Sharing and Reusing Data and Analytic Methods with LearnSphere

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

John Stamper
Carnegie Mellon University
jstamper@cs.cmu.edu

Paulo F. Carvalho
Carnegie Mellon University
pcarvalh@cs.cmu.edu

ABSTRACT: This workshop will explore LearnSphere, an NSF-funded, community-based repository that facilitates sharing of educational data and analytic methods. The workshop organizers will discuss the unique research benefits that LearnSphere affords. In particular, we will focus on Tigris, a workflow tool within LearnSphere that helps researchers share analytic methods and computational models. Authors of accepted workshop papers will integrate their analytic methods or models into LearnSphere's Tigris in advance of the workshop, and these methods will be made accessible to all workshop attendees. We will learn about these different analytic methods during the workshop and spend hands-on time applying them to a variety of educational datasets available in LearnSphere's DataShop. Finally, we will discuss the bottlenecks that remain, and brainstorm potential solutions, in openly sharing analytic methods through a central infrastructure like LearnSphere. Our ultimate goal is to create the building blocks to allow groups of researchers to integrate their data with other researchers in order to advance the learning sciences as harnessing and sharing big data has done for other fields.

Keywords: Learning metrics; data storage and sharing; data-informed learning theories; modeling; data-informed efforts; scalability.

1 WORKSHOP BACKGROUND

The use of data to improve student learning has become more effective as student learning activities and student progress through educational technologies are increasingly being tracked and stored. There is a large variety in the kinds, density, and volume of such data and to the analytic and adaptive learning methods that take advantage of it. Data can range from simple (e.g., clicks on menu items or structured symbolic expressions) to complex and harder-to-interpret (e.g., free-form essays, discussion board dialogues, or affect sensor information). Another dimension of variation is the time scale in which observations of student behavior occur: click actions are observed within seconds in fluency-oriented math games or in vocabulary practice, problem-solving steps are observed every 20 seconds or so in modeling tool interfaces (e.g., spreadsheets, graphers, computer algebra) in intelligent tutoring systems for math and science, answers to comprehension-monitoring questions are given and learning resource choices are made every 15 minutes or so in massive open online courses (MOOCs), lesson completion is observed across days in learning management systems, chapter/unit test results are collected after weeks, end-of-course completion and exam scores are collected after many months, degree completion occurs across years, and long-term human goals like landing a job and achieving a good income occur across lifetimes. Different paradigms of data-driven education research differ both in the types of data they tend to use and in the time scale in which that data is collected. In fact, relative isolation within disciplinary silos is fostered and fed by differences in the types and time scale of data used (cf., Koedinger et al., 2012).

Thus, there is a broad need for an overarching data infrastructure to not only support sharing and use within the student data (e.g., clickstream, MOOC, discourse, affect) but to also support investigations that bridge across them. This will enable the research community to understand how and when long-term learning outcomes emerge as a causal consequence of real-time student interactions within the complex set of instructional options available (cf., Koedinger et al., 2010). Such an infrastructure will support novel, transformative, and multidisciplinary approaches to the use of data to create actionable knowledge to improve learning environments for STEM and other areas in the medium term and will revolutionize learning in the longer term.

LearnSphere transforms scientific discovery and innovation in education through a scalable data infrastructure designed to enable educators, learning scientists, and researchers to easily collaborate over shared data using the latest tools and technologies. LearnSphere.org provides a hub that integrates across existing data silos implemented at different universities, including educational technology “click stream” data in CMU’s DataShop (Stamper et al., 2011), massive online course data in Stanford’s DataStage and analytics in MIT’s MOOCdb (Veeramachaneni et al., 2014), and educational language and discourse data in CMU’s new DiscourseDB (Jo et al., 2016). LearnSphere integrates these DIBBs in two key ways: 1) with a web-based portal that points to these and other learning analytic resources and 2) with a web-based workflow authoring and sharing tool called Tigris. A major goal is to make it easier for researchers, course developers, and instructors to engage in learning analytics and educational data mining without programming skills.

The main goal of this workshop is to provide attendees with hands-on experience using Tigris for learning analytics. We hope that this year we will be able to attract attendees that have been exposed to LearnSphere from these past events, although we will have some tutorial activities included for new attendees as well. This workshop builds off a successful LAK 2018 Tutorial, and workshop at AIED/EDM 2017.

2 ORGANIZATIONAL DETAILS

2.1 Type of event

Workshop

2.2 Proposed Schedule and Duration

Table 1: Proposed Half-Day Schedule.

Time	Item
1:30p	Introductions
2:00p	Tigris workflow tool (Lecture & Demos)
2:30p	Hands-on I: Build custom analysis workflows using existing Tigris components

3:30p	Coffee Break
4:00p	Hands-on II: Breakout sessions (upload your own data; create workflow components)
4:45p	5-minute participant talks about proposed or created workflows
5:15p	Closing/High Level Discussion

2.3 Type of Participation

Mixed participation will be through submission of reviewed abstracts, invited guests, and open registration. For participants who have accepted abstracts or are invited by the workshop committee, we have allocated approximately $20,000 from our grant funding to cover registration and travel costs.

2.4 Activities

Activities will include presentations from workshop organizers, invited guests, and short presentations from accepted abstract presenters. Hands on sessions will include demos and group work towards implementing analytics.

2.5 Expected Numbers

We expect 15-20 participants based on previous workshops.

2.6 Activities to Recruit Attendees

We will create a website to announce the workshop and method of submitting abstracts. The Learning Analytics, Educational Data Mining, and LearnLab mailing lists will be used to direct potential attendees to the workshop website. In addition, we will invite a number of invited guests. Both accepted submissions and invited guests will have the chance to receive funding to attend.

2.7 Required Equipment

Projector and screen will be required by organizers. Attendees will need to bring laptops and will need adequate internet connectivity.

3 OBJECTIVES AND OUTCOMES

Broadly, this workshop offers those in the Learning Analytics community an exposure to LearnSphere as a community-based infrastructure for educational data and analysis tools. In opening lectures, the organizers will discuss the way LearnSphere connects data silos across universities and its unique capabilities for sharing data, models, analysis workflows, and visualizations while maintaining confidentiality.

More specifically, we propose to focus on attracting, integrating, and discussing researcher contributions to Tigris, the web-based workflow authoring and sharing tool. Workshop submissions in the form of abstracts will involve a brief description of an analysis pipeline relevant to modeling educational data as well as accompanying code. Prior to the workshop itself, the organizers will coordinate with authors of accepted submissions to integrate their code into Tigris. A significant portion of the workshop will be dedicated to hands-on exploration of custom workflows and workflow modules within Tigris. Authors of accepted submissions will present their analysis pipelines, and everyone attending the workshop will be able to access those analysis pipelines within Tigris to a variety of freely available educational datasets available from LearnSphere. The goal is to generate -- for each workflow component contribution in the workshop -- a publishable workshop paper that describes the outcomes of openly sharing the analysis with the research community.

Finally, workshop attendees will discuss bottlenecks that remain toward our goal of a unified repository. We will also brainstorm possible solutions. Our goal is to create the building blocks to allow groups of researchers to integrate their data with other researchers we can advance the learning sciences as harnessing and sharing big data has done for other fields.

4 REFERENCES

Jo, Y., Tomar, G., Ferschke, O., Rosé, C. P., & Gašević, D. (2016, April). Pipeline for expediting learning analytics and student support from data in social learning. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 542-543). ACM.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2010). The knowledge-learning-instruction (KLI) framework: Toward bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*.

Stamper, J., Koedinger, K.R., Baker, R., Skogsholm, A., Leber, B., Demi, S., Yu, S., Spencer, D. (2011) Managing the Educational Dataset Lifecycle with DataShop. In Kay, J., Bull, S. and Biswas, G. (eds). *Proceeding of the 15th International Conference on Artificial Intelligence in Education* (AIED2011).

Veeramachaneni, K., Halawa, S., Dernoncourt, F., O'Reilly, U. M., Taylor, C., & Do, C. (2014). Moocdb: Developing standards and systems to support MOOC data science. *arXiv preprint*. arXiv:1406.2015.