Using Data to Inform Computing Education Research and Practice

Thomas Price (Moderator)
North Carolina State University
Raleigh, NC, USA
twprice@ncsu.edu

Baker Franke
Code.org
Chicago, Illinois, USA
baker@code.org

Shuchi Grover
Looking Glass Ventures &
Stanford University
Palo Alto, CA, USA
shuchig@cs.stanford.edu

Monica M. McGill
Knox College, csedresearch.org
Galesburg, Illinois, USA
monica@csedresearch.org

ABSTRACT

The analysis of data plays an increasingly critical role in computing education research, enabled by more and larger datasets, more powerful analysis techniques and better infrastructure for sharing. This panel brings together four panellists at various stages of work involving the collection and analysis of large datasets in different fields of computing education. The panellists will each discuss the current state of their work, the unique aspects of their data, and how that data fits into the larger landscape of computing education and research. Panellists will be asked to explain how they are employing AI and data mining techniques to learn about learners, the research methods they have used to make this happen, and any significant key findings they have discovered through this processes. The panel will discuss emerging topics, including: going beyond log data, handling global-scale datasets, efficiently collaborating with cross-dataset analysis, and ethical and privacy considerations. After the panelists present (5 minutes each), the moderator will pose follow-up questions and invite the audience to pose additional questions or provide other feedback. Key takeaways will include how data mining and artificial intelligence can contribute to improved insight and learning gains and how the larger computer education community can participate in data collection or analysis.

KEYWORDS

datasets, educational data mining, learning analytics

ACM Reference Format:
Thomas Price (Moderator), Baker Franke, Shuchi Grover, and Monica M. McGill. 2020. Using Data to Inform Computing Education Research and Practice. In *The 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20), March 11–14, 2020, Portland, OR, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3328778.3366967

SIGCSE '20, March 11–14, 2020, Portland, OR, USA

ACM ISBN 978-1-4503-6793-6/20/03.
https://doi.org/10.1145/3328778.3366967

1 SUMMARY

Programming environments increasing collect data from students as they work, and the largest of these datasets contain work from hundreds of thousands of learners [2, 12]. This has enabled researchers to apply techniques from Educational Data Mining (EDM) and Learning Analytics (LA) to derive new insights [5] and to build powerful AI- and data-driven systems, including adaptive models that predict student knowledge [13] and automated hints to support struggling students [7]. Complementing this are collections of survey data, evaluation instruments and literature, which enable researchers to more easily build on each others work (e.g. [6]). This increasing role of data in computing education research raises important questions around its collection and use, including ethics and privacy.

This panel will bring together four panelists at various stages of research who have been using computing data to enable research, gain insight and develop new tools to improve computing education. Panellists will discuss how they have used data mining and AI techniques to learn about learners, and improve their outcomes by developing new data-driven systems. They will outline their infrastructure for collecting and analyzing data and highlight significant key findings from their research. In addition to speaking on their own work, panellists will discuss broader questions on role that data should play in computing education, ethical and privacy concerns, and the future of the field. Key takeaways will include how data mining and analytics can lead to improved teaching and learning in computing, and best practices for applying these techniques in research. Attendees will come away with clear next steps for how they can begin or improve the way they collect, analyze and share computing datasets as well as how they can participate in the data collection or analysis of these datasets.

2 PANEL STRUCTURE

Each panelist will have five minutes to introduce their work, their datasets, and how these are being used to inform computing education research and practice. We will then pose prepared questions to the panelists who will each have an chance to answer (approximately 25 minutes). Questions may include:

- What innovations and insights has data enabled for the field of computing education?

- For researchers new to analytics and data mining, where can one learn best practices for collecting and analyzing data?
- What programming data sources and data-driven tools are already available to researchers?
- How can we ground data analysis in educational theories?
- What are ethical and privacy considerations for the collection and use of data?
- How will data shape the future of computing education?

We will take questions from the audience for the last 30 minutes.

3 THOMAS PRICE

As a researcher, I develop tools to support students as they learn to program, such as hints [7] and examples. I use AI- and data-driven techniques to automate this support (e.g. [9]), to make it easier to scale to new classrooms and contexts. In addition to bringing my own perspective as a researcher, I have co-organized two workshops on Educational Data Mining in Computer Science Education [1]. I also work closely with the Standards, Protocols, and Learning Infrastructure for Computing Education group (SPLICE; cssplice.org), where we developed a standard format for programming log data, ProgSnap2 [8].

4 BAKER FRANKE

I am the research and evaluation manager for Code.org. Participating in CS Education research efforts is a priority area for Code.org, since a research base is critical for broad adoption of CS at a national scale. My role is to form research collaborations with the academic community to (1) share Code.org's large datasets collected from students and teachers engaged in CS education activities; (2) implement educational interventions in our platform and programs at scale for the purposes of research; and (3) turn those research findings into evidence-based learning tools that we can build into our platform, and disseminate more broadly. Code.org has roughly 1M students engaged in CS courses and lessons across K-12 and roughly 10K teachers active in our professional development programs with a truly national spread. Examples of successful collaborations are [10–12], with several others in the works .

5 SHUCHI GROVER

I am a CS education researcher and learning scientist working to understand how K-12 students learn programming and computational thinking in the context of block-based environments. The motivation for using EDM/LA to analyze log data stems from a desire to design better curricula and environments to support such learning. I have advocated for hybrid approaches to examining student processes that bring learning theory to bear on EDM/LA efforts and which combine hypothesis-driven (top-down) approaches with data driven (bottom-up) ones to reach a better understanding of learner behaviors and the learning process [3]. I will share our research on examining existing datasets of middle school students' programs in Alice, creating a shareable dataset captured in Exploring Computer Science high school classrooms using Alice (available on the DataShop - https://pslcdatashop.web.cmu.edu/), and analyzing high school students' computational modeling processes of scientific phenomena in a Snap! based modeling environment [4].

6 MONICA MCGILL

Over the course of computing education research, there have been various efforts to improve the practice. Having been involved in some of these efforts, my latest (along with Adrienne Decker) have been focused on developing and maintaining csedresearch.org [6]. The dataset contains data curated from over 500 K-12 computing ed articles and a collection of instruments for evaluating factors (both cognitive and noncognitive) related to academic achievement. During this panel, I will discuss datasets currently in place and our efforts to move the research community to open data so that we can make decisions about best practices based on data that is easier to compare across various studies and has more integrity and meaning.

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. National Science Foundation under Grant Nos. 1625005, 1625335, 1757402, and 1745199.

REFERENCES

[1] David Azcona, Yancy Vance Paredes, Sharon I. Han Hsiao, and Thomas W. Price. 2019. *Proceedings of the 2nd Workshop on Educational Data Mining in Computer Science Education.* Tempe, Arizona.

[2] Neil C. C. Brown, Michael Kölling, Davin McCall, and Ian Utting. 2014. Blackbox: A Large Scale Repository of Novice Programmers' Activity. In *Proceedings of the ACM Technical Symposium on Computer Science Education*. 223–228. https://doi.org/10.1145/2538862.2538924

[3] Shuchi Grover, Satabdi Basu, Marie Bienkowski, Michael Eagle, Nicholas Diana, and John Stamper. 2017. A framework for using hypothesis-driven approaches to support data-driven learning analytics in measuring computational thinking in block-based programming environments. *ACM Transactions on Computing Education (TOCE)* 17, 3 (2017), 14.

[4] Nicole Hutchins, Gautam Biswas, Shuchi Grover, Satabdi Basu, and Caitlin Snyder. 2019. A Systematic Approach for Analyzing StudentsâĂŹ Computational Modeling Processes in C2STEM. In *International Conference on Artificial Intelligence in Education*. Springer, Cham, 116–121.

[5] Petri Ihantola, Matthew Butler, Stephen H Edwards, Virginia Tech, Ari Korhonen, Andrew Petersen, Kelly Rivers, Miguel Ángel Rubio, Judy Sheard, Jaime Spacco, Claudia Szabo, and Daniel Toll. 2015. Educational Data Mining and Learning Analytics in Programming: Literature Review and Case Studies. In *Proceedings of the ACM Conference on Innovation and Technology in Computer Science Education.*

[6] Monica M. McGill and Adrienne Decker. 2017. Computer Science Education Resource Center. https://csedresearch.org

[7] Thomas W. Price, Yihuan Dong, and Dragan Lipovac. 2017. iSnap: Towards Intelligent Tutoring in Novice Programming Environments. In *Proceedings of the ACM Technical Symposium on Computer Science Education.*

[8] Thomas W Price, David Hovemeyer, Kelly Rivers, Austin Cory Bart, Andrew Petersen, Brett A Becker, and Jason Lefever. 2019. ProgSnap2: A Flexible Format for Programming Process Data. In *Proceedings of the Educational Data Mining in Computer Science Workshop in the Companion Proceedings of the International Conference on Learning Analytics and Knowledge*. 1–7.

[9] Thomas W. Price, Rui Zhi, and Tiffany Barnes. 2017. Evaluation of a Data-driven Feedback Algorithm for Open-ended Programming. In *Proceedings of the International Conference on Educational Data Mining.*

[10] Kevin Robinson, Keyarash Jahanian, and Justin Reich. 2018. Using online practice spaces to investigate challenges in enacting principles of equitable computer science teaching. *SIGCSE 2018 - Proceedings of the 49th ACM Technical Symposium on Computer Science Education* 2018-January (2018), 882–887. https://doi.org/10.1145/3159450.3159503

[11] David Weintrop, Heather Killen, and Baker E Franke. 2018. Blocks or Text? How programming language modality makes a difference in assessing underrepresented populations. International Society of the Learning Sciences, Inc.[ISLS].

[12] Mike Wu, Milan Mosse, Noah Goodman, and Chris Piech. 2018. Zero Shot Learning for Code Education: Rubric Sampling with Deep Learning Inference. (2018). arXiv:1809.01357 http://arxiv.org/abs/1809.01357

[13] Michael Yudelson, Roya Hosseini, Arto Vihavainen, and Peter Brusilovsky. 2014. Investigating Automated Student Modeling in a Java MOOC. In *Proceedings of the International Conference on Educational Data Mining*. 261–264.