DATA SCIENCE CURRICULUM: CURRENT SCENARIO

Duaa Bukhari

Taibah University, Almadinah Almunawarah, Saudi Arabia

ABSTRACT

Companies desires for making productive discoveries from big data have motivated academic institutions offering variety of different data science (DS) programs, in order to increases their graduates' ability to be data scientists who are capable to face the challenges of the new age. These data science programs represent a combination of subject areas from several disciplines. There are few studies have examined data science programs within a particular discipline, such as Business (e.g. Chen et al.). However, there are very few empirical studies that investigate DS programs and explore its curriculum structure across disciplines. Therefore, this study examines data science programs offered by American universities. The study aims to depict the current state of data science education in the U.S. to explore what discipline DS programs covers at the graduate level. The current study conducted an exploratory content analysis of 30 DS programs in the United States from a variety of disciplines. The analysis was conducted on course titles and course descriptions level. The study results indicate that DS programs required varying numbers of credit hours, including practicum and capstone. Management schools seem to take the lead and the initiative in lunching and hosting DS programs. In addition, all DS programs requires the basic knowledge of database design, representation, extraction and management. Furthermore, DS programs delivered information skills through their core courses. Moreover, the study results show that almost 40 percent of required courses in DS programs is involved information representations, retrieval and programming. Additionally, DS programs required courses also addressed communication visualization and mathematics skills.

KEYWORDS

Data Science, Information Retrieval, Curricula, Master's Programs, DS curriculum.

1. INTRODUCTION

Data science (DS) is one of the fastest-growing area in many countries around the world. This emerging field combines elements of many disciplines such as mathematics, statistics, computer science, and knowledge in a particular application domain, therefore, the data since is considered as interdisciplinary field. This combination of multidiscipline helps extracting meaningful information from the increasingly sophisticated array of data available in many settings. The National Academies of Sciences, Engineering, and Medicine [1] indicated that, massive investments have gone into building out wireless infrastructure and data centres (the cloud) and into leveraging such facilities. In addition, new methods have been developed to connect and understand the data being generated. According to Marr [2], there are 2.5 quintillion bytes of data created each day. And this massive volume of data is increasing every minute particularly with growth of the Internet of Things. Therefore, the challenge and struggle became real in order to deal with this exploding amount of data, which require highly trained professional data scientists.

The need for data scientists is rising and traditional courses and programs offered by statistics departments are not meeting the needs of those seeking training [3]. As a result, academic institutions have started to offer revised and new data science programs to produce qualified graduates. In order to explore these revised programs, this study investigates DS programs and its

DOI:10.5121/ijdkp.2020.10301

curricula in the USA institutions, to reveal, what the required core is, how may credit is required? and what school is hosting the DS program. This study addresses two questions. 1: What does the required core cover in the DS curriculum? 2: How IR is involved in DS master's programs? The study conducts a content analysis of the curricula of the DS master's programs in the United States academic institutions to explore what discipline DS programs covers at the graduate level and discuss how DS curricula relate or connect with the field of information retrieval.

2. Background and Literature Review

We live today in the era of big data, and professionals across industries as well as the educated citizenry have a need to understand and gain insights from the massive amount of data generated from various sources, including science, health care, education, social media, and governmental and non-governmental organizations. Digital data is now everywhere, in every organization, in every sector and every economy. As indicated by Manyika [4], European governmental administrators could save more than €100 billion in operational efficiency improvements alone by using big data. Using big data can create a substantial value in five broad ways. First, big data can unlock significant value by making information transparent and at a much higher frequency [4]. Additionally, organizations can collect more accurate and detailed performance information on everything from their stored transactional data in digital forms [4]. These digital transactions can be used to conduct controlled experiments to make better management decisions and improve their business level. Moreover, big data can help in producing much more precisely tailored products or services as big data permits ever-narrower segmentation of customers [4]. Furthermore, decision-making can be substantially improved and enhanced by sophisticated analytics. Last of all, big data can be utilized to improve the development of the next generation of products and services [4]. For example, manufacturers who use data obtained from sensors embedded in products can create innovative after-sales service offerings such as proactive maintenance or preventive measures that take place before a failure is noticed or even occurs [4]. The large amount of data is increasing. Therefore, analysing large data sets will become a key basis of competition, productivity growth, and innovation. According to Manyika [4], leaders in every sector will have to cope with the implications of big data. The increasing volume and detail of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future. The report produced in 2011 by McKinsey Global Institute [4] predicts that there will be a shortage of talent necessary for organizations to take advantage of big data. It was projected that by 2018, the United States alone would face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to utilize the analysis of big data to make effective decisions. GilPress [5] predicted that as of June 2014, there were more than 30 higher educational institutions in the US offering over 40 data science programs. The schools hosting these DS programs vary. The academic degrees offered would range from bachelor's, master's, and doctoral, to certificates, however, the majority are graduate degree programs as indicated by Tang & Sae-Lim [6]. The data science field is classified as an interdisciplinary field that involves the combining of two or more academic disciplines into one activity. Wladawsky-Berger [7] indicates that Data Science as an interdisciplinary field, employing methodologies and practices from across academia and industry. Having interdisciplinary nature, the DS programs vary greatly in their interdisciplinary characteristics, requirements, curricula, and technical expertise. For instance, the Master's degree in Business Analytics at Michigan State University comprises 31.5 credit hours and is completed within one calendar year. In contrast, the Master of Data Science program of the Illinois Institute of Technology is a two-year program with 34 credit hours of coursework and a practicum [9].

The demand for data scientists and engineers is growing enormously. To instance, the job title "Data Scientist" topped Glassdoor's 50 Best Jobs in America for the second consecutive year in

2017 [8]; [9]. In addition, the job titles "Data Engineer" and "Analytics Manager" both cracked the top five. A recent study by Louis Columbus and IBM projects that the demand for data scientists and data engineers will increase by 39 percent by 2020 [8], when the number of annual job openings for data professionals reaches 2.72 million. The study further states that annual demand for data-driven decision makers, which comprise one-third of the data savvy professional job market, is predicted to rise by 110,000 openings by 2020 [8]. Moreover, annual demand for the fast-growing new roles of data scientist, data developer, and data engineer will reach nearly 700,000 openings by 2020 [8].

Patil [9] mentions that it is a good thing to see the growing availability of data science programs in academia. This study indicated that as recently as 2011, there were no existing formal training programs. Now, there are solid data science or advanced analytics programs in place at Columbia's Institute for Data Sciences and Engineering; UC Berkeley's iSchool; Carnegie Mellon University; Illinois Institute of Technology; Imperial College, London; North Carolina State; Syracuse University; and the University of Tennessee. The study also indicated that big companies like IBM are partnering with universities to close the Big Data skills gap. Moreover, there are pioneers from early data science groups at Yahoo! and LinkedIn now scattered throughout the tech world, dedicating themselves to training and inspiring the next generation of data scientists. The Insight Data Science Fellowship Program started by Jake Klamka is a major example of that [9].

There are a few studies examined the data science curricula. Nolan & Temple Lang [10] provided a method for teaching computing skills in the field of statistics and data. The authors suggested that the statistic curricula need to be revised and reformed constantly. They stated that statisticians need to address what is missing from the curricula and take the lead in improving the level of students' data competence. In 2014, Anderson, Bowring, McCauley, Pothering & Starr [11] conducted a study to discuss the implementation of a four-year DS undergraduate program at the College of Charleston, Charleston, South Carolina, USA. Their study presents a ten-year status report detailing the program's origins, successes, and challenges, development effort and the results of deploying the DS programs. One year later, Hardin et al. [12], investigated the implementation of data science in their own statistics curricula. Their study examined the case of seven institutions at the undergraduate level. The study also included examples of assignments designed for courses that foster engagement of undergraduates with data and data science. In the same year, Asamoah, Doran, & Schiller [13], presented their experience in designing, developing and delivering an interdisciplinary Introduction to Data Science course for upper-level undergraduate and graduate level students in multiple colleges. Ben Baumer in his article "A Data Science Course for Undergraduates: Thinking With Data" [14], proposed and described an undergraduate course at a liberal arts college in data science that is atypical within the current statistics curriculum which provides students with the tools necessary to apply data science and prepare them to work with the modern data streams. His study [14] is not simply a collection of topics from existing courses in statistics and computer science, but rather an integrated presentation of something more holistic that might be used as a blueprint for a significant expansion of the existing statistics curriculum. In 2016, Tang & Sae-Lim [6], conducted an exploratory analysis of 30 randomly selected data science programs from eight disciplines. Their analysis is cantered on linguistic patterns of program descriptions, curriculum requirements, and DS course focus as pertaining to key skills and domain knowledge. In 2017, De Veaux et al. [15], developed a curriculum guideline for undergraduate programs in data science. This study included the required skills for data scientists and as result what data science should cover in its crucial. A very recent study by Yan & Davis [16] examined the undergraduate program offered by the University of Massachusetts Dartmouth that started in 2015. As it can be noticed, this study is similar to the studies conducted by Anderson, Bowring, McCauley, Pothering & Starr [11], De Veaux et al. [15], and Baumer [14] in terms of examining and documenting the implementation of DS programs at the undergraduate level. In addition, National Academies of Sciences, Engineering, and Medicine produced in 2018 a book with three chapters to outline some considerations and approaches for academic institutions and others in the broader data science communities to help guide the ongoing transformation of this field [1]. The only study found of data science programs at the graduate level was the study of Tang, & Sae-Lim [6]. What it can be drawn from reviewing the related literature is that most of the studies done were at the undergraduate level and very limited studies conducted at the graduate level. Also, most of the curriculum studies found were conducted in different discipline. Tang & Sae-Li's study is the only one found to be studying DS master's programs and analysing the DS curricula [6]. However, their study was analysing the program descriptions and course descriptions to identify the skills that covered by DS programs. Moreover, the study sample was chosen almost equally from different disciplines to identify the difference of DS programs in terms of hosting schools. More studies need to be conducted in order to fulfil the gap in the literature.

3. RESEARCH METHOD

This study applies content analysis that is a widely used qualitative research technique. The term content analysis refers to a research method that allows the qualitative data collected in research to be analysed systematically and reliably so that generalizations can be made from them in relation to the categories of interest to the researcher [17]. It is also used for analysing the content in a systematic and quantitative approach [18]. A review of the literature relating to data science and other curricula indicated that various research approaches have been used in studying curriculum such as interviews, focus group and content analysis. For instance, interviews and the focus group technique were utilized in curriculum research projects by Curran, et al. [19], and Curry [20] while a survey was the adopted technique in Al-Ansari, & Yousef [21] and Kaeshita & Otsuki [22] studies. Content analysis was the chosen method employed by Tang, & Sae-Lim for investigating data science curricula in higher education [6]. Chu [23] also explored the curricula of the library and information science LIS master's programs in the USA by conducted a content analysis technique. This study aims for analysing data science curricula; therefore, content analysis appears to be the most appropriate technique to use in this study. As this is an exploratory study, content analysis will be used to describe the current situation of 30 DS masters' programs in the USA and to provide a descriptive and quantitative analysis to DS programs. In addition, content analysis will help to address the second question of this study.

4. DATA COLLECTION

Data for the current study is collected from different websites to obtain DS master's programs and to gather the required information of individual course titles and descriptions. The following subsections specify the entire data collection process in three steps.

4.1. Data Science Programs

The first step in collecting data for this research involves the identification of data science master's programs in the USA. Two websites are used in this stage in order to obtain a list of programs. A list of such programs is extracted from http://www.mastersindatascience.org and https://www.discoverdatascience.org. Fifty master's programs in the United States are obtained at the beginning. After obtaining the list of programs, each of the 50 DS programs' websites was visited for locating its corresponding curriculum and locating the school hosting the program. Twenty DS programs are offered as a track or concertation and thus eliminated from the list. Only thirty DS programs meet the criteria of inclusion in this study. A list of those programs and their curricular URLs can be found in the Appendix.

4.2. DS Curricula

Each DS curricular URL listed in the Appendix was followed to retrieve its actual curriculum. Only required courses or core courses for DS master programs are obtained, elective courses are not in the scope of this study. Once a DS curriculum was located, the university’s name, programs’ title, school hosting the program and link to program curricula were entered in Microsoft Excel.

4.3. Course Titles & Descriptions

Course titles and related information (e.g., credits or descriptions) were entered into an Excel sheet. During this process, courses that appear unique (e.g., Introduction to Data Science & Data Visualization) were identified and marked accordingly after consultation with respective course descriptions. Unfortunately, syllabi were not available on the Web for most courses, so the course descriptions were analysed which greatly help determine contents of the courses.

5. DATA ANALYSIS

Exactly 290 individual core courses offered by 30 DS programs were examined based on their course titles, descriptions, and other curriculum-related information. First of all, DS programs were categorized based on the hosting schools. Second, required courses are extracted and examined to discern any pattern in the DS core curriculum. Then, core courses were analysed based on titles and descriptions to identify what area the course covers. After that, courses were grouped and coded according to the subject titles and descriptions. A coding schema was developed by the author with consulting to De Veaux et al. [15], Tang and Sae-Lim [6] studies and 12 top Data Science Skills reported by Data Flair Team [24]. Third, frequency of categorized courses based on the developed schema was computed to determine the prominent area or discipline that has control on DS master's programs. Finally, a further analysis was conducted in courses titled "Information/Data Retrieval" and "Text Mining" in order to examine the involvement of IR in DS. Both quantitative and qualitative methods are applied for analysing data gathered for this study as program requirements including the total number of credit hours, core courses, and capstone were recorded for statistical analysis. A number of schools did not provide detailed information about their credit hour requirements on their websites. Such cases were counted as missing values.

6. RESULT AND DISCUSSION

On average, DS master's programs require 18.3 credits and 9.7 courses to complete the core requirements. In terms of required cores number, New York University had the highest number of core courses where University of Maryland recorded the lowest number of required courses. Massachusetts Institute of Technology was recorded as having the longest DS master's program requiring 84 credits excluding capstone, project, practicum or internship, while the shortest program requires only 7 credits at Brown University. There seems to be inconsistency of the requirement across the 30 programs. This could be due to the length of the program and the academic calendar systems used in the school hosting the program if it is quarter, semester or trimester system. Table 1 provides several summary statistics for the required courses in the field.

Table 1. Summary Statistics of Required DS Courses and required Credit

	Number of Required Courses	**Required Credit**
Maximum	17 courses NYU	84 credits MIT
Minimum	4 courses University of Maryland	7 credits Brown University
Mean	9.7	18.3

6.1. Hosting School

As it is mentioned earlier, data science is characterized as an interdisciplinary field. Therefore, DS courses would reflect this interdisciplinary nature in the programs and would represent a combination of subject areas from several disciplines. This is also applying to the school hosting the programs. Schools that host DS master's programs in the United States are diverse: they include and are not limited to Business, Computer Science, and Sciences Schools. However, School of Management seems to take the lead and the initiative in lunching and hosting this program. School of Business comes second in line. Third, is School of Engineering by 5 and then School of Science, Computer Science and Data Science Institutes. School of Continuing Liberal and Professional Studies, School of Arts and Sciences and Graduate School, they all hosted a program each. The table below provides a list of universities organized by hosting school to DS programs.

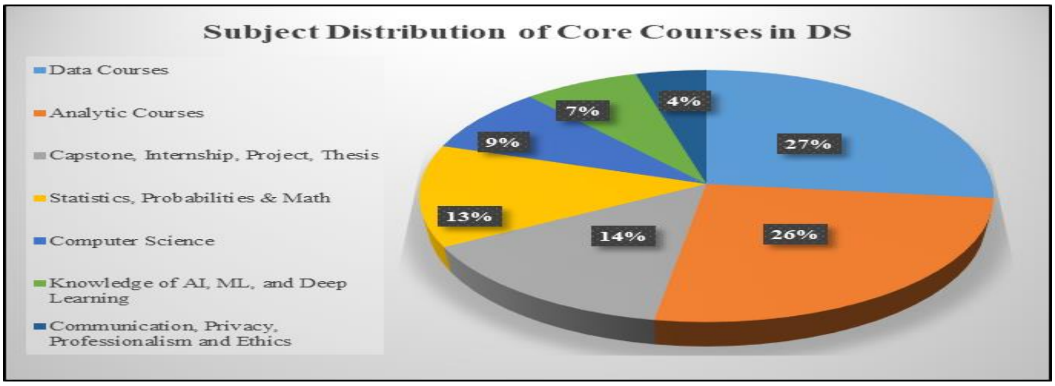
Table 2: list of university by disciplines or hosting school

School / Discipline	University
School of Management	1. Massachusetts Institute of Technology 2. University of Minnesota-Twin Cities 3. University of California-Davis 4. University of Iowa 5. Arizona State University 6. University of Connecticut 7. Southern Methodist University
School of Business	1. Southern Methodist University 2. University of Texas at Austin 3. New York University 4. Texas A&M University: business and Stats 5. University of Maryland 6. Duke University
School of Engineering	1. Northwestern University 2. California Baptist University 3. City College of New York 4. Grand Valley State University 5. Harvard University
Data Science Institutes	1. Columbia University 2. The University of Virginia 3. Brown University
School of Science	1. Embry-Riddle Aeronautical University 2. Georgetown University and art 3. Illinois Institute of Technology
Computer Science	1. College of Charleston and math 2. University of Central Florida 3. City University of New York
School of Continuing Liberal and Professional Studies	1. University of Chicago
School of Arts and Sciences	1. American University
Graduate School	1. Montclair State University

6.2. Course Distribution

Table 3 clearly demonstrates that data science related courses such as data retrieval, database management, data mining and data visualization are in the top core offerings, followed by analytical courses. The introduction courses represent the majority of DS curricula by almost one third of the DS curricula. Here is where the involvement of IR can be seen. Two courses have the word retrieval in their titles. These two courses are offered by DS master's program at Embry-Riddle Aeronautical University and Northwestern University. When an in-depth analysis is conducted for the two course descriptions, these courses teach students the basic of databases and systems such as administration, applications, data scripting, query processing and how to model, organize, store and analyze data in modern relational database. In addition, these two courses teach algorithm and how effectively student can use modern system optimizations (indexing, partitioning, memory hierarchy. These two courses clearly identify the connection between IR and DS. Basically, IR cannot exist without data, and data cannot make senesce without having a well-developed Information Retrieval System IRS. There is no surprise that analytical courses come as the second majority of DS curricula as the goal of DS programs is to train students to extract knowledge and insights from data. Also, analytical courses are connected to IR in term of teaching students the programing languages, techniques and applications that facilitate data extracting and data mining which is IR but in another form. Text mining and analytics course was also investigated in detail to determine how IR involve in DS. Text mining is defined by Krallinger & Valencia [25], as the automatic extraction of information. IR is concerned about finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). This course explores a breadth of Natural Language Processing (NLP) applications with a focus on deep learning techniques. Topics include word embeddings and common deep learning NLP architectures; approaches to a variety of NLP tasks such as text classification, named entity recognition, machine translation, information retrieval, term vectors, similarity, clustering which all used by IRS. It can be said that text mining or text analytics is a form of IR. Extracting knowledge and insights from data cannot be achieved by individual without intensive training. As a result, capstone, project, practicum and internship come as the third common area in DS curricula. Statistics, math and probability courses play an important role in DS curricula as well as IR as they used for retrieving techniques, weight, text mining and etc. There is also a need for students to have some knowledge about computer science, AI, ML, DL, NLP in order to train such an intelligent system which is capable to store and retrieve big data and also make a conclusion from it. Privacy and ethics are emerging as an issue in the world of big data. Therefore, privacy, ethics, professionalism and communicating with data courses need more attention in DS curricula as they represent only 4% of courses in DS programs. Below is a descriptive analysis of subject distribution of core courses in DS master's programs.

Figure 1. Course Distribution in Percentage



TTable 3. Subject Distribution of Required Courses in DS – Frequency

Required Course Clusters	**Frequency**
Data Science Courses • Introduction to Data Science • Database Design and Information Representation • Database management • Database Structure • Data/Information Retrieval • Information System • Data Visualization • Data Mining	**86**
Analytics Courses • Big Data • Decision Making Courses • Predictive Analysis • Business Analytics • Text Analytics	**83**
Capstone, Internship, Project, Thesis	**46**
Statistics, Probabilities & Math	**42**
Computer Science • Algorithm • Programing	**27**
Knowledge of AI, ML, and Deep Learning • Business Intelligence • Deep Learning • Natural Language Processing • Machine Learning	**23**
Communication, Privacy, Professionalism and Ethics	**14**

6.3. Information Retrieval & Data Science

According to Srivastava [26], information retrieval system is a network of algorithms, which facilitate the search of relevant data/documents as per the user requirement. It is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Information retrieval is involving the process of information extraction. Information extraction is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. Like IR, data science also involves the process of information extracting in order to obtain important pattern, features, knowledge from data. Somehow, data science needs IR in order to obtain data and make conclusion from it. Here is the involvement between the two fields that can be documented. Data Science and data analytics cannot make sense without IR as the base idea of DS is IR. Furthermore, IR is based on data, IR cannot work if data is not existed, represented and stored. Therefore, no wonder the word retrieval is repeated more than twice in the course title and more than seven times in the course descriptions of the selected sample. The synonymous word of retrieval “extract" also appears almost 13 times in the description of data management courses, data visualization courses, data mining, big data and computer science courses. This appearance of the word extract confirms the idea of IR being the core of data science field.

There is also relationship, connection or similarity can be identified between the two fields. The idea of IR is based on the way information is represented. According to Chu [27], Information representation and retrieval is known as abstracting and indexing, information searching, and information processing and management. The main objective of IR is to extract information resources that relevant to an information need. This mission of retrieving relevant information is achieved through hard coded rules or through feature-based models like in machine learning. Either way, the end goal is to get out the relevant resources. In machine learning, the end goal is to learn good models of reality in order to regress, classify, or describe the data. Here is a connection identified between IR and DS as they both make use of ML models in order to retrieve relevant information in IR and meaningful information in DS. As a result, both fields need to study ML as it is the only way to succeed and achieve their end goals. Another connection can be recognized is in building information retrieval systems. Information retrieval system according to Le & Gulwani [28], is a network of algorithms, which facilitate the search of relevant data/documents as per the user requirement. The IRS is basically extracting information and knowledge from data. In order to build such a system, developers need to be sophisticated in the area of database structure and algorithm. This knowledge is also needed by data scientists. Furthermore, Data analytics needs important information for processing, visualization. Raw data is not useful directly, once you extract important information out of it, that can give you better insight. So, the result indicates that almost 40 percent of required courses in DS is about IR. Data science cannot exist without information extraction and information retrieval system cannot exist without having indexed and organized data.

7. CONTRIBUTION & STUDY LIMITATION

Results of this study can be beneficial to other universities in the US and different countries that aim to lunch such a program. Moreover, results can be used by the selected sample in order to compare their program to similar one on the same area and for evaluation and improvement purpose. This study contributes by selecting a purposive sample of 30 DS master's programs in the U.S. in order to identify the most popular hosting school to DS programs. In addition, this study examines the course titles and course description to find out the disciplines that contribute in forming future data scientists. Moreover, this study investigates and discusses the connection or involvement of information retrieval in data science. As an exploratory study, the present examination has a number of limitations. First, the data collected was derived based only on what was published on program websites. Additionally, data collection was completed in about two weeks in late October 2019, So, factual information might not be published on the website at that time, or the website has not been updated with most recent data. A number of DS programs did not publish course description on their website. The sample was selected from the mentioned websites. So, there would be other disciplines and schools involving in offering the DS master’s programs that not yet discovered. A larger sample in similar study can be conducted in the future. Thirdly, the coding scheme that was adopted from De Veaux et al.; Tang and Sae-Lim and Data Flair Team, targets mainly business analytic skills and data scientist skills. There is a very limited number of classification schemes available in existing literature that address general data science skills or data science categories.

8. CONCLUSION

Data science centres on the notion of multidisciplinary and interdisciplinary approaches to extracting knowledge or insights from large quantities of complex data for use in a broad range of applications [1]. It incorporates knowledge from Statistics, Computer Science and Mathematics and hence can deal with challenging application domains which had remained out of reach because of a combined lack of data and computer power [29]. Data scientists are very much in demand as companies grapple with the challenge of making valuable discoveries from Big Data.

For that reason, academic institutions have started to lunch and establish new DS programs to prepare students to be data scientists. A number of studies have investigated data science programs within a particular discipline, such as Business (e.g. Chen et al.) [30]. However, there are very few empirical studies that explore DS programs and examine its curriculum structure across disciplines. The present study attempted to make a contribution to establish an understanding of the current state of DS education in the U.S. Due to the fact that DS programs are still in their early stage of development, there seems to be inconsistency in DS programs between program length and the core curricular. This idea was also supported by Tang & Sae-Lim [6]. There are also notable influence of the school hosting the programs in structuring the program curricula. For example, the DS programs in Business Schools attempted to devote more core courses towards covering more business skills, but still fall short of covering mathematics and statistics as well as communication and visualization and computer science skills. Large proportions of DS curricula were dedicated to data science, data retrieval, data visualization and analytics related courses. Mathematics, statistics, computer science and artificial intelligence related courses are also contributing to the DS programs. However, DS programs fall short of covering privacy and ethics course in their curricula as they are becoming a hot topic recently. This study presents a step-in describing data science educational practices in the United States. Needless to say, a lot needs to be done in order to meet the challenges the new age of big data presents in educating data professionals. Future research could have larger sample including more countries. Studies also should involve leaders, faculty, students, and graduates of DS programs in order to understand the development of these programs, their effectiveness in achieving stated goals, the operational structure of the DS curricula, the learning outcomes of core and elective courses, and, finally, the perceptions and experiences of stakeholders with DS programs.

ACKNOWLEDGEMENTS

I would like to express my special thanks of gratitude to Taibah University and Saudi Government for the sponsorship and support provided to my study. I would also like to extend my gratitude to Dr. Chu for her valuable advice and comments and to my family for offering all the support I needed.

REFERENCES

[1] National Academies of Sciences, Engineering, and Medicine. Data science for undergraduates: Opportunities and options. National Academies Press, 2018.

[2] Marr, Bernard. "How much data do we create every day? The mind-blowing stats everyone should read." In Forbes. 2018.

[3] Hicks, Stephanie C., and Rafael A. Irizarry. "A guide to teaching data science." The American Statistician 72, no. 4 (2018): 382-391.

[4] Manyika, James. 2011. "Big data: The next frontier for innovation, competition, and productivity". McKinsey Digital. Accessed April 22,2020 http://www. mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.

[5] GilPress. 2012, " Graduate Programs in Big Data Analytics and Data Science." Accessed February 20,2020. https://whatsthebigdata.com/2012/08/09/graduate-programs-in-big-data-and-data-science/

[6] Tang, Rong, and Watinee Sae-Lim. "Data science programs in US higher education: An exploratory content analysis of program description, curriculum structure, and course focus." Education for Information 32, no. 3 (2016): 269-290.

[7] Wladawsky-Berger, Irving. 2014. "Why Do We Need Data Science When We've Had Statistics for Centuries?". Wall Street Journal (blog). May 2, 2014. Accessed April 29, 2020. https://blogs.wsj.com/cio/2014/05/02/why-do-we-need-data-science-when-weve-had-statistics-for-centuries/

[8] Columbus, Louis. "IBM predicts demand for data scientists will soar 28% by 2020." IBM White Paper (2017).

[9] Patil, DJ. "Still the Sexiest Profession Alive." Harvard Business Review. Accessed February 22, 2020. https://hbr.org/2013/11/still-the-sexiest-profession-alive

[10] Nolan, Deborah, and Duncan Temple Lang. "Computing in the statistics curricula." The American Statistician 64, no. 2 (2010): 97-107.

[11] Anderson, Paul, James Bowring, Renée McCauley, George Pothering, and Christopher Starr. "An undergraduate degree in data science: curriculum and a decade of implementation experience." In Proceedings of the 45th ACM technical symposium on Computer science education, pp. 145-150. 2014.

[12] Hardin, Johanna, Roger Hoerl, Nicholas J. Horton, Deborah Nolan, Ben Baumer, Olaf Hall-Holt, Paul Murrell et al. "Data science in statistics curricula: Preparing students to "think with data"." The American Statistician 69, no. 4 (2015): 343-353.

[13] Asamoah, Daniel, Derek Doran, and Shu Schiller. "Teaching the foundations of data science: An interdisciplinary approach." arXiv preprint arXiv:1512.04456 (2015).

[14] Baumer, Ben. "A data science course for undergraduates: Thinking with data." The American Statistician 69, no. 4 (2015): 334-342.

[15] De Veaux, Richard D., Mahesh Agarwal, Maia Averett, Benjamin S. Baumer, Andrew Bray, Thomas C. Bressoud, Lance Bryant et al. "Curriculum guidelines for undergraduate programs in data science." Annual Review of Statistics and Its Application 4 (2017): 15-30.

[16] Yan, Donghui, and Gary E. Davis. "A First Course in Data Science." Journal of Statistics Education 27, no. 2 (2019): 99-109.

[17] Haggarty, Linda. "What is content analysis?." Medical Teacher 18, no. 2 (1996): 99-101.

[18] Allen, Mike, ed. The SAGE encyclopedia of communication research methods. SAGE Publications, 2017.

[19] Curran, Charles, Stephen Bajjaly, Patricia Feehan, and Ann L. O'Neill. "Using focus groups to gather information for LIS curriculum review." Journal of Education for Library and Information Science (1998): 175-182.

[20] Curry, Ann. "Canadian LIS education: Trends and issues 1." Education for information 18, no. 4 (2000): 325-337.

[21] Al-Ansari, Husain, and Nibal Yousef. "Coverage of competencies in the curriculum of information studies: An international perspective 1." Education for information 20, no. 3-4 (2002): 199-215.

[22] Kakeshita, Tetsuro, and Mika OHTSUKI. "Survey and Analysis of Computing Education at Japanese Universities: Non-IT Departments and Courses." (2019).

[23] Chu, Heting. "Curricula of LIS programs in the USA: A content analysis." (2006).

[24] DataFlair Team. "12 Top Data Science Skills - Want to be a Data Scientist in 2019?." Data Flair Training (blog). January 9, 2019. Accessed December 14, 2019. https://data-flair.training/blogs/data-science-skills/

[25] Krallinger, Martin, and Alfonso Valencia. "Text-mining and information-retrieval services for molecular biology." Genome biology 6, no. 7 (2005): 224.

[26] Srivastava, Tavish. "Information Retrieval System explained in simple terms!. " Analytics Vidhya (blog). April 7, 2015. Accessed February 22, 2020. https://www.analyticsvidhya.com/blog/2015/04/information-retrieval-system-explained/

[27] Chu, Heting. Information representation and retrieval in the digital age. Information Today, Inc., 2003.

[28] Le, Vu, and Sumit Gulwani. "FlashExtract: a framework for data extraction by examples." In Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 542-553. 2014.

[29] Ley, Christophe, and Stéphane PA Bordas. "What makes data science different? A discussion involving statistics2. 0 and computational sciences." International Journal of Data Science and Analytics 6, no. 3 (2018): 167-175.

[30] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business intelligence and analytics: From big data to big impact." MIS quarterly (2012): 1165-118

Appendix. DS Master's Program & Their URLs for Curricula

Program Name	URL for Curriculum
Massachusetts Institute of Technology	https://mitsloan.mit.edu/master-of-business-analytics#curriculum
University of Texas at Austin	https://www.mccombs.utexas.edu/Master-of-Science-in-BusinessAnalytics/Academics/Curriculum
University of Chicago	https://grahamschool.uchicago.edu/academic-programs/mastersdegrees/analytics/curriculum
Columbia University	https://datascience.columbia.edu/course-inventory
Duke University	https://www.fuqua.duke.edu/programs/mqm-businessanalytics/curriculum
University of Minnesota-Twin Cities	https://carlsonschool.umn.edu/degrees/master-science-in-businessanalytics/academics/one-year-msba/program-structure
New York University	https://www.stern.nyu.edu/programs-admissions/ms-businessanalytics/academics/course-index
Texas A&M University	https://mays.tamu.edu/ms-analytics/curriculum-overview/
The University of Virginia	https://datascience.virginia.edu/degrees/info/programs-and-courses
Northwestern University	https://www.mccormick.northwestern.edu/analytics/curriculum/
University of California-Davis	https://gsm.ucdavis.edu/msba-academics
University of Iowa	https://tippie.uiowa.edu/future-graduate-students/mastersprograms/business-analytics/full-time-msba/curriculum
Arizona State University	https://wpcarey.asu.edu/masters-programs/business-analytics/curriculum
University of Connecticut	https://msbapm.business.uconn.edu/academics/curriculum/
Southern Methodist University	https://www.smu.edu/cox/Degrees-and-Programs/MS-in-BusinessAnalytics/curriculum American University https://www.american.edu/programs/shared/data-science/admissions.cfm
Brown University	https://www.brown.edu/initiatives/data-science/mastersdegree/curriculum
California Baptist University	https://calbaptist.edu/programs/master-of-science-data-science-andknowledge-engineering/courses
City College of New York	https://www.ccny.cuny.edu/engineering/curriculum
City University of New York	https://www.gc.cuny.edu/Page-Elements/Academics-Research-CentersInitiatives/Masters-Programs/Data-Science/Curriculum-and-Courses
College of Charleston	https://catalog.cofc.edu/preview_program.php?catoid=13&poid=2778&hl =%22data+science%22&returnto=search
Embry-Riddle Aeronautical University	https://erau.edu/degrees/master/master-of-science-degree-in-data-science
George Washington University	https://datasci.columbian.gwu.edu/ms-degree
Georgetown University	https://analytics.georgetown.edu/academics/degree-requirements/#
Grand Valley State	https://www.gvsu.edu/catalog/2019-

University	2020/program/master-of-science-indata-science-and-analytics.htm
Harvard University	https://www.seas.harvard.edu/applied-computation/graduateprograms/masters-data-science/degree-requirements
Illinois Institute of Technology	https://science.iit.edu/programs/professional-masters/master-datascience/coursework
University of Maryland	https://www.rhsmith.umd.edu/programs/ms-businessanalytics/academics#required
Montclair State University	https://www.montclair.edu/graduate/programs-of-study/data-science-ms/
University of Central Florida	http://ucf.catalog.acalog.com/preview_program.php?catoid=15&poid=700 6

AUTHOR

Duaa Bukhari is currently a Doctoral Student and Graduate Assistant at LIU Post, New York. She has a scholarship from Taibah University as an IT Lecturer to continue her study and obtain PhD degree from the United States. Prior to this Duaa was KAUST's metadata specialist. A native of Saudi Arabia, Duaa holds a Bachelor's degree in Library and Information Science from Umm Al-Quraa University, a Diploma in Computer Science from Alalamia Institute for Computer and Technology, and a Master degree in Information Technology from the University of Technology, Sydney, Australia. She has taught and done a lot of work in the Kingdom on cataloguing and classification, including the management of e-resources. Her interest is on the area of Data Science and Big Data Analytics.