Journal of Behavioral Data Science, 2021, *1* (1), 1–16.
DOI: https://doi.org/10.35566/jbds/v1n1/p1

What is Data Science? An Operational Definition based on Text Mining of Data Science Curricula

Zhiyong Zhang[1] and Danyang Zhang[2]

[1] University of Notre Dame
zzhang4@nd.edu
[2] University of Texas–Austin
danyang.zhang@utexas.edu

Abstract. Data science has maintained its popularity for about 20 years. This study adopts a bottom-up approach to understand what data science is by analyzing the descriptions of courses offered by the data science programs in the United States. Through topic modeling, 14 topics are identified from the current curricula of 56 data science programs. These topics reiterate that data science is at the intersection of statistics, computer science, and substantive fields.

Keywords: Data Science · Topic Modeling · Data Science Curriculum.

1 Introduction

Data science has been a buzzword in the past two decades. However, the exact meaning of data science has never been clear. In a statement by American Statistical Association (ASA), it states "there is not yet a consensus on what precisely constitutes data science" (Van Dyk et al., 2015). Hayashi (1998) is probably the first formal attempt to define data science although the history of data science practice is considerably longer (Donoho, 2017; Tukey, 1962).[1] He argued that "the aim of data science is to reveal the features or the hidden structure of complicated natural, human and social phenomena with data" and data science consists of "design for data, collection of data, and analysis on data." To many, this sounds like the characteristics of applied statistics (e.g., Broman, 2013; Silver, 2013). Not surprisingly, some have also argued that data science is actually different from statistics. For example, Dhar (2013) pointed out that data science is different from statistics in terms of data types and skills required. The

[1] Tukey has used the term "data analysis" in his writing that is conceptually similar to what data science does. Naur (1966) coined the term "datalogy" to call "the science of the nature and use of data" and Naur (1974) provided a more detailed treatment of data largely from a computer science perspective.

current view of data science aligns more closely with what Cleveland (2001) has described – data science consists of 25% Multidisciplinary Investigations, 20% Models and Methods for Data, 15% Computing with Data, 15% Pedagogy, 5% Tool Evaluation, and 20% Theory. Regardless of how it is perceived, data science is now widely accepted as its own paradigm (Hey, Tansley, & Tolle, 2009).

Many data science degree programs emerged in the past few years. The Institute for Advanced Analytics (IAA) at North Carolina State University tracks the master's degrees in Data Science at universities based in the United States. By its count, there are 78 data science programs in 2020.[2] From a practical point of view, it is probably more informative to understand what data science offers and what it is constituted than its exact definition that might not be possible at all. In the same ASA statement, Van Dyk et al. (2015) identified three foundations to data science:

> (i) *Database Management enables transformation, conglomeration, and organization of data resources.*
> (ii) *Statistics and Machine Learning convert data into knowledge.*
> (iii) *Distributed and Parallel Systems provide the computational infrastructure to carry out data analysis.*

In a review of the history of data science, Donoho (2017) coined the "Greater Data Science" field with six divisions: data exploration and preparation, data representation and transformation, computing with data, data modeling, data visualization and presentation, and science about data science. Some empirical studies also tried to understand what skills and knowledge are needed in jobs (e.g., Cegielski & Jones-Farmer, 2016) and taught in degree programs (e.g., Gorman & Klimberg, 2014; Song & Zhu, 2016).

More recently, Fayyad and Hamutcu (2020) proposed a "Data Science Knowledge Framework" aiming to support industry standardization and building measurement and assessment methodologies for data science professionals. The framework identified two domains in analytics and data science: Science and Math, and Programming and Technology. Within the Science and Math domain, they identified the following seven fields: Scientific Method, Mathematics, Computer Science, Statistics, Operations Research & Optimization, Data Preparation and Exploration, and Machine Learning. The Programming and Technology domain has four fields: General Purpose Computing, Scientific Computing, Database & Business Intelligence, and Big Data. Fayyad and Hamutcu (2020) also provided a list of subjects for each field with example topics.

However, Fayyad and Hamutcu (2020) did not provide much empirical support to their knowledge framework. The existing empirical studies (e.g., Gorman & Klimberg, 2014; Song & Zhu, 2016) on data science curricula were conducted several years ago without considering the newly emerged programs. The goal of this study is to empirically examine the current data science programs to hopefully better understand and define what data science is. The rest of this paper

[2] We consider the data analytics and business analytics programs as different from data science programs.

is structured as follows. In Section 2, we present our data collection method and data analysis procedure. In Section 3, we report the results from our data analysis. In Section 4, we discuss our findings.

2 Methods

2.1 Data Collection

IAA keeps an up-to-date track of graduate degree programs in analytics, business analytics, and data science offered in the US.[3] From it, we identified 74 programs from 74 universities, one program from each university, with the term "data science" in their names.[4] The actual names of the programs have 17 different varieties such as M.S. in Data Science, M.S.E. in Data Science, M.S. in Computational Data Science, and M.S. in Data Science and Business Analytics. Many data science programs offer different concentrations. For example, Depaul University started its M.S. in Data Science program in 2010 and now has four concentrations: Computational Methods, Health Care, Hospitality, and Marketing.

For each program, we looked through its website and downloaded the information on the courses offered and the description of each course in one of the following two ways. For the majority of the programs, we used Python to download the course information automatically. For the rest, we saved the information manually.

2.2 Data Preprocessing

The 74 programs offered a total of 2,022 courses after removing the same courses listed in different concentrations by the same programs. Different programs can offer the courses with the same names. For example, *Machine learning*, *Data visualization*, and *Data mining* are offered by 28, 19, and 18 programs, respectively, with the exact same names. However, the contents taught in these courses can be different. Only 58 of the 74 programs provided descriptions of the courses they offered at the time of our data collection. In total, 1,383 courses were found to have description information. For some courses, the descriptions were very brief. For example, for one course *Scripting Languages*, its description was "Survey of current business analytics scripting languages." In this study, we removed such courses with short or uninformative descriptions, which eventually resulted in 1,276 courses from 56 programs.

Typical text data preprocessing steps (e.g., Hickman, Thapa, Tay, Cao, & Srinivasan, 2020; Vijayarani, Ilamathi, Nithya, et al., 2015) were taken to prepare the course descriptions for further analysis. First, all words were converted to lower cases and all numbers were removed. Second, we replaced abbreviations

[3] https://analytics.ncsu.edu/?page_id=4184

[4] After our data analysis, IAA added four more programs from Old Dominion University, University of Colorado Boulder, University of Miami, and Utah State University.

such as "GIS" with "geographic information systems", and "ML" with "maximum likelihood" so that the same forms of terms were used in all descriptions. Third, we combined some terms with the same or similar meaning such as both "C" and "C++" to "cprogram" and "SQL", "MySQL" and "NoSQL" to "sql". However, we did not conduct word-stemming except for changing all the words in the plural forms to their singular forms because different forms of the words might have different meanings. Fourth, we removed common stopwords such as "a", "the", "about", and "very". Some frequently used words such as "students", "semester", and "assignment" in course descriptions were not conventionally considered as stopwords. However, they did not provide useful information and, therefore, were removed before analysis.

2.3 Data Analysis

With the preprocessed data, we conducted both term frequency analysis and topic modeling.

2.3.1 Term Frequency Analysis. We first tokenized the course descriptions into individual words and analyzed the frequency of each word. A large frequency shows that a word is more frequently used in the course descriptions and indicates the importance of a topic that the word is associated with. After that, we investigated the frequency of short phrases including two-word phrases such as "data mining" and "machine learning", three-word phrases such as "natural language processing" and "support vector machine", and four-word phrases such as "Markov chain Monte Carlo" and "relational database management system".

2.3.2 Topic Modeling. Topic modeling or topic models can be used to investigate the topics and associated words through mining text information. We used topic modeling to identify the common topics in courses offered in data science programs. Latent Dirichlet allocation (LDA; Blei, Ng, & Jordan, 2003) is probably the most widely used method in topic modeling that allows the observed text, in our case the course descriptions, to be explained by latent topics. In LDA, each course description can be assumed as a mixture of a small number of topics, and each word's presence in the description is associated with one of the topics.

One may argue that the name of a course would summarize the main topic of the course. However, it is not necessarily the case based on our quick analysis of the course descriptions. For example, a course was named "Advanced Data Analysis." First, the name itself was not informative. Second, its description included topics on "data visualization techniques", "dimension reduction techniques", and the use of "computer packages." As we will show, these can be viewed as three different topics. Through LDA, we explored how many common topics the courses from many different data science programs cover.

Suppose there are a total of K topics in all courses. For a given course, it can consist of one or all of the K topics with different probabilities. Let z_{km} be the

kth ($k = 1, \ldots, K$) topic in the mth ($m = 1, \ldots, M$) course. z_{km} takes a value between 1 and K. The topic from which a word n associated with is assumed to be generated from categorical distribution

$$z_{mn} \sim Cat(\boldsymbol{\theta}_m)$$

with the topic probability $\boldsymbol{\theta}_m = (\theta_{m1}, \theta_{m2}, \ldots, \theta_{mK})'$ for the course m. Note that $\sum_{k=1}^{K} \theta_{mk} = 1$. For example, if there are two topics, $K = 2$. Let $w_{mn}, n = 1, \ldots, N_m, m = 1, \ldots, M$, be the nth word and N_m be the total number of words in the mth course description. w_{mn} would take a value between 1 and V with V being the total number of unique words used in all the course description. LDA specifies that

$$w_{mn}|z_{mn} = k \sim Cat(\boldsymbol{\beta}_k)$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kV})'$ is the probability that a word is used given the topic k is discussed in a course.

The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ in LDA models are typically not known and need to be estimated. Both frequentist and Bayesian methods are available to estimate the parameters. For example, Blei et al. (2003) proposed both efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. In this study, we used the Bayesian method based on Gibbs sampling for our data analysis.

In topic modeling, the number of topics is often unknown and needs to be determined. In this study, c-fold cross-validation (CV) was used. The basic idea of CV is to divide the data into c folds, or c subsets. Each time, one uses $c-1$ folds of data to fit a topic model and then uses the left out fold of data to calculate the statistic–perplexity–to evaluate the model fit (Grün & Hornik, 2011). This can be done for different numbers of topics and the model with the close-to-smallest perplexity can be chosen as the one with the optimal number of topics.

Although the LDA was initially conducted on individual words, research has shown that including phrases of a sequence of words can lead to improved topic quality (e.g., Lau, Baldwin, & Newman, 2013; Nokel & Loukachevitch, 2015). Therefore, in our study, we included individual words, two-word phrases, and three-word phrases in our topic model.

3 Results

3.1 Term Frequency

The word cloud in Figure 1 displays the 167 words that were used at least 50 times in the descriptions of the 1,276 courses after removing stopwords. Not surprisingly, the most widely used word was "data", for a total of 2,322 times. The words "analysis", "model", "method", "learning", and "system" were each used more than 500 times. Other commonly used words include "algorithm",

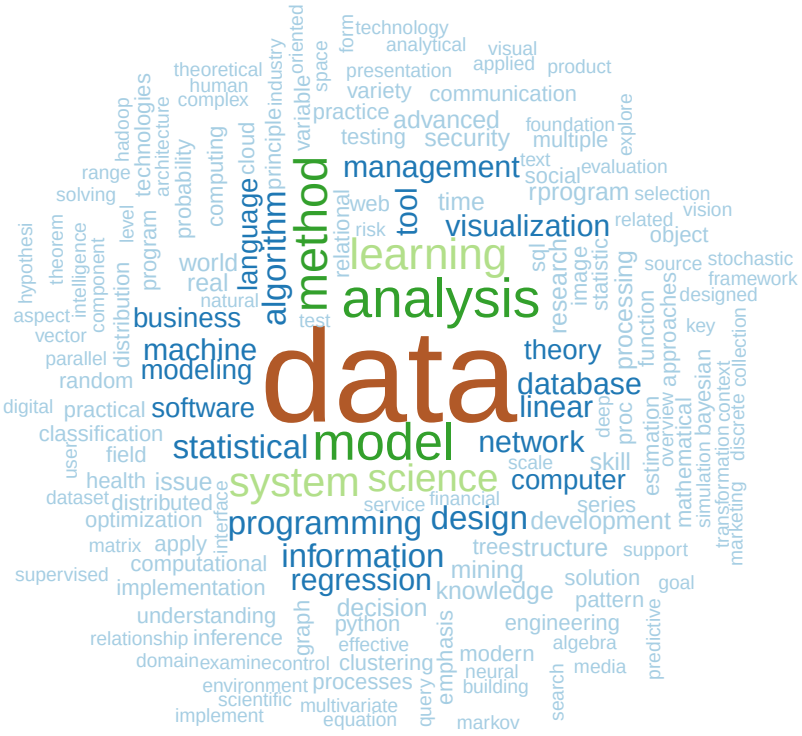


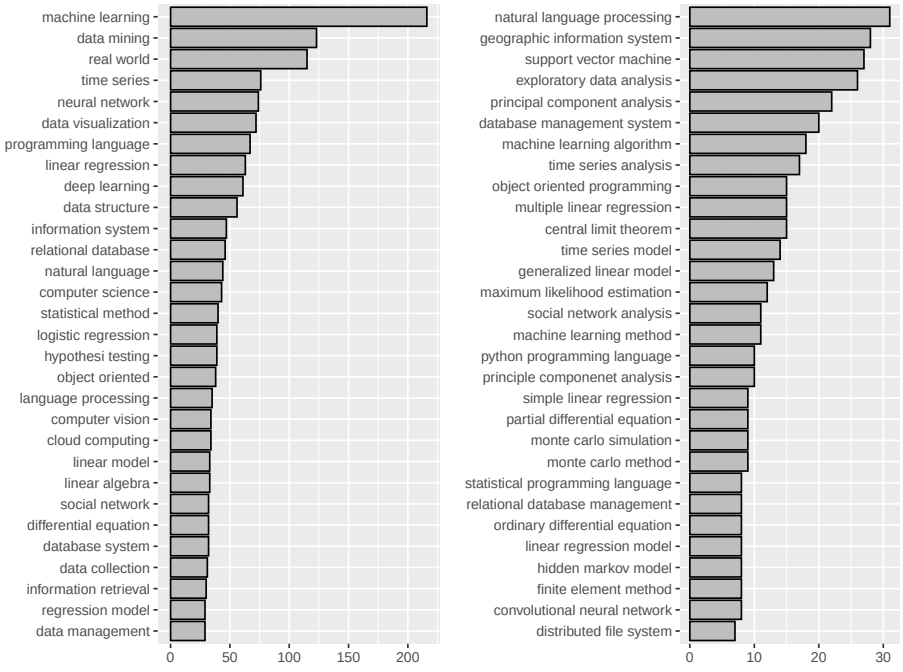
Figure 1: Frequency of individual words used in all course descriptions

"regression", "modeling", "programming", and "visualization". From the frequency analysis, we can see the basic focuses of the data science courses in this study were data analysis, modeling, and data visualization.

Figure 2 shows the top 30 most frequently used two-word phrases and three-word phrases in all course descriptions. Each two-word phrase was used at least 29 times and each three-word phrase was used at least 7 times. These phrases are generally names of analytical models and methods such as "machine learning", "data mining", "linear regression", "natural language processing", and "principal component analysis". We also looked into four-word phrases but only found several to be informative including "markov chain monte carlo" (8 times), "relational database management system" (8 times), "unsupervised machine learning algorithm" (3 times), and "multivariate time series analysis " (3 times).

3.2 Topic Modeling

To determine the number of topics, we first conducted a 5-fold cross-validation. We varied the number of topics from 2 to 25 and calculated the perplexity of the model with a given number of topics. The R package topicmodels (Grün &



(a) Two-word phrases (b) Three-word phrases

Figure 2: Most frequently used phrases

Hornik, 2011) we used for LDA estimation was sensitive to the seed used for the Gibbs sampling algorithm. Therefore, we tried 100 different seeds to get 100 sets of results and then evaluated each set of result to get the best number of topics. Figure 3 shows the perplexities of a topic model with different numbers of topics based on one seed. From it, the model with 14 topics had the smallest perplexity and the perplexity seemed to flatten out after 14 topics. For this particular seed, we would conclude that the best model was the one with 14 topics. Using the perplexity plot, we identified the number of topics for all 100 sets of analyses. Among the 100 sets of analyses, the models with 11, 12, 13, 14, 15, 16, 17, and 18 topics were best models for 1, 7, 35, 27, 20, 4, 5, and 1 times, respectively, based on the perplexity. Therefore, it suggested a model between 13 to 15 topics was probably the best for the course descriptions. We then fitted the models with 13, 14, and 15 topics and investigated the terms and courses associated with each topic. All considered, we found that the model with 14 topics gave a clear representation of different topics and therefore based our discussion on the model with 14 topics.

To understand what each of the 14 topics represented, we first studied the top 30 words and phrases associated with that topic. The topic words were

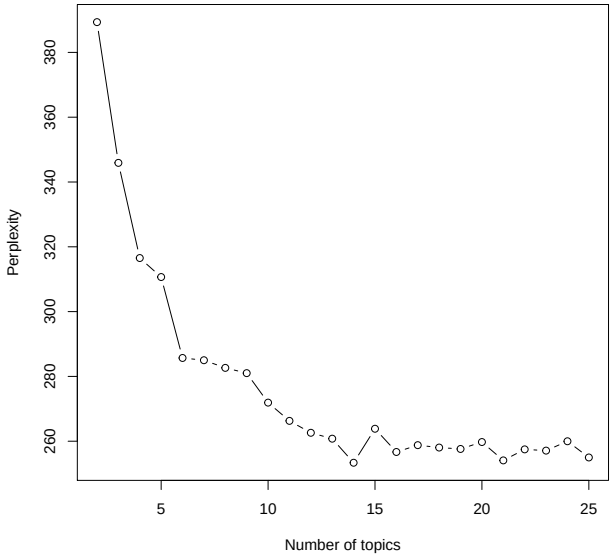


Figure 3: Perplexities of the topic model with different numbers of topics

selected based on the term-scores proposed by Blei and Lafferty (2009). Then, we assigned each course to the highest likely topic and looked through the names of the courses. Based on the analysis, we named the 14 topics, each of which can be viewed as a course to be taught in a data science program. We also identified the commonly taught subjects in each topic/course. We now discuss each of the topics in terms of the most frequently used words and phrases. We will also provide four example classes on each topic from the data science programs analyzed in our study.

Topic 1. Ethics, Privacy, and Security. The first topic is related to research ethics, privacy issues, and data security.[5] The top relevant words and phrases associated with the topic include information, management, security, system, technology, collection, risk, information system, policy, privacy, spatial, ethical, law, ethic, data collection, geographic information system, data management, change, impact, individual, market, access, cost, technical, environment, managing, operation, quantitative, internet, and document. A course in this topic can discuss subjects such as ethics and policy in data analysis, information policy and ethics, data privacy and security, particularly in security and governance of big data, and cyber data security and policy. Example classes include *Behind the*

[5] Note that the order of the named topics might not be the same as the output of the topic modeling in R and does not reflect the relative importance of the topics.

Data: Humans and Values, *Ethics of Big Data*, *Cyber Security Law & Policy*, and *Ethics, Privacy, Security and Governance of Big Data*.

Topic 2. Database Structure and Database Management. The second topic is on database/data structure and database/data management. The top relevant words and phrases associated with the topic include database, system, relational, sql, distributed, parallel, query, architecture, hadoop, relational database, processing, structured, database system, management, mapreduce, memory, transaction, unstructured, storage, query language, database design, management system, database management, file, warehousing, database management system, physical, unstructured data, managing, and selected. A course in this topic can discuss subjects such as different types of database systems, different types of data, database processing and information retrieval, database management systems, big data, and data warehousing. Example classes include *Big Data and NoSQL Program*, *Large-Scale Database Systems*, *Principles of Database Management Systems*, and *Databases and Data Management*.

Topic 3. Data Visualization. The third topic is mainly about visualization, graphical display of data, and exploratory data analysis. The top relevant words and phrases associated with the topic include visualization, tool, principle, communication, data visualization, effective, explore, visual, exploratory, graphic, interactive, insight, exploratory data, exploratory data analysis, critical, perception, technical, apply, aspect, dataset, biology, goal, complex, driven, finding, human, trend, quantitative, environment, and graphical. A course in this topic can discuss subjects such as data visualization techniques and tools, data preparation and preprocessing methods, and types of statistical graphs. Example classes include *Data Visualization*, *Information Visualization and Infographics*, *Visualization of Complex Data*, and *Data Presentation and Visualization with R*.

Topic 4. Algebra. The fourth topic mainly concerns algebra and optimization methods. The top relevant words and phrases associated with the topic include linear, system, function, space, component, matrix, transformation, vector, form, algebra, decomposition, map, reduction, properties, element, linear algebra, spectral, principal, rprogram, dimensional, standard, clustering, cross, dimension, computation, finding, theoretical, equation, primary, principal component analysis. A course in this topic can discuss subjects such as linear and matrix algebra, and numerical methods. Example classes include *Numerical Linear Algebra*, *Computational Algebra*, *Linear Programming*, and *Matrix Algorithms for Data Science*.

Topic 5. Mathematical Foundations and Modeling. This topic is on foundation of mathematics and mathematical modeling. The top relevant words and phrases associated with the topic include theory, mathematical, optimization,

processes, simulation, financial, discrete, stochastic, finance, economic, engineering, modeling, equation, numerical, differential, integration, procedure, differential equation, classical, operation, calculus, transform, continuous, generation, complexity, dynamic, function, complex, control, and matlab. A class on this topic would focus on basic knowledge and foundations of mathematics, optimization methods, and mathematical modeling. Example courses include *Fundamentals of Computational Mathematics*, *Mathematical Modeling*, *Mathematics for Data Scientists*, and *Simulation & Optimization*.

Topic 6. Probability Theory and Statistical Inference. This topic is about basic probability theory and statistical inference. The top relevant words and phrases associated with the topic include statistical, statistic, distribution, estimation, probability, inference, testing, random, bayesian, hypothesis, sampling, variance, sample, hypothesis testing, variable, likelihood, interval, maximum, conditional, parameter, nonparametric, bayes, maximum likelihood, measure, statistical method, limit, prior, statistical inference, confidence, and statistical analysis. A course in this topic would focus on traditional probability and inference topics such as different types of distributions, random variables, sampling distributions, hypothesis testing, and maximum likelihood method. Example classes include *Mathematical Statistics*, *Probability and Statistics for Data Science*, *Bayesian Statistics*, and *Statistical Inference for Data Science*.

Topic 7. Statistical Models. This topic focuses on different types of statistical models for data analysis. The top relevant words and phrases associated with the topic include model, regression, time series, multiple, linear regression, selection, linear, variable, simple, statistical, logistic, forecasting, logistic regression, linear model, parametric, response, regression model, factor, generalized, experimental, interpretation, time series analysis, hierarchical, modeling, multiple linear regression, comparison, sequence, nonlinear, statistical method, and classical. A course in this topic would discuss different types of statistical models such as linear and generalized linear models. Example classes include *Linear Models for Data Science*, *Multivariate Data Analysis*, *Applied Regression Analysis*, and *Experimental Design*.

Topic 8. Statistical Software and Programming. This topic is related to statistical software and basic programming for data analysis. The top relevant words and phrases associated with the topic include programming, algorithm, structure, graph, python, programming language, data structure, tree, matching, flow, efficient, dynamic, complexity, sequence, sorting, object oriented programming, matlab, framework, algorithmic, driven, advanced, code, operation, dataset, package, internet, ethical, measurement, program, and single. A course in this topic could teach how to use software such as R, MATLAB, and Python for data analysis, software programming, and data computing. Example classes include *Statistical Programming in R*, *Systems and Technologies: Python*, *Python for Data Analysis*, and *SAS Programming*.

Topic 9. Machine Learning and Deep Learning This topic is about machine learning and deep learning methods and techniques. The top relevant words and phrases associated with the topic include learning, machine, machine learning, deep, neural network, neural, deep learning, supervised, unsupervised, classification, clustering, tree, artificial, support vector, unsupervised learning, support vector machine, learning algorithm, reinforcement learning, feature, graphical, reinforcement, learning method, decision tree, supervised learning, training, support, mean, dimensionality, machine learning algorithm, and recognition. A course on this topic would introduce different machine learning and deep learning methods and techniques. Example classes include *Neural Networks and Deep Learning*, *(Applied) Machine Learning*, *Machine Learning and Big Data*, and *Deep Learning*.

Topic 10. Business Analytics and Data Mining. This topic is about data mining and business intelligence techniques and methods. The top relevant words and phrases associated with the topic include business, decision, mining, data mining, modeling, intelligence, pattern, predictive, classification, marketing, support, prediction, discovery, identify, association, customer, domain, bioinformatics, tool, healthcare, clustering, organizational, implementing, organization, enterprise, life, topic, descriptive, implement, and exploration. Such a course would be different from a course on machine learning and deep learning in terms of the subjects taught. Example classes include *Data Mining*, *Financial Data Mining*, *Business Analytics and Data Mining*, and *Business Analytics Fundamentals*.

Topic 11. Network Analysis and Text Mining. This topic is about network analysis and text mining/natural language processing. The top relevant words and phrases associated with the topic include network, language, social, web, text, natural, processing, human, search, media, retrieval, natural language, interaction, relationship, social network, natural language processing, language processing, information retrieval, topic, social media, document, probabilistic, indexing, extraction, graph, measure, standard, business, algorithm, and generation. A course on this topic can teach different types of network models, text mining, and graph theory. Example classes include *NLP: Computational Models of Social Meaning*, *Natural Language Processing*, *Text Mining*, and *Social Network Analysis*.

Topic 12. Cloud Computing and Big Data Analysis. This topic is on computing in the cloud and analysis of big data. The top relevant words and phrases associated with the topic include real, computing, world, program, cloud, rprogram, practical, industry, technologies, apply, scale, platform, cloud computing, real world data, dataset, life, aspect, framework, infrastructure, manipulation, cluster, language, cleaning, storage, computation, experimental, survey, internet, statistical method, and quantitative. A class on this topic would focus on how to conduct cloud computing and how to mining data in the cloud. Example courses

include *Cloud Computing*, *Big Data Technologies*, *Big Data Analysis*, and *Cloud Computing for Data Analytics*.

Topic 13. Software Design and Software Engineering. This topic is about software design and software engineering. The top relevant words and phrases associated with the topic include design, software, advanced, implementation, object, user, oriented, control, interface, implement, engineering, common, level, survey, art, software development, package, effect, quality, cycle, code, libraries, experiment, strategies, environment, access, model, exploration, relationship, and real. A course on this topic can teach how to design and develop software, software environment and fundamentals of programming. Example courses include *Programming for Data Science*, *Software Engineering*, *Introduction to Software Development*, and *Computer Systems Programming*.

Topic 14. Applications. This topic is related to the application of data science in different disciplines particularly heath. A notable area is computer vision and image processing. The top relevant words and phrases associated with the topic include computer, computational, image, health, field, foundation, vision, digital, processing, computer science, detection, public, level, goal, medical, theoretical, biological, domain, object, recognition, limited, measurement, extraction, quantitative, interpretation, discipline, feature, organization, filtering, and interpret. A course on this topic may focus on a particular area of applications. Example courses include *Computer Vision*, *Health Data Science*, *Introduction to Biomedical Informatics*, and *Genomics Analytics*.

4 Discussion

Through the analysis of the descriptions of more than 1,200 courses from 56 data science programs offered in the United States, we identified 14 topics or themes that are common in data science training. They are Ethics, Privacy, and Security, Database Structure and Database Management, Data Visualization, Algebra, Mathematical Foundations and Modeling, Probability Theory and Statistical Inference, Statistical Models, Statistical Software and Programming, Machine Learning and Deep Learning, Business Analytics and Data Mining, Network Analysis and Text Mining, Cloud Computing and Big Data Analysis, Software Design and Software Engineering, and Applications. All 14 topics contributed about equally to the contents of all the courses analyzed in the study, with Probability Theory and Statistical Inference contributing the most, 7.39% and Algebra the least, 6.94%.

Data science training or even a single course is often an integrated unit. Therefore, the 14 topics are more or less related and can share the same contents, as reflected in terms associated with the topics. For example, when teaching the discipline-specific applications, it cannot be avoided to discuss data mining and machine learning techniques, data visualization, and data management to

demonstrate their utilization. The 14-topic model in our study includes the following two topics – "Business Analytics and Data Mining" and "Machine Learning and Deep Learning". Although the two topics shared some same subjects, Business Analytics and Data Mining seemed to focus more on traditional big data techniques often developed in the statistics discipline such as classification and regression tree, mixture model, and discriminant analysis as well as business intelligence. Machine Learning and Deep Learning, on the other hand, covered more topics developed in the computer science discipline such as different types of learning methods, neural network, pattern recognition, and support vector machine techniques. Similarly, we identified a topic on Statistical Software and Programming as well as a topic on Software Design and Software Engineering. The former focused more on the use of software such as R, Python, and MATLAB for practical data analysis and the latter concerned more about software development.

Although fourteen topics provided the best result for our topic model based on cross-validation, the fourteen topics did not necessarily cover all the topics offered in all data science programs analyzed in the study. For example, in the process of understanding the meaning of each topic, we found that Computer Vision stood out as an important topic taught in the data science programs. In addition, some of the topics might be split into multiple topics. For example, the topic of Business Analytics and Data Mining can be split into two. Network Analysis and Text Mining can also be viewed as two separated but closely related topics.

Among the fourteen topics, Algebra, Mathematical Foundations and Modeling, Probability Theory and Statistical Inference, Statistical Models, and Statistical Software and Programming are arguably the traditionally strong areas of the discipline of statistics. Database Structure and Database Management, Machine Learning and Deep Learning, Network Analysis and Text Mining, Cloud Computing and Big Data Analysis, and Software Design and Software Engineering can be viewed as emerging and important areas in the discipline of computer science. Data Visualization and Data Mining have been the focuses of both statistics and computer science disciplines. Ethics, Privacy, and Security is becoming an important topic in both disciplines. The fourteen topics together speak unequivocally that data science is an interdisciplinary area that integrates statistics, computer science, and substantive fields (Applications).

We have chosen to focus on fourteen topics in the analysis. If only thirteen topics were kept, the topic "Software Design and Software Engineering" would drop out. On the other hand, if fifteen topics were used, then "Network Analysis" and "Text Mining" can be broken into two topics.

Although we arrived at the identified topics through empirical analysis of course descriptions, these topics aligned well with the existing literature. Particularly, they were consistent with the Data Science Knowledge Framework by Fayyad and Hamutcu (2020). Our results also reflected the finding by Gorman and Klimberg (2014). In their study, they analyzed the curriculum of 17 business analytics programs and interviewed 11 programs. The 14 subjects that

they identified, such as Introduction to Statistics, Regression, Multivariate, and Time Series Analysis, seemed to mostly fall within the scope of applied statistics. However, they also pointed out three trending developments including Big Data: Internet of Things, Unstructured Data and Semantic Analysis, and Network Analytics. In another study, Song and Zhu (2016) investigated both undergraduate (7 in total) and graduate (15 in total) curricula in data sciences and proposed several approaches for data science educations. The topics identified in our study can be combined with their approaches. Overall, our study provides additional empirical support to the existing literature for understanding what is data science.

4.1 Limitations

Our study has several limitations. First, the analysis used data only from the programs with "Data science" in their titles. There are many data analytics and business analytics programs tracked by IAA. In practice, the differences in "Data science" and "Business analytics" might not be large. It can be interesting to see whether the course information from the programs can be combined for more comprehensive data analysis. Second, the findings in this study were based on the analysis of course descriptions from data science programs in academic institutes. However, the results may or may not align with industry/applied/business applications of data science. In the future, the results can be compared to the analysis of other information, such as job postings for data science positions, to identify potential similarities and differences between academic training in data science and data science as practiced in industry. Third, we only considered the data science programs in the US. The findings may not be generalized to other countries.

4.2 Conclusion

The goal of this study is to understand what data science is through the mining of the courses offered by data science programs in the US to hopefully provide a better definition of data science. We adopted a bottom-up approach to mining the description information of individual courses taught in current data sciences programs. Although we identified fourteen topics among all the courses, it is still difficult to provide a concise and conclusive definition of data science. However, we believe our results can provide useful information on how to operate data science programs. The results of our study further reiterate the notion that data science is at the intersection of statistics, computer science, and applications. A major contribution of our study is to provide empirical support to a better understanding of data science.

Acknowledgment

This study is partly supported by a grant from the Department of Education (R305D210023). However, the contents of the study do not necessarily represent

the policy of the Department of Education, and you should not assume endorsement by the Federal Government. We thank Wen Qu and Tyler Wilcox for their helpful comments and suggestions that improved the study.

References

Blei, D. M., & Lafferty, J. D. (2009). Text mining: Classification, clustering, and applications. In A. Srivastava & M. Sahami (Eds.), (pp. 71–93). Chapman & Hall/CRC.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022. doi: https://doi.org/10.5555/944919.944937

Broman, K. W. (2013). *Data science is statistics.* Retrieved from https://kbroman.wordpress.com/2013/04/05/data-science-is-statistics/ (Retrieved on Mar 12, 2021)

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

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, *69*(1), 21–26. doi: https://doi.org/10.1002/sam.11239

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

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

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

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

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30. doi: https://doi.org/10.18637/jss.v040.i13

Hayashi, C. (1998). What is data science? fundamental concepts and a heuristic example. In *Data science, classification, and related methods* (pp. 40–51). Springer.

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Microsoft research Redmond, WA.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2020). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 1-33. doi: https://doi.org/10.1177/1094428120971683

Lau, J. H., Baldwin, T., & Newman, D. (2013). On collocations and topic models. *ACM Transactions on Speech and Language Processing (TSLP)*, *10*(3), 1–14. doi: https://doi.org/10.1145/2483969.2483972

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

Naur, P. (1974). *Concise survey of computer methods*. Petrocelli Books.

Nokel, M., & Loukachevitch, N. (2015). A method of accounting bigrams in topic models. In *Proceedings of the 11th workshop on multiword expressions* (pp. 1–9).

Silver, N. (2013). *What i need from statisticians.* Retrieved from https://www.statisticsviews.com/article/nate-silver-what-i-need-from-statisticians/ (Retrieved on 3/12/2021)

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

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, *33*(1), 1–67. doi: https://doi.org/10.1214/aoms/1177704711

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

Vijayarani, S., Ilamathi, M. J., Nithya, M., et al. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.