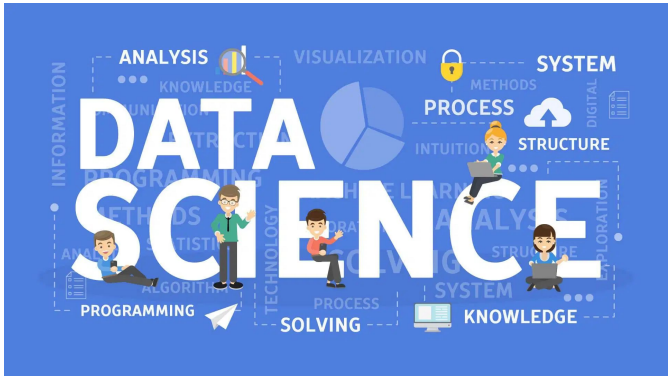
Exploring Interdisciplinary Data Science Education for Undergraduates: Preliminary Results

Fanjie Li[1], Zhiping Xiao[1], Jeremy Tzi Dong Ng[2], Xiao Hu[2]

[1]The University of Hong Kong, Hong Kong S.A.R.; Two authors had equal contributions
[2]The University of Hong Kong Shenzhen Institute of Research and Innovation, China

Introduction

- Data Science Community:
  - 624 data science related Higher Education programs as of Aug 2020
  - 69% master-level
  - 10% Bachelor’s
  - 3.7% PhD
- Bachelor’s programs in data science
  - **initial phase**
- This study:
  - probes existing literature and summarize experience
  - **systematic literature review**
  - **semi-structured interviews**



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?

Search Strategy

Scopus

ACM DL DIGITAL LIBRARY

- Literature indexed by three databases:
  - **Web of Science (Wos)**, **Scopus**: computing, engineering, education, and social science
  - **ACM Digital Library (ACM DL)**: scientific and educational computing
- Final query on Title, Abstract, Keywords:

  ("data science" OR "big data") AND (education OR teaching OR learning OR curriculum OR curricula) AND (undergraduate OR bachelor)
- No restriction on timeframe; English articles

Study Selection and Citation Search

- Initial database searching: 169 articles
- 19 passed our iterative screening process
- 1 additional paper identified after citation searching

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

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.
	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.

Semi-structured Interviews

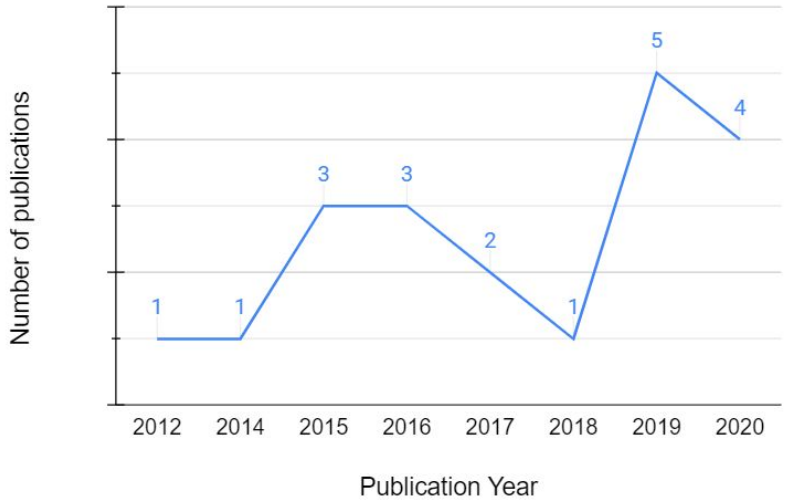
- To elicit **practical insights** and **suggestions** from these authors as **frontier educators** of data science
- Interviews with two authors on Zoom; 20 minutes each

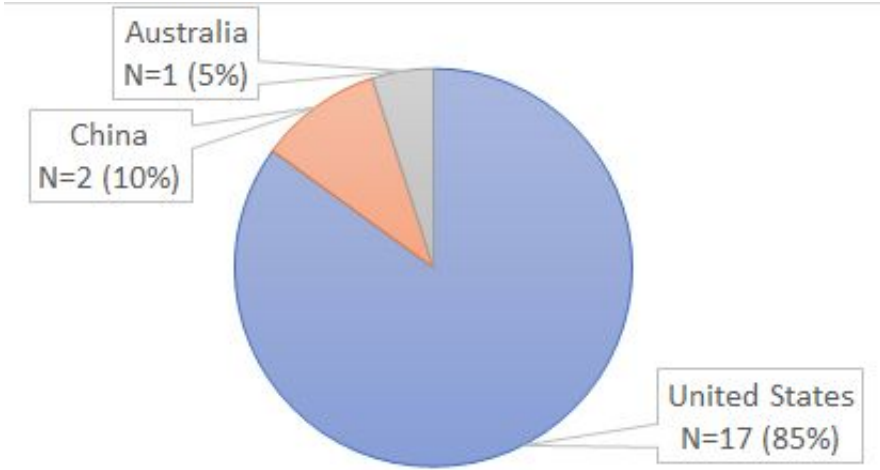
Highlights of interview questions

1. Roles (e.g., director) and courses they teach
2. Opinions on designing an interdisciplinary data science program
3. Challenges and difficulties in curriculum design and program implementation
4. Suggestions for other data science programs

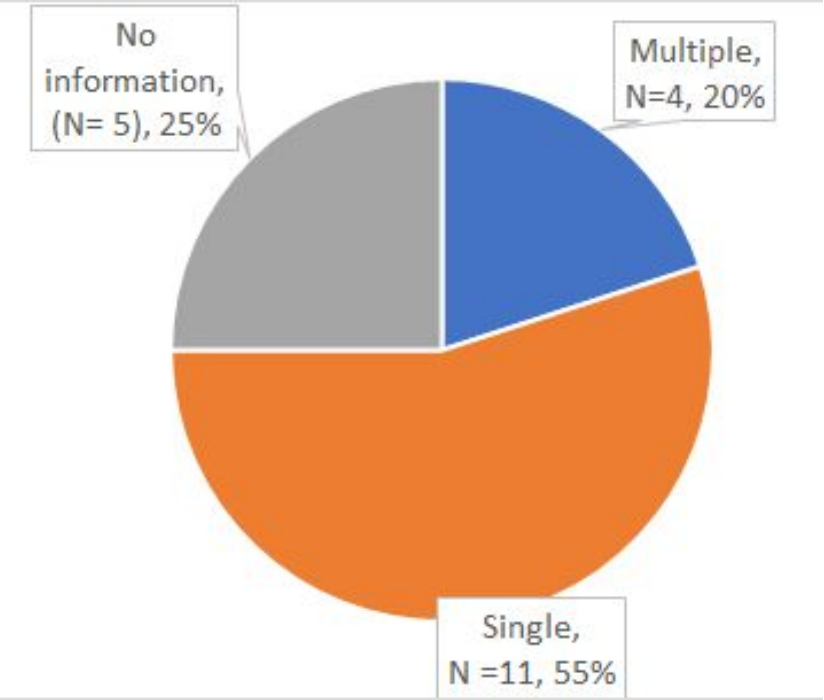
Identified Publications and Programs

- 20 papers, published in 2012 - 2020
- Designed and delivered in United States, China and Australia





Hosting Department: Diverse, Interdisciplinary



Jointly offered by multiple departments:

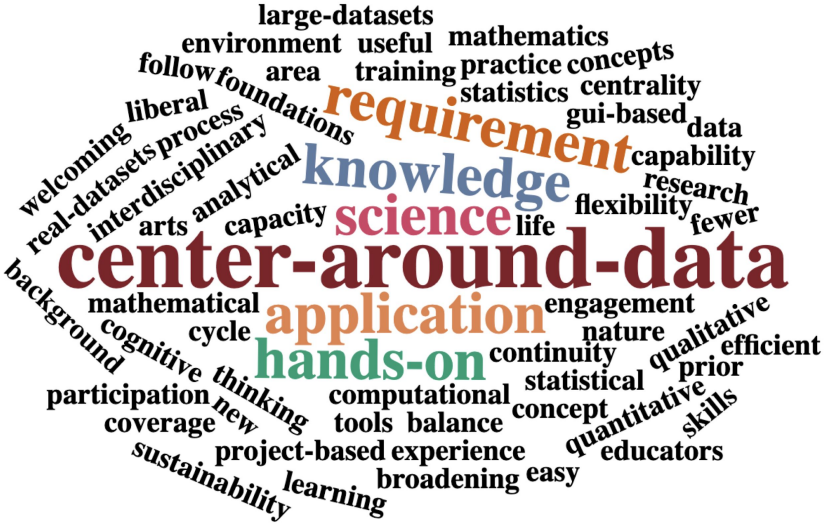
- Mathematics + Computer Science (N=2)
- Business + Science (N=1)
- Committee involving ten disciplines (N=1)

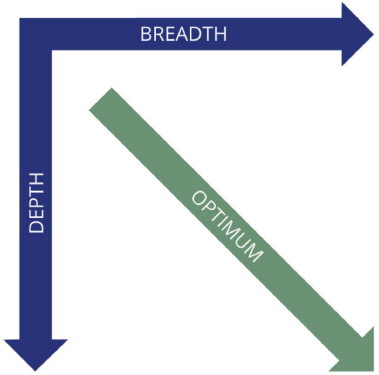
Single departments:

- Computer Science (N=3)
- Mathematics / Statistics (N=3)
- Information Management (N=2)
- Business (N=1)
- Journalism and communication (N=1)
- Liberal Arts (N=1)

Design Principles (N = 16)

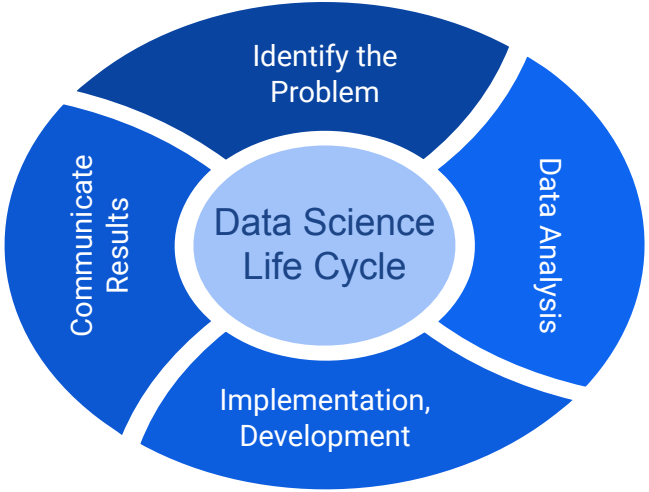
1. **Data at the core** (or “center around data”)
   a. Significance of data (N = 3)
   b. Large, real-world datasets (N = 2)
2. Opportunities for **hands-on practices** with big data (N = 3)
3. **Disciplinary knowledge**;
   a. Mathematical foundation; Statistical, computational thinking
   b. Avoid high-level requirements on mathematics and computer science
   c. Teaching with GUI-based analytic tools
4. **Balanced breadth and depth**





Curriculum Objectives and Learning Outcomes (N = 16)

1. to **acquire** comprehensive knowledge about data science concepts, methods, and tools (e.g., Machine Learning) essential statistical concepts and methods
2. to flexibly **apply** and **transfer** their knowledge and skills for data analysis and problem solving in real-world contexts
3. to **design** and **implement** a standard data processing pipeline in a data-intensive application
4. to effectively **communicate** and **present** the data analysis outcomes using text, table, or other visualization techniques



Courses

- All include courses on **data mining**, **data analytics**, or **big data**

Mathematics Courses	**# Programs**
Statistics	**9**
Calculus	5
Linear algebra	2
Probability theory	2
Discrete structures	1

Computer Science Courses	**# Programs**
Programming / computing	**11**
Machine learning / artificial intelligence	8
Data structures & algorithms	6
Database management system / DB design	4
Information system	3
Introduction to software design	1
Introduction to semantic technology	1
Internet of Things	1

Courses (cont'd)

Data Science Courses	# Prog
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 and management	1

Other Courses	# Prog
Ethics and privacy	**4**
Communication	**2**
Asking interesting questions	1
Quantitative decision making	1
Management & organization behaviour	1
Project management	3
Anthropology and sociology, biology, economics, philosophy, physics, political science, psychology	**1**

- Importance of **communication** and **presentation** skills for communicating **reproducible** data analysis
- Misalignment with recent research on **data governance & fair, accountable, and transparent DS**
- Lack of courses in **application areas** (e.g., political science)

Course Prerequisites

- Not assume any prior experience / background knowledge (N = 16)
  - **Novice-friendly** requirement for entering the field
  - **Application-driven** nature
- With courses having computer science / mathematics prerequisites (N = 4):
  - **Programming experience**: Python, Java, C++, or Linux environment
  - Basic knowledge of **mathematics** and **statistics**

Challenges and Opportunities

- Challenges in program design and development (N = 8)
  - Students' difficulty in **fulfilling course prerequisites**
  - **Limited faculty** for course delivery and maintaining engagement
  - Designing **experiential learning activities for intl' students**
  - Covering relevant knowledge within **limited credit hours**
- Problems in **multi-department administration**
  - Hard to control **course set-up** (Interviewee #2)
  - No student **identity** or **community** (Interviewee #1)
- Opportunities (Interviewee #1)
  - Possible **integration** with various other majors (e.g., liberal arts)
  - Focus on **institutional strengths**

Conclusion and Future Work

- Summary
  - Inherently **interdisciplinary** field
  - Joint **collaborations** (e.g., multi-departmental administration)
  - **Hands-on** practices and **experiential learning** opportunities
- Open questions:
  - How to achieve an effective balance between disciplinary **breadth** and **depth**
  - How to **integrate domain knowledge** effectively
- Future Works
  - In-depth analysis of **curriculum**, **syllabus** and **regulations**

References

1. DataScienceCommunity, http://datascience.community/colleges, last accessed 2020/08/15.
2. Mongeon, P., Paul-Hus, A.: The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics 106(1), 213-228 (2016).
3. Belyakova, E. G., Zakharova, I. G.: Interaction of university students with educational content in the conditions of information educational environment. The education and science journal 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, pp. 178-184. ACM, Portland, OR, USA (2020).
5. Adams, J. C.: Creating a Balanced Data Science Program. In: Proceedings of the 51st ACM Technical Symposium on Computer Science Education, pp. 185-191. ACM, Port-land, OR, USA (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, pp. 9-14. ACM, Minneapolis, MN, USA (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, pp. 145-150. ACM, Atlanta, Georgia, USA (2014).
8. Carter, T., Hauselt, P., Martin, M., Thomas, M.: Building a big data research program at a small university. Journal of Computing Sciences in Colleges 28(2), 95-102 (2012).
9. Eckroth, J.: A course on big data analytics. Journal of parallel and distributed computing, 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, pp. 169-174. ACM, Memphis, TN, USA (2016).
11. Baumer, B.: A data science course for undergraduates: Thinking with data. The American Statistician 69(4), 334-342 (2015).
12. Yan, D., Davis, G. E.: A First Course in Data Science. Journal of Statistics Education 27(2), 99-109 (2019).
13. Yavuz, F. G., Ward, M. D.: Fostering Undergraduate Data Science. The American Statistician 74(1), 8-16 (2020).

References (cont'd)

14. Li, X., Fan, X., Qu, X., Sun, G., Yang, C., Zuo, B., Liao, Z.: 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. Decision Sciences Journal of Innovative Education 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). Journal of Information Systems Education 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, pp. 42-47. ACM, Guangzhou, China (2019).
18. Mandel, T., & Mache, J.: Developing a short undergraduate introduction to online machine learning. Journal of Computing Sciences in Colleges 32(1), 144-150 (2016).
19. De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., ..., Kim, A. Y.: Curriculum guidelines for undergraduate programs in data science. Annual Review of Statistics and Its Application, 4, 15-30 (2017).
20. Leman, S., House, L., & Hoegh, A.: Developing a new interdisciplinary computational analytics undergraduate program: A qualitative-quantitative-qualitative approach. The American Statistician 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, pp. 183-188. ACM, Tacoma, WA, USA (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. Communications of the Association for Information Systems 36(1), 23 (2015).
23. Miah, S. J., Solomonides, I., Gammack, J. G.: A design-based research approach for developing data-focused business curricula. Education and Information Technologies 25(1), 553-581 (2020).
24. Yu, B., Hu, X.: Toward training and assessing reproducible data analysis in data science education. Data Intelligence 1(4), 381-392 (2019).
25. Bates, J., Cameron, D., Checco, A., Clough, P., Hopfgartner, F., Mazumdar, S., ... & de la Vega de León, A.: 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, pp. 425-435. ACM, Barcelona, Spain (2020).

Thank you!

TDG
TEACHING DEVELOPMENT GRANT

国家自然科学基金委员会
National Natural Science Foundation of China

Ms. Fanjie Li, Research Assistant - fanjie@connect.hku.hk
Mr. Zhiping Xiao, Master Student - jakexiao@connect.hku.hk
Mr. Jeremy Ng, PhD candidate – jntd@connect.hku.hk
Dr. Xiao Hu, Associate Professor, HKU-SIRI – xiaoxhu@hku.hk

http://ccmir.cite.hku.hk/