A Systematic Literature Review of Undergraduate Data Science Education Research

Mine Dogucu*
Department of Statistics, University of California, Irvine
and
Sinem Demirci†
Statistics Department, California Polytechnic State University
and
Harry Bendekgey‡
Department of Computer Science, University of California, Irvine
and
Federica Zoe Ricci§
Department of Statistics, University of California, Irvine
and
Catalina M. Medina¶
Department of Statistics, University of California, Irvine

March 17, 2025

Abstract

The presence of data science has been profound in the scientific community in almost every discipline. An important part of the data science education expansion has been at the undergraduate level. We conducted a systematic literature

*Dogucu has been supported by NSF IIS award #2123366. Dogucu completed an earlier part of this work in Department of Statistical Science at University College London.

†Demirci has been supported by the Scientific and Technological Research Council of Türkiye. Demirci completed an earlier part of this work in Department of Statistical Science at University College London.

‡Bendekgey has been supported by the HPI Research Center in Machine Learning and Data Science at UC Irvine.

§Ricci has been supported by the HPI Research Center in Machine Learning and Data Science at UC Irvine.

¶Medina has been supported by NSF IIS award #2123366.

review to (1) portray current evidence and knowledge gaps in self-proclaimed undergraduate data science education research and (2) inform policymakers and the data science education community about what educators may encounter when searching for literature using the general keyword 'data science education.' While open-access publications that target a broader audience of data science educators and include multiple examples of data science programs and courses are a strength, substantial knowledge gaps remain. The undergraduate data science literature that we identified often lacks empirical data, research questions and reproducibility. Certain disciplines are less visible. We recommend that we should (1) cherish data science as an interdisciplinary field; (2) adopt a consistent set of keywords/terminology to ensure data science education literature is easily identifiable; (3) prioritize investments in empirical studies.

Keywords: data science curriculum, data science programs, data science courses, educational technology, open access

1 Introduction

The emergence of data science in the last few decades has given scientists a lot to talk and write about, accumulating a wealth of scientific literature. The presence of data science has been profound in the scientific community in almost every discipline. The National Science Foundation (NSF) called "Harnessing the Data Revolution" one of its 10 big ideas which cuts across all NSF Directorates (2017).

The demand for data science skills in academia and industry resulted in many higher education institutions developing courses (e.g., Baumer 2015) as well as degree programs (e.g., Glantz et al. 2023, Adhikari et al. 2021, Stern et al. 2021). With newer educational opportunities, newer educational research questions have been formed and thus data science education emerged as a field. For instance, the American Statistical Association's (ASA's), *Journal of Statistics Education* changed its name to *Journal of Statistics and Data Science Education (JSDSE)* (Witmer 2020).

Over the years, a major focus of the data science education community, unsurprisingly, has been at the undergraduate level. Different professional organizations and groups have tried to describe the data science competencies including but not limited to the Park City Math Institute report (De Veaux et al. 2017), the framework of the National Academies of Sciences, Engineering, and Medicine (2018), the computing competencies guidelines of the Association for Computing Machinery's (Danyluk et al. 2021), the accreditation for data science programs of the Accreditation Board for Engineering and Technology (ABET) (2024), and the EDISON project of the European Union (Wiktorski et al. 2017). It is also worth noting that the majority of data science programs are at the master's level and there is still a lot of room for growth at the undergraduate level (Li et al. 2021). Given the global

need for data science education, it is important for us to understand undergraduate data science education as well as the scientific literature on this topic.

Even though there is not a consensus on what data science is and which disciplines are part of it, a common view is that data science is interdisciplinary and statistics, computer science, mathematics and other domains contribute to it (Donoho 2017). The interdisciplinary structure of data science naturally gets reflected in educational research as well, making data science education an interdisciplinary field with contributors from statistics education, computer science education, and other educational research communities.

While contributions from different educational research communities can make data science education richer, these contributions may not easily circulate across disciplines in the absence of a centralized community. For instance, Hazzan and Mike (2021) mention that, despite its name, JSDSE mainly caters to the statistics community and state that "no journal exists today that deals exclusively with data science education, let alone highlights data science education from an interdisciplinary perspective".

The aforementioned examples of courses, programs and curricular guidelines as well as the existence of conferences and journals with the title or keywords *data science education* show the growth of this field. In this manuscript, we detail a study conducted to understand undergraduate data science education through in-depth readings of the existing literature. As undergraduate teacher-scholars, our focus was on data science education at the undergraduate level. In this study, we did not intend to define data science education research or judge whether a study meets a specific definition of data science. Instead, we focused on publications that claimed to be on "data science education".

Our goals were to (1) specify current evidence and knowledge gaps in self-proclaimed un-

dergraduate data science education and (2) inform policymakers and the data science education community about what educators may encounter when searching for literature using the general keyword 'data science education.' We conducted a systematic literature review (Evans & Benefield 2001, Liberati et al. 2009) using criteria that we detail in Section 2. In Section 3 we share our findings and then we discuss their implications in Section 4.

2 Data Collection and Analysis

The target population of this literature review was publications on data science education that directly address the undergraduate level. Due to variations in terminology, keywords, and language used by teacher-scholars across different data science fields, identifying the entire target population was not feasible. Since there is no consensus on what data science is, and consequently what data science education is, we did not evaluate whether publications met a certain definition of data science. Rather, we considered publications that self-proclaim to focus on data science education. Therefore, we opted to identify the accessible population for this study as publications that included "data science education" in quotes in the title, abstract, or keywords.

Figure 1 summarizes the stages of our data collection and analysis processes which led to the sample of this study. As shown in the diagram, we extracted data from six databases that potentially include publications on data science education. These databases were: (1) ERIC ProQuest, (2) IEEE Xplore, (3) PubMed, (4) ScienceDirect, (5) Scopus, and (6) Web of Science. These databases were selected because they cover a large number of publications and they include multiple disciplines that are commonly linked to data science education. For instance, ERIC ProQuest is known to mainly include education-focused publications, IEEE Xplore focuses on engineering and Web of Science is multidisciplinary-focused.

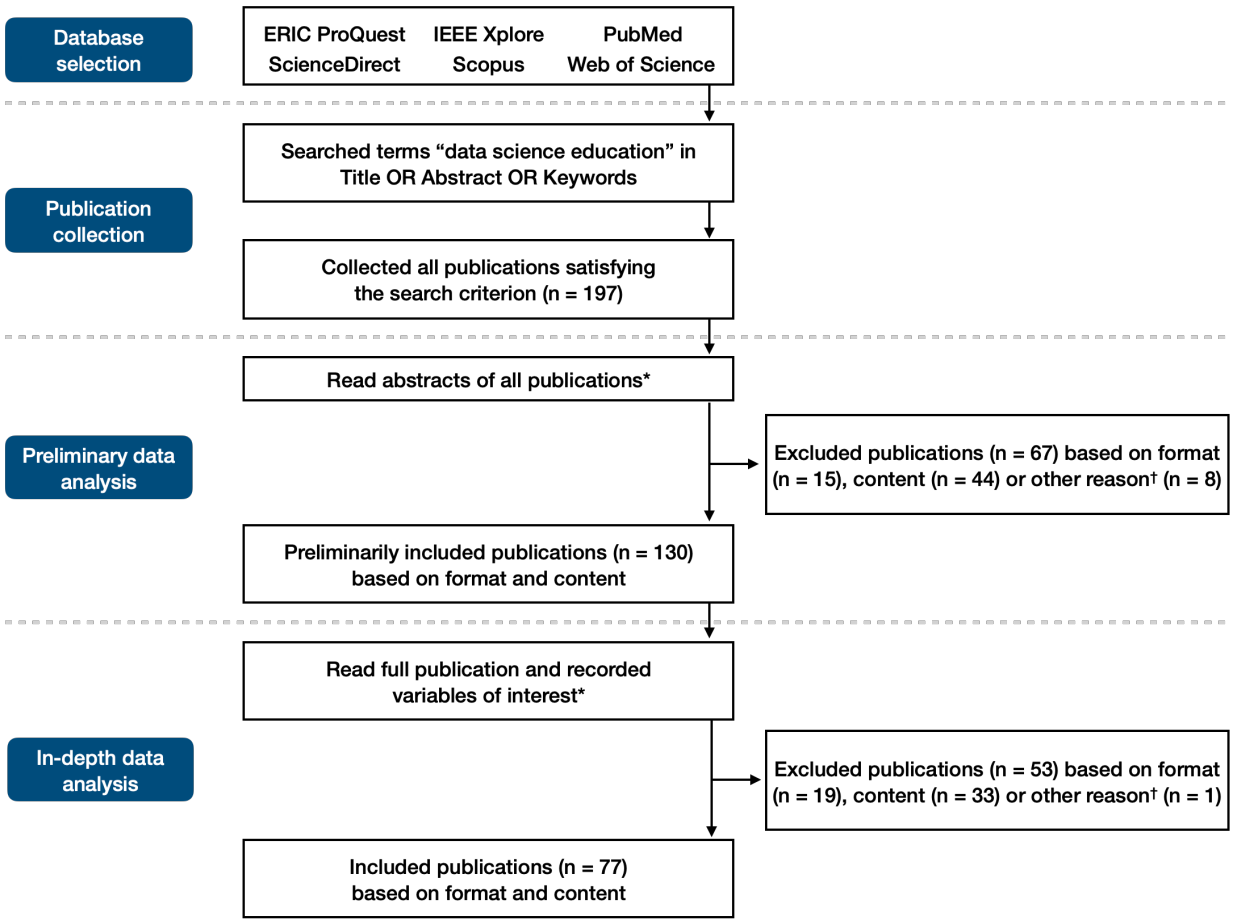

Figure 1: Flowchart of data collection and analysis. Publications were collected in December 2022.

From the selected databases, we collected publications including the term "data science education" (in quotes) in at least one of the following fields: title, abstract, keywords. We acknowledge that there are many other combinations of search terms that could result in publications related to data science education, including terms such as "data science courses", "data science pedagogy", "data science curriculum". One can generate numerous similar terms, with or without quotes. We specifically used the term "data science education" for its broadness and to set the scope of the research. We did not use "data science" and "education" separately to avoid publications in education that employ data science methods for their analyses.

Across the six databases, we found a total of 197 publications that met our search criteria. We extracted some variables including, but not limited to: author names; publication title; publication venue (e.g., journal title or conference title). We conducted the initial database search in December 2022, resulting in a pool of publications that were either published by that date or available online by that date but officially published in 2023.

The data analysis was conducted in two stages: 1) preliminary data analysis and 2) in-depth data analysis. Each publication was randomly assigned to two authors of the present manuscript. At both data-analysis stages, the two assigned reviewers first examined each publication independently and then discussed discrepancies between their analysis decisions, to reach a consensus. In cases where conflicts persisted, the entire group of five authors deliberated on the final decision.

During preliminary data analysis, we manually opened and read the abstracts of all publications. At this stage, we sought to exclude publications that did not meet our format and content criteria based on their abstracts. Specifically, we aimed to include journal/conference/magazine articles and book chapters (*format*) that focused on undergradu-

ate data science education (*content*) and were written in English. During in-depth analysis we examined the full publications and, upon confirming that they met our inclusion criteria, we recorded variables of potential interest. The number of publications that were excluded due to format, content or other reason, at each stage of our analysis, is shown in Figure 1.

Across both stages, we excluded 34 publications due to formatting (which included posters, panels, letters to journal editors and meeting highlights). Of the remaining 163 publications, 77 were excluded due to their content: 12 of them were not about data science education (including e.g., publications focused on data science methodology) and 65 of them focused exclusively on a different level of education than undergraduate. Among the latter excluded group, 18 publications focused on graduate level, 29 focused on K-12, middle school or high school and 18 publications focused on data science education for adults in non-academic programs (including, e.g., practitioners, citizen science and instructors). Publications that focused on both undergraduate and non-undergraduate levels were included. Of the remaining 86 publications, another 9 were excluded due to being: not written in English; duplicated in our dataset; retracted by their authors.

After excluding publications for the reasons detailed above, we were left with 77 publications, which we analyzed in-depth. In addition to the variables extracted from databases (e.g., title of the publication, author names, etc.) for each publication we collected data on:

- affiliation country of researchers
- open access status (i.e., whether the full publication is accessible for free from a Google Scholar search)

- year when the publication was first published online
- document type (conference article, journal article, magazine article or book chapter)
- whether there were explicit research questions stated in the publication
- whether there was any reporting of data collection in the publication and, if data were collected, the type of data (quantitative, qualitative or mixed)
- publication focus, that is a categorization of the subject matter of the publication (for example "pedagogical approach", "class activity" or "review of current state of data science education")
- discipline of the publishing source, determined by examining the call for contributions of the journal or conference, or the description of the book or magazine ("broad" when the publishing source called for contributions across all data science fields, otherwise a specific sub-field of data science, e.g., "computer science" or "statistics")
- the discipline of the target audience, as expressed by the authors in the publication ("broad" when the publication target were all data science educators, otherwise a specific sub-field of data science, e.g., "computer science" or "statistics")

3 Results

Publication years. Figure 2 shows the distribution of the years when publications in our review were made available online. Some of these publications were available during our database search in December 2022, but would ultimately be published in a 2023 journal edition (e.g., Pieterman-Bos & van Mil 2023, Schmitt et al. 2023). The body of literature that we identified is very recent, with the oldest paper available in 2015 and the volume of

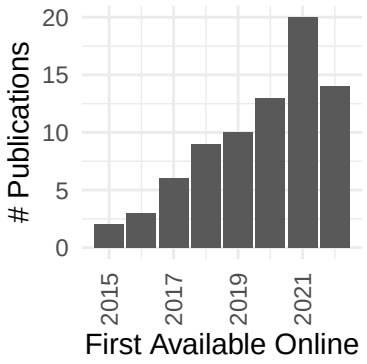

Figure 2: Undergraduate data science education publications ($n = 77$) by year they were first made available online. Note that initial data retrieval was in December 2022.

work increasing steadily in the following years.

Open access status and document types. A majority of work on undergraduate data science education published over these 8 years has been freely available to the public: the breakdown of articles' open access status is shown in Table 1 along with the document type. The vast majority of journal articles were open access, while conference publications showed a more even split. We reviewed only four publications which were neither journal nor conference articles, of which two were magazine articles (Hazzan & Mike 2021, Bonnell et al. 2022) and two were book chapters (Manzella & Emery 2022, Ryan 2016).

Table 1: Document type stratified by open access.

Document Type	Open Access	Not Open Access	Total
Conference Article	24	19	43 (56%)
Journal Article	26	4	30 (39%)
Book Chapter	0	2	2 (3%)
Magazine Article	1	1	2 (3%)
Total	51 (66%)	26 (34%)	77

Affiliation countries. We also investigated the geographic breakdown of institutions that are contributing to literature on undergraduate data science education. A majority of the publications analyzed included at least 1 author associated with an American institution (45 of 77), with European and Asian institutions providing the bulk of the remaining analyzed literature. Table 2 breaks down the analyzed publications by the institutional affiliations of their authors. The higher number of scholars' affiliations being in the United States may be at least partly attributed to the large size of the US relative to many other countries, as well as selection bias due to our search being restricted to English-written publications and to databases where US scholars might be more represented.

Table 2: For each country below, we report the number of publications which included at least one affiliated author. Affiliation country of authors was determined based on their home institutions as reported in the publication, and does not represent authors' nationalities.

Country	# Publications	Country	# Publications
United States of America	45	Australia	2
Netherlands	8	Egypt	2
Canada	5	Germany	2
Norway	5	Greece	1
United Kingdom	5	Hong Kong S.A.R.	1
Spain	4	Ireland	1
China	3	New Zealand	1
India	3	Portugal	1
Israel	3	Romania	1

Italy	3	Slovakia	1
Japan	3	Thailand	1
Switzerland	3	United Arab Emirates	1

Source and audience disciplines. Figure 3 shows the breakdown of articles by discipline. We report both the discipline of the publishing venue (such as the journal's or conference's discipline), and the audience that the publication stated (or implied) to target. About half of the articles (40/77) were written for the broad community of data science educators. The remaining articles were written for educators in particular disciplines, most commonly computer science (12 publications), statistics (5 publications), and information sciences (5 publications). We found that publication venues in all of these disciplines, as well as engineering, contributed articles aimed at their specific audience as well as articles aimed at the broad community of data science educators.

Research Question and Data Collection. Most publications (69 of 77 reviewed) either posed research questions *and* collected data or did neither, although there were exceptions. Four publications had research questions but no data collected (Vance et al. 2022, Pieterman-Bos & van Mil 2023, Hagen 2020, Robeva et al. 2020). On the other hand, four publications collected data but did not pose a specific research question or study goal (Hicks & Irizarry 2018, Rao et al. 2018, Liu & Wei 2020, Cuadrado-Gallego et al. 2021). Of the 40 publications that included collected data, 6, 9, and 25 publications had qualitative, quantitative, and mixed data respectively.

Publication focus. We also classified publications based on their focus. We found publications that reviewed the current state of data science education and provided guidelines. We also encountered publications that provided examples of programs, courses, class activ-

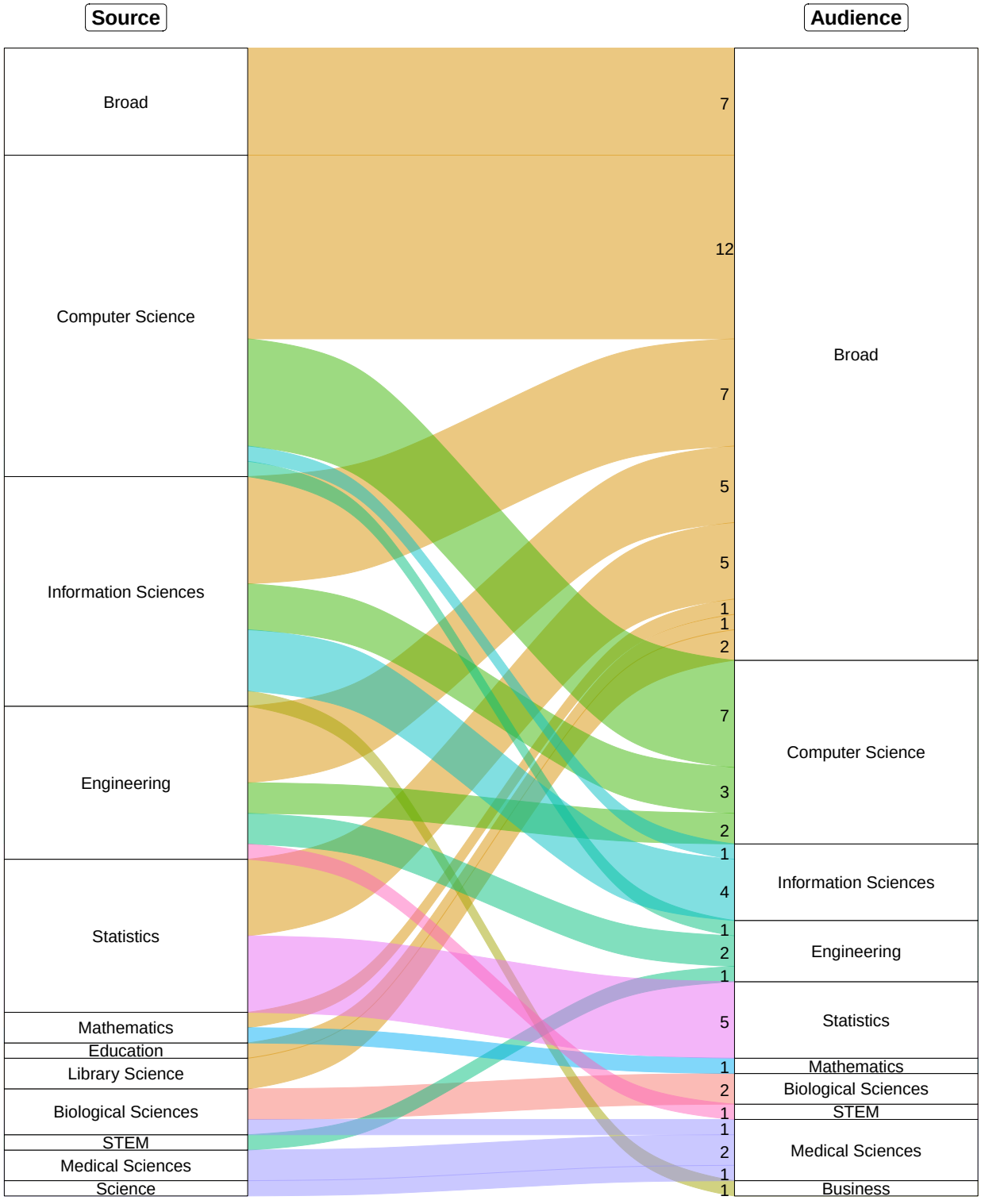

Figure 3: Composition of the discipline of targeted audiences of publications on data science education (right) by the discipline of the publishing source (left). A total of 2 publications were not included in the plot, one because it stated to target both statistics and computer science, the other one because it was not possible to determine its audience discipline.

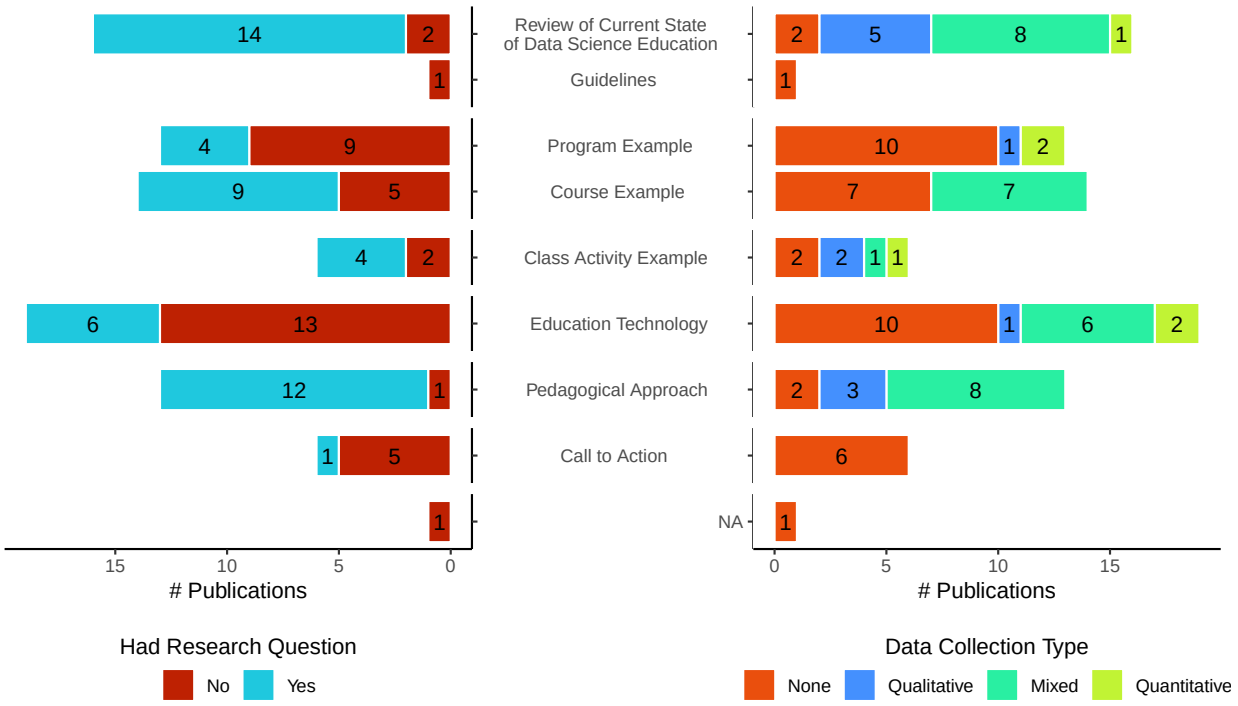

Figure 4: Number of publications with communicated research questions and with collected data stratified by publication focus.

ities, and extracurricular activities. Some publications specifically focused on educational technology or discussed a pedagogical approach. Some publications included more than one focus: in such cases, we noted down all the categories that the publication falls into, with the exception of *call to action*. Almost all publications we read were calling the scientific community to action and this is unsurprising for scientific publications. For instance, an educational technology paper might call to action to use a specific tool and a pedagogy paper might call to action to use a specific pedagogical approach. However, some publications solely made a call to action, without content falling into other publication focuses (e.g., education technology or course example). For these publications, we reserved the publication focus "call to action", that is, papers that we labeled in the "call to action" category did not fall in any other category for publication focus.

For each category of publication focus, Figure 4 shows how many publications featured that

publication focus, had a research question and included qualitative, quantitative or mixed type of data. Note that, in this figure, publication focuses are arranged and grouped to match the order in which we are going to illustrate them below, as we highlight contributions from each publication focus.

Similar to what we are doing in our study, many scholars wanted to understand the **current state of data science education**. Many of the studies conducted wanted to understand and evaluate data science programs (Wiktorski et al. 2016, Oliver & McNeil 2021, Song & Zhu 2016, Shao et al. 2021, Li et al. 2021, Raj et al. 2019) or curricula (Schmitt et al. 2023). Davis (2020) specifically looked at the current state of ethics in undergraduate data science education and Ceccucci et al. (2015) considered data science education from a scientific literacy perspective. Among these reviews, there were also comparative studies: for instance, Bile Hassan & Liu (2020)'s comparison of informatics and data science programs. Some studies also compared differences of data science approaches at country or regional levels such as in Japan (Takemura 2018), or in Middle Eastern countries (Zakaria 2023). In understanding the current state of data science education, scholars also wanted to understand the data science practice. For example, Belloum et al. (2019) developed semi-automated methods to determine the competencies needed in the job-market and Kross & Guo (2019) focused on understanding the skills that data science practitioners who teach data science in various settings pass onto their students. In addition to reviewing data science education, scholars also provided **guidelines** for data science degrees (Blair et al. 2021).

We found multiple data science **program examples**, including but not limited to general data science programs (e.g., Demchenko et al. 2017, 2019, Kakeshita et al. 2022). Some publications were about data science education in specific programs such as computer

science education (Bile Hassan & Liu 2020), microbiology (Dill-McFarland et al. 2021), information schools (Song & Zhu 2017, Hagen 2020) and business (Miah et al. 2020). Less traditional programs were also featured, such as the Data Mine which takes on data science education in a residential community of students (Betz et al. 2020).

The literature we read also included many **course examples** with some clever ways of including data science concepts in different courses. For instance, Fisler (2022) suggested to include data science and data structures in an introductory computing course, Rao et al. (2019) teach data science through the use of education data in their engineering course on modern technologies, and Haynes et al. (2019) teach data science in a general education IT course. Some institutions developed data science courses for specific disciplines including information schools (Hagen 2020) and medicine (Doudesis & Manataki 2022). Some scholars also described courses that merge data science with philosophy of science (Pieterman-Bos & van Mil 2023) or with humanities perspectives (Vance et al. 2022). Last but not least, data science educators also tried to provide real-life data science experiences through work-integrated learning (Bilgin et al. 2022), capstone projects (Allen 2021) and entirely case-study based courses (Hicks & Irizarry 2018).

Similar to program examples and course examples provided in the literature, some publications focused on activity examples. Among **class activity examples**, Yamamoto et al. (2021) developed a programming exercise to bring higher-order tensor content to undergraduate level by using a 3-D puzzle. Another example is by Banadaki (2022) who developed different activities that include applications of data science in Mechanical Engineering, Biomedical science, Quantum Physics, and Cybersecurity. In data science education, there is also room for learning outside the classroom as an **extra-curricular activity**. This included data hackathons (Anslow et al. 2016).

In terms of **education technology**, as much of data science education tackles with issues such as computing power, storage of large datasets, and sometimes automation, some of these publications focused on cloud-based data science platforms for teaching purposes (e.g., Demchenko et al. 2017). Many others focused on tools such as online platforms, learning environments, and apps that support learning data science (e.g., Hoyt & Wangia-Anderson 2018, Bornschlegl 2016, Liu & Wei 2020, Nair et al. 2020).

Studies that focused on **pedagogical approaches** were also frequently encountered. These studies focused on various subtopics including team-based learning (e.g., Vance et al. 2022), project-based learning (e.g., Mike et al. 2020) and large-scale product development which has both project- and team-based learning aspects (e.g., Bhavya et al. 2020). Scholars also studied social topics such as ethics and equity in the data science classroom (Alexander et al. 2022) and student self-regulation (Zhang & Wu 2020).

Almost all publications we read were calling the scientific community to action. For instance, through a systematic review, Davis (2020) **called the community to action** to include ethics in undergraduate data science education. We also encountered publications with an important message about data science education but without a review of programs, courses, activities etc. For instance, Robeva et al. (2020) argued for the inclusion of data science in quantitative biology education. Engel (2017) drew attention to the importance of statistical literacy in data science education. We have also seen a to-do list (i.e., action list) for the community including starting a multidisciplinary data science education journal (Hazzan & Mike 2021), building a consensus on data science education and curricula (Dinov 2019), developing and deploying degree programs for university students, having basic data science training for university students, and training instructors to teach data science (Bonnell et al. 2022).

4 Discussion

Turning our attention to the knowledge that was gathered and summarized in our study, this section examines the strengths and gaps in the literature authored by researchers who self-proclaim to focus on undergraduate data science education. Building upon the findings detailed in Section 3, we also extend comprehensive recommendations to both policymakers and the data science education community, aiming to enhance capacity building in undergraduate data science education studies.

Before proceeding, we believe that it is essential to address the limitations (1) inherent in the field of data science and (2) related to the scope of this study. Acknowledging these constraints provides crucial context and helps avoid overgeneralization, by enabling readers to more effectively contextualize the strengths, gaps, and recommendations presented in this section.

4.1 Inherent Limitations in the Field of Data Science (Education)

Data science is an interdisciplinary (Cao 2017) and relatively young field. While its emergent nature brings numerous opportunities, it also introduces inherent limitations that can influence the accumulation of knowledge in this field and, consequently, the findings of this systematic literature review.

Perhaps the most critical limitation is the lack of consensus on the definition of data science (Hazzan & Mike 2023). Although we acknowledge that the absence of a standard definition is not unique to the field of data science, this lack of agreement at such an early stage, especially in a field that is so interdisciplinary, may result in a deeper issue: the challenge of generating an *identifiable* cumulative body of knowledge in this field.

Indeed, a corollary of this definitional ambiguity is the difficulty in labeling research as data science education. Some studies may be considered data science education by certain scholars, but lack explicit labeling as such, while others that we included might not be categorized as data science education under different interpretations. Therefore, capturing all undergraduate data science education research papers is extremely challenging at this stage.

Arguably, the lack of clear identifiability and consistent self-labeling creates at least two additional challenges. First, it restricts practitioners' access to findings that could inform and enhance their data science teaching practices. Second, it may hinder researchers from identifying recurring patterns and comprehensively understanding the current state of data science education research.

In this study, we deliberately refrained from defining data science and data science education. Instead, we used "data science education" as a keyword to capture publications that self-proclaim as undergraduate data science education research, across all disciplines. We opted for this approach because we did not want to take the role of gate-keepers by assessing whether publications adhered to any specific definition of data science. However, this choice may lead to an *incomplete picture* of the undergraduate "data science education" literature, as we discuss in the subsequent section.

4.2 Limitations of the Scope of This Study

We acknowledge that, in this study, we did not reach the entire target population of research publications on undergraduate data science education. In addition to the inherent limitations of the field outlined above, methodological limitations may also contribute to this incomplete picture of undergraduate data science education literature.

One key methodological limitation is the use of a single term, "data science education", for our search. While the choice of this term was aimed at capturing a broad range of self-proclaimed undergraduate data science education research, it may have restricted the results. For instance, an article that includes terms such as "data science class", "data science activity", or "data science curriculum" but does not use the keyword "data science education" would be excluded from this study (e.g., Baumer et al. 2022). Definitional ambiguities and self-labeling challenges in the field may exacerbate this issue.

Our review may have missed relevant publications due to several other factors, including publications written in languages other than English and publications that are not included in the databases that we searched (e.g., Finzer 2013).

Despite these challenges, we believe our study offers valuable insights into the strengths and knowledge gaps within self-proclaimed undergraduate data science education research. It provides a realistic portrayal of what data science educators might encounter when searching for literature using the general keyword "data science education." Furthermore, it offers insights that can benefit the broader data science education community, including those whose work does not explicitly label itself as data science education.

4.3 Strengths in Undergraduate Data Science Education

The majority of published studies on undergraduate data science education are open access, marking a substantial strength in the field. The freely available information and scholars' insights regarding the status of undergraduate data science education not only contribute to overcoming barriers, but also facilitate the dissemination and application of knowledge. Over the eight years included in our analysis, a higher percentage of conference articles (56%) have been published compared to journal articles (39%). Education technology is

the publication focus that is studied most by scholars. Among educational technologies to support learning data science, learning environments (e.g., Bornschlegl 2016, Hoyt & Wangia-Anderson 2018) are one of the popular ones. Ethics is also one of the recurring themes among the studies that we reviewed (e.g., Davis 2020, Shapiro et al. 2020, Alexander et al. 2022).

As shown in Figure 3, the majority of self-proclaimed undergraduate data science publications are directed to the broad audience of data science educators generally. Specifically, 40 out of 77 publications target the undergraduate data science education community as a whole, rather than appealing to a specific discipline in data science (e.g., computer science) or a subset of disciplines in data science (e.g., STEM). This inclusivity can be seen as a potential strength, as the insights from these broader publications provide valuable perspectives that scholars and researchers can adapt and apply across diverse contexts in data science to enhance teaching practices, develop curricula, and foster a more comprehensive understanding of data science education.

Studies including multiple examples of data science programs and courses across various fields add another layer of strength to the undergraduate data science education literature. In addition to overall data science programs (e.g., Demchenko et al. 2019, Kakeshita et al. 2022), we have also seen data science education practices in different programs such as computer science education (Bile Hassan & Liu 2020), microbiology (Dill-McFarland et al. 2021), and business (Miah et al. 2020).

The course examples showcase a diverse array of strategies for incorporating data science concepts, ranging from introductory computing (Fisler 2022), modern technologies for computer science and engineering students (Rao et al. 2019) and general education IT (Information Technology) (Haynes et al. 2019), to medicine (Doudesis & Manataki 2022) and

introduction to psychological statistics (Tucker et al. 2023). This broad spectrum in both program and course examples serves as compelling evidence, illustrating that the intrinsic interdisciplinary nature of data science attracts the attention of scholars from diverse fields.

4.4 Knowledge Gaps in Undergraduate Data Science Education

While the strengths of the current data science education literature, as outlined above, are evident, our study also reveals some knowledge gaps. Knowledge gaps are areas or topics derived from the synthesis of an existing body of literature (Cooper 1998). Understanding these gaps is crucial because it adds a more structured and evidence-supported layer to our knowledge. In this section, we discuss some knowledge gaps in self-proclaimed undergraduate data science education.

Knowledge gap 1: Certain disciplines in data science are less visible in the current body of literature.

Despite the lack of consensus on defining what exactly data science is, there is agreement that data science encompasses various disciplines, including statistics, mathematics, computer science, and other relevant domains, as defined by many interpretations (e.g., Cao 2017, Hazzan & Mike 2023). However, these disciplines are not equally represented among the self-proclaimed data science education publications that we examined.

In terms of source discipline (i.e., the main discipline of a journal, conference, book chapter etc.) we have seen many publications published in venues related to computer science, information science, engineering, and statistics. The difference in quantity of publications published in sources related to computer science and statistics is worth remembering from Figure 3. This result partly aligns with the study of Wiktorski et al. (2017), who reported that mathematics and statistics departments are not at the forefront of data science. This

result can also potentially be explained by the fact that different disciplines have different publication rates in general. For instance, in the *Science and Technology Indicators* report published by NSF it is stated that 2.3% of the articles published in 2016 is in the field of mathematics and 8.3% are in computer sciences (2018).

We have seen even fewer publications in venues related to domain sciences. These were mainly related to education, biological sciences, library science, and medical sciences. The findings may indicate a need to have journals and conferences in other domain sciences (e.g., astronomy, economics, psychology) that provide more opportunities for disseminating works of data science education researchers to a broader data science community. It is also possible that such journals and conferences for domain sciences might exist, but the keyword "data science education" is not used in these venues. For instance, economists may continue to use the term "econometrics education" rather than "data science education".

Among source disciplines, we have also identified publication venues for the broader data science community. This, however, mainly consisted of conferences. The two journals that were identified as broad were *International Journal of Data Science and Analytics* and *Foundations of Data Science*. Both of these journals focus on data science while allowing for education-related publications, but neither of these journals focuses specifically on data science education. These findings reiterate the importance of the call made by Hazzan and Mike (2021) on having an interdisciplinary journal on data science education.

In terms of audience disciplines, scholars write for the broader data science education community as well as specific disciplines. Representation of both broad and specific disciplines is important and can enrich the data science education research. There are fewer publications written for the audiences in domain sciences. This might indicate further need for data science education research targeting these audiences, or this might be due to under-

utilization of the keyword "data science education" in these disciplines.

Statistics is underrepresented both as source and audience discipline in the body of literature that we reviewed. Arguably, statisticians are contributing to data science education research more than this, but their work was not prominently captured in this study. Several explanations may account for this finding. Statisticians might avoid using the term "data science" in their work, perceiving their research as strictly related to statistics and targeting their primary audience within that domain. Another possibility is that they consider data science as a subset or natural extension of statistics, and thus do not see the need to explicitly label their work as data science education research. Lastly, they might be using some other keywords in their title, abstract, and keywords instead of "data science education", making it less visible to the broader data science education community.

This limited visibility of statisticians in the current literature underscores the insufficiency of the existing body of data science education research. Most importantly, it highlights a critical gap: the lack of visible perspectives from statisticians in this identified body of literature. As statistics is one of the foundational disciplines in data science, its underrepresentation not only diminishes the diversity of insights but may also impede the development of comprehensive, interdisciplinary approaches to data science education.

Although statistics is underrepresented both as source and audience discipline in comparison to computer science, it is worth noting that both computer science and statistics sources write for broad data science education audiences as well as their corresponding specific discipline audiences. For foundational sciences of data science, such as statistics and computer science, writing for the broader data science community as well as the specific discipline is important and can continue to enrich data science education research in the years to come.

Knowledge gap 2: Within the identified body of data science education literature, there is lack of empirical data and identifiable research questions.

Research questions and data are two common elements in empirical research. Data for empirical research include both qualitative and quantitative approaches where the researchers collect 'observable information about or direct experience of the world'. And perhaps most importantly, empirical data are not just stored as in numbers but also words and categories (Punch & Oancea 2014).

One of the important functions of systematic literature reviews is to gain a deeper understanding and inform possible further research avenues that can be conducted in the field (Evans & Benefield 2001, Liberati et al. 2009). To facilitate this, we categorized the studies based on their publication focus, the existence of research questions, and the types of data collection. As stated earlier, 37 studies out of 77 did not collect data. Given the scopes of publication focus such as calls to action, educational technology, and program examples coupled with the emergent nature of the field, the lack of empirical data is not a surprising finding. However, it also suggests that undergraduate data science education researchers have not yet begun to systematically collect empirical data to assess, for example, the effectiveness of educational technologies, programs, learning outcomes and/or other pedagogical approaches.

A lack of empirical data could impede the development of systematic literature reviews or meta-analyses on a specific publication focus (e.g., educational technology), which are essential for identifying trends, studied variables, and recurring patterns in undergraduate data science education through qualitative and/or quantitative approaches. It is essential to clarify that our emphasis on lack of empirical data is not a promotion of empiricism over all other 'ways of knowing'. We acknowledge and appreciate alternative forms of knowledge,

such as expert opinions (Fraenkel et al. 2012), for their valuable contributions to the data science education community's know-how. These forms of knowledge are important catalysts in guiding researchers towards areas that require systematic data collection. What we are highlighting is the disproportionately high percentage of studies lacking empirical data and identifiable research questions, which complicates the literature's potential for gaining a deeper understanding and identifying recurring patterns.

Knowledge gap 3: Reproducibility is one of the potential challenges in undergraduate data science education research.

The corollary of a lack of empirical data and identifiable research questions may introduce another challenge: the reproducibility of certain studies. We speculate that the absence of critical information about research designs, such as the lack of research questions and non-collection of data, may contribute to the reduced reproducibility of available studies. This makes it challenging to replicate or modify research, impeding the identification of recurring patterns.

Lack of reproducibility is not unique to data science education research. Importance of reproducibility and lack thereof have been discussed in many disciplines including physics (Junk & Lyons 2020), economics (Chang & Li 2015), and psychology (Open Science Collaboration 2015). In recent years, this has even been referred to as the *reproducibility crisis*.

Potential explanations for the lack of reproducibility in data science education research might be similar to reasons seen in the broader science community. Namely, word limitations can result in lack of detailed information on research design and data collection in publications (Bausell 2021).

Another explanation might simply be the "publish or perish" culture in academic settings. There are even academics who publish a paper every five days (Ioannidis et al. 2018). Even if not at this rate, many academics might feel under pressure to get publications out, without having much time to focus on the reproducibility of their work.

Reproducibility is an important skill for teacher-scholars of data science both in their teaching and research (Dogucu & Çetinkaya-Rundel 2022, Dogucu 2024). Considering that much of the published research in the literature are written by those who also teach data science, closing the reproducibility gap both in research and teaching is extremely important. One potential reason for this gap may also be the minimal training that most instructors receive in reproducibility (Horton et al. 2022).

4.5 Recommendations

Considering the findings and our arguments, we present three recommendations for the future of undergraduate data science education studies: one for policymakers and funding agencies, and two for institutions and scholars whose research focus includes data science education.

Recommendation 1: Cherish data science as an interdisciplinary field.

Despite the existence of other key fields offering data science courses at the undergraduate level, there is a noticeable gap in studies reflecting their perspectives in data science education. To address this gap, it is imperative for undergraduate data science education studies to incorporate the viewpoints of scholars from foundational disciplines such as statistics, mathematics, and other application domains intersecting with data science education practices. We need to encourage scholars in all data science fields to maintain their visibility and contribute more to publications. We posit that future endeavors in this direction will

substantially enhance our understanding of the strengths and needs in undergraduate data science education.

As statistics community, we must take an active role in expanding the data science education research and other opportunities. For instance, despite having Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report (2016) and Curriculum Guidelines for Undergraduate Programs in Statistical Science (CGUPSS) (2014), ASA has not yet written any guidelines specific to data science education. However, ASA Board of Directors endorsed the Park City Math Institute report (De Veaux et al. 2017) and have provided input on the criteria for the data science program ABET accreditation (Liu 2022). At the time of writing of this manuscript, ASA is working to update GAISE. Further efforts by ASA similar to GAISE and CGUPSS or newer versions of these documents can also help distinguish statistics and data science programs and courses.

Recommendation 2: Adopt a consistent set of keywords/terminology to ensure data science education literature is easily identifiable. While there is no doubt that all data science fields are contributing to data science education research, some of their work was not prominently captured in this study. Speculatively, it appears that the keyword "data science education" is more commonly embraced by certain fields compared to others. As a result, a researcher broadly interested in data science education may not see the interdisciplinary diversity of insights within the field, if there is a lack of consistent use of keywords and terminology.

We believe that using both broad, cross-disciplinary terms and specialized terminology unique to each field/domain is essential for facilitating communication within individual disciplines/domains and across the wider data science community. One way this can possibly be achieved is by using "data science education" as a keyword across different disciplines

while also specifying discipline such as “nursing”, “sociology”, etc. Adopting a consistent set of keywords does not require a shared definition of what is and is not data science. If a set of keywords is consistently used by data science education researchers, it may facilitate the accumulation of collective knowledge and help identify what works or does not in the field. Therefore, it is important to standardize terminology to enhance the accessibility and comprehensiveness of the data science education literature.

Recommendation 3: Prioritize investments in empirical studies. Not having sufficient empirical data is one of the knowledge gaps in undergraduate data science education research. Accumulating empirical data is essential to be able to gain a sound understanding of data science education studies at a large scale. Hence, we recommend that policymakers and funding agencies prioritize investments in undergraduate data science studies dedicated to systematic data collection. Additionally, directing investments toward empirical studies in application domains of data science that were underrepresented in our study, such as astronomy, psychology, and economics could also help provide a more complete picture of data science education in the long run.

These strategic approaches will enable a comprehensive assessment of foundational disciplines and various application domains of data science, including the effectiveness of educational technologies, program impact, learning outcomes, and other pedagogical approaches, ultimately contributing to a more informed and robust understanding of data science education.

4.6 Closing Remarks

In summary, the results of this study show that data science education is an emerging field with much more room for growth. Scientific studies are an integral part of reviewing exist-

ing practices as well as of improving higher education institutions' data science practices. Therefore, we should diversify our research efforts by investing in more empirical studies and fostering scholars from key fields in data science, especially in statistics and domain sciences.

Further research studies may improve or try to replicate the findings of this study in multiple ways by utilizing different keywords, databases, and languages. An even larger endeavor can be undertaken if investigators want to work off on a definition of data science utilizing a systematic literature review.

Lastly, we also believe that the data collected as part of this study can help novice researchers in data science education find inspiration and examples from the existing literature. We presented findings of the study at the Electronic Conference on Teaching Statistics 2024 where we had participants design their own research study by grouping them according to their publication focus (e.g., educational technology) of interest. Readers interested in pursuing research in data science education may utilize our dataset to create a reading list for their research agenda.

Data Availability Statement

The data on all the publications and the associated codebook for the variables are publicly available in a GitHub repository at https://github.com/mdogucu/comp-data-sci and an OSF project at https://osf.io/b3u7y/.

References

ABET (2024), 'Criteria for Accrediting Applied and Natural Science Programs, 2024-2025'.

URL: *https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-applied-and-natural-science-programs-2024-2025/*

Adhikari, A., DeNero, J. & Jordan, M. (2021), 'Interleaving Computational and Inferential Thinking in an Undergraduate Data Science Curriculum', *Harvard Data Science Review* **3**(2).

URL: *https://doi.org/10.1162/99608f92.cb0fa8d2*

Alexander, N., Eaton, C. D., Shrout, A., Tsinnajinnie, B. & Tsosie, K. (2022), 'Beyond Ethics: Considerations for Centering Equity-Minded Data Science', *Journal of Humanistic Mathematics* **12**(2), 254–300.

URL: *https://doi.org/10.5642/jhummath.OCYS6929*

Allen, G. I. (2021), Experiential Learning in Data Science: Developing an Interdisciplinary, Client-Sponsored Capstone Program, *in* 'SIGCSE '21', p. 516–522.

URL: *https://doi.org/10.1145/3408877.3432536*

American Statistical Association (2014), 'Curriculum Guidelines for Undergraduate Programs in Statistical Science'.

URL: *https://www.amstat.org/docs/default-source/amstat-documents/edu-guidelines2014-11-15.pdf*

Anslow, C., Brosz, J., Maurer, F. & Boyes, M. (2016), Datathons: An Experience Report of Data Hackathons for Data Science Education, *in* 'SIGCSE '16', p. 615–620.

URL: *https://doi.org/10.1145/2839509.2844568*

Banadaki, Y. (2022), Enabling Data Science Education in STEM Disciplines through Supervised Undergraduate Research Experiences, *in* '2022 ASEE Annual Conference and Exposition'.

Baumer, B. (2015), 'A Data Science Course for Undergraduates: Thinking with Data', *The American Statistician* **69**(4), 334–342.

URL: *https://doi.org/10.1080/00031305.2015.1081105*

Baumer, B. S., Garcia, R. L., Kim, A. Y., Kinnaird, K. M. & Ott, M. Q. (2022), 'Integrating data science ethics into an undergraduate major: A case study', *Journal of Statistics and Data Science Education* **30**(1), 15–28.

URL: *https://doi.org/10.1080/26939169.2022.2038041*

Bausell, R. B. (2021), *Publishing Issues and Their Impact on Reproducibility*, Oxford University Press, p. 0. DOI: 10.1093/oso/9780197536537.003.0010.

URL: *https://doi.org/10.1093/oso/9780197536537.003.0010*

Belloum, A. S., Koulouzis, S., Wiktorski, T. & Manieri, A. (2019), 'Bridging the Demand and the Offer in Data Science', *Concurrency and Computation: Practice and Experience* **31**(17), e5200.

URL: *https://doi.org/10.1002/cpe.5200*

Betz, M., Gundlach, E., Hillery, E., Rickus, J. & Ward, M. D. (2020), 'The Next Wave: We Will All Be Data Scientists', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **13**(6), 544–547.

URL: *https://doi.org/10.1002/sam.11476*

Bhavya, B., Boughoula, A., Green, A. & Zhai, C. (2020), Collective Development of Large Scale Data Science Products via Modularized Assignments: An Experience Report, *in* 'SIGCSE '20', Association for Computing Machinery, p. 1200–1206.

Bile Hassan, I. & Liu, J. (2020), A Comparative Study of the Academic Programs between Informatics BioInformatics and Data Science in the U.S, *in* '2020 IEEE 44th Annual

Computers, Software, and Applications Conference (COMPSAC)', pp. 165–171.

URL: *https://doi.org/10.1109/COMPSAC48688.2020.00030*

Bilgin, A. A. B., Powell, A. & Richards, D. (2022), 'Work Integrated Learning in Data Science and a Proposed Assessment Framework', *Statistics Education Research Journal* **21**(2), 12–12.

URL: *https://doi.org/10.52041/serj.v21i2.26*

Blair, J. R. S., Jones, L., Leidig, P., Murray, S., Raj, R. K. & Romanowski, C. J. (2021), Establishing ABET Accreditation Criteria for Data Science, *in* 'SIGCSE '21', p. 535–540.

URL: *https://doi.org/10.1145/3408877.3432445*

Bonnell, J., Ogihara, M. & Yesha, Y. (2022), 'Challenges and Issues in Data Science Education', *Computer* **55**(2), 63–66. Publisher: IEEE.

URL: *https://doi.ieeecomputersociety.org/10.1109/MC.2021.3128734*

Bornschlegl, M. X. (2016), IVIS4BigData: Qualitative Evaluation of an Information Visualization Reference Model Supporting Big Data Analysis in Virtual Research Environments, *in* M. X. Bornschlegl, F. C. Engel, R. Bond & M. L. Hemmje, eds, 'Advanced Visual Interfaces. Supporting Big Data Applications: AVI 2016 Workshop', Lecture Notes in Computer Science, Cham, pp. 127–142.

URL: *https://doi.org/10.1007/978-3-319-50070-6_10*

Cao, L. (2017), 'Data Science: a Comprehensive Overview', *ACM Computing Surveys (CSUR)* **50**(3), 1–42.

URL: *https://doi.org/10.1145/3076253*

Ceccucci, W., Tamarkin, D. & Jones, K. (2015), 'The Effectiveness of Data Science as a means to achieve Proficiency in Scientific Literacy', *Information Systems Education*

Journal **13**(4), 64.

URL: *https://www.isedj.org/2015-13/n4/ISEDJv13n4p64.html*

Chang, A. C. & Li, P. (2015), 'Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'', *FEDS Working Paper No. 2015-083* .

URL: *https://dx.doi.org/10.17016/FEDS.2015.083*

Cooper, H. M. (1998), *Synthesizing Research: A Guide for Literature Reviews*, Vol. 2, Sage.

Cuadrado-Gallego, J. J., Demchenko, Y., Losada, M. A. & Ormandjieva, O. (2021), Classification and Analysis of Techniques and Tools for Data Visualization Teaching, *in* '2021 IEEE Global Engineering Education Conference (EDUCON)', pp. 1593–1599. ISSN: 2165-9567.

URL: *https://.doi.org/10.1109/EDUCON46332.2021.9453917*

Danyluk, A., Leidig, P., McGettrick, A., Cassel, L., Doyle, M., Servin, C., Schmitt, K. & Stefik, A. (2021), Computing Competencies for Undergraduate Data Science Programs: An ACM task force final report, *in* 'Proceedings of the 52nd ACM Technical Symposium on Computer Science Education', pp. 1119–1120.

URL: *https://doi.org/10.1145/3453538*

Davis, K. C. (2020), Ethics in Data Science Education, *in* '2020 ASEE Virtual Annual Conference Content Access'.

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R. et al. (2017), 'Curriculum Guidelines for Undergraduate Programs in Data Science', *Annual Review of Statistics and Its Application* **4**, 15–30.

URL: *https://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf*

Demchenko, Y., Belloum, A., de Laat, C., Loomis, C., Wiktorski, T. & Spekschoor, E. (2017), Customisable Data Science Educational Environment: From Competences Management and Curriculum Design to Virtual Labs On-Demand, *in* '2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)', pp. 363–368.

URL: *https://doi.org/10.1109/CloudCom.2017.59*

Demchenko, Y., Comminiello, L. & Reali, G. (2019), Designing Customisable Data Science Curriculum Using Ontology for Data Science Competences and Body of Knowledge, *in* 'ICBDE '19', p. 124–128.

URL: *https://doi.org/10.1145/3322134.3322143*

Dill-McFarland, K. A., König, S. G., Mazel, F., Oliver, D. C., McEwen, L. M., Hong, K. Y. & Hallam, S. J. (2021), 'An Integrated, Modular Approach to Data Science Education in Microbiology', *PLOS Computational Biology* **17**(2), e1008661.

URL: *https://doi.org/10.1371/journal.pcbi.1008661*

Dinov, I. D. (2019), 'Quant Data Science meets Dexterous Artistry', *International Journal of Data Science and Analytics* **7**(2), 81–86. PMID: 30923735 PMCID: PMC6433171.

URL: *https://doi.org/10.1007/s41060-018-0138-6*

Dogucu, M. (2024), 'Reproducibility in the Classroom'. Publisher: Annual Reviews.

URL: *https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-112723-034436*

Dogucu, M. & Çetinkaya-Rundel, M. (2022), 'Tools and Recommendations for Reproducible Teaching', *Journal of Statistics and Data Science Education* **30**(3), 251–260.

URL: *https://doi.org/10.1080/26939169.2022.2138645*

Donoho, D. (2017), '50 Years of Data Science', *Journal of Computational and Graphical Statistics* **26**(4), 745–766.

URL: *10.1080/10618600.2017.1384734*

Doudesis, D. & Manataki, A. (2022), 'Data Science in Undergraduate Medicine: Course Overview and Student Perspectives', *International Journal of Medical Informatics* **159**, 104668. PMID: 35033982.

URL: *https://doi.org/10.1016/j.ijmedinf.2021.104668*

Engel, J. (2017), 'Statistical Literacy for Active Citizenship: A Call for Data Science Education', *Statistics Education Research Journal* **16**(1), 44–49.

URL: *https://doi.org/10.52041/serj.v16i1.213*

Evans, J. & Benefield, P. (2001), 'Systematic Reviews of Educational Research: Does the Medical Model Fit?', *British Educational Research Journal* **27**(5), 527–541.

URL: *https://doi.org/10.1080/01411920120095717*

Finzer, W. (2013), 'The data science education dilemma', *Technology Innovations in Statistics Education* **7**(2).

URL: *https://doi.org/10.5070/T572013891*

Fisler, K. (2022), Data-Centricity: Rethinking Introductory Computing to Support Data Science, *in* '1st International Workshop on Data Systems Education', DataEd '22, Association for Computing Machinery, p. 1–3.

URL: *https://doi.org/10.1145/3531072.3535317*

Fraenkel, J., Wallen, N. & Hyun, H. (2012), *How to Design and Evaluate Research in Education (8th) ed.)*, McGraw-Hill.

GAISE (2016), 'Guidelines for Assessment and Instruction in Statistics Education

(GAISE): College report'.

URL: *http://www.amstat.org/education/gaise*

Glantz, M., Johnson, J., Macy, M., Nunez, J., Saidi, R. & Velez Ramirez, C. (2023), 'Students' Experience and Perspective of a Data Science Program in a Two-Year College', *Journal of Statistics and Data Science Education* **31**(3), 248–257.

URL: *https://doi.org/10.1080/26939169.2023.2208185*

Hagen, L. (2020), 'Teaching Undergraduate Data Science for Information Schools', *Education for Information* **36**(2), 109–117.

URL: *https://doi.org/10.3233/EFI-200372*

Haynes, M., Groen, J., Sturzinger, E., Zhu, D., Shafer, J. & McGee, T. (2019), Integrating Data Science into a General Education Information Technology Course: An Approach to Developing Data Savvy Undergraduates, *in* 'SIGITE '19', p. 183–188.

URL: *https://doi.org/10.1145/3349266.3351417*

Hazzan, O. & Mike, K. (2021), 'A Journal for Interdisciplinary Data Science Education', *Communications of the ACM* **64**(8), 10–11. Publisher: ACM New York, NY, USA.

URL: *https://doi.org/10.1145/3469281*

Hazzan, O. & Mike, K. (2023), What is Data Science?, *in* 'Guide to Teaching Data Science: An Interdisciplinary Approach', Springer, pp. 19–34.

URL: *https://doi.org/10.1145/3575663*

Hicks, S. C. & Irizarry, R. A. (2018), 'A Guide to Teaching Data Science', *The American Statistician* **72**(4), 382–391.

URL: *https://doi.org/10.1080/00031305.2017.1356747*

Horton, N. J., Alexander, R., Parker, M., Piekut, A. & Rundel, C. (2022), 'The Grow-

ing Importance of Reproducibility and Responsible Workflow in the Data Science and Statistics Curriculum'.

URL: *https://doi.org/10.1080/26939169.2022.2141001*

Hoyt, R. & Wangia-Anderson, V. (2018), 'An Overview of Two Open Interactive Computing Environments Useful for Data Science Education', *JAMIA Open* **1**(2), 159–165.

URL: *https://doi.org/10.1093/jamiaopen/ooy040*

Ioannidis, J. P. A., Klavans, R. & Boyack, K. W. (2018), 'Thousands of Scientists Publish a Paper Every Five Days', *Nature* **561**(7722), 167–169. Bandiera_abtest: a Cg_type: Comment Publisher: Nature Publishing Group Subject_term: Authorship, Publishing.

URL: *https://www.nature.com/articles/d41586-018-06185-8*

Junk, T. R. & Lyons, L. (2020), 'Reproducibility and Replication of Experimental Particle Physics Results', *Harvard Data Science Review* **2**(4). arXiv:2009.06864 [physics].

URL: *http://arxiv.org/abs/2009.06864*

Kakeshita, T., Ishii, K., Ishikawa, Y., Matsubara, H., Matsuo, Y., Murata, T., Nakano, M., Nakatani, T., Okumura, H., Takahashi, N., Takahashi, N., Uchida, G., Uematsu, E., Saeki, S. & Kato, H. (2022), Development of IPSJ Data Science Curriculum Standard, *in* D. Passey, D. Leahy, L. Williams, J. Holvikivi & M. Ruohonen, eds, 'IFIP Advances in Information and Communication Technology', pp. 156–167.

URL: *https://doi.org/10.1007/978-3-030-97986-7_13*

Kross, S. & Guo, P. J. (2019), Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges, *in* 'CHI '19', p. 1–14.

URL: *https://doi.org/10.1145/3290605.3300493*

Li, F., Xiao, Z., Ng, J. T. D. & Hu, X. (2021), Exploring Interdisciplinary Data Science

Education for Undergraduates: Preliminary Results, *in* K. Toeppe, H. Yan & S. K. W. Chu, eds, 'International Conference on Information', Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 551–561.

URL: *https://doi.org/10.1007/978-3-030-71292-1_43*

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J. & Moher, D. (2009), 'The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies that Evaluate Health Care Interventions: Explanation and Elaboration', *Annals of Internal Medicine* **151**(4), W–65.

URL: *https://doi.org/10.1371/journal.pmed.1000100*

Liu, D. (2022), Prepare Data Science Program Student Outcomes and Curricula for ABET Accreditation, *in* '2022 ASEE Annual Conference & Exposition'.

URL: *https://peer.asee.org/prepare-data-science-program-student-outcomes-and-curricula-for-abet-accreditation.pdf*

Liu, Y. & Wei, X. (2020), How to use stock data for data science education: A simulated trading platform in classroom, *in* '2020 IEEE 2nd International Conference on Computer Science and Educational Informatization (CSEI)', pp. 5–8.

URL: *https://doi.org/10.1109/CSEI50228.2020.9142534*

Manzella, G. M. & Emery, W. (2022), How Can Ocean Science Observations Contribute to Humanity?, *in* 'Ocean Science Data', Elsevier, pp. 319–335.

URL: *https://doi.org/10.1016/B978-0-12-823427-3.00005-0*

Miah, S. J., Solomonides, I. & Gammack, J. G. (2020), 'A Design-Based Research Approach for Developing Data-Focussed Business Curricula', *Education and Information*

Technologies **25**(1), 553–581.

URL: *https://doi.org/10.1007/s10639-019-09981-5*

Mike, K., Nemirovsky-Rotman, S. & Hazzan, O. (2020), Interdisciplinary Education - The Case of Biomedical Signal Processing , *in* '2020 IEEE Global Engineering Education Conference (EDUCON)', pp. 339–343.

URL: *https://doi.org/10.1109/EDUCON45650.2020.9125200*

Nair, R., Chugani, M. N. & Thangavel, S. K. (2020), MetaData: A Tool to Supplement Data Science Education for the First Year Undergraduates, *in* 'ICIET 2020', New York, NY, USA, p. 153–160.

URL: *https://doi.org/10.1145/3395245.3396409*

National Academies of Sciences, Engineering and Medicine (2018), *Data Science for Undergraduates: Opportunities and Options*, National Academies Press.

URL: *https://doi.org/10.17226/25104*

National Science Foundation (2017), 'NSF's 10 Big Ideas'.

URL: *https://www.nsf.gov/news/special_reports/big_ideas/*

National Science Foundation (2018), 'Science and Engineering Indicators'.

URL: *https://www.nsf.gov/statistics/2018/nsb20181/assets/nsb20181.pdf*

Oliver, J. C. & McNeil, T. (2021), 'Undergraduate Data Science Degrees Emphasize Computer Science and Statistics but Fall Short in Ethics Training and Domain-Specific Context', *PeerJ Computer Science* **7**, e441.

URL: *https://doi.org/10.7717/peerj-cs.441*

Open Science Collaboration (2015), 'Estimating the Reproducibility of Psychological Science', *Science* **349**(6251), aac4716. Publisher: American Association for the Advance-

ment of Science.

URL: *https://www.science.org/doi/10.1126/science.aac4716*

Pieterman-Bos, A. & van Mil, M. H. W. (2023), 'Integration of Philosophy of Science in Biomedical Data Science Education to Foster Better Scientific Practice', *Science and Education* **32**(6), 1709–1738.

URL: *https://doi.org/10.1007/s11191-022-00363-x*

Punch, K. F. & Oancea, A. E. (2014), 'Introduction to Research Methods in Education'.

Raj, R. K., Parrish, A., Impagliazzo, J., Romanowski, C. J., Aly, S. G., Bennett, C. C., Davis, K. C., McGettrick, A., Pereira, T. S. M. & Sundin, L. (2019), An Empirical Approach to Understanding Data Science and Engineering Education, *in* 'ITiCSE-WGR '19', p. 73–87.

URL: *https://doi.org/10.1145/3344429.3372503*

Rao, A., Bihani, A. & Nair, M. (2018), Milo: A Visual Programming Environment for Data Science Education , *in* '2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)', pp. 211–215.

URL: *https://doi.org/10.1109/VLHCC.2018.8506504*

Rao, A. R., Desai, Y. & Mishra, K. (2019), Data Science Education Through Education Data: an End-to-End Perspective , *in* '2019 IEEE Integrated STEM Education Conference (ISEC)', pp. 300–307.

URL: *https://doi.org/10.1109/ISECon.2019.8881970*

Robeva, R. S., Jungck, J. R. & Gross, L. J. (2020), 'Changing the Nature of Quantitative Biology Education: Data Science as a Driver', *Bulletin of Mathematical Biology*

82(10), 127.

URL: *https://doi.org/10.1007/s11538-020-00785-0*

Ryan, L. (2016), From Self-Service to Self-Sufficiency, *in* L. Ryan, ed., 'The Visual Imperative', Morgan Kaufmann, Boston, pp. 45–60.

Schmitt, K. R. B., Clark, L., Kinnaird, K. M., Wertz, R. E. H. & Sandstede, B. (2023), 'Evaluation of EDISON's Data Science Competency Framework Through a Comparative Literature Analysis', *Foundations of Data Science* **5**(2), 177–198.

URL: *https://doi.org/10.3934/fods.2021031*

Shao, G., Quintana, J. P., Zakharov, W., Purzer, S. & Kim, E. (2021), 'Exploring Potential Roles of Academic Libraries in Undergraduate Data Science Education Curriculum Development', *The Journal of Academic Librarianship* **47**(2), 102320.

URL: *https://doi.org/10.1016/j.acalib.2021.102320*

Shapiro, B. R., Meng, A., O'Donnell, C., Lou, C., Zhao, E., Dankwa, B. & Hostetler, A. (2020), Re-shape: A method to teach data ethics for data science education, *in* 'CHI '20', p. 1–13.

URL: *https://doi.org/10.1145/3313831.3376251*

Song, I.-Y. & Zhu, Y. (2016), 'Big Data and Data Science: What Should We Teach?', *Expert Systems* **33**(4), 364–373.

URL: *https://doi.org/10.1111/exsy.12130*

Song, I.-Y. & Zhu, Y. (2017), 'Big Data and Data Science: Opportunities and Challenges of iSchools', *Journal of Data and Information Science* **2**(3), 1–18.

URL: *https://doi.org/10.1515/jdis-2017-0011*

Stern, H. S., Richardson, D. J. & Papaefthymiou, M. (2021), 'Data science and computing: The view from a sister campus', *Harvard Data Science Review* **3**(2).

Takemura, A. (2018), 'A New Era of Statistics and Data Science Education in Japanese Universities', *Japanese Journal of Statistics and Data Science* **1**(1), 109–116.
URL: *https://doi.org/10.1007/s42081-018-0005-7*

Tucker, M. C., Shaw, S. T., Son, J. Y. & Stigler, J. W. (2023), 'Teaching Statistics and Data Analysis with R', *Journal of Statistics and Data Science Education* **31**(1), 18–32.
URL: *https://doi.org/10.1080/26939169.2022.2089410*

Vance, E. A., Glimp, D. R., Pieplow, N. D., Garrity, J. M. & Melbourne, B. A. (2022), 'Integrating the Humanities into Data Science Education', *Statistics Education Research Journal* **21**(2), 9–9.
URL: *https://doi.org/10.52041/serj.v21i2.42*

Wiktorski, T., Demchenko, Y. & Belloum, A. (2017), Model Curricula for Data Science EDISON Data Science Framework, *in* '2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)', IEEE, pp. 369–374.
URL: *https://doi.org/10.1109/cloudcom.2016.0109*

Wiktorski, T., Demchenko, Y., Belloum, A. & Shirazi, A. (2016), Quantitative and Qualitative Analysis of Current Data Science Programs from Perspective of Data Science Competence Groups and Framework, *in* '2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)', pp. 633–638. ISSN: 2330-2186.

Witmer, J. (2020), 'ASA Journal Gets New Name, Mission', *Amstat News* .
URL: *https://magazine.amstat.org/blog/2020/12/01/journal-gets-new-mission/*

Yamamoto, N., Ishida, A., Ogitsuka, K., Oishi, N. & Murakami, J. (2021), 'Development of

Online Learning Material for Data Science Programming Using 3D Puzzle', *International Journal of Information and Education Technology* **11**(4), 154–163.

URL: *https://doi.org/10.18178/ijiet.2021.11.4.1505*

Zakaria, M. S. (2023), 'Data Science Education Programmes in Middle Eastern Institutions: A Survey Study', *IFLA Journal* **49**(1), 157–179.

URL: *https://doi.org/10.1177/03400352221113362*

Zhang, J. & Wu, B. (2020), Self and Socially Shared Regulation of Learning in Data Science Education: A Case Study of "Quantified Self" Project, *in* 'ICLS 2020 Proceedings'.

URL: *https://repository.isls.org/handle/1/6745*