8th Educational Data Mining in Computer Science Education (CSEDM) Workshop

Yang Shi
North Carolina State University
Utah State University
yshi26@ncsu.edu

Peter Brusilovsky
University of Pittsburgh
peterb@pitt.edu

Bita Akram
North Carolina State University
bakram@ncsu.edu

Thomas W. Price
North Carolina State University
twprice@ncsu.edu

Juho Leinonen
Aalto University
juho.2.leinonen@aalto.fi

Kenneth R. Koedinger
Carnegie Mellon University
koedinger@cmu.edu

Andrew Lan
University of Massachusetts Amherst
andrewlan@cs.umass.edu

ABSTRACT

There is a growing community of researchers at the intersection of data mining, AI, and computing education research. The objective of the CSEDM workshop is to facilitate a discussion among this research community, with a focus on how data mining can be uniquely applied in computing education research. For example, what new techniques are needed to analyze program code and CS log data? How do results from CS education inform our analysis of this data? The workshop is meant to be an interdisciplinary event at the intersection of EDM and Computing Education Research. Researchers, faculty, and students are encouraged to share their AI- and data-driven approaches, methodologies, and experiences where data transforms how students learn Computer Science (CS) skills. This **full-day hybrid** workshop will feature paper presentations and discussions to promote collaboration.

Keywords

Computer Science Education, Educational Data Mining, AI in Education, Learning Analytics

1. WORKSHOP GOALS

Computing is an increasingly fundamental skill for students across disciplines. It enables them to solve complex, real, and challenging problems and make a positive impact on the world. Yet, the field of computing education is still facing a range of problems, from high failure and attrition rates to challenges in training and recruiting teachers to the under-representation of women and students of color.

Advanced learning technologies, which use data and AI to

improve student learning outcomes, have the potential to address these problems. However, the domain of CS education presents novel challenges for applying these techniques. CS presents domain-specific challenges, such as helping students effectively use tools like compilers and debuggers and supporting complex, open-ended problems with many possible solutions. CS also offers unique opportunities for developing learning technologies, such as abundant and rich log data, including code traces that capture each detail of how students' solutions evolved.

These domain-specific challenges and opportunities suggest the need for a specialized community of researchers working at the intersection of AI, data mining, and computing education research. The goal of this Educational Data Mining for Computer Science Education (CSEDM) is to bring this community together to share insights for supporting and understanding learning in the domain of CS using data. This field is nascent but growing, with research in computing education increasingly using data analysis approaches and researchers in the EDM community increasingly studying CS datasets. This workshop will help these researchers learn from each other and develop the growing sub-field of CSEDM.

The workshop will build on seven successful prior CSEDM workshops at:

- the International Educational Data Mining Conference (EDM) in 2018[1],
- the International Learning Analytics and Knowledge Conference (LAK) in 2019[2],
- the International Conference on AI in Education (AIED) in 2019[3],
- the International Educational Data Mining Conference in 2020[4],

[1] sites.google.com/asu.edu/csedm-ws-edm-2018/
[2] sites.google.com/asu.edu/csedm-ws-lak-2019/
[3] sites.google.com/asu.edu/csedm-ws-aied-2019/
[4] sites.google.com/ncsu.edu/csedm-ws-edm-2020/

- the International Educational Data Mining Conference in 2021[5],
- the International Educational Data Mining Conference in 2022[6],
- the International Learning Analytics and Knowledge Conference in 2023[7].

Each of these workshops was productive and well-attended. Our past in-person workshops have been well attended, and our virtual events have had *over 100 people registered and over 70 simultaneous attendees*! The proceedings were published in CEUR[8] and Zenodo[9].

We hope to keep our momentum with an 8th CSEDM Workshop, returning to EDM in 2024. Compared with prior versions of CSEDM, we specifically added a focusing area to include contributions of large language models and generative AI in the workshop paper presentations, which aligns with the theme of EDM 2024.

The CSEDM workshop is funded by the CS-SPLICE project[10]. The CSEDM workshop will serve as a hub for researchers in the EDM community to discuss potential collaborations and identify EDM challenges in computing education. We plan to provide need-based funding support for participants, covering the cost of lodging on the workshop day and the registration fees.

2. RELEVANT TOPICS

The workshop encourages contributions from the following topics of interest:

- Predictive and descriptive modeling for CS courses
- Adaptation and personalization within CS learning environments
- Intelligent support for collaborative CS problem solving
- Machine learning approaches to analyze massive CS datasets and courses
- Online learning environments for CS: implementation, design, and best practices
- Multimodal learning analytics and combination of student data sources in CS Education
- Affective, self-regulation, and motivational modeling of students as related to CS learning
- Adaptive feedback and adaptive testing for CS learning
- Discourse and dialogue research related to classroom, online, collaborative, or one-on-one learning of CS
- Teaching approaches using AI tools
- Visual Learning Analytics and Dashboards for CS
- Network Analysis for programming learning environments

- Classification of student program code
- Natural Language Processing for CS forums and discussions
- Analysis of programming design and trajectory paths
- Recommender systems and in-course recommendations for CS learning
- Adaptive educational technology and CS pedagogy for non-majors
- Deep learning approaches for analyzing, assessing, and scaffolding programming challenges
- Generative AI and Computing Education

We will invite researchers who are interested in further exploring, contributing, collaborating, and developing data- and AI-driven techniques for building educational tools for Computer Science to submit papers on any of these topics.

3. WORKSHOP ORGANIZATION

The workshop will be organized by a team with a history of CSEDM research:

Yang Shi is a Ph.D. student and Goodnight Doctoral Fellow at North Carolina State University. He has been working towards building data-driven methods for representing program code to enhance the ability of Intelligent Tutoring Systems and benefit student modeling processes for computing education. With a focus on DM/ML approaches applied to CS education, his research interests also include Programming Language Processing, Software Analysis, and Deep Learning. He has been serving as a program committee (PC) member in conferences across multiple disciplines, including EDM, LAK, KDD, AAAI, EAAI, SIGCSE, and ITICSE.

Peter Brusilovsky is a Professor of Information Science and Intelligent Systems at the University of Pittsburgh, where he also directs the Personalized Adaptive Web Systems (PAWS) lab. He has been working in the field of adaptive educational systems, user modeling, and intelligent user interfaces for more than 30 years. He published numerous papers and edited several books on adaptive hypermedia and the adaptive Web. He is a founder of CS-SPLICE and has advanced research and infrastructure for CSEDM.

Bita Akram is a research assistant professor with the Department of Computer Science at North Carolina State University. Her research lies at the intersection of artificial intelligence and advanced learning technologies with its application on improving access and quality of CS Education. She been actively developing data-driven approaches for assessing students' CS competencies as demonstrated through their interactions with educational programming activities. She has served as the organizer and program committee for venues focused on educational data mining including EDM and CSEDM.

Thomas Price is an Assistant Professor of Computer Science at North Carolina State University. His primary research goal is to develop learning environments that automatically support students through AI and data-driven help features.

[5] sites.google.com/ncsu.edu/csedm-workshop-edm21/
[6] sites.google.com/ncsu.edu/csedm-workshop-edm22/
[7] sites.google.com/ncsu.edu/csedm-workshop-lak23/
[8] ceur-ws.org/Vol-3051/
[9] zenodo.org/communities/csedm23/
[10] cssplice.github.io/

His work has focused on the domain of computing education, where he has developed techniques for automatically generating programming hints and feedback for students in real-time by leveraging student data. He has helped organized a number of efforts at the intersection of AIED, Data Mining and CS Education, including the CS-SPLICE working group on programming snapshot representation and prior CSEDM and CS-SPLICE workshops.

Juho Leinonen is an Academy Research Fellow at Aalto University. His research focuses on creating better insight into students' learning with fine-grained learning analytics; using educational technology and artificial intelligence for personalizing course content; and using learnersourcing to create ample learning opportunities for distinct student needs. He has served on the program committee of both computing education focused and educational data mining focused conferences.

Ken Koedinger is a professor of Human Computer Interaction and Psychology at Carnegie Mellon University. Dr. Koedinger has an M.S. in Computer Science, a Ph.D. in Cognitive Psychology, and experience teaching in an urban high school. His multidisciplinary background supports his research goals of understanding human learning and creating educational technologies that increase student achievement. His research has contributed new principles and techniques for the design of educational software and has produced basic cognitive science research results on the nature of student thinking and learning. Koedinger directs LearnLab, which started with 10 years of National Science Foundation funding and is now the scientific arm of CMU's Simon Initiative. LearnLab builds on the past success of Cognitive Tutors, an approach to online personalized tutoring that is in use in thousands of schools and has been repeatedly demonstrated to increase student achievement, for example, doubling what algebra students learn in a school year. He was a co-founder of CarnegieLearning, Inc. that has brought Cognitive Tutor based courses to millions of students since it was formed in 1998, and leads LearnLab, now the scientific arm of CMU's Simon Initiative. Dr. Koedinger has authored over 250 peer-reviewed publications and has been a project investigator on over 45 grants. In 2017, he received the Hillman Professorship of Computer Science and in 2018, he was recognized as a fellow of Cognitive Science.

Andrew (Shiting) Lan is an assistant professor in the Manning College of Information and Computer Sciences, University of Massachusetts Amherst. Before that, he was a postdoctoral research associate in the EDGE Lab at the Department of Electrical Engineering, Princeton University, and received his M.S. and Ph.D. degrees in Electrical and Computer Engineering in May 2014 and May 2016, respectively, from the Digital Signal Processing (DSP) group at Rice University. His research focuses on the development of artificial intelligence (AI) and especially natural language processing (NLP) methods to enable scalable and effective personalized learning in education, covering areas such as learner modeling, personalization, content generation, and human-in-the-loop AI.

3.1 Program Committee

The 8th CSEDM Workshop's program committee will draw from members of prior program committees, including:

- Austin Cory Bart (University of Delaware, USA)
- Barbara Ericson (University of Michigan, USA)
- Collin Lynch (North Carolina State University, USA)
- Didith Mercedes Rodrigo (Ateneo de Manila University, Philippines)
- Kelly Rivers (Carnegie Mellon University, USA)
- Cliff Shaffer (Virginia Tech University, USA)
- Alan Smeaton (Dublin City University, Ireland)
- Sergey Sosnovsky (Utrecht University, Netherlands)
- John Stamper (Carnegie Mellon University, USA)
- Michael Yudelson (Chegg Inc., USA)
- Kamil Akhuseyinoglu (University of Pittsburgh, USA)
- Satabdi Basu (SRI Education, USA)
- Adam Gaweda (North Carolina State University, USA)
- Julio Guerra (Austral University, Chile)
- Arnon Hershkovitz (Tel Aviv University, Israel)
- Nguyen-Thinh Le (Humboldt-University of Berlin, Germany)
- Tanja Mitrovic (University of Canterbury, New Zealand)
- Narges Norouzi (University of Californa Santa Cruz, USA)
- Benjamin Paaßen (Humboldt-University of Berlin, Germany)
- Andy Smith (North Carolina State University, USA)
- Khushboo Thaker (University of Pittsburgh, USA)
- Eric Wiebe (North Carolina State University, USA)
- Lauri Malmi (Aalto University, Finland)
- Andrew Petersen (University of Toronto, Canada)

4. CALL FOR PARTICIPATION
We will solicit three types of research contributions:

8-page Research Papers: Original, unpublished work, addressing any of the topics of interest above.

6-page Position Papers or Work-in-progress Papers:

- Critical meta-reviews of CSEDM research and practice putting forward discussions of the vision and future research and practice directions for the CSEDM community.
- Original, unpublished work-in-progress (incomplete or ongoing work, ready for feedback, but not yet fully developed), addressing any of the topics of interest above.

2-page Descriptions of CS Tools/Datasets/Infrastructure: Researchers will present their work at CSEDM in a conversational format. Presentations might include:

- Descriptions of shareable Computer Science (CS) datasets

- Descriptions of data mining / analytics approaches applied to specifically Computer Science datasets
- Descriptions of tools or programming environments that use/produce data
- Case studies of collaboration where reproducible practices were used to integrate or compose two or more data analysis tools from different teams
- Descriptions of infrastructures that could collect and integrate data from multiple learning tools (e.g. forum posts, LMS activity and programming data)

4.1 Timeline
The CFP will be released as soon as the workshop is accepted. An approximate timeline is as follows:

- May 3: Abstract Deadline for Papers from All Tracks
- May 10: Paper Deadline for Papers from All Tracks
- June 7: Notification of acceptance for Papers from All Tracks
- June 21: Camera Ready Deadline for Papers from All Tracks
- July 14: Workshop at EDM'24

5. WORKSHOP ACTIVITIES
The workshop will be a **full day** workshop. It will primarily consist of paper presentations and discussions to facilitate collaboration. Interactive sessions include multiple parallel, short presentations, where participants can float around to the presentations they are interested in, similar to a poster session (see Section 6 for details on remote attendees).

A tentative schedule is as follows:

- **09:00 - 09:30** Introductions and logistics
- **09:30 - 10:00** Networking
- **10:00 - 11:00** Paper Presentations
- **11:00 - 11:15** Coffee Break and Discussion
- **11:15 - 12:15** Paper Presentations
- **12:15 - 13:30** Lunch
- **13:30 - 14:30** Paper Presentations
- **14:30 - 14:45** Coffee Break and Discussion
- **14:45 - 15:45** Paper Presentations or Panel Discussion or Keynote
- **15:45 - 16:30** Wrap up Discussions

6. PLANS FOR SUPPORTING REMOTE ATTENDEES
We propose a hybrid format, where participants, including presenters, can participate remotely as needed. If the conference is held fully online, we can switch to an online format, as we did in CSEDM 2022. We will take the following steps to support remote attendees:

- The workshop will occur concurrently on Zoom (or another online platform) and in-person.

- All presenters will join the Zoom meeting and share their screen while presenting, giving remote attendees full access to presentations. If possible, we will integrate microphones into the Zoom meeting as well.
- For remote presenters, we will project the Zoom meeting to the in-person participants, so they can see it presented live, and then hold a live Q&A over Zoom.
- For interactive presentations, we will ask all presenters to bring computers and share their while presenting, so that remote attendees can join via Zoom breakout rooms. Remote presenters will be remote-only, and present their work in Zoom breakout rooms.

7. SOLICITATION PLAN
Building on our growing network of contributors to prior workshops, we intend to solicit participation on the workshop through the following mailing lists and research networks:

- ACM's Special Interest Group on Computer Science Education (SIGCSE)
- Computer Science Education (CSED) research list (from the ICER community)
- European Association of Technology-Enhanced Learning (EATEL) community
- User Modeling (UM) mailing list
- Asia-Pacific Society for Computers in Education (APSCE) community
- PSLC community list
- Relevant EU project consortia
- The International Educational Data Mining Society
- The Society for Learning Analytics Research (SoLAR)
- The Learning Engineering mailing list

We will also reach out to prior contributors to CSEDM Workshops to solicit additional submissions.