Future of Business and Finance

Patrick Glauner
Philipp Plugmann *Editors*

Innovative Technologies for Market Leadership

Investing in the Future

Springer

Future of Business and Finance

The Future of Business and Finance book series features professional works aimed at defining, describing and charting the future trends in these fields. The focus is mainly on strategic directions, technological advances, challenges and solutions which may affect the way we do business tomorrow, including the future of sustainability and governance practices. Mainly written by practitioners, consultants and academic thinkers, the books are intended to spark and inform further discussions and developments.

More information about this series at http://www.springer.com/series/16360

Patrick Glauner • Philipp Plugmann
Editors

Innovative Technologies for Market Leadership

Investing in the Future

Springer

Editors
Patrick Glauner
Applied Computer Science
Deggendorf Institute of Technology
Deggendorf, Germany

Philipp Plugmann
Interdisciplinary Periodontology
and Prevention
SRH University of Applied Health Sciences
Leverkusen, Germany

ISSN 2662-2467 ISSN 2662-2475 (electronic)
Future of Business and Finance
ISBN 978-3-030-41308-8 ISBN 978-3-030-41309-5 (eBook)
https://doi.org/10.1007/978-3-030-41309-5

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

Foreword

Digital technologies are dramatically changing the way we do business, the way we work, and increasingly the way we live. The speed of change, reflected in "Moore's Law" of exponential growth in computing power, is influencing all aspects of society, and not just in the developed world but globally. Sometimes the speed of change can be overwhelming in its breadth and scope, and therefore it is important to understand in detail how such technologies will specifically affect different industries and parts of society. This excellent book edited by Professors Patrick Glauner and Philipp Plugmann provides important guidance for understanding the impact of digital technologies now and in the future.

Some of the contributions focus on particular sectors such as **energy** (contribution by Sharafi), **health care** (contribution by Plugmann, contribution by Lerzynski, contribution by Trestioreanu et al., contribution by Ehsani), **transportation** (contribution by Franke), **plastics** (contribution by Krause, contribution by Stillings), and **construction** (contribution by Jacob). Other contributions focus on the impacts of digital technologies on infrastructures that will affect many industries such as **quantum technologies** (contribution by Akenine), **artificial intelligence** (contribution by Glauner, contribution by Falk), and **ubiquitous computing** (contribution by Panné). Another group of contributions focus on how digital technologies will affect **innovation processes** (contribution by Bludau, contribution by Ohlberg and Salmeron, contribution by Denkel), **mechanical engineering** (contribution by Thurner and Glauner), **data engineering** (contribution by Lu), and **autonomous driving** (contribution by Mund and Glauner).

The authors are very well qualified for addressing these important topics. For example, the health care contributions include a medical doctor, hospital CEO, and university professors. Other authors are relevant experts from industry and academia.

Taken together, these contributions provide an excellent “road map” to guide academics, industry representatives, and other interested readers to understand the large impact of digital technology today and its enormous potential for future development.

Chair of Innovation and Technology Management
University of Regensburg
Regensburg, Germany

Münchner Kreis
Munich, Germany

Michael Dowling

Preface

Our lives have entirely changed within the last 250 years as our society has developed from agricultural to high tech. Even within just the last 20 years, advances in such fields as artificial intelligence, systems biology, or surgery have revolutionized our society. For example, we now take for granted voice assistants and broadband connections on smartphones or have a relatively high chance to survive most types of cancer. Most of these advances are driven by novel and innovative technologies. The innovative technologies of the future will further change the world as we know it. The question is what new technologies to commit to in order to become the market leader of tomorrow.

This innovative book reflects these recent developments while providing comprehensive outlooks on which technologies will be of importance in the future. It provides an unparalleled mix of expertise of respected international authors from academia and the industrial world. The authors present their works on and expertise in innovative technologies spanning from the fields of biology and medicine through quantum technologies and augmented reality to digitalization in mechanical engineering and smart power grids. This book is aimed at investors, decision makers, entrepreneurs, researchers, and students.

Each chapter is self-contained and provides the necessary respective prerequisites. Some chapters are more business-oriented while others are more technical in order to address a diverse audience. In their chapters, the authors also make concrete recommendations on how to invest in their fields and demonstrate the potential of their technologies to create economic value in real-world applications.

This book would not have been possible without Mr. Philipp Baun, our commissioning editor. We would like to thank him and all the other Springer staff, in particular Ms. Irene Barrios-Kezic, involved for their professionalism, tireless ability to read multiple drafts, and help in improving the book.

Patrick would like to thank his wife Shengqin for her support and her endless patience and boundless wisdom. Philipp would like to thank his wife Julia who has been hugely supportive throughout the several months it took to write this book.

Regensburg, Germany — Patrick Glauner
Leverkusen, Germany — Philipp Plugmann
February 2020

About the Book

This book introduces the reader to the latest innovations in fields such as artificial intelligence, systems biology, or surgery and gives advice on what new technologies to consider for becoming a market leader of tomorrow. Companies generally acquire information on these fields from various sources such as market reports, scientific literature, or conference events, but find it difficult to distinguish between mere hype and truly valuable innovations. This book offers essential guidance in the form of structured and authoritative contributions by experts in innovative technologies spanning from biology and medicine to augmented reality and smart power grids. The authors identify high-potential fields and demonstrate the impact of their technologies to create economic value in real-world applications. They also offer business leaders advice on whether and how to implement these new technologies and innovations in their companies or businesses.

Contents

Smart Grid, Future Innovation and Investment Opportunities 1
Dean Sharafi

Quantum Technologies 11
Daniel Akenine

Security in Intelligent Transportation Telematics 21
Erich H. Franke

Innovation and Future Technology Scenarios in Health Care: Ideas and Studies 31
Philipp Plugmann

Unlocking the Power of Artificial Intelligence for Your Business 45
Patrick Glauner

Innovation Means: Asking the Right Questions 61
Oliver Bludau

Innovative Technologies in the Ageing Population: Breaking the Boundaries 75
Guido Lerzynski

Using Augmented Reality and Machine Learning in Radiology 89
Lucian Trestioreanu, Patrick Glauner, Jorge Augusto Meira, Max Gindt, and Radu State

Digitalization in Mechanical Engineering 107
Michael Thurner and Patrick Glauner

Lean Launch Data Engineering Projects with Super Type Power 119
Kenny Zhuo Ming Lu

Ubiquitous Computing: From 5G to the Edge and Beyond 133
André Panné

Autonomous Driving on the Thin Trail of Great Opportunities and Dangerous Trust 153
Sandro Mund and Patrick Glauner

Analytic Philosophy for Biomedical Research: The Imperative of Applying Yesterday's Timeless Messages to Today's Impasses 167
Sepehr Ehsani

Proposal-Based Innovation: A New Approach to Opening Up the Innovation Process 201
Karl H. Ohlberg and Jose L. Salmeron

Technologies and Innovations for the Plastics Industry: Polymer 2030 233
Michael Krause

How Do Innovative Business Concepts Enable Investment Opportunities in the Complete Construction Value Chain? 245
Christoph Jacob

Motivation, Employees, and Communication in the Start-Up Phase 265
Achim Denkel

AI to Solve the Data Deluge: AI-Based Data Compression 271
Eric Falk

Digital Transformation in Plastics Industry: From Digitization Toward Virtual Material 287
Christopher Stillings

Correction to: Analytic Philosophy for Biomedical Research: The Imperative of Applying Yesterday's Timeless Messages to Today's Impasses C1
Sepehr Ehsani

Editors and Contributors

About the Editors

Patrick Glauner is the Founder and CEO of skyrocket.ai GmbH, an artificial intelligence consulting firm based in Bavaria, Germany. In parallel, he is Full Professor of Artificial Intelligence at Deggendorf Institute of Technology, a position he is honored to hold since the age of 30. His research on AI was featured in *New Scientist* and cited by McKinsey and others. He is also Area Editor of the *International Journal of Computational Intelligence Systems* (IJCIS). Previously, he held managerial positions at the European Organization for Nuclear Research (CERN), at Krones Group, and at Alexander Thamm GmbH. He studied at Imperial College London and also holds an MBA. He is an alumnus of the German National Academic Foundation (Studienstiftung des deutschen Volkes).

Philipp Plugmann has been doing multidisciplinary work for the last 20 years in parallel to practicing as a dentist in his own clinic in Leverkusen, Germany. He is also Full Professor for Interdisciplinary Periodontology and Prevention at SRH University of Applied Health Sciences. His first book on innovation in medical technology published in 2011 was reviewed by Cisco. His second book on innovation published with Springer in 2018 got more than 50,000 chapter downloads in its first 15 months. Previously, he held multiple adjunct faculty appointments for more than 12 years and has won multiple teaching awards. He also holds an MBA, an MSc in Business Innovation, and an MSc in Periodontology and Implant Therapy (DGParo) and is currently pursuing his third doctorate. Plugmann has given research talks in the field of innovation at conferences at Harvard Business School, Berkeley Haas School of Business, Max Planck Institute for Innovation and Competition, and Nanyang Tech University, Singapore. Plugmann is a serial entrepreneur and advisor to several companies, including a global technology consultancy—DataArt.

Contributors

Daniel Akenine IASA, Stockholm, Sweden

Oliver Bludau Innovators Institute, Cologne, Germany

Achim Denkel CAPinside, Hamburg, Germany

Sepehr Ehsani Department of Philosophy, University College London, London, UK
Ronin Institute for Independent Scholarship, Montclair, NJ, USA

Erik Falk NIUGroup SARLS, Luxembourg, Luxembourg

Erich H. Franke AFUSOFT Kommunikationstechnik GmbH, Königsbach-, Stein, Germany

Max Gindt Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg, Luxembourg

Patrick Glauner Deggendorf Institute of Technology, Deggendorf, Germany

Christoph Jacob CASEA, AG, Neu-Isenburg, Germany

Michael Krause KIMW-Qualifizierungs gGmbH and KIMW-Forschungs gGmbH, Lüdenscheid, Germany

Guido Lerzynski St. Marien-Hospital, Cologne, Germany

Jorge Augusto Meira Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg, Luxembourg

Kenny Zhuo Ming Lu School of Information Technology, Singapore, Singapore

Sandro Mund Trier University of Applied Sciences, Trier, Germany

Karl H. Ohlberg EmpraGlob GmbH, Dusseldorf, Germany

André Panné TRADUM, Bonn, Germany

Philipp Plugmann SRH University of Applied Health Sciences, Leverkusen, Germany

Jose L. Salmeron Universidad Pablo de Olavide, Seville, Spain

Dean Sharafi Australian Energy Market Operator (AEMO), Perth, WA, Australia

Radu State Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg, Luxembourg

Christopher Stillings Covestro Polymers (China) Co., Ltd, Shanghai, China

Michael Thurner Regensburg, Germany

Lucian Trestioreanu Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg, Luxembourg

Smart Grid, Future Innovation and Investment Opportunities

Dean Sharafi

Abstract

The electricity industry is going through a massive transformation which is fundamentally changing the way the electrical energy is generated, transmitted, distributed and consumed. This change will bring about challenges and opportunities for the different players in this massive system of supply and demand. The drivers for this transformation are numerous, which include the desire to shift to a more sustainable energy supply, advancement in technology and reduction in cost of renewable energy. Power system operators around the world are dealing with the challenges of operating grids which behave differently compared to originally designed concepts. However, there are vast opportunities for innovation and investments which can benefit from low cost energy. This chapter explores different ideas on how this massive low-cost energy can be harnessed in order to create value.

1 Introduction

The electricity grid is commonly referred to as the largest machine mankind has ever created. It is a machine because it works in harmony in its entirety; and there are common figures and standard values that apply to the whole of the grid as a system of various systems. Since their inception more than a hundred years ago, the electricity grids have been merely passive systems or machines that performed as means of transferring the energy of large generators to the end customers. These passive machines started to change in many ways when power electronics enabled a shift from production of electricity of conventional generators to electronic devices

D. Sharafi (✉)
Australian Energy Market Operator (AEMO), Perth, WA, Australia
e-mail: dean.sharafi@aemo.com.au

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_1

commonly known as inverters. This shift itself was caused by the development of renewable energy and harnessing the power that exist in the wind and the Sun. Since around two decades ago and most importantly since the last decade, this shift has become so prominent that it is now the single most important factor in transformation of the electricity system from the conventional passive grid to an active smart grid. The smart grid of today enables energy to flow bidirectionally from large-scale generators to the consumers, and from the consumers, who are now also producers of energy, or for a better word, prosumers, to other consumers and to the grid itself. These prosumers do so via their rooftop photo-voltaic devices, batteries electric vechicles, smart appliances and other devices which we now call Distributed Energy Resources (DER).

2 Energy Transformation

To manage such a complex system of energy flow, the electricity grids have evolved and gone through a journey of transformation, enabled by advancement in information and communication technology (IT), operational technology (OT) and increasingly importantly data technology (DT). Digitalisation has brought about ample opportunities to innovate the grid and make investments in new business models.

The changes in electricity grids and the shift from production of energy using fossil fuels, to renewable energy resources is generally referred to as energy transformation. This changing energy landscape has transformed many aspects of how we consume energy and the time energy is consumed. It has also transformed, in many ways, the concepts historically used in regulation of energy services as well as their applicable standards. This transformation has created many challenges for the grid operators, as well as many opportunities for innovation and investment. We will discuss these opportunities further in this chapter, but let us see how these changes have affected the energy ecosystem.

Renewable energy resources by nature create a variable supply of energy, because wind does not blow all the time and when it blows its intensity and speed is not constant. The same is true for the Sun, as there are different levels of irradiation in different times of the day, different seasons, and the energy reaching the Earth depends on cloud coverage and other environmental conditions. Currently, grid operators manage the electricity demand levels using a mix of conventional generators which can operate up their maximum capacity levels when required, and the variable renewable generators, which can only operate to a level that their fuel (wind, solar energy, etc.) allows them. The trend in future generation in most advanced countries is a shift from conventional generators to renewable variable energy resources; therefore, in future we will have much more variable generators and just enough fast-moving conventional generators which can compensate for the variability of renewable generators, in order to keep the supply of electricity at the demand level all the time.

Solar and wind energies are in abundance at certain times, and with renewable sources which can supply more than the required demand at these times, there will be an excess in total energy produced during these times. Similarly, there is an energy deficit at times of low energy production. These two effects can create opportunities for innovation and future investment.

3 Smart Grid and Renewable Energy

Smart Grid is a concept enabled with the electricity grid transitioning from a passive low-intelligence asset-intensive grid to a high-intelligence active grid. Smart Grid refers to an electricity system which can intelligently integrate the activities of all players involved in the energy ecosystem including generators, consumers and prosumers in order to deliver a secure, sustainable and economically efficient electricity system using intelligent communication, monitoring and control devices, and innovative products and services. The need for Smart Grid came about when customers started producing energy and taking control of their energy needs. In most countries this was due to customers installing rooftop Photo-Voltaic (PV) panels on their roofs to harness the energy of the Sun, or installing smart meters to distinguish between time-of-use of energy or other smart functionality. For some time, Smart Grid was synonymous with Smart Meters, but Smart Grid is now much wider in concept and functionality than a grid which is just equipped with Smart Meters. Smart Grid is an interactive grid capable of variety of functions which may include, but are not limited to, measurement of energy usage, control of appliances and orchestration of the load for various purposes. Smart Grid is both necessitated by, and most importantly an enabler of, integration of renewable energy. Smart Grid has been made possible by advancement in communication technology, enabling the grid operators to understand load consumption and patterns of energy usage. Smart Grid is now a modernised hybrid energy system integrating the whole components of the grid from transmission down to distribution and home appliances. Smart Grid will be much more interactive and modernised in the future by necessity and due to the rapid growth of the DER.

One important aspect of Smart Grid is decentralisation. In most advanced grids which have enabled customers to take an active part in the energy supply and consumption chain, the electricity system and its critical components have shifted from large central power stations to energy generated by small and distributed resources. While in the past decades, the system planners were mostly concerned about load ensuring adequacy of large generators to meet the peak demand, in present times and most critically in the future, the interaction between load and generators at both ends of the electricity supply chain will be important. Given the changing source of energy production from conventional fuels which were controllable, to the natural environment as a major source of fuel, which is uncontrollable, the future generation will be highly variable.

By 2050 wind and solar will make up around 50% of generation according to forecasts of global generation mix, while renewables collectively will form more

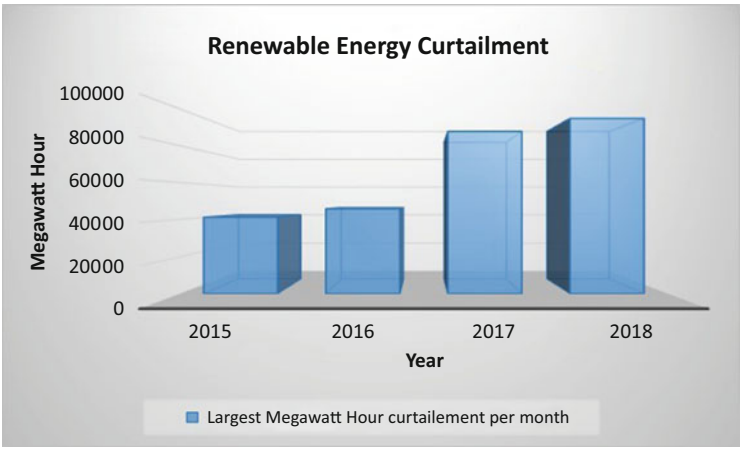

Fig. 1 Curtailment of renewable energy. Data from California Independent System Operator (2019). Source: author

than 60% of the generation fleet. Nuclear energy will fall to the levels we saw in 1980s and coal generation will have a share of even less than 1970s. The dominant share of generation mix will then become renewables and gas generators such as open-cycle gas turbines. Therefore a massive amount of energy production in energy mix of the future is atmosphere dependent and will be subject to significant curtailment, because at times the produced energy will be in excess of need. The energy systems which have a large share of renewables in their current energy mix have already faced this curtailment. For example, California has curtailed increasing amount of energy from variable sources, year after year since 2014. This excess energy translates into negative electricity market prices during the times of abundance and creates opportunities for many energy-intensive technologies which previously were not economically viable. More details are depicted in Fig. 1.

4 Harnessing Variability

An ideal electricity system is one which is secure, reliable, cost effective and environmentally clean. In making an electricity system reliable, the supply (generation) and demand (load) must be in equilibrium all the time. Until about two decades ago, the variable side of this equation was only the load side; the load changed when consumers varied their energy consumption; generators adapted their output to satisfy the demand. With Variable Renewable Energy (VRE), the variability is now on both sides of the equation, namely both the generation and the load sides. In harnessing this variability, there are opportunities on both side of this supply/demand equation that can be used for businesses investment and innovation.

4.1 Harnessing Variability at Distribution Grid Level (Small Energy Level)

Distribution Electricity Markets: Electricity markets are now established in many countries in the world. These electricity markets are wholesale markets which can be characterised as a one-way pipeline model, in which distribution connected producers and consumers have no, or limited access to participate in the localised energy systems. There are other markets known as Essential Reliability Services (or Ancillary Services) which can only be accessible to large generators. The energy produced in the distribution grid is exported into the network, which acts as an infinite storage. The PV owner who produces this energy can only trade the energy with the local utility/retailer. This one-way pipeline model has worked relatively well since the inception of the electricity markets under the traditional structure of the electricity supply system, because the share of DER in the supply–demand equation has been relatively small. With the rapid growth of DER (PVs, electric vehicles, battery storage, etc.) these technologies are increasingly being connected by many businesses and homes who were traditionally passive energy consumers. These energy prosumers are now able to generate, convert and store energy. Aggregation to a large level will enable these small energy producers to become active participants in the future distribution energy and ancillary services markets. The current pipeline model shifts all energy trading to the wholesale markets which were designed based on traditional supply–demand equilibrium using price and quantity as the only metrics for energy trading. This has worked so far because supply has always been dispatchable and the demand has always been assumed inflexible. This situation has changed drastically with VRE, such that supply is variable (not dispatchable) and demand can become flexible to match the variable supply. Current reliability standard requires supply of the load almost 100% of the time, assuming an inflexible load, whereas in reality, there are many types of loads with varying degrees of tolerance to supply reliability. The examples of such loads are pool pumps, electric vehicle charging, water heaters, home energy storage, etc. for which time of service/reliability can be flexible. In order to realise the full potential of the DER and their variable nature, new business models and innovations need to be developed. Energy trading can be made more sophisticated than the traditional wholesale markets through innovative approaches, enabling reliability as a new metric for trading of energy in addition to quantity and price. In such a model, distribution connected generators can trade locally with other distribution customers to maximise the full consumption of their energy resources, as well as full financial benefit. A business model which can facilitate such interactions using digitalisation (cloud computing, machine learning, data science, blockchain technology, etc.) can derive the full value of excess energy of renewables which currently may only be exported into the transmission grid. This excess energy currently is exported to the grid reducing daytime operational demand, creating an undesired phenomenon commonly known as the Duck Curve as depicted in Fig. 2.

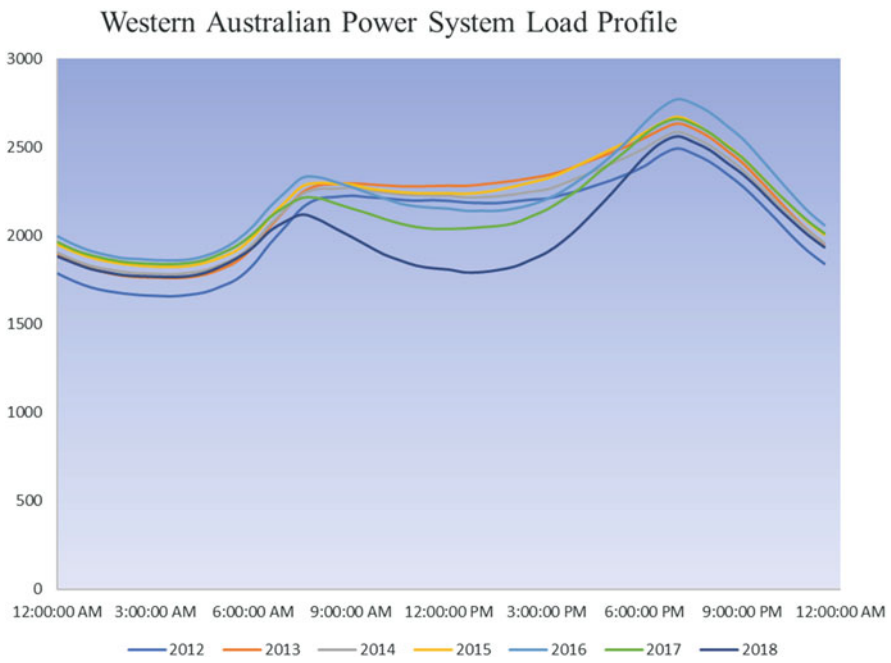

Fig. 2 Sketch of the Western Australian power system duck curve. Reduced operational demand during daytime due to increased penetration of DER. Source: author

During the low operational demand times, the electricity prices are very low or even negative. Any business model able to create flexibility in load, both behind-the-meter and at large-scale levels, in order to match it in real time with the supply, can gain from the low or negative energy prices. This can be done by understanding load patterns and characteristics at the grid and home level using accurate forecasting, Data Science and Machine Learning. The objective will be creating a flexible load through communication and control of home smart appliances, routing energy to certain types of load when required and redirecting the energy generated inside the house to the grid, when this energy can help to stabilise the power system. An example of a load which has a large tolerance to low reliability is a water heater acting as a resistive load and an effective energy storage. Energy routers which can act like internet routers can facilitate this objective. For example, an electric water heater can be fed through an energy router capable of measuring the grid metrics such as frequency and voltage in real time. This router can supply part of the energy generated by DER to the water heater; and redirect the energy from the water heater to the grid in a fraction of a second, acting as a means of firming variable energy. Further innovation can be the transformation of the current technology of home cooling into a technology based on freezing liquids such as water, which can capture a large amount of energy when it turns into ice. This cooling process can work with

variability of DER generation as it has a large tolerance to low energy reliability and is not strictly time critical.

There are also many opportunities to harness the variability of renewable energy at large energy levels. Some of these technologies are mentioned below.

4.2 Hydrogen Production

Hydrogen can be produced by electrolysis of water and splitting it into the atoms that make up water molecules, namely oxygen and hydrogen. This process is very energy intensive and the high cost of this process has so far made it uneconomical. However, with the advent of renewable energy and their zero-emission technology as well as abundance and low cost of production of energy from these sources, hydrogen production has become more viable in the last few years. The performance of hydrogen production and the efficiency of the process has also considerably increased. In Australia, where the proliferation of renewable energy industry has transformed the power system and has opened a new era in clean energy, hydrogen production using zero-emission energy has been given the name of "Liquid Sunshine". This is due to abundance of solar energy in Australia where practical research, trials and technology innovation has vastly emerged around hydrogen production value chain. Conversion of hydrogen as a gas into liquid has also created other opportunities for storage and transport of this high-energy density fuel, using existing gas networks. Moreover, hydrogen as a gas can be directly injected into the domestic gas network and used as a natural domestic fuel for burning and cooking purposes (Fig. 3).

Hydrogen production at a large scale can be a practical way of converting power into gas (P2G) which can then be used in many different ways, including using the gas to produce other types of energy. For example, hydrogen can be used again to generate power for grid stability functions such as ancillary services (P2G2P), or as fuel cells for transport (P2G2T), or heating purposes (P2G2H).

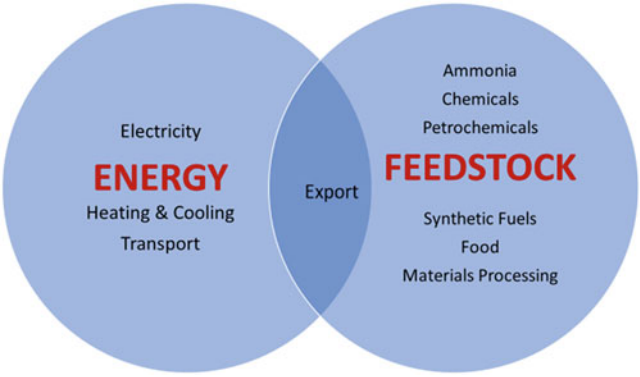

Fig. 3 Various applications of hydrogen. Source: author

Hydrogen can also be used for optimisation of electricity, gas and transport sectors.

Through a relatively simple process hydrogen can be turned into another useful product, ammonia. Ammonia, which consists of one nitrogen atom bonded to three hydrogen atoms, has many applications in the industry. It has historically been used as a fertiliser, but as a fuel its energy density by volume is nearly double that of liquid hydrogen and it is easier to store, transport and distribute. Ammonia can also be converted back into hydrogen and nitrogen.

4.3 Desalination Plants

Water scarcity will be a feature of many economies in the coming years. Population growth, climate change and industrialisation will compound this problem for some countries in the next decade. Some of these countries, such as Middle Eastern or African nations have the potential for clean energy production due to abundance of renewable energy resources. The two major technologies for desalination plants, namely Thermal Desalination and Reverse Osmosis Desalination are both energy intensive and energy price is a major factor in their operating costs. Figure 4 shows the detail and breakdown of the operational costs of desalination plants,

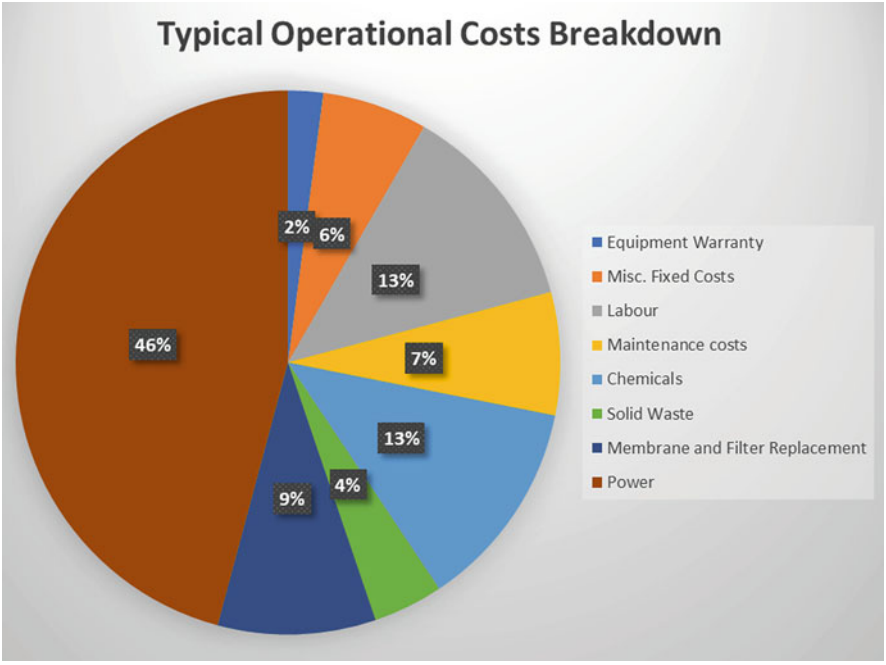

Fig. 4 Price of energy compared with other operational costs of reverse osmosis desalination plants. Data from Advision (2019). Source: author

demonstrating the major share of energy costs in operation of these facilities. The business cases for these plants are now much more attractive due to renewable energy and abundance of energy during certain times. Furthermore, due to advancement of technologies related to this industry, the capital cost of desalination plants has decreased in recent years. In future, the desalination plants will be using free energy to turn unconsumable water into clean, drinkable water, and sometimes they are paid to do so, when the energy prices are negative. Some of the desalination technologies are more reliant on higher reliability power and some have a degree of tolerance to lower reliability energy. The opportunity to use free energy is maximised by development of small-scale desalination plants with low-reliability energy requirements, such that these plants can operate when the price of energy is low and temporarily stop production when the energy is in high demand. A desalination plant capable to switch off instantly when required can also take part in the Ancillary Services markets and provide power system support functions.

4.4 CO_2 Extraction from Nature

Chemical industry has a large carbon footprint. Many chemical substances are produced using energy-intensive processes. It is now possible to extract CO_2 directly from the atmosphere and through electrochemical conversion, turn it into chemical products and fuels. This process is depicted in Fig. 5 and has two benefits, firstly reducing the impact of CO_2 in the nature and secondly closing the carbon loop and turning the waste into useful products. The key to viability of such a process

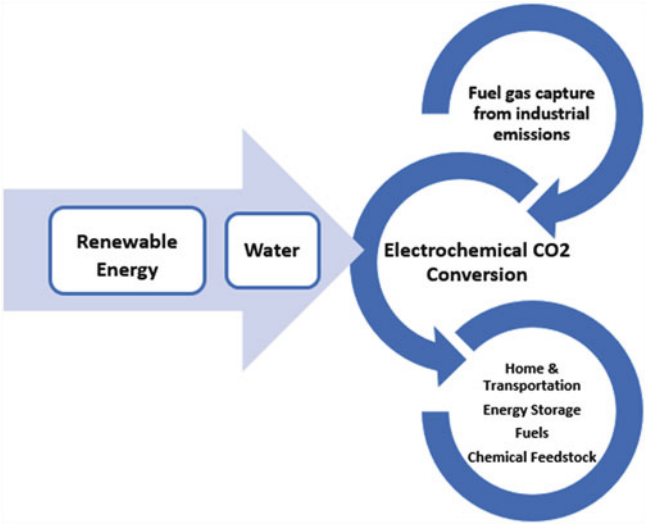

Fig. 5 Electrochemical CO_2 conversion: a negative emission process involving renewable energy sources. Source: author

is the renewable energy technology which can produce energy from zero-emission sources; therefore, this whole process will become a negative emission footprint. Recent research in this area has highlighted a variety of substances can be made in this process. These include alcohols, oxygenates, synthesis gas and other products. The CO_2 extraction from the atmosphere using renewable energy has the potential to act as a long-term storage of energy generated from renewable sources in the form of other products and fuels, a process that decarbonises the atmosphere and provides new clean sources of energy and feedstock.

Fuels generated from this process can be stored long term and turned into another forms of energy when required. The electrochemical conversion process can be conducted in times of energy abundance such that the process running costs are as low as possible. Decarbonisation of the atmosphere is a business that will be very profitable should the price for carbon reflect the true cost of its effects on the natural environment. Renewable energy and its clean production of abundant energy will soon make this business opportunity viable and attractive.

5 Conclusion

Renewable energy and its continued penetration into the energy supply mix has caused an energy transition that will continue to transform the power grids and energy ecosystem. This transformation has created opportunities for investment and innovation that can effectively utilise the surplus of energy that otherwise would be wasted, and turn it into value and opportunities for further decarbonisation of atmosphere. These opportunities will make energy-intensive industries more viable compared to the past when the energy was produced by conventional generators. The surplus of renewable generation can provide abundant inexpensive energy for innovative and future-looking industries to become sustainable and return a stable profit.

References

Advision. (2019). *The cost of desalination*. Retrieved August 12, 2019 from https://www.advisian.com/en-gb/global-perspectives/the-cost-of-desalination

California Independent System Operator. (2019). *Managing oversupply*. Retrieved August 12, 2019 from http://www.caiso.com/informed/Pages/ManagingOversupply.aspx

Quantum Technologies

Daniel Akenine

Abstract

Many have heard about quantum computing, but very few understand how the technology works and it is common with misunderstandings. Will it make my computer go faster? Will it change how AI works? Will I soon have a quantum computer in my mobile phone? Is there an app for that? Why is it not here already? Will it ever be? This chapter will discuss some of the core concepts of quantum technologies. We will see that quantum is not only about computing and discuss some possible new applications on the horizon.

1 Introduction

Winter days are short in southern Sweden, and it was one of those dark days in December 1995 when I first encountered quantum mechanics. I was taking a class in quantum physics, studying for a degree in Engineering Physics at Lund Technical University. At the time, I was in my 20s and had for a couple of years gone deep on topics like material science, laser physics, astrophysics, and electromagnetism. I felt these topics fitted together like pieces of a bigger puzzle and made a lot of sense. But quantum mechanics were different. It did not make sense to me at all.

Why did it not make sense? Common sense is based on your experience of the world surrounding you. Things you can see, hear, smell, and read. All these inputs create knowledge, and by using this knowledge, you form an understanding of how the world works. But quantum mechanics is not describing the world you can see. It is describing the world for very tiny things, which means it is challenging to understand quantum mechanics based on common sense. In fact, common sense

D. Akenine (✉)
IASA, Stockholm, Sweden
e-mail: daniel@akenine.net

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_2

may be a burden. To understand quantum mechanics and the reality for the world of the very small, it helps to use mathematics.

However, as this chapter will mostly focus on potential future applications using quantum technology, we will not use any mathematics. Instead, you need to trust that the rules of quantum mechanics have been the result of mathematical predictions and interpretations, verified (more or less) during years of experimentation.

2 Concepts

Let us start with a question: What does "*quantum*" in "*quantum mechanics*" mean? The word sometimes is used in phrases as "*a quantum leap*," which means something similar to a "*big move forward,*" but "*quantum*" does not mean big. Instead, quantum refers to the smallest amount of something that you can have. It means something that cannot be split or divided, something with discrete values.

The understanding of quantum mechanics is not old. In the year 1900, some physicists believed that all important things in physics had already been discovered, and the future was all about refining and gathering facts. Achieving better precision. However, there were annoying facts that were difficult to explain using classical physics. One of these things was how light can travel through space from the sun to the earth? Light is a wave, and waves need a medium to propagate. As an example, sound waves need air (or some other medium) to spread. Another annoying thing was black body radiation. I will not get into details of the difficulty in explaining black body radiation using classical physics, but the study of black body radiation at the beginning of the twentieth century by Max Planck, a German theoretical physicist, became the start of quantum mechanics.

When I took my course in quantum mechanics in 1995, it was almost 100 years after the start of an intense period in modern physics where people like Albert Einstein, Niels Bohr, Werner Heisenberg, Erwin Schrödinger, and many others formed the mathematics and foundations of quantum mechanics. A theory that challenged the way we think about concepts like causality, locality, and determinism.

As said before, this chapter will not go further into the mathematics of quantum mechanics. Instead, we will look into the possible future use of quantum mechanics when it comes to developing new technologies like quantum computers, quantum communication, blind quantum clouds, and more.

However, to be able to understand possible future innovations, we need to go deeper into three (actually four, but more on that later) concepts that are essential pieces in any quantum technology. These are *superposition*, *measurement*, and *entanglement*.

2.1 Superposition: Life Is Uncertain

On a typical day, I spend my time either at home or at the office. However, if I was smaller, particle size small, I could be at the office and home at the same time. One

of the surprising facts in quantum mechanics is that a particle can exist at two places at the same time and behave both like a wave and a particle at the same time. This is difficult to understand, and science has not yet fully understood superposition in all its fascinating details. But one thing is sure; small objects have to follow the rules of quantum mechanics such as the Heisenberg uncertainty principle "*. . . the position and the velocity of an object cannot both be measured exactly, at the same time, even in theory*" (Encyclopedia Britannica 2005).

The concept of superposition and measurement will be critical factors for building secure communications, something we will see later in this chapter.

2.2 Measuring: To Measure or Not to Measure Is the Question

When you measure something in the classical world, it usually does not change the object you measure. As an example, if you are a doctor measuring the body temperature of a patient using an IR device, or if you are a police officer measuring the speed of a car with a clock, you do not change the temperature of the patient or the speed of the vehicle. However, in the quantum world, the concept of measurement is critical and comes with consequences. As we discussed, in the quantum world, small things like particles can exist in a state of superposition (uncertain state). But if you measure the particle to gain insight into things like position, you will force the superposition to collapse and give a definitive answer to the question you are asking. It would be the same as me working both at home and at the office at the same time—but if you check my house to see if I am at home, I have to make a decision where I am in reality. But as we are much bigger than particles, the concept of measurement does not influence us, right?

You may have heard of the thought experiment called "*Schrödinger's cat.*" Erwin Schrödinger was a Nobel Prize-winning physicist working on Quantum Mechanics at the beginning of the twentieth century. His thought experiment is as follows.

You take a box and put a cat in a box together with some deadly poison in a closed glass bottle, some radioactive material, and a Geiger counter (measures radioactivity) connected to a hammer. If the Geiger counter measures any radioactive activity (radioactive decay is in a state of superposition), it will activate the hammer which will then smash the glass bottle with the poison and kill the cat.

But we cannot know if the Geiger counter has measured any decay from the radioactive material if we do not open the box to look. In the meantime, the superposition of the Geiger counter, hammer, glass, and cat will be in a spooky state where the cat is both dead or alive at the same time—until we look and the state collapses into one or the other reality. There are so many questions to be asked about what is really happening and what the concept of reality means here, but let us not go down that path as it is a big topic. Just one example to illustrate the complexity, one of the possible interpretations is the "*many-world interpretation,*" meaning that the universe splits into both realities, one with the cat dead and one with the cat alive.

What we need to remember about this discussion is that if you measure a particle that is in a superposition, the state gets destroyed, and the superposition is lost.

2.3 Entanglement: Spooky Action at a Distance

Entanglement does not have any good counterpart in classical physics, so it is difficult to create useful analogies. Einstein is said to have called it "*spooky action at a distance.*" Entanglement means that particles under certain circumstances become "*entangled*" with each other, which means they are connected and share a common state. As an example—if you have two entangled electrons and you measure the spin on one of the particles, then the spin of the other electron will be decided as well. The interesting (and strange) effect is that any measurement made on one of the entangled particles will affect the other particle immediately, even if it is in another galaxy light-years away.

Can we produce entangled particles that we can use in applications or for an experiment? That is certainly possible; for instance, you can create entangled photons by a laser beam and certain crystals. Creating and controlling entanglement is a critical component of any quantum computer.

As a side note, it is tempting to think that entangled particles can be used to send information "*faster than light*" over the universe. Unfortunately, that is not possible.

3 Applications

What kind of new inventions can be made using entanglement, superposition, and the impact of measurement? Let us start with the most discussed, Quantum Computers.

3.1 Quantum Computers

To understand how a quantum computer works, we need to introduce the concept of "*quantum bits*" or "*qubits.*" In classical computing, we process bits. Information on a computer or mobile phone is stored using bits that can have only two different states, either 0 or 1. To be useful, you need to develop some way to represent the value of 0 or 1, and there are many ways to do that. You could create bits using transistors, or you could use magnetism to store bits on a hard drive. There are many other techniques you could use to store bits on as well. Punch cards are an early computing example, or you could use flags in your garden. If the flag is up, it would mean 1; if it is down, it would mean 0. Some of these methods are more convenient and faster than others, of course. A bit is not connected to any specific technology but rather a concept on how to store information using two states.

A qubit is a concept that uses more complex rules than regular bits. It can be both 0 and 1 at the same time as illustrated in Fig. 1. It seems like a thought experiment,

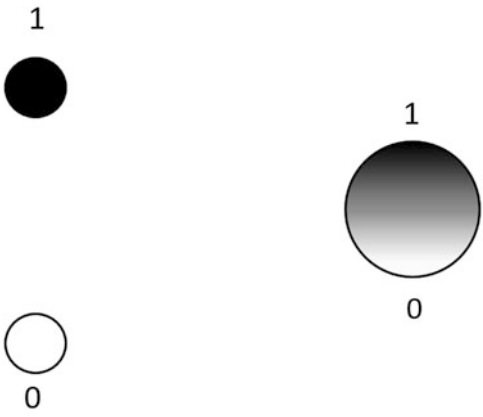

Fig. 1 Classical bit to the left and a qubit to the right. Source: author

but we can produce qubits that support this concept in reality. To store this particular state, we can use things like electrons or photons that exist in a superposition. Superposition makes it possible to process a lot more information by using relatively few numbers of particles.

If one qubit can be in a superposition of two states, then two qubits can be a superposition of four states. Three qubits can be in a superposition of eight states, etc. The number of states increases logarithmically by 2 to the power of N, so when N gets bigger, you will end up with some very, very, very big numbers. This means that just using a small number of qubits; you can hold many more possible states than regular bits can handle.

It is essential to understand that a quantum computer is not a faster or better version of a classical computer. It is a different technology, and quantum computing comes with both advantages and disadvantages compared to traditional computing.

You can create qubits today using several different techniques; examples include atoms, quantum dots, or superconducting circuits. These qubits can then be processed by the quantum computer.

If we can create those qubits, why do we not have quantum computers around us already today? The big challenge with quantum computers is not necessarily the creation of a qubit. The problem is to maintain the superposition state of all the qubits. This means that the calculations performed today with a quantum computer have a lot of errors, noise, and sudden loss of the quantum state. If the qubits interact with anything, they lose their superposition and become bits instead of qubits.

There is intensive research on ways to create less noise and better error corrections. And the future contains a lot of promises. As an example, Microsoft is researching something called topological qubits (Alexander 2019) that could make qubits much more resistant to noise. The power of a quantum computer is a combination of both the number of qubits as well as the quality of them. Soon we could reach the point of "*quantum supremacy,*" the point in time when quantum computers can start solving problems that classical computing cannot do today. Some argue that we are already there, true or not, the significant change will come

when we are able to do useful things with quantum computers that we cannot do today.

As we have understood, a quantum computer is different from a classical computer in many ways and is of little use in many real-world problems. However, sometimes, the problem you are trying to solve matches the capabilities of a quantum computer perfectly.

As an example, creating computer models of molecule behavior becomes easier if the quantum computer used to model the molecules is based on the same rules as the molecules themselves. These types of problems are difficult to model on classical computing because the difficulties increase exponentially. If you are looking into solving these types of exponential issues, you may be interested in quantum computing.

Let us discuss some problems quantum may be used to solve in the future.

3.2 Shor's Algorithm: The End of Encryption?

Shor's algorithm is one of the most discussed when it comes to quantum algorithms. The reason is that it breaks one of the most essential encryption schemes that we use today on the Internet (Loeffler 2019). RSA encryption is based on asymmetric encryption with public and private keys and is used to encrypt the information from your bank or your healthcare provider. In fact, it is used in many other situations as well, from password-free logins to digital signatures, etc. It is safe to say that RSA is one of the most used and well-known encryption algorithms used today.

Breaking this algorithm would cause a lot of harm to security and privacy. Yet, in 1994, Peter Shor published an article describing a quantum algorithm that could break RSA encryption. When Peter wrote the piece in 1994, quantum computing was still mostly a theoretical subject, and there were few ideas on what such a computer could be used for. Peter Shor's algorithm showed that a working quantum computer would be able to do useful things classical computing could not do, and as such, it sparked a lot of interest in quantum algorithms.

There will probably be an end to the traditional RSA algorithm, so before we have quantum computers advanced enough to use Shor's algorithm, we need to develop and deploy new post-quantum encryption algorithms. If you are interested in the topic, the NIST Post-Quantum Cryptography Standardization initiative is worth following (NIST 2019). It could be worth noting that today's digital signatures based on RSA will not be safe in the future and needs to be updated. There is also a need to analyze the effect on cryptocurrencies like Bitcoin that relies heavily on encryption and hashes.

3.3 Quantum Networks

Quantum computing is exciting, but not the only exciting area that can be addressed by quantum mechanics. Another area is new quantum networks/internets that could

change the way we build secure communications. We have already been discussing the concepts of entanglement, qubits, and measurement, so we have the toolbox to understand how quantum networks will work.

1. In *classical networks*, we send bits of information, today usually as light through fiber optics. As the bits travel through cables, routers, and repeaters over the Internet, they are susceptible to eavesdropping. Therefore, we need to encrypt sensitive data on networks, usually with RSA public/private key encryption already discussed. Also, if we have very secret data, it is likely that the eavesdropper has substantial computing resources and world-class mathematicians trying to crack existing encryptions or find vulnerabilities. Today's networks are quite safe but not immune to different types of attacks.
2. In *quantum networks*, as you may guess, we send qubits instead of bits. As discussed, qubits cannot be measured without losing superposition, which means we can use them to design networks that can detect any attempt to eavesdrop. We could, for instance, send an encryption key over the network and be sure that no unauthorized entity has manipulated or viewed the key. This key could then be used for symmetric encryption of other data.

It is still early days in building these networks, and one of the challenges is to amplify the signal for longer distances. Easier said than done if you want to amplify and repeat the qubits without measuring them! However, it can be done, and pre-quantum networks are in operation today in different places in the world using pieces of the technology needed to build a full quantum network (TUDelft 2019).

3.4 Quantum "*Blind*" Clouds

In the last decades, there have been massive investments in building large cloud data centers all around the world. Technology like AI, Internet of Things, digital identities, and data storage has moved workloads from the local data center to the clouds. One challenge is to use the cloud for extremely sensitive scenarios where the user wants to make it technically impossible for the cloud provider to access any data. If we combine quantum computing with quantum communication, we could imagine future computing clouds that are impossible to eavesdrop. Physics would stop any attempt to do so.

3.5 Other Applications with Quantum Mechanics

- True anonymity

  "*On the Internet, Nobody Knows You're a Dog*" is a famous meme from the early days of the Internet. Even though there are several technologies and algorithms designed to keep a sender in a network anonymous (like the Tor

network), it is hard to be sure you are 100% anonymous. By using things like quantum anonymous transmission protocols, there are possible ways to set up true anonymous communications in the future (Christandl and Wehner 2005).

- New digital signatures

  Quantum computers are a threat to the digital signatures we use today, but future quantum computers could make it possible to create new types of digital signatures that are safe in a post-quantum world.
- Secure voting

  No electronic voting systems are 100% safe from possible manipulation and hacking. This causes a challenge to maintain trust in a digital voting system. Using quantum technologies for proving identity and securing communications could mean much safer electronic voting systems.
- Better GPS systems

  The current GSP systems rely on atomic clocks in satellites. They have very high precision, and some atomic clocks would not even lose a second for 300 million years (NIST 2014). However, the clocks between different GPS satellites need to be synchronized, and with quantum communication, the clocks could get better synchronization. This means better location accuracy, down to centimeters instead of meters. Future location services may also use other things than GPS satellites, like quantum accelerometers and quantum gyroscopes, making it possible to locate your position without the use of satellites. By using a fixed starting point as a reference and, by extreme accuracy, detect any acceleration and movement, you could determine your location.

4 Conclusions

In many ways, quantum has the potential to improve things that we are doing today as well as making new things possible. If we see what quantum may be capable of doing in the next 5–10 years, the impact will likely be an evolution in things that we do today, like solving more sophisticated algorithms and creating more secure networks. Your laptop will hardly transform into a quantum computer; quantum capabilities will instead be delivered over the Internet through cloud platforms convenient to use when the problems are suitable. Today we can see this in the field of AI and machine learning with AI hardware designed to solve specific issues delivered through the cloud.

In the decades after that, it is hard to say what quantum could mean for innovation. Just as the Internet created a platform for innovations that were hard to predict, we will see new applications that are hard to predict building on quantum technology. As the technology is so different, many of these applications are likely to be disrupting existing ecosystem and prove to be game changers.

As always, the future is uncertain but with some outcomes more probable than others—just as quantum mechanics teaches us.

References

Alexander, L. (2019). *Topological quantum computing*. Retrieved December 12, 2019, from https://medium.com/swlh/topological-quantum-computing-5b7bdc93d93f

Christandl, M., & Wehner, S. (2005). *Quantum anonymous transmissions*. Retrieved December 12, 2019, from https://arxiv.org/abs/quant-ph/0409201

Encyclopedia Britannica. (2005). *Uncertainty principle*. Retrieved December 12, 2019, from https://www.britannica.com/science/uncertainty-principle

Loeffler, J. (2019). *How Peter Shor's algorithm dooms RSA encryption to failure*. Retrieved December 12, 2019, from https://interestingengineering.com/how-peter-shors-algorithm-dooms-rsa-encryption-to-failure

NIST. (2014). *NIST launches a new U.S. time standard: NIST-F2 atomic clock*. Retrieved December 12, 2019, from https://www.nist.gov/news-events/news/2014/04/nist-launches-new-us-time-standard-nist-f2-atomic-clock

NIST. (2019). *Post-quantum cryptography*. Retrieved December 12, 2019, from https://csrc.nist.gov/Projects/Post-Quantum-Cryptography

TUDelft. (2019). *Quantum internet | The internet's next big step*. Retrieved December 12, 2019, from https://issuu.com/tudelft-mediasolutions/docs/quantum_magazine_june_2019

Security in Intelligent Transportation Telematics

Erich H. Franke

Abstract

More than 10 years ago, the European Telecommunications Standards Institute (ETSI) began to standardize the communication between vehicles and infrastructure in so-called "Intelligent Transportation Systems (ITS)." This communication is supposed to be self-organized, which means, it has to be operated without the assistance from an access network. However, since most of the communication components are deployed in an inhomogeneous manner by vehicle operators "in the wild," the security aspects are—for the lack of a better word—challenging at least. We will examine the security of ITS networks by discussing different modes of possible attacks.

1 The ITS Ecosystem

In August 2008, the European Commission decided the harmonization of the use of the radio spectrum for safety-related applications of Intelligent Transport Systems (ITS). Furthermore, starting around September 2010, the EC published its general view about the communication architecture of ITS systems, as defined in ETSI EN 302 665 V1.1.1 (2010-09). The communication between the different types of ITS stations may use either direct point-to-point communication in single or multiple hops between the source and destination stations, but the access to public and private networks, including the global Internet and even existing infrastructure and satellite broadcast have been considered in the document.

E. H. Franke (✉)
AFUSOFT Kommunikationstechnik GmbH, Königsbach-Stein, Germany
e-mail: erich.franke@afusoft.com

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_3

The participants of the ITS ecosystem can be assigned to four "subsystems," counted in the number of their appearance:

- Vehicular subsystem: i.e., ITS components can be deployed in motor cars, trucks, security vehicles or similar, in motion or parked. Vehicular subsystems are expected to be deployed in high volume, compared to other types of subsystems. Therefore, sometimes Intelligent Transportation Systems (ITS) are—slightly incorrectly—labelled V2X, an acronym, designating communication "vehicle-to-anything." In most of the cases, the ITS components used in vehicles are owned and operated by the owners of the vehicles themselves.

  A common application is the traffic accident warning, which we will discuss later.
- Roadside subsystem: consisting of components, mounted on traffic lights, gantries, poles, etc. In most cases, the deployed ITS components are owned and operated by the authorities or companies responsible for road maintenance.
- Personal subsystem: e.g., hand-held devices for pedestrians or bicyclists. ITS components in this domain are presently not very common. However, as soon as they become part of mobile phones or tablets, their number is expected to increase significantly.
- Central subsystem: traffic control centers, emergency warning centers, adverse weather forecast centers, etc.

In theory, a variety of existing networks may be used to convey ITS messages between subsystems. However, since particularly the mobile components require to communicate with each other, a decentralized, globally available, low-latency and self-organizing communication scheme is required.

To fulfill these requirements, ETSI standardized a communication scheme, based on IEEE 802.11p, which is known in Europe under the label "ETSI-G5." The protocol stack of G5 is similar to standard WIFI; however, its parameters have been modified, to allow connectionless communication even between speeding vehicles and particularly defines the use of the frequency band between 5.85 and 5.925 GHz, which has specifically assigned to ITS safety critical applications. In Europe, this protocol is known as ETSI-G5 and considered as the basis of the Dedicated Short-Range Communication (DSRC) standard.

In this context, it is very important not to confuse the license-free broadcast, WIFI-like, short-range communication protocol ETSI *G5* with the "next-generation mobile communication standard" *5G*.

The latter is a mobile communication scheme, similar to 4G/LTE which requires an operated cellular network and requires, needless to say, communication contracts to be closed between users and network operators and operation fees paid.

Of course any public mobile network can be used to convey V2X messages, technically. The security, integrity, and privacy, however, depend on the security properties of the respective network and its operator.

The underlying network layer structure of public mobile networks, such as 5G, as well as the different data and voice communication capabilities thereof are beyond the scope of this document.

However, two important properties have to be discussed, regardless of the actual network been used: latency and channel capacity. We have to have an eye on them, since these parameters are important for the understanding of operation integrity and of the vulnerability against denial-of-service attacks:

- Latency: Denotes the travelling time of a data packet from the information source to the destination. For highly dynamic traffic information systems, minimizing this latency is desirable.
- Channel capacity: The amount of information, which can be conveyed through any given communications channel.

The requirements for data transfer in ITS communications are different from general data communications, though:

- Point-to-multipoint: most of the ITS data packets are generated in one station, but intended to be received and processed by many recipients, if not for all, in a given range. The transmission from a roadside station, for instance, is subject to be received by all passing vehicles. Emissions from vehicles shall be received by all other vehicles, roadside stations, etc. in the respective range. Point-to-point operation is merely uncommon in ITS.
- Highly dynamic environment: Since the traffic scenario is spatially dynamic, so is the structure of information links for example between speeding cars. In the ITS domain, the term "geo-networking" is commonly been used to describe this kind of scenario. Therefore, ITS communication is practically always connectionless and considered self-managed.
- Information Relaying: Important messages will be relayed by other vehicles in order to extend the information range.

Very well! With ITS, we have to deal with an ecosystem consisting of inhomogeneous entities that dynamically provide each other with relevant traffic-related information. In a perfect world, where all human beings are brothers, everything would be fine so far. In the real world, however, action must be taken to prevent malicious participants from abusing the system. Let us investigate this aspect by examining a typical scenario.

2 Application Scenario Versus Vulnerability

One of the simplest, yet highly important applications derives from the *car crash/traffic jam warning* scenario. Given the—unfortunately inevitable—fact that two vehicles collided on a highway. The airbag sensor will trigger an emergency message, which can be received by all other vehicles in the given range of—

say—200 m. Based on the transmitted geo information, the onboard computer of approaching vehicles—particularly when autonomously driving—may then decide, to perform an evasive maneuver, emergency breaking or such. This measure is more or less automatically controlled ("Direct Control") and intended to prevent another collision by supporting cameras and laser scanners of autonomous vehicles or the eyeballs and/or the brain of a human driver.

The risk level reduces with the actual distance to the crash site. In—say—500 m distance, an approaching vehicle gets a "Collision Risk Warning" (LCRW, ICRW), a bit further upstream, a "Road Hazard Signal" (RHS) is raised for awareness and further away, an information is given, e.g. to be displayed as an "In Vehicle Signage" (IVS). All these steps are considered "Primary Road Safety Applications." Related scenarios include warnings of construction sites or slow-moving maintenance vehicles.

In each case, the "warning" will very likely trigger some automatic actions in the receiving vehicles, from emergency braking, evasive maneuvers, and speed reduction. Since these actions have a severe impact on the traffic flow, the "warning" messages have to be protected adequately.

But there is an even more important scenario: Think about an ambulance, a fire truck, or a police car in a city, which try to drill through the common urban traffic jam during rush hour. Not an easy task and also rather risky! An "electronic lightbar," (e.g., the German "blue light") using ITS communication, as we are currently testing it, could help to warn the other drivers from an approaching rescue vehicle.

But an even better solution is to offer the emergency vehicles free travel at traffic lights by controlling them accordingly. The traffic lights have to be equipped with ITS receivers—so called "roadside units" which examine specific messages denoting the approaching rescue vehicle and identifying their trajectory using the vehicle's geo position and movement vector.

When the trajectory of the rescue vehicle crosses an intersection, the traffic light controller can switch to a special signal phase in order to stop other vehicles and therefore give the rescue vehicle an appropriate right of way.

Both the *car crash/traffic jam warning* and the rescue vehicle right of way scenario can be very helpful. However, since they can have a severe impact on the traffic average flow—on a motorway as well as in a city—any abuse must efficiently be prevented.

3 Signature as the Primary Security Measure

The term "abuse" implies an intentional attack of some kind. This leads our discussion toward the security of such a system: Information Technology Security and Information System Security denotes the protection of the system components "... from intentional theft of or damage to their hardware, software, or electronic data, as well as from the disruption or misdirection of the services they provide ..." as this term is generally defined.

The creators of the ETSI G5 standard decided to cryptographically protect communications, however by *signing* the transmitted messages, rather than *encrypting* them.

The reason for this decision is that each participating station is basically able to receive every message, even if it does not currently have possession of a valid crypto key. In ETSI's opinion, each recipient is responsible for deciding whether or not to use any unsigned or erroneously signed message received.

Of course, a valid message may only be created properly if the sending party has used a valid certificate for signature.

The key distribution system has to take into account, that participants receive their certificates through different channels. This means that in the background, a hierarchical and geographically distributed Public Key Infrastructure (PKI) has to be implemented. Consider, for example, that a French vehicle on a German motorway has to exchange messages with an Italian truck easily, even though all the certificates have been issued in different countries!

The actual certificates, however, have to have a limited validity lifetime. However, the actual certificates must have a limited validity period. Since, for example, a stolen RSU could be used for a hostile attack on the road infrastructure and the vehicles on it, it is important to keep the time window for nay successful abuse as short as possible.

In present systems, the certificate used for signing actual messages, is valid 24 h maximum. From a security point of view, although a much shorter interval—e.g., 1 h—would be desirable, this would have a significant impact on the practical usability of the system. Keep in mind that the certificates must be requested online from a PKI, as they cannot be kept in stock for security reasons. This implies that each participant must have more or less permanent online access to the Internet, which may not be a problem for a stationary road unit (RSU), but might have a massive impact on the mobile users, for example, cars and trucks.

The process of requesting the certificates has to be secure—too. Presently, a four-step scheme is implemented. On the top level, every manufacturer of ITS equipment has to be registered with his/her local security agency. Using the "manufacturer's certificate," the provider may request intermediate certificates to create a "trust anchor." All certificates in the chain have different, limited lifetimes, and of course, all requests are cryptographically protected. If all intermediate steps have been processed successfully, the "Authorization Ticket" can be requested, which is then used for signing during the next 24 h period.

4 Security, Safety, Integrity—and Privacy—Issues

Let us now discuss the different implications of the security issue.

Virtually all mobile components of an Intelligent Transportation System, i.e., particularly those mounted in vehicles, are owned and operated by the users themselves. Therefore, the *physical* security, e.g., theft prevention, requires appropriate precautions of the very users. This is particularly important, since the

ITS components contain credentials, e.g., crypto keys. Although these credentials have limited validity, they may still be misused by malicious users. The same consideration applies to road side units. Malevolent users, who succeed in gaining access to a road side unit, might exploit the RSU's transmissions in order to simulate traffic congestion or the like. This kind of misdirection is one of the most likely scenarios for malevolent attacks on ITS.

Another security issue, to be considered, is the classic "Denial of Service" attack. In ITS, this jamming scenario is not always caused by an intentional misbehavior of users or attackers. Since ITS messages are conveyed by radio frequencies, mostly on 5.9 GHz, these can easily be disrupted. This means that any ITS system must be aware that the probability of reception of messages is anything but stable. The conditions are rather like in any Wi-Fi network.

However, malicious attacks are usually in the minority in this scenario.

In general, the reception probability is significantly reduced simply because the vehicular users move relatively quickly on motorways or in an urban area, yielding undefined a nonstationary signal reflection patterns.

Given the fact that ITS uses broadcast transmissions rather than connected ones, the information redundancy is created by repeating the same message with a relatively high information bandwidth.

The latter is important in order to get at least some of the data through the channel. The so-called Cooperative Awareness Message (CAM), which tells all other users of the state of a vehicle, its location, its movement vector and a lot of information more, is repeated with a variable rate between ten times a second down to one per second. The actual transmission rate depends on the speed of the vehicle, if it follows a curve or any state change.

Now consider a scenario on a typical motorway intersection, with several thousand vehicles per hour speeding in different directions. Even if only a few percentage of vehicles are actually equipped with ITS components, there is a lot of radio transmissions on the air.

For information "self-defense," each ITS component uses so-called congestion control functionality, reducing its transmissions if the channel is highly populated.

Therefore, it can be said that traffic flow restrictions are not always the result of an external attack.

The term "safety" denotes "... a state in which ... you are safe and not in danger or at risk..." according to the Cambridge Dictionary.[1] We might discriminate "safety" from "security" in conjunction to unintentional, accidental, and even random risks. In ITS, designers have to take precautions not only against the result of malicious attacks, but also against effects caused by mere system malfunctions. A good example is the GNSS position, which is included in the geo-networking portion of virtually every ITS message. If the satellite navigation signals is jammed, e.g., by multipath reception in an urban environment in one vehicle, the geodetic position may be erroneously distorted. All recipients of this information hence have

[1] https://dictionary.cambridge.org/de/worterbuch/englisch/safety, accessed on December 4, 2019.

to examine the integrity information supplied, in order to be sure, where the sender's vehicle is actually located. This is important, since all moving participants record their recent positions in so-called "traces," posting them over-the-air. All other receivers may exploit these "traces" in order to assure, that their current movement vector is compatible, to provide a form of early warning to a possible collision hazard.

Failing this precaution might lead to unintended behavior, particularly in autonomous vehicles, such as unnecessary lane changes or even braking maneuvers.

In ITS, we have furthermore to assure privacy. The Business Directory defines this term as "... the right to be free from secret surveillance and to determine whether, when, how, and to whom, one's personal or organizational is to be revealed ...".[2] Privacy is one of the key factors to assure user acceptance of this technology.

However, full protection of users' privacy is not an easy task at first. ETSI G5 communication works similarly to other Wi-Fi networks. Each message contains a unique identifier, the MAC address (Media Access). This MAC address is used to ensure that informational messages or replies can be forwarded to the intended recipient.

Since G5 usually operates based on broadcast messages, this MAC address is far less important than in a Wi-Fi network. As a consequence, ETSI defined a scheme to modify the MAC address of each sender frequently during operation in a way, which cannot be reversed. This method, called "pseudonymization," is intended to technically ensure, that it is not possible to identify a mobile user after a short period of time.

5 Why May Someone Want to Attack Any ITS Network?

When we talk about an abnormal, unwanted, or even illegal use of data in an ITS network, we have to keep in mind that the nature of these "nonstandard users" and their intentions can be very different.

We can identify the following groups:

- Road service providers
- Law enforcement agencies
- "Curious persons"
- Malicious attackers

When discussing privacy issues in relation to the first two parties, conflicting aspects must be taken into account.

The transport service providers want to use the data of mobile users to monitor the utilization of the road and the traffic flow. These traffic data are valuable

[2] http://www.businessdictionary.com/definition/privacy.html, accessed on December 4, 2019.

for the actual road users, e.g., in order to derive suggestions for diversions or alternative routes. Of course, these data could also be sold by road service providers, but this would not harm the privacy of a single user. Usually, anonymized or "pseudonymized" data are sufficient to provide this task.

A little bit off are requirements brought in the discussion by law enforcement organizations to use ITS movement data, e.g., for speed violation ticketing or the prosecution of red light violations at traffic lights. Technically, it would be possible to implement these—somewhat strange—ideas with ITS, since all information necessary are already included in cooperative awareness messages (CAM). However, automatic ticketing systems, based on ITS technologies are, not yet considered legal, at least today in western countries. In Saudi Arabia, however, a similar system has already been fielded, as stated by Jan (2014). It is expected that ITS will be used in the future to determine the travel time on motorways, as it is done today by using the Bluetooth emissions of cars, as reported in Spangler et al. (2010).

Pure passive monitoring of ITS activities on highways or in urban environments is easy for any group of curious people to accomplish. A simple WLAN stick capable of receiving IEEE 802.11p emissions along with a simple DIY processor board and a small piece of software is sufficient to perform this task. The specification of the ETSI G5 protocol stack is almost completely in the public domain, and since G5 implements signatures rather than encryption, no certificates are required to exploit an ITS message. However, this is not completely uncritical: Remember that with these rather simple techniques you can monitor ambulances, fire engines, and even police cars. Even without evaluating the MAC address, recognizing these types of vehicles is far from impossible if you exploit the metadata contained in CAM and DENM: vehicle length, width, load, type, speed, acceleration, even the state of the headlights, everything is easy to exploit. Our company has already received requests from law enforcement agencies to "anonymize" at least the speed of its vehicles.

Much more critical than the pure monitoring would be the injection of phony or malicious messages in an ITS network. This could be considered the equivalent of throwing stones from a bridge onto a populated motorway. Attackers can be mere "script kiddies," using openly published malevolent application programs from the Internet. But also criminals or even terrorists could use such devastating techniques, for example, to bind or slow down law enforcement during their illegal activities.

Of course, the ETSI-defined cryptographic signature scheme should protect the network from such malicious user attacks. However, experience tells us that malicious hackers rarely enter the system through the front door. ITS developers therefore have not only to rely on cryptography, but also on sanity checks and data fusion in their software.

6 Conclusions

Intelligent Transportation Systems are currently being created and deployed worldwide in order to orchestrate the various transportation systems in a sustainable and environmentally friendly way. Reasonable operational and communication security,

however, are the key elements for the user acceptance and, ultimately, for their success in the real world.

References

ETSI EN 302 665 V1.1.1. (2010-09). Intelligent transport systems (ITS); Communications architecture. *European Standard (Telecommunications series)*.

Jan, Y. (2014). *Drivers' perception of Saher traffic monitoring system in Jeddah, Saudi Arabia*. Masters theses and specialist projects paper 1438. Retrieved March 12, 2019, from http://digitalcommons.wku.edu/theses/1438

Spangler, M., Leonhardt, A., Busch, F., Carstensen, C., & Zeh, T. (2010). *Deriving travel times in road networks using Bluetooth-based vehicle re-identification: Experiences from Northern Bavaria*. Conference paper: FOVUS - Networks for Mobility. Stuttgart, Germany. Retrieved March 12, 2019, from https://www.researchgate.net/publication/266064514

Innovation and Future Technology Scenarios in Health Care: Ideas and Studies

Philipp Plugmann

Abstract

Innovations are turning the health care market into a technology-dominated sector. Artificial intelligence (AI), preventive medicine and all variations of upcoming technology will influence the health care market of the future. In the short term, it seems unthinkable in times when all expect higher costs in the future that we will accept that our economies need to lower health care costs by 20% or even more. People are getting older, leading to many additional treatments in the next decades, e.g. knee replacement treatments, cancer therapy treatments, coronary heart disease treatments and many other therapies that will emerge should people reach ages of, on average, 100 or more. The economy can only meet the requirements for successful international economic competition if we stop the upcoming explosion in health care costs, which are mostly financed by health insurance contributions and taxes. However, the reality is that our health care systems in Europe will not be able to deliver the best medicine to all people, especially not in those countries with demographic change, if the costs rise in the next decades as forecasts predict. We would like to present several ideas for and long-term scenarios of a health care revolution as well as present scientific studies, including the acceptance of users and patients, because we think acceptance by users and patients is a key factor for success in the future. Automation, BIG DATA combined with AI and patient cooperation are the requirements for a highly efficient and cost-effective health care system. Another key factor is preventive medicine. Let's start with the idea of logic automation in health care.

P. Plugmann (✉)
SRH University of Applied Health Sciences, Leverkusen, Germany
e-mail: philipp.plugmann@srh.de

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_4

1 Self-Driving Hospital Beds

Most hospitals in Germany will have problems recruiting enough nurses and physician assistants in the next years. The reason for this is the demographic transformation. It's not just a quantitative problem of hiring enough workforce in the future, but also a qualitative problem as these jobs also require social and emotional skills, which must reach a certain standard without exception. Care assistants and nurses are currently exhausted and the percentage of sick days is, based on Deutsche Ärzteblatt (2019), growing. So how can this problem be handled to help this workforce deliver their service to the patient?

Much time is lost in inefficient transportation of patients between different departments, for example from orthopedics to radiology and back again. As personally experienced nearly 30 years ago while working at the University Clinics of Cologne for several years as medical student help to finance my studies, the hospital bed itself is heavy, the brakes are difficult to release, transportation is slow because the bed frequently moves to the right or left, often due to worn tires, and doors need to be opened. The elevator waiting time combined with partly filled elevators is inefficient. Moreover, when you have finally arrived at the new department, it can happen that appointments have been changed, cancelled or transferred to another department, meaning you need to return with the bed and waste a lot of time. Sometimes two persons are needed to transport a hospital bed.

So, what about an efficient, fully-integrated concept of self-driving hospital beds in conjunction with self-opening doors and coordination with elevators and the schedules of all departments? The time won will lead to more time for care and less stress for the workforce. Staff sick days may decline, and employee and patient satisfaction indicators increase. The following research question is important for this idea: "How will patients accept transportation in a self-driving hospital bed?"

In a questionnaire-based multicenter study between January and December 2018, we surveyed 264 patients in two dental and two general medical clinics in North-Rhine Westphalia, Germany. The inclusion criteria were: 30–70 years old, with at least one period of inpatient hospitalization of at least one night due to a health crisis incident or for another reason. We also categorized by age groups, education, income and experience with technology, e.g. smartphone use. The exclusion criteria were: older than 70, younger than 30, no experience in hospitals and no technology experience.

The results of this quantitative empirical research study showed that overall 94.32% (n = 249) of the 264 patients interviewed would have no problem being transported by self-driving hospital beds. Even in the group of 60–70-year-old-patients, the acceptance rate was 91%. At over 85%, the three most important factors for the patients were the safety of the technology, insurance for accidents with the bed and an emergency button. Generally, there was a high acceptance rate for self-driving hospital beds in this study. With more than 2000 hospitals and hundreds of private clinics and rehabilitation centers in Germany, there is a huge market

for hospital beds. Germany has all the necessary technological components and production options, so it could develop the next big thing in health care technology.

2 Reorganization of Medical Studies with Contests and Crowdsourcing

Competition is a natural status quo, also in health care. Medical students should therefore get used to it. Contests, research projects and interdisciplinary projects all offer a good opportunity to train competition. The power of contests and crowdsourcing is a tool that can be used by individuals and organizations interested in starting new companies or pushing projects for new products and services in HC.

A few years ago, we started a study on the "Willingness of German Medical Students and Experienced Physicians to Participate in Online Contests to Innovate in the Field of Health Care". Presenting the first results at the Open and User Innovation (OUI) Conference at the Harvard Business School (Boston, USA) in August 2016, we discussed the further impact of contests and crowdsourcing in health care.

Crowdsourcing can be a powerful tool to solve problems in the field of health care (Lakhani et al. 2013). In the age of digital transformation, the technologies allow individuals and networks to interact in an efficient way characterized by high-speed performance per time unit and the irrelevance of geographic distances. It even no longer seems essential to be a long-time working expert in a certain field of knowledge to create innovative solutions. The strategy of using the crowd as an innovation partner (Boudreau and Lakhani 2013) can help to find a better innovative solution in a shorter time.

The influence of users, user communities (Harhoff and Mayrhofer 2010), lead-users (Herstatt and von Hippel 1992) and producers on the innovation processes from different perspectives means that we live in times of democratized innovation (von Hippel 2005). Through online contests in the field of health care to innovate with crowds, where you can participate as an individual with your own idea or work on a special problem as part of a crowd, access to participation is generally open. In 2016, there were a range of national and international online contests in the field of health care, such as the Health Acceleration Challenge of the Harvard Business School and Harvard Medical School in the USA, the Medical Valley Open Innovation Platform Contests in Germany and the Business Plan Competition of the China Healthcare Investment Conference (CHIC). The research area we focused on in this study concerned the breadth of the crowd involved in the online contests in the field of health care because the participants are usually also experienced physicians acting in different executive capacities in clinics or in the health care industry. The specific research question we were interested in was to compare the willingness of a group of German medical students to participate in online contests in the field of health care with that of a group of experienced German physicians.

For this study, we surveyed two groups in the timeline November 2014–January 2016. The first group were German medical students ($n = 258$) from five universities

in Germany without regard to the progress in their studies. The second group were experienced German physicians of different disciplines who had worked clinically for a minimum of 10 years ($n = 184$). The research process was organized differently for both groups. The medical students were interviewed on university campus, especially at times of conferences or exhibitions on campus when we could expect to meet and interview a large number of students with a standardized question paper at one time. The second group of experienced physicians were first invited to participate in this study by e-mail with a repeated 2-week reminder e-mail. Then, if they replied positively, we sent them a standardized question paper by e-mail. Both groups were asked to name three factors which could lead to a higher willingness of their peers to participate in online contests.

Regarding the first group ($n = 258$), we found that 78.29% ($n = 202$) of the German medical students asked were willing to participate in online contests to innovate in the field of health care, compared to 30.43% ($n = 56$) of the experienced German physicians asked in the second group ($n = 184$). The three most important factors named by group 1 that could encourage their peers to participate more in online contests to innovate in the field of health care were: credit points for extra-curricular activities related to innovation (77.91%), more information from the medical and other faculties about online contests (71.32%) and interdisciplinary events on campus (57.75%). In group 2: more detailed response to the ideas submitted (60.87%), real-life meetings face-to-face with a part of the participating network (52.17%) and more interaction in sharing the final results of the contest (45.11%).

The empirical results of this study clearly showed that the group of German medical students has a significantly higher (78.29%) willingness to participate in online contests to innovate in the field of health care than the group of experienced German physicians (30.43%). Asked about factors to improve willingness to participate within their peer group, the most important factor for the German medical students was credit points for extra-curricular activities related to innovation (77.91%) and for the experienced German physicians more detailed response to the ideas submitted (60.87%). For the future of online contests regarding innovation in the field of health care, some thought should be given to widening the crowd to include a higher number of medical students. On the one hand, the contest itself would benefit from the additional perspective of this group of participants, who have less expert knowledge while studying medicine, but may in some ways have a less complex view. On the other hand, the medical students themselves would learn from their very first years at a medical faculty to think in innovative patterns.

The results of this first study on the topic posed the question of what results would be obtained if students of IT and engineering were asked. We therefore set up a follow-up study in 2017 and 2018 focusing on this group of students titled “Willingness of Students of IT and Engineering to Participate in Health Care Contests”. The factor of crowdsourcing was not evaluated in this follow-up-study.

We surveyed two groups in the timeline August 2017–November 2018. The first group were German IT students ($n = 114$) from three universities in Germany without regard to the progress in their studies. The second group were German

engineering students ($n = 129$). The students of this follow-up study were interviewed directly on university campus. Both groups were asked to name the main factor which could lead to a higher willingness of their peers to participate in online contests especially in the field of health care. For 87% of the IT students and nearly 90% of the engineering students, this was raising the visibility of their personal performance to industry.

Regarding the first group, the IT students ($n = 114$), we found that 90.35% ($n = 103$) were willing to participate in online contests to innovate in the field of health care, compared to 71.32% ($n = 92$) of the engineering students in the second group ($n = 129$). The empirical results of this study clearly showed that both groups have a high willingness to improve health care and participate in such contests. If we take the first study with medical students and experienced physicians and the follow-up study with IT and engineering students, it is clear that any platform open to these groups would be very successful driving solutions and innovations in the field of health care.

3 All-In-Data-Approach in Health Care for New Business Models

The future health care business model could deliver guaranteed full health care service for a certain amount of money. For private companies, one of the requirements needed to deliver this full-service business model is 100% availability of all individual data of a patient, not just medical records, but really all data, in order to deliver the best possible risk analysis to the customer. This depends on the willingness of individuals to support this "All-In-Data-Approach". I have worked in various international projects as an advisor on research into the acceptance of such a business model from the perspective of the user. In November 2015 I presented the results of my study, conducted while helping to build a new service for users for an industry partner I was working with, with a poster at the 2nd World Open Innovation Conference (WOIC) at Santa Clara/Silicon Valley (USA), which was organized by Prof. Chesbrough (Berkeley Haas School of Business, University of California, Berkeley). Titled "Users (Patients) willingness to transfer personal data to a future-IT-service of Open Innovation driven IT Health Care companies to receive an efficient service—a follow-up study", I would like to share it with you.

3.1 Introduction

It has been found that technology companies, also in the health care sector, use a limited open innovation approach (West 2003) to reduce costs in research and development, and to achieve higher profits (Chesbrough 2006). The integration of the public in research and development in health care is seen as essential for the advancement of innovation (Bullinger et al. 2012). Higher profits can only be achieved with new products and services for the market that are ahead of competitors

and meet user needs. Today this means mobile-Health (Estrin and Sim 2010) because by opening the m-Health architecture, the barriers to entry are lowered, and the development of new tools through participation of the community helps design of new m-Health apps.

These IT companies involve users and lead-users as innovators (Bogers et al. 2010) to develop new products and services. Lead-users, in particular, can help to create new products and services that are not on the radar of market research or internal innovation teams to achieve breakthroughs (Von Hippel et al. 1999). We present the high percentage of users (patients) willing to transfer their entire personal data, medical and non-medical, to a future IT service of an open innovation driven health care IT company in return for a better health care service.

3.2 Theoretical Background

In this chapter we analyze user (patient) willingness to transfer all data to a future IT service of an open innovation driven IT health care company to contribute radical innovation (Lettl et al. 2006). Although the open-innovation approach with users and lead-users has been very successful in the past under certain circumstances (Reichwald and Piller 2007; Baldwin and von Hippel 2011; Van de Vrande et al. 2009), the process of the open innovation architecture needs to be reconstructed, and there is still uncertainty as to whether the user (patient) will support a future IT service that collects not only medical data, but also non-medical data.

The results of this study can help entrepreneurs in the health care IT industry to prototype a future IT service based on an open innovation approach and to decide how open it can be (West 2003). For a holistic health care IT service approach, it helps to understand user requirements for the transfer of medical and non-medical data.

3.3 Research Design

This follow-up study is based on the first study presented in 2014. The first study analyzed user (patient) willingness to transfer medical data to an innovative health care app. This follow-up study analyzes willingness to transfer medical and non-medical data to an open innovation app in return for a better health care service.

3.4 First Study

Reporting on the first study on the willingness of users (patients) to transfer medical data to a health care app at the 12th Open and User Innovation (OUI) Conference at Harvard Business School (HBS) from July 28–30, 2014, we presented our view that there will be various future scenarios of interaction between the user (patient) and health care IT companies and their applications, as depicted in Fig. 1. The future

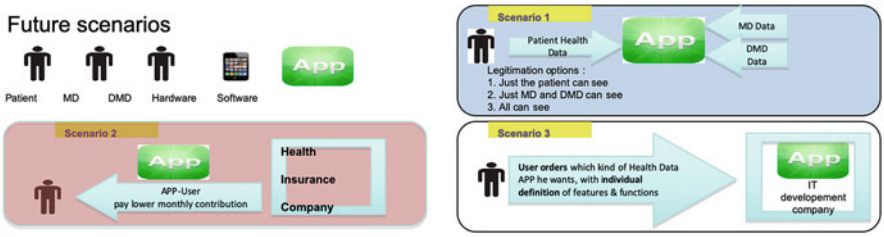

Fig. 1 Three future scenarios as presented at the 12th OUI Conference in July 2014. Source: author

scenarios and prototypes help to understand and involve the users and create new products and services (Kanto et al. 2014; Parmentier and Mangematin 2014; Steen et al. 2014):

First future scenario: The user (patient) decides if the medical and dental doctor may enter information into the app in addition to the patient himself and who is legitimated to view the data and the results of data input in dependence on the IT application. Second future scenario: The health care insurance company (or other service delivery companies in the health care industry) offer app users a lower monthly premium/price if they use the company's app. Third future scenario: The IT development company provides and the user orders which kind of health care data app he wants, with individual user definition of features and functions.

These three future scenarios and the results of the study on user (patient) willingness to open personal health data to an innovative app in return for a more efficient health care service were presented in July 2014. The material and methods were in that from January–December 2013, two groups of patients totaling 528 with a history of periodontal disease were interviewed in Germany in a multicenter study (four dental clinics). In the first group ($n = 244$), no user had a prior general disease (e.g. diabetes, coronary heart disease), while in the second ($n = 284$), they had a minimum of one general disease or more.

We found that 93% of the second group would significantly ($p < 0.02$) open their individual health care data to such an innovative app, allow the MD and DMD to enter certain medical parameters and would also enter information daily/weekly on, for example, how they feel, what they eat and if they still smoke. In the first group, 32% would open their individual health care data to such an innovative app.

3.5 Follow-Up Study

The first study merely determined the flow of health data from the patient, MD and DMD in interaction with software (app) and use of a mobile device. Following the very high rate of user (patient) willingness (93%) to open personal health data to an innovative app in return for a more efficient health care service in the first study, the next question was what would be the case if data collection were expanded to a

holistic approach. The holistic approach to deliver a better health care service to the user (patient) would need medical and non-medical data on the user. The influence of such a future user community and the potential results of the research data based on the future IT service could also help develop open innovation processes and future research in the open innovation field (Chesbrough and Bogers 2014).

Research Question We prototyped a future IT health care service that would be offered by an open innovation driven health care IT company. This IT product (service) would collect all the individual medical and non-medical data it could obtain with the permission of the individual and depending on electronics and sensor system technologies. The research question was whether the user would transfer all medical and non-medical data to a future IT service of an open innovation driven health care IT company. Would such a future IT service prototype meet the users' needs and lead to a high percentage of willingness to transfer all data to an open innovation driven health care company that offers this service?

Secondary Data We interviewed 821 patients in a multicenter study in Cologne and Bonn (four dental clinics and six medical practices) from February 2014–February 2015 and asked them about the importance of several factors. Of more than 2439 patients, just 821 met the inclusion criteria. These were: history of dental and medical illness in the past; age 20–75 years; at least one chronic medical disease (e.g. diabetes or coronary heart disease); experienced in using IT; and a positive attitude to IT services.

Data is defined in this study as all data that can be collected in a way that make sense for a holistic health care IT service approach: e.g. food, food preparation, weight, sport, health data and history, stress profile, genetic risks if test available, environment, sleep time and quality, regeneration profile, hygiene profile as well as sun exposition and protection. Connection of various electronic tools is required for this. We asked 67 directors of small and midsized technology companies in Germany and Belgium from the health care industry by e-mail for interviews. Just 17 answered and 8 accepted an interview.

Primary Data For this follow-up study, we chose a multicenter study in two steps. First, we applied a qualitative research method, in which we interviewed eight directors of small and midsized German and Belgian technology companies from the health care sector about their view of future scenarios of technological products and services for patients based on present or future technologies and concepts. At the same time, we interviewed 16 patients with a combined history of dental and medical illness about their expectations of such products in the future and their willingness to transfer their personal data to an open innovation driven health care company.

After clustering the interviews in three main sectors each on the industry and patient side, we designed a prototype IT model and presented it in step 2 of the study to patients who met the inclusion criteria of our study and asked them (quantitative

research method) questions with a standardized question paper, designed on the experiences of the interviews.

Data Analysis The interviews (step 1) were followed by writing the main subjects on paper. Later the main subjects of the interviews were coded. This coding helps to identify patterns and to develop a list of standards from the point of view of the industry and of the user.

The question paper for step 2 was based on the results of the interviews and covered the most important subjects that emerged from the interviews. Finally, the answers of the users to the standardized question paper were analyzed with the statistical software IBM SPSS 22.0.

3.6 Future Health Care IT Service Prototype Model

This future health care IT service prototype model included IT applications available at present in combination with sensor systems technology and electronics available at present. The combination concept, however, is currently not available and is a future technology approach. This future concept allows users (patients), as the legal owner of their data, to transfer all dental, medical and other data that they and the company define as relevant to the health care company (consulting need of the user) in return for an efficient health care service.

The open innovation process allows every single user to see anonymized data of other customers, to participate in research results based on an outcome of the common data pool of this specific user community and to interact directly with the company to communicate user wishes, which can be used to serve as an individualized evolutionary model in a very short time to meet user needs.

3.7 Findings

The results showed that the most important factor for users (patients) at 91.1% ($n = 748$) was the ability to influence the future IT service in health care through an open innovation process. The security of the IT data came second at 89.4% ($n = 734$), followed by the possibility to benefit from the scientific research results based on the data pool of the future IT service community at 86.6% ($n = 711$). If these three important standards of open innovation process, IT security and scientific results from the data pool of the community would be guaranteed, overall 87.8% ($n = 721$) of the patients would transfer their entire medical and non-medical data as mentioned above in return for an efficient service in health care.

3.8 Conclusions of the Study

The results of the study showed that patients fulfilling the inclusion criteria named three important factors as a necessary standard before they would transfer their medical and non-medical data: open innovation process which integrates the user and his thoughts, IT security and a benefit from the data pool (research results) of the users of this service. If these three standards are met, the empirical study showed clearly that there is a high user (patient) willingness to transfer personal medical and non-medical data to a future IT service of open innovation driven health care companies in return for an efficient health care service.

This chapter makes a contribution to what relevance the user is and his willingness to participate in the open innovation process (Von Hippel et al. 1999), what standards are expected in the open innovation process from the users' perspective (Chesbrough and Bogers 2014) and how far companies themselves have to open up in the open innovation process (West 2003) to succeed in the future health care services market.

4 Drone-Supported Emergency Concepts in Combination with Automotive Health Systems

One of the most vital factors when a person suffers a heart attack or stroke is how quickly a physician or qualified help can intervene. The new automotive health concepts and systems will implement and integrate health services, meaning companies like Audi or Bosch could combine their competencies in sensor systems technology and offer this additional service. Depending on where you have your heart attack, it could save your life. Usually the first 15–30 min after a heart attack decides on the prognosis of the patient surviving. This can be a very short time if you are driving in a rural area, where much time can be lost in communication, transportation and finally treatment.

First of all, therefore, sensors in the car must be able to detect a significant problem and interact with the driver if there is a need to start an emergency process or the user himself can call an emergency hotline. If the driver is unable to make the call himself because his condition is getting progressively worse, then the car sends the emergency signal. The big questions are what happens next:

1. Can the car interact with other cars to ask if there is a physician nearby, who could maybe come to help even if the physician has a day off and is not on duty?
2. Is a drone dispatched by a hospital with medical equipment, drugs and technology the fastest way of getting the package to where it is needed? The speed of the fastest drones is currently 200–300 mph. The advantage of a drone is that it can start immediately from the hospital or another place. A helicopter needs preparation time, the pilot must get to the helicopter, it must get starting permission and needs a certain take-off time. Landing zones for helicopters are

also limited. There are, therefore, many arguments speaking for use of drones in the whole process of patient rescue following a heart attack or stroke.

3. If there are individuals nearby or in the car who could help, should the car have drugs and medical equipment on board so that a physician could, via telemedicine, instruct the individuals on what to do until professional help arrives?

There is a combined approach of cars (autonomous driving, emergency status, sensor systems, new medical technology such as the "in-car stethoscope"), drones (speed, weight, autonomous flying), hospitals (telemedicine, drones), physicians, availability of patient data, legal requirements and digital health crisis management that could help save lives in the long run. Further, this kind of holistic approach needs 5G communication and powerful servers and computers, an infrastructure which could possibly be shared by a group of hospitals and car manufacturers.

It is also an option to let a drone land on a fast autonomous vehicle driving to a hospital so that the people in the vehicle with the person suffering a health crisis can help with first steps until arrival at the hospital. There are various new business models for use of drone technology in health care, also for delivery of medications and other health care items (Scott and Scott 2019) or to exchange lab samples as a service innovation for hospitals and clinics (Mion 2019).

5 Conclusions

For investors with venture capital or a mergers and acquisition strategy, it is clear that the health care market is one of the most interesting fields both at present and in the future. The health care market in the G20 countries is growing, people are getting older and the procedures for diagnosis and therapies are improving every year.

The car industry, the sensor systems technology industries, the drone industry, the high speed computer processor industry, the communications industry, BIG DATA companies and the software development sector will lead this future health care market, which will be completely different to anything we have known so far. Even if the big players of Big Data such as Apple, Amazon, Facebook and Microsoft do not seem so strong in this market at present, it is clear that they will also lead this market and maybe become the health care management deliverers of tomorrow.

Ideas such as self-driving hospital beds, companies that guarantee a certain health status if you go "all-in" with your data, and students who participate in contests to innovate health care are scenarios of the new horizons, which will be realized. The question is which countries and companies will dominate this health care market.

Further, it is necessary to consider that the domination of the health care market will have a long-term impact on the technology domination of space travel because, besides the transportation and construction problems, the main challenge in space will be how to keep our human species alive and healthy in this endless universe as we travel to new horizons with new environments that will test endlessly the biology

and mental fitness of humankind. The knowledge for the health care market on earth is the key factor for domination of research and travel in space.

It is not just about infection prevention during human space travel (Mermel 2012), alterations of sympathetic control by space travel (Eckberg 2003) or the mars colonization (Levchenko et al. 2019), it is also about the combination of collecting and analyzing tons of data at high speed on the technology side and healthy and innovative individuals on the human side. This means that without outstanding health care in space travel and planet colonization, there will be no long-term innovation.

Digitalization, transformation as well as technological and pharmaceutical innovations are already a big step for humankind to enable us to deal with the challenges and problems awaiting us in space and the colonization of planets. The health care market on earth today is a part of space travel tomorrow.

References

Ärzteblatt, D. (2019). *Hoher Langzeitkrankenstand bei Pflegekräften*. Retrieved August 26, 2019, from https://www.aerzteblatt.de/nachrichten/89238/Hoher-Langzeitkrankenstand-bei-Pflegekraeften

Baldwin, C., & von Hippel, E. (2011). Modeling a paradigm shift: From producer innovation to user and open collaborative innovation. *Organization Science, 22*(6), 1399–1417.

Bogers, M., Afuah, A., & Bastian, B. (2010). Users as innovators: A review, critique, and future research directions. *Journal of Management, 36*(4), 857–875.

Boudreau, K. J., & Lakhani, K. R. (2013). Using the crowd as an innovation partner. *Harvard Business Review, 91*(4), 60–69.

Bullinger, A. C., et al. (2012). Open innovation in health care: Analysis of an open health platform. *Health Policy, 105*(2), 165–175.

Chesbrough, H. (2006). *Open innovation: The new imperative for creating and profiting from technology*. Brighton, MA: Harvard Business Press.

Chesbrough, H., & Bogers, M. (2014). Explicating open innovation: Clarifying an emerging paradigm for understanding innovation. In *New frontiers in open innovation* (pp. 3–28). Oxford: Oxford University Press.

Eckberg, D. L., & Neurolab Autonomic Nervous System Team. (2003). Bursting into space: Alterations of sympathetic control by space travel. *Acta Physiologica Scandinavica, 177*(3), 299–311.

Estrin, D., & Sim, I. (2010). Open mHealth architecture: An engine for health care innovation. *Science, 330*(6005), 759–760.

Harhoff, D., & Mayrhofer, P. (2010). Managing user communities and hybrid innovation processes: Concepts and design implications. *Organizational Dynamics, 39*(2), 137–144.

Herstatt, C., & von Hippel, E. (1992). From experience: Developing new product concepts via the lead user method: A case study in a "low-tech" field. *Journal of Product Innovation Management, 9*(3), 213–221.

Kanto, L., et al. (2014). How do customer and user understanding, the use of prototypes and distributed collaboration support rapid innovation activities? In *Management of Engineering and Technology (PICMET), 2014 Portland International Conference*. IEEE.

Lakhani, K. R., Boudreau, K. J., Loh, P. R., et al. (2013). Prize-based contests can provide solutions to computational biology problems. *Nature Biotechnology, 31*(2), 108.

Lettl, C., Herstatt, C., & Gemuenden, H. G. (2006). Users' contributions to radical innovation: Evidence from four cases in the field of medical equipment technology. *R&D Management, 36*(3), 251–272.

Levchenko, I., Xu, S., Mazouffre, S., Keidar, M., & Bazaka, K. (2019). Mars colonization: Beyond getting there. *Global Challenges, 3*(1), 1800062.

Mermel, L. A. (2012). Infection prevention and control during prolonged human space travel. *Clinical Infectious Diseases, 56*(1), 123–130.

Mion, F. U. (2019). Flying drones to exchange lab samples: Service innovation by the Swiss Multisite Hospital EOC. In *Service design and service thinking in healthcare and hospital management* (pp. 463–479). Cham: Springer.

Parmentier, G., & Mangematin, V. (2014). Orchestrating innovation with user communities in the creative industries. *Technological Forecasting and Social Change, 83*, 40–53.

Reichwald, R., & Piller, F. (2007). Open innovation: Customers as partners in the innovation process. Retrieved August 26, 2019, from http://www.impulse.de/downloads/open_innovation.pdf

Scott, J. E., & Scott, C. H. (2019). Models for drone delivery of medications and other healthcare items. In *Unmanned aerial vehicles: Breakthroughs in research and practice* (pp. 376–392). Pennsylvania: IGI Global.

Steen, M., Buijs, J., & Williams, D. (2014). The role of scenarios and demonstrators in promoting shared understanding in innovation projects. *International Journal of Innovation and Technology Management, 11*(01), 1440001.

Van de Vrande, V., et al. (2009). Open innovation in SMEs: Trends, motives and management challenges. *Technovation, 29*(6), 423–437.

Von Hippel, E. (2005). Democratizing innovation: The evolving phenomenon of user innovation. *Journal für Betriebswirtschaft, 55*(1), 63–78.

Von Hippel, E., Thomke, S., & Sonnack, M. (1999). Creating breakthroughs at 3M. *Harvard Business Review, 77*, 47–57.

West, J. (2003). How open is open enough? Melding proprietary and open source platform strategies. *Research Policy, 32*(7), 1259–1285.

Unlocking the Power of Artificial Intelligence for Your Business

Patrick Glauner

Abstract

Every day, we deal dozens of times with artificial intelligence applications such as autonomous cars, spam filters, or voice recognition systems. Often we do so without explicitly knowing that those applications are based on AI because AI technology has become mainstream in the last 10 years. From a business perspective, AI enables us to automate human decision making. We can thus cut costs and waiting times as well as increase revenue and profit margin. However, we have just started to scratch the surface as there are many more AI opportunities for companies to be exploited. This chapter first provides a gentle, intuitive introduction to AI for the interested business decision maker. In the second part, this chapter provides advice and best practices on how to rethink your business in order to become an AI-driven business that prospers in an ever more competitive environment.

1 Introduction

There is not a day that passes on which we do not hear news about artificial intelligence (AI): autonomous cars, spam filters, Siri, chess computers, killer robots, and much more. What exactly is AI? I have been in and around AI for about 10 years. What does AI mean to me? Humans make decisions dozens of times an hour such as when we have a coffee break, picking a marketing strategy, or whether to buy from vendor A or B. In essence, humans are great in making a lot of very different decisions. While we have seen automation of repetitive tasks in industry

P. Glauner (✉)
Deggendorf Institute of Technology, Deggendorf, Germany
e-mail: patrick@glauner.info

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_5

for about the last 200 years, we had not experienced automation of multifaceted decision making. That is exactly what AI aims at.

In my view, a simple definition of AI would therefore be:

> AI enables us to automate human decision making.

In contrast, Peter Norvig, Research Director of Google, describes AI as follows:

> AI is the science of knowing what to do when you don't know what to do. (Computer History Museum and KQED television 2016)

Admittedly, at first glance, that description is somewhat confusing, but secondarily reasonable: The goal of AI is to solve complex problems that are often associated with uncertainty.

I have led various AI transformation projects, in particular in utilities (Glauner 2019) and mechanical engineering. In this chapter, I share my experience and best practices with you. The first part of this chapter enables you to understand the fundamentals of AI. In the second part, I will show you how you need to rethink your business in order to become an AI-driven business that increases revenue and margin and reduces costs and waiting times.

This book has a number of other chapters that address the field of artificial intelligence. You can learn more about some of my work in the chapter by Thurner and Glauner and in the chapter by Trestioreanu et al., which discuss advances in mechanical engineering and medical technology, respectively, using AI.

2 Motivation: China Is Spearheading AI Innovation

You may wonder whether you should actually invest in AI so soon. Probably your business is going very well at present time. On top of that, there may be a limited number of competitors that so far have not been able to outrank you. All of that may be true—today. In the coming years, however, completely new competitors will emerge. Most likely, they will be based in China. I often feel that most people in the Western world, including decision makers, see China mainly as an export market or a place for cheap labor. In the last couple of years, however, and unnoticed by most Westerners, China has become the world's leading country in AI innovation. You can learn more about China's AI innovation ecosystem and its strong support from both the government and industry in Kai-Fu Lee's book "*AI Superpowers: China, Silicon Valley, and the New World Order*" (Lee 2018). Lee's book is both, encouraging and shocking in my opinion.

How Quickly Is China Innovating in AI?
Let me tell you more about my own experience. I travel to Shanghai at least once a year. I kept noticing an old factory in the Yangpu district. It seemed to have been closed down a long time ago and the land appeared unused. Every single year I passed by, nothing had changed. In 2017, however, the factory was suddenly gone. Furthermore, the factory was not only teared down, the entire land has been turned into an AI innovation hub named "Changyang Campus." The office space also already seemed to be entirely taken, predominantly by start-ups. All of that had happened in less than 12 months! Imagine how many years it even takes in the Western world in order to tear a factory down and get a new construction permit.

In my opinion, we need to radically rethink innovation and agility in the Western world in order to remain competitive. AI's ability to automate human decision will play a crucial role in the future of nearly every company's value chain, be it in research and development, procurement, pricing, marketing, or sales, just to name a few parts. Therefore, the companies that invest in AI early on will be the leaders of their sector in the coming decades. Those that do not invest now are likely to be put out of business by a new AI-driven competitor. After I share the insights of Lee's book and my own experience, I typically manage decision makers to rethink their business and how AI can help them to remain competitive in the long term. Take some time to read Lee's book, it will be a truly rewarding experience.

3 Artificial Intelligence

This section provides a brief introduction to the field of artificial intelligence, its history, recent advances, and most relevant questions for its future.

3.1 History

The first theoretical foundations of AI were laid in the mid-twentieth century, especially in the works of British mathematician Alan Turing (Turing 1950). The actual year of birth of AI is the year 1956, in which the 6-week conference Summer Research Project on Artificial Intelligence at Dartmouth College took place. For that purpose, an application for funding was made in the previous year. The research questions contained therein proved to be indicative of many of the long-term research goals of AI (McCarthy et al. 1955). The conference was organized by John McCarthy and was attended by other well-known scientists such as Marvin Minsky, Nathan Rochester, and Claude Shannon.

Over the following decades, much of AI research has been divided into two diametrically different areas: expert systems and machine learning. **Expert systems** comprise rule-based descriptions of knowledge and make predictions or decisions based on input/data. In contrast, **machine learning** is based on recognizing patterns in training data.

Over the past decades, a large number of innovative and value-adding applications have emerged, often resulting from AI research results. Autonomously driving cars, speech recognition, and autonomous trading systems, for example. Nonetheless, there have been many setbacks. These were usually caused by too high and then unfulfilled expectations. In that context, the term of an "AI winter" has been coined, with which periods of major setbacks in recent decades, the loss of optimism and consequent cuts in funding are referred to. Of course, this section can only provide an overview of the history of AI. The interested reader is referred to a detailed discussion in (Russell and Norvig 2009).

3.2 Machine Learning

The principle of machine learning is further outlined in Definitions 1 and 2.

Definition 1 "[Machine learning is the] field of study that gives computers the ability to learn without being explicitly programmed." (Samuel 1959)

Definition 2 "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 1997)

Concretely, a machine learning algorithm finds ("learns") patterns from examples. These patterns are then used to make decisions based on inputs. Both, expert systems and machine learning, have their respective advantages and disadvantages: Expert systems, on the one hand, have the advantage that they are understandable and interpretable and that their decisions are therefore comprehensible. On the other hand, it often takes a great deal of effort, or sometimes it even turns out to be impossible to understand and describe complex problems in detail.

Example 1 Machine Translation To illustrate this difficulty, an example of machine translation, the automatic translation from one language to another, is very helpful: First, languages consist of a complex set of words and grammar that are difficult to describe in a mathematical form. Second, one does not necessarily use languages correctly, which can cause inaccuracies and ambiguities. Third, languages are dynamic as they change over decades and centuries. Creating an expert system for machine translation is thus a challenge. The three factors of complexity, inaccuracy, and dynamics occur in a variety of fields and prove to be a common limiting factor when building expert systems.

Machine learning has the advantage that often less knowledge about a problem is needed as the algorithms learn patterns from data. This process is often referred to as "training" an AI. In contrast to expert systems, however, machine learning often leads to a black box whose decisions are often neither explainable nor interpretable. Nonetheless, over the decades, machine learning has gained popularity and largely replaced expert systems.

3.3 The Three Pillars of Machine Learning

The field of machine learning can broadly be separated into three so-called pillars that are depicted in Fig. 1.

An interconnection can be made between each pillar and human learning: Imagine when you were a kid, you walked through the park with your parents. Your parents then pointed at various animals, say a cat, a dog, and a bird. You perceived the visual and audio signals from your eyes and ears, respectively. In addition, you got an explanation of what type of animal you were seeing. That pillar is called **supervised learning**, in which you get an explicit explanation or "label." Mathematically speaking, supervised learning uses pairs $\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})\}$, where $\mathbf{x}^{(i)}$ is the input and $y^{(i)}$ the label, respectively. The goal is to learn a function f: $y^{(i)} = f(\mathbf{x}^{(i)})$ that infers the label from the input. This is also called function induction, because rules from examples are derived. In any case, the labels y give an unambiguous "right answer" for the inputs $\mathbf{x}$.

When you continued your walk through the park, you perceived more cats, dogs, and birds of different colors and sizes. However, that time you did not get any supervision from your parents. Instead, you intuitively learned how to distinguish cats, dogs, and birds regardless of their individual attributes. That is an example of **unsupervised learning**, which aims to find hidden structures in unlabeled data $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\}$.

In many problems, it is essentially impossible to provide an explicit supervision to a learning problem. In **reinforcement learning**, we mainly think in terms of states, actions, transition between states and rewards, or penalties you get subject to your performance. That is how humans actually learn most of the time. One great example of how humans learn in a reinforced way is riding a bicycle. It is awfully difficult to explain someone else how to ride a bicycle. Instead, as kids, we tried

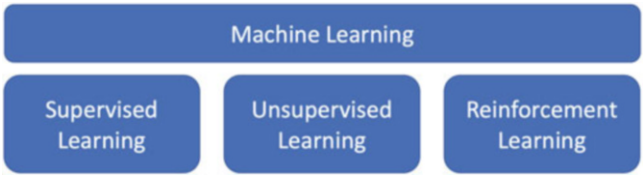

Fig. 1 The three pillars of machine learning. Source: author

out how to ride it. If we did the wrong moves, we got hurt. Concretely, we are in different states and tried to find the right transitions between states in order to remain on the bicycle.

3.4 Neural Networks

Of particular historical significance are so-called (artificial) neural networks. These are loosely inspired by the human brain and consist of several layers of units—also called "neurons." An example of a neural network is shown in Fig. 2.

This type of neural network falls into the category of supervised learning. The first layer (on the left) is used to enter data and the last layer (on the right) to output labels. Between these two layers are zero to several hidden layers, which contribute to the decision making. Neural networks have experienced several popularity phases over the past 60 years, which are explained in detail in (Deng and Yu 2014). In addition to neural networks, there are a variety of other methods of machine learning, such as decision trees, support vector machines, or regression models, which are discussed in detail in (Bishop 2006).

3.5 Recent Advances and Deep Learning

Although AI research has been conducted for over 60 years, many people first heard of AI just a few years ago. This, in addition to the "Terminator" movie series, is largely due to the huge advances made by AI applications over the past few years. Since 2006, there have been a number of significant advances, especially in the field of neural networks, which are now referred to as deep learning (Hinton et al. 2006). This term aims to ensure that (deep) neural networks have many hidden layers. This type of architecture has proven to be particularly helpful in detecting hidden relationships in input data. Although this was already the case in the 1980s, there was a lack of practical and applicable algorithms for training these networks from data first and, secondly, the lack of adequate computing resources. However,

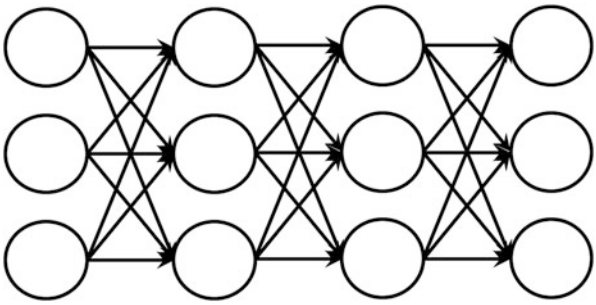

Fig. 2 Neural network. Source: author

today there is much more powerful computing infrastructure available. In addition, significantly better algorithms for training this type of neural network have been available since 2006 (Hinton et al. 2006).

As a result, many advances in AI research have been made, some of which are based on deep learning. Examples are autonomously driving cars or the computer program AlphaGo. Go is a board game that is especially popular in Southeast Asia, where players have a much greater number of possible moves than in chess. Traditional methods, with which, for example, the IBM program Deep Blue had beaten the then world chess champion Garry Kasparov in 1997, do not scale to the game of Go, since the mere increase of computing capacity is not sufficient due to the high complexity of this problem. It was only until a few years ago the prevailing opinion within the AI community that an AI, which plays Go on world level, was still decades away. The UK company Google DeepMind unexpectedly revealed their AI AlphaGo to the public in 2015. AlphaGo beat South Korean professional Go play Lee Sedol under tournament conditions (Silver et al. 2016). This success was partly based on deep learning and led to an increased awareness of AI world-wide. Of course, in addition to the current breakthroughs of AI mentioned in this section, there have been a lot of further success stories and we are sure that more will follow soon.

3.6 Frontiers

We would now like to discuss some current issues concerning AI. While many recent accomplishments are based in part on deep learning, this new kind of neural network is only one of many modern techniques. It is becoming increasingly apparent that there is a hype about deep learning and more and more unrealistic promises are being made about it (Dacrema et al. 2019; LeCun et al. 2015). It is therefore essential to relate the successes of deep learning and its fundamental limitations. The "no free lunch theorem," which is largely unknown both in industry and academia, states that all methods of machine learning averaged over all possible problems are equally successful (Wolpert 1996). Of course, some methods are better suited to some problems than others, but perform worse on different problems. Deep learning is especially useful for image, audio, video, or text processing problems and when having a lot of training data. By contrast, deep learning, for example, is poorly suited to problems with a small amount of training data.

We have previously introduced the notion of an AI winter—a period of great setbacks, the loss of optimism, and consequent reductions in funding. It is to be feared that the current and hype-based promise could trigger a new AI winter. It is therefore essential to better understand deep learning and its potential and not neglect other research methods. A major limitation of deep learning—and neural networks in general—is that these are black box models. As a consequence, the decisions made by them are often incomprehensible. Some advances have been made in this area recently, such as local interpretable model-agnostic explanations (LIME) (Ribeiro et al. 2016) for supervised models. However, there is still great

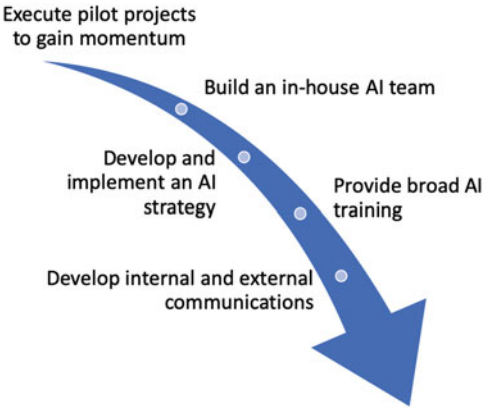

Fig. 3 Steps of an AI transformation. Source: author

research potential in this direction, as future advances may also likely increase the social acceptance of AI. For example, in the case of autonomously driving cars, the decisions taken by an AI should also be comprehensible for legal as well as software quality reasons.

4 AI Transformation of a Company

AI projects typically address the value chain of companies or the company's products. Which one should you focus on in the beginning? Employing AI to optimize your value chain usually leads to a reduction of waiting times and costs and allows to re-allocate resources within your company. Typically, you can optimize parts of your value chain quickly because you have access to your internal databases and processes. Optimizing your value chain allows to generate the budget needed for shipping AI through next generation products to customers, which typically takes longer. Once you ship AI-based products to your customers, you can further increase your revenue.

Andrew Ng, one of the leading scientists in the field of machine learning, suggests five major steps[1] that compose a successful AI transformation (Ng 2018). These are depicted in Fig. 3.

Most successful AI transformations typically follow those steps, yet they need to be tailored to a company's requirements and structure. In the beginning, it is important to execute pilot projects in order to gain momentum. When choosing pilot projects, it is helpful to automate tasks that have previously not been able to do or that turned out to be lengthy and expensive.

[1] http://landing.ai/ai-transformation-playbook.

Best Practice 1: Hire a Chief Digital Officer That Is Independent from Your Chief Information Officer

An AI transformation of a company is part of the company's digital transformation. Any corporate AI strategy needs to align to the corporate digitalization strategy. Who is in charge of digitalization? Most companies have a Chief Information Officer (CIO) who heads the company's IT department and defines the corporate IT strategy. IT departments are usually very conservative and seem to favor the status quo instead of innovation. That makes sense as the most important duty of an IT department is to provide key services such as internet access, emails, enterprise resource planning systems, data storage, or phone services as reliable as possible. Focusing on reliability typically comes with a limited user experience. For example, users may not be able to install any third-party software on their computers or browse random websites. All of that makes sense from a reliability and IT security point of view.

In contrast, in order to take full advantage of modern digitalization technology, your digitalization specialists need more freedom and must be able to try out new tools, programming languages, frameworks, and cloud services. A CIO usually does not have that sort of mindset. Successful corporate digital transformations typically have one thing in common: There is a **Chief Digital Officer (CDO)** who runs the company's digital department and does not report to the CIO. CDOs typically report to a board member[2] and thus have the autonomy to run a department that utilizes its own IT infrastructure and is able to develop new products and services quickly. The separation of and interaction between those two departments is depicted in Fig. 4.

Most companies do not only have a central data infrastructure. Instead, data is typically spread all over the company in many different locations and formats (e.g., spreadsheets, databases, etc.). This makes it difficult for machine learning projects to take full advantage of the data of company. Therefore, one of the CDO's responsibilities includes to establish a central and harmonized data storage that can be used in subsequent projects that aim to get insights from the entire corporate data.

Who actually implements the AI applications used in your company? On the one hand, you can buy solutions from vendors or consulting companies. This will mainly require a one-time investment. However, integration into your existing infrastructure and processes as well as maintenance will usually become a challenge if you fully rely on external partners.

[2]With digitalization and AI becoming ever more crucial to the success of a company, more companies will likely appoint CDOs directly to their board in the future.

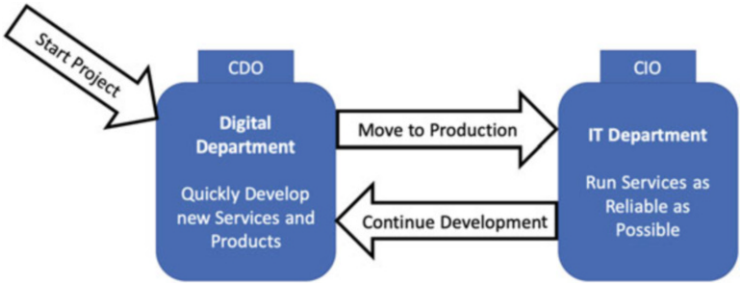

Fig. 4 Separation, responsibilities of and interactions between a Chief Digital Officer (CDO) and a Chief Information Officer (CIO). Source: author

Best Practice 2: Establish an In-House AI Competence Center
Ideally, you establish not only an in-house AI team but also an in-house AI competence center. The competence center spearheads the AI transformation of your company. In order to do so, it should employ a number of different talents:

- AI experts such as data scientists who implement AI models.
- Software engineers who are responsible for software architecture and software quality.
- Data architects who build large-scale data infrastructure.

The competence center should be run by **AI innovation managers**. What does a regular innovation manager do? Most innovation managers in industry have mainly a business background and analyze requirements and business opportunities with customers and partners. They then draft concepts in slides. Yet, in reality, most of those concepts are never turned into reality due to the discrepancy between the innovation managers' business background and the real-world technological environment.

AI innovation managers have a background in both, AI and business. AI projects need to have a positive return on investment. AI innovation managers must come up with feasible and sustainable AI business cases. AI innovation managers also have to be self-critical of the field of AI. Thus, they should strive to build solutions that are as simple as possible and also use alternatives to AI when necessary. In many cases, simple if-else rules or simple control loops may be sufficient.

Hiring AI innovation managers may be challenging as they are a scarce human resource and typically draw big salaries. However, that is a necessary investment as your AI competence center will only succeed if it is run by managers that have experience in both, AI and business.

Table 1 Proportional time investment of a machine learning project

Step	Books and courses	Reality
Defining KPIs	Small	Large
Collecting data	Small	Large
Exploratory data analysis	Small	Large
Building infrastructure	Small	Large
Optimizing ML algorithms	Large	Medium
Integration	Small	Large

Best Practice 3: Use Open Source
When choosing the underlying libraries of your AI applications, you have to decide between commercial and open source software. Making that choice is crucial in terms of quality, reliability, costs, and support. The most popular AI frameworks and libraries are open source and thus do not come with any license fee. `TensorFlow` (Abadi et al. 2016) or `scikit-learn` (Pedregosa et al. 2011) is among the most popular libraries that are the foundation of plenty of commercial AI applications. Those libraries have a strong community behind and come with high software quality. For example, `TensorFlow` was initially developed by Google for their in-house AI applications. Later, `TensorFlow` was released to the public and Google maintains an active role in the development of this library. Most students who study AI have used those libraries during their studies. Therefore, recruiting experts who know how to use those libraries is also much easier than hiring experts that know how to use some rarely used commercial AI software.

Best Practice 4: Successful Machine Learning Projects Do More Than Just Learning Patterns
When reading machine learning books, e.g., (Bishop 2006), or taking a course, you notice a strong focus on learning patterns/optimizing machine learning models. However, there are other necessary steps that make a machine learning projects successful. Most books or courses have a strong focus on academia and thus largely ignore those other aspects as depicted in Table 1.

However, the reality is very different. As the other steps often get ignored in books or courses, a large number of industrial machine learning projects actually fail. Bearing the following points in mind will help to successfully execute machine learning projects:

(continued)

1. Define the purpose of the project and define KPIs that the project would improve.
2. Collect your data from different sources and aggregate the data.
3. The quality of the data plays a crucial role for the success of a machine learning project. I often refer to machine learning as "*garbage in, garbage out*." That is why you need to perform a process called **exploratory data analysis** in which you not only check for missing values but also you rather perform a much deeper quality check approach. That approach includes checking distributions of your data and finding anomalies. In this step, you need to include domain experts that can tell you whether the data matches the reality of their domain.
4. Build the necessary computing infrastructure for training and running your machine learning models.
5. Train the actual machine learning models. However, you do not need to be an expert in the hundreds of different models reported in the literature. In practice, you can use **automated machine learning (AutoML)** libraries such as `auto-sklearn` (Feurer et al. 2015). AutoML libraries evaluate a large number of possible models for your data set and then choose the one that performs the best.
6. Integrate your machine learning models into a legacy IT infrastructure. That work is often less interesting, yet challenging. Often it turns out to be useful to provide machine learning models through web services that can quickly be integrated into existing code without significantly amending the code base of the legacy systems.

Best Practice 5: Everyone in Your Company Needs Some Understanding of Digitalization and AI

Once you have established your in-house competence center, you need everyone in your company to learn about it and what its staff do. I tell decision makers all the time:

> Everyone, from the board members down to the factory workers, needs to acquire some understanding of what AI is.

That understanding will help every employee to identify tasks that can be automated using AI. For example, a factory worker may see an opportunity for how to improve a process using data that was previously collected. Their supervisor also needs some understanding of how AI works in order to assess their worker's proposal. If the assessment is positive, they can forward it to their supervisor or directly to the AI competence center. Also the top

(continued)

management of a company need to be aware of what AI is and how it can improve the company. That understanding helps the top management to challenge the status quo and to assign resources accordingly to innovation. Training your staff is cheap as there are plenty of free massive open online courses (MOOCs) available. Popular MOOC platforms include Coursera,[3] Udacity[4], and edX.[5]

5 The Fear of an Out-of-Control AI Is Exaggerated

When looking at the rapid progress of AI, the question arises as to how the field of AI will evolve in the long term, whether 1 day an AI will exceed the intelligence of a human being and thus potentially could make mankind redundant. The point of time when computers become more intelligent than humans is referred to in the literature as the *technological singularity* (Shanahan 2015). There are various predictions as to when—or even if at all—the singularity will occur. They span a wide range, from a period in the next 20 years, to predictions that are realistic about achieving the singularity around the end of the twenty-first century, to the prediction that the technological singularity may never materialize. Since each of these predictions makes various assumptions, a reliable assessment is difficult to make. Overall, today it is impossible to predict how far away the singularity is. The interested reader is referred to a first-class and extensive analysis on this topic and a discussion of the consequences of the technological singularity in (Shanahan 2015).

In recent years, various stakeholders have warned about so-called killer robots as a possible unfortunate outcome of AI advances. What about that danger? Andrew Ng has set a much-noticed comparison (Williams 2015): Ng's view is that science is still very far away from the potential killer robot threat scenario. In his opinion, the state of the art of AI can be compared to a planned manned trip to Mars, which is currently being prepared by researchers. Ng further states that some researchers are also already thinking about how to colonize Mars in the long term, but no researcher has yet tried to explore how to prevent overpopulation on Mars. Ng equates the scenario of overpopulation with the scenario of a killer robot threat. That danger would also be so far into the future that he was simply not able to work productively to prevent it at the moment, as he first had to do much more fundamental work in AI research. Ng also points to potential job losses as a much more tangible threat to people by AI in the near future.

[3] http://www.coursera.org.

[4] http://www.udacity.com.

[5] http://www.edx.org.

6 Conclusions

The first part of this chapter provides a gentle introduction to the field of artificial intelligence, its history, recent advances, and most relevant questions for its future. The center of the world-wide AI innovation has been established in China. In the coming years, Western companies of any sector will face an unparalleled level of competition because their Chinese competitors will be AI-driven and thus increasingly automate human decision making in software and hardware. As a consequence, Western companies should invest in AI as soon as possible in order to remain competitive. The second part of this chapter provides the necessary steps for an AI transformation of a company: executing pilot projects, building an in-house AI competence center, defining and implementing an AI strategy, providing broad AI training, and developing internal and external communications. This chapter also provides a number of best practices that will help to turn an AI transformation into a success.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., 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).

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Computer History Museum and KQED Television. (2016). *CHM Revolutionaries: The Challenge & Promise of Artificial Intelligence*. http://www.youtube.com/watch?v=rtmQ3xlt-4A [Online]. Accessed 18 July 2018.

Dacrema, M. F., Cremonesi, P., & Jannach, D. (2019). Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys 2019)*.

Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and Trends in Signal Processing, 7*(3–4), 197–387.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28* (pp. 2962–2970). Red Hook: Curran Associates, Inc.

Glauner, P. (2019). Artificial Intelligence for the Detection of Electricity Theft and Irregular Power Usage in Emerging Markets. Ph.D. Thesis, University of Luxembourg, Luxembourg.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation, 18*(7), 1527–1554.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436.

Lee, K.-F. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. Boston: Houghton Mifflin Harcourt.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.

Ng, A. (2018). *AI transformation playbook: how to lead your company into the AI era*. Landing AI. https://landing.ai/ai-transformation-playbook/ [Online]. Accessed 18 July 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). New York: ACM.

Russell, S. J., & Norvig, P. (2009). *Artificial intelligence: a modern approach* (3rd ed.). Englewood Cliffs: Prentice Hall.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210–229.

Shanahan, M. (2015). *The technological singularity*. Cambridge: MIT Press.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature, 529*(7587), 484.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Williams, C. (2015). *AI guru Ng: fearing a rise of killer robots is like worrying about overpopulation on Mars*. Situation Publishing, London. http://www.theregister.co.uk/2015/03/19/andrew_ng_baidu_ai/ [Online]. Accessed 1 Aug 2018.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation, 8*(7), 1341–1390.

Innovation Means: Asking the Right Questions

Oliver Bludau

Abstract

In your professional career as an employee of a company which is not like the typical start-up, you have probably been facing several "innovation projects" every now and then. Projects which have the goal to reinvent the company, to create new added value to stay ahead of competition. Innovation seems to be the apparent answer to a slowdown of the growth of a company. But how many initiatives were successful in the end? Not many I guess. One of the reasons is the wrong approach that management starts off with. Instead of beginning with a question, it starts with the answer. The answer to what? Learn with how little effort you can increase the success rate of innovation projects significantly by asking the right questions.

1 Introduction

During my 25-year career I founded or led companies in various sizes and industries. I was an officer in the German air force, founded an investment broker company in the 1990s. I later sold it to the credit card company American Express where I served as a member of the board for several years. I was managing director of Germany's largest pure show and event park, took acting classes, and played a main

The author "Oliver Bludau" was deceased at the time of publication.

O. Bludau (✉)
Innovators Institute, Cologne, Germany

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_6

role full-time on stage for 4 years in front of around 750,000 people over all. In 2006 I switched industries again and took over a family-owned machining business. I developed it from an average small job shop with 2 mio. Euro in revenue and 20 employees in 13 years to a first tier supply chain partner for the semiconductor, aerospace, and mechanical engineering industries with 60 mio. Euro in revenue and about 250 employees in 3 countries. I am a proud Harvard Business School alumnus and studied Artificial Intelligence at the MIT in Boston.

One main factor for the success of my companies has been the consequent focus on sustainable innovation and life-long learning—regardless of industry, company size, economic climate, or political insecurity.

Of course I had defeats and throwbacks. Innovations that did not go out well and my company lost lots of money. Sometimes I was just too naïve, sometimes too optimistic, sometimes I bet on the wrong horse, sometimes I planned not well enough, and sometimes I just had the totally wrong timing. Those setbacks hurt and gnaw on your self-confidence. You start doubting and it took a lot of courage to start the next innovation adventure when the last one was a total failure. While failed innovation in large corporations may lead to the end of your personal career in that company, failed innovation in an SME may threaten the whole company, all employees, and their families. But once you are infected with the innovation virus you just cannot stop to always think about the next mountain you will climb. Putting it all together for me it worked in every company, in every business and I kept getting better and better from one innovation journey to the next.

What I learned over the years is that innovating does not have to be that risky, you do not have to reinvent the world again. Your responsibility as a manager is just to do something at all. What you need is an innovation culture, an innovation strategy, and an innovation method. In this chapter I would like to convince you of becoming an innovation enthusiast and I will give you some really good stuff how you can become one. This is not theory, this is not consultant talk, this is pure experience from one entrepreneur to another, from one manager to another, from one guy who has the duty to secure the future of his company to another.

2 So Let Us Innovate

When I work with companies on their attitude towards innovation I mostly start with a meeting of the 10 most influential people in the company. That can be top level managers, but also people from the informal communication network in a company—the administration lady who has great connections to everybody, the production worker which is most admired because of his expertise, the opinion leader among the employees, although he has no leadership responsibility.

The first question I ask is quite rhetorical: “Who of you thinks that innovation has to be a top priority for the sustainability and future viability of your company?” Guess the answer—it is always a 100% agreement or at least very close to that. Of course I ask this question only to prepare the next one: “Well, great, then let us start. You should start innovating right now. What will be your first step then?” Suddenly

it gets quiet in the room and people turn their head away elegantly to just avoid being cold called on this.

People are faced with that buzzword "innovation" every single day and start nodding their heads automatically when the importance is questioned. What would your boss think of you if you said "No, we do not need innovation. It is all fine as it is and it will be like that for the next 100 years." Guess, you should be prepared to make your next career step maybe in a different company.

3 What Is Innovation Anyway?

When I ask the participants of the meeting what comes to their mind when they hear the term "innovation" the problem becomes obvious. While innovation for one person is purely incremental and only related to a physical product of a company (e.g., efficiency) the other person claims that this is development, not innovation. For him innovation starts with new services, new organization forms, new management structures, new business models, radical or disruptive technologies, or dealing with future megatrends like "Gender Shift," "New Work," "Silver Society," "E-Mobility," "Urbanization," etc.—everything but definitely not incremental improvements.

4 Do You Really Need to Constantly Innovate?

There is no progress without innovation. And without progress there is no future—in general and for your company. If you are successful in what you are doing, others are eager to take a share from your core market. The paradox is that the more successful you are the more you need to innovate, because your rivals do not even just imitate you. Their only goal is to make your offer more efficient, easier, cheaper, or better in order to take market share from you. You have to innovate to differentiate yourself. You are the hunted one! But are you fast enough?

5 Where Is Your Game Plan?

A company is like a sports team. The coaching team consists of several people responsible for several tasks, e.g., a team manager, a tactical coach, a fitness coach, a psychologist, a head coach, and maybe even more. They all have the same goal, they want the team to win, and they align their activities on the purpose with the agreed measures to do so. Imagine there was only a head coach who tells the team: "I think our goal is clear. We want to win. So do it please." Probably this team would not get very far. Players might be in different physical shape. While one has played in the Champions League, the other might have played in a lower class. Now the coach expects the same performance from both. Players do not even know what their exact task is and how they should behave. Do they play offense or defense, when should they leave their position, when should they stay? Who should they

back up in case of an opponent's breakthrough. I know—this story sounds absurd to you. But actually this is how a lot of companies still manage their innovation activities.

Does it sound familiar to you when the boss says: "I think our goal is clear. We need to innovate. So do it please." Mostly some dedicated group of employees—more or less experienced in managing innovation—is expected to arrive at promising innovations. But what? And how? Should they just grab into the chest of "Kaizen ideas," pick one by random and work on it? Should they react on market demand and make up some new fancy ideas? How can they know that this is the right approach? How do they know that their idea is the top choice? What are the expectations of the members of the innovation team towards innovation? What is the expectation of the company itself?

As long as innovation is just a placeholder for anything and everything "new" in a company, you will not be successful in your innovation activities. Some of the companies I work with already got aware of this and renamed the position "Head of Innovation" back to "Head of Product Development" as it used to be called years ago.

6 Innovation Is Hard

Innovation is not invention, nor is it an idea. Innovation is the commercially successful implementation of "something" in the market. And it is not only about the innovation itself that is hard. It is even much harder to align the whole organization on innovation. An organization of individuals, where everybody has his own agenda. An organization who went in one direction for the last 30 years and that I now want to completely turn around into a different direction with a new mindset. Concretely, Professor Gary P. Pisano from Harvard Business School said that "If innovation were easy, everyone could do it, and then innovation could no longer be a source of advantage" (Pisano 2019).

7 Innovation Is Not Rocket Science Though

The good news is that—contrary to popular opinion—every company, every organization, and every person can innovate. Large corporations excuse their failure in innovation by their inertia, myopia, and bureaucracy; small companies excuse their inactivity by a lack of money or the stress from daily business. And if nothing is reasonable it is the fear of a "could be" recession which stops them. All this is nonsense. Innovation in most companies shows symptoms of an inability to keep up the pace in innovating which they themselves had been set in the time of their birth. But it is not the industry, company size, or situation that matters, but the management practice and leadership. Innovation is a combination of strategy, organization, and culture, shaped by leadership. All what you find in this book is not much worth if you do not have the supporting culture in your company,

the commitment of the leadership, the methods to implement, and the strategy for sustained innovation. But if you have them, you are able to skyrocket your innovations and your company. And how you do this, I will show you in this chapter.

8 It All Starts with a Question

Kids are great questioners. Why is the sky blue? Why can I not touch the air? Why do my soccer shoes stink like cheese? All children are curious and eager to understand their world. And their most important tools are questions. Brandy Frazier from the University of Michigan has worked with her team on how and why children ask so many questions. She found out that children never want to annoy their parents with their questions, but that they are really interested in all this. Even the infamous question "And why?" is not because they did not understand something—they ask if they were not satisfied with the first answer, or think they need to learn more to see through the whole subject. Unfortunately, our school system is structured in such a way that children forget to ask smart questions. It completely focuses on answers and information.

The Right Question Institute,[1] an NGO that deals with clever questions, has found that just a few weeks after school children ask far fewer questions than before. They are bombarded with fixed answers until their hunger for new knowledge fades. Every exam is only about giving the correct answer. Nobody has the time and desire to research the "why." At the latest when we accept our first job, we have all forgotten to ask the right questions and to be as curious as a child. Asking questions in a business environment is often considered as a weakness. The guy with all the answers is the hero. Especially top managers are often too vain to show vulnerability by asking questions. Instead they just claim some half-truth. A strong claim is better than a weak evidence. That is a pity, because the older we get, the more we could do with clever questions.

Think of your brainstorming sessions. How do they work. Normally it goes like this: "Oh, we could do this and that," "Here is my idea." and "We should try this!" Dozens of ideas lie on the table. So far, so good. What a creative bunch of people. But what you might forget is: Are those ideas really the answers to relevant questions for the future viability of your company? Would it not make much more sense to brainstorm for questions before brainstorming for answers? Questions are a powerful tool for unlocking value in organizations. They support learning and the exchange of ideas, they catalyze innovation and performance improvement, they build rapport and trust among the team members. And they can lower the business risk of doing the completely wrong innovation by uncovering unforeseen pitfalls and hazards.

[1] http://www.rightquestion.org.

9 Innovate by Gut Feeling

There are so many marvelous innovations presented in this book. Every single one justifies an own industry and you might be tempted to jump on one or another. You probably had several touch points with some of them and never even taken a closer look to others. After reading this book you will have a great overview of ideas that you could realize. This is like the brainstorming you are used to. "Oh, interesting, we could do Augmented Reality or let us consider Artificial Intelligence. Maybe we should invest in autonomous driving, and online education is also just a great idea." And you are totally right—they are all great and you could go for any single idea. But does this make sense? First of all, you know that it needs focus. You just cannot go for every approach. No focus means no success. But what should you focus on?

My observations in companies are that a quite frequently used approach for choosing an innovation is "Innovate by gut feeling." A management team or even worse only the CEO judge innovations with their exaggerated self-confidence and feel enlightened to know exactly what the market demands. USD 500,000 later they realize that this horse was not the winning one. Well, who cares, there is another race to bet on. What I would like you to do while reading this book and you are fascinated by an innovation, ask a simple "Why?" question to yourself. Why are you fascinated by the innovation? Why do you think, this is the idea you should follow? Stress test the innovation. It could save you USD 500,000.

9.1 Learn to Ask Questions Again

Fortunately, it is relatively easy to learn to ask clever questions again. Just give it a try. First of all, start with yourself and your closest people. Ask the following question: "Thinking of the future viability of our company what are the most important questions we need to have an answer for?" Take 15 min and let the participants write down as many questions as they have. Do not comment, do not discuss, do not judge. Next let everybody present the questions that are so important for him personally and write them on a board. Use the Eisenhower-Matrix to classify the questions in important and/or urgent (to start). Finally focus only on the questions that most of the participants evaluate as important and urgent for the future of the company. Keep the questions in mind while reading the book and match them with the innovation that is presented to you.

10 The Innovation Canvas

In our work at the Innovators Institute,[2] we have developed a very simple method that will help you to assess which innovation fits best to your company before you invest. It is depicted in Fig. 1. I would recommend to fill out this innovation canvas

[2] http://www.innovators-institute.com.

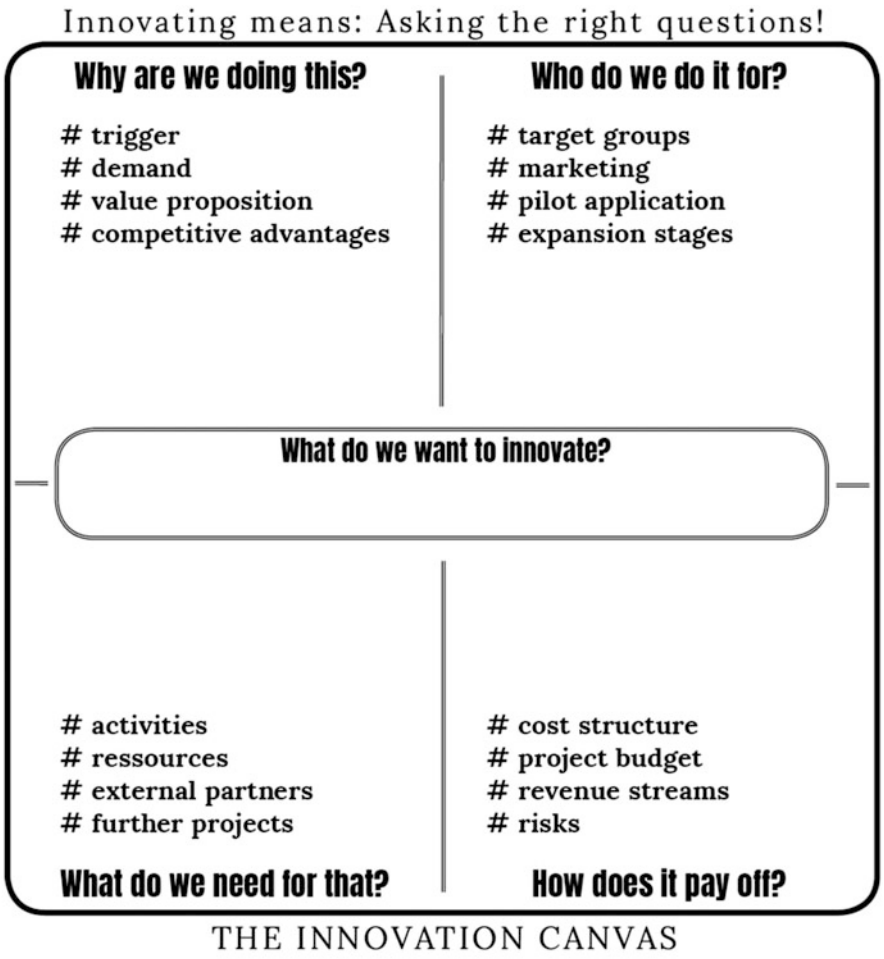

Fig. 1 Innovation canvas. Source: author

for every single chapter. Every part of the innovation canvas is separated in four subtitles which are supportive to get a more detailed look on that innovation itself.

10.1 What Do We Want to Innovate?

Write the title of each chapter in the center of the canvas. The following questions all spin around this. Let us ignite the innovation turbo!

10.2 Why Are We Doing This?

10.2.1 Trigger

As a start in the development of the innovation, it helps to define which triggers led to the idea. Are those rather concrete wishes by customers who have proactively approached you with a request or are they rather passive triggers like articles that you read or technologies that are considered to become state of the art? Maybe you are triggered by megatrends or you just like the idea driven by fascination rather than by reason.

- What impulses contribute to the attractiveness of this innovation?
- Do these impulses come from the direct environment of the company (e.g., customer, supplier, competitor, management, employees) or are they more general developments (e.g., digital transformation, megatrends)?
- Which specific effects do these impulses have on your company?

10.2.2 Demand

The demand ensures that the focus of the innovation is actually on future needs. Innovations often result in important strategic changes. They might pivot a whole company.

- What advantages does a user have from this innovation?
- What drawbacks can a user avoid through this innovation?
- How will the necessary activities and approach of your company change in the future?
- Which players in the company's environment are relevant in the future?

10.2.3 Value Proposition

An important success factor of an innovation is that the user gets more or pays less. The decisive question is what can be promised to the user with this innovation. An innovation can deliver a wide range of added value, e.g., a new product/service, better or more performance, individualization and adaptation to user wishes, facilitation of work for the user, attractive design, a strong brand and more status, an attractive price, lower costs, lower risk, better user friendliness, and availability.

- How can this idea be transformed into a convincing innovation?
- Which problem of the user do you solve with this innovation?
- What added value do you convey to the user?

10.2.4 Competitive Advantages

Innovations can include small improvements or be something completely new. Therefore, there may be several advantages over competitors, e.g., more attractive products, lower costs, etc. Above all, it may be important to consider how to build up advantages over competitors and how best to maintain them when competitors follow.

- What advantages do you have over competitors in terms of execution?
- Is there a time-to-market-advantage over competitors?
- How can you protect yourself against imitation of your innovation by other companies?
- What entry barriers can you build up in a targeted manner?

10.3 Who Do We Do It For?

10.3.1 Target Groups

A key aspect is the identification of relevant target groups and the most detailed market analysis possible. Typical types of target groups are internal users, mass market, niche market, market segments with slightly different desires, and little coherent segments with different demands.

- For which users do you create the most value, and which segments do you concentrate on?
- What are the actual needs of typical users?
- What is the current users' behavior and how will it develop in the future in the context of relevant trends?
- How large is the market potential estimated for the various target groups?
- What about the development of the markets, and which opportunities and risks should be taken into account when estimating market size?

10.3.2 Marketing

This aspect includes a systematic analysis of the marketing channels for the innovation. It requires both, a clear understanding of the users and a detailed overview of all players in the marketing process. Even if a precise financial analysis is carried out later, at least an initial assessment of the business model and at least a basic way of thinking for the management of the relationship with the user (e.g., attention—evaluation—purchase—mediation—after the purchase) must be taken into consideration. Typical categories of user relationships are personal support, self-service, automated services, communities, and participation through interaction.

- Can a viable business model basically be developed?
- What kind of relationship does each of your user segments expect from you?
- Which channels are preferred by your user, and who are the most important marketing partners then? Which work best and most cost-effective?
- Which competitors, new entrants, and other stakeholders need to be considered, and how will these external stakeholders respond?

10.3.3 Pilot Application

Basically, it can be assumed that there is a high added value in the involvement of users in the innovation process. Of course, considerations regarding the confiden-

tiality of new solutions should always be taken into account. Internal experts from R&D, sales, etc. can often identify some potential pilot customers for whom a new solution is of particular relevance relatively quickly. However, it is important to be aware that close cooperation with pilot users may lead to an excessive focus on their specific needs.

- Which pilot users can you work with to develop your innovation?
- How should potential pilot users be addressed and what is their status?
- How do you ensure that you do not focus too much on the pilot users?
- Which target segment is particularly promising in the launch phase?

10.3.4 Expansion Stages

During the innovation process, exciting ideas are often developed as to how the planned innovation can be further improved. Many of these remarks can be considered directly in the ongoing innovation process. Often, however, there are also numerous ideas that only concern the same subject area, but go significantly further, are not yet technically feasible, etc. Beyond an initial solution, the ideas for the further development of the innovation or for completely new functionalities should therefore be carefully documented.

- Which basic functionalities does a solution have to offer?
- Which possibilities for the further development of your solution are to be expected?
- Which functionalities are particularly desirable from the user's point of view?
- Which functions should currently not be considered and why?
- Which new technical possibilities will arise in the foreseeable future?

10.3.5 Activities

A concretization of the activities is necessary to create the planned innovation. Typical types of activities include production, purchasing, integration of required platforms, networks, etc.

- Which activities does your innovation require?
- Which activities do your relationships with users require?
- Which activities do your distribution channels require?
- Which activities do your revenue streams require?

10.3.6 Resources

This aspect includes a systematic analysis of resources for innovation. This does not only mean financial resources, but all necessary resources, e.g., physical resources (machines, etc.), intellectual resources (brands, patents, copyrights, data, etc.), human resources (time budget of key persons, etc.), and financial resources (development budget, etc.). The creation of a first prototype will certainly give further indications of the resources needed.

- What resources does your innovation require?
- What resources do your relationships with users require?
- What resources do your distribution channels require?
- What resources do your revenue streams require?
- Which resources do you already provide? Which do you have to invest in?

10.3.7 External Partners

Often, even large companies cannot carry out all the necessary activities for an innovation on their own. Rather, they need external cooperation partners. There are numerous reasons for such external partnerships, e.g., optimization and economies of scale, reduction of risks and uncertainties, and acquisition of certain resources and activities.

- Who are your most important partners?
- Who are your most important suppliers?
- What resources do you source from partners?
- Which relevant activities do partners carry out?

10.3.8 Other Projects

Within the framework of the further development of your concepts and solutions, you will often find links to other innovation projects and activities of your company and possibly also to external partners. Therefore, these positive or negative synergies with other projects should be analyzed in detail. Sometimes other projects are necessary for the success of your own solution, sometimes savings in the project budget can be achieved by combining the activities in your company, and sometimes other projects can also destroy the added value of a solution.

- Which other projects contribute to the success of this innovation?
- Which other solutions are relevant for your innovation from the user's point of view?
- What are the concrete interdependencies with your project?
- With which contact persons should you coordinate concretely?

10.4 How Does It Pay Off?

10.4.1 Cost Structure

Now you develop a rough cost-benefit analysis for the planned innovation. An important part of this is an overview of the costs associated with the innovation. This is not about the development costs for the innovation (these are taken into account in the project budget), but about the costs associated with the innovation after implementation. They include, for example, an overview of the fixed costs (wages, rents, and operating resources) and variable costs (e.g., components) as well as an assessment of volume advantages and economies of scale. In principle, innovations can often be cost-oriented (lean cost structure, low price promise,

maximum automation, and outsourcing) or value-oriented (focus on added value, premium).

- What are the most important costs associated with your solution, and how high are the total costs?
- Which resources and activities are most expensive?
- Who in the company can give as good an estimate as possible for the various costs?
- What external information can be used to estimate relevant costs, and who could be contacted?
- What does a rough cost planning look like as a basis for the cost-benefit analysis of the current project status?

10.4.2 Project Budget

This aspect includes a rough analysis of the budget required to develop this innovation. It is not a question of the costs accumulated, for example, for the components of a new product, but of the costs accumulated in the course of the project, for example, for market analyses, R&D, prototypes, external experts, etc. The assessment should be as realistic as possible, taking risks into account. It often makes sense to coordinate with other areas within the company in order to take into account any overlaps with other budgets in order to exploit synergies.

- How high is the necessary project budget until successful implementation and introduction?
- Which other budgets (e.g., R&D, digitization, coaching, fairs and events, etc.) overlap?
- How does a rough planning of the project budget look like at the current project status?

10.4.3 Revenue Streams

The analysis of possible sources of income based on this innovation is part of the cost-benefit analysis. Typical sources of income are: cost savings (e.g., in process innovations), internal transfer pricing, sales, usage fees, membership fees, lending/renting/leasing, license fees, brokerage fees, and advertising. In many cases, short discussions with potential users can lead to appropriate price estimates. The extent of the use of the innovation, e.g., through scenarios for market shares, must also be estimated.

- For what added value are users really willing to pay, or how can internal cost savings be achieved?
- What and how do users currently pay for, how would they prefer to pay, and can the prototype be further improved in this respect?
- How much does each revenue source contribute to the total revenue or how quickly does the solution pay for itself through internal cost savings?

- Which optimistic or pessimistic scenarios for the use of innovation are realistic?
- What does rough planning of revenues look like at the current project status?

10.4.4 Risks

In addition to the cost-benefit analysis and the required project budget, an assessment of the uncertainty and risks associated with a project is also required. These can be market risks, so that an innovation does not assert itself on the market or internally. Technical difficulties can also arise, so that the feasibility is unclear. There may also be financial risks because it is unclear whether a viable business model can be achieved. In addition, further problems can arise within the company or in the external environment.

- What are the market-related risks for the acceptance of your solution?
- What are the technical and organizational risks in the development process?
- What other financial risks exist during the introduction?
- Which general economic risks have to be considered?

11 Nothing Worth Without Culture

Generally, innovations that disrupt certain industries are more likely to arise from new entrants. This green field approach is tough enough. Even tougher is refreshing the innovative spirit of an existing company. It is like improving or repairing your race car while you are driving in a Formula 1 race at top speed. And it is even more demanding when you consider the incredible complexity of your existing company. Processes, functions, technologies, and people (not everybody will see the benefit of an innovation) need to be synchronized and aligned. A spanner in the works can vaporize your complete innovation effort. Well, it would be too easy (but unfortunately very common) to resign and accept that existing companies cannot really innovate. There is a solution. Not a quick fix, but a sustainable measure.

12 Drive Your Innovation Culture!

No strategy and no tool set can overcome the barriers from an unhealthy innovation culture. So before using strategy and tools make sure you are healthy. Anything else would be a waste of your assets and very frustrating.

- Align the organization on the importance of innovation first—top down, not bottom up. Your managers must be your innovation ambassadors.
- Make sure that everybody knows that it is part of his/her job to innovate. Get everybody on-board from the janitor to the president.
- Make sure that innovation is considered as a core process in your company and not a “weekly-1-h-Friday-afternoon-creativity-event.”

- Make sure that everybody knows that innovation is more than product innovation. It is also about processes, services, and business models.
- Make sure that innovation does not need to be disruptive every time. What might be incremental for you could be a major benefit for your customer.

13 Conclusions

We have been talking about the culture, the strategy, the methods, and the importance of the commitment of the leaders. Now you are well prepared not only to evaluate the innovations in this book, but also every innovation in your company you want to stress test.

Reference

Pisano, G. P. (2019). *Creative construction: The DNA of sustained innovation*. New York: PublicAffairs.

Innovative Technologies in the Ageing Population: Breaking the Boundaries

Guido Lerzynski

Abstract

In the coming years, the generation of senior citizens will be able to benefit from the possibilities offered by digitalisation and artificial intelligence to a considerable extent. A decisive factor for the success of the use of new technologies in old age is the attainment of digital independence. Key areas of application for digital innovations include improving care and mobility for the elderly, personalised medicine and social exchange. New challenges in the context of demographic change and the remoteness of rural and economically weak regions can thus be mitigated.

1 Introduction

Brigitte Lehmann is 72 years old and lives on the edge of the Eifel, a rural region in western Germany. She has two daughters aged 40 and 45, who moved with their families to metropolitan areas far away from Mrs. Lehmann's hometown. Her husband died 3 years ago of a heart attack. Mrs. Lehmann lives in an analogue environment. She has no knowledge of digital technologies. To date, this has had no negative impact on her life. She has also had hardly any health problems so far. However, she was diagnosed with adult-onset diabetes a few weeks ago. In addition, the two daughters discovered that their mother was increasingly forgetting things in daily life. An examination by her family doctor indicated the suspicion of senile dementia. Mrs. Lehmann now has to take medication in the morning and evening. However, due to the onset of dementia, taking medication regularly is not guaranteed. Mrs. Lehmann's two children are worried about their mother's health problems, but due to the distance involved, they can only keep in touch by

G. Lerzynski (✉)
St. Marien-Hospital, Cologne, Germany

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_7

telephone. In the future, Mrs. Lehmann will be dependent on external help for taking medication, but also for organising everyday life. In Mrs. Lehmann's rural region, it is becoming increasingly difficult to organise day-to-day assistance.

Renate Hermann is 76 years old and lives near Hamburg. She, too, has been widowed for 2 years and has no children. Her health situation has also worsened in recent years. In addition to cardiovascular disease, her mobility is increasingly limited. However, Mrs. Hermann is constantly gaining digital competence, in order to maintain her quality of life. She uses a tablet computer to check her vital signs daily, which are then supervised online by her family physician. Her medication intake is also monitored by an application on her smartphone. Mrs. Hermann orders her purchases, especially beverage crates, on the Internet. Her friends and acquaintances increasingly communicate with her via social media. Despite her health restrictions, Mrs. Hermann has been able to continue her life without major restrictions.

These two examples show how different life situations of elderly people can be, and the potential impact of digital innovations as well as artificial intelligence. The following chapter describes the opportunities and risks of digital innovation for older people, and presents a perspective for the future.

2 Demographic Shift

Population development has long been a much-discussed topic in social research. The change in the age structure caused by the decline in births and rising life expectancy is one of the greatest sociopolitical challenges in most industrial nations (Küpper and Peters 2019).

Demographic change is more advanced in Europe than on other continents. Europeans have the lowest birth rates in the world and live the longest. The increasing ageing of the population will shrink the working population in the coming years, and pose new challenges for the economy and social security funds (Kröhnert et al. 2019). However, demographic risks are by no means equally distributed across the continent. The aging process will hit economically weak regions, which are often rural, the hardest in the coming years.

In Germany, this applies to regions in eastern Germany as well as rural areas of Rhineland-Palatinate, Saarland, Lower Saxony, Hesse and North Rhine-Westphalia. The relationship between generations will change as a result of demographic change. This can be clearly seen from the age ratio for the coming decades (Table 1). In the years up to 2050, the over-60s will dominate German society.

In Europe, the ageing of the population and thus the demographic risks are greatest in the eastern and southern member states.

A decline in the population is not a problem if the next generation is well qualified and open to new technological innovations. Some gaps can also be filled with immigrants, provided they are well integrated. Young people are increasingly moving to places where there are jobs and attractive incomes. This further increases the regional disparities, because mainly older people remain behind in the econom-

Table 1 Youth and old-age dependency ratio from 2000 to 2050

	Youth ratio[a]	Old age ratio[b]
2000	38.1	42.8
2010	33.2	48.3
2020	31.3	59.9
2030	33.1	81.3
2040	32.1	85.9
2050	31.9	91.4

Birg (2011)
[a]Under 20-year-olds per 100 people aged 20–60
[b]Over 60-year-olds per 100 people aged 20–60

ically weaker regions. Here, there is ultimately a population decline, and the older generation dominates the remaining younger ones (Kröhnert et al. 2019).

The loss of the younger generation leads to a downward spiral. With a shrinking young population, the use of public facilities also declines, leading to high fixed costs in order to maintain operations. At the same time, revenues are reduced due to the lower take-up of local community services (Birg 2011). As a result, there will be demographically induced emigration of businesses and utilities, closures of administrative facilities, kindergartens, healthcare and leisure facilities such as swimming pools and sports fields (Birg 2011). Finally, locations for social life such as restaurants, pubs and clubs will also be affected.

In the coming years, much will depend on whether and how new technologies can replace the loss of traditional public places for the generation 60+. The outlook seems promising. Nevertheless, it is only the beginning of an integration of digital applications in the society of the over 60s. In order to pave the way for this, it is important to be able to let go of tried and tested things, and perhaps accept something new that is unusual. This is the only way to create an openness for digital solutions.

In Germany, many institutions are currently working on creating the conditions for the use of digital infrastructures (e.g. Bertelsmann and the Robert Bosch Foundation). However, in addition to highlighting the opportunities, the risks must not be concealed. Data protection and the protection of privacy are extremely important and must be adequately protected.

3 Digital Sovereignty

On an individual level, the term "digital sovereignty" refers to the competent handling of digital technologies (Zukunftspfade Digitales Deutschland 2014). Behind this lies the social requirement to create sovereignty. Digital sovereignty is basically an issue for all generations, but the focus is particularly on the older generation, with the question of where people must be engaged in order to become digitally sovereign (Stubbe et al. 2019).

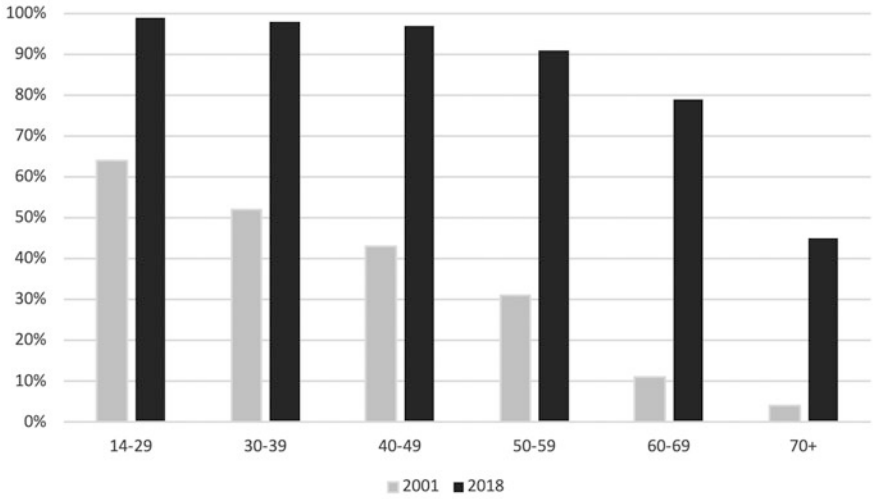

Fig. 1 Internet usage—age groups compared between 2001 and 2018 (Initiative D21 2019). Source: author

Between 2001 and 2018, the rate of Internet users in the individual age groups rose significantly. In 2018, almost 100% of the younger generations used the Internet. But even in the 70+ age group, almost one in two uses the Internet today (Fig. 1). With the ageing of the younger generations, a further increase in Internet use in old age is predictable. However, information on Internet usage does not yet provide enough details for the understanding of digital technologies.

Therefore, the topic of "digital competencies" is the most prominent element of digital sovereignty; competence here can be understood as the ability of the individual to behave appropriately in professional, social and private situations. Digital skills thus include the proper use of digital technologies and the ability to reflect on how to deal with them (Stubbe et al. 2019). The Deutsches Institut für Vertrauen und Sicherheit im Internet (German Institute for Trust and Security on the Internet, DIVSI) defined different Internet milieus (DIVSI 2016). People with a high level of digital skills, for instance, belong to a milieu called "sovereign realists" or "net enthusiasts". People in these milieus are mostly young and well educated. They are intensive users of the Internet (ibid). They emphasise the personal responsibility of the users, and therefore prefer to take care of their own security on the Internet instead of leaving it to others (Borgstedt et al. 2016).

The group of the older population is more likely to belong to the milieu of "the Internet wary", who are sceptical about the Internet and digital technologies (DIVSI 2016). This group represents almost one-fifth of German society. They feel overwhelmed regarding use of the Internet, and perceive significantly more risks than opportunities. This leads to extremely cautious use or rigorous avoidance of the Internet. Because they are hardly familiar with the Internet, they delegate responsibility for security above all to the state and companies, and take little

responsibility for themselves (ibid). People in this milieu feel overwhelmed by the significant social developments of the last decades (individualisation, digitalisation) (ibid).

Digital competencies include forms of operating knowledge, such as the use of office applications or messenger services, right through to the mastery of programming languages, which is relevant for the professional work environment. They also include skills that represent digital orientation knowledge—for example, the awareness that Internet services and apps pass on personal data, or the recognition of fake news (Initiative D21 2019). In general, digital competence in the German population is increasing significantly. However, this does not apply to the older generation, which still has significant potential for improvement in their awareness of digital skills (ibid). The main task for the coming years is to make digital competencies available and learnable for older people. Digital competencies may soon be crucial for participation in everyday life for the elderly.

4 Digital Education and Social Interaction

Access to digital education becomes a democratic prerequisite in the context of digital sovereignty. The education system faces the challenge of allowing individual learning paths, in order to enable full participation in society in the future.

Digital sovereignty requires changes in education, training and continuing education. Competence development can be achieved sustainably through direct experience (Loroff et al. 2017). The Digital Opportunities Foundation (Kubicek 2018) has drawn up a guideline on how competencies can be conveyed with a constant reference to the everyday life of older people. It emphasises aspects such as adapting learning content towards the everyday life of older people, only increasing complexity step by step, or not conveying knowledge exclusively digitally, but always offering an analogue alternative (Kubicek 2018). In principle, however, support services must be initiated across the country to teach digital skills to older people, to offer help with questions, and fundamentally promote self-confidence regarding what is to be learned.

Digital sovereignty is a social attitude. It can be related to the individual, but the term is only meaningful if it describes the attitude of the individual towards a group or towards society. In this respect, the term describes the conscious perception of one's own ability to shape a social environment. Thus, digital sovereignty can be understood as a form of social sovereignty (Stubbe 2017). General importance of social interaction for digital sovereignty is demonstrated by the success of social networks and services such as WhatsApp, as well as by the implementation of offers for further digital education in communities and neighbourhoods. Therefore, dealing competently with technology is essential. Among the elderly, communication-based applications such as WhatsApp are spreading more and more. However, obtaining information is still at the forefront of their Internet use: only 21% of over-65s are active in social networks, while 86% search for information about goods and services (Federal Statistical Office 2018). See Fig. 2 for details.

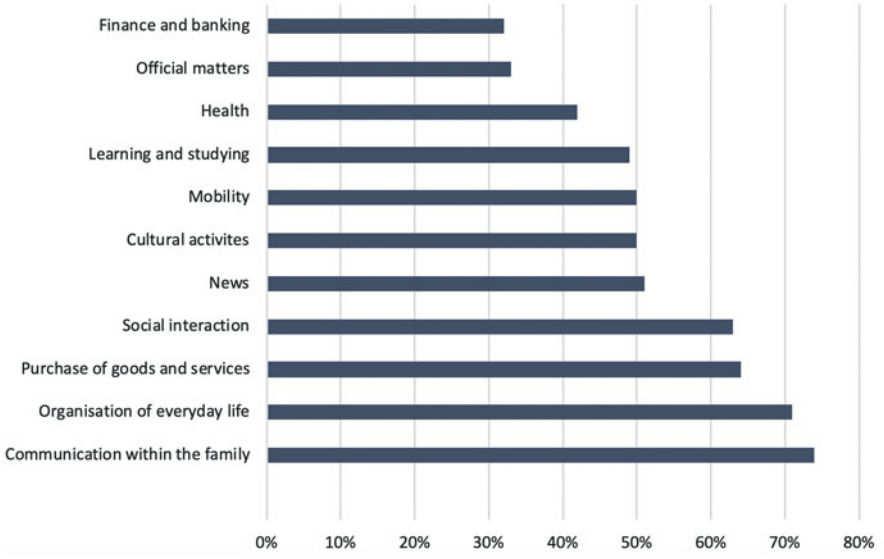

Fig. 2 Why do elderly persons acquire digital skills? Survey of 97 participants (Stubbe et al. 2019). Source: author

5 Data Security and Trust

Digital sovereignty presupposes that the legislator creates a regulatory framework in which people can act safely. In the past, the state has often undermined this trust through its actions. Since the beginning of global networking, questions of security, freedom and control have always been at the centre of attention (Castells 2005). The public's trust in the network society has moved to the centre of digital sovereignty (Friedrichsen and Bisa 2016).

The implementation of the General Data Protection Regulation (GDPR) by the European Union in 2018 was a milestone in the public's confidence in institutions. For the first time, it lays down uniform EU-wide rules on how private companies and public bodies must process personal data. This safeguards personal data within the EU and ensures the free movement of data within the European Single Market. These developments support the thesis that trust in the stakeholders who create the regulatory framework has become a cornerstone of digitalisation. This includes trust in the technological security that personal data is secure and protected against misuse by digital service providers, as well as social debate on ethical aspects of digitalisation and open and fair communication (German Ethics Council 2017). For example, conventions such as the use of open source software in private and public applications could facilitate the development of trusted systems (Gräf et al. 2018).

6 Usability

Digital sovereignty should also include products that most of the population can use. Usability describes the extent to which a technology can be used by a person, and whether this use is easy and convenient (user-friendly). Usability can be measured according to standards, but also includes a subjective dimension. The usability of digital products for older people has been a widely discussed topic in research for some time (Weiß et al. 2017). Meanwhile, older people are among the most popular subjects of research projects dealing with the design of user interfaces. The design of specific interfaces for older people is quite ambivalent, especially regarding digital sovereignty. It contradicts the idea of sovereignty being dependent on specially adapted offers that may be more expensive and not connectable. It is much more important to enable older people to use mainstream technologies, such as tablets or smartphones, and applications such as WhatsApp, Facebook, etc. (Stubbe 2017). This corresponds to the requirements of older people and promotes their cultural participation.

7 Artificial Intelligence as an Innovation Driver of Digital Technologies Amongst the Elderly

Artificial intelligence (AI) is increasingly becoming a catalyst for digital innovation. But which aspects of AI can empower us, and which can limit us in a self-determined way of life? While the defining characteristics of AI and machine learning make everyday life easier for older people, they also pose challenges to digital sovereignty (Stubbe et al. 2019).

Digital technologies are constantly changing. Interviews with experts asked which technologies will be particularly relevant for older people in the future. The answers were condensed into a questionnaire that was answered online by people who convey digital skills to older people (Fig. 3). In the survey, these people stressed the innovation potential of personal assistance systems, smart home technologies and e-health for older people (Stubbe et al. 2019). In addition to the other technologies on the list, the application potential of these three areas is strongly influenced by AI. In the digitalised world, AI is becoming a cross-sectional technology that provides an innovation boost to applications that are already widespread.

The great potential of artificial intelligence and digital innovation is evident in the ageing generations to come, who have already acquired a high level of digital competence. It can be assumed that problems caused by demographic change can be mitigated by the targeted use of artificial intelligence.

Most challenges in life are mastered by people using their cognitive abilities to process information in the brain. We can use the term intelligence as a measure of the extent of this problem-solving ability. Because computers also process information, the desire arose to equip them with human problem-solving abilities.

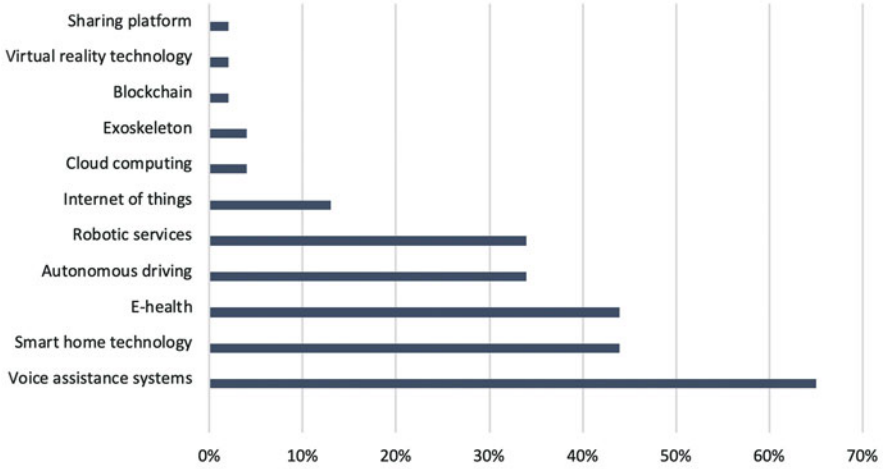

Fig. 3 Which digital technologies will have the greatest impact on the lives of the elderly in the future? Survey of 97 participants (Stubbe et al. 2019). Source: author

In the early days of AI research, researchers used logic-based and knowledge-based approaches to explicitly programme the knowledge required for intelligent problem-solving in computers (Stubbe et al. 2019). The technical possibilities of the past years and decades led to the current dominance of so-called data-based approaches of supervised machine learning, in which computers are supposed to draw the necessary knowledge from sample data of human problem-solving. Knowledge is extracted from the data sample provided, which can then be used for specific problem-solving. With a large amount of structured and representative sample data, algorithms trained for specific problems can perform human-like tasks: for example, the error rate of machine image recognition has been the same as that of humans since 2015, and has been continuously improving since then (Russakovsky et al. 2015). This is also successfully used in medical applications for radiological image recognition. Due to the large number of radiological image files to which the algorithm is exposed, the diagnosis of pathological findings is constantly being improved, and is at least comparable with human diagnostic findings. Moreover, it can even be assumed that computer algorithms will soon outperform human potential in terms of diagnostic accuracy. This results in a wide range of applications for artificial intelligence in the field of health sciences, which can contribute to an improvement in quality of life, but can also detect and treat diseases much more accurately. Some of the diminishing medical services in rural areas can be mitigated by digital applications and artificial intelligence, which are available at any time and at any place.

The increasing availability of data also leads to the use of AI in applications that are already part of everyday life for many people. Nonetheless, more than one in two respondents (54%) of a BITKOM study (BITKOM Research GmbH 2018) has

already used voice assistants on smartphones, and a further 21% would like to do so. Two-thirds of the respondents (68%) would like AI to support elderly people. Only 6% are of the opinion that artificial intelligence will not noticeably change society.

In addition to the opportunities, respondents also see risks of using AI in certain areas, such as childcare and relationships. This opinion coincides with a Bertelsmann Stiftung survey on the use of autonomous computer decisions (Fischer and Petersen 2018), which are often made by AI systems: the more people are directly affected when using these systems, the greater their rejection. The risks of AI systems often only become apparent when they are used. This can be seen, for example, in e-health applications, where—according to the online survey—digitalisation will have a particularly significant impact on the lives of older people. Especially in the health sector, it is a prerequisite for their use that AI systems are not only controllable, but also transparent and comprehensible. The adaptivity of applied AI systems must also be geared towards the well-being of those affected, and not towards economic criteria. The data dependency of dominant methods of machine learning brings additional risks, especially when dealing with sensitive data. In order to preserve the digital sovereignty of users, it should thus be possible to show for which goals AI systems use their autonomy, and how these goals are implemented (Stubbe et al. 2019). The criteria of data protection and privacy must be considered as early as the development of the algorithm. This then also requires new interactive relationships between humans and technology.

8 The Future of Humans and Technology

The convergence of humans and technology is not a new model. Rather, the idea that humans are optimised by technology is a utopia that is not only conceived in philosophical transhumanism, but is already becoming reality today under the label of "enhancement". Which characteristics should be technically enhanced, and to what extent these enhancements conflict with social perceptions of "being human", is not historically determined, but part of discussions and negotiations (Dickel 2016).

In contrast to the United States, the European canon of values has so far been shaped by the attitude that technology should compensate for human deficits, and not improve people. A change of this attitude must not take place with a radical shift, but can take place slowly and gradually between compensating and improving (Bovenschulte and Stubbe 2018). Sensor technology close to the body is a preliminary stage of implants that lie under the skin. The best-known example to date is the cochlear implant, a sensory neuroprosthetic for people with impaired hearing that is already being used today. Retina implants, brain pacemakers or cardiac pacemakers have attained a similar degree of maturity. Important research topics that have already been incorporated into national funding programmes include, for example, active implants to make information tailored to patients' needs transparent and available; interfaces to implanted systems that allow physicians and medical professionals easier access to information for participatory decision-

making; technologies for the conscious control and management of implanted systems; and the improvement and functional expansion of current implants through new interaction options (BMBF 2018). Some of these technologies are already being used by people in order to enhance their own capabilities. There are now non-medical products on the market or instruction manuals for applications on the Internet, all of which aim to improve the ability to concentrate through electric stimulation of certain brain regions.

The debate about opportunities and risks, about the "good" or "bad" of these technologies is also already taking place, as is the discussion about the supposed advantages of prosthetic legs over healthy legs in competitive sports.

The health sector will initially remain the area in which social acceptance of physically integrated technologies will emerge. Here, older people will also encounter invasive technologies—probably earlier than younger people. Older people will play a pioneering role here. In the more distant future, developments will go beyond the health sector and include everyday and formal activities: payment processes, identification of drivers or residents. Interactive implants will not only interact with their environment, but will also be adaptively configurable. Users can determine the purpose of the sensor technology under their skin by independently teaching and further developing the AI of the implants.

9 The Digital Divide

The extent to which the digitalisation of society leads to the exclusion of certain population groups is discussed under the term "digital divide". There has been a development within this debate: although for many people in Germany, the general accessibility of Internet connections and broadband continues to be a prerequisite for digitalisation that has not yet been achieved (BMVI 2019), the question of digital competence as an indicator of a digital social divide is gaining more and more importance. Digital skills go beyond the operation of new devices, and increasingly concern aspects such as awareness that personal data is passed on through Internet services, the recognition of fake news, or the independent handling of hostility in social networks. Even though older people are increasingly living digitally, they still lag behind in terms of digital skills (Initiative D21 2019).

The design of digitalisation requires the interaction of stakeholders at different levels. National and European funding programmes set formative guidelines that steer research and development over several years, while innovations are sustainably put into practice at the local level. The success of digital sovereignty depends above all on the level of civil society: this is where the social debate about values and goals in digitalisation takes place. Civil society representatives must engage people with their demands and needs and at the same time invite them to participate constructively. Elderly people are called upon as partners to achieve digital sovereignty: based on their life experience, they can convey important skills, participate in innovation projects, or expand the data pool that can be used by society through open data and data solidarity. Sovereignty means well-reflected decision-

making, for which life experience and education are important prerequisites (Stubbe et al. 2019). Openness and commitment are often apparent in many older people: there are many elderly people who want to help shape digitalisation. To achieve this, the deficits that the elderly have in dealing with new technologies must be eliminated, by providing digital competence.

The success of digital sovereignty requires a rethink in the development of digital technologies. This applies above all to large corporations, but also to science and research. For a long time, the low-friction and visually appealing interaction between users and technology has been an ideal of technology design. This ideal has produced innovations that can exploit the potential of personal assistance systems or smart home applications. However, digital sovereignty requires more: a low-friction interaction must not obscure algorithmic functions, but should present them transparently and comprehensibly in the interaction. The technical prerequisites for this are already in development. It is essential that the test users involved in developing the technology actively follow the ideals of critical citizens. The participation of people in innovation processes has been strengthened in recent years. In fundamental research, the integration of users is now part of a widespread design approach. However, the discussions on network policy also show that the institutions have not yet fully grasped the social requirement of participation.

10 Conclusions

Digital technologies and the key AI with its diverse applications offer considerable added value for the lives of older people. More and more data resources are being collected for this purpose, but sensitive personal data is also being used. Technical design principles can be effective mechanisms to meet the need for the highest possible level of data protection. They are thus the prerequisite for the social benefits of digital technologies to be unfolded through a solidarity-based approach to data. As a result, it is becoming increasingly important for older citizens to have the appropriate knowledge and digital competence so that they can confidently decide whether they want to disclose their own sensitive data. Digital skills are thus becoming even more of a prerequisite for social participation than they are today. The conditions for older people to acquire digital skills are good: interest and curiosity are the most important motivators. In addition, a wide range of low-threshold educational opportunities are already emerging, which should be expanded in the coming years. As the digitally experienced generations age, the penetration of digital applications and artificial intelligence will continue to increase.

If properly understood and applied, digital technologies and artificial intelligence are important partners in the effort to ensure the quality of life of older people, and to positively shape the consequences of demographic change. Medical diagnostics and treatment will change radically in the coming years, due to the possibilities offered by artificial intelligence and digital technologies. It will be an exciting and promising path.

References

Birg, H. (2011). *Soziale Auswirkungen der demographischen Entwicklung*. Bundeszentrale für politische Bildung. Accessed September 2, 2019, from www.bpb.de/izpb/55920/soziale-auswirkungen-der-demographischen-entwicklung

Bitkom Research GmbH. (2018). *Künstliche Intelligenz* (Research Spotlight, 2018-06). Accessed August 18, 2019, from www.bitkom-research.de/epages/63742557.sf/de_DE/?ObjectPath=/Shops/63742557/Categories/Presse/Spotlight/Research_Spotlight_201806_Kuenstliche_Intelligenz

BMBF – German Federal Ministry of Education and Research. (2018). *Technik zum Menschen bringen*. Forschungsprogramm zur Mensch-Technik-Interaktion. Accessed August 17, 2019, from www.bmbf.de/de/technik-zum-menschen-bringen-149.html

BMVI – German Federal Ministry of Transport and Digital Infrastructure. (2019). *Der Breitbandatlas*. Accessed September 2, 2019, from www.bmvi.de/DE/Themen/Digitales/Breitbandausbau/Breitbandatlas-Karte/start.html

Borgstedt, S., Resch, J., von Schwartz, M., & Ernst, S. (2016). *DIVSI Ü60-Studie. Die digitalen Lebenswelten der über 60-Jährigen In Deutschland. (Ed.)*. Hamburg: Deutsches Institut für Vertrauen und Sicherheit im Internet (DIVSI).

Bovenschulte, M., & Stubbe, J. (2018). Intelligenz ist nicht das Privileg von Auserwählten. In V. Wittpahl (Ed.), *Künstliche Intelligenz* (pp. 215–220). Berlin: Technologie| Anwendung| Gesellschaft.

Castells, M. (2005). *Die Internet-Galaxie*. Internet. Heidelberg: Wirtschaft und Gesellschaft.

Dickel, S. (2016). Der neue Mensch. Ein (technik)utopisches Upgrade. In *Aus Politik und Zeitgeschichte* 37/38 (pp. 16–21).

DIVSI. (2016). *DIVSI Internet-Milieus 2016 Die digitalisierte Gesellschaft in Bewegung*. Hamburg: Deutsches Institut für Vertrauen und Sicherheit im Internet. Accessed July 31, 2019, from https://www.divsi.de/wp-content/uploads/2016/06/DIVSI-Internet-Milieus-2016.pdf

Fischer, S., & Petersen, T. (2018). *Was Deutschland über Algorithmen weiß und denkt. Ergebnisse einer repräsentativen Bevölkerungsumfrage*. Pub. Bertelsmann Stiftung. Accessed August 28, 2019, from www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/Was_die_Deutschen_ueber_Algorithmen_denken.pdf

Friedrichsen, M., & Bisa, P. J. (Eds.). (2016). *Digitale Souveränität*. Wiesbaden: Vertrauen in der Netzwerkgesellschaft.

German Ethics Council. (2017). *Big Data und Gesundheit—Datensouveränität als informationelle Freiheitsgestaltung*. Berlin.

German Federal Statistical Office. (2018). *Die Hälfte der Generation 65 plus surft im Internet*. Press release from 18 October 2018. Accessed September 3, 2019, from www.destatis.de/DE/Presse/Pressemitteilungen/2018/10/PD18_407_p001.html

Gräf, E., Lahmann H., & Otto, P. (2018). *Die Stärkung der digitalen Souveränität. Wege der Annäherung an ein Ideal im Wandel*. iRights.Lab; Deutsches Institut für Vertrauen und Sicherheit im Internet (DIVSI).

Initiative D21 e.V. (2019). *D21 Digital Index 2018/2019 Jährliches Lagebild zur digitalen Gesellschaft*. Accessed September 2, 2019, from https://initiatived21.de/app/uploads/2019/01/d21_index2018_2019.pdf

Kröhnert, S., Hoßmann, I., Klingholz, R. (2019). *Die Demographische Zukunft von Europa* (p. 6). Berlin: Berlin-Institut für Bevölkerung und Entwicklung, DTV. ISBN: 978-3-423-34509-5.

Kubicek, H. (2018). *Leitfaden—Digitale Kompetenzen für ältere Menschen. So plane ich und gestalte ich Angebote zur Unterstützung von Senioren*. Berlin: Pub. Jutta Croll. Stiftung Digitale Chancen.

Küpper, P., & Peters, J. C. (2019). *Entwicklung regionaler Disparitäten hinsichtlich Wirtschaftskraft, sozialer Lage sowie Daseinsvorsorge und Infrastruktur in Deutschland und seinen ländlichen Räumen* (p. 168). Braunschweig: Johann Heinrich von Thünen-Institut, Thünen Rep 66. https://doi.org/10.3220/REP1547565802000.

Loroff, C., Lindow, I., & Schubert, M. (2017). Bildung als Voraussetzung digitaler Souveränität. In: V. Wittpahl (Ed.), *Digitale Souveränität. Bürger, Unternehmen, Staat* (pp. 151–175). Berlin: Springer.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.

Stubbe, J. (2017). Von digitaler zu soziodigitaler Souveränität. In: V. Wittpahl (Ed.), *Digitale Souveränität. Bürger, Unternehmen, Staat* (pp. 43–60). Berlin: Springer.

Stubbe, J., Schaat, S., & Ehrenberg-Silies, S. (2019). Digital souverän? Kompetenzen für ein selbstbestimmtes Leben im Alter. *Bertelsmann Stiftung*. https://doi.org/10.11586/2019035.

Weiß, C., Stubbe, J., Naujoks, C., & Weide, S. (2017). *Digitalisierung für mehr Optionen und Teilhabe im Alter*. Gütersloh: Pub. Bertelsmann Stiftung. Accessed September 2, 2019, from www.bertelsmannstiftung.de/de/publikationen/publikation/did/digitalisierung-fuer-mehr-optionen-und-teilhabe-im-alter/

Zukunftspfade Digitales Deutschland 2020. (2014). *Eine Studie des IT-Planungsrates, durchgeführt von TNS Infratest*, October 2013 (p. 34). Accessed September 2, 2019, from www.cio.bund.de/SharedDocs/Publikationen/DE/Aktuelles/studie.pdf?__blob=publicationFile

Using Augmented Reality and Machine Learning in Radiology

Lucian Trestioreanu, Patrick Glauner, Jorge Augusto Meira, Max Gindt, and Radu State

Abstract

Surgeries are one of the main cost factors of health care systems. To reduce the costs related to diagnoses and surgeries, we propose a system for automated segmentation of medical images in order to segment body parts like liver or lesions. The model is based on convolutional neural networks, for which we show promising results on real computed tomography scans. The deep learning algorithm is part of a larger system that aims to support doctors by visualizing the segments in a Microsoft HoloLens, an augmented reality device. Our approach allows doctors to intuitively look at and interact with the holographic data rather than using 2D screens, enabling them to provide better health care. Both the machine learning algorithm and the visualization utilize high-performance GPUs in order to enable doctors to interact efficiently with our system.

1 Introduction

According to a 2016 Goldman Sachs report, the combined virtual reality (VR) and augmented reality (AR) market is forecast to value 80 billion USD in revenue, by 2025. It is expected new markets to be created and existing markets to be disrupted (Bellini et al. 2016). Meanwhile, the USA spent 17% of its 2015 gross domestic product on health care, with approximately 3% related to surgery (Douglas

L. Trestioreanu (✉) · J. A. Meira · M. Gindt · R. State
Interdisciplinary Centre for Security, Reliability and Trust University of Luxembourg, University of Luxembourg, Luxembourg, Luxembourg
e-mail: lucian.trestioreanu@uni.lu; jorge.meira@uni.lu; radu.state@uni.lu

P. Glauner
Deggendorf Institute of Technology, Deggendorf, Germany
e-mail: patrick@glauner.info

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*, Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_8

et al. 2018). As such, there is a strong demand for increasing efficiency and effectiveness in the operating rooms in order to both decrease cost and improve patient care, which can be addressed by taking advantage of the latest technological developments. Despite advancements in medical imaging, the current pipeline used by the practitioners is not automated and the quality of results depends on the limited resources and subjectivity related to the human factor processing the images. Recently, less time-consuming, automatic solutions which do not rely on the human factor have been investigated, like using machine learning to perform the image *segmentation.*

Segmentation of images is the process of partitioning the image into semantically meaningful parts and classifying each part into predetermined classes (pixels representing bone, liver, arteries, kidney, and tumor). As segmentation of medical images is a time-consuming manual process, involving machine learning into this use-case has gained a high interest in the last years.

Also, the 3D volumetric images are currently viewed on 2D displays with a mass of information remaining hidden from the viewer, which increases the effort required to interpret the results (Gotra et al. 2017). It is estimated that humans retain approximately 10% of written information, 20% of audio, 30% of visual, 50% of combined audio-video information, 70% of discussions, and 80% of personal experiences (Jung 2019). As such, the potential of AR and VR in improving the cognitive experience is high. Today there are different approaches to present the 3D medical images in a true 3D manner, but few of them address aspects of practicality and ergonomics such as those offered by augmented reality and more specifically by the Microsoft HoloLens, depicted in Fig. 1 (left): surroundings awareness and interaction using mechanisms unbound by traditional input hardware, which opens the possibility of using this technology in the medical operating theaters, which have strict requirements.

Besides aspects like almost instant access to segmented data, this work aims to leverage practical advantages of the AR technology in order to create a valuable product for real-life and real-time medical usage. More specifically, we aim to address aspects like:

Fig. 1 3D visualization of liver segmentation in Microsoft HoloLens (left): the segmented liver volume (right) was separated from the input volume—a scanned torso (center). Source: authors

- Being able to access the information (3D medical image and patient data) while during the situation (surgery, medical emergency) and get (near-) real-time updates;
- Information sharing (multiple users working in the same time on the same data, marking and pointing) which can also be useful during the planning of the operation;
- Strict hygiene rules in the operations rooms—the ability to control the device without having to touch any dedicated hardware;
- Surroundings awareness (being able to see the real-world environment in the same time with the medical data).

One of the challenges is the inherently limited computing power associated with any mobile device, worsened by the high requirements for the GPU to render the 3D medical volumetric images, which we address by using dedicated external hardware. Our main contributions are:

- We propose and evaluate a machine learning algorithm based on convolutional neural networks for liver and lesion segmentation. For the example volume depicted in Fig. 1 (center), it segments the liver as shown in Fig. 1 (right);
- We propose and evaluate a platform for a fast holographic visualization of 3D volumetric organ segmentations inside the Microsoft HoloLens headset using a dedicated high-performance GPU server;
- We provide a system to connect all the pipeline modules, to create an end-to-end automatic pipeline from image acquiring through holographic visualization.

The rest of this chapter is organized as follows. In Sect. 2, we review the state of the art of medical image segmentation and holographic data visualization. In Sect. 3, we propose our methodology, which we subsequently evaluate in Sect. 4. We summarize our work in Sect. 5.

2 Related Work

Many methods for liver segmentation employ statistical shape and intensity distribution models (Hoogi et al. 2017a), while others rely on classifiers and low-level segmentation (Hoogi et al. 2017b). Support vector machines and random forests, although having a higher discriminative power than intensity-based techniques, can lead to coarse segmentation and leakage (Gotra et al. 2017). Because of the flexibility and complexity of the learned features, neural networks hold the most potential to infer high-level features with acceptable errors. Convolutional neural networks (CNN) are robust to image variation and allow building automatic segmentation of heterogeneous liver scans in under 100 s (Gotra et al. 2017). Some challenging training characteristics like the imbalance of data labels require some modifications of the classic approaches though. 3D medical images can be processed using several methods, like 2D fully convolutional networks (FCN) as in deep retinal

image understanding (DRIU) (Maninis et al. 2016) or U-net (Ronneberger et al. 2015). Both output the 3D segmentation by processing each 2D slice and have a disadvantage at z-axis spatial correlation. DRIU (Maninis et al. 2016), the basis of the 2.5D model presented in this chapter, segments the optical disc and blood vessels of the eye. The FCN used by DRIU has side outputs with supervision at different convolution stages (Bellver et al. 2017) and the final output is obtained by combining the multiscale side outputs.

Current methods for medical imaging visualization of computed tomography (CT) or magnetic resonance images (MRI), such as 2D slice-by-slice viewing, 3D surface rendering (Douglas et al. 2017, 2018), and 3D volume rendering (Hohne et al. 1988; Calhoun et al. 1999) are limited to 2D displays (Fishman et al. 2006). While 3D images have certain advantages over 2D, presenting them on 2D displays still involves certain disadvantages like not being able to achieve true depth perception or cognitive limitations in the case of overlapping structures (Johnson et al. 1996; Heath et al. 2006). This can be overcome using depth three dimensional (D3D) imaging, i.e., displaying stereo 3D images on AR or VR headsets, which involves a GPU-intensive stereo rendering engine providing options to move, rotate, or scale using a hardware or touch-less controller based on voice, gaze, and gestures. AR and VR provide enhanced viewing including depth perception and improved human machine interface (HMI). AR, mixed reality (MR), and VR head mounted displays (HMDs) present a unique image for each eye, thus achieving stereoscopy and depth perception.

VR is fully immersive, displaying only the virtual image while the real surroundings are occluded. Full immersion and the presence of hardware controllers limit its usage for live medical interventions due to hygiene and surroundings awareness reasons. AR simultaneously displays the virtual image and real environment and, for some medical applications, the real-world image can be the patient's body (Douglas et al. 2017). Generally, VR is suitable for an off-site experience while AR is used for enhancing on-site experience. We should also mention that by the generalized definition, AR can make use of all other senses like touch, hearing, or smell (Carmigniani and Furht 2011).

Summarizing, the advantages of AR and VR over traditional medical imaging display methods are:

- True depth perception, which considerably improves the diagnostician's interpretation (Beydoun et al. 2017);
- The 3D characteristic, which subsequently provides the possibility of adding novel user interface (UI) and tools (Douglas et al. 2017), for example, 3D cursors and markers, and an overall increased efficiency in processing high amounts of information (patient data);
- Introduces the possibility of improved human control (Beydoun et al. 2017) interface capabilities like motion tracking (the scene camera view angle and position updates with head movement and rotation) (Douglas et al. 2017), use of more ergonomic controllers inspired from gaming, or even voice and gestures which would provide a complete hygienic environment in the case of live intra-

operative situations. We explore the voice and haptic interaction in surgical settings which would allow the medical images to be viewed and manipulated without contact by making use of the HoloLens' gesture recognition capability;
- In the case of MR (e.g., the HoloLens), the possibility of co-registration of the medical image to the real patient's image gives the impression of seeing into the patient, creating a live perception of the medical image inside the patient (Douglas et al. 2017).

Also, there are still some challenges faced by AR and VR in the medical imaging area:

- The perception of structures overlapping in the image is reduced but not eliminated (Douglas et al. 2017);
- Only light headsets stand a chance of being accepted into the operations room (HoloLens may still be considered somewhat bulky and heavy) (Douglas et al. 2017);
- Motion sickness is still a potential problem which can hamper the medical personnel's capacity to best performing the medical act (Douglas et al. 2017);
- The HoloLens, being still an early product, has a limited field of view (FOV), which forces the user to turn his head (possibly away from a live surgical intervention) if the displayed image is outside FOV (Beydoun et al. 2017);
- The HoloLens tinted visor that covers the display dampens the ambient light and decreases the efficacy of other potential diagnostic monitors involved (Beydoun et al. 2017).

Despite some inherent challenges faced by any new technology, AR and VR will continue to develop to finally put their footprints into real life, including the medical industry. If augmented by the PC's computing power, we believe that at this time HoloLens offers the best mix of advantages and disadvantages for the medical visualizations usage. The Microsoft HoloLens AR headset depicted in Fig. 1 (left) provides surroundings awareness, good tracking, and multiple touch-less input options such as voice, gestures, and gaze. To date, these are the most promising answers to the strict hygiene requirements in the operating rooms. The HoloLens is a relatively lightweight, standalone device with no attachments hampering movement. It offers the possibility of image co-registration and collaborative work, i.e., multiple users view, point, and mark in the same time.

NOVARAD's "Opensight" and DICOM Director are worth mentioning as just some examples of companies and projects activating in the medical field and which are proposing promising solutions. One more notable example of using AR or VR in medicine is EchoPixel, which proposes an interesting holographic solution, although with the limitation that it cannot be used during a live situation due to the visualization method and hardware controller used (Aguirre 2019).

3 Methodology

We combine a machine learning algorithm for image segmentation with the advantages of AR, in particular Microsoft HoloLens, in order to advance the technology into the live medical operating rooms. A general overview of the whole proposed system is illustrated in Fig. 2.

The images acquired by the CT scanner are stored in a shared storage. A script monitors it and queues a new liver segmentation job as soon as it detects a new CT scan. A second script manages the new segmentation results: as soon as a new segmentation is available, it copies it in a second shared storage for classic visualization and starts the file format conversion. Furthermore, a 3D volumetric image rendering server based on the Unity engine powered by one or more NVIDIA GPUs renders the 3D liver segmentation and encodes the video frames. Next, it sends them through a 5 GHz 802.11ac router towards the HoloLens client for visualization and manipulation as a hologram. The client sends back to server the camera position and controls updates. Through the user interface, the user can browse, load, or unload for visualization the segmentations already prepared in Unity format on the second shared storage.

The advantage of having two dedicated machines is twofold: as both the 3D visualization and the medical volumetric image segmentation using a CNN are resource-intensive, the visualization performance will not be affected by a (possibly overlapping) running segmentation (ensure no hardware resource sharing), and also, the machine learning works at its best on Linux. Our experiments have also shown that for best latency, the router should be a model capable of running in the 802.11ac WiFi standard at 5 GHz, and channel bandwidth should be set as large as possible (we used 80 MHz).

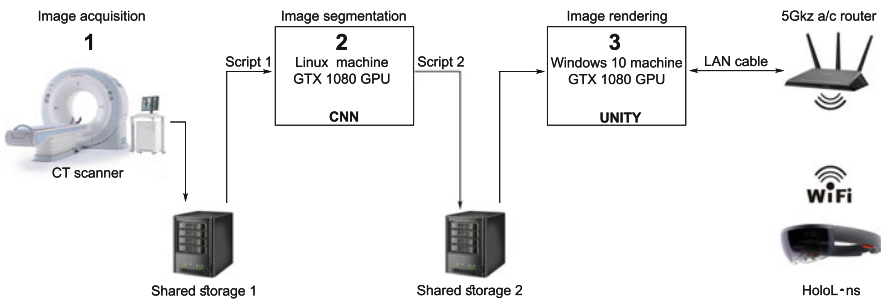

Fig. 2 System architecture. Source: authors

3.1 The Machine Learning Algorithm

Because of noise in the images, lesions variability, and low contrast between the image components (organs, lesions, and other tissue), a quality automatic segmentation can prove to be a complex task to achieve.

The machine learning algorithm used for segmentation is inspired by DRIU (Maninis et al. 2016), and the framework is TensorFlow (Bellver et al. 2017). This model has already demonstrated good performance during a liver segmentation dedicated competition, and also its generality is demonstrated on different anatomical structures from the Visceral data set. The main steps of the algorithm are depicted in Fig. 3.

After the liver segmentation, the segmented volume is cropped slice by slice around the liver region of interest (ROI). The resulting smaller liver segmentation volume is fed to both the lesion segmentation network and the lesion detector. Finally, the output predicted by the lesion segmentation network is compared to the output of the lesion detector, and only if both agree, the lesion localization is kept. Some of the most important aspects of the algorithm are:

- Preprocessing,
- Binary cross entropy (BCE) loss weighting (solve data imbalance),
- Using 3 consecutive slices as input (2.5D approach),
- Masking, and
- Post-processing using 3D conditional random fields.

3.1.1 Preprocessing

The pixel intensities in the data set have values exceeding |1000|. Many pixels with value −1024 belong to the background. Preprocessing means clipping the pixels' intensities values at min–max values that belong to the liver and liver lesions (−150 to 250), from a statistical point of view. A min–max normalization is done on all

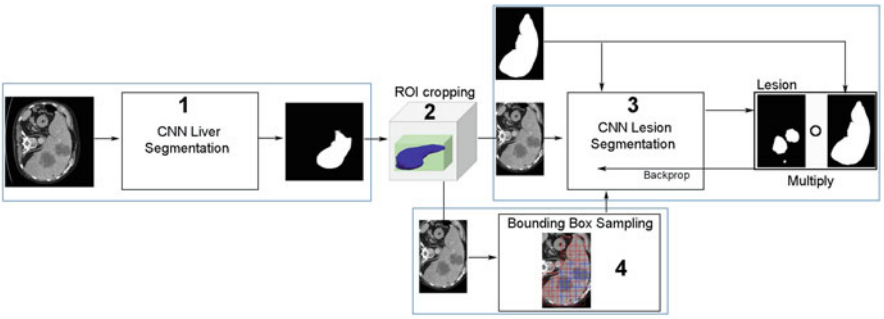

Fig. 3 Segmentation algorithm overview: liver segmentation (1), region of interest cropping (2, green box), lesion segmentation network (3), and the lesion detector (4). Source: authors

volumes afterwards:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \tag{1}$$

which means the range between the old min and max values is mapped to the new range between the new min and max values (−150 to 250).

3.1.2 Loss Objective

For the loss objective, the binary cross entropy (BCE) loss is used (Rothman 2018):

$$L(y, \widehat{y}) = -y \log \widehat{y} - (1 - y) \log(1 - \widehat{y}), \tag{2}$$

where y is the actual truth and $\widehat{y}$ is the predicted value. The BCE offers an individual per pixel loss which allows knowing which loss comes from positive or negative pixels. As such, the positive and negative loss can be balanced separately.

3.1.3 Input Multiple 2D Slices to Take Advantage of 3D Data

Originally the algorithm parsed the data as if the slices were independent. Also, the algorithm was pretrained with Imagenet, and it uses 3 channel images for training—RGB (red, green, and blue). Because actually the medical image is 3D-coherrent, as shown in Fig. 4, the three channels can be fed simultaneously with three slices from the input image, for each RGB channel (2.5D approach). This is done both during the liver and lesion segmentation. During testing only the central output slice is kept.

3.1.4 ROI Cropping

After the liver segmentation, the outputted volume is fed to the ROI cropping, where the liver segmentation volume is cropped slice by slice around the liver ROI, as shown in Fig. 3 (module 2). The resulting smaller volume is afterwards used as input for the lesion detector and the lesion segmentation modules. The number of positive pixels in each slice of the predicted liver masks resembles a Gaussian, so after a fitting of a Gaussian, a mean and variance are computed (Bellver et al. 2017).

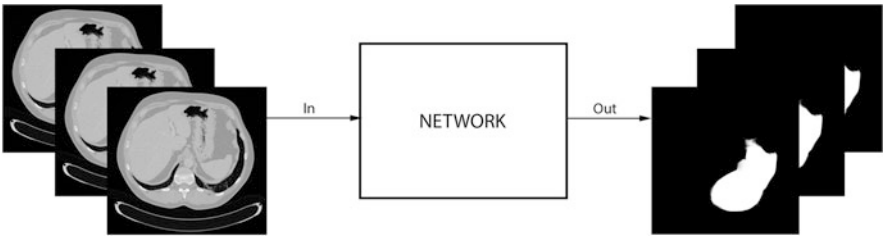

Fig. 4 The input uses a series of three consecutive slices. For the test, only the middle output image is used. Source: authors

The fitting is used to remove false positives, as all images outside a certain threshold are not likely to contain any lesion. As such, a significant number of false positives are removed, in return for just a small number of false negatives.

3.1.5 Using the Liver Segmentation for the Lesion Segmentation

The segmented liver is used for segmenting the lesions as a mask which constrains back-propagation only to those pixels which belong to the liver segmented ROI. This way, only pixels belonging the liver are used in the lesion segmentation process, and also, the process gets more balanced because a significant number of negative pixels in the image have been eliminated. The balancing term includes just the liver pixels. This process is illustrated in Fig. 3 (module 3). The possible disadvantage of this approach is that if the liver segmentation is not of good quality, this will negatively affect also the process of lesion segmentation.

3.1.6 Lesion Detector Module

Because as such, the original algorithm was not considering a global view when performing lesion segmentation, it was often returning false positives and, to help it get a larger context, a lesion detector was added (Bellver et al. 2017). The difference between segmentation and detection is that image semantic segmentation classifies each pixel of the image as belonging to one class or another, while detection searches for some object (dog, car, lesion, and smile) and localizes it generally with a bounding box. The detector discovers from a larger point of view, in which parts of the image there is actually a lesion. Then, the segmentation result is compared with the detector's result, and only those locations where the results are in sync are kept. The lesion detector module is built from the pretrained ResNet 50 model, without the Imagenet classification. Finally, only one neuron is used to take the healthy or non-healthy decision.

3.1.7 3D Conditional Random Fields

A 3D fully connected conditional random field (3D-CRF) is used for final processing. Conditional random fields make use of all the input available in order to model the conditional distribution of the result. The 3D CRF refines the segmentation by taking into account the spatial coherence and the pixels' intensities (Bellver et al. 2017). The 3D CRF uses the algorithm from (Christ et al. 2017), and its input is the soft prediction of the network output and the preprocessed volume.

3.1.8 Loss Balancing

Data imbalance comes from the fact that not all liver scans contain lesions (most are healthy), and also from the fact that inside the liver, the healthy (negative) pixels are more than the positive pixels (lesion pixels). Imbalance of data falls into the general topic of biases in data sets, which is discussed in detail in (Glauner et al. 2017, 2018a,b). Because the negative pixels are the majority, the algorithm may tend to output all the image as negative, and hence, a new variable, w, is introduced in the

BCE formula to compensate:

$$L(y, \hat{y}) = -(1 - w)y \log \hat{y} - w(1 - y) \log(1 - \hat{y}). \quad (3)$$

This is a general balance factor, taking into account just the positive samples for each class, and as such, all medical image volumes participate in the process. Also, the different factors take into account only those images which contain the class (Bellver et al. 2017).

3.2 Using Unity and WebRTC to Deliver PC Rendering Power to HoloLens

In a 2D plane, a pixel has one length in the x direction and one length on the y direction. By adding a third dimension to the pixel, a 3D volume object is created, which is called a voxel. Each pixel has a grayscale value called a Hounsfield unit (HU), which in medical imaging is a function of tissue composition. For example, the water's HU is zero, and soft tissue like the brain, kidney, and muscle has an HU between 30 and 40. Bones can have an HU of 400 while air (less dense than water) has -1000.

Being made out of voxels, medical imaging data volumes can be rendered by using a technique called raycasting. In Unity, to a simple cube geometry, is assigned a material which in its turn loads a shader. A shader is, in fact, a program with the difference that it is written for and it runs on the GPU. The shader loads the texture, in our case the 3D volumetric image made out of voxels, and tells the GPU how to render the object, in this case using the technique often called volumetric raycasting, ray-tracing, or ray marching. As depicted in Fig. 5, for each pixel of the final image, a "ray" is sent through the volume and the values of the nearby pixels intersected by the ray are interpolated to compute the final image pixel.

As illustrated in Fig. 6, a usual medical volume size in the DICOM "Digital Imaging and COmmunication in Medicine" format is about $512 \times 512 \times 1024$ voxels. That is more than 250 million voxels, hence computationally intensive in terms of GPU performance. Therefore, we needed to externalize these computations to a dedicated Windows server machine.

As a result, our visualization architecture comprises three interconnected applications, running in the same time: the HoloLens client, the Windows desktop server, and a Signaling Server which manages the communication and connection between the first two, as depicted in Fig. 7.

Our solution makes use of the "3D Toolkit" which in turn uses the Web Real-Time Communications (WebRTC) protocols and API as well as the NVEncode hardware encoding library from NVIDIA. The 3D Streaming Toolkit provides server-side libraries for remotely rendering 3D scenes, client-side libraries for receiving streamed 3D scenes, low-latency audio and video streams using WebRTC, as well as high-performance video encoding and decoding using NVEncode (Ermilov et al. 2019).

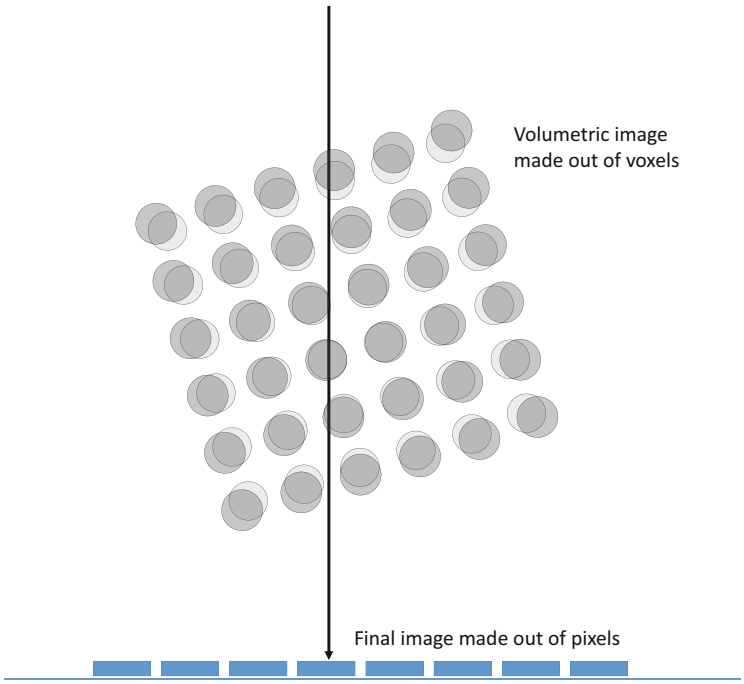

Fig. 5 Volume raycasting. Source: authors

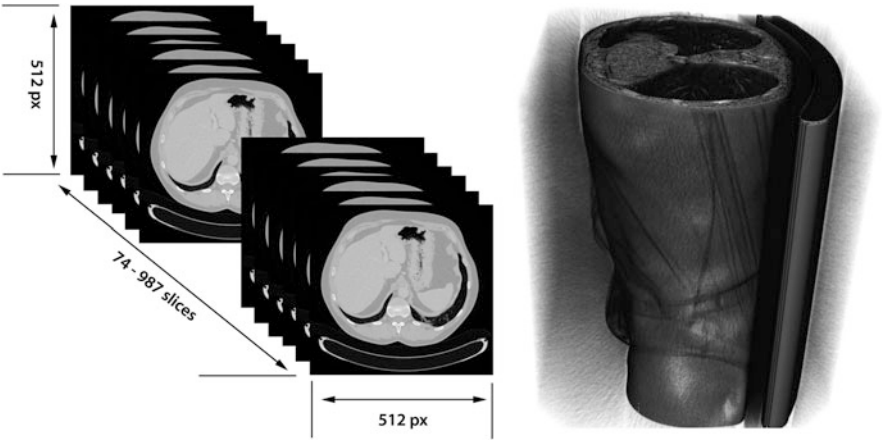

Fig. 6 Structure of data volumes used (left), and actual 3D visualization (right). Source: authors

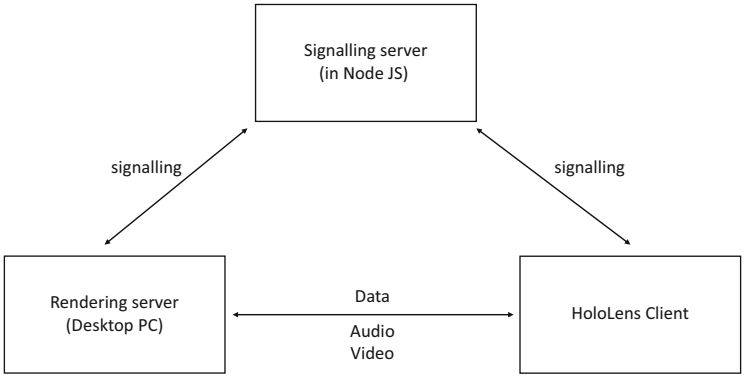

Fig. 7 System logical architecture. Source: authors

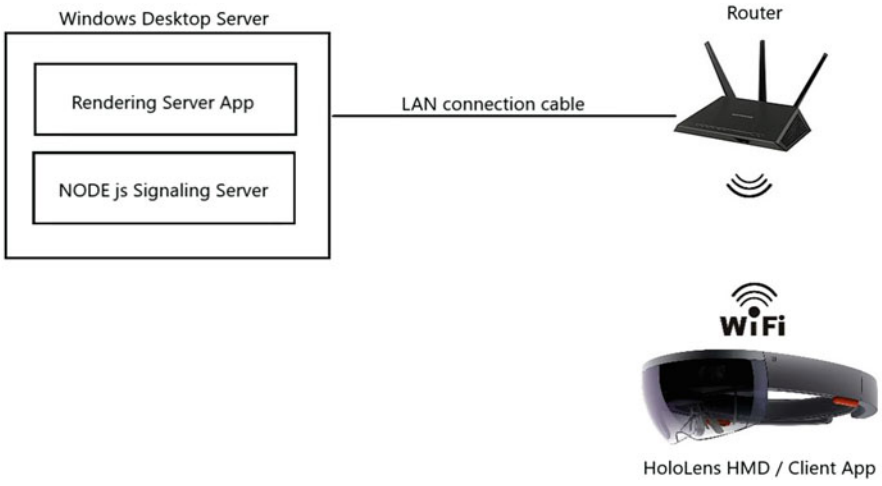

Fig. 8 Hardware architecture. Source: authors

The hardware architecture, presented in Fig. 8, comprises 3 components: a router, the Desktop Windows Server (hosting the Rendering Server app and the Signaling Server app), and the HoloLens HMD running the DirectX HoloLens client.

3.2.1 Server

The server is built using the Unity game engine and is meant to offload the heavy GPU rendering task from the HoloLens client. It is meant to run in a Windows OS, and makes use of the following technologies:

- NVIDIA drivers and CUDA library to render and encode the scene frames which will be sent to the HoloLens client. Most NVIDIA graphics cards include dedicated hardware for video encoding, and NVIDIA's NVEncode library pro-

vides complete offloading of video encoding without impacting the 3D rendering performance.

- The WebRTC open source project, released by Google in 2011 for the development of real-time communications between apps, including low-latency VOIP audio and video applications. Communication between peers is managed through one or more data channels. The Video Engine comes as a middleware service establishing a video data channel and automating buffer, jitter, and latency management. The Audio Engine does the same in regard to audio transmission and is conceived for efficient processing of voice data. Applications can open different data channels for custom messages (Ermilov et al. 2019).

The 3DStreamingToolkit's additions to the typical WebRTC usage are:

- The NVIDIA NVEncode hardware encoder library for real-time encoding of 3D rendered content was added to the video encoders.
- A dedicated data channel manages the camera transforms and the user interaction events. This channel is used to update the HoloLens camera position in the rendering server when the user moves through the room.

These were implemented by means of plugins that engage Unity or native DirectX rendering engines. The Unity server makes use of a native plugin produced by the 3D Toolkit build pipeline. The plugin negotiates with clients to configure a stream, and for encoding and sending visual frame data from Unity to the client. The core scripts employed by the server are the *StreamingUnityServerPlugin*, which provides a wrapper around the native plugin that powers the experience. An instance of the wrapper is created by the *WebRTCServer*, and exposed publicly. The *WebRTCServer* is the main WebRTC component, which configures the native plugin and handles client input data. Finally, the *WebRTCServerDebug* enables detailed logging data on request.

3.2.2 Client

We intended to use the Unity HoloLens client, but because it does not feature frame prediction yet, we decided to switch to the DirectX (DX) client. The DX client connects to the Signaling Server for handshaking, to finally establish a peer-to-peer connection with the Rendering Server via WiFi in order to receive the rendered frames as a stream, and send back to the Rendering Server updates concerning the HMD's position and rotation via the dedicated data channel. The Rendering Server updates the view per the newly received coordinates of the HoloLens HMD in the world. A UI allows movement, rotation, and scaling of the 3D medical volumetric image visualized. The scripts involved are acting directly on the data cube. Another set of scripts which activates at the shader level allows volume slicing and image luminosity adjustments.

After also considering other options, we have decided that for our high-requirement medical purposes, combined with latest GDPR laws in Europe, a dedicated local Windows desktop server equipped with an NVIDIA GTX 1080

GPU and an AMD octa-core CPU from the FX 8000 family running at 3.5 GHz, and communicating with the client over local WiFi, would be the solution offering the best mix of security, reliability, scalability, and performance.

4 Evaluation and Discussion

The liver data set is obtained from the Liver Tumor Segmentation (LiTS) challenge (Christ 2017) and consists of 131 CT scan volumes for training and validation, and 70 volumes for testing. We carried out our evaluation using two different graphic cards, GTX950M and GTX1080. The training takes 135 h and 30 h, respectively. Using the trained model, the average time to segment a volume takes 159 and 35 s using these GPUs, respectively. For the example volume depicted in Fig. 1 (center), our methodology segments the liver. The output is depicted in Fig. 1 (right). The metric used for assessment is the dice score (Aljabar et al. 2009):

$$\text{DICE\_score} = 2\frac{|P \cap T|}{|P| \cup |T|}, \tag{4}$$

which measures the similarity between P, the volume predicted by the network and T, the "truth" volume manually segmented by a radiologist, respectively. This metric is inspired by the F1 score. When testing on the LiTS test volumes, we achieve a liver dice score of 0.936. Also, the lesion segmentation test dice score is 0.586. For comparison, Table 1 presents the challenge results.

For the liver segmentation, we notice that more than half of the models participating in the challenge perform close to the maximum score. DRIU was originally intended to segment the blood vessels of the eye. We are therefore impressed that our DRIU-based model also performs in that range without having to do any customizations specific to the segmentation of the liver. As we work on a unified approach including visualization, we are satisfied with the scores so far. Furthermore, as the algorithm is derived from DRIU which was able to perform also the blood vessels' segmentation, the model appears promising to perform similarly well on the segmentation of other organs or body parts like arteries, in the future.

While we are used to encountering in our machine learning tasks much larger data sets, the data sets specific to this task are usually much smaller. For example, the original authors (Bellver et al. 2017) have successfully trained and tested this algorithm also on the "Visual Concept Extraction Challenge in Radiology" (Visceral) data set (Hanbury 2018) which comprises only 20 volumes, out of which

Table 1 LiTS challenge: liver and lesion segmentation dice scores (Christ 2017)

Dice score quantile	0.025	0.25	0.5	0.75	0.975	Best
Liver	0.043	0.927	0.943	0.959	0.967	0.97
Lesion	0.289	0.498	0.613	0.643	0.699	0.702

18 were used for training and 2 for validation. The reasons for such small data sets are related to the difficulty of producing these labeled volumetric images data sets, data privacy, and data size.

In regard to visualization, we started from less than 1 frame per second (fps) for rendering our test volume of approximately 260 million voxels directly on the HoloLens' GPU. Finally, the average frame rate obtained by rendering the same volume using a raycasting volume rendering technique with a "brute force" shader (unoptimized) on GTX1080 GPU was 23 fps, with a latency of 250 ms. The experience involves stereoscopy, which means the scene is rendered twice, once for each eye. When viewed on a dedicated headset, the two rendered images are combined by the brain to give the perception of 3D depth. The 23 fps obtained are stereo fps, hence the GPU has rendered in fact 46 mono fps. Using 2 GTX 1080 in SLI mode should almost double the stereo fps to around 45. As such, we are convinced that this is a pertinent, scalable solution, whose performance can be improved with further optimizations. Other possibilities to increase performance except using superior GPUs would be adding shader improvements like:

- GPU cache efficiency and memory access: as the volume data is linearly loaded into memory, a ray that is cast through the volume has poor chances to have the fastest possible memory access to neighboring voxels information as it traverses the volume. This can be improved by converting the layout to a block-based layout, as shown in Fig. 9. Thus, as we travel through the volume, we will be more likely to faster access in memory the neighboring voxels.

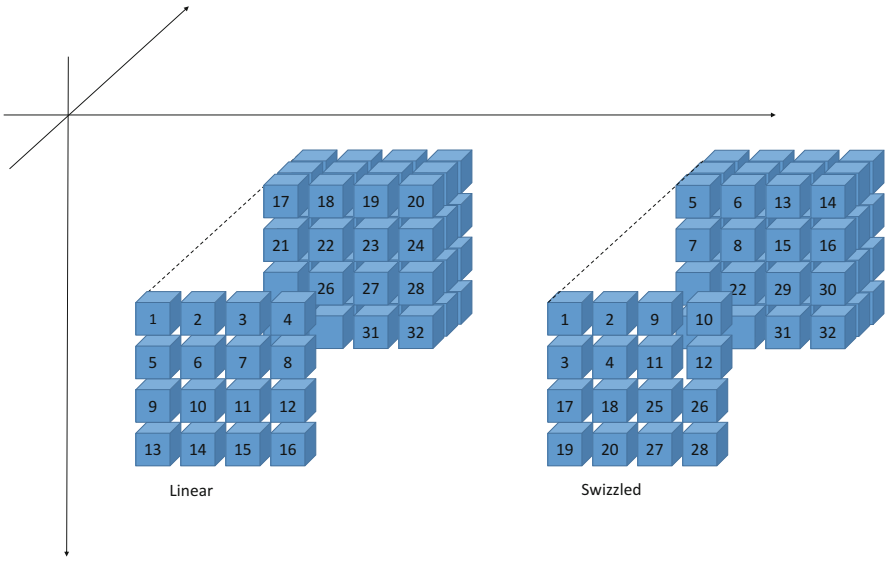

Fig. 9 Improving memory and cache efficiency. Source: authors

- Eliminating the "0 alpha" voxels: right now, we are ray-casting through the whole volume in the scene, i.e., all voxels, even if the volume contains areas with zero alpha voxels. The zero alpha voxels can be skipped and we can render only the volume parts which contain the see-able voxels (alpha non-zero).

Although not experimented during this project, the 3D Toolkit authors state that an unlimited number of peers can connect to a single instance of a server. However, this number will eventually be limited by the hardware, as NVIDIA enforces a maximum of 2 GPU encoding sessions on desktop series graphics cards. The number of peers can be limited via signaling, more specifically, through configuring a ".json" file.

While cost evaluation is not an obvious task in this case, we can state though that the liver segmentation is a complex procedure which can take up to 90 min for a single patient (Gotra et al. 2017), whereas the ML algorithm is able to perform this task in less than 30 s on a good GPU. Being able to "see inside" the patient should also shorten the planning and operation times, improve the outcome, and decrease the odds of a subsequent intervention.

5 Conclusions and Outreach

In this chapter, we proposed a solution for automated segmentation of medical images in order to both reduce health care costs and improve patient care. We first described the model that employs convolutional neural networks. We have then evaluated it on the segmentation of livers and lesions and demonstrated promising results on 200 scans provided through an on-going challenge. Second, we have included our deep learning algorithm in a holistic approach to support doctors. Our system visualizes the segments in a Microsoft HoloLens and thus allows doctors to intuitively look at the holographic data rather than using 2D screens. We have also achieved both, fast training and a promising fps rate, using GPUs and thus we deliver a ready-to-use system. We are currently preparing an evaluation from medical doctors and radiologists in a real-world environment. While the liver segmentation is automatic, some minor steps, such as transferring or converting data are still manual. Therefore, our future work will fully automate the pipeline.

We are also planning to evaluate the usage of multiple HoloLenses in parallel for one hologram, a cooperative working feature that is of importance to doctors in operating rooms. While being just a sample of how new technologies like machine learning and augmented reality can be harnessed towards improving effectiveness, cost efficiency, quality, and safety in a real-life scenario, this project can also be an indicator as of how business and technology together can disrupt an existing status quo on a given market.

Alongside the medical field, other fields where AR and VR seem promising so far are the video games industry, the military, education, real-estate, engineering, retail, events, and entertainment (Jung 2019). For 2025, their predicted market shares are 33% for video games, 15% for healthcare, 13% for engineering, 12% for live

events, 9% for video entertainment, and 7.5% for real estate (Bellini et al. 2016). As more companies enter the space and ramp up the work to create the software, hardware, and services, the combined VR and AR units shipments is forecast to grow up to 65.9 million units. In the AR market, the drivers will be the Microsoft HoloLens and Magic Leap One. Tethered (connected by cable to PC or smartphone) devices will tend to form the low-cost market segment. A 2018 forecast published by the International Data Corporation (IDC) estimates a combined market potential of shipping more than 60 million AR and VR devices in 2022 (International Data Corporation 2018).

References

Aguirre, S. (2019). *True 3D Viewer*. http://www.echopixeltech.com [Online]. Accessed 9 Sep 2019.

Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J.V., & Rueckert, D. (2009). Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage, 46*(3), 726–738.

Bellini, H., Chen, W., Sugiyama, M., Shin, M., Alam, S., & Takayama, D. (2016). *Equity Research – Profiles in Innovation. Virtual & Augmented Reality. Understanding the race for the next computing platform.* http://www.goldmansachs.com/insights/pages/technology-driving-innovation-folder/virtual-and-augmented-reality/report.pdf [Online]. Accessed 19 Aug 2019.

Bellver, M., Maninis, K.-K., Pont-Tuset, J., Giró-i-Nieto, X., Torres, J., & Van Gool, L. (2017). Detection-aided liver lesion segmentation using deep learning. arXiv:1711.11069v1.

Beydoun, A., Gupta, V., & Siegel, E. (2017). DICOM to 3D Holograms: Use Case for Augmented Reality in Diagnostic and Interventional Radiology. In *SIIM scientific session posters and demonstrations*.

Calhoun, P., Kuszyk, B., Heath, D., Carley, J., & Fishman, E. (1999). Three-dimensional volume rendering of spiral CT data: Theory and method. *Radiographics: A Review Publication of the Radiological Society of North America, Inc., 19*(3), 745–764.

Carmigniani, J., & Furht, B. (2011). Augmented reality: An overview. In B. Furht (Ed.) *Handbook of Augmented Reality*. New York: Springer.

Christ, P. (2017). *Liver Tumor Segmentation Challenge (LiTS)*. http://competitions.codalab.org/competitions/17094 [Online]. Accessed 18 Jul 2018

Christ, P., Ettlinger, F., Grun, F., Elshaera, M. E. A., Lipkova, J., Schlecht, S., et al. (2017). Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. arXiv 1702.05970.

Douglas, D., Venets, D., Wilke, C., Gibson, D., Liotta, L., Petricoin, E., et al. (2018). Augmented reality and virtual reality: initial successes in diagnostic radiology. In *State of the art virtual reality and augmented reality knowhow*. London: IntechOpen Limited.

Douglas, D., Wilke, C., Gibson, D., Boone, J., & Wintermark, M. (2017). Augmented reality: advances in diagnostic imaging. *Multimodal Technologies and Interact, 1*, 29.

Ermilov, A., Gibson, T., Cao, P., Greenier, B., & Zolochevska, A. (2019). *Real-Time Streaming of 3D Enterprise Applications from the Cloud to Low-Powered Devices*. http://www.microsoft.com/developerblog/2019/03/19/real-time-streaming-of-3d-enterprise-applications-from-the-cloud-to-low-powered-devices/ [Online]. Accessed 3 Sep 2019.

Fishman, E., Ney, D., Heath, D., Corl, F., Horton, K., & Johnson, P. (2006). Volume rendering versus maximum intensity projection in CT angiography: What works best, when, and why. *Radiographics: A Review Publication of the Radiological Society of North America, Inc., 26*(3), 905–922.

Glauner, P., Migliosi, A., Meira, J. A., Valtchev, P., State, R., & Bettinger, B. (2017). Is big data sufficient for a reliable detection of non-technical losses? In *2017 19th International Conference on Intelligent System Application to Power Systems (ISAP)*.

Glauner, P., State, R., Valtchev, P., & Duarte, D. (2018a). On the reduction of biases in big data sets for the detection of irregular power usage. In *Proceedings of the 13th International FLINS Conference on Data Science and Knowledge Engineering for Sensing Decision Support (FLINS 2018)*.

Glauner, P., Valtchev, P., & State, R. (2018b). Impact of biases in big data. In *Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2018)*.

Gotra, A., Sivakumaran, L., Chartrand, G., Vu, K. N., Vandenbroucke-Menu, F., Kauffmann, C., et al. (2017). Liver segmentation: indications, techniques and future directions. *Insights Imaging, 8*(4), 377–392.

Hanbury, A. (2018). *Visceral*. http://www.visceral.eu/contact-us [Online]. Accessed 19 Jul 2018.

Heath, D., Corl, F., Horton, K., Fishman, E., & Johnson, P. (2006). Volume rendering versus maximum intensity projection in CT angiography: What works best, when, and why. *Radiographics: A Review Publication of the Radiological Society of North America, Inc., 26*(3), 905–922.

Hohne, K., Bomans, M., Tiede, U., & Riemer, M. (1988). Display of multiple 3D-objects using the generalized voxel-model. In *Medical Imaging II. Newport Beach: SPIE*.

Hoogi, A., Beaulieu, C. F., Cunha, G. M., Heba, E., Sirlin, C. B., Napel, S., et al. (2017a). Adaptive local window for level set segmentation of CT and MRI liver lesions. *Medical Image Analysis, 37*, 46–55.

Hoogi, A., Lambert, J. W., Zheng, Y., Comaniciu, D., & Rubin, D. L. (2017b). A fully automated pipeline for detection and segmentation of liver lesions and pathological lymph nodes. arXiv 1703.06418.

International Data Corporation (2018). *Augmented Reality and Virtual Reality Headsets Poised for Significant Growth, According to IDC*. http://www.idc.com/getdoc.jsp?containerId=prUS44966319 [Online]. Accessed 20 Jul 2018.

Johnson, P., Heath, D., Kuszyk, B., & Fishman, E. (1996). CT angiography with volume rendering: Advantages and applications in splanchnic vascular imaging. *Radiology, 200*(2), 564–568.

Jung, T. (2019). The power of AR and VR for business. In M. Claudia, T. Dieck, & T. Jung (Eds.) *Augmented reality and virtual reality*. Berlin: Springer.

Maninis, K.-K., Pont-Tuset, J., Arbelaez, P., Van Gool, L. (2016). Deep retinal image understanding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 140–148). Berlin: Springer.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Berlin: Springer.

Rothman, D. (2018). *Artificial intelligence by example: Develop machine intelligence from scratch using real artificial intelligence use cases*. Birmingham: Packt Publishing.

Digitalization in Mechanical Engineering

Michael Thurner and Patrick Glauner

Abstract

A high level of industrial automation of repetitive tasks allows companies to efficiently produce products at large scale. Digitalization is the subsequent step of industrial automation and aims to further reduce costs and waiting times. Digitalization also aims to automate individual decision making. Key to both goals is to transform business processes from the analog to the digital world and then to analyze and thus to take advantage of digitized information. In this chapter, we provide an intuitive introduction to digitalization in mechanical engineering. We then present various business opportunities and discuss the related challenges. Next, we propose how mechanical engineering companies need to align their mindset with the digital transformation. Last, we present some of our works on digitalization in mechanical engineering and share a number of best practices. As an outcome, you will be able to employ digitalization in order to create real value in your business. That increase of efficiency will allow you to remain competitive in an environment that keeps becoming more and more competitive.

1 Introduction

Industrial automation has become a reality during the last 200 years. That transformation has happened in multiple stages, starting from mechanization through mass production using electricity to applications of electronics and information technology. As an outcome, a large number of repetitive tasks are automated in

M. Thurner (✉)
Regensburg, Germany

P. Glauner
Deggendorf Institute of Technology, Deggendorf, Germany
e-mail: patrick@glauner.info

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_9

a number of industrial sectors, including mechanical engineering. **Digitalization** is the next step in this chain. It aims to not only further reduce costs, resource allocation, and waiting times. In addition, digitalization also has the goal to automate multifaceted decision making, a discipline in which humans excel.

In recent years, the three terms **digitization**, **digitalization**, and **digital transformation** have been coined and are often used interchangeably. Yet, all of them employ digital information processing. The three terms are distinct and have different meanings. In addition, they build on top of each other as depicted in Fig. 1. We therefore aim to provide meaningful and delimiting definition as follows:

Definition 1 Digitization: Transitioning business processes from analog to digital.

Definition 2 Digitalization: Analyzing and taking advantage of digitized information.

Definition 3 Digital Transformation: Transforming the entire value chain of a company using digitalization applications. This includes creating completely new business models.

The concept of digitization is therefore not new, as it has been around since approximately the 1960s. In that time, computers were introduced in various industries, including mechanical engineering, to store and process data. Digitalization allows us to take advantage of and analyze that data to further speed up processes and cut costs. The term digital transformation is also often referred to as **industry 4.0** and also has applications in other fields such as power engineering (Glauner 2019).

In this chapter, we will provide an extensive introduction to the field of digitalization from a mechanical engineering perspective. First, we show how digitalization already creates value in the mechanical engineering industry. Second, we will highlight future business opportunities and discuss related challenges. Digitalization in mechanical engineering could thus offer the same degree of impact like mechanization and electrical power did about two centuries ago. Last, we

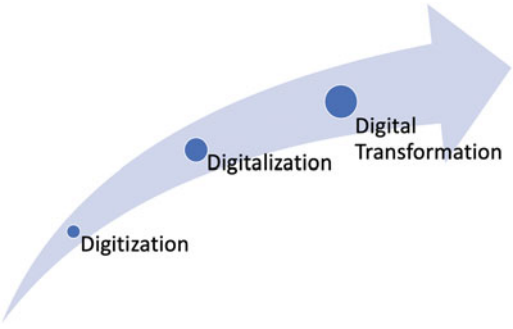

Fig. 1 Digitization, digitalization, and digital transformation. Source: authors

present some of our works on digitalization in mechanical engineering and share a number of best practices.

2 Comparison to Traditional Industrial Automation

It is important to highlight that the three terms digitization, digitalization, and digital transformation are diametrically different to traditional industrial automation, as the latter entirely aims to automate physical behavior employing sensors and actors. One prominent example includes programmable logic controllers (PLCs). The scope of PLCs is mainly to turn sensor values into actor signals in (near) real time (Bolton 2015). PLCs are essentially a transistor-based variant of the previously used relays that employ simple binary (yes/no) rules for controlling physical processes, automating or moving actors. In contrast, PLCs initially were not meant to store a large amount of data for post-mortem analyses. Even if today's PLCs may process data and provide functions for digitalization, the main concept of a PLC will restrain your digital transformation projects. Further elaboration on this topic can be found in this chapter.

3 Opportunities and Challenges

Due to new technological opportunities, a vacuum of possibilities and challenges has been created in mechanical engineering: Digitization, digitalization as well as the digital transformation may be performed by many different players on the market. Typical mechanical engineering companies can develop and turn into digital performers. Likewise, leaders in information technology and tech companies could conquer a new market share and fill this vacuum. Tech companies put pressure on this open field while manufacturers try to grow from the down side up to technological strength as depicted in Fig. 2. According to McKinsey, 58% of manufacturers and 84% of suppliers expect outside competitors to enter manufacturing industries (Wee et al. 2015).

A major part of this vacuum will be filled by artificial intelligence (AI) applications. AI allows to automate human decision making, mainly by examining historical data and finding statistical patterns in it. There is an enormous potential for AI in engineering as a whole as argued by McKinsey (Bughin et al. 2017). You may wonder whether AI will be relevant for you so soon. In the coming years, however, completely new competitors will emerge. These competitors will be strong in AI as AI allows to automate human decision making. Therefore, the AI leaders of today will the winners of tomorrow. Most likely, these competitors will be based in China. You can learn more about China's AI innovation ecosystem and its strong support from both the government and industry in Kai-Fu Lee's book "*AI Superpowers: China, Silicon Valley, and the New World Order*" (Lee 2018). Lee's book is both, encouraging and shocking in our opinion.

Fig. 2 Existing vacuum with opportunities and challenges for manufacturers and tech companies. Source: authors

Later in this chapter, we describe some of the cutting-edge AI-based applications we work on. These go far beyond predictive maintenance, the automated determination of equipment's condition. We often feel that most stakeholders in AI for mechanical engineering mainly think that predictive maintenance was the holy grail to reach. However, in our opinion, predictive maintenance only scratches the surface of the potential that AI provides.

In the previously defined vacuum, mechanical engineering companies are challenged by the drop of prices and a shift of their value chain. Suppliers provide them with out of the box solutions and their own in-house production depth shrinks. Customers see them mainly as steel suppliers and assembly companies, but not anymore as mechanical engineering solution providers. Automation is fulfilled by suppliers and has become more easy to implement in the last years. As an example let us point out various mechanical cam disks in synchronous machines. Fifty years ago that was high level engineering while today software cams are implemented easy and fast on different PLCs. Added value can today be offered with new digital features and not by slightly faster machines. Further challenges are placed in existing companies by established parameters. These consist, for example, of the workforce with its skills and prejudice as well as internal or external IT services. Frequently, we get to know companies where development aside old trails and thinking outside of the box is not desired.

In order for digitalization to succeed in mechanical engineering, companies need to invest in training their entire staff to think in terms of digitalization. In addition, companies need to rethink their information technology. This includes hiring a Chief Digital Officer (CDO) who works independently of the Chief Information Officer (CIO). In essence, the CDO heads a company's digital department which is agile and uses its own infrastructure in order to quickly build and deploy prototypes. We provide more details and best practices on these topics in the chapter by Glauner.

4 The Way of Thinking: People, Processes, and Technology

Let us imagine the digital transformation as a three-legged table. Leg one represents the **people** working either actively or passively on or with digitalization. The second leg represents all internal and external **processes** around digitalization. The last leg stands for the most obvious part of digitalization: **technology**. These three legs of a successful digital transformation have to grow similarly and well-balanced.

Why do we need this metaphor of a table for digitalization in mechanical engineering? Because you can see the same slate table in too many companies. In the following, we take a deeper look into the way we need to see digitalization and what way of rethinking is necessary to fulfill our goals.

People are the number one factor of every company, likewise for digitalization. In traditional mechanical engineering, artful machines and products are designed by engineers working for months or even years on one product in order to find an ideal solution. Neither is this the way how digital development works nor is this the path our human brain connects neurons (Sporns 2010). Digitalization and neurons link by trial and error, fast proofs of concepts and retrial. The first leg of our digitalization table is often built up of mechanical and electrical engineers who work as reskilled digitalization specialists. Now we come across the important way of thinking: Did a real mind change take place in the whole company to shift engineers from perfectionists of technically mature products to computer scientists that deliver quickly? Do they feel digitalization as the most important part of their job or is it just the next trend to follow halfheartedly? Is the main driver of your digitalization strategy the strong belief of your teammates or the fear of missing your competitor who sells fancy digital features? It is important that you do not attempt to force engineers to push digitalization. Better persuade the open-minded part of your team and let the others stick with their area of expertise.

Processes define how and in what efficiency people work in companies. As the responsible person for digitalization, you should avoid digitizing analog processes just as is. A benefit is only provided by rethinking the whole process as a digital process with added value, for example, for customers. Processes contain the risk of excessive use and complicate easy and fast solutions. Digitalization gives you the opportunity to rethink processes, trim them down to a reasonable scale, and highlight new benefits.

Technology is what you first think about when you hear digitalization. There exist variously shaped technologies for digitalization in mechanical engineering. In every sector, it is common to use for your digital transformation what you already buy and assemble. This step leads to a number of issues and challenges. Like already mentioned previously, PLCs are widespread regardless of the concrete supplier, hardware or software. They mainly work cyclic and are rarely interrupt triggered. All of them are designed to operate fast, predictable, and traceable under a clear amount of input data. This is contradictory to the concept of digitalization and creating value from a big amount of historical data. If you want to add value, make use of what already exists in computer engineering and data engineering:

Single board computers (SBC) provide sufficient processing power, the necessary amount of memory, calculation speed, and open source software for your digital transformation. Furthermore prices and availability are unbeatable. Even process control and software PLC functions can be handled by SBCs using real-time patches. Typical examples of SBCs in Rasberry Pi (Upton and Halfacree 2014) and Arduino (Banzi and Shiloh 2014). However, it is important to highlight that SBCs will not fully replace PLCs in the foreseeable future: Some legacy tasks as well as new tasks will be performed by SBCs. PLCs will keep an important role in traditional industrial automation, though.

5 Selected Use Cases and Applications

In this section, we present four use cases that we have worked on previously. We also share best practices that allow others to build their own digitalization use cases that create real value in mechanical engineering.

5.1 Reducing the Number of Simulation Runs

During R&D and order processing, mechanical engineering companies typically run a large number of different simulations, for example, to evaluate machine behavior and reliability under different conditions. However, running simulations often requires large computational resources and comes with long waiting times for them to run. Advances in high performance computing make it possible to speed up simulations. However, as a consequence, companies usually run even more simulations. As a consequence, companies still need a large amount of resources for their simulations.

We suggest to entirely rethink simulations: A lot of companies have been running variants of the same type of simulations for many years. From a data science point of view two questions arise:

1. Do we really still need to run that many simulations?
2. Can we use the results of previous simulations in order to predict the outcome of simulations without running them?

We provide an introduction to the field of artificial intelligence and its subfield machine learning in the chapter by Glauner. We encourage you to read that chapter if you want to learn more about those fields. In essence, machine learning gives computers the ability to find (“learn”) patterns from data. These patterns are then used to make decisions based on new inputs (Bishop 2006; Russell and Norvig 2009).

We have built an artificial intelligence that employs machine learning to predicting the outcome of simulations. For that, we find correlations between inputs (parameters) and results of previously run simulations. Once those are found, the

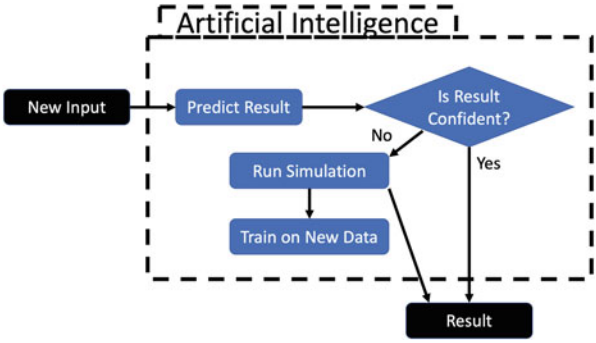

Fig. 3 Workflow of reducing the number of simulation runs. Source: authors

patterns can be used to efficiently predict simulation results within milliseconds. We do so before running a simulation. However, machine learning typically employs statistics for finding patterns. As a consequence, some predictions may be wrong. Our AI therefore also reports the confidence of its prediction. If it is very confident about its prediction, we report the prediction result. In contrast, for less confident predictions, we eventually run the actual simulation and report the simulation result. The outcome of newly run simulations is fed back to the AI which incrementally becomes better. We depict this workflow in Fig. 3.

We have applied this framework to a fluid dynamics simulation that is run approximately 1000 times a year. We have used data from the last 2 years for the training of the AI. We can now instantaneously and confidently predict the outcome of 2/3 of the simulations. We only need to run the remaining 1/3. Our approach can be applied to a large number of other types of simulations that are typically used in mechanical engineering.

5.2 Intelligent Mechatronic Modules: Cyber-Physical Systems

Intelligent mechatronic modules are the reinvented combination of mechanic and electric parts building up a reasonable entity. The concept of an entity from the perspective of intelligent mechatronic modules is diametrically different to traditional mechanical engineering. In the first industrial wave, the focus was on invention and speed improvement in the final product. The second wave brought production and the improvement of production into focus. Now in the digitalization phase, the functionality of a single module is focused. This change has been triggered by having reached physical limitations and highlighting economic reasons to produce more efficiently and in higher quality.

With a change towards function-oriented modules, we emphasize human and non-human interaction with every intelligent mechatronic module. So, the main feature of every intelligent mechatronic module is the ability to interact and the

expected benefits of interaction. More about these benefits are described in the next use case of Self-X.

As a best practice, we recommend to not simply cut existing machines into modules and retrofit them into intelligent mechatronic modules. This will lead to higher costs without added value. The other way round, you should build the machine up from new intelligent modules and focus the main aims of the machine and highlight how different modules could support in the complete product life cycle. Independent modules can save time in each of these steps and bring the value faster to the customer.

5.3 Self-X and Organic Computing

The field of "organic computing" is an interdisciplinary research field which studies how complex, distributed systems work and how they can be controlled (Würtz 2008). The term "organic" refers to the distributed behavior that is composed of parts. Self-X is a part of organic computing. This term is a verbal composition of the word "self" and further different supplements represented by X. This technology targets machines interacting in organic behavior, e.g., self-organization, self-configuration, self-healing, self-optimization, or self-protection (Schmeck 2005). Transferred to mechanical engineering, intelligent modules of a cyber-physical system could have the following characteristics:

- Self-X is a tool to link and interface intelligent modules in a network based on mechanical, electrical, and software behavior.
- Self-X modules communicate over a standardized interface.
- Self-X is not only a technology but also an integrated part of every product's life cycle management.
- Self-X needs a hardware and software component in each module.
- Self-X correlates with modularization and decomposition in mechanical engineering to generate small controllable components. This property is referred to as "divide and conquer" in computer science.

Providing these characteristics, Self-X also enables self-configuration. That is the property of technical self-awareness and interaction with other modules to gain a stable condition without operator interference. Self-organization is enabled by communication about necessary circumstances for each module to find an ideal solution. Self-optimization can be delivered combining the former capabilities. Self-healing and self-protection can be implemented using the knowledge of each module what is best and necessary to function in a machine network. If not all functions are available, the module can search for other best possible solutions.

In mechanical engineering, Self-X could drive the further development on machines. Imagine a production site as simple to run and maintain like a current smartphone. Without deep knowledge of the production technology, the modules

and machines provide all needed information for themselves, for self-configuration, and for interacting humans. Self-optimization would allow lines to reconfigure themselves in order to become more efficiently.

We have investigated the field of Self-X in mechanical engineering and built various prototypes that will inspire future product development. These include a three-axes milling machine which is composed of intelligent mechatronic modules. One can easily plug in and plug out individual axes without having to reconfigure the software or settings. Concretely, the axes communicate with each other and understand what they can do together. We have also investigated other approaches such as mobile robotics for self-healing lines by removing broken or misplaced objects from a line. This approach allows to efficiently resume production without the need of manual human interaction.

5.4 Automatically Layouting New Machine Variants

Variant management is a matter of many subfields of mechanical engineering. In special purpose machinery, variant management has a particularly strong importance since every machine built and sold may slightly be different to the previously sold ones. This has also impact on quoting in which an approximate layout of a new machine variant is needed for estimating prices, costs, and delivery times before actually building the new variant.

Choosing the layout as well as estimating the consumption of energy and other physical resources of a new variants is challenging. This is typically done by using one of the following approaches:

- Deriving equations: Domain experts derive equations (or rules) that incorporate physical machine properties subject to customers' requirements. The derived concepts can then be applied to new customers' requirements in order to predict the corresponding variant layout. This approach is challenging as these physical properties are often not fully understood by experts and can thus hardly be formalized.
- Comparing new variant requirements to previously sold machines: Domain experts manually check databases that contain entries of previously sold machines, their layouts, and corresponding customers' requirements. Once related variants are found, domain experts estimate how the difference to the new customers' requirements affects the layout of the new variant. While this approach is feasible, it is time-consuming, expensive, error-prone, and irreproducible in practice.

Our novel approach utilizes machine learning to find correlations between previously sold machines and customers' requirements. These patterns essentially reflect physical properties of machines and can now be used for predicting the layout of new machine variants.

We have applied this framework to the quoting process of stretch blow molding machines that turn so-called preforms into PET bottles. We can now instantaneously and reliably predict the consumption of resources such as electricity, heating, pressure, and cooling of new stretch blow molding machine variants subject to customers' requirements. Following the aforementioned approaches, doing so had previously required a lot of manual decision making by domain experts that took days, was expensive, and proved to be error-prone and irreproducible.

6 Conclusions

The first part of this chapter provides an introduction to digitalization in mechanical engineering. We distinguished digitalization from traditional industrial automation. While the latter mainly focuses on increasing repetitive tasks by controlling actors using sensor values and binary logic, the former aims to take advantage of and analyze stored digital information in order to create new value. In the second part, we provided a review and analysis of opportunities for digitalization in mechanical engineering as well as the respective challenges. In summary, digitalization in mechanical engineering allows companies to reduce costs and increase efficiency by automating human decision making. We also described how mechanical engineering companies need to align their mindset with the digital transformation by rethinking how people, processes, and technology interact. In the third part of this chapter, we described four use cases we have previously worked on: reducing the number of simulations, intelligent mechatronic modules, Self-X, and automatically layouting new machine variants. Some of those use cases are based on artificial intelligence and machine learning. We also presented a number of best practices on how to apply the underlying methodology to other use cases.

References

Banzi, M., & Shiloh, M. (2014). *Getting started with Arduino: The open source electronics prototyping platform*. San Francisco: Maker Media.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Bolton, W. (2015). *Programmable logic controllers*. Boston: Newnes.

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., et al. (2017). Artificial intelligence: The next digital frontier. McKinsey Global Institute.

Glauner, P. (2019). Artificial intelligence for the detection of electricity theft and irregular power usage in emerging markets. PhD Thesis, University of Luxembourg, Luxembourg.

Lee, K.-F. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. Boston: Houghton Mifflin Harcourt.

Russell, S. J., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River: Prentice Hall.

Schmeck, H. (2005). Organic computing - a new vision for distributed embedded systems. In *Eighth IEEE International Symposium on Object-Oriented Real-Time Distributed Computing (ISORC'05)* (pp. 201–203).

Sporns, O. (2010). *Networks of the brain*. Cambridge: MIT Press.

Upton, E., & Halfacree, G. (2014). *Raspberry Pi user guide*. Hoboken: Wiley.
Wee, D., Kelly, R., Cattell, J., & Breunig, M. (2015). Industry 4.0 how to navigate digitization of the manufacturing sector. McKinsey Digital.
Würtz, R. (2008). *Organic computing - understanding complex systems* (1st edn.). Berlin: Springer.

Lean Launch Data Engineering Projects with Super Type Power

Kenny Zhuo Ming Lu

Abstract

Data is ubiquitous. Many modern software and data analytics applications rely on robust and quality datasets. Data engineering becomes a common pipeline in systems running in start-up and enterprise businesses. Data engineering projects in the past were perceived as a set of programming scripts which were typically in a "build-then-scrap" cycle. As the data analytics applications became parts of the main trends, such projects require a serious planning and development to minimize the overhead of integration and maintenance due to scaling up. In this article, we discuss how to use type systems and formal methods to reduce these overheads.

1 Introduction

Lately in the software industry, the need of data engineering is on the rise due to the increased adoption of AI and data analytics applications.

Like any other software engineering applications, data engineering application requires a sound design, a correct implementation, and thorough testing to ensure the quality of the end products. In the presence of big data, many modern data engineering applications often incur resource overhead. One of the cost centers come from the development overhead due to late-discovered or undiscovered errors. For instance, we would like to eliminate most of the run-time errors, in particular those in the latter stages of data processing pipelines. It is obvious that errors arising at these stages are extremely costly due to higher computation overhead and higher utilization of development resources such as man power and infrastructure. Testing with sampled/scaled-down dataset might not be sufficient to detect all errors. In this

K. Z. M. Lu (✉)
School of Information Technology, Nanyang Polytechnic, Singapore, Singapore

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_10

article, we propose to use some static analysis techniques with the assistance of the type systems, to eliminate errors in the data engineering applications as much as we could and retain the code readability, modularity, and re-usability.

This article is organized as follows. In Sect. 2, we walk through a simple batch data processing example in Spark and we illustrate how to use static type system to eliminate the need of extra run-time checking in the data processing stage and yet retain the code re-usability. In Sect. 3, we look into a different example in the streaming data context and study how to use phantom types and type constraints to eliminate the possible run-time errors. In Sect. 4, we will conclude this article with some further discussion.

2 Towards Type Safe and Reusable Spark Applications

As Apache Hadoop (Hadoop 2019) became the widely adopted system for big data infrastructure, Apache Spark (Spark 2019) became the most popular engine by lifting data parallelism to the next level. Spark offers a simple computation model which abstracts over a MapReduce-like data batch processing run-time. In many data engineering applications, it out-performs the "battery included" MapReduce library shipped with Apache Hadoop.

In this section, we will highlight some of the common pit-falls using Spark for data processing and the remedies. For convenience, we use Scala as the main language for the discussion, but the concepts are not language specific. We are aware of projects that attempting to enhance the type system of dynamic type languages such as Python (MyPY 2019). The main idea can be applied to these system/language extensions.

2.1 Loosely Typed Data

Let us consider a simple example which constructs a language model from some textual data using the two-gram operation.

```
1 def twogram(s:String):List[String] = {
2     val words = rmpunc(s).toLowerCase.split(" ").toList
3     val pairs = words.flatMap(w=>w.zip(w.drop(1)))
4     pairs.map(p=>s"${p._1}${p._2}").toList
5 }
```

Function `twogram` takes a paragraph of text, splits it into words, and stores them in a list as stated in line 2. For simplicity, we assume there exists some helper function `rmpunc` that removes punctuation from the string. For each word, we construct the two-gram character lists by zipping the word with itself having the first character dropped as stated in line 3. `s1.zip(s2)` forms a new sequence by pairing up elements from `s1` with those from `s2`. For example, `"abc".zip("123")` yields a sequence of `(a,1), (b,2), (c,3)` `.w.drop(1)` removes the first character in `w`. Let `l` as a sequence, method invocation `l.flatMap(f)` applies

the function `f` to each element in `l`. Assuming the results of `f` are sequence, the inner sequences are concatenated (or flattened). In line 4, we use `map` to apply the anonymous function `p=>s"$p._1$p._2"` to each pair in `pairs`. Given a pair `p`, `p._1` extracts the first component and `p._2` gives us the second one. `s"$x$"` embeds a Scala value referenced by variable `x` in a string. For instance, `twogram("hello world")` yields a list

```
1 List(he, el, ll, lo, wo, or, rl, ld)
```

Next, we load the input data from a CSV file into a Spark dataframe. Dataframe offers a high-level abstraction and operations to manipulate data. It is particularly appealing to Python programmers due to its similarity to Pandas dataframe.

```
1 import org.apache.spark._
2 import org.apache.spark.sql._
3
4 val sparkSession = SparkSession.builder().getOrCreate()
5 val mydata:DataFrame = sparkSession.read.csv("/data/input.csv")
6 mydata.show
```

At line 4 above we instantiate a spark session, followed by a read from a CSV file. The last statement allows us to take a look at the data, which likes the following.

```
1 +--------------------+----------+--------------------+
2 |              Source|      Date|                Text|
3 +--------------------+----------+--------------------+
4 | www.dailymail.co.uk|2019/09/05|Meet 'the most be...|
5 |       www.rdasia.com|2019/09/05|The most chilling...|
6 |   theindependent.sg|2019/09/05|Netizens lash out...|
7 +--------------------+----------+--------------------+
```

Next, we would like to apply function `twogram` to the text field from the above dataframe.

```
1  import scala.util.Try
2
3  def hasColumn(df: DataFrame, path: String) = {
4    Try(df(path)).isSuccess
5  }
6
7  def twogram_df(df:DataFrame):Option[Dataset[List[String]]] =
8  {
9    if (hasColumn(df, "Text")) {
10     val text_df = df.select("Text")
11     Some(text_df.map((row:Row) => twogram(row.getAs[String]("Text
           "))))
12   } else { None }
13 }
```

Function `twogram_df` applies function `twogram` to the "Text" column of the input dataframe. Note that we need to check for the existence of the "Text" column. In case of failure, a `None` value is returned. Thanks to this check, the returned value has to be of type `Option[Dataset[List[String]]]`. This has some implication to the subsequent data processes in the pipeline, that is, all the use of

twogram_df would need to check for potential None as incoming values. Though in a strongly typed language such as Scala, the compilation process would have enforce these checks in an earlier stage. In the next section, we pursue this direction.

2.2 Type-Setting the Data

An alternative approach would be use Dataset instead of Dataframe as the main collection data structure.

```
1 case class Article(source:String, date:Date, text:String)
2
3 val mydata:Dataset[Article] = sparkSession.read.csv("/data/input.
      tsv").as[Article]
4 mydata.show
```

In Spark, Dataset is a generic/polymorphic version of dataframe. A Dataset type constructor expects a type argument which represents the type of the contained elements. In this running example, we make it explicit that the elements in the Dataset are of type Article. Dataset offers a set of data processing APIs subsuming those offered by dataframe. In Spark, Dataframe is a type alias of Dataset[Row], where Row is a closed world encoding of universal data representation. A Row object can be seen as a non-generic Vector whose elements of type Object. Using Row as the element type imposes certain limitations as we have seen in Sect. 2.1, i.e., we need to keep checking for existence of a given column in the rows.

At line 1 of code snippet above, the case class keyword introduces a singleton algebraic data type Article, which has a constructor Article (same as the type's name) taking three arguments, source, date, and text. The use of algebraic data type gives us multiple advantages over the Row representation.

1. Firstly, it frees us from restricting the data to be structured data. Algebraic data type is a natural encoding of semi-structured data as studied by previous works (Wallace and Runciman 1999; Sulzmann and Lu 2008).
2. Secondly, it provides a type level declaration of the incoming data's structure. It can be regarded as a contract enforcing the roles and responsibilities between different processors in the data processing pipeline, i.e., the data must be cleansed and parsed before being used by the current processing step, none of the columns should be null or absent.
3. The third advantage is that the "down-stream" processors are greatly simplified, for example, we rewrite twogram_df into twogram_ds as follows:

```
1 def twogram_ds(ds:Dataset[Article]):Dataset[List[String]] = {
2   ds.map(a => twogram(a.text))
3 }
```

In the body of twogram_ds, we simply apply twogram to the text fields of all articles in the dataset. This simplification applies to all subsequent steps which use twogram_ds since the Option type is no longer required.

However, there remains a draw-back in this approach. Unlike the dataframe approach, the current dataset approach demands a fix schema to the input type of the function `twogram_ds`. It limits the re-usability of this function.

2.3 Sending Them for Classes

To resolve the issue of lacking of re-usability, we resort to a well-known programming language concept, type class (Peyton Jones et al. 1997). Type class was introduced in some functional programming languages, such as Haskell. It allows programmers to have a type-safe mixture of parametric polymorphism (such as generic) with ad-hoc polymorphism (such as function overloading). Type class is not a first class citizen in Scala. There exist some well-known tricks to encode type classes in Scala.

```
1  object HasTextTypeClass extends Serializable {
2    trait HasText[A] extends Serializable {
3      def getText(x:A):String
4    };
5
6    object HasTextOps extends Serializable {
7      def instance[A](fn:A => String):HasText[A] = new HasText[A] {
8          override def getText(x:A):String = fn(x)
9      }
10   }
11 }
```

At line 1, we declare an object `HasTextTypeClass` which has two declarations, a trait `HasText` and an accompanying object `HasTextOps`. A trait in Scala is similar to a Java interface, which defines an abstract contract. All implementations of the trait need to fulfill the contract by providing a concrete implementation of the trait member functions. In our case, all implementations of `HasText[A]` are obliged to provide some concrete implementation of `getText`. The object `HasTextOps` in line 6 defines a helper function `instance`. Function `instance` takes a higher-order function argument `fn` to define an implementation of the `HasText[A]` trait, by overriding `getText`.

Referencing `HasTextTypeClass` as a library, we can re-define a generic version of `twogram_ds`.

```
1  import HasTextTypeClass._
2
3  def twogram_ds[A](ds:Dataset[A])(implicit ev:HasText[A])
4                  :Dataset[List[String]] = {
5    ds.map(a => twogram(ev.getText(a)))
6  }
```

In the above adjusted definition of `twogram_ds`, we generalize the input argument type as `Dataset[A]`, where `A` is a type variable (A.K.A. generic). In addition, we include an implicit argument `ev` (short for "evidence"), which indicates an implementation of `HasText[A]` must be provided / inferred in the current context.

In the function body, we use the trait member function `ev.getText` to extract the text field from `a`, which is an element in the dataset. The above implementation of `twogram_ds` and `HasTextTypeClass` can be packaged as a library module for re-usability purposes.

Let us get back to our data processing application. To apply `twogram_ds` to our actual argument of type `Dataset[Article]`, we must first provide an instance of `HasText[Article]`.

```
1 implicit val articleHasText = HasTextOps
2     .instance[Article]( (a:Article) => { a.text });
3
4 twogram_ds(mydata)
```

At line 1, we define an instance of `HasText[Article]` as an implicit value. In Scala, when a function with implicit arguments is invoked, the compiler will search for all the matching implicit values based on their types in the current context. At line 4, we apply `twogram_ds` to the dataset `mydata` without the need of explicitly applying `articleHasText` as the argument.

2.4 A Quick Summary

In this section, we illustrated how to make a simple data processing task type safe and yet retaining re-usability by using algebraic data types and type classes. The same trick is applicable to other distributed data structures such as resilient distributed dataset (RDD) and larger scaled tasks.

3 Sailing Safe Through the Storm

It is common that at certain point in time, a data engineering team sorts out most of the historical data via batch processing and turns to developing a strategy to process the new incoming data. Compared to the historical data, these new incoming data are much smaller in size and refreshed in every second. It is inefficient and in-economic to apply the batch processing models to this tiny little "delta."

Spark streaming (Spark 2019) was introduced to address this issue by providing simple wrappers around existing Spark batch mode data processors. This greatly reduces the need of re-development due to shift of data processing mode. However, Spark streaming supports data parallelism but not task parallelism. Data parallelism parallelizes computation by applying common instructions to different subsets of data. Task parallelism parallelizes computation by identifying non-interfering sub-tasks (with different instructions) in a main task and executes the sub-tasks in parallel.

Apache Storm (Storm 2019) is a robust framework for handling real-time infinite stream of data. It supports both data parallelism and task parallelism. In a nutshell, Apache Storm models a flow-based programming paradigm where the processes

are represented as objects of class spout and class bolt. A spout denotes a data-generator and a bolt denotes a data processing/transforming possess. Such an architecture allows flexible configuration and is designed for a stream processing system. However, in Apache Storm, data are propagated across different processors in an untyped manner. This leads to potential run-time errors.

3.1 An Untyped Storm Topology

To illustrate the idea, we adopt the example, "word-count topology" from Apache Storm tutorials such as (Gkatziouras 2017),

```
1  object WordCountTopologyUntyped {
2    def main(args: Array[String]): Unit = {
3      val builder = new TopologyBuilder
4      builder.setSpout("spout", new RandomSentence, 1)
5      builder.setBolt("split", new SplitSentence, 2).
           shuffleGrouping("spout")
6      builder.setBolt("count", new WordCount, 2)
7            .fieldsGrouping("split", new Fields("word"))
8      val conf = new Config()
9      conf.setDebug(true)
10     conf.setNumWorkers(3)
11     StormSubmitter.submitTopology(args(0), conf, builder.
           createTopology())
12   }
13 }
```

We present the word-count Storm topology in Scala together with a directed-graph representation. In the graph representation, double-line rectangle denotes a spout which generates data. single-line rectangles denote bolts which are data processors. The Scala implementation was given in the `main` function in object `WordCountTopologyUntyped`. We use a `TopologyBuilder` to "assemble" the three components via `setSpout` and `setBolt` methods. `setSpout` method takes a string as the name of the spout instance, the spout object, and a non-negative integer which denotes the number worker threads that the spout will spawn. The method `setBolt` has a similar set of arguments and it connects bolt objects to the topology. Methods `shuffleGrouping` and `fieldGrouping` specify the names of the up-stream processor/generator. The difference of the two will be discussed in the next few paragraphs.

In Fig. 1, we provide the complete definitions of `RandomSentence` spout, `SplitSentence` bolt, and `WordCount` bolt. In the class `RandomSentence`, the `open` method defines the initialization routines. The main routines of the spout is defined in the method `nextTuple`, which randomly picks one sentence out of the predefined set and passes it to the next process in the pipeline using `_collector.emit`. The method `declareOutputFields` creates a simple

```
1  class RandomSentence extends BaseRichSpout {
2    var _collector: SpoutOutputCollector = _
3    var _rand: Random = _
4    override def open(conf: java.util.Map[_, _],context:
         TopologyContext,
5        collector: SpoutOutputCollector): Unit = {
6      _collector = collector
7      _rand = Random
8    }
9    override def declareOutputFields(declarer: OutputFieldsDeclarer)
         : Unit = {
10     declarer.declare(new Fields("sentence"))
11   }
12   override def nextTuple(): Unit = {
13     Utils.sleep(100)
14     val sentences = Array( "the cow jumped over the moon",
15       "an apple a day keeps the doctor away",
16       "four score and seven years ago",
17       "snow white and the seven dwarfs",
18       "i am at two with nature" )
19     val sentence = sentences(_rand.nextInt(sentences.length))
20     _collector.emit(new Values(sentence))
21   }
22 }
23 class SplitSentence extends BaseBasicBolt {
24   override def execute(input: Tuple, collector:
         BasicOutputCollector): Unit = {
25     val sentence = input.getString(0)
26     sentence.split(" ").foreach {
27       word => collector.emit(new Value(word))
28     }
29   }
30   override def declareOutputFields(declarer: OutputFieldsDeclarer)
         : Unit = {
31     declarer.declare(new Fields("word"))
32   }
33 }
34 class WordCount extends BaseBasicBolt {
35   val counts = scala.collection.mutable.Map[String, Int]()
36   override def execute(input: Tuple, collector:
         BasicOutputCollector): Unit = {
37     val word = input.getString(0)
38     val optCount = counts.get(word)
39     if (optCount.isEmpty) { counts.put(word, 1) }
40     else { counts.put(word, optCount.get + 1)   }
41     collector.emit(new Values(word, counts))
42   }
43 }
```

Fig. 1 A simple word count topology (spout and bolts definitions)

label for each generated data field, which can be referenced by the processors in the downstream. The class `SplitSentence` takes the output from its up-stream processors in the pipeline and splits it by spaces. Each word from the split will be forwarded to the downstream, as defined in the `execute` method. The class `WordCount` maintains a mapping from words to their numbers of occurrence in a

variable named `counts`. When `execute` is invoked, the incoming value is treated as a word. The word's count will be increased by 1 if the word already exists in `counts`, otherwise the count is initialized to be 1.

Let us go back to the main topology construction. When a bolt is "connected" to its up-stream, an input declaration has to be made. In case of the `SplitSentence` bolt, the distribution of the inputs from `RandomSentence` does not matter, hence `shuffleGrouping` is used. In case of the `WordCount` bolt, the input assignment is crucial, as we would like to ensure that the same word must always go to the same bolt instance otherwise the counting is meaningless. Thus, `fieldsGrouping` is used.

3.2 Storm Is Dangerous

Everything is neat and tidy except that we spot two problems.

1. The connection between spout and bolts is dynamically typed, i.e., if there is a type mismatch, it will only be discovered during run-time. For instance, let us consider there is a slight adjustment of the definition of `SplitSentence` by replacing lines 27–29 in Fig. 1 by the following:

```
1       sentence.split(" ").foreach {
2         word => collector.emit(new Value(word.toList))
3       }
```

   As a result, the bolt emits `List[Char]` instead of `String`.[1] The compiler happily accepts it and type checks it. The type mismatch between the output from `SplitSentence` and input to `WordCount` will only be discovered as a run-time error. This is going to be potentially costly and dangerous. In production code, we are dealing with large set of data and large-scaled topologies with reusable components. There are cases where this type of mismatch run-time errors become hard to trace. Testing and debugging will drain the team resources.
2. The construction of the topologies requires that there should be at least one spout followed by zero or more bolts within a topology. It does not make sense to add a bolt into an empty topology. However, the `TopologyBuilder` does not enforce this constraint statically. For instance, if line 4 from the `WordCountTopologyUntyped` topology in Sect. 3.1 was omitted, the compiler does not report errors, but a run-time exception will be raised.

[1] In Scala, `String` type is not an alias of `List[Char]`.

3.3 Phantom Types to the Rescue

To solve these problems, we adopt a well-known technique called *phantom type* (Finne et al. 1999; Cheney and Hinze 2003), which is often used to provide type safety in domain specific language embedding. In a nutshell, phantom types are data type whose type variables are not fully mentioned in its data constructors.

We follow style of phantom type encoding described here (Iry 2010). First of all, we introduce two phantom types to capture the input and output types of the spout and the bolt.

```
1  trait StormSpoutT[Out]{def spout:IRichSpout}
2  trait StormBoltT[In,Out]{def bolt:IBasicBolt}
```

In the above, we define two traits which embed a spout (or a bolt, respectively). `StormSpoutT` expects a type parameter `Out` which describes the output type of the embedded spout and `StormBoltT` expects two type parameters `In` and `Out` which define the input and output type of the embedded bolt. Note that none of these type parameters are mentioned in the body/member of the traits. Hence, we call them phantom types.

The next step is to provide some extra type annotations to the `RandomSentence` spout, `SplitSentence` bolt, and `WordCount` bolt. Note that the definitions of these spout and bolts remain unchanged. For brevity, we omit the repetition of the definition and provide only the signatures.

```
1  case class RandomSentenceT (spout: RandomSentence)
2       extends StormSpoutT[String]
3  class RandomSentence extends BaseRichSpout { ... }
4
5  case class SplitSentenceT (bolt:SplitSentence)
6       extends StormBoltT[String,String]
7  class SplitSentence extends BaseBasicBolt { ... }
8
9  case class WordCountBoltT (bolt:WordCount)
10      extends StormBoltT[String,(String,Int)]
11 class WordCount extends BaseBasicBolt{ ... }
```

Here we provide concrete implementation of the `StormSpoutT` and `StormBoltT` traits. `RandomSentenceT` extends the `StormSpoutT` trait by declaring the underlying spout to be `RandomSentence` and at the same time specifying the output type of the spout is `String`. `SplitSentenceT` embeds `SplitSentence` as the underlying bolt and specifies the input type of the bolt is `String` and the output type is `String`. Similarly, `WordCountBoltT` embeds `WordCount` which has `String` as input type and `(String, Int)` as output type.

Next, we define three possible states of constructed topologies, an ordering among the three states, and a phantom type representing a state of a topology. Note that the ordering is implicitly enforced via a sub-typing relation `TopWithBolt <: TopWithSpout <: TopEmpty`.

```
1  abstract class TopEmpty
2  abstract class TopWithSpout extends TopEmpty
3  abstract class TopWithBolt extends TopWithSpout
4
5  case class TopologyBuilderT[+State,+Out]
6            (builder:TopologyBuilder,output_name:String) {
7    def createTopology = builder.createTopology
8
9    def init: TopologyBuilderT[TopEmpty, Out] =
10       new TopologyBuilderT(builder, output_name)
11
12   def >>[NextOut](
13       spout_name: String,
14       ts: StormSpoutT[NextOut],
15       threadMax: Int
16   )(
17       implicit evS: State <:< TopEmpty
18   ): TopologyBuilderT[TopWithSpout, NextOut] = {
19     builder.setSpout(spout_name, ts.spout, threadMax)
20     new TopologyBuilderT[TopWithSpout, NextOut](builder,
           spout_name)
21   }
22
23   def >>>[NextOut, State <: TopWithSpout](
24       bolt_name: String,
25       tb: StormBoltT[Out, NextOut],
26       threadMax: Int
27   )(inDecl: BoltDeclarer => BoltDeclarer)(
28       implicit evS: State <:< TopWithSpout
29   ): TopologyBuilderT[TopWithBolt, NextOut] = {
30     val i = builder.setBolt(bolt_name, tb.bolt, threadMax)
31     inDecl(i)
32     new TopologyBuilderT[TopWithBolt, NextOut](builder, bolt_name
           )
33   }
34 }
```

`TopologyBuilderT` expects two type parameters, `State` and `Out`. The plus `+` sign suggests that both type parameters are covariant. This case class embeds the actual `TopologyBuilder` instance. In addition, we define two combinators `>>` and `>>>` for the ease of topology construction, which enforce the matching constraint between the output type of current processor and the input type of its following processor. Furthermore, as the topology being constructed, we also keep track of the state of the topology. For example, when a topology is initiated, its state should be `TopEmpty`, until a spout is added, the state becomes `TopWithSpout`. Codes that add a bolt to a topology with a `TopEmpty` state will be rejected by the type system.

`>>` adds a spout into a topology whose state is bounded by `TopEmpty`, which is enforced via the implicit type constraint at line 17. As a formal argument type annotation, `<:<` sets an upper bound to the left-hand side type variable. It then updates the topology state to the result topology type by annotating it with

`TopWithSpout` at lines 18 and 20. In addition, it also ensures that the resulting topology shares the same output type as the spout by annotating with type variable `NextOut` at lines, 14, 18, and 20.

`>>>` adds a bolt into a topology whose state is bounded by `TopWithSpout` and returns a topology with bolt. This is enforced by type constraints at lines 23 and 27. Note that `<:` serves the same purpose as `<:<` except that it is used at the type variable declarations instead of formal argument type signatures. The result must be a topology with bolt added, as enforced by annotations at lines 29 and 32. In addition, it ensures the input type of the bolt to be added agrees with the output type of the current topology being extended, i.e., `Out`, as in line 25. The output type of the resulting topology has to be `NextOut`, which is the output type of the bolt. This is enforced in lines 25, 29, and 32.

With these new phantom type and combinators, we can rewrite the topology construction in Sect. 3.1 as follows:

```
1  object WordCountTopologyTyped {
2    def main(args: Array[String]): Unit = {
3      val builderT = (new TopologyBuilderT(new TopologyBuilder, "")
           ).init
4        .>>("spout", new RandomSentenceT(new RandomSentence), 1)
5        .>>>("split", new SplitSentenceT(new SplitSentence), 2)(
6          _.shuffleGrouping("spout")
7        )
8        .>>>("count", new WordCountT(new WordCount), 2)(
9          _.fieldsGrouping("split", new Fields("word"))
10       )
11     val conf = new Config()
12     conf.setDebug(true)
13     conf.setNumWorkers(3)
14     StormSubmitter.submitTopology(args(0), conf, builderT.
           createTopology())
15   }
16 }
```

where the two issues mentioned in Sect. 3.2 are fixed via the type system. The code verbosity of the type annotations and constraints can be reduced by using techniques such as macros and meta-programming (Burmako 2017).

4 Conclusion

In this article, we walk through a few examples to illustrate how static typing tricks can be applied to reduce testing and run-time checking overhead during data engineering project development. It is well-known that

> Program testing can be used very effectively to show the presence of bugs but never to show their absence. (Dijkstra 2017)

With type system, static analysis, and logical proof, we are able to verify whether the given properties hold in the program codes. It follows that a large subset of test cases/run-time checks can be eliminated to stream-line the data engineering projects as long as we verify these properties using type system, static analysis, and logical proof. As a resource, development resource and infrastructure can be reduced or re-allocated for other more important tasks.

All the code examples in this article can be found in (Lu 2019).

References

Burmako, E. (2017). Unification of compile-time and runtime metaprogramming in Scala. Infoscience, p. 240.

Cheney, J., & Hinze, R. (2003). First-class phantom types. Technical report, Cornell University.

Dijkstra, E. W. (2017). The manuscripts of Edsger W. Dijkstra. Retrieved December 1, 2019, from https://www.cs.utexas.edu/users/EWD/transcriptions/EWD03xx/EWD303.html

Finne, S., Leijen, D., Meijer, E., & Peyton Jones, S. (1999). Calling hell from heaven and heaven from hell. In *Proceedings of the Fourth ACM SIGPLAN International Conference on Functional Programming, ICFP '99* (pp. 114–125). New York: ACM.

Gkatziouras, E. (2017). Java code geeks: Wordcount with Storm and Scala. https://www.javacodegeeks.com/2017/02/wordcount-storm-scala.html

Hadoop (2019). Apache hadoop. Retrieved December 1, 2019, from http://hadoop.apache.org/

Iry, J. (2010). Phantom types in Haskell and Scala. Retrieved December 1, 2019, form http://james-iry.blogspot.com/2010/10/phantom-types-in-haskell-and-scala.html

Lu, K. Z. M. (2019). Code examples in the article - lean launch data engineering projects with super type power. Retrieved November 1, 2019, from https://github.com/luzhuomi/lean_launch_data_engineering_ex

MyPY (2019). Optional static typing for python. Retrieved December 1, 2019, form http://mypy-lang.org/

Peyton Jones, S., Jones, M., & Meijer, E. (1997). Type classes: An exploration of the design space. In *Haskell Workshop*.

Spark (2019). Apache spark. Retrieved December 1, 2019, form https://spark.apache.org/

Storm (2019). Apache storm. Retrieved December 1, 2019, form https://storm.apache.org/

Sulzmann, M., & Lu, K. Z. M. (2008). Implementation and application of functional languages. In O. Chitil, Z. Horváth, & V. Zsók (Eds.), *Implementation and application of functional languages*, chapter XHaskell — Adding Regular Expression Types to Haskell (pp. 75–92). Berlin: Springer.

Wallace, M., & Runciman, C. (1999). Haskell and XML: Generic combinators or type-based translation? In *International Conference on Functional Programming ICFP '99* (pp. 148–159). New York: ACM Press.

Ubiquitous Computing: From 5G to the Edge and Beyond

André Panné

Abstract

This is an invitation to a journey from our communication technology's past to the technological borders of today and beyond to the unknown. We will jump back in time for about three generations to become aware of the major steps of progress we have achieved in the past 50 years. From there we will move forward in three main paths, covering the invention of the (inter-networking) network, the development of hard- and software, and the advance of mobile communications. We will have a look at how these three streams of development eventually merged into one and how it led us to the technological reality of today. If you have been part of this story yourself, or if you already know all about it, you may want to read it anyway with a smile of remembrance. If you do not want to repeat this part, please feel free to jump ahead a couple of pages to Sect. 4.

1 Ubiquitous Computing

In the beginning was the word, and the word was about the vision of a new economy:

> It begins and ends with people. Our people are the architects of the new Internet economy. Our clients grasp the magnitude of the opportunity. Together, we're changing everything. Anything less would simply be another job. (Proxicom 1997)

The quote above is out of Proxicom's Little Red Book, a brochure for new hires laying out the values and the vision of the company. Proxicom was a full-service e-business professional services provider, one of the so-called "Little Five," besides Razorfish, Sapient, Scient, and Viant, which were challenging the consulting arms

A. Panné (✉)
TRADUM, Bonn, Germany
e-mail: andre.panne@tradum.de

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_11

of the "Big Five" accounting firms (Sadler 2001). Everything, everywhere on every device was the motto based on the vision of "ubiquitous computing" (Weiser 1991). And nothing less did these pioneers try to achieve by breaking the rules and by rebuilding processes, entire industries, and markets.

Ubiquitous Computing eventually started to become true by the end of the 1990s, when the Internet was exploding. New technologies allowed to scale and implement more and more sophisticated websites and web-based solutions. The upcoming mobile Internet was an early glimpse into a wireless interconnected world. Looking backward it turns out that the ideas behind ubiquitous computing were on the spot. Many, if not most of the correlating ideas were to become true. Then, in 1997, we just did not expect that implementing all of these would take us so long.

2 The Journey: Or How We Got Here

Sometimes when you deal with a complex situation, it is helpful to step back and look what happened in the past, to better understand where you are standing today. What we are calling the Internet of Everything, which encompasses technologies like 5G and architecture concepts like Edge Computing, did not drop out of a box, but it was a long-lasting iterative development process and a progress taking several steps and hurdles throughout the years.

The three main components building today's communications ecosystem are:

1. The (inter-networking) network
2. Hard- and software
3. Mobile communications

2.1 The Becoming of the Inter-Networking Network

Let us celebrate a 50th birthday: It was on October 29, 1969, when the first technologies of what would once become the Internet were turned into action forming the basic elements of the "ARPANET" (Advanced Research Projects Agency Network). This computer communication network was initially established between universities and research institutes and it eventually expanded over the entire United States. With the possibility to connect computers across the country, another challenge needed to be solved: the question how to address other users in the network. In 1971, a computer scientist named Ray Tomlinson eventually had the idea to use the @ symbol for this purpose (Allman 2012). Also and already in 1971, the University of Hawaii developed a wireless network, the so-called AlohaNet (McClelland 2017) which would connect computer systems via radio. With the invention of TCP/IP (Cerf and Kahn 1974), a new era was getting prepared to launch. This technology was based on the idea of packet switching, allowing several computers to share a single network without the data packages interfering with each other. Together with the Ethernet standard, TCP/IP was to become the norm of interconnection between computer systems. Finally, on January 1, 1983 all 400 hosts of what had been the ARPANET until then were migrated to the

new TCP/IP protocol. From an inter-networking perspective, this was one of the founding pillars of the Internet of today.

With computers getting connected all over the world, efficient processing and structuring of information became more and more important. Since the switchover to TCP/IP, it took another 6 years for the most essential building elements of the Internet to be defined:

- HTML—the Hyper Text Markup Language. It was Sir Tim Berners-Lee, who wrote a memo to his manager at CERN, a European research facility based in Geneva, suggesting the introduction of a general information system (Berners-Lee 1989). Eventually his concept was agreed upon and Berners-Lee was writing a browser software to access his html data, which he called:
- "World Wide Web." What was initially only the name of a piece of software should become the synonym of what has become globally available by now: The World Wide Web as a technology platform of interconnected content created what we call "the Internet" today.

But to create a ubiquitous computing world, it needed more than a standard for inter-networking, content access, and distribution. Way before software was eating hardware for breakfast, it required loads of hardware available to many, many people. These nerds, programmers, software engineers, coders, were the ones to develop the ecosystem of programming languages, resources, and tools around, as we have them available today.

2.2 Hard- and Software Evolution

Machine-based computing was piloted by Konrad Zuse, a German engineer prior to the Second World War. Eventually he created the Z3—the first binary 22-bit floating point calculator the world had ever seen. The company Konrad Zuse founded—the Zuse KG—was going to sell a total of 251 computers before it was eventually acquired by Siemens in 1967. This shows how exponential curves start: with small numbers. We know the rest of the story: With the invention of semiconductor-based transistors, vacuum tubes were being replaced and things started to get rolling and scaling.

Moore defined his law in 1965, stating that semiconductor performance would double every 18 months. It should be proven true to this day (Hiremane 2005). This incredible increase in performance accompanied by falling prices drove the computing world from mainframes to home computers such as the Apple II, the IBM PC, SUN, the Commodore C64, the Sinclair ZX82, followed by the Apple Macintosh, and many more ever since. The important aspect of this part of technology history is the sudden and unprecedented availability of computing power at hands of hundreds of thousands of people. Not only professional software engineers were able to afford these new machines, but everyone was able to own one. Once the Genie was out of the bottle, there was no stopping anymore for the exploding software industry.

When machine computing started, coding software was basically punching holes into punch cards, putting them into a reader and hoping that you had not punched the wrong hole. Since then, the art of writing code has leap-jumped along with the invention of supporting operating systems. Examples are IBM's OS/360, CP/M, Unix, MS-DOS, just to name the most relevant ones of the early years. It was this development of standardized systems that allowed the portability of software and therefore the scale of the young industry. With the rise of the Apple Macintosh and its revolutionary operating system in 1984, yet another era began: graphical user interfaces changed the way of interacting with computers forever. WYSIWYG—what you see is what you get—was the new interface standard. Microsoft would need another 6 years before releasing their first graphical user interface with their Windows operating system, soon to be followed by Linux, BeOS, and NeXtStep. The latter, also a Unix derivative, eventually should become the core of the current Apple Mac OS-X with the returning of Steve Jobs to the company he founded.

To program all these mainframes and computer systems with their individual operating systems, it needed suitable and stable programming languages and frameworks. Since the invention of computers, mankind actually invented an incredible number of programming languages. On checking the Internet, you easily end up with over 600. Actually, one would be surprised, how many you have not heard of. It is actually a simple self-test to write down the ones you do know, which may become a rather revealing exercise given our growing dependency on software.

2.3 Mobile Telecommunications Everywhere

Yet, it needed one more element for achieving ubiquity: wireless/mobile communication. Mobile communication turned out to have the largest imprint on the world as we know it today. Elements of this technology actually have been foreseen in articles in a book called "Die Welt in 100 Jahren" (The world in 100 years) by German journalist Arthur Bremer (1910). Already at the beginning of the twentieth century, this book spoke about devices to transport voice and video via radio. It even predicted the availability of wireless handheld devices in everybody's pocket. And it may be hard to believe, but in 1926, a fast train connecting Berlin and Hamburg already had a public phone cell providing "wireless" phone calls from the train to fixed line receivers. Today we know that these engineers were on the right track back then already.

This promising development in the early years of the last century took a long break during World War II, but eventually picked up again in the "Wirtschaftswunderland" Germany, once the war was over. With the economic success of the country, the need for mobile telecommunication surged and resulted in a series of mobile network technologies to be launched in the market. In Germany, these networks were named by the alphabet: In 1958 the A-Network was launched, followed by the B-Network (1972). Finally, the C-Network, the last analog mobile communications network, was launched in 1986. These analog networks were considered the first generation of mobile telecommunication networks and therefore

named 1G networks. The actual 1G telephones had little to do with today's mobile phones, and typically were built into cars and other vehicles, as their sheer size and weight did not allow them to be carried around much.

The analog mobile communications age came to an end with 2G—second-generation—networks. Mobile communications started being based on digital compression algorithms, enabled by fast low energy number crunching processors. For the transmission of data, these mobile networks started operating on a tiered network architecture similar to the one of the Internet. These networks were operating on a different protocol: the SS7 protocol (Techopedia 2019). In 1987, the GSM (Groupe Speciale Mobile) released a MoU introducing the GSM standard (GMSA 2019a). The first German 2G GSM network was launched by DeTeMobil, a subsidiary of Deutsche Post, in 1992. Then, the German post office (Deutsche Post) was in charge of all of the country's telecommunication services. DeTeMobile was followed by the mobile network operator Mannesmann Mobilfunk, a diversification of the Mannesmann steel portfolio. Following in the alphabetic order, the E-Net was launched with the mobile network operators E-Plus and Viag Interkom. Today, DeTeMobil has become Deutsche Telekom and is widely known in some markets as T-Mobile. Mannesmann Mobilfunk eventually got acquired by Vodafone. And Viag Interkom was first acquired by O2, then by Telefonica, which eventually swallowed the fourth provider E-Plus as well.

When communication started, it was all about data already. The Morse code was the way to transport characters and numbers over wires across the land or over radio at sea. Still, for a long time after, telecom communication was perceived mainly as a voice-based service. Of course, there were still telegrams, Telex and later fax machines, mainly in offices and in the post office. But these were expensive and bulky machines.

It should take until the introduction of GSM, that another chapter of mobile communication was opened: sending and receiving data from and to everybody. Utilizing some leftover signaling capacity in the SS7 network protocol, the Short-Messaging-Service—SMS was invented by a Deutsche Post employee. It was introduced to the market und the brand name "D1-Alpha." And nobody expected what should turn out of a service limited to 160 characters, which tediously needed to be typed on a numerical keyboard. Not only did SMS at an initial price tag of 39 German Pfennig, contribute large parts of profit to mobile operator's balance sheets in the coming years. It also opened the doors for large scale application-based data transmission via wireless telecommunication networks. SMS enabled the first M2M—machine-to-machine—data use cases on GSM networks.

3 Mixing The Dough

It took 25 years to get the ingredients prepared for a market and technology mixture that would expand like yeast dough. We had the World Wide Web, the Internet, and the network technology to connect millions of computers being programmed by an ever-growing number of software developers. We had affordable computers

and we had a global mobile communications network, which was about to grow exponentially with the accelerating globalization and with the rising numbers of the World Wide Web users. It was the beginning of what would later be called the dot.com bubble: an overhyped market with lots of technical and business phantasy, setting the example of hockey stick business case dreams and resulting thereof, exaggerated company evaluations.

Along with these new technologies, corresponding associations and standardization bodies were created. The W3C—the World Wide Web Consortium—was founded by Tim Berners-Lee in 1994 (W3C 2019). The GSM Association in 1995 (GSMA) and ICANN—the California-based nonprofit "Internet Corporation for Assigned Names and Numbers"—in 1998. ICANN is the organization which is home to the Internet Assigned Numbers Authority (IANA) group, which is in charge of Internet domain names and IP numbering. The market was creating its own rules with a new paradigm of interconnection, international inter-operability and—step by step—with inter-exchangeable data formats. This was a major breakthrough in a computer world that was protected by walled gardens for the longest time. These days, you could not even send e-mails from the AOL online service to their competitor Compuserve.

But of course, it was exactly companies like AOL to make the Internet a common asset. They gave millions of private households access to the Internet with their "You have got mail" alert on receiving new e-mails, a soundbite, which eventually became famous with the identically named movie. Many others were using the Compuserve or Prodigy services to go online, and yet others again simple bought themselves a 3.5 in. disk with the necessary PPP drivers to connect to the next Internet server at their university. Mosaic, the first commercial browser, was pushing the window to the World Wide Web open, soon to be overtaken by the Netscape Communications browser. Evers since, several other browsers have followed, competing on speed, html-interpretation and X-platform compatibility until today. With the World Wide Web and browsers available and websites piling up on the Internet, finding content became a challenge. It was the opportunity of web crawlers like Yahoo! and AltaVista, which launched in 1995, soon to be followed by Google in 1997. The latter eventually to become the synonym of web search in today's languages.

Toward the end of the 1990s, the Internet was the driving force of business. E-commerce startups and e-business consultancies such as the previously mentioned Little Five were challenging brick-and-mortar business models. Disintermediating established value chains was the motto. A famous example is the Polymerland online marketplace project at GE Plastics. Following Jack Welsh's motto "Destroy your own Business" (Martinson 2000), the world's largest plastic pellet manufacturer was kicking out the intermediaries in their cascaded distribution chain. On the B2C (business to consumer) side, Amazon was attacking the booksellers' market, and eBay was inspiring many other marketplaces in the B2C and B2B (business to business) segment to offer similar online auctioning platforms.

On the network architecture side, the addressable space in the Internet was getting tight with the strong increase of web sites. To solve this issue, in 1998, IPv6 was introduced to expand the address space to 2^{128} possible addresses. In parallel

security became an ever-growing concern on the web and TLS (Transport Layer Security) was introduced as successor to SSL (Secure Socket Layers). Companies like Thawte started their Web of Trust model, providing web security certificates via their network of registered and certified notaries.

In the same year, on the mobile communications side of things, 3GPP, the 3rd Generation Partnership Project, united several regional standardization bodies. Their aim was to lay out the technical specifications of the third generation of mobile networks. It was the initiation of 3G/UMTS networks (3GPP 2019). Today, the 3GPP organization is headquartered in Sophia Antipolis, a small town in the Provence in southern France. This is Europe's high-tech center and has been a sweet spot of technical excellence for a long time.

In these last years of the ending millennium, Siemens delivered the first GSM data module to the market. This allowed the use of mobile data transmission with an industry-proof device. Another mobile web "killer" application called WAP (web-access-protocol) was less successful. It delivered an awful user experience on SMS-style text web pages and first implementations failed due to lacking business cases. Yet another technology to influence our every day's life should be defined by a newly built association—the Wireless Ethernet Compatibility Alliance (WECA) branded their new technology Wi-Fi and should later on rename themselves to Wi-Fi Alliance (Information Gatekeepers 2002). Today, the first question of every kid in the world entering a new building is: "Do you have Wi-Fi? What is the password?"

With beginning of the new millennium, our civilization survived the year 2000 bug without major damages and started into a new chapter of a mobilized Internet world. Microsoft released Windows Mobile and the widely spread Handspring Palm PDAs (personal digital assistants) were enhanced with GSM functionality under the Palm treo product line. Nokias Communicator was running the mobile operating system Symbian S60. And Nokia, having some 30% market share of all mobile phone sales worldwide, was Europe's most valuable company (Young 1999). With introducing the data transmission protocol GPRS, also called 2.5G, the GSM networks expanded their data transfer capacities to staggering 40 kbit/s, which was similar to what you would get via a regular phone line modem then.

The stage was set for a new term to show up: Internet of Things (IoT). A term coined by Kevin Ashton at Procter & Gamble (Cole 2018), IoT was originally intended for describing a solution to track goods in the supply chain via RFID chips. But even before that, first devices were connected via M2M (machine to machine communication), such as connected Coke vending machines (IBM 2018) or connected toasters (Rebaudengo 2012). What IoT needed to flourish were further technological frameworks to simplify the architectural setup. As such, the introduction of the XML (Extensible Markup Language) through W3C, and later REST (Representational State Transfer), as well as JSON (JavaScript Object Notation) would give developers the architectural stability to build what was called Web Services. Eventually Amazon would become the leading web services provider with its subsidiary Amazon Web Services (AWS), which it founded in 2006 to outsource its increasing internal IT demands. Cloud Computing was born and has diversified into Software as a Service (SaaS)—e.g., SAP S/4HANA. Later came

Platform as a Service (PaaS), e.g., IBM Bluemix/Watson, and Infrastructure as a Service (IaaS), e.g., the Oracle Cloud Infrastructure (OCI). Meanwhile, entire corporate IT system landscapes are being delivered as cloud-based applications, with the advantage of scalability, reliability, constantly upgraded and bug-fixed, and providing more security than most IT departments could ever achieve themselves.

On the mobile network side, data bandwidth and radio coverage have been continuously increasing. The leading mobile phone suppliers of the early millennium, Nokia, Motorola, and Siemens Mobile, were offering smart phones with included e-mail, calendar and web applications and further smart features. In 2002, first 3G/UMTS networks were launched and promised speeds up to 384 kbit/s. However, even in highly developed countries, EDGE, or 2.75G, still is the best connection speed you can get in some areas until today. For short-range connectivity of smart devices, the standards ZigBee and Z-Wafe were established and triggered new Internet-of-things product lines around sensors and actors collecting themselves in local mesh networks, such as smart home automation, steering air conditioning, heat, light and other appliances.

In the years to come, the GSMA was pushing the mobile carrier market with further, even faster transmission technologies. On 3G followed "fake" LTE in its 3.9G version in 2010, even though true LTE$^+$/4G would not be around before 2014. The Internet received a major feature update by the standardization of HTML5 in 2008. This led to many new website designs with new features like responsive web design. Eventually HTML5 put an end to energy-consuming flash websites and flash-based banner advertising. Further wireless data transmission technologies for connected things like LoRa, NB-IoT and LTE-M were established, paving the way of connecting the foreseen 50 billion smart devices as predicted by Cisco for the year 2020 (Cisco 2011).

With all focus on technologies and standards, we must not disregard the maybe most important aspect of this story: The human contribution. Along with all these developments in 50 years, a continuously growing number of software developers was required to do the job. What was once an elite knowledge limited to scientists in white coats and to electrical engineers has become a job for many more people: Jeans and t-shirt-dressed global nomads, sitting on palm beaches and making a living with laptops on their knees. By the end of 2018, according to IDC, the number of software developers worldwide has reached 22.3 million with a growing tendency (IDC 2018). Soon we are about to reach the point, when more people on this planet are developing software, than are involved in building cars (Wickham 2017). This is a paradigm change we will need to consider: The transformation of value creation—from building things to building code.

Finally, one major disruption altered the world into a before and after. On January 9, 2007 at the Macworld Conference & Expo in San Francisco, Steve Jobs announced "... a product that comes along that changes everything ..." The first iPhone should ring the final bell for the old league of mobile phone producers. With its market availability in 2008 and together with the Google mobile operating system Android, which was released the same year, the term "smart phone" should now be something very different than ever before. For consumers, the iPhone and Android

based devices made the dream of ubiquitous computing come true: all content of this world at your hands, at any place, at any time.

4 Status Quo 2019

Not everybody needs to be able to cook a stew, but maybe everybody should be capable of appreciating the work that has been done, as much as it would be helpful for everybody to be able to identify the ingredients of the soup. Exactly this was the intention of the previous chapters and I hope you enjoyed the cooking session.

Here we are in the year 2019 and we have come a long way, eventually arriving at a cross-junction of fixed and mobile networks. Here end users and applications and devices are sharing the Internet for all thinkable sorts of communication and data. The underlying architectures and standards have achieved a stable, cross-operational environment for the use cases we are seeing today. The industry behind hardware, software, and networks has gone through several cycles and leap-jumps of innovation and consolidation. This resulted in a global footprint of thousands of technology firms delivering elements to the entire ecosystem. The market is led by a winning pack of a few global players, which dominate the playing field: Amazon, Apple, Google, IBM, Microsoft, and not to forget the forward-storming Chinese players Alibaba, Baidu, Huawei, Tencent (WeChat), and Xiaomi.

On the mobile network side of things, operators are still busy with upgrading their 3G and 3.9G systems to 4G-LTE^{+} (LTE-Advanced), with entire regions outside areas of high population being stuck without any acceptable data coverage (Opensignal 2019). With only 65% availability of high-speed LTE connections, Germany is making a disappointing rank 70 in an international comparison. The average max download speed in Germany is lower than 15 Mbit/s. That is less than half of the 35 Mbit/s the Netherland is achieving, not to talk about the 50Mbit/s of South Korea (Speedcheck 2019). In this context, Germany is a developing country.

The good news is that the number of mobile telephony subscribers first time ever has bypassed the number of fixed network subscribers in Germany, and this trend is going to continue. Regarding data consumption in Germany, fixed networks still are in the lead with 45.000 million Gigabyte of data. Mobile has a share of 1.993 million GB only, but with a 44% year-over-year growth rate, it is going to eat into the big chunk of the pie chart (Bundesnetzagentur 2019a). Talking about appetite: data is eating voice, and since the introduction of VoLTE, the difference between voice and data foreseeably is going to be void in the future. With the auctioning of the 5G spectrum (Bundesnetzagentur 2019b), the legal framework is prepared for the now again four mobile network operators to offer 5G in Germany. Already, the first 5G networks have been switched live in dedicated areas. 5G is starting to make its way into our connected reality, telling big promises of a high-speed mobile network in the near future.

5 The Evolutionary Revolution to 5G

When it comes to keeping marketing terminologies such as "LTE" or "5G" apart from the physical reality, you have to understand that the technical developments of cellular or mobile networks are happening based on subsequent specification releases worked out in the working groups of the previously cited 3GPP organization. The intention behind 3GPPs activities is to reduce complexity and to avoid fragmentation of technologies with each progressive 3GPP radio access technology. There will be no switch turning off 4G and turning on 5G tomorrow, but it will be a side-by-side developing, kind of overlaying and morphing the one network into the other over the years to come. In this sense, 5G actually is not a revolution, but more of an evolution and an extension of performance plus the add-on of new technologies to existing setups.

From a use case perspective, 5G is intended to combine these three different scenarios:

(a) Ultra-fast data transmission capability by **e**nhanced **M**obile **B**roadband **C**apability (eMBB)
(b) The M2M network of Internet of Things by offering the capability of addressing and communicating with literally billions of devices through **m**assive **M**achine **T**ype **C**ommunications (mMTC), and finally
(c) The support of time or mission-critical applications through **U**ltra-**R**eliable and **L**ow **L**atency **C**ommunications (URLLC)

Let us elaborate further on the most important features of 5G. The major difference to preceding GSM technologies is the significantly higher maximum data transmission rate. LTE (LTE^+ or LTE-Advanced) has a theoretical max data transmission rate of 500–1000 Gbit/s. The fully fledged real-world $5G^+$ is promised to turn out up to 20 Gbit/s. To put it in more illustrative words: downloading a UHD movie will need a few seconds only. The secret behind this huge increase is for one the improved overall technological architecture, but also the extension of utilized frequencies. While the old mobile networking technologies were operating on a spectrum between 0.8 and 2.6 Ghz, 5G is designed to also work on frequencies between 6 and 300 Gigahertz (Ghz) to achieve its full advantages. As physics teaches us, increasing frequencies is coming along with shortening wavelengths, so we are moving the spectrum to mm-waves in future. This will result in much denser information packaging, so much more data to be carried along on each wave segment. On the downside, the higher the frequencies, the shorter the reach. Millimeter waves cannot easily pass through objects or walls, which limits such theoretical 5G use case significantly. Anyway, the performance improvement is significant, even when used on today's frequencies.

Besides speed improvements, 5G will support technologies like Beamforming, Massive Multiple Input Multiple Output (MIMO) and Networkslicing.

- **Beamforming** allows to change the expansion of the radio waves in a way to maximize coverage in a defined target area. Instead of radiating the radio wave in a spherical way, a beamforming signal would look more like a baseball club hitting the ball (the receiving device) right where it flies.
- **Massive Multiple Input Multiple Output (MIMO)** is the bundling of several antennas of a wireless network, which allows transmitting and receiving of several data signals simultaneously over the same radio channel.
- **Networkslicing** will allow to create a multilayered bundle of virtual networks on the same antenna, providing different qualities of service levels and thereby serving different purposes. Conflicts between data packages of different importance and criticality can be kept apart—such as e-mails that are being sent from real-time traffic information.

Being more efficient and faster in sending to and receiving from end-devices will allow for another major improvement of 5G networks: the optimization of latency. This will drop from best case 10 ms today to under 2 ms, allowing for time critical systems, such as, e.g., remote surgery, remote control of fast-moving objects, to be using near-real-time data connectivity.

After first pilots in 2018, the first commercial 5G networks were actually launched in 2019. The first 5G devices were made available at the industry's major fair the Mobile World Congress (MWC) in Barcelona this year. At the time this article was written, about one dozen 5G handsets had made it to the market. Apple as one the major players in the smartphone market has not yet implemented it into its brand-new iPhone 11, though.

Along with all the increased data transfer rates of 5G, one more phenomenon is growing in importance: the necessity to operate on these massive amounts of data right after or before such data enters or leaves the network. Computing capacity is required at exactly this point, to guarantee the low latency promise. Sending all this information first through the mobile operator's network, on to the Internet into the cloud, would ruin the time advantage right away. This is where the term "Edge Computing" comes into play, to unlock the key benefits of 5G—speed, low latency, and the capability to handle huge amounts of connected devices in the network.

6 Edge Computing

To put it in simple words: we need to minimize the travel time for data. Edge Computing is about installing computer hardware at the edge of the network, this is at the antenna and base stations. After we have stored our information in the cloud on platforms, we now we need to bring the number crunching power of such services closer to the application needing it in minimum time. Which works just fine with a technology called network and server virtualization. This allows the dynamic allocation of computing power dependent on various criteria to the most suitable location and addressing it as a virtual entity, still.

With the implementation of network virtualization, it does not matter anymore, where you actually locate your systems. Consequently, for the key 5G benefits latency and speed they can be located as close as possible to the Edge. This is exactly what Edge computing is all about (Ericsson 2019). Besides reducing data travel time to and from end application or device, Edge computing also reduces the overall load on the core network of mobile operators, as well as on the backbones of the Internet as such. Therefore, following the goal, that data should not have to travel far in the network, reduces overall data load on the Internet.

One other data hungry technology requiring Edge Computing is the field of AI—Artificial Intelligence and machine learning. Huge data amounts are required for both—for training AI systems, as well as for using AI solutions. While, e.g., text-based language processing, so called chat bots, are running on relatively small amounts of data, requirements are getting bigger once you start using spoken language, as entire recordings need to be transferred. With the analysis of video content, for example to implement face recognitions technologies, or even interpretations of mood and situational context, even more data will need to be exchanged.

The same is true for the thousands of applications not yet invented around the use of sensor data of all kind in all circumstances of human life, industry, environment, farming, etc., which is supported by another recent extension of mobile networks: Narrowband-IoT (NB-IoT) and LTE-M, two standards specifically targeting the connection of billions of smart devices.

Our world is changing into a Matrix of information. Not yet like the movie "Matrix," but with the invention of digital twins of physical realities, we are heading strongly into this direction.

7 Benefits of 5G and Edge Computing

Business is about making money, and so is building 5G networks, installing Edge Computing capacity, or developing AI platforms. The GSMA is expecting mobile network subscriber base to grow up to 5.8 billion subscribers, then representing 71% of the world population. In the same time horizon until 2025, the number of IoT devices is about to scale up to 25 billion. Even if the latter is only half of what Cisco predicted in 2011, it still correlates with revenue expectations of 1.1 trillion USD (GSMA 2019b). There will be literally no aspect of human society that is not going to be connected with what we call the Internet of Things, or to express it more correctly—the Internet of Everything. Digital is the new normal. Or to say it in the words of the German digital transformation insider Karl-Heinz Land: "Everything that can be digitalized, will be digitalized."

And the industry has understood this very well. According to IDG, data analytics is the top IoT benefit for enterprises. The largest target of future investments will go in there, along with IoT Security and IoT Management and Connectivity (IDG 2018). Never before, data had such value for all aspects of human society. May it be your online or offline purchasing behavior, or your medical information, or the

availability of the next bus to come, or the traffic jam prediction of your connected car—all of this needs IoT and data analysis. And thinking beyond the human interaction layer, even more data is being used in background processes steering and controlling machines, our infrastructure, industries, simply our entire civilization.

8 Top Five Use Case Categories

In its "The Internet of Things: Mapping the Value beyond the Hype" report from June 2015, the McKinsey Global Institute was framing nine use case scenarios for the Internet of Things, together with a forecast of revenues and data usage (McKinsey 2015). Many of the IoT cases are can be applied seamlessly for 5G, even though for some this may not always make much sense. The key USPs of 5G may not be even required for the respective IoT application. Let us have a look at the short list of application scenarios for 5G and Edge Computing which seem to be most promising.

8.1 Human Beings

Everything is all about people, so this category deserves to be named first. When talking about IoT, 5G and human beings, some of the older folks may start thinking about "The Borg," a technological enhanced collective society and race from the TV series Star Trek. I guess it does not have to be that scary. Indeed, there is a lot of promising potential in connecting human data with the cloud. We all know the simplest scenario of such data application: the emergency room in a hospital, where somebody is connected to the monitors and in case something develops into the wrong direction, an alarm goes off, calling the doctors for help. This self-explanatory scenario can be easily extended to real life outside of hospitals today. Especially for elderly people, a real-time connection via an IoT monitoring devices can save lives or avoid further complications. In a medical emergency situation, using smart AI analysis on a combination of recent and historical body data jointly with the background of the patient's health files, would drastically improve the quality of medical treatment decisions.

But what works for grandma, does as well for your newborn, while you are watching TV in the other room. And it works for all of us, when going running in the park, hiking the mountains, or riding our bikes. Maybe it is only used as a gamification of our daily activity, reminding us to be more health and body conscious. Fitness trackers like fitbit together with fitness apps like Runtastic, heart-frequency and pulse rate measurement devices, blood-sugar sensors, and much more, are available on the market, as all-in-one smart watches or as specialized application. The value of the combination of 5G and Edge in this context is the possibility to connect real time and at any place where there is coverage to corresponding data sources, AI-health-analysis and further connected features in the cloud. Certainly, a clear case, where each of us would benefit from.

But without doubt, the most convincing argument for 5G is even more simple: billions of mobile subscribers love to use bandwidth, unlimited data packages and transmission speed for all kinds of entertainment: mainly movies and games, which thanks to advanced Edge-based Content Delivery Networks (CDNs) will be delivered as real-time 3D experience, soon. And asking the experts in the GSM arena, that is exactly the main reason why their networks are being upgraded.

8.2 Smart Mobility

When riding an ICE train in Germany about 5 years ago, one could be stunned by the fact that communicated reason for not processed seat reservations, was the missed handover of a 3.5″ floppy disk at the previous train station. Hopefully this is managed by wireless solutions in the meantime, and it shows, how smart mobility starts with such simple services. Of course, trains, planes, trucks, and cars are at the core of smart connected devices, where all of the qualities provided by 5G and Edge Computing phenotypically will come to play. Mobility has bypassed the limits of sustainable growth a long time ago. Traffic jams, parking nightmares, accidents are only the tips of the iceberg of a severely dysfunctional feature of human society. On the contrary, a globalized economy, an interacting world population is not imaginable without mobility either. As a consequence, mobility needs an upgrade.

Together with the advance of electric cars such as Tesla's, a new concept had its breakthrough: An over-the-air, remotely updatable car operating system, with a promise of real self-driving level 5 capability in the foreseeable future. But not only the remote software updates of such cars are requiring high-speed, reliable, and secure mobile communications. A whole bunch of other features and functions in today's cars are using online connectivity. For example, eCall, an automatic accident detector, is alerting first aid and police. Real-time information on delicate functional elements of the car can be forwarded for pattern recognition purposes to the manufacturers' databases and AI systems for failure analysis.

Already today, traffic information is being received and sent by the cars navigation system, informing the driver and contributing to the cloud-based traffic information status on the same hand. Real time and latency critical data for controlling and for steering the car will pull from the cloud and from the surrounding mesh and this will be by far exceeding the perception bandwidth of the human driver. A whole flood of near-real-time information on surrounding traffic situations, special weather conditions, accident warnings and traffic congestion updates, and similar information will be processed by smart vehicles. The human society will remain to be mobile. But in future, humans may be passengers only, no matter if they are using cars, trains, planes, or flying taxis.

8.3 Smart Logistics

One of the oldest IoT Case Studies is the Power-by-the-hour service offering of the jet-turbine manufacturer Rolls-Royce. It was in 1962 that Rolls-Royce started to sell its engines by the hours they were used, instead of selling the machine as a piece. Back then, this was a revolutionary approach which changed the market dynamics and it was hard to handle. You needed quite some effort to pull the usage data together, though (Garvey 2016). Today, accessing such kind of data is possible as a result of global GSM network coverage. Even with plenty of white spots in between, it is pretty much impossible for a plane not to pass a mobile network base station in reach. Today it is not the hours anymore which are being counted alone. It is all data the machine produces, which is of interest for the manufacturer running on a usage-based business model. Predictive maintenance, the early indication of things going wrong in a machine way before it really happens, is based on number crunching pattern recognition, and will get even more reliable with smarter AI to be developed.

Of course, smart logistics encompasses all the other aspects of mobile connectivity, such as tracking and tracing, route planning, avoiding weather conditions by integrating real-time weather information, controlling temperature and humidity of transported goods and much more. But smart logistics may be also about smart contracting, merging the field of export banking and logistics to new set of services. This could be done by deploying blockchain technologies to parcel tracking in combination with Bank Payment Obligations (BPOs) or the older Letter of Credits (LoCs) and a real-time insurance package on top of it. A trustful combination of exactly knowing, where your goods are, where your money is, and that everything is properly insured.

In many of these scenarios it is not the speed and the bandwidth of the network, which is relevant, but it is the promise of 5G of handling a huge amount of individually connected devices, which make the difference to the earlier network setups. Tracking everything, everywhere and receiving a stream of information which can be processed in minimal time to cause a counterreaction based on the given data, is what will change logistics and mobility of the future.

8.4 Smart Environment

Within the context of IoT and 5G, a lot is being talked about smart farming, which really should be considered to be part of a Smart Environment category. Without a doubt, this is a use case where mobile connectivity can help feed the world. The ever-updated information on temperature, moisture in the ground, growth of plants development, the precise identification of need of countermeasures against insects or plant diseases, will help the agricultural industry feeding the world with improved sustainability, less chemistry, and with the outcome of healthier food. The

combination of sensors, smart machinery such as smart harvesters, weed eliminating robots and more are already finding their way from startup status to production.

But there is another aspect, which is not making headlines until we read about disasters: measuring the activity of this planet, of its climate, its forests, its oceans. Scientists have been covering the planet with sensors for all kinds of information and it has been complicated and costly to maintain such data aggregation in the past. With 5G and accompanying IoT network technologies, the concept of Smart Dust (Marr 2018) is becoming reality. Smart dust are smallest devices, also called micro-electro-mechanical systems (MEMS) for collecting and communicating information. The vision is that such can be distributed out of a helicopter or plane in even most difficult terrains. With such technology, our knowledge about what is going on in our entire environment will grow exponentially. Much, which is based on limited data and constructed theories nowadays, will become factual knowledge in the future.

8.5 Smart Industry

What Germany calls "Industry 4.0" is an industry-focused perspective of IoT. Consequently, there is a Germany-only 5G-specification with the possibility to set up nonoperator so-called "Campus Networks." Such are networks limited to a specific and limited area, such as an industrial site. For such locations, dedicated 5G networks can be setup after receiving a special license within the 3700–3800 MHz range from Germanys Federal Network Agency (Bundesnetzagentur). Deutsche Telekom has already launched a campus network prototype together with OSRAM (Telekom 2019) and other industrial players are evaluating the feasibility of such projects.

The reasoning behind using 5G instead of fixed cabling of assets or standard Wi-Fi technology is the combination of secure connectivity with the hope of less interferences, and therefore more network reliability. And—of course—building and operating a wireless network for a whole industrial complex with potentially tens of thousands of connected devices and machines is not an easy job, not even for large corporate IT departments. That is a skill predominantly available at the mobile network operator technology departments. It remains to be proven if the increased cost of building such networks (with niche market equipment being necessary due to the special frequency range) is paying off with the expected business cases and with the productivity gains.

Wrapping it all up, we are truly looking into a future of ubiquitous computing, as once envisioned at the end of the last millennium. The amount of use cases of interconnected devices and applications, the pervasiveness of this technology into all aspects of human live, human society, into business and science will without a doubt represent a next step of development of the human race. 5G and Edge computing are only in-between technologies on this way. And considering the developments of the last 50 years, today's development are rather evolutionary ways of rolling-out ever improving components and successor technologies. Trusting in Garners most recent

hype cycle, 5G is at the peak of inflated expectations, so regarding networks and connectivity, the 5G market is on a train with clear destination: downwards, before it goes up again (Gartner 2019).

9 Conclusions/What Is Left to Do

Being a step away from the cliff of the valley of despair—or to stay with Gartners terminology—the Trough of Disillusionment—5G will develop from a hyped marketing terminology to a very real part of our technological ecosystem. No doubt, the implementation of all of 5G capabilities in their full spread, especially the high frequency, low latency, and max speed vision, is going to take years to be built. Initially, we will be seeing 5G networks mainly in highly populated regions or in specific industrial areas, where either many subscribers are being served, or where IoT use cases will be prototyped. This all will go in parallel with the existing and still being implemented setup of 4G coverage, which will be more than sufficient for many use cases for quite a while. 4G coverage of remote areas, uninterrupted coverage on highways, train lines, and in architectural difficult parts of the city may require preferred attention in the year to come, compared to getting it all on 5G, which may need significantly more antennas and masts than today's network setup.

Rolling out 5G networks across entire countries will be a challenge. Not only does it require enormous investments into many more base stations and antennas, due to the shorter range and reach of high frequency 5G networks, it also requires connecting all of these network elements with fiber cables. Considering the fact that today, at least in Germany, we by far do not have full 3G or 4G coverage across the nation, not to talk about full coverage of every network operator in every corner of the nation, we can develop a feeling of how long it may take to get the 5G job done.

Besides this, it may be a legitimate question to ask, if we really want to entrust any kind of mission critical, security involving data-based decision to be dependent on the availability and quality of the mobile Internet connection at the given time. Just imagine, sitting in a car that requires a continuous high-speed online connection to ensure its self-driving functions. Based on our everyday experience of failing connectivity, or of difficulties with connections in between different mobile network providers, this would not feel very good.

Accepting this dilemma, other solutions need to be considered. Staying with mobility, self-driving cars, automated flying taxis and whatever else the future might bring, will need sufficient AI-power on board, to deliver the essential Level 5 functions also without network availability. Connectivity to Internet-based computing power or other real-time information therefore may be an add-on benefit to improve the user experience of such autonomous systems, only.

Way more important will be a possibility to gather information from your immediate surroundings, beyond sight, but within reach of possible interaction based on direction and speed of your own system's movement. 5G is promising some peer-to-peer P2P functionality down the road, but this remains to be validated. And it would come with the same bad gut feeling, described above. Looking at

the history of peer2peer networks, most prominent is probably the music sharing platform Napster, we do have working concepts of direct connectivity between network elements at hand. During the recent protests in Hong Kong in 2019, a peer2peer chat platform app called Bridgefy has successfully bypassed mobile network-based control mechanisms and spontaneously generated mesh networks of the involved chat participants.

Such mesh networks may have to be developed for mobility and other time critical, moving or regional systems as well. It should be possible to come up with a safe, spontaneous network building standard based on NFC or Bluetooth or any other wireless protocol, or a combination of several protocols, to provide what is obviously needed. Of course, such would be bypassing the centralized approach and control regime of the established mobile network kingdoms of today.

May it be for mobility solutions, or for democracy and freedom of speech, or for simple financial or technological reasons, I have no doubts that we will see this becoming real. We will be living the ubiquitous dream.

Even more than we can imagine today.

References

3GPP. (2019). *About 3GPP*. Accessed September 30, 2019, from https://www.3gpp.org/

Allman, W. F. (2012). The accidential history of the "@" Symbol". *Smithsonian Magazin*. Accessed September 30, 2019, from https://www.smithsonianmag.com/science-nature/the-accidental-history-of-the-symbol-18054936/

Berners-Lee, T. (1989). *Information management: A proposal*. CERN. Accessed September 30, 2019, from https://www.w3.org/History/1989/proposal.html

Bremer, A. (1910). *Die Welt in 100 Jahren*. Verlag Georg Olms.

Bundesnetzagentur. (2019a). *Jahresbericht 2018*. Accessed September 30, 2019, from https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Allgemeines/Bundesnetzagentur/Publikationen/Berichte/2019/JB2018.pdf?__blob=publicationFile&v=6

Bundesnetzagentur. (2019b). *Mobile broadband – Spectrum for 5G*. Allotment. Accessed September 30, 2019, from https://www.bundesnetzagentur.de/EN/Areas/Telecommunications/Companies/FrequencyManagement/ElectronicCommunicationsServices/ElectronicCommunicationServices_node.html

Cerf VG, Kahn RE (1974). *A protocol for packet network intercommunication*. IEEE/Princeton. Accessed September 30, 2019, from http://www.cs.princeton.edu/courses/archive/fall06/cos561/papers/cerf74.pdf

Cisco. (2011). *The internet of things. How the next evolution of the internet is changing everything*. Accessed September 30, 2019, from https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

Cole, T. (2018). *Interview with Kevin Ashton – inventor of IoT: Is driven by the users*. Accessed September 30, 2019, from https://www.smart-industry.net/interview-with-iot-inventor-kevin-ashton-iot-is-driven-by-the-users/

Ericsson. (2019). *Edge computing and 5G*. Accessed September 30, 2019, from https://www.ericsson.com/assets/local/digital-services/doc/edge-computing-5g-report.pdf?

Gartner. (2019). *Gartner hype cycle for emerging technologies*. https://www.gartner.com/smarterwithgartner/5-trends-appear-on-the-gartner-hype-cycle-for-emerging-technologies-2019/

Garvey, W. (2016). *More Than 2.000 BizJets Enrolled In Rolls-Royce CorporateCare Program*. Accessed September 30, 2019, from http://aviationweek.com/nbaa-2016/more-2000-bizjets-enrolled-rolls-royce-corpo- ratecare-program

GSMA. (2019a). *Brief history of GSM and the GSM*. Accessed September 30, 2019, from https://www.gsma.com/aboutus/history

GSMA. (2019b). *The mobile economy 2019*. Accessed September 30, 2019, from https://www.gsmaintelligence.com/research/?file=b9a6e6202ee1d5f787cfebb95d3639c5&download

Hiremane, R. (2005). From Moore's law to intel innovation – prediction to reality. *Technology@Intel Magazine*, Intel.

IBM. (2018). *The first connected coke vending machine*. Accessed September 30, 2019, from https://www.ibm.com/blogs/industries/little-known-story-first-iot-device/

IDC. (2018). *Worldwide developer census, 2018*. Accessed September 30, 2019, from https://www.idc.com/getdoc.jsp?containerId=US44363318

IDG. (2018). *State of the network survey*. Accessed September 30, 2019, from https://resources.idg.com/thank-you/research/snapshot-2018-state-of-the-network-executive-summary?

Information Gatekeepers. (2002). *Wireless access 2000*. Accessed September 30, 2019, from https://books.google.de/books?id=-peYIFTCW74C&pg=PA111&redir_esc=y#v=onepage&q&f=false

Marr, B. (2018). *Smart dust is coming. Are you ready?* Accessed September 30, 2019, from https://www.forbes.com/sites/bernardmarr/2018/09/16/smart-dust-is-coming-are-you-ready/

Martinson, J. (2000). $5m advance for GE years of Jack Welch. *The Guardian*. Accessed September 30, 2019, from https://www.theguardian.com/business/2000/jul/12/books.booksnews

McClelland, C. (2017). *8 Things you didn't know about WiFi. Medium*. Accessed September 30, 2019, from https://medium.com/iotforall/10-things-you-didnt-know-about-wifi-fe638076c0c

McKinsey. (2015). *The internet of things: Mapping the value beyond the Hype*. Accessed September 30, 2019, from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/the-Internet-of-things-the-value-of-digitizing-the-physical-world

Opensignal. (2019). *The state of mobile network experience*. Accessed September 30, 2019, from https://www.opensignal.com/sites/opensignal-com/files/data/reports/global/data-2019-05/the_state_of_mobile_experience_may_2019_0.pdf

Proxicom. (1997). *Proxicom's Little Red Book. A vision for the company and its people*. Reston, VA: Proxicom Corporate Brochure.

Rebaudengo, S. (2012). *Addicted products*. Accessed September 30, 2019, from http://www.simonerebaudengo.com/project/addictedproducts

Sadler, P. (2001). *Management consultancy: A handbook for best practice*. London: Kogan Page.

Speedcheck. (2019). *Mobilfunk report 2019 – Deutschland, EU und USA*. Accessed September 30, 2019, from https://www.speedcheck.org/de/reports/mobilfunk-report-2019.pdf

Techopedia. (2019). *Signaling system no. 7 (SS7)*. Accessed September 30, 2019, from https://www.techopedia.com/definition/25119/signaling-system-no7-ss7

Telekom. (2019). *5G technology in industrial campus networks*. Accessed September 30, 2019, from https://www.telekom.com/en/company/details/5g-technology-in-campus-networks-556692

W3C. (2019). *Facts about W3C – history*. Accessed September 30, 2019, from https://www.w3.org/Consortium/facts#history

Weiser, M. (1991). *The computer for the 21st century*. Accessed September 30, 2019, from https://www.scientificamerican.com/article/the-computer-for-the-21st-century/

Wickham, A. (2017). *The Automotive industry employs more people than you think*. Accessed September 30, 2019, from https://www.fircroft.com/blogs/the-automotive-industry-employs-more-people-than-you-think-71462610395#0

Young, B. (1999). FOCUS-Nokia market cap overtakes BP, No.1 in Europe. Forbes. Accessed September 30, 2019, from https://www.forbes.com/1999/12/07/mu6.html#2c99fff62135

Autonomous Driving on the Thin Trail of Great Opportunities and Dangerous Trust

Sandro Mund and Patrick Glauner

Abstract

Achieving fully autonomous driving cars is a considerable technological milestone that will have significant impact on many lives and the adaption of new technologies. The question of when this milestone will be achieved is currently being debated and contradictory forecasts are increasingly being made. In this chapter, the most important components of self-driving cars are presented and different approaches are discussed. We show what makes autonomous driving so challenging and what misjudgments have been made in the past. In particular, the role of artificial intelligence will be illuminated to give a clear picture of what progress is realistic in the coming years. Next, we discuss related challenges that need to be solved in the coming years. Based on our own research, we will also show how hard it is to interpret models, like neural networks, i.e., understanding why they make the decisions they make in the context of self-driving.

1 Introduction

The conditions and degrees of autonomy under which a vehicle is described as self-propelled differ greatly. A commonly used classification for this are six levels (Society of Automotive Engineers 2014) depicted in Fig. 1.

Starting at zero, i.e., a normal car, a standard vehicle that already contains systems that can intervene briefly in the control system, for example, to prevent the wheels from locking when braking. The first level includes assistance systems

S. Mund (✉)
Trier University of Applied Sciences, Trier, Germany

P. Glauner
Deggendorf Institute of Technology, Deggendorf, Germany
e-mail: patrick@glauner.info

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_12

Fig. 1 Six levels of autonomous driving. Source: authors

that are already commercially successful but require the driver to be constantly and attentively observed. Examples are automatic speed control, lane keeping, and parking. Vehicles reaching the second level can drive most of the time on their own but require attention from the driver to catch mistakes. The following level extends this intervention by a time window and only from the fourth level the control of the driver is no longer necessary. Level four, however, is limited to the fact that this safety can only be guaranteed under certain circumstances or on certain routes. Only with level five a fully autonomous vehicle will be spoken of. The jump to the last level is the most challenging one, because all imaginable traffic situations have to be dealt with.

The social impact of the jump to level 5 is also the greatest. For example, it has been argued that every 23 s someone dies in traffic (World Health Organization 2019). In the future, many lives could be saved by this new technology. Cheap autonomous taxis will threaten jobs; however, the shared use of vehicles will in turn reduce environmental pollution. Many potential consequences are possible and whether they are positive or negative is often debatable. Autonomous driving not only affects society as a whole, but driving itself is also a social act. People give each other hand signals and disregard traffic rules to react to extraordinary situations. Equally important is interaction with passengers for vehicles that cannot drive fully autonomously to ensure that controls are not neglected. Many questions from different disciplines are therefore important to make statements about the acceptance and use of self-driving cars. In order to answer the question of whether a technical implementation is possible at all, there is the particularly important area known as artificial intelligence (AI).

This chapter is structured as follows: In Sect. 2, we discuss a number of challenges that autonomous cars need to master in order to understand their surrounding environment. We argue in Sect. 3 why AI is key to doing so. Meanwhile, in Sect. 4, we first review the history of autonomous driving and then discuss the state of the art as well as a number of predictions that have been made for the foreseeable future. In Sect. 5, we discuss how easy it has become in recent years for a large number of people to acquire the knowledge of how to build and use the complex technologies needed for building autonomous cars. Next, we look at interpretability of machine learning models in the context of autonomous cars in Sect. 6. We also present some of our research results on this topic in the framework of convolutional neural networks. Last, we summarize this chapter in Sect. 7.

2 Understanding the Environment

The architecture of a system for autonomous control of a vehicle is complex. Various components such as sensors, powerful hardware for computations, or the control of the vehicle bus must communicate with each other in real time and ensure reliability. Contradictory signals have to be handled and risks are not always avoidable. Among the most important tasks of such systems are perception, localization, planning, and control. By combining the various sensors, an overall representation of the environment of a vehicle can be obtained that is required for following steps. An intermediate step here is to bring the differently coded information into a uniform shape in order to have a consistent representation of the outside world. In the next step, the system uses this overall information to localize the vehicle. This means to include the position of the vehicle in this model of the outside world. Based on this, the effects of different control signals over a certain distance are calculated and the necessary control signals for this selection are determined.

Different sensors with their own strengths and weaknesses are used in the process. Regular cameras are cheap and have a good range but cannot measure distances. Light detection and ranging (LIDAR) systems fire millions of laser steels per second and measure how long it takes for them to jump back (Cracknell 2007). With this information an accurate 3D map can be created. The costs of a LIDAR are very high and the sensor is not robust enough to be used on a large scale for commercial vehicles. However, since this sensor is preferred for use, a lot of money is currently being invested to solve these problems. In contrast, radars use radio waves to generate images of the environment. Comparing both approaches, LIDARs are more accurate but radar sensors are much cheaper and not prone to fog, rain, or snow.

As the environment becomes more complex, the implementation becomes more challenging. If only machines were involved in road traffic, no AI would have to attempt to predict human behavior. Because of the enormous complexity, the movement in public road traffic presents research and development with a number of challenges. A further dimension in distinguishing autonomous systems is their robustness, e.g., whether they need to be intervened immediately, within defined time windows or not at all by humans. A key threshold here is the achievement of safe driving without human control, so-called fully autonomous driving.

3 The Critical Role of Artificial Intelligence

Systems for controlling autonomous vehicles consist of a multitude of components with different tasks. For example, recognizing the road, predicting the actions of other drivers, or planning the way through the next curve. These components work with various types of information, which can be provided by sensors. The information obtained in this way must be processed, evaluated, and merged in order to enable independent driving. AI is used at various points in such complex systems

and in particular machine learning, a branch of the AI that creates models from data. A particularly known method is deep learning (LeCun et al. 2015), which are neural networks that consist of multiple intermediate layers and are therefore called "deep." For an autonomous vehicle to be able to safely participate in road traffic, many sub-tasks with different requirements have to be mastered in order to ensure the safety of people has priority. Some aspects of deep learning lack a well-founded theory (Lin et al. 2017) and it is generally challenging to verify whether a statistical model works well under all circumstances. Unlike with model-based development, where system correctness can be proven mathematically, there is no absolute certainty. Redundancy in the execution of sub-tasks and their control by testable systems helps but leads to new challenges. It is practically not possible to define abstract rules that cover all potential situations. Data-based models are more scalable because they become better with more data and can thus address situations that can hardly be described by rules.

The increase in performance (i.e., accuracy) comes with the loss of interpretability. A combination of methods of machine learning with knowledge-based systems could solve this problem, but how to do so is a contemporary research challenge. Generally, the fundamental problems of AI have not been solved yet. Human thinking is hardly understood and there is no method that is promising to simulate this intelligence. Artificial general intelligence (AGI) does not appear to be achievable in the foreseeable future without an unexpectedly large breakthrough (Shanahan 2015). Some cases in traffic require abstract thinking to understand complex situations. A person can understand when a stop sign is painted over, stolen, mirrored, or just printed on a T-shirt. Whether such a level of intelligence is necessary to control a car safely enough or whether it is possible to collect sufficient data is questionable. With the focus on methods of machine learning for autonomous driving, it is therefore crucial to have good data that contains rare special cases.

Controlling a car automatically in standard situations is a relatively simple task today. Small amounts of data are sufficient to teach a model how to stay on track and avoid objects. Mastering the remaining fraction of cases is much more challenging as it was assumed several times in the past, though. The question of when fully autonomous vehicles are roadworthy can be answered by determining when enough of these situations can be considered so that self-driving cars are statistically safer than humans. Reaching this threshold of safety is tried feverishly. The commercial success of these vehicles in turn depends on many more factors. There are laws, insurance and production costs, or the acceptance and trust, as well as the social change to mobility-as-a-service, just to name a few.

4 Ambitious Goals and Their Consequences

Today, machine learning algorithms and especially neural networks are a key component for self-driving cars. But this was not always the case. In this section, we review advances in autonomous driving, AI, and contemporary R&D challenges.

4.1 Advances in Autonomous Driving and Artificial Intelligence

Experiments with autonomous vehicles have existed since the beginning of the twentieth century. As early as 1939, General Motors sponsored radio-controlled electric cars powered by electromagnetic fields, generated by circuits embedded in the roadway. Already at that time there were optimistic estimates to have completely autonomous cars in a few decades. It was around this time that a small number of scientists from various disciplines began to discuss how artificial brains could be created, which led to the founding of the field of AI research (McCarthy et al. 1955).

At that time, one was just as optimistic to achieve good results quickly. That optimism, however, was to be paid off. In the 1970s, it became clear that many problems were much more challenging than expected and the high expectations could not be fulfilled, so that no further funding was provided. The time from 1974 to 1980 is often referred to as the first so-called AI winter (Russell and Norvig 2009). However, within this time further research was done and some progress was made. Neural networks at that time were known as perceptrons and only consisted of one unit (Minsky and Papert 1969). It was later mathematically shown that this model's learning capabilities are severely limited (Blum and Rivest 1989). The discovery of how parameters of a multi-layered perceptron (i.e., a neural network) can be trained changed the field substantially (Rumelhart et al. 1985). The broad consequence of following improvements in this learning process was later to be called deep learning and brought the networks back more attention in research (Hinton et al. 2006). Before neural networks regained importance again, however, experts systems therefore were very dominant. But again, the expectations were too high, which led to the second AI winter, which lasted from 1987 to 1993. Within this time in 1989, Carnegie Mellon University had pioneered the use of neural networks to steer autonomous vehicles forming the basis of contemporary control strategies (Pomerleau 1989).

The big start for the development of autonomous vehicles did not happen until a decade later. The second DARPA Grand Challenge was launched in 2005. To win the prize money of 1 million US dollars, more than 200 miles had to be driven autonomously. While in the previous year not a single car completed the course, this time five teams made it (Thrun et al. 2006). Again, there were optimistic voices that autonomous driving would be possible soon and big car manufacturers like BMW, Volkswagen, Audi, and many others started with their own experiments. Google also began in 2009 to secretly work on driving its own cars.

The same year, ImageNet (Deng et al. 2009), a very large and freely available database of more than 14 million labeled images, was launched. The availability of ImageNet has simplified access to data, increased the accessibility to train models with deep learning, and encouraged further research. More and more public libraries for machine learning algorithms appeared and were optimized for calculations on end-user graphics cards, making hardware much cheaper. NVIDIA is a leading supplier of hardware optimized for machine learning. In 2016, they demonstrated in a paper how a car can be controlled by a neural network in order to promote new

products for autonomous driving (Bojarski et al. 2016). Their approach was not new but inspired further projects.

In the same year, it also became known that an AI named AlphaGo had defeated one of the world's best professional Go players (Borowiec 2016). Due to the complexity of the board game, it was assumed that this would only be possible in a couple of decades later. This breakthrough strongly supported the hype of machine learning that continues to this day. AI is finding more and more economic applications and many advances in research have not yet arrived in the wider economy. Whether there will be a new AI winter is questionable, the continuous results speak against it. Further breakthroughs are not unlikely, especially due to the large investments of the automotive industry. But every success seems to be followed by ever greater expectations.

4.2 Contemporary Forecasts and Challenges

Billions are currently being invested and the entire automotive industry is taking AI seriously. Tesla wanted to launch a fully autonomous car in 2019 (Siddiqui 2019). Nissan, Honda, Toyota, and Hyundai have made announcements for 2020 and Volvo, BMW, and Ford-Chrysler for 2021 (Connected Automated Driving Europe 2019). These deadlines appear to be unrealistic, though. Other scientists doubt that it is possible in the near future or at all. Elon Musk, the CEO of Tesla, on the other hand predicts that it will be unusual soon to produce cars that are not fully autonomous. In an AI podcast at MIT, Musk said that Tesla has a big lead (Fridman 2019a). A message on Twitter that 1 billion miles were driven in autopilot mode supports this picture. Tesla is able collect the most data due to a large fleet of sold cars, which are already equipped with cameras (Fridman 2019b). Nevertheless, Musk has previously often postponed his forecasts and many experts are critical of his statements.

The first company that did 10 million miles of autonomous driving is Waymo, which continues the Google Car project (Ohnsman 2018). Their self-propelled car program was initially led by the winner of the 2005 DARPA Challenge (Thrun et al. 2006). What is particularly interesting by comparing the two companies is that their approaches are entirely different. While Waymo uses LIDAR in combination with maps, which is by far the prevailing opinion, Tesla has a strong focus on cameras in combination with computer vision and deep learning. Musk said at an event for investors that LIDAR had no future for autonomous driving and every company focusing on it was doomed. Their approach is also viewed with a lot of skepticism because the camera lacks depth information and is sensitive to bad weather conditions (Templeton 2019).

There is currently no vehicle suitable for the mass market that allows autonomous driving without constant control of the human driver and important questions have not been answered. There is no algorithm that can understand complex traffic situations and many sensors are too expensive or unreliable. In the past, false promises have often been made, but unexpected breakthroughs have also occurred.

To have fully autonomous vehicles in a few years is a very optimistic estimation, though.

5 The Challenge of Easy Access to Complex Technologies

Easy access to new technologies is basically a good thing for a research discipline. More investments lead to more research and economic applications. This process is self-reinforcing and particularly pronounced in the field of deep learning. The flattening out of this hype does not seem predictable yet, which is why the entry into this technology is becoming easier and easier. Meanwhile, there is a multitude of online training courses and public libraries that make it possible to use complex systems within a short period of time without understanding them. This is not a problem in itself, but it can be a source of problems. What makes matters worse is that it is not obvious when gross mistakes are made. The deceptive certainty of an apparently good result can then lead to further damage. In the long run, it can also be just as harmful for a company not to use new technologies that could work really well with a little more specialist knowledge. It also makes sense to have in-house experts to check external results. This also protects against paying huge amounts of money for projects that have already been better solved with freely available software. Even large and established companies assign new employees to customer projects after only a short time. It is therefore risky not to thoroughly check the results of suppliers, but this often occurs in the AI area because the necessary knowledge is lacking.

Equally critical is the one-sided cooperation with universities. One-sided means in this context that students do not have any technical contact persons in the company. It is then dangerous that there is a discrepancy between practice and theory in many areas of machine learning. Data sets provided for teaching are usually unrealistically clean, so, for example, there are few statistical outliers, errors, and missing data points. In addition, the literature hardly deals with these practical problems, which are decisive for project success. In software development, requirements management (Nuseibeh and Easterbrook 2000) is regarded as particularly decisive, while many AI projects are still carried out in a relatively unstructured manner.

The area of autonomous driving gives the impression that these problems are not affected by the high entry barriers. But how realistic is this impression? In fact, there is an entire market for development in this area that makes it easy for small teams to get started. Hardware and software are available in different price ranges. Suppliers have entire kits in their assortment, e.g., (NVIDIA Corporation 2019). This is sensible because relatively little data is available for public research, even if this situation continues to improve. Therefore, it is necessary to conduct experiments under real conditions to develop robust models. However, it becomes problematic when investors are presented with supposed research projects that basically only consist of hardware and software that were bought, installed, and configured. Just as critical are unrealistic expectations and misjudgments.

The entry into the development of a competitive entire autonomous car should not be the goal of a small research team. Ensuring the road capability of a vehicle takes on dimensions that cannot be achieved without large teams that have a lot of expert knowledge and budget. In addition, cutting-edge research is not public and the greatest breakthroughs are therefore made in industry. In turn, companies can see which public research results are published. How far these two worlds are apart is often shown by the attempt to reproduce the results. Often only a small part of research results works under real conditions. Ironically, there are also publications that confirm these findings (Ioannidis 2005). Scientists are under great pressure to publish successful results and mistakes hardly have any consequences. Blind trust is also inappropriate for scientific publications. The evaluation of the conferences in which a paper has appeared can be an indicator for non-experts, but the evaluation of a subject expert is preferable. General authority should not have to be trusted blindly. In order to make sure that quality can also be examined in topics such as deep learning, the following section presents a number of corresponding guidelines.

6 Interpreting Deep Learning Models in Self-Driving Cars

Autonomous cars come with great risks to health and safety of drivers and pedestrians. Therefore, it is important to look into the underlying models and why they make the decisions they make based on input from the environment. Neural networks are often referred to as "black boxes," meaning that is hardly understood why they make specific decision. However, approaches exist to gain a better insight into how they work (Samek et al. 2017). In particular, in the field of image processing there are several methods available (Erhan et al. 2009). In this section, we will analyze this topic in greater detail for autonomous driving.

6.1 Convolutional Neural Networks for End-to-End Driving

Convolutional neural networks (CNN) (Krizhevsky et al. 2012) have been regarded as one of the best methods in the field of image processing for years. CNNs extend neural networks by several layers of filters that learn from the input a hierarchy of increasingly more complex features. Their approach is in some sort inspired by how the human vision system works. The processing of camera images is an important part of autonomous driving. Therefore, CNNs have also gained popularity in this area, because success can be achieved quickly, such as letting a vehicle drive autonomously over a test track. Using only image processing is not suitable for road traffic. For example, the sirens of an ambulance also need to be recognized. Pure CNNs also lack the ability to include previous events in forecasts, i.e., each camera image is viewed individually. For industrial applications, they are therefore only suitable for partial tasks.

Nevertheless, research in this direction is being carried out. In so-called end-to-end driving (Bojarski et al. 2017), a single model controls a vehicle, such as a CNN

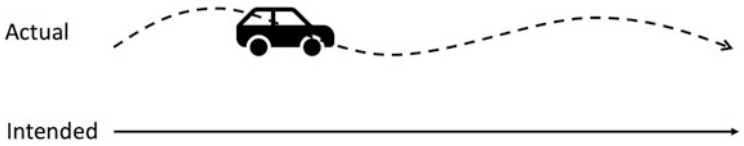

Fig. 2 Sinusoidal trajectory of an autonomous car. Source: authors

that has been trained with historical data. A person drives the vehicle first and the control signals are stored. These signals are the labels, i.e., solutions to the states of the outside world that have been reacted to. The technically simplest way to record these states is to use cameras. The final data set would then consist of a large number of images with the associated control signals. If the problem of driving is broken down to the fact that an action of the driver is to be assigned to every state of the world, it is a classic problem that can be solved using image processing. The final model gets the frames of a camera while driving and returns a control signal for each image. This approach works surprisingly well in simple environments. Simple here means that the weather conditions remain constant and few shadows and reflections influence the forecasts. However, there is a variance in how the car can be controlled. Since for each frame a slightly different control signal is returned, the steering wheel tends to tremble very strongly and the car drives sinusoidally, like being drunk as depicted in Fig. 2. This driving style does not give a feeling of safety. And this kind of safety would not be appropriate either. The steering movements of neural networks are ultimately only statistical statements about similar situations in the training data.

6.2 Visualizing What Deep Learning Models Learn

So why is it useful to deal with CNNs in the context of end-to-end driving? The properties of neural networks can be exploited here to understand how they learn from the data. What makes driving so interesting for image processing is that some elements in the images only change the output in a specific context. Most images show cars, but how a model drives the car is more influenced by where these cars are. For example, a wrong-way driver should have a significantly different effect on a control signal than a vehicle printed on a billboard. Only the context gives the picture elements a meaning that is broken down to a steering signal. In order to understand this particular context and how different models react to it, it is helpful to adapt the data sets so that better models can be trained. A further interesting aspect is the label, i.e., the direction in which the data is directed. In this way, it is possible to determine how certain sections of the input influence the overall result. In other words: Which pixels of a camera image lead to left or right steering? Areas of an image can have multiple clusters, with different influences on the overall result. How a model deals with contradictions or brings them into context is another promising research approach, also for models in other areas of image processing

or deep learning. In safety-critical systems it is particularly important to be able to make statements about how reliable a system is. Visualizing what a model sees is therefore important not only to develop better models faster, but also to understand and test them. Autonomous driving requires a lot of data to deal with as many special situations as possible. Finding blind spots in this data can be accomplished by visual analysis.

Visualizations are well understandable for humans and enable plausibility checks for complex models. For instance, a car drives particularly well on a test track. Also light and weather changes do not influence the performance. Is the model safe now? When visualizing the most important regions of the camera images by coloring using so-called heat maps, it turns out that the distinctive shape of the surrounding trees is the most important influencing factor. A simple approach to create such visualizations is to systematically modify inputs to observe their effects, such as setting the color values of pixels in an area to zero. Through many of these repetitions with different areas, each pixel can then be assigned a relative relevance to the overall image. The resulting heat maps are also called occlusion maps (OM) (Zeiler and Fergus 2014).

One question that we investigated previously was how meaningful OMs are for road traffic situations. Cutting out image areas creates new side effects, since a black pixel also matters to the model. A person would also react if there was a black, indefinable square on the road. One approach to neutralize these unwanted side effects in our research was to invert the OMs (Mund et al. 2018). This inversion is done by removing the areas that were not hidden during the creation of the maps and the other way round. The area remains the same in both cases and only whether the pixels inside or outside this window are masked changes. The two resulting maps can be combined by multiplication. The resulting map can then be used again to color areas on images according to their relevance. We sketch the outcome of the method in Fig. 3.

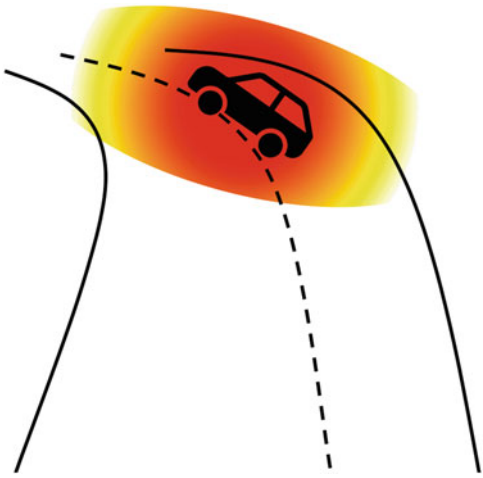

Fig. 3 Occlusion map sketch. Source: authors

It depicts a camera view towards the front of the car shortly before turning left. The generated occlusion map highlights the relevance of pixels for automatically deciding how to steer the car. The color should be interpreted as follows: The redder the region, the greater the importance of the corresponding pixels. The depicted occlusion map shows that the underlying neural network makes reasonable decisions as it mainly considers to region around the street turn ahead of it (where a different car is currently located) for steering left. Readers are advised to read our corresponding paper (Mund et al. 2018) for more detailed real-world visualizations.

It is particularly interesting to use this visualization for videos to highlight in real time what is currently important. It was noticeable that even with only few training data, the important regions of the images are cluster-like, even if they did not make sense at the beginning. However, the analyses are empirical and it is therefore challenging to make general statements. Especially in neural networks, observations often depend on hyperparameters, such as the network structure, or data as general insights into the learning process. We therefore repeated these visualizations and different models that were trained with different data for multiple times. With the addition of more training data, clusters started to cover more parts of the road and became more traceable. However, the visualizations became more unstable in our experiments. This means that per frame very different areas of the camera images became relevant for the model. This showed that although the learned features are sometimes logically comprehensible, the model could not generalize. A human being pays attention to different things during driving, but does not evaluate their relevance very differently many times in a second. Both the evaluation of the model predicts on the test set and the inspection of the visualizations gave a wrong impression. Once again it turned out that the previously chosen metrics were not sufficient. The critical testing of quality criteria remains a crucial and creative process that makes an important contribution.

7 Conclusions

In this chapter, we first introduced the different stages of autonomous driving and the criteria by which they can be distinguished. We then presented the most common sensors and what the most important tasks are for self-propelled cars. Different technologies also have different strengths, but a number of questions of feasibility are still open. We then highlighted the relevance of artificial intelligence and discussed its limitations and why large data sets are important for success. Next, we reviewed the relationship between autonomous vehicles and AI in a historical context. The roots of modern approaches like deep learning are older than often assumed and have a past with heights and depths. We showed that overly high expectations already led twice to so-called AI winters and how the current hype about artificial intelligence was triggered, as well as which optimistic forecasts were made up to date. We then discussed why it is important to build up expertise to maintain controls that exist for other technologies already. Finally, our own research has shown how the decisions of neural networks can be visualized in order to better

understand supposedly safe systems. The plausibility of complex systems can be visualized with simple means.

The future of self-driving cars depends on significant breakthroughs that are not yet predictable. It remains to be seen whether the enormous effort and expense involved will make this success possible.

References

Blum, A., & Rivest, R. L. (1989). Training a 3-node neural network is NP-complete. In *Advances in neural information processing systems* (pp. 494–501). http://papers.nips.cc/paper/125-training-a-3-node-neural-network-is-np-complete.pdf

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., et al. (2016). End to end learning for self-driving cars. arXiv:1604.07316.

Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., et al. (2017). Explaining how a deep neural network trained with end-to-end learning steers a car. arXiv:1704.07911.

Borowiec, S. (2016). Alphago seals 4-1 victory over go grandmaster lee sedol. Retrieved August 1, 2018, from http://www.theguardian.com/technology/2016/mar/15/googles-alphago-seals-4-1-victory-over-grandmaster-lee-sedol.

Connected Automated Driving Europe. (2019). How do automakers perform with their self-driving car timeline? Retrieved December 15, 2019, from http://connectedautomateddriving.eu/mediaroom/how-do-automakers-perform-with-their-self-driving-car-timeline/

Cracknell, A. P. (2007). *Introduction to remote sensing*. Boca Raton: CRC Press.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). Piscataway: IEEE.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal, 1341*(3), 1.

Fridman, L. (2019a). Elon musk: Neuralink, AI, autopilot, and the pale blue dot. Retrieved December 15, 2019, from http://lexfridman.com/elon-musk-2/

Fridman, L. (2019b). Tesla vehicle deliveries and autopilot mileage statistics. Retrieved December 15, 2019, from http://lexfridman.com/tesla-autopilot-miles-and-vehicles/

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*(7), 1527–1554.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105). http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436.

Lin, H. W., Tegmark, M., & Rolnick, D. (2017). Why does deep and cheap learning work so well? *Journal of Statistical Physics, 168*(6), 1223–1247.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. https://web.archive.org/web/20070826230310/; http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Minsky, M., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge: MIT Press.

Mund, S., Frank, R., Varisteas, G., & State, R. (2018). Visualizing the learning progress of self-driving cars. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 2358–2363). Piscataway: IEEE.

Nuseibeh, B., & Easterbrook, S. (2000). Requirements engineering: A roadmap. In *Proceedings of the Conference on the Future of Software Engineering*, pp. 35–46. New York: ACM.

NVIDIA Corporation. (2019). NVIDIA DRIVE: Scalable AI platform for autonomous driving. Retrieved December 15, 2018, from http://www.nvidia.com/en-us/self-driving-cars/drive-platform/

Ohnsman, A. (2018). Waymo racks up 10 million test miles ahead of launching its robotaxi business. Retrieved December 15, 2018, from http://www.forbes.com/sites/alanohnsman/2018/10/10/waymo-racks-up-10-million-test-miles-ahead-of-launching-its-robotaxi-business/#46de3097eb1f

Pomerleau, D. A. (1989). ALVINN: An autonomous land vehicle in a neural network. In *Advances in neural information processing systems* (pp. 305–313). http://papers.nips.cc/paper/95-alvinn-an-autonomous-land-vehicle-in-a-neural-network.pdf

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California University San Diego La Jolla Institute for Cognitive Science.

Russell, S. J., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd edn.). Upper Saddle River: Prentice Hall.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv:1708.08296.

Shanahan, M. (2015). *The technological singularity*. Cambridge: MIT Press.

Siddiqui, F. (2019). Tesla floats fully self-driving cars as soon as this year. Many are worried about what that will unleash. Retrieved December 15, 2019, from http://www.washingtonpost.com/technology/2019/07/17/tesla-floats-fully-self-driving-cars-soon-this-year-many-are-worried-about-what-that-will-unleash/

Society of Automotive Engineers. (2014). Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. Warrendale: Society of Automotive Engineers. https://www.sae.org/standards/content/j3016_201401/

Templeton, B. (2019). Elon Musk's war on LIDAR: Who is right and why do they think that? Retrieved December 15, 2019, from http://www.forbes.com/sites/bradtempleton/2019/05/06/elon-musks-war-on-lidar-who-is-right-and-why-do-they-think-that/#305b7af32a3b

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., et al. (2006). Stanley: The robot that won the DARPA grand challenge. *Journal of field Robotics, 23*(9), 661–692.

World Health Organization. (2019). Global status report on alcohol and health 2018. World Health Organization.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* (pp. 818–833). Berlin: Springer.

Analytic Philosophy for Biomedical Research: The Imperative of Applying Yesterday's Timeless Messages to Today's Impasses

Sepehr Ehsani

Abstract

The mantra that "the best way to predict the future is to invent it" (attributed to the computer scientist Alan Kay) exemplifies some of the expectations from the technical and innovative sides of biomedical research at present. However, for technical advancements to make real impacts both on patient health and genuine scientific understanding, quite a number of lingering challenges facing the entire spectrum from protein biology all the way to randomized controlled trials should start to be overcome. The proposal in this chapter is that philosophy is essential in this process. By reviewing select examples from the history of science and philosophy, disciplines which were indistinguishable until the mid-nineteenth century, I argue that progress toward the many impasses in biomedicine can be achieved by emphasizing theoretical work (in the true sense of the word "theory") as a vital foundation for experimental biology. Furthermore, a philosophical biology program that could provide a framework for theoretical investigations is outlined.

The current collection of chapters in this book is organized around the theme of innovative technologies and "investing in the future." It might therefore appear as peculiar to have a chapter ostensibly focused on philosophy in such a collection.

The original version of this chapter was revised. A correction to this chapter can be found at https://doi.org/10.1007/978-3-030-41309-5_20

S. Ehsani (✉)
Department of Philosophy, University College London, London, UK

Ronin Institute for Independent Scholarship, Montclair, NJ, USA
e-mail: ehsani@uclmail.net; ehsani@csail.mit.edu

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_13

What I hope to achieve here is to make the case that despite many technological advances in biology and medicine over the past 50 years, current biomedical research paradigms are facing significant and seemingly insurmountable challenges in their theoretical foundations. This has resulted in widespread impasses in finding treatments for major categories of human diseases that might come anywhere close to being called a "cure." If we are to make real inroads, the theoretical branch of biology should be reinvigorated so that *theory*, in the true sense of the word, can be reintroduced to and complement experimental biology. This approach is nothing new: it worked quite well in biology in the first half of the twentieth century, and still defines the two-pronged approach that is taken in physics, i.e., theoretical and experimental physics. The history of philosophy, which until the mid-nineteenth century was indistinguishable from "science," can have many relevant lessons for how theoretical strands could be pursued in biomedicine. As such, I will first start by providing a brief overview of the current state of research outcomes in biology and medicine (disciplines which I am considering as interchangeable for the purposes here), followed by an analysis of the place of theory in biomedical research, before moving on to the discussion of the philosophical biology program.

1 Successes and Lingering Challenges in Biomedicine Today

As just noted, biology, particularly molecular biology, has experienced significant changes and technical innovations in the past several decades. Whereas novel insights and descriptive modes of understanding began to emerge from the early applications of molecular biology in the 1970s and 1980s, the widespread popularity of high-throughput techniques and genome sequencing in the 1990s and 2000s have led to the generation of a tremendous amount of new insights and descriptive data about the workings of the cell under normal and disease conditions (Ehsani 2013b). The human impact of these findings has been especially pronounced in the case of monogenic and/or relatively rare Mendelian diseases. A case in point here is the success of gene therapy for β-thalassemia patients (Thompson et al. 2018). Moreover, among human cancers, one can point to successful primary treatments of lymphomas and leukemias [see, e.g., Greaves (2018), Schaapveld et al. (2015)] and to overall "annual reductions of 1 to 2% in age-adjusted cancer mortality rates in the United States for many years" (Varmus 2016). Relatedly, "childhood cancer was once a death sentence, but today more than 80% of children and teenagers survive long term," with the important caveat that "adults who survive cancer as children can suffer long-term health effects" (Couzin-Frankel 2019). In the domain of infectious diseases, the recently rising cure rates in hepatitis C cases are especially noteworthy (Rehermann 2016).

Despite these positive examples, contrary to initial expectations, most common human diseases have remained refractory to various (non-symptomatic) therapeutic interventions, mostly because we have not been able to unify the diseases under common *causative models* or *mechanisms*. In fact, we may often observe in molecular biology research that finding a new "mechanism" in a cellular process

comes to only mean finding "a molecule that is involved in the process" (Garfinkel 2015), which is clearly not in the true spirit of what a "mechanism" entails (Ehsani 2019). Let us consider a few examples. Mutations in the gene coding for the 3685-amino-acid Dystrophin protein implicated in Duchenne muscular dystrophy (DMD) were reported more than three decades ago (Monaco et al. 1986), but as of yet "there is no specific treatment for DMD cardiomyopathy, in large part due to a lack of understanding of the mechanisms underlying the cardiac failure" (Yucel et al. 2018). In the case of type 1 diabetes, in addition to the important issue of access to insulin, Linda DiMeglio and colleagues note that "clinicians, investigators, and patients have gained a better appreciation of the true complexity of type 1 diabetes, and humility in the face of many unsuccessful trials aimed at inducing a durable disease remission" (DiMeglio et al. 2018). In the field of Parkinson's disease research, Heiko Braak and co-authors write that "despite remarkable progress in the management of its motor symptoms by pharmacologic dopamine replacement or deep brain stimulation, there is still no cure and all attempts to develop treatments that halt or slow down the relentless progression of the disease have so far failed" (Braak et al. 2018). A similar diagnosis can be said of Alzheimer's disease (AD) research. Commenting on the halt of a trial on inhibiting the β-secretase (BACE) enzyme, Paul Aisen writes:

> This is one more terribly disappointing result in our field. BACE inhibition would seem to be a powerful tool against the accumulation of toxic amyloid species that drives AD neurobiology in the early stages of the diseases. Secretase inhibition has been a leading candidate for primary prevention. But both γ-secretase and β-secretase inhibition are associated with a risk of cognitive worsening and other adverse effects that have halted all major development programs of these drugs to date. Off-target and on-target mechanisms behind these effects remain uncertain. It is possible that substantial inhibition of amyloid peptide generation interferes with synaptic function. (Aisen 2019)

Moving on, what about the case of infections and the immune system? Writing in 2012, the microbiologist Kim Lewis conveyed the important message that "the more we know about antibiotics, the fewer we can discover" and that "this is essentially why the field is in trouble—only one antibiotic belonging to a new class, the narrow-spectrum daptomycin, was discovered and made it into clinical practice in the past 50 years" (Lewis 2012). Advising on treatment strategies for malaria, the WHO Strategic Advisory Group on Malaria Eradication warned that "even with our most optimistic scenarios and projections, we face an unavoidable fact: using current tools, we will still have 11 million cases of malaria in Africa in 2050" (WHO 2019). In the case of cancer immunotherapy, Alex Jaeger and colleagues make the following observation: "Despite the accumulation of extensive genomic alterations, many cancers fail to be recognized as 'foreign' and escape destruction by the host immune system. Immunotherapies designed to address this problem by directly stimulating immune effector cells have led to some remarkable clinical outcomes, but unfortunately, most cancers fail to respond, prompting the need to identify additional immunomodulatory treatment options" (Jaeger et al. 2019). This is compounded by the fact that "the majority of proposed anticancer treatments do not succeed in advancing to clinical use because of problems with efficacy or toxicity, often for unclear reasons" (Lin et al. 2019).

The overview of challenges in the pathobiological research just provided may not be surprising because many seemingly "lower-hanging-fruit" problems in biology remain open and appear to be intractable, or at least quite difficult, for us to gain a clear understanding of in the context of current research frameworks. Just to illustrate this point, let us review some examples at different levels of "complexity."[1] At the basic, *chemical level*, deciphering the chemical structure of water, which is of paramount importance for questions such as the elucidation of protein structure, is a matter of great challenge in physical chemistry and an area of active research (Tulk et al. 2019; Yang et al. 2019). At the *protein level*, a big mystery is the presence of "intrinsically disordered" domains in the structure of proteins (Robustelli et al. 2018), with different functions attributed to their presence [see, e.g., Shrinivas et al. (2019)]. Also at the protein level, there is the question of the enigmatic in vivo formation of protein crystals (e.g., Charcot-Leyden crystals formed of the Galectin-10 protein)[2] implicated in inflammatory and other unknown (and "normal") biological processes (Allen and Sutherland 2019; Persson et al. 2019). At the *cellular level*, cell structures that might be sensitive to magnetic field changes remain controversial (Servick 2019). At the *organism level*, new types of puzzlement come to light. One such puzzle is stated as follows by the immunologist Simrit Parmar: "A fetus is 50% mum and 50% dad—it's effectively the biggest allogeneic transplant in biology [. . .] Nature itself has created a mechanism to stop the baby being rejected by using [regulatory T] cells to protect against the inflammatory onslaught from the mother's side" (Arney 2018). Understanding facets of this natural puzzle better would have implications for conditions such as graft-versus-host disease. Moving on to the *pharmacological level*, there are still many research questions to be asked concerning the "inactive" ingredients (excipients) in pharmacological formulations that are meant to have one "active" ingredient, and, more generally, what makes for an "active" versus "inactive" ingredient (Reker et al. 2019). Of course, cells in the body do not discriminate between active and inactive compounds, and as such excipients have individual and combined (side) effects on the body and on the active ingredient (and its putative molecular targets).

Finally, and as the last example here, at the *methodological level*, questions of (1) when an effect is really an effect, and (2) certainty, such as with the use of *P*-values, are salient in that they remain debatable. Matthew Kramer, as a case in point, puts forward the following suggestion: "The real question is whether a treatment effect is important, not whether it differs 'significantly' from a control. To answer this, the researcher should justify beforehand how large the effect size needs to be. Then, if a 10% improvement over the control is required, the probability that this has been attained can be calculated from the data using familiar statistical tools for hypothesis testing and sample-size determination" (Kramer 2019). John Ioannidis proposes that "focusing on effect sizes can often be better than determining whether an effect

[1] By "complexity" I simply mean to indicate the greater number of known (and potentially unknown) variables involved in a given level of investigation of a biological phenomenon.

[2] These crystals were identified in the 1850s (Su 2018).

exists." Overall, perhaps Albert Einstein (1879–1955) said it best in 1922, albeit in a different context, when he observed that "as far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality" (Einstein 1922).

What I think the various challenges at different levels of investigation reviewed in this section point to is that substantial work needs to be done on the theoretical underpinnings of biological questions: theoretical work that should start with the simpler questions, ones for which fewer potential variables could be envisaged.

2 The Current State of Theory in Biomedical Research

A reaction to the need for theory reintroduction to biomedicine might be to state that, surely, theory *does* exist in the field today. It is evident that theory has always been an indistinguishable part of the modern biological sciences, from ecology (Odenbaugh 2013), evolutionary theory, and microbiological basis of disease (Shou et al. 2015) to the elucidation of DNA and protein structures and networks of gene regulation (Britten and Davidson 1969; O'Malley 2010). Lymphocytic V(D)J recombination in adaptive immunity (Dong et al. 2015) and the elucidation of friction reduction by bacteria in their medium (Hatwalne et al. 2004; Lopez et al. 2015; Marchetti 2015) are two instances of the successful implementation of a strong theoretical model through to experimental validation. Another example of the usage of sound theoretical arguments in advance of establishing an experimental paradigm is the focus on siderophore quenching strategies to avoid the emergence of antibiotic resistance in a bacterial community (Ross-Gillespie et al. 2014), which in essence shift the burden of antibiotic resistance from individual bacterial cells or colonies to a microbial community. In fact, it appears that the immune system may utilize a similar strategy as part of its own defensive mechanisms (Nakashige et al. 2015).

Nevertheless, the theoretical underpinnings that do currently exist, barring some exceptions, are not part of a systematic framework of investigation that is adhered to consciously. In fact, in the era of high-throughput and big-data experiments, a notion has become prevalent that observations and data collection can be pursued independently of prior theories. This cannot be the case, since no process of data collection, however carefully planned, can be completely devoid of bias (MacCoun and Perlmutter 2015). Again in the words of Einstein, "it is the theory that determines what we can observe," a statement which was followed by the theoretical physicist Werner Heisenberg's (1901–1976) comment that "we have to remember that what we observe is not nature in itself, but nature exposed to our method of questioning" (Bodner 1986). A pertinent example here could be the utility of Alan Turing's (1912–1954) theoretical reaction–diffusion model to the observation of "Turing-like features in the periodic pattern of digits" in developing limb buds (Raspopovic et al. 2014; Zuniga and Zeller 2014).

Presently, "theoretical biology," with some exceptions, has become mostly synonymous with computational biology and the application of mathematical

models to various forms of data structures.[3] Congruent to this conclusion, let us consider the five modalities of theoretical biology at work today as identified by Massimo Pigliucci: (1) analytical modelling (e.g., mathematical/formal models in population genetics); (2) statistical modelling (e.g., quantitative genetics); (3) computer modelling (e.g., genetic networks); (4) verbal-conceptual models (e.g., conceptual diagrams based on experimental results); and (5) philosophy of biology (Pigliucci 2013). At the moment, however, because of a host of issues such as the great number of unknowns in even "simple" biological phenomena, "theoretical" work in the tradition of the history of rational thought (which will be touched upon in the following section) is few and far between, and even in those works that do fall into this category, it is important to point out that "approximating observational phenomena is very different from formulating an explanatory account of a significant body of empirical data" (Everaert et al. 2015).

The time is therefore ripe to reintroduce genuine theoretical analysis back into biology. But where would new theories, or, better put, *inspirations and approaches toward new empirical questions and theories*, come from? One source could be philosophy (i.e., philosophy of science, philosophy of mathematics, philosophy of language, the history of philosophy, etc.), in the form of *philosophical biology*.[4] Such an approach, distinct from (but still utilizing) the philosophy *of* biology, could endeavor to search for, propose, and develop questions and answers in the true spirit of the theoretical sciences using a vast array of tried-and-tested analytical philosophical tools that have been developed over many centuries. It has to be emphasized that the goal is not to produce *only* theories or theoretical narratives, because, in the words of the evolutionary biologist Richard Lewontin, "there is no end to plausible storytelling" (Lewontin 1998). Rather, the theories should be accompanied by inherently testable sets of questions and possible solutions.

3 Lessons from the History of Philosophy and Rational Thought

As the aim of this chapter is to suggest a framework whereby philosophical approaches can find their way back into mainstream biological research, a brief survey of the apposite history of philosophy and rational inquiry is presented in this section. Each subsection here, particularly Aristotle's biology, can be a vast and separate strand of investigation. Nonetheless, the purpose here is to look at snippets of inspiration for a philosophical biology framework. Before proceeding, however, it is important to point out why looking for lessons in philosophy applicable to

[3]For some examples, see Armiento et al. (2016), Asatryan and Komarova (2017), Bertsch et al. (2017), Editors (2016), Leek et al. (2010), Tadrist and Darbois-Texier (2016), Weiss et al. (2003).

[4]The phrase "philosophical biology," or "philosophical science" in general, would have seemed pleonastic to the scientists of the Enlightenment and later periods, but today this pleonasm may be necessary.

current problems in biomedicine could be a fruitful strategy: first, philosophy does not have a "state-of-the-art." This is fortunate, I think, because philosophy is simply a mode of thought and inquiry that once arrived at, can stand the test of time and be applicable to different situations and scenarios. This is not surprising, because modern humans' cognitive capacities have not changed much since the emergence of our language faculty (Berwick and Chomsky 2015, 2017), and as such the philosophical achievements of Plato (ca. 428–348 BC) or Aristotle (384–322 BC) more than 2400 years ago may represent some of the limits of what could be achieved theoretically in certain domains of thought. Second, true philosophy is not based on mere debates,[5] where there is no room for the interlocutors to change their minds and learn from the other, but rather, philosophy is based on arguments that could build on each other and that allow for thought experiments to advance one's knowledge.

3.1 Ancient Philosophy

Plato's dialogues offer a wealth of concepts relevant to the discussion here. The rational question-and-answer-based method Plato uses in the dialogues is usually called the Socratic (or "elenctic")[6] method, which is of a "maieutic"[7] nature (Leigh 2007). In other words, a sequential and adaptive question-based style of reasoning can lead one (or a group of individuals) to introspect toward new and improved reasoning. This can be said to parallel the process of hypothesis generation in a scientific inquiry. Each dialogue, such as the *Theaetetus*, provides specific instances of the usages of this methodology. Clark Glymour and colleagues point out that the dialogue *Meno* is "the source of a [philosophical] method: conjecture an analysis, seek intuitive counterexamples, reformulate the conjecture to cover the intuitive examples of the concept and to exclude the intuitive non-examples; repeat if necessary" (Glymour et al. 2010). This seems to be the perfect recipe for thought experiments.

One can also find hints of the use of simple models for testing or observation before moving on to the actual phenomenon in question. This can be read in the *Sophist*, when the Eleatic Stranger/Visitor says to Theaetetus, a mathematics pupil (and later of great fame as a geometer): "when it comes to grappling effectively with any of the big subjects, everyone has long thought it best to practise on small and easier things before moving on to the big ones themselves"[8] (218c5/d1) and also that "we should pursue something of no consequence and try to establish it as a

[5] Any point, however absurd or incorrect, could theoretically be debatable or arguable, and if this becomes the basis of a philosophical and scientific investigation, there would be no room for true progress.

[6] From the Greek *elegkhos* "refutation."

[7] From the Greek *maieutikos* "acting as midwife."

[8] The *Sophist* translations are from Christopher Rowe's edition (Plato 2015).

model for the more important subject" (218d5). There is an analogous message in Aristotle's *Parts of Animals* (*PA*): "If any person thinks the examination of the rest of the animal kingdom an unworthy task, he must hold in like disesteem the study of man. For no one can look at the primordia of the human frame—blood, flesh, bones, vessels, and the like—without much repugnance" (*PA* I.5).[9]

In addition to philosophy and logic, Aristotle's perceptive and meticulous observations of nature make him a foremost naturalist (Romanes 1891). His writings on biology have received varying levels of attention from scholars in different periods. Sophia Connell notes that "because Aristotle himself does not attempt to distinguish the biological from the philosophical, it makes sense to read all Aristotelian texts as potentially representative of the same philosophical outlook" (Connell 2001). A great portion of Aristotle's observations are detailed descriptively which may be followed by inferred conclusions. The observations themselves may be firsthand or referenced from others. Commenting on Aristotle's *History of Animals* (*HA*), for example, I. M. Lonie notes: "In a celebrated passage [*HA* III.2 and III.3] [Aristotle] describes the theories of Syennesis, Diogenes of Apollonia, and Polybus, on the blood vessels, in all of which the heart is subordinate to the brain. After recording their views, Aristotle remarks [*HA* III.3] that these men and other natural philosophers were mistaken: the blood vessels begin from the heart, not from the brain" (Lonie 1964).

But there are also general accounts to be found in Aristotle's biology. A case in point is his four categories of traits which could frame one's investigation of life in nature: ways of life (*bioi*), actions and activities (*praxeis*), dispositions and character (*ethê*), and parts (*moria*) (Depew 1995). Of these general accounts, I would like to point to a few methodological proposals. First, in *PA* I.4, Aristotle indicates *analogy* and *difference measurements* as two modes of comparison: "Groups that only differ in degree, and in the more or less of an identical element that they possess, are aggregated under a single class; groups whose attributes are not identical but analogous are separated." The method of "the more and the less" is quite reminiscent of the earlier discussion on *P*-values and effect size.[10]

In the *Generation of Animals* (*GA*), Aristotle makes an important distinction between the *potential* and the *actual*, stating that "all three kinds of soul [nutritive, sensitive and rational] ... must be possessed potentially before they are possessed in actuality" (*GA* II.3).[11] In *De Anima* (*On the Soul*) III.5, he explains this notion to a greater extent, writing that "in a sense light makes potential colours into actual colours."[12] Connell provides a further helpful example from *Metaphysics* IX.7: "is

[9]From William Ogle's translation (Aristotle 1882) (see also: online text by D. C. Stevenson, The Internet Classics Archive, MIT; classics.mit.edu/Aristotle/parts_animals.html).

[10]There are commonalities here with John Stuart Mill's "method of difference" (Mill 1843).

[11]From Arthur Platt's translation (Aristotle 1910) (see also: online text at Wikisource digital library; en.wikisource.org/wiki/On_the_Generation_of_Animals).

[12]From J. A. Smith's translation (Aristotle 1931) (see also: online text by D. C. Stevenson, The Internet Classics Archive, MIT; classics.mit.edu/Aristotle/soul.3.iii.html).

earth potentially a human being? No [. . .] just as earth is not yet potentially a statue, because it must undergo a change before it becomes bronze" (Connell 2001). The potential/actual dichotomy may first and foremost bring to mind today's fields of developmental biology and genetics/inheritance. But it also brings into discussion the potential role of a living being's environment, and epigenetics. Although beyond the purview of our present discussion, I think it is important to mention that reading the *GA* with an eye on epigenetic development (Henry 2018) should always be accompanied by the determinants of "scope and limit": A fish embryo, although susceptible to certain variations, cannot naturally develop into a bird, or another species: a scope comes hand-in-hand with limits, and therefore epigenetic variation in development is considerably constrained.[13]

A third, and perhaps the most famous of Aristotle's methodological proposals, is the categorization of causes. In *Physics* II.3, he introduces the four as follows[14]: *Material cause* is "that out of which a thing comes to be and which persists," e.g., "the bronze of the statue." *Formal cause* is "the form or the archetype." *Efficient or moving cause* is "the primary source of the change or coming to rest" or "what makes of what is made and what causes change of what is changed." *Final cause* is "in the sense of end or 'that for the sake of which' a thing is done, e.g. health is the cause of walking about." In giving illustrations of each of the causes, one could rely on examples from the crafts, but as Connell points out, "the natural world is not constructed and does not work just like the crafts; indeed, the reverse seems to be the case—crafts copy nature. Natural objects take priority in Aristotle's ontology, possessing properties that crafts will never be able to exemplify" (Connell 2001). Can one or more types of causes be reduced to each other under some circumstances? This appears plausible, particularly for biological applications. John Cardwell relayed a similar message more than a century ago: "as 'form' includes, by definition, *all* the properties of a material thing, the 'formal cause' may, in some instances, include both the 'efficient' and 'final' causes, thus reducing the four to two, and bringing one back to the primal dual postulate, i.e., matter with 'form'" (Cardwell 1905).

This dual theme of matter and form is quite important and pertinent to some of the current impediments in biomedical research, for the focus in the discipline—for practical or other reasons—has usually solely been on material causation. Here in particular I have in mind the mechanistic framework of investigation in contemporary cellular and molecular biology. Might it be possible to augment

[13]Noam Chomsky made a very pertinent comment related to this point in 1983: "Consider something that everybody agrees is due to heredity—the fact that humans develop arms rather than wings. Why do we believe this? Well, since nothing in the fetal environments of the human or bird embryo can account for the differences between birds and men, we assume that heredity must be responsible. In fact, if someone came along and said that a bird embryo is somehow 'trained' to grow wings, people would just laugh, even though embryologists [context: 1983] lack anything like a detailed understanding of how genes regulate embryological development" (Chomsky 1983).

[14]From R. P. Hardie and R. K. Gaye's translation (Aristotle 1930) (see also: online text by D. C. Stevenson, The Internet Classics Archive, MIT; classics.mit.edu/Aristotle/physics.html).

mechanisms with biological "binding principles," with the former acting as the material and the latter as the formal causes in an intelligible account of a biological phenomenon (Ehsani 2019)? Moreover, a research area today where ascriptions of causality are in need of significant attention and work might be the field of randomized controlled trials, which "seem poorly suited for answering questions related to why therapies work in some situations and not in others and how therapies work in general" (Carey and Stiles 2016). In these trials, it is not uncommon to have a classification called "all-cause mortality." In a trial published in 2018 (McNeil et al. 2018), for example, the list of all-cause mortality included: cancer, cardiovascular disease, major hemorrhage, "other," and "insufficient information" (12 out of 1052 patients). For concepts such as "all-cause mortality" and related (and derived) theoretical notions, much can be done along the theme of this section.

Aristotle's methodology may often be thought to revolve around *aporíā* (i.e., difficulties, impossibilities or puzzles). Michael Frede makes a connection between Aristotle's approach to such puzzles and that of Plato in the *Sophist*: "[The *Sophist*] sets out carefully constructing a series of puzzles, *aporiai* [. . . and] then it turns toward a resolution of these *aporiai*. In this regard the procedure of the dialogue reminds one of the methodological principle Aristotle sometimes refers to and follows, the principle that on a given subject matter we first of all have to see clearly the *aporiai* involved before we can proceed to an adequate account of the matter, which proves its adequacy in part by its ability both to account for and to resolve the *aporiai*" (Frede 1992, p. 423).

A few centuries after Aristotle, Galen (of Pergamon, (ca. 129–210 AD)) also made his own lasting impressions on the philosophical pursuit of human biology.[15] Ronald Christie reminds us that "what Galen taught is of great importance since his writings dominated medical education for the next 1500 years" (Christie 1987). Eva Del Soldato, writing on the "Renaissance debate over the superiority of Aristotle or Galen," observes that "Aristotle was regarded by physicians as an important authority because of his philosophical system, but Galen had offered in his works more precise observations of the human body. Nonetheless, since many points of their disagreement (e.g., the localization of the brain functions) were merely founded on speculation, some doctors preferred to demonstrate the harmony between Aristotle and Galen in order to overcome this impasse" (Del Soldato 2019). Galen himself thought highly of Aristotle and Hippocrates[16]: "All these and many other points besides in regard to the aforesaid faculties, the origin of diseases, and the discovery of remedies, were correctly stated first by Hippocrates of all writers whom we know, and were in the second place correctly expounded

[15]Interestingly, and of relevance here, one of Galen's writings is entitled "The Best Doctor is Also a Philosopher" (*Quod optimus medicus sit quoque philosophus*) (Singer 2016).

[16]One would expect, perhaps, that given Aristotle's focus on health and disease, he would have discussed the work of Hippocrates the physician (ca. 460–370 BC) (distinguished from the geometer/astronomer Hippocrates of Chios) in some fashion, but I have only found one reference in the Aristotelian corpus to Hippocrates in *Politics* VII.4.

by Aristotle" (*On the Natural Faculties* II.4).[17] It is in this context that I would like to briefly move back a few centuries before Galen and mention Erasistratus (ca. 304–250 BC), who "was regarded by his followers as a successor to Aristotle and Theophrastus" (Lonie 1964). Christie, writing on Galen's critical reception of the teachings of Erasistratus, remarked that "the school of the Erasistrians survived until after the time of Galen, who did a great disservice to medical progress by destroying its credibility with rhetoric based on sarcasm and ridicule" (Christie 1987). This is due to the fact that "Erasistratus discarded most of the humoural theory of disease in favor of one based on changes in individual organs," which is closer to modern medical approaches. But in certain other areas, Erasistratus did not make progress *from today's point of view*: Lonie, for example, points out that "what prevented Erasistratus, and any other ancient physiologist, from advancing a systematic hypothesis on the circulation of the blood was not so much the failure to realize how it might be possible mechanically. The obstacle was more deep-seated than that: it was a failure to see beyond the analogies which they employed in their physiological systems [e.g., the blood supply vs. an irrigation system]" (Lonie 1964). Here one is again reminded of the machine analogy that is inherently tied to mechanistic accounts in modern biology. Other commonly-used analogies such as protein "folding," "intrinsic disorder," and "interaction" are further instances of this phenomenon. Overall, being cognizant of the limitations of analogies borrowed from their ordinary language usage may limit their potential pitfalls.

3.2 After the Galilean Revolution in Science

If openness toward natural puzzles, paradoxes, and thought experiments is one message from the preceding section, then perhaps Galileo Galilei (1564–1642) is a quintessential figure in the history of rational thought adhering to this method. Galileo allowed himself to be puzzled by seemingly mundane phenomena, leading him to perform "scientific" (i.e., rational) thought experiments where few others—as far as can be ascertained—had made any significant advances. Nowhere is this more pronounced than his thought experiment about a moving ship (1632), which convincingly demonstrated that to a person present on a stationary ship versus one on a moving ship with constant velocity, all types of motion would appear the same in both scenarios. The result of Galileo's work, along with those of René Descartes (1596–1650) and Cartesian philosophers following him was the establishment of the mechanical philosophy as an intelligible overarching account of natural phenomena and the appreciation that the world was *directly* understandable (Chomsky 2009). The new science of mechanics of Isaac Newton (1643–1727) changed all of this, whereby *action at a distance* could no longer allow for a cogent account of "matter" and "physical" to be given. Thus, the effect of the new Newtonian mechanics was to

[17] From Arthur John Brock's translation (Galen 1916) (see also: online text by D. C. Stevenson, The Internet Classics Archive, MIT; classics.mit.edu/Galen/natfac.html).

revive some of the Aristotelian (and Scholastic) notion of "mysteriousness" in the science of the day. Indeed, this mysteriousness about the nature of matter remains to this day.

What Newton tried to avoid was "explaining what is 'unknown' by what is 'more unknown'" (Cohen and Smith 2004, p. 25). This mantra, along with Christie's caution against "dogmatism [which] can be a dishonest or insincere substitute for ignorance" (Christie 1987), become especially apropos if the implications of Newton's undermining of the mechanical philosophy are to be implemented. Namely, post-Newton, it became clear that the world cannot be directly intelligible and hence only the intelligibility of *theories about the world* could be contemplated (Chomsky 2014b). When one attempts to—in the words of the physician Paracelsus (1493–1541)—"inquire of the world" (Lister 1957), rather than merely recording and measuring it, the end result of the process would be to attain simpler, more intelligible and better explanatory theories. Therefore, theory is not *removed from* the reality of nature; rather, as far as we as human beings can tell, theory *is* our reality of nature.

Let us now pick one important theory of nature, i.e., that of causality, and briefly investigate how it changed post-Galileo compared to Aristotle's account of the four causes. John Locke (1632–1704), in the first volume of his 1689 *An Essay Concerning Human Understanding*, offered an account of causality that may still inform today's use of mechanistic explanations in biology: "for to have the idea of cause and effect, it suffices to consider any simple idea or substance, as beginning to exist, by the operation of some other, without knowing the manner of that operation" (Locke 1894).[18] A few decades later (in 1748), David Hume (1711–1776) offered an account of causality, and a commentary on its "ultimate springs and principles," that rings true with an enduring tone:

> Hence we may discover the reason why no philosopher, who is rational and modest, has ever pretended to assign the ultimate cause of any natural operation, or to show distinctly the action of that power, which produces any single effect in the universe. It is confessed, that the utmost effort of human reason is to reduce the principles, productive of natural phenomena, to a greater simplicity, and to resolve the many particular effects into a few general causes, by means of reasonings from analogy, experience, and observation. But as to the causes of these general causes, we should in vain attempt their discovery; nor shall we ever be able to satisfy ourselves, by any particular explication of them. These ultimate springs and principles are totally shut up from human curiosity and enquiry. Elasticity, gravity, cohesion of parts, communication of motion by impulse; these are probably the ultimate causes and principles which we shall ever discover in nature; and we may esteem ourselves sufficiently happy, if, by accurate enquiry and reasoning, we can trace up the particular phenomena to, or near to, these general principles. The most perfect philosophy of the natural kind only staves off our ignorance a little longer: as perhaps the most perfect philosophy of the moral or metaphysical kind serves only to discover larger portions of it.

[18] See also: online text by Project Gutenberg; gutenberg.org/files/10615/10615-h/10615-h.htm.

> Thus the observation of human blindness and weakness is the result of all philosophy, and meets us at every turn, in spite of our endeavours to elude or avoid it. (Hume 1902)[19]

4 Precedents of "Philosophical Biology"

Looking back over the past two centuries, how can we characterize the philosophical development of the as-yet non-specialized field of "biology" among the natural sciences? Cécilia Bognon-Küss and Charles Wolfe note in their volume *Philosophy of Biology Before Biology* that "the word 'biology' was simultaneously and independently coined by several authors from different national and disciplinary backgrounds ([Michael Christoph] Hanov 1766, [Marie François Xavier] Bichat 1800, [Jean-Baptiste] Lamarck 1809, [Gottfried Reinhold] Treviranus 1802–1822)" (Bognon-Küss and Wolfe 2019, p. 5). Although biology was thus beginning to diverge from the more "practical" field of medicine and was perhaps more philosophical due to its theoretical origination, philosophy's role in the discipline varied contextually. For example, the eighteenth-century Scottish surgeon John Hunter (1728–1793) is said to have suggested to his vaccine-pioneering student Edward Jenner (1749–1823): "Don't think. Try" (Barry 2005; Bartley 1999). In the 1850s, Rudolf Virchow (1821–1902), the pioneer of cellular pathology (Bagot and Arya 2008), is famous for having quoted the physician Salomon Neumann (1819–1908) that "medicine is a social science" (Kottke 2011). He also stated that "medicine as a social science, as the science of human beings, has the obligation to point out problems and to attempt their theoretical solution; the politician, the practical anthropologist, must find the means for their actual solution" (JRA 2006). As a further illustration, a commentary appearing in the *North American Review* in 1868 stated that the "great questions of biology, considered in its philosophical aspect, are three: What is the origin of life in the first instance? What is the origin of species or the different forms of life? What are the causes of organic evolution in general?" (Abbot 1868).

It may be worthwhile to point to a number of more specific cases in this period where philosophy, theory and experimentation demonstrate an intertwined relationship. To begin with, in 1806, the chemist Theodor Grotthuss (1785–1822) proposed a theory of proton tunneling across hydrogen bonds (Cukierman 2006). The Grotthuss mechanism remains an enigmatic and very relevant question and phenomenon in studies of water structure and water-protein interactions. Theoretical investigations into hydrogen bonding in water remain an active area of research, as for example the theoretical chemist David Clary notes that "the excellent detailed agreement between the quantum dynamical calculations and experimental data shows that theory is getting much closer to a highly accurate description of water and, thus, to providing a detailed quantitative understanding of hydrogen-bond

[19] Based on the posthumous edition of 1777, Section IV (see also: online text by Project Gutenberg; gutenberg.org/files/9662/9662-h/9662-h.htm).

dynamics" (Clary 2016). In 1872, the physician Casimir Davaine (1812–1882) put forth the idea of "passages" in microbiology by studying *Bacillus anthracis* virulence in blood samples, a concept that has been a mainstay of any microbiological experiment up to this day. The duality of "humoralism" versus "cellularism" began to take shape in 1882, with the development of Ilya Metchnikoff's (1845–1916) theory of phagocytosis (Tauber 2003). Although initially derided by some as a "fairy tale" (Vikhanski 2016), this theory still resonates today in immunological research on phagocytic cells and inflammation (Gordon 2016). Around the same period, Louis Pasteur's (1822–1895) "discovery that microbes discriminate between *D*- and *L*-substrates [. . .] had been given little attention until taken up by [Emil] Fischer [1852–1919], who suggested that 'the yeast cells with their asymmetrically formed agent are capable of attacking only sugars of which the geometrical form does not differ too widely from that of *D*-glucose'" (Barnett and Barnett 2011), leading to Fischer's proposition of a "lock and key" metaphor in 1894. The interaction between theory and observation was not always harmonious, in hindsight. For example, *Bacillus icteroides* was proposed in 1896 (by Giuseppe Sanarelli, 1864–1940) as a bacterial cause of yellow fever which fulfilled Koch's postulates. Similarly, based on the prevalent germ theory of the early twentieth century, investigators' finding of the bacterium *Haemophilus influenzae* in influenza patients was a perfectly reasonable and suitable answer for the cause of influenza. Only a few decades later was a virus identified as the cause through the work of Richard Shope (1901–1966) and colleagues (Van Epps 2006). In fact, Oswald Avery's (1877–1955) work on DNA was a result of his decades-long work on influenza and pneumonia in the same period [for an in-depth discussion, see Barry (2005)]. These examples show approaches that started with a philosophical theory followed by selective experimentation, leading eventually to a more refined theory. Nevertheless, there were instances where this model did not apply. For example, in 1909, Paul Ehrlich (1854–1915) and his collaborators tested 900 chemical compounds for a syphilis treatment to eventually identify Salvarsan, which we could venture to call one of the earliest precursors (since the heydays of alchemy) of today's high-throughput compound screens. This approach relied more on trial-and-error than *ab initio* theorizing.

The first half of the twentieth century saw the formalized emergence of theoretical and experimental branches in physics, a division that might not have seemed necessary beforehand, but one which continues to the present. Biology saw similar developments. In the 1930s, after the work of investigators such as J. B. S. Haldane (1892–1964), Ronald Fisher (1890–1962) and Sewall Wright (1889–1988) had established population genetics and the modern synthesis in evolutionary biology, C. H. Waddington (1905–1975) and Ludwig von Bertalanffy (1901–1972), among others, proceeded toward formalizing "theoretical biology" and "systems biology," respectively [see, e.g., Moya (2015)]. In 1968, the philosopher Marjorie Grene published "Approaches to a Philosophical Biology" on the state and outlook of the philosophy of biology and, over the next several decades, elements of philosophical approaches to biology were further extended into medical humanities (e.g., in Hans Jonas's 1966 publication on phenomenology and bioethics in "The

Phenomenon of Life: Toward a Philosophical Biology"), philosophical psychology, philosophy of chemistry, philosophical chemistry (tracing its roots at least back to the work of Joseph Black, 1728–1799), physical oncology (Frieboes et al. 2011), healthcare improvement theory (Davidoff et al. 2015) and other related disciplines. Nevertheless, with the advent of recombinant DNA technology in the 1970s followed by molecular biology and the widespread adoption of relevant technologies such as flow cytometry (Robinson and Roederer 2015), theoretical/philosophical biology did not have an opportunity to reach the same level of attention as its counterpart in physics, and today, as noted earlier, many consider theoretical biology to be synonymous with computational and mathematical biology. This is not to say that questions pertaining to theoretical biology have been forgotten. Some of these questions have indeed been rigorously pursued under the domain of philosophy of science/biology, focusing on problems in evolutionary theory or population genetics, amongst others [see, e.g., Gare (2008) and Gouvêa (2015)].

Among the developments in philosophical biology in the first half of the twentieth century, there is perhaps a singular work which stands out given the scope of our analysis: Joseph Henry Woodger's (1894–1981) *Biological Principles* (1929), which in many respects reads farsighted from today's vantage point [see also Nicholson and Gawne (2014)]. Woodger writes in the preface that he takes inspiration from works of the philosophers C. D. Broad (1887–1971) and Alfred North Whitehead (1861–1947). He shines a spotlight, for example, on Whitehead's statement in his *Science and the Modern World* (1925) that "the progress of biology and psychology has probably been checked by the uncritical assumption of half-truths. If science is not to degenerate into a medley of *ad hoc* hypotheses, it must become philosophical and must enter into a thorough criticism of its own foundations." Although Woodger does not use the phrase "philosophical biology," he refers to "theoretical biology" more than 20 times, and writes: "Only two types of theoretical biology have so far been devised, both involving using the analogy of a humanly constructed machine: (1) vitalism (with a mechanic), and (2) the 'machine theory' (without a mechanic). This provides no independent biological way of thinking, because machines presuppose organisms" (Woodger 1929, p. 441). Again on mechanisms, and responding to Haldane, he remarked that "it is always possible to defend microscopic mechanism *in principle*, if any one wishes to do so, by making your mechanism complicated enough, and by 'postulating' enough sub-mechanisms to meet all contingencies. It *cannot* then be refuted, but neither can it be verified. All I have undertaken to do is to show the undesirability of *restricting* biological thought in this way" (Woodger 1929, p. 485). This is germane to our earlier focus on mechanisms. Lastly, he observed the following of the state of biology at the time: "biology [. . .] is still in the metaphysical stage: too eager to press on to startling 'conclusions' rather than to devote critical attention to the purification of its concepts and making more sure of its foundations. The accumulation of data, still less the erection of speculative theories, is not enough" (Woodger 1929, p. 84).

5 The Imperative for a Coherent and Unified Theoretical and Philosophical Biology

Based on the historical precedents expounded thus far, and the current state of biomedical research outcomes, it appears vital to renew the application of philosophical reasoning to theoretical biology research. Philosophical biology would be biology through philosophy, with the aim of gaining insights into foundational questions in biology using a philosophical approach. Its objectives would, in essence, be similar to the ideals of the physical sciences community in the early period of theoretical physics in the 1920s and 1930s. In fact, Max Born (1882–1970) commented in 1963 that "I am now convinced that theoretical physics is actually philosophy." Nevertheless, although the goals would be similar, it is evident that biology and physics are dissimilar in many ways and not necessarily reducible to each other. As Robert Berwick and Noam Chomsky point out, "biology is more like case law, not Newtonian physics" (Berwick and Chomsky 2015, p. 36).

5.1 Contours of a Revived Philosophical Biology

A primary concern of philosophical biology would be the development of models in biology. Baruch Spinoza (1632–1677) pointed out a common fallacy regarding models in his magnum opus *Ethics* (1677): "for men are wont to form general ideas both of natural phenomena and of artifacts, and these ideas they regard as models, and they believe that Nature [. . .] looks to these ideas and holds them before herself as models. So when they see something occurring in Nature at variance with their preconceived ideal of the thing in question, they believe that Nature has then failed or blundered and has left that thing imperfect" (Spinoza 1992). In other words, testing a model *against* a natural phenomenon is different than testing the said natural phenomenon *against* the said model, an issue which is as relevant today as it was more than 300 years ago, and harks back to the historical discussion post-Newton about intelligible theories about the world—a world which is not directly intelligible. Jeremy Gunawardena defines a model as "some form of symbolic representation of our assumptions about reality" (Gunawardena 2014a), with "assumptions" here being a key term. Gunawardena further describes the duality between informal models (mental, verbal, etc.) and formal models (mathematical) in biology. Now, *whereas assumptions about reality are tested daily in the laboratory, and formal models are developed in computational/mathematical biology, informal models, as a bridge between our assumptions about reality and symbolic representations of those assumptions, are ripe for philosophical investigation*.[20] A second primary concern of philosophical biology would be on studying the "limits" of our current understanding in biology. What is accessible to us today and what is not? What can we reasonably expect to find and understand

[20] See also Orzack (2012).

about the cell, given that not finding something does not indicate its nonexistence? Again, drawing on the common themes with theoretical physics, the following observation from the physicist Jean Baptiste Perrin (1870–1942) in his 1926 Nobel Lecture is pertinent:

> Certain scholars considered that since the appearances on our scale were finally the only important ones for us, there was no point in seeking what might exist in an inaccessible domain. I find it very difficult to understand this point of view since what is inaccessible today may become accessible tomorrow (as has happened by the invention of the microscope), and also because coherent assumptions on what is still invisible may increase our understanding of the visible. (Perrin 1926)

On the whole, *philosophical biology would entail an analysis that endeavors to be historically, biologically, and philosophically informed*. It would also rely upon all three types of knowledge among which Spinoza drew a distinction: (1) "knowledge of the first kind" being *opinion* or *imagination*, (2) "knowledge of the second kind" being *reason*, and (3) *intuition* forming "knowledge of the third kind" (Spinoza 1992). Whether or not such an approach leads to a new or modified theory of some facet of biology, it would in any case result in greater understanding of biological phenomena. The distinction here is crucial. Understanding is a process toward the intelligibility of a concept or theory; but why or how that concept becomes intelligible to us, be it through internal analogies, comparisons with similar past experiences, etc., might be beyond the reach of our introspection. To put it differently, nature is full of surprises; to increase understanding of some natural phenomenon is to experience fewer surprises, while certainly always allowing for new puzzlements to arise. Along the same line, the philosopher and psychologist Wilhelm Dilthey (1833–1911) noted in 1894 that "we explain nature, [whereas] we understand psychic life" (Dilthey 1977), drawing a distinction between explanation and understanding.

It might be useful to suggest a few specific examples of questions and topics that a philosophical biologist could consider. These examples are divided into two categories: Theoretical Methods and Tools (TMT) and Theoretical Problems and Solutions (TPS) (Fig. 1).

5.2 Theoretical Methods and Tools (TMT)

When the phrase "philosophical methods" is mentioned, the logical operations of *deduction* and *induction* are usually one of the first to be invoked. To illustrate their utility using the case of the cellular basis of time as an example (Ehsani 2012b), we could say the following: (A) If (1) *all biological reactions in cellular environments are synchronous* and (2) *all synchronous phenomena need an internal or external pacemaker*, then we can syllogistically deduce that (3) *all biological reactions in cellular environments are definitely driven by a pacemaker*. (B) If (1) *cell line X has a pacemaker* and (2) *cell line Y also has a pacemaker*, then we can induce that (3) *other cells may also have a pacemaker*. These two methods are very useful and are in fact indispensable components of human rational reasoning in general.

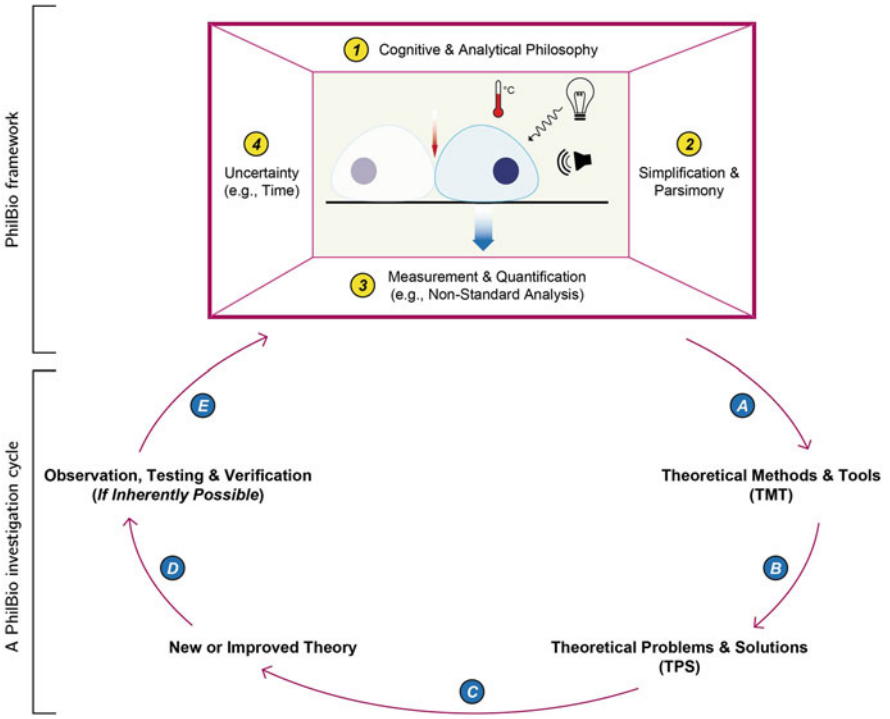

Fig. 1 A proposed outline for philosophical biology investigations. Philosophical biology ("PhilBio") can be framed as a set of perspectives to approach what is known and unknown about a given topic in biology. These perspectives could be cognizant of (**1**) analytical, cognitive, and rational philosophical reasoning; (**2**) a general goal toward simplification and parsimony; (**3**) novel mathematical, logical or other means of measurement; and (**4**) a general appreciation of uncertainty around the interface between our cognitive capacity and different hard facets of nature. An investigation that takes philosophical biology into account can (**A**) use these perspectives and choose one or more philosophical tools to (**B**) approach the problem at hand, using those tools to refine, redefine, or even dismiss the initial question. If the question is not dismissed, (**C**) a set of possible solutions could be proposed. The set of possible solutions could eventually be amalgamated into a new explanatory theory or account, which (**D**) may or may not be verifiable based on the current experimental paradigms of the period. (**E**) This process is repeated as more is understood about the said topic (Source: author)

Nevertheless, in proposing a set of methods and tools for philosophical biology, we could develop approaches that are more tailored to the kinds of questions that are investigated [see, e.g., Nesse (2013)]. The TMT category could include approaches that are analytical, following the works of philosophers such as Charles Sanders Peirce (1839–1914), Gottlob Frege (1848–1925) and Alfred Tarski (1901–1983), or could follow nonanalytical and nontraditional reasoning methods. If a philosophical biology investigation foresees a direct or an immediate human impact, the philosophical approach should be grounded in moral philosophy as a first step. Furthermore, as much as possible, one could aim to initially avoid using

philosophical methods that provoke competing or nontrivial definitions (e.g., mereological, teleological, epistemological, tautological, phenomenological, ontological, normative, etc.) and to appeal to as-simple-as-possible rational and evidentiary approaches. Nevertheless, certain "simplified" components of the concepts in the former category should necessarily be used. Some examples include:

1. When we attempt to understand and describe the behavior of a protein or lipid membrane in a cell, how do we begin to offer a "good" explanation? Is a molecular "descriptive" account an intelligible "explanation" nonetheless? Here contemporary analytical philosophical methods that have been developed at least beginning with the work of Rudolf Carnap (1891–1970) with regards to "explication" can have utility (Friedman 1974; Weber et al. 2013).
2. In offering a descriptive or causal explanation, how do we move beyond providing a statistical view of the phenomenon at hand (Lewontin 2006)? Given current trends in biology, and the natural sciences in general, toward the expansion of numerical models and big-data science (Bauer et al. 2015), this question becomes especially important, as no natural process can have a "statistical nature"; a natural process just has a *nature*, which we may choose to model statistically in the absence of a suitable explanatory theory. In fact, some historians of science rightly point out the fact that "big data" is not a new notion in the sciences, as large collections of data have been a staple of astronomy, in the form of astronomical tables, for many centuries (Mozaffari 2016; Toomer 1968).
3. Because many biological interactions happen at infinitesimal scales where exact measurements give way to approximations, can "non-standard analysis" and the theory of infinitesimals (developed by Abraham Robinson, 1918–1974; published in 1966) along with hyperreal numbers (Robert 2011) be used instead of standard calculus? Can this be combined with Gödel numbering, mereology and set theory?
4. In trying to establish causal relationships in gene/protein circuits, how can we use deontic logic (which focuses on the notion of "obligation") (McNamara 2019)? Would deontic logical approaches help with questions such as *does something that looks like an effect really need a cause*? What role could nonclassical logic play? Here one should note that although deontic and nonclassical logical systems are themselves divided into various subcomponents, the applicability to philosophical biology would not necessarily be in the closed formal proofs that these systems allow, but more in the processes and connections that they can hint at or disprove.
5. What *biological* principles could be envisaged to augment mechanistic accounts in biology and medicine (Ehsani 2019)? An example here could be Xue-Xin Wei and colleagues' report on using mathematical modeling to propose that a "principle of economy predicts the functional architecture of grid cells" (Wei et al. 2015). Moreover, how can the parameters for such principles be discovered? And, how can analogous developments in the field of biolinguistics since the 1950s (Chomsky 2007) be translated to biological investigations?

5.3 Theoretical Problems and Solutions (TPS)

The TPS category comprises a group of newly formed questions and possible sets of solutions that could be proposed using philosophical approaches. Moreover, the goal may at times be to come up with new or improved explanatory theories. A "good" theory is an adaptable theory, one that would allow for incremental growth in understanding while also hinting at gaps that a new and improved theory could bridge (Brigandt 2016), and, in a sense, produce a leap in understanding. Furthermore, such a theory should be developed in "abstraction from the full complexity" of what is being studied (Chomsky 1986; Martin 1980). TPS also includes existing questions or paradigms that are expanded or revised. Some examples include:

1. What is the structural difference between two helical or beta-sheet domains of equal length in two proteins arising from different amino acid sequences? Is a "disordered" domain of a protein really "disordered," or do such domains adopt a limited set of structures that are "appropriate to, but not caused by"[21] the protein/lipid microenvironment around the protein? Can new protein folding theories become alternatives to molecular dynamics simulations (Chung et al. 2015; Robustelli et al. 2018)? Is it conceivable that in some circumstances protein folding, rather than proceeding to minimize free energy (Neupane et al. 2016), proceeds primarily to minimize search efficiency only (i.e., "computational" efficiency from the perspective of the amino acid sequence)? These questions are not only important to understanding the structure and dynamics of proteins but are also indispensable in deriving new theories to account for protein aggregation in neurodegenerative diseases or prion-like propagation of proteins. Furthermore, these questions all have unresolved theoretical underpinnings that can be resolved piece by piece using philosophy. For example, if one tries to conceptualize the number of possible atom-to-atom "interactions" (a vague and problematic notion that needs resolution itself) as a nascent polypeptide chain emerges from the ribosome, the number of possibilities can easily escape finite bounds, whereas it is evident that protein folding takes place in a finite amount of time in the cell (or even under artificial conditions). An "infinite" number of possibilities resolving in a finite amount of time is reminiscent of a "supertask" in mechanical philosophy (Manchak and Roberts 2016), a concept that has been explored since the Antiquity.
2. As noted previously, given the many ambiguities about the structure of water molecules (Thamer et al. 2015; Yang et al. 2019), how do hydrophilic residues of proteins really interact with water molecules in their vicinity? Are such interactions always electrostatic in nature, or could non-electrostatic interactions, such as hydrogen-hydrogen (H–H) bonding—distinct from electrostatic hydrogen bonding (Matta 2006)—also play an important role? It should be

[21] Using the Cartesian terminology referred to by Chomsky in Chomsky (2009).

noted that water is by no means the only "simple" ubiquitous molecule for which deep ambiguities remain. The C–H bonds of the seemingly simple methane molecule (CH_4) are another case in point (Smith et al. 2016), a strand of investigation which could have implications for C–H bonds in amino acids. Moreover, given the essential interaction of many proteins with metal ions, organometallic compounds and other small molecules, do current theories satisfactorily account for the unique interaction of amino acids and these compounds?[22] It should be noted that these questions are essential *primary questions* not only to understand protein folding, but also protein interactions with other proteins and macromolecules. Before finding suitable answers to these questions, it is doubtful that an explanatory framework can ever reach the point where broader concepts, such as cross-species "inter-interactomes" of protein–protein interaction networks (Zhong et al. 2016), could be addressed.

3. What is the concept of time in the cell (cellular time as opposed to circadian time) (Ehsani 2012b)? Does a cell need a sub-second timekeeping or pacemaking mechanism to arrange the plethora of simultaneous functions taking place in the cytoplasm and other subcellular compartments? If so, what could such a mechanism be (Fig. 2)?
4. Can one deduce whether the time frame of a given cellular process increases polynomially with the increasing "complexity" of the task? This problem could borrow from extensive research in theoretical computer science under the famous *P* (polynomial time) versus *NP* (nondeterministic polynomial time) paradigm. Such questions can initially draw from existing research in computational biology regarding the protein folding problem (Berger and Leighton 1998) or RNA structure prediction (Smit et al. 2008), to name a few. One can eventually expand the scope of these investigations to include kinetic studies of enzymes and process timing in the framework of the Michaelis–Menten equation (Gunawardena 2014b; Xie 2013).
5. What exactly is aging on an organism level, and what accounts for the diversity of aging profiles across species (Jones et al. 2014; McCormick et al. 2015; Moger-Reischer and Lennon 2019) and different phyla (Hug et al. 2016; Levin et al. 2016)? As evidenced by work on the *Prochlorococcus* genus (Biller et al. 2015), this line of investigation would require an analysis of what it really means to be a species, a genus, etc. (Buchanan 2015; Nei and Nozawa 2011).
6. For the phenomenon of antimicrobial resistance, as alluded to earlier, can one devise a strategy where the emergence of resistance would theoretically be impossible? Since the introduction of sulfonamide antibacterial drugs in the 1930s, the emergence of resistant subpopulations of bacteria or fungi has become inevitable (Balaban et al. 2019; Fisher and Mobashery 2016). Nevertheless, notwithstanding the re-emerging bacteriophage research field (Rohwer and Segall 2015), strategies are beginning to be refurbished or de

[22] As an example, see Seeman and Cantrill (2016) for a discussion of the challenges in understanding the structure of the organometallic compound ferrocene.

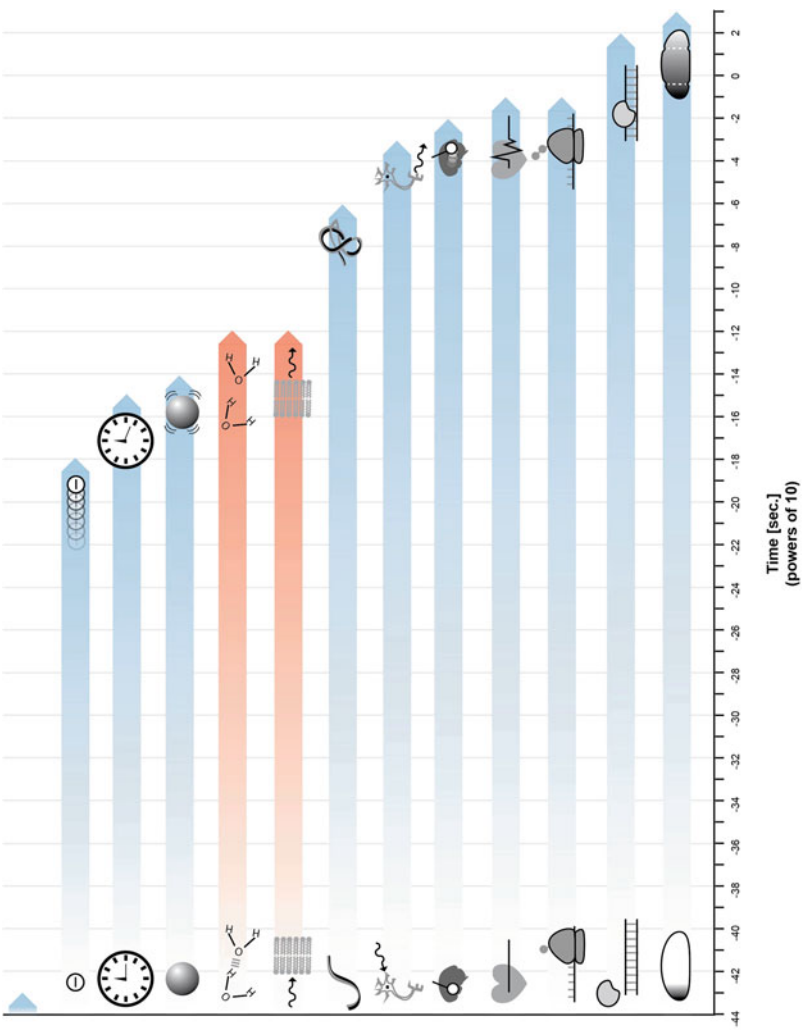

Fig. 2 Timescale of select cellular processes. A logarithmic depiction of essential "chemical" and "biological" cellular tasks, based on published results, demonstrates a picosecond to second time range. If the goal of a research question is to posit a cellular timekeeping mechanism, then the proposed mechanism should at least be as fast as the fastest *biological* process, if not faster. Energy transfer across phospholipid membrane leaflets appears to be a common denominator of slower biological functions (Source: author)

novo devised where compound-based antimicrobial resistance would become avoidable (Baym et al. 2016; Kolter and van Wezel 2016; Lazar et al. 2013; Szybalski and Bryson 1952; Toprak et al. 2012). This is a critical and ripe area for practical philosophical contributions.

7. In cancer biology, what are the theoretical underpinnings of the occasionally paradoxical nature of metastasis (Piskounova et al. 2015), heterogeneous origins (Ling et al. 2015), differing outcomes (Nikiforov et al. 2016), and spontaneous regression (Brodeur and Bagatell 2014; Diede 2014; Hoption Cann et al. 2002; Saade Lemus et al. 2019)? How can some of these paradoxes be used as "natural experiments" (Feder and Mitchell-Olds 2003; Louis 2007; Sekikawa et al. 2003) in cancer research? Since the postulation of the Warburg effect more than 90 years ago, our understanding of cancer metabolism, and oncology in general, has greatly advanced. Nevertheless, an all-encompassing theory is still lacking. For example, in reviewing a metabolomic analysis of cancer cell proliferation (Hosios et al. 2016), Jason Tanner and Jared Rutter pose the following questions that await explanation:

> Although the majority of cell protein is comprised of amino acids imported from the environment, why do cultured cells—awash in amino-acid-rich culture medium—utilize glutamine to synthesize other amino acids de novo even when those amino acids are available for import? Might this have to do with limited import capacity, or is there a separate unforeseen advantage to biosynthesis? Finally, the finding that glucose-derived carbon contributes to a small fraction of cell mass raises still more questions. Why don't proliferating cells utilize the large amounts of carbon consumed as glucose to meet their biosynthetic needs? What is the purpose of such a carbon-wasting metabolic program? Is it simply that this program enables the rapid production of adequate ATP while maintaining the NAD/NADH redox balance, or is there more to it?" (Tanner and Rutter 2016)

   One could speculate that these and other questions will only yield to new experimental studies if a novel explanatory theory is provided to frame and structure the plethora of pieces of knowledge already available. In fact, certain areas within the cancer research field (e.g., cancer stem cells or immunotherapy) have already benefited noticeably from philosophical and theoretical approaches (Laplane 2016; Ledford 2016; Willyard 2016; Wong and Slavcev 2015).

8. More broadly, what are the inherent differences between correlation and causation (Ehsani 2013a; Karmon and Pilpel 2016)? Do they have structural differences? For what questions might causal thinking and the notion of agency not be necessary? In areas such as neural networks, to paraphrase the pre-Socratic philosopher Heraclitus of Ephesus (ca. 535–475 BC), "is a hidden connection better/stronger than a visible connection"? Could a hidden or unobservable variable in an experiment, to borrow from economic theory, "share covariance properties with the observed variables" (Oster 2015)?[23] Additionally, similar to the notion mentioned earlier for disordered domains in

[23] See Elf (2016), Hilfinger et al. (2016), Justman (2016) for further discussion related to this topic.

proteins, are there biological processes that are "appropriate to but not caused by" (Chomsky 2009) the stimuli that are currently thought to be the causes of those processes? These questions are fundamental to all areas of biology, from investigating the still-unraveling workings of organelles (Delling et al. 2016; Norris and Jackson 2016) to metabolic processes (Scholl and Nickelsen 2015), cell-death pathways (Wallach et al. 2016), designing protocols that allow a smoother transition of findings from model organisms to humans (Anders et al. 2016), and human speech fluency (Lieshout et al. 2014), to name a few. It is also evident that a more thorough understanding of causative structures has direct applicability to research on disease mechanisms regardless of the degree to which the exact etiology is known [e.g., Mendelian (Chen et al. 2016; Steffan 2016) or infectious diseases (Byrd and Segre 2016)].

9. Do cellular processes that seem chaotic, stochastic or random (Capp 2012; Kadelka et al. 2013; Losick and Desplan 2008; Uphoff et al. 2016), such as bursts of transcription or Brownian-like motion of different macromolecules in the cytoplasm, in fact follow as-yet unrecognized deterministic pathways (Ehsani 2012a; Kryazhimskiy et al. 2014)?
10. What exactly is "uncertainty," and is it possible to postulate theories that go beyond a statistical description of uncertainty? For example, the American Statistical Association emphasized in 2016 that a *P*-value is "a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself" (Wasserstein and Lazar 2016). In addition, are there concepts that "ought to be true" but that we cannot describe or observe with any certainty? Are there aspects of biological cells which one can never in principle be certain about? In other words, are there limits to our understanding in certain areas of biology [e.g., see Lewontin (1998)]? Some of the "low-hanging fruits" amongst these questions may initially be found in computational biology. For example, there have been efforts to identify inherent upper limits in accelerating search speeds in biological datasets (Kannan and Tse 2015). Moreover, notions of "loose and tight limits" have been defined for computational problems (Markov 2014). Furthermore, similar questions can be asked in chemistry, as for example noted by the biochemist Christopher Walsh: "how much new chemistry is yet to be found [and] what kinds of biosynthetic enzymatic transformations are yet to be characterized?" (Walsh 2015).
11. The field of neuroscience is readily conducive to philosophical and theoretical inquiries [see, e.g., Casebeer (2003), Fleming (2016), Greene (2003), Gregory (2000)]. However, in light of the numerous unsolved, lower-hanging-fruit problems in "simpler" organisms such as *D. melanogaster* or *C. elegans*, many questions regarding human cognition, the primate nervous system or the mouse brain (the circuitry of which is beginning to be mapped) may remain outside the purview of philosophical biology for some time to come. Nevertheless, there are questions that could be further refined in human cognitive science and neurobiology using philosophical biology. For example, why are "our brains [. . .] preprogrammed to misread certain images" (Chatterjee 2015) or what is the initiating mechanism of voluntary movements (Bizzi and Ajemian 2015)?

12. Lastly, certain philosophy of biology threads could be investigated from a philosophical biology perspective with immediate application to both strands. The philosopher of science Paul Griffiths, for instance, writes that "when addressing [conceptual puzzles within biology], there is no clear distinction between philosophy of biology and theoretical biology" (Griffiths 2018). As a case in point, if chemistry is arguably not reducible to physics (yet unifiable at the same time) (Chomsky 2009), in the same spirit could we ask if the properties of a biological system (e.g., a cell) can ever be reduced to the properties of its component parts (e.g., proteins)? As another example, if one excludes some obvious explanations, why are certain findings either in molecular biology or clinical medicine not reproducible (Niepel et al. 2019; Open Science 2015)? Approaches to tackle these and other questions of this kind are well established in the philosophy of science literature, and therefore philosophical biology and philosophy of science are not exclusive of each other.

5.4 Inherent and Experimental Verifiability

Given the various lines of study suggested above in the TMT and TPS sections, what would be an endpoint to, or natural progression of, a philosophical biology investigation? Is empirical verification a necessary touchstone? Although certain outcomes of philosophical biology studies can and should be tested computationally or in a cellular biology laboratory, one can posit that experimental validation should not be the ultimate standard to validate or invalidate such strands of investigation. Again to draw an analogy with theoretical physics, early quantum physicists realized that some theoretical paradigms will, at least for the foreseeable future, remain outside the purview of experimental falsifiability in light of the inherent limits within physical experimental approaches [see, e.g., Deutsch (2015)]. The Laser Interferometer Gravitational-Wave Observatory (LIGO) experimental results act as a case in point (Abbott et al. 2016). The nonobligatory interaction between theory and experiment is not limited to physics but can also be observed in, for example, economics (Abreu et al. 2012) or biolinguistics (Hauser 2016). It is therefore to be anticipated that philosophical biology would also generate hypotheses or questions that might not be testable in the laboratory *immediately*, but their value would only be shown in time and in perhaps not-so-predictable manners.

Theories developed using the framework described here are certainly not endpoints in a given theoretical investigation. They are merely stepping stones toward more complete frameworks and programs, which could generate questions that are inherently empirical. A relevant example here is the culmination of many biological theories and empirical validation attempts that now form the exhaustive set of transcriptional and translational programs in developmental biology (Oates et al. 2009). This process can also be observed in biolinguistics, whereby linguistic theories that led to our current state of understanding of Universal Grammar were included in a Minimalist Program, which was then followed amongst other things by the development of the Strong Minimalist Thesis (SMT). SMT is now a robust

explanatory framework that can allow linguists to discover the extent to which one can "account for the relevant phenomena of language" (Chomsky 2014a).

6 Conclusions

For the modern biological disciplines to produce genuine instances of understanding of the workings of the cell, philosophy must regain its rightful place in the theoretical foundations of biology. This is in line with the development of the natural sciences at least since the Galilean revolution in science and the Enlightenment. The overarching aim of a philosophical biology program could be to define suitably innovative and worthwhile horizons for individual parts of biomedical research, horizons that are not mere pedantic extrapolations of current technical information. Furthermore, solutions that arise from these investigations may be isomorphic, such that their theoretical structure could be applicable to other areas of the sciences, in the same line that a number of frameworks from modern linguistics have been utilized in this chapter in the context of philosophical biology.

It may be apt to end on a note that the complexities and "deep truths" of cellular processes could remain hidden even in spite of philosophical, theoretical and other sincere efforts. In the words of Attar of Nishapur (ca. 1145–1221), "the sea will be the sea, whatever the drop's philosophy." Nevertheless, one can at least be assured that a philosophical approach to biology will constantly question our questions and provide a framework for reassessing and improving our perspectives of the workings of the cell in normal and disease biology.

Acknowledgments Earlier developments of some of the ideas presented in this work have appeared in two preprint articles (Ehsani 2016, 2019). I would like to thank Philipp Plugmann for the invitation to contribute to this collection. The author's work is supported by a University College London Overseas Research Scholarship.

References

Abbot, F. E. (1868). Philosophical biology. *The North American Review, 107*, 377–422.

Abbott, B. P., Abbott, R., Abbott, T. D., Abernathy, M. R., Acernese, F., Ackley, K., et al. (2016). Observation of gravitational waves from a binary black hole merger. *Physical Review Letters, 116*(6), 061102.

Abreu, D., Pearce, D., & Stacchetti, E. (2012). One-sided uncertainty and delay in reputational bargaining. *Economic Theory Center Working Paper No. 45-2012*.

Aisen, P. (2019). *Comment on 'End of the BACE inhibitors? Elenbecestat trials halted amid safety concerns' (Alzforum)*. Retrieved September 13, 2019, from http://alzforum.org/news/research-news/end-bace-inhibitors-elenbecestat-trials-halted-amid-safety-concerns#comment-32966.

Allen, J. E., & Sutherland, T. E. (2019). Crystal-clear treatment for allergic disease. *Science, 364*(6442), 738–739.

Anders, H. J., Jayne, D. R., & Rovin, B. H. (2016). Hurdles to the introduction of new therapies for immune-mediated kidney diseases. *Nature Reviews Nephrology, 12*(4), 205–216.

Aristotle. (1882). *De Partibus Animalium (On the parts of animals)* (W. Ogle, Trans.). Kegan Paul, Trench & Co.

Aristotle. (1910). *De Generatione Animalium (On the generation of animals)* (A. Platt, Trans.). Clarendon Press.
Aristotle. (1930). *Physica (Physics)* (R. P. Hardie & R. K. Gaye, Trans.). Clarendon Press.
Aristotle. (1931). *De Anima (On the soul)* (J. A. Smith, Trans.). Clarendon Press.
Armiento, A., Doumic, M., Moireau, P., & Rezaei, H. (2016). Estimation from moments measurements for amyloid depolymerisation. *Journal of Theoretical Biology, 397*, 68–88.
Arney, K. (2018). Solving lymphoma's stem-cell problem. *Nature, 563*(7731), S48–S49.
Asatryan, A. D., & Komarova, N. L. (2016). Evolution of genetic instability in heterogeneous tumors. *Journal of Theoretical Biology, 396*, 1–12.
Bagot, C. N., & Arya, R. (2008). Virchow and his triad: A question of attribution. *British Journal of Haematology, 143*(2), 180–190.
Balaban, N. Q., Helaine, S., Lewis, K., Ackermann, M., Aldridge, B., Andersson, D. I., et al. (2019). Definitions and guidelines for research on antibiotic persistence. *Nature Reviews Microbiology, 17*(7), 441–448.
Barnett, J. A., & Barnett, L. (2011). *Yeast research: A historical overview*. Washington, DC: ASM Press.
Barry, J. M. (2005). *The great influenza: The story of the deadliest pandemic in history*. London: Penguin.
Bartley, S. (1999). John Hunter – the scientific surgeon 'don't think, try; be patient, be accurate...'. *Journal of Investigative Surgery, 12*(6), 305–306.
Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature, 525*(7567), 47–55.
Baym, M., Stone, L. K., & Kishony, R. (2016). Multidrug evolutionary strategies to reverse antibiotic resistance. *Science, 351*(6268), aad3292.
Berger, B., & Leighton, T. (1998). Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Journal of Computational Biology, 5*(1), 27–40.
Bertsch, M., Franchi, B., Marcello, N., Tesi, M. C., & Tosin, A. (2017). Alzheimer's disease: A mathematical model for onset and progression. *Mathematical Medicine and Biology, 34*(2), 193–214.
Berwick, R. C., & Chomsky, N. (2015). *Why only us: Language and evolution*. MIT Press.
Berwick, R. C., & Chomsky, N. (2017). Why only us: Recent questions and answers. *Journal of Neurolinguistics, 43*(B), 166–177.
Biller, S. J., Berube, P. M., Lindell, D., & Chisholm, S. W. (2015). Prochlorococcus: The structure and function of collective diversity. *Nature Reviews Microbiology, 13*(1), 13–27.
Bizzi, E., & Ajemian, R. (2015). A hard scientific quest: Understanding voluntary movements. *Daedalus, 144*(1), 83–95.
Bodner, G. M. (1986). Constructivism: A theory of knowledge. *Journal of Chemical Education, 63*(10), 873–878.
Bognon-Küss, C., & Wolfe, C. T. (2019). The idea of 'philosophy of biology before biology': A methodological provocation. In C. Bognon-Küss & C. T. Wolfe (Eds.), *Philosophy of biology before biology* (pp. 4–23). Routledge.
Braak, H., Del Tredici-Braak, K., & Gasser, T. (2018). Special issue "Parkinson's disease". *Cell and Tissue Research, 373*(1), 1–7.
Brigandt, I. (2016). Do we need a 'theory' of development? *Biology and Philosophy, 31*, 603–617.
Britten, R. J., & Davidson, E. H. (1969). Gene regulation for higher cells: A theory. *Science, 165*(3891), 349–357.
Brodeur, G. M., & Bagatell, R. (2014). Mechanisms of neuroblastoma regression. *Nature Reviews Clinical Oncology, 11*(12), 704–713.
Buchanan, M. (2015). Bacterial complexity. *Nature Physics, 11*, 887.
Byrd, A. L., & Segre, J. A. (2016). Infectious disease. Adapting Koch's postulates. *Science, 351*(6270), 224–226.
Capp, J. P. (2012). Stochastic gene expression stabilization as a new therapeutic strategy for cancer. *BioEssays, 34*(3), 170–173.

Cardwell, J. C. (1905). The development of animal physiology: The physiology of Aristotle. *Medical Library and Historical Journal, 3*(1), 50–77.
Carey, T. A., & Stiles, W. B. (2016). Some problems with randomized controlled trials and some viable alternatives. *Clinical Psychology and Psychotherapy, 23*(1), 87–95.
Casebeer, W. D. (2003). Moral cognition and its neural constituents. *Nature Reviews Neuroscience, 4*(10), 840–846.
Chatterjee, R. (2015). Out of the darkness. *Science, 350*(6259), 372–375.
Chen, R., Shi, L., Hakenberg, J., Naughton, B., Sklar, P., Zhang, J., et al. (2016). Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nature Biotechnology, 34*(5), 531–538.
Chomsky, N. (1983). *Things no amount of learning can teach* (interview by J. Gliedman), from chomsky.info/198311__/
Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger Publishers.
Chomsky, N. (2007). Biolinguistic explorations: Design, development, evolution. *International Journal of Philosophical Studies, 15*(1), 1–21.
Chomsky, N. (2009). The mysteries of nature: How deeply hidden. *The Journal of Philosophy, 106*(4), 167–200.
Chomsky, N. (2014a). Preface to the 20th anniversary edition. *The Minimalist Program (20th Anniversary Edition)* (pp. vii–xiii). MIT Press.
Chomsky, N. (2014b). *Science, mind, and limits of understanding*. The Science and Faith Foundation, Pontifical Council for Culture.
Christie, R. V. (1987). Galen on Erasistratus. *Perspectives in Biology and Medicine, 30*(3), 440–449.
Chung, H. S., Piana-Agostinetti, S., Shaw, D. E., & Eaton, W. A. (2015). Structural origin of slow diffusion in protein folding. *Science, 349*(6255), 1504–1510.
Clary, D. C. (2016). Quantum dynamics in the smallest water droplet. *Science, 351*(6279), 1267–1268.
Cohen, I. B., & Smith, G. E. (2004). Introduction. In I. B. Cohen & G. E. Smith (Eds.), *The Cambridge companion to Newton* (pp. 1–32). Cambridge University Press.
Connell, S. M. (2001). Toward an integrated approach to Aristotle as a biological philosopher. *The Review of Metaphysics, 55*(2), 297–322.
Couzin-Frankel, J. (2019). Beyond survival. *Science, 363*(6432), 1166–1169.
Cukierman, S. (2006). Et tu, Grotthuss! and other unfinished stories. *Biochimica et Biophysica Acta, 1757*(8), 876–885.
Davidoff, F., Dixon-Woods, M., Leviton, L., & Michie, S. (2015). Demystifying theory and its use in improvement. *BMJ Quality and Safety, 24*(3), 228–238.
Del Soldato, E. (2019). Natural philosophy in the Renaissance. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
Delling, M., Indzhykulian, A. A., Liu, X., Li, Y., Xie, T., Corey, D. P., & Clapham, D. E. (2016). Primary cilia are not calcium-responsive mechanosensors. *Nature, 531*(7596), 656–660.
Depew, D. J. (1995). Humans and other political animals in Aristotle's history of animals. *Phronesis, 40*(2), 156–181.
Deutsch, D. (2015). The logic of experimental tests, particularly of Everettian quantum theory. *arXiv*, 1508.02048.
Diede, S. J. (2014). Spontaneous regression of metastatic cancer: Learning from neuroblastoma. *Nature Reviews Cancer, 14*(2), 71–72.
Dilthey, W. (1977). Ideas concerning a descriptive and analytic psychology (1894) *Descriptive psychology and historical understanding (R. M. Zaner & K. L. Heiges, Trans.)* (pp. 21–120). Martinus Nijhoff Publishers.
DiMeglio, L. A., Evans-Molina, C., & Oram, R. A. (2018). Type 1 diabetes. *Lancet, 391*(10138), 2449–2462.
Dong, J., Panchakshari, R. A., Zhang, T., Zhang, Y., Hu, J., Volpi, S. A., et al. (2015). Orientation-specific joining of AID-initiated DNA breaks promotes antibody class switching. *Nature, 525*(7567), 134–139.

Editors. (2016). So long to the silos. *Nature Biotechnology, 34*, 357.
Ehsani, S. (2012a). Simple variation of the logistic map as a model to invoke questions on cellular protein trafficking. *arXiv*, 1206.5557.
Ehsani, S. (2012b). Time in the cell: A plausible role for the plasma membrane. *arXiv*, 1210.0168.
Ehsani, S. (2013a). Correlative-causative structures and the 'pericause': An analysis of causation and a model based on cellular biology. *arXiv*, 1310.0507.
Ehsani, S. (2013b). Macro-trends in research on the central dogma of molecular biology. *arXiv*, 1301.2397v2.
Ehsani, S. (2016). A framework for philosophical biology. *arXiv*, 1605.00033.
Ehsani, S. (2019). The challenges of purely mechanistic models in biology and the minimum need for a 'mechanism-plus-X' framework. *arXiv*, 1905.10916.
Einstein, A. (1922). *Sidelights on relativity* (G. B. Jeffery & W. Perrett, Trans.). Methuen & Co.
Elf, J. (2016). Staying clear of the dragons. *Cell Systems, 2*, 219–220.
Everaert, M. B., Huybregts, M. A., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences, 19*, 729–743.
Feder, M. E., & Mitchell-Olds, T. (2003). Evolutionary and ecological functional genomics. *Nature Reviews Genetics, 4*(8), 651–657.
Fisher, J. F., & Mobashery, S. (2016). Endless resistance. Endless antibiotics? *Medicinal Chemistry Communications, 7*, 37–49.
Fleming, S. M. (2016). Changing our minds about changes of mind. *eLife, 5*, e14790.
Frede, M. (1992). Plato's *Sophist* on false statements. In R. Kraut (Ed.), *The Cambridge companion to Plato* (pp. 397–424). Cambridge University Press.
Frieboes, H. B., Chaplain, M. A., Thompson, A. M., Bearer, E. L., Lowengrub, J. S., & Cristini, V. (2011). Physical oncology: A bench-to-bedside quantitative and predictive approach. *Cancer Research, 71*(2), 298–302.
Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy, 71*(1), 5–19.
Galen. (1916). *On the natural faculties* (A. J. Brock, Trans.). William Heinemann.
Gare, A. (2008). Approaches to the question, 'what is life?': Reconciling theoretical biology with philosophical biology. *Cosmos and History: The Journal of Natural and Social Philosophy, 4*, 53–77.
Garfinkel, A. (2015). *Bad philosophical ideas that are driving modern biology and medicine*. MIT Philosophy Colloquium.
Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., et al. (2010). Actual causation: A stone soup essay. *Synthese, 175*(2), 169–192.
Gordon, S. (2016). Phagocytosis: An immunobiologic process. *Immunity, 44*(3), 463–475.
Gouvêa, D. Y. (2015). Explanation and the evolutionary first law(s). *Philosophy of Science, 82*(3), 363–382.
Greaves, M. (2018). A causal mechanism for childhood acute lymphoblastic leukaemia. *Nature Reviews Cancer, 18*(8), 471–484.
Greene, J. (2003). From neural 'is' to moral 'ought': What are the moral implications of neuroscientific moral psychology? *Nature Reviews Neuroscience, 4*(10), 846–849.
Gregory, R. (2000). Reversing Rorschach. *Nature, 404*(6773), 19.
Griffiths, P. (2018). Philosophy of biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
Gunawardena, J. (2014a). Beware the tail that wags the dog: Informal and formal models in biology. *Molecular Biology of the Cell, 25*(22), 3441–3444.
Gunawardena, J. (2014b). Time-scale separation – Michaelis and Menten's old idea, still bearing fruit. *The FEBS Journal, 281*(2), 473–488.
Hatwalne, Y., Ramaswamy, S., Rao, M., & Simha, R. A. (2004). Rheology of active-particle suspensions. *Physical Review Letters, 92*(11), 118101.
Hauser, M. D. (2016). Challenges to the what, when, and why. *Biolinguistics, 10*, 1–5.

Henry, D. (2018). Aristotle on epigenesis: Two senses of epigenesis. In A. Falcon & D. Lefebvre (Eds.), *Aristotle's generation of animals: A critical guide* (pp. 89–107). Cambridge University Press.

Hilfinger, A., Norman, T. M., & Paulsson, J. (2016). Exploiting natural fluctuations to identify kinetic mechanisms in sparsely characterized systems. *Cell Systems, 2*, 251–259.

Hoption Cann, S. A., van Netten, J. P., van Netten, C., & Glover, D. W. (2002). Spontaneous regression: A hidden treasure buried in time. *Medical Hypotheses, 58*(2), 115–119.

Hosios, A. M., Hecht, V. C., Danai, L. V., Johnson, M. O., Rathmell, J. C., Steinhauser, M. L., et al. (2016). Amino acids rather than glucose account for the majority of cell mass in proliferating mammalian cells. *Developmental Cell, 36*, 540–549.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nature Microbiology, 1*, 16048.

Hume, D. (1902). *Enquiries concerning the human understanding and concerning the principles of morals*. Clarendon Press.

Jaeger, A. M., Stopfer, L., Lee, S., Gaglia, G., Sandel, D., Santagata, S., et al. (2019). Rebalancing protein homeostasis enhances tumor antigen presentation. *Clinical Cancer Research, 25*(21), 6392–6405.

Jones, O. R., Scheuerlein, A., Salguero-Gomez, R., Camarda, C. G., Schaible, R., Casper, B. B., et al. (2014). Diversity of ageing across the tree of life. *Nature, 505*(7482), 169–173.

JRA. (2006). Virchow misquoted, part-quoted, and the real McCoy. *Journal of Epidemiology and Community Health, 60*(8), 671.

Justman, Q. (2016). The power of logic and reason. *Cell Systems, 2*, 215.

Kadelka, C., Murrugarra, D., & Laubenbacher, R. (2013). Stabilizing gene regulatory networks through feedforward loops. *Chaos, 23*(2), 025107.

Kannan, S., & Tse, D. (2015). Fundamental limits of search. *Cell Systems, 1*(2), 102–103.

Karmon, A., & Pilpel, Y. (2016). Biological causal links on physiological and evolutionary time scales. *eLife, 5*, e14424.

Kolter, R., & van Wezel, G. P. (2016). Goodbye to brute force in antibiotic discovery? *Nature Microbiology, 1*, 15020.

Kottke, T. E. (2011). Medicine is a social science in its very bone and marrow. *Mayo Clinic Proceedings, 86*(10), 930–932.

Kramer, M. (2019). Stats: Is this therapy useful? *Nature, 569*(7755), 192.

Kryazhimskiy, S., Rice, D. P., Jerison, E. R., & Desai, M. M. (2014). Microbial evolution. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science, 344*(6191), 1519–1522.

Laplane, L. (2016). *Cancer stem cells: Philosophy and therapies*. Harvard University Press.

Lazar, V., Pal Singh, G., Spohn, R., Nagy, I., Horvath, B., Hrtyan, M., et al. (2013). Bacterial evolution of antibiotic hypersensitivity. *Molecular Systems Biology, 9*, 700.

Ledford, H. (2016). Cocktails for cancer with a measure of immunotherapy. *Nature, 532*(7598), 162–164.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics, 11*(10), 733–739.

Leigh, F. (2007). Platonic dialogue, maieutic method and critical thinking. *Journal of Philosophy of Education, 41*(3), 309–323.

Levin, M., Anavy, L., Cole, A. G., Winter, E., Mostov, N., Khair, S., et al. (2016). The mid-developmental transition and the evolution of animal body plans. *Nature, 531*(7596), 637–641.

Lewis, K. (2012). Antibiotics: Recover the lost art of drug discovery. *Nature, 485*(7399), 439–440.

Lewontin, R. C. (1998). The evolution of cognition: Questions we will never answer. In D. Scarborough, S. Sternberg, & D. N. Osherson (Eds.), *An invitation to cognitive science: Methods, models, and conceptual issues* (Vol. 4, pp. 107–132). MIT Press.

Lewontin, R. C. (2006). Commentary: Statistical analysis or biological analysis as tools for understanding biological causes. *International Journal of Epidemiology, 35*(3), 536–537.

Lieshout, P., Ben-David, B., Lipski, M., & Namasivayam, A. (2014). The impact of threat and cognitive stress on speech motor control in people who stutter. *Journal of Fluency Disorders, 40*, 93–109.

Lin, A., Giuliano, C. J., Palladino, A., John, K. M., Abramowicz, C., Yuan, M. L., et al. (2019). Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science Translational Medicine, 11*(509), eaaw8412.

Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., et al. (2015). Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proceedings of the National Academy of Sciences of the United States of America, 112*(47), E6496–E6505.

Lister, J. (1957). The medical traveler. *The New England Journal of Medicine, 257*, 878–879.

Locke, J. (1894). *An essay concerning humane understanding, Book II*. Clarendon Press.

Lonie, I. M. (1964). Erasistratus, the Erasistrateans, and Aristotle. *Bulletin of the History of Medicine, 38*, 426–443.

Lopez, H. M., Gachelin, J., Douarche, C., Auradou, H., & Clement, E. (2015). Turning bacteria suspensions into superfluids. *Physical Review Letters, 115*(2), 028301.

Losick, R., & Desplan, C. (2008). Stochasticity and cell fate. *Science, 320*(5872), 65–68.

Louis, E. J. (2007). Evolutionary genetics: Making the most of redundancy. *Nature, 449*(7163), 673–674.

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature, 526*(7572), 187–189.

Manchak, J., & Roberts, B. W. (2016). Supertasks. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Metaphysics Research Lab, Stanford University.

Marchetti, M. C. (2015). Soft matter: Frictionless fluids from bacterial teamwork. *Nature, 525*(7567), 37–39.

Markov, I. L. (2014). Limits on fundamental limits to computation. *Nature, 512*(7513), 147–154.

Martin, R. M. (1980). *Primordiality, science, and value*. State University of New York Press.

Matta, C. F. (2006). Hydrogen-hydrogen bonding: The non-electrostatic limit of closed-shell interaction between two hydrogen atoms. A critical review. In S. J. Grabowski (Ed.), *Hydrogen bonding – new insights*: Springer.

McCormick, M. A., Delaney, J. R., Tsuchiya, M., Tsuchiyama, S., Shemorry, A., Sim, S., et al. (2015). A comprehensive analysis of replicative lifespan in 4,698 single-gene deletion strains uncovers conserved mechanisms of aging. *Cell Metabolism, 22*(5), 895–906.

McNamara, P. (2019). Deontic logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

McNeil, J. J., Nelson, M. R., Woods, R. L., Lockery, J. E., Wolfe, R., Reid, C. M., et al. (2018). Effect of aspirin on all-cause mortality in the healthy elderly. *The New England Journal of Medicine, 379*(16), 1519–1528.

Mill, J. S. (1843). *A system of logic, ratiocinative and inductive*. John W. Parker.

Moger-Reischer, R. Z., & Lennon, J. T. (2019). Microbial ageing and longevity. *Nature Reviews Microbiology, 17*(11), 679–690.

Monaco, A. P., Neve, R. L., Colletti-Feener, C., Bertelson, C. J., Kurnit, D. M., & Kunkel, L. M. (1986). Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature, 323*(6089), 646–650.

Moya, A. (2015). General systems theory and systems biology. In *The calculus of life: Towards a theory of life* (pp. 25–30). Springer.

Mozaffari, S. M. (2016). Planetary latitudes in medieval Islamic astronomy: An analysis of the non-Ptolemaic latitude parameter values in the Maragha and Samarqand astronomical traditions. *Archive for History of Exact Sciences, 70*(5), 513–541.

Nakashige, T. G., Zhang, B., Krebs, C., & Nolan, E. M. (2015). Human calprotectin is an iron-sequestering host-defense protein. *Nature Chemical Biology, 11*, 765–771.

Nei, M., & Nozawa, M. (2011). Roles of mutation and selection in speciation: From Hugo de Vries to the modern genomic era. *Genome Biology and Evolution, 3*, 812–829.

Nesse, R. M. (2013). Tinbergen's four questions, organized: A response to Bateson and Laland. *Trends in Ecology and Evolution, 28*(12), 681–682.

Neupane, K., Manuel, A. P., & Woodside, M. T. (2016). Protein folding trajectories can be described quantitatively by one-dimensional diffusion over measured energy landscapes. *Nature Physics, 12*, 700–703.

Nicholson, D. J., & Gawne, R. (2014). Rethinking Woodger's legacy in the philosophy of biology. *Journal of the History of Biology, 47*(2), 243–292.

Niepel, M., Hafner, M., Mills, C. E., Subramanian, K., Williams, E. H., Chung, M., et al. (2019). A multi-center study on the reproducibility of drug-response assays in mammalian cell lines. *Cell Systems, 9*(1), 35–48 e5.

Nikiforov, Y. E., Seethala, R. R., Tallini, G., et al. (2016). Nomenclature revision for encapsulated follicular variant of papillary thyroid carcinoma: A paradigm shift to reduce overtreatment of indolent tumors. *JAMA Oncology, 2*(8), 1023–1029.

Norris, D. P., & Jackson, P. K. (2016). Cell biology: Calcium contradictions in cilia. *Nature, 531*(7596), 582–583.

Oates, A. C., Gorfinkiel, N., Gonzalez-Gaitan, M., & Heisenberg, C. P. (2009). Quantitative approaches in developmental biology. *Nature Reviews Genetics, 10*(8), 517–530.

Odenbaugh, J. (2013). Searching for patterns, hunting for causes - Robert MacArthur, the mathematical naturalist. In O. Harman & M. R. Dietrich (Eds.), *Outsider scientists: Routes to innovation in biology* (pp. 181–198). Chicago: University of Chicago Press.

O'Malley, B. W. (2010). Masters of the genome. *Nature Reviews Molecular Cell Biology, 11*(5), 311.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Orzack, S. H. (2012). The philosophy of modelling or does the philosophy of biology have any use? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1586), 170–180.

Oster, E. (2015). *Unobservable selection and coefficient stability: Theory and evidence*. Brown University.

Perrin, J. B. (1926). *Discontinuous structure of matter*. Nobel Lecture.

Persson, E. K., Verstraete, K., Heyndrickx, I., Gevaert, E., Aegerter, H., Percier, J. M., et al. (2019). Protein crystallization promotes type 2 immunity and is reversible by antibody treatment. *Science, 364*(6442).

Pigliucci, M. (2013). On the different ways of 'doing theory' in biology. *Biological Theory, 7*(4), 287–297.

Piskounova, E., Agathocleous, M., Murphy, M. M., Hu, Z., Huddlestun, S. E., Zhao, Z., et al. (2015). Oxidative stress inhibits distant metastasis by human melanoma cells. *Nature, 527*(7577), 186–191.

Plato. (2015). Theaetetus and Sophist. In C. Rowe (Ed.), *Cambridge texts in the history of philosophy* (pp. 99–177). Cambridge: Cambridge University Press.

Raspopovic, J., Marcon, L., Russo, L., & Sharpe, J. (2014). Modeling digits. Digit patterning is controlled by a Bmp-Sox9-Wnt Turing network modulated by morphogen gradients. *Science, 345*(6196), 566–570.

Rehermann, B. (2016). HCV in 2015: Advances in hepatitis C research and treatment. *Nature Reviews Gastroenterology and Hepatology, 13*(2), 70–72.

Reker, D., Blum, S. M., Steiger, C., Anger, K. E., Sommer, J. M., Fanikos, J., & Traverso, G. (2019). "Inactive" ingredients in oral medications. *Science Translational Medicine, 11*(483), eaau6753.

Robert, A. M. (2011). *Nonstandard analysis*. Dover Publications.

Robinson, J. P., & Roederer, M. (2015). Flow cytometry strikes gold. *Science, 350*(6262), 739–740.

Robustelli, P., Piana, S., & Shaw, D. E. (2018). Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences of the United States of America, 115*(21), E4758–E4766.

Rohwer, F., & Segall, A. M. (2015). In retrospect: A century of phage lessons. *Nature, 528*(7580), 46–48.

Romanes, G. J. (1891). Aristotle as a naturalist. *Science, 17*(422), 128–133.

Ross-Gillespie, A., Weigert, M., Brown, S. P., & Kummerli, R. (2014). Gallium-mediated siderophore quenching as an evolutionarily robust antibacterial treatment. *Evolution, Medicine, and Public Health, 2014*(1), 18–29.

Saade Lemus, P., Anderson, K., Smith, M., & Bullock, A. (2019). Spontaneous regression of pancreatic cancer with liver metastases. *BMJ Case Reports, 12*(5), e229619.

Schaapveld, M., Aleman, B. M., van Eggermond, A. M., Janus, C. P., Krol, A. D., van der Maazen, R. W., et al. (2015). Second cancer risk up to 40 years after treatment for Hodgkin's lymphoma. *The New England Journal of Medicine, 373*(26), 2499–2511.

Scholl, R., & Nickelsen, K. (2015). Discovery of causal mechanisms: Oxidative phosphorylation and the Calvin-Benson cycle. *History and Philosophy of Life Sciences, 37*(2), 180–209.

Seeman, J. I., & Cantrill, S. (2016). Wrong but seminal. *Nature Chemistry, 8*, 193–200.

Sekikawa, A., Horiuchi, B. Y., Edmundowicz, D., Ueshima, H., Curb, J. D., Sutton-Tyrrell, K., et al. (2003). A "natural experiment" in cardiovascular epidemiology in the early 21st century. *Heart, 89*(3), 255–257.

Servick, K. (2019). Humans may sense Earth's magnetic field. *Science, 363*(6433), 1257–1258.

Shou, W., Bergstrom, C. T., Chakraborty, A. K., & Skinner, F. K. (2015). Theory, models and biology. *eLife, 4*, e07158.

Shrinivas, K., Sabari, B. R., Coffey, E. L., Klein, I. A., Boija, A., Zamudio, A. V., et al. (2019). Enhancer features that drive formation of transcriptional condensates. *Molecular Cell, 75*(3), 549–561 e7.

Singer, P. N. (2016). Galen. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Metaphysics Research Lab, Stanford University.

Smit, S., Rother, K., Heringa, J., & Knight, R. (2008). From knotted to nested RNA structures: A variety of computational methods for pseudoknot removal. *RNA, 14*(3), 410–416.

Smith, K. T., Berritt, S., Gonzalez-Moreiras, M., Ahn, S., Smith, M. R., 3rd, Baik, M. H., & Mindiola, D. J. (2016). Catalytic borylation of methane. *Science, 351*(6280), 1424–1427.

Spinoza, B. (1992). *Ethics: With the Treatise on the emendation of the intellect and Selected letters*. (S. Shirley, Trans.). Hackett Publishing Company.

Steffan, J. S. (2016). A cause for childhood ataxia. *eLife, 5*, e14523.

Su, J. (2018). A brief history of Charcot-Leyden crystal protein/galectin-10 research. *Molecules, 23*(11), 2931.

Szybalski, W., & Bryson, V. (1952). Genetic studies on microbial cross resistance to toxic agents. I. Cross resistance of Escherichia coli to fifteen antibiotics. *Journal of Bacteriology, 64*(4), 489–499.

Tadrist, L., & Darbois-Texier, B. (2016). Are leaves optimally designed for self-support? An investigation on giant monocots. *Journal of Theoretical Biology, 396*, 125–131.

Tanner, J. M., & Rutter, J. (2016). You are what you eat... or are you? *Developmental Cell, 36*, 483–485.

Tauber, A. I. (2003). Metchnikoff and the phagocytosis theory. *Nature Reviews Molecular Cell Biology, 4*(11), 897–901.

Thamer, M., De Marco, L., Ramasesha, K., Mandal, A., & Tokmakoff, A. (2015). Ultrafast 2D IR spectroscopy of the excess proton in liquid water. *Science, 350*(6256), 78–82.

Thompson, A. A., Walters, M. C., Kwiatkowski, J., Rasko, J. E. J., Ribeil, J. A., Hongeng, S., et al. (2018). Gene therapy in patients with transfusion-dependent beta-thalassemia. *The New England Journal of Medicine, 378*(16), 1479–1493.

Toomer, G. J. (1968). A survey of the Toledan tables. *Osiris, 15*, 5–174.

Toprak, E., Veres, A., Michel, J. B., Chait, R., Hartl, D. L., & Kishony, R. (2012). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature Genetics, 44*(1), 101–105.

Tulk, C. A., Molaison, J. J., Makhluf, A. R., Manning, C. E., & Klug, D. D. (2019). Absence of amorphous forms when ice is compressed at low temperature. *Nature, 569*(7757), 542–545.

Uphoff, S., Lord, N. D., Okumus, B., Potvin-Trottier, L., Sherratt, D. J., & Paulsson, J. (2016). Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation. *Science, 351*(6277), 1094–1097.

Van Epps, H. L. (2006). Influenza: Exposing the true killer. *The Journal of Experimental Medicine, 203*(4), 803.
Varmus, H. (2016). The transformation of oncology. *Science, 352*(6282), 123.
Vikhanski, L. (2016). *Immunity: How Elie Metchnikoff changed the course of modern medicine*. Chicago Review Press.
Wallach, D., Kang, T. B., Dillon, C. P., & Green, D. R. (2016). Programmed necrosis in inflammation: Toward identification of the effector molecules. *Science, 352*(6281), aaf2154.
Walsh, C. T. (2015). A chemocentric view of the natural product inventory. *Nature Chemical Biology, 11*(9), 620–624.
Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133.
Weber, E., Van Bouwel, J., & De Vreese, L. (2013). How to study scientific explanation? In E. Weber, J. Van Bouwel & L. De Vreese (Eds.), Scientific explanation (pp. 25-37): Springer.
Wei, X., Prentice, J., & Balasubramanian, V. (2015). A principle of economy predicts the functional architecture of grid cells. *eLife, 4*, e08362.
Weiss, J. N., Qu, Z., & Garfinkel, A. (2003). Understanding biological complexity: Lessons from the past. *The FASEB Journal, 17*(1), 1–6.
WHO. (2019). *Malaria eradication: Benefits, future scenarios and feasibility* (Strategic Advisory Group on Malaria Eradication).
Willyard, C. (2016). Cancer therapy: An evolved approach. *Nature, 532*(7598), 166–168.
Wong, S., & Slavcev, R. A. (2015). Treating cancer with infection: A review on bacterial cancer therapy. *Letters in Applied Microbiology, 61*(2), 107–112.
Woodger, J. H. (1929). *Biological principles: A critical study*. London: Kegan Paul, Trench, Trubner & Co.
Xie, X. S. (2013). Biochemistry. Enzyme kinetics, past and present. *Science, 342*(6165), 1457–1459.
Yang, N., Duong, C. H., Kelleher, P. J., McCoy, A. B., & Johnson, M. A. (2019). Deconstructing water's diffuse OH stretching vibrational spectrum with cold clusters. *Science, 364*(6437), 275–278.
Yucel, N., Chang, A. C., Day, J. W., Rosenthal, N., & Blau, H. M. (2018). Humanizing the mdx mouse model of DMD: The long and the short of it. *NPJ Regenerative Medicine, 3*, 4.
Zhong, Q., Pevzner, S. J., Hao, T., Wang, Y., Mosca, R., Menche, J., et al. (2016). An inter-species protein-protein interaction network across vast evolutionary distance. *Molecular Systems Biology, 12*(4), 865.
Zuniga, A., & Zeller, R. (2014). Development. In Turing's hands – the making of digits. *Science, 345*(6196), 516–517.

Proposal-Based Innovation: A New Approach to Opening Up the Innovation Process

Karl H. Ohlberg and Jose L. Salmeron

Abstract

A basic principle of innovation is synthesis, a recombination of previously unconnected concepts. A known framework in this regard is Open Innovation (OI), which is widespread in nonmanufacturing industries such as software. In manufacturing, however, OI is largely rejected mainly due to high intellectual property (IP) protection requirements. This chapter describes an approach that can break through this aversion of manufacturers. The first step is to make the innovation activities of manufacturers transparent in a wide variety of industries and in a structured way, based on facts freely available on the Internet. The concern is only with WHAT companies do, have achieved, or intend to do, and not their intellectual property, i.e., HOW they do it. Remote locations (worldwide) and outside industries are particularly important because the information deficit is greatest in these axes, not least because search engines offer only limited and inefficient help when carrying out research in this regard. The second step is to use artificial intelligence (AI) to turn the extracted structured facts into concrete creative proposals for innovation and cooperation between different manufacturers (regardless of location and industry) in order to stimulate the innovation process in new ways. This principle is what Karl H. Ohlberg calls Proposal-based Innovation (PBI). The advantage lies in the fact that sign up, be a member, share your ideas, the idea of the Web 2.0 and Open Innovation from the end of the last millennium, is replaced by the use of the immense amount of

K. H. Ohlberg (✉)
EmpraGlob GmbH, Dusseldorf, Germany
e-mail: karl@empraglob.com

J. L. Salmeron
Universidad Pablo de Olavide, Seville, Spain
e-mail: salmeron@acm.org

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_14

data that has accumulated over the past 20 years. Discussions with manufacturers reveal there is great interest in this principle, so that the goal is to develop a cross-national prototype.

1 Introduction

In recent decades, manufacturing industries have contributed enormously to the prosperity of people and economies through the globalization of manufacturing ecosystems. These manufacturing ecosystems essentially relate to supply chain networks as well as marketing, service, and recycling networks that were established in different, distinct industries. However, the picture is different in R&D and innovation. Most of such work is done in silos. Although there are regional business clusters that promote innovation cooperation on a local basis and there are also supra-regional initiatives in individual sectors, there is an almost complete lack of global innovation ecosystems. Cross-industry links, in particular, are practically nonexistent. The popular concept of *Open Innovation* (OI) (Chesbrough 2003), which has been successful in nonmanufacturing industries since the beginning, is also not well accepted in manufacturing industries (see Sect. 6).

This chapter describes a new innovation approach in manufacturing industries, which was developed in a practical project of EmpraGlob GmbH, a German consulting company. This project involved hundreds of discussions with C-level executives of medium and large manufacturers and their top innovation experts, entrepreneurs, and academics from 15 different countries, including the United States, China, Germany, the United Kingdom, France, India, Spain, Australia, Switzerland, Austria, and the UAE.

2 A New Approach to Innovation: Task and Goal

The starting point was the idea of making external knowledge, technologies, ideas, and concepts more accessible and usable in the innovation process of companies across borders and industries. The challenge began with providing a facility to search for innovation activities worldwide, including analytical tools for all types of manufacturing industries.

Insights into innovation activities on a global scale are, however, urgently needed in the age of digitization and global structural change in order to minimize innovation risks and achieve entrepreneurial advantages (e.g., reduction of development costs or time to market). Greater use of recombination of previously unconnected concepts is also an important topic. Recombination, today's most important driver of innovation, is also known as synthesis and is one of the basic principles of innovation. Overall, manufacturing companies want and need to understand which innovation activities are currently being carried out in other regions and industries worldwide, and must learn how they can benefit from them. Existing innovation

networks do not reach far enough and hence are inadequate to allow companies to thrive in the current innovation landscape. A new approach is therefore needed.

Another concern of this chapter is to understand why OI has not written a similar success story in the long history of manufacturing industries as it has in nonmanufacturing, i.e., software or service industries (what economists call the tertiary sector).

The overriding goal of the underlying project is to find a new approach to innovation based on the requirements of manufacturers; an approach aimed at building a novel platform to open up innovation processes and create multinational innovation networks and ecosystems. Another concern is to encourage manufacturers to rethink current approaches, to help them create a new mindset, at least for those that want to or have to do so. The developed new approach can also be a catalyst to help manufacturing industries move toward pervasive OI at some time in the future, as is already the case today in typical information technology industries.

The practical implementation of such a new platform is not covered in this chapter. This is a topic in its own right, for which a detailed concept has already been developed.

This chapter should also provide an opportunity to stimulate discussion about this approach at the most international level possible. Karl H. Ohlberg has detailed additional concepts and plans. Readers are invited to contact him for more information on the underlying project. Cross-border exchange of ideas and a global picture are the most important and essential foundations of this work.

3 Manufacturing Industries: A Definition for This Chapter

There is a large variety of manufacturing industries, which are summarized in this chapter under *manufacturing*. Since the term *manufacturing industries* is used in very different contexts, it is defined below for use in this chapter. In addition to *manufacturing*, there is also the term *production*, and the two are sometimes used synonymously. There are various definitions of the terms *manufacturing* vs *production*. A common distinction is simplified: Manufacturing refers to tangible products, production to tangible as well as intangible products, services included. Since tangible products play the decisive role in this chapter, the term *production* is not used here, but exclusively *manufacturing*. Particularly important in this context is the distinction between *manufacturing* and *nonmanufacturing,* that is to say *services*, e.g., software development. The reason for this is that both groups have completely different innovation behavior (see Sect. 5). A borderline case is the Industrial Internet of Things (IIoT), software technology that works very closely with hardware, directly supports the automation process in manufacturing environments, and is often deeply integrated into the manufacturing process, and is included here.

In this chapter, the term *manufacturing* is used for manufacturing as a whole. This includes everything where something solid, liquid, or gaseous is produced or processed, or where knowledge or concepts in this regard are developed or

researched. So it includes all discrete or process industries (approximately what economists call the primary and secondary sector), including logistics and the circular economy.

Examples of these sectors are:

- Aerospace/Aviation/Defense
- Agriculture
- Apparel/Footwear
- Automotive/Transportation
- Chemicals/Materials
- Construction/Construction materials
- Consumer goods
- Electrical and electronic
- Energy
- Food/Beverages
- Healthcare/Life sciences
- Heavy industry
- Industrial automation/IIoT
- Logistics
- Machining/Industrial
- Packaging
- Plastics
- Shipbuilding
- Textiles
- The circular economy

4 Challenges and Opportunities in Manufacturing Industries

Many industries are already forced to adapt at an extremely high pace. Based on an example, it will be explained how important it is for companies to open up the innovation process. This process is also about the priority of precise information and analyses concerning the innovation activities of other market participants in one's own and in foreign industries, taking a global perspective.

4.1 An Example

At present, the automotive supplier industry is an example of an industry that is affected by enormous changes. In a news article dated September 25, 2019, the German automotive supplier Continental AG announced that 20,000 of its 244,000 jobs would be cut, and at least five factories would be closed over the next 10 years due to the slowdown in the global automotive sector. The program is expected to

lead to costs of around $1.21 billion by 2022. From 2023, the supplier wants to save around $550 million per year in gross costs (Sims 2019).

This example is a warning signal to an entire industry. Declining vehicle sales, the weakening economy and the transformation of the automotive industry are forcing suppliers around the world to make adjustments. Companies in this situation will not have a good future simply by implementing cost reductions such as process optimization and plant closures.

4.2 The Options

This example shows two things:

1. Once a company is in such a challenging situation, an agile innovation process can help in growth areas that are close to the current business and can help achieve profit as quickly as possible. In this example, these options would then be forward-looking solutions such as autonomous, networked, or electric driving.
2. It is better if a company does not get at all involved in such a situation by recognizing trends in its own business in good time beforehand and by opening up new lucrative business areas much earlier. These new fields may then be further away from the core business—both thematically and geographically—because more time is available for business development.

It is immediately clear that an efficient innovation process is necessary in both of the scenarios mentioned above in order to achieve the best results with existing budgets. Ideally, this process starts with an analysis of all innovation activities that all the different actors have in all relevant geographies. This is the best foundation to develop options for action. Without question, in today's world of increasingly complex products, external ideas, knowledge, technologies, and concepts must also be considered, whether in partnerships or in M&A processes. At the same time, it is also important to examine which elements of the company's own intellectual property may be of interest to other companies. This can result in an additional revenue stream through the granting of licenses or access to completely different industries in development partnerships. Another possibility is to sell one's own technologies or business units.

4.3 Insight Is Essential

In any case, the best possible target companies in a global scenario must be identified. The thematic search area must not only be limited to the own industry, but must also include other suitable ones. The identification of these targets is a crucial process, which is very challenging because there is little transparency about the details of the activities of companies. In particular, cross-industry search and analysis poses a huge difficulty because the search process requires a great amount

of creativity and is particularly complex due to the demands for multinational searches.

4.4 Early Warning Indicators

Companies should also know in which fields early warning indicators could be found that need to be monitored. These include some fields that may not immediately be thought of. These indicators usually receive little attention, particularly in the tunnel vision of operationally focused managers. This is due not only to the fact that precisely coordinated controlling processes in companies sometimes hinder the innovation process, but also to the fact that such analyses have so far been very time-consuming and can hardly ever be carried out completely by even the largest companies, and small- and medium-sized manufacturers cannot even consider them.

4.5 Need for Worldwide Intelligence?

The question now arises as to whether the innovation intelligence described above, i.e., the search for and analysis of innovation activities, really must take place worldwide. The answer is: Yes, it must, it is important even for smaller companies, given that innovation plays a role. The reason is that we are at the beginning of a fourth stage of globalization, in which not only products, but also knowledge is becoming fluid due to the enormous increase in global communication opportunities. In the exchange of ideas, knowledge, technology, and concepts, the probability of a high level of innovation arbitrage—i.e., the different valuation of performances—is much greater internationally than nationally. It is by no means just a matter of purchasing concepts at lower prices; it is also possible that exclusive innovations from a high-wage country can be particularly highly rewarded in a low-wage country. As far as internationality is concerned, national partners can of course also prove to be the most suitable. However, innovation intelligence should always take international aspects into account. The fourth stage of globalization is discussed in more detail in Sect. 4.6.1.

4.6 Changes and Influences

Massive changes in technological, economic, political, and social structures are one of the main reasons for entrepreneurial challenges, and they can often occur quite suddenly. A characteristic feature of the current era is that many of these changes have multinational dimensions or at least a multinational origin, and in some cases there is global influence. What is striking is that shocks in the global economy are occurring at ever shorter intervals and, above all, can overlap.

Although many of these changes usually have little impact on a particular business in the short term, they can become major risks in the medium and long

term if ignored. However, if they are recognized and strategically exploited, they can offer enormous opportunities. However, recognition requires careful observation. Questions are then, for example: Which regional or thematic markets or partnerships will become obsolete in the future, and which new ones will emerge?

It therefore quickly becomes clear that manufacturers must make a special effort to observe all essential points, to define relevant early warning indicators, and to derive innovation activities from them. In the following, some points are described in which particularly strong changes already exist, or are expected soon. In the compilation, care was taken to also include those points that are not necessarily on the radar of managers. The aim is to show how great the variety of changes is. These points are particularly important for understanding the concept presented in Sect. 8.

4.6.1 Changes in the Nature of Globalization

Experts on the world economy see a new stage of globalization emerging, in which borders continue to disappear and, after a long phase of an intensive exchange of raw materials and goods, human brainwork is now also becoming fluid worldwide (Baldwin 2018). Another description of this development is that political geography, that is, the way we legally divide the world, develops into a functional geography that then describes the reality of how we actually use the world (Khanna 2016). This development is described as an evolution that is currently only taking place to a small extent, but is progressing unstoppably.

It is easy to understand that greater connectivity based on immediate brainpower can provide great opportunities for opening up the innovation process. It will offer great competitive advantages to those companies that can recognize and deal with developments and, above all, that know where and how to find suitable opportunities. It is not to be expected that cooperation or traditional personnel recruitment will be replaced on a large scale in a disruptive manner. However, in the future, it will be much more common for locally separated, but internationally collaborating experts to come together in digitally linked working groups. This will allow companies to recruit complete development teams from remote and scattered locations. There have already been initial attempts to adopt this approach and it has the potential to become mainstream. Traditional development departments will need to rethink if they are to take advantage of these benefits.

4.6.2 Changing World Order

The change in the nature of globalization described in the last section is closely linked to the change in the world order, but there is an important difference. While the first one describes how the economy and people use the world to do business efficiently, the second one describes how leaders of major countries shape and transform their economies, creating competition between different systems and their economic networks (*multiconceptual*, see below). There are also tendencies toward controlling economies with business management methodologies, e.g., long-term planning and KPIs.

As far as the world order is concerned, a bipolar power structure in the Cold War and foreclosure period followed a period of unipolar US power that was then replaced by the complex world order in which we live today. This new order not only has multipolar power structures, but is also *multiconceptual* (Collins 2019). The development toward multipolar power structures with several very different concepts is a consequence of the enormous development of emerging economies. China's rise as an economic and geostrategic competitor to the United States is a sign of changing geopolitics (Fayd'herbe 2019). The China–United States trade war, for example, is directly linked to the changed world order. Developments toward major changes in regulatory structures are also taking place in Europe with Brexit. Interestingly enough, this power struggle is taking place at the same time as the China–United States trade war.

The new form of globalization discussed in the previous section, however, is being met with rejection by some political decision-makers. There are novel disruptive anti-globalization forces and tendencies toward isolationism. US President Donald Trump, for example, said in his 2019 speech at the U.N.: "The future does not belong to globalists. The future belongs to patriots" (Crowley and Sanger 2019). However, election results can also quickly change the situation. "We will be back" said Former US Vice President Joe Biden at the Munich Security Conference 2019 in Germany (Rogers and Sanger 2019).

These changes in the world order are very difficult to predict and can occur suddenly. Especially in manufacturing industries, the effects are obvious. It is not only the case that existing partnerships in trade and supply chains and even entire sales and procurement markets can be destroyed, but also technological cooperation, as the example of the Chinese technology company Huawei shows. Changes in the world order can be associated with the greatest uncertainties.

4.6.3 Innovation Centers Are Shifting

One of the most remarkable developments of the last 30 years has certainly been the rise of Asia. Western observers assume that Asia will continue to have a strong growth future. By 2040, Asia is expected to contribute 50% of the global gross domestic product (GDP) and to account for 40% of the world's total consumption. In 2018, 210 of the 500 largest companies (by turnover) came from Asia, as the Fortune Global 500 ranking shows (Khanna 2019).

While Asia was previously primarily important as a procurement and sales market, today it has the potential to become a hub for innovation (Pande and Anil 2018). The German government also reports on the shift of innovation centers: "Europa verändert sich, die Welt wächst zusammen. Globale und regionale Innovationszentren verschieben sich nach Zentralasien, Lateinamerika und Afrika" [Europe is changing, the world is growing together. Global and regional innovation centers are shifting to Central Asia, Latin America and Africa] (Bundesministerium für Bildung und Forschung (BMBF) n.d.). This quote from the Federal Ministry of Education and Research of the German government describes a development that companies must observe and deal with. The "China Strategy 2015–2020" of the same ministry mentions further details: "The China Strategy of the Federal Research

Ministry plays a key role in implementing the comprehensive strategic partnership between Germany and China which was jointly announced by President Xi Jinping and Chancellor Angela Merkel in March 2014" (Bundesministerium für Bildung und Forschung (BMBF) 2015).

Existing partner networks in any area will have to change as a result.

4.6.4 How Innovation Has Changed

Manufacturing innovation arises either through discovery, or experimentation, or through the recombination of existing knowledge and concepts, also known as synthesis. Here is an example: the production of charcoal using the heat of fire under certain conditions was a discovery. The development of the process for the rational mass production of charcoal was an experiment. The development of briquettes was then a synthesis. The problem with the mass production of charcoal was that many small pieces of waste were formed. The aim now was to use this waste, because it had a value. It was known that charcoal was a natural fiber material, just like wood. The other knowledge came from the paper industry, which had discovered that natural starches could be used as good binders for natural wood fibers. By mixing a starch and charcoal slurry, the process for the production of briquettes was created.

The possibilities for discovering new knowledge or concepts have been declining for some time. The same applies to experiments. Big ideas are getting harder to find. For that reason, innovation is moving at a great speed toward synthesis (see Fig. 1).

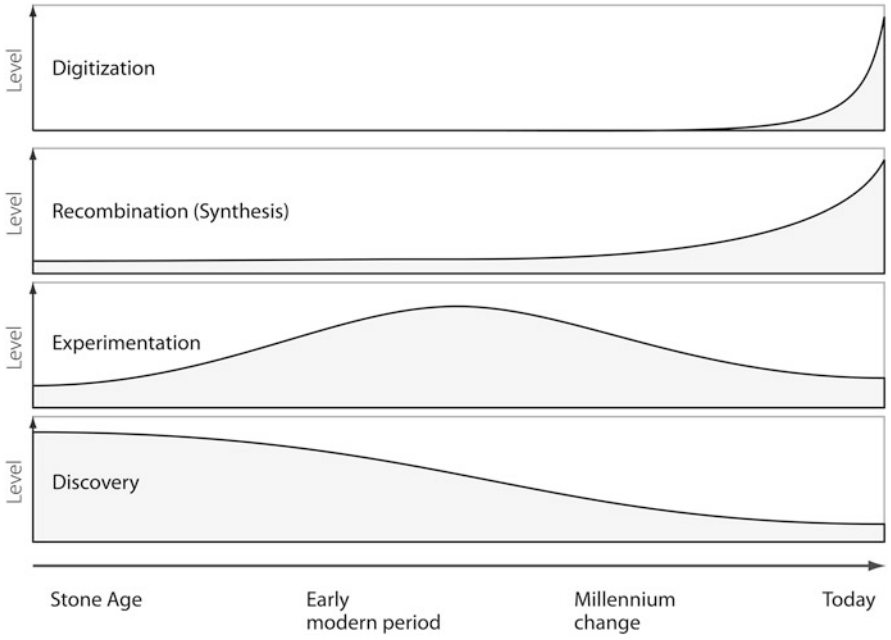

Fig. 1 How innovation has changed. Categorization from Burnett (2009). Source: authors

The result: products are becoming increasingly complex (see next section), and the possibilities of the innovation process are increasing. In addition, this process is accelerated by digitization, which offers even greater possibilities through many new technologies and business models.

4.6.5 Increasing R&D Expenditures

A very common topic of our time is the fact that products are becoming more and more complex, and the R&D effort required for the related innovation process is becoming ever greater. Research has shown that an economy has to double its research expenditure every 13 years in order to achieve a constant growth rate. This means that there is an exponential increase in R&D costs (Bloom et al. 2017). As companies in national economies are the largest contributors to R&D expenditure, it follows that firms are also affected by the increase. It is certainly the case that there are manufacturing sectors that are not strongly technology-oriented and in which the increase is less pronounced. However, even in these cases, the increase is considerable and in any case, great ideas are increasingly difficult to find.

These connections clearly show how important external knowledge is. The big questions are:

- Who does what exactly, and where? (Overview)
- Where can external concepts be sought that may be relevant? (Concepts and location)
- What are the best options for recombining existing concepts? (Application)

4.6.6 Time to Market

"If the rate of change on the outside exceeds the rate of change on the inside, the end is near." The message from this quote of Jack Welch, the famous former CEO of General Electric, is applicable in many cases, and can even help in recognizing whether a company is bringing new products to market quickly enough.

Time to market describes the period of time that passes from the idea for a product to its market launch. The term has its origin in industries with products with a short life cycle, e.g., technology industries. Also in manufacturing industries, the speed has increased with the ever-increasing expansion of digital technologies and the stronger individualization of products. Today, an adequate time to market has become very important for the competitiveness of most companies.

Large companies have been using a precisely coordinated time to market for quite a long time. Now the time has come for medium-sized companies to take advantage of this approach, as cooperation with development or manufacturing partners can significantly shorten time to market without companies having to invest in their own infrastructure.

Fig. 2 Comparison of global competitors. Data from Pisano and Shih (2012, p. 30). Source: authors

4.6.7 Digital Divide

In today's economy, the general trend clearly indicates that some major organizations are making use of digitization almost completely while others only benefit partially, or even worse, are just focusing on traditional strengths.

Throughout the world, a large majority of traditional companies with predominantly traditional structures are active in all manufacturing sectors. It has long been apparent that small and medium-sized companies in these sectors are adapting only hesitantly to the age of digitization. In the electrical industry, the leading sector for digitization in manufacturing industries, a study was carried out in 2016 by the German Electrical and Electronic Manufacturers' Association confirming this observation (Frietsch et al. 2016). There is a danger of a digital divide between a few large companies that are characterized by intensive use of digitization, and a large number of small and medium-sized enterprises (SMEs) that are unable or unwilling to take advantage of the opportunities offered by digital products, processes, or business models (Frietsch et al. 2016). As primary obstacles to digitization, the study states that manufacturers are taking too long to adopt digital technologies, and that digitization is mainly successful when networked with other companies, but necessary collaboration is not happening enough.

4.6.8 Increasing Competition

Increasing competition in the last two decades is a worldwide phenomenon. This subject is so well known that it will not be discussed at length. As an example, Fig. 2 shows the United States with its competition based on the number of qualified workers in manufacturing industries. Although the data are older, the development is likely to have continued, especially due to the rise of Asia.

4.6.9 Corporate Social Responsibility (CSR)

Laurence D. Fink, founder and chief executive of the American global investment management corporation BlackRock, Inc., has informed business leaders that their companies need to make more than just profits. They need to contribute to society as well if they want to receive the support of BlackRock. "Society is demanding that companies, both public and private, serve a social purpose," Laurence D.

Fink wrote. "To prosper over time, every company must not only deliver financial performance, but also show how it makes a positive contribution to society." He is seeing "many governments failing to prepare for the future, on issues ranging from retirement and infrastructure to automation and worker retraining" (Sorkin 2018).

Another aspect of contributing to society is that some advanced economies, such as the United States, have the potential for a manufacturing renaissance. In some cases, for instance, governments are already demanding that companies take back production. Companies can actively support this process. Such a development can be advantageous for societies because those industries often employ appropriately paid staff from the middle class of societies (Pisano and Shih 2012).

Many manufacturers will have to adapt and invest heavily in CSR to meet requirements from governments or investors such as BlackRock.

4.6.10 Growing Middle Class

Primarily through the rise of Asia, in the coming decades, the growing global population with higher incomes will lead to a sharp increase in global demand. The OECD states in this context: "Global gross domestic product (GDP) is projected to quadruple between 2011 and 2060, according to the central baseline scenario projected by the OECD ENV-Linkages model. By 2060, global average per capita income is projected to reach the current OECD level (around USD 40 000)" (OECD 2019, p. 3).

After September 2018, "for the first time ever, the poor and vulnerable will no longer be a majority in the world. . . . Compared to today, the middle class in 2030 will have 1.7 billion more people" (Kharas and Hamel 2018).

As a result, manufacturers have strong growth expectations, not only in terms of production, but also in terms of recycling. Since the circular economy also belongs to the manufacturing sectors in our definition, there is a strong potential for growth overall (IndustryWeek Custom Research and Kronos Incorporated 2016).

4.6.11 Aging Society

"According to data from World Population Prospects: the 2019 Revision, by 2050, one in six people in the world will be over the age of 65 (16%), up from one in 11 in 2019 (9%)" (United Nations 2019). This development represents an enormous opportunity for manufacturers to conquer these future markets through innovation processes.

4.6.12 Megacities

There is a major trend for people to live in megacities. By 2030, two-thirds of the world's population will live in megacities. There will then be 50 megacity clusters worldwide (Khanna 2016). For manufacturers, this is a great opportunity for growth in often completely new business areas.

It is also expected that governments will increasingly promote megacities because they may be the only way to reduce the world's overpopulation. Experience has shown that people living in megacities reproduce comparatively little.

This development will also provide enormous opportunities for manufacturers of all kinds if they are innovative enough to develop appropriate products for this way of life.

5 About Startups and Manufacturing

When the term *innovation* is used today, it is in many cases linked to digitization and startups. If we now look at the industries in which startups are active, we see that they are almost exclusively active in service industries.

5.1 Services Startup Environment

A current international overview is provided by US business magazine Inc. with the list of *50 World-Changing Startups to Watch in 2019* (Meyer 2018). Forty-nine of the 50 listed companies come from the *Services Startup Environment* outlined in Fig. 3. Companies in this environment are a large variety of digital services providers who are opening up a very large market with innovative business models. For example, the US company Uber Technologies, Inc. is a well-known example in the Mobility category (Fig. 3).

The only manufacturing-oriented startup on the list of *50 World-Changing Startups to Watch in 2019* is Hex Labs,[1] a US company that claims to be developing a nano-material. However, this small research startup only builds prototypes and is not a manufacturer of market-ready mass hardware products.

A similar picture emerges worldwide: startups are actually services startups, for the most part in nonmanufacturing industries, whose share is about 98% of all startups, as published lists of startups show.

5.2 Manufacturing Services Startups

Digital technologies play a role in all types of services startups in manufacturing industries (see Fig. 3), such as the Industrial Internet of Things (IIoT), which includes maintenance and predictive technologies (e.g., predictive analytics). Other types include simulation tools, machine learning and artificial intelligence (AI), as well as cyber security. It is very clear that these startups can help increase manufacturing efficiency, but the actual R&D process and mass production of physical products is hardly addressed here.

It is important to note that for this reason, manufacturing services startups more often than not depend on partnerships with other, usually larger, often very large manufacturers. For those startups, it is not easy to sell their new concepts

[1] https://www.hexlabsco.com

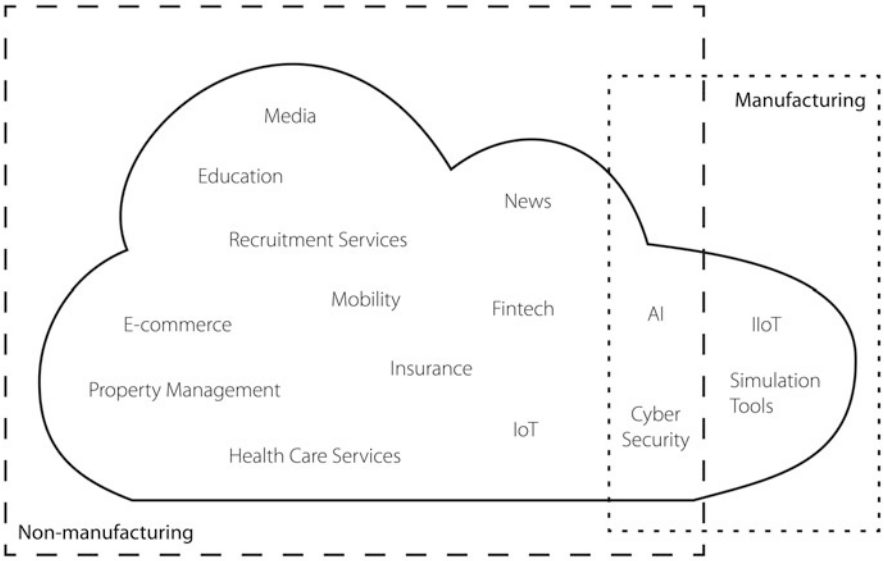

Fig. 3 Services startup environment. Source: authors

to traditional manufacturers. In the new ways that such startups go, not only in technology, but also in business models, they frequently encounter various obstacles. Traditional industrial companies are used to working in silos, have long innovation cycles, are risk-averse, are organized hierarchically top-down, and have very specific requirements. Industrial companies often initially show a great interest in certain services startups and raise hopes, but then become organizationally very sluggish, demand long project durations, and often present a great challenge for those startups in terms of financing sales cycles. This situation stands in contrast to nonmanufacturing services startups, which have consumers as their target group. These customers can bring fast sales successes and enable quick feedback from the market, which is what makes possible the typical structured course correction in the development phase, the so-called startup pivoting.

5.3 Mass Manufacturing Startups

In searching for mass products startups related to innovation and manufacturing, the path leads to e-scooter-sharing systems. Here, innovative mass-produced hardware plays a major role. One company in this area, for example, is the US company Bird Rides, Inc. (active in United States, Europe, Middle East). This company belongs to the mobility category of services startups. Although it owns and rents out electric scooters, it is a pure service company, which is not involved in the development and production of the material. The same applies to other startups in this industry.

They leave the hardware mass production to other companies. The search for manufacturers of these rental scooters leads to the world of real manufacturing startups. Interestingly enough, almost every electric scooter produced on earth comes from a single Chinese manufacturer named Ninebot Inc. (Bergen and Brustein 2018). The vehicle maker Ninebot Inc. is headquartered in Beijing, China, was founded in 2012, has 3000 employees and took over US rival Segway Inc. in 2014. The main investors are Chinese smartphone maker Xiaomi Inc. and venture capital firm Sequoia Capital China (Shih 2015). Other mass manufacturing startups include Tesla, Inc. (USA), Faraday Future (USA), Byton (China), and StreetScooter GmbH (Germany). All these companies are characterized by occupying the new field of electro-mobility, by being OEMs (finished products, not suppliers) and by having received high investment from other major companies in the early stages.

In rare cases, however, there are also smaller young production companies. Cepton Technologies, Inc., a mass manufacturing startup from Silicon Valley, USA, which manufactures Lidar devices (for autonomous driving and other applications), is an example. The following quotation from a trade journal gives a good insight into this small mass manufacturing startup:

> The Cepton principals are old for Silicon Valley companies. They adopt what's there. This is a more practical approach—a traditional business model: make and sell a real product at a profit. Therefore, there is a rhythm in business developments. This philosophy determines the technology approach. Cepton is not developing fancy technology without revenue—this doesn't give cash flow. The goal is to bring real products to solve meaningful market problems today. (Lukas et al. 2018, p. 40)

Other Mass Manufacturing startups are in the robotics, drones, and additive manufacturing sectors.

In principle, all mass manufacturing startups are likely to be acquired by a bigger company—in many cases a traditional one—such as the German e-vehicle startup StreetScooter GmbH, which was founded in 2010 and acquired by Deutsche Post DHL Group in 2014, and is currently for sale again.

5.4 Conclusion

The international startup scene is characterized by very innovative services startups in nonmanufacturing industries. This typical startup scene can hardly be found in real manufacturing, if at all, only in the field of manufacturing services startups (see Sect. 5.2). However, there are only a few of them (2% of all startups), and they depend on cooperation with traditional manufacturers. This restricts their actions.

Mass manufacturing startups that manufacture products are even rarer. There are no figures available for the quantity or percentage of all startups. However, it can be assumed that their share of all startups is far below 1 in 1000 startups. Mass manufacturing startups are breaking new ground in their respective areas, e.g., electro-mobility. Some of them also have new business models (e.g., online sales at Tesla, without traditional dealerships). However, in all cases, they do have traditional components in their business, such as supply chains. A radical rethinking

of business approaches, as with services startups in nonmanufacturing, is not at all possible.

In the core business activities of the majority of traditional mass manufacturers, startups do not play a role. This situation is very similar worldwide. As a result, the traditional manufacturer scene is not really mixed up, changed, or challenged by startups. Their innovation processes remain essentially unchanged. If at all, manufacturing services startups may be able to play a role, but traditional manufacturers usually treat them like suppliers and require lengthy project times when collaborating with them.

The considerations in this section show one thing quite clearly: only nonmanufacturing services startups, which exist in large numbers worldwide, have real freedom with regard to their innovation process. In contrast, manufacturing services startups, which exist in small numbers, certainly also have highly innovative, digitization-driven startup ecosystems that can help improve manufacturing, but usually only contribute to increasing efficiency in the factory (e.g., Smart Factory). As a result, manufacturing services startups are dependent on projects from large traditional manufacturers and have to adapt to them. They only rarely get to the heart of product development innovation.

It is important to recognize the fundamental differences between manufacturing services startups and manufacturers of physical mass products. The actual mass manufacturers, who are often traditional in their innovation processes, digitization processes and business models, have no startup culture at the core of their business, and startups have little chance of breaking into it. On the other hand, it is precisely these mass manufacturers who face major challenges, as described in Sect. 4. They therefore need a new approach to their innovation process, which is at the heart of this chapter, and will be described in Sect. 8.

6 Open Innovation (OI)

Before the new *Proposal-based Innovation* (PBI) concept is presented (Sect. 8), there will first be a discussion of which application the well-known approach of *Open Innovation* (OI) has in manufacturing industries.

OI is a methodology to open up the innovation process of organizations of all kinds. This method is based on the willingness of participants in a community to be open to the ideas of others and to share knowledge with them. The idea seeker, usually an organization, actively involves customers, suppliers, business partners, and other experts with different backgrounds in an idea development process. Ideas, knowledge, and innovative concepts are jointly generated in this setting.

In short, with this method, external knowledge can be imported and internal knowledge can be made available to other organizations. It originates geographically from Silicon Valley, USA and the technology industries there, in the former environment of companies such as Xerox, IBM, Lucent, and Intel. The term OI was coined by Henry Chesbrough (Chesbrough 2003). The opposite of OI is *Closed*

Innovation, in which competitive advantages are achieved through the exclusivity of one's own knowledge (Vrontis 2013, p. 169).

In retrospect, it is interesting to note that OI was created in 2003 at the same time and place (Silicon Valley, USA) as the Web 2.0 (Aced 2013). The Web 2.0 gradually offered end users the opportunity to use the previously static websites dynamically and to get involved in the participatory Web with their own contributions and in digital communities.

OI has proven itself in many projects, especially in the heyday of Web 2.0. Most users can be found in services industries. In addition, almost all technology companies, especially community-based software companies, are driven by OI. This penetration, however, did not take place in manufacturing companies. Only a few large companies are among the users (e.g., General Electric, Lego) for special applications.

6.1 General Limitations

Sixteen years after the emergence of OI and the Web 2.0—a long time, especially in the enormously fast digital environment—the world is very different. The virtually empty Web of the early days is now full of valuable facts and figures. As a result, more and more data-driven business models have emerged. Mobile communication and apps have conquered the world. Invitations like *Sign up, be a member, share your ideas, communicate and collaborate*, with the initial community spirit of the Web 2.0, are no longer attractive to most Web users today. On the contrary, there is even aversion because business people are harder pressed for time than ever before.

Basically, OI projects are about communication and relationship building, and an exchange of ideas, knowledge, technologies, and concepts should take place. These activities require at least two partners for the exchange, usually more. For that reason, in OI projects, networks must be established in addition to core competencies. The big question is where these network partners should come from. OI (tacitly) assumes that the partners meet more or less automatically, e.g., on the Internet or in real communities. Consequently, OI requires active behavior by all actors. In particular, potential projects must be known (awareness of projects) and individuals must contact each other. In addition, these connections must be honestly desired by all parties. Awareness, especially, cannot be guaranteed. Not every potential stakeholder gets the message. Important prospective contributors, who might also like to participate, simply do not notice a project. Particularly today, in times of increasing information overload, such a communication deficit can be a real issue. In addition, business people today have a laser-sharp focus, narrower than ever before. This focus is being promoted by the principle of *Closed Innovation* that still prevails in most manufacturing industries worldwide.

The central question is how to gain an appropriate base of potential partners. Those who are passive are left out. If in OI projects it is assumed that there is a suitable base of potential partners available, but in reality it is too small or only conditionally suitable, e.g., due to regional restrictions, then the project cannot run

optimally. This is especially true when the expected community has great cultural differences, e.g., between the western world and Asia, and thus the hurdle is even higher.

The author Paul Sloane also reports on concrete barriers in OI:

> However, there are still large swathes of businesses which remain untouched by the Open Innovation initiative. They refuse to join the bandwagon. The leaders of these organizations pay lip service to Open Innovation and claim that their people are open to outside ideas and collaboration but the reality is very different. There appear to be a number of very real impediments. (Sloane 2015)

These cultural and process barriers will be described in more detail in the next two sections.

6.2 Cultural Barriers

Cultural barriers in intercultural communities have already been mentioned above. However, there are also cultural barriers that can occur within a company with a unified culture in an OI process. This is the case when companies concentrate on themselves and they are convinced that they know their customers and their demands. They work very efficiently and have no time for network meetings or the search for partners. The old assumption *we know best* applies there. Outsiders are viewed skeptically and there is a fear of losing IP. These people are not used to working and sharing in transparency (Sloane 2015).

6.3 Process Barriers

Many large companies have processes that define the development of new products. Various internal committees and decision-makers are supposed to protect respective companies from risks, fraud, and unnecessary costs. In addition, there are legal departments that are involved in the execution of those business processes. There are usually no executives with OI responsibility. The task is divided between R&D and marketing, which leads to rivalries. In addition, there is usually no budget for OI, so it has to be taken from other overburdened budgets. In addition, the calendars of all those involved are overcrowded, leading to delays and frustration for partners (Sloane 2015).

6.4 Intermediaries

In order to avoid the problem of cultural and process barriers, intermediaries have established themselves to support the innovation projects of companies in so-called crowd-sourcing processes. An example of such a company is the US company

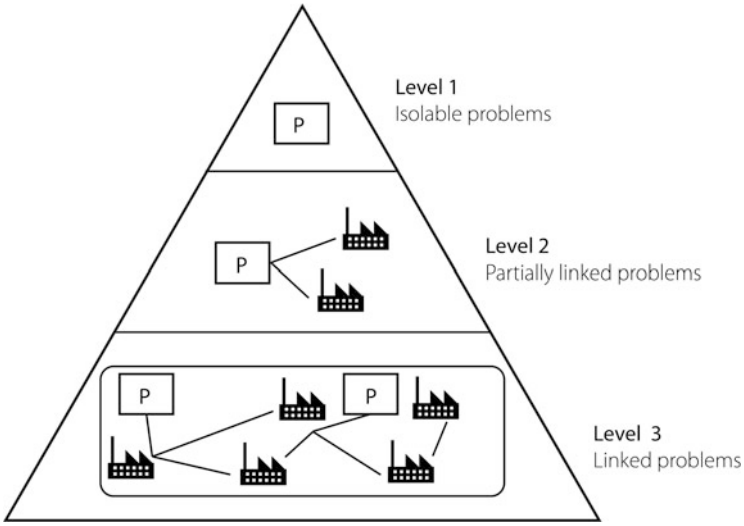

Fig. 4 Innovation intermediaries—different levels of problems. Source: authors

InnoCentive, Inc.[2] To put it simply, companies interested in ideas and concepts can submit their requests in a document (so-called *challenges*) to a worldwide network of people and companies for a fee, in which so-called *solvers* are the experts. The solvers are often scientists in low-wage countries. The inquiring companies then receive a number of solutions, from which they can choose one, whose solvers then receive a bonus. The process has its origins primarily in the pharmaceutical and chemical sectors.

However, the limitation of intermediaries is that the described procedure can only solve problems that can be isolated (Fig. 4, Level 1) and, with restrictions, problems that are partially linked (Level 2). The development of a chemical formula, for example, is an isolable problem (Level 1) and therefore suitable for intermediaries. However, the innovation projects of manufacturers are in most cases related to at least partially linked problems (Level 2) and in many cases heavily linked problems with related tasks in other companies (Level 3). These Level-3-problems require networked structures at the enterprise level that are not compatible with the crowd sourcing process of intermediaries.

Intermediaries are therefore in most cases not attractive for manufacturers. The described restriction is also reflected in the fact that intermediaries have not reached the momentum that tech companies often reach. They exist, but remain at the same level of relevance. Among intermediaries, the US company InnoCentive, Inc. is the most renowned and oldest one. This company has gone rather quiet recently. As a

[2] https://www.innocentive.com

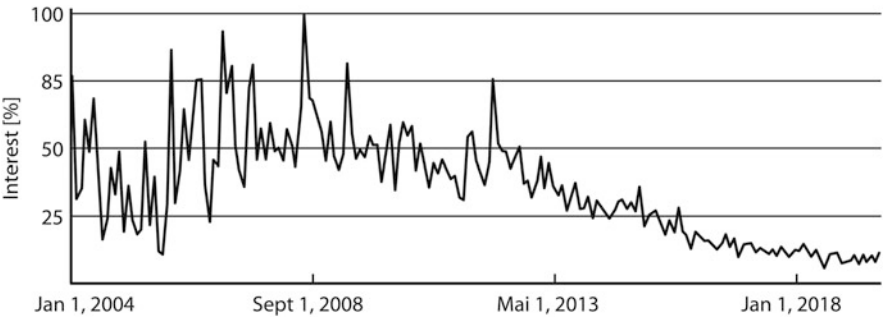

Fig. 5 Interest over time (indexed to 100%) for the term "InnoCentive," 2004–2019, worldwide. Data from Google Trends. Source: authors

confirmation, Fig. 5 shows the development of interest in the term "InnoCentive" in Google Trends.[3]

According to Google Trends, interest declined from 2012 onwards and has been stable since mid-2018. This can be an indication that the need for Level 3 solutions is increasing and that Innocentive is not addressing this need. Other intermediaries show similar trends.

7 Barriers in Web Search

Identifying suitable partners is an important detail in the process of establishing innovation networks, as clearly demonstrated in Sect. 6 in this chapter. In regional projects where personal relationships exist, such partners are quickly found. The same applies to superregional projects if there is an existing community. The situation is different if connections in distant, or even intercontinental contexts are desired, where there is no community. Challenges in this respect can also arise in the establishment of cross-industry innovation networks, or in the international identification of targets for mergers and acquisitions. In all these cases, it is a common method to search for information on the World Wide Web using search engines.

The notion is that with such search engines, it is very easy to find all the information people need. However, this idea is wrong, as the next two sections will show. The danger is that most Internet users are not even aware of these limitations.

[3] https://trends.google.com/trends/explore?date=all&q=Innocentive

7.1 The Language Barrier Web

Searching the Internet with search engines like Google, Baidu, Bing, Yahoo, or Yandex is a daily task for many people worldwide. The functionality of these services is well known to every Web user. Less conspicuous in everyday use is the fact that searching for content in foreign languages is practically possible only if the target language is known and the user has at least a basic command of that language. If the target language of searched information is unknown, the user is lost in the Internet, although the content they are looking for may exist. The part of the Web that lies behind this barrier is what Karl H. Ohlberg calls the *Language Barrier Web* (see Fig. 6). By the way, the limitation of this barrier did not play a role in the early days of the Internet because quite simply all content was in English. Later, other languages were quickly added, such as German, French, Japanese, and Spanish, but the situation stayed manageable. In recent years, however, when the Internet has gained increasing global momentum, the situation has changed. Especially the rise

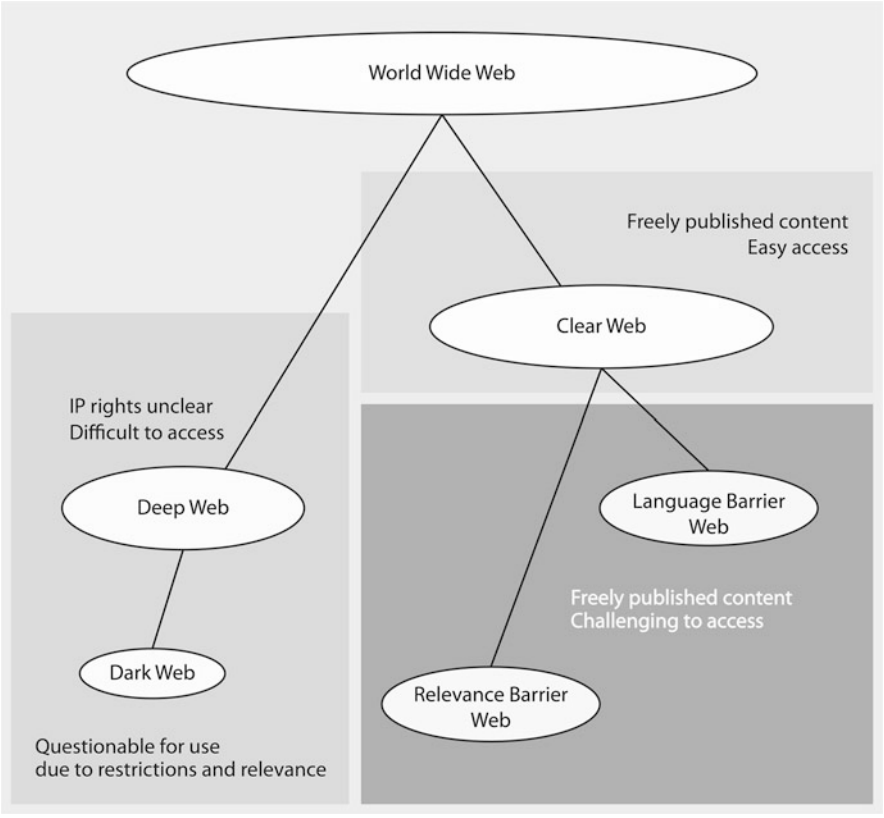

Fig. 6 The Language and Relevance Barrier Web. Source: authors

of Asia has led to a lot of Asian content, mostly Chinese, but also other languages are spreading more and more on the Web, e.g., Eastern European languages. As users prefer their native language when using the Internet and adding content, the *Language Barrier Web* is increasing in size.

For the sake of completeness, it should be mentioned that an attempt was made a long time ago to remove this barrier with the so-called *Cross-Language Information Retrieval* (CLIR) approach. For example, in 2007, Google started their CLIR project with a *Translated Foreign Pages* search filter for their search engine (Sterling 2007). This option was removed in 2013 due to lack of usage (Schwartz 2013). At that time, Google had not given any specific reasons for the decision. However, it can be assumed that it was because Google's primary target group are consumers, who are often interested in comments from other users that are difficult to interpret from a linguistic point of view (e.g., colloquial language, ambiguities, and irony). Since there is machine translation behind the CLIR approach, it is clear that consumers get frustrated. For example, if a colloquial restaurant rating is found in a foreign city, it will be more a source of amusement than information.

A general search engine with CLIR does not yet exist, probably for the reasons mentioned above. However, CLIR is realized in very specialized applications, such as the information retrieval project for international law enforcement of MIT Lincoln Laboratory's Human Language Technology (HLT) Group (Tantillo and Ryan 2016).

7.2 The Relevance Barrier Web

There is, however, another barrier to the Web, which arises from the preference for relevant content. Try to search for a technical term in your language—e.g., a specific technology in which manufacturers of very different sizes and industries are involved—with the special feature of obtaining results from companies with less than 100 employees or any specific turnover threshold. This will be challenging, because the variety of company sizes and industries makes the search very difficult. Search engines will now focus on the content of the major, mostly large, manufacturers and display the first pages with dozens of results of big companies. Of course, suitable results might be displayed—possibly far down the list of results—but their identification is a time-consuming process, not acceptable for most business people. Actually, there should be a filter for this task. Such filters, which are common in business databases, do not exist in search engines. The relevance criteria are defined by the makers of the search engines. To put it simply, relevance is based on frequent retrieval, and is not at all related to the technical correctness of content.

The part of the Web that lies behind the described barrier is what Karl H. Ohlberg calls the *Relevance Barrier Web* (see Fig. 6). Depending on the task at hand, every Web user can experience the hurdle of the *Relevance Barrier Web*, even in the context of a single language.

In summary, it can be said that until about 2010 the ecosphere on the Internet was still quite manageable in terms of languages as well as quantity of content. Then the

amount of content and the variety of languages exploded. The notion of getting an overview of what is happening all over the world with a quick and simple search on the Internet has become obsolete. By the way, the terms *Language Barrier Web* and *Relevance Barrier Web* are apparently not yet in use.

8 Proposal-Based Innovation (PBI)

The enormous importance and associated potential of synthesis (recombination of previously unconnected concepts) in the innovation process has already been demonstrated in Sect. 4.6.4. In nonmanufacturing industries, this advantage is already being exploited more and more by a lively, digitized, disruptive, and globalized startup culture. OI and digitization are the big drivers, based on tech and software industries. This globalized spirit of intensive startup cultures does not exist in manufacturing industries. In our current world, manufacturing industries can actually be described as traditionally innovating and digitally forgotten (see Sect. 5). However, because of strong changes (Sect. 4) the pressure for new approaches in innovation is high.

This is the situation that gave rise to the idea of a completely new innovation methodology based on global insight from which the concept of *Proposal-based Innovation* (PBI) was developed.

8.1 PBI in the Global Environment

The basic idea of PBI is that it is not crucial for synthesis to know how a technology or innovation is designed (e.g., how a product or process works), but only that this technology or innovation exists somewhere in the world or is planned or being developed. This decoupling from the Intellectual Property (IP) and the reduction to the WHAT level—in contrast to the HOW-it-works-level of IP—enables the provision of an enormous amount of facts (description of innovations) and related metadata, which are completely legal and freely available on the Internet, but not accessible by search engines in a worldwide context. These current limitations of global searching also make clear that this approach brings the greatest benefit in a global environment because companies are quite well informed about developments in their own local environment. Although supra-regional information is usually also available, at least in the own industry, the amount of it is usually limited. Another important point is that insight into relevant information on cross-industry innovation is often completely lacking.

PBI is a data-driven approach to help open up the innovation process through detailed insight into innovation activities and a suitable configuration of selected partners on a global level. Ecosystems and communities as in OI are not required. The approach is designed to initialize and leverage the possibilities of collaboration over very long distances and between technically distant industries. The concept is aligned with the new form of globalization that is now emerging (see Sect. 4.6.1).

Table 1 Summary of the number of manufacturers on three continents (30 largest economies)

	Number of manufacturers	
	<500 employees	>500 employees
Americas	1,238,000	7000
Asia	906,000	33,000
EMEA	718,000	11,000
Total	2,862,000	51,000

PBI is designed to achieve maximum compatibility with corporate structures in their current, predominant *Closed Innovation*, in all manufacturing sectors. This approach can also be applied to nonmanufacturing sectors. The focus in this chapter is on manufacturing industries, because from today's point of view these kinds of industries can benefit the most. It is important to look at this concept in its entirety, considering all the aspects described in this chapter, to understand it well.

The aim of PBI is to fundamentally change innovation in manufacturing industries. Many manufacturers want to improve their innovation process, become more open, use more external ideas, knowledge, and conceptual resources for faster innovation, make unused knowledge available to others, collaborate, and become more digital, not only in products, but also in business models. They want to understand which innovation activities are currently emerging in other regions and industries worldwide and how they can benefit from them.

The potential of companies for which this approach may be of interest was also calculated. Table 1 summarizes the number of manufacturers in the three main regions. It is clear that not all manufacturers today are ready to reimagine their innovation process. Nor is it the majority. Analyses and discussions have shown that 25% of the manufacturers with above 2000 employees, 15% of companies with between 500 and 2000 employees, and 2% of companies with under 500 people are open to a new approach. In total, more than 50,000 companies want to break new ground today. For these companies, existing innovation networks do not reach far enough and are therefore inadequate. They want to go further.

With regard to research into innovation activity data, it should be noted that a large proportion of manufacturers are active in the market with several innovative products. This means that a worldwide potential of several million freely available facts and metadata on innovation activities can be identified.

8.2 The Concept of PBI

In the following, the core concept of PBI is described, referring to the structure outlined in Fig. 7. Further details are available on request from Karl H. Ohlberg.

1. *Languages and regions*

   The initial step is to define languages (L_1, L_2, L_3, ... L_n) and also regions in which information on innovation activities is to be sought. Since a fully automated search for information (crawling and scraping) will not be sufficient

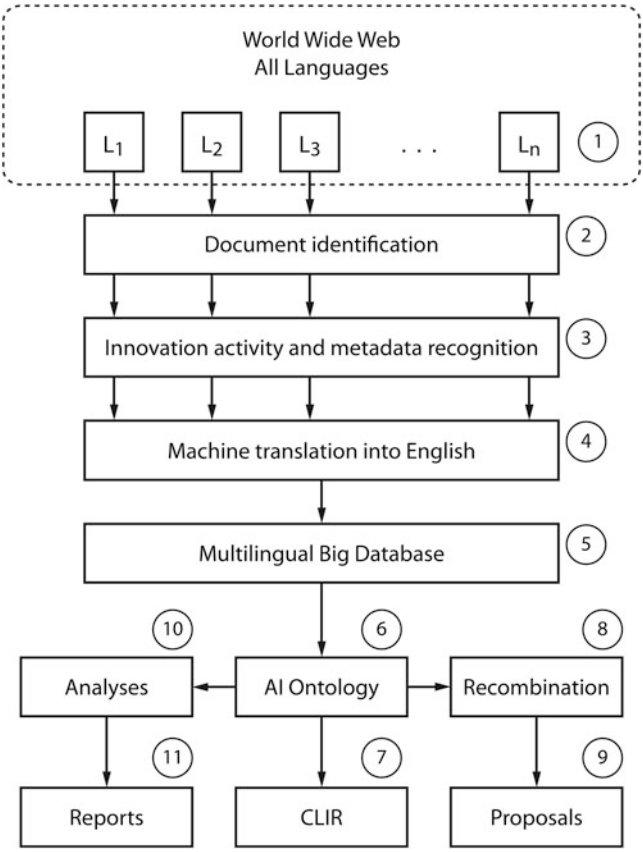

Fig. 7 The structure of proposal-based innovation (PBI). Source: authors

to get the required results, human research teams must be assembled in the defined regions.

2. *Document identification*

   Next, all publicly available, legally published information on innovation activities in manufacturing industries must be identified in all defined languages and regions. These source documents could be, for example, articles in professional journals, websites, brochures, annual reports, patents, research documents, and information from trade fairs.

3. *Innovation activity and metadata recognition*

   This step is about automatically recognizing or researching the core of the innovation activities (descriptive text in the source language) and the corresponding metadata from the previously identified documents. Metadata can include:

   (a) Web link of the source document
   (b) Language

(c) Innovator (e.g. company name)
(d) Website address
(e) Country and province
(f) Legal entity type
(g) Employee count
(h) Turnover
(i) Industry and sub-industry
(j) Materials and processes
(k) Targeted industry

4. *Machine translation into English*

   The innovation activity description in the source language must be translated into English using machine translation. Since the descriptions are high-quality editorial texts, a good result can be expected from the translation, although a quality check is mandatory. All other metadata must also be standardized into English.

5. *Multilingual Big Database*

   All collected data on the different decentralized (regional) activities must be summarized in a Big Database, which serves as a starting point for subsequent data processing. The innovation activity description is stored in both the source language and English machine translation.

6. *Artificial Intelligence (AI) ontology*

   An ontology of all English language data must be created. An ontology is a formally ordered representation of a set of conceptualities and the relationships between them in a given subject area. In principle, we are already familiar with this same procedure in search engines in the so-called semantic web. Thus, terms and correlations that are similar or related can also be found in search processes.

7. *Cross-Language Information Retrieval (CLIR)*

   By using the English language representations of all information in the ontology, a true CLIR is possible, also known as a cross-language search. However, this search will not only refer to terms and contexts, but will also offer further possibilities for much more targeted search processes on the basis of the numerous metadata (see above). Examples are filters for certain subsectors, company sizes or geographies.

8. *Recombination*

   The ontology must be extended so that innovation activities are recognized by their meaning. A specialized AI will then recognize patterns in innovation activities. This procedure will serve in particular to recognize the same or similar innovation processes in different industries and at different locations (Similarities). In addition, another specialized AI will identify innovation activities that complement other activities (Complementarities).

9. *Proposals*

   The procedure described above makes it possible to automatically generate concrete innovation proposals. These proposals represent a global as well as a cross-industry view that is not yet available anywhere else. The importance

of these proposals becomes clear when we look at the changes, challenges, and opportunities in all the many manufacturing industries (Sect. 4). This high value creation leads Karl H. Ohlberg to call this new innovation methodology Proposal-based Innovation (PBI).

10. *Analysis*
    Numerous analyses of previously unavailable facts will be possible based on the data obtained.
11. *Reports*
    Industry- and user-specific reports will be created. Specialized reports can also be used to mitigate risks in the R&D and innovation process.

8.3 Artificial Intelligence (AI)

The foundation of the presented concept will be intensive research on innovation activities, database applications, and machine translation, all state-of-the-art technologies. However, in order to understand the innovation activities described in the previous section and to automatically generate these associated proposals, a sophisticated artificial intelligence (AI) technology will be required. In the following, Jose L. Salmeron, summarizes the basics of such an application.

In general, unstructured data accounts for about 80% of the data that companies process on a daily basis. Structured data has clearly defined data types whose patterns make it very easy to search. In contrast, unstructured data is usually not as easy to search. This data includes data formats such as documents, websites, digital news, audio, video, and social media postings.

For the detection of innovation activities, the field of AI known as Natural Language Processing (NLP) will be needed. NLP must extract the meaning from the text (unstructured data). This process includes formal grammars that identify the relationship between text units—especially between language elements such as nouns, pronouns, adjectives, and verbs—that mainly address syntax.

Grammars can be extended to address natural language semantics by greatly expanding subcategorization, with additional rules/constraints (e.g., "innovation" only applies to new products or services). It should be noted that the rules can become very numerous and can often interact unpredictably with more common ambiguous parses (multiple interpretations of a word sequence are likely).

In addition, handwritten rules will poorly handle *ungrammatical* spoken prose and the highly telegraphic prose of real-world notes, even though such prose is human comprehensible.

The NLP approach for the PBI concept will be as follows (see Fig. 8):

1. Gathering: The data must be gathered from the Web using Web scraping or human research.
2. Understanding: NLP techniques will be needed to understand the text gathered.

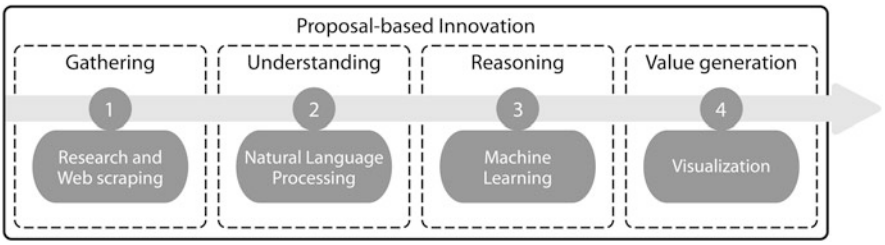

Fig. 8 The basic NLP approach for the PBI concept. Source: authors

3. Reasoning: Using machine learning tools, the system must carry out a reasoning process about innovation activity detection.
4. Value generation: Visualization tools will increase the value of the previous phases.

8.4 The Vision

The vision of the concept of PBI described in this chapter is to help create a platform for the largest source of information on all innovation activities in all manufacturing sectors (Sect. 3) in the world. It would be a platform where companies can find everything they are looking for in terms of analysis to mitigate risk, as well as in the field of innovation, and where they can innovate faster and more efficiently with concrete proposals for using external ideas, knowledge, and concepts.

In short, the new platform will enable companies to:

- Identify the invisible
- Reach out to the unreachable
- Collaborate with the best suitable

9 Conclusion and a Special Concern

Proposal-based Innovation (PBI) is a large-scale, data-driven approach that can help all manufacturing industries (Sect. 3) meet today's and tomorrow's business challenges and that can also help solve fundamental issues affecting humankind.

Innovation in manufacturing industries today is still carried out very much in-house, especially in medium-sized and small companies. In this definition, medium-sized companies may well have several thousand or even some tens of thousands of employees. Only the largest enterprises have globalized innovation processes, mostly limited to their own industries. And even in these companies, the potential benefits of external concepts and cross-industry innovation are usually only exploited to a limited extent.

In this historically evolved environment, the pressure on manufacturing companies has increased enormously, especially in recent times. Today, numerous factors are coming together at the same time. Moreover, it is increasingly external factors that are forcing practically all manufacturers in all 20 main manufacturing sectors (Sect. 3) to become much more agile and efficient in the innovation process. These external factors are essentially the following (Sect. 4.6):

- Changes in the nature of globalization
- Changes in global power and leadership structures
- Increasingly complex products and digitization
- Demand for decreased time-to-market
- Challenges of constantly increasing R&D costs
- Massive changes in social structures, as well as climate change

The methodology of Open Innovation (OI), which has enabled many nonmanufacturing industries (e.g., e-commerce, social media, fintech, and the sharing economy) to develop companies and sectors enormously by opening up innovation processes, has been largely rejected by manufacturing sectors (Sect. 6).

In order to resolve this dilemma in the innovation process of manufacturing industries, a new methodology is proposed which Karl H. Ohlberg calls *Proposal-based Innovation* (PBI). This has emerged from practical discussions with hundreds of manufacturing companies as well as academia and other stakeholders. In contrast to the community approach of OI from the time of the empty Web (Sect. 6), PBI is a data-driven approach based on an enormous amount of information available worldwide on innovation activities, but unfortunately hidden for the most part due to deficits in search engines and language barriers (Sect. 7). Another important point concerning PBI is the decoupling of the WHAT level (what is done technologically) from the HOW level (how something is done technologically). The WHAT level is the decisive factor in developing a rapid and efficient innovation process that integrates external concepts. The intellectual property (the HOW level) can then be licensed.

The last and important component of PBI is the automatic generation of concrete proposals for innovation, based on innovation activity data using artificial intelligence (AI); hence the term PBI. Concrete proposals have always been among the main demands of manufacturing companies, as related conversations have shown. This is because a search for options, no matter how convenient the search, is always a creative process for which there is often insufficient time in everyday life. To reject a proposal quickly, on the other hand, is much faster and more efficient. With a large number of automatically generated proposals, it is easy to find some that arouse interest.

Finally, yet importantly, a special concern of this chapter is to stimulate discussion on the presented concept at as international a level as possible. Karl Ohlberg has developed additional detailed concepts and plans regarding a new platform for putting the concept of PBI into praxis. Readers are invited to contact Karl Ohlberg

for more information on the underlying project. He is particularly happy to receive requests from those who are geographically and/or culturally distant.

References

Aced, C. (2013, July). *Web 2.0: The origin of the word that has changed the way we understand public relations*. Retrieved December 30, 2019, from https://www.researchgate.net/publication/266672416_Web_20_the_origin_of_the_word_that_has_changed_the_way_we_understand_public_relations

Baldwin, R. (2018, December 22). *If this is Globalization 4.0, what were the other three?* Retrieved December 30, 2019, from https://www.weforum.org/agenda/2018/12/if-this-is-globalization-4-0-what-were-the-other-three/

Bergen, M., & Brustein, J. (2018, December 5). *Almost every electric scooter in the world comes from this Chinese company*. Retrieved December 30, 2019, from https://www.bloomberg.com/news/features/2018-12-05/almost-every-electric-scooter-comes-from-this-chinese-company

Bloom, N., Jones, C. I., Reenen, J. V., & Webb, M. (2017, September 8). *Are ideas getting harder to find?* Retrieved December 30, 2019, from https://www.nber.org/papers/w23782

Bundesministerium für Bildung und Forschung (BMBF, Federal Ministry of Education and Research, a cabinet-level ministry of Germany). (n.d.). *Vernetzung weltweit*. Retrieved December 30, 2019, from https://www.bmbf.de/de/vernetzung-weltweit-268.html

Bundesministerium für Bildung und Forschung (BMBF). (Ed.). (2015, December). *The China strategy 2015-2020*. Retrieved December 30, 2019, from https://www.bmbf.de/en/the-china-strategy-2015-2020-2345.html

Burnett, B. (2009). *Building new knowledge and the role of synthesis in innovation*. Retrieved December 30, 2019, from https://de.scribd.com/document/47345390/Bill-Burnett-2009Building-New-Knowledge-and-the-Role-of-Synthesis-in-Innovation-by-Bill-Burnett

Chesbrough, H. W. (2003). *Open innovation: The new imperative for creating and profiting from technology*. Boston: Harvard Business School Press.

Collins, A. (2019, January 15). *The global risks report 2019*. Retrieved December 30, 2019, from https://www.weforum.org/reports/the-global-risks-report-2019

Crowley, M., & Sanger, D. E. (2019, September 24). *Trump celebrates nationalism in U.N. speech and plays down Iran crisis*. Retrieved December 30, 2019, from https://www.nytimes.com/2019/09/24/us/politics/trump-nationalism-united-nations.html

Fayd'herbe, N. H. (2019, January 16). *A multipolar world brings back the national champions*. Retrieved December 30, 2019, from https://www.weforum.org/agenda/2019/01/a-multipolar-world-brings-back-the-national-champions/.

Frietsch, R., Beckert, B., Daimer, S., Lerch, C., Meyer, N., Neuhäusler, P., et al. (2016, November 15). *Die Elektroindustrie als Leitbranche der Digitalisierung – Innovationsstudie*. Retrieved December 30, 2019, from https://www.zvei.org/presse-medien/publikationen/die-elektroindustrie-als-leitbranche-der-digitalisierung-innovationsstudie/

IndustryWeek Custom Research and Kronos Incorporated. (2016). *The future of manufacturing: 2020 and beyond*. Retrieved December 30, 2019, from https://www.kronos.com/resources/future-manufacturing-2020-and-beyond

Khanna, P. (2016). *Connectography: Mapping the future of global civilization*. New York: Random House.

Khanna, P. (2019). *The future is Asian: Commerce, conflict, and culture in the 21st century*. New York: Simon and Schuster.

Kharas, H., & Hamel, K. (2018, September 27). *A global tipping point: Half the world is now middle class or wealthier*. Retrieved December 30, 2019, from https://www.brookings.edu/blog/future-development/2018/09/27/a-global-tipping-point-half-the-world-is-now-middle-class-or-wealthier/

Lukas, V., Walker, A. S., KukkoDr, A., NeishDr, C., OsinskiDr, G., Zanetti, M., et al. (2018, March). *Lidar Magazine*, vol 8(2). Retrieved December 30, 2019, from https://lidarmag.com/issue/volume-08-issue-02/

Meyer, A. (2018, December 13). *50 world-changing startups to watch in 2019*. Retrieved December 30, 2019, from https://www.inc.com/anna-meyer/top-emerging-companies-2018-global-affordability.html

OECD. (2019, February 12). Global material resources outlook to 2060: Economic drivers and environmental consequences. OECD Publishing, Paris. Retrieved December 30, 2019, from doi:https://doi.org/10.1787/9789264307452-en

Pande, S., & Anil, A. M. (2018, November 19). South Asia can become an innovation hub. Here's how. Retrieved December 30, 2019, from https://www.weforum.org/agenda/2018/11/here-s-how-south-asia-can-harness-the-power-of-emerging-technologies/

Pisano, G. P., & Shih, W. C. (2012). *Producing prosperity: Why America needs a manufacturing renaissance*. Boston, MA: Harvard Business Press.

Rogers, K., & Sanger, D. E. (2019, February 16). *Among European allies, Americans offer competing visions*. Retrieved December 30, 2019, from https://www.nytimes.com/2019/02/16/world/europe/mike-pence-joe-biden.html

Schwartz, B. (2013, May 20). *Google drops "translated foreign pages" search option due to lack of use*. Retrieved December 30, 2019, from https://searchengineland.com/google-drops-translated-foreign-pages-search-option-due-to-lack-of-use-160157

Shih, G. (2015, April 15). *Xiaomi-backed Chinese firm acquires iconic scooter maker Segway*. Retrieved December 30, 2019, from https://www.reuters.com/article/us-ninebot-xiaomi-investment/xiaomi-backed-chinese-firm-acquires-iconic-scooter-maker-segway-idUSKBN0N60GN20150415

Sims, T. (2019, September 25). *Germany's continental to cut jobs and close plants as auto sector slows*. Retrieved December 30, 2019, from https://www.reuters.com/article/us-continental-strategy/germanys-continental-to-cut-jobs-and-close-plants-as-auto-sector-slows-idUSKBN1WA1F6

Sloane, P. (2015, March). *What is stopping open innovation?* Retrieved December 30, 2019, from https://www.destination-innovation.com/stopping-open-innovation/

Sorkin, A. R. (2018, January 15). *BlackRock's message: Contribute to society, or risk losing our support*. Retrieved December 30, 2019, from https://www.nytimes.com/2018/01/15/business/dealbook/blackrock-laurence-fink-letter.html

Sterling, G. (2007, May 24). *Google launches 'cross-language information retrieval (clir)'*. Retrieved December 30, 2019, from https://searchengineland.com/google-launches-cross-language-information-retrieval-clir-11296

Tantillo, A., & Ryan, D. (2016, May 27). *Finding relevant data in a sea of languages*. Retrieved December 30, 2019, from http://news.mit.edu/2016/finding-relevant-foreign-language-data-0527

United Nations. (2019). *Ageing*. Retrieved December 30, 2019, from https://www.un.org/en/sections/issues-depth/ageing/

Vrontis, D. (2013). *Innovative business practices prevailing a turbulent era*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Technologies and Innovations for the Plastics Industry: Polymer 2030

Michael Krause

Abstract

The plastics industry is changing. Not only are global influences leading to greater competition; a number of sectors that rely heavily on the plastics industry are also undergoing radical transformations. Consider the automotive industry: in this sector, a radical process of transformation and change is underway as new, more environmentally friendly drive systems are being developed—in particular, the move to electric vehicles. The CEO of Volkswagen AG has set out a clear position on electric cars, thereby triggering a dramatic upheaval. This issue has been given additional impetus through to the "Fridays for Future" movement and the diesel emissions scandal.

1 Technologies and Innovations for the Plastics Industry: Polymer 2030: Fit for the Future Thanks to Innovations and Technology

The plastics industry is changing. Not only are global influences leading to greater competition, a number of sectors that rely heavily on the plastics industry are also undergoing radical transformations. Consider the automotive industry: in this sector, a radical process of transformation and change is underway as new, more environmentally friendly drive systems are being developed—in particular, the move to electric vehicles. The CEO of Volkswagen AG has set out a clear position on electric cars, thereby triggering a dramatic upheaval (Tagesspiegel 2019). This issue

M. Krause (✉)
KIMW-Qualifizierungs gGmbH and KIMW-Forschungs gGmbH, Lüdenscheid, Germany
e-mail: krause@kunststoff-institut.de

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_15

has been given additional impetus through to the "Fridays for Future" movement and the diesel emissions scandal.

The plastics industry supplies components and products for vehicle interiors and exteriors: oil sumps, tanks, panels, etc. The recent developments in the automotive industry are giving rise to drastic changes for the suppliers and plastics manufacturers, with entirely new requirements. This is a result of changes to the components and products in the vehicle—for example, the battery for an electric car needs a new kind of casing, and its engine compartment differs significantly from that of a combustion engine.

The automotive industry is only one of the markets important to the plastics industry; the packaging industry is another. Debates on ocean pollution, as well as the "Fridays for Future" movement, are likewise bringing this sector under pressure. In the public perception, packaging is often what comes to mind when one thinks of the plastics industry, and this strong association means that concerns about packaging create a negative image for the entire sector. In reality, plastics are very diverse and are also used as technical components where they can be highly durable, e.g. as described above for vehicle panels. Polymers are also elements of fibre-composite materials, which have considerable potential for lightweight designs. These save on resources, and thus yield environmentally friendly effects. Clearly, taking a one-sided view and only considering plastic in terms of its negative effects is somewhat short sighted.

The packaging industry is facing additional pressure as a result of new legislation such as new packaging law, or regulations for waste sacks in corporate enterprises. Given the perceptible transformation in mindset that is underway in our society, businesses in this sector will find innovation and sustainable business models essential to their future survival.

However, there are always obstacles to innovation, particularly for small- and medium-sized enterprises: they lack the necessary human and financial resources; the risks of creating something new and establishing it on the market are too high; often, they lack the necessary infrastructure (Industrie- und Handelskammer Nordrhein-Westfalen 2014). One potential key to resolving these problems can be collaboration—with other companies, start-ups or research institutes—leading to synergies and interdisciplinary competences that can be exploited. The plastics industry can also find opportunities in other sectors—for example, medical technology: as well as routine products such as cannulas and syringes, new market opportunities might be found in, e.g. 3D printing of finished products such as individual prostheses.

The negative trends described above are reflected in the latest economic data for businesses in the plastics sector: the business climate index plunged dramatically in the second quarter of 2019 to −12.4%, while capacity utilisation only reached 77.4%. The economic figures for the plastics industry are reinforced by a number of expert statements and studies that have also indicated a recession in the sector. Figure 2 shows how the industry depends on other sectors (Manager Magazin 2019).

For these reasons, it is becoming increasingly important to understand new customer requirements early on, identify new sectors, megatrends and technologies, and

to recognise where revenues are being lost. This means that it becomes important to develop new products, new services, new moulded parts and new business models, alongside establishing innovative technologies. However, businesses face a number of challenges when it comes to introducing these changes. As described above, there is a lack of financial and human resources, infrastructures are inadequate, and there is often insufficient initiative for developing the new products.

The next sections describe the structure of the plastics sector, then outline megatrends and new models relevant to the industry, before going on to give some best practice examples and finally derive recommendations for businesses.

2 Structure of the Plastics Sector

The Kunststoff-Institut Lüdenscheid is a plastics institute boasting a network which includes 350 businesses, from small- and medium-sized enterprises to large corporations. This network includes companies throughout the plastics industry value chain and thus models the structure of the sector. Plastics manufacturers and polymer compounding companies are typically large international corporations that produce polymers and enrich them with filler materials. These polymers, e.g. PET, are then used by the plastics manufacturers, who are themselves implementing the requirements of their clients, generally on a business-to-business (B2B) basis. For example, a large manufacturer (OEM) might order specific semi-finished products (e.g. rods, sheets or pipes) or finished products such as windows, toys or garden furniture. Many plastics manufacturers produce semi-finished products and moulded parts/products to order ("service providers"), thus effectively putting their machine capacities, expertise, and established processes and quality standards at the disposal of their customers. Other enterprises manufacture their own products and then market them B2B to various manufacturers.

The plastics manufacturers use various techniques to process polymers. One such is injection moulding. This is a primary shaping process, i.e. it uses a formless material such as powder to create solid products (Maschinenbauwissen.de 2019). In this case, the granules are brought into a fluid state and guided through several steps into a tool. The various processes ultimately generate a solid moulded part or product. The advantages of injection moulding include the direct route to a finished moulded part/product, with only minor post-treatment generally necessary. In addition, it is an automated process, meaning that large batch sizes can be manufactured.

Machine construction companies who produce injection moulding machines, such as Engel, Fanuc, or Arburg, can therefore also be considered to be market players in the plastics sector; similarly, so can companies who manufacture accessories relating to the injection moulding process. For example, we can include moulded parts dealers and tool manufacturers, who specialise in building new tools for developing new products/moulded parts. They test construction options and check whether they meet customer requirements. Vendors for peripheral

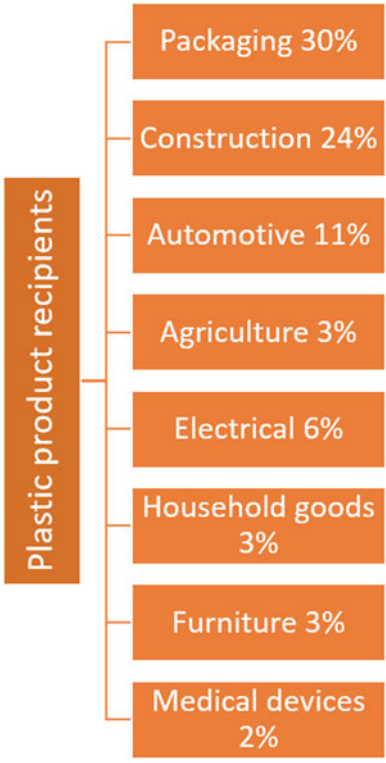

Fig. 1 Plastic product recipients. Data from Gesamtverband Kunststoffverarbeitende Industrie (2019). Source: author

products, tempering, maintenance, surface finishing, etc. must also be taken into consideration.

The diversity of these companies and their various functions indicates the high potential of the sector, which has annual revenues of around 90 billion euros (Gehalt.de 2019). A study from the German plastics industry's umbrella organisation, the GKV, emphasises this, with a breakdown of the diverse industries where the plastic semi-finished and finished products are used: approximately 30% end up in the packaging sector, 24% in the construction sector, 11% in the automotive sector, plus around 6% in the electrical sector, 4% in agriculture, 3% for household goods, 3% for furniture and 2% in medical devices (Fig. 1).

3 Megatrends and New Business Models for the Plastics Industry

Small- and medium-sized businesses find it particularly difficult to tackle new ideas, business models and innovations. Requirements are becoming steadily more complex, due to the increase in digitalisation and individual customer requirements. Large enterprises such as Nokia or Kodak failed after years of success when

their newest innovations did not take the right direction. One way of approaching megatrends, new business models and technologies is described in the St. Gallen business model (Gassmann et al. 2017). Businesses can begin by structuring their existing business model. The first step is to specify the target customers and their requirements. For example, a plastics producer might consider an OEM from the automotive industry. The next feature in the model is the business's value proposition, which needs to take into account the customer's problem and the potential solutions offered by the business. How does my business stand out from the competition? What is our expertise; what are our quality standards? This is how the business sets itself apart and wins customers.

A third feature is resource usage: the processes my business needs to offer a particular option; the expertise and human resources that the business needs; what the current product marketing looks like. The final feature for the business model to define at this stage is the pricing mechanism. This means answering more questions: What is the customer prepared to pay? What costs arise? Are there additional options that could bring in extra revenue—services, for example? (Gassmann et al. 2017).

Once the business has established these four factors, further influencing factors can be identified. These include stakeholders (whether competitors or cooperative partners), but also—crucially—new trends and technologies that could have repercussions for the business and its business models. Adapting new technologies and trends could even lead to new business models being developed. In the next section, we will address a number of megatrends that have the potential to have a significant effect on the evolution of the plastics industry.

The trend research institute Zukunftsinstitut defines megatrends on the basis of various studies. These megatrends include connectivity, new working models, individualisation, mobility, urbanisation, the "Silver Society", neo-ecology and health. The individual megatrends can be further divided into subtrends (Fig. 2). For example, the megatrend "Connectivity" can be broken into a number of topics such as "Platform economy", "Artificial Intelligence" and the "Internet of Things".

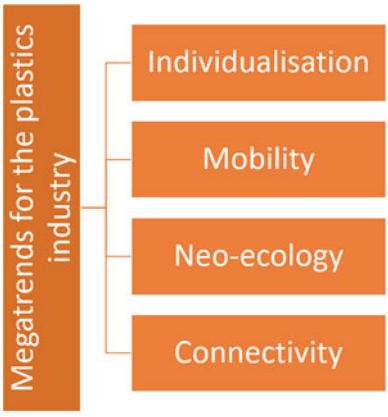

Fig. 2 Megatrends for the plastics industry. Data from Zukunftsinstitut (2019). Source: author

Some of these megatrends have particularly important roles to play for the plastics industry. It is important to analyse these, adapt them and draw appropriate conclusions for future business models.

Connectivity is a key megatrend with particular importance for the plastics industry: new platforms, for example, could create opportunities to link up different target groups, products and demand. Such platforms—which already exist, at least to a limited extent—might focus on sales of injection moulding machines, accessories or plastics. Other possibilities include offering machine capacity, or bringing together project partners who want to collaborate on a new research or innovation project, thereby exploiting synergies and saving capacity. Other trends of interest to the plastics industry are artificial intelligence and the IoT, which could be used, e.g. to evaluate parameters from injection moulding machines and thus achieve early identification of process disruptions, to create feedback links and transmit error reports to the machine operator. Virtual reality and augmented reality, meanwhile, could be used to help visualise injection moulding processes and workflows, for example when training new entrants. Another connectivity subtrend focuses on the use of social media channels. These are currently still only given rudimentary attention in the plastics industry, which is largely focused on B2B relationships where access to decision makers is still routed via analogue channels.

Along with connectivity, individualisation is another megatrend with potentially very interesting consequences for the plastics industry. Demand for individual products and services is growing. This means a move away from mass production, towards batch sizes of 1. There is a wide range of technologies for surface treatments—digital printing, for example, or in-mould decoration—and these can be used to implement individual solutions, such as adding customised motifs to products, and therefore to print or manufacture small batch sizes, whether for a vehicle panel or an individualised glasses case (Fig. 3).

Additive manufacturing also gives rise to other possibilities: manufacturing initial prototypes for new plastic components/products, and manufacturing small quantities with 3D printers, which could have particular relevance for, e.g. replacement parts or individualised products such as prostheses.

Mobility is a third megatrend with significant impact value for the plastics industry. As described above, 11% of the finished and semi-finished plastic products are produced for the automotive sector. The German Federal Government has launched various initiatives to promote electric vehicles, such as the “Umweltprämie”, a “green bonus” to promote sales of new vehicles with electric drives (Bundesamt für Wirtschaft und Ausfuhrkontrolle 2019). So far, the country is falling a long way short of the Federal Government’s stated goal of having a million electric cars registered by 2020 (Bundesministerium für Wirtschaft und Energie 2019). In 2018, around 68,000 new electric cars were registered in Germany (Smarterfahren.de 2019). Realistically, the “green bonus” cannot balance out a lack of acceptance. Consumers have reservations about the limited range for the vehicles; the infrastructure networks (charging points) are not sufficiently dense; it takes too long to charge the battery. However, there are indications that the tide may be beginning to turn. For example, the public pressure through the “Fridays for

Fig. 3 Digital printing example. Source: KIMW-Forschungs gGmbH

Future" movement, the recent statements from the CEO of VW. These trends go hand-in-hand with developments in self-driving cars. As these segments change, the plastic industry can expect significant repercussions in terms of plastic products for interiors and exteriors. Furthermore, a change in attitude in society, combined with technical influences such as self-driving cars, may lower the overall demand for vehicles, with corresponding knock-on effects for suppliers to the automotive industry.

Another megatrend that is important to the plastics industry is neo-ecology: the sustainability issues that are currently centre stage for the ongoing social transformation. The plastics industry is attempting to counteract its current image problems, stemming from hotly debated topics such as ocean pollution, with innovations in other sectors—recycling, biodegradable plastics and resource efficiency. The main theme chosen for the major plastics trade fair "K-Messe" is further evidence of this: Recycling. This covers questions such as: How can we increase the proportion of recycled manufactured and raw materials (approx. 46.7% in 2017)? How can we optimise processes such as collection, crushing, sorting, etc.?

In addition to the megatrends outlined above, a number of other trends also have important repercussions for the plastics industry. For example, globalisation and the associated increase in competition that businesses face—which, however, comes hand-in-hand with opportunities for new markets.

Businesses in the plastics sector must identify individually which trends and technologies are important for them, based on their corporate structures. They can then investigate the possibilities for adapting their business models based on the new trends. For example, introducing additive manufacturing makes it possible to respond to individual customer requirements. It is not enough, however, for businesses to consider the changes to their business models that result from adapting

new trends and technologies; they must also weigh up the opportunities and risks of individual megatrends/subtrends, for example by drawing up a number of scenarios and carrying out scenario analyses.

4 Trends and Technologies: Best Practice Examples for the Plastics Industry

The first step for corporate managers to take is to pay attention to these megatrends and subtrends and consider what the new trends and technologies yield in terms of objectives, new business models and the opportunities and risks for their business. Equally essential, they must define concrete operative objectives and projects to transfer the changes into the corporate culture. The next section presents a number of individual projects that can be categorised under particular megatrends and subtrends, and the related technologies.

4.1 Technologies for Individualisation in the Plastics Industry

Film insert moulding (FIM) is a surface treatment procedure that can be used to individualise components and products. To create decorative surfaces, colours and motifs are applied to plastic film with a screen printing process, and the film is then subjected to a moulding process and cut to the desired dimensions. The resulting decorative film is then placed in a special injection moulding tool where it is injected with plastic. This causes a bond to be created between the film and the component, yielding the finished product. By using different films, it is possible to create different components.

This in-mould decoration process is being used for a research project at the Lüdenscheid plastic institute, in which moulded organic light-emitting diodes (OLEDs) are integrated into a 3D-moulded component. Potential applications include control panels for, say, a vehicle interior. The treated OLEDs and a functional, decorative film are injected with plastic to produce a final component or product that not only has control functions and lamps, but is also decorative. This facilitates manufacture of customised components, while saving resources thanks to the integrated functionality (Fig. 4).

4.2 Resource Efficiency

The tools used for the injection moulding process are highly cost intensive, but very important. When plastics manufacturers produce new components and designs for their customers, they generally need prototypes and, consequently, new tools or mould inserts. Additive manufacturing not only makes it possible to create prototypes, but also facilitates development of new tools. This in turn means that long-term material requirements and processes can be better modelled, for example

Fig. 4 DekOLED (Librizzi 2019). Source: KIMW Forschung gGmbH

by incorporating materials in the process that will be used later on. That is, mould inserts for tools could be produced with a 3D printer and used in the injection moulding tools and machines, so that small product batches can be manufactured under real conditions. The same plastics used for these products will later be used for manufacturing larger batches. Depending on the material for the mould insert, it is likely that only small quantities can be produced: generally, the effects of temperature and other parameters make the mould inserts usable after producing a small batch. The plastic tool inserts are therefore only suitable for producing prototypes. Plastic inserts also result in problems when compared with metal tools, due to the longer cooling times that result from the low thermal conductivity of the plastic inserts. Rapid cooling of the molten plastic is essential to ensure mechanical characteristics and product quality.

A research project is focusing on the topic of improving mechanical product characteristics when using plastic tool inserts. To this end, cooling systems are being designed such that the end products have the same mechanical characteristics as products created using conventional tools. The advantage of using plastic tools manufactured with an additive process comes from the material cost reductions (plastic in place of metal) and the rapid development of the tools. Plastic tools take a few days to produce, while it takes several weeks to manufacture metal tools for prototype production (Clemens 2019).

As described above, these tools are a key element in the injection moulding process. Protective technologies and maintenance intervals for the tools are therefore very important for economic reasons. Steel tools can be given a protective coating using technologies such as chemical vapour deposition (CVD), for example to protect against corrosion—this helps avoid expensive repairs (Formalczyk 2019).

4.3 Digital Transformation

Process reliability is highly important for the plastics industry. Product or component quality must be reliable and wastage kept to a minimum. Various parameters of the injection moulding process affect the quality of the product/component: temperature, pressure, cooling time, flowing time, etc. During another research project, a software platform has been developed that can evaluate process data regardless of the industry sector. The platform is structured with two levels: firstly, historical data and events are evaluated and used to generate predictions. The resulting insights are then transferred into prediction functions; secondly, data is collected and evaluated in real time (Schlutter 2017).

This allows process changes to be made with very little delay, quality issues to be identified in plenty of time, and reduces wastage through rejects.

5 Recommended Approaches for the Plastic Industry

In the light of the current situation in the plastic industry, new business models and innovations are essential, and must be combined with constant adaptation to immediate circumstances. On this basis, we can derive the following recommendations:

1. Businesses must regularly review their business models. This entails reviewing customer segments, industries and customer requirements. It is furthermore important not to simply rely on existing customer relations, which may date back many years, but for the business to stand out from the competition with new technologies and services.
2. Developing new business models does not always require new technology: restructuring business models and adapting new ideas from other industries can also lead to innovations.
3. Collaboration with other partners creates synergies. For example, working with start-ups can introduce fresh interdisciplinary ideas and mindsets to the company.
4. Businesses in the plastics industry must also constantly be ready to address new legal framework conditions, such as the new packaging laws, and develop products that comply with the legal requirements in plenty of time.
5. It is crucial for businesses to make early predictions about developments in their sector. For example, businesses that supply to the automotive sector must make assumptions about what new products might be required based on the new drive techniques.
6. In the light of current trends and the social transformation, environmentally friendly plastic processing is another essential factor.
7. Businesses must build on their structures to develop new strategies and scenarios in response to megatrends.

8. The objective must be to develop new hybrid business models based on subtrends such as artificial intelligence.
9. Businesses must use their expertise and structures to supply new sectors and open up new possibilities. To this end, there is a need to develop standards that can be applied when manufacturing plastic components for new sectors.
10. Plastics manufacturers must evaluate their own role: are you a service provider or a product manufacturer and distributor?
11. It is essential to develop unique selling points, for example, by making use of new technologies.

References

Bundesamt für Wirtschaft und Ausfuhrkontrolle. (2019). *Elektromobilität*. October 1, 2019, from https://www.bafa.de/DE/Energie/Energieeffizienz/Elektromobilitaet/elektromobilitaet_node.html

Bundesministerium für Wirtschaft und Energie. (2019). *Elektromobilität*. Accessed October 1, 2019, from https://www.bmwi.de/Redaktion/DE/Dossier/elektromobilitaet.html

Clemens, N. (2019). *OptiCool, Jahresbericht 2018*. KIMW Forschung gGmbH 3D.

Formalczyk, G. (2019). *Corrosion protective coatings, Jahresbericht 2018*. KIMW Forschung gGmbH.

Gassmann, O., Frankenberger, K., & Csik, M. (2017). *Geschäftsmodelle entwickeln: 55 innovative Konzepte mit dem St. Galler Business Model Navigator*. Carl Hanser Verlag.

Gehalt.de. (2019). *Kunststoffindustrie*. Accessed October 1, 2019, from https://www.gehalt.de/branche/kunststoffindustrie

Gesamtverband Kunststoffverarbeitende Industrie. (2019). *Branchenüberblick*. Accessed October 1, 2019, from http://www.gkv.de/de/branchen/ueberblick.html

Industrie- und Handelskammer Nordrhein-Westfalen. (2014). *Industrie- und Innovationsreport*.

Librizzi, A. (2019) *DekOLED, Jahresbericht 2018*. KIMW Forschung gGmbH.

Manager Magazin. (2019). *Konjunktur-Hiobsbotschaften aus der Industrie*. Accessed October 1, 2019, from https://www.manager-magazin.de/politik/konjunktur/konjunktur-hiobsbotschaften-aus-der-industrie-a-1280824.html

Maschinenbauwissen.de. (2019). *Kunststoffverarbeitung*. Zugegriffen October 1, 2019, from http://www.maschinenbau-wissen.de/skript3/werkstofftechnik/kunststoffe/387-kunststoffverarbeitung

Schlutter, R. (2017). *Entwicklung eines modellbasierten Steuerungskonzeptes zur unternehmensweiten Optimierung datenintensiver Prozesse, TECHNOMER 2017*, Chemnitz.

Smarterfahren.de. (2019). *1 Million E-Autos*. Accessed October 1, 2019, from https://www.smarter-fahren.de/1-million-e-autos/

Tagesspiegel. (2019). *Vergleichbar mit dem ersten Käfer, VW stellt E-Auto Id3 vor*. Accessed October 1, 2019, from https://www.tagesspiegel.de/wirtschaft/vergleichbar-mit-dem-ersten-kaefer-vw-stellt-e-auto-id-3-vor/24997756.html

Zukunftsinstitut. (2019). *Megatrends*. Accessed October 1, 2019, from https://www.zukunftsinstitut.de/dossier/megatrends/

How Do Innovative Business Concepts Enable Investment Opportunities in the Complete Construction Value Chain?

Christoph Jacob

Abstract

Our world is changing and the world population is growing rapidly. For all people we need buildings to live, to shop, to work and to enjoy to feel safe and protected. We need a reliable infrastructure to travel, to connect people with people. One of the consequences of growing population is that major cities are incrementally getting bigger. The way we are building needs to change. Conventional infrastructure and living environments are emerging and creating the need for faster, smarter, and lower cost setups. Concepts for Smart Cities and modular housing are born. In most of the local ecosystems, the construction industry is the key driver of growth, wealth, and security. Compared to other industry sectors, the fragmented traditional building industry participants are decades behind adapting to process and lean factory production driven manufacturers. New digital innovative technologies through the Internet of Things (IoT), artificial intelligence (AI), augmented reality (AR), roboting, automatization, and new materials can help to make the required changes. Pioneers, Innovation, Software and Technology are reinventing our construction world. It is time to be the change to come. It is time to change the way the world is building. There are significant investment opportunities in the complete construction value chain through innovative start-up companies. These entrepreneurs and start-up founders will use technology, data, and engaged people to drive the change in the global construction market.

C. Jacob (✉)
CASEA AG, Neu-Isenburg, Germany
e-mail: Christoph.Jacob@casea.com

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_16

1 Introduction to the Global Construction Market

The construction industry entails architectural—civil—and other engineering services, as well as the physical process of erecting buildings and infrastructure projects, including all products and services required. The construction industry spending worldwide amounted to 11.4 trillion U.S. dollars in 2018. It is also expected that construction expenditures will reach 14 trillion U.S. dollars in 2025 (Wang 2017).

The construction industry is representing more than 13% of the global GDP and is the largest consumer of raw materials and plays a major role in every local economy. There is a close link between spending in construction and growth of the national economy. Project—time—and cost controls are the key performance indicators of success. The industry is very diversified and the production methods differ significantly from the so-called industrial production. The traditional way of building has no future as their productivity increase was only 1% per year over the last 20 years, where other industries improve by more than 3%.

The McKinsey Global Institute highlighted ten root causes of poor productivity classified in three different overarching principles (Barbosa et al. 2017).

Principle 1: External Forces
- Increasing project and site complexities
- Extensive regulation, land fragmentation, and the cyclical nature of public investment
- Informality and potential for corruption distort the market

Principle 2: Industry Dynamics
- Construction is opaque and highly fragmented.
- Contractual structures and incentives are misaligned.
- Bespoke or suboptimal owner requirements.

Principle 3: Firm-Level Operational Factors
- Design processes and investment are inadequate.
- Poor project management and execution basics.
- Insufficiently skilled labor at frontline and supervisory levels.
- Industry underinvests in digitization, innovation, and capital.

The construction industry can get closer to increase their productivity by improving the dynamics and quality of the complete value chain.

2 What Is the Construction Value Chain?

The value chain of the construction industry is composed of distinct stages—raw material, building material, construction design, construction project production, conversion, demolition, and recycling. In order to bring a project to function and

realization significant interactions between each process stakeholder are required. In some assignment there are more than 100 companies involved. As every building and infrastructure project is different, there is a unique project-based nature of relationships along the value chain, resulting in a highly fragmented industry structure.

The McKinsey Global Institute identified potential global productivity improvement of the external forces by reshape regulation and raise transparency. Industry dynamics could improve their cost structure by 17% through better collaboration and contracting as well by improving design, planning, and engineering. The Firm-Level Operational Factors do have the biggest impact on improvements by more than 20%. There are four areas identified:

1. Procurement and supply-chain management
2. On-site execution and management
3. Using digital technology and smarter building material
4. Improving the skills of the workforce

We are not building fast enough, we are not building bright enough, people and technology can help to close the gap. Human beings are the heart of change across cultures, languages, and regions.

3 How Is the World Population Developing?

At about 8000 years before Christ, around 5 million people were living on earth. Until Christ was born, the population growth rate per year was not more than 0.05%. By the eighteenth century the world population reached 1 billion and by the time of the industrial revolution in 1930 the second billion was reached. The growth rate got faster while the time horizon for the next billion was getting shorter. A population of 6 billion was reached in 2000. It is expected to reach 8 billion by 2027. Furthermore, the world population is set to reach 10 billion by 2057 (United Nations 2019). Asia, Africa, and Latin America are the primary drivers of growth.

The constant growth of the world population generates a growing need for housing, dwellings, and infrastructure projects. Next to population growth there are trends that are describing important changes in areas of society.

4 Globally, What Are the Major Impacting Trends on Construction?

1. Wealth will be generated mainly in cities. People are moving to cities for improved personal, social, and economic possibilities.
2. In the last 50 years the household sizes have shown significant declining trends. The United Nations informed that the average household size across the globe ranges from 2 to 9 persons per household. Small average household sizes, of

fewer than three persons per household, were found in most countries of Europe, Japan, and Northern America. Largest household sizes are found in Senegal and Oman, averaging 9 persons. The trend is that household is getting smaller all over the world.

3. Of the world's two billion households, approximately 15%—or 300 million—are one-person households. Among European countries of 40% or more are reported in Denmark, Finland, Germany, and Norway. Moderately high levels of one-person households are also observed in Japan (32%), the United States (28%), and Australia (24%).
4. There is a global shortage of skilled construction workers. The lack of skilled workers has become a growing issue as the ageing, technically experienced workforce that has been relied upon for the last few decades are heading into retirement and no new people want to learn these professions.
5. In the year 1937, Toyota started to develop the "just in time" concept for the Car manufacturing industry. In 1978 the Lean production theory was published in Japan and 10 years later translated into English. During the 1990s the Car industry started to establish lean manufactory. In 1993, the IGLC,[1] an international network of researchers from practice and academia in architecture, engineering, and construction (AEC), was founded and the first ideas of Building Information Modelling (BIM) were launched. After the millennium (year 2000) most of the other industry producers started to establish lean production models in their processes. The building industry is significant far behind.
6. The population is getting older. The ageing Society is recoding the economy, which will become apparent in the coming decade. People in the second half of their life have a different view and different needs on performance, innovation, and growth than younger generations.
7. We live in a global network. Everyone is connected with everyone and everything. The interaction between people and technology and the handling of new opportunities will change our social systems. Now it is important to use the enormous potential correctly and develop new business and revenue models.
8. In our complex world, knowledge is fluid, which is why implicit skills that allow us to be agile and respond to change are in focus. Holistic, systemic thinking, context formation and observation become core competencies as well as deeply interpersonal qualities in our knowledge culture.
9. Through new digital and innovative technologies like GPS, the Internet of Things (IoT), artificial intelligence (AI), augmented reality (AR), roboting, and automatization the construction industry has the necessary tools which are required to overcome the high complexity and improve productivity.

[1] http://www.iglc.net/

5 Is the Technology Breakthrough There?

In 2019, KPMG initiated a survey on the global construction Survey (Armstrong and Threlfall 2019). The majority of participants acknowledged the importance and impact of technology and innovation, but few were adopting it significantly, with even fewer reaping the benefits.

MGI's productivity survey also indicated that the biggest barriers to innovation by construction companies are underinvestment in IT and technology more broadly, and a lack of R&D processes (Barbosa et al. 2017).

We are at an early stage of lean manufacturing and at a very early stage of digital transformation of the construction industry. Looking at the global trends and the need of human beings, we can see that the fast transformation is more than needed. We need the change now. The digital revolution touches construction companies, as well as every participant in the value chain and is creating a massive potential for new business models, new products, and services through innovation. At present, globally there are more than 1.000 relevant start-ups in the construction industry focused on analyzing newly appearing technologies to push the transformation. Most of them are based in the United States, others are scattered all over the world. These start-ups are driving new technologies and some of the most relevant and influential initiatives are introduced below.

5.1 Smart Building Material and Green Technology

The much awaited and anticipated revolution in construction is gaining momentum. Researchers and various institutes are taking construction material—and product technologies to the next level. Developments in construction materials has been intense and has managed to offer a very convincing answer to the burning question of how modern construction materials could look like in the near future. A selection of the most interesting start-ups from Germany, Austria, Norway, and the Netherlands are giving answers.

5.1.1 Interpanel GmbH

The interpanel GmbH[2] is producing and selling a unique product technology which combines cooling, heating, acoustics, lighting in one prefabricated, modular and ready-to-install solution for buildings. The company was founded in 2016 in Crossen, Thüringen, Germany. interpanel is spin-off of the Fraunhofer Institute.

The global demand for room space cooling increases exponentially due to global warming, higher comfort requirements and economic development. The prevailing principle is an often inefficient, unhealthy, and noisy air conditioning. Due to a lack of regular maintenance, appropriate sizing and installation AC leads to health

[2] http://www.interpanel.com/

problems and discomfort. Studies resulted in a discomfort rate of up to 60% (VBG 2018). Discomfort leads to increased labor costs and reduced productivity.

Interpanels technology cools the room by absorbing thermal radiation, guaranteeing optimal comfort with zero draft. As first of its kind the radiant cooling functions below the dew-point of the ambient air. Due to the high cooling capacity this decreases the necessary covered ceiling area by up to 70% compared to conventional systems. The surface can be activated as human-centric surface lighting system for workspaces. To reduce noise all panels are acoustically equipped. As a result, the essential room climate user needs are covered with one solution: heating, cooling, light, and acoustics. In addition, smart controls and sensors are used to generate data for customers. This automatically optimizes the systems performance. The data is accessed via app. The interpanel solution enables significant lower greenhouse gas emissions, reduced building complexity, and highest flexibility.

Team members are Mr. Alexander Buff (CEO) who is essentially responsible for marketing, sales, IP and process and product improvements. Mr. Daniel Himmel (CFO) handles financing, sales, investments, and legal. Mr. Dominik Donderer, (CTO), production—quality—and product management. Mrs. Andrea Keisers (COO) takes care on process documentation, sales, database maintenance, and operational project management. The interpanel product integrates the functions of cooling, heating, human-centric light, and acoustic absorption in a multifunctional ceiling-sail solution. The "acoustically effective climate light" is the only surface cooling system that is free of condensation water and drafts and offers reliable, healthy and quiet high-performance cooling even in hot summer months. The interpanel technology was developed in cooperation with the Fraunhofer Institute for Building Physics and is patented worldwide. Interpanel is not only presenting a product innovation but as well the complete company is fully digitalized and using innovative lean management concepts, which enables them an early stage profitability. All product components are standardized and designed that construction projects are delivered with module dimensions.

All operational processes are mapped through fully digitized workflows in cloud-based ERP systems. The consistent database gives a complete overview of all-important processes in the warehouse, production and CRM from the first day.

The lean decentralized organizational structure supports an intensive relationship with suppliers and customers. Customers are in focus by every member of the management team and presentations and meetings are held and accessible online. The constantly updated activity documentation is an internal knowledge and process database creating transparency and clarity.

Interpanel is currently active in the mid- to high-end construction sector with a focus on office and commercial properties. Development opportunities are in the area of product diversification and innovative developments in new product areas like medical facility buildings and global expansion.

5.1.2 Nuki Home Solutions GmbH

The Nuki Home Solutions GmbH[3] provides smart home solutions and makes access controls smarter and physical keys irrelevant. Nuki is based in Graz, Austria and was founded in 2014 by Mr. Martin Pansy.

The company aims to make a change from standard physical keys to access controls based on technology. They believe in the power of mobile technology innovations and use it to facilitate our daily lives. Nuki turns a smartphone into a smart key in only a few minutes. The Smart Lock is mounted on the inside of the existing door lock and has or requires permanent Internet access. The smart door lock delivers maximum convenience to its customers. By using Bluetooth, Nuki opens the entrance door automatically as soon as an authorized person approaches and closes the door at the push of a button, when the building is left.

5.1.3 Airthings

Airthings[4] designs and sells digital radon detectors and was founded in 2008. The Management Team Mr. Øyvind Birkenes (CEO), Mr. Koki Yoshioka (COO), and Mr. Erlend Bolle (CTO) are based in Oslo, Norway. Airthings uses accurate technology to create user-friendly digital radon detectors. An estimated 20,000 people die from radon exposure every year, and it is the second leading cause of lung cancer after smoking. The primary customers are homeowners living in radon-prone areas in the United States, Canada, and Europe. Homeowners can use their device to see if they are at risk of high radon exposure. Radon is a radioactive gas that seeps into the foundation of many homes. Airthings also aims to replace traditional charcoal detection devices that are cumbersome, need to measure for months, and then have be sent to a lab to obtain any results. They also differentiate themselves from their competitors through better build, quality, and design.

5.1.4 Breeze Technologies

Breeze Technologies[5] provides air quality sensors as well as data and analytics services. The company is based in Hamburg, Germany and was founded in 2015 by Mr. Robert Heinecke (CEO), Sascha Kuntze (CTO). Breeze Technologies pushes the limits of environmental sensor development. Their small-scale air quality sensors can measure common pollutants like carbon and nitrogen oxides, ozone, particulate matter and many more. Their competitive price point allows for new applications like smart air quality management in a buildings or large-scale environmental sensing in your urban environment. Their environmental analytics cloud platform gathers real-time data from Breeze Technologies air quality sensors as well as external data sources. Based on machine learning and big data technologies, they use their proprietary Adaptive Cloud Calibration Engine to increase data reliability and accuracy. Their cloud platform allows to achieve an arbitrarily high data resolution

[3] http://nuki.io/de/

[4] http://www.airthings.com/

[5] http://ww.breeze-technologies.de/

and can assist facility management, environmental scientists and even municipality management and governments in understanding air quality, its influences and how to improve it.

5.1.5 Field Factors

Field Factors[6] is a developer and supplier of rainwater management solutions for urban applications, with special attention to spatial quality. They are a water technology provider, founded in 2016 in Delft, the Netherlands. Mrs. Karina Peña (CEO) and Mr. Wilrik Kok (CCO) are the founding partners and Management Team with the vision to restore the natural water cycle in cities. With an integral approach, Field Factors delivers innovative nature-based solutions for on-site water management, which guarantee high performance and reliability. They have developed Bluebloqs Technology, a circular water system for rainwater treatment, storage, and reuse. Bluebloqs can be applied to ensure green parks, playable sporting pitches, and good quality water for industrial use. Field Factors works with architects and construction teams on urban development and construction projects in the private and public markets in Europe. Installations are done by selected partners with infiltration and recovery systems competences.

5.2 Artificial Intelligence, Data Analytics, and Internet of Things

Huge improvements by a structured on-site execution are possible. Using a precise planning process to ensure that key activities are achieved on time and on budget. The use of integrated planning tools will at least achieve 50% increase in project's efficiency. Next is the reshaping of relationships and interactions between all stakeholders like owner, architects, contractors, and subcontractors agreeing and executing regular performance meetings with additional forward-looking plan metrics to identify and reduce variance is critical, ensuring that all preworks have been completed in time. Lean construction manufacturing with the principles to reduce waste and variability need to be established. To change from processes that rely on command-and-control to a more responsible holistic operating system with educated and reliable participants. The high complexity of projects requires an operating approach that integrates technical, logistical, and management competences to maximize resources.

There are some very exiting start-ups representing jobsite construction management mobile software and AI platforms to increase speed, quality, communication, and collaboration between all stakeholders.

[6] http://fieldfactors.com

5.2.1 Fieldwire

Fieldwire[7] offers the easiest way for construction companies to stay organized on their jobsites. It is a construction field management platform which connects owners, architects, engineers, general and specialty contractors, and suppliers together in real time.

Fieldwire's Headquarters is based in San Francisco, California, USA. Fieldwire has a regional US office in Scottsdale, Arizona, an international office in Paris, France, and is considering an additional office in the JAPAC region. Mr. Yves Frinault (CEO) and Mr. Javed Singha (COO) are the founding partners. Since founding the company in 2013 Fieldwire has supported more than 500,000 projects active in more than 100 countries with more than 2000 paying customers in 13 local languages. The Fieldwire platform is helping construction companies of all sizes by empowering clear communication on projects. With its easy-to-use mobile application, Fieldwire saves each user 1 h every day by enabling more efficient information sharing onsite. Fieldwire's key features include plan viewing, task management, document control, custom forms, progress photos, instant messaging, reporting, and more—all accessible from one place. Fieldwire is venture-backed by top investors in the industry and is already transforming the way dispersed teams communicate and collaborate on projects.

Fieldwire's construction management software is a fully featured blueprint management solution, so people in the field can view, edit, and share drawings. The construction app makes it easy for everyone to use the latest information and drawings. It is a paperless approach, which renders the need to print new paper copies useless. You can track changes and record your markups and verify with photos and videos. After changes are done, you will receive the latest updated plan as auto-hyperlinking in and real-time sync.

Fieldwire's mobile and web-based construction management software connects field and office teams and gives everyone on the jobsite the tools to execute day-to-day work. Through real-time communication the tools are helping to accelerate decision-making and resolution. These improve craftsperson productivity in the field by putting the information they need right on their phones.

Fieldwire's construction scheduling software focuses on the day-to-day execution of the project plan, bringing both speed and structure to the entire team. Organize, assign, and distribute work from any device while making sure nothing falls through the cracks. Easy crew scheduling by coordinating all upcoming items via a Kanban priority and calendar view. Accurate progress tracking. It is possible to dispatch work to each specialty contractor.

The punch list app adds both speed and structure to the closeout process. It allows to run a better walkthrough process and assign work directly to the people responsible for getting things done. It automatically generates PDF reports and enables sharing them with the project team, owner, or architect with the result of faster closeouts.

[7] http://www.fieldwire.com

5.2.2 INDUS.AI

INDUS.AI[8] is an Artificial Intelligence Platform for Construction based in San Francisco, California, USA and in Toronto, Canada founded by Matt Man (CEO) and Navin Kaminoulu (COO) in 2017. The company has clients globally in Hong Kong, San Paulo (Brazil), and the continental United States and Canada.

The INDUS.AI's platform enables developers, project managers, and general contractors to optimize labor, equipment, and materials deployment on commercial construction projects. INDUS.AI is an advanced construction intelligence solution provider, who enables real estate investors, owners, developers, Architects and general contractors to have real-time visibility and actionable insights into all activities, productivity, and risks at their construction sites. They are enabling active proactive monitoring with continuous (1) live site streaming by permanent cameras (including safety and intrusion alerts); (2) artificial intelligence by computer vision (Truck analysis, material arriving, data storage) and machine learning; and (3) with real-time dashboards, predictions (forecasts) and reports (Truck and stuff reports, claim reports) detecting both anomalies and progress. These videos show how INDUS.AI captures, interprets, and analyzes video streams and time-lapse images and turns them into actionable insights. Understand the unit production rate for each trade and compare it to your planned assumptions, track the construction progress and predict whether the construction is on schedule.

Labor analysis for contractor and subcontractor coordination. Get ahead of resource loading and timesheets. Track progress against your Building Information Model (BIM). Ensure that your site and workers are compliant with all safety standards. Track equipment uptime to better manage schedule dependencies. Track and automatically collect construction data on material. Track material arrival and departure insights for project cost controls. Clear visibility in prefab quality and production.

5.2.3 Building Radar

Building Radar[9] empowers companies to discover and win potentials building projects through digital tools. By that their customers can realize new market opportunities. Building Radar informs its customers about construction projects worldwide at a very early stage. State-of-the-art AI technology is searching in the internet around the clock and customers are able to find new and existing construction projects before their competitors and are able to pitch first. With the help of Building Radar, customers are informed at an early stage about activities in the construction industry and can thus use this opportunity to position their products and services first.

Building Radar was founded in 2015 by Mr. Leopold Neuerburg (MD), Mr. Paul Indinger (MD), and Mr. Raoul Friedrich (CPO). Member of the management team is Mr. Julian Scharf (COO). The team is supported by 45 employees. Customers

[8] http://www.indus.ai

[9] http://www.buildingradar.com

are companies and individuals whose products or services are needed throughout the construction value chain. Similar services are already offered, but on a manual basis only rather than by machine. In General, manual information are much later available and not scalable. A search algorithm searches in real time the Internet for information on new construction projects. Every day, hundreds of thousands of websites, online newspapers, and information portals (for example, companies involved in construction, such as architects or engineering offices, public bidding platforms, local news sites, architecture blogs, etc.) are searched for new construction projects. The information is extracted using Machine Learning, Data Mining, and Natural Language Processing algorithms. More than 100,000 sources will be generated daily, resulting in approximately 5000 new projects per day globally. Another product line is the new digital information tool: Market Intelligence generates information that updates users about current market trends and movements. This information usually represents another competitive edge for their users. The next step will be the international expansion, with branches in the United States and the United Kingdom and new product development to enlarge their service offering.

5.2.4 bGrid

bGrid[10] is a technology and innovation company based in Amsterdam. Mr. Wouter Kok (CEO) founded the company in 2015. bGrid develops, markets, and sells smart building products and services that use state-of-the-art communications technology, with focus on controls, remote monitoring, and "Internet of Things" solutions. To support the delivery of these products and services, bGrid develops mobile devices and sensor units, data management systems, data processing algorithms, and data interfaces. bGrid works in cooperation with international technology partners and integrators, and executes projects for commercial real estate, education, airports, hospitals, laboratories, and other buildings.

The bGrid Smart Building Solution is an answer for intelligent buildings. A network of bGrid Nodes senses everything that happens in the building and enables fast accurate positioning of people and assets. The bGrid Smart Building Solution is open enough to connect and communicate with everything in your building from the lighting and climate system to the coffee machines and even people through their smart devices. It also enables controlling light, climate, blinds, etc. based on the collected and analyzed data. The open API enables third party smart building hardware and software developers to easily connect to bGrid and develop new innovative building applications.

5.2.5 reINVENT Innovation GmbH

The reINVENT innovation GmbH[11] simples communication and planning processes for construction projects via a central digital platform. They are based

[10] http://www.bgridsolutions.com

[11] http://www.re-invent.io/

in Munich, Bavaria in Germany. They offer a software platform to connect all major parties during and after construction. It is their mission to redesign all customer processes with digital solutions for communication, data transfer, and documentation to generate significant cost and time savings for building contractors and project developers. Mr. David Uhde (CEO) and Mr. Julian Stieghorst, Mr. Valentin Rüchardt (CTO) and Mr. Christian Brachert are representing the founders and the management team.

5.3 Building Information Modeling (BIM), Virtual (VR) and Augmented Reality (AR)

Building Information Modeling (BIM) describes the way in which the entire value chain of the construction industry is interconnected, if planning, execution management and demolition of buildings are using BIM software. All relevant building data are digitally modelled, combined, updated, and recorded. The building is also geometrically visualized as a virtual 3D model. Benefits are significantly improved quality of data, as they all go back to a common database and are constantly synchronized by all involved with immediate and continuous availability of all current and relevant data. This results in a significant improvement on how to exchange information and coordination throughout the life cycle of a building.

Whether architects, planners, or building owners, all benefit from the possibilities of virtual reality (VR) configurators. With them objects in planning are made accessible and adaptable. This greatly simplifies decision-making, for example in the selection of materials or the layout of the room, and allows early detection of errors before they are costly or even irreversible. The applications only need to be developed once. Then they can be viewed and used from any location with VR glasses.

Augmented Reality (AR) is a computer-aided extension of the perception of reality. This information can appeal to all human sensory modalities. Visual presentation by supporting information, such as images or videos with additional computer-generated information or virtual objects by means of fade-in or overlay. AR is used to help with complex tasks such as building construction, industrial applications, navigation, digital cameras, geology, architecture, simulation, and learning.

5.3.1 Finalcad

Finalcad[12] provides Collaboration platform as mobile apps and predictive analytics that helps all project's stakeholders fix issues found on the building process. Mr. Jimmy Louchart (CEO), Mr. Joffroy Louchart, and Mr. David Vauthrin founded the company in 2011 in Paris, Ile-de-France. They already helped more than 24,000 projects in 35 countries with 9 local offices and keeps on advancing the

[12] http://www.finalcad.com/

digital transition of the construction industry. Finalcad is focusing on the segments buildings, infrastructure, and energy. It also wants to expand its activities for companies working in other industries. Their vision is based on lots of hours spent on the field, people within the company who are coming from the field in their past professional experiences. This vision is to facilitate the construction with their software, AI analysis tools, and their people. Finalcad offers a guide along the complete construction cycle by Mobile App for site and field engineers, by Web App for the construction site for projects managers and an Analytics App for the general management and project lead. Finally, you will have a complete BIM model with all drawings and information.

Finalcad helps its customers at any stage of their project. Some use cases are:

1. Defect management during and after the construction
2. Digitalized Quality controls
3. Progress follow-up during the whole construction
4. Improved proactivity and reactivity with Health and Safety (H.S.E.) matters

Defect management can easily add observations directly onto drawings (architect, plumbing, technical drawings), assign defects to companies and trades and enrich observations with pictures, comments, and schemes. Track the defect solving rate via a validation workflow. Generate reports from your mobile app or from your web-app. Send your list of defects to the right persons with our PDF and Excel reports. Get a site overview of your defects through the dashboard. With this tool time savings and quality increases in real time are the biggest benefits. Digitalized quality control uses all control forms from the site by using templates. Localization of control sheets on drawings or on a BIM object (checklists, measures, pictures) and get signature from anybody on site or office. Track the whole list of controls from the dashboard. Get real-time information about your team's progress with drawing elements that need to be done based on your daily production. Assign them to the team. Add details about each element (concrete type, element type, openings). Get an overview of the potential orders. Score the daily progress and add information such as concrete consumption, teams, number of hours, etc. Get a weekly progress report and export all data in order to compare progress with planning. Automatically calculate productivity ratios. Capture any field incident, share it instantly and record all safety events and identify risks to take preventive actions. Stay informed about all safety matters via real-time notifications. Automatically generate safety reports to measure the incident rates. Monitor and improve continuously to reach a zero-accident rate.

5.3.2 Matterport

Matterport[13] develops and operates cloud-based platform that enables users to create cloud-based 3D and virtual reality models of real-world spaces that can be experi-

[13] http://www.matterport.com

enced, changed, and shared online. Its 3D model allows people to walk through, modify, and share digital environments on devices, such as laptops and iPads. Matterport also provides Matterport Pro 3D camera to capture visual and spatial data and the appearance and dimensions of a space, Matterport Cloud, for cloud hosting and processing, Matterport Spaces for playing web player, and Matterport 3D player, a web-based viewer that enables users to see and navigate through their 3D model. It also offers 3D Showcase that enables users to navigate and visualize homes and other buildings with multiple floors. Matterport serves home improvement, furnishings and decor, property insurance, real estate management, real estate photography, construction management, hotels and vacation rentals, retail space planning, forensics animation, travel, engineering and construction, public and private security, and other industries. Mr. David Gausebeck (CTO), Mr. Matthew Bell, and Mr. Michael Beebe founded it in 2011, with its headquarters in Sunnyvale in California. Next to David the members of the Management Team are Mr. RJ Pittman, (CEO), Mr. Chris Bell, (CMO), Mr. Dave Lippman, (CDO), Mr. JD Fay (CFO), Mrs. Jean Barbagelata (CHRO), and Mr. Jay Remley (CRO). Additional offices are in Chicago, New York and London.

5.3.3 IrisVR

IrisVR[14] is the leading software for immersive design review and collaboration in virtual reality. The company was founded by Mr. Shane Cranton (CEO) and Mr. Nate Beatty (CTO) in 2014 in New York, USA. It is used by BIM and VDC teams, design firms, and engineers who coordinate 3D models and implement design and construction processes. Because IrisVR integrates with Revit, Rhino, Navisworks, SketchUp, and other 3D tools out of the box, you can instantly create an immersive VR experience that allows you to present to clients and work more effectively with your team. IrisVR offers a desktop product called Prospect that works with the HTC Vive and Oculus Rift and Windows MR headsets. Prospect makes it easy to host model coordination and QA/QC meetings in a true-to-scale environment. Quickly catch clashes, spot issues, and solve problems with Prospect's built-in tools before they make it onto the job site. Driving hours to the job site and screen sharing kills productivity. Prospect streamlines collaboration by connecting remote teams around the globe in a true to scale environment. Keep the team present and productive with Prospect. A mobile product called Scope that is compatible with the Samsung GearVR, Google Daydream, and Cardboard. Both products are available for free evaluation at www.irisvr.com IrisVr enhance BIM and VDC workflows.

5.3.4 XYZ Reality

XYZ Reality[15] is supporting the construction industry with a unique Augmented Reality solution that is able to reduce construction costs by up to 20%. Users are able to walk on site and view their 3D BIM Model, in context, to mm accuracy,

[14] http://www.irisvr.com

[15] http://ww.xyzreality.com

using Augmented Reality. No more disputes, no more out of tolerance errors and real-time validation. The company was founded by Mr. David Mitchell (CEO), Mr. Murray Hendriksen (CTO), and Mr. Umar Ahmed (COO) in London, UK in 2017.

5.4 Robotics, Drones, and 3D Printing

Robotic technology provides the construction industry with many opportunities and advantages. Automating processes with increasing productivity, robotics is being used to get work done cheaper, fast, and with a higher quality than manual labor can do. With drones the construction site inspection is easy to monitor. Structured progress monitoring understanding the status of projects. Easy-to-use tools enable measurements and annotations. Communication and getting information by drones are important factors which increase efficiency and reduce labor cost as well you improve the overall security.

A 3D-printed house can be a prefabricated house that can be manufactured off-site or produced on the construction site. With a 3D printer static constructive structures of the house like frame are produced, as well as the walls and the cover construction. Layer-by-layer, using 3D printing technologies entailing a big robotic arm with a nozzle that extrudes specially formulated material, the structures are created. All other building elements, including doors, windows, stairs, flooring, tiles, plumbing and many other elements will be installed in the factory. The advantages of 3D printing a home are the low cost and high speed in which the frame can be built. It is a very sustainable green way of building, as you do not produce any waste. 3D printers today are not limited to concrete but can be used to print metal which the following company is presenting.

5.4.1 MX3D

MX3D[16] is a highly innovative company that developed a groundbreaking additive manufacturing method based in Amsterdam, The Netherlands. MX3D was founded in 2014 by Mr. Gijs van der Velden (CEO), Mr. Tim Geurtjens, and Mr. Joris Laarman. Next to Gijs there are Mr. Jelle van Kleef (CTO) and Mr. René Backx (CCO) part of the actual Management Team. MX3D developed a groundbreaking additive manufacturing method with robots (WAAM, Wire Arc Additive Manufacturing).

They can 3D print metals and resin in midair, without the need for supporting structure. It looks like they invented a technology, which can be the beginning for a manufacturing revolution. As this technology opens up endless possibilities, the digital design and fabrication are changing rapidly. They are showing that digital fabrication is entering the world of large-scale, functional objects made of durable materials. Their main competences are based on their proprietary software which is guiding the welding robots of the metal 3D printer.

[16] http://www.mx3d.com

What are the main USPs?:

1. Scaling of machines through software.
2. Low capital investment cost and low operational cost.
3. The freedom of producing complex and large components by 6-axis robots.
4. Access to vast library of off-the-shelf metal materials.
5. Most important is that MX3D robots are not bounded to a building envelop and are able to produces and repair big units and parts.

MX3D executes two different business concepts:

1. Printing and manufacturing services for industrial companies (Best practice)
2. Selling and licensing software for other 3D metal printing companies (Scaling competences)

5.4.2 KEWAZO

KEWAZO[17] offers a smart scaffolding logistics robot system that transports scaffolding parts in a flexible, cost-efficient, and safe way. KEWAZO has a strong, interdisciplinary, and international team with six co-founders with diverse background, combining broad knowledge in robotics, civil engineering, automation, software and business disciplines: Artem Kuchukov (CEO), Ekaterina Grib (CFO), Alimzhan Rakhmatulin (mechanical engineering), Eirini Psallida (software engineering), Leonidas Pozikidis (electrical engineering), and Sebastian Weitzel (product development). The first pilot projects are running and in 2020 the company aims to begin with the series product sales.

From 2021, they plan to present a 2D solution for the transport in vertical and horizontal directions as well will introduce robot as a service business model (RaaS—Pay-per-use). The patent-pending solution of KEWAZO is a scaffolding logistics robot "Liftbot" that transports scaffolding parts during scaffolding assembly. The solution is composed of robotic modules and rails. The rails are installed on standard scaffolds using standard connectors, which allows for rapid installation. Robotic modules are installed on the rails within several seconds. They move on the rails vertically and horizontally and transport scaffolding parts from the ground to the assembly point just-in-time and just-in-sequence, securing a constant material flow. The proposed system automates the logistics process by employing autonomous navigation and workers detection. Compared to existing solutions the robotic system requires only two workers for operation, is easy to install, and can perform vertical and horizontal transportation of parts. The solution addresses labor shortage, saves at least 33% of labor costs, and increases productivity by 20%. The robotic system also records operational data from the construction site by using computer vision and sensors data. By creating a digital twin of on-site operations, KEWAZO provides customers with better controlling, planning, and suggestions.

[17] http://www.kewazo.com

KEWAZO provides a fully integrated smart solution—a package that contains a robotic system empowered by the data analytics platform. The customers have an option of buying and renting the system. The data analytics platform is offered on the SaaS basis.

5.4.3 XtreeE

XtreeE[18] is a large-scale 3D printing system for the architectural and the construction industry with design requirements. They offer collaborative design approaches combined with large-scale 3D prototype manufacturing with concrete, clay, and polymers. The company was founded end of 2015 and is based in Rungis, Ile-de-France in France. Mr. Alban Mallet is co-founder (CEO) and contributes to the design of innovative large-scale additive manufacturing systems. Mr. Alain Guillen and Mr. Jean-Daniel Kuhn, managing directors, make large-scale 3D-printing technologies available in the building and infrastructure sector in France and international as well as service- and rental business. XtreeE's team brings together a vast array of complementary skills and experience, i.e. architecture, civil engineering, robotics, computer science, and material science. This allows the company to master 3D printing's complete production chain, from design to manufacturing, and gives the ability to intervene at every of an architecture or design project. They are active in art and design, developing new products together with their customers as well designing new house concepts.

5.4.4 Apis Cor

The company Apis Cor[19] develops mobile construction 3D printers that work in polar coordinates and is originally from Moscow City, Russian Federation founded by Mr. Nikita Chen-iun-tai (CEO), who lives close to Boston, Massachusetts, USA today. They call themselves "robotics in construction" as they 3D print the whole structure right on the site from start to the end. They are working on different new functions like inter-story floors, roof printing, automatic horizontal wall, and on foundation reinforcements placement.

5.4.5 Yingchuang Building Technique

First Chinese companies were actively driving the development of 3D-printed buildings. Spectacular are the achievements of the Chinese construction company called Yingchuang Building Technique (Shanghai) Co. Ltd. (Winsun)[20] that 3D printed a 6-story apartment building. Winsun has built more than 400 buildings and developed as well as created new building material which helps to increase speed, quality and reduce cost. Winsun improved the development of new building material like glass fiber-reinforced gypsum board and cement.

[18] http://www.xtreee.eu/

[19] http://ww.apis-cor.com

[20] http://www.winsun3d.com/

5.4.6 ICON3D

Another pioneer is the company ICON3D[21] based in Austin, Texas, USA. ICON developed concrete 3D printers as individual robots, the required design and application software as well the advanced concrete material.

5.4.7 RedWorks Construction Technologies Inc.

The company RedWorks Construction Technologies Inc.[22] based in Lancaster, California, USA is developing an on-site 3d printing system that only use on-site sources of sand/dust/dirt to create materials that can match the strength of existing masonry. This technology will cut overhead costs, reduce site logistics, and let builders print custom materials without impacting costs or build-time. The founders are Mr. Keegan Kirkpatrick (CEO), Mr. Paul Petros (CDO), and Mrs. Susan Jennings (Building Material Specialist).

5.5 Smart and Mobile/Modular Homes

Modular homes production can reduce the construction time by more than 50%. Site construction build houses can take many months to complete, while modular homes are assembled on site only and can be finished between 1 day and 1 week. In a controlled factory environment modular homes are built in sections, or modules, and then transported to the construction site. There, they are installed on permanent foundations and completed.

There are some very exciting start-ups on the way to change the way of construction buildings. Haus.me is based in the Ukraine and in the United States. Mighty Buildings Inc. in the USA as well, where the Project Milestone is founded in the Netherlands and Containerwerk Hall eins in Germany.

5.5.1 haus.me

haus.me[23] is the first fully self-sustainable mobile house provider. The portable home does not require an electric grid, propane, natural gas, firewood, or any other fuel. It uses solar energy for heating, cooling, and electricity and it works in both hot and cold climates. The benefits are no energy bills as the house is fully off-the-grid and self-sustainable. As the house will be delivered turnkey no site construction is required and is ready to use. The house is fully automated including operating system and smart home applications. The house withstands hurricanes and earthquakes and is zombie proof and is offered in different models like a family home.

[21] http://www.iconbuild.com

[22] http://www.redworks3d.com

[23] http://www.haus.me

haus.me is an integrated, intelligent, and smart home. The house possesses 24 intelligent subsystems that work together with house AI system to ensure the living safety and comfort. The home self-diagnosis system informs about the problem before it appears. haus.me can be used as a primary residence, a vacation home, a guest house, studio, or income-producing secondary dwelling—like an Airbnb rental—or as an autonomous hotel unit.

haus.me team was founded in 2016 as PassivDom and has years background and deep knowledge in energy-efficient construction, 3D printing, physics, chemistry, IoT, electronics software, real estate marketing, and sales. Julia Gerbut and Max Gerbut are the founders. Sergii Tychyna is the COO of the company. The haus.me business is like SaaS for Construction industry, including new products development, engineering smart IoT systems, logistics, software and cloud cervices. In the next phase they will look for international distributors and investors to build a global business.

5.5.2 Mighty Buildings Inc.

The start-up company Mighty Buildings Inc.[24] is located in Oakland, California, USA. Their concept is to build beautiful, affordable, and sustainable modular housing. They developed a 3D printing opportunity to enable a green way of building houses. Through their automation process in a factory, they are able to speed up the building process and are able to reduce cost simultaneously. Actually, they are in a stealth-mode and will inform about their strategic moves during 2020. The founders are Mr. Slava Solonitsyn (CEO), Mr. Alexey Dubov (COO), and Mr. Dmitry Starodubtsev (CTO). At present they have more than 80 employees and are present on Facebook.com and on YouTube.com with video footage.

5.5.3 Containerwerk eins GmbH

Containerwerk eins GmbH[25] is building modular homes and offices by using old/used containers. They are designers and create innovative portable housing solutions. They are based in Wasserberg, Germany and are founded and managed by Mr. Ivan Mallinowski and Mr. Michael Haiser. They searched for answers to housing shortages, inexpensive, resource-saving, and contemporary living and therefore introduced the method of using containers. In the interests of the circular economy, ecological and social sustainability, Containerwerk acquires used sea freight containers and refines the corpus into high-quality and inexpensive living space. Convinced and fascinated by the idea of building with disused sea freight containers, they are first dealt with the building block, the "brick" of this architecture, and developed a process that turns a container into a universal, sustainable housing module.

[24] http://www.mightybuildings.com

[25] http://ww.containerwerk.com/

6 Conclusion

Globally digital technology and services have affected a number of industries. The traditional construction industry can take the huge advantages of it. Digitalization has finally reached architects, building material producer, contractors and their value chain. And it is about time to get up to speed to use the tools which are available. Most of the contractors still work the same way they operated decades ago. A shift from the physical and manual work to the digital world is extremely beneficial and the construction industry seems to be slowly coming to this realization. Better profit margins, less cost, improve time efficiency, better collaboration and communication with advanced reporting, a higher productivity and a healthier security—digital technology and services are promoting the construction industry on a performance level, it never did before. Implementing it is not a choice, it is a necessity. If organizations are reluctant to innovative change, they will no longer be in business. The innovation train is driving every day faster to reinvent the construction industry, say yes to innovation and remain competitive. Supplementary material for this chapter is provided at www.christophjacob.com.

References

Armstrong, G., & Threlfall, R. (2019). *Future-Ready Index, leaders and followers in the engineering and construction industry*. Global Construction Survey 2019, Publication number: 136218-G, KPMG International Cooperative, Swiss, www.kpmg.com

Barbosa, F., Woetzel, J., Mischke, J., Ribeirinho, M. J., Sridhar, M., Parsons, M., Bertram, N., & Brown, S. (2017). *Reinventing construction: A route to higher productivity*. McKinsey Global Institute.

United Nations. (2019). *World population prospects: The 2019 revision*, New York City.

VBG. (2018). *Studie Gesundheit im Büro*. Wiesbaden: BC GmbH Verlags- und Mediengesellschaft.

Wang, T. (2017). *Global construction expenditures 2014–2025*. Accessed December 26, 2019, from http://www.statista.com/statistics/788128/construction-spending-worldwide/

Motivation, Employees, and Communication in the Start-Up Phase

Achim Denkel

Abstract

A challenging task for a start-up is the development of employees. Finding people who are interested in unpredictable overtime and bad pay sounds bad in any job description. Economically, there is no reason to refuse a job with a secure, successful company. There is also no economically understandable reason to give up a permanent position unless there is a better paid one. And there is no economic reason to change the last years of one's professional life into the uncertainty of self-employment, at least not voluntarily. Nevertheless, there are thousands of examples that refute exactly that. For a reason that cannot be measured in figures.

1 Motivation: The Engine of Our Founder Scene

It is impressive to see which subculture of the established economy has formed in Startup Hubs, Co-Working Spaces and the one or other free workrooms. A spark ignited the inner fire of inspiration and self-realization. With some entrepreneurs, the fire then also passes to co-founders and employees, but that does not succeed without a plan or talent. The vision must exist from the beginning, form or be created, otherwise the blazing fire falls on wet wood and goes out. How to get employees to peak performance, even though they have to work more in a badly paid job, have to carry more responsibility, possibly have to look for a new job in a few months. Financial incentives?

A. Denkel (✉)
CAPinside, Hamburg, Germany
e-mail: ad@capinside.com

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_17

For founders and co-founders planning an exit, the question is not whether there are financial incentives, but these are originally difficult to transport to employees. The rain of money, which is supposed to come up during the sale, is the dream of the exit-driven founder, but the employee in the team of 25 who is responsible for the acquisition of new customers on the phone or by mail bot, has not yet had this vision (Deutsche Startups 2013).

Work–Life Balance drops flat at probable workload? In theory, it is a nice thing, but it cannot be sustained in the development of work processes, because they are not yet foreseeable in terms of time expenditure. Nowadays all add-ons like Playstation, Kicker, or Superfood have become an indispensable standard and make no difference. Thus, visions, missions, and a good team have more weight than ever what makes the working day an experience (European Commission 2019).

2 Creating a Basis for the Team

> My model for business is the Beatles: They were four guys that kept each other's negative tendencies in check; they balanced each other. And the total was greater than the sum of the parts. (Steve Jobs) (Curtin 2019)

The beginnings of team building are the most challenging times in the team building phase. You need leaders, leading figures, and self-runners who absorb visions and missions like an elixir of life and carry the spirit into the team. It needs guidelines, written and felt, experienced and imagined. The first team members are crucial to maintaining these guidelines, while external pressure from investors and potential customers increases. The very beginning is the meaningful design of written corporate goals. Top down they let themselves interpret and stimulate these from the beginning, like a philosophy, which promotes creativity with procedures to the goal reaching which was before indefinite. Even if this has potential for misinterpretation, it is a clearly solution-oriented approach rather than making any decision dependent on a key person, such as a founder or co-founder, who ultimately makes opportunistic decisions by flooding tasks.

2.1 Create Room for Feedback

In addition to all fixed and established processes of meetings, group meetings and staff appraisals, team moods can best be achieved outside the working environment. If the team leaves its familiar environment into "free terrain", the role of all team members must also be reassigned. A new starting line is created for limited time for all. Be it a bar, a restaurant, a visit or a hike, whatever it is, it is a useful special situation. If the meetings are unplanned and voluntary, it is an award for the team spirit, but of course not only for start-ups. What makes the difference here is the discussion about setting up the various necessary business areas—only then can talents be identified who have fully understood and consumed the business model.

Feedback then becomes talent scouting and opens up new opportunities. Compared to established companies with fixed structures, there is a blatant difference here that distinguishes itself from the standard in agility and resource planning (Rericha 2006).

2.2 Agile Planning

Using a temporary work instruction as agile management for teams not only makes sense in project management, but also in the management of teams whose workflows have not yet been defined. Mistakes must be made known, causes and effects discussed and changed into modified workflows. Thus the department grows out of itself and finds the way to a stronger, better self. The biggest obstacle is the disclosure of mistakes, because even in our Western culture, the admission of poor performance is seen as a scar on the career ladder. This is countered by short-term planning, because the achievement of goals is more tangible and there is only the flight forward. If something has not worked, we discuss and adapt without letting bad feelings arise in the participants. They see themselves more as part of a process than as a source of error. This does not mean a release from responsibility, but a liability for the misappropriation of mistakes (Projektmanagement 2015).

In fact it motivates additionally, because the feeling to help shape something has a positive effect on people. People, on the other hand, who have sold themselves very well, quickly show their true core, because focusing solely on their career does not help them to error message or process optimization (Arndt 2004). Creative employees, on the other hand, drive processes forward.

2.3 Fast Communication

Decisions that are proverbially postponed can be contrary to the first intention. An everyday means to put something literally "on the long bank." If it is a masterful tactic in politics, it can lead to the end of a business idea in the young economy. But how can this danger of delay be avoided?

Not only in private life, but also in project management, short messages and text messages in chat format have been established for years. It sometimes paralyses the process in the decision-making process, but brings all parties involved, insofar as they are in the chat, to the same, current state of knowledge. Strict rules increase the speed (Rach 2019). The most important thing is to make final decisions, because they should be discussed in the chat format. Looking back, everyone can read how the decision was made, without influencing a decision in the aftermath. Thus the understanding of decision-making contributes as a further building block to the common picture of a vision or mission.

2.4 Limitation of Freedom

> That's been one of my mantras—focus and simplicity. Simple can be harder than complex: you have to work hard to get your thinking clean to make it simple. But it's worth it in the end because once you get there, you can move mountains. (Steve Jobs) (British Broadcasting Corporation 2011).

The challenge in transferring responsibility to employees is to avoid the loss of focus of the responsible person. In contrast to the founder who regularly gets feedback on his ideas, the responsible employee usually lacks the relevant environment. This does not mean that there is a lack of opinions in his environment, but the usefulness is doubtful.

Limitation by strict, disciplinary measures with creative, motivated personalities work deterring or even motivational destructive. The common mission supports the way of focusing and at the same time offers the possibility to be perceived not only as a superior but also as a counsellor! The way to support focusing and to be perceived as counselor, superior, is found in the common mission. The more the common mission is promoted, the less a limitation is necessary. If this no longer works through the common orientation, two questions arise. If the further mission is too undefined or communication no longer fits. Thus the focus turns back to the start, the team building.

3 Limits of Capabilities

As operation/working life progresses, skills are increasingly needed that were previously ignored. This is logical, because previously there were no concrete departments or defined positions to fill. Here there is the chance to further develop employees who have already attracted attention in indirect talent scouting through qualification and mission understanding and to equip them with additional tasks. One species decoupled from this are the creative minds.

It is more difficult to lead them and the more difficult it is for them to lead other people with exclusively creative talent. Creatively gifted team leaders with mission understanding and applied agile management skills are as rare as a four-leaf cloverleaf. So the goal is to set the creativity and the initiated self-realization within not noticeable limits.

Also the self-reflection in the management plays an important role and should be a firm component of the process, because the own abilities can reach limits like with each other coworker in the enterprise and that is the most serious brake which an enterprise can have.

4 Motivation Comes from Success

Whether a goal has been achieved can be measured in advance by determining it. A vision is too big to be measurable and a mission too long-term to have noticeable success, so the solution lies in the achievement of partial goals and that brings motivation. This brings the incentive to repeat the same experience at shorter intervals. Thus a cycle can be recognized from the sum of all things, which can be supplemented in its arrangement by further methods. The cycle is depicted in Fig. 1.

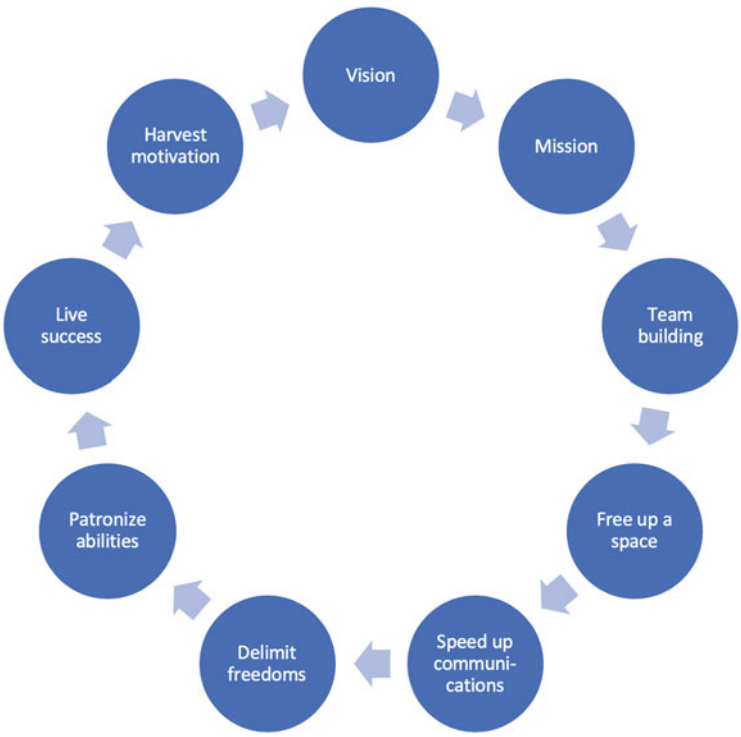

Fig. 1 From motivation to success. Source: author

References

Arndt, H. (2004). *Supply chain management*. Wiesbaden: Gabler Verlag Springer Fachmedien Wiesbaden GmbH.

British Broadcasting Corporation. (2011). *In quotes: Apple's Steve Jobs*. Accessed December 22, 2019, from https://www.bbc.com/news/world-us-canada-15195448

Curtin, M. (2019). *33 Steve Jobs quotes that will inspire you to success*. Accessed December 22, 2019, from https://www.inc.com/melanie-curtin/33-steve-jobs-quotes-that-will-inspire-you-to-achieve-massive-success.html

Deutsche Startups. (2013). *So wird der Exit erfolgreich*. Accessed December 22, 2019, from http://www.deutsche-startups.de/2013/04/29/exit-erfolgreich/

European Commission. (2019). *Work-life balance*. Accessed December 22, 2019, from http://ec.europa.eu/social/main.jsp?catId=1311&langId=en

Projektmanagement. (2015). *Agiles Projektmanagement*. Accessed December 22, 2019, from http://projektmanagement-definitionen.de/glossar/agiles-projektmanagement/

Rach, P. (2019). *Teamregeln - 10 erfolgreiche Teamregeln für mehr Erfolg*. Accessed December 22, 2019, from http://www.rach-team-kommunikation.de/newsroom/69-teamregeln.html

Rericha, N. (2006). *Psychologie der Kommunikation: Feedback(regeln)*. University of Klagenfurt. Accessed December 22, 2019, from http://wwwu.edu.uni-klu.ac.at/nrericha/lv-epp/nrericha.pdf

AI to Solve the Data Deluge: AI-Based Data Compression

Eric Falk

Abstract

The massive amounts of data, growing as we speak, are one of the, if not the, most accountable reasons of today's AI systems which on many tasks exhibit human grade performance. Thanks to the enormous amounts of image data that machines can be trained to recognize scenes and steer cars. Quantities of medical imagery lead to machine provided diagnostics, sensor data allows us to detect natural disasters before they occur, and to prepare for them. Times are exciting since researchers find new applications to AI at astonishing pace. However, there is a small concern. How will we handle the ever-growing amounts of data? The consensus is that storage is cheap, yet with load of data it is expensive and unsustainable. The amount of live streamed data is also increasing. In other words, we are well advised to consider data compression again. In this chapter, we will introduce traditional compression terminology and techniques, before surveying novel approaches proposed by industry and academia. It sounds contradictory, but AI may just as well help us to address this problem.

1 Introduction

At the crowning of King Charles the VII of France, in 1422, the Duke of Uzès proclaimed the famous sentence: "The king is dead, long live the king!". This saying is particularly adequate in our technology driven world. In similar fashion a king among technology, polarizing industry and academia, fades and gives way to the new king. Typically, in such transitions all problems introduced by the previous monarch are regarded as solved, and the new king can build on the legacy

E. Falk (✉)
NIUGroup SARLS, Luxembourg, Luxembourg
e-mail: e.falk@niugroup.lu

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_18

of the former. In the last decade, we saw the rise of Big Data enabling first the querying, and later the analysis of extremely large amounts of data. Data analysis has become the new king technology under the designation Data Science then evening the path for Machine Learning. It was not sufficient to analyse data and extract information anymore, the information has to be leveraged in an automated way to render decisions and call to actions. Today we are in the era of artificial intelligence (AI), where machines begin to tackle tasks requiring human cognitive and decision-making skills. This outline and classification is as correct as it is wrong, minds collide on the topic of properly defining the above terms. We are not aiming at entering this discussion but at making the point that the initial enabler of the current information and communication technologies realm is data. The large whale companies such as Google, Facebook, Amazon, and Apple understood this early on. There, data related tasks are addressed with plenty of computing power, oftentimes with specific hardware. GPUs are the most prominent representative of the high performance computing (HPC) clusters. With similar tools at their disposable through various cloud computing offerings, more and more companies and also public institutions start assessing and leveraging their data gold mines. It is not as trivial as it may sound.

To illustrate the complexity and the costs, we describe a use case out of our experience. Without disseminating more details, an entity was in possession of data generators, dumping 1 TByte of industry specific standardized binary data per day. To exploit the data, to extract actionable insights, the files have to be parsed which in the same time expands the data by a factor three. Now a single day worth of data represents 3 TBytes. The company owns thirty units generating that 1 TByte per day so we are looking at ingesting thirty times 3 TBytes of data per day, 90 TBytes. With a data retention of only 1 month, $30 * 90\,\text{TBytes} = 2700\,\text{TBytes}$, or 2.7 PBytes are stored at all times. Unfortunately, the story does not end here, inserting the data into Elasticsearch, a widespread distributed search engine often used for business intelligence (BI), comes at the cost of additional overhead. According to the article on the Elasticsearch blog (Kim 2015) the expansion rate for structured data, similar to the data we consider here, is between 0.553 and 1.118 per replica. We are almost at the end of our demonstration. Let us take the golden middle as an expansion rate for our example, namely 0.836. As a consequence, we now store 2.7 PBytes/0.836, 3.23 PBytes per replica. To ensure at least a little load balancing, a replication factor of 2 should be selected. All of a sudden we store 6.46 PBytes at all times. The current offer of a leading cloud storage provider is USD 0.0021/GB for storage requirements above 500 GBytes. Storing all this data costs USD 13,566/month and represents a yearly commitment of USD 162,792. Of course the estimation does not consider special discounts or building a private architecture. Therefore, some other aspects are estimated favourably such as the low data retention period and the low replication factor. Furthermore, operating a private architecture can be cheaper in terms of price per GByte, but infrastructure, electricity, and costs for staff must be considered.

We want to direct the attention of the reader that storage is cheap when compared to other components. Although, if the calculation is made, storage is not that cheap

after all when the data volumes are large. We should also think of streaming data. Sometimes the data is streamed from remote locations, where each byte has a price, for example, on a mobile plan. Altogether we are convinced that the democratization of AI is only possible if the cost factors mentioned above are mitigated. One way is to judge carefully what data should be kept and when it should be discarded. This is an active area of research, interesting angles have been outlined by Prof. Tova Milo, from Tel Aviv University, in her 2019 VLDB (Conference for Very Large Databases) Keynote talk entitled "Getting Rid of Data!" (Milo 2019). Although this approach is interesting, it is also tied to the risk of accidentally and irreversibly deleting the wrong datasets. The second idea would be to investigate data compression techniques. Data compression is ubiquitous but mostly unnoticed. Since decades generic data compression techniques are employed. While before it was an active research area, today only technology leaders built their own methods. Companies such as YouTube, Facebook, Google, and Dropbox run their custom compressors under the hood. Only in the last couple of years, codecs and methods have been made available to a larger audience. In this chapter, we will introduce basic compression terminology and concepts before introducing AI-based compression but also methods where machine learning and compression work in symbiosis.

2 Data Compression Preliminaries

Before introducing the exciting new concepts, we need to eat our theory vegetables by establishing basic vocabulary and concepts. The most important concept is the notion of *information entropy*. It designates the amount of information carried by data. An intuitive example would be that we tell you: "Today the sun is shinning!". This sentence contains important information for you. Although if we repeat the sentence, the amount of information you can extract is negligible since you already know the fact. Information entropy is denoted as the letter H, introduced by Claude Shannon in his famous paper (Shannon 2001). For a random variable X:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \cdot \log_b P(x_i) \tag{1}$$

For computers, the base b of the logarithm is set to 2 since the binary number system is used, and information is represented with zeros and ones. Intuitively you can think of this as how many yes/no questions you would need to ask until receiving a distinctive answer. The equation from 1 was proven to set the lowest bound that can be reached when compressing information. Data can never be compressed more than the entropy. Besides information entropy, context is another important notion. As we are based in Luxembourg, if we tell a foreign visitor: "In Luxembourg it rains a lot.", he can extract more information than a local person already aware. This is leveraged when encoding images, videos, or sound. Modern audio encoders, for example, remove sound waves not audible by listeners. Context is an important

notion since it is unnecessary to encode information the receiver does not require to understand the content.

In data compression two axes can be optimized. The first one is compression ratio, in which case it is important to come as close to the entropy as possible. Here, the target is clearly to save as much storage space as possible. Compression ratio is defined by $Compression\ Ratio = Uncompressed\ Size/Compressed\ Size$. The other metric to optimize is speed, where compression must be as fast as possible. This is especially important in critical loops such as data streams, where many packets have to be managed in short intervals. It can be tolerable to have a weaker compression ratio as long as time overhead is low. Actually, we could say that tree axis can be optimized because we should differentiate between compression speed and decompression speed. Depending on use cases, slow compression has close to no impact but decompression must be fast. Websites are one example. They can be pre-compressed once, which can be a long process, but should be decompressed at lightning speed so site visitors have an enjoyable experience.

A generic data compression codec is typically constituted of two consecutive steps. The first is the de-duplication phase, where repeating patterns are detected and replaced with shorter index and offset pairs. For the character string composed by the first letters of the words from the "99 bottles of beer on the wall..." song:

```
NNBOBOTWNNBOBIOOTBSHTFNEBOBOTWNE...
^^^^^    ^^^^^

NNBOBOTW(0;5)IOOTBSHTFNEBOBOTWNE...
```

The **NNBOB** substring was already seen by the encoder and is therefore replaced by an index and offset combination for the second occurrence. This base algorithm was introduced by Ziv and Lempel (1977). Today a multitude of variants exist, all still making use of the original findings, which is why so many compression algorithms include the "LZ" initials in their names.

Although the de-duplication step already provides compression, it is worth going further by subsequently applying entropy encoding. For a long time, two entropy coding mechanisms dominated, Huffman coding (Huffman 1952) and arithmetic coding (Witten et al. 1987). In the former, a binary tree is built based on the probabilities of symbol occurrences. Given the alphabet $\Sigma = \{a, b, c\}$ the respective probabilities of occurrence of each symbol are as follows: $P_{i \in \Sigma} = \{.5, .2, .3\}$. A Huffman encoding for Σ is shown in Fig. 1.

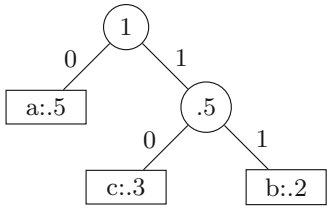

Fig. 1 Example of a Huffman code for $\Sigma = \{a, b, c\}$ and $P_{i \in \Sigma} = \{.5, .2, .3\}$. Source: author

The symbols are ordered by their probabilities, first the least probable symbol is connected to form a new inner node. This process is repeated until the shape of a binary tree is achieved. The most probable symbol a is encoded with the single bit 0, while c and b will be encoded with 2 bit short codes. All we have to do is walk up the tree starting from the symbol we want to encode. Applying the information entropy formula 1 on the Σ distribution $H(\Sigma) \approx 1.49$ indicates that the ideal minimal encoding corresponds to 1.49 bits per symbol. Encoding the sequence abc results in the binary string 00111. We required a total of 5 bits, instead of $8{\cdot}3 = 24$ as in ASCII characters are encoded on 8 bits. We want to point out that with an average of 1.49 bits per character we should only require $\approx$4.5 bits for the encoding of abc. This highlights one weakness of Huffman coding, it is not optimal as it does encode an alphabet by substitutions on per symbol basis. Therefore, Huffman coders can only encode symbols with a natural number of bits; the required 4.5 bits are rounded up to 5 bits. However, the strength of Huffman coding is its low complexity and fast run-times.

Arithmetic coding on the other hand encodes sequences of symbols and can therefore converge towards entropy much better. Taking the same example with the alphabet $\Sigma = \{a, b, c\}$, probabilities P_a, P_b, P_c corresponding to $P_{i \in \Sigma} = \{.5, .2, .3\}$. The arithmetic encoding process of the sequence abc is shown in Fig. 2.

In the first step of arithmetic coding, shown on top of Fig. 2, the probability interval of the sequence abc is computed, the interval is called $[L; U[$ for upper and lower bounds. In the second step, intervals are divided by two until an interval $[L'; U'[$ is found, that is completely contained in $[L; U[$. It gives us the encoding of abc, which is the binary string 0000001. The reader may have notice that with arithmetic coding more bits are required, but please keep in mind that the examples are chosen to be illustrative. Therefore, the performance of the encoders is not representative, over longer sequences arithmetic coding exhibits better compression. The weakness of arithmetic coding became obvious: it is the run-time of the algorithm. It heavily relies on one of the most expensive operations, namely floating point multiplications, with additional overhead due to interval re-normalization. The latter is necessary because long sequences let interval numbers overflow.

The typical conception is that Huffman and arithmetic coding are diametrically opposed. Where one encoder is slower with better compression ratio, the other is fast with a worse compression ratio. This already brings us to the first innovation in the field we would like to introduce. A new kind of entropy coding came to life with the *Asymmetric Numeral Systems* (ANS) (Duda 2013) paper from Jarek Duda in 2014. A variety of entropy coders have been derived from ANS, such as tANS or rANS, all having their own purposes. Explaining the algorithm would probably be a chapter on its own, the innovation of the ANS encoder family is that it provides arithmetic coding compression quality at Huffman coding speed. Several new, now open sourced compression codecs have been proposed with ANS used in Facebook's ZStandard (zstd) (Collet 2015), in Apple's LZFSE (Apple 2015), or JPEG XL (Rhatushnyak et al. 2019).

Now that the basic compression processes are understood, we hope the reader has developed an intuition what data compression processes consist of. Compressors

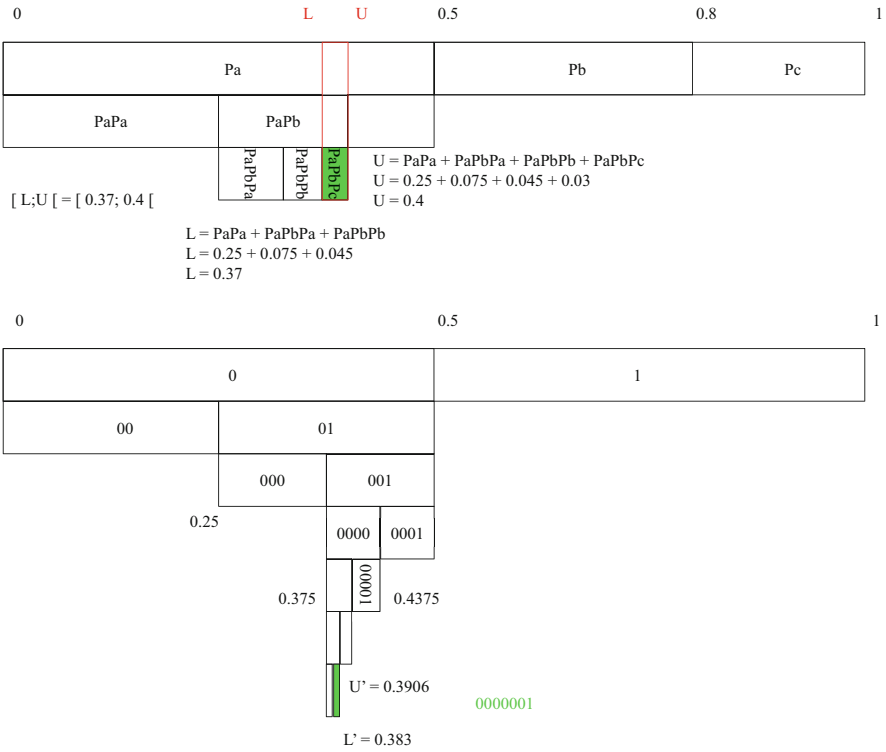

Fig. 2 Example of arithmetic coding for the sequence *abc*. Source: author

build statistical models of occurrences of symbols and assign short codes to frequent symbols and longer codes to rare symbols. Another method is also to remove information altogether if the context allows it. In image compression, for example, information is removed as it is not relevant to see the picture. If the information cannot be recovered to 100% after compression, we are talking about lossy compression. Whereas when data is recovered to be identical, the codec is said to be lossless. With this information, we can now move over to AI-based data compression mechanisms.

3 AI for Data Compression

One special type of data is particularly difficult to compress. We are talking about sequences of scientific numerical data, mostly encoded as double precision floating point numbers. This is because of the way they are constructed, namely using 8 Bytes containing a 52 bits significant, also called mantissa, an 11 bits exponent,

plus one bit for the sign $+/-$.

$$0 \quad 01111111111 \quad 0000000000000000000000000000000000000000000000000000 = 1 \tag{2}$$

$$0 \quad 10000000000 \quad 0000000000000000000000000000000000000000000000000000 = 2 \tag{3}$$

In Eqs. 2 and 3, we see two double precision floating numbers. The first bit is the sign, the 11 following bits the exponent and the final 52 bits for the mantissa. It becomes obvious that two subsequent integer numbers as 1 and 2 have a very different shape when we look at the bits. The first bytes differ whereas the final bytes are identical. With authentic scientific numbers, we rarely obtain matching bytes in two consecutive numbers, even when the changes are far into the decimal places. Below, two examples are given this time split as bytes. The corresponding numbers are 34.5 and 34.6, respectively.

$$01000000 \; 01000001 \; 01000000 \; 00000000 \; 00000000 \; 00000000 \; 00000000 \; 00000000 = 34.5 \tag{4}$$

$$01000000 \; 01000001 \; 01001100 \; 11001100 \; 11001100 \; 11001100 \; 11001100 \; 11001101 = 34.6 \tag{5}$$

We can see that even with numbers that appear to be mathematically close, their representations are divergent (see Eqs. 4 and 5). At best the first two bytes are identical. In (Falk et al. 2017), we addressed the task of identifying users of smartwatches by their movement data. The purpose of this work was to detect when a person different from owner is wearing the smartwatch. For that we used the data from the integrated inertial measurement unit (IMU). These sensors can be found in smartwatches, smartphones, and also in airplanes. To illustrate how irregular scientific data is we generated Fig. 3. We took one column of our data and displayed it as bitmap picture to show how much entropy scientific data exhibits (strong irregularity). The data corresponds to 500 data rows (one per column). This in return corresponds to a time windows of approximately 10 s as we have a sampling rate of 60 Hz. All data points are contained in the interval [0.46312; 0.99993].

In the upper parts, we can see some regularity, but quickly, as visible in the lower parts no patterns can be detected at all. Even over this short time window, and the tiny interval, entropy is extremely high. Since compression consists of finding regularities in data, modelling a statistical representation, and finally encoding, this type of data gives a challenge to compression codecs. This is unfortunate if we consider Internet of Things (IoT) devices. A large plurality of said devices are

Fig. 3 Sequence of floating point data from IMU sensor, as it can be found in smartphones. Columns correspond to one double precision floating point number. Source: author

sensors such as the IMU units we mentioned above. They are almost exclusively logging and transferring numerical data. Next we will present a neural network-based approach for the compression of numerical data. Subsequently, we outline the methodology we engineered, derived from the knowledge we acquired while working on the project described in (Falk et al. 2017).

3.1 DeepZip: Compressing Numerical Data with Neural Networks

The authors of DeepZip (Goyal et al. 2018) leverage the qualities of neural networks to predict the next symbol in a sequence. The particular type of neural networks are called recurrent neural networks (RNNs). They found large spread adoption for tasks such as speech synthesis, text or music generation, and automatic translation. DeepZip proposed a lossless compression codec, composed of two building blocks: a neural network-based *probability prediction block* and an arithmetic encoder block.

3.1.1 Probability Prediction Block

For the probability prediction block, several variants of neural networks have been evaluated. They range from fully connected neural networks to the more sequence adequate methods such as RNNs, in particular LSTMs, GRU, and biGRU. According to our experimental results which are in line with the findings in the DeepZip paper, we achieved better results with the LSTM/GRU/biGRU variants. In fact they offer more options for hyperparameter tuning and can represent longer sequence. Therefore, they can be adapted to different data more easily. For example, with our IMU datasets and data from IoT sensors deployed in client projects, we achieved better compression when we could map the relation between multiple features. Let us assume we want to compress data where a data row consists of three double precision floating point numbers. Each data row consists of $3 \cdot 8 = 24$ Bytes. This is a rather long sequence for neural networks. With LSTMs, GRU/biGRU we can still represent them, thanks to their elaborated forgetting mechanisms. The probability prediction block foresees symbols on a per byte basis we work with 24 Bytes sequences.

3.1.2 Arithmetic Encoder Block

We provided an extensive description of arithmetic encoding in Sect. 2. The difference to classical compression codecs is that symbol probabilities are not feed from a deterministic calculation of symbol occurrences but come from the trained neural network. Because the arithmetic encoding model differs from symbol to symbol, we could almost talk about an adaptive encoder. Yet, we are not using one model that evolves over time, but several finite models derived from each layer of the neural network. In that sense, it is more adequate to talk about a sequence of finite arithmetic coders.

3.1.3 DeepZip Compression

Overall the DeepZip compression functions as follows: (a) starting from the trained neural network and an arithmetic coder assuming a uniform symbol distribution, the first symbol is encoded. (b) For the next symbol, the neural network can provide more information, it knows the probability of the current symbol given the previous symbol. The current symbol is encoded with an optimal arithmetic coder. (c) Step (b) is repeated until the data row is encoded. (d) The encoding continues as described in the previous steps for the totality of the data rows. The process is shown in Fig. 4, we assume the symbol sequence $S^N = \{S_1, S_2, ..., S_N\}$.

Keep in mind that with every trained encoder, a matching decoder must be trained, so the data can be recovered from the compression. In the paper, the authors evaluate their method on real world, freely available datasets. Although, the most intriguing part for us was the performance on pseudo-randomly generated number sequences. In Table 1, the results are shown.

Here it becomes obvious, where standard generic compressors fail at capturing data distributions, DeepZip can learn even complex structures such as randomly generated numbers. As a result, the compression is drastically improved.

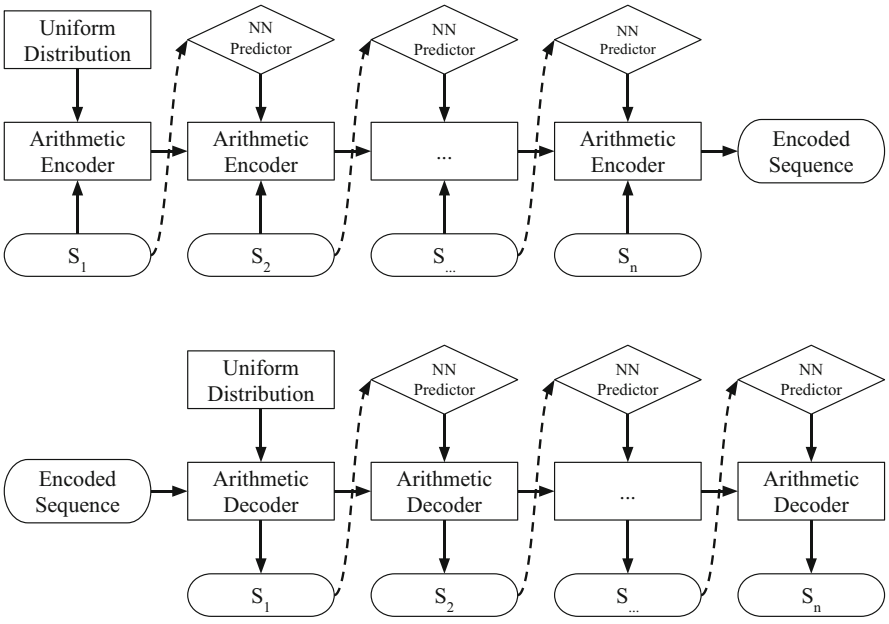

Fig. 4 DeepZip compression and decompression frameworks and process. The top panel illustrates the encoding, where we start from a uniform distribution and the neural network probability prediction block. The lower panel illustrates the decoding process. Inspired by (Goyal et al. 2018). Source: author

Table 1 Results of different DeepZip compression scenarios

Dataset	Seq. size	Gzip	BSC	DeepZip		
				FC	biGRU	LSTM-multi
IID	10	0.81	**0.60**	0.98	0.76	1.20
XOR20	10	1.51	**0.06**	0.40	0.18	0.63
XOR30	10	1.51	1.26	0.40	**0.18**	1.87
XOR40	10	1.49	1.26	**0.40**	1.43	1.87
XOR50	10	1.48	1.26	0.40	**0.18**	0.63
HMM20	10	1.49	0.87	0.98	**0.76**	1.87
HMM30	10	1.49	1.26	0.98	**0.76**	1.21
HMM40	10	1.49	1.26	**0.98**	1.42	1.87

Data volumes are in MBytes. IID: Independent and identically distributed random variables, HMM Hidden Markov model and XOR, the number corresponds to the entropy rate. Without going into details, the entropy rate describes the complexity of the probabilistic model. Source: (Goyal et al. 2018). The bold values indicate the best performance for the respective dataset

3.1.4 Thoughts on DeepZip

We experimented with DeepZip, and obtained good to excellent compression results, depending on how large the time windows for training were. The aspect hindering an adoption for IoT use cases is the required time to train a proper model. On the several experiments the authors made with their datasets, a single training epoch takes between 1 and 2 h. Models are trained for 4 epochs. The experiments from the paper were conducted on a rather powerful GPU (12 GBytes Nvidia TITAN X). On the opposite, we experimented with the NVIDIA Jetson, a credit card sized PC for the use with remotely located IoT devices. We trained the models for a day; however, we realized that for our data we reached a satisfying convergence faster, probably because our data has more regularity than the randomly generated data from the DeepZip paper. The fact that the compression is lossless however makes this tool a solution to consider for the archiving of large amounts of sensor data.

3.2 Classification and Anomaly Detection on Lossy Compressed Data

The final topic we want to address in this chapter is the compression we more or less stumbled upon during our work on (Falk et al. 2017). We had to transfer sensor data to a remote server for the evaluation by our models. We are also in a setup where the transmission costs money or at least considerably impacts the data plans of users.

The problem we solved in this paper is the recognition of a smartwatch user based on his movement data, recorded by IMU sensors. From the different degrees of freedom (DoF) such as accelerometer, gyroscope, and magnetometer, we extracted quaternions. Quaternions give a precise description of the device orientation. They are composed by four double precision floating numbers contained in the interval $[-1; 1]$. Our approach was to apply tensor decomposition for the anomaly detection and classification.

3.2.1 Tensor Decomposition for Natural Compression

When using tensor decomposition, the data is represented as a rank n tensor. A matrix, for example, is a rank 2 tensor, in three dimensions we are taking about a rank 3 tensor. In Fig. 5, we see a rank 3 tensor decomposition. The vectors on the different axis are called fibres.

As can be seen in Fig. 5, a tensor decomposition described in simple terms is a matrix factorization but at higher dimensions. This way we can create a signature tensor of normal behaviour. Novel data is fit into tensor shape. Using mathematical distance measures, the deviation from the signature tensor is computed. Decomposing tensors is an iterative process where fibres are factorized. We will not focus on that process, our intention was to notify the reader that it is not sufficient to fit data points in a tensor. Using this methodology, we could solve the task of identifying smartwatch users with high accuracy, and in a second use case detect anomalies with typical environment monitoring sensors, for example, temperature, moisture, and atmospheric pressure. These measures are relevant for a variety of use cases where conditions are monitored on remote sites with bad connectivity, such as in cargo transports and construction. Without going into more details about the machine learning aspects of tensor decomposition, we move on and show you how we achieve compression. Actually the compression is a step in the data to tensor pre-processing. We will show you how we add context to the equation to achieve a trade-off between prediction accuracy and compression ratio.

Tensors have a predefined format. There are only as many free slots in a tensor as agreed at the instantiation. As we said earlier, our IMU motion data consists of 4 features where each value is in the interval I, $I = [-1; 1]$. This requires us to map data into bins. How large the said bins are will impact the accuracy of the predictions and the compression ratio. This allows users to tweak the model for optimal results on both aspects. The set of features is discretized with a certain span. If this span s is $s = 0.1$, it means the interval $[-1; 1]$ is now treated as several subintervals for which the union equals I, such as $I = I_1 \cup I_2 \cup \cdots \cup I_{20} = [-1; -0.9[\cup[-0.9; -0.8[\cup \cdots \cup [0.9; 1]$. For each of the subintervals, the number of values falling into that range is counted. The amount of considered data rows is static. The aggregated sums compose the tensor $\mathscr{X} \in \mathbb{R}^{M_1 \times M_2 \times M_3 \times M_4}$ with M_n representing the nth feature. If the subintervals are decomposed with a 0.1 span,

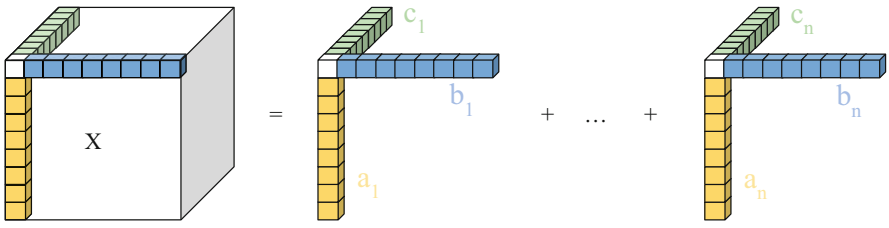

Fig. 5 Example of a tensor decomposition. The Tensor X is decomposed into a sum of n factorized fibres. Source: author

the resulting tensor has a size of $20 \times 20 \times 20 \times 20$. A step length of 0.04 would result in a $50 \times 50 \times 50 \times 50$. The first option yields better compression but is less precise in their predictions, whereas the second option provides more accuracy in predictions but requires more space.

Let us consider the below data rows with an interval span set to 0.5, the data rows with two features are both in the interval [0; 1]. The way they are arranged is shown in Fig. 6.

As can be seen, we are representing six 8 Bytes values with the rank two tensor shown in Fig. 6. With the 2×2 rank 2 tensor (matrix), we require only to represent $2 \times 2 = 4$ integer values. What we need to communicate in order to use our tensor is the following: (a) the shape of the tensor, at most 4 Bytes (one integer) per rank. In most of the cases we have seen so far, 1 Byte was sufficient since we can represent 256 values to be used as dimensions (i.e. 256 columns). (b) The span size with which values are discretized. As we will see, we implemented a flexible mechanism in this regard for a use case diverging from the IMU usage, but the results were not optimal. (c) The number of data rows contained in the tensor. We decide to communicate this configuration information in a sort of a handshake when network nodes initiate the connection. All the subsequent communication is only transferring the important values because both sides agreed on the model beforehand. In our example case, with all parties aware of the context, we can represent a single tensor with $4 \cdot 2 = 8$ bits $= 1$ Byte instead of $4 \cdot 8 = 32$ Bytes. This is because we know that in one tensor we can have at most three rows, so we need to represent four symbols at most, namely $\{0, 1, 2, 3\}$, which requires 2 bits per symbol. By that we achieve a compression ratio of $32/1 = 32$.

Let us now consider our second usage scenario, which is the monitoring of environmental conditions. Here, we leverage additional context information to achieve good compression. We are using the Bosch BME680 (Bosch 2015) sensor unit, reporting metrics such as temperature, pressure, and humidity. Let us consider the case in which we want to monitor all the metrics, with the temperature being of particular interest. From the data sheet we learn that the operation range of the temperature sensor is between -40 and $85\,^{\circ}$C. So firstly the integer values to represent are 126 symbols, hence 7 bits. The second crucial information from the data sheet is that the measurement resolution is of 0.01. It indicates us that the sensor is precisely up to two decimal digits, we need to represent 100 symbols for the decimals (at most .99). It does not make sense to literally waste 8 Bytes for a

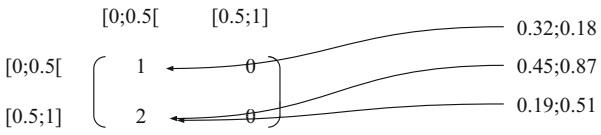

Fig. 6 Three data rows are represented in the rank two tensor. The column corresponds to the first feature, and the rows to the second feature. The total number of occurring data rows are the values at the said indexes. Source: author

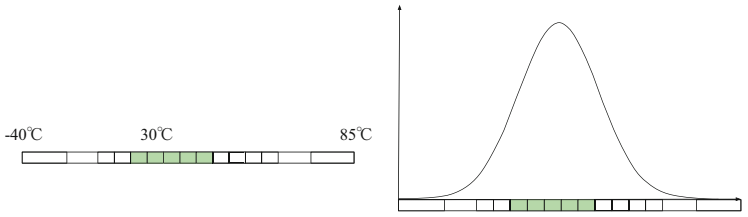

Fig. 7 Different bin sizes over the data value space lead to overfitting the data distribution. Source: author

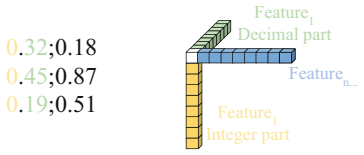

Fig. 8 Feature to dimension mapping illustrating the splitting of integer and decimal parts. Source: author

double precision floating point number, if the required information requires at most 14 bits. With this additional knowledge, we discretized all the sensor measurements into tensors as described before. Yet for the use case at hand, it was tolerable to lose precision, although for certain ranges a higher precision in the transmitted data was required.

Our first intuition was to use different discretization spans over the total data area. For example, we are less interested in the data at the edges ([−40 °C; 85 °C], but for the data around the 30 °C mark. Hence we choose different interval sizes depending on the area, as shown in Fig. 7 (left panel).

The problem with this approach was that by doing this our approximation of the normal distribution of the data was skewed (see Fig. 7 right panel). We probably underfitted our model, because with smaller bin sizes in the centre the data was more equally distributed in the tensor. For this reason, we examined another approach which was to split numbers into a integer part and a decimal part, each of which represented as one dimension of the tensor (see Fig. 8). This way we achieved a precise anomaly detection with outstanding compression ratio.

With the said feature to dimension mapping, we can maintain precise predictions and compression ratio. In this regard, we are currently investigating two different angles. One topic is an automated method to determine the best mappings and allocations for optimal prediction precision and compression ratio. The second angle we investigate is a descriptive language with which manufacturers can describe the properties of their sensors, so ideal data compression schemes can be derived. We believe that these steps are a necessity since evidence proves that data volumes increase faster and faster. This impacts storage space on disks but also bandwidth in real-time networks.

4 Conclusion

To conclude: will AI help us to solve the data deluge by storing and transferring data in more compact formats? We would say yes, but not for now. We presented one method purely based on AI by employing neural network methods, and our approach in which we leverage context information and robust machine learning methods to achieve compression. Both of the proposed methods align themselves in the ongoing feud between compression speed and compression performance.

Where DeepZip can score with exceptional compression ratios on numeric data, training of the models is slow and tedious. The methodology is not close to ready for widespread adoption. Still, it is worth investigating the case of data archiving with DeepZip. Commercial solutions already exist. Compression AI (Francis and Tissera 2018) has picked up the concept for their products.

On the opposite with our methodology, compression speed is there but it comes at the price of compression quality. Although the compression ratio is good, the lossy nature of the method makes it unusable in cases where the precise source data is required. We use it in combination with proper raw value logging as backup. For the usage in our scenarios the method performs well, of course because the machine learning method of tensor decomposition is robust enough to cope with the information loss. This method is suitable for real-time analysis.

We hope we could direct the reader's attention to data compression and that sooner than later generic compression codecs will be overwhelmed. With simple measures data storage usage can be optimized. Maybe more importantly network bandwidth can be unburdened which is particularly important with nascent technologies such as IoT, self-driving cars, augmented reality, and blockchain.

References

Apple. (2015). LZFSE compression library and command line tool. Retrieved December 19, 2019, from https://github.com/lzfse/lzfse

Bosch. (2015). Bosch BME680. Retrieved December 19, 2019, from https://www.bosch-sensortec.com/bst/products/all_products/bme680

Collet, Y. (2015). Zstandard - Real-time data compression algorithm. Retrieved December 19, 2019, from http://facebook.github.io/zstd/

Duda, J. (2013). Asymmetric numeral systems as close to capacity low state entropy coders. CoRR, from https://arxiv.org/abs/1311.2540/

Falk, E., Charlier, J., & State, R. (2017). Your moves, your device: Establishing behavior profiles using tensors. In *Advanced Data Mining and Applications - 13th International Conference, ADMA 2017*.

Francis, D., & Tissera, M. (2018). *Compression AI.* Retrieved December 19, 2019, from https://compression.ai/

Goyal, M., Tatwawadi, K., Chandak, S., & Ochoa, I. (2018). DeepZip: Lossless data compression using recurrent neural networks.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE, 40*(9), 1098–1101.

Kim, P. (2015). Elasticsearch storage overhead. Retrieved December 19, 2019, from https://www.elastic.co/blog/elasticsearch-storage-the-true-story

Milo, T. (2019). Getting rid of data, vldb2019 keynote talk. Retrieved December 19, 2019, from https://vldb.org/2019/?program-schedule-keynote-speakers#Keynote_2

Rhatushnyak, A., Wassenberg, J., Sneyers, J., Alakuijala, J., Vandevenne, L., Versari, L. et al. (2019). Committee draft of JPEG XL image coding system.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review, 5*(1), 3–55.

Witten, I. H., Neal, R. M., & Cleary, J. G. (1987). Arithmetic coding for data compression. *Communications of the ACM, 30*(6), 520–540.

Ziv, J., & Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory, 23*(3), 337–343.

Digital Transformation in Plastics Industry: From Digitization Toward Virtual Material

Christopher Stillings

Abstract

Digital transformation occurs by digitalization or digitization, simply meaning the conversion of information into a digital data format. Up to 90% of all the digital data available today is estimated to been generated just over the past 2 years! At the same time, the processing power of computers is increasing exponentially, with the result that existing data can be processed in entirely new ways. Computational power increases further with development of new technologies, e.g., quantum computing. Today we are talking about the digital revolution or digital era, also in the sense of a technological paradigm that fuels innovation and affects society and economy, leading to digital transformation in all sectors. The chapter intends to first give some insights on the impact of digital transformation and the rising opportunities in the plastics industry, using partially the example of Covestro as a global supplier of plastic material and chemicals. It furthermore relates to the increasingly relevant topic of virtual materials. In the second part, this chapter provides some suggestion in how to explore, exploit, and experiment, proactively taking part in the digital transformation, again using Covestro as an example.

1 Introduction: What Is an Innovative Technology?

In which way digitalization and digital transformation can and may impact an industry that is the opposite to virtual in the sense of its main product by definition been a very physical, tactile, and tangible one? The intention of this

C. Stillings (✉)
Covestro Polymers (China) Co., Ltd, Shanghai, China
e-mail: christopher.stillings@covestro.com

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_19

chapter is to share the perspective on digital transformation with a focus on plastic materials and related technologies, respectively the chemical and plastics industry as such using where applicable the example of Covestro as a global player in this industry. My personal perspective and interest are driven by me being a material scientist and innovation manager by education and having gained academic and business experience in the area for over 20 years. It is not the intention to offer a comprehensive and scientific research-based review on the topic, thus it is to be seen as the elaboration of personal thoughts and opinions without targeting at all completeness and taking plastics industry and the case of Covestro as an example but not articulating any message or opinion representing the company. Joining my fellow coauthor in this book Michael Krause, I can only echo his statement that the plastics industry is undergoing transformation and change. In this chapter I will focus on how digital transformation based on innovative Information and Communication technologies (ICT) is changing drastically the plastics industry driving digitalization toward virtual materials.

Relating to the title of this book "innovative technologies," I start with a definition of what actually is an innovative technology in my understanding and to elaborate on why the broader set of technologies summarized under Information Communication Technologies (ICT) has been chosen as an innovative technology in this chapter.

An innovative technology in my understanding is a technology that due to its rather recent (novelty as a dimension of innovation) development in at least a certain aspect triggers or enables an innovation. Singular individual technology-based innovations can be interconnected in technology systems, the latters furthermore can be interconnected into a technological revolution. A *technological revolution* according to Perez (2009) can be defined as a set of interrelated (often radical) breakthroughs, building a major set of interdependent technologies, or so to say a system of systems. ICT, often defined as the fifth industrial revolution, opened such a first technology system around semiconductors (material-based) and microprocessors. This initiated the formation of specialized suppliers and initial adoption in, e.g., calculators, gaming, and digitalizing of control panels, driven further more by miniaturization. This also gave rise to further increasing production of digital data. Processing and handling of this increasingly accessible and produced data led to an overlapping sequence of minicomputers and personal computers, software, telecoms and Internet that have each opened on their own new system's trajectories in this system, initiating the formation of systems in system all being strongly interrelated and interdependent. According to Perez (2009), Five such meta-systems can be identified since the initial "Industrial Revolution" originating in England, followed by the age of steam and railway as a second, the age of steel, electricity and heavy engineering as third, the age of oil, automobile and mass production as fourth and finally the age of ICT, often seen as fifth industrial revolution.

According to Perez (2004) each new technology system not only modifies the business space but also the institutional context and even the culture (as disposable plastics did in the past and Internet and Internet of Things does now). New user behaviors and regulations are likely to be required and developed, as well as new

competence building and other institutional facilitators (potentially disrupting the established ones). Each can be seen as inaugurated by an important technological breakthrough that opens a new paradigm of opportunity for innovation. A good example for this is the case the microprocessor, initiating the ICT revolution. The massive and increasing impact of the Information Revolution and the visibly increasing importance of innovation and entrepreneurship (e.g., IT start-ups) triggered a great interest in Schumpeterian ideas. In his definition of innovation Alois Schumpeter is one of the few economists who puts technical change and entrepreneurship at the base of economic growth. Schumpeter clearly differentiates between *innovation*, seen as the profitable commercial introduction of a new product or service, and *invention*, which relates to the fields of research and development or science.

The ICT revolution, understandable best as system of systems of technology which are interdependent and present various feedback loops and their role in innovation in the sense of profitable growth can thus be seen as truly innovative technology topic and furthermore some of the underlying technologies have been chosen as the subject for this chapter in the sense of enabler for the digital transformation in plastics industry. ICT is massively transforming all sectors and can be understood as also a technical economic paradigms, which according to Perez (2004) potentially fuel or imply also socioeconomic changes as mentioned above.

Whereas in the past megatrend(s) related to information and communication technology have been a lot formulated from a rather pure technology point of view (Internet of Things, supercomputing, robotics, etc.) or the adjacent area of applying the technology (e.g., dataism, big data, online and real time, etc.) on can observe recently that the megatrend is including much more the output or impact of ICT application and diffusion in all sectors (Z-punkt 2019; TrendOne 2019). The term "Digital Transformation" clearly indicate this.

2 The Perspective of Material Science

Material Science is a highly interdisciplinary field, including natural science-based disciplines and engineering among others. Concerning the megatrends impacting society, economy and environment new and improved materials are seen as key to solutions to cope with challenges of humankind nowadays and in future. New and advanced materials as well as related synthesis and production processes are expected to increase efficiency and effectiveness in production and usage over the whole product life cycle and beyond. In order to achieve this, the field of materials science relies nowadays on experiments and simulation-based models to understand the structure property relationships of different materials and their characteristics better and ideally in a fundamental way. The overall target is the discovery of new materials with improved or tailored properties and specific applicability, respectively efficient production and processing processes as well as recyclability.

Following the description of Agrawal and Choudhary (2016), in the beginning, for thousands of years, curiosity and belief led to experiments and further more

to science. Historically science was empirical, and in early stage it followed more or less the trial and error approach. A few centuries ago came the paradigm of models, theories, and concepts in order to formulize of "laws" and concert these into mathematical equations. As science progressed and scientific problems became more complex, also the theoretical models turned to reach limits to deliver an analytical and accurate solution. Some decades ago, with the surge of semiconductors and computers as well as the relevant software developments the paradigm of computational (based or aided) science become more and more popular. The ever-increasing calculation performance of computers fueled by dynamic development of chips and their capacity allowed the computational simulation of complex real-world phenomena based on the theoretical models of the second paradigm (e.g., molecular dynamics simulations). Today these are popular as the branches of theory, experiment, and computation in almost all scientific domains, scientific methods and analytics as well as scientific results have been more and more digitized. The latter obviously leads to an ever-increasing amount of (digital) data which initiated a fourth paradigm of science over the last years—(big) data-driven science. It unifies the previous three paradigm (experiment, theory, and computation/simulation) and is becoming increasingly adopted in material science leading to a new field of what Agrawal and Choudhary (2016) call materials informatics.

The field of materials informatics can still be considered as in its early and emerging stage, much like what bioinformatics and genomics were 20 years ago. Interdisciplinary collaborations bringing together competencies and expertise from (analog) materials science and (digital) computer science. Thus, building an interdisciplinary skilled workforce are crucial to leverage the opportunities materializing and to enable timely discovery and application of new advanced materials for the benefit of mankind and solutions for the challenges mankind and the planet are facing.

In many areas of materials science in the past, there was more of a lack of data than a "big data" issue. Open, accessible and high quality data has been rather limited to be sourced, similar to the accessibility and usability of modern simulation tools and related. The Materials Genome Initiative is one of several examples of initiatives that are supporting and promoting around the world the availability and accessibility of digital data and the relevant tools in materials science. The activities include combining experimental and simulation data into a searchable materials data infrastructure and encouraging researchers to make their data available to the community. A subproject for example is the Materials Project (2019), which is leveraging the power of increasingly available supercomputing together with latest quantum mechanical theory to compute the properties of all known inorganic materials. The intention is to design novel materials apart of making the data available to the community, adding online analysis and design algorithms. Meanwhile it contains data for more than 70,000 materials and millions of associated materials properties (Jain et al. 2018).

Improvements in computational resources over the last decade are enabling a new era of computational characterization, prediction, and design of novel materials. This will further more ad to the (big) data pools or "lakes" as recently called and

been build up on materials and processing data. The next but already present frontier is leveraging technologies and methods related to artificial intelligence and machine learning to harnessing the increasingly available data for automated learning and accelerated as well as more accurate discoveries. On artificial intelligence and machine learning please also refer to the related chapter in this book written by Patrick Glauner. Nevertheless analog experiments in particular in materials science will be needed to proof predictions and to transfer between the analog and digital world accordingly.

3 The Perspective of Covestro

Covestro is among the world leading suppliers of premium polymers. Covestro's materials and application solutions are found in nearly every area of modern life. Innovation and sustainability are the driving forces behind the continuous development of Covetsro's products, processes, and facilities. The backbone of its organization are in total 16,800 employees, who work at around 30 sites across the globe—from smaller technical centers to innovation hubs to large-scale production plants. All activities are coordinated from the corporate headquarters in Leverkusen, Germany. Covestro's core business comprises three segments that produce and continuously advance raw materials for polyurethanes and their derivatives, the premium plastic polycarbonate as well as coatings, adhesives, and other specialties. Covestro develops sustainable solutions to the greatest challenges of mankind: climate change, resource depletion, urban expansion, population growth, and the resulting increase in awareness of environmental issues.

The ITC-related technologies and innovation trends are opening up new options and opportunities for any industry, thus also for the plastics industry. Covestro addresses this in three dimensions of its activities: processes (both business and chemical), digital customer experience(s), and further more digital business models. In the following will describe these three dimensions in more detail and give some concrete examples (Covestro 2019).

- **Digital Processes for Internal Operations**

  As the chemical production and the related sites obviously are crucial for chemical and plastic manufacturing players, digitalization of the related process and the site's infrastructure including logistics is one of the biggest initiatives related to this dimension. The global *Optimized System Integration 2020* project targets digital operating processes in production. Its goal is to make the design, operations, and maintenance of global production plants more efficient and transparent. This is to be achieved within the next 3 years by means of data integration, coupled with the new thought processes and operating procedures associated with it. Predictive maintenance is to be carried out in the plants using mobile devices that deliver real-time data. Further digitalization of the production facilities will make planning, operation, and maintenance much easier.

The so-called *predictive maintenance* of systems, for example, becomes even more reliable in combination with machine learning and artificial intelligence. This is shown by a pilot project of the company. The temperature and vibration sensors installed in a large engine of the production plant transmit their collected data on the condition of the engine during operation to software. This enables to predict possible engine failure 8 months in advance. The aim is to be able to intervene precisely in the production processes on the basis of a clear presentation of all information and thus continuously optimize them. To this end, Covestro comprehensively analyzes data from ongoing production and maintenance in order to be able to assess the behavior of machines and materials in advance and make appropriate recommendations. The system learns automatically. The Integrated Plant and Engineering Platform (IPEP) creates a virtual data model and a *digital twin* of each production plant. The entire technical documentation of each plant is brought together in digital form in this type of database. This will benefit all production employees. IPEP will enable to work even more securely and efficiently in the future and to access all data quickly and easily.

- **Customer Experience—Communicating with Customers on All Channels**

  Covestro is also leading the chemical industry with its plans for fully integrated digital communication with business customers. They are to receive more effective after-sales support, from the first product idea to ongoing service, above and beyond all the digital channels. In 2018 the first milestone was a new internet presence with an enhanced product search function. Covestro is continuously enhancing its digital offerings to customers supported by intelligent analytics right where they are looking for solutions to their business challenges. Additionally, the company is increasingly making usage of automated marketing solutions and social networks such as WeChat, Facebook, LinkedIn & Co.

  Apart from the communication through digital channels and conducting (alternative) business or transaction processes by digital tools to create a digital customer journey and related experience, the challenge is how to virtualize the main element of the product, the material itself and being as analog and physical as it can be. A good example on bringing a virtual material-based experience to the customer is to introduce the digital twins approach not only for sites and devices but for material itself. One approach is the total appearance capture technology by X-rite Pantone that is been presented as an example in this chapter (see further proceedings in this chapter). Beside an increased digital customer experience of an analog material, the analogy to a pdf concerning written and picture content makes a digital data format such as AxF simulating the appearance of a material and the ability to usage in different rendering, simulation and design related software solutions potentially beneficial for improving speed and efficiency of workflows, higher quality in communication and ultimately interesting for new services and eventually business models, basically targeting to realize the concept of a digital twin of the material.

- **New Business Models Focus on the Customer**

  In addition to the digital commerce platform, there is for example a new business model called “digital technical services” that is critical to support

efficient production processes for customers. Together with customers in the foam manufacturing industry, Covestro has been gaining experience for 10 years now with analyzing data to optimize production conditions. Algorithms are now supporting the expansion of these services. Customers will be in a position as a result to significantly cut production costs and operate more efficient and reliable.

A current emphasis of the new business models is the digitalization and optimization of process flows. In the simulating process steps, development times at customers and along value chains can be reduced considerably, and process flows can be designed more efficiently. With an easy-to-use web-based calculation tool, customers can enter the desired physical properties of the foam and wait for the matching formulas to be calculated based on our raw materials. To develop the digital tool, an interdisciplinary team at Covestro first manufactured various viscoelastic foams with the aid of predefined formulas and identified their properties. Based on these data sets, the team then generated an algorithm, which uses the properties of these foams to calculate other foam densities, hardness levels, and viscoelastic behaviors.

4 Virtual Customer Experience of Materials

Virtual customer experience is seen as one of the leading (mega) trends related to ICT according to the trend research consultancy TrendOne (TrendOne 2019). Virtual created spaces and simulations feel close and real when accompanied not only by sounds and images, but also enhanced by olfactory and haptic elements. Innovative input devices allow more and more to explore virtual situations and stories, as well as interact in them with others or even with the content. Experiences worth remembering are then no longer restricted to reality. Smartphones and special head-mounted displays like the Oculus Rift and Microsoft HoloLens are the first generation of devices to open the gateway to immersive worlds. But it remains a challenge already to "simulate" in virtual space in most realistic way the analog world (Guarnera and Guarnera 2018). Furthermore the challenge for a material supplier is how to transfer the customer experience related to it major product—the material, by definition purely analogous—to the virtual space. Digitization of information about the material or digital photography is not adequate. Potentially augmented reality (AR) and virtual reality solutions offer cost reduction and resource economies from social interaction and entertainment, to learning and working—through a remote presence, time savings through dynamic collaboration, danger minimization through simulation and empathy through immersion.

Capturing the material appearance data and using it to create customer experience in the digital or virtual space is a challenge for industries like fashion, cinema, gaming, automotive, and of course for material manufacturer in particular when it comes to rendering of virtual worlds in 3D (Fang 2011). The capture of all relevant data in a single, editable, portable file format is an obstacle in the virtualization of products, especially when consistency in appearance is required. AxF is the a format designed for system-independent communication of digital appearance and

has been introduced by X-rite Pantone (2019). In this sense in the following part I will use it to showcase innovative digital technology as an example of how digital transformation and related technologies are affecting furthermore plastics industry.

Headquartered in Grand Rapids, Michigan, X-Rite Pantone is a global company with locations around the world. Experts in combining the art and science of color, focus on providing complete end-to-end color management solutions in every industry where color matters. In addition Pantone provides color systems and leading technology for the selection and accurate communication of color across many of industries. The company has around 800 employees and 17 offices worldwide. Founded in 1958, X-Rite was actually created by data-driven color scientists. Based on the technology developed by "sensible graphics" a start-up of the University of Bonn, Germany X-rite, who acquired the start-up in 2012, developed the Total Appearance Capture (TAC) system and commercialized it in 2016. Sensible Graphics is nowadays the development center for the TAC technology. In addition with the sponsorship of X-Rite the university established a Graduate School of Material Appearance and installed a new professorship focusing on digital material appearance.

The Total Appearance Capture (TAC) ecosystem, with the TAC7 Scanner at its core and the AxF file format as data format, enables designers to capture reality in a physically precise way. From special effects pigments to synthetic fabrics, it enables to capture and communicate physical appearance properties, such as color, gloss, and texture, digitally to experience a high degree of realism in 3D designs. The TAC7 scanner collects the appearance-relevant data of the materials sample to be compressed and translated by algorithm into the axf file format. The AxF stands for an universal file format that offers a way to capture, store, edit, and communicate complex color and appearance data of a material and make it accessible to a wide range of rendering, simulating, design, and engineering software solutions. AxF is used in product design, development, manufacturing, sales, and marketing. It is an industry first, and is helping brands reduce cycle time, control cost, and ensure consistency in color and appearance (Mueller et al. 2019).

To capture the exact physical appearance properties such as color, texture, gloss, translucency, and transparency in a digital format that makes it easy to experience unmatched realism in the virtual world enabling and advanced virtual material base customer experience and leading toward the realization of the concept of digital twin of a physical (analog) material. This adds an additional dimension to the topics of prediction, simulation, and characterization mentioned as digital enabled approaches in material science as well. This furthermore leads to different business models of nonmaterial supplier or apart data format and scanning hardware supplier (such as X-rite Pantone) providing data, e.g., for specific target groups such as creative industry and industrial designers (Brain of Materials 2019), processing-related materials data or offering in combination the necessary software tools and platforms to process them (Optis Ansys 2019). I presume we will see much more of this kind of service and platform (ecosystem) based business models in the near future.

5 Suggestions

As it has been stated already in the foreword of this book, this book intends to give insights and suggestions on how to invest in innovative technologies). Regarding the topics in this particular chapter, written from a view point of a material scientist and employer of a global plastic materials supplier and looking into examples of implementation and adoption of Covestro and beyond I believe that companies in many different sectors and industries are facing similar challenges and opportunities when it comes to digitalization. Challenges because new business models based on ICT paradigm have disruptive potential for any traditional players in general and opportunities because digital transformation leads to new options and opportunities for core businesses and new ventures. Thus, there is no way of not investing in the ICT technologies and their adoption mentioned in this chapter as examples. This appears to be common sense among basically all industries and common practice already, once considering the advanced and further developing adoption that can be observed.

But how to invest with which priority and timing is a challenging decision to make, as resources such as capital, knowledge (experts), and time are always limited—often scarce. Key for success to invest as a company is to holistically use the technologies for both (short term) improvement and efficiency gains, AND innovation in all areas. In the case of Covestro this is realized by the three dimensions approach plus the initiatives concerning digital R&D. The challenge is to create the set up and mindset, the culture of doing both exploring and exploiting and embracing experiments to try new ways out. The latter actually loops back in a way to the empirical age of science and key to success for former innovation driven by chemical industry.

Invest, Explore, and Exploit! On a macroscopic level to do so successfully organizational ambidexterity is needed. The term organizational ambidexterity refers to the ability of any organization to be efficient in its management of the traditional or today's business but at the same time also to be agile and adaptable with changes induced, in particular when it comes to a new transformational path or paradigm—just such as the ICT paradigm and the digital transformation. Similar to being ambidextrous (meaning to be able to use the right and the left hand equally) this in the case of any organization requires use both exploration and exploitation approaches to be successful. Pretty much it can be seen as a holy grail of modern management, in particular innovation management.

Using Covestro as an example but considering general applicability, organizational ambidexterity can be fostered by different approaches and initiatives. The following suggestions should give an idea but of course applicability depends on the organizational and industry context (and culture).

- Cross-Functional, Community, and Ecosystem Approach in Workstyle

  Looking into the challenges that come along with interdependencies of technologies and related system and the increasing complexity of processes

cross-functional collaboration becomes crucial for efficient integration but also for any new approaches to be adopted. The later requires total system analysis which is enabled by internal and external collaboration. Entrepreneurial activities more and more also lead to ecosystems, e.g., in the start-up environment, thus open innovation and external collaboration is crucial.

- Intrapreneurship Initiatives and Entrepreneurial Mindset

  Coming back to earlier definition of innovation by Schumpeter and his focus on the role of the entrepreneur as driver of innovation, the entrepreneurial mindset and the organizational options to address intra and entrepreneurial activities in the company becomes important for topics and technologies' adoption that are more distant form the today's (core) business. Covestro addresses this through its "start-up challenge" wherein employees can suggest an entrepreneurial activity and topic as a team and are trained to formulate a business plan. One team is selected and gets awarded capital and time to realize the project. In addition internal ventures with respective organizational settings are in place. Another option are internal incubators and/or the collaboration with external incubators and start-ups. Furthermore Covestro collaborates on the Business Model Think Tank approach initiated by the University of St. Gallen (BMI Lab 2019).
- Design Thinking as Mindset and Methodology

  Design thinking puts the customer in the focus and aims to design the customer experience and to define the customer's problem first, then to work out the solution in the sense of a product. It is based on iterative approaches focusing on learning cycles by rapid prototyping and utilizing the concept of minimum viable product to be tested regarding the related assumptions. Training employees in design thinking (at Covestro this is called We create) and building a related culture and mindset is an important element of investing further in activities in digital transformation, as customer centricity and repaid learning cycles are essential for realization and exploration.
- Trend Watching and Foresight—Thinking in Scenarios and Options

  As pointed out, digital transformation is driven by a set of innovative technologies in many different fields. It is a dynamic and complex field regarding what actually is happening in development and in terms of adoption and innovation and on top the impact on many areas of business and ways of doing business is expected to be high and not always predictable. Apart from trend watching, scenario management is recommended as toolset for any kind of company. The adoption of innovative technologies in the ICT requires contextual thinking and the thinking in options, thus scenario management is viable tool to foster digital transformation. The future development is not foreseeable, to think in different scenarios enable an organization in effectuation beside the traditional causal logic. In addition it helps to look further ahead, creating ideas for new options and opportunities as well as been supportive for formulating and communicating a related vision internally and externally.
- Infrastructure and Hardware

  Investing means in a traditional way also investing in assets. Similar to other players in the industry Covestro is also increasing the processing capacity and

is investing in advanced hardware as well as IT infrastructure and software including supercomputing.

- Experiments (!)

  Giving employees and partners the time and limited resources to conduct experiments as well as encouraging peoples curiosity and confidence to do experiments on where to apply digital technologies and tools in their area or adjacent is one of the key drivers to benefit from any kind of new technologies and to make them innovative. Covestro has curiosity and courageousness as key elements of its company culture.

6 Conclusion

As work makes up for a good portion of our personal and social life time and of course in many way affects the private life also outside working hours it becomes clear on how the digital transformation truly can be seen not "just" as technical revolution but as a sociotechnical economic paradigm as stated by Perez. One can observe improvements, disruption, and new opportunities basically in all sectors: science, economy and society and politics. For an material scientist by education and innovation manager in chemical industry it is an interesting field to invest with engagement, openness, and the experimental mindset to explore. For any chemical and materials manufacturer it has a great potential to innovate on product, service, and business model dimension and furthermore can be seen as an enabler to drive sustainability by system integration and efficiency gains in all areas of activities. Thus totally worthy to invest time, and capital and continue experimenting, learning and innovating.

References

Agrawal, A., & Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials, 4*, 053208.

BMI Lab. (2019). *Think Tank*. Accessed December 27, 2019, from https://bmilab.com/think-tank

Brain of Materials. (2019). *Home page*. Accessed December 27, 2019, from https://www.brainofmaterials.com/

Covestro. (2019). *Digitalization*. Accessed December 27, 2019, from https://www.covestro.com/en/innovation/digitalization

Fang, J. (2011). An analysis on virtual material making and its application in 3D animation designs. *Advanced Materials Research, 211–212*, 1172–1175.

Guarnera, D., & Guarnera, G. C. (2018). Virtual material acquisition and representation for computer graphics. *Synthesis Lectures on Visual Computing, 10*(1), 1–101.

Jain, A., Montoya, J., Dwaraknath, S., Zimmermann, N. E. R., Dagdelen, J., Horton, M., Huck, P., Winston, D., Cholia, S., Ong, S. P., & Persson, K. A. (2018). The materials project: Accelerating materials design through theory-driven data and tools. In W. Andreoni & S. Yip (Eds.), *Handbook of materials modeling*. Cham: Springer.

Materials Project. (2019). *Home page*. Accessed December 27, 2019, from https://materialsproject.org/

Mueller, G., Tautges, J., Gress, A., & Lamy, F. (2019). AxF – Appearance exchange Format, Version 1.6, X-Rite, Inc., 4300 44th St. SE, Grand Rapids, MI 49505.

Optis Ansys. (2019). *Overview*. Accessed December 27, 2019, from http://www.optis-world.com/OPTIS-revealed/Overview

Perez, C. (2004). Technological revolutions, paradigm shifts and socio-institutional change. In *Globalization, economic development and inequality: An alternative perspective*, pp. 217–242. Edward Elgar.

Perez, C. (2009). Technological revolutions and techno-economic paradigms. In *Working papers in technology governance and economic dynamics.* The Other Canon Foundation.

TrendOne. (2019). *Home page*. Accessed December 27, 2019, from https://www.trendexplorer.com/en/

Xrite Pantone. (2019). *Home page*. Accessed December 27, 2019, from www.xrite.com

Z-punkt. (2019). *Connected reality*. Accessed December 27, 2019, from http://www.z-punkt.de/en/studien/studie/connected-reality-2025/55

Correction to: Analytic Philosophy for Biomedical Research: The Imperative of Applying Yesterday's Timeless Messages to Today's Impasses

Sepehr Ehsani

Correction to:
Chapter 13 in: P. Glauner, P. Plugmann (eds.),
***Innovative Technologies for Market Leadership*,**
https://doi.org/10.1007/978-3-030-41309-5_13

The chapter "Analytic Philosophy for Biomedical Research: The Imperative of Applying Yesterday's Timeless Messages to Today's Impasses" was previously published non-open access. It has now been changed to open access under a CC BY 4.0 license and the copyright holder has been updated to "The Author(s)."

The updated online version of this chapter can be found at
https://doi.org/10.1007/978-3-030-41309-5_13

P. Glauner, P. Plugmann (eds.), *Innovative Technologies for Market Leadership*,
Future of Business and Finance, https://doi.org/10.1007/978-3-030-41309-5_20