An Integrated Approach to Data Science Foundations in Computing, Mathematics and Statistics

Yuanlin Zhang
Department of Computer Science,
Texas Tech University
Lubbock, TX, USA
y.zhang@ttu.edu

Hanxiang Du
School of Teaching and Learning,
University of Florida
Gainesville, FL, USA
h.du@ufl.edu

Wendy Staffen
San Felipe Del Rio Consolidated
Independent School District
Del Rio, TX, USA
wendv.staffen@sfdr-cisd.org

Wanli Xing
School of Teaching and Learning,
University of Florida
Gainesville, FL, USA
wanli.xing@coe.ufl.edu

Joshua Archer
Department of Computer Science,
Texas Tech University
Lubbock, TX, USA
josh.archer@ttu.edu

ABSTRACT

To address the challenge of teaching the interdisciplinary foundations of data science in computing, mathematics and statistics, we propose a mathematical logic based framework to seamlessly and coherently integrate these foundations. A 8-week module based on the framework is implemented in a high school. The results show an overall feasibility of the integrated approach.

Introduction. A National Academies of Sciences, Engineering, and Medicine report [2] highlights the foundations of data science in *computing*, *mathematics* and *statistics* among others. As a result, data science provides excellent opportunities for high school students to develop college-ready STEM competencies and prepare their future careers. However, these foundations are rarely addressed explicitly in the existing work. We propose a *mathematical logic* based framework naturally unifying those foundations and study its impact on students' learning of statistics and computing.

Theoretical Framework. The *data* in data science in general are *sets* of interesting objects in human activities, and the *relations* among them. *Set theory* and *logic*, core components of *mathematical logic*, provide the language to represent and reason with sets and relations. Statistics has developed important concepts about data which can be precisely defined using set theory and logic. Computing complements statistics by automating the entire reasoning process with data using computers. *Abstraction* and *programming* are at the core of computing [1]. Set theory and logic offer a precise language for abstraction (of relevant information for solving intended problems) and for programming too (e.g., *tuples* in set theory provides a direct and explicit support of data types such as *vectors* and *data frames* in programming language R). Finally, all concepts are introduced in the context of solving real life problems.

Study Design. We develop an 8-week module (15 100-minute lessons) aligned with the Texas standards on computer science and statistics. It is implemented inside an elective course *Computer Science II* of a pubic Texas high school in Spring 2021. The module covers abstraction (sets, relations and logic), programming, basics of statistics, central tendency and two categorical/quantitative variables. The course has 53 10-12th grade students. We conduct pre- and post-tests (before and after the module) to assess students' learning outcomes on abstraction (of computing) and statistics. The statistics questions are based on AP practice questions from a popular AP statistics textbook. The development of abstraction questions is guided by the *argument based validation* with mainly theoretical evidences. Each test has 13 short answer questions. At the end of the course, four semi-structured group interviews, with a total of 11 volunteered students, were conducted via zoom and recorded. The interview questions are to help reveal their thoughts of the course and cover the course structure and the impacts of this approach on learning set theory, statistics and computing.

Results. 44 valid responses were obtained for both pre- and post-test. The difference between students' pre- and post-test scores is statistically significant ($t(43) = -5.82, p < .05; d = 2.08$). As for interviews, we transcribed the recordings, reviewed the transcripts, labeled students' responses, and merged similar labels to several meaningful ones. Regarding complexity of the course, four students thought the content is easy to understand, while some suggested that they felt lost. As for the impact of learning statistics and computing on each other, at least six interviewed students identified the benefits of learning statistics and computing at the same time, as they thought one subject helps the understandings of another one.

Acknowledgement. This work is partially supported by NSF (USA) grant 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.

REFERENCES

[1] K-12 Computer Science Framework. 2017. http://www.k12cs.org. Retrieved on September 4 2018.
[2] National Academies of Sciences, Engineering, and Medicine. 2018. *Data science for undergraduates: Opportunities and options.* Washington, DC. National Academies Press.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGCSE '23, March 15–18, 2023, Toronto, ON, Canada.
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9433-8/23/03.
https://doi.org/10.1145/3545947.3576240