The Birth of a New Discipline: Data Science Education

Koby Mike¹ Benny Kimelfeld² Orit Hazzan¹

¹Faculty of Education in Science and Technology, Technion – Israel Institute of Technology, Haifa, Israel,
²Faculty of Computer Science, Technion – Israel Institute of Technology, Haifa, Israel

Published on: Oct 27, 2023
DOI: https://doi.org/10.1162/99608f92.280afe66
License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)
ABSTRACT

Data science is emerging from an interdisciplinary integration of mathematics, statistics, computer science, and many other application domains such as business, transportation, biology, and education. Parallel to the emergence of data science, a new field is developing as well—data science education.

Based on our expertise in data science education, gained in our extensive research, the main research topics of the emerging field of data science education have not yet been mapped, and its scope, as reflected by the papers published so far on data science education, has not yet been defined. Several systematic reviews on topics in data science education have been written in recent years, but they tend to focus on specific topics or specific populations, such as K-12 learners.

To partially close this gap, we conducted a systematic machine learning–based literature review. We collected 1,048 papers in the field of data science education using both keyword searches in well-established scientific databases as well as semantic searches in semantic scientific databases. Using SPECTER to generate document-level embedding, we clustered the papers by applying the k-means algorithm. The titles and abstracts of the papers were examined manually to identify main cluster topics. The result is a framework of 26 clusters, organized into five superclusters: (a) curriculum, (b) pedagogy, (c) STEM skills, (d) domain adaptation, and (e) social aspects.

The large body of literature on data science education that exists and the framework of its research topics that we revealed indicate the birth of a new scientific discipline—data science education.

Keywords: data science education, systematic literature review, SPECTER embeddings

1. Introduction

Data science is emerging from an interdisciplinary integration of mathematics, statistics, computer science, and many other application domains such as business, health, education, and transportation. In parallel, the new field of data science education is developing as well. As of today, data science education is growing mainly in the context of other educational communities, such as computer science education, statistics education, engineering education, business analytics education, social science education, and many other educational and research communities (Hazzan and Mike, 2021). Only a handful of journals are dedicated to data science education (for example, the Journal of Statistics and Data Science Education, the special issue of the Statistics Education Research Journal on Data Science Education, and HDSR’s Data Science Education collection).
Since the research and knowledge regarding data science education is spread over many communities and journals, it does not seem trivial to ask the following question: What are the main research topics of data science education? Several reviews were written in recent years on specific topics within data science education. Such topics include, for example, data science curricula (Wu, 2017), data science skills (Gurcan & Cagiltay, 2019), data science for non-STEM majors (Barboza & Teixeira, 2020), and machine learning for K-12 pupils (Foster & Tasnim, 2020; Lupusoru et al., 2021; Martins & Gresse Von Wangenheim, 2022). These reviews, however, focused on specific aspects of data science education, but did not reveal a comprehensive picture of what data science education is. To partially close this gap, this article presents a systematic review of the literature on data science education.

While a systematic literature review is a well-established method for academic research, the case of data science education poses several challenges. First, no designated journals on data science education exist, nor has even a single research community been established that can help focus the search. Second, any systematic literature review is limited by the capacity of the human researcher and, therefore, it may be appropriate for a narrow topic of interest with several dozen to several hundred papers. Data science education, however, as presented later in this article, is a wide field of research with thousands of published research papers, far beyond the surveying capacity of a human researcher. We therefore recruited data science itself for this task, and specifically the emerging technology of producing machine learning–based systematic literature reviews.

The result is a framework of 26 clusters, organized into five superclusters: (a) curriculum, (b) pedagogy, (c) STEM skills, (d) domain adaptation, and (e) social aspects. The first supercluster addresses the components of data science programs, that is, what to teach. The second focuses on how to teach data science, which is a complex task, as will be described later. The third and fourth categories refer, respectively, to the teaching of specific components of data science: teaching data science within an application domain and the teaching of the STEM components of data science: mathematics, statistics, and computer science. The fifth category, social aspects, focuses on the learners as data science users and practitioners.

The rest of this article is organized as follows: Section 2 presents the background for this research, including a review of the short history of data science education and a theoretical background regarding manual and algorithmic systematic literature reviews. Section 3 describes the research method and section 4 presents the results. We discuss the results in section 5 and conclude in section 6.

2. Background: The Short History of Data Science Education

In recent years, several committees were formed to discuss data science education, and specifically to formalize a data science curriculum. In 2015, the National Science Foundation (NSF), together with the ACM (Association for Computing Machinery) Education Board and Council, organized a workshop on data science education (Cassel & Topi, 2016). The product of this workshop was a report titled Strengthening Data Science Education Through Collaboration. One of the main motivations for organizing this workshop was the shortage
of data science professionals, which in turn prevented organizations and societies from enjoying the potential benefits of data science. Although by 2015 several educational programs had already been launched to address this shortage, many workshop participants felt that the quality and future direction of these programs, as well as their structures and practices, did not provide a sufficiently broad perspective on data science and should be defined better. The goal of this workshop was, therefore, “to start a conversation to address these concerns and develop a deeper integrated understanding of the best ways to offer data science education, ultimately leading to a better prepared workforce” (Cassel & Topi, 2016, p. 3).

In the summer of 2016, the NSF and the Institute for Advanced Study at Princeton University funded another workshop that focused on formulating curriculum guidelines for an undergraduate data science degree. This workshop was held at the Park City Math Institute at Princeton University in New Jersey. The workshop product was a report titled *Curriculum Guidelines for Undergraduate Programs in Data Science* (De Veaux et al., 2017) that, among other things, emphasized that the guidelines were not prescriptive, but rather, designed to inform and enumerate the core skills that a data science major should have.

In December 2016, the US National Academies of Sciences, Engineering, and Medicine established the Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective, charging it with the goal of setting forth a vision for the emerging discipline of data science at the undergraduate level (National Academies of Sciences, Engineering, and Medicine, 2018). The committee recognized two challenges in designing data science programs: the diversity of the domains in which data science programs are developed and the complex nature of data science itself. The committee’s final report stated that:

> Current data science courses, programs, and degrees are highly variable in part because emerging educational approaches start from different institutional contexts, aim to reach students in different communities, address different challenges, and achieve different goals. This variation makes it challenging to lay out a single vision for data science education in the future that would apply to all institutions of higher learning, but it also allows data science to be customized and to reach broader populations than other similar fields have done in the past. (National Academies of Sciences, Engineering, and Medicine, 2018, p. 16)

In parallel to the efforts made in the United States, the European Commission funded the EDISON project in 2017 to define the data science profession and promote the education and training of data scientists (EDISON, n.d.). The EDISON project produced the EDISON Data Science Framework (EDSF), which is composed of four documents that define the data science profession and its required competences and knowledge. In addition, the EDSF also presents a model for a data science curriculum (EDISON, 2017).

The last committee we review here is the ACM Data Science Task Force formed in 2017 by the ACM Educational Council. The task force’s mission was to “explore a process to add to the broad, interdisciplinary
conversation on data science, with an articulation of the role of computing discipline-specific contributions to this emerging field” (Danyluk & Leidig, 2021, p. 6).

Following the development of data science courses and data science curricula for undergraduates, an effort was made to adjust and develop data science courses and data science curricula for school pupils as well. In 2017, a symposium on data science curricula for schools was held in Paderborn, Germany (Biehler & Schulte, 2018), and soon afterwards, a draft of a data science curriculum for school pupils was published (Heinemann et al., 2018). Following this publication, the International Data Science in Schools Project (IDSSP) published a curriculum framework for introductory high school data science (Fisher et al., 2019). In addition, the Mobilize Introduction to Data Science program was developed aiming to enhance computational and statistical thinking skills so that pupils will be able to access and analyze data from a variety of traditional and nontraditional sources (Gould et al., 2018). In Israel, a data science course for high school computer science pupils has since been integrated into the current, official Israeli high school computer science curriculum (Mike et al., 2020). Rosenberg and Jones (2022) recently completed a comprehensive systematic review of data science programs for K-12.

Beyond the academic attempts to define data science curricula and programs, the industry also launched several initiatives to define the data science body of knowledge and skills, from a commercial perspective. One example is the Initiative for Analytics and Data Science Standards (IADSS), developed by Usama Fayyad and Hamit Hamutcu (2020). This initiative was motivated by the need of the data science ecosystem to address the needs of data science education and to promote cooperation between academic institutions and nonacademic initiatives that also train data scientists.

3. Method

3.1. Overview

The aim of this research is to characterize the field of data science education by identifying the main research topics explored in the field using a systematic literature review. According to Kitchenham (2004), “Systematic review is a means of evaluating and interpreting all available research relevant to a particular research question, topic, area, or phenomenon of interest” (p. 1). To fairly evaluate a research topic, a trustworthy, rigorous, and auditable methodology is required, and indeed, over the years several methods have been proposed for systematic literature reviews, for example, by Kitchenham (2004), Nightingale, (2009), and Torres-Carrión et al. (2018).

The fast growth in the number of scientific publications and the concurrent advances in natural language processing (NLP) have led to the development of new systematic literature review techniques (Tandjung & Fudholi, 2022), which largely fall under the umbrella of topic modeling. In general, topic modeling refers to algorithms that can extract latent variables from large data sets and can, for example, support the structuring of
databases of academic papers into groups based on similar focus areas (Vayansky & Kumar 2020). In recent years, several NLP-based systematic literature reviews have been published in several domains, such as machine learning research (Sharma et al., 2019), mobile learning (Hamzah et al., 2020), and artificial intelligence in education (Paek and Kim 2021).

Running an NLP-based systematic review requires numerical representation of documents. Although several methods for representing documents as numbers exist, for example, TFIDF (term frequency-inverse document frequency) (Joachims, 1996), the current state-of-the-art method for representing words and documents is deep learning. SPECTER is a method based on training transformer language models to generate document-level embedding of scientific documents (Cohan et al., 2020). The training is based on pretraining a transformer language model (BERT) on the citation graph, which is a powerful signal of document-level relatedness. Recent research demonstrated using SPECTER embeddings for academic research. For example, Yamazaki et al. (2022) experimented with SPECTER representation for paper mining and Mozgai et al. (2022) used the technique for scoping reviews.

To cluster the documents, we use the k-means algorithm (Hartigan, 1975) and in order to label the clusters, we applied a qualitative manual process of coding and categorization for revealing a grounded theory (Glaser & Strauss, 2017).

The research follows the following steps (see Figure 1):

First, we planned the research according to the systematic literature review methodology.

The second stage, data collection, is common to systematic literature review and to grounded theory. The data science technique used in this stage was a semantic search, which is a subfield of NLP (see Section 3.2).

The third stage, data synthesis, was performed by applying data science methods and the grounded theory practice of coding and categorization. The data science methods used in this step were document embedding using SPECTER and the k-means algorithm (see Section 3.3). The coding and categorization practice of grounded theory is an iterative process that aims to enhance category coherence, and in which the researchers identify categories and refine them as the data is rereviewed. The iterative process of coding and categorization is reflected in the co-refinements of the k-means hyperparameters and the semantic understanding of the clusters.

Finally, we report our findings, which reflect a first attempt at defining and mapping the emerging field of data science education (see Section 4). Using grounded theory terminology, this mapping is the emerged theory, or in less formal terminology, it is the theoretical framework.

The complete code for this research can be found here: https://github.com/Data-Science-Education-Review/litreture-review.
3. Data Collection

As mentioned, a systematic literature review is a common academic research method and several guidelines have been published for its implementation, for example, Kitchenham (2004), Nightingale (2009), and Torres-Carrion et al. (2018). These guidelines, however, were written with manual systematic reviews in mind, and do not consider algorithmic systematic reviews. We therefore followed the Procedures for Performing Systematic Reviews (Kitchenham, 2004) with the necessary adaptations for the case of an algorithmic systematic review. According to these guidelines, after defining the research question(s), the following steps are required for data collection: defining the search strategy, defining the selection criteria and procedures, and defining a quality assessment measure for each study.

3.2.1. Search Strategy

Gusenbauer and Haddaway (2020) investigated which academic search systems are suitable for systematic reviews or meta-analyses. Their analysis included 28 resources, including Google Scholar and Web of Science, which they classified either as being suitable to serve as a principal source or as a supplementary source. In our
case, we collected data from two principal sources that are relevant for data science education research: the ACM Digital Library (ACM-DL) and Web of Science (WoS). As supplementary sources we selected Google Scholar because it contains references to a body of research that is not covered by Web of Science (Martín-Martín et al., 2018), Semantic Scholar database because it allows a semantic search of academic papers (Ammar et al., 2018), and the IEEE Xplore database as a major source of engineering education research.

3.2.2. Selection Criteria

To answer our research question ‘What are the main research topics of data science education?’, the research aims to collect all papers published in academic journals and conferences that discuss data science education. Kitchenham (2004) proposed searching for various terms that represent various facets of the research question. To limit the scope of the review, we searched only for the term ‘data science,’ omitting related terms such as ‘data analytics’ and ‘data engineering.’ At the same time, to include the core components of education, we did include the terms ‘teaching,’ ‘curriculum,’ and ‘pedagogy’ (Shulman, 1986). Selection criteria included the terms ‘data science’ and one of the terms ‘education,’ ‘curriculum,’ ‘pedagogy,’ and ‘teaching’ in the paper title or the paper abstract (see Appendix). We did not search for the search terms in the full text because a full-text search retrieved many non-relevant papers (for example, papers about data science applications in the domain of education).

3.2.3. Study Quality Assessment

For this research, which aimed to identify trends of a large body of research, we were interested only in the research topic and not the research quality. Therefore, we did not look at parameters of research quality such as research method or number of participants. We did, however, verify text validity, and all paper titles and abstracts were manually reviewed by all authors of this article. Papers with nonvalid titles or abstracts were excluded.

3.2.4. Collection Procedure

All search queries were executed in September 2022. ACM-DL, WoS, and IEEE Xplore support Boolean search in specific fields such as paper title and paper abstract. Google Scholar and Semantic Scholar do not support Boolean search and so for those databases, we run full-text database queries, retrieved all returned documents, and excluded papers that did not meet the selection criteria after they were retrieved (see Section 3.2.5).

3.2.5. Inclusion and Exclusion Criteria

Table 1 presents the total number of papers retrieved, excluded, and included in the research. Papers were excluded if the search terms were not found in the paper title or paper abstract (see Section 3.2.4). Papers were included only if we could retrieve their SPECTER embeddings from the Semantic Scholar database. Even though about 20% of the papers found in the other sources were missing from Semantic Scholar, Semantic
Scholar does cover a wide variety of academic sources (see https://www.semanticscholar.org/about/publishers) and can, therefore, be considered a reasonable sampling frame for this research. We did not choose a larger number of clusters to maintain cluster sizes of at least 20 papers.

Table 1. Total number of papers retrieved, included, and excluded.

<table>
<thead>
<tr>
<th>Database</th>
<th>Retrieved by Query</th>
<th>Meet Exclusion Criteria</th>
<th>Do Not Meet Inclusion Criteria</th>
<th>Included in Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM-DL</td>
<td>335</td>
<td>46</td>
<td>77</td>
<td>212</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>2,446</td>
<td>1,228</td>
<td>602</td>
<td>616</td>
</tr>
<tr>
<td>IEEE Xplore</td>
<td>179</td>
<td>15</td>
<td>21</td>
<td>143</td>
</tr>
<tr>
<td>Semantic Scholar</td>
<td>29,771</td>
<td>29,030</td>
<td>53</td>
<td>688</td>
</tr>
<tr>
<td>Web of Science</td>
<td>675</td>
<td>222</td>
<td>31</td>
<td>422</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>2,081</strong></td>
</tr>
<tr>
<td><strong>Duplicates</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>1,033</strong></td>
</tr>
<tr>
<td><strong>Total unique</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>1,048</strong></td>
</tr>
</tbody>
</table>

3.3. Data Analysis

Data were analyzed using the iterative coding and categorization process of grounded theory, with the support of machine-learning techniques (k-means) to cluster the data. As the k-means algorithm is a hard-clustering algorithm that clusters each sample to a single cluster, it does not treat multiple-topic documents well (Mozgai et al., 2022). For example, a document that discusses teaching methods for machine learning for K-12 pupils should belong both to the teaching methods cluster, the machine learning cluster, and the K-12 cluster. To mitigate this shortcoming of the k-means algorithm, we used a large number of clusters, based on the elbow method (Syakur et al., 2018). As the number of clusters increased, the accepted clusters were smaller and more homogeneous. Also, a larger number of clusters represents more topics and minimizes the chance of overlooking a topic. Ultimately, we selected 26 clusters, each of which ended up containing 21 to 73 papers. The clusters were manually labeled by all authors of this article using the grounded theory methodology (Glaser & Strauss, 2017). This procedure enabled us to find connections between the clusters and build categories as well as superclusters. Disagreements between the authors regarding cluster labels were discussed until agreement was achieved. About two-thirds of the clusters were focused and were easy to label and to
agree upon. About one-third of the clusters were broader, more difficult to label, and longer discussions were needed to reach agreed-upon labels.

3.4. Limitations

We divide the limitations into two sections: SPECTER and clustering.

SPECTER

- Even through SPECTER embeddings achieve superior results on various NLP tasks performed on academic papers, compared with other document representation methods, SPECTER representation has not been tested for the task of literature review and only a few researchers have used it for similar tasks such as scoping review (Mozgai et al., 2022).
- SPECTER embeddings have been tested mainly on interdomain tasks. In our research, the papers were all collected from one domain (data science education) and so it was harder to differentiate between the papers.
- We could not include all of the retrieved papers in the research since we could not retrieve the SPECTER embeddings of all papers.

Clustering

- Cluster generation was sensitive to the random initialization of the k-means algorithm, indicating that other options of clustering were possible.
- Some of the clusters were difficult to label and long discussions between the authors were needed to reach agreed labels. While the labels do fit most of the papers in each cluster, some of the papers seem to not fit the agreed-upon label, and we treat them as outliers.

4. Results

The 1,048 papers collected were grouped into 26 clusters, which were then coded (labeled) by the authors to extract their topics (see Table 2; full details of the papers appear in the spreadsheet). After the clusters were labeled, they were organized in five superclusters: (a) curriculum, (b) pedagogy, (c) STEM skills, (d) domain adaptation, and (e) social aspects.

Table 2. Clusters topics.

<table>
<thead>
<tr>
<th>Cluster Number</th>
<th>Supercluster</th>
<th>Number of Papers</th>
<th>Cluster</th>
<th>Number of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Topic</td>
<td>Page</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------</td>
<td>------</td>
<td>--------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Curriculum</td>
<td>297</td>
<td>Principles of data science curriculum design</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td></td>
<td></td>
<td>Approaches to data science education</td>
<td></td>
</tr>
<tr>
<td>1.3</td>
<td></td>
<td></td>
<td>Introduction to data science course</td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td></td>
<td></td>
<td>Curriculum design for K-12 pupils</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td></td>
<td>Curriculum design for data science majors</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>Pedagogy</td>
<td>162</td>
<td>Teaching AI and machine learning</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td></td>
<td></td>
<td>Teaching methods</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td></td>
<td></td>
<td>Online teaching</td>
<td></td>
</tr>
<tr>
<td>2.4</td>
<td></td>
<td></td>
<td>Tools and methods</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>STEM skills</td>
<td>190</td>
<td>Statistics education</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td></td>
<td>Computer science for data science</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td></td>
<td></td>
<td>Cloud computing in data science education</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td></td>
<td></td>
<td>Data engineering</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td></td>
<td></td>
<td>Statistics for data science</td>
<td></td>
</tr>
</tbody>
</table>
4.1. Curriculum

‘Data science curriculum’ is the largest supercluster defined, containing 297 papers out of 1,048. Papers in this supercluster discuss various aspects of data science curricula, including principles of data science curriculum design (Cluster 1.1) and approaches to data science education (Cluster 1.2). While papers in Cluster 1.1 address programs for various formats and diverse populations, special attention is given to three cases: the Introduction to Data Science course (Cluster 1.3), data science for K-12 pupils (Cluster 1.4), and data science programs for data science majors (Cluster 1.5).

The principles of curriculum design (Clusters 1.1 and 1.4) include various formats of data science curriculum frameworks, ranging from one-semester introductory courses to full degrees for diverse populations, for example, school pupils, undergraduates, teachers, graduates, and non-STEM students. The topics discussed in
the papers include the required knowledge, competence, and skills, based on the academia’s and industry’s requirements; defining data science literacy, for example, data science skills required for anyone; balancing computer science and statistics knowledge with application domain knowledge; generating a collaborative and interdisciplinary curriculum, involving different faculties in academia, industry, and community; organizing teaching according to the data science workflow; integrating theory and practice using real-life data; developing lab environments to support working with real life data; using data from students’ domains; and assessment methods for data science.

4.2. Pedagogy

Data science education faces many challenges that result from the integration of a variety of application domains and core STEM subjects: mathematics, statistics, and computer science. Consequently, on the one hand, non-STEM students are required to learn STEM subjects on a higher level than customary prior to the data science era, and on the other hand, STEM students are required to pay more attention to the (usually) non-STEM application domains. New teaching methods have, therefore, been developed for teaching data science. Four topics revealed within this supercluster are: teaching AI and machine learning (Cluster 2.1), teaching methods for data science (Cluster 2.2), online teaching (Cluster 2.3), and tools and methods for data science education (Cluster 2.4).

4.3. STEM Skills

Data science is an integration of computer science, mathematics, statistics, and an application domain. While the papers in Supercluster 4 (domain adaptation) focus on data science education from the application domain’s viewpoint, the papers in this supercluster address data science education from the perspective of the STEM components of data science, that is, computer science and statistics. Clusters 3.1 and 3.5 are connected to statistical education and Clusters 3.2, 3.3, and 3.4 are connected to computer science education. We did not identify a cluster that focuses on mathematical education for data science.

The literature on statistics education discusses the following main topics: integrating data science into statistics curriculum and teaching data science for statisticians, teaching computer science as part of the statistics curriculum, teaching statistics as part of the data science curriculum, and methods of teaching statistics. One topic that recurs in many of these clusters addresses the interrelationship between statistics and data science, including similarities and differences between the two disciplines (see, e.g., MacGillivray, 2021).

Papers in Cluster 3.2 (computer science for data science) discuss the relationship between computer science education and data science education, similarly to the focus of Clusters 3.1 and 3.5 on statistics. Papers in this cluster address computer science curriculum for data science, data science curriculum for computer science, and the role of software development in data science. Clusters 3.3 (cloud computing in data science education) and 3.4 (data engineering) focus on more technical aspects of statistics and computer science tools and skills required for data science, such as notebooks, databases, Hadoop, cloud computing, computing environments
for data science education, and big data infrastructure. We note that although ‘data engineering’ was not explicitly included in the search terms for data collection (3.2.2), the semantic nature of the search did retrieve papers on this topic.

4.4. Domain Adaptation

Data science has become an essential tool for research and practice in a variety of application domains. Accordingly, this supercluster includes papers regarding the need to educate professionals in these domains as well as methods to embed data science into the curricula of these application domains. Papers in this supercluster address data science education in the following domains: business analytics (Cluster 4.1), health (Cluster 4.2), digital technologies (Cluster 4.3), biomedicine (Cluster 4.4), education (Cluster 4.6), and nursing (Cluster 4.7). Additional domains are aggregated in Cluster 4.5 and include geography, astronomy, hydrology, environmental science, biodiversity, and physics.

4.5. Social Aspects

This supercluster aggregates different topics that relate to human and social aspects of data science education: ethics (Cluster 5.1), data science as a skill (Cluster 5.2), engagement (Cluster 5.3), and diversity (Clusters 5.4 and 5.5).

Ethics (Cluster 5.1) is a major issue in data science education, and 62 papers were grouped under this topic. Main themes that are discussed in this cluster are the importance of understanding the human aspects of data and the importance of data for understanding humankind as well as privacy, fairness, accountably, transparency, ethics in the curriculum, teaching methods for ethics, data science for social good, critical thinking, politics of data, feminist views of data, data and equality, and social and cognitive biases. Several papers use the COVID-19 pandemic as a case study for discussing many of the above-listed issues.

Papers in Cluster 5.2 (data science as a skill) discuss the importance of bringing data science education to everyone, since it is nowadays considered to be an important skill for any profession. Similar to this cluster, papers in Clusters 5.4 and 5.5 discuss the large variety of professionals who should learn data science. Unlike clusters in Supercluster 2 (integrating data science education into application domain programs), which focus mostly on curriculum aspects, papers in these clusters (5.4 and 5.5) focus on the motivation for teaching data science to different populations. Thus, Cluster 5.4 addresses the diversity of students and Cluster 5.5 focuses on enhancing diversity in data science and its relevance for all.

Papers in Cluster 5.3 discuss several engagement issues in data science education, for example, making programs more interesting for students by incorporating real-life projects and real-life data, and overcoming barriers for data science education, such as the required high level of mathematics, which not all learners are comfortable with.
5. Discussion

Data science is an interdisciplinary domain that not only captures the interdisciplinary integration of computer science and statistics, but also encapsulates an integration of technical domains (computer science, statistics) with a real-world application domain (Adhikari & Jordan, 2021). This interdisciplinarity creates special challenges in data science education that are represented by the categories found in the literature review, as described below.

In the curriculum category, many papers discuss the challenge of generating a real interdisciplinary course that is feasible within the time constraints, on the one hand, and includes the broad required knowledge base, on the other. This topic was raised at the very beginning of the discussions regarding data science education. In an interview Robert A. Lue held with Laura Haas and Alfred Hero, co-chairs of the Committee on Envisioning the Data Science Discipline, Haas stated that the new element in data science is data acumen, the skill of selecting the proper data and tools for the problem at hand, and that:

Getting meaningful, correct, and useful answers from data requires skills that are typically not fully developed in traditional mathematics, statistics, and computer science (CS) courses. (Haas et al., 2019)

In the pedagogy category, many papers discuss methods for interdisciplinary teaching. The ACM Data Science Task Force Final Report summarizes this point nicely:

Each component of the data science environment: the domain that provides the data; statistics and mathematics for analysis, modeling, and inference; and computer science for data access, management, protection, as well as effective processing in modern computer architectures, is essential. However, a random collection of the three elements does not constitute a meaningful data science program. Data science is interdisciplinary and requires the effective integration of the three components to produce meaningful results. (Danyluk & Leidig, 2021, p. 10)

The large body of research that discusses domain adaptation and diversity further demonstrates the unique challenge of data science education, that is, to accommodate multiple diverse audiences. This diversity of audiences is reflected both in the level of learners (from kindergarten to postdoc students) and in their backgrounds (from humanities to exact sciences and engineering). This challenge was also referred to in the above-mentioned interview held by Alfred Hero who said:

Other findings [of the report], like the need to focus on data science ethics and on broadening participation in the field (and classroom) may be less familiar. (Haas et al., 2019)

The last category, social aspects, demonstrates unique challenges and opportunities in data science education. For example, since data science is also a research paradigm, it is relevant for many different academic
researchers from diverse domains; as a result, it increases inclusion in STEM education and enables diversity in STEM subjects. Thus, according to Laura Haas,

> The many types of problems to which data science is applicable will be appealing to diverse students, and this range of educational opportunities will be critical in engaging them, and beneficial to the industry as a whole. (Haas et al., 2019)

Another aspect of this cluster is data science ethics, which involves two different topics—data ethics and application domain ethics—which are integrated in the context of data science. One category, STEM skills, represents challenges that are common to data science education and to the domains forming data science: mathematics, statistics, computer science, and engineering.

Data science education, therefore, includes unique challenges and opportunities: teaching computer science and statistics within the context of application domains, teaching the importance of the domain knowledge to learners who specialize in computer science and statistics, teaching advanced computer science and statistics concepts to domain specialists who lack sufficient mathematical background, teaching research skills to data science learners and practitioners, teaching ethical considerations that derive from both data ethics and application domain ethics, and, finally, leveraging new opportunities in data science education to increase inclusion in STEM education. While most of the research on data science education is published in the context of other domains, such as computer science education and statistics education, many of the challenges and opportunities of data science education are unique to data science and stem from the interdisciplinarity nature of data science and the diversity of its learners. More research on data science education should, therefore, be performed in the unique context of data science education.

### 6. Conclusion

Data science is emerging from the interdisciplinary integration of mathematics, statistics, computer science, and many other application domains such as business, transportation, biology, and education. In parallel to the emergence of data science, a new field is emerging as well—data science education. This article mapped the growing body of research on data science education and found five major themes: (a) curriculum, (b) pedagogy, (c) STEM skills, (d) domain adaptation, and (e) social aspects.

While the curricular aspects of data science draw most of the attention, research on pedagogical issues attracts less attention (Mike, 2020). Nevertheless, the large body of literature on data science education and the framework of research topics revealed here indicate the birth of a new scientific discipline—data science education.

### Disclosure Statement

This research was supported by a VATAT grant to the Technion’s Artificial Intelligence Hub (Tech.AI).
References


https://doi.org/10.1162/99608f92.1a99e67a


https://doi.org/10.1162/99608f92.38f16b68


## Appendix

### Databases and Search Queries

<table>
<thead>
<tr>
<th>Repository</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM-DL</td>
<td>(Title: “data science” OR Abstract: “data science”) AND (Title: education OR Title: curriculum OR Title: pedagogy OR Title: teach OR Abstract: education OR Abstract: curriculum OR Abstract: pedagogy OR Abstract: teach)</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>“data science” education</td>
</tr>
<tr>
<td></td>
<td>“data science” curriculum</td>
</tr>
<tr>
<td></td>
<td>“data science” pedagogy</td>
</tr>
<tr>
<td></td>
<td>“data science” teach</td>
</tr>
<tr>
<td>IEEE Xplore</td>
<td>((&quot;Document Title&quot;: &quot;data science&quot;) OR (&quot;Abstract&quot;: &quot;data science&quot;)) AND (&quot;Document Title&quot;: education) OR (&quot;Document Title&quot;: curriculum) OR (&quot;Document Title&quot;: pedagogy) OR (&quot;Document Title&quot;: teach) OR (&quot;Abstract&quot;: education) OR (&quot;Abstract&quot;: curriculum) OR (&quot;Abstract&quot;: pedagogy) OR (&quot;Abstract&quot;: teach))</td>
</tr>
<tr>
<td>Semantic Scholar</td>
<td>data science education</td>
</tr>
<tr>
<td></td>
<td>data science curriculum</td>
</tr>
<tr>
<td></td>
<td>data science pedagogy</td>
</tr>
<tr>
<td></td>
<td>data science teaching</td>
</tr>
<tr>
<td>Web of Science</td>
<td>(TI=&quot;data science&quot; AND (TI=education OR TI=curriculum OR TI=pedagogy OR TI=teach)) OR (AB=&quot;data science&quot; AND (AB=education OR AB=curriculum OR AB=pedagogy OR AB=teach))</td>
</tr>
</tbody>
</table>
References


Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources.


