Chapter 1
Data Science in Action

In recent years, *data science* emerged as a new and important discipline. It can be viewed as an amalgamation of classical disciplines like statistics, data mining, databases, and distributed systems. Existing approaches need to be combined to turn abundantly available data into value for individuals, organizations, and society. Moreover, new challenges have emerged, not just in terms of size ("Big Data") but also in terms of the questions to be answered. This book focuses on the *analysis of behavior based on event data*. *Process mining* techniques use event data to discover processes, check compliance, analyze bottlenecks, compare process variants, and suggest improvements. In later chapters, we will show that process mining provides powerful tools for today's data scientist. However, before introducing the main topic of the book, we provide an overview of the data science discipline.

1.1 Internet of Events

As described in [73], society shifted from being predominantly "analog" to "digital" in just a few years. This has had an incredible impact on the way we do business and communicate [99]. Society, organizations, and people are "Always On". Data are collected *about anything*, *at any time*, and *at any place*. Nowadays, the term "Big Data" is often used to refer the expanding capabilities of information systems and other systems that depend on computing. These developments are well characterized by *Moore's law*. Gordon Moore, the co-founder of Intel, predicted in 1965 that the number of components in integrated circuits would double every year. During the last 50 years the growth has indeed been exponential, albeit at a slightly slower pace. For example, the number of transistors on integrated circuits has been doubling every two years. Disk capacity, performance of computers per unit cost, the number of pixels per dollar, etc. have been growing at a similar pace. Besides these incredible technological advances, people and organizations depend more and more on computerized devices and information sources on the Internet. The IDC Digital Universe Study of April 2014 confirms again the spectacular growth of data [134].

W. van der Aalst, *Process Mining*, DOI 10.1007/978-3-662-49851-4_1

This study estimates that the amount of digital information (cf. personal computers, digital cameras, servers, sensors) stored in 2014 already exceeded 4 Zettabytes and predicts that the "digital universe" will to grow to 44 Zettabytes in 2020. The IDC study characterizes 44 Zettabytes as "6.6 stacks of iPads from Earth to the Moon". This illustrates that the long anticipated *data explosion* has become an undeniable reality.

From Bits to Zettabytes
A "bit" is the smallest unit of information possible. One bit has two possible values: 1 (on) and 0 (off). A "byte" is composed of 8 bits and can represent $2^8 = 256$ values. To talk about larger amounts of data, multiples of 1000 are used: 1 Kilobyte (KB) equals 1000 bytes, 1 Megabyte (MB) equals 1000 KB, 1 Gigabyte (GB) equals 1000 MB, 1 Terabyte (TB) equals 1000 GB, 1 Petabyte (PB) equals 1000 TB, 1 Exabyte (EB) equals 1000 PB, and 1 Zettabyte (ZB) equals 1000 EB. Hence, 1 Zettabyte is $10^{21} = 1{,}000{,}000{,}000{,}000{,}000{,}000$ bytes. Note that here we used the International System of Units (SI) set of unit prefixes, also known as SI prefixes, rather than binary prefixes. If we assume binary prefixes, then 1 Kilobyte is $2^{10} = 1024$ bytes, 1 Megabyte is $2^{20} = 1048576$ bytes, and 1 Zettabyte is $2^{70} \approx 1.18 \times 10^{21}$ bytes.

Most of the data stored in the digital universe is unstructured, and organizations have problems dealing with such large quantities of data. One of the main challenges of today's organizations is to *extract information and value from data* stored in their information systems.

The importance of information systems is not only reflected by the spectacular growth of data, but also by the role that these systems play in today's business processes as the digital universe and the physical universe are becoming more and more aligned. For example, the "state of a bank" is mainly determined by the data stored in the bank's information system. Money has become a predominantly digital entity. When booking a flight over the Internet, a customer is interacting with many organizations (airline, travel agency, bank, and various brokers), often without being aware of it. If the booking is successful, the customer receives an e-ticket. Note that an e-ticket is basically a number, thus illustrating the alignment between the digital and physical universe. When the SAP system of a large manufacturer indicates that a particular product is out of stock, it is impossible to sell or ship the product even when it is available in physical form. Technologies such as RFID (Radio Frequency Identification), GPS (Global Positioning System), and sensor networks will stimulate a further alignment of the digital universe and the physical universe. RFID tags make it possible to track and trace individual items. Also note that more and more devices are being monitored. Already 14 billion devices are connected to the Internet [134]. For example, Philips Healthcare is monitoring its medical equipment (e.g., X-ray machines and CT scanners) all over the world. This helps Philips to

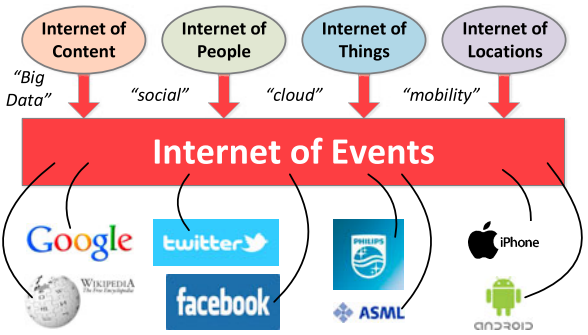

Fig. 1.1 Internet of Events (IoE): Event data are generated from a variety of sources connected to the Internet

understand the needs of customers, test their systems under realistic circumstances, anticipate problems, service systems remotely, and learn from recurring problems. The success of the "App Store" of Apple illustrates that location-awareness combined with a continuous Internet connection enables new ways to pervasively intertwine the digital universe and the physical universe.

The spectacular growth of the digital universe, summarized by the overhyped term "Big Data", makes it possible to record, derive, and analyze *events*. Events may take place inside a machine (e.g., an X-ray machine, an ATM, or baggage handling system), inside an enterprise information system (e.g., an order placed by a customer or the submission of a tax declaration), inside a hospital (e.g., the analysis of a blood sample), inside a social network (e.g., exchanging e-mails or twitter messages), inside a transportation system (e.g., checking in, buying a ticket, or passing through a toll booth), etc. Events may be "life events", "machine events", or "organization events". The term *Internet of Events* (IoE), coined in [146], refers to all event data available. The IoE is composed of:

- The *Internet of Content* (IoC), i.e., all information created by humans to increase knowledge on particular subjects. The IoC includes traditional web pages, articles, encyclopedia like Wikipedia, YouTube, e-books, newsfeeds, etc.
- The *Internet of People* (IoP), i.e., all data related to social interaction. The IoP includes e-mail, Facebook, Twitter, forums, LinkedIn, etc.
- The *Internet of Things* (IoT), i.e., all physical objects connected to the network. The IoT includes all things that have a unique id and a presence in an Internet-like structure.
- The *Internet of Locations* (IoL) which refers to all data that have a geographical or geospatial dimension. With the uptake of mobile devices (e.g., smartphones) more and more events have location or movement attributes.

Note that the IoC, the IoP, the IoT, and the IoL are overlapping. For example, a place name on a webpage or the location from which a tweet was sent. *Process mining aims to exploit event data in a meaningful way*, for example, to provide insights, identify bottlenecks, anticipate problems, record policy violations, recommend countermeasures, and streamline processes. This explains our focus on event data.

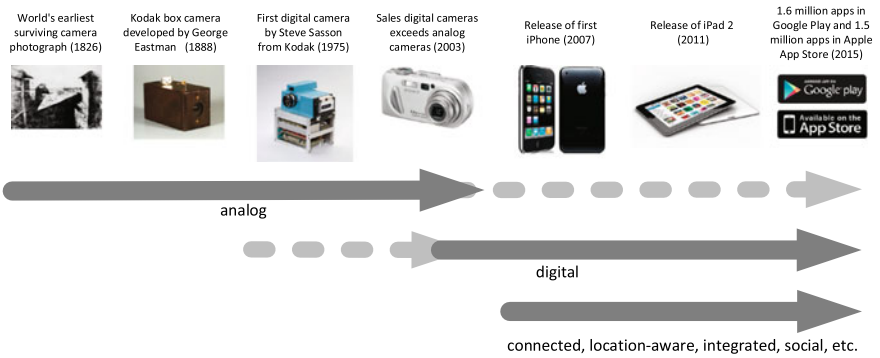

Fig. 1.2 The transition from analog to digital dramatically changed the way we create and share photos. This is one of the factors contributing to the rapid expansion of the Internet of Events (IoE)

To illustrate the above developments, let us consider the development of photography over time (see Fig. 1.2). Photography emerged at the beginning of the 19th century. Around 1800, Thomas Wedgwood attempted to capture the image in a camera obscura by means of a light-sensitive substance. The earliest remaining photo dates from 1826. Towards the end of the 19th century, photographic techniques matured. George Eastman founded Kodak around 1890 and produced "The Kodak" box camera that was sold for $25, thus making photography accessible for a larger group of people. The company witnessed the rapid growth of photography while competing with companies like Fujifilm. In 1976, Kodak was responsible for 90% of film sales and 85% of camera sales in the United States [57]. Kodak developed the first digital camera in 1975, i.e., at the peak of its success. The Kodak digital camera had the size of a toaster and a CCD image sensor that only allowed for 0.01 megapixel black and white pictures. It marked the beginning of digital photography, but also the decline of Kodak. Kodak was unable to adapt to the market of digital photography. Competitors like Sony, Canon, and Nikon better adapted to the rapid transition from analog to digital. In 2003, the sales of digital cameras exceeded the sales of traditional cameras for the first time. Today, the market for analog photography is virtually non-existent. Soon after their introduction, smartphones with built-in cameras overtook dedicated cameras. The first iPad having a camera (iPad 2) was presented on March 2nd, 2011 by Steve Jobs. Today, the sales of tablet-like devices like the iPad exceed the sales of traditional PCs (desktops and laptops). As a result of these developments, most photos are made using mobile phones and tablets. The remarkable transition from analog to digital photography has had an impact that goes far beyond the photos themselves. Today, photos have GPS coordinates allowing for localization. Photos can be shared online (e.g., Flickr, Instagram, Facebook, and Twitter) and changed the way we communicate and socialize (see the uptake of the term "selfie"). Smartphone apps can detect eye cancer, melanoma, and other diseases by analyzing photos. A photo created using a smartphone may generate a wide range of events (e.g., sharing) having data attributes (e.g., location) that reach far beyond the actual image. As illustrated by

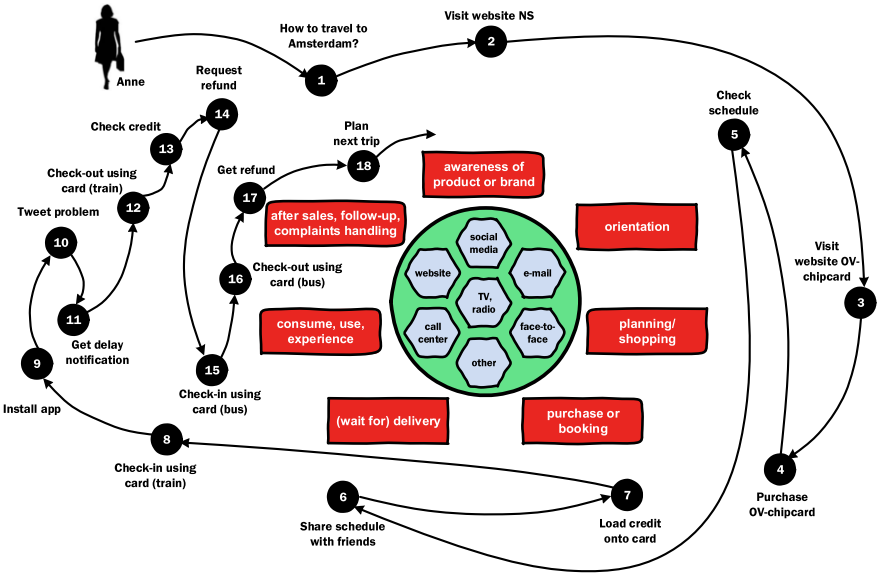

Fig. 1.3 An example of a customer journey illustrating the many (digital) touchpoints generating events that allow us to understand and serve customers better

Fig. 1.2, developments in photography accelerated since the first digital camera, and the transition from analog to digital photography contributed significantly to the growth of the Internet of Events (IoE). Digitalization resulted not just in content (e.g., photos) but also in new ways to capture "events" showing what is really happening.

Let us now consider another development: the digitization[1] of the *customer journey* where customers interact in multiple novel ways with organizations [36]. In the digital era, there are many *touchpoints* using different media. The center of Fig. 1.3 shows the different media: social media, e-mails, websites, face-to-face contacts, call-centers, etc. Although there are significant differences between the wide variety communication channels, "content" tends to become less dependent on the media used (phone, laptop, etc.). Smartphones and iPads can make photographs, computers can be used to make phone calls (through Skype or FaceTime), and customer complaints can be expressed via a website or call-center. Different devices and services co-exist in an integrated ecosystem. Consider, for example, the capturing, managing, publication, viewing, and sharing of photos using digital cameras, mobile phones, computers, websites, media-players, printers, televisions, interactive whiteboards, etc.

[1] Some distinguish between the terms digitization and digitalization where *digitization* is the process of converting an analogue reality into digital bits and *digitalization* refers to the way in which social life and businesses are impacted by digital communication infrastructures. In this book, we do not use this distinction as both are too much intertwined.

Figure 1.3 distinguishes seven stages for an archetypal customer journey:

1. *Awareness of product or brand.* The customer needs to be aware of the product and/or brand to start a customer journey. For example, a customer that does not know about the existence of air purifiers will not consider buying one. (An air purifier removes contaminants from the air in a room to fight allergies, asthmatics, or tobacco smoke.)
2. *Orientation.* During the second stage, the customer is interested in a product, possibly of a particular brand. For example, the customer searches for the differences between air purifiers, e.g., there are devices that use thermodynamic sterilization, ultraviolet germicidal irradiation, HEPA filters, etc.
3. *Planning/shopping.* After the orientation phase the customer many decide to purchase a product or service. This requires planning and/or shopping, e.g., browsing websites for the best offer.
4. *Purchase or booking.* If the customer is satisfied with a particular offering, the product is bought or the service (e.g., flight or hotel) is booked.
5. *(Wait for) delivery.* This is the stage after purchasing the product or booking the service, but before the actual delivery. For example, the air purifier that was purchased is unexpectedly out of stock, resulting in a long delivery time and an unhappy customer. Events like this are an integral part of the customer journey.
6. *Consume, use, experience.* At the sixth stage, the product or service is used. For example, the air purifier arrived and is used on a daily basis. While using the product or service, a multitude of events may be generated. For example, some air purifiers are connected to the Internet measuring the air quality. The user can control the purifier via an app and monitor the air quality remotely. The recorded event data can be used to understand the actual use of the product by the customer.
7. *After sales, follow-up, complaints handling.* This is the stage that follows the actual use of the product or service. For example, the customer may want to return the air purifier because it is broken or does not deliver the performance expected. At this seventh stage, new add-on products may be offered (e.g., air filters).

Given a particular product or organization, many customer journeys are possible. The customer journey is definitely *not* a linear process. Stages may be skipped and revisited. Moreover, customers may use many products of the same brand leading to an overall customer experience influencing future purchase decisions.

Figure 1.3 shows one particular customer journey to illustrate the different touchpoints potentially providing lots of event data for analysis. Consider a teenager (let us call her Anne) that wants to make a trip from Eindhoven Central Station to Amsterdam to visit the Van Gogh Museum. Anne first explores different ways to travel to Amsterdam (1) followed by a visit to the website of NS (Dutch railroad company) (2). Anne finds out that she needs to buy a so-called "OV-chipcard". Such a card gives access to a contactless smartcard system used for all public transport in the Netherlands. Using the card Anne can check-in at the start of a trip and check-out at the end of trip. After vising the OV-chipcard website (3), Anne purchases

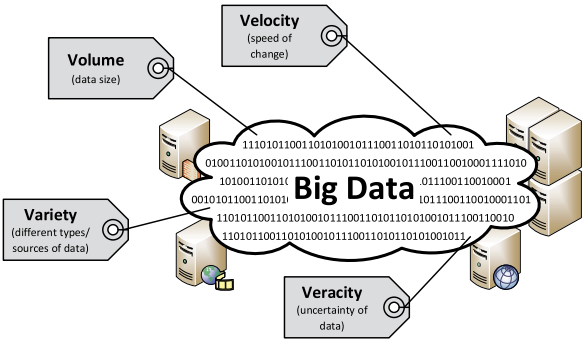

Fig. 1.4 The four V's of Big Data: Volume, Velocity, Variety, and Veracity

the OV-chipcard from a machine in the train station (4), and checks the schedule (5) using her mobile phone. She shares the selected schedule with her friends (6). Before checking in using the card (8), she first loads 100 euro credit onto her OV-chipcard (7). While traveling she installs the NS app obtained from iTunes (9). Due to a broken cable, the train gets a 90 minute delay. Anne tweets about the problem while mentioning @NS_online to express her disappointment (10). A bit later, she gets a push message from her newly installed app (11). Customers build expectations based on experiences, and Anne is clearly not happy. Due to the digitization of the customer journey, such negative sentiments can be detected and acted upon. Finally, Anne reaches Amsterdam Central Station and checks out (12). Anne checks her credit on the card using a machine (13) and requests a refund using the app on her mobile phone (14). She takes the bus to the Van Gogh Museum. When entering the bus she checks in (15) and checks out (16) when exiting. A few days later she gets the requested refund (17) and starts planning her next trip (18).

During Anne's journey many events were recorded. It is easy to relate all events involving the OV-chipcard. However, some of the other events may be difficult to relate to Anne. This complicates analysis. *Event correlation*, i.e., establishing relationships between events, is one of the key challenges in data science.

The seven customer journey stages in Fig. 1.3 illustrate that the journey does not end after the 4th stage (purchase or booking). The classical "funnel-oriented" view towards purchasing a product is too restrictive. The availability of customer data from all seven stages helps shifting attention from sales to loyalty.

The development of photography and the many digital touchpoints in today's customer journey exemplify the growing availability of event data. Although data science is definitely not limited to Big Data, the dimensions of data are rapidly changing resulting in new challenges. It is fashionable to list challenges starting with the letter 'V'. Figure 1.4 lists the "four V's of Big Data": *Volume*, *Velocity*, *Variety*, and *Veracity*. The first 'V' (Volume) refers to the incredible scale of some data sources. For example, Facebook has over 1 billion active users and stores hundreds of petabytes of user data. The second 'V' (Velocity) refers to the frequency of incoming data that need to be processed. It may be impossible to store all data or the data may change so quickly that traditional batch processing approaches cannot cope with high-velocity streams of data. The third 'V' (Variety) refers to the differ-

ent types of data coming from multiple sources. Structured data may be augmented by unstructured data (e.g., free text, audio, and video). Moreover, to derive maximal value, data from different sources needs combined. As mentioned before, the correlation of data is often a major challenge. The fourth 'V' (Veracity) refers to the trustworthiness of the data. Sensor data may be uncertain, multiple users may use the same account, tweets may be generated by software rather than people, etc.

Already in 2001, Doug Laney wrote a report introducing the first three V's [87]. Later the fourth 'V' (Veracity) was added. Next to the basic four V's of Big Data shown in Fig. 1.4, many authors and organizations proposed additional V's: Variability, Visualization, Value, Venue, Validity, etc. However, there seems to be a consensus that *Volume*, *Velocity*, *Variety*, and *Veracity* are the key characteristics.

Later in this book we will focus exclusively on event data. However, these are an integral part of any Big Data discussion. Input for process mining is an event log which can be seen as a particular view on the event data available. For example, an event log may contain all events related to a subset of customers and used to build a customer journey map.

1.2 Data Scientist

Fueled by the developments just described, *Data science* emerged as a new discipline in recent years. Many definitions have been suggested [48, 112]. For this book, we propose the following definition:

> ***Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.***

The above definition implies that data science is broader than applied statistics and data mining. *Data scientists* assist organizations in turning data into value. A data scientist can answer a variety of data-driven questions. These can be grouped into the following four main categories [146]:

- (Reporting) *What happened?*
- (Diagnosis) *Why did it happen?*
- (Prediction) *What will happen?*
- (Recommendation) *What is the best that can happen?*

The definition of data science given is quite broad. Some consider data science as just a fancy term for *statistics*. Clearly, data science has its roots in statistics,

a discipline that developed over four centuries. John Graunt (1620–1674) started to study London's death records around 1660. Based on this he was able to predict the life expectancy of a person at a particular age. Francis Galton (1822–1911) introduced statistical concepts like regression and correlation at the end of the 19th century. Although data science can be seen as a continuation of statistics, the majority of statisticians did not contribute much to recent progress in data science. Most statisticians focused on theoretical results rather than real-world analysis problems. The computational aspects, which are critical for larger data sets, are typically ignored by statisticians. The focus is on generative modeling rather than prediction and dealing with practical challenges related to data quality and size. When the data mining community realized major breakthroughs in the discovery of patterns and relationships (e.g., efficiently learning decision trees and association rules), most statisticians referred to these discovery practices as "data fishing", "data snooping", and "data dredging" to express their dismay.

A few well-known statisticians criticized their colleagues for ignoring the actual needs and challenges in data analysis. John Tukey (1915–2000), known for his fast Fourier transform algorithm and the box plots, wrote in 1962: "For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. ... I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." [133]. This text was written over 50 years ago. Also Leo Breiman (1928–2005), another distinguished statistician, wrote in 2001 "This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics." [25]. David Donoho adequately summarizes the 50 year old struggle between old-school statistics and real-life data analysis in [48].

Data science is also closely related to data processing. Turing award winner Peter Naur (1928–2016) used the term "data science" long before it was in vogue. In 1974, Naur wrote: "A basic principle of *data science*, perhaps the most fundamental that may be formulated, can now be stated: The data representation must be chosen with due regard to the transformation to be achieved and the data processing tools available" [107]. Earlier, Peter Naur also defined *datalogy* as the "science of the nature and use of data" and suggested to use this term rather than "computer science". The book from 1974 also has two parts considering "large data": "Part 5—Processes with Large Amounts of Data" and "Part 6—Large Data Systems" [107]. In the book, "large amounts of data" are all data sets that cannot be stored in working memory. The maximum capacity of magnetic disk stores considered in [107] ranges between 1.25 and 250 megabytes. Not only the disks are orders of magnitude smaller than today's disks, also the notion of what is "large/big" has changed dramatically since the early 1970s. Nevertheless, many of the core principles of data processing have remained invariant.

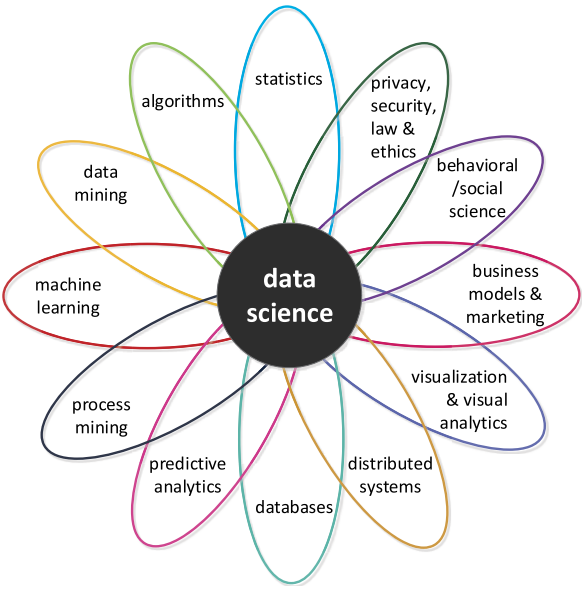

Fig. 1.5 The ingredients contributing to data science

Like data science, computer science had its roots in a number of related areas, including mathematics. Computer science emerged because of the availability of computing resources and the need for computer scientists. Data science is now emerging because of the omnipresence and abundance of data and the need for data scientists that can turn data into value.

Data science is an amalgamation of different partially overlapping (sub)disciplines. Figure 1.5 shows the main ingredients of data science. The diagram should be taken with a grain of salt. The (sub)disciplines are overlapping and varying in size. Moreover, the boundaries are not clear-cut and seem to change over time. Consider, for example, the difference between data mining and machine learning or statistics. Their roots are very different: data mining emerged from the database community, and machine learning emerged from the Artificial Intelligence (AI) community, both quite disconnected from the statistics community. Despite the different roots, the three (sub)disciplines are definitely overlapping.

- *Statistics* can be viewed as the origin of data science. The discipline is typically split into *descriptive* statistics (to summarize sample data using notions like mean, standard deviation, and frequency) and *inferential* statistics (using sample data to estimate characteristics of all data or to test a hypothesis).
- *Algorithms* are crucial in any approach analyzing data. When data sets get larger, the complexity of the algorithms becomes a primary concern. Consider, for example, the Apriori algorithm for finding frequent items sets, the MapReduce approach for parallelizing algorithms, and the PageRank algorithm used by Google search.
- *Data mining* can be defined as "the analysis of (often large) data sets to find unsuspected relationships and to summarize the data in novel ways that are both

understandable and useful to the data owner" [69]. The input data are typically given as a table and the output may be rules, clusters, tree structures, graphs, equations, patterns, etc. Clearly, data mining builds on statistics, databases, and algorithms. Compared to statistics, the focus is on scalability and practical applications.

- *Machine learning* is concerned with the question of how to construct computer programs that automatically improve with experience [102]. The difference between data mining and machine learning is equivocal. The field of machine learning emerged from within Artificial Intelligence (AI) with techniques such as neural networks. Here, we use the term machine learning to refer to algorithms that give computers the capability to learn *without* being explicitly programmed ("learning from experience"). To learn and adapt, a model is built from input data (rather than using fixed routines). The evolving model is used to make data-driven predictions or decisions.
- *Process mining* adds the process perspective to machine learning and data mining. Process mining seeks the confrontation between event data (i.e., observed behavior) and process models (hand-made or discovered automatically). Event data are related to explicit process models, e.g., Petri nets or BPMN models. For example, process models are discovered from event data or event data are replayed on models to analyze compliance and performance.
- *Predictive analytics* is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. To generate predictions, existing mining and learning approaches are applied in a business context. Predictive analytics is related to business analytics and business intelligence.
- *Databases* are used to store data. The database discipline forms one of the cornerstones of data science. Database Management (DBM) systems serve two purposes: (i) structuring data so that they can be managed easily and (ii) providing scalability and reliable performance. Using database technology, application programmers do not need to worry about data storage. Until recently, relational databases and SQL (Structured Query Language) were the norm. Due to the growing volume of data, massively distributed databases and so-called NoSQL databases emerged. Moreover, in-memory computing (cf. SAP HANA) can be used to answer questions in real-time. Related is OLAP (Online Analytical Processing) were data are stored in multidimensional cubes facilitating analysis from different points of view.
- *Distributed systems* provide the infrastructure to conduct analysis. A distributed system is composed of interacting components that coordinate their actions to achieve a common goal. Cloud, grid, and utility computing rely on distributed systems. Some analysis tasks are too large or too complex to be performed on a single computer. Such tasks can be split into many smaller tasks that can be performed concurrently on different computing nodes. Scalability may be realized by sharing and/or extending the set of computing nodes.
- *Visualization & visual analytics* are key elements of data science. In the end people need to interpret the results and guide analysis. Automated learning and

mining techniques can be used to extract knowledge from data. However, if there are many "unknown unknowns" (things we don't know we don't know),[2] analysis heavily relies on human judgment and direct interaction with the data. The perception capabilities of the human cognitive system can be exploited by using the right visualizations [178]. Visual analytics, a term coined by Jim Thomas (1946–2010), combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets [83].

- *Business models & marketing* also appear in Fig. 1.5 because data science is about turning data into value, including business value. The market capitalization of Facebook in November 2015 was approximately US $300 billion while having approximately 1500 million monthly active users. Hence, the average value of a Facebook user was US $200. At the same time, the average value of a Twitter user was US $55 (market capitalization of approximately US $17 billion with 307 million users). Via the website www.twalue.com one can even compute the value of a particular Twitter account. In November 2015, the author's Twitter account (*@wvdaalst*) was estimated to have a value of US $1002.98. These numbers illustrate the economic value of data and the success of young companies based on new business models. Airbnb (helping people to list, find and rent lodging), Uber (connecting travelers and drivers who use their own cars), and Alibaba (an online business-to-business trading platform) are examples of data-driven companies that are radically changing the hotel, taxi, and trading business. Marketing is also becoming more data-driven (see Sect. 1.1 describing the increase in digital touchpoints during a customer journey). Data scientists should understand how business considerations are driving the analysis of new types of data.
- *Behavioral/social science* appears in Fig. 1.5 because most data are (indirectly) generated by people and analysis results are often used to influence people (e.g., guiding the customer to a product or encouraging a manager to eliminate waste). Behavioral science is the systematic analysis and investigation of human behavior. Social sciences study the processes of a social system and the relationships among individuals within a society. To interpret the results of various types of analytics, it is important to understand human behavior and the social context in which humans and organizations operate. Moreover, analysis results often trigger questions related to coaching and positively influencing people.
- *Privacy, security, law, and ethics* are key ingredients to protect individuals and organizations from "bad" data science practices. Privacy refers to the ability to seclude sensitive information. Privacy often depends on security mechanisms which aim to ensure the confidentiality, integrity and availability of data. Data should be accurate and stored safely, not allowing for unauthorized access. Privacy and security need to be considered carefully in all data science applications. Individuals

[2] On February 12, 2002, when talking about weapons of mass destruction in Iraq, United States Secretary of Defense Donald Rumsfeld used the following classification: (i) "known knowns" (things we know we know), (ii) "known unknowns" (things we know we don't know), and (iii) "unknown unknowns" (things we don't know we don't know).

need to be able to trust the way data are stored and transmitted. Next to concrete privacy and security breaches, there may be ethical notions related to "good" and "bad" conduct. Not all types of analysis possible are morally defendable. For example, mining techniques may favor particular groups (e.g., a decision tree may reveal that it is better to give insurance to middle-aged white males rather than other groups). Moreover, due to a lack of sufficient data, minority groups may be wrongly classified. A data scientist should be aware of such problems and provide safeguards for "irresponsible" forms of data science.

Figure 1.5 shows that data science is quite broad and located at the intersection of existing disciplines. It is difficult to combine all the different skills needed in a single person. Josh Wills, former director of data science at Cloudera, defined a data scientist as "a person who is better at statistics than any software engineer and better at software engineering than any statistician". It will be a challenge to find and/or educate "unicorn" data scientists able to cover the full spectrum depicted in Fig. 1.5. As a result, 'unicorn" data scientists are in high demand and extremely valuable for data-driven organizations. As an alternative it is also possible to form multi-disciplinary teams covering the "flower" of Fig. 1.5. In the latter case, it is vital that the team members are able to see the bigger picture and complement each other in terms of skills.

1.3 Bridging the Gap Between Process Science and Data Science

In Fig. 1.5, we listed process mining, the topic of this book, as one of the essential ingredients of data science. Unfortunately, this is not a common view. The process perspective is absent in many Big Data initiatives and data science curricula. We argue that *event data should be used to improve end-to-end processes.* Process mining can be seen as a means to *bridge the gap between data science and process science*. Data science approaches tend to be process agnostic whereas process science approaches tend to be model-driven without considering the "evidence" hidden in the data.

We use the umbrella term "*process science*" to refer to the *broader discipline that combines knowledge from information technology and knowledge from management sciences to improve and run operational processes*. Figure 1.6 shows the ingredients of process science. Just like Fig. 1.5, the diagram should be taken with a grain of salt. The (sub)disciplines mentioned in Fig. 1.6 are also overlapping and varying in size.

- *Stochastics* provides a repertoire of techniques to analyze random processes. The behavior of a process or system is modeled using random variables in order to allow for analysis. Well-known approaches include Markov models, queueing networks/systems, and simulation. These can be used to analyze waiting times, reliability, utilization, etc. in the context stochastic processes.

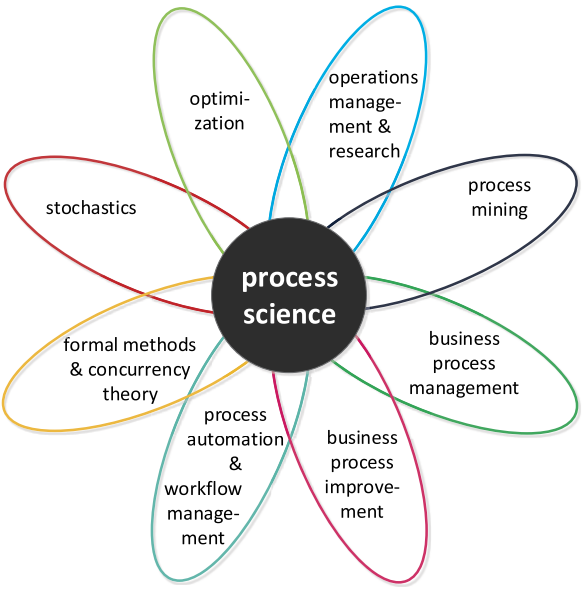

Fig. 1.6 Process science is an umbrella term for the broader discipline that combines knowledge from information technology and knowledge from management sciences to improve and run operational processes

- *Optimization* techniques aim to provide a "best" alternative (e.g., cheapest or fastest) from a large or even infinite set of alternatives. Consider, for example, the following question: Given a list of cities and the distances between each pair of cities, what is a shortest possible route that visits each city exactly once and returns to the origin city? Numerous optimization techniques have been developed to answer such questions as efficient as possible. Well-known approaches include Linear Programming (LP), Integer Linear Programming (ILP), constraint satisfaction, and dynamic programming.
- *Operations management & research* deals with the design, control and management of products, processes, services and supply chains. Operations Research (OR) tends to focus on the analysis of mathematical models. Operations Management (OM) is closer to industrial engineering and business administration.
- *Business process management* is the discipline that combines approaches for the design, execution, control, measurement and optimization of business processes. Business Process Management (BPM) efforts tend to put emphasis on explicit process models (e.g., Petri nets or BPMN models) that describe the control-flow and, optionally, other perspectives (organization, resources, data, functions, etc.) [50, 143, 187].
- *Process mining* is also part of process science. For example, process mining techniques can be used to discover process models from event data. By replaying these data, bottlenecks and the effects of non-compliance can be unveiled. Compared to mainstream BPM approaches the focus is not on process modeling, but on exploiting event data. Sometimes the terms Workflow Mining (WM), Business Process Intelligence (BPI), and Automated Business Process Discovery (ABPD) are used to refer to process-centric data-driven approaches.

- *Business process improvement* is an umbrella term for a variety of approaches aiming at process improvement. Examples are Total Quality Management (TQM), Kaizen, (Lean) Six Sigma, Theory of Constraints (TOC), and Business Process Reengineering (BPR). Note that most of the ingredients in Fig. 1.6 ultimately aim at process improvement, thus making the term business process improvement rather unspecific. One could argue that the whole of process science aims to improve processes.
- *Process automation & workflow management* focuses on the development of information systems supporting operational business processes including the routing and distribution of work. Workflow Management (WFM) systems are model-driven, i.e., a process model suffices to configure the information system and run the process. As a result, a process can be changed by modifying the corresponding process model.
- *Formal methods & concurrency theory* build on the foundations of theoretical computer science, in particular logic calculi, formal languages, automata theory, and program semantics. Formal methods use a range of languages to describe processes. Examples are transition systems, Petri nets, process calculi such as CSP, CCS and π-calculus, temporal logics such as LTL and CTL, and statecharts. Model checkers such as SPIN can be used to verify logical properties such as the absence of deadlocks. Concurrency complicates analysis, but is also essential: In reality parts of a process or system may be executing simultaneously and potentially interacting with each other. Petri nets were the first formalism to model and analyze concurrent processes. Many BPM, WFM, and process mining approaches build upon such formalisms.

As mentioned earlier, Fig. 1.6 should not be taken too seriously. It is merely a characterization of process science and its main ingredients. Note, for example, that stochastics and optimization are partly overlapping (e.g., solving Markov decision processes) and that BPM can be viewed as a continuation or extension of WFM with less emphasis on automation.

Process mining brings together traditional model-based process analysis and data-centric analysis techniques. As shown in Fig. 1.7, *process mining can be viewed as the link between data science and process science*. Process mining seeks the confrontation between event data (i.e., observed behavior) and process models (hand-made or discovered automatically). Mainstream data science approaches tend to be process agnostic. Data mining, statistics and machine learning techniques do not consider end-to-end process models. Process science approaches are process-centric, but often focus on modeling rather than learning from event data. The unique positioning of process mining, as sketched in Fig. 1.7, makes it a powerful tool to exploit the growing availability of data for improving end-to-end processes.

Process mining only recently emerged as a subdiscipline of both data science and process science, but the corresponding techniques can be applied to any type of operational processes (organizations and systems). Example applications include: analyzing treatment processes in hospitals, improving customer service processes in a multinational corporation, understanding the browsing behavior of customers

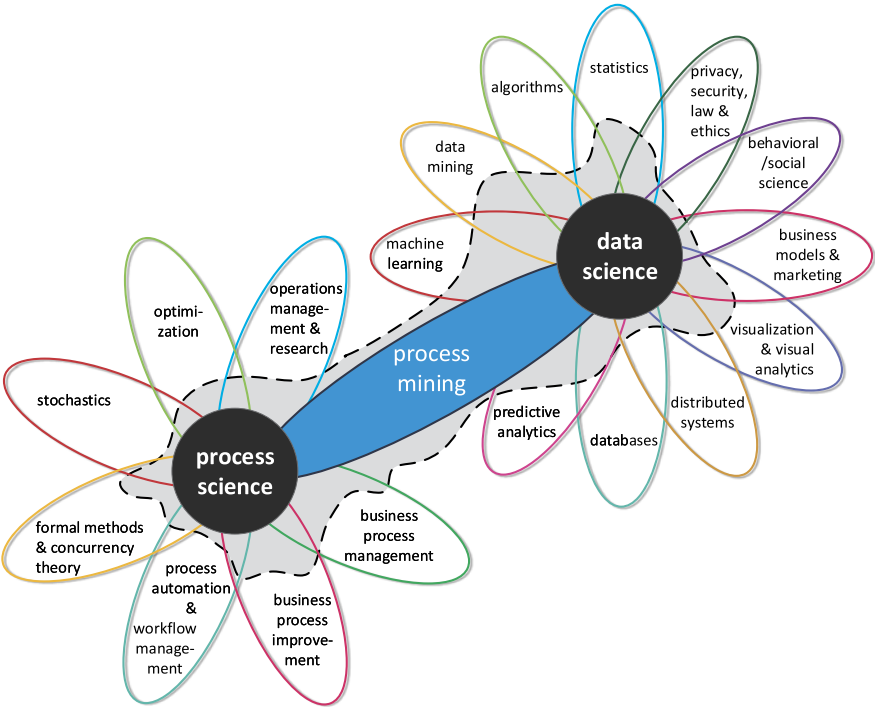

Fig. 1.7 Process mining as the bridge between data science and process science

using a booking site, analyzing failures of a baggage handling system, and improving the user interface of an X-ray machine. What all of these applications have in common is that dynamic behavior needs to be related to process models. Hence, we refer to this as "data science in action".

Spreadsheets: Dealing with numbers rather than dynamic behavior
Spreadsheet software can be found on most computers, and over the last 25 years many computers have been purchased just to be able to create and use spreadsheets. A spreadsheet is composed of cells organized in rows and columns. Some cells serve as input, other cells have values computed over a collection of other cells (e.g., taking the sum over an array of cells). The expression associated to a cell may use a range of arithmetic operations (add, subtract, multiply, etc.) and predefined functions. For example, Microsoft's Excel provides hundreds of functions including statistical functions, math and trigonometry functions, financial functions, and logical functions. Most organizations use spreadsheets in financial planning, budgeting, work distribution, etc. Hence, it is interesting to view process mining against the backdrop of this widely used technology.

The first widely used spreadsheet program was VisiCalc ("Visible Calculator") developed by Dan Bricklin and Bob Frankston, founders of Software Arts (later named VisiCorp). VisiCalc was released in 1979 for the Apple II computer. It is generally considered as Apple II's "killer application" because numerous organizations purchased the Apple II computer just to be able to use VisiCalc. When Lotus 1-2-3 was launched in 1983, VisiCalc sales dropped dramatically. Lotus 1-2-3 took full advantage of the IBM PC's capabilities and better supported data handling and charting. What VisiCalc was for the Apple II, Lotus 1-2-3 was for the IBM PC. For the second time, a spreadsheet program generated a tremendous growth in computer sales. People were buying computers in order to run spreadsheet software: A nice example of the "tail" (VisiCalc/Lotus 1-2-3) wagging the "dog" (Apple-II/IBM PC). Lotus 1-2-3 dominated the spreadsheet market until 1992. The dominance ended with the uptake of Microsoft Windows. After decades of spectacular IT-developments, spreadsheet software can still be found on most computers (e.g., Excel is part of Microsoft's Office) and can be accessed online (e.g., Google Sheets as part of Google Docs).

The situations in which spreadsheets can be used in a meaningful way are almost endless. In short, *spreadsheets can be used to do anything with numbers.* However, spreadsheets are *not* suitable for analyzing event data. One can count frequencies, sums, and the number of events per case using a so-called pivot table, but spreadsheets cannot be used to analyze bottlenecks and deviations (see Fig. 1.8). Consider questions like:

- What are the most frequent paths in my process? Do they change over time?
- What do the cases that take longer than 3 months have in common? Where are the bottlenecks causing these delays?
- Which cases deviate from the reference process? Do these deviations also cause delays?

Obviously, these questions cannot be answered using spreadsheets because the process perspective is completely absent in spreadsheets. Processes *cannot* be captured in numerical data and operations like summation. Process models and concepts such as cases, events, activities, timestamps, and resources need to be treated as first-class citizens during analysis. Data mining tools and spreadsheet programs take as input any tabular data without distinguishing these key concepts. As a result, such tools tend to be process-agnostic. Nevertheless, there is an obvious need for spreadsheet-like technology tailored towards processes and event data.

Where spreadsheets work with *numbers*, process mining starts from *event data* with the aim to analyze processes. Instead of pie charts, bar charts, and tables, results include end-to-end process models, conformance diagnostics, and bottlenecks.

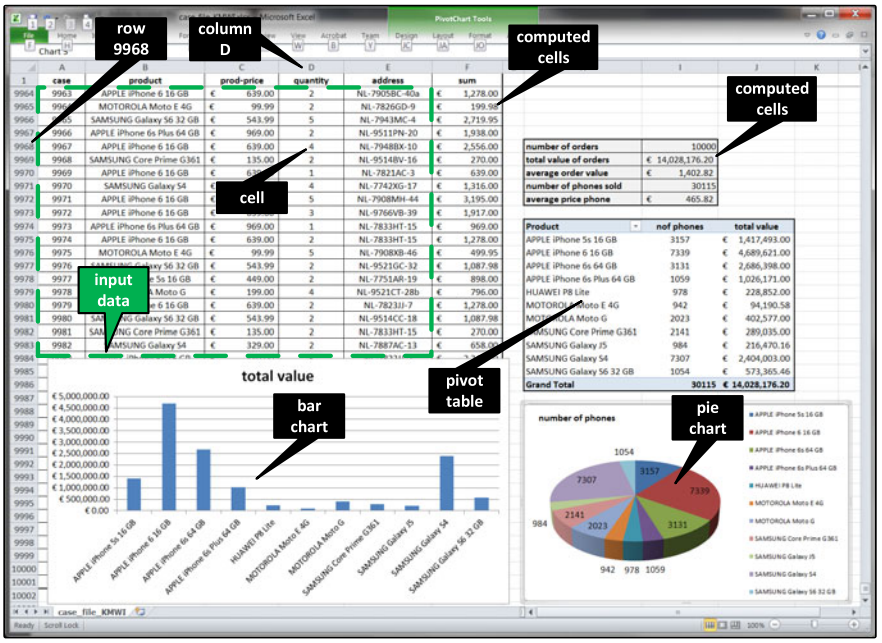

Fig. 1.8 Spreadsheets can be used to do anything with numbers, but have difficulties adequately capturing dynamic behavior

As will be demonstrated in later chapters, the process mining spectrum is quite *broad*. It is not limited to automated process discovery: Process mining can also be used to check compliance, diagnose deviations, pinpoint bottlenecks, improve performance, predict flow times, and recommend actions. Process mining techniques are also *generic*: just like spreadsheet software. Event logs and operational processes can be found everywhere and the analysis techniques are not limited to specific application domains. Just like Excel can be used in finance, production, sales, education, and sports, process mining software can be used in a variety of application domains.

1.4 Outlook

Process mining provides an important bridge between data mining and business process modeling and analysis. Process mining research at TU/e (Eindhoven University of Technology) started in 1999. At that time there was little event data available and the initial process mining techniques were extremely naïve and hence unusable in practice. Over the last decade event data have become readily available and process mining techniques have matured. Moreover, process mining algorithms have been implemented in various academic and commercial systems. Today, there is an active group of researchers working on process mining, and it has become one of the

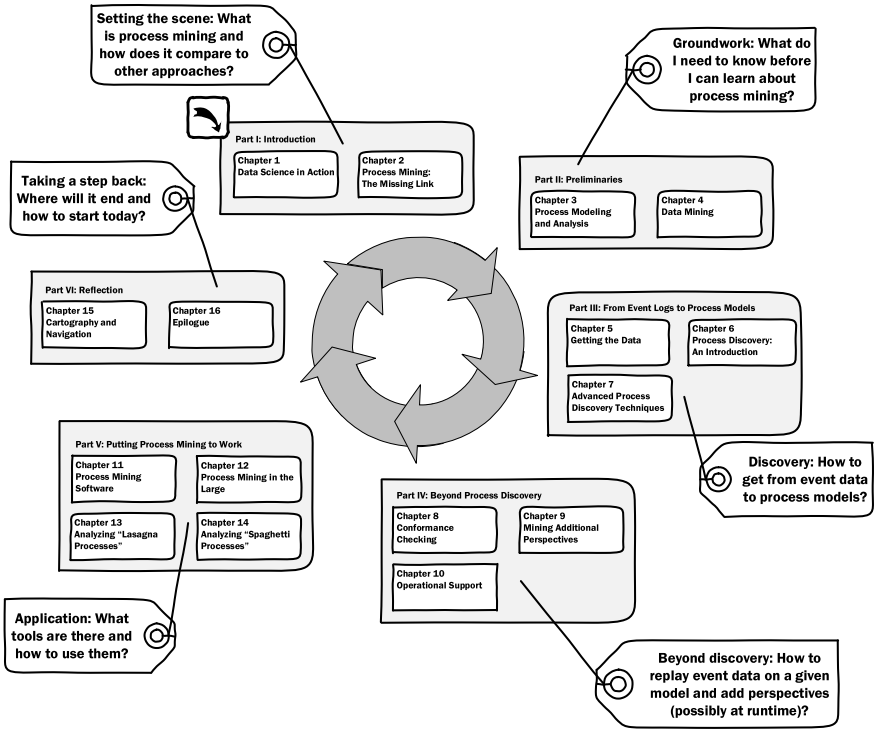

Fig. 1.9 Outline of the book

"hot topics" in BPM research. Moreover, there is a rapidly growing interest from industry in process mining. More and more software vendors started adding process mining functionality to their tools. Our open-source process mining tool ProM is widely used all over the globe and provides an easy starting point for practitioners, students, and academics. These developments are the main motivation for writing this book. There are many books on data mining, business ~~un~~intelligence, process reengineering, and BPM, but these rarely address process mining.

This book aims to provide a comprehensive overview of process mining. The book is intended for business process analysts, business consultants, process managers, graduate students, and BPM researchers. On the one hand, the book avoids delving into unnecessary details. On the other hand, the book does not shy away from formal definitions and technical issues needed to fully understand the essence of process mining. As Einstein said: "Everything should be made as simple as possible, but not one bit simpler".

Figure 1.9 provides an overview of the book. *Part I* introduces process mining and positions this emerging discipline in the context of data science and process science. *Chap. 2* discusses the role of process models, introduces the notion of event logs, and illustrates the main process mining tasks using a small example.

Part II provides the preliminaries necessary for reading the remainder of the book. *Chap. 3* introduces different process modeling languages and provides an overview of model-based analysis techniques. *Chap. 4* introduces standard data mining techniques such as decision tree learning and association rule learning. Process mining can be seen as a bridge between the preliminaries presented in both chapters.

Part III focuses on one particular process mining task: process discovery. *Chap. 5* discusses the input needed for process mining. The chapter discusses different input formats and issues related to the extraction of event logs from heterogeneous data sources. *Chap. 6* presents the α-algorithm step-by-step in such a way that the reader can understand how it works and see its limitations. This simple algorithm has problems dealing with less structured processes. Nevertheless, it provides a basic introduction into the topic and serves as a "hook" for discussing more advanced algorithms and general issues related to process mining. *Chap. 7* introduces more advanced process discovery approaches. This way the reader gets a good understanding of the state-of-the-art and is guided in selecting suitable techniques.

Part IV moves beyond process discovery, i.e., the focus is no longer on discovering the control-flow. *Chap. 8* presents conformance checking approaches, i.e., techniques to compare and relate event logs and process models. It is shown that conformance can be quantified and that deviations can be diagnosed. *Chap. 9* focuses on other perspectives: the organizational perspective, the case perspective, and the time perspective. *Chap. 10* shows that process mining can also be used to support operational processes at runtime, i.e., while cases are running it is possible to detect violations, make predictions, and provide recommendations.

Part V guides the reader in successfully applying process mining in practice. *Chap. 11* provides an overview of the different process mining tools. Data science is often related to Big Data. The "four V's of Big Data" (Fig. 1.4) are obviously also relevant for event data and their analysis. *Chap. 12* shows that process mining problems can be decomposed in various ways and many of the techniques can be adapted to provide scalability. The next two chapters are based on the observation that there are essentially two types of processes: "Lasagna processes" and "Spaghetti processes". Lasagna processes are well-structured and relatively simple. Therefore, process discovery is less interesting, but the techniques presented in Part IV are highly relevant for Lasagna processes. The added value of process mining can be found in conformance checking, detailed performance analysis, and operational support. *Chap. 13* explains how process mining can be applied in such circumstances and provides various real-life examples. Spaghetti processes are less structured. Therefore, the added value of process mining shifts to providing insights and generating ideas for better controlled processes, but advanced techniques such as prediction are less relevant for Spaghetti processes. *Chap. 14* shows how to apply process mining in such less-structured environments.

Part VI takes a step back and reflects on the material presented in the preceding parts. *Chap. 15* provides a broader vision on the topic by comparing process modeling with cartography, and relating BPM systems to navigation systems provided by vendors such as TomTom, Garmin, and Navigon. The goal of this chapter is to provide a refreshing view on process management and reveal the limitations of existing

information systems. *Chap. 16* concludes the book by summarizing improvement opportunities provided by process mining. The chapter also discusses some of the key challenges and provides concrete pointers to start applying the material presented in this book.