

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 14:06:19

PAGE 1

REFERENCE NO: 248

This contribution was submitted to the National Science Foundation as part of the NSF CI 2030 planning activity through an NSF Request for Information, https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17031. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

Author Names & Affiliations

- Ryan Shaun Baker - University of Pennsylvania and International Educational Data Mining Society
- Mingyu Feng - SRI International and International Educational Data Mining Society
- Neil Heffernan - Worcester Polytechnic Institute and International Educational Data Mining Society
- Collin F. Lynch - North Carolina State University and International Educational Data Mining Society
- Sidney D'Mello - University of Notre Dame and International Educational Data Mining Society
- John Stamper - Carnegie Mellon University and International Educational Data Mining Society

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Education, Education Research, Learning Sciences, Science of Learning, Educational Data Mining, Learning Analytics

Title of Submission

Research Opportunities and Challenges for Cyberinfrastructure in Education

Abstract (maximum ~200 words).

The increasing quantity and quality of data available in education has created an opportunity for enriching the science of learning and educational practice. Recent work has used automated models of student knowledge, as well as complex models of affect, engagement, and self-regulation to study both moment by moment learning and longitudinal student success. However, work in this area remains limited by the available cyberinfrastructure. Enriching facilities for analyzing data that is infeasible to deidentify, and improving support for sharing data analysis software has the potential to expand the reach of educational data mining and learning analytics research. In particular, improving educational cyberinfrastructure will make it possible to understand how scientific findings on engagement and learning apply across populations, contexts, and domains and how these insights can be used to improve learning outcomes more broadly.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Over the last decades, the science of learning and education has increased in rigor and quantitative precision, due (in no small part) to the advent of large-scale educational data sets. These advances have been facilitated by the support of the National Science Foundation for Cyberinfrastructure initiatives like the Pittsburgh Science of Learning Center DataShop (now part of LearnSphere), the ADAGE project for logging game-based learning, and the ASSISTments project for conducting educational experimentation rapidly, in the real-world, and at

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 14:06:19

PAGE 2

REFERENCE NO: 248

scale.

Just over the last few years, there have been seminal findings in understanding many key phenomena, such as what forms of engagement, meta-cognition, and self-regulated learning impact learning, and how short-term learning during a learning activity relates to that student's retention of what they have learned, their preparation for future learning, and their long-term interest and involvement in the domain.

Much of this work has been facilitated by the increasingly large data sets becoming available from online learning, where each individual student may generate tens of thousands of educationally meaningful actions over the course of a year of schooling (or across a more brief online course or while using an online learning environment at home).

This has led education researchers and learning scientists to begin to connect educational and learning phenomena across several orders of magnitude, from the level of second-by-second student behaviors and cognition during learning, to the development of robust knowledge that prepares students for future learning, to a student's eventual choices about their career and their academic preparation to enter into that career.

At the micro-scale, there has been considerable work over the last decade to better measure student knowledge, understanding, and learning. An overly-simple model, Bayesian Knowledge Tracing, was dominant both in research and practice from the early 1990s until around 2008. However, at that point, a frenzy of research (which has continued to this day) has focused on integrating knowledge from fields such as cognitive science, psychometrics, and machine learning in order to better model student knowledge in real-time. The latest approaches have bridged around 70% of the difference between the state of the art in the 1990s and perfect inference -- indicating that there is still much work to do.

These computational knowledge models have been applied within online learning platforms that optimize student activity to improve skills, memory, and conceptual understanding. They have also become a component of assessment models that many think will eventually replace standardized examinations with less disruptive ongoing learning activities that measure knowledge with equal accuracy.

Furthermore, knowledge models have become a key component of computer models of student engagement, affect, meta-cognition, and self-regulated learning which help us to understand when, why, and how a student becomes disengaged or shifts their learning strategies. These models can be used to respond to student disengagement in real time; to provide information to a teacher on which students need greater support, or information to a regional curricular coordinator on which teachers need further professional development on teaching with technology; and to infer which aspects of a student's behavior or approach to learning today may have consequences for their success tomorrow. For example, prediction models integrating across student knowledge and engagement in middle school mathematics have been able to predict which students are at risk of not attending college or of dropping out of the STEM pipeline. Beyond this, models predicting which students will achieve a poor grade or drop their current course have impacted practice in undergraduate education, with large-scale vendors such as Civitas providing services to dozens of colleges.

Despite these successes, many of the more complex models and approaches have not yet scaled to the full range of students, domains, and platforms.

As these models advance, they are increasingly used to answer questions relevant to basic understanding but targeted towards practical outcomes, such as:

? How does the design of learning content impact student affect, engagement, self-regulated learning strategy, and ultimately robust learning?

? What are the processes that shape the development of interest and positive emotions towards effortful learning over time?

? Which student strategies, in which situations, for which students, lead to better long-term learning outcomes?

? And ultimately, what are the processes that determine long-term student success, both academically, and in terms of developing the habits of mind that will lead to participating in careers in STEM and other fulfilling and societally productive work? And where can we make nudges that can have positive impacts on student outcomes?

By developing frameworks that can help us study these research questions across contexts, for the full diversity of American learners, we may be able to achieve transformative effects on students' lives and in turn major positive impacts on American economic competitiveness over the next generation.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

As discussed above, there has been considerable progress in the science of learning and education research over the last decade. Much of the recent progress has been due to the advent of large-scale educational data sets, facilitated by the availability of research

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 14:06:19

PAGE 3

REFERENCE NO: 248

cyberinfrastructure supported by the National Science Foundation. Considerable progress has been made on fundamental scientific questions in this area in large part due to the support of the National Science Foundation in developing data repositories and research frameworks such as the Pittsburgh Science of Learning Center DataShop, LearnSphere, and ADAGE.

However, several limitations remain and these have hampered efforts in this area to reach their full potential. At the current time, the cyberinfrastructure available to support education research and research into the science of learning remains insufficient for the challenges and needs of the next generation of the science of learning. Fortunately, it may be possible to address many of these limitations through enhancements of the field's cyberinfrastructure.

One key issue is that the phenomena of interest are both general in relevance, and longitudinal in scope. Yet many of the programs of research working in this area have remained localized to single learning platforms, whether intelligent tutoring systems like the Cognitive Tutors, homework and teacher support platforms like ASSISTments, open educational resources such as Khan Academy, learning-by-teaching environments like Betty's Brain, games like Impulse and Physics Playground, MOOC platforms like edX and Coursera, or online learning platforms like Moodle and Blackboard.

Although reasonably large number of findings -- primarily related to basic cognition and memory, or to the relationship between specific behaviors and learning -- have been shown to generalize across platforms and contexts, many other findings have been discovered in a single context and have proven difficult to test for generalizability. This is particularly common in cases where sophisticated computer models of affect, meta-cognition, or self-regulated learning behavior are leveraged to answer questions about these constructs. Models or algorithms often have to be re-implemented several times by different research groups, slowing progress, hindering understanding of failed replication (Is a failed replication due to a genuine difference between contexts or model implementation differences?), and obscuring which phenomena are general and which are limited to local aspects of student populations or design choices.

The difficulty in replicating and studying the generalizability of findings comes in part from the complexity of current analyses and models used in modern learning analytics and educational data mining research. Although some journals in other fields, such as Science, require the sharing of code samples as part of the ground rules for publication, this is neither a standard expectation for education journals, nor is it fully sufficient for replicating, as analyses are equally dependent on data formats as software, and transforming a data set into the format relevant to an arbitrary researcher's code may be challenging. Difficulties can also occur when new data has slightly different attributes than older data despite having an identical format; poorly engineered research software produce unknown errors or incorrect output in these new contexts.

Furthermore, while a small number of learning platforms such as Cognitive Tutors and ASSISTments share their data widely through the auspices of open platforms like the Pittsburgh Science of Learning Center DataShop, most online learning data is only available to a small number of researchers with direct connections to the organizations producing data.

There are several reasons for why data is not shared more widely. One of the key reasons is concerns about student privacy. It is challenging to fully deidentify educational data, a necessity for some types of educational data (governed by legislation such as FERPA), and an important privacy consideration for all forms of educational data. Discussion forum data from online courses, for instance, is currently technically intractable to fully deidentify, as students may disclose a considerable amount of personal information about themselves in their posts to discussion forums. However, even data from online homework where a student simply enters numerical answers may be reidentifiable -- if for, example, the student posts on twitter that they obtained a relatively uncommon answer. Some have argued that it is impossible to fully deidentify educational data, as it is usually possible to reidentify a data set by connecting it with other data sources. Additionally, the public-private intersection of educational practice can make data sharing difficult -- fully sharing some data sets may create the risk of an adaptive learning or computer adaptive testing algorithm, or a knowledge-engineered domain structure, being reverse-engineered. Many companies consider this information part of their intellectual property.

Addressing these challenges through enhanced cyberinfrastructure for education research would both facilitate researchers in conducting analyses on learning and education, as well as speeding the replication and generalization of findings across contexts. In particular, cyberinfrastructure is needed that:

? Makes it possible for a range of researchers to work with data that cannot be fully deidentified, in a fashion that enables analysis while preventing researchers from being able to access personally identifying information

? Supports the application of novel and context-specific deidentification methods where feasible, to forward the study of how to effectively deidentify student data

? Facilitates the sharing and reuse of data analysis software and processes across a range of sources of data and types of analysis

? Facilitates storing a range of diverse data in common formats, including contextual data and meta-data, and linking heterogeneous data sources involving the same students (such as both learning process and learning outcome data) to each other, to facilitate longitudinal analysis

? Facilitates validating models of complex constructs (such as affect, engagement, self-regulated learning) across populations, domains, and learning platforms

? Preserves the interests of all of the individuals and organizations (learning platforms, researchers contributing analytical software, educational organizations, the students themselves, and research funders) who contribute to the shared endeavor

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 14:06:19

PAGE 4

REFERENCE NO: 248

By developing education research cyberinfrastructure that achieves these goals, we will create a scientific community that is better able to answer the scientific questions of the next generation of the science of learning, and resultantly is better able to positively impact student outcomes.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Another key consideration for the future of the science of learning is better resources for building capacity and workforce for educational data mining and learning analytics. Currently, despite literally hundreds of job openings in industry for these areas, not to mention a large number of job openings in universities, there remain only a small number of Masters programs, no focused doctoral programs, two Massive Online Open Courses (Big Data and Education and Practical Learning Analytics), and a summer institute. Practitioners and scientists in this area need a range of skills: data mining and data science, educational design, educational psychology, and the learning sciences, sociology and educational policy, and the ability to connect to needs and practices in educational settings. Building capacity in this field is a necessity for growing this research community to the size it will need to be to fully serve all the organizations looking to use big data to enhance education.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."
-