Preprint from Braschler, Stadelmann, Stockinger (Eds.): *“Applied Data Science - Lessons Learned for the Data-Driven Business”*, Springer , 2018 (to appear) .

# Introduction

by Thilo Stadelmann, Martin Braschler and Kurt Stockinger.

*What is data science? Attempts to define it can be made in one (prolonged) sentence, while it may take a whole book to demonstrate the meaning of this definition. This book introduces data science in an applied setting, by first giving a coherent overview of the background in Part I, and then presenting the nuts & bolts of the discipline by means of diverse use cases in Part II; finally, specific and insightful lessons learned are distilled in Part III. This chapter introduces the book and provides an answer to the following questions: What is data science? Where does it come from? What are its connections to big data and other mega trends? We claim that multidisciplinary roots and a focus on creating value lead to a discipline in the making that is inherently an interdisciplinary, applied science.*

## 1. Applied data science

It would seem reasonable to assume that many readers of this book have first really taken notice of the idea of "data science" after 2014. Indeed, while already used sparingly and with different meanings for a long time, widespread use of the term "data science" dates back to only 2012 or thereabouts (see Section 2). Of course, the "substance" of the field of data science is very much older, and goes by many names. To attest to this fact, the institute at which the editors of this book are located has given itself the mission to "build smart information systems" already back in 2005. And the main fields of work of the respective editors (Information Retrieval, Information Systems/Data Warehousing and Artificial Intelligence/Machine Learning) all have traditions that span back for decades. Still, a fundamental shift has been underway since 2012.

This chapter traces this fundamental shift by giving a historical account of the various roots of data science[1]. However, what do we understand by the term data science? As for this book, we adopt a definition that attests to the breadth and history of the field, is able to discriminate it from predecessor paradigms, and emphasizes its connection to practice by having a clear purpose:

*Data science refers to a unique blend of principles and methods from analytics, engineering, entrepreneurship and communication that aim at generating value from the data itself.*

These principles and methods are diverse (e.g., spanning disciplines from IT to legal studies), and are applied to all kinds of data (from relational to multimedia) to explicitly achieve a specific end: added value. This makes data science inherently an interdisciplinary and applied science, and connects the term closely with the definitions of a *data product* (an exploitable insight derived from the collected facts) and the *data scientist* (the one carrying out data science endeavours). All three terms will be more thoroughly treated, rooted and discussed in Chapters 2 to 4.

[1] Based on updated, translated and considerably extended versions of (Stockinger & Stadelmann, 2014; Stockinger et al., 2016).

# 2. The history of data science

This section intends to give a concise answer to two questions: what is the connection between data *science* and business, especially in the presence of a massive media hype? And what fueled the availability of (and trust in) large-scale data analysis? A more complete overview of the most influential publications leading to data science as we know it is given by Press (2013).

## 2.1 Data science, business and hype

The understanding of data science as the field concerned with all aspects of making sense of data goes back to discussions in the scientific community that started with Tukey (1962)[2] and where summarized by Cleveland (2001) in requiring an independant scientific discipline in extension to the technical areas of the field of statistics. Notable mentions go to the foundation of the field of "knowledge discovery in databases" (Fayyad et al, 1996) after the first KDD workshop in 1989[3], the first mentioning of "data science" in the title of a scientific conference in 1996 (Hayashi et al., 1996), and Leo Breiman's (2001) famous call to unite statistical and computational approaches to modeling data. These events lead to the foundation of the first scientific data science journals[4] in 2002 and the first data science research centers in 2007[5].

However, widespread recognition beyond a scientific community, including the dynamics we see today, only started after certain contributions from business: Hal Varian tells the McKinsey Quarterly in 2009 that the *"sexy job of the next 10 years will be statisticians"* (Manyika, 2009). This view broadens beyond statistics after the introduction of the term "data scientist" by Patil and Hammerbacher in 2008 during their collaboration at LinkedIn and Facebook (Patil, 2011). Both felt the need for a new job description for their team members that, on the one hand, got deep engineering know-how and, on the other hand, were directly shaping the economic value of the company's core products: *"those who use both data and science to create something new".* Earlier, Davenport and Harris (2007) and others had prepared the way for an acceptance of those terms (Smith, 2011) by influential popular scientific books that continued to accompany the development of data science (Siegel, 2013; Domingos, 2015).

The two concepts - the field of data science as well as the job description of a data scientist - in their now popular form (Loukides, 2010) together with their fame per se thus ultimately resulted from a need and development within businesses[6] (Patil, 2011). The scientific

[2] About the same time as Tukey used "data science" in reference to statistics, Peter Naur in Sweden used the term (interchangeably with "datalogy") to refer to computer science (Sveinsdottir & Frøkjær, 1988).
[3] See https://www.kdnuggets.com/meetings/kdd89/index.html.Since 1995, the term "data mining" has risen to prominence: http://www.aaai.org/Conferences/KDD/kdd95.php.
[4] See e.g. http://www.codata.org/publications/data-science-journal (inaugurated 2002, relaunched 2015) and http://www.jds-online.com/ (since 2003).
[5] See e.g. http://datascience.fudan.edu.cn/ (under the term "Dataology and Data Science").
[6] This anchoring of the modern understanding of data science more in business than in academia is the main reason for many of the references in this work pointing to blog posts and newspaper articles

discussion, once in a leading role, had difficulty to keep up with the dynamics of 2010-2015 and followed with some delay (Provost & Fawcett, 2013; Stadelmann et al., 2013). It is currently accelerating again (see Brodie (2015b) and his chapters later in this book).

Omnipresent, however, since the adoption of the topic in mass media has been a hype (see some of its expressions e.g. in (Humby, 2006) or (Davenport & Patil, 2012)) strong enough to provoke scepticism even in benevolent experts. While hype can lead to unreflected and hence bad decisions on all levels (from job choice to entrepreneurial and legislative agenda setting), it should not cloud the view on the real potential and challenges of data science:

- *Economic potential*: the McKinsey Global Institute estimates the net worth of the open data market alone to be three trillion dollars (Chui et al., 2014). A recent update explains that this potential is not realized yet, and certainly not overhyped (Henke et al., 2016). Earlier, Manyika et al. (2011) estimated a total shortcoming of 190,000 new data scientists.
- *Societal impact*: data analytics affects medical care (Parekh, 2015), political opinion making (Harding, 2017; also Krogerus & Grassegger, 2016 and the aftershocks of the US presidential election 2016 with regards to the involvement of the company Cambridge Analytica) and personal liberty (Li et al., 2015; see also Clemens Cap's later chapter on risks and side effects).
- *Scientific influence*: data-intensive analysis as the fourth paradigm of scientific discovery promises breakthroughs in disciplines from physics to life sciences (Hey et al., 2009; see also Brodie's later chapter "on developing data science").

Hype merely exclaims that *"data is the new oil!"* and jumps to premature conclusions. The original quote continues to be more sensible: *"[...] if unrefined, it cannot really be used. It has to be changed [...] to create a valuable entity that drives profitable activity"* (Humby, 2006). This already hints at the necessity of the precise and responsible work of a data scientist, guided by a body of sound principles and methods maintained within a scientific discipline. However, how did individual voices of "big data evangelists" grew into a common understanding of the power and usefulness of the resource of data by means of analytics?

## 2.2 Different waves of big data

Data science has profound roots in the history of different academic disciplines as well as in science itself (see also the detailed discussion in the next chapter). The surge of large-scale science in the second half of the 20th century, typified by facilities like CERN[7] or the Hubble Space Telescope[8], is the direct enabler of data science as a paradigm of scientific discovery based on data. These facilities have arguably enabled the *first wave* of big data: a single experiment at CERN for example would generate hundreds of terabytes of data in just one second, if not for a hardware filter that would do a preselection of what to record.

instead of scientific journals and conference papers. It reflects current reality while not making the point that academia is subordinate. Data science as a field and business sector is in need of the arranging, normative work of academia in order to establish solid methodical foundations, codes of conduct etc. This book is meant as a bridge builder in this respect.

[7] See https://home.cern/ (the website of the web's birthplace).

[8] See http://hubblesite.org/.

Consequently, specific projects like RD45 (Shiers, 1998) where launched already in the nineties to manage these high volumes of data before any commercial database management system was able to host petabytes (Düllmann, 1999)[9]. This rise in scientific data volumes was not just due to technical ability, but due to a change of paradigm (Hey et al., 2009): The first paradigm of science basically was to perform theoretical studies; the second paradigm added empiricism: the experimental evaluation of theoretical hypotheses. Because of the complexity and expensiveness of large-scale scientific experiments like at CERN, computer simulations emerged as the third paradigm of scientific discovery (called computational science). Now, the fourth paradigm is data-intensive science: evaluating, for any given experiment, *all* the facts (i.e., the complete data set, not just sub-samples and hand-engineered features), combining them to *all possible* probabilistic hypotheses (see also Brodie’s chapter on “what is data science” later).

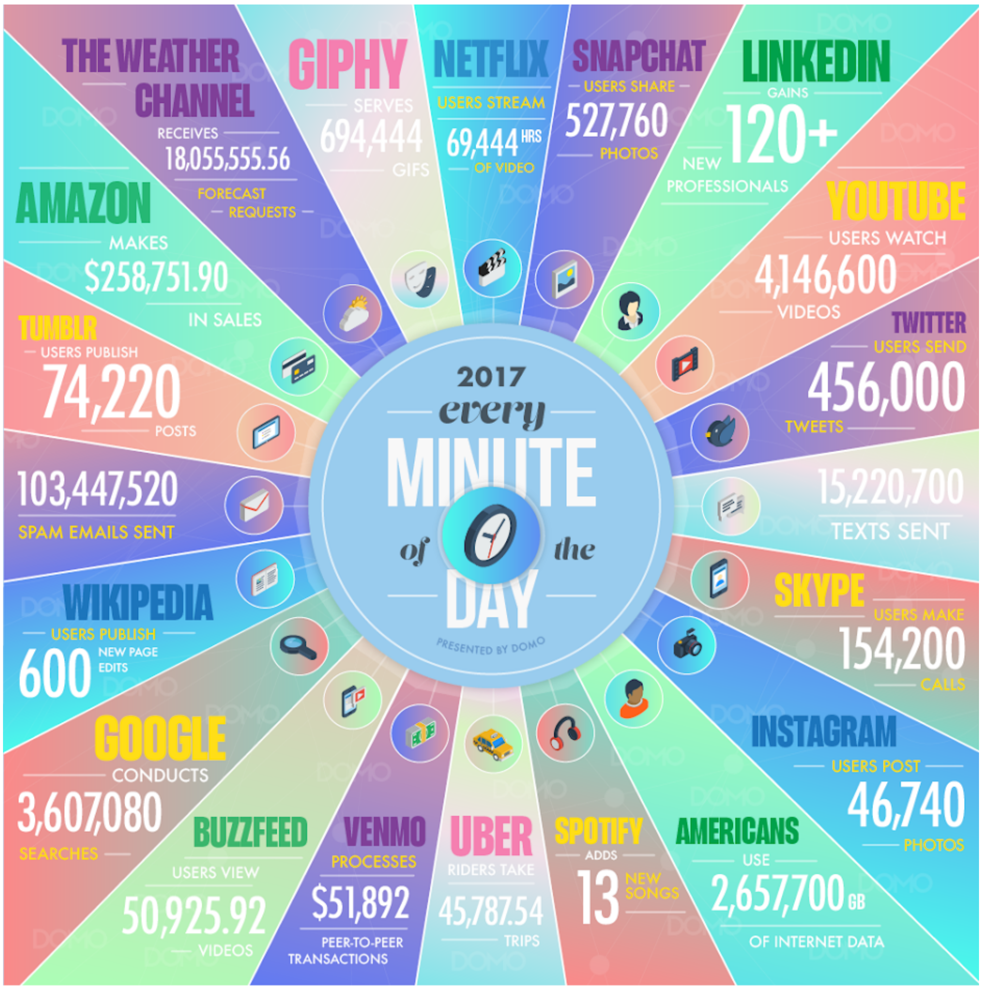


*Figure 1: The amount of data created every minute publicly on the web, as of July 2017. Shown is a part of the “data never sleeps 5.0” infographic by Domo (2017).*

[9] Additionally, unlike with relational databases in industry at that time, the types of data to be stored for scientific experiments frequently comprised numerical data (such as temperature, velocity or collision counts for particles), often stored in object-oriented database systems or images (e.g. from stars or galaxies), both at higher volumes and speed.

Following large-scale science, the *second wave* of big data was triggered by internet companies[10] like Google, Yahoo or Amazon at the beginning of the 21st century. In contrast to scientific data, the web companies originally focused on managing text data; the continuing data explosion (see Figure 1) is still fueled by additionally indexing more and more images and videos. Social media companies such as Facebook and LinkedIn gave individuals - instead of large scale scientific facilities - the ability to contribute to the growth of data; this is regarded as the *third wave* of the data tsunami. Finally, the *fourth wave* is currently rolling up based on the rise of machine-generated data, such as log-files and sensor data on the internet of things.

## 3. Data science and global mega trends

The scientific and commercial development of data science has been accompanied by considerable buzz in the public press. In this section, we review the contemporary trends of big data, AI, and digitalization, respectively, and put the terms into the context of the professional discussion. Our goal is to disentangle the meaning of the terms as hype words from their scientific definition by showing discrepancies in public understanding from what experts refer to when using largely overlapping vocabulary, thus contributing to a successful dialog.

*Figure 2: Snapshot from the preparations of a business road show in illustration of the hopes and dreams connected with the term "big data" as used by the public press around 2013 (picture courtesy of T.S.). The hopes and dreams have been seamlessly transferred to other wording as of the writing of this book. This is in stark contrast to the continuous scholarly work on big data discussed e.g. by Valerazo et al. (2016) and the scientific community that adopted the same name[11].*

[10] All of the following three example companies have transformed themselves considerably since the times of the second wave (see https://abc.xyz/ and https://www.oath.com/ for Google and Yahoo, respectively; https://www.amazon.com/ still looks familiar, but compare http://www.visualcapitalist.com/jeff-bezos-empire-chart/).

[11] See e.g. https://journalofbigdata.springeropen.com/.

## 3.1 Big data

In 2013-2014, "big data" was the favourite buzzword in business: the new oil (McAfee et al., 2012) of the economy. It alluded exclusively to the technical origins of the term, namely Laney's (2001) "3 Vs" (variety, velocity and volume) as attributes of the growing flood[12] of data: the amount, the number of sources and the speed at which new data arrives is very large, i.e., "big" (Soubra, 2012). It is this technical definition that is also eponymous of the scientific sub-discipline of information systems researchers that develop database management systems at scale[13]. Data science, on the contrary, is far from focusing exclusively on "big" in the sense of "large" data: while large itself is always a relative term as compared to the computational resources of the day[14], data science is concerned with generating value from *all kinds of* data (see Braschler's later chapter on "small data").

There is a different, more economical than technological connotation to the public discussion on big data that conveys some meaning for the data scientist: it is best seen in a historic version of the Wikipedia article on big data from 2014[15]. In its "See also" section, the term "big data" is brought into relationship with terms like "big oil", "big tobacco", "big media" etc. The connotation is as follows: at least as much as from the description of the phenomenon of increased variety/velocity/volume, the term big data might stem from a whole economy's hope of having found the new oil (Humby, 2006), of getting new business opportunities and of launching the "next big thing" after the internet and mobile revolution[16].This explains why the term has been hyped a lot in business for several years. Additionally, this also expresses the fear of individuals being ushered into the hands a "big brother", just as it is expressed as a present reality in the other "big*" terms from above.

The term "big data" up to here thus contributes a two-fold meaning to the professional discussion on data science: first, *technologically*, it gives a description of the growing state of data (and as such is a scientific sub-discipline of research in databases, information systems and distributed systems). Second, *economically*, it expresses a hope for business opportunities and voices a subtle concern with respect to the attached dangers of this business. Both dimensions are worthy to be explored and have to be researched. The greatest potential of the term, however, may lie in pointing to the following *social* phenomenon[17]:

[12] In a variation of Naisbitt and Cracknell's (1982) famous quote on megatrends, Eric Brown (2014) said: "Today, we are drowning in data and starved for information".

[13] The top league of international database researchers meets under the umbrella of "very large data bases" (http://www.vldb.org/) since 1992.

[14] See e.g. the seminal book by Witten et al. (1999) on "managing gigabytes" that was allegedly influential in building Google, but would not qualify as discussing "big" a couple of years later. Similarly, the test collection at the first TREC conference, at 2 gigabytes of size, posed a considerable challenge to the participants in 1992 (Harman & Voorhees, 2006).

[15] The following discussion is based on https://en.wikipedia.org/wiki/Big_data as of May 01, 2014. It is accessible via Wikipedia's history button.

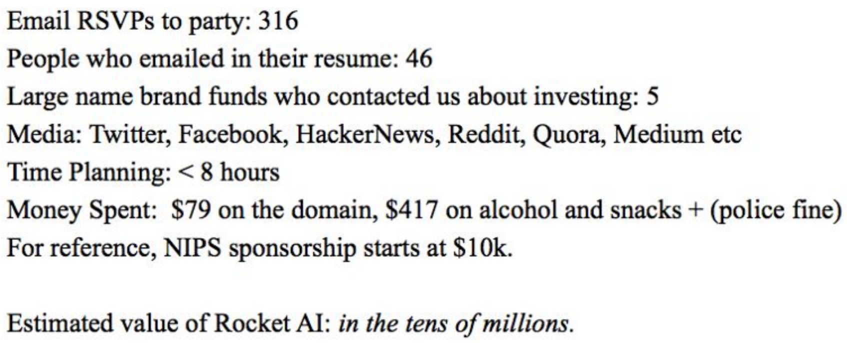
[16] Dan Ariely expressed this in a Facebook post gone viral on January 06, 2013: "Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everybody else is doing it, so everybody claims they are doing it..." (see http://t.co/tREI1mRQ).

[17] This thought was first formulated by Michael Natusch in his Keynote at SDS|2014, the 1st Swiss Conference on Data Science (see also Figure 1 in the preface to this book):

People (and organizations) have changed their mindset in the last decade. Data is now regarded as being available (cheap if not for free) about virtually any aspect of the world, meaning that we can have facts for potentially any phenomenon on the planet. Additionally, we have the technology ready to automatically make use of these facts via the principles and methods of data science, for almost all existing business processes. This enables optimized (i.e., fact-based) decisions, which have a measurable value independent of the data having properties of up to *n* Vs (Vorhies, 2014). This social value of "big data" thus is this: it refers to *a big change in thinking about the possibilities of data-driven automated decision making.*

## 3.2 Artificial intelligence

The term "big data" was largely forsaken by the public press after circa two years of constant use (White, 2015), but it did not leave a vacuum: unforeseen breakthroughs in deep learning technology as of 2012 ended the last AI winter[18] around 2015 (see also Stadelmann et al.'s later chapter on "deep learning in industrial practice") and directly turned it into the next "AI" hype. This unreasonable cycle of popularity can be traced in the open, too: for example, Simard et al (2003) display the AI winter of the 2000s with the following quote: "[...] it was even pointed out by the organizers of the Neural Information Processing System (NIPS) conference that the term "neural networks" in the submission title was negatively correlated with acceptance". On the other hand, the RocketAI story (Tez, 2016) illustrates the peak of the hype: at the NIPS conference of 2016, expectations in neural networks where again so high that the joke of two PhD students of a "launch party" for a fake, fancy AI start-up attracted large unsolicited funding, applications and attendees with minimum effort within a day (see Figure 3).



Email RSVPs to party: 316
People who emailed in their resume: 46
Large name brand funds who contacted us about investing: 5
Media: Twitter, Facebook, HackerNews, Reddit, Quora, Medium etc
Time Planning: < 8 hours
Money Spent: $79 on the domain, $417 on alcohol and snacks + (police fine)
For reference, NIPS sponsorship starts at $10k.

Estimated value of Rocket AI: *in the tens of millions.*

*Figure 3: AI hype at its worst. The metrics of the "Rocket AI launch party" prank at NIPS 2016. Picture from Tez (2016).*

AI in the public press as of 2018 mainly refers to the expectation to "do with computers anything that a human could do - and more". It often ascribes human-like properties to AI systems as reflected in larger parts of the discussion revolving around terms like robots (Kovach, 2017), digital assistants (Kremp, 2018), self-driving cars (Roberts, 2018), chatbots (Sprout Social, 2018), and neural networks as digital brains (Gruber, 2017). What

https://www.zhaw.ch/en/research/inter-school-cooperation/datalab-the-zhaw-data-science-laboratory/sds2014/michael-natusch/.

[18] See https://en.wikipedia.org/wiki/AI_winter.

contributes to (if not even causes) the high, even inflated expectations is the use of terms originally coined for intelligent live ("intelligence", "learning", "cognitive", "social"). Everybody has an intuitive understanding of what "intelligence" means in a human context, and subconsciously ascribes said properties to the technical system.

The scientific community is not innocent of this dilemma, even if it tries to clarify things (Brooks, 2017): in private communication[19], one of the fathers of AI regretted coining the term "artificial intelligence" at the Dartmouth workshop in 1956 exactly for the outgrowths described above, mentioning that it would have been wiser (but maybe less successful) to have gone with the second proposed name for the field - *"complex computer applications"*. It is exactly this what defines the scientific discipline of AI today (Russell & Norvig, 2010): a collection of diverse techniques, from efficient search algorithms to logic to various shades of machine learning, to solve tasks of everyday life that usually can only be solved by humans, in an attempt that *might look intelligent* from the outside. As such, the field of AI is not concerned with researching intelligence per se nor in reproducing it; but it delivers practical solutions to problems e.g. in business for many decades (Hayes-Roth et al., 1983), even with deep learning (LeCun et al., 1997).

## 3.3 Digitalization

Different technological and industry-specific trends[20] additionally get summarized under the unifying term "digitalization"[21] in the public discourse. The added value of this term per se can reasonably be questioned: things arguably became digital - digitized - since the IT revolution of businesses and societies in the last century. The modern use extends this mega trend by emphasizing increased interconnectedness (social networks) and automation. This trend is specifically enabled by data science, based on the availability of digital data ("big data" in the social sense above) and analytics technologies ("AI"). It spans almost all industry and societal branches, and hence the discussion not only involves data science professionals - technical people -, but also sociologists, politicians etc. with valuable contributions to the phenomenon et large.

The missing selectivity of the public use of "digitalization" and the abovementioned "buzz" words create a problem: experts and laypeople speak in the same terms but mean different things. The new "AI algorithm" inside a company's product is potentially more statistics than AI, speaking in technical terms; the just purchased "big data platform" might likely refer to an analytics tool not specifically designed to handle large data (as would be the case if called such by a big data researcher); and digitalization changes our education, but likely not predominantly in the sense that we now have to replace human teachers, but by teaching students skills in handling a digitalized society (including skills in handling digital media and basic understandings of data science technology) (Zierer, 2017).

[19] The statement was orally witnessed at an AI planning conference in the nineties by a colleague who wishes to remain anonymous.

[20] Compare terms like FinTech (Dapp et al., 2014), MedTech (MedTech Europe, 2018), EdTech etc. (Mayer, 2016) as well as Industrie 4.0 (Kagermann et al., 2011).

[21] Not "digitization" - compare the article by Clerck (2017).

The missing selectivity in the use of terms cannot be removed from the discourse[22]. It is thus important for data professionals - data scientists - to understand what experts and laypeople mean and hear when speaking of such terms, in order to anticipate misunderstandings and confront unreasonable expectations.

## 4. Outlook

What is the future of data science? Data science as a discipline of study is still *in its infancy (Brodie, 2015a)*, and the principles and methods developed in its underlying disciplines have to be furthered in order to adapt to the phenomenon we called big data in the previous section. This maturing of data science will be addressed in two later companion chapters by Michael Brodie in Part II of the book.

From a business perspective, data science will continue to deliver value to most industries[23] by introducing possibilities for automation of repetitive tasks[24]. A recent overview of successful data-driven innovations in Switzerland, presented at the first "Konferenz Digitale Schweiz", showed the overall potential by demonstrating that the innovation depth in current data science projects is surprisingly low (Swiss Alliance for Data-Intensive Services, 2017): one third of business innovations was achieved by deploying technology and processes that have been well-known for decades; another third was achieved by consulting companies on recent innovations in data-driven business models and technologies; and for only one third, applied research projects fostered the foundation for the business innovation. Part II of this book reports on numerous case studies emerging from that latter third.

While the benefit to businesses is fairly obvious and easily measurable in terms of profit, the effect of data science on our societies is much less well understood. Recent reports warn about potential blind spots in our core technologies (Rahimi & Recht, 2017) and go as far as suggesting to treat AI technology similar to nuclear weapons in limiting access to research results (Brundage, 2017). The recent emergence of the word "techlash" might indicate society's first large-scale adverse reaction on the dawn of digitalization (Kuhn, 2018). Clemens Cap explores such issues in his chapter towards the end of Part I of this book, while Widmer and Hegy shed some light on the legal space in which a data scientist operates.

The next three chapters will continue defining what data science, a data scientist and a data product is, respectively. As stated in the preface, they are best read in order to get a

[22] Some experts suggest to use all the discussed terms synonymously for the sake of simplicity, e.g. Brodie in his later chapter on "developing data science" speaks of "AI/data science". While this might be appropriate in certain situations, maintaining proper attribution certainly helps in other situations by maintaining concise and effective communication. Using precise terminology prevents inflated expectations, describes true expertise, and gives guidance in where to find it (e.g., in what discipline).

[23] François Bancilhon put it frankly in his ECSS 2013 keynote: *"most industries will be disrupted"* (see http://www.informatics-europe.org/ecss/about/past-summits/ecss-2013/keynote-speakers.html).

[24] See (Brooks, 2017) for a counter argument on the hopes (or fears) that too many jobs could be automated in a nearer future. But as Meltzer (2014) points out, repetitive tasks even in jobs considered high-profile (e.g., medical diagnosis or legal advice) could be automated quite well: automation potential lies in repetitiveness per se, not the difficulty of the repetitive task.

coherent picture of the frame for this book. The remaining chapters can be approached in any order and according to personal interest or project need. A concise summary of all lessons learned will be presented in Part III. We intend this part to form best practices for applying data science that you will frequently refer to as you start your own professional data science journey.

## References

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science, 16(3)*, 199-231.

Brodie, M. L. (2015a). The emerging discipline of data science. Keynote at the 2nd Swiss Workshop on Data Science SDS|2015. Available online (May 03, 2018): https://www.youtube.com/watch?v=z93X2k9RVqg

Brodie, M. L. (2015b). Doubt and Verify: Data Science Power Tools. Available online (March 23, 2018): http://www.kdnuggets.com/2015/07/doubt-verify-data-science-power-tools.html.

Brooks, R. (2017). The Seven Deadly Sins of AI Predictions. *MIT Technology Review.* Available online (March 28, 2018): https://www.technologyreview.com/s/609048/the-seven-deadly-sins-of-ai-predictions/.

Brown, E.D. (2014). Drowning in Data, Starved for Information. Available online (March 27, 2018): http://ericbrown.com/drowning-in-data-starved-for-information.htm.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Anderson, H. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv preprint arXiv:1802.07228*.

Chui, M., Farrell, D., & Jackson, K. (2014). How Government can Promote Open Data. Available online (March 23, 2018): https://www.mckinsey.com/industries/public-sector/our-insights/how-government-can-promote-open-data.

Clerck, J. (2017). Digitization, digitalization and digital transformation: the differences. i-SCOOP. Available online (March 23, 2018): https://www.i-scoop.eu/digitization-digitalization-digital-transformation-disruption/.

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1), 21-26.

Dapp, T., Slomka, L., AG, D. B., & Hoffmann, R. (2014). *Fintech–The digital (r) evolution in the financial sector*. Deutsche Bank Research, Frankfurt am Main.

Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Harvard Business Press.

Davenport, T.H., & Patil, D. (2012). Data scientist: the sexiest job of the 21st century. Available online (March 23, 2018):
http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1.

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world.* Basic Books.

Domo (2015). Data never sleeps 5.0. Available online (March 23, 2018):
https://www.domo.com/learn/data-never-sleeps-5.

Düllmann, D. (1999). Petabyte databases. In ACM SIGMOD Record (Vol. 28, No. 2, p. 506). ACM.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Gruber, A. (2017). Wenn Maschinen lernen lernen. *Spiegel Online*, January 17, 2017. Available online (May 10, 2018):
http://www.spiegel.de/netzwelt/web/kuenstliche-intelligenz-wenn-maschinen-lernen-lernen-a-1130255.html

Hayashi, C., Yajima, K., Bock, H. H., Ohsumi, N., Tanaka, Y., & Baba, Y. (Eds.). (1996). Data Science, Classification, and Related Methods: *Proceedings of the Fifth Conference of the International Federation of Classification Societies* (IFCS-96), Kobe, Japan, March 27–30, 1996. Springer Science & Business Media.

Harding, C. (2017). Digital Participation - the Advantages and Disadvantages. Available online (March 23, 2018):
https://www.polyas.de/blog/en/digital-democracy/digital-participation-advantages-disadvantages.

Harman, D. K., & Voorhees, E. M. (2006). TREC: An overview. *Annual review of information science and technology*, 40(1), 113-155.

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., & Sethupathy, G. (2016). The age of analytics: Competing in a data-driven world. McKinsey Global Institute report.

Hey, T., Tansley, S., & Tolle, K. M. (2009). The fourth paradigm: data-intensive scientific discovery (Vol. 1). Redmond, WA: Microsoft research.

Hayes-Roth, F., Waterman, D. A., & Lenat, D. B. (1983). *Building expert system.*

Humby, C. (2006). Data is the new Oil!. ANA Senior marketer's summit, Kellogg School, November 2006. http://ana.blogs.com/maestros/2006/11/data_is_the_new.html.

Kagermann, H., Lukas, W. D., & Wahlster, W. (2011). Industrie 4.0: Mit dem Internet der Dinge auf dem Weg zur 4. industriellen Revolution. *VDI nachrichten*, 13, 11.
Kuhn, J. (2018). "Techlash": Der Aufstand gegen die Tech-Giganten hat begonnen. *Süddeutsche Zeitung*. Available online (April 03, 2018): http://www.sueddeutsche.de/digital/digitalisierung-techlash-der-aufstand-gegen-die-tech-giganten-hat-begonnen-1.3869965.

Kremp, M. (2018). Google Duplex ist gruselig gut. *Spiegel Online*, May 09, 2018. Available online (May 10, 2018): http://www.spiegel.de/netzwelt/web/google-duplex-auf-der-i-o-gruselig-gute-kuenstliche-intelligenz-a-1206938.html

Krogerus, M., & Grassegger, H. (2016). Ich habe nur gezeigt, dass es die Bombe gibt. *Das Magazin*, (48-3). Available online (May 11, 2018): https://www.tagesanzeiger.ch/ausland/europa/Ich-habe-nur-gezeigt-dass-es-die-Bombe-gibt/story/17474918

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6(70).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

Li, R., Lu, B., & McDonald-Maier, K. D. (2015). Cognitive assisted living ambient system: a survey. *Digital Communications and Networks*, 1(4), 229-252.

Loukides, M. (2010). What is data science? Available online (March 23, 2018): https://www.oreilly.com/ideas/what-is-data-science.

Manyika, J. (2009). Hal Varian on how the Web challenges managers. McKinsey Quarterly. Available online (March 23, 2018): https://www.mckinsey.com/industries/high-tech/our-insights/hal-varian-on-how-the-web-challenges-managers.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A.H. (2011). Big data: the next frontier for innovation, competition, and productivity. Available online (March 23, 2018): https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

Mayer, M. (2016). Fintech? Edtech? Adtech? Duriantech? - the 10 buzziest startup sectors. Available online (March 23, 2018): https://techsauce.co/en/startup-2/fintech-edtech-adtech-duriantech-the-10-buzziest-startup-sectors/.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. *Harvard business review*, 90(10), 60-68. Available online (March 23, 2018): https://hbr.org/2012/10/big-data-the-management-revolution.

MedTech Europe (2018). The European medical technology industry in figures 2018. *MedTech Europe Brochure*. Available online (May 10, 2018): http://www.medtecheurope.org/EU-medtech-industry-facts-and-figures-2017

Meltzer, T. (2014). Robot doctors, online lawyers and automated architects: the future of the professions? *The Guardian.* Available online (March 28, 2018): https://www.theguardian.com/technology/2014/jun/15/robot-doctors-online-lawyers-automated-architects-future-professions-jobs-technology.

Naisbitt, J., & Cracknell, J. (1984). *Megatrends: Ten new directions transforming our lives (No. 04; HN59. 2, N3.).* New York: Warner Books.

Parekh, D. (2015). How Big Data Will Transform Our Economy And Our Lives. Available online (March 23, 2018): http://techcrunch.com/2015/01/02/the-year-of-big-data-is-upon-us/.

Patil, D. (2011). Building data science teams. Available online (March 23, 2018): http://radar.oreilly.com/2011/09/building-data-science-teams.html.

Press, G. (2013). A very short history of data science. Available online (March 23, 2018): https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science.

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data Vol. 1, No. 1, March 2013.

Rahimi, A., & Recht, B. (2017). Reflections on Random Kitchen Sinks. Acceptance speech for Test of Time Award at NIPS 2017. Available online (March 28, 2018): http://www.argmin.net/2017/12/05/kitchen-sinks/.

Roberts, D. (2018). Here’s how self-driving cars could catch on. *Vox* article, May 09, 2018. Available online (May 10, 2018): https://www.vox.com/energy-and-environment/2018/5/8/17330112/self-driving-cars-autonomous-vehicles-texas-drive-ai

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach, Third Edition*. Upper Saddle River, New Jersey. Pearson Education, Inc.

Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR* (Vol. 3, pp. 958-962).

Shiers, J. (1998). Building a multi-petabyte database: The RD45 project at CERN. Object Databases in Practice, 164-176.

Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die.* John Wiley & Sons Incorporated.

Smith, D. (2011). “Data Science”: What’s in a name?. Available online (March 27, 2018): http://blog.revolutionanalytics.com/2011/05/data-science-whats-in-a-name.html.

Soubra, D. (2012). The 3Vs that define Big Data. Available online (March 23, 2018): www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data.

Spout Social (2018). The Complete Guide to Chatbots in 2018. *Sprout blog*, available online (May 10, 2018): https://sproutsocial.com/insights/topics/chatbots/

Stadelmann, T., Stockinger, K., Braschler, M., Cieliebak, M., Baudinot, G.R., Dürr, O., & Ruckstuhl, A. (2013). Applied Data Science in Europe – Challenges for academia in keeping up with a highly demanded topic. In *European Computer Science Summit ECSS 2013*. August 2013, Amsterdam, The Netherlands, Informatics Europe.

Stockinger, K., & Stadelmann, T. (2014). Data Science für Lehre, Forschung und Praxis. HMD Praxis der Wirtschaftsinformatik, 51(4), 469-479.

Stockinger, K., Stadelmann, T., & Ruckstuhl, A. (2016). Data Scientist als Beruf. In Fasel, D., Meier, A. (eds.), *Big Data*, Edition HMD, DOI 10.1007/978-3-658-11589-0_4.

Sveinsdottir, E., & Frøkjær, E. (1988). Datalogy—the Copenhagen tradition of computer science. *BIT Numerical Mathematics, 28(3)*, 450-472.

Swiss Alliance for Data-Intensive Services (2017). Digitization & Innovation through cooperation. Glimpses from the Digitization & Innovation Workshop at “Konferenz Digitale Schweiz”. Available online (March 28, 2018): https://www.data-service-alliance.ch/blog/blog/digitization-innovation-through-cooperation-glimpses-from-the-digitization-innovation-workshop.

Kovach, S. 2017. *We Talked To Sophia — The AI Robot That Once Said It Would 'Destroy Humans'*. Tech Insider youtube video, available online (May 10, 2018): https://www.youtube.com/watch?v=78-1MlkxyqI

Tez, R.-M. (2016). *Rocket AI: 2016’s Most Notorious AI Launch and the Problem with AI Hype.* Blog post, available online (May 10, 2018): https://medium.com/the-mission/rocket-ai-2016s-most-notorious-ai-launch-and-the-problem-with-ai-hype-d7908013f8c9

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics, 33(1)*, 1-67.

Valarezo, U. A., Pérez-Amaral, T., & Gijón, C. (2016). Big Data: Witnessing the Birth of a New Discipline. *Journal of Informatics and Data Mining*, 1(2).

Vorhies, W. (2014). *How Many "V's" in Big Data? The Characteristics that Define Big Data.* Available online (March 23, 2018): https://www.datasciencecentral.com/profiles/blogs/how-many-v-s-in-big-data-the-characteristics-that-define-big-data.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing Gigabytes: Compressing and Indexing Documents and Images.* Morgan Kaufmann.

White, A. (2015). *The end of Big Data - It’s all over now.* Available online (March 23, 2018): https://blogs.gartner.com/andrew_white/2015/08/20/the-end-of-big-data-its-all-over-now/.

Zierer, K. (2017). Warum der Fokus auf das digitale Klassenzimmer Unfug ist. *Spiegel Online*, December 27, 2017. Available online (May 10, 2018): http://www.spiegel.de/lebenundlernen/schule/digitales-klassenzimmer-die-schueler-muessen-wieder-in-den-mittelpunkt-a-1181900.html#ref=meinunghpmobi