Infrastructure for Data-Driven Innovation in Education: Challenges and Future Directions

White Paper for Data Infrastructure Building Blocks (DIBBs) Workshop, Jan 12-13, 2017

By Ken Koedinger (PI, CMU, attending), Kalyan Veeramachaneni (MIT, attending), Candace Thille (Stanford), Phil Pavlik (Memphis), Una-May O'Reilly (MIT), Carolyn Rose (CMU), John Stamper (CMU)

Background and Accomplishments. *LearnSphere* is an infrastructure for sharing and collaboration across the diverse variety of educational data and analytics available today. 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, massive online course data in Stanford's DataStage and analytics in MIT's MOOCdb, and educational language and discourse data in CMU's new DiscourseDB. 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.

Developments and Discoveries. Bridging data silos from a) tutor interaction, b) MOOC resource use and outcomes, and c) discussion boards, we have employed LearnSphere to explore the variations in student behaviors that are most highly associated with learning outcomes. An analysis across 4 online courses involving over 5 million interactions of over 12,000 students revealed *a striking discovery*: Active learning activities (e.g., answering questions with feedback) are associated with *6x more learning* outcomes than are passive learning activities (e.g., lecture watching or text reading). Since the start of the LearnSphere project about two years ago, the number of data sets available has doubled to over 1,300 now, DiscourseDB was developed from scratch, MOOCdb has grown, a distributed version of DataShop was developed and demonstrated with a new instances at Memphis, and we released the alpha version of Tigris, the workflow development tool. Workflow components have been contributed by multiple researchers and these correspond with published learning science research results or practical insights from these analytics.

Challenges & Future Directions. For an R&D community so broadly inclusive of learning scientists, innovators, and educational practitioners in many disciplines, it is not practical to deliver a single data schema that will ideally satisfy many diverse analysis needs. So, a key challenge for us is how to balance uniformity (i.e., toward maximum reusability) and flexibility (e.g., allowing representations best adapted to nature of data; fit to user experience in existing languages or representations). With Tigris we are making progress on this challenge. Flexibility is facilitated by allowing the contribution of any analytic component that processes any data representation and that component can be written in any programming language (e.g., Java, C, R, Python, Matlab). Representational uniformity (aka standards) is seeded by pre-existing representations (from DataShop, DiscourseDB, and MOOCdb) and will further emerge, or usefully diverge, as the community of users contributes new components. A future direction is to scale-up workflow tool use to test whether these strategies will meet this *uniformity-flexibility challenge*.

A second challenge in this learning data space is how to maximize sharing of human data without sacrificing privacy. Many privacy control technologies have been proposed and are potentially relevant, but better understanding of the privacy concerns and needs of the controllers of educational data is critical and is an immediate future direction for this *sharing-privacy challenge*.

A third challenge is balancing a *sophistication-understanding* trade-off between advancing the sophistication and variety of analytic DIBBs available to users versus having a smaller set of well-documented set of analytics that users can understand and trust. A fourth challenge is offering a *flexible "need for speed"* when a complex job is impractical within the default processing capabilities -- a challenge that might be addressed by a smooth integration of cloud services that is transparent to the user.