Beyond Unicorns: Educating, Classifying, and Certifying Business Data Scientists

Thomas Davenport^{1,2}

¹Division of Operations and Information Management, Babson College, Babson Park,
Massachusetts, United States of America,

²International Institute for Analytics, Portland, Oregon, United States of America

Published on: May 19, 2020

DOI: <https://doi.org/10.1162/99608f92.55546b4a>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

There is increasing recognition that the data scientist ‘unicorn’—one who can master all the necessary skills of data science required by businesses—exists only rarely, if at all. Successful data science teams in business organizations, then, need to assemble people with a variety of different skills. This is only possible at scale with clear classification and certification of skills. While such certifications and classifications are in their early days, some firms are beginning to create them, and they are beginning to emerge in professional associations as well. Ideally, universities and other education providers and certifiers of data science skills would also employ standard skill classifications to communicate the skills they intend to inculcate.

Keywords: data science skills, data science job definitions, analytical skills, skill classification, certification

1. Data Science Unicorns Really Don’t Exist

Data science is a new and popular, but difficult-to-define field. Its true age (Donoho, 2017), its relationship to previously existing fields like statistics (Gelman, 2013), and the nature of its ‘true’ practitioners (Gutierrez, 2019) are widely discussed and debated. As Alan Garber, the Provost of Harvard University, put it in the first issue of this journal, “the pervasive use of the term ‘data science’ in academic settings reflects both the appeal of the intellectual activities it encompasses and the capaciousness—or vagueness—of its meaning” (Garber, 2019). And since it is probably safe to say that academics care more about clear definitions of disciplines and terms than do businesspeople, there may be even less clarity about what constitutes data science and data scientists in the business domain.

But, however it is defined, data science is increasingly a mission-critical activity for businesses and organizations, and one involving a variety of tasks and skills. As firms employ data science, big data, and artificial intelligence (AI) to enable and redesign important products and processes, the tools become a major component of business change. In a [2018](#) Deloitte survey of U.S. executives, 56% believed that AI would transform their companies within three years (Loucks et al., 2018).

Business transformation—bringing about radically improved business capability and performance—through data science requires talented people, either hired with the needed skills or trained to possess them. And as data science becomes a driver or enabler of business transformation, the skills required become broader. A common assumption behind hiring and educating data scientists is that each individual will need all the skills to perform data science—statistics, data engineering, systems development, people management, and even organizational change management. One list of required skills for data scientists, for example, includes 19 analytical skills, nine “open-mindedness skills,” 15 communications skills, 10 mathematical skills, 11 programming skills, and if that weren’t enough, an additional 26 “More Data Science Skills” (Doyle, 2019). Mastering the total of 90 skills would seem to be beyond even a data science superhero.

The ‘self-sufficient unicorn’ assumption ignores the reality of modern data science in business, in which no one can possibly be qualified at all of these required capabilities, and individuals work in teams in which members have a variety of different roles and skills. In practice, then, the data scientist takes on a variety of different forms and specializations: the statistical data scientist, the computational data scientist, the strategy consultant data scientist, and so forth. Indeed, this specialization in backgrounds and activities has been consistently present since the beginning of the field. As an anecdote, when I interviewed more than 30 data scientists almost a decade ago for an article with D. J. Patil (who says he co-suggested the first use of the data scientist term for a business role at LinkedIn in 2008), I found that the most common academic background was experimental physics, but there were also data scientists with backgrounds in astrophysics, statistics, sociology, meteorology, artificial intelligence, and many others (Davenport & Patil, 2012). And they performed a variety of different types of tasks at that time as well.

As with unicorns in myth and legend, the word is out that self-sufficient data science unicorns don’t actually exist in the business world. As one IT press account put it:

The data science unicorn is a somewhat mythical person who is a leader in data science, technology, and business. Of course, these candidates practically don’t exist, nor do they necessarily make strong team members. As data science teams have grown, businesses have moved away from trying to find that one person to fill different roles; instead, companies have realized the benefits of hiring employees with specialized, complementary skills. (Zhang, 2019)

However, postunicorn thinking has not yet penetrated how individuals and organizations think about and structure data science education and capability-building. For that thinking to take hold, individuals need to educate themselves in preparation for specialization and collaboration. Companies and organizations that use data science need to think about the educational needs of teams rather than individuals, and to create classifications of skills and jobs to make their resources and needs clear. Educational institutions need to structure their data science offerings for different specializations. And companies—and ideally the entire society—need to develop certification and classification structures that make visible and reliable the different types of skills that data scientists possess.

2. Business Data Science Teams Require Multiple Skillsets

Data science teams’ assignments and responsibilities vary across organizations, of course, but in general require that the following types of skills be present across the entire team (modified from Davenport, 2018b):

- *Quantitative and technical skills* are the primary differentiating skills for the profession. A data science team must have members who are proficient in both inferential and predictive statistical analysis and modeling, and their advanced applications in fields like deep learning neural networks. Some members will also need familiarity with the quantitative disciplines specific to their industry or business function: lift or attribution analysis in marketing, stochastic volatility analysis in finance, biometrics and population health analyses in

pharmaceuticals, and informatics in health care, for example. Data scientists must also know how to use (and often write code in) the specific software associated with their type of analytical work, whether it be to build statistical or AI models, generate visual analytics, define decision-making rules, create simulations, or embed analyses into a business dashboard. The most commonly employed data science software tools in the current environment, according to Robert Muenchen at the University of Tennessee, who compiles an annual ranking, are Python, SQL, Java, Amazon Machine Learning, and R (Muenchen, 2019). This is a very different list from that of a decade ago—when proprietary statistical packages were at or near the top of the list—which means that data science teams will need substantial retraining offerings for all but the most recent graduates.

- *Business knowledge and consulting skills* enable data scientists to be more than data analysts and coders who generate well-fitting models for their data. They must be familiar with the business functions and processes to which data analysis is being applied—marketing, finance, HR, new product development, and so forth—in order to diagnose and frame the problem to be solved and understand whether data science models will be effective in solving it. They need enough general business background and acumen to work at the ‘coal face’ of business processes and problems. They also must have insight into the key opportunities and challenges facing their companies, and know how analytics and AI can be used to drive business value. Of course, these behaviors were often practiced by effective industrial statisticians and operations researchers before the data scientist term became popular.
- *Data management skills* are perhaps even more important to data scientists than statistical and mathematical expertise, at least when ‘big data’ (also a difficult term to define, but generally involving large and relatively less-structured data formats) is their primary focus. A 2012 (non-peer reviewed) study of quantitative analysts found that they spent 23% of their time on average on data collection and preparation—more than any other activity. The most data management-focused analysts spent 46% of their time on that activity (Roberts & Roberts, 2013). Python, SQL, and Java—the three most in-demand software tools for data science—are as much data management tools as analytical ones. For many data scientists, data management activities will constitute a major component of their jobs.
- *Relationship and communication skills* enable data scientists to work effectively with their business counterparts to conceive, specify, pilot, and implement analytical and AI applications. Relationship skills—advising, negotiating, and managing expectations—are vital to the success of all data science projects. Furthermore, a data scientist needs to communicate the results of analytical and computational work (Malone, 2020) for a variety of purposes. Within the business, there is a need to share best practices and to emphasize the value of data science projects; outside the business, to shape working relationships with customers and suppliers. It may also be necessary to explain the role of analytics and AI in meeting regulatory requirements (e.g., utility company rate cases or credit decisions in financial services). This skill has been described as “telling a story with data” (Davenport, 2014).
- *Coaching and staff development skills*, as in other types of teams, are essential to a data science organization. The need is particularly great when a company has a large or fast-growing pool of data scientists, or when its

quantitative talent is spread across business units and geographies. All data science professionals may not need these skills, but they are certainly required for supervisors and managers of large teams. Such individuals who lead data science teams must be data scientists themselves, or at least very familiar with data science issues and the concerns of data scientists. When talent isn't centralized, coaching can ensure that best practices are shared across the company. Good coaching not only builds quantitative and technical skills, but also helps people understand how data-driven insights can drive business value.

Some individuals may focus on one or two of these skillsets, and be conversant with others. A degree of overlap and redundancy in skills can often make innovation-oriented teams more effective (Nonaka, 1990). This perspective is consistent with the idea of a “T-shaped” data scientist, who has a breadth of relevant skills but depth in one area (Vaisman et al., 2013). There may also be a need for “pi-shaped” data scientists, with two or more areas of depth (Friedlein, 2012).

While technical and analytical skills have been the hallmark and primary differentiator of the data scientist role, the balance of needed skills is likely to change over time. Such changes have taken place throughout the history of data analysis, such as when commercial statistical packages were first marketed in the late 1960s and early 1970s. Today, the development of automated machine learning tools, for example, may mean that many repetitive and time-consuming tasks in machine learning are performed automatically by software, while the more human-oriented skills like problem framing, business and data acumen, relationship development, and consulting remain and become more important for some data scientists to possess (Abbasi et al., 2019). These automated tools are of most value to ‘citizen data scientists’ who have limited quantitative skills and who perform less complex analyses.

Today there are many different analytics and data science programs offered by educational institutions with the goal of creating data scientists for employment in business—over 200 in accredited business schools alone (Davenport, 2018a)—and many others in computer science and engineering schools. However, there is little consensus among these institutions as to which of the data science skills should be taught. The programs generally do not tell students which types of skills they will learn in that program and what types of jobs students will be prepared to take when they graduate. And many students know too little about the field to know what topics and required courses to look for in a curriculum. Irizarry (2020) has argued for at least three different tracks of data science education—data engineer, data analyst, and machine learning engineer—but these are not yet reflected in curricula.

3. Enterprise Data Science Job Role and Skill Structures: An Example at a Large Bank

It is unlikely, of course, that any individual would possess all data science skills to a high degree. And in practice, there are many different types and levels of data scientists within organizations (Berthold, 2019). Therefore, it's important to form or educate teams that have different types of data scientists in the needed

types and quantities. In order to do this effectively, firms need to have clear definitions of jobs and the skills that are needed to perform those jobs successfully. This is particularly critical for large organizations that have placed a high strategic priority on analytics and data science. Some employers are beginning to create such definitions and classifications.

For example, a large bank headquartered in North America realized that it had a poor understanding of its analytics and data science talent, so it embarked upon an “enterprise talent workstream” for data and analytics talent. The first task was simply to identify all the data and analytics talent. Through a “snowball sampling” approach, the enterprise talent initiative found almost 100 different teams comprising about 2,000 people (out of over 80,000 employees overall). Another key component involved the creation of standard job families and classifications across the bank. Seven different job families were identified, including:

- Advanced Analytics
- Business Application Ownership/Management
- Business Information Management
- Business Insights and Analytics
- Business Intelligence Reporting
- Data Governance and Data Management
- Visualization

Within each of the families, specific jobs were defined in detail—in all, 65 different roles. For each job, several attributes were described, including the primary purpose of the role, the numerical level within the bank’s HR system, key accountabilities (to internal customers or business partners, shareholders, and other bank employees), the breadth and depth of the role, and the experience and education required to perform it. Individual contributor attributes were identified as well as those for people management (e.g., coaching and staff development).

Sixteen different competencies were also identified across the job classifications, with competency assessments and a self-assessment process. The roughly 2,000 people in data and analytics jobs within the bank were then mapped to the job families and specific jobs. For the first time the bank was able to understand what skills its teams had and lacked, and how to combine individuals into teams with the skills to complete data science projects. While such a classification requires considerable time and effort—and close collaboration with the organization’s HR function—it is necessary in order to accurately assess, and take effective action on, data science human resources. Since the creation of the job role and skill structures, the bank has become significantly more focused on data science, and has significantly increased its ranking among the most desirable employers of data scientists in North America.

4. The Value of Enterprise Job Role and Skill Structures

With an enterprise classification structure for data science jobs like the one that the bank created, an organization can ensure that all the needed skills are present on a data science team, and can provide different capabilities for different types of data science projects. It allows organizations to assess data science teams to determine if they have the needed diversity of skills and backgrounds. Diversity on data science teams is critical, as Rob Casper, the chief data officer of JPMorgan Chase, put it in an interview with *McKinsey*:

If you have a team that's very similar in nature, you're not going to get that necessary healthy tension. You want somebody who's strong with technology. You want somebody who's strong with business process. You want somebody who's strong with risk and regulatory. You want people who can communicate effectively, both in writing and verbally. If you have that, then you have the healthy tension that makes for a good team. (Díaz et al., 2018)

Firms with a classification structure can also make educational offerings available to fill needed skill gaps. If a firm learns, for example, that it doesn't have enough skills to enable the broad creation and management of Hadoop-based data lakes, it can invest in educational programs for its data scientists.

There may also be different skill requirements at different stages of a data science project. Early stages, for example, are more likely to involve problem-framing skills; later stages involve coding and data management. A clear classification structure allows firms to tailor the composition of teams to the stage of the project. Methodologies for data science projects, such as Microsoft's Team Data Science Process (Microsoft, 2020) can work well in conjunction with a skill and job classification model.

Because data science skills are scarce, they should be carefully allocated to projects and tasks. A classification model makes it more likely that the needed skills will be available on teams when and where they are most needed. Data scientists who are excellent at algorithm generation, for example—and less skilled on other tasks—won't have to spend a lot of time on understanding and redesigning the business process into which the algorithm will fit—and vice versa.

One alternative to the top-down enterprise structure approach is to create a bottom-up analysis of data science job roles and skillsets across a variety of employers. This approach was taken by one set of researchers, who created from an analysis of online job postings a set of four job families, nine groups of what they refer to as “Big Data skills,” and a mapping of job families to the level of competence on each skillset (De Mauro et al., 2018). The authors argue that the analysis is replicable and could be used by organizations to create their own typologies.

5. Classifying and Certifying Data Science Skills

Classification and certification structures for jobs and skills are valuable within individual firms, but would be even more desirable at an interorganizational, societal level. If there were widely employed standards about what constituted different types and levels of data scientists, companies would be able to hire one with confidence about what capabilities they are getting in that person. They would be able to ensure that their data science teams had the complement of skills needed to be successful on projects.

Many other professions—doctors, attorneys, engineers, and so on—have well-defined classification and certification approaches, and their fields have gained trust and influence as a result. Of course, classification and certification structures could have negative effects as well. They might, for example, exclude talented but less-credentialed individuals from entering the data science field. They might also ‘lock in’ a set of ideas about what constitutes effective data science from a group powerful enough to create and institutionalize them. However, I believe that the current low barriers to entry into the data science field, and the confusion about the skills and job roles involved in the field, make it more desirable to move in the direction of greater classification and certification.

Today, no such societywide classification and certification approach for data scientists exists. However, there are efforts underway to create one, and there are certification programs in the related domain of analytics. The Initiative for Analytics and Data Science Standards (IADSS) is a recently created body formed to try to create a broad set of standards for data science qualifications (Fayyad & Hamutchu, 2020). Its website describes the current situation:

almost every company in the industry has a unique way of defining roles and assigning titles in data analytics related positions which have resulted in a chaotic market that is confusing to employers, academic and training institutions, and candidates; with a large number of unqualified candidates calling themselves “data scientist,” “data architect,” “data engineer” or “analytics professional.” (Initiative for Analytics and Data Science Standards [IADSS], 2019)

IADSS is conducting a research study to learn what leaders and practitioners in the profession think about needed skills and job standards. It is also conducting workshops at prominent data science events. However, it may be some time before standards are agreed upon and certainly before they are widely adopted.

In the analytics field, INFORMS, the professional association for operations researchers, has developed a certification for the Certified Analytics Professional, or CAP (Nestler et al., 2012). The test consists of an online knowledge test of different phases of analytics projects, as well as a certification from employers or consulting clients of ‘soft skills.’ There are about 600 CAPs thus far; the slow growth of program certifications (it was established in 2013) attests to the difficulty of establishing any standard and promulgating it throughout a profession.

In addition to the INFORMS certification, there is a variety of certification programs (Olavsrud, 2020) offered by particular universities or vendors. None of these have the breadth of acceptance of the CAP program, and vendor independence would seem to be a positive certification attribute. These programs may well be useful, but they fall short of a societywide certification approach.

Despite widespread agreement that data science unicorns don't exist, and a consensus that teams are necessary with members possessing multiple backgrounds and skillsets, the world isn't currently constructed to form such teams easily. We typically have only a data scientist's own word for it that he or she possesses a certain type and level of particular skills. Some potential employers of data scientists actually require job candidates to demonstrate coding abilities during job interviews, although it would be more difficult to assess softer skills like consulting or relationship-building in this fashion.

There is hope for eventual society-level classifications and certifications, but until then each organization that wants to employ data scientists will need to develop its own. If your organization has not yet developed an enterprise classification and certification structure like the bank's, it can at least provide greater detail on the specific skills and job activities involved in a particular job role. A perusal of job boards can provide a large number of examples. Eventually, however—even after society-level approaches become available—it will be important for organizations that focus heavily on data science capabilities to create job families, roles, and required skills that help to advance their particular strategies and objectives.

Disclosure Statement

Thomas Davenport has performed paid consulting and speaking work for the North American bank described in the article. He is an unpaid volunteer advisor to the INFORMS CAP program and the IADSS.

References

- Abbasi, A., Kitchens, B., & Ahmad, F. (2019, October 24). The risks of autoML and how to avoid them. *Harvard Business Review*. <https://hbr.org/2019/10/the-risks-of-automl-and-how-to-avoid-them>
- Berthold, M. (2019). What does it take to be a successful data scientist? *Harvard Data Science Review* 1(2). <https://doi.org/10.1162/99608f92.e0eaabfc>
- Davenport, T. (2014, May 5). 10 kinds of stories to tell with data. *Harvard Business Review*. <https://hbr.org/2014/05/10-kinds-of-stories-to-tell-with-data>
- Davenport, T. (2018a, December 31). Analytics in business education: analyzing the future. *BizEd*. <https://bized.aacsb.edu/articles/2019/january/analyzing-the-future>

- Davenport, T. (2018b). The analytics team. In *INFORMS Analytics Body of Knowledge* (pp. 49–76). Wiley. <https://doi.org/10.1002/9781119505914.ch3>
- Davenport, T., & Patil, D. (2012, October). Data scientist: Sexiest job of the 21st century. *Harvard Business Review*. <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>
- De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). Human resources for big data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management*, 54(5), 807–817. <https://doi.org/10.1016/j.ipm.2017.05.004>
- Díaz, A., Rowshankish, K., & Saleh, T. (2018, September). Data culture: Where organization meets analytics. *McKinsey Quarterly*. <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/why-data-culture-matters>
- Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(7), 745–766. <https://doi.org/10.1080/10618600.2017.1384734>
- Doyle, A. (2019, June 3). Important job skills for data scientists. *The Balance Careers*. <https://www.thebalancecareers.com/list-of-data-scientist-skills-2062381>
- Fayyad, U., & Hamutchu, H. (in press). Analytics and data science standardization and assessment framework. *Harvard Data Science Review*.
- Friedlein, A. (2012, November 8). Why modern marketers need to be pi-people. *Marketing Week*. <https://www.marketingweek.com/why-modern-marketers-need-to-be-pi-people/>
- Garber, A. (2019). Data science: What the educated citizen needs to know. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.88ba42cb>
- Gelman, A. (2013, November 14). Statistics is the least important part of data science. *Statistical Modeling, Causal Inference, and Data Science blog*. <https://statmodeling.stat.columbia.edu/2013/11/14/statistics-least-important-part-data-science/>
- Gutierrez, D. (2019, April 25). *Data scientists versus statisticians*. Open Data Science. <https://opendatascience.com/data-scientists-versus-statisticians/>
- Initiative for Analytics and Data Science Standards. (2019). *About us: research governance and steering parties*. <https://www.iadss.org/about-us>
- Irizarry, R. (2020, January 31). The role of academia in data science education. *Harvard Data Science Review*, 2(1). <https://doi.org/10.1162/99608f92.dd363929>

- Loucks, J., Davenport, T., & Schatsky, D. (2018). *State of AI in the enterprise* (2nd ed.). Deloitte Insights. <https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/state-of-ai-and-intelligent-automation-in-business-survey.html>
- Malone, K. (2020, January 31). When translation problems arise between data scientists and stakeholders, revisit your metrics. *Harvard Data Science Review*, 2(1). <https://doi.org/10.1162/99608f92.c2fc310d>
- Microsoft. (2020, January 10). *What is the team data science process?* Microsoft. <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview>
- Muenchen, R. (2019). *The popularity of data science software*. R4stats. <http://r4stats.com/articles/popularity/>
- Nestler, S., Levis, J., & Klimack, B. (2012, October). *Certified analytics professional*. INFORMS. <https://www.informs.org/ORMS-Today/Public-Articles/October-Volume-39-Number-5/Certified-Analytics-Professional>
- Nonaka, I. (1990, April). Redundant, overlapping organization: A Japanese approach to managing the innovation process. *California Management Review*, 32(3), 27–38. <https://doi.org/10.2307/41166615>
- Olavsrud, T. (2020, Jan. 14). The top 9 big data and analytics certifications for 2020. *CIO*. <https://www.cio.com/article/230388/big-data-certifications-that-will-pay-off.html>
- Roberts, P., & Roberts, G. (2013, July). *Research brief: Four functional clusters of analytics professionals*. Data Science Central. <https://www.datasciencecentral.com/profiles/blogs/research-brief-four-functional-clusters-of-analytics>
- Vaisman, M., Harris, H., & Murphy, S. (2013, June). *Analyzing the analyzers: An introspective survey of data scientists and their work*. O'Reilly Media.
- Zhang, V. (2019, August 9). Stop searching for that data scientist unicorn. *InfoWorld*. <https://www.infoworld.com/article/3429185/stop-searching-for-that-data-science-unicorn.html>

©2020 Thomas H. Davenport. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.