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

# Data Science

by Martin Braschler, Thilo Stadelmann and Kurt Stockinger.

## Abstract

*Even though it has only entered public perception relatively recently, the term "data science" already means many things to many people. This chapter explores both top-down and bottom-up views on the field, on the basis of which we define data science as "a unique blend of principles and methods from analytics, engineering, entrepreneurship and communication that aim at generating value from the data itself". The chapter then discusses the disciplines that contribute to this "blend", briefly outlining their contributions and giving pointers for readers interested in exploring their backgrounds further.*

## 1. Introduction

"Data science" is a term that has entered public perception and imagination only since the first half of the decade. Even in the expert community, fundamental treatments such as "What is Data Science?" (Loukides, 2010) were first published as recently as 2010. Yet, the substance of what constitutes data science has been built up for much longer. An attempt to define the term "data science" can follow either a top-down or a bottom-up philosophy. On the one hand, looking "top-down", data science is the research field that studies mechanisms and approaches necessary to generate value and insights from data, enabling the building of data products. Importantly, a "data product" is not just a product "dealing" with data, but it is a product deriving its value from the data and producing data itself (Loukides, 2010). On the other hand, adopting the "bottom-up" view, data science is an interdisciplinary research field (Stockinger et al., 2015) that adopts a new, holistic way of exploiting data, looking beyond single aspects such as how to store data, or how to access it. It follows that we need to integrate competencies from many older disciplines of study: technical-mathematical disciplines such as "computer science" and "statistics", but also disciplines such as "entrepreneurship" and "art".

No one view, top-down or bottom-up, is superior. In fact, there was and is considerable disagreement exactly where the boundary of data science is to be drawn (Warden, 2011). Rather than engage in this "war of definitions", we think it is helpful to view the different approaches as complementary. For our own work in talking to audiences as diverse as students, colleagues and business partners, we found the following definition most helpful, and thus adopt it for this book on applied data science (Stadelmann et al., 2013):

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

What makes this phrasing stand out for us is threefold:

1. The definition distinguishes data science well from preceding paradigms: it is not equal to its individual parts, such as analytics, engineering, etc. (or their sub-disciplines, such as AI, algorithms, or statistics), i.e., no single sub-discipline "owns" data science. Nor is it simply the sum of these parts, i.e. it does not include any sub-discipline entirely. Instead, it refers to a *unique blend of principles & methods* from them. We arrived at this conclusion through intensive collaboration between computer scientists and statisticians at the ZHAW School of Engineering, and are convinced that it holds generally. Unlike e-science and other domain-specific paradigms, data science is universal in applying to all kinds of data and application areas. Finally, unlike in data mining, which concentrates on exploratory data analysis, there is a clear goal in data science to *generate value from data*. Reflecting on the top-down view given above, the data product guides this value generation.
2. The definition connects science to practice: by emphasizing data science as an applied science (encompassing *entrepreneurship*, having the goal of *generating value*), the important aspect of it being “grounded in reality”[1] is highlighted. Again, the applicability of this has been verified many times over in our work in applied research and development over the past years, and distinguishes data science from some of the fundamental work done in the constituting disciplines (e.g., establishing the laws of probability remains a fundamental result in math, not data science).
3. The definition testifies to the breadth and history of the field: the *unique blend of methods & principles* explicitly acknowledges that data science is "standing on the shoulders of giants" instead of "reinventing the wheel". It also acknowledges the fact that it is more than an umbrella term: the blend we refer to is unique, and not just a universal collection, but a tailored selection of relevant methods, principles and tools from the constituting disciplines.

The remainder of this chapter will trace these "giants" and their contributions to data science.

Critically, to turn "data science" into more than a label or a hype (see e.g. Davenport & Patil, 2012), and rather into a real discipline on its own right, a re-thinking of the whole process of leveraging data is necessary, from data acquisition all the way to the building of a "data product". Data science is more than an umbrella term precisely because it not only allows to bundle all the individual disciplines from the constituting fields; the term also finally allows a convenient way to express this idea of working at the so far uncovered interfaces of the different subfields. Data science is very much about creating synergies. The remainder of this chapter will highlight clearly that data science is an *applied* and *interdisciplinary* endeavour: the case studies covered in Part II could not be feasible otherwise, and would suffer greatly from the lack of a concise, accurate term to describe them.

## 2. Applied data science

When discussing the "added value" of combining traditional academic disciplines such as statistics and computer science, but also economics, the notion of "generating value from data" stands out. Data becomes a product - inherently making data science an applied

[1] See also Brodie’s later chapter “on developing data science”

science. On the other hand, an endeavour becomes *scientific* if it examines a phenomenon by use of the scientific method[2] with the goal to gain knowledge. It becomes *applied science* if the scientific method is applied not just to any phenomenon, but to problems that arise in "everyday life" and the solution of which directly improves issues at home, at work, in business, or in the society at large.

The distinction between basic and applied research thus is the origin of the research question – the phenomenon or hypothesis to consider. The majority of research in data science is applied, being directly motivated by use cases; and it can be argued that without this demand from use cases the more fundamental questions (e.g., how to scale the developed methods to all relevant domains[3]) would basically not arise. In turn, use cases provide a means to test hypotheses arising from purely fundamental work - much like test set examples in machine learning help evaluating the generalization capabilities of an established (trained) model.

More use cases than ever await solutions due to more data than ever being available to more actors than at any time in history[4] - but if one believes in the age-old saying "knowledge is power", leveraging that data becomes ever more pressing, lest the competition might glean insight from it first. Many companies discover that they are in fact more data-driven than they may have previously perceived, and that, as their respective fields of business transform in the information age, they need to "activate" their data if they want to continue to prosper. This shift had already been seen previously in data-intensive academic fields, such as physics and astronomy, and thus there is much that industry can learn from earlier endeavours.

## 3. Interdisciplinarity in data science

The key to becoming a business player in today's supercharged "online market" is the ability to build the necessary "data products". Successful data science projects often capitalize on the interface between industry and science by relying, on the one hand, on a successful interpretation of the use case and the customer needs, coupled with an attractive, effective design, and on the other hand on building on top of the right, state-of-the art techniques and tools.

It would be a tall order to cover all the many diverse disciplines for such a project in equal depth. In practice, this is not necessary in every undertaking. Ideally, a team of data scientists bundles the required skills, with the individual team members having different profiles - more on the question of what makes a successful data scientist can be found in the following chapter. Importantly, one could argue that a fair amount of fascination for the field of data science derives precisely from this bridging of business and engineering aspects.

[2] The scientific method refers to the cycle of theory formation and experimentation introduced by Newton - see https://en.wikipedia.org/wiki/Scientific_method.

[3] Refer also to Brodie's later chapter on "what is data science?".

[4] See e.g. https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/.

Figure 1 shows the different academic subfields that the data scientists at the ZHAW Datalab[5] use to describe their main lines of research. The figure, a tag cloud or word cloud (Bateman et al., 2008), nicely illustrates the sheer diversity of (sub-)disciplines that contribute to data science. We will in the following sections briefly describe the most important of the contributing academic fields, starting with the technical-mathematical disciplines, and extending to the additional fields that have to be covered to truly produce "data products".

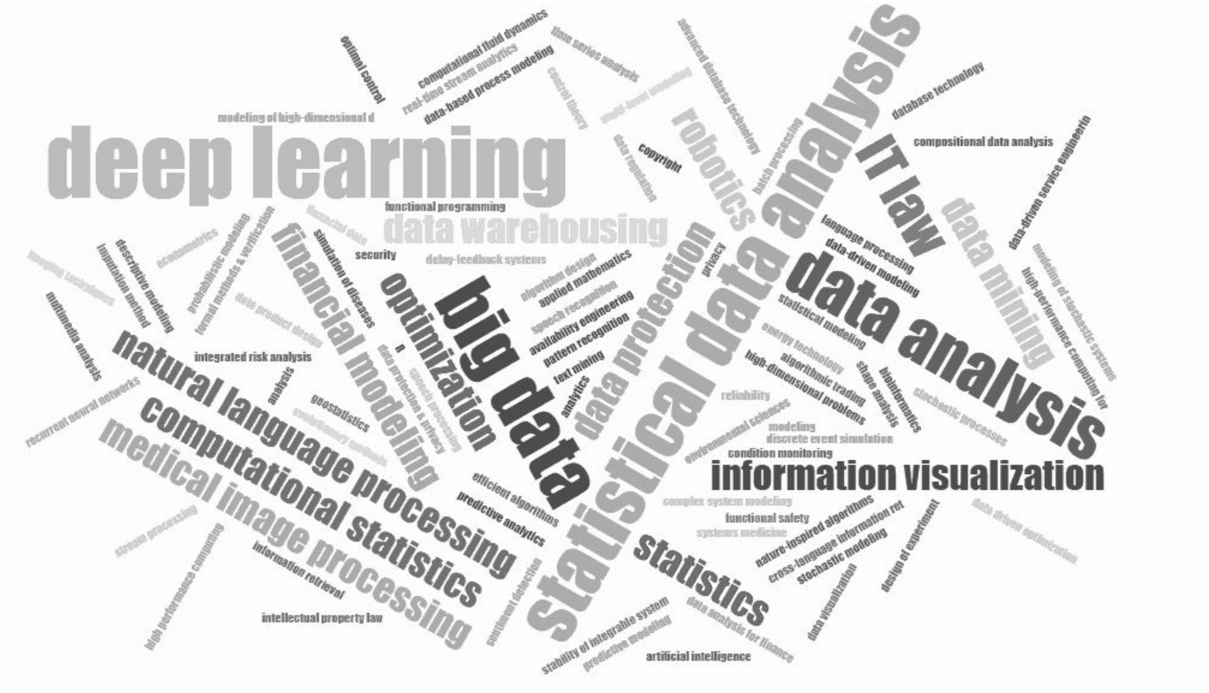

Figure 1: A tag cloud compiled from the tags that researchers at the ZHAW Datalab use to describe their research[6].

## 3.1 Computer science

Computer science, the academic discipline that covers all aspects related to the design ("hardware") and use ("software") of computers (Aho & Ullman, 1992), is a frequent first career path for data scientists. This is on the basis of the two subfields of data-processing algorithms (Knuth, 1968) and information systems (Ramakrishnan & Gehrke, 2002). The former is the study of the way that computers process data: algorithms describe how computers carry out tasks. The latter deals with storage, handling and processing of large amounts of (digital) data - something that is impossible without the use of computers. Processing and handling data stands at the core of every (digital) computer. Starting with the introduction of the von Neumann architecture in 1945 (von Neumann, 1993), even the instructions for the computer are handled equally to the data it processes - both are stored in the same form, namely in the volatile main memory and on external storage. Everything is thus "data" from a computer science perspective. However, not all aspects of computer science are of equal importance to data science: aspects such as design of hardware, or

[5] See the preface of this book for more information on the ZHAW Datalab.

[6] Produced with the generator available at https://www.jasondavies.com/wordcloud/.

software engineering, take backseat to those research lines directly addressing data and information:

- the storage of data: here, mainly research on database systems (Silberschatz et al., 1997), i.e. the persistent storage of structured data, is relevant. Classically, the "relational model" for databases (Codd, 1970) has been the main approach for storing structured data for a long time. However, in the context of big data systems, new exciting developments are also pertinent, such as NoSQL databases (Stonebraker, 2010).
- the handling of data: mostly tools-driven. Scripting languages, such as Python (van Rossum & Drake, 2003) or Perl (Wall et al., 1999) are often used for "data wrangling"; i.e., the transformation of data between various formats or representations.
- the processing or accessing of data and information: here, in addition to algorithmic work that we will treat below under the umbrella of artificial intelligence, the most important research subfields are data warehousing (for structured data) (Chaudhuri & Dayal, 1997, Inmon, 2005) and information retrieval (for unstructured data) (Schütze et al., 2008). Data warehousing is mostly concerned with the methods to arrange data for efficient and effective analysis, where information retrieval extends the research on accessing unstructured textual or multimedia data to questions on how to interpret the data to satisfy information needs by the users.

Much of the subfields listed above can be subsumed under the heading "information systems". It should be noted here that the names of these subfields are somewhat plagued by inconsistent use of the terms "data" and "information". Often, they deal in actuality with both aspects - with the "raw" data and with information, i.e. the data coupled with an interpretation (Bellinger et al., 2004).

## 3.2 Statistics

While computer science delivers the tools to store, process and query the data, which is the "fuel" of data science, statistics is at the core of the academic fields that support the transformation of data into value, e.g. in the form of insight or decisions (Wilcox, 2009). When consulting common definitions of the field of statistics, some of the same boxes we mentioned for information systems are ticked: statistics deals, much like information systems, with the collection and organization of data. However, the viewpoint is a fundamentally different one: while in information systems, we refer to the storage and processing of data at large in the "mechanical sense", here we have the focus on the selection and organization of data in the mathematical sense. This collection is the precursor to analysis, interpretation and presentation of data. Statistics provides tools to describe properties of data sets ("descriptive statistics" (Holcomb, 1997)), as well as drawing conclusions from data sets that are a sample of a larger population ("inferential statistics" (Wasserman, 2013)). Crucially, statistics provides the tools (tests) to verify hypotheses as to the relationship between variables and data sets, and provides a different angle to work done in computer science on machine learning.

## 3.3 Artificial intelligence

Artificial intelligence (AI) (Russell & Norvig, 2010), and especially its branch machine learning, is typically treated as a subfield of computer science, and sits nicely at the intersection of computer science, statistics and several other disciplines. AI generally studies solutions to complex problems arising in areas such as human perception, communication, planning and action (Luger, 2008). Most relevant for data science, but not uniquely so, is the branch of machine learning that studies algorithms that can "learn" from data, based on pattern recognition in data sets (Bishop, 2007). There is potential in increasingly combining this with logic-based AI that reasons over ontologies[7]. Ideally, the learning improves performance on a given task without the need for a human to program an explicit solution (Samuel, 1959). This is both attractive in cases where such an explicit solution is very complex or if the task deals with constantly changing parameters.

Supervised (machine) learning is based on providing the learning algorithm pairs of possible inputs and their corresponding outputs. Based on this "training data", a desired mapping function or algorithm is learnt. A key problem in this context is potential overfitting, i.e. if the learning process picks up undesired artifacts present in the training data that are not representative of the larger population of possible inputs. More fundamentally, suitable training data may be difficult and costly to obtain. In unsupervised (machine) learning, the aim is to find hidden structure in data sets without the use of training labels.

## 3.4 Data mining

Another subfield straddling the boundaries of computer science and statistics is data mining. The term is used in somewhat different ways, with many different forms of information processing being labelled as data mining (Witten et al., 2016). Generally, it applies principles from machine learning, statistics and data visualization, where the goal is the detection of previously unknown patterns in data[8]. The differentiation to data science lies in the focus on the extracted patterns themselves, whereas data science covers a broader range of topics already beginning at data collection with the explicit goal of a data product in the end. Data mining can thus be thought of as a predecessor paradigm to interdisciplinary work on data by applying fundamental results from e.g. the analytics fields of statistics or machine learning.

## 3.5 Additional Technical Disciplines

There are multiple additional technical disciplines that contribute to the "umbrella field" of data science. We want to make a special note of some of these: on the one hand, there is business intelligence (Chen et al., 2012), which stands at the interface between technical aspects and management and aims to leverage the data to provide a view on business

[7] Pointed out, e.g., by Emmanuel Mogenet from Google, at the Zurich Machine Learning & Data Science Meetup in February 2017, talking about combining subsymbolic (machine learning) approaches with symbolic (logic-based) AI.

[8] Rather than detection of the data itself. One could thus argue that the term is unfortunate, and an alternative along the lines of "mining on data" would be more appropriate.

operations. On the other hand, there are several disciplines that dissolved from AI in the last decades and now form their own communities, usually centered around separate types of data: this includes speech as well as natural language processing, which deals with processing natural (human) language by computer (Manning & Schütze, 1999; Yu & Deng, 2014); computer vision, which deals with images and video[9]; or pattern recognition, which deals with the problem of automatizing human perception tasks (Stork et al., 2000)

## 3.6 The "knowledge discovery in databases (KDD)" process

An alternative viewpoint of data science can be taken by referencing the "knowledge discovery in databases (KDD)" process (Fayyad et al., 1996). Data typically sits at the bottom of a "knowledge pyramid", illustrated in Figure 2 (Frické, 2009). Simply put, data can be viewed as a collection of codes, that can be stored, organized and accessed. Only if a (contextual) meaning is affixed to data does it become information. When information is then analyzed and interpreted, and new inferences are drawn, the result is knowledge. This is, as we have previously pointed out, the goal of data science, insofar as value is created through new knowledge.

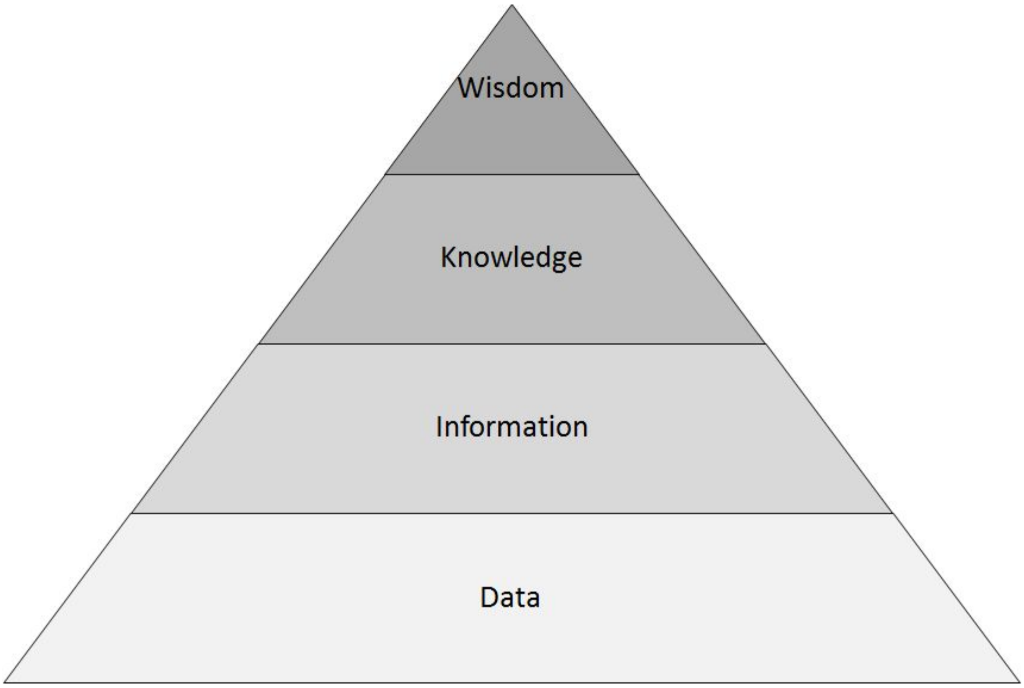

Figure 2: The knowledge pyramid.

On a conceptual level, the KDD process models this progression from data to knowledge through a series of stages[10]:

1. data is first selected (measured, recorded, then initially stored), and then
2. pre-processed (noise removal, filling in of missing values, outlier detection, etc.);

[9] Since 2012, the field of computer vision has been largely changed by recent research on deep neural networks, see e.g. (Goodfellow et al., 2016) and (LeCun, 2013).

[10] compare also the cross-industry standard process for data mining (Shearer, 2000)

3. it is transformed into a form suitable for analysis (often a 2D table format, after doing including feature selection or transformations like Fourier transform), and then
4. analyzed (by statistical or machine learning methods to find patterns of correlation).
5. Finally, the result of potentially multiple analyses is interpreted/evaluated by a human (or human-deveised decision mechanism)

These five stages again tie in nicely with the technical disciplines discussed so far. An alternative rendering of the KDD process, that puts a slightly different emphasis on the different stages, aligning it more with the disciplines, could thus read:

1. data recording
2. data wrangling (incl. data cleaning and storage, i.e. in databases or information retrieval systems)
3. data analysis (incl. statistics, artificial intelligence and data mining)
4. data visualization and/or interpretation, and lastly,
5. decision making

Data Science as an interdisciplinary academic field goes far beyond only technical-mathematical aspects as covered by disciplines such as computer science or statistics. We like the quote by Hilary Mason, who concisely described data science as "statistics, computer science, domain expertise, and what I usually call 'hacking" (Woods, 2012). The chapter has not covered the "domain expertise" bit so far, but such expertise is crucial for understanding the unique value proposition of treating data science as an unified pursuit of leveraging data. "Domain expertise" both addresses the need for knowledge of the different domains that the data originates from (legal domain, medical domain, etc.), but also more generally the possession of non-technical skills such as arts and entrepreneurship. The KDD process as outlined above culminates in data visualization/interpretation and decision making, which both heavily rely on non-technical expertise as outlined below.

## 3.7 Data or information visualization

At the interface between computer science (computer graphics, see e.g. Hughes et al., 2013) and arts (see below) lies data or information visualization (Ware, 2012). In both cases, the goal is a rendering of the data to support better human understanding. When choosing the term "data visualization", more weight is given to the aspect of raw data visualization (Tufte, 2001), whereas "information visualization" more stresses the aspect of supporting interpretation of the data through visualization[11]. Often the terms are used (rather confusingly) interchangeably. In both cases, the visualization is a communication tool: large amounts of data are either compressed visually into a rendering that can provide an overview, or are rendered to be explored interactively, with the user zooming in and out of the data to discover the needed information. The rendering of the results of data analysis is crucial to feed the KDD stages of interpretation or decision making - data has to be rendered in such a way that the desired information or knowledge becomes prominent.

[11] See also the emerging genre of data journalism: https://en.wikipedia.org/wiki/Data_journalism.

## 3.8 Arts

The renderings of data or information visualization often combine usefulness with attractive presentation - giving the resulting graphics a new, artistic dimension. The term "new digital realism" is used - data being the medium to visualize what exists "but cannot be seen" (Sey, 2015). The analogy to "realism" in classical arts, such as painting, implies that reality is rendered "as it is" - without emotional interpretation or subjective frames. This is of course a tall order, insofar as any visualization consciously puts certain aspects of the data in the forefront, and thus influences subsequent interpretation. In its most pure form, art in the context of data science may pursue the finding of "beauty"[12], not value, in data.

## 3.9 Communication

A totally different, yet important artistic aspect of data science lies in the general challenge of the communication of results. Data products - any findings, any value in the data - rely on interpretations, and this discipline and their proponents need the skills to effectively and truthfully communicate well to stakeholders. We will explore this angle of "communication as a skill" more in the next chapter, where we discuss the profile of a successful data scientist.

## 3.10 Entrepreneurship

Our definition of data science puts the data product in the center: the data product leverages data to produce value and insight (see Chapter 4). The design of a data product is much more than a technical exercise, or even an exercise in leveraging domain expertise in the narrow sense. Questions of how to frame the value proposition of the product, how to identify the right audience and how to find the matching business model arise. These directly address the business-savvy of the data scientist. Generally speaking, a successful data scientist does well to display a degree of entrepreneurial spirit: opportunities to seize value have to be anticipated, based on a deep enough understanding of all three of the following aspects: the technical possibilities, the potential in the data itself, and the need of some “customer”.

# 4. Value creation in data science

The discussion of entrepreneurship brings the overall exposition on data science as a new field of academic study full circle. Whether data science is perceived as an amalgamation of different disciplines, essentially harvesting the synergies of combining technical foundations of computer science with analysis insight from statistics and extending these skills to domain expertise and business-savvy, or whether data science is more seen as a holistic approach to leverage the value of data: the results of applied data science projects typically culminate in data products, that is products that derive their value from the data they are built on. Data products often come in the form of data-driven services. Consequently, the discipline of service science (Spohrer, 2009), which is concerned with, among other things, service innovation, and sits at the intersection of business and information technology, also contributes to the development of data products (see more details in Chapter 4).

[12] For some examples, see (Pickover, 2001).

## 5. Conclusions

In closing the chapter, we want to reflect again on the idea of data science both as a field of study in its own right, and as an umbrella term that allows to describe interdisciplinary endeavors at the interfaces of the disciplines covered above. As stated in the introduction, these two views can be interpreted as "top-down" and "bottom-up". Both views are complementary and enhance the insight into the nature of data science. By covering the various (sub-)disciplines, the bottom-up view of data science as an "unique blend" of the disciplines is represented well. As regards the top-down view, the essence of data science undertakings is the creation of value from data (or information). That thought is not necessarily new. If we look at older fields of study that nowadays contribute to data science projects, such as information retrieval, then similar themes emerge: by making access to relevant information possible, information retrieval "turns information useful". Is data science therefore really more than the sum of its parts? Or could this creation of value take place under different, potentially even more specific, labels in all cases? Or, put differently, if we start with the concept of data science and then remove all the (sub-)disciplines in this chapter, will we be left with something meaningful?

The last of these questions borders on the philosophical, and the answer is probably influenced by where we draw the boundaries of the disciplines. But it may be precisely at the interfaces of the disciplines, or even in the gaps between them, that data science is an enabler for new concepts. The rewards for venturing into these spaces between disciplines and finding new, exciting combinations may be greater than ever. The case studies in the chapters of Part II are a nice testament of the diversity of research questions or business cases that can be pursued.

## References

Aho, A. V., & Ullman, J. D. (1992). Foundations of computer science. Computer Science Press, Inc..

Bateman, S., Gutwin, C., & Nacenta, M. (2008). Seeing things in the clouds: the effect of visual features on tag cloud selections. In Proceedings of the nineteenth ACM conference on Hypertext and hypermedia (pp. 193-202). ACM.

Bellinger, G., Castro, D., & Mills, A. (2004). Data, information, knowledge, and wisdom.

Bishop, C. M. (2007). Pattern Recognition and Machine Learning. Springer.

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. ACM Sigmod record, 26(1), 65-74.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. MIS quarterly, 1165-1188.

Codd, E. F. (1970). A relational model of data for large shared data banks. Communications of the ACM, 13(6), 377-387.

Davenport, T.H., & Patil, D. (2012). Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review, https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

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

Frické, M. (2009). The knowledge pyramid: a critique of the DIKW hierarchy. Journal of information science, 35(2), 131-142.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning. Cambridge: MIT press.

Holcomb, Z. C. (1997). Fundamentals of descriptive statistics. Routledge.

Hughes, J. F., Van Dam, A., Foley, J. D., McGuire, M., Feiner, S. K., Sklar, D. F., & Akeley, K. (2013). Computer graphics: principles and practice. *rd edition. Addison Wesley Professional.

Inmon, W. H. (2005). Building the data warehouse. John Wiley & sons.

Knuth, D. E. (1968). The Art of Computer Programming: Fundamental Algorithms. Addison-Wesley.

LeCun, Y. (2013). "Hi Serge". Google+ post. Available online (May 23, 2018): https://plus.google.com/+YannLeCunPhD/posts/gurGyczzsJ7.

Loukides, M. (2010). What is data science. Available online (June 12, 2018): https://www.oreilly.com/ideas/what-is-data-science.

Luger, G. F. (2008). Artificial intelligence: structures and strategies for complex problem solving, 6th edition. Pearson.

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.

Ramakrishnan, R., & Gehrke, J. (2002). Database management systems, 3rd edition. McGraw Hill.

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

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

Schütze, H., Manning, C. D., & Raghavan, P. (2008). Introduction to information retrieval (Vol. 39). Cambridge University Press.

Sey, M. (2015). Data visualization design and the art of depicting reality https://www.moma.org/explore/inside_out/2015/12/10/data-visualization-design-and-the-art-of-depicting-reality/.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (1997). Database system concepts (Vol. 4). New York: McGraw-Hill.

Spohrer, J. (2009). Editorial Column—Welcome to Our Declaration of Interdependence. Service Science 1(1):i-ii. https:// doi.org/10.1287/serv.1.1.i.

Stadelmann, T., Stockinger, K., Braschler, M., Cieliebak, M., Baudinot, G., Dürr, O., Ruckstuhl, A. (2013). Applied Data Science in Europe: Challenges for Academia in Keeping Up with a Highly Demanded Topic. European Computer Science Summit, ECSS 2013, Informatics Europe, Amsterdam, August 2013.

Stockinger, Kurt; Stadelmann, Thilo; Ruckstuhl, Andreas (2015). Data Scientist als Beruf. Big Data – Grundlagen, Systeme und Nutzungspotenziale, Springer Verlag., Edition HMD 59-81.

Stonebraker, M. (2010). SQL databases v. NoSQL databases. CACM, 53(4), 2010.

Tierney, B. (2013). Type I and Type II Data Scientists. http://www.oralytics.com/2013/03/type-i-and-type-ii-data-scientists.html.

Van Rossum, G., & Drake, F. L. (2003). An introduction to Python. Bristol: Network Theory Ltd..

Von Neumann, J. (1993). First Draft of a Report on the EDVAC. IEEE Annals of the History of Computing, 15(4), 27-75.

Wall, L., Christiansen, T., & Schwartz, R. L. (1999). Programming perl.

Warden, P. (2011). http://radar.oreilly.com/2011/05/data-science-terminology.html.

Ware, C. (2012). Information visualization: perception for design. Elsevier.

Wasserman, L. (2013). All of statistics: a concise course in statistical inference. Springer Science & Business Media.

Wilcox, R. R. (2009). Basic statistics: understanding conventional methods and modern insights. Oxford University Press on Demand.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Woods, D. (2012). bitly's Hilary Mason on "What is A Data Scientist?", Forbes Magazine. https://www.forbes.com/sites/danwoods/2012/03/08/hilary-mason-what-is-a-data-scientist/.