An Empirical Approach to Understanding Data Science and Engineering Education

Rajendra K. Raj*
Rochester Institute of Technology
Rochester, NY, USA
rkr@cs.rit.edu

Allen Parrish†
Mississippi State University
Mississippi State, MS, USA
aparrish@research.msstate.edu

John Impagliazzo†
Hofstra University
Hempstead, NY, USA
john.impagliazzo@hofstra.edu

Carol J. Romanowski†
Rochester Institute of Technology
Rochester, NY, USA
cjr@cs.rit.edu

Sherif G. Aly
The American University in Cairo
Cairo, Egypt
sgamal@aucegypt.edu

Casey C. Bennett
DePaul University
Chicago, IL, USA
cbenne33@depaul.edu

Karen C. Davis
Miami University
Oxford, OH, USA
karen.davis@miamioh.edu

Andrew McGettrick
Strathclyde University
Strathclyde, UK
andrew.mcgettrick@strath.ac.uk

Teresa Susana Mendes Pereira
Inst. Politécnico de Viana do Castelo
Viana do Castelo, Portugal
tpereira@esce.ipvc.pt

Lovisa Sundin
University of Glasgow
Glasgow, UK
l.sundin.1@research.gla.ac.uk

ABSTRACT

As data science is an evolving field, existing definitions reflect this uncertainty with overloaded terms and inconsistency. As a result of the field's fluidity, there is often a mismatch between what data-related programs teach, what employers expect, and the actual tasks data scientists are performing. In addition, the tools available to data scientists are not necessarily the tools being taught; textbooks do not seem to meet curricular needs; and empirical evidence does not seem to support existing program design. Currently, the field appears to be bifurcating into data science (DS) and data engineering (DE), with specific but overlapping roles in the combined data science and engineering (DSE) lifecycle. However, curriculum design has not yet caught up to this evolution. This working group report shows an empirical and data-driven view of the data-related education landscape, and includes several recommendations for both academia and industry that are based on this analysis.

CCS CONCEPTS

• **Social and professional topics** → **Model curricula**; **Computing education**; **Computing education programs**;

KEYWORDS

ITiCSE 2019 working group; data science education; data engineering education; multidisciplinary education; global standards; accreditation.

ACM Reference Format:
Rajendra K. Raj, Allen Parrish, John Impagliazzo, Carol J. Romanowski, Sherif G. Aly, Casey C. Bennett, Karen C. Davis, Andrew McGettrick, Teresa Susana Mendes Pereira, and Lovisa Sundin. 2019. An Empirical Approach to Understanding Data Science and Engineering Education. In *2019 ITiCSE Working Group Reports (ITiCSE-WGR '19), July 15–17, 2019, Aberdeen, Scotland Uk.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3344429.3372503

1 INTRODUCTION

Since its inception from more than a half century ago [79], the broad field of the synthesis and analysis of large data sets has evolved radically, leading to the current demand for data science programs. The early focus on data analysis included procedures, techniques, evaluation and interpretation of those data, deeply embedded in the mathematics and statistical domains. Over the years, the forms of data have changed dramatically and the volume continues to increase exponentially [62]. The proliferation of data and technologies to organize, manage and analyze data has resulted in new academic disciplinary content and new competencies for jobs that must be achievable.

*Working Group Leader
†Working Group Co-Leader

ITiCSE-WGR '19, July 15–17, 2019, Aberdeen, Scotland Uk

ACM ISBN 978-1-4503-6895-7/19/07…$15.00
https://doi.org/10.1145/3344429.3372503

As is common in many nascent fields of study, there is a proliferation of activity among academic institutions to offer data science programs, and among prospective employers to identify required competencies, so as to align hiring decisions with appropriate and available academic program graduates. At the same time, the degree of alignment among academic programs–or between academic programs, academic materials and the job market–is still unclear.

This working group report develops an organizing framework to structure the broad field of data science education. The intention, to some extent, is to supplement the ACM Data Science Task Force's goal of providing "guidance for undergraduate data science programs of study" [21] by adding an empirical basis using existing data science and engineering programs. This framework is characterized by an appropriate use of words and labels to define concepts and subsets of the field, with the intent of achieving clarity of the fundamental ideas that underpin data science. The work utilizes an empirical approach to formulate definitions from observations of the educational and career space, rather than simply developing arbitrary definitions. However, the framework may motivate changes to empirically derived definitions or concepts to improve clarity and consistency.

This report begins by exploring the background underlying data science including possible definitions for the field, supported by tables and visualizations. Section 2 focuses on empirical results from the data science job market, current degree programs in data science, data science textbooks, and online data science course content. This report then analyzes these data about data science empirically.

The salient aspect of the report is the discussion surrounding data disciplines, curricular issues, pedagogical concerns, the need for security, privacy, ethics, communication, and other aspects encompassing the fields of data science (DS) and data engineering (DE), or when viewed together as data science and engineering (DSE). Recommendations generated from this work include ongoing clarity of definitions, evolving visualizations, recurring themes, fundamental data science courses, data science resources, the skills gap between academia and industry, and the need for industry-academic cooperation.

Given the fluidity of the DSE-related fields at present, this report aims to provide a platform for further discussion, resulting in a better understanding of the DSE-related fields and their impact on undergraduate education. datae Many different terms have been introduced in the DSE field to reflect fundamental concepts and categories within the broad area of DSE-oriented competencies. Figure 1 reflects the relative frequency in Google search results of many of these terms, such as big data, data mining, data science, and data engineering. Note that numbers on the y-axis represent search interest relative to the highest point on the chart for the given region and time.

From these data, clearly *data science* has the most momentum by far. While *data mining* and *informatics* dropped rapidly once the early buzz started to settle, *data engineering* never really caught on, *big data* is past its peak, and *data analytics* is growing at a slower rate than *data science*. What accounts for the term's surge in popularity? Though possibly coincidental, 2012 - the year the term began ballooning - was also when data science was dubbed the "sexiest job in the 21st century" by a Harvard Business Review

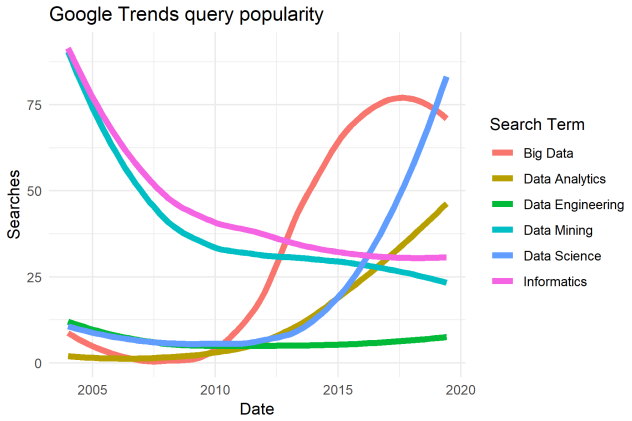

Figure 1: Search term popularity (Note: 100 means peak popularity for the term; 50 means half as popular; and 0 means insufficient data.)

article [24], perhaps a title so alluring that data science became a "catch-all" term.

Of the various definitions in the popular literature, it is important to distinguish between *data science* and *data engineering*. Data science is so popular a term that it may be overloaded, while data engineering represents a different perspective that is emerging as a separate concept.

This report formalizes and strengthens the notion of a bifurcated space within the broad area of data-oriented competencies by defining two broad notions: one notion for *data science* and the other for *data engineering*. Hypothetical definitions for these two ideas are developed below, along with an organizing framework to view the two as complementary. The rest of the paper attempts to validate these definitions as being meaningfully descriptive and complementary through an analysis of empirical data on educational programs and careers.

1.1 Definitions of data science and data engineering

A number of reports provide recommendations for data science (DS) education [10, 25, 26, 41, 57, 85]. There are commonalities and differences about how DS is defined, some of which is synthesized in this discussion. DS is not entirely a new research field; rather it has evolved as a multidisciplinary field that integrates approaches from mathematics, statistics, machine learning, data mining, and data management, while incorporating other areas such as communication skills and knowledge of a particular domain. DS extracts semantic value from the data–both structured and unstructured–to obtain meaningful information. In addition, DS activities include acquiring, integrating, cleaning, analyzing and using data. In this context, data scientists are expected to have a solid mathematical and statistical foundation, the ability to apply existing techniques and design new algorithms, and computational skills to work with massive amounts of data. In addition, they should develop some expertise in the specific knowledge domain as well as professional

skills such as communication and ethical reasoning. Data science educators should seek to develop and inculcate a data science mindset that encompasses these essential features.

The number of definitions of data engineering (DE) is considerably fewer than what exists for data science [3]. Generally, these definitions involve the management of data acquisition infrastructure, as well as the staging, cleaning and architectural flow of data. While data science is largely a process driven by analytics and conclusions, data engineering is largely a process driven by construction of a data infrastructure.

1.2 Relationship between data engineers and data scientists

There have been many discussions and even debates contrasting the meaning of engineering with the meaning of science. Engineers like James Watt design and build things, while scientists conduct scientific experiments and do research to advance knowledge. Biologists, physicists, chemists and mathematicians such as Albert Einstein, Charles Darwin, and Archimedes fall into this category.

Clearly, the data field has not reached a level of maturity when compared to classical engineering and science fields. Notwithstanding, from a practical viewpoint, it is possible to understand the relationship between what data engineers design and build and what data scientists explore, related to data.

One viewpoint considers data engineers as workers whose "primary job responsibility involves preparing data for analytical or operational uses" [65]. Another viewpoint is that data engineers are "responsible for the maintenance, improvement, cleaning, and manipulation of data in the business's operational and analytics "databases" [18]. A third viewpoint is that data engineers design and "develop tools to ensure clean, reliable, and performative access to data and databases" [35].

Engineering thinking can be described as a flow of thought from theory to the concrete; scientific thinking is a flow of thought from the concrete to theory. Engineering design and scientific inquiry are complementary; their mutual synergies support each other. Engineers seek options for working designs; if one or more options work, then the design is successful. Scientists seek unique, correct theories; if several theories seem plausible, all but one must be wrong. That is, in engineering one successful design can validate a concept, no matter how many previous versions have failed; in science one failure can disprove a theory, no matter how many previous tests it passed. Design applies information through tools; inquiry extracts information through instruments. Design shapes the physical world to fit its descriptions; inquiry shapes its descriptions to fit the physical world. The contrasts are between designs and theories, leading to different ways of thinking [47].

It is possible to view engineering design as a top-down structure working from theoretical principles to the construction of physical things. Likewise, it is possible to view scientific inquiry as a bottom-up structure working from physical things of interest to provable theoretical abstractions. Figure 2 illustrates this idea based on Drexler [27].

The right part of the figure shows the flow of information from the abstract to the physical. The left part of the figure shows the flow of information from the physical to the abstract. Each part has an intermediate "concrete description" phase based upon two different foundations. For engineering design (right) the foundation is specification; for scientific inquiry (left) the foundation is data.

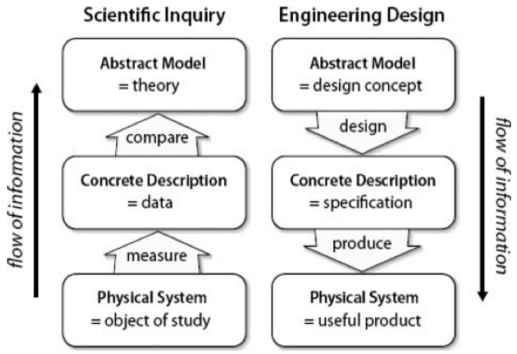

Figure 2: Scientific inquiry versus engineering design

1.3 Are there additional subfields?

Fundamentally, data scientists use algorithms to analyze and predict data behavior based on current information; data analysts use mathematics, statistics, probability, spreadsheets, and business intelligence tools to describe and to categorize data. For example, for an online insurance company, a data scientist might analyze data on customers who purchase insurance online and then use that information to predict a 'perfect' insurance policy for each new visitor to the website. A data analytics specialist for the same company might create visuals to help marketing track who is buying each insurance policy and the amount of profits the company is making. It is thus possible to broaden the above scenario to categorize three different but related fields: data engineering, data analytics, and data science.

Data engineering (DE) is a field that involves working with the lifecycle of data sets, helping to make raw data useful to a user and to a project. DE professionals in this field find efficient methods to collect, aggregate, and store raw data. They also design and implement useful, data structures to benefit the objectives of a cause or project [74].

Data science (DS) is a broad field of interest that refers to the collective processes, theories, concepts, tools and technologies surrounding large-scale data. DS specialists have tools to enable them to review, analyze and extract information from raw data. The DS field seeks to help individuals and organizations make better decisions from stored, consumed and managed data [75].

This report treats data analytics as a subfield within DS that focuses on drawing conclusions and generating actionable knowledge, within the broader DS field. Data analytics enable organizations to make more-informed business decisions and assists scientists and researchers to verify or disprove scientific models, theories and hypotheses [66].

In summary, data engineers support data scientists and data analysts by providing infrastructure and tools used to deliver end-to-end solutions to business problems. Data engineers build scalable, high performance infrastructures for delivering clear insights from

raw data sources. They design and implement complex analytical projects with a focus on collecting, managing, analyzing, and visualizing data toward developing batch and real-time analytical solutions for real-time complex problems. On the other hand, data scientists engage in a constant interaction with the data infrastructure built and maintained by the data engineers. They are not responsible for building and maintaining that infrastructure. Instead, their task is to conduct high-level market and enterprise research [53]. So, data engineers design and build data systems; data scientists, who include data analysts, study data through statistics and mathematics, as well as through theory and abstraction.

1.4 A working hypothesis

The working hypothesis thus is that data-oriented competencies form a two-dimensional space where one dimension is a spectrum from raw data to information through the data lifecycle, and the other dimension reflects the context in which the competencies are executed, ranging from hardware through various levels of software and eventually to the human enterprise. Figure 3 reflects the nature of competencies in each of the four quadrants of this space.

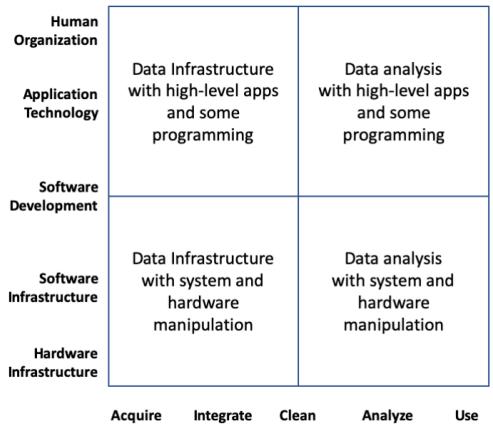

Figure 3: Two-dimensional data-oriented competency space: the x-axis shows Data Activities and the y-axes shows Specific Areas of Application

The range of competencies exceeds the span of what one job function can reasonably expect to accomplish, and can be legitimately divided between science and engineering. The initial hypothesis is that the division between data scientist and data engineer is approximately an even division, as in Figure 4. Note that the division, shown by a dotted line, is not a hard boundary. Further, data science and data engineering (DSE) should be viewed as separate, yet related, disciplines within this space.

In this division, data engineers focus on earlier parts of the data lifecycle, combined with the use and manipulation of hardware, systems infrastructure and software code. Data scientists, on the other hand, focus more on information and applications to the organization, primarily using vanilla and customizable application software and code.

CC2005 [68] provides similar two-dimensional diagrams for several computing disciplines. This paper extends that notion to DSE in

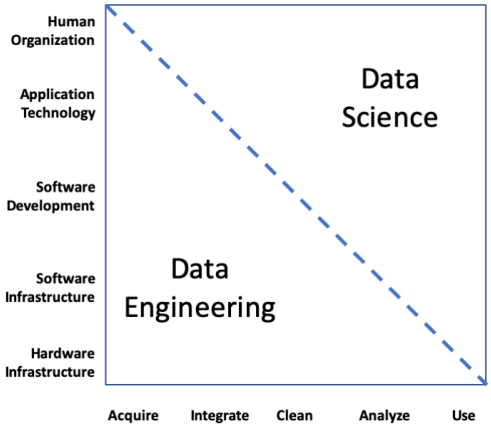

Figure 4: Data science vs. engineering: the x-axis shows Activities and the y-axes shows the Specific Area of Application

a relatively straightforward way to capture the essential elements of the two disciplines. The paper's argument is that just as computer engineering, computer science, information systems, information technology, and software engineering are all co-disciplines within the broad computing space, data science and data engineering are co-disciplines within a similar space of DSE competencies.

2 EMPIRICAL DATA AND ANALYSIS

To understand the DSE space better, this report looked beyond academic distinctions to the online data-related community. It took advantage of various surveys and data available in online technical forums and communities from those who self-describe as working in a DSE area, using this information to explore how DSE professionals saw themselves and what they deemed important. The authors also examined available DSE textbooks for coverage of DSE content. Finally, a machine learning model was built to predict if an individual would self-identify as either a data scientist or data engineer, based on tool usage.

2.1 Analyzing the DSE space

The analysis conducted for this report was conducted using several approaches, as discussed below. It revealed several interesting gaps and mismatches between what is currently *taught* and what is currently *practiced*.

Education. One large part of the online DSE ecosystem is *Stack Exchange*, an online community for sharing knowledge and gaining reputation; within this community, there is a dedicated branch called *CrossValidated* [70]. Figure 5 provides an insight into the areas of focus among DSE learners who post queries on Stack Exchange. An even split was observed within the use of traditional methods, such as ANOVAs and regression, as well as within more recently popularized topics such as machine learning, neural networks and Bayesian statistics.

Studyportals is an online platform that aggregates English-taught degree programs from 3200+ universities worldwide [71]. While

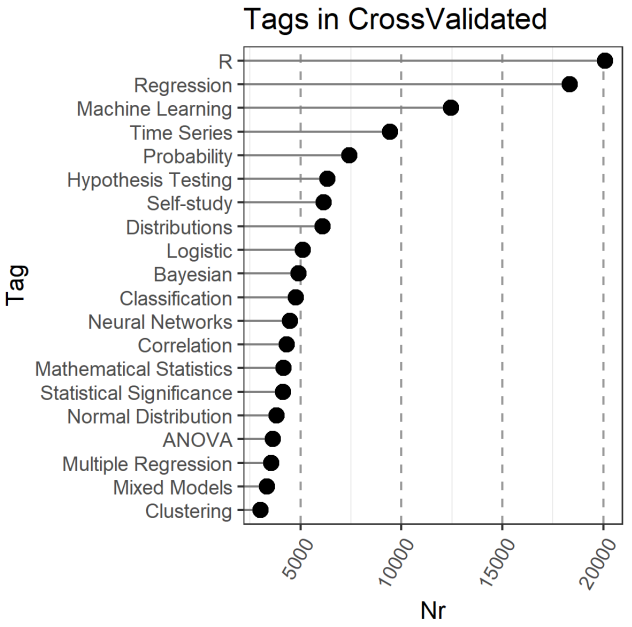

Figure 5: Top 20 popular tags of the data science online community *CrossValidated* versus the number of responses (Nr)

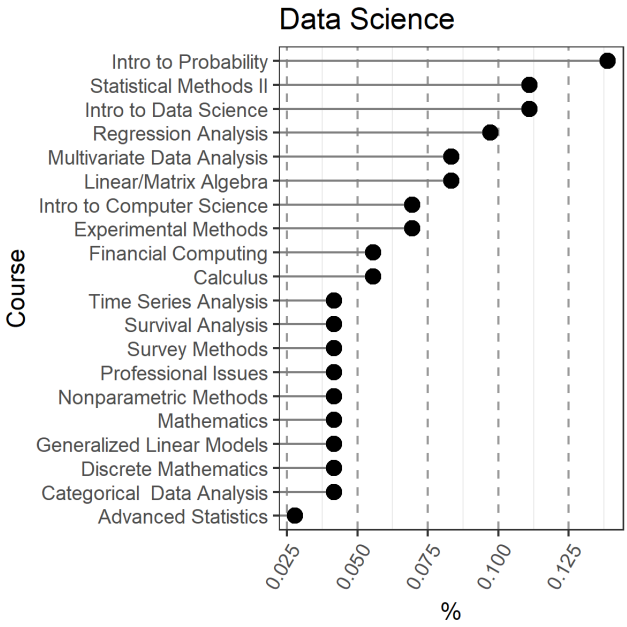

Figure 6: Top 20 popular courses in (English-taught) data science programs (n=288) worldwide

nothing guarantees its catalog to be exhaustive, this report views it as a fair sampling of degrees worldwide. Using a web spider, data from all bachelor programs categorized by *Studyportals* as "data science" was collected. This categorization was manually refined to exclude joint degree programs, specializations and statistics course variants, leaving an n of 288 programs. For each program, course modules advertised on the website were extracted and manually classified. For example, *Foundations of data science* and *Introduction to data statistics* were collapsed into one, while the distinction between module names *Data mining* and *Data analytics* was retained. There was no easy way of comparing modules by their content, and so only their titles could be used as reflective of the content. Figure 6 shows the normalized frequencies for the 20 most popular modules. Note that, for data science, *Statistical Methods I* were excluded as they were included in nearly every program.

The most frequent courses, as shown in Figure 6, are introductory courses in DSE, statistics, and computer science. Common fixtures also include mathematical modules covering discrete mathematics, calculus, and linear algebra, as well as traditional statistical approaches such as regression analysis and generalized linear models. It is noteworthy that no explicitly labeled machine learning or Big Data course made it to the Top 20.

Job market. For a curriculum to meet the demands of the job market, it is worth tracking the most employable skills. Therefore, the search engine Indeed [42] was crawled for UK job ads seeking data scientists. From a sample of 226 ads, all listings of essential or desirable criteria were extracted, as shown in Figure 7. Consistent with other analyses, such as that by Muenchen [56], Python is the most requested language, making it a natural choice for an introductory language, later to be complemented with R and Java. For curricular purposes, two things stand out. First, distributed computing and cloud computing are highly sought-after specializations, but Big Data remains in demand. Secondly, many ads requested the ability to develop dashboard interfaces to datasets: Tableau is the sixth on the list, but the R dashboard library Shiny is also featured, along with web technologies like HTML, Flask and D3.

Working professional job tasks. Existing industry-wide surveys provide details of actual practice. For example, a 2018 survey distributed via social media by software vendor JetBrains [43] for people "involved in data analysis," another 2018 survey by data community Kaggle [44], one by Rexer Analytics (n=1123) [63] and O'Reilly (n=800) [72].

Among job roles, respondents working as "data analysts" (33%) and/or "data scientists" (32%) exceeded that of "data engineers" (21%). When self-perceived roles were considered instead of formal job titles, 62% identified as developers, 28% as data analysts, 22% data engineers, and only 18% as data scientists [43]. Note that respondents were allowed to select more than one role, and so the numbers do not add to 100%.

There was bifurcation in tool use among practitioners. Of the Kaggle respondents, 55.6% (out of 1477) reported they used Big Data tools. Among these, the top technologies were Apache Spark, Hadoop and Hive (39%, 37%, and 26%, respectively).

In terms of languages, there was a clear dominance of Python. JetBrains [43] reports that 57% use Python versus 15% for R. Additionally, there is little mobility in terms of switching from one to another (51% do not intend to adopt another language in the next year) and 56% expect Python to remain dominant (versus 9%).

Finally, although only 14% count Java as their main language, and few (7%) predict it will dominate anytime soon, 62% regularly use

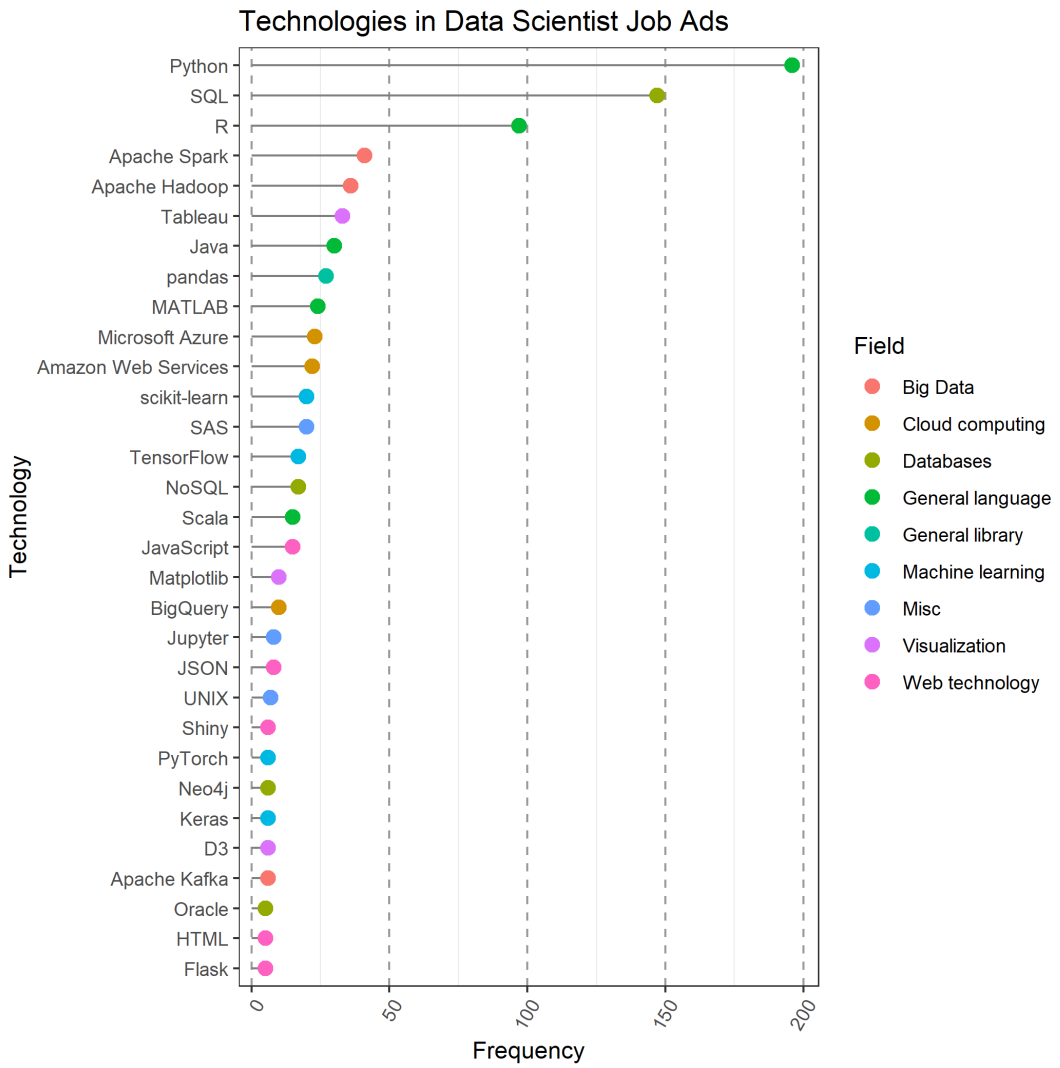

Figure 7: The most requested technologies in UK job advertisements for data scientists (n=226)

it. Java appears to inhabit a different niche for associated software engineering tasks.

2.2 Existing degree programs

Another rich source of empirical data is the set of existing degree programs in this space. One online source [19] enumerates 466 programs in the US, based on the "Awesome Data Science Colleges" list posted on Github. Most of these programs are master's degrees (301) while the others are bachelors-degree programs (46). The list also includes one associate's degree program from Brigham Young University, 19 doctoral programs and 99 graduate certificates. As this list was populated by Github users, it may safely be assumed that there are other programs than those appearing on this list.

The US is not the only country with the largest number of DSE-related degree programs. As presented by Yu [86], perhaps the greatest growth in the DSE space has been in at universities in China. In 2016, the Chinese Ministry of Education approved an official degree called "Data Science and Big Data Technology" for possible undergraduate degree offerings at all universities throughout China. By 2018, 283 universities in China had begun offering the degree. In 2019, an additional 196 universities began offering the degree. Despite the names, these 481 programs may inherently have elements of DE content. In China, official degrees such as "Computer Science and Technology" are de facto computer engineering programs where approximately 10% of the students specialize in computer science while 90% specialize in computer engineering. The landscape for the study of big data includes application, models and algorithms, platforms and tools, and data. The large growth of Chinese DSE programs is summarized in Table 1.

Table 1: Bachelor's DSE programs in China, per Yu [86]

Year	Universities	Programs
2016	3	3
2017	32	32
2018	248	250
2019	196	196
Total	**479**	**481**

2.3 Online content

No discussion of DSE-related courses is complete without considering the impact of massively open online courses (MOOCs). ClassCentral.com lists 491 data science courses in 11 languages, 368 of which are part of certification programs. Providers include MOOC platforms such as Coursera, edX, and Udacity as well as industry offerings from Microsoft, Google, and IBM. The popularity of online courses is shown by Johns Hopkins' data science specialization, which by 2015 had enrolled over one million students [38]. As of August 2019, their offering of this set of ten courses had almost 250,000 students enrolled [20].

2.4 DSE textbooks

The working group conducted an analysis of DSE-related textbooks, a list of which is available in a separate online supplement [61]. The three prime sources of information for this textbooks' analysis were:

(1) textbooks used in various academic programs worldwide,
(2) textbooks mentioned in research publications pertaining to DS education as seen on Google Scholar, and
(3) relevant textbooks available for purchase online.

Of these, the last was seen as most relevant to study the supply of textbooks available to the professional and academic community. The course titles and descriptions of DS-related programs at several selected universities (Waterloo, Northeastern, Columbia, Warwick, Essex, Chinese University of Hong Kong, Tsinghua, Melbourne, Copenhagen, and the American University in Cairo) were used to formulate search phrases. Amazon was manually searched for textbooks using key phrases including but not limited to: data science, machine learning (ML), statistical computing, computational statistics, business analytics, data ethics, data regulation, statistical models, data processing, data mining, data management, data engineering, and information visualization. Default search parameters were used, and resulting textbooks were noted along with other recommendations of similar textbooks. Where available, the structure of the textbooks was browsed either through the online vendor, publisher, or author sites. Surveyed textbooks were manually classified as either fundamental or advanced based on the type and depth of topics covered. Figure 8 shows the breakdown of content coverage noted in this analysis.

Over thirty textbooks associated with DSE education were reviewed, and the list is available in the separate online supplement [61]. This study revealed a notable variance of coverage of knowledge areas. In fact, the lack of a global reference model for DSE education is evident, specifically as it relates to the perception of required knowledge units needed to build some form of capacity

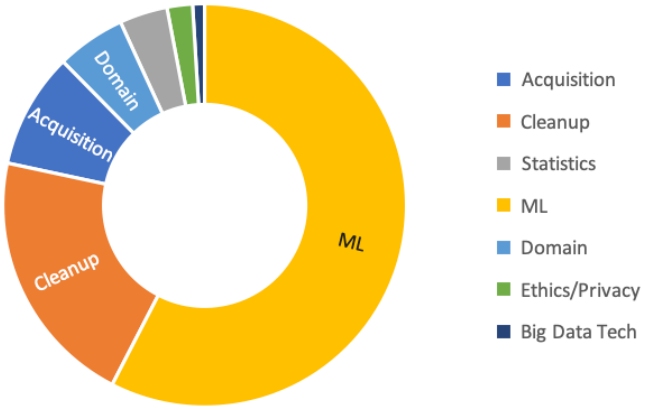

Figure 8: Coverage of DSE areas in textbooks

in this domain. Even though aspects of computational sciences, statistics, and domain knowledge are generally seen as three necessary pillars, only a few of the surveyed textbooks addressed these areas and how they fit to form a well-rounded data scientist. The more notable observation is that the textbooks are primarily themed around one of three areas: a programming language, machine learning, or statistics.

Surveyed textbooks that provide fundamental exposure generally start with an attempt to define the boundaries of the domain, followed by an exposure to some of the major building blocks such as data importation, tidying, transformation, visualization, modelling, and communication. The usage of different terminology or combination of the aforementioned building blocks is not uncommon. Almost all textbooks that delve into some practical exposure would use Python and R as the preferred programming languages, along with their most commonly used libraries in support of the domain. Some of the knowledge units covered at the introductory level include but are not limited to basic coverage of linear algebra, probability and statistics, and hypothesis and inference. Basic machine learning concepts for classification and clustering are also covered including regression, k-nearest neighbor, naive Bayes, decision trees, and occasionally some exposure to neural networks. Types of regression are covered at various levels of depth, including concepts of linear, multiple, and logistic regression. Few introductory technical textbooks cover ethical and legal issues, data regulation, and privacy, let alone provide any coverage of domain application of data science.

Other fundamental textbooks focus on statistics as the foundation for data science, with typically little emphasis on technology or domain. However, they generally do help to build a solid statistical foundation. Topics related to the structure of data, estimation of location, variability, distribution, correlation, sampling, statistical experimentation, and significance are some of the prominent topics covered. Nevertheless, a well-rounded coverage of other pillars for the domain is usually overlooked. For example, concepts of machine learning in such settings would generally be very abstract with little or no practical application using current technology, let alone any coverage of the knowledge use in various domains.

More notably, an interesting textbook [37] titled "Data Science from Scratch" considers data science to be an intersection of "hacking skills" (computational skills), "math and statistics knowledge," and "substantive expertise" (domain knowledge). This textbook explicitly refrains from covering the needed "substantive expertise" citing how huge this coverage may be, but does cite some interesting domain examples in the textbook. The textbook covers many of the topics mentioned earlier as a fundamental textbook in the domain, however it barely scratches the surface of each topic.

Other types of textbooks with fundamental coverage focus on the knowledge domain. For example, data science for business, for executives, for supply chain forecast, for medicinal research, for finance, and for fraud detection are all examples of textbooks that focus on domain. A data science for business textbook, for example, will be more focused on the usage of DSE within organizations to build competitive advantage, treating data as a business asset, mining data for organizational purposes, and also building human capacity in the domain. Other textbooks that are domain oriented will follow suit in their respective domains. Similarly, a textbook focused on finance will include topics such as the usage of data for forecasting financial trends, analysis of customer sentiment, and fraud detection. Moreover, a textbook in data science for medicinal research will focus on topics such as the computation of medical risks, identification of hazards, event history analysis, and disease prediction.

More advanced DSE-related textbooks in an independent domain are very difficult to spot. Textbooks that are explicitly affiliated to a domain typically cover advanced modeling topics such as additive and mixed models as well as more advanced regression techniques. Other topics in parallelism, dimension reduction, and time series analysis are also covered. Focus on trending technology, such as Spark and Hadoop, is very evident when depth is encountered. Other textbooks would direct attention to more advanced machine learning, specifically deep learning with some in-depth coverage.

A survey of advanced textbooks relevant to data science clearly indicates a lack of maturity in the comprehensive coverage of advanced knowledge units relevant to the domain. To be specific, advanced textbooks in DSE are simply more detailed coverage of topics in the subdisciplines. No clear affiliation to DSE as a domain is encountered when such advanced topics are covered.

This textbooks study shows that while many DSE-related textbooks have emerged over the past five years, the lack of a global perspective of DSE education is quite evident. Few textbooks of a fundamental nature acknowledge the need for a well-rounded coverage of computational sciences, statistics, and domain. Those that do fall short of achieving any notable coverage of the latter, let alone any coverage of ethical, regulation, or privacy issues. There exists a notable deficit of textbooks that provide in-depth coverage of DSE as an independent discipline of its own. As a result, major and swift strides need to be made to keep up with the fast pace at which DSE is evolving into an independent, and hopefully standardized, discipline.

In general, there exists an obvious mismatch in the coverage of topics presented in surveyed textbooks, and the necessary areas of DS identified by the NSF workshop [10] and shown in Figure 8. From the analysis, machine learning evidently appears to have the most coverage, followed by data cleanup. Domain and professional issues on the other hand demonstrate the least coverage. More importantly, there is little or no coverage of topics in data publishing and preservation/destruction [10].

As stated earlier, the list of DSE textbooks used for this analysis is available as an online supplement [61].

Table 2: Top 25 DSE job skills reported by practitioners

Job Skill	Combined Feature Average Use %
Python	83.5%
Batch data processing	78.1%
Local machine	75.9%
NumPy	67.6%
pandas	66.7%
Matplotlib	61.0%
Windows	59.0%
TensorFlow	58.4%
Linux	54.6%
scikit.learn	48.3%
Jupyter.Ipython SW	48.3%
R	46.7%
Spreadsheet V	46.0%
Streaming.Real.Time DP	44.1%
Mac	43.5%
PyCharm SW	41.3%
SciPy	40.0%
Apache Spark	38.7%
Keras	36.8%
Cloud Service	36.8%
ggplot2 R	34.9%
RStudio	34.9%
Apache Hadoop.MapReduce	34.0%
Tableau V	34.0%
Developer	32.7%

2.5 Self-reported job tasks of DSE practitioners

The working hypothesis underlying this report was supported by evaluating the tasks actually performed by practicing DSE professionals in their day-to-day jobs. This analysis used data from the JetBrains survey [43], which included responses from 1500 practicing analytics professionals, with self-reported job titles and skills/tasks from their daily job functions. From this total sample, self-identified data scientists or data engineers were selected resulting in a final sample of n=315. The survey focused on 101 specific skills related to data analytics functions. The top 25 most commonly reported job skills, combined across data scientists and engineers, are shown in Table 2. Note the differences between Table 2 and Figures 5, 6, and 7, as this appears to be further evidence of a disconnect in program curriculum design, something that needs to be addressed in DSE-related programs.

This data was analyzed using a machine learning (ML) approach to check whether the model could predict that an individual was either a data scientist or data engineer, based purely on their daily job tasks as reported in the survey. If so, the results would support

Table 3: Machine learning results for predicting data scientists vs. data engineers from self-reported job tasks

Classifier	Accuracy	AUC
Random Forest	0.75	0.86
Neural Network	0.70	0.82
SVM	0.73	0.82
Gradient Boosting	0.74	0.81

our hypothesized emerging bifurcation of the DSE field seen in Figure 3. Our approach followed a standard ML methodology [4], similar to prior work [9]. First, data from the JetBrains survey was extracted into Python and evaluated via the Scikit package. As the initial data was not balanced, with 207 data scientists and 108 data engineers, the data was rebalanced by undersampling the majority class, then performing SMOTE to arrive at a balanced dataset with the same total sample count (n=315) [17].

After the data was balanced, multiple classifiers were run on the dataset using basic default parameters: random forests, neural networks, SVMS, and gradient boosting. Performance was evaluated using standard 10-fold cross validation, which conforms to best practices [28]. The results are shown in Table 3.

Table 3 shows that it can be predicted whether an individual was a data scientist or data engineer 75% of the time, based on their actual job tasks. This result supports the hypothesis that the roles in DSE are indeed bifurcating, similar to what occurred early on in the computer science domain between roles such as software developers and information systems professionals.

Feature selection was also performed to investigate the features were driving this prediction. This effort was approached in multiple ways, including using wrapper-based approaches, mutual information filters, and calculating odds ratios [11]. Table 4 summarizes the results of this analysis, showing the top 15 languages/tools used by data engineers (left column) and those used by data scientists (right column). The table was generated by taking the top features from the wrapper method and calculating odds ratios, representing the degree of over-representation of the skill in data science vs. data engineering (and vice versa). For instance, data scientists were three times more likely to report using Seaborn while data engineers were three times more likely to use Java. The data engineering list contains more developer-heavy languages (Java, Scala, Visual Studio), Linux, and distributed computing infrastructures (Hadoop, Spark, Cloud setups), while the data science list is focused on more analytical software, including various Python and R libraries, deep learning packages (Tensorflow, Keras), and data visualization tools (Seaborn, Matplotlib, Plotly). Note that Python does not appear explicitly in either group, because, as can be seen in the list of commonly reported DSE job skills in Table 2, Python has become virtually ubiquitous in software development across a variety of roles.

Finally, note that the promising results presented in this paper are preliminary, and require additional validation through hyperparameter tuning and other advanced modeling techniques.

Table 4: Top 15 tools used in DSE

Rank	Data Engineering Tool	Data Science Tool
1	Linux	Matplotlib
2	Developer	TensorFlow
3	Apache Hadoop.MapReduce	scikit.learn
4	Apache Spark	R
5	Small cluster	SciPy
6	Java	Rstudio
7	Scala	Keras
8	Visual Studio Code SW	seaborn
9	SAS V	dplyr R
10	Medium cluster	Plotly
11	Google Cloud SW	xgboost
12	Dataiku	randomForest R
13	Colaboratory SW	Bokeh
14	Alteryx	JupyterLab SW
15	Domino	NLTK

3 THE ETHICAL FOUNDATION OF DSE

There is now recognition in both computing and non-computing circles of the critical need for professionals in data science and engineering to have a strong ethical foundation. This awareness parallels broader developments in society around data privacy and ownership rights. Several collaborative, multi-disciplinary reports from academia and professional organizations have proposed curricular recommendations for data science programs [10, 25, 26, 41, 57, 85]. The necessity of including a study of ethics is present in all reports (except Demchenko et al. [26]), although the depth of coverage varies. For instance, the NSF report states that one of the goals for ethical training of students is to address challenges "that will render data-driven systems useful, effective, and productive, rather than intrusive, limiting, and destructive" [10, p. 20]. The Park City Math Institute report mentions teaching ethics in some of the ten recommended core courses, but lacks specifics [25], while a draft report for two high school data-related courses by the International Data Science in Schools Project intersperses technical topics with ethical discussions [41]. The Data Science Leadership Summit summarizes one of their goals as "taking collective responsibility in the broader effort to prepare next generation data scientists to contribute in the best interests of society" [85, p. 1]; they recommend defining a code of ethics and integrating ethical case studies into research and education programs. The report also includes a link to a list of university courses on ethics and technology and other resources [31].

The report [57] by the U. S. National Academies of Science, Engineering, and Medicine (NASEM) describes data science as a hybrid discipline requiring analytical skills, communication skills, and problem-solving skills for both technical and ethical challenges. The report states that data scientists should develop data acumen, "the ability to make good judgments and decisions with data and use tools responsibly and effectively" [57, p. 1] by both tool developers and tool users, which can be thought of as ethical practice. Data acumen "is increasingly important, especially given the large volume of data typically present in real-world problems, the relative ease

of (mis)applying tools, and the vast ethical implications inherent in many data science analyses" [57, p. 11]. The report concludes with a data science version of the Hippocratic Oath. In particular, the NASEM report explains why ethics is important, provide illustrative examples and affirms the importance of societal context.

A typical starting point for raising awareness of ethical behavior is through codes of ethics. There are many possible ways to describe the overlapping terminology, principles, and tenets of the codes of ethics published by various professional societies and organizations [2, 5–7, 22, 23, 32, 33, 36, 57, 58, 60]. Most include (1) usage of data (protecting privacy, for example), and (2) considering impact of algorithms and techniques. Regardless of which code is selected for inclusion in a course activity, the important points are that students have some appropriate set of normative behaviors for their profession and that they learn to conduct and continue to practice ethical reasoning. Tractenberg and FitzGerald [78] suggest that exposure to prerequisite knowledge about ethics, such as in a code of ethics, is the first step toward developing ethical reasoning skills.

Burton et al. [13] state that "a good technology ethics course teaches students how to think, not what to think, about their role in the development and deployment of technology, as no one can foresee the problems that will be faced in a future career" [13, p. 54]. In addition to teaching students to solve technical challenges, it is imperative that they develop skills to engage with ethical challenges arising from their professional work. A goal of teaching ethics is to equip students with the means to discuss, reason, and reflect on ethical issues. Codes of ethics define normative behavior for a professional practitioner, but a code cannot solve all problems and may even have conflicting concepts for a given situation. "Ethics education often requires a different kind of education from understanding and applying an established body of knowledge" [13, p. 58]. By also exposing students to different kinds of ethical schools of thought (descriptive ethics) and having them practice interpreting ethical issues using these theories, they have the opportunity to question and explore beyond their own assumptions.

One practical tool for classroom teaching of ethics to data science and data engineering students is the Ethical Reasoning Mastery Rubric (MR-ER) [76]. There are five knowledge/skill/ability categories (KSAs) and four proficiency levels (novice, beginner, competent, and proficient). The KSAs are the components of developing ethical reasoning capabilities: (1) recognizing a moral issue, (2) identifying decision-making frameworks, (3) identifying and evaluating alternative actions, (4) making and justifying a decision, and (5) reflecting on a decision. This is a process that can be applied to examining case studies or topics encountered during specific stages in the data lifecycle throughout a student's education in data science, as shown in Figure 9.

This leads to a suggested call to action: DSE needs good case studies (or even good fiction [13]) generally associated with stages in the pipeline and more specifically with techniques and algorithms deployed in technical courses. Consistent with the NASEM recommendations for intertwining ethical and technical studies [57], Tractenberg suggests that "a one-time ethics training 'vaccine'" is not ideal [77], but rather creating a culture of ethical practice through repeated and consistent exposure, aligned with their technical training, would produce data scientists and engineers better prepared to engage with the challenges to be faced in their future professional lives.

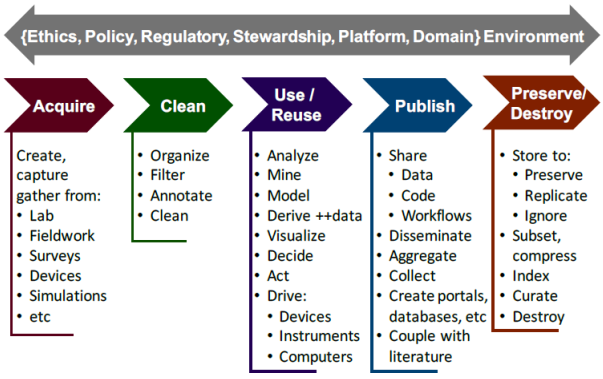

Figure 9: Data lifecycle, from NSF workshop report [10]

4 OVERLAPS WITH OTHER DISCIPLINES

As stated earlier, DS and DE are nascent fields that draw from other established disciplines. Therefore, these overlaps need to be acknowledged when discussing DS and DE education.

4.1 Artificial intelligence and machine learning

DS overlaps considerably with the field of artificial intelligence (AI) [80], but how DSE, machine learning (ML), and AI fit together is a matter of some debate. Among some practitioners and researchers in these fields, AI is seen as an umbrella term that includes machine learning, but not DS–although the part of DS that uses machine learning is contained within the AI universe.

The definition of AI is still rather fluid, as is the definition of DS. However, one difference between AI, machine learning, and DS is due to the reasoning and automated capabilities of AI [34]. At this time, machine learning is the approach used in both AI and DS [81]. The output from a DS project is usually presented to a human user, whose decision on a course of action is based primarily on the discovered knowledge.

For non-computing professionals, the boundary between AI and DS is more blurry. Carlos et al. [15] conflate the two terms in discussing the future of radiologists' jobs once machine learning/DSE/AI takes over their field. Elish and Boyd [29] claim that AI is just a repackaging of big data and machine learning into a more palatable term that avoids the "Minority Report" flavor of the term "big data" and discuss AI as a field practiced by data scientists and engineers.

Are data scientists also machine learning experts? Kozyrkov lists ten roles in a data science team [49], and states that the data scientist encompasses the roles of expert analyst, statistician, and applied machine learning engineer. Although data scientists develop and implement machine learning algorithms, machine learning engineers tend to write production level, scalable code [67], which reinforces the applied focus.

4.2 Overlaps with cybersecurity

There are two distinct ways in which DSE overlaps with cybersecurity. First, DSE needs to pay careful attention to issues of data privacy and security for legal and ethical reasons. Second, many DSE techniques can help to solve a variety of cybersecurity problems. Successful practitioners will need to make use of the synergies between the two.

The increased use of internet-of-things (IOT) devices caused increased access to the internet. Consequently, this access has created a large amount of data. This phenomenon has challenged organizations to implement adequate methods to process these data. Data processing aims to extract interesting and meaningful patterns as well as knowledge from all types of data. In addition, due to their lack of expertise with data analytical skills, the information security community has found it difficult to analyze logs, network flow, and system events for forensics and intrusion detection.

Moreover, common technologies are not always adequate to support long-term, large-scale analytics. An example that supports this fact is the storage of large quantities of data for a long period of time; event logs and other recorded computer activities are often eliminated after a defined retention period, e.g., 60 days [14].

Sensitive data collection and storage have brought challenges to an organization's privacy and confidentiality; they also have accentuated security risks. In fact, an attacker usually intends to compromise organizational information security goals, namely confidentiality, integrity, availability, authentication, authorization, and non-repudiation. On the other hand, attackers are continuously improving their attack techniques and strategies, which requires adequate data analytical skills in security to place data patterns in context, to formulate hypotheses and to estimate or predict security cyber attacks. The large amount of data generated by automatic logs and sensors, traces from intrusion detection, malware analysis, insider attack analysis and phishing detection requires efficient and automated DSE techniques. The literature shows that the techniques generally used in the scope of cybersecurity are statistics, data mining, machine learning and natural language processing for security [82]. In practice, this knowledge's basic requirements comprise mathematics, statistics, and computer programming through data structures and algorithms, which are typically part of a computer science or computer engineering degree.

Cybersecurity deals with data. However, one should emphasize that data might have missing or corrupted values or different types of attributes. These require a cleaning procedure and decision mechanisms for handling missing or corrupted values. To understand the types of data and attributes, cybersecurity professionals and students must know how to preprocess data and their attributes [82]. In fact, recent big data applications are starting to become part of security management software since they support efficient cleaning, preparing, and querying data in heterogeneous, incomplete and noisy formats [14].

Regarding applied basic statistics, practitioners and students should study content to include parameter estimation, confidence intervals, hypotheses tests and Bayesian techniques' essential to understand machine learning techniques required for computer security data analysis [82].

Data mining techniques such as association rules, clustering and anomaly detection are applicable to cybersecurity. For example, intruders usually use botnets to send spam and to host phishing websites, which are difficult to trace and blacklist. Association rules are applicable in intrusion detection for anomaly detection through the frequency of item-sets and generated association rules [52]. The malware can come in a wide range of forms and variations such as in viruses, worms, botnets, rootkits, Trojan horses and denial-of service attacks. In practice, malware exploits software vulnerabilities in browsers and operating systems, or using social engineering techniques to deceive users to run and execute malicious code. Additionally, one can support malware analysis performed through the identification of samples that exhibit similar behaviors through automated clustering techniques. Furthermore, execution traces of malware programs generate behavioral profiles, which are useful as input to efficient clustering algorithms, thereby allowing them to handle sample malware sets [8].

In machine learning, cybersecurity students should know topics such as nearest neighbor, decision trees, neural networks and time-series prediction to help them in new applications such as in filtering spam email (nearest neighbor) [4], intrusion detection (neural networks) [73] and prediction attacks and threats, thus contributing to a more preventive posture regarding security.

Regarding security applications, natural language processing (NLP) concepts might include basic knowledge of information retrieval techniques such as retrieval models, web search, and information retrieval metrics (recall and precision) [82]. Example applications of these concepts involve malware detection and authorship attribution methods, which includes spam filtering, fraud detection, computer forensics [46], cyber bullying, and plagiarism detection [30, 48, 51]. Word sense disambiguation and knowledge bases such as Wordnet [55] are applicable to phishing email detection through a classifier design, which combines semantics and statistics in analyzing text in emails [41, 83]. NLP is also useful for security application in URL detection and in sophisticated phishing attacks aimed to install malware on computers or hijacking a website [46]. Cybersecurity students with diverse backgrounds can use NLP techniques in areas such as information technology, information systems management, computer engineering and computer science [82].

In conclusion, cybersecurity professionals and students should have some knowledge in DSE and its techniques to provide them with a set of skills to address cybersecurity challenges. However, it is difficult to ensure satisfactory depth of knowledge for all DSE-related topics. One strategy is to combine an understanding of cybersecurity with DSE knowledge. Another strategy is to have cybersecurity professionals collaborate with DSE experts and develop efforts to evolve basic knowledge in data science to ensure a successful cooperation.

5 RECOMMENDATIONS

The following discussion provides some suggestions that may be useful in sustaining and developing DSE as a robust field of interest.

5.1 Thinking like a data scientist and engineer

The first recommendation that this paper makes is that students needed to learn "how to think like a Data Scientist" [39], and Data Engineer. This is a lesson learned from other fields that have grappled historically with attempting to teach a mindset, such as engineering and law. For instance, if one examines how law school curriculum in many US universities is structured, the first thing students learn at many law schools is that faculty are not there to teach them "the law." Students learn that when they are studying for the bar exam. Rather, the emphasis in law school is to teach them how to "think like a lawyer," otherwise referred to as the concept of legal reasoning [40, 50]. Empirical research has shown its effectiveness in completely transforming how students mentally approach legal questions in a matter of months, across a diverse array of student backgrounds [50]. In much the same way, the ultimate goal of DSE education programs ought to be to teach students how to think like a data scientist and engineer. Although it is important for students to learn technical skills, DSE is not just knowing how to program or how to apply statistical techniques: it is a *way of thinking* where an analytical mindset that takes advantage of technical know-how.

In real-world practice, this mindset manifests as an empirical approach to formulating problems and testing hypothetical answers to them, using computational tools.

From a pedagogical standpoint, the above dictates that the focus must be on creating a consistent educational experience across students: a shared common thought process in which a student from Chicago and one from Germany or China can communicate about a problem to be solved, and quickly understand each other. This consistent experience is rooted in some ways in the notion of constructivist pedagogy – the theory that people learn by constructing their own understanding and knowledge of the world [16], experiencing things and then reflecting on those experiences. The fundamental idea behind this approach is that people learn through their own curiosity, not by being told or talked to, nor through rote memorization of programming code or equations or the like. There is heavy evidence of this kind of curiosity-based learning via studies of how human infants learn [59, 84], as well as in more formal science and math education [69]. In reality, effective teaching has to be built on this principle: that students will take knowledge from the instructor and reconstruct their own meaning [64].

5.2 Non-technical skills

Another recommendation is the need for DSE students to learn non-technical skills, an important aspect of a technical education in DSE. Along with the DSE mindset described above, the list of these skills important to data scientists and engineers includes teamwork, communication, intellectual curiosity, domain understanding, and problem solving, among others [12, 38, 39, 45, 54]. Two of the most crucial of these skills are the ability to work in teams and the ability to communicate results and explain findings to a largely non-technical audience [39]. While both are useful skills for any technical professional, communication and "knowledge transfer" skills are crucial for data scientists and engineers. Data scientists and engineers often need to be "storytellers" who can describe results and findings to a non-technical audience in a relatable manner. The Park City Report [25] recommends that undergraduate curricula include communication opportunities across the curriculum, instead of separating them into dedicated courses. This report concurs with that recommendation; information, regardless of delivery method, should be clearly conveyed to non-technical users, and written reports should be geared to an appropriate reading level. However, Hardin et al. [38] raise the concern that because statistics faculty lack the prowess needed to assess technical writing and presentation skills, students may not be receiving sufficient instruction in that area. However, as communication skills need to be developed within the DSE disciplines, DSE faculty, who feel unprepared in this space, should be encouraged to collaborate with faculty in liberal arts to develop proper communication skills for DSE students.

Working in teams is also important for DSE specialists, although in start-ups or small projects the data scientist may have to wear multiple hats. Again, data scientists and engineers must be able to communicate with members in other roles (such as business managers) who are not necessarily as familiar with DSE terminology and methods. Data scientists and engineers may also benefit from project management training, which are usually not thought of as "human skills" but are helpful for timely and efficient completion of DSE projects.

As part of the overall communications ability, DSE students need to make use of visuals extensively to reach broader audiences. It is useful to have visuals to represent the different aspects of the data fields of interest. This paper proposes an initial visual that bisected the data space into two equal regions: one for data engineering and one for data science, as depicted in Figures 3 and 4. The visualization of Figure 4 is not absolute. As data science and engineering evolve, their meaning and their graphic representations will require periodic updates to reflect the fields more precisely.

5.3 Recurring themes

Within the DSE fields, there are currently several themes that recur. Of course these include computational thinking, mathematical thinking and statistical thinking. The data lifecycle shown in Figure 9 should permeate DSE course assignments, with different aspects of the lifecycle given greater attention than others depending on the topic of the exercise.

Technical ethics should pervade the whole approach, covering privacy and confidentiality, for example, and ensuring that bias does not distort interpretation or application of results. As cybersecurity is needed to ensure that data, programs, and systems are protected and not compromised, DSE students must at minimum have a working understanding of cybersecurity issues.

The themes of ethics and cybersecurity should, in our opinion, not be relegated to single courses but woven throughout a DSE curriculum to ensure that students encounter these concepts at multiple points throughout their course of study.

5.4 Fundamental DSE courses

The current situation at colleges and universities suggests that a void may exist in providing a consistent picture of the DSE space that will attract students and help inform their decision on which academic pathway to follow. Educators should address this void,

where it exists, by developing a general course in DSE that follows a standardized curriculum for an introductory data-related course. For example, the course could be named *DSE-0 Fundamentals of Data Science and Engineering*.

Such a course should be accessible to all students, regardless of their specialty. Its purpose should be to acquaint students with the discipline but in a way that emphasizes motivation for studying DS and DE, highlighting interesting insights and developments arising from study of the topic. A follow up to this introductory course would be one for students who intend to specialize in the DSE field; for instance, it might introduce students to the elements of machine learning.

5.5 Resources

A fundamental and important requirement for any institution offering programs in DSE is a group of staff and a range of faculty who have data experience and expertise, and can motivate students; faculty members should have good relationships with faculty from statistics and mathematics as well as traditional electronic and electrical engineering. Other requirements include the need for sufficient hardware to be able to perform DSE-related activities at a scale approximating what graduates will find in industry, and access to data collections (to be updated regularly and collected in diverse ways) for use by students. Datasets should exhibit various characteristics - for instance data to be cleaned, data to be used in exercises, very large date collections that span more than one machine.

In addition, software should include the availability of languages with libraries that support DSE, visualization, statistics, and machine learning as well as tools that support the teaching of mathematics. Relevant online resources include an appropriate range of web services, access to online materials (to be seen as one method of keeping current) such as MOOCs and up-to-date DSE textbooks.

In general, the environment should be supportive and encouraging of study in DSE-related areas, and curricula should include illustrations (to be updated regularly) of machine learning in practice and use. An outward facing perspective can be facilitated by having contacts from industry or elsewhere with those involved in data fields for possible internships, speakers who can motivate students, and so on.

5.6 Shared understanding of DSE degree programs

A growing skills gap exists between the products produced by academia (its graduates) and the needs of business, industry, and government. This phenomenon permeates many DSE areas and beyond. Although universities are not training grounds for industry, they cannot generate DSE curricula in isolation. DSE curricula should reflect market needs so graduates are able to secure meaningful positions upon graduation.

As data is pervasive, few areas are void of data. For the DSE field, this fact underscores how important it is for academic institutions to have ongoing communication with government and industry. All governments in the modern world maintain multitudes of records from which one could extract data. Business and industry are increasingly data-intensive in multiple domains, whether the domain involves payroll, inventory, maintenance, or other aspects of a modern enterprise. By having some connections with industry and government, academic programs and institutions can develop synergistic areas of cooperation and mutual benefit. Employers would learn that DSE expertise needs to come from the entire team, not necessarily be expected to come from just one member.

A consequence of such interaction between academia and industry would be the development of shared expectations of the competencies of DSE graduates upon graduation from an academic program. For example, at the undergraduate level in the United States, ABET has played a role in defining such expectations through accreditation criteria for various computing programs [1]. Establishing accreditation criteria by ABET and other similar bodies in different countries and regions would allow DSE programs to gain the legitimacy needed for the DSE disciplines. Moreover, as DSE are relatively new fields of study, accreditation criteria would help educators and practicing professionals define and evolve these disciplines using approaches that have worked for computing and engineering fields for decades.

ACKNOWLEDGMENTS

This work builds on many of the prior efforts in data science and data engineering education. The authors would like to thank Guido Rößling and Michalis Giannakos for their patience and feedback to earlier drafts.

In addition, Karen Davis acknowledges the support of an undergraduate research assistant, Jon Zanillo, in the section on data science ethics. Teresa Pereira acknowledges the support of FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2019. Rajendra Raj acknowledges support provided by the US National Science Foundation under Awards DGE-1433736 and 1922169.

REFERENCES

[1] ABET, Inc. 2017. Criteria for Accrediting Computing Programs. Effective for Review During the 2017-18 Accreditation Cycle. http://www.abet.org/wp-content/uploads/2016/12/C001-17-18-CAC-Criteria-10-29-16-1.pdf

[2] Accenture Labs. 2016. Building Digital Trust: The Role of Data Ethics in the Digital Age. https://www.accenture.com/us-en/insight-data-ethics. Accessed: 2019-06-29.

[3] Saeed Aghabozorgi. 2016. Data Scientist vs Data Engineer, What's the Difference? https://cognitiveclass.ai/blog/data-scientist-vs-data-engineer/. Accessed: 2019-06-15.

[4] Ethem Alpaydin. 2014. *Introduction to Machine Learning, Third Edition.* The MIT Press, Cambridge, MA.

[5] American Statistical Association. 2018. Ethical Guidelines for Statistical Practice. https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx. Accessed: 2019-06-29.

[6] Data Science Association. 2013. Data Science Code of Professional Conduct. https://www.datascienceassn.org/code-of-conduct.html. Accessed: 2019-06-29.

[7] Association for Computing Machinery. 2018. ACM Code of Ethics and Professional Conduct. https://www.acm.org/code-of-ethics. Accessed: 2019-06-29.

[8] Ulrich Bayer, Paolo Milani Comparetti, Clemens Hlauschek, Christopher Kruegel, and Engin Kirda. 2009. Scalable, behavior-based malware clustering. In *Network and Distributed Systems Security Symposium*, Vol. 9. Citeseer, Internet Society, San Diego, 8–11.

[9] Casey Bennett and Thomas Doub. 2011. Data Mining and Electronic Health Records: Selecting Optimal Clinical Treatments in Practice. In *IEEE International Conference on Data Mining (ICDM)*. IEEE, Vancouver, Canada, 313–318.

[10] Francine Berman, Rob Rutenbar, Henrik Christensen, Susan Davidson, Deborah Estrin, Michael Franklin, Brent Hailpern, Margaret Martonosi, Padma Raghavan, Victoria Stodden, and Alex Szalay. 2016. Realizing the Potential of Data Science: Final Report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data

Science Working Group. https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf.
[11] Verónica Bolón-Canedo, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera. 2014. A review of microarray datasets and applied feature selection methods. *Information Sciences* 282 (2014), 111–135.
[12] Robert J. Brunner and Edward J. Kim. 2016. Teaching data science. *Procedia Computer Science* 80 (2016), 1947–1956.
[13] Emanuelle Burton, Judy Goldsmith, and Nicholas Mattei. 2018. How to teach computer ethics through science fiction. *Commun. ACM* 61, 8 (2018), 54–64.
[14] Alvaro A. Cárdenas, Pratyusa K. Manadhata, and Sreeranga P. Rajan. 2013. Big data analytics for security. *IEEE Security & Privacy* 11, 6 (2013), 74–76.
[15] Ruth C. Carlos, Jr Charles E Kahn, and Safwan S Halabi. 2018. Data Science: Big Data, Machine Learning, and Artificial Intelligence. *Journal of the American College of Radiology* 15 (03 2018), 497–498. https://doi.org/10.1016/j.jacr.2018.01.029
[16] Malcolm Carr, Miles Barker, Beverley Bell, Fred Biddulph, Alister Jones, Valda Kirkwood, John Pearson, and David Symington. 2013. *The Constructivist Paradigm and Some Implications for Science Content and Pedagogy.* Taylor and Francis Group, London, Chapter The Content of Science: A Constructivist Approach to its Teaching and Learning, 159–172.
[17] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.
[18] Cleverism. 2019. Objectives and Responsibilities of the Data Engineer. https://www.cleverism.com/job-profiles/data-engineer/. Accessed: 2019-09-05.
[19] Data Science Community. 2017. College and University Data Science Degrees. http://datascience.community/colleges. Accessed: 2019-07-16.
[20] Coursera. 2019. Data Science Specialization. https://www.coursera.org/specializations/jhu-data-science?. Accessed: 2019-07-14.
[21] Andrea Danyluk, Paul Leidig, Lillian Cassel, and Christian Servin. 2019. ACM Task Force on Data Science: Draft Report and Opportunity for Feedback. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19)*. ACM, New York, NY, USA, 2.
[22] Data4Democracy. 2019. Ethics Resources. https://github.com/Data4Democracy/ethics-resources. Accessed: 2019-06-29.
[23] DataEthics. 2017. Data Ethics Principles. https://dataethics.eu/data-ethics-principles/. Accessed: 2019-06-29.
[24] Thomas H Davenport and D J Patil. 2012. Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review* 90, 5 (2012), 70–76.
[25] Richard D De Veaux, Mahesh Agarwal, Maia Averett, Benjamin S Baumer, Andrew Bray, Thomas C Bressoud, Lance Bryant, Lei Z Cheng, Amanda Francis, Robert Gould, et al. 2017. Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application* 4 (2017), 15–30.
[26] Yuri Demchenko, Adam Belloum, Wouter Los, Tomasz Wiktorski, Andrea Manieri, Holger Brocks, Jana Becker, Dominic Heutelbeck, Matthias Hemmje, and Steve Brewer. 2016. EDISON data science framework: a foundation for building data science profession for research and industry. In *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE, Luxembourg, 620–626.
[27] K. Eric Drexler. 2013. The difference between science and engineering. https://fs.blog/2013/07/the-difference-between-science-and-engineering/. Accessed: 2019-09-09.
[28] Alain Dupuy and Richard M Simon. 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute* 99, 2 (2007), 147–157.
[29] M C Elish and Danah Boyd. 2018. Situating Methods in the Magic of Big Data and AI. *Communication Monographs* 85 (2018), 57–80.
[30] Hugo Jair Escalante, Thamar Solorio, and Manuel Montes-y Gómez. 2011. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies–Volume 1*. Association for Computational Linguistics, Portland, OR, 288–298.
[31] Casey Fiesler. 2018. Tech Ethics Curricula: A Collection of Syllabi. https://medium.com/@cfiesler/tech-ethics-curricula-a-collection-of-syllabi-3eedfb76be18.
[32] Principles for Digital Development. 2017. Principles. https://digitalprinciples.org/principles/. Accessed: 2019-06-29.
[33] Alan Fritzler. 2018. An Ethical Checklist for Data Science. https://dssg.uchicago.edu/2015/09/18/an-ethical-checklist-for-data-science/. Accessed: 2019-06-29.
[34] Vincent Granville. 2017. Difference between Machine Learning, Data Science, AI, Deep Learning, and Statistics. https://www.datasciencecentral.com/profiles/blogs/difference-between-machine-learning-data-science-ai-deep-learning. Accessed: 2019-06-18.
[35] Hillary Green-Lerman. 2018. What is Data Engineering? https://www.datacamp.com/community/blog/data-engineering. Accessed: 2019-09-05.
[36] Peter Grindrod and Juan Bernabe Moreno. 2018. Code of Conduct for Professional Data Scientists. http://www.code-of-ethics.org/. Accessed: 2019-06-29.
[37] Joel Grus. 2015. *Data Science from Scratch: First Principles with Python.* O'Relly Media, Sebastopol, CA. http://my.safaribooksonline.com/97814919-01427
[38] J. Hardin, R. Hoerl, Nicholas J. Horton, D. Nolan, B. Baumer, O. Hall-Holt, P. Murrell, R. Peng, P. Roback, D. Temple Lang, and M. D. Ward. 2015. Data Science in Statistics Curricula: Preparing Students to "Think with Data". *The American Statistician* 69, 4 (2015), 343–353.
[39] Stephanie C Hicks and Rafael A Irizarry. 2018. A guide to teaching data science. *The American Statistician* 72, 4 (2018), 382–391.
[40] Larry O. Natt Gantt II. 2007. Deconstructing Thinking Like a Lawyer: Analyzing the Cognitive Components of the Analytical Mind. *Campbell Law Review* 29 (2007), 413.
[41] International Data Science in Schools Project. 2019. IDSSP: the International Data Science in Schools Project: Abbreviated Topics List. http://www.idssp.org/files/IDSSP_DraftFramework_AbbreviatedLists.pdf. Draft Curriculum Framework.
[42] Indeed. 2019. https://www.indeed.co.uk).
[43] JetBrains. 2018. Data Science Survey. https://www.jetbrains.com/research/data-science-2018/. Accessed: 2019-07-14.
[44] Kaggle. 2017. Kaggle Machine Learning & Data Science Survey 2017. https://www.kaggle.com/kaggle/kaggle-survey-2017. Accessed: 2019-07-14.
[45] Daniel Kaplan. 2018. Teaching Stats for Data Science. *The American Statistician* 72, 1 (2018), 89–96.
[46] Vlado Kešelj, Evangelos Milios, Andrew Tuttle, Singer Wang, and Roger Zhang. 2006. DalTREC 2005 Spam Track: Spam Filtering using N-gram-based Techniques.
[47] Oleksii Kharkovyna. 2019. Who Is a Data Engineer & How to Become a Data Engineer? https://towardsdatascience.com/who-is-a-data-engineer-how-to-become-a-data-engineer-1167ddc12811. Accessed: 2019-09-09.
[48] J Zico Kolter and Marcus A Maloof. 2006. Learning to detect and classify malicious executables in the wild. *Journal of Machine Learning Research* 7, Dec (2006), 2721–2744.
[49] Cassie Korzyrkov. 2018. Top 10 roles in AI and data science. https://www.kdnuggets.com/2018/08/top-10-roles-ai-data-science.html. Accessed: 2019-06-18.
[50] Jess M Krannich, James R Holbrook, and Julie J McAdams. 2008. Beyond thinking like a lawyer and the traditional legal paradigm: Toward a comprehensive view of legal education. *Denv. UL Rev.* 86 (2008), 381.
[51] Maarten Lambers and Cor J Veenman. 2009. Forensic authorship attribution using compression distances to prototypes. In *International Workshop on Computational Forensics*. Springer, The Hague, The Netherlands, 13–24.
[52] Wenke Lee and Salvatore J. Stolfo. 1998. Data Mining Approaches for Intrusion Detection. In *Proceedings of the 7th Conference on USENIX Security Symposium - Volume 7 (SSYM'98)*. USENIX Association, Berkeley, CA, USA, 6–6. http://dl.acm.org/citation.cfm?id=1267549.1267555
[53] Yaniv Leven. 2017. Data Engineer Vs Data Scientist. https://blog.panoply.io/what-is-the-difference-between-a-data-engineer-and-a-data-scientist. Accessed: 2010-06-15.
[54] Adam Loy, Shonda Kuiper, and Laura Chihara. 2019. Supporting Data Science in the Statistics Curriculum. *Journal of Statistics Education* 27, 1 (2019), 2–11.
[55] George A Miller. 1998. *WordNet: An electronic lexical database.* MIT Press, Cambridge, MA.
[56] Robert Muenchen. 2019. The Popularity of Data Science Software. http://r4stats.com/articles/popularity/. Accessed: 2019-04-04.
[57] Engineering National Academies of Sciences and Medicine. 2018. *Data Science for Undergraduates: Opportunities and Options.* The National Academies Press, Washington, DC. https://doi.org/10.17226/25104
[58] Institute of Electrical and Electronics Engineers. 2014. IEEE Code of Ethics. https://www.ieee.org/about/corporate/governance/p7-8.html. Accessed: 2019-06-29.
[59] Jean Piaget. 1954. *The Construction of Reality in the Child.* Basic Books, New York.
[60] The Linux Foundation Projects. 2019. Data Values and Principles. https://datapractices.org/manifesto/. Accessed: 2019-06-29.
[61] Rajendra K. Raj, Allen Parrish, John Impagliazzo, Carol J. Romanowski, Sherif G. Aly, Casey C. Bennett, Karen C. Davis, Andrew McGettrick, Teresa Susana Mendes Pereira, and Lovisa Sundin. 2019. *A Listing of Data Science and Engineering (DSE) Textbooks, circa July 2019.* Technical Report. The American University in Cairo. https://bit.ly/2lUhQJh.
[62] David Reinsel, John Gantz, and John Rydning. 2018. The Digitization of the World from Edge to Core. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Accessed: 2019-11-11.
[63] Karl Rexer. 2017. A Decade of Surveying Analytic Professionals: 2017 Survey Highlights. http://www2.cs.uh.edu/~ceick/UDM/RexerDSSOV17.pdf. Accessed: 2019-07-14.
[64] Virginia Richardson. 2003. Constructivist pedagogy. *Teachers college record* 105, 9 (2003), 1623–1640.
[65] Margaret Rouse. 2016. Definition: data engineer. https://searchdatamanagement.techtarget.com/definition/data-engineer. Accessed: 2019-09-05.

[66] Margaret Rouse. 2019. Data Analytics: definition. https://searchdatamanagement.techtarget.com/definition/data-analytics. Accessed: 2019-09-09.

[67] Aravind Sekar. 2018. What Is The Difference Between Data Science And Machine Learning? https://analyticstraining.com/what-is-the-difference-between-data-science-and-machine-learning. Accessed: 2019-05-25.

[68] Russell Shackelford, Andrew McGettrick, Robert Sloan, Heikki Topi, Gordon Davies, Reza Kamali, James Cross, John Impagliazzo, Richard LeBlanc, and Barry Lunt. 2006. Computing curricula 2005: The overview report. *ACM SIGCSE Bulletin* 38, 1 (2006), 456–457.

[69] Martin A Simon. 1993. *Reconstructing mathematics pedagogy from a constructivist perspective*. ERIC, New York.

[70] Stack Exchange. 2019. Cross Validated. https://stats.stackexchange.com/.

[71] Studyportals B.V. 2019. Studyportals. https://www.studyportals.com/.

[72] Brian Suda. 2018. 2017 Data Science Salary Survey. https://learning.oreilly.com/library/view/2017-data-science/9781491997079/ch04.html. Accessed: 2019-07-14.

[73] Andrew H Sung and Srinivas Mukkamala. 2003. Identifying important features for intrusion detection using support vector machines and neural networks. In *Proceedings of the 2003 Symposium on Applications and the Internet*. IEEE, Orlando, FL, 209–216.

[74] techopedia. 2019. Definition - What does Data Engineer mean? https://www.techopedia.com/definition/33707/data-engineer. Accessed: 2019-09-09.

[75] techopedia. 2019. Definition - What is Data Science? https://www.techopedia.com/definition/30202/data-science. Accessed: 2019-09-09.

[76] Rochelle Tractenberg, Kevin FitzGerald, and Jeff Collmann. 2016. Evidence of sustainable learning from the mastery rubric for ethical reasoning. *Education Sciences* 7, 1 (2016), 2.

[77] Rochelle E. Tractenberg. 2016. Institutionalizing Ethical Reasoning: Integrating the ASA's Ethical Guidelines for Professional Practice into Course, Program, and Curriculum. In *Ethical Reasoning in Big Data: An Exploratory Analysis*, Jeff Collmann and Sorin Adam Matei (Eds.). Springer, Berlin.

[78] Rochelle E Tractenberg and Kevin T FitzGerald. 2012. A Mastery Rubric for the design and evaluation of an institutional curriculum in the responsible conduct of research. *Assessment & Evaluation in Higher Education* 37, 8 (2012), 1003–1021.

[79] John W. Tukey. 1962. The Future of Data Analysis. *The Annals of Mathematical Statistics* 33, 1 (03 1962), 1–67. https://doi.org/10.1214/aoms/1177704711

[80] Nanyang Technological University. 2018. Bachelor of Science in Data Science and Artificial Intelligence. Brochure. http://scse.ntu.edu.sg/Programmes/CurrentStudents/Undergraduate/Documents/2018/DSAIbrochure.pdf

[81] Iliya Valchanov. 2018. Data Science versus Machine Learning versus Data Analytics versus Business Analytics. https://www.kdnuggets.com/2018/05/data-science-machine-learning-business-analytics.html. Accessed: 2019-06-01.

[82] Rakesh Verma, Murat Kantarcioglu, David Marchette, Ernst Leiss, and Thamar Solorio. 2015. Security analytics: essential data analytics knowledge for cybersecurity professionals and students. *IEEE Security & Privacy* 13, 6 (2015), 60–65.

[83] Rakesh Verma, Narasimha Shashidhar, and Nabil Hossain. 2012. Phishing email detection the natural language way. In *Computer Security – ESORICS 2012*, Sara Foresti, Moti Yung, and Fabio Martinelli (Eds.). Springer, Berlin, Heidelberg, 824–841.

[84] Lev Semenovich Vygotsky. 1980. *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.

[85] Jeannette M Wing, Vandana P Janeja, Tyler Kloefkorn, and Lucy C Erickson. 2018. Data Science Leadership Summit: Summary Report.

[86] Ge Yu. 2019. The Core Courses of Data Science and Big Data Technology: The Computing in Data Science. In *ACM TURC 2019 (SIGCSE China)*. ACM, Chengdu, China, 1.