The Ambiguity of Data Science Team Roles and the Need for a Data Science Workforce Framework

Jeffrey S. Saltz
Syracuse University
Syracuse, NY
jsaltz@syr.edu

Nancy W. Grady
Advanced Analytics, SAIC
Oak Ridge, TN
nancy.w.grady@saic.com

***Abstract*—This paper first reviews the benefits of well-defined roles and then discusses the current lack of standardized roles within the data science community, perhaps due to the newness of the field. Specifically, the paper reports on five case studies exploring five different attempts to define a standard set of roles. These case studies explore the usage of roles from an industry perspective as well as from national standard big data committee efforts. The paper then leverages the results of these case studies to explore the use of data science roles within online job postings. While some roles appeared frequently, such as *data scientist* and *data engineer*, no role was consistently used across all five case studies. Hence, the paper concludes by noting the need to create a data science workforce framework that could be used by students, employers, and academic institutions. This framework would enable organizations to staff their data science teams more accurately with the desired skillsets.**

Keywords-big data; data science; project management; data science roles

I. INTRODUCTION

Big data represents a significant change in the techniques and technologies used for data-intensive computing. The move to parallelization has added complexity to big data solutions and introduced the need for several specialized skills. Related, the term "data science" has become ubiquitous, used to describe any activities that touch data. Consequently, it is difficult to ascertain what skills are needed to perform the specific tasks required to build and deploy big data analytics (BDA) systems. This is compounded by the fact that the field is evolving from work performed by an individual that does data science to a team that does data science [1]. Within this new context, we lack a vocabulary to describe the roles and skills required for an effective data science/big data team. This lack of vocabulary creates many issues (e.g., identifying the appropriate person that should be hired for a specific role within a data science team). To address this challenge, this paper frames and provides data science workforce definitions with examples.

Section II provides some background and motivation for the development of standard skill categories and definitions. Section III provides several case studies in the use of different job titles. Section IV provides a comparison between the case studies as well as our review of the usage of these roles within online job postings. Section V describes future trends that will affect the data science workforce. Finally, Section VI presents our conclusions and next steps.

II. BACKGROUND

The paradigm shift known as big data occurred in the mid-2000s with the advent of new techniques for large data file storage (Hadoop Distributed File System [HDFS]), physically distributed logical datasets (Hadoop), and parallel processing on the distributed data (MapReduce). The Hadoop ecosystem relies on horizontal scaling to distribute data across independent nodes, with scalability arising from the addition of nodes. The result is the use of parallelism for scalable, data-intensive application development. While it has been given a number of differing conceptual definitions [2], big data is not "bigger data" than common techniques can handle but rather data that requires parallel processing to meet the time constraints for end-to-end analysis performance at an affordable cost.

The role of *data scientist* is often assigned to anyone who performs any activity that touches data, including data management, data processing systems, data analytics, and so on. However, the skills required for these different tasks vary greatly.

As data science grows in usage, and teams grow in size, specialization within the team naturally occurs. For example, data science teams often have people that focus on analytics (often called data scientists) and others that focus on collecting/cleaning data (often known as data engineers). In reality, many specializations are for "vertical" subject matter experts, such as data architects, big data engineers, data analysts, or machine learning experts. Being a "horizontal" data scientist refers to one having general expertise in several disciplines sufficient to guide the work of a diverse team of specialists. While these roles are starting to be commonly used, little has been published on them. In fact, in a literature review on team data science processes, the concept of roles was not identified [3].

A. The Need for Workforce Descriptions

The main motivation for workforce descriptions is the need to identify, recruit, train, develop, and maintain an appropriately skilled workforce by providing a common language to categorize and describe the type of data science work that needs to be done.

While the roles and vocabulary are data science specific, the need for workforce descriptions is not limited to data science. In other words, data science is not the only discipline that has required clarification of roles and skills. For example, cybersecurity is another domain where this

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

need has existed. For data security, the U.S. National Institute for Science and Technology (NIST) developed the National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework [4] that clarifies the categories, specialty areas, and work roles for cybersecurity practitioners. In addition, they provided lists of tasks, knowledge, skills, and ability descriptions, mapping them to work roles. Another effort for to provide workforce definitions was the U.S. Department of Defense Cyber Workforce Framework [5]. This work is ongoing, including revisions to a companion document A Role-Based Model for Federal Information Technology/Cyber Security Training [6].

The benefits listed in the NICE report apply equally well to the domain of data science and include the following:

- ***Employers***–track staff skills, training, and qualifications; improve position descriptions; develop career paths; and analyze proficiency
- ***Educators***–develop curriculum and conduct training for programs, courses, and seminars for specific roles
- ***Technology Providers***–identify work roles, tasks, and knowledge, skills, and abilities associated with their products.

Of course, this work would also be of value for students (in understanding how their education maps to different possible roles) and employees (understanding roles where their skills can be leveraged most effectively). Hence, providing job titles and job descriptions that more clearly identify tasks, knowledge, skills, and abilities would benefit the data science community and remove the overloading of the term data scientist.

B. Skills vs Roles

While there are some skills in common across different types of data science roles, some skills might be specific to a particular role. Just as the NICE workforce framework has knowledge, skills, and abilities that can apply to multiple work roles, it will be important to ensure that each data science work role is similarly described. More generalist practitioners would be able to fit into a number of roles, but non-overlapping role descriptions are important to ensure clarity. In addition, skills might vary based on the type of data science project, such as projects with more or less discovery required within the analysis [7].

C. Challenge Due to Lack of Process Model

There are several challenges related to the development of a data science workforce framework but perhaps the most significant one is that there is not an agreed upon process model for data science [1, 8]. In the late 1990s, the Cross-Industry Standard Process model for Data Mining (CRISP-DM) [9] was developed by a consortium to resolve the conflicts between individual data mining process models, to promote communication in the discipline, and to ensure greater data mining success. This model is still the framework followed by the largest number of practitioners [10], but it predated cloud, big data, machine learning, agile, "the Internet of Things", and so on, and it did not consider system development or management processes [8].

D. In Relation to Software Development Lifecycles

Most analytic system development results in situational awareness through reports or business intelligence. Software development lifecycles (SDLCs) are geared toward this kind of requirements-driven analytics system development. Advanced analytics systems are, however, outcomes-driven and require experimentation for the choice of data and data features, model building, and model evaluation and optimization. Some data science skillsets would overlap those used in SDLCs, but for completeness, roles would need to be distinctly described. Care should be taken, however, to align the data science and SDLC models, in particular lining them up with agile and DevOps standard methodologies.

III. Methodology

To understand how roles are currently defined within the field of data science, qualitative case studies were performed on selected organizations that defined roles for teams working on data science projects. To help ensure that a representative view of the current thinking and usage across the field of data science was captured, a cross section of organizations were explored. Specifically, two standards bodies, two organizations from industry, and one consultanting/advisory firm were selected for our case studies.

For each case study, the analysis of the defined roles were based on written documention that was collected from each organization. Where documentation was not as robust, discussions were also held with individuals from the identified organizations.

IV. Case Studies

In this section, we explore, via a series of case studies, some of the roles used/defined within different organizations. We first explore the work that has been done to date by standards organizations (NIST and EDISON). We then explore framework used by two industry organizations (SAIC and Springboard). Finally, we explore an advisory company's view (Gartner). The goal of these case studies is to explore the commonality and diversity of the vocabulary used to describe roles within data science teams.

A. NIST

To advance progress in big data, the NIST Big Data Public Working Group (NBD-PWG) aims to develop consensus on important, fundamental concepts related to big data. The results are reported in the *NIST Big Data Interoperability Framework* series of volumes [11].

One of the NBD-PWG's core activities is to develop a big data reference architecture (RA) that categorizes the components of big data systems. It also describes, at a high level, the roles of those whose tasks are contained in that component. The RA consists of five components and identifies their respective roles, as shown in Fig. 1 and as follows:

- ***System Orchestrator***–defines and integrates the required data application activities into an operational vertical system.
- ***Data Provider***–introduces new data or information feeds into the big data system.
- ***Big Data Application Provider***–executes a life cycle to meet security and privacy requirements as well as system orchestrator-defined requirements.
- ***Big Data Framework Provider***–establishes a computing framework in which to execute certain transformation applications while protecting the privacy and integrity of data.
- ***Data Consumer***–includes end users or other systems who use the results of the big data application provider.
- ***Security and Privacy***–interacts with the system orchestrator for policy, requirements, and auditing, and with both the big data application provider and the big data framework provider for development, deployment, and operation.
- ***Management***–management of big data systems should handle both system- and data-related aspects of the big data environment. That is system management and big data life-cycle management. System management includes activities such as provisioning, configuration, package management, software management, backup management, capability management, resources management, and performance management. Big data life-cycle management involves activities surrounding the data lifecycle of collection, preparation/curation, analytics, visualization, and access.

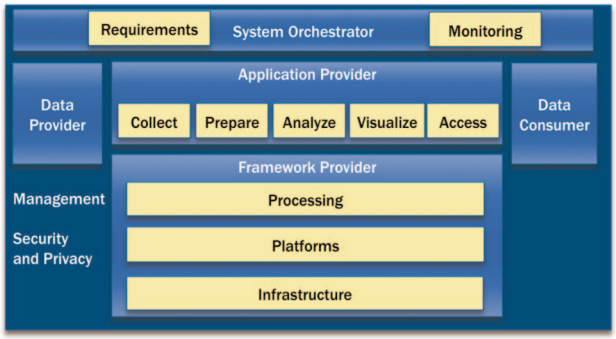

Figure 1. NIST big data reference architecture

The NIST 1500-2 taxonomy provides additional descriptions of these overarching roles and the activities within these roles. Data science activities would reside primarily under the application provider where the analytics life cycle takes place, with tight coupling to the resources in the processing and platform frameworks. Data science work roles could span all of those NIST reference architecture roles.

B. EDISON

The EDISON project [12] is an European Union (EU)-funded effort to "speed-up the increase in the number of competent and qualified Data Scientists across Europe and beyond". The focus of the collection of information is the EDISON Data Science Framework (EDSF), which comprises several related documents, including the Competence Framework, the Data Science Professional (DSP) Profiles, and the Model Curriculum.

The DSP Profiles define four main occupational groups: data science infrastructure managers, data science professionals, data science technology professionals, and data and information entry and access. Each of these profiles has descriptions for specific roles within that occupational group. Of particular focus is the data science professional, which has the following roles:

- ***Data Scientist***–data scientists find and interpret rich data sources, manage large amounts of data, merge data sources, ensure consistency of data-sets, and create visualizations to aid in understanding data. Build mathematical models, present and communicate data insights and findings to specialists and scientists, and recommend ways to apply the data.
- ***Data Science Researcher***–applies scientific discovery research/process, including hypothesis and hypothesis testing, to obtain actionable knowledge related to a scientific problem, business process, or reveal hidden relations between multiple processes.
- ***Data Science Architect***–designs and maintains the architecture of data science applications and facilities. Creates relevant data models and processes workflows.
- ***Data Science Programmer***–designs, develops, and codes large data (science) analytics applications to support scientific or enterprise/business processes.
- ***Data/Business Analyst***–analyses large variety of data to extract information about system, service, or organization performance and present them in usable/actionable form.

C. SAIC

SAIC is a systems integrator that works primarily for the federal government, including civilian, defense, and intelligence customers. To increase efficiency in developing and deploying BDA systems, SAIC developed an internal process model known as Data Science Edge™ [8], shown in Fig. 2.

This model extends the earlier limited data mining process of CRISP-DM to add in big data, systems development, and data-driven decision-making considerations, including the alignment with agile process models.

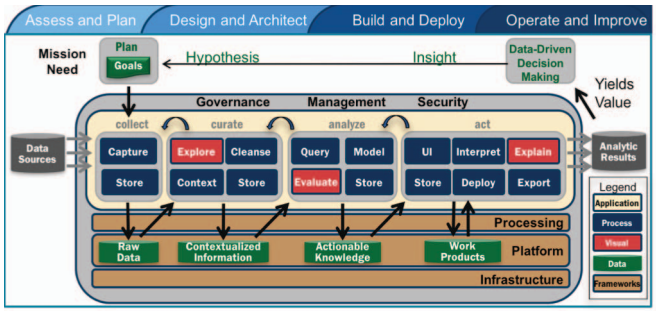

Figure 2. SAIC's Data Science Edge BDA process model

This overarching process model aligns with the general data science roles SAIC uses. In addition to the traditional roles for software and systems development, SAIC has specific roles for data science, big data platforms, and data management:

- ***Information Architect***–designs shared information environments involving models or concepts of information. Develops data models for optimal performance in databases. Designs data structures for data interchange. Develops data standards and converts data to controlled vocabularies.
- ***Data Scientist***–works in cross-functional teams with data at all stages of the analysis lifecycle to derive actionable insight. Follows a scientific approach to generate value from data, verifying results at each step.
- ***Metrics and Data***–develops, inspects, mines, transforms, and models data to raise productivity, improve decision-making, and gain competitive advantage. Conducts statistical analysis of data and information to ensure correct predictive forecasting or classification. Manages all aspects of end-to-end data processing.
- ***Knowledge and Collaboration Engineer***–designs and implements tools and technologies to promote knowledge management and collaboration within the enterprise.
- ***Big Data Engineer***–develops parallel data-intensive systems using big data technologies. Works with the full open-source Hadoop stack from cluster management, to data repositories, to analytics software to schedulers.

In addition, SAIC has management roles for each of these positions. Note that while DSE calls out visualization explicitly in three distinct types, there are no specific roles for visualization or business intelligence specialists, which are usually found among software engineers or data scientists

SAIC recognizes the importance of understanding the provenance of data in a given domain. BDA systems are developed by teams that collectively contain all the skills needed for agile BDA system development [13], as shown in Fig. 3.

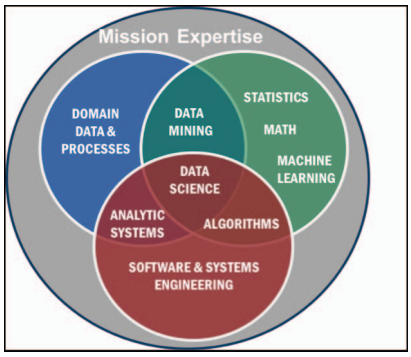

Figure 3. SAIC's data science team expertise model

D. *Springboard*

Springboard, an online data science education startup, defines three roles: data engineer, data scientist, and data analyst. As one can see from Fig. 4, all of these roles require software engineering, math/stats, and data communication skills.

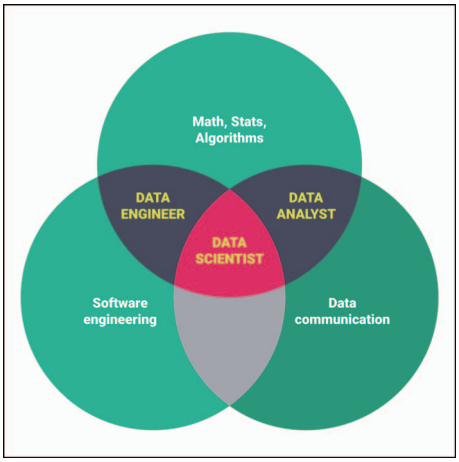

Figure 4. Springboard career paths

Below we describe some possible roles, based on springboard's definitions:

- **Data Engineer**–relies mostly on his or her software engineering experience to handle large amounts of data at scale. Typically focuses on coding, cleaning up data sets, and implementing requests that come from data scientists. Typically knows a broad variety of programming languages, from Python to Java. When somebody takes the predictive model from the data scientist and implements it in code, they are typically playing the role of a data engineer.
- **Data Scientist**–bridges the gap between the programming and implementation of data science, the theory of data science, and the business implications of data. Can take a business problem and translate it to a data question, create predictive

models to answer the question, and tell a story about the findings.

- **Data Analyst**–looks through the data and provide reports and visualizations to explain what insights the data are hiding. When somebody helps people from across the company understand specific queries with charts, they are filling the data analyst role.

One role not shown in the diagram, but mentioned by springboard is a **data architect,** who focuses on structuring the technology that manages data models.

E. Gartner

Gartner is a consulting firm specializing in strategic advice to business officers such as a chief information officer. In a Gartner Report entitled "Staffing Data Science Teams" [14] they considered that a number of individuals are doing data science but are not identified as such. The set of roles they suggested are as follows:

- **Data Scientists**–critical key staff members that can extract various types of knowledge from data, have an overview of the end-to-end process, and can solve data science problems.
- **Data Engineers**–make the appropriate data accessible and available for data scientists and, as such, can be instrumental in big productivity gains.
- **Business Experts**–individuals that understand the business domain really well. These can sometimes be the business leaders or sometimes a range of key specialists.
- **Source System Experts**–those that have intimate knowledge of the data at the business application level.
- **Software Engineers**–needed sporadically when custom coding is required (special visualization, data integration, or deployment of certain results, for example).
- **Quant Geeks**–excel in a specific range of quantitative skills. In certain situations, they are a "nice-to-have", but in rare situations, they are a "must-have".
- **Unicorns**–data scientists that are extremely well versed in the whole range of skills and are romanticized occasionally in the literature. They are super-rare.

Gartner broke down the main skill areas into Domain Understanding, IT Skills, and Quantitative Skills, with Fig. 5 indicating the strength of their expertise.

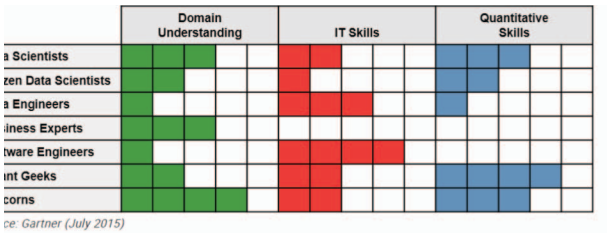

Figure 5: Team roles versus mapped skills

V. Discussion

A. An Integrated View of Roles Used

Table I shows the roles used by more than one of our case studies. As one can see, there is not a consistent use of roles used. Other roles mentioned ranged from COMMONLY used roles such as information architect and software engineers to unique roles not commonly seen, such as data providers and source system experts.

TABLE I. Summary Of Roles Used Across Multiple Case Studies

	EDISON	NIST	SAIC	Gartner	Springboard
Data Science Researcher	✓				
Data Scientist	✓		✓	✓	✓
Data Architect	✓				✓
Data Analyst	✓				✓
Data Science Programmer	✓				
Data Engineer			✓	✓	✓

B. Role Usage in Industry–Open Jobs

To see how these roles are currently used, we searched dice.com, a technology-focused jobs website. In this analysis, we used each of the roles as the search phrase (e.g., "data architect"). Table II shows the number of jobs found for each role. Not surprisingly, some of the roles had no hits, as these roles (e.g., "data science researcher", "system orchestrator") are not currently used in industry. Other roles, such as "data science programmer", also had no hits, which might be a bit more surprising, since there is programming within a data science team (we believe this role is currently being filled by data engineers).

TABLE II. Summary of Selected Search Results

Role	Number of Jobs
Data Science Researcher	0
Data Scientist	440
Data Architect	467
Data Analyst	696
Data Science Programmer	0
Data Engineer	378
Big Data Engineer	107
Information Architect	123
Data Science Manager	7
System Orchestrator	0
Data Provider	28
Big Data Application Provider	0
Big Data Framework Provider	0
Data Consumer	3
Security and Privacy	148

One can also see that there is roughly the same number of positions using “data scientist” and “data engineer”, and the most common phrase was “data analyst”, since a data analyst is a more common role, beyond use within data science teams. The information architect is another role that is used extensively outside of a data science context.

To gain additional clarity on the two most popular roles directly related to data science (data scientist and data engineer), we explored the key phrases used in the job description. As can be seen in Fig. 6, the data engineer is more focused on the data pipeline, data sets, and data ingestion. In contrast, the data scientist role is more focused on skills such as machine learning and data analytics.

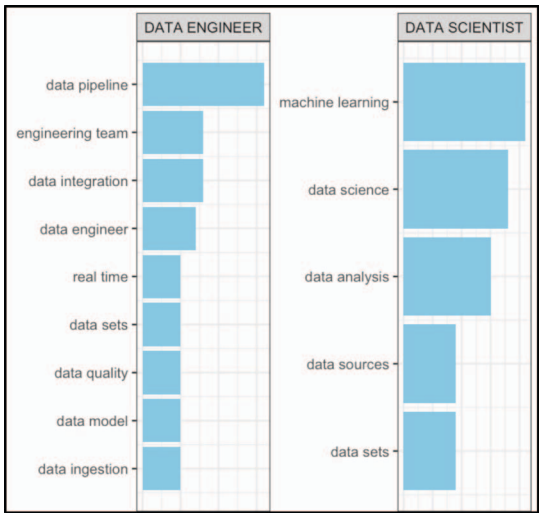

Figure 6. Popular phrases for different roles

VI. FUTURE

Data science is continuing to mature and evolve from the significant shift that occurred in techniques and technologies around 2007 for big data, deep learning, cloud, and Internet applications. The next big change in data science will likely occur in the eventual blending of data-intensive and compute-intensive applications. Compute-intensive applications for high-performance computing (HPC) traditionally have had either the data needed on each node (known as shared nothing) or the data the compute nodes needed resided on one data node (shared everything), as shown in Fig. 7. Big data developed initially through MapReduce, which is an “embarrassingly parallel” data application—meaning there is one compute node that sends the same code to any number of data-nodes, which each have a portion of the data. The HPC community is seeking to access larger datasets (becoming known as High Performance Data Analytics or HPDA) while the big data community is leveraging parallel computation such as GPU computing to develop large machine learning modules for BDA. While the crossover between compute- and data-intensive applications is recent, this is a relationship to consider when developing work role models.

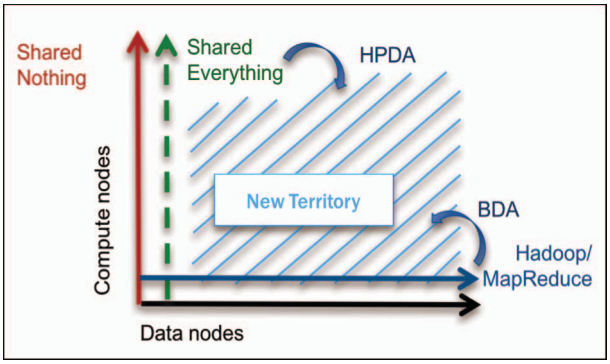

Figure 7. Team roles versus mapped skills required

In addition, as team data science becomes more common, the need for an accepted data science process methodology will grow in importance. Initial work in this area has shown significant differences in team performance based on the process methodology used by the team [14]. We believe any comprehensive workforce development framework and/or vocabulary, as well as any data science process models, will need to consider the blend of data-intensive tasks with compute-intensive tasks.

VII. CONCLUSION

Providing clarity and greater specificity for the term data scientist would assist in the training and employment of specialists with the required skills. One way to achieve a comprehensive workforce description for data science would be via a consortium of interested parties from government, industry, and academia. The first task of that consortium would be to develop a detailed data science process model that reflects all participants’ consensus on what defines a data science activity before beginning to determine requisite data scientist skills and job titles. This work could then follow the NICE model to provide categories, specializations, work roles, and tasks to clarify the differences in roles.

One potential challenge will be that traditional analytics systems that need straightforward summary statistics, reporting, or business intelligence are developed almost exclusively by software and systems engineers, along with traditional roles for data modeler, database analyst, and database administrators. The challenge will be to describe data science work roles for the activities they perform, which in many cases may overlap with the traditional systems development roles.

Given the continuing evolution of big data and data science, we note that current usage might not show how the industry is evolving, so a different, complementary next step may be to rerun an analysis of role usage in the industry in the future (perhaps every six months) to identify trends over time.

VIII. REFERENCES

[1] Saltz, J. (2015). The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness, Big Data Conference

[2] NIST SP 1500-1, "Big Data Interoperability Framework: Volume 1, Definitions" version 2 (201x), available at https://bigdatawg.nist.gov/.

[3] Saltz, J. S., & Shamshurin, I. (2016). Big data team process methodologies: A literature review and the identification of key factors for a project's success. In *Big Data (Big Data), 2016 IEEE International Conference on* (pp. 2872-2879). IEEE.

[4] NIST Draft SP 800-181, "NICE Cybersecurity Workforce Framework (NCWF), National Initiative for Cybersecurity Education (NICE)", (2016).

[5] National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework. (2017). available at https://niccs.us-cert.gov/workforce-development/cyber-security-workforce-framework

[6] NIST SP 800-16, "A Role-Based Model for Federal Information Technology/ Cyber Security Training", revision 1, second draft version 2, retrieved September 19, 2017, from https://www.nist.gov/sites/default/files/documents/2017/09/06/draft_sp800_16_rev1_2nd-draft.pdf

[7] Saltz, J., Shamshurin, I., and Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects. *Journal of the Association for Information Science and Technology*.

[8] Grady, N. W. (2016). KDD meets Big Data. In Big Data (Big Data), IEEE International Conference on. IEEE.

[9] Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining," Journal of Data Warehousing, 5(4)

[10] Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Available: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

[11] NIST 1500 series; available at https://bigdatawg.nist.gov/V2_output_docs.php

[12] Payne, J., Grady, N., Parker; H. (2017). AnalyticsOps for Data Science, In *Big Data (Big Data), 2016 IEEE International Conference on* (in press).

[13] Gartner, Staffing Data Science Teams. (2015). Refreshed: 02 September 2016.

[14] Saltz, J., Shamshurin, I., Crowston, K. (2017). Comparing data science project management methodologies via a controlled experiment. *Hawai'i International Conference on System Sciences* (HICSS-50).