International Society of the Learning Sciences

A New Approach to High School Data Science: Set Theory and Logic

Yuanlin Zhang, Texas Tech University, y.zhang@ttu.edu
Hanxiang Du, Western Washington University, duh2@wwu.edu
Wanli Xing, University of Florida, wanli.xing@coe.ufl.edu

Abstract: There is a consensus to introduce data science to secondary schools. However, data science is interdisciplinary in nature and not easy to teach in K−12 settings. We proposed a new approach to integrate mathematics, statistics, and programming—the foundations of data science—for high school students based on set theory and logic. We developed an 8-week data science foundation course and implemented it in a public high school. We conducted semi-structured group interviews to collect students' feedback on the course and the new approach. Students thought the approach could well connect the topics and helped them learn the interdisciplinary content.

Introduction

K−12 data science research remains an emerging field that has only started to receive attention within the last decade (Du et al., 2022; Mobasher et al., 2019). The interdisciplinary nature of data science raised challenges for high school data science course development. Most of the existing work provides summer camp, after-school programs or a more generalized framework/program which includes some data perspectives (e.g., Grover et al., 2015; Mobasher et al., 2019; Weintrop et al., 2016). A handful of studies introduced discrete math and logic programming to high school classrooms, covering topics such as classical mathematic theorems, properties, and proofs (e.g., Bouhnik & Giat, 2009). We aimed to prompt high school data science education by proposing a new approach to integrate mathematics, statistics, and programming—the foundations of data science—to help students learn data science and develop problem-solving skills. The current study reported a high school data science foundation course developed based on the proposed approach and the preliminary results of the course implementation, addressing the following research question: *How do students reflect on their learning experiences with our proposed course?*

Study design

We proposed a new approach to teach data science in high school by integrating mathematics, statistics, and programming based on set theory and logic. We first introduced the basics of set theory and logic, including set builder notation, logic concepts such as negation, and function mappings. Mathematics knowledge serves as the tool to represent statistical concepts, such as mean, frequency, relative frequency, and joint frequency. When introducing statistical concepts, we introduced the set theory representation of the concepts to help students learn both. Meanwhile, set theory provides a bridge to connect mathematics and programming practices: tuples, which are very much like the data structure vector of R programming language in terms of definition and behaviors. Using tuples and vectors is expected to lower the challenge of transferring written-on-paper solutions to programming programs. Statistics provides the method for students to solve real-world problems: they represent a problem using set theory and logic, then transfer the representation into a programming project, and eventually, use statistical methods to solve the problem in programming environments.

Based on the approach, we developed an 8-week long face-to-face data science foundation course for high school students, covering topics on computing knowledge, statistical concepts, and programming practice. Students first learned how to develop a mathematical expression (e.g., set theory) to abstract a problem, then to implement the mathematical expression using R programming language to solve the problem. A total number of 53 students voluntarily registered for the course in a public high school in the United States. The study has gained IRB approval.

We developed a pre- and post-test to measure students' understanding of covered topics. To ensure the validity of our assessment, we designed our tests based on the questions designed for Advanced Placement test from the textbook (Starnes & Tabor, 2018). At the end of the course, we conducted four semi-structured group interviews with 14 students to understand students' thoughts on the new approach and the course. The interview questions cover two categories: (1) the course structure (e.g., In general, how do you feel about this approach helping you understand data science? How do you feel about learning concepts from set theory, computing, and statistics together? Any suggestions on how we teach this in the future?), (2) the impacts of this approach on

learning set theory, statistics, and computing (e.g., Do you think this course help you learn set theory/statistics/computing concepts? How well do you think set theory, statistics or computing concepts help you learn one another?). All interviews were conducted via Zoom and recorded upon approval.

Result and conclusion

Although we conducted four group interviews, one recording was lost due to an equipment breakdown. Only three group interview recordings (11 students) were transcribed and analyzed. Based on interview protocols, one researcher first reviewed the transcripts and labeled students' responses individually, then merged similar labels to several meaningful ones: (1) the complexity of the course, (2) the impact of the approach on learning data science, and (3) other interesting comments.

Regarding complexity of the course, four students thought the content is easy to understand, and explicitly suggested that *"they (the learning materials) were actually very well structured"* and *"looks nice and connected and like eventually to all the different things together.*" Meanwhile, two students felt they *"don't have like a complete grasp on both of those subjects"* or *"couldn't really find a good ... you know ... connection between that (the tuples, set theory and standard deviation stuff like that) and coding."* Another student who also felt they *"kind of learned something and it would not really come up until significantly later if at all."*

As for the impact of the approach, at least six students identified the benefits of learning statistics, set theory, and programming at the same time, as they thought one subject helps their understanding of another one. A student who had not learnt statistics before thought *"this course and the computing is a good introduction to statistics."* Another student stated that *"coding helped me more than anything to understand statistics."* Even students who thought the computing is hard found the statistics content help them better understand computing: *"understanding of statistics helps me understand the computing more."* It is very interesting that while some students thought their *"understanding of statistics kind of helped me understand the computing more,"* some other students suggested that *"usually for me it was the opposite."*

The semi-structured interviews also led to unexpected discoveries. One student pointed out that the course *"helps you to better understand how, like other companies use data to, like find out what the customer is like and so,"* bridging the gap between school education and industries. Several students also discussed their learning difficulties. Two students explicitly suggested that they have *"really bad memory"* and need *"more reinforcement."* One student thought *"when assigning variables, it would be harder,"* while another one sometimes had *"a little bit of trouble kind of understanding what the question is asking."*

The semi-structured interviews showed that in general, students thought the approach could well structure and connect the course content, and more than half of interviewed participants identified the benefits of this approach: understanding any one subject of set theory, statistics, and programming may help students better understand the other two subjects.

References

Du, H., Xing, W., Pei, B., Zeng, Y., Lu, J., & Zhang, Y. (2022, April). A Descriptive and Historical Review of STEM+ C Research: A Bibliometric Study. In *International Conference on Computer Supported Education* (pp. 1-25). Cham: Springer Nature Switzerland.

Bouhnik, D., & Giat, Y. (2009). Teaching high school students applied logical reasoning. *Journal of Information Technology Education. Innovations in Practice*, 8, 1.

Grover, S., Pea, R., & Cooper, S. (2015). Designing for deeper learning in a blended computer science course for middle school students. *Computer science education*, 25(2), 199-237.

Mobasher, B., Dettori, L., Raicu, D., Settimi, R., Sonboli, N., & Stettler, M. (2019). Data science summer academy for chicago public school students. *ACM SIGKDD Explorations Newsletter*, 21(1), 49-52.

Starnes, D. S., & Tabor, J. (2018). *The practice of statistics.* New York: WH Freeman.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of science education and technology*, 25, 127-147.

Acknowledgments

This work is supported by the National Science Foundation (NSF) of the United States under grant number DRL-1901704 and DRL-2201393. We thank C. Birchall-Roman, J. Daniel, J. Ketring, S. Kumar, S. B. Nooka, R. Rodriguez, P. B. Sivaraju, R. Varan, R. Yalavarthi and J. Zhao for their help.