Next steps for "Big Data" in education: Utilizing data-intensive research

Next Steps for "Big Data" In Education: Utilizing Data-Intensive Research.

Chris Dede
Harvard Graduate School of Education
13 Appian Way
Cambridge MA, 02138
617-495-3830
chris_dede@harvard.edu

Dede, C. (2016). Next steps for "Big Data" in education: Utilizing data-intensive research. *Educational Technology* LVI(2): 37-42.

Author's Final Version

ABSTRACT

Data-informed instructional methods offer tremendous promise for increasing the effectiveness of teaching, learning, and schooling. Yet-to-be-developed data science approaches have the potential to dramatically advance instruction for every student and to enhance learning for people of all ages. Next steps that emerged from a recent NSF funded Computing Research Association workshop on data-intensive research in education were: 1) mobilize communities around opportunities based on new forms of evidence, 2) infuse evidence-based decision-making throughout a system, 3) develop new forms of educational assessment, 4) re-conceptualize data generation, collection, storage, and representation processes, 5) develop new types of analytic methods, 6) build human capacity to do data science and to use its products, and 7) develop advances in privacy, security, and ethics. If these steps are taken, participants agreed that data science approaches have the potential to dramatically advance instruction for every student and to enhance learning for people of all ages. This article briefly summarizes three of these themes that are particularly relevant, yet have not received as much attention as they deserve.

INTRODUCTION

In June, 2015, the National Science Foundation (NSF) sponsored a Computing Research Association (CRA) workshop on data-intensive research in education. This article summarizes insights about next steps from that workshop, articulated in its report, *Data-Intensive Research in Education: Current Work and Next Steps* (Dede, 2015). 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. 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 many 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. This article articulates workshop participants' of next steps needed to overcome these challenges and realize these educational opportunities.

DEFINITIONS

The following definitions for "big data," "data-intensive research," and "data science" are used in this article, with the understanding that delineations for these terms are not universally accepted, are still developing, and are 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 realized previously).

Data-intensive research involves data resources that are beyond the storage requirements, computational intensiveness, or complexity that is currently typical of the research field. Recently, data-intensive research has been described as the fourth paradigm of scientific discovery where data and analysis are interoperable (Hey, Tansley, & Tolle, 2009). The first two

paradigms of traditional scientific research refer to experimentation and theory, while the third encompasses computational modeling and simulation (Strawn, 2012).

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

Big Data in the Context of Education Research

Education research could greatly benefit from increased investment in the data and computing revolution. Less than 1% of total national K-12 expenditures are targeted to research and development, which deprives the educational community of tools and strategies to provide students with the best possible education. While Internet companies have devoted significant resources to analyze large volumes of consumer data and provide a more personalized experience, researchers are looking to explore whether similar techniques are applicable to education. To better support these innovations and next-generation learning technologies, the White House Administration has proposed several data-intensive actions in educational research. In its February 2015 report, "Investing in America's Future: Preparing Students with STEM Skills," (White House Office of Science and Technology Policy, 2015), the Administration announced its continued support for the Department of Education's Institute of Educational Sciences (IES) initiative, the Virtual Learning Laboratory, which explores "the use of rapid experimentation and 'big data' to discover better ways to help students master important concepts in core and academic subjects."

In December 2013, The President's Council of Advisors in Science and Technology (PCAST) noted that research support for Massive Open Online Courses (MOOCs) and related educational technologies offer opportunities to capture massive amounts of real-time data to expand research opportunities in learning, including those associated with gender, ethnicity, economic status, and other subjects (Executive Office of the President, 2013). PCAST recommended sponsoring a national center for high-scale machine learning for these growing educational data sets, as well as the development of competitive extramural grants to accelerate the improvement of educational materials and strategies to lead to customizable curricula for different types of students. Through these reports and recommendations, the Administration recognizes that capitalizing on America's STEM investments requires increased support for data-intensive research in education.

As illustrated in sciences and engineering, federal agencies have played an important role in the development of data-intensive research. 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 also could apply to federal efforts aiding data-intensive research in education.

PERVASIVE THEMES ABOUT DATA-INTENSIVE RESEARCH IN EDUCATION

The report documents strategies that repeatedly emerged across multiple briefing papers and in workshop discussions. Seven themes that surfaced as significant next steps for stakeholders such as scholars, funders, policymakers, and practitioners; these themes are illustrative, not inclusive of all promising strategies. They are:

- Mobilize communities around opportunities based on new forms of evidence
- Infuse evidence-based decision-making throughout a system
- Develop new forms of educational assessment
- Re-conceptualize data generation, collection, storage, and representation processes
- Develop new types of analytic methods
- Build human capacity to do data science and to use its products
- Develop advances in privacy, security, and ethics

This article summarizes three of these themes that are particularly relevant, yet have not received as much attention as they deserve.

Mobilize communities around opportunities based on new forms of evidence

Data-intensive educational research is a means, not an end in itself. Data science applied to education should not be framed as a solution looking for a problem, but instead as a lever to improve decision-making about perennial issues in teaching, learning, and schooling. Briefing paper authors and workshop participants identified important educational issues for various educational data types (e.g., MOOCs, games and simulations, tutoring systems, and assessments) where richer evidence would lead to improved decision-making. They suggested various areas of "low-hanging fruit" for data-intensive research: important educational problems about which data is already being collected and stored in repositories that have associated analytic tools. To advance data science in education, proofs of concept seem an important next step for the field, and studying perennial educational challenges brings in other stakeholders as both advocates and collaborators.

Often, when it comes to integrating data from multiple sources or when dealing with extremely large data sets, the data producers are not the data consumers, and sometimes the distinction between producer and consumer is unclear. For example, when citizen scientists are involved, they can be seen as the data producer, but also as a data consumer, in the sense that, for citizen science to be successful, the research team has to provide the raw data to be analyzed, provide compelling research questions that clearly need human processing, and keep the citizens informed and up to date on the findings. In education, networked improvement communities are an example where it is unclear who is the producer and who is the consumer.

The types of partnerships where data consumers use data produced from a range of sources come with benefits and drawbacks. When the producer is a single, well-established public database or analytical toolset, the data are usually more standardized, trustworthy, and indefinitely accessible; however, there may be less room for customization. Alternatively, when the "producers" are patients, students, or retail consumers, they are often unaware that they are data producers, leading to a potential for both societal harm and good from this situation.

Students as producers. Although students are typically thought of as consumers in an educational setting, in terms of data produced to study learning and pedagogy, they are also producers. One of the central issues to this producer/consumer relationship is protection of the "producer" and the level of data that should be shared. This issue highlights how data scientists need both highly technical and highly ethical training. Aside from privacy concerns, there is the issue that both academia and industry are consumers of student success data and the associated variables, and these two groups of consumers have different goals. However, there are times when their objectives align, and there is a gap to bridge between good research findings and how to get them to market.

Teachers as producers. Another complex type of educational data produced is teacher evaluations, which are mainly "consumed" by administrators. At some universities, such as Harvard, incoming students are also able to "consume" these evaluations to inform their course selections. In this type of producer/ consumer relationship, as with the first example above, the producers and two types of consumers may have quite different, and sometimes opposing, interests. For example, deans and faculty may prefer highly challenging courses with high student engagement, while students may prefer courses that are enjoyable, educational, and a potential boost to their GPA, on which much of their future may depend.

Even when the interests of educational producers and consumers are aligned, it is often difficult to identify the true causal relationships when there are so many covariates, and if data are missing at key points in the system. An example of missing data could be factors outside the classroom that lead to poor performance. Poor performers may not need remediation; instead, the issues preventing success may lie beyond the classroom door. Moreover, even when the statistically significant metrics are identified and can be accurately measured, it is important to anticipate how the metrics change over time.

Education technologies as producers. There is a large amount of innovative technology being produced to enhance teaching and learning. This particular type of producer/ consumer relationship will break down if the technology simply "bounces off the walls of education." The consumers (teachers or students, depending on the specific technology) need straightforward strategies to implement the technologies, whether via data coaches in schools or train-the-trainer courses. There is a gap in research on how to implement new technologies, especially those that face barriers related to perceived threats in privacy or security.

The overall goal of all of these producer/consumer relationships and partnerships in education is to efficiently use big data to optimize student success. For example, when interpreted and used correctly, data-intensive research can inform for which students it is best to use educational games, under what circumstances it is best to use flipped classrooms, how best to implement new technologies, and how to "intervene" when the implementation is not working. It can be said, in data science, often producers want to do better things while the consumers want to do things better. That is, consumers may use the results of data-intensive research to drive new types of production with new sources of data.

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, which will be a difficult task considering the many producers and consumers with unique goals. Also, some decisions for setting the trajectory will be immediate, tactical choices, while others need to be "mission decisions," which may not be immediately beneficial, but will pay off in the long run. For example, taking time to establish standards and ontologies may immensely slow progress in the short-term, but once established, would pay off. 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.

Essentially, defining consumers and building tailored products creates a pull from the community, rather than pushing new data-products onto them. Anticipating community needs in advance is important, as opposed to waiting until a problem exists and then asking groups, like the government, to intervene. As part of developing the common trajectory for data-intensive educational research, federal agencies could provide resources and policies to assist the field in defining consumers and creating products based on their needs.

Infuse evidence-based decision-making throughout a system

Like many other innovations, "build it and they will come" is not a good way to achieve widespread utilization for data-intensive research in education. Instead, new forms of evidence should be infused throughout educational decision-making systems. As an illustration, data-intensive research could help educational stakeholders make better decisions, obtain deeper and better insights, and find new patterns that can help provide new understandings about learning and cognition through predictive and prescriptive analytics. Predictions are of greater institutional value when tied to treatments and interventions for improvement and to evaluations to ensure results are being delivered. Additionally, outputs from these systems are valuable for their users when presented with intuitive visualizations and embedded workflows.

In a briefing paper for the workshop, Piotr Mitros discusses how big data in education has the potential to provide a variety of opportunities to: (1) individualize a student's path to content mastery, through adaptive learning or competency-based education; (2) improve learning as a result of faster and more in-depth diagnosis of learning needs or course trouble spots, including the assessment of skills such as systems thinking, collaboration, and problem-solving in the context of deep, authentic subject-area knowledge assessments; (3) target interventions to improve students' success and reduce overall costs to students and institutions; (4) use game-based environments for learning and assessment, where learning is situated in complex information and decision-making situations; (5) provide a new credentialing paradigm for the digital ecosystem, integrating micro-credentials, diplomas, and informal learning in ways that serve individuals and employers; and (6) enable academic resource decision-making, such as managing costs per student credit hour; reducing D, Fail, Withdraw (DFW) rates; eliminating bottleneck courses; aligning course capacity with changing student demand; and more.

This perspective is a substantial shift from current practices in using data for educational decision-making. Data analytics can be used on the small scale, to provide real-time feedback within one classroom, or on the large scale. Increasing the uptake of evidence-based education could be achieved in several ways. One way could be to focus first on the small percentage of teachers who are readily willing to use evidence-based techniques. Then gradually it will become evident to others that, even if a study on a successful teaching practice was conducted at a different institution or in a different domain, the methods and findings may still be applicable to them. Another way to increase uptake is to send the message that evidence-based education benefits both students and teachers, with continuous data-based improvement for classroom management and teaching, often coupled with a decrease in workload for the teachers.

This diffusion of innovation needs to be both top-down (new technologies produced and implemented) and bottom-up (teachers exhibiting a strong need for a new system and taking action). The bottom-up approach is a way to ensure that the tools produced are actually the ones in highest demand. In order to encourage this bottom-up approach, IES is funding competitions for networks to look at college completion at community colleges, with the goal of understanding what practitioners need from researchers. Even with a top-down approach, personal communication is most effective. Teachers generally take recommendations on new tools from other teachers or individuals they trust. Therefore, increasing interactions among teachers, researchers, and practitioners would be highly beneficial.

Adopting evidence-based education requires a common set of assessments, one of the themes discussed in the full report. The rate of degree-completion is an insufficient measurement because many professions require skills instead of a degree, so a different metric for competency is necessary. Standardized assessments would allow for straightforward aggregation and comparison across studies, thus strengthening findings from data-intensive research in education.

Re-conceptualize data generation, collection, storage, and representation processes

An opportunity for data science in education is to extend the range of student learning data that is both generated and collected. Mitros' briefing paper discusses how large volumes of data can be gathered across many learners (broad between-learner data), but also within individual learners (deep within-learner data). Data derived from MOOCs includes longitudinal data (dozens of courses from individual students over many years), rich social interactions (such as videos of group problem solving over videoconference), and detailed data about specific activities (such as scrubbing a video, individual actions in an educational game, or individual actions in a design problem). The depth of the data is determined not only by the raw amount of data on a given learner, but also by the availability of contextual information.

In a briefing paper for the workshop, Andrew Ho indicates that an emphasis on "data creation" is crucial because it focuses analysts on the process that generates the data. The development of a MOOC, an online educational game, a learning management system, or an online assessment enables the creation of data in a manner that enables its collection. Projects that focus on "data intensive" or big data orientation need to describe the contexts and processes that generate the data.

An additional approach for determining what data to generate is Evidence Centered Design (ECD). In a briefing paper for the workshop, Eric Klopfer delineates how ECD defines four relevant models: (1) the *student* model (what the student knows or can do); (2) the *evidence* model (what a student can demonstrate and we can collect to show what they know); (3) the *task* model (the designed experience from which we can collect data); and (4) the *presentation* model (how that actually appears to the student).

Although developers originally conceived of ECD to create better and more diverse assessments, it has become popular among learning game designers for its ability to create a framework for collecting and interpreting assessment data in games.

Beyond what educational data to generate, creating infrastructures and tools for collecting and sharing this data is an important next step. In a briefing paper for the second workshop, Rick Gilmore points out that data repositories can help translate insights from scientific research into applications. Open data sharing policies bolster transparency and peer oversight, encourages diversity of analysis and opinion, accelerates the education of new researchers, and stimulates the exploration of new topics not envisioned by the original investigators. Additionally, data sharing and reuse increases the return on public investments in research and leads to more effective public policy. Researchers share interpretations of distilled, not raw, data, almost exclusively through publications and presentations. The path from raw data to research findings to conclusions cannot be traced or validated by others. Other researchers cannot pose new questions that build on the same raw materials. Open data sharing addresses these challenges and promotes a culture of transparency.

George Siemens, in his briefing paper for the workshop, describes how personal learning graphs (PLeGs) provide another example of ways data might be generated, collected, and shared to create a profile of what a learner knows exists. PLeGs would enable all members involved in an educational process, including learners, faculty, and administrators, to see what a learner knows and how this is related to the course content, concepts, or curriculum in a particular knowledge space. Four specific elements included in the multipartite graphs that comprise PLeGs include: (1) social learning activities and networks; (2) cognitive development and concept mastery; (3) affectiveness and engagement; and (4) process and strategy (meta-cognition). A PLeG shares attributes of the semantic web or Google Knowledge Graph: a connected model of learner knowledge that can be navigated and assessed and ultimately "verified" by some organization in order to give a degree or designation. All stakeholders in the education system today have access to more data than they can possibly make sense of or manage. The development of PLeGs and an open data-sharing platform are critically needed innovations to contribute to and foster a new culture of learning sciences research.

New types of analytic methods are needed to enable rich findings from complex forms of educational data; breakthroughs in this area are clearly a necessary advance for data science in education. Further, there is a pressing need for both more people expert in data science and data engineering, as well as the challenge of helping all stakeholders become sophisticated consumers of data-intensive research in education Moreover, as the field of data-driven educational research expands, researchers will need to be sensitive to privacy, confidentiality, and ethical issues in their analyses. The full report includes discussions on all three of these themes.

CONCLUSION

Data science is transforming many sectors of society through an unprecedented capability for improving decision-making based on insights from new types of evidence. Workshop participant presentations and discussions emphasized that data-informed instructional methods offer tremendous promise for increasing the effectiveness of teaching, learning, and schooling. In recent years, education informatics has begun to offer new information and tools to key stakeholders in education, including students, teachers, faculty, parents, school administrators, employers, policymakers, and researchers. Yet-to-be-developed data science approaches have the potential to dramatically advance instruction for every student and to enhance learning for people of all ages.

The next step is to accelerate advances in every aspect of education-related data science so we can transform our ability to rapidly process and understand increasingly large, heterogeneous, and noisy data sets related to learning. Sections in the report offer visions of mobilizing communities around opportunities based on new forms of evidence, infusing evidence-based decision-making throughout educational systems, and developing new forms of educational assessment, and adapting data science models from STEM fields.

When something is not working well in education, doing it twice as long and twice as hard is too often the strategy tried next. Unfortunately, at present many uses of digital technologies in schooling center on automating weak models for learning rather than developing innovative, effective approaches. 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, in and out of classrooms.

I believe that is happening with data-intensive research in education is comparable to the inventions of the microscope and the telescope. Both of these devices revealed new types of data that were always present, but never before accessible. We now have the equivalent of the microscope and the telescope for understanding learning, teaching, and schooling in powerful ways. What was previously invisible can be studied and shaped, if we take the next steps outlined in the CRA report.

1. ACKNOWLEDGMENTS

The viewpoints expressed are those of the participants in the Computing Research Association workshop, not official positions of the National Science Foundation as the funder.

REFERENCES

Dede, C., Editor. (2015). *Data-intensive research in education: Current work and next steps*. Arlington, VA: Computing Research Association. http://cra.org/wp-content/uploads/2015/10/CRAEducationReport2015.pdf

Executive Office of the President. (2013). The president's council of advisors in science and technology report. https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_edit_dec-2013.pdf.

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research. Redmond, Washington.

Strawn, G.O. (2012). Scientific research: How many paradigms? *EDUCAUSE Review* 47(3). 26-34.

White House Office of Science and Technology Policy. (2015). Investing in America's future: Preparing students with STEM skills. https://www.whitehouse.gov/sites/default/files/microsites/ostp/stem_fact_sheet_2016_budget_0.pdf.