Received 8 May 2019
Revised 8 June 2019
Accepted 19 June 2019

Data science from a library and information science perspective

Sirje Virkus
School of Digital Technologies, Tallinn University, Tallinn, Estonia, and
Emmanouel Garoufallou
Department of Library Science and Information Systems,
Alexander Technological Educational Institute of Thessaloniki,
Thessaloniki, Greece and
Deltos Group, Thessaloniki, Greece

Abstract

Purpose – Data science is a relatively new field which has gained considerable attention in recent years. This new field requires a wide range of knowledge and skills from different disciplines including mathematics and statistics, computer science and information science. The purpose of this paper is to present the results of the study that explored the field of data science from the library and information science (LIS) perspective.
Design/methodology/approach – Analysis of research publications on data science was made on the basis of papers published in the Web of Science database. The following research questions were proposed: What are the main tendencies in publication years, document types, countries of origin, source titles, authors of publications, affiliations of the article authors and the most cited articles related to data science in the field of LIS? What are the main themes discussed in the publications from the LIS perspective?
Findings – The highest contribution to data science comes from the computer science research community. The contribution of information science and library science community is quite small. However, there has been continuous increase in articles from the year 2015. The main document types are journal articles, followed by conference proceedings and editorial material. The top three journals that publish data science papers from the LIS perspective are the *Journal of the American Medical Informatics Association*, *the International Journal of Information Management* and the *Journal of the Association for Information Science and Technology*. The top five countries publishing are USA, China, England, Australia and India. The most cited article has got 112 citations. The analysis revealed that the data science field is quite interdisciplinary by nature. In addition to the field of LIS the papers belonged to several other research areas. The reviewed articles belonged to the six broad categories: data science education and training; knowledge and skills of the data professional; the role of libraries and librarians in the data science movement; tools, techniques and applications of data science; data science from the knowledge management perspective; and data science from the perspective of health sciences.
Research limitations/implications – The limitations of this research are that this study only analyzed research papers in the Web of Science database and therefore only covers a certain amount of scientific papers published in the field of LIS. In addition, only publications with the term “data science” in the topic area of the Web of Science database were analyzed. Therefore, several relevant studies are not discussed in this paper that are not reflected in the Web of Science database or were related to other keywords such as “e-science,” “e-research,” “data service,” “data curation” or “research data management.”
Originality/value – The field of data science has not been explored using bibliographic analysis of publications from the perspective of the LIS. This paper helps to better understand the field of data science and the perspectives for information professionals.

Keywords Data science, Data scientist, Skills, Business value, Data management, Information science, Library science, Bibliographic analysis, Literature review, IoT

Paper type Research paper

1. Introduction

Data science is a relatively new term which has gained considerable attention in recent years. The search of this phrase provides now more than 76m hits in Google. The data science field has emerged in response to the increased amount of data. Huge amounts of data have become available to people at all levels of society, through social networks, mobile devices and various sensor devices (i.e. “the Internet of Things”). These new types of data, in enormous volume, in various forms, often complex, unstructured and volatile, are being generated at an accelerating

Data Technologies and
Applications
Vol. 53 No. 4, 2019
pp. 422-441
© Emerald Publishing Limited
2514-9288
DOI 10.1108/DTA-05-2019-0076

rate (Virkus *et al.*, 2018). The majority of digital data is generated by consumers, in the form of movie downloads, VOIP calls, e-mails and cell-phone location readings (Regaldo, 2013). Cukier and Mayer-Schönberger (2013), defining this process as datafication, note that transforming all things under the sun into a data format and thus quantifying them is at the heart of the current world. Just as electricity changed industrial processes and domestic practices in the nineteenth century, a data-driven paradigm is the core of twenty-first century processes and practices (Schäfer and Van Es, 2017, p. 11). Yet only about 0.5 percent of data is ever analyzed (Regaldo, 2013). There is so much more data out there than anyone can capture or analyze and therefore the concept of data overload has been suggested (Virkus *et al.*, 2018).

At the same time, computers have become much more powerful as technology has advanced, "networking is ubiquitous, and algorithms have been developed that can connect data sets to enable broader and deeper analyses than previously possible" (Provost and Fawcett, 2013, p. 51). This has led to the emergence of data science (Cervone, 2016). van der Aalst (2016, p. 4) notes: "Data abundance combined with powerful data science techniques has the potential to dramatically improve our lives by enabling new services and products, while improving their efficiency and quality." This presents an opportunity for better decision making and strategy development (Aristodemou and Tietze, 2018, p. 37).

European library and information science (LIS) education has met a number of challenges in recent years including the financial crisis, negative demographic trends in some countries, emerging technologies, internationalization and globalization. For this reason, innovative ways to survive and achieve the educational goals are constantly needed (Virkus, 2015). Data science is an opportunity that will provide new interdisciplinary perspectives for LIS professionals as well as for LIS education to address new societal needs including e-science and research data management. Therefore, there is a growing interest in data management and data science among library and information professionals (Garoufallou *et al.*, 2008; Antell *et al.*, 2014).

The purpose of this paper is to present the results of the study that explored the field of data science from LIS perspective on the basis of papers published in the Web of Science database. The structure of this paper is organized as follows: the second section describes the research methodology adopted. The third section discusses the concepts of data science and data scientists, the necessary skills required from data scientists and data science-related activities. In the fourth section, the results of the bibliographic analysis of the data science from the LIS perspective are presented. In the fifth section, conclusions are presented.

2. Methodology

Analysis of research publications on data science was made on the basis of papers published in the Web of Science database. Web of Science™ Core Collection provides access to the world's leading citation databases and its authoritative, multidisciplinary content covers over 12,000 of the highest impact journals worldwide, including Open Access journals and over 150,000 conference proceedings across more than 250 disciplines with coverage to 1,900 (Virkus, 2016). Therefore, it seemed reasonable to start exploring this emerging field on the basis of this database.

The following research questions were proposed:

RQ1. What are the main tendencies in publication years, document types, countries of origin, source titles, authors of publications, affiliations of the article authors and the most cited articles related to data science in the field of LIS?

RQ2. What are the main themes discussed in the publications from the LIS perspective?

Searches were carried out in the database by topic in April 2019 using the term "data science." The search strategy discovered 80 publications. The following categories were explored: the years in which the documents were published; the document types of the publication; the countries of origin; the journals in which the documents were published; the authors of the publications; affiliations of the article authors; the most cited articles; and disciplinary affiliation of publications (in addition to information science and library science). A statistical descriptive analysis of these categories of data are presented. The main themes discussed in the publications from the LIS perspective are presented, but the detailed content analysis of 80 publications of this study will be reported in another publication.

The limitations of this research are that this study only analyzed research papers in the Web of Science database and therefore only covers a certain amount of scientific papers published in the field of LIS. In addition, only publications with the term "data science" in the topic area of the Web of Science database were analyzed. Therefore, several relevant studies are not discussed in this paper that are not reflected in the Web of Science database or were related to other keywords such as "e-science," "e-research," "data service," "data curation" or "research data management" (e.g. Osswald, 2008; Garoufallou and Papatheodorou, 2014; Borgman, 2015).

3. Literature review

In recent years, there has been an increasing amount of literature on big data and data science (Wang, 2018, p. 1244). However, making sense of data has a long history and has been discussed by scientists, statisticians, librarians, computer scientists and others for years (Press, 2013a, para. 1). Kelleher and Tierney (2018, p. 6) suggest that on the one hand data science draws on the history of data collection and on the other hand on the history of data analysis. They point out that the earliest form of writing around 3,200 BC was used for commercial record keeping and the first national census was carried out in Egypt in 3,000 BC. The reason that early states put so much effort and resources into large data-collection operations was that these states needed to raise taxes and armies. Someone must have been analyzing the data collected at these times as well, if only to ensure proper tax collection (p. 7). Statistics was developed around 1,700 and it emerged as a distinct and mature discipline around 1,900 (Stigler, 1986). However, data science is a quite recent phenomenon.

3.1 The concept of data science

According to Wainer (2015, p. 2) the term "data science" was first used by Peter Naur in 1960 as a substitute for computer science: "Data science is the study of the generalizable extraction of knowledge from data." According to Ratner (2017, p. 16) in 1971 within the International Federation for Information Processing (IFIP) Guide to Data Processing, a brief definition of data science can be found: "Data science is the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences." In 1974 a Danish computer science pioneer and Turing award winner Peter Naur used the same "data science" definition in his book *Concise Survey of Computer Methods* (cited in Press, 2013a). The roots of the book are in his presentation of a paper at the IFIP Congress in 1968 titled *Datalogy, the Science of Data and of Data Processing and its Place in Education* (Akerkar and Sajja, 2016, p. 2).

However, there is still considerable disagreement about what data science is. Data science has been referred to as a:

- new scientific paradigm, emerging as the fourth scientific paradigm in terms of the previous three: empirical, theoretical and computational science (Bell *et al.*, 2009; Hey *et al.*, 2009; Chen and Zhang, 2014, p. 2);
- new scientific discipline providing techniques, methods and tools to gain value and insights from new and existing data sets (Responsible Data Science, 2016);
- set of disciplines necessary to solve big data challenges (Song and Zhu, 2016, p. 364).

- research field concerned with processes and systems that extract knowledge from massive amounts of data (Zuo *et al.*, 2017, p. 1795);
- practice of collecting data, analyzing data within a problem domain, interpreting findings with graphics and drawing conclusions (Ratner, 2017, p. 127);
- area of work concerned with the collection, preparation, analysis, visualization, management and preservation of large collections of information (Stanton, 2012, p. ii);
- set of principles that support and guide the principled extraction of information and knowledge from data (Provost and Fawcett, 2013, p. 52);
- process of discovering interesting and meaningful patterns in data using computational analytics methods (Amirian *et al.*, 2017, p. 16);
- set of processes for extracting non-obvious and useful patterns from large data (Kelleher and Tierney, 2018); and
- new profession that is expected to make sense of the vast stores of big data (Walker, 2015, p. 8).

Since it is a new field there is both excitement and confusion about it (Vohra, 2013, para. 1) with some people decrying it as equivalent to statistics, and others accepting it as a new scientific discipline. For example, several authors have argued that data science is just a new name for statistics (Diggle, 2015, p. 794). C. F. Jeff Wu, an academic statistician, used the term "data science" in his professorial inaugural lecture and argues that data science is identical to statistics and therefore statistics needs to be replaced by data science and statisticians renamed data scientists (Press, 2013a; Ratner, 2017). In 2001, William S. Cleveland, an academic statistician, views data science as an independent discipline that has emerged from the field of statistics by incorporating advances in computing with data. Cleveland (2001, pp. 21-22) suggested six main areas of data science: multidisciplinary investigations, models and methods for data, computing with data, pedagogy, tool evaluation and theory. However, Ratner (2017) notes that Cleveland's data science is just statistics. Foreman (2013, p. xiv) indicates that "to an extent, data science is synonymous with or related to terms like business analytics, operations research, business intelligence, competitive intelligence, data analysis and modeling, and knowledge extraction (also called knowledge discovery in databases or KDDD). It's just a new spin on something that people have been doing for a long time."

However, Frické (2015, p. 660) argues: "The ability to cheaply and easily gather large amounts of data does have advantages: Sample sizes can be larger, testing of theories can be better, there can be continuous assessment, and so on. But data-driven science, the 'fourth paradigm,' is a chimera. Science needs problems, thoughts, theories, and designed experiments. If anything, Science needs more theories and less data."

Other terms closely related to data science include big data, data analytics and business analytics. Kelleher and Tierney (2018, pp. 1-2) indicate that the terms data science, machine learning and data mining are also often used interchangeably. Provost and Fawcett (2013, p. 52) believe that the most closely related concept to data science is data mining – the actual extraction of knowledge from data via technologies. They suggest that in order to better understand and explain data science as well as to serve business effectively, it is important to understand its relationships to other related concepts, and to begin to identify the fundamental principles underlying data science.

3.2 Defining data science

Numerous authors have attempted to define data science. However, Provost and Fawcett (2013, pp. 51-52) note that today "there is confusion about what exactly is data science, and this confusion could well lead to disillusionment as the concept diffuses into meaningless buzz."

Vincent Granville (2014), who is data scientist himself, believes that the term has been much abused and a lot of hype surrounds big data and data science. However, Voulgaris (2014, p. 16) argues that “data science is not a fad, but something that is here to stay and bound to evolve rapidly” and is “a response to the difficulties of working with big data and other data analysis challenges.” Cao (2018, p. 7) notes that “the abuse, misuse and over-use of the term ‘data science’ is ubiquitous, and essentially contribute to the buzz and hype. Myths and pitfalls are everywhere at this early, and somehow impetuous, stage of data science.”

According to Dhar (2013, p. 64) “data science is the study of the generalizable extraction of knowledge from data.” Provost and Fawcett (2013, p. 52) define data science as “a set of fundamental principles that support and guide the principled extraction of information and knowledge from data.” Foreman (2013, p. xiv) defines data science as “the transformation of data using mathematics and statistics into valuable insights, decisions, and products.” The Data Science Association (2017) defines data science as “the scientific study of the creation, validation and transformation of data to create meaning.” NIST (2018) in the USA (p. 6) defines data science as “the extraction of useful knowledge directly from data through a process of discovery, or of hypothesis formulation and hypothesis testing.” Kelleher and Tierney (2018, p. 1) define data science as “a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large data sets.” They also mention that data science also takes up other challenges such as the capturing, cleaning, and transforming of unstructured social media as well as web data, and the use of big data technologies to store and process large data sets. In addition, questions related to data ethics and regulation are also included (p. 2).

Thus, data science has many aspects, dimensions and various definitions reflect different emphases. According to Greenberg (2017), Dhar (2013) focuses on the predictive capabilities of data, emphasizing application of statistical methods, Stanton (2012) offers a broader definition, explaining that data science encompasses a full range of processes and activities. Greenberg finds that the “unifying factor across various definitions is the ‘science’ that comprises defining appropriate questions, selecting and obtaining suitable data, and applying the correct, at times often innovative, modeling, and statistical methods” (p. 22).

Although different opinions still exist, there appears to be some consensus that data science is an emerging, interdisciplinary field that concerns identifying and extracting valuable patterns from big data, converting data into information and knowledge through data analysis and mining (Wang, 2018, p. 1244). Provost and Fawcett (2013) note that the goal of data science is to discover relationships, trends and patterns extracted from large data sets to gain valuable knowledge in order to support decision making and conclude that each of us sees the field from a different perspective and thereby forms a different conception.

3.3 *Business value of data science*

However, business leaders from all over the world have realized the importance of analyzing big data sets due to its huge operational and strategic potential and the demand for data scientists (Manyika *et al.*, 2011; Provost and Fawcett, 2013). Recent years have seen increasingly rapid application of data science in many fields, including business, economics, industry, education, physics, health care, agriculture, public policy, management, marketing, public transport, urban living, space science and sociology. According to Voulgaris (2014, pp. 13-14) the following industries appear to have benefited, or are inclined to benefit the most from big data: retail (e.g. in terms of productivity boost), telecommunications (in terms of revenue increase), consulting, health care, air transportation, construction, food products, steel and manufacturing, industrial instruments, automobile industry, customer, care, financial services, publishing and logistics. Data scientists can help decision makers to gain valuable insights from varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data (Elgendy and Elragal, 2014) that can be converted into decisions and actions (Pauleen, 2017). The extracted knowledge may help to "describe what happened, explain why something happened, and predict what may or is likely to happen" (Baškarada and Koronios, 2017, p. 65). It is believed that data science is about bridging the different components that contribute to business optimization and productivity, and eliminating the silos that slow down business efficiency (Granville, 2014, p. 11).

Fosso Wamba *et al.* (2015, p. 236) presented five dimensions related to business value creation from big data: creating transparency; enabling experimentation to discover needs, expose variability, and improve performance; segmenting populations to customize actions; replacing/supporting human decision making with automated algorithms; and innovating new business models, products and services. According to the report of the McKinsey Global Institute, "the effective use of big data has the potential to transform economies, delivering a new wave of productivity growth and consumer surplus" (Manyika *et al.*, 2011, p. 13). McAfee and Brynjolfsson (2012, pp. 63-64) provided the results of the study which explored whether data-driven companies are more effective in their performance. Interviews with executives at 330 public North American companies about their organizational and technology management practices, and analysis of performance data from their annual reports and independent sources showed that the more data-driven the companies were, the better they performed. For example, companies in the top third of their industry in the use of data-driven decision making were, on average, 5 percent more productive and 6 percent more profitable than their competitors.

Davenport and Patil (2012, p. 70) have declared a data scientist as the sexiest job in the twenty-first century in the *Harvard Business Review* article. However, Chatfield *et al.* (2014, p. 1) note that despite that assertion, there is no rigorous definition of a data scientist in the literature and what job skills this hottest job title may require. Many organizations lack clear understanding of the required knowledge and skills of data professionals (Kennan, 2017) and this can lead organizations to waste effort when trying to find data professionals (Harris *et al.*, 2013, p. 5). Furthermore, academic institutions are putting together programs to train data scientists without exactly understanding what knowledge and skills are required in different types of organizations (Provost and Fawcett, 2013, p. 51).

3.4 Skills of data scientists

Data scientists and their skills have become one of the most significant current discussions (Davenport and Patil, 2012; Waller and Fawcett, 2013; Wang, 2018). Costa and Santos (2017, p. 726) note that in order to adequately describe the profile of data scientist and distinguish it from other professions it is important to understand its origin, knowledge base and skills set.

The term "data scientist" first appeared in the literature in 2005, when the National Science Board (2005) report "Long-lived digital data collections: enabling research and education in the 21st century" (p. 27) defined it as "the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection." However, several authors believe that the term "data scientist" was coined in 2008 by D. J. Patil and J. Hammerbacher (Davenport and Patil, 2012).

Most people would agree that data scientists require a wide range of knowledge and skills in diverse fields including mathematics and statistics (e.g. hypothesis testing, and predictive analytics), computer science (e.g. data structures and algorithms, machine and algorithmic techniques), information science (e.g. storage and preservation of data, classification, indexing, data curation and management, metadata management and data quality). Data science also relies heavily on probability models, data mining, methods and methodology of data visualization and machine learning in order to understand and use the huge amount of data (Provost and Fawcett, 2013; Granville, 2014; Cervone, 2016; Marchionini, 2016; Song and Zhu, 2016). In addition, it depends on specifics of the domains to which it is applied (e.g. biology, economics, finance, medicine, sociology and psychology) and therefore requires domain expertise. Carter and Sholler (2016, p. 2316) note that critics have argued that data science approaches often ignore the importance of domain knowledge that does play an important role in data science. Song and Zhu (2016, p. 364) claim that data, technology and people are the three pillars of data science. Thus, data science brings together a number of principles, techniques, processes, and methodologies from various disciplines. Successful data scientists must be able to view business problems from a data perspective and support of data-analytic thinking (Provost and Fawcett, 2013, p. 52; Granville, 2014, p. 11). Davenport and Patil (2012, para. 12) note that it is appropriate to "think of him or her as a hybrid of data hacker, analyst, communicator, and trusted adviser" (para. 12). Furthermore, Cao (2018, p. 6) believes that data science has emerged as an inter-, cross- and transdisciplinary new field.

Several national and international reports have tried to identify the skills of data scientists. The Data Science Association (DSA) indicates that a data scientist is somebody who uses "scientific methods to liberate and create meaning from raw data" and "who can play with data, spot trends and learn truths few others know" (Ortiz-Repiso *et al.*, 2018, p. 770). Davenport and Patil (2012) note that "data scientists are high-ranking professionals with the training and curiosity to make discoveries in the world of big data; they make discoveries while swimming in data; they communicate what they've learned and suggest its implications for new business directions." Song and Zhu (2017, p. 5) refer to the three dimensions of data scientists: they should be able to understand roles of big data and big data technologies, work with all the steps of a data science lifecycle – discover problems, solve problems, and communicate solutions and use a range of tools to solve big data problems.

Swan and Brown (2008, pp. 1, 8) suggested four main roles of the data professional: data creator or data author (people with domain expertise who produce data and may have a high level of expertise in handling, manipulating and using data); data scientist (people involved in creative enquiry, and analysis of data, enabling others to work with digital data and developments in database technology); data manager (computer scientists, information technologists or information scientists who take responsibility for computing facilities, storage, continuing access and preservation of data); data librarian (people originating from the library community, trained and specialising in the curation, preservation and archiving of data). Lyon and Brenner (2015, p. 114) proposed the following roles and typical organizational locations: data analyst (people involved in business/scientific analytics, mathematics, statistics, modeling in the corporate sector); data archivist (people involved in long term preservation, repository management in the national archive); data engineer (people involved in software development, coding, programming and tools in the information technology (IT) company); data journalist (people telling stories and providing news using visualizations in the newspaper publisher); data librarian (people involved in advocacy, research data management, and training in the university/research institute); data steward/curator (people involved in curation, cleansing, annotation, selection and appraisal in the data center).

Based on a survey with 250 respondents, Harris *et al.* (2013, p. 11) identified five main skill groups on the basis of a set of 22 generic skills that are applicable to data scientists: business (product development, business), machine learning and Big Data (unstructured data, structured data, machine learning, big and distributed data), math and operations research (optimization, math, graphical models, Bayesian/Monte Carlo statistics, algorithms, simulation), programming (system administration, back end programming, front end programming) and statistics (visualization, temporal statistics, surveys and marketing, spatial statistics, science, data manipulation, classical statistics). The respondents of their study fell into four self-identified groups who had distinctive skills: data businesspersons, data creatives, data developers and data researchers. Data business people focused on the organization and how data projects yield profit. They rated themselves as leaders and entrepreneurs, and reported managing an employee. They had often done contract or consulting work, and started a business. They had technical skills, often undergraduate engineering degrees and about 90 percent had worked on gigabyte-scale problems (Harris *et al.*, 2013, p. 14). Data creatives applied a wide range of tools and technologies to a problem, or created innovative prototypes at hackathons. They had substantial academic experience, undergraduate degrees were often in areas like economics and statistics. They had substantial business expertise, having done contract work (80 percent) or started a business (40 percent), and had the deepest open source experience. Data developers focused on the technical problem of managing data – how to get it, store it, and learn from it. They rated themselves as scientists, particularly those who were closely integrated with the machine learning and related academic communities. They often were writing code in their day-to-day work, contributed to open source projects and had computer science or computer engineering degrees (Harris *et al.*, 2013, p. 15). Data researchers started with academic research in the physical or social sciences, or in statistics. They often had published in peer-reviewed journals and had a PhD. They were least likely to have started a business, and only half have managed an employee (Harris *et al.*, 2013, p. 16).

Stanton *et al.* (2012) suggested the concept of a T-shaped data scientist who should fall into three categories: data curation, analytics and visualization, and networks and infrastructure. The concept of T-shaped skills was first proposed by Iansiti (1993), who refers to experts with specific disciplinary knowledge (vertical bar on the T) who has the ability to interact across disciplines with other experts in other areas (horizontal bar).

Several general competencies are also important for data scientists. Provost and Fawcett (2013, p. 52) find that intuition, creativity and common sense are necessary in certain areas. Davenport and Patil (2012, para. 14) indicate that one of the main characteristics of data scientist is an intense curiosity. Song and Zhu (2017, p. 11) note that the ability to collaborate with domain experts or other members of a team is an essential skill as most data science problems are complex and highly interdisciplinary and require a team effort.

Emmert-Streib *et al.* (2016, p. 2) conclude that "metaphorically, the result of a data analysis process is like a cocktail having a taste that is beyond its constituting ingredients." In addition, it is often difficult to find these divergent skills needed from a data scientist in a single human being and "it's a little bit like looking for a unicorn" (Bertolucci, 2013, para. 9). Nobody is an expert in everything and therefore it makes more sense to have teams of people who have different profiles and different expertise (Schutt and O'Neil, 2014, p. 10).

3.5 Data science-related activities

During the last decade, many authors have published books and articles on data science, the term has been included in the title of conferences, workshops and journals, academic institutions have developed programs to educate and train data scientists, and companies have started to recognize the importance of data scientists and hire them. The following sections provide some examples of these activities.

The first conference that included the term "data science" in its title was a biennial conference "Data Science, Classification, and Related Methods of the International Federation of Classification Societies (IFCS)" which was organized in Kobe, Japan in 1996 (Press, 2013a). Today, the term is often used in the title of conferences, workshops and symposiums all over the world. For example, in 2014 the first "International Conference on Data Science and Advanced Analytics (DSAA)" was organized in Shanghai, China and the sixth DSAA conference will be held in Washington, USA in October 2019. "Applied Machine Learning & Data Science Conference for Developers" was held in Tallinn, Estonia in March 2019.

There will be more than 20 conferences on data science around the world in 2019. For example, "2nd International Data Science Conference" was held in May 2019 in Salzburg, Austria, the "DATA 2019: 8th International Conference on Data Science, Technology and Applications" in Prague, Czech Republic in July, (Springer CCIS) "Research School on Statistics and Data Science 2019" in Melbourne Australia in July, "ACM-2019 2nd International Conference on Data Science and Information Technology" in Seoul, Korea in July, "15th International Conference on Data Science" in Las Vegas, NV USA in July–August, "EEE International Conference on Data Science and Systems" in Zhangjiajie, China in August, "International Conference on Data Science, E-learning and Information Systems" in Dubai, Arab Emirates in December, etc. (www.wikicfp.com/cfp/call?conference=data%20science). In addition, there are a number of other conferences that focus on data science, although their titles are somewhat different; for example, "International Conference on Metadata and Semantic Research" that has taken place since 2005, clearly covers this subject area (published by Springer CCIS, https://link.springer.com/conference/mtsr).

Several data science journals have been established since 2002. In 2002 the Committee on Data for Science and Technology (CODATA) of the International Council for Science (ICSU) started publishing *the Data Science Journal* (www.codata.org/publications/data-science-journal). In 2003 Columbia University began publishing *The Journal of Data Science* (www.jds-online.com/) and in 2012 *EPJ Data Science* journal (https://epjdatascience.springeropen.com/) was established. In 2014 *the Data Science Journal* (www.codata.org/publications/data-science-journal) was relaunched in partnership with Ubiquity Press. In 2015 *The International Journal of Data Science* (www.inderscience.com/jhome.php?jcode=ijds) was launched by Springer and *International Journal on Data Science and Technology (IJDST)* (www.ijdst.org/index) was established. In 2015 *The Statistical Analysis and Data Mining* journal (established in 2008) changed its name to *Statistical Analysis and Data Mining: The ASA Data Science Journal* (http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1932-1872). In 2015 *Data Science and Engineering* (https://link.springer.com/journal/41019) was published under the brand SpringerOpen, on behalf of the China Computer Federation (CCF) and in 2016 *International Journal of Data Science and Analytics* (https://link.springer.com/journal/41060) was established. *The Journal of Finance and Data Science* (www.keaipublishing.com/en/journals/the-journal-of-finance-and-data-science/) was founded in 2016 by two of the world's leading STM publishers, China Science Publishing & Media and Elsevier. The *Journal of Data and Information Science* (JDIS) (http://manu47.magtech.com.cn/Jwk3_jdis/EN/2096-157X/home.shtml), sponsored by the Chinese Academy of Sciences (CAS), was launched in 2016 and is published by the National Science Library of CAS. In 2017 *Data Science* (www.iospress.nl/journal/data-science/) journal and *the International Journal of Population Data Science (IJPDS)* (https://ijpds.org/index) were established. In 2018 *Journal of Data Science and its Applications (JDSA)* (http://commdis.telkomuniversity.ac.id/jdsa/index.php/jdsa) was published by the School of Computing, Telkom University, Bandung, Indonesia. Since 2017, the journal *Program: Electronic Library and Information Systems* bears the name *Data Technologies and Applications (DTA)* (www.emeraldinsight.com/loi/dta). There are three topic areas such as: open and social data: data sharing; semantic data: computational semantics and interoperability; and big data: scale and analytics. In 2019 *International Journal of Data Science and Advanced Analytics (IJDSAA)* (http://ijdsaa.com/index.php/welcome), and the *Journal of Data Science, Statistics, and Visualisation* (https://jdssv.org/index.php/jdssv) of the International Association for Statistical Computing (IASC) were published. In addition, Springer publishes Series on Data Science and the Data Analytics.

Many universities have begun to offer data science educational programs at all levels of undergraduate, Master's and doctoral study. According to Piatetsky (2013) graduate degrees in data science started to emerge in 2007, with a big spike in 2012 (cited in Press, 2013b).

In 2014 Granville (2014, p. 2) notes that "graduate degrees in data science are spreading like mushrooms after the rain." There are 602 entries of colleges and universities providing degrees related to data science in April 2019 on the data science community portal (http://datascience.community/); 101 certificate, 57 bachelor, 420 masters and 23 doctoral programmes. For example, in the USA University of San Francisco offers bachelor in Data Science and Master of Science in Data Science, University of California-Berkeley the Master of Information and Data Science (MIDS) delivered online and New York University PhD in Data Science. In Europe, for example, at bachelor level the University of Warwick in the UK offers Data Science and Maastricht University in the Netherlands Data Science and Knowledge Engineering. At the master level, there are Data Science programmes at the City University London in the UK, University of Barcelona in Spain, Technical University of Denmark and TU Dortmund in Germany. At doctoral level, for example, the data science is offered at the University of Bristol, University of East London and in the University of Edinburgh. Data science programmes are also offered in Australia, Brazil, Canada, Hong Kong, Israel, Mexico, Russia, Singapore, Turkey, Ukraine and New Zealand.

A study by McKinsey Global Institute estimated that 140,000 to 190,000 data science positions would be needed in the USA in 2018 (Manyika *et al.*, 2011, p. 10). Average annual salary for an R programming language expert in the USA was $115,000 and for a Python programming language expert $101,000 according to a Dice Tech Salary Survey (2015). Thus, it is not surprising that numerous data science programs have emerged at major universities in the USA (Brunner and Kim, 2016) and around the world.

However, the content of data science degree programmes varies considerably. Every institution has its own vision and ways of teaching data science (Song and Zhu, 2017, p. 6). Data science programmes involve a blend of courses from mathematics, computer science and statistics, with additional emphasis on workflow, reproducible research, communication, and visualization (Reid, 2018, p. 43). Critical curricular topics include mathematical foundations, computational thinking, statistical thinking, principles of effective data management, techniques for data description and curation, data modeling approaches, effective communication skills, reproducibility challenges and current best practices, exposure to ethical dilemmas and problem-solving skills, and a range of domain-specific topics (National Academies of Sciences, Engineering, and Medicine, 2018). Cady (2017) notes that most programmes seem to fall into one of two camps: introductory courses that focus on the application of data science in a particular arena, such as text analytics, or advanced courses that dive deeply into the mathematical underpinnings of statistical and machine learning. However, they are not taking always into account the full lifecycle of working with data: data acquisition, data cleaning, data analysis, data visualization, and data modeling (Cady, 2017). Granville (2014, p. 2) even notes that often it is just repackaging old material (such as statistics and R programming) with the new label "data science." Organizational affiliation of the data science education programs is ranging from computer science over mathematics to art and sciences and also influence the content of the programmes.

Several frameworks for the education of data professionals have been developed (Waller and Fawcett, 2013; Granville, 2014; Song and Zhu, 2016, 2017). For example, the EU funded EDISON project developed a Data Science Competence Framework (Demchenko *et al.*, 2017). Foster *et al.* (2018, pp. 1423-1424) note that none of these frameworks encompass the whole information life and the data value cycle and data work is largely considered in a decontextualized fashion.

Therefore, there are still many challenges for universities to decide what to teach in their data science programmes, and for students who decide which topics are useful for their careers (Kennan, 2017). Kim *et al.* (2011, p. 133) present a top ten list of recommended courses which include: digital data curation, optimally in a course specialized toward the curation of large scientific or engineering data sets; database design and management, focusing on large scale relational databases; project management, including project planning and budgeting; essentials of scientific research, including literature review, study design, and descriptive statistical analysis; overview of cyberinfrastructure, including cloud and grid computing; geographically distributed collaboration, with a judicious division of time between the human issues and the technological issues; web content management and web interaction design; scripting or practical introductory programming; data mining, with a focus on either quantitative data for the natural sciences or mixed data types for the social sciences and humanities; and information system management and server administration, including general IT and computer knowledge.

However, while there is no doubt that the demand for data scientists is increasing, the supply of knowledge and skills tends toward training. Foster *et al.* (2018, pp. 1423-1424) note that there is a need to shift from the training to the education of the data professional.

4. Data science from the library and information science perspective

In the Web of Science database 2,350 publications were received under the topic area "data science" in the period 1980–2019. The highest contribution to data science comes from the computer science research community. In total, 44.8 percent (1,052) of publications come from the subject area of computer science, followed by engineering 18.2 percent (427), mathematics 7.4 percent (175), science technology and other topics 5.6 percent (131), business economic 4.6 percent (108), education and educational research 3.6 percent (85), information science and library science 3.4 percent (80), physics 2.9 percent (68), telecommunications 2.9 percent (68), medical informatics 2.7 percent (64), materials science 2.6 percent (60) and operations research and management science 2.6 percent (60). Table I illustrates the percentage of articles by subject area.

Next, the analysis of the bibliographic information of the 80 articles (n1) in the subject area of information science and library science is analyzed. The first paper on data science in the field of information science and library science was published in 2005. Table II shows the number of papers per year since the year 2005. The number of articles has increased over the last few year reaching to 23 articles published in 2018; it appears that there has been continuous increase in articles from the year 2015.

The main document types are journal articles 50 (62.5 percent), followed by conference proceedings 12 (15.0 percent), editorial material 9 (11.3 percent), reviews 6 (7.5 percent) and book chapters 3 (3.7 percent) (Table III).

Table IV reveals the top journals in which relevant articles are published. The top three journals, which account for 26.4 percent of the articles are the *Journal of the American Medical Informatics Association* (13.8 percent), the *International Journal of Information*

Subject area	Number of publications	Percentage of publications
Computer Science	1,052	44.8
Engineering	427	18.2
Mathematics	175	7.4
Science Technology and Other Topics	131	5.6
Business Economics	108	4.6
Education and Educational Research	85	3.6
Information Science and Library Science	80	3.4
Physics	68	2.9
Telecommunications	68	2.9
Medical Informatics	64	2.7
Materials Science	60	2.6
Operations Research/Management Science	60	2.6

Table I.
The subject areas of the publications

Management (6.3 percent) and the *Journal of the Association for Information Science and Technology* (6.3 percent). These are followed by *Scientometrics* (3.7 percent) and *Communications in Computer and Information Science*, *Frontiers in Artificial Intelligence and Applications*, *Information Research*, *Information Systems Research*, the *Journal of Academic Librarianship*, the *Journal of Data and Information Science*, *Legal Knowledge and Information Systems*, *Library Hi Tech*, *Online Information Review*, each 2.5 percent. The top four journals account for a total of 24 articles, the other one or two articles are divided between the other 47 sources. This indicates that that the field of data science and its core journals in the field of information science are still in its infancy.

The Table V shows that the top five countries are USA, China, England, Australia and India, with 38, 7, 6 and 4 articles respectively. It is evident from the analysis that USA is the leading continent in the field of data science from the LIS perspective. From European

Year	Number of publications	Percentage of publications
2019	6	7.5
2018	23	28.8
2017	17	21.3
2016	16	20.0
2015	9	11.3
2014	3	3.7
2013	4	5.0
2012	1	1.2
2005	1	1.2

Table II.
Years of data science publications

Document types	Number of publications	Percentage of publications
Article	50	62.5
Proceedings paper	12	15.0
Editorial material	9	11.3
Review	6	7.5
Book chapter	3	3.7

Table III.
Document types

Document types	Number of publications	Percentage of publications
Journal of the American Medical Informatics Association	11	13.8
International Journal of Information Management	5	6.3
Journal of the Association for Information Science and Technology	5	6.3
Scientometrics	3	3.7
Communications in Computer and Information Science	2	2.5
Frontiers in Artificial Intelligence and Applications	2	2.5
Information Research	2	2.5
Information Systems Research	2	2.5
Journal of Academic Librarianship	2	2.5
Journal of Data and Information Science	2	2.5
Legal Knowledge and Information Systems	2	2.5
Library Hi Tech	2	2.5
Online Information Review	2	2.5

Table IV.
The main sources of publications

Country	Number of publications	Percentage of publications
USA	38	47.5
China	7	8.8
England	6	7.5
Australia	4	5.0
India	4	5.0
Canada	3	3.7
Germany	3	3.7
Greece	2	2.5
Hungary	2	2.5
The Netherlands	2	2.5
Pakistan	2	2.5
Portugal	2	2.5
Saudi Arabia	2	2.5
South Korea	2	2.5
Sweden	2	2.5

Table V.
The main countries of publications

countries, only England (6; 7.5 percent), Germany (3; 3.7 percent), Greece (2; 2.5 percent), Hungary (2; 2.5 percent), Netherlands (2; 2.5 percent), Portugal (2; 2.5 percent) and Sweden (2; 2.5 percent) have some publications.

The main authors are Lucila Ohno-Machado (four publications; 5 percent), who is the editor in chief of the *Journal of the American Medical Informatics Association*, Victor Chang (3; 3.7 percent) from the Xi'an Jiaotong Liverpool University (Suzhou, China) and the Center for Mobile Cloud Computing Research, University of Malaya (Kuala Lumpur, Malaysia), H Frank Cervone (2; 2.5 percent) from the University of Illinois at Chicago (Chicago, Illinois, USA), Jan Nolin (2; 2.5 percent) from the University of Borås (Borås, Sweden) and Jane Greenberg (2; 2.5 percent) and Yongjun Zhu (2; 2.5 percent) from the Drexel University College of Computing and Informatics (Philadelphia, USA).

Table VI shows the top affiliations of the article authors. The main contributions come from the Drexel University (4; 5.0 percent), followed by the University of Illinois (3; 3.7 percent) and Xi'an Jiaotong Liverpool University (3; 3.7 percent).

Table VII shows the top 10 cited articles. The most cited article is "Big data, data science, and analytics: the opportunity and challenge for IS research" by Agarwal and Dhar (2014) with 112 citations (average citations per year 18.7 percent) in the journal *Information Systems Research.* This is followed by article "The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data" by Margolis *et al.* (2014) with 102 citations (average citations per year 17.0 percent) in the *Journal of the American*

Organization	Number of publications	Percentage of publications
Drexel University	4	5.0
University of Illinois	3	3.7
Xi'an Jiaotong Liverpool University	3	3.7
King Abdulaziz University	2	2.5
New York University	2	2.5
Syracuse University	2	2.5
University California Los Angeles	2	2.5
University of Florida	2	2.5
University of Maryland	2	2.5
University of Texas Austin	2	2.5
University of Washington	2	2.5

Table VI.
The main affiliation of publications

Authors of the paper	Number of citations	Average citations per year
Agarwal and Dhar (2014)	112	18.7
Margolis *et al.* (2014)	102	17.0
Park and Leydesdorff (2013)	40	5.7
Sundararajan *et al.* (2013)	37	5.3
Larson and Chang (2016)	29	4.7
Kumar *et al.* (2015)	15	2.4
Antell *et al.* (2014)	14	2.3
Si *et al.* (2013)	14	2.3
Intezari and Gressel (2017)	8	2.0
Yousafzai *et al.* (2016)	8	2.0

Table VII.
The top 10 cited articles

Medical Informatics Association and "Decomposing social and semantic networks in emerging 'big data' research" by Park and Leydesdorff with 40 citations (average citations per year 5.7 percent) in the *Journal of Informetrics*. Sundararajan *et al.* (2013) paper "Information in Digital, Economic, and Social Networks" in the journal *Information Systems Research* has received 37 citations (average citations per year 5.3 percent) and the paper "A review and future direction of agile, business intelligence, analytics and data science" by Larson and Chang (2016) in the *International Journal of Information Management* 29 citations (average citations per year 4.7 percent).

The analysis also revealed that the data science field is quite interdisciplinary by nature (Table VIII). Only 30 (37.5 percent) publications belonged to the research area of Information Science and Library Science (ISLS). In total, 21 (26.3 percent) of publications belonged both to the ISLS and Computer Science research area. In total, 11 (13.8 percent) of publications belonged to the ISLS, Computer Science, Health Care Sciences and Services and Medical Informatics research area. Four (5.0 percent) of publications belonged to the ISLS and Business and Economics research area or ISLS, Computer Science and Business and Economics research area. Two (2.5 percent) of publications belonged to the ISLS and Education and Educational Research area and ISLS and Computer Science and Government and Law research area. One (1.2 percent) publication belonged to ISLS, Computer Science and Social Sciences – other topics research area, ISLS, Computer Science and Environmental and Occupational Health research area, ISLS, Computer Science and Operational Research

Research areas	Number of publications	Percentage of publications
Information Science and Library Science (ISLS)	30	37.5
ISLS and Computer Science	21	26.3
ISLS; Computer Science; Health Care Sciences and Services; Medical Informatics	11	13.8
ISLS; Business and Economics	4	5.0
ISLS; Computer Science; Business and Economics	4	5.0
ISLS; Education and Educational Research	2	2.5
ISLS; Computer Science; Government and Law	2	2.5
ISLS; Computer Science; Social Sciences – other topics	1	1.2
ISLS; Computer Science; Environmental and Occupational Health	1	1.2
ISLS; Computer Science; Operational Research and Management Science	1	1.2
ISLS; Computer Science; Geography; Physical Geography	1	1.2
ISLS; History; Social Sciences – other topics	1	1.2
ISLS; Computer Science; Astronomy and Astrophysics	1	1.2

Table VIII.
Disciplinary affiliation

and Management Science research area, ISLS, Computer Science, Geography and Physical Geography research area, ISLS, History, Social Sciences – other topics research area and ISLS Computer Science and Astronomy and Astrophysics research area.

In total, 80 papers in the research area of information science and library science of the Web of Science database published in 2005–2019 can be divided into six main categories: data science education and training, knowledge and skills of the data professional, the role of libraries and librarians in the data science movement, tools, techniques and applications of data science, data science from the knowledge management perspective, data science from the perspective of health sciences. There were several other topics that were discussed only in few papers. For example, the relations of data science and LIS, the role of metadata in data science, data science from the perspective of information systems and from the legal science, suggested research directions. The detailed content analysis of 80 publications will be reported elsewhere.

5. Conclusions

During the last decade, many authors have published books and articles in the field of data science, the term has been included in the title of conferences, workshops and journals, academic institutions have developed programs to educate and train data scientists and companies have started to recognize the importance of data scientists and hire them. Since it is a new field there is still both excitement and confusion about it.

In the Web of Science database 2,350 publications were received under the topic area "data science" in the period 1980–2019. The highest contribution to data science comes from the computer science research community (44.8 percent). The contribution of information science and library science community is quite small (3.4 percent).

The first paper in the area of data science, reflected in the Web of Science database, was published in 2005. The number of articles has increased over the last few years. It appears that there has been continuous increase in articles from the year 2015. The main document types are journal articles, followed by conference proceedings and editorial material. The top three journals that publish data science papers from the LIS perspective are the *Journal of the American Medical Informatics Association*, the *International Journal of Information Management* and the *Journal of the Association for Information Science and Technology*.

The top five countries publishing are USA, China, England, Australia and India. It is evident from the analysis that USA is the leading continent in the field of data science from the LIS perspective. From European countries, only England, Germany, Greece, Hungary, the Netherlands, Portugal and Sweden have some publications. The most cited article has got 112 citations.

The analysis also revealed that the data science field is quite interdisciplinary by nature. In addition to LIS, the papers belonged to several other research areas: for example, computer science, medical informatics, health care sciences and services, business and economics, operational research and management sciences, government and law, environmental and occupational health, geography and physical geography, astronomy and astrophysics, education and educational research, history and social sciences – other topics research areas.

The reviewed articles were very diverse in content and belonged to the six broad categories: data science education and training; knowledge and skills of the data professional; the role of libraries and librarians in the data science movement; tools, techniques and applications of data science; data science from the knowledge management perspective; data science from the perspective of health sciences.

It is hoped that the findings of this research will help LIS educators and practitioners understand the educational challenges triggered by the advent of the data-driven society and the opportunities of the field of data science. They can also recognize their potential to contribute in the field of data curation and management, data analytics, visualization and presentation. This study only analyzed publications with the term "data science" in the topic area of the Web of Science database, but the future research could expand the field of research and explore publications also reflected by Scopus and Google Scholar and add other keywords such as "e-science," "e-research," "data service" "data curation" and "research data management."

References

Agarwal, R. and Dhar, V. (2014), "Big data, data science, and analytics: the opportunity and challenge for IS research", *Information Systems Research*, Vol. 25 No. 3, pp. 443-448.

Akerkar, R. and Sajja, P.S. (2016), *Intelligent Techniques for Data Science*, Springer International Publishing, Cham.

Amirian, P., van Loggerenberg, F. and Lang, T. (2017), "Data science and analytics", in Amirian, P., Lang, T. and van Loggerenberg, F. (Eds), *Big Data in Healthcare, SpringerBriefs in Pharmaceutical Science & Drug Development*, Springer, Cham, pp. 15-37.

Antell, K., Foote, J.B., Turner, J. and Shults, B. (2014), "Dealing with data: science librarians' participation in data management at association of research libraries institutions", *College & Research Libraries*, Vol. 75 No. 4, pp. 557-574.

Aristodemou, L. and Tietze, F. (2018), "The state-of-the-art on intellectual property analytics (IPA): a literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data", *World Patent Information*, Vol. 55, pp. 37-51.

Baškarada, S. and Koronios, A. (2017), "Unicorn data scientist: the rarest of breeds", *Program*, Vol. 51 No. 1, pp. 65-74.

Bell, G., Hey, T. and Szalay, A. (2009), "Beyond the data deluge", *Science*, Vol. 323 No. 5919, pp. 1297-1298.

Bertolucci, J. (2013), "Are you recruiting a data scientist or a unicorn?", *InformationWeek,* available at: www.informationweek.com/big-data/big-data-analytics/are-you-recruiting-a-data-scientist-or-unicorn/d/d-id/899843 (accessed April 4, 2019).

Borgman, C.L. (2015), *Big Data, Little Data, No Data: Scholarship in the Networked World*, The MIT Press, Cambridge, MA.

Brunner, R.J. and Kim, E.J. (2016), "Teaching data science", *Procedia Computer Science*, Vol. 80, pp. 1947-1956.

Cady, F. (2017), *The Data Science Handbook*, John Wiley & Sons, Hoboken, NJ.

Cao, L. (2018), *Data Science Thinking: The Next Scientific, Technological and Economic Revolution*, Springer, Cham.

Carter, D. and Sholler, D. (2016), "Data science on the ground: hype, criticism, and everyday work", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 10, pp. 2309-2319.

Cervone, H.F. (2016), "Informatics and data science: an overview for the information professional", *Digital Library Perspectives*, Vol. 32 No. 1, pp. 7-10.

Chatfield, A.T., Shlemoon, V.N., Redublado, W. and Rahman, F. (2014), "Data scientists as game changers in big data environments", *Proceedings of the 25th Australasian Conference on Information Systems, Auckland University of Technology*, pp. 1-11.

Chen, C.P. and Zhang, C.-Y. (2014), "Data-intensive applications, challenges, techniques and technologies: a survey on big data", *Information Sciences*, Vol. 275, pp. 314-347.

Cleveland, W.S. (2001), "Data science: an action plan for expanding the technical areas of the field of statistics", *International Statistical Review*, Vol. 69 No. 1, pp. 21-26.

Costa, C. and Santos, M.Y. (2017), "The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age", *International Journal of Information Management*, Vol. 37 No. 6, pp. 726-734.

Cukier, K. and Mayer-Schönberger, V. (2013), “The rise of big data: how it’s changing the way we think about the world”, *Foreign Affairs*, Vol. 92 No. 3, pp. 28-40.

Davenport, T.H. and Patil, D.J. (2012), “Data scientist: the sexiest job of the 21st century”, *Harvard Business Review*, Vol. 90 No. 5, pp. 70-76, available at: www.hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century (accessed November 3, 2018).

Demchenko, Y., Belloum, A. and Wiktorski, T. (2017), “EDISON data science framework: part 1. Data science competence framework (CF-DS) release 2”, available at: www.edison-project.eu/sites/edison-project.eu/files/filefield_paths/edison_cf-ds-release2-v08_0.pdf (accessed November 1, 2018).

Dhar, V. (2013), “Data science and prediction”, *Communications of the ACM*, Vol. 56 No. 12, pp. 64-73.

Dice Tech Salary Survey (2015), available at: www.marketing.dice.com/pdf/Dice_TechSalarySurvey_2015.pdf (accessed November 1, 2018).

Diggle, P.J. (2015), “Statistics: a data science for the 21st century”, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 178 No. 4, pp. 793-813.

Elgendy, N. and Elragal, A. (2014), “Big data analytics: a literature review paper”, in Perner, P. (Ed.), *Advances in Data Mining: Applications and Theoretical Aspects*, Springer International Publishing, Cham, pp. 214-227.

Emmert-Streib, F., Moutari, S. and Dehmer, M. (2016), “The process of analyzing data is the emergent feature of data science”, *Frontiers in Genetics*, Vol. 7, p. 12.

Foreman, J.W. (2013), *Data Smart: Using Data Science to Transform Information into Insight*, John Wiley & Sons, Hoboken, NJ.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G. and Gnanzou, D. (2015), “How ‘big data’ can make big impact: findings from a systematic review and a longitudinal case study”, *International Journal of Production Economics*, Vol. 165, pp. 234-246.

Foster, J., McLeod, J., Nolin, J. and Greifeneder, E. (2018), “Data work in context: value, risks, and governance”, *Journal of the Association for Information Science and Technology*, Vol. 69 No. 12, pp. 1414-1427.

Frické, M. (2015), “Big data and its epistemology”, *Journal of the Association for Information Science and Technology*, Vol. 66 No. 4, pp. 651-661.

Garoufallou, E. and Papatheodorou, C. (2014), “A critical introduction to metadata for e–science and e–research”, *International Journal of Metadata, Semantics and Ontologies*, Vol. 9 No. 1, pp. 1-4.

Garoufallou, E., Balatsoukas, P., Siatri, R., Zafeiriou, G., Asderi, S. and Ekizoglou, P. (2008), “Greek academic librarians’ perceptions of the impact of Google on their role as information providers”, *Education for Information*, Vol. 26 No. 2, pp. 133-145.

Granville, V. (2014), *Developing Analytic Talent: Becoming a Data Scientist*, John Wiley and Sons, Incorporated, Hoboken, NJ.

Greenberg, J. (2017), “Big metadata, smart metadata, and metadata capital: toward greater synergy between data science and metadata”, *Journal of Data and Information Science*, Vol. 2 No. 3, pp. 19-36.

Harris, H., Murphy, S. and Vaisman, M. (2013), *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*, O’Reilly Media, Sebastopol, CA.

Hey, A.J., Tansley, S. and Tolle, K. (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, WA.

Iansiti, M. (1993), “Real-world R&D: jumping the product generation gap”, *Harvard Business Review*, Vol. 71 No. 3, pp. 138-147.

Intezari, A. and Gressel, S. (2017), “Information and reformation in KM systems: big data and strategic decision-making”, *Journal of Knowledge Management*, Vol. 21 No. 1, pp. 71-91.

Kelleher, J.D. and Tierney, B. (2018), *Data Science*, MIT Press, Cambridge, MA.

Kennan, M.A. (2017), "'In the eye of the beholder': knowledge and skills requirements for data professionals", *Information Research*, Vol. 22 No. 4, available at: www.informationr.net/ir/22-4/rails/rails1601.html (accessed January 18, 2019).

Kim, Y., Addom, B.K. and Stanton, J.M. (2011), "Education for eScience professionals: integrating data curation and cyberinfrastructure", *International Journal of Digital Curation*, Vol. 6 No. 1, pp. 125-138.

Kumar, S., Abowd, G.D., Abraham, W.T., al'Absi, M., Gayle Beck, J., Chau, D.H., Condie, T., Conroy, D.E., Ertin, E., Estrin, D. and Ganesan, D. (2015), "Center of excellence for mobile sensor data-to-knowledge (MD2K)", *Journal of the American Medical Informatics Association*, Vol. 22 No. 6, pp. 1137-1142.

Larson, D. and Chang, V. (2016), "A review and future direction of agile, business intelligence, analytics and data science", *International Journal of Information Management*, Vol. 36 No. 5, pp. 700-710.

Lyon, L. and Brenner, A. (2015), "Bridging the data talent gap: positioning the iSchool as an agent for change", *International Journal of Digital Curation*, Vol. 10 No. 1, pp. 111-122.

McAfee, A. and Brynjolfsson, E. (2012), "Big data: the management revolution", *Harvard Business Review*, Vol. 90 No. 10, pp. 60-68.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H. (2011), *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, New York, NY, available at: www.bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf (accessed January 14, 2019).

Marchionini, G. (2016), "Information science roles in the emerging field of data science", *Journal of Data and Information Science*, Vol. 1 No. 2, pp. 1-6.

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J. and Green, E.D. (2014), "The national institutes of health's big data to knowledge (BD2K) initiative: capitalizing on biomedical big data", *Journal of the American Medical Informatics Association*, Vol. 21 No. 6, pp. 957-958.

National Academies of Sciences, Engineering, and Medicine (2018), "Envisioning the data science discipline: the undergraduate perspective: interim report", The National Academies Press, Washington, DC, available at: www.nap.edu/catalog/24886/envisioning-the-data-science-discipline-the-undergraduate-perspective-interim-report

National Science Board (2005), "Long-lived digital data collections: enabling research and education in the 21st century", available at: www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf (accessed January 16, 2019).

NIST (2018), "National institute of standards and technology (NIST) special publication 1500-1r1. NIST Big Data Interoperability Framework: Volume 1, Definitions. Version 2. NIST Big Data Public Working Group (NBD-PWG)", available at: www.nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1r1.pdf (accessed January 14, 2019).

Ortiz-Repiso, V., Greenberg, J. and Calzada-Prado, J. (2018), "A cross-institutional analysis of data-related curricula in information science programmes: a focused look at the iSchools", *Journal of Information Science*, Vol. 44 No. 6, pp. 768-784.

Osswald, A. (2008), "E-science and information services: a missing link in the context of digital libraries", *Online Information Review*, Vol. 32 No. 4, pp. 516-523.

Park, H.W. and Leydesdorff, L. (2013), "Decomposing social and semantic networks in emerging 'big data' research", *Journal of Informetrics*, Vol. 7 No. 3, pp. 756-765.

Pauleen, D.J. (2017), "Davenport and Prusak on KM and big data/analytics: interview with David J. Pauleen", *Journal of Knowledge Management*, Vol. 21 No. 1, pp. 7-11.

Piatetsky, G. (2013), "Analytics education boom – trends and overview", available at: www.kdnuggets.com/2013/02/education-analytics-data-mining-trends-overview.html (accessed January 17, 2019).

Press, G. (2013a), "A very short history of data science", *Forbes*, available at: www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#121ba83755cf (accessed January 14, 2019).

Press, G. (2013b), "Data science: what's the half-life of a buzzword?", *Forbes*, available at: www.forbes.com/sites/gilpress/2013/08/19/data-science-whats-the-half-life-of-a-buzzword/#3e86a69c7bfd (accessed January 19, 2019).

Provost, F. and Fawcett, T. (2013), "Data science and its relationship to Big Data and data-driven decision making", *Big Data*, Vol. 1 No. 1, pp. 51-59.

Ratner, B. (2017), *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data*, Chapman and Hall/CRC Press, Boca Raton.

Regaldo, A. (2013), "The data made my do it", *MIT Technology Review*, July/August, available at: www.technologyreview.com/news/514346/the-data-made-me-do-it/ (accessed January 14, 2019).

Reid, N. (2018), "Statistical science in the world of big data", *Statistics & Probability Letters*, Vol. 136, pp. 42-45.

Responsible Data Science (2016), available at: www.redasci.org (accessed January 19, 2019).

Schäfer, M.T. and Van Es, K. (Eds) (2017), *The Datafied Society: Studying Culture Through Data*, Amsterdam University Press, Amsterdam.

Schutt, R. and O'Neil, C. (2014), *Doing Data Science: Straight Talk from the Frontline*, O'Reilly Media, Sebastopol, CA.

Si, L., Zhuang, X., Xing, W. and Guo, W. (2013), "The cultivation of scientific data specialists: development of LIS education oriented to e-science service requirements", *Library Hi Tech*, Vol. 31 No. 4, pp. 700-724.

Song, I.Y. and Zhu, Y. (2016), "Big data and data science: what should we teach?", *Expert Systems*, Vol. 33 No. 4, pp. 364-373.

Song, I.Y. and Zhu, Y. (2017), "Big data and data science: opportunities and challenges of iSchools", *Journal of Data and Information Science*, Vol. 2 No. 3, pp. 1-18.

Stanton, J. (2012), *Data Science*, Syracuse University, Syracuse, NY, available at: www.storage2.ischool.syr.edu/media.ischool.syr.edu/oldmedia/documents/2012/3/DataScienceBook1_1.pdf (accessed January 14, 2019).

Stanton, J.M., Palmer, C.L., Blake, C. and Allard, S. (2012), "Interdisciplinary data science education", *Special Issues in Data Management (ACS Symposium Series, Vol. 1110), American Chemical Society, Washington, DC.*

Stigler, S.M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press, Cambridge, MA.

Sundararajan, A., Provost, F., Oestreicher-Singer, G. and Aral, S. (2013), "Information in digital, economic, and social networks", *Information Systems Research*, Vol. 24 No. 4, pp. 883-905.

Swan, A. and Brown, S. (2008), "The skills, role and career structure of data scientists and curators: an assessment of current practice and future needs. Report to the JISC, Key Perspectives, Playing Place", available at: www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.8960&rep=rep1&type=pdf (accessed January 14, 2019).

The Data Science Association (2017), "About data science", available at: www.datascienceassn.org/about-data-science (accessed January 12, 2019).

van der Aalst, W.M. (2016), "Responsible data science: using event data in a 'people friendly' manner", *International Conference on Enterprise Information Systems, Springer*, *Cham*, pp. 3-28.

Virkus, S. (2015), "Change and innovation in European library and information science education", BiD: textos universitaris de biblioteconomia i documentació núm 35 (desembre), available at: http://bid.ub.edu/en/35/virkus.htm (accessed January 12, 2019).

Virkus, S. (2016), "Knowledge management and information literacy: an exploratory analysis", *European Conference on Information Literacy, Springer*, *Cham*, pp. 119-129.

Virkus, S., Mandre, S. and Pals, E. (2018), "Information overload in a disciplinary context", in Kurbanoğlu, S., Boustany, J., Špiranec, S., Grassian, E., Mizrachi, D. and Roy, L. (Eds), *Information Literacy in the Workplace*, Springer, Cham, pp. 615-624.

Vohra, G. (2013), "Myriad opportunities in data science", *Deccan Herald*, available at: www.deccanherald.com/content/316957/myriad-opportunities-data-science.html (accessed January 12, 2019).

Voulgaris, Z. (2014), *Data Scientist: The Definitive Guide to Becoming a Data Scientist*, Technics Publications, Westfield, NJ.

Wainer, H. (2015), *Truth or Truthiness: Distinguishing Fact from Fiction by Learning to Think Like a Data Scientist*, Cambridge University Press, Cambridge, MA.

Walker, M.A. (2015), "The professionalisation of data science", *International Journal of Data Science*, Vol. 1 No. 1, pp. 7-16.

Waller, M.A. and Fawcett, S.E. (2013), "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management", *Journal of Business Logistics*, Vol. 34 No. 2, pp. 77-84.

Wang, K. (2018), "Twinning data science with information science in schools of library and information science", *Journal of Documentation*, Vol. 74 No. 6, pp. 1243-1257.

Yousafzai, A., Chang, V., Gani, A. and Noor, R.M. (2016), "Directory-based incentive management services for ad-hoc mobile clouds", *International Journal of Information Management*, Vol. 36 No. 6, pp. 900-906.

Zuo, H., Zhang, G., Pedrycz, W., Behbood, V. and Lu, J. (2017), "Fuzzy regression transfer learning in Takagi–Sugeno fuzzy models", *IEEE Transactions on Fuzzy Systems*, Vol. 25 No. 6, pp. 1795-1807.

Further reading

Zhang, J., Fu, A., Wang, H. and Yin, S. (2017), "The development of data science education in China from the LIS perspective", *International Journal of Librarianship*, Vol. 2 No. 2, pp. 3-17.

Corresponding author

Sirje Virkus can be contacted at: sirje.virkus@tlu.ee