Exploring Interdisciplinary Data Science Education for Undergraduates: Preliminary Results

Fanjie Li[1], Zhiping Xiao[1], Jeremy Tzi Dong Ng[2], and Xiao Hu[2](✉)

[1] University of Hong Kong, Pokfulam Road, Hong Kong, Hong Kong S.A.R.
{fanjie,jakexiao}@connect.hku.hk
[2] University of Hong Kong Shenzhen Institute of Research and Innovation, Shenzhen, China
jntd@connect.hku.hk, xiaoxhu@hku.hk

Abstract. This paper reports a systematic literature review on undergraduate data science education followed by semi-structured interviews with two frontier data science educators. Through analyzing the hosting departments, design principles, curriculum objectives, and curriculum design of existing programs, our findings reveal that (1) the data science field is inherently interdisciplinary and requires joint collaborations between various departments. Multi-department administration was one of the solutions to offer interdisciplinary training, but some problems have also been identified in its practical implementation; (2) data science education should emphasize hands-on practice and experiential learning opportunities to prepare students for data analysis and problem-solving in real-world contexts; and (3) although the importance of comprehensive coverage of various disciplines in data science curricula is widely acknowledged, how to achieve an effective balance between various disciplines and how to effectively integrate domain knowledge into the curriculum still remain open questions. Findings of this study can provide insights for the design and development of emerging undergraduate data science programs.

Keywords: Data science education · Undergraduate · Interdisciplinarity · Curriculum design

1 Introduction

With the growing demand of data-savvy workforce in various sectors, data science programs designed for a variety of degree levels are on the rise. According to the statistics published by the Data Science Community, as of August 2020, 624 data science-related higher education programs have been implemented and delivered around the world [1]. Currently, data science education has been dominated by master-level programs (69%, 432 out of 624), while the bachelor's and PhD programs only count for 10% (N = 66) and 3.7% (N = 23) respectively [1].

Compared to master's program, the bachelor's program in data science is still in its emerging phase. Data science educators are still exploring effective solutions to designing curricula for undergraduate students, as their limited technical experience and prior

K. Toeppe et al. (Eds.): iConference 2021, LNCS 12645, pp. 551–561, 2021.
https://doi.org/10.1007/978-3-030-71292-1_43

background knowledge both pose more challenges and constraints on the curriculum structure. Hence, this study aims to probe existing literature on undergraduate-level data science programs and summarize their experience. Specifically, we are interested in the following research questions:

RQ1: How may the existing literature inform the curriculum design of emerging undergraduate data science programs, especially their practical insights on developing students' interdisciplinary learning experience?
RQ2: What are the challenges and opportunities in designing and implementing an interdisciplinary data science program for undergraduate students?

2 Methods

To answer these research questions, we conducted a systematic literature review and semi-structured interviews with some authors of the reviewed articles so as to further elicit their insights. This section outlines the search strategy, study selection procedure, and coding schema employed, as well as the subsequent interview with invited authors.

2.1 Search Strategy

To depict the current situation of undergraduate data science education, we searched and reviewed the literature indexed by the following databases: (1) Web of Science (WoS), (2) Scopus, and (3) ACM Digital Library (ACM DL), as these databases are representative of the fields closely related to the topic under discussion. Specifically, we chose WoS and Scopus because they broadly cover the computing, engineering, education, and social science fields [2], while another subject specific database (i.e., ACM DL) was selected given its focus on the scientific and educational computing literature.

To form the search query, we identified a set of keywords related to three facets: (1) data science, (2) education, and (3) undergraduate. Particularly, the third facet (i.e., undergraduate) was added because the curriculum design for undergraduate- and graduate-level program can differ significantly given their differences in program duration and student background. After including the alternative terms relevant to each facet, the final query was: ("data science" OR "big data") AND (education OR teaching OR learning OR curriculum OR curricula) AND (undergraduate OR bachelor). The query was applied to the title, abstract, and keyword fields. We did not restrict the publication timeframe but limited our search to English articles. Our literature search was performed in May 2020.

2.2 Study Selection and Citation Search

169 articles were obtained via this initial database searching. These articles then went through an iterative screening process where the inclusion and exclusion criteria were continuously refined. Table 1 summarized the final inclusion and exclusion criteria for study selection.

Table 1. Inclusion and exclusion criteria

Category	Inclusion and exclusion criteria
Study purpose	1. Included: Studies that focus on education, including program/course design, curriculum guideline, or pedagogical suggestions, etc. 2. Excluded: Studies that discuss programs in fields other than data science 3. Excluded: Studies that use data science methods for analyzing educational data (e.g., [3])
Target students	1. Included: The program mentioned should be designed for undergraduate students or be applicable to both undergraduate and graduate students 2. Excluded: Studies discussing programs designed for graduate students
Article type	1. Included: Full articles 2. Excluded: Articles with only an abstract or a brief introduction (e.g., posters)

During the study selection phase, we first assessed the relevance of the article based on its title and abstract. In case the information presented in the title or abstract is insufficient for reaching an inclusion decision, we read the full article to check whether the paper is aligned with our inclusion criteria. Out of the 169 articles obtained via database searching, 19 papers passed the study screening, while the other 150 papers failed to meet the inclusion criteria.

To complement the database searching results, we further performed citation searching for those 19 relevant papers, obtaining papers citing them which then went through the same study selection procedure. Only one new paper was identified as relevant, which resulted in 20 papers for our subsequent analysis.

2.3 Coding Schema

For the selected articles, a coding schema was designed based on iteratively reviewing the 20 studies. The coding schema covers (1) the basic information of the publication and the data science program being described, (2) program details, and (3) challenges and opportunities in program design and program development (see Table 2).

Table 2. Coding schema

Category	Codes	Description
Basic information	Year of publication	Year in which the article was published
	Country (program)	Country where the program was developed
Program details	Involved department(s)	Department(s) involved in the program

(*continued*)

Table 2. (*continued*)

Category	Codes	Description
	Design principles	Principles that guided the program design
	Learning outcomes	Expected learning outcomes of this program
	Courses	Courses included in the curriculum
	Course prerequisites	Prerequisites of courses in the program
Challenges & Opportunities	Challenges	Challenges in program design/development
	Opportunities	Opportunities in program design/development

2.4 Semi-structured Interviews

In addition to the document analysis of the literature, we further conducted semi-structured interviews with authors of these studies to elicit practical insights and suggestions from those frontier educators. An invitation email together with an informed consent form were sent to the corresponding author of each paper. Two of them have accepted our invitation and signed the consent form. A Zoom-based remote interview was conducted with each of them. Both of them have participated in the design of their corresponding programs.

The interview protocol contains questions on (1) interviewee's roles in the program (e.g., designer, director, regular teacher), (2) the courses they teach (e.g., a particular course, internship, capstone), (3) their opinions on designing an interdisciplinary data science program, (4) how their programs integrate domain knowledge and address the needs of interdisciplinary training in the curriculum design, (5) challenges and difficulties in curriculum design and program implementation, and (6) suggestions for other data science programs. The interviews lasted 20 min each and were audio-recorded and transcribed for further analysis.

3 Results and Discussions

3.1 Overview of Identified Publications and Programs

The papers obtained from this study were published between 2012 and 2020, and half of them (N = 10) were published in last three years (2018–2020). This suggests that undergraduate data science education is still in its emerging phase. A majority of these data science programs (85%, N = 17) were designed and delivered in the United States. 10% of programs (N = 2) were in China and the remaining one was in Australia.

3.2 Program and Curriculum Design

With our focus on how existing literature informs the curriculum design of emerging undergraduate data science programs, we further analyzed the following features of identified programs.

Hosting Department. The department in which the data science program was hosted varied from program to program. Particularly, 4 out of 20 (20%) programs were jointly offered by multiple departments [4–7], while the rest of them either did not provide relevant information in the publication (N = 5) or were hosted by departments from a single discipline (N = 11). For the programs jointly hosted by multi-departments, 2 out of 4 were resulted from collaborations between mathematics and computer science departments [5, 7]. Besides, one program [4] was offered by the departments of business and science. Another program [6] was brought about by a more interdisciplinary endeavor, which involved a committee of faculty representing ten disciplines. For the programs hosted in a single department, 3 out of 11 (27%) were offered by the computer science department [8–10], 3 (27%) by the mathematics or statistics department [11–13], 2 (18%) by departments in the field of information management [14, 15], while the rest of them were hosted by departments from other disciplines, including business (1) [16], journalism and communication (1) [17], and liberal arts (1) [18]. Such diversity in the hosting departments indicates the interdisciplinary nature of the data science field, which calls for collaborations between experts in different fields to offer interdisciplinary training to future data scientists.

Design Principles. Statements describing the design principles of the program (e.g., "the principles of this program are …", "this program is based on the rules of …") were identified in 80% of the articles (N = 16). Based on these statements, we plotted a word cloud diagram to extract and visualize the key principles that guided the program design (Fig. 1). As presented in Fig. 1, the frequently appearing principles mainly center around three aspects: (1) data at the core (or "center around data"), (2) opportunities for hands-on practice, and (3) disciplinary knowledge. Specifically, three articles explicitly highlighted the significance of data [11, 12, 19], with another two programs underscored that students should be trained with large and real-world datasets [13, 20]. Closely related to this principle, three programs further propounded that the curriculum should offer rich opportunities for hands-on practices with big data [9, 13, 14], which is in line with the pedagogical principle of experiential learning. Finally, we identified several design principles in relation to the coverage of disciplinary knowledge. For the breadth of disciplinary coverage, one article acknowledged that the curriculum design should reflect the interdisciplinary nature of the data science field [19]. As for the depth of disciplinary coverage, this article suggested that students should receive sufficient training in mathematical foundations as well as statistical and computational thinking, while other programs (e.g., [6, 13]) suggested that the curriculum should assume little prior background knowledge and avoid a high level of computer science and mathematics requirements. In line with this concern, [21] encouraged teaching with GUI-based analytics tools (e.g., RapidMiner) to reduce the programming requirement. Nonetheless, despite the ever-growing discussions on the breadth and depth of disciplinary coverage, how to achieve an effective balance of breadth (exposure to multiple disciplines) and depth (knowledge of pertinent disciplines) in the curriculum design still remain open questions. Last but not least, the importance of reaching an effective balance between disciplines has also been taken into consideration. For instance, [5] stressed that the program should provide balanced training in statistics and computer science.

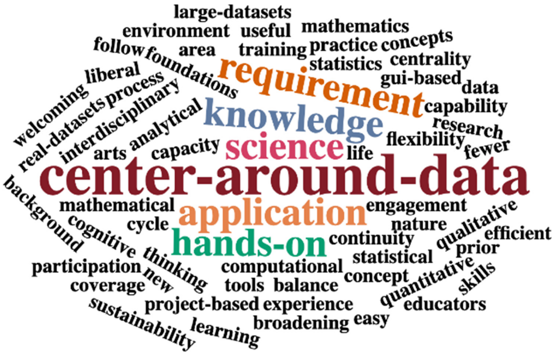

Fig. 1. Word cloud generated from the program design principles

Learning Outcomes. The expected program learning outcomes often play an important role in shaping the program curriculum design. 16 articles (80%) explicitly described their curriculum objectives and expected learning outcomes. We plotted the frequent words using another word cloud diagram (Fig. 2). Being consistent with the visualization shown in Fig. 2, our coding also revealed that the expected learning outcomes of these data science programs were mainly fourfold: (1) Students are expected to acquire comprehensive knowledge about data science concepts, methods, and tools, especially developing familiarity with machine learning (e.g., predictive analytics) and essential statistical concepts and methods (e.g., probability, statistical inference); (2) Students should be able to flexibly apply and transfer their knowledge and skills for data analysis and problem solving in real-world contexts; (3) Students should be able to design and implement a standard data processing pipeline in a data-intensive application; (4) Students should be able to effectively communicate and present the data analysis outcomes using text, table, or other visualization techniques. These curriculum objectives generally aligned with the essential knowledge and skills throughout the data science life cycle [19]. Moreover, apart from the mathematical and computing knowledge, several programs (e.g., [6, 10, 20, 22, 23]) further included the domain expertise (e.g., business, political science) as part of their curriculum objectives.

Courses. Aligned with the aforementioned curriculum objectives, courses on statistics, machine learning, data analytics, programming, and data visualization were covered by most of the program curricula. Table 3 summarizes courses included in the reviewed programs. It is noteworthy that all of the 20 programs include courses on data mining, data analytics, or big data, confirming that it is the core knowledge and skillset of data science.

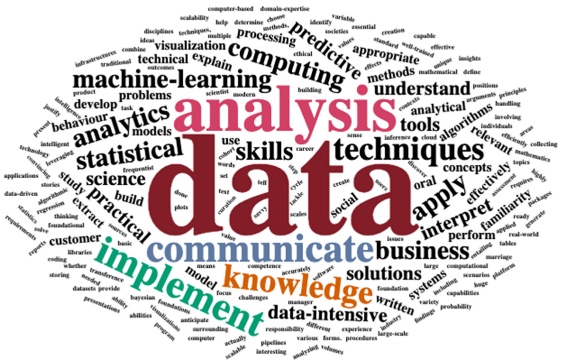

Fig. 2. Word cloud generated from the expected program learning outcomes

Besides mathematics and computing, more than half of the reviewed programs have also acknowledged the importance of communication and presentation skills, which hence offered training targeted at this competence (data visualization: N = 9, communication: N = 2). This echoes the findings of a recent study [24] that emphasized the importance of training data science students in communicating reproducible data analysis.

Courses on data curation and management are included in four programs, so are those on ethics and privacy. This indicates a misalignment with recent prevalent research on data governance, particularly on fairness, accountability, and transparency in data science [25]. Although these issues have been discussed in research, and in the context of data science education most recently [25], the actual inclusion of these into the curricula, especially on the undergraduate level has yet to be reflected in the literature. One program explicitly includes courses in application areas of data science such as sociology, economics, political science, and psychology [6]. Although it is well acknowledged that data science is an application-oriented field, most programs do not include courses in application areas.

Course Prerequisites. Though a few programs presented in the articles (20%, N = 4) contain courses with computer science or mathematics prerequisites, the majority of them (80%, N = 16) do not assume any prior experience or background knowledge. Among those programs with course prerequisites, several require programming experience in Python, Java, C++, or Linux environment (e.g., [9, 15]), while the others assume basic knowledge of mathematics and statistics (e.g., [11]). While this observation further reflects the significant role of mathematics and computer science in the data science curricula, it is also noteworthy that the majority of reviewed programs do not have course prerequisites. This might reflect the novice-friendly requirement for entering the data science field and its application-driven nature.

3.3 Challenges and Opportunities

Our document analysis of the literature and interviews with invited authors further revealed several challenges and opportunities in designing and implementing an interdisciplinary data science program.

Table 3. Overview of program curricula

Category	Course	Covered by # of programs
Mathematics	Statistics	9
	Calculus	5
	Linear algebra	2
	Probability theory	2
	Discrete structures	1
Computer science	Programming/computing	11
	Data structures & algorithms	6
	Machine learning/artificial intelligence	8
	Database management system/DB design	4
	Information system	3
	Introduction to software design	1
	Introduction to semantic technology	1
	Internet of Things	1
Data science	Introduction to data science	5
	Data analytics/big data/data mining	20
	Regression and forecasting models	1
	Business intelligence	1
	Data visualization	9
	Data curation	3
	Data manipulation	1
	Data organization & management	1
Others	Ethics and privacy	4
	Communication	2
	Asking interesting questions	1
	Quantitative decision making	1
	Management & organizational behaviour	1
	Project management	3
	Anthropology and sociology, biology, economics, philosophy, physics, political science, psychology	1

Challenges. 8 out of 20 articles discussed the challenges in program design and development, such as (1) students' difficulty in fulfilling mathematics and computer science course requirements, especially for those with no programming experience and those who studied liberal arts [4, 21], (2) limited faculty for course delivery and for maintaining active engagement with students considering the increasing class size [4], (3) challenges in designing experiential learning activities for international students due to the work visa problem [5], and (4) difficulties in covering relevant knowledge components within limited credit hours [17]. Besides, our interviews also revealed some problems encountered in multi-department administration: (1) It is hard to control how courses were set up in another department such as prerequisites (Interviewee #2); (2) Without a departmental home, students may not have an identity or community as a data science student. (Interviewee #1).

Opportunities. Despite the challenges mentioned above, there are also opportunities for the emerging undergraduate data science programs. Given the interdisciplinary nature of data science, the programs can be integrated with various social science (e.g., business) and liberal arts (e.g., journalism) majors. As pointed out by interviewee #1, it is flexible for each data science program to focus on strengths of their own institutions.

4 Conclusion and Future Work

To extract practical insights for the design and implementation of undergraduate data science programs and identify the challenges and opportunities in program design and development, we conducted a systematic literature review and performed semi-structured interviews with two frontier data science educators. Through analyzing the hosting departments, design principles, curriculum objectives, and curriculum design of the existing undergraduate data science programs, our findings reveal that (1) the data science field is inherently interdisciplinary and requires joint collaborations between various departments. Multi-department administration was one of the solutions to offer interdisciplinary training, but some problems has also been identified in its practical implementation (c.f. Sect. 3.3); (2) data science education should emphasize hands-on practices and experiential learning opportunities to prepare students for data analysis and problem-solving in real-world contexts; and (3) although the importance of comprehensive coverage of various disciplines in data science curricula is widely acknowledged, how to achieve an effective balance of breadth (exposure to multiple disciplines) and depth (knowledge of pertinent disciplines), especially the effective integration of domain knowledge, still remain open questions.

As a preliminary review of the status quo of undergraduate data science education, this study discussed the practical experience from existing literature and frontier educators, which may shed light on the design and development of emerging undergraduate data science programs. In future work we will continue interviewing more data science educators for an in-depth analysis of the curriculum, syllabus, and regulations.

Acknowledgments. This study is supported by a Teaching Development Grant sponsored by the University of Hong Kong and a grant (No. 61703357) by National Natural Science Foundation National Natural Science Foundation of China.

References

1. DataScienceCommunity. https://datascience.community/colleges. Accessed 15 Aug 2020
2. Mongeon, P., Paul-Hus, A.: The journal coverage of web of science and scopus: a comparative analysis. Scientometrics **106**(1), 213–228 (2015). https://doi.org/10.1007/s11192-015-1765-5
3. Belyakova, E.G., Zakharova, I.G.: Interaction of university students with educational content in the conditions of information educational environment. Educ. Sci. J. **21**(3), 77–105 (2019)
4. Rosenthal, S., Chung, T.: A data science major: building skills and confidence. In: Proceedings of the 51st ACM Technical Symposium on Computer Science Education, Portland, OR, USA, pp. 178–184. ACM (2020)
5. Adams, J.C.: Creating a balanced data science program. In: Proceedings of the 51st ACM Technical Symposium on Computer Science Education, Portland, OR, USA, pp. 185–191. ACM (2020)
6. Havill, J.: Embracing the liberal arts in an interdisciplinary data analytics program. In: Proceedings of the 50th ACM Technical Symposium on Computer Science Education, Minneapolis, MN, USA, pp. 9–14. ACM (2019)
7. Anderson, P., Bowring, J., McCauley, R., Pothering, G., Starr, C.: An undergraduate degree in data science: curriculum and a decade of implementation experience. In: Proceedings of the 45th ACM Technical Symposium on Computer Science Education, Atlanta, Georgia, USA, pp. 145–150. ACM (2014)
8. Carter, T., Hauselt, P., Martin, M., Thomas, M.: Building a big data research program at a small university. J. Comput. Sci. Coll. **28**(2), 95–102 (2012)
9. Eckroth, J.: A course on big data analytics. J. Parallel Distrib. Comput. **118**, 166–176 (2018)
10. Ramamurthy, B.: A practical and sustainable model for learning and teaching data science. In: Proceedings of the 47th ACM Technical Symposium on Computing Science Education, Memphis, TN, USA, pp. 169–174. ACM (2016)
11. Baumer, B.: A data science course for undergraduates: thinking with data. Am. Stat. **69**(4), 334–342 (2015)
12. Yan, D., Davis, G.E.: A first course in data science. J. Stat. Educ. **27**(2), 99–109 (2019)
13. Yavuz, F.G., Ward, M.D.: Fostering undergraduate data science. Am. Stat. **74**(1), 8–16 (2020)
14. Li, X., et al.: Curriculum reform in big data education at applied technical colleges and universities in China. IEEE Access **7**, 125511–125521 (2019)
15. Asamoah, D.A., Sharda, R., Hassan Zadeh, A., Kalgotra, P.: Preparing a data scientist: a pedagogic experience in designing a big data analytics course. Decis. Sci. J. Innov. Educ. **15**(2), 161–190 (2017)
16. Wymbs, C.: Managing the innovation process: infusing data analytics into the undergraduate business curriculum (lessons learned and next steps). J. Inf. Syst. Educ. **27**(1), 61 (2016)
17. Liao, H.T., Wang, Z., Wu, X.: Developing a minimum viable product for big data and AI education: action research based on a two-year reform of an undergraduate program of internet and new media. In: Proceedings of the 2019 4th International Conference on Big Data and Computing, Guangzhou, China, pp. 42–47. ACM (2019)
18. Mandel, T., Mache, J.: Developing a short undergraduate introduction to online machine learning. J. Comput. Sci. Coll. **32**(1), 144–150 (2016)
19. De Veaux, R.D., et al.: Curriculum guidelines for undergraduate programs in data science. Ann. Rev. Stat. Appl. **4**, 15–30 (2017)
20. Leman, S., House, L., Hoegh, A.: Developing a new interdisciplinary computational analytics undergraduate program: a qualitative-quantitative-qualitative approach. Am. Stat. **69**(4), 397–408 (2015)

21. Haynes, M., Groen, J., Sturzinger, E., Zhu, D., Shafer, J., McGee, T.: Integrating data science into a general education information technology course: an approach to developing data savvy undergraduates. In: Proceedings of the 20th Annual SIG Conference on Information Technology Education, Tacoma, WA, USA, pp. 183–188. ACM (2019)
22. Gupta, B., Goul, M., Dinter, B.: Business intelligence and big data in higher education: status of a multi-year model curriculum development effort for business school undergraduates, MS graduates, and MBAs. Commun. Assoc. Inf. Syst. **36**(1), 23 (2015)
23. Miah, S.J., Solomonides, I., Gammack, J.G.: A design-based research approach for developing data-focused business curricula. Educ. Inf. Technol. **25**(1), 553–581 (2020)
24. Yu, B., Hu, X.: Toward training and assessing reproducible data analysis in data science education. Data Intell. **1**(4), 381–392 (2019)
25. Bates, J., et al.: Integrating FATE/critical data studies into data science curricula: where are we going and how do we get there? In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, pp. 425–435. ACM (2020)