**Harvard Data Science Review**

# From Unicorn Data Scientist to Key Roles in Data Science: Standardizing Roles

**Usama Fayyad[1,2] Hamit Hamutcu[3]**

**[1]Institute for Experiential Artificial Intelligence, Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, United States of America,**

**[2]Open Insights, Bellevue, Washington, United States of America,**

**[3]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:** Jul 28, 2022

**DOI:** https://doi.org/10.1162/99608f92.008b5006

**ABSTRACT**

The lack of an agreed-upon classification of job roles related to data science is causing much confusion that is challenging to the industry, educational sector, and practitioners. Prior work in this area has considered different aspects from different fields or points of view and has shown that more detail is needed in subcategorizing data science professionals. However, other prior work has also shown that avoiding the detailed subcategorization leads to challenging problems, for example, the pursuit of the elusive 'data science unicorns.' In this article, we target a simplification of prior work and an anchor categorization of job roles with clear definitions and expectations from each. We achieve this through analysis of survey results, LinkedIn profiles and job descriptions, and in-depth interviews with managing and hiring executives in data science. We also use our judgment as long-term practitioners and employers of data scientists to provide a practice-guided view of the problem.

Our analysis has led to a simplification into three key role families with complementary skills: data analyst, data scientist, and data engineer. We believe this anchor categorization helps resolve several problems, including recruiting, forming, training, managing, and retaining effective data science teams. Although we realize there are and will continue to be many variations of these proposed anchor roles, this simplification is an effective tool to bypass the data science unicorn issue, and it can be used as a basis to establish more specialized or domain-specific roles. The combined skills in these role categories converge on the body of knowledge specification from the Initiative for Analytics and Data Science Standards (IADSS) data science knowledge framework (Fayyad & Hamutcu, 2020). The concise and familiar role categories simplify the problem and decompose it into more solvable subchallenges. We describe the essential knowledge required for each role and how, when, and in what ways it can be varied and extended. This description helps align expectations and serves as a step to tackle the pressing issue of training, evaluating, and building effective data science teams.

**Keywords:** data science, job roles, skill requirements in data science, categorization of roles, data science unicorns

---

# 1. Introduction

Coined by DJ Patil and Jeff Hammerbacher (Davenport & Patil, 2012) to describe the roles of their team members at LinkedIn and Facebook and popularized by Thomas Davenport and DJ Patil as the "sexiest job of the 21st century" (Davenport & Patil, 2012), the data scientist job title was deemed quite demanding in the early 2010s. Back then, professionals holding the title were expected to blend wide ranging knowledge with a fairly diverse set of skills—machine learning, numerical computing, data management, statistical inference, distributed programming, business, and communication. These professionals were described as, for instance,

half hackers, half analysts and also product and final insight builders (Woods, 2011), or engineers who find insights in data employing the scientific method and applying data discovery tools (Press, 2012) as reviewed by Chatfield et al. (2014).

An instant critique of the initial idea of a data scientist is that it takes a 'unicorn' to master all necessary knowledge and skills relevant to data science. The unicorn idea poses a significant challenge for the field: unicorns are rare. In fact, they may not exist, as Davenport (2020) puts it. Realistically, such a broad set of skills can only be acquired at a generic level or expected from a single person in small organizations (Garten, 2018; Jung, 2020).

However, the confusion about what exactly a data scientist does remains. As Bowne-Anderson (2018) analyzed, whether the data scientist title requires acquiring a generic set of skills or refers to a specialization is still unclear. Due to its vagueness, the professionals with entirely different knowledge, skills, and jobs might claim the data scientist title (Miller, 2014). The liberal and inflated usage of the data scientist title, as a result, might cause the professionals to abandon using it completely.

On the other hand, the multidisciplinary nature of data science is undisputed. It is acknowledged as an umbrella term but considered a set of activities more complex than a single professional— the data scientist—can perform (Irizarry, 2020). Reliance solely on data scientists in all aspects of data science, however, puts the organizations at risk of underperforming, for instance, in complex data-engineering problems (Anderson, 2018, 2019; Loukides & Lorica, 2017), in specific data analytics tasks (Henke et al., 2018; Kozyrkov, 2018b), or in making use of domain expertise (Viaene, 2013). As numerous previous opinions and studies agree, instead of looking for data science unicorns, it is better to form teams with individuals specializing in different aspects of data science (Olavsrud, 2015; O'Neil & Schutt, 2013; Zhang, 2019).

In summary, although *data scientist* has emerged as a job title, it has had different meanings over the past decade. The rarity of data science unicorns and the buzz around data science have driven every industry, function, and business to look for their definition of the role. Building a construct with clear descriptions for a set of key roles in a data science organization has thus become both an essential and urgent necessity.

This article is the second in a series by IADSS, an initiative established to provide a framework for professional standards in data science and analytics.[1] In the precursor of this article, we introduced a *data science knowledge framework* that includes various knowledge areas under *science & math* and *programming & technology* domains, from computing essentials, machine learning theory, and spreadsheets to big data technology (Fayyad & Hamutcu, 2020). The purpose of this article is to discuss what our analysis showed to be the main categories of roles that are key to a data science organization in three main job categories: *data scientist*, *data analyst*, and *data engineer*. The list of key roles is kept concise, ensuring generality across organizations. Furthermore, we believe the key roles can act as anchors of an extended set of data science job

titles as well—some job titles are already very close to the key roles, and some of them, we argue, will potentially converge toward them.

We compile our results from (i) the findings of a global public questionnaire that we conducted with a broad range of roles represented—from engineering to managerial, and (ii) the skills that the professionals self-declare on their LinkedIn profiles or are listed in job postings. The findings are also guided by our insights from individuals leading data science teams and divisions in some of the major data science employers—obtained via a set of private interviews and various interactions with the data science community at conferences and workshops.

We discuss the key roles through a comparative lens, providing pairwise comparisons of a data analyst and data engineer with a data scientist and summarize their differences. A closer look into self-declared data scientists also reveals different data scientist subgroups, highlighting the confusion of professionals on their own job descriptions. Finally, combining our sources and experience, we map the subject areas and fields in the knowledge framework to the key roles in data science and develop precise specifications.

We believe our specification will reduce the confusion around data science roles and guide employers in their recruiting and organizational efforts, the aspiring and existing data science professionals in their career planning and development, and the educational organizations in designing better and more consistent data science curricula.

The rest of this article is organized as follows. Section 2 summarizes the data sources that helped us build our view of data science roles. Section 3 reviews some role classifications found in the literature and introduces the three key roles: data analyst, data engineer, and data scientist. Section 4 develops a specification of the key roles characterized by the subject areas and fields in the IADSS knowledge framework. Finally, we summarize the findings, list the implications, and conclude the article in Section 5.

## 2. Data Sources

Later in the article, we build a list of key roles in data science and describe them. While synthesizing our descriptions, we benefited greatly from data sources that provide insights into how data science is practiced in the industry.

We rely heavily on the responses to a global questionnaire that we conducted with data science professionals. We also benefited from interviews with executives leading data science teams across the globe. Finally, we conducted an analysis of public professional profiles and job postings on LinkedIn. The questionnaire closely follows the IADSS Data Science Knowledge Framework, which we presently review for completeness.

## 2.1. A Review of the IADSS Data Science Knowledge Framework

*IADSS Data Science Knowledge Framework* (Fayyad & Hamutcu, 2020) proposes a classification of relevant knowledge areas as a foundation for a complete body of knowledge for analytics and data science. The knowledge framework is organized into fields under two major knowledge domains: *science & math* and *programming & technology*. The *science & math* domain is a taxonomy of relevant subject areas classified into fields within the proximity of related academic disciplines to analytics and data science. The subject areas include statistics, operations research and optimization, computing essentials, data mining, and machine learning. Programming & technology, on the other hand, is a compilation of computational tools and technologies associated with designing, building, utilizing, and operating analytics and data science practice. It is organized into groups based on the use cases they address, including scientific computing, databases, business intelligence, and big data technology. For completeness, basic knowledge of mathematics (within the *science & math* domain) and general-purpose computing (within the *programming & technology* domain) are included in the framework, as they constitute a common background for analytics and data science professions.

In summary, the knowledge framework includes knowledge areas upon which there appear to be a broad consensus and the common prerequisites to the rest of data science knowledge. The framework leaves the domain expertise out of classification. The workplace skills, such as leadership or communication, are also excluded.

## 2.2. The Questionnaire

We conducted a global 50-question survey asking for the extent of knowledge required to practice a data science and analytics profession based on the IADSS Knowledge Framework. The survey participants were asked to judge the relevance of these specific knowledge and skills to their occupations. They provided, to each, a degree of relevance from “not relevant” to “must- have” in a total of five levels.

Out of more than 800 people who responded to the survey, we filtered reliable responses with complete answers to 536 respondents.[2] In addition to the 50 questions addressing the relevance of specific knowledge and skills, we asked respondents about their job titles, the industry of the companies they are working for, the size of their companies, the amount of their experience, and geographic regions they are based in. The survey was advertised and distributed in professional networks and platforms, and we should note that there might be bias, such as our sample consisting of people who are more likely to respond to surveys.

The respondents represented a variety of roles and organizations. Table 1 lists the most frequent 10 titles among the survey responses.

**Table 1: Top-10 titles among survey responses.**

| |
|---|
| Data scientist |
| Data analyst |
| Analytics manager/director |
| BI analyst/specialist |
| Chief data/analytics officer |
| Data science director |
| Data engineer |
| Machine learning engineer |
| Big data engineer |
| Machine learning scientist/expert/specialist |

Figure 1 breaks down the respondents by industry (a), organization size (b), and geography (c).

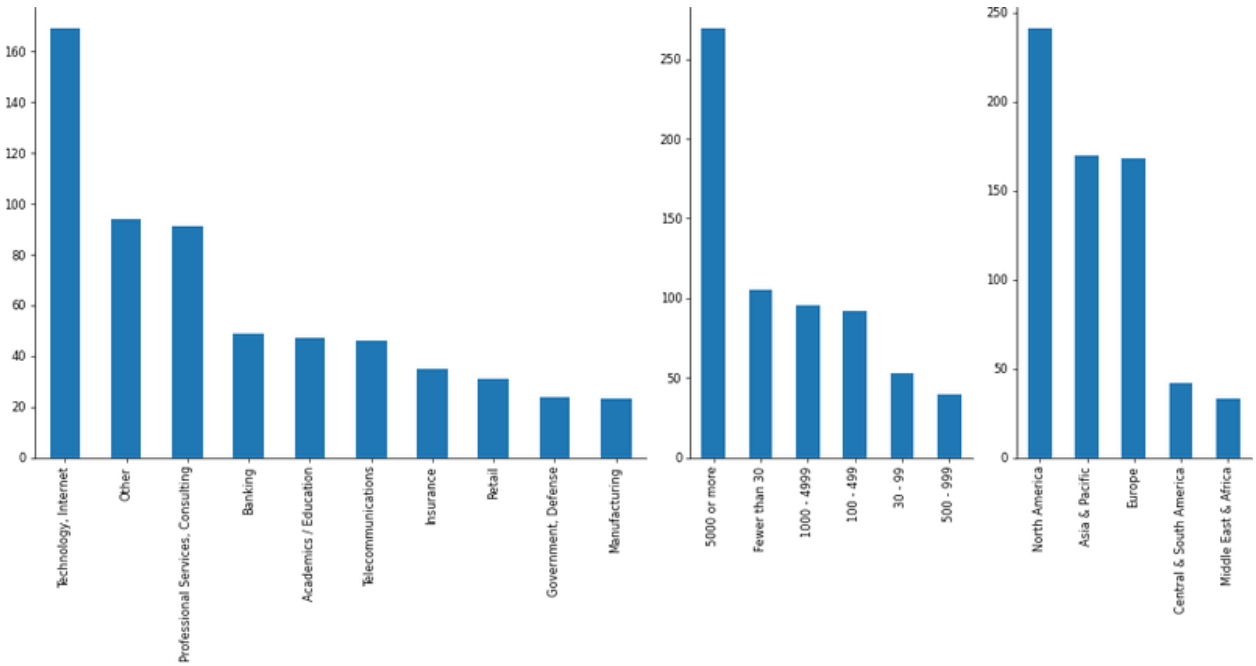



**Figure 1.** Number of respondents of the questionnaire by industry (left), by organization size (middle), and by geography (right).

The 50-question survey covers the topics in the IADSS Knowledge Framework, including data preparation and exploration, statistics, machine learning, computer science fundamentals, computing (general-purpose, mathematical/statistical), databases, and several other areas under *science & math* and *programming & technology*.

We refrained from mentioning specific technologies and instead focused on concepts. For instance, for generating visual descriptions of data, we ask for the ability to generate and interpret histograms, time-series visualization, and so on, not a particular tool to generate visualizations.

A sample list of questions and the corresponding topics in the knowledge framework are listed in Table 2. The full question set is provided Appendix A.

**Table 2. Sample questions and corresponding IADSS Knowledge Framework areas.**

| Subjects listed in IADSS Knowledge Framework | Questionnaire items |
|---|---|
| Data preparation and transformation | Data transformation/reduction (discretization/binning, sampling, normalization/scaling, engineering new variables |
| Machine learning | Predictive modeling (decision trees, linear models for regression and classification, model validation, etc.)<br><br>Machine learning (more recent theory, e.g., neural networks, deep learning, Support Vector Machines (SVM), boosted trees, etc.<br><br>Natural language processing/text mining |
| Statistical computing | Statistical programming languages (e.g., R, SAS Language, MATLAB/Octave, etc.) |
| RDBMS & SQL | SQL, understanding of relational databases<br><br>Enterprise relational DBMS (e.g., Oracle (PL/SQL), Teradata, DB2, SQL Server, Vertica)<br><br>Open-source relational DBMS (e.g., MySQL, Postgres) |
| Big data infrastructure | Big data development and maintenance for infrastructure and applications<br><br>Cluster resource managers, other big data technologies (e.g., Mesos, YARN) |
| Big data access | Distributed databases, data warehousing, and storage (e.g., Hive/Pig, HBase, Cassandra, Parquet) |

| | |
|---|---|
| Querying and Presentation | Reporting and visualization software (e.g., Tableau, QlikView, Power BI, TIBCO Spotfire, MicroStrategy, Hyperion, Cognos, Business Objects)<br><br>Insight presentation--building a data-driven story line and present to an executive/client audience |

The questionnaire extends the knowledge framework with a set of questions that address common workplace skills, asking about *self-sufficiency*, *cross-functional collaboration* skills, and proficiency in *office productivity* tools. Another set includes leadership and management questions, that is, *project management*, *peer leadership*, and *executive leadership*. Finally, domain expertise is addressed with questions about an experience in applying analytics in a domain, knowledge of the terminology, practices, and the key performance indicators (KPIs) of a domain, and expert understanding of a domain or industry.

## 2.3. LinkedIn Professional Profiles and Job Postings

As part of our data collection efforts, we searched for the commonly encountered job titles on LinkedIn, a popular online network for professionals.

We collected 2,726 professional profiles and 487 job postings (primarily in the United States and United Kingdom). One-fourth of the collected job postings are for a *data scientist* position, followed by *big data engineers* and *machine learning engineers*. *Data scientists* are most commonly encountered in professional profiles, too, and they are followed by *data analysts*, *machine learning engineers*, and *data engineers*.

LinkedIn members list a set of key phrases as their skills on their profiles. Similarly, job postings may include a set of skills—depending on the poster's preferences. In the analysis, we used the self-declared skills by the members (not skills claimed by others for them) and the desired skills if provided by the job posters.

Similar to the questionnaire, the key phrases are mapped to the subjects and fields in the knowledge framework. A sample list can be found in Table 3 and the complete list in Appendix B.

**Table 3. Sample key phrases extracted from LinkedIn professional profiles and job postings and the corresponding IADSS Knowledge Framework areas.**

| IADSS Knowledge Framework | | |
|---|---|---|
| **Knowledge area** | **Subject** | **LinkedIn key phrases** |

| | | |
|---|---|---|
| Computer science | Data structures & algorithms | Data structures |
| | Databases and data processing systems | SQL, NoSQL, databases, relational databases, database design, database administration, data warehousing, data warehouse architecture, storage, OLAP, data marts, data processing, databases, data architecture, data modeling, data management |
| | Software engineering & development | Test automation, testing, solution architecture, software project management, software engineering, software documentation, software development lifecycle, software development, software design, agile, scrum, requirements gathering, requirements analysis, object-oriented programming, object-oriented design |
| | Operating systems | Linux |
| | Parallel and distributed computing | MapReduce, parallel computing, high-performance computing, distributed systems |

| | | |
|---|---|---|
| DB & BI | RDBMS & SQL | SQL, T-SQL, PL/SQL, SQL tuning, relational databases, Oracle SQL, Teradata, PostgreSQL, Oracle database, MySQL, Microsoft SQL Server, Microsoft Access, DB2 |
| | NoSQL/NewSQL | SAP Hana, MongoDB, Apache HBase, Elasticsearch, Cassandra, NoSQL |
| | Data warehousing | Data warehouse architecture, data warehousing, Netezza, SAP Business Warehouse, SAP R/3, Microsoft SQL Server Analysis Services, Oracle, Teradata |
| | Querying & presentation | Tableau, SQL Server Reporting Services, SAP BusinessObjects, SAP BI, reporting & analysis, QlikView, Qlik, Microsoft Power BI, pandas, Oracle Business Intelligence Suite, MicroStrategy, Microsoft Excel, dashboard, Crystal Reports, Cognos, Business Objects, Spotfire, Looker |
| Machine learning | Supervised learning | Predictive modeling, predictive analytics, regression, logistic regression, linear regression, decision trees, convolutional neural networks, recommender systems |
| | Unsupervised learning | Cluster analysis |
| | Deep learning (DL) | Artificial neural networks, deep learning, convolutional neural networks |
| | Text mining and natural language processing (NLP) | NLP, text mining, text analytics, information retrieval |

## 2.4. Private Interviews

We conducted in-person interviews with 25 executives heading data science teams and divisions in major data science employers from various industries, including technology, financial services, and health care.

An interview takes 45 minutes and is organized into questions asking about the distinctive characteristics of different professionals in their teams and their typical activities, technical skills expected from these professionals and how these skills are assessed, recruitment processes, and transitions between roles.

The interviewees put their perspectives on these aspects, giving us an idea of how a typical data science team of professionals, organized into multiple roles, operates.

## 3. Key Roles in Data Science

As discussed at the outset of the article, contrary to the initial framing of the title *data scientist*, there is more to data science than a single professional can perform alone. A successful data science organization requires professionals who manage, analyze, and connect data to the business goals, reconcile data sets, ensure the data model's integrity, and make inferences and predictions from data through computational methods. The industry, indeed, transitions from the idea of relying on data science unicorns to forming data science teams as a pragmatic and more effective approach.

Forming a data science team starts with identifying a set of complementary roles needed to achieve the goals of a data science project. In the attempts to formulate the ingredients of such teams, several prior efforts have considered multiple dimensions: classifying data scientists, standardizing job titles, characterizing unicorns, and data-driven classification, that is, emerging titles by analyzing data on the tasks performed and roles played by the people performing the data science work. We provide context for these below and discuss our approach to the problem.

**Classification of Data Scientists:** Numerous studies have attempted to provide a data scientist typology without assigning role specifications to titles. Instead, they focus on the variety of data scientists. Mayank (2017), in their "The 7 Types of Data Scientists," compiles their typology from job postings. "The Kinds of Data Scientists" classifies them based on their 'deliverables' presented (Garten, 2018). Berthold (2019) classifies data scientists based on their expertise. Of particular note is "Analyzing the Analyzers," where Murphy et al. (2013), based on their analysis of a questionnaire conducted with professionals among analytics, data science, big data, applied statistics, machine learning spaces, coin four major professional profiles: *data businesspeople*, *data creatives*, *data researchers*, and *data developers*.

**Standardizing Job Titles:** Some studies propose a set of standard job titles, or families, associated with the corresponding competencies. As an example, take the "Data Science Professional Profiles" (DSPP) report released within the EDISON Data Science Framework (Demchenko & Cuadrado-Gallego, 2020). The DSPP report provides a list of occupations for inclusion in the third level occupation groups in the European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy. The list includes the titles *data scientist*, *data science researcher*, *data science architect*, *data science application programmer/engineer*, *(big) data analyst*, and *business analyst*. They also list a number of titles for managerial, 'data handling,' technician, and support working roles. However, detailed definitions of these roles are not provided.

**Characterizing Unicorns:** Data scientists are described as data science unicorns (under different names) in some specifications. For instance, Murphy et al. (2013) characterize their "data-creative" by being a "jack of all trades." Among the extensive collection of the titles they propose, Demchenko & Cuadrado-Gallego (2020) still describe a data scientist as a generalist, where the job holders are expected to perform daily activities "ranging from finding and interpreting rich data sources, managing large amounts of data, creating visualizations, building mathematical models, and presenting and communicating data insights and findings, to recommending ways to apply data." Nelson (2018), too, adds a "data science generalist" title, in addition to the data analysts, data engineers, and machine learning engineers. Kozyrkov (2018a), among nine roles, describes the data scientist as an expert analyst, statistician, and machine learning engineer. Willems (2015) refers to data scientists as "as rare as unicorns." Although such data science unicorns also exist among our questionnaire respondents, we attribute this description of a data scientist to "post-unicorn thinking not being yet penetrated in how individuals and organizations think about and structure data science education and capability-building," as Davenport (2020) accurately states it. We also observed during our private interviews that data science professionals are usually expected to perform the tasks of multiple roles (such as data scientist and data engineer) in small organizations, typically one person starting out and then hiring other team members as the organization grows.

**Data-Driven Classification:** The methodologies of coming up with a classification also vary across studies. Some studies are purely curated by experts. Another group of studies is purely data driven. First, subjects are extracted from self-declared skills and labeled post facto. Later, groups of subjects are identified and assigned to titles, types of data scientists, or job families. DSPP classification is provisionally curated by the experts from professional profiles collected from job advertisements, blogs, and discussions in different forums. In "Analyzing the Analyzers," (Murphy et al., 2013) on the other hand, the authors apply a *non-negative matrix factorization*, as described by Brunet et al. (2004) in their questionnaire responses.

De Mauro et al. (2017) provides another example of a purely data-driven classification, where they investigate a classification of roles in data science systematically from a human resources perspective. They analyze 2,700 job posts that contain the keywords 'Big Data' in either the title or the description, and apply *latent Dirichlet allocation* (Blei et al., 2003) to extract nine groups of keywords that frequently co-occur, then label the groups with titles corresponding to subjects, and finally combine them into job families, namely, business analyst, data scientist, developer, and engineer. As a result, they report data analysts and data engineers as frequently occurring job titles for data scientists. In their "The Elusive Data Scientist," Nelson and Horvath (2017) also organize data science activities under five job families, namely, business analysis, statistical analysis, technical analysis, leadership, and project management.

**Our Approach:** We believe the data scientist typologies only contribute to the confusion by labeling professionals with distinct roles with the same 'data scientist' title. One anecdotal example we saw was an organization relabeling all of their data analysts to data scientists in the hopes of attracting more talent, which

obviously is problematic in many ways. On the other hand, introducing new titles or types that do not reflect the actual titles commonly encountered in the industry would risk widespread adoption. Therefore, toward standardization of their roles and to ensure widespread adoption by the industry, we propose a concise list of three 'post-unicorn' titles, *data scientist*, *data analyst*, and *data engineer*, as key to the data science teams and organizations. We find the list complete in the sense that (i) the roles associated with the three titles are complementary and (ii) the knowledge and skills associated with these roles fully cover the IADSS data science knowledge framework (Fayyad & Hamutcu 2020).

We develop our proposal in multiple steps, based on the previously mentioned data sources and our judgment. This section starts by unifying multiple job titles into one based on the similarities in the knowledge and skills that jobholders declare and our opinion on the permanence of the titles. Through pairwise comparisons of the data analysts and data engineers with the data scientists, we then list some of the knowledge and skills relevant to data science and that should be attributed to the data analysts and engineers rather than data scientists. We also highlight the diversity among the data scientists, with a particular focus on *data science unicorns*, who still exist. In what follows, we provide a complete listing of the knowledge and skills associated with the three key roles and a guide to job specifications.

## 3.1. Data Analysts and Data Engineers

The questionnaire respondents identifying themselves as data analysts, business intelligence (BI) analysts, or in adjacent roles provide responses so similar that they are indistinguishable and can be considered as members of a single, unified profession. Roughly half of data analysts have their 'closest neighbors' among BI analysts and vice versa.[3] This observation bases our first unification. We will combine data analyst and BI analyst titles into one, collectively refer to it as a **data analyst**, and list the distinctions based on this unification onward.

A similar unification can be made for data engineers. For example, 'big data engineer' is not a rarely encountered title; in fact, it is particularly common among LinkedIn job postings. We, however, find such a specification transitory for our purpose of listing the key roles. Without a need to be specific in their job titles, data engineers already assume big data technology as part of their jobs. Popular big data technologies such as Apache Hadoop and Apache Spark and the big data keyword itself are among the top-10 skills of LinkedIn professionals who do not call themselves a big data engineer, but a data engineer. This line of thinking leads us to unify related engineering titles under the data engineer role for standardization purposes.

Let us now discuss in detail how a data analyst and a data engineer are different from a data scientist, with reference to their questionnaire responses. The professionals' self-declaration of their roles reveals a stark distinction of a data scientist from a data analyst: data scientists have superior knowledge of the fundamentals of machine learning; that is, theoretical knowledge corresponding to common machine learning algorithms and architectures and inclusion of scripting languages and libraries for numerical computing and machine learning

in their arsenal of data science tools. Proportions of *must-haves* and *should-haves* in knowledge areas related to machine learning among data scientists and data analysts are summarized in Table 4.

**Table 4. Proportion of must-have/should-have responses to the questionnaire items related to machine learning by data analysts and data scientists.**

| | The Proportion of Must-Have/Should-Have Responses | |
|---|---|---|
| **Questionnaire Items Directly Related to Machine Learning** | **Data Analysts** | **Data Scientists** |
| Machine learning (more recent theory, e.g., neural networks, deep learning, support vector machines, boosted trees, etc.) | 23% | 74% |
| Libraries for machine learning / numerical computing (e.g., TensorFlow, scikit-learn, NumPy, SciPy, pandas, Theano, Torch, Keras) | 29% | 76% |
| Predictive modeling (decision trees, linear models for regression and classification, model validation, etc.) | 39% | 90% |

Based on the previous observation, one might be tempted to conclude that the professionals call themselves data analyst if they lack theoretical machine learning knowledge and development skills to an extent expected from a data scientist. Due to the popularity of machine learning, it is easy to believe that data scientists are upgraded versions of data analysts. Based on our observations, this conclusion is incorrect. Although they consider machine learning knowledge and skills only as a nice-to-have for the analyst profession, a data analyst is typically stronger in their expertise in using enterprise databases, reporting and visualization software, and office productivity tools. They can leverage these skills along with knowledge of the demands of a business user to formulate analyses that are easier to understand or relate to a familiar business/domain setting.

That said, data analysts and scientists are closer when it comes to statistics and statistical programming languages. Sixty-three percent of the data analysts who responded to the questionnaire find statistics knowledge at least as a *should-have*, and 45% of them knowledge in statistical programming languages, also as a *should-have*. The proportion of should-haves and must-haves of the knowledge and skills provided by the data analysts and data scientists to selected questionnaire items are shown in Table 5.

**Table 5. The proportions of must-have/should-have responses to selected questionnaire items by data analysts and data scientists.**

| | | The Proportion of Must-Have/Should-Have Responses | |
|---|---|---|---|
| | **Questionnaire Items** | **Data Analysts** | **Data Scientists** |
| **Databases** | SQL, understanding of relational databases | 87% | 79% |
| | Enterprise relational DBMS | 56% | 43% |
| | Open-source relational DBMS | 39% | 38% |
| **Reporting and visualization tools** | Office productivity, spreadsheet tools | 90% | 66% |
| | Reporting & visualization software | 75% | 47% |
| | Enterprise analytics and data mining software | 26% | 19% |
| **Statistics** | Statistics | 63% | 85% |
| | Statistical programming languages | 45% | 64% |

We now focus on the *data engineer*, the next key title that distinguishes itself in a particular set of skills, which, still, some data scientists (and organizations) might develop a misconception that data scientists can easily acquire and practice, as Anderson (2019) observes.

Most of the interviewees who head data science and analytics teams expressed their views on the level of programming required for a data scientist as at a prototype level, saying, “a data scientist does not need to develop production-ready software, but they should be able to implement solutions at a prototype level.” We attribute deeper software development/engineering knowledge to data engineers (although they are still not software engineers). Not surprisingly, data engineers are also stronger in developing and maintaining databases and data warehouses, big data, and streaming data frameworks, according to their responses to the questionnaire. On the other hand, the data scientists found theoretical knowledge of machine learning more relevant, as one would expect. Table 6 shows the proportion of should-haves and must-haves of the knowledge and skills to selected questionnaire items, as rated by the data engineers and data scientists.

**Table 6. The proportions of must-have/should-have responses to selected questionnaire items by data engineers and data scientists.**

| | | The Proportion of Must-Have/Should-Have Responses | |
|---|---|---|---|
| | **Questionnaire Items** | **Data Engineers** | **Data Scientists** |
| **Big Data** | Big data development and maintenance for infrastructure and applications | 50% | 24% |
| | General knowledge of distributed storage/computing frameworks (e.g., Spark, Hadoop) | 54% | 38% |
| **Databases** | Database and data warehouse maintenance and development | 64% | 26% |
| | NoSQL (e.g., MongoDB, Redis) | 57% | 22% |
| **Machine learning** | Machine learning (more recent theory, e.g., neural networks, deep learning, support vector machines, boosted trees, etc.) | 22% | 74% |

## 3.2. The Diversity Among Data Scientists

Examining the responses of self-described *data scientists* reveals interesting patterns in the data. In this analysis, we examine the data to see what subclusters of data scientists are apparent. A group of data science unicorns can be spotted through a clustering exercise on the questionnaire responses. They express all sorts of knowledge and skills to be very relevant to their jobs. Notably, at least half of them rated almost all knowledge and skills (except four of them) as at least should- have, as illustrated by their median responses in Figure 2.

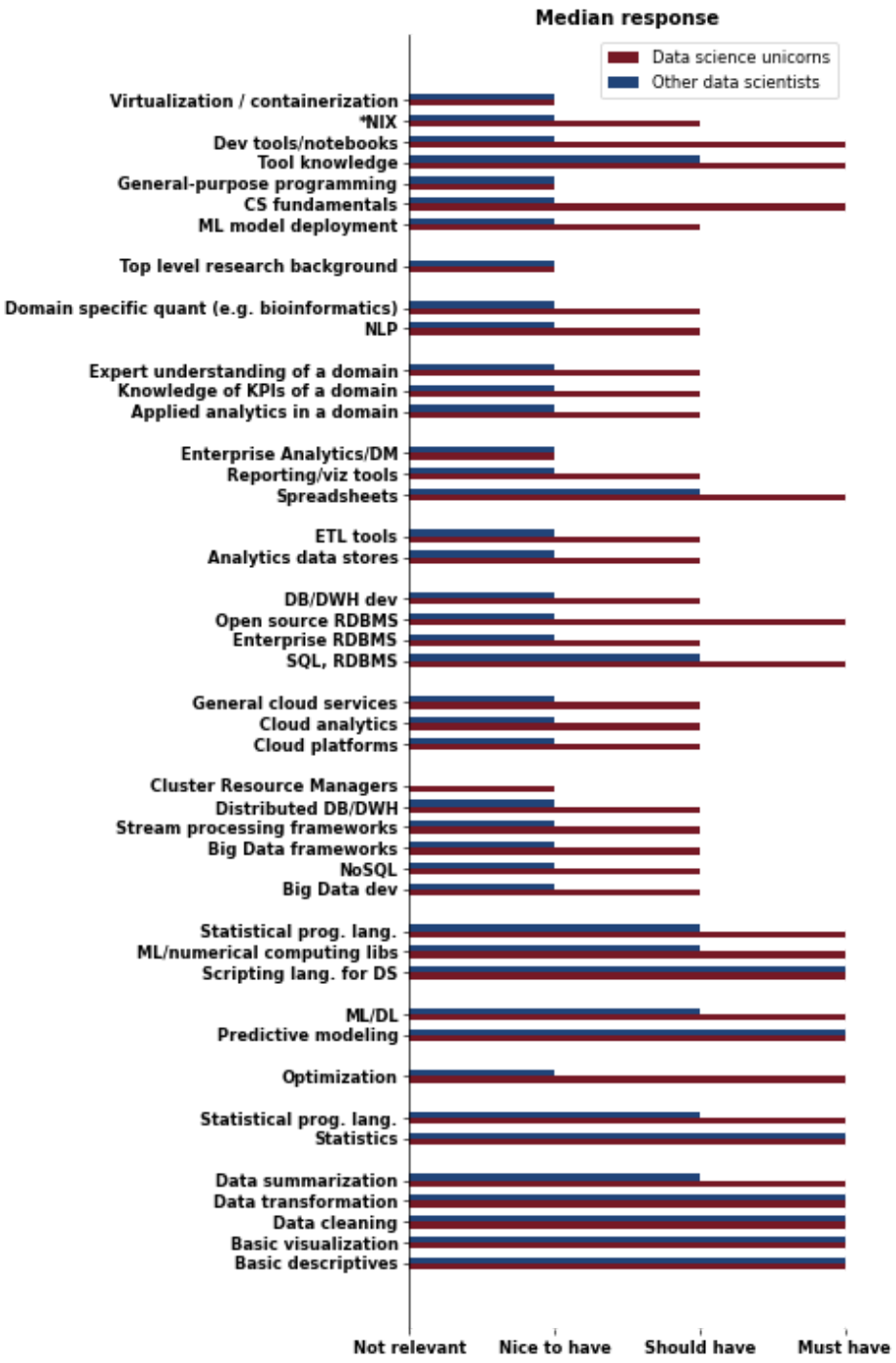



**Figure 2. Median responses of the identified data science unicorns to the questionnaire items in comparison with the rest of self-declared data scientists.** Questionnaire items are grouped for clarity, and they are labeled with short explanations for brevity. See Appendix D for an explination of all abbreviations used.

The individual responses by the identified data science unicorns, along with a summary of their responses collectively, are depicted in Figure 3.

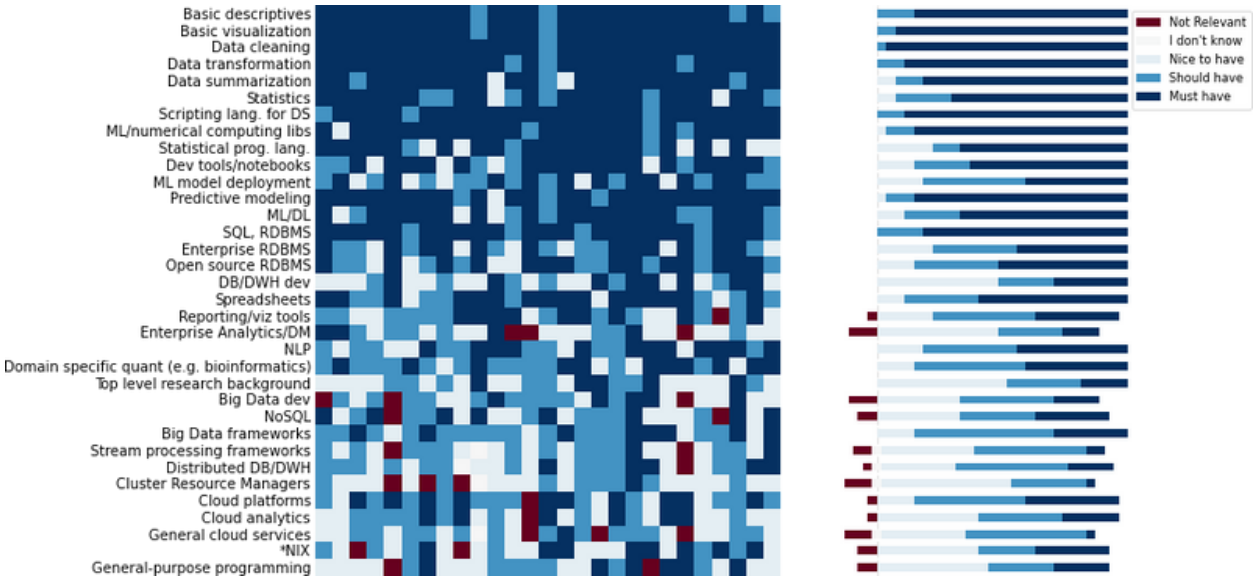



**Figure 3. The listing and a summary of questionnaire responses from the identified data science unicorns.** The left figure represents the set of all members of this subgroup, where a row corresponds to a questionnaire item and a column to an individual survey participant. The cells are color-coded based on the rating by the professional to the corresponding questionnaire item, where red denotes not-relevant, and blues denote nice-to-have, should-have, and must-have (from light to dark). The right figure is a diverging bar chart that aggregates the responses. For each bar, the proportion of professionals who find the corresponding knowledge/skill "not-relevant" are on the left of the origin (which corresponds to "I don't know"), whereas the proportion of professionals stating "nice-to- have," "should-have," or "must-have" are stacked on the right.
See Appendix D for an explination of all abbreviations used.

A diametrically opposing group, roughly the same size, reports much lower expertise in every aspect. Even data preparation and summarization skills (although regarded as relatively more important than other classes of knowledge) are not considered must-haves. In virtually all data science knowledge areas, they declare below-average knowledge requirements. We believe the existence of this group is a result of the liberal usage of the data scientist title, where anybody engaged in any activity related to data science claims the title. When forming a team to mitigate against the rareness of unicorns, we further believe that specialization in at least some areas (including domain or subject matter expertise) is a necessary ingredient. In this regard, this *lower-expertise* group, lacking any specialization, should not be considered 'data scientists' by our standard.

The rest of the data scientists who participated in the questionnaire forms the largest group. They have a more concentrated set of skills compared to data science unicorns and are clearly differentiated from the lower expertise group. In Figure 4, we compare all groups in terms of the proportion of ratings on the relevance of knowledge and skills.

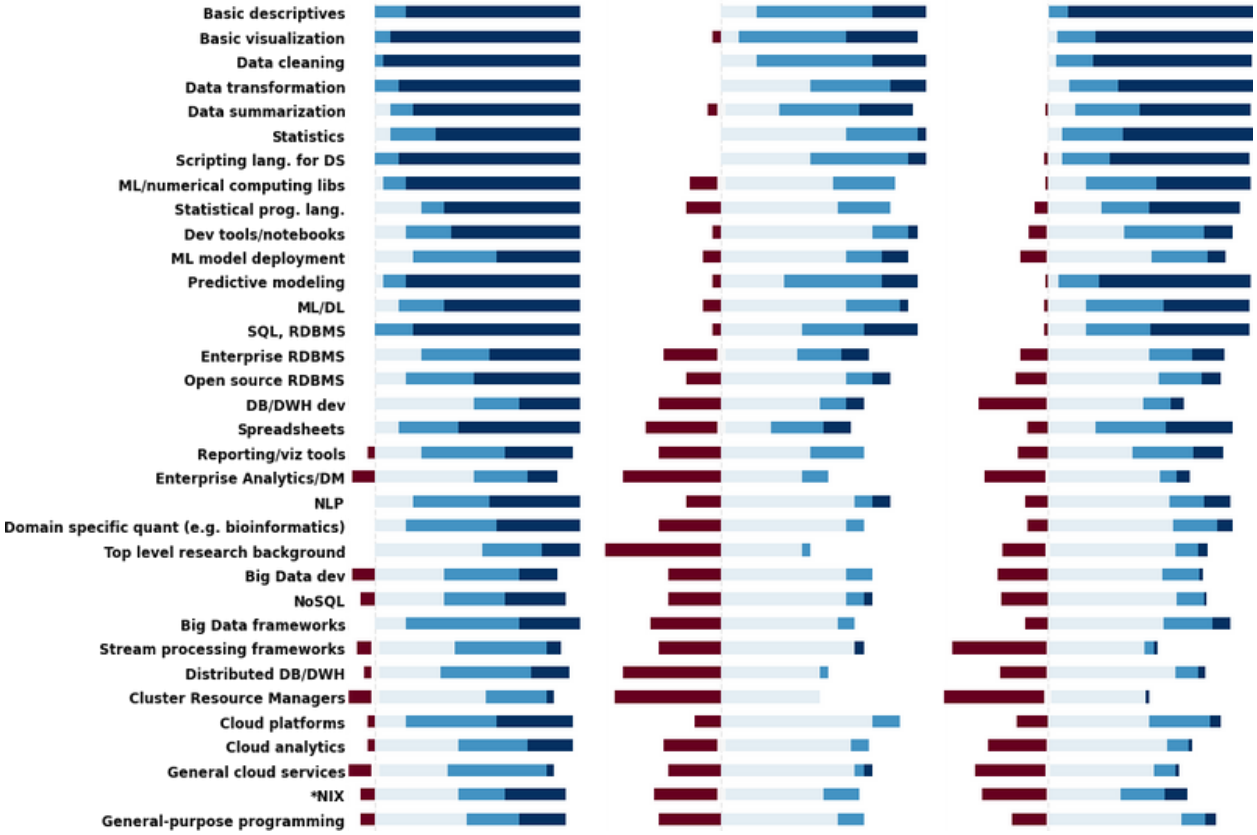



**Figure 4. A summary of the questionnaire responses of the three subgroups of data scientists.** The leftmost block of bars represents the identified data science unicorns. The middle block represents the low-expertise group, whereas the rightmost one represents the larger group, with a more concentrated set of skills than data science unicorns and clearly differentiated from the lower expertise group. The bars were generated in the same way as in Figure 3, and colors should be interpreted similarly. The bars are equi-length across the groups and the questionnaire items.
See Appendix D for an explination of all abbreviations used.

Cluster assignments were inferred through a mixture model on a numeric version of the rating- scale questionnaire data. In Figure 5, how the three groups of data scientists have distinct skill- sets can be seen at a glance where the prevalence of the blue color in the first block points to the many skills attributed to a data science unicorn.

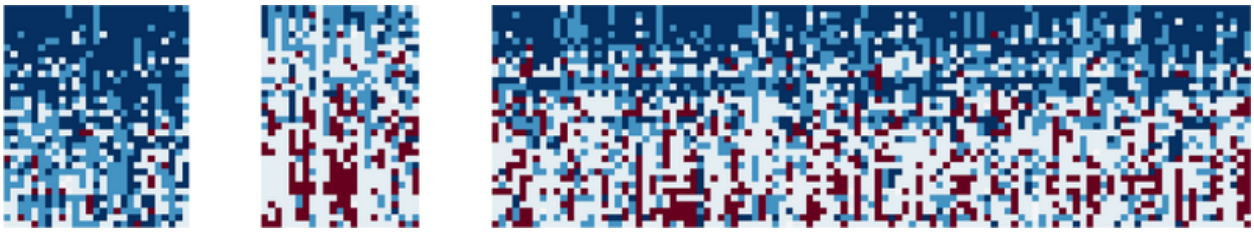


**Figure 5. Questionnaire responses and the three clusters of data scientists.** A block represents a cluster assignment (constructed exactly like that of Figure 3), where the first block is the cluster of data science unicorns. We find the third (and the largest) group more representative of the modern data scientist and completely ignore the lower-expertise group of self-declared data scientists in the second block. The ordering of the rows (i.e., the questionnaire items) is the same as it is in Figures 2 and 3, and the labels are omitted for ease of exposition.

## 4. A Classification of Key Roles

The pairwise comparisons are illustrative of why data analysts and data engineers should be considered key to data science teams---alongside data scientists. However, the approach does not produce a competence framework for the three key roles. In this section, we build job specifications for the three key roles and describe when and in what ways they can be extended. We specify the subjects essential for the three roles, highlight the distinctive technical competence for each role by filtering the essential skills, and specify the roles for which an organization can reasonably expect competence for the subjects that might additionally be found relevant for an organization.

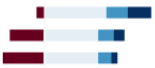
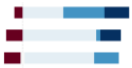
We refer to the knowledge framework for the subjects. The subjects are mapped to the key roles through matching the subject areas and fields in the knowledge framework with the key phrases in LinkedIn professional profiles, the skills listed in LinkedIn job postings, and the questionnaire.

Table 7 summarizes the questionnaire responses for items related to subjects in *science & math* and *programming & technology* domains, respectively.

### Table 7. Summary of the questionnaire.

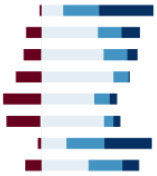
#### a. *Science & math*

| **Knowledge framework** | **Questionnaire summary** | | | | |
|---|---|---|---|---|---|
| **Subject Area** | **Questionnaire items (short)** | **Data Science Unicorns** | **Data Scientists** | **Data Analysts** | **Data Engineers** |
| **Statistics** | Statistics | | | | |

| | | | | | |
|---|---|---|---|---|---|
| **Operations research & Optimization** | Optimization |  |  |  |  |
| **Data preparation & exploration** | Data transformation<br>Data cleaning<br>Basic descriptives<br>Basic visualization |  |  |  |  |
| **Machine learning (ML)** | Predictive modeling<br>ML/DL<br>NLP |  |  |  |  |

## b. Summary of the questionnaire. *Programming & technology.*

| **Knowledge framework** | **Questionnaire summary** | | | | |
|---|---|---|---|---|---|
| **Subject Area** | **Questionnaire items (short)** | **Data Science Unicorns** | **Data Scientists** | **Data Analysts** | **Data Engineers** |
| **General-purpose computing** | General-purpose programming<br>CS fundamentals<br>Virtualization / containerization<br>Cloud platforms |  |  |  |  |

| | | | | | |
|---|---|---|---|---|---|
| **Scientific computing** | Statistical computing languages<br><br>ML/numerical computing libraries<br><br>Development tools/notebooks |  |  |  |  |
| **Database (DB) and business intelligence (BI)** | RDBMS & SQL<br><br>Enterprise RDBMS<br><br>Open-source RDBMS NoSQL<br><br>DB-DWH development<br><br>Enterprise analytics/DM tools<br><br>Insight presentation<br><br>Reporting/viz tools |  |  |  |  |
| **Big data (BD)** | Big data dev<br><br>Cluster resource managers<br><br>Big data frameworks<br><br>Stream processing<br><br>Distributed DB-DWH |  |  |  |  |

*Note.* The bars are vertically aligned at the origin—denoting the response "I don't know." The red color on its left represents "not relevant." The blue colors on the right, from light to darker, represent "nice to have," "should have," and "must have."

To quantify if a subject is relevant for a role, we take the average proportion of the professionals who rate the questionnaire items about a subject in the knowledge framework as a should-have or a must-have.

In LinkedIn professional profiles and job postings, we take the average count of the listed key phrases and skills relevant to the subject, respectively. The key phrases identified as related to the subjects in the knowledge framework and their average frequencies among the professional profiles and job postings for the three main roles are presented in Appendix C. We have two crucial observations in using LinkedIn data, particularly as we compare it to the findings from the questionnaire. The first one is that the skills listed in LinkedIn professional profiles are heavily slanted toward specific tools and technologies relevant to data science, found in our *programming & technology* knowledge areas in the IADSS Knowledge Framework. In other words, knowledge and skills grouped under *science & math* in the IADSS Knowledge Framework, such as fundamental topics in statistics, math, and computer science, are not featured as prominently as these technology-specific skills. The same observation is valid for job postings, which might reflect the assumption that core skills in areas such as math are prerequisites for higher-level requirements. The second observation is that, in most instances, knowledge and skills listed in professional profiles list the entire skillset that the individual has accumulated over their professional lives, not necessarily relevant to their current position in data science. For example, if someone moved into data science from software engineering, we expect to see many software skills listed and should not assume that they are necessarily related to their job in data science.

Despite these distinctions with the survey data, we observe that our exercise using LinkedIn data yields results that are supportive of findings in the survey to a great extent. See Figure 4 for an illustration. We observe, for example, that data analysts differentiate themselves in Querying & presentation and Databases, whereas data engineers do so in Big data pProcessing and Cloud platforms. It is also clear that Machine learning, Statistics, and Statistical computing are skills mainly expected from data scientists.

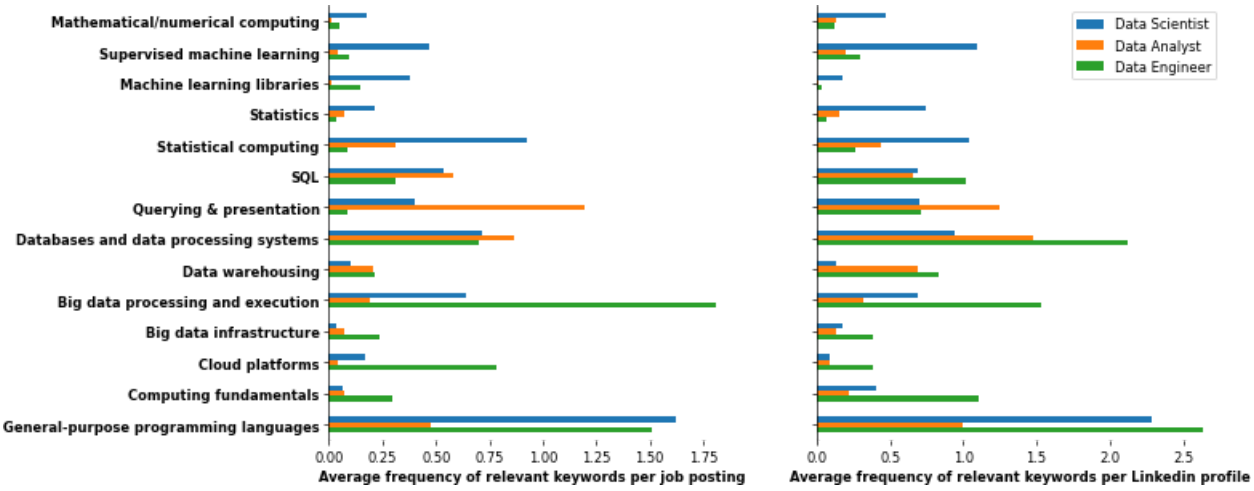



**Figure 6. Average total frequencies of key phrases in LinkedIn profiles (left) and job postings (right), as identified as relevant to selected subjects in the IADSS knowledge framework.** Note that the results mostly align with the questionnaire responses (see, for instance, pairwise comparisons in Tables 4–7).

## 4.1 Essential Technical Knowledge

A subject is included in the essentials list when categorically high rated (relative to the other subjects) in our sources. Some of the essential skills are found to be common. In particular, most participants in the survey agree on the relevance of data preparation and exploration and SQL to their jobs. On the other hand, some of them are found to be distinctive for a role. Examples include querying and presentation tools (e.g., Business Objects, Power BI, QlikView, Qlik Sense, MicroStrategy) for data analysts, machine learning theory and software (e.g., machine learning

libraries such as scikit-learn or TensorFlow, numerical computing libraries such as NumPy) for data scientists, and big data–processing platforms/software for data engineers. Table 8 compiles the list of essential subjects for the three key roles. For completeness, we also include prerequisites to the rest of the list of essentials.

### Table 8. Essential technical knowledge for the key roles.

#### a. Data Scientists

| Subject Area | Subject | Example Topics |
|---|---|---|
| **The scientific method** | The scientific method and problem formulation | Framing a problem and a research question, literature review, research methods |
| **Mathematics** | Calculus and linear algebra | Univariate and multivariate calculus, vectors, matrices, matrix decompositions |
| **Computer science** | Databases & data processing systems | Data and database modeling, online analytical processing (OLAP) |
| **Data preparation & exploration** | Data exploration & visualization | Data exploration strategies, histograms, scatterplots, box plots, etc. |
| | Data Cleaning | Handling missing/inconsistent data, outlier analysis |
| | Data Preparation and Transformation | Normalization, discretization, feature selection, and extraction |

| | | |
|---|---|---|
| **Scientific computing** | Numerical computing libraries | MATLAB, Octave, Julia, GAMS, Netlib, GSL, NumPy, SciPy |
| | Machine learning libraries | TensorFlow, scikit-learn, Keras, PyTorch |
| | Statistical computing | R, SAS, SPSS |
| | Visualization | Matplotlib, ggplot, D3.js |
| **Database and business intelligence** | SQL | |
| **Statistics** | Probability basics | Basic combinatorics, probability distributions, random variables expected values |
| | Descriptive statistics | Measures of central tendency, measures of variability, correlation, contingency tables |
| | Inferential statistics | Hypothesis testing, parameter estimation, design of experiments |
| **Machine learning** | Unsupervised learning | Clustering, association rule mining, dimensionality reduction |
| | Supervised Learning | Regression models, generalized linear models, kernel methods, Gaussian processes, decision trees |

## b. Data Analysts

| Subject Area | Subject | Example Topics |
|---|---|---|
| **The scientific method** | The scientific method and problem formulation | Framing a problem and a research question, literature review, research methods |
| **Mathematics** | Basics, calculus, and linear algebra | Basic arithmetic and algebra, differentiation and integration, vectors, matrices |

| | | |
|---|---|---|
| **Computer science** | Databases & data processing systems | Data and database modeling, OLAP |
| **Data preparation & exploration** | Data exploration & visualization | Data exploration strategies, histograms, scatterplots, box plots, etc. |
| | Data Cleaning | Handling missing/inconsistent data, outlier analysis |
| | Data Preparation and Transformation | Normalization, discretization, feature selection, and extraction |
| **Scientific computing** | Visualization | Matplotlib, ggplot, d3.js |
| **Database and business intelligence** | SQL | |
| | Data warehousing | Intellica, SAP, Teradata, Microsoft |
| | Querying and presentation | Business Objects, Power BI, QlikView, Qlik Sense, MicroStrategy |
| **Statistics** | Probability basics | Basic combinatorics, random variables, probability distributions |
| | Descriptive statistics | Measures of central tendency, measures of variability, correlation, contingency tables |

### c. Data Engineers

| Subject Area | Subject | Example Topics |
|---|---|---|
| **The scientific method** | The scientific method and problem formulation | Framing a problem and a research question, literature review, research methods |

| | | |
|---|---|---|
| **Computer science** | Data structures and algorithms | |
| | Databases & data processing systems | Types of data storage and processing systems, data, and database modeling |
| | Software development | Software development, testing, and maintenance |
| **Data preparation & exploration** | Data transformation | Combining and aggregating data sets, extract- transform-load. |
| | Data exploration & visualization | Data exploration strategies, histograms, scatterplots, box plots, etc. |
| | Data Cleaning | Handling missing/inconsistent data, outlier analysis |
| **Database and business intelligence** | SQL | |
| | Relational database management systems | MS SQL Server, Oracle, Teradata, DB2, Postgres, SQLite, MySQL |
| **Big data** | Big data processing & execution | Hadoop, Spark |
| **General-purpose computing** | Computing fundamentals | *NIX shells, shell scripting, git |
| | General-purpose programming languages | C/C++, C#, Java, Scala, Python |
| | Cloud platforms | AWS, Azure, Google Cloud |

## 4.2. Additional Technical Knowledge

It is natural for an organization to require knowledge beyond the collection of essential subjects. Here, we list the subjects not included as essentials, but the professionals of a particular role are considered or expected to have a reasonably high degree of knowledge in our sources. Take, for instance, text mining or natural language processing. Among the three key roles, knowledge on this subject should be expected from data scientists, not from the analysts or engineers. Similarly, as it lays the foundations of operations research and most modern machine learning algorithms, mathematical optimization is associated with data scientists.

On the other hand, knowledge of operating systems, parallel and distributed computing, big data infrastructure, virtualization/containerization, and NoSQL tools extend the essential skills of a data engineer when necessary

for an organization. Data analysts, too, can be expected to have knowledge in particular relational database management system technology, analyzing big data through querying interfaces, and statistical computing languages. In Table 9, we provide a curated list of additional knowledge that might be required for the key roles.

### Table 9. Additional technical knowledge for the key roles.

### a. Data Scientists

| Subject Area | Subject | Example Topics |
|---|---|---|
| **Operations research & 0ptimization** | Optimization | Model building (objective, constraints), gradient descent, Newton and quasi-Newton methods |
| **Machine learning** | Reinforcement learning | Bandits, Markov decision processes, policy and value iterations, Q-Learning |
| | Deep learning | Deep feedforward networks, autoencoders, deep recurrent networks (Long Short-Term Memory, Gated Recurrent Unit), deep generative models (Variational Autoencoder [VAe], Generative Adversarial Networks [GAN], etc.) |
| | Text mining/NLP | Entity recognition, sentiment analysis, topic modeling, content classification |
| **Scientific computing** | Development environments | Jupyter, RStudio |
| **Statistics** | Bayesian statistics | Bayesian modeling, approximate inference, Markov chain Monte Carlo, model selection |
| | Stochastic Processes, Time Series, Survival Analysis | Random walks, Markov chains, AR/MA processes, ETS, ARIMA, state-space models |
| | Statistical sampling | Sampling methods and biases, sample sufficiency metrics |

### b. Data Analysts

| Subject Area | Subject | Example Topics |
|---|---|---|
| **Database and business intelligence** | Relational database management systems | MS SQL Server, Oracle, Teradata, DB2, SQLite |
| **Scientific computing** | Statistical computing | R, SAS, SPSS |
| | Numerical computing libraries | NumPy, MATLAB |
| | Development environments | RStudio, Jupyter |
| **Statistics** | Inferential statistics | Hypothesis testing, multivariate analysis, parameter estimation |
| | Statistical sampling | Sampling methods and biases, sample sufficiency metrics |
| **Machine learning** | Supervised Learning | Regression models, generalized linear models |
| **Big data** | Big data access | Hive, Pig, Spark SQL |

### c. Data Engineers

| Subject Area | Subject | Example Topics |
|---|---|---|
| **Computer science** | Parallel & distributed computing | Levels of parallelism, parallel and distributed computation models |
| | Operating systems | Memory management, threads, file systems, process scheduling |
| **Databases and business intelligence** | NoSQL/NewSQL | HBase, Cassandra, MongoDB, Redis |
| | Data warehousing | Intellica, SAP, Teradata, Microsoft |
| **Big data** | Big data infrastructure | YARN, HDFS |

| | | |
|---|---|---|
| | Big data integration | Flume, Sqoop, Kafka |
| **General-purpose computing** | Virtualization/containerization | Docker, Kubernetes |

It is crucial to correctly identify if a role can be reasonably expected to have knowledge of these additional subjects. Unreasonable expectations would be harmful to the organizations and frustrating for the professionals. As Kozyrkov (2018b) puts it, for instance, to ask the data analysts to develop their machine learning skills instead of excelling in analytics would result in mediocrity instead of excellence (yet we see 29% of the data analysts in the survey assumes knowledge in machine learning libraries, as presented before). Alternatively, as Anderson (2019) observes, a mistaken belief that data scientists can easily build complex data pipelines would hurt the organizational capabilities. We see some examples of this in job postings on LinkedIn, where big data–processing and execution technology are listed as requirements for data scientists, which we find dissuading potential candidates from applying.

## 4.3. Further Distinctions

As the base set of essential skills can be extended, the three key roles can also act as anchors for other roles. We expect further diversification that will be mainly based on the organization's main activity domain or industry. A professional would possess fundamental characteristics of one of the key roles and extend it with specialized knowledge elemental for their organization. The professional can then be given a new title derived from one of the main roles. As an example, take the "biomedical data scientist" Dunn and Bourne (2017) describe, where professionals can be given the title when they are expected to have fundamental knowledge attributed to a data scientist, and "to obtain basic biomedical knowledge in order to eventually specialize, *e.g.*, in cancer biology."

Another area of diversification is when specifying a newly coined job title. Take, for instance, the machine learning engineer title. Sculley et al. (2015) discuss that a significant fraction of the tasks for real-world machine learning systems are composed of tasks such as configuration, machine resource management, serving infrastructure, and monitoring, most of which we can associate with the data engineering profession. In this regard, one can see the machine learning engineer title converging to or derived from the data engineer.

Further distinctions can be made based on seniority. Executive respondents of the questionnaire (such as *analytics directors/managers*, *data science directors*, *chief data officers*) value general knowledge of a diverse set of theoretical and practical skills. That is to say, the managers, for instance, do not find engineering tools as relevant as data engineers, but they find such tools more relevant than the analysts. As expected, they also find leadership and domain knowledge more relevant.

Finally, one can extend the essentials with skills not included in the knowledge framework, such as common workplace skills. Note that these skills are rated relatively higher by the data analysts in our questionnaire. In

our interviews, too, their role is regarded as an intermediary to the business-specific roles. In that sense, they are the closest representatives of the recently arisen "analytics translator" role described by Henke et al. (2018).

Our characterization for the three key roles serves as a guide for a job specification from the ground up. Figure 7 schematically illustrates the process.

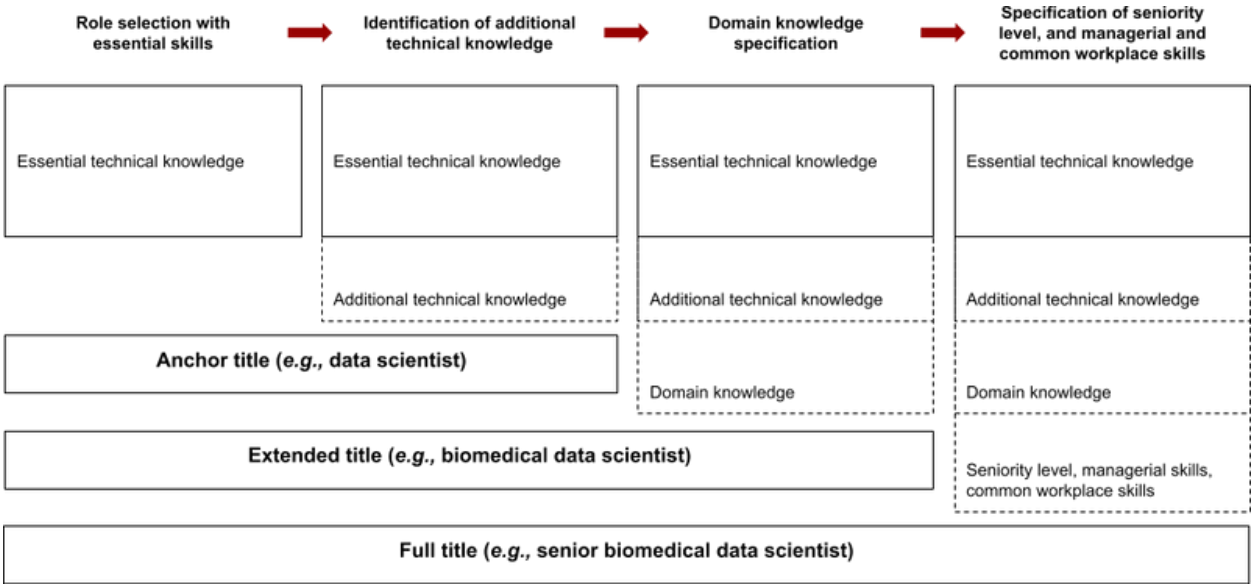



**Figure 7. A guide to the job specification process shown as a sequence of steps.** At each step, the list of expected knowledge and skills from the professional might be extended. The full title originates from one of the three key titles and might be revised to make specific the extended expectations.

## 5. Conclusion

The data science space suffers from confusion emanating from the lack of an agreed-upon classification of job roles. This confusion negatively impacts employers, educational institutions, and practitioners in the field. In prior work, we discussed how expensive this confusion could be in recruiting efforts (economically), but it also has human resources and managerial implications.

This article targets a simplification and an anchor categorization of job roles with clear definitions of roles. We achieve this through analysis of survey results, LinkedIn profiles, job descriptions, and in-depth interviews with managers of data science teams and efforts. We also use our judgment as long-term practitioners and employers of data scientists to provide a practice- guided view of the problem.

Our analysis has led to a simplification into three key roles with complementary skills: data analyst, data scientist, and data engineer. We believe this anchor categorization helps resolve several problems, including:

1. The difficulty of recruiting the right candidates for the right roles

2. Overloading the term ‘data scientist’ to the point where disappointment is inevitable
3. Chasing rare unicorns
4. Lack of structure around evaluating performance

We believe a clear categorization with precise role requirements and expectations can go a long way in creating effective data science teams. It is also an effective input in planning successful data science projects. What is interesting here is that as we combine the skills in these role categories, we converge on the body of knowledge specification from the IADSS Data Science Knowledge Framework (Fayyad & Hamutcu, 2020). We believe this is not a coincidence but further evidence for how we should be thinking about these positions.

Although we realize there will continue to be a variety of titles in data science, and we see new titles emerging such as ‘data translators,’ we believe the industry practitioners can easily adopt the three key roles as anchors. We further believe this simplification is useful not just to hiring managers and executives, but also to educators and aspiring data science professionals. It helps align expectations and serves as a step to tackle the pressing issue of training, evaluating, and building effective data science teams. A simplified (consolidated) categorization will help stakeholders: employers, educational institutions, and practitioners get better clarity on the role of the data scientist. Further subclassification of data scientists could and will evolve, but at least we can get to a better agreement with the core category.

Our work can certainly be extended in a variety of directions, including industry-specific role classifications. We also believe objective, skills-based assessment is essential to any standardization effort to ensure internal and external alignment for an organization and have this topic on our agenda for future work. Another interesting follow-up work would be to take these definitions and compare/contrast them with practices in the industry, for example, how do skill and knowledge requirements in LinkedIn job posts for ‘data analyst’ compare to our standardized definition of this role. Even more interesting would be benchmarking these definitions with existing role definitions at a variety of employers. We hope that we and others can do this follow-up analysis and share findings in future work. We aim to work on creating awareness and driving adoption of our knowledge and standardization framework while continuing our efforts in further understanding and developing the data science profession.

## Disclosure Statement

Usama Fayyad and Hamit Hamutcu have no financial or non-financial disclosures to share for this article.

## References

Anderson, J. (2018, April 11). *Data engineers vs. data scientists*. O’Reilly Media. https://www.oreilly.com/radar/data-engineers-vs-data-scientists/

Anderson, J. (2019, April 9). *Why a data scientist is not a data engineer*. O'Reilly Media. https://www.oreilly.com/content/why-a-data-scientist-is-not-a-data-engineer/

Berthold, M. R. (2019). What does it take to be a successful data scientist? *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.e0eaabfc

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, *1*(1), 121–143. https://doi.org/10.1214/06-BA104

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

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

Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, *101*(12), 4164–4169.

Chatfield, A. Takeoka., Shlemoon, V. Najem., Redublado, W. & Rahman, F. (2014). Data scientists as game changers in big data environments. Proceedings of the 25th Australasian Conference on Information Systems (pp. 1-11). New Zealand: Auckland University of Technology.

Davenport, T. (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, DJ. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, *90*(10), 70–76. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2017). 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

Demchenko, Y., & Cuadrado-Gallego, J. J. (2020). Data science professional profiles. In J. J. Cuadrado-Gallego & Y. Demchenko (Eds.), *The data science framework* (pp. 109–134). Springer International Publishing. https://doi.org/10.1007/978-3-030-51023-7_5

Dunn, M. C., & Bourne, P. E. (2017). Building the biomedical data science workforce. *PLOS Biology*, *15*(7), Article e2003082. https://doi.org/10.1371/journal.pbio.2003082

Fayyad, U., & Hamutcu, H. (2020). Toward foundations for data science and analytics: A knowledge framework for professional standards. *Harvard Data Science Review*, *2*(2). https://doi.org/10.1162/99608f92.1a99e67a

Garten, Y. (2018, November 6). The kinds of data scientist. *Harvard Business Review*. https://hbr.org/2018/11/the-kinds-of-data-scientist

Henke, N., Levine, J., & McInerney, P. (2018, February 5). You don't have to be a data scientist to fill this must-have analytics role. *Harvard Business Review*. https://hbr.org/2018/02/you-dont-have-to-be-a-data-scientist-to-fill-this-must-have-analytics-role

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

Jung, J. (2020, August 19). *Machine learning engineer vs data scientist (Is data science over?)*. Towards Data Science. https://towardsdatascience.com/mlevsds-3c89425baabb

Kozyrkov, C. (2018a, July 27). *Top 10 roles in AI and data science.* Hacker Noon. https://hackernoon.com/top-10-roles-for-your-data-science-team-e7f05d90d961

Kozyrkov, C. (2018b, December 4). What great data analysts do—And why every organization needs them. *Harvard Business Review*. https://hbr.org/2018/12/what-great-data-analysts-do-and-why-every-organization-needs-them

Loukides, M., & Lorica, B. (2017, June 6). *What are machine learning engineers?* O'Reilly

Media. https://www.oreilly.com/radar/what-are-machine-learning-engineers/

Mayank, M. (2017). *7 Types of data scientist job profiles*. KDnuggets. Retrieved November 27, 2020, from https://www.kdnuggets.com/7-types-of-data-scientist-job-profiles.html/

Miller, S. (2014). Collaborative approaches needed to close the big data skills gap. *Journal of Organization Design*, *3*(1), 26–30.

Murphy, S., Vaisman, M., & Harris, H. (2013). *Analyzing the analyzers*. O'Reilly Media.

Nelson, G., & Horvath, M. (2017). *The elusive data scientist: Real-world analytic competencies*. https://support.sas.com/resources/papers/proceedings17/0832-2017.pdf

Nelson, S. (2018, January 27). *4 types of data science jobs*. Udacity. https://blog.udacity.com/2018/01/4-types-data-science-jobs.html

Olavsrud, T. (2015, December 3). *Don't look for unicorns, build a data science team*. CIO. https://www.cio.com/article/3011648/dont-look-for-unicorns-build-a-data-science-team.html

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

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Press, G. (2012, September 27). Data scientists: The definition of sexy. *Forbes*. https://www.forbes.com/sites/gilpress/2012/09/27/data-scientists-the-definition-of-sexy/

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Proceedings of the 28th International Conference on Neural Information Processing Systems–Volume 2* (pp. 2503–2511). NIPS.

Viaene, S. (2013). Data scientists aren't domain experts. *IT Professional*, *15*(6), 12–17. https://doi.org/10.1109/MITP.2013.93

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

Woods, D. (2011, November 27). LinkedIn's Monica Rogati on "What is a data scientist?" *Forbes*. https://www.forbes.com/sites/danwoods/2011/11/27/linkedins-monica-rogati-on-what-is-a-data-scientist/

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

# Appendices

## Appendix A. Full Set of Survey Skill and Knowledge Area Questions

The list of skill/knowledge areas that were included in the survey is presented below. All questions had the following answer choices for the corresponding role: *not relevant*, *nice to have*, *should have*, *must have*, *I don't know*.

| | |
|---|---|
| 1 | Generating and interpreting basic statistical descriptions of data (e.g., means, medians, measures of variance and range, quantiles, etc.) |

| | |
|---|---|
| 2 | Generating and interpreting basic visual descriptions of data (histograms, density plots, time-series visualization, etc.) |
| 3 | Data cleaning (outlier detection, de-duplication, identifying data value conflicts, imputing missing values) |
| 4 | Data transformation/reduction (discretization/binning, sampling, normalization/scaling, engineering new variables) |
| 5 | Data summarization (e.g., through data aggregation, querying, and business intelligence tools) |
| 6 | Statistics (e.g., hypothesis testing, regression, probability distributions, estimation, etc.) |
| 7 | Optimization (linear programming, numerical optimization, genetic algorithms etc.) |
| 8 | Predictive modeling (decision trees, linear models for regression and classification, model validation, etc.) |
| 9 | Machine learning (more recent theory, e.g., neural networks, deep learning, support vector machines [SVM] boosted trees, etc.) |
| 10 | Natural language processing/text mining |
| 11 | Other domain-specific quantitative skills (time series analysis, image, audio, signal processing, bioinformatics, etc.) |
| 12 | Top-level research background in a specific scientific field—the ability to contribute novel algorithms, models to top conferences |
| 13 | Software engineering skills (e.g., embedding a model as a feature into a production IT system) |
| 14 | Big data development and maintenance for infrastructure and applications |
| 15 | Database and data warehouse (DWH) development and maintenance |

| | |
|---|---|
| 16 | Computer science fundamentals (e.g., data structures, algorithms, computer architecture, etc.) |
| 17 | Development and maintenance of analytics-oriented data stores (e.g., OLAP) |
| 18 | Insight generation—building a data-driven business narrative (ROI, financial model, sensitivity, and scenario analyses) |
| 19 | Insight presentation—building a data-driven story line and present to an executive/client audience |
| 20 | Cross-functional collaboration—with internal/external customers and providers to build solutions and drive adoption |
| 21 | Self-sufficiency and task management—understanding requirements and delivering solutions independently, managing the process end-to-end |
| 22 | Project management—leading a cross-functional project |
| 23 | Peer leadership—leading, coaching, and developing others on the same team |
| 24 | Executive leadership—leading the data & analytics agenda at the CxO level |
| 25 | Experience applying analytics in a specific domain/industry (e.g., credit risk, customer relationship management, human resources, etc.) |
| 26 | Knowledge of standard terminology and KPIs of a specific domain (e.g., natural language processing for banking, click-through rate [CTR] for internet, etc.) |
| 27 | Expert understanding of a business domain/industry—i.e., can easily switch to a non- data-science role in the organization |
| 28 | How important is it for the candidate to have experience on a specific tool (programming language, database etc.) |
| 29 | Scripting languages for data science (e.g., Python, Scala) |

| | |
|---|---|
| 30 | Statistical programming languages (e.g., R, SAS Language, MATLAB/Octave, etc.) |
| 31 | General-purpose programming languages (e.g., Java, C/C++, C#) |
| 32 | Libraries for machine learning / numerical computing (e.g., TensorFlow, scikit-learn, NumPy, SciPy, pandas, Theano, Torch, Keras) |
| 33 | Developer tools (specific version control systems, CI, IDE, etc.) (e.g., Git, Jenkins, Jupyter, Zeppelin) |
| 34 | SQL, understanding of relational databases (e.g., SQL) |
| 35 | Enterprise relational DBMS (e.g., Oracle (PL/SQL), Teradata, DB2, SQL Server, Vertica) |
| 36 | NoSQL (e.g., MongoDB, Redis) |
| 37 | Open-source relational database management systems [DBMS] (e.g., MySQL, Postgres) |
| 38 | Integration, ETL, Automation, Design tools (e.g., Informatica, SSIS, Talend, SSRS, Erwin) |
| 39 | General knowledge of distributed storage/computing frameworks (e.g., Spark, Hadoop) |
| 40 | Knowledge of specific (e.g., stream processing) frameworks or services (e.g., Storm, Flink, Flume) |
| 41 | Distributed databases, data warehousing, and storage (e.g., Hive/Pig, HBase, Cassandra, Parquet) |
| 42 | Cluster resource managers, other big data technologies (e.g., Mesos, YARN) |
| 43 | Enterprise advanced analytics/data mining software (e.g., SAS Enterprise Guide, SAS Enterprise Miner, SPSS Modeler, Alteryx, KNIME, RapidMiner) |

| | |
|---|---|
| 44 | Reporting and visualization software (e.g., Tableau, QlikView, Power BI, TIBCO Spotfire, MicroStrategy, Hyperion, Cognos, Business Objects) |
| 45 | Office productivity, spreadsheet tools (e.g., MS Excel, Google Sheets, MS Access) |
| 46 | Experience with cloud computing platforms (e.g., AWS, MS Azure, Google Cloud Platform) |
| 47 | Cloud-based data warehousing (e.g., Amazon Redshift, Big Query) |
| 48 | Other specific cloud-computing services (e.g., EMR, S3, Elasticsearch, Lambda) |
| 49 | Specific knowledge of *NIX systems, shell scripting (e.g., Linux, UNIX, POSIX shells, Perl, etc.) |
| 50 | Virtualization, containerization, etc. (e.g., Docker, Kubernetes) |

## Appendix B. Full List of LinkedIn Keywords and Knowledge Areas

Table 3 in the main text includes a sample of key phrases extracted from LinkedIn professional profiles and job postings. In Table B1, we list the complete set of key phrases extracted from LinkedIn profiles and job postings.

### Table B1. LinkedIn key phrases.

### a. *Programming & technology* subjects in the IADSS Knowledge Framework

| Knowledge framework subject area | Subjects | Frequent key phrases |
|---|---|---|
| ***General-purpose computing*** | General-purpose programming languages | Visual basic, Scala, Ruby, Python, PHP, Perl, JavaScript, Java, Fortran, C#, C, C++, .NET |
| | Computing fundamentals | UNIX, Linux, bash, shell scripting, git, DevOps, subversion |
| | Virtualization/Containeriz ation | Docker, Kubernetes, virtualization |

| | | |
|---|---|---|
| | Cloud platforms | Cloud computing, Microsoft Azure, Amazon web services, Google Cloud, Amazon EC2 |
| ***Scientific computing*** | Statistical computing | Stata, SPSS, SAS, R |
| | Mathematical/numerical computing | SciPy, NumPy, MATLAB, Mathematica |
| | ML libraries | TensorFlow, scikit-learn, Keras, Theano, Spark MLlib, PyTorch, MXNet, Caffe |
| | Development environments | Jupyter, Apache Zeppelin |
| ***Database (DB) and business intelligence (BI)*** | RDBMS & SQL | SQL, T-SQL, PL/SQL, SQL tuning, relational databases, Oracle SQL, Teradata, PostgreSQL, Oracle database, MySQL, Microsoft SQL Server, Microsoft Access, DB2 |
| | NoSQL/NewSQL | SAP Hana, MongoDB, Apache HBase, Elasticsearch, Cassandra, NoSQL |
| | Data warehousing | Data warehouse architecture, data warehousing, Netezza, SAP Business Warehouse, SAP R/3, Microsoft SQL Server Analysis Services, Oracle, Teradata |
| | Querying & presentation | Tableau, SQL Server Reporting Services, SAP BusinessObjects, SAP BI, reporting & analysis, QlikView, Qlik, Microsoft Power BI, pandas, Oracle Business Intelligence Suite, MicroStrategy, Microsoft Excel, dashboard, Crystal Reports, Cognos, Business Objects, Spotfire, Looker |

| | | |
|---|---|---|
| ***Big data*** | Big data processing and execution | Apache Spark, Apache Hadoop, MapReduce, Apache Storm, PySpark, Apache Flink |
| | Big data access | Big data analytics, Apache Hive, Apache Pig, Spark SQL, Athena, Impala, Presto, |
| | Big data integration | Apache Sqoop, Apache Oozie, Apache Kafka, Apache Flume |

## b. *Science & math* subjects in the IADSS Knowledge Framework

| **Knowledge framework subject area** | **Generic keywords found for the subject area** | **Subjects** | **Frequent key phrases for the subjects** |
|---|---|---|---|

| | | | |
|---|---|---|---|
| ***Computer science*** | | Data structures & algorithms | Data structures |
| | | Databases and data processing systems | SQL, NoSQL, databases, relational databases, database design, database administration, data warehousing, data warehouse architecture, storage, OLAP, data marts, data processing, databases, data architecture, data modeling, data management |
| | | Software engineering & development | Test automation, testing, solution architecture, software project management, software engineering, software documentation, software development lifecycle, software development, software design, agile, scrum, requirements gathering, requirements analysis, object-oriented programming, object-oriented design |
| | | Operating systems | Linux |
| | | Parallel and distributed computing | MapReduce, parallel computing, high-performance computing, distributed systems |

| | | | |
|---|---|---|---|
| ***Statistics*** | Statistics, statistical modeling, statistical analysis | Probability basics | |
| | | Descriptive statistics | |
| | | Inferential statistics | Regression, regression analysis, multivariate statistics |
| | | Bayesian statistics | Bayesian statistics |
| | | Stochastic processes | Forecasting, time-series analysis |
| | | Statistical sampling | Monte Carlo simulation |
| ***Operations research & optimization*** | Optimization, operations research | Linear programming | |
| | | Non-linear optimization | |
| ***Data preparation & exploration*** | | Data preparation and transformation | ETL, transform, data migration |
| | | Data cleaning | Data quality |
| | | Data exploration & visualization | Data profiling, data visualization, visualization |

| | | | |
|---|---|---|---|
| ***Machine learning (ML)*** | Pattern recognition, machine learning, machine learning algorithms | Supervised learning | Predictive modeling, predictive analytics, regression, logistic regression, linear regression, decision trees, convolutional neural networks, recommender systems |
| | | Unsupervised learning | Cluster analysis |
| | | Reinforcement learning | |
| | | Deep learning (DL) | Artificial neural networks, deep learning, convolutional neural networks |
| | | Text mining and natural language processing (NLP) | NLP, text mining, text analytics, information retrieval |

The average frequencies of these key phrases base our subject classification for the roles. In Figure B1, we show the average frequencies of the subject-related key phrases in LinkedIn job postings and professional profiles.

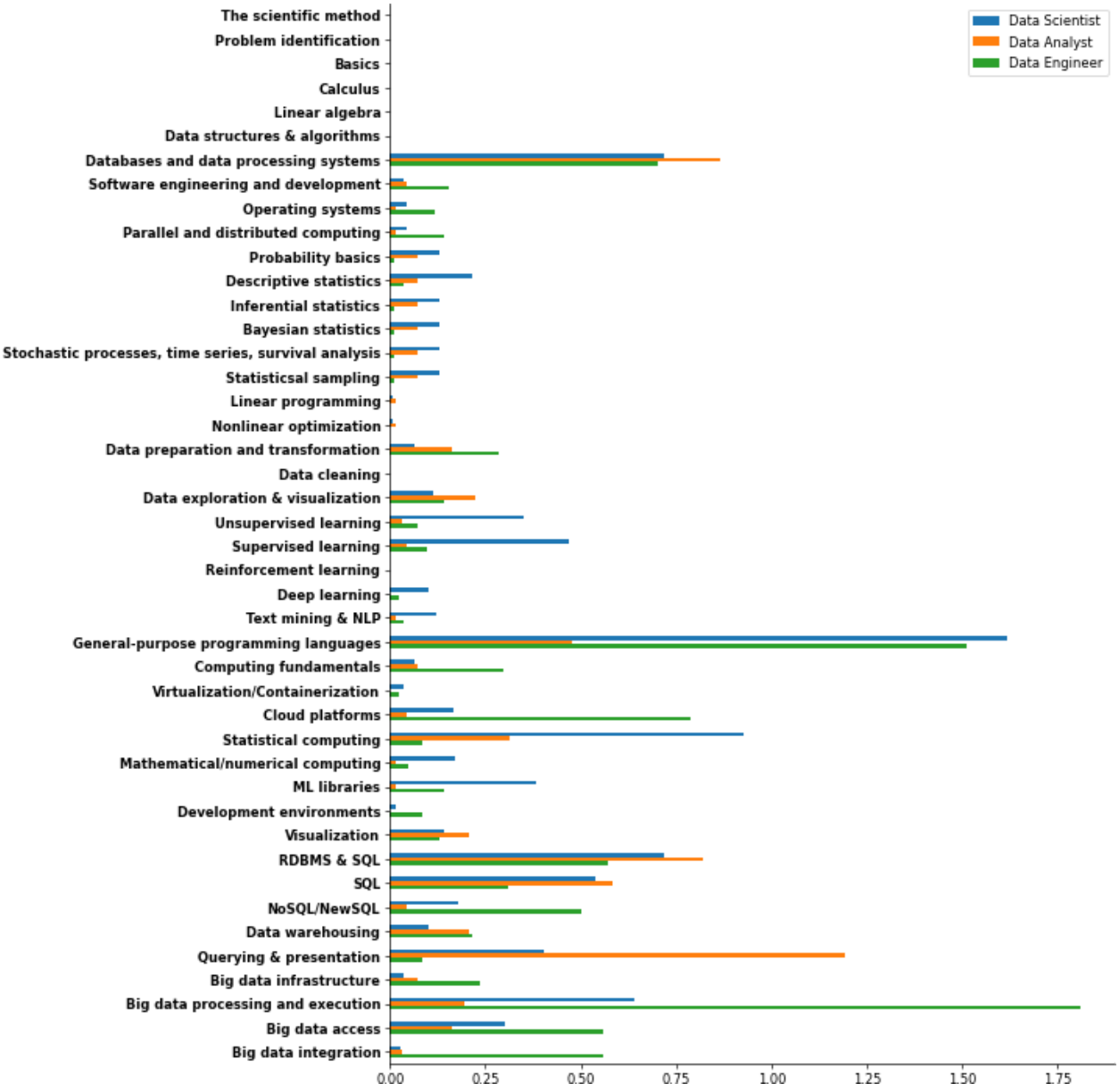



**Figure B1a.** The average frequency of key phrases found in LinkedIn job postings.

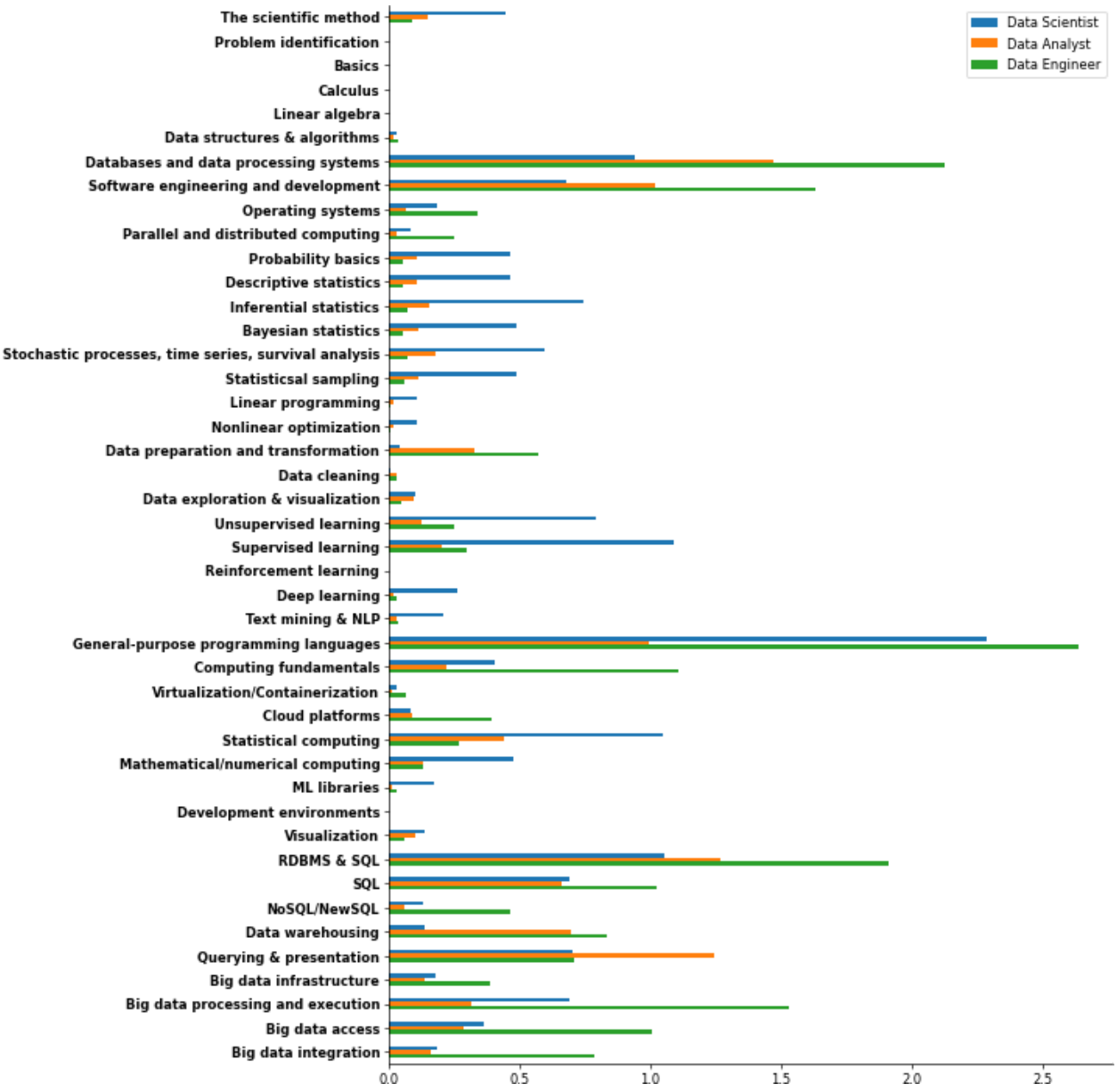



**Figure B1b.** The average frequency of key phrases found in LinkedIn professional profiles.

**Figure B1.** Average frequency of key phrases (grouped by subjects based on the assignments in Table 1) in LinkedIn job postings (a) and professional profiles (b).

## Appendix C. Details on the Clustering Exercise

In Section 3.2 of the main text, we have discussed three clusters of data scientists who participated in our questionnaire. Here, we provide details on the clustering exercise that resulted in the subgroups of data scientists mentioned there.

We note that the responses are in a rating scale, where the participants provide one of the following five responses: *not relevant*, *I don't know*, *nice to have*, *should have*, *must have*, where the response "I don't know"

is very rare.

The statistics and all comparisons provided in the main text are based on proportions of different rating levels or median responses. However, for the clustering exercise, the rating levels were converted to numerical (from 0 to 4, respectively), and standardization was applied for the set of responses to each question. Then a Gaussian mixture model was fit, where each cluster is assumed to have its own covariance matrix. The number of clusters was identified with a separate experiment where an infinite mixture model was assumed with a Dirichlet process prior (Blei & Jordan, 2006), followed by our verification. We used scikit-learn software library (Pedregosa et al., 2011) for clustering. That said, we are aware that a Gaussian mixture model assumption might raise questions. Indeed, we tried several other common models and algorithms for clustering and got similar results.

## Appendix D. Figure Abbreviations

**Figure 2. Median responses of the identified data science unicorns to the questionnaire items in comparison with the rest of self-declared data scientists.**

- Big Data dev = Big Data Development
- CS Fundamentals = Computer Science Fundamentals
- DB/DWH dev = Database/Datawarehouse development
- Dev tools/notebooks = Development tools/notebooks
- Enterprise RDBMS = Enterprise Relational Database Management System
- Enterprise Analytics/DM = Enterprise Analytics/Data Mining
- ETL tools = Extract, Transform, Load tools
- Knowledge of KPIs of a domain = Knowledge of Key Performance Indicators of a domain
- ML Model Deployment = Machine Learning Model Deployment
- ML/ numerical computing libs = Machine Learning/numerical computing libraries
- ML/DL = Machine Learning/Deep Learning
- NLP = Natural Language Processing
- NoSQL = Non-relational Structured Query Language
- Open source RDBMS = Open source Relational Database Management System
- Reporting/viz tools = Reporting/visualization tools
- Scripting lang. For DS = Scripting language for data science
- SQL = Structured Query Language
- Statistical prog. lang. = Statistical programming language

**Figure 3. The listing and a summary of questionnaire responses from the identified data science unicorns.**

- Big Data dev = Big Data Development
- DB/DWH dev = Database/Datawarehouse development

- Dev tools/notebooks = Development tools/notebooks
- Enterprice RDBMS = Enterprice Relational Database Management System
- Enterprise Analytics/DM = Enterprise Analytics/Data Mining
- ML/numerical computing libs = Machine Learning/numerical computing libraries
- ML Model Deployment = Machine Learning Model Deployment
- ML/DL = Machine Learning/Deep Learning
- NLP = Natural Language Processing
- NoSQL = Non-relational Structured Query Language
- Open source RDBMS = Open source Relational Database Management System
- Reporting/viz tools = Reporting/visualization tools
- Scripting lang. For DS = Scripting language for Data Science
- Scripting lang. For DS = Scripting language for data science
- SQL/RBMDS = Structured Query Language/Relational Database Management System

**Figure 4. A summary of the questionnaire responses of the three subgroups of data scientists.**

- Big Data dev = Big Data Development
- DB/DWH dev = Database/Datawarehouse development
- Dev tools/notebooks = Development tools/notebooks
- Dev tools/notebooks = Development tools/notebooks
- Distributed DB/DWH = Distributed Database/Datawarehouse
- Enterprise RDBMS = Enterprise Relational Database Management System
- Enterprise Analytics/DM = Enterprise Analytics/Data Mining
- ML Model Deployment = Machine Learning Model Deployment
- ML/numerical computing libs = Machine Learning/numerical computing libraries
- ML/DL = Machine Learning/Deep Learning
- NLP = Natural Language Processing
- NoSQL = Non-relational Structured Query Language
- Open source RDBMS = Open source Relational Database Management System
- Reporting/viz tools = Reporting/visualization tools
- Scripting lang. For DS = Scripting language for Data Science
- SQL/RBMDS - Structured Query Language/Relational Database Management System
- Statistical prog. lang. = Statistical programming language

---

## Footnotes

1. More information on the effort can be found at www.iadss.org. ↩

2. Researchers can request access to the survey data set through the contact form at www.iadss.org. ↩

3. According to "silhouette scores" (Rousseeuw, 1987) calculated from individuals' responses to the questionnaire. ↩