

Classroom Companion: Business

Werner Liebrechts
Willem-Jan van den Heuvel
Arjan van den Born *Editors*

Data Science for Entrepreneurship

Principles and Methods
for Data Engineering, Analytics,
Entrepreneurship, and the Society

 Springer

Classroom Companion: Business

The Classroom Companion series in Business features foundational and introductory books aimed at students to learn the core concepts, fundamental methods, theories and tools of the subject. The books offer a firm foundation for students preparing to move towards advanced learning. Each book follows a clear didactic structure and presents easy adoption opportunities for lecturers.

Werner Liebregts • Willem-Jan van den Heuvel
Arjan van den Born
Editors

Data Science for Entrepreneurship

Principles and Methods for Data Engineering, Analytics,
Entrepreneurship, and the Society



Springer

Editors

Werner Liebrechts
Jheronimus Academy of Data Science (JADS)
's- Hertogenbosch, The Netherlands

Tilburg University
Tilburg, The Netherlands

Arjan van den Born
Jheronimus Academy of Data Science (JADS)
's- Hertogenbosch, The Netherlands

Tilburg School of Economics and
Management (TiSEM), Tilburg University
Tilburg, The Netherlands

Willem-Jan van den Heuvel
Jheronimus Academy of Data Science (JADS)
's- Hertogenbosch, The Netherlands

Tilburg University
Tilburg, The Netherlands

Section Editors

Willem-Jan Van den Heuvel
Jheronimus Academy of Data Science
(JADS)
's-Hertogenbosch, , The Netherlands

Damian A. Tamburri
Jheronimus Academy of Data Science
(JADS)
's-Hertogenbosch, The Netherlands

Florian Böing-Messing
Jheronimus Academy of Data Science
(JADS)
's-Hertogenbosch, The Netherlands

Werner Liebrechts
Jheronimus Academy of Data Science
(JADS)
's-Hertogenbosch, The Netherlands

Anne J. F. Lafarre
Jheronimus Academy of Data Science
(JADS)
's-Hertogenbosch, The Netherlands

ISSN 2662-2866

ISSN 2662-2874 (electronic)

Classroom Companion: Business

ISBN 978-3-031-19553-2

ISBN 978-3-031-19554-9 (eBook)

<https://doi.org/10.1007/978-3-031-19554-9>

© Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Data science is a fairly young but rapidly evolving scientific discipline. In the past two decades, it has attracted the attention of many scholars and practitioners alike, and so, one can find many educational and research programs focusing on data science nowadays. However, surprisingly, few of them make an explicit link with entrepreneurship, business development, or management more broadly. Even though a deeper understanding of pure data engineering and analytics topics is very important, data only become *truly* valuable when used for new value creation, for business and/or society at large.

This is a role that entrepreneurship-minded people typically fulfill. Like the late Joseph Schumpeter, the famous political economist (1883–1950), once said: “... the inventor produces ideas, the entrepreneur ‘gets things done’, ...” (Schumpeter, 1947, p. 152). Hence, the Schumpeterian entrepreneur turns inventions into economically viable business activities. Likewise, data entrepreneurs exploit opportunities that emerge from technological inventions in the data science domain. Think of improved ways to collect, store, and analyze data, which are then utilized for process, product, and service innovations.

The Jheronimus Academy of Data Science (JADS), a joint initiative of Tilburg University and the Eindhoven University of Technology, still is one of the very few initiatives that are truly multidisciplinary in nature. More specifically, at JADS, we not only conduct research and offer education at the intersection of the data science and entrepreneurship disciplines, but also help businesses of all sorts and sizes to take (more) advantage of the unprecedented opportunities that data science brings. No wonder there are many (former) colleagues among the authors who contributed to this book.

It is not an easy task to bring together so many people with a wide variety of backgrounds, but I am glad we succeeded. In my humble opinion, this book has become a coherent and very complete overview of the latest scientific knowledge on data science for successful, data-driven entrepreneurship (including corporate entrepreneurship forms). As such, it complements existing books by establishing a clear link between data science on the one hand and entrepreneurship on the other hand, while not forgetting the legal and ethical side of data usage.

I hope that you enjoy reading the book as much as we did writing and editing it. Do not hesitate to contact me about any issue related to the book.

Werner Liebrechts

's-Hertogenbosch, The Netherlands

Acknowledgments

The editors would like to thank all section editors and chapter authors for their invaluable contributions to this book. Without them, this book would not have been as complete and insightful as it is now. In particular, many thanks to Damian Tamburri as a co-editor of the Data Engineering section, and Florian Böing-Messing and Anne Lafarre for coordinating and editing the Data Analytics and Data and Society sections of this book, respectively. Your valuable input during regular alignment meetings and your clear and timely communication with us as well as the many contributing authors in your sections are very much appreciated. We sometimes had to put a little pressure on you, and at other times your patience was tested. The latter also holds for quite some of the contributing authors, who managed to deliver their manuscripts (way) earlier than some others. Here too, the negative impact of the Covid-19 pandemic should not be underestimated. We thank you for your flexibility and understanding and sincerely hope that you are as proud of the end result as we are. A final word of thanks goes out to Prashanth Mahagaonkar and Ruth Milewski, both of whom have always been very helpful and supportive during the entire writing and editing process.

About the Book

The textbook in front of you is the culmination of a years-long combined effort of many scholars coming from a large variety of disciplines. Therefore, we can rightly say that it is multidisciplinary in nature and, as such, offers you a comprehensive overview of the latest knowledge on a broad spectrum of related topics. The book follows a conceptual framework, bringing together the two disciplines of data science and entrepreneurship, which until recently have been treated fairly separately. This particular framework also forms the basis of research and education at the Jheronimus Academy of Data Science (JADS), a joint initiative of Tilburg University and the Eindhoven University of Technology (both located in the Netherlands), although framing and labeling may vary slightly. At JADS, among many other things, data science knowledge and skills are being utilized to explore and exploit entrepreneurial opportunities to create new value.

In essence, the conceptual framework nicely illustrates that one needs (at least basic) knowledge of the data engineering, data analytics, and data entrepreneurship disciplines, as well as the business and societal context in which all this happens (labeled data and society, think of legislation and ethics), in order to transform bright, data-driven ideas into value for business and/or society. Conceptually, this looks like a simple, linear process, but in practice, such processes are complex and highly dynamic with many feedback loops. Continuous iteration is expected to lead to improved if not optimal performance. Let us refer to Section 5 of the introductory chapter for a more detailed explanation of the conceptual framework that determines the structure of this book.

Clearly, the Data Engineering and Data Analytics sections cover all kinds of relevant topics in the data science domain, whereas the Data Entrepreneurship section contains accessible chapters about topics that are important for successful, data-driven entrepreneurship or business development. The subject areas in the Data and Society section touch upon the legal and ethical aspects, which are crucial for both data science and entrepreneurship, so along the entire value chain.

All chapters introduce a subject area and its fundamentals and provide a foundation for the reader to proceed to advanced learning. Even though all chapters are to some extent self-contained, we do not recommend to read them in isolation. For example, entrepreneurs and business developers need to know about (ethical) data management and governance and machine learning (ML) processes, in case they aim to deploy an accurate predictive ML model for improved, data-driven decision-making.

Most of the chapters include a number of recurring style elements, for example definition, example, and important. These are separate boxes to highlight certain parts of the main text body. All chapters start with a set of learning objectives, depicting your knowledge level after having read the chapter. One could test this level of knowledge by trying to answer the questions and/or by intensively discussing the points raised at the end of each chapter. Every chapter also provides a number of take-home messages, summing up its key takeaways.

This textbook is primarily intended for upper undergraduate or graduate students, who would like to combine data science and entrepreneurship knowledge and skills, in their pursuit of a role as entrepreneurial data scientist, data entrepreneur, or data-driven business developer upon graduation. At the same time, the book is also very interesting for practitioners, who would like to obtain a deeper understanding about how data science can be utilized for improved entrepreneurial or business performance. Finally, data science and entrepreneurship researchers will find the latest scientific knowledge concerning topics in their respective domains. Regardless of to which audience you belong, enjoy reading!

Werner Liebregts

Willem-Jan van den Heuvel

Arjan van den Born

's-Hertogenbosch, The Netherlands

Contents

1	The Unlikely Wedlock Between Data Science and Entrepreneurship	1
	<i>Arjan van den Born, Werner Liebrechts, and Willem-Jan van den Heuvel</i>	
I	Data Engineering	
2	Big Data Engineering	25
	<i>Damian Tamburri and Willem-Jan van den Heuvel</i>	
3	Data Governance	37
	<i>Indika Kumara, A. S. M. Kayes, Paul Mundt, and Ralf Schneider</i>	
4	Big Data Architectures	63
	<i>Martin Garriga, Geert Monsieur, and Damian Tamburri</i>	
5	Data Engineering in Action	77
	<i>Giuseppe Cascavilla, Stefano Dalla Palma, Stefan Driessen, Willem-Jan van den Heuvel, Daniel De Pascale, Mirella Sangiovanni, and Gerard Schouten</i>	
II	Data Analytics	
6	Supervised Machine Learning in a Nutshell	105
	<i>Majid Mohammadi and Dario Di Nucci</i>	
7	An Intuitive Introduction to Deep Learning	121
	<i>Eric Postma and Gerard Schouten</i>	
8	Sequential Experimentation and Learning	147
	<i>Jules Kruijswijk, Robin van Emden, and Maurits Kaptein</i>	
9	Advanced Analytics on Complex Industrial Data	177
	<i>Jurgen van den Hoogen, Stefan Bloemheuvel, and Martin Atzmueller</i>	
10	Data Analytics in Action	205
	<i>Gerard Schouten, Giuseppe Arena, Frederique van Leeuwen, Petra Heck, Joris Mulder, Rick Aalbers, Roger Leenders, and Florian Böing-Messing</i>	

III Data Entrepreneurship

11	Data-Driven Decision-Making	239
	<i>Ronald Buijsse, Martijn Willemsen, and Chris Snijders</i>	
12	Digital Entrepreneurship	279
	<i>Wim Naudé and Werner Liebrechts</i>	
13	Strategy in the Era of Digital Disruption	305
	<i>Ksenia Podoynitsyna and Eglé Vaznyté-Hünermund</i>	
14	Digital Servitization in Agriculture	331
	<i>Wim Coreynen and Sicco Pier van Gosliga</i>	
15	Entrepreneurial Finance	353
	<i>Anne Lafarre and Ivona Schoonbrood</i>	
16	Entrepreneurial Marketing	383
	<i>Ed Nijssen and Shantanu Mullick</i>	

IV Data and Society

17	Data Protection Law and Responsible Data Science	415
	<i>Raphaël Gellert</i>	
18	Perspectives from Intellectual Property Law	443
	<i>Lisa van Dongen</i>	
19	Liability and Contract Issues Regarding Data	461
	<i>Eric Tjong Tjin Tai</i>	
20	Data Ethics and Data Science: An Uneasy Marriage?	483
	<i>Esther Keymolen and Linnet Taylor</i>	
21	Value-Sensitive Software Design	503
	<i>Paulan Korenhof</i>	
22	Data Science for Entrepreneurship: The Road Ahead	523
	<i>Willem-Jan van den Heuvel, Werner Liebrechts, and Arjan van den Born</i>	

Editors and Contributors

About the Editors

Werner Liebrechts

is an Assistant Professor of Data Entrepreneurship at the Jheronimus Academy of Data Science (JADS, Tilburg University, the Netherlands) and a secretary of the Dutch Academy of Research in Entrepreneurship (DARE). His research focuses on how entrepreneurs and entrepreneurship scholars can leverage data science and AI for new value and new knowledge creation, respectively.

Willem-Jan van den Heuvel

is a Full Professor of Data Engineering at the Jheronimus Academy of Data Science (JADS, Tilburg University, the Netherlands) and the Academic Director of the JADS' Data Governance lab. His research interests are at the cross-junction of software engineering, data science and AI, and distributed enterprise computing.

Arjan van den Born

is a Full Professor of Data Entrepreneurship and the former Academic Director of the Jheronimus Academy of Data Science (JADS), a joint initiative of Tilburg University and the Eindhoven University of Technology, both located in the Netherlands. He is also the Managing Director of a Regional Development Agency (RDA) in the Utrecht region (the Netherlands).

Contributors

Rick Aalbers Radboud University, Nijmegen, The Netherlands

Giuseppe Arena Tilburg University, Tilburg, The Netherlands

Martin Atzmueller Osnabrück University, Osnabrück, Germany
Tilburg University, Tilburg, The Netherlands

Stefan Bloemheuel Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands
Tilburg University, Tilburg, The Netherlands

Florian Böing-Messing Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Arjan van den Born Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Regionale Ontwikkelingsmaatschappij (ROM) Regio Utrecht, Utrecht, The Netherlands

Ronald Buijsse Tilburg University, Tilburg, The Netherlands

Giuseppe Cascavilla Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Wim Coreynen Vrije Universiteit, Amsterdam, The Netherlands

Lisa van Dongen Tilburg University, Tilburg, the Netherlands

Stefan Driessen Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Robin van Emden Tilburg University, Tilburg, The Netherlands

Martin Garriga YPF, Neuquén, Argentina

Raphaël Gellert Radboud University, Nijmegen, the Netherlands

Petra Heck Fontys University of Applied Sciences, Eindhoven, The Netherlands

Willem-Jan van den Heuvel Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Tilburg University, Tilburg, The Netherlands

Jurgen van den Hoogen Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Tilburg University, Tilburg, The Netherlands

Maurits Kaptein Tilburg University, Tilburg, The Netherlands

A. S. M. Kayes La Trobe University, Melbourne, Australia

Esther Keymolen Tilburg University, Tilburg, the Netherlands

Paulan Korenhof Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, the Netherlands

Wageningen University & Research (WUR), Wageningen, the Netherlands

Jules Kruijswijk Tilburg University, Tilburg, The Netherlands

Indika Kumara Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Anne Lafarre Tilburg University, Tilburg, The Netherlands

Roger Leenders Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Tilburg University, Tilburg, The Netherlands

Frederique van Leeuwen Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands

Werner Liebrechts Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Tilburg University, Tilburg, The Netherlands

Majid Mohammadi Vrije Universiteit, Amsterdam, the Netherlands

Geert Monsieur Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Joris Mulder Tilburg University, Tilburg, The Netherlands

Shantanu Mullick Coventry University, Coventry, United Kingdom

Paul Mundt Adaptant Solutions AG, Munich, Germany

Wim Naudé RWTH Aachen University, Aachen, Germany

IZA Institute of Labor Economics, Bonn, Germany

Maastricht School of Management, Maastricht, The Netherlands

Ed Nijssen Eindhoven University of Technology, Eindhoven, the Netherlands

Dario Di Nucci University of Salerno, Fisciano, Italy

Stefano Dalla Palma Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Daniel De Pascale Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Sicco Pier van Gosliga Eindhoven University of Technology, Eindhoven, The Netherlands

Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Ksenia Podoyntsyna Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Eric Postma Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Mirella Sangiovanni Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

Ralf Schneider University of Stuttgart, Stuttgart, Germany

Gerard Schouten Fontys University of Applied Sciences, Eindhoven, The Netherlands
Fontys University of Applied Sciences, School of ICT, Eindhoven, The Netherlands

Ivona Schoonbrood AllianceBlock, Utrecht, The Netherlands

Chris Snijders Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands
Eindhoven University of Technology, Eindhoven, The Netherlands

Damian Tamburri Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands
Eindhoven University of Technology, Eindhoven, The Netherlands

Linnet Taylor Tilburg University, Tilburg, the Netherlands

Eric Tjong Tjin Tai Tilburg University, Tilburg, the Netherlands

Eglé Vaznyté-Hünermund Sandoz, Copenhagen, Denmark

Martijn Willemsen Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands
Eindhoven University of Technology, Eindhoven, The Netherlands



The Unlikely Wedlock Between Data Science and Entrepreneurship

*Arjan van den Born, Werner Liebrechts,
and Willem-Jan van den Heuvel*

Contents

- 1.1 Introduction – 2**
- 1.2 Defining Data Science
and Entrepreneurship – 3**
- 1.3 Towards a Definition of Data
Entrepreneurship – 4**
- 1.4 Processes of Data Science
and Entrepreneurship – 7**
 - 1.4.1 The Data Science Process – 7
 - 1.4.2 The Entrepreneurial Process – 10
 - 1.4.3 Comparing Data Science and
Entrepreneurial Processes – 13
- 1.5 The Data Entrepreneurship
Framework – 14**
- References – 18**

Learning Objectives

After having read this chapter, you will be able to:

- Comprehend how the two seemingly unrelated disciplines of data science and entrepreneurship are actually strongly connected.
- Recognize how the refinement of data (into information, knowledge, understanding, and wisdom) and data value are interconnected.
- Understand a number of basic process models underpinning both data science and entrepreneurship separately.
- Understand how one can bring together data science and entrepreneurship in a conceptual framework concerning data entrepreneurship.
- Know the structure of the book in terms of its main sections, the order in which they appear, and their relationships.

1.1 Introduction

This book is about the linkages and integration of two unlikely academic bedfellows, viz. data science and entrepreneurship, shortly alluded to as data entrepreneurship. From the point of view of the ordinary practitioner and business person, however, the amalgamation of these two areas is only natural. Their question seems rather mundane: How to create value with data? However, for an academic, this devilish simple question of value creation comes with a myriad of dimensions and numerous challenges. How can we merge these almost diametrically opposing strands of science and build a common framework, which leverages both academics and practitioners alike to generate *real* value with data?

These have been the questions we have been dealing with for over the past five years in establishing a unique collaboration between a university of technology (the Eindhoven University of Technology) and one of social sciences (Tilburg University), named the Jheronimus Academy of Data Science (JADS) and located in 's-Hertogenbosch, the Netherlands. This book not only gives an overview of this exciting journey towards shaping and aligning education and research revolving around data entrepreneurship, but also shares the latest insights regarding plenty of underlying research fields. Now let us go onward defining and reviewing this substrate of data entrepreneurship.

The remainder of this introductory chapter is structured as follows. In ► Sect. 1.2, we start with defining what both data science and entrepreneurship are. Only then can we move on trying to integrate both disciplines and defining the concept of data entrepreneurship in ► Sect. 1.3. In ► Sect. 1.4, we extensively discuss processes of data science and entrepreneurship, respectively, and conclude that, despite a number of key differences, a few important elements are strikingly similar. In ► Sect. 1.5, we introduce our newly developed data entrepreneurship framework, which synthesizes various models that have been discussed in the previous section. The remainder of this book will be structured in accordance with this conceptual framework. ► Section 1.6 provides conclusion.

1.2 Defining Data Science and Entrepreneurship

The integration of the fields of data science and entrepreneurship is not easy at all. While data science is widely regarded as an emerging yet already significant discipline (Van der Aalst, 2016), it is sometimes also simply perceived as a collection of established scientific disciplines including statistics, data mining, databases, and distributed systems. Entrepreneurship is said to be omnipresent, but that also makes it hard to clearly demarcate what can be deemed entrepreneurial and what not.

Now what have the theory and practice of entrepreneurship to offer to data science and vice versa? At a first glance, it seems important to determine whether or not value creation is already part of the definition of data science. At the same time, while the scientific and practical definitions of data science diverge, the disparity between the scientific and practical definitions of entrepreneurship is likewise tidy.

Let us start with considering Wikipedia's definition of entrepreneurship.

Definition of Entrepreneurship (Wikipedia)

The process of designing, launching, and running a new business, which is often initially a small business.

The people who create and manage these (small) businesses are called entrepreneurs. Furthermore, the business dictionary defines entrepreneurship as follows:

Definition of Entrepreneurship (Business Dictionary)

“The capacity and willingness to develop, organize, and manage a business venture along with any of its risks in order to make a profit.”

While the everyday definitions of entrepreneurship stress the importance of business venturing, risk-taking, and profit seeking, scientific definitions see entrepreneurship in a more abstract fashion. In a historical overview of the academic field of entrepreneurship, Bruyat and Julien (2001) state the following: “The scientific object studied in the field of entrepreneurship is the dialogic between individual and new value creation, within an ongoing process and within an environment that has specific characteristics” (p. 165). In this definition, there is no mentioning of business creation, nor is there any notion of risk taking, profit, or even management. In their seminal article, Shane and Venkataraman (2000) argue that entrepreneurship researchers ought to study individuals, “opportunities,” and their fit, i.e., the individual-opportunity nexus. In short, it is the task of the entrepreneurship researcher to understand more about the entrepreneurial process, i.e., the process how entrepreneurs discover, evaluate, and exploit opportunities.

While it seems to be appealing to equate the process of entrepreneurship to the value-generating process mentioned often in definitions of data science, this is,

however, too hasty a conclusion. In fact, it is certainly not the case that all entrepreneurs add value. Baumol (1990) already famously argued that entrepreneurship can be productive, unproductive, and destructive. While the entrepreneur ideally plays an innovative, constructive, and therefore productive role in society, this is not always the case. Many entrepreneurs fail to contribute to society, and some entrepreneurs may even be parasitical and damaging to economy and society.

► **Example: Productive, Unproductive, and Destructive Entrepreneurship**

Entrepreneurship is deemed productive if an entrepreneur creates new value (for society). This is, for example, the case when innovations are being developed. Entrepreneurship is unproductive if an entrepreneur engages in rent-seeking. Rent-seeking is an attempt to obtain economic rent by manipulating the social or political environment in which the entrepreneurs operate. A good example of destructive entrepreneurship is organized crime, something which often requires a strong entrepreneurial mindset, but these activities are clearly at the detriment of society. ◀

The distinction between productive, unproductive, and even destructive forms also applies to the concept of data entrepreneurship; data can be translated into something beneficial to society but may also be used as a destructive force.

Up till now, we have discussed everyday definitions of entrepreneurship, but as we are working towards a definition of data entrepreneurship, we need to combine the definition of entrepreneurship with a proper definition of data science. Below, we therefore give the definition of data science as found on Wikipedia.

— **Definition of Data Science (Wikipedia)** —

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data and apply knowledge and actionable insights from data across a broad range of application domains.

This definition stresses a couple of interesting elements. First, it emphasizes the interdisciplinary character of data science. Data science is a mix of various sciences, such as computer science, mathematics, and statistics. Second, the definition accentuates the importance of applying these underlying scientific methods and processes towards certain application areas. As we will see in the next section, in our case, this application area is, first and foremost, entrepreneurship. Third, the definition highlights the importance of obtaining knowledge and actionable insights from data, which can be both structured and unstructured (at least, in the first instance).

1.3 Towards a Definition of Data Entrepreneurship

When considering the various definitions of data science and entrepreneurship, one may wonder how we should define data entrepreneurship. Before we give a basic definition of data entrepreneurship (this will be further elaborated and refined

once we have attained more insights in her substrate), we will now review the fundamental fabric of data science, i.e., data.

The classical data-information-knowledge-wisdom (DIKW) hierarchy of Ackoff (1989) may help us to better appreciate data entrepreneurship as an emerging scientific discipline at the intersection of data science and entrepreneurship.

► **Important: Ackoff's DIKW Hierarchy**

According to Ackoff (1989), data are the product of factual observations and are of no value until they are processed into a usable form and become information. Information is contained in descriptions, or in answers to who, what, where, and when questions. Knowledge, the next layer, further refines information by transforming information into instructions and thus consists of answers to how questions. In turn, understanding is contained in explanations and in answers to why questions. For Ackoff, wisdom means an ability to understand the long-term consequences of any act, which includes the ability to assess and correct for all sorts of errors (i.e., evaluated understanding). ◀

Or, in other words:

- » An ounce of information is worth a pound of data. An ounce of knowledge is worth a pound of information. An ounce of understanding is worth a pound of knowledge. An ounce of wisdom is worth a pound of understanding. Russell Lincoln Ackoff (1999)

At around the same time as Ackoff, Zeleny (1987) also introduced a data taxonomy. He even suggests an extra, rather arcane, hierarchical level, viz. enlightenment. We compare Zeleny's model with that of Ackoff in ■ Table 1.1.

■ **Table 1.1** Comparing Ackoff's and Zeleny's hierarchies

	Ackoff (1989)	Zeleny (1987)
Data	Symbols that represent the properties of objects and events	Know nothing
Information	Answers to who, what, where, and when questions	Know what
Knowledge	Answers to how questions	Know how
Understanding	Answers to why questions	N/A
Wisdom	Evaluated understanding	Know why
Enlightenment	N/A	Attaining a sense of truth, of right and wrong, and having it socially accepted and respected

Note: Table compiled by authors

Whereas Ackoff's and Zeleny's pyramid-like taxonomic structures appear very similar, in reality, scholars deeply disagree about the fundamental definitions of data, knowledge, and wisdom (Zins, 2007). However, at the same time, in her overview of the literature on data, information, and knowledge, Rowley (2007) asserts that most scholars do adopt the notion of such a hierarchy or taxonomy to conceptually link the notions of data, information, and knowledge.

The higher up in the DIKW model, the more value can be attached. After all, information adds more value to data, as it makes it more structured and organized. In turn, knowledge adds value to information, since it can be used to actually address a particular business opportunity or problem. Wisdom pertains to the highest value distillation, explaining when to use which method to resolve a business problem.

At the same time, there is still a lot of misconception about the exact nature of the transformation process, i.e., how can we convert data into value? Nevertheless, many scholars acknowledge that the data transformation process implies that human input becomes more important in the upper transformational stages of the hierarchy, while the value of algorithms and AI appear less important.

Given the everyday definitions of data science and entrepreneurship, and the hierarchy of information, what would be a great starting definition of data entrepreneurship? Clearly, such a definition should encompass the ability to transform data into information, knowledge, and actionable insights to support the design, launch, or running of a new business venture. This information should improve the understanding of the business. It may improve the capacity of an organization to run a business, it may assist in discovering and managing the entrepreneurial risks, or it may lead to enhanced profits.

Definition of Data Entrepreneurship

Data entrepreneurship is an interdisciplinary field that lies on the crossroads of data science and entrepreneurship. It actively applies the scientific methods, processes, algorithms, and systems of data science to develop, organize, and manage a business venture along with any of its risks in order to make a profit.

In other words, data entrepreneurship is the process of new value creation by refining data into information, knowledge, understanding, and/or wisdom in order to exploit a business opportunity. This definition is still quite general. It encompasses the use of data in any business to support its launch and/or existence. In the next section, we will see that the nature of data entrepreneurship will change depending on the size of the enterprise, its growth path, the exact application domain, and the dynamics of the business environment. For instance, in pretty stable, operational settings with loads of data, data entrepreneurs have different challenges than in more dynamic environments with scarce or limited datasets. A change of scenery also implies the usage of different methods, processes, and techniques to find the appropriate answers. To get a better understanding on how this works, we will discuss the processes of data science and entrepreneurship in more detail below.

1.4 Processes of Data Science and Entrepreneurship

Now that we have defined what data science is, what entrepreneurship is, and how both disciplines come together in a concept called data entrepreneurship, we can move on to highlighting various well-known processes of data science and entrepreneurship, respectively. Subsequently, we will compare them in order to find common ground as well as some striking differences.

1.4.1 The Data Science Process

The ideas behind the DIKW pyramid have been translated into many more detailed and dynamic process views of data science. Notably, Hanspeter Pfister and Joe Blitzstein (2013) promoted a rather simplistic, yet very practical process of data science in their Harvard CS109 class on Data Science. In their view, the following five essential steps are taken in any data science assignment:

1. Ask an Interesting Question
2. Get the Data
3. Explore the Data
4. Model the Data
5. Communicate and Visualize the Data

Probably the best-known process model used to resolve data science assignments, and thus to refine data into something (more) valuable, is the CRISP-DM process (Wirth & Hipp, 2000, also see ■ Fig. 1.1). Here, CRISP-DM stands for CROSS-Industry Standard Process for Data Mining. This process model originates from industry and has been initially developed by a consortium consisting of Daimler-Chrysler, SPSS, and NCR. The CRISP-DM process comprises the following six stages:

1. **Business Understanding**—This initial phase focuses on understanding the project objectives and requirements from a business perspective and then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the business objectives.
2. **Data Understanding**—The data understanding phase starts with an initial data collection and proceeds with activities to get more familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.
3. **Data Preparation**—The data preparation phase covers all activities to construct a structured and well-understood dataset from the initial raw data.
4. **Modeling**—During this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values with the highest accuracy.
5. **Evaluation**—At this stage, the model (or models) obtained are more thoroughly evaluated, and the steps executed to construct the model are reviewed to be certain that it properly achieves the business objectives.

■ Fig. 1.1 CRISP-DM cycle.
(Source: Wirth and Hipp (2000))

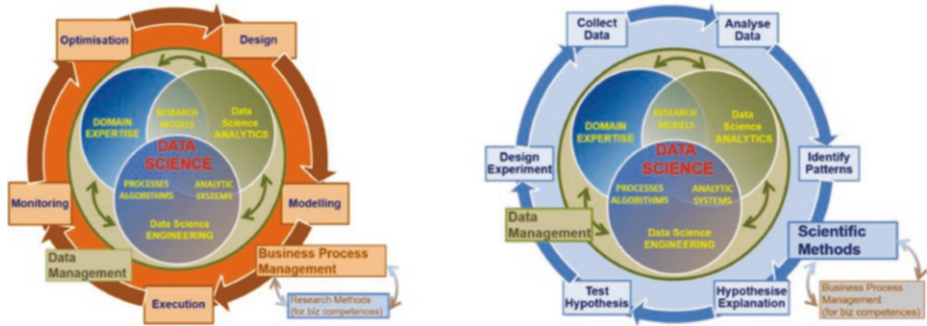


6. **Deployment**—Creation of the model is generally not the end of the project; in many cases, it is considered the start of a lengthy software engineering exercise to factor the model into the existing information systems landscape. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can effectively interpret and apply.

When one compares the above two process views, one can notice the similarity with the basic DIKW pyramid as both approaches try to translate rough data into information that is valuable. But there are some interesting differences. Let us now consider the three most relevant differences.

Firstly, the CRISP-DM process prioritizes a good business understanding at the beginning, while Pfister and Blitzstein (2013) emphasize the importance of a research question at the beginning of the process. Secondly, the CRISP-DM process ends with the actual deployment of a model, whereas Pfister and Blitzstein (2013) stress the importance of communication and visualization of results. Finally, the CRISP-DM process accentuates the dynamic character of data science. It is often an ongoing process of exploration, and oftentimes, after concluding that a model does not perform as desired, the whole process needs to be rerun.

Another well-known model that is largely complementary to the above process models is the EDISON Data Science Competence Framework (CF-DS) by Demchenko et al. (2016) (also see ■ Fig. 1.2). This framework does not emphasize a process view of data science, but rather gives a list of soft and hard skills needed to be a professional data scientist. By doing so, it gives some interesting clues about



■ Fig. 1.2 EDISON Data Science Competence Framework. (Source: Demchenko et al. (2016))

the process of data science. According to the CF-DS, the skills needed by a data scientist consist of the following three core competence groups:

1. **Data Engineering:** Data and software engineering, distributed computing, batch and stream processing data architectures, and data warehousing
2. **Data Analytics:** Descriptive, diagnostic, predictive, and prescriptive analytics, with techniques ranging from classical mathematics and statistics to more state-of-the-art machine learning and deep learning
3. **Domain-Specific Competences:** Including domain knowledge and expertise

On top of these three core competence groups, there are two additional, supporting competence groups identified by the EDISON project:

4. **Data Management and Data Governance:** Strategic and tactic plans for the storage and maintenance of data collections, and data quality assurance (roles needed: data custodians and data stewards)
5. Either **Scientific or Research Methods** (for academics) or **Business Process Management** (for practitioners)

Interestingly, Demchenko et al. (2016) describe these two supportive competence groups as dynamic processes rather than static competences. For instance, the research process includes the following eight basic phases:

1. Design Experiment
2. Define Research Questions
3. Collect Data
4. Analyze Data
5. Identify Patterns
6. Hypothesize Explanation
7. Test Hypothesis
8. Refine Model (and Start New Experiment Cycle)

Indeed, these eight steps describing the scientific method are not much different from the five steps mentioned by Pfister and Blitzstein (2013) in their Harvard

CS109 class. There is just somewhat extra attention to experimental design and hypothesis testing.

According to the EDISON framework, the business process management life cycle for data science includes the following six straightforward phases:

1. **Define the Business Target:** Such as the product/market combination.
2. **Design the Business Process:** As a logically structured collection of business activities.
3. **Model/Plan:** Develop executable business process models, including planning and scheduling.
4. **Deploy and Execute:** Deploy the business processes on, for example, a business process or workflow engine.
5. **Monitor and Control:** Exploit business process activity monitors to oversee and measure progress against performance indicators.
6. **Optimize and Redesign:** Continuously adapt and optimize the business process to improve its performance.

Here, we see some overlap with the CRISP-DM framework with its emphasis on business understanding, deployment of models, and continuous improvement.

Now let us try to map the DIKW and the above process models to be able to better appreciate what data science in the context of value creation entails. Firstly, data engineering methods are applied to turn data into information by curating, structuring, and mapping it into a harmonized format that is easily interpretable. Secondly, data analytics methods, such as Bayesian statistics and machine learning, may then help to distill knowledge from information, e.g., through clustering and classification of data. Novel approaches such as AutoML—aimed at the automation of machine learning—and neuro-evolution may then be mapped to the highest level of value (i.e., wisdom). The loops in the process models pertain to the highly iterative nature of data value refinement. For example, training deep learning models is an immensely repetitive exercise, involving many cycles with feedback loops to improve the accuracy of knowledge and/or wisdom obtained.

1.4.2 The Entrepreneurial Process

In the beginning of this chapter, we have already concluded that the definitions of data science and entrepreneurship are rather ambiguous. We argued that the typical practitioner and academic definitions of entrepreneurship are quite diverse and that some of the definitions of data science explicitly included the creation of value, while other definitions were not so adamant. To get a deeper understanding and to uncover the true links between data science and entrepreneurship, we will now focus on various entrepreneurial processes.

In the work of Moroz and Hindle (2012), the authors have conducted a systematic review of studies on the entrepreneurial process analyzing 32 different process models. Surprisingly, only 9 of the studied 32 process models were actually based on an empirical study, while the vast majority could be described as unsubstantiated artifacts. On the positive side, the study reveals that there are six core elements

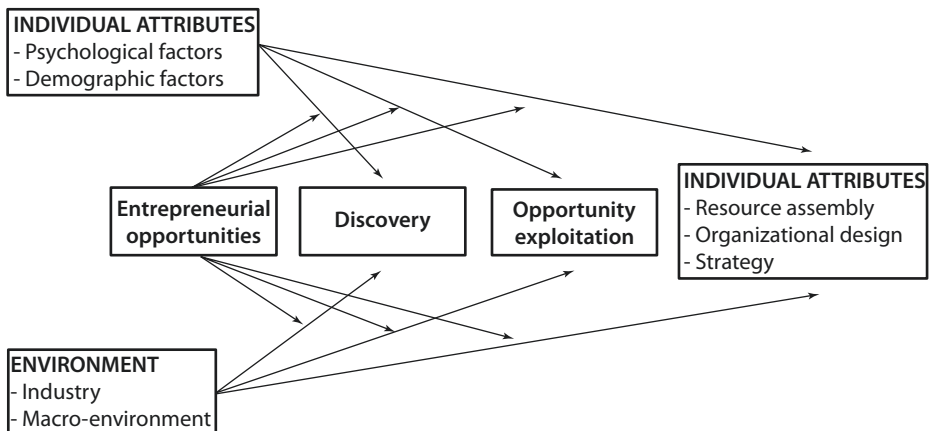
that are to be found in all process models of entrepreneurship. These are the following:

1. There needs to exist a match between individual and opportunity: Not every opportunity can be processed by every hopeful entrepreneur.
2. There is a need to critically assess the transformative and disruptive value of knowledge.
3. There is an emphasis on the creation of new business models in contrast to enhancing existing business models.
4. Timeliness matters: Opportunities do not last forever, and market receptiveness can differ over time.
5. Action matters: Formulating a plan is merely part of the process, and action by the entrepreneurs is critical.
6. Context matters: Understanding the broader environment is imperative.

A particularly relevant model of the entrepreneurial process is Shane's (2003) model that we will shortly describe below as a prototypical example of the models of entrepreneurial venture (also see ■ Fig. 1.3).

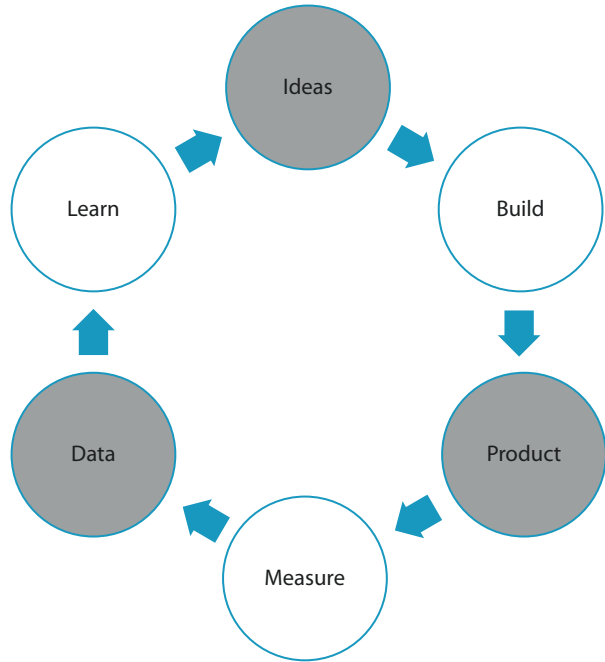
According to Shane (2003), the individual-opportunity nexus is of crucial importance. Here, it is the combination of environment and individual that determines the start of the entrepreneurial process and its subsequent course. Moreover, Shane is adamant to show that entrepreneurship has several distinct stages, viz. from opportunity discovery to exploitation of opportunities and the organization thereof. Interestingly enough, entrepreneurial success is not a *conditio sine qua non* for the entrepreneurial process.

In the last decade, "The Lean Startup" by Eric Ries (2011) has caught a tremendous amount of attention. The book prescribes a set of practices that aim to help entrepreneurs increase their odds of building a successful startup. Inspired by the Lean Six Sigma methodology, it aims to find waste in the business development



■ Fig. 1.3 Shane's model of the entrepreneurial process. (Source: Shane (2003))

■ **Fig. 1.4** The build-measure-learn loop of the lean startup method. (Source: Ries (2011))



process and weed this out. Ries (2011) defines a startup as a human institution that operates in conditions of extreme uncertainty (p. 8).

While any business operates under uncertain conditions, in the case of a startup, these uncertainties explode. There is typically a great uncertainty about the value proposition, the target customers, the pricing, the business model, the organization, etc. To cope with this enormous level of fuzziness and uncertainty, Ries (2011) proposes the adoption of rigorous scientific methods. The startup shall devise and run short-term, small-scale experiments that either support or reject the hypotheses underlying the business model. This iterative process continues in a loop where startups build stuff (preferably in the form of the so-called minimum viable product, or MVP), measure if it gains traction, and learn from this experience (also see ■ Fig. 1.4).

Since the publication of “The Lean Startup,” the principles of this method have been widely adopted, but academic scrutiny of this method remains relatively scarce (Shepherd & Gruber, 2020). In a scientific reflection, Frederiksen and Brem (2017) find strong support for many of its underlying principles, such as the early involvement of users and the iterative development process. However, the academic support for embracing experimentation and prototyping in startups is not yet as strong. Some have even shown that the lean startup method may actually be harmful (Mollick, 2019). For example, Felin et al. (2020) state that a focus on getting fast feedback from customers may lead to incremental improvements only instead of disruptive innovations. Prominent entrepreneurship scholars therefore promote

further research, for example into the design of experiments in relation to the further development of nascent businesses (Frederiksen & Brem, 2017; Shepherd & Gruber, 2020).

1.4.3 Comparing Data Science and Entrepreneurial Processes

To improve our understanding of the linkages between entrepreneurship and data science, we will now discuss the process models underpinning them both in more detail. Does a typical data science process relate to the entrepreneurial process?

If we carefully consider the processes of both data science and entrepreneurship, we can see some overlap, but also some important differences in our quest to reconcile them into one process model of data entrepreneurship (also see ■ Table 1.2).

If we look at the similarities, we first notice that both processes tend to be very iterative with many steps going back and forth. In data science, this will often lead to adapted research questions and methods, while in startups, this will lead to pivots which may lead to novel value propositions. Secondly, and intriguingly, in both processes, the creation of value is the overriding objective. Although what value exactly means remains rather vague in both cases. Finally, both data scientists and data entrepreneurs regard data as the key resource that is indispensable to create real value.

There are also a number of areas where the data science and entrepreneurial processes clearly overlap. Firstly, both the data scientist and the data entrepreneur embrace the scientific method to further learning. The data scientist articulates questions, formulates hypothesis, sets up experiments, and tests the hypothesis to get a better understanding of the “true” world or model. The data entrepreneur does roughly the same, but applies these techniques to test business aspects, such as the value proposition, the pricing, the target audience, etc. The aim of the data entrepreneur is to minimize uncertainty before putting in extra time and effort.

■ Table 1.2 Comparing data science and entrepreneurial processes

Largely similar	Somewhat similar	Significantly different
Dynamic, iterative processes	Adoption of scientific method	New business creation versus obtaining insights
Ultimate goal is value creation	Action matters	Radical new design versus optimization
Data as a unique asset	Time matters	Role of the entrepreneur
		Organizing for implementation

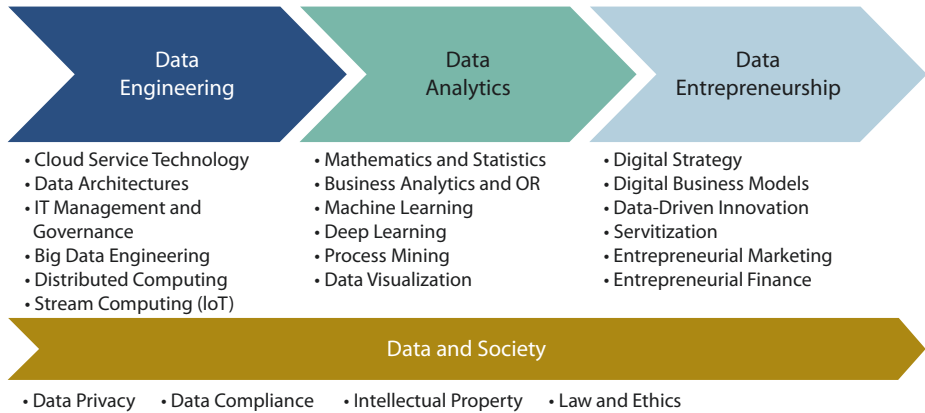
Note: Table compiled by authors

Moreover, where data scientist experiments are typically based on big data and obtaining quantitative insights, the experiments of the data entrepreneur are both large-scale quantitative and small-scale qualitative. Secondly, both processes are action based. It is not enough to design something or come up with a plan. The data scientist and the data entrepreneur are both action oriented. They both engage in action (*do* stuff), and, if possible, come up with novel insights and solutions. Thirdly, time and timing are essential. This is especially the case for the data entrepreneur, who needs to explore the market, as success is determined not only by the characteristics of the product and service itself, but also partly by the timing strategy. But also, the data scientist understands that models need to be dynamic and continuously trained and adapted, as new data may require new models and algorithms to be effective.

Yet, there are a number of significant differences between processes of data science and entrepreneurship. Where data science processes tend to focus on using data to obtain better insights or, at most, to build novel products or services, the most important goal of entrepreneurship is to build a new venture from scratch (Timmons et al., 1994). The interest of the entrepreneur goes typically beyond obtaining new insights or developing a new product or service. The data entrepreneur wants to develop or scale up a venture. This could be a startup, or it could be a corporate venture or even a social venture, but the objective goes beyond simple insights. Moreover, where the typical data scientist merely aims to optimize an existing process or way of working, the data entrepreneur wants to bring something new and innovative to the world. The data entrepreneur aims to disrupt the current way of working in the various industries. Partly because the purpose of the data entrepreneur is so far-reaching, the role of the entrepreneur (or members of the entrepreneurial team) is very important. There thus needs to be a fit between the data entrepreneur and the opportunities offered by data and digital technologies. Finally, where the data scientist often stops when the insight has been obtained or the product or service has been deployed, the work of the data entrepreneur is just beginning. The entrepreneur has to make sure that the venture is properly managed or organized. This goes beyond the standard methods for data and IT governance and management, and the typical approaches to software engineering and software development such as Scrum (Schwaber & Beedle, 2002) and DevOps (Hüttermann, 2012). A true data entrepreneur needs to go beyond these IT management methods and think thoroughly about the strategy, the business model, how to obtain funding (entrepreneurial finance), how to brand and market the venture and its services (entrepreneurial marketing), how to attract users, how to price the products and services, how to deal with ethical and legal issues, etc.

1.5 The Data Entrepreneurship Framework

Based on the above considerations, we will now introduce our data entrepreneurship framework (henceforth DEF), on which the Jheronimus Academy of Data Science (JADS) has grounded its academic programs (n.b., both education and



■ Fig. 1.5 Data entrepreneurship framework (DEF) of JADS. (Note: Authors' own figure)

research). The framework synthesizes a competence model and a high-level process model and is depicted in ■ Fig. 1.5.

The DEF of JADS is pretty similar to many existing data science process models but extends them by paying ample attention to the competences needed to build a new, data-driven venture. Of course, as with any model, it is a simplification of reality. As we have observed before, real-life data science processes and entrepreneurial processes are highly dynamic and complex, where learnings lead to ongoing changes in statistical models as well as business models. For instance, if models or services need to be deployed, the data engineers will come into play again, just like individuals concerned with data management and governance. Moreover, it can be the case that new insights may lead to better fitting algorithmic models and statistics. So, while this model may be a waterfall model in disguise at first sight, in reality, this model is highly iterative in nature, with many feedback loops, embracing change and continuous improvement.

While most aspects of our framework can be found in the earlier process models of data science and entrepreneurship, we see one particular type of activities that is not explicitly addressed in any of the earlier models, namely those related to the societal (and business) context (for example, law and ethics). As with any new and powerful technology, one has to think about the ethical side of these technologies. Ethical principles that guide our behavior are becoming evermore important. Related to ethics, law is also becoming an important aspect. More and more, the rules of the game, especially around data protection and privacy (also see the GDPR) and intellectual property (IP) of data and algorithms, determine the outcomes of the game and which value is created and destroyed. The ancient Greek philosophers already understood that to really understand a technology one sometimes needs to take a step back and observe the essence of technology from a distance.

Let us now revisit the key components of the framework and explain how they are addressed in the remainder of this book.

Our basic framework starts with all the competences to create valuable data (i.e., information) from raw data, i.e., **Data Engineering**. This entails getting the basic infrastructure in place that connects various data streams and making sure that the data is of great quality and that all the management procedures are in place to safeguard the security and integrity of the data. This is typically the domain of a data engineer. Data engineering (or big data engineering), including its ramifications for the data entrepreneur, is further explored in **Part I** of this book.

Right after, we find a section on **Data Analytics**. Here, data scientists use mathematics and various forms of statistics to uncover (probabilistic) patterns in the data. Numerous techniques are used, from machine learning to deep learning, and from process mining to Bayesian network analysis. In this phase, one also needs to select proper data science methods that best fit to the hypothesis (or hypotheses) to be validated and/or business problem(s) to be resolved. This phase roughly corresponds to the translation of information into knowledge in the DIKW hierarchy. More detailed materials on descriptive, diagnostic, predictive, and prescriptive data analytics methods, including some real-life use cases that apply them, can be found in **Part II** of this book.

The third section covers a variety of topics related to **Data Entrepreneurship**. The section starts with a broad overview of data-driven decision-making, puts emphasis on its benefits, and discusses many different forms and processes of such decision-making. Subsequently, we not only introduce different forms of data (or digital) entrepreneurship, including digital servitization, but also discuss some of the most important aspects of owning and managing a successful data-driven business, such as strategy development and implementation, finance, and marketing (and sales). The entrepreneurship tenet of data entrepreneurship is further explored and illustrated in **Part III** of this book.

The above three phases—all of them separate sections in this book—are embedded in society—the missing strain in most of the existing process models. The fourth and last section of this book, which is essentially cutting through all three other phases, thus entails **Data and Society**. This block addresses societal aspects, including—but not restricted to—legal and ethical issues (e.g., how to deal with intellectual property in an exceedingly international context, in which data is shared and traded), data compliance, data protection and privacy, and philosophical underpinnings of data science. These pervasive, substantive societal aspects are treated and further examined in **Part IV** of this book.

Conclusion

In this chapter, we have given a definition of data entrepreneurship and showed how it relates to the definitions of data science and entrepreneurship. We have discussed the processes of data science and entrepreneurship in more detail and outlined how they relate to each other. This, amongst other things, showed that while there is a large overlap between data science and entrepreneurship, such as its common goal to create new value, there are also some significant differences, especially when the focus shifts from obtaining insights to creating new businesses. This all implies that

data entrepreneurship is a new skill, not often taught in the curricula of current data science programs, a skill that embraces radical change and uses data to change the world rather than to optimize the world (bringing about incremental change). Therefore, to use data to really transform businesses and society, schools need to teach about data entrepreneurship: about new, digital business models, about digital strategy, about digital forms of organization, about data-driven marketing and finance, etc. Moreover, data entrepreneurship cannot be taught in a standalone manner. The ability to transform business and society with data requires an in-depth understanding not only of data entrepreneurship, but of data engineering, data analytics, and societal context (amongst others, determining ethical norms and values) as well. Only by integrating these various disciplines and competences can data be converted into transformational new business activities.

Discussion Points

1. As we have seen, it is not so straightforward to properly define the concepts of data science and entrepreneurship. Would you agree with one of our conclusions that both concepts at least seem to concern processes aimed at new value creation? Argue why (not).
2. Suppose that you have access to sensor data of the manufacturing process of a particular product. How can these data be made valuable? What is needed for these data to reach the levels of information, knowledge, understanding, and wisdom in Ackoff's hierarchy? Be precise.
3. Now knowing how to define data entrepreneurship, try to come up with at least three good examples of data entrepreneurs (or organizations engaging in data entrepreneurship). Explain why these entrepreneurs (or organizations) meet the definition of data entrepreneurship.
4. Back to the example of having access to sensor data of the manufacturing process of a particular product. What part of handling these data would be considered data engineering, and what data analytics? At what point would you call using these data a typical example of data entrepreneurship?
5. Throughout the entire value chain, so from data engineering to data entrepreneurship via data analytics, it is very important to keep all sorts of legal and ethical issues in mind and adhere to the prevailing rules and regulations (for example, regarding data protection and privacy). Name and explain at least two of such issues and discuss what they could entail in each of the three stages of the conceptual framework presented in ► Sect. 1.5.

Take-Home Messages

- Both data science and entrepreneurship are defined in many different ways, but both at least seem to concern processes aimed at new value creation.
- One can only create new value with data if these data are refined into information, knowledge, understanding, or even wisdom.

- Data entrepreneurship actively applies the scientific methods, processes, algorithms, and systems of data science to develop, organize, and manage a business venture along with any of its risks in order to make a profit.
- The CRISP-DM and CF-DS are well-known examples of process models of data science, and the lean startup method is a renowned and nowadays often applied process model of entrepreneurship.
- A newly developed conceptual data entrepreneurship framework brings together common elements in process models of both data science and entrepreneurship and explicitly adds the societal and business context.
- The remainder of this book is structured in accordance with the aforementioned conceptual framework and, hence, includes sections on data engineering, data analytics, data entrepreneurship, and data and society, respectively.

References

- Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16(1), 3–9.
- Ackoff, R. L. (1999). *Ackoff's best*. John Wiley & Sons.
- Baumol, W. J. (1990). Entrepreneurship: Productive, unproductive, and destructive. *The Journal of Political Economy*, 98(5), 893–921.
- Bruyat, C., & Julien, P. A. (2001). Defining the field of research in entrepreneurship. *Journal of Business Venturing*, 16(2), 165–180.
- Demchenko, Y., Belloum, A., Los, W., Wiktorski, T., Manieri, A., Brocks, H., Becker, J., Heutelbeck, D., Hemmje, M., & Brewer, S. (2016). EDISON Data Science Framework: A foundation for building data science profession for research and industry. In: *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*. pp. 620–626.
- Felin, T., Gambardella, A., Stern, S., & Zenger, T. (2020). Lean startup and the business model: Experimentation revisited. *Long Range Planning*, 53(4), 101889.
- Frederiksen, D. L., & Brem, A. (2017). How do entrepreneurs think they create value? A scientific reflection of Eric Ries' lean startup approach. *International Entrepreneurship and Management Journal*, 13(1), 169–189.
- Hüttermann, M. (2012). *DevOps for developers: Integrate development and operations, the Agile way*. Apress.
- Mollick, E. (2019). What the lean startup method gets right and wrong. *Harvard Business Review*, 10, 1–4.
- Moroz, P. W., & Hindle, K. (2012). Entrepreneurship as a process: Toward harmonizing multiple perspectives. *Entrepreneurship Theory and Practice*, 36(4), 781–818.
- Pfister, H., & Blitzstein, J. (2013). Course: CS109 data science.
- Ries, E. (2011). *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses*. Crown Business.
- Rowley, J. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science*, 33(2), 163–180.
- Schwaber, K., & Beedle, M. (2002). *Agile software development with scrum*. Prentice Hall.
- Shane, S. A. (2003). *A general theory of entrepreneurship: The individual-opportunity nexus*. Edward Elgar Publishing.
- Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *Academy of Management Review*, 25(1), 217–226.
- Shepherd, D. A., & Gruber, M. (2020). The lean startup framework: Closing the academic-practitioner divide. *Entrepreneurship Theory and Practice*, 1–31. <https://doi.org/10.1177/1042258719899415>

- Timmons, J. A., Spinelli, S., & Tan, Y. (1994). *New venture creation: Entrepreneurship for the 21st century*. Irwin.
- Van der Aalst, W. (2016). Data science in action. In W. Van der Aalst (Ed.), *Process mining* (pp. 3–23). Springer.
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pp. 29–39.
- Zeleny, M. (1987). Management support systems: Towards integrated knowledge management. *Human Systems Management*, 7(1), 59–70.
- Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, 58(4), 479–493.

Data Engineering

*Willem-Jan van den Heuvel and
Damian Tamburri*

Over the past decade, big data has been in the spotlights of the scientific, business, and governmental world. Big data can be basically characterized as (raw) data that comes in larger volumes than ever before, at unprecedented levels of speed, and with ever-growing variety. With the uptake of novel big data processing technologies and architectures, the engineering of data-intensive applications has become a daunting task. This task involves engineering, technical, organizational, and business-driven considerations within a huge, swiftly expanding design space.

At the same time, data engineering plays a pivotal role in entrepreneurship and lays the low-level “plumbing” work for data analytics. On the one hand, data engineering helps to unlock, integrate, refine, and process big data sources so they can be used for advanced data analytics. On the other hand, big data engineering is a valuable tool for data entrepreneurship to go from experimental and scientifically sound AI/ML models to well-engineering products and services that can be deployed and operated in production environments.

This module serves as a general introduction to the fundamentals of the data-intensive design process, focusing on the design constructs, patterns, and experimentation involved thereto, as well as practical examples of the process itself.

In ► Chap. 2, entitled *Big Data Engineering*, we firstly lay the foundation of data engineering, explaining the basic principles, concepts, and technologies. In addition, key challenges to data engineering will be highlighted both from a scientific and practical vantage point.

Subsequently, ► Chap. 3, entitled *Data Governance*, is introduced as a means to effectively organize decision-making around the development and maintenance of big data architectures and infrastructure, defining authorizations and obligations for stakeholders. New roles have emerged over the past few years, notably that of data custodians and data stewards, to realize them, e.g., in the context of—among others—novel rules, regulations, and policies with respect to safety and security. The concepts underpinning data governance are then explored and illustrated based on several realistic scenarios, drawn from A H2020 EU project entitled SODALITE.

► Chapter 4, namely *Big Data Architectures*, then sets out on architectural principles, patterns, and models underpinning modern day big data processing infrastructure, such as the Lambda and Kappa architecture. In addition, it introduces the SEI-CMU reference architecture that is widely recognized as the de facto standard reference point defining standard nomenclature and data modules adopting a wider system-of-systems perspective involving data providers and consumers.

► Chapter 5, namely *Data Engineering in Action*, concludes this module with three real-world cases that demonstrate the ramifications of data engineering in action. Firstly, this explores the data engineering aspects involved in cybercrime fighting on the dark net. This comes with specific desiderata, e.g., surrounding the collection, integration, and storage of mined data. The authors demonstrate the end-to-end data pipeline from identification of data sources up to advanced, real-time analytics. Secondly, it describes the challenges of data collection and harmonization for an EU H2020 project on the protection of cyber-physical places, called PROTECT. Lastly, it explores data engineering aspects of setting up smart beehives that serve as real-time “thermometers” for measuring (and predicting) biodiversity.

Contents

Chapter 2 **Big Data Engineering -- 25**

*Damian A. Tamburri and
Willem-Jan van Den Heuvel*

Chapter 3 **Data Governance -- 37**

*Indika Kumara, A. S. M.
Kayes, Paul Mundt, and
Ralf Schneider*

Chapter 4 **Big Data Architectures -- 63**

*Martin Garriga,
Geert Monsieur, and
Damian Tamburri*

Chapter 5 **Data Engineering in Action -- 77**

*Giuseppe Cascavilla,
Stefano Dalla Palma,
Stefan Driessen, Willem-Jan
van den Heuvel, Daniel de
Pascale, Mirella
Sangiovanni, and
Gerard Schouten*



Big Data Engineering

*Damian Tamburri
and Willem-Jan van den Heuvel*

Contents

- 2.1 Introduction: The Big Data Engineering Realm – 26**
 - 2.1.1 Data Engineering Challenges in Theory and Practice – 27
- 2.2 (Big) Data Engineering to Leverage Analytics – 29**
 - 2.2.1 Value-Driven Big Data Engineering – 29
 - 2.2.2 Key Fabric of Data Engineering – 29
 - 2.2.3 MLOps: Data Engineering (Finally) Meets AI/Machine Learning – 33
- References – 34**

Learning Objectives

After reading this chapter, the readers will be able to:

- Understand and explain the key concepts of big data and (Big) data engineering.
- Understand and describe the key activities and roles involved in data engineering, including their key challenges.
- Have a basic understanding of data lakes, data pipeline, and key data architectures supporting data analytics.
- Have a basic understanding of data engineering process models, notably MLOps.
- Have sufficient basis to understand how the topics discussed in the remainder of this chapter are related to each other.

2.1 Introduction: The Big Data Engineering Realm

Big data concerns large-amount, complex, and dynamically growing data collections with multiple, autonomous sources, networking, data storage, and data processing capacity. These data are rapidly expanding in all science and engineering stream, including physical, biological, and medical sciences (Senthil Kumar & Kirthika, 2017). The speed of generating data is growing in a way that makes it exceedingly challenging to handle such amount of large data. The main challenge is that the volume of data is ever-growing with respect to the capabilities of computing resources.

Big data requires salable, robust, and safe technologies to efficiently process large quantities of data within tolerable elapsed times. Technologies being applied to big data include massively parallel processing (MPP) databases, data mining grids, distributed file systems, distributed databases, cloud computing platforms, the Internet, and scalable storage systems—just to mention a few of the most predominant ones (Sun & Wen, 2019).

Quoting from “Practical DevOps for Big Data Applications” book by the EU H2020 DICE consortium,¹ “*Big Data [engineering] as an emerging scientific discourse reflecting the digitization of business systems at an unprecedented scale.*”

Low scale in the above quote pertains to more “controllable,” “smaller sized,” batch-oriented, and structured data repositories. Data methods and scientific methods or technologies for “lower” scales quickly become obsolete when scale increases. By now, several new concepts, techniques, and technologies have emerged, including software development methods that effectively single out collaborative multidisciplinary behaviors from both the software engineers (including software analysts, developers, and maintainers) and domain experts (e.g., financial experts, marketing specialists).

Indeed, there is an ever-growing urgency of “controlling” the current, continuous, and bulky wave of big data by meaningful abstractions and automation capa-

¹ ► https://en.wikibooks.org/wiki/Practical_DevOps_for_Big_Data.

ble of taming the scale involved and being able to deal with another imposed layer of complexity (Artac et al., 2018). Nowadays, a number of different heterogeneous technologies for big data appear—a non-exhaustive overview of which is available later in this module—at a very high rate whilst they at the same time become more and more intricate.

Lastly, we observe that data-intensive developers need to consider that a data-intensive application is much akin of a biology-inspired *complex adaptive system* (CAS) with different highly distributed, discrete, and autonomous computing and storage nodes (which typically live in the cloud or “at the edge”) that closely collaborate for the purpose of data processing tasks. Such modern-day data-intensive systems increasingly demonstrate nonlinear and emergent behaviors and are governed by specific communication mechanisms and associated policies. Some mundane examples of such policies include privacy-concerned regulations such as GDPR and the privacy-by-design (Guerriero et al., 2018) paradigm (Tamburri, 2020).

This means that in order to effectively use such technologies at the best of their potential, one needs to deeply understand them, effectively design for their users and usage scenarios, and continuously verify for such policies and constraints. This has serious implications for the baseline of technologies that can be considered for specific data-intensive applications, adding yet another layer of complexity. Lastly, data-intensive applications grow even more complex as the number and qualities of data-intensive application components increase in the architecture as well.

Ergo, the above depicts an intrinsically challenging realm of big data, with layers upon layers of abstraction and complexity. To deal with this, systematic, tractable, and disciplined methodologies, which in themselves constitute mixed methods and tools, are of critical importance: the realm of (big) data engineering.

2.1.1 Data Engineering Challenges in Theory and Practice

At this stage, therefore, the real data engineering problem has become to develop novel methods and technologies (Perez-Palacin et al., 2019) that lead to continuously deployable, data-intensive blueprints by having big data application abstracts, on top of more specific technologies (Artac et al., 2018).

From a data engineering perspective, among the many *vs* which are typically ascribed to the big nature of data, in the scope of this book, we focus on the following:

1. **Variety:** Data stems from different distributed and heterogeneous data sources like web pages, web logs, social media, and sensor devices. In addition, such data varies in terms of format: some data is highly structured and well formed (e.g., data stemming from relational database repositories), whilst other data is stale and semi- or unstructured (e.g., data from social media or JSON application interfaces). At the same time, the data is increasingly generated at varied and unpredictable levels of speed—e.g., consider a high-resolution thermal camera attached to a police drone with a shake connection with backend data processing resources on the ground. It is exceedingly difficult, time consuming,

and cumbersome for traditional computing resources to deal with this variety in data.

2. **Volume:** Nowadays, data storage needs are growing to astronomical numbers such as petabytes of data—a petabyte of data equals one thousand million million. It is supposed to jump into zettabytes (cf. 1000 petabytes) in the next few years (Sun et al., 2018). Notably, social networks nowadays produce terabytes of data each day, and the World Economic Forum estimated in 2019 that the total Internet is expected to reach 40 zettabytes by 2020 (Desjardins, 2019). Obviously, it is hard to handle this volume of data by using the traditional data processing techniques due to limitations in their scalability and elasticity.
3. **Velocity:** Velocity references to the velocity of data arriving from a data source (e.g., a sensor device) or the speed in which data may be processed by a node (e.g., a cloud service). The speed in which data is being generated is extreme. Sensors, signal receivers, machine learning algorithms, and so on are precipitately generating and processing massive streams of data at lightning-fast paces instead of in overnight or hourly batches (Ciavotta et al., 2019; Susanto et al., 2019).

At the same time, the aforementioned dimensions incur organizational and technical challenges that pose yet another data engineering challenge. More specifically, among others:

Data representation: Many datasets have certain levels of heterogeneity in type, structure, semantics, organization, granularity, and accessibility. Data representation aims to make data more meaningful for computer analysis and user interpretation. Nevertheless, an improper data representation will reduce the value of the original data and may even obstruct effective data analysis.

Redundancy reduction and data compression: Generally, there is a high level of redundancy in datasets. Redundancy reduction and data compression are effective to reduce the indirect cost of the entire system on the premise that the potential values of the data are not affected.

Data life cycle management: Compared with the relatively slow advances of storage systems, pervasive sensing and computing are generating data at unprecedented rates and scales. We are faced with several pressing challenges, one of which is that current storage systems may not support such massive data.

Analytical mechanisms: The multi-model and streaming nature of big data puts serious demands on the analytical mechanisms in place. In contrast to traditional RDBMS, modern-day analytical mechanisms require dealing with various data models (i.e., graph based and column stores), dynamic querying techniques that can deal with continuously changing and streaming data from various heterogeneous and vastly distributed sources.

Expendability and scalability: The analytical system of big data must support present and future datasets. The analytical algorithm must be able to process increasingly expanding and more complex datasets.

Cooperation: The analysis of big data is inherently interdisciplinary in nature and requires experts from different complementary disciplines (such as quality and

performance engineering) to closely cooperate to effectively harvest the potential of big data. A comprehensive big data network architecture must be established to help scientists and engineers in various fields access different kinds of data and fully utilize their expertise, so as to cooperate to complete the analytical objectives.

2.2 (Big) Data Engineering to Leverage Analytics

2.2.1 Value-Driven Big Data Engineering

In 2006, Michael Palmer, a marketing pundit, was one of the first to promote data to be just like crude oil. This analogy effectively demonstrated how data in its rawest and purest form has actually no value at all. However, just like crude oil, data may turn into a valuable commodity through proper processing and refinery. That is exactly what data engineering aims to support.

In its purest form, data engineering can be perceived as a way to (1) *identify* data sources (oil fields), (2) *extract* the data from them, and develop pipelines to transport it, to further (3) *transform* it in a uniform and semantically enriched format of sufficient quality. Then, the data will be (4) *stored* and (5) *managed and governed* in a data repository, like a data lake, so it can be safely and routinely consulted to unlock potential business value. Data analytics logically follows these data engineering tasks to refine this data into information to improve decision-making.

The tantalizing value proposition of big data in tandem with data analytics (e.g., using ML and AI techniques) seems nowadays fully recognized by commercial ventures, governmental institutes, and society at large.

Whilst much work in both the data engineering and data analytics research and practice has gone into each of the fields individually, much work still needs to be done to combine these domains efficiently and effectively as well as by means of varied and highly distributed heterogeneous cloud infrastructure.

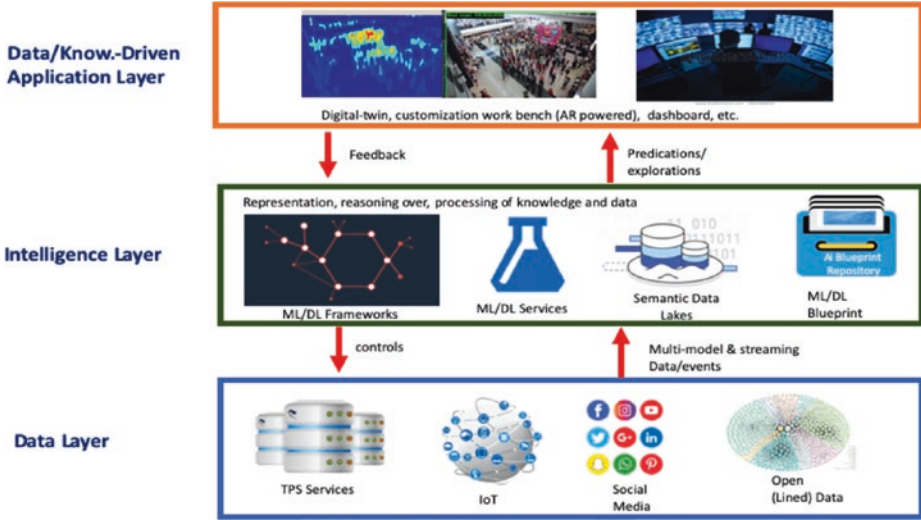
2.2.2 Key Fabric of Data Engineering

2.2.2.1 Intelligent Enterprise Application Architecture (iA)²

Novel enterprise applications are increasingly infused with machine learning and/or deep learning code in order to make their systems more “intelligent.” This imposes additional requirements on traditional enterprise applications.

The Intelligent Enterprise Application Architecture (iA)² (van den Heuvel & Tamburri, 2020) constitutes a stratified architecture encompassing three layers promoting logical separation of concerns, loose coupling, and reuse.

In normal cases, employing the basic data layer in iA², a data engineer designs, develops, and deploys a data pipeline including data preparation, feature engineering, data transformation, data management, and governance functionality.



■ Fig. 2.1 The intelligent enterprise application architecture

The intelligence layer of the iA^2 encompasses necessary roles and functionality for developing and deploying ML/DL models that collectively embody the key “intelligence.”

Typically, data scientists/data analysts and/or AI experts will exploit existing DL/ML frameworks such as Google’s AI, Microsoft’s Azure, and IBM’s Watson—and increasingly AutoML platforms—that provision (semi-) automated AI/DL services such as sentiment analysis, recommendation systems, purchase predication, spam detection, and others (■ Fig. 2.1).

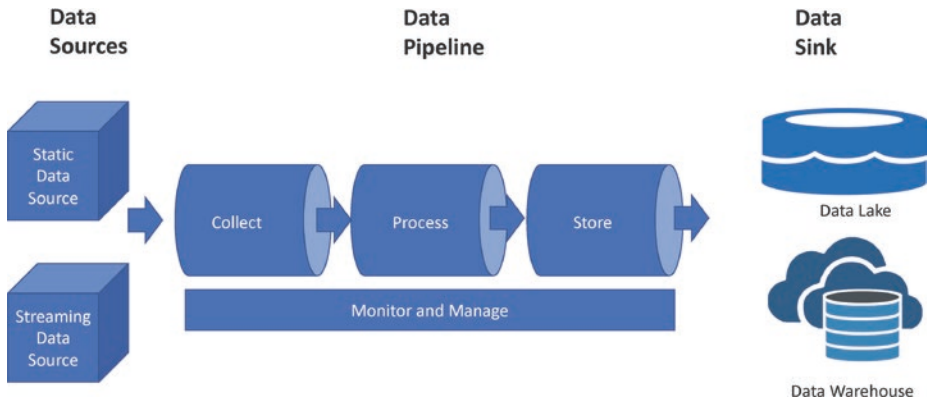
The intelligence layer makes use of semantic data lake that comprises meta-data descriptors, which may range from simple label identifiers to full-fledged semantic ontologies. The purpose of the semantic data layer is to capture and convey the meaning in the context of the AL/DL models and their associated datasets.

Since data lakes are the key data reservoirs in this setup, and a first-class citizen in data engineering scenarios, we will now further explain them.

2.2.2.2 Data Pipelines

From an abstract perspective, a data pipeline may be basically defined as a series of data processing tasks connected in a series, where the output of one task is the input of the next task. Data pipelines in real-world settings typically consist of multiple micro-services aka tasks, leveraging different technologies to meet required design goals or considerations.

From a high level of abstraction, tasks within (big) data pipelines may be categorized into three main types: collect, process, and store (see ■ Fig. 2.2). The collect task accommodates data from the source into the data pipeline. Depending on the design of source, data is pulled or pushed into the data pipeline. During the



■ Fig. 2.2 High-level architectural overview of a data pipeline. (“Author’s own figure”)

next main task, data may be transformed and/or processed. The final main task involves storing data to a *data lake* or *data warehouse*.

Data ingestion involves extracting data from data sources, where data sources are the endpoints from where the data pipelines consume data and may be further categorized as streaming or static (stationary) sources.

Streaming data sources involves data that is poured into the data pipeline simply continuously when new data is available. For example, consider a light sensor that detects movements in a room; this sensor streams data into the data source once movement is detected by the sensor. Static data on the other hand is typically entered or changed sporadically rather than continually and may be gathered from the data source in a periodic, batch-oriented manner, for example once a day or week.

Once data is collected from the data source (either continuously or not), it can be either buffered or streamed directly to the sink (data lake or data warehouse).

The collector has a streaming layer to accommodate data from streaming sources and a batch layer to gather data from stationary sources. The data bus acts as a buffer for the incoming messages. However, in those cases where a data bus results in a large overhead, the data processor may receive the data directly from the data collector. Once the data processor picks up incoming data, it transforms the data and writes the result through the output driver to a data store or sink.

The data bus allows the data collection and data processing to operate asynchronously, with some dynamical scaling capacity.

By implementing a publish/subscriber broker and a centralized architecture, for example, multiple processors can consume data from a single point. This simplifies complicated tasks such as governing, routing, and managing data.

To this end, the data bus is typically instrumented with a controller that is responsible for the monitoring and management of the data pipeline, including logging, monitoring, and dynamic adaptation. Once transported, the data is stored into the data sink.

2.2.2.3 Data Lakes and Data Warehouses

Big data processing will support the processing and integration of data into a unified view from disparate big data sources ranging from business data warehouse, customer and product data, ERP and CRM systems, sensors (at the edge) and smart devices, and social platforms to databases, whether structured or unstructured, to support big data-driven AI analytics.

The aim is to select, aggregate, standardize, analyze, and deliver data to the point of care in an intuitive and meaningful format.

To achieve this, the following data engineering steps may be pursued:

- Embrace semantic- and/or knowledge-based metadata techniques to structure and reconcile disparate business datasets and content, annotate them, link them with associated business processes and software services, and deliver or syndicate information to recipients. These are the mechanisms that transform stale data into more value-bearing and actionable data—aka knowledge.
- To improve business data interoperability, data engineering is geared toward developing a systematic representation and interoperability language. Typically, this is achieved by defining semantic reference data structures, sometimes referred to as blueprints, that are instrumental in defining the action application-level interfaces needed to unlock and extract data from data sources.
- The structurally and semantically enhanced data collection is purposed to create and manage a data lake that provides the basis for actionable insights on emerging concerns that can be highly relevant to improving enterprise value.
- The data in the data lake may be organized and made accessible when needed and subsequently made actionable for analytics using late binding.

In summary, a data lake may be defined as an open reservoir for the vast amount of data inherent in enterprises, e.g., traditional sources of data (comprehensive business records or master ERP and CRM systems, product life cycle management systems), from new sources of data (mobile apps, sensor networks, and wearables) and sources that are usually created for other purposes such as environmental and contextual data, which can be integrated into an analytics platform to improve decision-making. A data lake can ensure that data can employ data security and privacy mechanisms to ensure safety, confidentiality, and anonymity of data transfer to avoid misinterpretation and inappropriate conclusions by using proper annotation methodologies of the data.

A common misinterpretation is that a data lake is simply another instantiation of a data warehouse. On the contrary, a data lake entails a reusable building block of an early-binding data warehouse, a late-binding data warehouse, and a distributed data processing (Hadoop or Sparc) system. The early-binding mechanism in a data warehouse guarantees that all the data are organized and harmonized before it can be consumed.

A data lake thus brings value as it provides companies with a single, multimodal data repository system allowing on-demand and early-binding data to its users.

2.2.3 MLOps: Data Engineering (Finally) Meets AI/Machine Learning

Over the past few years, data science in general and artificial intelligence (AI) more in particular have swiftly grown in a key tech driver that is currently reshaping the way in which we conduct business and live our daily lives (Artac et al., 2018), witnessing a plethora of tantalizing AI-driven innovations that are explained throughout this book.

Indeed, AI is now quickly maturing and finally delivering industry-strength applications—popularly referred to as *AI software*—breaking away from the early-day, experimental, non-scalable “toy” AI prototypes devoid of practical enterprise value.

This latest AI trend has led to software engineering academia and industry to seriously turn their attention to infuse AI techniques, technologies, and platforms (such as Google AI platform, TensorFlow, IBM’s Watson Studio, and Microsoft’s Azure) in their software development practices.

Notably, AI techniques are increasingly deployed to foster automatic code generation, continuous testing and integration, and mapping software designs in executable, deployable, and scalable code. This has paved the way for machine learning applications with full-fledged DevOps pipelines to maintain them: a piece of software engineering fabric termed as *MLOps*.

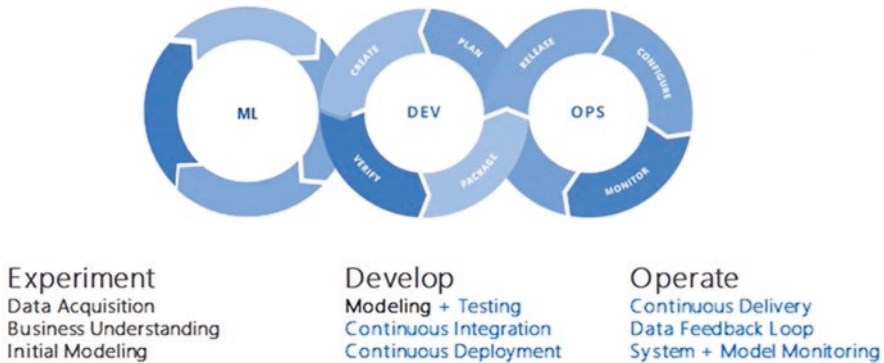
MLOps has emerged from the DevOps philosophy and associated practices that streamline the software development workflow and delivery processes. Like DevOps, MLOps adopts the continuous integration and continuous testing cycle to produce and deploy production-ready new micro-releases and versions of intelligent enterprise applications.

The basic philosophy underpinning has been graphically depicted in ■ Fig. 2.3. Essentially, MLOps compounds the cycle of machine learning, with a Dev(elopment) and Op(eration)s cycle.

The ML cycle is concerned with model training, testing, and validation in a highly iterative and experimental manner in order to find the best fit to the business problem at hand and generate the most business value. Once a model has been calibrated yielding the optimal business fit with business performance (e.g., in terms of accuracy), it is handed off to the second cycle. During the Dev(elopment) cycle, the models are coded and tested in a continuous manner, embracing well-understood best practices from software engineering including automated code (integration) testing, unit testing, and source code analysis. The operations cycle puts the tested code into the production environment, after having packaged the code, deploying it, and monitoring it in real time.

Clearly, MLOps implies a major culture shift between data analysts, data engineers, deployment and system engineers, and domain experts, with improved dependency management (and thus transparency) between model development, training, validation, and deployment. As such, MLOps clearly requires sophisticated policies based on metrics and telemetry such as performance indicators like F1, and accuracy scores, as well as software quality.

MLOps = ML + DEV + OPS



■ **Fig. 2.3** MLOps: combining machine learning development and operations. (Adopted from ► <https://www.c-sharpcorner.com/blogs/mlops>)

Whilst the exact boundary between MLOps and DevOps is blurry, a prominent example of MLOps can be found in Amazon Web Services that offers an integrated ML workflow—albeit vendor locked—across build, testing, and integration, supporting continuous delivery with source control and monitoring services.

Take-Home Messages

The reader can take the following key points from this chapter:

- Developing data-intensive applications and AI software requires understanding of (big) data engineering.
- Big data engineering is concerned with methods, tools, and process models (such as MLOps) that help to systematically and repeatedly engineer data-centric applications, including data pipelines, data lakes, and models, into commercial products and services fit for a production environment.
- Big data engineering is inherently challenging requiring seamless collaboration between data scientists, AI modelers, and data/software engineers.

References

- Artac, M., Borováqal, T., Di Nitto, E., Guerriero, M., Perez-Palacin, D., & Tamburri, D. A. (2018). Infrastructure-as-code for data-intensive architectures: A model-driven development approach. In *2018 IEEE International Conference on Software Architecture (ICSA)* (pp. 156–15609).
- Ciavotta, M., Krstic, S., Tamburri, D. A., & van den Heuvel, W.-J. (2019). Hyperspark: A data-intensive programming environment for parallel metaheuristics. In E. Bertino, C. K. Chang, P. Chen, E. Damiani, M. Goul, & K. Oyama (Eds.), *BigData Congress* (pp. 85–92). IEEE.
- Desjardins, J. (2019). How much data is generated each day?
- Guerriero, M., Tamburri, D. A., & Di Nitto, E. (2018). Defining, enforcing and checking privacy policies in data-intensive applications. In *Proceedings of the 13th International Conference on*

- Software Engineering for Adaptive and Self-Managing Systems*, SEAMS '18 (pp. 172–182), New York, NY, USA. Association for Computing Machinery.
- Perez-Palacin, D., Merseguer, J., Requeno, J. I., Guerriero, M., Di Nitto, E., & Tamburri, D. A. (2019). A uml profile for the design, quality assessment and deployment of data-intensive applications. *Software Systems Modeling*, 1–38.
- Senthil Kumar, S., & Kirthika, M. V. (2017). Big data analytics architecture and challenges, issues of big data analytics. *International Journal of Trend in Scientific Research and Development*, 1(6), 669–673.
- Sun, P., & Wen, Y. (2019). Scalable architectures for big data analysis. In S. Sakr & A. Y. Zomaya (Eds.), *Encyclopedia of big data technologies*. Springer.
- Sun, Z., Strang, K. D., & Li, R. (2018). Big data with ten big characteristics. In ICBDR (pp. 56–61). ACM.
- Susanto, H., Leu, F.-Y., Rosiyadi, D., & Kang, C. C. (2019). Revealing storage and speed transmission emerging technology of big data. In L. Barolli, M. Takizawa, F. Xhafa, & T. Enokido (Eds.), *AINA* (Advances in intelligent systems and computing) (Vol. 926, pp. 571–583). Springer.
- Tamburri, D. A. (2020). Design principles for the general data protection regulation (GDPR): A formal concept analysis and its evaluation. *Information Systems*, 91, 101469.
- van den Heuvel, W.-J., & Tamburri, D. A. (2020). Model-driven ml-ops for intelligent enterprise applications: Vision, approaches and challenges. In B. Shishkov (Ed.), *Business modeling and software design* (pp. 169–181). Springer International Publishing.



Data Governance

*Indika Kumara, A. S. M. Kayes,
Paul Mundt, and Ralf Schneider*

Contents

- 3.1 Introduction – 39**
- 3.2 Motivational Case Studies – 40**
 - 3.2.1 SODALITE Vehicle IoT – 40
 - 3.2.2 SODALITE Clinical Trials – 42
- 3.3 Data Governance in a Nutshell – 43**
- 3.4 Data Governance Dimensions – 45**
 - 3.4.1 Data Principles – 45
 - 3.4.2 Data Quality – 47
 - 3.4.3 Metadata – 49
 - 3.4.4 Data Access – 49
 - 3.4.5 Data Life Cycle – 50
- 3.5 Data Governance Structure – 51**
 - 3.5.1 Executive Sponsor – 52
 - 3.5.2 Data Governance Council – 52
 - 3.5.3 Data Custodian – 53
 - 3.5.4 Data Steward – 53
 - 3.5.5 Data User Groups – 54

3.6	Contemporary Data Governance – 54
3.6.1	Big Data Governance – 54
3.6.2	IoT Data Governance – 55
3.7	Case Studies with Data Governance – 56
3.7.1	SODALITE Vehicle IoT Architecture – 56
3.7.2	SODALITE Clinical Trial Architecture – 58
	References – 60

Learning Objectives

After reading this chapter, the readers will be able to:

- Understand and explain the key concepts of data governance.
- Understand and describe the key decision domains of data governance.
- Understand and describe a desired organizational structure for data governance, in terms of key roles and their responsibilities.
- Explain and analyze the key implications of big data and IoT on data governance.
- Incorporate data governance into designs of data products and services.

3.1 Introduction

The organizations are increasingly producing and consuming a massive amount of data at a rapid pace. Turning these big data to a value or strategic asset for the organizations is a key objective of data-intensive products and services. While the utilization of data enables gaining advantage and maximizing value of products and services, there are also associated cost and risk of using data, for example, economic and reputational risks of storing sensitive data, and storage, energy, maintenance, and software costs of storing and analyzing data (Tallon, 2013). The quality of cooperate data such as financial data, customer data, and supplier data is important for the business. For example, the inconsistencies in customer data in different systems across an organization can create data integrity issues that affect the accuracy of data-driven decision-making. The poor quality of data can also complicate maintaining of the compliance of data with respect to the regulations and laws such as General Data Protection Regulation (GDPR). Thus, the appropriate management and governance of organizational data are key to strike a balance between risk/cost and value creation (Malik, 2013; Tallon, 2013; Wilkinson et al., 2016; Cumbley & Church, 2013; Khatri & Brown, 2010).

Data governance considers the organizational entities that hold the decision rights and are held accountable for the decision-making about the data assets of the organization. It specifies the decision rights and accountability framework to support and encourage desirable behaviors in the use of data by the products and services in the organization (Khatri & Brown, 2010). A data governance program of an organization establishes the required decision-making structures, processes, policies, standards, architecture, and evaluating metrics that enable the strategic objectives for data and its quality to be implemented and monitor how well these strategic objectives are being achieved.

The literature (Khatri & Brown, 2010; Otto, 2011) differentiates the terms *data governance* and *data management*. Governance considers the decisions concerning the effective use of data (what), the personnel responsible for making decisions (who), and the methods by which the management actually makes and realizes the decisions (how). Thus, data governance is a prerequisite for data management.

In this chapter, we discuss the data governance and its application for data-intensive products and services. In ► Sect. 3.2, we present two industrial case studies from vehicle IoT and clinical trial domains to highlight the needs for data governance in data products and services. Next, ► Sect. 3.3 defines the data governance and provides an overview of the key components of a data governance framework. Next, in ► Sect. 3.4, we discuss the five key decision domains or dimensions in a data governance program/framework: data principles, data quality, data access, data life cycle, and metadata. ► Section 3.5 focuses on the organizational structure for supporting a data governance program, in terms of key roles such as executive sponsor, data governance council, data custodian, data steward, and data user. In ► Sect. 3.6, we present threats, opportunities, and approaches for governing big data and IoT data. To guide the design of the data governance-aware data-intensive products and services, we finally present the designs of two case studies that support a proper governance of the data produced and consumed by them.

3.2 Motivational Case Studies

In this section, we motivate the needs for data governance in data products/services using the two industrial case studies from IoT and clinical trial domains. They are from a European Union project, namely SODALITE (Di Nitto et al., 2022).¹

3.2.1 SODALITE Vehicle IoT

The SODALITE Vehicle IoT use case involves the provisioning and delivery of data-driven services from the cloud to a connected vehicle (or across a fleet of vehicles), leveraging a combination of data both from the vehicle itself (e.g., GPS-based telemetry data, gyroscope and accelerometer readings, biometric data from driver monitoring) and from external sources that can enrich the vehicle data and provide additional context to the service (e.g., weather and road condition data based on the location and heading of the vehicle).

Vehicle services can be deployed in a number of different ways:

- From the cloud to the edge (in this case, the vehicle itself) directly
- Directly at the edge (self-contained within the vehicle itself)
- From a self-contained fleet cloud to managed vehicles within the fleet
- From the cloud to one or more fleet clouds to managed vehicles within each vehicle fleet (multi-cloud federation)

A unique characteristic of the use case is that the vehicle is not a stationary object and may, at any time, cross over into another country—subjecting the data process-

¹ ► <https://sodalite.eu/>: Software Defined AppLIcation Infrastructures management and Engineering.

ing activities carried out by the service to the regulatory compliance requirements of not only the country where it started its journey, but also every country it enters along the way. Data service providers, therefore, must be able to demonstrate end-to-end compliance of their processing activities in each territory where these activities are carried out. Furthermore, the data protection authorities (DPAs) in each country may request the service provider to provide evidence of compliance with their country-specific regulations pertaining to the processing activities that occurred while the vehicle fell under their territorial scope in a subsequent audit.

An additional point of dynamism falls with the driver, who may change their privacy preferences or withdraw their consent for a given service at any time throughout a journey. In a long journey, the driver may also periodically change, with each driver having their own unique attitudes to privacy and services they have consented to. These changes require not only that deployed services reconfigure and adapt themselves to match the changes in preferences, but also the provisioning of additional services the driver wishes to receive as well as the deprovisioning of currently active services the driver has not (or has no longer) consented to. Data service and platform providers must, therefore, be able to demonstrate end-to-end compliance not only for the service itself, but also for each driver (as the data subject) individually across the duration of the journey.

Regulations such as Regulation (EU) 2016/679 (the GDPR) (Protection Regulation, 2016) and the newer Regulation (EU) 2018/1807 on the free flow of nonpersonal data (the FFD) (Protection Regulation, 1807) have done much of the heavy lifting in providing comprehensive regulatory frameworks for enabling cross-border data flows in data-driven services, but are not without their own limitations. Besides the basic compliance requirements of the GDPR, a number of additional factors must also be considered:

- **GDPR variance across EU member states:** While the GDPR is often presented as a consistent regulation with consistent application, it contains over 50 provisions where member states can derogate. Many of these derogations apply to restrictions on specific processing activities (specifically, Articles 23 and 85-91).
- **Supplemental legislation across EU member states:** Member states must enact supplemental legislation in order to bring their existing regulatory environment in line with the GDPR. The supplemental nature of these laws should not be understated, as they are often far longer and more complex than the GDPR itself.
- **Level of data protection adequacy of the country entered:** Service providers must consider not only travel between GDPR-compliant countries, but also travel to third countries with an adequate level of data protection, as well as those for which no data protection adequacy decision has yet been made.
- **Data types that are not covered by either the GDPR or the FFD, or which may be subject to additional national-level legislation:** Biometric data, as used in driver monitoring, for example, is classified as special category data pursuant to Article 9 GDPR and may be subject to additional national-level health data legislation that prevents this data from leaving the country, regardless of whether the data subject has consented to this or not.

Based on these criteria, data governance can be seen to play an integral role in the operability of the data service and must be engineered specifically to the service and its unique data processing activities while accounting for the dynamic operational environment it is ultimately deployed into.

3 3.2.2 SODALITE Clinical Trials

SODALITE in silico clinical trial case study targets the development of a simulation process chain supporting in silico clinical trials of bone implant systems in neurosurgery, orthopedics, and osteosynthesis. It deals with the analysis and assessment of screw-rod fixation systems for instrumented mono- and bi-segmental fusion of the lumbar spine by means of continuum mechanical simulation methods. The simulation chain consists of a number of steps that need to be fulfilled in order and can be considered a pipeline. The output of each step serves as an input to the next step as shown in Fig. 3.1.

The simulation process helps to optimize screw-rod fixation systems based on clinical imaging data recorded during standard examinations and consequently target the lowering of the reported rates of screw loosening and revisions, enhance safety, expand the knowledge of the internal mechanics of screw-rod fixation systems applied to the lumbar spine, and finally reveal optimization potential in terms of device application and design. Once established, the proposed method will make the treatment outcome and development of implants for orthopedic surgery also more robust, since doctors can recognize irregularities in the healing process much earlier.

As can be seen in Fig. 3.1, the simulation process chain relies on a central database component acting as the top-level data instance in which the clinical

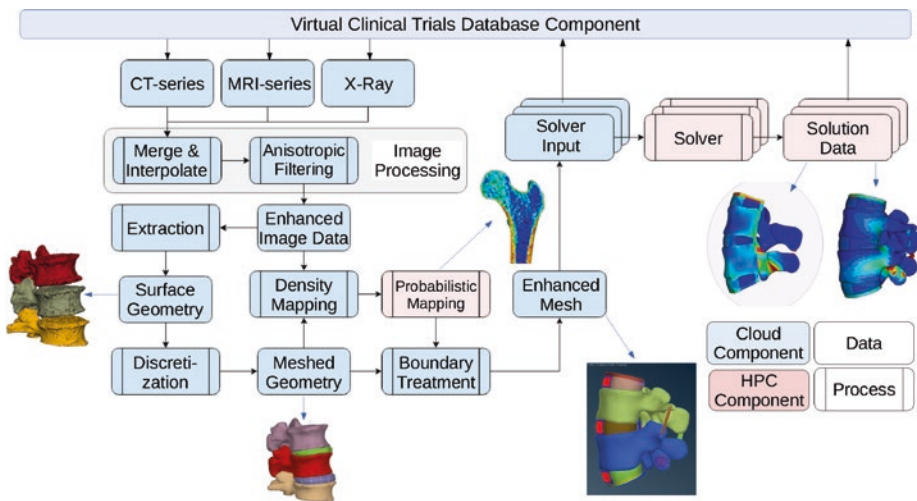


Fig. 3.1 Simulation process chain of in silico clinical trial use case. (“Author’s own figure”)

imaging data are stored that form the basis for the complete simulation process. From this database, imaging data are loaded and passed on to the different components of the process chain. The solution data are finally stored back into the database for further evaluation. Additionally, the solution data are used in a feedback loop to continuously improve the quality and accuracy of the simulation results. Another characteristic of the implemented process chain is that parts of it are deployed on cloud as well as on HPC infrastructure, between which the intermediate data have to be transferred. As can be seen, not only the input data of the process result from data containing sensitive information and hence should be governed properly, but also the treatment of the intermediate and result data requires data governance in the following respects:

- **GDPR compliance:** The case study involves the processing of personal data from several types of data subjects such as patients and investigators. The storage and handling of such data must be compliant with GDPR regulations.
- **Anonymization:** Since the simulation process is based on clinical imaging data, which in most cases contain header information including sensitive patient data, proper anonymization procedures have to be available.
- **Policy-based data transfer:** Given that the process chain is composed of cloud and HPC components, policy-based data transfer between different computing resources is a key issue, as it must comply with the respective legal requirements for the treatment of medical data.
- **Data quality:** Not only on data input but also in intermediate steps like image processing, it is important to consistently maintain a high quality of the processed data to build better prediction models and to ease regulatory compliance effort.

3.3 Data Governance in a Nutshell

There are many definitions of the term *data governance*. Gartner² defines data governance as *the specification of decision rights and an accountability framework to ensure the appropriate behavior in the valuation, creation, consumption, and control of data and analytics*. Data Governance Institute (DGI)³ defines data governance as *a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods*. In sum, data governance specifies and enforces rules and regulations over capture, retention, value creation, sharing, usage, and retirement of data.

In the literature (Khatri & Brown, 2010), data governance is differentiated from data management. The former defines types of decisions about the strategic use of data within an organization as well as the roles for making those decisions. In contrast, the latter refers to the process of making and implementing those decisions.

2 ► <https://www.gartner.com/en/information-technology/glossary/information-governance>.

3 ► http://www.datagovernance.com/adg_data_governance_definition/.

For example, governance includes establishing who in the organization holds decision rights for determining standards and metrics for data quality assessment and defining privacy and protection policies. Management involves measuring data quality against the given metrics using data profiling and implementing and enforcing privacy and protection policies using anonymization and encryption techniques.

A proper data governance strategy helps the organizations to strike a balance between value creation and risk exposure of data (Tallon, 2013). For example, the clinical trial data or the personal data of drivers need to be securely stored and managed to limit the risks of violation of data protection laws (GDPR) without undermining creating the desired form of value from the data. The retention and access of data should be governed based on explicitly and carefully defined policies to ensure compliance with the relevant regulations.

There exist several data governance frameworks, for example, DGI,⁴ Deloitte,⁵ and Informatica.⁶ The key components of such framework include the following:

- **Policies, Standards, Processes, and Procedures:** A data governance policy defines the rules that are encoded to ensure that the data assets in an organization are managed and used properly by balancing risk and value creation. There may be individual policies for different decision domains such as data quality, data compliance, and data access. Standards also serve a similar purpose (i.e., rules and guidelines for protecting and using data) in an interoperable way. Sample data standards include metadata management standards, naming standards, data modeling standards, data architecture standards, data quality standards, and other regulatory standards (e.g., GDPR). Processes and procedures are to ensure that policies and standards will be enacted and enforced continuously and consistently.
- **Roles and Responsibilities:** Organizations should select decision makers and define their respective roles and responsibilities at different levels such as executive, strategic, tactical, and operational. The roles span from executive sponsor, who supports and coordinates data governance activities and programs, to data governance council, which is responsible for establishing policies, standards, processes, and procedures, to data users, who access data.
- **Technology Capabilities:** To implement a data governance process, an organization needs appropriate platforms and tools, for example, metadata management tools/platforms, data profiling tools, data cleansing tools, and compliance checking tools. Sample companies that provide such tools include Collibra, Truedat, Talend, Informatica, and IBM.

4 ► <http://www.datagovernance.com/the-dgi-framework/>.

5 ► <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology/us-big-data-governance.pdf>.

6 ► https://www.informatica.com/nl/lp/holistic-data-governance-framework_2297.html.

3.4 Data Governance Dimensions

There are five main decision domains within the data governance (Khatri & Brown, 2010): data principles, data quality, metadata, data access, and data life cycle. In this section, we discuss each decision domain/dimension in detail.

3.4.1 Data Principles

Data principles are the foundation of any successful data governance framework. They aim to make the data an enterprise-wide asset that has values to both data providers and the organization. Data principles also determine the strategies and rules for facilitating reuse of data, enforcing security and privacy, assessing impacts of changes to data, and so on. In the literature, several principles and guidelines were proposed for the data and their use and governance, such as FAIR principles (Wilkinson et al., 2016) and FACT principles (van der Aalst, 2017).

Among the data principles proposed in the research literature, in the context of scientific data management and stewardship, FAIR data principles aim to make data findable, accessible, interoperable, and reusable (Wilkinson et al., 2016).

■ Figure 3.2 shows the detailed guidelines of FAIR principles.

- **Findable:** Data have sufficiently rich metadata and unique and persistent identifiers that allow their discovery by both humans and computer systems.
- **Accessible:** Once the users (humans and machines) find the required data, they should be able to easily access the data. The users should be able to easily understand data as well as things controlling data access such as licenses and other conditions, and authentication and authorization policies.
- **Interoperable:** The users (humans and machines) should be able to integrate a given dataset with other datasets, as well as applications or workflows that analyze, store, and process the datasets.
- **Reusable:** FAIR aims to optimize the reuse of data, and the users (humans and machines) should be able to use, replicate, and/or combine the data in different usage contexts.

FACT principles (van der Aalst, 2017; van der Aalst et al., 2017) aim to make data science research and practices responsible. Data science is an interdisciplinary field aiming to turn data into real value. ■ Figure 3.3 illustrates the FACT principles in the context of a data science pipeline.

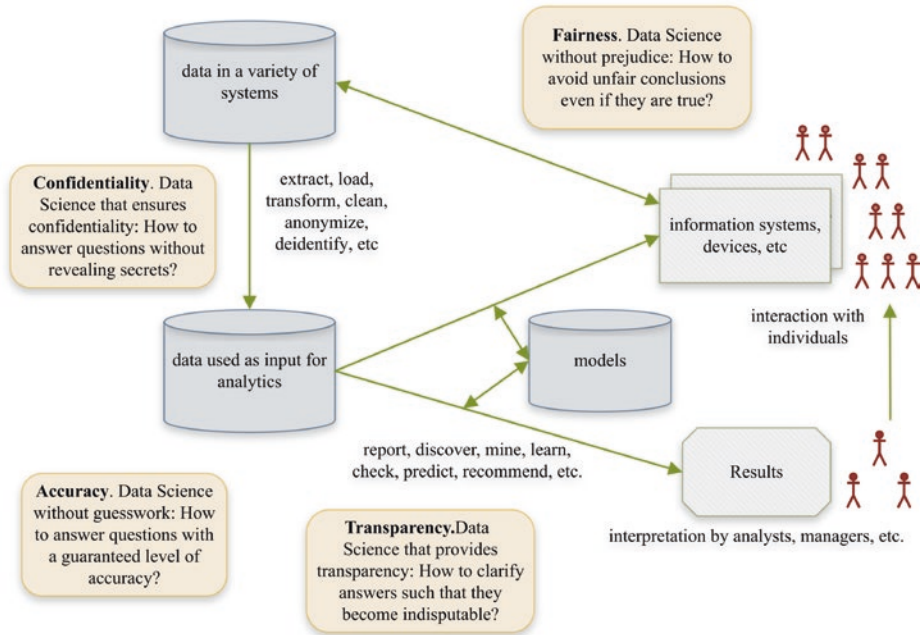
- **Fairness:** Automated decisions and insights should not be used to discriminate in ways that are unacceptable from a legal or ethical point of view. An example for the age discrimination in clinical trials is promoting the inclusion of older participants with multiple comorbidities. The process for achieving fairness includes **discrimination discovery** and **discrimination prevention**. The former aims to identify individuals or groups that are discriminated based on sensitive variables such as name, birth date, gender, driving experience, and age. The lat-

The FAIR Guiding Principles	
To be Findable:	<p>F1. (meta)data are assigned a globally unique and persistent identifier</p> <p>F2. data are described with rich metadata (defined by R1 below)</p> <p>F3. metadata clearly and explicitly include the identifier of the data it describes</p> <p>F4. (meta)data are registered or indexed in a searchable resource</p>
To be Accessible:	<p>A1. (meta)data are retrievable by their identifier using a standardized communications protocol</p> <p>A1.1 the protocol is open, free, and universally implementable</p> <p>A1.2 the protocol allows for an authentication and authorization procedure, where necessary</p> <p>A2. metadata are accessible, even when the data are no longer available</p>
To be Interoperable:	<p>I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.</p> <p>I2. (meta)data use vocabularies that follow FAIR principles</p> <p>I3. (meta)data include qualified references to other (meta)data</p>
To be Reusable:	<p>R1. meta(data) are richly described with a plurality of accurate and relevant attributes</p> <p>R1.1. (meta)data are released with a clear and accessible data usage license</p> <p>R1.2. (meta)data are associated with detailed provenance</p> <p>R1.3. (meta)data meet domain-relevant community standards</p>

■ Fig. 3.2 FAIR data principles (Wilkinson et al., 2016)

ter focuses on the development of algorithms that do not discriminate using sensitive variables.

- **Confidentiality:** Sensitive data such as personal information and company secrets should not be revealed at any stage of the data science pipeline. The regulations such as GDPR govern the disclosure of such data. The de-identification techniques can be used to anonymize, remove, or obscure sensitive data. Such techniques should strike a balance between the disclosure of protected data and the usefulness of analysis results. To project the access to data within the data science pipeline, the authentication and authorization policies should be specified and enforced.
- **Accuracy:** The results of the data analysis should guarantee a level of accuracy and prevent misleading users. To ensure the accuracy, the analysis techniques should take into account the various issues such as overfitting the data, testing multiple hypotheses, uncertainty in the input data, and hidden uncertainty in the results.



■ Fig. 3.3 FACT principles in the context of a data science pipeline (van der Aalst, 2017)

- **Transparency:** The automated decisions based on the rules learned from historic data and the manual decisions based on analysis results need to be explainable, understandable, auditable, undisputable, and trustworthy. Automated decision-making using black box machine learning models and communication of analysis results to decision makers in an unintelligible and vague manner can harm transparency.

3.4.2 Data Quality

The quality of data refers to its ability to satisfy its usage requirements (Strong et al., 1997; Batini et al., 2009; Khatri & Brown, 2010). Data quality is usually described using multiple dimensions whose precise interpretations depend on the context in which data is used (Liu & Chi, 2002; Batini et al., 2009; Strong et al., 1997; Khatri & Brown, 2010). ■ Table 3.1 defines and provides examples for five common data quality dimensions: accuracy, completeness, consistency, timeliness, and credibility.

There exist many methodologies and techniques to assess and improve the quality of data, for example, Total Data Quality Management (TDQM), Data Quality Assessment (DQA), and Comprehensive methodology for Data Quality Management (CDQ). By qualitatively analyzing these existing methodologies, Batini et al. (2009) identified common activities of a data quality assessment and

Table 3.1 Sample data quality dimensions. “Table compiled by author”

Dimension	Definition	Example
Accuracy	The degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use	The measured temperature of a device is 50 °C, and the real value is 55 °C
Completeness	The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use	Missing the value for the attributes’ email and phone numbers in a patient record
Consistency	The degree to which a set of data items violate semantic rules/constraints defined over them	The violation of the constraint. Age must range between 0 and 120
Timeliness	The extent to which data are sufficiently up to date for a task	2-min delay in the backend receiving the changed location of the vehicle
Credibility	Indicates the trustworthiness of the source as well as its content	Temperature data from sensors in the same room have significant mismatches

improvement methodology, which consists of three key phases: state reconstruction, measurement/assessment, and improvement.

- **State Reconstruction Phase:** The contextual information and metadata required by the activities in the assessment and improvement phases are collected in this phase. If there is necessary information already, this phase can be skipped.
- **Measurement/Assessment Phase:** The key objective of this phase is to measure the values of a set of data quality dimensions relevant to the processes in the organization. The activities include understanding the data in the organization as well as the policies for their use and management, identifying quality issues and desired quality targets, identifying data sources and consumers (processes and systems), and identifying the quality dimensions affected by the quality issues.
- **Improvement Phase:** This phase aims to identify the strategies, processes, and techniques that need to be employed to achieve the desired quality targets. It starts with estimating the cost of data quality; identifying data owners, process owners, and their roles and responsibilities; and identifying the root causes of quality issues. Next, the solutions for data improvements are formulated and enacted, and the data production processes are redesigned to enable the desired level of data quality monitoring. Finally, the improvement processes are also continuously monitored and adapted to ensure that the desired levels of improvements are delivered.

3.4.3 Metadata

Metadata is simply data about data and provides a mechanism for a concise and consistent description of the data and its context. It helps to organize, find, and understand data, the meaning or “semantics” of data. In the literature, there exist different catalogs of metadata (Singh et al., 2003; Greenberg, 2005; Riley, 2017; Khatri & Brown, 2010).

Gurmeet et al. (Singh et al., 2003; Khatri & Brown, 2010) identified five levels of metadata: physical, domain independent, domain specific, virtual organization, and user.

- **Physical metadata:** These include information about the physical storage of data. Database management systems and file systems are examples for the systems that maintain these types of metadata.
- **Domain-independent metadata:** These define the generic attributes that can be used to describe some aspects of the data items generated by different applications in different domains, for example, logical names, creator and modifier of data content, and access information.
- **Domain-specific metadata:** These include the attributes that can only be used to characterize domain-specific datasets, for example, metadata for clinical trial datasets and metadata for different types of applications developed at different organizational units.
- **Virtual organization metadata:** A virtual organization consists of geographically dispersed individuals, groups, organizational units, or entire organizations. The datasets used by such organizations can have organization-specific metadata.
- **User metadata:** These describe user-specific properties pertaining to the use of data such as user preferences and usage context and history.

The metadata can also be categorized into descriptive, structural, and administrative (Riley, 2017):

- **Descriptive metadata:** These metadata include the descriptive information about a resource to enable easy discovery and identification of the resource, for example, title, abstract, author, and keywords of a book or paper.
- **Structural metadata:** These metadata are to characterize the structure and composition of resources such as resource types and relationships, for example, the metadata that is necessary to describe how Web pages are connected and grouped to form a web site.
- **Administrative metadata:** These metadata include information necessary to manage a resource properly, such as resource type, when and how it was created, and access permissions.

3.4.4 Data Access

Data should be made available to the data consumers securely. The authorized consumers need to be able to retrieve, modify, copy, or move data from different data sources. The techniques for data access need to ensure the confidentiality,

integrity, and availability of data, as well as auditability, provenance, and compliance of data access operations (Khatri & Brown, 2010).

There are two basic forms of data access: sequential access and random access (or direct access) (Zomaya & Sakr, 2017). In the sequential access model, the data source only allows reading and writing data sequentially. In the random access model, the data source enables reading or writing arbitrary data items in any order. A linked list is an example of a data structure providing sequential access, and an array is an example of a data structure providing random access. Databases and batch processing systems generally support the random access model, and streaming processing systems generally support sequential access model.

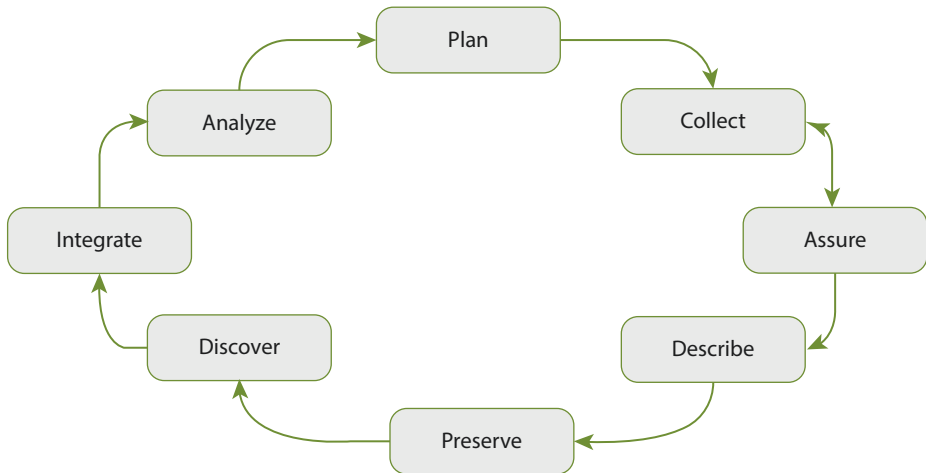
In the literature, different types of access control models have been introduced, such as the role-based access control (RBAC) (Sandhu et al., 1996), the attribute-based access control (ABAC) (Servos & Osborn, 2017), and the context-aware access control (CAAC) (Kayes et al., 2014, 2020b). In these access control models, the users' roles, attributes, and contexts have been considered, respectively, for authenticating and authorizing data access operations. These models can define and enforce various permissions and levels of security required for data access.

3.4.5 Data Life Cycle

The data life cycle presents the sequence of stages that data objects move from their capture/generation to their retirement/deletion. By gaining a proper oversight of data throughout its life cycle, organizations can develop approaches to optimizing its usefulness, minimizing or eliminating the potential for errors, and minimizing the total cost of storing data (Khatri & Brown, 2010).

There exist many life cycle models (Ball, 2012). In this section, we present an overview of the DataONE data life cycle model, which is a domain-independent model (Michener et al., 2012). ■ Figure 3.4 shows the stages of DataONE model.

- **Plan:** To meet the goals of the data-intensive project, this activity develops a data management plan covering the entire data life cycle.
- **Collect:** This activity decides the appropriate ways to get the desired data from different sources and to structure the collected data properly, for example, decision of collecting vehicle device temperature using sensors and some clinical trial data using computational simulations.
- **Assure:** In this activity, the quality of the data are measured based on metrics and standards and profiling techniques, and the discovered issues (inconsistencies and other anomalies) are fixed using data cleansing techniques (e.g., removing outliers, missing data interpolation).
- **Describe:** This activity documents the data by characterizing the data accurately and thoroughly using the metadata standards and tools.
- **Preserve:** This activity decides to archive the data based on the needs for short-term and long-term preservation of data, by considering and balancing differ-



■ Fig. 3.4 DataONE data life cycle model (Michener et al., 2012)

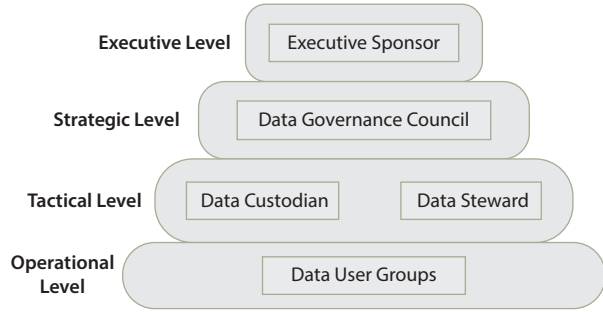
ent factors such as volume (thus storage cost), sensitivity (thus risk of exposure), usefulness, ease of access, and raw data or processed data (analysis results).

- **Discover:** In this activity, the relevant datasets that can create a value for the concerning data-intensive project are identified and retrieved.
- **Integrate:** This activity integrates and consolidates the collected datasets (including the data generated by the project) to enable different types of ad hoc analysis of data (constrained by the goals of the project). There exist data integration tools that can work with different types of big or small data, for example, Talend Data Integration, Informatica PowerCenter, AWS Glue, and Pentaho Data Integration.
- **Analyze:** Data are explored, analyzed, and visualized. Many tools are available for data analysis and visualization, for example, Python and R libraries, Apache Spark ecosystem, and Elasticsearch.

3.5 Data Governance Structure

The data governance structure of an organization defines roles and their responsibilities of different actors participating in the data governance program in the organization. In general, a data governance structure has multiple levels: executive, strategic, tactical, and operational (Cheong & Chang, 2007; Khatri & Brown, 2010; Weber et al., 2009). In the literature (Al-Ruithe et al., 2019; Korhonen et al., 2013), there are different variations of data governance structures. ■ Figure 3.5 shows a common structure, highlighting key roles at different levels. In the rest of this section, we discuss each role in detail. It is important to note that responsibilities of the roles can vary between different organizations and institutions. Such differences are evident in the research literature too.

■ **Fig. 3.5** A common data governance structure. (“Author’s own figure”)



3.5.1 Executive Sponsor

Data governance programs often fail when there is no strong support from the top management. The executive sponsor (senior leadership team) is the highest level of data governance and provides the support from top management. He/she is a member of the top management, such as the CEO, CFO, or CIO (Wende, 2007). The executive sponsor regularly participates in data quality council meetings and authorizes essential decisions, such as the data quality strategy and the data management plan.

The major responsibilities of the executive sponsor include the following:

- Sponsor approval, and get funding and support for the data governance program.
- Chair the data governance council and focus its work.
- Make key decisions when consensus within the council members cannot be reached.
- Identify and prioritize data quality initiatives across the organization.
- Nominate staff for projects and advocacy roles, and ensure accountability.

3.5.2 Data Governance Council

A data governance council (or committee/group) is responsible for providing strategic guidance for the data governance program of the organization. It usually includes the data owners and the lead data steward (Otto, 2011). A data owner is accountable for the quality of a defined dataset, for example, finance director as the data owner for finance data. The lead data steward uses one or more data stewards to formulate the rules for handling this dataset and its quality.

The major responsibilities of a data governance council include the following:

- Decide how data stewards are assigned to processes, data types, and business units.
- Enforce the adoption of standards, help establish data quality metrics and targets, and ensure that regulatory, privacy, and information sharing policies are followed.

- Implement organization policies and ensure compliance with government laws related to data governance.
- Manage, protect, and ensure the integrity and usefulness of organization data.
- Staff and supervise all data stewards.

3.5.3 Data Custodian

A data custodian is concerned with the safe custody, moving, and storage of the data and implementation of business rules. They are generally IT professionals (e.g., a database administrator and an ETL developer) who are responsible for the management and operation of the systems that produce, store, and transport organizational data (Otto, 2011).

The major responsibilities of the data custodian include the following:

- Provide a secure infrastructure in support of the data, which includes aspects such as physical security, backup and recovery processes, and secure transmission of the data.
- Implement and manage data access policies.
- Ensure system availability and adequate response time and not violation of the relevant service-level agreements.
- Participate in setting data governance priorities by providing details on technical, system, and staffing requirements related to data governance initiatives.

3.5.4 Data Steward

Data stewards are responsible for carrying out the tactical plans set by the data governance council. Each steward may have a responsibility for a subset of organization's data (e.g., customer data, supplier data, and product data). There are two types of data stewards: technical data stewards and business data stewards (Marco, 2006). The role of former is tactical, and the role of latter is operational and thus is often put under operational level. Technical data stewards (e.g., data architects and data modelers) are responsible for the data model and data life cycle across IT systems. Business data stewards are the business leaders accountable for definition, accuracy, consistency, and timeliness of critical information within their business scope (Villar, 2009; Wende, 2007).

The major responsibilities of the data steward include the following:

- Implement data standards and train the staff, who maintain data to ensure that they follow standards.
- Monitor data quality, which involves establishing a process for identifying data quality issues such as inconsistencies and violations of the selected quality standards.
- Respond to the inquiries about datasets they are accountable for, for example, questions on access, standardization, definition, and usage of data.
- Define data elements and values according to business requirements.

3.5.5 Data User Groups

A data user is an individual, who has access to the organization's data as part of assigned duties or in fulfillment of assigned roles or functions within the organization. They include people who collect, process, and report on the data (Wende, 2007). They need to use the data in accordance with the organization's policies and procedures regarding security, integrity, quality, consistency, usage, and sharing of data.

The major responsibilities of the data user groups include the following:

- Attend training and follow the organization's policies and procedures related to data management and protection.
- Report any data-related issues including those related to data management and protection.
- Request the functionality that would help them use data more efficiently.

3.6 Contemporary Data Governance

In this section, we discuss the data governance challenges and approaches in the context of big data and Internet of Things (IoT).

3.6.1 Big Data Governance

Nowadays, organizations have massive heterogeneous datasets that come from sources like ERP systems, Web server logs, social media, click streams, and sensor data. These datasets are growing at a rapid pace. In order to extract any form of value from this big data and to minimize the increasing cost and risk of storing data, the organizations need to adopt appropriate and scalable data management and governance practices (Sinaeepourfard et al., 2016; Malik, 2013; Taleb et al., 2016; Cumbley & Church, 2013; Tallon, 2013).

According to Malik (2013), the challenges and opportunities offered by big data, pertaining to data governance, include the following:

- **Confluence of mobile, Internet, and social activity:** Like enterprise data, the data being generated by social media and Web browsing must be governed and used appropriately, for example, recommendation systems for offering customized offers for users.
- **Evolving consumer behavior:** The purchase behaviors of customers are being influenced by that of their social contacts and other online content (e.g., reviews and rating). To understand this evolving customer behaviors, data from diverse sources such as in-store activities, click streams, and social media must be collected, integrated, and managed.
- **Rise of social commerce:** Social shopping is increasingly becoming popular. To use the massive unstructured data generated from social shopping, these social data must be integrated with structured corporate data.
- **Security and privacy:** As the amount and variety of sensitive data such as personal, health, and financial data increase, the safe sharing and use of these data

become critical. The handling of data needs to be compliant with not only local data regulations but also cross-border data regulations such as GDPR.

- **Technology advancements and open-source issues:** New systems and tools are being invented to cope with the unique characteristics of the big data, for example, NoSQL databases, MapReduce-based data processing systems, and streaming data processing systems (Zomaya & Sakr, 2017). Most of these software systems are open source.
- **Quality and uncertainty:** Compared with cooperate data, social data have many unique quality issues, for example, accuracy, completeness, and credibility of human sentiments and expressions.
- **Public datasets and data consortiums:** A huge amount of open datasets are available at different online sites. The governance of these open datasets and their providers is essential for a safe and reliable use of open data for organizational decision-making, which often requires the integration of open data with cooperate data.

While the big data does not necessarily change the basic structure of a conventional data governance program, the unique properties of big data require refining the existing tools, standards, processes, and practices. The research literature on big data governance includes topics such as big data quality dimensions, techniques for measuring big data quality (Merino et al., 2016), big data life cycle models (Sinaeepourfard et al., 2016), big data privacy and security models (Alshboul et al., 2015), and big data preprocessing (Taleb et al., 2015).

3.6.2 IoT Data Governance

The Internet of Things have created the world of connected experiences by the convergence of multiple technologies including a variety of sensors, devices, actuators, and embedded systems. However, with the proliferation of these IoT device technologies, there are still challenges in integrating, indexing, and managing data from multiple IoT sources (Doan et al., 2020). Thus, data governance in the Internet of Things (IoT) is crucial to support better decision-making using dynamic IoT data.

The quality of IoT and sensor data has been a key research topic (Karkouch et al., 2016; Perez-Castillo et al., 2018). By analyzing the existing relevant literature, Karkouch et al. (2016) identified six key IoT data quality issues:

- **Dropped readings:** Intermittent communication and lack of resources cause failures in the delivery of readings from IoT devices to the application or system that consumes those data.
- **Unreliable readings:** The credibility of the data collected from IoT devices can become low as some nodes may produce erroneous data either intentionally (e.g., a hacked node) or unintentionally (e.g., due to device calibration failures).
- **Multisource data inconsistencies:** The heterogeneity in devices and data types and the dynamic nature (addition or removal of smart objects) of IoT introduce inconsistencies in IoT data.

- **Data duplication:** A large number of devices may be deployed to collect data, and thus, a data consumer can often receive similar data from many devices.
- **Data leakage:** The data consumer may collect and store IoT data more than necessary, creating the environment for data leakages. IoT data can have sensitive information such as insights about daily life and routines of the users.
- **Multisource data time alignment:** IoT applications combine both dynamic (real-time data) and static from multiple sources. Lack of time alignment of these data sources complicates their integration.

As regards the secure data access, the IoT computing environment can provide data access services at the edge of the IoT network layers (Kayes et al., 2020a). For example, a fog-based access control framework has been proposed in the earlier research to reduce the latency of computation and the cost of processing (Kayes et al., 2018). A secure IoT data governance protocol is an ultimate need to provide mutual authorization and authentication services among three layers: cloud, fog, and IoT (Kayes et al., 2020b).

3.7 Case Studies with Data Governance

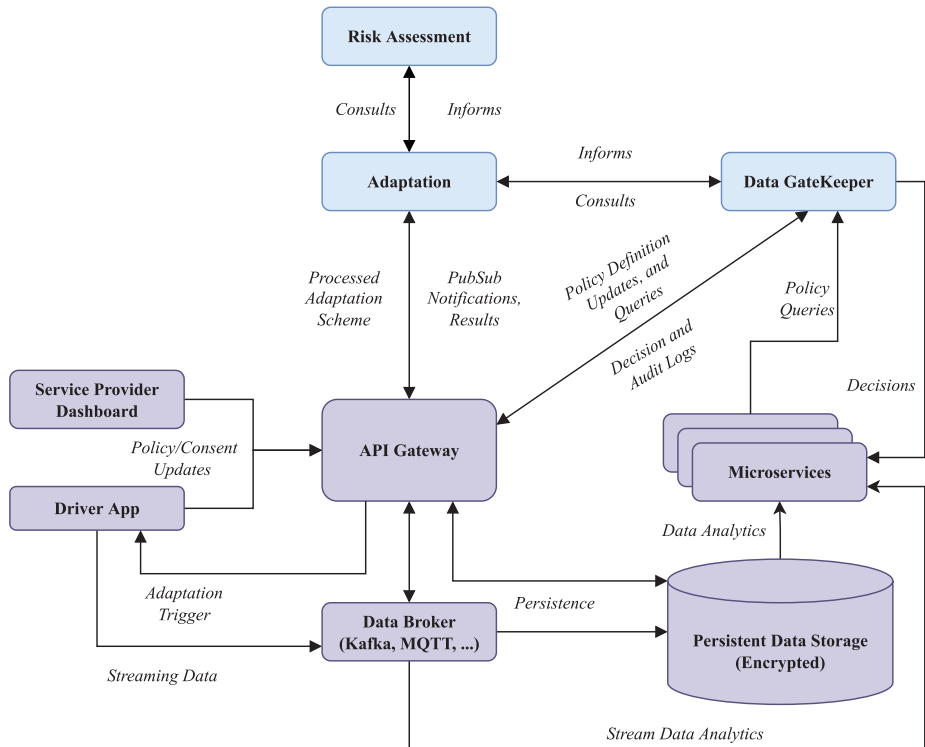
In this section, we present the architectural designs of the data-driven products in our case studies, highlighting architectural support for data governance.

3.7.1 SODALITE Vehicle IoT Architecture

■ Figure 3.6 shows the high-level architecture of the vehicle IoT use case including the components responsible for data governance.

While the deployment can be instantiated hierarchically for the different deployment scenarios discussed earlier, we present a flattened view of the architecture here in order to highlight the roles of specific components that contribute directly to data governance within the system, as well as the overall compliance methodology used by the system. Components with a data governance role in the system include the following:

- **Service Provider Dashboard:** This component provides the service provider's view into the system, pertaining to their deployed service. This component is used for the definition of processing purposes by the data service, including the different types of data desired by the service, as well as the management of consent receipts from users of the data service. Each service is required to break down its data requests between needed and wanted data, such that end users can understand the fundamental data requirements of the service and can also provide additional data in order to derive increased value from the service, in line with each user's privacy preferences.
- **Driver App:** This component reflects the interface that the data subject interacts with to manage their consent to purpose-specific data access by service providers. From the consent definition, a consent receipt is generated and provided to



■ **Fig. 3.6** Inclusion of data governance components into the vehicle IoT use case. (“Author’s own figure”)

both the service provider and the end user, such that consent can be verified by either party at any time. Based on the needed/wanted data specified by the service, a range of consent options can be made by the end user. The codification of processing purposes by the data service provider further requires an explanation for each purpose of processing to be provided in a clear and unambiguous way, allowing the user to provide informed consent.

- **API Gateway:** The API Gateway is the main entry point for different applications into the system. It is placed behind a region-aware router and instantiated per regulatory domain, allowing client connections to be routed to a compliant endpoint. In the event where no suitable deployment is found, it may signal an adaptation event to the Driver App, providing the application with the opportunity to reconfigure itself for device-local processing or to stop any flows of sensitive data that is not able to flow across borders.
- **Data GateKeeper:** This component provides a central policy decision point (PDP) for the deployed system as a whole, with multiple policy enforcement points (PEPs) spreading throughout the system. In addition to providing decisions, the Data GateKeeper can also determine the type of granularity for data access to a service provider, allowing for sensitive data to be treated in line with the data subject’s conditional consent (e.g., a service provider is given consent

to access geolocational data, but only if other identifying information has been filtered out or otherwise anonymized). This enables data to be dynamically anonymized, minimized, and filtered and can be applied to both persisted data at rest and data in motion (streaming data).

- **Risk Monitor:** The run-time risk monitor aims to obtain situational awareness of the deployment and the environment in order to assess end-to-end compliance needs for a given point in time and makes adaptation recommendations to the adaptation manager in order to bring the deployment into compliance. The component itself is event triggered and is triggered on events such as a change in country, deployment, end-user consent, and introduction of sensitive category data.
- **Adaptation Manager:** The adaptation manager is the component responsible for carrying out adaptations and run-time reconfiguration at both the application and infrastructure levels, based on the recommendations of the run-time risk monitor.

In order to facilitate auditing, logs from critical decision points are taken and persisted. This includes changes in end-user consent, all decisions made by the PDPs, enforcement of these decisions by the PEPs, compliance risks identified by the run-time monitor, and adaptations carried out as a result. Non-repudiation of the audit logs is provided by blockchain-based timestamping, allowing log data and timestamp proofs to be transferred to service providers (or any other relevant third party) for independent attestation.

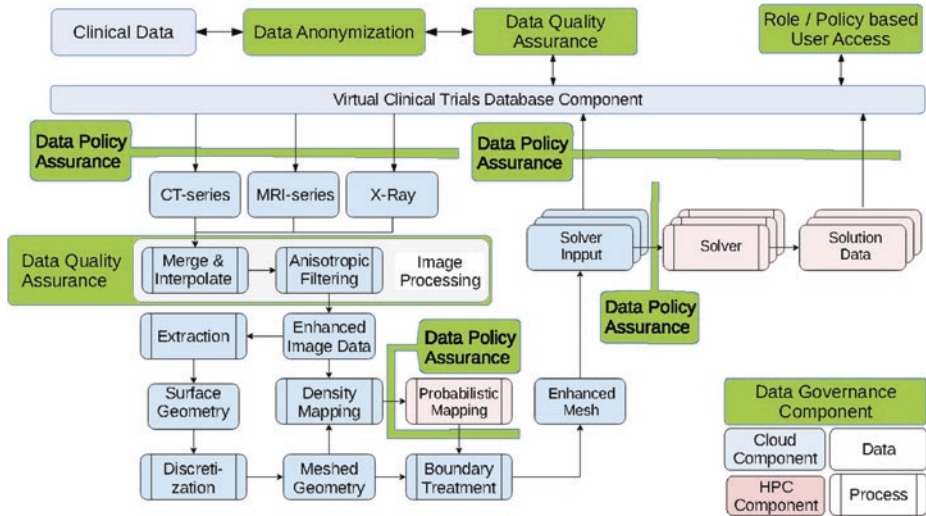
3.7.2 SODALITE Clinical Trial Architecture

Figure 3.7 shows the envisaged architecture of the virtual clinical trial simulation process chain including data governance components.

As can be seen, already the first step on data input is an anonymization component through which the clinical imaging data are passed. In this component, GDPR-sensitive metadata have to be removed which normally are stored by a producing entity like a computer tomography scanner in the so-called DICOM header.⁷ This component has to ensure that the data passed into the central database are correctly anonymized and fully GDPR compliant. If the simulation procedure is in future applications also applied to the cervical spine, besides the treatment of metadata, more complex tasks like data encryption might become necessary in this component, since from clinical 3D imaging data containing the head of the patient, it is in principle possible to reconstruct the face and by that again GDPR-relevant data of the patient.

Due to the nature of the DICOM image standard which allows original equipment manufacturers to put custom header tags into the images, which might prevent images from being parsed properly, a data quality assurance component is

7 ► <https://www.dicomstandard.org>.



■ Fig. 3.7 Inclusion of data governance components into the in silico clinical trial use case. (“Author’s own figure”)

necessary, directly after the anonymization step. This component will ensure standard conformity of data entering the central database component.

In addition to the aforementioned data governance components which ensure the data integrity of the database, a GDPR-compliant role and policy-based access component will be necessary to ensure that only authorized users have access to the database.

Within the simulation process itself, basically two different data governance components are foreseen to be necessary:

- **Data policy assurance component:** This component has to ensure that the data loaded from the database, treated by the different process parts and passed on between different computing resources, are actually authorized for these operations on targeted resources.
- **Data quality assurance component:** A second data quality assurance component is necessary within the process chain and will be part of the image processing steps, which does not ensure standard conformity of the data a second time but is instead intended to ensure the quality of the data by means of statistical methods.

Conclusion

The proper data governance practices enable data products and data services to create values from data while minimizing or eliminating risk exposure of data, and thus are crucial for gaining a competitive advantage and maximizing value from the use of data.

This chapter presented the key aspects of a data governance model for an organization. We first motivated the needs for data governance with two real-world case

studies (data products/services). Next, we provided an overview of a data governance framework, followed by a detailed discussion of data governance decision domains as well as roles and responsibilities of decision makers. We also discussed the contemporary data governance areas such as big data governance and IoT data governance. Finally, we presented the architectural designs of the two case studies that incorporate data governance elements.

Take-Home Messages

The reader can take the following key points from this chapter:

- The data produced and used by the data-intensive products or services should be properly governed in order to achieve the competitive advantages from the data while minimizing the cost and risk of keeping and using the data.
- When organizations are establishing data governance programs, they need to identify the domains of governance decision-making as well as the responsible organizational roles. The common decision domains include data principles, data quality, metadata, data access, and data life cycle. The common roles include executive sponsor, data governance council, data custodian, data steward, and data users.
- Big data and Internet of Things (IoT) bring novel challenges to the contemporary data governance practices in terms of heterogeneity, size, and quality of data.

Acknowledgements This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825480 (SODALITE project).

References

- Al-Ruithe, M., Benkhelifa, E., & Hameed, K. (2019). A systematic literature review of data governance and cloud data governance. *Personal and Ubiquitous Computing*, 23(5–6), 839–859.
- Alshboul, Y., Nepali, R., & Wang, Y. (2015). Big data lifecycle: threats and security model. In *21st Americas Conference on Information Systems*.
- Ball, A. (2012). *Review of data management lifecycle models*. University of Bath, IDMRC.
- Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3).
- Cheong, L. K., & Chang, V. (2007). The need for data governance: a case study. In *ACIS 2007 Proceedings* (p. 100).
- Cumby, R., & Church, P. (2013). Is “big data” creepy? *Computer Law Security Review*, 29(5), 601–609.
- Di Nitto, E., Cruz, J. G., Kumara, I., Radolović, D., Tokmakov, K., & Vasileiou, Z. (2022). *Deployment and operation of complex software in heterogeneous execution environments: The sodalite approach*. Springer.
- Doan, Q., Kayes, A. S. M., Rahayu, W., & Nguyen, K. (2020). Integration of IOT streaming data with efficient indexing and storage optimization. *IEEE Access*, 8, 47456–47467.

- Greenberg, J. (2005). Understanding metadata and metadata schemes. *Cataloging & Classification Quarterly*, 40(3–4), 17–36.
- Karkouch, A., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications*, 73, 57–81.
- Kayes, A. S. M., Han, J., Colman, A., & Islam, Md. S. (2014). Relboss: A relationship-aware access control framework for software services. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"* (pp. 258–276). Springer.
- Kayes, A. S. M., Rahayu, W., Dillon, T., & Chang, E. (2018). Accessing data from multiple sources through context-aware access control. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)* (pp. 551–559).
- Kayes, A. S. M., Kalaria, R., Sarker, I. H., Islam, M., Watters, P. A., Ng, A., Hammoudeh, M., Badsha, S., Kumara, I., et al. (2020a). A survey of context-aware access control mechanisms for cloud and fog networks: Taxonomy and open research issues. *Sensors*, 20(9), 2464.
- Kayes, A. S. M., Rahayu, W., Watters, P., Alazab, M., Dillon, T., & Chang, E. (2020b). Achieving security scalability and flexibility using fog-based context-aware access control. *Future Generation Computer Systems*, 107, 307–323.
- Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, 53(1), 148–152.
- Korhonen, J. J., Melleri, I., Hiekkänen, K., & Helenius, M. (2013). Designing data governance structure: An organizational perspective. *GSTF Journal on Computing*, 2(4), 11–17.
- Liu, L., & Chi, L. (2002). Evolutional data quality: A theory-specific view. In *ICIQ* (pp. 292–304).
- Malik, P. (2013). Governing big data: Principles and practices. *IBM Journal of Research and Development*, 57(3/4), 1–13.
- Marco, D. (2006). Understanding data governance and stewardship, Part 1. *Information Management*, 16(9), 28.
- Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A data quality in use model for big data. *Future Generation Computer Systems*, 63, 123–130.
- Michener, W. K., Allard, S., Budden, A., Cook, R. B., Douglass, K., Frame, M., Kelling, S., Koskela, R., Tenopir, C., & Vieglaiss, D. A. (2012). Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, 11, 5–15.
- Otto, B. (2011). A morphology of the organisation of data governance.
- Perez-Castillo, R., Carretero, A. G., Caballero, I., Rodriguez, M., Piattini, M., Mate, A., Kim, S., & Lee, D. (2018). Daqua-mass: An ISO 8000-61 based data quality management methodology for sensor data. *Sensors*, 18(9), 3105.
- Protection Regulation. (1807). Regulation (EU) 2018/1807 of the European parliament and of the council. *REGULATION (EU)*, 2018, 2018.
- Protection Regulation. (2016). Regulation (EU) 2016/679 of the European Parliament and of the council. *REGULATION (EU)*, 679, 2016.
- Riley, J. (2017). *Understanding metadata* (p. 23). National Information Standards Organization. Retrieved from <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>
- Sandhu, R. S., Coyne, E. J., Feinstein, H. L., & Youman, C. E. (1996). Role-based access control models. *Computer*, 29(2), 38–47.
- Servos, D., & Osborn, S. L. (2017). Current research and open problems in attribute-based access control. *ACM Computing Surveys*, 49(4).
- Sinaeepourfard, A., Garcia, J., Masip-Bruin, X., & Marin-Torder, E. (2016). Towards a comprehensive data lifecycle model for big data environments. In *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, BDCAT '16 (pp. 100–106). Association for Computing Machinery.
- Singh, G., Bharathi, S., Chervenak, A., Deelman, E., Kesselman, C., Manohar, M., Patil, S., & Pearlman, L. (2003). A metadata catalog service for data intensive applications. In *Proceedings of the 2003 ACM/IEEE Conference on Supercomputing*, SC '03 (p. 33). Association for Computing Machinery.
- Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, 40(5), 103–110.

- Taleb, I., Dssouli, R., & Serhani, M. A. (2015). Big data pre-processing: A quality framework. In *2015 IEEE International Congress on Big Data* (pp. 191–198).
- Taleb, I., Kassabi, H. T. E., Serhani, M. A., Dssouli, R., & Bouhaddioui, C. (2016). Big data quality: A quality dimensions evaluation. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/IATC/ScalComl/CBDComl/IoP/SmartWorld)* (pp. 759–765).
- Tallon, P. P. (2013). Corporate governance of big data: Perspectives on value, risk, and cost. *Computer*, *46*(6), 32–38.
- van der Aalst, W. M. P. (2017). Responsible data science: Using event data in a “people friendly” manner. In S. Hammoudi, L. A. Maciaszek, M. M. Missikoff, O. Camp, & J. Cordeiro (Eds.), *Enterprise information systems* (pp. 3–28). Springer International Publishing.
- van der Aalst, W. M. P., Bichler, M., & Heinzl, A. (2017). Responsible data science. *Business & Information Systems Engineering*, *59*(5), 311–313.
- Villar, M. (2009). Establishing effective business data stewards. *Business Intelligence Journal*, *14*(2), 23–29.
- Weber, K., Otto, B., & Österle, H. (2009). One size does not fit all—a contingency approach to data governance. *Journal of Data and Information Quality*, *1*(1).
- Wende, K. (2007). A model for data governance-organising accountabilities for data quality management.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9.
- Zomaya, A. Y., & Sakr, S. (2017). *Handbook of big data technologies*. Springer.



Big Data Architectures

*Martin Garriga, Geert Monsieur,
and Damian Tamburri*

Contents

- 4.1 Introduction – 64**
- 4.2 Background – 65**
 - 4.2.1 Key Attributes of Big Data Systems – 65
 - 4.2.2 From Structured Data to Semi-structured Data – 66
- 4.3 Lambda Architecture – 69**
- 4.4 Kappa Architecture – 71**
- 4.5 SEI-CMU Reference Architecture – 72**
- References – 76**

Learning Objectives

After reading this chapter, readers will be able to:

- Identify and describe different types of big data architectures and their main characteristics.
- Analyze a business domain as a big data problem, and compare and contrast big data architectures to solve it.
- Ultimately sketch an architecture as an instance of any of the big data architectures, to solve any given big data problem.

4.1 Introduction

It is already true that big data has drawn huge attention from researchers in information sciences and policy and decision makers in governments and enterprises. As the speed of information growth exceeded Moore's law at the beginning of this new century, excessive data is making great troubles to businesses and organizations. Nevertheless, great potential and highly useful value are hidden in the huge volume of data.

Many business cases exploiting big data have been realized in recent years. For example, LinkedIn (Sumbaly et al., 2013) collects data from users and offers services such as “People you may know,” skill endorsements, or news feed updates to end users based on the analysis of the data. Netflix uses big data for providing recommendations and ranking-related services to customers (Amatriain, 2013). Twitter uses collected data for real-time query suggestion and spelling corrections of their search algorithm (Mishne et al., 2013). Moreover, the implementation architectures for these use cases have been published in an effort to ease future solutions.

In this chapter, we take an evolutionary view on big data from the perspective of academics, entrepreneurs, and practitioners alike, by discussing fundamentals on big data architectures. We analyze technology-independent reference architecture for big data systems, stemming from the analysis of published implementation architectures of outstanding big data use cases. Those are represented by three reference architectures that encompass main features and challenges on big data: Lambda and Kappa architecture from the industry practitioners' point of view and the reference architecture from SEI-CMU from the academic point of view.

The rest of this chapter is organized as follows. ► Section 4.2 presents useful background on big data systems. ► Section 4.3 presents the Lambda architecture. ► Section 4.4 details the Kappa architecture. ► Section 4.5 discusses the SEI-CMU reference architecture. ► Section 4.6 presents the main discussion points stemming from the analysis of the architectures. Finally, ► Sect. 4.6 concludes the chapter with main take-home messages.

4.2 Background

4.2.1 Key Attributes of Big Data Systems

The properties architects strive for in big data systems are as much about complexity as they are about scalability. Not only must a big data system perform well and be resource efficient, it must be easy to reason about as well. These properties are as follows (Marz & Warren, 2015):

Definition 4.1 (Robustness and Fault Tolerance)

Distributed systems need to behave correctly despite machines going down randomly, complex semantics of consistency in distributed databases, duplicated data, concurrency, and more. These challenges make it difficult even to reason about what a system is doing. Part of making a big data system robust is avoiding these complexities so that you can easily reason about the system.

Definition 4.2 (Low Latency Reads and Updates)

The vast majority of applications require reads to be satisfied with very low latency, typically between a few milliseconds to a few hundred milliseconds.

Definition 4.3 (Scalability)

The ability to maintain performance in the face of increasing data or load by adding resources to the system.

Definition 4.4 (Generalization)

A general system can support a wide range of applications.

Definition 4.5 (Extensibility)

You do not want to have to reinvent the wheel each time you add a related feature or make a change to how your system works. Extensible systems allow functionality to be added with a minimal development cost.

Definition 4.6 (Ad Hoc Queries)

Nearly every large dataset has unanticipated value within it. You must be able to mine a dataset arbitrarily. The problems with fully incremental architectures give opportunities for business optimization and new applications. Ultimately, you cannot discover interesting things to do with your data unless you can ask arbitrary questions of it.

Definition 4.7 (Minimal Maintenance)

This includes anticipating when to add machines to scale, keeping processes up and running, and debugging anything that goes wrong in production. An important part of minimizing maintenance is choosing components that have as little implementation complexity as possible.

Definition 4.8 (Debuggability)

The key is to be able to trace, for each value in the system, exactly what caused it to have that value.

4.2.2 From Structured Data to Semi-structured Data

Information systems typically rely on a database management system (DBMS) to store and retrieve data from a database. A DBMS is based on a database model for describing the structure, the consistency rules, and the behavior of a database. Traditional information systems and database management systems (DBMSs) typically are developed to cater for the processing of *structured data*. The de facto standard database model for structuring data is the *relational model*, which essentially represents data in a tabular fashion. A relational database typically consists of a set of related tables, in which each table has a static structure of columns that can hold data of one specific type (e.g., numbers, text values). The blueprint of a database structure is referred to as the *database schema*, which is very static in the case of relational database, implying that each record in a relational database has to comply to the structure formalized in the database schema.

Relational DBMSs, which perform very well with structured data, often are less appropriate for big data applications. This can be understood by reconsidering the three-dimensional Vs that are associated with big data and were already discussed in the ► Chap. 1 on big data engineering:

Definition 4.9 (Volume)

The most straightforward way to provide data consistency and realize a rich query model is to have all data stored in a database on a single machine. However, the data volumes in big data applications tend to overtake the limits of scaling up machines running database systems. Hence, it simply becomes impossible to store all data on one single machine, and, therefore, database systems for big data applications require the ability to *scale horizontally* and organize data in a much more *distributed* fashion.

Definition 4.10 (Variety)

Traditional DBMSs have weaker support for the increasing variety present in big data. First, the variety demands for more flexible or even the absence of database schemas. Dynamic business environments force systems to deal with data that do not always obey the formal static structure of data models associated with relational databases. DBMSs for big data applications require better support for *semi-structured data*, in which entities may still be grouped together, although they do not share the exact same attributes, and data is self-describing by including meta-data (e.g., tags in XML documents) and other markers to distinguish different fields and records. Second, modern big data applications may also include *unstructured data* like natural language, video, and images.

Definition 4.11 (Velocity)

The high velocity of data arriving from a data source results in massive streams of data that require time-consuming cleansing and transformation processes when needed to be inserted in structured tables in a relational database. Clearly, DBMSs specifically designed for storing unstructured and streaming data are more appropriate in this context.

DBMSs addressing the above challenges are often referred to as non-SQL, nonrelational, or not only SQL (NoSQL) DBMSs. Most NoSQL DBMSs are distributed systems and especially designed for semi-structured and schemaless data storage, high performance, availability, data replication, and scalability, as opposed to immediate data consistency, powerful query languages, and structured data storage in relational DBMSs (Elmasri & Navathe, 2015).

It is beyond the scope of this book to present a complete overview on NoSQL DBMSs. In general, NoSQL DBMSs can be classified into four main categories:

Definition 4.12 (Key-Value Store)

Data is organized in associative arrays, a.k.a. a dictionary or hash table, resulting in a collection of objects, which in turn have many different fields within them, each containing data. Popular key-value stores are Redis, Amazon DynamoDB, and Memcached. Key-value stores focus on high performance, availability, and scalability. A key serves as a unique identifier associated with a data item. The value can be structured (e.g., tuples similar to rows in a relational database), unstructured (string of bytes), or semi-structured self-describing data (e.g., JSON). Key-value stores have fast data retrieval using keys, but do not come with query languages.

Definition 4.13 (Wide-Column Store)

Data is structured in dynamic columns, as opposed to rows with static columns in relational databases. Wide-column databases can be interpreted as multidimensional key-value stores, in which the key is composed of the table name, row key, and column name. Additionally, columns can be composed of a column family and column qualifier. While a column family needs to be defined at table creation, column qualifiers are variable and may differ greatly between rows. Popular wide-column DBMSs are Cassandra, HBase, and Google Cloud Bigtable.

Definition 4.14 (Graph Database)

Data is represented as a graph, which basically is a collection of schema-free objects (vertices or nodes) and relationships between the objects (edges). Data is stored as properties of nodes or edges and can be unstructured, structured, or semi-structured. Graph DBMSs come with query languages optimized for graph-like queries such as shortest path calculation or community detection. Popular graph DBMSs are Neo4j and OrientDB.

Definition 4.15 (Document-Oriented Database)

Data is stored in documents, typically in some standard format or encoding such as XML or JSON. A document-oriented database groups data into collections of *similar* documents. Documents are self-describing. Although more and more document-oriented DBMSs support languages like XML schema or JSON schema, schemas are not required, allowing documents in one collection to have different data elements. Popular document-oriented DBMSs are MongoDB and Couchbase.

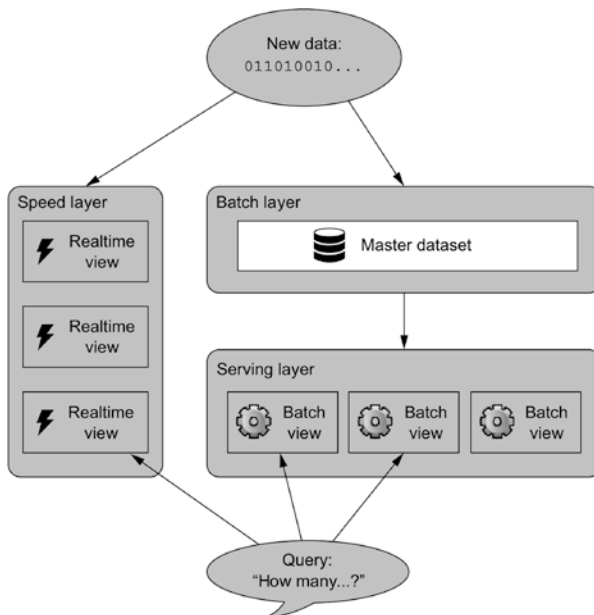
4.3 Lambda Architecture

Ideally, one could run any data processing functions (independently of how big is the data) on the fly to get the results. Unfortunately, even if this were possible, it would take a huge amount of resources to do and would be unreasonably expensive.

The Lambda architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem and the target big data system into three layers: the batch layer, the serving layer, and the speed layer (Philip Chen & Zhang, 2014), as shown in ■ Fig. 4.1. Each layer satisfies a subset of properties and builds upon the functionality provided by the layers beneath it.

The **batch layer** stores the master copy of the dataset and precomputes batch views on it. The master dataset can be thought of as a very large list of records. The batch layer needs to be able to do two things: store an immutable, constantly growing master dataset, and compute arbitrary functions on that dataset. This type of processing is best done using batch processing systems. Hadoop is the canonical example of a batch processing system (Marz & Warren, 2015).

The batch layer emits batch views as the result of its functions. The next step is to load the views somewhere so that they can be queried. This is where the **serving layer** comes in. The serving layer is a specialized distributed database that loads in a batch view and makes it possible to do random reads on it.



■ Fig. 4.1 Lambda architecture diagram. (Adapted from (Marz & Warren, 2015))

The serving layer updates whenever the batch layer finishes precomputing a batch view. This means that the only data not represented in the batch view is the data that came in while the precomputation was running. What is still missing to have a fully real-time data system—that is, to compute arbitrary functions on arbitrary data in real time—is to compensate for those last few hours of data. This is the purpose of the **speed** layer. As its name suggests, its goal is to ensure that new data is represented in query functions as quickly as needed for the application requirements. The speed layer updates the real-time views as it receives new data instead of recomputing the views from scratch like the batch layer does. The speed layer does incremental computation instead of the recomputation done in the batch layer (Marz & Warren, 2015).

Lambda architecture can be deployed for those data processing enterprise models where (Samizadeh, 2018):

- User queries are required to be served on an ad hoc basis using the immutable data storage.
- Quick responses are required and system should be capable of handling various updates in the form of new data streams.
- None of the stored records shall be erased and it should allow addition of updates and new data to the database.

The benefits of data systems built using the Lambda architecture go beyond just scaling. As the system handles much larger amounts of data, it becomes possible to get more value out of it. Increasing the amount and types of stored data will lead to more opportunities to mine the data, produce analytics, and build new applications. Another benefit of using the Lambda architecture is how robust the applications will be, for example, able to run computations on the whole dataset to do migrations or fix things that go wrong.

One can avoid having multiple versions of a schema active at the same time. When the schema changes, it is possible to update all data to the new schema. Likewise, if an incorrect algorithm is accidentally deployed to production and corrupts the data, one can fix things by recomputing the corrupted values, making big data applications more robust. Finally, performance will be more predictable. Although the Lambda architecture as a whole is generic and flexible, the individual components comprising the system are specialized. Very little “magic” happens behind the scenes, as compared to something like a SQL query planner. This leads to more predictable performance (Marz & Warren, 2015).

However, it is also clear that this architecture cannot fit all the big data applications (Philip Chen & Zhang, 2014). The problem with the Lambda architecture is maintaining two complex distributed systems (batch and speed layer) to produce the same result. Ultimately, even if one can avoid coding the application twice, the operational burden of running and debugging two systems is going to be very high (Kreps, 2014).

Additionally, intermediate results of the batch layer are written to the file system, resulting in higher latency as job pipelines grow in length. Despite many

efforts to reduce access latency, the coarse-grained data access of a MapReduce and Distributed File System stack is only appropriate for batch-oriented processing, limiting its suitability for low-latency backend systems (Fernandez et al., 2015).

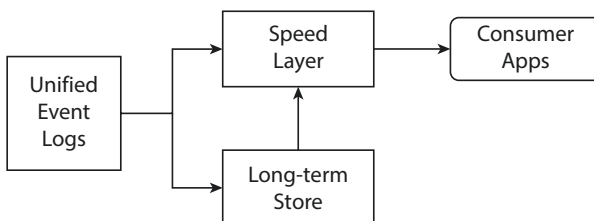
4.4 Kappa Architecture

Lambda architectures allowed to build complex, low-latency processing systems, featuring a scalable high-latency batch system that can process historical data and a low-latency stream processing system that cannot reprocess results (Kreps, 2014).

However, as the technology and frameworks for stream processing evolved, the challenges also changed: Why not handle the full problem set with stream processing? Why do one need to glue on another system (i.e., the batch one)? Why can't we do both real-time processing and also handle the reprocessing when data or code changes? Stream processing systems already have a notion of parallelism: Why not just handle reprocessing by increasing the parallelism and replaying history at speed? The answer to these challenges gave birth to a novel architecture, namely the *Kappa* architecture.

The intuition behind *Kappa* architecture is that stream processing is inappropriate for high-throughput processing of historical data based mostly on the limitations of early stream processing systems, which either scale poorly or lose historical data. In that case, stream processing system is inherently something that computes results of some ephemeral streams and then throws all the underlying data away. But there is no reason this should hold true. The fundamental abstraction in stream processing to represent data flows is directed acyclic graphs (DAGs). Stream processing is just a generalization of this data flow model that exposes checkpointing of intermediate results and continual output to the end user.

■ Figure 4.2 shows the key components of the *Kappa* architecture. The unified event log uses a distributed messaging system to retain the full log of the input data. The speed layer processes the data in real time as they become available. A copy of the data is also stored in a persistent storage. To do the reprocessing of data, start a second instance of the stream processing job that starts processing from the beginning of the retained data, but direct this output data to a new output table.



■ Fig. 4.2 Kappa architecture. (“Author’s own figure”)

Unlike the Lambda architecture, the reprocessing occurs only when the processing code changes, and thus one actually needs to recompute results. Note that the job of doing the recomputation may be just an improved version of the same code, running on the same framework, taking the same input data. Naturally, the reprocessing job should ramp up the parallelism, so it completes very quickly. Of course, one can optimize this further, e.g., by combining the two output tables. However, having both for a short period of time allows to revert back instantaneously to the old logic by just redirecting the application to the old table. And in critical use cases, one can control the cutover with an automatic A/B test or bandit algorithm to ensure that bug fixes or code improvements do not degrade performance in comparison to the prior version.

A stream processor in a Kafka architecture just maintains an “offset,” which is the log entry number for the last record it has processed on each data partition. So, changing the consumer’s position to go back and reprocess data is as simple as restarting the job with a different offset. Adding a second consumer for the same data is just another reader pointing to a different position in the log.

4.5 SEI-CMU Reference Architecture

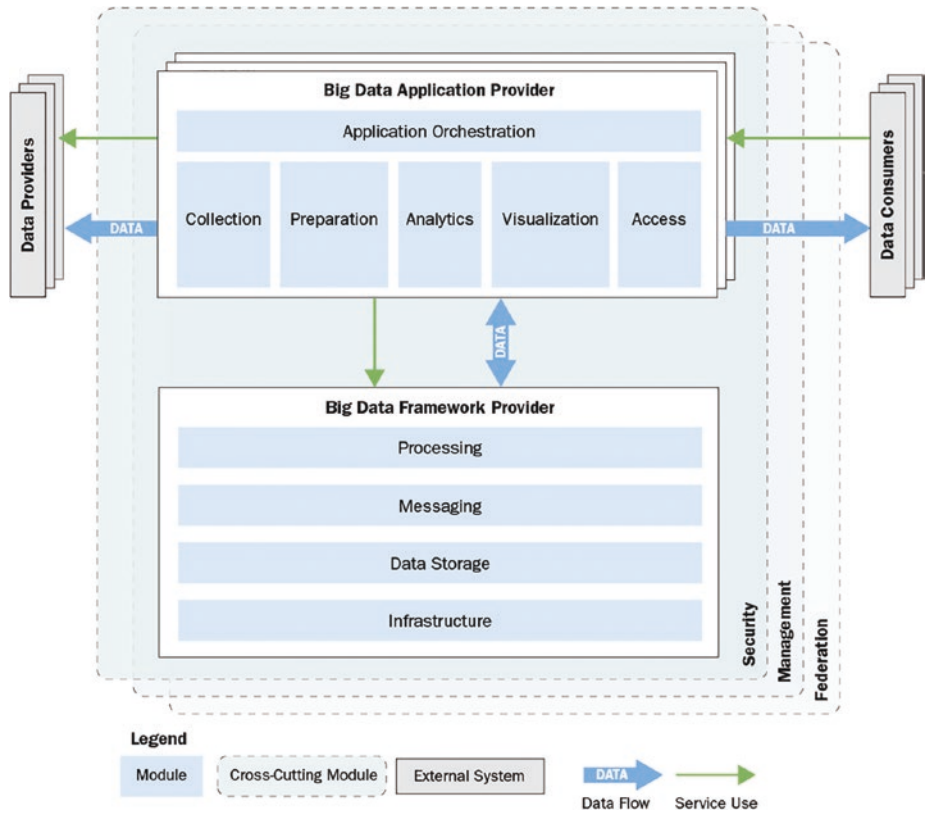
Different big data architectures have been proposed also from the academic side. One of the first and most widespread ones is the reference architecture (RA) for big data systems in the national security domain (Klein et al., 2016), from the Software Engineering Institute, Carnegie Mellon University (SEI-CMU).

A reference architecture facilitates software engineering activities—in this case, building a big data architecture—by standardizing nomenclature, defining key solution elements and their relationships, collecting relevant solution patterns, and classifying existing technologies. Within the national security domain, existing reference architectures for big data systems such as Lambda and Kappa have not been useful because they are too general or are not vendor neutral. The reference architecture for big data systems is focused on addressing typical national defense requirements and is vendor neutral. Furthermore, this reference architecture can be applied to define solutions in other domains.

The reference architecture assumes a system of systems context, where data providers and data consumers are external systems that are not under the same design or operational authority as the big data system. These systems may be instances of big data systems developed using this RA (or another architecture). The architecture is shown in ■ Fig. 4.3. The 13 modules are grouped into three categories.

Big data application provider modules: This includes application-level business logic, data transformations and analysis, and functionality to be executed by the system.

- *Application orchestration* configures and combines other modules of the big data application provider, integrating activities into a cohesive application. An application is the end-to-end data processing through the system to satisfy one or more use cases.



■ Fig. 4.3 SEI-CMU reference architecture. (Adapted from (Klein et al., 2016))

- The *collection* module is the interface to external data providers, matching the characteristics and constraints of the providers and avoiding data loss.
- The *preparation* module transforms data to make it useful for the other downstream modules, in particular analytics. Preparation performs the transformation portion of the traditional extract, transform, load (ETL) cycle, including tasks such as data validation, cleansing, optimization, schema transformation, and standardization.
- The *analytics* module extracts knowledge from the data, typically working with multiple datasets and characteristics. Analytics can contribute further to the transform stage of the ETL cycle by performing more advanced transformations and enrichment to support knowledge extraction.
- The *visualization* module presents processed data and outputs of analytics to a human data consumer, in a format that communicates meaning and knowledge. It provides a “human interface” to the big data.
- The *access* module interacts with external actors. Unlike visualization, which addresses “human interfaces,” the access module is concerned with “machine interfaces” (e.g., APIs or Web services). The access module is the intermediary between the external world and the big data system to enforce security or provide load balancing capability.

Big data framework provider modules: This includes the software middleware, storage, and computing platforms and networks used by the big data application provider. As shown in ■ Fig. 4.3, the system may include multiple instances of the big data application provider, all sharing the same instance of the big data framework provider:

- The *processing* module provides efficient, scalable, and reliable execution of analytics. It provides the necessary infrastructure to support execution distributed across tens to thousands of nodes, defining how the computation and processing are performed.
- The *messaging* module supports reliable queuing, transmission, and delivery of data and control functions between components. While messaging is common in traditional IT systems, its use in big data systems creates additional challenges.
- The *data storage* module provides reliable and efficient access to persistent data. This includes the logical data organization, data distribution and access methods, and data discovery (using, e.g., metadata services, registries, and indexes).
- The *infrastructure* module provides the infrastructure resources necessary to host and execute the activities of the other BDRA modules.

Cross-cutting modules: This includes concerns that impact nearly every module in the other two categories:

- *Security* module controls access to data and applications, including enforcement of access rules and restricting access based on classification or need-to-know.
- The *management* module addresses two main concerns, namely *system management*, including activities such as monitoring, configuration, provisioning, and control of infrastructure and applications, and *data management*, involving activities surrounding the data life cycle of collection, preparation, analytics, visualization, and access.
- The *federation* module provides interoperation between federated instances of the RA. These concerns are similar to typical system of systems (SoS) federation concerns (Maier, 1998); however, existing SoS federation strategies may not support the scale of big data systems.

Finally, a set of *common concerns* do not map to any specific modules but should be considered in the architecture of a big data system. These include the following:

- *Scalability* (increase or decrease the processing and storage provided, in response to changes in demand), *availability* (remain operational during fault conditions such as network outages or hardware failures), and *data organization* (design of data considering downstream performance)
- *Technology stack* decisions (both hardware and software)
- *Accreditation* (cybersecurity qualities of the system)

Conclusion

Big data computing is an emerging platform for data analytics to address large-scale multidimensional data for knowledge discovery and decision-making. In this chapter, we have discussed foundational aspects and examples of big data architectures. Big data technology is evolving and changing the present traditional databases with effective data organization, large computing, and data workload processing with new innovative analytics tools bundled with statistical and machine learning techniques. With the maturity of cloud computing technologies, big data technologies are accelerating in several areas of business, science, and engineering to solve data-intensive problems. We have enumerated a few case studies that apply big data technologies and architectures. The domain applications are manifold, ranging from healthcare studies, business intelligence, and social networking to scientific explorations (Kune et al., 2016). Further, we focus on illustrating how big data architectures differ from traditional databases.

However, more research needs to be undertaken, in several areas like data organization, decision-making, domain-specific tools, and platform tools to create next-generation big data infrastructure for enabling users to extract maximum utility out of the large volumes of available information and data (Kune et al., 2016). At the same time, we have a growing social understanding of the consequences of big data. We are only beginning to scratch the surface today in our characterization of data privacy and governance, as discussed in the previous chapter. Our appreciation of the ethics of data analysis is also in its infancy. Mistakes and overreach in this regard can very quickly lead to backlash that could close many things down. But barring such mishaps, it is safe to say that big data may be hyped, but there is more than enough substance for it to deserve our attention (Jagadish, 2015).

Take-Home Messages

The reader can take the following key points from this chapter:

- Organizations have to design big data architectures to handle the ingestion, processing, and analysis of huge volumes of their business data. Several data architectures have been proposed by the researchers and practitioners. Depending on the functional and nonfunctional requirements of their big data applications, organizations need to adopt the most suited data architecture to commence their big data journey.
- Lambda architecture and Kappa architecture are originated from the industry. Both are capable of efficient data processing of massive datasets. Compared with Lambda, Kappa uses a unified layer for both batch and stream data processing.
- SEI-CMU reference architecture can be used as a template to create custom data architectures.

References

- Amatriain, X. (2013). Big & personal: Data and models behind Netflix recommendations. In *Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications* (pp. 1–6).
- Elmasri, R., & Navathe, S. B. (2015). *Fundamentals of database systems* (7th ed.). Pearson.
- Fernandez, R. C., Pietzuch, P. R., Kreps, J., Narkhede, N., Rao, J., Koshy, J., Lin, D., Riccomini, C., & Wang, G. (2015). Liquid: Unifying nearline and offline big data integration. In *CIDR*.
- Jagadish, H. V. (2015). Big data and science: Myths and reality. *Big Data Research*, 2(2), 49–52.
- Klein, J., Buglak, R., Blockow, D., Wuttke, T., & Cooper, B. (2016). A reference architecture for big data systems in the national security domain. In *2016 IEEE/ACM 2nd International Workshop on Big Data Software Engineering (BIGDSE)* (pp. 51–57). IEEE.
- Kreps, J. (2014). *Questioning the lambda architecture*. Retrieved from <https://www.oreilly.com/radar/questioning-the-lambda-architecture/>
- Kune, R., Konugurthi, P. K., Agarwal, A., Chillarige, R.-d. R., & Buyya, R. (2016). The anatomy of big data computing. *Software: Practice and Experience*, 46(1), 79–105.
- Maier, M. W. (1998). Architecting principles for systems-of-systems. *Systems Engineering: The Journal of the International Council on Systems Engineering*, 1(4), 267–284.
- Marz, N., & Warren, J. (2015). *Big Data: Principles and best practices of scalable real-time data systems*. Manning Publications.
- Mishne, G., Dalton, J., Li, Z., Sharma, A., & Lin, J. (2013). Fast data in the era of big data: Twitter's real-time related query suggestion architecture. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 1147–1158).
- Philip Chen, C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275, 314–347.
- Samizadeh, I. (2018). A brief introduction to two data processing architectures—Lambda and kappa for big data. Retrieved from <https://towardsdatascience.com/a-brief-introduction-to-two-data-processing-architectures-lambda-and-kappa-for-big-data-4f35c28005bb>
- Sumbaly, R., Kreps, J., & Shah, S. (2013). The big data ecosystem at LinkedIn. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 1125–1134).



Data Engineering in Action

*Giuseppe Cascavilla, Stefano Dalla Palma,
Stefan Driessen, Willem-Jan van den Heuvel,
Daniel De Pascale, Mirella Sangiovanni,
and Gerard Schouten*

Contents

- 5.1 Introduction – 79**
- 5.2 The ANITA Project for the Fighting of Cybercrime – 79**
 - 5.2.1 Data Collection – 80
 - 5.2.2 ANITA Architecture – 80
 - 5.2.3 Data Extraction – 81
 - 5.2.4 Data Management and Analysis – 82
- 5.3 The P_RoTECT Project for the Protection of Public Spaces – 83**
 - 5.3.1 Objectives of P_RoTECT – 84
 - 5.3.2 P_RoTECT and the Data Fusion Approach – 86
- 5.4 The Beehives Project for the Quality of Urban Biodiversity – 89**
 - 5.4.1 Problem Description – 90
 - 5.4.2 Objectives – 91

5.4.3	Data Gathering – 91
5.4.4	Big Data Analytics for Biodiversity – 93
5.4.5	Systemic Change – 94
5.4.6	Bringing It All Together: The IoT Bee- hive Stratified Architecture – 94
	References – 96

Learning Objectives

After reading this chapter, readers will:

- Understand how big data and data engineering may support in building value-added end-to-end pipelines for data entrepreneurship.
- Identify the trade-offs to take into account when selecting suitable technologies for a big data pipeline.
- Recognize the qualities of real-world big data problems that can lead to decisions concerning such trade-offs.
- Ultimately sketch a data pipeline to handle large flow of data during the system life cycle.

5.1 Introduction

In this chapter, we will guide the reader through the concepts introduced in the previous chapters by introducing three real-case scenarios: one from cybercrime fighting in the form of illegal trafficking activities on the web (► Sect. 5.3), one concerning the protection of public spaces against terrorist attacks (► Sect. 5.4), and one scenario about the monitoring of urban biodiversity exploiting and processing streaming data from IoT at the edge.

For each scenario, the goal of the project is described followed by the data generation/collection approach and the architecture used for processing that data. Each scenario shows a different level of technology in the context of big data and demonstrates the relevant trade-offs and decisions that were made to resolve the challenges at hand. This allows the reader to relate the concepts and ideas introduced in the previous chapters to real-world problems and to follow the decision process of data engineering “in action.”


We believe that these scenarios are illustrative and useful examples that introduce the reader to the remarkable and complex word of “data engineering in practice” because they come with very clear distinctive approaches for the presented data engineering concepts, techniques, methods, and tools in each of the previous chapters.

5.2 The ANITA Project for the Fighting of Cybercrime

The Advanced Tools for Fighting Online Illegal Trafficking (ANITA) is a project funded by the European Commission to ameliorate the investigation capabilities of the law enforcement agencies (LEAs) by delivering a set of tools and techniques to efficiently address online illegal trafficking activities on the web, such as the sale of counterfeit/falsified medicines, drugs, and weapons.

To support LEAs in more powerful investigation activities, ANITA furnishes a reusable platform system that compounds big data analysis and knowledge

management technologies to analyze the online content of a variety of sources (e.g., “surface,” “deep,” and “dark” web; offline LEAs’ databases; and more). Analysis ranges over different data formats that include text, audio, video, and image and helps discovering so-called black markets: underground economy activities characterized by some form of noncompliant behavior with an institutional set of rules.¹

The ANITA system is based on the design and implementation of a scalable and big data-oriented infrastructure, which revolves around a reusable end-to-end pipeline able to crunch large volumes of data in near real time and to summarize analysis results to provide LEAs with relevant insights on illegal trafficking-related phenomena. This end-to-end pipeline basically supports three main steps, which are illustrated in  Fig. 5.2 and described below.

5.2.1 Data Collection


As explained in the previous chapters, big data is typically characterized with (at least) three *Vs*: *volume*, *velocity*, and *variety*. We are in a flood of data that is investing large volumes of high-speed data with a lot of variety. With all this data comes information and with that information comes the potential for innovation. However, this does not automatically imply that more data means more or better information. The data might be noisy and must be collected and managed properly.


The web in general, and dark web in particular, is delineated by a huge volume of data coming from different sources and varying every minute. Therefore, the first step of the pipeline comprises the *detection* and *assessment* of relevant sources about counterfeit medicine, drugs, guns, weapons, terrorism funding, and more.

This assumes the development of tools to crawl raw (dark) web pages related to illegal marketplaces. Specifically, dedicated services are responsible for the fast download and storage of the crawled pages and for the dynamic integration of new marketplaces in real time, and the collected web pages are then stored for later processing.

The output of this module constitutes the input for the next ones, which will proceed with the extraction and analysis of the data under different aspects.

5.2.2 ANITA Architecture

In  Fig. 5.1 is depicted an overview of the architecture behind the Trend Analysis tool developed as part of the ANITA framework.

In the bottom of  Fig. 5.1, we have the four modules involved. The *import module* in charge of loading into the system the HTML files and validate them. The *scrape module* that oversees the extraction of textual information from the product

1 ► <https://www.normshield.com/deep-web-and-black-market/>.

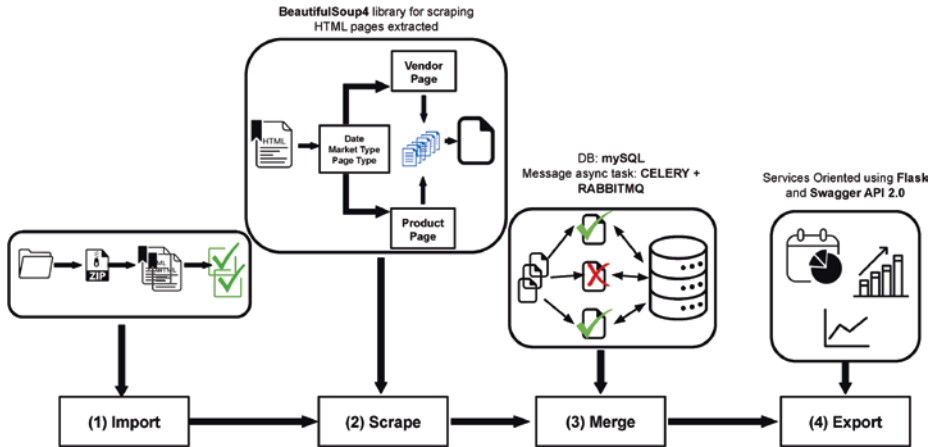


Fig. 5.1 Graphical representation of the Trend Analysis tool from ANITA framework and related technologies involved. (“Author’s own figure”)

pages and the vendor pages. Hence, the *merge module* is responsible for storing the data and removing duplicates. Lastly, the *export module* creates the final visualization of data and analysis. On top of each module are briefly mentioned the technologies involved.

5.2.3 Data Extraction

Once relevant data sources have been listed, the actual extraction of meaningful data can take place. A scraper, a computer program that extracts data from human-readable outputs, is used to parse and refactor the content of the web pages to convert data into a more suitable format for actual data analytics, visualization, and user interpretation. Nevertheless, an improper data representation will reduce the value of the original data and may even obstruct effective data analysis.

Generally, in black marketplaces, there exists a relatively high level of redundancy where the same product appears in different pages. For example, a gun may be sold by the same vendor in different pages. When that happens, the only way to identify this redundancy is by analyzing the content of both pages. To deal with the problem of redundancy of equivalent data and improper representation in ANITA, we have instrumented the merge module with mechanisms.

The merge module takes as input the collected data from the scraper and is in charge of consolidating and loading it into a target relational database system (the sink). It merges all the scraped page information into one specific dataset with two tables: one for the vendors and the other for the products. While doing the merge operation, it purges all duplicates.

5.2.4 Data Management and Analysis

We have observed that the analytical system of big data shall process masses of heterogeneous data within a limited, acceptable time frame. Unfortunately, however, traditional relational databases are strictly designed for more conventional, small-scale data, with a lack of scalability and expandability, while non-relational databases have shown their advantages in the processing of unstructured data.

Nevertheless, although the data coming from the web is mainly unstructured by its nature, and as such suitable for non-relational databases, ANITA relies on relational databases. The reason is that the data was already preprocessed and structured in the previous step of data extraction. Therefore, the non-relational database in this context is typically useless and inefficient.

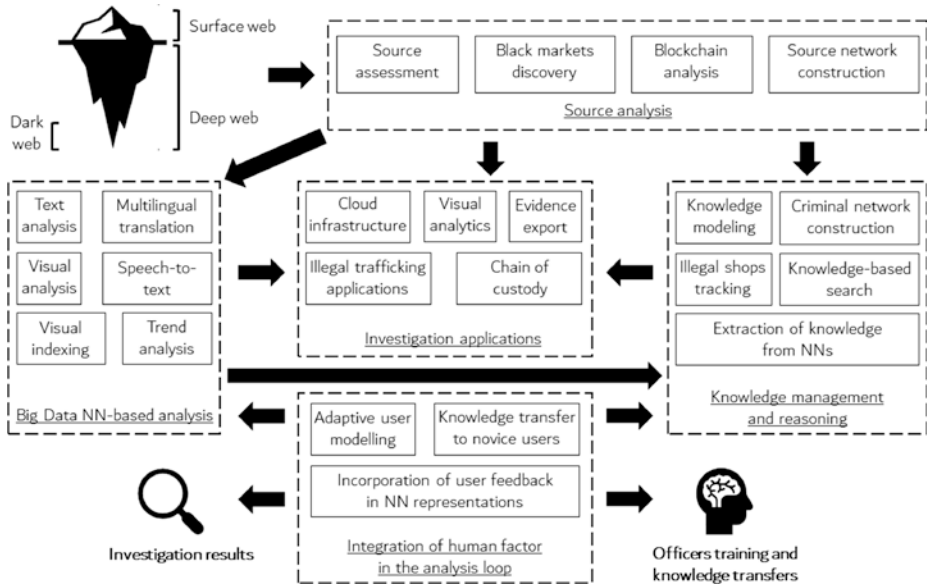
5

After the identification and collection of a vast amount of multimedia material related to illegal trafficking, sophisticated big data analytic services are applied to manipulate, analyze, and semantically organize the acquired information and to detect meaningful events. In this context, “meaningful events” are anomalous increments of illegal products sold, for example, an increment of 1000% of guns sold in a given time range.

Afterward, a semantic-based engine, able to automatically extract and categorize entities from the web contents, is used to deliver *text analysis services*. Among others, these services allow for stylometric analysis to link criminal and terrorist groups. Then, a process called *visual content analysis* is performed to identify interesting information and evidences in the formed databases. *Multilingual translation services* are also developed to support the processing of documents written in different languages. These comprise the automatic translation of segments of speech between different languages, as well as the transformation of audio streams to written documents. This is achieved by machine learning and deep neural network (DNN) algorithms.

Finally, trends of illegal trafficking are analyzed in the latest process named *trend analysis*. The goal of such an analysis is to get insights on the user’s buying habits and behavior. The trend analysis is particularly interesting from a data engineering point of view, as it can be performed using either stream or batch processing techniques. However, in the context of ANITA, a batch processing approach was preferred, because the crawler would be used for fresh data gathering at most every 24 h, thus limiting the eventual advantages of the stream processing. The trend analysis process is further supported by a visualization-as-a-service approach in charge of providing a customizable and interactive dashboard for the visualization of trafficking event patterns.

In general, the collected knowledge (i.e., the application domain expertise) renders feasible the realization of complex and highly demanding tasks, like criminal network construction, illegal shop tracking, and knowledge-based search and retrieval that are vital for analyzing different aspects of the illegal trafficking incidents. The data is elaborated by reasoning and knowledge mining services and tools for correlating high-level information, events, and facts and supports investi-



■ Fig. 5.2 Graphical representation of the ANITA framework and data-intensive pipeline. (“Author’s own figure”)

gators in understanding the dynamics of illegal trafficking activities and reconstructing criminal and terrorist groups involved.

Overall, the fundamental consideration of ANITA to inject human feedback in the analysis pipeline serves the following two fundamental goals: (a) to significantly boost the efficiency of the investigation process, by continuously improving the robustness of the feature detectors through the incorporation of the explicit and implicit user feedback, while also updating and expanding the knowledge infrastructure for the selected application domain and (b) to remarkably speed up the training process of new/novice crime investigators, practitioners, and officers for the application domain at hand, by reusing and transferring knowledge that has been collected and combined from multiple expert users (■ Fig. 5.2).

5.3 The PRoTECT Project for the Protection of Public Spaces

The EU International Security Fund (ISF) sponsors the PRoTECT project (2018–2021) to strengthen local authorities’ capabilities in safeguarding citizens in urban spaces. Five European cities, namely Eindhoven (Netherlands), Brasov (Romania), Vilnius (Lithuania), Malaga (Spain), and Larissa (Greece), cooperate to achieve the common goal of putting in place big data and artificial intelligence technologies to provide effective and direct responses to better assure safety in public places, e.g., by preventing and/or reacting to terrorist threats.

With this goal in mind, the main expected outcomes of PRoTECT are threefold:

- Reduced risk of terrorist attacks and costs to prevent or react to them
- Training materials related to the protection of public spaces for municipalities and LEAs
- A practical tool suite consisting of intelligent, data-driven tools to effectively support and improve the European cities' capabilities to counteract public threats

The first step in achieving these outcomes focuses on (a) the delivering of a comprehensive overview of the state of the art in the current information technologies based on a systematic literature review (SLR); (b) the definition of quantitative indicators for further risk assessment and research in underdeveloped technologies; (c) the assessment of different techniques, prevention measures, and methods for targeted information technologies; and (d) the elicitation of best practices adopted by governments and private organizations.

Here, we provide the results of our analysis applied on 112 documents from academics and practitioners over the last 15 years. The *encoded concepts* represents the five major categories found by our analysis. This group of concepts represents the main research topics from the analyzed studies. In the *clustered concept*, we have the main branches the research is focusing on.

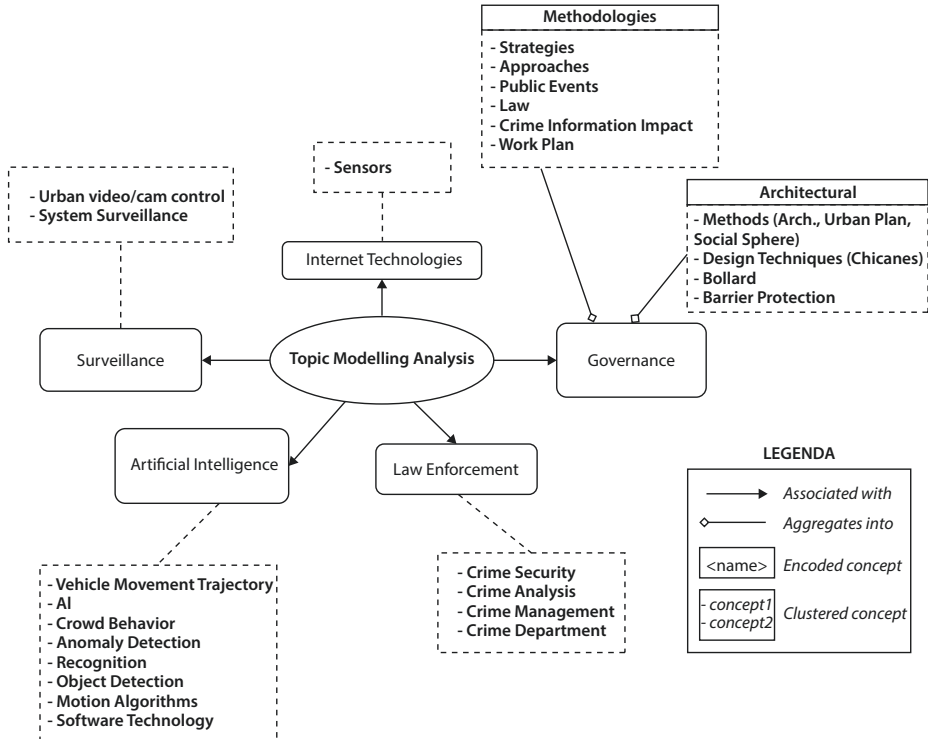
The second step consists of developing a European Technology Evaluation Framework (EU-TEF) comprising eight methodological steps to gather and evaluate technologies based on the vulnerabilities of the EU VAT. The steps go from prioritizing vulnerabilities, performing a request for information (RFI), and evaluating the technology results to the demonstration of relevant technologies in each city. The EU-TEF can be used to evaluate technologies, based on preset criteria and specific scenarios.

Using the European Technology Evaluation Framework, critical threats of each city involved have been identified, the risks to be mitigated, and the security measures that need to be enhanced. An RFI has been disseminated using various channels (project web site, social media, networking, written invitations to networks, industry, projects, etc.) to collect relevant solutions. These solutions have been selected based on their potential given the outcome of the vulnerability assessment (VA) within the project. Hence, we will have a “demo session” phase to test some selected solutions in the five cities. After the demo session, will be drafted the results and how the selected technologies have been able to mitigate the risks found and enhance the protection of the public spaces in the involved cities (■ Fig. 5.3).

5.3.1 Objectives of PRoTECT

The strategic objective of the project is to provide all members of the European Forum for Urban Security (EFUS)² actionable European municipalities an actionable perspective on the protection of their public spaces and other soft targets, by

2 ► <https://efus.eu/it/>.



■ Fig. 5.3 Topic modeling analysis taxonomy result. ("Author's own figure")

providing them with good practices and access to technology concepts and the knowledge to tailor them to their needs. The approach to achieve this is defined as follows:

1. To improve the protection of public spaces and other soft targets in five European cities and respective LEAs by providing them with direct support from the project to conduct both a vulnerability self-assessment and technology assessment in their municipalities
2. To let municipalities organize peer-to-peer exchange with other municipalities (EFUS members) with technology solutions and best practices in place, to disseminate their experiences with implementing good practices and technology concepts, including validation through tabletop exercises and a technology road map

At long term, we expect all EU local municipalities to be familiar with the vulnerability self-assessment and with a broad range of good practices and technology concepts and willing to improve the protection of their public spaces. The European Network for Law Enforcement Technology Services (ENLETS) network will be positioned to play a major role in providing technology advice via the organization of targeted workshops in close collaboration with the ENLETS national contact points (NCPs) and working groups.

Moreover, the effects of PRoTECT project can be summarized as follows:

5

- **To facilitate and reinforce** the exchange of practices on the topic of the protection of public spaces and soft urban targets in line with the EU action plan to improve the protection of public spaces
- **To contribute** to the reinforcement and creation of networks of stakeholders (EFUS and ENLETS), of local and regional authorities (PRoTECT workshops with the participation of local authorities from many European MS), as the project results will contribute to draw lessons from past attacks, develop guidance, and share innovative solutions to enhance the protection of public spaces
- **To develop** recommendations and concrete solutions, in particular regarding the link between LEAs and local and regional authority's partnerships in the field of protecting public spaces and soft targets, transferable to other local/regional authorities
- **To use** already existing results from European initiatives such as innovative solutions as part of EU research projects in the security domain and upcoming tools such as the EU Commission's guidance material (risk assessment tools and security by design solutions) in order to strengthen local authorities' capacities

The contribution of PRoTECT project is twofold and can be briefly summarized as follows. From one side, the aim of the PRoTECT project is to improve the protection of public spaces and other soft targets in five European cities and respective LEAs by providing them with direct support from the project to conduct both a vulnerability self-assessment and technology assessment in their municipalities. Meanwhile, on the other side, PRoTECT aims to let municipalities with technology solutions and best practices in place organize peer-to-peer exchange with other municipalities (EFUS members) to disseminate their experiences with implementing good practices and technology concepts.

5.3.2 PRoTECT and the Data Fusion Approach

Video surveillance has become a fundamental element in many activities to guarantee human security. It is omnipresent already in many urban (public and private) places such as banks, prisons, airports, parking lots, and petrol station. Video surveillance is also essential in multiple occasional events which attract large number of people, such as live festival, football games, and concert. Unfortunately, most visual surveillance heavily relies on a human judgment to analyze these videos (Bheechook et al., 2019). This is due to the fact that understanding uncommon behavior automatically from videos is a very challenging problem (Isupova et al., 2015; Remagnino et al., 2007).

Due to this, there exists an urgent need related to automating certain surveillance tasks to evaluate security and allow officers to develop their work in a more efficient way (Roberto Arroyo et al., 2015). Uncommon behavior—sometimes referred to as anomalous behavior—is one which deviates from all behaviors.

However, uncommon behavior can only be defined in a specific situation; therefore, it cannot be subjective but rather context sensitive (Harrigan et al., 2008).


Computer vision studies have been studying human behavior for a long time. For example, several studies have been conducted on recognizing and analyzing basic human actions such as hand movements, small gestures, and gait (Moeslund et al., 2006). However, more recently, sophisticated research has been carried out in order to obtain maturity and promising results for automatic detection of uncommon behavior gained from the footage of CCTV surveillance systems (Ibrahim, 2016).

Automated visual surveillance has been advanced to the third-generation surveillance system (3GSS) (Raty, 2010) that installs a high number of cameras in geographically diverse locations by a distributed manner for establishing a multi-modal camera network. The 3GSS has several capabilities: automatic event understanding for alerting operators, computer-based monitoring of a target, and tracking of targets among cameras (Wang, 2013).

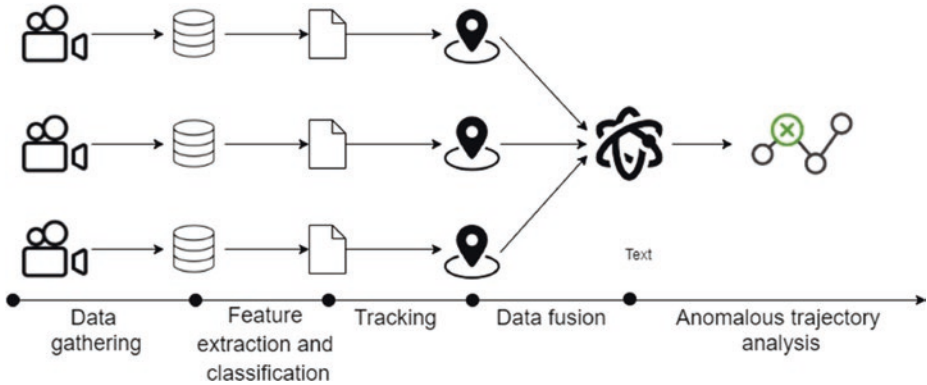
Uncommon behavior surveillance is important not only for detection but also for prevention in a lot of situations ranging from incidents such as thefts, watching movements of old people in homes, and avoiding the need of babysitting every time to spotting a potential terrorist in crowded public places like airports (Baig & Jabeen, 2016; Ibrahim, 2016). The reason why automatic detection is being given so much importance in surveillance system recently is because of the monotonous, tiring, and continuous focus required from the human brain (Bheechook et al., 2019).

The PROTECT techniques and tools are grounded on the assumption of data that is composed from multiple data sources, spread over various geographically dispersed locations, and combining IoT data and other data, such as open-source intelligence (OSINT).

The advantages of a data fusion approach instead of a single-source application are several: to better explain the benefits related to the tracking activity, let us make an example. In a railway station, there are different video surveillance cameras that monitor the activities. For the pickpocketing activity recognition, it is necessary to recognize the entire action. Analyzing a single camera, it is possible that the entire action is not captured and, for this reason, the prediction model built behind the video surveillance system may not be able to recognize the illicit action. This problem can be fixed with a data fusion approach, because all this information is merged in the fusion step in order to rebuild the single action from different perspectives.

The pipeline in  Fig. 5.4 shows the steps for the data fusion process. At first, in the data gathering process, all the data processed by the devices are collected. In this process, for each video, the information about the recording time and the type of the video is saved (e.g., normal video, infrared video).

Then, in the feature selection, the data gathered are processed, in order to extract the most important features for the tracking process. Each video is split into 30 frames, and information about the coordinates of each object recognized is stored. A practical implementation of this process is provided by TensorFlow (Abadi et al., 2016), an AI/machine learning platform that can be used for the



■ **Fig. 5.4** Processing steps. In this chart is shown the flow of data through the nodes and the procedures involved. (“Author’s own figure”)

object detection process as well. Once the system recognizes a person, it saves this information as a set of object coordinates.

During the tracking step, all the object coordinates gathered are assembled and cross-correlated in order to re-create their object path.

Until now, all the sensors/devices have worked separately. The difference from a naive approach and a data fusion approach starts at this point. Instead of analyzing the path collected separately, they are combined together in order to generate a single path containing the information of all paths.

Data fusion can be done in different ways (e.g., classification based on the relations between the data sources, classification based on the abstraction levels) (Castanedo, 2013). The PRoTECT data fusion proposes an unsupervised approach adopting dimensionality reduction technique (UMAP) and the clustering algorithm K-means to extract common pattern in the data that is to be fused.

These algorithms automatically cluster trajectories that have common statistic behavior and cluster pattern with uncommon behavior. In particular, dimensionality reduction techniques may be deployed to bring back the number of features (aka input variables) in a dataset (like video footage) without too much loss of information. K-means clustering refers to one of today’s most basic unsupervised ML techniques, which groups similar data points together in a number of k cohesive clusters. The data points are grouped based on the central point in the cluster space (called the centroid), keeping their number as low as possible, and data points as close as possible to the centroid in each cluster. So, the data fusion process aims to discover common paths of an object in order to correlate the path tracked on a source with another path tracked on another source. With this information, it is possible to associate these paths with the same object and rebuild the scene recorded from different points of view.

Results show that the tracking can be achieved deploying unscented Kalman filters (UKF).

The results from the UKF are first compared to a geometrical approach called baseline. Eventually, two methods to validate the UKF approach and the baseline

approach are presented: a visual validation at a specific time during the experiment (Snidaro et al., 2004) and validation adding white noise to the measurement (Wan & Van Der Merwe, 2000). Results show that the UKF outperforms the baseline approach. Blending data sources, i.e., smartphone S10 on the right position and smart glasses Vuzix on the left position, indeed demonstrates no significant loss in performance.

The unsupervised anomaly detection technique clusters instances of uncommon behavior of a particular trajectory. A real-world case study has been used to validate the anomaly detection technique, namely, a video of a live festival in the Netherlands. The case renders footage of festival goers leaving the premises. Their trajectory is generated using fuzzy fixed points.

An unsupervised detector is deployed to recognize and cluster common and uncommon behavior. This approach correctly separates behavior accordingly.

Experimental results of the data fusion from the PRoTECT approach showed the great accuracy achievable by the proposed approach in comparison with simple geometrical data fusion systems. Two validations have been performed for testing the reliability and precision of the trajectory.

The fusion approach has produced trajectories that are more reliable and useful for a surveillance system. More detailed information on PRoTECT can be found here: ► <https://protect-cities.eu/>.

5.4 The Beehives Project for the Quality of Urban Biodiversity

Over the last few centuries, traditional science has made startling progress resulting in elegant “laws of nature” and rule-based paradigms or mechanisms to better understand the world around us.

However, in practice, these laws are hard to apply and often fall short for wicked societal and environmental problems, such as managing a pandemic, guaranteeing safety in cities, or counteracting the detrimental effects of climate change (Sinha et al., 2020).

Meanwhile, recent developments and practical applications of big data clearly demonstrate that the predictive power of data may be surprisingly good for these kinds of actual quandaries.

Indeed, data analytics in combination with big data are a valid alternative for dilemmas that are difficult to solve and concepts that are hard to grasp. Algorithms that mine, unravel, and expose patterns in large and varied datasets allow us to get a grip on complex natural phenomena. Whether we like it or not, likes on social media platforms are strong indications of political preference or other personality traits. We all probably have heard about the notorious and somewhat embarrassing example of shopping behavior, which appears to be linked to pregnancy status (Oviatt & Reich, 2019). Besides tracking plain whereabouts, wearables or smartphones are able to retrieve other data about us well, without us being even aware of it, including data about sleep quality, stress level, sport performance, driver behavior, or even road quality.

So why not exploit a big data approach to help us with the biodiversity crisis which is unfolding at a fast pace? For this case, we outline a possible solution infused with (IoT) instruments—intelligent beehives—that gather unprecedented amounts of data which provide the basis for tantalizing new insights in the quality of life in the urban and rural areas in which we live our daily lives.

5.4.1 Problem Description

5

Nature is the sine qua non for human survival and quality of life. Ecosystems are shaped by the constellations of organisms (biotic components) in a physical, non-living environment (a system of non-biotic components), providing a number of critical processes, such as photosynthesis and carbon cycles, across the world.

Nevertheless, urbanization, one of the most important land-use and land-cover changes, has over time incurred a multitude of deep and pervasive effects on biodiversity, human health, and well-being (Guetté et al., 2017; Reeves et al., 2019).

The concept of biodiversity is about the variety and variability of organisms, incorporating intraspecific, interspecific, and ecosystem diversity, in relation to time (evolution) and space (geographical distribution) (National Research Council, 1999).

It is generally accepted that biodiversity is rapidly declining across the world (Hallmann et al., 2017). The demographic growth and the intensive exploitation of natural resources accelerate not only the degradation of natural environments but also the extinction of species. Indeed, urbanization entails a major cause of insect decline including the bee colonies; the severity of the event has consequences on the pollination process (Jones et al., 2013).

Governments and societal organizations of the entire world are developing environmental protection guidelines and policies to halt or reverse this trend (Sand, 1992).

In short, nature is under severe threat. The recent UN report from the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services warns that nature is declining globally at rates unprecedented in human history (Brondizio et al., 2019). This can be witnessed by the rapid decline of both biodiversity and biomass, as has been substantiated for instance in the famous Krefeld study about insect free fall (Hallmann et al., 2017). Human activities, such as reuse of natural land surface for housing, agriculture, or economic activity; use of pesticides and herbicides; pollution of soil, water, and air; and unsustainable harvesting of natural resources, are the primary causes of this decline (Stuart Chapin III et al., 2000). As a result, human existence and quality of life are increasingly threatened (Chiesura, 2004). Underlying these direct factors are deeper causes relating to the contemporary human perception of and relationship with nature. The implicit assumption that nature is there to be exploited to our benefit combined with a fading of the notion of how our consumption, safety, and comfort are entwined with the natural world forms the basis of a value system supporting the unsustainable practices mentioned (Aeon Co, n.d.).

In this case study, we aim at changing this deeply rooted social attitude and detrimental culture of exploiting the ecosystems in which we live for our own benefit, without any regard of the surrounding and global environment.

Cities play a central role in the problem of declining nature, but also in potential solutions with the ecological function of green spaces (Lepczyk et al., 2017). On the one hand, the rapid urbanization has negative effects on nature. Globally, only 220 million people (13%) lived in urban areas in 1900; this increased to 3.2 billion (49%) by 2005 and is projected to reach 4.9 billion (60%) by 2030 (Maksimovic et al., 2015). However, it is estimated that by 2050, two-thirds of the global human population will live in urban areas (Reeves et al., 2019). This continued, rapid growth of cities often heavily draws on natural resources, consumes and fragments biodiversity hotspots, increases pollution, and may stimulate efficient but non-sustainable production methods in agriculture and industry.

On the other hand, cities can be part of the solution. Rich ecosystems and biodiversity, including endangered and threatened species, can exist in cities and provide ecosystem services within and across urban boundaries (Aronson et al., 2017; Baldock et al., 2015). In recent years, several initiatives have been taken to make cities greener and improve urban biodiversity. In this regard, it is necessary to obtain contextual data and build models for urban biodiversity to plan and execute policies and interventions, assess their effectiveness, and make the process more dynamic (Nilon et al., 2017).

5.4.2 Objectives


This case study aims to improve the quality of the living environment by increasing urban biodiversity.

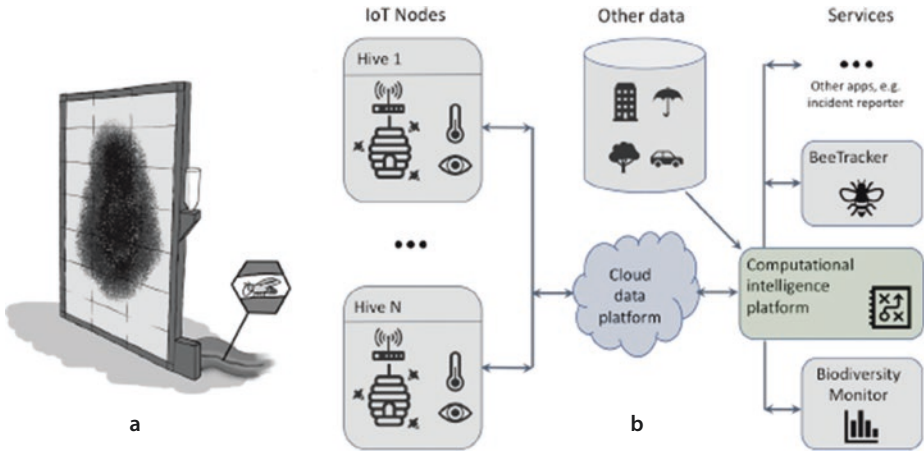
The project has two lines of research. One focuses on technology: creating an innovative system for monitoring urban biodiversity using sensor-equipped beehives combined with AI and big data technology.

The other focuses on systemic change: setting a transition process in motion that will enable two stakeholder sets—school children and urban planners—to adopt new behaviors that stimulate urban biodiversity.

The ultimate goal is that we can design our future cities taking into account biodiversity to maintain the quality of our natural habitat. As a first step, we should be able to measure biodiversity (or part of it) in a holistic way (1) with high-tech systems (such as beehives equipped with sensors) that tap into our natural environment and (2) by combining this with relevant open environmental data (such as weather, water and air quality, databases with infrastructure, buildings, and public urban green).

5.4.3 Data Gathering

In this case study, high-tech equipment and advanced technologies (IoT grid with various sensors and fast 5G communication protocol) are used to collect data from beehives and the surrounding environment. We foresee a data collection and storage architecture as conceptualized in  Fig. 5.5.



■ **Fig. 5.5** (a) Schematic impression of the intelligent beehive. (b) Conceptual architecture with data collection at IoT nodes, use of other (open) data, data storage, and computational services. (“Author’s own figure”)

The application of big data in biodiversity is a valid approach for building predictive models, providing unexplored solutions, creating breakthrough improvements in biodiversity monitoring, and comparing quickly and efficiently huge amounts of data with significantly lower cost (Klein et al., 2015).

Honey bees are excellent bioindicators; they are highly sensitive to their natural environment. Hence, it is interesting to mine data about the health of the bee colony (such as weight, brood size, empty holes in the comb, honey store). They visit flowers and collect nectar and pollen in an area of approximately 3 km around the hive. The bees bring interesting information to the hive about the nature and biodiversity of the surrounding area (e.g., through pollen and nectar). Another source of valuable data that can be used for estimating the quality of the natural environment around the hive. In addition, bees tell each other about the landscape and the available forages through a waggle dance, an effective form of animal communication that offers advantages for foraging success (Nürnberg et al., 2017). This interesting and amazing behavior will be detected by cameras and image analysis in the beehive. More specifically, the dance encodes the direction, distance, and quality of a food resource in the field (Nürnberg et al., 2019). Also, the project aims to collect data about pollution and contamination (e.g., toxic metals, PAHs) which are highly present in urban areas and evaluate-estimate their effects with predictive models (Jovetić et al., 2018). The hypothesis that we will test at large is the following: We can estimate the biodiversity of the surrounding area using the data assembled and decoded by the intelligent beehive, in conjunction with mining other open data sources (e.g., regarding traffic, pollution, databases with infrastructure, buildings, and public urban green).

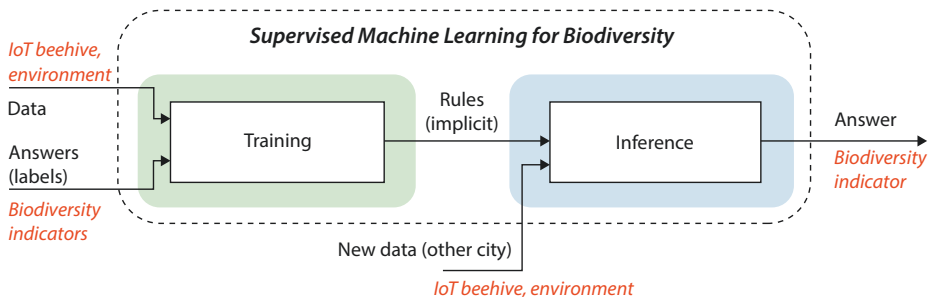
In order to test the hypothesis that we can estimate the biodiversity of the surrounding area using data assembled by an Internet of Things (IoT) beehive, we need to collect two types of data: (1) ground-truth measurements of biodiversity

collected through traditional methods (counts of plant and animal species) and (2) beehive data collected by ten intelligent hives over at least one season, enriched with other environmental data from public databases.

5.4.4 Big Data Analytics for Biodiversity

In general, the purpose of big data analytics (a.k.a. AI or machine learning) is to predict nontrivial, emergent, and nonlinear phenomena that are difficult to gauge. These models demand huge amounts of data to do so. The most common approach is to apply supervised learning, i.e., learning a function that maps an input to an output based on example input-output pairs. In this case, we try to infer the biodiversity of an urban area on the basis of two sets of input data: IoT-instrumented beehive data (e.g., processed waggle dances, pollen, brood size, honey store, weight) and open environmental data (e.g., weather, water and air quality, databases with infrastructure, buildings, and public urban green). Before the predictive model can be used, it needs to be trained. For the training, we use a third dataset: ground-truth biodiversity measurements. Collecting the ground-truth biodiversity is a laborious process. We plan to do this in various student projects. This manual counting should be supported by a sound protocol and the usage of species identification apps. The training of the model follows a common script: (1) collect the data and annotate the records with labels; (2) split the data in a training set, validation set, and test set; (3) train the model on the training set and tune the model on the validation set; and (4) evaluate its accuracy on the holdout test set. The training, tuning, and evaluation steps are depicted in **■** Fig. 5.6, the left-pointing red arrow corresponding to the training and tuning phases and the right-pointing one to the evaluation phase.

The labels used to annotate the ground-truth biodiversity measurements determine the output of the final model. If we train the model on the raw species counts, the final system deployed in a new environment will also produce estimated species counts. Although we plan to experimentally run this scenario during the project, we expect the reliability of the resulting estimates to strongly vary across species groups. For example, we expect insect and flower biodiversity to be estimated more accurately with IoT beehives than, say, amphibians, because bees, being insects,



■ Fig. 5.6 Engineering a supervised machine learning model. (“Author’s own figure”)

collect information that is more relevant to insects than to amphibians. We will therefore, prior to training, recode the ground-truth data into aggregate biodiversity indicators that integrate across species groups. These indicators are determined by the projection of ground-truth species surveys on an ecological quality system (Dale & Beyeler, 2001).

Once the ecological quality system is established and standardized and the model is trained, we can measure biodiversity in new urban environments. The model will predict the biodiversity indicators for new input data (IoT beehive data and open environmental data). Well-established modeling practices such as K-fold validation, dimensionality reduction, and normalization are used to prevent overfitting and to ensure that you end up with a model that has good generalization capabilities for new urban areas.

5

5.4.5 Systemic Change

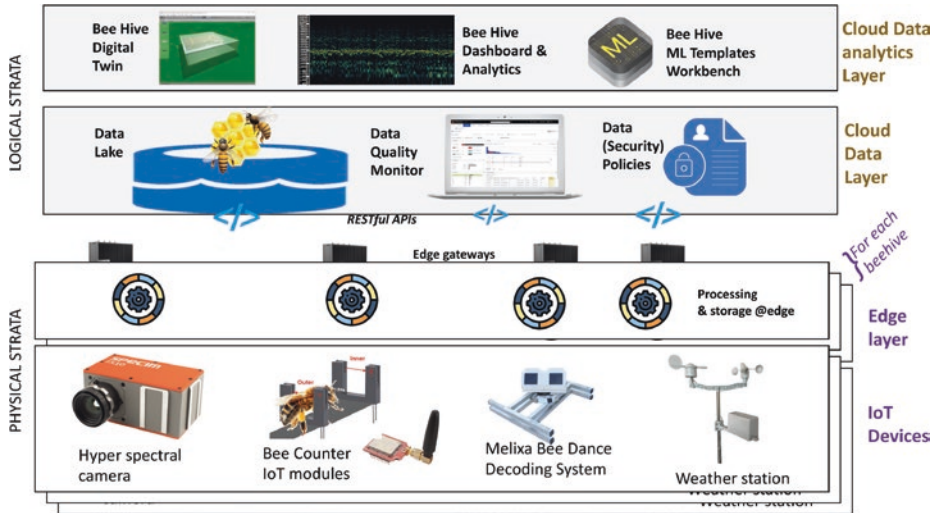
The systemic change approach is based on co-design and is developed with identified stakeholder groups in materials and products that will educate and activate. These sessions have an iterative nature, repeatedly going through design-thinking cycles including frame, ideate, build, and discover phases. We specifically target two groups for our systemic change approach: (1) primary school children, who are the future adult citizens, and (2) urban planners. The multidisciplinary co-design sessions aim at producing the following solutions: A cross-curricular educational program that creates awareness and engagement with the surrounding nature. The program will be based on self-directed, problem-based learning and will improve the children's twenty-first-century skills such as problem-solving ability, creative thinking, engaged citizenship, and digital literacy. The materials and products that will be developed include tools supporting teachers and pupils; apps and games; and building materials for making simplified IoT hives or insect hotels.

5.4.6 Bringing It All Together: The IoT Beehive Stratified Architecture

The Smart Beehive project embraces a stratified data architecture, with well-defined layers that are separated in terms of the type of data and logic they combine, but built on top of each other. Each layer is self-contained and addresses a specific concern, such as data collection and storage, or data analytics.

The stratified architecture assumes several logical and physical layers that abstract away from IoT devices and edge gateways at the physical layers to the cloud data and analytics in the logical (software) layers.

In this way, the architecture gracefully allows for continuous integration where changes in one layer have minimal ramifications for the other layers, lowering maintenance efforts and encouraging continuous rather than incidental changes and improvements (such as extensions).



■ Fig. 5.7 The IoT beehive stratified data architecture (Sangiovanni et al., 2020)

We have graphically depicted the IoT beehive stratified data architecture in

■ Fig. 5.7.

In the following, we will now further explain our architecture starting from the physical layers and working up in the stack to the logical (software) layer.

In particular, IoT beehive stratified architecture encompasses the following physical layers:

- **IoT Beehive Device Layer:** This layer consists of IoT-enabled beehives. IoT devices can monitor important parameters such as the number of bees entering and leaving the beehive, the type of pollen collected by bees entering the beehive, and weather conditions.
- **IoT Beehive Edge Layer:** To process the data generated by IoT devices efficiently, edge devices can be used. The edge layer allows data processing at or near the source of data generation, enabling real-time status monitoring at each beehive.

The following logical layers sit on top of the physical data architecture stratosphere:

- **Cloud Data Layer:** The cloud data lake receives and stores all data generated by the IoT devices and edge devices. It can also collect the data from open sources (e.g., social media and weather APIs). The quality of the collected data needs to be assessed and enforced. To enable secure and timely access to the data, the data governance service should be implemented.
- **Cloud Analytics Layer:** This layer enables turning the collected beehive data into value through machine learning and artificial intelligence. It also supports building beehive digital twins that constitute a digitally designed virtual system to support efficient monitoring and management of the beehive.

Take-Home Messages

The reader can take the following key points from this chapter:

- Data engineering principles and practices can be successfully applied to real-world big data use cases.
- Depending on the function and nonfunctional requirements of the target applications, the practitioners need to use different data architectures, data processing, and analytic techniques and tools.

5

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). TensorFlow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (pp. 265–283).
- Aeon Co. How protestantism influenced the making of modern science. Retrieved from <https://aeon.co/essays/how-protestantism-influenced-the-making-of-modern-science>
- Aronson, M. F. J., Lepczyk, C. A., Evans, K. L., Goddard, M. A., Lerman, S. B., Scott MacIvor, J., Nilon, C. H., & Vargo, T. (2017). Biodiversity in the city: Key challenges for urban green space management. *Frontiers in Ecology and the Environment*, *15*(4), 189–196.
- Baig, A. R., & Jabeen, H. (2016). Big data analytics for behavior monitoring of students. *Procedia Computer Science*, *82*, 43–48.
- Baldock, K. C. R., Goddard, M. A., Hicks, D. M., Kunin, W. E., Mitschunas, N., Osgathorpe, L. M., Potts, S. G., Robertson, K. M., Scott, A. V., Stone, G. N., et al. (2015). Where is the UK's pollinator biodiversity? The importance of urban areas for flower-visiting insects. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1803), 20142849.
- Bheechook, L., Baichoo, S., & Khan, M. H.-M. (2019). The need for automatic detection of uncommon behaviour in surveillance systems: A short review. In S. C. Satapathy et al. (Eds.), *Information systems design and intelligent applications* (pp. 411–419). Springer.
- Bronzizio, E. S., Settele, J., Díaz, S., & Ngo, H. T. (2019). *Global assessment report on biodiversity and ecosystem services*. Global assessment report. United Nations Organisation.
- Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal*, *2013*, 704504.
- Chiesura, A. (2004). The role of urban parks for the sustainable city. *Landscape and Urban Planning*, *68*(1), 129–138.
- Dale, V. H., & Beyeler, S. C. (2001). Challenges in the development and use of ecological indicators. *Ecological Indicators*, *1*(1), 3–10.
- Guetté, A., Gauzcre, P., Devictor, V., Jiguet, F., & Godet, L. (2017). Measuring the synanthropy of species and communities to monitor the effects of urbanization on biodiversity. *Ecological Indicators*, *79*, 139–154.
- Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., Stenmans, W., Muller, A., Sumser, H., Horren, T., et al. (2017). More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS One*, *12*(10), e0185809.
- Harrigan, J., Rosenthal, R., Scherer, K. R., & Scherer, K. (2008). *New handbook of methods in non-verbal behavior research*. Oxford University Press.
- Ibrahim, S. W. (2016). A comprehensive review on intelligent surveillance systems. *Communications in Science and Technology*, *1*(1).
- Isupova, O., Mihaylova, L., Kuzin, D., Markarian, G., & Septier, F. (2015). An expectation maximisation algorithm for behaviour analysis in video. In *2015 18th International Conference on Information Fusion (Fusion)* (pp. 126–133). IEEE.
- Jones, E. L., Leather, S. R., et al. (2013). Invertebrates in urban areas: a review. *European Journal of Entomology*, *109*(4), 463–478.

- Jovetić, M. S., Redžepović, A. S., Nedić, N. M., Vojt, D., Đurđić, S. Z., Brčeski, I. D., & Milojković-Opsenica, D. M. (2018). Urban honey the aspects of its safety. *Arhiv za Higijenu Rada i Toksikologiju*, 69(3), 264–274.
- Klein, D. J., McKown, M. W., & Tershy, B. R. (2015). Deep learning for large scale biodiversity monitoring. In *Bloomberg Data for Good Exchange Conference*.
- Lepczyk, C. A., Aronson, M. F. J., Evans, K. L., Goddard, M. A., Lerman, S. B., & MacIvor, J. S. (2017). Biodiversity in the city: Fundamental questions for understanding the ecology of urban green spaces for biodiversity conservation. *BioScience*, 67(9), 799–807.
- Maksimovic, Č., Kurian, M., & Ardakanian, R. (2015). *Rethinking infrastructure design for multi-use water services*. Springer.
- Moeslund, T. B., Hilton, A., & Kruger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2–3), 90–126.
- National Research Council, et al. (1999). *Perspectives on biodiversity: Valuing its role in an everchanging world*. National Academies Press.
- Nilon, C. H., Aronson, M. F. J., Cilliers, S. S., Dobbs, C., Frazee, L. J., Goddard, M. A., O'Neill, K. M., Roberts, D., Stander, E. K., Werner, P., et al. (2017). Planning for the future of urban biodiversity: A global review of city-scale initiatives. *BioScience*, 67(4), 332–342.
- Nürnbergger, F., Steffan-Dewenter, I., & Härtel, S. (2017). Combined effects of waggle dance communication and landscape heterogeneity on nectar and pollen uptake in honey bee colonies. *PeerJ*, 5, e3441.
- Nürnbergger, F., Keller, A., Härtel, S., & Steffan-Dewenter, I. (2019). Honey bee waggle dance communication increases diversity of pollen diets in intensively managed agricultural landscapes. *Molecular Ecology*, 28(15), 3602–3611.
- Oviatt, J. R., & Reich, S. M. (2019). Pregnancy posting: Exploring characteristics of social media posts around pregnancy and user engagement. *mHealth*, 5.
- Raty, T. D. (2010). Survey on contemporary remote surveillance systems for public safety. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(5), 493–515.
- Reeves, J. P., Knight, A. T., Strong, E. A., Heng, V., Cromie, R. L., & Vercammen, A. (2019). The application of wearable technology to quantify health and wellbeing co-benefits from urban wetlands. *Frontiers in Psychology*, 10, 1840.
- Remagnino, P., Velastin, S. A., Foresti, G. L., & Trivedi, M. (2007). Novel concepts and challenges for the next generation of video surveillance systems.
- Roberto Arroyo, J., Yebes, J., Bergasa, L. M., Daza, I. G., & Almazán, J. (2015). Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls. *Expert Systems with Applications*, 42(21), 7991–8005.
- Sand, P. H. (1992). Unced and the development of international environmental law. *Journal of Natural Resources and Environmental Law*, 8, 209.
- Sangiovanni, M., Schouten, G., & van den Heuvel, W.-J. (2020). An IOT beehive network for monitoring urban biodiversity: Vision, method, and architecture. In S. Dustdar (Ed.), *Service-Oriented Computing* (pp. 33–42). Springer International Publishing.
- Sinha, A., Sengupta, T., & Alvarado, R. (2020). Interplay between technological innovation and environmental quality: Formulating the SDG policies for next 11 economies. *Journal of Cleaner Production*, 242, 118549.
- Snidaro, L., Foresti, G. L., Niu, R., & Varshney, P. K. (2004). Sensor fusion for video surveillance.
- Stuart Chapin, F., III, Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., Hooper, D. U., Lavorel, S., Sala, O. E., Hobbie, S. E., et al. (2000). Consequences of changing biodiversity. *Nature*, 405(6783), 234–242.
- Wan, E. A., & Van Der Merwe, R. (2000). The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)* (pp. 153–158). IEEE.
- Wang, X. (2013). Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34(1), 3–19.

Data Analytics

Florian Böing-Messing

In the previous module, we have presented and discussed various methods, tools, and concepts with respect to *data engineering* aspects of data-intensive applications. As was explained, data engineering activities (e.g., developing data pipelines, data quality assessment, data storage) prepare the grounds for doing *data analytics*. Data analytics has been exploited over the past decades by organizations to acquire more faithful business insights, for example, through OLAP Business Intelligence (BI) toolkits. This type of data analytics is also known as *descriptive (data) analytics*. Essentially, descriptive analytics utilizes statistical methods on historical data (e.g., from application and execution logs) in order to reveal patterns and assign meaning.

Over the past decade, a new generation of data analytics, sometimes referred to as *advanced analytics*,¹ has emerged with the capability to make predictions and recommendations even in (close to) real-time big data environments. *Predictive analytics* has taken descriptive analytics one step further and processes historical data for analysis of potential future scenarios, such as trends in consumer behavior. In a typical case of predictive analytics, historical transaction data is fed into machine learning or data mining models to make forecasts about specific phenomena under the likelihood of actual occurrence. More classical approaches for predictive analytics include the application of time series and regression models, for example, logistic regression, which allocates probabilities to observations for classification.

The course of action that is needed to attain particular forecasted scenarios may be analyzed

1 ▶ <https://www.gartner.com/en/information-technology/glossary/advanced-analytics>.

with another branch of advanced data analytics, labeled *prescriptive analytics*. Prescriptive analytics is generally perceived as the highest level of analytics as it builds on top of the outcomes of predictive analytics to suggest actionable decision options for seizing future opportunities (e.g., optimized operational efficiency in patient care) and/or avoiding risks (e.g., problematic medical interventions). Like descriptive and predictive analytics, prescriptive analytics can be realized with a cocktail of computational and mathematical/statistical models, including signal processing, recommendation engines, and neural networks (including deep learning).

In the *Data Analytics* module, we introduce, explain, and explore various (advanced) descriptive, predictive, and prescriptive analytics approaches. The first chapter in the module (i.e., ► Chap. 6) is entitled *Supervised Machine Learning in a Nutshell*. The authors present the fundamental principles and workings of regression and classification using a supervised machine learning approach, which assumes a preexisting, labeled set of input-output pairs (e.g., dog pictures as inputs and their breed as labeled outputs). Note that, next to supervised machine learning, there is also an area called *unsupervised* machine learning, which deals with learning patterns in *unlabeled* data. An example of a technique from this area is cluster analysis, which attempts to group objects (e.g., shopping items) based on their similarity. Unsupervised machine learning is not covered in detail in this module. For an accessible and hands-on introduction to this topic, we refer the interested reader to Géron (2019).

► Chapter 7 provides *An Intuitive Introduction to Deep Learning*. Deep learning is a generic header denoting deep multilayer neural networks that are typically used for prescriptive analytics. In particular, the authors provide a fundamental understanding of the internal structure of neural networks, including some intuitive formal underpinnings that revolve around the concept of a perceptron. Special emphasis is placed on a particular type of deep neural networks called convolutional neural networks (CNNs), which have attracted

much attention over the past years. One of the main applications of CNNs is in the areas of image recognition and computer vision. The authors discuss an example where CNNs are used in a medical image analysis context to diagnose skin cancer in images of skin lesions.

► Chapter 8, entitled *Sequential Experimentation and Learning*, discusses the contextual Multi-Armed Bandit (cMAB) problem. The cMAB problem falls under the umbrella of reinforcement learning, which is—next to supervised and unsupervised learning—a third fundamental machine learning paradigm. Simply speaking, reinforcement learning is about agents learning to reach the best result (e.g., reward) with a trial-and-error approach in contexts which are not completely known, where each error is a penalty and each desirable result is a reward. The cMAB problem is a special case of reinforcement learning. Its name derives from the situation where a gambler repeatedly plays a slot machine with multiple arms (i.e., a multi-armed bandit) without knowing which arm is best (in the sense of providing the highest monetary reward in the long run). This situation illustrates the crucial dilemma that the gambler faces when deciding which arm to play next: Should he or she exploit the knowledge about the arms from earlier tries by playing the arm that is the best one so far? Or would it be better to explore other arms in the hope of finding an even better arm? This dilemma is also referred to as the exploration-exploitation trade-off. The authors discuss different strategies—also called policies—for the gambler to select arms with the goal of maximizing the cumulative reward over the course of multiple interactions with the slot machine. A lot of real-world sequential decision-making situations can be modeled as a cMAB problem, which makes policies for selecting actions in such situations powerful methods in practice. To facilitate the use of bandit algorithms, the chapter presents two software tools that can assist data scientists and researchers in their sequential learning and experimentation applications.

► Chapter 9 discusses *Advanced Analytics on Complex Industrial Data*. The authors provide an

overview of several popular approaches for dealing with streaming and complex industrial data, including multivariate time series, application log data, and multimodal sensor data. In particular, three advanced analytics approaches are introduced and explored: graph pattern mining, graph signal processing, and analytics for fault diagnosis. A broad understanding of these approaches, and the way in which they are related, is created through several real-world examples drawn from the smart industry domain—popularly referred to as Industry 5.0—with an emphasis on sensor data that may be exploited for the purpose of preventive maintenance.

The final chapter in this module (i.e., ► Chap. 10) is entitled *Data Analytics in Action*. The chapter presents three real-life case studies, in which the above arsenal of data analytics—ranging from descriptive to predictive and prescriptive techniques—is further applied and exemplified, and practical implications are considered. In the first case study, entitled *BagsID: AI-powered Software System to Re-identify Baggage*, the authors present an end-to-end case from the inception of a business idea to an actual AI-powered software system. This BagsID AI solution embraces CNNs as the backbone for reidentifying mishandled luggage at airports. The second case study in the chapter is entitled *Understanding Employee Communication with Longitudinal Social Network Analysis of Email Flows*. The case study revolves around a commercial company that implemented measures for promoting innovation among its employees. The authors examine the effect of these measures by analyzing email communication about innovation between employees using a social network analysis approach. In the final case study, entitled *Using Vehicle Sensor Data for Pay-How-You-Drive Insurance*, the author considers how state-of-the-art technologies in cars (e.g., the so-called controller area networks) can be leveraged for insurance purposes. These technologies generate large amounts of data while the car is on the road. The resulting time series can be analyzed to understand the driving behavior of individual drivers. The case study emphasizes practical considerations and implications of dealing with streaming data that is generated while driving.

References

- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). Sebastopol, CA: O'Reilly Media.

Contents

- Chapter 6 Supervised Machine Learning in a Nutshell – 105**
Majid Mohammadi and Dario Di Nucci
- Chapter 7 An Intuitive Introduction to Deep Learning – 121**
Eric Postma and Gerard Schouten
- Chapter 8 Sequential Experimentation and Learning – 147**
Jules Kruijswijk, Robin van Emden, and Maurits Kaptein
- Chapter 9 Advanced Analytics on Complex Industrial Data – 177**
Jurgen O. D. van den Hoogen, Stefan D. Bloemheuvel, and Martin Atzmueller

Chapter 10 Data Analytics in

Action – 205

*Gerard Schouten, Giuseppe
Arena, Frederique C. A. van
Leeuwen, Petra Heck, Joris
Mulder, Rick Aalbers, Roger
Th. A. J. Leenders, and
Florian Böing-Messing*



Supervised Machine Learning in a Nutshell

Majid Mohammadi and Dario Di Nucci

Contents

- 6.1 Introduction – 106**
- 6.2 Supervised Learning: Classification – 107**
 - 6.2.1 Motivating Example: Credit Card Fraud Detection – 107
 - 6.2.2 An Overview of Classifiers – 108
 - 6.2.3 Evaluating a Classification Model – 110
 - 6.2.4 Designing a Pipeline for Machine Learning Classification – 112
- 6.3 Supervised Learning: Regression – 114**
 - 6.3.1 Simple Linear Regression – 114
 - 6.3.2 Regression Methods: An Overview – 116
 - 6.3.3 Evaluating a Regression Model – 117
 - 6.3.4 Designing a Pipeline for Machine Learning Regression – 118
- References – 119**

Learning Objectives

- Understand the basic concepts concerning supervised machine learning.
- Explain the differences between classification and regression.
- Describe the core principles for validating and evaluating a machine learning model.
- Explain the main components of a machine learning pipeline for classification and regression.

6.1 Introduction

6

In 1959, Arthur Samuel described machine learning as “The field of study that gives computers the ability to learn without being explicitly programmed” (Samuel, 1959). More recently, Tom Mitchell provided the following definition: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E ” (Mitchell, 1999). In other words, machine learning (ML) is the subset of artificial intelligence devoted to defining computer algorithms that automatically improve through experience. Machine learning algorithms create mathematical models based on previous observations to make predictions “without being explicitly programmed” (Samuel, 1959) to do it. In the last decades, machine learning algorithms have been applied to many fields.

In this chapter, we provide two simplified examples of (1) credit card fraud detection and (2) stock price prediction. We use such examples to introduce a class of machine learning techniques, namely “supervised learning,” which comprises the algorithms fed with labeled data (i.e., data with a tag, or a type, or a number) to create models that make predictions over given previously unseen data.

In principle, the aim of supervised learning is to approximate a mapping function f from a set of input variables X to obtain the output variable y . Supervised learning can be leveraged to solve both classification and regression problems, the former of which includes the prediction of categories given novel observations, while the latter entails the prediction of a number (e.g., stock price) for a given observation. As a result, the output variable y for classification is discrete, while it is continuous for regression. For both classification and regression, we outline a practical example. Then we provide an overview of the essential techniques used to solve the problems. The models created by using such techniques strongly depend on the data used to feed them (i.e., training set). Therefore, it is essential to evaluate model performance correctly. Finally, we provide an overview to help the reader understand how to design a basic machine learning pipeline.

6.2 Supervised Learning: Classification

Classification methods lie within the supervised learning where the label of samples is categorical. Hence, labels need to be inspected to use a classifier for the data at hand. A categorical variable is a variable that takes on a value from a finite (and usually fixed) number of possible values. An important and distinct feature of categorical variables is that its possible values are not mathematically comparable to each other. For example, if there are two categories c_1 and c_2 , one cannot say if $c_1 > c_2$ or the other way around. Therefore, the categories are solely the representative of some classes, and we can only understand that the samples of one category are different from those of the others. A categorical variable with two possible values is called binary (or dichotomous), and the corresponding supervised learning methods are called binary classifications. By the same token, if a variable can take on more than two values, then it is called polytomous, and the associated supervised learning algorithm is called multi-class classifications.

In this section, we focus on the binary classification with an example in detecting fraud in credit card transactions. We first explain the types of transactions in credit cards and illustrate the applicability of classifiers by a simplified two-dimensional example. An overview of the most popular classifiers is then presented, followed by a discussion concerning the proper ways to evaluate a classifier given a dataset and to create a machine learning pipeline for data classification.

6.2.1 Motivating Example: Credit Card Fraud Detection




► Example


As the number of cashless transactions continues to grow, the number of fraudulent transactions also increases, making the detection of such operations of utmost importance for the credit card companies. Two primary issues that compound fraud detection are the massive number of transactions conducted every moment, as well as a high similarity between fraudulent and regular operations. Thus, automatically detecting fraudulent transactions requires proper computational tools.

The credit card transactions are either fraudulent or normal. Hence, they can be divided into two mutually exclusive sets. Since the label of data is dichotomous, the problem of detecting fraud in credit card transactions can be transformed into a supervised binary classification problem. Before applying any classifier, it is required to accumulate a sufficient number of transactions, each labeled as regular or fraudulent. Having such transactions is the primary prerequisite of using classification methods for detecting frauds.

Another important step is to extract features (or attributes) from the transactions. A feature describes a characteristic in the data (here, credit card transactions) that can help discriminate fraudulent from non-fraudulent transactions. Features allow a computer

to learn and understand the dynamics of the transactions and to discern and detect fraud. Features are inspired by the way fraudsters conduct abnormal transactions. For example, fraudsters try to conduct many transactions in a short period before the credit card companies suspend the card for more investigations. Thus, one feature that helps to detect abnormal transactions could be the aggregated number of transactions over a period, which helps detect the abrupt changes, i.e., many transactions in a short period, in the usage of a credit card. In reality, companies use a plethora of more complicated features to detect fraud in credit card transactions. There are also several efforts and articles that have put forward appropriate features for detecting fraudulent transactions (Bhattacharyya et al., 2011; Whitrow et al., 2009; Paasch, 2008).

After extracting features from both regular and fraudulent transactions, a classifier can be applied to discriminate fraudulent from regular transactions. To show how a classifier discrimination work, we demonstrate an example in 2-dimensional space based on two randomly selected features, as shown in  Fig. 6.1. Now, the classification task boils down to partitioning the space into two areas, each of which includes the transactions of one type (either regular or fraudulent). Based on  Fig. 6.1, we need to find the *decision boundary* to do such partitioning. A decision boundary is the area of a feature space where the output label of a classifier is not clear, but any area beyond that takes on a specific label. Therefore, the goal of a classifier is to find such a boundary, according to which future unseen samples will be classified into one of the predefined classes. For the example in  Fig. 6.1, the samples to the right of the decision boundary are labeled as non-fraudulent, while those to the left are assigned to the fraudulent transactions.

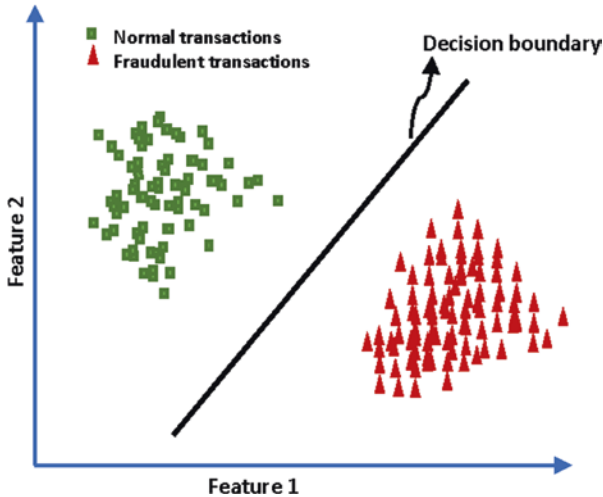
For the simple example in  Fig. 6.1, two features could discriminate the types of transactions. However, in more sophisticated real situations, more features are required for building classifiers with more discriminatory power. As an instance, there is a well-known Kaggle dataset¹ which contains around 285,000 transactions and 30 features. This dataset includes only 492 fraudulent transactions, forming 0.17% of all transactions. The features for each transaction entail the time of conducting transactions as well as its amount. The remaining features are the result of a transformation (based on principal component analysis, or PCA) and have thus no tangible meaning. ◀

6.2.2 An Overview of Classifiers

In the literature of supervised learning, there are several classifiers with plenty of variations. In this section, we provide an overview of the most popular classifiers, the implementation of which is also freely available in almost any programming language.

Logistic regression in contrast to its name is a binary classification technique. Logistic regression computes the likelihood that a given data point belongs to either of the two classes (Wright, 1995). The data points with a likelihood value below 0.5

1 ▶ <https://www.kaggle.com/mlg-ulb/creditcardfraud>.



■ **Fig. 6.1** An example of classifying the normal transactions, shown by green squares, and the fraudulent transactions, shown by red triangles. The decision boundary separates the sample of the two classes, the below and above regions of which correspond to the samples of one of the classes. (“Author’s own figure”)

are assigned to one class, while the remaining are designated to the other class. Interestingly, logistic regression can be extended with some limited effort to address multi-class classification problems.

Naive Bayes is a class of probabilistic classifiers, which are based on the strong conditional independence assumptions between the features (Rish et al., 2001). Though the independence assumption is likely to be violated in real-world situations, its simplicity and scalability to large-scale problems have made it one of the most popular classification techniques.

K-nearest neighbors or KNN is another popular classification technique whose underlying idea is arguably the simplest, among other methods. KNN first computes the distance of a new data point to other available training data points and then selects the K -nearest data points, where K can be any integer that is identified by the user (Cunningham & Delany, 2020). Finally, the data point is assigned to the class belonging to the majority in the identified K samples. In contrast to other learning algorithms, KNN does not have a separate training stage, and the training samples are just used to determine the K -nearest neighbors. The drawback of KNN is its high time complexity, where it needs to compute the distance of new data points to all the data points in the training set. It makes it particularly complicated for large datasets.

Support vector machines (SVMs) are a class of algorithms and methods that provides a decision boundary with the maximum distance from both classes. The term *support vector* refers to the data points which are close to the hyperplane, whose removal

would alter the separating decision boundary (Mohammadi et al., 2019; Suykens & Vandewalle, 1999). One of the salient features of SVMs is their ability to classify datasets with nonlinear boundaries, since it maps the input data to another space with a higher dimension. The underlying idea is that when data are mapped to a higher dimension, they will be linearly separable. The mapping of input data to another higher dimensional data is done by using *kernel functions* (Shawe-Taylor et al., 2004), and the resulting algorithm is called a kernel method. This feature has made SVMs one of the most popular supervised learning algorithms that have been applied to many problems in various fields.

Decision trees are a class of supervised learning, which includes several methods such as C4.5 and CART. The training process of a decision tree entails creating a treelike structure with decision nodes and leaf nodes. While decision nodes state some conditions on features and have at least several branches, they only represent the class labels to which the given instance belongs. Each decision node has at least two branches; moving to each of the branches is based on the value of a set of features. The leaf nodes represent the label of a given data point. A distinctive property of decision trees, in contrast to many classifiers, including SVMs and logistic regression, is its transparency, which means that it is explainable why a data point is assigned a specific label based on the values of its features.

6

6.2.3 Evaluating a Classification Model

Classifiers are typically evaluated based on performance metrics, which are based on basic statistics. To explain such statistics, we use the confusion matrix, a contingency matrix, as tabulated in ■ Table 6.1. Let us consider the previous example concerning fraudulent transactions. In this case, the table includes four simple values, which are extracted from the following sets:

- **True Positive (TP)**: The set of fraudulent transactions correctly predicted as fraudulent.
- **False Negative (FN)**: The set of fraudulent transactions wrongly predicted as non-fraudulent.
- **False Positive (FP)**: The set of non-fraudulent transactions wrongly predicted as fraudulent.
- **True Negative (TN)**: The set of non-fraudulent transactions correctly predicted as non-fraudulent.

■ Table 6.1 A confusion matrix for credit card frauds. “Table compiled by author”

		Predicted label	
		Fraudulent	Non-fraudulent
Actual label	Fraudulent	TP	FN
	Non-fraudulent	FP	TN

The most popular performance metric for classification is accuracy, which is defined as the ratio of the number of correctly classified data points to the number of all data points. Based on the contingency table, accuracy can be written as

$$\text{Accuracy} = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}, \quad (6.1)$$

where $|\cdot|$ is the cardinality of a set. Although accuracy computes the fraction of correct predictions, it is not appropriate in several circumstances, like credit card fraud prediction, where the number of data points for one of the classes, e.g., regular, is much higher than that of the other class, i.e., fraudulent. For example, in the Kaggle dataset, the fraudulent transactions form only 0.17% of all the transactions. It means that if a classifier assigns a regular label to any given data point, it achieves an accuracy of around 99.8%. Ironically, such a classifier has not detected any fraudulent transactions, which has been the main aim of devising a classifier in the first place. Thus, it is required to use some other metrics, such as precision, recall, and F-measure (also called F_1 score). Precision is an indicator of the correctness of a classifier in detecting fraudulent transactions and is defined as

$$\text{precision} = \frac{|TP|}{|TP| + |FP|}. \quad (6.2)$$

A higher precision value indicates that most of the transactions detected as fraudulent by a classifier are indeed fraudulent. Recall is a complementary metric that indicates the percentage of correctly detected fraudulent transactions out of all the fraudulent transactions in the given set and is defined as

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|}. \quad (6.3)$$

Recall indicates the completeness of a classifier in detecting abnormal transactions. Hence, a higher value for this metric shows the extent to which the fraudulent transactions are detected. Both precision and recall have some ignorance; the former considers only false positives and true positives, but not false negatives, and the latter does not take into account false positives. That is why each of these metrics highlights one aspect of a classifier. As a combination of precision and recall, F-measure is defined as the harmonic mean of the metrics, i.e.,

$$F\text{-measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \quad (6.4)$$

6.2.4 Designing a Pipeline for Machine Learning Classification

Figure 6.2 provides an overview of a pipeline for classification instantiated for detecting credit card frauds. It consists of three main components: (1) a repository of credit card transactions, (2) a repository miner, and (3) a credit card fraud detector based on a machine learning classifier.

6.2.4.1 Data Mining

First of all, both fraudulent and non-fraudulent credit card transactions should be available in a repository. Please remember that both classification and regression are supervised problems. For each transaction, we have several input variables and an output variable (i.e., fraudulent or non-fraudulent). Our goal is to create a function to map the input to the output. Data can have different formats that are (1) structured, (2) unstructured, or (3) semi-structured.

- **Structured data** are tabular data that are very well defined in rows and columns. The format is rigorous: we know how many columns there are and what kind of data they contain. This kind of data can be stored in databases that represent the relationships between the data as well.
- **Unstructured data** are the rawest form of data whose data extraction is usually hard. Unstructured data must be abstracted to understand which features to consider and how to transform it into a readable format. An example is the extraction of topics from movies to label them as positive or negative.
- **Semi-structured data** are composed of both structured and unstructured data. Although a consistent format is defined, it is not rigorously applied. Semi-structured data are often stored as files.

6.2.4.2 Data Preprocessing

Before feeding a machine learning classifier, data should be transformed in a structured format: therefore, we need a repository miner. In our example, this software component mines the transactions, asks experts to manually label the transactions as fraudulent or non-fraudulent, and extracts their features. Finally, a machine learning model is trained on the data (i.e., extracted features from the gathered transactions): the obtained model will enable us to classify the instances. Before

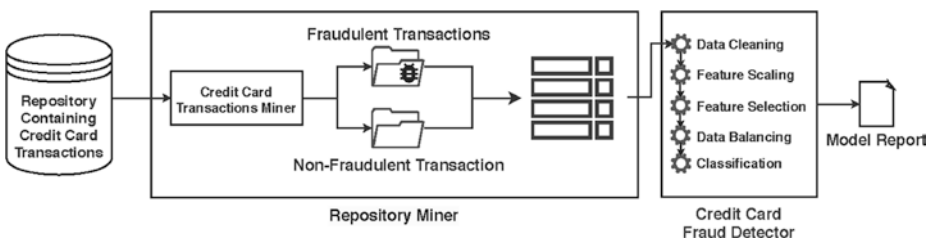


Fig. 6.2 A pipeline for ML classification of credit card transactions. (“Author’s own figure”)

training the model, several further steps are required: (1) data cleaning, (2) feature scaling, (3) feature selection, and (4) data balancing.

Data Cleaning A classifier cannot work with missing values, and hence an extra step is required. In particular, given a feature with some missing value, there are three alternatives: (1) remove the feature; (2) remove all the instances containing missing values on that feature; and (3) replace the missing values with other values (e.g., zero, the mean).

Feature Scaling First of all, the data instances should be preprocessed to homogenize data coming from different sources of information. This method is called data normalization (Han et al., 2011) and is used to normalize the range of independent variables or features of data. The most common data normalization techniques are (1) min-max normalization and (2) standardization. The former rescales the values for a specific feature in the interval $[0, 1]$. For each feature, the minimum value will be transformed into 0, while the maximum will be a 1:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}. \quad (6.5)$$

Standardization transforms the values for each feature so that their mean is 0 and their standard deviation is 1. In other words, given for each feature the mean of its values (\bar{x}) and the standard deviation (s), standardization is computed as follows:

$$x' = \frac{x - \bar{x}}{s}. \quad (6.6)$$

Feature Selection Not all features mined in the dataset may be helpful for classification. For example, they could be constant or do not provide useful information exploitable by a learning method for a particular dataset. Feature selection (Guyon & Elisseeff, 2003) reduces the size of the dataset, speeds up the training, and selects the optimal number of features that maximizes a given performance criterion.

Data Balancing Once feature selection is finished, the training data are balanced such that the number of fraudulent instances equals the number of non-fraudulent instances. Data balancing (He & Garcia, 2009) can be introduced by resampling/transforming the training set or by using meta-classifiers (e.g., cost-sensitive classifiers).

6.2.4.3 Data Classification

The normalized data and the learning algorithms described in ► Sect. 6.2.2 are used to build the learner. A validation step must be carried out to assess the performance of the machine learning pipeline. There are many validation strategies: the simplest one is holdout validation. The data is split into three subsets: (1) the training set to train the model, (2) the validation set to tune the classifier through feature selection and data balancing, and (3) the test set to perform a final estimation of the model performance after training and validation. The test set should never be used to make decisions regarding the training. Before the learner is tested, the orig-

inal test data are normalized, applying the same technique used for the training data. The dimension is reduced to the same subset of attributes from *feature selection* set. After comparing predicted and actual values, the performance is obtained in terms of the metrics shown in ► Sect. 6.2.3.

Holdout validation has several limitations. First of all, partitioning the available data into three sets could drastically reduce the number of samples. Then, the model performance can depend on a particular random choice for the train, validation, and test sets. For this reason, *cross-validation* (Stone, 1974) is used to verify model generalizability. Its goal is to evaluate to what extent a model can predict unseen data (i.e., data not used in either training or validation) and to give insights concerning the model generalization on independent datasets. There are several ways to perform *cross-validation*. Among these, *k-fold cross-validation* (Stone, 1974) is one of the most common. This methodology randomly partitions the data into k folds of equal size. A single fold is used as the test set, while the remaining ones are used as the training set. The process is repeated k times, using each time a different fold as the test set.

6

6.3 Supervised Learning: Regression

Regression is one of the most popular techniques to investigate the relationship between a dependent variable (or label) and one or multiple independent variables (or predictors). In contrast to classification, in regression, dependent variables are numerical.²

We first focus on the case where there is a linear relationship between a dependent and a set of independent variables. Then, we briefly discuss the framework of kernel methods, regression with multiple independent variables, as well as nonlinear relationship between the dependent and independent variables. In the subsequent section, we first explain the regression by simple linear regression, where there is only one independent variable. Then, we present an overview of the most well-known regression methods, followed by the metrics to evaluate the quality of a regression. Finally, we provide an introduction on how to design a machine learning pipeline for data regression.

6.3.1 Simple Linear Regression

► Example

Simple linear regression is the case where there is only one independent variable, with a linear relation with the dependent variable. For example, assume that finding the relationship between the annual returns to the Standard & Poor's 500 (S&P 500) and the annual returns to Apple stock is sought. The S&P 500 contains the 500 most valu-

² There is another type of regression, named ordinal regression, where the dependent variables are of ordinal type, each showing a rank assigned to a sample within the dataset.

able stocks in the United States, including the Apple stock. As a result, it is realistic to assume that the annual returns to S&P 500 have something to say regarding the annual returns to the Apple stock.

The relationship between the annual returns can be analyzed by using simple regression, where Apple stock is treated as a dependent variable (y) and S&P stock as the independent variable (X). The outcome of a regression analysis determines how a change in S&P stock affects Apple stock. For simplicity, we assume that there is a linear relationship between S&P and Apple stocks. Intuitively, it means that the graph between the two variables is a straight line, which can be written as

$$y = \beta X + \beta_0, \quad (6.7)$$

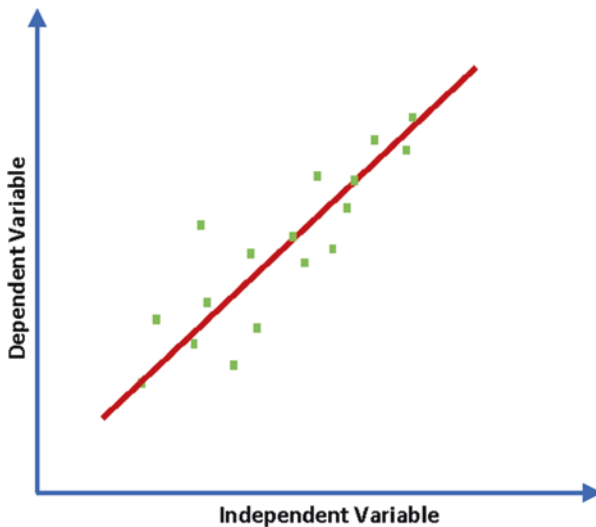
where β is the slope of the line and β_0 is the y -intercept. Evidently, β and β_0 are not known in advance, and the goal of regression is to estimate these parameters based on some observed data. Therefore, the regression equation for n observed data points can be stated as

$$y_i = \beta X_i + \beta_0 + \epsilon_i, \quad i = 1, \dots, n, \quad (6.8)$$

where n represents the number of samples, and ϵ_i is a random noise or error term. The existence of ϵ_i is essential since it is virtually always impossible to find a line that crosses through all the samples. Instead, we assume that there is a noise that contaminated the observations, and the desired line is the one that has the minimum distance to all the samples.

For the case of estimating the annual returns, the samples could be the records of the previous years regarding the annual returns of Apple and S&P stocks. The linear relationship between X and y can be investigated for the simple regression by a scatterplot.

■ Figure 6.3 shows a scatterplot of a dependent and independent variable with a linear relationship. ◀



■ Fig. 6.3 Regression of a dependent and an independent variable with a linear relationship. (“Author’s own figure”)

6.3.2 Regression Methods: An Overview

There are different regression techniques whose implementation is also freely available in almost any software or programming language. In the following, we give a brief overview of the essential techniques known in the literature, each with specific advantages/drawbacks.

Ordinary least square (OLS) is arguably the most well-known and straightforward regression method. The principle of OLS is particularly simple: minimizing the squares' sum between the observed dependent variable and the predicted value by the linear function. The optimization regarding OLS provides a closed-form solution, requiring to compute the inverse of the covariance matrix of the independent variables.³ However, the matrix is not necessarily invertible, making the overall computation in OLS unstable.

Ridge regression is another popular technique in regression analysis, which also provides a closed-form solution. The difference between ridge regression and OLS is subtle but important: It only has an extra ℓ_2 regularization term, which adds a penalty of the square of the magnitude of coefficients (Hoerl et al., 1975). This simple adjustment resolves the problem with inverting of the matrix by adding a positive diagonal matrix, making the computation of the regression coefficient stable. However, its drawback is to specify a regularization parameter, which is a trade-off between the least square error and the regularization term.

Lasso is similar to ridge regression with the difference that the ℓ_2 regularization term is replaced by an ℓ_1 term, so it has a penalty of the absolute value of the magnitude of coefficients. The change is seemingly infinitesimal, but it has many positive and negative consequences. The ℓ_1 norm motivates the regression coefficient to have many zero elements. Therefore, the nonzero coefficients show the importance of the corresponding independent variables (Hastie et al., 2015). This interesting feature allows us to use lasso both as a regression and variable selection technique (or what is called feature selection in machine learning). The main drawback of lasso is that, on the other hand, the ℓ_1 norm is not differentiable. In other words, not only lasso does not yield a closed-form solution, but even the standard algorithms for convex programming cannot be directly applied to this problem. Such a drawback places a serious obstacle to its use for real-world problems, but there are many efficient solvers that can provide a reliable solution in a reasonable time (Beck & Teboulle, 2009; Kim et al., 2007; Mohammadi, 2019).

Support vector regression (SVR) is the SVM extension for regression and utilizes the concept of support vectors for regression (Drucker et al., 1997). SVR is a powerful regression technique and has shown promising performance on different problems. Similar to SVM, working out a regression problem using SVR also

³ In fact, it is the inverse of the covariance matrix if the data is normalized.

entails solving constrained quadratic programming, which can be time consuming for large-scale problems. However, efficient solvers exist in the literature that can be used for the analysis.

Except for SVR, other regression methods are introduced for linear regression in their original form. However, SVR handles nonlinear regression by using proper kernel methods. The similar technique used by SVR can be applied to other regression techniques as well to handle nonlinear relationships between independent and dependent variables.

6.3.3 Evaluating a Regression Model

The performance of regression techniques is also evaluated by using different metrics. For measuring the fitness of a regression method, we show the dependent variable by y and the predicted values by the regression method by \hat{y} . The proximity of these two values indicates the goodness of fit, while a significant difference is a sign of an unreliable regression.

Several metrics for evaluating the performance of a regression method are based on the prediction error, which is the difference between the true and the predicted values across all the data points. Since the prediction can be above or below the true value, a popular way of calculating the prediction error is to compute the sum of squared errors (SSE) as

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (6.9)$$

A higher SSE value shows that the regression method's prediction deviates significantly from the true values, making the overall regression less reliable. On the other hand, a lower SSE value is an indicator of a good fit of the regression method. A problem of SSE is that it squares the error term, magnifying the influence of larger errors. This issue is particularly unwanted when only some of the data points are contaminated with large errors, because even those few samples can profoundly influence SSE. A remedy for such a case is to use the absolute value of errors (AVE) defined as

$$AVE = \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (6.10)$$

AVE does not increase the influence of larger errors. But both SSE and AVE virtually always increase when there are a higher number of data points in the dataset. That is why the mean SSE (MSSE) and mean AVE (MAVE) are usually used, which are the average of squared or absolute value of errors over the data points, respectively.

6.3.4 Designing a Pipeline for Machine Learning Regression

Similar to classification, we can build a pipeline for regression as well. ■ Figure 6.4 provides an overview of such a pipeline for predicting stock prices. It consists of three main components: (1) a repository of stock prices, (2) a repository miner, and (3) a stock price estimator based on a machine learning regression model. In the following, we provide an overview on the data mining and preprocessing steps.

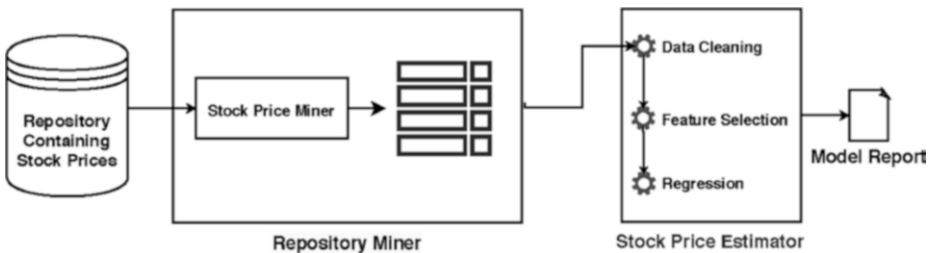
6.3.4.1 Data Mining

First of all, previous instances of stocks with their features and corresponding price should be available in a repository. As for classification, for each stock, we have several input variables and an output variable (i.e., the price). The goal is to create a function to estimate the stock prices based on the extracted features.

6.3.4.2 Data Preprocessing

Before feeding a machine learning regression model, data should be transformed in a structured format: therefore, we need a repository miner. In our example, this software component mines the stocks along with their prices and extracts their features. Finally, a machine learning model is trained on the data: the obtained model will enable us to estimate the stock prices. Before training the model, several further steps are required: (1) data cleaning and (2) feature selection. Data cleaning is performed similarly as in classification models, while feature selection differs. Choosing which feature to select and their order is essential to apply regression models to datasets with many features. Indeed, given a dataset with p features, the number of possible models is 2^p . In other words, given a dataset with 100 features, it is possible to create 1, 267, 650, 600, 228, 229, 401, 496, 703, 205, and 376 distinct models. Therefore, better alternatives must be found: one of them is stepwise regression, which comprises three approaches:

- **Forward Selection:** The selection starts with a model with no predictor ($Y = \beta_0$). Then, p linear regressions are executed, and the predictor for which the error is minimized is added. The process is repeated until adding more predictors does not improve the model in a statistically significant manner.
- **Backward Elimination:** This algorithm commences with a model containing all feature variables $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$ and which accuracy is evaluated. At each step, the algorithm removes the most detrimental variable to the regression. The deletion step is repeated until the accuracy of the model improves.



■ Fig. 6.4 A pipeline for ML regression of stock prices. (“Author’s own figure”)

Concluding Remarks

In this chapter, we discussed supervised learning and studied two types of such algorithms, namely classification and regression. While classification and regression are explained through tangible examples, such techniques can be used in many other business problems and can provide more insights based on the available data. For doing so, proper data engineering pipelines are also discussed for classification and regression, which enable the practitioners to implement any classification or regression algorithm for their problem readily and evaluate the results of the technique being utilized by using proper performance metrics.

Take-Home Messages

- Machine learning creates models that automatically improve through experience “without being programmed.”
- Supervised learning is the subset of machine learning, where the observation must be labeled to feed the model.
- Classification methods predict the category of unseen observations.
- Regression methods investigate the relationships between a dependent variable and one or multiple independent variables.
- Designing and evaluating a machine learning pipeline are essential to crafting effective machine learning models.

References

- Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1), 183–202.
- Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3), 602–613.
- Cunningham, P., & Delany, S. J. (2020). k-nearest neighbour classifiers. *arXiv preprint arXiv:2004.04523*.
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems* (pp. 155–161).
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.
- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. CRC Press.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
- Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistics-Theory and Methods*, 4(2), 105–123.
- Kim, S.-J., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2007). An interior-point method for large-scale ℓ_1 -regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4), 606–617.
- Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, 42(11), 30–36.

- Mohammadi, M. (2019). A projection neural network for the generalized lasso. *IEEE Transactions on Neural Networks and Learning Systems*, 31, 2217–2221.
- Mohammadi, M., Mousavi, S. H., & Effati, S. (2019). Generalized variant support vector machine. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51, 2798–2809.
- Paasch, C. A. W. (2008). *Credit card fraud detection using artificial neural networks tuned by genetic algorithms*. Hong Kong University of Science and Technology (Hong Kong).
- Rish, I., et al. (2001). An empirical study of the naive bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 3, 41–46.
- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210–229.
- Shawe-Taylor, J., Cristianini, N., et al. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.
- Stone, M. (1974). Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111–133.
- Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3), 293–300.
- Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery*, 18(1), 30–55.
- Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 217–244). American Psychological Association.

Further Reading

- Aldridge, I. (2013). *High-frequency trading: A practical guide to algorithmic strategies and trading systems* (Vol. 604). John Wiley & Sons.
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.
- Mitchell, T. M. (1997). *Machine learning* (Vol. 45(37), pp. 870–877). McGraw Hill.



An Intuitive Introduction to Deep Learning

Eric Postma and Gerard Schouten

Contents

- 7.1 Brief Historical Overview – 123**
- 7.2 Datasets, Instances, and Features – 123**
- 7.3 The Perceptron – 124**
 - 7.3.1 The Decision Boundary – 125
 - 7.3.2 The Delta Learning Rule – 126
 - 7.3.3 Strengths and Limitations of the Perceptron – 127
- 7.4 The Multilayer Perceptron – 128**
 - 7.4.1 Combining Decision Boundaries – 129
 - 7.4.2 The Generalized Delta Learning Rule – 130
- 7.5 Deep Neural Networks – 131**
 - 7.5.1 Combinations of Combinations of ... Decision Boundaries – 131
 - 7.5.2 The Generalized Delta Learning Rule in Deep Networks – 132

7.5.3	From Two- to High-Dimensional Feature Vectors – 133
7.6	Convolution: Shifting a Perceptron Over an Image – 133
7.6.1	The Basic Convolution Operation – 133
7.7	Convolutional Neural Networks – 136
7.7.1	Convolutional Layers – 136
7.7.2	Pooling Layers – 137
7.7.3	Combinations of Combinations of ... Features – 137
7.7.4	Dense Layers – 138
7.7.5	From AlexNet to Modern CNNs – 138
7.8	Skin Cancer Diagnosis: A CNN Application – 138
7.8.1	Introduction – 138
7.8.2	Data Collection and Preparation – 139
7.8.3	Baseline and Multitask CNN – 141
7.8.4	Experiments and Results – 142
7.8.5	Conclusion on the CNN Application – 143
	References – 144

Learning Objectives

After having read this chapter, you will be able to:

- Identify the different types of neural networks: perceptron, multilayer perceptron, and convolutional neural networks (CNNs).
- Understand the different levels of processing and abstraction associated with neural networks of increasing depth.
- Understand the training and operation of deep networks in terms of combinations of decision boundaries.
- Understand the practical application of CNNs.

7.1 Brief Historical Overview

The recent upsurge in artificial neural networks, deep learning, has led to major advances in a wide variety of domains. In this chapter, we present a gentle introduction to deep learning, using elementary high school mathematics only. The emphasis is on feedforward types of neural networks, namely perceptrons, multilayer perceptrons, and convolutional neural networks. For a historical overview of deep learning, the interested reader is referred to Chap. 1 of the main deep learning textbook (Goodfellow et al., 2016) or to a slightly alternative review (Schmidhuber, 2015).

7.2 Datasets, Instances, and Features

Machine learning algorithms learn by means of examples, also called instances. The prevalent type of learning is called supervised learning, which means that each instance is accompanied by a label that represents the class or value associated with that instance. ■ Figure 7.1 shows an example for an image classification task: an image showing a dog as an instance and the accompanying label DOG. Machine learning algorithms often incorporate parameterized models of the input-output relation, where the input represents the instance and the output the model's estimate of the accompanying label. In the case of deep learning, the parameters consist of connection weights. The aim is to tune these weights in such a way that given an instance, the accompanying label is predicted correctly.

In traditional machine learning, which was prevalent before the recent advent of deep learning, algorithms were unable to automatically classify images as shown in ■ Fig. 7.1. Instead, machine learning relied on features, numerical descriptors of images that represent the visual characteristics in terms of numbers. Two examples of such descriptors are histograms of color values and proportions of contours in a certain orientation.



■ Fig. 7.1 An image of a dog. The image is the instance, and a possible label is DOG. (Author's own)

7

7.3 The Perceptron

The perceptron is the building block of deep neural networks. In this section, we provide some intuition and explanation of its structure and function. Throughout our discussion, we limit ourselves to two-dimensional feature vectors. In other words, our instances consist of two numbers. Although realistic machine learning problems typically involve high-dimensional features, our restriction to two features allows us to visualize instances as points on a two-dimensional plane. The two feature values of an instance define the coordinates in the feature plane. The labels are represented by symbols. A scatterplot is an example of such a two-dimensional feature space. The horizontal axis represents the value of feature 1 (F1), and the vertical axis represents the value of feature 2 (F2). A dataset consists of a cloud of points in the scatterplot. In case of a binary classification task, the points have two different colors, each representing one of the classes. For instance, the negative instances could be represented by black dots and the positive ones by white dots.

The equation to compute the output of the perceptron is

$$o = \text{sgn} \left(\sum_0^N w_i F_i \right), \quad (7.1)$$

where $N = 2$, $F_0 = 1$, and w_0 represents the bias. The use of N for the number of inputs makes explicit that our two-dimensional input readily generalizes to higher dimensions ($N > 2$). The transfer function sgn returns $+1$ in case its input is larger or equal to zero, and -1 otherwise. In other words, this equation takes the weighted sum of all features and returns $+1$ if the sum exceeds zero, and -1 otherwise.

7.3.1 The Decision Boundary

As we will see, training the perceptron with two inputs corresponds to the proper positioning of a straight line in feature space that separates the positive instances from the negative ones. This line is called a *decision boundary* that has a positive and a negative side. Initially, before training, the two input weights and the bias weight of the perceptron have random values. As a consequence, the decision boundary has a random orientation and intercepts with the F2 axis. It is quite easy to demonstrate that the perceptron with two features incorporates the equation of a straight line. As you may remember from high school, the general equation of a line is

$$y = ax + b, \quad (7.2)$$

where x and y represent the horizontal and vertical axes, respectively; a the slope of the line; and b the intercept with the y -axis. By varying a and b , the orientation and intercept with the y -axis can be manipulated. We now reformulate the perceptron Eq. (7.1) into the form of Eq. (7.2). We start by observing that the decision boundary is at the transition from the perceptron output of -1 to $+1$ and hence defined by the equality

$$\sum_{i=0}^N w_i F_i = 0 \quad (7.3)$$

which for two inputs corresponds to

$$w_0 + w_1 F_1 + w_2 F_2 = 0, \quad (7.4)$$

where we have used $F_0 = 1$. This can be rewritten as

$$w_2 F_2 = -w_0 - w_1 F_1, \quad (7.5)$$

and by dividing both sides by w_2 and rearranging the right-hand-side terms as

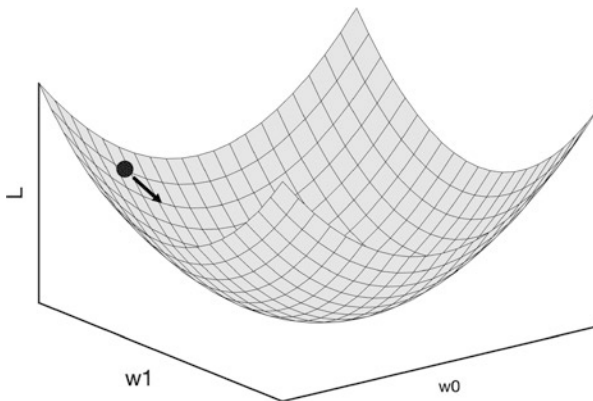
$$F_2 = -\frac{w_1}{w_2} F_1 - \frac{w_0}{w_2}. \quad (7.6)$$

Given that in our two-dimensional feature space the F1-axis corresponds to the x -axis and the F2-axis to the y -axis, we have obtained the equation of the decision boundary in the form of Eq. (7.2), with $a = -w_1/w_2$ and $b = -w_0/w_2$. So, the perceptron with two inputs represents a line, i.e., the decision boundary, in two-dimensional feature space. The perceptron learning rule attempts to position and orient the line in such a way that it separates the positive instances from the negative ones.

7.3.2 The Delta Learning Rule

In order to adapt the weights of the perceptron, a measure of the quality of its predictions is needed. An error function, also referred to as a *loss function*, measures the deviations between the actual output generated by the perceptron and the desired output, i.e., the label. For example, when a positive instance is presented to the input of the perceptron and the perceptron output value is negative, their difference is larger than zero. The error provides a cue to determine how much the individual weights have to be changed and in what direction. This is done by means of the delta learning rule. Intuitively, this rule is applied to each weight in such a way that weight values are updated to decrease the loss. The mathematical term for such updating is gradient descent. In optimization problems, such as finding the shortest route or finding the best schedule given a set of constraints, gradient descent is a standard approach to find a solution. To illustrate how gradient descent works for the perceptron's delta learning rule, we consider a perceptron with two parameters, a single input weight w_1 and a bias weight w_0 . ■ Figure 7.2 shows the loss surface that specifies the loss L for each combination of values of w_0 and w_1 .

The loss surface is convex, implying that there is a single optimal combination of values for both weights. For this combination, the perceptron gives outputs with the best correspondence to the labels. Since the perceptron is randomly initialized, the initial weight values, prior to learning, correspond to a random location on the loss surface. From this random start location, the delta learning rule moves the weights along the negative gradient of the loss surface. It can be likened to a ball that is pulled downward by gravity. Ultimately, the weights will converge to the optimal location. It is important to note that the depiction of the entire loss surface is a bit misleading. At any moment, only a local region surrounding the current location is known. The situation is comparable to a walk in the Scottish Highlands in a very dense fog. You cannot see much by looking around, but in order to find the lowest point, you simply take small steps in the directions of largest descent.



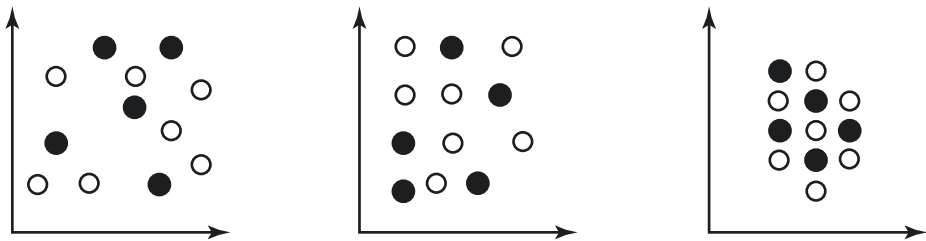
■ Fig. 7.2 Illustration of a convex loss surface L as a function of w_0 and w_1 . The current state of the network is represented by the black sphere. The gradient is indicated by the arrow. (Author's own)

7.3.3 Strengths and Limitations of the Perceptron

The perceptron can be trained on simple tasks, such as determining if mushrooms are poisonous from a number of mushroom descriptors, such as its color and shape.¹ Training the perceptron on a labeled set of instances, where the labels are poisonous versus edible, it is able to give a prediction for unseen instances. In the visual domain, perceptrons can be trained on the classification task of handwritten digits (0 to 9) or characters (a to z). In both cases, multiple perceptrons process the input simultaneously, yielding what we will refer to as a “parallel perceptron.” Each of the constituent perceptrons is responsible for the recognition of a single digit or character. In such applications, the discontinuous transfer function of the perceptron is generally replaced by a smoother function, e.g., a linear function, to allow for graded outputs. This allows to determine which of the 10 (for digits) or 26 (for characters) has the largest activation.

As emphasized by Minsky and Papert (1969), perceptrons suffer from several limitations, the main of which originates from the incorporation of a linear decision boundary. Whenever positive and negative instances (in a binary classification task) can **not** be separated by a single straight line, the perceptron cannot solve the task. ■ Figure 7.3 shows three examples of such tasks that are not linearly separable. In all three examples, the white and black circles cannot be perfectly separated by a single straight line.

This main limitation of the perceptron can easily be alleviated. As was already clear to Rosenblatt (1958), combining multiple decision boundaries (perceptrons) would enable to deal with the tasks depicted in ■ Fig. 7.3. Stacking perceptrons is quite straightforward. However, the main obstacle is to find a generalization of the delta learning rule for such stacked perceptrons. The discontinuity imposed by the sgn transfer function imposes a mathematical obstacle, because the derivative is undefined at the threshold. The derivative is important to compute the gradient of the loss surface in order to find the weights that achieve the best performance or minimal loss.



■ Fig. 7.3 Three examples of classification tasks that are not linearly separable. Each task comprises two features, represented by the horizontal and vertical axes, and two classes (the black and white circles). (Author’s own)

1 ► <https://archive.ics.uci.edu/ml/datasets/mushroom>.

7.4 The Multilayer Perceptron

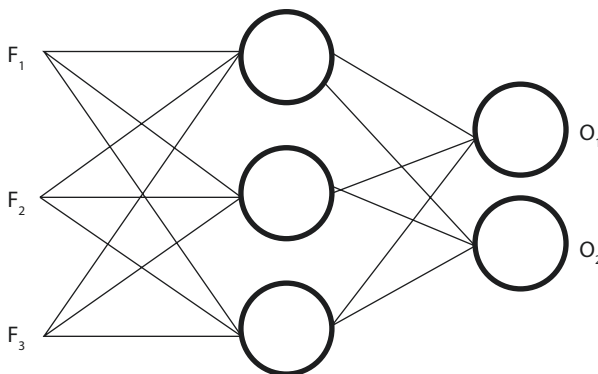
The multilayer perceptron combines perceptrons by stacking them. The most successful multilayer perceptrons in the 1990s consisted of two layers of perceptrons. The first layer connected the input layer to a so-called hidden layer, and the second layer connected the input layer to the output layer. In the terminology of multilayer perceptrons, the hidden layer consists of hidden neurons or units, but in fact they correspond to the parallel perceptrons described in the previous section. The number of hidden units in the multilayer perceptron is a so-called hyperparameter. The value of the hyperparameter has to be determined beforehand and is typically optimized empirically. ■ Figure 7.4 is an illustration of a multilayer perceptron with three inputs, three hidden units, and two output neurons. All inputs have weighted connections to all hidden units, and all hidden units have weighted connections to all outputs. In addition, all hidden and output neurons have a bias weight. In the deep learning community, such fully connected layers are often referred to as dense layers, to distinguish them from layers with fewer connections.

The equation for the value of the k -th output of the multilayer perceptron is

$$o_k = f \left(\sum_{h=0}^H w_{hk} f \left(\sum_{i=0}^I w_{ih} F_i \right) \right), \quad (7.7)$$

where h is the index over the H hidden neurons, i is the index over the I input neurons, and the weights w_{ok} and w_{oh} represent the bias weights. The $f()$ s are sigmoid (S-shaped) functions, typically of the form

$$f(x) = \frac{1}{1 + e^{-x}}. \quad (7.8)$$



■ **Fig. 7.4** Example of a multilayer perceptron consisting of three inputs (F_1 , F_2 , and F_3), three hidden units (the three circles in the middle), and two output units (O_1 and O_2). The weighted connections are represented by the lines connected to the circles. Bias weights are not shown. (Author's own)

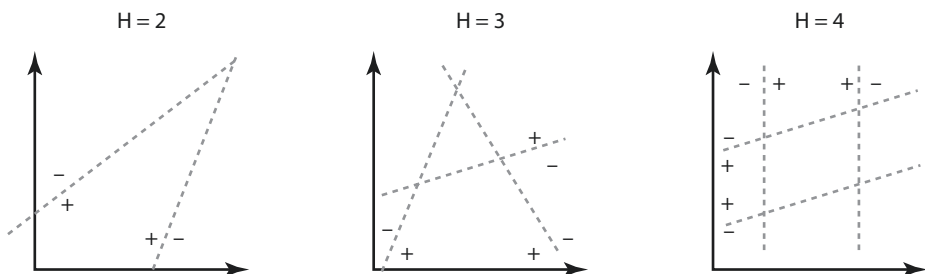
The sigmoid function is a smooth version of the sgn function. As stated in the previous section, such a smooth function² is needed to be able to realize a learning rule for the multilayer perceptron. In terms of decision boundaries, the effect of the sigmoid function is that the decision boundary becomes a bit fuzzier. Instead of a sharp line, it is smeared out a bit. In what follows, we assume that the decision boundaries are sharp so as to facilitate the intuitive understanding of multilayer networks and their deeper variants.

It is important to note that Eq. (7.7) is a nested version of perceptrons according to Eq. (7.1), with the sgn function replaced by the sigmoid function $f()$. By stacking perceptrons, the multilayer perceptron can combine multiple decision boundaries to deal with tasks that are not linearly separable.

7.4.1 Combining Decision Boundaries

The ability of stacked perceptrons to combine decision boundaries is already evident from the parallel perceptron discussed before. Given an input layer representing the feature vector, a parallel perceptron consisting of H perceptrons represents H decision boundaries. For $H = 3$, the combination of three decision boundaries can separate a triangular area in two-dimensional feature space. ■ Figure 7.5 shows three examples of combinations of decision boundaries for three different values of H . Boundaries have a polarity in the sense that all instances on one side of the boundary lead to activation of the associated perceptron and all instances on the other side to inactivation. In the figure, the polarities are indicated by + and – signs. As a result of the polarities, for example, the middle example ($H = 3$) creates a triangular area in which all instances activate all three perceptrons.

A parallel perceptron is not sufficient to deal with tasks that are not linearly separable. At least one perceptron per final output should be placed on top of the parallel perceptron to integrate the activations of H perceptrons. This yields the multilayer perceptron with a single hidden layer (formed by the parallel perceptrons).



■ Fig. 7.5 Three examples of decision boundaries. The number of perceptrons (H) determines the number of boundaries. Each boundary has a polarity (a + and a – sides); these are indicated next to each boundary. (Author’s own)

² In mathematical terms, the sigmoid function is continuous and differentiable.

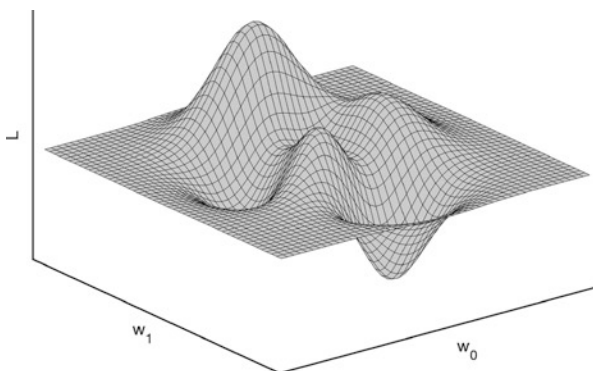
The number of hidden units H is a hyperparameter that should match the complexity of the tasks. Each of the examples shown in Fig. x requires at least the specified number of H to be solved. For two-dimensional feature spaces, this is easy to see. In realistic machine learning applications involving often much more than two features, the optimal value of H is determined empirically by training the multilayer perceptron on the same task for different values of H . The value for which the best prediction performance is obtained is considered to be the optimal one.

7.4.2 The Generalized Delta Learning Rule

7

The main innovation underlying the multilayer perceptron is the development of the generalized delta learning rule (Goodfellow et al., 2016; Rumelhart et al., 1986; Schmidhuber, 2015). The learning rule culminates in an algorithmic procedure called backpropagation that consists of three steps. The first step is forward propagation, which corresponds to the application of Eq. (7.7) to a given feature vector of instance I . The second step is comparing the resulting output to the label of instance I , yielding a loss value. The third step is the backward propagation of the loss through the network, starting at the hidden to output weights, and subsequently moving to the input to hidden weights. In this step, each weight will be updated in a direction that reduces the loss. So, also in this case, the Scottish Highlands metaphor applies. However, this time, there is not a single valley but there are many ones with different depths. Following the negative gradient by taking small steps in the downward direction is not guaranteed to lead to the lowest loss value. You may get stuck in a local minimum. Whatever direction you move, you will always go up. Neighboring valleys that may be much deeper are unreachable.

In terms of optimization problems, whereas the loss function of the perceptron is a convex function (single valley), the loss function of the multilayer perceptron is a non-convex function (many valleys with different depths). ■ Figure 7.6 illustrates this in a two-dimensional sketch of the loss surface. The horizontal axes represents the weight values and the vertical axis the loss value. Initially, the weights



■ Fig. 7.6 Example of a non-convex loss surface. (Author's own)

of the multilayer perceptron are set to random values, and hence the start position of the multilayer perceptron on the loss surface is random as well.

Training multilayer perceptrons consists of repeated presentation of instances according to the three backpropagation steps. The term epoch is used to indicate the application of the three steps to all instances in the training set. Multilayer perceptrons are trained for several hundreds or thousands of epochs, because at each location of the landscape the gradient has to be determined anew (this is due to the dense fog).

Multilayer perceptrons have been successfully applied to a wide variety of tasks, such as handwritten zip code recognition, phoneme recognition, and object classification from radar echo. In terms of prediction performance, multilayer perceptrons are typically outperformed by non-perceptron approaches, such as support vector machines or random decision forests.


7.5 Deep Neural Networks

Increasing the depth of multilayer perceptrons is straightforward. Initial attempts in the 1990s to train deeper multilayer perceptrons failed. The common assumption was that this failure was due to fundamental limitations of deep networks. In recent years, it turned out that this assumption was wrong. Under the generic header of *deep learning*, deep multilayer networks outperform traditional machine learning on a wide range of tasks, such as image recognition, speech recognition, translation, and many more. Three reasons for the success of deep learning are (1) availability of more data (required to tune the large increase in the number of parameters); (2) availability of more powerful computers, especially the parallel processing capacity of graphical processing units (GPUs); and (3) innovations in the details of deep (multilayer perceptron) networks, most notably the introduction of *rectified linear units* (ReLUs (Jarrett et al., 2009)), that replace the sigmoid transfer function for the hidden layers. A ReLU is a piecewise linear function defined as

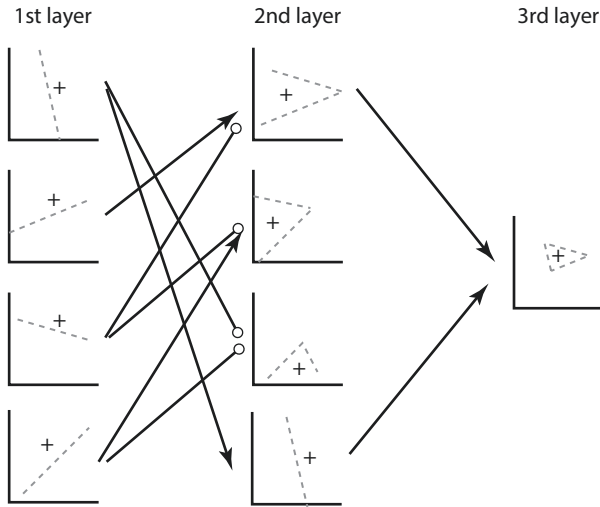
$$\text{ReLU}(x) = \max(0, x), \quad (7.9)$$

in which x represents the weighted sum entering the perceptron.

7.5.1 Combinations of Combinations of ... Decision Boundaries

Increasing the depth of multilayer networks results in deep learning networks in which each hidden layer acts upon the outputs of previous hidden layers. In terms of decision boundaries, this means that the first hidden layer defines decision boundaries, the second hidden layer defines combinations of decision boundaries, the third layer combinations of combinations of decision boundaries, and so forth. An illustration of the combinatorial advantage of deep layered networks is provided in  Fig. 7.7.

This recursive or nested application of decision boundaries is partially responsible for the tremendous power of deep learning.



■ **Fig. 7.7** Illustration of the recursive combination of decision boundaries in a three-layered multilayer perceptron. Each plot represents a perceptron and its associated decision boundary. The first layer consists of four perceptrons with the familiar straight decision boundaries. The second layer combines decision boundaries of the first layer. The arrows indicate a positive combination, and the arrows terminating in a circle indicate a negative combination in which the polarity of the first-layer decision boundary is swapped. The single perceptron in the third layer forms a combination of decision boundaries of the second layer. (Author's own)

7.5.2 The Generalized Delta Learning Rule in Deep Networks

The generalized delta learning rule in deep networks does not differ from that of multilayer perceptrons. However, the large depth of deep networks requires a more careful consideration of the random initialization of the weights. Assuming sigmoid transfer functions in the hidden layers constituting the deep network, the weighted sum feeding into each hidden neuron should be kept within bounds. For instance, if the weighted sum is negative and too large, the sigmoid function is pushed asymptotically towards zero. This implies that the information flow through this particular hidden neuron is shut off for subsequent layers higher up in the deep network. If this happens to many hidden neurons in the early layers, the forward propagation of inputs to outputs is disrupted and learning fails. Since the generalized delta rule updates weights by multiplying them with the derivative of the sigmoid, asymptotic values of the sigmoid are problematic as well. The derivatives at the almost horizontal extremes of the sigmoid are near zero. As a consequence, weight updates are almost zero as well and backpropagation learning effectively stops. This problem is known as the *vanishing gradient* problem (Goodfellow et al., 2016).

7.5.3 From Two- to High-Dimensional Feature Vectors

Up to this point, we have assumed that our perceptrons, multilayer perceptrons, and deep networks have two-dimensional feature vectors as input. The reason for this was that it allowed us to illustrate the functioning of these networks in terms of decision boundaries. Increasing the number of features (i.e., the dimensionality of the feature vector) changes the perspective. For instance, if we move from two-dimensional feature vectors to three-dimensional feature vectors, our decision boundaries become decision planes. Increasing the dimensionality even further leads to decision hyperplanes. Our imaginatory skills for high-dimensional spaces are very limited, and our intuitions are demonstrably false. However, mathematics, in particular linear algebra, allows us to generalize the notion decision boundaries to high dimensions. Perceptrons, multilayer perceptrons, and deep networks typically operate on high-dimensional feature vectors as input. Mathematically, this does not pose any obstacle. The only price to pay is that we cannot easily visualize the functioning of these networks in terms of decision boundaries.

7.6 Convolution: Shifting a Perceptron Over an Image

The early success of deep learning on the ImageNet classification problem relied on a variant of deep learning that involves convolution. The basic idea of using convolution goes back to the seminal work of LeCun more than 30 years ago (LeCun et al., 1989). Convolution is a well-known operation in signal and image processing. As we will see, the convolution operation can be considered as an alternative use of the basic perceptron. However, the interpretation in terms of decision boundaries is no longer applicable, because the associated perceptrons typically have more than two inputs.

7.6.1 The Basic Convolution Operation

Convolution is similar but not identical to cross-correlation. In our current context, we roughly interpret convolution as a kind of “template matching.” In deep learning, given an image, convolution acts as a local template that is much smaller than image. The template typically consists of a small square matrix of weights. For instance, a gray-level image may have dimensions 512×512 , representing the width and height (in pixels), whereas the local template is typically much smaller, i.e., 3×3 weights. Application of convolution to all image locations proceeds as follows. Initially, the template is positioned in the upper left corner of the image. In our example, it would cover the upper-left 3×3 pixels of the image. The convolution operation corresponds to the pixel-wise multiplication of the pixel values with the corresponding weight values. All 3×3 pairwise multiplications are summed to yield a convolution value for the upper left corner of the image. Subsequently, the template is shifted one position to the right and again a convolution value is computed. This procedure is

repeated for the entire row, and then repeated for the next row, up to lower right corner of the image. At that point, convolution values have been computed for the entire image, yielding a *convolution image* that highlights all locations where the local pattern of the image is similar to the template. The described convolution procedure is known as “window sliding” and can also be performed in parallel, which is typically performed in graphical processing units (GPUs).

Convolution in the context of convolutional neural networks that are trained on images consists of a perceptron with a linear transfer function that takes a small sub-image as input and is shifted over the entire image. The weights of the perceptron form the template. We start by considering a gray-valued (single-channel) image G that consists of X columns and Y rows of pixels. We assume that the perceptron takes $T \times T$ sub-images as input. For each $T \times T$ sub-image centered at (x, y) , the perceptron computes a convolution output $C(x, y)$ according to the equation

7

$$C(x,y) = \left(\sum_{i=x}^{x+T-1} \sum_{j=y}^{y+T-1} w_{ij} I_{i,j} \right), \quad (7.10)$$

where T is the linear dimension of the square template and $x \in \{1 : X - T + 1\}$ and $y \in \{1 : Y - T + 1\}$ to ensure that the perceptron only samples valid pixel locations. The convolution image C has almost the same size as the input image ($(X - T + 1) \times (Y - T + 1)$). It is important to note that this equation is almost identical to Eq. (7.1). The only two differences are the absence of the transfer function and the introduction of two-dimensional indices i and j over image I , instead of a one-dimensional index i over feature vector F .

A crucial aspect of convolutional neural networks is that the same perceptron weights are applied to every image location. Depending on the implementation, the convolution by means of a perceptron can be executed by shifting the perceptron from the upper left corner of the image up to the lower right corner of the image (this is what typically happens in sequential algorithms), or it can be executed by invoking a massively parallel perceptron consisting of $(X - T + 1) \times (Y - T + 1)$ perceptrons. In both cases, *weight sharing* is applied, which means that the same weights are applied to each image location.

The weights w_{ij} of the perceptron define the template. Roughly speaking, the weights should be tuned to local visual characteristics that support the overall classification task. For instance, if the convolutional neural network is trained on recognizing animals in images, the perceptron may become tuned to circular black circles in order to detect the nose and eyes of a dog. To illustrate convolution, we consider a convolution perceptron with a 4×4 weight matrix \mathbf{w} , i.e., $T = 4$. The first incorporates a “horizontal filter” and is defined as follows:

$$\mathbf{w}_{horizontal} = \begin{pmatrix} -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 \\ +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & +1 \end{pmatrix} \quad (7.11)$$

The “vertical filter” perceptron is defined as follows:

$$\mathbf{w}_{vertical} = \begin{pmatrix} -1 & -1 & +1 & +1 \\ -1 & -1 & +1 & +1 \\ -1 & -1 & +1 & +1 \\ -1 & -1 & +1 & +1 \end{pmatrix} \quad (7.12)$$

Figure 7.8 illustrates an input image (left) and two convolution images: one that is obtained by convolution with $\mathbf{w}_{horizontal}$ highlights the presence of horizontal contours (middle), and one obtained by convolution with $\mathbf{w}_{vertical}$ highlights the presence of vertical contours.

Convolution is typically applied to inputs that have temporal or spatial neighborhood relations. For instance, in a one-dimensional time series consisting of once-per-minute temperature measurements at the same location, neighboring samples of the time series are related. In general, they are highly correlated because the temperature does not change that much over the time course of 1 min. For two-dimensional inputs, such as images, neighboring pixels sample adjacent locations in space. Also here, in general, you will find highly correlated values for neighboring samples. Convolution can also be applied to three-dimensional data, for instance, MRI data, and even higher dimensional data, but in this chapter, we focus on the application to two-dimensional images that may have a single channel (e.g., gray-valued images) or multiple channels (color images).

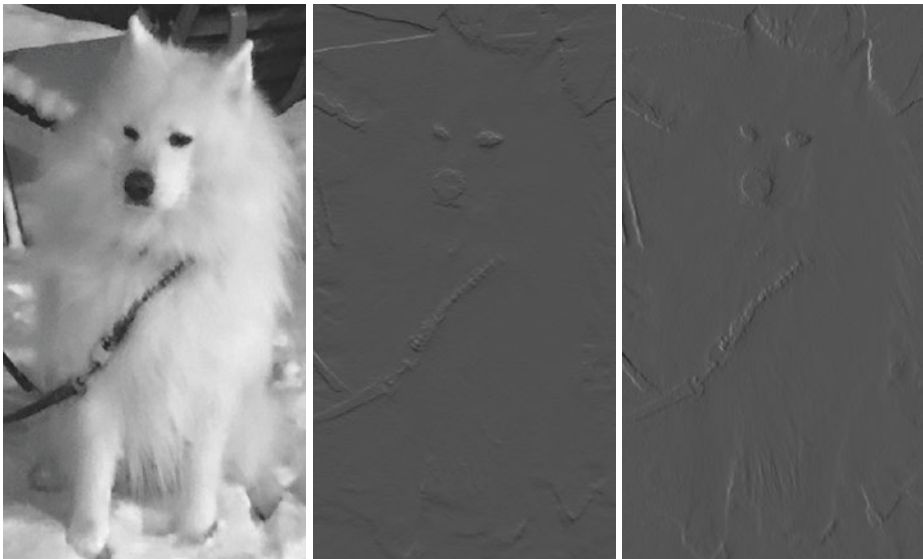


Fig. 7.8 Illustration of convolution applied to an image. Left: Source image of a dog. Middle: Convolution image obtained by convolving the image with a horizontal filter. Right: Convolution image obtained by convolving the image with a vertical filter. (Author’s own)

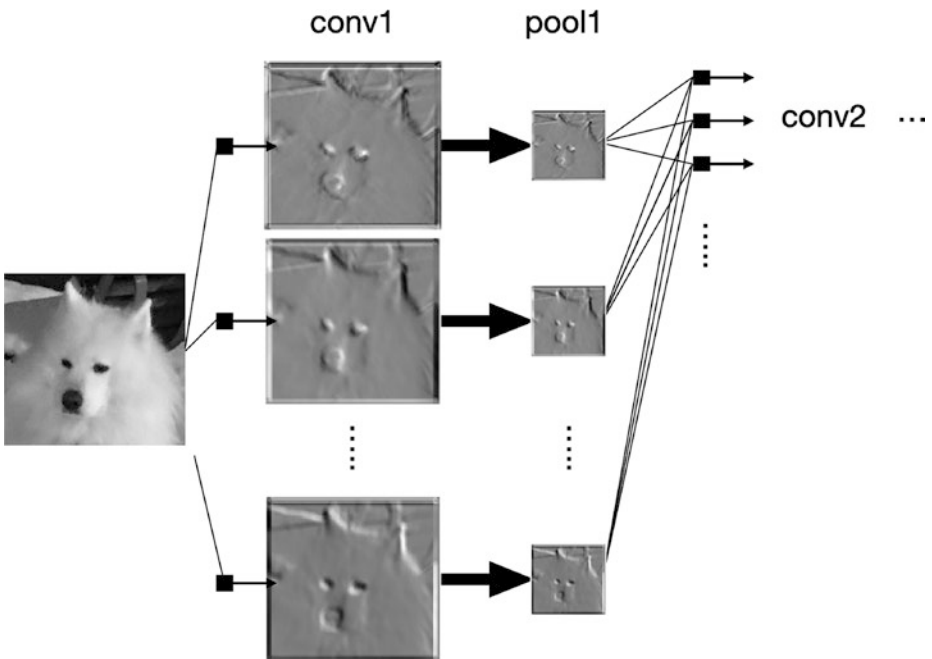
7.7 Convolutional Neural Networks

Convolutional neural networks (CNNs) consist of multiple convolution layers interspersed with pooling layers finally feeding into so-called dense layers, which essentially form a multilayer perceptron. We describe each type of layers in turn.

■ Figure 7.9 provides an illustration of the main building block of CNNs consisting of an input image (shown on the left of the figure), a convolution layer (**conv1**), and a pooling layer (**pool1**).

7.7.1 Convolutional Layers

A source image of $N \times M$ pixels is fed into a convolution layer, which consists of multiple convolution filters, each of which is essentially a perceptron. In ■ Fig. 7.9, the convolution filters are represented by small black squares. The inputs of the perceptron cover a very small sub-image of 2×2 or 7×7 pixels. The same perceptron is applied to all image locations. This can be done in two ways, either by slid-



■ Fig. 7.9 Illustration of the two main building blocks of CNNs. (Author's own)

ing the perceptron across the image, e.g., by starting in the top left and ending in the bottom right, or by having many ($M \times N$) identical perceptrons each applied to a sub-image centered around a pixel. In the latter case, the many perceptrons share their weights.

A convolutional layer consisting of P perceptrons translates an input image into P convolution images. Referring to the example image of a dog, one filter may become tuned to contours in a certain direction, another filter to a detail of the pointy ears, and yet another to the visual texture of the dog's fur. ■ Figure 7.9 shows three examples of such convolution images (the images below **conv1**). Each convolution image highlights different visual features of the input image on the left. The convolution images are submitted to a nonlinear threshold function, such as the ReLU, to suppress small convolution values.

The weights of each of the P perceptrons are tuned using the generalized delta rule. So, the minimization of the loss function determines the nature of the P templates.

7.7.2 Pooling Layers

Pooling layers reduce the dimensionality of the convolution images outputted by the convolution layers by pooling over subregions of the convolution images, for example regions of size 2×2 convolution values, and only propagating the maximum convolution value within the subregion (i.e., max pooling) or the average convolution value (i.e., average pooling). In case of a convolution image of size 100×100 , applying pooling on subregions of size 2×2 results in pooling outputs of size 50×50 . ■ Figure 7.9 shows a pooling layer (**pool1**) that reduces the size of each convolution image in conv1 to a quarter of its size.

7.7.3 Combinations of Combinations of ... Features

The recognition abilities of CNNs are due to the recursive application of convolution and pooling layers. As suggested in ■ Fig. 7.9, the pooled convolution images in **pool1** are convolved by a new set of filters, the column of three black squares on the right. Importantly, this second stage of convolution applies each filter to all pooled convolution images of **pool1**. In this way, the second convolution layer (**conv2**) develops convolution images that represent combinations of features. Adding another convolution and pooling stage gives rise to convolution images that represent combinations of combinations of features. The parallel with combinations of decision boundaries is not entirely coincidental. After all, visual features are just another view on what happens in a deep neural network.

7.7.4 Dense Layers

Dense layers, sometimes referred to as fully connected layers, are like the input to hidden layers of the multilayer perceptron. “Dense” refers to the density of connections. All units of the source layer are connected with adaptive weights to the target layer. It is important to note that a convolution layer can be considered to be a “sparse” layer, given the limited number of adaptive weights due to weight sharing. Dense layers are typically found at the top of convolutional neural networks after several sequences of convolution and pooling. The final dense layer connects the last hidden layer to the output layer.

7.7.5 From AlexNet to Modern CNNs

7

AlexNet was the first convolutional neural network that achieved a breakthrough performance on the ImageNet classification task (Krizhevsky et al., 2012). The ImageNet task requires the automatic classification of natural images into one of 1000 classes. An important innovation of AlexNet was the replacement of the sigmoid function of Eq. (7.8) by the ReLU function of Eq. (7.9). The introduction of the ReLU function improved the stability and efficiency of training in deep networks. The development of convolutional neural networks after AlexNet led to a range of innovations that enhanced the accuracy of CNNs, often by reducing the number of parameters in smart ways or by creating very deep networks (see, e.g., Simonyan and Zisserman (2015)). The interested reader is referred to Rawat and Wang (2017) for a comprehensive review of recent convolutional architectures.

7.8 Skin Cancer Diagnosis: A CNN Application

In this section, we turn to a specific application domain task, namely skin cancer diagnosis, to provide an illustration of the way CNNs are applied in practice.


7.8.1 Introduction

Most state-of-the-art vision applications use CNN-like deep learning networks. Today’s performance of CNNs for specific object recognition tasks is on par or even better than human perception (Brinker et al., 2019). With CNNs, we can count whales from satellite images and unlock our smartphones, and it is a major component in self-driving cars. In this final section, we illustrate how CNN tech-

nology can be used in the medical domain, in particular to detect skin cancer. For this, we used the public dataset of the ISIC 2017 challenge (Codella et al., 2018). A more extended version of this case study, which is carried out by the TU/e and Fontys ICT, in cooperation with UMC Utrecht, can be found in the work of Raumanns et al. (2020).

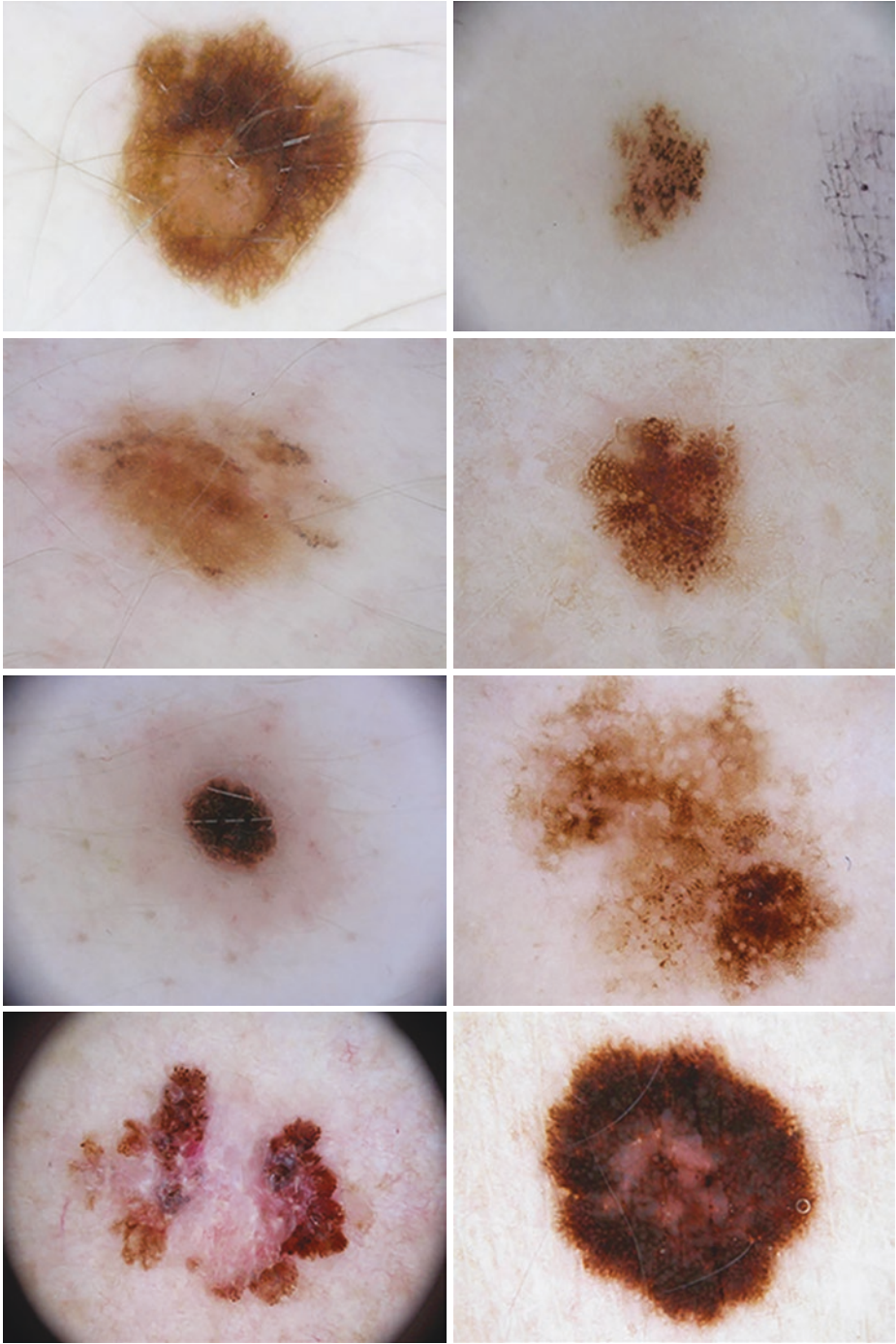
Supervised learning algorithms need labeled data. For instance, in the case of X-ray, CT, or MRI scans, a radiologist has to establish a ground-truth diagnosis for thousands of images. With these labeled examples, a model is trained that can be used to diagnose reliably new unseen scans. Labeling is usually done by experts, a laborious and costly process. In this case study, we explore the question: Can information given by the crowd significantly improve the diagnosis of skin cancer by means of a CNN? Such an alternative removes an important roadblock and might accelerate the adoption of fast and well-grounded computer-aided diagnosis with deep learning in health care. More specifically, we compare the performance of a baseline CNN model with a multitask CNN model. In the baseline model, we modify the weights in such a way that the diagnosis is predicted well. In the multitask model, we optimize the network weights for both diagnosis and additional information obtained from crowd annotations.

7.8.2 Data Collection and Preparation

The data is extracted from the ISIC 2017 challenge (Codella et al., 2018). An advantage of public datasets is that they are usually labeled. The ISIC 2017 challenge contains 2000 skin lesion images. The labeled classes are melanoma (374 lesions), seborrheic keratosis (254 lesions), and nevus (1372 lesions). We combine the images with melanoma and seborrheic keratosis and label them as *malignant*, and the nevus images are labeled *benign*. Note that the dataset is still somewhat unbalanced; for each malignant skin lesion, there are about two benign lesions in the dataset. Some examples are shown in  Fig. 7.10.

The labeled data is enriched by so-called ABC features that were collected from the crowd, in our case undergraduate students. The students assessed the ABC features (Abbasi et al., 2004) as used by dermatologists: A for asymmetrical shape, B for border irregularity, and C for color of the assessed lesion. We followed a similar protocol as in the work of Cheplygina and Pluim (2018); that is, the students scored the strength of each feature on an ordinal scale. All images are rescaled to (384,384) pixels to match the input requirements for both the baseline CNN and multitask CNN.

7



■ Fig. 7.10 Examples of benign (upper row) and malignant (lower row) skin lesions from the ISIC 2017 challenge dataset. (Author's own)

7.8.3 Baseline and Multitask CNN

We apply both a baseline CNN and a multitask CNN to the dataset, as shown in **Fig. 7.11**. The baseline CNN outputs a binary classification (*benign* or *malignant*). The goal of multitask learning is to perform multiple related tasks at the same time. In multitask models, more than one objective is optimized. The idea behind multitask learning is that multiple tasks reinforce each other, i.e., one task helps in learning another task. In this way, better performance can be achieved; see, e.g., Murthy et al. (2017). Both the baseline CNN and the multitask CNN are built on the same pre-trained encoder, i.e., transfer learning is applied. As encoder we use the VGG16 (Simonyan & Zisserman, 2015) convolutional and pooling layers with ImageNet weights. The baseline model is extended by adding two fully connected layers to implement a single classification head; the multitask model is extended by adding three fully connected layers to implement the classification head as well as the annotation head (for asymmetry, border, and color). Note that in this setup, the ABC features are not used as predictors but as outcomes. Only the weights of the additional fully connected layers are trained with the prepared skin lesion images. A cross-entropy loss function is used for the classification task, and a mean square-less loss is used for the annotation regression task.³ For the multitask CNN, both loss values are summed and minimized during training. We implemented both CNN models in Keras using the TensorFlow backend (Geron, 2019).

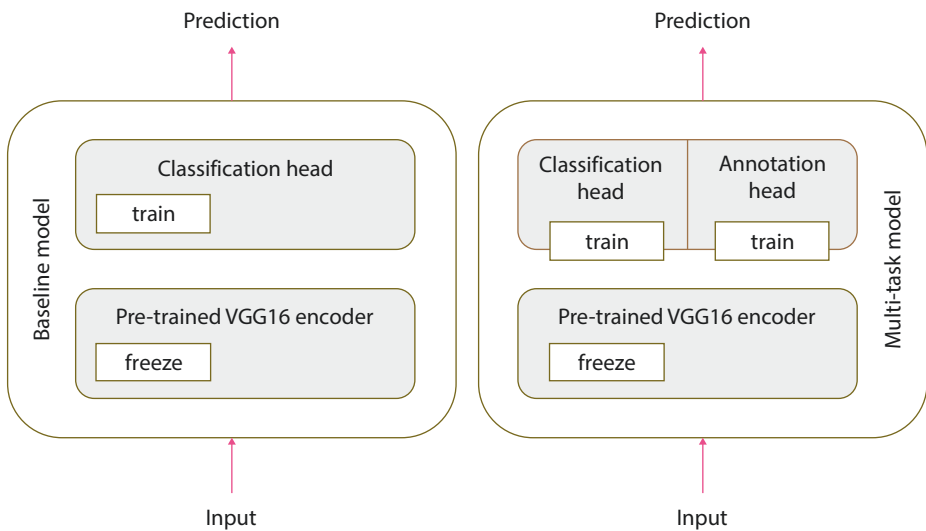



Fig. 7.11 Architecture of baseline CNN (left) and multitask CNN (right). Both models are built on top of the VGG16 encoder that is pre-trained with ImageNet weights. (Author's own)

³ For a tutorial on how neural networks can be used for a regression task, see ► <https://towardsdatascience.com/deep-neural-networks-for-regression-problems-81321897ca33>.

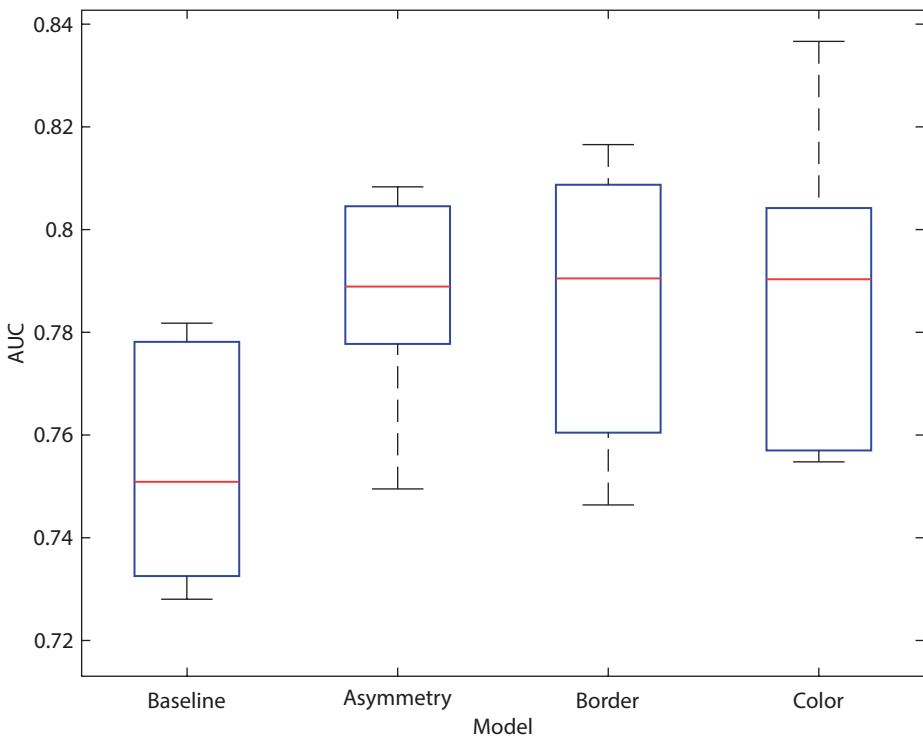
7.8.4 Experiments and Results


We compared four experimental settings:

1. Baseline model with binary classification
2. Multitask model with the asymmetry feature
3. Multitask model with the border feature
4. Multitask model with the color feature

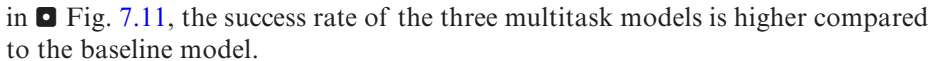
We apply fivefold cross-validation, with the dataset split in a stratified fashion, keeping the malignant-benign ratio equal over the training, validation, and test subsets. More specifically, 70% of the dataset is used as the train subset (1400 lesions) and 17.5% as the validation subset (350 lesions), leaving 12.5% as the test subset (250 lesions). We trained both CNNs iterating over 30 epochs with a learning rate of 0.00002. Since the dataset is unbalanced, we use the area under the ROC curve (AUC) as evaluation metric.⁴ The AUC score of each fold is logged; the results are summarized in the boxplot of  Fig. 7.12. The average AUC is

7



 **Fig. 7.12** AUC scores of the four experiments: baseline model, multitask with asymmetry, multitask with border, and multitask with color. (Author's own)

⁴ For an explanation of AUC, see ► <https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/>.

0.75 for the baseline model and around 0.79 for multitask models. As can be seen in  Fig. 7.11, the success rate of the three multitask models is higher compared to the baseline model.

7.8.5 Conclusion on the CNN Application

Many of today's CNN applications are single task. A lot of effort is put to reach marginal performance improvements, often in the order of subdecimal percentages. In this case study, we demonstrate that a multitask approach leads to a substantial improvement for skin cancer detection (about 4% in AUC). Nonexperts provided the labels for the additional ABC task. The popularity of Mechanical Turk crowdsourcing marketplace shows that there is a need for easy-to-use platforms to collect cheap human labels. Whether and how crowd judgements can be employed to replace expert judgements depends of course on the criticality of the application. This is still an active and ongoing field of research.

Conclusion

In the first part of this chapter, we presented an intuitive explanation of deep learning. In the last section, we gave a single concrete example of the way CNNs are being applied. It may be clear that the range of possible applications is very large. We have used CNNs in such diverse tasks as the analysis of artworks (van Noord 2015; 2017) or the automatic detection of plastic waste (van Lieshout et al., 2020). We hope to have inspired the reader to delve into the theory (Goodfellow et al., 2016) and practice (Geron, 2019) of deep learning to apply it to novel innovative applications.

Take-Home Messages

Deep neural networks essentially consist of stacked perceptrons. Each perceptron incorporates a decision boundary. Convolution filters can be considered to act as perceptrons that receive inputs from small image regions. The modeling power of deep neural networks and convolutional neural networks arises from their nested use of decision boundaries.

? Questions

1. What would happen if the transfer functions in deep neural networks are linear functions?
2. Why is a step function not acceptable as a transfer function in deep neural networks?
3. What is the vanishing gradient problem?
4. What is multitask learning?

✓ Answers

1. Compositions of linear transformations reduce to a single linear transformation. Hence, it does not make sense to have deeper networks than a single-layered one.

2. The discontinuity of the step function does not allow for the computation of gradient. Hence, backpropagation is not applicable.
3. The vanishing gradient problem arises if the propagated activations and errors become too small, due to the small gradients or near-zero weight updates.
4. Multitask learning is a subfield of machine learning in which multiple learning tasks are solved at the same time while exploiting commonalities and differences across tasks. This can result in improved learning efficiency and prediction accuracy, when compared to training the models separately.

References

- 7
- Abbasi, N. R., Shaw, H. M., Rigel, D. S., Friedman, R. J., McCarthy, W. H., Osman, I., Kopf, A. W., & Polsky, D. (2004). Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria. *JAMA*, 292(22), 2771–2776.
- Brinker, T. J., et al. (2019). Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer*, 119, 11–17.
- Cheplygina, V., & Pluim, J. P. W. (2018). Crowd disagreement about medical images is informative. In M. J. Cardoso, T. Arbel, S.-L. Lee, V. Cheplygina, S. Balocco, et al. (Eds.), *Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis* (pp. 105–111). Springer International Publishing.
- Codella, N. C. F., et al. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (pp. 168–172).
- Geron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. Retrieved from <http://www.deeplearningbook.org>
- Jarrett, K., et al. (2009). What is the best multi-stage architecture for object recognition? (pp. 2146–2153).
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In: *Neural Information Processing Systems* (p. 25).
- LeCun, Y., et al. (1989). Back-propagation applied to handwritten zip code recognition. *Neural Computation*, 1(4), 541–551.
- Minsky, M., & Papert, S. (1969). *Perceptrons*. MIT Press.
- Murthy, V. et al. (2017). Center-focusing multitask CNN with injected features for classification of glioma nuclear images (pp. 834–841).
- Raumanns, R. et al. (2020). Multi-task learning with crowdsourced features improves skin lesion diagnosis. *arXiv*.
- Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9), 2352–2449.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). MIT Press.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *CoRR*. Retrieved from <https://arxiv.org/abs/1409.1556>.

- van Lieshout, C., et al. (2020). Automated river plastic monitoring using deep learning and cameras. *Earth and Space*. <https://doi.org/10.1029/2019ea000960>
- van Noord, N., Hendriks, Ella, & Postma, E. (2015). Toward discovery of the artist's style: learning to recognize artists by their artworks. *IEEE Signal Processing Magazine*, 32(4), 46–54.
- van Noord, N., & Postma, E. (2017). Learning scale-variant and scale-invariant features for deep image classification. *Pattern Recognition*, 61, 583–592.



Sequential Experimentation and Learning

*Jules Kruijswijk, Robin van Emden
and Maurits Kaptein*

Contents

- 8.1 Introduction – 149**
- 8.2 The Multi-Armed Bandit Problem – 151**
- 8.3 Solutions to Bandit Problems: Allocation Policies – 153**
 - 8.3.1 ϵ -First – 153
 - 8.3.2 ϵ -Greedy – 155
 - 8.3.3 Upper Confidence Bound Methods – 155
 - 8.3.4 Thompson Sampling – 157
 - 8.3.5 Bootstrapped Thompson Sampling – 157
 - 8.3.6 Policies for the Contextual MAB Problem – 158
- 8.4 Evaluating Contextual Bandit Policies: The Contextual Package – 159**
 - 8.4.1 Formalization of the cMAB Problem for Its Use in Contextual – 159

8.4.2	Class Diagram and Structure – 160
8.4.3	Context-Free Versus Contextual Policies – 161
8.4.4	Offline Policy Evaluation with Unbalanced Logging Data – 163
8.5	Experimenting with Bandit Policies: StreamingBandit – 166
8.5.1	Basic Example – 167
8.5.2	StreamingBandit in Action – 170
	References – 173

Learning Objectives

After reading this chapter, you should be able to:

- Recognize sequential learning problems as (contextual) multi-armed bandit problems and understand the core challenge involved: in a cMAB problem, the objective is to optimally balance exploration and exploitation.
- Implement and reason about various effective bandit policies such as ϵ -greedy, UCB, and Thompson sampling.
- Use the contextual package to run simulations of bandit policies and to conduct effective, unbiased, offline policy evaluation given a logged dataset.
- Understand the perils involved in unbalanced logging data: you should be able to use Simpson's paradox to illustrate the use of propensity score weights.
- Use streaming bandit package to deploy bandit policies in the wild.

8.1 Introduction

It is fairly common in many business contexts to wonder what the outcomes will be of a change to our course of action. For example, a manager of an e-commerce store might wonder whether the website performs better when the UI is changed or when the underlying recommender system is updated. Alternatively, a medical practitioner might wonder whether some novel treatment or behavioral intervention will be effective for the current patient. In each of these cases, it is tempting to turn to data to answer these prospective questions: Can the rich data already in our possession, potentially by using one of the various data science methods introduced in this book, inform our future actions?

In this chapter, we will argue that surprisingly often the answer to this question is “no.” Despite years of data collection, web companies and hospitals alike will often encounter the problem that their potential terabytes of data are ill-suited to inform a possible change of action. Why? Well, simply because the collected data contains no instances of the envisioned new future. The data of the e-commerce store simply does not contain any records detailing the purchase behavior of customers when faced with the new UI or recommender system, nor does the electronic patient record inform us about “what would happen if we do something else” simply because this alternative choice of action was never executed. Thus, unless we are willing to resort to generalizations from other contexts in which useful data might be available (in which case the validity of the generalization is always a potential problem), we simply have no data to rely on.

When no data is present, it is paramount to collect new data that allows us to learn and make an informed future decision. Simply put, if we want to know how effective our new UI or our new treatment is, we will have to resort to trying it out. We can set up a study, experiment, or whatever name we would like to give the period in which we try out our new course of action on a subset of customers or patients. Subsequently, once enough data is available, we might indeed turn to some of the methods discussed elsewhere in this book to make an informed deci-

sion regarding the potential change in our course of action. However, this approach immediately raises questions regarding the setup and length of this experimental period: How many customers do we need to confront with the new UI before we can confidently change our course of action? Which patient, and how many of them, should receive the new treatment before we can safely deduce that the new treatment is better than the old?

Taking a step back, the crux of the problem is, informally, easy to state: on the one hand, we would like to have an experimental period that is sufficiently long to make a well-informed decision. Thus, we want to sufficiently *explore* our new course of action. On the other hand though, we would like to make sure that we make the best choice as often as possible: we want to *exploit* our knowledge and not expose too many patients or customers to our suboptimal choices (Sutton & Barto, 2011; Lattimore & Szepesvári, 2018). Assuming that we are in the position to learn sequentially,¹ e.g., to try out a course of action and see its result on one customer or patient before moving to the next, we can rephrase our questions regarding the length of our experiments to a more general one in which we wonder what decision strategy—or *policy*—balances trying out new actions with using the action we think is best in such a way that over a finite number of interactions we attain the best outcome (Eckles & Kaptein, 2014). Clearly, exploring with *all* our interactions is suboptimal as we will subscribe the suboptimal action to a fixed proportion of customers (or patients). Such a policy is said to over-explore. Alternatively, exploring *very little* and quickly choosing the action that we believe is most successful risk choosing an action that seemed optimal in the collected data but in reality is not: such a policy is said to over-exploit.

Phrased in terms of exploration and exploitation, it is clear that the traditional randomized experiment (or A/B test) leading to a subsequent choice of the best action is just one out of many potential policies: in the traditional experiment, we first try out how effective our actions are by choosing actions with probability $1/K$ where K is the total number of future actions under consideration (i.e., we explore) and subsequently choose the action that performs best with probability 1 for the remaining customers or patients (i.e., we exploit). There is however, when the problem is approached as a sequential decision problem, nothing stopping us from more smoothly changing these probabilities as we interact sequentially. We can subsequently wonder which allocation policy is most effective (Perchet et al., 2013). In this chapter, we will, after introducing the above problem more formally, detail several policies that have appealing theoretical properties and/or are useful in practice. Next, we will discuss several software tools developed by the Computational Personalization Lab at JADS that researchers and practitioners can use to tackle the type of sequential decision problem described above. We close off with a short discussion of the applied merits of this sequential experimentation approach embraced in this chapter.

¹ Many of the approaches we describe in this chapter are easily generated to a so-called *batch* setting in which not all observations manifest one by one. However, we will focus primarily on the fully sequential setup.

So, our main aim in this chapter is to introduce several policies that are useful to tackle sequential learning problems, which we feel are commonplace in many situations in which decision makers aim to change their future course of action. In essence, our topic of study is the performance of different allocation policies (i.e., the strategy by which a decision maker should, at each point in time, decide between trying the new course of action and using one that he or she feels works best in the given context).

8.2 The Multi-Armed Bandit Problem

Imagine facing a row of slot machines. You have been told that one of the machines pays out more often than the others, but you are unsure which one has the highest payout probability. You have a fixed set of coins that allow you to play the machines, and your goal is to earn as much as possible. Assuming that you cannot reinvest your earnings, how do you go about playing the different machines using your fixed budget such that you earn as much as possible?

The situation described above provides a simplified version of the type of decision problem introduced in the previous section: you are faced with a number of future actions (i.e., which machine to play, which UI to choose, or which treatment to administer), and sequentially (coin by coin, customer by customer, or patient by patient) you can try a course of action and see the result (a win, a click, or a healthy patient). This problem is known as the *multi-armed bandit problem* (MAB) based on the colloquial term “one-armed bandit” to relate a single slot machine.² The multi-armed bandit problem provides an abstract representation of a sequential decision that is relatively simple (e.g., we are not yet considering differences between customers and patients) but still sufficiently challenging to provide fruitful analysis. The study of effective policies—strategies that determine into which machine to put the next coin—is extremely large and has given us numerous insights into the nature of this challenging and omnipresent decision problem.

Multi-armed bandit problems can be formalized as follows. At each time $t = 1, \dots, T$, we have a set of possible actions (i.e., arms, treatments) \mathcal{A} . After choosing $a_t \in \mathcal{A}$, we observe reward $r_t(a_t)$. The aim is to select actions so as to maximize the cumulative reward $\mathcal{R}_c = \sum_{t=1}^T \gamma_t r_t$, where γ_t is a discount rate. Note that here we often assume $r(a)$ to be *i. d.* $\Pi(a)$ and further note that we assume that at time t only $r_t(a_t)$ is revealed; the rewards of alternative actions at that point in time are not observed.³ We aim to examine the performance of various allocation policies π , which are mappings from the historical data up to time $t-1$, $\mathcal{D}_{t-1} = (a_1, r_1, \dots, a_{t-1}, r_{t-1})$ to the next action a_t . The MAB problem, with various assumptions regarding properties of the actions and (distribution of) rewards, as well as with various dis-

2 The term one-armed bandit is in turn set to be derived from the idea that slot machines, *bandits*, are “as efficient in parting a human from his money as their human counterpart.”

3 Note on the potential outcome notation/fundamental problem of causal inference.

count schemes, is extensively studied (see Eckles & Kaptein, 2014, for additional references): we will discuss a number of effective policies and their properties in the next section.

An often-used extension to the standard MAB problem is the *contextual MAB* (cMAB) problem (Perchet et al., 2013): in this case, before choosing an action, the “state of the world” (or *context*) x_t is revealed prior to choosing an action. It is assumed that in this case, the (stochastic) rewards are a function of both the action and the context (i.e., $r_t(a_t, x_t)$). This fruitful extension of the problem formalization maps to many important real-world problems: What if prior to choosing the new or old UI of the e-commerce store, we first observe a number of properties of the customer (the context)? Or, what if prior to choosing a treatment, we observe the state of the patient?


Bandit problems were initially studied in the 1940s providing a useful model for choosing amongst competing medical treatments (Berry & Fristedt, 1985). However, the (non-contextual) bandit problem formalization is quite distant from the actual problems faced, e.g., in medicine, wherein the rewards of the action often do not immediately manifest, actions might have very different cost structures, and ignoring the context seems to be a gross oversimplification of the problem. After a period of reduced interest in the problem, its popularity surged again with the advent of online advertising: the theoretical (contextual) bandit problem is, in this case, very close to the actual application. Currently, most of the larger web companies use “bandit approaches” to select the contents they display to customers.

The cMAB problem is a special version of a broader type of decision problems that is studied under the heading of reinforcement learning. The standard reinforcement learning framework assumes that each action taken by the policy changes the state of the world (potentially with some probability), after which a reward is associated with the resulting state. The framework of reinforcement learning, which is introduced in detail by Sutton and Barto (2011), contains challenges that are similar to the MAB problem. First of all, the learner needs to learn how actions (and associated state changes) relate to the observed rewards, and second the learner needs to balance trying out new actions to learn more with using the actions he or she believes are most successful to maximize his or her cumulative rewards. On top of this, the RL formalization adds the challenging problem of reward attribution: in an RL setting, often multiple actions lead to a penultimate outcome which has (large) positive rewards. This high ultimate reward should be allocated to previous actions in a meaningful way (for example, when using an RL framework to teach a computer how to play chess (Sutton & Barto, 2011), it is paramount to allocate rewards to the initial moves of an eventually won game). Furthermore, the RL literature has focused more on situations in which data collection is relatively “cheap” (i.e., the process can be repeated time and time again), whereas the bandit literature has focused on situations in which there is only one sequence of actions to learn and exploit. Despite these differences in emphasis, bandit problems are simply a subset of RL problems (Langford & Zhang, 2008). In the remainder of this chapter, we focus on cMAB problems; a number of intuitions developed, however, hold for more general RL problems.

8.3 Solutions to Bandit Problems: Allocation Policies

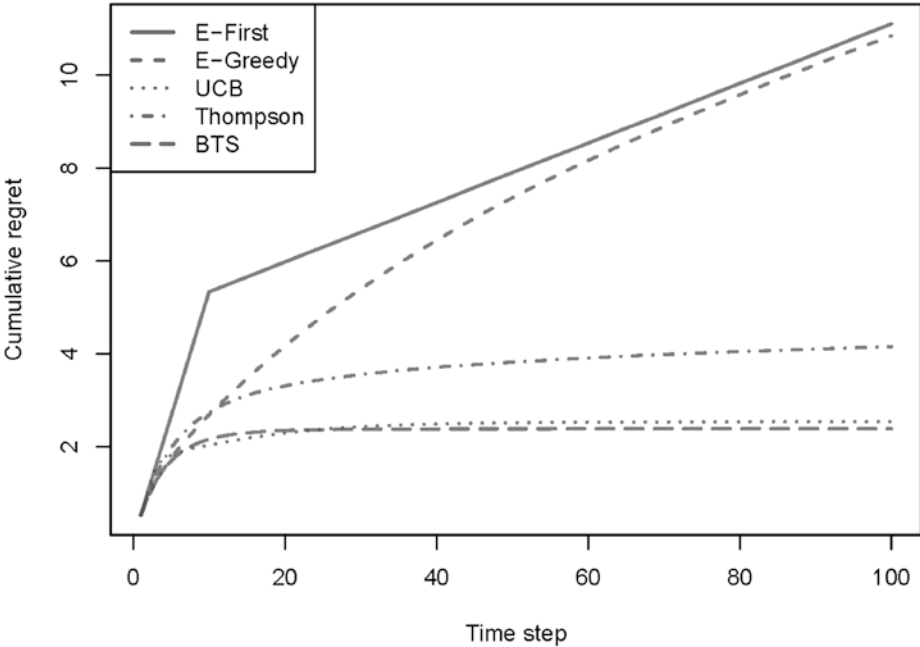
Allocation policies (e.g., rule sets that map medicines to patients or pages to customers) can formally be defined as a mapping from all historical data \mathcal{D}_{t-1} (all data until time point $t - 1$) and possibly a context x_t to a new action $a_t : \pi(x_t, \mathcal{D}_{t-1}) \rightarrow a_t$. It is often the aim to choose the policy such that it maximizes the cumulative reward $\mathcal{R}_c = \sum_{t=1}^T \gamma_t r_t$. Alternatively, instead of assessing the performance of a policy using the cumulative reward, we can also use the expected regret. The expected regret is the sum of the differences in reward between the most optimal policy (i.e., the policy that always plays the action with the highest expected reward according to some oracle) and the allocation policy that is being assessed. More formally, the expected cumulative regret is defined as $\mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\sum_{t=1}^T r_t^* - r_t\right]$ with r_t^* is the reward of the optimal action (i.e., the action with the highest expected reward) and where the expectation is over the stochastic nature of the reward distribution and (possibly) the stochastic nature of the policy itself.

Using the expected regret allows us to investigate the behavior of allocation policies and derive so-called regret bounds that describe how the regret of an allocation policy behaves in the long run (i.e., as T is not fixed). Some policies have been shown to have asymptotically optimal regret bounds, meaning that their regret is a certain constant factor lower than the regret of the best possible policy. Intuitively, an allocation policy that (perhaps by chance) always plays the action with the highest expected reward will incur zero regret (as r^* and r are then equal). Furthermore, a policy that continually has some degree of random exploration will incur linear regret as it has a constant, nonzero, probability of choosing the wrong action. Because of these properties, in the literature, the usage of regret is preferred over cumulative reward, especially when we talk about simulated environments. Duly note, however, that we are not able to compute the regret in every scenario. For instance, in field experiments, we often have no prior information on what the most optimal policy should look like, and we thus have to resort to using cumulative reward. We will however use regret in the remainder of this section when evaluating different allocation policies because of its ease of interpretation.

In  Fig. 8.1, we show the performance of each policy that we will now discuss in the coming sections. These simulations were done using a three-armed bandit with dichotomous rewards distributed according to a Bernoulli distribution. The actions had the following probabilities of returning a reward of 1: [0.9, 0.1, 0.1]. This means that the first action is by far the best action to play. We used $T = 100$ and simulated this for 10^5 times. We will now discuss each represented policy in turn (see also Dudík et al., 2011).

8.3.1 ϵ -First

As a first example, we discuss ϵ -first. In this policy, the experimenter starts with a random exploration of all possible actions for a specified ϵ interactions—which we



■ **Fig. 8.1** The expected cumulative regret of five different policies on a three-armed Bernoulli bandit. (Author’s own figure)


call the exploration phase. In this exploration phase, the actions are sampled uniformly at random. After these interactions have taken place, we select the action with the highest expected reward for the remainder of the $T \epsilon$ interactions—which we call the exploitation phase. Formally, this can be defined as follows:

$$\Pi(\mathcal{D}_{t-1}) := a_t = \begin{cases} \text{Rand}(a \in \mathcal{A}) & \text{if } t < \epsilon \\ \text{argmax}_a \left(\hat{\theta}^a \right) & \text{otherwise} \end{cases} \quad (8.1)$$

where the first part of the equation is the exploration phase: as long as $t \leq \epsilon$, we choose a random action. When $t > \epsilon$, we pick the action for which the expected reward is the highest—in this case, denoted by $\hat{\theta}^a$ which can for example be a mean.

This policy is the same as doing an A/B test or randomized controlled trial (RCT). In ■ Fig. 8.1, we can see how ϵ -first behaves. In the exploration phase, the policy incurs linear regret as it is just randomly selecting actions. Then, in the exploitation phase, it will select the expected optimal arm based on N interactions ($N = 10$ in the simulation above). In some of the 10^5 simulated cases, it will select a suboptimal action (and incur maximum regret) and sometimes select the optimal action (and incur zero regret). Averaging over all these simulations, we end up with an average regret that is in between of those two. We can also see that it has the highest regret of all the policies discussed here.

8.3.2 ϵ -Greedy

Instead of having a separate exploration and exploitation phase, we can just explore other actions in between exploiting our perceived best action with some small probability. In the cMAB literature, this is called ϵ -greedy: for each interaction with probability ϵ , we randomly select an action, and with probability $1 - \epsilon$, we select the action we believe that is most optimal. A typical setting is $\epsilon = 0.1$. As we are always doing some random exploration in between and never stop to do that, we will incur linear regret as t increases beyond 100—as can be seen in  Fig. 8.1. Formally, ϵ -greedy can be defined as follows:

$$\Pi(\mathcal{D}_{t-1}) := a_t = \begin{cases} \text{Rand}(a \in \mathcal{A}) & \text{if } u < \epsilon \\ \operatorname{argmax}_a \left(\hat{\theta}^a \right) & \text{otherwise} \end{cases} \quad (8.2)$$

where u is a draw from $\text{Uniform}(0, 1)$. So, if $u < \epsilon$, we will select a random action; otherwise, we pick the action with the highest $\hat{\theta}^a$ again.

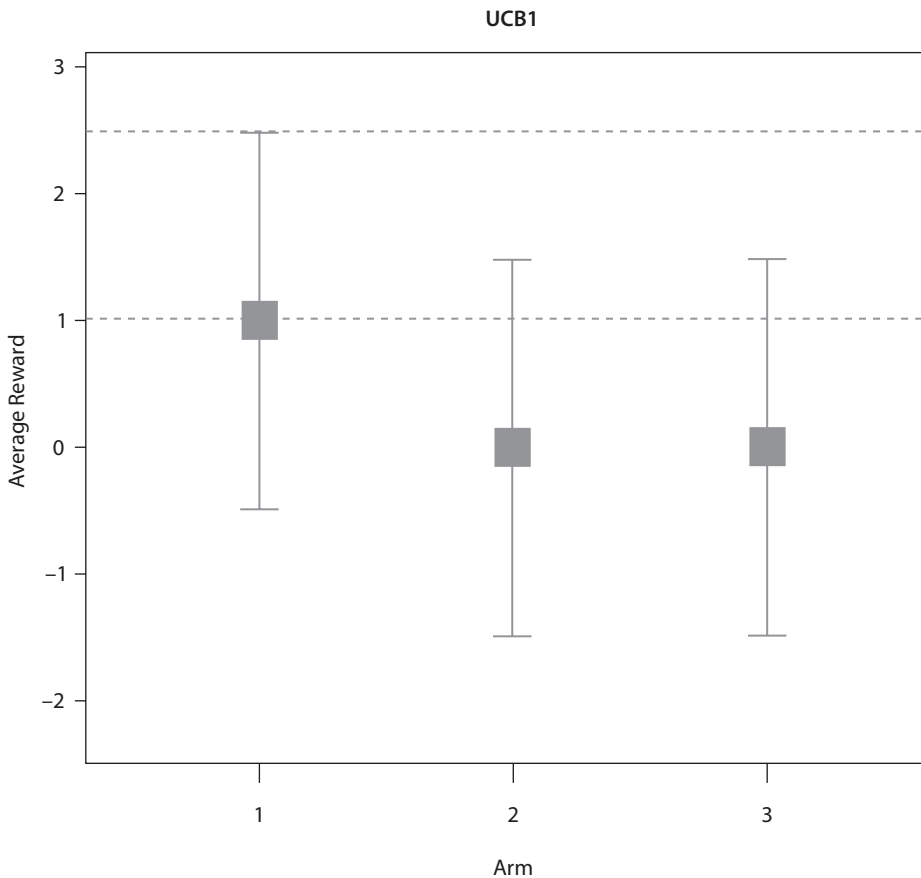
8.3.3 Upper Confidence Bound Methods

We can recognize that repeatedly doing some form of random exploration will not maximize the utility of the knowledge that we get when exploring. Rather than doing random exploration, we can explore with more intent. One way of doing this is taking into account the uncertainty we have about the expected reward of the different actions. As an example, let us say we have a bandit experiment with three different actions. For two of those actions, we have played them a few times and are quite certain about their expected reward. We have not played the third action that much because mainly the expected reward is low. However, as we have not played it often, we are not confident about our estimate on the expected reward. Therefore, it is in our interest to play this action to gather more confidence about the expected reward. One way of addressing this is using upper confidence bound (UCB) methods. In UCB methods, an upper bound for the confidence interval of our actions is computed and the policy selects the action that maximizes the sum of the expected reward and the confidence bound (Auer & Ortner, 2010). Translating to our example, this means that if an action has a low expected reward but a very high confidence bound—because it has not been played often—UCB will play this action if the sum of the two parts totals over all of the other actions. These bounds can be computed in different ways under different assumptions and there is a whole body of literature that derives new policies. Formally, a general formulation for UCB can be defined as follows:

$$\Pi(\mathcal{D}_{t-1}) := a_t = \begin{cases} a & \text{if } n^a = 0 \\ \operatorname{argmax}_a \left(\hat{\theta}^a + \text{CB} \right) & \text{otherwise} \end{cases} \quad (8.3)$$

where we first have to play each action of the complete action set once (first line). After this, we take the action with the maximum of sum of the highest expected reward ($\hat{\theta}^a$) and the confidence bound (CB). An example of a confidence bound could be $\sqrt{2 \log(t) / n_t^a}$ —this confidence bound decreases with the number of interactions that the action has been played. ■ Figure 8.2 illustrates how the confidence bounds and the means of each arm relate to each other.

UCB methods have been shown to be asymptotically optimal—when deployed under the right assumptions—which means that we always select the best action as t grows large. ■ Figure 8.1 shows that UCB has no trouble learning the three-armed bandit setting and has zero regret after only a few interactions.



■ **Fig. 8.2** An example of the confidence bounds in the UCB for the three different arms of the experiment. We see that the first arm has the highest sum of the confidence bound and the mean. This means that the policy will select this arm in the next iteration. (Author's own figure)


8.3.4 Thompson Sampling

Another way of dealing with uncertainty is by using the Bayesian point of view for estimates of the expected rewards. In Bayesian statistics, the estimated expected reward would be regarded as a probability distribution given the evidence that we have gathered so far. More specifically, Bayesian methods would use posterior probability distributions to give not just a point estimate of the expected rewards, but a range of most probable expected rewards given the data at hand. As more evidence is gathered during an experiment, the posterior probability distribution is updated and centers more around the true value of the reward. Furthermore, if there is a priori knowledge about the experiment or data, the Bayesian paradigm allows to tune the posterior distribution to account for that information via the prior distribution (Chapelle & Li, 2011).

Thompson sampling is a policy that uses the Bayesian paradigm and models rewards using an appropriate distribution. The policy was first described by Thompson (1933). For example, in the case of dichotomous rewards, the policy typically is modelled using a beta-Bernoulli distribution. Then for each interaction, the policy obtains a single draw from the posterior distribution for each action (i.e., for each action, a distinct distribution with its own set of parameters is tracked) and then selects the sample with the highest draw. This automatically ensures that Thompson sampling explores in the beginning and as it gathers more certainty about the expected rewards, it will stop exploring. A Thompson sampling policy modelled with a beta-Bernoulli distribution could formally be described as follows:

$$\Pi(\mathcal{D}_{t-1}) := a_t = \underset{a}{\operatorname{argmax}}(\theta'^a) \quad (8.4)$$

where θ'^a is a single draw from the $\operatorname{Beta}(\alpha^a + R_c^a, \beta^a + n^a - R_c^a)$ posterior for arm a —which is a conjugate distribution of a beta prior with a Bernoulli likelihood function.

In  Fig. 8.1, we can see that Thompson sampling has a slightly higher regret than UCB and BTS. If we would have a priori knowledge, we could use an informative prior distribution and in turn have a lower regret than UCB and BTS.


8.3.5 Bootstrapped Thompson Sampling

Although Thompson sampling is easy to implement when sampling from the posterior distribution is straightforward, there are situations where sampling from the posterior distribution is not feasible. In that case, we would have to resort to approximations using, e.g., Markov chain Monte Carlo (MCMC) sampling methods. The huge drawback of using MCMC sampling in bandits is that it is computationally too inefficient to carry out in a one-by-one (or online) fashion (see, e.g., Michalak et al., 2012). Bootstrapped Thompson sampling tries to solve this problem by replacing the Bayesian posterior distribution by a bootstrap distribution around the point estimates of the expected rewards. In the case of BTS, the double-or-nothing bootstrap is typically used. How this works is as follows: instead of

only computing one point estimate of the expected reward (denoted as $\hat{\theta}$), BTS computes J replicates of the point estimate. This means that we have a set of parameters $\hat{\theta} = \{\hat{\theta}_1, \dots, \hat{\theta}_J\}$. Each time a reward is observed, BTS randomly updates a portion of its bootstrap replicates, such that not all replicates are equal—typically, in expectation, half of the replicates are updated. An action is subsequently selected by randomly sampling one of the bootstrap replicates for each action and selecting the action with the highest $\hat{\theta}$. Formally, BTS differs not that much from Thompson sampling, only that how we sample $\hat{\theta}$ is different:

$$\Pi(\mathcal{D}_{t-1}) := a_t = \underset{a}{\operatorname{argmax}} \left(\hat{\theta}_j^a \right) \quad (8.5)$$

where we take the action with the highest $\hat{\theta}$ from a uniformly randomly sampled bootstrap replicate j .

Next to the fact that BTS solves some of the computational issues involved in Thompson sampling, it is also often more robust to model misspecification. See Eckles and Kaptein (2014, 2019) for more details. In  Fig. 8.1, we see that using BTS with $J = 100$, we have the lowest regret of all policies.

8

8.3.6 Policies for the Contextual MAB Problem

The examples introduced thus far are policies that try to solve the multi-armed bandit problem. However, as discussed before, there are times we deal with scenarios where a context of the environment has a potential influence on the reward. Discussing these policies in detail is out of the scope of this chapter, but there are multiple ways of incorporating the context for potential reward gains. If the environment is not overly complicated—for example if the context only consists of a single variable and the action set exists of two different actions—it can suffice to use different estimates of $\hat{\theta}$ for each feature in the context (i.e., we track the mean of each action for both actions for both levels of the context separately). When more complex environments are considered, such strategies may not suffice, because the number of actions that are considered for each estimate of $\hat{\theta}$ will then be too low or the assumptions on the relation between the context and the action do not hold in practice. The literature considers different approaches to the contextual bandit problem (see, e.g., Zhou, 2015). For example, one of the most popular policies is an UCB-inspired policy called LinUCB, which assumes that the reward of actions is linearly dependent on the context and models it using a set of linear predictors (Li et al., 2010).

We hope that the above sections, and the references therein, have provided readers with sufficient information to understand the utility of the MAB problem formalization and to understand a number of the most popular policies. Next to the generalization of the contextual MAB problem, there are many more variants of the MAB problem which are studied in the literature. See Lattimore and Szepesvári (2018) for a detailed introduction on some of these variants. In the next sections in this chapter, we will discuss various software packages that allow readers to easily experiment with bandit policies both in simulated settings (i.e., to effectively

explore the effectiveness of different policies) and in the field (i.e., to deploy bandit policies in applied problems).

8.4 Evaluating Contextual Bandit Policies: The Contextual Package

As already mentioned above, the recent development of a particularly versatile MAB generalization known as the *contextual* multi-armed bandit (cMAB) problem has invigorated MAB research. cMAB policies differentiate themselves, by definition, from their previously introduced MAB cousins in their ability to make use of side information that reflects the current state of the world—information that can then be mapped onto available options or actions (Langford & Zhang, 2008). cMAB policies have proven to be successful in many different areas: from recommendation engines (Lai & Robbins, 1985) to advertising (Tang et al., 2013) and (personalized) medicine (Katehakis & Derman, 1986; Tewari & Murphy, 2017), healthcare (Rabbi et al., 2015), and portfolio choice (Shen et al., 2015)—inspiring a multitude of new bandit algorithms or policies. However, although cMAB algorithms have found more and more applications, comparisons on both synthetic and, importantly, real-life, large-scale offline datasets (Li et al., 2011) have relatively lagged behind. The R package `contextual` facilitates such offline analysis of various bandit policies (van Emden & Kaptein, 2020). In this section, we introduce `contextual` and use it to illustrate how to carry out unbiased, offline, policy evaluation even when the logging data is not balanced (and hence a naive application of Li’s replay method fails).

8.4.1 Formalization of the cMAB Problem for Its Use in Contextual

As the structure of the class or of the R package stays close to its formal roots, we briefly reintroduce the contextual bandit problem: A bandit B is defined as a set of arms $k \in \{1, \dots, K\}$ where each arm is itself described by some reward function that maps d -dimensional context vector $x_{t,k}$ to some reward $r_{t,k}$ (Auer et al., 2002; Langford & Zhang, 2008; Kruijswijk et al., 2016) for every time step t until horizon T . A policy π seeks to maximize its cumulative reward $\sum_{t=1}^T r_t$ (or minimize its cumulative regret)

by sequentially selecting one of bandit B ’s currently available arms (Bubeck et al., 2012), here defined as taking action a_t in $\mathcal{A}_t \subseteq K$ for $t = \{1, \dots, T\}$.

At each time step t , policy π first observes the current state of the world as related to B , represented by d -dimensional context feature vectors $x_{t,a}$ for $a_t \in \mathcal{A}_t$. Making use of some arm selection strategy, policy π then selects one of the available actions in \mathcal{A}_t . As a result of selecting action a_t , policy π then receives reward $r_{a_t,t}$. With observation $(x_{t,a_t}, a_t, r_{t,a_t})$, the policy can now update its arm selection strategy. This cycle is then repeated T times. That is, for each round $t = \{1, \dots, T\}$:

1. Policy π observes current context feature vectors $x_{t,a}$ for $\forall a \in \mathcal{A}_t$ in bandit B .

2. Based on all $x_{t,a}$ and θ_{t-1} , policy π now selects an action $a_t \in \mathcal{A}_t$.
3. Policy π receives a reward r_{t,a_t,x_t} from bandit B .
4. Policy π updates arm selection strategy parameters θ_t with $(x_{t,a_t}, a_t, r_{t,a_t})$.

Overall, it is policy π 's goal to minimize *cumulative regret* or maximize *cumulative reward* $R_T = \sum_{t=1}^T (r_{t,a_t,x_t})$.

8.4.2 Class Diagram and Structure

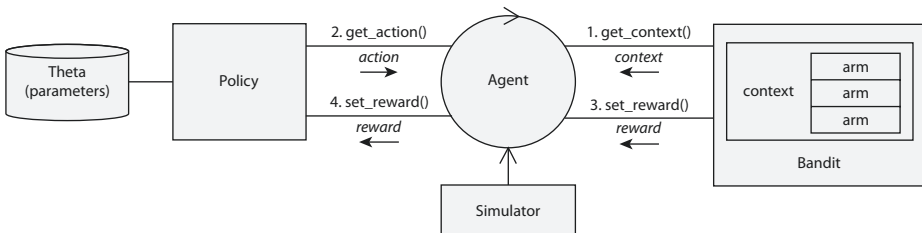
The current section will show that contextual's structure does indeed closely mirror the previous section's formal description of the cMAB problem. In contextual, the bandit and policy superclasses expose, respectively, contextual's reward generation and its decision allocation strategy API. For custom of bandits or policies, the two classes to subclass and extend are the following (■ Fig. 8.3):

8

- **Bandit:** R6 class bandit is the parent class of all bandit subclasses. It exposes k arms and is responsible for the generation of a chosen arm's reward, and, in the case of contextual policy evaluation, current d-dimensional or $k \times d$ -dimensional context.
- **Policy:** R6 class policy is the parent class of all policy subclasses. For each $t = \{1, \dots, T\}$, it has to choose one of a bandit's k arms and update its parameters theta in response to the resulting reward, and, in the case of contextual policy evaluation, the current d-dimensional or $k \times d$ -dimensional context.

The four remaining classes constitute contextual's parallel evaluation, logging, and visualization routines and are generally not subclassed or extended:

- **Agent:** R6 class agent is responsible for the running of one bandit/policy pair. Multiple agents can be run in parallel, where each agent keeps track of t for its assigned policy and bandit pair. To keep agent simulations replicable and comparable, starting seeds are set equal and deterministically for each agent.
- **Simulator:** R6 class simulator is the entry point of any contextual simulation. It encapsulates one or more agents, creates agent clones (each with its own deter-



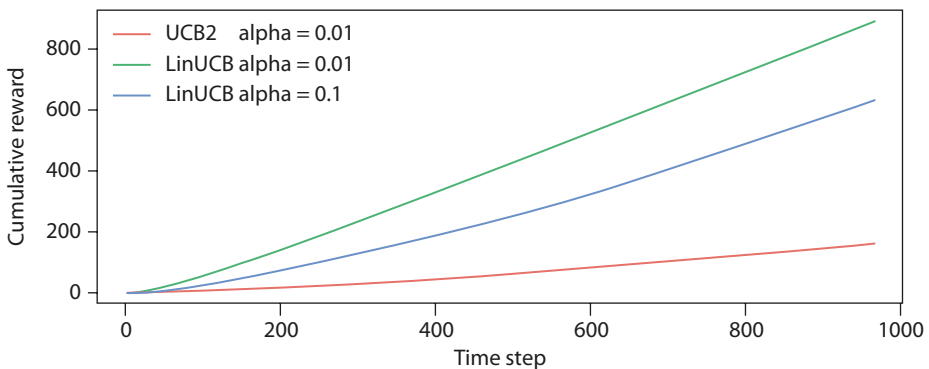
■ Fig. 8.3 Diagram of contextual's basic structure. The context feature vector or matrix returned by `get_context()` is only taken into account by contextual policies and may be ignored by context-free policies. (Author's own figure)

ministic seed) for each to be repeated simulation, runs the agents in parallel, and saves the log of all agent interactions to a history object.

- History: R6 class history keeps a data table-based log of all simulator interactions and several performance measures, such as policies' cumulative reward and regret. Optionally, it also keeps context and theta logs. It allows several ways to interact with these logs, provides summaries, and can save and load simulation logs.
- Plot: R6 class plot generates plots from history logs. It is usually invoked by calling the generic plot(h) function, where h is a history class instance.

8.4.3 Context-Free Versus Contextual Policies

The following code brings all of the classes described in the previous section together by comparing an upper confidence bound method as described above with two contextual linear UCB policies (the latter only differ with respect to their meta-parameter alpha). The code clearly highlights how contextual's comprehensive class structure enables researchers to construct offline policy comparisons with ease (■ Fig. 8.4).



■ Fig. 8.4 Cumulative reward for a context-free UCB2 (Auer et al., 2002) and two contextual LinUCB policies (Li et al., 2010) with differing α -values (determining the width of the upper confidence bound) when evaluated against a “Replay” bandit using offline data. (Author’s own figure)

```

library(contextual); library(data.table)
# Load data, 0/1 reward, 10 arms, 100 features, arms always start from 1
dt <- fread("http://dlie9wlkzgsxr.cloudfront.net/data_cMAB_basic/data.txt")

#
# z y x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x100
# 1: 2 0 5 0 0 37 6 0 0 0 0 25 0 0 7 1 0 0 0
# 2: 8 0 1 3 36 0 0 0 0 0 0 0 0 1 0 0 0 0 10
# 3: . . . . . . . . . . . . . . . . . .

# Set up formula: y ~ z | x1 + x2 + ..
# In bandit parlance: reward ~ arms | covariates of contextual features
f <- y ~ z | . - z
# Instantiate Replay Bandit (Li et al., 2010)
bandit <- OfflineReplayEvaluatorBandit$new(formula = f, data = dt)
# Bind Policies with Bandits through Agents, add Agents to list agents <- list(
Agent$new(UCB2Policy$new(0.01), bandit, "UCB2 alpha = 0.01"),
Agent$new(LinUCBDisjointPolicy$new(0.01), bandit, "LinUCB alpha = 0.01"),
Agent$new(LinUCBDisjointPolicy$new(0.1), bandit, "LinUCB alpha = 0.1")
# Instantiate and run a Simulator, plot the resulting History object history <- Simulator$new(agents,
horizon = nrow(dt), simulations = 5)$run()
plot(history, type = "cumulative", regret = FALSE, legend_border = FALSE)

```

The R package `contextual` is openly available at ► <https://github.com/Nth-iteration-labs/contextual>.

8.4.4 Offline Policy Evaluation with Unbalanced Logging Data

In this section, we demonstrate how `contextual` can be used to perform offline policy evaluation even when the logging policy is not balanced. We use a modern reincarnation of Simpson's paradox (Blyth, 1972) to illustrate the issues involved.

► Example

In the context of this demonstration, imagine a popular website with sports and movie-related articles. The unbiased click-through rate (CTR) per article category for both male and female visitors of this website is presented to the left in Table 8.1.

Clearly, both male and female visitors prefer sports over movie-related articles, which is reflected by the overall CTR per article category. This scenario can be converted to a contextual bandit problem without much difficulty, with article categories for arms and male and female visitors for context:

```
horizon      <- 5000
L simulations <- 1L
#   S----M-----> Arm 1:  Sport
#   |    |           Arm 2:  Movie
#   |    |
weights <- matrix( c(0.4, 0.3, #----> Context: Male
                    0.8, 0.7), #----> Context: Female
                  nrow = 2, ncol = 2, byrow = TRUE)

policy      <- RandomPolicy$new()
bandit      <- ContextualBasicBandit$new(weights = weights)
agent       <- Agent$new(policy, bandit)
simulation  <- Simulator$new(agent, horizon, simulations,
                             save_context = TRUE
                           )
history     <- simulation$run()
u_dt       <- history$get_data_table()

print(paste("Sport:", sum(u_dt[choice==1]$reward)/nrow
            (u_dt[choice==1])))
print(paste("Movie:", sum(u_dt[choice==2]$reward)/nrow
            (u_dt[choice==2])))

[1] "Sport: 0.592259577795152"
[1] "Movie: 0.502457002457002"
```

It is clear that the simulation indeed generates the correct expected unbiased CTR per article category. Let us now run an offline policy evaluation on the generated data and evaluate the CTR of both arms using this dataset:

```
bandit <- OfflineReplayEvaluatorBandit$new(u_dt, 2, 2)
policy <- EpsilonGreedyPolicy$new(0.1)
```

■ **Table 8.1** To the left, a table with unbiased per category click-through rates for male and female visitors to a website. Here, visitors are randomly assigned to article categories. For example, male visitors click through to sports articles four out of ten times and three out of ten times to movie articles—displaying each type of article 50% of the time. To the right, a table representing the same site with the same visitors where the assignment of visitor types to categories is skewed. For instance, here, male visitors are shown sports-type articles 75% and movie articles 25% of the time, and female visitors vice versa

	Sports	Movie		Sports	Movie
Male	0.4×0.5	0.3×0.5	Male	0.4×0.75	0.3×0.25
Female	0.8×0.5	0.7×0.5	Female	0.8×0.25	0.7×0.75
Per category CTR	0.6	0.5	Per category CTR	0.5	0.6

8

```
agent      <- Agent$new(policy, bandit, "OfflineLinUCB")
simulation <- Simulator$new(agent, horizon, simulations,
  reindex = TRUE)
history    <- simulation$run()
ru_dt     <- history$get_data_table()

print(paste("Sport:", sum(ru_dt[choice==1]$reward)/nrow
  (ru_dt[choice==1])))
print(paste("Movie:", sum(ru_dt[choice==2]$reward)/nrow
  (ru_dt[choice==2])))

[1] "Sport: 0.590882178804026"
[1] "Movie: 0.483119906868452"
```

Again, the correct, unbiased CTR estimate per category.

Let us now suggest that the editor of this website just “knows” that men like sport-type articles and women like movie-type articles, without taking a look at the actual data. So, the editor has some lines of code added to the site that assigns movie-related articles, on average, to 75% of female visitors and sport articles, on average, to 75% of male visitors. See ■ Table 8.1, to the right, for an overview of this setup. Surprisingly, even though overall both male and female visitors still prefer sports over movie-related articles, the resulting (now biased) click-through rate (CTR) estimate per article category is suddenly reversed: overall, visitors now seem to prefer sports articles. This unexpected result, known as “Simpson’s paradox” (Blyth, 1972), serves as a perfect backdrop for a demonstration of the use of propensity scores in offline bandit evaluation. To do so, we first need to implement a policy that assigns male and female visitors to articles according to the editor’s wishes:

```
BiasedPolicy <- R6::R6Class(
  ...
  public = list(
    ...
    get_action = function(t, context) {
      if(context$X[1,1]==1) { # 1: Male, 0: Female.
```

```

        prob <- c(0.75,0.25)
      } else {
        prob <- c(0.25,0.75)
      }
      action$choice <- sample.int(context$k, 1, prob =
        prob, ...)
      action$propensity <- prob[action$choice]
      ...
    }
  )
)

```

Note that, importantly, next to `action$choice`, this policy also calculates the probability p_{t,a_t} of selecting this action, returning it in `action$propensity`. This propensity can then:

Be used for inverse propensity weighting (Austin, 2011) to estimate the action's causal effect by accounting for contextual covariates (Imbens & Rubin, 2015; Pearl, 2009)

Be stored in the log, resulting in $D = (x_{t,a_t}, a_t, r_{t,a_t}, p_{t,a_t})$, and then be used to estimate average rewards using inverse propensity scoring (Horvitz & Thompson, 1952) by computing

$$ips(\pi) = \frac{1}{N} \sum_{t=1}^N \mathbb{1}\{\pi(x_t) = a_t\} r_t / p_t \quad (8.6)$$

where the indicator is 1 when π 's action matches the action in the logs (Kruijswijk et al., 2018).

Inverse propensity weighting can reduce bias in the policy evaluation by controlling for the existence of confounding factors that skew random arm assignment. So let us again generate data for the same bandit—but this time round, generated by our editor's biased policy:

```

policy <- BiasedPolicy$new()
bandit <- ContextualBasicBandit$new(weights = weights) agent
<- Agent$new(policy, bandit, "Random")
simulation <- Simulator$new(agent, horizon, simulations,
save_context= TRUE)
history <- simulation$run()
b_dt <- history$get_data_table()

print(paste("Sport:", sum(b_dt[choice==1]$reward)/nrow(b_
dt[choice==1]))) print(paste("Movie:", sum(b_
dt[choice==2]$reward)/nrow(b_dt[choice==2])))

[1] "Sport: 0.497419610956729"
[1] "Movie: 0.60217654171705"

```

Clearly, Simpson's paradox is at work here: overall, visitors do indeed seem to prefer sports articles now. Even worse, when we use this data to evaluate another policy without paying attention to the propensities, the bias propagates itself:

```

bandit <- OfflineReplayEvaluatorBandit$new(b_dt, 2, 2)
policy <- EpsilonGreedyPolicy$new(0.1)

```

```

agent      <- Agent$new(policy, bandit, "rb")
simulation <- Simulator$new(agent, horizon, simulations,
reindex = TRUE)
history    <- simulation$run()
rb_dt     <- history$get_data_table()

print(paste("Sport:", sum(rb_dt[choice==1]$reward)/nrow(rb_
dt[choice==1]))) print(paste("Movie:", sum(rb_
dt[choice==2]$reward)/nrow(rb_dt[choice==2])))

[1] "Sport: 0.511053315994798"
[1] "Movie: 0.618181818181818"

```

However, when we make one minor change to our original replay evaluator, implementing *ips* through the multiplication of rewards by $1/\text{action}\$propensity$ (see contextual's `OfflineReplayEvaluatorBandit` for the full implementation), we are able to fully correct for the editor's bias:

8

```

bandit     <- OfflinePropensityWeightingBandit$new(b_dt, 2, 2)
policy    <- EpsilonGreedyPolicy$new(0.1)
agent     <- Agent$new(policy, bandit, "prop")
simulation <- Simulator$new(agent, horizon, simulations,
reindex= TRUE)
history   <- simulation$run()
prop_dt  <- history$get_data_table()

print(paste("Sport:", sum(prop_dt[choice==1]$reward)/nrow(prop_
dt[choice==1])))
)
print(paste("Movie:", sum(prop_dt[choice==2]$reward)/nrow(prop_
dt[choice==2])))
)

[1] "Sport: 0.601543859649123"
[1] "Movie: 0.519125683060109" ◀

```

The current discussion and the above example offered but a short introduction to offline policy evaluation. For more information on inverse propensity scoring, doubly robust evaluation, and other methods in this field, see for example Dudík et al. (2011) and Swaminathan and Joachims (2015).

8.5 Experimenting with Bandit Policies: StreamingBandit

To take the next step and to start experimenting with policies *in the field*, `StreamingBandit` is a useful tool. `StreamingBandit` is an open-source RESTful web application for developing and deploying sequential experiments in field and simulation studies. It allows designers to easily and quickly implement a policy $\pi()$ on a web server. It is designed such that when set up, it alleviates the technical hurdles for researchers to deploy different policies in the field and thus to enable sequential experimentation to be used within a broader research community.

Just as in contextual, in StreamingBandit, we translate the cMAB problem into two important steps. To ensure the computational scalability of StreamingBandit, we assume that, at the latest interaction t , all the information necessary to choose an action can be summarized using a limited set of parameters denoted θ_{t-1} , the dimensionality of θ often being (much) smaller than that of the historical data \mathcal{D}_{t-1} . Given this assumption, we identify the following two steps of a policy:

1. The decision step: In the decision step, using x_t and θ_{t-1} , and often using some (statistical) model relating the actions, the context, and the reward, which is parametrized by θ_{t-1} , the next action a_t is selected. Making a request to StreamingBandit's *getaction* REST endpoint returns a JSON object containing the selected action.
2. The summary step: In each summary step, θ_{t-1} is updated using the new information $\{x_t, a_t, r_t, p_t\}$. Thus, $\theta_t = g(\theta_{t-1}, x_t, a_t, r_t, p_t)$, where $g()$ is some update function. Effectively, all the data \mathcal{D}_t are summarized in θ_t . This choice means that the computations are bounded by the dimension of θ and the time required to update θ instead of growing as a function of t . Note that this effectively forces users to implement an online policy (Michalak et al., 2012) as the complete dataset \mathcal{D}_t is not revisited at subsequent interactions.

Making a request to StreamingBandit's *setreward* endpoint containing a JSON object including a complete description of $\{x_t, a_t, p_t\}$, and the reward r_t , allows one to update θ_t and subsequently to influence the actions selected at $t + 1$ and further.⁴

For the basic usage of StreamingBandit, the experimenter—or rather an external server or mobile application—sequentially executes requests to the *getaction* and *setreward* endpoints (more details will follow next) and allocates actions accordingly. Using this setup, StreamingBandit can be used to sequentially select advertisements on web pages, for example, allocate research subjects to different experimental conditions in an online experiment, or sequentially optimize the feedback provided to users off a mobile eHealth application. The complete details of how the software is set up and how it should be installed, configured, and prepared can be found in the original paper and the online documentation.⁵ In the remainder of this section, we assume that StreamingBandit is installed.

The python package StreamingBandit is openly available at ► <https://github.com/Nth-iteration-labs/streamingbandit>.

8.5.1 Basic Example

When StreamingBandit is running, a researcher can use some of the default implementations of policies that are shipped with the software. As an example, we run

4 It is also possible to use the *advice_id* functionality, but this is not discussed here for simplicity sake. Full details can be found in the paper.

5 See ► <https://nth-iteration-labs.github.io/streamingbandit> for the complete documentation.

through how ϵ -first would be deployed within StreamingBandit. We will show the code for the *getaction* and *setreward* endpoints and run through them line by line. The *getaction* code for ϵ -first looks as follows:

```
n = 100
mean_list = base.List(
    self.get_theta(key="treatment"),
    base.Mean, ["control", "treatment"]
)
if mean_list.count() >= n:
    self.action["treatment"] = mean_list.max()
else:
    self.action["treatment"] = mean_list.random()
```

This code uses a number of libraries implemented in StreamingBandit. First, the sample size n of the exploration phase of the experiment is set. The next line of code generates a list of *base.Mean* objects from the *libs.base* library. This object provides the functionality to compute streaming updates of sample averages, and the list contains one such average for each of the possible treatments specified by name, using ["control", "treatment"]. The *self.get_theta()* call is used to retrieve θ_t , which in this case thus contains two *base.Mean* objects named “control” and “treatment.” A count, n , and mean reward, \bar{r} , are contained within each *base.Mean* object.

The resulting *mean_list* object thus, in this case, contains two *base.Mean* objects, each of which contains a mean value and a count that can be updated and manipulated. In the next lines, the total count of the number of observations over all mean elements in the list is retrieved. If this is larger than n , the treatment with the highest average value is returned; otherwise, a random element of the list is returned.

Then we have the code for the *setreward* endpoint:

```
n = 100
mean_list = base.List(
    self.get_theta(key="treatment"),
    base.Mean, ["control", "treatment"]
)
if mean_list.count() < n:
    mean = base.Mean(
        self.get_theta(
            key="treatment", value=self.action["treatment"])
        )
    mean.update(self.reward["value"])
    self.set_theta(
        mean, key="treatment",
        value=self.action["treatment"]
    )
```

First, again a *mean_list* is created. After this, the θ_t that is associated with the played action is retrieved and the associated mean object is updated using *mean.update* as long as the exploration phase is ongoing. The last line stores θ_{t+1} such that it can be retrieved again for future decision-making using the *self.set_theta* function. In this implementation, after the experiment when $n > t$, θ is no longer updated.

Once the experiment has been created with this code, it receives an `<exp_id>` and a key `<key>`. This enables the REST endpoints

```
http://HOST/getaction/<exp_id>?key=<key>&context={}
```

and

```
http://HOST/setreward/<exp_id>?key=<key>&context={}&reward={}&action={}
```

where HOST is the location of the hosted StreamingBandit instance. Within the {}s, we can supply the information that is needed by StreamingBandit to select actions and update parameters.

Making a call to `http://HOST/<exp_id>/getaction?key=<key>` and filling in the correct *exp_id* and *key* for the experiment will return a JSON object that looks as follows:

```
{"action":
 {"treatment": "control"},
 "context": {}}
```

A JSON object is a widely accepted internet standard of formatting data, which is both human readable and machine readable. We see here that the object contains an action, which contains an object called *treatment*. This treatment now is equal to the value of *control*. This means that *e*-first has now randomly selected the control condition to be allocated as the first treatment. As long as $n \leq t$, the value of *treatment* will be either randomly *control* or *treatment*. If a *context* were supplied when making the call to the endpoint, then StreamingBandit automatically also returns these values—which may be valuable when integrating StreamingBandit in a web service. To then set a reward, we would call the endpoint for the *setreward* with the action and reward filled in as such: `http://HOST/<exp_id>/getaction?key=<key>&context={}&action={"treatment":"control"}&reward={"value":1}` and this would return the following JSON object:

```
{"action": {"treatment": "control"},
 "context": {},
 "reward": {"value": 1},
 "status": "success"}
```

Again StreamingBandit returns the *action*, *context*, and *reward* values for administrative purposes, but it also returns a *status* object which states whether or not the updating succeeded. In this case, the code ran successfully.

And that is the beginning of your first experiment in StreamingBandit. We have now once requested an action and updated θ . Fortunately, this is not the only functionality that the software supports. StreamingBandit has also implemented a multitude of endpoints that support the use of it, of which two useful endpoints

facilitate the user with testing their code within the software before it is being deployed: the *getcontext* and *getreward* endpoints. These two endpoints can be used to simulate how contexts and rewards are generated and facilitate a full feedback loop within the software. Furthermore, the standard libraries of StreamingBandit support a multitude of default policies, such as Thompson sampling and its bootstrapped variant. In the next part, we show an example of how StreamingBandit has been implemented in a real-world experiment.

8.5.2 StreamingBandit in Action

To demonstrate StreamingBandit in action, we discuss a recent case in which the package was used for online policy evaluation.

► Example

In this example, an online rebate company wanted to examine the effects of their pricing scheme. The company offered customers a rebate on online purchases for numerous e-commerce stores if the customers would purchase products through their website—which is similar to affiliate marketing. The rebate company negotiated different rebates with these e-commerce stores and offered part of the discount to its customers. After a customer signed up, they were shown the different stores and their different discount rates. By default, the company would offer half of the negotiated rebate to the customer. This split was chosen arbitrarily, with no way of knowing if this 50/50 split would be optimal in terms of generating revenue or maximizing profit.

To explore the effects of different splits of the discount on their profits, the company integrated StreamingBandit within their internal system to generate random splits. The split runs from 0 to 1, where a split of 0 would mean that the company would keep the whole discount and with a split of 1 the company would offer the discount completely to the customer. The company would send the maximum possible percentage that they could offer in a context. The eventual discount was then calculated by multiplying the split with the maximum percentage—if the split was 1, StreamingBandit would return the maximum percentage as offered rebate. Using the data of random splits, the company could explore different policies using offline evaluation to see if there was any effect of handling the splits differently.

To generate the random splits, the *getaction* API call was set up with the following code:

```
maxpercentage = self.context['maxpercentage']
split = np.random.uniform()
discount = split * maxpercentage
self.action['split'] = split
self.action['discount'] = discount
```

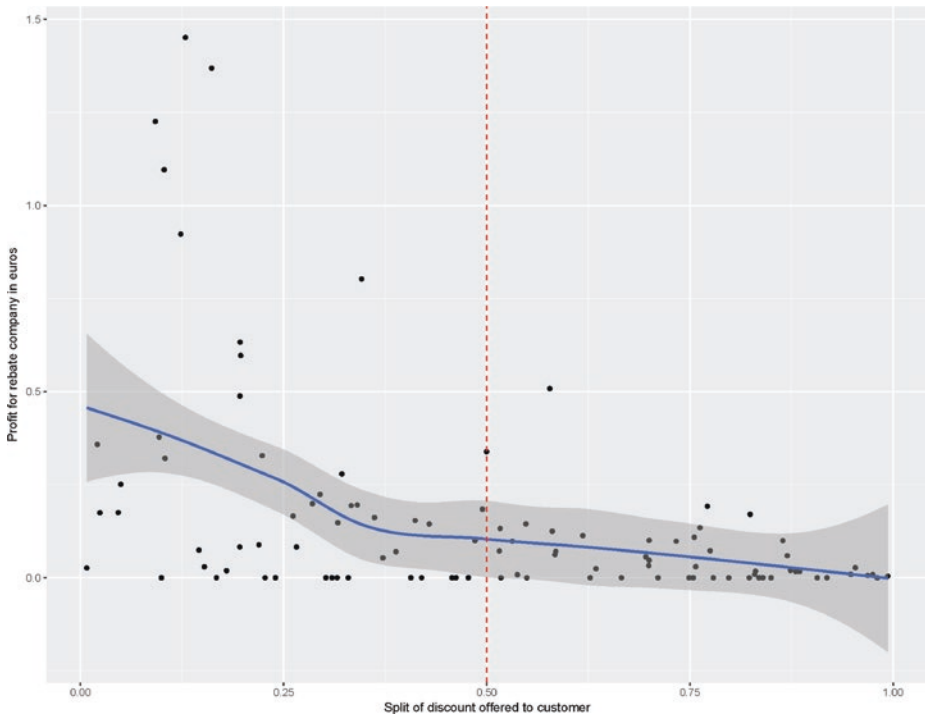
The code is quite straightforward. The *getaction* API call requires a context with a supplied *maxpercentage*. This is the maximum percentage that the company could possibly offer as a rebate—if the random split would be 1, StreamingBandit would return the maximum percentage as offered discount. Then in the second line, a random number would be generated as the split. The total discount is calculated after that and that gets

returned in the responding JSON. Now when the *getaction* API would be called (using for example a maximum percentage of 10), the response will be as follows:

```
{"action": {"split": 0.14300822412482905, "discount": 1.4300822412482905}, "context": {"maxpercentage": 10}}
```

StreamingBandit saves all the data that is incoming with the API calls, so a returning revenue will automatically be saved. Therefore, we do not need to set up the *setreward* call, as in this stage we are merely interested in collecting random data and not updating any parameters of a policy. This is also where one of the main strengths of using *StreamingBandit* lies: since the software is already integrated into their platform, if in a next stage the company would be interested in setting up a multi-armed bandit policy, we would only need to change the two API calls.

■ Figure 8.5 shows the results for the first round of field data collection. On the x-axis, we have the split of the discount that is offered to the customer, ranging from 0 to 1. The red line shows the de facto split of $\frac{1}{2}$ that was used by the company. On the y-axis, we have the profit for the company in euros, which was calculated by 1 minus split times revenue. Each dot shows a completed purchase by a customer, which possibly contains



■ **Fig. 8.5** The collected data for the rebate company, showing the revenue in terms of the discount split. The red dotted line shows the de facto $\frac{1}{2}$ split. Extrapolating from this data, the company should lower their splits to increase the overall revenue (of course not taking into account the number of purchases). (Author's own figure)

multiple products, with $n = 103$ dots. From this limited data (we limited the results to a single e-commerce store), we can see that now a higher discount for the customer leads to a lower overall revenue. A lower discount seems to lead to a higher revenue.

Using an integrated StreamingBandit, the company can now experiment with multiple different policies to deploy into their service. For example, the relation between different stores and different user features can be examined with offline evaluation (Li et al., 2011; Kruijswijk et al., 2019). Also, before deploying any further experiments, the company can also use offline evaluation to compare different policies with each other.

This is one of many examples where a multi-armed bandit problem can be used in practice. A few other examples are using bandits for news article recommendation on websites (Li et al., 2010), customer acquisition via display advertising (Schwartz et al., 2017), and dynamic online pricing (Misra et al., 2019). ◀

Conclusions

In this chapter, we have introduced the multi-armed bandit problem formalization as a useful formalization to look at, and think about, sequential experimentation problems. This formalization provides an extremely fruitful framework to study situations in which an experimenter, by sequentially choosing actions, discloses their outcomes. Sequential experimentation problems are often encountered when trying to use data to make new policy choices: each policy choice constitutes an action, and only the outcome of the selected course of action is disclosed. We have tried to introduce how in such situations the core of the problem is finding a balance between exploring (trying out new actions) and exploiting (playing the most successful actions). Furthermore, we have introduced several strategies that are useful in dealing with bandit problems: strategies such as Thompson sampling are actively used to tackle real-world bandit problems.

After introducing the theory and rationale behind bandit problems, we have introduced two software packages that allow the practitioner to directly start using the gained knowledge:

1. The package contextual allows for easy, offline, experimentation with bandit policies. Effectively, it allows users to study the question: “What would have happened if we had deployed an alternative allocation strategy?” We used our discussion of contextual to illustrate the dangers of unbalanced logging policies and suggested inverse propensity weights as a potential solution. We hope that this discussion encourages readers to actively experiment with (simulated) bandit policies and validly evaluate alternatives based on logging data.
2. Next, StreamingBandit was introduced; this package allows for field experimentation with different policies. Once a policy has been chosen, it is often challenging to actually deploy the policy in the field. This is exactly where StreamingBandit helps: it allows deploying bandit policies at a large scale.

Obviously, a single book chapter is too short to properly introduce a topic as rich as the MAB problem; we, however, content to have given a first practical introduction. The references in this chapter should allow the interested reader to gain a more thorough theoretical understanding.

Take-Home Message

Contextual multi-armed bandits provide a useful problem formalization for sequential learning problems in which exploration and exploitation need to be balanced. Clearly, the traditional experiment (i.e., ϵ -first) is not the only way to address such problems: more effective treatment allocation policies exist and are easy to implement given modern software packages. This chapter introduced contextual for easy simulation and offline policy evaluation, and it introduced StreamingBandit for large-scale policy deployment.

? Questions

1. In sequential learning, the exploration-exploitation trade-off quickly arises: please give a description of this problem in the context of medication testing.
2. The UCB policy balances exploration and exploitation quite explicitly: explain how both components are used in the policy.
3. Offline policy evaluation is tricky when the logging policy is not known: Which information regarding the logging policy is essential to conduct proper, unbiased, offline policy evaluations?
4. What is the main purpose of the contextual package?
5. What is the main purpose of the StreamingBandit package?

✓ Answers

1. When comparing two competing treatments, one would like to (a) explore the effectiveness of each by randomly allocating the treatments to patients and observing the results (exploration), but one would also like to (b) choose the treatment that has the most favorable outcome most often (exploitation).
2. The UCB policy selects the arm with the highest confidence bound. The confidence bound is high because of two reasons: First, the expected reward of an arm might be high; playing this arm equates to making an exploitation choice. Second, the confidence bound of an arm might be large; playing such an arm equates to making an exploration choice. UCB explicitly combines these two objectives.
3. It is essential to know the propensity score (i.e., $\Pr(a_t | x_t, \forall t)$) of the logging policy.
4. The contextual package allows users to quickly evaluate bandit policies through simulation and offline policy evaluation.
5. The StreamingBandit package allows users to deploy bandit policies in the wild.

References

-
- Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3), 235–256.
- Auer, P., & Ortner, R. (2010). UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2), 55–65.
- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3), 399–424.

- Berry, D. A., & Fristedt, B. (1985). *Bandit problems: Sequential allocation of experiments (monographs on statistics and applied probability)* (Vol. 5, pp. 71–87). Chapman & Hall.
- Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338), 364–366.
- Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1), 1–122. <https://doi.org/10.1561/22000000024>
- Chapelle, O., & Li, L. (2011). An empirical evaluation of Thompson sampling. In: *Advances in neural information processing systems* (pp. 2249–2257).
- Dudik, M., Langford, J., & Li, L. (2011). *Doubly robust policy evaluation and learning*. arXiv preprint arXiv:1103.4601.
- Eckles, D., & Kaptein, M. (2014). *Thompson sampling with the online bootstrap*. arXiv preprint arXiv:1410.4009.
- Eckles, D., & Kaptein, M. (2019). Bootstrap Thompson sampling and sequential decision problems in the behavioral sciences. *SAGE Open*, 9(2), 2158244019851675.
- Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press. ISBN: 9780521885881. Google-Books-ID: Bf1tBwAAQBAJ.
- Katehakis, M. N., & Derman, C. (1986). Computing optimal sequential allocation rules in clinical trials. In *Lecture Notes-Monograph Series* (pp. 29–39). <https://doi.org/10.1214/lnms/1215540286>.
- Kruijswijk, J., Parvinen, P., & Kaptein, M. (2019). *Exploring offline policy evaluation for the continuous-armed bandit problem*. arXiv preprint arXiv:1908.07808.
- Kruijswijk, J., Parvinen, P., van Emden, R., & Kaptein, M. C. (2018). Streamingbandit: Experimenting with bandit policies. *Journal of Statistical Software*. <https://doi.org/10.18637/jss.v094.i09>
- Kruijswijk, J., van Emden, R., Parvinen, P., & Kaptein, M. (2016). *StreamingBandit: Experimenting with bandit policies*. arXiv preprint arXiv:1602.06700.
- Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1), 4–22. [https://doi.org/10.1016/0196-8858\(85\)90002-8](https://doi.org/10.1016/0196-8858(85)90002-8)
- Langford, J., & Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems* (pp. 817–824).
- Lattimore, T., & Szepesvári, C. (2018). Bandit algorithms. Preprint (p. 28).
- Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 661–670). ACM. <https://doi.org/10.1145/1772690.1772758>
- Li, L., Chu, W., Langford, J., & Wang, X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM'11* (pp. 297–306). ACM. ISBN: 9781450304931. <https://doi.org/10.1145/1935826.1935878>
- Michalak, S., DuBois, A., DuBois, D., Wiel, S. V., & Hogden, J. (2012). Developing systems for real-time streaming analysis. *Journal of Computational and Graphical Statistics*, 21(3), 561–580. <https://doi.org/10.1080/10618600.2012.657144>
- Misra, K., Schwartz, E. M., & Abernethy, J. (2019). Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2), 226–252.
- Pearl, J. (2009). *Causality*. Cambridge University Press. ISBN: 9780521895606. Google-Books-ID: f4nuexsNVZIC.
- Perchet, V., Rigollet, P., et al. (2013). The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2), 693–721.
- Rabbi, M., Aung, M. H., Zhang, M., & Choudhury, T. (2015). MyBehavior: Automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 707–718). ACM. <https://doi.org/10.1145/2750858.2805840>
- Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4), 500–522.

- Shen, W., Wang, J., Jiang, Y.-G., & Zha, H. (2015). Portfolio choices with orthogonal bandit learning. *IJCAI*, 15, 974–980.
- Sutton, R. S., & Barto, A. G. (2011). *Reinforcement learning: An introduction*. MIT Press.
- Swaminathan, A., & Joachims, T. (2015). Batch learning from logged bandit feedback through counter-factual risk minimization. *Journal of Machine Learning Research*, 16(1), 1731–1755.
- Tang, L., Rosales, R., Singh, A., & Agarwal, D. (2013). Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management* (pp. 1587–1594). ACM. <https://doi.org/10.1145/2505515.2514700>
- Tewari, A., & Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile health* (pp. 495–517). Springer. https://doi.org/10.1007/978-3-319-51394-2_25
- Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4), 285–294.
- van Emden, R., & Kaptein, M. (2020, March). *Nth-iteration-labs/contextual: v0.9.8.3*. <https://doi.org/10.5281/zenodo.3697236>.
- Zhou, L. (2015). *A survey on contextual multi-armed bandits*. arXiv preprint arXiv:1508.03326.



Advanced Analytics on Complex Industrial Data

*Jurgen van den Hoogen,
Stefan Bloemheugel,
and Martin Atzmueller*

Contents

- 9.1 Introduction – 179**
- 9.2 Data Analytics for Fault Diagnosis – 180**
 - 9.2.1 Maintenance of Equipment – 180
 - 9.2.2 Preparing the Data – 182
 - 9.2.3 Machine Learning Classifiers – 184
 - 9.2.4 Deep Learning Techniques – 184
 - 9.2.5 Fault Diagnosis in Practice – 185
 - 9.2.6 Simulating a Real-World Situation – 186
 - 9.2.7 Summary – 187
- 9.3 Graph Signal Processing (GSP) – 187**
 - 9.3.1 GSP Background – 188

Jurgen O. D. van den Hoogen, Stefan D. Bloemheugel and Martin Atzmueller contributed equally with all other contributors.

9.3.2	GSP Applications – 190
9.3.3	Summary – 193
9.4	Local Pattern Mining on Complex Graph Data – 193
9.4.1	Overview – 193
9.4.2	Local Pattern Mining on Graphs – 194
9.4.3	Local Pattern Mining on Attributed Graphs – 195
9.4.4	MinerLSD: Local Pattern Mining on Attributed Graphs – 195
9.4.5	Application Example – 196
9.4.6	Summary – 197
	References – 199

Learning Objectives

- Understand the increased complexity of data analytics in the industrial setting with the usage of sensor data.
- Understand modeling and analysis methods on complex industrial data, covering both sequential (time series) and relational data (networks/graphs).
- Analyze and explain machine conditions with the use of specialized deep learning applications.
- Analyze and re-create signals with less information to optimize the information gathering using graph signal processing.
- Learn how to mine local patterns on complex data in order to enable interpretable machine learning and computational sensemaking.

9.1 Introduction

In the world of today, more and more data are captured via sensors and logs in machinery. As data becomes the new “oil” in different domains—for example in Industry 4.0 (Lu, 2017; Xu et al., 2018) and the Internet of Things (Atzori et al., 2010; Wortmann & Flüchter, 2015)—according methods are necessary to process and actually make sense of the collected complex data, as the “motor” working on, for example, multivariate time series, log data, and multimodal sensor data. With the ever-increasing amounts of large, heterogeneous as well as richly structured datasets, which are often also called *big data* (Wu et al., 2013), advanced analytics methods for modeling, processing, and analyzing such complex data are required (cf. Atzmueller et al., 2016b; Folmer et al., 2017; Gebhardt et al., 2016).

In particular, this relates to, for instance, variety, volume, and veracity of the data, which need to be taken into account. In this chapter, we specifically focus on data covering advanced sequential as well as relational features. This relates mainly to time series as well as complex network and graph structures.

In these contexts, the respective complex data requires both according modeling and analysis methods. Such advanced analytics methods are then able to handle vast amounts of data and allow modeling of inherently complex data while being adaptive to changing environments. In this chapter, we introduce three exemplary topics covering a set of methods and approaches for advanced analytics on complex industrial data:

1. First, we focus on analytics for fault diagnosis, which can handle complex large-scale data, e.g., gathered from sensors in ► Sect. 9.2. Fault diagnosis is a prominent field in advanced analytics on complex data, where we specifically describe a data-driven approach with a focus on deep learning applications. Here, we also tackle explainability and transparency of such methods. We summarize results where the presented methods are demonstrated on a standard benchmark dataset for fault diagnosis.
2. Second, ► Sect. 9.3 briefly presents graph signal processing, outlining the connection between complex graphs (or networks) and signal processing, allowing

the analysis of signals from non-uniformly structured domains (Angelo Medeiros Fonini, 2019). As an application, we consider structural health monitoring using a complex real-world dataset.

3. Finally, in ► Sect. 9.4, we consider local pattern mining for modeling and analyzing complex data in the form of (attributed) networks represented as graphs in order to mine structural as well as descriptive patterns. These patterns are, in particular, simple to interpret and to understand and therefore provide interpretable and explainable machine learning—leading to computational sense-making on complex data.

The remainder of this chapter is structured as follows: we first introduce the aforementioned topics and their relation to complex data. Afterwards, every topic will be discussed separately to outline the respective approaches and methods and their usability in industrial applications.

9.2 Data Analytics for Fault Diagnosis

9

With improved computational power and increase in data gathering, industrial applications become more and more “intelligent.” This section extends on our previous research (van den Hoogen et al., 2020), exploring data-driven approaches for fault diagnosis on machinery and equipment as a part of maintenance strategies in industrial processes. We address the method for data gathering and preparation and the use of machine learning classifiers and finally focus on developments in automated learning methods using deep learning applications. After that, we present data-driven fault diagnosis in practice together with simulating real-world situations.

9.2.1 Maintenance of Equipment

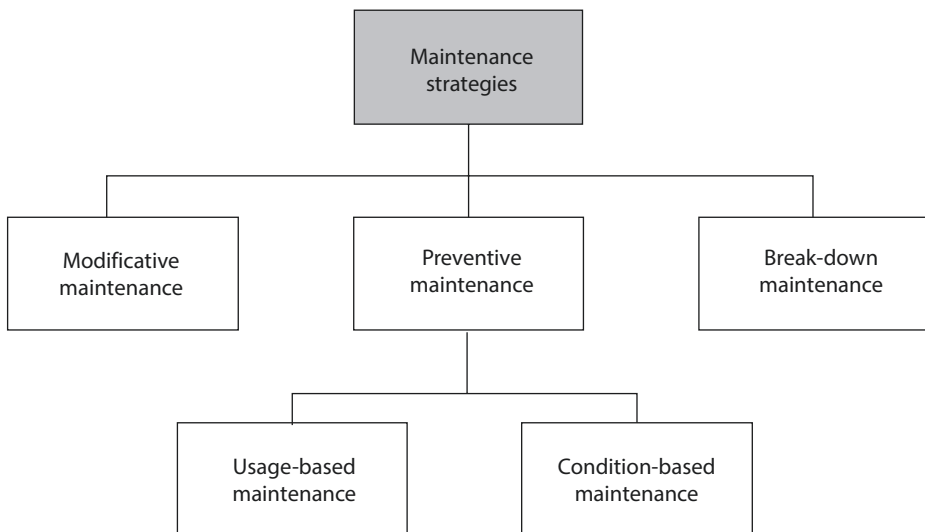
Industrial applications depend on the use of machinery. One of the most common issues in industry is the breakdown of these machines due to wear of the underlying parts. Maintenance of the equipment is vital in reducing these breakdowns and therefore lowering the downtime. According to Mobley (2002), costs for maintenance vary between 15% and 60% of the overall costs of produced products. Within these margins, around 33% of the maintenance costs are directly linked to redundant and inaccurate maintenance of equipment. Therefore, reducing the costs for expensive maintenance could drastically reduce the overall production costs by increasing productivity of the equipment (Alsyof, 2007).

Maintenance Strategies There are three different strategies defined for maintaining equipment in industrial applications (Arts, 2017; Jardine et al., 2006): (1) modificative maintenance where parts are being replaced by an upgrade to boost productivity and performance of the machine, (2) preventive maintenance that replaces a part just before a failure, and (3) breakdown corrective maintenance that occurs right after

failure, which leads to downtime of the machine. In this chapter, we focus on preventive maintenance, which can be divided into two sub-strategies, usage-based maintenance (UBM) and condition-based maintenance (CBM). ■ Figure 9.1 shows the different maintenance strategies.

Preventive Maintenance Within preventive maintenance, the UBM strategy solely focuses on scheduling maintenance visits by the engineer when a certain threshold of usage is reached. In practice, this means that visits are planned with fixed time in between, similar to a yearly checkup for cars. This strategy results in very low downtime of the equipment, which is beneficial for the productivity. However, this strategy comes with one major drawback due to the high costs of maintenance visits and replacement of parts that are still useful. Therefore, CBM is the preferred maintenance strategy in many industrial applications (Jardine et al., 2006). CBM monitors the current condition of equipment to determine what type of maintenance is needed. The idea behind CBM is to perform maintenance only when certain indicators, e.g., deviations in data, show a decrease in performance or an expected increase in failures. This results in less maintenance visits and optimal use of the underlying parts.

Fault Diagnosis Fault diagnosis is a prominent technique for industrial machinery and is a part of CBM. Traditionally, fault diagnosis was initially done using physics-based models that require prior knowledge of the underlying processes and were unable to update to new measurements (Yin et al., 2014). The developments of the industrial Internet of Things (IoT) and data-driven analytics techniques changed the field of fault diagnosis in a more intelligent manner (Zhao et al., 2019). These techniques are able to automatically process data with little prior knowledge on technical aspects of machinery and are adaptable to a changing environment.



■ Fig. 9.1 Different types of maintenance strategies according to Arts (2017)

Specifically, the use of deep learning applications has become increasingly popular due to the availability of large-scale datasets and improved computational power. Nowadays, fault diagnosis relies more and more on sensors that record large-scale time series data. For example, when parts of the machinery degrade over time, this will not be directly seen in the analogue metrics of the machine itself. However, things such as increasing power consumption or vibrations of parts of the machine measured with external technology, e.g., sensors, could indicate that the underlying parts need to be replaced.

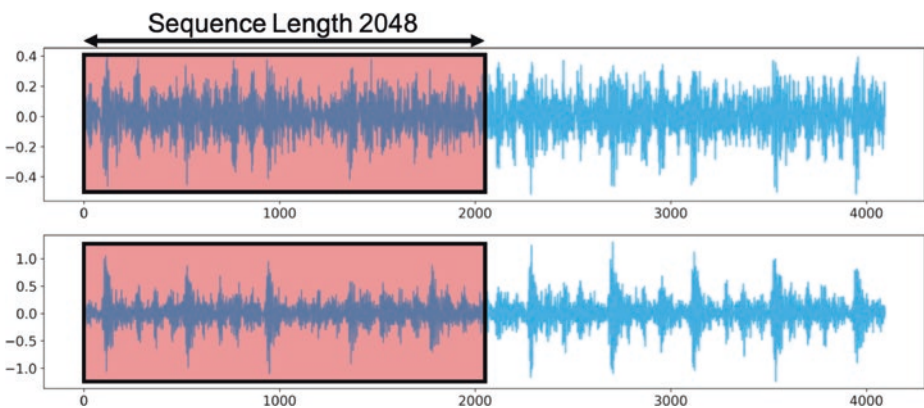
Fault diagnosis is therefore reliant on the use of sensors in industrial applications and is most applicable on equipment with a low failure frequency and a high downtime (Scarf, 2007). Fault diagnosis consists of three tasks (Liu et al., 2018b), where most research primarily focus on the first task:

1. Determining the state of equipment
2. Detecting failures
3. Forecasting fault development over time

9.2.2 Preparing the Data

9

Data for diagnosing fault conditions can be gathered using sensors that record vibration signals. Digitizing these signals transforms them into one-dimensional time series data, represented into one large vector for every sensor. Depending on the amount of sensors placed on the machine, the data can be represented as univariate or multivariate time series. However, the data is not directly usable for traditional machine learning classifiers. First, it needs to be segmented into sequences with a corresponding condition label before training the model, as can be seen in **■** Fig. 9.2. This segmentation is of arbitrary length that is frequently derived as a power of 2 for implementing the widely used fast Fourier transformation (FFT) algorithm (Zhang et al., 2017, 2019).



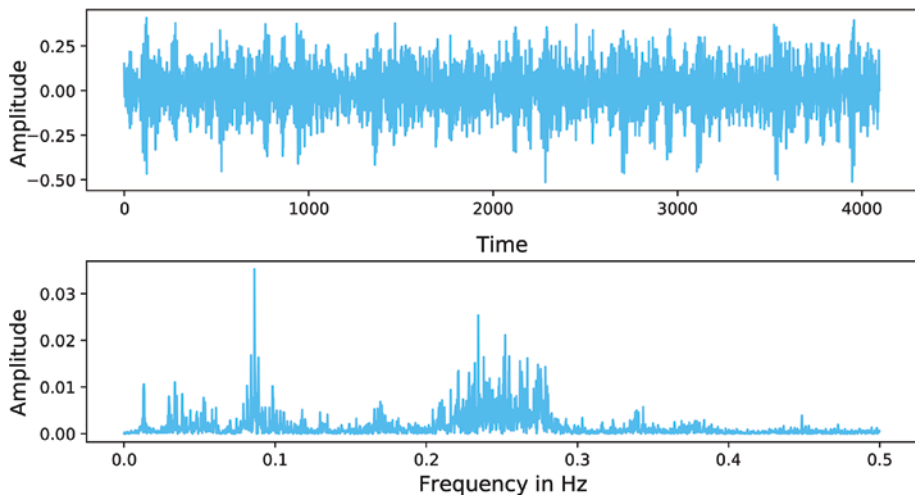
■ Fig. 9.2 Example signal of the vibrations from two sensors. The red box indicates the segmentation of signals into sequences (van den Hoogen et al., 2020)

After segmentation, features need to be extracted from the time series data because raw data on its own is not informative enough. This can be done in several ways by analysis on the time domain (e.g., statistical measures such as mean and standard deviation), frequency domain using Fourier transformations (Lei et al., 2016), or time-frequency domain with wavelet transformations (You et al., 2014).

■ Figure 9.3 shows an example of a signal in the time domain transformed to the frequency domain using the FFT algorithm. After transformation, features in the frequency domain can be extracted/engineered, e.g., peak detection, power, and energy of the signal. Over the years, enhancements and hybrid combinations of these analyses in the different domains were carried out regularly to improve the model's performance.

One can imagine that the amount of features derived from the different domains commonly results in a high-dimensional dataset. To accommodate for the curse of dimensionality, algorithms are used to reduce the dimensionality of these features, such as principal component analysis (PCA) (Malhi & Gao, 2004; Zhang et al., 2005) or linear discriminant analysis (LDA) (Jin et al., 2013). These techniques are used to compress the vast amount of features into smaller representations. When the dimensions of the features are reduced to an acceptable level, the data can be fed to a classifier.

The necessary preprocessing steps require a significant amount of time and high-level expertise in signal processing and data processing, before the final dataset can be fed to a classifier. In addition, the feature extraction process is dependent on the type of machinery and sensors used for fault diagnosis. It is therefore that deep learning applications with automatic feature extraction have become increasingly popular in the field of fault diagnosis.



■ Fig. 9.3 Example of the original signal in the time domain (upper) transformed to the frequency domain (lower) using FFT. (Author's own figure)

9.2.3 Machine Learning Classifiers

As described in the previous paragraph, raw signals need to be preprocessed before using traditional machine learning classifiers for fault diagnosis. These new representations of the features can be used to train the classifier efficiently and improve performance drastically.

In the field of fault diagnosis, many different classifiers are used such as K-nearest neighbors (KNN) (Pandya et al., 2013), support vector machines (SVMs) (Huang et al., 2011; Konar & Chattopadhyay, 2011; Santos et al., 2015; You et al., 2014), artificial neural networks (ANNs) (Chow et al., 1991; Cococcioni et al., 2013), and less common techniques in fault diagnosis such as random forest (Wang et al., 2017). The performance of these techniques varies a lot depending on the data quality, thoroughness of the feature extraction process, and complexity of the classification task. Therefore, it is often particularly difficult to find the right classifier for the task at hand. Previous research has shown that there is not one particular machine learning classifier that is most capable of distinguishing different fault conditions. Therefore, a comparison between classifiers is deemed necessary for every fault diagnosis task to find the most optimal model.

9

9.2.4 Deep Learning Techniques

The use of deep learning approaches and methods has significantly increased efficiency and performance in applications for fault diagnosis. Their ability for automatic feature extraction and classification by processing (raw) input data saves time, and they are often less sensitive to errors. Additionally, deep learning techniques do not require extensive prior knowledge on feature extraction techniques in the signal processing domain and are able to scale up when the amount of data increases. In the field of fault diagnosis, many deep learning models have been tested.

One of the first deep learning models applied for fault diagnosis was the multi-layer perceptron (MLP) (Hajnayeb et al., 2011). This model learns feature representations of the input by stacking multiple layers. Unfortunately, to create representative features from raw signals, one must design a network with a certain amount of depth. For the MLP, this resulted in a drastic increase in computation time which made its capabilities limited for diagnosing fault conditions.

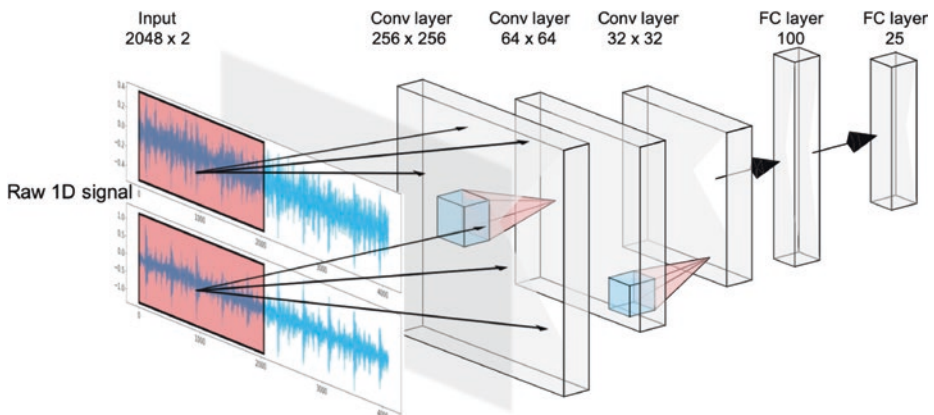
The use of recurrent neural networks (RNNs) showed promising results since they are able to handle long-term dependencies from time series data well (Malhotra et al., 2016). Unfortunately, taking these dependencies into account results in slow training times due to excessive memory use. On the other hand, autoencoders were used to reduce the dimensionality of the data while creating useful representations (Meng et al., 2018; Sun et al., 2016). By combining an autoencoder as feature extractor with a RNN that is able to account for long-term dependencies, researchers were able to solve the problems with slow training times (Liu et al., 2018a). However, this also led to increased complexity and reduced interpretability of the model.

Convolutional neural networks (CNNs) had already shown positive results in the field of computer vision such as image classification and video recognition (Simonyan & Zisserman, 2014). Usually, a CNN processes two-dimensional data that represents image pixels and color channels. The convolutional layer convolves the input with filter kernels followed by the activation unit to generate output features. Each of these filters uses the same kernel to extract local features from the input's local region, which is called weight sharing. Results of the convolutional operations across the input are fed to the activation function that leads to the output features.

To use a CNN for time series data, one needed to transform the data into a two-dimensional representation in the time-frequency spectrum (Hoang & Kang, 2019). This would indicate that the data needs to be preprocessed, which is something that collides with the benefits of using deep learning applications over traditional machine learning techniques. The development of the one-dimensional (1D) CNN solved this problem by combining the automatic feature extraction with classification specified for time series data. These models tend to handle noise in the time series well and are able to be trained with limited data. It is therefore that 1D CNNs are considered the best option in fault diagnosis. ■ Figure 9.4 shows an example of a 1D CNN that is able to process multivariate signals.

9.2.5 Fault Diagnosis in Practice

Previous research has shown that around half of the broken machinery is caused by rolling bearing element faults (Group et al., 1985; Zhou et al., 2007). A rolling bearing element is part of the rotating mechanism of industrial machinery. This element is subject to degradation due to the rotations caused by an electric-driven motor and is one of the key components for determining the condition of the machine. To measure the condition of a rolling bearing element, one can place



■ Fig. 9.4 Example of a one-dimensional CNN for processing multivariate signals with three convolutional layers and two fully connected layers. (Author's own figure)

vibration sensors on the designated parts to record the vibrations in the form of a continuous signal. These vibration signals indicate the underlying condition of the rolling bearing element well (Jing et al., 2017) and can be digitized into numerical time series data.


For accurate rolling bearing fault diagnosis with the use of deep learning, CNNs have proven to perform very well. Especially the use of a wide kernel in the first convolutional layer followed by small kernels in the following convolutional layers has shown to handle signal data from sensors particularly well (Zhang et al., 2017). The most optimal models usually contain around five convolutional layers combined with pooling layers (Zhang et al., 2017, 2019). This particular model architecture is able to handle noisy data, which makes them suitable for monitoring the conditions of industrial machinery. van den Hoogen et al. (2020) proposed an improved wide-kernel CNN specifically designed for classifying multivariate signals with the use of multichannel CNNs.

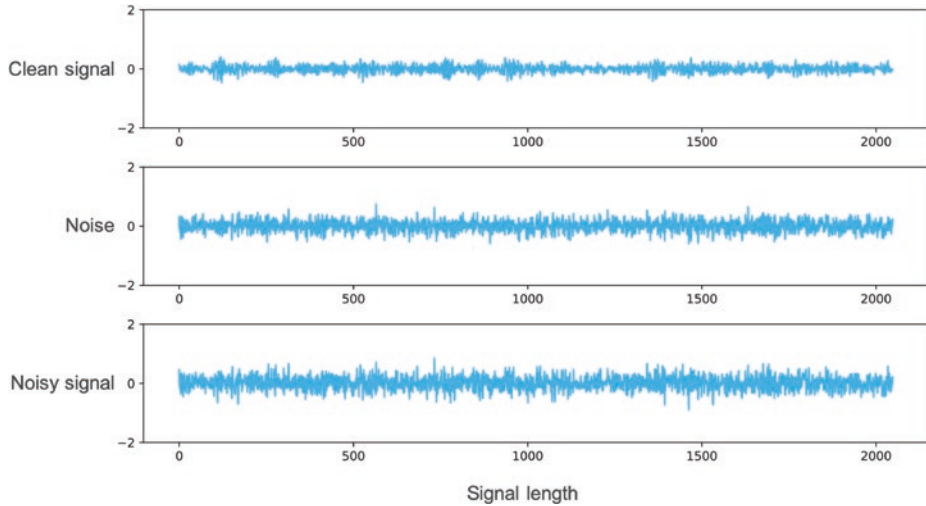
The Case Western Reserve University (CWRU) bearing dataset (Case School of Engineering, n.d.) is widely used to test a model's ability for diagnosing fault conditions. This dataset consists of various sub-datasets, each containing a specific fault condition, and is divided into several categories. Behind the dataset, CWRU conducted experiments where they inflicted damage to the bearing element on different locations. By adding sensors to the machine, they measured the vibrations.

9.2.6 Simulating a Real-World Situation

In real-world applications, clean signals such as the data from the CWRU bearing experiment are rarely available due to external factors that influence the sensor (e.g., pulses from other machinery). To create signals that better reflect a real-world situation, one can exploit random noise. By adding additive white Gaussian noise (AWGN) to the signal, a clean signal can be transformed into a noisy signal. We use the signal-to-noise ratio (SNR) calculating the proportion of noise compared to the clean signal, measured in dB. The SNR is denoted as

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{Noise}}} \right)$$

A SNR of 0 dB reflects that the noise signal is of equal power as the original signal. When the SNR becomes negative, the noise becomes stronger than the original signal, which indicates that lowering the SNR usually results in a lower performance of the model. We recommend to test several SNR levels for evaluating a model so that you get a good reflection of the model's performance under different noise environments. However, there is no general rule for knowing which SNR level is representing a real-world situation most accurately. Nonetheless, one can reason that when the power of the noise signal is much higher than the original signal, this might indicate that the designated sensor is not working properly. An example of a constructed noisy signal is shown in  Fig. 9.5.



■ **Fig. 9.5** Example of constructed noisy signal with SNR level of -4 dB. (Author's own figure)

9.2.7 Summary

Fault diagnosis is a method used in condition-based maintenance to determine the state of equipment, locate the origin of faults, and forecast fault development over time. With more computational power and data available, fault diagnosis has become more and more data driven, especially towards using deep learning methods and applications, which we covered in this section.

9.3 Graph Signal Processing (GSP)

Traditional signal processing can be extremely powerful in uniform, euclidean domains such as sampled audio or power circuits. In such situations, the data can be represented in euclidean space defined by R^n for n dimensions. However, not all domains have such a desirable property. For example, when the data at hand are sensors placed along a piece of land, the topography will most likely not resemble a perfect, uniform square grid. There could be mountains that influence the altitude of the sensor and bodies of water where no sensor can be placed. Moreover, transportation networks also resemble complex connections that are not structured uniformly. Some locations will serve as hubs in the network, while there will be less dense connections in more non-urban areas.

9.3.1 GSP Background

To start, a graph is a data structure of ordered pairs of nodes connected by edges. The edges between nodes can be either directed or undirected and weighted or unweighted. Depending on the type of data that is available, a researcher can tune each of these parameters in the according modeling process: what are nodes, what are the connections, and what resembles the weights. Often, it then turns out that a striking way to assign weights to the graph structure (edges) can already solve a problem on its own.

Commonly, GSP focuses on weighted graphs where the edge between two nodes resembles the amount of trust in the relationship between the sensors (Stankovic et al., 2019a). For example, in the case of analyzing a network of temperature sensors irregularly placed around a country, distance would be the property that defines the relationship between two nodes. A graph signal is then the set of scalar values that represent the temperature at each sensor location (the nodes). See **Fig. 9.6** for a simple example.

To demonstrate and explain the theoretical concepts in a bit more detail, consider the graph shown in **Fig. 9.7**. Here, the degree (D), adjacency (A), and Laplacian (L) matrix are

9

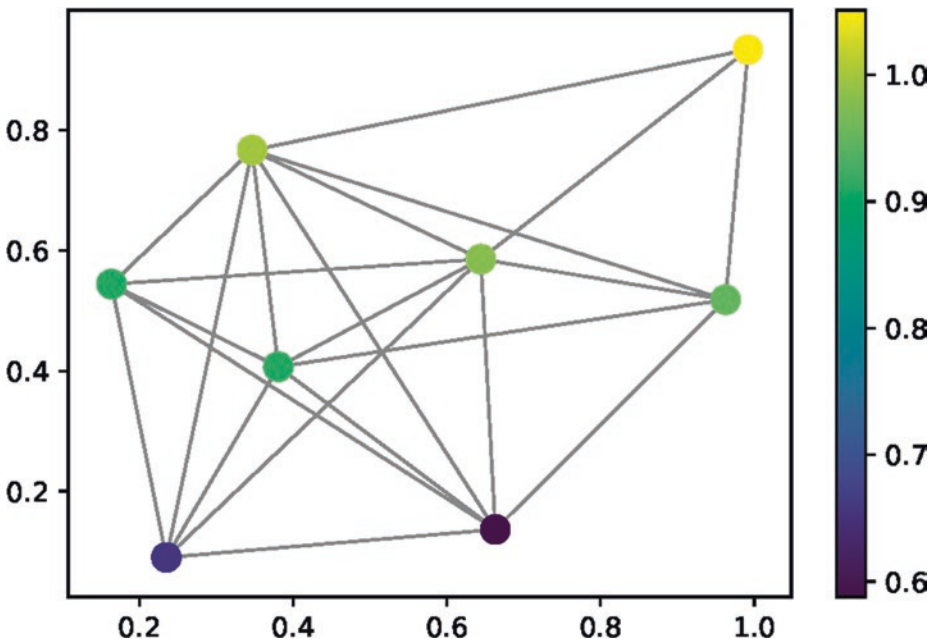
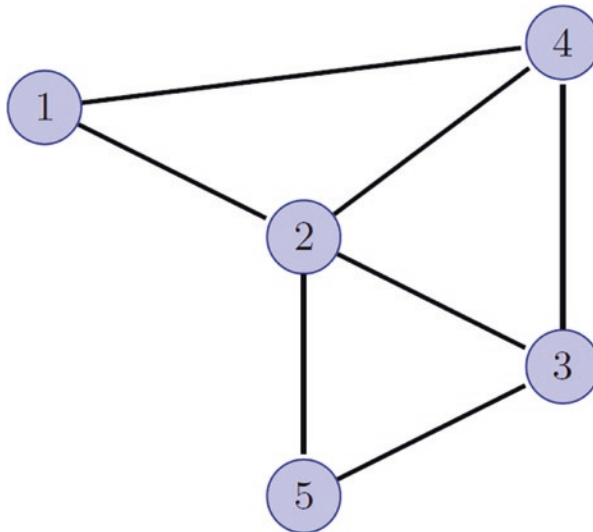


Fig. 9.6 An example sensor network with 8 nodes and 22 edges. Each node corresponds to a sensor, and each edge represents the distance between the edges. A cutoff was made to remove edges that had a longer distance than the threshold. The x- and y-axes resemble the x- and y-locations where the sensors were placed. (Author's own figure)

$$D \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix} - A \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} = L \begin{bmatrix} 2 & -1 & 0 & -1 & 0 \\ -1 & 4 & -1 & -1 & -1 \\ 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & -1 & 3 & 0 \\ 0 & -1 & -1 & 0 & 2 \end{bmatrix} \quad (9.1)$$

For a graph with n nodes, the adjacency matrix is an $n \times n$ matrix where each element indicates whether pairs of nodes are connected by the respective edges. For example, the first row represents the connections of node 1, $[0, 1, 0, 1, 0]$, respectively. Therefore, node 1 has connections with nodes 2 and 4. The degree matrix is a diagonal matrix which contains information about the number of edges that a node has. Considering node 1, the connections with nodes 2 and 4 add up to 2. The *Laplacian* matrix is then the degree matrix minus the adjacency matrix. It is thus a $n \times n$ symmetric matrix (and shift operator) that has n real eigenvalues $\lambda_1 < \lambda_2 \dots < \lambda_n$ (the spectrum) and eigenvectors v_1, \dots, v_n . These eigenvalues and eigenvectors tell us a lot about the graph (Stankovic et al., 2019b). For example, they can be used to calculate the Fourier transformation of a graph-based signal.

In classical signal processing, one of the most prominent tools is the Fourier transform. The Fourier transform is a decomposition of a signal in its containing frequencies. The transformation also contains information about the magnitudes of the available frequencies in a signal. In GSP, this transform is achieved by evalu-



■ Fig. 9.7 A graph with five nodes and seven edges. (Author's own figure)

ating the eigenvalues and eigenvectors of the Laplacian matrix. This graph Fourier transform enables a wide range of applications for graph signal processing, which will now be discussed below.

9.3.2 GSP Applications

There is a wide range of applications of graph signal processing in various domains. Below, we sketch some application scenarios and provide examples, specifically targeting the sampling of graph signals, monitoring signals in order to detect specific patterns, as well as applying special filtering on graph signals in the spectral domain employing Fourier transformation techniques.

Sampling One of the first applications that was investigated considered the sampling of graph signals. Imagine a sensor network of humidity sensors (see [Fig. 9.8](#)). It could be the case that each of these sensors runs on battery power, and to increase the lifetime of the sensors, the sensors in the network may not operate simultaneously. What would then be an appropriate (optimal, minimal) subset of sensors to infer the original signal back?

9

Such an approach has been performed on bridge sensor data collected in the InfraWatch project (Bloemheuveld et al., 2020). Sensors were placed on the girders (long bars that carry the strain) and the deck of the bridge. In a practical application, it can take a while to analyze all the sensors simultaneously, and not all sensors need to be turned on all the time. Therefore, calculating an optimal subset of sensors to reconstruct the entire signal could save bandwidth and decrease computation time, as well as decrease overall maintenance costs and thus improve overall equipment efficiency. In addition, future projects could inspect the optimal positions of sensors to decrease the amount of sensors that need to be installed, saving costs of planning and purchasing such sensor systems.

Monitoring In addition, the sensor network can be monitored to detect certain patterns in the strain signals (Bloemheuveld et al., 2020). For example, [Fig. 9.9](#) shows

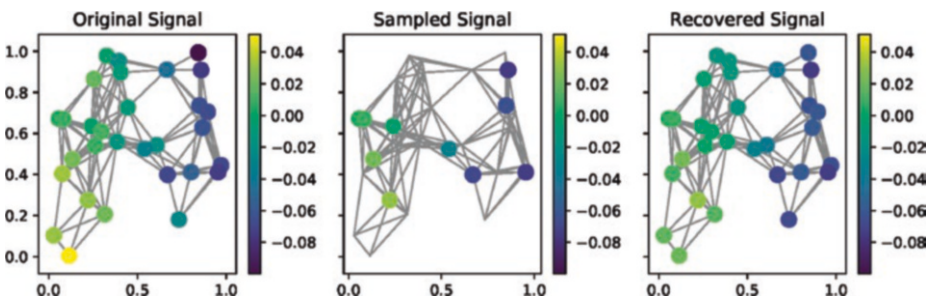
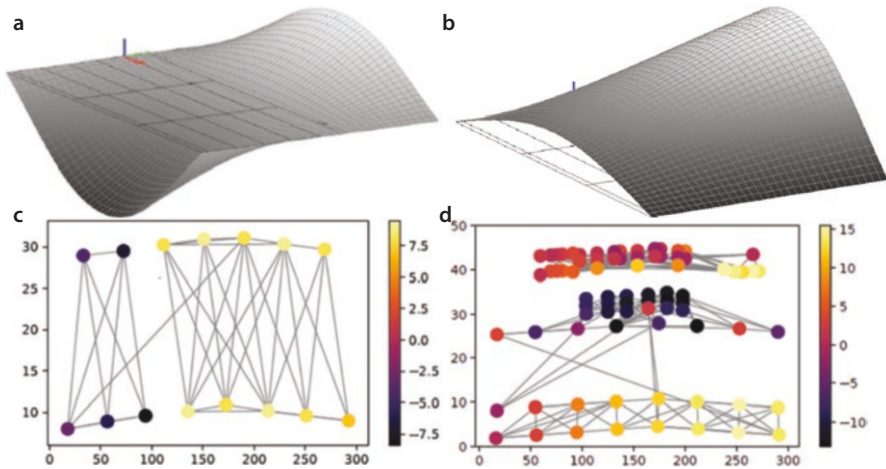


Fig. 9.8 Example case of graph signal sampling. From left to right: (1) The original signal of 30 nodes representing the 30 sensors in the network. (2) Sampled signal of 10 nodes. (3) The recovered signal from extrapolating the sensors that were sampled in (2). (Author's own figure)



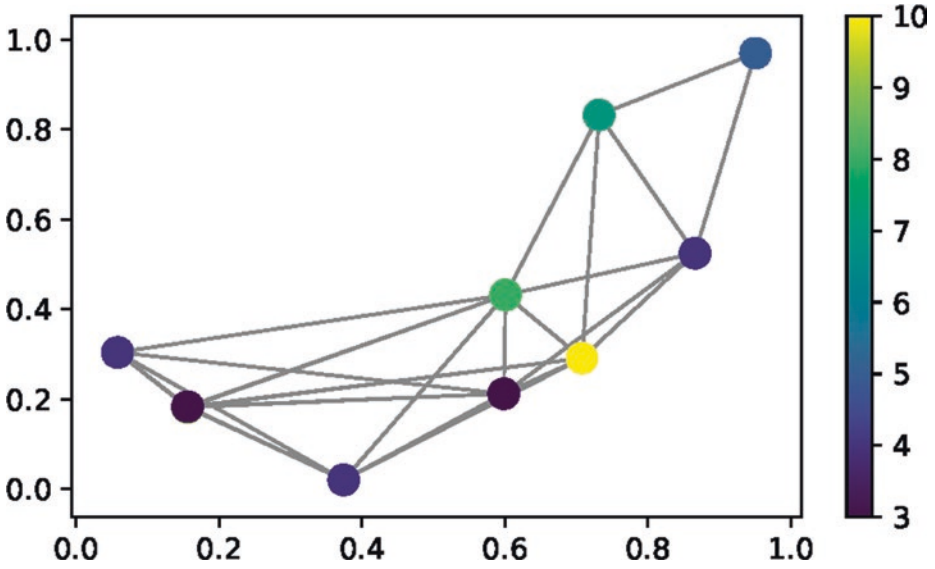
■ **Fig. 9.9** a Shows a FEM-based (FEM—finite element method—from Miao et al., 2013) combination of mode shapes and the corresponding graph signal in c. b Shows a FEM-based (from Miao et al., 2013) combination of torsional mode shapes in the girders and the corresponding graph signal in d. (Author's own figure)

the bridge sensors in two conditions: (a, c) resemble the sensor network when no activity is measured, and (b, d) resemble a truck driving over the right side of the bridge. The strain sensors at the bottom-right side of the bridge measure a huge increase in strain, whereas the sensors on the deck of the bridge show a decrease in strain. This is exactly what the girders should do, so engineers can inspect animations of such sensor networks to monitor the behavior of the bridge. The behavior of the bridge to certain events can act as key indicators for the stiffness and damping of the bridge, which are indicators for structural health.

Filtering Another application domain for GSP is filtering signals, which is possible by performing the Fourier transformation on a graph signal. The fundamental approach is to transform the graph signal into the graph spectral domain with the Fourier transformation, weaken unwanted or magnify wanted frequencies of the signal by altering the Fourier coefficients, and convert the signal back to the vertex domain (a graph signal). By applying a low-pass filter defined as

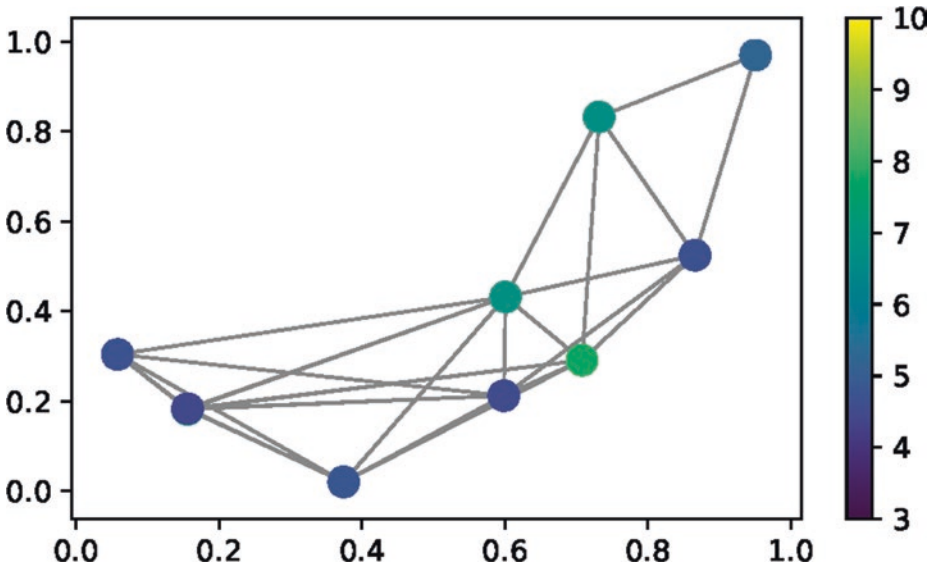
$$g(x) = \frac{1}{1 + \tau x}, \quad (9.2)$$

the lower frequencies of the signal are kept. The result of such a filter is visible in ■ Figs. 9.10 and 9.11. The sensor networks show a sensor reading in the middle of the network that is off from the rest of the network. Once the signal is filtered, this anomaly is spread out over the network and a new signal is created. Low-pass filtering can be used to remove outliers from signals, and high-pass frequencies could be used in contrast, in order to spot such anomalies.



■ Fig. 9.10 The original signal of temperature values in a sensor network. (Author's own figure)

9



■ Fig. 9.11 The low-pass filtered data in the sensor network. (Author's own figure)

9.3.3 Summary

The measuring of complex systems with sensing technology and signal processing has seen significant progress in the last two decades. Here, recent advances in network science formulated new challenges for linking huge amounts of data collected in hundreds or even thousands of sensors working in networked structure. This triggered scientists to develop new approaches to work with such complex data. As a promising solution to such challenges, GSP already showed its value in a lot of different research areas. In particular, GSP has the advantage over classical data domains that graphs account naturally for irregular data relations. After successfully modeling a particular problem as a GSP problem, this framework captures both techniques from the fields of signal processing and graph theory. This section outlined the basic concepts as well as examples of the application of GSP.

9.4 Local Pattern Mining on Complex Graph Data

The detection of local patterns is a prominent approach in knowledge discovery and data mining (e.g., Knobbe et al., 2008; Morik, 2002; Morik et al., 2005), in particular for mining complex graph data. We first provide an overview, before we discuss methods and applications (Atzmueller et al., 2019a; Atzmueller & Kloepper, 2018).

9.4.1 Overview

In general, pattern mining aims at discovering a set of *novel*, *potentially useful*, and ultimately *interesting* patterns from a given (large) dataset (Fayyad et al., 1996). Common methods include those for association rule mining (Agrawal & Srikant, 1994) or subgroup discovery (e.g., Atzmueller, 2015; Klösgen, 1996; Knobbe et al., 2008; Lemmerich et al., 2012; Wrobel, 1997). The goal is typically to detect a set of the most interesting patterns according to a given quality function, e.g., with a quality above a certain threshold, or the top-k patterns according to their interestingness.

In general, local pattern mining aims at identifying locally interesting structures, also often called “nuggets in the data” (Klösgen, 1996) with an interpretable description. Thus—compared to global modeling which aims at modeling the complete dataset—local pattern mining focuses on identifying such locally relevant (and interesting) patterns which can be easily interpreted and understood by humans. Subgroup discovery, for example, is an exploratory approach for discovering interesting subgroups, at the intersection of descriptive and predictive data mining (Lavrac, 2005). For example, regarding two sensors SX and SY with respect to a target parameter *scrapRate*, there could be a pattern (rule) like the following:

$$\textit{pressure}_{SX} = \textit{low} \text{ AND } \textit{temperature}_{SY} = \textit{high} \text{ THEN } \textit{scrapRate} = \textit{high}.$$

Here, we see that the pattern is made up of simple conditions which are typically combined by a conjunction (“AND”)—with respect to a specific target concept. Such rules can then be discovered using local pattern mining, e.g., using subgroup discovery as described above. Applications involve, for example, technical (fault) analysis (e.g., Atzmueller & Lemmerich, 2009; Atzmueller et al., 2005; Jin et al., 2014), like mining service processes (Natu & Palshikar, 2014), analysis of smart electrical meter data (Jin et al., 2014), or fault analysis of production processes (Atzmueller & Lemmerich, 2009; Atzmueller & Sternberg, 2017; Sternberg & Atzmueller, 2018). The latter, for example, has been implemented using the VIKAMINE system (Atzmueller & Lemmerich, 2012) by identifying patterns (as combination of certain factors) that cause a significant increase/decrease in, e.g., the fault/repair rates of certain products (cf. Atzmueller & Lemmerich, 2009). Since the discovered patterns are typically quite understandable, local pattern mining also facilitates *computational sensemaking* (Atzmueller, 2018), which is quite important for practical applications. Computational sensemaking aims to “make sense” in the context of complex information and knowledge processes. This is enabled using computational methods for *analysis*, *interpretation*, and *intelligent decision support*. While the latter is mostly supported by human-computer interaction techniques, the former two are supported by data mining approaches, e.g., by local pattern mining methods.

9

For analyzing complex data, such as sequential (e.g., Atzmueller, 2016; Atzmueller et al., 2017) and in particular relational data (e.g., Atzmueller et al., 2016a, 2019a, b), we need to apply methods that not only are able to mine simple tabular data but can also directly work on complex graph data representations such as complex networks and temporal graphs. Below, we first outline the basic concepts before we describe application examples.

9.4.2 Local Pattern Mining on Graphs

Complex networks or graphs can be applied for modeling complex data, such as sequential, temporal, and multi-relational concepts. As briefly outlined in the previous section, typically a complex network represented as a graph is made up of a set of vertices (often also called nodes) and a set of edges (links) connecting the nodes. The links can then be defined according to various criteria; for example, when analyzing event log data consisting of a set of events that are captured in log data with specific timestamps, an according complex network can be constructed with the events (as vertices) and links between those indicating sequential relationships. Local pattern mining then considers the extraction of (structural) patterns contained in the graphs, i.e., subgraphs with specific properties.

For pattern mining on networks and graphs, there thus exist several quality measures, usually taking into account the *support* of the pattern, i.e., its size. Furthermore, the topological structure of the subgraph induced by the pattern is also taken into account. The goal then is to enumerate the set of all patterns that satisfy some constraint, e.g., a minimal support in terms of the number of covered

objects, their (topological) connectivity, or a topological property or constraint. Regarding a topological property of a graph, for example, a popular approach consists of extracting a *core subgraph* from the network, i.e., some essential part of the graph whose nodes satisfy a local property. The k -core definition was first proposed by Seidman (1983). It requires all nodes in the core subgraph to have a degree of at least k .

9.4.3 Local Pattern Mining on Attributed Graphs

Besides the topological structure of the graph, often further information can be included for local pattern mining, whenever properties of the graph's vertices are available. Then, we can consider an attributed graph made up of the vertices and edges as before, but also including property vectors on vertices and/or edges.

In particular, local patterns on attributed graphs allow the characterization in terms of their structural (topological) as well as attributive features. Then, either structural features such as a specific structure in the graph, e.g., a subgraph with many connections, or a conjunctive description, i.e., as a conjunction of descriptors such as *conditionA AND conditionB AND conditionC*, can characterize a pattern. We could, for instance, consider the example discussed above in the context of a sensor network with labels (*low, medium, high*) and then also investigate a pattern like *pressureSX = low AND temperatureSY = high*. As we will also outline below, this can also relate to events such as *event1 AND event2* or a specific pattern in our event log example below, *Warning AND cell6 AND line20*, for example relating to warning condition(s) occurring in (production) lines 6 and 20, relating to the specific subgraph.

Pattern mining on attributed graphs specifically aims at getting a description-oriented view on the pattern, making them interpretable and explainable. A specific instance of local pattern mining on attributed graphs is given by the *MinerLSD* algorithm (Atzmueller et al., 2019b) described below.

9.4.4 MinerLSD: Local Pattern Mining on Attributed Graphs

The *MinerLSD* algorithm allows efficient and effective pattern mining. It applies efficient pruning techniques and utilizes effective constraints using graph abstraction (cf. Atzmueller et al., 2016a; Soldano et al., 2015), based on efficient pattern enumeration strategies (Soldano et al., 2017). In order to prevent the typical pattern explosion in (naive) pattern mining, *MinerLSD* employs closed patterns (cf. Atzmueller et al., 2019b) for a detailed discussion.

As input parameters, *MinerLSD* requires a graph; a set of items, e.g., corresponding to events, conditions, or measurements (as in the example above); and a dataset describing vertices as sets of items and a special operator, which focuses on a topological property, e.g., focusing on k -cores. The algorithm outputs the frequent pairs consisting of a specific pattern and a subgraph—corresponding to the associated (k -)core of the pattern. For pattern selection and ranking, *MinerLSD*

applies the local modularity quality function. As a simplified intuition on this, this quality function favors patterns having many (more) connections in the subgraph as expected by chance.


9.4.5 Application Example

For an example of an industrial use case, we refer to Atzmueller et al. (2019a) regarding a framework for human-centered exploration of event log data. We will briefly summarize the use case below, following the presentation by Atzmueller et al. (2019a): The dataset applied there is a real-world dataset provided by the company ABB in an anonymized version—for research. It provides a real-world event log of an industrial process, capturing about 4 million events on several production lines and production cells. For analysis, the event log is transformed into an attributed graph, where we utilize information about the event type, timestamp, and some descriptive information about the event. Using the event log, event sequences can be constructed, by ordering events in time. Then the graph can be constructed using the sequences, where two nodes (denoting events) are connected if the respective events occur one after another. After that, the graph can optionally be further summarized applying clustering, etc. (cf. Stefan Bloemheugel & Atzmueller, 2019). The graph is then labeled by adding properties/labels using additional information about the event, for example, the type of the event or the respective production line the event happened in. Given this representation, local pattern mining is applied in order to extract “interesting” patterns, which can help in identifying anomalies, diagnostics, process optimization, and general data exploration. Graph patterns can then always be inspected in context of the complete graph.

In the following, we summarize some results of Atzmueller et al. (2019a) and refer to the latter for more details and a comprehensive discussion. In the context of our application example, the graph pattern *warning* AND *cell6* AND *line20*, e.g., indicates *warning* conditions occurring in production lines 6 and 20. Thus, the pattern relates to a specific subgraph in the complete event log graph, capturing nodes which are labeled with the specific elements of the graph pattern. Then, a sequence which can be extracted from this attributed (sub-)graph is

71414 : *ConcurrentChangesOfSignalValue* → 80002 : *UserDefinedEvent3*

which indicates a certain problem and its respective (root) cause. Please note that in the data, the “user-defined events” (e.g., *UserDefinedEvent3*) actually indicate very interesting events, which however we cannot report due to anonymization. Exploiting the applied graph representation, we can then directly inspect the pattern and the respective sequences in context of the graph, e.g., for identifying and exploring the pattern (and its covered events) in context.

For the event “80002:UserDefinedEvent3,” for example, we can create a graph as shown in  Fig. 9.12, visualizing the paths leading to that event within a specific timeframe (such as a day). The frequency of a path is visualized using the size of the edges and nodes, respectively.



■ **Fig. 9.12** Example graph (Atzmueller et al., 2019a) with respect to the event “80002:UserDefinedEvent3.” The graph is compiled from the respective sequential log data, i.e., preceding and consecutive events. Then, given a specific event, all paths to this event can be analyzed in more detail. We refer to Atzmueller et al. (2019a) for a detailed discussion

9.4.6 Summary

Local pattern mining on complex graph data is a versatile and flexible approach for industrial data analytics. Its specific advantage is its interpretability and explainability of the patterns, both for predictive and for descriptive approaches. The interestingness can be flexibly defined using a quality function, and the discovered patterns can always be inspected in context. This allows for computational sense-making on complex industrial data for a wide range of applications.

Conclusion

The increasing amount and complexity of data ask for advanced analytics methods in order to make sense of the data and to extract valuable information and knowledge. Depending on the task at hand, several different methods can be applied to

analyze the gathered data. In this chapter, we provided such methods and their practical application focusing on data from industrial applications, i.e., Industry 4.0 and Internet of Things, exemplified by time series data, sensor networks, as well as real-world log data. We focused on (1) fault diagnosis using machine learning—in particular deep learning methods, (2) graph signal processing extending a signal-driven analysis approach on complex relational domains provided by graphs and networks, and (3) finally, using local pattern mining on such richly structured graph-based representations.

In particular, industrial machinery requires a maintenance strategy to prevent the number of breakdowns and therefore reduce the amount of downtime of the machine. If the data from sensors represent the condition of a certain machine, there are many analysis techniques that can be used to derive useful information from these signals. Fault diagnosis is a method used in condition-based maintenance to determine the state of equipment, locate the origin of faults, and forecast fault development over time. In the last decade, fault diagnosis has become more and more data driven due to improvements in computational power and gathering of large-scale data. In addition, in the last years, data-driven fault diagnosis steered more towards deep learning applications because of their ability to automatically extract features from the data. Automated feature extractors combined with classification are recommended to use for diagnosing fault conditions. Especially the use of one-dimensional CNNs has drastically improved the performance in fault diagnosis. These models save time in complex feature extraction and are often less prone to errors, as we have sketched in this chapter. In addition, they are able to handle noisy data well and can be updated to changing environments. Nonetheless, we need to keep in mind that most of these models perform most optimally in supervised settings where every fault condition is known, something that still remains challenging to acquire in real-world situations such as industrial applications.

When the data contains signals from nonuniform domains, graph signal processing can be employed. In general, GSP is a promising solution for various challenges using heterogeneous data. It has already shown its value in several different research areas. In particular, by using graphs, GSP can naturally account for irregular data relations. Then, after successfully modeling a particular problem as a GSP problem, this framework captures both techniques from the fields of signal processing and graph theory. We introduced basic concepts and methods and demonstrated those in the context of specific examples.

Finally, for identifying patterns in the data, we proposed local pattern mining for identifying interesting (descriptive) structures such as exceptional subgroups captured via interpretable rules. Pattern mining approaches can be applied for predictive as well as descriptive approaches and allow for powerful modeling options. In particular, regarding complex representations such as graphs and networks, pattern mining on attributed graphs specifically aims at getting a description-oriented view on the pattern, making them interpretable and explainable. A specific instance of local pattern mining on attributed graphs is given by the *MinerLSD* algorithm, which we briefly described and for which we also summarized an example case using a real-world dataset. Ultimately, such approaches enable computational sensemak-

ing—as computational methods for *analysis*, *interpretation*, and *intelligent decision support*—which are in particular important for complex industrial applications.

Overall, in this chapter, we have presented a diverse set of powerful methods for a wide range of application scenarios, which have already demonstrated their impact in several prominent use cases and thus provide the potential for further successful application onto advanced analytics on complex industrial data.

Take-Home Messages

1. Fault diagnosis:
 - Over the last decade, fault diagnosis has changed from physics-based techniques towards more intelligent data-driven approaches. These are mostly applied in a supervised manner.
 - Data for fault diagnosis can take many forms, e.g., sequential or rich relational representations. Typical scenarios for fault diagnosis, for example, rely on time series data acquired using sensors.
 - Data-driven approaches using deep learning, such as one-dimensional CNNs, have proven to be more powerful in fault diagnosis than traditional machine learning with manual feature extraction.
2. Graph signal processing:
 - Graph signal processing enables powerful techniques of signal processing on richly structured representations such as networks and graphs for modeling complex data.
 - Typical application scenarios include sampling, monitoring, and filtering of graph signals.
 - In particular, for large amounts of heterogeneous data, e.g., for large sensor networks, graph signal processing proved to be an effective method for a large number of applications.
3. Pattern mining on complex graph data:
 - Local pattern mining provides a powerful framework for both predictive and descriptive analytics focusing on locally interesting structures (“nuggets”) in the data.
 - Methods on feature-rich representations such as graphs and local pattern mining can be directly applied for obtaining patterns in terms of features as well as structure.
 - One particular advantage of local patterns is their interpretability and explainability in order to allow human-centered approaches, leading to computational sensemaking.

References

-
- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, & C. Zaniolo (Eds.), *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)* (pp. 487–499). Morgan Kaufmann.

- Alsyouf, I. (2007). The role of maintenance in improving companies' productivity and profitability. *International Journal of Production Economics*, 105(1), 70–78.
- Angelo Medeiros Fonini, P. (2019). *A didactic introduction to graph signal processing techniques and applications*. Instituto Alberto Luiz Coimbra.
- Arts, J. (2017). *Maintenance modeling and optimization* (Vol. 526). TU/e Eindhoven.
- Atzmueller, M. (2015). Subgroup discovery. *WIREs DMKD*, 5(1), 35–49.
- Atzmueller, M. (2016). Detecting community patterns capturing exceptional link trails. In *Proceedings of the IEEE/ACM ASONAM*. IEEE.
- Atzmueller, M. (2018). Declarative aspects in explicative data mining for computational sensemaking. In *Proceedings of the International Conference on Declarative Programming (DECLARE)*. Springer.
- Atzmueller, M., Arnu, D., & Schmidt, A. (2017). Anomaly detection and structural analysis in industrial production environments. In *Proceedings of the International Data Science Conference (IDSC 2017)*, Salzburg, Austria.
- Atzmueller, M., Bloemheugel, S., & Kloeppe, B. (2019a). A framework for human-centered exploration of complex event log graphs. In *Proceedings of the International Conference on Discovery Science (DS 2019)*. Springer.
- Atzmueller, M., Doerfel, S., & Mitzlaff, F. (2016a). Description-oriented community detection using exhaustive subgroup discovery. *Information Sciences*, 329, 965–984.
- Atzmueller, M., & Kloeppe, B. (2018). Mining attributed interaction networks on industrial event logs. In *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning*. Springer.
- Atzmueller, M., Kloeppe, B., Mawla, H. A., Jäschke, B., Hollender, M., Graube, M., Arnu, D., Schmidt, A., Heinze, S., Schorer, L., Kroll, A., Stumme, G., & Urbas, L. (2016b). Big data analytics for proactive industrial decision support. *ATP Edition*, 58(9).
- Atzmueller, M., & Lemmerich, F. (2009). Fast subgroup discovery for continuous target concepts. In *Proceedings of the 18th International Symposium on Methodologies for Intelligent Systems (ISMIS 2009)* (LNCS) (Vol. 5722, pp. 1–15). Springer.
- Atzmueller, M., & Lemmerich, F. (2012). VIKAMINE—Open-source subgroup discovery, pattern mining, and analytics. In *Proceedings of the ECML/PKDD*. Springer.
- Atzmueller, M., Puppe, F., & Buscher, H. P. (2005). Profiling examiners using intelligent subgroup mining. In *Proceedings of the IDAMAP, Aberdeen, Scotland* (pp. 46–51).
- Atzmueller, M., Soldano, H., Santini, G., & Bouthinon, D. (2019b). MinerLSD: Efficient mining of local patterns on attributed networks. *Applied Network Science*, 4(43).
- Atzmueller, M., & Sternberg, E. (2017). Mixed-initiative feature engineering using knowledge graphs. In *Proceedings of the 9th International Conference on Knowledge Capture (K-CAP)*. ACM Press.
- Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer Networks*, 54(15), 2787–2805.
- Bloemheugel, S., Van den Hoogen, J., & Atzmueller, M. (2020). Graph signal processing on complex networks for structural health monitoring. In *Proceedings of the Complex Networks*. Springer.
- Case School of Engineering. (n.d.). CWRU dataset; Case Western Reserve University bearing data center. Retrieved from <https://cseggroups.case.edu/bearingdatacenter/home>
- Chow, M., Mangum, P. M., & Yee, S. O. (1991). A neural network approach to real-time condition monitoring of induction motors. *IEEE Transactions on Industrial Electronics*, 38(6), 448–453.
- Cococcioni, M., Lazerini, B., & Volpi, S. L. (2013). Robust diagnosis of rolling element bearings based on classification techniques. *IEEE Transactions on Industrial Informatics*, 9(4), 2256–2263.
- Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 1–34). AAAI Press.
- Folmer, J., Kirchen, I., Trunzer, E., Vogel-Heuser, B., Pötter, T., Graube, M., Heinze, S., Urbas, L., Atzmueller, M., & Arnu, D. (2017). Big and smart data—Challenges in the process industries. ATP Edition (pp. 1–2).
- Gebhardt, J., Froese, T., Krüger, A., Appel, J., Benner, R., Hammer, M., Altermann, A., Hochrein, T., Kugler, C., Jatzkowski, P., Gloy, Y. S., Saggiomo, M., Roth, R., Elixmann, I., Tapken, H., Weber, W., Atzmueller, M., Garcke, J., Pielmeier, J., Rosen, R., & Tercan, H. (2016). *Status report*:

- Chances with big data—Best practice. Tech. Rep.* VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik.
- Group, M. R. W., et al. (1985). Report of large motor reliability survey of industrial and commercial installations, part I. *IEEE Transactions on Industry Applications*, 1(4), 865–872.
- Hajnayeb, A., Ghasemloonia, A., Khadem, S., & Moradi, M. (2011). Application and comparison of an ANN-based feature selection method and the genetic algorithm in gear-box fault diagnosis. *Expert Systems with Applications*, 38(8), 10205–10209.
- Hoang, D. T., & Kang, H. J. (2019). Rolling element bearing fault diagnosis using convolutional neural network and vibration image. *Cognitive Systems Research*, 53, 42–50.
- Huang, J., Hu, X., & Yang, F. (2011). Support vector machine with genetic algorithm for machinery fault diagnosis of high voltage circuit breaker. *Measurement*, 44(6), 1018–1027.
- Jardine, A. K., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7), 1483–1510.
- Jin, N., Flach, P., Wilcox, T., Sellman, R., Thumim, J., & Knobbe, A. (2014). Subgroup discovery in smart electricity meter data. *IEEE Transactions on Industrial Informatics*, 10(2), 1327–1336.
- Jin, X., Zhao, M., Chow, T. W., & Pecht, M. (2013). Motor bearing fault diagnosis using trace ratio linear discriminant analysis. *IEEE Transactions on Industrial Electronics*, 61(5), 2441–2451.
- Jing, L., Zhao, M., Li, P., & Xu, X. (2017). A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox. *Measurement*, 111, 1–10.
- Klösgen, W. (1996). Explora: A multipattern and multistrategy discovery assistant. In *Advances in knowledge discovery and data mining* (pp. 249–271). AAAI Press.
- Knobbe, A. J., Cremilleux, B., Fürnkranz, J., & Scholz, M. (2008). From local patterns to global models: The LeGo approach to data mining. In *From Local Patterns to Global Models: Proceedings of the ECML/PKDD-08 Workshop (LeGo-08)* (pp. 1–16).
- Konar, P., & Chattopadhyay, P. (2011). Bearing fault detection of induction motor using wavelet and support vector machines (SVMS). *Applied Soft Computing*, 11(6), 4203–4211.
- Lavrac, N. (2005). Subgroup discovery techniques and applications. In T. B. Ho, D. W. Cheung, & H. Liu (Eds.), *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (LNAI)* (Vol. 3518). Springer.
- Lei, Y., Jia, F., Lin, J., Xing, S., & Ding, S. X. (2016). An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. *IEEE Transactions on Industrial Electronics*, 63(5), 3137–3147.
- Lemmerich, F., Becker, M., & Atzmueller, M. (2012). Generic pattern trees for exhaustive exceptional model mining. In *Proceedings of the ECML/PKDD*. Springer.
- Liu, H., Zhou, J., Zheng, Y., Jiang, W., & Zhang, Y. (2018a). Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders. *ISA Transactions*, 77, 167–178.
- Liu, R., Yang, B., Zio, E., & Chen, X. (2018b). Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mechanical Systems and Signal Processing*, 108, 33–47.
- Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. *Journal of Industrial Information Integration*, 6, 1–10.
- Malhi, A., & Gao, R. X. (2004). PCA-based feature selection scheme for machine defect classification. *IEEE Transactions on Instrumentation and Measurement*, 53(6), 1517–1525.
- Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., & Shroff, G. (2016). LSTM-based encoder-decoder for multi-sensor anomaly detection. arXiv preprint arXiv:1607.00148.
- Meng, Z., Zhan, X., Li, J., & Pan, Z. (2018). An enhancement denoising autoencoder for rolling bearing fault diagnosis. *Measurement*, 130, 448–454.
- Miao, S., Veerman, R., Koenders, E., & Knobbe, A. (2013). Modal analysis of a concrete highway bridge: Structural calculations and vibration-based results. In *Proceedings of the Conference on Structural Health Monitoring of Intelligent Infrastructure, Hongkong*.
- Mobley, R. K. (2002). *An introduction to predictive maintenance*. Elsevier.
- Morik, K. (2002). Detecting interesting instances. In D. Hand, N. Adams, & R. Bolton (Eds.), *Pattern detection and discovery (LNCS)* (Vol. 2447, pp. 13–23). Springer.

- Morik, K., Boulicaut, J., & Siebes, A. (Eds.). (2005). *Local Pattern Detection, International Seminar, Dagstuhl Castle, Germany, April 12–16, 2004, Revised Selected Papers* (LNCS) (Vol. 3539). Springer.
- Natu, M., & Palshikar, G. (2014). Interesting subset discovery and its application on service processes. In K. Yada (Ed.), *Data mining for service* (Studies in big data) (Vol. 3, pp. 245–269). Springer.
- Pandya, D., Upadhyay, S., & Harsha, S. P. (2013). Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN. *Expert Systems with Applications*, 40(10), 4137–4145.
- Santos, P., Villa, L. F., Reñones, A., Bustillo, A., & Maudes, J. (2015). An SVM-based solution for fault detection in wind turbines. *Sensors*, 15(3), 5627–5648.
- Scarf, P. (2007). A framework for condition monitoring and condition based maintenance. *Quality Technology & Quantitative Management*, 4(2), 301–312.
- Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, 5, 269–287.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- Soldano, H., Santini, G., & Bouthinon, D. (2015). Local knowledge discovery in attributed graphs. In *Proceedings of the ICTAI* (pp. 250–257). IEEE.
- Soldano, H., Santini, G., Bouthinon, D., & Lazega, E. (2017). Hub-authority cores and attributed directed network mining. In *Proceedings of the ICTAI* (pp. 1120–1127). IEEE.
- Stankovic, L., Mandic, D., Dakovic, M., Brajovic, M., Scalzo, B., & Constantinides, T. (2019a). Graph signal processing—Part I: Graphs, graph spectra, and spectral clustering. arXiv preprint arXiv:1907.03467.
- Stankovic, L., Mandic, D. P., Dakovic, M., Kisl, I., Sejdic, E., & Constantinides, A. G. (2019b). Understanding the basis of graph signal processing via an intuitive example-driven approach [lecture notes]. *IEEE Signal Processing Magazine*, 36(6), 133–145.
- Stefan Bloemheugel, B. K., & Atzmueller, M. (2019). Graph summarization for computational sense-making on complex industrial event logs. In *Proceedings of the Workshop on Methods for Interpretation of Industrial Event Logs, International Conference on Business Process Management, Vienna*.
- Sternberg, E., & Atzmueller, M. (2018). Knowledge-based mining of exceptional patterns in logistics data: Approaches and experiences in an industry 4.0 context. In *Proceedings of the 24th International Symposium on Methodologies for Intelligent Systems (ISMIS)* (LNCS). Springer.
- Sun, W., Shao, S., Zhao, R., Yan, R., Zhang, X., & Chen, X. (2016). A sparse auto-encoder-based deep neural network approach for induction motor faults classification. *Measurement*, 89, 171–178.
- van den Hoogen, J., Bloemheugel, S., & Atzmueller, M. (2020). An improved wide-kernel CNN for classifying multivariate signals in fault diagnosis. In *ICDMW*. IEEE.
- Wang, Z., Zhang, Q., Xiong, J., Xiao, M., Sun, G., & He, J. (2017). Fault diagnosis of a rolling bearing using wavelet packet denoising and random forests. *IEEE Sensors Journal*, 17(17), 5581–5588.
- Wortmann, F., & Flüchter, K. (2015). Internet of things. *Business & Information Systems Engineering*, 57(3), 221–224.
- Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery* (pp. 78–87). Springer.
- Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1), 97–107.
- Xu, L. D., Xu, E. L., & Li, L. (2018). Industry 4.0: State of the art and future trends. *International Journal of Production Research*, 56(8), 2941–2962.
- Yin, S., Li, X., Gao, H., & Kaynak, O. (2014). Data-based techniques focused on modern industry: An overview. *IEEE Transactions on Industrial Electronics*, 62(1), 657–667.
- You, D., Gao, X., & Katayama, S. (2014). WPD-PCA-based laser welding process monitoring and defects diagnosis by using FNN and SVM. *IEEE Transactions on Industrial Electronics*, 62(1), 628–636.

- Zhang, A., Li, S., Cui, Y., Yang, W., Dong, R., & Hu, J. (2019). Limited data rolling bearing fault diagnosis with few-shot learning. *IEEE Access*, 7, 110895–110904.
- Zhang, W., Peng, G., Li, C., Chen, Y., & Zhang, Z. (2017). A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors*, 17(2), 425.
- Zhang, X., Xu, R., Kwan, C., Liang, S. Y., Xie, Q., & Haynes, L. (2005). An integrated approach to bearing fault diagnostics and prognostics. In *Proceedings of the 2005, American Control Conference* (pp. 2750–2755). IEEE.
- Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115, 213–237.
- Zhou, W., Habetler, T. G., & Harley, R. G. (2007). Bearing condition monitoring methods for electric machines: A general review. *2007 IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics and Drives* (pp. 3–6). IEEE.



Data Analytics in Action

*Gerard Schouten, Giuseppe Arena,
Frederique van Leeuwen, Petra Heck,
Joris Mulder, Rick Aalbers,
Roger Leenders,
and Florian Böing-Messing*

Contents

- 10.1 Introduction – 207**
- 10.2 BagsID: AI-Powered Software System to Reidentify Baggage – 208**
 - 10.2.1 Business Proposition – 209
 - 10.2.2 System Overview – 210
 - 10.2.3 AI Engine – 212
 - 10.2.4 Software Engineering Aspects – 214

Gerard Schouten and Petra Heck contributed the *BagsID* case study in ► Sect. 10.2. Giuseppe Arena, Joris Mulder, Rick Aalbers, and Roger Leenders contributed the *Understanding employee communication with longitudinal social network analysis of email flows* case study in ► Sect. 10.3. Frederique van Leeuwen contributed the *Using vehicle sensor data for Pay-How-You-Drive insurance* case study in ► Sect. 10.4. Florian Böing-Messing provided feedback on the case studies and helped shape the chapter.

Shared first authorship: The first authors of the case studies (Gerard Schouten, Giuseppe Arena, and Frederique C. A. van Leeuwen) are co-first authors of this chapter.

10.3	Understanding Employee Communication with Longitudinal Social Network Analysis of Email Flows – 216
10.3.1	Digital Innovation Communication Networks – 217
10.3.2	The Relational Event Modeling Framework – 218
10.4	Using Vehicle Sensor Data for Pay-How-You-Drive Insurance – 223
10.4.1	Time Series – 224
10.4.2	Driving Behavior Analysis – 226
	References – 231

Learning Objectives

- Understand the characteristics of reidentification deep learning and how this technique can be applied to promptly identify mishandled luggage at airports.
- Understand information sharing dynamics among employees of an organization by means of longitudinal social network analysis.
- Understand what controller area network bus technology is and what the possibilities are with respect to driving behavior analysis.

10.1 Introduction

In this chapter, we present three case studies that cover a broad spectrum of problems and methods in the area of data analytics. We begin with the BagsID case study in ► Sect. 10.2, which is carried out in collaboration with Vanderlande, PTTTRNS.ai, and Eindhoven Airport. The case study illustrates how computer vision and reidentification deep learning can be applied to reidentify mishandled luggage at airports. The approach uses Re-ID neural networks that can be trained to predict the degree of similarity between individual objects (pieces of luggage in this case) rather than categorizing objects. The BagsID case study emphasizes that getting robust AI-powered software systems into production is quite different from building proof-of-concept AI prototypes.

The second case study in ► Sect. 10.3 analyzes the effect of a business intervention strategy on the employees of a multinational service company. More specifically, a European branch of the company implemented multiple interventions aimed at stimulating its employees to open their minds to innovation. The efficacy of these interventions can be assessed by investigating how they shape communication and discussions about innovation between the employees. To this end, the case study analyzes email communication between employees using longitudinal social network analysis.

The third case study in ► Sect. 10.4 considers how vehicle sensor data can be used for insurance purposes. Through the standardization of the controller area network bus technology in modern cars, a large amount of sensor data is generated every day. This enables insurance industries to obtain more reliable and direct characterizations of driving styles for their Pay-How-You-Drive models. If used wisely, accidents can be prevented instead of restituted. This is beneficial for both the customers and the insurance industry.

10.2 BagsID: AI-Powered Software System to Reidentify Baggage

BagsID¹ is a Dutch company that aims at improving baggage handling systems worldwide by using the bag itself as an ID. At the core of their technology stack, they employ computer vision, powered by deep learning. The company is currently moving towards initial deployment to showcase its potential, in close collaboration with three organizations. These organizations are (1) Vanderlande, the global market leader for logistic process automation at airports; (2) PTTRNS.ai, a software company that specializes in developing and integrating artificial intelligence (AI) solutions to accelerate digital innovation; and (3) Eindhoven Airport. A joint project is set up at Eindhoven Airport to prove the proposition that baggage can be identified with state-of-the-art vision AI. A scale-up of the system to other European airports, and in a later stage to airports worldwide, is foreseen. This case study describes one possible application of the BagsID reidentification system: that of mishandled baggage. To illustrate this application, we begin this case study with a short user story:

► Example

March 4, 2020: Just after midnight, Jane lands at Tromsø Airport with the last flight from Oslo Gardermoen. A few hours ago, she departed from Amsterdam Schiphol. After descending from the aircraft staircase and a short walk outdoors on the slippery platform, she enters the terminal. The arrival hall is divided into two public spaces. The first area is dominated by a conveyor belt to pick up luggage, and the other area hosts a few offices of car rental companies and holds the exit doors as well as a few uncomfortable seats. As in most airports in northern Europe, the hall is decorated with huge posters showing local wildlife and snowy winterscapes with northern lights skies. The conveyor belt runs already, and soon the first suitcases appear. One by one, the passengers of SAS flight SK4438 pick up their bags and leave the hall facing the freezing cold. After 20 min, the conveyor belt stops and all fellow travelers are gone. Jane's suitcase did not appear. She is all alone at the completely deserted airport. ◀

This is no fantasy. Regular travelers could easily feel the unease of the situation sketched above. Being the last person at the airport's conveyor belt and slowly realizing that your bag is not coming is a traveler's nightmare. Better baggage handling is not just about keeping passengers happy. Claims due to lost or mishandled luggage cost airlines around the world 2.4 billion US dollars in 2018 (Air Transport IT, 2019). Over the past few years, most airlines have introduced a baggage track and trace at key points in the journey—check-in, loading onto the aircraft, transfers, and arrival—in response to IATA's Resolution 753 (IATA, 2020). Now, most bags are tracked from start to finish. Despite these efforts, the number of mishandled bags rose to 24.8 million in 2018, a figure that translates to 5.7 bags per 1000

1 ► <https://bagsid.com>.

passengers (Air Transport IT, 2019). Of all mishandled bags in 2018, 77% is delayed, about 17% is seriously damaged or pilfered, and 5% is stolen. Transferring bags from one aircraft to another, or one airline to another, is a major cause for delays of flights as well as late delivery of luggage.

This case study shows work in progress. It illustrates how an initial business idea is translated into a software-based AI solution. The BagsID case is beyond schoolbook AI. It clearly demonstrates that machine learning algorithms cannot be applied “just like that” to practical cases. We argue that a componentized extendable architecture, an iterative planning approach, and a solid software engineering process for AI embodiment are all needed for successfully building professional and maintainable AI-powered software solutions.

10.2.1 Business Proposition

The current handling of baggage depends on stickers and paper tags, which are wrapped around handles of suitcases, trolleys, or other luggage items (ski boxes, bike bags, etc.) at check-in. These stickers and tags are labeled with a printed barcode and a three-letter abbreviation of the destination airport.² The barcode is uniquely coupled to the traveler. At depots where mishandled baggage is gathered, a human-centric exception handling process—i.e., people scanning the tags with line-of-sight barcode readers³ and initiating logistic actions—is in place to identify the bags and resend them to their legitimate owners, either to the destination airport or to their home address. Serious problems with the current track and trace functionality arise when these tags have become unreadable or are even detached from the luggage. Relying on physically attached labels makes the current system inherently vulnerable.

In recent years, vision AI has drastically improved (Krizhevsky et al., 2012; LeCun et al., 2015; Howard et al., 2017; Canziani et al., 2017). With state-of-the-art deep learning, using convolutional neural networks (CNNs) as a backbone, a reliable machinery can be built to detect and identify objects in images. The ubiquitous use of face recognition, from unlocking your smartphone to crowd security management, is probably the best known example of this progress (Ye et al., 2020). So, why not apply this technology to reidentify baggage? In this way, the bag itself can become an ID. It removes the abovementioned bottleneck, that is, the problem of ripped-off tags. This business opportunity of suitcase fingerprinting has the potential to further improve efficiency and reduce the chances of a bag being mishandled. It saves not only money for airlines but also agony and discomfort for travelers. It is to be expected that the magic number of 5.7 mishandled bags per 1000 items can significantly be lowered by implementing this idea.

2 Each airport in the world is characterized with a unique three-letter combination.

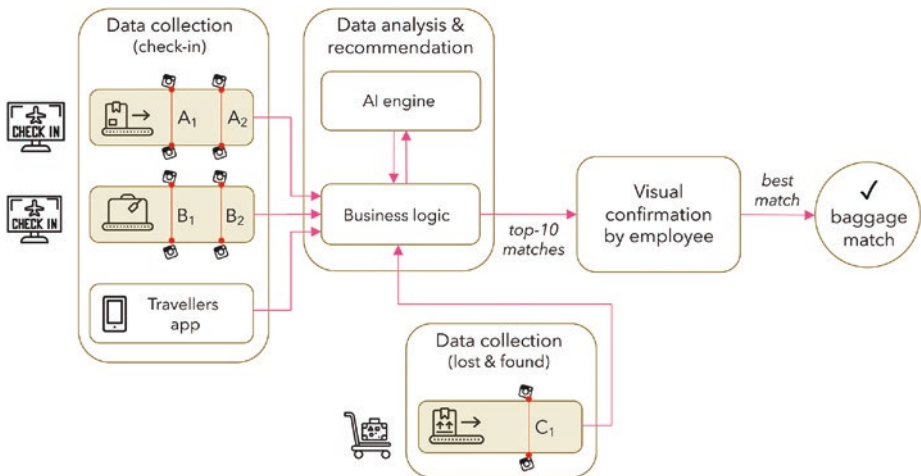
3 RF-ID tags and near-field RF-ID scanners could solve some of the issues, but this solution is too expensive.

10.2.2 System Overview

The task of the BagsID system is simple: Find for each mishandled luggage item its legitimate owner as fast as possible. The technical concept to do this is object fingerprinting, that is, find for a mishandled item the matching or corresponding item at check-in. A simplified overview is depicted in ■ Fig. 10.1. This system is divided into three major building blocks: (1) data collection at check-in, (2) data collection of lost and found bags, and (3) data analysis and recommendation. Note that there is a human in the loop for the final visual inspection. An airport or airline baggage handler is still needed to narrow down the top 10 matches of the system to an accepted best match.

Data Collection and Storage The hardware setup of the system consists of multiple camera-equipped conveyor belts. Directly after either self check-in or desk check-in, the luggage that is put on the belt is registered with two multi-camera systems that are placed just after one another. The registration establishes a link between captured suitcase images and a boarding pass. The data of this imaging system is optionally enriched with luggage information that can be entered via a traveler's app (see below). Each multi-camera system photographs passing luggage from different viewpoints. In the simplified schematic system overview of ■ Fig. 10.1, two camera-equipped conveyor belts are envisioned at check-in, and only one camera-equipped conveyor belt is present to handle lost-and-found luggage. The multi-camera systems are indicated with A_1 and A_2 for conveyor belt A , B_1 and B_2 for conveyor belt B , and C_1 for conveyor belt C . When a lost or mishandled bag with an unknown destination (because of an unreadable or ripped-off tag) is found, it will be scanned with the multi-camera system C_1 . These query images are compared with images in the gallery set that were previously captured at the check-in.

10



■ Fig. 10.1 System overview of the baggage reidentification system. Happy flow: for a lost bag, the system will retrieve the top 10 matching luggage items, based on images which were taken at check-in. A final visual check is done by an airport or airline employee. (Author's own)

Business Logic and AI Engine The data analysis building block consists of two modules. The AI engine is responsible for finding, for each mishandled bag, the best K matches from the data collected at check-in. It will be discussed in more detail in the next section. The rule-based business logic module will be connected to the airport flight schedule system and takes into account various logical time-related constraints and statistics (e.g., performance monitoring). For instance, flights might be delayed or cancelled. The task of the business logic module is fourfold: (1) narrow down the search possibilities for the AI engine up front, that is, establish the gallery set; (2) filter out matches that are logically not possible; (3) monitor performance of the AI engine; and (4) inform airport personnel what can best be done with a positively reidentified bag. Can it still be boarded at the intended airplane in time? If not, what are the best options to send it to the final destination?

Traveler's App The system also comes with a user-friendly smartphone app for travelers. It is an extension of the onboarding process and will be developed in the second phase of the project. The idea of the traveler's app is to enrich the image information that is captured at the airport's check-in. Once travelers have registered for this app, they can create and maintain a list of personal luggage items. For each bag or suitcase, they can specify values for a number of characteristic attributes, like luggage type (suitcase, trolley, backpack, guitar case, ski box, etc.), brand, color, presence of a lock, numbers of wheels, hardcover or soft side, and presence of damage marks such as scratches. These attributes correspond to the IATA baggage ID chart.⁴ This information helps to identify unique luggage items. The app is optional, that is, the system should also work if this information is not, or only partly, available.

► Example

October 28, 2021: Jane attaches the printed tags to her red old suitcase and puts it on the conveyor belt at the luggage drop-off. The coronavirus pandemic is over, and she looks forward to a short autumn break in the Mediterranean. Transavia flight HV6607 to Faro is about to leave in an hour from Eindhoven Airport. The advantage of regional airports is that the waiting time is limited. After a cappuccino, Jane buys a magazine and walks to the gate. She looks out of the window and recognizes her suitcase on one of those special airport vehicles. The red suitcase is loaded to the waiting plane. What Jane did not know was that her suitcase fell from the conveyor belt and that the loosely attached tag was ripped off. An airport employee picked up the untagged suitcase and brought it to the lost-and-found depot. Luckily, Jane—as a frequent flyer—had registered her luggage item with the BagsID traveler's app. The BagsID system was able to show ten possible matches within 30 s based on the photos taken and the earlier registered suitcase details (such as the scratch near the handle). The best match linked the red suitcase to Jane. The airport employee confirmed this best match, and 15 min later, Jane's suitcase enters the waiting airplane that was prepared to leave to Faro in 25 min. One year ago, it was unthinkable to deliver a lost-and-found suitcase to the right airplane within such a short time frame. ◀

4 ► <https://www.iata.org/en/publications/store/baggage-id/>

10.2.3 AI Engine

CNNs are a known solution for categorizing images. These feed-forward neural networks are inspired by human vision. They can abstract from viewpoint and illumination variations and are able to capture the very essentials of objects that are present in images. However, *category-level* object classification—where two images are considered similar as long as they belong to the same semantic class of objects—is not sufficient for a *search-by-example* image application. Search by example requires a more fine-grained distinction between objects that belong to the same category (Wang et al., 2014). As a simplified and intuitive example, for classification, a “Red Samsonite Omni Spinner” (hardcover suitcase), “Green Travelpro Maxlite 5” (soft-side suitcase), “Black Karrimor Ridge 32” (outdoor backpack), “Delsey Luggage Helium Aero Blue” (hardcover trolley), and “Black Briggs & Riley Baseline Vista Print” (soft-side trolley) are all luggage items. For luggage reidentification (Re-ID) on the other hand, if the query image is characterized by the phrase “red hardcover 4-wheeled suitcase,” it is essential to rank the “Red Samsonite Omni Spinner” higher than the other gallery items. Stated more formally, the objective of the Re-ID AI engine is as follows: Given a query baggage item of interest, determine the K best recommendations from the luggage gallery set. The hypothesis is that the ranking of top- K matches contains the bag (captured by a different camera in another place at a distinct time) that corresponds to the query image.

10

Re-ID Neural Network Architecture The neural network architecture of the Re-ID AI engine that is able to generate a suitcase fingerprint is shown in Fig. 10.2. It consists of two parts: an encoder module and a reidentification module. This architecture is state of the art for reidentification learning or search-by-example problems (Ye et al., 2020; Wang et al., 2014). The encoder part can be seen as a feature engineering process. Captured images—i.e., “low-level” raw pixel data—are processed with

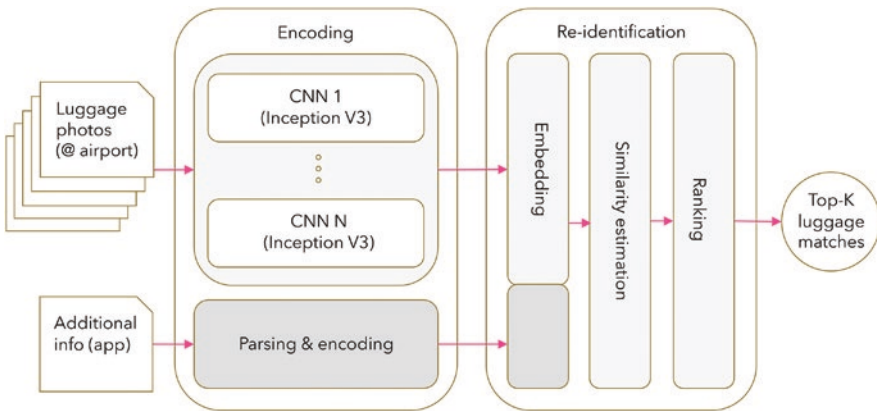


Fig. 10.2 Architecture of the AI engine. A top- K ranking is derived from N parallel CNN pipelines and app data. The app information (dark gray blocks) will be added in a later phase of the project. (Author’s own)

Google's Inception V3 CNN framework resulting in a number of feature maps or "high-level" encodings that are more suitable for visual tasks. Inception V3 is a widely used image recognition model. The model is the culmination of many ideas developed by multiple researchers over the years; it is made up of various building blocks, including convolutions, average pooling, max pooling, concatenation, dropouts, and fully connected layers. It is based on the original paper of Szegedy et al. (2015). In the proposed architecture, each camera viewpoint is coupled to a separate Inception V3 CNN encoding pipeline.

In the next step, these visual encodings are concatenated and combined with the information that travelers provide via the app to a so-called embedding layer. An embedding is a mapping of high-dimensional data, such as images (pixel data), to a vector. This vector is a relatively low-dimensional space that summarizes the relevant information in the data into a meaningful representation. Ideally, an embedding captures some of the semantics of the input by placing similar inputs close together in the embedding space. The embedding layer flattens, reduces, and normalizes the output of all CNNs (as well as the coded app information) to a fixed-size vector. In terms of the neural network, an embedding is just a hidden layer and is learned with backpropagation during the training process.

In practice, embeddings are often used to make recommendations or to rank possible match candidates, that is, to find nearest neighbors in the embedding space. To do this, a distance metric is needed. Several options are available for this, such as the standard euclidean distance, Manhattan distance, or a cosine similarity distance metric.⁵ A retrieved top 10 ranking list can then be obtained by sorting the calculated query-to-gallery similarity.

Training the Network: Triplets, Hinge Loss, Semiautomatic Labeling The standard CNN approach in supervised learning is to estimate a function $f(\cdot)$ that maps the entire set of input images as best as possible to probabilities for given category labels. This is done in the training phase by changing the weights of the CNN (usually a few million) in such a way that a so-called loss function is minimized. Usually, this is a cross-entropy loss or mean squared error between the CNN predictions and the actual labels. For Re-ID or ranking with deep learning, however, two accommodations are needed that go beyond this standard recipe (Wang et al., 2014; Hermans et al., 2017). First of all, it is common practice to train the network with triplets of input images. A triplet $t_i = (p_i, p_i^+, p_i^-)$ contains a query image p_i , a positive image p_i^+ , and a negative image p_i^- , where the positive image is more similar to the query image than the negative image (Wang et al., 2014).

Secondly, the loss function associated with ranking and triplets is a so-called hinge loss. It is defined as

$$l(p_i, p_i^+, p_i^-) = \max\{0, \Omega + D(g(p_i), g(p_i^+)) - D(g(p_i), g(p_i^-))\} \quad (10.1)$$

⁵ The latter is the dot product between the two normalized embeddings and ranges from -1 , most dissimilar, to $+1$, most similar.

where the function $g(\cdot)$ represents the embedding and D is a distance metric: in our case, the dot product between two normalized embeddings. As explained by Schroff et al. (2015), the hinge loss tries to bring the query image and the positive image close together in the embedding space and at the same time as far away as possible from the embedding of the negative image. As long as $D(g(p_i), g(p_i^-))$ is larger than $\Omega + D(g(p_i), g(p_i^+))$, there will be no gain for the algorithm to condense the query and positive image any further. The learning process boils down to finding the best embedding $g(\cdot)$ for generating fine-grained baggage sensitivity, that is, enabling similarity ranking. Note that the distance metric itself is given up front and not modified by the training process.

For the triplets, a semiautomatic labeling process will be bootstrapped, where a priori similarity information is exploited. This semiautomatic labeling process is extended with random triplet sampling from the large image database of stored luggage photographs combined with human labeling. A labeling service like the CloudFactory platform⁶ or the Amazon Mechanical Turk crowdsourcing marketplace⁷ can be used to obtain these human labels. Crowdsourcing is a good way to break down a manual, time-consuming task—such as labeling thousands of images—into smaller, more manageable “microtasks” to be completed by distributed workers over the Internet. Traditionally, tasks like this have been accomplished by hiring a large temporary workforce, which is time consuming, expensive, and difficult to scale or to undo. These platforms offer APIs to upload your data, to carefully describe the requested task, and to ask for specific skill levels.

10

Inference with the Network: Reidentify a Mishandled Bag Once the AI engine is trained, it can be used in inference mode, that is, in operation. The fixed weights of the model produce an embedding for a mishandled luggage item. This embedding will be compared with other embeddings of luggage items that are in the gallery. The business logic provides filters and other constraints for items that will be put in the gallery. Based on the chosen distance metric, a top- K ranking will be made of the most similar embeddings.

10.2.4 Software Engineering Aspects

The BagsID system includes several software components that interact with the AI engine: data collection software, business logic software, user interfaces, traveler’s app, etc. For the AI engine to be successful in a production environment with multiple airports and multiple camera systems per airport, it is of utmost importance that its deployment strategy is carefully designed and integrated with the deployment strategy for the other software components. This integrated approach is often referred to as MLOps.⁸

6 ► <https://www.cloudfactory.com>

7 ► <https://www.mturk.com>

8 MLOps is a practice for collaboration and communication between data scientists and operations professionals to help manage the production machine learning life cycle; see ► <https://en.wikipedia.org/wiki/MLOps>

The development of software systems with AI components has several intricacies (Heck, 2019) that also apply to the BagsID system. Some issues and questions that have to be addressed when designing the system are the following:

1. Collecting high-quality data is crucial for the success of the model training. That is why the BagsID system makes use of custom-built industry-grade camera systems (hardware and software). This ensures a constant image quality and robust recording from the luggage belts.
2. After the deployment of the trained AI model, it needs to be monitored for performance (so-called online testing and logging (Heck, 2020)) because with new images coming in, the generalization capability of the model may drop and periodical retraining might be needed.
3. When the system will be scaled up (i.e., will be installed at other airports), a multi-site deployment strategy will be needed. In particular, issues that have to be discussed and settled are when and how to introduce new models in the live systems and how to version both models and data.
4. Huge amounts of data will be collected. It needs to be decided how and where to store this (on premise or in the cloud?), for how long to keep the data, which privacy laws are applicable, how to be compliant with local legislation, if airports are willing to share data for better models, etc.
5. There needs to be a scalable way of serving the model to the BagsID system for inference purposes. The model needs to be decoupled from the rest of the system such that it can be more easily updated to new algorithms or new versions. It might be necessary to have multiple versions of the model running simultaneously, for example, for different countries. State-of-the-art software engineering practices will be used for this. It is planned to deploy the AI model(s) as a REST API with Docker containers in a cloud environment.

The project started with a data collection/data preparation phase (Rollins, 2015). For this, a first camera system is set up at Eindhoven Airport where real baggage is recorded. As said, the camera system is custom-built and also contains software to preprocess the recorded images. The collected images are used to train the AI model. The training is done using Jupyter Notebooks with Python and TensorFlow 2.0 in an AWS cloud environment. Amazon SageMaker is used to support the training process and deploy the trained models for testing purposes. Next to improving the model, time is spent on preparing the deployment phase of the project for both the AI engine and other software components.

Discussion and Conclusion

Reidentification learning is a challenging and fast-growing field within the computer vision community. Most Re-ID applications deal with human faces, persons, or vehicles. In this case study, it is applied to another use case: reidentification of luggage. A system based on this technology enables airports and airlines to provide more reliable information on the whereabouts of baggage at each step in the journey. The main message of this case study is that in practice AI innovations typically (1) combine or concatenate multiple known AI concepts in a specific setting that is new and

never tried before and (2) have a mixed design of known (well-established) algorithms (such as CNNs) and principles (such as embeddings) complemented with unique business rules that have to be derived from case-specific requirements. Like the wheels in a clockwork, hardware, software, and AI components should be synchronized and fit together smoothly. The design and implementation of these components should be balanced and tuned carefully. We would like to emphasize that taking an AI model into production and maintaining it demand a serious effort and might be as complicated as designing the model itself. We also see in practice that for various reasons—e.g., safety, security, accountability, and trust—AI-powered solutions often need a human in the loop.

10.3 Understanding Employee Communication with Longitudinal Social Network Analysis of Email Flows

Innovation is the spice of life for organizations and is generally seen as a requirement for long-term survival and attaining and sustaining above-average performance. Yet, innovation can be hard to accomplish.

In this case study, we consider the innovation struggle of a European branch of a multinational service company (referred to in the case study as STRATSERV). Innovation typically requires a company's employees to change the way they do their work, either by doing different things (such as providing a new service or engaging in new procedures) or by doing things differently (such as using new technology to do the work more efficiently). This means that, especially in service organizations, innovation can hardly be successful without the willingness of employees to change (the way they do) their work. This realization stimulated STRATSERV's management to attempt to open the minds of their employees to innovation. Hence, they organized various events where employees could suggest innovative ways of working, offered prizes for the best ideas, and provided resources to employees to explore their ideas further. In sum, the approach was to first open the minds of employees to the idea of innovation, stimulate the employees to come up with innovative suggestions, and then build on that joint openness to the innovation in order to implement new services and new procedures. Of course, this assumes that the minds of the STRATSERV employees would respond favorably and long-lasting to the company's innovative wishes.

Although the STRATSERV management believed in this approach, they also realized that they needed a way to test whether their approach was working. Did their efforts indeed create an innovation mindset in the heads of their employees and did that mindset last? Moreover, they wondered if all employees responded alike or whether the competitions, gatherings, newsletters, challenges, and other activities organized by the company's task force only affected certain employees but not others.

In this situation, it makes little sense to send out a survey to the employees, asking them whether they were thinking about innovation regularly. This would likely trigger socially acceptable answers and could not provide the detailed

insight into the effect of the activities that the company was looking for. In addition, surveys are poorly suited to monitor how employees respond over time, including repeated surveys. The company reached out for help to an external team of researchers. Below, we will show part of the analysis that was performed.

10.3.1 Digital Innovation Communication Networks

When employees discuss innovation, an innovation communication network emerges within the company. The structure and pervasiveness of this network are key indicators whether STRATSERV's approach is working. In addition, innovative activity is essentially a network activity (Aalbers & Leenders, 2016; Kratzer & Leenders, 2004; Leenders et al., 2003). Innovation is, by necessity, a collaborative effort. Existing knowledge and ideas merge into new combinations, and as formerly separated knowledge comes together, new knowledge emerges. Although the imagery of the lone inventor profoundly developing is appealing, it is an image rarely found in modern times. Innovation is a "team sport," where individuals work together in teams, teams work together in projects, organizations work together in alliances, and countries work together in international technology agendas. In fact, even the mythical lone inventor probably rarely operated in splendid isolation anyway, since it is likely that much of the inventor's inspiration came from interaction with other people or organizations, the financial resources may have been granted by banks or friends, the actual development of the product often involved the help of factories, and customers had to become involved in order to test the product for feasibility. No matter which (great) innovation one would look at, it is bound to be couched in network interaction of some sort (Leenders, 2016). In sum, an ideal approach to see if innovation was catching on as a core topic and activity inside STRATSERV was to measure how the innovation communication network developed.

Networks can be measured in a number of ways. The most common approach is to administer surveys to ask who communicates with whom. Alternatively, one could observe the interactions of employees throughout their working activities. These methods do not work in our case, since we wanted to follow the interactions of employees in real time for a full year. Alternative tools such as using video to see who interacts with whom or collecting data from proximity badges would not provide information on whether the conversation included innovation as a topic. Hence, the choice was made to analyze the email interaction between the employees over the course of a year.

Digital communication, in particular email, has become one of the most important means of communication in organizations. As email leaves digital traces about senders, receivers, and timing, these rich network data contain high-resolution information to understand how communication structures change when working teams reach deadlines, to understand new employee integration processes (and how these are affected by cultural differences and team compositions), or to under-

stand how ideas spread through a network of employees (and how this is affected by the actors' hierarchical positions, for example). Besides the academic/theoretical interest, these insights are also useful from a practical point of view as they can be used to optimize communication structures in deadline situations, they can be used to optimize the integration processes of new employees, and they can be used to reach all employees regarding certain working topics as fast as possible.

In this case study, we show one approach that can be used to study and understand how networks evolve over time, in real time, and how this knowledge can be leveraged in practice.

10.3.2 The Relational Event Modeling Framework

Description of the Data Our analysis focuses on the innovation communication networks in a European branch of STRATSERV. After developing and implementing procedures to ensure employee privacy and informed consent was received from the parties involved, we used text mining techniques to score the email messages on whether the exchanged text addressed innovation-related topics. The empirical data in this case study consist of a time-ordered sequence of $M = 1340$ email messages that were exchanged between 153 employees over the course of a year. An example of the data is given in ■ Table 10.1 where each row represents the 3-tuple (t_m, s_m, r_m) with, respectively, the time, the sender, and the receiver of the m th email in the sequence of emails $E = \{(t_1, s_1, r_1), \dots, (t_M, s_M, r_M)\}$.

We assume that email interaction is regulated and driven by factors that can depend either on workers' characteristics (e.g., one's status or outgoingness), on the dyadic characteristics of sender and receiver (e.g., hierarchy differences, co-location), on the history of workers' past interactions (e.g., the exchange of email that occurred in the past), or on the workers' location in the social structure (e.g., interaction with joint colleagues, norms of reciprocity). In particular, we will focus on modeling whether and how this email stream depends on working in the same building, the difference in hierarchy level between sender and receiver of the email, the tenure of the sender, the tendency of sender and receiver to continue to exchange email messages among each other (i.e., persistence or inertia), and the norms of

■ Table 10.1 Example of longitudinal network of emails

Time	Sender	Receiver
03 Jan 2010 08:21:33	Marco	Jane
03 Jan 2010 08:43:09	Jane	Marco
∅	∅	∅
31 Dec 2010 18:39:22	Paul	Jane

Compiled by authors

reciprocity between employees. Moreover, we allow a possible memory effect where recent email activity may have a relatively large effect on the future activity between actors.

The Model The novel modeling framework that is well suited to analyze time-to-event sequence data in networks is the so-called *relational event model* (REM) (Butts, 2008; Mulder & Leenders, 2019; Leenders et al., 2016). This framework aims to model the rate at which specific directed interaction (i.e., a given email being sent) between two actors (here: employees) occurs; in other words, we model the *emailing rate* among any pair of employees. In social network terms, such a pair is called a *dyad*. Within this framework, each email message constitutes a *relational event* characterized by the *sender* (s), who initiates the action (i.e., who sends the email); the *receiver* (r), to whom the action is targeted (i.e., who receives the email); and *time* (t), the exact time point at which the relational event occurs. At each time point in the sequence, 153 potential senders can send an email to 152 potential receivers (excluding email messages people send to themselves), which means that at any point in time $153 \times 152 = 23,256$ email dyads can potentially occur. The aim of the analysis is to model who sends an email message to whom at what point in time over the course of 1 year. Mathematically, the joint probability to model the whole sequence of emails is similar to the well-known event history model or survival model (Lawless, 2003; Cox, 1972).

In the REM, we model the rate at which an email is sent from a given sender to a given receiver at a given point in time as a loglinear model that (apart from the exponent that occurs in the equation) resembles the well-known linear regression structure. The model then takes into account every possible sender, every possible receiver, and every possible point in time, for the entire observation period. One of our substantive interests in this study is whether the emailing rates of employees depend only (or mainly) on the recent email interactions of the employees or whether they also take into account email exchanges that happened longer ago. This is important for STRATSERV, as it shows how long the effects of interventions last. If it turns out that employees mainly respond to innovation-related messages they received recently, and much less to messages received or exchanged longer ago, this is a sign that employees apparently need to be “reminded” of innovation constantly and that it has not become a routine part of their conversations.

In particular, we will investigate this for inertia and reciprocity (see ■ Table 10.2). In order to accomplish this, both the inertia and reciprocity variables are calculated according to two different event history lengths. For both variables, we include in the model a short-run version where only past events that occurred *until 30 days* before the time of the email are included (*recent past*) and a long-run version that includes the past events that occurred *more than 30 days* before the email was sent (*less recent past*) (cf. Quintane et al., 2013). A complete description of the variables used in our analysis can be found in ■ Table 10.2.

Model Comparison We estimate two models: in *Model 1*, all the variables in ■ Table 10.2 are embedded in the loglinear predictor; in *Model 2*, only the short-run and long-run versions of inertia and reciprocity are included. Via this model com-

Table 10.2 Predictor variables and their interpretations

Predictor variable	Description
ShortInertia	The number of messages a potential sender sent to a potential receiver in the last 30 days
LongInertia	The number of messages a potential sender sent to a potential receiver more than 30 days ago
ShortReciprocity	The number of messages a potential sender received from a potential receiver in the last 30 days
LongReciprocity	The number of messages a potential sender received from a potential receiver more than 30 days ago
SameBuilding	A binary variable which indicates whether potential sender and potential receiver work in the same building (1) or not (0)
DiffHierarchy	The hierarchical difference between the sender and receiver on a scale from 1 to 9
LogSenderTenure	The number of years a potential sender works in the organization on a log scale
Compiled by authors	

10

parison, we can learn whether a simpler model without exogenous effects may be enough for a good fit for the data. Considering the specification of Model 1, the email rate (λ) at time t_m for the dyad (sender, receiver) = (Marco, Jane) is

$$\lambda(t_m, \text{Marco, Jane}) = \exp\{\beta_{\text{Intercept}} + \beta_{\text{ShortInertia}} \text{ShortInertia}(t_m, \text{Marco, Jane}) + \beta_{\text{LongInertia}} \text{LongInertia}(t_m, \text{Marco, Jane}) + \beta_{\text{ShortReciprocity}} \text{ShortReciprocity}(t_m, \text{Marco, Jane}) + \beta_{\text{LongReciprocity}} \text{LongReciprocity}(t_m, \text{Marco, Jane}) + \beta_{\text{SameBuilding}} \text{SameBuilding}(\text{Marco, Jane}) + \beta_{\text{DiffHierarchy}} \text{DiffHierarchy}(\text{Marco, Jane}) + \beta_{\text{LogSenderTenure}} \text{LogSenderTenure}(\text{Marco})\} \quad (10.2)$$

where $\beta = (\beta_{\text{Intercept}}, \beta_{\text{ShortInertia}}, \beta_{\text{LongInertia}}, \beta_{\text{ShortReciprocity}}, \beta_{\text{LongReciprocity}}, \beta_{\text{SameBuilding}}, \beta_{\text{DiffHierarchy}}, \beta_{\text{LogSenderTenure}})$ is the vector of effects describing the impact of the variables on the rate of occurrence of an email being sent from a sender to a receiver. Positive effects (negative effects) imply that as the variable increases in value, it increases (decreases) the email rate. As regards Model 2, the rate of an email sent from Marco to Jane at time t_m becomes

$$\lambda(t_m, \text{Marco, Jane}) = \exp\{\beta_{\text{Intercept}} + \beta_{\text{ShortInertia}} \text{ShortInertia}(t_m, \text{Marco, Jane}) + \beta_{\text{LongInertia}} \text{LongInertia}(t_m, \text{Marco, Jane}) + \beta_{\text{ShortReciprocity}} \text{ShortReciprocity}(t_m, \text{Marco, Jane}) + \beta_{\text{LongReciprocity}} \text{LongReciprocity}(t_m, \text{Marco, Jane})\}. \quad (10.3)$$

The results of both models can be found in **Table 10.3**. Model 1 seems to be better supported by the data since the BIC and AIC for Model 1 are lower than for

Model 2. In addition to this, the email rate is mainly affected by recent email history, that is, by the short-run effects of inertia and reciprocity. Although the effect of long-run inertia (LongInertia) is statistically significant, the effects of long-run inertia and long-run reciprocity (LongReciprocity) are negligibly small and hence barely affect the email rate. The results of Model 2 (which only includes inertia and reciprocity) show that these effects are stable and unaffected by the other variables. In other words, the employees tend to repeat their recent behavior and mainly respond to innovation-related messages received in the recent past, while innovation messages that were sent or received more than 30 days ago seem to no longer affect emailing behavior today. In other words, employees appear to discuss innovation because it is what they recently discussed, not because it is something that is on their minds in the long run. This is a sign that STRATSERV has not been able to make innovation an integral part of their employees' mindset.

From Model 1, we see that employees send emails at lower rates to other employees who are lower in the organizational hierarchy than they are themselves and send their email messages at higher rates to those who have higher hierarchy levels than they have themselves ($\hat{\beta}_{\text{DiffHierarchy}} = -0.3003$). In other words, email messages about innovation are more readily sent up the organizational hierarchy than down. This is consistent with the idea that the STRATSERV employees are willing to inform their superior about potential innovation but are less likely to put their ideas into action themselves by discussing it with those lower in the chain of command. Conversely, employees who enjoy higher hierarchical positions are more popular targets for such email messages than are those who occupy low status positions in the organization. Again, innovation discussion is directed up the chain, but much less to the lower levels.

Except for DiffHierarchy, all other variables in Model 1 have positive effects on the emailing rates. For instance, the email rate of a sender to a receiver who works in the same building (SameBuilding = 1) is around two and a half times ($\exp\{\hat{\beta}_{\text{SameBuilding}}\} = 2.679$) higher than the email rate from that same sender to a colleague working in a different building, holding constant all the other variables. This is an important finding, as it suggests that physical boundaries (i.e., working in a different building) also appear to function as communication boundaries: STRATSERV employees more intensely discuss innovation-related topics with those whom they routinely meet at the coffee machine, and much less with those they do not run into that often.

We also observe that the rate at which employees send innovation-related email increases with the time they have been at the organization. Conversely, newcomers and juniors turn out less active in communicating about innovation than are the seniors of the firm, which makes sense.

Discussion and Conclusion

The relative importance of the different effects can be used to improve and optimize information sharing. For example, as there is a large positive (negative) effect of interaction when employees work in the same (in different) buildings, interaction may be greatly improved by setting up interventions in the organizations that stimu-

Table 10.3 Model 1 and Model 2: maximum likelihood estimates, standard errors, z-values, p-values, AIC, and BIC

Variable	Model 1					Model 2						
	$\hat{\beta}$	$se(\hat{\beta})$	z-value	p-value	$\hat{\beta}$	$se(\hat{\beta})$	z-value	p-value	$\hat{\beta}$	$se(\hat{\beta})$	z-value	p-value
Intercept	-11.6322	0.0862	-34.914	0.000	-9.2559	0.0249	-371.323	0.000				
ShortInertia	0.0831	0.0005	151.582	0.000	0.0869	0.0005	181.294	0.000				
LongInertia	0.0058	0.0005	10.871	0.000	0.0065	0.0005	14.025	0.000				
ShortReciprocity	0.0484	0.0104	4.628	0.000	0.0345	0.0101	3.406	0.0006				
LongReciprocity	-0.0070	0.0170	-0.409	0.682	-0.0094	0.0162	-0.579	0.563				
SameBuilding	0.9854	0.0401	24.591	0.000								
DiffHierarchy	-0.3003	0.0096	-31.307	0.000								
LogSenderTenure	0.9234	0.0378	24.413	0.000								
AIC	16,004.33					16,981.15						
BIC	16,045.93					17,007.15						

Compiled by authors

late discussions across employees in different buildings. In addition, it is important to know for managers that STRATSERV's employees are less likely to share innovation-related communication with colleagues they are not co-located with. Although this can partly be addressed by strategically placing employees in their various locations, it is also important for managers to realize where communication may flow more easily and where it is likely to be hampered.

Furthermore, STRATSERV learns from this analysis that a temporary silence in innovation-related activity tends to remove the topic from the active attention of its employees. This could potentially be addressed by organizing activities around innovation, but it also signals that the current activities have not been successful in making innovation part of the normal conversation of STRATSERV's employees. This may be a reason to reevaluate the effectiveness of the current strategy while, at the same time, taking into account that it may take a long time to establish an innovation mindset.

Thanks to the relational event model, we are able to understand which factors play a role in employee interaction. Specifically, the observed differences in intensities and signs of the relative effects showed that certain characteristics can impact the email rate to different degrees and in different directions. Using targeted interventions, these insights can be used to reach more employees in a shorter amount of time. For further reading on relational event models, we refer interested readers to Leenders et al. (2016), Schecter et al. (2017), and Pilny et al. (2016).

10.4 Using Vehicle Sensor Data for Pay-How-You-Drive Insurance

The emergence and growth of connected technologies and big data are changing the face of all industries. An example of an industry which is expected to avail tremendous benefits from the relevant data generated by the billions of connected devices is the insurance industry. One of the most popular cases of big data adoption within the insurance industry is the Pay-How-You-Drive (PHYD) paradigm (Carfora et al., 2019). This means that instead of calculating insurance premiums based on only demographic characteristics, personal driving characteristics—either exposure or behavioral—are also incorporated in the insurance models (Tselentis et al., 2016).

In order to understand people's driving behavior, data is gathered about, for example, the driver's speeding and braking behavior. State-of-the-art research about modeling human driving behavior is mostly based on GPS data (Grengs et al., 2008), including variables such as the GPS location, traveled distance, and coarse-grained speed profile. However, nowadays, the standardization of the controller area network (CAN) bus technology and the increase of the electronic control units (ECUs) in modern cars offer a large availability of sensor data, enabling a more reliable and direct characterization of driving styles (Fugiglando et al., 2017). Considering the car as a human body, the CAN bus is the nervous system enabling communication between the different body parts (ECUs). Modern cars

may have up to 70 ECUs, such as the cruise control, audio systems, and engine control unit. Hence, the ability to connect the different ECUs and sensors in a vehicle through CAN bus technology enables the gathering of valuable information about, for example, the state of the vehicle and the driving behavior of the driver.

Despite the useful data provided by the numerous sensors in modern cars, the interpretation of data is cumbersome due to the different implementations of the CAN messaging system (de Hoog et al., 2019). Whereas the CAN protocol is standardized, the actual implementation differs for every manufacturer and even for every car model. So, in order to obtain the useful information, CAN bus traffic has to be analyzed and reverse engineered for every car type separately (Huybrechts et al., 2017). As this is a very time-consuming task, the use of CAN bus data to model driving behavior for PHYD insurance is barely adopted so far (Fugiglando et al., 2018).

With the flexible CAN solutions established by *Beijer Automotive B.V.*,⁹ one is able to access the complex vehicle sensor data hidden in cars. This enables the analysis of an enormous amount of informative data about not only the drivers (e.g., speed, brake, steering position, wheel speed, odometer, left/right direction indicator), but also their surroundings (e.g., fog/hazard lights, wipers, ambient air temperature). Although this overload of data may be promising concerning the reliability of driving-style characterization, it remains a complex concept influenced by a burdensome number of factors and possible interpretations of the driver response (Martinez et al., 2017). In other words, due to many (external) conditions affecting the driving behavior, it is difficult to understand what factors exactly caused a certain driving behavior. Did the driver brake suddenly because of an unexpected event caused by another driver or because he or she was distracted by his or her phone? This and many other questions could arise while analyzing all the variables. What can actually be learned from all these variables and should they be analyzed separately or simultaneously?

10.4.1 Time Series

Before continuing with discussing some interesting applications, a bit more should be mentioned about the data. As the measurements from the CAN bus are collected at uniformly spaced time instants, the gathered data can be considered as a *time series*:

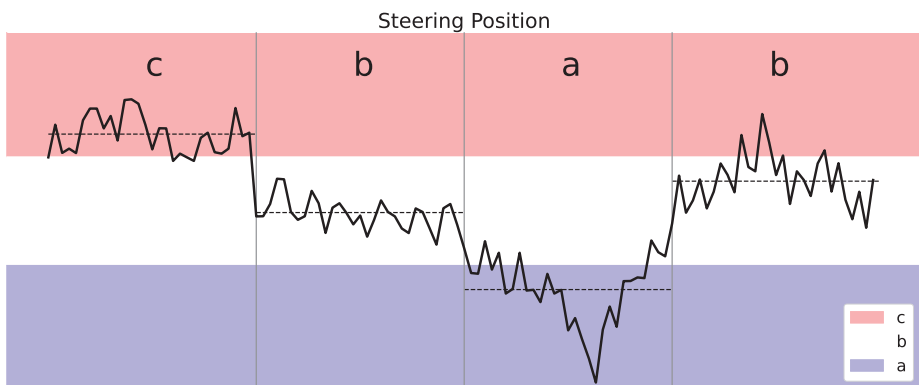
A time series $T = \{t_1, \dots, t_n\}$ is an ordered sequence of n real-valued numbers, often measured at fixed time intervals.

9 ► <https://www.beijer.com/en/>

The series can be univariate as described above or multivariate when several series simultaneously span multiple dimensions within the same time range (Esling & Agon, 2012). As all the data from the sensors and ECUs in the car are measured at the same time, they can be considered as multivariate time series.

While winning data from the CAN bus is already challenging, the actual problem begins when one wants to decode the gathered data. As mentioned before, this is due to the many different implementations of the CAN messaging system. However, imagine that you possess the information to make the right translation and thus that you can transform the raw data into long time series representing the variables of interest. Even when one is able to arrive at this stage, understanding the actual driving behavior remains challenging. This is due to the volume of the data; almost every 10 ms, a signal is sent through the CAN messaging system. Consequently, one ride of ± 1.5 h results in time series including over half a million data points. Hence, efficient algorithms are needed in order to analyze this data.

There are many different methods to analyze time series data, summarized by Esling and Agon (2012). As the obtained data is high-dimensional, algorithms directly applied to the raw time series would be computationally too expensive. To reduce the data dimensionality, one can use representation techniques. A widely used method for this is called Symbolic Aggregate approXimation (SAX) introduced by Lin et al. (2007). The method consists of two stages. First, the time series is converted into a piecewise aggregate approximation of a predefined number of segments. Afterwards, the average value of each segment is transformed into a symbol according to a set of break points. As a result, the time series is transformed into a *string* consisting of, for example, 3 symbols (see ■ Fig. 10.3). With string compression algorithms such as GrammarViz (Senin et al., 2018), grammar rules (e.g., *bba* in *acbbaacbbba*) can be inferred from the newly created string. These rules represent repeating patterns (*motifs*) in the time series. In a similar way, also anomalous patterns (*discords*) can be detected.



■ Fig. 10.3 SAX is used to transform a time series into a sequence of letters (a *string*). This figure illustrates a time series of 130 data points which is converted into a string *cbab* of 4 letters (i.e., segments). (Author's own)

Although dimensionality reduction techniques may increase the efficiency of time series data mining tasks, the downside is that details may be overlooked. In cases where those details play an important role, analysis can be better done on the raw time series. Depending on the application, the right technique should be chosen. Examples of applications in which motif or discord discovery could be of interest and situations in which dimensionality reduction techniques are not favorable are discussed in the coming sections.

10.4.2 Driving Behavior Analysis

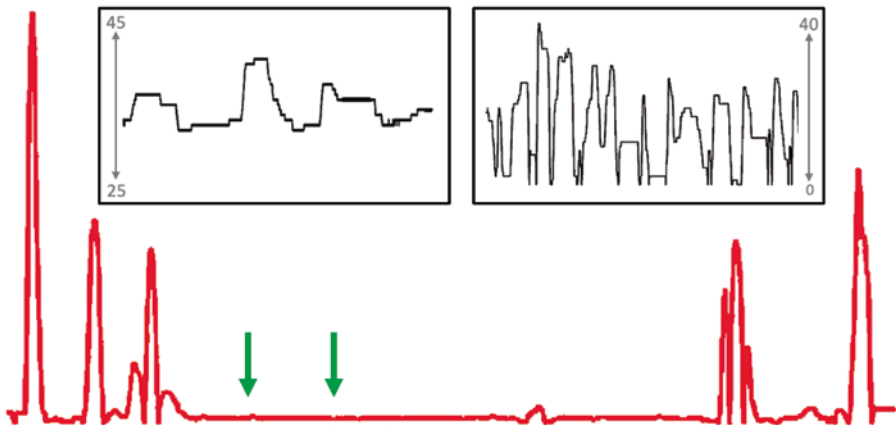
One insurance company in the Netherlands calculates its premiums based on their customers' driving behavior. For this task, they use four variables: speed, curves, brake, and acceleration. Although this provides insight into the driving style of a client, it is still very general. What exactly defines *safe* (or dangerous) driving behavior? Safety is a vague concept and could become more tangible when it is known what the patterns actually represent. In other words, only using those four variables does not include anything about the context of the ride. When more variables are included, maybe the cause of certain behavior can be detected and thus *safety* can be based on those events rather than on a variable like speed.

One of the main contributing factors to the road safety problem is an *inattentive driving style*, often caused by distracting activities (Meiring & Myburgh, 2015). Potential distracting activities may include attention to a person, object, or event outside the car, eating or drinking, talking, texting, and distracting weather conditions. Note that an inattentive driving style differs from an aggressive driving style due to its instantaneous and sporadic nature. Aggressive driving can be often observed as a pattern of misbehavior over a longer period of time (Meiring & Myburgh, 2015). The main challenge is how to use the gathered variables to detect such inattentive driving behavior. This contradiction serves well as an example for how the corresponding time series data should be analyzed: the detection of aggressive driving behavior may ask for motif discovery, while discords are of higher interest for the detection of inattentive driving behavior due to its anomalous nature.

Phone Usage Lately, especially the use of mobile phones is considered to be a threat to the safety on the road. Motivated by the impact on the overall safety, governments have enacted regulations that prohibit mobile phone usage while driving. But how can it be controlled? Is it possible to detect people being distracted in the car by using their mobile phone? Although previous methods promise to be effective in detecting the use of mobile phones while driving, they are dependent on either camera systems or on radars (Leem et al., 2017). As these attributes do not belong to the standard car equipment, they were specifically installed for the controlled experiment setup. The data from the CAN bus, however, is accessible in every car and could also be obtained from uncontrolled environments. Below, two variables from the CAN bus are described which could help to detect (or get insights into) phone usage or, more generally, driver inattentiveness.

Steering position. Beijer Automotive B.V. conducted an experiment in which they let people drive the same route twice: the first time without any instructions and the second time with the instruction to read a text message which was sent to them. The time interval in which the message was read is indicated by the green arrows in ■ Fig. 10.4. This figure shows the steering wheel position on the y-axis versus the time on the x-axis. A high peak corresponds to turning to the right or left.¹⁰ Although at first sight no difference was visible in the two different rides, when zoomed in, the difference came to light. Everyone who drives in a straight line moves the steering wheel lightly, resulting in a pattern similar to the left black curve in ■ Fig. 10.4. When distracted—in this case by reading the text message—people move the steering wheel more heavily, as shown in the right black curve in ■ Fig. 10.4.

As the motion is very detailed, it may not be advantageous to use dimensional-reduction techniques. On the other hand, when many rides need to be analyzed, it would become computationally too expensive to analyze the entire time series. Nonetheless, during this experiment, it became clear that one can use *one variable* (i.e., univariate time series) to get insights into the driving behavior of the driver. Although it was easy to identify the different patterns during this experiment, it becomes more challenging when no knowledge exists about the exact time slot in which a phone is used. When much data is generated in uncontrolled environments, it could be therefore useful to include more variables. In this way, the context can be used to understand a certain steering wheel action. Moreover, other variables like the brake may increase the accuracy of detecting people using their phone.



■ Fig. 10.4 The red curve shows the steering position during the route which was driven twice. The green arrows indicate a time interval, and the two black boxes show the steering position during that time interval for the two different rides. The left black curve represents a normal steering behavior when the driver did not receive a text message, and the right black curve represents the interval when the driver was distracted by his or her phone while driving. (Author's own)

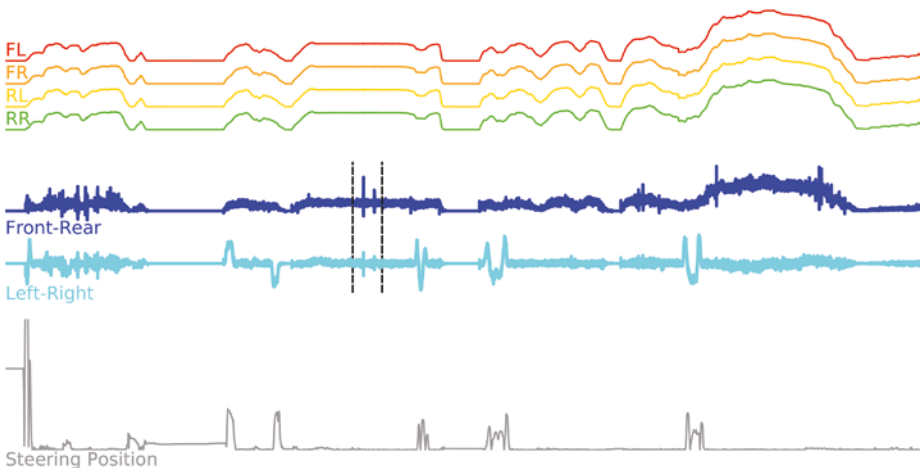
¹⁰ The peaks for both steering actions are positive as the signal is unsigned.

Wheel speed. Another way to detect potential inattentive driving behavior is by analyzing the wheel speeds. At the top of **■** Fig. 10.5, the speeds of the four wheels (front left, front right, rear left, and rear right) are visualized. While these variables separately may not seem informative, they include valuable information when analyzed simultaneously. The dark blue curve shows the difference in speed between the front and rear wheels (front–rear: $(FL + FR) - (RL + RR)$), and the other blue curve shows the difference in speed between the left and right wheels (left–right: $(FL + RL) - (FR + RR)$). The latter includes similar information as the steering position (gray curve): every time the steering wheel is moved to, for example, the right, the difference of the wheel speeds between left and right increases. While turns to the left or right are clearly visible through the big peaks, more detailed actions are also captured by the difference in the wheel speeds and can be used to detect anomalous driving patterns.

Not only the steering behavior of the driver is captured in the wheel speeds, but they also include additional information. In **■** Fig. 10.5, there are two anomalies visible in between the two black vertical lines halfway in the blue curves. When considering the separate wheel speeds, this is unexpected as the driver drove a continuous speed during that moment. Moreover, the steering position during that time period indicates that no steering action was performed. What these anomalies could represent is discussed in the next section.


10

Road Conditions Another important aspect of road safety is the monitoring of the road conditions (Meiring & Myburgh, 2015). Fazeen et al. (2012) demonstrated in their paper that by using mobile smartphones one is able to evaluate overall road conditions, including bumps, potholes, and rough, uneven, and smooth road. By



■ Fig. 10.5 At the top of this figure, the speeds of the four wheels are visualized. The blue curves show the differences in speed between the front and rear wheels (dark blue) and the left and right wheels (light blue). The gray curve represents the steering position. All curves share the same x-axis, which represents the time. The two black vertical lines highlight a time interval including two anomalies. (Author's own)

using the mobile phone's accelerometer, subtle or extreme vibrations were recorded inside the vehicle. Combining these accelerometer readings with GPS coordinates enabled them to make an accurate (85.6%) road condition mapping. However, to achieve accurate measurements, the location and orientation of five phones inside the car needed configuration.

By analyzing the wheel speeds, one is also able to detect road anomalies. Peaks as highlighted between the two black vertical lines in  Fig. 10.5 could indicate such anomalies. When exposed to a bump or pothole in the road, the speed of one wheel changes significantly compared to the other wheels. This leads to an anomaly in the differences between the wheel speeds. How accurate the detection of road anomalies via wheel speeds is has been hardly researched yet and is an interesting topic for future research. Nonetheless, with signals every 10 ms, small vibrations caused by either driving behavior or road conditions could be captured. Moreover, with many cars on the road, an enormous amount of data can be gathered and analyzed on a daily basis.¹¹ This enables a more reliable detection of road anomalies.

External Factors As all journeys differ considerably, the driver gets exposed every single journey to different external factors. Although variables like the wheel speed or steering position may include useful information, it still may be hard to detect anomalies in uncontrolled environments. One of the main advantages of using CAN bus data is that it includes informative data not only about the drivers, but also about their surroundings. Sensors like fog lights, hazard lights, wipers, and temperature provide insights into the climate, and other sensors like the direction indicator, the brake, and the throttle may include information about certain events. If, for example, the driver brakes heavily after driving 120 km/h, it may be more interesting to analyze the steering wheel position prior to this event than when someone is driving 30 km/h and uses the left direction indicator to turn to the left. Likewise, stormy days may elicit other driving responses than calm and sunny days, and so forth. By utilizing the information included in the overload of data retrieved from the CAN bus, one is able to understand the context of the driving scene and external conditions. This enables a more reliable and direct characterization of the driving behavior (Fugiglando et al., 2017). Note that in this case, there is no longer only dependency of one variable on its past values, but there is also some dependency between the other variables that has to be captured. Hence, techniques are needed which not only do have to deal with abnormal values or subsequences in each time series separately, but are also able to detect the relationships among the variables (Li et al., 2017).

Discussion and Conclusion

Whereas most car insurance companies quantify accident risk based on either demographic characteristics or GPS data, CAN bus data is expected to better characterize human driving behavior and thus accident risk (Fugiglando et al., 2018). Before

11 An example of a platform which brings together CAN bus data of many cars is *Vetuda* (► <https://www.vetuda.com/en/>). Not only road conditions can be analyzed, but it also provides information for applications such as incident, weather, and traffic management.

driving profiles of customers can be determined, many experiments should be conducted. By matching patterns in uncontrolled environments with the ground truth from controlled experiments, one is enabled to characterize *inattentive* and *aggressive* driving behavior. It is important to note that in uncontrolled environments, only rough proxies of inattentive driving behavior can be detected. Due to the lack of labels (ground truth), it is hard to determine the exact cause of anomalous driving patterns. Sometimes, people chose for an unsafe driving environment themselves (e.g., by using their phone), but also external factors such as other drivers can play a role in the decisions made by the driver. However, by focusing on steering actions such as corrections and unstable steering positions as depicted in ■ Fig. 10.4, it is possible to get a general overview of the driving behavior of customers.

Using this rich information not only is interesting for calculating the premiums of car insurance customers, but may also help insurance companies to understand the exact circumstances of accidents. Are there certain scenarios or places which cause many drivers to be distracted? Such information could be used to warn their customers and influence them to drive more safely. Ultimately, this could even lead to a shift in the core of their business model: a shift from restitution to prevention. Customers may also benefit from this new business model. With a reduction in restitution costs through prevention, discounts on premiums can be offered to those who drive safely. Using this information to adopt the Pay-How-You-Drive paradigm can be beneficial for the customers as they can now directly impact their paid premium. The safer you drive, the less you pay, and maybe even more importantly, the less we all pay.

10

Conclusion

In this chapter, we presented three case studies showing data analytics in action. The case studies considered diverse problems and provided an insight into the data analytical toolkit that is available to solve these problems. Of course, the data analytical toolkit is vast and there are many tools that we did not cover in this chapter. Nevertheless, the case studies illustrated how powerful modern data analysis techniques are for answering intricate questions that would otherwise remain open. We also emphasized that these techniques require careful adaption to the problem at hand in order to deliver the desired results. However, if this adaption is done right, data analytics can provide deep insights and produce practical outcomes that are highly valuable for businesses and consumers.

Discussion Points

1. AI education should be enriched with practical cases.
2. The inclusion of specific behavioral patterns in the dynamic social network analysis improves the understanding of the information flow between employees and helps refining business strategies.

3. In uncontrolled environments, only rough proxies of, for example, inattentive driving behavior can be detected. Due to the lack of labels (ground truth), it is hard to determine the exact cause of anomalous driving patterns. This should be taken into consideration when driving profiles are determined.

Take-Home Messages

- It takes a serious engineering effort to get an AI-powered software system into production. This is quite different from building AI demonstrators.
- It is an illusion to believe that a business intervention strategy affects all employees equally. Analyzing the communication between employees can help the management understand how, where, and for how long interventions carry an effect. Cutting-edge developments in longitudinal social network analysis can help target interventions more effectively and assess policy effectiveness realistically and in real time.
- By analyzing the enormous amount of informative data from CAN bus technology, human driving behavior—and thus accident risk—can be better characterized.

Acknowledgements Gerard Schouten and Petra Heck thank Erik van Breusegem of PTTRNS.ai for providing and reviewing the BagsID case study. We are more than happy that we can use this illustrative deep learning and software engineering case for educational purposes. We also thank Jesse Berger, graduate student at PTTRNS.ai, for his assistance and valuable support in digging up substantial business and technical information.

References

- Aalbers, R. H. L., Dolfsma, W., & Leenders, R. T. A. J. (2016). Vertical and horizontal cross-ties: Benefits of cross-hierarchy and cross-unit ties for innovative projects. *Journal of Product Innovation Management*, 33(2), 141–153. <https://doi.org/10.1111/jpim.12287>
- Air Transport IT. (2019). Insights, online. Retrieved March 2020, from <https://www.sita.aero/resources/type/surveys-reports/air-transport-it-insights-2019>
- Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, 38(1), 155–200. <https://doi.org/10.1111/j.1467-9531.2008.00203.x>
- Canziani, A., Culurciello, E., & Paszke, A. (2017). *An analysis of deep neural network models for practical applications*. arXiv:1605.07678.
- Carfora, M. F., Martinelli, F., Mercaldo, F., Nardone, V., Orlando, A., Santone, A., & Vaglini, G. (2019). A “pay-how-you-drive” car insurance approach through cluster analysis. *Soft Computing*, 23(9), 2863–2875.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B (Methodological)*, 34(2), 187–220. www.jstor.org/stable/2985181
- de Hoog, J., Castermans, N., Mercelis, S., & Hellinckx, P. (2019, November). Online reverse engineering of CAN data. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing* (pp. 776–785). Springer.

- Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1), 1–34.
- Fazeen, M., Gozick, B., Dantu, R., Bhukhiya, M., & González, M. C. (2012). Safe driving using mobile phones. *IEEE Transactions on Intelligent Transportation Systems*, 13(3), 1462–1468.
- Fugiglando, U., Massaro, E., Santi, P., Milardo, S., Abida, K., Stahlmann, R., Netter, F., & Ratti, C. (2018). Driving behavior analysis through CAN bus data in an uncontrolled environment. *IEEE Transactions on Intelligent Transportation Systems*, 20(2), 737–748.
- Fugiglando, U., Santi, P., Milardo, S., Abida, K., & Ratti, C. (2017, October). Characterizing the “Driver DNA” through CAN bus data analysis. In *Proceedings of the 2nd ACM International Workshop on Smart, Autonomous, and Connected Vehicular Systems and Services* (pp. 37–41).
- Grengs, J., Wang, X., & Kostyniuk, L. (2008). Using GPS data to understand driving behavior. *Journal of Urban Technology*, 15(2), 33–53.
- Heck, P. (2019). Software engineering for machine learning applications, online. Retrieved from <https://fontysblogt.nl/software-engineering-for-machine-learning-applications/>
- Heck, P. (2020). Testing machine learning applications, online. Retrieved from <https://fontysblogt.nl/testing-machine-learning-applications/>
- Hermans, A., Beyer, L., & Leibe, B. (2017). *In defense of the triplet loss for person re-identification*. arXiv:1703.07737.
- Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *MobileNets: Efficient convolutional neural networks for mobile vision applications*. arXiv:1704.04861.
- Huybrechts, T., Vanommeslaeghe, Y., Blontrock, D., Van Barel, G., & Hellinckx, P. (2017, November). Automatic reverse engineering of CAN bus data using machine learning techniques. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing* (pp. 751–761). Springer.
- IATA. (2020). Baggage tracking, online. Retrieved March 2020, from <https://www.iata.org/en/programs/ops-infra/baggage/baggage-tracking>
- Kratzer, J., Leenders, R. T. A. J., & Van Engelen, J. M. L. (2004). Managing creative team performance in virtual environments: an empirical study in 44 R&D teams. *Technovation*, 26(1), 42–49. <https://doi.org/10.1016/j.technovation.2004.07.016>
- Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *NIPS* (pp. 1106–1114).
- Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. John Wiley & Sons. <https://doi.org/10.1002/9781118033005>. Print ISBN: 9780471372158. Online ISBN: 9781118033005.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.
- Leem, S. K., Khan, F., & Cho, S. H. (2017). Vital sign monitoring and mobile phone usage detection using IR-UWB radar for intended use in car crash prevention. *Sensors*, 17(6), 1240.
- Leenders, R. T. A. J., & Dolfsma, W. A. (2016). Social networks for innovation and new product development. *Journal of Product Innovation Management*, 33(2), 123–131. <https://doi.org/10.1111/jpim.12292>
- Leenders, R. T. A. J., Contractor, N. S., & DeChurch, L. A. (2016). Once upon a time: Understanding team processes as relational event networks. *Organizational Psychology Review*, 6(1), 92–115. <https://doi.org/10.1177/2041386615578312>
- Leenders, R. T. A. J., Van Engelen, J. M. L., & Kratzer, J. (2003). Virtuality, communication, and new product team creativity: A social network perspective. *Journal of Engineering and Technology Management*, 20(1–2), 69–92. [https://doi.org/10.1016/S0923-4748\(03\)00005-5](https://doi.org/10.1016/S0923-4748(03)00005-5)
- Li, J., Pedrycz, W., & Jamal, I. (2017). Multivariate time series anomaly detection: A framework of Hidden Markov Models. *Applied Soft Computing*, 60, 229–240.
- Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2), 107–144.
- Martinez, C. M., Heucke, M., Wang, F. Y., Gao, B., & Cao, D. (2017). Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 19(3), 666–676.
- Meiring, G. A. M., & Myburgh, H. C. (2015). A review of intelligent driving style analysis systems and related artificial intelligence algorithms. *Sensors*, 15(12), 30653–30682.

- Mulder, J., & Leenders, R. Th. A. J. (2019). Modeling the evolution of interaction behavior in social networks: A dynamic relational event approach for real-time analysis. *Chaos, Solitons & Fractals*, 119, 73–85. ISSN: 0960-0779. <https://doi.org/10.1016/j.chaos.2018.11.027>.
- Pilny, A., Schechter, A., Poole, M. S., & Contractor, N. (2016). An illustration of the relational event model to analyze group interaction processes. *Group Dynamics*, 20(3), 181–195. <https://doi.org/10.1037/gdn0000042>
- Quintane, E., Pattison, P. E., Robins, G. L., & Mol, J. M. (2013). Short- and long-term stability in organizational networks: Temporal structures of project teams. *Social Networks*, 35(4), 528–540. ISSN: 03788733. <https://doi.org/10.1016/j.socnet.2013.07.001>
- Rollins, J. (2015). Online. Retrieved from <https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science>
- Schechter, A., Pilny, A., Leung, A., Poole, M. S., & Contractor, N. (2017). Step by step: Capturing the dynamics of work team process through relational event sequences. *Journal of Organizational Behavior*, 1–19. <https://doi.org/10.1002/job.2247>
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015). *FaceNet: A unified embedding for face recognition and clustering*. arXiv:1503.03832.
- Senin, P., Lin, J., Wang, X., Oates, T., Gandhi, S., Boedihardjo, A. P., Chen, C., & Frankenstein, S. (2018). Grammarviz 3.0: Interactive discovery of variable-length time series patterns. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(1), 1–28.
- Szegedy, C., Vanhoucke, V., Ioffe, S., & Shlens, J. (2015). *Rethinking the inception architecture for computer vision*. arXiv:1512.00567.
- Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2016). Innovative insurance schemes: Pay as/how you drive. *Transportation Research Procedia*, 14, 362–371.
- Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., & Wu, Y. (2014). *Learning fine-grained image similarity with deep ranking*. arXiv:1404.4661.
- Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. (2020). *Deep learning for person re-identification: A survey and outlook*. arXiv:2001.04193.



Data Entrepreneurship

Werner Liebrechts

The rise of technologies collecting, storing, analyzing, and visualizing data creates unprecedented opportunities for entrepreneurs. It is currently easier than ever before for startups to scale globally and to challenge incumbent organizations that used to dominate industries over the past decades. Hence, incumbents are also required to act more entrepreneurial to maintain their competitive advantage. Data entrepreneurs—nascent ones and those operating inside established firms—discover, evaluate, and exploit the opportunities offered by the ubiquity of ever-increasing amounts of data to create future goods and services. Data entrepreneurship can therefore be briefly defined as the process of new value creation by using data in order to exploit an opportunity.

Thus far, this book has covered various relevant topics under the umbrella of both *Data Engineering* and *Data Analytics*. In brief, data engineering has been described as the important preparatory work prior to data analytics activities, as it helps to unlock, integrate, refine, and process (big) data sources, such that they can be used for (advanced) data analytics. In turn, data analytics refers to all methods and techniques that are used to generate meaningful insights from those engineered data. However, from an entrepreneurial perspective, such insights are only deemed meaningful if they create new value for one or the other. Only then is there a potential market to be served and, hence, revenues or even profits to be gained. Besides, additional insights obtained from data may also lead to cost reduction, for example through an increase in (production) process efficiency.

In this *Data Entrepreneurship* section, we not only introduce different forms of data entrepreneurship, but also spend a great deal on discussing some of the most essential elements of owning and managing a successful data-driven or digital business. For instance, we devote ample space to explaining the state-of-the-art knowledge on topics like strategy development and implementation, entrepreneurial finance, and entrepreneurial marketing (and sales). We hereby always emphasize data-driven entrepreneurial activities. In what follows, we briefly explain what can be expected from each of the chapters in this section concerning data entrepreneurship.

► Chapter 11 titled *Data-Driven Decision-Making* covers the latest insights regarding the use of (big) data and data science in preparing, processing, executing, and evaluating decisions. In short, it is argued that, in order to create impact, the use of data science should both start and end with a thorough analysis of the related decision(s) to be made. This chapter opens the section on *Data Entrepreneurship*.

Data entrepreneurship is also often referred to as digital entrepreneurship. In the second chapter of this section—that is, ► Chap. 12 titled *Digital Entrepreneurship*—we therefore elaborate on the main concepts, central research questions, and latest theories and empirical evidence in the field of digital entrepreneurship research. Among others, the chapter deals with typical features of a digital economy, which have a number of important effects on the extent and nature of entrepreneurial activity in (mostly developed) economies.

► Chapter 13 titled *Strategy in the Era of Digital Disruption* then moves on to discussing strategies that can lead to the so-called digital disruption, that is, the increasing application of digital technologies by businesses as well as society more broadly, fundamentally changing competitive landscapes. Business model innovation is one of the core strategic concepts that may lead to such digital disruption, and is extensively discussed, among various other concepts.

One increasingly common form of business model innovation is digital servitization, which is about using digital technologies in order to shift from a product-centric to a more service-centric business model. This happens when firms add services to their core product offerings, thereby creating additional value for their customers. In ► Chap. 14 titled *Digital Servitization in Agriculture*, the authors explore how digital servitization benefits not only manufacturers, but also entrepreneurs in the agricultural sector.

Limited access to finance is seen as one of the most prominent problems for entrepreneurs willing to get their businesses off the ground. This is no less applicable to digital startups. ► Chapter 15 titled *Entrepreneurial Finance* therefore provides an overview of different types of investors and different types of funding for data- and/or technology-driven startups. The financial considerations and incentives for both entrepreneurs and investors are also discussed.

The marketing (and sales) of products and services based on radically new technology is being discussed in ► Chap. 16 titled *Entrepreneurial Marketing*. It stresses the importance of complementing a new product (or service) development process with a customer development process as to increase the chances of success. Later on, marketing efforts can be optimized using data collected from the firm's initial customer base. Think of improved ways to categorize customers into segments and to position one's products and/or services.

Contents

- Chapter 11 Data-Driven Decision-Making – 239**
Ronald Buijsse, Martijn C. Willemsen, and Chris Snijders
- Chapter 12 Digital Entrepreneurship – 279**
Wim Naudé and Werner Liebrechts
- Chapter 13 Strategy in the Era of Digital Disruption – 305**
Ksenia Podoyntsina and Egle Vaznyte-Hünermund
- Chapter 14 Digital Servitization in Agriculture – 331**
Wim Coreynen and Sicco Pier van Gosliga
- Chapter 15 Entrepreneurial Finance – 356**
Anne Lafarre and Ivona Skultétyová
- Chapter 16 Entrepreneurial Marketing – 383**
Edwin J. Nijssen and Shantanu Mullick



Data-Driven Decision-Making

*Ronald Buijsse, Martijn Willemsen,
and Chris Snijders*

Contents

- 11.1 Introduction – 241**
- 11.2 Introduction to Decision-Making – 244**
 - 11.2.1 Decision-Making Characteristics – 245
 - 11.2.2 The Decision-Making Process and Decision Rules – 248
 - 11.2.3 Decision-Making for Entrepreneurs – 249
- 11.3 Data-Driven Decision-Making – 251**
 - 11.3.1 What Is Data-Driven Decision-Making? – 252
 - 11.3.2 Maturity Levels of Data-Driven Decision-Making – 252
 - 11.3.3 Methodology Options for Data-Driven Decision-Making – 255
 - 11.3.4 Data-Driven Decision-Making by Entrepreneurs – 256

**11.4 Data-Driven Decision-Making:
Why? – 257**

11.4.1 Quality Reasons for Data-Driven
Decision-Making – 258

11.4.2 Capacity Reasons for Data-Driven
Decision-Making – 259

11.4.3 Mental Reasons for Less Data-Driven
Decision-Making – 261

**11.5 Data-Driven Decision-Making:
How? – 263**

11.5.1 Overview of Data-Driven Decision-
Making Solutions – 263

11.5.2 Data-Driven Decision-Making
Solutions for Programmed
Decision-Making – 265

11.5.3 Data-Driven Decision-Making
Solutions for Nonprogrammed
Decision-Making – 270

References – 275

Learning Objectives

After having read this chapter, you will be able to:

- Understand the main characteristics of decision-making, such as the decision-making situation, insights, rules, models, and processes.
- Identify the typical issues related to intuitive and rational decision making, and programmed and nonprogrammed decision making.
- Recognize and value the possible advantages and disadvantages of data-driven decision-making.
- Understand why and how to apply data science in decision-making and how data-driven decision-making relates to successful data entrepreneurship.

11.1 Introduction

Some scientists argue that data science is about extracting information or knowledge from data based on principled techniques (Provost & Fawcett, 2013). In the previous chapters, you have learned about data engineering and data analytics techniques to achieve this goal. However, others focus on the fact that this extraction is being done to drive or support decisions and actions: “Analytics is the extensive use of data, statistical, and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions” (Davenport & Harris, 2007: 9). In turn, business analytics is about delivering the right decision support to the right people at the right time (Laursen & Thorlund, 2010). In this chapter on data-driven decision-making, we take this second perspective and discuss how the knowledge gained from data engineering and analytics can be used to support decision-making and actions.

Data science techniques reveal patterns that inform us to take action, provide predictive models about what will happen next, and provide decision support (for example via recommendations) regarding what to act upon. This is where data-driven decision-making comes into play. How can we use the results of data science to make the *right* decisions, overcoming heuristics and cognitive limitations of the human decision maker? But even before data science is employed, it is important to think about the goals and purpose of the analysis. For securing data science impact, data science should start and end with an analysis of the related decision-making. The full embedding of data science in the decision-making process is what we label data-driven decision-making (DDDM).

Definition of Data-Driven Decision-Making

Data-driven decision-making is decision-making in which data- and model-based insights from data are used.

For a long time, the advanced use of data science solutions in intraorganizational decision-making has mainly been covered by operations research and management science. However, acceptance of operations research initially was rather low, and Russell Ackoff at some point declared that it should be considered as dead (Ackoff, 1979). However, the enormous growth in the availability of data, management techniques, and mathematical solutions to handle data, and the growing need to handle issues in a faster way, led to the increasingly high interest in data analytics and later data science (Donoho, 2017). At the same time, both the growth in data science and the need for improvements in decision-making were the main drivers behind the strong growth of data-driven decision-making. The use of DDDM in areas where processes are quite structured and have been quantified for administrative or fast decision-making reasons (e.g., finance, logistics, and manufacturing) gave way to the positive adoption of data science, also in areas where processes are much less structured and quantified (e.g., health care, human resource management, and marketing).

Most of these data science efforts can be seen as what Simon (1960, 1977) already long ago labeled as programmed decisions (see Fig. 11.1). These routine-based, repetitive decisions are easy to model, analyze, and improve by using data science techniques, offering large and easy gains by overcoming suboptimal habits and heuristics employed by human decision makers. However, Simon already understood that the real challenge is in how we can understand and support the more complex nonprogrammed decisions, in cases where the decision situation is novel and ill-structured. His classical work in understanding and modeling heuris-

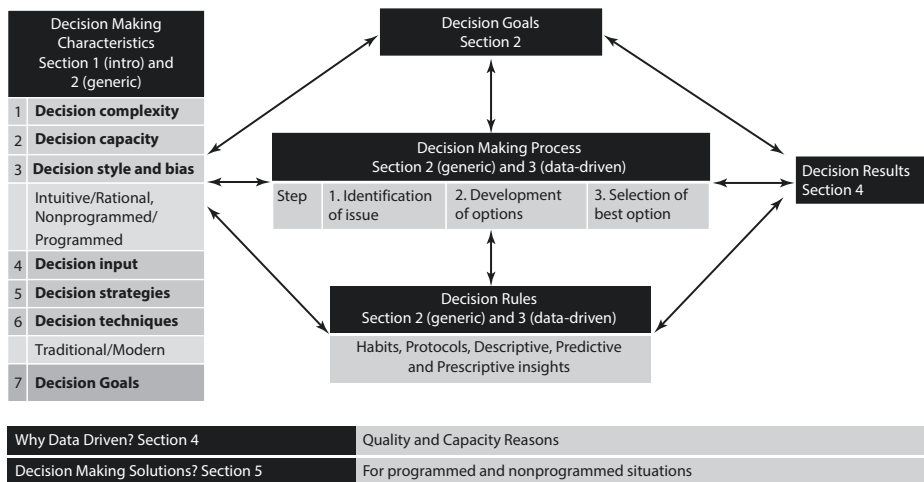
TYPES OF DECISIONS	DECISION-MAKING TECHNIQUES	
	Traditional	Modern
<p>Programmed:</p> <ul style="list-style-type: none"> • Routine, repetitive decisions • Organization develops specific processes for handling them 	<ol style="list-style-type: none"> 1. Habit 2. Clerical routine: <ul style="list-style-type: none"> • Standard operating procedures 3. Organization structure: <ul style="list-style-type: none"> • Common expectations • A system of subgoals • Well-defined information channels 	<ol style="list-style-type: none"> 1. Operations Research: <ul style="list-style-type: none"> • Mathematical analysis • Models • Computer simulation 2. Electronic data processing
<p>Nonprogrammed:</p> <ul style="list-style-type: none"> • one-shot, ill-structured novel, policy decisions • Handled by general problem-solving processes 	<ol style="list-style-type: none"> 1. Judgement, Intuition and creativity 2. Rules of thumb 3. Selection and training of executives 	Heuristic problem-solving techniques applied to: <ol style="list-style-type: none"> (a) training human decision makers (b) constructing heuristic computer programs

■ Fig. 11.1 Decision types and decision-making techniques by Simon (1977). (Source: Simon, 1977)

tic problem-solving (e.g., Newell & Simon, 1972) was a first attempt to quantify these more complex decision problems. But progress in that area has been much smaller and harder to make. However, with the renewed focus on and developments in DDDM, also for nonprogrammed decision-making by entrepreneurs, especially in the area of digital entrepreneurship, this is expected to change soon.

In this chapter, we will focus on DDDM in both programmed and nonprogrammed decision-making, following Simon's classification (1977). In his overview of decision types and decision-making techniques (see ■ Fig. 11.1), Simon (1977) clearly refers to the application areas of modern decision-making techniques for programmed and nonprogrammed types of decisions. For nonprogrammed decision-making, he already suggests two solutions for improvements, viz. (1) training human decision makers, with decision-making methods and best practices, and (2) constructing heuristic computer programs, which are now being delivered by data science solutions for nonprogrammed decision-making.

■ Figure 11.2 provides an overview of the different aspects of data-driven decision-making relevant to DDDM and how they are related to the different sections in the chapter. In ► Sect. 11.2, we start with a detailed description of decision-making in general, and the DDDM variant in particular. We describe various decision-making characteristics, which define whether DDDM is both possible and relevant. We discuss this for both generic and entrepreneurial decision-making. These characteristics define decision goals, process, and rules, both in a generic way, as we will discuss in ► Sect. 11.2, and in DDDM, which we will discuss in the section in more detail. ► Section 11.4 describes how these goals, processes, and



■ Fig. 11.2 Data-driven decision-making chapter overview. (Source: Authors' own figure)

rules contribute to actual decision results and what quality and capacity reasons trigger why DDDM is applied or not. ► Section 11.5 discusses the available DDDM solutions for both programmed and nonprogrammed decision-making. We end with conclusions and point for discussion.

11.2 Introduction to Decision-Making

Decision-making has been an important topic of study for many years (Buchanan & O’Connell, 2006). The encompassing discipline is the decision theory, the interdisciplinary approach to thinking about what constitutes sound decision-making, and is studied by economists, statisticians, psychologists, biologists, political and other social scientists, philosophers, and computer scientists.

In psychology, decision-making is defined as the cognitive process resulting in the selection of a belief or a course of action among several possible alternative options. Simon (1960) defined the task of rational decision-making as to select the alternative that results in the more preferred set of all the possible consequences. Mintzberg defined decision-making as a specific commitment to action (usually a commitment of resources) and a decision process as a set of actions and dynamic factors that begins with the identification of a stimulus for action and ends with the specific commitment to action (Mintzberg et al., 1976).

What these definitions have in common is that for decision-making we need some input on what to decide (goals, issues, a set of alternatives, and their prospective outcomes), and how to evaluate and decide among alternatives. We will evaluate different approaches to study decision-making along these lines.

11

Definition of Decision-Making

Decision-making is the selection of issues, alternatives, and procedure to choose a preferred alternative with a commitment to action, based on goals and insights.

The field distinguishes between normative theories, which determine the optimal rational decision-making given constraints and assumptions, and descriptive theories, which analyze how agents actually make the decisions.

Descriptive theories explain that decision makers do not use all information, avoid making extensive trade-offs and comparisons between alternatives but opt for shortcuts, also called heuristics, like rules of thumb or explicit or implicit protocols, introducing room for different types of biases. This is extensively covered in the bounded rationality theory of Simon (1955) and later extended in the work of Tversky and Kahneman, which is summarized adequately in the popular science work *Thinking Fast and Slow* (Kahneman, 2011).

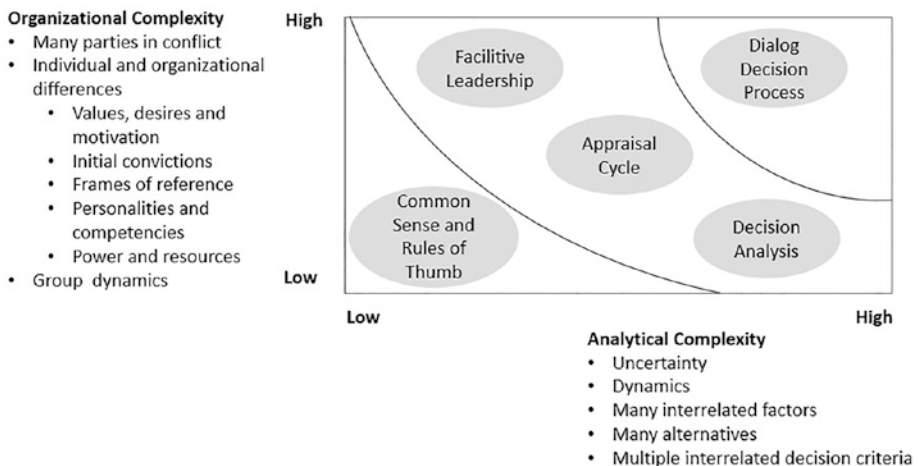
Busenitz and Barney (1997) discuss that the main reasons for the heuristic processing and simplified decision procedures are:

- High costs of rational decision-making efforts (Simon, 1977)
- Information processing limits of decision makers (Abelson & Levi, 1985)
- Differences in decision-making procedures adopted by managers (Shafer, 1986)
- Differences in values and strategies of decision makers (Payne et al., 1993)

For this introduction on decision-making, we will focus on the decision-making characteristics which are most relevant for our key topic, data-driven decision-making. These encompass situational characteristics that shape the decision process, such as the decision complexity, decision units, decision input, decision capacity, decision style, and techniques. Secondly, we will discuss perspectives on the decision process and decision rules that might help to improve DDDM, based on the decision process model of Mintzberg et al., (1976).

11.2.1 Decision-Making Characteristics

Decision complexity is the first situational aspect we consider. Spetzler (2016) positions decision complexity on two dimensions (see ■ Fig. 11.3): analytical complexity, related to the decision itself, and organizational complexity, related to the decision environment. Based on these dimensions, he positions several decision techniques, from more automatic decision techniques (using common sense or



■ Fig. 11.3 Decision complexity and decision solutions. (Source: Spetzler, 2016)

rules of thumb) to more deliberate decision techniques such as fact-based techniques like decision analysis (Parnell et al., 2013) and facilitative leadership. The technique for handling high decision complexity on both dimensions, the dialogue decision process (Spetzler, 2016), enables decision-making by making a split between a decision board and supportive decision project teams.

Directly related to these two dimensions is the **decision capacity** that is available and required to make a decision. The capacity is determined by the resources available to the decision maker. Decision-making resources are (1) decision owner(s), (2) potential decision expert(s) (i.e., one or more influencers) that can be consulted, (3) decision group(s) or organizations that influence the decision process, and (4) decision system(s) that support in collecting, processing, and/or evaluating data.

These elements define the decision capacity or cognitive capacity which is available for the decision-making process. To what extent these resources will be used will depend on the required cognitive capacity (how complex is the decision, how uncertain the input and expected consequences) and the extent to which the decision maker is motivated and able to use a rational style of decision-making rather than a more intuitive style. There are several barriers to cross before the full decision capacity is being used. First is the barrier to switch from an intuitive to a rational decision and increase the capacity by active research or learning. The second barrier to cross is the barrier to involve other people (experts, groups). The third barrier to cross is the barrier to involve extra data and decision systems.

Decision-making styles have a strong impact on the quality of the decision-making process, and they might lead to different types of decision biases (see ■ Table 11.1). Decision bias can be related to automatic associations, attention, and memory processes, as labeled by Kahneman as System 1 biases (availability, vividness, halo effects, anchoring). Decision bias can also be due to how the decision environment is shaped or if the decision is not purely individual, and bias can be related to social influences (conformity, groupthink, cascades). For details, see Spetzler (2016).

To what extent these biases play a role depends on the decision style. Several decision styles have been identified (Scott & Bruce, 1995): the rational style (with an in-depth search for information prior to making a decision), the intuitive style (with strong confidence in one's initial feelings and gut reactions), the dependent style (asking for other people's input), the avoidant style (averting responsibility), and the spontaneous style (make a quick decision). In line with these decision styles, choice scenarios have been defined in for example the ASPECT model of choice support (Jameson et al., 2015), including the more rational style-based attribute and consequence-based choice scenarios, the more intuitive-based experience and trial and error-based choice scenarios, and the more dependent policy- and social influence-based choice scenarios.

What type of **decision input** is being used, and to what extent this input is a complete representation of the decision problem or perhaps a biased perspective,

■ **Table 11.1** Overview of decision biases

Decision bias group	Bias
Protection of mindset	Avoiding dissonance, conformation bias, overconfidence, hindsight bias, self-serving bias, status quo, sunk cost
Personality and habits	Preference-based habits, habitual frames, content selectivity decision styles
Faulty reasoning	Selective attention, inability to combine many cues reliably, substitution heuristic, order effects, confusion about uncertainty
Automatic associations	Ease of recall, availability effects, vividness, narrative fallacy, halo effects, anchoring effects
Relative thinking	Framing effects, reference points effects, context effects
Social influences	Conformity, suggestibility, cascades, groupthink

Source: Spetzler (2016: 139)

depends also on the cognitive circumstances (time pressure, uncertainty, complexity) and cognitive capabilities of decision makers. Especially in complex, nonprogrammed decision situations, it might be hard to get an unbiased view of all the important goals, issues, alternatives, and decision constraints.

In *The Adaptive Decision Maker*, Payne et al. (1993) argue that decision makers adapt their **decision strategies** following an effort-accuracy trade-off. If decisions require more accuracy, exhaustive use of the available information is made, using compensatory decision strategies, whereas if decisions require less accuracy, simpler and less effortful decisions are made, for example using lexicographic choice (just look at the most important attribute) or satisficing (take the first option that passes all criteria).

Decision-making can also be characterized by the availability of traditional or modern **decision techniques**. For modern techniques, a split can be made for data-driven techniques (operations research, data science based on machine learning or artificial intelligence, recommender systems) and case-based techniques (case-based reasoning, competitive benchmarking, agent-based modeling, technology-assisted reviews, etc.).

Last but not least, the **decision goals** play a crucial role in defining and managing the decision-making process. Most of the time, decision goals will be defined explicitly, but when managing a decision process, also implicit decision goals should be taken into account. Especially Spetzler (2016) introduces several decision

framing techniques for defining and prioritizing the (shared) decision goals. Goals should be quantified for using certain data-driven decision-making techniques, such as prescriptive analytics or optimization, driven by operational research (OR) techniques.

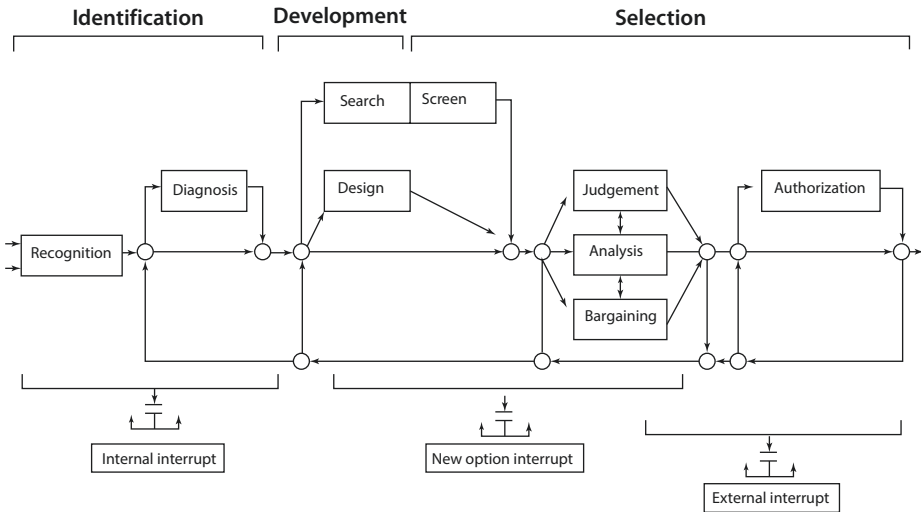
11.2.2 The Decision-Making Process and Decision Rules

By recognizing many cognitive limitations in decision-making and the related quick-fix solutions, such as heuristic decision-making and application of intuitive decision styles, many alternative methods have been proposed. These methods propose different approaches to slicing and structuring decision-making processes, such as using the dialogue decision process (Spetzler, 2016) or using additional information and extending the cognitive capacity, either by involving external parties or by using extra data and data science solutions, like decision support systems.

Good decision support requires a solid understanding of the underlying decision-making process by using decision process models (Robbins, 1996).

We will use the decision process model as defined by Mintzberg et al. (see

■ Fig. 11.4) as the main reference model for this chapter.



■ Fig. 11.4 Decision-making process model. (Source: Mintzberg, et al. 1976: 266)

The decision-making process as defined by Mintzberg et al. (1976) consists of three selection steps:

1. Identification (i.e., selecting the right issue to decide on)
2. Development (i.e., developing and selecting a set of viable alternatives)
3. Selection (i.e., selecting the best alternative from the set of viable alternatives)

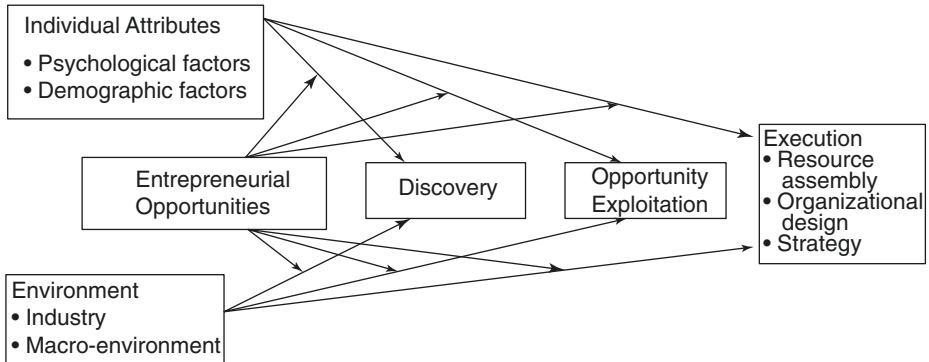
For each selection, data and insights, summarized in decision rules, are used. These decision rules for making selections are mainly deducted from experience, observations, and data and tuned towards the decision goals and preferred decision styles. An important consideration is if the insights found are descriptive (explaining the past), predictive (forecasting the future), or prescriptive (influencing the future). Another important consideration is if the insight is about a causal or correlated relationship. Insights are often translated in explicit or implicit protocols. Some are translated into descriptive, predictive, or prescriptive models.

The use of data science for deducting and applying insights for decision-making can increase the effectiveness, quality, and capacity of decision-making processes. A data-driven approach to decision-making can improve both the input and process of decision-making. It will extend the human cognitive system, increase the cognitive capacity by means of adequate analytics, reduce bias, improve the transparency and explainability of the decision, and make decisions more dynamic, personalized, and context driven. In the next sections, we will discuss in detail the data-driven decision approach. Before that, we first discuss decision-making as applied by entrepreneurs.

11.2.3 Decision-Making for Entrepreneurs

Entrepreneurial decision-making is one of the challenging areas for data-driven decision-making where still much nonprogrammed decision-making is applied. Entrepreneurship is about the discovery, evaluation, and exploitation of opportunities to create future goods and services (Shane & Venkataraman, 2000). In the view of Schumpeter, the entrepreneur is a change actor who is permanently seeking new opportunities. In other definitions, the capacity and willingness to develop, organize, and manage a business venture along with any of its risks in order to make a profit or realize another goal are included. The entrepreneur's role is described as to arrange or organize the human and capital assets under his or her control. So, entrepreneurship is also closely tied to resource ownership and employment relation.

The related types of decisions are grouped by Shepherd (Shepherd et al., 2014) in four groups: opportunity assessment decisions, entrepreneurial entry



■ Fig. 11.5 Entrepreneurial process model. (Source: Shane, 2003)

decisions, decisions about exploiting opportunities, and entrepreneurial exit decisions (see ■ Fig. 11.5).

While the entrepreneurial environment is characterized by a high level of uncertainty and risk, the entrepreneurial decision-making process, compared to non-entrepreneurs, is more often nonprogrammed and characterized by a higher level of bias and use of heuristics (Shepherd et al., 2014). Concerning bias: entrepreneurs tend to overestimate their prediction abilities (overconfidence bias) and overgeneralize from limited information (representativeness bias) (Busenitz & Barney, 1997). Concerning heuristics: these are used to increase the speed of decision-making.

Opportunities differ by the level of uncertainty, ranging from ultimate (in case of opportunity creation) to moderate (for opportunity discovery) to low (opportunity recognition) (Sarasvathy et al., 2010, based on Knight, 1921). In the cognitive continuum theory (CCT; Hammond et al., 1987), the levels of uncertainty are linked to types of decision-making. Highly uncertain tasks induce intuitive decision-making, moderate uncertainty induces quasi rationality, and low uncertainty induces analysis.

Because of the role of the entrepreneur, entrepreneurial decisions compared to management decisions are quite different on most of the characteristics described in ► Sect. 11.2.1. Entrepreneurs can rely less on decisions in the past, as entrepreneurial decisions are, for an important part, focused on creating a new future. Also, the obligation to report and to explain decisions is often lacking, which leads to different types of decision-making and different requirements for data-driven decisions. Especially because of the high level of uncertainty for both opportunity discovery and exploitation, entrepreneurial decision-making is more intuitive and nonprogrammed. For these reasons, the acceptance by entrepreneurs of data-driven decision-making is low. We will cover the role of DDDM for entrepreneurs after defining data-driven decision-making in general. ■ Table 11.2 provides an overview of entrepreneurial decision-making characteristics.

■ **Table 11.2** Entrepreneurial decision-making compared to managerial decision-making

Decision-making characteristics	Entrepreneurial and managerial decision-making compared
Type	Typical entrepreneurial decisions are opportunity evaluation decisions, entrepreneurial entry decisions, opportunity exploitation decisions, exit decisions. Managers focus more on exploitation decisions, which have less uncertainty but more organizational complexity.
Input	As entrepreneurs focus a lot on future situations, which do not exist yet, they often cannot rely on historical data, but they have to rely on their own vision and experience.
Decision-making unit	On average, entrepreneurs have smaller decision-making units as compared to managers, since managers need to involve entrepreneurs, peers, and employees to get approval and acceptance for their decisions.
Decision capacity	Entrepreneurs may have a large decision capacity, based on their experience and networks, but on average use less of it, from either experts, groups, or data systems, because of their need for speed and because they have their unique vision.
Decision area	The decision area of entrepreneurs is more strategic, related to opportunity creation or discovery, exploitation, and exit, and linked to all areas (product development, marketing, finance, human relations, etc.).
Decision style	On average, entrepreneurs are considered to have a more intuitive management style, because of the high-level uncertainty related to their ventures and the lower need to report.
Decision bias	Two important bias items for entrepreneurs are overconfidence and low representativeness (Busenitz & Barney, 1997).

Source: Table compiled by authors, mainly based on Busenitz and Barney (1997) and Shepherd et al. (2014)

11.3 Data-Driven Decision-Making

What is data-driven decision-making, and what characteristics of the decision-making situation create the need for this type of decision-making? To answer this question, we first need to specify what data we consider to be relevant for DDDM. The main purpose of data is to record activities or situations, to attempt to capture the true picture or real event and relationships. Therefore, all data are historical (Liew, 2007). For data-driven decision-making, it is important that data is to some extent objective, structured, and reusable, rather than mere subjective perceptions and attitudes that drive intuitive decision-making and that are not captured or recorded in any way.

11.3.1 What Is Data-Driven Decision-Making?

In the introduction, data-driven decision-making (DDDM) was defined as decision-making, in which we apply data- and model-based insights from data to support decision-making.

In the previous section, we discussed decision-making processes and the distinction between more rational decision styles and more intuitive styles that use less cognitive capacity and decision-making capacity. In relation to this, we see DDDM as the practice of making decisions based on the extensive use of data and insights deduced from data via modeling, rather than on subjective perceptions, intuition, or protocols (Provost & Fawcett, 2013). This is especially important when decisions have important consequences, are complex, and have many decision-making units involved and when a lot of data is available. In that case, decision rules or insights should be derived from the data by using descriptive, predictive, and prescriptive techniques, and data science techniques and models should be used to embed the derived insights and decision rules in the three steps of the decision-making process, to secure effective, high-quality, and well-accepted decisions. In fact, DDDM is data science applied (1) on the data input to find or create decision insights and decision rules and (2) on applying these insights and rules in the different steps of the decision-making process (identification, development, and selection). Depending on the complexity of the decision and the amount of data that is available, different levels of maturity of the DDDM can be identified, which we will discuss in the next section.

11.3.2 Maturity Levels of Data-Driven Decision-Making

The maturity of decision-making can be defined by the scope of the input (no or extensive use of data and models) and the decision-making capacity to handle uncertainty, risk, and complexity (low or high). Based on these dimensions, Davenport and Harris (2007) defined three levels for DDDM, labeled as descriptive, predictive, and prescriptive analytics. For a more complete picture, we extend their model to also include two non-DDDM levels, which are mainly based on human perceptions, habits, and intuitions instead of analytics. These levels are labeled as perceptions and protocols.

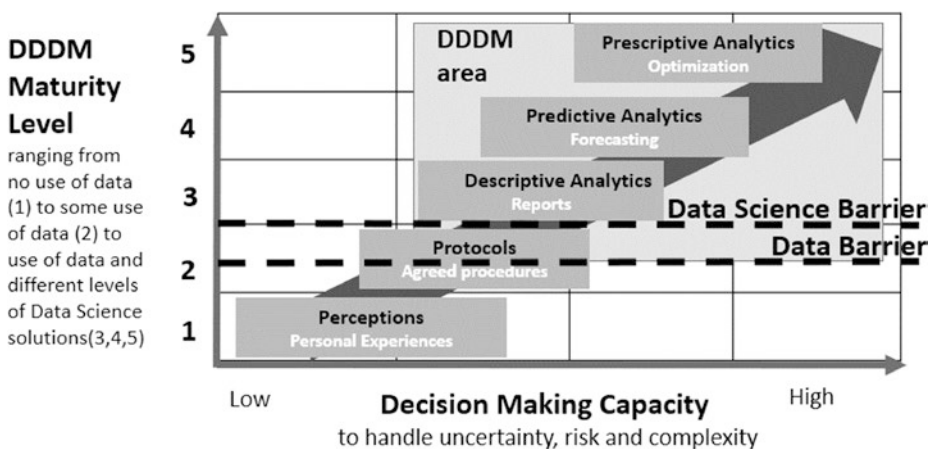
This leads to the following five maturity levels:

1. *Level 1*: Hardly any data is being used. Most decision-making is based on what is seen, heard, felt, or experienced before by the decision maker(s), and the decision-making process is merely intuitive, not rational. This way of decision-making is typically described as System 1 thinking by Kahneman (2011).
2. *Level 2*: Data is mainly being used as input instead of personal intuition or personal, non-stored perceptions. However, the decision-making process (creating and evaluating alternatives) is still classic, not digitalized, based on static procedures. See, for example, a large part of the medical decision-making and classical decision-making strategies (Payne et al., 1993).

3. *Level 3*: In addition to level 1, descriptive analytics is being used for both defining the issue and creating and evaluating alternatives. Especially accounting procedures can be related to this level.
4. *Level 4*: In addition to level 2, predictive analytics (forecasting) is being used in decision-making. This type of decision-making process can be found in utility services (for example, energy, gas, and waste collection) and many service industries (for example, public transport and nonfood products).
5. *Level 5*: In addition to level 3, prescriptive analytics (optimization) is being used in the decision-making process. This type of decision-making can be found in, for example, the finance and logistics industry.

What the right level of DDDM maturity and decision-making capacity is for a decision situation depends on the required cognitive capacity defined by the level of complexity of problem and organization, level of uncertainty, and available time. When applying DDDM, it should be taken into account that two barriers have to be crossed (see ■ Fig. 11.6). That is, (1) the barrier of using data instead of personal perceptions as the main input for decisions and (2) the data science barrier, where next to the use of data also data science concepts and technologies for making choices have to be accepted by decision makers, decision supporters, people executing the decisions, and people influenced by the decisions taken.

DDDM can also be seen as a solution to make decisions more transparent, repeatable, explainable, fact based, compensatory, and faster. Intuitive “System 1” decisions (Kahneman, 2011) are typically less structured, transparent and explainable, and often based on perceptions (DDDM level 1), or protocols and habits (DDDM level 2). More deliberate “System 2” decisions should at least partly be structured, transparent, and explainable, and therefore more related to DDDM levels 3, 4, and 5, but are often still limited in information processing (i.e., the num-



■ Fig. 11.6 DDDM maturity matrix. (Source: Authors’ own figure, partly based on Davenport & Harris, 2007)

ber of alternatives and attributes considered) due to the cognitive limitations of the (human) decision maker.

In fact, this is exactly the benefits DDDM can bring. By means of data and descriptive and predictive analytics, we can augment the human cognitive capacities to arrive at accurate and information-rich decision processes, driven by data science. A good example of such an approach can be found in the case of Dr. Reilly on the Cook County Hospital Emergency Room, in which human expertise combined with simple prescriptive models improved medical decision-making.

Less Can Be More with DDDM, the Case of Cook County Emergency Room

An interesting case that shows how a simple, data-driven decision-making model can improve health care was discussed by Blink (Gladwell, 2005). The Emergency Room (ER) of Cook County Hospital was overcrowded with patients; being located in a poor neighborhood of Chicago, it was the last resort for those without insurance. If resources are limited, how should one figure out who needs care? The compelling case is how to handle patients that enter with acute chest pain. Are they at risk of a heart attack and should they be admitted? Assessing the risk involves following an elaborate diagnosis protocol, taking a lot of clinical expertise, including an ECG. However, the results were often inconclusive, and when 20 case files were given to experts, there was hardly any agreement on their assessments of who was at risk. In 1996, Dr. Reilly took action to improve the decision-making in the hospital by implementing a simple decision tree once developed by Lee Goldman in the 1970s, which had been carefully designed but was never put to practice. That tree required only four pieces of information: an ECG, blood pressure,

whether there is fluid in the lungs, and whether the pain is felt as an unstable angina. Reilly took the tree and tested it for months, comparing it against the regular diagnosis of the staff. The Goldman algorithm won by a large margin, even though it used very little information and was much faster to assess: It was 70% better at recognizing patients that did not have a heart attack. More importantly, it did a better job than the doctors to find the cases that could lead to serious complications, with 95% against the doctors who were 75–89% of the time correct in identifying these.

The results show that a DDDM can overcome decision biases and intuitive judgments by identifying important diagnostic information and supporting the decision process with prescriptive analytics, improving the decisions while reducing effort and time investment (and costs) on the side of the doctors. The model makes use of the strengths of doctors (in assessing the evidence) and models (in combining the evidence into a good diagnosis), providing an excellent example of what DDDM stands for.

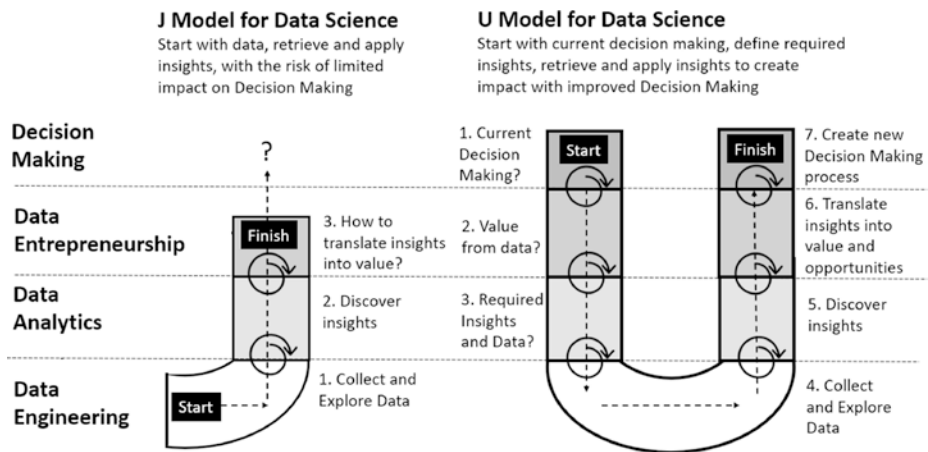
11.3.3 Methodology Options for Data-Driven Decision-Making

For some data science professionals, data science starts with data engineering (cleaning, aggregating data), followed by data analytics (to find insights) and then data visualization and insight implementation. We summarize this approach as the J model approach for data science. However, since in this approach the insights are coming bottom up from already available data, such insights might be based on incomplete datasets and not resonate with the actual decision-making challenges and goals of a decision maker or organization.

We argue that a good data-driven decision-making process should not start with data but with an assessment of the current insights and issues (similar to the process model of Mintzberg et al. 1976, ■ Fig. 11.4) and its related decision-making. For this reason, we propose the U model as a more complete methodology for data science. Both the J model and U model methodologies for data science are visualized in ■ Fig. 11.7.

The first and final step in the DDDM process based on the U model for data science is being considered as an important part of data entrepreneurship. In this way, the U model contributes to making the connection between DDDM and entrepreneurship.

In the first step of the U model, the problem or opportunity is being defined, including a description of the current decision-making data and processes. We would like to describe this as the entrepreneurial or management decision-making effort. In the second step, it should be defined if and how data science can solve the issue or capture the opportunity. We would like to describe this as a part of the data entrepreneurship effort. In the third step, the required insights and data should be defined, to guarantee that the engineering and analysis are done on the right



■ Fig. 11.7 The J and U model for data science. (Source: Authors' own figure)

dataset. In the fourth step, data engineering can start to collect or create the data required to discover the required insights. What data is analyzed is driven by an opportunity or problem and the insights it requires, not by the data which has been collected for other reasons in the past. The fifth step in the U model aligns with the second step in the J model and contains data analytics activities to discover relevant insights. Step six is all about data translation: when, where, and how can discovered insights be translated into solutions and opportunities, among those as defined in step two. Step seven is about embedding the discovered insights into the decision-making environment that was defined at the start, or removing the insights that were used, but proved not to be valid. This activity we would again describe as part of the data entrepreneurship effort. The U model approach for data science thus extends the more traditional J model for data science by starting with the real problem, diving into the required insights and data, and after that following the J model approach to arrive at new, data-driven insights for appropriate DDDM.

11.3.4 Data-Driven Decision-Making by Entrepreneurs

There is a growing volume of case evidence that DDDM has facilitated greater use of information in companies that has led to better company performance, at least in specific situations (Davenport & Harris, 2007; Loveman, 2003). The question is how DDDM can also be successfully applied in smaller businesses and to what extent DDDM can also contribute to the success of entrepreneurs.

Shane and Venkataraman (2000) state as the main reasons why some people discover entrepreneurial opportunities, and why others do not: (1) the possession of the prior information necessary to identify opportunities and (2) the cognitive properties necessary to value it. Concerning cognitive properties, people must be able to identify new means-ends relationships that are generated by a given change in order. This second reason seems to be more important than the first reason. Researchers have found that successful entrepreneurs see opportunities where other people tend to see risk (Busenitz & Barney, 1997; Kaish & Gilad, 1991; Shaver & Scott, 1992; Sarasvathy et al., 1998).

Exactly the expertise to discover and exploit relationships (insights) is the core value of data science. So therefore, DDDM, especially based on machine learning, for entrepreneurs, should be a huge opportunity for identifying, selecting, and exploiting opportunities.

Taking this into account, we extend our definition of data-driven decision-making towards data-driven decision-making for entrepreneurship.

Definition of Data-Driven Decision-Making for Entrepreneurship

Data-driven decision-making for entrepreneurship is the practice of finding, checking, and applying insights from data for discovering, evaluating, or exploiting opportunities for future goods and services.

However, as pictured in ► Sect. 11.2.3, entrepreneurial decision-making, on average, is still more intuitive than non-entrepreneurial decision-making, since data on new opportunities is often missing. Because of this, the adoption of DDDM by entrepreneurs may not be by entrepreneurs themselves, but by experts and other stakeholders in their network. There may be one exception, which is when the entrepreneur is involved in digital entrepreneurship. In that case, DDDM can be a more trusted source of decision capacity for the entrepreneur, as the entrepreneur knows about digital systems from his or her business activities. Once DDDM is part of the experience of the entrepreneur, DDDM can contribute a lot to the need for speed in decision-making, for at least those decisions that have enough certain data sources to create a data-driven decision situation.

Case Study About ► [Salesforce.com](https://www.salesforce.com)

A good example of digital entrepreneurship is ► [Salesforce.com](https://www.salesforce.com). This company is a full digital enterprise, with all services in the cloud. Because all customers are working in the cloud, Salesforce does not have to guess anymore if, how, and when their customers are using what type of their online facilities. Everything is measured continuously, both by the customer and by Salesforce. Whereas in the past suppliers of CRM systems delivered on premise packages and had to guess what was being used, Salesforce as a cloud-based CRM supplier can monitor all activities online, real time. By cre-

ating a platform for partner plug-ins, even the performance of these plug-ins can be measured by Salesforce, while the results can be used for fully data-driven decision-making. At the same time, customers are informed about best practices, based on the cloud measurements, so the customers can shift from intuitive sales and marketing to fully data-driven marketing too. So, when the entrepreneurial environment is fully data driven, there is less uncertainty, making the decision-making of entrepreneurs less intuitive, but more rational or analytical instead.

11.4 Data-Driven Decision-Making: Why?

There is a growing volume of case evidence that DDDM has facilitated greater use of information in companies that has led to better company performance, at least in specific situations (Davenport & Harris, 2007; Loveman, 2003). Research also shows that firms that adopt DDDM have a higher market value and that this value is most closely related to their level of IT capital. Brynjolfsson et al. (2011) did find that DDDM is associated with a 5–6% increase in the output and productivity, beyond what can be explained by traditional inputs and IT usage. In a more recent work, Brynjolfsson and McElheran (2016) show that from 2005 till 2010, DDDM in manufacturing has increased from 11% to 30%, but that adoption is mostly for larger plants, which have high usage of IT and educated workers and a high level of awareness of the usefulness of DDDM.

So, there are good reasons why it is important to make decision-making processes more data driven. In this paragraph, we will group the drivers for DDDM in quality reasons and capacity reasons. And we will introduce one concept, the Ladder of Inference, which helps to explain why, despite the presence of many drivers to opt for DDDM, DDDM is not adopted.

11.4.1 Quality Reasons for Data-Driven Decision-Making

The main advantages of DDDM are that the use of data and data science solutions can increase certainty about the relevant facts, can improve the recognition of an actual problem/opportunity, and can add information to the decision process beyond the perceptions of the decision maker, reducing bias and potentially discovering more insights. Moreover, it can make decision situations more transparent and structured and increase or extend the cognitive capacity by offloading part of the information processing to analytics. DDDM also provides the ability to store, tune, and reuse both explicit and tacit knowledge for decision-making for future uses.

When looking at the decision-making process, there are different reasons for using data and data science solutions in all steps of the decision-making trajectory:

- For triggering and scoping the process: collecting signals/facts (sensors)
- For preparing alternatives: data exploration
- For creating alternatives: simulation, what-if analysis
- For evaluating alternatives: finding insights, e.g., with machine learning
- For making choices: providing advice or recommendations
- For making, executing, and evaluating decisions: making use of for example artificial intelligence systems

However, DDDM might only be accepted when the relevant data is available, the extra cognitive effort for DDDM is diminished by adding the right tooling, and the possible positive impact makes it worth the effort to replace intuition- and habit-driven decision-making with data and data systems.

We will first discuss how DDDM can improve decision quality and what advantages but also limitations there are when moving from a less data-driven situation to a more data-driven situation. DDDM often replaces simpler, intuitive decision processes and might challenge existing beliefs, habits, and emotions surrounding the decision process. Not always are the tools used by data scientists trusted by the experts that feel being replaced by simple models.

11.4.1.1 DDDM for Decision Quality

Decision quality is the main subject for decision analysis (Parnell et al., 2013). Decision analysis is the process of creating value for decision makers and stakeholders facing difficult decisions involving multiple stakeholders, objectives, complex alternatives, uncertainties, and consequences. Decision analysis is founded on decision theory and uses insights from the study of decision-making.

■ **Table 11.3** Potential contribution of DDDM to drivers for decision quality

Driver for decision quality	Potential contribution for DDDM
1. Appropriate frame	1. Using more data and data analysis tools will help in defining a better frame, without getting lost in information overload
2. Creative, doable alternatives	2. By using data science, alternative models for prediction and simulation will become better
3. Meaningful, reliable information	3. By using facts and models instead of perceptions, intuition, and/or static protocols, information is easier to check, can be more real time, and will be less biased
4. Clear values and trade-offs	4. Data science models and evaluation tools will help in evaluations. See optimization systems (for best alternatives) and recommender systems (for range and/or best alternatives)
5. Logically correct reasoning	5. Logically correct reasoning is more secured when embedded in models and predictable processes
6. Commitment to action	6. When decisions are based on up-front agreed datasets and steps can be made more explicit and explainable with facts and models, there should be more commitment to action, unless the models restrict people in execution in using their capabilities

Source: Table compiled by authors, based on Spetzler (2016)

The ten decision analysis topics as identified by Parnell et al. are the selection of appropriate decision process, decision frame, decision objectives, decision alternatives, performance of deterministic analysis and development of insights, quantification of uncertainty, performance of probabilistic analysis and development of insights, optimization, communication of insights, and enablement of decision implementation.

Similarly, in his book on decision quality (Spetzler, 2016), Spetzler provides a more condensed list of six elements that offers an appropriate structure to discuss decision quality, which we will use to describe the effect of DDDM on decision quality. In ■ Table 11.3, we list the six elements and the potential contribution we see for DDDM on each of these elements.

11.4.2 Capacity Reasons for Data-Driven Decision-Making

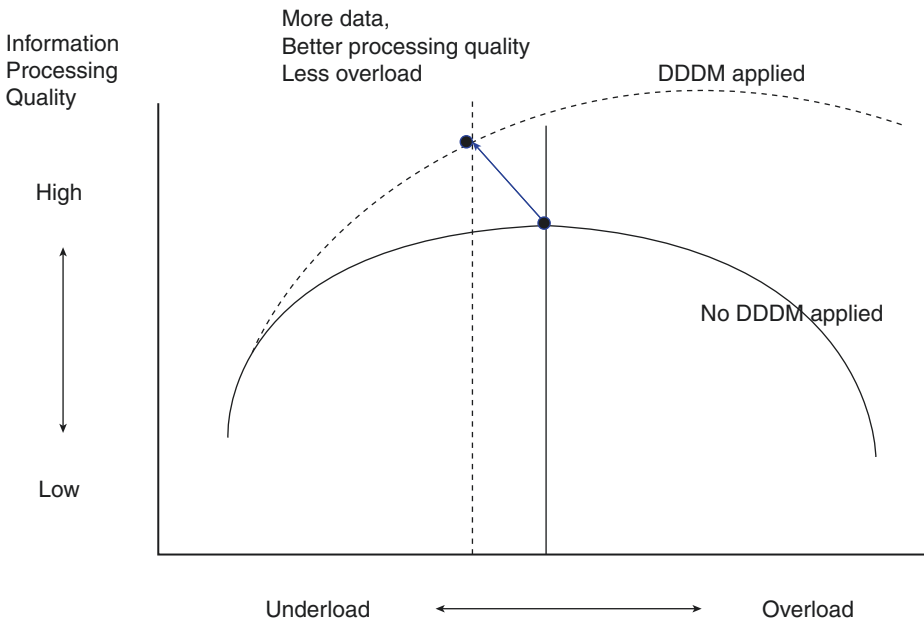
11.4.2.1 DDDM for Reducing Information Overload

For data-driven decision-making to be effective, the negative impact of increasing information overload should not surpass the positive impact of more and better information (Hwang & Lin, 1999). Researchers found that the amount of informa-

tion processing follows an inverted U shape: when increasingly more information is provided, people will initially improve in their information processing and decision-making, but at some point their cognitive capacities will be overloaded and information processing and decision quality will go down.

This means that for adequate DDDM, decision makers should be supported with adequate data science tools. If not, decision makers might revert to simple heuristic decision-making, using simplified non-compensatory decision strategies (Payne et al., 1993) that only process a limited amount of the information available. Ideally, DDDM should extend the decision makers' cognitive capacity, both in the scope of the data involved and in the capacity to review and evaluate alternatives.

So, when applied in the right way, DDDM solutions will reduce the information overload dramatically by focusing on the right input data and providing an optimal decision support. In this way, DDDM changes the inverted U shape, introducing both more information and information processing quality and at the same time reducing the risk of information overload. See ■ Fig. 11.8 and the case of Cook County ER in which a simple model was able to improve medical decisions.



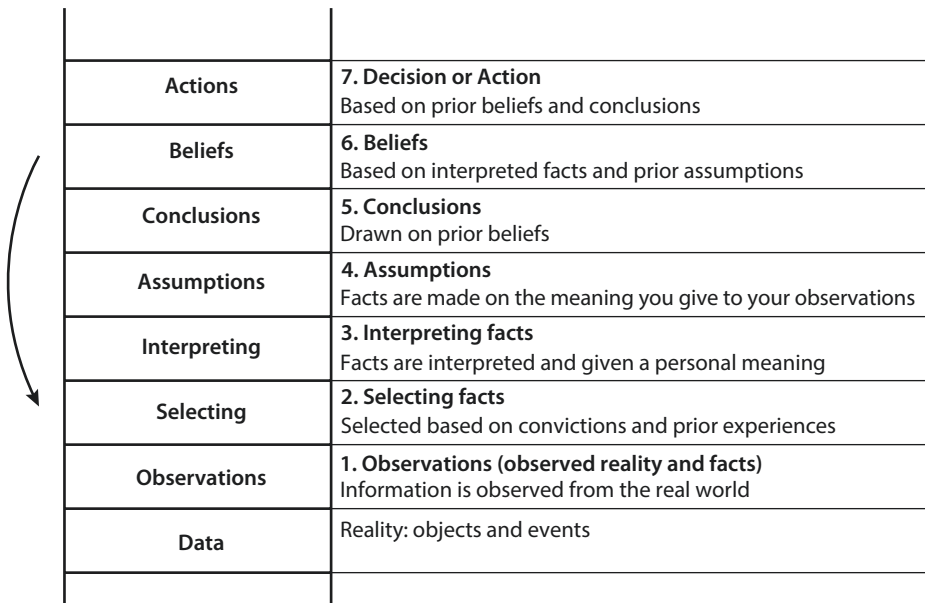
■ Fig. 11.8 Relationship between information processing quality, information load, and use of DDDM. (Source: Authors' own figure, based on Hwang & Lin, 1999)

11.4.3 Mental Reasons for Less Data-Driven Decision-Making

How do actors in decision-making make sure a data-driven insight or decision is accepted and executed? And how can it be explained when a data-driven insight or decision is not accepted or executed? For this topic, action research delivers some very useful concepts such as the Ladder of Inference of Chris Argyris and Peter Senge (Senge et al., 1994).

The Ladder of Inference (see ■ Fig. 11.9) represents the mental steps people make before they are committed to action. Next to data from observations, other mental steps are included such as interpretations, assumptions, conclusions, and beliefs. Often, these last four tend to be more driven by intuition than by data. So, there is a strong competition between these more intuitive features and data-driven features.

The Ladder of Inference can be seen as a data distillery process, in which there are data filters at each step of the ladder. In the cognitive processes related to decision-making, there are several “showstoppers” (see our case below) which diminish or even prevent the impact of data, as the data may not be observed, selected, or interpreted in the right way and the data has to compete with already existing assumptions, conclusions, and beliefs. Depending on the relevance and



■ Fig. 11.9 Ladder of Inference. (Source: Senge et al., 1994)

impact of the data, the data can pass by or even change interpretations, assumptions, conclusions, or beliefs. When DDDM solutions and procedures are being implemented in decision-making in the right way, there will be less reason and room for the use of nonfact-based assumptions and intuition.

Case Study on DDDM Showstoppers: Do Experts Trust Models?

Even though there are a plethora of reasons to want to use DDDM, there are certainly also many reasons why DDDM might not materialize. One could think of a timeline that a good DDDM idea should follow: from just an idea that might work at the start of the timeline to an implemented and applauded DDDM idea at the end of the timeline. In between, there are several hurdles to pass, of different sorts.

First, there can be reasons why it is difficult to come up with the appropriate data that is necessary. Although a lot of data is being collected nowadays, given the fact that a lot of our interactions leave digital traces somewhere, this need not imply that the data that one would need is necessarily part of it. Gathering appropriate data becomes more complicated, for instance, as the decision is more unique (less data to work with), or more qualitative (less data that can be used for predictive inference), or more personal (potential privacy issues). But even when we set aside the more technical hurdles and can come up with a proper predictive model, that in and of itself is just a start.

As soon as one would want to implement a DDDM, a certain amount of trust in the model is necessary. And there are certainly many reasons that users of models or the ones influenced by the decision of a model can come up with, to argue against the use of the model. For instance, whenever a DDDM is implemented in an existing

decision-making context, there is likely to be at least one decision-making actor whose task was to make the decision previously. Involving a DDDM implies overriding or at least hurting the autonomy, expertise, and experience of this previous decision maker, and people tend not to like this. The general issue of whether certain decisions are better taken by a human or a model is known in the literature as the “clinical-statistical” controversy. As many as 15 reasons that have been put forward why humans in general might not warm up to the idea of DDDM have been clearly formulated—together with their counterarguments—in a paper by Grove and Meehl that was written more than 25 years ago (Grove & Meehl, 1996).

We briefly summarize just a few:

- Aggregate statistics do not apply to the individual.
- Data-driven prediction models may defeat most decision makers, but not me
- The data-driven model cannot react to new information.
- Why not just compromise and use both?
- The data that have been used for the model do not apply in my case.
- A model just predicts, but a human wants more than just to predict.
- The model does not help us understand what is going on.

Many, if not all, of the counterarguments in Grove & Meehl's (1996) section "Replies to commonly heard objections" are as valid today as they were in 1996. We refer the unconvinced reader to their excellent expose for answers. And there are more. Ethical concerns could play a

role, privacy concerns, a need for transparency of decisions that might be hard to achieve, or worries about who or what is responsible in case the model happens to make mistakes. Coming up with a proper model is just the beginning of the journey.

11.5 Data-Driven Decision-Making: How?

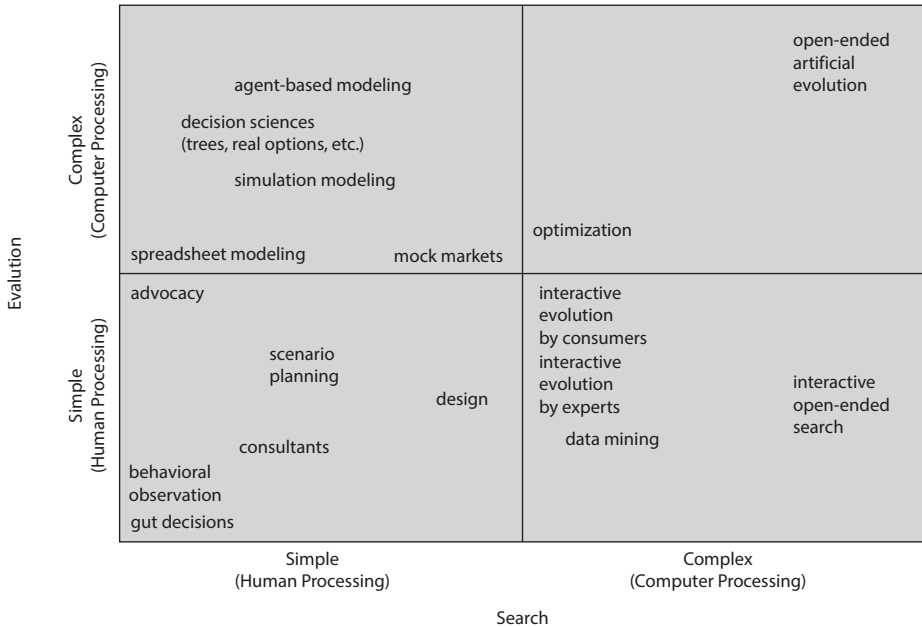
In the introduction to this chapter, we distinguished between programmed and nonprogrammed decision-making (Simon, 1977). In this section on DDDM solutions, we will discuss how DDDM can be performed in both cases.

11.5.1 Overview of Data-Driven Decision-Making Solutions

Bonabeau (2003) has pictured both programmed and nonprogrammed decision-making solutions in a matrix using complexity in number of options (data input) and evaluation complexity (decision process) as dimensions. Bonabeau maps about ten decision support options in his matrix (see ■ Fig. 11.10). When both the search and the evaluation of the alternatives are simple and easily quantifiable, these decisions can be classified as programmed decisions. When either the search or the evaluation is complex, a decision can be labeled as nonprogrammed decisions. The most difficult nonprogrammed decision-making issues are referred to as wicked problems. Whereas in nonprogrammed decision-making solutions can still be found in reference situations, in wicked problem situations, no reference situations can be found, which requires another set of DDDM solutions based on a combination of solutions for programmed and nonprogrammed solutions.

For programmed decision-making, many data science techniques and analytics are available to cope with these type of structured problems, and we already discussed some of them in the earlier sections. In their practical overview, Cukierski et al. (2015) group these techniques by phases in the decision-making process, from describing and discovering the issues and the related data to predicting and advising/recommending the preferred decision or action.

Programmed decisions are sometimes referred to as *routine* or *low-involvement* decisions because they do not require in-depth mental processing or complex solutions to reach a decision. However, that only applies if we indeed use DDDM to support these techniques as otherwise people might not overcome the data science



■ Fig. 11.10 Solutions for decision-making. (Source: Bonabeau, 2003)

barrier and keep using simple protocols and perceptions to make these type of decisions. In this section, we will discuss several programmed DDDM techniques, such as operations research (OR), data science techniques, and recommender systems and AI (see ■ Table 11.4) that will help decision makers to grow in their DDDM maturity level.

Nonprogrammed decisions on the other hand are novel, unstructured decisions that are generally based on criteria that are not well defined. With nonprogrammed decisions, information is more likely to be ambiguous or incomplete and the decision maker may need to exercise some thoughtful judgment and creative thinking to reach a good solution. This type of decision-making is strongly related to entrepreneurial decision-making. Nonprogrammed decisions are also sometimes referred to as *nonroutine* decisions or as *high-involvement* decisions, because they require greater involvement and thought on the part of the decision maker. There will always be unknowns in situations of nonprogrammed decisions. The best solution in this situation is to gather as much relevant information as possible and use some of the recently developed techniques for nonprogrammed decision-making, such as agent-based modeling, case-based reasoning, and others as listed in ■ Table 11.4.

In the next sections, for each of the two main areas (programmed and non-programmed decision-making, including decision-making for wicked problems), the main DDDM solutions (as summarized in ■ Table 11.4) will be briefly explained.

■ **Table 11.4** Overview of DDDM solutions

Programmed (data/formula driven)	<ul style="list-style-type: none"> Operations research Data science technologies, including machine learning Recommender systems Artificial intelligence systems
Nonprogrammed (driven by reference case(s))	<ul style="list-style-type: none"> Agent-based modeling Case-based reasoning Technology-assisted reviews Scenario-based decision-making Competitive benchmarking

Source: Table compiled by authors

11.5.2 Data-Driven Decision-Making Solutions for Programmed Decision-Making

Most of the DDDM solutions are based on data engineering and data analytics or operations research techniques and solutions as listed in ► Sect. 11.4.1. One of the advantages of programmed decision-making compared to nonprogrammed decision-making is the fact that both decision-making steps and input can be prepared and quantified up front and many mathematic concepts can be applied. That is one of the main reasons why operations research was already founded before 1940 and had been developed since then, despite the fact that for a long time, until 2000, there was a lack of data to exploit its potential. Data science, i.e., driven by the rise of machine learning to retrieve insights from large datasets as addition to the modeling techniques of operations research, became popular since 2000, because of the availability of many and large datasets. Recommender systems, with the techniques of data science (for retrieving insights) and operations research (for modeling a recommender system based on retrieved insights), became popular since 2005, due to the high rise of e-commerce applications for consumers and the need to support e-commerce customers in fast decision-making.

11.5.2.1 Operations Research Solutions

Operations research (OR), often referred to as management science, is a scientific approach to decision-making that seeks to design and operate a system, most often under conditions requiring the allocation of scarce resources (Winston & Goldberg, 2004). OR provides methods and techniques that support the decision-making process by evaluating every possible alternative and estimating the potential outcome (Sharma, 2006).

This approach usually involves the use of one or more mathematical models, which are made to understand the situation better or to make a better decision in a better way. Most OR models are prescriptive or optimization models, which help

organizations to meet their goals. Such models include an objective function, decision variables, and constraints.

A feasible solution that minimizes (or maximizes, if that is the goal) the objective function is called an *optimal solution*. When all the factors related to a problem can be quantifiable, only then operations research provides solution, otherwise not. OR solutions are very often used in finance and logistics for simulations, forecasting, and optimization decisions.

The nonquantifiable factors are not incorporated in OR models. Importantly, OR models do not take into account emotional factors or qualitative factors. For these type of decision situations, data science solutions like machine learning and deep learning are required.

11.5.2.2 Data Science Solutions

Data science solutions for decision-making are focused on finding, testing, and applying insights, derived from data. In other words, data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data (Provost & Fawcett, 2013). The most applied mathematical techniques in data science are clustering, classification, and regression. For an extensive list of data science solutions, see for example The Field Guide to Data Science (Cukierski et al., 2015).

A critical skill in data science for supporting decision-making is the ability to decompose a problem into pieces such that each piece matches a known task for which tools are available. Recognizing familiar problems and solutions avoids wasting time and resources for decision-making. It also allows people to focus attention on decision parts that cannot be automated so human decision-making comes into play.

For grouping familiar problems and their solutions, data science uses classification and clustering to group situations or objects based on their similarity. Classification is supervised, based on predefined classes, where clustering is unsupervised, without predefined classes. Next to that, regression supports decision-making by estimating or predicting the numerical value of some variable for that individual. Other data science techniques for decision-making are similarity matching to identify individuals or situations, based on data known about them, co-occurrence grouping, profiling (also known as behavior description), link prediction, data reduction, and causal modeling. Another solution area within data science is artificial intelligence, which represents methods for improving knowledge or performance of an intelligent agent over time. Within artificial intelligence, there is the field of machine learning, which supports decision-making by extracting models from data (Provost & Fawcett, 2013: 20).

A good example of a situation in which data science techniques can improve expert decision-making is the case of ICT transactions we will illustrate below.

Case Study on Assessing ICT Purchasing Transactions

Grove and Meehl (1996) have argued before that DDDM is likely to outperform human intuition and expertise typically when the following three conditions are fulfilled: (1) the topic is one where accumulated experience, intuition, and *Fingerspitzengefühl* are considered important; (2) decisions or predictions involve the incorporation of a relatively large number of dimensions; and (3) decisions or predictions involve the combination of dimensions in a “noisy environment.” That is, it is not clear which dimensions should be included, dimensions are hardly ever measured exactly, and it may very well be that combining the available measurements in even the most optimal way still leads to a decision or prediction that is only reasonable, and not good or even perfect. For this reason, Snijders et al. (2003) considered the comparison between DDDM and decision-making by human experts on a topic related to business processes: the assessment whether a certain purchasing transaction is likely to lead to a lot of problems.

The design of their research was as follows. Purchasing managers were given scenarios describing a transaction that described the procurement of IT products. They were then asked, among other things, to predict the amount of problems they would expect and how certain they were about their judgment. In fact, the scenarios were chosen from a larger database of real purchasing transactions, so that the correct answers were known to the researchers, and the purchasing managers’ answers can be compared to the actual answers. On a separate data-

set, a relatively straightforward formula was estimated, and this formula was used to likewise predict the “problem level” of these same transactions. Before, during, and after the experiment, the purchasing managers were asked whether they felt comfortable making these kinds of assessments—they were. However, it turned out that the prediction model, even when using only two predictors from the scenario, clearly outperformed the professionals. Neither the model nor the humans had perfect scores, but the model’s assessments were consistently better.

In a follow-up study, Tazelaar and Snijders (2004) have experimented with numerous ways in which perhaps the humans might nevertheless be able to outperform the computer model. Their results confirm the superiority of the prediction model in both these and numerous other slightly different circumstances. The model gets better when more information is available, humans do not. The model does not suffer under time pressure, humans do (and no, humans do *not* play their best game under pressure). In fact, Tazelaar and Snijders (2004) show that, apparently, this is not a situation that allows human expertise to shine: managers with more experience get more confident, but they show judgments that are actually slightly worse!

Academically, such findings are consistent with the results from different meta-studies. It is just that it strikes us as unlikely to be beaten by a model, especially if it happens in an area of expertise that we humans feel is our own.

11.5.2.3 Recommender Systems

Recommender systems are tools to support people in the entire DDDM process. Compared to other data science techniques, they are typically integrated more into the entire decision-making process and often interact with the decision maker directly.

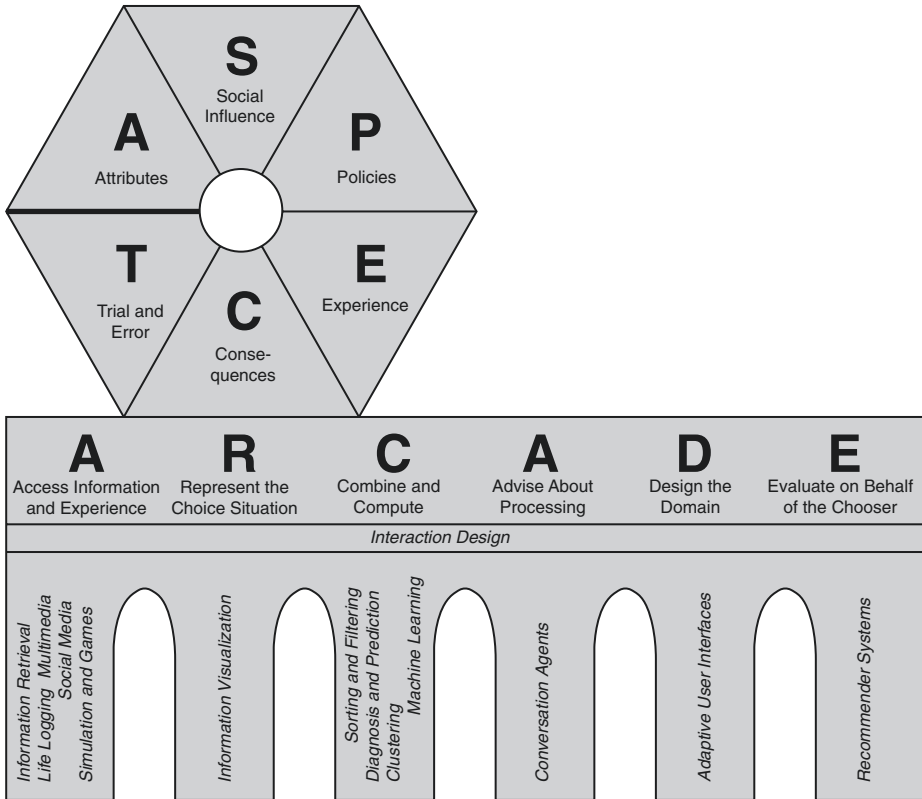
Whereas most programmed decision-making has a clear set of optimization criteria to improve, and predictive analytics can be used to advise on what a decision maker should do to optimize the workflow, recommender systems are often used in situations where a decision maker might not yet have a clear notion of what a good outcome or a good decision is. For example, when looking for a new movie to watch or song to play, historical data can be used to recommend new items, but that historical data is sparse and how much it represents the actual preferences of the decision maker depends.

Following the ASPECT and ARCADE models of choice support (Jameson et al., 2015) that were specially designed to describe the recommendation process, we see what value a recommender system can bring. As discussed before, the ASPECT model shows different choice patterns a decision maker can follow, and different types of recommendation techniques can be used to support these choices. Similar to the U model of DDDM, that proposes to start with the unique situation/problem, it is important that recommender systems start with understanding the choice pattern that needs to be supported, rather than just start with data and run a predictive model to generate a set of recommendations (■ Fig. 11.11).

To illustrate the role of the ASPECT model, take for example the attribute-based pattern. When the specific attributes on which the decision is based are clear, the recommender system can help the decision maker to find and evaluate all alternatives, based on what it learned about how important each of these attributes are. This reduces the information overload and extends the decision makers' cognitive capacity to make the appropriate trade-offs without relying on simple heuristics or simplified decision strategies.

Similarly, when the decision is socially influenced, the recommender can use collaborative filtering techniques to recommend based on what other people liked, but it can also use techniques to help find what people to take advice from or use a group recommender process.

Looking at the ARCADE model, we see many commonalities with decision process models we discussed earlier, such as the ones by Mintzberg (Mintzberg et al., 1976) and Spetzler (2016). Recommender techniques start with finding and representing data/information, combining and computing potential good items to recommend based on this, and then feeding this into an advice/recommendation process in which a user interacts with the recommendation. This process matches the DDDM process as we have proposed before. Recommender systems try to extend the space of possible alternatives a decision maker can choose from, propose what alternatives to consider by using machine learning and predictive techniques, and help users to evaluate these techniques. All of this happens without making the decision too complex, saving the decision maker from information



■ Fig. 11.11 Arcade model of choice support. (Source: Jameson et al., 2015)

overload, and also from mind-breaking intelligence collections, alternative listings, and choice making.

What is different from many other data science applications is that the advice and evaluation part is typically part of the system itself and that recommendations can be updated dynamically by means of user interaction. The accuracy of a recommendation is just one factor in the decision process (McNee et al., 2006), and diversity and serendipity are just as important in a good recommendation. Many recommender systems are interactive in some way and do not just recommend the best option, but provide a list of recommendations, often accompanied with some tools and visualizations to interact with the recommender engine (He et al., 2016) to update the list on specific user input.

We classified recommender systems as being part of the programmed type of decision-making, especially since they typically are used in situations in which historical data can quite accurately predict future item consumption, based on a set of predefined algorithmic approaches. Indeed, this model fits best with large-scale

recommender systems like Netflix or those used on e-commerce websites (Amazon) and social media (Facebook).

However, recommender systems also share some properties of nonprogrammed decision situations as the decision maker is typically uncertain about his or her real preferences and what the best decision might encompass. For example, Spotify already recognizes this in its discover weekly, which helps user find novel items, beyond the regular recommendations provided by the so-called daily mixes. Recommender systems are sometimes developed specifically to help users explore item spaces or even help people to develop new tastes or new healthier or more sustainable habits, for example by supporting their energy-saving decision-making (Starke et al., 2017, 2020). The field has recognized in recent years that purely building models on historical behavioral data is not enough (Ekstrand & Willemsen, 2016) and that a user-centric approach (Knijnenburg et al., 2012) is required to build systems that really support the (data-driven) decision-making of its users.

11.5.3 Data-Driven Decision-Making Solutions for Nonprogrammed Decision-Making

For nonprogrammed decision-making, it is not possible to define a structured suite of solutions as the decision situations are too diverse, and when a decision has to be made, most often there is a lack of time to collect data, structure the process, and build a solution (Ketter et al., 2016). For these reasons, DDDM can only be implemented in a nonprogrammed decision-making environment by (see Bonabeau, 2003) (1) creating decision-making building blocks, which can support the nonprogrammed way of decision-making, such as knowledge management systems, and (2) making reference situations (cases) available, e.g., by applying machine learning, agent-based modeling, or case-based reasoning on existing decision situations.

11.5.3.1 Agent-Based Modeling (ABM)

Agent-based models are being increasingly applied to the study of a wide range of social phenomena, often putting the focus on the macroscopic patterns that emerge from the interaction of a number of agents programmed to behave in a plausible manner (Francès et al., 2015). ABM is a technique based on models, dedicated to analyze the decision-making of different actors (Cian, 2017).

ABM provides an explicit representation of agent heterogeneity and of interactions across agents (Epstein & Axtell, 1996). ABMs are designed to capture the agents' perception of the relevant aspects of their environment and their decision-making according to their rationality. They often describe the interactions among different actors that operate according to prescribed behavioral rules and can capture emergent phenomena (Klimek et al., 2015). ABM applications are being used in problems like the energy transition.

ABM-based solutions can illustrate possible pathways of change at the level of individual decision-making, taking into account the behavioral implications of agents' heuristics and interactions with other agents.

ABM models depend on a diversity of technological and non-technological factors, while integrated assessment models (IAM) are cost-oriented models as the decisions in these models are based on choices regarding relative costs of technologies (e.g., capital, operation, and maintenance).

11.5.3.2 Case-Based Reasoning/Decision Analysis

Simon (1960), when making the distinction between programmed and nonprogrammed decision-making, discussed that modern decision techniques for nonprogrammed decisions would require heuristic problem-solving. At the time, heuristic problem-solving techniques like analogies and means-end analysis had just been identified and Newell and Simon (1972) even succeeded in programming some of these heuristics in their general problem solver. However, though the data science techniques we employ today indeed are useful to support and improve programmed decision-making, data-driven techniques for nonprogrammed decision-making are still rare.

However, one area that seems fruitful is to use a problem-solving technique that humans have mastered well, i.e., solve by analogy. Like we are able to solve novel problems by finding analogies, decision makers might be able to decide by looking at how similar decision cases have been handled in the past.

Courtney et al. (2013) discuss how decisions can be improved in these nonprogrammed situations, or cases in which the causal model is not known, as they label it. The basic idea behind case-based analysis is to take information from analogous past decision situations, which we label cases. The more similar the case is, the more its decision process and outcomes will help to predict the success of a decision strategy in the new situation. In their paper, Courtney et al. suggest that it is difficult for managers to look further than the most analogous case that comes to mind, but it is crucial that information from multiple analogous cases is integrated, weighted by the similarities to the new decision at hand.

This is where DDDM might come in, as it might provide techniques and tools to structure the data collection process to find appropriate cases that can serve as good analogies, using both qualitative and quantitative data collection techniques. Data science techniques for descriptive analytics might be used to identify the appropriate cases, and case-based reasoning algorithms can be used to support the decision-making in integrating the cases. Such case-based reasoning techniques have been developed for over 25 years in the area of recommender systems, but also in the medical domain, economics, education, energy, logistics, and workflows. To illustrate, one of the authors of this chapter used such techniques to help speed skaters to discover how to ride a better personal best with a race strategy (Smyth & Willemsen, 2020), and similar techniques have been used for pacing advice in marathon running (Smyth & Cunningham, 2017).

11.5.3.3 Technology-Assisted Reviews (TAR)

TAR is an application of machine learning. Machine learning processes like TAR have been used to assist decision-making in commercial industries since at least the 1960s, leading to efficiencies and cost savings in health care, finance, marketing,

and other industries. Now, the legal community is also embracing machine learning, via TAR, to automatically classify large volumes of documents in discovery.

TAR is conceptually similar to a fully human-based document review; the computer just takes the place of much of the human review workforce in conducting the document review. As a practical matter, in many document reviews, the computer is faster, more consistent, and more cost effective in finding relevant documents than human review alone. Moreover, a TAR review can generally perform as well as a human review, provided that there is a reasonable and defensible workflow. Similar to a fully human-based review where subject matter attorneys train a human review team to make relevancy decisions, the TAR review involves human reviewers training a computer, such that the computer's decisions are just as accurate and reliable as those of the trainers. A complete guide for implementing TAR can be found in the technology-assisted review (TAR) guidelines of the Bolch Judicial Institute (January 2019).

11.5.3.4 Scenario-Based Decision-Making

Scenario-based decision-making involves creation of descriptions of alternative future realities. The emphasis of these scenarios and the methods and techniques by which they are derived cover a diverse set of behaviors and involve methods that combine qualitative and quantitative and subjective and objective methodologies to different degrees (Harries, 2003).

Harries identifies five objectives of scenario planning: development of robust strategies, better understanding of the future, better perception of patterns and change, and transmission of management ideas through the use of these scenarios throughout the organization and leadership.

Scenario planning is based on identifying the drivers of change in an industry or domain, and the key uncertainties involved are picked out. The extremes of these are combined, and the resulting scenarios are described. Scenario planning is the process of generating causal story-like scenarios against which a strategy can be tested. In scenario-based decision-making, strategic decisions are tested for robustness against a series of scenarios describing possible/plausible future worlds.

11.5.3.5 Competitive Benchmarking

Competitive benchmarking (CB) (Ketter et al., 2016) is a research method that helps interdisciplinary research communities to tackle complex challenges of societal scale, e.g., wicked problems, by using different types of data from a variety of sources such as usage data from customers, production patterns from producers, public policy, and regulatory constraints for a given instantiation.

Further, CB data platforms most often generate data that can be used to improve operational strategies and judge the effectiveness of regulatory regimes and policies. CB is among other applications applied in complex decisions such as policy making related to banking, international trade, and health. A well-known example is also the global benchmarking related to the Sustainable Development Goals as defined in the Paris Agreement, in which more than hundreds of research groups from around the world jointly devise, benchmark, and improve sustainability policies and solutions.

Conclusion

This chapter discussed how data science, as discussed in the earlier sections on data engineering and data analytics, can support decision-making processes in organizations by means of data-driven decision-making. We first took a broad perspective on decision-making. Although huge improvements have been made in improving decision quality by implementing data-driven decision-making techniques, it is still a big challenge to implement DDDM for nonprogrammed decision-making and to make DDDM for programmed decision-making more often applied and more successful. This requires a good understanding of the “why” and the “how” related to DDDM.

We have listed the main reasons (i.e., the “why”) for applying DDDM next to the showstoppers that might prevent decision makers from applying DDDM. We have also given an overview of the solutions (i.e., the “how”). From these listings, we can learn that both the science and application of DDDM are making strong progress, driven by both the need for better, faster decision-making and the availability of new techniques and best practices. However, we also noted several aspects that deserve more attention to increase the use and success of DDDM.

For example, we have described the current data-first DDDM as the J model concept for data science projects and have argued that for DDDM to be successful, it might be advisable to follow the U model concept for data science, which starts and ends with decision-making analysis, to increase the success and impact of both data science projects and DDDM applications. Too often, current data-driven approaches start with the question “what can we get from the data?” rather than first defining and structuring the decision situation/problem that should drive the data science efforts.

Next to that, we have highlighted the growing importance of recommender systems as a solution for (partly) programmed decision-making, next to other solutions for fully programmed decision-making such as operations research and other data science tools that mostly focus on the descriptive and predictive analytics. The evaluation and action part of the decision-making cycle can be strongly supported by tools like recommender systems. More focus on a decision support or recommender systems approach, also for nonconsumer applications, could strengthen the successful use of DDDM and might become another area of strong growth. Recommender systems already have a strong impact on the data science domain, but yet have to become a much more discussed part of the decision quality literature as we reviewed it in paragraph 2.

Finally, we have focused on recent developments of different techniques to make also the more complex variants of nonprogrammed decision-making more data driven, which is also relevant for data entrepreneurs. When nonprogrammed decision-making situations have some kind of reference base of comparable decisions made before, there are already some successful cases that can be used as analogies. However, for decision situations where even all types of references are missing, such as the so-called wicked problems, there is still room to improve in both science and applications.

Again, we conclude that it is still a big challenge to implement DDDM for nonprogrammed decision-making and to make DDDM for programmed decision-making more often applied and more successful. This requires a good understanding of the “why” and the “how” related to DDDM. In this respect, this chapter has to be seen as a starting point for future research in those areas.

Discussion Points

1. Discuss the most important reasons that make the digital entrepreneurs (i.e., those working in the digital economy) adopt DDDM faster and more effective, compared to entrepreneurs working in other industries.
2. Suppose a company reveals a new data source from their production process that might be helpful in optimizing some business processes. Discuss from the J and U model of DDDM how the company should harvest this data to improve their DDDM. Also discuss which method would be the most successful in creating business value.
3. For a company that is still at the second level of DDDM maturity, but that already collected a lot of new data to improve their current protocols, what do you think are the most important quality reasons to do DDDM, and what barriers or showstoppers might they face? Would a company that is already at the fourth level of maturity (predictive analytics) have the same reasons and barriers or different ones?
4. The real challenge of DDDM is in nonprogrammed decision-making situations. Some recent data science techniques such as recommender systems and AI are still classified as programmed decision-making solutions. Discuss why this is the case, and compare these techniques with some of the nonprogrammed techniques discussed in the chapter. How do you think data science should progress to also support nonprogrammed decision-making?

Take-Home Messages

- Decision-making has many variants. Some variants, like programmed decision-making, already have a lot of support from data and data science solutions to become more data driven. For nonprogrammed decision-making, other concepts and data science solutions have been developed, but more research is needed.
- Data-driven decision-making requires a process in accordance with the U model for data science, not just the J model. According to the U model, data-driven decision-making should start with an analysis of the current decision-making situation and end with a proposal on how to implement the new insights derived from the analysis.
- Reasons for applying data-driven decision-making (i.e., the “why”) are mainly related to the quality of the decision and the decision-making capacity. Decision quality is improved by moving from intuitions and simple heuristics to decision based on the right data and the right decision process, and decision-making capacity is improved as long as the DDDM augments rather than overloads the cognitive capacity of the decision maker.
- Solutions for data-driven decision-making (i.e., the “how”) are mainly applied in programmed decision-making (e.g., descriptive, predictive, and prescriptive solutions, including recommender systems), but also for nonprogrammed decision-making, the expertise for applying data science solutions (e.g., agent-based modeling, case-based reasoning, and others) is growing.

References

- Abelson, R. P., & Levi, A. (1985). Decision making and decision theory. In I. G. Lindsey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. I, 3rd ed., pp. 231–309). Random House.
- Ackoff, R. L. (1979). The future of operational research is past. *The Journal of the Operational Research Society*, 30(2), 93–104. <https://doi.org/10.2307/3009290>
- Bonabeau, E. (2003, May 1). Don't trust your gut. *Harvard Business Review*. Retrieved from <https://hbr.org/2003/05/dont-trust-your-gut>
- Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). *Strength in numbers: How does data-driven decision-making affect firm performance?* (SSRN Scholarly Paper ID 1819486). Social Science Research Network. <https://doi.org/10.2139/ssrn.1819486>
- Brynjolfsson, E., & McElheran, K. (2016). The rapid adoption of data-driven decision-making. *American Economic Review*, 106(5), 133–139. <https://doi.org/10.1257/aer.p20161016>
- Buchanan, L., & O'Connell, A. (2006, January 1). A brief history of decision making. *Harvard Business Review*. Retrieved from <https://hbr.org/2006/01/a-brief-history-of-decision-making>
- Busenitz, L. W., & Barney, J. B. (1997). Differences between entrepreneurs and managers in large organizations: Biases and heuristics in strategic decision-making. *Journal of Business Venturing*, 12(1), 9–30. [https://doi.org/10.1016/S0883-9026\(96\)00003-1](https://doi.org/10.1016/S0883-9026(96)00003-1)
- Cian, E. D. (2017). *Actors, decision-making, and institutions in quantitative system modelling*. Fondazione Eni Enrico Mattei (FEEM).
- Courtney, H., Lovallo, D., & Clarke, C. (2013). Deciding how to decide. *Harvard Business Review*, 91(11), 62–70.
- Cukierski, W., Herman, K., & E., & Kherloplan, A. (2015). *The field guide to data science*. Booz.
- Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning* (1st ed.). Harvard Business Review Press.
- Donoho, D. (2017). 50 Years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745–766. <https://doi.org/10.1080/10618600.2017.1384734>
- Ekstrand, M. D., & Willemsen, M. C. (2016). Behaviorism is not enough: Better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 221–224). <https://doi.org/10.1145/2959100.2959179>
- Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up* (p. xv, 208). The MIT Press.
- Francès, G., Rubio-Campillo, X., Lancelotti, C., & Madella, M. (2015). Decision making in agent-based models. In N. Bulling (Ed.), *Multi-agent systems* (pp. 370–378). Springer International Publishing. https://doi.org/10.1007/978-3-319-17130-2_25
- Gladwell, M. (2005). *Blink: The power of thinking without thinking* (p. viii, 277). Little, Brown and Co.
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, 2(2), 293–323. <https://doi.org/10.1037/1076-8971.2.2.293>
- Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, and Cybernetics*, 17(5), 753–770. <https://doi.org/10.1109/TSMC.1987.6499282>
- Harries, C. (2003). Correspondence to what? Coherence to what? What is good scenario-based decision making? *Technological Forecasting and Social Change*, 70(8), 797–817. [https://doi.org/10.1016/S0040-1625\(03\)00023-4](https://doi.org/10.1016/S0040-1625(03)00023-4)
- He, C., Parra, D., & Verbert, K. (2016). Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 56, 9–27. <https://doi.org/10.1016/j.eswa.2016.02.013>
- Hwang, M. I., & Lin, J. W. (1999). Information dimension, information overload and decision quality. *Journal of Information Science*. <https://doi.org/10.1177/016555159902500305>
- Jameson, A., Willemsen, M. C., Felfernig, A., de Gemmis, M., Lops, P., Semeraro, G., & Chen, L. (2015). Human decision making and recommender systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 611–648). Springer US. https://doi.org/10.1007/978-1-4899-7637-6_18

- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Kaish, S., & Gilad, B. (1991). Characteristics of opportunities search of entrepreneurs versus executives: Sources, interests, general alertness. *Journal of business venturing*, 6(1), 45–61.
- Ketter, W., Peters, M., Collins, J., & Gupta, A. (2016). Competitive benchmarking: An IS research approach to address wicked problems with big data and analytics. *MIS Quarterly*, 40(4), 1057–1080. <https://doi.org/10.25300/MISQ/2016/40.4.12>
- Knight, F. H. (1921). Risk, uncertainty and profit (Vol. 31). Houghton Mifflin.
- Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4–5), 441–504. <https://doi.org/10.1007/s11257-011-9118-4>
- Laursen, G. H. N., & Thorlund, J. (2010). *Business analytics for managers: Taking business intelligence beyond reporting* (1st ed.). John Wiley & Sons.
- Liew, A. (2007). Understanding data, information, knowledge and their inter-relationships. *Journal of Knowledge Management Practice*, 8(2), 1–16.
- Loveman, G. W. (2003, May 1). Diamonds in the data mine. *Harvard Business Review*. Retrieved from <https://hbr.org/2003/05/diamonds-in-the-data-mine>
- McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems* (pp. 1097–1101). <https://doi.org/10.1145/1125451.1125659>
- Mintzberg, H., Raisinghani, D., & Théorêt, A. (1976). The structure of “unstructured” decision processes. *Administrative Science Quarterly*, 21(2), 246–275. <https://doi.org/10.2307/2392045>
- Newell, A., & Simon, H. A. (1972). *Human problem solving* (p. xiv, 920). Prentice-Hall.
- Parnell, G. S., Bresnick, T., Tani, S. N., & Johnson, E. R. (2013). *Handbook of decision analysis*. John Wiley & Sons.
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.
- Peter, K., Sebastian, P., Doyne Farmer J., & Stefan, T. (2015) To bail-out or to bail-in? Answers from an agent-based model. *Journal of Economic Dynamics and Control* 50144–154 S0165188914002097. <https://doi.org/10.1016/j.jedc.2014.08.020>.
- Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1), 51–59. <https://doi.org/10.1089/big.2013.1508>
- Robbins, S. (1996). *Organizational behavior*. Prentice Hall.
- Sarasvathy, S. D., Dew, N., Velamuri, S. R., & Venkataraman, S. (2010). Three views of entrepreneurial opportunity. In Z. J. Acs & D. B. Audretsch (Eds.), *Handbook of entrepreneurship research: An interdisciplinary survey and introduction* (pp. 77–96). Springer. https://doi.org/10.1007/978-1-4419-1191-9_4
- Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement*. <https://doi.org/10.1177/0013164495055005017>
- Senge, P. M., Kleiner, A., & Roberts, C. (1994). *The fifth discipline fieldbook: Strategies and tools for building a learning organization*. .
- Shafer, G. (1986). The combination of evidence. *International Journal of Intelligent Systems*, 1(3), 155–179. <https://doi.org/10.1002/int.4550010302>
- Shane (2003). *A general theory of entrepreneurship: The individual-opportunity nexus*. Northampton, MA: Edward Elgar Publishing.
- Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *The Academy of Management Review*, 25(1), 217–226. <https://doi.org/10.2307/259271>
- Sharma, S. C. (2006). *Introductory operation research*. Discovery Publishing House.
- Shaver, K. G., & Scott, L. R. (1992). *Person, Process, Choice: The Psychology of New Venture Creation*. *Entrepreneurship Theory and Practice*, 16(2), 23–46.
- Shepherd, D. A., Williams, T. A., & Patzelt, H. (2014). Thinking about entrepreneurial decision making: Review and research agenda. *Journal of Management*. <https://doi.org/10.1177/0149206314541153>
- Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118. <https://doi.org/10.2307/1884852>

- Simon, H. A. (1960). *The new science of management decision* (p. xii, 50). Harper & Brothers. <https://doi.org/10.1037/13978-000>
- Simon, H. A. (1977). *The new science of management decision*. Prentice Hall PTR.
- Smyth, B., & Cunningham, P. (2017). Running with cases: A CBR approach to running your best marathon. In *Case-Based Reasoning Research and Development—25th International Conference, ICCBR 2017, Trondheim, Norway, June 26–28, 2017, Proceedings* (pp. 360–374). https://doi.org/10.1007/978-3-319-61030-6_25
- Smyth, B., & Willemsen, M. C. (2020). Predicting the personal-best times of speed skaters using case-based reasoning. In *Proceedings ICCBR 2020*.
- Snijders, C., Tazelaar, F., & Batenburg, R. (2003). Electronic decision support for procurement management: Evidence on whether computers can make better procurement decisions. *Journal of Purchasing and Supply Management*, 9(5), 191–198. <https://doi.org/10.1016/j.pursup.2003.09.001>
- Spetzler, C. S. (2016). Decision quality: Value creation from better business decisions. Retrieved from <http://www.dawsonera.com/depp/reader/protected/external/AbstractView/S9781119144687>
- Starke, A., Willemsen, M., & Snijders, C. (2017). Effective user interface designs to increase energy-efficient behavior in a Rasch-based energy recommender system. In *Proceedings of the 11th ACM Conference on Recommender Systems* (pp. 65–73). <https://doi.org/10.1145/3109859.3109902>
- Starke, A. D., Willemsen, M. C., & Snijders, C. C. P. (2020). Beyond “one-size-fits-all” platforms: Applying Campbell’s paradigm to test personalized energy advice in the Netherlands. *Energy Research & Social Science*, 59, 101311. <https://doi.org/10.1016/j.erss.2019.101311>
- Tazelaar, F., & Snijders, C. (2004). The myth of purchasing professionals’ expertise. More evidence on whether computers can make better procurement decisions. *Journal of Purchasing and Supply Management*, 10(4), 211–222. <https://doi.org/10.1016/j.pursup.2004.11.004>
- Winston, W. L., & Goldberg, J. B. (2004). *Operations research: Applications and algorithms* (Vol. 3). Thomson/Brooks/Cole.



Digital Entrepreneurship

Wim Naudé and Werner Liebrechts

Contents

- 12.1 Introduction – 281**
- 12.2 What Is Digital Entrepreneurship? – 282**
- 12.3 What Is Different in the Digital Economy? – 284**
 - 12.3.1 How Do Digitization and Digital Artifacts Affect the Nature of Business and of New Venture Creation? – 284
 - 12.3.2 What Are the Implications for Entrepreneurship of the Nature of the Digital Economy? – 287
- 12.4 Digital Platforms and Digital Entrepreneurship – 289**
 - 12.4.1 Creating and Growing a Digital Platform Firm – 290
 - 12.4.2 Competing on Digital Platforms – 291

A condensed version of this chapter has appeared previously in Naudé, Wim; Liebrechts, Werner (2020): Digital Entrepreneurship Research: A Concise Introduction, IZA Discussion Papers, No. 13667, Institute of Labor Economics (IZA), Bonn. Reused with permission.

12.5	Supporting and Regulating Digital Entrepreneurship – 293
12.5.1	Understanding and Supporting Digital Entrepreneurial Ecosystems – 294
12.5.2	Regulating Digital Entrepreneurship – 296
	References – 301

Learning Objectives

After having read this chapter, you will be able to:

- Understand and explain the main concepts, central research questions, and latest theories and empirical evidence in the field of digital entrepreneurship.
- Discuss the effects of typical features of the digital economy on the extent and nature of entrepreneurial activity in such economies.
- Determine what it takes to create and grow a successful digital platform firm and/or to compete successfully on such a digital platform.
- Outline the ecosystems, in which digital entrepreneurs typically operate, and explain how they can be supported and regulated by policymakers.

12.1 Introduction

While the digitization of the economy started in earnest following advances in computing during and after the Second World War, and was given impetus by the commercialization of the personal computer in the 1980s and the invention of the World Wide Web in the 1990s, it was only around 2007 that the deep disruptive potential of the digital revolution became topical. As Friedman (2016) put it: “What the hell happened in 2007?” (p. 19).

“Digitization is the technical process, whereas digitalization is a socio-technological process of applying digitization techniques” (Sussan & Acs, 2017: 58). This digital revolution resulted in technologies such as ubiquitous computing, internet connectivity, digital devices, big data, artificial intelligence (AI), and digital platforms (Cavallo et al., 2019; Coyle, 2017). Consequently, the digital revolution has also made new forms of entrepreneurship possible, has accelerated the creation and scaling up of new businesses, and has changed the contours of competition. As Recker and Von Briel (2019) recognize, “through the infusion of digital technologies into entrepreneurship, entrepreneurial processes become more fluid and porous ... and entrepreneurial outcomes become increasingly malleable, extendable, and modifiable” (p. 4).

The “infusion” of digital technologies into entrepreneurship has resulted in what is known as digital entrepreneurship. Digital entrepreneurship research includes “those studies exploring and (possibly) theorizing on entrepreneurial processes, outcomes and agency transformed by digitization, or by rephrasing it as digital transformation of entrepreneurial processes, outcomes, and agency” (Cavallo et al., 2019: 24). Digital entrepreneurship research is in its infancy. There is a well-recognized need for more research on digital entrepreneurship (Nambisan et al., 2019; Sussan & Acs, 2017). In this light, this chapter provides an overview of the central research questions currently being pursued as well as comments on areas of neglect and avenues for future research.

The central research questions currently being pursued under the topic of digital entrepreneurship are the following: What is digital entrepreneurship? What is different in the digital economy from an entrepreneurial perspective? What is the

impact of digitalization—and big data—on business models and entrepreneurship? How can digital entrepreneurship be supported and regulated? These main research questions and the secondary questions they encompass will be discussed in ► Sects. 12.2–12.5 of this chapter. Then, ► Sect. 12.6 provides a brief summary of the most important conclusions we can draw at this point, including recommendations for addressing issues that are hitherto neglected and, hence, should be addressed in future research.

12.2 What Is Digital Entrepreneurship?

Recognizing who is a digital entrepreneur and who is not is not so straightforward. The digitalization of the economy may be changing the very concept of entrepreneurship. For example, Sussan and Acs (2017: 56) asked “what about Uber drivers and Airbnb renters? Are they digital entrepreneurs?” And what about mobile phone repair shops? Are these owned by digital entrepreneurs, or are they just traditional entrepreneurs benefiting from the rise of the information and communication technology (ICT) sector?

In a sense, one can argue that almost all entrepreneurship now is digital or data driven to the extent that it involves in one way or another computing and a computer. As Varian (2010) puts it: “Sometimes the computer takes the form of a smart cash register, sometimes it is part of a sophisticated point of sale system, and sometimes it is a Web site” (p. 2). As a consequence, virtually all entrepreneurial transactions in the economy are now tracked and stored digitally, as digital artifacts and trade on digital artifact stores.

Trying to further narrow down an answer to the question what digital entrepreneurship is, and to better recognize who a digital entrepreneur is, it is perhaps best to start off with one of the most widely accepted definitions of the field of entrepreneurship, viz. that of Shane and Venkataraman (2000).

Definition of Entrepreneurship

The field of entrepreneurship is defined as “the scholarly examination of how, by whom, and with what effects opportunities to create future goods and services are discovered, evaluated, and exploited” (Shane & Venkataraman, 2000: 218).

Hence, entrepreneurship refers to the process of discovering, evaluating, and exploiting opportunities to create future goods and/or services. These entrepreneurial processes often come with new value creation for the parties involved. To stay close to the definition by Shane and Venkataraman (2000), digital entrepreneurship should first include opportunity recognition and exploitation within the digital economy. The term *digital economy*, by the way, has been ascribed to Tapscott (1995).

Definition of Digital Entrepreneurship

Digital entrepreneurship then is “the pursuit of opportunities based on the use of digital media and other information and communication technologies” (Davidson and Vaast (2010: 2).

Second, digital entrepreneurship should explicitly include the “digital” dimension(s) of the opportunity. As Von Briel et al. (2018) point out, “one clear implication of Shane and Venkataraman’s (2000) framework is that characteristics of ‘that on which they act’ (the opportunity) should influence the venture creation process” (p. 279). In other words, digital entrepreneurship is distinct from traditional entrepreneurship in that the digital nature of the opportunity influences the process of entrepreneurship. To make clear how an opportunity in the digital economy influences the entrepreneurship process, the concept of a *digital artifact* is important.

Definition of Digital Artifacts

Digital artifacts are “man-made purposeful objects embodied in information and communication technology components of software and hardware” (Von Briel et al., 2018: 292).

Digital artifacts can be recombined, edited, and distributed, which can lead to new venture ideas and changes in prices, and in the nature of competition and strategy, in effect leading to what has been described as the “increasingly malleable, extendable, and modifiable” characteristics of entrepreneurial processes (Recker & Von Briel, 2019: 4). Because digital artifacts can be recombined, they offer unlimited scope for new artifact creation. A digital entrepreneur can, for example, offer a new set of products and/or services by recombining existing digital artifacts, such as application programming interfaces (APIs), in a novel manner or by introducing it in a new context. Digital entrepreneurs can therefore be defined as follows.

Definition of Digital Entrepreneurs

Digital entrepreneurs are entrepreneurs who pursue opportunities to produce and trade in digital artifacts on digital artifact “stores” or platforms and/or to create these digital artifact “stores” or platforms (also see Cavallo et al., 2019).

The most common forms of digital entrepreneurship thus include the creation and commercialization of new digital infrastructure, such as platforms, or the creation of value within existing digital platforms (Sussan & Acs, 2017). As such, Uber drivers and Airbnb hostesses are not digital entrepreneurs. Likewise, the owner of a mobile phone repair shop is also not a digital entrepreneur. Nor are millions of entrepreneurs who sell non-digital goods online. Thus, participation on digital

platforms or digital marketplaces is not sufficient to classify an entrepreneur as a digital entrepreneur (as, for instance, Sundarajan (2014) does), nor are using digital technologies in a business (e.g., 3D printing or mobile money). Digital entrepreneurship is recognized by the centrality of digital artifacts and the influence of these artifacts on the nature of the entrepreneurship process. Von Briel et al. (2018) label the ventures started by digital entrepreneurs as *digital ventures* and point out that some of the world's most valuable companies, including Apple, Facebook, Google, and Microsoft, started out as digital ventures whose offering consisted of a digital artifact.

Satisfactory cross-country measures of digital entrepreneurship as defined are lacking (Ojanperä et al., 2019). A number of initiatives in recent years that represent some progress in this direction include the World Bank's *Digital Indicators*¹ that provide comparable measures of digital infrastructure across countries, such as broadband connectivity, digital payment facilities, data privacy and security, and logistics (Chen, 2019). Other initiatives include the World Bank's *Knowledge Economy Index*, the *Digitalization Readiness Index* of UNIDO (2019), and the *Digital Knowledge Economy Index* (DKEI) by Ojanperä et al. (2019). The latter reflects more on digital entrepreneurship by including measures of content creation through digital platforms, such as GitHub (i.e., a code-sharing platform) and Wikipedia (i.e., the renowned crowdsourced encyclopedia).

12.3 What Is Different in the Digital Economy?

A second question that is explored in the current literature on digital entrepreneurship is the following one: What is different in the digital economy from an entrepreneurial perspective? This includes asking subquestions, such as the following: How do digitization and digital artifacts affect the nature of business and of new venture creation? What are the implications for entrepreneurship of the nature of the digital economy? Let us now turn to providing answers to the latter two subquestions one by one.

12.3.1 How Do Digitization and Digital Artifacts Affect the Nature of Business and of New Venture Creation?

In the previous section, it was pointed out that virtually all entrepreneurial transactions in the economy are now tracked and stored digitally, as digital artifacts and trade on digital artifact stores. This "mediation" of transactions by digitization affects entrepreneurship in many different ways, both digital entrepreneurship as defined and more traditional, non-digital entrepreneurship. Varian (2010: 2) discusses four broad types of impact, to which we can add two other relevant ones.

1 See ► <https://www.worldbank.org/en/research/brief/digital-business-indicators>.

One type of impact is that digital technologies allow for the creation of new forms of contracts. For example, revenue-sharing contracts, which are central in most business models of digital platforms, are possible due to the enhanced ability to monitor revenues in a digital space. Much more on digital business models can be found in ► Chap. 13.

A second broad type of impact of the digitization of the economy on entrepreneurship is that it generates data for storage and analysis. The analysis of data is central in many models used by digital ventures and digital platforms to model and influence consumers' behavior. Think of companies using classification or (supervised) segmentation techniques in order to predict customer churn, i.e., who is most likely to leave the company as a client. Then, one might rank prospects by their probability of leaving and allocate a certain incentive budget accordingly (e.g., to the highest probability instances or to the instances with the highest expected loss). Much more on optimizing customer segments using data can be found in ► Chap. 16.

A third impact is that digital spaces make experimentation, production, and diffusion faster, easier, and less costly. This is helpful for the startup of digital ventures, where the fundamental problem has always been that traditional planning methods, such as business planning based on “waterfall” product development and/or past performance, are not appropriate (Bortolini et al., 2018). Easier experimentation has allowed new practical approaches towards launching a new business, such as lean startup approaches (e.g., Blank, 2013; Ries, 2011), to become widely used by digital new ventures (Cavallo et al., 2019).

► Important: The Lean Startup Approach

The lean startup approach (LSA) is “a scientific, hypothesis-driven approach to entrepreneurship, where entrepreneurs translate their vision—i.e. business idea—into falsifiable hypotheses which are embedded in a first version of a business model. These hypotheses are then tested through a series of Minimum Viable Products (MVPs), which are ‘the smallest set of activities needed to disprove a hypothesis’ (Eisenmann et al., 2012: 2)” (Ghezzi & Cavallo, 2020: 521). Such hypothesis tests are very often performed on data gathered by the startup entrepreneurs themselves, for example by attracting (potential) customers to a very minimal first version of an app or website and requiring them to share their opinion and/or other data, either explicitly or implicitly by tracking their behavior (e.g., clicking behavior).

Digital ventures can also engage in much faster product development, making even better use of agile development (AD) practices than traditional firms. AD practices refer to “practices for software development that value the centrality of individuals and interaction, the incremental delivery of working software, collaboration with customers and response to change” (Ghezzi & Cavallo, 2020: 521). All this also changes the role and function of management, away from the importance of opinions, towards rational and decentralized decision-making based on experiments (Varian, 2010).

That is also why Jeff Bezos—founder and CEO of Amazon, and currently the richest man on earth, by far—once said: “The great thing about fact-based deci-

sions is that they overrule the hierarchy.” Hence, whereas top and middle managers used to make top-down decisions in the managerial economy, we now see more and more bottom-up initiatives from lower level employees, backed up with data, in what one could call the entrepreneurial economy. In other words, decisions are increasingly made based on (big) data analytics rather than managers’ intuition gut feeling, hereby shifting the power balance (to a certain extent) to lower level employees. One of the main reasons for this is that data-driven decision-making is believed to improve firm performance (Brynjolfsson & McElheran, 2016; Brynjolfsson et al., 2011; Wamba et al., 2017). Much more on data-driven decision-making can be found in ► Chap. 11.

A fourth impact of the digitization of the economy on entrepreneurship is that digitization and ubiquitous computing enable (hyper) personalization and mass customization. Differential pricing and consumer recommender systems are all based on digitization. Products and services can be developed based on concepts of a “digital twin.” The advent of personalization and custom-made recommendation systems, i.e., using methods of deep learning, has led to huge gains in consumer surplus. For instance, Brynjolfsson et al. (2003) calculated that consumers are benefiting significantly from online retail through paying lower prices, due to greater competition, as well as through the increased variety of offerings online, of which consumers are made (better) aware through recommender systems. Already back in 2000, the consumer surplus due to greater product variety offered online exceeded the welfare gain from increased competition and lower prices in the book market by seven to ten times (Brynjolfsson et al., 2003).

A fifth effect that can be added to the aforementioned ones is the ability to crowdsource inputs and solutions. In the context of the internet, crowdsourcing refers to the sourcing of “digital and material contributions from an on-demand workforce” (Howcroft & Bergvall-Kåreborn, 2018: 21). Platforms that are based on crowdsourcing include platforms that source capital (i.e., *crowdfunding*), ideas (i.e., *crowdsolving*), polling and voting (i.e., *crowdvoting*), and labor (i.e., *crowdwork*) (Howcroft & Bergvall-Kåreborn, 2018).

► Example: Amazon Mechanical Turk

A good example of a platform that sources labor is *Amazon Mechanical Turk* (MTurk). On its website, Amazon explains that MTurk is “a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. This could include anything from conducting simple data validation and research to more subjective tasks like survey participation, content moderation, and more.” They continue to explain that MTurk “enables companies to harness the collective intelligence, skills, and insights from a global workforce to streamline business processes, augment data collection and analysis, and accelerate machine learning development.” Also see ► <https://www.mturk.com/>. ◀

Crowdsourcing also underpins the so-called sharing economy, where digital platforms leverage under- or unutilized assets of users. The use of crowdfunding for entrepreneurial startups broadly—not only of digital ventures—has already generated a fairly large literature (Cavallo et al., 2019; Nambisan et al., 2019).

A sixth clear impact of digitization on entrepreneurship, which is related to the rise of crowdsourcing, is in the ways it is changing entrepreneurial agency. Cavallo et al. (2019: 24) argue that in the digital economy, there is a gradual shift away from the lone entrepreneur towards the community. Don Tapscott (2012), who coined the term *digital economy*, has referred to this as “Capitalism 2.0,” where we are “all collaborating as never before and in business the hottest concepts are social—collective intelligence, mass collaboration, crowd sourcing and collaborative innovation” (p. 3).

12.3.2 What Are the Implications for Entrepreneurship of the Nature of the Digital Economy?

In order to answer the question what is different in the digital economy from an entrepreneurial perspective, it is also necessary to consider the broader context and nature of the digital economy. Here, there are two aspects that are most crucial. The first one is the presence of (indirect) network effects in some markets. The second aspect is about the consequences for countries, regions, firms, and individuals when “certain costs fall substantially and perhaps approach zero” (Goldfarb & Tucker, 2019: 3) due to digitization.

► Important: Network Effects

Network effects (also called network externalities) arise when the number of participants in a market affects the value that everyone obtains on that market. There are both direct and indirect network effects, and they can be either positive or negative. The most prominent example of a direct (and positive) network effect is that of a telephone network, where it becomes more valuable to own a telephone the more people are connected to the telephone network. Indirect network effects refer to network economies, where the value to network members increases for one side of the market if there are more users on the other side of the market (or platform, see ► Sect. 12.4 below). For example, the value of a ride-hailing platform like *Uber* increases for taxi drivers if there are more ride users on the platform, and vice versa.

With indirect network effects being important, demand-side economies of scale—as opposed to supply-side economies of scale, as is the case in “traditional” markets—tend to determine how a market or platform will develop. As described by Parker et al. (2016), “demand economies of scale are the fundamental source of positive network effects, and thus the chief drivers of economic values in today’s world” (p. 20). Digital entrepreneurs therefore tend to put more effort into harnessing positive network effects. This, in turn, tends to make intangible capital and communities, including assets that the entrepreneur does not own—as is the case with *Uber* taxis and *Airbnb* apartments—a more critical focus for digital entrepreneurs than more traditional ones.

Moreover, the network effects and digital economy features described above bring the unfortunate consequence of raising the uncertainty and risk inextricably linked to entrepreneurship. Indeed, the environment, in which digital entrepre-

neurs operate, tends to be more uncertain than that of most traditional forms of entrepreneurship. One manifestation of this is that digital startups tend to go through fast change and innovation at earlier stages of their firms' life cycle, due to the dynamic and uncertain context they face (Ghezzi & Cavallo, 2020). Another manifestation is that, due to the unpredictable growth of new digital ventures, there has been an evolution in equity funding, such as the rise of (groups of) informal investors, also often referred to as angel investors or business angels (Cavallo et al., 2019). Such investors usually invest their own money—often obtained from one or more successful entrepreneurial exits concerning businesses they (co-) founded themselves—in a portfolio of (digital) startups in pursuit of a return on investment.

► Important: Lower Costs Due to Digitization

Digital technologies reduce the cost of storage, computation, and transmission of data. More specifically, in their literature review, Goldfarb and Tucker (2019) emphasize the reduction in five distinct categories of economic costs associated with the rise of digital technologies, viz. (1) search costs, (2) replication costs, (3) transportation costs, (3) tracking costs, and (5) verification costs. “Search costs are lower in digital environments, enlarging the potential scope and quality of search. Digital goods can be replicated at zero cost, meaning they are often non-rival.² The role of geographic distance changes as the cost for transportation for digital goods and information is approximately zero. Digital technologies make it easy to track any one individual's behavior. Last, digital verification can make it easier to certify the reputation and trustworthiness of any one individual, firm, or organization in the digital economy” (Goldfarb & Tucker, 2019: 3–4).

12

The reduction of the aforementioned five types of costs to very low levels, and sometimes even (close to) zero, has a number of important implications for the nature of digital economic activity (Goldfarb & Tucker, 2019). Most importantly, it is easier than ever before to adopt and use digital technologies. Firms are also encouraged to do so, because a large (and growing) number of studies report a positive direct link between digital technology adoption and usage on the one hand, and productivity growth at the firm level on the other hand (e.g., Brynjolfsson & Saunders, 2010; Draca et al., 2009). At the same time, quite some factors are found to enhance or mitigate this relationship—think of firm age, firm size, and potential for network effects—and thus, not every firm benefits to the same extent, and some not at all.

Lower costs due to digitization also provide unprecedented opportunities for (digital) entrepreneurs to create new value by means of innovative business models (e.g., Brousseau & Penard, 2007). Again, much more on digital business models can be found in ► Chap. 13.

² In case of non-rivalrous goods, increased demand does not affect the supply left for other individuals. A good example is Netflix, where more views of the movies and series offered by them do not have any effect on the opportunities for other people to also watch these movies and/or series.

12.4 Digital Platforms and Digital Entrepreneurship

Digital platforms have become one of the most discussed forms of digital entrepreneurship, as a growing literature attests to. This literature has studied the design and development of such platforms; their social, business, and economic impacts; and the regulatory challenges that they bring.

► Important: Digital Platforms

A straightforward definition of a digital platform is lacking, which also makes it more difficult to design policies to regulate them. Similar to “traditional” platforms, such as newspapers bringing together readers and advertisers, a digital platform fulfills an intermediate (or matching) function between various users, but then in the digital economy. Coyle (2017) defines a platform as “a business strategy as much as an organization” (p. R5), and many scholars share this idea that digital platforms can be considered both firms and markets (Chen, 2019).

Generally speaking, digital platforms contain four kinds of participants (who often switch roles or fulfill more than one role at once). These are (1) the owners of the platform, (2) the producers of the content, (3) the customers who consume the content, and (4) the providers of the interfaces between the owners, producers, and customers (Van Alstyne et al., 2016). A distinction can be made between one-way digital platforms (e.g., Spotify), two-sided ones (e.g., Uber), and multi-sided ones (e.g., Microsoft) (Litan, 2016). Multi-sided platforms tend to be intermediaries or matchmakers and often do not even produce their own content (e.g., Facebook) (Nuccio & Guerzoni, 2018).

Digital platforms have themselves changed the nature of competition in markets and disrupted many traditional so-called pipeline business models. Oft-quoted examples are of Amazon upending traditional booksellers (such as Borders) or Netflix upending traditional video-rental firms (such as Blockbuster). This has led Parker et al. (2016) to warn that “almost in every occasion when a platform enters a market, relative to a product, the platform almost always wins.” While digital platforms are prone to dominate their market, for reasons that will be explained below, they lead to further disruption through enabling third-party entrepreneurs to start new digital ventures on the platform. For this reason, Litan (2016) considers digital platforms as “launching pads for new and potentially disruptive firms” (p. 581).

Thus, digital platforms are an essential phenomenon in digital entrepreneurship. As it was defined in ► Sect. 12.2, digital entrepreneurs pursue opportunities to produce and trade in digital artifacts on platforms and/or to create these platforms themselves. In the remainder of this section, these two ways of using digital platforms for entrepreneurship will be discussed, but then in reversed order.

12.4.1 Creating and Growing a Digital Platform Firm

First, consider the entrepreneurial act of creating and growing a digital platform firm. In light of the fact that digital platforms offer unparalleled scope for rapid scale-up, due to network effects and demand-side economies of scale, the establishment of a new digital venture that can become a global digital platform has become somewhat of the *ne plus ultra* of digital entrepreneurship.

Both entrepreneurs and venture capitalists have come to chase after the next Facebook or Netflix. Startup accelerators explicitly aim to create the next “unicorn,” i.e., a startup firm that is valued in excess of a billion dollars. Indeed, the scaling up of digital platforms is often fast and their market valuation exorbitant. For example, Chen (2019) describes that “social media platform ByteDance and ride-hailing platform Didi Chuxing from China are valued more than the GDP of many developing countries such as Kyrgyzstan, Uganda and Zambia” (p. 5). Scaling up a digital venture to become a billion-dollar digital platform is—notwithstanding the aforementioned prominent examples—extremely difficult. As Sussan and Acs (2017) pointed out, “almost everyone who tries to build one fails” (p. 68).

The central challenge that a digital entrepreneur faces in creating and growing a digital platform firm is to maximize the indirect positive network effects. This is a source of both success and failure, since, as Rochet and Tirole (2003) pointed out, “platform owners or sponsors in these industries must address the celebrated ‘chicken-and-egg problem’ and be careful to ‘get both sides on board’” (p. 990).

► Important: Growing a Digital Platform Firm

Digital platform owners will typically attempt to grow the number of users on their platforms by actively building a community, encouraging collaboration between different users, maintaining good communications, working on extending connections, and perhaps most importantly curating the content of the platform. These are the so-called 5 Cs to building a successful platform (*community, collaboration, communication, connection, and curation*, respectively) (see Johnson, 2020).

The aforementioned 5 Cs are dependent on digital technologies, such as rating and recommendation systems, and on matching algorithms (Sutherland & Jarrahi, 2018). If successful, this can lead to a positive feedback loop between customers on both sides of the platform. In this feedback loop, the extent of data that platforms collect from their customers will determine their success and competitiveness; the more data, the better they can predict customer behavior, refine matching algorithms, and hence tailor their product(s) and service(s) to customers’ needs and wants (Nuccio & Guerzoni, 2018). As such, the development and use of data analytical tools, including artificial intelligence (AI), are key for digital entrepreneurs.

A second and related challenge that digital entrepreneurs face in the establishment of a successful digital platform is that it requires significant outlays on fixed costs. It is therefore mistaken to assume that, because many costs in the digital economy have fallen significantly (see ► Sect. 12.3.2 above) and many scholars and policymakers describe digital entrepreneurial entry as easy, there are not funda-

mental costs to incur. For instance, Nuccio and Guerzoni (2018) report that Google's capital expenditure peaked at US \$10.9 billion in 2016.

➤ **Important: Operating a Successful Digital Platform Firm**

If digital platforms can obtain large numbers of users—and hence generate big data—then their business models can become highly profitable over time. This is due to the fact that marginal costs of extending their service to a new customer or user are very low. It is this combination of high fixed costs of operating the platform and low marginal costs for each additional customer which results in successful digital platforms becoming very profitable, or at least having the promise of high profit growth in the (nearby) future. These digital platforms can then become superstar firms.

This also accentuates the first-mover advantage of establishing a digital platform (Nuccio & Guerzoni, 2018), which makes it very difficult for new entrants. Litan (2016) therefore advocates a new role for antitrust policy to ensure adequate competition, both between different platforms and on the platforms themselves.

12.4.2 Competing on Digital Platforms

Digital entrepreneurs pursue opportunities to produce and trade in digital artifacts on platforms and/or create these platforms. In the previous subsection, the creation and growth of digital platforms were discussed. This subsection discusses some of the key challenges and features of digital entrepreneurs when competing on digital platforms.

The growth and dominance of digital platforms in the digital economy have come to mean that “ultimately most firms will have no choice but to do business on somebody else's digital property, and to agitate for better terms if the owner gets too greedy. Call it the class struggle of platform capitalism” (The Economist, 2016). This has both positive and negative consequences for digital entrepreneurship. Some would even argue that the negative implications of “platform capitalism” outweigh the positive consequences. While this issue cannot be adjudicated in this chapter, some of the positive and negative consequences can be highlighted.

On the positive side, participation on digital platforms has offered many opportunities for micro entrepreneurs (Howcroft & Bergvall-Kåreborn, 2018). This includes opportunities for digital artifact creation, most often app development as on Apple's iOS platform or on Google's Android platform. For app developers, the platform is a marketplace to connect with the owners of computing devices, such as mobile phones, tablets, and computes (Van Alstyne et al., 2016). By 2015, there were already 1.4 million apps in Apple's App Store, generating revenue estimated at US \$25 billion for the developer entrepreneurs (Van Alstyne et al., 2016). By 2019, this number stood at 1.8 million apps. At the end of 2019, the four major app platforms offered over 5.5 million apps altogether, viz. Google Play (2.57 million apps), Apple App Store (1.84 million apps), Windows Store (669,000 apps), and Amazon App Store (489,000 apps).

Further on the positive side, digital platforms are also judged to hold out promise for (recombinant) innovation by entrepreneurs as a result of the possibilities of recombining digital artifacts that are “open, reprogrammable, and accessible by other digital objects” (Parker et al., 2017: 256). To harness this possibility, many of the largest digital platforms, such as Apple, Google, and Microsoft, have shifted part of their innovation outside of the core firms to developers (i.e., many micro entrepreneurs) in its platform ecosystem and provide their own platform resources and advantages to these entrepreneur-developers (Parker et al., 2017).

How various digital platforms govern their ecosystems to facilitate and control digital entrepreneur-developers to create and benefit from new apps depends on the platform’s strategic model, and whether or not it emphasizes openness (and “permissionless” innovation) or control. In this regard, Parker et al. (2017: 256–257) contrast the governance models of Apple iOS and Google Android, showing that while Google Android is more open and thus generates more app development and innovative activities by micro entrepreneurs, the more controlled Apple iOS environment is more profitable, but perhaps less innovative. This points to the fact that a key strategic decision by platform owners is how open they should be, and how to manage their openness in order to minimize negative (demand) externalities and bad behavior, such as scamming and spamming (Van Alstyne et al., 2016; Coyle, 2017).

Regarding the negative effects of digital platforms on entrepreneurship, a major fear is that as digital platforms gain market power they will drive traditional small businesses out of the market and will reduce the traditional and typical sources of work. Given these concerns, Howcroft and Bergvall-Kåreborn (2018) are of the opinion that “the claim that crowdwork is nurturing enterprise is highly questionable” (p. 24).

Another fear is that entrepreneurs on digital platforms may be especially prone to role conflict, which could increase their stress and reduce their performance (Nambisan & Baron, 2019). The reason for role conflict on digital platforms stems from the governance by the platform owner, which could conflict with the goals of the individual entrepreneurs on that platform. For instance, the platform owner faces the incentive to increase the number of users on the platform and may engage in actions to increase this, which may be detrimental to the revenues of independent entrepreneurs already operating on the platform. Think of the platform owner forcing them to certain price discounts. As such, the issue is that digital entrepreneurs operating on a platform may lose some of their independence and become (more) dependent on the platform owner (Nambisan & Baron, 2019).

Finally, digital platform entrepreneurship affects not only digital entrepreneurship, both between platforms and on platforms, but also traditional entrepreneurship. Again, the effects are both positive and negative. One positive effect is that many traditional firms are benefiting from the accumulation of data by digital platforms. Examples include the production of wearable devices (e.g., Fitbit), which increase in value (i.e., higher consumer surplus) through a connection to software driven by growing volumes of data on the cloud.

The most significant effect, however, is probably the impact of competition from digital platforms and on-platform entrepreneurs on traditional firms. Burtch

et al. (2018) study how digital platforms affect local entrepreneurial activity, particularly the entry and exit of entrepreneurs. They start from the idea that digital platforms may facilitate entry, for instance by offering work flexibility and by reducing entry costs but moving on to stating that they may also reduce entry, since they offer alternatives to self-employment in the gig economy.

Definition of Gig Economy Platform

Gig economy platforms are “digital, on-demand platforms that enable a flexible work arrangement” (Burtch et al., 2018: 5497).

In essence, gig economy platforms may raise the opportunity costs of entrepreneurship, meaning that getting involved in local entrepreneurial activity may become less attractive when such platforms enter the local market. Burtch et al. (2018) test this using data on the effect of Uber (i.e., the ride-hailing platform) and Postmates (i.e., an on-demand delivery platform) entering local areas on crowdfunding campaign launches on Kickstarter (i.e., a crowdfunding platform). Taking the rate and volume of crowdfunding campaign launches as a measure of entrepreneurial activity, the authors find a negative and significant relationship between platform entry and local entrepreneurial activity. Also, “gig-economy platforms predominantly reduce lower quality entrepreneurial activity, seemingly by offering viable employment for the unemployed and underemployed” (Burtch et al., 2018: 5497).

In conclusion, digital platforms have become one of the most discussed forms of digital entrepreneurship. Digital entrepreneurs create and grow digital platforms, but they also compete on such platforms. These kinds of digital entrepreneurship have become substantial and significant, with impacts extending to the traditional, non-digital entrepreneurial sphere. There is a growing concern that digital platforms are not all that good news for entrepreneurship (e.g., Howcroft & Bergvall-Kåreborn, 2018). Others have argued, however, that there is not yet sufficient research on the negative implications of digital platforms on entrepreneurship (e.g., Nambisan & Baron, 2019). Clearly, this is an avenue for future research.

12.5 Supporting and Regulating Digital Entrepreneurship

A third broad question that the emerging field of digital entrepreneurship has tried to answer is how digital entrepreneurship can be supported and regulated. Here, research has focused on two main aspects. That is, how to understand, describe, and strengthen digital entrepreneurial ecosystems, and how to regulate digital entrepreneurship, in particular given the tendency of network effects and demand-side economies of scale to lead to winner-takes-it-all outcomes and market dominance by only a few superstar firms. In this section, these two main aspects will be discussed in more detail.

12.5.1 Understanding and Supporting Digital Entrepreneurial Ecosystems

There are many ways to define what an *entrepreneurial ecosystem* (EE) is, and hence, we do not just share one of them. Instead, a number of complementary definitions should lead to a more comprehensive understanding of the concept.

Definition of Entrepreneurial Ecosystem

According to Acs et al. (2014), an entrepreneurial ecosystem refers to “the dynamic institutionally embedded interaction between entrepreneurial attitudes, abilities and aspirations, by individuals, which drives the allocation of resources through the creation and operation of new ventures” (p. 479). It consists of “sets of actors, institutions, social networks, and cultural values that produce and sustain entrepreneurial activity” (Roundy et al., 2018: 1). According to Stam (2014), an entrepreneurial ecosystem is “an interdependent set of actors that is governed in such a way that it enables entrepreneurial action. It puts entrepreneurs center stage but emphasizes the context by which entrepreneurship is enabled or constrained” (p. 1).

Current thinking in entrepreneurship support policy is that governments and other agencies should not try to identify and support individual, potential high-growth enterprises (or, put differently, pick winners), but rather provide an ecosystem that is conducive for the emergence of such firms. Modern entrepreneurship support policy is thus aiming at strengthening entrepreneurial ecosystems.

In the case of digital entrepreneurship, similar considerations apply. Hence, supporting policies for digital entrepreneurship need to understand the *digital entrepreneurial ecosystem* (DEE). However, the DEE is more complex, because, given that the production of and trade in digital artifacts are central in digital entrepreneurship, there is also a *digital ecosystem* to contend with.

Definition of Digital Ecosystem

A digital ecosystem is “a self-organizing, scalable and sustainable system composed of heterogeneous digital entities and their interrelations focusing on interactions among entities to increase system utility, gain benefits, and promote information sharing, inner and inter cooperation and system innovation” (Sussan & Acs, 2017: 58; Li et al., 2012).

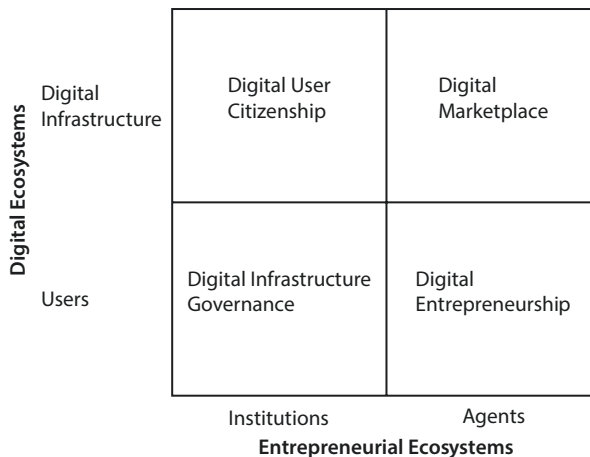
Digital entrepreneurs operate in the DEE, which is at the intersection of the EE and the digital ecosystem (Sussan & Acs, 2017: 62).

Definition of Digital Entrepreneurial Ecosystem

The digital entrepreneurial ecosystem is defined as “the matching of digital customers (users and agents) on platforms in digital space through the creative use of digital ecosystem governance and business ecosystem management to create matchmaker value and social utility by reducing transactions cost” (Sussan & Acs, 2017: 63).

As described by Sussan and Acs (2017), the DEE comprises *digital infrastructure* (DI), users of digital infrastructure and digital artifacts, entrepreneurial agents, and prevailing formal and informal institutions (“rules of the game in a society,” see North, 1990) that shape their interaction. Based on these components, Sussan and Acs (2017) provide a conceptual framework from which to approach understanding and researching the digital entrepreneurial ecosystem (DEE). Their conceptual framework can be explained with reference to ■ Fig. 12.1.

The two-by-two diagram in ■ Fig. 12.1 depicts the two dominant concepts, viz. digital ecosystems (left) and entrepreneurial ecosystems (bottom). In turn, digital ecosystems comprise digital infrastructure and users of such digital infrastructure, and entrepreneurial ecosystems include (entrepreneurial) agents and the institutional context in which they operate. The four quadrants of the conceptual framework are (clockwise, starting at top left) digital user citizenship (DUC), digital marketplace (DM), digital entrepreneurship (DE), and digital infrastructure governance (DIG). Each concept is interrelated with all the others, and a proper understanding is needed for a digital entrepreneurial ecosystem to function properly and to be sustainable.



■ Fig. 12.1 Conceptual framework of the digital entrepreneurial ecosystem. (Source: Sussan & Acs, 2017: 63)

First, digital user citizenship (DUC) addresses “the explicit legitimization and implicit social norms that enable users to participate in a digital society” (Sussan & Acs, 2017: 64). Here, users can be either entrepreneurial agents in the digital economy or customers. Anyway, better chances to participate in a digital society are expected to be congruent to and supportive of digital entrepreneurial activity in that society. Second, a digital marketplace (DM) represents “the combination of users and agents within the context of both ecosystems” (Sussan & Acs, 2017: 65). This quadrant is about value creation through new products, services, and/or knowledge resulting from entrepreneurial activities and about value capture by users that embrace them. DMs are seen as key to sustainable DEEs. Third, digital entrepreneurship (DE) in this conceptual framework includes “any agent that is engaged in any sort of venture, be it commercial, social, government, or corporate that uses digital technologies” (Sussan & Acs, 2017: 66). Hence, this is not in line with the way digital entrepreneurs have been defined in ► Sect. 12.2, since Uber drivers could now be seen as such agents. By leveraging digital technologies and seeking and acting on opportunities within DMs, digital entrepreneurs (as defined here) are believed to increase an economy’s efficiency by moving it closer to the technological frontier (Sussan & Acs, 2017). Fourth and last, digital infrastructure governance (DIG) is about “the coordination and governance needed in order to establish a set of shared technological standards that are related to entrepreneurial activities” (Sussan & Acs, 2017: 64). It is suggested that such governance “is likely the most open, transparent, and informal” at the beginning, when many digital entrepreneurs are “essentially forcing the creation of new rules,” until their disruptive activities reach a certain momentum, and the DIG suddenly becomes “less open, less transparent, and more formal” (Sussan & Acs, 2017: 64). In brief, DIG supposedly has an inverted U-shaped relationship with sustainable DEEs. The authors derive a number of propositions from the conceptual framework (► Fig. 12.1), which they offer for future researchers to evaluate.

12.5.2 Regulating Digital Entrepreneurship

The unique regulatory challenges posed by the emergence of digital entrepreneurship are due to the typical features and consequences of digital infrastructure and entrepreneurial ecosystems interacting. This subsection will explain these challenges and will indicate the conundrums they pose for regulators.

The first challenge is how to define digital entrepreneurship, and moreover how to define a digital platform for the purposes of regulation. As was argued in ► Sect. 12.2, digital entrepreneurship is distinctive due to the centrality of digital artifacts and their influence on the process of entrepreneurship. Traditional entrepreneurs who sell goods online or, for example, drive a taxi as part of the Uber ride-hailing platform do not produce or trade digital artifacts, and merely use a digital artifact (e.g., the Uber app) to facilitate a part of their business. The owners of the Uber platform, however, are digital entrepreneurs, as they have created a digital artifact and used it to establish and grow a firm. Regulating the Uber platform as distinct

from regulating the self-employed Uber drivers making use of the platform is a challenge. The Uber example given here is representative of the challenge. For instance, while the self-employed drivers are competing against each other, Uber may or may not be a monopoly, or it can become a monopoly if it would drive competitor taxi firms out of the market.

Therefore, the difficulty that policymakers face is to determine whether a digital platform firm is a monopolist or not. If prices are considered, these mostly do not show signs of price collusion or markup pricing, due to the tendency of consumer prices to decline in the digital economy (see ► Sect. 12.3.2 above). If market share is considered, it begs the question in which domain, since many digital platforms have spread their brand image across different domains. For example, Google not only provides a search engine, but also advertising space, translation services, and even driverless cars; Facebook provides not only connectivity, but also finance and a marketplace; and Amazon not only sells books, but also owns food stores (Rossotto et al., 2018; Van Alstyne et al., 2016).

A second major challenge that regulators face is precisely due to this domain crossing (also often referred to as “shape shifting” or “envelopment”). Shape shifting allows the digital platform to benefit from “regulatory arbitrage.” An example is that of the already mentioned Uber entering into taxi transportation, but without being subject to the regulations applying to more traditional taxi firms (Chen, 2019). In essence, shape shifting by digital entrepreneurs makes it difficult to define a digital platform. The lack of definitional clarity, compounded by the speed at which digital entrepreneurs can act and metamorphize, means that digital platforms can occupy “legal grey areas” (Coyle, 2017: R6) and that digital entrepreneurs may outrun the regulator (Sussan & Acs, 2017).

A third major challenge that regulators face with respect to digital entrepreneurship and digital platform entrepreneurs is due to them possessing substantial intangible assets, including their relative intangible physical presence. Digital entrepreneurs reside in a digital space and may not be tied to any one physical location. This, and the complexity in defining and delineating a digital platform as was discussed earlier on, allows digital platforms to avoid taxation through selection of jurisdiction for reporting profits and use of transfer pricing (Chen, 2019; Nuccio & Guerzoni, 2018; Rossotto et al., 2018).

A fourth challenge for regulators is due to the nature and extent of innovation by digital platforms and their entrepreneurs. Their innovation has been seen as a way to attain and sustain market dominance. Chen (2019) explains that this can be through proactive acquisition of possible rivals—that is, merger and acquisition (M&A) activities as a substitute for research and development (R&D) activities—and/or by copying a new rival’s product or service, also described as market consolidation (Rossotto et al., 2018). Other strategies could involve patent thickets and other defensive innovation strategies (Shapiro, 2001). The problem that regulators face in regulating this as anticompetitive behavior is that antitrust authorities generally consider innovation as a mitigating behavior of firms that enjoy monopoly profits. As Nuccio and Guerzoni (2018) point out, antitrust laws “punish not market power per se, but its abuse” (p. 317). Abuse would typically manifest itself in

higher prices, discriminatory prices, and large markups or margins without significant innovation. As few of the global digital platforms seem guilty of these abuses—and, in fact, engage in innovation and offer considerable consumer surplus—Nuccio and Guerzoni (2018) conclude that, in the specific case of digital platforms, the consequences of high levels of market concentration may not be that harmful.

A fifth challenge that regulating entrepreneurship poses is that abuse by digital entrepreneurs may be taking different forms than the abovementioned, more traditional types of monopolistic market power abuse. New forms of abuse include data privacy and security violations, and consumer and voter manipulation. As these abuses relate to data, regulators have focused scrutiny on the ability of digital platforms to accumulate big data. What is the implication when data becomes a valuable commodity? Could and should data be protected and shared? A major challenge is that “the market power obtained by access to or the holding of vast amounts of data connected to algorithms may create barriers to entry for second movers” (Lundqvist, 2017: 713). Other challenges in this regard include limiting cybercrime, data misuse, and a general lack of trust in the digital economy (Chen, 2019).

A further new form of abuse by digital entrepreneurs is the possible exploitation of workers that are active on gig economy platforms. There are growing concerns in this regard, because the gig economy has grown exponentially at the same time that there has been rising concern over the exploitation of workers on these platforms (Howcroft & Bergvall-Kåreborn, 2018). The situation of these workers, who are not employed, but are independent contractors or freelancers, is a concern as they are unregulated, they do mostly micro tasks (or gigs) at low rates of remuneration, their performance and evaluation management is often subject to “algorithmic control,” and they mostly have little legal recourse against poor labor practices and working conditions (Howcroft & Bergvall-Kåreborn, 2018).

Finally, a sixth challenge that is perhaps not so much a regulatory challenge, but a challenge of global governance and the outcome of the new challenges to regulation that digital entrepreneurship poses, is the existence and widening of digital gaps. While digital technologies can in principle diffuse instantaneously, practice has seen many obstacles to the diffusion and the adoption of digital technologies that support digital entrepreneurship. For instance, UNIDO (2019) found that the creation and diffusion of advanced digital production (ADP) technologies “remain concentrated globally” (p. 1). Also, “... ten economies—the frontrunners—account for 90 percent of all global patents and 70 percent of all exports directly associated with these technologies” (p. 1). Given digital gaps, concerns have been voiced about the dangers of “data colonialism” by the actions of global platform firms in emerging economies (Rossotto et al., 2018).

In conclusion, while the regulatory challenges posed by digital entrepreneurship are substantial, the generation of large volumes of data by entrepreneurs through and on the digital economy can, in fact, help authorities and support agencies in their governance functions. The digital footprints and digital shadows cast by entrepreneurs online will allow matching scarce resources with entrepre-

neers who are more likely to succeed. Indeed, as far as entrepreneurial success is concerned, the current consensus is still that it is largely unpredictable. With large datasets becoming available, a number of scholars have recently argued that it will become easier to predict success, and thus tailor support and other governance measures (Menon, 2018). Ng and Stuart (2016), for example, taking the career histories of two million entrepreneurs and using machine learning algorithms, classify entrepreneurs into “hobos” and “highfliers,” with hobos being “self-employed entrepreneurs who often depart relatively low-wage jobs and may further sacrifice income for the autonomy of self-employment” and highfliers being individuals who “exit high-wage, high-advancement careers to launch high potential companies” (p. 5). This is a promising line of future research that offers the potential to improve the allocation and efficiency of public support policies for all entrepreneurs.

Conclusion

The main purpose of this chapter was to provide an overview of state-of-the-art knowledge in the field of digital entrepreneurship research. With this goal in mind, a selection of latest theories and empirical evidence have been discussed with regard to a number of key research questions that are currently being pursued in this field.

The chapter started by defining the main concepts in the field. This is important, since it is not so clear-cut how to pinpoint digital ventures or digital entrepreneurs. In essence, digital entrepreneurship refers to the pursuit of opportunities based on the use of digital technologies. Digital entrepreneurs produce and trade in so-called digital artifacts on digital artifact “stores” (or platforms) or they create these digital platforms themselves.

The chapter then moved on to discussing the most important effects of the nature of the digital economy on entrepreneurial activity. The various impacts of digitization on entrepreneurship that have been discussed clearly illustrate why the digitization and digitalization of (mostly developed) economies have led to serious and lasting changes in the entrepreneurial landscape.

The next section has been devoted to describing one of the most discussed forms of digital entrepreneurship, namely digital platforms. Digital entrepreneurs create and grow such platforms or compete on it. The presence and impact of digital platforms have become substantial, with implications extending to traditional, non-digital entrepreneurship. Digital platforms come with both positive and negative consequences, but more research is needed on any of these issues to clearly judge which ones outweigh the others.

Finally, it is of the utmost importance to understand the main features of the context in which digital entrepreneurs typically operate. For this, the conceptual framework of the digital entrepreneurial ecosystem presented in ► Sect. 12.5.1 can be of help. However, future research should still focus on testing the various propositions that have been derived from it. Also, no matter how well policymakers’ understanding is, regulatory challenges posed by digital entrepreneurship remain substantial.

Discussion Points

1. As mentioned, recognizing who is a digital entrepreneur and who is not is not so straightforward. Now knowing how to define digital entrepreneurship and a digital entrepreneur, try to come up with at least three different types of entrepreneurs operating in the digital economy. Would you consider them as digital entrepreneurs? Argue why (not).
2. We have discussed six types of impact of the digitization of the economy on entrepreneurial activity. For example, that it has led to a plethora of data providing a lot of entrepreneurial opportunities (i.e., the second type of impact in ► Sect. 12.3.1). Name and explain at least one specific example of entrepreneurial activity (a certain firm, whether new or already established, a certain entrepreneur, etc.) per impact type.
3. Now suppose that you are willing to develop a successful digital platform firm. Describe as precise as possible what is required. Which conditions must be met? What activities should you undertake to build and grow the platform? What does your success (or failure) depend on?
4. The growth and dominance of digital platforms in the digital economy, and the consequences that come with it, have been referred to as “platform capitalism” by The Economist (2016). This relatively new form of capitalism has both positive and negative consequences for the digital entrepreneurs involved as well as for more traditional, non-digital entrepreneurs. Argue why you think that the positive consequences outweigh the negative ones, or vice versa.
5. Now step into the shoes of a regulator concerned with competition policy. How would you define a digital platform? When does a digital platform have monopolistic market power according to you? Is a digital platform having such market power actually a reason to intervene? Argue why (not). If yes, which policy measures can you implement?

Take-Home Messages

- In a digital economy, digital entrepreneurs pursue opportunities to produce and trade in digital artifacts on so-called digital artifact stores or platforms and/or to create these digital artifact “stores” or platforms themselves.
- Typical features of the digital economy, such as the presence of (indirect) network effects and digital technologies reducing a number of important economic costs, have a number of relevant effects on the extent and nature of entrepreneurial activity in such economies.
- Digital platforms are well-known and often-discussed forms of digital entrepreneurship, which typically attempt to become successful by taking care of the so-called 5 Cs, viz. community, collaboration, communication, connection, and curation.
- The growth and dominance of digital platforms come with both positive and negative consequences for digital entrepreneurs competing on digital platforms, but more research is needed to determine which ones outweigh the others.

- A better understanding of the characteristics of and actors in a digital entrepreneurial ecosystem, in which digital entrepreneurs typically operate, allows one to make appropriate policies that support and/or regulate digital entrepreneurship, if necessary at all.

References

- Acs, Z. J., Autio, E., & Szerb, L. (2014). National systems of entrepreneurship: Measurement issues and policy implications. *Research Policy*, 43(3), 476–494.
- Blank, S. (2013). Why the lean start-up changes everything. *Harvard Business Review*, 91(5), 635–672.
- Bortolini, R. F., Nogueira Cortimiglia, M., Danilevicz, A. D. M. F., & Ghezzi, A. (2018). Lean startup: A comprehensive historical review. *Management Decision*. <https://doi.org/10.1108/MD-07-2017-0663>
- Brousseau, E., & Penard, T. (2007). The economics of digital business models: A framework for analyzing the economics of platforms. *Review of Network Economics*, 6(2), 81–114.
- Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). *Strength in numbers: How does data-driven decision making affect firm performance?* SSRN Working Paper. Available at SSRN 1819486.
- Brynjolfsson, E., Hu, Y., & Smith, M. D. (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, 49(11), 1580–1596.
- Brynjolfsson, E., & McElheran, K. (2016). The rapid adoption of data-driven decision making. *American Economic Review*, 106(5), 133–139.
- Brynjolfsson, E., & Saunders, A. (2010). *Wired for innovation: How information technology is reshaping the economy*. MIT Press.
- Burtch, G., Carnahan, S., & Greenwood, B. N. (2018). Can you gig it? An empirical examination of the gig economy and entrepreneurial activity. *Management Science*, 64(12), 5497–5520.
- Cavallo, A., Ghezzi, A., Dell’Era, C., & Pellizzoni, E. (2019). Fostering digital entrepreneurship from startup to scaleup: The role of venture capital funds and angel groups. *Technological Forecasting and Social Change*, 145, 24–35.
- Chen, R. (2019). *Policy and regulatory issues with digital business*. Policy Research Working Paper No. 8948. Washington DC: The World Bank.
- Coyle, D. (2017). Precarious and productive work in the digital economy. *National Institute Economic Review*, 240, R5–R14.
- Davidson, E., & Vaast, E. (2010). Digital entrepreneurship and its sociomaterial enactment. In *Proceedings of the 43rd Hawaii International Conference on System Sciences* (pp. 1–10). IEEE Computer Society Press.
- Draca, M., Sadun, R., & Van Reenen, J. (2009). Productivity and ICTs: A review of the evidence. In C. Avgerou, R. Mansell, F. Quah, & R. Silverstone (Eds.), *The Oxford handbook of information and communication technologies* (pp. 100–147). Oxford University Press.
- Eisenmann, T. R., Ries, E., & Dillard, S. (2012). *Hypothesis-driven entrepreneurship: The lean startup*. Harvard Business School Entrepreneurial Management Case No. 812-095.
- Friedman, T. L. (2016). *Thank you for being late: An optimist’s guide to thriving in the age of accelerations*. Farrar Straus and Giroux.
- Ghezzi, A., & Cavallo, A. (2020). Agile business model innovation in digital entrepreneurship: Lean startup approaches. *Journal of Business Research*, 110, 519–537.
- Goldfarb, A., & Tucker, C. (2019). Digital Economics. *Journal of Economic Literature*, 57(1), 3–43.
- Howcroft, D., & Bergvall-Kareborn, B. (2018). A typology of crowdwork platforms. *Work, Employment and Society*, 33(1), 21–38.
- Johnson, N. L. (2020). What are network effects? *Applico*. Retrieved July 13, 2020, from <https://www.applico.com/blog/network-effects/>

- Li, W., Badr, Y., & Biennier, F. (2012). Digital ecosystems: Challenges and prospects. In *Proceedings of the International Conference on Management of Emergent Digital Ecosystems* (pp. 117–122). ACM.
- Litan, R. E. (2016). Entrepreneurship, innovation, and antitrust. *The Antitrust Bulletin*, 61(4), 580–594.
- Lundqvist, B. (2017). Standardization for the digital economy: The issue of interoperability and access under competition law. *The Antitrust Bulletin*, 62(4), 710–725.
- Menon, C. (2018). Mixing experimentation and targeting: Innovative entrepreneurship policy in a digitized world. In OECD (Ed.), *OECD science, technology and innovation outlook 2018* (pp. 297–295). OECD.
- Nambisan, S., & Baron, R. A. (2019). On the costs of digital entrepreneurship: Role conflict, stress, and venture performance in digital platform-based ecosystems. *Journal of Business Research*. <https://doi.org/10.1016/j.jbusres.2019.06.037>
- Nambisan, S., Wright, M., & Feldman, M. (2019). The digital transformation of innovation and entrepreneurship: Progress, challenges and key themes. *Research Policy*, 48, 1–9.
- Ng, W., & Stuart, T. E. (2016). *Of hobos and highfliers: Disentangling the classes and careers of technology-based entrepreneurs*. Unpublished Working Paper.
- North, D. C. (1990). A transaction cost theory of politics. *Journal of Theoretical Politics*, 2(4), 355–367.
- Nuccio, M., & Guerzoni, M. (2018). Big data: Hell or heaven? Digital platforms and market power in the data-driven economy. *Competition & Change*, 23(3), 312–328.
- Ojanperä, S., Graham, M., & Zook, M. (2019). The digital knowledge economy index: Mapping content production. *Journal of Development Studies*, 55(12), 2626–2643.
- Parker, G. G., Van Alstyne, M. W., & Choudary, S. P. (2016). *Platform revolution: How networked markets are transforming the economy and how to make them work for you*. W.W. Norton & Company.
- Parker, G. G., Van Alstyne, M. W., & Jiang, X. (2017). Platform ecosystems: How developers invert the firm. *MIS Quarterly*, 41(1), 255–266.
- Recker, J., & Von Briel, F. (2019). The future of digital entrepreneurship research: Existing and emerging opportunities. In *Fortieth International Conference on Information Systems, Munich*.
- Ries, E. (2011). *The lean startup*. Crown Business.
- Rochet, J. C., & Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4), 990–1029.
- Rosotto, C. M., Das, P. L., Ramos, E. G., Miranda, E. C., Badran, M. F., Licetti, M. M., & Murciego, G. M. (2018). Digital platforms: A literature review and policy implications for development. *Competition and Regulation in Network Industries*, 19(1–2), 93–109.
- Roundy, P. T., Bradshaw, M., & Brockman, B. K. (2018). The emergence of entrepreneurial ecosystems: A complex adaptive systems approach. *Journal of Business Research*, 86, 1–10.
- Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *Academy of Management Review*, 25(1), 217–226.
- Shapiro, C. (2001). Navigating the patent thicket: Cross licenses, patent pools, and standard-setting. In A. B. Jaffe (Ed.), *Innovation policy and the economy* (pp. 119–150). MIT Press.
- Stam, E. (2014). The Dutch entrepreneurial ecosystem. Available at: SSRN 2473475.
- Sundarajan, A. (2014). *Peer-to-peer businesses and the sharing (collaborative) economy: Overview, economic effects and regulatory issues*. Written testimony for the hearing titled *The Power of Connection: Peer-to-Peer Businesses* by the Committee on Small Business of the United States House of Representatives, January 15.
- Sussan, F., & Acs, Z. J. (2017). The digital entrepreneurial ecosystem. *Small Business Economics*, 49, 55–73.
- Sutherland, W., & Jarrahi, M. H. (2018). The sharing economy and digital platforms: A review and research agenda. *International Journal of Information Management*, 43, 328–341.
- Tapscott, D. (1995). *The digital economy promise and peril in the age of networked intelligence*. McGraw-Hill.
- Tapscott, D. (2012). Capitalism 2.0. Thought Leadership Quarterly (TLQ) Essay. Retrieved from <http://dontapscott.com/wp-content/uploads/TLQ-Capitalism2.0-1.pdf>

- The Economist. (2016). *The emporium strikes back*. Retrieved from <https://www.economist.com/business/2016/05/21/the-emporium-strikes-back>
- UNIDO. (2019). *Industrializing in the digital age. Industrial development report 2020*. United Nations Industrial Development Organization.
- Van Alstyne, M. W., Parker, G. G., & Choudary, S. P. (2016). Pipelines, platforms, and the new rules of strategy. *Harvard Business Review*, April (pp. 54–62).
- Varian, H. R. (2010). Computer mediated transactions. *American Economic Review*, 100, 1–10.
- Von Briel, F., Recker, J., & Davidsson, P. (2018). Not all digital venture ideas are created equal: Implications for venture creation processes. *The Journal of Strategic Information Systems*, 27(4), 278–295.
- Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J. F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, 70, 356–365.



Strategy in the Era of Digital Disruption

*Ksenia Podoyunitsyna and
Eglė Vaznytė-Hünermund*

Contents

- 13.1 Introduction – 307**
- 13.2 Disruption Driven by Business Model Innovations – 309**
 - 13.2.1 Freemium Business Models – 313
 - 13.2.2 Sharing Economy Business Models – 313
 - 13.2.3 Usage-Based Business Models – 314
- 13.3 Disruption Driven by Innovation Ecosystems – 315**
 - 13.3.1 Supply-Side Synergies – 318
 - 13.3.2 Demand-Side Synergies – 319
- 13.4 Disruption Driven by Platforms and Network Effects – 320**
- 13.5 Discussion – 323**

- 13.5.1 Trend 1: Industry Crossover Trends
in a Digital World – 324
 - 13.5.2 Trend 2: Changing Competitive
Landscape – 324
 - 13.5.3 Trend 3: Rising Customer
Expectations – 325
- References – 326**

Learning Objectives

After having read this chapter, you will be able to:

- Understand the concept of digital disruption and how it affects different industries and markets.
- Familiarize with the core business aspects that change with the digital transformation of different industries, including (1) business models, (2) innovation ecosystems, and (3) platforms and network effects.
- Discuss how both digital disruption and digital transformation affect the three aforementioned core business aspects.
- Gain insight into recent trends of intra-industry and inter-industry diversification, changing competitive landscape, and rising customer expectations that may considerably shape strategic decision-making.

13.1 Introduction

Digitalization is the increasing application of digital technologies by business and society (Thomas & Autio, 2020).

Digitalization has been rapidly transforming our economy. In fact, the process of digitalization is so intense that it is even referred to as **digital disruption**—a force breaking down industry boundaries, reshaping competitive landscapes, and fundamentally changing the historically sustainable logics for value creation and capture (Crittenden et al., 2019; Skog et al., 2018; Teece, 2018a; Weill & Woerner, 2015). This environmental turbulence caused by emerging digital technologies, new business models, and ever-increasing customer expectations has put a high pressure on traditionally organized companies to react quickly in order to remain competitive (Massa et al., 2017; Sampler, 1998). And even more importantly so, these causes of disruption “are like gravity—they are constant and always at work within and around the firm.”¹ As such, *digital disruption* has become a buzzword of a global phenomenon representing a fast-paced innovation-driven growth, comprising both promises and perils for today’s business and society.

Yet, as we are still in flux of digital transformation, the notion of *digital disruption* is oftentimes misused, which may restrict theoretical advancements and mislead practice (Christensen et al., 2018). For example, *digital disruption* is oftentimes confused with **digital innovation**—a process of combining digital and physical

1 A quote by prof. Clayton M. Christensen in his interview with guest editor Karen Dillon for a special issue of MIT Sloan Management Review, Spring 2020. Source: MIT Sloan Management Review (last time accessed on February 7, 2020): ► <https://sloanreview.mit.edu/article/an-interview-with-clayton-m-christensen/>

assets so as to create radically or incrementally new products, services, or business models (Skog et al., 2018). Although firms in nearly every industry are fiercely looking for digital innovative solutions potentially contributing to their revenue growth and strengthened market position, these innovation efforts are not necessarily disruptive. A concept of *disruption* rather illustrates a process whereby a new entrant can successfully challenge an established firm with a (new-to-the-world) product, technology, or business model by targeting customers “overserved” by incumbents or by creating new markets (Ansari et al., 2016; Christensen et al., 2015, 2018; Markides, 2006). As such, disruption refers to a more fundamental change in a system or an environment, when for example the majority of customers switch to a new offering that is considered to be superior, less expensive, or more accessible than the existing one (e.g., a switch to smartphones).

Digital transformation is yet another term that is often used interchangeably with *digital disruption*. That is, **digital transformation** should be perceived as “a process where digital technologies play a central role in the *creation* as well as the *reinforcement* of disruptions taking place at the society and industry levels” (Vial, 2019: 122). As such, digital transformation refers to a more gradual change happening either at the firm level in terms of an organizational change or strategic renewal (Agarwal & Helfat, 2009; Matt et al., 2015; Weill & Woerner, 2015; Westerman & Bonnet, 2015) or at the market level as a general development of digital technologies and their adoption within society (Nambisan et al., 2019; Vial, 2019).

► Example

The intriguing characteristic of a digital transformation is the fact that it tends to be overlooked or unduly downplayed by the incumbents, eventually becoming a *digital disruption* for the extant market players (Christensen et al., 2018). When, for example, *Apple* announced its introduction of an iPhone in 2007, many mobile phone producers were skeptical about this inter-industry entrance and did not see it as a potential disruptive threat. As one of the phone producers told: “We’ve learned and struggled for a few years here figuring out how to make a decent phone ... PC guys are not going to just figure this out.”² Yet, this view has drastically changed with *Apple* becoming a leading mobile phone producer worldwide.³ This impressive shift of market leadership positions could be mainly associated not only with *Apple’s* unique technology, but also with its timely and successful adoption of a platform-based business approach—i.e., the App Store, along its handset business (Van Alstyne et al., 2016), which altogether greatly contributed to *Apple’s* overall innovation ecosystem spanning several sectors (Adner, 2006). As later insightfully concluded by *Nokia’s* CEO: “our competitors aren’t taking our

-
- 2 The quote is by Palm company’s CEO Ed Colligan in 2006, right after the news that Apple is developing a phone. Source: CB Insights (last time accessed December 19, 2019): ► <https://app.cbinsights.com/research/big-company-ceos-execs-disruption-quotes/>.
 - 3 By the end of 2007, the mobile phone industry was dominated by five key players, viz. *LG*, *Motorola*, *Nokia*, *Samsung*, and *Sony Ericsson*, which altogether accounted for 90% of the global industry profits. However, a successful introduction of Apple’s iPhone shifted this market power distribution, with *Apple* being the leading mobile producer ever since.

market share with devices; they are taking our market share with an entire ecosystem.”⁴ Similarly, *Netflix*, which started off as a DVD-by-mail service and later as an online streaming provider, was not considered seriously, neither by video rental businesses nor by the broadcast television producers and movie theatres.⁵ However, by applying an innovative business model, offering a large on-demand video selection (including its own content in the meanwhile), and providing highly personalized recommendations (Aversa et al., 2017), *Netflix* has considerably shaken the whole entertainment industry with its shares rising over 4000% within a decade.⁶ *Airbnb* is yet another example of how an adoption of unique business model and community-based multi-sided platform has disrupted the hospitality industry (Christensen et al., 2018). For example, *Hilton* hotel’s representative stated that *Airbnb* is not “a major threat to the core value proposition,” or *Expedia*, one of the biggest travel technology companies, also considered that there is no direct effect on their business.⁷ Nevertheless, both hotel and travel companies since then have seen considerable changes in customer purchasing behavior, with *Airbnb* being at the forefront of their choices.⁸ ◀

There are various instances on why and how digital disruption takes place. Thus, in this chapter, we will further elaborate on the core concepts needed to understand the disruptive processes triggered by digitalization and we will suggest a few mechanisms that drive them.

13.2 Disruption Driven by Business Model Innovations

As we have already learned from the previous examples of *Apple*, *Netflix*, and *Airbnb*, digital transformation, and eventually a digital disruption, cannot be merely achieved through a digital or technological innovation itself. In fact, the economic value of any technological advancement remains latent until it is being exploited via an appropriate **business model** (Chesbrough, 2010).

4 The quote said by Nokia’s CEO Steven Elop in 2011. Source: Engadget (last time accessed February 20, 2020): ▶ <https://www.engadget.com/2011/02/08/nokia-ceo-stephen-elop-rallies-troops-in-brutally-honest-burnin/?guccounter=1>.

5 Source: Forbes (last time accessed February 4, 2020): ▶ <https://www.forbes.com/sites/aal-sin/2018/07/19/the-future-of-media-disruptions-revolutions-and-the-quest-for-distribution/#247aa68c60b9>.

6 Source: Bloomberg (last accessed February 4, 2020): ▶ <https://www.bloomberg.com/news/articles/2019-12-30/netflix-s-10-year-4-000-rally-underlines-shift-to-streaming>.

7 The first quote is by Hilton’s CEO Christopher Nassetta in 2015, and the second insight is by Expedia’s CEO Dara Khosrowshahi, also in 2015. Source: CB Insights (last time accessed December 19, 2019): ▶ <https://app.cbinsights.com/research/big-company-ceos-execs-disruption-quotes/>.

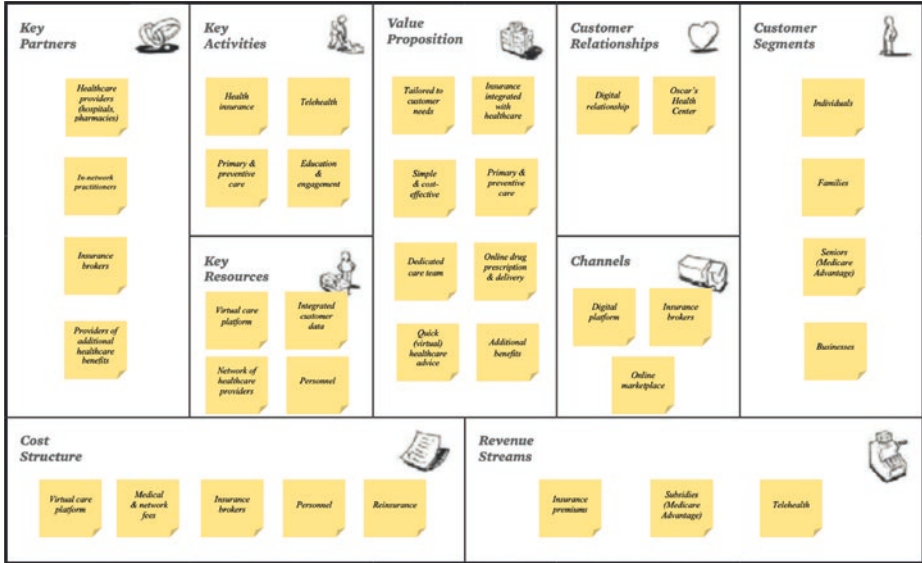
8 Source: Vox (last time accessed on February 6, 2020): ▶ <https://www.vox.com/2019/3/25/18276296/airbnb-hotels-hilton-marriott-us-spending> and Bloomberg (last time accessed on February 6, 2020): ▶ <https://www.bloomberg.com/news/articles/2019-05-02/expedia-shares-slump-as-short-term-rental-growth-slows>.

Business model represents an organizational and financial architecture of a business, which helps to explain the mechanisms through which a focal firm creates, delivers, and captures value (Teece, 2010).

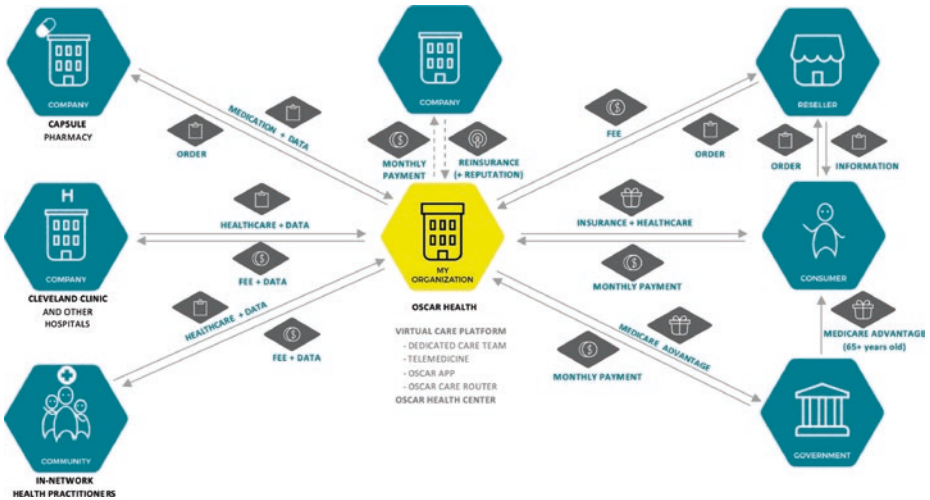
It encompasses both a cognitive and an action aspect in it, as it can serve as a cognitive scheme helping to think of a firm's essence as well as a constellation of actions that a firm executes in reality (Berends et al., 2016). If sufficiently differentiated (e.g., honed to meet particular customer needs) and difficult to replicate for other companies (e.g., contains many tightly interlinked activities, technologies, and organizations), business models contribute not only to firms' performance (Andries & Debackere, 2007; Chesbrough & Rosenbloom, 2002; Cucculelli & Bettinelli, 2015; Pauwels & Weiss, 2008; Zott & Amit, 2007), but also to their competitive advantage (Casadesus-Masanell & Zhu, 2013; Teece, 2010). As such, substantial research efforts have been devoted to a better understanding of how firms design, manage, and transform their business models (Massa et al., 2017; Morris et al., 2005; Priem et al., 2018).

There are two complementary perspectives on business models in both the scientific and the managerial world. First, the *component-based perspective* distinguishes a set of components to depict and explain what the business model is. The business model canvas developed by Osterwalder and Pigneur (2010) is by far the most famous set of business model components and is the de facto standard in the management and entrepreneurial world. Similar component-based frameworks were developed by Morris et al. (2005) and Chesbrough and Rosenbloom (2002). Scientific studies building upon the component-based perspective have been quite numerous (Andries et al., 2013; Berends et al., 2016; Bocken et al., 2015; Demil & Lecocq, 2010). Second, the *boundary-spanning perspective* focuses instead on a system of boundary-spanning transactions, activities, or (monetary and nonmonetary) value transfers between a set of actors necessary for firms to create, deliver, and capture value (Amit & Zott, 2001; Arend, 2013; Brehmer et al., 2018; Zott & Amit, 2007, 2008, 2010). While the component-based perspective largely focuses on what is happening "inside the box," the box being the firm or organization, the boundary-spanning perspective focuses on what is happening between the focal organization and other actors and stakeholders, essentially taking the "outside-the-box" view. Nevertheless, both perspectives and their conceptualizations demonstrate that business models are instrumental for grasping and reflecting how firms create and capture value (Priem et al., 2018). For illustration, please refer to a business model of *Oscar Health*, a health insurance company, depicted by both the "inside" and "outside" perspectives (see ■ Figs. 13.1 and ■ 13.2, respectively).

As digitalization continues to vigorously transform business processes, communication channels, and customer activities, companies are forced to re-evaluate their value propositions in order to be able to meet increasing customer expectations and be profitable in doing so (Teece, 2010, 2018a; Zott et al., 2011). Advancing through **business model innovation** has thus become a new way of experimentation allowing to achieve these goals.



■ Fig. 13.1 Illustration of a component-based business model, *Oscar Health*. (Source: Authors' own figure using Business Model Canvas (Osterwalder & Pigneur, 2010))



■ Fig. 13.2 Illustration of a boundary-spanning business model, *Oscar Health*. (Source: Authors' own figure using Business Model Connect tool (Brehmer et al., 2018))

Business model innovation represents a novel approach for commercializing underlying firm assets (Gambardella & McGahan, 2010).

More precisely, business models not only can act as a *vehicle for innovation* allowing to unlock the value potential from a firm's technology (Chesbrough, 2010), but they can also be a *subject of innovation* complementing the traditional products, processes, and organizational innovation (Foss & Saebi, 2017; Gambardella & McGahan, 2010; Massa et al., 2017; Zott et al., 2011). That is, firms can innovate by adding new activities (i.e., novel content), collaborating with new partners responsible for specific type of activities (i.e., novel governance), or by coupling these activities in novel ways (i.e., novel structure) (Snihur & Zott, 2020). Business model innovations can help to harness the digital disruption for a firm's advantage, by allowing to offer more (technologically) sophisticated, efficient, integrated, and personalized value propositions (Brock et al., 2019; Gambardella & McGahan, 2010; Teece, 2018b). This altogether also helps to reinforce the central role of a customer, and thus value creation potential (Priem et al., 2018).

➤ Innovating through business models becomes particularly important for companies active in industries exposed to a “perfect storm of two forces”—i.e., low entry barriers and great reliance on legacy business models that can be easily digitized (Grossman, 2016). According to a survey of the top executives from around the world, a great or moderate threat of digital disruption should therefore be most pronounced in media, telecommunications, technology, and consumer product sectors, and to an increasing extent in financial services, healthcare, and industrial sectors (Russel Reynolds Associates, 2015, 2017). Here, business model innovations may be pivotal.

While every entrepreneur aims at designing an innovative business model turning customer needs into a self-sustaining profit-making engine, these transformative business efforts might nevertheless be quite challenging due to a high complexity and uncertainty of its effectiveness (Berends et al., 2016). This is particularly true for large established firms, which despite their resource abundance are prone to bureaucratic inertia or unwillingness to cannibalize sales of their existing products (Christensen & Rosenbloom, 1995). Hence, incumbents' inability to properly innovate through business models, accompanied with customers' frustrations associated with old operating models (e.g., lack of price transparency, control, and convenience), provides a window of opportunity for new potentially disrupting entrants (Crittenden et al., 2019), who eventually can enter the markets by stealing a select few activities that customers are not satisfied with (Teixeira, 2019). As a result, new business opportunities brought about by new technologies, increasing competitive threats, and changing demands of stakeholders have prompted a development of novel business models helping firms to be more (digitally) agile and better address increasing (digital) customer expectations (Foss & Saebi, 2017). Among a large variety of new ways of creating and capturing value, we provide a closer look into several recent business model innovations—namely, *freemium*, *sharing economy*, and *usage based*—that are gaining momentum in today's digital environment.

13.2.1 Freemium Business Models

Well-developed business models are crucial for internet-based companies such as *LinkedIn*, *Dropbox*, or *Skype*, wherein customers expect to use their services free of charge. One core characteristic of a digital good is the fact that it has (near) zero marginal costs of production and distribution (Lambrecht et al., 2014). This gave rise to the abundance of **freemium-based** business models, in which a certain part of the offering is provided for free, and a more extended version of the same offering against a certain price, i.e., the premium (Tidhar & Eisenhardt, 2020; Van Angeren et al., 2022). Freemium comes in many forms, mainly differentiating between what is exactly being offered for free. Free could be provided (and limited) based on features, time, or customer segment (Anderson, 2009). Another distinction that gained greater popularity due to digitalization is the distinction between durable and consumable features. Durable features include levels and functionalities that do not expire, deteriorate, or decrease in quantity (e.g., new levels in games or drawing apps), while consumable features include coins, gems, and credits that decrease with consumption or use by customer (Van Angeren et al., 2017). A final noteworthy distinction is between bundled and fragmented freemium models (Tidhar & Eisenhardt, 2020). In the former case, a firm can add substantially more value to the customer by offering one or several highly interrelated and reinforcing features, which otherwise are difficult to be consumed separately (e.g., *Spotify's* premium features entailing an on-demand streaming and offline mode). In the latter case, a firm can create more value by offering multiple and diverse “fragmented” premium features, which can be purchased and consumed independently (e.g., *Udemy's* users can freely browse and read lecture reviews, but for a premium fee, they can choose from a wide variety of online courses, their respective modules, and other fragmented features). One of the greatest challenges of freemium models, nevertheless, is that the free and paid content has to be neatly balanced (Kumar, 2014; Pauwels & Weiss, 2008; Rietveld, 2018; Wagner et al., 2014). Too many features or content on the free side will boost the user base; however, this user base will not be likely to convert to the paid content. In similar vein, too much emphasis on the paid side will make the basic value proposition for the free side not interesting enough pressing the customer base down. All in all, with the increasing digitalization and competition, freemium business models can offer an effective design for customer acquisition and respective monetization of digital products or services.

13.2.2 Sharing Economy Business Models

Another tendency is that business models are shifting from resource ownership to resource sharing. Whereas traditional corporate theory posits a firm's tangible (e.g., high-tech equipment) and intangible resources (e.g., intellectual property rights) that are valuable, rare, imperfectly imitable, and non-substitutable as a prerequisite for a temporary or sustained competitive advantage (Barney, 1991), developments in a **sharing economy** prove this notion not necessarily true (Teece,

2018b). Companies like *Airbnb*, *Uber*, *Neighbor*, *Vinted*, *LendingClub*, *CrowdMed*, and *Upwork*, among others, do not possess any of the strategic resources, yet are among the most valuable (or trending) companies to date. By relying on sharing economy business models, they can instead provide a *temporal access to* and facilitate the *exchange of* the underutilized assets held by the individuals and firms alike, such as their apartments, cars, storage space, clothes, money, knowledge, or time (Amit & Han, 2017; Foss & Saebi, 2017). Depending on different types of transactions, sharing economy business models are sometimes called collaborative consumption, peer-to-peer networks, crowdsourcing, or crowdfunding activities (Laukkanen & Tura, 2020).⁹ These business models extensively draw on new information and telecommunication technologies, which help to enable customers' accessibility, flexibility, trust, and ease of asset sharing (Kathan et al., 2016). One of the key features of sharing economy business models is that firms engage customers or users already in the value creation phase, which is traditionally internalized within a given traditional company (Amit & Han, 2017; Kohler, 2015; Prahalad & Ramaswamy, 2004). To compensate for this, customers or users get a proportion of the value captured as well, which could be in terms of reduced price, increased convenience, financial return, or reputation (Kohler, 2015). Overall, sharing economy business models help to identify unique ways of value creation by linking heterogenous resources with heterogenous needs in a digitally enabled economy (Amit & Han, 2017).

13.2.3 Usage-Based Business Models

A final trend worth noting revolves around business model innovations that are more customer centric rather than product centric (Priem et al., 2018). **Usage-based** business models offer value propositions that are customer tailored, cost transparent, and highly flexible—features that are of paramount importance for today's digital consumers. Although the concept of these business models is not new, consider for example *Rolls-Royce* and its power-by-the-hour model in the 1960s (Teece, 2018b); it does nevertheless receive a whole new meaning in light of the digital economy. By leveraging big data (e.g., data from the Internet of Things, telematics, wearables, and other sensor- or GPS-enabled devices) and new technologies (e.g., big data analytics, artificial intelligence, and machine learning),

13

⁹ Crowdfunding activities are more common in the financial sector, where a focal firm (e.g., a startup) amasses a small amount of underutilized assets (e.g., money) from a large group of resource providers (e.g., individuals), such as in *LendingClub*, *Seedrs*, or *Kickstarter* (Amit & Han, 2017). The same principle has now been applied in other sectors, such as insurance, where a group of people agree to cover similar risks by creating a pool of financial resources comprised of their premium shares. Essentially, this model allows to insure customers' belongings or even health without the help from traditional insurance companies, such as *Friendsurance*, *Lemonade*, or *BoughtByMany*.

companies can align their strategies in meeting customer needs more astutely and, for example, offer products or services tailored to a customer's behavior. For instance, usage-based business models are getting popularity among car insurers, who can price insurance policies based on how a customer drives (e.g., *Root* insurance collects data on acceleration patterns, breaking, and time of the day), distance travelled (e.g., *Metromile* insurance offers fee-per-mile), or only when a car is used (e.g., *By Miles* insurance charges no fee if the car is parked). However indirectly, usage-based approach also gets traction among life and health insurers, who can offer policies at a discount rate for customers having a healthy lifestyle. *Health IQ*, for example, offers life insurance policies based on customers' regular sport activities such as running at least ten miles per week, cycling fifty miles per week, or providing a proof by a coach of indoor-class activities. Oftentimes, these usage-based business models are accompanied with an *on-demand feature*—i.e., a product or service that can be accessed and used only when a customer needs it. As such, business models that align pricing with usage or customer behavior are getting traction among a wide range of (physical and digital) products and services (Deloitte, 2016). When combined with on-demand business models, they can not only help to reduce up-front barriers (e.g., markets with significant asset requirements, high costs, or unmet demand), but also offer more customer-centric value proposition with greater personalization, flexibility, and cost transparency.

To conclude, in order to make digitization and big data to play at their advantage, firms should focus not only on how to create value in novel ways, but also on how to best capture value and monetize it. As such, business model innovations are vital for new and established companies willing to adapt and thrive in these changing environments (Christensen et al., 2015; Teece, 2018b).

13.3 Disruption Driven by Innovation Ecosystems

From the innovation literature, we already know that a successful commercialization of innovation requires not only a possession of strategic assets, but also their combination with other technologies and competencies in unique, value-enhancing ways that oftentimes exceed a mere additive nature of standard complements (Jacobides et al., 2018; Teece, 1986, 2018a). This observation becomes especially important in the age of rapidly evolving digital economy, where companies need to react quickly in acquiring and coordinating diverse and novel capabilities if they are to remain on a competitive edge and capture additional value (Fuller et al., 2019). As a result, firms tend to collaborate with more and more partners within and across industries as part of their business model, with **innovation ecosystems** being one of the increasingly prominent ways to engage for a joint value creation (Adner & Feiler, 2019; Jacobides, 2019; Talmar et al., 2018; Teece, 2018a; Williamson & De Meyer, 2012).

Innovation ecosystems are defined as “the alignment structure of the multilateral set of partners that need to interact in order for a focal value proposition to materialize” (Adner, 2017, p. 40). They represent an array of complementary elements (e.g., technologies, services, standards, and regulations) that must be in place for a value proposition to be delivered (Adner & Kapoor, 2016).¹⁰

- The business model and innovation ecosystems are two related concepts, especially in the boundary-spanning view on business models: both are linked to systems and focus on a network of actors (Adner & Kapoor, 2010; Zott & Amit, 2010). As opposed to business models however, innovation ecosystems span a greater set of (inter)dependent actors that are not necessarily related or dependent upon one central actor and go well beyond any individual focal firm (Adner, 2017; Fuller et al., 2019).

Innovation ecosystems provide firms with new ways of handling the trade-off between *flexibility* and *commitment* (Fuller et al., 2019). Specifically, they enable different firms to engage into multiple complex relationships characterized by simultaneous cooperation and competition for developing an interdependent product or service while at the same time preserving their autonomy and decision power (Adner & Kapoor, 2010; Jacobides et al., 2018; Williamson & De Meyer, 2012). Interdependencies among ecosystem actors are created by flows of information, materials, services, mutual influence, and/or funds (Adner, 2017) and can be viewed from technological, economic, or cognitive perspectives (Thomas & Autio, 2020). For example, companies can engage in technology-oriented collaborations via platforms or any other type of technological architecture in order to allow customers to assemble separate components for their final product (e.g., *Android OS*, *Apple iOS*). Interdependencies can also be realized through a simultaneous exchange of separate actors’ offerings that contribute to their economies of scope, economies of scale, knowledge, or risk sharing (Autio & Thomas, 2014). Finally, innovation ecosystems can be viewed from a cognitive perspective. As innovation ecosystems may attract many diverse participants, their coherent view about the core principles of the ecosystem—i.e., a collective identity, is pivotal for a smooth functioning of an ecosystem and ultimately for value creation and capture.

Another important characteristic distinguishing innovation ecosystems from other traditional organizational arrangements, such as supply chains, clusters, or value networks, is their strong reliance on noncontractual governance (Thomas &

¹⁰ Although mutually consistent, *innovation ecosystems* are distinct from *entrepreneurial ecosystems* in several important respects (see Adner, 2017; Thomas & Autio, 2020). Most prominently, innovation ecosystems pertain to an activity-centric view of interdependence, where the alignment of partners is critical for creating and delivering a certain value proposition. Entrepreneurial ecosystems, on the contrary, are more actor centric and are primarily concerned with interactions at macro level. The strategy of entrepreneurial ecosystems is focused on general governance and community enhancements, with only limited insights into a certain value creation (Adner, 2017).

Autio, 2020). Specifically, actors can specialize in certain roles within the ecosystem based on the *co-alignment structure*—a mutual agreement instead of vertical integration based on formal contracts or static configurations specifying the exact input of each partner (Adner, 2017; de Vasconcelos Gomes et al., 2018; Jacobides et al., 2018). The respective type of structure allows for an effective coordination of dynamic, coevolving, and semipermanent types of relationships between different actors. This in turn also helps firms to quickly adapt to changing customer needs, emerging technologies, unexpected market shifts, or regulatory changes compared to the vertically integrated ones.

- It is worth noting that the extent to which each party is attached to a certain ecosystem is defined by their investments or resources that cannot be easily redeployed in any other setting without a cost, such as cost of product configuration adjustment or coordination with other members' activities (Jacobides et al., 2018).

Yet, whether an ecosystem is worthwhile to be pursued primarily depends on the nature and degree of the *complementarity* of assets and capabilities between actors within an ecosystem (Jacobides et al., 2018). That is, if the underlying complementarity between separate actors is nongeneric and entails some degree of customization, then actors have interest to align and act as a group.¹¹ Additionally, for this complementarity to be effective, it also has to rest upon a *related diversification* strategy (Ahuja & Novelli, 2017; Markides & Williamson, 1994; Palich et al., 2000; Robins & Wiersema, 2003). That is, certain strategic assets and capabilities (those that cannot be easily copied by nondiversified competitors) should be highly relevant for both parties in order for the long-run value to accrue (Markides & Williamson, 1994).¹² In fact, the greater the level of strategic relatedness between different parties, the greater the potential gains from this diversification strategy.

-
- 11 Complementary assets can be broadly divided into generic and specific (Teece, 1986). Generic complementary assets, on the one hand, are the ones that do not need to be tailored to the innovation in question (e.g., electricity is needed for an innovation, but it can be purchased broadly and under generic terms). Specific complementary assets, on the other hand, have a certain *unique* unilateral (e.g., A cannot function without B) or bilateral dependence (e.g., A and B require each other) with the focal innovation, and therefore need to be coordinated for a unique value to be derived. Next to *unique* specific assets, there is another type of *supermodular* complementarity, which results in a greater value when two products are consumed together rather than in isolation (e.g., more of A increases value of B) (Teece, 2018a). As such, firms that want to engage in specific *unique* or *supermodular* complementarity would have the greatest interest to create the alignment structure (Jacobides et al., 2018).
- 12 Information (whether or not retrieved from data) is considered to be a strategic asset. Although if viewed from a traditional market-based perspective, information may be discarded due to its high “fungibility,” i.e., it is interchangeable and can be widely applicable in many situations. More precisely, information (or data) is a strategic asset if it enables a firm to implement strategies that improve its efficiency or effectiveness, and is idiosyncratic or difficult to imitate (Barney, 1991). In this digital age, information indeed has become one of the most critical assets for diversification, and thus pursuit of economic rents (Arend, 2013; Hartmann & Henkel, 2020; Hartmann et al., 2016; Sampler, 1998). In fact, firms possessing sufficient amount of critical data or a collection of data for the same market (e.g., customers) define the industry boundary (Sampler, 1998).

■ **Table 13.1** Supply- and demand-side synergies (compiled by the authors)

Supply-side synergies	Demand-side synergies
<ul style="list-style-type: none"> – Economies of scope and scale – Reduced transaction costs and risks – Increased (administrative) efficiency – Increased customer reach and potential demand – Access to more comprehensive data – Access to partners' intellectual property, technology, and talent pool – More regular engagement with customers via multiple channels – Additional streams of revenue – Possibility to scale up faster 	<ul style="list-style-type: none"> – Single access point for managing data – Seamless and efficient end-to-end journey providing real-time information – Personalized services, more accurate predictions – Greater access and choice of products and services – Cost-effective and transparent products and services

These complementarities contributing to *value creation* in terms of increasing customer benefits and *value capture* in terms of firm profitability (Priem, 2007) are mainly realized through the supply-side and/or demand-side synergies (Jacobides et al., 2018), which we both describe in more detail in the two subsections below and illustrate in ■ Table 13.1, respectively.

13.3.1 Supply-Side Synergies

A related diversification strategy primarily contributes to the **supply-side** or **producer-specific synergies** stemming from a strategic resource bundling (Markides & Williamson, 1994; Palich et al., 2000; Sirmon et al., 2008). By combining complementary assets and capabilities, firms can increase their value creation and appropriation potential (Adner & Kapoor, 2010; Jacobides et al., 2006; Kapoor, 2014), and thus gain a sustainable competitive advantage over other market participants (Aversa et al., 2017; Dyer & Singh, 1998; Teece, 2018a). For instance, by collaborating with complementors within and across innovation ecosystems, firms can benefit from economies of scope (Palich et al., 2000; Tanriverdi & Lee, 2008) and economies of scale (Van Alstyne et al., 2016); increase capacity utilization and (administrative) efficiency (Iansiti & Levien, 2004); access additional knowledge, technology, and talent pool (Williamson & De Meyer, 2012); improve performance of their products (Kapoor, 2014); and reduce risks and overall costs (Aversa et al., 2017; Jacobides et al., 2018), just to name a few.¹³ Additionally, the supply-side

13 According to Markides and Williamson (1994), synergies stemming from economies of scope are important only for offering a short-term advantage in terms of improved differentiation and reduced costs. Other types of production synergies that are directed towards existing asset improvement, new asset creation, or learning new competencies hold a greater promise of becoming a long-term competitive advantage.

synergies further contribute to the access of more comprehensive and reliable customer-related data (Aversa et al., 2017), and respectively to a greater customer reach in existing or new market segments (Kapoor, 2014). This in turn equips firms with a possibility to scale up faster and allows to generate additional streams of revenue (Tanriverdi & Lee, 2008). As such, participation in innovation ecosystems is particularly valuable for players from highly fragmented and knowledge- or capital-intensive industries (Aversa et al., 2017; Mitchell & Singh, 1996; Williamson & De Meyer, 2012).¹⁴

13.3.2 Demand-Side Synergies

Although the *supply-side perspective* of a firm's strategy has long been considered as one of the most influential perspectives, the *demand-side perspective* that advances the strategic relevance of consumer preferences is gaining traction too (Aversa et al., 2020; Priem, 2007; Priem et al., 2018). As digital transformation is changing the traditional customer-supplier balance, with customers demanding more complex and customized offers (Teece, 2010), the competition is no longer a zero-sum game within any particular market or industry. Specifically, the classic view of a competition within one industry, where “a discrete set of broadly similar players compete to produce a common end product in a vertically integrated fashion” (Fuller et al., 2019), is challenged. Firms do no longer compete solely for addressing customer needs within one particular industry, but are instead focused on how they can meet as many customer needs as possible irrespective of industry boundaries (de Vasconcelos Gomes et al., 2018; Thomas & Autio, 2020). As such, the **demand-side** or **customer-specific synergies**, representing an increased value for customers when certain products or services are consumed jointly rather than in isolation (Jacobides et al., 2018), become crucial for any customer-centric business strategy (Teixeira, 2019). For example, these demand-side synergies can be realized through facilitation of consumers' accomplishment of several tasks simultaneously, e.g., a one-stop shop concept (Ye et al., 2012), which increases convenience and reduces learning- and search-related costs (Klemperer & Padilla, 1997; Tanriverdi & Lee, 2008). Demand-side synergies can also be derived from a more personalized product or service offering, customers' involvement in a co-creation process, as well as a greater cost transparency, overview, and control over the purchase (Hienerth et al., 2014; Weill & Woerner, 2015; Williamson & De Meyer, 2012). Among various instances of the demand-side synergies, synergies stemming from the innovation ecosystems arranged around a (digital) platform, i.e., (*cross-side network effects*) (Ye et al., 2012), represent yet another powerful way of increasing customer value. Due to its prominence, we cover this topic more comprehensively in the following section.

14 When considering a digital environment, (asymmetric) information oftentimes becomes a crucial strategic asset, contributing not only to a firm's competitive advantage, but also to its value capture potential (Williamson & De Meyer, 2012). This is especially true in knowledge-intensive industries, such as finance, insurance, telecommunications, healthcare, and education.

- Demand-side synergies stemming from consumers' bundling preferences can help to create and capture value irrespective of whether there are any supply-side synergies (Ye et al., 2012). For instance, customer-specific synergies can contribute to companies not only in terms of superior knowledge about customer needs and reduced customer acquisition costs, but also in terms of increased customer demand and willingness to pay for their complex offerings (Schmidt et al., 2016). Yet, in this case, companies should have readily available mechanisms helping them to appropriate and monetize the value from these customer-specific synergies (Priem et al., 2018).

To conclude, innovation ecosystems are imperative for the age of digital transformation. By leveraging innovation ecosystems, companies can engage in multiple complex relationships and still maintain their corporate focus. Synergies stemming from resource complementarity further enhance value creation and value capture potential, which is in line with increasing customer expectations and changing competitive landscapes.

13.4 Disruption Driven by Platforms and Network Effects

A prominent mechanism behind the scenes of a digital disruption is the pervasiveness of **multi-sided platforms**. Although the concept itself is not new (consider for example brick-and-mortar stores, travel, or real estate agencies), digital multi-sided platforms are taking over the traditional type of businesses by storm (Eisenmann et al., 2011; Jacobides et al., 2019). To illustrate this, seven of the world's ten largest companies have a digital platform at the core of their business activities compared with only two a decade ago.¹⁵ In fact, many (digital) innovation ecosystems are arranged around one platform, or there are even multiple platforms in a given ecosystem (Jacobides et al., 2018). Thus, an interesting question then is what makes these multi-sided platforms so special.

Multi-sided platform is a (technological) interface that brings together different groups of actors such as producers and consumers in high-value exchanges (Jacobides, 2019; Van Alstyne et al., 2016).

The most straightforward interdependencies in innovation ecosystems are driven by functional or activity-based effects, i.e., hardware is needed for the software of a machine or gadget to function, and vice versa (Adner, 2017). Think about the iPhone and iOS as an example for them. The alternative mechanism driving such interdependencies is less visible and is triggered by **network effects**. A **direct** or **same-side**

¹⁵ By the end of 2019, the list of the top ten largest global companies included seven platform-based businesses, namely *Apple*, *Amazon*, *Alibaba*, *Alphabet (Google)*, *Facebook*, *Microsoft*, and *Tencent*, whereas by the end of 2009, only *Apple* and *Microsoft* were in the top ten of this list. Source (last accessed on May 8, 2020): ► https://en.wikipedia.org/wiki/List_of_public_corporations_by_market_capitalization.

network effect exists when the value of a given product, service, or solution is dependent upon its number of users, such as in case of fax machine, messenger apps, social networks, or traffic detection as part of the car navigation software and apps. As such, the more users join the platform (e.g., *Facebook*, *Microsoft*, *Waze*), the more valuable it becomes, making a switch to any alternative platform costly and difficult (Gawer & Cusumano, 2014). An **indirect** or **cross-side network effect** refers to dependencies with the number of users of a different yet related product, service, or solution, such as Blu-ray players that need films in Blu-ray format or operating systems that require apps in order to be valuable for the customers (Podoyntsyna et al., 2013; Schilling, 2002). Similarly, these indirect effects become stronger the greater the cross-side participation, e.g., the greater the choice of apps, the more attractive the platform for users, and vice versa. This respectively also enables platform owners to implement different business models for different platform sides allowing to create and capture most of the value (Parker & Van Alstyne, 2005). While direct and indirect network effects are conceptually related, they may have very distinct implications for firms' strategies and performance (Podoyntsyna et al., 2013). With the rise of digitalization, such network effects become even more pronounced and widely spread due to the increased connectivity between different digital products and services, and the dynamics driven by the expansion of digital platforms. Several examples of direct and indirect network effects are presented in ■ Table 13.2.

■ **Table 13.2** Examples of solutions with direct and indirect network effects (compiled by the authors)

Direct (same-side) network effects	Indirect (cross-side) network effects
<ul style="list-style-type: none"> - Dating websites/apps - Peer games (e.g., Counter-Strike, Call of Duty, Battlefield) - Telephone, fax, WhatsApp, Skype, Myspace, Instagram, Twitter, Facebook, LinkedIn, Stack Overflow, etc. - App developers on a platform (first positive due to info exchange and improvements in quality, then potentially negative due to increased competition) - Telecom services (calling inside the network of KPN or T-Mobile is free) - Exchanges in (crypto)currency - MS Office and other software packages using certain standard: I can make a document and other people can open it - Emails - Walkie-Talkie (a special case since it is capped at two devices) - Airbnb (reviews of tourists and owners) - Netflix (personalization algorithm gets better) - TomTom with traffic updates and Waze - Medications (more side effects are known and effectiveness is proven) 	<ul style="list-style-type: none"> - Games and game consoles - Android/iPhone/iPod/iPad and their accessories, iPhone/iPod/iPad users and apps - iTunes devices' (e.g., iPod) users and music tracks/music publishers that can be bought, same for Spotify and other music streaming services - Word and WordPlus - E-readers (e.g., Kindle) and e-books - Marktplaats, eBay, ► Aliexpress.com, Amazon - Television (ad providers and people watching TV) - Search engines (users and ad providers) - Uber, SnappCar, and other P2P car sharing services - Operating system and software that runs under it - Airbnb (hotel owners and tourists) - Netflix (users and film providers) - Oscar Health (insured people and health providers) - Smart watches and apps for them

In what follows, the main strategic goal of each multi-sided platform is to grow the relevant sides of the market so as to harvest most of the benefits from these (*cross-side*) *network effects* (Adner, 2017; Eisenmann et al., 2011).

► Example

Amazon, for example, started off as a two-sided online marketplace for books and consecutively expanded its business to e-commerce “of everything,” to entertainment (video and music online streaming), to web services (cloud computing platforms and APIs), and to financial services and several more. Cross-side network effects enabled *Amazon* to leverage its huge demand-side economies of scale spanning boundaries of several seemingly unrelated industries (Aversa et al., 2020; Van Alstyne et al., 2016).¹⁶ This respectively was instrumental for further expanding and scaling up the platform and its related innovation ecosystem. Yet, it is important to note that marginal utility derived from these intergroup externalities may not always be the same for different customer groups (Ye et al., 2012). While participation of more sellers on *Amazon* makes its platform more attractive to buyers (e.g., greater choice and convenience), and respectively more buyers on *Amazon* make a listing on platform more attractive to suppliers (e.g., greater exposure and customer reach), this may not be the case when considering a platform such as *Metro* newspaper. On the *Metro* platform, more advertisers benefit from a greater number of readers, but not the other way round. Therefore, platforms depending on the type of value exchange between separate parties can be regarded as *symmetric*—i.e., all customer groups (sides) derive value from the other side(s) of the platform, or *asymmetric*—i.e., only one customer group (side) derives value from the other side(s) of the platform (Brehmer et al., 2018). In this latter case, asymmetric platform owners must therefore “cross-subsidize” underserved customer groups if they are to maximize the overall number of users (Ye et al., 2012). ◀

An important quality of multi-sided platforms is that they help to facilitate and increase the degree of innovation on complementary products or services. That is, the greater the complementarity, the more value is created via network effects, and thus the greater the barrier to entry for any other new entrants or rivals (Gawer & Cusumano, 2014). When these network effects are strong, switching costs are high, and there is no benefit from niche specialization—the platform is considered to have strong **isolating mechanisms** (Sun & Tse, 2009). In this regard, a number of platforms in certain markets tend to be small, with the competition between platforms resulting in so-called winner-takes-all outcomes (Teece, 2018a). A competitor in this case is able to enter the market only if it can either offer substantial platform improvements followed by large investments covering user switching costs or successfully capture a market share through envelopment—i.e., by bundling its own platform’s functionality with an existing platform’s so as to leverage (and

16 When considering a platform, both its suppliers (e.g., app producers) and customers (e.g., people installing apps on their mobile devices) represent the *demand side* for the platform owner (e.g., *Apple iOS*). Therefore, (cross-side) network effects are considered as a special case of demand-side synergies.

eventually foreclose) shared user relationships (Eisenmann et al., 2011). As platform-to-platform competition can be fierce, a platform's governance decisions regarding its *access* (a degree of openness) and *attachment* (a degree of co-specialization) become pivotal for its success (Jacobides, 2019; Van Alstyne et al., 2016). That is, a platform's governance determines its attractiveness for its complementors as well as the extent to which it can directly capture value from the cross-side network effects.

► Example

While open platforms such as *Android* may attract complementors more easily, and thus offer a greater selection of apps, their quality may nevertheless vary; also, owners of open platforms may have difficulty in capturing value directly. Managed platforms such as *Apple App Store* or closed platforms such as *Philips' digital health* may offer greater quality apps, yet they may be more expensive or relatively few. The degree of attachment, or specific platform-related investments, is yet another important trade-off to be made by platform owners. While, for example, a platform with a strong market position can require its complementors to adjust to its ecosystem standards, it nevertheless runs the risk that not many complementors would be willing to co-specialize and would instead exploit opportunities elsewhere. ◀

Like almost everything in life, network effects come with pros and cons. The advantage from the focal firm's perspective is mostly related to the powerful, cost-efficient added value generated for the customers purely based on the customer base, which is on top of the value generated by the regular features of the product or service. This phenomenon is termed the "bandwagon effect" or the "critical mass point" (Goldenberg et al., 2010; Podoyntsyna et al., 2013; Tellis, 2010). The same phenomenon serves as the core disadvantage since it may take quite a bit of time and effort for firms to be able to reach it, having a "chilling effect" on customer acceptance (Goldenberg et al., 2010). Further disadvantages for the focal firm include the fact that network effects have a negative effect on the survival duration of pioneering products (Srinivasan et al., 2004), and decrease net product value as well as increase payback periods (Goldenberg et al., 2010).

13.5 Discussion

The main objective of this chapter was to provide an overview of strategies that can be used to fuel digital disruption and shed more light on the key mechanisms underlying digital dynamics. We distinguish business model innovation, innovation ecosystems, as well as platforms and the related network effects as the core concepts that are relevant for understanding strategy making in the era of digital disruption. We would like to conclude this chapter with several state-of-the-art insights into key trends that should be taken into consideration when implementing any of the respective strategies.

13.5.1 Trend 1: Industry Crossover Trends in a Digital World

In the age of digital disruption, companies are expanding their activities within and across industries. While the *intra-industry* diversification is more organic as it allows to expand the scope of offerings, collect more insights into customer preferences, and leverage current technologies for a new market niche within the same industry (Tanriverdi & Lee, 2008), the *inter-industry* diversification is nevertheless associated with greater synergies stemming from both supply and demand sides. The inter-industry diversification strategy aims at increasing the collaborations across traditional industry boundaries so as to meet as many customer needs as possible (de Vasconcelos Gomes et al., 2018; Thomas & Autio, 2020). For example, big platform-based businesses may enter the “unrelated” market by leveraging their superior network effects (e.g., *Amazon’s* entry into healthcare, *Google’s* entry into home-automation market), targeting an overlapping customer base with a new product offering (e.g., *Airbnb* competing with traditional hotel chains or *Uber* with traditional taxi services), or collecting the same type of data as existing businesses (e.g., *Apple* or *Fitbit* gathering health-related data) (Alstynne et al., 2016). The shift to an ecosystem thinking challenges the very idea of *industry*, where a discrete set of broadly similar players are competing to produce a common end product in a vertically integrated fashion (Fuller et al., 2019). Thus, the boundaries between traditional industries are fading, making the business model innovations, innovation ecosystems, and platforms imperative for a successful business strategy (Tece, 2018a).

13.5.2 Trend 2: Changing Competitive Landscape

In what follows, blurring industry boundaries have also prompted new paths to dominance. That is, the competitive landscape is changing and is increasingly becoming a competition of *ecosystem* against *ecosystem*, or *platform* against *platform*, which is counter to the winner-takes-it-all paradigm that was common thinking in the past (Schilling, 2002). This potentially has implications for what the optimal competitive strategy is. In particular, in the classical competitive landscape, where competition focuses on products and services, prior research established that the relationship between distinctiveness and firm performance follows an inverted U-shape (Zhao et al., 2017, 2018). In other words, there is an optimal distinctiveness point and the firm’s competitive strategy “should be as different as legitimately possible” in order to achieve superior performance (Deephouse, 1999: 197). At the same time, to this date, there has been only one study exploring similar effects on the platform level, and it found exactly the opposite. In particular, it states that the relationship between distinctiveness of platforms and performance is a simple U-shape so that the best performance is being achieved through either no differentiation with existing platforms or being very distinct from them (Cennamo & Santalo, 2013). To the extent that the whole competition landscape shifts more towards platforms and ecosystems, the optimal competitive strategy will also likely change. We firmly believe that more research on this topic is warranted.

13.5.3 Trend 3: Rising Customer Expectations

Increasing reliance on digital technologies is further rising customer expectations for more complex, tailored, and integrated value propositions (Amit & Han, 2017; Teece, 2010; Williamson & De Meyer, 2012). The increased appetite for online and mobile usage, flexibility (e.g., on-demand usage), and convenience (e.g., self-management, around-the-clock service) is thus shaping the types of products and services and the way they are consumed (IAIS, 2017). For example, customers expect to connect to multiple ecosystems using a single account (e.g., *Google* and *Facebook*), to receive an insurance policy while purchasing a car (e.g., *Tesla* and *BMW*), or to receive a medical advice, drug delivery, and health insurance while shopping online (e.g., *Amazon* and *Alibaba*). Respectively, companies that have successfully capitalized on the demand-side synergies can expect to lock-in their customers early on and become leaders in their ecosystems (Adner & Kapoor, 2010). However, as much as customers value the increased value propositions, they also value their privacy and data (Nambisan et al., 2019). According to the Accenture Global Financial Service Consumer Survey (2019), customers could potentially provide access to their personal data in exchange for benefits such as more competitive prices, faster and easier services, and more relevant advice for their personal circumstances. Yet, trust is critical to retaining them. Trust may not necessarily be a driver of loyalty, but the erosion of trust may certainly increase customer attrition.

Conclusion

The ongoing digital disruption is transcending and blurring the boundaries between many industries. While in this chapter we discussed the core factors contributing to the digital transformation happening across many industries, one should note that these forces are not acting in isolation, but rather in conjunction with one another. Amid such complex environments, established companies may risk losing their market positions if they fail to adapt. At the same time, this complex and rapidly changing landscape provides plentiful opportunities for breaking down traditional business boundaries and allowing for new ways of value creation and capture through business model innovations, innovation ecosystems, and platforms.

Discussion Points

1. What would you consider as the most influential digital disruption(s) that happened in the past decade? Is digitalization always good for business and for society? Elaborate on your answer.
2. How do you think environmental shocks, such as recent COVID-19 crisis, would affect digital transformation? Do you expect to see any new potentially disruptive trends? If so, what kind?
3. What key elements for creating a successful digital business strategy are different for established versus newly founded firms?

4. When considering innovation ecosystems, how do you think a firm should decide on whether and, if so, what type of innovation ecosystem to join? What are the key factors to be considered?
5. In the increasingly digital economy, should all firms build a digital platform at the core of their business? If so, how many platforms one should build in different markets? Think also about both the competition and policy implications.

Take-Home Messages

- Digital disruption is a process that often starts incrementally, in small niches, but eventually changes the power balance in the whole market and industry.
- Business model innovations, innovation ecosystems, and platform and network effects are crucial themes underlying the dynamics of digital disruption.
- Both supply- and demand-side synergies are driving the expansion of one business into other markets, the latter synergies being more important in the digital economy.

References

- Adner, R. (2006). Match your innovation strategy to your innovation ecosystem. *Harvard Business Review*, 84(4), 98.
- Adner, R. (2017). Ecosystem as structure: An actionable construct for strategy. *Journal of Management*, 43(1), 39–58.
- Adner, R., & Feiler, D. (2019). Interdependence, perception, and investment choices: An experimental approach to decision making in innovation ecosystems. *Organization Science*, 30(1), 109–125.
- Adner, R., & Kapoor, R. (2010). Value creation in innovation ecosystems: How the structure of technological interdependence affects firm performance in new technology generations. *Strategic Management Journal*, 31(3), 306–333.
- Adner, R., & Kapoor, R. (2016). Right tech, wrong time. *Harvard Business Review*.
- Agarwal, R., & Helfat, C. E. (2009). Strategic renewal of organizations. *Organization Science*, 20(2), 281–293.
- Ahuja, G., & Novelli, E. (2017). Redirecting research efforts on the diversification–performance linkage: The search for synergy. *Academy of Management Annals*, 11(1), 342–390.
- Amit, R., & Han, X. (2017). Value creation through novel resource configurations in a digitally enabled world: Novel resource configurations in a digitally enabled world. *Strategic Entrepreneurship Journal*, 11(3), 228–242.
- Amit, R., & Zott, C. (2001). Value creation in E-business. *Strategic Management Journal*, 22(6–7), 493–520.
- Anderson, C. (2009). *Free: The future of a radical price* (1st ed.). Hyperion.
- Andries, P., & Debackere, K. (2007). Adaptation and performance in new businesses: Understanding the moderating effects of independence and industry. *Small Business Economics*, 29(1–2), 81–99.
- Andries, P., Debackere, K., & van Looy, B. (2013). Simultaneous experimentation as a learning strategy: Business model development under uncertainty: New ventures' business model development under uncertainty. *Strategic Entrepreneurship Journal*, 7(4), 288–310.
- Ansari, S., Garud, R., & Kumaraswamy, A. (2016). The disruptor's dilemma: TiVo and the U.S. television ecosystem. *Strategic Management Journal*, 37(9), 1829–1853. <https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.2442>.
- Arend, R. J. (2013). The business model: Present and future—Beyond a skeumorph. *Strategic Organization*, 11(4), 390–402.

- Autio, E., & Thomas, L. D. W. (2014). Innovation ecosystems: Implications for innovation management. In M. Dodgson, D. M. Gann, & N. Phillips (Eds.), *The Oxford handbook of innovation management* (pp. 204–228). Oxford University Press. Oxford handbooks online.
- Aversa, P., Haefliger, S., Hueller, F., & Reza, D. G. (2020). Customer complementarity in the digital space: Exploring Amazon's business model diversification. *Long Range Planning*, 1–21 (in press).
- Aversa, P., Haefliger, S., & Reza, D. G. (2017). Building a winning business model portfolio. *MIT Sloan Management Review*, 58, 49–54.
- Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, 17(1), 99–120.
- Berends, H., Smits, A., Reymen, I., & Podoyntsyna, K. (2016). Learning while (re)configuring: Business model innovation processes in established firms. *Strategic Organization*, 14(3), 181–219.
- Bocken, N. M. P., Rana, P., & Short, S. W. (2015). Value mapping for sustainable business thinking. *Journal of Industrial and Production Engineering*, 32(1), 67–81.
- Brehmer, M., Podoyntsyna, K., & Langerak, F. (2018). Sustainable business models as boundary-spanning systems of value transfers. *Journal of Cleaner Production*, 172, 4514–4531.
- Brock, K., den Ouden, E., van der Klauw, K., Podoyntsyna, K., & Langerak, F. (2019). Light the way for smart cities: Lessons from Philips lighting. *Technological Forecasting and Social Change*, 142, 194–209.
- Casadesus-Masanell, R., & Zhu, F. (2013). Business model innovation and competitive imitation: The case of sponsor-based business models: Business model innovation and competitive imitation. *Strategic Management Journal*, 34(4), 464–482.
- Cennamo, C., & Santalo, J. (2013). Platform competition: Strategic trade-offs in platform markets. *Strategic Management Journal*, 34(11), 1331–1350.
- Chesbrough, H. (2010). Business model innovation: Opportunities and barriers. *Long Range Planning*, 43(2–3), 354–363.
- Chesbrough, H., & Rosenbloom, R. S. (2002). The role of the business model in capturing value from innovation: Evidence from Xerox Corporation's technology spin-off companies. *Industrial and Corporate Change*, 11(3), 529–555.
- Christensen, C. M., McDonald, R., Altman, E. J., & Palmer, J. E. (2018). Disruptive innovation: An intellectual history and directions for future research. *Journal of Management Studies*, 55(7), 1043–1078.
- Christensen, C. M., Raynor, M. E., & McDonald, R. (2015). What is disruptive innovation? *Harvard Business Review*. Retrieved from <https://hbr.org/2015/12/what-is-disruptive-innovation>
- Christensen, C. M., & Rosenbloom, R. S. (1995). Explaining the attacker's advantage: Technological paradigms, organizational dynamics, and the value network. *Research Policy*, 24(2), 233–257.
- Crittenden, A. B., Crittenden, V. L., & Crittenden, W. F. (2019). The digitalization triumvirate: How incumbents survive. *Business Horizons*, 62(2), 259–266.
- Cucculelli, M., & Bettinelli, C. (2015). Business models, intangibles and firm performance: Evidence on corporate entrepreneurship from Italian manufacturing SMEs. *Small Business Economics*, 45(2), 329–350.
- de Vasconcelos Gomes, L. A., Facin, A. L. F., Salerno, M. S., & Ikenami, R. K. (2018). Unpacking the innovation ecosystem construct: Evolution, gaps and trends. *Technological Forecasting and Social Change*, 136, 30–48.
- Deephouse, D. L. (1999). To be different, or to be the same? It's a question (and theory) of strategic balance. *Strategic Management Journal*, 20(2), 147–166.
- Deloitte. (2016). Align price with use. *Reducing up-front barriers with usage-based pricing*. Retrieved from <https://www2.deloitte.com/us/en/insights/focus/disruptive-strategy-patterns-case-studies/disruptive-strategy-usage-based-pricing.html>
- Demil, B., & Lecocq, X. (2010). Business model evolution: In search of dynamic consistency. *Long Range Planning*, 43(2–3), 227–246.
- Dyer, J. H., & Singh, H. (1998). The relational view: Cooperative strategy and sources of interorganizational competitive advantage. *Academy of Management Review*, 23(4), 660–679.
- Eisenmann, T., Parker, G., & Van Alstyne, M. (2011). Platform envelopment. *Strategic Management Journal*, 32(12), 1270–1285.
- Foss, N. J., & Saebi, T. (2017). Fifteen years of research on business model innovation: How far have we come, and where should we go? *Journal of Management*, 43(1), 200–227.

- Fuller, J., Jacobides, M. G., & Reeves, M. (2019). The myths and realities of business ecosystems. *MIT Sloan Management Review*, 11.
- Gambardella, A., & McGahan, A. M. (2010). Business-model innovation: General purpose technologies and their implications for industry structure. *Long Range Planning*, 43(2–3), 262–271.
- Gawer, A., & Cusumano, M. A. (2014). Industry platforms and ecosystem innovation. *Journal of Product Innovation Management*, 31(3), 417–433.
- Goldenberg, J., Libai, B., & Muller, E. (2010). The chilling effects of network externalities. *International Journal of Research in Marketing*, 27(1), 4–15.
- Grossman, R. (2016). The industries that are being disrupted the most by digital. *Harvard Business Review*.
- Hartmann, P., & Henkel, J. (2020). The rise of corporate science in AI: Data as a strategic resource. *Academy of Management Discoveries*. <https://doi.org/10.5465/amd.2019.0043>
- Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2016). Capturing value from big data—A taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management*, 36(10), 1382–1406.
- Hienerth, C., Lettl, C., & Keinz, P. (2014). Synergies among producer firms, lead users, and user communities: The case of the LEGO producer-user ecosystem. *Journal of Product Innovation Management*, 31(4), 848–866.
- IAIS. (2017). *FinTech developments in the insurance industry*. International Association of Insurance Supervisors (IAIS). Retrieved from <https://www.iaisweb.org/>
- Iansiti, M., & Levien, R. (2004, March 1). Strategy as ecology. *Harvard Business Review*. Retrieved from <https://hbr.org/2004/03/strategy-as-ecology>
- Jacobides, M. G. (2019). In the ecosystem economy, what's your strategy? *Harvard Business Review*, 19.
- Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems. *Strategic Management Journal*, 39(8), 2255–2276.
- Jacobides, M. G., Knudsen, T., & Augier, M. (2006). Benefiting from innovation: Value creation, value appropriation and the role of industry architectures. *Research Policy*, 35(8), 1200–1221.
- Jacobides, M. G., Sundararajan, A., & Van Alstyne, M. (2019, February). *Platforms and ecosystems: Enabling the digital economy*. World Economic Forum.
- Kapoor, R. (2014). Collaborating with complementors: What do firms do? *Advances in Strategic Management*, 30, 3–25.
- Kathan, W., Matzler, K., & Veider, V. (2016). The sharing economy: Your business model's friend or foe? *Business Horizons*, 59(6), 663–672.
- Klempere, P., & Padilla, A. J. (1997). Do firms' product lines include too many varieties? *The RAND Journal of Economics*, 28(3), 472–488.
- Kohler, T. (2015). Crowdsourcing-based business models. *California Management Review*, 57(4), 63–84.
- Kumar, V. (2014). Making “freemium” work. *Harvard Business Review*, 925, 27–29.
- Lambrecht, A., Goldfarb, A., Bonatti, A., Ghose, A., Goldstein, D. G., et al. (2014). How do firms make money selling digital goods online? *Marketing Letters*, 25(3), 331–341.
- Laukkanen, M., & Tura, N. (2020). The potential of sharing economy business models for sustainable value creation. *Journal of Cleaner Production*, 253, 120004.
- Markides, C. (2006). Disruptive innovation: In need of better theory. *Journal of Product Innovation Management*, 23(1), 19–25.
- Markides, C. C., & Williamson, P. J. (1994). Related diversification, core competences and corporate performance. *Strategic Management Journal*, 15(S2), 149–165.
- Massa, L., Tucci, C. L., & Afuah, A. (2017). A critical assessment of business model research. *Academy of Management Annals*, 11(1), 73–104.
- Matt, C., Hess, T., & Benlian, A. (2015). Digital transformation strategies. *Business & Information Systems Engineering*, 57(5), 339–343.
- Mitchell, W., & Singh, K. (1996). Survival of businesses using collaborative relationships to commercialize complex goods. *Strategic Management Journal*, 17, 169–195.
- Morris, M., Schindehutte, M., & Allen, J. (2005). The entrepreneur's business model: Toward a unified perspective. *Journal of Business Research*, 58(6), 726–735.

- Nambisan, S., Wright, M., & Feldman, M. (2019). The digital transformation of innovation and entrepreneurship: Progress, challenges and key themes. *Research Policy*, 48(8), 103773.
- Osterwalder, A., & Pigneur, Y. (2010). *Business model generation: A handbook for visionaries, game changers, and challengers*. Wiley.
- Palich, L. E., Cardinal, L. B., & Miller, C. C. (2000). Curvilinearity in the diversification-performance linkage: An examination of over three decades of research. *Strategic Management Journal*, 21, 155–174.
- Parker, G. G., & Van Alstyne, M. W. (2005). Two-sided network effects: A theory of information product design. *Management Science*, 51(10), 1494–1504.
- Pauwels, K., & Weiss, A. (2008). Moving from free to fee: How online firms market to change their business model successfully. *Journal of Marketing*, 72, 14–31.
- Podoyntsyna, K., Song, M., van der Bij, H., & Weggeman, M. (2013). Improving new technology venture performance under direct and indirect network externality conditions. *Journal of Business Venturing*, 28(2), 195–210.
- Pralhad, C. K., & Ramaswamy, V. (2004). Co-creation experiences: The next practice in value creation. *Journal of Interactive Marketing*, 18(3), 5–14.
- Priem, R. L. (2007). A consumer perspective on value creation. *Academy of Management Review*, 32(1), 219–235.
- Priem, R. L., Wenzel, M., & Koch, J. (2018). Demand-side strategy and business models: Putting value creation for consumers center stage. *Long Range Planning*, 51(1), 22–31.
- Rietveld, J. (2018). Creating and capturing value from freemium business models: A demand-side perspective. *Strategic Entrepreneurship Journal*, 12(2), 171–193.
- Robins, J. A., & Wiersema, M. F. (2003). The measurement of corporate portfolio strategy: Analysis of the content validity of related diversification indexes. *Strategic Management Journal*, 24(1), 39–59.
- Russel Reynolds Associates. (2015). *Digital pulse 2015*. Retrieved from <https://www.russellreynolds.com/insights/thought-leadership/digital-pulse-2015>
- Russel Reynolds Associates. (2017). *Digital pulse: 2017 outlook and perspectives from the market*. Retrieved from <https://www.russellreynolds.com/insights/thought-leadership/digital-pulse-2017-outlook-perspectives-from-the-market>.
- Sampler, J. L. (1998). Redefining industry structure for the information age. *Strategic Management Journal*, 19, 343–355.
- Schilling, M. A. (2002). Technology success and failure in winner-take-all markets: The impact of learning orientation, timing, and network externalities. *Academy of Management Journal*, 45(2), 387–398.
- Schmidt, J., Makadok, R., & Keil, T. (2016). Customer-specific synergies and market convergence: Customer-specific synergies and market convergence. *Strategic Management Journal*, 37(5), 870–895.
- Sirmon, D. G., Gove, S., & Hitt, M. A. (2008). Resource management in dyadic competitive rivalry: the effects of resource bundling and deployment. *Academy of Management Journal*, 51(5): 919–935.
- Skog, D. A., Wimelius, H., & Sandberg, J. (2018). Digital disruption. *Business & Information Systems Engineering*, 60(5), 431–437.
- Snihur, Y., & Zott, C. (2020). The genesis and metamorphosis of novelty imprints: How business model innovation emerges in young ventures. *Academy of Management Journal* (in press). <https://doi.org/10.5465/amj.2017.0706>.
- Srinivasan, R., Lilien, G. L., & Rangaswamy, A. (2004). First in, first out? The effects of network externalities on pioneer survival. *Journal of Marketing*, 68(1), 41–58.
- Sun, M., & Tse, E. (2009). The resource-based view of competitive advantage in two-sided markets. *Journal of Management Studies*, 46(1), 45–64.
- Talmar, M., Walrave, B., Podoyntsyna, K. S., Holmström, J., & Romme, A. G. L. (2018). Mapping, analyzing and designing innovation ecosystems: The ecosystem pie model. *Long Range Planning*. <https://doi.org/10.1016/j.lrp.2018.09.002>
- Tanriverdi, H., & Lee, C.-H. (2008). Within-industry diversification and firm performance in the presence of network externalities: Evidence from the software industry. *Academy of Management Journal*, 51(2), 381–397.

- Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, 15(6), 285–305.
- Teece, D. J. (2010). Business models, business strategy and innovation. *Long Range Planning*, 43(2–3), 172–194.
- Teece, D. J. (2018a). Profiting from innovation in the digital economy: Enabling technologies, standards, and licensing models in the wireless world. *Research Policy*, 47(8), 1367–1387.
- Teece, D. J. (2018b). Business models and dynamic capabilities. *Long Range Planning*, 51(1), 40–49.
- Teixeira, T. S. (2019, June). Disruption starts with unhappy customers, not technology. *Harvard Business Review*. Retrieved from <https://hbr.org/2019/06/disruption-starts-with-unhappy-customers-not-technology>
- Tellis, G. J. (2010). *Network effects: Do they warm or chill a budding product?* Available at SSRN 1536854. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1536854
- Thomas, L. D. W., & Autio, E. (2020). Innovation ecosystems. In R. Aldag (Ed.), *Oxford research encyclopaedia of business and management*. Oxford University. <https://doi.org/10.1093/acrefore/9780190224851.013.203>
- Tidhar, R., & Eisenhardt, K. M. (2020). Get rich or die trying... finding revenue model fit using machine learning and multiple cases. *Strategic Management Journal*, 41, 1245–1273.
- Van Alstyne, M. W., Parker, G. G., & Choudary, S. P. (2016). Pipelines, platforms, and the new rules of strategy. *Harvard Business Review*, 9.
- Van Angeren, J., Podoyntsyana, K. S., & Langerak, F. (2017). Proceed with caution: Analyzing the performance of freemium business models in the Apple App Store. In *Proceedings of the Annual Meeting of the Academy of Management, Atlanta* (p. 13593).
- Van Angeren, J., Vroom, G., McCann, B. T., Podoyntsyana, K., & Langerak, F. (2022). Optimal distinctiveness across revenue models: Performance effects of differentiation of paid and free products in a mobile app market. *Strategic Management Journal*, 43(10), 2066–2100.
- Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, 28(2), 118–144.
- Wagner, T. M., Benlian, A., & Hess, T. (2014). Converting freemium customers from free to premium—The role of the perceived premium fit in the case of music as a service. *Electronic Markets*, 24(4), 259–268.
- Weill, P., & Woerner, S. L. (2015). Thriving in an increasingly digital ecosystem. *MIT Sloan Management Review*, 56(4), 26–34.
- Westerman, G., & Bonnet, D. (2015). Revamping your business through digital transformation. *MIT Sloan Management Review*, 6.
- Williamson, P. J., & De Meyer, A. (2012). Ecosystem advantage: How to successfully harness the power of partners. *California Management Review*, 55(1), 24–46.
- Ye, G., Priem, R. L., & Alshwer, A. A. (2012). Achieving demand-side synergy from strategic diversification: How combining mundane assets can leverage consumer utilities. *Organization Science*, 23(1), 207–224.
- Zhao, E. Y., Fisher, G., Lounsbury, M., & Miller, D. (2017). Optimal distinctiveness: Broadening the interface between institutional theory and strategic management. *Strategic Management Journal*, 38(1), 93–113.
- Zhao, E. Y., Ishihara, M., Jennings, P. D., & Lounsbury, M. (2018). Optimal distinctiveness in the console video game industry: An exemplar-based model of proto-category evolution. *Organization Science*, 29(4), 588–611.
- Zott, C., & Amit, R. (2007). Business model design and the performance of entrepreneurial firms. *Organization Science*, 18(2), 181–199.
- Zott, C., & Amit, R. (2008). The fit between product market strategy and business model: Implications for firm performance. *Strategic Management Journal*, 29(1), 1–26.
- Zott, C., & Amit, R. (2010). Business model design: An activity system perspective. *Long Range Planning*, 43(2–3), 216–226.
- Zott, C., Amit, R., & Massa, L. (2011). The business model: Recent developments and future research. *Journal of Management*, 37(4), 1019–1042.



Digital Servitization in Agriculture

Wim Coreynen and Sicco Pier van Gosliga

Contents

- 14.1 Introduction – 332
- 14.2 Servitization – 333
- 14.3 Types of Services – 334
- 14.4 Servitization
in Agriculture – 336
- 14.5 Digital Servitization – 338
- 14.6 Digital Servitization
in Agriculture – 339
- References – 348

Learning Objectives

After having read this chapter, you will be able to:

- Understand the meaning of servitization, its potential benefits and pitfalls for companies.
- Understand the different types of services that companies can provide, ranging from basic repair services to complex customer solutions.
- Name and explain the challenges faced by farmers that urge them to rethink their current business model.
- Discuss the different ways in which technology—implemented in either the back end or the front end of the organization—can enable servitization.
- Discuss how farms and farm suppliers can leverage data to provide services to customers and tap new markets.

14.1 Introduction

Since the industrial revolution, new technologies are continuously transforming companies, sectors, and ultimately even entire economies. Starting from the mid-nineteenth century, new manufacturing technologies such as the steam engine, assembly line, and automation have changed the agricultural economy into an industrial economy. More recently, information technologies (IT) are increasingly transforming the industrial economy into a service economy through the Internet, computers, and other electronic devices (Rust & Huang, 2014). From a company's perspective, this recent transition is known as “servitization” (Vandermerwe & Rada, 1988).

Definition

Servitization describes the process whereby companies add services to their core product offerings to create additional customer value (Raddats et al., 2019).

Such combined product-service offerings have been referred to as “integrated solutions” (Davies, 2004; Matthyssens & Vandenbempt, 2008), “hybrid offerings” (Ulaga & Reinartz, 2011), and also “product-service systems” (PSS) (Baines et al., 2007; Tukker, 2004). For example, Atlas Copco, a manufacturer of air compressors, has started to provide machines-as-a-service (MaaS), and next, they plan to literally sell air by offering compressed-air-as-a-service (CaaS) (Link Magazine, 2018). Other well-known examples of servitization are Rolls-Royce's power-by-the-hour service, which offers guaranteed flight hours for their airplane engines, and Xerox' pay-per-copy service for their office printers (Kowalkowski et al., 2017).

Servitization was first suggested in the late 1980s as a trend that is “pervading all industries, is customer-demand driven, and is perceived by corporations as sharpening their competitive edges” (Vandermerwe & Rada, 1988, p. 314). Since

then, interest from both the industrial and academic community has been growing exponentially (Fliess & Lexutt, 2017). Though the core of servitization research takes place within industrial manufacturing (Baines et al., 2009a), there have been studies in other sectors as well, such as the maritime sector (Pagoropoulos et al., 2017), road transport sector (Bigdeli et al., 2017), and also the agricultural sector (Pereira et al., 2016; Vidickiene & Gedminaite-Raudone, 2018).

The aim of this chapter is to show how companies can leverage technology to unlock the opportunities of servitization. Drawing from three cases from the Dutch agricultural sector—one chicken farm and two equipment suppliers, one for the horticultural sector and one for dairy farmers—we discuss how companies can differentiate themselves from the market by supporting customers through data collection, analysis, interpretation, and reporting. We selected these cases for three specific reasons. First, farmers are under enormous pressure these days, because the traditional business model of farming—that is, increasing production volume while improving technical efficiency—has run its course. Second, though the literature so far has paid little attention to the agricultural sector, it has been found that moving into services potentially holds more benefits for agricultural players than companies from other sectors. Third, we purposefully selected smaller companies (one case is a family-owned farm and two are start-ups), because SMEs usually have limited time and resources to experiment with new technology-driven business models. This chapter shows that servitization provides opportunities not only for large manufacturers but also for SMEs. We will discuss these reasons more in depth later in this chapter. First, we dig deeper into the potential benefits and pitfalls of servitization and show different types of services that companies can offer.

14.2 Servitization

When the term servitization was first introduced by Vandermerwe and Rada (1988), the research field was mainly focused on answering the question: Why should product companies move into service? Overall, services are considered to offer several *strategic*, *marketing*, as well as *financial* opportunities for companies (Baines et al., 2009b).

First, servitization helps companies differentiate from the competition (Kamp & Parry, 2017). Particularly in commoditized markets, where customers perceive products as more or less similar, companies can (re)gain a superior market position by offering value-added services (Matthyssens & Vandenbempt, 2008). For example, in the switchboard manufacturing sector, companies compete mostly on price and delivery time. To differentiate themselves from the market, some manufacturers have started to also provide maintenance and upgrade services for their installed base (Coreynen et al., 2018).

Second, providing services on top of products generates loyalty among customers, and it can even influence their purchasing decisions later (Gebauer & Fleisch, 2007; Ulaga & Reinartz, 2011). For example, on top of manufacturing and delivering customized insoles (i.e., removable soles worn in shoes), an insole supplier also offers a whole range of other products and services for podiatrists, such as measur-

ing equipment (e.g., foot scanners), software (e.g., to design insoles) and specialist podiatrist training (Coreynen et al., 2018).

Third, moving into services leads to better performance and growth. Past studies have associated servitization with increased company market value (Fang et al., 2008), more stable revenue streams as well as higher profitability rates (Eggert et al., 2014), and even higher employment levels (Crozet & Milet, 2017). For example, despite a severe drop in the number of locomotives sold between 1999 and 2001, General Electric (GE) maintained its usual operating margins thanks to the growth of its service business (Welch & Byrne, 2003).

Despite these opportunities, servitization is not considered to be easy, and many companies struggle to move into the service business (Gebauer et al., 2005). Before we move on, a few words of caution about the potential pitfalls of servitization are in place.

First, different service strategies may be better suited for different business environments (Gebauer, 2008). For instance, when there is little industry growth, servitizing companies are more likely to grow, but when industry growth is high, companies are better off paying close attention to their core (product) business.

Second, servitization can also fail when the newly created services may not be what customers want (Valtakoski, 2017). For example, American farmers have started a right-to-repair movement against John Deere, and some have even turned to tractor hacking to fix broken-down tractors themselves (Koebler & Wanstreet, 2018).

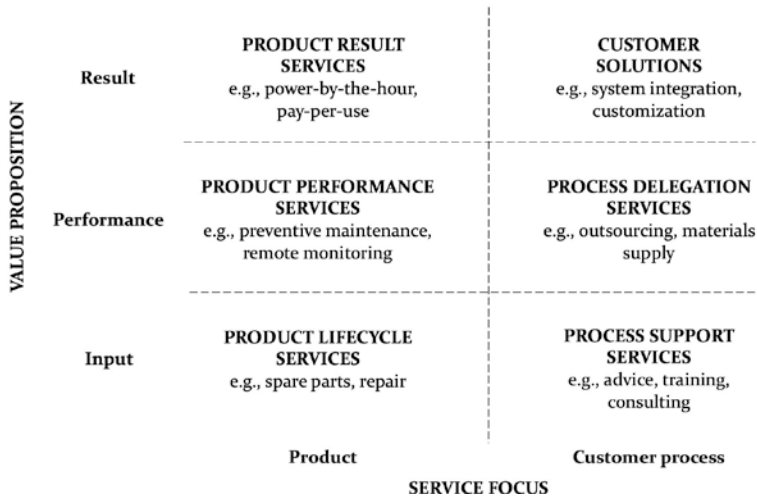
Third, companies should also develop new resources and skills to successfully create, sell, and deliver services (e.g., Kindström et al., 2013; Ulaga & Reinartz, 2011). For example, salespeople should be able to listen to the customer, be empathetic to their problems (even those outside the scope of the company), and think in terms of offering solutions rather than pushing products.

Finally, over time, technologies can change and severely disrupt the supply chain. For example, a capital goods manufacturer lost visibility to its installed base due to evolving product technology and decided to pull back from servitization. This reverse pathway is known as “deservitization” (Finne et al., 2013; Kowalkowski et al., 2017).

14.3 Types of Services

When Henry Ford introduced the Model T back in 1908, he said: “Any customer can have a car painted any color that he wants, so long as it is black” (Ford & Crowther, 1922, p. 72). More than a century later, people customize their car to the smallest detail, or they just use a car through leasing, renting, or sharing, without the hassle of buying or owning one (Tukker, 2004). In the near future, self-driving cars will hit the road and people will not even have to worry about the task of driving anymore (The Economist, 2018).

Though servitization can be a valuable pathway for companies to (re)gain market power and boost their performance, it does not simply imply that they follow a



■ Fig. 14.1 Service typology based on Coreynen et al. (2017), Kowalkowski and Kindström (2014), and Ulaga and Reinartz (2011)

linear transition from selling products to services (Oliva & Kallenberg, 2003). In fact, companies can offer many different types of services, and some even offer different services to different markets simultaneously.

► Important

There are two popular service-offering categorizations: One distinguishes between different types of service ‘focus’, that is, either on the *product* or the *customer’s process* (Oliva & Kallenberg, 2003; Ulaga & Reinartz, 2011), while the other distinguishes between different types of ‘value’ provided to customers, such as services that are offered either as an *input* (for customers who want to do it themselves), as a *performance* (for customers who want others to do it with them), or as a *result* (for customers who want others to do it for them) (Baines & Lightfoot, 2014; Tukker, 2004).

■ Figure 14.1 combines both categories and gives an overview of the different types of service offerings, including some practical examples.

► Example

Consider the fictitious example of a truck manufacturer. At first, the company builds and sells trucks and provides customers product life cycle services such as spare parts and repair, in case one of their trucks breaks down. Next, it gradually expands into services oriented towards the performance of the product. For example, through preventive maintenance, the company guarantees its customers that their trucks will no longer break down. Ultimately, the company also makes trucks available for customers to use, for instance, through leasing or renting. In this case, the company provides customers a result (e.g., a fully working truck for a period of time) while maintaining ownership of the product.

On top, the manufacturer further expands its offering by not simply servicing and leasing trucks, but also by focusing on customers and the wider goals they want to achieve. The company starts to support customers who prefer to maintain control over their own operations through training and consulting, for example, on how to drive safely or save on fuel. Next, customers can delegate particular activities to the company, such as truck cleaning, so they do not have to worry about these tasks anymore. Finally, the truck company offers total transport solutions to customers who want to outsource their entire transport operations, such as food companies, so they can focus on their core business. ◀

14.4 Servitization in Agriculture

The previous sections focused mostly on how product manufacturers can grow through services, but how is servitization relevant for the agricultural sector? For that, we first need to understand the pressures that farmers are facing today.

In the twentieth century, the agricultural sector was mainly concerned with ensuring a stable income for farms by increasing their production volume and improving technical efficiency. Yet over the last few decades, the sector has evolved drastically, and there are several reasons why this business model is no longer working. First, due to the consistent increase in productivity, food shortages have become a problem of the past (at least, in developed economies) (Koba, 2013), but overproduction has caused the markets to saturate, leaving little incentive for farms to produce as much as they can. Second, the need to continuously invest to become more efficient while also meeting the food industry's increased quality requirements has raised farms' costs of production. For example, American chicken farmers are continuously requested by food companies to upgrade their facilities, just so they can stay in the game (Lee & Cappellazzi, 2015). Third, farms need to comply with a growing body of legislation on animal welfare protection and environment preservation. For example, they need to adopt new practices that reduce the erosion caused by working farmland (Lichtenberg, 2019). Fourth, globalization has made the agriculture business riskier for farmers. Because of free trade, they need to compete with countries where labor costs are lower and legislation is less restrictive (Matheny & Leahy, 2007). Finally, the consequences of climate change (in part, caused by farming), such as the emission of greenhouse gases and conversion of forests into agricultural land (FAO et al., 2018), have made the outcome of agricultural activity less certain. In summary, being a farmer has become a more costly and high-risk profession, and the number of successors and individuals wanting to take over farms is severely dropping (Vidickiene & Gedminaitė-Raudonė, 2018).

Following industry, the agricultural sector is increasingly paying attention to servitization as a way to rekindle growth and improve the perception of local farming activity (Vidickiene & Gedminaitė-Raudonė, 2019).

Definition

From the end consumer's point of view, servitization is considered a way to respond to the increasing demand for fresh, high-quality, and locally produced food, complying with environmental and animal welfare standards. From a farmer's perspective, servitization has been defined as the "transformational process that requires rethinking all aspects of the business: production structure and methods, marketing, pricing, service delivery infrastructure and financial management" (Vidickiene & Gedminaite-Raudone, 2018, p. 1552).

Several agricultural service initiatives have already been reported. For example, some farmers are shortening the food supply chain by selling products straight to the end user through farmers' markets (Vignali et al., 2006). Others take part in community-based farming initiatives through teaching and renting farmland to city residents who want to grow vegetables and fruit in the countryside (Vidickiene & Gedminaite-Raudone, 2019). Farmers have also started to service other (mostly large-sized food) companies through consulting and even experimenting with new crop species.

Not only farms but also farm suppliers are increasingly expanding their service offering. For instance, manufacturers of farming equipment (e.g., reapers, dorsers) have started offering leasing services, so farmers can avoid the initial high investment cost of buying the equipment themselves (Corti et al., 2013). Also, suppliers are customizing their offerings more efficiently in order to better suit farmers' needs. For example, instead of making multiple engines with different levels of horsepower, John Deere offers engines for which farmers can modify the level horsepower through software alone (Porter & Heppelmann, 2014). Finally, other services by suppliers of agricultural equipment include (preventive) maintenance and upgrade services. For example, Foton Lovol built an Internet of Things (IoT) after-service platform to provide farmers with equipment monitoring and maintenance services (Wang et al., 2014).

Besides farms and suppliers of farming equipment, other players are entering the agricultural service sector as well. For example, pesticide companies have started to offer crop protection services to farmers, such as advice and training; some even take over the complete management of crops' health (Pereira et al., 2018). Also new network organizations are formed to support farmers by combining resources and building expertise. For example, dairy farm cooperatives offer farms heifer breeding and fodder production and delivery services (Pereira et al., 2016).

Servitization in agriculture creates several *economic*, *financial*, and *environmental* benefits for the agricultural sector in particular, and even for society at large. First, by outsourcing specific activities, such as the production of cattle fodder and crop protection, farms can spend more time and resources on their focal activity (Pereira et al., 2016). Second, experienced farmers can remain active longer by sharing knowledge with younger, aspiring farmers and also with other companies, for example, through training and consulting (Vidickiene & Gedminaite-Raudone,

2019). Third, it has been found that servitized manufacturers of agricultural equipment are more profitable and show higher employment levels (Crozet & Milet, 2017). Finally, servitization improves farms' eco-efficiency. For example, dairy farms have lowered their carbon footprint by using their land and farm machinery more efficiently (Pereira et al., 2016), and vineyards have reduced the amount of pesticides used by opting for pest management solutions. Here, the provider focuses on managing crops' health while striving for the most efficient use of pesticides (Pereira et al., 2018).

14.5 Digital Servitization

The antecedents of servitization in agriculture go back at least 170 years, as the earliest noted examples date from the mid-nineteenth century in the USA (Schmenner, 2009). The bundling of goods and services was led by companies without exceptional manufacturing skills as a way to compete. For example, McCormick, a manufacturer of reapers (the most complex piece of farming equipment at the time), reduced production at its Chicago factory during harvest to send workers into the field for repair services (Chandler & Hikino, 1994; Schmenner, 2009). Back then, offering services was made possible through the invention of new technologies, such as the telegraph and completion of the railroad network, which allowed for better, faster, and more complete communication and transportation (Schmenner, 2009). Without these new technologies, the management and coordination of geographically dispersed factories and offices would not have been possible.

Today, new technologies continue to transform businesses, sectors, and even entire economies.

Definition

Digital servitization is known as the use of digital technologies that enable companies to shift from a product-centric to a more service-centric business model (Sklyar et al., 2019).

14

Digitization essentially means the transformation from analogue to digital (Hsu, 2007; Storbacka, 2018), such as the digitization of administrative documents and procedures. Yet, because competitors are quick to follow suit by acquiring or copying similar technologies, digitization on its own offers limited possibilities for building a sustainable competitive advantage (Barney, 1991; Carr, 2003). Therefore, companies increasingly apply technology as an enabler for new service-driven business models, which are far more difficult to copy.

Important

There are three main digitally enabled service transitions. First, by using IT in the *back* end of the organization (e.g., software for system integration, program optimization), companies create scalability in the production and delivery of customized

solutions. For example, a metalwork supplier implemented a new software to produce metal components more efficiently, which later also enabled them to produce more customized and fully assembled solutions for customers, such as bed frames (Coreynen et al., 2017). Second, by leveraging IT in the *front* end (e.g., websites, online applications), they open new opportunities for (continuous) interaction with customers. For example, a switchboard manufacturer launched a user-friendly web app for small installation companies, so they can configure and order switchboards online (Coreynen et al., 2017). Third, by merging products with IT (e.g., through sensors, IoT), companies create smart products to stay connected with customers (Laudien & Daxböck, 2016; Santos et al., 2017). For example, a mining equipment manufacturer operates entire fleet of equipment far underground, and technicians are dispatched only when issues requiring human intervention arise (Porter & Heppelmann, 2014).

In the near future, big data, advanced data analytics, and artificial intelligence (AI) will further reshape service by gradually taking over increasingly complicated tasks in both the front and back end of the organization (Huang & Rust, 2018). The evolution from *descriptive* analytics (i.e., using past data to report about the past) towards *predictive* (i.e., using past data to predict the future) and ultimately *prescriptive* analytics (i.e., using analytical models to optimize current behavior and future actions) (Storbacka, 2018) will allow companies to move from merely monitoring individual products to controlling their functions, optimizing their performance, and providing fully autonomous systems of products. For example, John Deere has evolved to providing farming management systems that connect their tractors to other systems, such as weather maps and field sensors (Porter & Heppelmann, 2014).

However, the saying goes: “A fool with a tool remains a fool” (Thurlbeck, 2012). On top of investing in new technologies, digital servitization also requires companies to develop new IT-specific resources and competences. The collection of customer data, for one, and the ability to process and interpret data offer companies unique advantages to improve their own products while also creating new services that better address customer needs (Ulaga & Reinartz, 2011). For example, based on the data collected through the new web app, the switchboard manufacturer was later able to produce panels at a lower cost for installers than if they were to buy all components separately (Coreynen et al., 2017). Furthermore, on top of developing new digital skills and resources, digital servitization also requires more intense collaboration with other stakeholders. By collaborating with customers, suppliers, and also start-ups, companies can gain access to resources, skills, and knowledge that lie beyond their own abilities (Eloranta & Turunen, 2016), as shown by the following cases.

14.6 Digital Servitization in Agriculture

We present three cases from the Netherlands as illustrative cases on digital servitization in agriculture: Kuijpers Kip, a family-run chicken farm, and 30MHz and Connecterra, two start-ups that provide data services for the horticulture and dairy

sector, respectively. These cases are selected because they are representative for being at different stages of servitization at the time of writing, which was in 2019. Before the COVID-19 pandemic, Kuijpers Kip was exploring the benefits of adding services to its current business, 30MHz had recently transformed from a product-oriented to a primarily service-oriented business model, while Connecterra had opted for a service-oriented business from the start to differentiate itself on the market. Also, these cases view servitization from both a farm's and a supplier's perspective. In the next few sections, we discuss every case in detail, before further discussing their stories from a digital servitization lens.

Kuijpers Kip

Kuijpers Kip is a poultry chicken farm that is run as a family business, located in the south of the Netherlands. The business is dedicated to breeding chicken for meat production. Its mission is to produce tasty meat at an affordable price in the most sustainable manner under conditions that guarantee the health and well-being of the chicken.

Motivated by its goal to be both more environmentally sustainable and animal friendly, Kuijpers Kip decided in 2004 to grow its production up to a scale that allows more on-site facilities. Transport of livestock has a negative impact on the well-being of the animals, as well as the quality of meat due to increased levels of cortisol in the chicken after transportation. Also, transporting livestock from either hatcheries to broiler farms, or from farms to abattoirs, comes with an increased risk to infectious diseases. To fully eliminate the need for transport throughout the full life span of the chicken, it is planned to have a hatchery for in-house hatching and an abattoir at the farm. With Vencomatic, Stienen BE, and INNO+ (three local housing climate experts), Kuijpers Kip designed the barn in such a way that it can reuse

warmth for the hatchery. The gas emissions of the barn are filtered from smells, fine dust, and ammonia to limit the impact on its surroundings. In 2018, after a 15-year-long procedure, the plans were approved to start the construction of a new barn that can hold up to one million chickens. This scale is based upon the smallest size upon which it is economically feasible to have an on-site poultry abattoir. As a result of these plans, Kuijpers Kip is not only growing as a business, but, in effect, is also changing its business model from selling chicken to producing sustainable and animal-friendly food.

Technology plays a central role in enabling this transition. While scaling up, Kuijpers Kip is moving into a data-driven business, as the technologies that are deployed also enable the offering of some additional services. A wide variety of sensors are installed throughout the barn to allow for the continuous monitoring of livestock and experimentation. For example, tests have been conducted to measure chicken stress levels with acoustic sensors. Data is used to manage the farm itself, but it is also commercially offered to third parties.

The produced data is exchanged for specific expertise that benefits Kuijpers Kip or sold to interested stakeholders through 5-year contracts. By commercially sharing data, Kuijpers Kip collaborates with other companies on technological innovations. One of these contracts is with lightning company Signify to pilot new multispectral lighting in the barn. Kuijpers Kip also signed a contract with a veterinarian organization to track and ensure the health and welfare of the chicken—they are paid a fixed price per chicken rather than per treatment. The revenues from selling data in multi-year contracts help finance new innovations. Regardless of its financial ability to invest, Kuijpers Kip only invests in innovations at the farm that improve its margins and thereby are financially self-supporting.

By becoming a data provider to third parties that, in fact, develop innovations for all poultry farmers, Kuijpers Kip is indirectly benefiting its competitors. However, Kuijpers Kip does not see these innovations as a threat. They are considered primarily a threat to conservatively run farms, reasoning that early adopters—like themselves—will be less disrupted by innovative concepts and techniques. With the current investments, the farm aims to scale up its production of chicken meat to 2% of the national production in the Netherlands. Operating at this scale introduces some new hurdles,

as large-scale livestock farming is met with distrust from vocal parts of Dutch society. The scale on which Kuijpers Kip intends to operate and its new farming practices, which are intended to limit transport and keep a low ecological footprint, are uncommon in the Netherlands. These measures are not recognized by existing welfare and organic labels, despite their effectiveness. To earn the trust of consumers and other parties in the production system regarding respect for animal welfare and environmental sustainability, Kuijpers Kip is developing a limited distributed ledger to store information—a blockchain—that creates transparency and traceability throughout the production chain. It is also introducing its own label that is complementary to existing labels to address the demands of critical consumers. Kuijpers Kip has a partnership with Food Insights, a blockchain technology company, and an automation company to develop such a platform. Sensors enable the accumulation of information for each individual chicken through the whole production and welfare chain. All transactions are securely and immutably stored within the blockchain. Technical information, such as data from sensors, is stored outside the blockchain, yet will be 100% traceable. By pioneering this technology, Kuijpers Kip wishes to become a trustworthy provider of meat, as well as a trustworthy provider of data-as-a-service.

30MHz

The company 30MHz is based in Amsterdam, the Netherlands' capital. It was launched in 2014 as an IoT hardware company that also offers some supportive services. Over the years, 30MHz' offering evolved into a data platform for the horticultural sector. The company today has about 35 employees and more than 300 customers in over 30 countries.

In its initial years, 30MHz had a diverse customer base across multiple industries in the Netherlands. The company sold plug-and-play sensor solutions to a wide variety of customers, such as the Dutch Association of Mental Health and Addiction Care (GGZ), WTC Schiphol, and the Port of Amsterdam. 30MHz' business model was focused on selling its sensor solutions for real-time monitoring tasks, such as the management of facilities' availability. The hardware was supported by a cloud-based platform, which was offered as an additional service. Over time, 30MHz discovered that it had the most success in the agricultural sector and therefore decided to focus solely on specializing its products and services for farmers. After changing its focus, the company has been growing its customer base, which is now in majority made up of horticultural companies that are distributed globally. By targeting a single market, 30MHz was able to better focus its sales and product development efforts, reduce its cost of operation, and increase its growth and profitability. Their product evolved into an IoT solution for real-time crop monitoring that can be shipped directly to farmers and set up without special training or assistance, which lowers the costs of onboarding new customers.

Whereas 30MHz' primary focus originally was on the sales of plug-and-play

sensors, over the years, the supportive platform has become its main selling point. Since 2019, the company has been shifting from selling products to providing services, tools, and support for farmers and growers to digitalize their environments. 30MHz changed its business model from selling sensors to exclusively offering a cloud-based platform that caters the agricultural sector and its multiple stakeholders. The platform is ingested with sensor measurements based on in-house-developed IoT sensor technology. Farmers can interactively inspect the climatic conditions via the 30MHz Platform. This helps them by continuously monitoring and analyzing the conditions under which crops are grown, thereby enabling them to produce plants and vegetables in an increasingly sustainable manner. 30MHz offers a wide range of sensors that can be connected to its platform, for instance, to measure CO₂, airflow, moisture, and temperature. The sensors are connected to the 30MHz Platform via a gateway, which can connect to up to 4,000 individual sensors. Typically, one gateway per farm is enough. The company differentiates itself from the competition by letting farmers maintain the sole ownership of their data. The current business model is driven by a monthly subscription to the 30MHz Platform, with additional monthly fees for extra sensors. By introducing these recurring revenues, the business has become less volatile and now also has a lower cost of entry for potential new customers.

In recent years, the company has been actively looking for partners to add additional data sources to its platform, on top of its own in-house-developed

hardware, thereby focusing on becoming a more integrated and relevant platform for farmers. One of 30MHz' current partners is Fargro Ltd., a wholesaler and supplier of commercial horticultural products that has been offering 30MHz hardware in the UK since 2018. Proeftuin Zwaagdijk, an agricultural and horticultural research center in the Netherlands, became a research partner and customer of 30MHz in 2017. It tests new applications of 30MHz' technology and supports innovative, subsidized projects with the company to demonstrate the value of crop-level data in practice. Wageningen University & Research (WUR) also serves as a partner by advising 30MHz on its product road map to stay up to date of the changing needs of growers worldwide. Furthermore, in 2019, WUR formed a team with 30MHz and Delphy, a Dutch agricultural consultancy company, for the Autonomous Greenhouse Challenge to successfully demonstrate

that growing high-quality and profitable crops can be done autonomously with the 30MHz Platform.

The future ambition of 30MHz is to maximize its impact by digitalizing the entire indoor agriculture industry. In 2019, it received an investment of 3.5 million EUR from two agricultural related funds: SHIFT Invest and the Rabo F & A Innovation Fund. The investment is used to accelerate new product development and further improve worldwide distribution. With this new funding, 30MHz seeks to collaborate with other technology providers for the indoor agriculture industry to digitalize their products and services. Where growers now use separate tools and systems for climate control, irrigation, and pest management, it is 30MHz' ambition to integrate all this equipment and information flows into one digital platform that allows for a more sophisticated farm management plan.

Connecterra

Connecterra is a start-up company located in Amsterdam that addresses the needs of dairy farmers through IoT and AI technology. Since the company was founded in 2014, it has grown to 32 employees by 2019. Furthermore, it received 6 million EUR in two successive rounds of funding in 2016 and 2018, including a 1.7 EUR million grant from the European Union (EU).

Contrary to the prior two cases, Connecterra did not go through a servitization process but distinguished itself from incumbents by directly choosing for a service-oriented business model. The

company offers a digital intelligent assistant to dairy farmers called IDA, which monitors the health and fertility of cows, and alerts farmers when a cow needs attention. The service is offered at a monthly subscription fee per cow. As part of the subscription, each cow is equipped with a sensor that keeps track of all its behavior, such as rumination, eating, lying, walking, and more. This information is continuously sent to IDA's centralized server, where it is analyzed, and relevant findings are communicated back to the farmer and shown as easy-to-interpret insights via an app. If applica-

ble, the findings are bundled with a recommended action and a request for optional feedback. Since Connecterra started to offer its service commercially in early 2017, it was able to rapidly increase its customer base. IDA is already active in 14 countries and used by dairy farmers with 20 to 3,500 cows per farm.

One of the motivations for Connecterra to offer IDA as a subscription service is the flexibility and accessibility for farmers to start using the technology. Positioning IDA as a service creates a durable relationship between Connecterra and its customers, as farmers do not have to invest in the sensors themselves. Also, the subscription fee keeps users incentivized and engaged to respond to alerts, which helps improve the whole system as a result, and the recurring revenues are used to sustain the AI platform. Over time, gradual improvements are achieved by utilizing machine learning algorithms at the sensors and the centralized server. Having an adaptive, continuously updated system is a necessity because the dairy farms' operations, building structures, and composition of the herds change over time. IDA's machine learning capabilities enable Connecterra to expand its feature set and tailor it to the specific needs of not only farmers but also other users. Since 2018, Connecterra has been commercially offering a version of IDA specifically aimed at farms' supply-chain partners, such as dairy processing companies (e.g., Danone) and input suppliers (e.g., Bayer), with insights tailored to their needs. In 2019, the company also started trials in Africa in cooperation with IFC, a sister organization of the World Bank. This way, Connecterra targets small-holder farmers, larger farmers, as well as

their industrial stakeholders, all ingesting relevant data for the IDA platform to evolve.

Yet, for dairy farmers to sign up for a subscription service and receive advice is breaking with tradition and requires a new way of thinking. For one, farm suppliers, such as feed and breeding companies, commonly offer advice to farmers free of charge. Also, farmers are used to investing in the ownership of farm equipment with long-term loans. In comparison, a monthly paid subscription to IDA has a low cost of entry, and it is considered a low-risk investment accessible to a wider range of dairy farmers, particularly those with less financial resources. But farmers who previously invested in equipment may still be committed to paying off loans to systems whose functionality overlaps with IDA's features.

Next to the data pulled from farm management programs, Connecterra keeps on adding additional third-party data sources to its platform to offer a wider range of assistance and thus increase its value for farmers. For example, the company has partnered with Microsoft and Google to elevate its technology and expand its outreach. Likewise, Connecterra collaborates with WUR in the Netherlands and Aarhus University in Denmark on Horizon 2020 research projects funded by the EU to further understand and improve the impact of precision livestock farming on cow welfare, sustainability, and farm efficiency. With these partnerships, Connecterra aims to further build on an AI-driven technology that learns how to increase the productivity of farms, of all sizes and in all regions of the world, while simultaneously reducing the environmental impact of farming.

Discussion

From a farm's perspective, at first sight, it seems that there are little opportunities for digital servitization. It is highly unlikely that farms one day will offer vegetables, chickens, or perhaps even cows "as-a-service." Contrary to the equipment manufacturers, which can easily add sensors to their products, farmers create produce (e.g., grain, vegetables, meat) that is sold to customers (e.g., retailers, food companies) and subsequently gets lost down the supply chain. This is similar to other input-to-process suppliers (e.g., metal component manufacturers) that create products for other companies (e.g., equipment manufacturers) that transform their products in such a way that they cease to exist as separate entities (Storbacka et al., 2013). However, there are some opportunities for farms to extend into services to support customers in their process. Based on the case of Kuijpers Kip, we discuss two types of potential service transitions.

The first concerns taking control over and integrating different activities of the (food) supply chain (e.g., from hatching to broiling, slaughter, and finally food processing). This not only reduces the costs of farming (e.g., animal transports are no longer necessary), but also improves the quality of the final product (e.g., better animal welfare leads to higher quality meat). From a technological perspective, an important enabler for this transition is back-end digitization. By implementing technology in companies' back-end operations (e.g., sensors that measure chickens' stress levels, cameras that automatically count the number of hatched eggs), farmers can create further efficiency gains that may enable them to expand their position in the supply chain. The second service transition builds further on the first one, as it concerns the further expansion into new markets by providing data-as-a-service. By collecting data, farms create much valuable information for other companies, such as technology manufacturers, which can use these data to improve their own offering. An example of this is Kuijpers Kip's partnership with Signify to develop new multispectral lighting.

From an agricultural supplier's perspective, there seems to be a much wider range of digital servitization options to choose from. First, on top of selling hardware, such as farming machinery and equipment, suppliers can provide services to farms that want to remain in control over their own operations. For example, 30MHz at first only provided custom-built IoT sensor hardware to customers, and their platform was offered only as an additional service option. Second, they can provide services to farms that require more support. For example, Connecterra equips dairy farmers' cows with sensors that serve them as a digital assistant. IDA monitors their activities and informs the farmer when their intervention is required. Third, they can provide services for farms that want to completely outsource particular activities to better focus on their core business. For example, in the future, 30MHz and Connecterra could provide services that not only monitor but also guarantee crops and cows' health. However, this last option is a risky strategy, because it lays the burden and responsibility entirely on the supplier. Also, it would require suppliers to intervene quickly when action is necessary, which may not be feasible. 30MHz and Connecterra, for example, are both small companies that, despite their limited resources, serve customers worldwide. Perhaps other players may be interested to fulfil this role.

For example, private companies in Spain have started to offer crop protection services that guarantee vineyards' health (Pereira et al., 2018), and cooperatives take care of dairy farms' entire fodder production and delivery (Pereira et al., 2016).

From a technological standpoint, the emphasis for suppliers lies on front-end digitization, whereby companies apply technology (e.g., sensors, IoT, AI) in the front end to continuously connect with the farmers and offer them different kinds of services. Contrary to farmers, which use sensors in the back end to improve their *own* operations, suppliers use sensors to improve their *customers'* operations. Therefore, we consider their transition to be in the front end. Furthermore, simply using sensors to provide farmers descriptive services (i.e., reports about the past) is probably not enough to gain a sustainable competitive advantage, because other suppliers can easily offer similar technologies and services. Suppliers also need to build a solid customer base to collect huge amounts of data, which enables them to further move into predictive services (i.e., predict the future) and ultimately prescriptive services (i.e., optimize the future) through machine learning and AI. For that purpose, 30MHz has refocused the company entirely toward the horticultural sector, which can use their platform to continuously monitor crops in real time. Connecterra even goes one step further by also providing feedback to dairy farmers and recommending specific actions.

Conclusion

The purpose of this chapter is to provide insight in digital servitization as a potential pathway for companies, particularly farms and their suppliers, to build a sustainable competitive advantage and generate future growth. Three cases from the Dutch agricultural sector offer several illustrative examples of how such a technology-enabled transition may unfold. Applying technology not only holds benefits in terms of improving efficiency and reducing costs, but also enables companies to roll out different types of service strategies related to data collection, analysis, interpretation, and reporting.

For one, farmers can invest in technology in the back end of the company, making them better suited to perform different activities and strengthen their position in the supply chain. They can leverage new technologies to better monitor products and operations, and use the data gathered to collaborate with other sector players (e.g., technology manufacturers). The case of Kuijpers Kip is an example of such a service transition. Second, technology can also be used by companies in the front end to better connect with customers, maintain continuous relations, and offer different types of service offerings. These services can focus either on guaranteeing the proper functioning of the product or unburdening the customer in their process. 30MHz and Connecterra are two case examples of such a transition, from simply supplying hardware to providing fully integrated platforms that support farmers in managing their operations.

Take-Home Messages

- Adding services to products offers several strategic, marketing and financial opportunities, though companies should also be aware of the potential pitfalls of servitization.
- The use of digital technologies enables servitization, and *vice versa*, the transition into services is an enabler for further data science (i.e., data collection, analysis, interpretation, and reporting).
- Specifically, the use of sensors and IoT offers farms and farm suppliers several paths to extend into service, transform their business, and (re)generate growth. For instance:
 - Farms can leverage data to take control over (and integrate) different activities of the supply chain. They can also provide their data to other companies as a service and/or collaborate with them on new farming equipment and applications.
 - Suppliers can leverage insights from data to offer farmers either descriptive, predictive, or prescriptive services. To do so, they need to build a sufficiently large customer base and develop the necessary analytical skills to transform large amounts of data into suitable recommendations.

? Questions

1. What is servitization?
2. What types of services can companies offer?
3. Why should the agricultural sector consider servitization?
4. What is digital servitization?
5. How can agricultural companies adopt digital servitization?

✓ Answers

1. Servitization is the process of adding services to companies' core product offerings to create additional value for customers. It is a way for companies to differentiate from the competition, enhance customer loyalty, and (re)boost growth. In order to servitize successfully, companies should develop resources and skills to create, sell, and deliver services that customers want. Furthermore, they should be mindful of their business environment, which may enhance or disrupt servitization efforts.
2. Companies can offer many different types of services. There are two popular service categorizations: service focus (i.e., on the product/customer) and value proposition (i.e., an input/performance/result). Combining both categories creates six service types: product life cycle services (e.g., repair), product performance services (e.g., preventive maintenance), product result services (e.g., pay-per-use), process support services (e.g., training), process delegation services (e.g., outsourcing), and customer solutions (e.g., system integration).
3. The agricultural sector has evolved drastically over the last few decades, putting farmers under a lot of pressure. Servitization is a way to respond to the increasing demand for fresh, high-quality, and locally produced food, comply-

ing with environmental and animal welfare standards. Through servitization, farms can better focus on their focal activity by outsourcing specific activities, and experienced farmers can remain active longer by sharing knowledge with aspiring farmers and other companies. Also, it has been found that farm suppliers perform better through servitization, and that servitized suppliers improve farms' eco-efficiency.

4. Digital servitization is the use of digital technologies to enable companies to shift from a product-centric to a service-centric business model. Companies can implement IT in the back end to enable scalability in the production and delivery of customized solutions and/or in the front end to stay connected and continuously interact with customers. In the near future, big data, advanced data analytics, and AI will allow companies to move from merely monitoring individual products to controlling their functions, optimizing their performance, and providing fully autonomous systems.
5. On the one hand, farms can create efficiency gains by implementing technology in the back end, which enables them to take control over and integrate different activities of the (food) supply chain. They can also expand into new markets by collecting data that may be valuable to other companies and offering the data-as-a-service. On the other hand, farm suppliers can connect with farms by implementing technology in the front end. This enables them to offer services for farms that want to maintain control over their operations (e.g., custom-built hardware), that require support (e.g., sensors as digital assistants), and/or that want to completely outsource particular activities (e.g., crop protection).

References

- Baines, T., & Lightfoot, H. W. (2014). Servitization of the manufacturing firm: Exploring the operations practices and technologies that deliver advanced services. *International Journal of Operations & Production Management*, 34, 2–35. <https://doi.org/10.1108/IJOPM-02-2012-0086>
- Baines, T., Lightfoot, H. W., Evans, S., Neely, A., Greenough, R., Peppard, J., Roy, R., Shehab, E., Braganza, A., Tiwari, A., Alcock, J. R., Angus, J. P., Bastl, M., Cousens, A., Irving, P., Johnson, M., Kingston, J., Lockett, H., Martinez, V., Michele, P., Tranfield, D., Walton, I. M., & Wilson, H. (2007). State-of-the-art in product-service systems. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 221, 1543–1552. <https://doi.org/10.1243/09544054JEM858>
- Baines, T., Lightfoot, H. W., Peppard, J., Johnson, M., Tiwari, A., & Shehab, E. (2009a). Towards an operations strategy for product-centric servitization. *International Journal of Operations & Production Management*, 29, 494–519. <https://doi.org/10.1108/01443570910953603>
- Baines, T. S., Lightfoot, H. W., Benedettini, O., & Kay, J. M. (2009b). The servitization of manufacturing: A review of literature and reflection on future challenges. *Journal of Manufacturing Technology Management*, 20, 547–567. <https://doi.org/10.1108/17410380910960984>
- Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, 17, 99–120. <https://doi.org/10.1177/014920639101700108>
- Bigdeli, A. Z., Bustinza, O. F., Vendrell-Herrero, F., & Baines, T. (2017). Network positioning and risk perception in servitization: Evidence from the UK road transport industry. *International Journal of Production Research*, 0, 1–15. <https://doi.org/10.1080/00207543.2017.1341063>
- Carr, N.G. (2003). IT doesn't matter. *Harvard Business Review*.
- Chandler, A. D., Jr., & Hikino, T. (1994). *Scale and scope*. Harvard University Press.

- Coreynen, W., Matthyssens, P., De Rijck, R., & Dewit, I. (2018). Internal levers for servitization: How product-oriented manufacturers can upscale product-service systems. *International Journal of Production Research*, 56, 2184–2198. <https://doi.org/10.1080/00207543.2017.1343504>
- Coreynen, W., Matthyssens, P., & Van Bockhaven, W. (2017). Boosting servitization through digitization: Pathways and dynamic resource configurations for manufacturers. *Industrial Marketing Management*, 60, 42–53. <https://doi.org/10.1016/j.indmarman.2016.04.012>
- Corti, D., Granados, M. H., Macchi, M., & Canetta, L. (2013). *Service-oriented business models for agricultural machinery manufacturers: Looking forward to improving sustainability*. Presented at the 2013 International Conference on Engineering, Technology and Innovation (ICE) IEEE International Technology Management Conference (pp. 1–8). <https://doi.org/10.1109/ITMC.2013.7352612>.
- Crozet, M., & Milet, E. (2017). Should everybody be in services? The effect of servitization on manufacturing firm performance. *Journal of Economics and Management Strategy*, 26, 820–841. <https://doi.org/10.1111/jems.12211>
- Davies, A. (2004). Moving base into high-value integrated solutions: A value stream approach. *Industrial and Corporate Change*, 13, 727–756. <https://doi.org/10.1093/icc/dth029>
- Eggert, A., Hogreve, J., Ulaga, W., & Muenkhoff, E. (2014). Revenue and profit implications of industrial service strategies. *Journal of Service Research*, 17, 23–39. <https://doi.org/10.1177/1094670513485823>
- Eloranta, V., & Turunen, T. (2016). Platforms in service-driven manufacturing: Leveraging complexity by connecting, sharing, and integrating. *Industrial Marketing Management*, 55, 178–186. <https://doi.org/10.1016/j.indmarman.2015.10.003>
- Fang, E., Palmatier, R. W., & Steenkamp, J.-B. E. M. (2008). Effect of service transition strategies on firm value. *Journal of Marketing*, 72, 1–14. <https://doi.org/10.1509/jmkg.72.5.1>
- FAO, IFAD, UNICEF, WFP, & WHO. (2018). *The State of Food Security and Nutrition in the World 2018. Building climate resilience for food security and nutrition*. Food and Agricultural Organization of the United Nations.
- Finne, M., Brax, S., & Holmström, J. (2013). Reversed servitization paths: A case analysis of two manufacturers. *Service Business*, 7, 513–537. <https://doi.org/10.1007/s11628-013-0182-1>
- Fliess, S., & Lexutt, E. (2017). How to be successful with servitization—Guidelines for research and management. *Industrial Marketing Management*. <https://doi.org/10.1016/j.indmarman.2017.11.012>
- Ford, H., & Crowther, S. (1922). *My life and work*. Doubleday, Page.
- Gebauer, H. (2008). Identifying service strategies in product manufacturing companies by exploring environment–strategy configurations. *Industrial Marketing Management*, 37, 278–291. <https://doi.org/10.1016/j.indmarman.2007.05.018>
- Gebauer, H., & Fleisch, E. (2007). An investigation of the relationship between behavioral processes, motivation, investments in the service business and service revenue. *Industrial Marketing Management*, 36, 337–348. <https://doi.org/10.1016/j.indmarman.2005.09.005>
- Gebauer, H., Fleisch, E., & Friedli, T. (2005). Overcoming the service paradox in manufacturing companies. *European Management Journal*, 23, 14–26. <https://doi.org/10.1016/j.emj.2004.12.006>
- Hsu, C. (2007). Scaling with digital connection: Services innovation. In *2007 IEEE International Conference on Systems, Man and Cybernetics* (Vols. 1–8, pp. 4057–4061). IEEE.
- Huang, M.-H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, 21, 155–172. <https://doi.org/10.1177/1094670517752459>
- Kamp, B., & Parry, G. (2017). Servitization and advanced business services as levers for competitiveness. *Industrial Marketing Management*, 60, 11–16. <https://doi.org/10.1016/j.indmarman.2016.12.008>
- Kindström, D., Kowalkowski, C., & Sandberg, E. (2013). Enabling service innovation: A dynamic capabilities approach. *Journal of Business Research*, 66, 1063–1073. <https://doi.org/10.1016/j.jbusres.2012.03.003>
- Koba, M., 2013. *A hungry world: Lots of food, in too few places* [WWW Document]. Retrieved August 21, 2019, from <https://www.cnbc.com/id/100893540>

- Koebler, J., & Wanstreet, R. (2018). America's farmers are becoming prisoners to agriculture's technological revolution. *Motherboard*. Retrieved April 15, 2019, from https://motherboard.vice.com/en_us/article/a34pp4/john-deere-tractor-hacking-big-data-surveillance
- Kowalkowski, C., Gebauer, H., Kamp, B., & Parry, G. (2017). Servitization and deservitization: Overview, concepts, and definitions. *Industrial Marketing Management*, 60, 4–10. <https://doi.org/10.1016/j.indmarman.2016.12.007>
- Kowalkowski, C., & Kindström, D. (2014). Service innovation in product-centric firms: A multidimensional business model perspective. *The Journal of Business and Industrial Marketing*, 29, 96–111. <https://doi.org/10.1108/JBIM-08-2013-0165>
- Laudien, S. M., & Daxböck, B. (2016). The influence of the industrial internet of things on business model design: A qualitative-empirical analysis. *International Journal of Innovation Management*, 20, 1640014. <https://doi.org/10.1142/S1363919616400144>
- Lee, S., & Cappellazzi, M. (2015). Under contract: Farmers and the fine print.
- Lichtenberg, E. (2019). Conservation and the environment in US farm legislation. *EuroChoices*, 18, 49–55. <https://doi.org/10.1111/1746-692X.12214>
- Link Magazine. (2018). *Pionier Atlas Copco blijft in servitization investeren* [WWW Document]. Retrieved December 31, 2018, from <https://www.linkmagazine.nl/pionier-atlas-copco-blijft-in-servitization-investeren/>
- Matheny, G., & Leahy, C. (2007). Farm-animal welfare, legislation, and trade. *Law and Contemporary Problems*, 70, 325–358.
- Matthyssens, P., & Vandenbempt, K. (2008). Moving from basic offerings to value-added solutions: Strategies, barriers and alignment. *Industrial Marketing Management*, 37, 316–328. <https://doi.org/10.1016/j.indmarman.2007.07.008>
- Oliva, R., & Kallenberg, R. (2003). Managing the transition from products to services. *International Journal of Service Industry Management*, 14, 160–172. <https://doi.org/10.1108/09564230310474138>
- Pagoropoulos, A., Maier, A., & McAloone, T. C. (2017). Assessing transformational change from institutionalising digital capabilities on implementation and development of product-service systems: Learnings from the maritime industry. *Journal of Cleaner Production*, 166, 369–380. <https://doi.org/10.1016/j.jclepro.2017.08.019>
- Pereira, Á., Carballo-Penela, A., González-López, M., & Vence, X. (2016). A case study of servicizing in the farming-livestock sector: Organisational change and potential environmental improvement. *Journal of Cleaner Production*, 124, 84–93. <https://doi.org/10.1016/j.jclepro.2016.02.127>
- Pereira, Á., Carballo-Penela, A., Guerra, A., & Vence, X. (2018). Designing a policy package for the promotion of servicising: A case study of vineyard crop protection in Galicia (Spain). *Journal of Environmental Planning and Management*, 61, 348–369. <https://doi.org/10.1080/09640568.2017.1308317>
- Porter, M. E., & Heppelmann, J. E. (2014). How smart, connected products are transforming competition. *Harvard Business Review*, 92, 64–88.
- Raddats, C., Kowalkowski, C., Benedettini, O., Burton, J., & Gebauer, H. (2019). Servitization: A contemporary thematic review of four major research streams. *Industrial Marketing Management*. <https://doi.org/10.1016/j.indmarman.2019.03.015>
- Rust, R. T., & Huang, M.-H. (2014). The service revolution and the transformation of marketing science. *Marketing Science*, 33, 206–221. <https://doi.org/10.1287/mksc.2013.0836>
- Santos, M. Y., Oliveira e Sá, J., Andrade, C., Vale Lima, F., Costa, E., Costa, C., Martinho, B., & Galvão, J. (2017). A Big Data system supporting Bosch Braga Industry 4.0 strategy. *International Journal of Information Management*, 37, 750–760. <https://doi.org/10.1016/j.ijinfomgt.2017.07.012>
- Schmenner, R. W. (2009). Manufacturing, service, and their integration: Some history and theory. *International Journal of Operations & Production Management*, 29, 431–443. <https://doi.org/10.1108/01443570910953577>
- Sklyar, A., Kowalkowski, C., Tronvoll, B., & Sörhammar, D. (2019). Organizing for digital servitization: A service ecosystem perspective. *Journal of Business Research*. <https://doi.org/10.1016/j.jbusres.2019.02.012>
- Storbacka, K. (2018). Extending service-dominant logic—Outside marketing and inside managerial practice. In *The SAGE handbook of service-dominant logic* (p. 1237). SAGE.

- Storbacka, K., Windahl, C., Nenonen, S., & Salonen, A. (2013). Solution business models: Transformation along four continua. *Industrial Marketing Management*, 42, 705–716. <https://doi.org/10.1016/j.indmarman.2013.05.008>
- The Economist. (2018). What will be the biggest stories of 2019? | Part One | *The Economist*.
- Thurlbeck, W. M. (2012). Perestroika, fashion, and the universal glue. *The American Review of Respiratory Disease*. <https://doi.org/10.1164/ajrccm/139.5.1280>
- Tukker, A. (2004). Eight types of product–service system: Eight ways to sustainability? Experiences from SusProNet. *Business Strategy and the Environment*, 13, 246–260. <https://doi.org/10.1002/bse.414>
- Ulaga, W., & Reinartz, W. J. (2011). Hybrid offerings: How manufacturing firms combine goods and services successfully. *Journal of Marketing*, 75, 5–23. <https://doi.org/10.1509/jmkg.75.6.5>
- Valtakoski, A. (2017). Explaining servitization failure and deservitization: A knowledge-based perspective. *Industrial Marketing Management*, 60, 138–150. <https://doi.org/10.1016/j.indmarman.2016.04.009>
- Vandermerwe, S., & Rada, J. (1988). Servitization of business: Adding value by adding services. *European Management Journal*, 6, 314–324. [https://doi.org/10.1016/0263-2373\(88\)90033-3](https://doi.org/10.1016/0263-2373(88)90033-3)
- Vidickiene, D., & Gedminaitė-Raudonė, Z. (2018). Challenges for agricultural policy in the service-driven economic system. *Ekonomika Poljoprivrede (Economics of Agriculture)*, 65, 1545–1555. <https://doi.org/10.5937/ekoPolj1804545V>
- Vidickiene, D., & Gedminaitė-Raudonė, Z. (2019). Servitization as a tool to increase vitality of ageing rural community. *European Countryside*, 11, 85–97. <https://doi.org/10.2478/euco-2019-0006>
- Vignali, C., Guthrie, J., Guthrie, A., Lawson, R., & Cameron, A. (2006). Farmers' markets: The small business counter-revolution in food production and retailing. *British Food Journal*. <https://doi.org/10.1108/00070700610676370>
- Wang, J., Han, W., & Jia, G. (2014). *Servitization based on information system: The case of Foton Lovol*. Presented at the 2014 11th International Conference on Service Systems and Service Management (ICSSSM) (pp. 1–5). <https://doi.org/10.1109/ICSSSM.2014.6943403>.
- Welch, J., & Byrne, J. A. (2003). *Jack: Straight from the gut*. Hachette UK.



Entrepreneurial Finance

Anne Lafarre and Ivona Schoonbrood

Contents

- 15.1 Introduction – 355**
- 15.2 Pre-seed Financing and Support – 357**
 - 15.2.1 Family, Friends, and Fools – 357
 - 15.2.2 Accelerators, Incubators, and Startup Studios – 357
- 15.3 Early Sources of Funding (Seed and Startup Stage) – 361**
 - 15.3.1 Business Angels – 361
 - 15.3.2 Crowdfunding – 362
 - 15.3.3 Initial Coin Offerings (ICOs) – 363
- 15.4 Venture Capital and Private Equity (Growth Stage) (Da Rin & Hellmann, 2019) – 364**
 - 15.4.1 Ownership and Valuation – 365
 - 15.4.2 Preferred Shares – 367
 - 15.4.3 Staged Financing – 368
 - 15.4.4 Corporate Governance – 370
 - 15.4.5 Exit Routes – 372

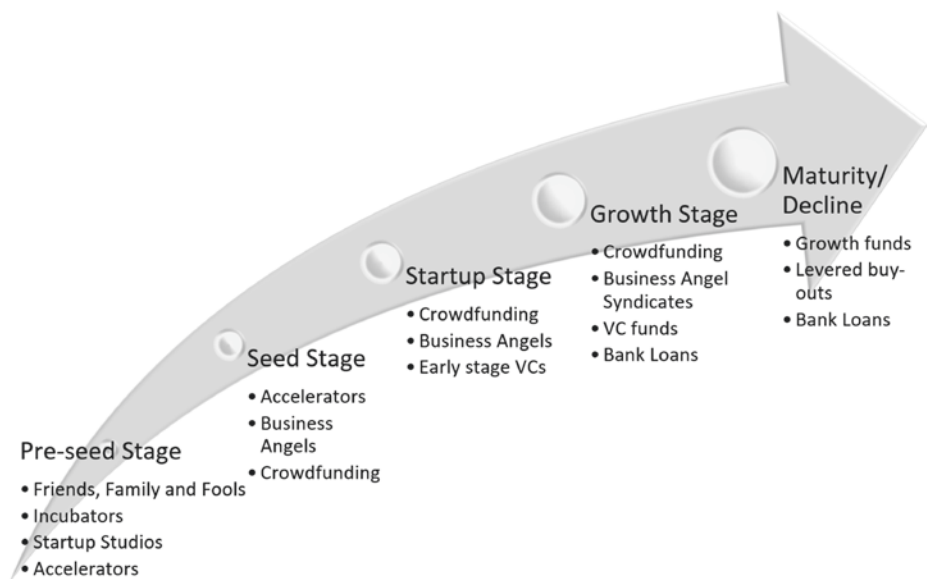
- 15.5 Tech Startup Financing in Practice – 372**
- 15.6 Answers to the Cases – 376**
 - References – 380**

Learning Objectives

- Identify the different types of investors, including the added value and potential problems that these investors bring.
- Evaluate features of different financing options at each stage of the innovative (tech) startup's life cycle.
- Understand the financial structure of a startup company, including the differences between debt and equity financing and the particular features of preferred shares, and the VC startup valuation technique.
- Know the most common financing methods chosen by innovative (tech) startups in the USA and Europe.

15.1 Introduction

This chapter provides an introduction to entrepreneurial finance, explaining the fundamental challenges entrepreneurs face when raising external capital. Obtaining financing is vital for a startup, and we consider the types of investors in the different stages of the startup. During its life cycle, a venture goes through different stages that are marked by particular business-related milestones. In this chapter, we distinguish among **five different stages** of a venture's evolution (■ Fig. 15.1). The first stage is a **pre-seed stage**, which covers the very inception



■ Fig. 15.1 The five stages of a venture's evolution. (Note to figure: Author's own figure)

of the business idea and initial steps towards developing an operational business model. In the second stage (the **seed stage**), founders of the ventures focus on validating a product/service, reformulating a business model, and acquiring first customers. The third **startup stage** is the period when ventures still improve their products/services, continuously grow their customer base, and get closer to the break-even point, a financial milestone that marks the moment when income of ventures can cover all their costs. In the fourth **growth stage**, ventures experience continuous growth and expand their activities by introducing new business lines or through an entry to new markets. When ventures reach the **maturity stage**, their growth slows down, while their existing business lines still provide rather stable revenue. This comfortable market position may sometimes prevent them from exploring riskier and potentially innovative business lines and thus lead to stagnation or even decline. Due to that, their business models become much more prone to disruption by long-term market competitors, which follow faster innovation product cycles or new market entrants, which bring superior and/or cheaper products or services.

As described above, every stage of a venture's evolution requires a firm to employ rather different strategies and utilize different assets. It is very difficult to determine the duration of each stage, as it very much depends on the type of a venture, access to capital, market conditions, and geographical location. For instance, pre-seed stage can take anywhere from several weeks for a software-based SaaS solution that can be developed quite quickly without substantial material costs to 5 years for a MedTech company that has to significantly invest into R&D efforts and obtaining of formal approvals (for instance for pharmaceuticals or medical devices) even before bringing a prototype to the market. In the context of entrepreneurial finance, recognizing an evolutionary stage of a particular venture is quintessential to determine which type(s) of investors may be suitable and most likely to invest required funds. Startup financing is however not limited to providing capital. Startup investors (depending on the type) fulfill also myriad of other functions such as (1) mentoring, (2) strategic advising, (3) resolution of conflicts among startup founders, (4) connecting startups to their network, (5) recruiting human capital, and (6) supporting ventures in subsequent rounds of financing.

Throughout this chapter, we will use a particular case study to make some important aspects of entrepreneurial finance more comprehensible and we include related case questions: **EnvironTECH**. The founders of EnvironTECH are Ana and Peter, who met during their master's in artificial intelligence. EnvironTECH is a promising startup that aims at building digital copies of physical environments, where artificial intelligence models can be used to help understand the parameters of that environment and provide valuable feedback. This AI is especially useful for construction companies that need to map the area of their construction sides. EnvironTECH was started by Peter and Ana during their master's studies at the Jheronimus Academy of Data Science (JADS) in the Netherlands. A bit later, Jan who has a background in business and finance joins the team.

15.2 Pre-seed Financing and Support

15.2.1 Family, Friends, and Fools

The very first people willing to invest capital in the very early stage of venture's life cycle are not investment professionals, but rather individuals that are affiliated to the founders through personal or family ties. Although they often provide the initial capital needed to establish a company or develop a minimum viable product, they rarely can fulfill other roles due to lack of the entrepreneurial or industry-specific experience. Moreover, their personal ties to founders prevent them from providing any relevant product validation, even if only from the perspective of target customers. Therefore, after raising money with **3Fs**, as they are colloquially known, founders should very quickly move towards other organizations and investors that can examine the product on objective merits and provide the valuable advice and capital.

Ana and Peter are currently in their pre-seed stage working on the original business idea and creating technology that will be the basis of their groundbreaking product. They can try to raise the very initial funds from their family members of close circle of friends, because it can help them bridge the very basic financial needs of a venture, while they are still working on the prototype. Nevertheless, they should not forget that mixing up personal and business relationships may bring additional complexities and dilemmas.

15.2.2 Accelerators, Incubators, and Startup Studios

In the past decade, business stakeholders and policymakers started to provide their support for startups in a much more intensive way than before, creating essential ingredients for building sustainable and productive startup ecosystems. Besides venture capital firm, which emerged roughly in the 1980s, business stakeholders recognized that startups need significantly greater supports in the most vulnerable stages of their existence, in the pre-seed and seed stages. This gave rise to new types of intermediaries that put emphasis on developing industry-specific and business skills of venture founders that give them much greater chance to survive the early years of their existence and materialize their high growth potential. The most important intermediaries are incubators, accelerators, and startup studios, which are discussed below. Usually, these intermediaries also provide a small amount of seed funding, but sometimes only in-kind services are provided.

15.2.2.1 Incubators

Incubators are organizations that provide startups with various resources in the very beginning of their entrepreneurial journey. They usually offer founders a physical location; an office or co-working space; access to network of mentors; various service providers such as law firms and tax or business consultancies; or

connections to big corporates that may become their customers or strategic partners. Incubators often organize a number of workshop sessions on different topics relevant for building a successful business. Another essential advantage of incubators is a peer learning that occurs naturally as founders share the physical space with other entrepreneurs going through the similar stage of their entrepreneurial journey. Some incubators have a very structured and time-limited programs, where participation of venture founders in scheduled activities and workshops is compulsory, and others provide their services on voluntary and ad hoc basis. The group activities in incubators may include general business knowledge workshops; legal, tax, or business seminars; and meetings with experienced entrepreneurs that walk the incubator participants through their success stories. The individual activities are focused on matching founders with suitable mentors, who can provide invaluable and industry-specific advice.

Due to these features of their program, they are generally considered to be the initial springboards of innovative startups. They usually do not invest into ventures directly, nor do they acquire shares of companies that take part in their programs. The added value of incubators largely depends on the quality of services that they provide, for instance on the expertise and connections of their mentors or quality of business-related seminars that they organize. One major difficulty in setting up an incubator program is its lack of business viability. Since incubators usually do not take equity and provide their services for free or small fee, they need to be supported by parties that are strategically interested in startups or parties that do not expect return on their investment of capital or time in the incubation program. Therefore, incubators are often set up or at least financially supported by universities, corporates, or public authorities that view the innovation or development of startup ecosystems as their primary goals (Lasrado et al. 2015). Furthermore, incubation programs typically do not provide any remuneration to mentors or service providers that support startups with their knowledge and network connections. Acquiring the most skilled mentors and service providers may thus be rather challenging. Another concern of the existing incubation model is related to disproportional shielding of the entrepreneurs. Incubators often provide very supportive and collegial environment that may inadequately shield entrepreneurs from often brutal market conditions and extend the life span of ventures that are destined to fail. Whether highly efficient or not, incubators provide a relatively safe space to explore various business ideas, especially for novice entrepreneurs that may be otherwise not incentivized to pursue entrepreneurial path. Ana and Peter may choose to take part in an incubator, because it can help them secure certain resources such as office space, connection to mentors, and some useful workshops related to the business skills they currently lack.

15

15.2.2.2 Accelerators

As evident from their name, the main objective of **accelerators** is to fast-forward the development of early-stage ventures to the point, where they are either (1) ready to put the product on the market or (2) ready to accept external financing or both. In comparison with incubators, accelerators standardly invest small amounts of capital in exchange for a percentage of equity (5–15%) and provide startups

with time-limited, highly structured, and intensive program of activities. Accelerators accept startups to their programs in group, e.g., in so-called batches or cohorts in order to provide startups with a peer learning environment. Startups are usually offered a combination of group activities, where founders have a chance to explore general business topics and individual company-specific mentoring from various mentors affiliated to the acceleration program. The program itself focuses on three essential goals: a product refinement and validation, creation of a viable business model, and preparation for promoting a venture to external investors (pitching). Startups rarely enter an acceleration program only with an idea “on the paper.” They often possess a prototype or minimum viable product that may need further modifications and improvements. Since their time in an accelerator is limited to several months, ventures cannot realistically change the product in a drastic manner, and there is simply no time for it. On the other hand, an accelerator will usually facilitate product validation in the target group of customers. If an outcome of such validation is not positive, venture founders may use the acquired knowledge to “pivot,” e.g., to modify a product in order to better address the needs of target customers.

Besides product focus, acceleration programs provide venture founders with essential business knowledge. Workshop related to business model building, writing of a business and financial plan, growth and go-to market strategies, and legal compliance often form integral parts of the program’s curriculum. Last but not least, accelerators thoroughly prepare venture founders for pitching, e.g., presenting their business idea and viability to external investors verbally in a short presentation and “on paper” (so-called pitch deck). Although this may seem as a marginal issue, many entrepreneurs fail to catch the attention of potential investors because they are unable to articulate their business proposition and properly explain the functioning of their product. Since startup investors standardly review hundreds if not thousands of proposals per year, they tend to immediately pass on products they do not properly understand. An effective pitching is therefore an essential ingredient in attracting suitable investors.

As mentioned before, accelerators invest small amounts of capital in exchange for a minority equity in a startup accepted to their program. The invested capital usually ranges between 10,000 and 50,000 EUR, for which accelerators ask from 5% to 15% of share ownership (equity). Sometimes, **convertible notes** are used, which is an easy-to-negotiate financial debt instrument that can be automatically converted to equity in a later professional funding round, for instance with a VC. Usually, there is a discount on the share price for the conversion, or a valuation cap is used, to provide the convertible note holder with the highest number of shares after conversion. Besides that, accelerators rarely ask for any special control rights or seat on a board of a company. Nevertheless, this investment has two important implications. Accelerators are oftentimes first formal external investors that appear on the capitalization tables of the ventures. Secondly, accelerators offer the same amount of money for the identical percentage to all their participants, which means that the first valuation of a venture will not be based on their specific features and future potential but is rather determined by a standardized offer of an accelerator.

In contrast with incubators, accelerators have their own investors that have several incentives to invest in acceleration programs. Venture capital firms may invest in order to acquire an exclusive access to most promising startups that can later on join their own startup portfolio. Big corporations may be interested in the startup-generated innovation that they can later acquire and integrate in their own product lines. Public authorities may invest in order to strengthen the local or national startup ecosystem.

Case 2.1: EnvironTECH

At the moment, the startup venture EnvironTECH is in its very early stage. After a period of bootstrapping, including some first few months in which they used their own savings and did not pay themselves any salary, Peter and Ana consider themselves ready to tap into external financing.

Peter and Ana were approached by representatives of an AI accelerator located in the Netherlands. This acceleration program is quite well known and reputable. The accelerator offers every startup a €100,000 convertible note investment, which includes €20,000 in cash and €80,000 in-kind funding consisting of, among others, office space and other facilities, expert workshops, international events, and intensive hands-on (mentor) support.

? Questions

What is a convertible note and why is it often used in early-stage financing? Based on the provided information, would you advise Peter and Ana to consider this financing option?

15.2.2.3 Startup Studios: Venture Builders

Startup studios or **venture builders** are organizations that recruit promising entrepreneurs and scientists, create a viable and dynamic startup team, and match them with a technology or a business idea, which is ready for a market adoption. In contrast to incubators or accelerators, startup studios are much more involved in building of the initial team and exploring and experimenting with different applications of a technology at hand. For instance, the venture builder HighTechXL in Eindhoven actively cooperates with CERN (HighTechXL, 2019). This world-known particle research center provided HighTechXL with an access to their unique technologies that were further developed and applied in products of a number of startups. Due to high amount of efforts that startup studios put into their ventures, they can usually facilitate much lower number of startups than regular acceleration programs. Moreover, they tend to take much larger percentage of equity (15–25%); therefore, in terms of share ownership, they are often perceived to have an equal role to a co-founder. Startup studios are perfect for starting entrepreneurs without experience and a particular business idea in mind or more expe-

rienced entrepreneurs that are looking for a next great project. Since Ana and Peter have already developed a viable business idea and they seem to possess necessary (technical) skills to execute it, a startup studio may not be a good match for them.

15.3 Early Sources of Funding (Seed and Startup Stage)

15.3.1 Business Angels

Business angels are wealthy private individuals that invest their own capital into innovative ventures. In contrast to venture capital firms, they do not pool the funds from institutional investors and thus do not have to be accountable to anybody for their investment decisions and monitoring actions. They are usually former entrepreneurs or corporate executives who beyond financial gain look to share the acquired knowledge and experience. At the same time, the connection to early-stage ventures provides them with an access to the cutting-edge products in their respective fields. Business angels tend to invest in the technology areas that correspond with their professional backgrounds. Furthermore, they often select companies, which are geographically in their vicinity, because they rely on personal contact with the venture team. Their main objective is to become an active investor with a close connection to the founders. Therefore, their portfolios are rather small, between 4 and 6 companies at once.

After incubators, accelerators, or startup studios (that not all early-stage ventures take part in), business angels are the first investors that provide external capital to startups. They bear significant risk, as they invest quite early on in the venture's life cycle. Like other seed funders, angel investors often use convertible notes that are automatically converted into equity in a later stage funding round. The investment amount of an average business angel per deal can range from €10,000 to €250,000, but some so-called super angels may invest even more. The quality of the mentoring, strategic advice, and networking always depends on the angels' experience and their fit with a startup.

An average angel investor is typically a successful (male) entrepreneur or corporate executive in the age between 45 and 65. Angels are sufficiently wealthy, but contrary to common belief, they do not tend to be superrich. Currently, there are several initiatives to attract other individuals into business angel investing and create more diversity in the angel landscape. For instance, women angels in Scandinavian countries founded the Nordic Female Business Angel Network that specifically focuses on connecting, supporting, and recruiting female business angels (NFBAN, 2019).

Besides diversity, the abovementioned examples demonstrate also another development in the business angel investing, e.g., the emergence of business angel groups and syndicates. While business angels standardly invest individually and alone, pooling the investment capacity of a number of business angels

has been significantly impacting the angel investment landscape. Firstly, business angel syndicates can better streamline and standardize the process of investing. Combining the investment amounts enables them to provide larger rounds and also follow-up investments that are not standard in angel investing. Moreover, especially in Europe, syndicating enables angels to invest across borders.

15.3.2 Crowdfunding

Crowdfunding was firstly used as an alternative financing model in art and music industries, where fans of particular artists/musicians helped to fund the recording of albums. Since then, crowdfunding has evolved into several types of financing with vastly different implications for ventures and their supporters. Mollick defines crowdfunding “as an effort by entrepreneurial individuals and groups—cultural, social, and for-profit to fund their ventures by drawing on relatively small contributions from a relatively large number of individuals using the internet” (Mollick, 2014). Crowdfunding platform plays the role of an intermediary agent, who matches the demand side of financing, e.g., campaigners and supply side of financing, e.g., crowdfunders. In principle, one can distinguish between two main categories of crowdfunding, financial and nonfinancial. Nonfinancial types of crowdfunding do not provide crowdfunders with any financial return or financial proceeds on their pledged funds and thus represent one-off transactions between crowdfunders and campaign owners. These include donation-based model, where funders of a campaign essentially donate money for no consideration at all, and reward-based crowdfunding, where funders receive a symbolic reward in return. Pre-order crowdfunding on the other hand provides an opportunity for the public to pre-purchase a product, which is at the time of the campaign still in the process of development. Nonfinancial crowdfunding is from a legal and financial perspective not too complex. Besides potential misuse of raised funds and fraud, it does not pose significant legal challenges. Pre-order crowdfunding can however put a significant production pressure on a venture, since an early-stage startup may not be fully equipped to accommodate orders from larger amount of crowdfunders at once. The financial types of crowdfunding include **loan-based crowdfunding**, where aggregated contributions of crowdfunders are provided to a company in the form of a repayable loan with fixed interest and **equity crowdfunding** (often called also **crowdinvesting**), where crowdfunders through various schemes invest into equity, e.g., shares of a company. The financial types of crowdfunding are significantly more complex and as opposed to donation- or reward-based crowdfunding create rather long-term commercial relations among involved parties. Besides providing the capital, crowdfunding may provide some evidence of product validation and crowdfunders often act as product supporters and ambassadors, which further increases the visibility of a product.

From the entrepreneurial finance perspective, loan-based and equity crowdfunding represents schemes that are qualified as an investment. In contrast to other types of venture funding, financial types of crowdfunding pool funds from professional and nonprofessional investors and therefore enable also broader public to

participate in the venture financing. One significant disadvantage of financial crowdfunding is that investors are numerous and dispersed and therefore may not be willing or capable to fulfill other important investor roles. Hence, crowdfunded ventures usually do not have the opportunity to use the network of their investors or get a strategic advice from their investors.

Case 3.1: EnvironTECH (Continued)

EnvironTECH BV is developing, and things are going great: Peter and Ana manage to have the very first prototype of their AI that is able to digitally map the garden of the mother of Peter (which is about 20 m²). Using this first result, Peter and Ana want to pitch their business to some interested angel investors. More specifically, they estimated that, in order to take their company to the next level in 2020 and be able to map a larger area and thus develop a minimum viable product for a few interested construction companies, they need an investment of €500,000.

During a network event, a friendly entrepreneur tells Peter and Ana that their interesting prototype will not be sufficient to convince an angel investor to invest in their company. Peter and Ana will also need to spend some time developing a financial plan according to this entrepreneur. Neither Ana nor Peter has a background in finance, and they ask their friend Jan who is currently doing his MBA to help them with this financial plan. Jan proves to be a great addition to the team, and with his help, the entrepreneurs are able to identify their concrete financing needs. Yet, Jan also heard about equity crowdfunding as a rather novel method of startup financing that is faster and more standardized. He convinces Ana and Peter to look into this brand-new financing method and consider running an equity crowdfunding campaign.

? Question

Would you advise the entrepreneurs to pursue equity crowdfunding to finance their startup funding needs?

15.3.3 Initial Coin Offerings (ICOs)

Initial coin offerings emerged only recently, as an alternative financing method of blockchain-based ventures. Simply put, an initial coin offering is an online call for purchase of digital cryptographic assets called “tokens” to a wider public. The sale is conducted through smart contract transactions that facilitate an exchange of widely used cryptocurrencies such as Bitcoin, Ether, Ripple, or Litecoin for a pre-determined number of tokens. Tokens may be defined as digital representations of certain rights bestowed upon the token holder. They usually carry quite a wide variety of rights ranging from an access to a platform providing specific service (utility tokens), through digital representation of a real-life asset (asset-backed tokens), to a security-like investment with profit expectations (investment tokens). Tokens are usually not defined by law, and they often do not fit definitions of shares, bonds, and derivatives of other financial instruments. Therefore, regulators

have been struggling with how to approach these types of offerings, which in extreme cases can raise several hundred million euros worth of cryptocurrencies.

The main features of ICOs are summarized below:

1. ICO initiators launch an ICO online call and publish the so-called white paper, a document providing basic information about the ICO project.
2. Token buyers transfer selected cryptocurrencies or fiat currencies to the wallet or an account of ICO initiators.
3. Token buyers receive in return digital assets called tokens that carry a bundle of rights of a financial or utility value (sometimes combined).
4. Transactions between ICO initiator and contributors are executed through a smart contract and recorded on the blockchain (usually Ethereum).
5. Tokens may become tradable on the secondary cryptocurrency exchanges, some of which have quite high liquidity.
6. ICOs are usually conducted without any intermediary comparable to a crowdfunding platform in crowdfunding.

From a pragmatic entrepreneurs' perspective, ICOs represent a very beneficial development on the landscape of alternative finance. They are fast and affordable, they raise amounts comparable to later rounds of VC financing or even initial public offerings, and they do not dilute the equity or reduce the control of the founders. Investors, retail or professional, may also be attracted to a new booming market that is easily accessible and provides a possibility of trading and thus instant exits.

So far, ICOs have been conducted only by blockchain-based businesses that rely on larger crypto community for support and funding. Since EnvironTECH BV does not utilize blockchain technology, ICO is most likely not a suitable funding method for Ana and Peter.

15.4 Venture Capital and Private Equity (Growth Stage)

(Da Rin & Hellmann, 2019)

Probably the most well-known startup investors are venture capitalists (often called VCs). These are professional investors who invest on behalf of other investors like institutional investors,¹ and particularly in the technology industry, there are now-

15

1 Usually, a limited partnership (LP) is established with these institutional investors using a limited partnership agreement (LPA) that determines the investment strategy of a particular VC fund. The VC is the general partner that manages the fund, while the institutional investors become the limited partners that do not engage in managing the fund but are protected from liability claims. A VC fund is often established for a predetermined period (of about 10 years), after which the LP will be terminated and the returns are paid to the investors. Since institutional investors diversify their portfolios by investing in multiple VC funds (and other debt and equity instruments in financial markets), they usually prefer a focused investment strategy. However, the VC (i.e., the general partner) may wish to diversify its investment strategy to diminish the risk sensitivity of the VC fund to a specific industry. Hence, whereas we will see that the main agency problems arise between VCs and entrepreneurs in this chapter, it is important to realize that VCs also have their diverging incentives that they have to deal with at the level of the fund.

adays mega investment rounds. Particularly, in the beginning of 2019, there was a record in the number of funding rounds totaling 100 million USD or more in the USA (Kruppa, 2019). For instance, after raising 535 million USD from SoftBank's Vision Fund in March 2018, after raising a down round a few years ago, the food delivery company DoorDash reached a valuation of 1.4 billion USD, and, only a few months later, 4 billion USD. Next, only 1 year later, with new rounds of funding of 400 million USD and 600 million USD from others Darsana Capital and Sands Capital, the company was valued at 12 billion USD in May 2019 (Bradshaw, 2019). In this section of the chapter, we outline the different aspects of venture capital and private equity funding for startups.

In the next section, we discuss (VC) ownership and valuation of startups (► Sect. 15.4.1), financial instruments they use (► Sect. 15.4.2), staged financing and anti-dilution protection mechanisms (► Sect. 15.4.3), and control of the entrepreneurial firm (► Sect. 15.4.4).

15.4.1 Ownership and Valuation

There are strong relationships between the investment, the investor ownership, and the valuation in entrepreneurial finance. Since the valuation of startups is even more uncertain and difficult compared to more mature companies, (Welch 2022) VCs usually use quite a simple method that matches the valuation of a startup with their expected exit values at the end of their investment period. This so-called VC method starts from the required rate of return: since many startups fail, but the VC fund needs to meet the expectations regarding the fund's rate of return of its institutional investors, this required rate of return is very high compared to other investments. This required rate of return adds to the normal investment return (which is usually the sum of the risk-free rate, the risk premium, and, in case of a large investment, an illiquidity premium) a premium for the large failure rate of startups and a premium that accounts for the services and advice these VCs provide to startups. Hence, the required rate of return can be about 40%, and sometimes even 60%, depending *inter alia* on the industry.

When the required rate of return is defined, the VC determines the expected exit value of the startup it wants to invest in. With this expected exit value and the time it takes to exit, it can calculate the ownership stake it wants in return for its investment. Let us see this with an example: suppose a VC has a required return rate of 50% and wants to invest €2 million in a startup that has an expected exit value of €50 million in 5 years. What would be the required ownership stake of this VC in this startup? With a required rate of return of 50%, the exit value should be €15.2 million in 5 years for this VC ($2 \text{ million} \times (1.5)^5 = 15.2 \text{ million}$), (Welch 2022) and €15.2 million is about 30.4% of €50 million. Hence, given its expectations, the VC wants to invest €2 million in this startup in return for an ownership stake of let us say 30%.

Next, we need to distinguish two valuation terms: the **pre-money valuation** and the **post-money valuation**. From the aforementioned example, the post-money valuation can be calculated in a simple way: if the VC wants to invest €2 million in return for a 30% ownership stake, the total value of the company should be 2 mil-

lion/30% = €6.7 million. This is also called the post-money valuation, i.e., the value of the startup company right *after* the investment of the VC. The pre-money valuation is the value of the firm right *before* the investment: in this case, the investment is €2 million, and thus the pre-money valuation should be €6.7 million – €2 million = €4.7 million. We can formalize these calculations using the following formulas, where V_{before} denotes the pre-money valuation, V_{after} the post-money valuation, I_i the investment of VC i , s_i the ownership stake of VC i after the investment, and P_s the share price:

Pre-money valuation:	$V_{before} = V_{after} - I_i$
Post-money valuation	$V_{after} = V_{before} + I_i$
	$V_{after} = I/s_i$
Number of shares pre-money:	$S_{before} = V_{before}/P_s$
Number of shares post-money:	$S_{after} = V_{after}/P_s$

For the first professional investment round, the amount of shares that is in the company and the price attached to these shares do not really matter. For instance, if a VC wants to invest €500,000 in your company, and would get an ownership stake of 20% in return, the post-money valuation is €2.5 million and the pre-money valuation is €2 million. Whether these amounts are divided by shares of €1 or €10 each does not really matter: if the share price is €1, the VC will receive 500,000 shares and the parties that were involved in the startup before this VC entered—including the entrepreneurs—own the remaining 2 million shares; if the share price is €10, the VC receives 50,000 shares and the others 200,000 shares. One may note that in both situations, the value of their investment stake remains the same. However, once a next professional financing round starts, the share price actually would matter as discussed in ► Sect. 15.4.3.

Most startups use share option plans for employees to attract and keep talented people. Since startups, in particular in the early stages, cannot pay these employees high salaries like mature corporations, share options provide a useful addition to normal pay. Employees receive share options from the company for a very low price, usually at par (the nominal value of a share). The share options can be exercised by the employee at a particular date in the future (or, if no particular date has been set, after a certain period, this is called *vesting*). If an employee exercises the share option, it in effect exercises the right to buy the shares of the company against the nominal value of the share (or another very low price). The employee gains because when the company grows, it becomes more valuable, and hence, the share price increases. However, these shares are usually not tradable yet, and therefore, a company loan can be used to enable employees to buy the startup shares. Usually, VCs require startups to set up an employee share option pool (usually called ESOP)—in case they did not set up one yet. These stock option pools are part of the pre-money valuation.

Case 4.1: EnvironTECH

Things are still going great for the entrepreneurs: only after a few years, Peter and Ana manage to develop their AI and launch a minimum viable product, which attracts the interest of several VCs. Eventually, Peter and Ana decide to negotiate a deal with a VC that is famous for investing in creative AI startups. They close a deal for an investment of €3,000,000 in return for an ownership stake of 20%. The VC, called CreativeVC, also suggests creating an employee stock option pool (“ESOP”) that includes 10% of the shares. The share price is set at €15 per share. Peter and Ana each receive twice as much shares as Jan, who joined the startup in a later phase.

? Question

What is the ownership structure of EnvironTECH after this investment round with CreativeVC (do not take into account the AI accelerator)?

15.4.2 Preferred Shares

Vcs are professional investors that are quite experienced with negotiating deals with startup companies. Before parties enter into a contract, term sheets are used that outline the terms of the investment deal. Term sheets are documents that are not legally enforceable (except for some of the clauses including confidentiality terms), but form the basis of the shareholder agreements (and the corporate charter, or the articles of association, and the bylaws) that are drafted after the deal is accepted by all parties. However, whereas these term sheets are not binding, there are reputational motives that drive party commitment to these negotiated terms.

In these term sheets, the investment amount and pre-money valuation are specified, just like the share price of the particular investment round. In addition, Vcs usually specify the special capital and control rights attached to their shares. In finance, there are different share classes. The plain vanilla share type is the common share, which carries one vote per share. Other share classes include preferred shares and cumulative preferred shares. These shares grant shareholders a priority claim over the startup’s proceedings upon a liquidation event. To understand these types of shares better and why VC investors often use this financial instrument, first let us consider the differences between **debt** and **equity**. Whereas debt has a fixed claim—i.e., a creditor gets the face value of his or her debt and the interest paid at maturity—equity has a residual claim on all the cash flows of the company after all the fixed claims are paid. Since startups usually have little to no debt financing (except from possible investments from angel investors that use convertible notes (see ► Sect. 15.3 of this chapter), perhaps some accounts payable to suppliers or friendly family members who provided an early loan), most of the cash flows upon liquidation will flow to the shareholders. Since preferred shares provide Vcs with a debt-like claim that has a priority over the claims of the common shareholders, Vcs are usually paid first upon liquidation: this is called the **liquidation preference**.

The debt-like claim of preferred shares in startup financing usually includes the investment amount and a particular agreed accumulated dividend rate that is paid upon exit. VCs usually negotiate the right to convert these preferred shares into common stock: then, if the exit value is very high, VCs have the incentive to convert their shares to common shares and participate on a pro rata basis with the other shareholders in the startup company. In some cases, the liquidation preference contains a multiple of the initial investment, which is very favorable to the VC. Another possibility is to have **participating preferred shares**. In this case, the VC will never have the incentive to convert its preferred shares into common stock as the participating preferred shares allow the VC to—after receiving its liquidation preference—participate in the residual value with the common stock on an as if converted basis.

Case 4.2: EnvironTECH (Continued)

As is usually the case, in return for its funding, CreativeVC has received preferred shares and has also negotiated a liquidation preference of “2 times participating.”

? Question

What is the effect of this liquidation preference on the payoff for CreativeVC? Would your answer be different if the liquidation preference would have been “2 times non-participating”?

Why would a VC usually use preferred shares with a liquidation preference? On the one hand, the VC receives some protection of its investment: if the startup turns out to be a failure and needs to be liquidated, the VC will be the first to get its money back. However, on the other hand, if the startup turns out to be a **unicorn** with a very high valuation at exit, the VC can convert its shares to common stock and participate in the high upside gains with the entrepreneurs. Since entrepreneurs usually do not receive this downside protection and they thus only start earning money from their startup after the liquidation preference has been paid, preferred shares provide entrepreneurs with the right entrepreneurial incentives. Note that, when VCs have participating preferred shares, they will not have any incentive to convert their shares into common stock. However, it is common practice that, in case of a **qualified IPO**—i.e., an IPO with a very high valuation—VC preferred shares are automatically converted into common shares.

15

15.4.3 Staged Financing

In the previous sections, we already referred several times to financing in multiple rounds. Since startup financing is very risky, investors are usually not willing to pay all needed funds up front in one lump sum, or, if they are prepared to do this, against very unfavorable terms for entrepreneurs. To reduce uncertainty, startup financing takes place in sequential rounds, in which entrepreneurs raise money to

achieve the next milestone. If everything goes well and milestones are reached, there is often no concern and staged financing may actually be beneficial to entrepreneurs as they keep higher ownership stakes in their company when the value of their startup and thus the share price increase (they are less **diluted**).

Case 4.3: EnvironTECH (Continued)

Suppose that Ana, Peter, and Jan are able to achieve their second milestone and they raise €4 million from a new VC called TechFund for a share price of €20 per share.

? Question

Show that the entrepreneurs are less diluted compared to the situation where they would raise this €4 million for a share price of €7.5 per share, for instance if they were not able to achieve their second milestone (“down round”). What is the post-money valuation of EnvironTECH in both situations?

The EnvironTECH case shows that the ownership stakes of entrepreneurs are less diluted in a successful next investment round compared to a **down round**. A down round can happen in many situations: most commonly, entrepreneurs fail to reach their next milestone. However, also external effects, such as worsened economic conditions in the market, can lead to less available VCs to take on the next investment round. Another reason may be that the startup was overvalued in the previous round. Not only the stakes of entrepreneurs are diluted in a down round, but also incumbent investors experience a decrease in value of their ownership stake. For instance, in the previous example of EnvironTECH, if the price in the next funding round would indeed become €10 per share, the value of the stake of CreativeVC decreases from €3 million to €2 million (the VC has a stake of 200,000 shares, which are now valued at €10 per share). To protect themselves from possible down rounds, VCs usually negotiate so-called **anti-dilution protection provisions**. Such a provision, in essence, compensates the incumbent VC for having paid a share price that turned out to be too high. The compensation consists of additional shares, and the particular terms of such an anti-dilution provision determine the amount of extra shares the VC receives.

Particularly, there are two different types of anti-dilution protection: **full ratchet** and **weighted average**. Both clauses have the purpose to compensate the old investor for having paid a share price that turned out to be too high. Yet, the full-ratchet protection provision offers the old investor full protection through the provision of a new conversion price that is equal to the new share price in the new funding round. For instance, in the EnvironTECH case, if the new share price is indeed €7.50 per share and TechFund would thus receive 533,333 shares, a full-ratchet provision inserted in the funding agreement with CreativeVC would lead to an extra 200,000 shares for CreativeVC: the full-ratchet provision provides CreativeVC with the amount of shares that this investor would have received if the price in the first round was the same as in this down round, and hence, this VC receives a total stake of 400,000 shares for its €3 million investment. In contrast, the weighted aver-

age anti-dilution provision provides the old investor with a conversion price that lies somewhere between the share price in the down round and the old price. Usually, it represents the ratio of the total number of shares that would have been issued at the old price, divided by the total number of shares that are actually outstanding after the new round. This ratio is then multiplied with the old share price.

Case 4.4: EnvironTECH (Continued)

Suppose that CreativeVC has a weighted-average anti-dilution protection provision negotiated in its term sheet.

? Question

What is the effect of the share price of €7.50 in the second investment round on the ownership stake of CreativeVC in EnvironTECH?

15.4.4 Corporate Governance

There are generally five fundamental characteristics of corporations all over the world. These include (1) legal personality, (2) limited liability, (3) transferability of shares, (4) a centralized board structure, and (5) investor ownership (Hansmann & Kraakman, 2017). These five core characteristics correspond to the most important economic needs of modern corporations and are shared by virtually every jurisdiction around the world. These characteristics contribute to the attractiveness of the corporate form. Since a corporation is a legal person, it can operate as a single contracting party with a perpetual life that is distinct from its corporate actors such as board members and shareholders. The corporation serves as the common counterparty in contracts with for instance suppliers, customers, and employees. It is the legal owner of the corporate assets, which are separated from the assets of the shareholders. As a result, the creditors of the shareholders cannot claim the firm's assets. Of course, in order to enter contracts or use its entitlements of ownership such as using or selling the assets, the corporation needs representatives to act on its behalf. Limited liability provides protection to shareholders as the corporate owners. Creditors of the corporation are limited to making claims against the corporation's assets and have no claims against those assets that are owned by the shareholders. Hence, in a startup setting, due to limited liability, founders (but also investors) are—in principle—not personally liable for the liabilities of their creditors. The transferability of shares permits the company to continue its business when the owners change: it has a **perpetual life**, independent from its owners.

Corporate governance does not have a set definition but can for example be explained as “*a set of relationships between a company's management, its board, its shareholders and other stakeholders. Corporate governance also provides the structure through which the objectives of the company are set, and the means of attaining those objectives and monitoring performance are determined*” (G20/OECD Principles

of Corporate Governance). Two of the five characteristics are about the governance of the corporation. Every jurisdiction requires the installment of a corporate board, which in general can follow one of the two typical board structures that are used all over the world: a **one-tier** or a **two-tier board model**. In a one-tier board, all directors are part of the same board. These directors can be **executive directors** who direct the company and engage in its daily management and determine the corporate strategy and **nonexecutive directors** who monitor the behavior of the executive directors on behalf of shareholders or its stakeholders, depending on whether the jurisdiction has a shareholder or a stakeholder orientation. As regards the managers below board level, the one-tier board can usually delegate the direction of the company including the initiation and execution of business decisions, but the monitoring function generally cannot be delegated.

In a two-tier board, the **supervisory board members** have a comparable function to the nonexecutive directors in a one-tier board, but they are formally separated from the so-called **management board members** (similar to the executive directors in a one-tier board) and sit on a different board than these management board members. Hence, in the two-tier board structure, a strict division of powers between the management board and the supervisory board is mandatory. Usually, in a one-tier board system, shareholders can elect both executive and nonexecutive directors. In a two-tier board system on the other hand, shareholders usually elect the supervisory board members, who in turn elect the management board members.

The last fundamental characteristic is investor (or shareholder) ownership. Ownership in this respect includes two elements: (1) the right to control the company with legal control rights and (2) the capital right to receive the company's net profits. Shares can thus be seen as a bundle of rights, containing both control and capital rights. Since the powers in a corporation are usually divided between the corporate board and the shareholders (and in some countries, other parties can play a role too, like the government or employees), VCs usually negotiate control at both levels in the term sheets. Since the board of directors has the responsibility to direct the company, VCs usually want to install their own board members (Broughman, 2013). In addition, their shares provide them with control over important decisions that require shareholder approval in the statutory laws of a particular jurisdiction. However, specific shareholder powers can also be determined in the articles of association (or the corporate charter and bylaws) and in shareholder agreements, based on the negotiated clauses in the term sheet. This can be the case for the aggregate of all shareholders, or for holders of a specific share class. For instance, a term sheet can allocate the veto right regarding particular decisions to shareholders holding a particular share class (i.e., the preferred shares as we discussed in ► Sect. 15.4.2). Another example is that holders of preferred shares have the right to elect a particular number of board members. Since control over a company is extremely important, it is crucial for entrepreneurs to not focus solely on the financial aspects of the VC deal, but also carefully consider the clauses in the term sheet that divide the powers of control over the corporation. The incentives of VCs and entrepreneurs are not always aligned, including in exit situations (Fried & Ganor, 2006). The next section of this chapter further briefly outlines the different exit possibilities for startups and the corresponding motives of entrepreneurs and VCs.

15.4.5 Exit Routes

Due to the nature of startup financing, exit is extremely important for investors. Particularly, investment in startups creates a lock-in effect for VCs and other investors for several years, until the investment is cashed out and a return can be realized. Generally, there are three main (successful) exit routes (in addition to a write-off if the venture fails). We start with the most desirable exit route and the dream for many startup entrepreneurs: the **initial public offering**, or **IPO**. In an IPO, part of the shares is sold to the larger public and the shares are listed on the stock exchange. This exit strategy is a very visible one, providing large reputation benefits, and is often considered most profitable of all exit options. Yet, there are also some disadvantages to an IPO. First of all, due to regulatory requirements, investors cannot sell all their shares immediately after the company's listing on the stock exchange, providing not a full exit route for VCs. Next, although IPOs can be very profitable, they also result in substantial costs due to regulatory and administrative requirements and are therefore only beneficial to ventures with large exit values.

The next exit strategy to be considered is an **acquisition**. In an acquisition, all shares of the venture are sold to an external party, for instance a corporation. Although an acquisition is substantially less complex than an IPO, there can be an important disadvantage for the entrepreneurs: as buyers often want to obtain full control and term sheets (and shareholder agreements) usually contain **drag along** clauses, entrepreneurs are usually forced to sell their entire stake in their venture in such an exit transaction. Lastly, **buyouts** are often used by investors to exit the company. In such a transaction, the VC often sells its shares to the entrepreneurs, or an external investor. Buyouts are usually considered to be last-resort exits, as the price paid in these transactions is often substantially lower.

15.5 Tech Startup Financing in Practice

The financing path of every startup is different and highly dependent on a team, type of a product, market conditions, and serendipity. In this section, we will take a closer look at the unique fundraising story of two data analytics companies that went through the full venture circle, from the seed funding to exit. These case studies by no means represent the ideal scenarios or textbook examples that should be blindly followed. Nevertheless, they serve us to demonstrate that the funding success stories may have very different forms.

Our first example is Cloudera that was founded in 2008 by former employees of internet giants Facebook, Yahoo, Google, and Oracle (McDaniel, 2019). Cloudera is a data analytics software company that provides businesses (B2B) a structured, flexible, and scalable platform, increasing their efficiency in collecting, analyzing, and creating actionable insights from overwhelming amount of data points (McDaniel, 2019). Due to the substantial human capital embodied in the founding team, Cloudera attracted venture capital investors from its very inception. In 2009, less than 6 months after its incorporation, Cloudera raised its first external financing round worth 5 million USD led by Accel, reputable early-stage and growth venture capital firm (Crunchbase,

2019a). Interestingly, the deal was executed as a syndicate between Accel, one business angel network, and seven individual angel investors. Accel, however, was a crucial partner and mentor of Cloudera from the very beginning, since it provided the company with an office and Accel's staff helped to incubate Cloudera until its very first VC round (Accel, 2019). Cloudera's solution was so convincing that it took the founding team only few more months to raise follow-up round worth 6 million USD led by Greylock Partners, a VC firm focusing on disruptive, market-transforming consumer and enterprise software companies (Greylock Partners, 2019). In the next few years, Cloudera managed to raise increasing funding rounds almost every year, including 25 million USD round led by Meritech Capital, 40 million USD round led by Ignition Partners, and 160 million USD round headed by T. Rowe Price (Crunchbase, 2019b). In March 2014, Cloudera secured 740 million USD financing from Intel Capital, a corporate venturing arm of Intel Corporation. At that point, Cloudera reached a unicorn valuation of 4.1 billion USD, and it became quite clear that the next logical step in terms of financing is an IPO (Crunchbase, 2019b).

As already mentioned, IPOs are quite expensive and labor-intensive transactions, for which corporations have to thoroughly prepare. Eventually, Cloudera was listed on New York Stock Exchange on 28th April 2017. The corporation decided to price its shares at 15 USD, raising 225 million USD with overall market capitalization of 1.9 billion USD. While still in the realm of unicorns, Cloudera's market capitalization was significantly lower than the valuation of 4.1 billion USD in their last private round with Intel Capital. Indeed, experiencing IPO as a down round is not a usual occurrence. There were several factors that led to this valuation decrease. Firstly, the competitors of Cloudera that already traded their shares publicly were experiencing significant decrease in share prices, and the market of data management companies was becoming more crowded. Moreover, Cloudera's revenue of 261 million USD and high capital burning rate did not warrant 4 billion valuation (Huston, 2017). Despite this, Intel Corporation participated in the IPO and purchased additional 10% of Cloudera shares. The main reason why Intel was willing to pay significant premium on the share price is the strategic cooperation with Cloudera, which was working on "optimizing" software in Intel's processors and other systems (Huston, 2017).

Not every startup will manage to raise multimillion financing rounds and reach the IPO stage. Actually, most of the successful innovative ventures are sooner or later in their life cycle acquired and integrated into product portfolios of larger corporations. In our second case study, we demonstrate this funding route on an example of Dutch data analytics startup SILK. This Amsterdam-based startup was founded in 2010 by Salar Al-Khafaji and Lon Boonen. SILK customers could use the platform to create basic free data visualizations or more complex tooling for visual data presentation that was behind the payment wall (Cook, 2016). The main motto of the product was to get the most out of your data. In 3 years since its incorporation, SILK managed to raise 3.7 million USD in three consecutive early-stage rounds (Crunchbase, 2019c). On 10th August 2016, the founders announced that the company was acquired for undisclosed amount by Palantir Technologies, rather secretive US government-backed data analytics company founded by serial entrepreneur and investor Peter Thiel (van Gool, 2016). In their

blogpost following the acquisitions, SILK founders explained that after meeting the Palantir team, they realized that they have a unique opportunity to move on to much more impactful projects within Palantir, even if it meant that they could no longer work on the core SILK application (Al Khafaji, 2017). In this case, the acquirer was obviously interested in capturing the data processing knowledge and skills of the co-founders rather than on expanding their original product, a transaction that is colloquially known as *acqui-hiring*. For some time, SILK application was still up and running until it finally shut down in 2017 (Al Khafaji, 2017).

Conclusion

The entrepreneurial path and financing opportunities of every startup are different. In this chapter, we outlined various forms of early and later stage financing of innovative ventures and described a number of strategic contractual arrangements between startup and investors in the context of venture capital deals. Since the early stages of a venture, selecting an optimal funding route is always a matter of strategic choice. Entrepreneurship is a complex undertaking; therefore, founders have to be aware of a number of skills and resources that they need to turn their initial innovative concept to a successful business. Understanding the needs of a particular startup is a crucial prerequisite for selecting the most suitable investor. If you possess technical knowledge but generally lack business acumen, you may want to participate in an incubator or accelerator program. On the other hand, if your startup team already covers all necessary and complementary skill sets, you may want to turn to an industry-savvy business angel or promote your consumer-oriented product through crowdfunding.

Take-Home Messages

- Always analyze the financing opportunity according to the added value it can bring to your venture.
- Do not focus only on your capital needs, but also on strategic advice and networking that the investor can provide to your team.
- When conducting negotiations with an investor, focus not only on the important financial aspects of the deal, but also on ownership and control characteristics, as there may be diverging interests between entrepreneurs and investors.
- There is no one-size-fits-all financing trajectory that you should aspire to achieve.

15

? Questions

1. What are the main differences between incubators and accelerators?
2. What is a typical profile of a business angel?
3. Explain the notions pre-money and post-money valuation?
4. What does a “liquidation preference” entail and what are the advantages for investors to have such a preference negotiated?
5. What are the most common exit routes?

✓ Answers

1. What are the main differences between incubators and accelerators?

There are several contrasting features of accelerators and incubators. First, accelerators are restricted in time (several months), while incubators usually do not have a well-defined timeline of support. Second, incubators most often do not take any equity in the supported companies, while accelerators do take a small percentage of shares (between 5% and 15% of shares). Thirdly, accelerators provide their startups with a very intensive program of coaching, mentoring, and group activities, while incubators significantly vary in the scope of the support they offer to startups. Last but not least, incubators are usually attached to corporates or universities, while accelerators often operate as independent entities.

2. What is a typical profile of a business angel?

Business angels are wealthy private individuals that invest their own capital into innovative ventures. In contrast to venture capital firms, they do not pool the funds from institutional investors and thus do not have to be accountable to anybody for their investment decisions and monitoring actions. They are usually former entrepreneurs or corporate executives who beyond financial gain look to share the acquired knowledge and experience. The investment amount of an average business angel per deal can range from €10,000 to €250,000, but some so-called super angels may invest even more. The quality of the mentoring, strategic advice, and networking always depends on the angels' experience and their fit with a startup.

3. The “pre-money valuation” contains the value of the company right *before* the investment takes place; the “post-money valuation” is the value of the company right *after* the investment. Hence, the difference between the pre-money and post-money valuation is the investment of the VC or another investor in a particular funding round. Usually, the investor offers a funding amount for a particular percentage of ownership, for instance, €250,000 for a 25% stake. In this case, the post-money valuation is €1 million, and the pre-money valuation is €750,000. See the formulas in ► Sect. 15.4.1 of this chapter for more information.
4. A liquidation preference is used by investors to protect themselves against downside losses if the startup is not performing well. Remember that investments in startups are very risky and many startups unfortunately go bankrupt. Yet, if the startup performs well, the upside gains are very high. Usually, VCs use preferred shares that include such a liquidation protection. These shares provide VCs with a debt-like claim that has a priority over the claims of the common shareholders. The debt-like claim of preferred shares in startup financing usually includes the investment amount and a particular agreed accumulated dividend rate that is paid upon exit. In this way, VCs are usually paid first upon liquidation and thus have a larger chance to receive their initial investment back. VCs usually negotiate the right to convert these preferred shares into common stock, so that they can share the profits if the startup turns out to be successful.
5. Successful exits are an initial public offering (IPO), acquisition, and a buyout. In addition, if the venture fails, there is a write-off.

15.6 Answers to the Cases

Answer to Case 2.1

Accelerator program may just what Ana and Peter need at this early stage. While the technology behind their venture idea is already fully formed and being executed, they need to start developing a viable business model and try to get first customers and first investors on board. Nevertheless, they should be very careful in selecting the accelerator that will be able to provide them with added value. Probably, an accelerator focused on artificial intelligence or deep tech would work for them the best. As regards the convertible note, this is a debt instrument that usually automatically converts into equity when a first “formal” investment round takes place. For instance, if the share price in an investment round with a VC is €10 per share, this AI accelerator receives in total €100,000/€10 = 10,000 shares. Depending on the terms negotiated, these can be **common shares** or **preferred shares** (see ► Sect. 15.4.2). Sometimes, parties negotiate a discount on the share price or a **valuation cap**, which results in a larger number of shares for the convertible note investor. For instance, if this AI accelerator negotiated a valuation cap of €2 million on the **pre-money valuation** (see ► Sect. 15.4.1 for an explanation), whereas the actual pre-money valuation is €4 million for an amount of €10 per share, there are 400,000 shares outstanding before this investment (*pre-money*). The AI accelerator now pays an amount of €2 million/400,000 shares = €5 per share (i.e., the share price for the valuation cap). Hence, due to the valuation cap, the AI accelerator receives 20,000 shares instead of 10,000 shares when a professional investment round takes place.

Answer to Case 3.1

EnvironTECH aims at building digital copies of physical environments that can serve primarily construction companies or even architects. Their business model indicates that their target customers are companies (B2B) rather than individual consumers (B2C). Particularly, retail investors may not fully understand the viability and attractiveness of B2B businesses; therefore, crowdfunding in general may not be a right choice for Ana and Peter. Nevertheless, there are some crowdfunding platforms that focus on B2B businesses, whose retail investors are much more likely to understand and invest in these types of businesses. In addition to that, equity crowdfunding may pose another set of complexities for early-stage startups. After an equity crowdfunding round, they may suddenly enable hundreds, if not thousands, of individual investors to become the company “owners.” Having to inform and communicate, and eventually buyout, a significant number of small investors may be an administrative task that startups are usually not ready to face.

Answer to Case 4.1 Below a possible answer to this case is provided. Note that the following assumption is made: the ESOP shares are part of the pre-money valuation. In practice, this may be different depending on the particular features of a deal (▣ Table 15.1)

Post-money valuation:	$V_{after} = I_i/s_i$
	= 3 million/20% = 15 million euro
Pre-money valuation:	$V_{before} = V_{after} - I_i$
	= 15 million – 3 million = 12 million euro.
Share price:	$p_s = 15$ euro.
Pre-money amount of shares:	12 million/15 euro = 800,000 shares.
ESOP:	10% of 800,000 shares = 80,000 shares.
Post-money amount of shares:	15 million/15 euro = 1,000,000 shares.
CreativeVC:	200,000 shares for $I_i = 3$ million euro.

▣ **Table 15.1** Overview of ownership stakes (Case 4.1)

Who?	Amount of shares?	Ownership stake (%)
Peter	288,000	28.8
Ana	288,000	28.8
Jan	144,000	14.4
ESOP	80,000	8.0
CreativeVC	200,000	20.0
<i>Total shares</i>	1,000,000	100

Note to table: Author’s own table

Answer to Case 4.2

CreativeVC invested €3 million in EnvironTECH, and with the liquidation preference of “2 times participating,” this VC will get first €6 million back upon a liquidation event and afterwards also participates with common stock in the remaining exit proceedings on an as if converted basis without converting to common shares. For instance, if the exit value would be €20 million, CreativeVC will first receive €6 million, and afterwards receives 20% of the remaining value of €12 million, making its total gain €8.4 million.

In contrast, if the liquidation preference would have been “2 times nonparticipating,” the VC needs to decide between receiving its liquidation preference and converting to and participating with common stock. If it would receive its liquidation preference, it receives €6 million. In contrast, if it participates with common stock with its 20% ownership stake, it would receive €4 million (which is 20% of €20 million).

The liquidation preference has an effect on the incentives of both entrepreneurs and VCs. When a VC has a participating liquidation preference, it has never an incentive to convert its preferred shares to common stock (except when the participation is capped or in the situation of a qualified IPO). On the other hand, in this example, if the VC has a “2 times nonparticipating” liquidation preference, it has no incentives to increase the exit value of EnvironTECH between €6 million and €30 million: for these exit values, its gains will still be €6 million, and only for an exit value that exceeds €30 million will CreativeVC convert to common stock and earn more than €6 million.

Answer to Case 4.3

The entrepreneurs raise €4 million for a share price of €20. Since before the second investment round there are 1 million shares outstanding (see Case 4.2), this means that the pre-money valuation for this new investment round is €20 million and the post-money valuation is €24 million. TechFund receives 200,000 shares of €20 each for this €4 million funding. Hence, after the funding round, there are in total 1,200,000 shares (see ■ Table 15.2). The entrepreneurs (Ana, Peter, and Jan) together still own the majority of the shares (60%). Yet, if the share price was only €7.50, and EnvironTECH would thus experience a “**down round**,” TechFund would receive 533,333 shares for the same investment. Hence, after the second round, the total number of shares is 1,533,333 and TechFund receives an ownership stake of almost 35%. Peter, Ana, and Jan together do not own a majority stake in EnvironTECH anymore (about 47%).

■ **Table 15.2** Overview of ownership stakes (Case 4.2)

Who?	€20 per share		€7.50 per share	
	Amount of shares?	Ownership stake (%)	Ownership stake	Ownership stake (%)
Peter	288,000	24.00	288,000	18.78
Ana	288,000	24.00	288,000	18.78
Jan	144,000	12.00	144,000	9.39
ESOP	80,000	6.67	80,000	5.22
CreativeVC	200,000	16.67	200,000	13.04
TechFund	200,000	16.67	533,333	34.78
<i>Total shares</i>	1,200,000	100.00	1,533,333	100.00

Note to table: Author's own table

Answer to Case 4.4

The calculation of the extra number of shares CreativeVC receives on the formula that is used for the weighted-average anti-dilution protection provision: Generally, it depends on which shares are included in the number of shares issued before the new investment round. In some narrow-based weighted-average anti-dilution protection formulas, the shares of the ESOP and/or those of the founders are not included in this ratio, whereas in the broad-based formula, they are. In this case, we use a broad-based approach and include all shares of EnvironTECH including those of the ESOP and of Ana, Peter, and Jan. The formula can be denoted as follows:

$$Conversion\ price = P_{s1} \left(\frac{S_{before} + \left(\frac{I_2}{P_{s1}} \right)}{S_{before} + \left(\frac{I_2}{P_{s2}} \right)} \right)$$

where P_{s1} is the price in round 1 (which is €15); S_{before} denotes the share base before the second investment round, which is 1,000,000 shares; P_{s2} is the price in the down round, which is €7.50; and I_2 is the investment amount of the second VC, TechFund, which is €4 million. Hence, this formula shows that the ratio of the shares that would have been outstanding of the new round would have had the same price as the previous round, compared to the amount of shares that are actually outstanding after this new down round. Filling out this formula results in a ratio of 0.83 and a conversion price for CreativeVC of €12.39 per share. This means that CreativeVC receives €3 million/12.39 per share = 242,105 shares, and thus 42,105 shares extra. The total amount of shares of EnvironTECH is therefore 1,575,438 after the second round (1,533,333 + 42,105 shares), and CreativeVC thus owns a stake of 15.37%.

References

- Accel (2019). *Cloudera*. Retrieved December 18, 2019, from <https://www.accel.com/companies/cloudera>
- Al Khafaji, S. (2017, November 5). *It's time to say goodbye*. Retrieved December 18, 2019, from <https://blog.silk.co/post/167155630197/its-time-to-say-goodbye>
- Bradshaw, T. (2019, May 23). DoorDash raises another \$600m at \$12bn-plus valuation. *The Financial Times*. Retrieved December 18, 2019, from <https://www.ft.com/content/65f08660-a762-11e9-984c-fac8325aaa04>
- Broughman, B. (2013). Independent directors and shared board control in venture finance. *Review of Law and Economics*, 9(1), 41–72.
- Cook, J. (2016, August 10). *Palantir acquired a Dutch startup that creates data visualisations*. Retrieved December 18, 2019, from <https://www.insider.com/palantir-acquires-silk-dutch-startup-data-visualisations-2016-8>
- Crunchbase. (2019a). *Accel profile*. Retrieved December 18, 2019, from https://www.crunchbase.com/organization/cloudera/funding_rounds/funding_rounds_list#section-funding-rounds
- Crunchbase. (2019b). *Cloudera profile*. Retrieved December 18, 2019, from <https://www.crunchbase.com/organization/cloudera>
- Crunchbase. (2019c). *Silk profile*. Retrieved December 18, 2019, from <https://www.crunchbase.com/organization/silk>
- Da Rin, M., & Hellmann, T. (2019). *Fundamentals of entrepreneurial finance*. Oxford University Press.
- DoorDash (2019, May 23) raises another \$600m at \$12bn-plus valuation. *The Financial Times*. Available at: <https://www.ft.com/content/dal20214-7d80-11e9-81d2-f785092ab560>.
- Fried, J. M., & Ganor, M. (2006). Agency costs of venture capitalist control in startups. *New York University Law Review*, 81, 967–1025.
- Greylock Partners. (2019, December 11). *A leading silicon valley venture capital firm*. Retrieved December 18, 2019, from <https://www.greylock.com/>
- Hansmann, H., & Kraakman, R. (2017). What is corporate law? In *The anatomy of corporate law* (3rd ed.). Oxford University Press.
- HighTechXL (2019). HighTechXL - Enabling Innovation & Entrepreneurship. Retrieved December 18, 2019, from <https://www.hightechxl.com/>
- Huston, C. (2017, April 18). *Four things to know about the Cloudera IPO*. Retrieved December 18, 2019, from <https://www.marketwatch.com/story/four-things-to-know-about-the-cloudera-ipo-2017-04-18>
- Kruppa, M. (2019, July 16). Number of mega-funding venture capital rounds hits record. *The Financial Times*. Retrieved December 18, 2019, from <https://www.ft.com/content/65f08660-a762-11e9-984c-fac8325aaa04>
- McDaniel, S. (2019, August 8). *Cloudera: Unlock the power of data | Talend*. Retrieved December 18, 2019, from <https://www.talend.com/resources/what-is-cloudera/>
- Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing*, 29(1), 1–16.
- NFBAN (2019) Nordic Female Business Angel Network – Harnessing Investment Potential of Nordic Women. Retrieved December 18, 2019, from <https://www.nfban.org/>
- van Gool L. (2016, August 11). *Data visualization startup silk acquired by Peter Thiel's Palantir*. Retrieved December 18, 2019, from <https://startupjuncture.com/2016/08/11/silk-acquired-peter-thiel-palantir-data-visualization-startup/>
- Welch, I. (2022). Corporate finance. Retrieved from <http://book.ivo-welch.info/home/>



Entrepreneurial Marketing

Ed Nijssen and Shantanu Mullick

Contents

- 16.1 Introduction – 383**
- 16.2 Defining Marketing and Sales – 384**
- 16.3 Customers Buy Solutions Rather Than Products – 386**
 - 16.3.1 The Means-End Chain – 386
 - 16.3.2 Trade-Offs Regarding Radically New Products and Services – 388
- 16.4 Co-developing and Positioning a New Product or Service – 389**
- 16.5 Organizing Customer Development as a Separate Process – 391**
- 16.6 A One-Page Marketing and Sales Plan – 394**
 - 16.6.1 The General Motivation and Objectives – 394
 - 16.6.2 Three Main Pillars – 395
 - 16.6.3 Building a Marketing Information System – 399

16.7	Leveraging Your Growing Customer Base – 399
16.7.1	Segmentation and Targeting – 400
16.7.2	Efficient A/B Testing of Value Proposition – 402
	References – 406

Learning Objectives

After having read this chapter, you will be able to:

- Describe and compare entrepreneurial marketing and sales.
- Apply market definition and explore customer segments' attitude towards your product/service based on the trade-offs involved.
- Develop a value proposition using the concept of positioning.
- Understand the importance of a separate customer development process and using it to experiment with customers to validate assumptions.
- Choose marketing and sales instruments to create a one-page marketing/sales plan and formulate measures to evaluate progress in the domains of building value for the customer, market presence, and customer relations.
- Evaluate and optimize customer segments and positioning of a new product/service using data.

16.1 Introduction

Although innovation and marketing are the two business functions that are most fundamental for new business creation (Drucker, 1973), the marketing literature has paid surprisingly little attention to the development of commercial capabilities of new ventures. A possible explanation is marketing scholars' bias towards large companies. However, the question regarding how new ventures can develop commercial capabilities and achieve excellent market performance is highly important, as the answer likely differs for new versus established firms.

Established firms often aim to maximize efficiency gains using traditional marketing planning (Read et al., 2009). Existing market data and familiarity with its customers allow the established firm to use this linear and planning-oriented approach to extend existing business and achieve firm growth.¹ In contrast, young, *entrepreneurial firms* typically aim to expand and achieve maximal growth in the market using experimentation to discover customers. In this endeavor, young firms face a particular set of challenges due to their liability of newness. In addition to having to overcome a lack of reputation and prospective customers' mistrust (DeKinder & Kohli, 2008; Read et al., 2009), they generally have limited human and financial resources to accomplish their goals. Therefore, effectively managing these scarce resources is pivotal to young firms' commercial growth. Through experimentation, the firm will try to match its technology and application with the value of these potential customers, i.e., innovatively minded *prospects*. The aim is

1 So, the suggestion is that established firms mostly engage in incremental innovation. They may engage in intrapreneurship and explore new ground, but this generally is hard, since representatives of the old technology are powerful and hinder or prevent such endeavors (e.g., Christensen, 1997).

to create a new product or service that they value but that will also resonate with the rest, i.e., majority of other customers in the marketplace.

Entrepreneurial marketing uses an effectuation perspective to help young firms face their market uncertainty and find (and build) customers. *Effectuation theory* is one of the major strands of thinking about organizations and the strategic behaviors of young firms (Sarasvathy, 2001; Sarasvathy & Dew, 2005).

Effectuation inverts the fundamental principles and overall logic of predictive rationality, which considers the environment exogenous but predictable and assumes that the firm uses foresight and planning to adjust to trends and capture opportunities in the market. In contrast, in the effectual view, the environment is endogenous to the actions of “effectuators” (which can be either firms or individuals/entrepreneurs) that can apply their resources in an attempt to contribute to and shape the future and environment through commitments with a network of partners, investors, and customers. Effectuation logic starts with the means the entrepreneurial firm has and uses an iterative cycle of experimentation with customers to address and control the uncertain environment and future (Sarasvathy & Dew, 2005). This iterative approach helps the entrepreneurial firm to discover and create (and deliver) value for this customer (Andries et al., 2013). In this process, a broader validation of the customer value with other potential customers is critical to be able to bridge with the rest of the market.

In this chapter, we continue by explaining the difference between marketing and sales. We then discuss how prospects usually perceive and evaluate a startup’s new product or service and introduce the concept of market definition to understand the evolving market and use it to discover the best, initial target segment or niche for the entrepreneurial firm to build its strategic position on. We then introduce the new customer development process, a separate process that complements the firm’s new product development process. It will help ensure that enough time, money, and resources are spent on customer creation (and the business case). Without users or customers, there generally is no business. Subsequently, we discuss the development of a one-page marketing plan and its link with different marketing instruments that can be used to build the new firm’s customer value, market presence, and customer relations, i.e., customer base. Finally, we will discuss how you can optimize your marketing and sales efforts once your number of customers and repurchases start to occur. We will discuss ways in which you can start leveraging this data to improve your product/service offering and customer experience.

16.2 Defining Marketing and Sales

Many businesses use big data to create new value (Hartmann et al., 2016). This activity is just as much about the creation of ideas, products, or services as about discovering and building customers. Without customers (and/or users), there generally is no *business case*, i.e., no money-earning capacity. Dragons’ Den, a TV show where entrepreneurs pitch their business ideas to a small group of wealthy investors (i.e., the Dragons), illustrates this well. Dragons’ interest always immediately surges when an entrepreneur mentions customer interest for his/her product/

service. To the Dragons, customer interest signals potential merit of the new product/service over alternatives that are currently available in the market. It suggests sales potential, which generally is a prerequisite to financial success and thus profits.

The tasks of discovering and building customers, and managing the relationship with an evolving customer base, are the domain of marketing and sales. Both marketing and sales are boundary functions of an organization that help a firm relate to its market. For an entrepreneurial firm, these functions can and should help discover and then manage the exchange relationships with prospects and customers (users).

Definition

Marketing concerns the discovering, creating, and expanding a firm's customer base. Its activities include defining and ensuring the core benefit for the customer.

It relies on the concept of value proposition, which refers to explaining the solution the new product entails for potential customers (users) and how this solution is different, i.e., better than alternatives already out there in the marketplace. The value proposition is not directed to the market at large, but generally to a specific customer segment that benefits most from it. Marketing thus is not just advertising or pricing, but rather concerns all activities for "getting to your market," and staying there. As mentioned, in contrast to traditional marketing, entrepreneurial marketing relies not on rational planning but effectuation logic. It begins with the technology and potential application(s) and then looks for the best customer segment and product/service configuration using experimentation with customers from the emerging segment, often also referred to as *co-creation*.

Particularly, if the entrepreneurial firm draws on new technology which may seriously shake up the current market, existing market data often is irrelevant. As a result, experimentation with customers will be necessary to be able to control the uncertainty involved and discover a viable new business model and its related customer segment(s). Systematic feedback from innovative prospects will allow the young organization to make necessary adjustments for its product or service to work properly and render value to the customer when applied in the customer's context, i.e., consumer *or* business processes. Only then will the entrepreneurial firm not just create but also in fact deliver value and thus create customer satisfaction.

Definition

Sales concerns relationship management towards prospects and customers over time. It aims to move customers towards transactions with the firm and focuses on activities of closing deals. It includes prospecting, approaching, developing, and negotiating with customers. It extends to maintaining the relationship after the deal is done.

Since a customer may repurchase and be a target for cross- or upselling, maintaining the relationship is important. While marketing generally uses a mid- or long-term view, sales has a shorter time horizon. Sales tries to sell the products/services the firm currently has on offer. Selling for an entrepreneurial firm is particularly difficult. Often, the physical product does not yet exist (or only exists on paper), and price information is lacking too. Being unaware of the ultimate demand for the new product, the potential market size is unknown and break-even calculations are hard to make (number of products sold after which all costs are recouped and after which profit will be made). Selling for an entrepreneurial firm concerns new business creation and typically has a long lead time and involves serious sales learning (Leslie & Holloway, 2006). This learning refers to discovering prospects and their needs in response to the ideas and prototypes the new firm puts forward. Sales can act as *knowledge broker* of customer and market information and thus help to connect the firm's R&D staff with these potential customers.

It should be noted that in young firms, the tasks of marketing and sales generally are combined and performed by a single person or small team. Later, when the firm grows, marketing and sales will become separate functions performed by different people. This is a natural process. Once the customer and customer value of the product have been identified, first and foremost sales activities become important; this is the way to scale up sales and increase cash flow. However, marketing will remain important to plan and support further development of the commercial side of the firm. By developing good communication support, prospecting, and talking to product development about further enhancing the product value, further improvements can be made.

16.3 Customers Buy Solutions Rather Than Products

16.3.1 The Means-End Chain

Entrepreneurs commonly believe very much in their idea and new product or service. They generally think that for prospective customers “seeing is believing” and thus that their new product will sell itself. They typically overestimate the value of their new product and underestimate the marketing challenge ahead by underestimating the level of customer conservatism and ignoring customers' switching costs. More importantly, entrepreneurs typically are product and not customer focused. Consequently, when they talk to customers, they stress product specifications (“specs”) rather than benefits. However, customers are generally less interested in the means and more in the ends. Customers are more interested in what your product can do for them than the technicalities. They are interested in the *solution* your product or service offers.

► Example

Imagine that you plan to develop a high-speed train that can attain a speed of 1000 miles/h. Potential customers may be enthusiastic. It suggests to them that they may be able to travel faster and thus save time. However, prospects will probably also wonder about other aspects: How much time will I ultimately save? Is the new option safe? So, customers may be interested in saving several hours' traveling time but will be unimpressed if they will only gain half an hour because they first have to travel to a central location your train is leaving from or because of poor connections. Moreover, they will definitely avoid your train for family travel if the technology is unproven and/or unsafe. Which parent would jeopardize his/her family? Other factors may also explain why people may be less inclined to switch, e.g., an excessive price or they may like their current mode of transport and customer loyalty program. ◀

To better understand how customers relate to a product, psychologists developed the theory of the *means-end chain* (e.g., Macdonald et al., 2016).

► Important

The means-end chain theory argues that customers conceptualize products as bundles of attributes, from which benefits are derived.

The benefits customers seek are those that help them move closer to achieving their ultimate goals or values in life. By focusing on a product's salient attributes and the benefits these offer, customers evaluate the value of different alternatives and make decisions. This is also how they compare new to existing alternatives. The result of the evaluation is a *customer's attitude* towards the new product. The attitude is the total score of weighted salient attributes multiplied by the evaluation scores per alternative on each attribute. If the attitude for the new product compares favorably to the attitude to existing alternatives, then the chance increases that the prospect will consider and buy the new product/service.

Key to the evaluation of new products, particularly those based on new technology, are the inherent trade-offs that are involved (Paluch & Wunderlich, 2016; Gourville, 2006). That is, like in the case of our high-speed train example, the new product or service may bring new benefits, but also involves serious drawbacks (risk) or extra costs (high price). However, often new technology also involves behavioral change. Many customers will already have a product in place and will be used to using it. Those unhappy with their current solution may be willing to consider your new product willingly, but those that are very happy with the current situation may be uninterested and ignore your alternative or prefer to wait. We will discuss this in more detail in the next section.

Understanding customer attitude towards the *trade-offs* involved in new product evaluation can help you to discover which customer segment will not be interested in your new product/service, but also help identify the (new or existing) segment or niche that is most likely to consider adoption.

16.3.2 Trade-Offs Regarding Radically New Products and Services

For many customers, the adoption of a new product involves important trade-offs. Particularly, new products that incorporate new technology often include new benefits but suffer important drawbacks too. For instance, a new algorithm may be faster and better but may be harder to implement due to a lack of data, or because the visualization of results is still lagging behind. Similarly, the first electrical cars offered environmental benefits (over petrol and diesel cars), but had an extremely limited driving range (e.g., less than 120 km). Consequently, these cars were useless for heavy users of the old technology and its vehicles (e.g., a salesperson using a diesel car to travel 40,000 km/year). For this group of individuals—potential customers—the new electrical car was simply no option.

These examples suggest that the first customers of a new technology and its application generally are people who appreciate the benefits but do not mind the drawbacks of the new technology's application. So, environmentalists with limited travel distance might seriously consider the new, first electrical cars, and thus should be identified and targeted (e.g., housewives with a favorable disposition towards sustainability using the family's second car to drive the kids to hockey, or firms with a need to behave environmentally friendly making deliveries in crowded and polluted inner cities). It refers to a possible (emerging) small segment (i.e., a niche), which the young firm could use to create inroads in the car market and from which it can expand its business. As soon as the technology evolves and matures, product performance will improve and drawbacks will be overcome. This will make the product more acceptable also for other, more pragmatic customers in the market.

Derk Abell's (1980) definition of the market and a firm's *business domain* is a useful instrument to conceptualize a young firm's evolving market and position in this market. He believed that a firm's mission was not determined by customers, benefits, or technology, but by all three simultaneously.

Definition

Customer segments concern *who* the firm aims to satisfy, benefits sought refer to *what* is being satisfied, and the technology involves *how* the seller or sellers try to satisfy customer needs.

This implies that the entrepreneurial firm's product (technology application) is in the matrix and connects the axes. The area the entrepreneurial firm fills in the matrix refers to its *mission* and concerns its *business domain*. For example, an app like BlaBlaCar and its technology might compete with public transport, hitchhiking, interliner bus services, etc. Benefits are speed, price/cost, availability, and particularly also social dimension of traveling. Segments could be the commuter, the traditional traveler, and young adventurous and social people. This is, of course,

just one possibility to conceptualize the market and the firm's business domain. Drawing on existing data, many alternative conceptualizations are possible. For each, assumptions about the impact of the new app technology on benefits and segments can be formulated and then verified. The objective of this exercise is to better understand the impact of the entrepreneurial firm's application on the market, to begin to discover the evolution that could take place, and why.

Abell (1980) suggested to look at the market and business domain using an evolutionary lens to understand the changes in the marketplace. The market may and probably will evolve along all three dimensions, with at any point in time multiple technologies coexisting and serving the market with their applications. Each technology has unique qualities (unique selling points or USPs) and limitations (points of disadvantage or disparity), which explains why one alternative better caters to the needs of one segment and another caters better to another segment. To return to our car example, diesel cars cater well to people who travel much, while the new electrical cars address a new (emerging) segment with environmental needs but less requirements regarding travel distance. So, a young firm selling electrical cars should look at its technology and application, determine its customer benefits, and from that try to discover the segment of prospective customers connected with these benefits that do not mind current disadvantages involved using them.

By approaching customers of these segments (in particular the segment you think will most favorably be dispositioned towards your new product), assumptions about the relationships between technology and customer needs can be validated. The validation will involve visiting prospects from this/these segment(s) to establish whether they indeed respond to the new product this way. If not, the exercise can be repeated. By testing customer response to the new product and its pros and cons, insights can be gained about the needs that are well and poorly covered and thus which segments are willing to consider the new option more.

After validation, the young firm can target the identified segment. By focusing on and collaborating with its most innovative customers from this *target segment*, it should then try to enhance the advantages (USPs) while reducing the new application's disadvantages (that is, changing points of disadvantage or disparity into points of parity). This will help make the application, i.e., new product/service, more competitive and more suitable for the niche, and hopefully for other segments of the market too. However, the outcome may benefit from a maturing of the technology over time. For example, the driving range of electrical vehicles has not only increased by offering more battery capacity, i.e., adding a range extender. Battery technology progressed, and as a result the capacity of batteries improved also.

16.4 Co-developing and Positioning a New Product or Service

The quality of your new product, from a customer's point of view, depends on its value-in-use.

Definition

Value-in-use concerns whether a product or service works in everyday practice and accounts for an initial target customer setting (Macdonald et al., 2016).

By working closely with innovative customers from the target segment, the entrepreneurial firm can learn and enhance its product's customer value. By showing its idea/product to innovative prospects, valuable feedback will be obtained. Moreover, by involving the potential customers in ironing out kinks or even repurposing the application, further improvements can be made. Such co-creation efforts are important to establish and increase customer value (Coviello & Joseph, 2012).

Unfortunately, however, most entrepreneurs forego the opportunity to interact with customers and first develop their product to then be disappointed about a lack of customer enthusiasm. In a large worldwide survey, the majority of entrepreneurs indicated regret having waited too long to get out there and obtain customer feedback (Onyemah et al., 2013). As a result, many had wasted time and money developing something no one wanted. In contrast, successful counterparts had sought feedback and involved strategic buyers. Involving these innovative customers as co-developers was an important driver of these entrepreneurs' success. Sometimes the co-development was limited to offering feedback, but often also covered joint engineering and cofounding.

Prospective customers interested in innovation and with a positive disposition to share and discuss problems are the best people to involve. These are people who love new technology and participating in experimentation. They are, for example, the customers camping out in front of a store to buy the first new iPhone, the die-hards. Working with these people, progress can be made to ensure customer value, but is not risk free. These tech-minded customers may add features that do not resonate with more pragmatic customers, which makes up the majority of the market. To reduce the danger of a misfit and thus chasm with the early majority, the entrepreneurial firm should validate the results of the co-creation process with a broader group of customers from its target segments. Sales and marketing should play a key role in this (Leslie & Holloway, 2006). Sales and marketing understand customers' business processes and can help identify the changes in cognitions and routines necessary for customers to enjoy the new product's value (Hartmann et al., 2018). They can identify all stakeholders involved and help make sure that routines to ensure customer value are well understood (including the behavioral change involved). They should also develop and implement sales/service procedures for securing the new product/service's value-in-use.

Definition

Product positioning refers to a statement regarding the presentation of your product in a favorable way that resonates with prospects of the young firm's target segment.

A core concept of product positioning involves identifying *unique selling points (USPs)* (Nijssen, 2022; Kaul & Rao, 1995). Three aspects are important regarding these USPs. First, they should be factual and sustainable. This implies that they should be strong points of your new technology and firm and thus be based on your young firm's competencies. Second, these points should indeed make the product stand out in the crowd. They should help communicate the difference between your product and alternatives from the competition. It means that the contrast effect should be large enough and clear (noticeable difference). The USP should be unique to your product and help to identify it. Finally, the points need to be salient to the customers of your target segment or niche. Only if these points refer to important attributes and benefits will they be recognized and positively affect customer attitude and decision-making in favor of your firm's new offering.

In accord with effectuation logic and marketing theory, product positioning will happen in multiple iterations and together with your strategic buyers. In this process, the young firm should carefully manage its co-development with its innovative customers and ensure value-in-use of its product for the early adopters and early majority of the market. This will benefit from salespeople's active engagement. Salespeople know the importance of excellent product positioning to successful selling.

16.5 Organizing Customer Development as a Separate Process

Although most firms invest all their money in new product development and often save on costs for commercializing the new product, research shows that both activities require equal investment to succeed (Colarelli O'Connor & Rice, 2013). Also, anecdotal evidence shows that young firms typically fail not because their technology fell short but simply because of their inability to find and build customers (Blank, 2007). Consequently, young firms should take customer development seriously and allocate resources to it.

Although new product development accounts for business aspects of new product creation, and also foresees some customer testing and launch activities in many startups, technology typically crowds out customer discovery and building tasks. By making customer development a separate process and managing it carefully, this can be resolved (Nijssen, 2022; Blank, 2007). By making it a separate process, it becomes, in fact, *a project* with separate management attention, goals, and resources/funding. The ultimate goal should be to reach a small yet stable customer base with predictive conversion rate of new prospects into customers. This formal approach will help the young firm to move learning about customers and their problems as early in its development process as possible. It avoids having no means left for commercialization after having researched and developed the new product or service.

Definition

A customer development process can be defined as a systematic attempt to discover, validate, and build customers, which is managed as a project (Nijssen, 2022).

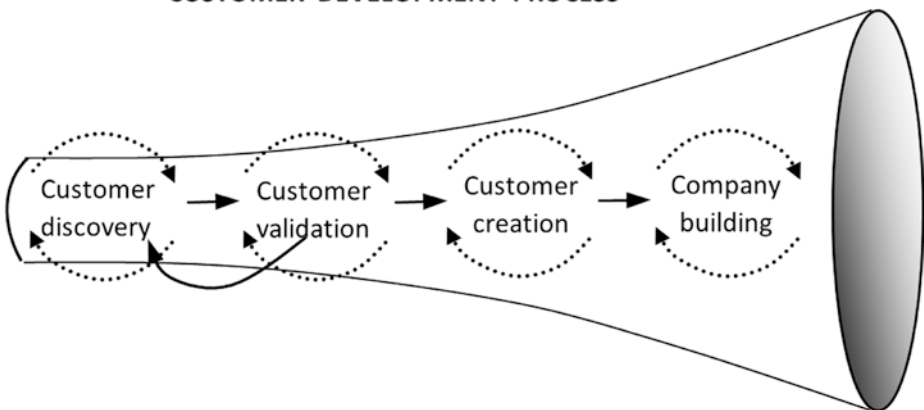
Important

The *customer development process* involves four stages, viz. customer discovery, customer validation, customer development, and company development (Blank, 2007).

We will now go through each of these four stages in more detail (also see Fig. 16.1).

First, the goal of *customer discovery* is just what the name implies, viz. finding out who the customers for your product are and your product can indeed solve the problem (manifest or latent need) you think they experience. It refers to discovering “the business you are in” and the customer niche or customer segment for your product. We suggest using Derek Abell’s (1980) market definition matrix to address this challenge.

Second, *customer validation* involves approaching prospects of the market niche that you identified and getting their responses to your idea and/or prototype. It then extends to co-creation with a set of innovative prospects to create a prototype and polish the original concept to make it more generalizable and thus suitable for other customers of the target segment. “In essence, Customer Discovery and Customer Validation corroborate your business model” (Blank, 2007, p. 21). As the product/service’s value-in-use becomes clearer, ideas about how to develop communications for the market, how to build distribution channels, and thinking about cost and price structure can be developed. Just like the multiple iterations in its product development, the firm will need to employ sales and marketing learning

CUSTOMER DEVELOPMENT PROCESS

16

Fig. 16.1 Customer development stages (Nijssen, 2022)

using experimentation and creativity. The customer development team will begin developing a sales road map.

Third, *customer development* builds on the success the young firm has had in its initial sales. The goal is to routinize the sales road map and begin expanding the customer base. By increasing the number of customers, replicability of the sales road map can be established and demand can be driven further into the sales channel that has been established. The result should be a sales road map that has been field-tested by successfully selling the improved product to subsequent customers. The sales activities and sales efforts will become embedded in a more general marketing perspective about the firms' approach towards its customers and market. The step precedes heavy (heavier) marketing spending and hiring extra sales staff to establish the firm by ramping up sales.

Fourth, *company development* is where the company transitions from its informal learning by the customer development team into formal departments with vice presidents of sales and marketing. These executives will ramp up sales by exploiting the approach developed and building and extending the firm's early market success. Marketing and sales now are regular functions and departments.

The early stages typically will require multiple iterations; two or three efforts for discovery and validation are normal. Unless the firm can discover and satisfy its initial customers and leverage on this effort to create a solution for a segment or market at large, there will be no business model, and thus no future. This "sales (and marketing) learning" process is tough, but an essential part of new business creation.

A *customer development team* should be appointed to manage this customer development. Apart from the entrepreneur, someone with some commercial experience or feeling should be involved, just like someone from the engineering team should be present to act as a linking pin to product development. The former will help to prevent conflict with the original entrepreneurs' points of view and objectives. The latter will ensure that customer feedback flows to new product development. The team will benefit from prior entrepreneurial and marketing and sales experience. If this is absent, then having a marketing and/or sales advisor on board will be useful. This advisor should coach the team and help it develop its own commercial capabilities. For example, a young startup developing a process to work up the waist of 3D printing (plastic) benefitted from such a coach. Of the two entrepreneurs, one led the young firm's product development process, while the other focused on manufacturing and selling. By accident, the latter ran into a retired senior salesperson, who offered to help. The advisor immediately looked at the few customers the young firm was working with to co-develop its product and recommended to extend this rapidly to reduce dependency and to generate more cash flow. He also investigated their price structure. He recommended ensuring at least a 30% price advantage over regular raw material to have a clear USP. It has led to a much more formal approach of the firm's marketing and sales, and actually the enactment of a separate customer development process. It benefitted the young firm's success and secured its future. This example stresses the importance of carefully thinking about a young firm's customer development and creating and empowering a team to lead all this.

16.6 A One-Page Marketing and Sales Plan

16.6.1 The General Motivation and Objectives

Definition

A marketing and sales plan identifies the goals, market strategy, and marketing and sales tactics that will be used to approach the market segment best fitting the startup's technology and application. It helps the customer development team's planning.

The reasons to only use one or a few pages is simple. Consistent with effectuation logic, predicting the future is very hard, if not impossible. Thus, marketing and sales will need to use experimentation and multiple iterations to discover customers, and to help ensure excellent value-in-use of the new product or service. Small plans that are regularly updated best align with this logic.

The development of the plan will be an ongoing effort and will be directly related to the customer development stages. It begins with the aim of discovering the startup's customers and obtaining first customer reactions to validate hypotheses about the new product and its market potential. At this time, the plan will be rather abstract and perhaps only includes a few actions. Slowly, attention will shift from validation to customer creation. By including several innovative customers in the startup's product development process, assumptions about the benefits of the new product application can be further scrutinized. Working with subsequent customers and using experimentation, the product configuration, price, distribution, and sales message, among other things, can then be developed. With a more complete profile of the target customers and their buying behavior, the marketing and sales program can be further detailed and routinized and sales scaled up. The customer development team should begin to systematically document and analyze customer responses in order to develop customer insights.

Content-wise, the plan should focus on three core dimensions: (1) specifying and optimizing the product and its customer value, (2) creating product awareness and market presence, and (3) identifying and converting prospects, and thus building a customer base. The most important elements of the plan should be written down. The further the organization progresses in the customer development process, the more detailed the plan will get. At the beginning, simple marketing and sales measures should be used to measure progress: burn rate (speed of spending money/resources), conversion cost (cost needed to identify and achieve a sale), number of new prospects and customers, etc.

Responsible for making this brief marketing and sales plan is the startup's customer development team. The team should use the plan to ensure proper customer

value creation and delivery. It should also build awareness of the market for the new venture, its technology, and application. In a similar way, it needs to build distribution as a way to get the products/services to the customers. Finally, it should develop a sales channel to develop the young firm's customer portfolio. Joint progress in all three domains will be necessary to turn the opportunity into a viable business and secure the firm's position in the market.

The one-page document should help ensure focus. A focus on a small set of simple marketing and sales goals will prevent spreading means thin. It will help the customer development teamwork in a professional way and show the progress made. This will help build trust and support of the engineers for the young organization's marketing and sales activities. It will help to internally secure the adequate investments in customer development too.

16.6.2 Three Main Pillars

The one-page plan should focus on creating the right customer experience for the startup's new product application. Three core elements exist: building (1) customer value, (2) market presence, and (3) customer relationships. Goals and activities in each domain need to be specified in each period and evaluated afterwards to monitor progress and learn from experience. We briefly review a number of key issues in each pillar.

Building customer value: This refers to the fact that the startup's product application represents a solution to customers and thus has customer value. This value should be identified and optimized, paying particular attention to how the product is used by the customer in his/her consumer or business process. This can be done by carefully considering the different product attributes and their impact on the actual value the customer experiences, i.e., the value-in-use (or lack thereof). By enhancing and highlighting the positive effects and reducing the negative effects, this customer value can be increased. The terms "points of parity" and "-difference" (USPs) are often used in this regard. Optimization of value will involve technical enhancement as well as finding the correct framing of product claims. The latter is particularly important if the technology is new, and customers may lack necessary cognitions (knowledge) to make sense and understand the innovation. To enjoy its value, often new procedures and routines may have to be created at the customer's end. This is a serious task for the sales team, but also for engineers of the young firm. It will involve educating the customer.

However, price plays an important role in this process too. Economic value refers to what a customer gains compared to what he/she has to give up. Only if the new product or service can be sold at a competitive price will its value be high and sales take off. In this sense, it is important to understand cost structure, market size, and learning curve effects.

At the early stages of development, a lot can still go wrong in the value delivery process, so paying attention to excellent customer expectation management and

service provision is important. Whereas customers understand that developing new technology and applications is difficult and risky, they do like to be taken seriously. So, managing customer expectations is an important part of the value-building process. It can reduce the hassle for customers and can be used if recovery is necessary. Thinking about contingency plans can help the young firm respond swiftly to delays and other problems.

A prototype is useful to discuss the new product with customers. It may help identify features and determine the optimal configuration for the target segment and the market at large, respectively. These discussions and customer reactions may hold important information for developing the selling road map too. As the final product configuration materializes, a better understanding of cost and price is possible. Whereas initially price information may be absent, salespeople will now need some kind of basic price list to approach customers. It requires an understanding of the young firm's *cost structure* and developing a *price structure*. Costs include variable costs, but also fixed costs. The former involve costs to make additional products (e.g., raw materials), while the latter refer to, for instance, overhead and R&D costs. Marketing expenses per product should be accounted for, just like warranty service costs and profit margin that retailers or resellers require.

To determine the price for the young firm's product or service, different methods can be used. Apart from a cost-plus (margin) approach, a customer-oriented, a competitor-oriented, and a mixed method exist. The cost-plus approach is the most straightforward one. However, with the actual costs dependent on the volume sold, and with the market size unknown, this may be harder to do. Customer-oriented pricing refers to what customers are prepared to pay. How much is our product worth for them? Competitive pricing looks at the main competitors and uses this as reference point. Generally, the price is set 10 or 15% lower, although it could also be higher if more benefits are offered. The mixed method combines all aforementioned approaches, and hence, it is the most complete method. Table 1 provides a brief overview of the different pricing methods.

Although price setting suggests that one is looking for a specific price, the firm should rather develop a price structure. This not only includes the direct price elements, such as costs and profit, but also accounts for the profit margin of retailers or resellers, potential discounts (if customers place large or repeated orders), product line decisions, etc. Also, business model like considerations of pricing should be considered. For instance, products may be sold in bundles with a single price, leased rather than sold, or combined with a subscription model.

Pricing decisions are complex and deserve serious attention. Price is a very sensitive instrument, and price decisions are not easily reversed. Particularly, it is difficult to raise prices over time. Consequently, for young firms, the general advice is to begin with a high price, an approach that is also often referred to as skimming. This approach is particularly likely to be successful if the focus is on a particular market niche that indeed appreciates the USPs of the product or service. These customers are likely to be more willing to pay (a premium) for your product than the average customer.

➤ Important

Pricing methods

- Cost-plus method: Focuses on the cost per unit and adds a margin on top.
- Customer-oriented pricing: Considers the customers' willingness to pay based on the benefits incurred.
- Competitor-oriented pricing: Looks at competitor prices and then sets the price lower or higher.
- Mixed method: Relies on a combination of the aforementioned methods to determine the price.

Building market presence: Apart from having a good product, market presence is needed for a young firm's business to succeed. The market needs to know about your technology and product to be able to act on it. A separate product category may even have to be “negotiated” with customers and thus developed (Rosa & Spanjol, 2005). Think, for example, again of electrical cars. Although the technology had existed since the early 1900s—even with some early attempts to develop the market in the 1970s and 1980s—the general market was unfamiliar with electrical cars. Therefore, and to ensure the fostering of the right and positive *brand associations* (e.g., Keller, 1999), Tesla invested in Lotus Elise (a sports car being equipped with an electrical engine) and used it on a road show to let the public familiarize with the car and the new technology. Building market presence requires support of opinion leaders and the press to write about the new technology and your innovation.

However, market presence also refers to gaining access to the market by building distribution channels. Although the Internet has made it much easier to reach customers and distribute your products, other channels may be important too. Moreover, possibly your new product or service is not plug and play and requires advice or installation as well as service support to function properly. It is thus important to establish which channels exist or can be built and which partners are required. It is important to ensure that the channel you plan to use is suitable for reaching the customers of your target segment, and whether the channel is available. The distributor will have his/her own business objectives and will ask: What is in it for me? A careful analysis of distribution option and opportunities is called for. This should focus on the buying motivations of these distributors (e.g., profit margin, complementary products, enhancing innovative image, a strategic move). There may be a potential conflict of interest—for example, if the distributor represents competing products—and signing exclusive deals can get a young firm locked in. Be sure that offering exclusive rights does not lock you in and paralyze your operations.

To create the biggest impact using communications and advertising, young firms have to be creative. They should visit events and try to generate free publicity by stimulating the press to write about them. The Internet and contributing to blogs on relevant subjects may help to create the necessary “buzz.”

Building customer relationships: To sustain your firm, you will need to develop a portfolio of customers or solid *customer base*. First, the right target segment for

the innovation needs to be identified. It involves developing a prospect list of innovative customers, the so-called early adopters, and approaching them with information. Getting in contact with these potential customers may be hard, and therefore, attending fairs can be useful. Innovative prospects tend to attend such events, because they like to stay informed about new trends and they are constantly looking for new opportunities for competitive advantage. In the process, the young firm's customer development team will learn about the customer's decision-making unit. This insight may make it easier to pursue subsequent customers. Referrals from initial customers can help generate interest from other prospects.

However, the sales task is not only towards the customer. The young firm's salespeople will need to interact with the organization's engineers to help optimize the product/service to ensure delivering the right customer value. As the new firm will need cash flow, the salespeople should try to get co-financing and/or secure early buy-in from strategic customers, even if the new product is still being developed and its ultimate shape and form may still be unclear and the actual price unknown.

Developing the young firm's sales message and approach is key. It probably benefits from a *value-based selling* approach.

Definition

Value-based selling (VBS) refers to sales activities that involve co-creating a solution with customers to ensure that the customer will enjoy the product's value in its business processes (Terho et al., 2017; Ulaga & Eggert, 2006).

By first focusing on innovative customers that act as strategic partners, the approach may allow the young firm to develop a thorough understanding of these customers' business processes and innovation goals (Andries et al., 2013). VBS will be particularly helpful to exploit the young firm's available resources well by using them to support the customer develop new work routines and goals to experience this newly created value (Hartmann et al., 2018). VBS will help identify all relevant stakeholders and thus members of the *decision-making unit*. To move the customer towards adoption, all members will need to be convinced and offered the right arguments. Approaching all of them with individualized and convincing information thus is called for. Demonstrating the value-in-use and its positive impact on the bottom line will be helpful to persuade everyone and make progress.

In conclusion, progress along all three elements is important for the young firm to move forward marketing- and sales-wise. For each element, goals need to be set (for the next period), activities identified, and budgets/resources determined. It should be complemented with information of who of the customer development team will be responsible for each activity and its goal. This will help ensure that the job gets done. The customer development team should monitor its progress and learn from this. By managing customer development as a separate process using a brief marketing and sales planning method, progress can be assured.

16.6.3 Building a Marketing Information System

The customer development team should systematically collect data about the market and its customers through the customer development process and store it. It will act as a memory for different marketing decisions but can also be used to make further analyses and decisions in the future. Customer information is very valuable for extending relations, but also for analyzing what went wrong. It can be used to help prepare for customer visits and to train new staff. By studying data across customers, patterns may be identified. It can benefit the development of the firm's sales approach, i.e., the firm's sales road map and message.

Definition

A marketing information system refers to systematically collected data about the market, competitors, and general customer developments that is used for understanding customer, competitor, and market dynamics and is used for supporting marketing (and sales) decisions.

The database will develop over time and can be used for developing and improving, for example, the firm's product positioning and estimates about the market size. Since positioning requires a good understanding of strengths and weaknesses of alternatives (or substitutes), information about the unique benefits and drawbacks of each alternative is important.

Generic market data is referred to as secondary data and is contrasted with primary data. Primary data are data that a researcher collects himself/herself for a specific purpose. Secondary data concerns data collected by other people for other purposes but that is still useful for your case. Secondary data are cheaper and quicker to obtain than better fitting primary data. Secondary sources are, for example, newspapers, magazines, Census Bureau or Eurostat (the statistical office of the European Community), societies of manufacturers' industry reports, industry analyses of banks, and companies' annual reports.

16.7 Leveraging Your Growing Customer Base

If you pay careful attention to the points that we discuss in the section about the one-page marketing plan, you will start acquiring customers, and your customer base will start to grow. In this section, we discuss some methods that you can use to leverage your incoming customer data to improve your product/service offerings. In some sense, these methods will be similar to the ones we use in more traditional settings, and yet, some of them will be more tailored to startups in that they are designed to keep in mind that the number of customers—while growing, hopefully—is still relatively small.

16.7.1 Segmentation and Targeting

As your customer base grows, you will realize that the benefits that they seek from your product or service are not the same. This opens the door for you to segment your customers into groups based on the type of value proposition that they find most appealing. Subsequently, you can make small modifications to your product or service so that you have multiple products or services, each designed to appeal to its own customer base. Before we continue, it is best to define segmentation in a more systematic way and to discuss two specific examples to illustrate the concept.

Definition

Segmentation is the process of categorizing customers into groups (or segments or clusters), such that customers within one group are similar enough for you to develop a value proposition that appeals to all of them highly (coherence), and customers within a segment should find the value proposition you develop for them much more appealing than the value proposition you offer to any other group (differentiation).

► Example

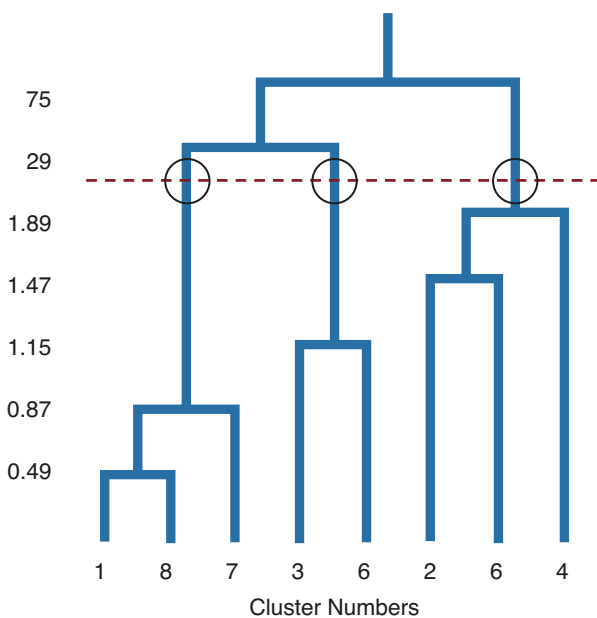
Let us take the example of Dropbox. When Dropbox—originally conceived as an online cloud storage company—started, it had two offerings, viz. one free and the other paid, which provided more storage. With the popularity of cloud increasing substantially, and as people started to use their platform more for working and collaboration (i.e., different people started deriving different benefits from it), they then moved to three plans, with the first two being for individuals and the third being for businesses. Dropbox currently has six plans, viz. three plans for individuals (Dropbox Basic, Dropbox Plus, and Dropbox Professional) and three for businesses. Now consider the three plans targeted at individuals. In terms of the value proposition, the difference between these plans is not just the amount of storage, but also a range of features related to recovery, search, and sharing that cater to the expanded benefits that customers derive out of Dropbox's service. The idea of segmentation is to craft these different value propositions in a manner that, for instance, the customers of Dropbox Plus are happy enough with its benefits that they are receiving (coherence) and would not want to move to Dropbox Basic or Dropbox Professional (differentiation). ◀

► Example

Now let us turn our attention to some apps that were launched recently and that have started building their customer base rapidly, viz. food waste apps such as Swipe Shark, My Foody, and OptiMiam. These apps typically work by linking retailers to consumers and making these consumers aware of perishables—that are close to expiry—that have been put on discount by the retailer. It is a win-win situation for all parties: (1) perishable food does not get wasted, (2) retailers get to salvage some money that would have been lost if the perishables had to be thrown away, and (3) consumers get a discount (Mullick et al., 2021). As more data flows in, some of these companies want to segment customers

to understand their underlying motivation for using the app. For instance, is it driven by a desire to save money or to reduce food waste? Based on these segments, companies want to develop interventions to involve (some of) their customers even more in the fight against food waste, thereby enhancing the experience of these customers. However, it is generally difficult to tease out these underlying motivations from the data you have on customers (related to purchasing behavior, socio-demographics, etc.). Thus, some of the companies focused on reducing food waste resort to surveys where they generally ask a subset of customers what their motives are to use their app. A typical way to collect data for segmentation is to do surveys where you ask your customers to rate on a scale of 1–7 (called the Likert scale; see Edmondson, 2005) how much they would like to experience a new attribute in your product/service. ◀

Now that we know about what segmentation is, we need to understand how it is done. Once you have collected the data, you use it to group “similar” customers into segments based on their needs, which you can deduce from the attributes they liked in your product. You start by grouping two customers that are most similar into one group and then choose a third very similar customer and add him/her to the newly created group and so on. *Similarity* is defined in a more formal way as the Euclidean distance between two customers based on their answers to, for example, the survey question. From a practical perspective, you have computer programs that can calculate and create groups based on similarity, and then you have a dendrogram like the one given in ◼ Fig. 16.2.



◼ Fig. 16.2 Dendrogram showing hierarchical clustering solution. Source: Authors’ own figure

To understand the dendrogram, we also need a concept called *loss of information*, which implies how much information is lost when you add a customer to an existing segment or when you merge two segments into one. In ■ Fig. 16.2, we plot *loss of information* on the y-axis, and we find that there is a sudden jump in the loss of information from around 1.5 to 20. This is a steep jump and one best avoided. As we avoid making this jump, we do not group beyond this particular level. To see the number of groups created, you can draw a line before the jump, and each time the line crosses the dendrogram, it represents one segment. Thus, in the figure below, we have three segments.

However, often in real-life applications, there is no clear jump in loss of information like the one we saw in ■ Fig. 16.2. How would you decide the number of groups then? As a rule of thumb, one can take into account not just the information loss that happens by cutting the dendrogram at a particular point, but also the number of segments that a company can practically manage. Note that each group requires a different value proposition, which requires significant resources to administer.

Once you have formed your segments, you can use the information gained from them to target potential customers. This implies offering a different value proposition to a different set of potential customers. But how would you know which potential customer to offer a particular value proposition to? In general, whenever you conduct a survey among your existing customers (for example, to gather data for the segmentation), you also collect some more information, such as socio-demographics in case of individuals and firm size, sector, and location in case of firms. The idea is to see if a combination of these factors can be used to predict the segment to which a group of potential customers belongs, who you did not include in the survey. This can be viewed as a classification exercise. How accurately you are able to classify the segment of a potential customer can be assessed using the ubiquitous confusion matrix, which helps visualize—in a tabular way—whether the segment predicted for a customer is the actual one he/she belongs to.

16.7.2 Efficient A/B Testing of Value Proposition

Imagine that you have formed your segments and that you are aware of the needs of customers in your segment. The segment that a customer is in should give you a reasonably good idea of the value proposition he/she would prefer, but you still need to convert your value proposition (for each segment) into a concrete product/service offering. As a value proposition encompasses different dimensions, many choices need to be made about, for example, the price, communication, and what attributes to include in the product/service. What typically happens in the industry is that managers develop two to three similar yet slightly different value propositions for each customer segment. The question then becomes how one knows which value proposition is optimal. This is generally achieved using A/B tests.

Definition

An A/B test is a methodology that allows one to compare two different value propositions—for example, two variants of an app or a website—by randomly assigning to different customers to see which one performs better, where perform can imply activities ranging from acquiring new customers to retaining existing customers (Sahni et al., 2018).

Thus, A/B testing allows one to see which one of the two value propositions are better suited for the customer and will perform better. Once you know which value proposition is better, you continue to use that. One of the less expensive communication media for startups to talk to prospective customers is emails. We will see below an example of how A/B tests were used in an email context.

► Example

In the fashion industry, customers buy a lot online, and this trend has really taken off since the last 5–7 years, with companies such as Asos, Zalando, and Shoeby among the leading players. Most companies in online fashion send regular emails to their clients to tell them what the new items on offer are. Customer can choose one or more items listed in the mail and return them for free if they do not like it. Having this policy was proving expensive for fashion retailers, and some of them—notably a fashion retailer in Germany—decided to experiment with the content of the emails (which can be viewed as their value proposition) they send to their clients. In the emails, they included messages related to social norms in product return as one value proposition, and as the second (or control) value proposition, the emails did not include any message related to social norms (Kihal et al., 2019). The authors find that it reduced customers' product return rate compared to the “control” condition. Hence, we see that A/B testing allows for a relatively inexpensive way to decide between two value propositions. This method has picked up quite well in the industry, and companies which have a digital component in their product or service—such as Facebook, Expedia, but also other businesses such as online food retailers—have embraced this method to optimize different parts of its product offering. ◀

One thing to keep in mind though is that the underlying method we use to determine which value proposition is better is classical hypothesis testing. Using such classical hypothesis tests implies that the tests need to be run on a large sample of customers in order to generate meaningful results. This, however, can be a clear limitation for startups, which usually only have access to relatively small groups of customers.

Fortunately, recent developments in the marketing literature have led to a way to overcome this limitation. The intuition behind the method called *test and roll* is that practitioners typically focus on big effects as they generate higher gains (imagine that the difference between two value propositions being considered for the same segment is higher than imagined), while academics (who generally use classi-

cal hypothesis tests) focus on identifying small effects with high confidence. The test and roll method (Feit & Berman, 2019) recognizes the trade-off between the opportunity costs of a test and the potential to deploy a wrong treatment (i.e., use a value proposition that is not the best suited for a segment). This allows them to devise a formula to calculate the sample size that generally leads to much lower numbers than demanded by classical hypothesis tests. Once you conduct the test, the value proposition that generates the higher demand is the “winner” that you use for your subsequent marketing activities.

Conclusion

This chapter explained entrepreneurial marketing and sales. Entrepreneurial marketing begins with an idea and uses iterative steps of experimentation and co-creation with positively minded innovative prospects to develop the idea and ensure that the resulting application entails adequate customer value and works in the customer’s particular context and thus business/consumer processes. In the process, the focus needs to be on obtaining feedback and creating a solution that will also resonate with the early majority in the market. Based on the experience with these first customers, a sales road map can be developed to approach and convince customers in the sales process.

It is important to note that most new products represent a trade-off to prospects. New products may come with new benefits, but in general also with risk and switching costs. While entrepreneurs overestimate benefits, they typically underestimate potential customers’ concerns. By actively searching for segments in the market for whom the aforementioned trade-off looks most positive, and then positioning the product specifically towards this segment, one can increase the chance of building a comfortable customer base and successfully launching the product or service.

To ensure that new product or service development does not crowd out this important customer-building process, a separate customer development process should be implemented and a customer development team installed. It will help ensure that enough attention and resources are allocated to customer development. It will also help ensure customer experimentation and co-creation with innovative prospects, and actions directed towards learning about customers and developing a sales road map to build a customer base. It will ensure attention to market presence-building activities too.

When data of initial customers become available, these data can be analyzed and leveraged to optimize marketing and sales decisions, for example with regard to segmentation, positioning, and customer experience and journey. These analyses and approach begin to resemble traditional marketing activities.

Take-Home Messages

- Focus less on the new product or service itself or its features (“specs”), but on customer benefits instead, because customers seek solutions for their problems.
- Customer development requires at least as much monetary investments as developing the product or service.

- The best chance for creating a solid customer base is (1) implementing a separate customer development process and customer development team and (2) following up with data analyses for optimizing decision-making when data of initial customers come in.
- Value-based selling can be used to ensure that your product/service can be integrated in the customer's business processes, and thus presents to the customer value-in-use. This ensures value creation and delivery and helps reduce perceived customer risk.
- By using marketing instruments for value creation, building market presence, and customer relations, a strategic position in the market is developed, which begins with the initially identified target segment or niche.

? Questions

1. How does entrepreneurial marketing differ from traditional marketing?
2. Do customers aspire a product or a solution, and why?
3. Why is having a formal customer development process important?
4. Why is building a marketing information system and thus database important?
5. Once you have a good customer base, how would you test which value proposition is more optimal for your customers?

✓ Answers

1. Traditional marketing uses a planning approach, whereas entrepreneurial marketing relies on effectuation, discovering the future via small steps and by responding to developments. Moreover, entrepreneurial marketing uses an inside-out approach, whereas traditional marketing uses an outside-in approach. The latter identifies latent needs and creates or adapts products or services for it. The former begins with the idea and then searches for the market segment that best fits with it; it uses co-creation to optimize customer value.
2. Prospective customers seek solutions, not products. Products or services are merely a means to an end. This can be understood using the means-end theory, which states that customers understand the value of a product to help them get closer to their goals in life through their attributes and the benefits that come from them.
3. Although regular new product development includes a feasibility stage and determining whether a market for the new product exists, engineering issues generally dominate commercial issues. This is particularly the case if the entrepreneur is an engineer. Installing a formal customer development process can help prevent this typical underinvestment in customer building.
4. Building a marketing information system is important to allow for making market-based decisions in the future. By systematically collecting data and then analyzing these data as it comes in, decisions can be checked and optimized.
5. When you have built a good customer base, you should use A/B testing to decide which value proposition works better with which part of your customer base. One of the ways to administer an A/B test that does not require a lot of budget is to customize the contents of the emails you send to your customers.

References

- Abell, D. F. (1980). *Defining the business: The starting point of strategic planning*. Prentice Hall.
- Andries, P., Debackere, K., & Van Looy, B. (2013). Simultaneous experimentation as a learning strategy: Business model development under uncertainty. *Strategic Entrepreneurship Journal*, 7, 288–310.
- Blank, S. G. (2007). *The four steps to the epiphany, successful strategies for products that win*. Retrieved November 1, 2018 from https://web.stanford.edu/group/e145/cgi-bin/winter/drupal/upload/handouts/Four_Steps.pdf.
- Christensen, C. M. (1997). *The Innovator's Dilemma*. HBS Press.
- Coviello, N. E., & Joseph, R. M. (2012). Creating major innovations with customers: Insights from small and young technology firms. *Journal of Marketing*, 76(6), 87–104.
- DeKinder, J. S., & Kohli, A. K. (2008). Flow signals: How patterns over time affect the acceptance of start-up firms. *Journal of Marketing*, 72(5), 84–97.
- Drucker, P. F. (1973). *Management: Tasks, responsibilities, practices*. Harper & Row.
- Colarelli O'Connor, G., & Rice, M. P. (2013). New Market Creation for Breakthrough Innovations: Enabling and Constraining Mechanisms. *Journal of Product Innovation Management*, 30(2), 209–227.
- Edmondson, D. R. (2005, April). Likert scales: A history. In: *Proceedings of the 12th conference on historical analysis and research in marketing (CHARM)* (pp. 127–133).
- Feit, E. M., & Berman, R. (2019). Test & Roll: Profit-Maximizing A/B Tests. *Marketing Science*, 38(6), 1038–1058.
- Gourville, J. T. (2006). Eager sellers, stony buyers: Understanding the psychology of new-product adoption. *Harvard Business Review*, 99–106.
- Hartmann, N. N., Wieland, H., & Vargo, S. L. (2018). Converging on a new theoretical foundation for selling. *Journal of Marketing*, 82(2), 1–18.
- Hartmann, P. M., Zaki, M., Feltmann, N., & Neely, A. (2016). Capturing value from big data—A taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management*, 36(10), 1382–1406.
- Kaul, A., & Rao, V. R. (1995). Research for product positioning and design decisions: An integrative review. *International Journal of Research in Marketing*, 12(4), 293–320.
- Keller, K. L. (1999). Brand mantras: Rationale, criteria and examples. *Journal of Marketing Management*, 15(1–3), 43–51.
- Kihal, S. E., Emrich, O., & Pfrang, T. (2019). After the nudge is gone: Keep good resolutions or return to old (return) habits? *Working Paper*.
- Leslie, M., & Holloway, C. A. (2006). The sales learning curve. *Harvard Business Review*, 84(7/8), 115–123.
- Macdonald, M. K., Kleinaltenkamp, M., & Wilson, H. N. (2016). How business customers judge solutions: Solution quality and value in use. *Journal of Marketing*, 80, 96–120.
- Mullick, S. Raassens, N., Haans, H., and Nijssen E.J. (2021). Reducing food waste through digital platforms: A quantification of cross-side network effects, *Industrial Marketing Management -SI*, 93 (February), 533–544.
- Nijssen. (2022). *Entrepreneurial marketing and effectual approach* (3rd ed.). Routledge Publishers.
- Onyemah, V., Pesquera, M. R., & Ali, A. (2013). What entrepreneurs get wrong. *Harvard Business Review*, 91(5), 74–79.
- Paluch, S., & Wunderlich, N. V. (2016). Contrasting risk perceptions of technology-based service innovations in inter-organizational settings. *Journal of Business Research*, 69(2016), 2424–2431.
- Read, S., Dew, N., Sarasvathy, S. D., Song, M., & Wiltbank, R. (2009). Marketing under uncertainty: The logic of an effectual approach. *Journal of Marketing*, 73(3), 1–18.

- Rosa, J. A., & Spanjol, J. (2005). Micro-level product-market dynamics: Shared knowledge and its relationship to market development. *Journal of the Academy of Marketing Science*, 33(2), 197–216.
- Sahni, N. S., Wheeler, S. C., & Chintagunta, P. (2018). Personalization in email marketing: The role of noninformative advertising content. *Marketing Science*, 37(2), 236–258.
- Sarasvathy, S. D. (2001). Causation and effectuation: Toward a theoretical shift from economic inevitability to entrepreneurial contingency. *Academy of Management Review*, 26(2), 243–288.
- Sarasvathy, S. D., & Dew, N. (2005). New market creation through transformation. *Journal of Evolutionary Economics*, 15(5), 533–565.
- Terho, H., Eggert, E., Ulaga, W., Haas, A., & Böhm, E. (2017). Selling value in business markets: Individual and organizational factors for turning the idea into action. *Industrial Marketing Management*, 66, 42–55.
- Ulaga, W., & Eggert, A. (2006). Value-based differentiation in business relationships: Gaining and sustaining key supplier status. *Journal of Marketing*, 70(1), 119–136.

Data and Society

Anne Lafarre

Data scientists are often called the engineers of the future, and a vast majority of innovations and research projects today are data driven. The ultimate goal is to create real value out of data, thereby establishing beneficial insights and changes for business and society. For instance, obtaining data about the quality of air can provide the basis for effective climate change policy actions removing dangerous pollution. Since the start of the COVID-19 pandemic in 2020, many governments and other parties have been cooperating with data science experts to develop apps and other data-driven solutions that should help identifying contamination patterns, tracing interpersonal contact and facial recognition, and providing immunity proofs and other ways to assist in controlling the pandemic. The creation of value out of data—including the relevant aspects of data entrepreneurship—has been widely discussed in the previous modules of this book. Yet, these COVID-19 apps, but also all other data science projects, require a strong sense of professional, legal, and ethical responsibility. In this final *Data and Society* module, these fundamental concepts for data entrepreneurs are extensively discussed.

First of all, there are very important data protection, privacy, and intellectual property rights issues that need to be considered. Knowledge of data protection and privacy rules is essential for data scientists nowadays, particularly when handling personal data. As will be explained in ► Chap. 17 titled *Data Protection Law and Responsible Data Science*, data protection is the European term used for the body of law that determines what type of operation can be performed on data, and under what circumstances—and thus safeguards—the appropriate use

of personal data should take place. The latter is often denoted as privacy or data privacy. The chapter provides a clear and practical road map of the comprehensive and much-debated General Data Protection Regulation (GDPR) that entered into force in European member states in 2016. For instance, it shows that the notion of personal data can be quite broad and that it is important to establish a valid consent when you are processing or using personal data.

► Chapter 18 titled *Perspectives from Intellectual Property Law*, in turn, addresses the important features of intellectual property rules that are particularly relevant for data scientists and when handling data. Intellectual property is quite abstract, as it provides property rights to intangible matters like ideas, books, and music. Like the previous chapter, this chapter takes a European perspective and provides practical explanations of the various intellectual property rights, including copyrights, database rights, and trade secrets rights. It is essential for data scientists to be aware of the fact that intellectual property rights may exist on the data they want to use, and whether the data may indeed be used by third party.

However, as an entrepreneur, you are particularly interested in protecting your valuable assets that are key to your business, including your data, algorithms, and developed software. When starting a business, it is important to set a proper fundament. This includes investing in a sound legal structure, in which your intellectual property rights are properly defined (including the ownership of these rights). Employment agreements may, for instance, specify that all intellectual properties generated by entrepreneurs and employees belong to the company. Moreover, once professional investors are joining a startup business, they may require entrepreneurs to sign over to the company any preexisting intellectual property. Another example from practice is that entrepreneurs are required to provide evidence that they indeed own the intellectual property that is valuable to their company. To set this proper fundament, dealing with contracting and liability issues is of the utmost importance for data scientists, as discussed in ► Chap. 19 titled *Liability and Contract Issues*

Regarding Data. This chapter clearly discusses the important special attention for data and data infrastructure when legal issues arise.

As ► Chap. 20 titled *Data Ethics and Data Science: An Uneasy Marriage?* explains, ethics revolves around the question “how one should act,” in which morality plays an important role. The authors discuss several fundamental moral theories and link the COVID-19 apps’ discussion to one of them, viz. consequentialism. Here, a central question may be whether having a COVID-19 app would maximize well-being or not, providing a trade-off between *safe-guarding health* and *safeguarding privacy*. Yet, the discussion also highlights other important ethical and societal questions going further than privacy issues. For instance, may an employer require its employees to install and use such an app? What about the public transportation services, restaurants, theaters, or sports facilities? But also, the use of data-driven solutions for COVID-19 (or data-driven solutions to other problems) by (totalitarian) governments may magnify discrimination and other severe human rights issues.

An important aspect here is that technology is not neutral, but that design choices and characteristics of the technology itself affect society. As duly noted in ► Chap. 21 titled *Value-Sensitive Software Design* of this module, “[t]he innovative use of technology can [...] be highly valuable, but also highly problematic. Technology has therefore been the inspiration for utopias and dystopias alike.” In this chapter, the importance of value-sensitive design (VSD) that takes into account human and ethical values is explained and explored, providing clear rules of thumb on how to deal with this difficult yet important task.

To summarize, the *Data and Society* module provides some important insights on how professionalism and ethical responsibility can be shaped in data science. It also stresses that data entrepreneurs, who want to create value out of data, need to take into account complex privacy, property, contractual, and liability rules related to data and data infrastructure in order to become successful. The authors of this module provide practical examples and advice where possible to shed light on these—for data scientists often unfamiliar—concepts.

Contents

- Chapter 17 Data Protection Law and Responsible Data Science – 415**
Raphaël Gellert
- Chapter 18 Perspectives from Intellectual Property Law – 443**
Lisa van Dongen
- Chapter 19 Liability and Contract Issues Regarding Data – 461**
Eric Tjong Tjin Tai
- Chapter 20 Data Ethics and Data Science: An Uneasy Marriage? – 483**
Esther Keymolen and Linnet Taylor
- Chapter 21 Value-Sensitive Software Design – 503**
Paulan Korenhof
- Chapter 22 Data Science for Entrepreneurship: The Road Ahead – 523**
Willem-Jan van den Heuvel, Werner Liebregts, and Arjan van den Born



Data Protection Law and Responsible Data Science

Raphaël Gellert

Contents

- 17.1 Introduction – 415**
- 17.2 A Few Words on the Meaning of Privacy and Data Protection – 415**
- 17.3 Material Scope of Data Protection Law: Defining Processing and Personal Data – 417**
 - 17.3.1 Defining Processing – 417
 - 17.3.2 Defining Personal Data – 417
 - 17.3.3 Conclusion: Personal Data and Non-personal Data – 420
- 17.4 Personal Scope of Data Protection: Controller and Processor – 420**
 - 17.4.1 The Three Main Actors of Data Protection – 420
 - 17.4.2 Data Controllers – 421

- 17.4.3 Data Processors – 421
- 17.4.4 Problematic Situations – 422
- 17.5 Art. 6, GDPR: The Need for a Legitimate Ground of Processing – 423**
 - 17.5.1 Consent – 423
 - 17.5.2 Contract – 426
 - 17.5.3 Vital Interests of the Data Subject – 427
 - 17.5.4 Performance of a Task Carried Out in the Public Interest or in the Exercise of Official Authority Vested in the Control – 427
 - 17.5.5 Compliance with a Legal Obligation to Which the Controller Is Subject – 427
 - 17.5.6 Legitimate Interests of the Data Controller or a Third Party – 428
- 17.6 Art. 5 GDPR: Principles to Be Applied to the Processing of Data – 432**
 - 17.6.1 Purpose Limitation Principle – 432
 - 17.6.2 Data Minimisation – 435
 - 17.6.3 Storage Limitation – 436
 - 17.6.4 Additional Obligations – 436
- References – 438**

Learning Objectives

- Understand the difference between privacy and data protection
- Understand how data protection law works
- Understand and be able to determine whether a piece of data is personal or not
- Understand and be able to determine the duties of actors under data protection law
- Understand and be able to choose the right ground for processing personal data
- Understand and be able to correctly apply the principles applying to the processing of data

17.1 Introduction

This chapter provides data scientists with an introduction to data protection law. Data science relies upon the collection and analysis of data. Data protection is a body of law that determines what type of operation can be performed on data, and under what circumstances. For this reason, it is crucial for data scientists to have some basic knowledge of the main principles of data protection law, so that they can do data science in a socially responsible way. This chapter provides a general explanation of the key principles in a way that allows you to use them when confronted with a data science application, like the following example shows.

► Key Questions

Company X is an online marketplace. It therefore needs a certain amount of customer information such as address, name, or credit card information. However, it appears that Company X also uses this information to create political affinity profiles (i.e. profiles concerning the political orientation of its customers). These are then sold or rented to political parties who can then send targeted advertising. What are the rules that apply to this kind of data processing? Is it legal at all? ◀

This chapter proceeds in four steps. First, it provides some preliminary remarks on what is meant by (European) **data protection law**, and how it differs from **the right to privacy**. Second, it looks at the scope of data protection. This includes both the **material scope** (what is personal data) and the **personal scope** (who are the actors). Third, it looks into the conditions under which it is possible to start processing data. Finally, it looks at the principles that must be respected when data are actually being processed.

17.2 A Few Words on the Meaning of Privacy and Data Protection

The right to privacy is a key right of (European) democratic states. It is a complex and multi-folded right. This right was first conceptualised as the right “to be let alone” associated with issues of intimacy, secrecy of correspondence, protection of

the domicile, etc. (see Gutwirth, 2002). In Europe, the right to privacy is safeguarded by two supranational institutions: the European Union (EU) and the Council of Europe (CoE).¹

The right to privacy has gone on to cover many more issues, such as the right to make essential personal choices like a person's name and sexual orientation, gender, health, and identity (see Gutwirth, 2002). For this reason, the right to privacy in Europe is now associated with self-determination and autonomy (see Gutwirth, 2002). This chapter focuses on only one aspect of the right to privacy, which has sometimes been referred to as “data privacy” (see, Hoofnagle, Sloat, Zuiderveen Borgesius, 2019, p. 70), namely, the privacy issues that arise when our personal data are processed by computers. This is what data protection is about. Data protection law can therefore be understood as the law/legal framework that determines how our personal data should be processed in order to avoid the violation of our right to privacy (and other fundamental rights such as anti-discrimination for that matter).

This chapter focuses on the EU, because it has recently adopted what can be qualified as the most comprehensive data protection law, which is directly applicable in all EU member states: the **General Data Protection Regulation (GDPR)**. Note that data protection is also a fundamental right of the EU, enshrined in Article 8 of the EU Charter for Fundamental Rights.

Finally, a few words should be said about the United States. Even though the US legal system has some legislation regulating the processing of data, it does not recognise the concept of data protection. Here, everything is labelled under the term privacy, information privacy, or more recently consumer privacy, even if it is about data protection (see Hoofnagle, Sloat, Zuiderveen Borgesius, 2019, p. 70). The GDPR can be conceived as an “omnibus” legislation. This means that everything is contained in one law, which applies to all activities and to everyone: administrative bodies, businesses, other private parties, etc. In contrast, the laws in the US system cannot be considered omnibus legislation: at the federal level, the United States has a few laws that only address certain sectors such as the Health Insurance Portability and Accountability Act (HIPPA) or the Fair Credit Reporting Act (FCRA). Business activities are mostly regulated by the Federal Trade Commission (FTC), which is the consumer protection regulator (watchdog). Since 1995, it has expanded its competence to regulate “unfair and deceptive practices” to include issues of personal data processing (see, Gellman, 2019).

In general, it is fair to say that the level of protection is much lower in the United States than in the EU. Beyond the complexity and inconsistencies of the legal framework and the lack of a real data protection watchdog, the actual protection is much lower. Some important elements are for instance the narrow definition of personal data (compared to the European definition which is extensively discussed in this chapter) and a system which is mostly based on consent (so-called

1 While fundamental rights are only a tiny part of the EU's competences, the Council of Europe is an international organisation that specialises in the protection of fundamental rights. It currently has 47 member states. This comprises all EU members states, but also many more; see ► <https://www.coe.int/en/web/about-us/who-we-are>.

notice and consent), which offers less protection in practice, especially when compared to the European framework (see, Gellman, 2019). Yet, change is on its way: for instance, California has adopted the California Consumer Privacy Act (CCPA) in 2018, which has taken inspiration from the GDPR (but is only limited to California). At the federal level, the “Consumer Data Privacy and Security Act of 2020” (the “CDPSA”) has been put forth in March 2020.²

The chapter considers the four following aspects of the GDPR. First, it will investigate its scope, both personal and material. That is, when does data protection law apply, and to whom does it apply? It will then look at some of the most important substantive provisions, the so-called core principles that are enshrined in Articles 5 and 6 of the GDPR. These determine under what conditions it is possible to start processing personal data, and what principles should be respected when such processing is taking place.

17.3 Material Scope of Data Protection Law: Defining Processing and Personal Data

The material scope can be said to refer to the “what”: To what does data protection apply, and what is the object of data protection? It is the processing ► Sect. 17.3.1 of personal data Sect. ► 17.3.2.

17.3.1 Defining Processing

Data protection law applies to “the processing of personal data wholly or partly by automated means” following Article 2(1) of the GDPR. Automated means include computers and any type of digital device. Processing in turn means “any operation or set of operations which is performed on personal data” (Article 4(2), GDPR). This means that the life cycle of a processing operation starts at the moment the data is collected and ends when the data is destroyed or anonymised (see ► Sect. 17.4.2). In between these two moments (and including them), any operation that is done on the personal data at stake will be considered a processing.

17.3.2 Defining Personal Data

Personal data is defined as “**any information relating to an identified or identifiable natural person (‘data subject’)**” in Article 4(1) of the GDPR. To put it more simply, personal data are the data of the data subject, which is being processed (i.e. your data). As one can see, the definition contains four key elements, which are briefly considered below.

2 At the moment of writing this chapter, this Bill has not been adopted (yet). Last moment of writing: 29 June 2020.

17.3.2.1 “Any Information”

It is not entirely clear why the GDPR considers that personal data is information (instead of data). To simplify, both terms are treated as synonyms in this chapter, but one may note that the legal explanation of the definitions is related to the official terms used in the GDPR. According to the Article 29 Data Protection Working Party (Art. 29 WP) – an advisory body made up of a representative from the data protection authority of each EU Member State, and issuing the most respected guidance on the topic, which is now replaced by the European Data Protection Board (EDPB) – what qualifies as information under the GDPR is very broad. It does not matter whether the information is private or public, true or false, and subjective or objective. Furthermore, as long as it is suitable for processing by automated means, any format will do (e.g. digital bytes, audio, video, drawing) (Art. 29 WP, 2007, pp. 7, 8).

17.3.2.2 “Relating to”

For data to relate to the data subject simply means that these data should be about the data subject. This can be the case in various ways. The simplest way is when the **content of the data** (clearly) relates to this data subject, for instance, when the data contains a person’s name, address, and social security number that relates to that person in content. This relation can be quite extensive: it can be argued that if a data set contains a person’s smartphone location, it also relates to a data subject (Art. 29 WP, 2007, p. 9).

However, personal data can also relate to data subjects in more complicated ways that go beyond the strict content of the data. This is the case when the data is used in specific ways. Then the data relates to, or is about, the data subject not because of its content but because of the way in which it is used. This will be the case in two situations: when the data relates in **terms of purpose** or in **terms of result** (Art. 29 WP, 2007, pp. 10–11). The data relates **in purpose** when it is processed with the purpose to evaluate, treat in a certain way, or influence the data subject (Art. 29 WP, 2007, p. 11). This can be the case if information about a house’s consumption of energy is used to generate a bill to the occupier for payment (Welfare & Carey, 2018, p. 10). Finally, the data can relate to the data subject **in result (or impact)** when the data is processed in a way, which at the end of the road will have an impact on the data subject. This impact does not need to be major (i.e. simply being treated differently can suffice) (Art. 29 WP, 2007, pp. 11–12). This would be the case of a test of machinery at a factory, which reveals differences in the productivity of two workers and which leads the factory to make changes to their workers’ working pattern (not to mention firing people, see Welfare & Carey, 2018, p. 10).

17.3.2.3 “Identified or Identifiable”

The data must not only relate to the data subject. The data subject must also be **identified** or **identifiable**. In other words, we must know who this data subject is. There can be some confusion between these two requirements (“relating to” and “identified/identifiable”) as they may seem similar, but they are quite different. The first—“relating to”—can be described as being from the perspective of the data subject: “Does the data relate to him/her?” The second—“identified/identifiable”—

can be described as being from the perspective of those who process the data (data controller and/or processor, see ► Sect. 17.4 of this chapter) and third parties. Here, the question “do we know who the person whose data we are processing is?” is relevant.

The GDPR distinguishes between data subjects that are identified and those that are identifiable (Art. 29 WP, 2007, p. 12). The data subject is **identified** when the data controller/processor already possesses in its data set information that identifies (i.e. singles out) the data subject. This can be the name or another (unique) identifier such as mobile phone number, car registration number, social security number, location data, and online identifier (Art. 29 WP, 2007, p. 13). This is the reason why the GDPR refers to identifiers as “one or more characteristics that are the expression of a physical, physiological, psychological, genetic, economic, cultural, or social identity” (Recital 26, GDPR). Such information has a close relationship with data subject and therefore allows for his/her identification. As one can see, many identifiers also qualify as data that relates in content to the data subject (e.g. a name relates to and identifies a data subject). It is however crucial to conceptually distinguish between the two.

Hence, note that for a data subject to be identified, it is not necessary to know his/her exact name. It suffices to be able to single out the data subject, that is, to know who this person is and to be able to distinguish him/her from others (Art. 29 WP, 2007, p. 14). On the contrary, the name is not always a reliable identifier (i.e. information that allows for identification) in a planet of nearly ten billion human beings (many people have the same name) (see Art. 29 WP, 2007, p. 13).

The data subject is **identifiable** when the data controller/processor does not possess information in its data set that identifies the data subject but is nonetheless able to identify the data subject (Art. 29 WP, 2007, pp. 13–14). In other words, the fact that the data subject is not identified but could be identified makes him/her identifiable. The GDPR argues that the identification of the data subject can be done by those who process the data or any other third party. It can be done in various ways (Art. 29 WP, 2007, pp. 15–17).

The criterion for determining whether a data subject is identifiable within a data set is by using all the **means** “reasonably likely to be used” (Recital 26 GDPR). The GDPR therefore excludes the hypothetical possibility of identifying a data subject (Art. 29 WP, 2007, p. 15). One can distinguish between so-called **technical** or **objective means** and **organisational** or **subjective means**. The GDPR requires to take into account the **state of the art at the time of processing**, but also the evolutions in technology (Recital 26 GDPR). Technical (or objective) means refer to the addition of information in the data set. This can be done by **combining** different information that in itself would not have traced back to the person but does so in combination, when linking data sets or when inferring information from existing data (e.g. because of the data set structure) (see e.g. Art. 29 WP, 2007, p.13,15). In all cases, additional information has been added and the data subject is therefore identified. In the case of organisational or subjective means, the data controller does not add additional information, but is nonetheless able to identify the data subject (see Art. 29 WP, 2007, p. 16). The context is key here. The context is relevant both for the technical and subjective aspects of identification. It can refer to

the concrete possibility of identification (e.g. how easy and little costly it is to find additional identifying information, the processing as such entails the merging of various data sets that will lead to identification), but also to other actions and motives of the data controller that can lead to the identification (see Mourby et al., 2018, p. 231). This is the case if the data set is shared or leaked to a third party, or when for instance the purpose of the processing entails *per se* the identification of the data subject: “to argue that individuals are not identifiable, where the purpose of the processing is precisely to identify them, would be a sheer contradiction in terms” (Art. 29 WP, 2007, p. 16).

17.3.2.4 “Natural Person (Data Subject)”

The person in “personal data”, or the person to whom the data relates, is known as the data subject, who must be a “natural person” (Article 4(1), GDPR). In essence, this means two things. First, the person must be alive. Second, the person cannot be a legal person. In law, it is possible to have the so-called **legal personhood** or **legal personality**, which is mostly used for companies but not only that (e.g. rivers and mountains under several indigenous peoples laws; perhaps robots one day considering the (European) discussion on the legal personality of AI, see European Parliament, 2017).

17.3.3 Conclusion: Personal Data and Non-personal Data

As a way to conclude this overview of the notion of personal data, one can say the following. It is a very broad notion encompassing most types of information for instance. It is also contextual, since the context will often help determine whether a data relates to an individual or whether the latter is identifiable. In this regard, the identification test of “the means reasonably likely to be used” shows that personal data is also probabilistic. In other words, in some cases, it is very straightforward whether a piece of data is personal or not. In other cases, however, it will depend on the possibility of relating the data to the data subject or on the possibility of identifying the data subject. The latter are heavily dependent upon each specific context. However, the threshold is quite low, which means that in practice, it is very difficult for a piece of data not to be personal. When that is the case however, such non-personal data is referred to as “**anonymous data**” and escapes the reach of data protection (see Recital 26 GDPR).

17.4 Personal Scope of Data Protection: Controller and Processor

17.4.1 The Three Main Actors of Data Protection

In essence, one can argue that there are three key actors in data protection law:

- **The data subject** is an identified or identifiable natural person whose personal data are being processed.
- **Data controllers** (natural or legal persons) are the main actors and duty bearers. They are responsible and liable for compliance with data protection law.
- **Data processors** are separate natural or legal persons, which process personal data on behalf of the controller. Even though their responsibility is towards the data controller, the GDPR now allows administrative supervisory authorities to directly impose fines on them, and under certain conditions, they can also be liable to data subjects (see Rodway & Carey, 2018, p. 178).

17.4.2 Data Controllers

The GDPR defines a (data) controller as “the natural or legal person (...) which, alone or jointly with others, determines the **purposes and means** of the processing of personal data” (following Article 4(7), GDPR). A natural or legal person can refer to a sole person, a self-employed person, or companies such as banks, insurance companies, law firms, supermarkets, medical practices, and Internet search engines (see Welfare & Carey, 2018, p. 18). This means that an actor will qualify as a data controller if they determine either the purpose or the means, or both. The **purpose of the processing** is the overall reason or goal why the data is processed in the first place: this will be examined further in detail in this section. The GDPR is silent on what **the means of the processing** actually mean. According to the Art. 29 WP, these means must be understood as the **essential means** (Art. 29 WP, 2010, p. 14). These essential means refer to the most crucial and substantive choices that have to be made. They therefore include the following choices: which and how much data will be processed, for how long, who can have access to the data, what type of processing operations will be performed, how many data subjects are concerned, etc. (Art. 29 WP, 2010, p. 14).

The **non-essential means** are referred to as “organisational means” (Art. 29 WP, 2010, p. 14). They include, for instance, the choice of hardware and software and which employee of the company will operate the computers (Art. 29 WP, 2010, p. 14). These means can be determined by the data controller or by the data processor and have no bearing on the determination of the roles.

17.4.3 Data Processors

A **data processor** is “a natural or legal person (...) which processes personal data on behalf of the controller” (following Article 4(8), GDPR). In other words, the processor carries out the processing of personal data on account of the controller. The idea here is one of delegation: the processor will implement the instructions of the controller (Art. 29 WP, 2010, p. 25). One must keep in mind that the processor is a **separate** person. In theory, this means that delegating the processing to another employee within the same company does not qualify as data pro-

cessor (Art. 29 WP, 2010, p. 25). This can include a subcontractor for payroll administration, data storage, IT management, website hosting, a cloud computing supplier, or a computing centre (Rodway & Carey, 2018, p. 175; Voigt & von dem Busche, 2017, p. 20).

It should be noted that the GDPR requires that only processors providing sufficient guarantees to comply with GDPR can be selected as processors (Art. 28(1) GDPR). Furthermore, the relationship between a controller and a processor should be governed by a contract (Article 28(3) GDPR). The latter should mention some of the key elements of the processing: nature and purpose of processing, type of data, categories of data subjects, etc. Further, it must also contain a number of obligations for the processor, such as taking appropriate security measures and to be transparent with the controller (Article 28(3), GDPR).

17.4.4 Problematic Situations

The respective definitions of controller and processor might seem straightforward, but their application is not always easy. One can flag two issues: the grey zone between a controller and a processor and how to deal with a multiplicity of controllers.

17.4.4.1 Controller or Processor?

Even though the processor only processes on behalf of the data controller, there can be some grey zones. As the processor processes data, it can be brought to make a number of choices in terms of how the processing should be conducted. This can be the case when the processor has much more resources than the controller, but not only. Are these choices a mere implementation of the controller's mandate, or are these real choices concerning the essential means of the processing (which is the prerogative of the controller)? The key question here is the **degree of autonomy**: Was the processor still acting on behalf of the controller, or was it exerting its own influence on the processing (Rodway & Carey, 2018, p. 182; Voigt & von dem Busche, 2017, p. 19)? Would the data processing still have taken place if the processor had acted in the absence of instructions from the data controller (i.e. is there room for this type of discretionary decision-making) (Voigt & von dem Busche, 2017, p. 19)? In other words, who has the decisional power? If the answer to this question is positive, then the processor should be considered a controller as well, and we are facing a situation of **joint controllership**.

17.4.4.2 Multiple Controllers

Data controllers define the means and purpose of the processing "alone or jointly with others" (following Article 4(7), GDPR). The GDPR introduced the concept of joint controllership to refer to situations with multiple controllers (Article 26 GDPR). Given that one is a controller by defining either the purpose or the means of the processing, there is a broad typology of joint controllership. Joint controllers can have a close relationship and be bound by the same purpose (and essential means), have no common purpose but jointly determine the means, or have a looser

relationship by sharing only parts of the purpose and/or the means (Art. 29 WP, 2010, pp. 17–23).

As joint controllers, all actors are responsible for compliance with the GDPR (see Article 26(3), GDPR). The latter however requires that they allocate their responsibilities (i.e. “who does what?”) (Article 26(1), GDPR). The GDPR grants discretion on how in practice such responsibility should be shared. It only requires controllers to conclude an arrangement about it, and to make it transparent, in particular for data subjects (Article 26(1)(2), GDPR).

17.5 Art. 6, GDPR: The Need for a Legitimate Ground of Processing

If a data controller wants to start processing personal data, it must look into Article 6 of the GDPR. This article contains six alternative grounds on which the processing of data can be based. In other words, it is not possible to start processing personal data unless a ground for processing has been chosen. One can choose among the following grounds:

- (a) The data subject has **consented** to the processing of his/her personal data (see ► Sect. 17.5.1 of this chapter).
- (b) The processing is necessary for the performance of a **contract** to which the data subject is party (see ► Sect. 17.5.2).
- (c) The processing is necessary for compliance with a **legal obligation** (see ► Sect. 17.5.5).
- (d) The processing is necessary in order to protect the **vital interests** of the data subject (see ► Sect. 17.5.3).
- (e) The processing is necessary for the performance of a task carried out in the **public interest** (see ► Sect. 17.5.4).
- (f) The processing is necessary for the purposes of the **legitimate interests** pursued by the controller or by a third party except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject (see ► Sect. 17.5.6).

As one can see, all the grounds except **consent** require that the processing be necessary to the ground relied upon. The requirement of **necessity** implies that the ground cannot be achieved by other means that are less restrictive to the fundamental rights of the data subjects (see Art. 29 WP, 2014, p. 13).

17.5.1 Consent

Consent refers to the freely given, specific, informed, and unambiguous indication of the data subject’s wishes by which he/she, at the latest at the start of the processing, signifies agreement to the processing of his/her personal data (following Article 4(11), GDPR).

Table 17.1 Unspecific and specific purpose

Example of an unspecific purpose:	“Improving users’ experience”
Example of a specific purpose:	“We will share your shopping history with third parties to provide you with tailored content for future buys”
Source: Table compiled by the author, based on Art. 29 WP (2018b, p. 9)	

17.5.1.1 Consent Must Be Given in Relation to a Specific Purpose

A purpose is sufficiently specific if it is detailed enough so that the data subject can determine what kind of processing is and is not included under it. In other words, a processing will be covered by consent only if it is included in the purpose. The idea is to avoid the so-called blank consent (i.e. like a blank check, we have consented, but we do not know to what exactly) (Art. 29 WP, 2018a, pp. 11–12). An example of both an unspecific and a specific purpose can be found in **Table 17.1**.

17.5.1.2 Consent Must Be Informed

Data controllers must be transparent so that data subjects are able to understand what they consent to (Art. 29 WP, 2018a, p. 13). The principle of informed consent is twofold. On the one hand, it requires the data controller to provide the data subject with additional information. Examples of information that must be provided include the name of the data controller, the purpose of the processing, the types of data, and the types of processing (Art. 29 WP, 2018a, p. 13). On the other hand, it pertains to the quality of the information provided. The data subject must be able to **easily understand** what the controller says (Art. 29 WP, 2018a, pp. 13–14). This rules out the “legal jargon” that is still too often found in terms and conditions. Similarly, the information must be **easily accessible**. This means that the data subject must be able to easily find the relevant information (e.g. do not write it in super small fonts at the very bottom of the terms and conditions) (Art. 29 WP, 2018a, pp. 14–15).

17.5.1.3 Consent Must Be Unambiguous

The goal is that there must be no doubt about the data subject’s intention to consent (Art. 29 WP, 2018a, pp. 15–16). The consent itself can be given in any form as long as the data subject **actively** signifies agreement. There are various possible ways to provide consent, including in writing, a recorded oral statement (e.g. on the phone), box ticking, browser parameters, etc. (Art. 29 WP, 2018a, pp. 16–17). The example below shows an example of pre-ticked boxes for cookies:

► Pre-ticked Boxes

Pre-ticked boxes (known as opt-out) are not a valid way to express consent since there is no active consent on the part of the data subject. The only valid boxes are opt-in.

Example of invalid consent for cookies: “preferences” and “statistics” boxes are pre-ticked, like in **Fig. 17.1**. ◀

■ **Fig. 17.1** Invalid consent for cookies. Source: Author's own figure

This website uses cookies

Preferences Statistics Marketing

17.5.1.4 Consent Must Be Free

For the consent to be free, the data subject must have a **real choice** when consenting (see Art. 29 WP, 2018a, pp. 5–11). This is not the case when the data subject is compelled to consent, is subject to detriment, or has no real options. Consent will be compelled in cases characterised by power inequalities (i.e. when the parties are not in an equal bargaining position) (Art. 29 WP, 2018a, pp. 6–7). This is typically the case in the employment context (risk of job loss) or with public authorities (how can you refuse to the plans of the government?) (Art. 29 WP, 2018a, pp. 6–7).

Consent is subject to detriment when the refusal to consent leads to negative consequences such as fees, loss of service, or any sort of deception (e.g. a user revokes consent for special permissions on an app, and the app ceases to function properly and there is no objective reason for that) (Art. 29 WP, 2018a, pp. 8–9).

The absence of real choice points to issues of conditionality or “bundling”. This refers to situations where consent to the processing operation is bundled with a contract for the performance of a service, even though the consent is not necessary as such for the performance of the service, and should therefore be given separately (Art. 29 WP, 2018a, p. 8). One should be able to enter into a contract for the acquisition of the service and to separately consent to all the additional data processing. The goal here is to avoid using data as a way to pay for, as a counter performance, for digital services (Art. 29 WP, 2018a, pp. 8–9).

► Bundling of Consent

A mobile app for photo editing asks its users to have their GPS localisation activated for the use of its services. The app also tells its users that it will use the collected data for behavioural advertising purposes.

Neither geo-localisation nor online behavioural advertising is necessary for the provision of the photo editing service and to go beyond the delivery of the core service provided. Hence, this type of consent is not valid.

Note: Example based on Art. 29 WP (2018a, p. 6). ◀

17.5.1.5 Special Categories of Data: Explicit Consent

For certain categories of (sensitive) data such as those revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, and biometric data, a regular consent is not sufficient, and an explicit consent is required (Art. 9, GDPR). Since a regular consent already requires a “clear affirmative action”, it is not exactly clear what else is required from an explicit consent (Art. 29 WP, 2018a, p. 18). At the very least, the consent should be expressly confirmed in a written statement. A possible valid explicit con-

sent is for instance “I consent to the processing of data for purpose [x]”, which is sent by e-mail or a signed pre-written explicit consent, using an e-signature, etc. (see Art. 29 WP, 2018a, pp. 18–19).

17.5.2 Contract

A contract between a data subject and a data controller can serve as a basis for the processing of personal data in two cases. Consider a customer who buys goods at an online retail shop. First, the processing of data is **necessary** for **the execution** of the contract (Art. 29 WP, 2014, p. 16). See the example in ■ Table 17.2.

The second case is when the processing is necessary at the **pre-contractual stage** (i.e. so that the contract can exist). This is only valid if it is at the request of the data subject (Art. 29 WP, 2014, p. 18). See the example in ■ Table 17.3.

■ **Table 17.2** Processing of data is necessary for the execution of a contract

	Valid	Invalid
Necessary processing	Name and address of the customer, types and amount of articles purchased, method of payment, shipping information	Parents' name, partner's age, other online shopping habits
Execution of contract	Deliver the products to the customer	Debt collection, going to court in case of dispute related to the contract

Source: Table compiled by the author, following Voigt and Von dem Busche (2017, p. 102)

■ **Table 17.3** Processing of data is necessary at the pre-contractual stage

	Valid	Invalid
Pre-contractual steps (1)	Send online advertisements to a customer who wants more information about the seller's products	
Pre-contractual steps (2)		Profiling-based direct marketing unbeknownst to the data subject

Source: Table compiled by the author, following Voigt and Von dem Busche (2017, p. 102)

17.5.3 Vital Interests of the Data Subject

This ground—the vital interests of the data subject—is marginal and residual (i.e. when other grounds cannot be relied upon) (Art. 29 WP, 2014, p. 20). It targets situations of life and death of the data subject such as humanitarian purposes, monitoring epidemics, and natural and man-made disasters. With increasing global warming and refugee issues, it might gain importance in the future.

17.5.4 Performance of a Task Carried Out in the Public Interest or in the Exercise of Official Authority Vested in the Control

Public bodies provide public services (e.g. education, transport). For these so-called **public interest tasks**, they might need to process personal data. This can be the case for a specific task (e.g. setting up a new electronic ID scheme by the government) or pursuant to their general competence (for instance, in order to correctly perform their duty, tax authorities need to process an individual's tax return in order to establish the amount of taxes to be paid) (Art. 29 WP, 2014, pp. 21–23).

With the increasing privatisation of public services, the notion of public authority has expanded to include bodies that are subject to hybrid private-public law regimes (e.g. railway companies), or in certain cases fully private bodies that still exert a public interest task (e.g. medical professional association) (see Art. 29 WP, 2014, p. 22).

Public authorities derive their authority and competence from national (or EU) law (or even an administrative act, the notion of law being rather large). Indeed, a public institution can only exist if there is a law that provides for it (or for its specific competence) (Voigt & von dem Busche, 2017, pp. 107–108). Such law must possess certain characteristics. In particular, the law must be compliant with data protection law, and it should grant competences to the public authority that are proportionate to the aim pursued (Voigt & von dem Busche, 2017, p. 108). So a law granting the power to tax authorities to web scrape social media accounts of its citizens in order to detect cues of tax fraud is probably not proportional.

17.5.5 Compliance with a Legal Obligation to Which the Controller Is Subject

Law imposes obligations on all of us. Sometimes, in order to comply with **legal obligations**, data controllers will have to process personal data.

This ground, also based on a law, just like the previous one, is however more stringent. For the processing to be necessary for this ground, the controller must have no choice but to process the data (i.e. no discretion) (Art. 29 WP, 2014, p. 19). This has consequences on the quality of the law at stake. In addition to fulfilling all the requirements seen under the previous ground, the law leading to a legal obliga-

tion to process personal data must also be sufficiently clear as to the processing of personal data it requires (Art. 29 WP, 2014, p. 19).

17.5.6 Legitimate Interests of the Data Controller or a Third Party

The last ground upon which data controllers can rely in order to initiate the processing of personal data involves a **balancing of interests**. Questions that are relevant include the following: “Which interests weigh more?” “Those of the data controller (or a third party) or those of the data subject?” Depending upon the answer of this very subjective and context-dependent question, it will be possible (or not) to process personal data. This ground is a complex one and requires further investigation.

17.5.6.1 The Interest of the Data Controller: A Legitimate One

For a data controller to have an interest means that he/she has a stake, a benefit in the processing (Art. 29 WP, 2014, p. 24). Such interest must be clearly articulated (i.e. clear to understand) and should be real and present (or not speculative). That is, the interest must correspond to the current activities of the data controller or those that can realistically be expected in the very near future (in other words, one cannot start processing data because maybe in 2 years they will have an interest that justifies it) (Art. 29 WP, 2014, p. 24). The interest can also be that of a third party, which is crucial for a lot of companies which process data on behalf of their clients (Voigt & von dem Busche, 2017, p. 105).

Most importantly however, the data controller’s interest must be **legitimate**. This means that it must be lawful (i.e. in accordance with the law): not only with data protection law, but also with laws in general (including legislation, judgments, codes of conduct), and beyond, with ethics and societal expectations of what it is legitimate to process (Art. 29 WP, 2014, p. 25). Note that this social context and social values can change over time (Art. 29 WP, 2014, p. 25). For instance, when the widespread use of closed-circuit television (CCTVs) for surveillance purposes in cities began in the 1990s, it was far from considered legitimate (see, e.g., Coleman & McCahill, 2011, p. 146). Nowadays however, we are all used to CCTVs, and the debate has moved to more intrusive forms of surveillance such as facial recognition.³ The Art. 29 WP provides various examples of interests that are legitimate or not (Art. 29 WP, 2014, p. 25, 63, 68):

Legitimate Interest:

- Exercise of the right of freedom of expression and/or information by newspaper or NGO
- Conventional direct marketing
- IT network security

³ See, e.g., ► https://edps.europa.eu/press-publications/press-news/blog/facial-recognition-solution-search-problem_en, last accessed 29 June 2020.

Illegitimate Interest:

- Employee monitoring for verification of productivity
- Combination of personal information across web services

17.5.6.2 Interests or Fundamental Rights of Data Subject

Data subject's interest includes all their fundamental rights and freedoms (e.g. privacy, non-discrimination, fair trial) and does not need to be legitimate to be recognised (Art. 29 WP, 2014, pp. 29–30). The threshold is therefore lower and applies even in cases where the data subject has potentially engaged in an illegal activity. For instance, the illegal downloading of copyrighted material does not justify as such the surveillance of a data subject's Internet traffic (Art. 29 WP, 2014, pp. 29–30). A balance of the various interests at play must still be performed.

17.5.6.3 Balancing of Interests

In order to determine whether the interest of the data controller is sufficiently legitimate, it must be **balanced (or weighed)** against the interest of the data subject. The balancing of interest will determine which interest carries more weight and, therefore, whether the processing operation can take place (Art. 29 WP, 2014, p. 30). The balancing of interests is concretely done through a number of steps, which are explained below.

Step 1: Qualify the Interests

The first step is to **qualify the interests**. Are they very serious and compelling, or just trivial? As far as the **data controller** is concerned, one can refer to the classification as depicted in **Table 17.4**.

As far as the data subject is concerned, their interest is always high since the processing of personal data involves their fundamental right by definition (see the introduction of this chapter). Which is why one must have a look at the potential impact the planned processing operation will have on their interests and rights (second step).

Table 17.4 Qualification of interests

Category of interest	Seriousness	Example
Fundamental right	Very	Investigative journalism
Public interest	Medium	Medical research
Personal interest	Low	Private profit

Source: Table compiled by the author, building upon Art. 29 WP (2014, pp. 34–36)

Step 2: Impact(s) on the Data Subject

An **impact** can be defined as “the various ways in which an individual may be affected—positively or negatively—by the processing of his or her personal data” (Art. 29 WP, 2014 p. 37). Impact can be of different nature. They can be emotional/moral (e.g. fear, distress, reputation), material (e.g. financial loss, employment or price discrimination, physical), political (chilling effect, self-censorship), etc. (see, Art. 29 WP, 2014, p. 37).

Taken as such, these impacts might seem difficult to apprehend. One can therefore look at a number of factors that will render their appraisal smoother.

Step 3: Factors for Appraising the Impacts

In order to better determine what the impact is, one can look at the following risk factors. One can look at the nature or type of personal data being processed (Art. 29 WP, 2014, p. 38). The more sensitive the data, the higher the impact. Sensitive data can refer to the special categories of data enshrined in the GDPR (health-related data, political affiliation, etc.), but can also be sensitive in the general sense (e.g. children’s data, precise location data). Conversely, some data can be considered as less sensitive such as data that the data subject has already made publicly available (e.g. professional online profile) (Art. 29 WP, 2014, p. 38).

Another factor is the type of processing at play (Art. 29 WP, 2014, p. 39). This includes a variety of factors such as the amount of data subjects, the amount of data, or the variety of data processed. It also includes the amount of data controllers and/or processors with whom the data is shared and who can process the data. Finally, what kind of processing operation is performed on the data? Is it subject to simple and relatively benign operations (such as collection and sharing), or is it integrated into high-dimensional databases (thus combined with other data), and is further subject to advanced analytics (Art. 29 WP, 2014, p. 39)?

Another type of factor is the likelihood. On the one hand, a very likely impact signals a high impact. On the other hand, a highly uncertain impact can also signal a high impact (since we have very little clue whether it will happen or not) (Art. 29 WP, 2014, p. 38).

Finally, the last type of factor is the reasonable expectations of the data subject (Art. 29 WP, 2014, p. 40). This is an important notion in data protection law. It refers to what a data subject can reasonably expect in a specific context about what will happen to their data. That is, would they be surprised if their data were processed or underwent a given processing operation (Dehon & Carey, 2018, p. 59)? In other words, this refers to the context from the data subject viewpoint. In order to determine the data subject’s reasonable expectations, one can look at the following sub-factors. What is the relationship, the balance of power, between the data subject and the data controller (or what is their respective status)? Are they involved in an employment relation? Is the data controller a public authority? Is the data controller providing a service in a quasi-monopolistic situation (e.g. popular social networking service), or is it on the contrary a small company with very little bargaining power? Is the data subject himself/herself vulnerable (e.g. child, asylum

mentally ill, elderly)? One can also look at the legal or contractual obligations existing between data subject and data controller: medical doctors or attorneys are subject to confidentiality obligations, and a contract might also provide for similar confidentiality duties. In general, the more specific and restrictive the context of collection, the more limited the data subject's reasonable expectations (Art. 29 WP, 2014, pp. 40–41).

Step 4: Provisional Balance

At this point, one can make a **first balancing** between the legitimate interest of the data controller and the impacts upon the data subject's interests and fundamental rights (Art. 29 WP, 2014, p. 41). The act of balancing itself is—like data science—more art than science. There is no “objective rule” that can guide the data controller to determine which elements of the balance have the heaviest weight. It is a contextual, case-by-case decision that has to be made on the basis of the factors described herein above. In view of the various factors at play, it should be clear that not all impacts have the same weight. Some cases are clear-cut, while others might rely upon a “rule of thumb” type of decision, knowing that another person might opt for the opposite solution if placed in the same situation. This is why providing an explanation of the decision and keeping a record thereof are crucial (Art. 29 WP, 2014, p. 43). In any case, in case of doubt, it is recommended to balance in favour of the data subject (Dehon & Carey, 2018, p. 58).

Step 5: Additional Safeguards

As mentioned, the balance struck is only provisional. This is so because if it tilts in favour of the data subject, it is still possible to improve the situation by resorting to so-called **safeguards**. Safeguards can be understood as additional legal mechanisms that provide further protection to the data subject, and in so doing, they might help tilt the balance in favour of the data controller (Art. 29 WP, 2014, pp. 41–42). These additional safeguards should not be seen as a “silver-bullet” solution. The heavier the impact, the more safeguards one will need, and there is a chance that they do not manage to tilt the balance back in favour of the data controller (Art. 29 WP, 2014, p. 42).

Some of these safeguards are already part of the GDPR (and will be examined herein below). They consist of an “enhanced” application of existing provisions: for instance, providing more transparency than is required normally in order to make it clear to the data subject why the legitimate interest ground is chosen, how the impacts have been assessed and balanced, etc. Another possibility is to reduce the amount of data collected beyond what is prescribed so as to reduce the impact on the data subject. Another possibility is to pseudonymise the data (Art. 29 WP, 2014, p. 42).

Other safeguards are not as such part of the GDPR. They include for instance the possibility to include an unconditional opt-out mechanism (the data subject ends the processing operation without any condition), or the creation of confidentiality clauses (the data controller will not share the data with others) (Art. 29 WP, 2014, pp. 42–43).

Step 6: Final Balance

At this point, a **final balance** is struck. It follows the same principles as the provisional balance and will determine whether the safeguards have sufficiently reduced the impacts so that the balance now tilts in favour of the data controller.

17.6 Art. 5 GDPR: Principles to Be Applied to the Processing of Data

Once an adequate ground for processing has been found, it is then possible to start **processing personal data**. This is where Article 5 of the GDPR comes into play. It contains a number of principles that will have to be respected if the processing is to be legal.


17.6.1 Purpose Limitation Principle

Article 5(1)(b) of the GDPR states that data shall be collected only for specified, explicit, and legitimate purposes, and not processed in any manner that is incompatible with those purposes. This principle can be deconstructed into two types of requirements. The first type of requirement concerns the **initial purpose**: it must be specific, explicit, and legitimate (“purpose specification”). The second type of requirement concerns the processing of **already collected data for a new purpose**: this new purpose must be compatible with the initial purpose of collection (“purpose limitation *stricto sensu*”).

17.6.1.1 Purpose Specification: Why?

It is necessary to start the processing operation by determining and specifying its **purpose** because otherwise that would mean that we can process data without a reason. The processing must always be necessary with regard to a purpose (Art. 29 WP, 2013, p. 11). Also, a number of other requirements rely upon the fact that the purpose is well defined (see herein below).

17.6.1.2 Specific Purpose

The first element of the purpose specification requirement is that the purpose is **clearly and specifically identified**. This means that the purpose is defined in a way that is sufficiently specific, with a sufficient level of detail, so that it is possible to determine what processing operations are included under it (Art. 29 WP, 2013, p. 15). One can find an example of both a vague and a specific purpose in  Table 17.5.

Specifying the purpose might entail the provision of additional information such as the type of processing operation, its duration, or the type of data at stake (Art. 29 WP, 2013, p. 16). One key issue is that of “**umbrella purposes**”. This refers to situations whereby a number of purposes are regrouped under one single broader

■ **Table 17.5** Vague and specific purpose

Vague purpose	“Improving users’ experience”
Specific purpose	“We will share your shopping history with third parties to provide you with tailored content for future buys”

Source: Table compiled by the author, based on Art. 29 WP (2018b, p. 9)

purpose. This can make sense if the various processing operations are linked together (although sufficient detail should still be provided for each of them). What should be avoided is the use of an umbrella purpose to justify various purposes that are not related (Art. 29 WP, 2013, p. 16).

17.6.1.3 Explicit Purpose

For the purpose to be explicit means that it should be clearly revealed, explained, or expressed in some intelligible form (Art. 29 WP, 2013, p. 17). Whereas the requirement that the purpose be specific is to be constructed from the data controller perspective (they have to make it specific), this requirement is constructed from the data subject perspective: the data subject should have no difficulty in understanding what the purpose is (Art. 29 WP, 2013, p. 17). For this reason, this requirement can be deconstructed into two sub-requirements.

First of all, the purpose should be **easily understandable** (Art. 29 WP, 2013, p. 17). This requirement concerns the quality of the language used, which must be understood by everyone (all the potential data subjects, data protection authorities, etc.). Data controllers should take into account the fact that they might have different data subjects with different needs (e.g. children, elderly, people with different literacy skills). In order to ensure that the language is clear and unambiguous, the data controller should make sure that there is sufficient detail (but not too much) and that the language itself is clear and plain. The perfect “bad example” is terms and conditions which are overly long and rely upon sophisticated legal jargon (Art. 29 WP, 2013, p. 17).

Secondly, the purpose should be **easily accessible** (Art. 29 WP, 2013, p. 18). This requirement concerns the ease with which data subjects can find the information, which should be clear and distinguishable. Taking the example of the terms and conditions of a service, this means that the data subject should have no difficulty in finding the information that renders the purpose explicit (highlighted, in bold, a link clearly evidenced, etc.) (Art. 29 WP, 2013, p. 18, 51–55).

17.6.1.4 Legitimate Purpose

The last element of the purpose specification requirement is that the purpose be legitimate. This means that the purpose must be lawful (i.e. in accordance with the law), not only with data protection law, but also with laws in general (including

legislation, judgements, codes of conduct). Beyond this, it must comply with ethics and societal norms, which are contextual and can change over time (Art. 29 WP, 2013, pp. 19–20).

17.6.1.5 Different Purpose

This principle concerns the situation whereby the data is collected for one purpose, but the data controller then wants to use it for a different purpose (e.g. the other purpose was not considered at the time of collection). Normally, this should not be possible if the new purpose is not encompassed within the original purpose (“purpose limitation *stricto sensu*”). However, the principle of purpose limitation says that such further processing will be possible as long as this new purpose is **compatible** with the purpose of collection. This means that the data controller will need to perform a so-called **compatibility test**. The main idea is to look at the relationship between the purposes. The closer the relationship, the higher the chance that the purposes are compatible (Art. 29 WP, 2013, p. 21).

The compatibility test is undertaken in a number of steps (see also Art. 6(4), GDPR). The first step is to determine whether there is an **obvious link** between purposes. This will be the case when there is some overlap between the purposes: for instance, if the further processing was already more or less implied in the initial purposes, or if there is a link (even if partial) between purposes (Art. 29 WP, 2013, p. 22).

If there is **no obvious link**, a more thorough test will have to be undertaken. As shown in ■ Table 17.6, it is similar to the balancing test performed under the legitimate interest ground for processing (see ► Sect. 17.5.6 of this chapter). The com-

■ Table 17.6 Balancing tests compared

Steps for balancing the interests	
1. Assessing the impact on the data subject: <ul style="list-style-type: none"> (a) List of impact on the data subject (b) Factors for appraising the impact: <ul style="list-style-type: none"> – Type of data – Type of processing – Likelihood of impacts – Reasonable expectations of the data subject 	1. Assessing the compatibility between purposes <ul style="list-style-type: none"> (a) Link between purposes <ul style="list-style-type: none"> – Obvious: no need to go further – Not obvious: need to look at other steps (b) Reasonable expectations of the data subject (c) List of impact on the data subject (d) Factors for appraising the impact <ul style="list-style-type: none"> – Type of data – Type of processing – Likelihood of impacts
2. Provisional balance	
3. Additional safeguards	
4. Final balance	
Source: Table compiled by the author	

patibility assessment is nearly identical to this balancing test. In addition to the determination of the existence of an obvious link between purposes, also the order of the steps for balancing the interests changes somewhat in the compatibility assessment. This is because the object of assessment changes: instead of assessing the impact on the data subjects, one must assess the compatibility between purposes.

One way of assessing such compatibility is to look at the reasonable expectations of the data subject in context (Art. 29 WP 2013, pp. 24–25). The second step will be to look at and list the impacts on the data subject (Art. 29 WP, 2013, pp. 25–26). These impacts can be better appraised through the same factors (i.e. type of data, type of processing, likelihood of impact), minus the reasonable expectations of the data subject of course (see, Art. 29 WP, 2013, p. 26).

At this point, the assessment of compatibility is performed and one can also perform a provisional balancing between the two purposes: Are the two purposes sufficiently compatible so that the further processing can take place (Art. 29 WP, 2013, p. 26)? Again, keep in mind that things are on a “spectrum” and that this is a “rule of thumb” type of decision, knowing that another person might opt for the opposite solution if placed in the same situation.

The next step is of course the use of additional safeguards, which could reduce the impact on the data subject. The same remarks apply as for the legitimate interest ground, keeping in mind that what is considered the most adequate safeguard will change depending upon the context (Art. 29 WP, 2013, pp. 26–27).

At this point, a final balance is struck. It follows the same principles as the provisional balance and will determine whether the safeguards have sufficiently reduced the impacts so that the balance now tilts in favour of the further processing. An overview is provided in ■ Table 17.6.

If the balancing tilts in favour of the further processing, then the processing can simply take place. In case the balancing tilts against the further processing, the latter cannot take place (Art. 29 WP, 2013, p. 36). This means that a new ground for processing should be found. Note that **not all grounds have the same weight** in this regard. Consent is widely regarded as the safest, whereas relying upon the legitimate interest of the data controller can be considered shaky and uncertain (see Article 6(4) GDPR). Furthermore, the GDPR itself provides exceptions. It is possible to further process for an **incompatible purpose**, when this purpose is one of the following: archiving in the public interest, scientific or historical research purposes, and statistical purposes (Article 5(1)(b), GDPR). This exception requires additional explanations, which go beyond the scope of this chapter.

17.6.2 Data Minimisation

Article 5(1)(c) of the GDPR provides that the processing of personal data should be “adequate, relevant, and **limited to what is necessary**” in relation to the purposes of processing. In other words, the data controller should minimise the data processing as much as possible in a way that still enables them to achieve the purpose of the processing (Voigt & von dem Busche, 2017, p. 90).

There are various ways to minimise the processing of data. One can limit the amount of data initially collected, limit the type of data, limit the number of data subjects, limit the number of people who have access to the data, limit the type of processing operations performed on the data, etc. (also see Carey, 2018a, pp. 35–36).

17.6.3 Storage Limitation

Article 5(1)(e) of the GDPR provides that the personal data shall not be kept for longer than necessary for the purpose of the processing. This principle can be seen as the continuation of the **data minimisation principle** as far as the **storage duration** of the data is concerned. The point is that the data should not be kept longer as soon as the purpose for which they were collected in the first place is achieved. They can either be deleted/destroyed or anonymised (in this case, make sure that anonymisation is irreversible).

It can be difficult to predict in advance for how long it will be necessary to store the data. In order to avoid keeping the data “just in case”, it is recommended to establish a “**storage policy**”. The latter determines the time limits for the storage, and these limits are subject to periodic review (Recital 39, GDPR, also see Carey, 2018a, p. 38).

17.6.4 Additional Obligations

As a way to conclude this overview of Article 5 of the GDPR, one can also briefly mention the remaining principles.

17.6.4.1 Data Accuracy

Following Article 5(1)(d) of the GDPR, data controllers must make sure that the data is accurate and, where necessary, kept up to date. This is to make sure that the data in possession of the data controller soundly reflect the reality of the data subject (also see Voigt & von dem Busche, 2017, p. 91).

17.6.4.2 Lawfulness, Fairness, and Transparency

Following Article 5(1)(a) of the GDPR, the processing can only take place if it is legal in the general sense and if it complies with data protection law (see Carey, 2018a, p. 33). Further, the requirement of fairness entails to be fair to the data subject among others by taking into account their reasonable expectations as to what the processing entails (see Dehon & Carey, 2018, p. 43). The principle of transparency is twofold (Dehon & Carey, 2018, p. 44). On the one hand, it aims to make sure that the data subject is adequately informed about the processing and therefore requires to provide them with a minimum amount of information concerning the identity of the data controller, the processing purpose, the type of data collected, etc. On the other hand, it pertains to the quality of the information,

which must be both easily understandable and easily accessible (also see ► Sects. 17.5.1 and 17.6.1 of this chapter).

17.6.4.3 Integrity and Confidentiality

Following Article 5(1)(f) of the GDPR, data controllers should ensure an adequate level of security of the data they process, which includes protection against unauthorised or unlawful processing, and against accidental loss, destruction, or damage (Voigt & von dem Busche, 2017, p. 92).

Conclusion

In order to be socially responsible, it seems paramount that data science complies with data protection law. Contrary to broad rights and values such as privacy that are sometimes difficult to delineate, data protection features a number of very concrete rules and principles. This chapter has shown that the notion of personal data is broader than what is commonly assumed to be the case. It has also shown that what is referred to as a data controller or a data processor is a technical notion that does not necessarily match the vernacular understanding we may project onto them. It has also inspected the grounds allowing the processing of personal data in detail. Special attention can be paid to what constitutes a valid consent. This is not as easy as we may believe. Equally, doing the balancing of the interests under the legitimate interest ground remains a delicate exercise that can be contested. Such balancing can also be found in the determination of what constitutes an acceptable further processing for a different purpose. This reminds us that adequately defining a purpose is key to data protection law as it will allow to assess the necessity and proportionality of the envisaged processing operation, which is itself key to a responsible practice of data science.

? Questions

1. What are the three ways in which information can relate to a data subject?
2. What are the two criteria that can design someone as a data controller?
3. What requirement applies to all the grounds for processing except consent?
4. Why is it important to specify the purpose of the processing operation?
5. What is the link between the principles of data minimisation and storage limitation?

✓ Answers

Example 1. Key questions

If Company X has only one purpose of processing, this amounts to a further processing for a different purpose. One must therefore do a compatibility test. Given the lack of obvious link between purpose, one must look at the data subject's reasonable expectations (no such expectations), impacts on the data subject (high: interference with their voting preference and political opinion), factors for appraising the

impact (very sensitive data, creation of profiles, transfer to third parties), and safeguards (none mentioned). This processing is therefore illegal.

End questions

1. It can relate in terms of content, purpose, or result (also referred to as impact).
2. The data controller is the actor who determines the purpose and/or the essential means of the processing. It suffices to fulfil one of these conditions to qualify as data controller (in which case it is most likely a joint controllership situation).
3. The necessity of the processing must be determined.
4. If there is no purpose (or no well-defined purpose), then this means that we can process data for any reason we want, without any limit. Also, a number of key requirements assessing the adequacy, relevance, or necessity of the processing can only be complied with if the purpose is well defined (e.g. necessity, data minimisation, storage limitation).
5. The storage limitation principle can be understood as the application of the data minimisation principle to the issue of the duration of the storage of data.

Take-Home Message

- Personal data is a broad notion that encompasses most of the data processed in contemporary data processing technologies.
- The distinction between a data controller and a data processor can be tricky.
- In order to start a processing of data, one has the choice between six different grounds; however, one ground must be chosen.
- When processing data, all the provisions of Art. 5 of the GDPR must be respected.

References

- Art. 29 WP. (2013). *Opinion 03/2013 on Purpose Limitation*.
- Art. 29 WP. (2014). *Opinion 06/2014 on the Notion of Legitimate Interests of the Data Controller under Article 7 of Directive 95/46/EC*.
- Art. 29 WP. (2007). *Opinion 4/2007 on the concept of personal data*.
- Art. 29 WP. (2010). *Opinion 1/2010 on the concepts of “controller” and “processor.”*
- Art. 29 WP. (2018a). *Article 29 Working Party Guidelines on consent under Regulation 2016/679*.
- Art. 29 WP. (2018b). *Guidelines on transparency under Regulation 2016/679*.
- Carey, P. (2018a). Data protection principles. In P. Carey (Ed.), *Data protection: A practical guide to UK and EU law* (5th ed., pp. 32–41). Oxford University Press.
- Coleman, R., & McCahill, M. (2011). *Surveillance & crime*. Sage.
- Dehon, E., & Carey, P. (2018). Fair, lawful, and transparent processing. In P. Carey (Ed.), *Data protection: A practical guide to UK and EU law* (5th ed., pp. 42–65). Oxford University Press.
- European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).
- Gellman, R. (2019). *FAIR INFORMATION PRACTICES: A Basic History*.
- Gutwirth, S. (2002). *Privacy and the Information Age* (Rowman & Littlefield 2002).
- Hoofnagle, C.J., Sloot, B van der., & Zuiderveen Borgesius, F. (2019). ‘The European Union General Data Protection Regulation: What It Is and What It Mean’s. *28 Information and Communications Technology Law* 65.

- Mourby, M., Mackey, E., Elliot, M., Gowans, H., Wallace, S. E., Bell, J., et al. (2018). Are “pseudonymised” data always personal data? Implications of the GDPR for administrative data research in the UK. *Computer Law and Security Review*, 34(2), 222–233.
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), [2016], *OJ L 119/1*.
- Rodway, S., & Carey, P. (2018). Outsourcing personal data processing. In P. Carey (Ed.), *Data protection: A practical guide to UK and EU law* (5th ed., pp. 175–183). Oxford University Press.
- Voigt, P., & von dem Busche, A. (2017). *The EU General Data Protection Regulation (GDPR), a practical guide*. Springer.
- Welfare, D., & Carey, P. (2018). Territorial scope and terminology. In P. Carey (Ed.), *Data protection: A Practical guide to UK and EU law* (5th ed., pp. 1–31). Oxford University Press.



Perspectives from Intellectual Property Law

Lisa van Dongen

Contents

- 18.1 Introduction – 442**
- 18.2 Meeting the Criteria – 443**
 - 18.2.1 The Formal Requirements of Copyright – 443
 - 18.2.2 *Sui Generis* Database Right – 445
 - 18.2.3 Trade Secret Right – 446
 - 18.2.4 Summary – 447
- 18.3 The Scope of Protection – 448**
 - 18.3.1 Copyright: Protected Subject Matter – 448
 - 18.3.2 *Sui Generis* Database Protection – 449
 - 18.3.3 Trade Secret Right – 451
 - 18.3.4 Summary – 451
- 18.4 Exceptions and Limitations – 452**
 - 18.4.1 Limitations of the Rights – 452
 - 18.4.2 Exceptions: Common Ground – 454
 - 18.4.3 Exceptions Specific to the Right – 454
- 18.5 Alternative Sources – 455**
 - Further Reading – 458**

Learning Objectives

- What copyright, sui generis database rights, and trade secrets entail and how to determine their beneficiaries.
- When and how the use of third-party datasets is restricted by these rights and when not.
- The potential and limitations of alternative sources to complement or substitute third-party datasets, such as data portability rights and public sector information.

18.1 Introduction

More and more information is collected via the use of smart devices (e.g., smart thermostat, smart phone), internet services (e.g., Google and Facebook), sensors (e.g., in cars, smart homes, and cities), and cameras. The resulting datasets contain a lot of information about individuals, but also about society at large. These datasets allow their observers to spot problems and explore ways to address them, but also to spot opportunities and explore how to exploit them. For example, by studying information from the sensors of cars, the sensors and cameras pointed at the roads, and traffic light systems, it is possible to identify the causes of car accidents and propose solutions to decrease the number of accidents on a certain block. However, the access to such datasets generated by others is often restricted. There is a big group of actors who do not want others to use “their” data. A very important factor that helps such actors restrict access to their datasets is intellectual property law. The holder of intellectual property rights on a dataset has the ability to restrict access of everyone else to (parts of) his/her dataset, as well as impose limits on its use. To understand how to navigate this field of law, it is important to first understand what purpose intellectual property rights serve.

As articulated in the Enforcement Directive, the main underlying reason in current intellectual property law systems is incentivizing (investment in) innovation. Intellectual property rights have been created as artificial property rights to correct certain market failures. Think of the market as a field filled with fruits. If everyone is free to use the field and its fruits without any restrictions, it is likely that many will do so. What is unlikely, however, is that everyone using the field will also individually invest in it. This is due to the uncertainty that it will yield them any results or even allow them to recoup their investment; after all, everyone is free to use the field without restrictions. When the projected proceeds are smaller, less people will be willing to invest. Moreover, any investments that are made are likely to be smaller. This is where intellectual property rights come in. They are tools to correct this “market failure” by rewarding those who invest in innovation with a set of exclusive rights for a limited time. These rights are tools for the right holder to legally restrict the access and use of his/her intellectual property. This allows the right holder to charge higher prices to recoup investments and make a profit.

This chapter aims to provide an introduction to the basics of intellectual property rights in the EU. It uses simplifications and does not always provide the entire picture to maximize understanding of the material. Such simplifications are generally pointed out, and sources on the topic have been included in the references for those wishing to gain a deeper understanding of such an underdeveloped concept. It is, therefore, not to be used as a substitute for legal advice or as a basis for academic debates. Furthermore, while there are many different types of intellectual property rights, only **copyrights**, ***sui generis* database rights**, and **trade secrets** will be discussed here. Under the legal framework of the EU, the subject matter and conditions of these intellectual property rights are closely related to data and software, as will become apparent further in the chapter. Other rights such as patents currently play a more complicated role in the EU in data and software *inter alia* due to limitations in patentability of subject matter such as mathematical methods and computer programs as such. Such limitations have also started playing more of a role in, for instance, the United States, as can be inferred from the case law of their Supreme Court (i.e., on the “abstract idea” concept) between 2010 and 2014. This subject thus requires more attention than it could receive in this limited contribution.

This chapter thus focuses only on these particular intellectual property rights from an EU perspective. The questions explored in the following sections will focus on establishing for each of these intellectual property rights when it would be applicable (► Sect. 18.2), followed by what this means for the data’s usage by a third party (► Sect. 18.3), as well as limitations and exceptions (► Sect. 18.4). This chapter concludes by discussion of ways to gain lawful access to datasets covered by one or more of these intellectual property rights and alternative sources.

18.2 Meeting the Criteria

18.2.1 The Formal Requirements of Copyright

Something might be protected by copyright if it meets the three cumulative criteria for copyright protection. Following the Berne Convention, these criteria require that it is (1) an expression (2) that is original (3) in the area of literature and art. In the EU’s copyright regime, factors such as labor or investment are not relevant. There are three elements of a dataset that are capable of meeting these requirements:

- The contents of the dataset
- The selection of the data
- The arrangement of the data

If one or more of these elements meet the formal requirements, there might be copyright protection on those elements of the dataset. In that case, there would be legal restrictions on its use. It is thus important to understand these criteria to be able to determine the likelihood of copyright protection on a dataset to ensure lawful use.

The underlying premise of the criterion **expression** is that facts and ideas are not created but discovered. This is also confirmed in *Feist Publications, Inc., v. Rural Telephone Service Co.*, which shows that the United States and the European Union approach this criterion in a similar fashion. What this means for copyright is that it does not protect *what* is said, but *how* it is said. A good rule of thumb is looking at it as a spectrum in which facts and ideas are on one side and expressions on the other based on specificity. In principle, the more specific a fact or idea becomes, the closer the needle generally moves towards expression. The reasoning behind this is that an author can convey a fact or idea choosing his/her own words, thereby creating something both beyond and separate from the fact or idea. To illustrate, look at the difference in detail in the following sentences in **■ Table 18.1**.

This requirement is a possible hurdle for copyright protection on a dataset. For example, data in such datasets together may create a very specific picture, but if the data is merely displayed as variables in a table, the data lack expression.

In *Football DataCo Ltd.*, the second element—**originality**—was understood as a margin of discretion to make free and creative choices that is utilized. In simpler terms, it requires that the creator has put his/her personal stamp on it. However, this of course should not be taken literally. For instance, putting your logo on something does not make it original. The bar for meeting this criterion is not very high in practice. Such creative choices can be as simple as selecting lighting, a background and an angle for making a picture, or word choice in a text or code. It is, however, important to emphasize that there should be room to make such choices by the creator. For instance, a passport photo has to meet a

■ Table 18.1 Expression

Example sentence	Level of detail	Fact/idea/ expression
This house is green	Very little detail and very general	Fact/idea
This three-story house is three different shades of green	More detail, but still quite general	Fact/idea, but already more towards expression
This three-story living accommodation is a mix of shades of green, amongst which olive, moss, and even some hints of metallic green around the corners of its windows and doors	A lot of detail and very specific	Expression

Note to table: Author's own table

number of strict requirements. Such predetermined settings affect the room the photographer has to make his/her own creative decisions. For a passport photo, it is thus highly unlikely that the photographer would be able to meet the originality requirement. Another example is functionality requirements. Software code is capable of attracting copyright protection since the Software Directive came about, but, as confirmed in *Bezpečnostní softwarová asociace*, the expression in the code cannot amount to originality if “dictated by their technical function.”

The circumstance that the author has room to make creative choices is thus vital for meeting the originality requirement. Moreover, in the absence of such requirements, there is still the matter of whether creative choices are actually made. The selection and/or arrangement of data in a dataset can, for example, meet the minimum threshold of creativity, but these choices are generally made based on utility in practice; the choices made in selecting data are often determined by a company’s primary business, and the data are arranged for practical reasons such as by alphabetical order or by date.

The last criterion requires that it is a work in the area of **literature and art**. What constitutes art or literature is understood very broadly in the copyright regime. For instance, literature for the purpose of copyright protection can include essentially anything involving the written word. As mentioned above, it can even cover the code in software. This means that data—whether numeric or text—also falls within this broad category. Some other examples of works that may be protected are books, paintings, sketches, maps, architecture, preparatory design material for software code, films, musical compositions, lyrics, topography, choreographic works, and so forth: Article 2(1) of the Berne Convention contains well over 20 examples of types of works falling within the ambit of literature and art.

18.2.2 *Sui Generis* Database Right

If materials such as datasets and preparatory design material for software code are part of a database, their use may be restricted by *sui generis* database protection. Due to the limited protection provided by the copyright regime in databases, the Database Directive was adopted in 1996 to further strengthen the information economy in the EU. To this day, this regime is still very much a European creation (there is, for instance, no equivalent in the USA). The *sui generis* database right thus protects databases without originality. However, that does not mean that if there is copyright protection on the contents, selection, and/or arrangement of the database, there cannot also be *sui generis* database protection. The two rights can coexist on a single database. A dataset is likely to be covered by this right if it is (1) a database for which a (2) relevant investment was made (3) that is substantial.

For a dataset to satisfy the first condition—that it is a **database**—it is first required that the dataset is a collection or compilation of materials. Such materials include copyrighted works, numbers, facts, and data, but are not limited to those categories. Next, such materials must then be organized, stored, and accessible via electronic or nonelectronic means. This means that a written document meeting all the other requirements could also qualify as a database. However, for a physical database, it is not necessary that the materials are physically stored in an organized manner.

The second criterion requires that a **relevant investment** is made. This means that the investment must be made in the collection, verification, and/or presentation of data for the database. As clarified in *BHB v William Hill*, investment in other categories such as in the creation of data is not relevant for meeting this criterion. Such an investment can be made by way of financial resources, human resources, and material resources. Investment via human resources can, for example, be made in effort or time. For material resources, the investment is made in equipment to build the database such as hardware and software. Of course, such type of investments also cost money. Moreover, human input is generally required in operating equipment to make a database. In reality, the connection between these three types of investment thus often makes for a combination of the three with the emphasis on financial resources. Moreover, such investments should not have been made for other purposes. For instance, computers used to create the database are often not solely bought for that purpose. In that case, the investment generally does not count towards the coming into being of *sui generis* database protection.

The last criterion—that the investment must be **substantial**—is a bit more ambiguous. The Database Directive does not provide conclusive guidance on what this criterion means or how it should be applied. Case law so far has mostly dealt with high sums of financial investment, so these cases do not provide much guidance on the substantial threshold either. Unfortunately, the exact ceiling and floor of this criterion are also still subject to heavy academic debate, but it would be beyond the purposes of this chapter to include these. This threshold should, different than the word substantial might suggest, not be interpreted as “high.” Instead, this criterion is best understood as requiring an investment that is not too unsubstantial. These perimeters in the main text—not high, just not too unsubstantial—are generally accepted in EU member states such as Germany. A clear example of such an insubstantial investment would be a single employee of a big company devoting only a few hours to making the database. An example of something that would qualify would be investments in verifying a great quantity of data with another dataset.

18.2.3 Trade Secret Right

Following the Trade Secret Directive, if the dataset consists of (1) information not known in the relevant circles, (2) is of commercial value, and (3) is kept secret by the company in question, the dataset may be protected as a trade

secret. The first criterion requires that the information in question is **not readily accessible or known in the relevant circles**. The relevant circles refers to people generally dealing with this type of information, which means that the relevant circle may differ per type of information if the protected subject matter consists of different types of information. It can, therefore, not cover insignificant information or the kind attained through normal employment experience. Information that can be covered by a trade secret right at least includes know-how, business information, or technological information, but may be defined broader in domestic law.

Second, the information should be of **commercial value**. It does not matter whether it does so actually or potentially. What is important is that the interests of the right holder of the trade secret—whether scientific, technical, business, or financial in nature—would be harmed if the trade secret would be compromised. It should thus have commercial value because it is secret. If the value would not be affected negatively if it would be misappropriated, satisfaction of the second criterion is questionable.

Finally, the holder of the trade secret right should make reasonable efforts in **keeping the information secret**. Of course, this is subject to the circumstances of the case. In some cases, it might be more difficult to keep the information a secret or the circumstances may require different measures than in others. The fact that many people know does not necessarily mean that the company has failed in its effort to satisfy this criterion. For instance, many employees might require knowledge of (parts of) the trade secret in order to be able to make a product. As long as they are under contractual obligations to secrecy, it does not matter how many know. The same is true for distributors who have received certain information under a nondisclosure agreement to be able to do their job.

18.2.4 Summary

The formal requirements of each of the intellectual property rights can be broken down into three basic components. Put next to each other in a table, that creates the following picture (■ Table 18.2).

■ Table 18.2 Formal requirements

Copyright	<i>Sui generis</i> database right	Trade secret right
<ol style="list-style-type: none"> 1. Expression 2. Originality 3. Literature and art 	<ol style="list-style-type: none"> 1. Database 2. Relevant investment 3. Substantial 	<ol style="list-style-type: none"> 1. Not readily accessible or known in relevant circles 2. Commercial value 3. Kept secret

Note to table: Author's own table

18.3 The Scope of Protection

18.3.1 Copyright: Protected Subject Matter

If a dataset would be protected via one or more of these routes, there is still the limitation of what these rights actually protect and against what. When a dataset or software code meets the requirements for copyright protection, this protection is limited to the **original expression** only. This means that the protection can never extend to, amongst other things, factual content or ideas. Additionally, if only the selection and/or arrangement of a dataset are protected by copyright and not the data itself, the expression exists only in the selection and/or arrangement. For software code, this means that copyright can only rest on code *not* dictated by technical functions. A third party would thus be able to use the contents of the dataset or such unprotected parts of the software's code.

Moreover, copyright only protects the original expression against certain types of use by others. In other words, the copyright holder has certain rights to *exclude*. Different from what the term “copyright” suggests, it constitutes not one right but a bundle of rights. The bundle of rights contains exploitation rights, otherwise known as economic rights. There are several economic rights included in the InfoSoc Directive, but only the right to reproduction and the right to make public are of particular relevance for data usage and software. The **right to reproduction** entails that, in principle, only the copyright holder has the right to make copies of his/her work. Furthermore, it is important to note that a reproduction does not have to be exact. Making a photo of a painting is also reproducing the work. The means used to make a copy do not matter for this right. Furthermore, it does not have to be a copy of the entire work. What is important is that enough should be copied to display the intellectual and creative work of the artist. A sample as little as 11 words from newspaper articles has been found capable of doing that in *Infopaq v Danske Dagblades Forening*. Consequently, it is arguable that a small part of the dataset or code could also convey the creative choices of the author. If so, in the absence of an applicable exception, even the use of such small excerpts requires authorization. Second, there is the **right to make public**. Think, for instance, of putting a protected content on a website or using hyperlinks to protected content. Take care that this is somewhat oversimplified. What should be understood as making available to the public and who should be understood to be conducting this act are still evolving due to certain recent legislative and judicial developments at the EU level. In most cases, a reproduction of some sorts is necessary to be able to make it public. Notable exceptions here are the use of hyperlinks or displaying the original (i.e., a painting in a museum).

What this means for third parties is that they cannot lawfully engage in these uses of the original expression without authorization. The copyright holder can, for instance, grant others permission to reproduce his/her work via a license. Since 11 words could already convey creative choices from the author, the requirement of having to obtain authorization kicks in fast. In principle, such authorization can

only be obtained from the copyright holder. The **right holder** is generally a natural person—the author or creator. When a work has been created in assignment, the allocation of the copyright depends on who has made the creative choices. In some cases, the creative choices may have been made by several actors, which generally leads to shared rights to a work. However, this is different in case of creation under employment. For instance, the exploitation rights on a work are located with the employer if created by an employee in the course of his/her employment upon instructions by the employer. Additionally, in the case of software, the Publications Office clarified in their summary of the Enforcement Directive that EU member states may lay down that legal persons or entities may also be the right holder. In some jurisdictions, not all rights may always be transferable from the author to another.

18.3.2 *Sui Generis* Database Protection

The *sui generis* database right was created with the investor in mind, so just being the factual maker is insufficient to be the right holder. According to the Database Directive, the **right holder** is the person who takes the initiative and the risk of investing. Subcontractors and work for hire are explicitly excluded from this definition. If a database is made by an employee, the allocation of the rights depends on the criteria in national law. If multiple people or entities have contributed to a database, there might be joint rights. Unlike copyright, the *sui generis* database right is fully transferable. Like in copyright, the *sui generis* database right is not a single right. When a database is covered by *sui generis* database protection, the right holder has the exclusive rights to (1) extraction and (2) reutilization. These rights should be understood as follows.

Extraction refers to the transfer of the database or a substantial part thereof. This transfer may be permanent or temporary. Moreover, the means through which it is transferred do not matter. It is also irrelevant where the database is transferred to (type of medium). What matters is that the database or a substantial part thereof is transferred. This means that any person other than the right holder in principle requires the authorization from the right holder to perform this act lawfully. However, authorization is also required for systematic extraction of insubstantial parts. This is included in the definition of extraction to combat “milking.” This is the process of repeatedly transferring small parts of a database until the entire database or a substantial part thereof is transferred.

The other type of use—**reutilization**—refers to making the database or a substantial part thereof available to the public. This includes the distribution or renting of copies, online transmission of the database, and other types of transmissions. Any way in which the database is made public falls under this definition. In essence, this right thus gives the right holder the sole right to conduct an incidental reutilization of (a substantial part of) the database. However, just like the right to extraction, the right to reutilization also protects against the systematic reutilization of insubstantial parts. Again, if this definition were limited to substantial

parts or the entire database, this would provide third parties the opportunity to still communicate (a substantial part of) the database, just a smaller part at a time. Finally, there is one last instance in which there is reutilization. It involves the use of a meta search engine with certain functionalities.

A meta search engine is a search engine that makes it possible to search through a number of other databases. Generally, it transfers the search query that is inserted by a visitor of the meta search engine to other search engines. It does not copy anything from the databases through which it searches, but shows the results from the search, including those from other databases. It was established in *Innoweb BV v Wegener* that such a meta search engine is *likely* to reutilize (a substantial part of) the database if the three following functionalities are present. First, the search forms offered to the end user by the meta search engine and the other database function essentially the same. Second, the queries are translated for the end user in real time to other search engines. This means that all the information of the other database is searched through in real time after the end user of the meta search engine has initiated the search. Third and finally, the results are presented all together in an order that reflects similar criteria to those used by the other database. To this end, the format of the meta search engine's own website is used in showing the results, showing duplicates together as a block item. To reiterate, if a meta search engine that searches other databases functions in the aforementioned way, the operator of this meta search engine is likely to engage in the reutilization of (substantial parts of) another database. Of course, this does not mean that if a meta search engine does not have these characteristics, there could not nevertheless be reutilization.

For both of these rights, the word **substantial** plays a role again. For the purposes of extraction and reutilization, the term "substantial" refers to the volume of data from a database, more specifically, the volume of data that is extracted or reutilized in relation to the whole database (see *BHB v William Hill*). There is a link here between the investment and the two rights. The easy way to approach this is quantitatively. Consider the following example. There was substantial investment in the collection, verification, and/or presentation of the data, but no significant differences in the investment across the data. A third party now extracts half the data of the database. That means half of the investment is represented by the extracted part. The part that is extracted is thus likely to be substantial. However, whether the extraction or reutilization is substantial can also be tested qualitatively. This is a bit more ambiguous. The circumstances of our example change somewhat. Now, there is certain data in the database that has required much more investment in their collection, verification, and/or presentation than the rest of the data. The more "expensive data" is only a small part of the entire database. A third party now reutilizes only the part of the database that contains the "expensive data." Even though it is less data, it represents a bigger part of the investment. This means that it is likely that such reutilization by a third party would be qualitatively substantial. In both examples, the third party probably cannot conduct these acts without authorization from the right holder or by law.

18.3.3 Trade Secret Right

The Trade Secret Directive stipulates that the **trade secret holder** is any natural or legal person lawfully controlling the trade secret. Like the *sui generis* database right, this right can be fully transferred. The trade secret right protects against the unlawful acquisition, use, and/or disclosure of protected subject matter. These acts are to be construed very broadly. Any act contrary to honest commercial practices, unauthorized access, and/or appropriation of any material that contains the protected subject matter falls under unlawful **acquisition**. The same is true for material from which the trade secret information can be derived. Of course, if a person then proceeds to **use and/or disclose** the trade secret, this too would be unlawful. Use or disclosure of protected subject matter in breach of a contractual duty—including a confidentiality agreement—or any other duty imposing limits on those acts is also unlawful. Moreover, unlawful use includes the production of infringing goods, or offering or placing those on the market. Storing, importing, and exporting infringing goods to that end also fall within that definition. A good is infringing if the unlawfully acquired, used, or disclosed protected subject matter contributes in a meaningful way to (the production process or marketing of) a product.

The trade secret right is arguably the most fragile intellectual property right. When copyright or *sui generis* database right is infringed, these intellectual property rights will continue to exist. Once the data covered by a trade secret right is misappropriated in a way that it no longer satisfies the conditions regarding its secrecy, the right lapses. However, it is important to reiterate that the trade secret protects against *unauthorized* acts. Consider the following example. Data covered by a trade secret is disclosed under a nondisclosure agreement against payment. If the duties of the provider and acquirer—contractual and otherwise—do not prevent this transfer of data under the circumstances, it is likely to be lawful. Such disclosure presumably leaves the trade secret intact. Trading under a nondisclosure agreement does not necessarily result in the loss of the trade secret right. Contracts such as employee contracts with confidentiality clauses and nondisclosure agreements are thus vital tools for the holder of the trade secret right.

18.3.4 Summary

If a dataset qualifies for copyright, *sui generis* database protection, and/or a trade secret right, the protection is still limited to certain subject matter. Moreover, it is only protected against certain unlawful acts performed by someone other than the right holder (see ■ Table 18.3). Such acts are unlawful without authorization provided by the right holder or law (i.e., exception).

Table 18.3 Scope of protection

	Copyright	<i>Sui generis</i> database right	Trade secret right
Protected subject matter	Original expression	Database	Trade secret
Protected against unlawful	<ul style="list-style-type: none"> • Reproduction • Making public 	<ul style="list-style-type: none"> • Extraction • Reutilization 	<ul style="list-style-type: none"> • Acquisition • Use • Disclosure

Note to table: Author's own table

18.4 Exceptions and Limitations

18.4.1 Limitations of the Rights

In some cases, a use by a third party falls outside the scope of the right. Limitations, as the word suggests, limit the protection. For instance, intellectual property rights do not last indefinitely. In the EU, copyright lasts up to 70 years after the death of the author following the Term Directive. According to the Database Directive, *sui generis* database protection lasts for 15 years starting from the day of completion of the database, but the clock restarts with every new substantial change and/or investment. Trade secret rights are the exception here: there is no **maximum term of protection** inserted in the Trade Secret Directive. The trade secret right will last until its protected subject matter no longer satisfies the criteria for this right.

As aforementioned, copyright does not extend to **facts and ideas**. Moreover, even subject matter that is neither fact nor idea can fall outside the scope of the protection when it is not part of the original expression. Furthermore, originality means that the creative choices are made by the author, not that it should be new. This means that it does not protect against **independent creation**.

For the *sui generis* database right, protection revolves around the investment. If a third party incidentally **extracts** and/or **reutilizes insubstantial parts**, in principle, that would be lawful. However, there are some boundaries there as well. In doing so, the Database Directive requires the third party to take care that its acts do not conflict with the normal exploitation of the database by the right holder or unreasonably prejudice his/her interests. In short, acts by a third party should not "harm" the investment.

The scope of protection offered by the trade secret right also has its limitations under the Trade Secret Directive. The trade secret right only protects against unlawful acts. This means that **independent creation** or **discovery** does not interfere with trade secret rights. Moreover, **reverse engineering** after lawfully obtaining a prod-

uct would also not breach the trade secret right. This means that it would be lawful for a third party to buy a product that was brought on the market in the EU by the right holder and study its functioning to improve his/her own production process, product, and/or service. A car manufacturer could, for instance, buy a car sensor offered on the market by a competitor to reverse engineer it and use the gained knowledge to improve its own car sensors.

Finally, several references have been made to the transferability of rights allocated to the right holder by these intellectual property rights. What this means is that it is generally possible to contractually “reserve” or transfer such rights or allow acts under certain circumstances. Right holders themselves can thus also contractually limit their own rights. For the reservation of rights, think, for example, of a situation of shared rights. It could be beneficial for parties to lay down contractually that the authorization of *all* right holders must be obtained, not just one. Alternatively, a right holder could transfer the sole right to an exclusive distributor to enforce the intellectual property right against (alleged) infringers, thereby freeing his/her own hands. An example of allowing acts under certain conditions can be found in many terms of service in the gaming industry. Such terms often contain a clause allowing their users to engage in acts such as live streaming themselves playing the game in question. Another fitting example is the use of a threshold, allowing users to use protected material as long as they do not gain profit over a certain established number or reach a set number of clients (■ Table 18.4).

■ Table 18.4 Limitations

	Copyright	<i>Sui generis</i> database right	Trade secret right
Maximum term	70 years after the author's death	15 years, but renewable	–
Outside the scope	<ul style="list-style-type: none"> • Facts • Ideas • Independent creation 	<ul style="list-style-type: none"> • Extraction of insubstantial parts • Reutilization of insubstantial parts 	<ul style="list-style-type: none"> • Independent creation • Independent discovery • Reverse engineering
Contractual limitations possible	<ul style="list-style-type: none"> • Yes, on exploitation rights^a 	<ul style="list-style-type: none"> • Yes 	<ul style="list-style-type: none"> • Yes

Note to table: Author's own table

^a As mentioned earlier, the bundle of rights is not always transferable in its entirety. However, it is important to note that this generally does not apply to exploitation rights

18.4.2 Exceptions: Common Ground

If an act is covered by an exception, it is authorized by the law. This means the right holder cannot authorize nor prevent the act. Following legal instruments such as the Berne Convention, the TRIPS Agreement, and the InfoSoc Directive, the exceptions should be limited to special cases and not interfere with normal exploitation of the work nor unreasonably prejudice the legitimate interests of the author. In general, these exceptions are thus applied narrowly across the EU in favor of high protection of intellectual property rights.

The exceptions vary somewhat per intellectual property right, but there is some common ground. For instance, the exception for **teaching and research** and **public security purposes or an administrative or judicial procedure** exist both in the InfoSoc Directive (on copyright) and the Database Directive. In the former, these are exceptions to the right holder's reproduction right. In the latter, these exceptions target both the extraction and reutilization right. In the *sui generis* database regime, however, these exceptions can only be relied upon by a "lawful user." Think, for instance, of circumventing the requirement of a subscription to gain access to a nonpublic database without authorization. Extraction and/or reutilization by such a user cannot fall within the scope of these exceptions.

An example of teaching and research could be the showing of clips, (preparatory design material for) software code, small texts, or parts of a database for illustration to students or researchers. In order to qualify, both regimes require that third parties must not perform such uses for commercial purposes. Where possible, the source should be referenced and the use should not go beyond what is required for the noncommercial purpose pursued. For the exception for public security purposes or an administrative or judicial procedure, an example could be the copying of a work or certain data from a database to verify imported goods. Another could be the inclusion of such materials in the written decision of a court case revolving around questions of infringement of copyright and/or *sui generis* database protection. Again, such acts may not have been conducted for commercial purposes.

18.4.3 Exceptions Specific to the Right

The most common and relevant exceptions specific to the EU copyright regime are journalism, citation for review and criticism, and caricature, parody, and pastiche. EU law, more specifically the InfoSoc Directive, does not provide any conditions for any of these exceptions. This means, for example, that member states were free to limit exceptions to certain circumstances or uses only.

Finally—and perhaps most importantly—there is the recently introduced **text and data mining** exception in the Digital Single Market Directive. This concept is best understood as any analytical technique that is automated. It is used to derive information by analyzing text and data in digital form. This exercise could, for example, be performed to discover patterns, trends, and correlations in a dataset. Two variations of this right have been introduced, one focusing on text and data

■ **Table 18.5** Summary of exceptions

	Copyright	<i>Sui generis</i> database right	Trade secret right
Exceptions	<ul style="list-style-type: none"> • Teaching and research • Public security purposes and administrative or judicial procedure • Journalism • Citation for review and criticism • Caricature, parody, and pastiche • Text and data mining 	<ul style="list-style-type: none"> • Teaching and research • Public security purposes and/or administrative or judicial procedure 	<ul style="list-style-type: none"> • Freedom of expression and right to information • Revealing misconduct, wrongdoing, or illegal activity • Legitimate tasks of worker(s) (representatives)

Note to table: Author's own table

mining for scientific purposes and a general one. Both require that there was lawful access to the works that is to be subjected to text and data mining. The first type allows for the storage and retention of reproductions of the works for scientific research. However, there should be an appropriate level of security present on the storage of the copies of the works. For the general exception, there is the precondition that the right holder has not explicitly reserved the use of his/her work. In the absence of such a reservation, the works may be “mined,” kept, and stored as long as is required for the aim pursued with the text and data mining.

Under the regime of trade secret rights, the most relevant exceptions are the following three. First, the act may not infringe the trade secret right if the **freedom of expression and right to information** can successfully be invoked. Moreover, the acquisition, use, or disclosure of subject matter protected by a trade secret right in the pursuit of **revealing misconduct, wrongdoing, or illegal activity**, such as whistleblowing, may also be lawful. Furthermore, linked to one of the previously mentioned limitations, if their **legitimate tasks as workers or workers' representatives** necessitated the disclosure, the holder of a trade secret right may also not be able to apply for remedies against them (■ Table 18.5).

18.5 Alternative Sources

These intellectual property rights may vary in scope and purposes, but it is entirely possible that several may be applicable to (parts of) the same dataset or code. The exceptions to these various rights are limited to specific rights and purposes. Therefore, it is possible that an act that falls under an exception for one intellectual property right is not allowed due to the presence of another right. If a third party requires access to datasets (partially) covered by these rights, there are several options to gain lawful access to datasets.

The most straightforward option is to obtain a **license** from the right holder to use his/her datasets. A license allows the licensee to use protected subject matter in accordance with the agreed-upon terms, usually against payment. Protected subject matter can be licensed for some or all of the uses covered by copyright and/or *sui generis* database protection, but the author or maker remains the owner. Another similar option here would be to enter into an **ad hoc agreement or partnership** either by paying a sum or by offering something in return.

Alternatively, it is sometimes also possible to gain access to comparable datasets via other sources, such as **public sector information** or “PSI.” This is a very interesting and useful opportunity to consider because the state possesses many data—think, for instance, of maps, court decisions, company data, citizen statistics, etc.—and might have an obligation to release that data and allow its reuse (i.e., Freedom of Information Acts), although not necessarily for free. Data are likely to be subject to a PSI regime when (1) linked to the execution of state activities, (2) there are no intellectual property rights owned by third parties on them, and (3) the data are not kept secret for reasons of public policy (including data protection).

Depending on the business model, another option to consider might be using software or data subject to **open licensing schemes** (“open source”). The use of such data or software is free, but, depending on the type of license, there may be other types of restrictions. The most common division made is between *permissive licenses* and (weak or strong) *copyleft licenses*. These types of licenses are best envisioned as a spectrum from least prescriptive to most prescriptive license. Both types of licenses do not restrict the use of the subject matter in terms of use, modification, and redistribution, but permissive licenses allow proprietary derivative works while copyleft licenses do not. This means, for instance, that a third party can make modifications to the subject matter under the permissive license and license and distribute it under a different type of license. A weak copyleft license, on the other hand, would not permit this. Such licenses contain a clause proscribing making material derived from its subject matter proprietary or relicensing this derived material. Strong copyleft licenses additionally require that its subject matter can also not be licensed against a different license than the original. This means that a work subject to a “normal” proprietary license cannot be combined with another work subject to a copyleft license.

The provision of **complementary services or products** on the market to create or gain access to a similar dataset is also a possibility. For instance, a third party wants similar data as generated by sensors brought on the market by a competitor. The third party could decide to offer software that could operate the sensors from the competitor or offer competing sensors. Another option here would be to turn your own clients into data collectors themselves by having them correct or report certain data. Think, for instance, of reporting additions to a map or modifications to a street.

Finally, if these datasets contain personal data, you can ask those individuals to use their **right to data portability** via a promotion for new or existing customers of their own services or products. According to the General Data Protection Regulation, this right gives natural persons the opportunity to move their personal data from one online service to another. The requirements that have to be met here are that the

data (1) are personal data and (2) have been provided to the controller by the person whom the personal data are about. For instance, an insurance company or municipality could offer benefits in exchange for their personal data transfer to you, such as a discount on the insurance fee or on services offered by the municipality.

In short, if intellectual property rights exist on a dataset and none of the exceptions are applicable, there are still several avenues to gain legal access. Moreover, alternative sources can be explored as a complementary source or substitute for the protected dataset.

Conclusion

To summarize, when dealing with subject matter such as datasets and software, it is important to first establish whether intellectual property rights may exist on them. If so, the use of such subject matter by third parties may be restricted. Which uses are restricted and under what conditions depend on which right applies and, in varying degrees, on which regime applies (i.e., EU or USA).

Second, it should be established who the right holder is. If you are the right holder, this means that you may be able to restrict the access and use of others of the protected subject matter. If it is someone else, there are several possible routes to lawful use of that party's protected subject matter or alternatives to this subject matter, from obtaining consent from the right holder to acting within the scope of limitations or exceptions to finding or creating alternative sources.

Take-Home Messages

- Copyright on the contents, selection, or arrangement of a dataset gives the right holder the sole right to reproduce and make the protected material public.
- The *sui generis* database right gives the right holder the sole right to extract and reutilize substantial parts of the database.
- Trade secret rights on data protect the right holder against unlawful acquisition, use, and disclosure of the protected material.
- A third party can only engage in lawful use of subject matter protected by these rights if it is authorized by the right holder or by law (if exceptions are applicable).
- In the absence of authorization, there are several ways in which legal access can alternatively be gained to (parts of) the dataset or software or to comparable sources.

? Discussion Questions

1. Why do we have intellectual property rights?
2. How would you explain the distinction between ideas and expressions in copyright law?
3. Please define all types of investments relevant for *sui generis* database protection, including the means through which such investments can be made.
4. Please briefly explain the status of reverse engineering under the trade secrecy protection.

Further Reading

- Agreement on Trade-Related Aspects of Intellectual Property Rights, Annex 1c of the Marrakesh Agreement Establishing the World Trade Organization, 1994.
- Berne Convention for the Protection of Literary and Artistic Works of 1886 as amended on September 28, 1979 ('Berne Convention').
- Case C-202/12 *Innoweb BV v Wegener* [2013] ECLI:EU:C:2013:850.
- Case C-203/02 *British Horseracing Board v William Hill* [2004] ECLI:EU:C:2004:695.
- Case C-393/09 *Bezpečnostní softwarová asociace* [2010] ECLI:EU:C:2010:816.
- Case C-5/08, *Infopaq v Danske Dagblades Forening* Case C-5/08, *Infopaq v Danske Dagblades Forening* [2009] ECLI:EU:C:2009:465.
- Convention on the Grant of European Patents of 1973 as amended last on 29 November 2000 ('European Patent Convention').
- Council Directive 93/98/EEC of 29 October 1993 harmonizing the term of protection of copyright and certain related rights [1993] OJ L290/9 ('Term Directive').
- Estelle Derclaye, 'Database sui generis right: what is a substantial investment? A tentative definition' (2005) IIC 36(1).
- Directive 2001/29/EC of the European Parliament and of the Council on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L167/10 ('InfoSoc Directive').
- Directive 2004/48/EC of the European Parliament and of the Council on the enforcement of intellectual property rights [2004] OJ L157/45 ('Enforcement Directive').
- Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs [2009] OJ L111/16 ('Software Directive').
- Directive (EU) 2016/943 of the European Parliament and of the Council on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure [2016] OJ L 157/1 ('Trade Secret Directive').
- Directive 2019/790 of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] OJ L130/11 ('Digital Single Market Directive').
- Directive 96/6/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L77/20 ('Database Directive').
- Feist Publications, Inc., v. Rural Telephone Service Co.*, 499 U.S. 340 (1991).
- Husovec, M. (2019). How Europe Wants to Redefine Global Online Copyright Enforcement. *TILEC Discussion Paper*, 2019–2016.
- Publications Office in their Summary of Directive 2009/24/EC—the legal protection of computer programs, 23 January 2017.
- Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L 119/1 ('General Data Protection Regulation').
- Rosati, E. (2017). GS Media and its implications for the construction of the right of communication to the public within EU copyright architecture. *Common Market Law Review*, 54(4), 1221–1242.



Liability and Contract Issues Regarding Data

Eric Tjong Tjin Tai

Contents

- 19.1 Introduction – 460**
- 19.2 General Characteristics of Private Law – 460**
- 19.3 What Is Data? – 462**
- 19.4 Contracts and Data – 465**
 - 19.4.1 Formation of Contracts – 466
 - 19.4.2 Content of Contracts – 467
 - 19.4.3 Contractual Remedies – 469
- 19.5 Tort Law and Data – 473**
 - 19.5.1 Fault Liability – 474
 - 19.5.2 Strict Liability – 475
 - 19.5.3 Causality and Defenses – 475
 - 19.5.4 Damages and Other Remedies in Tort – 476
- References – 479**

Learning Objectives

- Understand how private law rules work
- Understand the different meanings of “data” in law
- Assess a contract and identify important contractual clauses
- Understand the concept of pure economic loss and its relevance for data
- Become familiar with the most important grounds of liability, in particular those relevant to data issues

19.1 Introduction

This chapter provides an introduction to certain areas of law, insofar as relevant for data scientists. If you are working as a data scientist, you may encounter legal questions. You may need to negotiate a contract or may worry about potential liability. A regular introduction to law such as by Ventura (2005) or Wacks (2015) will only be of limited assistance, as data poses particular legal problems which the general literature will not answer (Mak et al. 2018).

The aim of this chapter is to equip you with the basic knowledge of contract law and liability law, which should familiarize you with the basic principles involved. Also, this chapter contains some pointers to avoid potential pitfalls. As this is only a brief introduction, it is not possible to enter into the detailed rules that may apply in actual cases. When in doubt, consult a lawyer.

First, we will start with a brief example and a few general remarks. This is followed by a legal analysis of what data is. Subsequently will follow a discussion of contract law in some detail. Finally, liability in tort law is discussed.

► Example

Alice has started a business that provides analyses of client profiles for large companies. Bob hires Alice to analyze his business data. The analysis is performed by Eve, an employee of Alice. After the analysis is completed, Eve accidentally deletes Bob’s database, and Bob did not have a backup. Does Alice have to pay damages to Bob for the loss of the database, and if so, how much? ◀

19.2 General Characteristics of Private Law

As you can see from the example, there arise various questions. Law, in particular the area called private law, deals with these questions. Private law is the part of law that covers claims and relations between private individuals: two major parts of private law are contract law and tort law (dealing with liability). If you have a claim or other dispute about private law, you can ultimately go to a court to obtain a decision in the dispute. The court may award your claim, leading to a *remedy* (see

► Sect. 19.4.3).

Within private law, we distinguish between cases where there is a contract between parties, or where there is a claim for liability while there is no contract between the victim and the person who allegedly acted wrongfully (the tortfeasor). The first kind of case is governed by contract law (► Sect. 19.4), and the second kind of case by tort law or delictual liability (see ► Sect. 19.5).

Private law, insofar as relevant here, consists of rules (and exceptions) that determine whether there is a specific legal consequence. A legal analysis usually consists of first finding the relevant legal rules, and then assessing whether these rules apply to the facts of the case and what the outcome is. For example, a contract is formed if there is offer and acceptance. But if there is a defect of will such as mistake, the contract, albeit formally valid, could be annulled (► Sect. 19.4.1). Furthermore, if the contract is valid, but an obligation from the contract is breached (► Sect. 19.4.3), the creditor (the person who has a right to get the obligation performed) may claim damages (► Sect. 19.4.3). In programming language pseudo-code, the structure of and relation between such rules could be phrased as follows (to give an example):

```

if (offer & acceptance) {
    contract.valid()
}
if contract.mistake == TRUE {
    contract.invalid()
}
if contract.valid() {
    if contract.breached(case) {
        /* further conditions */
        creditor.money+=
            contract.breached.damages(case)
    }
}

```

This shows how, dependent on several conditions, the outcome may be that a contract is valid or invalid, and that the creditor may have the right to receive damages. The actual rules of the law are far more complicated than this example shows, but at least it may give you an idea of how legal rules operate.

Some further information is required to avoid misunderstandings. First of all, private law imposes obligations on human individuals (**natural persons**) but also on companies and other organizations. Such **legal persons** are generally treated in the same way as human individuals: they can conclude contracts and may be held liable in tort. There are specific rules governing legal persons, which is the subject of business law and is not dealt with here. In practice, legal persons are represented by agents and authorized personnel (such as the CEO or the managing director).

Secondly, you have to be aware that private law is localized and time dependent: it is primarily national and may change over time. We can draw an analogy to how computer programs are created in a specific version of a specific programming language and operating system. You have to know what the environment is in which the program will be executed: even though a program may run in several dif-

ferent versions, it is not guaranteed to run in a version for which it has not been written or tested. Similarly, lawyers can only give detailed answers about the law in relation to the legal system that applies.¹ It is, however, possible to describe the broad outlines of the law as applicable in most systems, similar to how you can describe an algorithm in pseudo-code, abstracting from specifics of actual programming languages. In this chapter, we use such an abstract approach to the law.

An important distinction which we cannot abstract from is that between countries² that have a common law system and countries with a civil law system. **Common law** jurisdictions are England and Wales (*not* the United Kingdom, as Scotland has a different legal system), the United States, and former English colonies (mostly part of the so-called Commonwealth). Most other countries have **civil law** systems³: they have a code, a written law, that collects most rules of contract law and tort law. Common law has some characteristics which deviate significantly from the rules of civil law countries: we will discuss a few examples below. Generally speaking, common law stresses formalities and the literal meaning of contracts and holds parties responsible for drafting the contract to express precisely what they want. Civil law systems tend to stress the actual intention of parties and allow courts larger freedom to interpret the contract.

19.3 What Is Data?

Before we can discuss contractual aspects of data and liability for data, we need to make sure that we understand what data actually *is*, both in fact and at law.

As a first approach, you may consider in what way people actually work with data. Data may be used in the form of word processor documents as attachments to e-mail, music files, and digital photos uploaded to the cloud databases. Technically, these are all **data files**. Furthermore, data is also identified with the **information** contained in such files and databases, as when we speak of “personal data.” Finally, the phrase “big data” has come in vogue. **Big data** refers not so much to a well-defined database or file, as well as to an ongoing system whereby data is continually received and processed. To keep such a system running, an organization requires technical facilities (database management systems, servers), services (continuous feed of data), and human resources (data scientists, IT support staff), all of which require legal support (license contracts, employment contracts). This can be graphically represented as follows in ■ Fig. 19.1 (showing the data flows between various sources and storage facilities).

These three forms of data, that is, (1) information, (2) data files, and (3) big data, lead to various legal issues. In the following, I will focus on the first two forms

1 This is governed by what is called Private International Law, also called Conflict of Laws.

2 Or more specifically, parts of a country that have the same system: those are also called jurisdictions.

3 There are also some exceptions that do not fall in either category, in particular mixed systems like South Africa that have elements of common law and civil law and/or other systems.

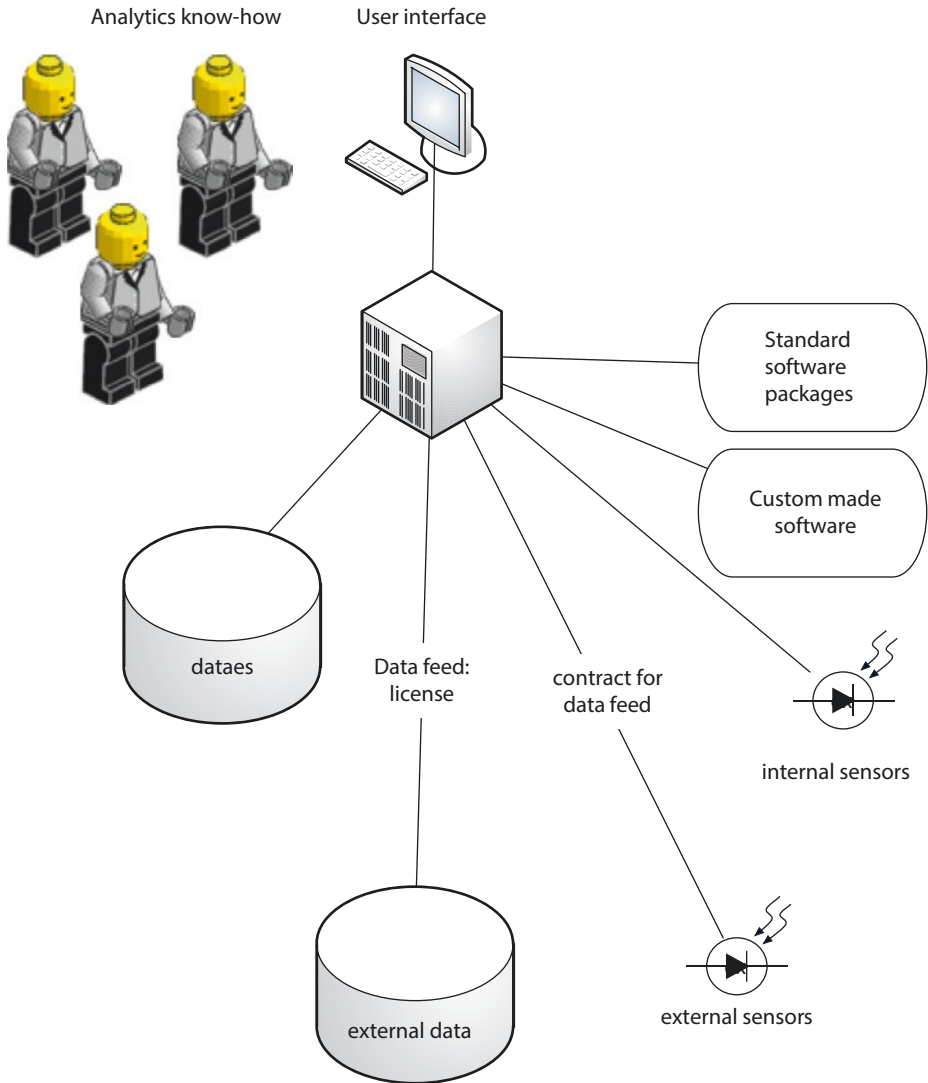


Fig. 19.1 Graphical representation. Note: Author’s own figure

of data, as big data is a separate issue that requires more space than is available here. Briefly put, big data can best be treated by looking at each element in turn.

The next question is what data is, legally speaking. How can data be qualified? Once we know that, we also know how to deal with data in contracts, or whether data can lead to liability.

As regards **big data**, that is not something specific in law. The various elements of big data can be qualified in turn.

Information is not as such an object of law. Information may give rise to liability, and it is possible to contract about information. We will see some examples below. But information as such is not a legally recognized object. It is not a tangible object and is not as such protected by law. It is possible that the information is partly protected by an intellectual property right because it is copyrighted or so, and trade secrets law does provide information on condition of it being commercially valuable and secret (see ► Chap. 18) that extensively discusses intellectual property rights. But information as such is not protected.

Data files require a bit more attention. Many lawyers tend to equate data files to information and similarly refuse to confer specific protection to data files. Indeed, IP rights and database rights themselves may cover databases, but they do not protect the data files as such: they only forbid the copying and distribution of the protected object. Trade secrets will often cover commercially relevant data (see ► Chap. 25).

However, one thing is not protected by IP rights. This is the actual *control* over a data file. Nonlawyers are quite happy to speak about data being someone's property or someone owning data. Lawyers are hesitant to do so. This is because data lacks characteristics that are common for normal objects of property, in particular tangible goods like cars, books, and vases. Data files are **intangible**, and data files are not exclusive: you can make a copy and use this copy without hindering the "owner" of the original data file.⁴ In law, usually only tangible objects are protected by property rights that give you the right to be restored in the possession of the object. Intangibles are usually only perceived from the viewpoint of intellectual property. As we have seen in ► Chap. 18, intellectual property rights do not operate like normal property rights as they only protect against infringement; they do not give a right to regain the control of its object (as that is the immaterial creation). For example, you cannot steal the song *Yesterday*. Even if you plagiarize it, the music and lyrics remain available to others. You cannot control the information as such.

Nonetheless, with data files, there does exist a form of factual control, simply by the fact that you prevent others from accessing the data file. This form of control of intangibles is new. In the past, you could only control information in your brain, but with data files, it is possible to control external information. The data file is a specific form of information, just like a printout of the data is a specific form (which is protected by property law, as a stack of paper is a tangible good). It is incorrect to equate the data file to the information as such: a data file provides advantages over the merely abstract information irregardless of the form. This becomes readily apparent if we consider the act of scanning a paper book and performing text recognition on the scan: although the resulting text file should not contain more or other information than the book, the text file may be useful for purposes such as data analysis in a way which the paper book is not.

⁴ Lawyers use the technical term "rivalrous" to denote the exclusive nature of ownership of tangible goods.

Currently, the law in most countries does protect the control over data files to a certain extent, and in some instances even allows you to claim the return of a data file that has been stolen. This would provide a protection similar to property rights to the data. However, this does not apply to all countries (Tjong Tjin Tai, 2018b). Other rights regarding property (for example security rights, such as mortgage and pledge) are difficult to arrange for data files.

Given the uncertain position of data files in law, the protection of data files is mostly indirect: because interference with the data is a tort, or because you have set up contractual obligations to treat data in the correct manner.

19.4 Contracts and Data

What is a contract? A contract, simply put, is an agreement between two persons, each of whom takes on certain obligations towards the other person.⁵ The classic example is the contract of sale: the buyer has to pay the price of purchase, and the seller is obliged to provide the buyer with the object that was sold.

When considering a contract, it is important to distinguish between the paper (or digital file) that provides proof of the contract and the legal contractual relation between the parties (which results from the act of signing, which proves the acceptance). The signed paper is also called a “contract.” However, if the paper would get lost, the contract itself (the legal relationship) would still exist and be valid. Here, we focus on the resulting legal relationship. General introductions are (Bix 2012, Cartwright 2013, Smits 2014).

There are situations where it is not immediately clear whether a contract is concluded. Whether there is actually a contract may depend also on the applicable law and other circumstances. An example is the terms and conditions (T&C) which govern the access to and use of a website: arguably, this is a contract as it seems to oblige you to adhere to these conditions while accessing the website, in return to which you get access. But the T&C can also be construed as being simply the conditions under which you are granted access, which do not provide additional obligations. The distinction between the two perspectives becomes clear when, for example, the T&C stipulate a fine if you post a negative review of the website elsewhere. In the first interpretation, this could be a binding condition; in the second form it is not, as it is not a condition of the access but an additional obligation, which cannot be imposed if there is no contract. We will not discuss these issues further as these lead to complicated legal discussions and cannot be explained or resolved easily.

Particularly important for data is the **license**. This refers primarily to permission to use someone’s intellectual property but is also used by extension for permission to use data. A license can be part of a broader agreement, and it can be the subject of a “sale,” for example when you “buy” an app for your smartphone. Such a contract involves a right to receive the program file of the app, and a license (right

⁵ There may also be more than two parties to a contract; we will not discuss those.

to use) for the app. You need both of these elements. It is possible to have a license while you have lost the file (because you have a new smartphone), in which case you would exercise your right to receive the program file again. Contrariwise, if you would have obtained an illegal copy of the program file, you naturally are not allowed to use it: you need permission, i.e., a license.

19.4.1 Formation of Contracts

Contracts are concluded by **offer and acceptance**. One party makes an offer for a contract, and the other party accepts the offer.⁶ In most cases, there are no formal requirements: a contract can be concluded in writing, digitally, or orally; consent to the contract can be deduced from actions (such as raising a hand) or even (in specific situations) from silence (tacit consent). However, for some types of contracts, there are additional formalities. An example is prenuptial agreements, which generally require some form of legal assistance (by a notary public or other legal service provider) in order to ensure that the consent was given freely and knowingly.

In common law, there is an additional requirement for a valid contract: **consideration**. This means that there has to be some kind of counterperformance in return. If one party unilaterally takes obligations upon himself/herself, without receiving anything in return, this is not considered to be a valid contract. The requirement of consideration could arguably lead to problems in the case of open-source software. Such software is provided for free, but may obligate the party who uses the software to accept certain restrictions in the use of the software, in particular may impose an obligation to make public under the same open-source license any modifications that are made to the open-source software.⁷ However, courts have accepted that the use of open-source software also involves some kind of consideration in the sense of compliance with the requirements of the license (Jacobsen v. Katzer, 2008). Hence, acceptance of the open-source license may create a valid contract under common law.

Although contracts generally do not have other formalities, there may be **information obligations** that require a party to inform the other party adequately when making an offer to contract. Examples are the identity and place of business when contracting on the Internet (for European cases).

A particular issue in contract law is the possibility that a contract which appears valid later turns out to be actually invalid. The two main reasons are that there is a **defect of will** (the consent was not formed in an appropriate way, for example by threat of violence) or that the contract violates **public policy** (such as contracting a hitman). Contracts against public policy are **null and void** (they have no binding effect; it is as if they were never concluded).

6 It is not necessary that parties make a clear explicit offer that is subsequently accepted as a separate action; it suffices that parties reach mutual consent on the contract, for example by drafting the contract together and subsequently signing the contract.

7 This is the so-called copyleft version of open-source licenses. There are other versions as well.

In practice, the most important defect of will arises in case of a misunderstanding between parties. In civil law systems, this may be caused by a **mistake** (a misunderstanding between parties about the object of the contract, for example whether a smartphone is brand new or a refurbished model). This implies that parties have to disclose relevant information of their own motion, and if a party is amiss in doing so, the other party may invoke mistake. In common law countries, the primary construct is **misrepresentation**: one of the parties represented in the contract (i.e., stated that something is the case) that the computer is new and not refurbished. If this turns out not to be true, the other party has a claim on the basis of misrepresentation. This implies that parties have to take the initiative to ask about the characteristics that they consider important. The consequence of mistake and misrepresentation is that the aggrieved party may **annul** the contract, which has the effect that parties are returned to the situation before the contract was concluded and everything that was already performed has to be undone (restored).

19.4.2 Content of Contracts

In a contract, you can specify what parties have to do and may expect from each other. A contract usually consists of several articles or clauses which describe these obligations and presumptions and regulate other issues as well. The interpretation of these clauses tends to focus on the text of the clauses (as it is usually assumed that parties deliberately chose a certain formulation), but the literal meaning may be corrected on the basis of the intentions of parties or other circumstances as well.⁸

An important distinction is between **obligations of means** and **obligations of result**. If you agree to do a data analysis, you may contract that the result will lead to savings of 10% of the current costs. However, you probably would not like to do so, as there is no way you can make sure that this happens. In that case, you would rather phrase it as an obligation of means: you will use your best efforts to perform a state-of-the-art analysis. This is similar to a doctor who only has to use care and skill, professional diligence, to attempt to improve the condition of a patient, without guaranteeing that the patient will fully recover his/her health. To give the other party some certainty, you can agree on some more objective ancillary obligations. For example, you agree that you will use a certain method of analysis and will provide at least three different analyses or reports.

A contract can also have as object the “sale” or licensing of data.⁹ The contract could contain clauses about the quality of the data (what is the source, to which level of detail/how many bits, which period, is it continuous, or is there a 99.99% uptime). It may concern a fixed dataset (such as the temperatures of a collection of sensors over a year) or a contract about a continuous data feed. It is important to

⁸ This applies in particular in civil law countries.

⁹ Strictly speaking, data cannot be sold as sale applies only to tangible objects. However, the law appears to develop to extend the meaning of “sale” to contracts regarding digital content as well.

realize that a license of data may have the effect that the licensee will have permanent use of the data or some form of transformation of the data. If the data, for example, is used to train an algorithm, it is in some way contained in the algorithm and cannot be removed from that. When drafting a temporary license for such use, it makes sense to explicate that parties agree about this definitive entrenchment of the data, while contracting that after the term of the license the licensee will delete the raw data. Furthermore, if a license is eternal and unrestricted (with a right to grant sublicenses), this amounts in effect to making the licensee the “owner” of his/her copy of the data. Even though you remain “owner” of your data (unless you provide an exclusive license, which would in effect make the licensee the actual new “owner”), you cannot stop the licensee from doing anything you could do with the data. This is one way in which companies may nominally say that you remain “owner” while actually obtaining owner-like powers through an extensive license.

Furthermore, contracts often contain a lot of clauses that regulate the kind of remedies that are available to parties. These clauses are particularly important in common law, as English and the US laws only provide remedies for specific kinds of contractual clauses, so-called **warranties** and **terms**,¹⁰ while the contract should also explicate which remedies apply to these clauses. In civil law, you do not need to be so precise: a civil law court will enforce any obligation that is contained in the contract, without needing a precise form.

An important category of clauses is the **choice of law** and **choice of forum**. A contract can (and often does) specify which law applies and which court you can go to. It can also contain a clause that says that you cannot go to court at all but (for example) have to arbitrate the case. The choice of law can have the effect that you lose a form of protection that you would have under your own system.

Some other kinds of clauses are discussed below (► Sect. 19.4.2) as they pertain to remedies.

The law does provide some checks on the content of a contract, through the doctrines of **unfair terms** and **unfair commercial practices**. Simply put, these disallow clearly unreasonable and socially unacceptable terms and may have as a consequence that terms which seem logical to be implicit are considered to be part of the contract.¹¹ For example, a contract of sale is presumed to imply that the object sold is fit for the purpose that it ordinarily has.¹² While this is an important form of control to ensure that contracts are reasonable, the legal rules are quite general. They have not been updated for the specific ways in which contracts that involve data may be unreasonable or unfair, and it is not always certain whether they can be reinterpreted to fit the data society.

The doctrine of unfair terms is related to a more general doctrine in civil law countries, namely that contracts are presumed to include a duty to interpret and execute the contract according to **good faith**: acting like a reasonable person would,

10 The precise terminology in English law differs from that in the US law.

11 In common law, this overlaps partly with the doctrine of implied terms.

12 In the United States, the Uniform Commercial Code § 2-315 codifies the implied warranty of fitness for a particular purpose.

and not merely focusing on the literal text of the contract. This instrument can also be used to fill in gaps in the contract. A drawback of the use of good faith is that it can result in courts deviating from the original intentions of parties. In common law, good faith is not generally accepted as an instrument, and there is debate as to whether it can be used to correct contracts. Nonetheless, US courts have also accepted that in every contract, there is an implied covenant of good faith and fair dealing.

19.4.3 Contractual Remedies

If the contract is not performed correctly (which means if the other party did not completely perform as he/she promised in the contract, i.e., he/she did **breach** one or more of his/her obligations), there is a so-called **breach of contract**. Upon breach, the party to whom the obligation was owed (we generally call this party the **creditor**) may invoke a **remedy**. A remedy is the legal response that the court offers the creditor, which is intended to motivate the **debtor** (the party upon whom the obligation rested) to actually perform his/her obligation.

The creditor usually has to prove that the debtor did not perform the obligation correctly. In case of an obligation of result that is easy, he/she only has to show that the result was not realized. In case of an obligation of means this is more difficult: he/she has to show that the debtor did not provide the care and skill that were required, which usually can only be determined indirectly.

19.4.3.1 Prerequisites for Invoking a Remedy

Before you may claim a remedy, most legal systems first require that the debtor is in **default**. This means that he/she definitely did not perform his/her obligation, and it is his/her fault that he/she did not do so. This implies two requirements: a notice of default and the absence of an excuse.

(a) The debtor should normally get a second chance to perform his/her obligation.

It is possible that he/she was not aware of the nonperformance and would willingly correct the situation if you complain. For example, if you order a smartphone from Amazon, the seller would not know if you did not receive the package. If you complain, he/she must get an opportunity to remedy the situation by sending another phone. Therefore, many legal systems require you to first send a **notice of default**: a clear statement that there was a breach of a contractual obligation, and a term within which you want the breach repaired. Only after the term has expired is the debtor in default and can you ask a court for a remedy. In some instances, when it is clear that the debtor will not respond to an opportunity to remedy the breach, or when the contract itself specified a fatal term (a moment on which the obligation should definitely be performed, such as the date of delivery of a wedding cake), it is not necessary to send a notice of default. Instead, the debtor will immediately become in default of his/her obligation once the term has passed and the obligation was not performed.

- (b) Even if an obligation is not performed as promised, it is possible that the debtor has a valid **excuse**. The actual cause of the breach may be an external factor. For example, a package was not delivered because the whole building was inaccessible due to a terrorist attack. In such a case, there is formally a breach of contract, but because the attack provides a valid excuse, the debtor is not at fault and the creditor may not have a remedy. The question is: How do you determine whether there is a valid excuse? This is usually done by looking at the actual cause of the nonperformance or breach, and considering whether this cause is **attributable** to him/her: he/she is to blame for that cause, or it is his/her responsibility. If he/she is not to be blamed, such a cause is considered to be **force majeure**. This is not a clear-cut issue. For example, if a manufacturer cannot supply you with computer chips because of a strike at his/her plant, you could consider this force majeure. However, if the strike started because the directors abused their personnel, you could also argue that they themselves are to blame for the strike. To avoid these kinds of discussions, contracts often contain a so-called **force majeure clause**, which lists the events that are considered to constitute force majeure. Such a list may be exhaustive or may only be illustrative (showing the kind of cases that constitute force majeure, and allowing that other similar cases may also constitute force majeure).

In civil law systems, generally all contractual obligations (as follow from interpretation of the clauses of the contract) give rise to a remedy when the obligation is not performed correctly. In common law systems, principally England and the United States, only specific contractual clauses may lead to breach of contract and remedies. In England, the distinction is between representations (which give rise to an action for misrepresentation, see ► Sect. 19.4.1) and terms (which lead to breach of contract). Such terms are further divided into conditions (which lead to termination), warranties (which are cause for damages), and intermediate terms (which may lead to termination and/or damages). In the United States, representations also give rise to an action for misrepresentation. However, while in English law a clause which is not a representation tends to be automatically a term which leads to remedies, in the US law, a clause which is not a representation does not automatically give rise to a remedy. Only so-called **warranties** allow the creditor to invoke a remedy. The contract often specifies the kind of remedies that are connected to breach of specific warranties. In civil law systems, it is not necessary to go into detail about which remedies apply and in which way. However, it is still useful to explicate this under civil law as well, as this may avoid complicated discussions in the advent of breach (see below for examples).

Hence, there are four possible requirements before you can invoke a remedy:

- Breach of a contractual obligation
- Default
- Attributability (no excuse, no force majeure)
- In common law: obligation is put in a clause that allows a remedy (such as a warranty or a term)

19.4.3.2 The Available Remedies

There are, generally speaking, three contractual remedies:

- Specific performance
- Award of damages
- Termination

These remedies can be claimed from a court. Termination may, depending on the jurisdiction, also be available out of court (termination by notice).

Specific performance means that the court orders the debtor to perform his/her obligation. For example, you have contracted to build a database, but refuse to do so, and the court subsequently orders you to do what you promised. In civil law systems, this is the primary remedy. In common law, specific performance is not always available: it depends on the kind of obligation. The reason for this attitude may be that in common law specific performance is in principle monitored by the court (which would cost the court time), as the breach of specific performance constitutes “contempt of court.” In civil law systems, specific performance is an obligation to the creditor and is often enforced through a private “fine” which the creditor can invoke without bothering the court.¹³

An **award of damages** is the primary remedy in most cases, both in civil law and in common law. This means that the court awards the creditor damages, i.e., the debtor has to pay damages to the creditor. Damages means a sum of money, with the purpose to compensate the creditor for the negative consequences of the breach. Take note: damages are a sum of money; damage (without an “s”) means the actual loss or injury. Damages are intended to compensate for the damage. Lawyers distinguish several kinds of damage.

An important distinction in forms of damage is between

1. personal injury,
2. property damage, and
3. pure economic loss.

Personal injury means the damage that follows from harm to a person’s body or health, such as medical costs and loss of income (because of the injury).

Property damage is the harm to physical property and the consequences of this harm. Examples are loss of value to a damaged car, repair costs for the car, as well as cost of alternative transport while the car cannot be used.

Pure economic loss encompasses all other kinds of damages: this means losses that do not follow from either personal injury or property damage. An example is a slanderous statement: this does not start with injury or property damage, instead starts with an immaterial harm. Examples are loss of profit, wasted time, and stock market losses. Also damage to a database or data file is pure economic loss. Hence, an economic loss (such as loss of income) is only pure economic loss if it is not the consequence of injury or property damage; otherwise, it would fall under one of the other two forms of damage.

¹³ In French law, this is the “astreinte,” a kind of fine that is due when the obligation is still not performed after a certain period.

The importance of this distinction is that the first two kinds of damage are commonly recoverable, including the consequential losses deriving from the primary injury or damage. In contrast to that, some jurisdictions (in particular common law) do not allow recovery of pure economic loss for breach of contract. Hence, even though you may formally have a remedy for breach of contract, in actuality, it is possible that your loss will not be compensated by an award of damages, as the loss constitutes pure economic loss. If you would like compensation for these kinds of loss, you would need to add a clause for liquidated damages for certain breaches (see below). An example that may illustrate the difference between these kinds of losses: your IT service provider accidentally deletes your database. As the deletion does not result from property damage or personal injury, this constitutes pure economic loss and may not be recoverable in England. If the service provider would have set fire to your server room, whereby the database and all backups (on hard disks) were lost, the loss of the database would be consequential economic loss, following from the property damage. This loss *would* be compensable. The qualification therefore does not depend on the loss itself, but on the way in which the loss occurred (as a *consequence* of property damage or personal injury, or not).

Limitation clauses or **exemption clauses** are clauses that limit the amount or kind of damage that the debtor has to compensate upon breach. An example found often in software licenses is that only damage from personal injury or physical property damage is compensated.¹⁴ Since software usually does not cause personal injury or property damage, the result is that the producer of the software will normally not have to pay any damages (unless the clause is unfair). A damages clause may also limit the amount of damage that may be awarded, for example to a maximum of €10,000.

Liquidated damages clauses are clauses that fixate the amount of damages awarded for specific kinds of breach. The advantage is that both parties will know in advance how much damages will have to be paid for certain breaches, whereby parties need not go into lengthy discussions about for example the value of an unsuccessful automation project. If the amount is much higher than the actual loss suffered, such a clause amounts to a penalty clause.¹⁵

Limitation clauses and liquidated damages clauses can amount to an unfair clause, in which case these are not allowed (► Sect. 19.4.2). For instance, it is usually not allowed to fully disclaim liability for death or personal injury.

Furthermore, damage is only compensated if it is **not too remote** from the breach. In English law, there is a requirement of foreseeability: the damage must be foreseeable. Breach of contract may for example require compensation of replacement costs, but not compensation for the costs of psychotherapy to an employee who feels personally disrespected by the breach. This limits the extent of liability.

¹⁴ In other words, compensation for pure economic loss is explicitly excluded.

¹⁵ Penalty clauses are usually allowed, but some jurisdictions do not allow them. In other jurisdictions, they may be mitigated by a court if they are disproportionately high.

Termination means that the contract comes to an end because of the breach. The result is that there are no further obligations between parties, except the obligations necessary to wind down the contract. The termination may have the effect that the contractual performance has to be restored: everything that has been done or given is undone. This may for example mean that a database that has been created and delivered to a company has to be returned to the developer, but that the payment that the developer has obtained may also have to be returned.¹⁶ A consequence could also be that the company may keep the database but has to pay the developer a certain amount of money for it. As this example makes clear, there are a lot of possibilities. It is useful to regulate the most important consequences of termination in some detail, to ensure that upon termination you will be in the position you want.

There are other kinds of termination that do not require breach (for instance, a clause that allows termination if your contract party is taken over by your competitor).

We can now return to the example in ► Sect. 19.1: deletion of the database is clearly not intended as part of the contract, while conceivably Alice should have taken precautions against such a mishap. Therefore, this may be said to constitute a breach of contract. If common law applies, it may also be necessary that there is a warranty or other term that provides a ground for breach of contract in case of deletion of the database. Although the mistake is made by Eve, Alice as her employer is liable as she cannot excuse herself by pointing to a fault by an employee (this does not constitute force majeure). However, the loss is pure economic loss, which under common law may not be recovered on the basis of breach of contract. This may be different if the contract contains a liquidated damages clause. In a civil law jurisdiction, the loss may be compensable, and the damages may be assessed at the market value of the database or the costs for the (re)creation of the database (see ► Sect. 6.4).

19.5 Tort Law and Data

If there is no contract between two persons, liability must be based on **tort law**.¹⁷ A tort, simply put, describes for specific cases under which conditions a person is liable. An example is defamation: this tort regulates when someone is liable for making defamatory statements. There are many different kinds of torts. In tort law, we distinguish between two forms of liability: **fault liability** and **strict liability**.

16 It does not necessarily mean that the database has to be entirely destroyed, which usually is not in the interest of either party: it may mean that one party may have to delete all copies of the database in his/her possession.

17 In civil law systems, alternative names can be found such as the law of delictual liability. General literature on tort law: Van Dam 2013, Tjong Tjin Tai 2022.

19.5.1 Fault Liability

The general idea of fault liability is that someone (lawyers call this person the **tort-feasor**) is liable if he/she committed a **fault**: he/she did something that he/she legally should not have done, and the fault caused harm to the victim. Therefore, there are three conditions:

- Wrongful behavior (fault)
- Harm
- A causal connection between fault and harm

Harm means some kind of disadvantage, an accident, or loss. This may be a physical accident, damage to a good, harm to reputation, and loss of privacy. The causal connection is further discussed in ► Sect. 19.5.3.

Most torts are defined by the kind of fault that they cover, but the kind of harm may also be relevant. An example is again defamation: this applies to certain kind of statements, but also requires the presence of harm to the reputation of the victim. Legal systems have various ways in which to determine whether certain conduct (which may consist of acts as well as omissions)¹⁸ is wrongful.

The most important tort is **negligence**.¹⁹ Strictly speaking, this is a tort in common law, but civil law systems also have rules that provide for liability in cases that are covered by the tort of negligence. Negligence means that you did not observe sufficient care or diligence towards the interests of the victim: it requires the breach of a duty of care. Negligence could for example apply if you developed an algorithm for a self-driving car without taking sufficient care during development to avoid mistakes in identifying obstacles, and the car killed a pedestrian because it thought it was a cardboard box. Negligence is useful as it provides an open-ended norm which can be applied to new developments. The disadvantage is that the actual application may be unclear: When is behavior negligent? The usual test is whether a reasonable person would have acted in the same way or not.

There are many torts besides negligence that may apply to data. Examples are breach of privacy and defamation. These are governed by specific rules and conditions; this overview is not the place to go into details. Furthermore, many abuses or wrongful interferences with data and computers constitute a crime.²⁰ A breach of criminal law is generally also a tort, under the tort of breach of a statutory duty.

18 In common law systems, pure omissions (which do not follow from an earlier negligent act) may not lead to liability.

19 This is the name in common law; in civil law systems, other names can be found, but negligence may be used to denote the local tort in English translation.

20 In particular because of the influence of the various Conventions on Cybercrime.

19.5.2 Strict Liability

Strict liability means that you are liable even though you did not personally commit a fault. Examples are:

- Liability of the employer for torts committed by an employee
- Liability of the car owner for accidents involving the car
- Product liability of the industrial manufacturer of tangible products

Most jurisdictions recognize these three forms of strict liability. Required for application of these forms of strict liability are at the very least:

1. A specific relation between the person held liable and the object or person that caused the damage (an employment relationship, ownership, control)
2. A specific form of behavior or activity of the object or person (tortious act, realization of a specific danger)

In many countries, there are also other forms of strict liability, such as liability for children, animals, dangerous objects, and dangerous activities. However, these forms are not accepted everywhere or to the same extent. The advantage of strict liability is that it provides further protection to victims, who can secure compensation from the party who actually profits from (or chose to enter into) the risky activity or the engagement of the person. In the literature, strict liability is often suggested as a model to regulate liability for robots and algorithms (Tjong Tjin Tai 2018a, and references therein).

To ensure that strict liability does not become too extensive, there are several limitations. In particular, the general limitations of causality and assessment of damages apply, as well as defenses (► Sects. 19.5.3 and 19.5.4).

19.5.3 Causality and Defenses

A requirement for obtaining a remedy for a tort is that there is harm that was caused by the fault: **causality**. The causal link between the fault and the harm is assessed first by looking at the **factual causal connection (factual causality)**: Would the damage also have occurred if the fault would not have taken place? This so-called but-for test (or *conditio sine qua non*, in civil law systems) is a necessary condition to assume a tort.

After that, a second causal connection is required: **legal causality**. The harm and damage should not be too remote. This is a similar test as with contractual liability: foreseeability or general remoteness may be used as a criterion. The details depend on the tort and legal system. Remoteness helps to limit possible exposure to liability. For example, if you happen to make a mistake in an update for an open-source routine, which ultimately causes servers all over the world to malfunction, your liability may be limited if these consequences are found to be too remote.

If there are several tortfeasors, causality tends to be **joint and several**: the tortfeasors are liable individually as well as collectively. Each may be sued on his/her own for the full amount of damages; afterwards, the tortfeasor who paid the damages may obtain a contribution from the other tortfeasors.

Liability is mitigated or limited if a **defense** applies. An important defense is **contributory negligence**: if the victim did something that also contributed to the occurrence of the harm or the extent of the damage, the court may reduce the award of damages by the proportion to which the victim contributed to the damage.

Other defenses may stand in the way to liability. An example is the defense of force majeure in strict liability.²¹ If for example an accident was not caused by the car or its driver, but rather because the car was rear-ended by a truck and thereby shoved into the car before it, this might disculpate the owner of the car from liability.

19.5.4 Damages and Other Remedies in Tort

If the conditions of the tort are fulfilled, the victim has a right to a remedy. The general notion of a remedy in tort is similar to remedies in contract (discussed in ► Sect. 19.4.3), but the requirements are different and not all remedies apply in both situations.

The principal remedy for a tort is an **award of damages**. Other important remedies may be given in the form of a **court order**. Court orders are proclamations by the court. There are many types of orders; the orders that are important as providing a remedy are those which command a party to perform a certain act or actions, or refrain from certain behavior. An example is a restraining order, prohibiting certain behavior. A particular type of order is an **injunction**, prohibiting or commanding the performance of a specific act.²²

An award of damages means that the loss that the tort caused has to be compensated by the tortfeasor. The court assesses the damage. However, there are several complications.

First of all, for many torts, not all kinds of damage are compensated. In particular, pure economic loss is quite often not compensated.²³ As cases involving data often only cause pure economic loss, such torts may effectively not have a proper remedy.

There are alternative forms of damages that supplement the usual forms of damage. A particular example is **punitive damages** (an award of a sum of money that exceeds the actual damage: this is imposed as a punishment of the tortfeasor, to serve as a deterrent).

21 This is somewhat similar to force majeure in contract law (► Sect. 19.4.3.1).

22 The exact definitions of orders and injunctions may vary between different legal systems.

23 For example, the tort of negligence in many cases does not allow compensation of pure economic loss, albeit such compensation is not always excluded.

What are the harms that may follow from data? Very generally speaking, we can consider the following issues.

- Lack of quality
- Errors in analysis
- Loss of data
- Data leaks/loss of privacy

As regards the first topic, there is very little known about this (Zeno-Zencovich, 2018). In big data, quality is not presumed as a given; rather, the idea of big data is often to work with what you have, polluted or not, in the assumption that slight blemishes will cancel out statistically. However, data which is incorrectly structured or is otherwise incorrect may cause negative consequences: the user may have to spend considerable effort to normalize or restructure the data, and he/she may draw incorrect conclusions which lead to bad decisions. Outside contract, there is little reason to assume that you have an action for lack of quality, as you did not pay for the data. If you make data available, it could be useful to explicate that you provide no guarantees as to the quality of the data. If the data is provided on the basis of contract, negative consequences could in theory be compensated, but these are usually pure economic loss and might fall outside compensation. Furthermore, most contracts would contain clauses regarding such damage, providing either clear liquidated damage rules or even exclusion/limitation clauses.

As regards errors in analysis: this will probably be a contractual issue. The rules of the contract will apply, in particular liquidated damage clauses or limitation clauses.

For loss of data, the general idea would be that the loss of value of the data could be compensated, or possibly the cost of reconstructing the database.

For data leaks and loss of privacy, there is often no effective remedy (Peters 2014, Varuhas 2018). Take the example of a social website that is hacked, whereby all your private data has been exposed to the hackers. While this may be a serious invasion of privacy, there is no material loss from the breach. Privacy itself has no clear value; at best, some symbolic or nominal amount of damages may be awarded for the immaterial injury, or in rare grievous cases punitive damages. While your private data could be used to hack other accounts and subsequently cause material loss (such as theft of money), this will usually be pure economic loss. Even if such loss can be compensated in your legal system, it will be hard to prove the causal connection between the data leak and the loss. Only in the case of privacy infringement of well-known people could there be a possibility of a substantial award of damages.

Hence, in case of torts regarding data, it may be difficult to obtain substantial damages.

It may be instructive to look again at the example of ► Sect. 19.1, this time from the perspective of tort law. Assume that Eve this time deleted a database of a third party (Charlie) that happened to be stored on the same server as the client's database. The deletion of the database by Eve probably constitutes a negligent action. Alice, as employer, is vicariously liable for the negligence of Eve. As the loss is pure economic loss (it does not result from property damage or personal injury),

it may not be recoverable in common law. In civil law systems, the loss may be compensated. It can be argued that there is contributory negligence from the part of Charlie: he could also have made a backup if the database is that important. In a contractual situation, this defense may not work if the contract obliges the debtor (Alice) to make the backups.

Conclusion

There are several issues that require attention when contracting about data and when considering liability for data. Besides the general legal points of attention (which were introduced in this chapter), there are also particular complications that are mostly a consequence of the intangible nature of data. It is important to realize that “data” is not a fixed concept and may also encompass elements of the surrounding infrastructure. Particular attention is required as to the limited compensation that may be obtained, and the more extensive kinds of harm that may be caused by data.

Take-Home Messages

- Data is not a well-defined concept in law. It may refer to the data files, information contained in files, or big data (a collection of various elements, including data files). Data files are not recognized as an entity in law; information is only protected by IP rights (including trade secrets law). For big data, each of its constituent elements has to be protected on its own.
- In contracts, it is important to explicate what you expect and what the consequences are when your expectations are not met. This applies to representations, warranties, and contractual obligations in general.
- Damage may consist of personal injury, property damage, or pure economic loss. Pure economic loss is in many cases not compensated, because of restrictions to compensation of such loss in contract law, and limitation or exclusion clauses. In case of data, there is often only pure economic loss; hence, in case only data is involved, you may not get any damages.
- Liability for data requires the presence of an applicable tort that deals with data. Liability requires a wrongful act or fault (something you should not have done), harm, and a causal connection between the act and the harm.
- You may be liable not only for direct wrongful acts, but also for the acts of someone else (vicarious liability) or consequences of objects which you control (and for which you could take precautionary measures).

? Questions

1. In what way are abusive contract clauses regulated?
2. What are the requirements for invoking a remedy for breach of contract?
3. What is pure economic loss?
4. How can you limit the risk of contractual liability?
5. What are the basic requirements of tort liability?

✓ Answers

1. By control of unfair terms or unfair commercial practices, or by good faith.
2. Breach of a contractual obligation, default, attributability of the cause of the breach. In common law also: violation of a warranty (i.e., the obligation that is breached is part of a warranty).
3. Losses that are not the consequence of physical injury or of property damages.
4. By limiting the extent of your obligations/warranties, with limitation/exemption clauses, by having an extensive force majeure clause, with liquidated damage clauses.
5. Fault, causality (factual and legal), and harm.

References

- Bix, B. H. (2012). *Contract law. Rules, theory, and context*. Cambridge University Press.
- Cartwright, J. (2013). *Contract law. An introduction to the English law of contract for the civil lawyer*. Hart, Oxford.
- Mak, V., Tjong Tjin Tai, T. F. E., & Berlee, A. (Eds.). (2018). *Research handbook in data science and law*. Elgar Publishing.
- Peters, R. M. (2014). So you've been notified, now what? the problem with current data-breach notification laws. *Arizona Law Review*, 56(4), 1171–1202.
- Smits, J. M. (2014). *Contract law: A comparative introduction*. Elgar.
- Tjong Tjin Tai, T.F.E. (2018a). Liability for (semi-)autonomous systems, in: Mak et al. (2018), pp. 55–82.
- Tjong Tjin Tai, T. F. E. (2018b). Data ownership and consumer protection. *Journal of European Consumer and Market Law*, 7(4), 136–140.
- Tjong Tjin Tai, T.F.E. (2022). *Tort Law: A Comparative Introduction*. Elgar.
- US Federal Court of Appeals, *Jacobsen v. Katzer*, 535 F.3d 1373 (Fed. Cir. 2008).
- van Dam, C. C. (2013). *European tort law* (2nd ed.). Oxford University Press.
- Varuhas, J. N. E. (2018). Varieties of damages for breach of privacy. In J. N. E. Varuhas & N. A. Moreham (Eds.), *Remedies for breach of privacy*. Hart Publishing.
- Ventura, J. (2005). *Law For dummies* (2nd ed.). Wiley.
- Wacks, R. (2015). *Law: A very short introduction*. Oxford University Press.
- Zeno-Zencovich, V. (2018). Liability for data loss, in: Mak et al. (2018), pp. 39–54.



Data Ethics and Data Science: An Uneasy Marriage?

Esther Keymolen and Linnet Taylor

Contents

- 20.1 Introduction – 482**
- 20.2 Data Ethics in Academia – 483**
 - 20.2.1 Moral Theories – 483
 - 20.2.2 Consequentialism – 484
 - 20.2.3 Deontological Ethics – 485
 - 20.2.4 Virtue Ethics – 486
 - 20.2.5 The Focus of Academic Data Ethics – 487
- 20.3 Data Ethics in the Commercial Domain – 489**
 - 20.3.1 Technological Level – 489
 - 20.3.2 Individual Level – 491
 - 20.3.3 Organizational Level – 492
- 20.4 Law and Data Ethics – 493**
- 20.5 Data Ethics and Data Science: Are They in It for the Long Run? – 495**
- References – 498**

Learning Objectives

By the end of this chapter, the reader should be able to:

- Distinguish between three key moral theories (consequentialism, deontological ethics, virtue ethics).
- Give examples of how moral theories can be used to analyze data science practices.
- Contrast academic and commercial data ethics.
- Explain the interaction between law and ethics in the data science domain.
- Critically assess different applications of data ethics.

20.1 Introduction

Over the last two decades, data-driven companies have reshaped the way in which society functions. Increasingly, citizens make use of social media platforms to build and maintain their social relations; governments base their interventions on off-the-shelf algorithmic tools; cities become “smart” by collecting, in close collaboration with private parties, all sorts of sensor data on their inhabitants; and in different fields—from health to law—there are initiatives to implement AI instruments to improve the decision-making process. Literally, the sky is no longer the limit, as tech entrepreneur Elon Musk is working on the possibility to transport large numbers of people to Mars as an answer to the climate crisis (New York Times, 2019).

Together with these high hopes on what technology and in particular data-driven innovations will bring us, we are at the same time also confronted with large and impactful incidents caused by these same technologies and innovations. From Cambridge Analytica intervening with elections around the world (New Statesman, 2018; Observer., 2020) to the company Clearview scraping social media to feed their facial recognition application (Wired, 2020), from Amazon’s poor treatment of their employees (Guardian, 2020) to Google’s racist search suggestions (Wired, 2018), it becomes clear that tech companies grapple with taking responsibility for their societal and political impact.

Against this backdrop, it should not come as a surprise that civil society around the world has called for data-driven companies to take their responsibility seriously and to work on becoming more fair, transparent, accountable, and trustworthy, to name just a few of the goals that have been set. But what exactly can we morally expect from these companies and what should companies do in order to better themselves? With these aspirations of making data-driven companies, as well as their products and services, more attuned to values central to a thriving, democratic society, **data ethics** enters the scene.

As we will see in this chapter, data ethics is a fuzzy concept that can mean different things to different people. The biggest part of this chapter is therefore dedicated to explaining data ethics from different angles. We will start with a brief introduction of data ethics as an academic discipline and illustrate how some of these academic viewpoints trickle down in the debate on data science and AI. Then,

we will focus on how data ethics has been put forward as a regulatory strategy by data-driven companies to formulate an answer to the above-described incidents and challenges. Next, we will explain how data ethics relates to law (which is also closely related to the other chapters in this module). We will address the risk of ethics washing, if in this entrepreneurial context data ethics is not properly embedded and is used as an escape from legal regulation. We end this chapter with a reflection on the future relation of data ethics and data science and provide some discussion questions to instigate further debate.

20.2 Data Ethics in Academia

Let us first look at the “ethics” part of data ethics. Ethics is a branch of philosophy that revolves around the question: “How should one act?” In a systematic way, this discipline studies the reasons and standards underpinning our actions and investigates what makes our actions morally right or wrong, and good or bad (Timmons, 2012, p. 4). “Morally” is an important adjective here, as it determines the scope of the kind of actions we are interested in. An example:

- You should eat your soup with a spoon.

This statement is prescriptive. It tells you how to act, probably based on reasons of efficiency and good manners. It is normative, in the sense that it tells you what the proper action is in a certain situation (use a spoon). However, though it is a normative statement, it is not necessarily a moral one. Whether or not you eat your soup with a spoon does not influence key values such as human dignity, freedom, or well-being. Generally, it also does not interfere with your or someone else’s chances of living a good life. Eating soup with a spoon is not a moral duty. It is rather a case of having good manners. Hence, whether or not you should eat your soup with a spoon is not a question ethics is concerned with.

- You should not tell lies.

Similar to the previous one, this statement too is prescriptive and it tells you what the right action is (not to lie). However, unlike the “soup with a spoon” statement, this statement obviously does touch upon some of the core aspects of a person’s life. It immediately interacts with key values such as your freedom, well-being, and human dignity. Whether or not you adhere to this statement has a significant impact on someone’s chances of living a good life. This is therefore a normative statement which includes a “moral ought.” Violating this normative statement is not a matter of bad manners, but severely unethical or immoral behavior.

20.2.1 Moral Theories

We might intuitively say that lying is unethical, but can we also give a coherent, rational explanation as to why it is wrong? Several moral theories have been developed to explain what makes a certain action morally (un)acceptable. We will very

briefly touch upon three influential moral theories: consequentialism, deontological ethics, and virtue ethics. For every theory, we added an example on how these specific viewpoints can play a role in understanding data science questions.

20.2.2 Consequentialism

Consequentialism holds that to morally judge certain actions, one has to solely focus on the consequences of those actions. In other words, this theory “embodies the basic intuition that what is best or right is whatever makes the world best in the future” (Sinnott-Armstrong, 2019). What matters is the impact of our actions.

In the eighteenth century, Jeremy Bentham ([1789] 1996) developed a well-known consequentialist theory called **utilitarianism**. In classical utilitarianism, an action is considered to be morally right if it **maximizes the good**. Bentham claims that the only value which is good in itself is pleasure. Therefore, this intrinsic value should guide all our actions. Thus, from a utilitarian point of view, it is our moral duty to maximize the good, to maximize pleasure.

In order to establish if an action effectively brings about the most pleasure or happiness for the most people, Bentham proposes to make use of a moral balance sheet, which is a kind of cost-benefit analysis trying to calculate how much happiness/unhappiness one can expect when executing an action. A balance sheet should enable the comparison of different possible actions and establish which one brings about the most happiness and therefore is, from a utilitarian point of view, the best action. Thus, if lying would maximize well-being or, in Bentham’s terms, happiness, it is morally the right thing to do.

► Consequentialism in Data Science Practices

In order to contain the devastating consequences of the Covid-19 pandemic, many data-driven solutions have been proposed and developed, e.g., smartphone apps to enable contact tracing, to provide a proof of immunity, or to regulate access to certain buildings and services. Also, AI applications such as facial recognition are being pitched as a solution to identify individuals who came in the vicinity of infected people. One recurring worry is that these data-driven solutions infringe on citizens’ privacy. For instance, people could lose control over their personal data and be identified and categorized without their consent.

Hoan Ton-That, co-founder of the facial recognition company Clearview AI, explains that these kind of solutions “take a little bit of our privacy” (NBC News, 2020) but are needed to solve the major health problem that is imposed on us. In other words, the good weighs out the bad here. Yes, we lose some of our privacy, but in the end, we are better off because health is more important in this situation.

This way of thinking is consequentialist in nature. What action maximizes well-being: “Safeguarding health” or “safeguarding privacy”? By framing the problem in this way, the focus lies on the consequences of the action, instigating a cost-benefit analysis. It presumes that in some way it is possible to quantify how much good we gain by safeguarding health and how much by safeguarding privacy. If safeguarding health by intro-

ducing these privacy-unfriendly technologies maximizes our overall well-being, then it is our moral duty to do so, even if in the process some people suffer from their privacy being violated. ◀

20.2.3 Deontological Ethics

In **deontological ethics**, an action is considered morally right if you act from your **moral duty**. An influential account of deontological ethics is Kantian ethics, developed by the eighteenth-century German philosopher Immanuel Kant ([1785] 1997). The focus in Kantian ethics lies not on the consequence of the action (as is the case with consequentialism) but on **the reason** for engaging in a certain action. According to Kant, an immoral action is an action that is contrary to reason. A morally right action is one that conforms to the moral law or what he called the “categorical imperative.” Two formulations of the categorical imperative include the following: (1) “act only on that maxim which you can at the same time will that it should become a universal law” and (2) “act to treat humanity, whether in your own person or in that of any other, always at the same time as an end, never merely as a means.” When considering a course of action, one should examine whether one’s reasons for one’s action—or what Kant calls one’s “maxim”—conform to the categorical imperative.

If we take the first formulation, it becomes clear that it is morally unacceptable to tell lies. We cannot imagine a world in which everyone was to have the “maxim” of lying, because there would be no world to imagine in which everyone tells lies and I am still able to act on my maxim. But also, the second formulation shows why from a Kantian viewpoint this would be morally wrong. Lying to people inherently means failing to treat them as an end—that is, failing to respect their **capacity for reason**—thus failing to respect their **humanity**.

▶ Deontological Ethics in Data Science Practices

In 2020, in a landmark ruling, the regional court in The Hague (the Netherlands) came to the judgment that the use of SyRI—a welfare fraud risk profiling system established by governmental agencies—was unlawful. The system analyzed different data sources—from data on income and house ownership to data on the use of water and energy—to score people. The highest score would lead to the label “worthy of investigation” (BVV, 2018). This investigation could for instance take the form of a house visit to check the legitimacy of a received benefit or be a hearing as a preparation for sanctions.

In its ruling, the court particularly emphasizes the lack of transparency of the used system. Citizens do not know that they are being scored, it is not clear to the public how the government came to a certain conclusion, and citizens cannot to a reasonable degree follow what happens with their data.

The importance the court attaches to transparency can be explained from a Kantian point of view. By not providing citizens with an explanation of how the system comes to a certain decision, it deprives them from the opportunity to reflect on it. Citizens cannot make up their mind whether they agree with the outcome. The opaque functioning of

the system also hampers their possibility to contest the decision as they cannot follow the steps that led to the outcome. All in all, this system does not appreciate the rationality of human beings and treats them merely as a means in the functioning of the system. ◀

20.2.4 Virtue Ethics

Virtue ethics, contrary to the two previous kinds of moral theories, does not focus on a moral rule or on consequences of actions, but on the character traits or **virtues** of a person. Virtue ethicists focus on the particularities of a certain situation and tailor one's actions to the demands of the specific context in which one acts. In this view, ethics is first and foremost about developing **practical wisdom**: the ability to determine what is morally required, even if it concerns a new or unusual situation where general rules cannot easily be applied. Virtue ethics is about being a good person. It is focused on developing the necessary virtues to live a good, flourishing life. When there is something ethically significant at stake, one has to ask the question “what should a virtuous person do in this situation?” and follow their example.

The Greek philosopher Aristotle (4th century BC, 1984) is one of the founding fathers of virtue ethics. He developed the theory of the golden means. The virtues needed to live a flourishing life, Aristotle finds, are located **in the middle of two extremes**. For instance, courage is in the middle of cowardice (one extreme) and foolhardiness (another extreme). It is cowardly to run away from all danger, yet it is foolhardy to risk too much. In order to become a virtuous person, it is not sufficient to merely possess these virtues; one also has to bring them into practice. **Moral exemplars**—people who already have developed these virtues and act virtuously—are important in this as they show us the way forward. **Training** is also an important aspect of virtue ethics. We try, fail, and learn from our mistakes in our quest of becoming more virtuous. In order to come to a morally sound evaluation and action, you will have to develop the necessary **character traits** and ask yourself “would a virtuous person lie?”

► Virtue Ethics in Data Science Practices

In 2018, Google employees protested against the involvement of their company in project Maven which would focus on improving the analysis of video images captured by drones in collaboration with the Defense Department in the USA (Hicks, 2018). Employees worried that this technology could be used for lethal purposes, for instance by picking out human targets for air strikes. Google employees first raised their concerns within the company. When this did not lead to a satisfying answer, their protest became more vocal. About 4000 Google employees signed a petition demanding to change Google's policy and to commit to never engaging in building warfare technology. Some of them resigned in protest against Google's activities. The protests of the Google employees resulted in Google ending the project.

From a virtue ethics perspective, the protesting employees of Google could be considered moral exemplars. In this very precarious situation, they were able to identify what was at stake and embrace their responsibility. They took into account the particu-

larities of the situation and tailored their actions accordingly. In doing so, they brought certain character traits into practice. They showed courage by talking truth to power, cooperativeness by working together to draw up the letter, and compassion with possible victims of the technology. ◀

The mere fact that we discuss three of these moral theories already indicates that there is not always a clear consensus on what makes an action morally right or wrong, and good or bad. This however should not lead to a kind of “anything goes” relativism. On the contrary, it should urge data scientists to put in a lot of time and effort in explicating the reasons and principles they base their decisions on. They should **actively look for others** to share their moral reflection with and see if there might be valuable counter-arguments they did not consider (also see Leonelli, 2016).

It unfortunately goes beyond the scope of this chapter to critically reflect on these moral theories, but hopefully you already thought of some critical questions yourself, when going through this short overview. For instance: How can we objectively operationalize and weigh how much happiness is brought forth by an action? Do I think it is acceptable if some people suffer if the majority benefits? Is it possible to always consider whether the reasons for my action are universalizable? If I have to act as a virtuous person, how can I identify a moral exemplar in everyday life?

As we will see later on in this chapter, ethics and ethical reflection are not the same as following a recipe. It does not guarantee that you end up with a morally excellent dish. It is an open-ended endeavor, requiring constant attention.

20.2.5 The Focus of Academic Data Ethics

Let us now turn to **data ethics**—or sometimes also referred to as AI ethics—as a specific subset of ethics. In the academic field, data ethics is defined as:

- » a new branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values) (Floridi & Taddeo, 2016, p. 1).

In this emerging field, academics coming from different backgrounds—ranging from full-fledged ethicists to historians, ethnographers, legal scholars, computer scientists, and mathematicians—focus on the particular challenges brought forth by data science practices. The moral theories, we briefly discussed above, fuel some of these data ethics studies. For instance, one can analyze moral questions brought forth by the introduction of self-driving cars through a moral theory lens or a combination thereof (Nyholm, 2018). However, it is good to note that sometimes these

moral theories are only implicitly working in the background or are simply absent, for instance, when the research is more descriptive in nature or takes a specific value—e.g., fairness or justice—or a human rights-based approach as a starting point.

One example of this background assumption of moral theory is the “Moral Machine” project (see ► Chap. 21) created by MIT researchers and later published in *Nature* (Awad et al., 2018). The project involved crowdsourcing online from 40 million people their opinion on decision-making by machines in relation to various ethical questions, starting from the example of a self-driving car. The project originated from a well-known thought experiment, the “trolley problem,” which involves an observer having to choose whether to save one person or another (or a group) from a runaway tram. The problem has traditionally been analyzed as one of doing versus allowing harm (Woollard & Howard-Snyder, 2002), but the MIT project lends a different dimension to the use of the problem by adding the subject of driverless cars. Now, instead of acting as “an observer” and thinking in the abstract, the experiment asks the participant to translate the thought experiment into the real world and apply it specifically to the development of a consumer good, namely a particular type of car.

This framing adds several assumptions not contained in the original: that the presence of cars on the roads is both inevitable and necessary, that people must accept a certain number of deaths as a result, and that the only element of moral choice involved is in shaping how those deaths occur by deciding between automated error and driver error.

An alternative approach might broaden the scope of the ethical question to ask what values should be central in shaping transport policy overall. For instance, we might ask whether transport policy should center on public forms of transport or incentivize car driving as the main mode of travel; whether all the deaths associated with the auto industry, rather than just traffic accidents, should be taken into consideration when deciding how people should travel (including pollution and contributions of the energy sector to climate change); or whether an individual rather than collective perspective on this problem is the most justified approach. By positioning a single level of the problem as the ethical dilemma, the researchers implicitly close off other possible routes of inquiry.

In the definition of data ethics from Floridi and Taddeo, the emphasis lies on **the technological aspects** (processing data and algorithms) on the one hand and **the practices** in which these technological aspects are embedded on the other. As to the former, data ethics may look into challenges related to data use (e.g., data collection, analysis, and dissemination can infringe on the privacy of people) or use of models (e.g., a model can misclassify people causing harm) (Saltz & Dewar, 2019, p. 206).

As to the latter, moral questions concerning practices can arise on both the individual and the organization level. On the individual level, enquiries such as “what virtues should a tech employee or a data scientist nurture?” (Vallor, 2016) or “how to be an accountable data scientist?” can arise. On the organizational level, questions related to how a company is structured can be relevant, for instance:

“what values are present in the code of conduct of a tech company?” (Hagendorff, 2020) or “does the business model of a data-driven company align with key values such as the human dignity and freedom of expression of citizens?”

This further refinement of the definition put forward by Floridi and Taddeo would mean that by and large, data ethics focuses on moral challenges arising on three different, yet related, levels: **the technological**, individual, and **organizational levels**.

20.3 Data Ethics in the Commercial Domain

Data ethics is not confined to academia; it also gained traction in the commercial domain. Pushed by all the incidents that surface—from data leaks and fake news to manipulating people online—and the popular backlash this has created, tech companies as well as individuals (including data scientists) working in the tech industry look for ways to improve their operations in the hope to counter these incidents.

It is a long-standing belief that if there is a decline in trust in a company or service, people will try to avoid using it. This would be detrimental to businesses as well as to society as a whole which has gotten increasingly dependent on the data-driven infrastructure. However, **trust without trustworthiness** is an empty shell. It is easy to persuade customers to trust your product or service by investing in shiny interfaces and easy-to-use products. If behind the interface, commercial parties try to manipulate people and data leaks away or is being sold, it is just a matter of time for the next malpractice to become public and trust to erode (Keymolen, 2016, 2017). If data-driven companies and data science as a profession want to be in it for the long run, they will need to become (more) trustworthy. Data ethics has been regarded as one important means to becoming more trustworthy and to establish, regain, and maintain consumer trust (Hasselbalch & Tranberg, 2016).

20.3.1 Technological Level

Corporate data ethics takes many shapes. Similar to the academic domain, also in the commercial domain, data ethics initiatives encompass the technological, individual, and organization levels.

On the technological level, both multinational companies and start-ups have formulated **ethical design principles** to ensure that their products and services are based on key values such as transparency, accountability, fairness, and nondiscrimination. At first sight, these are values which we can all easily agree upon (after all, who can be against fairness or nondiscrimination?). However, when one attempts to provide a definition of these values, let alone implement them in a product or service, it becomes complicated rather fast (Mittelstadt, 2019).

For instance, there is a lot of discussion on how to operationalize **fairness in AI** (Suresh & Guttag, 2019). Should fairness be defined on a group level (group parity)

or an individual level (individual parity) (Chouldechova, 2017)? Should fairness predominantly be about fair outcomes or fair procedures (Grgić-Hlača, 2018)? Is it possible to have both? In line with the diversity we already encountered in the brief overview of moral theories, we also find here, in the down-to-earth, practical-orientated commercial data ethics, that there is no clear answer to the question of what the meaning and content of these values should be (Jobin et al., 2019). When the political philosophy of fairness is brought into contact with computer scientific approaches to fairness, an irresolvable conflict appears to occur because “often, a contextually appropriate approach to fairness which truly captures the essence of the relevant philosophical points may hinge on factors which are not typically present in the data available” (Binns, 2018, p. 9).

On the problem of defining values and embedding them into technological design and use, academic ethics might come to the rescue as here different approaches have been developed to define and translate, in a systematic manner, values in technologies. These approaches could be of use to companies struggling with operationalizing their design principles. One of these approaches, **value-sensitive design** (Simon, 2017; Chen & Zhu, 2019), which will be addressed in ► Chap. 21, but also other methods have been developed, such as the **values that matter approach (VtM)** (Smits et al., 2019).

What oftentimes is emphasized in these approaches is that as values are not clear-cut concepts but mean different things to the various stakeholders involved, it is of utmost importance to get the **input of these stakeholders** in order to understand these different interpretations. Actually, already from the start, when defining the problem that will steer the data science activities, it is crucial to engage with the communities who will be affected by the tool or service you are developing. Their problem statement should be leading; the technology should follow.

As will be explained more extensively in ► Chap. 21, **technology is not a neutral instrument**. It can consolidate or break down existing power relations, it can strengthen the agency of end users or neutralize their actions, and it can give access to services or exclude people. You cannot foresee all these possible consequences on your own. Even when you include stakeholders, there will always be consequences you did not anticipate. Therefore, it is also important to consult experts of the domain or sector in which you are working. Such a consultation goes beyond a mere literature review. It is recommended to engage with experts who have hands-on knowledge and experience in your application field. As it turns out, real life is always much messier and more complex than what data can tell you. This interactive approach will give you a better chance at developing a rich understanding of the context you are working in and at grasping the meaning of the values you want your product or service to support.

Not merely in the academic domain there is attention to developing ways to base technology developments on values. Also, companies, public organizations (or together in public-private collaborations), as well as governmental actors are coming up with strategies to fruitfully implement values. These include guidelines and checklists (e.g., RSS and IFoA, 2019), questionnaires, interactive websites, case studies, frameworks (e.g., PartnershipOnAI, 2019), value canvasses, impact

assessments, and many more. Different motivations can be distinguished for developing these data science and AI ethics documents: ranging from being motivated by social responsibility and using it as a tool for change to seeing it as a competitive advantage to be acknowledged as “ethical” (Hasselbalch & Tranberg, 2016; Schiff et al., 2020).

While on the one hand it is promising that so many strategies are being developed, the current multiplicity of strategies makes it also difficult to decide which is the best fit. While there are some factors which could indicate success of these AI ethics documents—e.g., engagement with the law, specificity, reach, enforceability, and iteration and follow-up (Schiff et al., 2020, p. 156–157)—no extensive, comparative research has been done yet to establish the effectiveness of any of these approaches.

20.3.2 Individual Level

Empirical research indicates that AI practitioners (such as data scientists) see themselves as **partially ethically responsible** for the societal impact their applications have on society. However, they also express that their agency is to a large extent **constrained by powerful company and governmental forces** (Orr & Davis, 2020). Indeed, we witness a growing interest in **ethical** and **professional standards** to guide and steer the actions of data scientists. Codes of conduct—documents in which organizations (e.g., companies or professional associations) lay down guidelines for professional behavior—are being developed to guide professionals in the field, increase moral awareness, and stimulate ethical discussion amongst peers and within a company (Van de Poel & Royakkers, 2011, p. 32–42).

Different sorts of codes of conduct exist. **Aspirational codes** are directed to the outside world and express what a company stands for. For example, Microsoft (2020) has developed responsible AI principles that guide their business activities. **Advisory codes** are focused on professionals and aim at assisting them in making moral decisions in their work. For example, a professional code for data scientists has been developed to guide them in their work (Oxford-Munich, 2020). There are also **disciplinary codes**. These lay down some ground rules to ensure that the actions of employees meet certain standards and are predominantly focused on the internal functioning of the company (Van de Poel & Royakkers, 2011, p. 32–42). It is important to note that, generally speaking, codes of conducts are a **form of self-regulation**. They are generally not required nor enforced by law. They have no clear legal status. This means that customers cannot go to court when they believe that a company or a company employee has violated their own code of conduct. Furthermore, companies have no legal responsibility to make their internal codes of conduct accessible to users, underlining that such codes are not designed to serve as formal regulation.

Next to these codes of conduct, there is also another important development occurring on the individual data ethics level, which can be framed as “**data ethics from within.**” This refers to tech workers speaking truth to power by protesting

against their own companies when they believe that it operates in an unethical way. The Maven project, briefly discussed above in ► Sect. 20.2.4, is an example of such a “data ethics from within.” Movements of critical tech workers see the light, demanding to know which goal the technology they are working on actually is going to serve. They sign petitions, protest the executives of their companies, and sometimes even quit their jobs in order to put pressure on companies and society to intervene. Especially in a job market where there is a high demand for data scientists, computer scientists, and other key technical experts, the influence and power of these individual tech employees should not be underestimated, certainly not if they find a way to organize themselves and collectively voice their concerns and protest dubious corporate actions.

20.3.3 Organizational Level

Spurred by the commercial data ethics trend, companies are also investing in all sorts of **organizational innovations**. They set up ethics boards to review complaints from within as well as from outside the company. Ethicists are hired to enrich design teams, and ethics communities are installed to monitor the companies’ activities and to advise on ethically relevant issues.

In an ideal situation, the technological, individual, and organizational levels of commercial data ethics interlock. Data ethics is a “collaborative process” and is always “in flux” (Orr & Davis, 2020, p. 13). Codes of conduct and design principles are operationalized in a way that they can actually, in a meaningful way, guide the actions of employees. Critical employees are acknowledged as an asset of the company, and the company is structured to include their contributions in the decision-making process. Ideally, companies are set up to be accountable to both their employees and the outside world and more specifically to the communities their products and services are affecting.

However, as we already mentioned, all these data ethics initiatives are voluntary. Every company can publish its list of ethical design principles on their website, lure virtuous data scientists to join the company with raving ethics statements, and present themselves to the outside world as an ethical and sustainable company taking its end users’ interests at heart, while in fact they could not care less.

It is therefore crucial that all these initiatives come with **organizational enforcement mechanisms** in the form of structural accountability within and beyond the firm, such as reporting requirements and auditing of that reporting; otherwise, they are merely paying lip service to data ethics. Research indicates that more work needs to be done on that matter (Hagendorff, 2020). For instance, of more than 160 AI ethics guidelines that were collected, only 10 had proper enforcement mechanisms set in place. Moreover, these codes of conduct and ethical design principles remain rather vague and abstract, making it hard to actually implement them. This begs the question: guidelines that “can neither be applied nor enforced,” aren’t these “more harmful than having no ethical guidelines at all”?

20.4 Law and Data Ethics

In the commercial domain, data ethics becomes **ethics washing** when it is used to avoid legal regulation (Wagner, 2018). Ethics washing is a process where a firm performs ethical behavior to deflect criticism of harmful practices and thus “wash” its reputation, without changing its business model. One example of this is the US data analytics and surveillance firm Palantir sponsoring privacy law conferences while also developing surveillance systems used to separate immigrant families in the USA (Guardian, 2019). Data ethics as a self-regulation strategy is then not so much put to use to improve accountability but to avoid—stricter—top-down regulatory measures. A genuine data ethics strategy can only be developed if it knows its place in relation to the law. Ethics is, in the words of Hildebrandt (2020, p. 297),

- » both more and less than law: it is more because many ethical concerns are not addressed by the law and less because the outcome of ethical considerations are not necessarily transformed into legal norms and thus not enforceable by way of law.

Thus, when we apply this to commercial data ethics, we find that data ethics can be more than law as it can set standards that are not required by law and give guidance in situations which are not directly covered by law. For instance, data scientists can choose to treat anonymous data as personal data and adhere to the General Data Protection Regulation (GDPR, also discussed in ► Chaps. 17 and 21) because they find it important to foster privacy, not because they are legally obliged to. In that case, they go beyond what the law expects from them and enter the realm of data ethics.

Law provides an action space in which companies can develop data ethics practices; however, these data ethics practices cannot and should not replace law, because, most importantly, law provides a kind of **closure** which ethics cannot (Hildebrandt, 2020). In a well-functioning democratic society, it is of utmost importance that it is transparent and foreseeable what kinds of behavior are acceptable and which are not (Tamanaha, 2004, 2007).

For instance, the idea that you should not lie to someone is of such importance to a thriving democratic society that it has transformed from the ethical realm into a legal norm. For instance, when testifying in court, you have to promise to tell the truth and committing fraud (also a form of lying) is against the law and therefore punishable.

What makes a legal norm different from an ethical one is that it is both **foreseeable and enforceable**. In other words, you know beforehand what action is deemed as appropriate and what to expect if you do not adhere to the rule. This is a kind of clarity and power which ethics cannot provide. We already saw in our introduction to academic ethics that different moral theories exist, providing different rationales for what counts as a good action. This makes ethics more open-ended in character, enabling constant enquiry in and reflection on our decisions and actions.

In the most ideal situation, data ethics practices inform our actions within the action space provided by the law and encourage us to exceed a checkbox mentality in order to develop data science practices in which responsibility and accountabil-

ity are engrained. In these ideal situations, data ethics enables data scientists to develop and nurture their moral competences, and it facilitates companies to develop tailored data ethics approaches that fit their company's profile and organization.

However, in our current time frame, rapidly evolving technologies seem to challenge the ability of the law to provide and enforce this necessary action space. For instance, developments in AI may lead to complex questions concerning explainability, fairness, and accountability which require us to rethink and reinterpret existing legal norms. These democratic processes, however, take time.

At first sight, it therefore might look attractive to let data ethics, being much more agile and flexible than law, quickly fill that void. Data ethics then no longer is a way to develop data science practices **within this action space**, but it becomes the dominant provider **of that action space**.

However, because data ethics misses the power of enforcement and its flexibility does not match with the foreseeability, we expect from a democratic action space—people decide on the ethical path to follow through a process of reflection rather than following rules laid down in law, so that there might be a different “right” answer to a given question depending on the context—it can never provide the necessary closure. Consequently, an action space built on commercial data ethics will result in a scattered patchwork of many different, fluid action spaces in which companies can set their own rules of the game and other actors (both citizens and governments) will just have to play along. This leads to a **power imbalance** that will actually diminish key values such as equality, fairness, and justice instead of safeguarding them.

This severe misuse and even abuse of data ethics have led the policy domain to question whether data ethics as a cure is not actually worse than the illness itself. It even resulted in what Bietti (2020) refers to as **ethics bashing**: charging against the whole field of ethics because of these instances of abuse.

Ethics can **inform the democratic processes** that lead to an updated action space for data science practices, if and only if ethics is understood in a **broad sense** and not in a narrow sense of solely commercial data ethics (also see Taylor & Dencik, 2020). For instance, this broad perspective entails that there definitely is a role for ethicists—but also other academics, such as philosophers, social scientists, and scholars in science and technology studies (STS)—to share their knowledge with a wider audience and inform the public and political debate. Ethics can for instance “act as a meta-level perspective from which to consider any disagreement relating to the governance of technology” and it adds “a layer of rigorous principled thinking to value laden discussions” (Bietti, 2020, p. 5).

“Ethics from within” too should be part of this public debate as there is much value in the first-person experience and knowledge of tech employees—such as data scientists—who are actually building the data-driven technologies. And last but not least, the communities which might be affected by these data science practices should be heard and consulted in order to arrive at an action space which safeguards not just the interests of data-driven companies but first and foremost takes the interests of citizens at heart (Taylor & Dencik, 2020).

It goes without saying that organizing such democratic processes is hard and cumbersome. Providing a level playing field where all these voices are actually heard is a societal and political challenge that demands for a slower pace than some techno-optimists (including data-driven companies) are probably hoping for.

All in all, as data science (both as a field of research and commercial activity) already has and will continue to have a fundamental impact on the organization of society, the time is now to critically assess this power and to ensure that sufficient checks and balances are set in place to guarantee that it will actually contribute to a thriving society instead of abolishing it. Data ethics can certainly contribute to developing more responsible and accountable data scientists and data-driven businesses but should always be located within the action space provided by the law.

20.5 Data Ethics and Data Science: Are They in It for the Long Run?

This chapter is named “Data Ethics and Data Science: An Uneasy Marriage?” By now, you should be able to grasp what this “uneasy marriage” is referring to. On the one hand, it is conspicuously clear that data science and data-driven businesses have an ethically significant impact on our society. They mediate key aspects of everyday life: from education to social relations, from our interaction with the government to the way we get our news. Consequently, it seems only logical that we design and organize data science and the businesses it drives in a way that reflects the moral values we care about. Data ethics can help doing this. At first sight, data science and data ethics are a perfect match.

However, at the same time, we also established that data ethics is not a homogenous thing. It is a field of research in academic ethics and a business strategy, and sometimes these two come together. It can be focused on people, technology, and organization of a business. It can be used to interpret the action space law provides, and it can be abused to circumvent legal regulation.

What makes data ethics so attractive, especially in the commercial domain—is its flexibility! The chance of doing “the right thing!” turns out to be also its greatest weakness. As it turns out, data ethics becomes heavily overloaded when it is put to use to provide closure, as its open-ended nature actually does not support such an application.

Maybe, data science, being head over heels, rushed into this relation without fully understanding what data ethics is actually about. After all, data ethics as well as data science are both still rather young. However, in the darkest scenario, data ethics in the hands of data-driven businesses becomes a vehicle for intentional, malicious conduct, when it is used to hamper and ditch legal regulation. All in all, it becomes quite clear that data ethics and data science have not arrived at a calm place yet. Thus, what can we expect for the years to come? As a way of concluding this chapter, we will briefly touch upon three developments we expect this relation will take.

Conclusion

— New Regulation

Some issues are too important to leave to the discretion of companies. As long as data science is based on data sourced mainly from society, and the outputs of data science have effects upon society, both of which are inherent characteristics of its discipline and practice, data science must be regulated effectively to make sure that it is beneficial toward society. This includes its societal, political, and economic effects: debating what that beneficence consists of is, similarly, a task for society as much as for firms themselves. Regulation shapes this normative connection between business and society, but the effectiveness of existing regulation of data science and data technology in general is currently limited by a lack of coherence between modes of practice and bodies of law. A joined-up approach to regulation would take into account consumer and private law as well as data protection and would connect to human rights and internationally enforceable provisions for data science performed across borders and countries. Accountability needs to take on an international dimension, as can be seen by the failure of law, regulation, and politics to counter the violations of companies such as Cambridge Analytica.

— Responsible and Accountable Data Scientists

We may expect that data science as a profession will mature and that, over time, professional requirements will be established. These requirements will not only refer to the technical standards of the work, but also to its societal impact. This entails that data scientists will need to find ways to deal with their societal as well as their political agency. This goes beyond “good intentions” but will include “rigorous evaluations” and the need to explicate “political commitments” (Green, 2018, p. 45). This will also lead to a bigger focus on personal responsibility, but also accountability. In such a system, professionals will be held accountable on a personal level, not merely on a company level, when data-driven solutions fail to adhere to certain standards or it is proven that certain professional standards were not followed in the design process. Furthermore, the linking of accountability to responsibility is essential. The current computer scientific approach to “responsible data” is largely technical in nature, focusing on limited and formal requirements rather than the more complex demands of democratic and legal accountability. As such, it is currently insufficient to ensure beneficence or redress where rights or principles are violated.

— Societal Demand: Data-Driven Businesses 2.0

It is to be expected that in the years to come, societal awareness of the influence and contributions of data science will further grow. This could result in an increase in critical citizens who will demand from their governmental institutions stricter regulations and stronger enforcement. Moreover, in their role as consumers, citizens could press data-driven companies to invest in responsible business practices. Voicing their concerns and demanding new services might actually be a big push for business reform. However, if customers and employees lose trust, they might consider to “vote with their feet” (Hirschmann, 1970) and leave the

company and the services it offers, looking for more privacy-friendly and fair alternatives. In itself, this growing demand for more ethical data science practices will probably not be sufficient to adapt and/or terminate fishy business models such as the “paying with your data” approach or opaque data broker practices. However, it can instigate the development of more sustainable data science business models, steering away from the Silicon Valley start-up model that is solely focused on “fast growth.” New enterprises that get rid of this “move fast and break things” mentality can actually, from the very beginning, structure their business and develop their products and services with a genuine and clear data ethics mindset. This will help to ensure that data ethics becomes an equal partner in the relation and does not end up being just a fancy add-on.

Discussion Points

1. What do you see as the main ethical issues arising in data science? Consider how data science exerts power and influence in relation to society, how its impacts are distributed, and what kind of influence civil society can exert over the work of data scientists.
2. What do you think are the most important virtues a data scientist should develop? Review the moral theories outlined in this chapter, and how data scientists’ choices can lead to either beneficence or potential for harm.
3. Can data ethics be a competitive advantage for a company? Evaluate the potential disadvantages or advantages for firms in conducting data scientific work that, as a by-product, creates harm to society, and the ways in which not engaging in such work may affect firms.

Take-Home Messages

- Ethics is a branch of philosophy that revolves around the question: “How should one act?” In a systematic way, this discipline studies the reasons and standards underpinning our actions and investigates what makes our actions morally right or wrong, and good or bad.
- Data ethics is a branch of academic ethics as well as a business strategy data-driven companies develop to deal with the societal impact of their products and services.
- What makes a legal norm different from an ethical one is that it is foreseeable and enforceable. In other words, you know beforehand what action is deemed appropriate and what to expect if you do not adhere to the rule. This is a kind of clarity and power which ethics cannot provide.
- When data ethics is used to circumvent the development or enforcement of regulation, it is called ethics washing.

References

- Aristotle. (1984). *The complete works of Aristotle: Revised Oxford edition*. Edited by Jonathan Barnes. Princeton University Press.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). *The Moral Machine experiment*. *Nature* 563(7729):59–64.
- Bentham, J. ([1789] 1996). *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*: Clarendon Press.
- Bietti, E. (2020). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.
- Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In: *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency* (pp. 149–159).
- BVV. (2018). Wat is Syri. Retrieved from <https://bijvoorbautverdacht.nl/wat-is-syri/>.
- Chen, B., & Zhu, H. (2019). Towards value-sensitive learning analytics design. In: *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163.
- Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160360. <https://doi.org/10.1098/rsta.2016.0360>
- Green, Ben. 2018. Data science as political action: grounding data science in a politics of justice. *arXiv preprint arXiv:1811.03435*.
- Grgić-Hlača, N. (2018). Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian.
- Guardian. (2019). Palantir has no place at Berkeley: They help tear immigrant families apart. *The Guardian*. Retrieved from <http://www.theguardian.com/commentisfree/2019/may/31/palantir-berkeley-immigrant-families-apart>.
- Guardian. (2020). Hundreds of Amazon warehouse workers to call in sick in coronavirus protest. *The Guardian*. Retrieved from <http://www.theguardian.com/technology/2020/apr/20/amazon-warehouse-workers-sickout-coronavirus>.
- Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Mind*. Machines:1-22.
- Hasselbalch, G., & Tranberg, P. (2016). *Data ethics: The new competitive advantage*. Publishare.
- Hicks, Mar. (2018, November 9). The long history behind the Google Walkout. *The Verge*. Retrieved from <https://www.theverge.com/2018/11/9/18078664/google-walkout-history-tech-strikes-labor-organizing>.
- Hildebrandt, M. (2020). *Law for computer scientists and other folk*. Oxford University Press.
- Hirschmann, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Harvard University Press.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
- Kant, I. ([1785] 1997). *Groundwork of the metaphysics of morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press.
- Keymolen, E. (2016). *Trust on the line. A philosophical exploration of trust in the networked era*. Wolf Legal Publisher.
- Keymolen, E. (2017). Trust in the Networked Era: When Phones Become Hotel Keys. *Techné: Research in Philosophy and Technology*, 22(1), 51–75.
- Leonelli, S. (2016). Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160122.
- Microsoft. (2020). *Responsible AI principles*. Retrieved from <https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6>
- Mittelstadt, B. (2019). AI Ethics—Too Principled to Fail? *arXiv preprint arXiv:1906.06668*.
- NBC News. (2020). *Controversial tech company pitches facial recognition to track COVID-19*. Retrieved from <https://www.nbcnews.com/now/video/controversial-tech-company-pitches-facial-recognition-to-track-covid-19-82638917537>.

- New Statesman. (2018). *Cambridge Analytica and the digital war in Africa*. Retrieved from <https://www.newstatesman.com/world/2018/03/cambridge-analytica-facebook-elections-africa-kenya>.
- New York Times. (2019). SpaceX Unveils Silvery Vision to Mars: 'It's Basically an I.C.B.M. That Lands.' *The New York Times*. Retrieved from <https://www.nytimes.com/2019/09/29/science/elon-musk-spacex-starship.html>.
- Nyholm, S. (2018). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, 13(7), e12507.
- Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society*:1–17.
- Observer. (2020, January 4). Fresh Cambridge Analytica leak 'shows global manipulation is out of control.' *The Observer*. Retrieved from <https://www.theguardian.com/uk-news/2020/jan/04/cambridge-analytica-data-leak-global-election-manipulation>.
- Oxford-Munich. (2020). *Code of conduct*. Retrieved from <http://www.code-of-ethics.org/code-of-conduct/>.
- PartnershipOnAI. (2019). *Human-AI collaboration framework and case studies*. Retrieved from <https://www.partnershiponai.org/wp-content/uploads/2019/09/CPAIS-Framework-and-Case-Studies-9-23.pdf>.
- RSS, and IFoA. (2019). *A Guide for Ethical Data Science*. Retrieved from <https://www.actuaries.org.uk/system/files/field/document/An%20Ethical%20Charter%20for%20Data%20Science%20WEB%20FINAL.PDF>.
- Saltz, J. S., & Dewar, N. (2019). Data science ethical considerations: a systematic literature review and proposed project framework. *Ethics and Information Technology*, 21(3), 197–208.
- Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What's next for ai ethics, policy, and governance? A global overview. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.
- Simon, J. (2017). Value-sensitive design and responsible research and innovation. In S. O. Hansson (Ed.), *The ethics of technology: Methods and approaches* (pp. 219–236). Rowman & Littlefield.
- Sinnott-Armstrong, W. (2019). Consequentialism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Summer 2019)*. Stanford University. Retrieved from <https://plato.stanford.edu/archives/sum2019/entries/consequentialism/>.
- Smits, M., Bredie, B., van Goor, H., & Verbeek, P.-P. (2019). Values that matter: Mediation theory and Design for Values. In: *Academy for Design Innovation Management Conference 2019: Research perspectives in the era of Transformations*.
- Suresh, H., & Gutttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.
- Tamanaha, B. Z. (2007). *A concise guide to the rule of law*. FLORENCE WORKSHOP ON THE RULE OF LAW.
- Tamanaha, B. Z. (2004). *On the rule of law: History, politics*. Cambridge University Press.
- Taylor, L., & Dencik, L. (2020). Constructing commercial data ethics. *Technology and Regulation*: 1–10.
- Timmons, M. (2012). *Moral theory: An introduction*. Rowman & Littlefield Publishers.
- Vallor, S. (2016). *Technology and the Virtues. A philosophical guide to a future worth wanting*. Oxford University Press.
- Van de Poel, I. B. O., & Royakkers, L. (2011). *Ethics, technology, and engineering: An introduction*. Wiley.
- Wagner, B. (2018). Ethics as an escape from regulation: From ethics-washing to ethics-shopping. *Being profiling. Cogitas ergo sum*:84–90.
- Wired. (2018). When It Comes to Gorillas, Google Photos Remains Blind. *Wired*. Retrieved from <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>
- Wired. (2020). Scraping the Web Is a Powerful Tool. Clearview AI Abused It. *Wired*. Retrieved from <https://www.wired.com/story/clearview-ai-scraping-web/>
- Woollard, F., & Howard-Snyder, F. (2002). Doing vs. Allowing Harm. In: Zalta, E.N., & Nodelman, U. (eds). *The Stanford Encyclopedia of Philosophy (Winter 2022 Edition)*.



Value-Sensitive Software Design

Paulan Korenhof

Contents

- 21.1 Introduction – 502**
- 21.2 The Good, the Bad, and the Never Neutral – 503**
 - 21.2.1 Non-neutrality – 503
 - 21.2.2 Impact on a Micro-level – 504
 - 21.2.3 Impact on a Macro-level – 507
 - 21.2.4 In Sum – 509
- 21.3 Employing the Never Neutral – 510**
 - 21.3.1 A Challenge for Designers – 510
 - 21.3.2 Value-Sensitive Design – 511
 - 21.3.3 Values – 512
 - 21.3.4 Legal Values and Design – 513
- References – 518**

Learning Objectives

- Understand why technology is not a neutral instrument
- Be able to recognise the non-neutral impact of a particular technology
- Think about how to embed ethical values in software design

21.1 Introduction

Software is involved in almost everything in our daily lives: it is in our desktops, laptops, and smartphones, but we see it also implemented in an increasing range of common-use items like cars, bikes, electric toothbrushes, cookers, fitness devices, and toys. We use software to pay, communicate, shop online, plan a route, plan public transport, watch series and movies, decide which insurance to take, which political party to vote for, order food, check our health, and so on. And it is not just us as private citizens that so gladly use software for many things in our daily practice. Governmental institutions and business rely heavily on software for performing many of their processes. They at times even fully automate certain decision-making processes, like deciding whether someone should be given a fine for speeding, whether someone should be granted a loan or a credit card, or whether someone is a promising job applicant.

While there are many advantages to using software as an instrument to help us out with all sorts of tasks, there is a catch. The catch with software programs is that like all technology, they are inherently *not neutral*: every technology has a certain bias, a particular way in which it likely affects our practices, our choices, our perception, and how we interpret the world around us. Meanwhile, the impact of software on the lives of individuals can be high, especially as increasingly more elements of our lives are dependent on and intertwined with these applications.

The goal of this chapter is to draw the readers' attention to this non-neutral character of technology and to encourage them to try to utilise this non-neutrality in a beneficial manner. The chapter starts with discussing why technology is never a neutral instrument. With the help of various examples relating to software, the impact of technology on different elements of human life is discussed. Given the inherently non-neutral character of technology and its potential problematic impact, it is important to figure out how to reap the fruits of the technology while reducing its potential harms. The second half of the chapter therefore argues that, ideally, we should actively deal with this non-neutrality from the very beginning of technology design. In order to help designers with this, the chapter introduces the main ideas underlying value-sensitive design (VSD). Because it is not possible to provide here a complete instruction manual for value-sensitive software design, the chapter aims to give readers sufficient food for thought so that they can venture on the follow-up journey themselves.

21.2 The Good, the Bad, and the Never Neutral

In this section, we will delve into the non-neutrality of technology. First, the background of the non-neutrality perspective will be discussed. After that, the non-neutrality will be explained in more detail by approaching the technology from a micro- and a macro-level perspective.

21.2.1 Non-neutrality

The importance of technology for human life can hardly be overestimated: society and life as we know it today would not exist without the development and use of technology. Technology allows us to achieve certain goals, do and perceive things we could not do without the use of technology, and reveal the world in new ways to us. For example, we can see single cells of the body through a microscope, consult people at the other side of the planet over the phone, or look into the body with an echo device. By allowing us such new experiences, actions, and perceptions of the world, technology enables us to relate in the world in new manners and affects our interpretation of the world around us, as well as our practices and social conventions (Kiran & Verbeek, 2010; Verbeek, 2011). Due to the shaping influence of technology on our perception, experience, actions, goals, and understanding, technology transcends the role of being merely an instrument. In the last century, philosophers of technology therefore argued that technology is inherently *not neutral*: technologies can reveal the world to us in new ways; create new choices and possibilities for action; establish social identities, power relations, and occasions of inclusion and exclusion; and influence and inform us, our choices, our culture, and our world views (see, e.g., Heidegger, 1954; Ihde, 1983; Latour, 1993; Feenberg, 2002; Verbeek, 2005). Due to this non-neutrality, technology has a normative impact on the relation between human beings and their world (Hildebrandt, 2015).

The analysis of the impact and meaning of technology for human existence gave rise to different schools of thought in the philosophy of technology. Instead of dragging the reader into the discussion between the various ideas, it will be more valuable for the purpose of this chapter to take the two main directions of the perspectives into account, albeit in a simplified manner, and understand them as complementary to each other. Simplified, we can say that technology affects the way humans engage with the world that occurs stretching from a micro-level, an individual and empirical level (see, e.g., Ihde, 1983; Verbeek, 2005), to macro-level, a societal and abstract level (see, e.g., Stiegler, 1998; Feenberg, 2002). While taking the technology's impact on a more macro-level into account is indispensable for getting an understanding of the scope and depth of its impact on our lives, an analysis focused more on a micro-level can be very helpful to trace some of the problems back to particular concrete properties of the technology. However, I argue that the micro- and macro-level of the impact cannot be fully separated from each other because the micro-mechanics give shape to the impact on the macro-level, while what happens at the macro-level is bound to influence how the micro-mechanics are realised in practice. Despite this, I will for structural clarity start with a focus on the micro-level and from thereon move to a more macro-level of impact—but readers should note that these are linked.

21.2.2 Impact on a Micro-level

21

On a micro-level, technology affects the human perception, actions, practices, and goals, by letting us experience the world *mediated* by the technology (see, e.g., Ihde, 1983; Verbeek, 2005). For example, a thermometer can show us the temperature of our bodies. If the thermometer indicates 38.5 °C, we will likely conclude that we have a fever, even if we do not feel ill. By telling us that we are in fact ill, while we may feel fine, the technology affects how we understand our health. By doing so, the technology co-shapes our relation to the world (in this case, the human body). Technology affects our perception, our actions, and even how we think and what we remember. The latter was clearly shown by research into the effects of search engines on our memory: it turned out that once people know that they can rely on a search engine or the like for their information, they tend to remember where and how to find something instead of remembering the content itself that they needed to recall (Sparrow et al., 2011, p. 778).

When mediating our relation to the world, technology generally has a particular focus: it often reveals and highlights particular aspects of a technological co-shaped reality, while other elements are obscured or ignored (Verbeek, 2005, p. 131). Think for instance about what happens when you make a telephone call. When you make a telephone call, this makes the voice of the caller stand out, while the rest of the individual is concealed. The technology hereby establishes a particular relation between a human being and the world, a relation that is directed towards something (i.e. in the case of the telephone call, the technology is directed towards sound). We can therefore say that technology has a certain “directionality” (Verbeek, 2005, p. 115). This directionality is embedded in the material design of the technology. With its directionality, the technology takes a certain “stance”: it can “*suggest, enable, solicit, prompt, encourage, and prohibit certain actions, thoughts, and affects or promote others*” (Lazzarato & Jordan, 2014, p. 30). A designer will generally aim to give the technology a particular directionality by imbuing the technology with certain properties. For example, web applications of online stores and services are generally designed in such a manner that they render it impossible for users to place an online order for a product or service without accepting the company’s general terms and conditions. The directionality of the technology in this case is designed so that it ensures the legal base and protection of the company selling the product or service. However, there is a limit to the influence of the designer: in the end is the technology itself that is present in the world and with which users interact. Technology has a certain autonomous existence separate from its designers (Chabot, 2013, p. 15). As such, technology can also easily have effects or uses that are unforeseen or unintended by its designers. However, as the designers determine the material properties of the technology, they do play a pivotal role in the shaping of the non-neutral directionality of the technology they design.

The directionality of software thus depends on the characteristics and constraints of the technology itself, as well as on the choices made by its designer. The developer’s knowledge, expertise, cultural background, and limitations form the

background of the software design (Kitchin, 2017, p. 18). This takes shape in a two-step process: a developer needs (1) to interpret the task at hand and (2) translate this into code. As designers develop the software, their views on how to interpret and translate certain concepts, values, and goals form the base on which they encode procedures that determine what the software does and does not do. With this, the designer's assumptions and biases are incorporated in the software—whether it be intentional or accidental (Friedman & Nissenbaum, 1996; Goldman, 2008). Coded procedures are therefore inevitably value laden (see, e.g., Brey & Søraker, 2009; Mittelstadt et al., 2016).

► Example 1

Imagine that a company hires you to circulate a vacancy for a truck driver and select some potential candidates. You decide to develop an online form for the job application procedure. In order to apply, job applicants need to fill in their name and date of birth, upload a CV, tick a box with either male or female, and tick a box that they consent to the processing of their personal data. In order to prevent people from forgetting to add anything, you make all the fields mandatory in order to apply. While such a form may seem simple and straightforward, in this small application, we can already see many points on which the software has a certain directionality towards the world—and may even be problematic. First off, the fact that it is an online form immediately places the application process in the digital realm. As such, the application may bypass the less digital skilled or people who have limited access to the internet. Secondly, as the form needs to be fully completed before someone can apply, the software compels people to reveal all this information (or lie about it) and to identify themselves according to the options the form offers. The form expresses a certain view of the world: it presents certain elements as important about people who want to be a truck driver and expresses a binary gender perspective. Job applicants can experience this “identity fit to the software's box” as problematic: people may prefer not to give their date of birth for privacy reasons or as protection against age discrimination, people may not identify themselves as male or female or find it irrelevant to share in a job application, some people may prefer more creative freedom when applying for a job, etc. However, the choices people can make in applying for the vacancy are limited to and determined by the options offered by the online form. ◀

Giving shape to software can be especially tricky when designing software that needs to produce decisions based on particular laws. Examples are the automatic fines issued when someone is recorded by a traffic camera for speeding, or the allocation of child care subsidy based on a combination of data sets. In such cases, programmer will need to translate law or policy rules into code and may find themselves confronted with questions about when *exactly* a certain case falls under the definitions of a particular law or policy rule. When giving shape to such boundaries in code, the programmer fills in legal concepts and de facto establishes a policy rule by programming the software.

The role of design choices in software can hardly be overestimated: the code is what the software does and controls what users can and cannot do. This is what Lessig meant with his famous statement “code is law” (Lessig, 2006). However, as

Pariser explains, software code more forcefully controls user behaviour than law, at least at the outset:

21

» “If code is law, software engineers and geeks are the ones who get to write it. And it’s a funny kind of law, created without any judicial system or legislators and enforced nearly perfectly and instantly. Even with antivandalism laws on the books, in the physical world you can still throw a rock through the window of a store you don’t like. You might even get away with it. But if vandalism isn’t part of the design of an online world, it’s simply impossible. Try to throw a rock through a virtual storefront, and you just get an error” (Pariser, 2011, pp. 96–97).

The developer thus exercises a significant degree of power over users through the software’s architecture: users can only use the software as its architecture allows. In this, the software’s interface plays a key role on the one hand by suggesting to the user what software does and can do, while on the other hand it is also the realm of interaction by means of which users can operate the software. The interface thus shapes the perception of users, provides them with the know-how of the software, and offers them a particular set of actions—and without using tricks, the user’s perception and actions are commonly limited to that which is offered by the interface. Commonly, this is a graphic user interface (GUI) that hides the source code of the software and thereby often renders a substantial part of what the software actually does opaque. Moreover, the interface can be shaped to manipulate users to perform certain actions by means of persuasion or nudging (see, e.g., Fogg, 1999; Harjumaa & Oinas-Kukkonen, 2007; Thaler & Sunstein, 2009). Think for example of the various ways in which website designers try to nudge users into accepting marketing cookies by using a big green button to “accept all cookies”, versus a less visible small red button that allows a user to select a different setting (see the chapter by Gellert in this book for the explanation of consent and the use of cookies).

Depending on the interface, users are thus offered a more or lesser degree of insight into the operations performed by the software and are given certain choices with regard to the actions that they can perform. This affects the autonomy of users: their ability to self-govern, which entails the freedom to make informed decisions and shape their lives as they see fit.¹ For example, some online stores require users to classify themselves as “female” or “male” before they can place an order; other stores give them more options, like the extra option of “I’d rather not say”; while again other stores do not mark gender as a required field at all and leave it to the users to decide whether they want to fill in the field. The more freely users can choose and act, the more autonomy they have. Reducing the autonomy of users and forcing them down certain action paths can estrange human agents from the task they are performing with the software, especially if they have little insight into and know-how of what the software is actually doing behind the interface. De Mul and Van den Berg therefore argue: “Awareness of, and insight into the ‘scriptal

1 The exact definition of what autonomy entails differs somewhat per social and political perspective. For the purposes of this chapter, I kept the concept relatively open and phrased it in a manner that can give some handholds in relation to software design.

character’ of the artefact, and having the ability to influence that character, is crucial for users in the light of the delegation of their autonomy” (De Mul & van den Berg, 2011, pp. 59–60).

21.2.3 Impact on a Macro-level

Technology not only affects processes, practices, and perception on an individual level, but also influences our lives on a macro-level: it influences and even shapes societal organisation and transactions, institutions, governmental agency, politics, science, relations between individuals, and even our identity (Stiegler, 1998). Especially on the level of the functioning of companies and institutions, as well as the work of people therein, the use of software deeply affects the processes and output, which in turn can affect people outside of the organisation), and even society at large. Take for instance the use of automated decision-making software, like the automated issuing of fines for speeding. This entails a shift in decisionary power, whereby the main “decision maker” changes from a human agent who learned to employ their knowledge of legal rules and policy in order to make a contextual assessment, to software that strictly applies rules:

- » decisions are pre-programmed in the algorithms that apply the same measures and rules regardless of the person or the context (e.g., a speeding camera does not care about the context). Responsibility for decisions made, in these cases, has moved from ‘street-level bureaucrats’ to the ‘system-level bureaucrats’, such as managers and computer experts, that decide on how to convert policy and legal frameworks into algorithms and decision-trees (Noorman, 2020).

By shifting the decisionary power, such software generally reduces the space for individual discretion and gives rise to a workforce that mass produces decisions in a uniform production process on which they have little influence (Giritli Nygren, 2009; Wihlborg et al., 2016). As such, software “reframes relationships, responsibilities and competences” (Wihlborg et al., 2016, p. 2903).

Moreover, when know-how is embedded into software, the practical need for human agents to have this same know-how reduces and sometimes even disappears: a click on a button can be enough to provide users with what they need. An example of this is a bank where “customer advisers get predetermined interest rates from the IT system for their customers’ credit, but they do not know how this interest rate is calculated or what justifies it” (Spiekermann, 2015, p. 12). With the delegation of know-how to software, human agents, and society in general, are becoming increasingly dependent on software for many of their processes. Stiegler therefore argues that technology is in a sense a poison that is at the same time its own cure—a *pharmakon* (Stiegler, 2012): while software allows humans to forget knowledge and how to do certain things (poison), it at the same time remedies this loss of know-how by performing the actions for them (cure). Think for instance of phone numbers. In the pre-mobile phone era, the phone numbers were not stored in the telephone. This commonly meant that you automatically memorised the numbers of family and close friends because you had to consistently type the num-

ber in and, also, putting in some effort to remember a number was faster than having to look the number up in an address book. However, with smartphones, this need for remembering became virtually superfluous and typing in the number is unnecessary: the technology does this for us. The result is that we are far less likely to remember phone numbers unless we actively spend effort to memorise them. The effect of this becomes painstakingly clear when you lose access to the contact list in your phone.

The more dependent we become on particular software, the more power it holds over us. We can see this clearly in the use of search engines. Due to the abundance of information resources on the web, we have become highly dependent on their use to find online information. As such, this pivotal position of search engines imbues them with a significant power over the connection between users and content providers: search engines “are attention lenses; they bring the online world into focus. They can redirect, reveal, magnify, and distort. They have immense power to help and hide” (Grimmelmann, 2010, p. 435). Dropping out of a search engine’s search result list can render content nearly invisible to a significant part of the web users—with all due consequences for the web content owner as well as for searching users.

The power of software is strengthened by the trust people tend to have in the technology to fulfil its tasks properly: people tend to have an “automation bias” due to which they trust the output of software more than their own assessment (see, e.g., Skitka et al., 1999). As such, they may overly rely on software for their assessment or to quickly make decisions (Skitka et al., 2000). Combine the human inclination towards automation bias with an opaque interface that suggests an objectivity or neutrality of the software’s operations, while the software is in fact bound to harbour some (intentional or unintentional) biases and maybe even has some errors, and we have a recipe for disaster.

The scale of processing afforded by software can magnify the impact of its biases to a society-wide level. Software may “normalize the far more massive impacts of system-level biases and blind spots” (Gandy, 2010, p. 33). Take for example social media. This type of software led to changes in web culture by giving rise to new standards of what is considered “normal” (Van Dijck, 2013; Wittkower, 2014). One of the changes brought about by social media is a shift from relatively anonymous online communication to pattern communication where “individuals are increasingly known, and in fact willingly share a lot of their personal information online” (Sparrow et al., 2005, p. 283). A pivotal role here is played by the software’s default settings. The default settings set a standard for its use and require users with divergent preferences to invest time and effort in order to adjust the default (see, e.g., van den Berg & Leenes, 2013; Acquisti et al., 2015). The default settings thereby express a certain world view, a “normalcy”, with regard to its use. For example, originally on Facebook, the default setting of an account was that all user information and posts were publicly available. With these default settings, Facebook suggested that the standard was to be available, accessible, and identifi-

able as a particular offline person for a potentially worldwide audience.² Additionally, users tend to have an inclination to accept the default settings, because it “is convenient, and people often interpret default settings as implicit recommendations” (Acquisti et al., 2015, p. 512). The default settings thus strongly affect user behaviour and norms.

Last, the output of software can impact the lives of people who are not the software’s users—as well as society at large. For instance by placing certain groups of people at a disadvantage. Let us look a bit deeper into this by focusing on automated decision-making applications like those that issue fines for speeding, mark people as being fraud risks, calculate the amount people need to pay for their insurance policy, etc. In these cases, people who are not the initial users of the software are subjected to the output (decision) produced by the software. However, as the transparency of the software’s output is dependent on what is programmed *into* the software to show as output, people may be profiled and subjected to a decision of which the how, why, and what are not made clear to them. As such, it is difficult for them to figure out if an error was made, and if so, where and how. A lack of insight in what happens in software can be particularly problematic in the case of automated decision-making software used by government agencies because these agencies have the obligation to be transparent in their decisions and motivate them. Moreover, the lack of transparency and access to the same software makes it difficult for people to effectively challenge an automatically produced decision. It leads to a power imbalance by establishing an inequality of arms between a common citizen and the agency—which generally already holds a power position because people are dependent on the agency for one thing or the other. These are only some of the issues concerning automated decision-making software. The impact of automated decision-making software on our lives and world, a full discussion here is too extensive.

21.2.4 In Sum

This section discussed why software, like all technologies, is not a neutral instrument. Software has a certain directionality in which it likely affects and even changes the manner in which we work, decide, and interact. Its impact can stretch deep into society and especially into the lives of people. The question now is how should we deal with this non-neutrality.

2 Under pressure of public institutions and European legislation, Facebook eventually changed its default settings to a restricted audience and with that sets a somewhat more privacy-friendly standard.

21.3 Employing the Never Neutral

This section offers an approach on how to deal with the non-neutrality of technology. It will start by arguing for a proactive approach with regard to values in technology design. In order to give some handholds on how to start, the section then gives a general outline of “value-sensitive design”. Last, given the focus of this chapter on software, this section takes a look at the values promoted by the General Data Protection Regulation (GDPR) when it comes to the processing of personal data.

21.3.1 A Challenge for Designers

While technology is not necessarily good or bad, it is thus never neutral. The design of technology is therefore pivotal: at this stage of the process, a significant part of what a specific technology does and does not do, its directionality, is determined. De Mul and Van den Berg therefore point out that, despite the strong influence of technology on us and our world, “the *responsibility* for that world and what happens in it is still in the hands of human beings and not in the hands of the technologies. After all, human beings are the architects, designers and users of the technologies, and for that reason they are responsible for their creations and their creations’ output” (De Mul & van den Berg, 2011, p. 46). Technology is designed by us, and in many cases, we will be able to design the technology in such a manner that we can reduce, or even prevent, its problematic impact.

A way to deal with the inherent non-neutrality of technology is therefore to consciously *design* technology in a manner that it supports or promotes certain social or moral values, like freedom, safety, and privacy. Already in the design process, we should therefore be asking what the potential impact of a software application may be, and how the application should work if we want it to promote certain values, while repressing or even fully preventing the inscription of problematic biases into the technology. Of course, not all future effects and unintended uses are foreseeable (especially since real life is messy, see the point made by Keymolen and Taylor in this book), and not everything can be prevented. However, a good start is to consciously implement certain values from the first stages of the design and to try to become aware of the values that we are unconsciously building into the technology. With this, “[t]echnological innovation can become *responsible innovation*”[emphasis original] (van den Hoven et al., 2015, p. 3). This places an active responsibility on engineers. Consciously focusing on the values inscribed into the design can help to ensure that the technology meets societal needs and it helps to reduce the risk of unwanted, unintended, or harmful effects. This also beneficial for the designers and engineers, because it may avert damage to their reputation when people consider the technology to be untrustworthy or harmful. Moreover, in some cases, designing technology in such a way that it promotes particular values is even required by law. An important law in this regard is the GDPR

(see the chapter by Gellert in this book for more information about this regulation), which requires agents who process personal data to engage in *privacy by design* (Art. 25, GDPR, I will return to this later).

21.3.2 Value-Sensitive Design

One of the ways in which we can consciously aim to deal with the problematic, as well as beneficial, non-neutrality of technology is by engaging in a manner of designing that is *value sensitive*. Several approaches have been developed for explicitly taking human values into account when designing technology. These approaches “share at least four key claims: values can be expressed and embedded in technology, technologies have real and sometimes non-obvious impacts on those who are directly and indirectly affected, explicit thinking about the values that are imparted in technical design is morally significant, and value considerations should be surfaced early in the technical design process” (Friedman et al., 2017, p. 65). The most well known of these approaches is *value-sensitive design* (VSD) (for an extensive overview, see Friedman & Hendry, 2019).

The general idea of VSD was developed around the mid-1990s (Friedman et al., 2017, p. 64). VSD is an approach to technology design that takes human values into account during the whole of the design process (Friedman et al., 2008, p. 76). Van den Hoven describes it as “a proactive integration of ethics—the frontloading of ethics—in design, architecture, requirements, specifications, standards, protocols, incentive structures, and institutional arrangements” (Van den Hoven, 2008, p. 63). VSD is ongoing under development and may always be (which does not have to be a bad thing). Its general methodology still faces some challenges (see, e.g., Friedman et al., 2017; Winkler & Spiekermann, 2018)—a few of these will be discussed below. Despite the challenges, overall, VSD is a relatively practical approach concerning value-conscious technology design and can be of significant value to those who are at the heart of the design process. VSD’s methodology draws on inter alia the social sciences and human and computer interaction research (Friedman et al., 2017, p. 64). Its methodology **mixes empirical, technical, and conceptual studies** and applies these in an iterative and integrative manner throughout the design process (Friedman et al., 2008, p. 93). With this methodological mix, VSD takes an interactional stance and starts from the premise “that human beings acting as individuals, organizations, or societies shape the tools and technologies they design and implement; in turn, those tools and technologies shape human experience and society” (Friedman et al., 2017, p. 68).

► Example 2

Imagine that you want to develop software that helps people to spend less time looking at their smartphone. By means of **empirical analysis**, you can examine the context and experience of people’s current smartphone use and get an idea of their wishes and problems. In order to examine this, you would ideally use quantitative and/or quali-

tative research methods from the social sciences, like interviews, surveys, and statistical analysis. Additionally, you can use such empirical analysis to test your own design. However, these analyses are not all that is relevant. Users may not always know what they want (especially beforehand), or be aware of all the implications of what they are doing and using, nor may you have sufficient data to oversee the bigger picture. Doing a **conceptual analysis** is therefore important to get a full(er) picture of the concepts and issues involved, like the values that play a role or the broader individual and societal implications of the technology. For this **conceptual analysis**, you draw on theoretic and philosophic research that sees to the main concepts and issues that relate in one way or the other to the (to be designed) technology. Let us say that for this software, you will be reading up on philosophical accounts of agency, autonomy, nudging, manipulation, and privacy. This can in turn inform your further empirical inquiries and your technological design. It is thus important to also perform a **technological analysis** of the technology. A better understanding of the technology can be achieved by analysing its concrete mechanisms and results, as well as by looking at already existing technologies that share certain similarities and assessing their impact. Your findings of the technical mechanisms can further inform and specify your conceptual and empirical analysis, which in turn can help you to improve the design. And so goes the process back and forth until you end up with a design that is well rounded and backed by research. ◀

21.3.3 Values

In the context of VSD, the term “value” refers to “*what is important to people in their lives, with a focus on ethics and morality*”[emphasis original] (Friedman & Hendry, 2019, p. 24). The focus thus lies on social and moral values, and not on economic value. In this context, we can think of values like human welfare, trust, privacy, fairness, autonomy, universal usability, safety, health, and environmental sustainability. The values potentially covered by VSD range from those that can be found in the diverse moral philosophical theories like deontology, consequentialism, and virtue ethics (see the chapter by Keymolen and Taylor in this book), as well as personal values like preferences of taste and colour, and conventions like protocol standards (Friedman et al., 2008, p. 94).

VSD tends to base its value selection and assessment on the experiences and opinions of the stakeholders. A key element of VSD is therefore to identify the direct and indirect stakeholders and their corresponding values (Friedman et al., 2017, p. 69). Direct stakeholders are people, groups, or organisations who directly interact with the technology in question (Friedman et al., 2017, p. 76). The indirect stakeholders consist of people, groups, or organisations who are affected by the technology, but do not directly interact with it (Friedman et al., 2017, p. 76). An example of a method to get a sense of the values at stake of the direct and indirect stakeholders is to conduct semi-structured interviews (Friedman et al., 2008, pp. 100–101). However, identifying the stakeholders can be difficult, and a failure to identify a particular group of stakeholders can lead to their exclusion as well as to the exclusion of particular values (Manders-Huits, 2011; Winkler & Spiekermann, 2018). Moreover, stakeholders themselves may not always be able to oversee the

impact of particular technologies and be able to recognise which of their values may be at stake in a certain context.

In some cases, it turns out that two or more conflicting values are involved. These conflicting values do not necessarily have to originate from different stakeholders: the same stakeholder can have multiple values that may pull the design in different directions. If there is a tension between values, it is important to take this into account in the design process (Friedman et al., 2017, p. 69).

► Example 3

An example of a tension between values is the likely tension between privacy and national or public safety: while privacy generally benefits from collecting and revealing less personal data, safety generally benefits from having access to more personal data. This tension played a pivotal role in the introduction of body scanners in airports. The goal of the body scanner is to increase safety by visually showing airport security staff where on the body people carry objects. For this, they scan the surface of the body, and this can display a rather accurate view of what the naked body of the scanned person looks like. This deeply infringes the bodily privacy of those scanned. Many of the body scanner developers took this privacy infringement for granted as a plausible sacrifice in the name of safety (Spiekermann, 2015, p. 169). However, it turned out that neither the general public nor the security staff and airport operators (who had to deal with customer complaints and feared a drop in customers) were all too happy with this privacy infringement (Spiekermann, 2015, p. 169). One company took both values—that of safety and privacy—seriously and sought to design scanner software that reduced the privacy infringement while maintaining its safety goal. In the resulting design, the display of the body was replaced by an abstract stick figure outline of a body with areas in which an object was located on the body marked. With this design, the company reached its safety goal while at the same time building in privacy safeguards in the software. By taking both values seriously and embedding them in the design, the company managed to capture the majority of the market (Spiekermann, 2015, p. 169). ◀

21.3.4 Legal Values and Design

A good source to find values that value-sensitive software design should ideally take into account, is law. In the context of software design, the GDPR is of particular relevance due to its focus on data processing. The GDPR provides us with a set of values that need to be taken into account on behalf of (the protection of) “data subjects” (see the chapter by Gellert in this book for a full explanation of the term “data subject”) and society at large. Below are some of the main (derivative) values listed that can be found in the GDPR:

- Autonomy (see, e.g., informed consent, Recital 32, Art. 4(11), Art. 7)
- Privacy (see, e.g., control over own data, Recitals 7 and 68, Art. 17, Art. 21)
- Protection against power imbalance (see, e.g., Recital 43, automated individual decision-making, including profiling, Art. 22; purpose limitation, Art. 5(1)(b); data minimisation, Art. 5(1)(c); storage limitation, Art. 5(1)(e))
- Human dignity (see, e.g., Recital 4)

- Fairness (see, e.g., Recitals 39 and 60, 71, Art. 5(1)(a))
- Safety/protection (see, e.g., Recitals 1, 2, 51, 54, 78, and 108, Art. 1, Art. 6(d), Art. 6(f), Art. 25)
- Security (see, e.g., Recitals 2, 16, 19, and 49, Art. 5(f), Art. 32)
- Respect for the rights and freedom of individuals (see, e.g., Recitals 2 and 4)
- Human welfare (see, e.g., Recital 4)
- Transparency (see, e.g., Recitals 39, 60, and 71, Art. 5(1)(a))
- Economic prosperity (see, e.g., right to run a business, Recital 4)

What is interesting about the GDPR in light of this chapter, is that the GDPR even explicitly requires the embedding of some of its underpinning values in software design. Art. 25(1) of the GDPR on “Data Protection by Design and Default” states:

- » Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall (...) implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects (Art. 25(1), GDPR).

The data protection principles (the chapter by Gellert provides an extensive analysis of these principles; here, I will briefly touch upon them for clarity purposes) are listed in Art. 5 of the GDPR and state that personal data needs to be “processed lawfully, fairly and in a transparent manner in relation to the data subject (‘lawfulness, fairness and transparency’)” (Art. 5(1)(a), GDPR); the personal data can only be collected and processed “for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purpose (...) (‘purpose limitation’)” (Art. 5(1)(b), GDPR); the personal data needs to be “adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (‘data minimisation’)” (Art. 5(1)(c), GDPR); the personal data needs to be “accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay (‘accuracy’)” (Art. 5(1)(d), GDPR); the personal data needs to be “kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed (...) (‘storage limitation’)” (Art. 5(1)(e), GDPR); the personal data needs to be “processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures (‘integrity and confidentiality’)” (Art. 5(1)(f), GDPR); and last, the ones controlling the data are held responsible and need to be able to demonstrate that they comply with the data protection principles (‘accountability’) (Art. 5(2), GDPR).

A significant role of these principles is to curb the personal data that can be collected and retained about a specific data subject. Here, the adage “knowledge is

power” comes to mind. In this context, the purpose limitation principle, the data minimisation principle, and the storage limitation principle are important restrictions that curtail the power imbalance that may rise between citizens and institutions and/or corporations that can aggregate massive amounts of information about them (see, e.g., Brouwer et al., 2011). On the other side of the coin, we can find measures in the GDPR that aim to better balance the playing field by ensuring that the data subjects themselves have sufficient knowledge about how their data is processed. For example, Art. 13(2)(f) of the GDPR seeks to ensure fair and transparent processing in the case of automated decision-making, including profiling, by requiring the data controller to provide the data subject “meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject”. Art. 22 of the GDPR sees specifically to automated decision-making and requires human intervention in cases where the produced decision can make a significant impact on the life of an individual. An example of a significant impact is the automated refusal of an online credit application without human intervention (Recital 71). The automated decision-making “should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision” (Recital 71).³

The challenge of designing software value sensitively and trying to account for values like those referred to above is how to concretely embed these values in the designed product. For this, there is no one-size-fits-all method, as it depends on the values sought after and the software in question. Moreover, the “how” in itself is a topic of ongoing investigation. With regard to designing software in a privacy-enhancing manner, the article *A Critical Analysis of Privacy Design Strategies* by Colesky et al. (2016) can provide a valuable source of inspiration. The researchers identify several “privacy design strategies” that can be used to embed privacy by design. One of the suggested tactics is data minimisation; this links to Art. 5(1)(c) of the GDPR (see also the chapter by Gellert). It entails a selection in which the data that is not needed is excluded, stripped, or destroyed. Other strategies are access restriction and separation of data (Colesky et al., 2016). By isolating data collections, or by distributing them over different locations, the risk is reduced that the data is combined and provides a more detailed view on a specific individual and/or de-anonymises the individual. Another tactic they mention is abstraction (Colesky et al., 2016). If data is summarised or grouped on a more general level, the focus of the data shifts from particular individuals to a more generic level. Studies like the one performed by Colesky et al. can be an inspiration source for designers to come up with designs that realise their striven-for values.

3 However, requiring a human agent to be in the decision-making loop is no guarantee that the sought-after values are protected (Binns, 2019). Human intervention also has its up- and down-sides: humans can discriminate intentionally and unintentionally. Additionally, if a human is added into the decision-making loop, there is the risk that the human merely “rubber stamps” the made decisions to validate their outcome and bypass the further requirements of Art. 22 of the GDPR (Veale & Edwards, 2018, p. 400).

Conclusion

Technology is not neutral. It can influence and shape people's perception, what they know, what they can do, the way they engage with the world, and the way in which society, governments, companies, and others engage with them. The innovative use of technology can therefore be highly valuable, but also highly problematic. Technology has therefore been the inspiration for both utopias and dystopias alike.

Value-sensitive design is an approach that aims to deal with the non-neutrality of technology in a beneficial manner. It focuses on actively aiming to incorporate certain values in the design of technology. While VSD is not without problems and challenges, it is a promising start for designing technology that aims to be on the utopian, rather than on the dystopian, side of things. In some cases, law even requires the embedding of certain values in the software design: art. 25 of the GDPR calls for the implementation of “privacy by design” and “privacy by default”.

Designing value sensitively is not an easy task, especially because it is often difficult, if not impossible, to fully foresee the impact and use of a new technology. However, this should not stop us from trying. There is a pivotal role here for designers. Being aware of the views and assumptions that are necessarily built into the system design, they can try to do this in a conscious and value sensitive manner.

In order to start designing value sensitively, it can be of help to keep the following rules of thumb in mind:

1. Be aware of the inherent non-neutrality of what you are designing: think about what the technology adds to, takes away from, or changes in the current situation.
2. Identify which values you want to endorse with your design (e.g. you may want to design software in order to promote efficiency while at the same time safeguarding privacy).
3. Assess the impact of the design: does the technology benefit or negatively impact a specific group of people? And who are these people and what are the consequences for them?
4. Trace if you may unnecessarily or undesirably inscribe certain prejudices in the design (e.g. is the user you have in mind representative for the whole user group, or are you unconsciously designing the software in a way that only benefits a particular subset of users?).
5. Try to see if you can adjust the design in order to get rid of unintended bias or negative impact while promoting the values you want to endorse (i.e. test, evaluate, adjust).

How to (best) realise value-sensitive design (and in particular privacy by design and default given that these are required by law) is still—and with new technologies will always be—a topic of exploration and experiment. However, the first step is an awareness of a technology's non-neutrality, and a willingness to think about which values ideally should be protected in its design and how to achieve this. Hopefully, this chapter helps to readers with taking this first step.

? Question

How can designers influence the non-neutrality of software?

✓ Answer

Designers intentionally influence the non-neutral directionality of the technology by designing software to perform specific tasks and help users reach certain goals. They determine what a user can and cannot do with the program and thereby influence the non-neutrality on the user action level. This is set “in stone” in the code of the program. Additionally, the designers determine what a user perceives (in combination with the properties of the used hardware) when engaging with the application, thereby steering a user’s experience and interpretation of the technology.

An important role here is played by the image of the user that a designer has in mind. An example is the design of a website with little text and a lot of images. Users with impaired vision that depend for their web surfing on an application that reads text out loud, will have a hard time navigating the website.

Additionally, designers can influence the non-neutrality of software unconsciously by embedding their own assumptions into the design. An example is an online credit card request form that requires one to make a photo of an identity card with a smartphone directly, and not allow the uploading of files. This assumes that all internet users have a smartphone.

Furthermore, by means of user manuals and marketing, particular views on and uses of a technology can be influenced and promoted, for example by suggesting that using a particular software application can improve health, social status, friendship, and efficiency, can reduce human errors, etc.

? Question

What are the advantages of designing value sensitively?

✓ Answer

First and foremost, engaging in value-sensitive design will help designers to innovate in an ethically responsible manner by front-loading ethical values in the design of the software: in the interface, architecture, standards, specifications, incentive structure, institutional embedding, default settings, user requirements, etc.

Furthermore, taking into account the interest of the various direct and indirect stakeholders and trying as good as possible to account for their values in the design will help to generate more societal support for the use of the technology. More happy people generally means more users and user engagement.

Also, because a thorough reflection on the potential impact of the software is a necessary part of designing value sensitively, the designers (or commissioning company) are less likely to be surprised by unforeseen consequences of the technology.

? Question

Reread the text of Art. 25(1) of the GDPR cited in the text above. What is important for properly realising privacy by design?

✓ Answer

Art. 25(1) of the GDPR does not call for a flat-out implementation of privacy by design. Instead, it calls for a delicate balancing of the available technical options, interests of those involved, the purposes of the data processing, and its context, risks, and impact on people. It is thus not only privacy that should be taken into account as a value in the design: other values, like safety, human welfare, and economic prosperity, should also be taken into account and balanced with privacy. Privacy by design thus means, simply put, an active prevention of any possible privacy infringement that is not strictly necessary for realising a particular goal. As the case of the body scanners discussed in ► Sect. 21.3 shows, a careful balancing can result in a design that is able to respect conflicting values to a considerable degree: while maintaining their goal of safety, the “stick puppet” body scanners also significantly reduce the privacy infringement on those scanned. With this balance, these body scanners are a good example of privacy by design.

References

- Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, 347(6221), 509–514.
- Binns, R. (2019). Human judgement in algorithmic loops; individual justice and automated decision-making. *Individual Justice and Automated Decision-Making* (September 11, 2019).
- Brey, P., & Søraker, J. H. (2009). Philosophy of computing and information technology. In: *Philosophy of technology and engineering sciences*, Elsevier, pp. 1341–1407.
- Brouwer, E., et al. (2011). *Legality and data protection law: The forgotten purpose of purpose limitation*.
- Chabot, P. (2013). *The philosophy of Simondon: Between technology and individuation*. A&C Black.
- Colesky, M., Hoepman, J. H., & Hillen, C. (2016). A critical analysis of privacy design strategies. In: *2016 IEEE Security and Privacy Workshops (SPW)*, IEEE, pp 33–40.
- De Mul, J., & van den Berg, B. (2011). Remote control: Human autonomy in the age of computer-mediated agency. In: *Law, human agency and autonomic computing*, Routledge, pp. 62–79.
- Feenberg, A. (2002). *Transforming technology: A critical theory revisited*. Oxford University Press.
- Fogg, B. J. (1999). Persuasive technologies. *Communications of the ACM*, 42(5), 27–29.
- Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.
- Friedman, B., Hendry, D. G., Borning, A., et al. (2017). A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction*, 11(2), 63–125.
- Friedman B, Kahn PH, Borning A (2008). Value sensitive design and information systems. In: *The handbook of information and computer ethics*, pp. 69–101.
- Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3), 330–347.
- Gandy, O. H. (2010). Engaging rational discrimination: Exploring reasons for placing regulatory constraints on decision support systems. *Ethics and Information Technology*, 12(1), 29–42.
- Giritli Nygren, K. (2009). The rhetoric of e-government management and the reality of e-government work: The Swedish action plan for e-government considered. *International Journal of Public Information Systems*, 2, 135–146.
- Goldman, E. (2008). Search engine bias and the demise of search engine utopianism. In: *Web Search*, Springer, pp. 121–133.
- Grimmelmann, J. (2010). Some skepticism about search neutrality. *The next digital decade: Essays on the future of the Internet*, p. 435.

- Harjuma, M., & Oinas-Kukkonen, H. (2007). Persuasion theories and it design. In: *International Conference on Persuasive Technology*, Springer, pp. 311–314.
- Heidegger, M. (1954). *The question concerning technology. translated by William Lovitt in the question concerning technology and other essays*. 1977.
- Hildebrandt, M. (2015). *Smart Technologies and the End (s) of Law: Novel Entanglements of Law and Technology*. Edward Elgar Publishing.
- Ihde, D. (1983). *Existential technics*. SUNY Press.
- Kiran, A. H., & Verbeek, P. P. (2010). Trusting our selves to technology. *Knowledge, Technology & Policy*, 23(3–4), 409–427.
- Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1), 14–29.
- Latour, B. (1993). *We have never been modern*. Harvard University Press.
- Lazzarato, M., & Jordan, J. D. (2014). *Signs and machines: Capitalism and the production of subjectivity*. Semiotext (e) Los Angeles.
- Lessig, L. (2006). *Code Version 2.0*. Basic Books.
- Manders-Huits, N. (2011). What values in design? the challenge of incorporating moral values into design. *Science and Engineering Ethics*, 17(2), 271–287.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 2053951716679679.
- Noorman, M. (2020). Computing and moral responsibility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy, Spring 2020th Edition*. Metaphysics Research Lab, Stanford University.
- Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.
- Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, 52(4), 701–717.
- Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991–1006.
- Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776–778.
- Sparrow, B. C., Chapman, P., & Gould, J. (2005). *Social cognition in the internet age: Same as it ever was?* pp. 273–292.
- Spiekermann, S. (2015). *Ethical IT innovation: A value-based system design approach*. CRC Press.
- Stiegler, B. (1998). *Technics and time: The fault of Epimetheus* (Vol. 1). Stanford University Press.
- Stiegler, B. (2012). Relational ecology and the digital pharmakon. *Culture Machine*, 13.
- Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- van den Berg, B., & Leenes, R. E. (2013). Abort, retry, fail: Scoping techno-regulation and other techno-effects. In: *Human law and computer law: Comparative perspectives*, Springer, pp. 67–87.
- Van den Hoven, J. (2008). *Moral methodology and information technology. The handbook of information and computer ethics*, p. 49.
- van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). Design for values: An introduction. In: *Handbook of ethics, values, and technological design: Sources, theory, values and application domains* pp. 1–7.
- Van Dijk, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.
- Veale, M., & Edwards, L. (2018). Clarity, surprises, and further questions in the article 29 working party draft guidance on automated decision-making and profiling. *Computer Law & Security Review*, 34(2), 398–404.
- Verbeek, P. P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. Penn State Press.
- Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.
- Wihlborg, E., Larsson, H., & Hedström, K. (2016). “The computer says no!”—A case study on automated decision-making in public authorities. In: *2016 49th Hawaii International Conference on System Sciences (HICSS)*, IEEE, pp. 2903–2912.

- Winkler, T., & Spiekermann, S. (2018). *Twenty years of value sensitive design: A review of methodological practices in VSD projects*. *Ethics and Information Technology*. pp. 1–5.
- Wittkower, D. (2014). Facebook and dramauthentic identity: A post-goffmanian theory of identity performance on SNS. *First Monday*, 19(4).



Data Science for Entrepreneurship: The Road Ahead

*Willem-Jan van den Heuvel,
Werner Liebregts and Arjan van den Born*

Contents

22.1 Introduction – 522

22.2 The Road Ahead – 522

22.2.1 AI Software – 523

22.2.2 MLOps – 524

22.2.3 Edge Computing – 525

22.2.4 Digital Twins – 525

22.2.5 Large-Scale Experimentation – 526

22.2.6 Big Data and AI Opportunities – 527

22.2.7 Government Regulation – 527

References – 530

Learning Objectives

After having read this chapter, you will be able to:

- Identify and explain a few important ongoing developments that will (continue to) affect data entrepreneurship and, hence, data entrepreneurship research
- Outline how the field of data entrepreneurship practice will most likely evolve in the years to come
- Pinpoint a number of promising avenues for future research at the intersection of the data science and entrepreneurship disciplines

22.1 Introduction

This book has treated two domains that have—until recently—been treated in sheer isolation, viz. data science and entrepreneurship. As we have explored in this book, both disciplines find each other in the emerging discipline of data-intensive, data-driven, or data science entrepreneurship, also often shortly referred to as *data entrepreneurship*. Data entrepreneurship requires (at least basic) knowledge of the domains of data engineering and data analytics, and, in turn, of what we have coined data and society (i.e., the business and societal context, such as prevailing laws and generally accepted views on ethical behavior concerning data). Hence, the four sections of this book each covered multiple relevant topics in their respective domains.

Now that we have obtained a deeper understanding of the state-of-the-art knowledge in all these domains, it is about time for a glimpse into the (nearby) future. We see various important developments on the rise that will—sooner or later—influence data entrepreneurship, providing tantalizing new ways to generate more business value by exploiting the opportunities that data science brings: in (very) short, data science for entrepreneurship. The question central to this chapter further explains the subtitle of this book: How can entrepreneurs leverage big data and AI for new value creation? These developments also open up entirely new avenues for future research at the intersection of data science and entrepreneurship. Therefore, we also briefly discuss their implications for research by entrepreneurship scholars.

In this chapter, we first give an overview of what we believe are the most important developments, briefly discussing their implications and ramifications from an entrepreneurial and scientific perspective. Afterwards, we conclude this particular chapter and the book as a whole.

22.2 The Road Ahead

This section unfolds an itinerary of opportunities and challenges, of which we believe they will heavily influence the nascent data science and entrepreneurship discourse. Some of these developments are related to new technologies that enable

individuals and firms to develop new propositions and product-market combinations (product and service innovations), and other technology developments enable individuals and firms to pursue more effective or more efficient delivery (process innovation). Such firms can be both new and already existing. Five to ten years ago, when the interdisciplinary field of data entrepreneurship started to emerge, said technologies were nonexistent or mere blimps on the agenda. Currently, we see that these technologies have become more mature and that they play an increasing role for firms pursuing competitive advantage.

The most prominent technological development within the context of the so-called Fourth Industrial Revolution (or Industry 4.0) concerns artificial intelligence (AI). AI undeniably has the characteristics of a transformational general-purpose technology (Brynjolfsson & McAfee, 2014; Cockburn et al., 2018). And thus, as Chalmers et al. (2020) put it: “AI ... [has] profound implications for how entrepreneurs develop, design and scale their organizations” (p. 15). On top of these technological developments, we also see important social and economic changes in our landscape. Think of macro trends like globalization and internationalization.

As businesses have become more mature in terms of data adoption and usage, and have been developing new algorithms, and data-driven products and services, the competitive landscape is also becoming denser. The era of early discovery and exploration, and, hence, still limited competition, is rapidly coming to an end. Most firms and industries have been transforming their services into digital services. In this process, the low-hanging fruit solutions have already been developed and adopted, with only a limited number of markets as notable exceptions. Digital services based on algorithms have also become more of a commodity.

Finally, discussions about the regulation of platforms and/or algorithms were almost nonexistent, whereas nowadays regulation of platforms (Newman, 2019) and algorithms (Parikh et al., 2019) are both important fields of research. Moreover, governments are increasingly adopting new, often stricter rules to regulate digital markets and services (e.g., European Commission, 2020; U.S. House of Representatives, 2020).

In the remainder of this section, we will highlight a few recent advancements and trends in data science (in no particular order), which will—we believe—strongly impact data entrepreneurship in the years to come, and thus data entrepreneurship research as well. This is not meant to be an exhaustive list, but it does highlight a few major developments with substantial expected impact.

22.2.1 AI Software

After a few decades of growing up as a scientific and practical discipline, AI is now quickly maturing with a myriad of applications in business and society. Leveraged by the potential of AI, a new breed of software applications has emerged, often referred to as AI software. Indeed, AI software no longer limits itself to rather experimental, non-scalable “toy” applications devoid of any business value.

Here, the best of two worlds that used to operate separately until pretty recently have been bridged, namely that of AI and software engineering. AI has brought significant techniques and tools for exploring optimum solutions in highly unstructured, complex, fuzzy, unpredictable, and/or incomplete cases. Software engineering, on the other hand, has proven its value in factoring well-understood, relatively stable, and clearly demarcated solution spaces into code (Ford, 1987).

This latest AI trend has implied to the software engineering community to increasingly infuse AI technologies and platforms (such as Google AI platform, TensorFlow, IBM's Watson Studio, and Microsoft's Azure) and develop a new series of software engineering models and practices that foster automatic code generation, continuous testing and integration, and software design. Thus, exciting new business and research opportunities are to be found in this emerging area of AI software.

22.2.2 MLOps

Another important development we observe is turning AI and machine learning into an engineering discipline and improving the collaboration and coordination between data engineering professionals (including programmers and software maintenance staff), data scientists (including machine learning experts), and domain experts. This development is nicely reflected in the uptake of a new generation of disciplined, repeatable, and transparent machine learning operations (MLOps). MLOps comes with automated techniques for implementing the machine learning pipelines with software development, and a culture that advocates modeling teams that closely work together.

MLOps has been largely inspired by the DevOps philosophy (Ebert et al., 2016) and associated practices that streamline and tightly integrate the software development workflow and delivery processes. Like DevOps, MLOps adopts the continuous integration and continuous testing cycle to produce and deploy production-ready new micro-releases and versions of intelligent enterprise applications.

This implies a culture shift between data engineers, data analysts, deployment and system engineers, and domain experts, with improved dependency management—and thus transparency—between model development, training, validation, and deployment. As such, MLOps clearly requires sophisticated policies based on performance metrics and telemetry, such as F^1 , accuracy scores, and software quality (Nogueira et al., 2018).

With the exact boundaries between MLOps and DevOps being blurry, a seminal application scenario of MLOps is to be found in Amazon's Web Services offering, which supports an integrated ML workflow for building, testing, and integrating, supporting continuous delivery with source control and monitoring services.

22.2.3 Edge Computing

Edge computing is a new computing paradigm that allows for highly distributed processing and analysis of huge volumes of data at the edges of the network, and closest to the locus needed. In this way, processing is moved from the cloud to the edges of the network, furnishing highly decentralized processing, storage, and analytics. This illustrates that edge computing embraces a model of distributed instead of centralized computing, as is also the case with conventional cloud computing models (Khan et al., 2019).

Potential advantages include lower latency freeing up bandwidth, less reliance on the network, and proximity to the user, at the cost of decreased dependability with less processing capacity provided by edge devices (Bagchi et al., 2019). This requires scalable and robust security mechanisms to be distributed over the edge devices.

Edge computing systems are typically owned by different service providers and may operate under the provision of various business models. Every business runs according to different business strategies and management policies, while following different rules and regulations according to the organization of its operation (Khan et al., 2019). Similarly, edge devices are developed by different vendors and have their own interfaces, which affects the performance and entails high costs. In order to overcome aforementioned issues, a joint management and deployment business model is critical to ensure high performance and offer low-cost services to end users.

22.2.4 Digital Twins

While NASA has first practiced with so-called digital twin concepts since the 1960s to replicate and analyze, for example, the living conditions in spaceships like Apollo 13, the term itself was introduced by Michael Grieves in 2002 in the context of pitching a new product life cycle management institute (Grieves, 2005). The rapid uptake of digital twin technology has been fueled by the coming of age of various enabling technologies, including machine learning, data fusion, data communication, Internet of Things (IoT), augmented reality, virtual reality, and big data analytics.

In essence, digital twins may be defined as digital replicas of (non-)living physical objects (Shafto et al., 2012). They essentially go far beyond existing digital representations, such as CAD models, while supporting cyber-physical systems during the entire life cycle, that is, from their design to production and to actual execution and management.

As such, digital twins exploit AI and ML to visualize real-time, operational data gained from physical and virtual objects instrumented with IoT devices and, hence, to augment human decision-making. This makes information about objects

(e.g., buildings and production lines) or concepts (e.g., production planning) readily available for smoother and more intuitive communication between authorized stakeholders. Here, one can think of historical data, status reports, and contextual (meta-)data like weather reports. By adding AI-driven capabilities, digital twins can even simulate and reason about various situations and run for example “what-if” scenarios, leveraging diagnostic, predictive, and optimization capabilities.

22.2.5 Large-Scale Experimentation

A frequent critique on data science is that it focuses on correlations and associations instead of causal relationships and counterfactuals. Here, the argument goes along the lines that most correlations are by definition spurious, even with large databases (Calude & Longo, 2017). While this critique is valid to some extent, it misses the point that analysis of big datasets does not necessarily imply a focus on correlations. Big datasets can also be helpful in the discovery of causality and the elaboration of counterfactuals.

Today, most digital companies, including Airbnb, Amazon, ► [Booking.com](#), eBay, Facebook, Google, LinkedIn, Microsoft, Netflix, Twitter, and Uber, run online randomized controlled experiments at a (very) large scale (Kohavi et al., 2020). This enables them to use (big) data to find underlying causal factors. Typically, the larger companies run hundreds to thousands of such controlled experiments each day, sometimes on millions of users. Where so-called randomized controlled trials (RCTs) in medicine are often criticized for being expensive and complex, in digital environments, the marginal cost of such experiments is very low, and the added value of uncovering causal relations is not to be underestimated. If done right, the adoption of large-scale experimentation directly leads to (incremental) innovation and increased revenue. As a result of the year-on-year growth of the number of such experiments at large, digital companies have doubled, tripled, or even quadrupled.

Fortunately, while large firms may have an advantage in terms of the size and the costs of experiments (due to economies of scale), RCTs as such can also be of value for small and medium-sized enterprises (SMEs). Increasingly, the infrastructure to run such large-scale experiments have become available for all sorts of firms (Fabijan et al., 2018; Tang et al., 2010). For instance, Google Optimize is such an online split-testing tool that plugs into websites, thereby enabling SMEs to experiment with different ways of delivering content. Another example is IBM’s experimentation platform aimed at AI operations (Rausch et al., 2020). These developments will likely continue given the potential value of discovering causal relations, the continued rise in data (e.g., due to the rise of the IoT, also see Attaran (2017)), and the ever-decreasing costs of conducting experiments. The latter is partly due to the advent of experimentation platforms in different industries and domains.

22.2.6 Big Data and AI Opportunities

Five to ten years ago, the concept of big data was just starting to become mature (Provost & Fawcett, 2013), and AI solutions were predominantly used by the Big Technology firms with their access to heaps of data, huge amounts of funding, and their talented staff. In this period, all corporates as well as many relatively large firms among the group of SMEs have already been experimenting with data science. Typically, these firms started with setting up a project team, or even a data lab with professionals to see what data science could offer to them. Sometimes, these companies used external consultancy firms to help them develop their data capabilities. Without any exception, these companies discovered that data science is by no means easy. There are many cultural, technological, managerial, and organizational barriers to overcome.

However, the companies that persevered often found value in data. Insights obtained from data at least proved to be beneficial for internal decision-making. Other companies were even able to develop new data-based (digital) products and services. While exploring and developing digital products and services is hard, and monetizing data is often even harder (Bataneh et al., 2020; Wixom & Ross, 2017), in case these companies were able to overcome all barriers, they found ample opportunities with relatively limited competition (Zuboff, 2019). In these settings, companies were often able to exploit their first-mover advantage (Varadarajan et al., 2008).

Later on, many more companies have been able to overcome the initial barriers and have built their data science capabilities accordingly (Davenport & Ronanki, 2018; Fountaine et al., 2019). Therefore, today's chances of building a competitive advantage around a unique data science capability are slim. This probably requires access to unique and protected datasets, the use of state-of-the-art AI technologies, and/or the commitment of exceptionally talented workers. Firms with so-called big data analytics (BDA) capabilities appear to perform better overall, but effect sizes heavily depend on the firm's entrepreneurial orientation, the industry in which it operates, and the environmental dynamism (Dubey et al., 2020; Müller et al., 2018; Wamba et al., 2017). In any case, the application of data science by combining mundane and ubiquitous datasets and using standard statistics is not sufficient anymore. This is nowadays the ticket to the game, but it will no longer give firms a competitive edge.

22.2.7 Government Regulation

As mentioned above, the wide-open competitive plains of the first decades of this millennium are over. This not only applies to the level of competition, but it also applies to the presence of government agencies. History shows that government intervention and regulation are always lagging when new technology is being intro-

duced (Wiener, 2004). Data science and AI are by no means an exception to this rule. One can even say that the complexity and novelty of AI in combination with the opaqueness of the social impact of digital services have caused governments worldwide to adopt a wait-and-see approach. However, these days now seem to be over. The critique on digital technology and digital companies is increasing everywhere around the globe: from China to Europe and from Africa to the United States.

This critique is formed along multiple lines. First and foremost, there is criticism on the winner-takes-all characteristics of digital markets and the behavior of Big Technology giants, such as Apple, Google, Facebook, and Microsoft, as they use (abuse?) their market power to restrict competition. As said, market regulators are increasingly developing and adopting novel antitrust legislation. At the time of writing, a list of these new laws was still being debated, yet slowly but surely governments are targeting the market power of Big Tech. In Western democracies, governments worry about competition and consumer welfare, whereas in more authoritarian states, governments worry about the power of large technology firms, *vis-à-vis* the state.

Yet, market power is by no means the only reason why governments are willing to regulate data-intensive firms. Privacy has been an important argument for some years leading to the introduction of the GDPR framework in the European Union. In the United States, the new presidency is expected to come up with new policies on aspects such as federal privacy law, international data transfers, and net neutrality. Lately, more ethical and social issues, such as discrimination in algorithms and emerging filter bubbles in society, have become major points of interest. All in all, the days of unfettered access to data and unrestricted and unsupervised provision of digital services are over. Global distrust in Big Tech is skyrocketing, and governments will come with ever-increasing demands, laws, frameworks, and government bodies to supervise the digital marketplace. To compete in this market, it is crucial to be aware of these emerging rules and to build trusted, long-term relationships with governmental organizations at all levels.

Conclusion

Data entrepreneurship is here to stay. With a perceived annual growth of data of over 40% per annum, the importance and ubiquitousness of data will only increase in the years to come. IoT data, sensor data, genomic data, personal health data, audio data, video data, and many more (new) types of data will continue to boom. Data entrepreneurship, briefly defined as the exploration and exploitation of opportunities using data science, will therefore undoubtedly grow in importance as well.

However, the days of exploiting the low-hanging fruit seem to be over. Spotting and monetizing new opportunities with data science progressively require access to unique datasets, use of novel technologies, and/or involvement of exceptionally talented individuals. Today, simply mining a dataset and presenting descriptive statistics as insights are far from sufficient to obtain and (temporarily) maintain a competitive advantage. Luckily for the digital companies at the forefront of data science, and those lagging behind, technologies are still rapidly evolving. In addition, novel technology frameworks, such as AI software and MLOps, allow companies to scale up their AI capabilities. Lastly, (changes in) government-induced rules and

regulations with respect to data science adoption and usage will be a burden to some firms, but at the same time, it may provide new opportunities for those companies who can easily adapt to and perhaps even shape such (new) rules.

Obviously, all these developments with severe practical implications for data entrepreneurs and data-driven business developers (or data intrapreneurs) will also affect the field of data entrepreneurship research. This field of research is still in its infancy. A few topics have already received somewhat more attention though, albeit very recently. These topics include the opportunities that digital technologies like AI bring for entrepreneurship (e.g., Ransbotham et al., 2017; Townsend & Hunt, 2019; Von Briel et al., 2018; Von Krogh, 2018), the impact of AI on managers and their decision-making (e.g., Huang et al., 2019; Raisch & Krakowski, 2021; Shrestha et al., 2019), and the relationship between BDA capabilities and firm performance (e.g., Dubey et al., 2020; Müller et al., 2018; Wamba et al., 2017).

Nevertheless, in general, we still lack an in-depth understanding of how, when, and why entrepreneurs and intrapreneurs use data science to create new value (or not). Specific research topics that require further attention include (1) entrepreneurial and managerial awareness and competences with respect to new technological developments, (2) barriers to adoption and usage of data science among firms of all age and size categories, (3) determinants and consequences of firms having different levels of the so-called data maturity, (4) pros and cons of (open) data sharing for entrepreneurship and innovation, and (5) impact of new rules and regulation on data exploration and exploitation by firms. Other relevant examples have been discussed in detail by Chalmers et al. (2020).

We hereby call for thorough theorizing and extensive empirical scrutiny on any of the aforementioned topics. By definition, scholars then have to engage in multidisciplinary research building bridges between the disciplines of data science and entrepreneurship. A new era has begun (Obschonka & Audretsch, 2020), and let us contribute to an entrepreneurial society, in which big data and AI are being leveraged by entrepreneurs in the most productive ways.

Discussion Points

1. In this chapter, we have highlighted various recent advancements and trends that—according to us—will (continue to) impact data entrepreneurship. Which one of them will most likely have the strongest impact, do you think, and why?
2. It has been suggested that the enforcement of new rules and regulations (for example, to reduce the market power of Big Tech companies) can also provide new entrepreneurial opportunities to some. Name and explain at least one example of such an opportunity that might arise as a consequence of newly adopted antitrust legislation.
3. One of the promising avenues for data entrepreneurship research concerns the barriers to the adoption and usage of data science as perceived by firms. List all such barriers that you can think of, and discuss how each of these barriers could be alleviated or even taken away completely.

Take-Home Messages

- Since its inception 5–10 years ago, data entrepreneurship has proliferated very quickly and continues to gain importance in virtually all elements of our daily lives and society at large.
- New opportunities and challenges are to be found in emerging technologies, such as AI software, MLOps, edge computing, and digital twins.
- With data science quickly becoming more commonplace, simply applying it will not give firms a competitive edge (anymore), but one needs unique datasets, cutting-edge technologies, and/or exceptional talents instead.
- New rules and regulations are needed to better deal with the ever-increasing power of large tech companies, and with issues like data privacy and explainability of AI.
- Data entrepreneurship research is still in its infancy, so much more research is needed to better understand how entrepreneurs can leverage big data and AI for new value creation.

References

- Attaran, M. (2017). The internet of things: Limitless opportunities for business and society. *Journal of Strategic Innovation and Sustainability*, 12(1), 10–29.
- Bagchi, S., Siddiqui, M. B., Wood, P., & Zhang, H. (2019). Dependability in edge computing. *Communications of the ACM*, 63(1), 58–66.
- Bataineh, A. S., Mizouni, R., Bentahar, J., & El Barachi, M. (2020). Toward monetizing personal data: A two-sided market analysis. *Future Generation Computer Systems*, 111, 435–459.
- Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.
- Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of Science*, 22(3), 595–612. <https://doi.org/10.1007/s10699-016-9489-4>
- Chalmers, D., MacKenzie, N. G., & Carter, S. (2020). Artificial intelligence and entrepreneurship: Implications for venture creation in the fourth industrial revolution. *Entrepreneurship Theory and Practice*, 45, 1–26. <https://doi.org/10.1177/1042258720934581>
- Cockburn, I. M., Henderson, R., & Stern, S. (2018). *The impact of artificial intelligence on innovation*. National Bureau of Economic Research.
- Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, 96(1), 108–116.
- Dubey, R., Gunasekeran, A., Childe, S. J., Bryde, D. J., Giannakis, M., Foropon, C., Roubaud, D., & Hazen, B. T. (2020). Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organizations. *International Journal of Production Economics*, 226, 107599.
- Ebert, C., Gallardo, G., Hernantes, J., & Serrano, N. (2016). DevOps. *IEEE Software*, 33(3), 94–100.
- European Commission. (2020). *Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC. COM(2020) 825 final*. Brussels: European Commission.
- Fabijan, A., Dmitriev, P., McFarland, C., Vermeer, L., Holmström Olsson, H., & Bosch, J. (2018). Experimentation growth: Evolving trustworthy A/B testing capabilities in online software companies. *Journal of Software: Evolution and Process*, 30(12), e2113.
- Ford, L. (1987). Artificial intelligence and software engineering: A tutorial introduction to their relationship. *Artificial Intelligence Review*, 1, 255–273.

- Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-powered organization. *Harvard Business Review*, 97(4), 62–73.
- Grieves, M. W. (2005). Product lifecycle management: The new paradigm for enterprises. *International Journal of Product Development*, 2(1–2), 71–84.
- Huang, M., Rust, R., & Maksimovic, V. (2019). The feeling economy: Managing in the next generation of artificial intelligence (AI). *California Management Review*, 61(4), 43–65.
- Khan, W. Z., Ahmed, E., Hakak, S., Yaqoob, I., & Ahmed, A. (2019). Edge computing: A survey. *Future Generation Computer Systems*, 97, 219–235.
- Kohavi, R., Tang, D., Xu, Y., Hemkens, L. G., & Ioannidis, J. P. (2020). Online randomized controlled experiments at scale: Lessons and extensions to medicine. *Trials*, 21(1), 1–9.
- Müller, O., Fay, M., & Vom Brocke, J. (2018). The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics. *Journal of Management Information Systems*, 35(2), 488–509.
- Newman, J. M. (2019). Antitrust in digital markets. *Vanderbilt Law Review*, 72(5), 1497–1561.
- Nogueira, A.F., Ribeiro, J.C., Zenha-Rela, M.A., & Craske, A. (2018). Improving La Redoute's CI/CD pipeline and DevOps processes by applying machine learning techniques. In: *Proceedings of the 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*, pp. 282–286.
- Obschonka, M., & Audretsch, D. B. (2020). Artificial intelligence and big data in entrepreneurship: A new era has begun. *Small Business Economics*, 55, 529–539.
- Parikh, R. B., Obermeyer, Z., & Navathe, A. S. (2019). Regulation of predictive analytics in medicine. *Science*, 363(6429), 810–812.
- Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1), 51–59.
- Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*, 46(1), 192–210.
- Ransbotham, S., Kiron, D., Gerbert, P., & Reeves, M. (2017). Reshaping business with artificial intelligence: Closing the gap between ambition and action. *MIT Sloan Management Review*, 59(1).
- Rausch, T., Hummer, W., & Muthusamy, V. (2020). An experimentation and analytics framework for large-scale {AI} operations platforms. In: *2020 {USENIX} Conference on Operational Machine Learning (OpML20)*.
- Shafto, M., Conroy, M., Doyle, R., Glaessgen, E., Kemp, C., LeMoigne, J., & Wang, L. (2012). Modeling, simulation, information technology and processing roadmap. *National Aeronautics and Space Administration*, 32, 1–38.
- Shrestha, Y. R., Ben-Menahem, S. M., & Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), 66–83.
- Tang, D., Agarwal, A., O'Brien, D., & Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 17–26.
- Townsend, D. M., & Hunt, R. A. (2019). Entrepreneurial action, creativity, and judgment in the age of artificial intelligence. *Journal of Business Venturing Insights*, 11, e00126.
- U.S. House of Representatives. (2020). *Investigation of competition in digital markets: Majority staff reports and recommendations*. U.S. House of Representatives.
- Varadarajan, R., Yadav, M. S., & Shankar, V. (2008). First-mover advantage in an internet-enabled market environment: Conceptual framework and propositions. *Journal of the Academy of Marketing Science*, 36(3), 293–308.
- Von Briel, F., Davidsson, P., & Recker, J. (2018). Digital technologies as external enablers of new venture creation in the IT hardware sector. *Entrepreneurship Theory and Practice*, 42(1), 47–69.
- Von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries*, 4(4), 404–409.
- Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J. F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, 70, 356–365.

- Wiener, J. B. (2004). The regulation of technology, and the technology of regulation. *Technology in Society*, 26(2–3), 483–500.
- Wixom, B. H., & Ross, J. W. (2017). How to monetize your data. *MIT Sloan Management Review*, 58(3), 10–13.
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.