The Learning Difficulties Experienced by Introductory Data Science Students

Sinem Demirci, Ph.D.

July 11th, 2023

Hello!

Sinem Demirci, PhD

Postdoctoral Visiting Researcher/Lecturer - UCL

sinemdemirci.github.io

sinemdemirci

sinemmdemirci

drsinemdemirci

Today's Outline

In this talk, I will be talking about

- Data Science Education in Higher Education Context
- The role of Introductory Data Science (IDS) Courses in Data Science Education
- Aim of the Study
- Methodology
- Findings
- Conclusions and Discussions

What is Data Science?

- Data science is a field that blends multiple areas and demands expertise in a range of skills and concepts spanning statistics, computer science, mathematics, and other domains(Mike and Hazzan, 2023).
- An agreement for a single definition for data science is a difficult task because of its multifaceted nature.
- A Venn diagram (Figure 1) that integrates Application Domain, Mathematics & Statistics, and Computer Science is typically used to help illustrate the interdisciplinary nature of data science as a discipline.

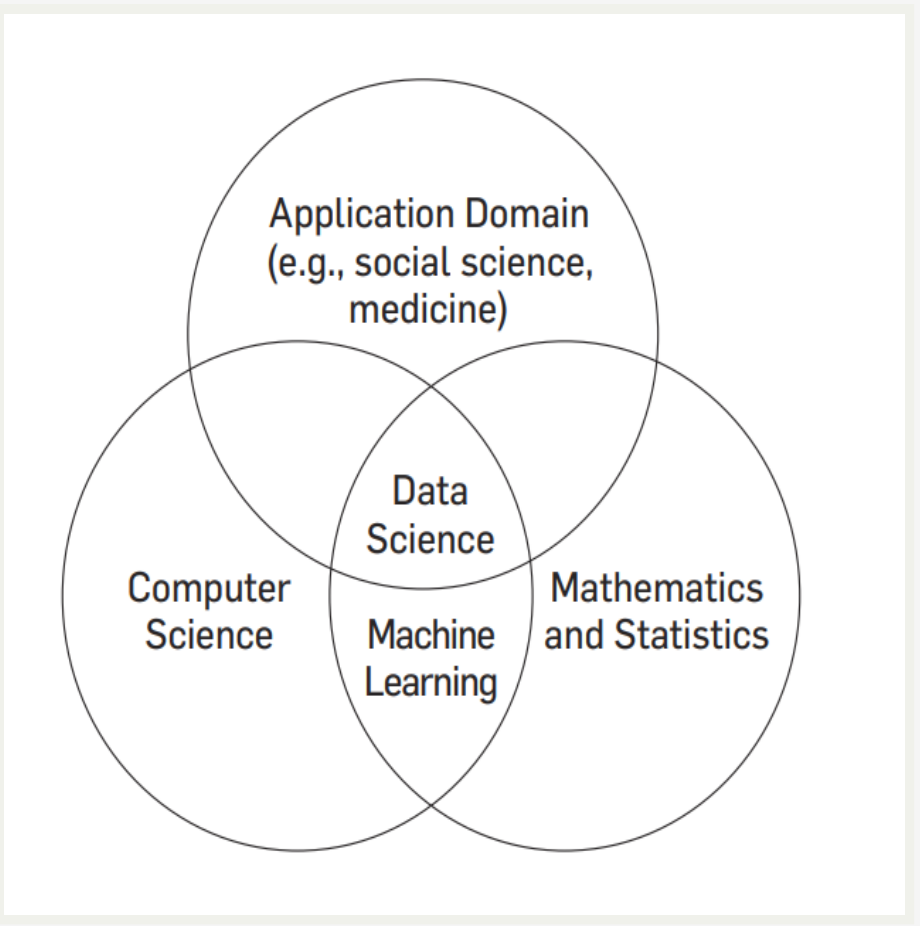

Figure 1. Venn diagram of data science (Mike and Hazzan, 2023)

Data Science in Higher Education Context

Data Science in Higher Education Context-II

- The growing demand for competent data scientists is being met by a significant number of students who have an interest in acquiring and applying data science skills. (Donoghue et al., 2021).
  - Students are interested in enrolling data science courses and/or pursuing a data science career.
- Introductory data science (IDS) course experiences have a potential to attract students to pursue majors, minors, tracks, and certificates offered by institutions (National Academies of Sciences, Engineering and Medicine Consensus Report, 2018)
  - Thus, IDS courses have an important role in students' decisions to become data scientists.

Introduction to Data Science Courses

- IDS courses are introductory courses offered by different departments such as mathematics, statistics, data science, or computer science aiming to enable individuals to grasp the basics of data science.
  - Even though majority of the enrolled students are from these departments, students are coming from almost every major/department to these courses.

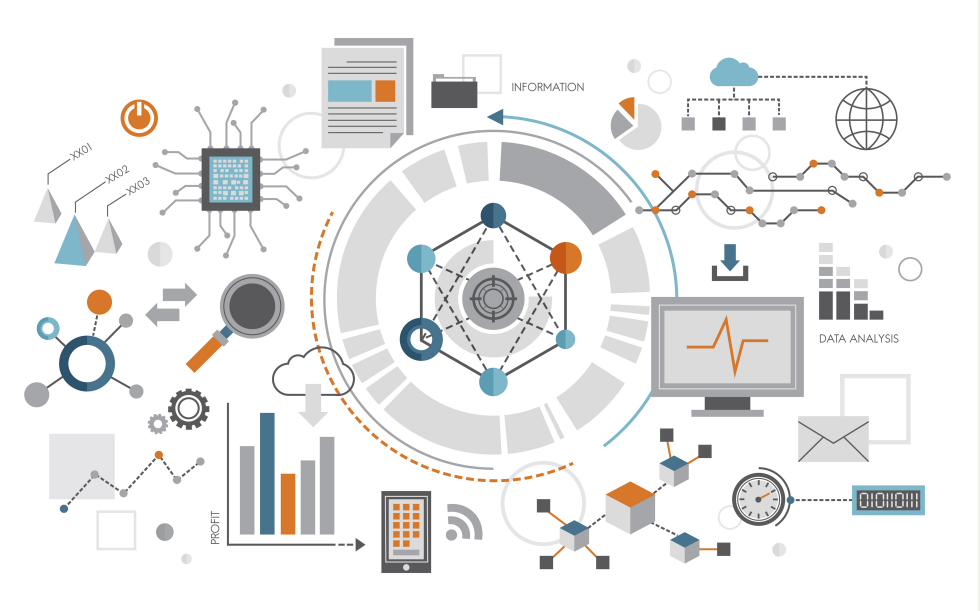

Image by rawpixel.com on Freepik

Aim of the study

- While data science education community strives to create curriculum guidelines and determine competencies for data scientists and data science students, we would like to direct attention to the importance of the learning difficulties of data science education students
  - Learning difficulties constrain them developing a sound understanding of data science and making progress (Qian & Lehman, 2017).
- This study aims to determine students' difficulties in IDS courses reported by their instructors.
  - It is evident that teachers' awareness of students' difficulties is an essential aspect of effective teaching (Sadler et al. 2013).
  - Thus, determining students' difficulties might provide an insight into how to enhance IDS teaching and revise the scope of IDS courses as well as data science majors/programs.

Theoretical Framework of the Study

- We adapted a framework that has been used in early computer science education (Bayman & Mayer, 1988; McGill & Volet, 1997; Qian & Lehman, 2017) to classify knowledge of programming into three categories as:
  - **Syntactic Knowledge:** "…knowledge of specific facts about a programming language and rules for its use" (McGill & Volet, 1997, p. 277)
  - **Conceptual Knowledge:** "…understanding of computer programming constructs and principles" (p. 277).
  - **Strategic Knowledge:** Integrating syntactic and conceptual knowledge of programming to solve novel problems (Qian & Lehman, 2017).
- We expanded its scope to make it compatible with *data science*.
  - We decided to retain the exact definitions of syntactic and strategic knowledge.
  - We extended the definition of *conceptual knowledge* to encompass understanding the constructs and principles of mathematics & statistics, computer science, subject-specific knowledge, and interdisciplinary knowledge which are included in defining data science (Mike & Hazzan, 2023) as a discipline.

Methodology

- In this study, we chose *qualitative research* design (Merriam & Tisdell, 2016)
  - Our aim was to understand how IDS instructors classify and interpret students' difficulties in IDS courses and
  - "what meaning they attribute to their experiences" (Merriam, 2009, p. 23).
- This study is a subset of a larger study.
- We completed the qualitative part of this study in the Term 1 and Term 2 of the 2022-2023 academic year.
- Quantitative part which is for testing the scale developed by the researchers will be administered in the next term.

Sampling Procedure

- We defined the target population to consist of instructors who **taught an introductory data science course at least twice** at the undergraduate level.
- Our rationale for recruiting participants was as following:
  - We tried to **standardize the name of the IDS courses**.
    - We selected participants who taught a course whose title include ‘Data Science’ and one of the following keywords:
      - Introduction, Principles, Elements or Fundamentals
  - When an instructor teaches a course for the first time, they focus on multiple aspects of the course as a novice.
    - Thus, instructors who have gone through the **second iteration** of the course would be able to reflect deeper about the course and the students.
- We recruited **16 participants** (2 pilot, 14 main study) via mailing lists and online forums with large teacher-scholar communities.

Sample Profile - IDS Instructors

- All participants were from Northern America
- Only 4 instructors were the sole instructor in their IDS course. The other participants have either co-instructors or PGTAs/graders.
- The instructors had terminal degrees in varying subjects including statistics, mathematics, computer science, genetics, and economics.
- They all had been teaching an introductory data science course for a varying number of years, with a range from 1 to 10 years of experience.

Formal training in Data Science

Self-taught – 4 participants

Workshops – 4 participants

Industry experience – 2 participants

Others – 5 (enrolling some DS courses in graduate years, graduated from closely related areas such as Stat and CS.

Formal Training in Teaching

Workshops – 3 participants

TA trainings – 3 participants

Course/internship – 3 participants

Degree – 1 participant

None – 5 participants

Sample Profile - IDS Classrooms

- Students are coming from almost every major/department.
  - Majority – Math, Stat, CS, Data Science
  - Others – Engineering, business school, social science, economics, humanities, life sciences, environmental science, political science, health science, undecided
- Prerequisite **Yes** – 6; **No** – 8
- Prerequisite to any other course **Yes** – 11; **No** – 2; **Not sure** – 1

Table 1: Class sizes reported by IDS instructors

Class Size	**n**
300+	2
200-299	1
100-199	2
…	
30-39	2
20-29	3
10-19	3
1-9	1

Data Collection and Data Analysis

- We collected data through online semi-structured interviews from 14 participants.
- We used *qualitative content analysis* (Merriam & Tisdell, 2016) for generating a comprehensive codebook.
- We analysed responses of questions and follow-up questions centered around students' difficulties on

1. Concepts
2. Conceptual Knowledge
3. Performing Data Science Tasks in IDS courses.

- We adapted the framework of Qian and Lehman (2017) which covers introductory programming students' difficulties and extended this framework to introductory data science courses.

Validity and Reliability Evidences of the Study

To enhance the trustworthiness of the study, we collected indicators for transferability, dependability, and credibility (Merriam & Tisdell, 2016).

- Particularly, we provided a detailed description for our participants' profile, data collection and data analysis procedures.
- We also had different participants (e.g., differed in terms of year of experience, terminal degree etc.) based on our selection criteria which enabled maximum variation in our sample.
- Our research team continuously compared and discussed to determine the extent of codebook based on the theoretical framework and data of the study.

Findings on Students' Difficulties

In this part, we present the findings of study which were categorized into three themes (1) Syntactic Knowledge Difficulties; (2) Conceptual Knowledge Difficulties; and (3) Strategic Knowledge Difficulties.

Syntactic Knowledge Difficulties I

- Within the theme of Syntactic Knowledge Difficulties, we identified two categories based on the reports from IDS instructors.
  - The first category related to students' difficulties with markup languages and reproducibility tools,
  - The second category related to difficulties with programming languages. The codebook for these difficulties is provided in Table 2.

Table 2. Knowledge of Students' Syntactic Difficulties

Categories	**Codes**
Markup Languages and Reproducibility Tools	HTML, R Markdown, Quarto Markdown, Jupyter Notebook,Linux, Git/GitHub
Programming Languages	Packages, Libraries, Misspelling, Adapting the Code, How to Read Data

Syntactic Knowledge Difficulties II

- As IDS courses utilized various markup languages, reproducibility tools, and programming languages, IDS instructors reported distinct syntactic difficulties that were specific to their course.
  - 11 out of 14 IDS instructors observed that students without prior coding experience encountered more syntactic difficulties.
  - To support these students, some IDS instructors offer additional sessions and/or office hours.

Conceptual Knowledge Difficulties-I

We categorized conceptual knowledge difficulties into five categories:

- mathematics
- statistics
- computer science
- domain-specific knowledge and
- interdisciplinary knowledge.

The codes that emerged from data are given in Table 3.

Table 3. Knowledge of Conceptual Difficulties

Category	**Concepts and Topics**
Mathematics	Algorithms, Permutation Testing
Statistics	Types of Variables, Confidence Interval, Principles of Data Visualization, Hypothesis Testing, Correlation vs. Causality, Bootstrapping, Inductive Inference, Statistical Modelling, p-value, Sampling Distribution
Computer Science	I/O File Management, Working Mechanisms of Markup Languages, Basics of Coding, Filter Function, Basics of Web Scraping, Select Function, Joining Data Sets, Mapping Functions, Loops, Creating Functions
Domain-Specific Knowledge	Understanding Technical Writing, Understanding the Nature of Data
Interdisciplinary Knowledge	Ethics, Machine Learning

Conceptual Knowledge Difficulties II

- Among the IDS instructors
  - Nine reported that students experienced difficulties in understanding statistical concepts,
  - Six reported difficulties in understanding computer science concepts.
  - Five IDS instructors mentioned difficulties in understanding either the nature of data or technical writing in a specific domain.
- The principles of data visualization were the most frequently mentioned among the statistical concepts.
- Understanding the basics of coding and joining data sets were two commonly reported difficulties among the computer science concepts.

Students' Strategic Knowledge Difficulties

- Except for 3 IDS instructors, 11 reported observing strategic knowledge difficulties in their IDS courses.
- The most frequently mentioned difficulties were debugging and data wrangling.
  - Additionally, some of them denoted that students tend to oversimplify data science tasks given in IDS course and try to run a statistical analysis without thinking about the content and examining data set accordingly.
- A sample excerpt was as following:

"…So certainly, so this so kind of so statistical analysis in so kind of correct statistical analysis in general is a problem. So, everyone is very tempted to just kind of throw any tool they can, they can at the problem and just like, look at the outputs to see if the if the p-value is significant. So, this so I try to instill this kind of skeptical mindset of like, you know, does that, does the model fit? Does the question make sense? … [conversation continues] So that, I would say, is kind of one of the more challenging things to teach."

Table 4. Knowledge of Students' Strategic Difficulties

Strategic Knowledge Difficulties
Debugging
Communication
Data Wrangling
Appreciating the complexity of Interdisciplinary Research
Making Appropriate Data Visualization Decisions
Creative Thinking
Proper Use of Descriptive Statistics
Conducting a Good Research
Deciding Statistical Analysis Methods-Modelling
Working with Real and Messy Data
Handling Missing Data
Asking Good Questions
Web Scraping
Setting up Data Science Pipeline

Discussion and Conclusion

- In summary of our key findings, we examined responses of IDS instructors on students' understanding specific to IDS courses.
- Most of them highlighted that students without prior programming knowledge tended to experience more syntactic difficulties and require additional support.
- Apart from students' difficulties, some IDS instructors in this study articulated that students have a tendency to oversimplify data science assignments in the IDS course, by attempting to run statistical analyses without adequately considering the content and carefully examining the dataset.
  - While some students may oversimplify IDS tasks, we suggest that this oversimplification may also be partially attributed to the difficulties that students face in these courses, which are not yet fully understood.
  - Therefore, further studies are needed to measure students' difficulties and identify the specific areas in which they struggle, to better understand the reasons for this "oversimplification".

Implications and Limitations

- It is noteworthy that while there were some commonalities among the IDS courses examined in this study, each course may have presented its own unique set of dynamics.
  - Therefore, our findings may serve as informative rather than generalizable constructs that can inform IDS instructors and the wider data science education community.
- This study was presented from the perspective of the instructors, not the IDS students.
  - Thus, it is essential to conduct more systematic research to assess students' perspectives in IDS courses as well as observe classroom environments to be able to inform policymakers and educators on how to improve learning environments in IDS courses.
- The sample of this study consisted of North American IDS instructors, even though we did not have such a specific aim within the context of the study.
  - The possible bias for sample selection might be related to the selection criteria (e.g., selecting participants based on similar course names).
  - In other country settings, there might be similar courses with different names. Thus, further studies in other country settings might also provide an insight into other students' difficulties that we were not able to capture in this study.
- We incorporated a framework utilized from early computer science education (e.g., Bayman & Mayer, 1988; McGill & Volet, 1997; Qian & Lehman, 2017) to analyze the data and adapted to IDS context.
  - While we found the framework useful in providing an initial understanding, we acknowledge the potential requirement for further modifications to actively classify knowledge in IDS classes.

Acknowledgements

- This study is funded by The Scientific and Technological Research Council of Turkey, TÜBİTAK and University College London.
- Collaborators of this project are Dr Mine Dogucu, Assist. Prof. Dr Joshua M. Rosenberg and Teaching Assoc. Prof. Dr Andrew Zieffler

References

Asamoah, D. A., Doran, D., & Schiller, S. (2020). Interdisciplinarity in data science pedagogy: a foundational design. Journal of Computer Information Systems, 60(4), 370-377, https://doi.org/10.1080/08874417.2018.1496803

Bayman, P., & Mayer, R. E. (1988). Using conceptual models to teach BASIC computer programming. Journal of Educational Psychology, 80(3), 291, https://psycnet.apa.org/doi/10.1037/0022-0663.80.3.291

Danyluk, A., & Leidig, P. (2021). Computing competencies for undergraduate data science curricula: ACM data science task force. Peer-Reviewed Publications, 8, https://scholarworks.gvsu.edu/cispeerpubs/8

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., … & Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. Annual Review of Statistics and Its Application, 4, 15-30.

Donoghue, T., Voytek, B., & Ellis, S. E. (2021). Teaching creative and practical data science at scale. Journal of Statistics and Data Science Education, 29(sup1), 27-39, https://doi.org/10.1080/10691898.2020.1860725

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). How to design and evaluate research in education (Vol. 7, p. 429). New York: McGraw-hill.

Merriam, S. B. (2009). Qualitative Research: A Guide to Design and Implementation. San Francisco: CA: Jossey-Bass.

Merriam, S. B., & Tisdell, E. J. (2016). Qualitative Research: A Guide to Design and Implementation (Fourth Edition). San Francisco.

Mike K. & Hazzan, O. (February 2023). What is data science? Communications of the ACM, 66(2), 12–13, https://doi.org/10.1145/3575663

National Academies of Sciences, Engineering and Medicine Consensus Report (2018). Data Science for Undergraduates: Opportunities and Options. Washington, https://nas.edu/envisioningds.

Qian, Y., & Lehman, J. (2017). Students' misconceptions and other difficulties in introductory programming: A literature review. ACM Transactions on Computing Education (TOCE), 18(1), 1-24, https://doi.org/10.3102/0002831213477680

Yan, D., & Davis, G. E. (2019). A first course in data science. Journal of Statistics Education, 27(2), 99-109, https://doi.org/10.1080/10691898.2019.1623136