Harvard Data Science Review • Issue 2.2, Spring 2020

Toward Foundations for Data Science and Analytics: A Knowledge Framework for Professional Standards

Usama Fayyad[1], Hamit Hamutcu[2]

[1]Applied Machine Learning Research Group, Institute for Experiential Artificial Intelligence, Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, United States of America; International Secure Systems Lab, Northeastern University, Boston, Massachusetts, United States of America,

[2]Initiative for Analytics and Data Science Standards, Institute for Experiential Artificial Intelligence, Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, United States of America

Published on: Jun 30, 2020

DOI: 10.1162/99608f92.1a99e67a

ABSTRACT

As the industry is racing to harness the power of data, demand for data science professionals is growing at an increasing rate. However, almost every organization has a unique way of defining roles in data science and associated skills and knowledge. This has resulted in a confusing industry landscape for employers, academic and training institutions, and existing and aspiring data science professionals.

This article is the first in a series authored by Initiative for Analytics and Data Science Standards (IADSS). We review the history of data science, which we trace back to 1974, and the emergence of data science as a profession in the industry, followed by a classification of knowledge and skills commonly associated with data science professionals, pointing to a lack of detailed and consistent treatment of the topic. We then present a Data Science Knowledge Framework, that we believe can support industry standardization and building measurement and assessment methodologies for data science professionals.

Keywords: data science roles, skills and knowledge, hiring and assessment, industry standards, data analysis, data science, analytics

---

1. Introduction

What do you think of when you hear the term ‘data science?’ More importantly, who are data scientists, data engineers, and analytics professionals? We have observed over the last decade a surge of interest in all things ‘data.’ Data collection, processing, and interpretation have evolved over the years to become increasingly sophisticated. These developments have helped transform industries and led to the inception of new business models that focus on creating data-driven products, features, and services.

Leading this transformation is a set of professional disciplines founded upon the principles of applied statistics, management science, and computer science, among several other fields. These disciplines help unlock the value of data in a spectrum of organizations, businesses, scientific research, NGOs, and governments worldwide. The data could be in stored and managed in a variety of formats: from relational databases to NoSQL databases to massive data stores, file systems, federated or totally fragmented stores, or BigData platforms. The data could be structured or unstructured. Collectively, these professional disciplines are referred to as ‘analytics’

and ‘data science.’ The demand for analytics and data science skills parallels the growth of interest and investment in data science. Some estimates predict the future demand for professionals with analytics and data science skills to exceed two to three million in the United States alone. As predicted by (Markow et al., 2017), there are now more than 250 data science programs just in graduate schools (North Carolina State University, 2019).

However, the explosive growth surrounding data science has left in its wake a state of confusion regarding the basic definitions of related tools, methods, skills, and roles associated with this discipline. The definition of data science itself is shrouded in uncertainty. Some data scientists are machine learning and algorithm experts, while others specialize in developing and maintaining data infrastructure. Yet another faction of data scientists comprises business-facing data strategists. Naturally, the educational backgrounds, skills, abilities, and professional relevance of these individuals also vary greatly by project as well as position.

The confusion stems not only from ‘which’ skills an analytics or data science professional must possess; it also extends to determining ‘what’ level of skill and knowledge is required for a professional to qualify for a particular title. Should practitioners that refer to themselves as data scientists hold advanced degrees? Or would several online courses suffice? Who is a ‘Senior Scientist’—a designation that is reserved for experienced scholars in every other scientific discipline?

Although ‘data scientist’ has emerged as a job title, every industry, function, and business appear to be looking for their definition of the role. This confusion leads to substantial costs for employers and professionals alike. Employers must build an understanding regarding the vague taxonomy of different data scientists to find the right talents for their business needs. They must then evaluate these professionals for whether they fit their requirements and appraise their contributions before deciding to retain them. Employees, on the other hand, must navigate through a sea of data science positions, only to find out that they lack a significant area of knowledge required for most of the roles advertised. Refer to Figure 1 for a list of titles that require skills in the analytics and data science spheres, and there are many other emerging titles such as ‘data translator,’ which make this an ever-growing list.

Data Scientist – Generalist	Business Analyst	Database Architect

Data Scientist - Machine Learning	Data and Analytics Manager	Business Intelligence Analyst
Data Scientist - Artificial Intelligence	Marketing Data analyst	Quantitative Analyst
Data Analyst	Reports and Data analyst	Computer and Information Research Scientist
Data Architect	Operations Analytics Manager	Data Expert
Data Engineer	Data and Analytics Solution Lead	Data Research Scientist
AI Engineer	Big Data Developer/Engineer	Chief Data Officers
Machine Learning Engineer	Database Administrator	Chief Analytics Officer

Figure 1. A quick overview of data science jobs and role. Adapted from Datacamp (Willems, 2015); Udacity (Nelson, 2018); ONET; and Linkedin.

Universities have responded to the demand for data scientists by creating schools, institutes, and centers and establishing degree programs for relevant disciplines. These suffer from the same confusion—some are housed in business schools and others are established within computer science departments, some are cross-disciplinary, and others are considered specializations of more established disciplines. Some institutes hire a dedicated faculty and admit their own students; others provide joint appointments to existing faculty members of traditional departments (Gorman & Klimberg, 2014). To add to the confusion, some institutes grant professional degrees, while others merely act as a research hub.

Other professions have addressed these issues by agreeing upon a common framework, a body of knowledge, and a generally accepted standard of qualifications (Bourque & Fairley, 2014; Project Management Institute, 2017). Data science has made limited progress in this regard, with several impediments standing in the way. For instance, a lack of common vocabulary in data science hampers the design and implementation of such standards and related measurement instruments to assess the abilities and competencies of data science professionals.

1.1. Motivation for This Article

We believe there is an urgent need in the data science ecosystem to address the confusion discussed. One big concern (mentioned previously) is the significant cost of this confusion to employers in terms of interviewing or hiring the wrong candidates or for aspiring data scientists having to deal with lack of clarity on requirements for knowledge and skills in job descriptions.

A second area of concern is that the rapid increase in market demand drove the creation of not just academic training programs, but a plethora of boot camps, 'meetups,' and many alternative means for 'self-training.' As an example, in the Silicon Valley area there are several meetups for data science ranging in sizes from 3,000 to 14,000 members, bigger than many main national and international conferences on this subject. Many academic institutions are not part of these ubiquitous activities and our concern is that a large population of practitioners and would-be 'future data scientists' are being trained outside of an academic environment, in ways that can compromise quality and consistency of the training.

A failure to reach agreement on a body of knowledge and standards for defining what we need in a data science professional runs the danger of causing negative impact on the entire ecosystem, with organizations making significant investments in data science resources and not achieving the desired results. This would lead employers to declare the whole area of data science as problematic and risk the health and reputation of the field itself.

As analytics and data science are in their formative period, it is natural for practice in the industry applications and context to drive the early stage of developments in the field, especially with the exponential growth of available data and corresponding demand for talent. We are confident that if the growth is managed well, data science will blossom into a rich space where academia will equip the future workforce with the right skills, employers will better understand how to define roles and assess candidates, and candidates will have higher quality training and well-understood expectations of the roles they are applying for. We hope that by creating a framework for skills and defining the initial knowledge framework, we can help advance the maturity of data science in practice and avoid some of the potentially harmful side effects of confusion. An effective characterization of the unreasonable expectations of employers in what they envision as a data scientist is provided in Davenport (2020), where he outlines that the problem of finding 'unicorns' is not just difficult, but

rather impossible. This leads to further disenchantment and unnecessary disappointment in what otherwise could be a great role.

Finally, we believe the current pandemic environment only compounds the urgency of this topic. As data scientists try to do their part in understanding, explaining, visualizing, and predicting the impact of COVID-19, the importance of data acumen and data-driven policy is even more evident. Moreover, the pandemic accelerated digital transformation as people and organizations attempt to achieve more over electronic channels, and as digital interactions replace human ones. This in turn generates more demand for insights from data to understand and act on what is working and what is not and, we believe, ultimately more demand for data science professionals.

1.2. Overview of This Article

This article is the first in a series authored by the Initiative for Analytics and Data Science Standards[1](IADSS) to present a framework for the knowledge and skills required in data science and support building a measurement and assessment methodology for analytics and data science professionals. Launched to develop industry standards for data science professionals, through a collaborative process with businesses, academia, and the data science community, IADSS is an independent initiative, aiming to bring together industry professionals and academics to conduct a global-scale research study on the professions associated with analytics and data science. Authors of this article are also co-founders of this initiative.

Through this article, we desire to advance the discussion on how data science skills should be defined and classified. We expect the contents of this article to be refined and extended through a series of expert interviews, workshops, and quantitative research conducted by IADSS. While in this article we do not try to define the field – as this is still too early in our opinion – we do review prior attempts at defining it and we provide a working definition to help us frame our effort to define the core body of relevant knowledge. We also made the practical choice of using ‘analytics and data science’ in defining our scope, to make sure that we cover the entire professional field regardless of existing organizational designations or job titles. This is driven by the fact that in many organizations, activities typically associated with data science are carried out by advanced analytics groups or people considering themselves in the analytics space.

The contents of this article provide a review of existing literature on analytics and data science skills. We have collected this information to enable its implementation under a common framework. At the onset of this article, we have reviewed the development and emergence of data science, followed by the various definitions and viewpoints associated with it. We then proceed to present a discussion on the term data science and propose a working definition. Beyond this, we have reviewed the literature available for describing the knowledge and skills that are typically attributed to data science professionals. We have collected the outcomes of this review and inserted the inputs provided by experts to create a framework that proposes a formal classification of data science disciplines. If integrated successfully, this framework, developed from an industry perspective, can prove to be an important milestone in standardizing data science and can also benefit similar efforts in academia, governments, and NGOs.

2. A Brief History of Data Science

Numerous works have attempted to define data science. Discussions around the topic abound, in scholarly articles, blog posts, and books. To our knowledge, the earliest mention of "the science of the use of data" dates to the 1960s (Naur, 1966). Irizarry (2020) points out that although the term data scientist is claimed to be coined by industry, the term was even suggested to rename the statistics discipline by Jeff Wu in 1997 (Irizarry, 2020). In 2015, the American Statistical Association published a "Statement on the Role of Statistics in Data Science" (Van Dyk et al., 2015), which we quote directly from:

> 1. There is not yet a consensus on what constitutes data science
>
> 2. Three professional communities, all within computer science and/or statistics, are emerging as foundational to data science:
>
>   a. *Database Management* enables transformation, conglomeration, and organization of data resources,
>
>   b. *Statistics and Machine Learning* convert data into knowledge, and
>
>   c. Distributed and Parallel Systems provide the computational infrastructure to carry out data analysis.

Peter Naur described data science rather narrowly and wrote in 1974, "The science of dealing with data, once they have been established, while the relation of the data to

what they represent is delegated to other fields and sciences" (Naur, 1974). Until the term entered the mainstream in the late 2000s, it made several other appearances in the statistics community, referred to as a proposed 'paradigm shift' in the field (Cleveland, 2001; Hayashi, 1998; National Science Foundation, 2005; Wu, 1997). As Donoho (2017) later observed, these attempts aimed to address a set of challenges "intellectual, rather than commercial." This entailed that applied statisticians had to increasingly adopt the approach of "data analysis," as Tukey (1962) had outlined. However, in this use of data science, computational challenges and business needs were secondary to 'intellectual' challenges. Meanwhile, many tasks and roles attributed to modern data scientists were associated with the data mining and knowledge discovery in databases (KDD) communities (Fayyad et al., 1996).

Some sources, such as National Institute of Standards and Society (2015), trace the roots of the term to a paradigm called "data-intensive science." Gray (2007) defines "data-intensive" science as the fourth paradigm in scientific discovery—following empirical, theoretical, and computational science, which is characterized by its application of the scientific process directly on the data. Grady and Chang (2015) based their standard definition of data science on this view, as the "extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing" (Grady and Change, 2015, p. 7).

The problem of lacking a clear definition of data science is exacerbated by the fact that the term seems to have emerged in job titles and descriptions from the tech industry, independently of its previous uses as a statistical science or as a new scientific paradigm. Coined by D. J. Patil of LinkedIn and Jeff Hammerbacher of Facebook, the data scientist title was used to encapsulate a set of activities in dealing with internet scale data (Davenport & Patil, 2012).

Before describing what makes a good data scientist, Patil (2011) offers insight into the development of the term. He highlights that 'data analyst' was perceived to be a limiting description of the first data science roles at LinkedIn. The title 'research scientist' seemed fitting for the skills required; however, its implied working on projects too "futuristic and abstract" as compared to projects with immediate impact on which the team worked.

Similarly, Hammerbacher (2009) writes that traditional titles such as "Statistician and Business Analyst" did not accurately capture what the team at Facebook did. What defined the newly christened role was that it combined a wide array of activities from engineering data platforms to designing and running rapid quantitative analyses and

deploying models and communicating findings. In both cases, data scientists often had graduate-level training and most of them held doctorates in natural sciences.

In their well-known *Harvard Business Review* article, Davenport and Patil (2012) describe the term data scientist as the sexiest job of the 21st century. They observe that data scientists bring a unique blend of skills, for "subduing" large amounts of data, applying quantitative analysis, and interfacing with business stakeholders. In this regard, they highlight how data scientists differ from the previous generation of quants, analytics professionals, and business analysts. This tells us that computational challenge and engineering skills that data scientists carry are essential to their use of the term. We can contrast this with the recent article by Davenport (2020), as discussed previously, in terms of seeking a "unicorn," which we believe captures the current market environment more accurately in our experience.

Provost and Fawcett (2013) present separate discussions of data science—they describe it as a field and as a profession. They characterize data science as the interface between data engineering and processing technologies and "data-driven decision making." As a result, they have laid down a set of shared fundamental principles in the practice of data science.

The data science "moniker," according to Blei and Smyth (2017), refers to a child discipline of computer science and statistics. Data science blends these two disciplines and "refocuses" them to address the needs of modern data analysis. In this sense, data science reflects the perspective of Tukey's "data analysis" (1962).

Donoho (2017) debates the discipline's inherent connection to 'big data' and argues that the dependence on any specific technology is "transitory." Citing Breiman's (2001) famous essay on "the two cultures" in statistical modeling, Donoho posits that a defining property of modern "consensus data science" is its focus on predictive modeling rather than inference. Finally, he points to a broad definition of data science that is "the science of learning from data," with all technologies, computational aspects, and human aspects that it comprises. He gives a framework for "greater data science," which we return to in the following. The question of a precise definition for data science was also raised by Cao (2017), Dhar (2013), Song and Zhu (2016), Wing (2019), and Van der Aalst (2014), each with differing viewpoints on the responsibilities, skills, tasks, and activities.

3. Defining Data Science

While our review clearly establishes that there is no commonly agreed on definition for data science, it also points out an additional finding: there is no consensus on what it is. Is it a new scientific paradigm? a new field of scientific study, or a practice—a trade—that appeared in Silicon Valley? Moving toward a working definition of data science, we provide a brief discussion based on some of the themes that were encountered in our review. It is important to clarify that the focus of this article is not on the definition but rather on establishing better understanding of the body of knowledge underpinning this growing area of activity.

There are several paths that one can use to define data science. For instance, some may suggest that the qualifications of a data scientist can form the basis of a working definition. However, we believe that such a definition would be insufficient, as qualifications tend to vary with time and context. For example, although a doctor's qualifications are largely agreed upon today, they have varied greatly throughout history, while the scientific study concerned with human health has been called 'medicine.' Another approach is to look at the tasks and activities of a data scientist. Indeed, Harris (2011) defined the discipline as "data science is what data scientists do." However, as Provost and Fawcett (2013, p. 14) point out, if we start identifying a profession with its tools, then we would expose ourselves to the risk of defining "chemists with test tubes." We, instead, start from the need—the challenge that gave rise to the practice. In the work of Gray (2007), for example, the concern is to extend the scientific method to meet the challenges of the day, and think of data exploration as a scientific paradigm. However, it appears that much of today's data science is rooted deep in the industry with the sole purpose of connecting the ever-growing data stores, of all shapes and sizes, to solving problems and decision making. This definition is also in line with the phenomenon we highlighted at the beginning of this article and is similar to the broader professional discipline that has also been referred to as 'analytics.'

The transdisciplinary aspect of data science is undisputed. All sources agree that this field incorporates principles, techniques, and methods from multiple existing academic disciplines and application domains. There is also a strong consensus that the methods employed by data science professionals, in business and research institutions alike, are a research area of their own. Indeed, there is a need for "science about data science" (Donoho, 2017, p. 758) that focuses on the pragmatic, exploratory, and highly tailored

methods of data analysis that are carried out and likely lie outside the purview of today's academic statistics departments.

Another emerging theme is the divide highlighted in Breiman's (2001) essay. He did not refer to the term data scientist but was concerned with the two different cultures in statistics. While academic statistics have mostly studied inference in well-specified statistical models, data scientists appear to be trapped in a world of ill-posed questions, wicked problems, and business challenges that call for pragmatic application of quantitative methods for predictive solutions.[2]

Finally, all sources agree that the need for computing with data, under varied computational constraints and in the presence of highly disparate, unstructured data, is a defining tenet of data science practice. Although tools, frameworks, and specific computational challenges may be transient, this fundamental computational challenge is expected to last. For the purposes of this article, we (IADSS) establish a *working definition* to guide the rest of the discussion. This working definition builds on a broad goal: connecting data to achieving goals. We take the practice of data science in industry as our starting point. We believe this goal is shared across data science's application domains, including its use in scientific discovery. Our working definition of data science is:

> ***"Using Data to achieve specified goals by designing or applying computational methods for inference or prediction."***

The terms are further defined as:

> – ***"Achieving Specified Goals":*** Depending on the domain and context, can mean exploration, discovery, decision-making, prediction, optimization, or similar objectives and tasks. This is very much related to the scientific method for building knowledge from observations.
>
> – ***"Designing or Applying":*** Means that activities such as designing, understanding, or examining inference methods (e.g., the study of learning from data in machine learning [ML]) or applying methods in a particular problem context (e.g., using statistical analysis or inference methods) are included as data science activities.
>
> – ***"Computational Methods":*** Refers to using computers either to directly conduct a search or to aid a human in formulating or optimizing a model; for example, the search can be for a model, a structure, or an explanation and can be

> achieved through a variety of techniques from many fields, including analytical, statistical, or machine learning tools and techniques.
>
> – ***"Inference or Prediction***": Inference can be conducted algorithmically, and this includes automated formulation of hypotheses, automated exploration of definitions of new attributes or representations, and so on. Prediction includes the cases where the goal is just to produce an optimized predictive model without necessarily gaining insights into how it works; for example, in deep learning when the goal is to achieve a certain level of performance on the predictive model output.
>
> – ***"Data":*** Can be structured or unstructured. *Achieving goals* or getting *computational inference* methods to work on data often require related tasks such as data cleaning or transformation.

In this broad view, data science implies a set of activities that, in today's view, can be considered transdisciplinary. Yet, we imply that data scientists deliver, to some degree, against three main challenges, that is, their daily goals are a combination of tackling computational, scientific, and organizational goals. This combination of goals implies a combination of skills and knowledge traditionally found in different professions. In the next section, we take a closer look at these knowledge areas and workplace skills associated with data science.

4. A Review of Data Science Knowledge and Skills

Many studies before ours have given a discussion of the knowledge, skills, and abilities associated with data science. While some studies highlight a mapping between academic fields of study and the knowledge required in data science practice, others have sought to build a set of required skills directly from job descriptions or reported tasks and activities of practitioners.

We first differentiate between two important terms. By knowledge, we will refer to an 'organized body of information usually applied directly to the performance of a function,' (Centers for Disease Control and Prevention, n.d.) whereas by skill, we will refer to "proficient manual, verbal or mental manipulation of data or things" (Office of Personnel Management [OPM], 2007, 2018; Organisation for Economic Co-operation and Development, 2018). In other words, knowledge refers to information often acquired through formal education, books, or other media. Skills, on the other hand, refer to the ability to apply this knowledge, often gained through practice. Many of the sources cited in this section focus on 'knowledge' connected to data science despite

the varying use of the word 'skill.'" We do this not merely to point out how they are interchangeably used in our literature review but also to serve as a basis for any assessment or measurement efforts to follow.

Take Drew Conway's (2010) popular Venn diagram and the skills diagram from Grady and Chang (2015) as an example. Figure 2 shows both diagrams provide a high-level view of the data science knowledge landscape. They identify content related to statistics and engineering as two of the three main pillars of data science. Both point at the third pillar as the domain expertise. This is known as the knowledge of the subject matter to which data science is applied. This broad classification is echoed in many of the works cited in the following.

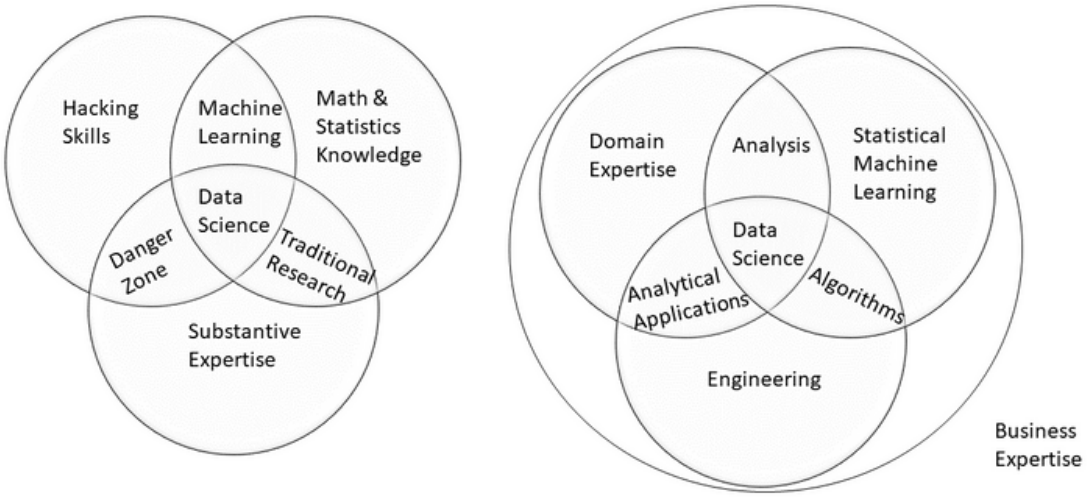

Figure 2. Data science Venn diagrams.

One group of studies for defining the knowledge areas of data science rely on job posts, expert opinions, or self-descriptions of employees. Harris et al. (2012) based their analysis on data collected from 250 data scientists via a web-based questionnaire. They defined the subfields of data science, yielding five main areas: Business, Machine Learning/Big Data, Mathematics/Operations Research, Programming, and Statistics.

Dubey and Gunasekaran (2015) identify two clusters, namely, "hard" and "soft" skills, based on quantitative research conducted with data and business analytics executives. Their list of hard skills, as shown in Figure 3.a, maps to our definition of knowledge.

Their classification also adds business domain expertise, Marketing, and Finance. Like Harris et al. (2012), they include topics of Operations Research, such as Optimization and Multiple Criteria Decision Making, into their research. They list Research Methods as required knowledge, echoing the definition of data science as a scientific paradigm.

Schoenherr and Speier-Pero (2015) provide an assessment of the current state of the field based on quantitative research, which was conducted with more than 500 supply chain management professionals. They group skills for data science as (1) enterprise business processes and decision making, (2) analytical and modeling tools, and (3) data management. De Mauro et al. (2018) identify eight big data skillsets based on an analysis of online job posts, as given in Figure 3.b. They list Systems Management, Distributed Computing, Database Management, and Cloud under the programming and engineering aspect of data science. Radilovsky et al. (2018) use text mining on job listings for data scientist and business analytics positions in the market. Their four-component knowledge model comprises (1) technical, (2) analytical, (3) business, and (4) communication as the main areas of skill required in analytics and data science.

Another group of studies review undergraduate and graduate degree requirements in analytics and data science, as well as the curricula of certification programs offered by academic institutions. Gorman and Klimberg (2014) survey some of the most established programs and interview their representatives. Their study aims to understand admission requirements, the required and elective course topics covered, and job opportunities for graduates. Their scope covers the union of three intersecting fields: (1) *information systems for business intelligence*, (2) *quantitative methods*, and (3) *statistics*. This classification is further broken down into 17 subfields as listed in Figure 3.e.

Mills et al. (2016) review information systems programs, which focus on content areas such as analytics, business intelligence, and big data. They adopt four pillars of skills: data preprocessing and retrieval (NoSQL, data warehousing), data exploration (statistical analysis and visualization), analytical models (machine learning, data mining, natural language processing), and data product, the final pillar, which encompasses data organization and application development.

In their research on undergraduate degree programs, Cegielski and Jones-Farmer (2016) identify three domains where skills and abilities are classified. Their conclusions combine an examination of other programs with experience in building academic programs in business analytics at two nationally ranked state universities. Their knowledge domains start with the business domain, which includes

communication skills, project management, client orientation, and time management. The analytical domain incorporates problem definition and problem-solving skills, predictive analysis, and integrative analysis. The technical domain has three subgroups with (1) applications like Excel, SAS, and Tableau; (2) programming languages like R, Python, SQL, and Java; and (3) infrastructure such as Hadoop, Cassandra, Oracle, Linux, MapReduce, Hive, and Pig.

Figure 3.a Business Analytics and Big Data Hard Skills according to Dubey and Gunasekaran (2015)	Figure 3.b Big Data Skills according to De Mauro et al. (2018)	Figure 3.c Data Science activities according to Donoho (2017)
• Forecasting • Optimization • Quantitative Finance • Financial Accounting • Multivariate Statistics • Multiple Criteria Decision Making • Marketing • Research Methods • Finance	• Business impact • Project Management • DB Management • Analytics • Coding • Systems Management • Distributed Computing • Cloud	• Data Exploration and Preparation • Data Representation and Transformation • Computing with Data • Data Modeling • Data Visualization and Presentation • Science about Data Science

Figure 3.d
Data Science activities according to O'Neil and Shutt (2013)

- Extract meaning from and interpret data which requires using tools and methods from statistics and machine learning
- Spends a lot of time in the process of collecting, cleaning, and munging data, because data is never clean. This process requires persistence, statistics, and software engineering skills.
- Exploratory data analysis, which combines visualization and data sense
- Find patterns, build models, and algorithms
- Design experiments
- Be part of data-driven decision making
- Communicate with team members, engineers, and leadership in clear language and with data visualization

Figure 3.e
Subfields according to Gorman and Klimberg (2014)

- Big Data
- Enterprise Systems
- Data Marts
- Databases
- Spreadsheets
- OLAP
- Programming
- Optimization
- Simulation
- Supply Chain
- Six Sigma
- Modelling
- Data Mining
- Forecasting
- Basic Statistics
- Multivariate Statistics
- Econometrics

Figure 3. Review of knowledge and skills in data science.

In another survey-based study, Song and Zhu (2016) provide a framework for data science based on current data science education in the United States. The four components named under the study include *big data infrastructure*, *big data analytics lifecycle*, *data management skills*, and *behavioral disciplines*. The first component corresponds to technologies such as Hadoop, NoSQL, in-memory and cloud computing, while the *analytics lifecycle* covers all stages of data analysis, understanding, preparation, integration, model building, and evaluation. *Data management skills* include data modeling and relational database knowledge. Finally, *behavioral disciplines* include soft skills, critical thinking, and communicating with domain experts.

An ambitious study conducted by International Data Science in Schools Project (2019) provides granular detail on topics to be covered by secondary schools in introducing data science. The study groups a wide range of knowledge areas under two main units, one covering foundations such as basic exploration techniques and graphical displays and the other specialized sub-areas such as machine learning and interactive visualization. The National Academies of Sciences Engineering Medicine (2018) define

four areas of focus as Foundational, Translational, Ethical, and Professional skills in their data science undergraduate study–focused interim report. In another study by De Veaux et al. (2017), key competencies for an undergraduate data science major are listed as Computational and Statistical Thinking, Mathematical Foundations, Model Building and Assessment, Algorithms and Software Foundation, Data Curation, and, finally, Knowledge Transference—communication and responsibility.

Donoho (2017) classifies topics of “greater data science” (GDS) in six categories, given in Figure 3.c. GDS includes advanced knowledge of exploratory data analysis, data visualization and data modeling. Both aspects of statistical modeling—'predictive’ and ‘generative’—are included in his framework.

A third group of studies approaches the discussion of knowledge, skills, and abilities associated with data science from the activity and task angle. That is, they identify the required knowledge by associating it with what data scientists *do*. Bowne-Anderson (2018) bases his findings on a study conducted with 35 data scientists who focus on their daily tasks and activities. The study analyzes activities involving *infrastructure*, *testing*, *machine learning for decision making*, and *data products*. It then lists the main activities in a typical workday as (1) data collection and cleaning, (2) building dashboards and reports, (3) data visualization, (4) statistical inference, (5) communicating results to key stakeholders, and (6) convincing decision makers of their results.

O’Neil and Shutt (2013) summarize the activities of data scientists and link some of the relevant knowledge and skills. Their list of activities is given in Figure 3.d. The list includes familiar activities such as data cleaning, but it also explicitly mentions exploratory analysis and design of experiments.

The data lifecycle by Grady and Chang (2015) is a four-step process covering the tasks of a data scientist. It includes *collection, preparation, analysis,* and *action.* Similarly, the Cross-Industry Standard Process for Data Mining (CRISP-DM) sheds light on data science tasks (Wirth & Hipp, 2000). The process starts with *business understanding* and *data understanding*, followed by *data preparation, modeling, evaluation,* and *deployment.* From the task and activity point of view, a general list of activities associated with analytics and data science roles can be summarized as in Figure 4.

- Business Understanding
- Data Exploration and Preparation
- Data Representation and Transformation
- Computing with Data
- Model Building
- Model Evaluation and Maintenance
- Visualization and Presentation
- Deployment and Application Development
- Communication

Figure 4. Tasks and activities in data science.

Yu and Kumbier (2020) suggest that data science requires involvement of humans who collectively understand the domain and tools and propose core principles to guide the "implicit and explicit judgment calls throughout the data science lifecycle" (Yu and Kumbier, 2020, p. 3920).

Perhaps the most extensive study toward a goal similar to this article was started by the EDISON project (Demchenko, Belloum, et al., 2016). The project developed the "competence framework" for data scientists, categorizing competence areas of data science as *analytics*, *engineering*, *applications to business analytics*, *data management and governance*, and *scientific research methods*. The project discusses the connections between these areas with other knowledge areas. It also relates these areas with multiple 'soft' and 'hard' skills. Moreover, the project also proposed the data science "body of knowledge" (EDISON, 2018b), which connects relevant subjects from computer science and statistics to the EDISON competence framework. These subjects are drawn from a wide spectrum of generally accepted bodies of knowledge, for example, in computer science, software engineering, ICT, and so on. Notably, it recognizes "business analytics" as a special application domain and gives a detailed taxonomy, but it also leaves a "placeholder" for other, domain-specific knowledge.

We drew up several conclusions after reviewing the large and diverse set of literature. First, although many works cited do not distinguish between *knowledge* and *skills,* some contrast *hard* skills with *soft* ones. Second, within hard skills, all distinguish between a 'science' and 'engineering' angle, just as suggested at the outset of this chapter. Finally, many academic disciplines and topics in engineering are cited, although there is no obvious structure of these topics beyond a high-level classification.

There appears to be no consensus on what depth and breadth of engineering knowledge is required of data scientists. Some works go as far as to connect data

scientists to advanced software engineering practice and application development. Others imply elementary knowledge of computing and just enough 'skills' to perform data-related tasks. Similarly, we observe that the breadth of quantitative disciplines associated with data science is highly varied. For example, some reviews only mention 'statistics' and 'machine learning' toward this aspect, while others attempt a fuller description by enumerating on topics of other fields, for example, operations research, economics, and fundamental knowledge of scientific research methods.

The final element, 'domain knowledge,' receives the most diverse treatment. Most sources implicitly assume that the application domain refers to 'business analytics.' However, only some associate this domain to existing bodies of knowledge such as in *project management* or *business analysis*, while others associate it with common workplace skills such as *communication*.

Let us also briefly address 'skills,' competencies that lie outside familiar domains of organized 'knowledge.' These competencies, sometimes called *workplace skills*, are shared between professions and are general enough to warrant their own area of study. A recent and comprehensive summary of general workplace skills literature can be found in National Research Council (2012). The study proposed the taxonomy of "21st-century" workplace skills with three primary groups: *cognitive, intrapersonal,* and *interpersonal.* Cognitive skills refer to the abilities of gathering, processing, synthesizing, and analyzing *knowledge* such as 'critical thinking,' or 'problem solving.' Interpersonal skills, on the other hand, contain communication, collaboration, influence, and leadership. Others, like responsibility or work ethic, fall under *intrapersonal* skills.

Many of our sources mention workplace skills in the context of data science. Some examples of the mentioned skills are *decision making* (O'Neil & Shutt, 2013), *leadership* and *teamwork* (Dubey & Gunasekaran, 2015), *critical thinking* (Song & Zhu, 2016), and *problem-solving* (Cegielski & Jones-Farmer 2016). Moreover, the data science competence framework of EDISON (2018a) makes explicit connections between 21st-century skills and data science. Notably, the DARE project (2017) also proposes a wide set of skills for data scientists, among which they list not only critical thinking, planning and organizing, and problem solving but also customer focus, flexibility, business fundamentals, and cross-cultural awareness.

There appears to be no comprehensive prior study that thoroughly focuses on skills required in analytics and data science. Many sources define the discipline through knowledge areas and implicitly with the accompanying cognitive skills. Moreover,

reminiscent of the data scientist's role in 'driving' decision making, interpersonal skills for communication and collaboration are mentioned frequently. In the rest of this article, we focus on the knowledge components of data science.

5. IADSS Analytics and Data Science Knowledge Framework

In this section, we present and explore the main contribution of our article—a knowledge framework for analytics and data science. Drawing from previous studies and our own research and data collection efforts, we classify the knowledge areas. Our classification can be viewed as a foundation for a complete 'body of knowledge' for the profession, similar to those outlined in other professions (e.g., Bourque and Fairley, 2014; Project Management Institute, 2017).

Our taxonomy is not intended to be exhaustive and does not imply that all data scientists should have knowledge of all included topics. The taxonomy rather includes a combination of topics for the "umbrella term" data science as Irizarry (2020) describes it, practiced by a team of data science professionals with complementary knowledge and skills. We detail knowledge areas upon which there appears to be a wide consensus, and that occur consistently in previous classifications, task and activity studies, and job description research. For completeness, we include topics that are common prerequisites to the rest of the data science knowledge. Finally, we identify essential subjects and topics within broader disciplines, giving a deeper hierarchy of knowledge than previous studies.

Our selection of subjects is guided by our definition of knowledge as given at the beginning of the last section. Selected knowledge areas address the broad goal of "*Using Data to achieve specified goals by designing or applying computational inference methods.*" Notably, we aim to strike a balance between academic descriptions of the field, surveys of industry job descriptions, and self-reported activities and knowledge.

In agreement with the reviewed literature, we seek a separation between practical engineering aspects and knowledge of related academic disciplines. As such, we classify *knowledge* in analytics and data science into two primary domains. We refer to the collection of well-established academic disciplines and fields of study originating from science, applied science, statistics, and mathematics as the Science and Math domain. Knowledge on software, programming languages, development frameworks, and system components are compiled under the Programming and Technology domain. We break these two knowledge domains down into *fields* and further into *subjects,*

respectively. We present *fields* of our knowledge classification in Figure 5. Nonexhaustive, lower level classifications of fields into subjects and example topics are given in the Appendix.

Science and Math	**Programming and Technology**
Formal and applied science disciplines and fundamental fields associated with analytics & data science.	*Tools and technologies of designing, building, developing, controlling, operating and deploying computational systems.*
• **Scientific Method**: Basics of the scientific method, research methods, hypothesis formulation and problem identification. • **Mathematics:** Basic math, calculus and linear algebra • **Computer Science:** CS essentials such as data structures and algorithms, DB, OS, parallel computing, software engineering • **Statistics:** Probability basics, descriptive, inferential, and Bayesian statistics, stochastic processes and time series, causality, sampling • **Operations Research & Optimization:** linear programming, nonlinear optimization • **Data Preparation and Exploration:** Practical knowledge related to 'data analysis,' feature extraction and transformation, data cleaning, data preparation, data exploration • **Machine Learning:** Unsupervised and supervised learning models and algorithms, reinforcement and deep learning, text mining and NLP	• **General Purpose Computing:** general purpose programming languages, shell basics, version control, virtualization and containerization, cloud platforms • **Scientific Computing:** Statistical, numerical programming languages and libraries, ML libraries, development environments, data visualization tools • **Database & Business Intelligence:** relational DBs and SQL, data warehousing, querying and presentation • **Big Data:** Big data infrastructure, processing and execution environments, big data access and integration tools

Figure 5. High-Level summary of knowledge domains and fields related to analytics & data science.

In the Science and Math domain, we aim to identify the taxonomy of subject areas that are identified to be within the purview of 'consensus' data science and can be clearly mapped to existing academic fields of study. We classify subjects based on their proximity in their respective fields, for example, subjects taught in the same department or topics that are likely to be found in the same course or textbook.

Inspired by earlier definitions of data science that we reviewed for this study, the *scientific method* is the first field of the Science and Math domain. Here, we include knowledge of the scientific method as well as basic research methods that are both critical to 'solving problems,' a key part of our definition for data science. Although *mathematics* is scarcely mentioned in the previous studies, we place it in our framework for completeness, as it is required background knowledge for most of the remaining subjects and topics. Our field of *mathematics* refers only to basics, calculus, and linear algebra—common prerequisites to other fields in *Science and Math.*

Few of our sources cite formal knowledge of computer science, often omitting it in favor of a broader mention of 'hacking skills.' EDISON report (2018b) links some of their knowledge areas to specific topics in computer science, within the study of databases, information security, and software engineering. We also include more basic subjects in *computer science*, such as data structures and algorithms, operating systems, database management systems, and software engineering. We feel these subjects serve as a prerequisite to gaining a holistic understanding of the engineering practices in data science.

We detail the study of *statistics* into several subjects, ranging from basics to stochastic processes, time-series analysis, Bayesian statistics and inference, and causality. The *operations research and optimization* subject includes *linear programming*—a practice that sometimes appears in daily tasks of data science professionals—as well as *nonlinear optimization*, a subject that often makes an appearance when studying machine learning.

Under *data preparation and exploration*, we organize formal study of topics such as data cleaning, preprocessing, exploration and visualization (see, e.g., Aggarwal, 2015). We leave 'learning'-related components to *machine learning*.

In the Programming and Technology domain, we take a slightly different approach. We compile tools and technologies used by data science professionals into four main groups based on proximity of the use-cases they address, and groups of engineering professionals they are identified with. Again, without implying knowledge of all the tools used by data science professionals, we give taxonomy of tools implied by a consensus of previous works. We start with a field for common background knowledge in engineering, under *general purpose computing.* We organize other tools and technology in *scientific programming and engineering, database and business intelligence,* and *big data.*

Our first field, *general purpose computing*, includes knowledge of fundamental computer skills such as version control (e.g., git), interaction with shells, basics of *NIX operating systems, basic knowledge of cloud services (e.g., AWS, Azure), and containerization (e.g., Docker). We also include knowledge of general-purpose programming languages, such as Java or Python, in this field.

We classify *scientific programming and engineering* into statistical programming environments (e.g., R, SAS), scientific computing libraries and environments (e.g., NumPy, SciPy, Octave, MATLAB), machine learning libraries (e.g., TensorFlow), and development environments. We also include familiarity with visualization libraries and frameworks (e.g., Matplotlib).

In *database and business intelligence tools*, we include knowledge of SQL and relational database management systems as well as NoSQL data stores. We also cite data warehousing solutions, and querying tools such as Business Objects or Power BI. Our final field in programming and technology relates to a popular subject encountered in our review: big data. While the term itself is increasingly challenged as an accurate representation of the computational challenges associated with data science and deemed transitory, we use it as an umbrella term for knowledge of popular systems, frameworks, and tools from the Apache big data stack: HDFS, Hadoop, Spark, Hive, and so on.

Note that our framework leaves out a third pillar: *domain expertise*. As highlighted earlier, many of the works we cite implicitly assume the application domain of data science to be business analytics, or a specific business function. Some hierarchies of knowledge point to organized bodies of knowledge, for example, in operations management, project management, financial risk management, and so on. Others, without making the distinction between knowledge and skill elements, use ‘business skills’ in a vague manner to include traits such as leadership or interpersonal skills.

We purposely leave the application domain out of our classification of data science knowledge. Our definition of data science considers businesses as the most likely domain of application. That said—it does not leave out other domains such as scientific research, defense and government, or journalism, where surely the object of interest is not a user, customer, or other business entities. We believe the bodies of knowledge that lie outside our two domains (Science and Math, Programming and Technology) are too broad and too deep to address in any one study focusing on analytics and data science. While we clearly agree that domain knowledge is integral to data science

practice, we do not associate command of any one domain to the body of knowledge shared by data science professionals, which we studied here.

6. Conclusion

Since its appearance in the late 2000s, the definition of data science has varied greatly. This is a costly confusion, which we have echoed in our review of existing studies, in a profession where three million people are expected to work in the near future.

Today, the term data science mostly refers to a role conceived within 'big tech,' one that is characterized by the rapid application of varied and advanced quantitative skills to very large and disparate data stores. These applications require a unique combination of skills, found in expert engineers and research scientists. In turn, these professionals are expected to interface with other stakeholders in the business, in increasingly agile organizations. These needs give rise to the modern data science profession.

The rapid emergence of the role and some of its traits were anticipated in earlier literature: in works regarding "data analysis" (Tukey, 1962), "predictive modeling" (Breiman, 2001), and "data mining and knowledge discovery" (Fayyad et al., 1996). However, in our view, what ultimately defines today's data science professionals is the need to embrace a growing responsibility to connect data to decisions and products, and a set of challenges that transcend the boundaries of traditional industrial roles, skills, and fields of academic study.

Exactly what topics and technologies are worthy of study for professionals in the broader analytics and data science fields? Our literature review found varied answers and conclusions. We organized our findings into a proposed hierarchy of knowledge, emphasizing subjects that are widely agreed upon among existing studies, giving a detailed view of topics complete with the required background knowledge. Our foundational body of knowledge, to be revised by further discussions and research findings, can be used as a reference in designing data science curricula as well as in developing measurement and assessment tools.

This article intends to contribute to the ongoing discussion on data science knowledge and skills in the industry. It is a precursor to a series of studies by IADSS that will further explore how data science is practiced and seek consensus on a standardized framework for defining and assessing professional competencies. In the next steps, we will synthesize the outcomes of a global-scale quantitative research study on industry practices conducted by IADSS. This will facilitate our aim of building a framework for

knowledge and skills required in data science, and support building a measurement and assessment methodology. Through job market analysis and in-depth interviews with expert data scientists and executives across the globe, our next study will provide an extensive analysis of how knowledge and skills vary across titles and roles.

Disclosure Statement

Usama Fayyad and Hamit Hamutcu are co-founders of Initiative for Analytics and Data Science (IADSS).

Appendix

Knowledge Hierarchy: Topics and Subjects

Science and Math – The Scientific Method		
#	**Subject**	**Example Topics**
1	The Scientific Method	Formulating a research question, hypothesis, experiment, analysis, research methods, literature review
2	Problem Identification	Problem formulation and framing, design thinking

Science and Math – Mathematics		
#	**Subject**	**Example Topics**
1	Basics	Set theory, basic arithmetic and algebra, analytic geometry, trigonometry, quadratic forms, polynomials, rational and real number systems

2	Calculus	Limits, continuity, differentiation, integration, multivariable calculus, series
3	Linear Algebra	Vectors, matrices, systems of linear equations, eigenvalue decomposition, matrix decompositions, least squares problems

Science and Math – Computer Science		
#	**Subject**	**Example Topics**
1	Data Structures and Algorithms	Basic data structures (e.g., stack, lists, maps, etc.), basic algorithms (e.g., sorting, searching, merging, etc.), asymptotic notation
2	Databases and Data Processing Systems	Types of data storage and processing systems. RDBMS, data and database modeling, multidimensional DBs, OLAP, SQL, NoSQL
3	Software Engineering and Development	Requirements engineering, software project management, documentation, software testing, software maintenance, software engineering economics
4	Operating Systems	Operating system architecture, memory management, processes, threads, processor scheduling, file systems, security

5	Parallel and Distributed Computing	Levels of parallelism, parallel computation models, distributed computation models

Science and Math – Statistics		
#	**Subject**	**Example Topics**
1	Probability Basics	Basic combinatorics, probability, random variables, probability distributions, expected values, moments
2	Descriptive Statistics	Measures of central tendency, measures of variability and spread, correlations, contingency tables
3	Inferential Statistics	Hypothesis testing, multivariate analysis, parameter estimation, design of experiments
4	Bayesian Statistics	Bayes rule, Bayesian modeling, Bayesian linear models, approximate Bayesian inference, Markov chain Monte Carlo, Bayesian model selection
5	Stochastic Processes, Time Series, Survival Analysis	Random walks, Brownian motion, Markov chains, AR/MA processes, ETS, ARIMA, state-space models
7	Statistical Sampling	Sampling methods and biases, stratified and uniform sampling, efficient sampling methods, sample sufficiency metrics

Science and Math – Operations Research and Optimization		
#	**Subject**	**Example Topics**
1	Linear Programming	Model building (decision variables, objective functions, constraints), feasibility, extreme points and optimality, simplex algorithm, integer programming, other elementary models (TSP, VRP, etc.)
2	Nonlinear Optimization	Optimality conditions, gradient descent, Newton and quasi-Newton methods, constrained and unconstrained problem variants, advanced topics in optimization (simulated annealing, stochastic optimization, etc.)

Science and Math – Data Preparation and Exploration		
#	**Subject**	**Example Topics**
1	Data Preparation and Transformation	Scaling, standardization, normalization, binning and discretization, combining and aggregating data sets, feature selection and extraction
2	Data Cleaning	Handling missing and inconsistent data, outlier analysis
3	Data Exploration and Visualization	Data exploration strategies, univariate and multivariate data visualization, histograms, scatterplots, box plots, etc.

Science and Math – Machine Learning		
#	**Subject**	**Example Topics**
1	Unsupervised Learning	Clustering, association rule mining, dimensionality reduction (FA, PCA), mixture models
2	Supervised Learning	Regression models, generalized linear models, kernel methods, Gaussian processes, decision trees and forests
3	Reinforcement Learning	Bandits, Markov decision processes, policy and value iterations, Q-Learning
4	Deep Learning	Deep feedforward networks, autoencoders, deep recurrent networks (LSTM, GRU), deep generative models (VAE, GAN, etc.)
5	Text Mining and Natural Language Processing	Entity recognition, sentiment analysis, topic modeling, content classification

Programming and Technology – General Purpose Computing		
#	**Subject**	**Example Topics**
1	General-Purpose Programming Languages	C/C++, C#, Java, Scala, Python
2	Computing Fundamentals	*NIX shells, shell scripting, git
3	Virtualization/ Containerization	Docker, Kubernetes

4	Cloud Platforms	AWS, Azure

Programming and Technology – Scientific Computing		
#	**Subject**	**Example Topics**
1	Statistical	R, SAS, SPSS
2	Mathematical/Numeric	MATLAB, Mathematica, Octave, Julia, GAMS, Netlib, GSL, numPy, sciPy
3	ML Libraries	TensorFlow, scikit-learn, Keras, PyTorch, caret, party, etc.
4	Development Environments	Jupyter, Zeppelin, RStudio
5	Visualization	Matplotlib, ggplot, D3.js

Programming and Technology – Database and Business Intelligence		
#	**Subject**	**Example Topics**
1	RDBMS and SQL	MS SQL Server, Oracle, Teradata, DB2, Postgres, SQLite, MySQL
2	NoSQL and NewSQL	MongoDB, Redis, SAP Hana
3	Data Warehousing	Intellica, SAP, Teradata, Microsoft
4	Querying and Presentation	Business Objects, Power BI, QlikView, Qlik Sense, MicroStrategy

Programming and Technology – Big Data		
#	**Subject**	**Example Topics**
1	Infrastructure	HDFS, YARN
2	Processing and Execution	Hadoop, Spark, Streaming
3	Access	Hive, Pig
4	Integration	Flume, Sqoop

References

Aggarwal, C. C. (2015). *Data mining: The textbook.* Springer. https://doi.org/10.1007/978-3-319-14142-8

Blei, D. M., & Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of Sciences, 114*(33), 8689–8692. https://doi.org/10.1073/pnas.1702076114

Bourque, P., & Fairley, R. (2014*). Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0.* IEEE Computer Society Press.

Bowne-Anderson, H. (2018, August 15). What data scientists really do, according to 35 data scientists. *Harvard Business Review*. https://hbr.org/2018/08/what-data-scientists-really-do-according-to-35-data-scientists

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Cao, L. (2017). Data science: Challenges and directions. *Communications of the ACM, 60*(8), 59–68. https://doi.org/10.1145/3015456

Centers for Disease Control and Prevention. (n.d.) *The importance of KSAs*. Retrieved June 24, 2020, from https://www.cdc.gov/hrmo/ksahowto.htm

Cegielski, C. G., & Jones-Farmer, L. A. (2016). Knowledge, skills, and abilities for entry-level business analytics positions: A multi-method study*. Decision Sciences Journal of Innovative Education, 14*(1), 91–118. https://doi.org/10.1111/dsji.12086

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review, 69*(1), 21–26. https://doi.org/10.2307/1403527

Conway, D. (2010). *The data science Venn diagram*. Retrieved June 24, 2020, from http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

DARE Project. (2017). Recommended APEC data science & analytics (DSA) competencies. https://www.apec.org/-/media/Files/Groups/HRD/Recommended-APEC-DSA-Competencies-Endorsed.pdf

Davenport, T. H. (2020). Beyond unicorns: Educating, classifying, and certifying business data scientists. *Harvard Data Science Review 2*(2). https://doi.org/10.1162/99608f92.55546b4a

Davenport, T. H., & Patil, D. J. (2012). Data scientist. *Harvard Business Review, 90*(10), 70–76. http://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

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., … Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application, 4*(1), 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

Demchenko, Y., Belloum, A., Los, W., Wiktorski, T., Manieri, A., Brocks, H., Becker, J., Heutelbeck, D., Hemmje, M., & Brewer, S. (2016). EDISON data science framework: A foundation for building data science profession for research and industry. In *2016 IEEE International Conference on Cloud Computing Technology and Science CloudCom* (pp. 620–626). https://doi.org/10.1109/CloudCom.2016.0107

Dhar, V. (2013). Data science and prediction. *Communications of the ACM, 56*(12), 64–73. https://doi.org/10.1145/2500499

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Dubey, R., & Gunasekaran, A. (2015). Education and training for successful career in big data and business analytics. *Industrial and Commercial Training, 47*(4), 174–181. https://doi.org/10.1108/ICT-08-2014-0059

EDISON Community Initiative. (2018a). EDISON data science framework part 1. *Data science competence framework (CF-DS)* (Rev. 3).

EDISON Community Initiative. (2018b). EDISON data science framework part 2. *Data science body of knowledge (DS-BoK)* (Rev. 3). Retrieved May 30, 2020, from https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_DS-BoK-release3-v06.pdf

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37–37. https://doi.org/10.1609/aimag.v17i3.1230

Gorman, M. F., & Klimberg, R. K. (2014). Benchmarking academic programs in business analytics. *INFORMS Journal on Applied Analytics, 44*(3), 329–341. https://doi.org/10.1287/inte.2014.0739

Grady, N., & Chang, W. (2015). National Institute of Standards and Technology (NIST) big data interoperability: *2015 NIST Big Data Public Working Group Definitions and Taxonomies Subgroup. Framework:* Vol. 1, Definitions, NIST Special Publication, 1500-1.

Gray, J. (2007). *Jim Gray on eScience: A transformed scientific method*. Microsoft Research.

Hammerbacher, J. (2009). Information platforms and the rise of the data scientist. In *Beautiful data: Stories behind elegant data solutions* (pp. 73–84). O'Reilly Media.

Harris, H. D. (2011). *Data science, Moore's Law, and Moneyball.* http://www.harlan.harris.name/2011/09/data-science-moore-s-law-and-moneyball/

Harris, H. D., Vaisman, M., & Murphy, S. (2012). *Data scientists survey results teaser*. Data Community DC. http://www.datacommunitydc.org/blog/2012/08/data-scientists-survey-results-teaser

Hayashi, C. (1998). What is data science? Fundamental concepts and a heuristic example. In C. Hayashi, K. Yajima, H.-H. Bock, N. Ohsumi, Y. Tanaka, & Y. Baba (Eds.),

Data science, classification, and related methods (pp. 40–51). Springer Japan. https://doi.org/10.1007/978-4-431-65950-1_3

International Data Science in Schools Project. (2019). *Curriculum frameworks for introductory data science IDSSP curriculum team*. http://www.idssp.org/files/IDSSP_Frameworks_1.0.pdf

Irizarry, R. A. (2020). The role of academia in data science education. *Harvard Data Science Review, 2*(1). https://doi.org/10.1162/99608f92.dd363929

Markow, W., Braganza, S., Taska, B., Miller, S. M., & Hughes, D. (2017). *The quant crunch: How the demand for data science skills is disrupting the job market.* Retrieved May 30, 2020, from https://www.ibm.com/downloads/cas/3RL3VXGA

Mills, R. J., Chudoba, K. M., & Olsen, D. H. (2016). IS programs responding to industry demands for data scientists: A comparison between 2011–2016. *Journal of Information Systems Education, 27*(2), 131–140. https://jise.org/Volume27/n2/JISEv27n2p131.pdf

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options.* National Academies Press. https://doi.org/10.17226/25104

National Research Council (U.S.), Pellegrino, J. W., Hilton, M. L., National Research Council (U.S.), National Research Council (U.S.), & National Research Council (U.S.) (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* National Academies Press. https://doi.org/10.17226/13398

Naur, P. (1966). The science of datalogy. *Communications of the ACM*, *9*(7), 485. https://doi.org/10.1145/365719.366510

Naur, P. (1974). *Concise survey of computer methods*. Lund: Studentlitteratur.

Nelson, S. (2018). *4 types of data science jobs.* Udacity. https://blog.udacity.com/2018/01/4-types-data-science-jobs.html

North Carolina State University. (2019). *Graduate degree programs in analytics and data science.* https://analytics.ncsu.edu/?page_id=4184

National Science Foundation. (2005). *Long-lived digital data collections: Enabling research and education in the 21st century.*

Office of Personnel Management. (2007). *Delegated examining operations handbook: A guide for federal agency examining offices.*

Office of Personnel Management. (2013). *Office of Personnel Management's multipurpose occupational systems analysis inventory–Close-Ended (MOSAIC) Competencies.*

O'Neil, C., & Schutt, R. (2013). *Doing data science*. O'Reilly.

Organisation for Economic Co-operation and Development. (2018). *Skills for jobs*. OECD Publishing.

Patil, D. J. (2011). *Building data science teams.* O'Reilly Media.

Project Management Institute. (2017). *PMBOK® guide* (6th ed.).

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data, 1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Radovilsky, Z., Hegde, V., Acharya, A., & Uma, U. (2018). Skills requirements of business data analytics and data science jobs: A comparative analysis. *Journal of Supply Chain and Operations Management, 16*(1), 82–110.

Schoenherr, T., & Speier-Pero, C. (2015). Data science, predictive analytics, and big data in supply chain management: Current state and future potential. *Journal of Business Logistics, 36*(1), 120–132. https://doi.org/10.1111/jbl.12082

Song, I.-Y., & Zhu, Y. (2016). Big data and data science: What should we teach? *Expert Systems, 33*(4), 364–373. https://doi.org/10.1111/exsy.12130

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics, 33*(1), 1-67. https://doi.org/10.1214/aoms/1177704711

Van der Aalst, W. M. P. (2014). Data scientist: The engineer of the future. In K. Mertins, F. Bénaben, R. Poler, & J.-P. Bourrières (Eds.), *Proceedings of the I-ESA Conferences: Vol. 7. Enterprise Interoperability VI* (pp. 13–26). Springer. https://doi.org/10.1007/978-3-319-04948-9_2

Van Dyk, D., Fuentes, M., Jordan, M. I., Newton, M., Ray, B. K., Lang, D. T., & Wickham, H. (2015, October 1). ASA statement on the role of statistics in data science. *AMSTAT News*. https://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/

Willems, K. (2015). *The data science industry: Who does what (Infographic).* https://www.datacamp.com/community/tutorials/data-science-industry-infographic

Wing, J. M. (2019). The data life cycle. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.e26845b4

Wirth, R. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (pp. 29–39). http://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Wu, C. F. J. (1997). *Statistics = data science?* Inaugural lecture for the H. C. Carver Chair in Statistics at the University of Michigan. https://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf

Yu, B., & Kumbier, K. (2020). Veridical data science. *Proceedings of the National Academy of Sciences, 117*(8), 3920–3929. https://doi.org/10.1073/pnas.1901326117

Footnotes

1. The reader can find more information and get engaged at www.iadss.org. ↩

2. As opposed to ‘data modeling.’ See Breiman’s essay (2001) and its discussion by Donoho (2017). ↩