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

Data Scientists

by Thilo Stadelmann, Kurt Stockinger, Gundula Heinatz Bürki and Martin Braschler.

What is a data scientist? How can you become one? How can you form a team of data scientists that fits your organization? In this chapter, we trace the skillset of a successful data scientist and define the necessary competencies. We give a disambiguation to other historically or contemporary definitions of the term, and show how a career as a data scientist might get started. Finally we will answer the above mentioned third question, i.e. how to build analytics teams within a data-driven organization.

1. Introduction

Reading contemporary press, one can come under the impression that data scientists are a rare (Harris & Eitel-Porter, 2015), almost mythical species[1], able to save companies by means of wonderworking (Sicular, 2012) if only to be found (Columbus, 2017). This chapter answers three questions: What is a data scientist? How to become a data scientist? And, how to build teams of data scientists? (see also Stockinger et al., 2015). Answering these questions will help companies to have realistic expectations towards their data scientists; will help aspiring data scientists to plan for a robust career; and will help leaders to embed their data scientists well.

What is a data scientist? As of spring 2018 the ZHAW Datalab, i.e. the data science research institute (DSRI) of the Zurich University of Applied Sciences, has more than 70 affiliated colleagues that "professionally work on or with data on a daily basis"[2]. The lab includes different kinds of researchers, ranging from computer scientists doing analyses with machine learning to domain experts in medical imaging or quantitative finance, to lawyers working on data protection law. Can these colleagues be considered as data scientists? From what we know, not all of these colleagues call themselves primarily data scientists.

However, what is a data scientist? Going beyond the trivial definition of data scientists being those who conduct data science, we can approach the definition by sketching the set of skills and qualities of a data scientist as two layers. Figure 1 contains these two layers of information: First, the blue bubbles show the contributions of several competence clusters to the skill set of the data scientist. Second, the grey labels attached to the data scientist in the center show important qualities of the personality that are paramount for a data scientist's professional success. While the academic (sub-)disciplines underlying these competence clusters were treated in the previous chapter, we will explore their relations to the work of a data scientist in conjunction with the character traits in the next section. In Section 3 we will disambiguate the definition of a data scientist from historical and contemporary alternative meanings. In Section 4 will will show career paths towards data science and finally discuss how to build effective data science teams in Section 5.

---

[1] The British recruiting firm Sumner & Scott was looking for a *"Seriously Fabulous Data Scientist"* on behalf of a customer from the game industry in April 2018.

[2] See http://www.zhaw.ch/datalab for a description of the lab, its statutes and a list of associates.

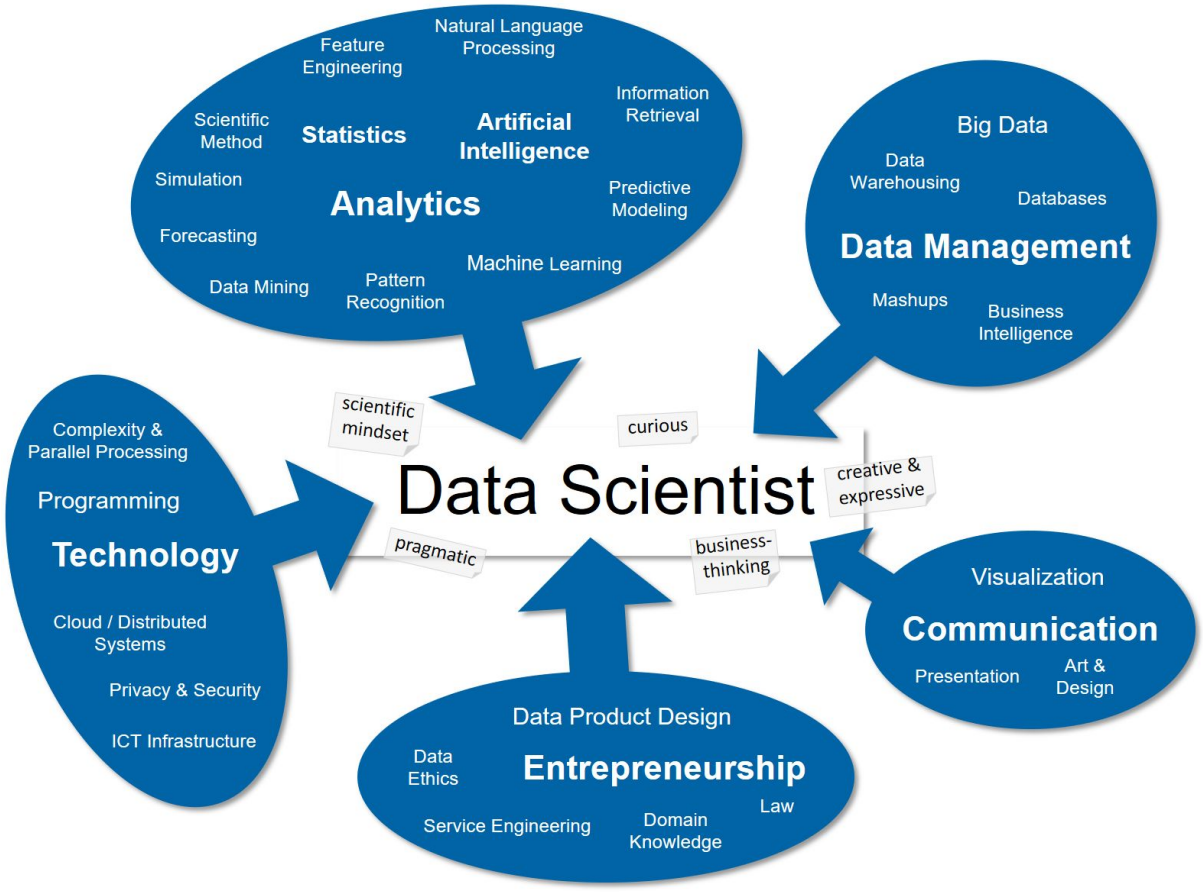

Figure 1: The definition of a data scientist by means of personal qualities (grey labels) and skills (in blue bubbles) as spanned up by this unique cut from several scientific (sub-)disciplines. Adapted from (Stadelmann et al., 2013).

2. The data scientist's set of skills and qualities

Data scientists are T-shaped people (Guest, 1991). They have broad interdisciplinary skills (the crossbar of the "T") and at the same time they have deep expertise (the T's stem) in a much narrower area of this skill set. This section will look at the crossbar and related soft skills, while Section 4 will look at the origins of the stem.

The blue areas in Figure 1 show competence clusters within the data scientist's set of skills. The appearing terms have been selected due to their high likelihood of being important in the daily work on almost any project of a data scientist, in the following sense: in any project, some skills from any bubble will likely be needed, i.e. some method(s) from the "Analytics" cluster but not all of them. We make no claim as regards to the completeness of this term set. Let us have a look at the individual clusters in more detail:

Technology and *Data Management*. Handling data well is crucial enough for any data scientist to make it a top-level competence cluster. Data management includes, but is not limited to, big data technologies, databases and respective query languages like SQL. A background in extract-transform-load processes for data integration and the fundamentals of

relational databases are relevant for many data science projects. The technology cluster includes various other areas from computer science such as the application and handling of software systems. Programming skills are paramount for a data scientist, however, not in the sense of large-scale software development, but in the form of scripting e.g. for data wrangling tasks. Combining small scripts in the spirit of the UNIX command line with each other (Peek et al., 1993) allows for rapid prototyping as well as repeatability of experiments[3]. It possibly also helps for executing analyses in different, even distributed environments.

Analytics. Skills in analytics, especially machine learning, are one of the core competencies of a data scientist to extract knowledge from data. The two main approaches to analytical methods come from the fields of statistics (Wasserman, 2013) and artificial intelligence (Russell & Norvig, 2010); the two fields often provide different individual approaches to similar methods. While discussions arising from these differences in viewpoint are challenging for any practitioner in a data science team, they are also a source for mutual interdisciplinary understanding, and hence are very valuable. This has been analyzed thoroughly by Breiman (2001).

Entrepreneurship. Data scientists are not only responsible for the implementation of an analytical solution for a given problem. Rather, they additionally need entrepreneurial skills to ask the right questions with respect to business cases, business models and the consequences of the data products on the business and society at large. This includes building up subject matter expertise in the application areas of the data product at hand, and the appreciation of the personal ethical responsibility. As many questions in data science touch on fundamental issues in privacy, data scientists must have knowledge of legal frameworks of their operating environments.

Communication. Being responsible for the complete analytical workflow, data scientists personally communicate their results to (senior) management. Needed skills thus range from targeted presentation to information visualization, in order to convey complex matters and conclusions in concise ways. It is questionable[4] if the creation of (graphical user interfaces for) web services for the final customers of data products should be a core part of the data science skill set.

The second layer of information in Figure 1 shows personality trait labels attached to the data scientist. Being more part of a person's character than the skill set, it seems a bit unfair to require them for a job as widespread as a data scientist. On the other hand, it is a matter of fact that certain jobs fit specific people (Fux, 2005). So what is the impact of these qualities on a practitioner's work?

---

[3] Experiments that are controlled purely by scripts are repeatable by archiving the code as well as data together with the results. They can be developed rapidly by re-using scripts from other projects (which is easier when every script serves exactly one purpose and uses a simple file-based API, as UNIX shell programs do) and automatizing parameter search.

[4] We see this better placed in the hands of software engineers; nevertheless, being able to build rapid prototypes in the way presented in the "Developing Data Products" course by the Johns Hopkins University (see https://www.coursera.org/learn/data-products) is an interesting additional skill for any practitioning data scientist.

Creativity & expressiveness. Both traits help in giving convincing presentations of the data scientist's work for internal stakeholders and potential customers. Creativity reaches even farther in also being a necessity for creating novel results. This plays into the next point:

Curiosity and *scientific mindset*. Curiosity pairs well with enthusiasm. A scientific mindset will balance utter positivism with basing one's hopes and findings on facts through critical hypothesizing, thorough experimentation[5], and precise formulation of results. Doubt and amazement are both important ingredients for novel solutions.

Business thinking. Thinking economically helps to have a clear goal in mind, on several levels: it contributes to not losing sight of the complete development process of a data product when concentrating, for example, on the analytical challenges at hand; it also helps in allocating resources to the various steps in a project and weigh options in order to produce business-relevant results. This will ultimately drive the success of analytics endeavors, since most stakeholders (in businesses, research or society) will be convinced not by the coolness of the engineering, but by actually helpful results delivered on time and on budget.

Pragmatism. The quality to do rapid prototyping and quick experimentation cannot be underestimated. The analytical work of a data scientist is inherently empirical, and having a drive towards experimenting and getting one's hands dirty with code and messy data is paramount in making progress in many projects. A special sort of pragmatism with respect to coding - specifically, to be able to abstain from undue perfectionism in software engineering in early project phases in favour of "hacking"[6] - and system design (specifically, to use simple scripts in a command-line like fashion) helps in keeping efficiency high in usually very complex tool landscapes.

3. Disambiguation

The following paragraphs deal with disambiguating the definition of a data scientist as presented above from other meanings used previously or contemporary.

3.1 The history of a job description

Probably the first one publicly speaking about data scientists was Jeff Wu (1997), who suggested to use the term as a replacement for "statistician". The modern use presented in the previous section arguably emerged out of discussions between DJ Patil and Jeff Hammerbacher on how to call their team members at Facebook and LinkedIn, respectively, that were closer to product development than a usual "research scientist", but with more

[5] The scientific method of theory formation (thinking) and collecting empirical evidence (experimenting) in a closed loop is directly applicable in the daily work of a data scientist. See a longer exposition of this thought in (Stadelmann, 2017) and an extension in a later chapter by Brodie ("On developing data science").

[6] Take this with a grain of salt: as much as we plead for hacking to facilitate rapid prototyping, especially for our audience with a background in computer science, the more we know about the importance of careful software engineering for production-ready code. See also (Zinkevich, 2018) for good advice for the latter (and the former).

technical skills than a typical "business analyst" (Patil, 2011). Previously, this profile had been called "deep analytical talent" in a noteworthy report from the McKinsey Global Institute (Manyika et al., 2011) and was famously rendered in graphic form by Drew Conway (2010) in his "data science Venn diagram". The diagram conveyed the notion that a data scientist works at the intersection of hacking, math (or statistics) and substantive expertise, thereby discriminating it from traditional research (no hacking), pure machine learning (no subject matter expertise) and a "danger zone" (no math). Patil and Hammerbacher added that their data scientists should also be able to communicate their own results to business owners on top of the deep engineering know-how embraced also by Conway.

The following years saw a race to ever more elaborate versions of the skill set of a data scientist, packed into Venn-like diagrams (Taylor, 2014). In very short time, deep analytical talent was inflated to unicorns (Press, 2015), marketed towards C-level executives as the ones finally being able to *"align IT and business"* (Jones, 2014). Additionally, expectations towards technical skills grew as well. For example, Brodie (2015a) pointed out the importance of data curation that involves the work on and with large IT systems at scale in preparation of the actual data analysis[7].

Moreover, data scientists were supposed to carry huge responsibilities due to the disruptive potential of the paradigm of data-driven decision making (Needham, 2013). This raised the requirement on them to make the attached risks of their work explicit, for example, by attaching common measures of correctness, completeness and applicability to data science results (Brodie, 2015a) such as confidence intervals for all quantitative results[8]. The necessity for some measures to this effect becomes apparent when regarding analysis results from higher-dimensional data: in dimensions beyond three, human intuition even of experts fails completely, known under the term "curse of dimensionality" (Bellman, 1961). Accordingly, the audience in a presentation of respective results could be easily and unintentionally mislead (Colclough, 2017)[9], drawing fatal business decisions from misinterpretations of analytical results.

However, an informal survey amongst the ca. 190 participants of the SDS|2015 conference[10] revealed that only about 50% of the practicing data scientists apply any counter-measures against misinterpretation or illusory certainty - also because this is not required of them by their customers (internally or externally). However, a data scientist is a scientist: this means following sound scientific practice to not let one's own biases and presuppositions overrule experimentally established facts (Brodie 2015b).

---

[7] Such systems are used to find, gather, integrate and manage potentially heterogeneous data sources. This adds up to about 80% of the daily work of a practicing data scientist (Brodie, 2015a).

[8] However, the debate in the *Journal of Basic and Applied Social Psychology* on the removal of p-values from all published articles because of a theoretical invalidity of the null hypothesis significance testing procedure (Trafimow & Marks, 2015) shows: reporting confidence intervals per se is no panacea as it "suffers from an inverse inference problem" as well.

[9] Colclough (2017) notes that just putting a data visualization on a slide often brings credibility to its statement, no matter the content of the visualization nor its correctness.

[10] See https://www.zhaw.ch/en/research/inter-school-cooperation/datalab-the-zhaw-data-science-laboratory/sds2015/.

3.2 Insightful debates

Two additional debates provide insight on what can or cannot be expected from a modern data scientist:

First, the trend in the mid-2010s to make data science results and careers more easily accessible for a larger number of people (and customers) who might not have formal education in computer science, maths or statistics. As a side effect, the profile of the profession might dilute as the work of a data scientist is reduced to the operation of self-service BI tools. On the other side of the same medal, complex and scientifically unsolved problems like social media monitoring (Cieliebak et al., 2014) are promised to get solved at the push of a button. While natural language processing has certainly made progress to the point of applicability in many cases, it is not solved in general - and which business owner can distinguish his very special use case that requires a great deal of generality from the superficially similar demonstration that works in a quite constrained environment[11]?

Seen in relationship with the above mentioned responsibility of a data scientist for potential good or harm *at scale*, this development might be considered dangerous. It needs certain skills to draw correct conclusions from data analytics results; it is thus important to keep the science as an important ingredient in the data scientist. Business analytics is a part and not a superset of data science; vice versa, not all data science challenges could and should be approached using readymade BI tool boxes or BI engineers[12]. This leads over to a second debate:

Second, there is a notion of data scientists "type A" and "type B"[13]. While "type A" are basically trained statisticians that have broadened their field ("data science for people"), "type B" have their roots in programming and contribute stronger to code and systems in the backend (B for "build", "data science for software"). So, two of the main influences for data science as an interdisciplinary field - computer science and statistics - are taken apart again to emphasize a less interdisciplinary profile[14].

Seen from the viewpoint of interdisciplinarity as a key concept for the data scientist makes this (and similar) distinctions between mono-disciplinary rooted types of data scientists useless. The whole point of interdisciplinarity, and by extension of the data scientists, is for proponents to think outside the box of their original disciplines (which might be statistics, computer science, physics, economics or something completely different), and acquire skills

[11] See http://xkcd.com/1425/. While the described phenomenon might be easy to solve in the year 2018, a contemporary example would be chatbots.
[12] The same applies to automated machine learning, although such systems have a real value for certain applications.
[13] Jaokar (2016) refers to a quora post by Michael Hochster for the origins: https://www.quora.com/What-is-data-science.
[14] T-shaped people will have their roots - their depth - mostly in one field; hence, the problem described here arises not from different expressions of the T-shape per se, but from specifically differentiating what lead to the notion of a data scientist in the first place: combining computer science and statistical know-how (see Section 3.1).

in the neighboring disciplines in order to tackle problems outside of intellectual silos. Encouraging practitioners to stay in their silos, as suggested by the A/B typology, is counterproductive, as it is able to quench this spirit of out-of-the-box thinking.

The debate, however, is well suited in that it challenges the infamous - and often unrealistic - "unicorn" description of a data scientist who is supposed to be an "expert in each and everything". A concept that addresses the same concern but arrives at different conclusions is the one of data scientists "Type I" and "Type II"[15]. "Type II" data scientists are managers, concerned with hiring and leading data practitioners and having a more high level view of data sciences' potentials and workings. On the other hand,"Type I" data scientists know how to "do the stuff" technically. This opens up the way to combined curricula for manager-type people and technically-oriented people (different roles) while not compromising the interdisciplinary profile of either of them.

On the other hand, the attempt to isolate sub-types of a "Type I" comes down to merely re-labeling traditional job titles like statistician, business analyst, BI specialist, data miner, database engineer, software engineer, et cetera. If these titles fit the role, i.e. accurately describe the skill set and breadth of the job description, they are still appropriate and very precise. If the job, however, requires the broader experience of a data scientist - the crossbar of the T instead of just the stem -, this could be indicated using the proper description of data scientist. Problems arise if an expected but missing crossbar experience leads to weakening the credibility of the discipline of data science.

4. Starting a data science career

If data scientists are interdisciplinary by nature with T-shaped skill profiles, trying to define what a data scientist is comes down to giving bounds on the width of that T's crossbar (how much interdisciplinary experience is necessary?) and the height of the T's pole (to what degree is this a specialist in some subset of skills?). The following bounds are subjective (as being based on personal experience), but can serve in giving guidelines as to what to expect from a senior data scientist, with our definition of "coverage" following in the next but one paragraph.

As for the crossbeam of the "T", a senior data scientist should cover a majority - we guesstimate ca. 80% - of the terms on the competency map in Figure 1, distributed over all five of the blue competence clusters. This usually means that individual senior data scientists should be firmly anchored in one of these clusters and having a solid understanding in at least two others without avoiding the balancing act between the technical-analytical and entrepreneurial-communicative hemispheres. Thus, the necessary interdisciplinary breadth of knowledge is ensured without calling upon mythical beasts.

[15] The naming itself is not important. The concept has been incorporated into the academic MSc Data Science programme of the University of Sheffield (see https://www.sheffield.ac.uk/postgraduate/taught/courses/sscience/is/data-science) and seems to go back to Tierney (2013).

The intended "covering" means that the data scientist should have an understanding of the respective terms (e.g., "Natural Language Processing" within the "Analytics" cluster, or "Law" within "Entrepreneurship") deep enough to recognize certain opportunities or potential issues arising in the current project from this domain, and in case of doubt can involve an expert. This level of understanding is usually gained by the equivalent of working hands on with the topic for a limited time or doing a typical one semester introductory course, i.e., it is not expert-level knowledge. The necessary skills can be trained, given a disposition to quantitative, complex and technical matters.

Regarding the stem of the "T", a typical career starts with undergraduate studies, for example, in statistics, computer science or another data-intensive science. From there, interdisciplinary skills can be built either by a data science master's degree, hands-on collaborations with other disciplines or continuing education. If personal interests suggest a closer look into research, a PhD is a good option, but not all education has to be formal[16]. In our experience, it is more important to show a good track record of projects one was engaged with in order to qualify for advertised data scientist positions. Projects in this regard is a loose term - included are also personal projects or those that are part of course work. What counts is the demonstration of gained experience, e.g. by cultivated personal GitHub pages or blogs, published research articles, or by contributions to publicly available products. Certificates themselves are not sufficient due to them becoming more and more omnipresent amongst candidates.

A data science curriculum - whether offered by any institution in the higher education sector, or self composed - should address the following three levels (measured in terms of distance to actual cases studies that could be solved). The content of the *business* layer is close to the case study that needs to be grasped in detail by the data scientist. This influences the choice of *algorithms* in the next layer, but is more or less independent from the technical *infrastructure* in layer 3.

1. *Business*:
    - visualization & communication of results
    - privacy, security & ethics
    - entrepreneurship & data product design
2. *Algorithms*:
    - data mining & statistics
    - machine learning
    - information retrieval & natural language processing
    - business intelligence & visual analytics
3. *Infrastructure*:
    - databases, data warehouses & information systems
    - cloud computing & big data technology

[16] Especially in the context of data and computer science, online courses like the ones offered by Coursera (https://www.coursera.org/specializations/jhu-data-science) or Udacity (https://www.udacity.com/nanodegree) have a good credibility.

Ideally, such a curriculum considers this intimate connection between the application of data science in actual cases studies, on the one hand, and the fundamentals of data science like details of methods, on the other hand, already in the coursework. This can be achieved by connecting the relevant theory with project work in certain problem domains like e.g. predictive maintenance (industry), medical imaging (health), algorithmic trading (finance), churn prediction (marketing), speech processing (technology), building control (energy) etc. These case studies run cross to all three layers from above.

We see the analytical aspects as central to any data science education: machine learning, statistics and the underlying theories have to be solidly mastered by any data scientist in order to decide on feasibility and perform impact assessment. These skills - the "deep analytical talent" or "deep engineering know-how" as it has been called by various thought leaders - are the ones most deeply learned early on in one's vocational career (i.e., better studied thoroughly for years than acquired using a crash course). They are also the ones that host the greatest potential both in terms of risks and opportunities when unleashed on the world. Data scientists thus owe a responsible mastership of the engineering aspects to their environment.

5. Building data science teams

Finding a senior - mature, "complete" - data scientist as described in the previous section might be difficult for an employer. Even if it was not so, it is advisable to let data scientists work in teams where the crossbows of the respective team members' T's overlap considerably, but the poles dig into different territory of the skill set map (Olavsrud, 2015). This way, not only can the less wide crossbars of less senior data scientists be integrated; rather, the full potential of interdisciplinarity can be leveraged. How should such teams be embedded into the organization?

Executive-level support for establishing data and analytics as a strategic capability is one of the key success factors for enabling a company to do data-driven, automated decision making. We will look at the following two main aspects of building data science teams[17]:

1. Shape an adequate operational model for the organization's advanced analytics capabilities and associated governance.
2. Identify data-driven use cases that have a big impact on the business and therefore the most added value.

5.1 Operational models for advanced analytics

Three main operational models exist for building a common data-driven culture in an organization (Hernandez et al., 2013). For an organization to decide for a specific one, this operational model has to align with the enterprise strategy first. Second, the complexity and maturity of the enterprise regarding data-driven decision making is relevant. The choice of model thus depends on the organization's structure, size and experience in this topic. The three models are as follows:

[17] Other aspects are highlighted for example by Stanley & Tunkelang (2016).

(a) Centralized unit within the IT or finance department:
The structure of such an organization is simple and focuses on allocating limited resources to strategic topics. The typical enterprise choosing this model already has mature reporting and analytics capabilities, with both the IT and the finance department having already acquired the necessary skills. The centralized unit thus has the technical prerequisites of the IT or statistical knowledge of the finance department because of this previous work and provides the expertise to the business units. This model fits well to most small and simple organizations.

(b) Cross-business unit with data scientists:
Again, experienced data scientists belong to a centralized group, where they are responsible for analytical models, standards and best practices. But these data scientists establish contacts to domain experts or even other analytical groups in the business units, as all business units have mature basic analytical skills. This model can be seen as a "hub and spokes" approach compared to the purely central model (a). It fits to moderately more complex organizations that see data as a core competitive advantage.

(c) Decentralized data science teams in several business units:
Here, every business unit engages its own data science team because the necessary business knowledge is domain specific. This business-specific analytical knowledge is significant to succeed. This model thus fits to highly complex, mostly large organizations with autonomous business units.

Due to its low requirements, the *operational model (b)* has the potential to be broadly implemented in practice. For its successful realization, it is relevant to consider which resources are available in the data science teams, such as data science skills, technology and data with sufficient quality. In the beginning, a central interdisciplinary team consisting of experts with different deep focuses such as machine learning, natural language processing or spatial analysis is formed. Business domain experts support these data scientists to implement high-quality business-related solutions. Scripting capabilities are amongst the core competencies, as they come into play in all phases of solution creation, from data extraction and transformation to system integration and finally building interactive dashboards for the user.

A good collaboration with the IT department is essential to ensure the work with an analytical sandbox that results in high-quality prototypes and products. A framework "from pilot to production" and defined architectures are decisive for becoming sustainably successful. Crucially, the unit should be supported by an adequate governance. A *steering board* assists with strategic decisions and work prioritization. An internal *data science meetup* presents exciting use cases to interested employees. Additional, a close relation to renowned local universities is beneficial to learn about the newest methodologies and remain at the state of the art.

5.2. Data-driven use cases

To start with the data-driven journey, organizations need to identify their crucial challenges with the most impact on their business. Then, analytics can support the process of finding a solution towards a new or updated product or service.

To spot relevant use cases, enterprises often get input from the market via the support of different consultants. Another opportunity is visiting industry-related conferences[18]. Design thinking approaches with cross-disciplinary teams, consisting of business people and data scientists, additionally help to detect use cases with strategic impact. A significant collection of key use cases can inspire an enterprise for the further journey. For the use case prioritization, a framework based on two dimensions is usually applied: estimated business benefits vs. the effort of investment in time or money (or complexity). The result is a roadmap of prioritized, high-value use cases together with the anticipated benefit, and consequently, it is possible to define quick wins: new problem-solving approaches that could be implemented very quickly. In addition, this method allows for the efficient identification of the most critical and therefore most valuable use cases for the company. By considering all existing resources, the use case portfolio can be managed well.

After use case prioritizations, it is helpful to start the first pilot with the support of an excited business sponsor. In the future he or she can be designated as an ambassador for the new analytical approach. The CRISP-DM approach (Wirth & Hipp, 2000) is often adopted in data science projects. When the first phase of piloting confirms the benefits for the business, a handover to the IT department helps to sustainably maintain the solution. Finally and step by step, the results and the knowledge about the new methodologies conquer the daily business (a more detailed overview of the creation of data products is presented in the next chapter).

6. Summary

We have presented the modern data scientist as a truly and inherently interdisciplinary professional with deep analytical and engineering know how, able to think entrepreneurially and to communicate results in various appealing ways. No matter if one wants to hire or to become a data scientist - there are two pitfalls attached to this definition of a data scientist:

1. The danger of *canonization*: unicorns, as data science all-rounders are often called, do not exist. Any attempt to become or find one are headed for disappointment. The solution is to acknowledge that a senior data scientist should have a reasonable understanding of the majority of the data science skill set map (the crossbeam in "T-shaped people"), while going deep in only a very restricted area of the map (the stem of the "T"). Rather than chasing unicorns, one should view data science as teamwork, with the team of data scientists together covering the skill set map with complementary specializations.

[18] For example, Strata (https://conferences.oreilly.com/strata) and Predictive Analytics World (https://www.predictiveanalyticsworld.com/) are internationally renowned.

2. The danger of *trivialization*: as finding data scientists becomes harder and being one becomes more profitable, there are natural market tendencies to dilute the skill profile and misuse the fashionable name for conventional job descriptions. Both may lead to wrong decisions due to mishandled complexity.

We presented a data science career as one rooted in one of many potential undergraduate degrees (computer science, industrial mathematics, statistics, or physics are prime candidates) that lays a solid disciplinary foundation (likely connected with the stem of this data scientist's "T"). On this, a data science master's degree can be pursued, or skills can be extended through other, more informal ways of continuing education in order to establish the crossbeam and the personal specialization.

Finally, we gave insights into the development of data science teams. Three operational models for advanced analytics depending on the organization's structure, size and experience were presented. Besides associated governance, the exploitation of strategic use cases is key to be sustainably successful.

References

Bellman, R. (1961). Curse of dimensionality. Adaptive control processes: a guided tour. Princeton, NJ.

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 - Principles and Techniques for Data-Intensive Analysis. Keynote talk the 2nd Swiss Workshop on Data Science SDS|2015. Available online (April 04, 2018): https://www.youtube.com/watch?v=z93X2k9RVqg.

Brodie, M. L. (2015b). Doubt and Verify: Data Science Power Tools. *KDnuggets*. Available online (April 05, 2018): https://www.kdnuggets.com/2015/07/doubt-verify-data-science-power-tools.html.

Cieliebak, M., Dürr, O., & Uzdilli, F. K. (2014). Meta-Classifiers Easily Improve Commercial Sentiment Detection Tools, LREC.

Colclough, A. (2017). When Data Visualization Goes Wrong and Numbers Mislead. Available online (April 04, 2018): https://www.dwrl.utexas.edu/2017/12/29/when-data-visualization-goes-wrong/.

Columbus, L. (2017). IBM Predicts Demand For Data Scientists Will Soar 28% By 2020. Available online (April 05, 2018): https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#6cf2d57e7e3b.

Conway, D. (2010). The Data Science Venn Diagram. Available online (April 04, 2018): http://www.dataists.com/2010/09/the-data-science-venn-diagram/ (figure available from http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram).

Fux, S. J. (2005). *Persönlichkeit und Berufstätigkeit: Theorie und Instrumente von John Holland im deutschsprachigen Raum, unter Adaptation und Weiterentwicklung von Self-directed Search (SDS) und Position Classification Inventory (PCI)*. Cuvillier Verlag.

Guest, D. (1991). The hunt is on for the Renaissance Man of computing. *The Independent* (London), September 17, 1991. Quoted by (available online: April 04, 2018): https://wordspy.com/index.php?word=t-shaped.

Harris, J. G., & Eitel-Porter, R. (2015). Data scientists: 'As rare as unicorns'. *The Guardian*. Available online (April 05, 2018): https://www.theguardian.com/media-network/2015/feb/12/data-scientists-as-rare-as-unicorns .

Hernandez, J., Berkey, B., Bhattacharya, R. (2013): Building an Analytics-Driven Organization. Available online (June 09, 2018): https://www.accenture.com/dk-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Industries_2/Accenture-Building-Analytics-Driven-Organization.pdf.

Jaokar, A. (2016). How to Become a (Type A) Data Scientist. *KDnuggets*. Available online (April 04, 2018): https://www.kdnuggets.com/2016/08/become-type-a-data-scientist.html.

Jones, A. (2014). Data Science Skills and Business Problems. *KDnuggets*. Available online (April 04, 2018): http://www.kdnuggets.com/2014/06/data-science-skills-business-problems.html.

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.

Needham, J. (2013). Disruptive Possibilities - How Big Data Changes Everything. O'Reilly Media, ISBN 978-1-449-36567-7.

Olavsrud, T. (2015). Don't look for unicorns, build a data science team. *CIO.* Available online (May 28, 2018): https://www.cio.com/article/3011648/analytics/dont-look-for-unicorns-build-a-data-science-team.html.

Patil, D. (2011). Building data science teams. Available online (April 04, 2018): http://radar.oreilly.com/2011/09/building-data-science-teams.html.

Peek, J., O'Reilly, T., & Loukides, M. (1993). *UNIX Power Tools*. Oreilly & Associates Incorporated.

Press, G. (2015). The Hunt For Unicorn Data Scientists Lifts Salaries For All Data Analytics Professionals. Available online (May 28, 2018): https://www.forbes.com/sites/gilpress/2015/10/09/the-hunt-for-unicorn-data-scientists-lifts-salaries-for-all-data-analytics-professionals/.

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach, 3rd Edition*. Pearson Education Inc., New Jersey.

Sicular, S. (2012). The quest for data scientists. *The Australian Business Review*. Available online (April 05, 2018): https://www.theaustralian.com.au/business/business-spectator/the-quest-for-data-scientists/news-story/eab27147e92d0011520f5adb32010e43.

Stadelmann, T. (2017). Science, applied. Die wissenschaftliche Methode im Kern des Produktentwicklungsprozesses. AIssays blog. Available online (April 06, 2018): https://stdm.github.io/Science-applied/

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.

Stanley, J., & Tunkelang, D. (2016). Doing Data Science Right - Your Most Common Questions Answered. *First Round Review.* Available online (June 11, 2018): http://firstround.com/review/doing-data-science-right-your-most-common-questions-answered/.

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.

Taylor, D. (2014). Battle of the Data Science Venn Diagrams. Available online (April 04, 2018): http://www.prooffreader.com/2016/09/battle-of-data-science-venn-diagrams.html.

Tierney, B. (2013). Type I and Type II Data Scientists. Oralytics Blog, March 22, 2013. Available online (April 04, 2018): http://www.oralytics.com/2013/03/type-i-and-type-ii-data-scientists.html.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37:1*, 1-2, DOI: 10.1080/01973533.2015.1012991.

Wasserman, L. (2013). *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39).

Wu, J. (1997). Statistics = Data Science? Inaugural Lecture at University of Michigan, Ann Arbor. Available online (April 04, 2018): https://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf.

Zinkevich, M. (2018). Rules of Machine Learning: Best Practices for ML Engineering. Available online (April 06, 2018): https://developers.google.com/machine-learning/rules-of-ml/.