2022

Data Science: A Study from the Scientometric, Curricular, and Altmetric Perspectives

Michael Joseph King

Data Science: A Study from the Scientometric, Curricular, and Altmetric Perspectives

A Dissertation

Submitted to the Faculty

of

Long Island University

by

Michael Joseph King

In partial fulfillment of the requirements for the degree

of

Doctor of Philosophy

April 2022

Dedication

To my parents for giving me the heart and strength to continue moving forward, always buying me books, and answering the phone for the good calls or the long hard car rides home. Also, to my brother for always supporting me and knowing that I will have someone who has my back no matter the time or the distance.

Acknowledgments

As my road through education has gotten longer, my respect for those at the pinnacle of academia has only grown. My committee has exemplified what it means to be an educator, and I can’t thank them enough for the countless hours they have deemed me worthy of giving.

Firstly, a deep heart-felt thanks to my main advisor, Dr. Heting Chu. Draft upon draft, her work with me has helped me learn about myself, my voice, and my education in innumerable ways. As the first person to read all my drafts, I can only thank you for guiding me and being patient with my pacing and writing.

Dr. Nicholas Olijnyk, my external advisor, thank you for being someone that has been not only a supportive committee member during my dissertation but before. I found you to be a role model for years now and appreciate your continued help here on this endeavor. During our discussions, Dr. Selenay Aytac always posed questions that have helped me shape my vision of research and the value each and every element holds in understanding the greater context, both to the work at hand and my education. Dr. Bea Baaden, thanks for so much even before this dissertation; your presence on my committee has provided continuity to my education and a pragmaticism to my schooling. Dr. David Jank always presented novel ways of looking at exciting topics, inspiring me to look beyond my own two hands and at the possibilities of the “could be.”

Finally, thank you to all my friends and family for giving me the support, the space, and sometimes the kick to complete my ventures and follow my aspirations.

Abstract

This research explores the emerging field of data science from the scientometric, curricular, and altmetric perspectives and addresses the following six research questions:

1. What are the scientometric features of the data science field?
2. What are the contributing fields to the establishment of data science?
3. What are the major research areas of the data science discipline?
4. What are the salient topics taught in the data science curriculum?
5. What topics appear in the Twitter-sphere regarding data science?
6. What can be learned about data science from the scientometric, curricular, and altmetric analyses of the data collected?

Using bibliometric data from the Scopus database for 1983 – 2021, the current study addresses the first three research questions. The fourth research question is answered with curricular data collected from U.S. educational institutions that offer data science programs. Altmetric data was gathered from Twitter for over 20 days to answer the fifth research question. All three sets of data are analyzed quantitatively and qualitatively.

The scientometric portion of this study revealed a growing field, expanding beyond the borders of the United States and the United Kingdom into a more global undertaking. Computer Science and Statistics are foundational contributing fields with a host of additional fields contributing data sets for new data scientists to act, including, for example, the Biomedical and Information Science fields. When it comes to the question of salient topics across all three aspects of this research, it was revealed that a large degree of coherence between the three resulted in highlighting thirteen core topics of data science. However, it can be noted that

Artificial Intelligence stood out among all the other groups with leading topics such as Machine Learning, Neural Networks, and Natural Language Processing.

The findings of this study not only identify the major parameters of the data science field (e.g., leading researchers, the composition of the discipline) but also reveal its underlying intellectual structure and research fronts. They can help researchers to ascertain emerging topics and research fronts in the field. Educational programs in data science can learn from this study about how to update their curriculums and better prepare students for the rapidly growing field. Practitioners and other stakeholders of data science can also benefit from the present research to stay tuned and current in the field. Furthermore, the triple-pronged approach of this research provides a panoramic view of the data science field that no prior study has ever examined and will have a lasting impact on related investigations of an emerging discipline.

Keywords: Bibliometrics, Plural Methodology, Qualitative Research, Quantitative Research, Social Study of Science

Table of Contents

List of Figures

List of Tables

1. Introduction

The deluge of information flooding societies and institutions has been changing the face of the world. Whether the conversation is about private business, academic discoveries, or government organizations, the dialog often contains discussions and assessments of information. Often labeled as big data, this data results from people's innate need to record and quantify all things, from posts on social media to records of the movement of celestial bodies through the sky. To meet the new demands of this tremendous quantity of data, an entire field of data science has emerged from academia. The development of a host of new sensors, software, and other technology to assist in recording audio, video, and other data has made the process of data collection so streamlined that data volume has grown exponentially. Leading news outlets have been heralding the call to investors with the succinct statement: "Data is the new oil." (Rotella, 2012). It was not only the private industry that recognized the value in utilizing data science to harness big data. Setting the trend, the more recent Trump 2016 campaign decision to hire data science company Cambridge Analytics has been a massive signal to the political world that data science is here to stay in the political arena. Cambridge Analytics implemented a hybrid approach to breaking the population down by traditional political metrics, psychological profiles, and consumer habits, providing a unique and new approach to data-driven campaigning that has played a large role in his election (Tett, 2017). It is safe to say that data science has permeated various aspects of society. The field’s future position is safely solidified in a tomorrow full of data science-powered technologies.

Data sciences' unprecedented rise to meet the fast exponential and accumulating amounts of data is essential to understanding the approaches of many aspects of science, business, and government. Big data is this human-constructed tidal wave that data science is designed to meet.

Understanding what is meant when discussing big data is often illustrated by utilizing the 5Vs Model of Big Data. This model describes big data as being data unique in volume, velocity, variety, veracity, and value. For some researchers, the definition of big data is data that is not manageable by traditional approaches or systems (Dumbill, 2012; Tambe, 2014). These unique attributes of big data and the inability of older traditional techniques to handle this new breed of data forced practitioners to look to computer science, mathematics, and other academic fields, searching for answers and tools.

Data science is not something genuinely new; many researchers, governments, and corporations have sought methodologies and techniques to use large data sets for decades. However, the ease at which these massive, fast-moving, and highly complex data sets are now being created has expanded, and data science has flourished alongside it. Both data science and big data have been sensationalized in the media and attributed to new technologies and novel approaches as if they have appeared from nothingness. The truth is more akin to the fact that big data has deep, historical roots. Department stores, international organizations, and governments constitute the early stages of data science. From a more historical perspective, it is easy to understand why in large part, private industry has been quick to adopt big data to harness its competitive edge through data science (Jin, Wah, Cheng, & Wang, 2015).

The term data science refers to the work that big data precipitated. It encompasses a host of subfields focusing on dealing with the various facets and approaches to handling big data in various implementations and environments. Data science's tools range from analyzing text data to complex, highly technical work dealing with the computation and network configurations designed to support massive databases. Data science's growth has been highly accelerated, largely thanks to the hype around big data. However, the field itself has now taken the spotlight.

That community focus has seen the expansion of university classes and the creation of departments preaching the word of data science. Private industries' interest must be taken note of as well, as the large-scale investment of companies in data scientists and their departments with the sole purpose of understanding internal and external data sources. For companies, the idea of data-driven decision-making (DDD) is also a trend that has been empowered by the new wealth of mission-critical data being parsed by data science from big data.

Alongside the data explosion, research regarding data science has also grown commensurately. The full range of academic disciplines that have contributed to the base knowledge in developing this research body is even more exciting and of importance to this research. As a relatively new area of research, understanding the contributing fields to data science research is poised to be valuable to understanding current research and perhaps even forecasting future research direction. The need for researchers and practitioners to have a firm grasp on the various data science research occurring worldwide can only aid the field. This research can be a means to provide professionals with data science awareness and proffer a more global understanding of where research is coming from, generated by who, and through which institutions. Coupled with the growth of research is the development of educational programs geared toward preparing the next generation of practitioners and academics. Research presented in this study provides a strategic resource for program designers and instructors alike.

Scientometrics will be the primary technique utilized for this research, along with the secondary use of and supplemented using content analysis and altmetrics. The more macroscopic view provided by bibliometric techniques examining the relationships between research publications, citations, institutional affiliations, and even national contributions can provide a deep and meaningful overview of where data science is in terms of subject matter composition and how it might continue to expand. Scientometrics is a well-established method that many research approaches have utilized in studying a scientific discipline. In addition, the use of content analysis for examining educational materials from institutions offering data science programs and degrees will provide an additional perspective to the bibliometric view. To further assist in laying out the research and conceptual understandings of data science from a greater social scope, the inclusion of altmetrics seeks to harness the plethora of data found on Twitter. The coupling of traditional scientometric/bibliometric data, curricular analysis, and altmetric data enabled the present researcher to obtain a fuller picture of the data science landscape. Therefore, this research seeks to answer the following research questions:

RQ1. What are the scientometrics features of the data science field?

RQ2. What are the contributing fields to the establishment of data science?

RQ3. What are the major research areas in the data science discipline?

RQ4. What are the salient topics taught in the data science curriculum?

RQ5. What topics appear in the Twitter-sphere regarding data science?

RQ6. What can be learned about data science from the bibliometric, curricular, and altmetric analyses?

2. Background

In this chapter, a brief excerpt of background knowledge will be presented to assist in understanding how data science relates to big data. Additionally, this section will provide some background descriptions of scientometrics and altmetrics, two of the methods adopted for this research. The third is addressed further in chapter three. This section will not discuss the implementation of methods in this research, as that is the topic for the methodology chapter of this proposal.

2.1 Data Science

As society finds itself adrift in a sea of information, institutions have come to terms with navigating these new oceans of data with various tools and approaches. Big data has accelerated the emergence of the field of data science. This section seeks to illustrate the role data science has in research, enterprise, and academia and provide a basic understanding of what it is and its potential. This section will also highlight some of the more prominent elements of data science but by no means attempt to be comprehensive in its discussion of data science tools, procedures, and methods.

Cao (2017, 43) presents a tremendous amount of research on data science along with its definition:

> "From the disciplinary perspective, data science is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments (including domains and other contextual aspects, such as organizational and social aspects) in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology."

Operating under Cao's definition, two things are clear. First and foremost, the contributing fields of data science are varied and interconnected. Second, data science is chiefly concerned with the "study of data." Therefore, data science encompasses a wide range of data-centric research topics, including data management, data visualization, data privacy, social aspects of big data, and many other data-related topics. However, data science, at its core, is the practice of extracting knowledge and information from data and has been developed from the theories and practices of several other scientific fields (Wu and Chin, 2014).

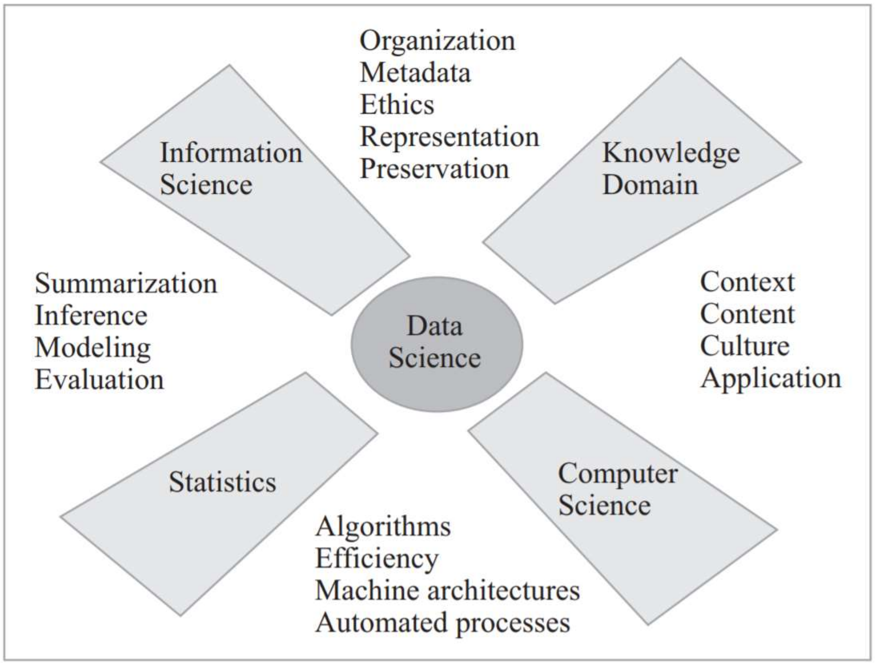


Figure 2.1. Contributing fields to Data Science (Marchionini 2016, 3)

Figure 2.1 from Marchionini (2016, 3) presents a visually clear example of four of the core academic domains that have contributed to the development of what is now deemed data science and some of the major contributing areas of study.

As far as data science is concerned, its notoriety has been in the shadow of big data as a term. However, this is becoming less and less the case. The meteoric rise of data science as a relevant term can be seen clearly through Google Trends' charting of the terms "big data" and "data science" from January 2004 to the present. Figure 2.2 illustrates a strong case for data science growth and its potential overtake of big data as a search term at Google, the world's largest web search engine (Google Trends, 2018).

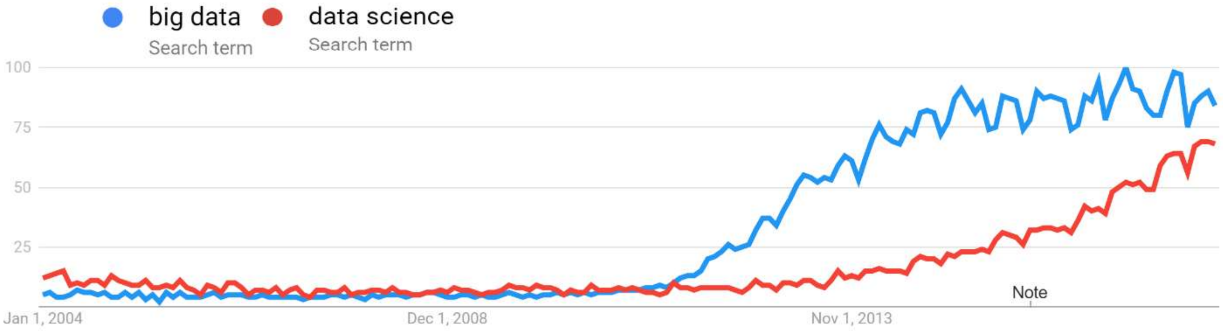


Figure 2.2. Google Trends chart of "big data" and "data science"

In this research, it is crucial to understand that the scope of implementation of data science is massive. Data science work is occurring across various topics, domains, and industries. Each of these areas of work and study is developing techniques and their applications, which presents a uniquely vast range of work being done by data scientists. This extensive assortment of work is a substantial contributing factor to the need for this research. This research will help those concerned with its development and evolution to understand the changes it is incurring across industries and domains of study.

This wide range of research has brought studies like Waller and Fawcett's (2013) to explicitly note that domain knowledge of the research focus in today's data science practice cannot be separated. It is now commonly understood that data scientists need to have the skills to analyze data and understand the context and environment in which this data resides to leverage analysis and findings fully. Data science adapts its methods and techniques to study data-centric phenomena within the context of other related fields. Data scientists must be acutely aware of the relevant research topics within which they operate to provide usable and actionable data analysis. Many researchers refer to this data environment as context and note that domain knowledge pertinent to the data is paramount to proper data analysis.

Understanding the context is important because the range of these topics has become enormous. Chen and Zhang (2014) mention several scientific fields that are highly data-driven: astronomy, meteorology, social computing, bioinformatics, and computational biology. Additionally of note is the cross-disciplinary trend that Cao (2017) points out as a new trend in data-driven discovery and science in what he calls the phenomenon of *x-informatics.* Astro-informatics, behavior informatics, bioinformatics, biostatics, brain informatics, health informatics, and medical informatics are just a few he brings up. The range, scope, and depth of influence that data science and its methodologies have brought to fields are on their own a powerful and revolutionary effect on science. The ability of data scientists to harness large data sets structure and do cursory analysis is alone a potent technique. However, their ability to programmatically harvest metadata and automate the analysis of trends, understandings and relationships from the data deluge is changing how business and science are being done.

One of the compelling aspects of big data is its potential for increasing efficiency and effectiveness when it comes to organizational structures and systems. In scientific research, this massive data flood has resulted in data-discovery techniques and approaches, and in many industries, this is referred to as data-driven decision-making. For corporations, this can increase productivity by analyzing patterns and data (Manyika et al. 2011). This potential for enhancement is perhaps one of the biggest reasons so many organizations are scrambling to find and utilize the big data "edge" over their competition. The industry-level demand for data scientists has exploded with the range of industries attempting to implement its newest methodologies. Along with implementing existing industries, the rise of information analytics as an industry has been equally fast and massive.

Overall, the influence of data science has been staggering. An in-depth scientometric study was needed to objectively document and explain the impact of data science across society, be it industry, education, or scientific study. This research can give scholars a look at the intellectual structure of data science as it has developed and allow for insights into the domain's future trends. The data science field will continue to expand into other sciences and industries, which will also be able to use the information provided within this study to better position themselves.

2.2 Data Science vs. Big Data

Big data has served as a powerful motivator for the development of data science. The defining lines separating big data and data science have been blurred. The difference has hardened far more in academia since the earlier days of the information explosion known as the information deluge. First and foremost is the conceptual idea of defining big data. Manyika et al. (2011) define big data as data too big and moving too fast to be processed by conventional database systems and other technologies. This research will expand on big data's definition and use an over-arching framework, common referred to as the 5V model, about big data's characteristics to help facilitate the delineation from the data science perspective.

Understanding big data through its more prominent traits will provide a solid working definition for this research, especially where there might be confusion in discussions between data science and big data. The present researcher will be utilizing the 5V framework of big data as a lens to understand its role in data science. These fundamental core concepts provide big data with its conceptual shape. While the specifics of what, where, and how these boundaries may exist from an exact standpoint are ambiguous, this framework provides a skeletal framework to better understand.

Paramount to this research is the understanding that big data is not the same as data science. Instead, data science is how big data is stored, retrieved, searched, and analyzed by researchers and scientists in academia and industry.

2.3 Scientometrics and Altmetrics

2.3.1 Scientometrics

Through qualitative and quantitative means, scientometrics provides an examination of scholarly domains and fields regarding, among other things, their formation, development, and interactions internally as well as externally. Scientometrics has moved from an obscure region of study to a very important sub-field of Information Science, playing a major role in understanding scientific development and related topics (Vinkler 2010; Mingers and Leydesdorff 2015). Derek de la Solla Price (1965) was one of the pioneer researchers to see the value of studying the communications between researchers using scientometric approaches.

Since those early days in the field's development, many research papers have referred to scientometrics as the science of science or the social study of science. Scientometrics has been defined as having an integral role in understanding the development of science, especially under the perspective of science "as an informational process" (Nalimov 1971). However, of particular importance to this research is the value that scientometric studies can provide in identifying subtle interactions among academic domains.

One feature of scientometric analysis is the visualization and mapping of relationships in a domain or field that can otherwise be difficult to observe. Scientometrics has also been used to study science policies, research collaboration, individual researchers, institutions, and countries (Anson 2016; Perron et al. 2016). This research seeks to instantiate scientometric methods to aid in a greater understanding of data science.

Even scientometric work has benefited from data science in developing close inter-disciplinary practices that utilize and harness new, more powerful computing science approaches and computing power. The recent development of technologies and tools around scientometrics has revolutionized the method but still primarily revolves around the “core notion” of citations and publications as a means of measure (Mingers and Leydesdorff 2015). While scientometrics is not confined to citations and publications, bibliometrics is used quite commonly in scientometric research. In this study, the terms scientometric and bibliometrics will be treated interchangeably.

Citations and publications, two major kinds of bibliometric data, provide an effective way to examine research at various levels. Within scientometrics, the scope of studies looking at scholarly works can range from examining publications attributed to an individual scholar to the relationships between nation-states and their academic production. The addition of much of this new technology-aided work has helped researchers visualize areas of study. These mappings have gone a long way in helping researchers conceptualize and understand the clustering and relatedness of fields, topics, and concepts. Even today’s standard computers can handle these complex networks that may contain upwards of thousands of data points, all having interconnecting relationships. The ability of researchers to visualize these connections with ease and great speed has brought these nuanced techniques to the mainstream.

The intended use of bibliometric methods to perform this scientometric study carries some distinct boundaries between the two concepts. Olijnyk’s (2014) diagram of the relationship between scientometrics and bibliometrics sets the two apart and emphasizes the fact that scientometrics employs bibliometric techniques for its analysis of science (see Figure 2.3).

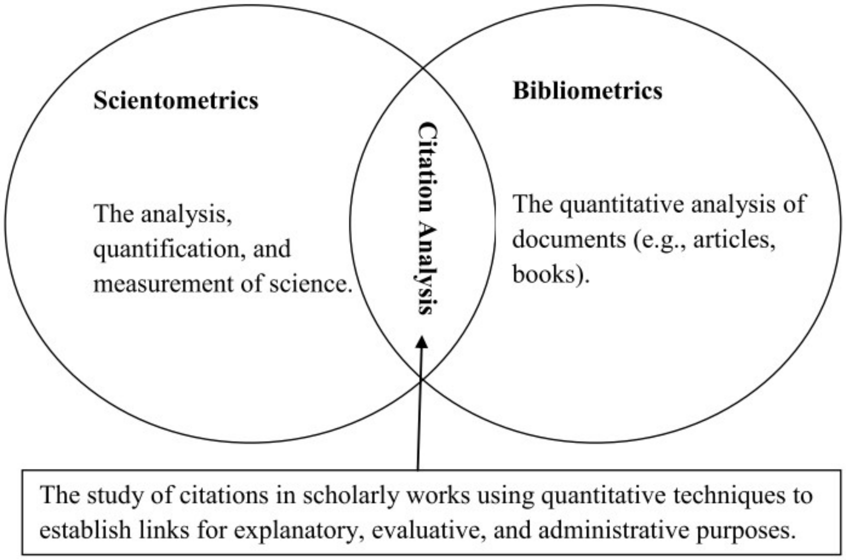

Scientometrics
The analysis, quantification, and measurement of science.
Citation Analysis
Bibliometrics
The quantitative analysis of documents (e.g., articles, books).
The study of citations in scholarly works using quantitative techniques to establish links for explanatory, evaluative, and administrative purposes.

Figure 2.3. The relationship between Scientometrics and Bibliometrics (Olijnyk 2014, 16)

Bibliometrics has been routinely attributed to the novel work of Price and Eugene Garfield (Godin 2006). Researchers have explained bibliometrics as applying quantitative tools to study scientific communications (Pritchard, 1969; Leydesdorff, 1995; Liu et al., 2015). Narin, Olivastro, and Stevens (1994) define bibliometrics as counts of publications, patents, and citations used to generate scientific indicators. Much of bibliometrics's raw power comes from examining citations and generating article-level relational maps of research, lists, and theories based on groups of publications. These relational maps are called citation diagrams and are a means by which bibliometrics can visualize relationships between authors, journals, academic institutions, and countries to help reveal scientific work's underlying structures.

2.3.2 Altmetrics

Altmetrics, short for alternative metrics, is a research technique geared towards harnessing data and metadata collected through social networking platforms. It aims to accurately analyze and examine scholarly output through the measurement of shares, likes, downloads, saves, tweets, reviews, and other measures beyond the traditional citations and publications (Zahedi, Costas, and Wouters 2014; Costas, Zahedi, and Wouters 2015; Galligan and Dyas-Correia 2013). Altmetrics was a term coined in 2010 by Jason Priem in a tweet (Priem 2010); it would become a research term that quickly gained traction across academia. Galligan and Dyas-Correia (2013) examined a cross-section of varying definitions for altmetrics, highlighting key aspects characterizing it to harness the data derived from the readership, sharing, likes, bookmarking, and various other interactions to mine for relevant communication and scholarship on web driven platforms. Additional value is derived from the fact that academic citations, traditionally the crux of bibliometric analysis, among other forms of impact research, only track scholarly communication. Lin and Fenner (2013) point out that only a small percentage of document engagement is truly manifest in its citations, as many people may download a paper and never cite it. Many researchers have agreed with Lin and Fenner, suggesting that altmetrics provides faster data than traditional citation methods that rely on a comparatively slower system of review and publication, and these take a variety of user interactions (Adie and Roe 2013; Bar-Ilan et al. 2013; Bornmann 2014; Haustein et al. 2013; Piwowar 2013;). While there is a discussion about the measure of such data, the variability between nodal types collected from various platforms, and how altmetrics relates to bibliometrics, the reality is that altmetrics are providing new insights and an entirely new frontier for examining the life of scholarly works.

Much of these data measurements are taken from social media and scholarly web platforms. Everything from Twitter and Facebook to the now relatively standard shares and like buttons found on many of the major database systems is utilized to gather data to inform scientists better. Not only are public social media systems being used, but researchers have developed several academic-specific community social platforms like Mendeley and

Academia.edu to collect altmetric data relating to scholarly communication and work. While still developing an understanding of altmetrics, researchers see the wealth of data and the broader scope of information that social media and other platforms can provide, especially when considering the massive data created at those sites (Bornmann 2015; Zahedi, Costas, and Wouters 2014).

Altmetrics was developed out of some significant advantages and some significant needs. The realization is that in this internet age, the speed at which information is moving is significantly faster and through significantly more communication channels when it comes to research. These new digital spaces have provided a new means for researchers to measure academic impact and scientific discovery. The addition of the internet and the expanding nontraditional means of readership and viewership of materials has disrupted classical communication models. The question is now: how do researchers harness this new data that is being generated?

The addition of the internet's massive data and speed advantage has become a focus for academia, especially concerning one of its foremost and controversial discussions: the impact of scientific publications and researchers. For altmetrics, the more significant focus has been on determining these metrics in forums that other methods like bibliometrics would not be able to track, for example, shares and hashtags.

Not only does this method look at many of these other indicators of access, assessment, or action, but this method also works in a much quicker timeframe than many other traditional research methods. In the case of a bibliometric approach, the ability to track a document's effect through citations can be a long and arduous process simply because of the publication and readership lifecycle. However, altmetric-type data like downloads, views, and time on page, just to mention three, can start accruing the moment a resource is posted online. This type of data collection allows for a compelling new view of scholarly work and provides a whole new perspective and timeline for looking at the ingestion of academic works by readership.

Altmetric methods will be used in this research to examine Twitter data. Twitter provides several advantages, the first of which is its massive userbase. The sheer scope of information from users' input into the system provides a plethora of information regarding any topic imaginable. Another advantage of using Twitter is the self-organization that Tweets have with hashtags to denote the messages related to topics, major ideas, and other discussions. Lastly, Twitter has a vibrant ecosystem for developers and researchers to collect data utilizing a robust API. The targeted collection of Twitter posts and their subsequent analysis will hopefully yield insights that can be analyzed alongside the other methodologies in this research to help inform a powerful dimension to the overall scientometric study.

3. Literature Review

This chapter has been divided into three overarching sections relating to the critical aspects of this research. The first contains scientometric research on domains and fields and is further divided into fields unrelated to data science and specific to data science. The second section will resolve to examine previous studies whose work has utilized methods of reviewing higher education curriculums as a means of academic investigation. The third section will contain studies that have executed altimetric methodologies in their research.

3.1 Scientometric Research on Fields

As defined by its intellectual structure, the nature of science has been the primary aspect of information studies in attempts to garner a greater understanding of science, its growth, and its evolution over time for years. The seminal work, *The Structures of Scientific Revolutions* by Kuhn made it clear that scientific development would grow in complexity, as did the corpus of publications in each domain (Kuhn 1962). Similarly, the realization and explosive growth of publications were evident, as was the need to index and grasp the tracking of publication production (Garfield and Sher 1963). This complexity was projected to grow as citation rates increased alongside publication rates, but so too was the need to look deeply at scientific networks developing through citations (Price 1965). This call to action and desire to understand scholarly communication networks in fields has expanded over the years. These studies now range across decades and topics, from research on industrial science (Johnston and Robins 1977) to more recent examinations of the human microbiome (Coccia 2018). The intellectual structure of a field has been characterized by examining the social communication of scholar discourse on various scopes to ultimately give insights to academics, institutions, companies, and organizations to make informed decisions on actions to be taken on impacted interests.

Scientometric research aims to understand the characteristics, processes, and relationships through scholarly networks that ultimately lead to a field's development over time. So, through scientometrics, many researchers have found a means to grasp the complex nature of scientific evolution in publications. Salzano (2018) took an in-depth look at Brazil's scholarly publications from the 1930s to 1999 relating to genetics and genomics and paints an optimistic vision for Brazil's contributions and future contributions to the forefront of genetic research. While historical analysis is compelling, many studies focus on smaller time frames. One such example is the analysis of computer supported cooperative work (CSCW) research, which examined a more compact 15-year time frame from 2001 to 2015 in which key scientists and publications were identified as well as some reasonable projections for the continued growth of the field (Correia, Paredes, and Fonseca 2018). These scientometric studies often employ bibliometrics methods to harness the power of the citation. The citation has proven to be one of the most valuable nodes of research in understanding the topography of scientific communities. Over the years, derivations on citation studies have occurred even from the earliest days, with many studies taking into account various granularities 'larger' than the citation, like authors, journals, institutions, and even topics. Leveraging these different levels of citation metrics, analysis of domain interactions has been carried out in studies like that done to examine the relationship between information systems and College of Business publications (Pratt, Hauser, and Sugimoto 2012). Determining some of these cross-discipline borders gives powerful insights into cross-over research areas and can lead to exciting new fields and understandings for both disciplines.

Its underlying structure traditionally defines a field of science. As fields of study mature over time, they develop topics of study. In a study of information science trends between 2009 and 2016, researchers attempted to identify emerging trends and revealed changes in publication rates of some previously identified core topics and new research topics (Hou, Yang, and Chen 2018). If topics gain enough traction in a field or amongst multiple fields, they can even lead to newly emergent fields. Emergent field research can be powerfully supported through scientometric research, like the bibliometric study of public relations intelligence as an emergent field evolving from the border research between strategic intelligence research and public relations research (Santa Soriano, Lorenzo Álvarez, and Torres Valdés 2018). These fields and topics are tied together through their scholarly discourse in research communities. A study examining physics took a look at defining the research communities by examining their publication activity and their community size and how that might be further related to the age or lifetime of a community. Researchers concluded that it does seem older, more established fields tend to have larger communities and active topics within themselves (Herrera, Roberts, and Gulbahce 2010). Not only is scholarly production key in the development of fields of science, but so too are scholarly communities that develop around crucial aspects of active research operating together on scholarly discourse over a shared core selection of past research. In some cases, these studies even focus on singularly influential community members, like the famous researcher Eugene Garfield. An entire research tribute mapping and analyzing the scientometric fingerprint of Garfield and his work is one such example (Jacso 2018).

Scientometric studies can be significant in providing evidence of newly emerging communities in established scientific areas and provide mappings that define the topology of intellectual structures and publication patterns. Anson (2016) produced a compelling analysis of emerging technology landscapes through citation analysis and argued for the power of scientometric studies to aid subject matter experts in determining where finite resources and efforts should go in studying these new frontiers. Additionally, many scientometric studies seek to shape these fields through visualizations. In a 2017 publication of a scientometric study examining the research fronts in the field of magnetic nanoparticles, researchers (Liu et al. 2017) leveraged their bibliometric citation research over fifteen years to create visual mappings of the literature's cocitations and coword analysis to determine the leading edges of the field. These visualizations can also provide an in-depth look into the history of a field, as the histography constructed around glaucoma research in a 2016 study represents a complex mapping of influential papers and their effects on later publications (Ramin et al., 2016). These visualizations can be constructed in numerous ways but often attempt to group publications into more macroscopic clusters either by topic, journal, or other similar means to aid in understanding field topography. Examining author collaboration networks in the field of scientometrics in one study revealed that the majority of prolific authors were, in fact, members of sub-networks existing within the overall network structure of the scientometric field (Hou, Kretschmer, and Liu 2008). Though, all of these methodologies have similar goals: understanding the nature of the field, laying the groundwork of where research has derived from or making conjectures on its future direction.

The meteoric rise of data science has left some researchers claiming its roots here and others claiming it there. What has mostly been accepted in the literature is that data science's roots are coming from mainly a combination of mathematics, especially the discipline of statistics, computer science, and information science but has leveraged a tremendous amount from a variety of other disciplines (Agarwal and Dhar 2014; Donoho 2017). However, what has become readily apparent is that data sciences methodologies and applications in other domains of science have made it a complex domain with a hugely varied number of contributing fields attempting to harness it for their own field's data sets (Cao 2017). Scientometric has been a powerfully implemented research tool designed to help understand some of these relationships and has been used and validated across a plethora of fields other than data science already.

3.1.1 Scientometric Studies on Fields Other than Data Science

Scientometric methodologies serve as a powerful tool to uncover and examine fields of science, and as such, their processes have been widely applied across the domains of science. While data science is the focus of this research, scientometric studies on other fields, especially in novel and fast-growing fields, in a range of different manners have been carried out and provide an in-depth look at the powerful research methods scientometric geared studies can employ to understand science holistically.

Foundational baseline metrics of scientometrics often come in the form of bibliometric analysis, uncovering some key structural aspects when examining a corpus of publications. In a study of 10,942 records examining 25 years of Antarctic work, Dastidar and Ramachandran (2008) applied scientometric approaches to look at overall productivity in the field, finding an increase of threefold. Not only was overall productivity noted, but so too were prominent authors, organizations, countries, and journals giving a robust map of increasing collaborative efforts among countries and individual authors. A highly focused study by Fatt, Ujum, and Ratnavelu (2010) on the Journal of Finance also applied co-authorship analysis to examine the state of the journal author collaborations, and researchers constructed an author-centric network model of the journal. Many researchers use bibliometric studies to elucidate the productivity and contribution metrics of larger entities like countries or states to understand their position compared to each other in scientific contributions and production. A study of Odisha, India, is an excellent example of a scientometric profile of a single nations state that attempts to accurately understand its scientific output and community through bibliometric analysis uncovering among other insights the fact that nearly 40% of their top-cited works were the result of international collaboration (Garg and Kumar 2016). Many similar studies have utilized scientometrics to characterize nations and nation-states internally, as well as on the national and global stage, like computer science research in India, China, and other South Asian countries (Gupta, Kshitij, and Verma 2011; Kumar and Garg 2005; Uddin and Singh 2014), stem cell research in India and other countries (Karpagam et al. 2012), scientist evaluations in Brazil (Wainer and Vieira 2013), research production in Colombia (Bucheli et al. 2012).

In the broadest definition, cocitation is the utilization of citations to relate publication elements to determine relatedness and distance, whether that be through articles or journals. Cocitation is becoming more critical as the size and sheer quantity of publications are growing past the point that anyone can effectively review all publications coming out of a field, let alone multiple fields. For many researchers, bibliometric studies, including cocitation, now provide a concise way to have a grander overarching view of a field of science to guide research and understand field dynamics from a higher perspective. Alongside these studies, today's participating researchers are increasingly relying on information discovery systems like Google Scholar to stay abreast of their field's growth. Accordingly, researchers have noted that care needs to be taken in designing these systems and their expansion (Ding et al. 2014). Amidst all of these publications, researchers note that sub-groups of researchers working closely together within fields or topic-focused networks in larger fields are emerging, often called "invisible colleges" (Crane 1972). Cocitation leads to identifying these social structures in science (Teixeira and Ferreira 2013). These clusters are used to draw mappings of underlying scientific networks that can provide insights into the leading edges of research fields, ultimately providing precious information to policymakers and stakeholders. Small (1973) also brought to light the application of this clustering strategy explicitly with authors, now commonly referred to as author cocitation; it provides yet more insight into the social networks that underly scholarly publication production. Alongside other cocitation studies, author cocitation has been used in many fields to help ascertain their intellectual structure, as was done in 2002 in the research field of knowledge management, where the field of computer science was determined to be contributing less than what many thought at the time (Ponzi 2002). The coupling of author cocitation with other research methods has also yielded encouraging results like that done on information science by Zhao and Strotmann (2014), which used author bibliographic coupling analysis to examine other research fronts and trends. While these cocitation and citation-based metrics provide deep insights into academic structures, it does seem that in combination with other methods, research resolution can be tightened and help align findings with greater accuracy.

3.1.2 Scientometric Studies of Data Science

Data science is relatively new, especially compared to some far older fields of science, and as such, it has a much smaller collection of scientometric-focused publications to draw on. Amidst a smaller representation of scientometric or bibliometric studies, an examination of big data studies is also included here as, in many cases, the overlap is significant, and more recently, the idea that data science acts on big data with the significant interplay between the two concepts occurring has been outlined in the research.

In an examination of big data, Halevi and Moed (2012) identify the exponential growth curve of publications, primarily starting from 2008, highlighting that much of this work came from the computer science subject area, with the majority of publications originating in the

United States. Halevi and Moed (2012) go on to make note that the literature goes back to the 1970s and has been chiefly computer science dominated but is increasingly gaining contributions from subject areas in the earth, environmental sciences, arts and humanities, engineering, and health fields. Park and Leydesdorff (2013) examined big data under a semantic network lens and found that much of the international collaborative work examined classical techniques. However, they suggested a potentially emerging research area pointing to newer research focused on digitally-supported techniques. As research continued and big data and data science research continued to proliferate through the 2000s and 2010s, the lexigraphic lines of big data quickly became fuzzier. In 2015 Huang et al. (2015) systematically approached the need for highly constructed database queries needed for accurate bibliographic studies; researchers in this study examined "big data" and generated an accurate database query to harness appropriate publications. In this study, the appearance of "data science" as part of an expanded "big data" query construction (Huang et al., 2015) suggests new words and research linkage. A scientometric study by Singh et al. (2015) on big data in 2015 provides a comprehensive look at authorship, countries, universities, domain contributions, and control terms; while the authors mention data science, it does not appear in their results. What does appear, though, are the terms "data analysis" and "big data analysis" (Singh et al. 2015) alongside many other data science terms that seem to be shared. In fact, in 2016, a literature review on "big data" and "big data science" was published categorizing the research from 186 journals into 20 categories outlining the domain (Chen et al. 2016), and while the research did not explicitly reference "data science" alone, the use of the term "big data science" itself is telling to how closely the terms were being used. In a 2017 bibliometric study of big data, however, data science appears again as researchers' most relevant keywords for big data, carried out on Web of Science on 6,572 documents the authors took note to mention that gaps in big data's bibliometric analysis existed in the literature (Kalantari et al. 2017).

The connection between data science and big data is hinted at early in the literature; even in the early 2010s, the connections were being vocalized more and more, and it was becoming clear these two terms would come together. In 2013 an article in the journal *Big Data* by Provost and Fawcett (2013) tried to break down and analyze this connection as the rise of data science in the media, academia, and private industry. The authors clarify that data science and big data technologies will be critical systems in business analytics going forward. In the same year, another study examining the overlap of big data and data science in the field of supply chain management published makes it clear that the influence of big data and data science on future supply chain managers will be something that needs to be prepared for (Waller and Fawcett 2013). A 2018 scientometric study of both big data and data science utilized both terms together and continued to confirm early studies in highlighting the United States as the most significant contributing nation and computer science as the largest contributing domain; however, the range of contributing domains is markedly expanded in comparison to earlier studies (Papi 2018).

It was not until 2020 that a study by Raban and Gordan was carried out that examined the relationship between big data and data science and where and how they related to one another (Raban and Gordon 2020). The research revealed big data to be a far more fixed term with a much larger corpus of documents when compared to data science, even though publications on data science seem to have started earlier (Raban and Gordon 2020). Though, what is uniquely engaging is that those publications that utilize both terms had a higher likelihood of becoming "highly cited" in their findings (Raban and Gordon 2020). Additionally, the authors comment that big data has seemingly developed faster and that this may reveal that data science needs more extensive data to act on, so data science publications have followed the rise of big data publications. At the same time, data science may serve as a theoretical ‘toolbox’ for big data uses (Raban and Gordon 2020). The two ideas of big data and data science are closely related, and the interplay between the two fields is an active occurrence.

While it may appear that disentangling data science and big data is perhaps approaching impossible, researchers are taking the time to examine each individually. For many areas of study, the implications of data science in their fields are paramount to understanding the integration of large data sets and new technologies. A study of data science’s influence on policy analysis represents one such field grappling with data science's impact. Utilizing bibliometric methodologies, the researchers determined that data science was still emerging within the policy analysis research but showed promising trends toward artificial intelligence and econometrics (Y. Zhang et al. 2018). Interestingly, in a 2019 study utilizing Google Scholar to analyze scientists who showed interest in data science, researchers found that “machine learning” and “artificial intelligence” cooccurred with “data science” as interests (Emmert-Streib and Dehmer 2018). Ultimately, the researchers discovered 20 fields that found data science interesting, with “machine learning” being the most important, followed by high energy physics and bioinformatics (Emmert-Streib and Dehmer 2018). Researchers in a 2019 scientometric study of data science identified the leading researchers, universities, and nations highlighting the United States as the largest contributing nation and *ACM* as the most contributing journal (Prakash and Arumugam 2019). The authors also noted that the field is growing and that data science is a booming area with respect to big data research (Prakash and Arumugam 2019). These findings are further supported by another scientometric study done in 2019 by Sarkar and Pal, who delineates that the English language is the leading publication language, which correlates well to the United States being the leading contributor; they too see a growth trend in data science citations (Sarkar and Pal 2019).

While the character and dimensions of data science have been tangentially touched in several of these studies and many other studies focused on other scientific research, those that have focused more specifically on data science have been looking at the relationship with big data, data analysis, and the co-evolution and potential entanglement of these research paradigms. What does seem to be lacking is any scientometric study that handles data science in its own right. Unlike the scientometric studies on fields other than data science, the research field of data science has not provided the same perspective on the intellectual structure, and this absence is strange given the term has grown deeply into the academic and the public world.

3.2 Content Analysis of Curriculums

The ecosystem between academic research and educational teaching merges the scholarship of publications and research with classes and curriculums in a delicate balance. The interaction of research and education is further connected and complicated through the private sector, job markets, and private industry. Classically determination of the intellectual structure of a field has relied mainly on the interpretation of academics almost exclusively through their scholarly communication; the reliance on citations and focus on the nature of science is indicative of this. However, the ability to reveal how other aspects of society characterize intellectual structure has become more practical in recent years, and content analysis techniques have played a significant role in accomplishing this.

Content analysis is a well-established flexible research method utilized across scientific fields. Definitional opinions on the differing forms content analysis can take and whether it should be more qualitative vs. quantitative exist. Neuendorf (2020) acknowledges the breadth and range of applications that content analysis has exhibited over the years and presents an inclusive definition that characterizes content analysis from the top down:

> "Content analysis is a summarizing, quantitative analysis of messages that follows the standards of the scientific method (including attention to objectivity–intersubjectivity, a priori design, reliability, validity, generalizability, replicability, and hypothesis testing based on theory) and is not limited as to the types of variables that may be measured or the context in which the messages are created or presented (2020, pg 17)."

Content analysis of curriculums primarily focuses on educational curriculums and associated components. These components are often course titles, descriptions, syllabi, degree requirements, and even departmental course offerings and documentation. Callison and Tilley (2001) took a look at course titles, job announcements, and self-described teaching and research to examine the change over time in the education of library and information sciences (LIS) students. Latham (2002) utilized job advertisements for both teachers and students as well as course information to examine effective communication education in the LIS curriculum. In yet another similar study of data-specialist skills, Si et al. (2013) used content analysis to examine university job offerings and curriculum elements to uncover the state of data-specialist skills taught in LIS programs. Each of these studies exhibits curriculum analysis's ability to use job market information to determine the relationship between classroom teaching and industry need. These studies offer immense value to stakeholders in education and industry by crafting powerful examinations of topics and how they relate to one another across learning and practice.

These highly-targeted curriculum analysis studies often examine a specific aspect of a curriculum to garner a deeper understanding. Curriculum analysis has been used to target a host of specific topics like palliative care in nursing education by Martins Pereira and Hernandez-Marrero (2016); information literacy and library use were analyzed in the business curriculum (Boss and Drabinski 2014); offerings of data curation courses in LIS programs (Harris-Pierce and Liu 2012); and ethics education in psychology programs (Griffith, Domenech Rodríguez, and Anderson 2014). Each of these studies examines integrating a topic of study into the existing curriculum. The adoption and inclusion of topics across these studies look at how courses evolve through time and how curriculums merge these topics into practice, be it from the standpoint of a single program or a collection of institutions. Additionally, many of these studies attempt to triangulate consistency and attribute the value of topics, subjects, or courses by analyzing the presence throughout educational programming. Another such study was performed by White (2005), examining the curricular content of LIS programs regarding business information courses. The study collected course syllabi for each business-oriented course and used content analysis methodologies to examine the extent to which these courses were business-information courses. White's study is broader than examining a single topic but similarly implements curriculum analysis to examine how educational content is adapting to new, changing, and different learning needs through curricular change.

Usage of curriculum analysis runs the gamut; however, the most related to this research is its usage in determining intellectual structure vis-a-vis educational materials and content. The more extensive mapping of sciences and content areas through curriculum analysis is like those studies focused on specific topics using course titles, descriptions, syllabi, and program structure. However, these intellectual structure studies differ in their goal to illustrate a broader picture as a snapshot through change over time or to attempt to solidify the core structure of a field. Irwin (2002) examined the effects of computer studies components on LIS curriculums and the traditional course utilizing curriculum analysis, ultimately finding that computer studies' effects are far less impactful than previously thought. As a result, Irwin carefully mapped out 49 accredited LIS programs courses providing an insightful look into the LIS intellectual structure and its changes between 1994 and 2002. Introducing new topics, tools, job requirements, and other novel elements frequently foreshadow alterations in education programs. Like Irwin's 2002 study, Varvel, Bammerlin, and Palmer (2012) undertook a curriculum study looking at the changing curricular designs of LIS programs in response to more data-intense job requirements. In another study of LIS curricula, Chu (2006) extensively examined 2,757 courses from 45 programs offering the research community an in-depth examination of courses. Looking at past research, Chu also went a step further in analyzing LIS educations' shifting intellectual structure utilizing changing course names, descriptions, course requirements, elective courses, and overall educational structural changes to mark differences in design and intent. The result of Chu's research is similar to Irwin's research in echoing that many of the changes occurring result from shifts to meet the demands of shifting job landscapes and requirements.

Several curriculum analysis studies focusing specifically on data science, big data, and data analytics have been carried out as researchers have noted that data science has begun to establish courses, programs, and degrees in universities across the globe. Song and Zhu (2016) clarify in a publication on big data and data sciences' education that the determination of what to teach and how is still developing and will require a unique approach of theory and pragmatism in preparation for the workforce. Underscoring the rise of these data science-related programs, Han (2017) gives a brief state of data programs and some of these emerging educational offerings' essential characteristics. With that in mind, many curriculum analysis studies around data science actively consider job postings and workforce needs alongside the academic underpinnings of scholarship. Behpour, Hawamdeh, and Gourarzi (2019) do just that in their content analysis directly on job ads in an attempt to provide a data set to support the rapidly growing population of data science degrees and programs and assist them in more impactfully crafting their curriculums.

Heeding the call for a more in-depth examination of these quickly developing data science programs and considering the co-evolution of the workforce's job requirements, researchers have only just begun to organize mappings of the intellectual structure of data science programs through curriculum analysis. Tang and Sae-Lim (2016) reviewed a random 30 data science programs from 8 different disciplines through curriculum analysis to map and characterize the intellectual structure belying these data science educational programs and determine similarities and differences. The research represents a cross-section of questions necessary to understand the educational designs of current curriculums and comprehend the intellectual structure in a manner conducive to future curricular improvements and design. Tang and Sae-Lim's research sets an in-depth mapping of the 2016 state of data science education, especially regarding examining programs across disciplines.

Later in a more focused study looking directly at iSchools and data-related curricula, Ortiz-Repiso, Greenberg, and Calzada-Prado (2018) compared research across 65 institutions. The study analyzed the universities to determine if the schools were teaching data science, big data, or data curation in their programs and delved further into examining how even these sub-categories were being taught and differed. Studies of this nature continually help researchers and educators delineate the characteristics of some of these very closely related topics like big data and data science. Additionally, the deep curricular examination and comparison among universities help bring to light inconsistencies at the program and course levels, especially regarding topics or credit/degree design for future educational designers and faculty.

In another curricular-centric study that combined job offerings, Washington Durr (2020) compared iSchool curriculums and job postings to each other, utilizing seven iSchool curriculums and over 1600 job postings. The study utilized a text analytic approach to find intersections between educational offerings and job requirements. Ultimately finding differences amongst some key characteristics, the study concludes by determining a high degree of similarity between the curricula and job postings. It also suggests that further research is needed and that curriculum is, in fact, mapping to job ads. The addition of job ads in a number of these studies helps researchers understand the balance between the education and job markets and provides insights into what core elements the average educational program and degree may be comprised. Perhaps most importantly, these curriculum studies can be seen to provide stakeholders on all sides of the educational world, researchers, teachers, or students, with valuable information with which to inform future studies, course design, or simple program choices.

Investigating intellectual structure through scientometric research is well established and backed by countless research studies; this is not the case for curriculum analysis. Most curriculum studies look at curriculum elements pertaining to the job market, institutional adoption, or satisfying new emerging research topics in a domain. Scientometric research on domains, including curriculum analysis, is near non-existent, especially with the objective of understanding intellectual structure. Additionally, this research's need is extended even farther when the small number of curriculum studies looking specifically at data science is not expansive or geared towards intellectual structure on their own. The scarcity of all this research and applying curriculum analysis alongside traditional scientometric research methods reveals a gap in the scientific literature.

3.3 Scientometric Research Using Altmetric Data

The internet's explosive growth has led to some of the most extensive and mappable networks of human connection ever fashioned, and with its spectacular rise, the scope of measurable communications has expanded to incalculable amounts. It was only a matter of time before researchers recognized the sheer possibility and wealth of data generated first by websites themselves and then by blogs, microblogs, and social networking platforms. Social media sites and systems like Twitter, Reddit, Mendeley, Facebook, LinkedIn, and many more have led researchers to embrace a new breed of techniques for measuring scholarly exchange -- altmetrics.

The raw quantity of data that altmetric studies can harvest and the range of topics it can be applied to has resulted in its being used across disciplines and in conjunction with a tremendous cross-section of research questions, especially given its relative youth as a research methodology. The scope of altmetrics studies has a great deal of variety, allowing it to cover huge fields as well as going as narrow as a single topic within a domain. Kim et al. (2015) utilized altmetric text mining to over 7 million tweets to determine topic coverage and topic life span about Ebola virus posting and news. Park, Youn, and Park (2018) used Mendeley and Twitter to examine cross-national academic networks to identify major scholarly groupings operating across borders. In an attempt to increase accuracy in gathering tweets on Twitter relating to specific events, Zheng and Sun (2019) published a proposed system of tweet categorization to help organize tweets more effectively than just location opting for a system of relevance, coverage, and involvement. Research of this type highlights the magnitude of information coming through systems like Twitter and the potential ability to harness it; also, this research brings to light the need for tools to effectively use all of this data.

In addition to an event or specific topic research, the study of networks and communication is at the front of altmetric research, and the quantity of these studies vouches for the perceived research potential these social networks like Twitter may provide. In a unique study that straddles the digital and physical world, Lee et al. (2017) examined the evolution of a digital community represented by those tweeting at an annual conference over three years. The study of these digital communities is not isolated to Twitter either; in a bibliometric and altmetric study of Anatolia's scientific output (Mokhtari et al. 2020), researchers used Scopus and an aggregation company Altmetric LLP to gather a cross-section of altmetric reports. Today researchers looking to leverage multiple altmetric sources at once can turn to private companies like PlumX or altmetrics.com, to name two. Nonetheless, Twitter has become many scholars' favorite research platform, and utilizing Twitter to study organization-level understandings is also prevalent; Zhang, Sheu, and Zhang (2018) used collected data to analyze how five major LIS organizations were employing Twitter. In an effort to understand altmetric data, researchers have reached and examined systems like Twitter at all levels. The complexity of the different actors and the systems themselves have left a lot of discussion and the understanding that many aspects of the networks themselves change potential user implementation and interaction. Yu et al. (2019) took the challenge of characterizing Twitter users, specifically those who post scientific tweets and provided an example of research towards characterizing members of a select community and platform. In addition, to provide a snapshot of scientific tweeters, the researchers also highlight that these communities are highly dynamic and changing; even just the number of scientific Twitter users is rapidly growing.

Altmetric research has found great success in integrating previously established research techniques. Coword analysis is a research technique that supplements citation-based approaches, especially in examining research fields' intellectual structure. Coword analysis is much older than altmetrics and finds its origins outside altmetrics but has found new purchases in the rich data sets of altmetric studies. Coword analysis has roots starting in 1983 (Callon et al. 1983) and is designed to utilize the cooccurrence of words to characterize document content and link similar documents together (Janssens et al. 2006). Whittaker highlights three aspects of coword analysis that serve as the basis of its foundation: first, that terms authors use are judiciously chosen; second, words used together share a meaningful relationship; and third, words used in conjunction often across authors and publications demonstrate meaning within a field (Whittaker 1989). The linking and clustering of these documents allow for some unique inspection of publications for relevant research topics and themes, ultimately understanding the intellectual structures underlying publication corpora. The structures emerging from coword analysis also help provide an intellectual structure of research separated from citations' unique life cycle aspects and potential pitfalls of citation-based confluences that can affect citations themselves (Callon et al. 1983). These pitfalls can result in some strange effects on the perceived evolution of a field and lead to the "clustering" of citations under how articles and scholarly discourse are executed in some research communities (Callon et al. 1983). Similar to cocitation analysis, visualizations can be derived from these word cooccurrences and provide discerning concept mappings for research fields (Callon, Courtial, and Laville 1991). Researchers have found great value in using coword analysis, and it has been applied across fields to give indispensable insights; one such study looked at scientometrics itself between the years of 2005 and 2010, generating some powerful mappings that show thematic changes in research occurring over the period (Ravikumar, Agrahari, and Singh 2014).

Coword analysis has been used in many fast-moving technology fields. Additionally, in an international anticancer study, coword methods were coupled with cocitation to visualize metrics on production and trends in topics across the world (Xie 2015). Outside of medicine but still within the realm of emerging and growing technology fields, cocitation has been used to map the field of research around software engineering (Coulter 1998), renewable energy (Romo-Fernández, Guerrero-Bote, and Moya-Anegón 2013), and the Internet of Things (Yan, Lee, and Lee 2015) to name a few. These studies are just a handful that has solidified coword usage outside altmetrics. However, the application of coword as an altmetric technique is now emerging across academia, especially when analyzing the data-rich messages of altmetric data in search of intellectual structure understanding.

Studies mapping the underlying intellectual structure of a field or domain have always been influential in understanding such fields and domains more deeply. Altmetrics have provided another manner to approach this understanding, especially with the inclusion of a different, broader community via altmetric data. Moreover, mappings of scientific fields can go a long way in understanding where scientific communities have moved and shifted attention, and this, in turn, can even provide evidence to make a conjecture on future emerging trends. A great deal of attention for altmetrics has been its value as a research tool coupled with one or more methods and in the case of intellectual structure mappings, which often is a combination with bibliometric methods. Biljecki (2016) combined Scopus citation data with altmetric (https://www.altmetric.com) and Mendeley data to characterize the output of geographical information science journals, countries, top articles, and even collaborative efforts providing a structure of the publishing community. In another multi-method study, Bhattacharya and Singh (2020) collected data from the Dimensions database, which provides altmetric data, for an insightful look at citations alongside altmetric data for an early view of COVID-19 research and how the larger public was interacting with highly cited documents relating to the novel virus. In addition, the inclusion of Google trends data took a look at some of the most active topics and how Google search numbers represented an interest in the public. All this provides an additional aspect of research interaction thanks to altmetrics around a specific topic in a time of great public interest and concern. Such multi-method approaches to understanding underlying research networks, interests, and intellectual structure help anchor altmetric findings against other methods and sometimes provide exciting disparities.

Recently, a growing number of studies focused more heavily on the altmetric data and analyzed it mainly on its own, separating it a bit farther from other more traditional practices. While the access to different sets of altmetric data sometimes results in difficulty in comparing studies, the variation does yield interesting sets of data and results to cross-examine. Arroyo-Machado et al. (2020) utilized Wikipedia to map the humanities and create detailed mappings of those citations to understand the intellectual structure as seen through the site. Another example of a large-scale study was creating overlay maps derived from Mendeley readership data by Bornman and Haunschild (2016).

In addition to these large scoped intellectual structure studies, a range of smaller studies highlights the adaptability of altmetrics when it comes to the granularity of research. In a study on Austria politics on Twitter, Ausserhofer and Maireder (2013) pose several research questions ranging from the interactions between politicians and citizens to important topics appearing in tweets and even how those tweet-based topics relate to news outlet topics. Specific field-based altmetric research is also common. Examples include research on social media attention of microbiology based on Twitter data (Robinson-Garcia, Arroyo-Machado, and Torres-Salinas

2019) and the analysis of dental journals and articles through Twitter (Kolahi et al. 2019). Chae (2015) leveraged the common usage of hashtags in Twitter to direct a research focus on tweets containing #supplychain to determine how it was being used and find topics related to it to understand further the community of users participating in Twitter discussions.

Data science-focused altmetric studies are equally rare as they are in scientometric studies and content analysis studies. Similarly, many of the studies that appear are effectively attempting to apply data science techniques to large altmetric data sets, which, interestingly enough, is a perfect application of data science itself. However, a study on big data by Lyu and Costas (2020) examined big data as it appears through differing altmetric sources. They concluded that Twitter ultimately had a higher concordance of hashtag cooccurrences with author keyword selection amongst blogs, news, Wikipedia, and other social platforms.

Regardless of the variability in altmetric studies or the data streams or platforms they originate from, the value of more data to reveal different pockets of society both within the academic community and outside of it only gives more insight into the movement of information. Moreover, the value garnered from examining communities falling outside the classical academic community provides insights into how outside societal groups relate to scientific fields, topics, and even journals, authors, and articles.

Altmetric research has risen quickly and has expanded in application across science, pertaining to the domain of data science; however, it is still lacking. A tremendous amount of research in the altmetric space looks at the impact and influence of documents or authors through social media platforms. Other researchers have utilized hashtags to narrow research and capture the conversation on topics or events. However, not many studies attempt to decode the intellectual structure of a domain. This research aims to understand the intellectual structure of data science and apply altmetrics techniques to Twitter to understand leading topics and related topics, which are conspicuously absent from the literature. The uniqueness of this research is further compounded when examining data science by combining altmetrics with scientometrics.

3.4 Chapter Summary

This chapter has presented a cross-section of research highlighting the research power of studies utilizing scientometrics, curriculum analysis, and altmetrics. What is more, is the fact that there is little research currently characterizing the intellectual structure of data science. While this chapter has attempted to present these research methods as powerful on their own, the advantages of research that employs multiple approaches should be clear. This study seeks to harness the power found in the confluence of all three methods to provide researchers and other interested parties with a strongly backed mapping of the domain of data science through these three unique perspectives and examine the three's interplay.

4. Objectives and Scope

The objective of this study is to characterize and understand the scientometrics features and intellectual structure of data science. The literature review shows an apparent scarcity of work when it comes to examining data science from the scientometric standpoint using bibliometric data, curricular data, or altmetric data. The lack of literature only further deepens when investigating data science from an educational or broader social perspective with almost no curriculum or altmetric studies. Currently, no study exists that relates the three aspects together. Additionally, the lack of research on this proposed topic only seems to echo louder when data science appears to be an exploding topic everywhere: academic research, public news outlets, and private industry. Therefore, this study aims to understand data science from a scientometric, curricular, and altmetric perspective and comprehend its research fronts, develop a scientific profile, and visualize its intellectual structure.

4.1 Research Questions

Determining and understanding the intellectual structure of data science from scientometric, curricular, and altmetric perspectives is geared toward providing stakeholders with a clear understanding of the state of the data science field. Under that goal, this study has formulated the following research questions (RQ):

RQ1. What are the scientometrics features of the data science field?

RQ2. What are the contributing fields to the establishment of data science?

RQ3. What are the major research areas in the data science discipline?

RQ4. What are the salient topics taught in the data science curriculum?

RQ5. What topics appear in the Twitter-sphere regarding data science?

RQ6. What can be learned about data science from the bibliometric, curricular, and altmetric analyses?

The primary objective of understanding and characterizing data science has been to address these research questions. However, RQ1, RQ2, and RQ3 will be the questions that represent the core scientometric goals to be approached using bibliometric data. The analysis of publication and citation data to reveal the scientometric features will provide essential metrics for understanding the current state of data science literature. These metrics and cocitation analysis will lay the groundwork for RQ2 and RQ3. RQ2 seeks explicitly to help ascertain which fields have and are contributing the most to the field. This research question also intends to understand which fields data science originates from and what fields are currently contributing most to its future. Analysis and categorization of citation data will help to identify the contributing fields. Along with that, cocitation analysis will additionally provide a look at the underlying network of authors, the dynamics of scholarly communication, and the focus of research fronts. These findings will assist in determining the topics and research fronts to address RQ3.

The use of curriculum analysis to map a domain or field's intellectual structure is established and growing, and the extent of program development for data science has been extraordinary. Examining the curriculums and their materials to determine how these programs can reveal the intellectual structure of data science through RQ4 will provide a unique educational perspective on the development of the data science field. Understanding the topics and variety of focus areas for these curricula put forth by universities will contribute additional evidence in crafting a more comprehensive data science profile.

Twitter allows for a far broader examination of the conceptual relationships of data science in a broader social setting. RQ5 aims to take advantage of Twitter's less academically-focused population and affords a look at how tweets are being used to relate topics and issues when it comes to data science. Far less specific than traditional bibliometrics, this will also grant a much more global look at how society relates, promotes, and discusses data science. This viewpoint adds another piece to data science's characterization and relevant topics and domains as perceived in the Twittersphere.

Finally, the goal of RQ6 is to take these three unique perspectives (i.e., scientometrics, curriculums, and tweets) of data science derived from the first five research questions and examine them together. Understanding the relationship between these three unique perspectives and how they may affect each other can give stakeholders valuable insights. The confluence of these three different approaches may also help speak to a more unified core conceptual map of data science. Additionally, these different communication systems' perceived importance or focus areas may shed light on ambiguous definitions, debates, and even future research fronts in the field of data sciences.

4.2 Concepts, Variables, and Operational Definitions

Scientometric features of data science will refer to the measure and metrics associated with the scholarly literature of data science. These features specifically include prolific authors, top-cited authors, publication year, and author affiliations, to name a few.

The intellectual structure of data science is defined in this study as a scholarly network of authors and interconnected research areas. This structure consists of the interrelated connections that exist among cocited authors and publications, supplemented by cohashtags from Twitter and topics from data science curriculums. The final network of these correlated materials will be a more comprehensive map of the intellectual structure of data science.

The term curriculum throughout this study refers to the planned content of an education program. The curriculum may include comprehensive curricular documents, course descriptions, syllabi, and other components.

Twittersphere is defined in this study as all tweets and users available from the Twitter platform. Twitter hashtags within tweets are the target source data for the altmetric component of this study.

4.3 Research Scope

This study will collect bibliometric data from the Elsevier database, Scopus. All publications on data science from the available years will be gathered for analysis. Additionally, only data in English will be obtained utilizing a structured search query to limit results to those about data science.

Every effort will be made to collect curricula from higher education institutions with data science programs. However, accessibility, permission, and formatting of any curriculum documentation will unavoidably affect this research. There have been many researchers who have stated that there is a lack of standardization when it comes to curriculum layout, whether in the form of whole program design or course description. Additionally, only those curricula available digitally and in English will be included in this study. The study will also be limited to only universities within the US with dedicated data science programs.

Altmetric data will be collected from Twitter, a widely used social media platform. Two significant aspects of the Twitter system will attempt to control the amount of data and data collection times. This research will seek to limit tweets through filtering via specific hashtags.

Additionally, Tweet data collection will be scheduled to occur at intervals over three weeks, and every effort will be taken to limit tweets to include only those written in English.

5. Methodology

5.1 Research Methods Selection and Justification

This research was designed to answer six research questions utilizing three distinctly different research methods and ultimately weave them together. This sub-section of Chapter 5 presents the research methodology of the current study, along with the rationale for choosing it. The chosen methodology of this study consists of scientometrics, content analysis and altmetrics.

5.1.1 Scientometrics

Scientometrics has a long history of being utilized to help profile and uncover the underlying structure of science fields (Garfield 1979; Price 1965). This research was conducted to determine the scientometric features of data science, and as such, this study was well suited for the application of scientometrics. The understanding and mapping of these features by analyzing citations gathered from publications of data science are highly valuable to stakeholders across the field. Furthermore, scientometrics has been deemed "essential" for scientific communities to understand and harness the research, productivity, specialization, and networks of a field (Perron et al., 2016). Importantly, scientometric research provides context and shape to a field and brings light to potentially under-recognized topics and themes occurring within a field.

Scientometrics was the research method chosen for this study to depict the salient features of data science through analysis of publication and citation data gathered. Anson (2016) sees scientometrics as a means to understand and investigate the research landscape based on its output, volumes, and origins. This understanding of scientometrics echoes countless other researchers' viewpoints and speaks back to the value of scientometric research in the earliest days and the reason for its continued development over the decades.

Scientometrics relies on fundamental concepts of how science develops over time and represents a combination of applied quantitative analysis and theoretical understanding of scientific development. Scientometrics has been built on the idea that science is a social endeavor and that the self-organization of science can be revealed by understanding the communications via publications within science. This concept is truly the merging point of both the social constructs and intellectual organization of science as a social phenomenon. The treatment of science as a series of communications across a domain through documents is the crux of the scientometric process.

However, recognizing scientific documents or scholarly publications as vehicles for understanding a more significant body of science is clear. The value that citations to scholarly publications hold in comprehending links between documents and authors allow scientometrics a depth and insight in examining fields. Publications and citations are the primary linking factor in this social world of science; scientometrics can scale, group, and aggregate publications and citations in incredible numbers to find relationships that lead to understanding science by analyzing these complex groupings and hierarchies. In this way, grouping publications and citations at various levels provide another variable to view scientific development. This understanding can be coupled with an account of time and when documents were published to reveal science's growing and dynamic structure. The features derived from a scientific field through this scientometric study from a corpus of bibliometric data are extensive, in-depth, and exceptionally valuable.

Cocitation analysis is the key tool in finding links through grouping research publications and relatedness mappings. Scientific publications under the lens of scientometric analysis reveal intellectual structures by using cocitation data. These structures result from scientific endeavors and scholarly conversations between researchers, manifesting through their publications and the contained citations. Thus, researchers can "visualize" these links in the highly structured, quantitatively based scientometric methods and further their understanding of a field and its "structure".

Equally important, rates of publications on several grouping levels are also readily available for study and add another dimension of understanding. For example, scientometrics can determine critical documents, top authors, crucial institutions, and influential nations in a field of study. These levels provide useful data for describing the field and drawing the initial intellectual framework of its layout. The features and structure of a field are multidimensional and come out when viewing the aggregate connections between publications.

5.1.2 Content Analysis

As a field develops through scholarly work in academia, it also develops in education through academic institutions and the educational materials they create. Therefore, this research employed content analysis and, more specifically, content analysis techniques for studying curricular materials to leverage a deeper understanding of the field of data science. Furthermore, examining how a field may be structured by investigating its educational programs can provide a seldom-seen perspective on prominent topics.

Content analysis is the analysis of communications in various formats. In this instance, communications are text-based documents, with the explicit understanding that these words carry meaning. Content analysis can extend to both qualitative and quantitative realms of research, and it relies on statistical analysis to describe the word use and organization of those words within documents. A great deal of the appeal with basic content analysis is that it rigorously provides validity, reliability, and objectivity in its analysis (Drisko and Maschi 2015).

Content analysis operates from the premise that words have meaning directly and in their specific organization within communications; with this understanding, inferences can be made systematically. Content analysis can cover a tremendous range of communication types, including published articles, books, radio, and television; this research focuses on curricular texts. This content analysis employs descriptive statistics that characterize a set of documents allowing for the summation of content and categorizing a subset of words, phrases, and even paragraphs. This ability to reduce researchers' data into more manageable and usable data sets is exceptionally powerful, given larger data sets (Drisko and Maschi 2015). Researchers can look at this summated data and make inferences and conjectures upon the greater set of documents.

As a specified form of content analysis, curriculum analysis provides this research with a technique to approach the unique documents that education produces. The highly varied and specialized forms of educational curriculums, course titles, course descriptions, and publications require a method like content analysis to decode the underlying topics and bridge the gap between highly mutable formats, layouts, and designs inherent in education. Therefore, the ability of content analysis to digest such varied and information-laden documents into manageable and categorizable data is paramount to this research.

Educational materials provide critical insights into what concepts, topics, tools, and theories dominate the scientific field at hand and what educators have deemed worthy of focus, time, and resources for students to learn. The current research's primary data set for content analysis includes course titles and course descriptions in the curriculum. This study is designed to determine the salient topics of data science in education and delineate its educational structure by examining these course titles and course descriptions. These curriculums can provide a window into data science education and are a helpful representation of the more complex and lengthy education process in the classroom. Utilizing content analysis on curricula also helps reduce the volume of data while providing insights. To investigate data science education on a deeper level, examining materials across institutions provides a summative analysis of how the field is developing on a larger scale and what value the educational community collectively is putting on the field when it comes to student learning.

5.1.3 Altmetrics

The rapid expansion of social media use for scholarly discourse, and broader general scientific readers has resulted in researchers' ability to look at how science is being discussed inside and outside the purely academic walls of journals and peer-reviewed publications. For many researchers, altmetrics is a complementary method, often in conjunction with bibliometric analysis (Kolahi, Iranmanesh, and Khazaei 2017). However, researchers believe that altmetrics has had a transformative effect on the social study of research by providing a new perspective (Ortega 2015). This research utilizes altmetrics to examine data science from another perspective to address its structure by investigating topics appearing in the Twittersphere.

The diverse and massive data sets of social media fuel altmetric studies so much that very few other sources can reach similar quantities, let alone equal them. As a result, raw altmetric data is also often more accessible with application programming interfaces (API) for collection and analysis purposes at a cost and sometimes free of charge (Thelwall 2016). The availability of data through Twitter is exceptionally valuable for this research. Altmetric research often analyzes the communications from collected social media data sets through a host of techniques. Using the altmetric method in this study leverages Twitter's hashtags to understand which individual research topics are most popular on Twitter and the relationships between these topics based on hashtags' co-appearance in tweets.

Moreover, widespread social media creates massive data sets and facilitates a growing engagement and communication network. Sugimoto et al. (2017) see the rise of altmetrics derived from increased scholarly use of social media as more than just a fad. Altmetric data sets often span much farther than the academic community. This research on data science targets identifying topics within data sets from a broader community participating in Twitter discourse. Twitter offers a tremendous amount of altmetric research data as its community is vast. As such, the coverage of Twitter data is diverse and extensive. This research utilized Twitter data to compare the topics found regarding data science with those discovered via the scientometric method.

The research methodology for this study comprises scientometrics, curriculum analysis, and altmetrics. The triangulation of these three approaches provides individual views and reaches a holistic understanding of the core features of data science as a scientific discipline. The beneficial effects of multi-method research are well-manifested throughout science, and it offers several touchstone elements that contribute to the research regarding both the studied topic and the methods. Using qualitative and quantitative data, combining different methods can significantly facilitate the reliability and validity of research findings (Zavaraqi and Fadaie 2012). Seawright (2016) identifies that while multi-methods should, in theory, afford advantages over single method research, that is not always a foregone conclusion. Like this study, research needs to be designed and constructed systematically to leverage the additional tests that a multi-method approach delivers.

One research question formulated by this study is to determine what can be learned about data science from different perspectives obtained via the three distinct research methods. The goal of understanding data science through different lenses: academic, educational, and a broader social media perspective, is to find both commonalities and differences. In many respects, this investigation may shed light on what core elements of data science are translated across communities and which are not. The design of this research's methodology has enabled the current researcher to address the six research. From these individual analyses, this research addresses whether data science has a single common core across all three dominions or if the perspectives are too disparate to formulate a single-core vision.

5.2 Data Sources

Each of the three unique research methods outlined above relies on unique data sources, respectively. The current section depicts each of the data sources utilized in this research.

5.2.1 Scientometric Data

This research utilizes data from the Elsevier Scopus database for the scientometric portion of this study. Scopus is one of the two premier databases for scientometric studies, the other being Web of Science. Initially launching in 2004, Scopus boasts over 77.8 million records in January 2020 and claims to have the most comprehensive overview of fields ranging from science and technology to arts and humanities (Elsevier 2020). Moreover, each year Scopus grows and attempts to overcome coverage weaknesses. Furthermore, Scopus makes available standard website search access and API access. Finally, Scopus provides data in various formats, including Bibtex, Excel, and JSON, thus allowing for the use of many analytical tools.

Within its exports, Scopus offers a tremendous amount of document information. Five categories of document information are available for users to customize their data export: citation information, bibliographic information, abstract and keywords, funding details, and other information. These five categories each have their own set of specific data to be collected, including but not limited to: authors, document title, publication year, source title, affiliations, publisher, abstract, author keywords, index keywords, sponsor, conference information, and associated citations. All categories of data can be downloaded. This wealth of data ultimately allows for the scientometric analysis in this research.

5.2.2 Curricular Data

Curricula data were primarily collected from data science programs in colleges and universities in the United States in the form of course titles and course descriptions. The website datascience.community[1] has a comprehensive list of data science programs at higher education institutions across degree levels and private boot camps; however, this research only collected programs at colleges and universities. The datascience.community website is geared towards lending resources and learning options to students and opportunities for those who want to find jobs in data science.

The website datascience.community lists and aggregates information about data science-related programs such as institution name, degree, country, state, location, and department. Curriculum data for this study was then collected from each institution's website, although such data from individual data science programs was not uniform across institutions. While the general form of curriculum, course titles, and course descriptions is established across academia, each document's content is typically varied depending on the professor, department, or institution of a program.

5.2.3 Altmetric Data

Twitter was the altmetric data source for this study. Thelwall et al. (2013) concluded that Twitter primarily stands above the others in terms of coverage in a broader look among ten other

[1]. http://datascience.community now redirects to its alias site: http://ryanswanstrom.com/ where users are required to navigate to the "Colleges" link at the top of the site. At this point the url is http://ryanswanstrom.com/colleges/ and allows for the searching of data science programs.

social web services. Coverage advantages and the relative ease of gathering Twitter data are the primary reasons why this study uses Twitter as the data source for its altmetric part. Twitter was founded in 2006 as a microblogging system; it has evolved into a global communication platform.

Single individual messages on Twitter, known as tweets, serve as the base unit of altmetric data in this study. Previously 140 characters per tweet, the text of a tweet can now be up to 280 characters. In addition, Twitter has a self-organizing system of topics utilizing the hashtag ("#"), allowing users to group their messages into larger threads of posts. Similarly, tweets can be tagged to a user ("@"), again providing loose connections somewhat analogous to scholarly publications' citations but not as significant. While lacking the associated value that citations have, they represent linkages between users and tweets.

The number of tweets that can be harnessed is a massive boon to this research. Twitter's API allows for the filtering of specific aspects of the tweet. These filtrations can be done based on user, time, language, hashtag, and many more tweet characteristics. In this research, the focus was on hashtags and the cooccurrence of hashtags in relation to one another, relying on the grouping feature of the hashtagging to "categorize" a tweet and connect it with a broader discourse community on Twitter. This hashtag system implements one of the effective means by which Twitter facilitates topical discourses across the globe. The composition of individual tweets collected through the Twitter developer API presents a clear and concise way to gather these hashtags for analysis alongside each corresponding tweet's metadata (e.g., author, language, creation date).

5.3 Data Collection

The following section describes how data was collected using the three research methods selected for this study. In addition, APIs and other tools were used throughout the data collection process.

5.3.1 Scientometric Data Collection

Data collection for the scientometric aspect of this research was conducted via Scopus. Table 5.1 depicts the collection process, which includes two stages: search and download, aggregation, and cleaning. While scientometric data is much cleaner than altmetric data, these stages were conducted similarly to ensure the data had no unforeseen issues.

Table 5.1 Scientometric data collection process

Process	**Task**
Search and Download	● Query Scopus database for "data science" as a phrase ● Set language filter to English only ● Download results by year in the BibTex format, including all the available data fields from all five categories
Aggregation and Cleansing	● Check each downloaded file to make sure that it contains the number of records as indicated in Scopus ● Use a simple text editor to conjoin all the downloaded files into a single file, covering all years of the present study (i.e., 1983-2021) ● Check the merged file to ensure that all the records are correctly joined from each downloaded file ● Check for record duplication and remove if present

Scopus provides several options for exporting data but limits users to export up to 2000 records each time. Because of this limit, all the Scopus data were collected in blocks of 2000 records organized by year and in BibTex file format. The BibTex file format was chosen for its functionality with many data analytic tools and provides a structured system for organizing collected data.

The Scopus data search and downloading were performed on January 13th, 2021. The separately exported files were merged into a single BibTex file of 8,458 records. R-Studio and Bibliometrix, the software for statistical and bibliometric analyses, were utilized to remove duplicates and other abnormalities to ensure that all the records in the merged file are clean and ready for data analysis. R-Studio, an essential tool used throughout this study, is a free, open-source development environment to implement the R scripting language for data analysis. Bibliometrix is a library of R scripts explicitly developed for bibliometric analysis.

5.3.2 Curricular Data Collection

This researcher completed curricular data collection through a systematic approach. The data has unique features, primarily the unstructured and non-uniform nature of institutional websites hosting curricular data. While some major organizational aspects of colleges and universities attempt to structure their respective websites, this is not by any means consistent. Therefore, a plan was constructed to help control the incongruities across institutions (see Table 5.2).

Table 5.2 Curricular data collection process

Process	Task
Document Collection	● Query datascience.community for programs relating to "data science" ● Extract search results and limit them to institutions in the United States ● Utilize each link provided in the results, visit each website to collect course titles and course descriptions, and place them into an Excel spreadsheet o Priority was given to data found directly on pages describing the program (generally the department or degree-specific pages) o If no data was found, two other sources were consulted wherever available ▪ Course Catalogue database ▪ Course Catalogue publication (pdf)
Cleansing	● Data spelling check for all data ● Duplication checking ● Categorical data, including elective vs. required courses and program type, checked for consistency

The datacience.community site offered a single repository of programs ranging from certificate-level education programs to doctoral-level offerings. In addition, the site provided a range of information highly pertinent to this research, including a direct link to the program page, the institution offering the program, the degree type, and the institution's location. The data collection began by scraping all the data science programs located within the United States. This listing resulted in a total of 150 programs, which was further pared down to exclude doctoral programs as they tended to have far less structured courses and included many "research" and "writing" oriented classes. Those courses appear vague as they depend heavily on individual students' research orientation. As a result, those doctoral course titles and descriptions offer little in helping conduct this research in the realm of data science.

Each of the chosen institutions' websites was then visited utilizing the link from the datascience.community site. At this point, links that were broken and provided no redirect were searched on Google to see if the programs still existed. If the program had simply moved locations within the institution, data was still gathered; however, the program was skipped for those that the additional searching provided no results. Some programs listed at datacience.community are gone and cannot be found.

If the provided link from datacience.community did work, data was collected into a spreadsheet by collecting data first from the program's webpage. Most programs' webpages were often organized within their departments' websites. Searching was conducted through the institution's course catalog system if course titles and descriptions were not provided on the department or program page. In a few cases, certificate and degree programs were only located through the course catalog; if that was the case, only the current year was utilized for collecting data. If this catalog search yielded no results or outline of a degree program, then that program was removed as well from the curricular data collection list.

The majority of links were obtained from datascience.community worked. Programs, course titles, and description information were found. Since each website was designed differently, data collection required manually copying and pasting the course title and description information into the data set held in an Excel spreadsheet. Courses were included when they appeared explicitly in the program description, regardless of if they were required or elective courses. If the program structure suggested, for example, a "Level 200 MATH course", this was not included in the data set even if it is a required course. This issue of core requirement courses consisting of a wide range of classes like the above example was primarily encountered in bachelor's programs.

The collection of curriculum data started on January 19th, 2021, and concluded on Feb 19th, 2021, with 3128 courses collected from 125 programs. The data set was spell-checked, and the vernacular, abbreviations, and duplications were removed. However, in some rare cases, duplications were kept in situations where a single institution may provide two degrees, and each one had the same course title in its program. The cleaning process was done within Microsoft Excel, where data was collected, stored, and assembled.

5.3.3 Altmetric Data Collection

Altmetric data gathering was accomplished through a process summarized in Table 5.3 utilizing the Python API provided directly by Twitter in conjunction with the Tweepy Python package.

Table5.3 Altmetric data collection process

Process	**Task**
Document Collection	● Data collection run on 2-hour scheduled blocks ● Datastream filtered to collect only tweets containing #datascience ● Data further filtered to generate CSV-formatted files containing: o Tweet author o Tweet date o Tweet message o Retweet author (if applicable) o Retweet date (if applicable) o Retweet message (if applicable)
Aggregation and Cleansing	● Individual collected data files merged into a single CSV file ● Due to the Twitter API design, Tweet messages and retweet messages, if applicable, were merged into a single full-text message data point for all messages. ● All links removed from full messages ● Hashtags were parsed from each message by filtering each full-text message for words starting with “#”

The Tweepy Python package is an open-sourced Python library designed to efficiently implement the Twitter API protocols. Tweepy is the primary code package used for the collection of tweets in this study.

According to Twitter’s data collection policies, the easiest data set collection would be through their live streaming API. The stream was designed only to collect those tweets with the hashtag: #datascience. As these tweets came in, data about the date, author, and message were recorded. In addition, data was collected on the original tweet if the tweet was a retweet. In the cases of a retweet, data about the original author, tweet date, and message that the retweet referenced were also saved.

Tweet data collection was conducted in two-hour intervals that started on a randomly chosen date, January 20th, 2021, at 10 pm and shifted ahead by two hours each subsequent day, as shown in Table 5.4.

Table 5.4 Twitter data collection schedule

Date	Start Time	End Time
1/20	10:00pm	12:00am
1/21	12:00pm	2:00am
1/22	2:00am	4:00am
1/23	4:00am	6:00am
1/24	6:00am	8:00am
1/25	8:00am	10:00am
1/26	10:00am	12:00pm
1/27	12:00pm	2:00pm
1/28	2:00pm	4:00pm
1/29	4:00pm	6:00pm
1/30	6:00pm	8:00pm
1/31	8:00pm	10:00pm
2/1	10:00pm	12:00am
2/2	12:00am	2:00am
2/3	2:00am	4:00am
2/4	4:00am	6:00am
2/5	6:00am	8:00am
2/6	8:00am	10:00am
2/7	10:00am	12:00pm
~~2/8~~	~~12:00pm~~	~~2:00pm~~
2/9	2:00pm	4:00pm

Unfortunately, on one day, February 8th, 2021, the recording scheduled to happen at noon lost connection to the stream, so no data was collected before this researcher was aware of its connection failure. February 9th thus was added to replace February 8th, the missed date of data collection, while all other collection sessions were completed uninterrupted. At this point, the data was backed up, and an additional column was created to indicate if a tweet was a retweet in the merged file.

The collected tweets cover 20 days for a total of 40 hours. The data set in its entirety is comprised of 41,838 tweets by 5,281 users. Moreover, there were 2,212 retweeted users collected in this data set. These users represented the community of contributors this research examined.

The usage of Twitter is by no means uniform from day to day or from hour to hour. Figure 6.14 shows the number of tweets collected each day in the 20-day period. The daily difference ranges between just over 1500 tweets and approaching 3,000 tweets in a two-hour collection period. The absence of data for the 8th is the result of a failed collection stream

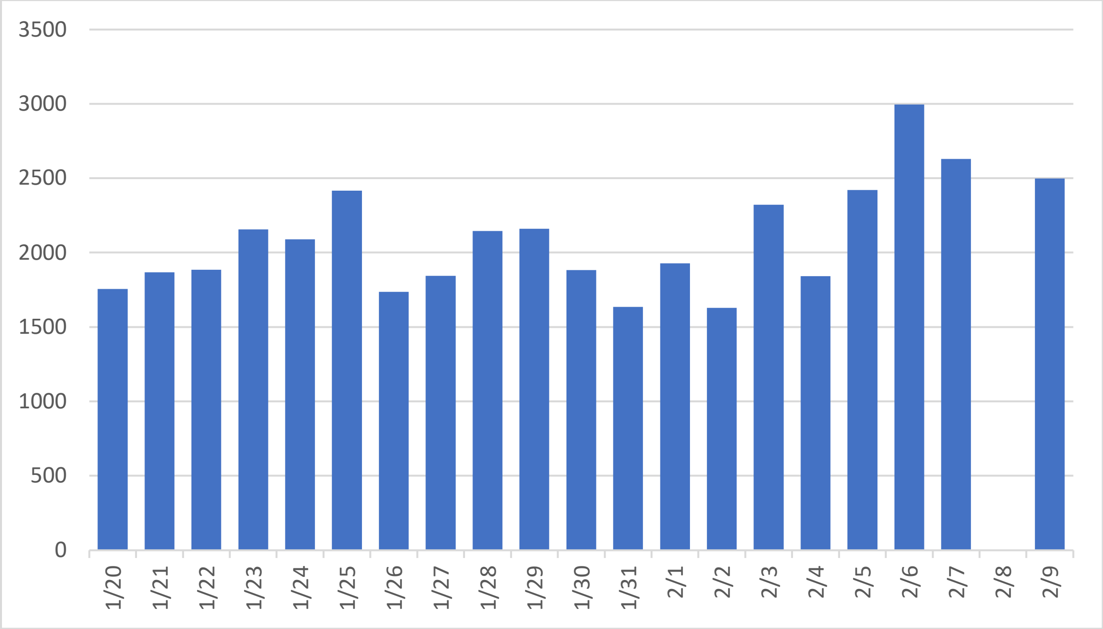


Figure 5.1 Daily tweet collection counts

Since the data collection was performed over three weeks, a quick examination reveals that many Twitter users post more frequently on the weekend (see Figure 5.2), while Monday appears to be the day with the lowest postings.

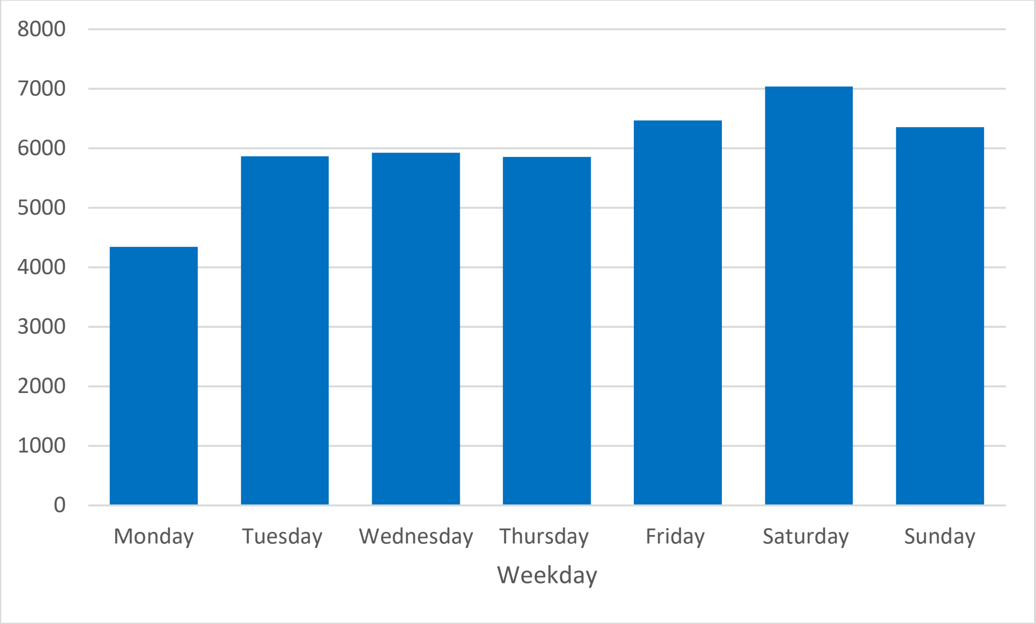


Figure 5.2 Tweet counts by weekday

Last but perhaps the most pertinent aspect of the Twitter data collected is that this data set contains 532,288 hashtags, and 7,505 of them are unique ones. These hashtags and their cooccurrences were later analyzed to address the fifth research question of this study.

5.4 Data Analysis

This section presents analyses of the data collected in the current study aligned to each related research question and in the order of scientometric analysis, content analysis, and altmetrics.

5.4.1 Scientometric Analysis

The scientometric analysis was conducted to address the first three research questions of this study:

1) What are the scientometric features of the data science field?
2) What are the contributing fields to the establishment of data science?
3) What are the major research areas of the data science discipline?

Descriptive statistics were performed on the collected bibliographic records utilizing the Bibliometrix's Biblioshiny app to validate complete and readable records merged and collected from Scopus. The Bibliometrix library in R Studio provides a powerful tool in its Biblioshiny app implementation that offers an efficient analysis of core parameters of records (Aria and Cuccurullo 2017). The data set from Scopus was analyzed to identify the top authors, top author affiliations, top-cited publications, top publishing countries, and other related parameters to address Research Question 1 regarding the scientometric features of data science. The number of common Asian last names was notable, complicating disambiguation from the data set provided. It is a significant challenge to ensure that authors are not being separated or merged. A tremendous amount of research has gone into addressing this issue (Kim, Jeong, and Song 2016; Liu et al. 2013; Müller, Reitz, and Roy 2017). Researchers have even begun applying data science methodologies to help combat these issues, like machine learning to assist in disambiguation techniques (Kim, Kim, and Owen-Smith 2021).

Each of the top 100 source publications in the data science field identified from the data set was analyzed and categorized into a scientific discipline (e.g., computer science, statistics). This analysis helped to find out which disciplines contributed to the creation and development of data science for addressing the second research question of this study. These disciplines were named based on established practice and were limited to twelve categories.

Cocitation analysis was also performed to discover the intellectual structure of data science based on the data set from Scopus. Scientometrics is often concerned with understanding and revealing the underlying intellectual structure of a field (Leydesorff 2015). Cocitation analysis is the premier means to accomplish this task. Cocitation analysis focuses on cooccurrence frequencies among cocited authors or documents in the data set to reveal the subject relationship of a field(Leydesdorff and Milojević 2015). Cocitation data in this study were analyzed using factor analysis, cluster analysis, and multidimensional scaling (MDS)from the Bibliometrix library with R-Studio. The three multivariate analytic techniques were used in conjunction to provide distinctly different perspectives on the cocitation data gathered for this study.

Factor analysis helps identify underlying variables or research topics in data science based on their cocitation frequency. On the other hand, cluster analysis facilitates the aggregation of individual cocited documents to form clusters that represent a group of related documents under analysis. It complements the factor analysis technique in identifying each cluster's members (i.e., cocited documents). MDS primarily assists in visualizing and mapping the relationships among the cocited documents onto a two or three-dimensional plane (Zhao and Strotmann 2015). These three multivariate analytic techniques are often adopted in combination to illustrate the underlying intellectual structure or identify research areas of a discipline, which constitutes the third research question the current study attempts to address.

5.4.2 Curricular Analysis

Curricular data collected for this study was analyzed using a combination of software to address the fourth research question: What are the salient topics taught in the data science curriculum?

The NumPy and Pandas packages in Python allow for easier manipulation of data sets programmatically with a pythonic data frame design. This initial statistical analysis laid the groundwork for understanding the basic metrics of the curricular data and yielded other results like the number of courses, number of programs, courses by program, and courses by requirement.

Further analysis was performed via WordStat 8, a program for qualitative data analysis, which enabled the current researcher to navigate and group the curricular data by parameters. Content analysis was also conducted on the qualitative part of the curricular data (i.e., course titles and descriptions). Content analysis allows a robust interpretation of course titles and descriptions by categorizing courses into the same group, even if they have different titles and descriptions. For example, although some courses are titled differently as “Introduction to Data Science” or “Data Science I”, they are all introductory courses in their curriculums. This value of content analysis is demonstrated in the grouping of introductory courses in data science and how they are described differently across various institutions. In addition, the purpose of content analysis of curricular data was not just to group courses on data science. Instead, this study intends to uncover topics and themes underlying those course titles and descriptions.

In addition, Wordstat 8 was used to identify the top 100 keywords by frequency in the curricular data set. The factor analysis procedure in Wordstat 8 was employed to group those keywords based on association strength via their cooccurrence frequencies. The cluster analysis procedure of Wordstat 8 was also adopted for analyzing the same curricular data set to visualize the groups formed in the factor analysis in the dendrogram. MDS was performed in conjunction to generate a two-dimensional map of the top 100 keywords, displaying the groupings from the cluster analysis with a color-coded presentation.

5.4.3 Altmetric Analysis

The altimetric data set was analyzed to answer the fifth research question: What topics appear in the Twitter space regarding data science? Python was primarily used to analyze and visualize the altmetric data. The analysis examined the basic elements of the data set, such as the number of tweets, number of hashtags used, number of authors, and other parameters.

Python was also used to separate hashtags from the collected tweets. The stripped hashtags from the tweets were then analyzed with WordStat 8 to examine the cooccurrence of hashtags across the data set. This process was handled similarly to the curricular analysis. A set of the 100 most frequent tweeted hashtags were further analyzed and visualized.

Additionally, from these top 100 hashtags, WordStat's topic analysis, which is based on factor analysis, was performed. Cluster analysis was also performed to help group the tweets into meaningful groupings. Together with mapping based on MDS analysis, these groupings were visualized similarly to the curricular data. These groups and mapping were used in revealing the subject relationships among the hashtags, ultimately discovering the significant topics of discussion occurring on Twitter.

5.4.4 Comparative Analysis of Results from the Three Data Sets

The sixth and last research question of this study seeks to determine what can be learned from these three unique research approaches. The topics that emerged from the analysis of each section were compiled, and the three sets of topics were then compared and integrated to form a holistic answer to the last research question of this study.

From there, three resulting visualizations were presented to identify those topics and to find out if they appeared across all three, in two of the three, or only in one of the three analyses performed in this study. This exploration of topics was then used to address the final research question.

6. Results and Discussion

This chapter presents and discusses the findings of the current research in four parts based on the scientometric, curricular, and altmetric data collected individually as well as in combination. Results obtained for each part of this study are then compared to address the six research questions of this study.

6.1 The Scientometric Perspective of Data Science

This section of the chapter is centered on the scientometric portion of the present study. The scientometric analysis has been divided into three parts, focusing on a unique research question. The first part examines the scientometric features of the data science field (RQ1). The second briefly examines the research fields that contribute to the field of data science (RQ2). The third one is devoted to determining the salient topics of research in data science (RQ3).

6.1.1 Scientometric Features

Key scientometric features of the data set collected for this study, including countries, sources, institutions, authors, documents, and keywords, are to be presented as follows.

6.1.1.1 Publication Output and Top Contributing Countries

The world of data science publications in academia exists in a rapidly expanding environment; therefore, this research is best presented with an understanding of the major players and the state of data science as a developing and growing field. Data science's growth as a domain has been rapid over the past decade, exemplified by Figure 6.1, where the annual production of documents relating to data science is visualized.

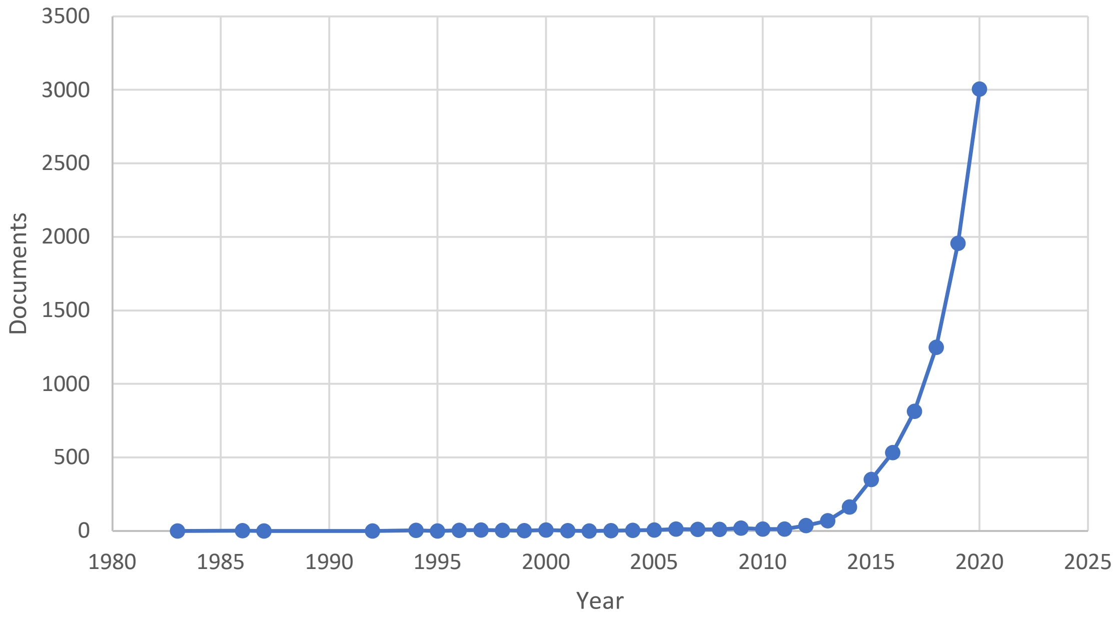


Figure 6.1 Distribution of publications in data science by year

It is essential to note that the bibliometric data for this study was collected via Scopus and contains documents from 1983 to 2021. Figure 6.1 shows the increasing rate of production starting around 2012. Upon examination of the year 2012, the rate of publications begins to increase. By 2015, the growth is noteworthy as publications begin to rise exponentially.

When it comes to data science's development across the academic globe, the United States has been a leader in terms of scholarly output between 1983 and 2015, with a massive 72% of the publication total out of the top five contributing countries being attributed to the United States (see the left half of Table 6.1). This global lead of the United States is in line with many other technology-focused disciplines (e.g., computer science). However, the right half of Table 6.1 illustrates a change in the years between 2016 and 2020, as the scientific production of the United States has markedly decreased concerning the other top data science producing countries.

Table 6.1 Top 5 countries data science publications production comparison for 1983-2015 and 1983-2021

1983-2015		**1983-2021**	
United States	72%	United States	59%
China	12%	China	12%
United Kingdom	8%	United Kingdom	12%
Germany	4%	Germany	9%
Japan	4%	India	8%

This re-distribution of publication percentages indicates that more countries are now playing a role in data science's development. This change in contributions can in part be attributed to the continued computer revolution occurring across the globe with expanding infrastructures like high-speed internet and computer access. In addition, more countries are participating in data science research, resulting in the declining control of the United States in the same area. On the other hand, China's retention of their 12% is also interesting as it means they are continuing to contribute an increasing amount of data science work in proportion to the rest of the globe. The United Kingdom and Germany both exhibited an increase in contribution with growths of 4% and 5%, respectively. The case of India and Japan is a bit different, with India replacing Japan by 2021 as a top-five producer. This change is almost certainly in part by the increased availability of technology in India and the influx of tech companies looking at India as one of the largest markets to join the world economy in the coming decades. Coupled with the expansion and increased internet access in growing countries like India, the value of generated data is being recognized.

As more data is generated by populations around the globe and with these countries seeing the value, there has been an increase in emphasizing procurement of data, efficient storage, and ultimately analysis with data science techniques. Each one of these tasks is now steeped in data science and the techniques derived from its research. The expectation is that an even more comprehensive range of countries will participate in future research, and those countries that continue to invest further in access to technology like India, will continue to generate a higher percentage of the world's data science research.

6.1.1.2 Top Contributing Institutions

Understanding and knowing the leading prominent institutions and organizations associated with data science is helpful to current academics, especially for prospective students of the field or those looking to hire. Their contributions to data science highlight the critical role that institutions play in academic progress. These contributions take the role of fostering environments of research through hiring, funding, and organizational support. Figure 6.2 is a chart highlighting the leading institutions of data science researchers.

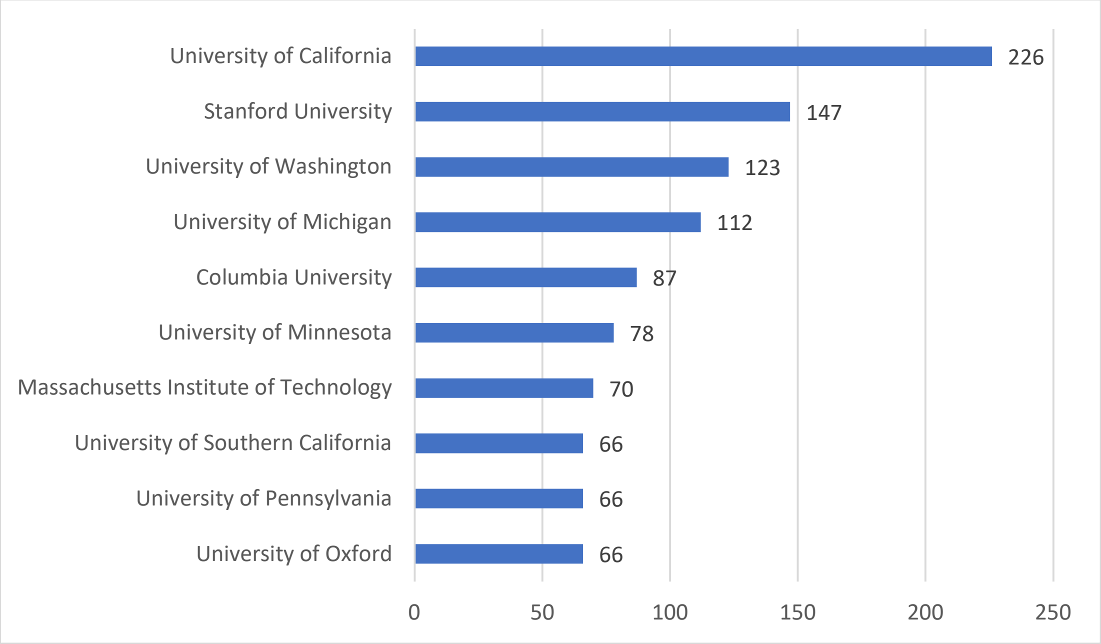


Figure 6.2 Top 10 institutions by publications on Data Science

Unsurprisingly and congruent to the findings of the top contributing countries, the data shows that United States-based institutions dominate the list of the top 10 institutions, as nine of the top 10 institutions are universities located in the United States. The only remaining institution, the University of Oxford, represents the United Kingdom, another leading country that contributes to research in data science (see Table 6.1). These findings again support that the United States is the leading contributing country in data science research output despite some decrease after 2015.

Interestingly, the top three universities, the University of California[2], Stanford University, and the University of Washington, are located on the West Coast of the United States. The common location of these three institutions is also not surprising, as the West Coast of the United States is home to some of the biggest tech companies dealing with the biggest data sets in human history. As a home base to these global tech companies, California and Silicon Valley boasts residences like Google, Apple, HP, Facebook, Netflix, Adobe, eBay, and Cisco Systems. Each of these companies is a leading tech giant around the world. Additionally, all three universities boast strong computer science and associated departments; all such disciplines have played a vital role in developing data science. The other seven institutions also have a highly technical background and have been homes to solid statistics and computer science departments. These departments conduct and contribute to data science research, as prior studies (Baumer 2015; Kim 2017; Zhang et al. 2018; Zhu and Xiong 2015) have shown the intensely close relationship of statistics and computer science to data science.

6.1.1.3 Top Contributing Authors

At a more microscopic level, the top contributing authors of a field can be highly influential in a discipline's structure, focus, and composition. The key authors of a field can frequently produce a significant amount of publications. Data science is similar in this regard

[2] Due to a change in subscriptions by the author's affiliated university, access to Scopus data for further analysis was cancelled. So the University of California, although consisting of many well-known individual campuses, is presented as one institution.

because the top twenty-five authors give a cursory cross-section of research in the domain of data science. Table 6.2 represents the top 25 authors based on their publication numbers.

Table 6.2 Top 25 Authors by publications

Author	Affiliation	Article #
JS Salt	Syracuse University, United States	22
Jörn Lötsch	Goethe-Universität Frankfurt am Main, Germany	21
Simon Elias Bibri	Norwegian University of Science and Technology, Norway	14
Carson K. Leung	University of Manitoba, Canada	14
Feras A. Batarseh	Virginia Tech, United States	9
Frank Emmert-Streib	Tampere University, Finland	9
Yuri Demchenko	University of Amsterdam, Netherlands	8
Ricardo-Adán Salas-Rueda	Instituto de Ciencias Aplicadas y Tecnología, Universidad Nacional Autónoma de México, Mexico	8
Austin Cory Bart	Virginia Tech, United States	7
LM. Chen	University of the District of Columbia, United States	7
Marco Spruit	Utrecht University, Netherlands	7
K. Takahashi	National Institute for Materials Science, Japan	7
Tomasz Wiktorski	University of Stavanger, Norway	7
Ben Williamson	University of Edinburgh, United Kingdom	7
Wil van der Aalst	Aachen University, Germany	6
Longbing Cao	University of Technology Sydney, Australia	6
B. Chen	University of Central Arkansas, United States	6
Matthias Dehmer	Swiss Distance University of Applied Sciences, Switzerland	6
Connie White Delaney	University of Minnesota, United States	6
Karina. Gibert	Intelligent Data Science and Artificial Intelligence Research Center, Spain	6
Michael Hahn	University of Stuttgart, Germany	6
S.R. Kalidindi	Georgia Institute of Technology, United States	6
Mary Beth Kery	Carnegie Mellon University, United States	6
Luca Pappalardo	University of Pisa, Italy	6
F. Piccialli	University of Naples Federico II, Italy	6
Iqbal H. Sarker	Swinburne University of Technology, Australia	6

The top 25 authors in Table 6.2 represent thirteen different countries. The United States accounts for eight authors (32%), with three (12%) from Germany as the second-highest contributing country. This distribution further supports the finding that the United States is the leading country in data science research. The economic classification of these countries is worth noting because the technological demands of data science are relatively high. The need for infrastructure to leverage the data generated by its populace is vital. So too is the development at the academic level, both in terms of education and institutional development. These factors are critical to providing the environment for students to learn data science or even be presented with opportunities in advanced technologies. Countries that lack these systems do not foster growth or implementation of data science. In the United States, these necessities are met so that businesses can realize the value offered by data science. This data-driven value helps reinforce the education, employment, and research of data science.

The top producing author in Table 6.2, J.S. Sal, focuses his research on frameworks and the organization of big data projects. His work experience is based in business and finance, which has led him to research how well-structured projects can help projects succeed. Sal's research is having a notable impact as more and more industries adopt techniques to use their data. While Sal's experience is based in private industry, his work undoubtedly extends into academics and governments.

Jörn Lötsch, the second most productive author, listed in Table 6.2, looks to tie data science, pain, and clinical pharmacology together from a medically influenced pathway. His research represents one of the fields generating some of the largest data sets, biology. It is important to note that data science is being widely applied to understanding biology in cellular functions and medical applications at many levels. Lötsch's work examines biometrics, pharmacometrics (drug effects), genomics, and next-generation gene sequencing. Handling the scope and size of data that can be extricated from medical studies with today's techniques and technologies can be sizeable and challenging for researchers. This complexity can be seen in the study of intricate interactions like proteins and genetics; however, this line of research sees some of the most exciting applications of data science.

Tied for third as the top researchers in Table 6.2, Simon Elias Bibri and Carson K. Leung look at vastly different topics. Bibri's work is around smart cities and is a great example of how data science, while often applied in fields like medicine and biology, can also be applied in other domains. The expansion of the IoT market and producers has led to exciting new data-gathering opportunities. Bibri's work on sustainability and data-driven urbanism offer an exciting glimpse at how the world humanity physically develops may be shaped by data science techniques.

On the other hand, Carson K. Leung's research focuses on data mining and analysis techniques. Leung's work attempts to combine data mining and human intelligence at a more fundamental level to provide focused and filtered data mining techniques. More specifically, his research relates to another hot topic within data science: work with images. His research investigates more effective querying of images and portions of those images to allow for greater search accessibility for users when users are searching for more specific aspects of the image. Research in image identification, manipulation, and querying utilizing machine learning and artificial intelligence approaches is important in data science.

The expansive nature of data science is evident in the examination of just these top 25 authors and only reinforced with the closer inspection of these top four authors' specific works. Also, it is important to note that only 1,646 documents, or roughly 20% of the publications, were single-author documents. This collaborative environment makes sense as the interdisciplinary nature of data science requires knowledge and techniques from multiple disciplines and domains that constitute data science. It appears that data science is a collaborative field of study that can leverage skills and knowledge from multiple disciplines. The collaborative efforts of researchers in data science are finding traction around the world, both geographically and academically, across domains.

6.1.1.4 Key Documents

Similar to the authors, the most cited documents illustrate what is being read, used as foundational work, or potentially as the roots of future works. Table 6.3 lays out the top ten most cited documents in the data set.

Table 6.3 Top 10 most cited documents

Title	Authors	Year	Citations
Machine learning: Trends, perspectives, and prospects	Jordan and Mitchell	2015	1288
Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2	Bolyen et al.	2019	971
Process Mining: Data Science in Action (Book)	van der Aalst	2016	792
Deep learning applications and challenges in big data analytics	Najafabadi et al.	2015	693
MedRec: Using Blockchain for Medical Data Access and Permission Management	Azaria et al.	2016	619
Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management	Waller and Fawcett	2013	580
Big Data: Astronomical or Genomical?	Stephens et al.	2015	537
MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories	McGibbon et al.	2015	466
Data Science and its Relationship to Big Data and Data-Driven Decision Making	Provost and Fawcett	2013	452
The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery	Swan	2013	441

The documents in Table6.3 represent the topmost cited documents and provide some interesting characteristics of their own. Firstly, nine of the ten articles, with van de der Aalst being the only book (Van der Aalst 2016). Two of the documents are papers presenting tooling for big data, specifically with biological data sets (Bolyen et al. 2019; Swan 2013). Interestingly, five of the top ten directly relate to the medical and biological fields (Azaria et al. 2016; Bolyen et al. 2019; McGibbon et al. 2015; Stephens et al. 2015; Swan 2013). This representation in itself suggests that the data being utilized, collected, and stored in biological and medical fields is a prominent field of collaboration for data science. Artificial Intelligence topics appear central in three of the ten documents, with discussions on deep learning, machine learning, and predictive analytics (Jordan and Mitchell 2015; Najafabadi et al. 2015; Waller and Fawcett 2013). These overall characteristics seem to adequately represent artificial intelligence as a top topic, as it appears throughout this study; however, it highlights a less leading topic: biological and medical implementations of data science.

The most cited document in the collection is a 2015 publication by Jordan and Mitchell titled *Machine learning: Trends, Perspectives, and Prospects.* As a review paper, the high citation count may be one of the main reasons why this paper appears as one of the most cited key papers. The paper's main focus is on machine learning, an extremely “hot topic” across the data science industry, media, and academia. Machine learning has been implemented across industries and disciplines such as health informatics (Gu et al. 2017), disaster response (Ofli et al. 2016), and brain-interface systems (Schreuder et al. 2013). As a sub-topic within the greater discussion of artificial intelligence, machine learning has been a huge topic for discussion, reflection, and continued research.

Additionally, the research paper titled *Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2* by Boylen et al. is remarkable because it was the most recently published paper (i.e., in 2019) among all the top-cited ones in Table 6.3. Like the publication by McGibbon et al., Boylen and colleagues present a powerful software tool. The relative speed with which this publication has accrued citations is also a testament to the rapid adoption rate that practical tools can have in data science. This paper’s citation growth underscores the importance of software tooling to data science, especially when designed with flexibility, growth, and specific topics in mind to leverage data science the most effectively.

Researcher Van der Aalst's book, published in 2016, a unique format for highly cited publications in a rapidly developing field such as data science, presents the relationship between data science and process science; taking event data and enterprise systems to generate models and deeper understandings of information systems as the link to process systems. This book dives deeply into the transition that businesses and organizations are undergoing as they begin to coordinate and record everything throughout the systems they have developed. The mining and optimization of these systems through data science techniques like data mining are providing shareholders with undeniable advantages in the form of greater efficiency across practices.

Najafabadi et al. (2015) published *Deep learning applications and challenges in big data analytics* as a work examining big data analytics and deep learning. Both topics are high profile in data science, and Najafabadi et al. present the current state of big data as more and more organizations are collecting data across systems. The article goes on to explore the state of big data analytics and the role and support deep learning now is playing in analyzing and supporting the investigation of large data sets.

The fifth most cited work, *MedRec: Using blockchain for Medical Data Access and Permission Management,* written by Azaria et al. (2016), presents one of the newest technologies that has gained fame and notice: blockchain. As a means of security, anonymization, and veracity, blockchain has been prominent as cryptocurrency; however, the implementation of it to handle big data sets, especially those of exceptionally high-security needs like medical records, has been a topic on the rise. This article presents a blockchain system for medical record management called MedRec. It is both compelling and insightful to state that data science will certainly continue to find roots both in the academic world and certainly in future private endeavors.

6.1.1.5 Top Author Keywords

Analyzing the author keywords (i.e., keywords chosen by authors) in the data science data set collected has revealed an insightful cross-section of data science as it stands today. Figure 6.3 lists the top 25 author keywords used in the current study's data set, and they paint another perspective of research focused on data science.

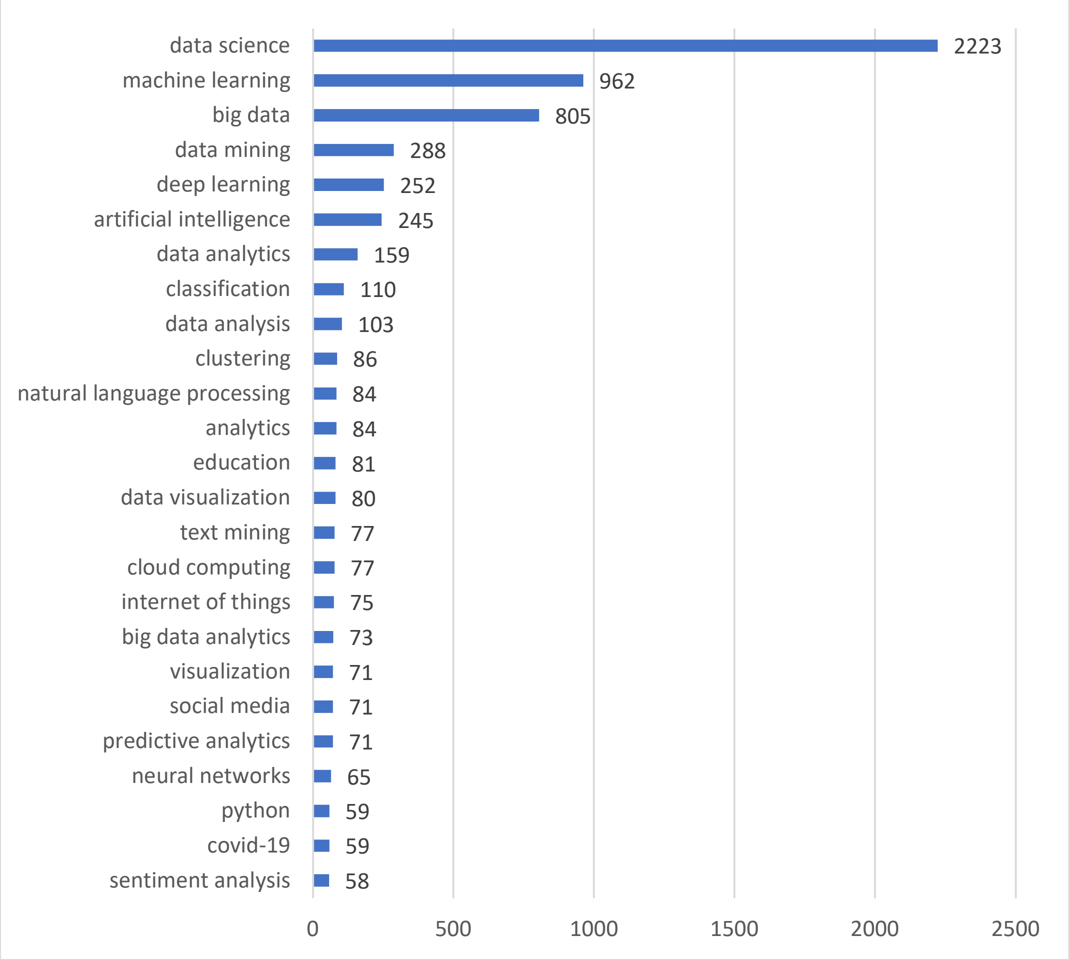


Figure 6.3 Top 25 author keywords

Looking at all the keywords in this study's data set can be problematic because of the sheer volume, but Figure 6.3 provides a snapshot of the most prominent topics across data science and coincides with much of the research discussed in previous works.

The top five keywords in Figure 6.3 are “data science”, “machine learning”, “big data”, “data mining”, and “deep learning”. Each of these keywords is high-profile in data science. The connection to big data has been discussed extensively in this research. Data mining, as one of the most foundational methods for data analysis and one of the oldest terms in data science, also has a strong history. The inclusion of machine learning and deep learning as representatives of artificial intelligence techniques and their growing popularity is a recurring theme throughout this study.

It also needs to be mentioned that the keyword list in Figure 6.3 does allow for some exceptionally time-sensitive topics like COVID-19. COVID-19 appears on the keyword list because the data sets generated by countries and health-related organizations have been enormous during the pandemic, lending themselves to data science techniques and contributing to the creation of academic publications.

As crucial topics of data science, the appearances of machine learning, data mining, deep learning, and artificial intelligence among the top author keywords confirm the status of these topics as popular and of high interest. The list of 25 top author keywords also includes keywords related to other uses, tools, and research sources of data science. Congruently as research topics, these keywords, representing related technologies and implementations, have found a way into applications across various fields and industries. For example, natural language processing (NLP), cloud computing, the Internet of Things (IoT), and social media are all areas of technology that generate sizable data sets and find excellent footing in data science.

It is important to note the highly debated argument between R and Python, as well as the inclusion of Python on the top 25 keyword list (see Figure 6.3). Both R and Python have users across academia and industry that promote each as the premier programming language. Python’s presence on this keyword list might suggest it is being used more frequently in data science research publications than R, which is used more for data analysis and visualization.

Figure 6.4 presents the top ten keywords picked by authors between 2010 and 2021, which shows that the positions of the 10 keywords on the distribution have not been strictly static.

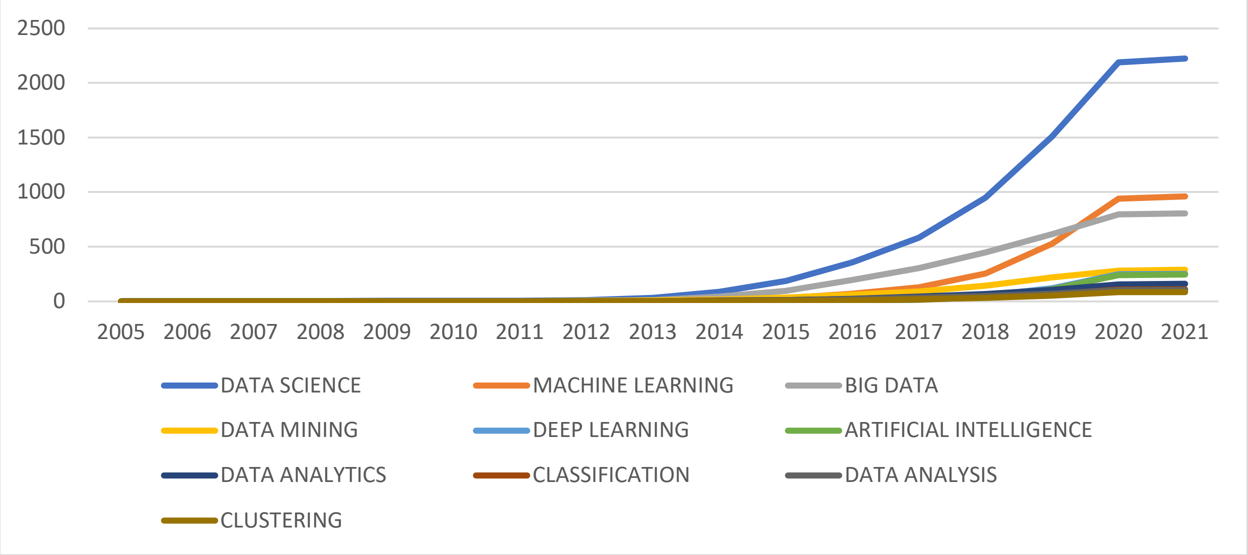


Figure 6.4 Distribution of the top ten author keywords in 2010-2021

Data science as a keyword has shown a strong upward trend over time since its emergence as a discipline around 2010. One notable change is the overtaking of the “big data” keyword by “machine learning” in approximately 2018. This change in positions coincides with several other aspects of the present research in which “machine learning” has more recently become a leading topic in the data science field.

Another unique point of Figure 6.4 is the growth of “deep learning” and “artificial intelligence”, both of which are intensely researched and covered in mainstream media and academia. As a result, they both were increasingly assigned as author keywords in research publications since 2018. The overtake of “data mining” by “deep learning” and “artificial intelligence" demonstrates a change in keyword selection as the slope of all three keywords changes at approximately the same time.

6.1.2 Contributing Publication Sources and Fields of Data Science

The emergence of data science is due primarily to, among other factors, the development and integration of a range of other disciplines and technologies. Their contributions are exceptionally important from the perspective of the field's continued evolution. They help shed light on research areas that may arise or flourish in the coming years and offer insights into the development and roots of data science's traditions and culture.

In the previous chapters of this research report, several disciplines have been mentioned repeatedly; computer science and statistics are at the forefront. Additionally, biology, medicine, physics, astronomy, and other related fields such as information science constitute the remaining contributing disciplines to establishing data science, these fields each deal with some aspects of data science in collaboration and application. Figure 6.5 illustrates the lead contributing fields to data science based on examining the top 100 publication sources.

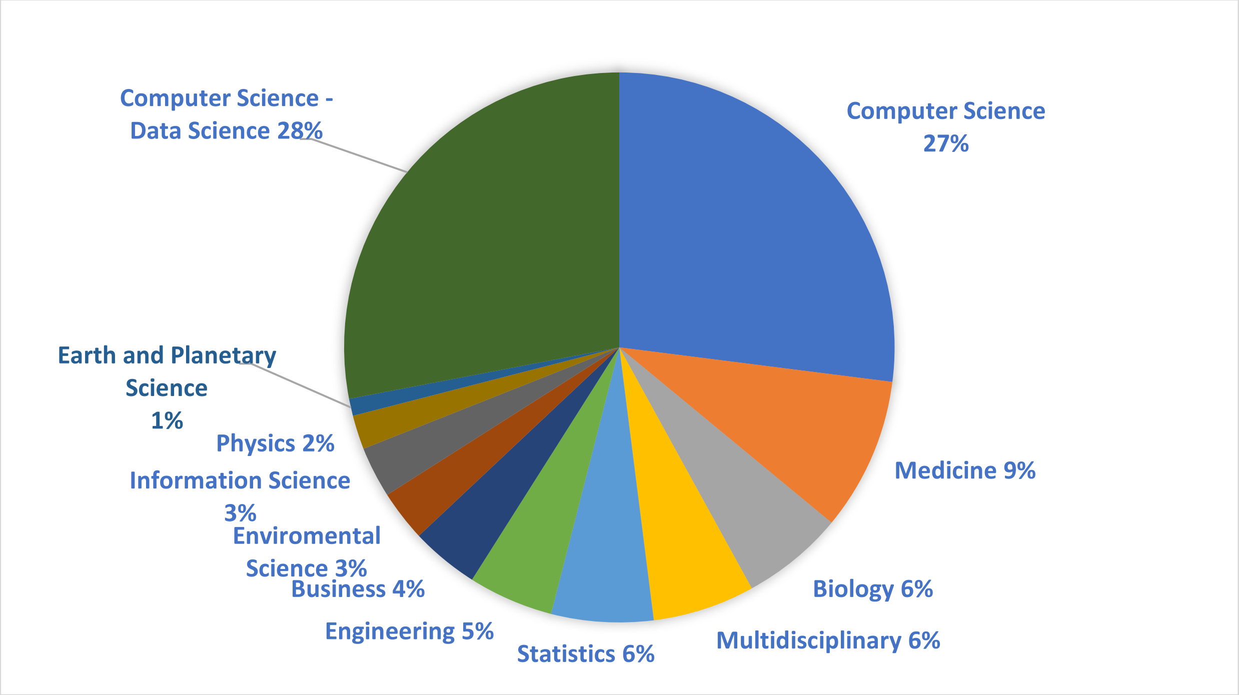


Figure 6.5 Contributing fields to data science based on the top 100 sources

In looking at Figure 6.5, it is apparent that the emergence of data science is based on the contributions of a host of disciplines. Specifically, computer science is the dominant field as it contributes the most to data science academically and practically. A combined 55% of the publications this study examines are from computer science, an overwhelming indicator of its role in the development of data science. Of the 55% of publications collected for the present research, 28% belong to computer science publications focusing directly on data science, while 27% are traditional computer science resources.

Big data is a synonym for data science in a certain sense. As data sets grow in size and scope, the need for data science to harness the most value and understanding also grows. Much of this data accumulation is a direct consequence of computer adoption and integration across society. Alongside data growth, the systems needed to store and manage those resources also expanded; the handling of this development falls mainly in the investigation of computer database systems, a core aspect of computer science. Data science relies heavily on the working space that computer science has developed and created. The ability for data science to operate effectively and, to some degree, exist is therefore very dependent on the realm computer science has generated over the years. In that regard, it is not surprising that computer science plays such a vital role in the development of data science.

For example, medicine and biology together account for 15% of the total, with a wide range of subfields in these two disciplines. These two fields appear as two of the largest fields contributing massive data sets that data science can then be applied to. New approaches in medicine like personalized medicine, precision medicine, and genetics have only heightened the need for data science to parse through findings. Some of the most common implementations of data science have been seen in pharmacological studies and genomic research. More recently, it is often to see data science's rapid deployment and use in the analysis of COVID-19 through machine learning, time-series analysis, and data visualization. Tthe various data generated from the pandemic has provided researchers with opportunities to deploy data science techniques to fight against this rapidly developing disease across the globe.

Statistics has a long history with data science and provides techniques for analyzing and modeling data. Data analysis and mining are comprised mainly of statistical approaches applied to very large data sets. Closely following statistics in Figure 6.5, information science provides structure and techniques for data processing in data science. The need for approaches and models of information to handle the variety and volume of data is a pivotal element that information science has contributed to data science.

Some other fields provide significant amounts of unique data for data science to operate on. Physics, environmental science, and business all contribute to the development of data science by offering unique, large data sets. The application of data science in business has been remarkable, and its use in fast-moving, wealth-generating industries like tech companies has undoubtedly spurred interest in the public. Along with that, physics, including astronomy and data-intensive research projects (e.g., the Large Hadron Collider at CERN), has also facilitated the emergence of data science. It routinely generates incredible amounts of sensory, measurement, and recording data that require processing and analysis. Each of these disciplines forms new territory for data science to analyze and hone techniques and for those working in the field to contribute further publications.

It can be difficult for many stakeholders to stay up to date with current research in data science as the rate of the field increases yearly and drastically. For those looking to focus on research and reading, the analysis of data science publications is a helpful tool. Conference proceedings papers make up 44.5% of all document types, with a total of 3,768 documents. Second to conference proceedings papers are journal articles, representing about 35.2% of the documents, or 2,981 documents. Data science explores topics like AI, IoT, and cloud computing which are leading edge; it is, therefore, anticipated that conference proceedings and journal articles are the two major forms of publications and together account for a significant portion of these publications in the data set. Additionally, the emergence of data science as a multi-disciplinary conglomerate has resulted in the subsequent advent of many conferences dedicated to data science.

Moreover, at many conferences of various academic disciplines, data science is a topic of discussion as it applies to other research within a wide range of domains. Thus, examining the most used publication sources is paramount to understanding the data science research environment. Computer sciences' connection to data science is not novel or hidden from view. The overlap is well documented across research examining data science and big data. This close connection to computer science only makes sense as much of data science exists within the confines of the digital space dominated and developed through computer science. Therefore, it is not surprising to see publication sources like *ACM International Conference Proceedings* and *Lecture Notes in Computer Sciences* leading the list of top publication sources in Figure 6.6.

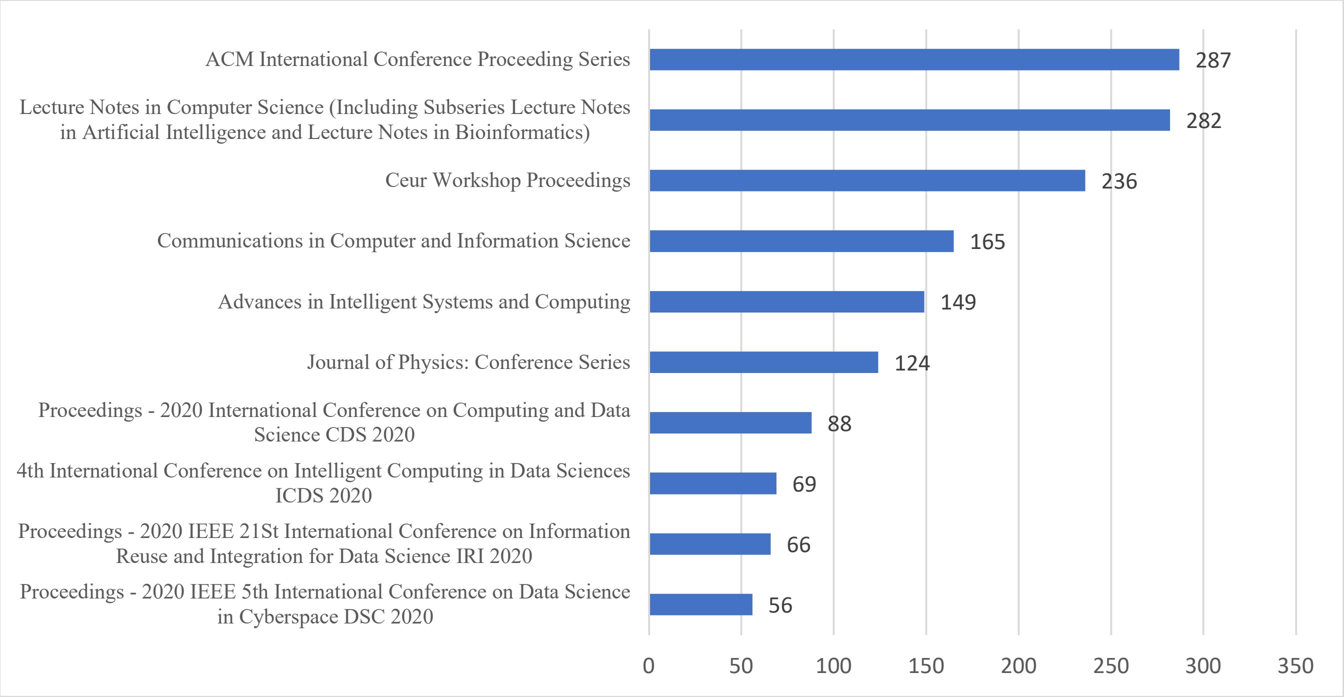


Figure 6.6 Top 10 publication sources

The top three sources, *ACM International Conference Proceedings Series, Lecture Notes in Computer Science,* and *Ceur Workshop Proceedings,* are strongly dedicated to providing the latest computer science, information science, and information technology research. Additionally, nine of these ten publication sources explicitly belong to computer science. Only the Journal of Physics: Conference Series is not a journal focused on computer science or data science. However, in the case of the Journal of Physics, the use of computers for processing and manipulating data sets that are often large is routine and has classically relied on computer science approaches. Physics is also a contributing field to computer science in principle. Furthermore, it should be noted that the majority of this research is coming out of 2020 conferences which is highly suggestive of a dynamic and rapidly growing research community. The composite of these journals is also of note, as IEEE and ACM are leading journals of computer science covering a wide range of topics, many of which become components of other fields or their own fields outright.

The overwhelming representation of computer science sources solidifies computer science at the forefront of contributing disciplines to data science. In the realm of data science, the connection to big data is only possible through computer systems, and its storage and cloud systems are all based on server environments relying on cutting-edge database constructions. Data analysis is often an implementation of algorithms coming out of computer-assisted mathematics. Simply put, because big data sets and analyses of them occur almost exclusively on computer systems, data science finds much of its lineage and future work deeply integrated with computer systems. While early work influenced what data science could accomplish, newer research is more mutually beneficial between computer science and data science as more recent computer systems are designed and constructed with data science and big data in mind.

6.1.3 Major Research Areas of Data Science

The wide range of applications of data science across so many disciplines results from its various research avenues and opportunities. This study identifies such major research topics via two approaches. One uses the cocitation data derived from the top-cited documents in the field, while the other involves grouping author keywords based on their cooccurrence.

6.1.3.1 A Cocitation Approach

Cocitation analysis in this study utilized multidimensional scaling, factor analysis, and cluster analysis to organize and present the intellectual structure underlying the data science field (see Figure 6.7). While this figure was built using the top 100 cited documents, it was filtered to remove any of those documents that did not cluster with any other.

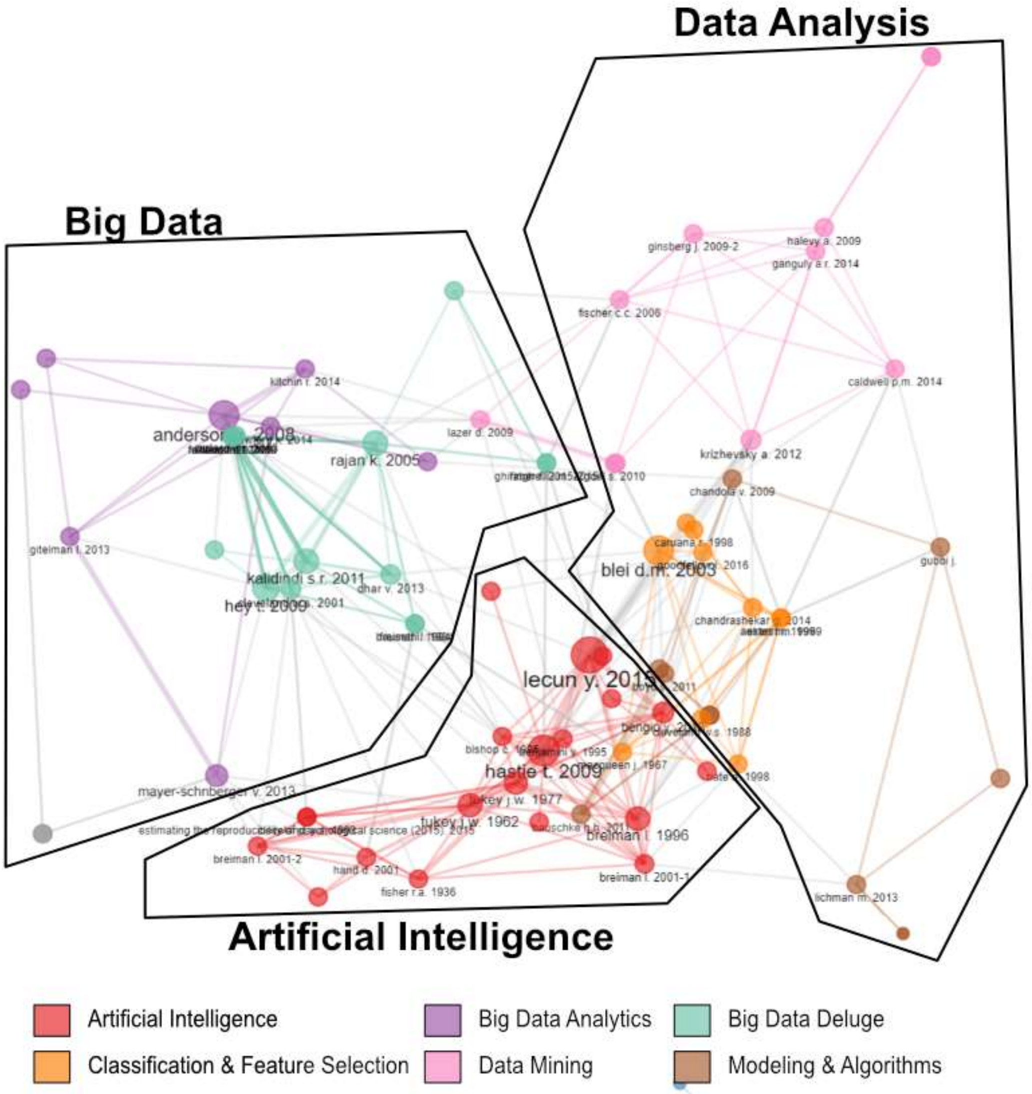


Figure 6.7 Intellectual structure of data science: A cocitation perspective

Figure 6.7 shows the six clusters further grouped into three major themes. The first theme Artificial Learning covers three clusters: deep learning, statistical learning, and machine learning, as they are all comprised of work with neural nets, classification problems, and optimization of models. The re-emergence of artificial intelligence topics and specifically machine learning is a clear indication that data science as a field is highly motivated in this area of study. These three clusters, while approaching the same topic in different manners, all are for enhancing the overall understanding and implementation of these highly complex learning networks.

The second theme is represented by two clusters: big data and the data deluge. These two clusters illustrate the importance of large data sets that have forced the rapid evolution of data science. These clusters present a very similar close link between data science foundational research and big data. The big data cluster examines more closely the history and future of data, how it will affect society, privacy, surveillance, and how data could be approached. The data deluge cluster approaches similar topics and touches on neural networks and statistical mechanics. Both clusters emphasize the importance of these topics within data science and strike a surprising balance between the philosophy of big data and technical aspects aimed at applying methodologies. They also show a solid connection to mathematics by diving deeply into implementing statistics and modeling in their research.

The third theme is represented by the data analysis cluster. As the core aspect of data science, it is a cluster that examines the features, variance, categorization, and evaluation of the data sets under study. These techniques include creating prediction models and adopting heavy use of the statistical methodological tradition. In addition, some members of this cluster step into machine learning to understand large data sets and discover features and models of data, but the cluster is far more focused on exploratory data analysis.

6.1.3.2 A Keyword Approach

Alternatively, author keywords were also analyzed to capture additional topics emerging from the data science corpus. The keyword data set was extracted from the top 100 cited publications. Table 6.4 presents eight groups, each composed of topics and their corresponding keywords, based on the keywords from the chosen documents. Keywords were analyzed using the techniques of cluster analysis and factor analysis to group and form major research topics in data science.

Table 6.4 Groupings of keywords in the top 100 cited documents

Grouping	**Topic**	**Keyword**	**% of Cases**
Data Science	Data Science	Data Science; Big Data;	55.76
Artificial Intelligence	Machine Learning	Machine Learning; Deep Learning	22.01
	Neural Networks	Neural Networks; Neural Network; Deep Learning; Convolutional Neural; Artificial Neural; Convolutional Neural Network; Artificial Neural Network; Artificial Neural Networks; Convolutional Neural Networks; Recurrent Neural; Deep Neural	5.57
	Artificial Intelligence	Artificial Intelligence;	4.62
	Natural Language Processing	Natural Language; Natural Language Processing; Text Mining	1.83
Data Analytics	Data Analytics	Predictive Analytics; Data Analytics; Business Intelligence; Business Analytics; Predictive Modeling; Big Data Analytics; Visual Analytics	3.68
	Data-Driven Decisions	Decision Support; Decision Making; Clinical Decision Support; Decision Support Systems; Decision Support System; Support Vector; Support Vector Machine; Decision Tree; Driven Decision Making; Clinical Decision Support Systems	2.56
	Data Mining	Knowledge Discovery; Data Mining; Text Mining; Knowledge Management; Text Classification; Knowledge Graph; Knowledge Representation	2.04
	Time Series	Time Series; Time Series Analysis; Time Series Prediction	1.54
	Feature Selection	Feature Selection; Feature Extraction;	0.89
Computer Science	Cloud Computing	Cloud Computing; Performance Computing; High-Performance Computing; Mobile Computing	2.4
	Software Engineering	Software Engineering; Software Analytics; Software Repositories	1.05
Statistics	Statisticals & Models	Methods; Statistical; Models; Computational	0.38
Education	Education	Science Education; Computer Science; Computer Science Education; Data Science Education; Curriculum Development; Data Science Applications In Education; Higher Education; Distance Education And Online Learning; Secondary Education; Computing Education; Engineering Education; Curriculum Design; Data Science Curriculum	1.61
Ethics and Security	Data Ethics	Data Ethics; Security	0.78
	Social Media	Social Media; Social Science; Computational Social Science; Social Network; Social Data; Social Network Analysis; Social Media Analytics	2.21
	Internet Of Things	Internet Of Things;	1.72

Emerging Technologies	Smart Cities	Smart Cities; Smart Sustainable; Urban Science; Smart City; Smart Sustainable Cities; Urban Data; Urban Analytics; Urban Planning; Smart Sustainable Urbanism; Urban Intelligence; Urban Sustainability; Urban Data Science; Big Data Computing	0.96
Health & Medicine	Health Informatics	Electronic Health; Health Care; Electronic Health Records; Health Informatics; Health Data; Public Health; Health Information; Digital Health; Medical Informatics; Electronic Health Record; Information Systems; Information Science; Mental Health	2.48
	Precision Medicine	Precision Medicine; Personalized Medicine	1.02

Data science and big data, the two words almost synonymous throughout the field, appear in Table 6.4 as being highly correlated and account for over 50 percent of keywords listed in the top 100 cited documents. Artificial intelligence, computer science, data analysis, and statistics are also major groupings identified through the examination of the keyword data set obtained. Specifically, artificial intelligence boasts several of the most popular subtopics: machine learning, neural networks, and natural language processing. Machine learning, in particular, seems to be perhaps the most popular topic in data science at the moment of this writing. The groupings of computer science and statistics are both in alignment as not only major topics but also major disciplines that contribute to the formation of data science. In the case of data analysis, data feature selection, data mining, and data-driven decision-making are the top topics that emerged from the keywords.

When it comes to emerging technologies and health and medicine, it is concerned more about how technologies are utilized in data science or data science is applied to large data sets to optimize, analyze, and model information. For example, smart cities and the internet of things are for huge data collection systems to expand as information systems with more sensors. In the case of health and medicine, the topics revolve around handling the incredible amount of medical records that exist in the healthcare system as well as the new movement towards personalized medicine. These two groupings suggest a strong affinity for research in these areas alongside data science, especially as prime sources for data perfect for data scientists. Each of these regions of study is extremely important as industries and society are attempting to harness fast-moving, highly varied data.

Figure 6.8 visualizes the groupings and related keywords based on the keywords taken from the top 100 cited documents. It helps reveal some of the underlying intellectual structure of data science yet not seen easily in Table 6.4. Furthermore, each of the groupings is encircled to provide visualization of the intellectual structure underlying the field of data science from the perspective of keywords.

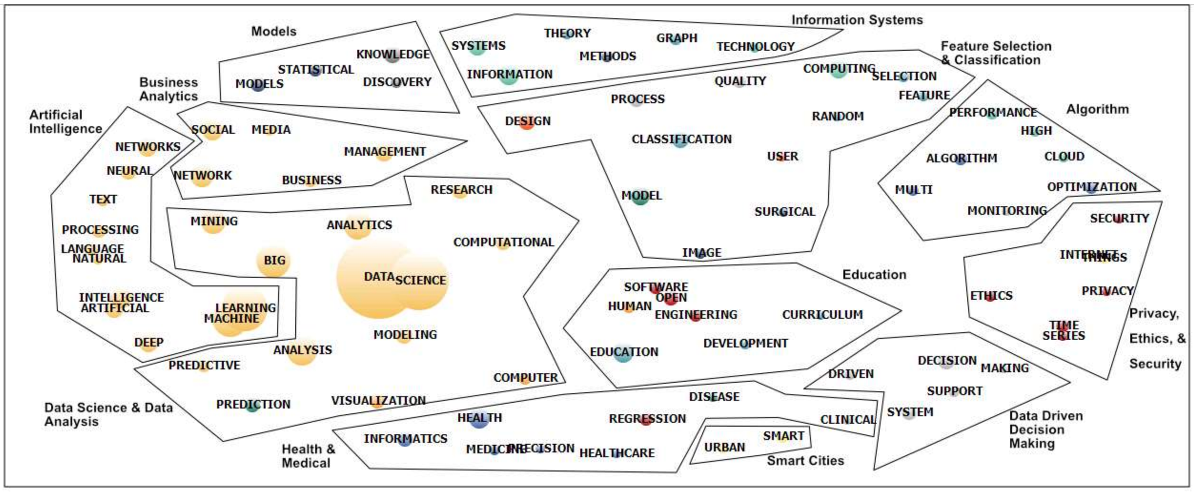


Figure 6.8 Intellectual structure of data science: a keyword perspective

This keyword mapping yields many more subject clusters than the one based on cocitation data. While the clusters are highly subjective, they do help make sense of these keywords included in the analysis. Moreover, these groupings facilitate elucidating the most pertinent salient topics from the keyword data set collected for the current study. What appears evident is the proximity of the Artificial Intelligence grouping and data science and data analytics grouping, which helps reveal the research focuses in data science, especially around artificial intelligence and its sub-topics. Moreover, those keywords in the business analytics cluster are also adjacent to the core data science grouping, suggesting a high degree of connectivity between the two clusters. Again, the position of the smart city, as well as healthcare and medical groupings in Figure 6.8, also hints at crucial applications of data science. These two particular groupings show how other disciplines utilize and affect data science research.

This section of the study approached three unique research questions, the first of which was the identification of the scientometric features of data science. To this end, the characterization of the field through 8,458 records. Top authors, institutions, and documents have shown that while the majority of research is occurring in the United States, the distribution of work, especially top-cited work, is becoming more and more global.

In a similar pattern, in the examination of the second research question regarding the contributing fields of data science, a few key aspects appeared. One was the fact that computer science plays a leading role in this regard with support from the statistics, biomedical, and information science fields. However, what also became apparent is that many more fields are contributing their data to be acted on by data science methodologies and technologies, fields like the biomedical, business, and engineering fields are a few examples.

The third research question addressed in this portion of the research was seeking to understand how the scientometric data would present the leading research topics. Nine major groupings appeared in the analysis, but among them, the most evident across the whole section with artificial intelligence, with topics like deep learning, machine learning, and neural networks. Artificial intelligence appeared in top documents, top authors, keywords, and intellectual structures. Other topics included data analysis, with subtopics in data mining and visualization. Also, the closely tied topics of computer science appeared with specific topics like databases and cloud computing. These three topics represent a few of the topics, with more being addressed in Section 6.4 of this chapter.

6.2 The Curricular Perspective of Data Science

The scientometric perspective of data science, presented in Section 6.1, addresses the first three research questions formulated for this study. The current section aims to address the fourth research question of this study: What are the salient topics taught in the curricula of data science. This study examines the course titles and descriptions in the curricula of data science from 102 institutions. This section reports the characteristics of the curricular data set to provide a unique perspective for the present study. Table 6.5 provides an overview of the curricular data set.

Table 6.5 Parameters of the curricular data set

Parameter	**Frequency**
Academic Institutions	102
Programs	125
Masters Bachelors Certificate	68 35 22
Courses	3128
Masters Bachelors Certificate	1892 966 270
Avg. Courses offered Per Program	25.0
Required Courses	1256
Avg. Number of Required Courses	
Masters Program Bachelors Program Certificate Program	16.4 18.3 9.4

Table 6.5 indicates that far more courses are provided at the master's level than at the bachelor’s because those programs can quickly produce graduates to meet the demand of the job market. Furthermore, as discussed previously, data science features interdisciplinarity, and many fields contribute to its emergence and development. Hence it is natural for data science to have master's programs with students from related or contributing fields such as computer science and statistics.

The University of Michigan in Ann Arbor had the most course offerings listing 94 courses in their master's program, including elective options. Moreover, the University of Michigan program was offered through the statistics department, which is interesting to note as statistics is an essential aspect of data science. Still, it is not always the case that these master's programs reside in the same department as data science programs can be offered through many different departments. Furthermore, not all schools structure their universities with similar department hierarchies. The range of departments and department-like groupings that universities utilize for their data science master’s is extensive. Data science programs can be hosted in disciplines such as computer science, statistics, mathematics, engineering, and information sciences in many universities. Some institutions (e.g., New York University) may create a new academic unit specifically for data science.

Similarly, certificate programs are offered at many universities and through different departments. However, these certificate programs often run parallel to master's programs as an option for those who have already obtained their master's degree to update themselves. In many cases, academic units offer certificates with degree programs (e.g., Syracuse University). The certificate programs connected to a master’s program allow certificate students to take master’s level courses that can later be counted toward the completion of a certificate. While these certificate programs are minimal and do not tend to discuss more advanced topics beyond the introductory concepts, the value they provide prospective students is precious.

Surprisingly, nearly 26.9% of all courses included in this study only had a title on their website. This statistic seems alarming because data science is currently a burgeoning and competitive subject for students, and these course descriptions offer meaningful insight into the course. This absence of information is even more concerning because the curricular data collection relied solely on institution and program websites, which is now also a major means for students to quickly gather information about programs, schools, and potential enrollment information. Course titles and descriptions from all levels of the data science curricula are analyzed to examine the topics included in the educational programs. Course titles are concise, informative, and highlight the course theme. On the other hand, course descriptions offer a plethora of information about the courses but can vary greatly from course to course. By virtue of their length, course descriptions frequently offer more information on how the course is laid out, what subtopics may be covered, and how students will go through the course. Examination of these courses through content analysis seeks to identify the features of data science through the curricular lens.

6.2.1 Analysis of Required and Elective Courses

Based on the curricular data set, data science core courses are identified across institutions. These courses have been grouped into nine topic-oriented core data science courses and five topics that appear frequently but are not required across programs. Table 6.6 displays the breakdown of required courses.

Table 6.6 Required courses in data science

Offered in All Programs	**Frequency**
Statistics	274
Data Analysis	171
Computer Science	157
Data Science	126
Databases	125
Artificial Intelligence	83
Big Data	39
Visualization	30
Ethics, Privacy, Security	19

Offered in Some Programs	**Frequency**
Business	58
Information Science & Informatics	27
Communication	20
Medicine	19
Research Methods	12

Among the core courses offered in all programs, some are foundational courses in computer science, such as databases and data analysis. These required courses illustrate a conceptual core in course design at the higher education level. These course topics also help highlight the most important content in the data science curriculums across the United States. Likewise, statistics is another crucial and fundamental aspect of data science. As shown in Table 6.6, data analysis also emerges as a top-tier topic for core courses.

The other data science core course topics identified in this research include databases, artificial intelligence, big data, and visualization. The importance of each of these topics is reflected primarily through the course descriptions both in electives and core courses.

The required courses offered in some programs (see Table 6.6) represent courses that are considered fundamental in some of the data science programs included in this study. Among them are courses about business-focused data analytics and other business-centric topics like economics. While these courses have not been grouped with the core courses in all data science programs, they are regarded by some programs as highly relevant topics required for data science students to understand.

Similarly, an examination of elective courses can provide insights into the data science curriculum's wide variety of topics. Most of the courses gathered for the present study are from master's level programs with an extensive range of electives. There may be multiple contributing factors to the range of electives; the idea that data science can be applied to almost any large data set is a testament to this. Table 6.7 presents an in-depth grouping and breakdown of elective courses in the curricular data set collected for this study. These groupings were constructed by analyzing the course titles and descriptions. These groups were organized into groups based on each course's relationship to data science's most popular topics.

Table 6.7 Elective courses groupings by topic for data science

Grouping	Course	Frequency
Statistics	Statistics	238
	Probability	98
Data Analysis	Data Analysis	205
	Data Mining	34
	Computer Vision	12
	Geographic Information Systems	9
	Time Series	5
Computer Science	Computer Science	118
	Programming	57
	Algorithms	39
	Web	9
AI	Machine Learning	73
	Natural Language Processing	30
	Artificial Intelligence	27
	Deep Learning	16
	Neural Networks	5
Databases	Databases	98
	Cloud	7
Data Science	Data Science	54
Ethics & Security	Security	27
	Ethics	6
Big Data	Big Data	31
Visualization	Visualization	21
Other Disciplines	Medicine & Biology	73
	Business	67
	Information Science	46
	Econometrics	24
	Research Methods	28
	Communication	9

Table 6.7 provides both groupings and course topics to signify the depth and breadth of the elective courses offered in data science while showing the wide variety and finer points of electives offered across all programs. Moreover, Table 6.7 enables the noting of several well-entrenched topics (e.g., natural language processing) in data science that are often overshadowed by more well-known topics (e.g., artificial intelligence). Artificial intelligence is, of course, one

such overshadowing topic, with it being prevalent in data science and beyond. However, natural language processing is an overshadowed topic, gets listed in Table 6.7, indicating that NLP is an elective course in a good number of data science programs. Additionally, several less-known subtopics such as computer vision, geographic information systems, and time series analysis are also offered as elective courses in some data science programs.

In addition, Table 6.7 has one part named other disciplines that comprises elective course topics that fall outside of data science. The three that emerged most distinctly were medicine, business, and information science. This researcher also separately included econometrics as many courses explicitly mentioned it. The analysis and applications of financial and economic data using statistical methods are a perfect partner for data science. Econometrics plays a far different role than classical business-related courses that are often more focused on pragmatic implementations of data science. Research methods and communication are topics often integrated with other topics in courses; however, some courses are explicitly dedicated to their explicit discussion.

Table 6.8 lists some of the more unique elective course topics. These courses are topics that only showed up in a few data science curriculums and may represent topics in either emerging technologies or research. Topics that going forward will almost certainly leverage data science in the future. Each of these emerging topics can eventually become a course or course session, implying that they will be taught and discussed in data science in the future.

Table 6.8 Potentially emerging electives

Courses
Biomarkers
Blockchain
Computer Architecture
Cyberwarfare
Design Thinking
Digital Archaeology
Entrepreneurship
Genetics
Internet of Things
Policy
Social Networks

These emerging topics represent some of the leading edges of larger fields. For example, cyberwarfare is directly related to the work being done around fields addressing technological privacy, security, and ethics. In a similar manner, biomarkers and genetics are closely related to the ever more important field of biomedical. Each of these topics is worth watching as they represent greater fields and their own expansion and the potential future uses and understanding in data science.

6.2.2 Further Analysis of Course Titles and Descriptions

It is essential to recognize that course titles, similar to course descriptions, are a representation of the curriculum. They provide a window looking into the ‘full’ curriculum, including the most critical topics of the data science field that educators have deemed essential for preparing the next generation of the workforce. Course titles follow the standard conventions of titling, with verbiage that does not typically surpass ten words, while many often have one to several words. The result of this short title structure, coupled with the fact that the purpose of course titles is to give students a glimpse into the content of the course, determines that course titles highlight the most significant topics to be taught, and some even offer an indication of the course level. For example, “Statistical Inference for Data Science I” and “Statistical Inference for Data Science II”. However, the majority of course titles simply indicate the most prominent topic being discussed (e.g., “Data Visualization”).

On the other hand, course descriptions contain more content than titles. In examining the course description data collected for this study, it became apparent that the range and depth of course descriptions varied significantly from program to program or even at the course level. When course titles and descriptions are used in combination for data analysis, they offer a full overview of important topics of individual courses as well as about the entire program.

6.2.2.1 Analysis of Keywords

To visualize some of the most prominent words utilized in data science courses, Figure 6.9 presents a word cloud based on the course titles and descriptions in the data set collected for the current research.

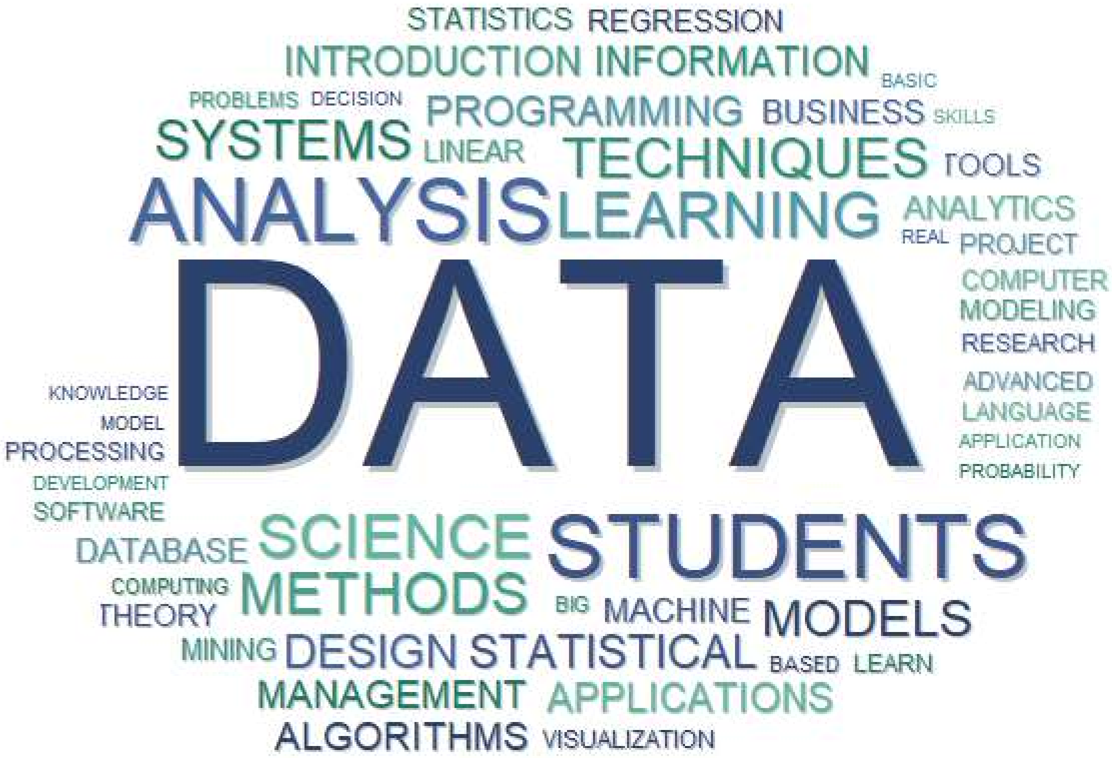


Figure 6.9 Word cloud of the top 100 keywords in course titles and descriptions

Figure 6.9 includes some of the most noticeable words that stand out from the cloud and correlate closely to the topics derived from the content analysis of the courses (see Tables 6.6 – 6.8). While the keywords “data” and “science” appear pronounced in the cloud, the word “analysis” provides a more meaningful correlation to courses. Data analysis as a course topic is entwined with nearly every course in data science. Additionally, ‘visualization’ is also easily connectable to the topic of data visualization, which, similarly to data analysis, is a basic topic in data science courses across programs.

Figure 6.9 demonstrates that statistics and probability represent an essential topic in the data science curriculum. More specifically, it can be seen from Figure 6.9 that statistics plays a prominent role in data science with related words like “statistical”, “mathematical”, “models”, “algorithms”, “regression”, “linear”, and “probability”.

Computer science courses constitute a good portion of courses in nearly every data science curriculum. “Databases” is perhaps one of the most critical keywords connected to computer science courses, and so are “programming”, “application”, “applications”, “algorithms”, and “software”. Additionally, Figure 6.9 shows a set of words connected closely to computer science like “computer”, “processing”, “computational”, “computing”, and “engineering”.

It is interesting to note in Figure 6.9 that “business” is a word not necessarily attributed to data science in the traditional sense. When it is included in data science course titles and descriptions, it becomes a major topic with a focus on the application of data science for the optimization of businesses.

6.2.2.2 Analysis of Key Phrases

While Section 6.2.2.1 describes and discusses the results of keyword analysis of course titles and descriptions collected for this study, the current section presents the results of key phrases extracted from the course data set. These phases were collected based on the top cooccurrence of words by extracting relevant phrases in the course titles and descriptions. Figure 6.10 displays the top 25 key phrases in the curricular data of this research.

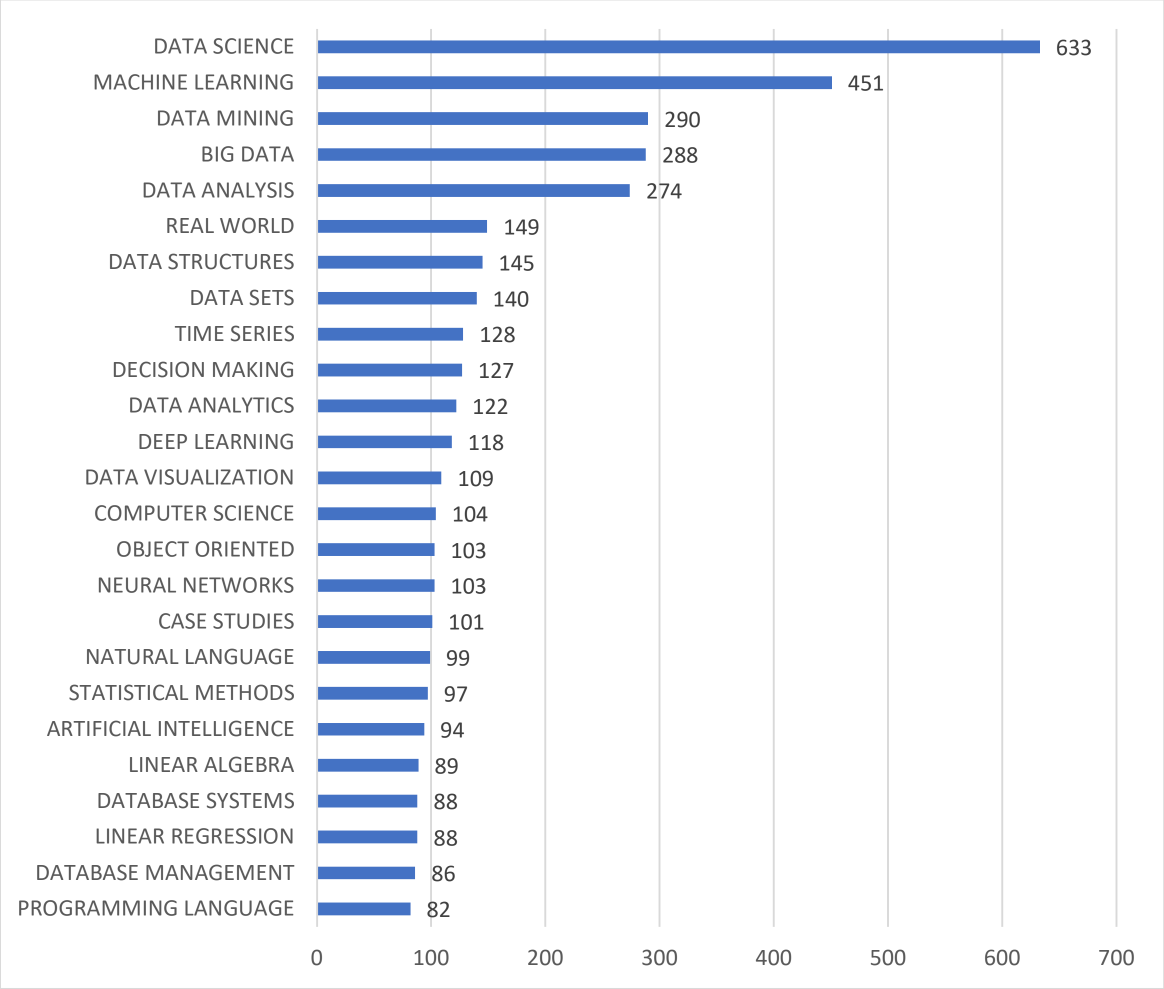


Figure 6.10 Top 25 phrases extracted from course titles and descriptions

Figure 6.10 further expands on the role of words in depicting the content of courses designed to be consumed by current and prospective students besides adding the context for understanding each course in the curriculum. One such example is the phrase "machine learning". As a topic of data science, machine learning is at the forefront of current research. With applications across disciplines, machine learning is a topic strongly supported in Figure 6.10 as a key phrase second only to "data science".

Compared to the analysis of the single keywords, these phrases represent a much clearer picture of the major topics in data science. Following machine learning are three more topics of data science in Figure 6.10: data mining, big data, and data analysis. Each topic is essential to the field and well represented in courses; some other crucial data science topics in the top 25 consist of time series analysis, deep learning, data visualization, neural networks, natural language processing, artificial intelligence, database systems, and database management.

In the case of AI, the representation of it is even more robust, with machine learning, deep learning, neural networks, natural language processing, and artificial intelligence all among the top 25 phrases. Those topics account for 20% of the top 25 phrases, which strongly indicates that AI is one of the cornerstones in data science education and research.

In addition, phrases concerning statistics are also accounted for as statistical methods, linear algebra, and linear regression, all among the top 25 key phrases in Figure 6.10. This confirms the close relationship between statistics and data science.

6.2.2.3 Analysis of Keyword Groupings

In performing a cluster analysis of cooccurrences of the keywords, thirteen data science topics are identified and labeled(see the Topic column in Table 6.9). These topics are representative of broader data science topics as reflected in the curricular data set. Table 6.9 has been augmented and grouped to illustrate the keywords, topics, and overarching conceptual relationships as larger groupings. Additionally, the percent of courses for each topic grouping is also presented in Table 6.9.

Table 6.9 Subject and topic groupings by course title and description keywords

Subject	Topic	Keyword	% Cases
Data Science	Data Science	Data Science; Big Data; Introduction To Data Science; Data Analytics	12.6
Databases	Database Systems	Database Systems; Database Management; Information Systems; Database Management Systems; Geographic Information Systems; Advanced Database	3.42
	Data Structures	Data Structures; Data Structures And Algorithms	0.48
Statistics	Statistics	Applied Statistics; Mathematical Statistics; Probability And Statistics; Statistics; Statistics For Data; Bayesian Statistics	2.88
	Linear Algebra	Linear Algebra; Linear Models	1.95
	Calculus	Calculus II	0.74
Artificial Intelligence	Artificial Intelligence	Artificial Intelligence; Intelligence; Artificial	1.28
	Machine Learning	Machine Learning; Deep Learning; Applied Machine Learning; Advanced Machine Learning	6.3
	Natural Language Processing	Natural Language Processing	1.21
Data Analytics	Time Series	Time Series; Time Series Analysis	1.31
	Business Analytics	Business Analytics; Marketing Analytics; Advanced Business; Business Intelligence	1.41
Education	Special Topics for Data Science	Special Topics; Data Science; Special Topics In Data Science; Computer Science; Topics In Computer Science; Advanced Topics; Computer Vision	1.82
	Research Methods	Statistical Methods; Research Methods; Statistical Inference	2.17

The topic groupings in Table 6.9 offer a more specific, focused, and aggregated presentation of the word usage in the curricular data. However, these thirteen topic groupings have been further grouped into six more extensive and prominent subjects: data science, databases, statistics, artificial intelligence, data analytics, and education. These subjects have been created to provide a framework to represent the thirteen groups coherently. The subject cluster for data science is the only one having the same name as the topic column in Table 6.9. Within that group, the keyword “big data” appears to show its close connection to big data, which is, in fact, a synonym of data science.

The education subject comprises the "Special Topics" and "Research Methods" clusters. The special topics courses run the gamut of individual, emerging data science topics. For example, the University of San Francisco offers special topics courses: geographic information systems (GIS), political analytics, sports analytics, supply chain analytics, marketing analytics, and simulations. Northeastern lists a course called "Special Topics in Data Science" that covers a wide range of topics such as machine learning, data mining, bioinformatics, information retrieval, and natural language processing. These examples do not exhaust the entire listings but showcase the kinds of topics taught in these courses. These special topic courses are essential in the educational landscape and offer flexibility for these data science programs. Moreover, they allow for course adjustments from semester to semester based on department choices, course rotations, and certainly staffing alignments.

Of the remaining four groups, databases, data analytics, statistics, and artificial intelligence, each one bolsters the standings of these topics as core topics within data science. The scientometric findings presented in Section 6.1 strongly correlate with the significance of these groupings being essential topics in data science. These subject groupings also highlight some of the most influential specific topics. In the case of statistics, for instance, the keywords reflect the importance of Bayesian statistics, applied statistics, linear algebra, and calculus as highly relevant topics in data science.

To further examine the curricular keywords, Figure 6.9 was created to visualize the top 100 keywords via MDS on course title keyword cooccurrences. This visualization helps explore the intellectual structure of data science while providing a new perspective with which to view these keywords and how they might be clustered.

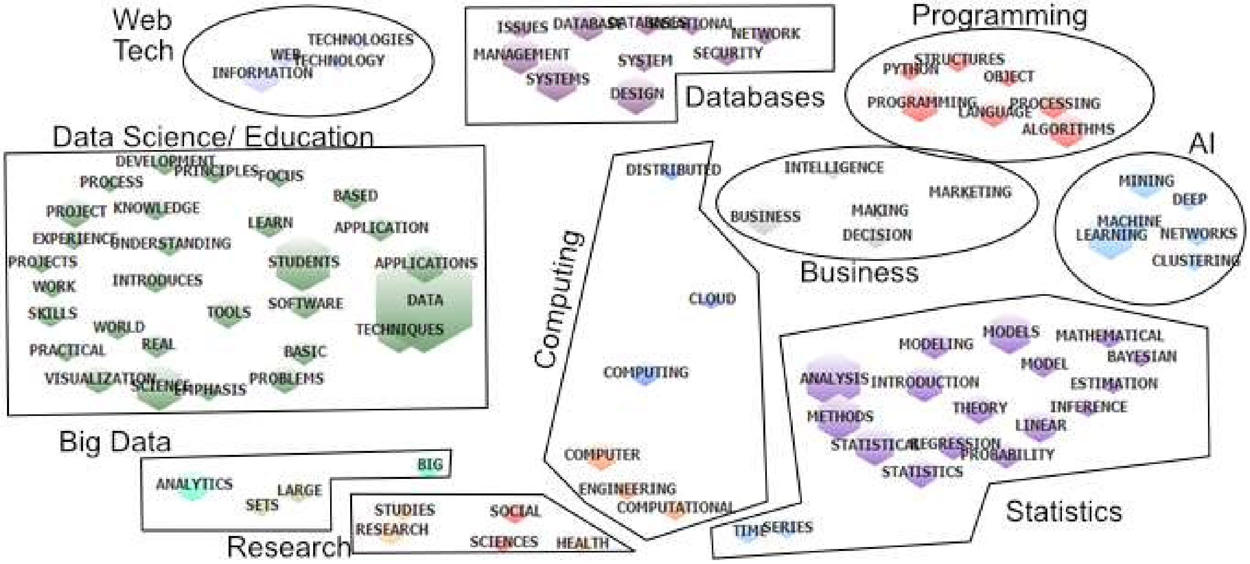


Figure 6.11 Intellectual structure of data science-based on cooccurrences of course title keywords

The mapping of these 100 course title keywords in Figure 6.11 allows for a more thorough and distinct description of the subject relationships among the important topics in the data science curriculum. The ten groupings collectively represent the overarching topics of data science. The clusters shown in Figure 6.11 relate very closely to those already presented in Table 6.9, as well as represent those keywords seen in Figure 6.10. All these findings together reveal the intellectual structure underlying data science.

Three points are of note from Figure 6.11. Firstly, the major topics covered in the curricular data set include artificial intelligence, big data, business, computing, databases, data science, programming, research, statistics, and web technologies. Of these eleven clusters, big data, databases, time series analysis, artificial intelligence, and statistics are all topics that have been highlighted in this section before and fit well in the field of data science.

Secondly, the grouping of web technologies appears to be a fascinating data science topic primarily because it does not show up so obviously in other analyses. Its presence, without a doubt, hints at further internet and network integration both on the technical side as well as the data side of data science. In course descriptions, this topic is often included in terms of web-based systems, programming, and networking.

Thirdly, the statistics group represents multiple statistics-based topics such as bayesian, linear, and regression. As a cluster, the size and position of statistics indicate that it is a critical topic in data science nestled among AI, business, and computing.

This section was focused on the discovery of the salient topics in the data science curriculum. Artificial intelligence and its accompanying topics were a leading topic in the data science curriculum, very similar to what was seen in the scientometric portion of this study in Section 6.1. Computing, business, and statistics also were uncovered as major topic groups for the curriculum. The examination of the curriculum also provided topics less obvious, like the continued appearance of the database, data structures, and programming aspects required of new data scientists. Also insightful was the examination of topics emergent of electives, and while far less popular across institutions, could be indicative of areas of growth and will be interesting to watch over the coming years.

6.3 The Altmetric Perspective of Data Science

This section presents findings using the altmetric data set collected for the present study. While this research part focuses explicitly on analyzing hashtags within tweets, the hashtags were examined to understand how data science is communicated and discussed within the Twittersphere. The hashtags in this study are treated similarly to the keywords in Section 6.2, although, as known, hashtags are inherently different from keywords.

6.3.1 Analysis of Tweet Hashtags

As a straightforward mechanism of grouping tweets based on a user-directed topic, hashtags provide a powerful mechanism of Twitter that allows for the easy connection of topics and discussions. From a user's standpoint, these hashtags can serve many purposes, but the primary purpose is to generate topic-oriented open conversations. Nevertheless, one aspect of this topic grouping and community building on Twitter is shorthand and abbreviations that refer to topics and are readily discernible for the community but may be cryptic for those unfamiliar with a conversation's unique jargon. While some abbreviations are relatively standard, like "AI" being short for "artificial intelligence", others are more technology-oriented, such as "IoT" being shorthand for "Internet of Things". Hence, a list of abbreviations relating to data science can be found in Appendix D.

6.3.1.1 Visual Analysis of Top Hashtags

To analyze the conversations through Twitter regarding data science, this portion of the present study takes a deep look at the top hashtags and attempts to derive topics from those hashtags in the Twitter discussion. As a starting point, Figure 6.12 displays the top 25 hashtags; all variations in capitalization and abbreviations have been considered in the hashtags shown.

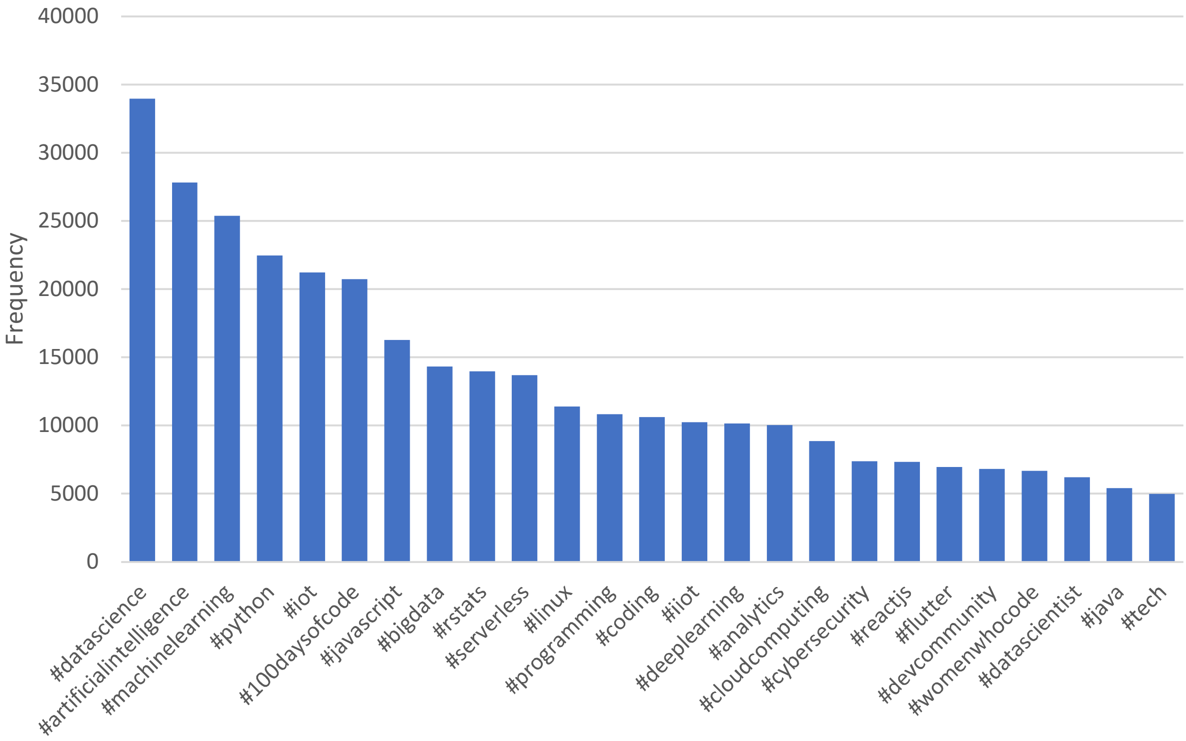


Figure 6.12 Top 25 Twitter hashtags on data science

Artificial intelligence and its subtopics are the most discussed conversation on Twitter, and this is very much in line with the findings from the scientometric and curricular data sets. AI is represented via three hashtags: #artificialintelligence, #machinelearning, and #deeplearning. Machine learning and deep learning are principal aspects of artificial intelligence work in data science. AI even managed to edge out #bigdata and #datascientist, which would intuitively seem to be far more closely related to #datascience. Even more impressive is #artificialintelligence was so close to #datascience, which is the hashtag used to collect all of the tweets and so has a 100% occurrence in the Tweets collected. This high representation puts #artificialintelligence and its abbreviation #ai at approximately 82% of all tweets collected.

Following AI in Figure 6.12 is programming-related topics. Those hashtags include not only the straightforward, high-frequency ones such as #programming and #coding seen but also hashtags for programming languages, communities, and technology. For language-related hashtags, #python, #rstats, #javascript, and #java are common choices. Python and R both represent data analysis languages, while Javascript is more closely related to web technologies. Moreover, two web programming frameworks (i.e., #reactjs and #flutter) are all present in the top 25. The communities dedicated to learning and developing programming skills help expand the programming topic discussion on Twitter with hashtags #100daysofcode, #devcommunities, and #womenwhocode.

Along with artificial intelligence and programming, emerging technologies are also a popular topic on Twitter, although it is surprising to see that #IoTis among the top 25 hashtags on data science. In both curricular and scientometric investigations, the internet of things is mentioned but not in an overwhelming or even truly meaningful manner. Additionally, while slightly different from #IoT, the presence of #IIoT, "Industrial Internet of Things", in Figure 6.12, only adds to the surprise previously noted about #IoTat the top of the hashtag list. Both refer to the implementation of the internet or networked devices that historically have not been. Also, other technology-related hashtags in Figure 6.12 include #serverless and #cloudcomputing.

One hashtag, #cybersecurity, ranked the 18th in Figure 6.12, is noteworthy as it stands apart from the others and is growing in importance. It is also analyzed in both the curricular and scientometric portions of this study. This hashtag, alongside #datascience, gives the topic of cyber security a much greater attribution than it receives in the other parts of the current study.

Figure 6.13 presents the word cloud created with the top 100 hashtags from the Twitter data set. This word cloud comprises hashtags without the "#" symbol and provides a more visual representation of the most popular hashtags in Twitter conversation around data science.

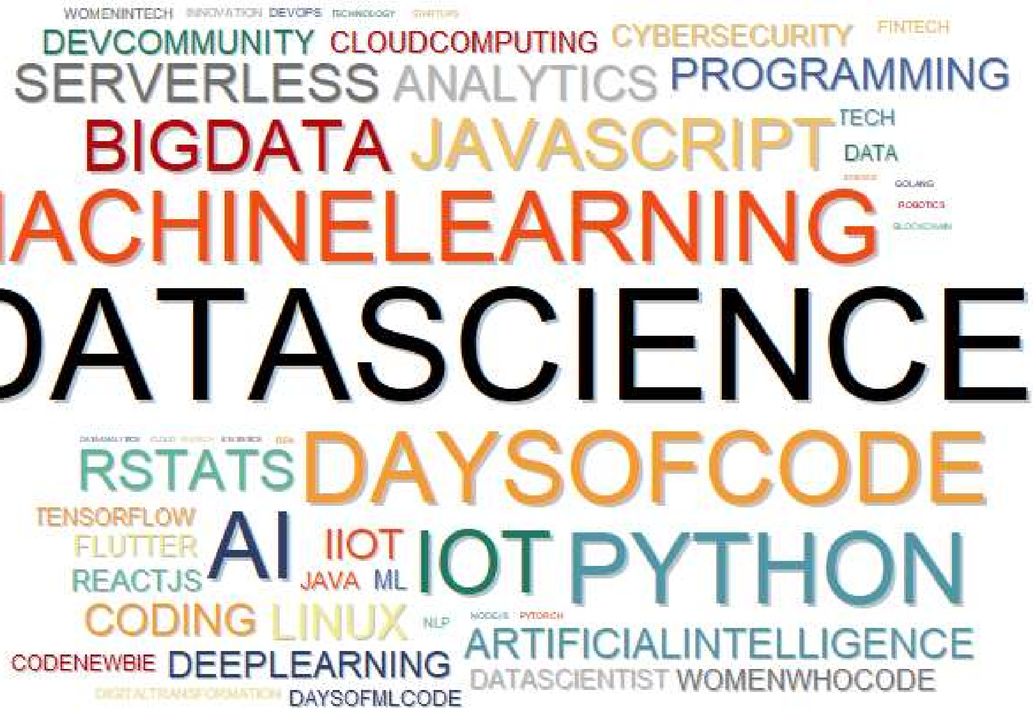


Figure 6.13 Word cloud of the top 100 Twitter hashtags

Highlighted in Figure 6.13 are several data science topics that appear front and center. The largest hashtags are machine learning, artificial intelligence, and deep learning. In addition, TensorFlow and Python represent programming language platforms dedicated to artificial intelligence work. The conspicuity of machine learning, particularly across the data science spectrum of conversation, is impressive, and its ability to be consistently at the top of discussions regardless of the medium seems eminently essential to notice.

Big data is notably present as a significant hashtag, indicating that big data and data science are entangled in discussions across Twitter as it seems to be everywhere. Although Figure 6.13 includes 75 more hashtags than Figure 6.12, the topics of discussion remain consistent with more hashtags about programming languages and communities in Figure 6.13. Fintech, short for financial technologies, does make an appearance and is the first strong hashtag for the Twitter discussion on business and finances in conjunction with data science.

6.3.1.2 Factor Analysis of the Top 100 Hashtags

Figure 6.14 contains results that are not structured or organized like those obtained from in other two sections of this research. In general, hashtags are created more casually and with little regulation and supervision. Often, tweets are filled with hashtags both pertinent and tangential to data science. Figure 6.14 visualizes the intellectual structure of the discourse using the top 100 hashtags in the Twittersphere based on hashtag cooccurrences. Although what is presented in Figure 6.14 seems less coherent than the results obtained from the scientometric and curricular data gathered for this study, Figure 6.14 still facilitates a better and more comprehensive understanding of data science from the altmetric perspective.

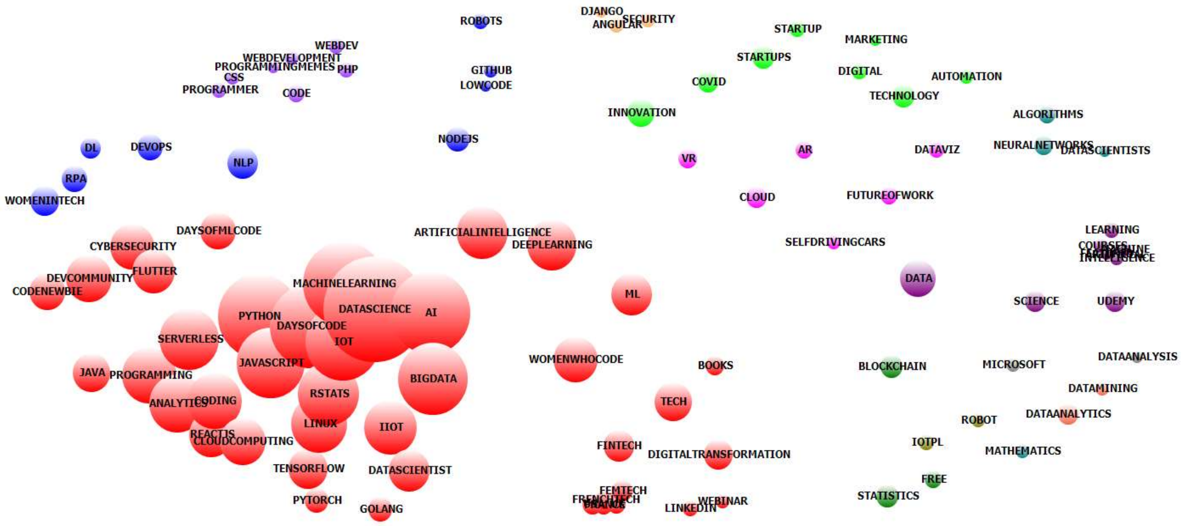


Figure 6.14 Intellectual structure of data science: a hashtag perspective

While the topical groupings with Twitter hashtags form far less meaningful groups, they provide insight into how data science is discussed in the social media environment, although less rigorously. Side by side with Twitter abbreviations is the appearance of many online communities utilizing hashtags of their own. The overwhelming message of these cooccurrence groupings is that nearly every group contains a hashtag that denotes a community; unsurprisingly, Twitter as a social network would foster such behavior. Topics in Figure 6.14 reflect a wider range of topics with less focus and perhaps of a great variety in terms of significance. Artificial intelligence is again represented strongly throughout the groups and its supporting topics: machine learning, deep learning, and natural language processing.

The hashtags like “machinelearning”, “AI”, and “bigdata” are immediately close to “datascience”, displaying how closely related these hashtags are. The most tightly packed aspect of Figure 6.14 is the red data science cluster that has a tremendous amount of artificial intelligence, as is to be expected, and contains a good amount of programming elements like #python, #rstats, #pytorch, #golang, #programming. Again, this suggests that much more of the Twitter conversation is focused on the application of techniques and tools in data science.

Decoding the hashtag use, the element of education comes through. The appearance of hashtags like #udemy, #courses, and #free all insinuate a search for knowledge and skill development. Coupled with learning-oriented communities like #daysofcode, #codenewbie, and #devcommunity, the case for data science education as a major topic of Twitter discourse is quite compelling. While other communities represented here may also serve as turnkeys for education and learning, these are the three focused on new programmers and those looking to learn.

Also more readily apparent here is the extensive discussion occurring around emerging technologies. These hot-topic items are the elements that get in the news, and so they would also seem to be some of the top choices for Twitter debate and dialogue. Augmented reality, virtual reality, robotics, blockchain, and self-driving cars are all emerging technology topics being discussed on Twitter. Each one represents a popular sector of technology and almost certainly technologies that will generate substantial data sets and employ data science to handle it all.

Additionally, the proximity of not just programming but also cyber security, it seems strange that security clusters are so close to the more core data science cluster when some aspects like #dataanalytics, #datamining, and #dataviz are quite far away. The shape and distribution of the topics suggest that overall discussion and topics appearing on Twitter have a very different underlying structure than those obtained via the examinations of scientometric and curricular data sets.

6.3.1.3 Content Analysis of the Top 100 Hashtags

Table 6.10 displays the subject groupings by performing a content analysis of the top 100 hashtags from another perspective. In contrast to factor/multivariate analysis of the top 100 hashtags, this approach attempted to group hashtags on their topic-relatedness above anything else with the aim to identify groupings of data science topics discussed on the Twittersphere.

Table 6.10 Content analysis of the top 100 hashtags

Subject	**Topic**	**Hashtag**
Data Science		data, datascience, datascientist
	Big Data	bigdata
	Visualization	dataviz
Artificial Intelligence		ai, artificial, artificialintelligence, intelligence
	Deep Learning	deeplearning, dl
	Machine Learning	daysofmlcode, machine, machinelearning, ml
	Natural Language Processing	nlp
	Neural Networks	neuralnetworks
	Tools	tensorflow, pytorch
Computer Science	Programming	code, coding, lowcode, programmer, programming, programmingmemes, webdev, webdevelopment
	Languages	css, golang, java, javascript, php, python, rstats
	Frameworks/Libraries	angular, django, flutter, nodejs, reactjs
	Communities	codenewbie, codenewbies, daysofcode,devcommunity, femtech, womenintech, womenwhocode
	Tools	github, linux
	Cloud Computing	cloud, cloudcomputing, serverless
Emerging Technologies	Internet of Things	iot, iiot, iotpl
	Robotics	robot, robotics, robots, selfdrivingcars
	Blockchain	blockchain
	Virtual Reality	Vr
	Augmented Reality	Ar
Data Analytics		analytics, dataanalysis, dataanalytics
	Data Mining	datamining
	Algorithms	algorithms
Business		business, devops, digitaltransformation, fintech, innovation, linkedin, marketing, rpa, startup, startups
Statistics		statistics
Security		cybersecurity, security
Education		udemy, books, learning, courses, webinar
Other	Countries	france, frenchtech, uk, usa
	Current Events	covid, futureofwork
	Uncategorized	featured, digital, free, tech, microsoft, g, science, technology

Categorizing the hashtags based on the implied topics makes it easier to identify topic groups, as presented in previous sections of this chapter. It should be noted that this categorization process is sometimes toucher than expected because the connotation of each hashtag is not always as clear. Though, Table 6.10 demonstrates that the subjects and topics as represented by hashtags overall are similar to those obtained from the scientometric and curricular examinations (see Section 6.1 and Section 6.2) despite some distinct differences.

In addition, Table 6.10 exemplifies the presence of artificial intelligence, business, and computer science in data science. It supports the Twitter discussions regarding data science education besides a sizeable amount of Twitter hashtags devoted to programming. Furthermore, Table 6.10 helps to bolster some of the topics that get overshadowed, like security and those found in the emerging technologies. Robotics, self-driving cars, augmented reality, and virtual reality are all topics derived from these emerging technologies hashtags that are not typically associated with data science but present potential implementations and data that data scientists could leverage. Business as a Twitter topic becomes easier to see, claiming many unique hashtags. These topics, lesser represented in the scientometric and altmetric parts of the current research, do get representation in the Twittersphere.

Lastly, the “Other” grouping comprised hashtags that did not fit easily into the other groups. Of note are the hashtags representing countries: USA, France, and the UK. While these hashtags may be more closely tied to the usage of Twitter than the association between countries and data science, they provide a context for the discussion as each of these countries are heavy Twitter user countries. Moreover, the USA and UK are both heavily invested in data science, so it makes sense for them to appear as hashtags and subsequently affect the discussions.

Examination of the Twittersphere’s discussion on data science to address the topics of research question five led to some fruitful take-aways. Firstly, it seems as though topics like artificial intelligence, computer science, and data analysis remain well represented, especially in comparison to the scientometric and curricular sections of this research. Hot topics like cloud computing and machine learning are still present. However, the nature of Tweets and their propensity to be far fuzzier in their connections between hashtags added a tremendous amount of hashtags that didn’t fit within the same rigid structure of academia or education. The result has been a perspective of hashtags more related to practitioners and those implementing data science in emerging technologies like blockchain or learning programming languages, or even those individuals examining data science tools and programming libraries. The expanse that the Twittersphere offers is both messy and more inclusive, and the benefit of that combination is a much broader view. Even with that messy view, it must be re-stated, it does seem as though core topics relating to data science translate to Twitter, and that is discussed further in Section 6.4.

6.4 Comparative Analysis of Results from the Three Research Methods

The final research question of this study is the comparison of the results obtained from the three unique research methods of this study: scientometrics, content analysis, and altmetrics. This analysis will compare the topics identified from the three kinds of analyses based on three different data sets. Tables 6.11, 6.12, and 6.13 present the three sets of topics obtained from the above analyses side by side, depending on whether a topic appears in one, two, or three kinds of results (i.e., scientometric, curricular, or altmetric results).

Specifically, Table 6.11 shows the topics yielded through all three research methods. Table 6.12 displays those topics appearing on two of the three lists, while Table 6.13 lists those topics identified via only one of the three kinds of results.

Topics that appeared precisely the same across lists were easily matched to a single term. In a few cases, topics were less exact. For example, the topics “data ethics” and “cyber security” were matched to one topic: “ethics, privacy, and security”. Similarly, “medicine and health” was also simplified down to “medicine”. Each of these simplifications was done only in situations where it appeared to this author that the intent and topic itself were close enough that they warranted interpretation for the sake of representation.

As indicated earlier, Table 6.11 represents the topics that occur across all three parts of this research. These thirteen topics represent a core set of research topics identified using the three different data sets.

Table 6.11 Core topics appearing on the scientometric, curricular and altmetric topic lists

Scientometric	**Curricular**	**Altmetric**
Artificial Intelligence	Artificial Intelligence	Artificial Intelligence
Big Data	Big Data	Big Data
Cloud Computing	Cloud Computing	Cloud Computing
Data Analytics	Data Analytics	Data Analytics
Data Mining	Data Mining	Data Mining
Data Science	Data Science	Data Science
Data Visualization	Data Visualization	Data Visualization
Deep Learning	Deep Learning	Deep Learning
Ethics, Privacy & Security	Ethics, Privacy & Security	Ethics, Privacy & Security
Machine Learning	Machine Learning	Machine Learning
Natural Language Processing	Natural Language Processing	Natural Language Processing
Neural Networks	Neural Networks	Neural Networks
Statistics	Statistics	Statistics

The importance of denoting which topics appeared across how many sections in Table 6.11 helps provide a more meaningful picture of the true core of data science. These topics are those that all three of these perspectives have not only agreed on but have also been represented by this research in expressive ways. The importance of this is perhaps best illustrated by Artificial Intelligence and its subtopics, which have appeared as exceptional standout topics throughout this study. Artificial intelligence, deep learning, machine learning, natural language processing, and neural networks all appeared in each section of this study. It unequivocally sets artificial intelligence as a core and primary topic within data science. In addition to artificial intelligence, data analytics, data mining, data visualization, and statistics were all core tenets of data science from an academic standpoint, and it was expected they retain that role through this study. Big data and its exceptional close historical and semantic ties to data science also made it to the top. Cloud computing and ethics, privacy, and security also appeared on all three lists. The appearance of ethics, privacy, and security reveals a growing refocusing on the proper use, storage, and protection of data and the corresponding systems. Cloud computing similarly is becoming a larger and larger component to data science work, especially in highly scalable environments.

Adding to the core list of topics are those topics that appeared across at least two of the sections of this research. Table 6.12 represents this secondary topic listing; the topics have been arranged to show which aspect of the research they were derived from.

Table 6.12 Topics appearing on two of the scientometric, curricular, and altmetric topic lists

Scientometric	**Curricular**	**Altmetric**
	Business	Business
Computer Science		Computer Science
Data-Driven Decision Making	Data-Driven Decision Making	
	Education	Education
Emerging Technologies		Emerging Technologies
Internet of Things		Internet of Things
Medicine	Medicine	
	Programming	Programming
Time Series Analysis	Time Series Analysis	
	Web Technologies	Web Technologies

Topics that appeared on two lists also created a list of solid data science topics. Computer science, data-driven decision making, databases, and time series analysis have classically been data science-associated topics. Less attributed to data science are emerging technologies and web technologies; however, both of these topics hint at future growth in data. Similarly, the topics of business, medicine, and the internet of things strongly relate to areas where data science is needed to handle the amount of data that is currently being generated and is expected to grow to even larger quantities. Business and medicine were seen as popular areas of study in the educational course and program design as fields already gearing up for data scientist needs. These topics represent those that are important to the field and are worth monitoring as they may be topics that are expanding into the core topics from Table 6.11.

Lastly, Table 6.13 is a collection of all the topics appearing in only one part of this study. Each of these topics represents a topic that helps understand and conceptualize data science but also did not appear strong enough across two of the three parts of this study. The nature of this interpretation also was also not to exclude these topics in any way but to give a voice to these topics as emerging or as being more represented in a significant way in some sections over others. These topics' relation to each other and even their potential ability to be simplified was considered and left structured this way to best demonstrate their unique representations in their respective data sets when compared to the other two.

Table 6.13 Topics appearing on only one of the scientometric, curricular, or altmetric topic list

Scientometric	**Curricular**	**Altmetric**
Data Deluge	Advanced Analytics	Augment Reality
Feature Selection	Algorithms	Autonomous Vehicles
Health Informatics	Communication	Blockchain
Information Systems	Computer Vision	Data Science Communities
Precision Medicine	Computing	Data Science Education
Sentiment Analysis	Data Structures	Data Science Networking
Smart Cities	Databases	Financial Technologies
Social Media	Geographic Information Systems	Industrial Internet of Things
Software Engineering	Information Science	Robotics
Statistical Learning	Research	Virtual Reality

Finally, when it comes to those topics that appeared on only one list, they are far more specific. Many of them are actual specific topics of emerging technologies, to point out a few: augmented reality, blockchain, robotics, and virtual reality. Comparably, several are techniques specific to a larger topic, like data structures and feature selection are for data analysis. These topics are potentially leading-edge technologies but are also those topics that may find more support in each of these data sets. In the case of research and data structures, these are topics that perhaps are more important to the education of future data scientists and so appear in curricular data where they just may never appear in Twitter discussions or academic publications in any sizable way. Similarly, Twitter topics push the boundaries of where data science is being used, and as a forum for people to discuss ideas and technologies, some of these conversations may be the machinations of ideas doomed to fail or disappear from the conversation in six months or a year. These topics, while leading topics in their own right, are still away from the core group seen in Section 6.11, and their destiny with data science is far less certain.

The final research question of this study was to determine what can be learned from the use of the three types of analysis used in this study. The comparison has revealed that there is a high degree of corroboration among this analysis and techniques. The revelation of thirteen core topics proves this strong data science identity that exists across forums. At the same time, the subsequent three-tiered representation of topics seen in Tables 6.11, 6.12, and 6.13 show that these methods do exhibit different topic topographies unique to their data. This, in its own regard, is a valuable finding in understanding that multiple perspectives of a field can yield a panoramic picture while still maintaining a strong central message. Data science and its conceptual understanding are not limited to one area of use, and the confluence of these varying identities undoubtedly influences each other. Setting each of these data science images side by side allows for some interesting insights and shall also shed light on the dynamics and development of the field in the years to come.

7. Conclusion and Future Research

This study was conducted because, in large part, data science has had a sparse history of being examined as a field even with its radical growth. The need for understanding the characteristics of data science to assist researchers, educators, students, and stakeholders is only growing. The current researcher thus has undertaken this task by using scientometric, curricular, and altmetric data and presented the findings of this study in Chapter 6. This chapter will expand on the implications of this research and outline its limitations and directions for future research.

7.1 Conclusions

This research reveals the intellectual structure underlying data science. The identification of the prime research topics and composition of data science can provide researchers, students, and educators with a valuable tool to help advance the field in their corresponding perspectives.

In the case of researchers, the scientometric examination of data science enables them to identify leading researchers, leading institutions, leading countries, and research fronts in the field. For both current and future researchers, the findings of this study will facilitate their further exploration of data science. For example, machine learning is emerging from the AI domain that has been making significant contributions to the development of data science. Additionally, while not as extensively represented as AI, privacy, security, and ethics are other examples of topics of importance in data science. While not as common as some topics, privacy, security, and ethics appear in classes, discussions, and research focusing on their importance.

Specific to education, this research portrays the common and unique courses in the curricula developed for bachelor's, master's, and Ph.D. programs across the United States. The common topics constitute the fundamentals that must be covered in data science education, while the unique topics represent those subjects that might shape the data science field in the near future. Specifically, those common courses among the curriculums this research examines are essential for the data science program to have a core while producing adequately prepared data scientists ready to traverse a rapidly growing field. The elective courses offered by different programs inspected in this study signify the emerging topics of the future curriculums in data science and help educate students that are competitive and forward-looking. Hence, the curricular findings of this study provide timely suggestions for updating and revising data science curriculums.

Furthermore, this study's Twitter data findings show that data science is perceived and discussed in the altmetric environment. Although altmetric data is much messier and less reliable than scientometric or curricular data, it still yields an interesting view of the data science field and highlights the topics that people in the Twitter-sphere discuss. For instance, security and the internet of things are among the many topics intensely discussed via Twitter. This view shows that participants in the Twittersphere have a discussion focus different from their counterparts in the research or education domains.

In addition, the triple-pronged approach of this research provides a panoramic view of the data science field using scientometric, curricular, and altmetric data. This plural methodology is novel in examining data science as no prior research has ever attempted to adopt two simultaneously, let alone three distinctly different research methods to study a discipline such as data science. In this sense, the present study sets a fine example for exploring a scholarly field holistically.

7.2 Limitations

This research provides a novel examination of data science; nevertheless, it has not come without limitations. A few distinctive elements of this research emerged as limitations of note and required addressing.

In particular, one unexpected limitation was a loss of access to the Scopus database. Amid this study's data collection process and completion, this researcher's affiliated library discontinued its subscription to Scopus, which served as the only means by which bibliometric data was being collected. The resulting loss of access made keeping the bibliometric data updated to its absolute latest difficult and created a barrier to re-extracting data to address export format issues.

In terms of the curricular data collection of this study, two limitations exist. The first major issue was that gathering course titles and descriptions was not trivial. How schools present their course information on their websites is inconsistent or unstructured, and how courses are outlined and stored is different from institution to institution. The second limitation was that, ultimately, this study examined curriculum data through course titles and course descriptions alone without any access to course syllabi or full curriculum descriptions. The latter handicaps the curriculum data collection and analysis process in this study as those data would have greatly helped the present researcher's understanding of data science curriculums.

Additionally, Twitter presents a unique challenge to both the scope and time of data collection. Tweets can only be reasonably collected historically if funding is available to buy records of past Tweets. The purchasing of past tweets, in fact, becomes extremely expensive and requires a great deal of specificity in search parameters. Hence, this study only collected Tweets in real-time, resulting in a data set that is a snapshot in time. Even though the data collection was done over 20 days, a more extensive cross-section of time could have yielded additional insights.

7.3 Recommendation for Future Research

The need for future research has become abundantly evident through this study. In the case of bibliometric data, continued research into the growth of data science research will always be a powerful addition to understanding this field as it will continue to change, develop, and reshape over the years. Continuous updating of the data science topology about research topics and major points of discussion and debate will continually provide insights for stakeholders across society. Moreover, future research will provide snapshots to help those who may look at the overall evolution and trajectory of data science as it passes through the years. The evolution of specific topics in this regard will also be an interesting and intricate examination of research themes over time.

In the case of curricular-based data, focusing on course titles and course descriptions to conceptualize deeper meaning from complete programs cannot fully understand the curriculums. Expansion beyond the use of only course titles and descriptions could provide additional insights into the curriculum. While at the same time, examination of data science beyond the scope of simply the United States could also give a greater perspective on data science as a global phenomenon.

Lastly, the part of the research using altmetric data seems only to touch on the surface of what social media could provide. Two characteristics that can be expanded upon would be the depth to which messages were categorized and the length of the study. This study focused on hashtags, while future studies could go beyond this. Each message and the interactions between users are massive and given the size and scope of social media platforms, users' interactions can become exceptionally complex. The mapping and composition of networks within these systems are rife with insights into how a greater public consumes data science.

Future research that is more longitudinally oriented will also be interesting to perform using altmetric data. This research looked at a small sub-section of time, but to chart that research over months or even years would be interesting for gaining insights on user interests and topic consumption for stakeholders and their networks. The mutability of conversations occurring on Twitter is highly variable. As such, data collection over time could show changes over time and help home in on long-standing conversations that may appear over more prolonged periods.

References

Adie, Euan, and William Roe. 2013. “Altmetric: Enriching Scholarly Content with Article-Level Discussion and Metrics.” *Learned Publishing* 26 (1): 11–17. https://doi.org/10.1087/20130103.

Agarwal, R, and V Dhar. 2014. “Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research.” *Information Systems Research* 25 (3): 443–48. https://doi.org/10.1287/isre.2014.0546.

Anson, Ian I. 2016. “Scientometric Analysis of the Emerging Technology Landscape.” *Qualitative and Quantitative Methods in Libraries* 5: 1–10.

Arroyo-Machado, Wenceslao, Daniel Torres-Salinas, Enrique Herrera-Viedma, and Esteban Romero-Frías. 2020. “Science through Wikipedia: A Novel Representation of Open Knowledge through Co-Citation Networks.” *PLoS ONE* 15 (2): 1–21. https://doi.org/10.1371/journal.pone.0228713.

Ausserhofer, Julian, and Axel Maireder. 2013. “NATIONAL POLITICS ON TWITTER: Structures and Topics of a Networked Public Sphere.” *Information Communication and Society* 16 (3): 291–314. https://doi.org/10.1080/1369118X.2012.756050.

Azaria, Asaph, Ariel Ekblaw, Thiago Vieira, and Andrew Lippman. 2016. “MedRec: Using Blockchain for Medical Data Access and Permission Management.” *Proceedings - 2016 2nd International Conference on Open and Big Data, OBD 2016*, 25–30. https://doi.org/10.1109/OBD.2016.11.

Bar-Ilan, Judit, Cassidy R. Sugimoto, William Gunn, Stefanie Haustein, Stacy Konkiel, Vincent Larivière, and Jennifer Lin. 2013. “Altmetrics: Present and Future – Panel.” *ASIST 2013 Annual Meeting*, 4. https://doi.org/10.1002/meet.14505001013.

Baumer, Ben. 2015. “A Data Science Course for Undergraduates: Thinking With Data.” *American Statistician* 69 (4): 334–42. https://doi.org/10.1080/00031305.2015.1081105.

Behpour, Sahar, Suliman Hawamdeh, and Abbas Gourarzi. 2019. “Employer’s Perspective on Data Science; Analysis of Job Requirement & Course Description.” *ALISE 2019 Proceedings*, 177–82.

Bhattacharya, Sujit, and Shubham Singh. 2020. “Visible Insights of the Invisible Pandemic: A Scientometric, Altmetric and Topic Trend Analysis.” *Journal of Petrology* 369 (1): 1689–99. http://arxiv.org/abs/2004.10878.

Biljecki, Filip. 2016. “A Scientometric Analysis of Selected GIScience Journals.” *International Journal of Geographical Information Science* 30 (7): 1302–35. https://doi.org/10.1080/13658816.2015.1130831.

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. “Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2.” *Nature Biotechnology* 37 (8): 852–57. https://doi.org/10.1038/s41587-019-0209-9.

Bornmann, Lutz, and Robin Haunschild. 2016. “Overlay Maps Based on Mendeley Data: The Use of Altmetrics for Readership Networks.” *Journal of the Association for Information Science and Technology* 67 (12): 3064–72. https://doi.org/10.1002/asi.23569.

Bornmann, Lutz. 2014. “Do Altmetrics Point to the Broader Impact of Research? An Overview of Benefits and Disadvantages of Altmetrics.” *Journal of Informetrics* 8 (4): 1–24. https://doi.org/http://dx.doi.org/10.1016/j.joi.2014.09.005.

Bornmann, Lutz. 2015. “Alternative Metrics in Scientometrics: A Meta-Analysis of Research into Three Altmetrics.” *Scientometrics* 103 (3): 1123–44. https://doi.org/10.1007/s11192-015-1565-y.

Boss, Katherine, and Emily Drabinski. 2014. “Evidence-Based Instruction Integration: A Syllabus Analysis Project.” *Reference Services Review* 42 (2): 263–76. https://doi.org/10.1108/RSR-07-2013-0038.

Bucheli, Victor, Adriana Díaz, Juan Pablo Calderón, Pablo Lemoine, Juan Alejandro Valdivia, José Luis Villaveces, and Roberto Zarama. 2012. “Growth of Scientific Production in Colombian Universities: An Intellectual Capital-Based Approach.” *Scientometrics* 91 (2): 369–82. https://doi.org/10.1007/s11192-012-0627-7.

Callison, Daniel, and Carol L. Tilley. 2001. “Descriptive Impressions of the Library and Information Education Evolution of 1988-1998 as Reflected in Job Announcements, ALISE Descriptors, and New Course Titles.” *Journal of Education for Library and Information Science* 42 (3): 181. https://doi.org/10.2307/40324010.

Callon, M., J. P. Courtial, and F. Laville. 1991. “Co-Word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemsitry.” *Scientometrics* 22 (1): 155–205. https://doi.org/10.1007/BF02019280.

Callon, M., J.-P. Courtial, W. A. Turner, and S. Bauin. 1983. “From Translations to Problematic Networks: An Introduction to Co-Word Analysis.” *Social Science Information* 22 (2): 191–235. https://doi.org/10.1177/053901883022002003.

Cao, Longbing. 2017. “Data Science: A Comprehensive Overview.” *ACM Computing Surveys* 50 (3): 1–42. https://doi.org/10.1145/3076253.

Chae, Bongsug. 2015. “Insights from Hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter Data for Supply Chain Practice and Research.” *International Journal of Production Economics* 165: 247–59. https://doi.org/10.1016/j.ijpe.2014.12.037.

Chen, C.L. Philip, and Chun Yang Zhang. 2014. “Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data.” *Information Sciences* 275: 314–47. https://doi.org/10.1016/j.ins.2014.01.015.

Chen, Yong, Hong Chen, Anjee Gorkhali, Yang Lu, Yiqian Ma, and Ling Li. 2016. “Big Data Analytics and Big Data Science: A Survey.” *Journal of Management Analytics* 3 (1): 1–42. https://doi.org/10.1080/23270012.2016.1141332.

Chu, Heting. 2006. “Curricula of Lis Programs in the Usa : A Content Analysis.” *Proceedings of Asia-Pacific Conference on Library & Information Education & Practice*, 328–37.

Coccia, Mario. 2018. “General Properties of the Evolution of Research Fields: A Scientometric Study of Human Microbiome, Evolutionary Robotics and Astrobiology.” *Scientometrics* 117 (2): 1265–83. https://doi.org/10.1007/s11192-018-2902-8.

Correia, António, Hugo Paredes, and Benjamim Fonseca. 2018. “Scientometric Analysis of Scientific Publications in CSCW.” *Scientometrics* 114 (1): 31–89. https://doi.org/10.1007/s11192-017-2562-0.

Costas, Rodrigo, Zohreh Zahedi, and Paul Wouters. 2015. “Do ‘Altmetrics’ Correlate with Citations? Extensive Comparison of Altmetric Indicators with Citations from a Multidisciplinary Perspective.” *Journal of the Association for Information Science and Technology* 66 (10): 2003–19. https://doi.org/10.1002/asi.23309.

Coulter, Neal. 1998. “Software Engineering as Seen through Its Research Literature: A Study in Co-Word Analysis.” *Journal of the American Society for Information Science* 49 (13): 1206–23. https://doi.org/10.1002/(sici)1097-4571(1998)49:13<1206::aid-asi7>3.3.co;2-6.

Crane, Diana. 1972. *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago: The University of Chicago Press.

Dastidar, Prabir G., and S. Ramachandran. 2008. “Intellectual Structure of Antarctic Science: A 25-Years Analysis.” *Scientometrics* 77 (3): 389–414. https://doi.org/10.1007/s11192-007-1947-x.

Ding, Ying, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. 2014. “Content-Based Citation Analysis: The next Generation of Citation Analysis.” *Journal of the Association for Information Science and Technology* 65 (9): 1820–33. https://doi.org/10.1002/asi.23256.

Donoho, David. 2017. “50 Years of Data Science.” *Journal of Computational and Graphical Statistics* 26 (4): 745–66. https://doi.org/10.1080/10618600.2017.1384734.

Drisko, James W, and Tina Maschi. 2015. *Content Analysis*. Pocket Guides to Social Work R.

Dumbill, Edd. 2012. “What Is Apache Hadoop?” *O’Reilly*. http://strata.oreilly.com/2012/02/what-is-apache-hadoop.html.

Elsevier. 2020. “Content Coverage Guide,” 1–24. https://www.elsevier.com/__data/assets/pdf_file/0017/114533/Scopus_GlobalResearch_Factsheet2019_FINAL_WEB.pdf.

Emmert-Streib, Frank, and Matthias Dehmer. 2018. “Defining Data Science by a Data-Driven Quantification of the Community.” *Machine Learning and Knowledge Extraction* 1 (1): 235–51. https://doi.org/10.3390/make1010015.

Fatt, Choong Kwai, Ephrance Abu Ujum, and Kuru Ratnavelu. 2010. “The Structure of Collaboration in the Journal of Finance.” *Scientometrics* 85 (3): 849–60. https://doi.org/10.1007/s11192-010-0254-0.

Galligan, Finbar, and Sharon Dyas-Correia. 2013. “Altmetrics: Rethinking the Way We Measure.” *Serials Review* 39 (1): 56–61. https://doi.org/10.1016/j.serrev.2013.01.003.

Garfield, Eugene, and I H Sher. 1963. “New Factors in the Evaluation of Scientific Literature through Citation Indexing.” *American Documentation* 14 (3): 195–201. https://doi.org/10.1002/asi.5090140304.

Garfield, Eugene. 1979. *Citation Indexing - Its Theory and Application in Science, Technology, and Humanities*. New York: John Wiley & Sons.

Garg, K. C., and S. Kumar. 2016. “Scientometric Profile of an Indian State: The Case Study of Odisha.” *Collnet Journal of Scientometrics and Information Management* 10 (1): 141–53. https://doi.org/10.1080/09737766.2016.1177950.

Godin, Benoît. 2006. “On the Origins of Bibliometrics.” *Scientometrics* 68 (1): 109–33. https://doi.org/10.1007/s11192-006-0086-0.

Google Trends. Accessed July 02, 2018. https://trends.google.com/trends/?geo=US.

Griffith, Shannon M., Melanie M. Domenech Rodríguez, and Austin J. Anderson. 2014. “Graduate Ethics Education: A Content Analysis of Syllabi.” *Training and Education in Professional Psychology* 8 (4): 248–52. https://doi.org/10.1037/tep0000036.

Gu, Dongxiao, Jingjing Li, Xingguo Li, and Changyong Liang. 2017. “Visualizing the Knowledge Structure and Evolution of Big Data Research in Healthcare Informatics.” *International Journal of Medical Informatics* 98: 22–32. https://doi.org/10.1016/j.ijmedinf.2016.11.006.

Gupta, B. M., Avinash Kshitij, and Charu Verma. 2011. “Mapping of Indian Computer Science Research Output, 1999–2008.” *Scientometrics* 86 (2): 261–83. https://doi.org/10.1007/s11192-010-0272-y.

Halevi, Gali, and Henk F. Moed. 2012. “The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature.” *Research Trends* 1969 (30): 3–6. http://www.researchtrends.com/issue-30-september-2012/the-evolution-of-big-data-as-a-research-and-scientific-topic-overview-of-the-literature/.

Han, W T. 2017. “A Current Situation Analysis of Data Science-Related Programs in North American ISchools in the Big Data Age.” *IConference-Workshop*, 3–5.

Harris-Pierce, Rebecca L., and Yan Quan Liu. 2012. “Is Data Curation Education at Library and Information Science Schools in North America Adequate?” *New Library World* 113 (11): 598–613. https://doi.org/10.1108/03074801211282957.

Haustein, Stefanie, Isabella Peters, Judit Bar-Ilan, Jason Priem, Hadas Shema, and Jens Terliesner. 2013. “Coverage and Adoption of Altmetrics Sources in the Bibliometric Community.” *Scientometrics*, 1–19. https://doi.org/10.1007/s11192-013-1221-3.

Herrera, Mark, David C. Roberts, and Natali Gulbahce. 2010. “Mapping the Evolution of Scientific Fields.” *PLoS ONE* 5 (5): 3–8. https://doi.org/10.1371/journal.pone.0010355.

Hou, Haiyan, Hildrun Kretschmer, and Zeyuan Liu. 2008. “The Structure of Scientific Collaboration Networks in Scientometrics.” *Scientometrics* 75 (2): 189–202. https://doi.org/10.1007/s11192-007-1771-3.

Hou, Jianhua, Xiucai Yang, and Chaomei Chen. 2018. “Emerging Trends and New Developments in Information Science: A Document Co-Citation Analysis (2009–2016).” *Scientometrics* 115 (2): 869–92. https://doi.org/10.1007/s11192-018-2695-9.

Huang, Ying, Jannik Schuehle, Alan L. Porter, and Jan Youtie. 2015. “A Systematic Method to Create Search Strategies for Emerging Technologies Based on the Web of Science: Illustrated for ‘Big Data.’” *Scientometrics* 105 (3): 2005–22. https://doi.org/10.1007/s11192-015-1638-y.

Irwin, Ray. 2002. “Characterizing the Core: What Catalog Descriptions of Mandatory Courses Reveal about LIS Schools and Librarianship.” *Journal of Education for Library and Information Science* 43 (2): 175. https://doi.org/10.2307/40323978.

Jacso, Peter. 2018. “The Scientometric Portrait of Eugene Garfield through the Free ResearcherID Service from the Web of Science Core Collection of 67 Million Master Records and 1.3 Billion References.” *Scientometrics* 114 (2): 545–55. https://doi.org/10.1007/s11192-017-2624-3.

Janssens, Frizo, Jacqueline Leta, Wolfgang Glanzel, and Bart de Moor. 2006. “Towards Mapping Library and Information Science.” *Information Processing & Management* 42 (6): 1614–42. https://doi.org/10.1016/j.ipm.2006.03.025.

Jin, Xiaolong, Benjamin W. Wah, Xueqi Cheng, and Yuanzhuo Wang. 2015. “Significance and Challenges of Big Data Research.” *Big Data Research* 2 (2): 59–64. https://doi.org/10.1016/j.bdr.2015.01.006.

Johnston, Ron, and Dave Robins. 1977. “The Development of Specialties in Industrialised Science.” *The Sociology Review*.

Jordan, M. I., and T. M. Mitchell. 2015. “Machine Learning: Trends, Perspectives, and Prospects.” *Science* 349 (6245): 255–60. https://doi.org/10.1126/science.aaa8415.

Kalantari, Ali, Amirrudin Kamsin, Halim Shukri Kamaruddin, Nader Ale Ebrahim, Abdullah Gani, Ali Ebrahimi, and Shahaboddin Shamshirband. 2017. “A Bibliometric Approach to Tracking Big Data Research Trends.” *Journal of Big Data* 4 (1): 1–18. https://doi.org/10.1186/s40537-017-0088-1.

Karpagam, R, S Gopalakrishnan, B Ramesh Babu, M Natarajan, Puthiya Parvai, and Tamil Arasi Publications. 2012. “Scientometric Analysis of Stem Cell Research : A Comparative Study of India and Other Countries,” no. December: 1–24.

Kim, Erin Hea Jin, Yoo Kyung Jeong, Yuyoung Kim, Keun Young Kang, and Min Song. 2015. “Topic-Based Content and Sentiment Analysis of Ebola Virus on Twitter and in the News.” *Journal of Information Science* 42 (6): 763–81. https://doi.org/10.1177/0165551515608733.

Kim, Ha Jin, Yoo Kyung Jeong, and Min Song. 2016. “Content- and Proximity-Based Author Co-Citation Analysis Using Citation Sentences.” *Journal of Informetrics* 10 (4): 954–66. https://doi.org/10.1016/j.joi.2016.07.007.

Kim, Hyunjung. 2017. “A Study on the Intellectual Structure of Data Science Using Co-Word Analysis.” *Journal of the Korean Society for Information Management* 34 (4): 101–26. https://doi.org/KOSIM.2017.34.4.101.

Kim, Jinseok, Jenna Kim, and Jason Owen-Smith. 2021. “Ethnicity-Based Name Partitioning for Author Name Disambiguation Using Supervised Machine Learning.” *Journal of the Association for Information Science and Technology*, no. April 2020: 1–16. https://doi.org/10.1002/asi.24459.

Kolahi, J., P. Iranmanesh, and S. Khazaei. 2017. “Altmetric Analysis of 2015 Dental Literature: A Cross Sectional Survey.” *British Dental Journal* 222 (9): 695–99. https://doi.org/10.1038/sj.bdj.2017.408.

Kolahi, Jafar, Saber Khazaei, Pedram Iranmanesh, and Parisa Soltani. 2019. “Analysis of Highly Tweeted Dental Journals and Articles: A Science Mapping Approach.” *British Dental Journal* 226 (9): 673–78. https://doi.org/10.1038/s41415-019-0212-z.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

Kumar, Suresh, and K. C. Garg. 2005. "Scientometrics of Computer Science Research in India and China." *Scientometrics* 64 (2): 121–32. https://doi.org/10.1007/s11192-005-0244-9.

Lee, Mi Kyung, Ho Young Yoon, Marc Smith, Hye Jin Park, and Han Woo Park. 2017. "Mapping a Twitter Scholarly Communication Network: A Case of the Association of Internet Researchers' Conference." *Scientometrics* 112 (2): 767–97. https://doi.org/10.1007/s11192-017-2413-z.

Lin, Jennifer, and Martin Fenner. 2013. "Altmetrics in Evolution: Defining and Redefining the Ontology of Article-Level Metrics." *Information Standards Quarterly* 25 (2): 20. https://doi.org/10.3789/isqv25no2.2013.04.

Liu, Jialu, Kin Hou Lei, Jeffery Yufei Liu, Chi Wang, and Jiawei Han. 2013. "Ranking-Based Name Matching for Author Disambiguation in Bibliographic Data." *Proceedings of the 2013 KDD Cup 2013 Workshop*. https://doi.org/10.1145/2517288.2517296.

Liu, Ping, Bao-li Chen, Kan Liu, and Hao Xie. 2017. "Magnetic Nanoparticles Research : A Scientometric Analysis of Development Trends and Research Fronts." *Scientometrics* 108 (3): 1591–1602. https://doi.org/10.1007/s11192-016-2017-z.

Liu, Zhigao, Yimei Yin, Weidong Liu, and Michael Dunford. 2015. "Visualizing the Intellectual Structure and Evolution of Innovation Systems Research: A Bibliometric Analysis." *Scientometrics*, 135–58. https://doi.org/10.1007/s11192-014-1517-y.

Lyu, Xiaozan, and Rodrigo Costas. 2020. "How Do Academic Topics Shift across Altmetric Sources? A Case Study of the Research Area of Big Data." *Scientometrics* 123 (2): 909–43. https://doi.org/10.1007/s11192-020-03415-7.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2011. "Big Data: The next Frontier for Innovation, Competition, and Productivity." *McKinsey Global Institute*, no. June: 156. https://doi.org/10.1080/01443610903114527.

Marchionini, Gary, and Gary Marchionini. 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.

Martins Pereira, Sandra, and Pablo Hernández-Marrero. 2016. "Palliative Care Nursing Education Features More Prominently in 2015 than 2005: Results from a Nationwide Survey and Qualitative Analysis of Curricula." *Palliative Medicine* 30 (9): 884–88. https://doi.org/10.1177/0269216316639794.

McGibbon, Robert T., Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee Ping Wang, Thomas J. Lane, and Vijay S. Pande. 2015. "MDTraj: A Modern Open Library for the Analysis of Molecular

Dynamics Trajectories." *Biophysical Journal* 109 (8): 1528–32. https://doi.org/10.1016/j.bpj.2015.08.015.

Mingers, John, and Loet Leydesdorff. 2015. "A Review of Theory and Practice in Scientometrics A Review of Theory and Practice in Scientometrics 1." *European Journal of Operational Research*, no. 1934: 1–47. https://doi.org/10.1016/j.ejor.2015.04.002.

Mokhtari, Heidar, Nima Soltani-Nejad, Seyedeh Zahra Mirezati, and Mohammad Karim Saberi. 2020. "A Bibliometric and Altmetric Analysis of Anatolia: 1997–2018." *Anatolia* 31 (3): 406–22. https://doi.org/10.1080/13032917.2020.1740285.

Müller, Mark Christoph, Florian Reitz, and Nicolas Roy. 2017. "Data Sets for Author Name Disambiguation: An Empirical Analysis and a New Resource." *Scientometrics* 111 (3): 1467–1500. https://doi.org/10.1007/s11192-017-2363-5.

Najafabadi, Maryam M., Flavio Villanustre, Taghi M. Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. 2015. "Deep Learning Applications and Challenges in Big Data Analytics." *Journal of Big Data* 2 (1): 1–21. https://doi.org/10.1186/s40537-014-0007-7.

Narin, F., D. Olivastro, and K. a. Stevens. 1994. "Bibliometrics/Theory, Practice and Problems." *Evaluation Review* 18 (1): 65–76. https://doi.org/10.1177/0193841X9401800107.

Neuendorf, Kimberly A. 2020. *Defining Content Analysis*. *The Content Analysis Guidebook*. https://doi.org/10.4135/9781071802878.n1.

Ofli, Ferda, Patrick Meier, Muhammad Imran, Carlos Castillo, Devis Tuia, Nicolas Rey, Julien Briant, et al. 2016. "Combining Human Computing and Machine Learning to Make Sense of Big (Aerial) Data for Disaster Response." *Big Data* 4 (1): 47–59. https://doi.org/10.1089/big.2014.0064.

Olhede, Sofia C., and Patrick J. Wolfe. 2018. "The Future of Statistics and Data Science." *Statistics and Probability Letters*. https://doi.org/10.1016/j.spl.2018.02.042.

Olijnyk, Nicholas Victor. 2014. "Information Security: A Scientometric Study of the Profile, Structure, and Dynamics of an Emerging Scholarly Specialty."

Ortega, José Luis. 2015. "Relationship between Altmetric and Bibliometric Indicators across Academic Social Sites: The Case of CSIC's Members." *Journal of Informetrics* 9 (1): 39–49. https://doi.org/10.1016/j.joi.2014.11.004.

Ortiz-Repiso, Virginia, Jane Greenberg, and Javier Calzada-Prado. 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): 768–84. https://doi.org/10.1177/0165551517748149.

Papi, Anita. 2018. “Big Data and Data Science : A Scientometrics Approach,” no. May: 233–40.

Park, Han Woo, and Loet Leydesdorff. 2013. “Decomposing Social and Semantic Networks in Emerging ‘Big Data’ Research.” *Journal of Informetrics* 7 (3): 756–65. https://doi.org/10.1016/j.joi.2013.05.004.

Park, Hyo Chan, Jonghee M. Youn, and Han Woo Park. 2018. “Global Mapping of Scientific Information Exchange Using Altmetric Data.” *Quality and Quantity* 53 (2): 935–55. https://doi.org/10.1007/s11135-018-0797-3.

Perron, B. E., B. G. Victor, D. R. Hodge, C. P. Salas-Wright, M. G. Vaughn, and R. J. Taylor. 2016. “Laying the Foundations for Scientometric Research: A Data Science Approach.” *Research on Social Work Practice*, no. October. https://doi.org/10.1177/1049731515624966.

Piwowar, Heather. 2013. “Altmetrics: What, Why and Where?” *Bulletin of the American Society for Information Science and Technology* 39 (4): 8–9. https://doi.org/10.1002/bult.2013.1720390404.

Ponzi, Leonard J. 2002. “The Intellectual Structure and Interdisciplinary Breadth of Knowledge Management: A Bibliometric Study of Its Early Stage of Development.” *Scientometrics* 55 (2): 259–72. https://doi.org/10.1023/A:1019619824850.

Prakash, M, and J Arumugam. 2019. “Scientometric Mapping of Data Science Research: A Global Perspective.” *International Conference on Enhancement of Technology and Innovations in Contemporary Libraries*, no. October: 243–51.

Pratt, Jean a., Karina Hauser, and Cassidy R. Sugimoto. 2012. “Defining the Intellectual Structure of Information Systems and Related College of Business Disciplines: A Bibliometric Analysis.” *Scientometrics* 93 (2): 279–304. https://doi.org/10.1007/s11192-012-0668-y.

Price, Derek J. 1965. “Networks of Scientific Papers.” *Science* 149 (3683): 510–15. https://doi.org/10.1126/science.149.3683.510.

Provost, Foster, and Tom Fawcett. "Data science and its relationship to big data and data-driven decision making." *Big data* 1, no. 1 (2013): 51-59.

Provost, Foster, and Tom Fawcett. 2013. “Data Science and Its Relationship to Big Data and Data-Driven Decision Making.” *Big Data* 1 (1): 51–59. https://doi.org/10.1089/big.2013.1508.

Raban, Daphne R., and Avishag Gordon. 2020. “The Evolution of Data Science and Big Data Research: A Bibliometric Analysis.” *Scientometrics* 122 (3): 1563–81. https://doi.org/10.1007/s11192-020-03371-2.

Ramin, Shahrokh, Mohammad Pakravan, Gholamreza Habibi, and Roghayeh Ghazavi. 2016. "Scientometric Analysis and Mapping of 20 Years of Glaucoma Research." *International Journal of Ophthalmology* 9 (9): 1329–35. https://doi.org/10.18240/ijo.2016.09.17.

Ravikumar, S., Ashutosh Agrahari, and S. N. Singh. 2014. "Mapping the Intellectual Structure of Scientometrics: A Co-Word Analysis of the Journal Scientometrics (2005–2010)." *Scientometrics* 102 (1): 929–55. https://doi.org/10.1007/s11192-014-1402-8.

Robinson-Garcia, Nicolas, Wenceslao Arroyo-Machado, and Daniel Torres-Salinas. 2019. "Mapping Social Media Attention in Microbiology: Identifying Main Topics and Actors." *FEMS Microbiology Letters* 366 (7): 1–8. https://doi.org/10.1093/femsle/fnz075.

Romo-Fernández, Luz M., Vicente P. Guerrero-Bote, and Félix Moya-Anegón. 2013. "Co-Word Based Thematic Analysis of Renewable Energy (1990-2010)." *Scientometrics* 97 (3): 743–65. https://doi.org/10.1007/s11192-013-1009-5.

Rotella, Perry. 2012. "Is Data The New Oil?," no. Letzter Zugriff 05.02.2014. http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/.

Salzano, Francisco M. 2018. "The Evolution of Science in a Latin-American Country: Genetics and Genomics in Brazil." *Genetics* 208 (3): 823–32. https://doi.org/10.1534/genetics.118.300690.

Santa Soriano, Alba, Carolina Lorenzo Álvarez, and Rosa María Torres Valdés. 2018. "Bibliometric Analysis to Identify an Emerging Research Area: Public Relations Intelligence—a Challenge to Strengthen Technological Observatories in the Network Society." *Scientometrics*, 1–24. https://doi.org/10.1007/s11192-018-2651-8.

Sarkar, Arindam, and Ashok Pal. 2019. "Where Does Data Science Research Stand in the 21st Century: Observation from the Standpoint of a Scientometric Analysis." *Library Philosophy and Practice* 2019: 0–2.

Schreuder, Martijn, Angela Riccio, Monica Risetti, Sven Dähne, Andrew Ramsay, John Williamson, Donatella Mattia, and Michael Tangermann. 2013. "User-Centered Design in Brain-Computer Interfaces-a Case Study." *Artificial Intelligence in Medicine* 59 (2): 71–80. https://doi.org/10.1016/j.artmed.2013.07.005.

Seawright, Jason. 2016. "Better Multimethod Design: The Promise of Integrative Multimethod Research." *Security Studies* 25 (1): 42–49. https://doi.org/10.1080/09636412.2016.1134187.

Si, Li, Xiaozhe Zhuang, Wenming Xing, and Weining Guo. 2013. "The Cultivation of Scientific Data Specialists: Development of LIS Education Oriented to e-Science Service Requirements." *Library Hi Tech* 31 (4): 700–724. https://doi.org/10.1108/LHT-06-2013-0070.

Singh, Vivek Kumar, Sumit Kumar Banshal, Khushboo Singhal, and Ashraf Uddin. 2015. “Scientometric Mapping of Research on ‘Big Data.’” *Scientometrics* 105 (2): 727–41. https://doi.org/10.1007/s11192-015-1729-9.

Small, Henry. 1973. “Co-Citation in the Scientific Literature: A New Measure of the Relationship between Two Documents.” *Journal of the American Society for Information Science* 24 (4): 265–69. https://doi.org/10.1002/asi.4630240406.

Song, Il Yeol, and Yongjun Zhu. 2016. “Big Data and Data Science: What Should We Teach?” *Expert Systems* 33 (4): 364–73. https://doi.org/10.1111/exsy.12130.

Stephens, Zachary D., Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. 2015. “Big Data: Astronomical or Genomical?” *PLoS Biology* 13 (7): 1–11. https://doi.org/10.1371/journal.pbio.1002195.

Sugimoto, Cassidy R., Sam Work, Vincent Larivière, and Stefanie Haustein. 2017. “Scholarly Use of Social Media and Altmetrics: A Review of the Literature.” *Journal of the Association for Information Science and Technology* 68 (9): 2037–62. https://doi.org/10.1002/asi.23833.

Swan, Melanie. 2013. “The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery.” *Big Data* 1 (2): 85–99. https://doi.org/10.1089/big.2012.0002.

Tambe, Prasanna. 2014. “Big Data Investment, Skills and Firm Value.” *Management Science* 60 (6): 1452–69.

Tang, Rong, and Watinee Sae-Lim. 2016. “Data Science Programs in U.S. Higher Education: An Exploratory Content Analysis of Program Description, Curriculum Structure, and Course Focus.” *Education for Information* 32 (3): 269–90. https://doi.org/10.3233/EFI-160977.

Teixeira, Aurora A C, and Elsa Ferreira. 2013. “Intellectual Structure of the Entrepreneurship Field: A Tale Based on Three Core Journals.” *Journal of Innovation Management Teixeira* 1: 21–66. http://www.open-jim.orghttp//creativecommons.org/licenses/by/3.0.

Tett, Gillian. 2017. “Trump, Cambridge Analytica and How Big Data Is Reshaping Politics.” *Financial Times*, 9–12. https://www.ft.com/content/e66232e4-a30e-11e7-9e4f-7f5e6a7c98a2.

Thelwall, Mike, Stefanie Haustein, Vincent Larivière, and Cassidy R. Sugimoto. 2013. “Do Altmetrics Work? Twitter and Ten Other Social Web Services.” *PLoS ONE* 8 (5): 1–8. https://doi.org/10.1371/journal.pone.0064841.

Thelwall, Mike. 2016. “Data Science Altmetrics.” *Journal of Data and Information Science* 1 (2): 7–12. https://doi.org/10.20309/jdis.201610.

Uddin, Ashraf, and Vivek Kumar Singh. 2014. “Mapping the Computer Science Research in SAARC Countries.” *IETE Technical Review* 31 (4): 287–96. https://doi.org/10.1080/02564602.2014.947527.

Uddin, Shahadat, and Arif Khan. 2016. “The Impact of Author-Selected Keywords on Citation Counts.” *Journal of Informetrics* 10 (4): 1166–77. https://doi.org/10.1016/j.joi.2016.10.004.

Uddin, Shahadat, Arif Khan, and Louise A. Baur. 2015. “A Framework to Explore the Knowledge Structure of Multidisciplinary Research Fields.” *PLoS ONE* 10 (4). https://doi.org/10.1371/journal.pone.0123537.

Van der Aalst, Wil MP. 2016. *Process mining: data science in action*. Springer.

Varvel, Virgil E., Elin J. Bammerlin, and Carole L. Palmer. 2012. “Education for Data Professionals: A Study of Current Courses and Programs.” *ACM International Conference Proceeding Series*, 527–29. https://doi.org/10.1145/2132176.2132275.

Vinkler, P. 2010. “Indicators Are the Essence of Scientometrics and Bibliometrics.” *Scientometrics* 85 (3): 861–66. https://doi.org/10.1007/s11192-010-0159-y.

Wainer, Jacques, and Paula Vieira. 2013. “Correlations between Bibliometrics and Peer Evaluation for All Disciplines: The Evaluation of Brazilian Scientists.” *Scientometrics* 96 (2): 395–410. https://doi.org/10.1007/s11192-013-0969-9.

Waller, Matthew a, and Stanley E Fawcett. 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.

Waller, Matthew A, and Stanley E Fawcett. 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.

Washington Durr, Angel Krystina. 2020. “A Text Analysis of Data-Science Career Opportunities and US ISchool Curriculum.” *Journal of Education for Library and Information Science* 61 (2): 270–93. https://doi.org/10.3138/jelis.2018-0067.

White, S.M. 2005. “Improving the System/Software Engineering Interface for Complex System Development.” In *12th IEEE International Conference and Workshops on the Engineering of Computer-Based Systems (ECBS’05)*, 281–88. IEEE. https://doi.org/10.1109/ECBS.2005.45.

Whittaker, John. 1989. “Creativity and Conformity in Science: Titles, Keywords and Co-Word Analysis.” *Social Studies of Science* 19 (3): 473–96. https://doi.org/10.1177/030631289019003004.

Wu, Zhaohui, and Ooi Beng Chin. 2014. “From Big Data to Data Science: A Multi-Disciplinary Perspective.” *Big Data Research* 1: 1. https://doi.org/10.1016/j.bdr.2014.08.002.

Xie, Ping. 2015. “Study of International Anticancer Research Trends via Co-Word and Document Co-Citation Visualization Analysis.” *Scientometrics* 105 (1): 611–22. https://doi.org/10.1007/s11192-015-1689-0.

Yan, Bei-Ni, Tian-Shyug Lee, and Tsung-Pei Lee. 2015. “Mapping the Intellectual Structure of the Internet of Things (IoT) Field (2000–2014): A Co-Word Analysis.” *Scientometrics*. https://doi.org/10.1007/s11192-015-1740-1.

Yu, Houqiang, Tingting Xiao, Shenmeng Xu, and Yuefen Wang. 2019. “Who Posts Scientific Tweets? An Investigation into the Productivity, Locations, and Identities of Scientific Tweeters.” *Journal of Informetrics* 13 (3): 841–55. https://doi.org/10.1016/j.joi.2019.08.001.

Zahedi, Zohreh, Rodrigo Costas, and Paul Wouters. 2014. “How Well Developed Are Altmetrics? A Cross-Disciplinary Analysis of the Presence of ‘alternative Metrics’ in Scientific Publications.” *Scientometrics*, no. Haustein 2010: 1–16. https://doi.org/10.1007/s11192-014-1264-0.

Zavaraqi, Rasoul, and Gholam-Reza Fadaie. 2012. “Scientometrics or Science of Science: Quantitative, Qualitative or Mixed One.” *Collnet Journal of Scientometrics and Information Management* 6 (2): 273–78. https://doi.org/10.1080/09737766.2012.10700939.

Zhang, Min, Feng Ru Sheu, and Yin Zhang. 2018. “Understanding Twitter Use by Major LIS Professional Organisations in the United States.” *Journal of Information Science* 44 (2): 165–83. https://doi.org/10.1177/0165551516687701.

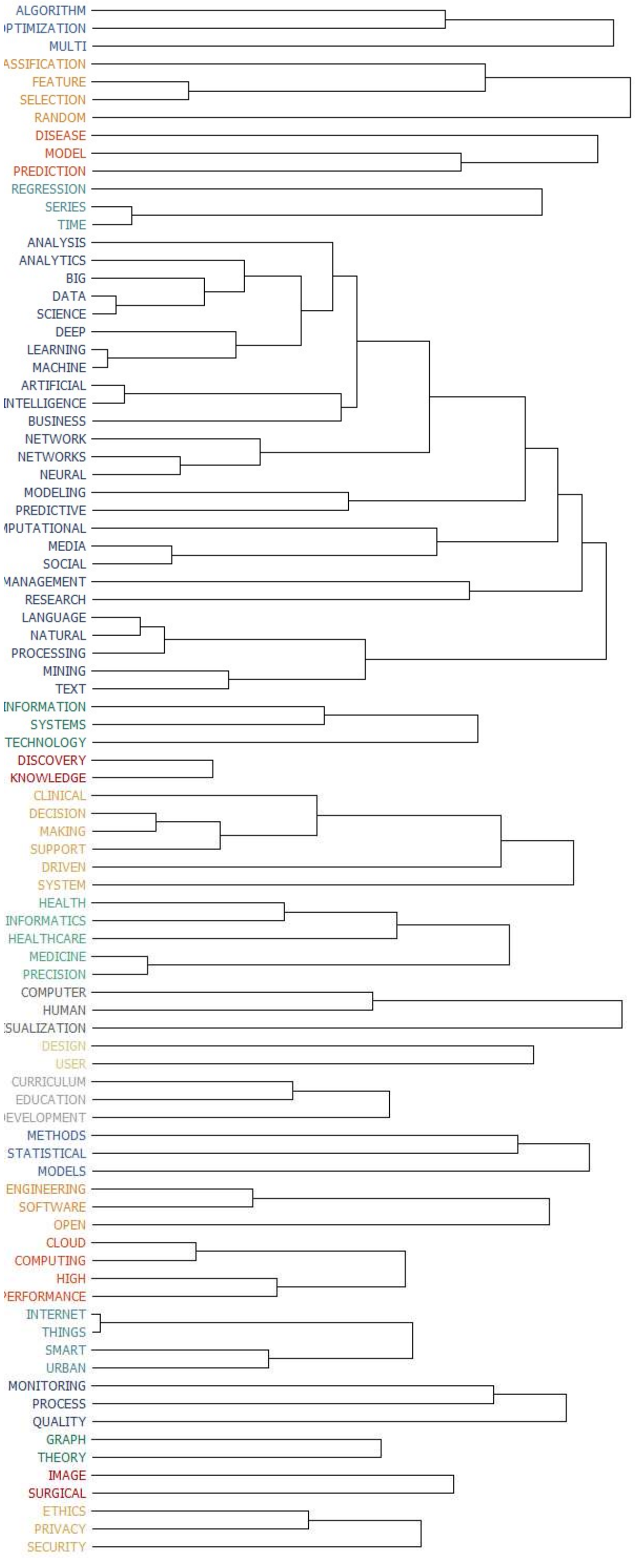
Zhang, Yi, Alan L. Porter, Scott Cunningham, Denise Chiavetta, and Nils Newman. 2018. “How Is Data Science Involved in Policy Analysis?: A Bibliometric Perspective.” *PICMET 2018 - Portland International Conference on Management of Engineering and Technology: Managing Technological Entrepreneurship: The Engine for Economic Growth, Proceedings*, no. August. https://doi.org/10.23919/PICMET.2018.8481979.

Zhao, Dangzhi, and Andreas Strotmann. 2014. “The Knowledge Base and Research Front of Information Science 2006-2010: An Author Cocitation and Bibliographic Coupling Analysis.” *Journal of the Association for Information Science and Technology* 65 (5): 995–1006. https://doi.org/10.1002/asi.23027.

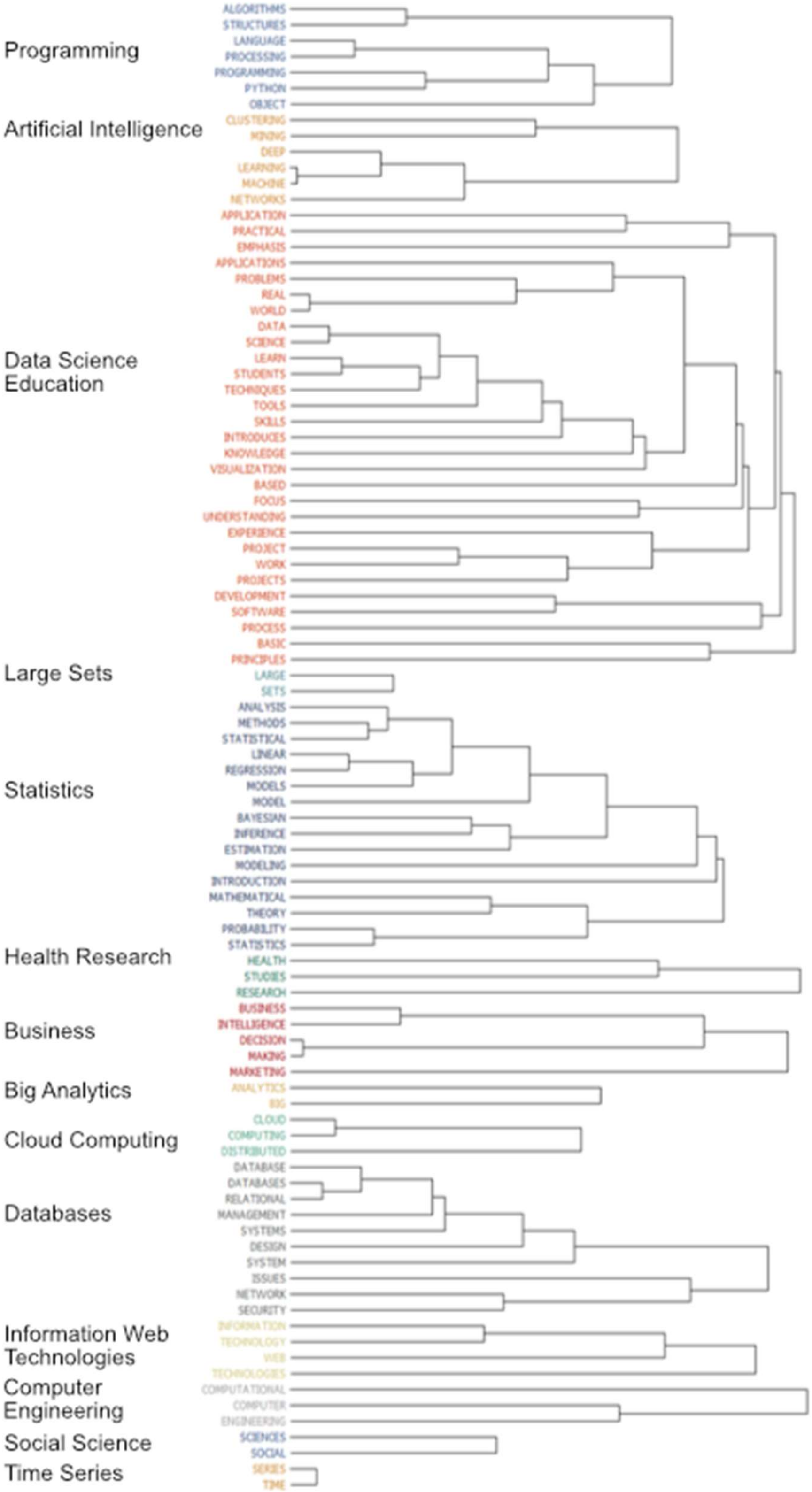
Zheng, Xin, and Aixin Sun. 2019. “Collecting Event-Related Tweets from Twitter Stream.” *Journal of the Association for Information Science and Technology* 70 (2): 176–86. https://doi.org/10.1002/asi.24096.

Zhu, Yangyong, and Yun Xiong. 2015. “Towards Data Science.” *Data Science Journal* 14: 1–7. https://doi.org/10.5334/dsj-2015-008.

Appendix A Scientometric Dendrogram



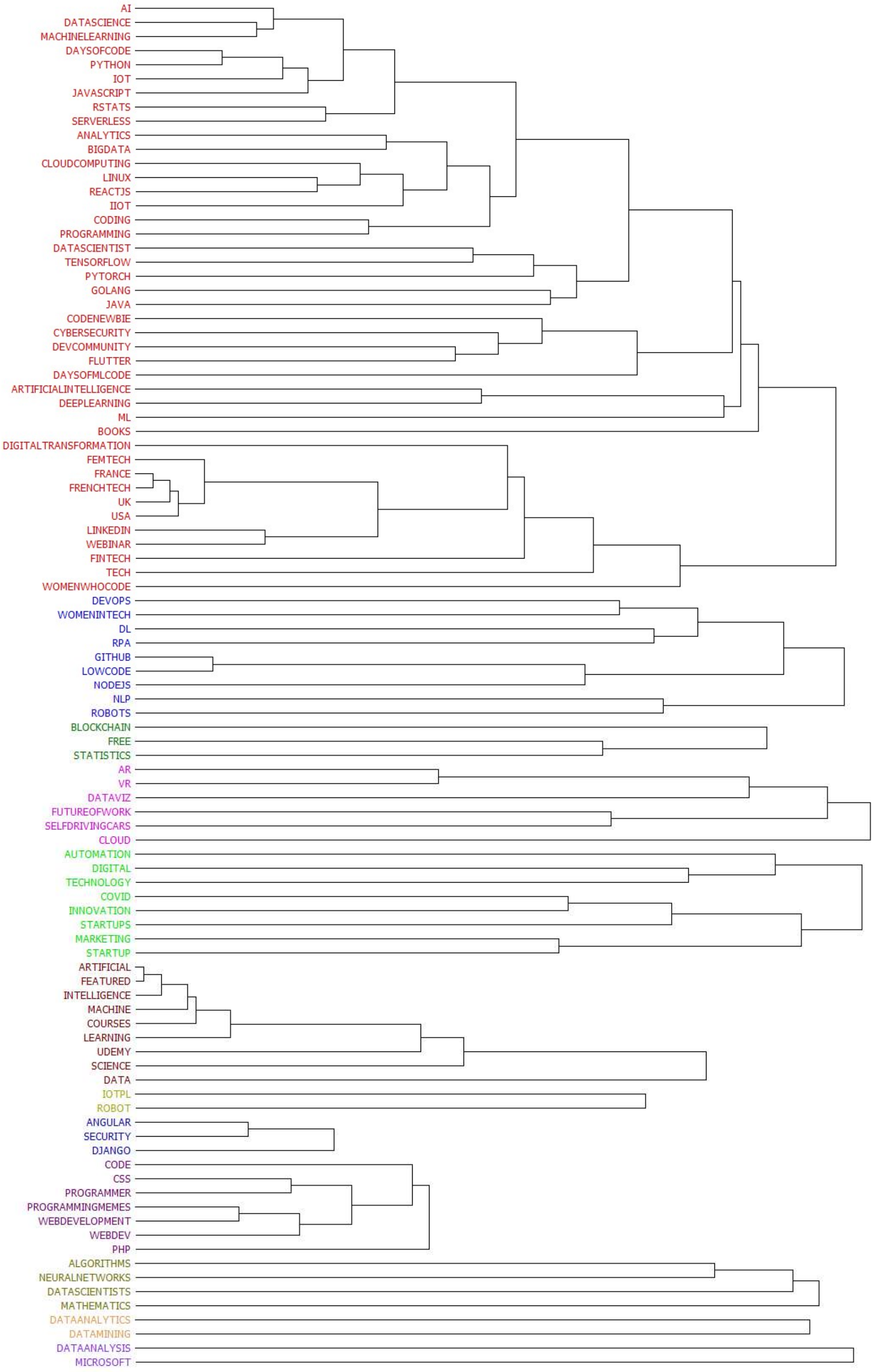
Appendix B Curricular Dendrogram



Appendix C Twitter Abbreviations

Abbreviation	Expanded Form
AI	Artificial Intelligence
AR	Augmented Reality
CodeNewbie	New programmer
DevOps	Development Operations
FinTech	Financial Technologies
IIoT	Industrial Internet of Things
IoT	Internet of Things
LowCode	Simple code or pseudocode
ML	Machine Learning
NLP	Natural Language Processing
RPA	Robotic Process Automation
RStats	R Statistics (Programming)
Udemy	A commercial learning site
VR	Virtual Reality
WebDev	Web Development

Appendix D Hashtag Dendrogram


AI
DATASCIENCE
MACHINELEARNING
DAYSOFCODE
PYTHON
IOT
JAVASCRIPT
RSTATS
SERVERLESS
ANALYTICS
BIGDATA
CLOUDCOMPUTING
LINUX
REACTJS
IIOT
CODING
PROGRAMMING
DATASCIENTIST
TENSORFLOW
PYTORCH
GOLANG
JAVA
CODENEWBIE
CYBERSECURITY
DEVCOMMUNITY
FLUTTER
DAYSOFMLCODE
ARTIFICIALINTELLIGENCE
DEEPLEARNING
ML
BOOKS
DIGITALTRANSFORMATION
FEMTECH
FRANCE
FRENCHTECH
UK
USA
LINKEDIN
WEBINAR
FINTECH
TECH
WOMENWHOCODE
DEVOPS
WOMENINTECH
DL
RPA
GITHUB
LOWCODE
NODEJS
NLP
ROBOTS
BLOCKCHAIN
FREE
STATISTICS
AR
VR
DATAVIZ
FUTUREOFWORK
SELFDRIVINGCARS
CLOUD
AUTOMATION
DIGITAL
TECHNOLOGY
COVID
INNOVATION
STARTUPS
MARKETING
STARTUP
ARTIFICIAL
FEATURED
INTELLIGENCE
MACHINE
COURSES
LEARNING
UDEMY
SCIENCE
DATA
IOTPL
ROBOT
ANGULAR
SECURITY
DJANGO
CODE
CSS
PROGRAMMER
PROGRAMMINGMEMES
WEBDEVELOPMENT
WEBDEV
PHP
ALGORITHMS
NEURALNETWORKS
DATASCIENTISTS
MATHEMATICS
DATAANALYTICS
DATAMINING
DATAANALYSIS
MICROSOFT