Data-Intensive Research in Education: Current Work and Next Steps

Report on Two National Science Foundation-Sponsored Computing Research Education Workshops

Edited by Chris Dede, Harvard University

Executive Summary

A confluence of advances in the computer and mathematical sciences has unleashed an unprecedented capability for enabling decision-making based on insights from new types of evidence. However, while the use of data science has become well established in business, entertainment, and science, technology, engineering, mathematics (STEM), the application of data science to education needs substantial research and development. Beyond the potential to enhance student outcomes through just-in-time, diagnostic data that is formative for learning and instruction, the evolution of educational practice could be substantially enhanced through data-intensive research, thereby enabling rapid cycles of improvement. The next step is to accelerate advances in every aspect of education-related data science so that we can transform our ability to rapidly process and understand increasingly large, heterogeneous, and noisy datasets related to learning.

That said, there are puzzles and challenges unique to education that make realizing this potential difficult. In particular, the research community in education needs to evolve theories on what various types of data reveal about learning and therefore what to collect; the problem space is too large to simply analyze all available data and attempt to mine it for patterns that might reveal generalizable insights. Further, in collecting and analyzing data, issues of privacy, safety, and security pose challenges not found in most scientific disciplines. Also, education as a sector lacks much of the computational infrastructure, tools, and human capacity requisite for effective collection, cleaning, analysis, and distribution of big data.

In response to these opportunities and challenges, the Computing Research Association held a two-workshop sequence on data-intensive research for the National Science Foundation (NSF) and the field. Insights from relatively mature data-intensive research initiatives in the sciences and engineering (first workshop) could aid in advancing nascent data-intensive research efforts in education (second workshop). Details about the agendas for these events and their presentations can be found at http://cra.org/towards-big-steps-enabled-by-big-data-science/. This report summarizes ideas and insights from these workshops, focusing on the second workshop.

The following definitions for "big data," "data-intensive research," and "data science" are used in this report, with the full understanding that delineations for these terms are not universally accepted, are still developing, and are heavily contextual:

> Big data is characterized by the ways in which it allows researchers to do things not possible before (i.e., big data enables the discovery of new information, facts, relationships, indicators, and pointers that could not have been previously realized).

Data-intensive research involves data resources that are beyond the storage requirements, computational intensiveness, or complexity that is currently typical of the research field.

Data science is the large-scale capture of data and the transformation of those data into insights and recommendations in support of decisions.

The four “Vs” often used to describe what makes data big are (1) the size of data (*volume*); (2) the rate at which data is produced and analyzed (*velocity*); (3) its range of sources, formats, and representations (*variety*); and (4) widely differing qualities of data sources, with significant differences in the coverage, accuracy, and timeliness of data (*veracity*).

Held in January 2015, the first workshop, "Towards Big Steps Fostered by Big Data Science," focused on determining the conditions for success in data-intensive research by studying effective partnerships within science and engineering. NSF-sponsored exemplary projects from geological, engineering, biological, computational, and atmospheric sciences were featured in order to categorize data-intensive research within these fields. This report presents five case studies from earth sciences, biological sciences, health sciences informatics, computer sciences – visualization, and astronomical sciences.

The report’s discussion of those cases is focused primarily on promising strategies for data-intensive research in education, which include:

- **Collaborate With Other Fields.** Data-intensive research, even for one specific goal, requires interdisciplinary collaboration, but often methods developed for data-intensive research in one field can be adopted in *other* fields, thus saving time and resources, as well as advancing each field faster.
- **Develop Standards, Ontologies, and Infrastructure.** In addition to common language among research groups through ontologies, the interoperability of standards, and shared infrastructure for data storage and data analysis is key. Also, it is highly beneficial when companies have incentives to make their data available and collaborate with academics.
- **Provide Structure, Recognition, and Support for Curation.** This includes (1) facilitating the exchange of journal publications and the databases, (2) developing a recognition structure for community-based curation efforts, and (3) increasing the visibility and support of scholarly curation as a professional career.
- **Transfer and Adapt Models From the Sciences and Engineering.** Data-intensive research strategies effective in the five STEM cases in the first workshop provide insights for educational researchers who face similar challenges with the nature of the data they collect and analyze.

Federal agencies have played an important role in the development of data-intensive research in STEM fields. Key activities have included supporting the infrastructure needed for data sharing, curation, and interoperability; funding the development of shared analytic tools; and providing resources for various types of community-building events that facilitate developing ontologies and standards, as well as transferring and adapting models across fields. All of these strategies could also apply to federal efforts aiding data-intensive research in education.

Held in June 2015, the second workshop, "Advancing Data-Intensive Research in Education," focused on discussing current data-intensive research initiatives in education and applying heuristics from the sciences and engineering to articulate the conditions for success in education research and in models for effective partnerships that use big data. The event focused on emergent data-intensive research in education on these six general topics:

- Predictive Models based on Behavioral Patterns in Higher Education
- Massively Open Online Courses (MOOCs)
- Games and Simulations
- Collaborating on Tools, Infrastructures, and Repositories
- Some Possible Implications of Data-intensive Research for Education
- Privacy, Security, and Ethics

Breakout sessions focused on cross-cutting issues of infrastructure, building human capacity, relationships and partnerships between producers and consumers, and new models of teaching and learning based on data-rich environments, visualization, and analytics. A detailed analysis of each of these topics is presented in the body of this report.

Overall, seven themes surfaced as significant next steps for stakeholders such as scholars, funders, policymakers, and practitioners; these are illustrative, not inclusive of all promising strategies. The seven themes are:

Mobilize Communities Around Opportunities Based on New Forms of Evidence: For each type of data discussed in the report, workshop participants identified important educational issues for which richer evidence would lead to improved decision-making. The field of data-intensive research in education may be new enough that a well-planned common trajectory could be set before individual efforts diverge in incompatible ways. This could begin with establishing common definitions; taking time to establish standards and ontologies may immensely slow progress in the short-term, but would pay off once established. In addition, if specific sets of consumers can be identified, targeted products can be made, motivated by what's most valuable and most needed, rather than letting the market drive itself.

Infuse Evidence-Based Decision-Making Throughout a System: Each type of big data is part of a complex system in the education sector, for which pervasive evidence-based decision-making is crucial to realize improvements. As an illustration of this theme, data analytics about instruction can be used on a small scale, providing real-time feedback within one classroom, or on a large scale, involving multiple courses within an organization or across different institutions. In order to determine and thus further increase the level of uptake of evidence-based education, a common set of assessments is necessary for straightforward aggregation and comparison across experiments in order to reach stronger conclusions from data-intensive research in education.

Develop New Forms of Educational Assessment: Novel ways of measuring learning can dramatically change both learning and assessment by providing new forms of evidence for decision-making to students, teachers, and other stakeholders. For example, Shute's briefing paper describes **"**continually collecting data as students interact with digital environments both inside and, importantly, outside of school. When the various data streams coalesce, the accumulated information can potentially provide increasingly reliable and valid evidence about what students know and can do across multiple contexts. It involves high-quality, ongoing, unobtrusive assessments embedded in various technology-rich environments (TREs) that can be aggregated to inform a student's evolving competency levels (at various grain sizes) and also aggregated across students to inform higher-level decisions (e.g., from student to class to school to district to state, to country)."

Reconceptualize Data Generation, Collection, Storage, and Representation Processes: Many briefing papers and workshop discussions illustrated the crucial need to change how educational data is generated, collected, stored, and framed for various types of users. Micro-level data (e.g., each student's second-by-second behaviors as they learn), meso-level data (e.g., teachers' patterns in instruction) and macro-level data (e.g., aggregated student outcomes for accountability purposes) are all important inputs to an infrastructure of tools and repositories for open data sharing and analysis. Ho's briefing paper argues that an important aspect of this is, "'data creation,' because it focuses analysts on the process that generates the data. From this perspective, the rise of big data is the result of new contexts that create data, not new methods that extract data from existing contexts."

Develop New Types of Analytic Methods: An overarching theme in all aspects of the workshops was the need to develop new types of analytic methods to enable rich findings from complex forms of educational data. For example, appropriate measurement models for simulations and games—particularly those that are open ended—include Bayes nets, artificial neural networks, and model tracing. In his briefing paper, Mitros writes, "Integrating different forms of data—from peer grading, to mastery-based assessments, to ungraded formative assessments, to participation in social forums—gives an unprecedented level of diversity to the data. This suggests a move from traditional statistics increasingly into machine learning, and calls for very different techniques from those developed in traditional psychometrics." Breakthroughs in analytic methods are clearly a necessary advance for data science in education.

Build Human Capacity to Do Data Science and to Use Its Products: More people with expertise in data science and data engineering are needed to realize its potential in education, and all stakeholders must become sophisticated consumers of data-intensive research in education. Few data science education programs currently exist, and most educational research programs do not require data literacy beyond a graduate statistics course. Infusing educational research with data science training or providing an education “track” for data scientists could provide these cross-disciplinary opportunities. Ethics should be included in every step of data science training to reduce the unintentional emotional harm that could result from various analyses.

Develop Advances in Privacy, Security, and Ethics: Recent events have highlighted the importance of reassuring stakeholders in education about issues of privacy, security, and ethical usage of any educational data collected. More attention is being paid to explicit and implicit bias embedded in big data and algorithms and the subsequent harms that arise. Hammer’s briefing paper indicates that “[e]ach new technology a researcher may want to use will present a unique combination of risks, most of which can be guarded against using available technologies and proper information policies. Speaking generally, privacy can be adequately protected through encrypted servers and data, anonymized data, having controlled access to data, and by implementing and enforcing in-office privacy policies to guard against unauthorized and exceeded data access.” A risk-based approach, similar to the approach taken by the National Institute of Standards and Technologies in guidelines for federal agencies, would allow for confidentiality, consent, and security concerns to be addressed commensurate with the consequences of a breach.

In summary, this report documents that one of the most promising ways society can improve educational outcomes is by using technology-enabled, data-intensive research to develop and apply new evidence-based strategies for learning and teaching, inside and outside classrooms. By adapting and evolving from the foundations for data-intensive research in the sciences and engineering, educators have a golden opportunity to enhance research on learning, teaching, and schooling. To realize the potential of these approaches, the many strategies described in this report will be most effective if applied together rather than in a piecemeal manner. Further, progress will be most rapid if these strategies are implemented in a coordinated manner by all stakeholders (i.e., funders, policymakers, researchers, practitioners, families, and communities), rather than in relative isolation.

Table of Contents

Data-Intensive Research in Education: Current Work and Next Steps