A Text Analysis of Data-Science Career Opportunities and US iSchool Curriculum

Angel Krystina Washington Durr
ad1205@ymail.com

Data-science employment opportunities of varied complexity and environment are in growing demand across the globe. Data science as a discipline potentially offers a wealth of jobs to prospective employees, while traditional information science–based roles continue to decrease as budgets get cut across the United States. Since, historically, data are related closely to information, this research will explore the education of US iSchool professionals and compare it to traditional data-science roles being advertised within the job market. Through a combination of latent semantic analysis of over 1,600 job postings and iSchool course documentation, the aim of the article is to explore the intersection of library and information science and data science. It is hoped that these research findings will guide future directions for library and information science professionals into data science–driven roles, while also examining and highlighting the data-science techniques currently driven by the education of iSchool professionals. In addition, the aim is to understand how data science could benefit from a mutually symbiotic relationship with the field of information science, since, statistically, data scientists spend too much time working on data preparation and not nearly enough time conducting scientific inquiry. The results of this examination will potentially guide future directions of iSchool students and professionals toward more cooperative data science roles and guide future research into the intersection between iSchools and data science and possibilities for partnership.

Keywords: data science, information science, iSchools, text analysis

"Data" is defined by Cambridge International Examinations (2017, p. 3) as "a collection of text, numbers or symbols in raw or unorganized form." This definition will serve as the theoretical basis for discussion of "data" in this article. "Data" appears to be a term that many industry professionals are quickly becoming familiar with in one capacity or another, at least from a terminological standpoint. However, just because a term is discussed does not mean it is theoretically or historically understood by the individuals who use it most frequently. "Data" can be defined in the simplest terms as information lacking any contextual understanding. "Information" is defined by Cambridge International Examinations (2017, p. 3) as "data that has been processed, e.g. grouped . . . to give it meaning and make it interpretable." Data are collected on a variety of subject matter and across industries and languages: "As data continues to grow within our lives, businesses, organizations, and governments, there is a greater need to make sense of all the data. With this increase in data there is suddenly a great need for people to learn and up-skill themselves with key data science and analytics techniques and tools to make better informed

decisions" (Anslow, Brosz, Maurer, & Boyes, 2016, p. 620). Data-analysis techniques and tools must evolve with the rapidly changing data landscape: "The growth in the quantity and diversity of data has led to data sets larger than is manageable by the conventional, handson management tools. To manage these new and potentially invaluable data sets, new methods of data science and new applications in the form of predictive analytics, have been developed" (Waller & Fawcett, 2013, p. 77).

Waller and Fawcett (2013, p. 78) go on to say, "Generally, data science is the application of quantitative and qualitative methods to solve relevant problems and predict outcomes. One of the salient revelations of today, with the vast and growing amount of data, is that domain knowledge and analysis cannot be separated." As a result of the influx of data across industries and professions, professional data science, while often perceived to be rooted in business analytics, is growing in importance across industries and disciplines, which in turn has substantially increased the number of job opportunities available within this field: "Data scientists need deep domain knowledge and a broad set of analytical skills. Developing a broad set of analytical skills requires consistent investments of time. Developing deep domain knowledge requires similar dedication of effort. To that end, typically there is no single individual that can possibly have all of what is needed by a data scientist" (Waller & Fawcett, p. 78). For the second year in a row, data scientist was declared the best job in America according to Glassdoor. This ranking is based on "the number of job openings, the job satisfaction rating, and the median annual base salary" (Zhang & Neimeth, 2017, p.1). Despite this rating, there remains a shortage of qualified applicants compared to the number of currently available positions. According to Zhang and Neimeth (2017, p. 1), "the role of the data scientist is evolving, and organizations desperately need professionals who can take on data organizing as well as preparing data for analysis. Data wrangling or cleaning data and connecting tools to get the data into a usable format, is still highly in demand."

KEY POINTS:

- iSchools are often considered to drive innovation among the Information Science community; therefore, they often set trends and lead change within this group.
- With a booming career industry currently driving the market, iSchool students and professionals could potentially consider some data-science roles based on the alignment between the two found in this examination.
- With demand for data-driven professionals growing across industries, there is a need for continued exploration of the intersection between iSchools and data science and possibilities for partnership.

While data have existed for thousands of years in various formats across the globe, many organizations are quickly learning that with a rapid influx in electronic data comes an ever-increasing demand to properly prepare it for usage for rapid decision making: "Data preparation may require many steps, from translating specific system codes into usable data to handling incomplete or erroneous data, but the costs of bad data are high. Some research shows that analyzing bad data can cost a typical organization more than $13 million every year" (Zhang & Neimeth, 2017, p.2). It's not simply the influx of data; it's also the fact that the data are heterogeneous, with some structured and some unstructured. It is extremely important that this influx of data be properly maintained and prepared for analysis, not only for potentially becoming information, but also from a cost perspective.

Information science attempts to scientifically deconstruct the theories around and provide meaning to the overall understanding of information in its varied formats and enviroments. In the same way that data science takes a scientific approach to data, information science seeks to define and standardize information collection and organization for dissemination, consumption, and usage, whereas library science has been more historically focused around the scientific approach to library management and organization. Due to the increasing presence of information within libraries, these two disciplines are often discussed in conjuction with each other in academic literature, despite perceived differences in many academic communities. Over a hundred years of scholarly effort have been devoted to maximizing the organization of both libraries and information. Basing their analysis on the already well-established theoretical and historical understanding of information science, many scholars have discussed the perceived strategic financial and academic value of training library-science and information-science professionals for a variety of data-science and curation tasks, especially as data resources continue to grow in numbers at a rapid pace across industries: "Data management services are becoming increasingly prominent for library service developments in the future" (Sutherland & Wildgaard, 2016, p.17). Some may argue that data management, like many other library managed tasks, should be the work of computer science. However, arguments regarding the distinction amongst library science, information science, and computer science, and related disciplines are not new: "There has long been discussion about the distinctions of library science, information science, and informatics, and how these areas differ and overlap with computer science" (Marchionini, 2016, p. 1).

Many information schools teach information-science students in a very general sense. A specialized group of higher education information schools officially joined together around a common information education goal and formed a group now formally known as the iSchools. According to their official website, the topics of study covered in iSchools comprise

a variety of disciplines, including both information science and library science. Members of the iSchool group work together around the common goal of training future leaders in the information field, and member information institutions are often regarded as pioneer institutions for new information-science initiatives: "iSchools, a community of institutions that aim to lead education and research in information science, are embracing this challenge" (Ortiz-Repiso, Greenberg, & Calzado-Prado, 2018, p. 1). From a data perspective specifically, as Ortiz-Repiso et al. (2018, p. 1) continue, "the iSchool community's attention to data-focused education is not surprising, as core processes of information science (collecting, organizing, managing, accessing and supporting the use and manipulation of information) are acutely relevant to data-driven disciplinary areas such as data science." It has been reasoned that information schools, especially iSchools, should act strategically and logically if they want to declare their relevancy in the widely diverse data community: "It could be argued that data science is a subset of information science and some data science training programs may be housed in information schools, however, it is more strategic to view information science as an essential component of data science so that the emerging field can benefit from the diversity of perspectives that interdisciplinary collaborations bring" (Marchionini, 2016, p. 5).

> According to Thomas Schutz, SVP and GM of Experian Data Quality, Organizations need to be able to effectively leverage their information to make decisions and drive new initiatives.... Good data is good for business. We find that when organizations make improvements to their data, they see positive results. However, organizations need to speak a common language around data and prove its value so that investments can be made in data management practices, (Chen, Schütz, Kazman, & Matthes, 2017, p. 19).

Making improvements to data quality can be difficult, as this issue is multifaceted and often takes breaking down various silos that exist within an organization: "94% of companies across all levels have experienced internal challenges when trying to improve their data quality" (Carmody, 2016, para. 4). Additionally, many organizations have yet to recognize data-management and data-quality activities as an organization-wide priority and therefore lack integral data-science functions, which can lead to extremely costly mistakes if actionable insights are drawn from bad data: "The opportunities associated with data and analysis in different organizations have helped generate significant interest in business intellegience and analysis, which is often referred to as the techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions" (Chen, Chiang, & Storey, 2012, p. 1166).

Information-science and other iSchool professionals may be integral to managing and guiding the academic and professional research process

not only from an information standpoint but from a data perspective as well. Since the steps in the research process are fairly universal, it can be asserted that iSchool and information-science professionals could be utilized across industries and displines for the purpose of research and therefore should be knowledgeable in a variety of research techniques and methods (Chen et al., 2012). As Burton and Lyon (2017, p. 33) point out, "Universities are making significant investments in data infrastructure across their campuses and librarians will need to be prepared to contribute to these efforts…However, many librarians lack the technical skills to be effective in a data-rich research environment. We call this the skills gap." The traditional reference and instruction offerings may very likely need to evolve to meet the needs of the growing data-rich scientific research community. The impact of the expanding importance of data in research processes can be seen across both the academic information science and professional data science communities alike: "Current data science efforts will benefit from a more coordinated approach, where critical mass advantages can be realized, and network effects may be catalyzed" (Burton & Lyon, p. 35). Burton and Lyon also note that "as society is increasingly infused with data, librarians will have a crucial role in the future development of the data science ecosystem across multiple sectors" (p. 33). In addition, as stated by Morgan, Duffield, and Walkley (2017, p. 304), "RDM can be understood as a natural extension of the role of the Academic Librarians. Libraries have always managed research outputs, albeit in the form of books and journals, so librarians have many of the intellectual frameworks already in place." With data outputs becoming larger and larger across not only the professional sector but also the academic sector, iSchools appear to be leading a growing trend among education institutions to infuse data science into various aspects of the curriculum: "Among the organizations, iSchools are becoming a leading community that actively promotes data science education across many disciplines. . . . In order to cope with the latest trends, we argue that iSchools . . . empower students with information computing by educating them to create values, information, and knowledge" (Song & Zhu, 2017, p. 17). These interconnected areas of study "explain why data-related curricula across iSchool programmes are pursuing a number of different data-related emphases" (Ortiz-Repiso et al., 2018, p. 2), so it should come as no surprise that "iSchools are among the array of institutions supporting data science curricula and also interconnect to Big data and data analytics" (Ortiz-Repiso et al., 2018, p. 3).

Significance of the study

Research questions

The research questions are as follows:

1. What are the job characteristics and requirements mentioned in data-science job postings?

2. How does iSchool curriculum map to the job requirements specified in data-science job postings?

An initial step in the research investigation process was to examine the definitions of iSchools, library science, information science, and data science presented within academic literature across these fields and determine where, if it all, overlap existed. Following this literature review, the research will then to turn to job postings to examine the career opportunities in the data-science profession. Additionally, iSchool learning objectives were extracted from iSchool syllabi in master's-level iSchool organization membership institutions. Since course syllabus format can vary greatly from course to course across institutions, examining course learning objectives provides a somewhat universal framework for examining what topics were expected to be mastered within iSchool master's-level course offerings across the United States. The initial goal was to conduct exploratory research to understand the relationship between the academic information science and the professional data science through the analysis of professional data-science job postings and iSchool course syllabi to determine if a relationship already exists between the two distinct communities. Since this research is exploratory in nature, it is hoped that future research will further investigate the changing demands of professionals within either discipline, both academically and professionally, especially in terms of how information-science professionals from iSchools see their own role as a part of the larger discussion of professional data science. This initial research is purposefully broad so as to gather an introductory sample of the current data-science opportunities potentially available to iSchool students entering the job market upon successfully completing their required coursework.

Additionally, this research will attempt to measure the potential for information science and other existing iSchool professionals to fulfill data-science professional roles based on assumed academic training. Unlike other research in this area, the goal was to understand whether iSchool course learning objectives align to professional data-science needs and, if they do, to determine what specific professional data-science roles currently available best align with the knowledge that iSchool information science students are already exposed to in their required coursework. Instead of recommending that iSchool and other IS faculty teach future library-science and information-science professionals new skills that more closely relate to professional data-science employment offerings, it is hoped that this research will articulate how the academic information-science and professional data-science fields are already connected, thus providing additional channels of nontraditional employment prospects to iSchool and other information-science graduates and existing professionals.

The competency-based education framework (Burke, 1989) was used to determine the specific skills and competencies specifically requested in

data-science professional job postings. The competency-based education framework is intended to improve curriculum and instruction for students by translating student performance into specific learning outcomes. These values were measured against complete 2017 academic calendar-year syllabi from the seven iSchools that provided the necessary course syllabi to participate in the study. These schools vary in size and geographic location. All 37 US iSchools were contacted and asked to participate multiple times; however, only seven institutions responded. Most iSchool institutions appear to currently lack a centralized repository of course syllabi, which makes mass collection of these documents difficult.

Methodology

The methodology used in this study includes a variety of coordinated techniques of analysis because the subject matters examined are multifaceted and far too varied as concepts to tie to one specific method of analysis. Therefore, through a combination of latent semantic analysis of text via related job postings and US iSchool master's-level course syllabus learning objective information, academic information-science and professional data-science subjects were evaluated and analyzed for existing and potential partnership opportunities.

To understand the current professional data-science career opportunities potentially available to LIS graduates, this research presents textual analysis of data-science job postings. The postings included contain both traditional library-science and information-science roles, as well as postings for employment opportunities outside the LIS and traditional iSchool communities. Over 1,500 job postings were collected and used to perform textual topical analysis using SAS Enterprise Miner. In collecting employment postings from the data-science profession, it was important to understand how often the terms "data science" and "LIS" appear with each other on job postings as a potential gauge of the public perception of the overlap between the academic iSchool and professional data-science communities. Studying the positions being offered to potential graduates is integral to determining what skills and abilities are shared across the academic information-science and professional data-science communities. Additionally, collecting job postings from across the United States is essential for understanding the complete job market that graduates will encounter upon graduation.

Since education is integral to advancing the data-science profession and related subfields within the traditional LIS community, textual analysis was also performed on current course offerings at iSchools. This specialized group of member institutions was selected as a primary area of study because "the iSchools organization is a consortium of Information Schools dedicated to advancing the information field" (iSchools, n.d.). A total of 37 of the 83 worldwide iSchool member institutions are in the United States. All 37 U.S. iSchool institutions were contacted multiple times over

a three-month period. A total of seven, or 19%, responded and agreed to participate in this research by providing syllabi from courses offered in the 2017 school year. The initial goal was to get a minimum of 10 iSchools to respond and agree to participate in this research by providing complete syllabi materials. However, based on the feedback received, many institutions appear to currently lack a centralized repository of course syllabi, which makes collection en masse difficult. All syllabi are from the 2017 academic school year and are limited to graduate-level courses; job postings are from the summer period directly following, in an effort to gauge the potential knowledge and skills obtained by students entering the workforce after successfully completing their degrees.

Text from the course syllabi and job postings was used to perform textual analysis, specifically frequency analysis of applicable professional data-science–related terms and skills, via the software SAS Enterprise Miner. In addition to textual analysis of iSchool course syllabi, frequency-based textual analysis was conducted on job postings with the keyword "data science" or related terms contained either in the job title or under required skills. The aim was to better understand the correlation between data-science job postings and academic iSchool course syllabi to determine how well the two relate to each other in order to better understand the relationship between academic information science and professional data science. Data-science–related job postings were collected from the job search engine Indeed.com, which was selected because it serves as a search engine for other major job-search websites: "Indeed is the #1 job site worldwide, with over 200 million unique visitors per month. Indeed is available in more than 60 countries and 28 languages, covering 94% of global GDP" (Indeed, 2018). Therefore, Indeed.com job results are robust in number and can be searched using Boolean operators. Over 1,500 job postings were analyzed to determine what, if any, academic information-science skills were most often mentioned in conjunction with "data science" and related terms.The same keyword search terms were reused throughout the job-search process to ensure uniformity and accuracy in research scope. The keywords utilized for the job-collection process are documented in the research results section below. These job postings were collected from Indeed.com daily over a consecutive 60-day period to ensure that a varied sample was collected.

Frequency analysis between job postings and iSchool syllabi were measured against each other in order to better understand the relationship between the two that currently exists and ideally to guide future directions in one or both communities. However, this represents only part of the complete professional data-science picture. Terms that appear within the job postings and course information were measured by cross-tabulations to determine and compare frequency. The Vector Space Model (Salton, Wong, & Yang, 1975) was selected for text analysis, including filtering of unnecessary terms and words, term weighting, and reduction

(Coussement & Van Den Poel, 2008), to streamline topics to only those with value associated via term-frequency analysis. Additionally, Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) was performed via Singular Value Decomposition (SVD) (Landauer, 2007) through reduction of the resulting term-frequency matrix. The skills and abilities from job postings were compared to course syllabus learning objective skills and abilities to determine weight in either category. The skills and abilities most discussed in job postings were compared to course syllabi to determine the similarities and differences of weighted terms. Value of skills and abilities were measured using chi-square analysis (Winson-Geideman & Evangelopoulos, 2013).

Additionally, a "greedy" approach algorithm was applied to results based on chi-square scores to eliminate terms based on value until the top 10 skills and abilities within the job postings remained (Cormen, Leiserson, & Rivest, 1995). Other mentioned professional data-science skills and abilities were eliminated from the analysis at this point and the remaining 10 were isolated and compared to iSchool course-learning objectives. Elimination through attribute reduction was conducted a second time to determine what skills and abilities align to course learning objectives, if any, and which skills and abilities, if any, do not appear to be represented (Winson-Geideman & Evangelopoulos, 2013).

Course syllabi contain specific data related to course learning objectives. This area is often labeled "learning objectives." Competency-based learning is now standard practice at most higher education institutions, so data should be available within course syllabi to provide specific detail about the specific knowledge and skills that students should obtain from a course. Students are expected to reach certain levels of competency and then advance to higher levels of understanding as their knowledge of information grows. Therefore, ideally learning objectives would build upon learning objectives already presented in previous courses so that students eventually reach a point at which it is assumed they have obtained a higher and more complete understanding of topics or knowledge covered in coursework upon successfully graduating (Burke, 1989).

Results and analysis

As illustrated in Figure 1, while all identified iSchools in the United States were contacted multiple times via multiple channels to gain access to course syllabi over a span of more than six months, only seven of the total 37 US member iSchool institutions provided syllabi for this research. Therefore, there is just over an 18% participation rate among US iSchool institutions for this research study. Syllabi were categorized by institution and the learning objectives in each iSchool course syllabus were broken into sentence format, with each sentence obtaining its own unique identification. During the months of April, May, and June 2018, over 1,500 US job postings were collected using the job search engine Indeed.com, with

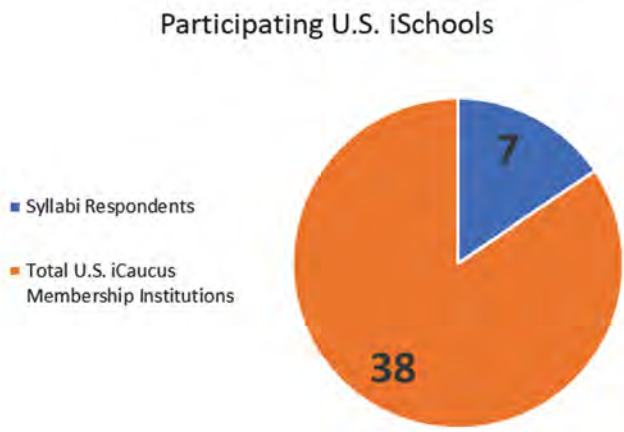
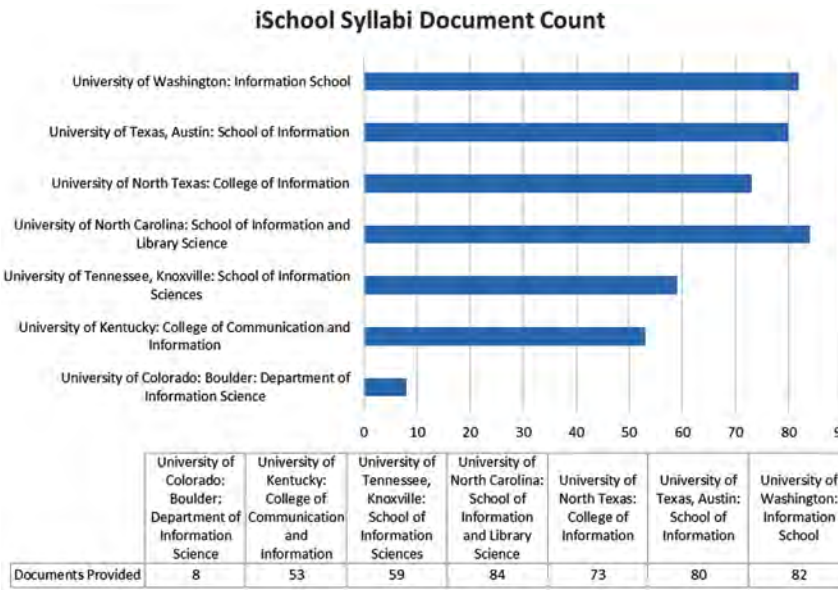


Figure 1: U.S. iSchool research participant count



	University of Colorado: Boulder: Department of Information Science	University of Kentucky: College of Communication and Information	University of Tennessee, Knoxville: School of Information Sciences	University of North Carolina: School of Information and Library Science	University of North Texas: College of Information	University of Texas, Austin: School of Information	University of Washington: Information School
Documents Provided	8	53	59	84	73	80	82

Figure 2: iSchool syllabus document counts

the keyword "data science." Like iSchool syllabi, data-science job postings were stripped down and categorized in sentence format, complete with unique identifications in order to maintain a separate classification from the syllabi for later comparison. The seven iSchools that were included in the analysis are outlined in Figure 2, along with the number of syllabi provided from each institution for the purposes of analysis.

VIEWTABLE: TMP1.datainfo51072

	Document_ID	File_Name	Sentence_ID	Text	CREATED
1	JPS1	(Senior) Consultant for Data Science job - MHP – A Porsche Company - Atlanta, GA _ Indeed.com	JPS1-45	A popular media company is seeking an experienced data science research engineer to join our growing analytics practice and help build their business intelligence capacity	04/19/2018
2	JPS1	(Senior) Consultant for Data Science job - MHP – A Porsche Company - Atlanta, GA _ Indeed.com	JPS1-32	Architect works with Application Architects, Data Architects, DBAs, and business partners to provide recommendations and best practices for the data warehouse, data models and data services required to support the client's strategic initiatives.	04/19/2018
3	JPS1	(Senior) Consultant for Data Science job - MHP – A Porsche Company - Atlanta, GA _ Indeed.com	JPS1-71	As Facebook continues to groW and connect people, having a platform thaT allows businesses to connect meaningfully to their customers matters more now than ever	04/19/2018
4	JPS1	(Senior) Consultant for Data Science job - MHP – A Porsche Company - Atlanta, GA _ Indeed.com	JPS1-50	At Nuna, our mission is to make high quality healthcare affordable for everyone.	04/19/2018

Figure 3: Screenshot of syllabus and job-posting data converted to sas format

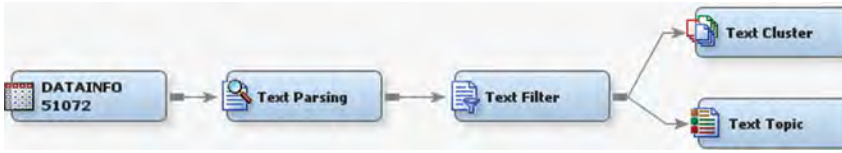


Figure 4: Analysis diagram screenshot 1

Figure 5: Analysis diagram screenshot 2

Graduate-level iSchool syllabi from the seven participating iSchools and job posting data were cleaned and prepared for analysis using Microsoft Excel during July and August 2018, and then it was loaded into SAS Enterprise Miner for topical analysis and comparison. A complete textual topical analysis was conducted using SAS Enterprise Miner in August 2018. Text was converted from Excel format to SAS format as depicted in Figure 3. The analysis diagrams used in SAS Enterprise Miner are depicted in Figures 4 and 5.

A total of 46,966 sentences were collected from 1,603 job postings, and a total of 4,072 sentences were collected from 439 syllabi, for a total of 51,038 sentences analyzed from 2,042 documents, as depicted in Figure 6.

After a topical analysis was performed on the sentences, Eigenvalues were plotted using Minitab and Excel, as depicted in Figure 6. It was determined based on this analysis that the first 18 eigenvalues were significant

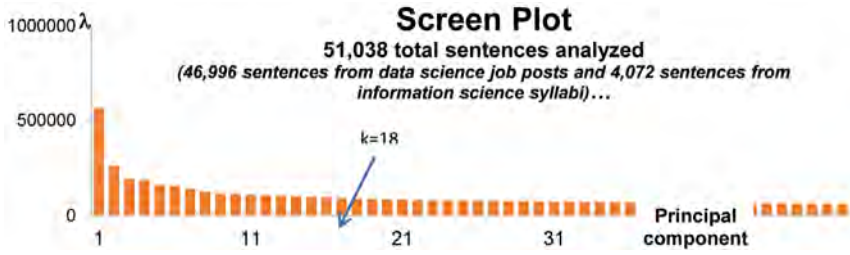


Figure 6: Eigenvalue screen plot

Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.082	0.019	+analysis,statistical,data,+data analysis,+statistical analysis	126	3906
Multiple	2	0.107	0.018	+degree,+statistic,+computer,+field,+bachelor	60	2617
Multiple	3	0.092	0.018	+skill,+communication,written,+strong,excellent	51	3395
Multiple	4	0.087	0.018	python,+program,+language,sql,java	88	3633
Multiple	5	0.107	0.018	+science,+data science,data,+computer,+team	55	3934
Multiple	6	0.092	0.018	+learning,+machine,+machine learn,+technique,+deep	59	2658
Multiple	7	0.086	0.019	+work,+ability,+environment,+team,+fast	125	5541
Multiple	8	0.097	0.018	+scientist,+data scientist,data,+senior,senior data scientist	62	2945
Multiple	9	0.105	0.018	+experience,+year,+work,minimum,+prefer	69	6588
Multiple	10	0.101	0.019	data,+tool,+visualization,+big,+data visualization	157	6647
Multiple	11	0.079	0.019	information,+technology,+system,management,+understand	122	3908
Multiple	12	0.083	0.018	+problem,+solve,+solution,analytical,+ability	105	3065
Multiple	13	0.090	0.018	analytics,advanced,+team,related,+related field	78	3599
Multiple	14	0.089	0.018	+business,+analyst,+intelligence,+understand,+solution	104	4921
Multiple	15	0.083	0.019	+project,management,+ability,+manage,+lead	171	4797
Multiple	16	0.089	0.019	+team,+engineer,+product,+development,+work	117	6453
Multiple	17	0.084	0.019	+model,predictive,+develop,+build,statistical	108	3903
Multiple	18	0.079	0.020	+customer,+company,+help,+market,+service	216	6327

Figure 7: Topic analysis from SAS Miner

and were therefore extracted for further topical analysis. The original topics and values associated with each from SAS Enterprise Miner are provided in Figure 7.

The 18 topics extracted for analysis were synthesized and summarized, as outlined in Table 1. Topics are identified by the number assigned to them, ranging from 1 to 18. For the purposes of this examination, these topic numbers were used for the remainder of the analysis. These topics vary in complexity and type and range from general employment skills to more specific data-science–related topics identified in the literature review.

After the topic mapping outlined in Table 1, topics were compared to one another by occurrence in collected iSchool graduate-level syllabi and data-science job postings as displayed in Figure 8. Topic frequencies are ranked according to usage by job postings in Figure 9 and collected iSchool syllabi in Figure 10 to show ranking as it pertains to each group of documents separately. This allows readers to examine topic frequency separately, since there is a large divide between the number of sentences included in each grouped text analysis. However, these data are integral to providing transparency in how the numbers were prepared and transformed from sentence-count measurement to topic-frequency measurement for a more accurate overall analysis.

Table 1: Topic analysis key

Topic ID	Summarized analysis topic	SAS topic
T1	Statistical Analysis	+analysis, statistical, data, +data analysis, +statistical analysis
T2	Degrees	+degree, +statistic, +computer, +field, +bachelor
T3	Communications Skills	+skill, +communication, written, +strong,excellent
T4	Programming Languages	python, +program, +language, sql, java
T5	Data Science	+science, +data science, data, +computer, +team
T6	Machine Learning	+learning, +machine, +machine learn, +technique, +deep
T7	Work Environment	+work, +ability, +environment, +team, +fast
T8	Data Scientist	+scientist, +data scientist, data, +senior, senior data scientist
T9	Experience	+experience, +year, +work,minimum, +prefer
T10	Data Visualization	data, +tool, +visualization, +big, +data visualization
T11	Information & ICT	information, +technology, +system, management, +understand
T12	Problem Solving	+problem, +solve, +solution, analytical, +ability
T13	Advanced Analytics	analytics, advanced, +team, related, +related field
T14	Business Understanding	+business, +analyst, +intelligence, +understand, +solution
T15	Project Management	+project, management, +ability, +manage, +lead
T16	Team Work	+team, +engineer, +product, +development, +work
T17	Predictive Modeling	+model, predictive, +develop, +build, statistical
T18	Customer Service	+customer, +company, +help, +market, +service

Following a separate and consolidated general comparison of topic frequency, additional analysis was conducted with an emphasis on topic as a percentage of relevant documents for both collected syllabi Figure 11 and job postings Figure 12. Then frequency was compared according to percentage of emphasis in either group of collected documents. Based on this comparison, one can clearly see that a relationship exists between

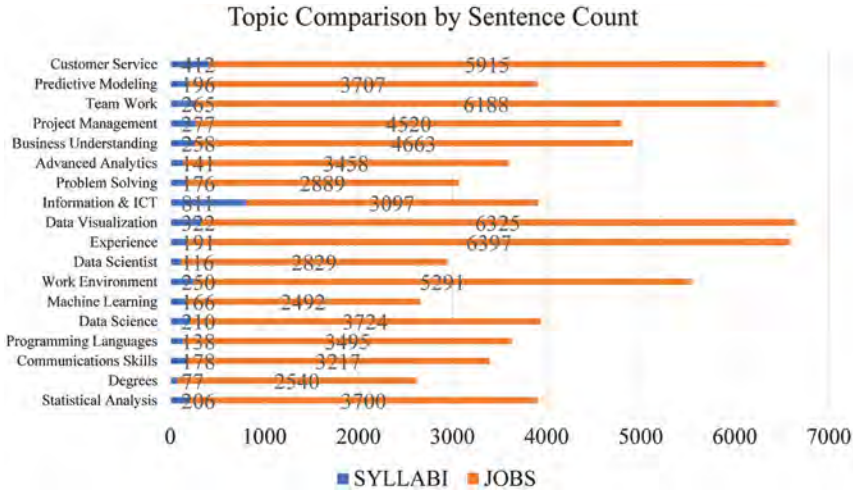


Figure 8: Topic comparison by sentence count

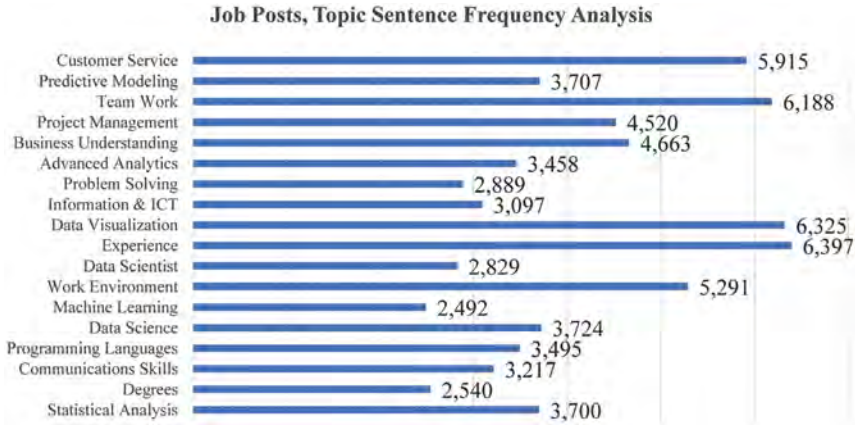


Figure 9: Job postings: Topic frequency analysis

collected LIS iSchool syllabi and data-science job postings in the context of topic comparison. Topics are displayed independently and combined in the sentence-count data collection to provide an equal frame of reference for analysis. As a final step, a chi square analysis was conducted upon examining the bar charts shown in the figures. This extra step was necessary because a much smaller number of syllabi were collected in comparison to the extremely large amount of job posting text analyzed. This additional step in the research analysis procedure allows ranking based on percentage for a more equal and unbiased comparison. After a chi square test was conducted

Figure 10: Syllabi: Topic frequency analysis

using Minitab and the frequency counts of both the collected job postings and syllabi were obtained, a chi square of 1955.325 with 17 degrees of freedom was discovered, as illustrated in Figure 13. After this is taken to the *t* table, the null hypothesis can be rejected based on the value being lower than the chi square obtained. The chi square null hypothesis measured was that iSchool syllabus learning objectives and related text collected have a topic frequency equal to that in the data-science job postings collected. In this context, the alternative hypothesis for the chi square analysis would be that the two groups of sentences have unequal distribution topic frequency.

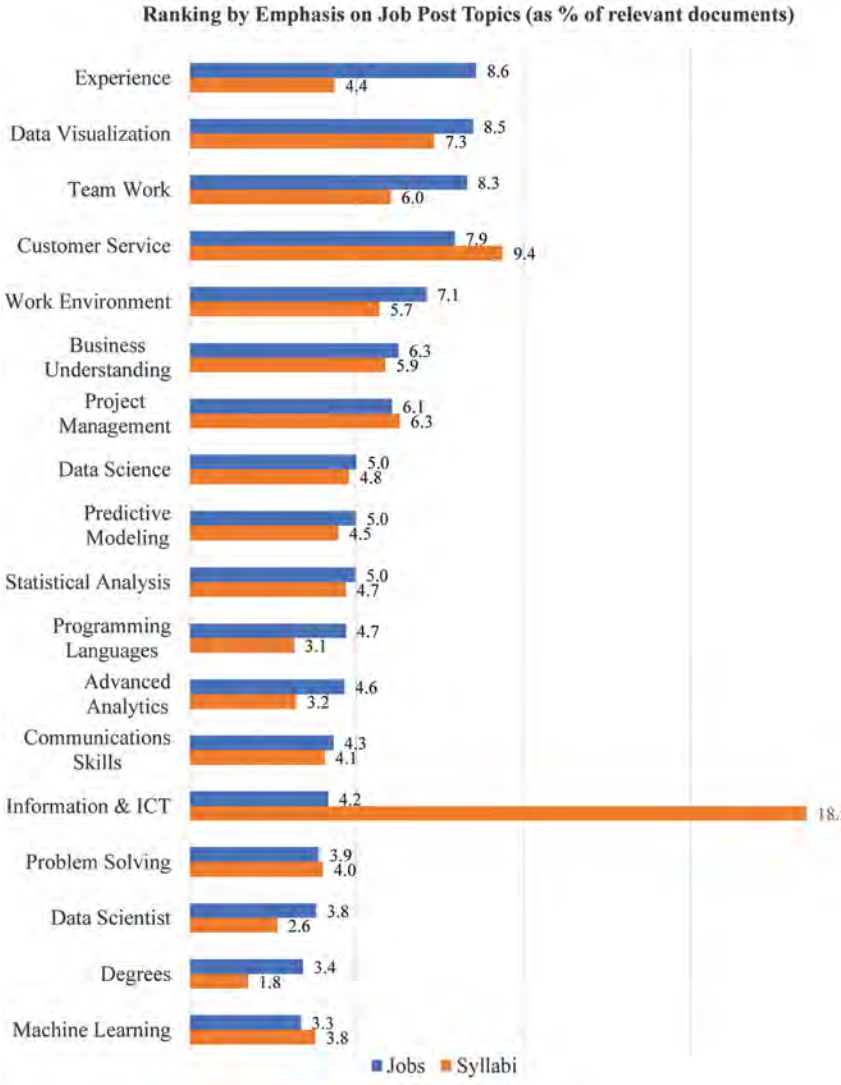


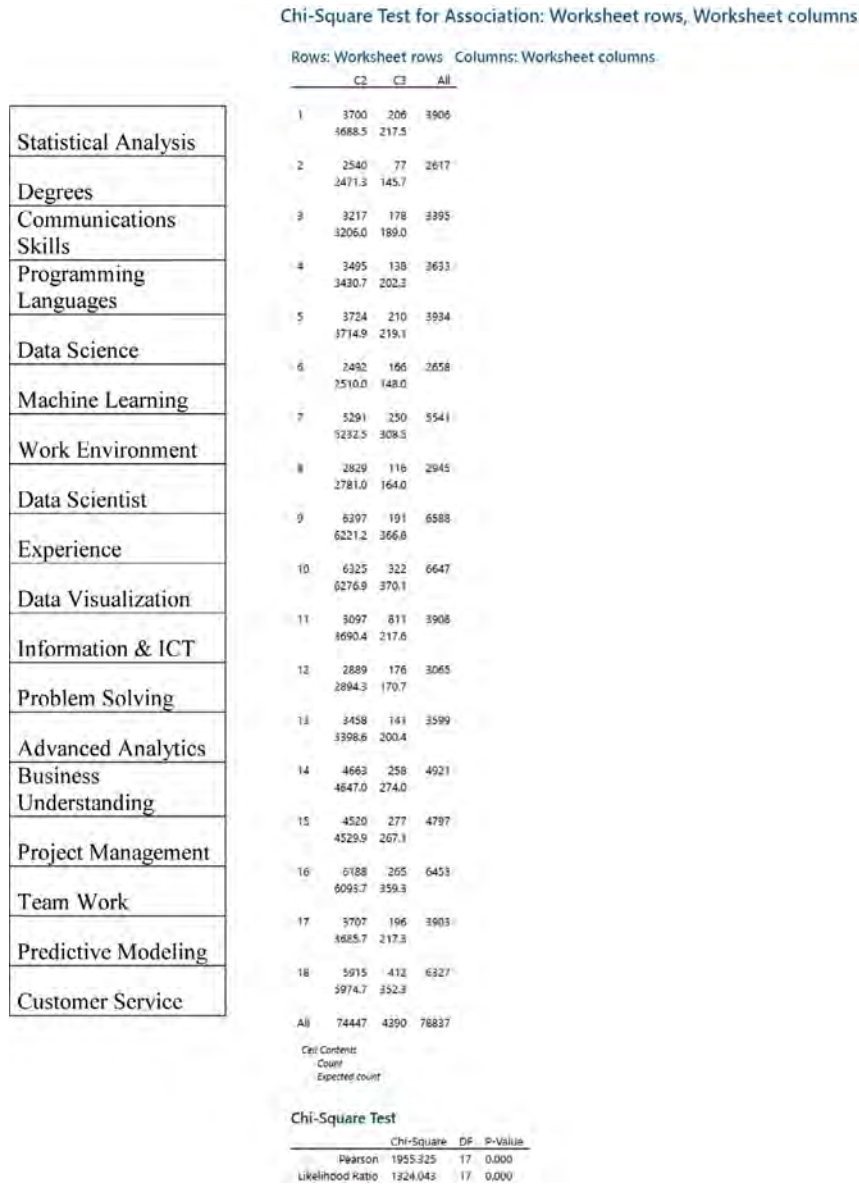
Figure 11: Ranking by emphasis on job posting topics

Based on this result it appears there is quantitive support that the data measured do not follow a standard normal distribution. The two sentence clusters against which text analysis was performed were confirmed by a chi square analysis to have unequal topic frequency. Measurements of topic frequency and ranking will be discussed in the following section.

Figures 14 and 15 display ranked percentiles of topic occurrence to provide an additional visual representation of topic ranking by frequency

Figure 12: Ranking by emphasis on syllabi topics

in both the job postings and the syllabi. The results show that 11.11% of professional data-science topics have the same ranking between data-science job posts and iSchool syllabi learning objectives by emphasis on topic. The topic "Business Understanding" is ranked sixth in frequency by topic in both groups. Additionally, the topic "Data Science" was ranked eighth



Statistical Analysis
Degrees
Communications Skills
Programming Languages
Data Science
Machine Learning
Work Environment
Data Scientist
Experience
Data Visualization
Information & ICT
Problem Solving
Advanced Analytics
Business Understanding
Project Management
Team Work
Predictive Modeling
Customer Service

Chi-Square Test for Association: Worksheet rows, Worksheet columns

Rows: Worksheet rows Columns: Worksheet columns

	C2	C3	All
1	3700	206	3906
	3688.5	217.5	
2	2540	77	2617
	2471.3	145.7	
3	3217	178	3395
	3206.0	189.0	
4	3495	138	3633
	3430.7	202.3	
5	3724	210	3934
	3714.9	219.1	
6	2492	166	2658
	2510.0	148.0	
7	5291	250	5541
	5232.5	308.5	
8	2829	116	2945
	2781.0	164.0	
9	6397	191	6588
	6221.2	366.8	
10	6325	322	6647
	6276.9	370.1	
11	3097	811	3908
	3690.4	217.6	
12	2889	176	3065
	2894.3	170.7	
13	3458	141	3599
	3398.6	200.4	
14	4663	258	4921
	4647.0	274.0	
15	4520	277	4797
	4529.9	267.1	
16	6188	265	6453
	6095.7	359.3	
17	3707	196	3903
	3685.7	217.3	
18	5915	412	6327
	5974.7	352.3	
All	74447	4390	78837

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	1955.325	17	0.000
Likelihood Ratio	1324.043	17	0.000

Figure 13: Chi-square analysis of sentence counts

within both groups of collected documents. Once it was determined there were no other topics that shared the same ranking in either document collection, topics were measured to determine if any were close in ranking. The results of this analysis confirmed that 33.3% of the topics were ranked within one row of each other. These topics were "Data Visualization,"

Machine Learning	3.35%	3.78%
Totals	100.00%	100.00%

Figure 14: Topic frequency ranking by job postings

Topic	Jobs	Syllabi
Information & ICT	4.16%	18.47%
Customer Service	7.95%	9.38%
Data Visualization	8.50%	7.33%
Project Management	6.07%	6.31%
Team Work	8.31%	6.04%
Business Understanding	6.26%	5.88%
Work Environment	7.11%	5.69%
Data Science	5.00%	4.78%
Statistical Analysis	4.97%	4.69%
Predictive Modeling	4.98%	4.46%
Experience	8.59%	4.35%
Communications Skills	4.32%	4.05%
Problem Solving	3.88%	4.01%
Machine Learning	3.35%	3.78%
Advanced Analytics	4.64%	3.21%
Programming Languages	4.69%	3.14%
Data Scientist	3.80%	2.64%
Degrees	3.41%	1.75%
Totals	100.00%	100.00%

Figure 15: Topic frequency ranking by syllabi

"Statistical Analysis," "Predictive Modeling," "Communication Skills," "Data Scientist," and "Degrees." These close rankings suggest that there is an existing relationship between what is being presented in iSchool graduate-level syllabus academic learning objectives and professional data-science job postings. Lastly, the topic analysis confirmed that 27.8% of topics in either the syllabus or job posting sentence analysis were within two rows of one another. These topics were "Team Work," "Customer Service," "Project Management," "Work Environment," and "Problem Solving." Therefore, a total of 72.2% of the topics collected were related in the syllabus learning objective content and the job posting requirements for data-science roles. This leaves only 27.8% of the topics ranked in either group unrelated to one another in terms of ranking by topic frequency. These were "Experience," "Advanced Analytics," "Information & ICT," "Programming Languages," and "Machine Learning."

A comparison of topic frequency reveals that the topic "Information and ICT" is more emphasized in the collected iSchool syllabi than in data

science. The topic "Experience" appears to be weighted more heavily in the collected data-science job postings than in the syllabi. Since many iSchool institutions require an internship component for students completing their information-science degree, this is perhaps an opportunity to provide information-school students with additional hands-on experience with professional data-science skills and tools. The tools aspect of professional data science seems also to be further emphasized by the fact that the topic "Programming Languages," while ranked in both document groups, is ranked higher in the job postings than in the syllabi. On the other hand, the syllabi appear to have more emphasis than the job postings on the topic "Machine Learning," which could be perceived as an opportunity for iSchools to more formally partner with data-science employers and education providers from other disciplines such as data science and business intelligence to emphasize the importance of machine learning on the future growth of the professional data-science community. This is especially important since machine learning was presented as an extremely in demand skill for data scientists by Jeff Hale (2018) in his article "The Most in Demand Skills for Data Scientists." Figure 14 outlines the response to Research Question #1 by providing a comprehensive list of the skills and competencies outlined in data-science job postings. Additionally, Figure 15 outlines the collected text-response frequency to Research Question #2. In Figures 14 and 15 one can see the frequency of collected skills and competencies in either area that overlap in frequency between syllabi and job posts. Therefore, one can safely assume that there is a high degree of overlap between the skills and requirements for professional data-science roles and the academic iSchool curriculum for information-science professionals. This leads to the conclusion that iSchools' curriculum does successfully map to data-science job postings available in the US job market. However, additional evaluation still needs to be done to gain a more accurate picture of how academic information science and professional data science are related.

Discussion and future research

Based on the conclusions reached in the previous sections, there appears to be a relationship between specific topics presented in both graduate-level iSchool syllabi and data-science job postings when sentences are broken down and analyzed for topic frequency in either document group. Since this research was intended to be exploratory, being the first of its kind, additional research is necessary to fully grasp the scope of the relationship and determine how iSchools can best meet the current needs of the professional data-science community. Additionally, it could be beneficial to further investigate surveys conducted with data-science professionals and employers to determine if there are professional needs that they believe are not being met by current academic data-science course offerings.

The initial hypothesis is that there are areas within the collected data-science job postings that are not being sufficiently covered by data-science learning objectives, which possibly presents an opportunity for iSchool students to fill this gap in the employment sector. This would likely require a formal partnership between data science and iSchools both in the employment sector and with respect to the education requirements for either degree. Perhaps both professional and academic data-science and information-science communities can reach a common and shared theoretical framework based on similar, if not the same, foundational concepts, whether the context is data or information.

Both the data-science and information-science professional and academic communities could benefit from a mutually beneficial relationship that defines both areas as a set of related subject matter with a shared set of separate defining concepts. Without conducting additional research, it is difficult to speculate about what areas, if any, are not currently being covered within the learning objectives of data-science courses, so the results of this type of research would define the next steps in this research area.

As with any study, there were limitations to this analysis. This research does not include all worldwide iSchool graduate-level syllabi. Additionally, a single measure was used to assess both data-science career opportunities and iSchool syllabi. For professional data-science topic evaluation, job postings were selected to measure data-science skills and competencies, while iSchool syllabi learning objectives were used to measure iSchool standards of learning. Due to time limitations, this research did not include any additional measures, but this could be expanded in future research. While every U. iSchool institution was contacted by email multiple times through multiple email addresses at each institution, not every institution responded, and additionally, many of those who did respond did not have a formal process in place for sharing syllabi. Some states, such as Texas, have strict laws for public universities about this type of information sharing and require schools to publicly share electronically, while other institutions and states do not require this type of public information sharing. This made it extremely easy to access some syllabi and difficult to access others. Additional research could easily be conducted to compare additional iSchool syllabi if a formal repository were to be set up via the formal iSchool executive leadership. I was informed that something like this was already being discussed, so this a reasonable possibility in terms of future research possibilities. If more schools digitize their syllabi in the United States, this research could easily be built upon to include additional iSchools on a case-by-case basis as well, or expanded to include non-iSchools with information-science and/or data-science programs, both at the master's level and otherwise. Also, as this analysis was conducted using only US iSchool graduate-level syllabi and US data-science job postings, it could be replicated for other countries to determine what differences, if any, exist

between the US-only version of this research and data-science markets in other countries.

Another potential limitation of this research is that analysis was conducted using data-science job postings only. Information-science job postings could also be included in the analysis to further determine what gaps, if any, exist between current traditional information-science job postings and course offerings, especially nontraditional iSchool graduate opportunities. Additionally, it would be interesting to survey recent iSchool graduates and current students to determine their perspectives on future employment prospects and current ones for those students already employed. Also, it is important to track the unemployment rates of recent iSchool graduates to determine if they have any difficulties finding job opportunities upon graduation within the traditional iSchool career tracks. As a follow-up, it is equally important to track iSchool students who have selected to take their iSchool training and enter the data-science career realm. Tracking these types of iSchool students using longitudinal methods is essential for determining if additional nontraditional career opportunities exist for iSchool graduates.

Conclusion

In conclusion, an analysis of topics within current US iSchool graduate-level syllabi course objectives and data-science job postings was conducted. These documents were collected over a period of nine months and comprised over 400 syllabi and over 1,500 data-science job postings. Each was broken down to the sentence level to determine if a relationship existed between the two document groups, with a topic comparison conducted using SAS Enterprise Miner. Based on the results of this analysis, a relationship is evident between data-science job postings and iSchool graduate-level syllabi. This presents a tremendous opportunity for additional research in this area and provides additional opportunities for iSchool institutions to react to the increasing growth of the data-science profession.

Through the collection and analysis of the text of data-science job postings and iSchool course syllabi, the original research questions were answered. The skills and competencies most sought after by the professional data-science community were indicated by the textual analysis performed on over 1,500 US data-science job postings in response to Research Question #1. Additionally, the skills and competencies taught in the iSchool curriculum during the same period provided a data set for Research Question #2. In conclusion, it appears that there is indeed overlap between iSchool learning objectives and current data-science professional job offerings.

Angel Durr is a first-generation college student who works as a strategic data consultant as well as nonprofit founder now that she has completed her PhD. Her research interests include data and information literacy, data science, information systems, entrepreneurship, community development, education, and career exploration.

References

Anslow, C., Brosz, J., Maurer, F., & Boyes, M. (2016). Datathons: An experience report of data hackathons for data science education. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (pp. 615–620). ACM. https://doi.org/10.1145/2839509.2844568

Burke, J. W. (1989). *Competency based education and training*. London, England: Falmer Press.

Burton, M., & Lyon, L. (2017). Data science in libraries. *Bulletin of the Association for Information Science and Technology*, *43*(4), 33–35. https://doi.org/10.1002/bul2.2017.1720430409

Cambridge International Examinations. (2017). Topic support guide: Topic 1.1—Data, information and knowledge. Retrieved from http://www.cambridgeinternational.org/images/285017-data-information-and-knowledge.pdf

Carmody, B. (2016). Biggest problem with big data management in 2016. Retrieved from https://www.billcarmody.com/biggest-problem-big-data-management-2016/

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), 1165–1188. https://doi.org/10.2307/41703503

Chen, H.-M., Schütz, R., Kazman, R., & Matthes, F. (2017). How Lufthansa capitalized on big data for business model renovation. *MIS Quarterly Executive*, *16*(1), 19–34.

Cormen, T. H., Leiserson, C. E., & Rivest, R. L. (1995). *Introduction to algorithms*. Cambridge, MA: MIT Press.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, *34*(1), 313–327. https://doi.org/10.1016/j.eswa.2006.09.038

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic indexing. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(sici)1097-4571(199009)41:6%3C391::aid-asi1%3E3.0.co;2-9

Hale, J. (2018). The most in demand skills for data scientists. *KD Nuggets*. Retrieved from https://www.kdnuggets.com/2018/11/most-demand-skills-data-scientists.html

Indeed (2017). Our company. Retrieved from https://www.indeed.com/about/our-company

iSchools. (n.d.). iSchools. Retrieved from https://ischools.org/

Landauer, T. (2007). LSA as a theory of meaning. In T. Landauer, D. McNamara, D. Simon & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 300–306). Mahwah, NJ: Lawrence Erlbaum.

Marchionini, G. (2016). Information science roles in the emerging field of data science. *Journal of Data and Information Science*, *1*(2), 1–6. https://doi.org/10.20309/jdis.201609

Morgan, A., Duffield, N., & Walkley, H. L. (2017). Research data management support: Sharing our experiences. *Journal of the Australian Library and Information Association*, *66*(3), 299–305.

Ortiz-Repiso, V., Greenberg, J., & Calzada-Prado, J. (2018). A cross-institutional analysis of data-related curricula in information science programmes: A focused look at the iSchools. *Journal of Information Science*, *44*(6), 768–784. https://doi.org/10.1177/0165551517748149

Robinson, L., & Bawden, D. (2017). "The story of data": A socio-technical approach to education for the data librarian role in the CityLIS library school at City, University of London. *Library Management*, *38*(6/7), 312–322. https://doi.org/10.1108/lm-01-2017-0009

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Song, I.-Y., & Zhu, Y. (2017). Big data and data science: Opportunities and challenges of iSchools. *Journal of Data and Information Science, 2*(3), 1–18. https://doi.org/10.1515/jdis-2017-0011

Sutherland, L., & Wildgaard, L. (2016). The evolution of the information professional: Thoughts on innovation in "librarianship" and practical solutions for future candidates from IVA. *Revy, 39*(1), 17–18. Retrieved from https://rauli.cbs.dk/index.php/revy/article/view/4977

Tonta, Y. (2016). Developments in education for information: Will "data" trigger the next wave of curriculum changes in LIS schools? *Pakistan Journal of Information Management & Libraries, 17*. Retrieved from https://www.researchgate.net/profile/Yasar_Tonta/publication/283722068_Developments_in_Education_for_Information_Will_'Data'_Trigger_the_Next_Wave_of_Curriculum_Changes_in_LIS_Schools/links/5644f33808ae9f9c13e5a88b/Developments-in-Educationfor-Information-Will-Data-Trigger-the-Next-Wave-of-Curriculum-Changes-in-LIS-Schools.pdf

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics, 34*(2), 77–84. https://doi.org/10.1111/jbl.12010

Winson-Geideman, K., & Evangelopoulos, N. (2013). Topics in real estate research, 1973–2010: A latent semantic index. *Journal of Real Estate Literature, 21*(1), 59–76.

Zhang, V., & Neimeth, C. (2017). 3 reasons why data scientist remains the top job in America. InfoWorld. Retrieved from https://www.infoworld.com/article/3190008/big-data/3-reasons-why-data-scientist-remains-the-top-job-in-america.html