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

REPORT ON TWO NATIONAL SCIENCE FOUNDATION-SPONSORED COMPUTING RESEARCH EDUCATION WORKSHOPS

CRA
Computing Research
Association

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

CRA
Computing Research Association

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

Introduction by the Editor

Recently, the Computing Research Association's (CRA's) Computing Community Consortium commissioned a series of white papers on the subject of data science: the large-scale capture of data and the transformation of those data into insights and recommendations in support of decisions[1]. 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.

The potential impact of these advances on education is beginning to be studied. Examples of research on this topic include "A Roadmap for Education Technology" by Beverly Park Woolf[2] and "Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda"[3]. In addition, the U.S. Department of Education has recently released "Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief"[4]. These and other studies conclude that data-informed instructional methods offer tremendous promise for increasing the effectiveness of teaching, learning, and schooling. 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 accelerated through data-intensive research, thereby enabling rapid cycles of improvement.

The Internet, social and immersive media, and rich interfaces (including sensors) allow us to capture much more information about learners than ever before—and the quantities of data are growing at a rapidly accelerating rate. In recent years, education informatics (an approach to education focused on collecting, mining, and analyzing large data sets about learning) has begun to offer new information and tools to key stakeholders in education, including students, teachers, parents, school administrators, employers, policymakers, and researchers. 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. 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.

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 gather 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 that involves issues of volume, velocity, variety, and veracity. The sciences and engineering are ahead of education in developing models for how to apply data science and engineering in their

[1] "Big Data and National Priorities." CCC-Led White Papers. Ed. Helen V. Wright. Computing Community Consortium, 5 Apr. 2006. Web. 15 Sept. 2015. <http://cra.org/ccc/resources/ccc-led-whitepapers/>.

[2] Wright, Helen V. "Learning Technology - Computing Community Consortium." Learning Technology. Computing Community Consortium, 5 Apr. 2006. Web. 15 Sept. 2015. <http://archive2.cra.org/ccc/visioning/visioning-activities/221-learning-technology>.

[3] Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda. Washington, D.C.: Computing Research Association, 2005. Computing Research Association, 5 Apr. 2005. Web. 15 Sept. 2015. <http://cra.org/uploads/documents/resources/rissues/cyberinfrastructure.pdf>.

[4] Gianchandani, E. "Dept. of Education Releases Learning Analytics Issue Brief." Review. Web log post. CCC Blog. Computing Community Consortium, 10 Apr. 2012. Web. 15 Sept. 2015. <http://www.cccblog.org/2012/04/10/dept-of-education-releases-learning-analytics-issue-brief/>.

fields; this report documents that insights from STEM can offer guidance for the evolution of data-intensive research in education.

TWO WORKSHOPS ON DATA-INTENSIVE RESEARCH

In response to these opportunities and challenges, CRA held a two-workshop sequence on data-intensive research for NSF and the field. The rationale for a two-workshop sequence was that insights from relatively mature data-intensive research initiatives in the sciences and engineering (the first workshop) could aid in advancing nascent data-intensive research efforts in education (the second workshop). Both workshops were deliberately small and invitational, to enable rich discussions among researchers in the field with substantial experience in this area.

The first workshop was designed to accomplish two objectives:

1. To articulate, across the sciences and engineering, the conditions for success to achieve effective usage of data-intensive research (infrastructures; analytic methods; data scientists; research agendas; partnerships; and policies on access, security, and privacy issues)
2. To study models of effective partnerships between sources of big data and its consumers, including researchers, practitioners, and policymakers

Based on lessons learned in the sciences and engineering, the second workshop's objective was to build capacity at NSF and in the field for effectively conducting and utilizing data-intensive research in education, particularly in higher education.

THE FIRST WORKSHOP: BIG DATA IN THE SCIENCES AND ENGINEERING

The opening workshop, "Towards Big Steps Fostered by Big Data Science," focused on the first two objectives: in the sciences and engineering, determining the conditions for success in data-intensive research by studying effective partnerships. Funded through a NSF grant to CRA, this workshop was held on January 29-30, 2015, at the Virginia Tech's Arlington campus. The event centered on relatively mature data-intensive research partnerships in six areas of the sciences and engineering:

- Data-Intensive Research Projects in Earth Sciences
    - Big Data in Open Topography
    - Climate Modeling and Big Data
- Data-Intensive Research Projects in Biology
    - Big Data in Plant Genomics
- Data-Intensive Research Projects in Astronomy
    - Astrophysics as a Data-Intensive Science
    - Big Data Lessons and Citizen Science from the Zooniverse
- Data-Intensive Research Projects in CISE/Health IT
    - Electronic Health Record and Patient Social Media Research Related to Diabetes
    - The Role of Medical Experts in Health Data-Science
- Data-Intensive Research Projects in Computer Science and Engineering
    - Dynamic Data Analyses in Engineering and Computer Science
    - Immersive Exploration of Big Data Through Visualization
- Methods and Analytics for Understanding Big Data
    - Computational Capacity and Statistical Inference
    - Machine Learning

Presenters at this workshop were asked to focus on procedural aspects of interest to data-intensive researchers in other fields:

1. What type of data do your consumers want? By what mechanisms do you determine this?
2. What makes this type of data big (e.g., volume, velocity, variety, veracity)?
3. What infrastructure, funding, and policies are needed to generate this data?

4. Do you confront issues of data standards and interoperability?

5. Do you confront issues of privacy, security, and ethics?

6. What types of methods are used to analyze your data by producers? By consumers?

7. Do you confront issues of limited current capacity in the data sciences?

8. Does your project involve partnerships or other types of sustained organizational relationships?

9. How has your big data work changed your field?

10. What advice do you have for others running big data projects?

Beyond these talks and subsequent discussions, breakout sessions centered on cross-cutting issues of infrastructure; data sharing; data standards and interoperability; privacy, security, and ethics; building capacity in the data sciences; and producer and consumer relationships and partnerships. A concluding discussion focused on what's needed in research and federal initiatives, as well as advice for data-intensive research projects in education.

This workshop yielded a rich set of ideas about the conditions for success in data-intensive research in the sciences and engineering, as well as effective models for partnerships between producers and consumers of big data. Further, heuristics gained from this workshop suggested emphases, issues, and structures for the subsequent workshop on data-intensive research in education. More details about the agenda for this workshop and its speakers, including their presentation slides, can be found at http://cra.org/towards-big-steps-enabled-by-big-data-science/.

THE SECOND WORKSHOP: ADVANCING DATA-INTENSIVE RESEARCH IN EDUCATION

The succeeding 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 in education the conditions for success and models for effective partnerships that use big data. Funded through a NSF grant to CRA, this workshop was held on June 1-2, 2015 at the Waterview Conference Center in Arlington, Virginia. The event focused on emergent data-intensive research in education on these 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, producer and consumer relationships and partnerships, and new models of teaching and learning based on data-rich environments, visualization, and analytics.

Unlike the first workshop, presenters at this event prepared 1,500–2,500 word briefing papers in advance, which participants were asked to read before attending. This enabled more focus in the workshop on discussion, and the briefing papers also provided a rich foundation for this report. The questions below were a guideline for issues that speakers might choose to address in their briefing papers:

1. In your area, what type of education-related data, analyses, and findings are most valuable for scholars, practitioners, and policymakers? By what mechanisms do you determine this?

2. What makes this type of data big or "intensive" (e.g., volume, velocity, variety, veracity)?

3. What infrastructure, funding, and policies are needed to generate this data, analyses, and findings?

4. What are the key issues in data standards and interoperability?

5. What are the key issues in privacy, security, and ethics?

6. What types of methods are used to analyze this type data by producers? By consumers?

7. What technologies and tools are used and needed?

8. What are the current limits on inference and practice?

9. Do you confront issues of limited current capacity in the data sciences? What types of workforce development and training would be helpful?

10. Does your work involve partnerships or other types of sustained organizational relationships?

11. How has data-intensive research on this type of educational data changed the field thus far?

12. In your area, are there parallels to models for data-intensive research that have emerged in the sciences and engineering or in the social sciences?

13. What advice do you have for others running data-intensive research projects?

More details about the agenda for the second workshop and its speakers, including their presentation slides, can be found at http://cra.org/towards-big-steps-enabled-by-big-data-science/. Modified versions of the briefing papers are interwoven in the body of this report.

ACKNOWLEDGMENTS

CRA is grateful to the National Science Foundation for providing funding for these workshops; the viewpoints expressed are those of the participants, not official positions of the funding agency. In writing the report, the Editor is very grateful for the guidance of NSF staff: Lida Beninson, AAAS S&T Policy Fellow at NSF, Division of Computer and Networked Systems (CISE/CNS); Elizabeth Burrows, AAAS S&T Policy fellow at NSF, Division of Mathematical Sciences (MPS/DMS); John Cherniavsky, Senior Advisor, Directorate of Education and Human Resources (EHR); Anthony E. Kelly (Office of the Assistant Director, EHR/OAD); Gül Kremer, Program Officer, Division of Undergraduate Education (EHR/DUE); Taylor Martin, Program Officer, Division of Research on Learning (EHR/DRL); and Susan Singer, Division Director, Division of Undergraduate Education (EHR/DUE). The Editor also thanks the authors of the briefing papers for their time and effort in modifying their writing for this report, and the participants in the workshops for the many insights they contributed. Finally, the Editor offers special thanks to Helen Wright and Andrew Bernat at CRA; their leadership has been very important in this effort.

Insights for Education from Data-Intensive Research in the Sciences and Engineering

Elizabeth Burrows (AAAS S&T Policy fellow at NSF); Lida Beninson (AAAS S&T Policy fellow at NSF); Benjamin Cash (George Mason University); Doreen Ware (U.S. Department of Agriculture and Cold Spring Harbor Lab); Hsinchun Chen (University of Arizona); Ari Kaufman (State University of New York at Stony Brook); Kirk Borne (George Mason University); Lucy Fortson (University of Minnesota); Chris Dede (Harvard University)

INTRODUCTION

The first workshop featured NSF-sponsored exemplary projects from geological, engineering, biological, computational, and atmospheric sciences in order to categorize data-intensive research within these fields. Prior to delving into the content of the second workshop focused on education, it is fruitful to conduct a comparative analysis of the status of big data in the projects presented in the first workshop. Increases in the capacity of computing technologies, coupled with cost reductions in computing infrastructures, have enabled unprecedented efficiencies in scientific discovery through the collection, curation, analysis, and distributed interpretation of massive datasets. However, big data opportunities are developing in different ways across the fields that span the sciences and engineering; this variation is caused by the nature of the data studied within these disciplinary communities. The sections below: 1) provide an overview of data-intensive research via five case studies across the sciences and engineering, summarizing salient lessons learned to inform all disciplines; and 2) discuss big data and data-intensive research in the context of education research, providing reflections on the applicability to education of insights about data-intensive research in the sciences and engineering.

Before delving into a thorough discussion of data-intensive research and big data, it is useful to define the terms as we are using them in this report. Big data can be broadly defined as the intersection of computer science, statistics, and ethics or policy. The ethical issues are particularly pertinent to fields such as education research and healthcare. The first workshop largely used the term “big data” while the second workshop largely used the term “data-intensive research.” With full understanding that definitions for these terms are not universally accepted, are still developing, and are heavily contextual, we will use the following definitions for big data and data-intensive research:

> 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[5]. The first two paradigms of traditional scientific research refer to experimentation and theory, while the third encompasses computational modeling and simulation[6].
>
> Data science is the large-scale capture of data and the transformation of that data into insights and recommendations in support of decisions.

Although writing and talking about big data is popular, interdisciplinary discussions on this subject are challenging. One of the reasons for this might be that our mental model on the meaning of big data is informed by our disciplinary boundaries. In other words, the volume of data in one discipline may be the defining factor that deems it big data, while the velocity and variety of dealing with that big data may not be as challenging compared to other disciplines. Generally, the “life cycle” of any dataset involves acquisition, storage,

[5] “Hey T, Tansley S, and Tolle K. 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. Redmond, Washington.

[6] Strawn GO. 2012. Scientific research: How many paradigms? *EDUCAUSE Review*, 47(3).

distribution, and analysis, and big data challenges can arise in any or all of these components[7].

Often involved in describing big data is the collection of "V" words associated with it. We have seen at least 12 V-words used to define or describe big data, while the carefully chosen original three that appeared within the 2001 Gartner Report are (1) the increasing size of data (volume), (2) the increasing rate at which it is produced and analyzed (velocity), and (3) its increasing range of sources, formats, and representations (variety). We have added "veracity" to our discussion, encompassing widely differing qualities of data sources, with significant differences in the coverage, accuracy, and timeliness of data[8]. While it is true that veracity applies to every dataset, and that veracity is inherent in variety, we feel that big data presents unique veracity challenges. In addition to the veracity issues that are characteristic of all data collection and the propagation of error associated with combining datasets, there are issues related to reusing data for purposes different from those for which it was collected, recreating historic datasets, and relying on crowdsourcing.

In the following sections we will use these four Vs, not as any form of a definition of big data, but rather as a framework for comparison of the status of big data across disciplines.

STATUS OF BIG DATA IN CASE STUDIES FROM FIVE DISCIPLINES

The first workshop was designed to understand the current levels of exposure and needs in data storage, manipulation, and analysis in representative case studies from various disciplines, in order to uncover potentials for sharing lessons learned. Such a comparison can support discussions on interdisciplinary research and development in this domain, and can cross-fertilize further development, with the benefit of lessons learned informing all disciplines.

In order to choose model case studies, NSF program officers from different directorates were asked to identify exemplary big data projects. Nominees were then invited to attend the first workshop, where they introduced their projects and discussed challenges and opportunities (as described in this report's introduction). The discussion below is drawn from the presentations by these speakers (who were project principal investigators), supplemented by notes taken during the presentations and Q&A sessions, as well as related literature. Each of five case studies is split into two subsections: "Big Data Classification," using the Velocity-Variety-Veracity-Volume typology, and "Potential Shared Challenges and Opportunities With Education."

EARTH SCIENCES CASE STUDY

Climate Modeling and Big Data: Current Challenges and Prospects for the Future;. Benjamin Cash, Research Scientist, Center for Ocean-Land-Atmosphere Studies (COLA), George Mason University

Big Data Classification. Climate modeling pushes the current volume boundaries for both storage and computing power. From 2010-2011 Project Athena, and from 2012-2014 Project Minerva, examined the importance of horizontal resolution in seasonal predictions models. This research was conducted by large collaboration efforts among COLA at George Mason University, the European Center for Medium-Range Weather Forecasts (ECMWF), the University of Oxford, and Japan Agency for Marine-Earth Science and Technology (JAMSTEC). These projects used the Athena supercomputer at the National Institute for Computational Science at Oak Ridge National Laboratory and the Yellowstone supercomputer at the National Center for Atmospheric Research Wyoming Supercomputing Center (NWSC), respectively. At the time of these projects, the Athena supercomputer ranked at #30 in the world, and Yellowstone ranked at #13 on the Top500 list of supercomputing sites. Both projects consumed 10s of millions of core hours and generated more than 1 petabyte of data.

[7] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical? PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195.

[8] Dong XL and Srivasta D. 2013. "Big Data Integration", ICDE Conference, pg. 1245-1248.

One conclusion that showcases the importance of this work is the discovery that, when resolution increased from 125 km to 16 km, the predicted decrease in precipitation due to global warming in Southern Europe was vastly different (Figure 1). This indicates that further increases in resolution, which are extremely demanding from a big data and computational perspective, may be needed for accurate predictions.

Veracity is also a central issue for climate models, due not only to the high-stakes predictions they generate, but also to the vast number of different fluxes that need to be quantified among different pools in the system (Figure 2a). Each and every flux in the system often has great uncertainty, not only because a finite number of often labor-intensive measurements have to be used to get an estimate for all other parts of the globe currently, but also because climate models often strive to incorporate all collected historic data. A clear illustration of uncertainty in climate prediction is shown when comparing models of the same event; for example, weather for the coming months in the same region (Figure 2b). The results of the U.S. National Multi-Model Ensemble (NMME) suggested that combining the results of many models can offset biases within any one model, and creates probabilistically more reliable predictions[10].

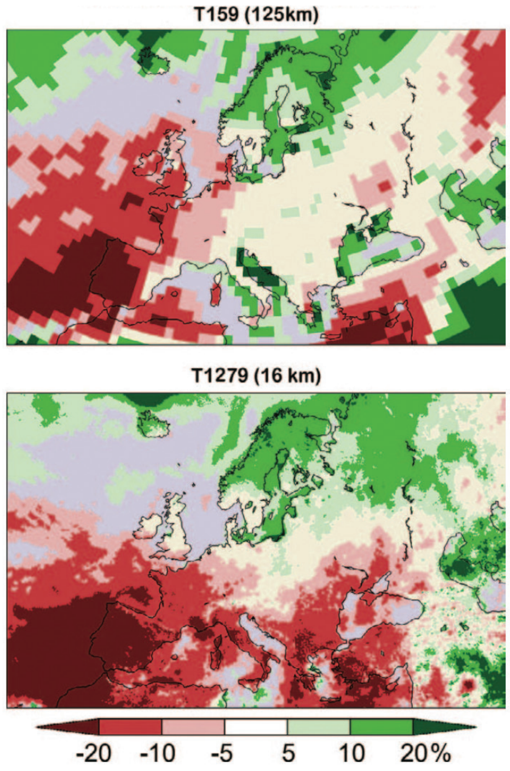


Figure 1. Growing season (April-October) precipitation change from the 20th to 21st centuries in Europe, as predicted by the ECMWF IFS global atmospheric model[9].

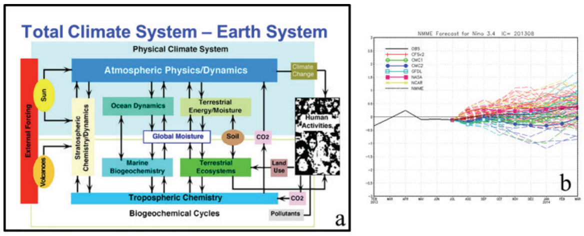


Figure 2. a) An example model of the climate system, from Earth System Science: An Overview, NASA, 1988; b) Predictions of change in temperature (ºC), starting in May 2013, for the Niño 3.4 Region, bounded by 120°W-170°W and 5°S- 5°N, figure from Kirtman et al. 2014 [11].

Potential Shared Challenges and Opportunities with Education. The challenge of determining optimal resolution is pertinent to education data, where the focus is generally temporal rather than spatial resolution. In education, the resolution chosen to analyze micro-learning behaviors across time has a large impact on the determination of macro-learning behaviors, just as geospatial resolution of precipitation data has a large impact on global atmospheric models. If it were determined that the highest resolution was needed to analyze the various forms of informal learning data such as games, social media, and makerspaces, then a volume issue would certainly arise, as it has with climate data. Another resolution challenge in education comes from determining the maximum resolution that still allows for anonymization of data.

[9] Kinter JL, Cash BA, et al. 2013. Revolutionizing Climate Modeling – Project Athena: A Multi-Institutional, International Collaboration. *BAMS*, 94, 231–245, doi: http://dx.doi.org/10.1175/BAMS-D-11-00043.1

[10] Kirtman BP, et al. 2014. The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bull. Amer. Meteor. Soc.*, 95, 585–601.

[11] Kirtman BP, et al.2014. The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bull. Amer. Meteor. Soc.*, 95, 585–601.

Climate models are designed to answer the high-stakes question of what the future climate will be, just as predictive models in education are designed to answer high-stakes questions such as what students will succeed. The PAR framework described in the next section (Predictive Models Based on Behavioral Patterns in Higher Education) incorporates 77 student variables in a model to predict student success. Some of the data needed to predict student success may come from data on home life, extracurricular activities, and social media presence, and data may need to be linked from tax returns, the census, and other datasets that are currently difficult to obtain and incorporate. There may be comparable methods for predictive modeling from different disciplines that all need to use variable, often incompatible datasets. Interoperability of disparate datasets is a central challenge for climate modeling (see Figure 2a), and it is also crucial for expanding predictive models in education, comparing MOOC data, and using data from multiple repositories.

BIOLOGICAL SCIENCES CASE STUDY

Research Challenges and Resource Needs in Cyberinfrastructure & Bioinformatics: BIG DATA in Plant Genomics; Diane Okamuro, NSF Program Officer for the Plant Genome Research Program AND Big Data: Challenges and Opportunities for Plant Sciences; Doreen Ware, Computational Biologist, U.S. Department of Agriculture, Agriculture Research Service and Adjunct Associate Professor at Cold Spring Harbor Lab

Big Data Classification. The National Plant Genome Initiative (NPGI) coordinates databases and tools such as Gramene, EnsemblPlants, Plant Ontology, and iPlant, which are all devoted to gaining a full understanding of plant systems, ranging from genomics and proteomics and metabolic models to phenotypes and large-scale production. While they are advancing capabilities in big data science with relation to all four Vs, variety is perhaps their greatest challenge. Particularly with NPGI's current five-year objectives of increasing open-source resources that span the data to knowledge to action continuum, their goal is to enable translation of all types of plant data (NPGI Five-Year Plan: 2014–2018[12]). Not only are the types of plant data varied in themselves, within each type of data, standards are still developing in order to ensure that all data are fully comparable. Examples include the Planteome Project that aims to establish a common semantic framework for sequenced plant genomes and phenotype data, the Plant Ontology Consortium that provides a glossary of anatomical and morphological plant components, and the Taxonomic Name Resolution Service (TNRS) that is a tool for automatically standardizing, updating, and correcting plant scientific names. NPGI has more than 16 partners in providing open access resources, including NSF's iPlant Collaborative, which in itself houses bioinformatics databases, high-performance computing platforms, and image storage and analysis capabilities, and has a data storage capacity of 427 TB. In addition, iPlant alone provides new registrations at a velocity of almost 500 per month. Data created through NPGI comes from industry, academia, government, and NGOs, and arrives in many different forms at varied but ever-increasing velocities.

Potential Shared Challenges and Opportunities With Education. Given the large variety of education data, there are several lessons that could be learned from the trajectory of the plant genomics field. Overall, the scientific community successfully formed, resolved differences in definitions, built shared infrastructure, pooled datasets, and conducted collective research on a grand challenge. Resources analogous to iPlant could be developed for education data and analysis. However, in any scientific field there needs to be a balance between question-driven and data-driven science; patterns identified using big data resources, such as genomes, could be used to further theory development in a cycle in which questions are raised, data are produced, and further questions are thus raised. The more tightly coupled this cycle is, the better. An education-specific version of this cycle was presented in the report for the "Advancing Technology-Enhanced Education" workshop

[12] NATIONAL PLANT GENOME INITIATIVE FIVE-YEAR PLAN: 2014 – 2018. https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/npgi_five-year_plan_5-2014.pdf

that was held in July 2013[13]. Once plant genomics data became readily available, there was a plethora of "low-hanging fruit" in terms of annotating each genome, attempting to find the function of each individual gene, and comparing the DNA base pairs of each genome to address questions in evolutionary biology. Conceptual and theoretical thinking about groups of genes, population analysis, and time-series analysis continues, in addition to the reductionist-level work. In education research it might be tempting to follow a similar path; for example, easily collecting large volumes of clickstream data for each student to infer learning trends, but the variety of data, rather than the volume, needs to be remembered.

HEALTH SCIENCES INFORMATICS CASE STUDY

783 Project in CISE/Health IT: From DiabeticLink to SilverLink; Hsinchun Chen, Ph.D., Regents' Professor, Thomas R. Brown Chair Professor, Director, Artificial Intelligence Lab, University of Arizona

Big Data Classification. Health science informatics has projects that exemplify both velocity and variety. Tracking of infectious diseases is an ideal "velocity" problem, since many kinds of real-time analytics are required. The BioPortal™ for Disease Surveillance program, developed at the University of Arizona, has been used to track SARS, West Nile virus, and Foot-and-mouth disease[15]. This program includes live tracking of new cases and hotspot analysis, as well as phylogenetic tree analysis to visualize how the organism causing the disease is changing (Figure 4).

The goal of the Smart and Connected Health program is to shift the focus of the healthcare industry from reactive cures to preemptive total health of patients. The types of data that are required to assess health range from sensors measuring physical characteristics of a person's current state, to detailed health and family histories, as well as past and present psychological information gained from sources such as social media presence. Sensors measuring someone as they are running could include blood pressure, pulmonary function, EEG, ECG, SpO_2, posture, gait, balance, step size

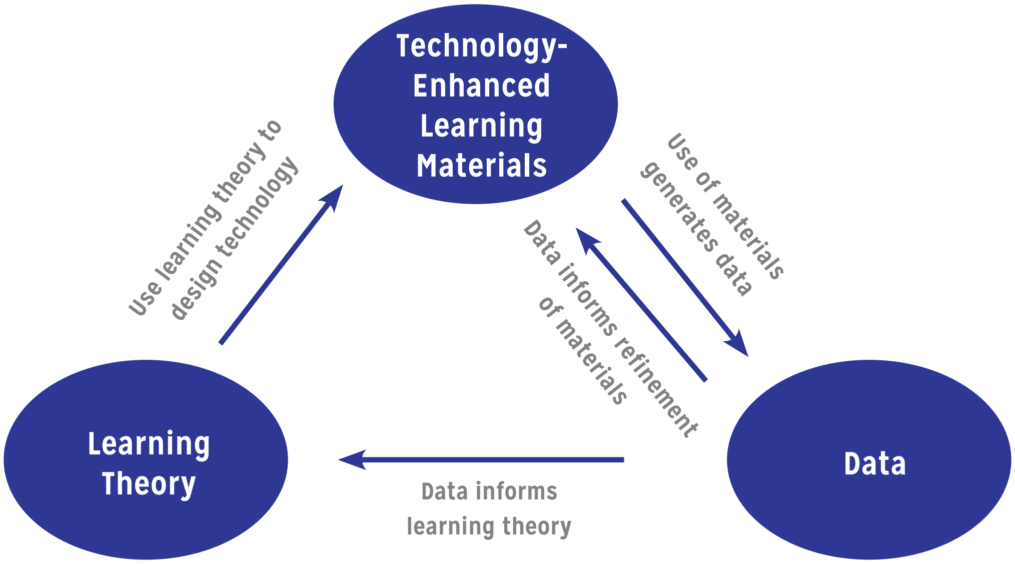

Technology-Enhanced Learning Materials
Use learning theory to design technology
Use of materials generates data
Data informs refinement of materials
Learning Theory
Data
Data informs learning theory

Figure 3. Learning Engineering Paradigm[14]

[13] Zuckerman BL, Azari AR, and Doane WEJ. 2013. Advancing technology-enhanced education: a workshop report. IDA Science & Technology Policy Institute. http://www.nsf.gov/ehr/Pubs/Advan_Tech_Ed_Wkshp.pdf

[14] Zuckerman BL, Azari AR, and Doane WEJ. 2013. Advancing technology-enhanced education: a workshop report. IDA Science & Technology Policy Institute. http://www.nsf.gov/ehr/Pubs/Advan_Tech_Ed_Wkshp.pdf.

[15] Chen H, Zeng D, and Yan P. 2010. Infectious Disease Informatics: Syndromic Surveillance for Public Health and Bio-Defense. Springer.

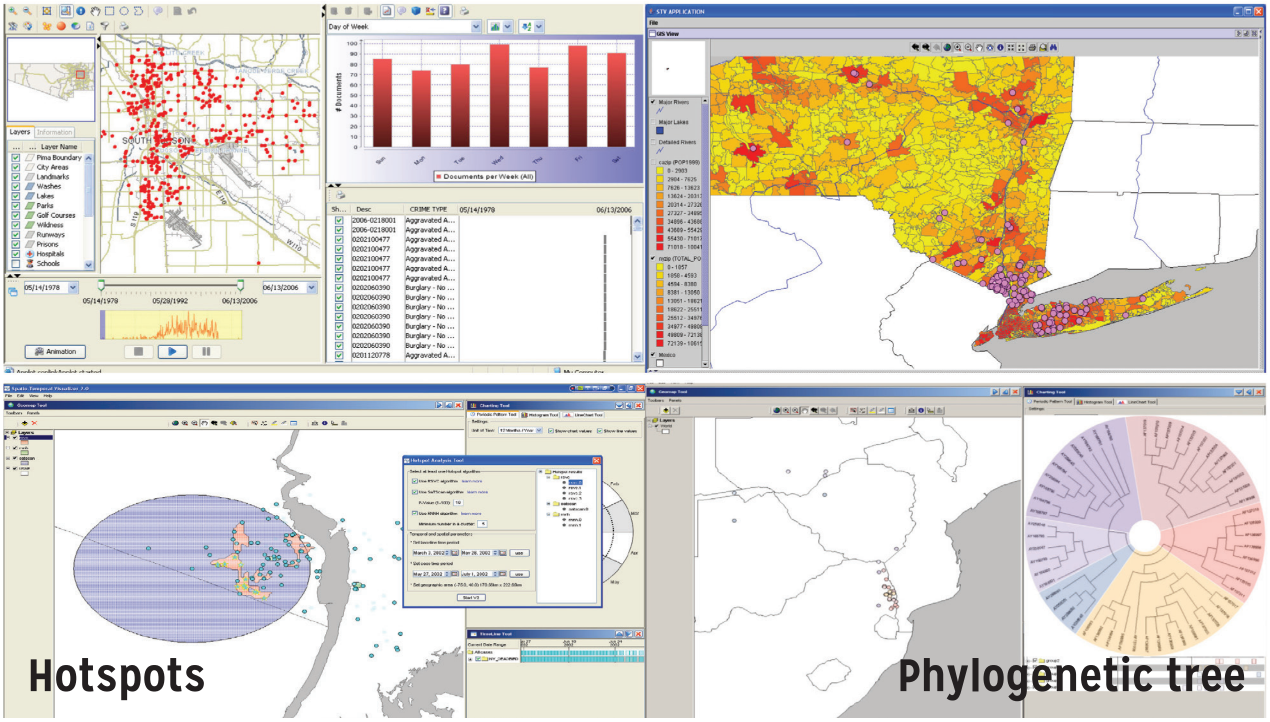


Figure 4. Screenshots from the BioPortal^{TM} for Disease Surveillance program

and height, and GPS; but this in itself is not necessarily "variety," because all of these sensors produce similar kinds of data. The variety could come from incorporating the runner's training regime, chronic care, mindset at the time of running, and more.

Potential Shared Challenges and Opportunities With Education. The two main commonalities between health informatics and education research are the privacy issues and the complexity of dealing with humans. With respect to privacy, conducting data analytics on patients' health records is similar to conducting data analytics on students' transcripts and related files, as well as behavioral and learning data contained in repositories. With respect to complexity, a person's overall health can only be assessed when data is incorporated from a variety of sources including health history, current diet and exercise, and mental state, as described above, just the same way a student's "academic health" can only be assessed when a large variety of factors are incorporated.

COMPUTER SCIENCES CASE STUDY: THE ROLE OF VISUALIZATION

Immersive Exploration of Big Data; Ari Kaufman, Distinguished Professor and Chairman of the Computer Science Department, Director of the Center of Visual Computing (CVC), and Chief Scientist of the Center of Excellence in Wireless and Information Technology (CEWIT) at the State University of New York at Stony Brook

Big Data Classification. Sophisticated high-resolution visualizations, such as those supported by the Reality Deck at Stony Brook University, could be used to bring new capabilities to several velocity problems. The Reality Deck is a 1.5 gigapixel, 4-walled room lined with screens[16]. It uses 240 CPU cores, with a speed of 2.3 TFLOPS and 1.2 TB of distributed memory. This chamber can be used for real-time analytics during disaster response or manhunt situations by capturing images of a cityscape at much higher resolution than is achievable with aerial photography (Figure 5).

[16] Papadopoulos C, Petkov K, Kaufman AE, and Mueller K. 2015. The Reality Deck – Immersive Gigapixel Display. IEEE Computer Graphics and Applications. 35(1), pp 33-45.

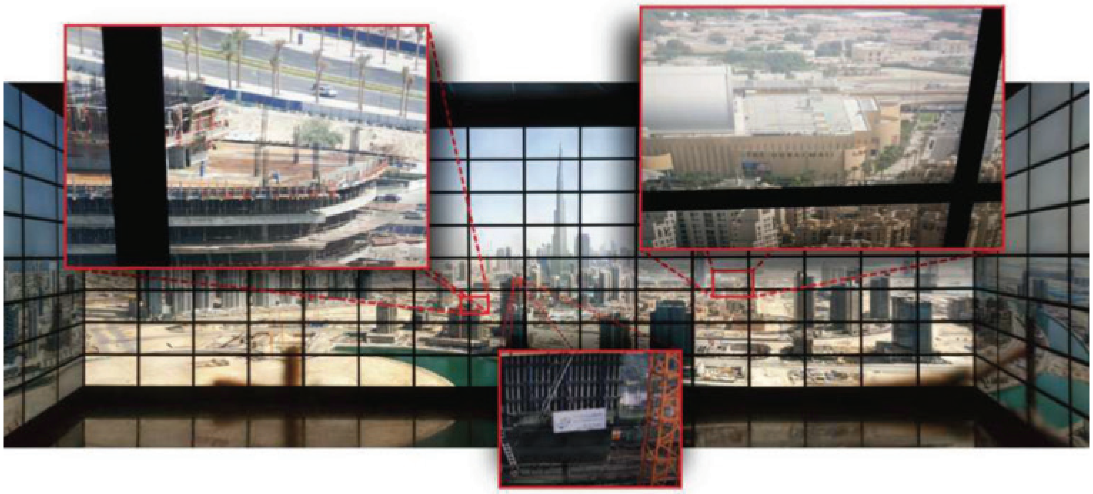


Figure 5. Examples of the level of resolution possible in the Reality Deck, using the Dubai skyline.

Potential Shared Challenges and Opportunities with Education. Issues central to the Reality Deck consist of access considerations (i.e., cost of the system might limit its wide adoption) and the prioritization of increasing engineering capability versus providing adequate information to answer specific research questions. The Reality Deck could be used for training purposes where the terrain or the physical environment is an important component of the learning (e.g., expected human behavior or chosen strategy is dependent on the physical environment). The real-time analytics capabilities of the Reality Deck could be used for several education applications such as intelligent tutoring systems or individualized guidance systems, immediate analysis of data-driven learning from social media platforms, and continuous and seamless student assessment followed by adaptation of instruction.

ASTRONOMICAL SCIENCES CASE STUDIES

Data Literacy For All: Astrophysics and Beyond; Kirk Borne, Professor of Astrophysics and Computational Science, George Mason University (now at Booz Allen Hamilton) AND Big Data Lessons from the ZOONIVERSE; Lucy Fortson, Associate Professor of Physics, University of Minnesota

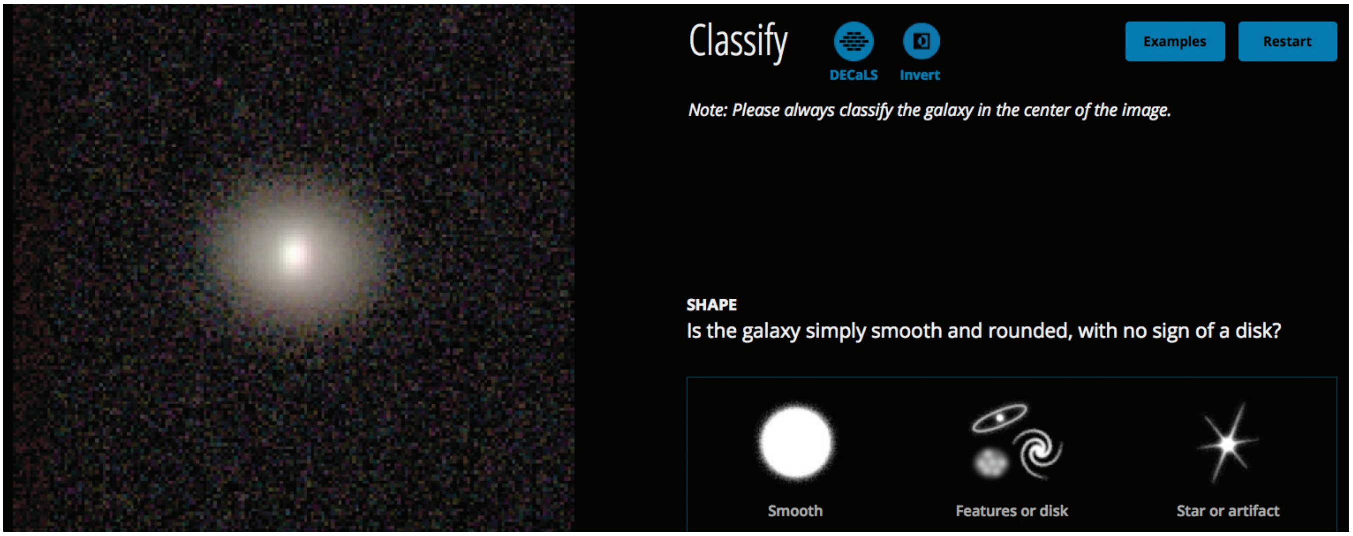
Big Data Classification. The Large Synoptic Survey Telescope, being built in Chile, is a key example of volume, velocity, and variety challenges. Construction began in August 2014, and it is projected to be fully operational in seven years, at which point a full 10-year survey will begin. The telescope will produce repeat imaging coverage of the entire night sky every three nights, producing one 10-square degree 6-GB image every 20 seconds. This will result in a 100–200 PB image archive, from which a searchable 20–40 PB database catalog will be produced; the entire database will be publicly available. While full data collection is several years out, there is intense focus on planning the data science aspects of the project, in addition to the focus on building the telescope. The volume, and velocity with which the data will need to be analyzed, will push the bounds of current capabilities. Real-time event mining will be crucial for identifying the velocity and trajectory of near-Earth objects, and for fully capturing dynamic time-varying phenomena such as exploding supernovae. With 10 million events per night every night for 10 years, event mining, to rapidly identify which events deserve more focus, will be crucial. The high-dimensional multi-PB database will present an enormous "big data variety" challenge also, as the source catalog alone will consist of more than 30 trillion rows (source measurements) with more than 200 columns of diverse attributes for each source observation[17].

Launched in 2007, the Galaxy Zoo project is an excellent example of the benefits of citizen science. Citizens were

[17] Ivezic Z, et al. 2014 (updated). LSST: from Science Drivers to Reference Design and Anticipated Data Products, arXiv:0805.2366.

encouraged to classify a collection of 1 million galaxies based on images from the Sloan Digital Sky Survey (Figure 6), and within 1.5 years, 35 million classifications had been submitted, by approximately 150,000 users. This may at first seem like a veracity issue, since citizens aren't as well trained as astronomers, but each galaxy was classified more than 35 times, and many data analytics techniques were applied to assess the quality of the data. Instead, the need for citizen science is indicative of a volume issue. While the volume may not be pushing data storage capabilities, it is pushing data analysis capabilities across a wide range of disciplines, and the Galaxy Zoo project led to the development of the Zooniverse platform for data-analysis citizen-science tasks, which now engages more than 1.3 million volunteers on 40-plus projects. This is an example of human computation applied to big data tasks[18]. Human computation is any computational process that takes advantage of the cognitive abilities of humans to solve tasks that are very difficult, or impossible, for computers alone, including complex image classification, complicated pattern detection, and unusual anomaly discovery within very large and diverse data collections (thus addressing the volume and variety challenges of big data).

Potential Shared Challenges and Opportunities with Education. In addition to the need for real-time analytics, as discussed in the previous case study, the astronomical sciences have the challenge of needing more humans to perform the first stages of data analysis, which leads to a need for including a greater focus on data science in curricula across education (see the "Summary of Insights" section). The citizen science described in the Galaxy Zoo example is similar to the need in the biological sciences to rely on undergraduate students to collect data, and there are parallels in education where educational scholars, practitioners, and policymakers could be trained to use data repositories, thus generating intermediate conclusions from data. One of the key contributions that citizen scientists, including educators and students, provide is a valuable complementary data set of characterizations of diverse objects and events in large data sets. These characterizations are simply features (e.g., shapes, colors, sizes, textures, anomalies, changes, and other components) that human cognitive abilities are innately tuned to recognize in any situation. This set of characterizations can then be added to any project's data collection as additional input to researchers' analyses, discoveries, and theory development. If educators and their students were able



Note: Please always classify the galaxy in the center of the image.

SHAPE

Is the galaxy simply smooth and rounded, with no sign of a disk?

Figure 6. Screenshot from galaxyzoo.org showing the active galaxy classification screen

[18] Law E, and Von Ahn L. 2009. Input-agreement: a new mechanism for collecting data using human computation games. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1197-1206). ACM.

to provide preliminary data characterization and analysis of large data sets, it would greatly accelerate the field. Citizen science provides one mechanism for enabling this acceleration while, at the same time, engaging the public, including students, in an authentic research experience enabling them to learn about the process of research. As part of the Galaxy Zoo project, a study was conducted on the motivations of the participants, and a common response was the desire to contribute to research[19]. This motivation suggests that citizen science would work in many fields if the questions are compelling and it is clear that human processing (i.e., human computation) is required.

SUMMARY OF LESSONS LEARNED FROM THE FIVE DISCIPLINES

The most common advice from presenters of the case studies highlighted in this chapter was to collaborate. Data-intensive research, even for one specific goal, requires interdisciplinary collaboration; and methods developed for data-intensive research in one field can often be adopted in *other* fields, thus saving time and resources, as well as advancing each field faster. Thus, while it is crucial for individuals to become domain experts, it is also highly beneficial to spend some time keeping aware of the state-of-the-art in other fields; hence, the reasoning behind the content of the two workshops summarized in this report.

Stressing the importance of standards and ontologies from the beginning is also critical. Even though it is tedious and takes time away from making immediate "progress," funding agencies and reviewers should understand that the long-term benefit is enormous. In addition to common language, shared infrastructure for data storage and data analysis to ensure interoperability of the data is key. Once the ability to share data is established, it is highly beneficial when companies have incentive to make their data available and collaborate with academics. In genomics, this incentive occurred when patent laws changed so that proof of gene function, and not simply gene sequence, is required for patents, which requires a much larger, often collaborative effort.

Echoing the importance of data-intensive work in the field of bioinformatics, but applicable to many fields, Howe et al. (2008)[20] direct attention to the need for structure, recognition, and support for biocuration—"the activity of organizing, representing, and making biological information accessible to both humans and computers." Further, they urge the scientific community to (1) facilitate the exchange of journal publications and the databases, (2) develop a recognition structure for community-based curation efforts, and (3) increase the visibility and support of scientific curation as a professional career. The importance of biocuration is evident in the urgency and complexity in researchers' need to locate, access, and integrate data. Howe et al. (2008) provide examples of such complexities. For example, papers sometimes report newly cloned genes without providing GenBank IDs, the human gene CDKN2A has 10 literature-based synonyms, etc. Indeed, efforts in interoperability and standards-based curation exemplified in the NSF investments in this field could be modeled by others.

BIG DATA AND DATA-INTENSIVE RESEARCH IN THE CONTEXT OF EDUCATION RESEARCH

How does the field of education compare to the sciences and engineering in the use of data-intensive 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

[19] Raddick MJ, Bracey G, Gay PL, Lintott CJ, Murray P, Schawinski K, Szalay AS, and Vandenberg J. 2010. Galaxy zoo: Exploring the motivations of citizen science volunteers. Astronomy Education Review: 9(1).

[20] Howe D, Rhee SY, et al. 2008. "Big Data: The Future of Biocuration", Nature, Vol.455/4, pg. 47-50.

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,” 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[21].” This report also renewed its support for the Advanced Research Projects Agency-Education (ARPA-ED), which was formed to fund projects with the potential to create a dynamic breakthrough in learning and teaching (*Winning the Education Future: The Role of ARPA-ED*; Department of Education, 2011[22]). ARPA-ED aims to catalyze the use of massive data to guide traditional classroom instruction, rethink curricula, and increase the pace of “learning about learning.”

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[23]. 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.

REFLECTIONS ON INSIGHTS FOR EDUCATION FROM THE LESSONS LEARNED IN THE FIRST WORKSHOP

The structure of the first workshop proved very productive in revealing the differential nature of data-intensive research challenges.

The first of the two earth science presentations focused on big data in open topography, which is a mature area in which data is easily and unobtrusively obtained via light detection-and-ranging measurements from laser sensors. A large user community draws on this data, which is collected, transformed, optimized, and organized in a central repository. The development of tools for analyzing this data is an important part of the cyberinfrastructure. Exponential growth in data and rapidly evolving scientific findings are emerging challenges in this field, but at present there are no major issues. *Models from this type of data-intensive research may be of value for comparable types of big data in education, such as student behavior data in higher education and the growing use of predictive model to derive insights from this for issues such as student retention. Another parallel in education is multi-modal data about student learning behaviors such as that available from sensors, video gesture recognition, and logfiles.*

Also in the earth sciences, but facing much more immediate challenges, is big data in climate modeling. The amount of data now available is pushing both computational and storage capabilities to their limits, and the important next step of improving the fidelity of climate models will necessitate a million fold increase in computing capability, with comparable impacts on data storage, transfer, and other parts of cyberinfrastructure. *Models from this type of data-intensive research may be of value for comparable types of big data in education, such as the massive amounts of learning data that could be collected outside of formal educational settings via games,*

[21] 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.

[22] Department of Education. 2011. *Winning the Education Future: The Role of ARPA-ED*, http://www.ed.gov/sites/default/files/arpa-ed-background.pdf.

[23] 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.

social media, and informal learning activities such as makerspaces.

In biology, data-intensive research in plant genomics required a multi-decade series of five-year plans, developed and actualized across the entire scholarly community in this field. These coordinated activities focused on translating basic knowledge into a comprehensive understanding of plant performance, studying the effects of local climate variations, and accelerating the field's processes of discovery. The evolution of systems and data interoperability and standards was crucial to success, and substantial cyberinfrastructure challenges remain in data aggregation, computational power, and analytic methods. *Models from this type of data-intensive research may be of value for comparable types of big data in education, such as the collection of data to inform student performance from widely varying sources such as home life, extracurricular activities, social media presence, financial situation, and health information, in addition to standard education data.*

In health informatics, data-intensive research requires collecting and integrating data from a wide variety of sources, posing considerable challenges of interoperability and standardization. Further, unlike the types of scientific data discussed thus far, issues of privacy and security are paramount in medicine and wellness, greatly complicating the processes of collection, storage, and analysis. *Models from this type of data-intensive research may be of value for comparable challenges of big data in education, such the development and management of repositories containing all types of behavioral and learning data from MOOCs, intelligent tutoring systems, and digital teaching platforms.*

Both engineering and astronomy confront challenges of needing more human capacity in data sciences to cope with the amount of data being collected and stored. In engineering, the development of centers that specialize in access to big data, the creation of specialized analytical tools, and the use of visualization are aiding with many of these problems. In astronomy, the recruitment, training, and usage of citizen scientists to aid in data analysis is essential to advancing the field, given the enormous and growing amounts of data being collected. *Models from these types of data-intensive research may be of value for comparable challenges of big data in education, such the involvement of educational scholars, practitioners, and policymakers in understanding and utilizing findings from the data repositories listed above.*

Developing new types of analytic methods tailored to the unique characteristics of big data is an important, cross-cutting issue across all fields of research. In the sciences and engineering, new approaches to statistical inference are developing, and machine learning is making advances on handling types of information outside the kinds of quantitative data for which statistical methods are appropriate. *Advances in these and other types of analytics may be of value for comparable challenges of big data in education.*

Overall, these insights from the first workshop illustrate emphases, issues, and structures for the subsequent workshop on data-intensive research in education.

Predictive Models Based on Behavioral Patterns in Higher Education

Ellen Wagner (PAR Framework) and David Yaskin (Hobsons)

WAGNER: THE EDUCATIONAL CHALLENGE

Although 90 percent of students enter college with the intention of completing a degree or certificate[24], only 59 percent of full-time students earn their bachelor's degrees within six years, and only 31 percent of community college students earn their degrees or certificates within 150 percent of the time allotted to do so[25]. Despite much investment and myriad solutions for improving student success, postsecondary education completion rates have generally remained unchanged for the past 40 years. Of all students who enroll in postsecondary education, less than half (46.1 percent) completes a degree within 150 percent of "normal time" to degree[26]. While online learning offers a legitimate path for pursuing a college education and provides students with a convenient alternative to face-to-face instruction, it, too, is laden with retention-related concerns[27], with even lower rates of completion than their on-the-ground counterpart courses and programs.

Thus, it is not surprising that higher education institutions are being pressured, either by regulation or law, to submit "student success[28]" data to state, regional, or federal agencies in order to receive funding[29]. Currently, 34 states are either using or in the process of implementing performance-based funding[30].

Habley and Randy located more than 80 programs and practices that institutions have implemented to help students, including supplemental learning, academic advising, tutoring, and first-year experience programs[31]. Even so, student completion rates have not significantly changed, leading Tinto and Pusser to suggest that higher education institutions must shift their attention from simply responding to students' attributes to evaluating how institutional policies and structures affect student success[32].

THE POTENTIAL VALUE OF DATA-INTENSIVE RESEARCH

Interest in exploring the potential value of data-intensive research in education comes at the same time that the pressures on education to show greater returns on educational investments and improved outcomes from students makes a very strong case for optimization. This means both augmenting and expanding beyond descriptive reporting and clusters of small *n*

[24] Ruffalo Noel-Levitz. 2013. *2013 national freshman attitudes report.* Retrieved from https://www.noellevitz.com/documents/shared/Papers_and_Research/2013/2013_National_Freshman_Attitudes.pdf.

[25] National Center for Education Statistics. 2014, May. *The condition of education*. Retrieved from http://nces.ed.gov/programs/coe/indicator_cva.asp.

[26] Knapp LG, Kelly-Reid JE and Ginder SA. 2012. "Enrollment in Postsecondary Institutions, Fall 2010; Financial Statistics, Fiscal Year 2010; and "Graduation Rates, Selected Cohorts, 2002–2007," NCES 2012-280 (Washington, D.C.: National Center for Education Statistics, 2012); U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS), Spring 2009, Graduation Rates component (Table 33).

[27] Wagner ED, and Davis BB. 2013. The Predictive Analytics Reporting (PAR) Framework, WCET *(December 6, 2013)* http://www.educause.edu/ero/article/predictive-analytics-reporting-par-framework-wcet.

[28] Kuh, Kinzie, Buckley, Bridges, and Hayek. (2006) define student success as persistence, satisfaction, academic achievement, education and skills/knowledge/competency attainment, education engagement, and performance post-college.

[29] Hayes R. 2014, October. Digital engagement: Driving student success. *EDUCAUSE Review Online*. Retrieved from http://www.educause.edu/ero/article/digital-engagement-driving-student-success.

[30] National Conference of State Legislatures. (2015, January 13). *Performance-based funding for higher education*. Retrieved from http://www.ncsl.org/research/education/performance-funding.aspx.

[31] Habley W, and McClanahan, R. 2004. *What works in student retention?* Iowa City, IA: ACT, Inc.

[32] Tinto V, and Pusser, B. 2006, June. *Moving From theory to action: Building a model of institutional action for student success.* Washington, DC: National Postsecondary Education Cooperative. Retrieved from https://nces.ed.gov/npec/pdf/Tinto_Pusser_Report.pdf.

experimental and quasi-experimental designs. Instead, the great opportunity for data-intensive research is to help educational stakeholders make better decisions, obtain deeper and better insights, and to find new patterns that can help provide new understandings about learning and cognition through predictive and prescriptive analytics (Figure 7), actively seeking out risk and actively mitigating risks through targeted treatments, interventions, and supports.

From Hindsight to Foresight

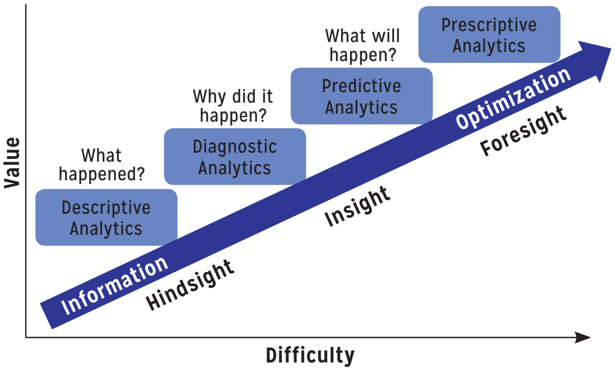


Figure 7. From Descriptive Analytics to Predictive Analytics

USING STUDENT-CENTERED DATA VERSUS SYSTEM-CENTERED DATA

Institutions collect vast amounts of data about their students, and significant benefit can be received from an enterprise success platform that makes better use of the data that institutions already have. At most institutions, student data lives in disparate data stores, such as the student information system (SIS) and the learning management system (LMS). Data is also being captured through tutoring logs, attendance records, and student self-assessments, among many other sources. As Hayes (2014) argues, this "system-centered approach makes it difficult to uncover the relationships among the data that, taken together, provide critical insight into the plans, progress, and needs of individual students."

Thus, Hayes recommends a student-centered approach for institutional data use. By focusing on the students instead of the systems that generated the data, stakeholders see a comprehensive view of an individual student's experience. According to Conrad and colleagues, a student-centered approach facilitates greater understanding of the relations between students, their college experiences, and outcomes[33]. To make data collection and integration straightforward, an enterprise success platform can be utilized. Such a platform will leverage the time and money that the institution has invested to implement and maintain its existing technology systems.

CREATING A DIGITAL ENGAGEMENT STRATEGY FOR STUDENT SUCCESS

Hayes argues that digital engagement—defined as the "use of technology and channels to find and mobilize a community around an issue and take action[34]"—is the "logical extension" of Tinto and Pusser's (2006) suggestion. Hayes (2014) continues to state that "[i]n this context, digital engagement involves using data and online tools to inform and motivate the entire campus community in order to underscore its student success efforts and drive change in completion outcomes." We also argue that a digital engagement strategy for student success will improve student outcomes by giving institutions greater insight into how students are performing and how the institution is responding.

To facilitate a digital engagement strategy, institutions can: (1) leverage an enterprise success platform (e.g., the Starfish® platform) to analyze student performance data using predictive analytics, (2) solicit feedback from key members of a student's success network, (3) deliver information to the right people who can help the student, and (4) collectively keep track of their efforts along the way—all of which leads to a continuous data-informed process-improvement cycle.

[33] Conrad, C., Gaston, M., Lundberg, T., Commodore, F., & Samayoa, A. C. 2013. *Using educational data to increase learning, retention, and degree attainment at minority serving institutions.* Philadelphia, PA. Retrieved from http://www.gse.upenn.edu/pdf/cmsi/using_educational_data.pdf

[34] Visser, J., & Richardson, J. 2013. *Digital engagement in culture, heritage, and the arts.* Retrieved from http://www.slideshare.net/MuseumNext/digital-engagement-in-culture-heritage-and-the-arts

EXAMPLES OF STATE-OF-THE-ART WORK

The Predictive Analytics Framework (PAR) PAR was born when members of the Western Interstate Commission for Higher Education's Cooperative for Educational Technology (WCET) proposed using predictive analytics to address the ongoing problem of student loss in U.S. post-secondary education. As described by Ice et. al.[35], PAR worked with six forward-thinking post-secondary institutional partners who contributed student and course data into one dataset, and a managing partner that built predictive models, managed the data, and managed all project operations. These collaborators determined factors contributing to retention, progression, and completion of online learners with specific purposes of (1) reaching consensus on a common set of variables that inform student retention, progression, and completion; and (2) exploring advantages and/or disadvantages of particular statistical and methodological approaches to assessing factors related to retention, progression and completion.

Using the results of this initial study as evidence, the PAR team continued to develop predictive modeling and descriptive benchmarking, adding an additional 16 colleges and universities to the collaborative and an additional 44 variables in the dataset in the following three years. From this data, PAR continued to develop and refine institutional predictive models for finding students at risk, national benchmarks showing comparative outcomes data, and an intervention insight platform for inventorying, tracking, measuring, and managing interventions.

PAR differentiated itself from other analytics providers in the post-secondary educational ecosystem by actively leveraging its common and openly published student success data definitions. PAR then further differentiated itself by connecting predictions of risk to solutions that mitigate risk as measured by improved student retention. PAR predictions of student risk are linked to information about interventions shown to work with specific risks with specific students at specific points in the college completion life-cycle. For example, Bloemer et al.[36], note that predictions of students at risk are of greater value when tied to interventions that have been empirically shown to mitigate risks for "students like them" at a specific point of need.

PAR Framework Current Status: PAR currently holds more than 2.6 million anonymized student records and 24 million institutionally de-identified course level records, working with more than 350 unique member campuses. PAR provides actionable institutional-specific insight to member institutions from 2-year, 4-year, public, proprietary, traditional, and progressive institutions. Participating institutions, each one committed to student success, actively engage in the collaborative by voluntarily sharing their assets and experience and benefitting from the member insight tools and exchange of best practices, all in the service of measurably improving student outcomes.

How PAR Works: The PAR Framework identifies factors that affect success and loss among all undergraduate students, with a focus on at-risk, first-time, new, and nontraditional students. While attention had initially been paid only to online students, the sample now includes records of all students from on-the-ground, blended, and online programs attending partner institutions. PAR focuses on 77 student variables that are available for each student in the massive data set. Viewing normalized data through a multi-institutional lens and using complete sets of undergraduate data based on a common set of measures, with definitions, provides insights that are not available when looking at records from a single institution (Figure 8).

PAR works with institutional partners to gather data according to the PAR Framework common data

[35] Ice, P., Diaz, S., Swan, K., Burgess, M., Sherrill, J., Huston, D., Okimoto, H. 2012. The PAR Framework Proof of Concept: Initial Findings From A Multi-Institutional Analysis Of Federated Postsecondary Data Vol 16, n.3 (2012) http://olj.onlinelearningconsortium.org/index.php/jaln/article/view/277

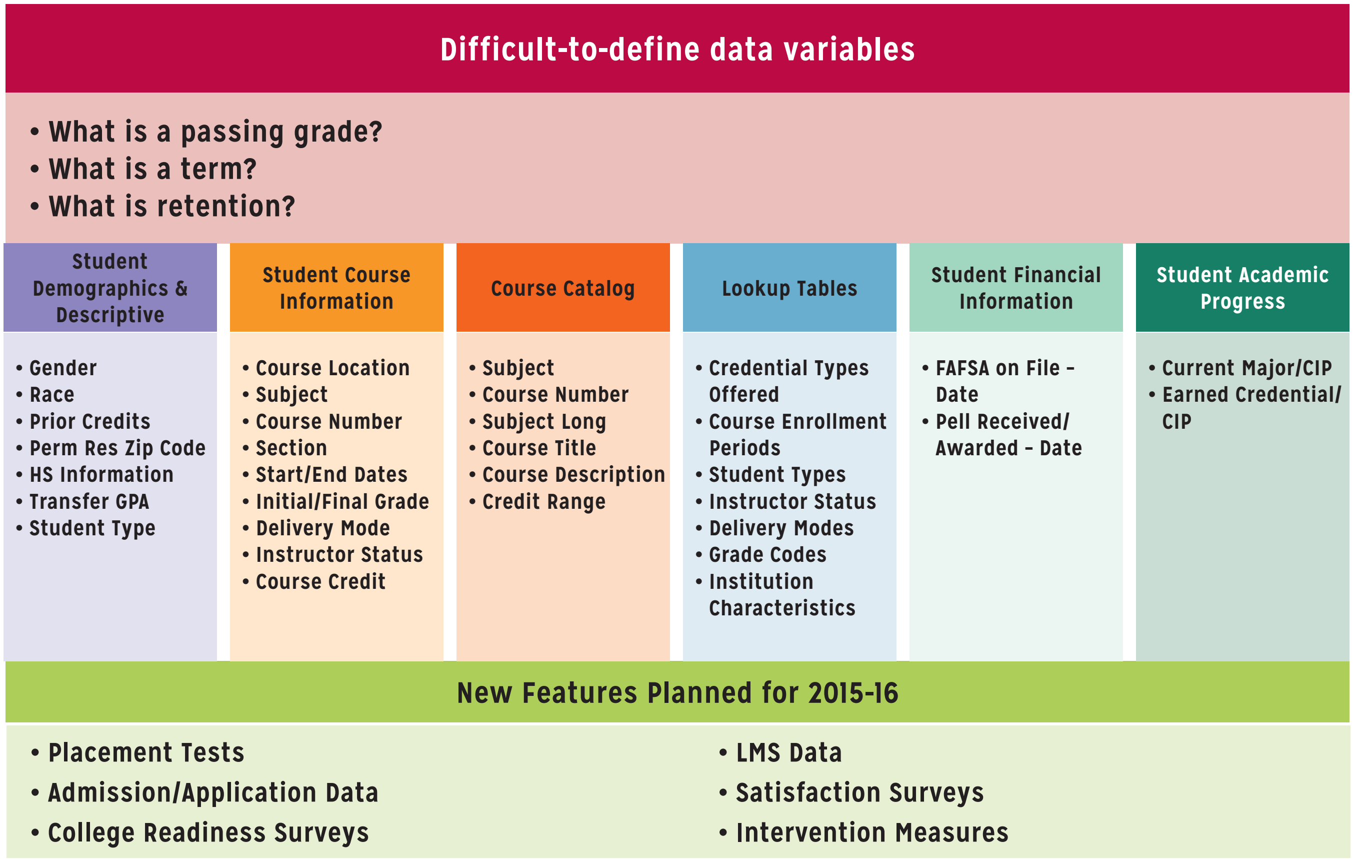
[36] Bloemer B, Swan K, Cook V., Wagner ED., Davis B. 2014. The Predictive Analytics Reporting Framework: Mitigating Academic Risk Through Predictive Modeling, Benchmarking, and Intervention Tracking, Illinois Education Research Conference, Bloomington IL, Oct 7, 2014.
Breiman, L. 2001. Random forests. *Machine Learning*, 45, 5-32.

definitions and a detailed file specification. As a last step before data submission, institutions remove any personally identifiable data, including date of birth, Social Security number, and local student ID number, and replace those items with a PAR student ID. Institutions maintain a translation table of their internal student ID to PAR Student ID, which is used to easily re-identify those students after PAR analyzes the data. PAR puts the data through more than 600 quality assurance tests as it is prepared for inclusion in PAR's Amazon Web Services-hosted data warehouse. Data are then analyzed to develop institutional-, program-, course- and student-level descriptive analytics and predictive insights contained in predictive analytic dashboards and in national benchmark reports built using SAS Visual Analytics, a choice made thanks to unlimited institutional visualization software licenses at PAR partner institutions. PAR members provide incremental data updates at the end of each term or course enrollment period to measure changes over time, evaluate the impact of student success interventions, and enable the PAR predictive models to be adjusted and tuned for current data.

PAR's framework for gathering student-level data based on common definitions helps member institutions:

- Understand their local data issues and challenges
- Develop capacity for reaching across systems and silos to create meaningful longitudinal student-level record sets
- Organize data across campuses consistently using common definitions and data types, making campus-level comparisons possible

Common Data Elements



Difficult-to-define data variables

- What is a passing grade?
- What is a term?
- What is retention?

Student Demographics & Descriptive	Student Course Information	Course Catalog	Lookup Tables	Student Financial Information	Student Academic Progress
• Gender • Race • Prior Credits • Perm Res Zip Code • HS Information • Transfer GPA • Student Type	• Course Location • Subject • Course Number • Section • Start/End Dates • Initial/Final Grade • Delivery Mode • Instructor Status • Course Credit	• Subject • Course Number • Subject Long • Course Title • Course Description • Credit Range	• Credential Types Offered • Course Enrollment Periods • Student Types • Instructor Status • Delivery Modes • Grade Codes • Institution Characteristics	• FAFSA on File - Date • Pell Received/ Awarded - Date	• Current Major/CIP • Earned Credential/ CIP

New Features Planned for 2015-16

- Placement Tests
- Admission/Application Data
- College Readiness Surveys
- LMS Data
- Satisfaction Surveys
- Intervention Measures

Figure 8. PAR Common Data Elements

- Uncover gaps, errors, and overlaps in student data elements across institutional systems

- Isolate and remedy anomalies in student cohort reporting generated by student exception handling

- Improve the capture and reporting of student military and veteran statuses across the multiple systems where that data is recorded

Linking Predictions to Action:- The PAR Framework Student Success Matrix

Most institutions have more than 100 student success services in effect at any one time. PAR Framework's Student Success Matrix (SSMx) application uses a validated mechanism to inventory, track, manage, and measure those student success activities in use across the institution. PAR SSMx gives users the tools to capture, measure, and compare ROI at the individual intervention level (Figure 9).

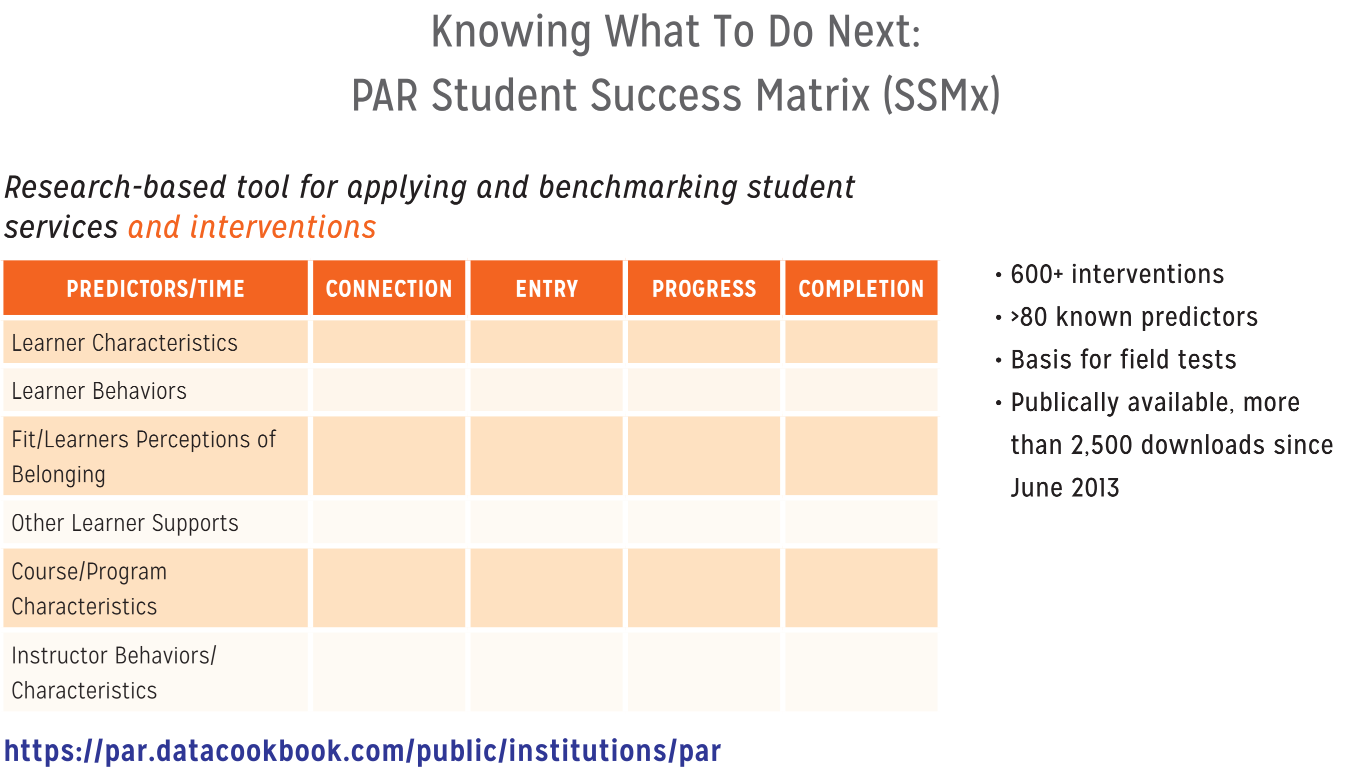
The PAR SSMx helps institutional members:

- Eliminate duplicate or redundant programs. Most campuses find that at least 10% and as many as 30% of intervention programs are serving the same audience and the same goal.

- Understand the scale of their student success programs. Many student success initiatives are upside down in terms of the institutional resources attached to the program, relative to the students served. The SSMx helps institutions right-size their investments to the student need and potential impact on retention and graduation.

- Match interventions with causes of student academic risk. Together with PAR predictive models that identify which students are at-risk and why, the SSMx identifies which key risk factors lack any success program counterparts. For example, while low GPA and student withdrawals often contribute to student risk for course success and retention, many campuses lack initiatives that flag for and address those behaviors.

- Measure the impact of student success programs.



Knowing What To Do Next:
PAR Student Success Matrix (SSMx)

Research-based tool for applying and benchmarking student services and interventions

PREDICTORS/TIME	CONNECTION	ENTRY	PROGRESS	COMPLETION
Learner Characteristics				
Learner Behaviors				
Fit/Learners Perceptions of Belonging				
Other Learner Supports				
Course/Program Characteristics				
Instructor Behaviors/ Characteristics				

- 600+ interventions
- >80 known predictors
- Basis for field tests
- Publically available, more than 2,500 downloads since June 2013

https://par.datacookbook.com/public/institutions/par

Figure 9: The PAR Student Success Matrix

Even among the most data-driven institutions, only about 10% of the many intervention programs are properly evaluated for effectiveness while millions are invested campuswide with a limited understanding of returns. The PAR SSMx enables institutions to measure their investments and the number of students reached for every intervention. More importantly, PAR analysis statistically measures intervention effectiveness, enabling a ROI comparison of impact to students at the intervention level.

- Respond to budget cuts with informed decisions. Aided by a comprehensive understanding of programs and their impact, institutions can make knowledgeable decisions on how to eliminate waste and redundancy during times of budget contraction without worrying they are cutting the wrong programs.

Yaskin: The Starfish Enterprise Success Platform

Starfish Retention Solutions, which became part of Hobsons in early 2015, has provided an enterprise student success system to more than 250 colleges and universities and supported more than 11,880,506 students, since being founded in 2007. The Starfish platform supports the philosophy that a student who is engaged with a connected, informed academic community will be more successful. The Starfish platform brings together data from people (e.g., faculty, advisors, residence hall staff, tutors, and counselors) and systems (SIS, LMS, advising systems, tutoring systems, and student surveys) at an institution. Although the value of a digital engagement strategy may be obvious for large public universities or open-access community colleges, even small colleges have saved millions of dollars in student revenue through the use of the Starfish platform[37].

In 2014, the Starfish platform processed more than 171 million automated and user-initiated flags representing actionable concerns about students. That same year, it facilitated more than 2.5 million meetings between students and staff members, which were engaged to help the student in their studies or in their lives. Similarly, more than 1.7 million notes, referrals, and plans were documented, creating personalized pathways for students to achieve their goals.

Having so much data about so many students can be challenging. Which student do you help first, second, and third? Predicative analytics shows promise to help prioritize who to assist. Starfish began creating its predictive analytics solution in 2014. During development, Starfish worked with two existing clients to build models from their data and provide success scores to indicate risk levels among their students. At the White House College Opportunity Day of Action in December 2014, Starfish committed to offer complimentary predictive analytics services to Davidson County Community College, Northeast Wisconsin Technical College, and Morgan State University.

Starfish's first predictive model was designed to answer the question, "Which students are most at risk of leaving the institution before the next term without completing their degrees?" The model produces a success probability for each student, where success means continuing at the institution in a future term. Registered students are scored against the model once per term, early in the term, and these predictive scores are available to advisors in the Starfish platform. Each student gets a score (e.g., 80 percent chance of continuing).

Starfish employs both machine-learning techniques and random forest models, a type of nonlinear, nonparametric regression model known for its versatility, performance, and ability to scale to large amounts of data[38]. Because these models are nonlinear, they find patterns such as discontinuities, threshold effects, break points in predictor variables, and interaction effects. These effects cannot be discovered automatically by generalized linear models (GLMs) such as linear regression or logistic regression.

The predictive model is built from data in the Starfish database, which includes data from the institution's

[37] Taylor Land McAleese V. 2012, July. Beyond retention: Using targeted analytics to improve student success. *EDUCAUSE Review Online*. Retrieved from http://www.educause.edu/ero/article/beyond-retention-using-targeted-analytics-improve-student-success.

[38] Breiman L. 2001. Random forests. *Machine Learning, 45*, 5-32.

SIS, the LMS, and the Starfish application itself. For new clients who do not have historical Starfish data, an initial model is constructed from a one-time upload of historical SIS data. Data available from the SIS includes admissions data, term and cumulative GPAs, credit hours attempted, term and cumulative credit hours earned, term and cumulative credit hours attempted but not completed, age, gender, ethnicity, program, time in program, financial aid and tuition data, and term GPA relative to past performance. Some of the strongest predictors come from the SIS data.

Once students have scores, the Starfish platform provides a variety of options for follow-up. For example, students may be flagged based on their predictive scores. The Starfish platform tracks these flags and records follow-up actions taken. The Starfish platform can define cohorts that represent students with predictive scores in a certain range, and follow-up for students in these cohorts can be managed as a group.

As the Starfish platform is used to advise and monitor students, it generates additional behavioral data that can define or refine future models, such as appointment types, reasons for making appointments (e.g., tutoring or advisement), topics discussed in meetings (as documented with its SpeedNotes), instructor-raised flags for attendance or other concerns, and system-raised flags (e.g., low assignment grades in LMS). Some of this behavioral data has been incorporated into Starfish predictive analytics models.

Behavioral metrics are difficult to standardize and interpret when moving from the context of one institution to another. Our models, therefore, do not use behavioral data from one institution to build models for use at a different institution. Just because making appointments of type X is predictive of persistence at one institution, the Starfish models do not assume that appointments of type X will necessarily have predictive value at another institution.

In addition to providing predictive scores, the Starfish platform provides more visibility into the reasons that certain students received certain scores. Having the ability to cluster or group students who received low scores for similar reasons can help guide different intervention strategies for different groups. For example, one identified group might be “non-traditional students (part-time adult learners) who are experiencing below-average progress toward completion.” These students might need a different type of intervention than, for example, traditional students who have received an academic warning.

LESSONS LEARNED ABOUT IMPLEMENTATION

- Scale requires reliable, generalizable outcomes and measures that can be replicated in a variety of settings with a minimal amount of customization. In the case of PAR, common definitions and look-up tables served as a Rosetta Stone of student success data, making it possible for projects to talk to one another between and within projects.
- Common data definitions are a game changer for scalable, generalizable, and repeatable learner analytics.
- Predictions are of greater institutional value when tied to treatments and interventions for improvement, and to intervention measurement to make sure that results are being delivered.
- Change happens when fueled by collaboration, transparency, and trust.
- Data needs to matter to everyone on campus. While data professionals will be needed to help construct new modeling techniques, ALL members of the higher education community are going to need to “up their game” when it comes to being fluent with data-driven decision-making, everyone from advisors to faculty to administrative staff to students.
- Using commercial software stacks already in place on campuses and data exchanges to extend interoperability with other Integrated Planning and Advising Systems (IPAS) systems extends the value and utility of technology investments.
- It takes all of us working together toward the same goal in our own unique ways to make the difference.

ISSUES IN MOVING FORWARD

Even with a student-centered approach in place, there are still some issues that need to be addressed when using an enterprise success platform. These include:

1. **Student Data Permissions.** Inadequate attention to who requires access to student data can expose inappropriate student information to staff. Thus, robust permissions schemas must exist that allow permissions to be tailored to the campus, college, and departmental levels per their policies.
2. **Data Overload.** Access to too much data can overwhelm staff. While it is useful to gain access to rich data for a single student, it can be difficult for staff to determine how to prioritize their time with students based on this data.
3. **When to Act?** Software applications and predictive analytics are needed to triage mountains of student data into actionable to-do items for staff members. The staff needs to know when they should act. Should it be as soon as a student misses a class? Or when a student receives a mid-term grade below a D? By analyzing historical data, a predictive model can be created to determine which characteristics and behaviors require the most urgent action.

The answer to these three issues requires the use of predictive analytics based on historical data, instead of snapshot reports of student data at any single point in time.

OTHER ISSUES INCLUDE:

- Helping first-year students discover which field is a good fit for their strengths in order to deepen engagement and motivation
- Getting students with problems to a staff member who actually figures them out; the symptom (e.g., poor initial grades) can have many root causes: academic, social, or financial
- Making the outputs from these systems more helpful for their users through better visualizations and imbedded workflows
- Connecting back to instructional methods

FUTURE OPPORTUNITIES

Technology for student success is still in its infancy. Based on vendor-reported market data, less than 25 percent of colleges have adopted third-party technology to support student success. Less than 10 percent have central systems for all undergraduates. Therefore, there are many ways to improve the technology and its use. A few exciting areas of future possibility include:

- Using outcomes data relating to employment after graduation to build better models
- Using activity and performance data from pre-college institutions to enhance predictions early in college
- Cross-institutional data sharing could lead to generalizable insights

SUMMARY

Arguably, there is no more important issue to engage the campus community in than student success. As Hayes (2014) mentions, switching to a student-centered approach for improving student outcomes will require a paradigm shift[39].

Hayes (2014) also argues that the right enterprise success platform offers tools "to identify at-risk students, offer academic advising and planning, and facilitate connections to campus support" using this student-centered data. Taylor and McAleese (2012) found that such an approach can contribute to significant gains in grades, persistence, and graduation rates. Such capabilities can also affect student success, support student needs, and promote student persistence outcomes[40] [41] [42]. Overall, an enterprise success platform

[39] Vuong Band Hairston, C C. 2012, October. *Using data to improve minority-serving institution success.* Washington, DC: Institute of Higher Education Policy. Retrieved from http://www.ihep.org/sites/default/files/uploads/docs/pubs/mini_brief_using_data_to_improve_msi_success_final_october_2012_2.pdf.

combined with predictive analytics that are based on historical student data can make institutional staff more effective at helping students succeed.

[40] Center for Community College Student Engagement. 2013. *A matter of degrees: Engaging practices, engaging students (high-impact practices for community college student engagement)*. Austin, TX: The University of Texas at Austin, Community College Leadership Program. Retrieved from http://www.ccsse.org/docs/Matter_of_Degrees_2.pdf.

[41] Kuh G D, Kinzie J, Buckley J A, Bridges B Kand Hayek, J C. 2006, July. *What matters to student success: A review of the literature*. Washington, DC: National Postsecondary Education Cooperative. Retrieved from http://nces.ed.gov/npec/pdf/kuh_team_report.pdf.

[42] Tinto Vand Pusser B. 2006, June. *Moving From theory to action: Building a model of institutional action for student success*. Washington, DC: National Postsecondary Education Cooperative. Retrieved from https://nces.ed.gov/npec/pdf/Tinto_Pusser_Report.pdf.

Massively Open Online Courses (MOOCs)

Andrew Ho (Harvard), Piotr Mitros (MIT), Diana Oblinger (EDUCAUSE) and Una-May O'Reilly (MIT)

THE EDUCATIONAL OPPORTUNITY

There are many types of big data that can be collected in learning environments. Large amounts of data can be gathered across many learners (broad between-learner data), but also within individual learners (deep within-learner data). Data in 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.[43] These types of big data in education potentially provide a variety of opportunities, such as:

- Individualizing a student's path to content mastery, through adaptive learning or competency-based education.
- Better learning as a result of faster and more in-depth diagnosis of learning needs or course trouble spots, including assessment of skills such as systems thinking, collaboration, and problem solving in the context of deep, authentic subject-area knowledge assessments.
- Targeted interventions to improve student success and to reduce overall costs to students and institutions.
- Using game-based environments for learning and assessment, where learning is situated in complex information and decision-making situations.
- A new credentialing paradigm for the digital ecosystem, integrating micro-credentials, diplomas, and informal learning in ways that serve the individual and employers.
- 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, etc.

MOOCs provide many of these opportunities, but substantial challenges in data-intensive research must be resolved to realize their full potential.

MITROS: THE POTENTIAL VALUE OF MOOCS FOR ASSESSING COMPLEX SKILLS

Historically, assessment in classrooms was limited to instructor grading or machine grading for problems that lend themselves well to relatively simple automation, such as multiple-choice questions. Progress in educational technology, combined with economies of scale, provides tools that digitally measure student performance on authentic assessments, such as engineering design problems and free-form text answers, radically increasing the depth and the accuracy of measurements of what students learn, allowing the tailoring of instruction to specific students' needs, and giving individualized feedback for an increasing range of learning issues. In addition, social interactions have increasingly moved online. This provides traces of a substantial portion of student-to-student interactions. By integrating these and other sources of data, the learning scientist has data to estimate complex skills, such as mathematical maturity, complex problem solving, and teamwork for large numbers of students. The next grand challenge in big data in education will be finding ways to analyze complex data from heterogeneous sources to extract such measurements.

Twenty years ago, most digital assessments consisted of multiple choice questions and most social interactions happened in person. Data was spread out across multiple systems with no practical means of integration.

[43] Thille C, Schneider DE, Kizilcec RF, Piech C, Halawa SAand Greene D.K. 2014. The Future of data–enriched assessment. *Research & Practice in Assessment*, *9*(2), 5-16. http://www.rpajournal.com/dev/wp-content/uploads/2014/10/A1.pdf.

Over the past two decades, we have seen fundamental progress in educational technology, combined with broad-based adoption of such technology at scale.[44] Digital assessment has increasingly moved toward rich authentic assessment. Previously, widely available data for large numbers of students principally came from standardized exams or standardized research instruments, such as the Force Concept Inventory. These assessments are limited to a short time window; as a result, they either contain a large number of small problems (which ensures the results are statistically significant but generally fail to capture complex skills that require more than a minute or two to demonstrate), or a small number of large problems (which on a per-student basis lack any statistical significance).

In contrast, today's researchers are increasingly collecting data on students who are doing large numbers of complex problems as part of their regular coursework. For example, the first edX/MITx course[45], 6.002x was implemented entirely with authentic assessment. Students completed circuit design problems (verified through simulation) and design-and-analysis problems (with answers as either equations or numbers)[46]. Since these types of questions have a near-infinite number of possible solutions, answers cannot be guessed. Students could submit an answer as many times as necessary in order to completely understand and solve a problem. The assessments were complex; most weeks of the course had just four assessments, but completing those four required 10–20 hours of work[47]. Similarly, there are rich assessments in courses such as chemistry, biology, physics, computer science, and many others. Such complex assessments, taken together across many courses, give rich data about problem-solving skills, creativity, and mathematical maturity.

Furthermore, researchers collect microscopic data about individual student actions. This gives the learning route for both correct and incorrect answers. Extensive research shows differences in problem-solving strategy between novices and experts. For example, experts can chunk information; for example, an expert looking at an analog circuit will be able to remember that circuit, whereas a novice will not[48] [49]. Thus microdata combined with rich assessments provides information on the development of expertise. Continuing with the example of chunking, we record how many times a student flips between pages of a problem set, looks up equations in a textbook, and similar activities that are proxies for expertise.

Further, social interactions are increasingly moving online. As increased amounts of digital group work are introduced in courses, more traces of social activity appear in logs. And the logs can then help in identifying students who underperform or overperform in group tasks, and directly measure students' group contributions. These systems provide enough data to begin to look for specific actions and patterns that lead to good overall group performance. Feedback can be provided to students by using these patterns to improve group performance. Natural language processing frameworks, such as the open-source edX EASE and Discern, are still used primarily for short-answer grading, but were designed to also apply to the analysis of social activities, such as emails and forum posts. Such frameworks will begin to give insights into soft skills, writing processes, communications styles, and group dynamics[50].

[44] We define at-scale learning environments as ones where thousands of students share common digital resources, and where we collect data about such use. This includes MOOCs, but also many educational technologies predating MOOCs, as well as formats such as small private online courses (SPOCs).

[45] Used both in a pure online format, as well as in a blended format in a number of schools.

[46] Mitros P, Affidi K, Sussman G, Terman C, White J, Fischer L, and Agarwal, A. 2013. Teaching electronic circuits online: Lessons from MITx's 6.002x on edX. In ISCAS, pages 2763–2766. IEEE.

[47] These are accurate estimates – they are consistent between both analysis of student activity in the courseware, and student surveys.

[48] Schneider W, Gruber H, Gold A, Opwis K. 1993. Chess expertise and memory for chess positions in children and adults. J Exp Child Psychol, 56, 328-49.

[49] Egan D E and Schwartz, B J. 1979. Chunking in recall of symbolic drawings. Memory and Cognition, 7(2), 149-158.

[50] Southavilay V, Yacef K, Reimann P, Calvo RA."Analysis of Collaborative Writing Processes Using Revision Maps and Probabilistic Topic Models" Learning Analytics and Knowledge - LAK 2013. Leuven, Belgium, 8-12 April, 2013.

Finally, aside from looking just within individual courses, longitudinal analysis across a student's educational career can be performed. In most cases, a single group design project does not provide statistically significant information. However, all of the projects over the duration of a student's schooling are likely to be significant. Learning analytics systems are increasingly moving in the direction of aggregating information from multiple sources across multiple courses. Open analytics architectures, such as edX Insights or Tin Can, provide a common data repository for all of a student's digital learning activities[51] [52]. That said, going from this type of massive data collection to the measurement of complex skills is a difficult problem.

HO: BEFORE "DATA COLLECTION" COMES "DATA CREATION"

Where does data come from? The phrases "data collection" and "data mining" both suggest that data simply exists for researchers to collect and mine. In educational research, I think a more useful term 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. If I create a massive open online course (MOOC), or an online educational game, or a learning management system, or an online assessment, I am less enabling the collection of data than creating data in a manner that happens to enable its collection.

This is a consequential perspective because it discourages lazy generalizations and false equivalencies. In previous work, my coauthors and I described MOOCs not as new courses but as new contexts, where conventional notions of enrollment, participation, curriculum, and achievement required reconceptualization[53]. We tempered early optimism about MOOCs as labs for researching learning by focusing on what made MOOCs different from seemingly analogous learning contexts in residential and online education: heterogeneous participants, asynchronous use, and low barriers to entry. Note that a completion rate is one minus a browsing rate, and browsing is a desired outcome for many MOOC participants[54]. Research that tries to increase completion rates (and, by definition, decrease browsing rates) is both poorly motivated and unlikely to inform dropout prevention where it matters in residential institutions and selective online courses.

Beyond MOOCs, I am arguing that one should be critical of any line of work that touts its "data intensive" or big data orientation without describing the contexts and processes that generate the data. When the context and process are particular, as they often are in big data educational research, applicants that promise general contributions to "how we learn" are likely to damage or at least muddy a field already overpopulated with mixed findings.

DEFINING (AND COMMITTING TO) THE MOOC "STUDENT"

In the previous section, I argue that we should view many "data-intensive" contexts in education not as familiar contexts with data but as unfamiliar contexts, else why would there be so much data? I believe this perception can productively refocus research on describing these contexts and determining whether, not just how, research findings within them generalize to contexts more familiar. In the context that I have studied most closely, which is Harvard and MIT open

[51] Siemens G, Gasevic D, Haythornthwaite C, Dawson S, Shum SB, Ferguson Rand Baker R. 2011. Open learning analytics: an integrated & modularized platform (Doctoral dissertation, Open University Press).

[52] Mitros P, Affidi K, Sussman G, Terman C, White J, Fischer L, and Agarwal, A. 2013 Teaching electronic circuits online: Lessons from MITx's 6.002x on edX. In ISCAS, pages 2763–2766. IEEE.

[53] DeBoe r J, Ho AD, Stump G Sand Breslow, L (2014). Changing "course": Reconceptualizing educational variables for Massive Open Online Courses. Educational Researcher, 43, 74-84.

[54] Reich J. 2014. MOOC completion and retention in the context of student intent. Educause Review Online.

online courses[55] [56], my colleagues and I do indeed find a "classroom" like no physical classroom on earth, with considerable variation in participant age, education, and geography, along with many teachers[57] and varying levels of initial commitment[58]. We and others have argued that this makes evaluating MOOCs extremely difficult, with the uncritical use of "completion rates" as an outcome variable being particularly problematic[59]. In this section, I make a normative argument that this difficulty should not exempt MOOCs from critical evaluation, and I point a path forward, coming full circle to completion rates.

I believe that many MOOC platforms, instructors, and institutions feel accountability to the first "M," for "Massive," and therefore report undifferentiated numbers of registrants whether they ultimately use or are interested in completing the course. Unsurprisingly, given the context I describe, completion rates for these registrants are very low. Unfortunately, the response by some MOOC insiders has been to rely on the contextual argument to exempt themselves from accountability to any metrics at all. I think this is bad science and bad pedagogy. Without a mutual sense of accountability, from students and instructors alike, I would describe MOOCs not as Massive Open Online Courses but Massive Open Online Content.

Content alone is a contribution, and content alone is indeed all that many instructors and institutions may be interested in providing. However, providing open content alone makes MOOC completion likely for a particular kind of learner—those who know what they need, those who are self-motivated, and those who have the time and skills necessary to keep themselves in the zone of proximal development as the course progresses. The general finding that MOOC registrants are disproportionately college educated is a testament to this. I consider this less "teaching" than "providing content to learners," a distinction that can also be described as that between "active teaching" and "teaching," similar to that between "active learning" and "learning." The consequence of passive teaching is that MOOCs will not close achievement gaps and provide a very limited definition of "access."

All MOOCs that commit to "active teaching" should embrace a common definition of a "committed learner" and make this clear to registrants and the public. My proposed definition of a "committed learner" is those registrants who: a) state a commitment to completing the course and b) spend at least 5 hours in the courseware. I choose this cutoff because it seems a sufficient amount of time for a student to understand what she or he is getting into (the "shopping period") and because it results in a completion rate of 50 percent in the Harvard and MIT data (tautologically, this maximizes variance in the dichotomous outcome variable). Instructors and institutions should publish counts of committed learners along with their completion rates and strive to improve them from baseline rates.

Importantly, this definition of "committed learner" does not exclude other participants. Under this model, browsers who are curious, auditors who merely wish access to videos, and teachers who are seeking materials may use MOOCs as they please. In other words, the natural response to the heterogeneity of the MOOC population is not to decide that measurement and accountability is impossible. It is the opposite: Now that we know who our participants are, the teacher's instinct is to hold oneself accountable to helping them achieve their goals.

[55] Ho AD, Reich J, Nesterko S, Seaton D, Mullaney T,Waldo J.and Chuang I. 2014. HarvardX and MITx: The First Year of Open Online Courses (HarvardX and MITx Working Paper No. 1).

[56] Ho AD, Chuang I, Reich J, Coleman C, Whitehill J, Northcutt C, Williams J J, Hansen J, Lopez G and Petersen R. 2015. HarvardX and MITx: Two years of open online courses (HarvardX Working Paper No. 10).

[57] Seaton DT, Coleman C, Daries J, and Chuang I. 2015. Enrollment in MITx MOOCs: Are We Educating Educators? Educause Review Online.

[58] Reich Jand Ho AD. 2014, January 24. Op-Ed; The tricky task of figuring out what makes a MOOC successful. The Atlantic.

[59] Ho AD, Reich J, Nesterko S, Seaton D, Mullaney T,Waldo Jand Chuang I. 2014. HarvardX and MITx: The First Year of Open Online Courses (HarvardX and MITx Working Paper No. 1).

TRAINING GRANTS FOR GRADUATE-LEVEL RESEARCH USING DIGITAL LEARNING DATA

I like to say that, in academia, the unit of work is not the professor but the graduate student. Graduate students also facilitate collaborations between research groups and push their advisers to learn new analytic methods and ask new questions. Some of the best researchers riding the recent wave of data-intensive research in education have been graduate students or recent graduates, and many of them have organized cross-disciplinary communities that could benefit from structured financial support and training. As big data in education is attracting those with little background in causal inference, assessment, or educational research, inferential and analytic errors will remain common and will include confusing correlation with causation, assuming all assessment scores are valid for their intended uses, assuming all distributions are normal, confusing statistical significance with substantial effect sizes, and generally wielding hammers without first asking whether there are any nails.

I think that a targeted investment in ongoing research training for doctoral students would be very wise long-term. As always, the keys to practical research training include granting students access to real data and training them in hands-on analytic methods. The Institute of Education Science (IES) Research Training Programs could serve as a model here, except that the particular focus would be on rigorous methods for drawing relevant inferences from digital learning data.

THE PURPOSE OF EDUCATION IS NOT PREDICTION BUT LEARNING

The most common questions I see being asked of digital learning data involve prediction, including prediction of certification, attrition, and future outcomes like course-taking patterns. I think it's worth remembering that, in any formative educational process, the criterion for prediction is not accuracy, as measured by the distance between predictions and outcomes. Instead, it is impact, as measured by the distance between student learning with the predictive algorithm in place, and student learning had it not been in place. I find the emphasis on technically sophisticated predictive models and intricate learning pathways to be disproportionate, and I think there is too little attention to rigorous experimental designs to ascertain whether students and instructors can use these tools to increase learning.

In short, we want educational predictions to be wrong. If our predictive model can tell that a student is going to drop out, we want that to be true in the absence of intervention, but if the student does in fact drop out, then that should be seen as a failure of the system. A predictive model should be part of a prediction-and-response system that a) makes predictions that would be accurate in the absence of a response and b) enables a response that renders the prediction incorrect. In a good prediction-and-response system, all predictions would ultimately be negatively biased. The only way to empirically demonstrate this is to exploit random variation in the assignment of the system, as in random assignment of the prediction-and-response system to some students, but not all.

MITROS: CHALLENGES IN ASSESSING COMPLEX SKILLS IN MOOCS

PEDAGOGICAL DESIGN

Making measurement an objective of instructional design can create substantial challenges. Assignments and assessments in courses have several objectives:

- *Initial and formative assessment as an ongoing means of monitoring what students know.* This allows instructors and students to tailor teaching and learning to problematic areas[60].
- *The principal means by which student learn new information.* In many subjects, most learning happens through assignments where students manipulate,

[60] Sadler D. 1989. Formative assessment and the design of instructional systems. Instructional Science 18: 119-144.

derive, or construct knowledge–not lectures, videos, or readings[61].

- *A key component of grading*. Grading itself has multiple goals, from certifying student accomplishment to providing motivation for desired student behaviors.
- *Summative assessment of both students and courses*. Summative assessment has many goals, such as student certification and school accreditation.

Historically, different research communities emphasized diverse objectives and gave very different principles around how good assessments ought to be constructed. For example, the psychometrics community principally relies on metrics such as validity and reliability. These suggest a high level of standardization in assessments. In contrast, the physics education research community emphasizes concepts such as the trade-off between authentic assessment and deliberate practice, as well as principles such as rapid feedback, active learning, and constructive learning[62]. As another point of view, educational psychology and gamification emphasize mastery learning[63].

Numerical techniques, which presume that assessments are designed based on principles which optimize for measurement, often fail when applied to the much broader set of classroom assessments. There is an inherent friction between:

- Having a sufficient number of problems for statistical significance vs. long-form assessments, which allow students to exercise complex problem-solving and mathematical maturity.
- Measuring individual students vs. group work.[64]
- Standardized assessments vs. diversity in education. The U.S. economy benefits from a diverse workforce, and the educational system, especially at a tertiary level, is designed to create one. There are more than 10,000 distinct university-level courses.
- Aiming for 50 percent of questions correct (maximize measurement) vs. 100 percent of concepts mastered (mastery learning)

To give an example of how friction comes into play, the MIT RELATE group applied item response theory (IRT), a traditional psychometric technique, to calibrate the difficulty of problems in 6.002x, the first MITx/edX course[65]. However, IRT presumes that problem correctness is a measure of problem difficulty. 6.002x is based on mastery learning, and students can continue trying until they answer a question correctly, and any sufficiently dedicated student could answer all questions correctly. To apply IRT in this context, RELATE had to substantially adapt the technique[66].

DIVERSITY AND SAMPLE BIAS

Many traditional psychometric techniques rely on a relatively uniform data set generated with relatively unbiased sampling. For example, to measure learning gains, we would typically run a pretest and a posttest on the same set of students. In most at-scale learning settings, students drop out of classes and take different sets of classes; indeed, the set of classes taken often correlates with student experience in previous classes. We see tremendous sampling bias. For example, a poor educational resource may cause more students to drop out, or to take a more basic class in the future. This shifts demographics in future

[61] Chi., MT. 2011. Differentiating four levels of engagement with learning materials: the icap hypothesis. *International Conference on Computers in Education.*

[62] Ericsson K, Krampe Rand Tesch-Romer C. 1993. The Role of Deliberate Practice in the Acquisition of Expert Performance. Psychological Review, 3, 363-406.

[63] Bloom, B. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16.

[64] At this point, we have overwhelming evidence that well-structured group work leads to improved student outcomes.

[65] Embretson S Eand Reise S. 2000. Item response theory for psychologists. Mahwah, NJ: Erlbaum Publishers.

[66] Champaign J, Colvin K F, Liu A, Fredericks C, Seaton Dand Pritchard D. E. 2014. Correlating skill and improvement in 2 MOOCs with a student's time on tasks. In *Proceedings of the first ACM learning@scale conference* (pp. 11–20). ACM.

assessments to stronger students taking weaker courses, giving a perceived gain on post-assessment unless such effects are taken into account.

Likewise, 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.

DATA SIZE AND RESEARCHER SKILLSET

Traditionally, statisticians in schools of education conducted big data educational research with tools such as spreadsheets and numerical packages such as R. This worked well when data sets were reasonably small. However, a typical data set from a MOOC is several gigabytes, and the data at a MOOC provider is currently several terabytes. While this is not big data in a classic sense, the skills and tools required for managing this data go far beyond those found at many schools of education. And with continuing moves toward technologies such as teleconferencing, we expect data sets to grow many-fold.

As a result, most data science in MOOCs has been conducted in schools of computer science by researchers generally unfamiliar with literature in educational research. This shortcoming is reflected in the quality of published results; for example, in many cases, papers unknowingly replicate well-established, decades-old results from classical educational research. Meaningful research requires skillsets from both backgrounds, but few researchers possess such skillsets, and collaborations are sometimes challenging due to substantial cultural differences between schools of education and schools of computer science.

EARLY SUCCESSES DESPITE THESE OBSTACLES

An early set of high-profile successes in this sort of data integration came from systems that analyzed data across multiple courses in order to predict student success in future courses. These systems include Purdue Course Signals[67], Marist Open Academic Analytics Initiative[68], and Desire2Learn Student Success System[69].

There have been early successes with systems that look at different types of data as well. For example, the first prototype of the edX Open-ended Response Assessment (ORA1) system integrated:

- *Self-assessment:* students rate their own answers on a rubric
- *Peer assessment:* students provide grading and feedback for other students' assignments
- *Instructor assessment:* the traditional form of assessment
- *AI assessment:* a computer grades essays by attempting to apply criteria learned from a set of human-graded answers.

In the theoretical formulation, each of the four grading systems contributes a different type and amount of information[70]. The system routes problems to the most appropriate set of grading techniques. An algorithm combines responses from graders to individual rubric items into feedback and a final score. A simplified form of this algorithm was experimentally validated.

[67] Arnold K Eand Pistilli MD. 2012. Course Signals at Purdue: Using Learning Analytics to Increase Student Success. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 267-270.

[68] Lauría E, Moody E, Jayaprakash S, Jonnalagadda Nand Baron J. 2013. Open academic analytics initiative: initial research findings. Proceedings of the Third International Conference on Learning Analytics and Knowledge.

[69] Essa A and Hanan A. 2012. Improving student success using predictive models and data visualisations. Research in Learning Technology. Supplement: ALT-C 2012 Conference Proceedings.

[70] Mitros Pand Paruchuri V. 2013. An integrated framework for the grading of freeform responses. The Sixth Conference of MIT's Learning International Networks Consortium.

TRANSCENDING THE LIMITS OF CURRENT ASSESSMENTS

In summary, while many of the goals of an educational experience cannot be easily measured, data-intensive educational research can make it much easier to improve, control, and understand those that can. The breadth and depth of data now available has the potential to fundamentally transform education.

Students and instructors are incentivized to optimize teaching and learning to measured skills, often at the expense of more difficult-to-measure skills. While we have seen tremendous progress in education with the spread of measurement, limited or inaccurate assessments can cause actual harm if relied on too much. Measurement in traditional education is tremendously resource-constrained, which severely restricts what can be measured. Standardized high-stakes tests are typically 3–4 hours long, and must be graded for millions of students in bulk. In most cases, such high-stakes exams can only accurately measure some skills and use those as proxies for more complex-to-measure skills. Many tests completely fail to capture skills such as mathematical maturity, critical thinking, complex problem-solving, teamwork, leadership, organization, time management, and similar skills.

While time constraints in traditional classroom settings are somewhat more relaxed than in high-stakes exams, instructors still often rely on proxies. For example, when measuring communication skill, a common proxy is an essay, a medium relatively rare outside of the classroom. Instructors cannot effectively critique longer formats of communications, such as email threads, meetings, and similar without extreme student/faculty ratios–but computers can.

Digital assessments have long been effective means to liberate instructor time, particularly in blended learning settings, as well as for providing immediate formative feedback[71 72 73]. Building on this work, we are increasingly seeing a move to authentic assessment, approaches where humans and machines work in concert to quickly and accurately assess and provide feedback to student problems, where data is integrated from very diverse sources, and where data is collected longitudinally[74].

With this shift, for the first time, we have data about virtually all aspects of students' skills, including complex abilities that are, ultimately, more important than simple factual knowledge[75]. We have the potential to provide new means to assess students in ways which can improve depth, frequency, and response time, potentially expanding the scope with which students and instructors can monitor learning, including assessment of higher-level skills, and proving personalized feedback based on those assessments. However, the tools for understanding this data (edX ORA, Insights, EASE, and Discern, in our system, and their counterparts in others) are still in their infancy. The grand challenge in data-intensive research in education will be finding means to extract such knowledge from the extremely rich data sets being generated today.

O'REILLY: ENABLING TECHNOLOGIES FOR DATA-INTENSIVE RESEARCH ON MOOCS

ALFA's research centers on elucidating the general design principles of data science workflows that enable rapid data transformation for analytic and predictive purposes. We currently have a project called MOOCdb (url: MOOCdb). One of the project's overarching goals is to identify and develop enabling technology for data-intensive research into MOOCs. The project is intended

[71] National Research Council. 2000. *How people learn.* (pp. 67–68, 97–98). National Academy Press.

[72] KURT VanLEHN (2011): The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems, Educational Psychologist, 46:4, 197-221.

[73] Novak GM, Patterson ET, Gavrin ADand Christian W. 1999. *Just-in-time teaching: Blending active learning with web technology.* New Jersey: Prentice-Hall.

[74] Basu S, Jacobs C and Vanderwende L. 21013, Oct. Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading, in *Transactions of the ACL,* ACL – Association for Computational Linguistics.

[75] Sternberg R. 2013, June 17. Giving employers what they don't really want. *Chronicle of Higher Education.*

to unite education researchers and technologists, computer science and machine learning researchers, and big data experts toward advancing MOOC data science. It also supports our specific learning science research into MOOC student online problem-solving, resource usage behavior, and persistence prediction.

One high-profile ambition of the project revolves around developing the means to efficiently study MOOC student behavior across multiple MOOCs released on different platforms (specifically Coursera and edX). This capability will allow cross-platform comparisons of learning behavior across or within institutions. It will facilitate the detection of universal aspects of behavior, as well as tease out the implications of important differences. Other project activities have goals such as enabling a collaborative, open-source, open-access data visualization framework, enabling crowdsourced feature discovery (featurefactory), and preserving the privacy of online student data. ALFA's team is currently working to openly release a number of its tools and software frameworks.

A short description of the multiple raw data streams of MOOC edX platform data that are supplied for data science and analytics can be found in Section 2 of "Likely to Stop? Predicting Stopout in Massive Open Online Courses" (arXiv#1408.3382). By far the largest is clickstream data. To analyze this data at scale, as well as write reusable analysis scripts, it is first organized into a schema designed to capture pertinent information and make cross-references to the MOOC's content. That schema is exhaustively described in the MOOCdb report. Chapter 2 of "Modeling Problem Solving in Massive Open Online Courses" provides a summary.

In the course of answering learning science questions, like "Who is likely to stop?" we add interpretation and knowledge to transform and enhance the data that lies in MOOCdb tables. This data is of a higher-level nature or of a particular data abstraction and is stored in new tables. For example, we might efficiently express each learner's trajectory of actions when solving each problem or a learner's navigation sequence through material in each module. See Chapter 4 of "Modeling Problem Solving in Massive Open Online Courses" for a clear example explaining the transformation of data to form student trajectories for every problem of a MOOC.

When we develop predictive models of learner behavior, we use the transformed data directly, or with some logic, populate yet another table that consisting of predictive features and labels for each machine learning training or testing example. The training data is input to the machine learning algorithm where the label acts as a supervisory signal and the features as explanatory model variables. The testing data is used to gauge generalized model accuracy.

We are computer scientists, so we fairly routinely develop software and use open source and/or commercial software. Our software operates at every part of the data science workflow. We execute our analyses on workstations and the cloud and databases. To achieve interoperability with MOOC data from different platform providers, we initiated the MOOCdb schema and the open source release of translation software. (For more information, see the MOOCdb documentation.) We develop software that we intend to share once it is release-ready. For machine learning, we largely use open source libraries situated within our own research frameworks that allow rapid scaling and result comparison.

STRATEGIES FOR BUILDING PARTNERSHIPS BETWEEN BIG DATA PRODUCERS AND CONSUMERS

The ideal time for a partnership is before or during education technology design and implementation. If education technologists and instructors can explicitly communicate their learning goals and desired learning outcomes and articulate the intent of their assessments AHEAD OF IMPLEMENTATION, the "producers" will be able to instrument the technology in a way that captures the appropriate feedback to validate hypotheses and outcome success.

One strategy is to encourage development projects where the stakeholders work together toward a deliverable, rather than the consumers receiving the data after digital learning. One goal of such projects, from a software technology perspective, should be open source middleware that hides layers of functionality that are necessary but not central to the consumer's mission. This is much as Amazon Web Services (AWS) does with a lot of its services. AWS services always handle compute scalability, elasticity, and reliability. This allows their "consumer" to focus on the tasks central to their business without attending to aspects, like scaling, that are not central to their mission. The AWS services also provide convenient interface abstractions and design patterns that are very common to their consumers. AWS develops the patterns for its internal business, gets them "right," and then offers them externally where they significantly help save development time.

Another way to answer this question is to list explicit examples of producers and consumers. In the MOOCsphere, the producers are the platform providers: edX and Coursera. The consumers of data are stakeholders: students, instructors, education technologists, institutional registrars, and learning scientists. In the MOOCsphere, relationship building has been driven by the platform providers because they possess the data.

ISSUES OF PRIVACY AND SECURITY

MOOC learners will require privacy during personalized learning interventions. We need to deeply explore different positive and negative scenarios in this context so we can inform and keep policy up to date, then define policy-dictated boundaries to inform capabilities, and, finally, develop the required privacy technology. One technology question would be: How do we design the algorithms and personalization technology to be accountable to policy?

MOOC learners also require privacy protection long after their learning interaction is completed and logged. In the digital learning enterprise, multiple stakeholders have legitimate reasons to retrospectively access logged data and analyze it. In the current context of digital learner data being shared, the concerns for learner privacy must be respected. Even when personally identifiable records within the data are removed and the learner's identity is replaced with a randomized value, there remains a risk of re-identification (i.e., the recovery of a specific learner's identity).

From the learner's perspective, despite contributing MOOC data, they do not directly receive or control it. The learner acknowledges this arrangement by accepting a terms of use agreement in exchange for using the site. They are briefed about the reasonable protections that will be afforded to their personally identifiable information via a site privacy policy.

As a result of the learner's activity, the data passes to the platform and content providers. From this perspective, their responsibility is to oversee and control its further transmission. They are entrusted, by the learner, to respect relevant parts of the terms of use and privacy policy. In transmitting the data, their current practice is essentially to ensure the receiver is trustworthy while transmitting the minimum data required in order to minimize potential privacy loss. They further bind the parties to whom they transfer data with some form of data use agreement.

Finally, from the perspective of those who receive learner data from platform or content providers, they agree to the data use agreement that commits them to fundamental measures that protect the learner. These include not ever attempting to re-identify anyone from the data, not contacting a learner they might recognize, and not transmitting the data onwards.

Overall, this process relies on direct verification of people and institutions, which provides the basis for the system's integrity.

The process culminates, indirectly, in trust that the best efforts of the parties involved to honor their commitments will be sufficient. This endpoint exposes the process's vulnerability; it assumes the data won't inadvertently fall into the wrong hands, when, in fact, it may. This problem could happen when the data is held by any of the data controllers. This danger implies a need for the development of new, practical, scalable privacy protection technology to mitigate the risk arising should the data fall into the wrong hands. This need is arguable because, to date, there is only one MOOC-related dataset in general open release, the HarvardX-MITx Person-Course Academic Year 2013 De-Identified data set.

It holds aggregate records, one per individual per single edX course for five MOOCs offered by HarvardX and eight by MITx. The dataset is "sanitized" for release by two complementary privacy protection technologies. It achieves k-anonymity (for k = 5), a measure of degree of de-identification, by a means called "generalization of quasi-identifiers." Using a second mechanism, it checks for L-diversity along sensitive variables and, if all values of a variable are the same, redacts the value. In fact, the release is not completely open because a terms of use agreement is required to download the data set; however, it provides a solid starting point for future open releases. The k-anonymity measures and L-diversity redaction don't provide a quantitative tradeoff, measuring risk of re-identification and utility.

One option that does offer this tradeoff measure is differential privacy. While research in differential privacy is largely theoretical, advances in practical aspects could address how to support the content and platform providers who transmit the data when they want to choose a tradeoff between risk of re-identification and utility. Subsequently, effort would be required to mature the demonstrations for regular use by the development of prototypes that have user-friendly interfaces to inform controller decisions. Controller acceptance will require a set of technology demonstrations that, in turn, require major effort and resources. Demonstrations would be feasible if a "safety zone" could be set up where technology can be explored and validated against friendly re-identification adversaries who try to crack identities without causing any threat of real harm to the learners' data. Data scientists in the MOOC analytics sphere who develop variables and analytic models should be encouraged and supported to explore differential privacy mechanisms and bring them to practice.

OBLINGER: CHALLENGES IN SCALING UP SUPPORT SYSTEMS FOR DATA-INTENSIVE EDUCATIONAL RESEARCH

Previous sections of this chapter have addressed MOOCs as a "context" rather than a course. Other parts have looked at assessing complex skills as well as enabling technologies that will allow us to study these big data sets. This section deals with scaling up the systems that support data-intensive educational research.

What this sampling of issues related to scaling up the outcomes of data-intensive research in education have in common is that we are dealing with a complex system. One illustrates the challenges of integration in complex systems (an integrated competency management system for students, higher education, and employers); another focuses on technical infrastructure (a next-generation digital learning environment). Two issues illustrate challenges in human capacity, specifically awareness and adoption and workforce development. The final area illustrated deals with policy. Note that some issues (e.g., policy ones) may not lend themselves to NSF-supported research; however, they must be addressed to achieve the potential of data-intensive environments.

INTEGRATED COMPETENCY MANAGEMENT SYSTEM FOR STUDENTS, HIGHER EDUCATION, AND EMPLOYERS

Big data in education can be used to build knowledge, enhance skills, and document an individual's or group's capabilities. Just as assessment can be longitudinal, the competencies learned throughout a lifetime can be longitudinal. This data can be useful for individuals, institutions, and employers. Today, however, the systems that document these competencies are not integrated, thus limiting its potential educational value. Adam Newman has described both the challenge and the opportunity[76].

According to Newman, there is an opportunity to use big data capabilities to create an integrated competency management system that supports students, higher education, and employers. Such a system would integrate "the body of knowledge, skills, and experience achieved through both formal and informal activities that an individual accumulates and validates during their lifetime[77]."

[76] Newman A. 2015, February. Evidence of Learning: The Case for an Integrated Competency Management System. http://tytonpartners.com/library/evidence-learning-case-integrated-competency-management-system/ .

[77] Ibid.

The current environment for skills, credentials, and employment opportunities is disconnected. Students attend multiple institutions and can assemble experience and credentials that go beyond a degree. Students use non-institutional career development networks, in part because institutions do not have enough career services professionals. Students and employers are turning to LinkedIn, Monster, and CareerBuilder. For example, LinkedIn reports hosting 300 million individual profiles. More than 75 percent of employers use social networks for employee recruitment. The opportunity appears to be significant. For example, investors have dedicated more than $700 million to education businesses focused on ventures that disaggregate and re-aggregate credentials[78].

"Foundational lifelong skills such as critical thinking, teamwork and collaboration, and problem solving are climbing to the top of employers' wish lists, and yet few institutional measures capture these attributes. These dynamics are pushing students and employers to explore alternative platforms for both presenting and evaluating profiles that capture an individual's evidence of learning."[79]

Newman cites at least five elements that involve big data:

- Experience: The process of learning, formally or informally, including MOOCs, adaptive learning, social learning models, etc. Also included are non-course-based learning activities.
- Validate: Assessing and recognizing experiences for credit or qualifications, including non-cognitive attributes of students, badging or micro-credentialing, credit for prior learning and training experiences.
- Assemble: Capturing and curating evidence of learning, including transcripts, assessments, outside learning experiences, and so on.
- Promote: Marking the assembled evidence, which may include social media analytics, behavioral assessment, and other data-mining techniques, to link candidates with opportunities.
- Align: Using feedback loops to constantly evaluate performance and make improvements at the individual and enterprise level.

Today, this emerging cross-segment competency management system appears to be developing outside of higher education. Colleges and universities can bridge students and the workplace by aligning learning outcomes across institutions and employers. But developing scalable systems will also require technical integration and workflow processes. Research could advance individual elements (e.g., adaptive learning, non-cognitive skill assessment, etc.) of this framework. Research may catalyze the necessary data exchanges among institutions and employers that will be required for such a system to be successful.

NEXT-GENERATION DIGITAL LEARNING ENVIRONMENT

If MOOCs represent a new context, not just content, and the purpose is learning, then researchers should explore how the technical infrastructure should be designed. The LMS is the most ubiquitous digital tool in higher education. In spite of its prevalence, the LMS is largely designed to administer learning (e.g., distribution of materials, gradebooks, etc.) rather than enabling it. It is also predicated on a course-centric and instructor-centric model. That model is being replaced with a focus on learning and the learner, moving beyond courses and today's credentialing systems. The LMS needs to be replaced by a new digital architecture and components for learning. This "next-generation digital learning environment" may not be a single application like today's LMS but be more of a "mash-up" or "Lego set."[80] EDUCAUSE research suggests that the next-generation digital learning environment (NGDLE) will be an ecosystem of sorts, characterized by:

[78] Ibid.

[79] Ibid, page 6

[80] Brown M, Dehoney J, and Millichap N. 2015. The next generation digital learning environment: a report on research. (EDUCAUSE Learning Initiative Paper). http://net.educause.edu/ir/library/pdf/eli3035.pdf.

- Interoperability and integration: Interoperability is the linchpin of the NGDLE. The ability to integrate tools and exchange content and learning data enables everything else.
- Personalization: Personalization is the most important user-facing functional domain of the NGDLE.
- Analytics, advising, and learning assessment: The analysis of all forms of learning data is a vital component of the NGDLE and must include support for new learning assessment approaches, particularly in the area of competency-based education.
- Collaboration: The NGDLE must support collaboration at multiple levels and make it easy to move between private and public digital spaces.
- A cloud-like space to aggregate and connect content and functionality, similar to a smartphone, where users fashion their environments directly with self-selected apps.[81]

In addition, there may be a host of additional NGDLE components, such as:

- Learning environment architectures: A set of exemplary NGDLE architecture designs, which could serve as models for the community.
- Smart tools: A set of learning-tool designs that explicitly incorporate learning science and universal design and are fully NGDLE compliant.
- Learning measurement rubrics: A set of designs to effectively integrate new rubrics for learning measurement and degree progress (e.g., competency) into the NGDLE.[82]

Research is needed to validate these elements and document best practices in architectures, tools, rubrics, etc.

HUMAN FACTORS

Achieving the promise of data-intensive educational environments hinges on human factors, as well as technological ones. Two examples include awareness and adoption and workforce development.

AUDIENCE, AWARENESS, AND ADOPTION

Awareness and adoption of data-intensive educational tools is very uneven. MOOCs are an example. EDUCAUSE surveys found that about three in four faculty (76 percent) said they are either conceptually or experientially familiar with MOOCs; compare this to only one in four undergraduates (24 percent) who say they know what a MOOC is. Although few faculty reported having actually taught a MOOC (3 percent), they are much more likely than students to know about this alternative model for online learning.[83]

Part-time faculty (53 percent) expressed more support than full-time faculty (38 percent); furthermore, non-tenure-track faculty (46 percent) were more supportive than tenured (34 percent) or tenure-track faculty (39 percent). About two in five faculty (43 percent) with less than 10 years of teaching experience were supportive, whereas somewhat fewer faculty (37 percent) with 10 or more years of experience were supportive. Not surprisingly, the picture painted here is that newer faculty have more positive perceptions of MOOCs adding value to higher education.[84]

The population enrolling in MOOCs may be somewhat different than earlier predictions. Young learners are a rising proportion of the MOOC population, according to University of Edinburgh research, with those under 18 rising 50 percent. While they are still only 5 percent of the learners on average, the increase may be tied

[81] Ibid

[82] bid

[83] Dahlstrom E and Brooks DC. 2014, July. ECAR Study of Faculty and Information Technology, 2014. (ECAR Research Report) http://net.educause.edu/ir/library/pdf/ers1407/ers1407.pdf .

[84] Dahlstrom E and Brooks DC. 2014, July. ECAR Study of Faculty and Information Technology, 2014. (ECAR Research Report) http://net.educause.edu/ir/library/pdf/ers1407/ers1407.pdf .

to teachers.[85] Recent research from edX and HarvardX illustrated that a major audience for MOOCs are teachers (28 percent of enrollees in 11 different MOOCs were former or active teachers).[86] As we understand more about MOOC audiences and motivations, we may need to shift the design of MOOCs to better align with audiences served. Ongoing research on audiences, experiences, and outcomes will be important.

WORKFORCE DEVELOPMENT

Data-intensive environments demand a new type of professional that some call data scientists. No matter what the name, higher education needs to develop the skills of these professionals as well as a pipeline into the profession. Data science is a blend of fields, including statistics, applied mathematics, and computer science. Qualities of data scientists who can address data-intensive challenges include:

- Technical skills: Mathematics, statistics, and computer science skills to work with data and analyze it.
- Tool mastery: Complex software tools are critical to analyzing massive amounts of data.
- Teamwork skills: Almost all of the data science roles are cross-disciplinary and team-based; hence, teamwork skills are critical.
- Communication skills: Deriving insights from data, communicating the value of a data insight, and communicating in a way that decision makers can trust what they're being told.
- Business skills: Understanding the business and bringing value from contextual understanding to the data analysis.[87]

Developing an understanding of the skills essential in data scientists and others who support big data systems will be important so that institutions can develop the appropriate training and education programs, as well as attract students.

POLICY AREAS

Most data-intensive environments represent risks and challenges in policy areas, particularly privacy and security. While there may be model policies in place at some institutions, the appropriate policy infrastructure is not in place at many institutions. In addition, many policy discussions are hampered by misinformation and fear. And federal regulations, such as FERPA, are often misunderstood. Appropriate policies must address privacy, security, and data sharing.

Good information security practices are essential to reduce risk; safeguard data, information systems, and networks; and protect the privacy of the higher education community. Good institutional information security practices encompass the technologies, policies, and procedures, and the education and awareness activities that balance the need to use information to support institutional missions with the need to protect the institution from internal and external threats and ensure the privacy of the campus community. These practices constantly evolve as the threat landscape evolves.[88]

All individuals associated with colleges and universities, whether faculty, staff, or students, need to protect their privacy and control their digital footprint. Big data environments escalate the importance of ensuring that protecting privacy and data are everyone's priority. There are different types of privacy that should be recognized. For example, autonomy privacy is an individual's ability

[85] Macleod H, Haywood J, Woodgate A and Alkhatnai M. 2015. Emerging patterns in MOOCs: Learners, course designs and directions. *TechTrends, 59*(1), 56-63. doi:10.1007/s11528-014-0821-y.

[86] Pope J. 2015. What Are MOOCs Good For? *Technology Review, 118*(1), 68-71. http://www.technologyreview.com/review/533406/what-are-moocs-good-for/ .

[87] Woods D. 2012, March. What Is a Data Scientist?: Michael Rappa, Institute for Advanced Analytics. *Forbes Magazine*. http://www.forbes.com/sites/danwoods/2012/03/05/what-is-a-data-scientist-michael-rappa-north-carolina-state-university/3/ .

[88] EDUCAUSE. 2014, August. Foundations of Information Security: Institutional Implications for Securing Data. http://net.educause.edu/ir/library/pdf/pub4011.pdf.

to conduct activities without any concern of or actual observation. Information privacy is the appropriate protection, use, and dissemination of information about individuals. Information security supports, and is essential to, autonomy and information privacy.[89]

Institutions must be aware of many ramifications of big data, such as:

- Legal and compliance issues: The consequences of compliance failure in analytics systems may be significant. Regulatory compliance (such as FERPA and HIPAA), e-discovery rules, open records laws, student privacy expectations, and the role of the institutional review board may all come into play.
- Unintended consequences of third-party data access and use: The use of big data systems may raise concerns about third-party misuse of data or its use for anything other than its intended purpose.
- Inappropriate use of data: Institutions may make the inappropriate use of the data presented in dashboards or reports, or misunderstand their limits.[90]
- Data ownership: Arguments exist for students to control data about themselves, as they do for institutions. The success of analytics depends on institutions accessing, curating, harvesting, and controlling multiple sources of data. Lack of control over the data might compromise the integrity of data-driven initiatives.[91]

Research associated with data-intensive applications must be based on an understanding of the relevant policy factors. And institutional implementation of these systems will only be successful if there is a solid policy framework at the institution, as well as at federal levels.

[89] Ho L. 2015. Privacy vs. Privacy. http://www.educause.edu/blogs/lisaho/privacy-vs-privacy.

[90] EDUCAUSE. (2014, April). What Leaders Need to Know about Managing Data Risk in Student Success Systems. http://www.educause.edu/library/resources/what-leaders-need-know-about-managing-data-risk-student-success-systems .

[91] Jones KML, Thomson, J and Arnold K. 2014, August 25. Questions of data ownership on campus. *EDUCAUSE Review Online*. http://www.educause.edu/ero/article/questions-data-ownership-campus .

Games and Simulations

Matthew Berland (University of Wisconsin-Madison), Chris Dede (Harvard University), Eric Klopfer (MIT) and Valerie Shute (Florida State University)

KLOPFER: THE EDUCATIONAL CHALLENGE

Few people learn anything *from* playing games, but there is a potential for many people to learn things *with* games. The distinction comes from how we situate game play into the learning experience. Games have the greatest potential impact on learning when they are part of an experience that also involves reflection, abstraction, and the application of concepts. This differentiates what has come to be known as the *game* (the digital distributed experience) from the *Game* (the entire experience including what happens on and off screen, such as interactions with peers and mentors, and the use of complimentary media like websites and video).

While some learners may possess the skills necessary to consciously reflect on what they are doing in a game in order to be able to abstract from specific instances to more general concepts, and then apply that understanding in a different context, in practice this is rare. Most students take the game play at its face value and would, on their own, struggle to connect that experience to learning goals. Instead this process typically needs to be scaffold by teachers (or peers, mentors, etc.). The question is then, How can we better support teachers in making that cycle of learning more efficient and more effective?

Addressing this challenge requires us to first take a step back and ask a series of questions about the goals and nature of game-based learning in classrooms today:

- What are the experiences that we want to provide students through games? And how are those situated in the learning experience?
- How do we design targeted experiences that focus on the learning activities that we are interested in? And how do we collect the relevant data from those experiences to guide teachers and learners?
- What kind of *actionable* data do we provide to teachers?

I argue that games have their greatest potential as learning experiences when they precede formal instruction, providing a concrete and common reference point upon which to build formal concepts. They further provide value as a touchpoint that students periodically return to as they iteratively build their knowledge in increasingly complex ways. Games provide meaningful learning experiences, and provide feedback to the learner on their understanding and engagement in that system.

THE POTENTIAL VALUE OF DATA-INTENSIVE RESEARCH ON GAMES AND SIMULATIONS

GAMES AND ASSESSMENT

Games may also play a role as a means for summative assessment, as they provide rich and complex problem spaces. But, more important, games can play a role as vehicles of formative assessment, where performance on tasks generates actionable information that guides their experience—and ultimately leads to enhanced learning. The data generated by games, and in games, creates a tremendous opportunity for supporting better learning experiences. As is often the case with data, the opportunities also bring their own challenges. Though we may be able to “fish in the exhaust” (as HarvardX researcher Justin Reich says) of the keystrokes and data trails of games to recognize successful patterns and differentiate them from those of players who struggle, that methodology is not yet sufficient for realizing this potential.

Instead, we must design for the learning experiences of games, the data they can generate, and specifically how we make sense of that data to inform further learning. Through offering specific activities and corresponding outcomes that can generate the data we need, not only can we differentiate success from failure, but we can identify why particular students are succeeding or struggling to support those students and allow all students to master the essential concepts. This means we need to follow an approach that helps designers create game-based tasks that elicit this useful data.

Evidence-Centered Design[92] (ECD) is one useful–and thus far highly popular among learning game designers–way of approaching this. ECD defines four relevant models:

- the student model (what the student knows or can do);
- the evidence model (what a student can demonstrate and we can collect to show what they know);
- the task model (the designed experience from which we can collect data); and
- the presentation model (how that actually appears to the student)

Though ECD was originally conceived by assessment developers to create better and more diverse assessments, it has become quite popular among learning game designers for its ability to create a framework for collecting and interpreting assessment data in games. Though the details of this methodology may seem onerous to a game designer seeking to create an experience, the ECD framework not only embodies the potential to create useful data, but also serves as a design lens that can provide engagement and challenges that draw players into the game.

In reality, a game is much more likely to become part of a larger educational experience if it can provide useful and actionable data to teachers, and this can really only come from an initial thoughtful and intentional design. Variations on ECD for the design of educational games may make this methodology easier and more effective to follow. As a lens for instructional design that aligns both content and assessment data in games, Groff et al.[93] have proposed a simplified version that reduces this to a Content Model (the relevant knowledge and skills), an Evidence Model (the information needed to know if someone has that knowledge) and a Task Model (what the person is engaged in doing to elicit that data). This model further links each of these sub-models in a more cyclic fashion, rather than a linear fashion as ECD typically provides, which is better aligned to how game designers think about their craft.

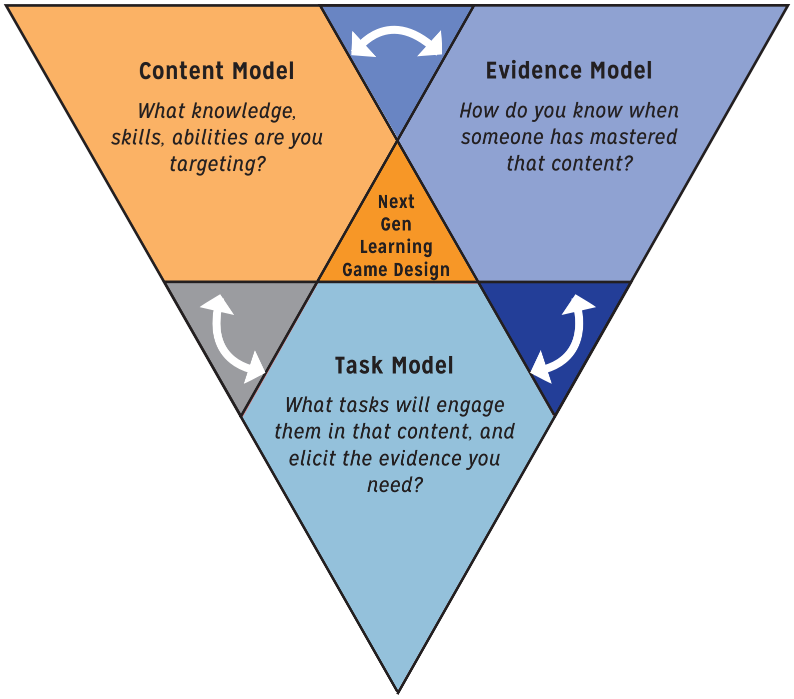

Content Model
What knowledge, skills, abilities are you targeting?
Evidence Model
How do you know when someone has mastered that content?
Next Gen Learning Game Design
Task Model
What tasks will engage them in that content, and elicit the evidence you need?

Figure 10: A simplified version of ECD known as XCD[52]

[92] Mislevy R, Almond Rand Lukas J. 2003. A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), 1-29.

[93] Groff J, Clarke-Midura J, Owens E, Rosenheck Land Beall M. 2015. *Better Learning in Games: An Expanded Framework for a New Generation of Learning Game Design*. A whitepaper by the Learning Games Network and the MIT Education Arcade.

Similarly, Conrad, Clarke-Midura, and Klopfer[94] have created a variant called Experiment Centered Design, in which the tasks are thought of specifically as series of experiments conducted by the learners. This works well for science-based games and math-based games in which players conduct experiments that both model the practices of those disciplines and provide a foundation upon which to design a series of tasks that can elicit relevant data. It is important in this methodology that we think of data not at the grain size of individual actions but rather as a series of related and predefined actions that comprise an iteration of an experiment. In spaces where the information is complex, mirroring authentic learning environments, single actions are not sufficient for accomplishing a task, and a priori defined chunks (e.g., experimental iterations) may make analysis both easier and more relevant to the learning outcomes.

For example, in a game designed around genetics experiments, the data may be thought of not as what a player does in a single breeding experiment, or even as the action taken based on the outcomes of such an experiment, but rather as an iterative series of experiments. In this case, the learner conducts an experiment, receives an outcome, and performs a new experiment based upon that outcome. Learners may in fact need to perform a fairly extensive sequence of these experiments, based upon both the complexity of the task and the random variation that may occur in those experiments.

THE RADIX GAME AS A CASE ILLUSTRATING FORMATIVE ASSESSMENT

We apply this Experiment Centered Design methodology in an educational massively multiplayer online (MMO) game called The Radix Endeavor (Radix). In Radix, players are set in an earth-like world in a Renaissance era state of knowledge, and must use math and science to help the world improve. Players get quests, which are tasks with proximate goals, along the way. Quests range from using geometry skills to fix buildings to diagnosing disease based on understanding of body systems. One of the quest lines is about genetics, and players get tasks such as delivering a “true breeding” strain of a medicinal plant. Starting with a stock of seemingly similar plants in the field, they must breed pairs of plants and observe the outcomes. A single outcome such as two plants producing identical offspring may not be sufficient for determining whether the plants are dominant or recessive, or even if they might be homozygous; and one has a sample too small to show diversity.

After the first breeding outcome, it is important to see what the player does next. Do they conduct the same experiment again, breed the offspring, or breed with one of the parental generation plants? From this sequence we can begin to uncover what the student understands about genotype and phenotype. Even in systems such as geometry, which are not stochastic, the series of measurements and building activities can be informative. An initial guess at the angles in a triangle may need to be adjusted in a second iteration, and it is key to observe which way they are adjusted. Based on these models, we can diagnose specific misconceptions and send players on “side quests” that specifically address their learning challenges.

In practice, doing this effectively is a significant challenge. Determining the student learning challenges, defining the models with sufficient specificity, implementing them, interpreting the data, and feeding it back to students and teachers is a lot of work. That work translates into cost, which is a challenge within the research space and a bigger challenge within the commercial space that might bring these games to scale. But it also provides an opportunity for creating games with increased value.

Looking at extended sequences of actions is also important in complex spaces to allow for exploration, times when players are simply orienting themselves and pursuing their own interests, which may not be targeting a particular learning outcome. Exploration is often seen as a desirable activity. In Radix, if players are exploring and conducting additional experiments on their

[94] Conrad S, Clarke-Midura Jand Klopfer E. 2014. A Framework for Structuring Learning Assessment in a Massively Multiplayer Online Educational Game: Experiment Centered Design. *International Journal of Game Based Learning*, *4*(1), 37-59.

own or just exploring the game's flora and fauna, we as game designers would view that as a positive. But it is difficult to detect when players are exploring, rather than simply not knowing what to do next. This is an area in which we may be able to examine players' patterns, past performance, and other factors to help nudge truly confused players in the right direction.

Related to exploration is the notion of productive failure—situations in which a player tests the bounds of the system, such as jumping off a cliff just to see what happens. The simple action of jumping off the cliff is not sufficient information to deduce whether the player is on a pathway to success or failure, even when that action leads to an outcome that may be perceived as negative (the player's temporary death). But such testing of the boundaries is important for players' understanding how the world in which they exist works. Longer sequences of actions–what the player does after that event–may provide a rich description of the learner's experience.

As an MMO, Radix provides us with the opportunity to also examine multiplayer interactions. This is a rich area to explore. The current iteration only provides optional multiplayer interactions–data sharing, "partying," chatting, etc. Structured interactions, in which players are differentiated by roles and given tasks, will provide better ways of examining these interactions from a data perspective, where we can infer some intentionality by role, as well as a player perspective.

Evidence-Centered Design, or any of these variants, allows us to identify the data of interest in advance. Rather than collecting every bit of data and parsing it after the fact, one can collect the necessary sequences based upon the defined tasks and provide real-time feedback on success or the lack thereof. However, there is still a roll for "fishing in that additional exhaust." As mentioned previously, we may be able to identify correlates of productive or counterproductive behaviors that we can pick up easily and use to provide additional feedback. In the formative case, we need not be certain that the person is on the right or wrong path; we need only to make a best-guess probe that guess and make a correction if it is not correct. We may also be able to identify additional behaviors or revise our theories on student understanding for the next iteration of a task. But in these cases, we should think of the data revising our theories, which in turn can influence our design and data collection, rather than the data itself directly informing students.

In some cases, the data can inform students about their progress, directly (by providing information about how they could improve) or indirectly (by giving them increasingly more difficult tasks when they are succeeding or breaking complex tasks down into simpler ones when they are not). But to turn the game into a Game that is a truly productive learning experience, the data must get out of the game and into the hands of the teacher in a useful way. This is a significant challenge, balancing the depth and complexity of information that we can provide, with the simplicity and immediacy that teachers need to make use of that data.

The first wave of simple dashboards that just show green, yellow, and red do not provide teachers with enough information to be useful on a case-by-case basis, other than knowing whether the class understands the material or not. The other end of the spectrum, which shows the outcome of every game action for every player, provides too much information to be useful. A teacher with 100 or more students cannot use such information to address individual or even class wide issues[95].

Additionally, as many of these models are probabilistic, we should provide teachers with the skills that they need to correctly interpret the incoming data. In fact, most assessment measures require a fairly sophisticated interpretation, but we don't usually convey this nuanced feedback. While we may not need to turn teachers into data scientists, we should provide them with a baseline of skills needed to interpret data.

[95] Ifenthaler D, Pirnay-Dummer P and Seel, NM. 2010 .Computer-based Diagnostics and Systematic Analysis of Knowledge. New York: Springer. N. pag. Print.

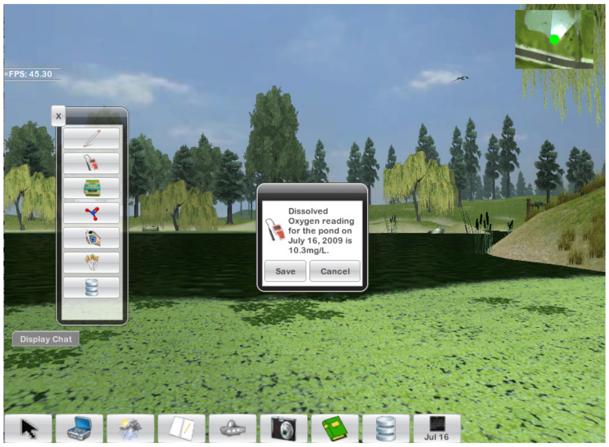
This all means that using games to learn what students know is not an activity that falls solely within the domain of data scientists; it is something that must draw upon the skills of learning scientists, instructional designers, game designers, and teacher educators. These roles are all required to define the necessary learning outcomes and challenges, develop effective and engaging tasks, and provide that data to teachers in actionable ways.

DEDE: IMMERSIVE AUTHENTIC SIMULATIONS

Building on work in educational games, multi-user virtual environments (MUVEs) and augmented realities (ARs) offer ways for students to experience richly situated learning experiences without leaving classrooms or traveling far from school[96]. By immersing students in authentic simulations, MUVEs and ARs can promote two deeper-learning strategies, apprenticeship-based learning and learning for transfer, which are very important in developing cognitive, intrapersonal, and interpersonal skills for the 21st century[97]. However, complex tasks in open-ended simulations and games cannot be adequately modeled using only classical test theory and item response theory[98]. More appropriate measurement models for open-ended simulations and games include Bayes nets, artificial neural networks, and model tracing; new psychometric methods beyond these will likely be needed as well.

ECOMUVE AS AN EXAMPLE OF IMMERSIVE AUTHENTIC SIMULATIONS IN MULTI-USER VIRTUAL ENVIRONMENTS

The EcoMUVE middle grades curriculum teaches scientific concepts about ecosystems while engaging students in both collaborative and individual scientific inquiry and helping them learn complex causality (http://ecomuve.gse.harvard.edu). The curriculum consists of two MUVE-based modules, allowing students to explore realistic, 3-dimensional pond and forest ecosystems. Each module consists of 10 45-minute lessons and includes a complex scenario in which ecological change is caused by the interplay of multiple factors[99]. Students assume the role of scientists, investigating research questions by exploring the virtual environment and collecting and analyzing data from a variety of sources over time (Figures 11,12). In the pond module, for example, students can explore the pond and the surrounding area, even venturing under the water; see realistic organisms in their natural habitats; and collect water, weather, and population data. Students visit the pond over a number of virtual days and eventually make the surprising discovery that, on a day in late summer, many fish in the pond have died. Students are then challenged to figure out what happened—they travel backward and forward in time to gather information to solve the mystery and to understand the complex causality of the pond ecosystem.



Dissolved Oxygen reading for the pond on July 16, 2009 is 10.3mg/L.

Figure 11. Students can collect pond and weather data

[96] Dede C. 2014. *The role of technology in deeper learning.* New York, NY: Jobs for the Future. http://www.studentsatthecenter.org/topics/role-digital-technologies-deeper-learning.

[97] National Research Council. 2012. *Education for life and work: Developing transferable knowledge and skills in the 21st century.* Washington, DC: The National Academies Press. http://www.nap.edu/catalog.php?record_id=13398

[98] Quellmalz ES, Timms, MJand Schneider SA. 2009. *Assessment of student learning in science, simulations, and games.* Paper prepared for the National Research Council Workshop on Gaming and Simulations. Washington, DC: National Research Council.

[99] Metcalf S, Kamarainen A, Grotzer Tand Dede C. 2013. Teacher perceptions of the practicality and effectiveness of immersive ecological simulations as classroom curricula. *International Journal of Virtual and Personal Learning Environments,* 4(3), 66-77.

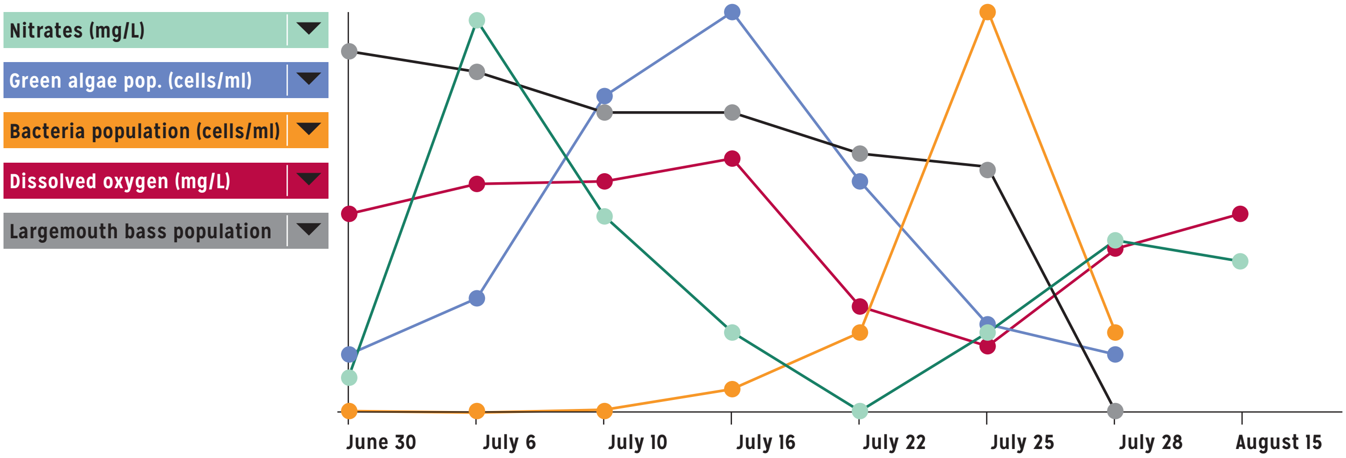


Figure 12. Summarizing and interpreting data

The EcoMUVE curriculum uses a "jigsaw" pedagogy in which students have access to differing information and experiences; they must combine their knowledge in order to understand what is causing the changes they see. Working in four-person teams, students are given roles that embody specific areas of expertise (naturalist, microscopic specialist, water chemist, and private investigator) and that influence how they participate and solve problems. Using the differing methods of their roles, students collect data, share it with teammates via tables and graphs that they create within the simulation, and then work collaboratively to analyze the combined data and figure out how a variety of interconnected parts come together to produce the larger ecosystem dynamics. The module culminates with each team creating an evidence-based concept map—representing their understanding of the causal relationships at work in the ecosystem—which they present to the class.

The types of big data about motivation and learning for each student that EcoMUVE can generate include: time-stamped logfiles of movements and interactions in the virtual world with artifacts, computer-based agents, data sources, guidance systems, and other students; chat-logs of utterances; and tables of data collected and shared. Other digital tools can provide data from concept maps that chart the flow of energy through the ecosystem and, for each team of students, document their group's assertions about its systemic causal relationships, with adduced supporting evidence. Using GoPro cameras, students' collaborative behaviors outside of digital media can be documented. Combined, these data are big in their collective volume, velocity, variety, and veracity. We would like to use this data to provide near real-time feedback to students and teachers, which requires various forms of visualization.

This guidance about instruction and learning could include "low-hanging fruit" types of feedback relatively easy to implement, such as:

Paths and heat maps. The paths that a student takes in exploring a virtual world to determine the contextual situation, identify anomalies, and collect data related to a hypothesis for the causes of an anomaly are an important predictor of the student's understanding of scientific inquiry. In our prior River City curriculum[100], we used logfile data to generate event paths (Figure

[100] Ketelhut D J, Nelson BC, Clarke J Eand Dede C. 2010. A multi-user virtual environment for building and assessing higher order inquiry skills in science. *British Journal of Educational Technology*, 41, 56–68.

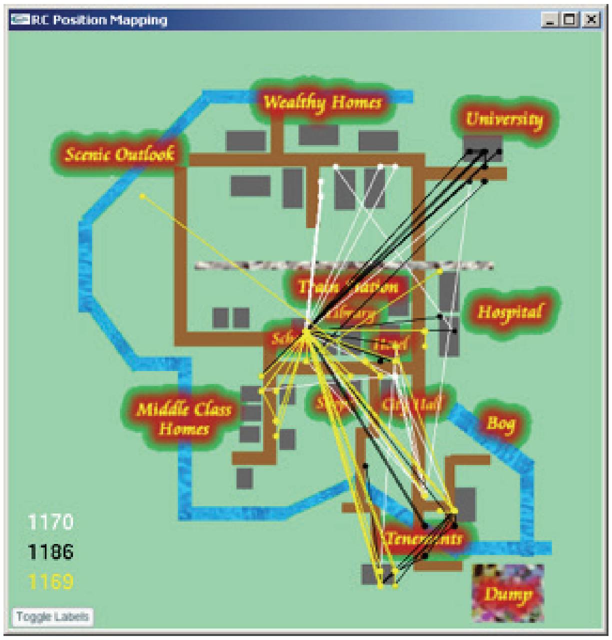


Figure 13. Event paths in RC for a three-person team

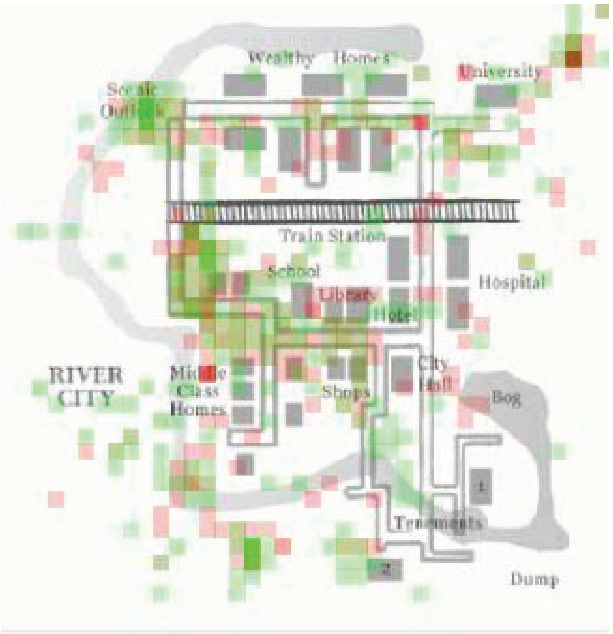


Figure 14. A heat map showing high-performing and low-performing students

13) for both individual students and their three-person teams. Students and teachers found this a useful source of diagnostic feedback on the relative exploratory skills—and degree of team collaboration—that these performances exhibited.

Dukas extended this research by developing an avatar log visualizer (ALV), which generates a series of slides depicting the relative frequency events of one or more subpopulations of students, aggregated by user-specified locations and time bins[101]. Figure 14 displays an ALV visualization that contrasts the search strategies of the high-performing and low-performing students in a class, displaying the top 10 scores on the content post-test (in green) and the lowest 10 scores (in pink).

The high-performing students' preferred locations provide an expert model usable in diagnostic feedback, formative about their search strategies, to students in subsequent classes. The low-performing students' locations may offer insights into what types of understanding they lack. Path analysis is a potentially powerful form of unobtrusive assessment, but choosing the best way to display student paths through a learning environment is a complex type of visualization not well understood at present. The utility of this diagnostic approach also depends on the degree to which exploration in the virtual world is an important component of learning.

Accessing an individualized guidance system. Nelson developed a version of River City that contained an interwoven individualized guidance system (IGS). The guidance system utilized personalized interaction histories collected on each student's activities to generate real-time, customized support[102]. The IGS offered reflective prompts about each student's learning in the world, with the content of the messages based on in-world events and basic event histories of that individual. For example, if a student were to click on the admissions chart in the River City hospital, a predefined rule stated that, if the student had previously visited the tenement district and talked to a resident there, then a customized guidance message would be shown

[101] Dukas G. 2009. *Characterizing student navigation in educational multiuser virtual environments: A case study using data from the River City project* (Unpublished doctoral dissertation). Harvard Graduate School of Education, Cambridge, MA.

[102] Nelson B. 2007. Exploring the use of individualized, reflective guidance in an educational multi-user virtual environment. *Journal of Science Education and Technology* 16(1), 83–97.

reminding the student that they had previously visited the tenement district, and asking the student how many patients listed on the chart came from that part of town.

Multilevel multiple regression analysis findings showed that use of this guidance system with our MUVE-based curriculum had a statistically significant, positive impact ($p < .05$) on student learning. In addition to using the logfiles to personalize the guidance provided to each student, we conducted analyses of guidance use. We knew if and when students first chose to use the guidance system, which messages they viewed, where they were in the virtual world when they viewed them, and what actions they took after viewing a guidance message. This data potentially provides diagnostic information that could guide instruction in immersive simulations.

Asking and answering questions of an agent. Animated pedagogical agents (APAs) are "lifelike autonomous characters [that] co-habit learning environments with students to create rich, face-to-face learning interactions"[103]. Beyond engaging students and providing a limited form of mentoring, APAs have two advantages for interwoven diagnostic assessment in a wide variety of learning environments: First, the questions students ask of an APA are themselves diagnostic; typically, learners will ask for information they do not know but see as having value. A single question asked by a student of an APA may reveal as much about what that learner does and does not know than a series of answers the student provides to a teacher's diagnostic questions. As an example, EcoMUVE and EcoMOBILE can embed APAs of various types for eliciting a query trajectory over time that reveals aspects of students' understanding and motivation, as well as aiding learning and engagement by the APA's responses.

Second, APAs scattered through an immersive authentic simulation can draw out student performances in various ways. As an illustration, in EcoMUVE and EcoMOBILE, a student can meet an APA who requests the student's name and role. Even a simple pattern recognition system could determine if the student made a response indicating self-efficacy and motivation ("ecosystems scientist" or some variant) versus a response indicating a lack of confidence or engagement ("sixth grader" or some other out-of-character reply). As another example, an APA can request a student to summarize what the student has found so far, and some form of latent semantic analysis could scan the response for key phrases indicating the student's understanding of terminology and relevant concepts. The design heuristics of this method for evoking performances are that (a) the interaction is consistent with the overall narrative, so it is not disruptive of flow, (b) the measurement is relatively unobtrusive, and (c) the interactions themselves deepen immersion.

But what about more complex types of feedback based on big data less easily analyzed? As examples, teachers and researchers would benefit from analyses of aggregated data that delineated learning trajectories of sophisticated skills (e.g., causal reasoning) in relation to which individual students' progress could be diagnostically assessed. In turn, students would benefit from multi-modal data analysis that could be used to alter, in real time, the context and activities of the immersive simulation to make salient what each student needs to understand next in their learning trajectory. Further, over a series of learning experiences, students' growth in intrapersonal and interpersonal skills (e.g., engagement, self-efficacy, tenacity, and collaboration) could be assessed. These functionalities are well beyond current capabilities, but are aspirational within the next decade.

EcoMOBILE as an Example of Augmented Realities

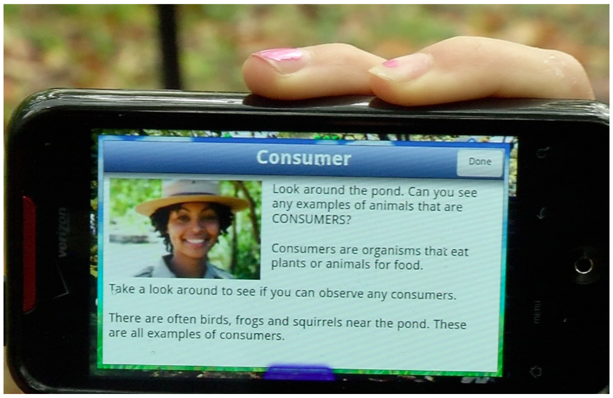
Designed to complement EcoMUVE, the EcoMOBILE project explores the potential of augmented reality (as well as the use of data collection "probeware," such as a digital tool that measures the amount of dissolved oxygen in water), to support learning in environmental

[103] Johnson W L, Rickel JWand Lester JC. 2000. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11, 47–78.

science education (http://ecomobile.gse.harvard.edu). The EcoMOBILE curriculum is a blend of the EcoMUVE learning experiences with the use of geolocated digital experiences that enhance students' real-world activities[104]. As an example of a three-day curriculum, during the first class period, a group of middle school students participated in an EcoMUVE learning quest, completing a 5–10 minute online simulation in which they learned about dissolved oxygen, turbidity, and pH. The following day, the students went on a field trip to a nearby pond, in order to study the relationship between biological and non-biological factors in the ecosystem, practice data collection and interpretation, and learn about the functional roles (producer, consumer, decomposer) of organisms in the life of the pond. At a number of spots around the pond, students' handheld devices showed them visual representations—overlaid onto the real environment—of the natural processes at work in the real environment, as well as interactive media including relevant text, images, audio, video, 3D models, and multiple-choice and open-ended questions. Students also collected water measurements using Vernier probes (Figures 15 and 16).

On the next school day after the field trip, back in the classroom, students compiled all of the measurements of temperature, dissolved oxygen, pH, and turbidity that had been taken during the excursion. They looked at the range, mean, and variations in the measurements and discussed the implications for whether the pond was healthy for fish and other organisms. They talked about potential reasons why variation may have occurred, how these measurements may have been affected by environmental conditions, and how to explain outliers in the data. Our research shows that virtual worlds and augmented realities are powerful complements to enable learning partnerships for real-world, authentic tasks.

Parallel to EcoMUVE, EcoMOBILE devices capture and store big data about motivation and learning for each student, which includes time-stamped logfiles of paths through the real world and data collected in that ecosystem (e.g., images, sound files, and probeware), as well as geolocated interactions with digital augmentations (e.g., simulations, guidance systems, and assessments). Using GoPro cameras, students' collaborative behaviors outside of digital media can be documented. Other digital tools can provide data from concept maps charting the flow of energy through the ecosystem and, for each team of students, documenting their group's assertions about its systemic causal relationships, with adduced supporting evidence. As with EcoMUVE, the combination of these data could support rich types of feedback to students, teachers, and researchers.



Look around the pond. Can you see any examples of animals that are CONSUMERS?

Consumers are organisms that eat plants or animals for food.

Take a look around to see if you can observe any consumers.

There are often birds, frogs and squirrels near the pond. These are all examples of consumers.

Figure 15. Handheld device delivering information

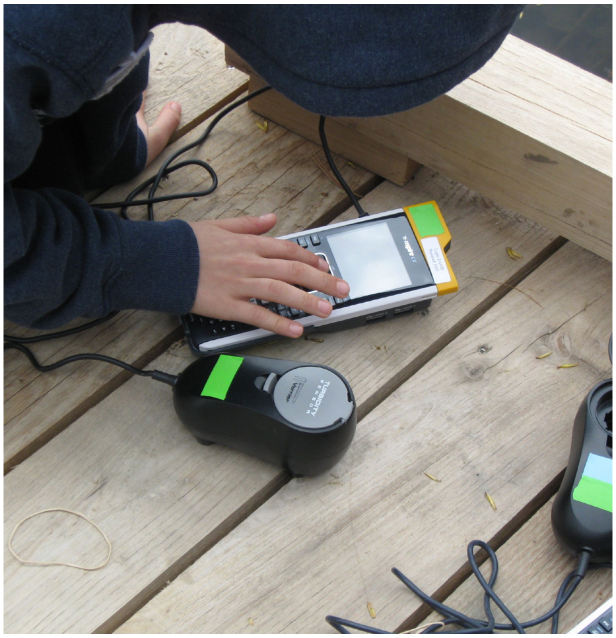


Figure 16. Collecting water data on turbidity

[104] Kamarainen A, Metcalf S, Grotzer T, Browne A, Mazzuca D, Tutwiler Sand Dede C. 2013. EcoMOBILE: Integrating augmented reality and probeware with environmental education field trips. Computers & Education, 68 545-556.

THE DATA SCIENCE CHALLENGE

Quellmalz, Timms, and Schneider (2009) examined issues of embedding assessments into games and simulations in science education[105]. Their analysis included both tightly structured and open-ended learning experiences. After studying several immersive games and simulations related to learning science, including River City, they noted that the complex tasks in simulations and games cannot be adequately modeled using only classical test theory and item response theory. This shortfall arises because these complex tasks have four characteristics[106]:

1. Completion of the task requires the student to undergo multiple, nontrivial, domain-relevant steps and/or cognitive processes.
2. Multiple elements, or features, of each task performance are captured and considered in the determination of summaries of ability and/or diagnostic feedback.
3. The data vectors for each task have a high degree of potential variability, reflecting relatively unconstrained work product production.
4. Evaluation of the adequacy of task solutions requires the task features to be considered as an interdependent set, for which assumptions of conditional independence do not hold.

Quellmalz et al. (2009) concluded that, given the challenges of complex tasks, more appropriate measurement models for simulations and games—particularly those that are open ended—include Bayes nets, artificial neural networks, and model tracing. They added that new psychometric methods beyond these will likely be needed. Beal and Stevens (2007) used various types of probabilistic models in studying students' performance in simulations of scientific problem solving[107]. Bennett, Persky, Weiss, and Jenkins (2010) described both progress in applying probabilistic models and the very difficult challenges involved[108]. Behrens, Frezzo, Mislevy, Kroopnick, and Wise (2007) described ways of embedding assessments into structured simulations; and Shute, Ventura, Bauer, and Zapata-Rivera (2009) delineated a framework for incorporating stealth assessments into games[109] [110].

In summary, games and immersive learning experiences can collect an impressive array of evidence about what a learner knows (and does not know), what he or she can do (and cannot do), and whether he or she knows when and how to apply disciplinary frames and prior knowledge to a novel problem. Immersive environments—because of their situated nature and because they generate logfiles—make it possible to elicit performances, to collect continuous data, and to interpret structures of evidence. In a virtual world, the server documents and timestamps actions by each student, including movements, interactions, utterances, saved data, and so on. In an AR, the mobile device can save moderately detailed information about movements and actions, and using GoPro cameras to record learners' visual perspectives and verbal utterances as

[105] Quellmalz E S, Timms M Jand Schneider SA. 2009. *Assessment of student learning in science, simulations, and games.* Paper prepared for the National Research Council Workshop on Gaming and Simulations. Washington, DC: National Research Council.

[106] Williamson D M, Bejar I land Mislevy R J. 2006. *Automated scoring of complex tasks in computer-based testing. Mahwah, NJ: Erlbaum.*

[107] Beal C Rand Stevens RH. 2007. Student motivation and performance in scientific problem solving. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education: Building technology rich learning contexts that work* (pp. 539–541). Amsterdam, Netherlands: IOS Press.

[108] Bennett RE, Persky H, Weiss Aand Jenkins F. 2010.Measuring problem solving with technology: A demonstration study for NAEP. *Journal of Technology, Learning, and Assessment,* 8(8), 1–45.

[109] Behrens JT, Frezzo D, Mislevy R, Kroopnick Mand Wise D. 2007. Structural, functional, and semiotic symmetries in simulation-based games and assessments. In E. Baker, J. Dickieson, W. Wulfeck, & H. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 59–80). Mahwah, NJ: Lawrence Erlbaum Associates.

[110] Shute VJ, Ventura M, Bauer M Iand Zapata-Rivera D. 2009. Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295-321). Mahwah, NJ: Routledge, Taylor and Francis.

their team interacts can provide another resource for analysis. Given the engagement, evocation, and evidence that immersive learning provides, these media are among the most powerful and valid instructional and assessment experiences available—but we can realize their full potential only via new methods for collecting, analyzing, and communicating findings from complex types of big data.

BERLAND: CREATING CREATIVE DATA SCIENTISTS

Building capacity in data science and data engineering is vital for attaining the potential of data-intensive research in education. The Learning Games Play Data Consortium (PDC) is made of game designers, the education industry, learning scientists, computer scientists, data scientists, students, and startups. The core mission of the PDC is to bring people together to facilitate collaboration and to advance understanding on how to create the next generation of data-driven learning games, learning theories, and learning tools. The PDC has developed and maintains several tools to help people implement advanced data analysis for learning in game design, research, and industry. We have currently identified six imminent challenges in data science for learning and education: training learning data scientists, building analytic tools, developing learning theory for design, designing new models of assessment, visualizing learning data, and innovating curricula.

Training a diverse and wide-ranging set of designers and researchers to think about new possibilities with data is difficult. Training people to think creatively about the possibilities for data in education turns out to be really difficult, and there are few good examples of how to do it. Imminent possibilities, given well trained people, include: pioneering new modes of assessment; new tools for teachers, students, and administrators; innovating design for games and learning environments; and new understandings of how people learn and how schools work. Those possibilities lie at the overlap of several disciplines (computer science, statistics, education, design, and information studies), but creative data-driven design thinking is not necessarily the purview of any single discipline. This makes the problem even harder; the data scientists trained by, say, industry, startups, or computer science tend (quite reasonably) to hew closely to their missions.

When examining the landscape of what is possible in education, few people are using data to design or create new things that were not possible before modern data science. Both industry and academia (the author of this text included) often use data to reify and reinforce classical ways of doing things, but it seems likely that the "killer apps" of data-driven learning are not going to come from deeper investment in, say, computer-based tests. As data-driven thinking is democratized, those models will likely seem more problematic to learners.

Part of the problem with training is that we know relatively little about creative data analysis and visualization toward creating educational learning environments. Learning sciences–the author's home academic field and a place in which novel work is being done–is small and only modestly funded, but computer scientists (another group in which the author considers himself a member) are usually not trained in how people learn and tend to replicate traditionalist models of education, thinking, and learning while vastly improving models of how to work with a lot of data. Information studies, design schools, arts, journalism, and applied mathematics have pioneered new ways of visualizing data, but they frequently lack training in either computer science or learning sciences. In short, very few people are training students (and faculty) to consider new modes of how to understand, visualize, and change how people learn and how education works.

It is simpler and cheaper to reproduce classical modes of education with big data than it is to develop new modes of giving learners agency. As a result, many of the data scientists in education are being trained to do very careful large-scale analyses of inherently problematic assessments. A scenario in which schools are optimized to produce the most available and easily parsable data would presumably result in a situation worse than the testing-driven model we are seeing now: it can (and may) become a model in which students are constantly tested, evaluated, and all opportunities to productively fail (in other words, learn) are eliminated. This is a real, if dystopian, possibility. I have heard many successful friends and colleagues say something to the effect of "I do not know how I would have learned anything if Twitter had been around when I was learning–I would hate to have all of my mistakes archived forever." When all mistakes are evaluated, people are more afraid to make mistakes.

That said, teenagers read and write more than they did before social media, they make fewer grammatical mistakes, and they "connect" with many more people[111]. The utopian promise of data in education is that students will be able to learn from their mistakes in real-time and authentic situations. Social media provides instant feedback–it is a novel mode of "big data analysis"–and one of the most salient introductions to data-driven learning comes from the kinds of simple analytics that Twitter, Facebook, and Google Analytics give to people. People like creating, and they would like to use the data they create to better understand their world. By giving the data back to people, we will be making people happy, helping them learn more quickly, and creating the next generation of data scientists.

The author's group at the University of Wisconsin-Madison, together with our many wonderful colleagues across the U.S., has attempted to do this in a few ways. One way is by developing tools through which the creation of data collection and analytics can be open to a much wider group of game designers and people designing creative learning environments. For instance with ADAGE (2014; 2015), we have developed a widely used, free, open source platform for collecting and analyzing learning data from games. We have also been developing ways to look at many different forms of data through our multi-modal PDC Dashboard. Our view is that learning data looks fundamentally different from the types of data that people look at in most data dashboards, that learning data happens over time, and outliers should be focused on and explored rather than ignored[112]. We also have several games in which data analytics are used to inform students and teachers about what is happening in their classrooms and understand why students are successful[113] [114] [115]. In all of these games, it is important that some element of analysis is structured and driven by the teachers and students themselves.

The group has learned several factors of successful data-driven learning: students love analyzing data about themselves; teachers understand better than we do when data would be helpful for teaching; and using advanced data analytics on constructive, creative learning environments is both possible and not nearly as hard as we had thought. In short, we learned that training novice data scientists through real constructive work–as researchers on my team, designers on my team, teachers we work with, and students themselves–is not only possible, but can enjoyable for all parties. We have found that people become deeply engaged and understand complex data analytic content more fully when they are deeply connected to that content. From there, it is possible for both learners and researchers to think differently about data by connecting and visualizing many different modes of those data, such as transcripts, game play, pre- and post-tests, and more longitudinal data. Those connections to both the data and across different types and modes of data seem essential to more deeply understanding learning trajectories.

[111] Boyd D. 2007. Why youth ♥ social network sites: The role of networked publics in teenage social life. *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning*, 119–142.

[1112] Berland M, Bake, Rand Blikstein P. (2014). Educational Data Mining and Learning Analytics: Applications to Constructionist Research. *Technology, Knowledge and Learning*, 1–16. http://doi.org/10.1007/s10758-014-9223-7.

[113] Berland M, Martin T, Benton T, Petrick Smith Cand Davis D. 2013. Using Learning Analytics to Understand the Learning Pathways of Novice Programmers. *Journal of the Learning Sciences, 22*(4), 564–599. http://doi.org/10.1080/10508406.2013.836655.

[114] Berland M, Smith C Pand Davis D. 2013. Visualizing Live Collaboration in the Classroom with AMOEBA. In *Proceedings of the International Conference on Computer-Supported Collaborative Learning*.

[115] Berland Mand Wilensky U. 2015. Comparing Virtual and Physical Robotics Environments for Supporting Complex Systems and Computational Thinking. *Journal of Science Education and Technology*, 1–20. http://doi.org/10.1007/s10956-015-9552-x.

Some recommendations for supporting the growth of data analytics to learning: 1) bring interested, diverse novices into your groups and let them be wrong; and 2) build tools that help students understand how they are creating (think: Twitter) rather than evaluating them post-hoc (think: standardized testing). Novices will frequently have terrible, unimplementable ideas, and the process will be horribly inefficient, but it will lead to a better solution. In artificial intelligence, this is how many optimization algorithms, such as simulated annealing, work, not by evaluating every possible branch forever but by finding pathways around and through local maxima. We are all stuck in our local maxima; we are all hindered by the activation energy to make big changes. To find new spaces in which to grow, we have to listen to what novices say when they are most totally wrong: What do they want to say? What information do they think might help them? Leverage their misunderstanding to reshape your own understanding, and teach them to use data to understand how they learn. By training new people to think creatively with data, you will be exposed to new ways of thinking by people who might use those data.

SHUTE: A VISION OF THE FUTURE OF ASSESSMENT

Imagine an educational system where high-stakes tests are no longer used. Instead, students would progress through their school years engaged in different learning contexts, all of which capture, measure, and support growth in valuable cognitive and noncognitive skills. This is conceivable because, in our complex, interconnected, digital world, we're all producing numerous digital footprints daily. This vision thus involves 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).

The primary goal for this vision of assessment is to improve learning, particularly learning outcomes and processes necessary for students to succeed in the 21st century[116] [117]. Most current approaches to assessment and testing are too disconnected from learning processes. That is, the typical classroom cycle is: Teach. Stop. Administer test. Go loop (with new content). But consider the following metaphor representing an important shift that occurred in the world of retail outlets (from small businesses to large department stores), suggested by Pellegrino, Chudhowsky, and Glaser[118]. No longer do these businesses have to close down once or twice a year to take inventory of their stock. Instead, with the advent of automated checkout and barcodes for all items, these businesses have access to a continuous stream of information that can be used to monitor inventory and the flow of items. Not only can businesses continue without interruption, but the obtained information is far richer, enabling stores to monitor trends and aggregate the data into various kinds of summaries, as well as support real-time, just-in-time inventory management. Similarly, with new assessment technologies, schools should no longer have to interrupt the normal instructional process at various times during the year to administer external tests to students. Instead, assessment should be continual and invisible to students, supporting real-time, just-in-time instruction, and other types of learning support.

The envisioned ubiquitous nature of assessment will require a reconceptualization on the boundaries of the educational system. That is, the traditional way of teaching in classrooms today involves providing lectures and giving tests in class, then assigning homework to

[116] Black Pand Wiliam D. 1998. Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice, 5*(1), 7-74.

[117] Shute V J. 2009. Simply assessment. *International Journal of Learning, and Media, 1*(2), 1-11.

[118] Pellegrino JW, Chudowsky Nand Glaser R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

students to complete outside of class (usually more reading on the topic and perhaps answering some topical questions). Alternatively, consider a relatively new pedagogical approach called "flipped classrooms." This involves a reversal of the traditional approach where students first examine and interact with a target topic by themselves at home and at their leisure (e.g., viewing an online video and/or playing an educational game); and then in class, students apply the new knowledge and skills by solving problems and doing practical work[119]. The flipped classroom is already operational for core courses at some schools and universities across North America. The teacher supports the students in class when they become stuck, rather than delivering the initial lesson in person. Flipped classrooms free class time for hands-on work and discussion, and permit deep dives into the content. Students learn by doing and asking questions, and they can also help each other, a process that benefits a majority of learners[120].

CHALLENGES AND FUTURE RESEARCH

For this vision of the future of assessment—as ubiquitous, unobtrusive, engaging, and valid—to gain traction, there are a number of large hurdles to overcome. Following are four of the more pressing issues that need more research.

QUALITY OF ASSESSMENTS

The first hurdle relates to variability in the quality of assessments within TREs. That is, because schools are under local control, students in a given state could engage in thousands of TREs during their educational tenure. Teachers, publishers, researchers, and others will be developing TREs, but with no standards in place, they will inevitably differ in curricular coverage, difficulty of the material, scenarios and formats, and many other ways that will affect the adequacy of the TRE, tasks, and inferences on knowledge and skill acquisition that can justifiably be made from successfully completing the TREs. Assessment design frameworks (e.g., ECD[121], Assessment Engineering[122]) represent a design methodology but not a panacea, so more research is needed to figure out how to equate TREs or create common measurements (i.e., standardized) from diverse environments. Toward that end, there must be common models employed across different activities, curricula, and contexts. Moreover, it is important to figure out how to interpret evidence where the activities may be the same but the contexts in which students are working are different (e.g., working alone vs. working with another student).

INTERPRETING DIFFERENT LEARNING PROGRESSIONS

The second hurdle involves accurately capturing and making sense of students' learning progressions. That is, while TREs can provide a greater variety of learning situations than traditional face-to-face classroom learning, evidence for assessing and tracking learning progressions becomes heterogeneous and complex rather than general across individual students. Thus, there is a great need to model learning progressions in multiple aspects of student growth and experiences, which can be applied across different learning activities and contexts[123]. However, as Shavelson and Kurpius point out, there is no single absolute order of progression as learning in TREs involves multiple interactions between individual students and situations, which may be too

[119] Bergmann Jand Sams A. 2012. *Flip your classroom: Reach every student in every class every day.* International Society for Technology in Education (ISTE).

[120] Strayer J. 2012. How learning in an inverted classroom influences cooperation, innovation and task Orientation. *Learning Environments Research, 15*(2), 171-193.

[121] Mislevy R J, Steinberg L Sand Almond R G. 2003. On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

[122] Luecht RM. 2013. An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 59-76). New York: Routledge.

[123] Shavelson R Jand Kurpius A. 2012. Reflections on learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 13-26). Rotterdam, the Netherlands: Sense Publishers.

complex for most measurement theories in use that assume linearity and independence. Clearly, theories of learning progressions in TREs need to be actively researched and validated to realize TREs' potential.

EXPANDED EDUCATIONAL BOUNDARIES

The third problem to resolve involves impediments to moving toward the idea of new contexts of learning (e.g., flipped classrooms). One issue concerns the digital divide where some students may not have access to a home computer. In those cases, students can be allowed to use library resources or a computer lab. Alternatively, online components can be accessed via a cell phone as many students who do not have computers or Internet access at home do have a phone that can meet the requirements of online activities. In addition, some critics argue that flipped classrooms will invariably lead to teachers becoming outdated. However, teachers become even more important in flipped classrooms, where they educate and support rather than lecture (i.e., "guide on the side" rather than "sage on a stage"). This represents an intriguing way to take back some of the very valuable classroom time, and serve as a more efficient and effective teacher. Much more empirical research is needed to determine how this pedagogical approach works relative to traditional pedagogies.

PRIVACY AND SECURITY

The fourth hurdle involves figuring out a way to resolve privacy, security, and ownership issues regarding students' information. The privacy and security issues relate to the accumulation of student data from disparate sources. The recent failure of the $100 million inBloom initiative showcases the problem[124]. That is, the main aim of inBloom was to store, clean, and aggregate a wide range of student information for states and districts, and then make the data available to district-approved third parties to develop tools and dashboards so the data could be easily used by classroom educators. The main issue boils down to this: Information about individual students may be at risk of being shared far more broadly than is justifiable. And because of the often high-stakes consequences associated with tests, many parents and other stakeholders fear that the data collected could later be used against the students.

What would it take to implement the vision once the hurdles are surmounted? I'll use ECD to illustrate. In addition to ECD's ability to handle multivariate competency models, it is able to accumulate evidence across disparate sources (e.g., homework assignment, in-class quiz on an iPad, and a high score on a video game). This is possible as ECD provides assessment designers with processes that enable them to work through the design trade-offs that involve multiple competency variables—either within one assessment or across multiple assessments. The "alchemy" involves turning the raw data from various sources into evidence. Evidence models will need to be able to interpret the results of all incoming data for the purposes of updating the student model. The rules of evidence must describe which results can be used as evidence, as well as any transformation that needs to be done to those results (e.g., averaging, rescaling, setting cut scores)[125]. As sufficient data (i.e., outcomes from students' interactions with a collection of tasks) become available, Bayesian inference can be used to replace the prior distributions for parameters with posterior distributions. This should improve the quality of inferences that come from the system.

Despite the foregoing hurdles, constructing the envisioned ubiquitous and unobtrusive assessments across multiple learner dimensions, with data accessible by diverse stakeholders, could yield various educational benefits. First, the time spent administering tests, handling make-up exams, and reviewing test responses is not very conducive to learning. Given the importance of time on task as a predictor of learning, reallocating those test-preparation activities into ones that are

[124] McCambridge R. 2014. Legacy of a failed foundation initiative: inBloom, Gates and Carnegie. In Nonprofit Quarterly, Retrieved from https://nonprofitquarterly.org/policysocial-context/24452-legacy-of-a-failed-foundation-initiative-inbloom-gates-and-carnegie.html.

[125] Almond R G. 2010. Using Evidence Centered Design to think about assessments. In V. J. Shute, & B. J. Becker. (Eds.), *Innovative assessment for the 21st Century: Supporting educational needs* (pp. 75-100). New York: Springer-Verlag.

more educationally productive would provide potentially large benefits to almost all students. Second, by having assessments that are continuous and ubiquitous, students are no longer able to "cram" for an exam. Although cramming can provide good short-term recall, it is a poor route to long-term retention and transfer of learning. Standard assessment practices in school can lead to assessing students in a manner that is in conflict with their long-term success. With a continuous assessment model in place, the best way for students to do well is to do well every day. The third direct benefit is that this shift in assessment mirrors the national shift toward evaluating students on the basis of acquired competencies. With increasing numbers of educators growing wary of pencil and paper, high-stakes tests for students, this shift toward ensuring students have acquired "essential" skills fits with the idea of my envisioned future of assessment.

The time is now ripe for such assessments given the dire need for supporting new 21st century skills and the increased availability of computer technology. New technologies make it easy to capture the results of routine student work—in class, at home, or wherever. It could be that 21st century assessment will be so well integrated into students' day-to-day lives that they don't even know it's there. This represents quite a contrast to our current testing contexts. However, while the benefits of using a seamless-and-ubiquitous model to run a business have been clear for more than four decades, applying this metaphor to education may require adjustments as we are dealing with humans, not goods. For instance, one risk associated with the vision is that students may come to feel like they are constantly being evaluated, which could negatively affect their learning and possibly add stress to their lives. Another risk of a continuous assessment vision could result in teaching and learning turning into ways to "game the system" depending on how it is implemented and communicated. But the aforementioned hurdles and risks, being anticipated and researched in advance, can help shape the vision for a richer, deeper, and more authentic assessment to support the learning of students in the future. How many current businesses would elect to return to pre-barcode days?

Collaborating on Tools, Infrastructures, and Repositories

Rick Gilmore (Databrary), Edith Gummer (EdWise), and Ken Koedinger (LearnLab Datashop)

THE OPPORTUNITY DATA REPOSITORIES PROVIDE FOR DATA-INTENSIVE RESEARCH IN EDUCATION

Open data sharing can help to translate insights from scientific research into applications serving essential human needs. Open data sharing bolsters 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. Data sharing and reuse increases the impact of public investments in research and leads to more effective public policy. Although many researchers in the developmental, learning, and education sciences collect video as raw research data, most research on human learning and development remains shrouded in a culture of isolation[126]. 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.

The National Science Foundation's *Ideas Lab to Foster Transformative Approaches to Teaching and Learning* was an activity intended to bring together a range of STEM education developers and researchers to think about how large data sets might be leveraged to improve STEM teaching and learning. The central premise of data in the announcement was that new advances in data analysis, coupled with rich and complex data systems, would enable us to develop and study new formal and informal learning environments. The focus on data in the announcement was deliberately quite wide.

These new approaches will require the generation and use of data that range from micro-level data on individual learners, to data from online learning sources (such as massively open online courses), to meso-level data from the classroom that provide information to students and teachers about how learning is progressing, to macro-level data such as school, district, state, and national data, including data from federal science and policy agencies.

The Databrary and LearnLaB cases below describe the value of repositories for micro-level data, and the EdWise tool case illustrates mining repositories for meso- and macro-level data.

EXAMPLES OF STATE-OF-THE-ART WORK IN DATA REPOSITORIES

GILMORE: THE DATABRARY REPOSITORY

Video is a uniquely rich, inexpensive, and adaptable medium for capturing the complex dynamics of behavior. Researchers use video in home and laboratory contexts to study how infants, children, and adults behave in natural or experimenter-imposed tasks[127]. Researchers record videos of students in classrooms to understand what teachers do and how students respond[128]. Because video closely mimics the multisensory experiences of live human observers, recordings collected by one person for a particular purpose may be readily understood by another person and reused for a different purpose. Moreover, the su ccess of YouTube and other video-based social media demonstrates that web-based video storage and streaming systems are now sufficiently well developed to satisfy large-scale demand. The question for researchers and policymakers

[126] Adolph KE, Gilmore R O, Freeman C, Sanderson Pand Millman D. 2012. Toward open behavioral science. *Psychological Inquiry, 23*(3), 244–247.doi:10.1080/1047840X.2012.705133 .

[127] Karasik L B, Tamis-LeMonda C Sand Adolph KE. 2014. Crawling and walking infants elicit different verbal responses from mothers. *Developmental Science, 17* (3), 388–395. doi: 10.1111/desc.1212.

[128] Alibali MWand Nathan M J. 2012. Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the Learning Sciences, 21* (2), 247-286. doi: doi:10.1080/10508406.2011.611446.

is how to capitalize on video's potential to improve teaching and learning.

Imagine a time in the near future when researchers interested in studying classroom teaching and learning can mine an integrated, synchronized, interoperable, open and widely shared data set. The components include video from multiple cameras, eye tracking, motion, and physiological measurements, and information from both historical and real-time student performance measures. Imagine that this classroom-level data can be linked with grade, school, neighborhood, community, region, and state-level data about education practice, curriculum, and policy. Then, imagine training a cadre of experts with skills in the data science of learning and education who are sensitive to privacy, confidentiality, and ethical issues involved in research with identifiable information. We empower these learning scientists to extract meaningful insights from the data about how educational practice and policy might be improved. In short, imagine a science of teaching and learning that can be personally tailored to individuals in ways analogous to the impact of big data on medicine. The barriers to realizing this vision are similar to those that confront the vision of personalized medicine: the development of technologies that enable data to be collected, synchronized, tagged, curated, stored, shared, linked, and aggregated; policies and practices that ensure security and individual privacy; and the cultivation of professional expertise needed to turn raw data into actionable insights.

CHALLENGES OF SHARING RESEARCH VIDEO DATA

There are significant technical, ethical, practical, and cultural challenges to sharing research video. File sizes and diverse formats present special challenges for sharing. Video files are large (one hour of HD video can consume 10+ GB of storage) and come in varied formats (from cell phones to high-speed video). Many studies require multiple camera views to capture desired behaviors. Research video creates a data explosion: A typical lab studying infant or child development collects 8-12 hours of video per week[129]. Thus, sharing videos requires substantial storage capacity and significant computational resources for transcoding videos into common, preservable formats.

Technical challenges involved in searching the contents of videos present barriers to sharing. Videos contain rich and diverse information that requires significant effort by human observers to extract. Researchers make use of videos by watching them and, using paper and pencil or more automated computerized coding software, translating observations into ideas and numbers. In many cases, researchers assign codes to particular portions of videos. These codes make the contents of videos searchable by others, in principle. However, researchers focus on different questions from varied theoretical perspectives and lack consensus on conceptual ontologies. So, in practice, most coded data are not easily shared. Although human-centered video coding capitalizes on the unique abilities of trained observers to capture important dimensions of behavior, machine learning and computer vision tools may provide new avenues for tagging the contents of videos for educational and developmental research[130 131 132 133 134].

Open video sharing must overcome ethical challenges linked to sharing personally identifiable data. Although policies exist for sharing de-identified data, video contains easily identifiable data: faces, voices, names,

[129] Gilmore R Oand Adolph K E. 2012. Video Use Survey of ICIS and CDS listserv members.

[130] Amso D, Haas S, Tenenbaum E, Markant Jand Sheinkopf S. (2014). Bottom-up attention orienting in young children with autism. *Journal of Autism and Developmental Disorders, 44* (3), 664-673. doi: 10.1007/s10803-013-1925-5.

[131] Yu Cand Smith LB. 2013, 11. Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLoS ONE, 8* (11), e79659. doi: 10.1371/journal.pone.007.

[132] Fathi A, Hodgins Jand Rehg, J. 2012, June. Social interactions: A first-person perspective. In 2012 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 1226-1233). doi: 10.1109/CVPR.2012.6247805.

[133] Google Research. (2014). A picture is worth a thousand (coherent) words: Building a natural description of images. Retrieved 2015-05-08, from http://bit.ly/1wTMbk7.

[134] Raudies Fand Gilmore RO. (2014). Visual motion priors differ for infants and mothers. *Neural Computation, 26* (11), 2652–2668.

interiors of homes and classrooms, and so on. Removing identifiable information from video severely diminishes its reuse value and poses additional burdens on researchers. So, open video sharing requires new policies that protect the privacy of research participants while preserving the integrity of raw video for reuse by others.

Open video sharing faces practical challenges of data management. Developmental and education research is inundated by an explosion of data, most of which is inaccessible to other researchers. Researchers lack time to find, label, clean, organize, and copy their files into formats that can be used and understood by others[135]. Study designs vary widely, and no two labs manage data in the same way. Idiosyncratic terms, record-keeping, and data management practices are the norm. Few researchers document workflows or data provenance. Although video requires minimal metadata to be useful, video files must be electronically linked to what relevant metadata exist, including information whether participants have given permission to share.

Perhaps the most important challenge is cultural–community practices must change. Most researchers in the education, learning, and developmental sciences do not reuse their own videos or videos collected by other researchers; they neither recognize nor endorse the value of open sharing. Contributing data is anathema and justifications against sharing are many. Researchers cite intellectual property and privacy issues, the lack of data-sharing requirements from funding agencies, and fears about the misuse, misinterpretation, or professional harm that might come from sharing[136] [137]. Data sharing diverts energy and resources from scholarly activities that are more heavily and frequently rewarded. These barriers must be overcome to make data sharing a scientific norm.

DATABRARY.ORG

The Databrary project has built a digital data library (http://databrary.org) specialized for open sharing of research videos. Databrary has overcome the most significant barriers to sharing video, including solutions to maintaining participant privacy, storing, streaming, and sharing video, and for managing video data sets and associated metadata. Databrary's technology and policies lay the groundwork for securely sharing research videos on teaching and learning. In only a year of operation, Databrary has collected more than 7,000 individual videos, representing 2,400 hours of recording, and featuring more than 1,800 infant, child, and adult participants. Databrary has more than 130 authorized researchers representing more than 75 institutions across the globe. Video data is big data, and the interest in recording and sharing video for research, education, and policy purposes continues to grow.

The Databrary project arose to meet the challenges of sharing research video and to deliver on the promise of open data sharing in educational and developmental science. With funding from NSF (BCS-1238599) and NIH (NICHD U01-HD-076595), Databrary has focused on building a data library specialized for video, creating data management tools, crafting new policies that enable video sharing, and fostering a community of researchers who embrace video sharing. Databrary also developed a free, open-source video annotation tool, Datavyu (http://datavyu.org). The project received funding in 2012-2013, began a private beta testing phase in the spring of 2014, and opened for public use in October 2014.

SYSTEM DESIGN

The Databrary system enables large numbers of video and related files to be uploaded, converted, organized, stored, streamed, and tagged. Databrary is a free, open-source (http://github.com/databrary) web application

[135] Ascoli G A. 2006a. Mobilizing the base of neuroscience data: the case of neuronal morphologies. *Nature Reviews Neuroscience*, *7* (4), 318–324. doi: 10.1038/nrn1885.

[136] Ascoli G A. 2006b. The ups and downs of neuroscience shares. *Neuroinformatics*, *4* (3), 213-215. doi:10.1385/NI:4:3:213.

[137] Ferguson L. 2014. How and why researchers share data (and why they don't). Retrieved from http://bit.ly/1A5mmEW.

whose data are preserved indefinitely in a secure storage facility at New York University. Databrary can house video and audio files, along with associated materials, coding spreadsheets, and metadata. Video and audio data are transcoded into standard and HTML5-compatible formats. This ensures that video data can be streamed and downloaded by any operating system that supports a modern browser. Copies of original video files are also stored. Databrary stores other data in their original formats (e.g., .doc, .docx, .xls, .xlsx, .txt, .csv, .pdf, .jpg, and .png).

The system's data model embodies flexibility. Researchers organize their materials by acquisition date and time into structures called sessions. A session corresponds to a unique recording episode featuring specific participants. It contains one or more videos and other file types and may be linked to user-defined metadata about the participants, tasks or measures, and locations. A group of sessions is called a volume. Databrary contributors may combine sessions or segments with coding manuals, coding spreadsheets, statistical analyses, questionnaires, IRB documents, computer code, sample displays, and links to published journal articles.

Databrary does not enforce strict ontologies for tagging volumes, sessions, or the contents of videos. Video data are so rich and complex that in many domains, researchers have not settled on standard definitions for particular behaviors and may have little current need for standardized tasks, procedures, or terminology. Indeed, standardized ontologies are not necessary for many use cases. Databrary empowers users to add keyword tags and to select terms that have been suggested by others without being confined to the suggestions. Moreover, Databrary encourages its user communities to converge on common conceptual and metadata ontologies based on the most common keyword tags, and to construct and enforce common procedures and tasks wherever this makes sense.

Future challenges include enhancing the capacity to search for tagged segments inside videos. Some search functionality exists in the current software, with more extensive capabilities on the near horizon. A related challenge involves importing files from desktop video coding tools. This will allow for the visualization of user-supplied codes independent of the desktop software deployed in a particular project. We envision a parallel set of export functions that permit full interoperability among coding tools. The priority will be to create interoperability with tools using open, not proprietary, file formats. Databrary also recognizes the need to develop open standards and interfaces that enable Databrary to link to and synchronize with outside sources that specialize in other data types (see the LearnLab and EdWise sections below).

POLICIES FOR SAFE AND SECURE VIDEO SHARING

Policies for openly sharing identifiable data in ways that securely preserve participant privacy are essential for sharing research video. Databrary does not attempt to de-identify videos. Instead, we maximize the potential for video reuse by keeping recordings in their original unaltered form. To make unaltered raw videos available to others for reuse, Databrary has developed a two-pronged access model that (a) restricts access to authorized researchers, and (b) enables access to identifiable data only with the explicit permission of participants.

To gain access to Databrary, a person must register on the site. Applicants agree to uphold Databrary's ethical principles and to follow accepted practices concerning the responsible use of sensitive data. Each applicant's institution must co-sign an access agreement. Full privileges are granted only to those applicants with independent researcher status at their institutions. Others may be granted privileges if they are affiliated with a researcher who agrees to sponsor their application and supervise their use. Ethics board or IRB approval is not required to gain access to Databrary because many use cases do not involve research, but IRB approval is required for research uses. Once authorized, a user has full access to the site's shared data, and may browse, tag, download for later viewing, and conduct non- or pre-research activities.

Unique among data repositories, the Databrary access agreement authorizes both data use and contribution.

However, users agree to store on Databrary only materials for which they have ethics board or IRB approval. Data may be stored on Databrary for the contributing researcher's use regardless of whether the records are shared with others or not. When a researcher chooses to share, Databrary makes the data openly available to the community of authorized researchers.

In addition to permitting access to only authorized researchers, Databrary has extended the principle of informed consent to participate in research to encompass permission to share data with other researchers. To formalize the process of acquiring permission, Databrary has developed a Participant Release Template[138] with standard language we recommended for use with study participants. This language helps participants to understand what is involved in sharing video data, with which the data will be shared, and the potential risks of releasing video and other identifiable data to other researchers.

MANAGING DATA FOR SHARING AND BUILDING A COMMUNITY

When researchers do share, standard practice involves organizing data after a project has finished, perhaps when a paper goes to press. This "preparing for sharing" after the fact presents a difficult and unrewarding chore for investigators. It makes curating and ingesting data sets challenging for repositories, as well. Databrary has chosen a different route to curation. We have developed a data management system that empowers researchers to upload and organize data as it is collected. Immediate uploading reduces the workload on investigators, minimizes the risk of data loss and corruption, and accelerates the speed with which materials become openly available. The system employs familiar, easy-to-use spreadsheet and timeline-based interfaces that allow users to upload videos; add metadata about tasks, settings, and participants; link related files; and assign appropriate permission levels for sharing. To encourage immediate uploading, Databrary provides a complete set of controls so that researchers can restrict access to their own labs or to other users of their choosing. Datasets can be openly shared with the broader research community at a later point when data collection and ancillary materials are complete, whenever the contributor is comfortable sharing, or when journals or funders require it.

Data sharing works only when the scientific community embraces it. From the beginning, Databrary has sought to cultivate a community of researchers who support data sharing and commit to enacting that support in their own work flows. Our community building efforts involve many interacting components. They include active engagement with professional associations, conference-based exhibits and training workshops, communications with research ethics and administration staff, talks and presentations to diverse audiences, and one-on-one consultations with individual researchers and research teams. These activities are time- and labor-intensive, but we believe that they are critical to changing community attitudes toward data sharing in the educational and learning sciences. Looking ahead, it will be critical to engage funders, journals, and professional organizations in the effort to forge community consensus about the importance, feasibility, and potential of open video data sharing.

SUMMARY

As Gesell once noted, cameras can record behavior in ways that make it "as tangible as tissue"[139]. The Databrary team contends that video has a central role to play in efforts to make tangible the anatomy of successful teaching and learning. In fact, we argue that video can be the core around which other measures of teaching and learning cluster. This requires reducing barriers to sharing video and fostering new community values around data sharing that make it indispensible. The Databrary project has built technology and policies that overcome many of the most significant barriers to widespread sharing within the developmental sciences community. Databrary suggests ways that video and other identifiable data collected in the context of

[138] Databrary Project. 2015. Databrary Release. Retrieved from http://databrary.org/access/policies/release-template.html

[139] Scott CS. 2011. 'Tangible as tissue': Arnold Gesell, infant behavior, and film analysis. *Science in Context, 24* (3), 417-42.

education research might also be shared. Technologies and policies for providing secure access to videos for broader use cases will have to be developed, tools that allow desktop coding software files to be seamlessly converted to and from one another will have to be perfected, and ways of synchronizing and linking disparate data streams will have to be created. Equally important, communities of scholars dedicated to collecting, sharing, and mining education-related video data will have to be cultivated. But we believe that the widespread sharing of high value, high impact data of the sort that video can provide promises to achieve this ambitious vision to advance education policy and improve practice. Databrary is working toward a future where open video data sharing is the norm, a personalized science of teaching and learning is the goal, and what optimizes student learning is as tangible as tissue.

KOEDINGER: LEARNLAB'S DATASHOP

One concern to raise regarding data intensive research is the question of whether we are currently using data as effectively as possible? In education we sometimes seek data to confirm our intuitions rather than looking at data closely to try to determine what is really going on in student learning. It is a bit like looking around and deciding the world is flat and then seeking data to confirm that. We need to take the position that learning is not plainly visible and that we will not be able to gain insight by simply reflecting on our classroom experiences. We sometimes act as though we can easily observe learning. For example, in response to data (e.g., Duckworth et al., 2011[140]; Ericcsson et al., 1993[141]) indicating that a substantial amount of deliberate practice is needed to acquire expertise, Hambrick et al. (2014)[142] quote Gardner (1995)[143] as suggesting "the deliberate practice view 'requires a blindness to ordinary experience' (p. 802)." Instead of relying on our "ordinary experience," we need to couple careful investigation of data, along with theoretical interpretation, to get at what is unseen and not immediately apparent. While ordinary experience suggests the world is flat, it took a combination of data and geometric theory to infer that the world is round and initially measure its circumference. New data opportunities in education, especially ones afforded by technology use, will increasingly allow us to get beyond our ordinary experience and yield insights into what cognitive, situational, motivational, and social emotional factors cause the unseen changes in learners' minds that lead to desired educational outcomes.

We have pursued this idea in LearnLab, an NSF-funded Science of Learning Center[144]. A major output of LearnLab has been the creation of DataShop, the world's largest open and free repository of educational technology data and analytic methods[145]. One of the many insights that can be drawn from the vast amount of data we have collected in DataShop is evidence on the rate at which learning occurs (Figure 17). We see across many data sets that each opportunity to practice or learn a skill in the context of problem-solving reveals a rather small average improvement in student success (or, equivalently, drop in error rate, as shown in Figure 17).

[140] Duckworth AL, Kirby TA, Tsukayama E, Berstein Hand Ericsson, KA. 2011. Deliberate practice spells success: Why grittier competitors triumph at the National Spelling Bee. Social Psychological and Personality *Science, 2,* 174–181. http://dx.doi.org/10.1177/1948550610385872.

[141] Ericsson KA, Krampe, RTand Tesch-Römer C. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

[142] Hambrick DZ, Oswald FL, Altmann EM, Meinz EJ, Gobet Fand Campitelli G. 2014. Deliberate practice: Is that all it takes to become an expert? *Intelligence*, 45, 34-45.

[143] Gardner H. 1995. "Expert performance: Its structure and acquisition": Comment. *American Psychologist, 50,* 802–803. http://dx.doi.org/10.1037/0003-066X.50.9.802.

[144] Koedinger KR, Booth JLand Klahr D. 2013. Instructional complexity and the science to constrain it. *Science*, 342, 935-937.

[145] Koedinger KR, Baker R, Cunningham K, Skogsholm A, Leber Band Stamper J. 2011. A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, R.S.J.d. Baker (Eds.). *Handbook of Educational Data Mining* (pp. 43-55). Boca Raton, FL: CRC Press.

These changes in student success across opportunities to practice or learn (get as-needed feedback or instruction on a skill) can be modeled as learning curves. DataShop provides a statistical modeling technique for estimating the shape of the learning curves that uses a logistic regression generalization of item-response theory called AFM[146]. The predictions of AFM are shown in blue dotted lines in Figure 17, with the actual data (average error) shown in red solid lines.

The average error rate increases about .15 in log odds (or "logit" scale used in item response theory and, more generally, logistic regression) for every opportunity to practice. That means if a group of students are at about 50 percent correct on a skill, after one opportunity of practice they will now be at about 54 percent correct. This 4 percent increase diminishes as correctness increases toward 100 percent. Thus, to get from 50 percent correct to 95 percent correct requires about 20 practice opportunities.

This learning rate estimated from educational technology data seems faster (indicating about 15 minutes of accumulated learning time per skill) than data from self-reports on expertise acquisition (Ericsson et al., 1993) that suggests it takes about 10,000 hours to become an expert[147]. Other estimates that expertise involves about 10,000 chunks of knowledge (or skills), yields a learning rate of about 1 hour per skill. The faster learning rate apparent in educational technology data might be an indication that deliberate practice in the context of educational technology is more effective than it is in the typical real-world learning environment. These estimates are rough at this point, so more careful work would need to be done to make such a point firmly and rigorously. Nevertheless, it does open the possibility for interesting further research. Might it be possible to establish some baselines on which to compare learning rate achieved by different instructional approaches or learning supports?

We do see large variations for different skills (see the first row in Figure 17). For example, in a unit on geometric area, learning rate for finding the area of triangles is .03 logits whereas the learning rate for the planning skill of identifying what regular shapes to use to find the area of an irregular shape is .15 logits. However, there

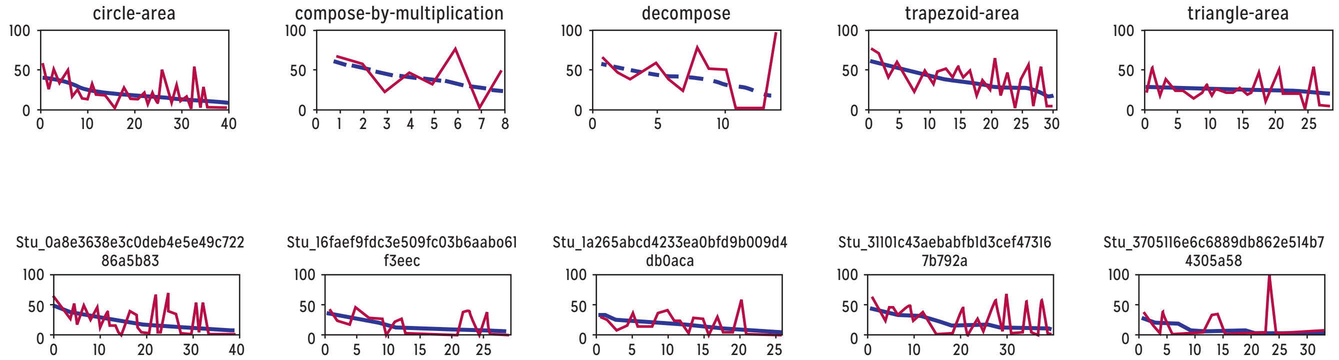


***Figure 17** Learning curves showing a decrease in error rate (y-axis) for each successive opportunity (x-axis) to demonstrate or learn a skill, averaged across students for different skills in the first row and averaged across skills for different students in the second row. The variations in learning rate (how much the error changes for each opportunity) are much bigger for skills than for students (the curves in the first row have more variation in their slopes than the curves in the second row). (Respective learning rates in log odds for the five skills shown are .07, .27, .15, .09, and .03.)*

[146] Koedinger KR, McLaughlin E Aand Stamper JC. 2012. Automated Student Model Improvement. Yacef, K., Zaïane, O., Hershkovitz, H., Yudelson, M., and Stamper, J. (eds.) *Proceedings of the 5th International Conference on Educational Data Mining*. (pp. 17-24) Chania, Greece. Best Paper Award.

[147] Ericsson K A, Krampe RTand Tesch-Römer C. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

is a relatively small variation across students[148]. At least relative to skills, it seems that most students learn at about the same rate. In contrast, some skills are much harder to learn than others and these skill difficulty variations are common across all students. We do find some variation in learning rate across students and this variation is quite interesting. What accounts for these student differences in learning rate? Is it innate ability, differences in domain-specific prior knowledge, or in general but malleable, metacognitive learning skills, motivational dispositions, identity self-attributions? To the extent that student learning rate differences are not innate, might it be possible to increase the learning rate of some students through instruction that addresses one of these causes? In other words, is there a data-driven path to helping students learn how to learn?

Returning to the larger point, given the relatively consistent and, frankly, relatively slow rate at which learning generally occurs across students, we can ask whether it might be better to focus attention on learning supports or instructional methods that increases learning for all. These methods may still be highly student adaptive to the large variations in student learning progress (how much students know), despite our observation above about the relatively small variations in student learning rate (how quickly they can change what they know). (Note: Large student variations in learning progress and achievement are clearly apparent in DataShop data sets even as only small variations in learning rate are seen.) Which of the trillions of different combinations of learning supports is best for what kinds of student learning outcomes?

The increasing availability of large-scale data, for instance, from massively open online courses (MOOCs) brings further opportunities to address these and other questions. For example, a recent analysis of a psychology MOOC data set explored how variations in students' choices to use different learning resources was associated with learning outcomes. Students who choose to do more interactive activities (tasks with as-needed feedback and instruction) had six times better learning outcomes (total quiz and final exam scores) than students who chose to watch more videos or read more web pages[149]. Many questions remain unanswered including: What particular patterns of learning resource use did students engage in? Do significant differences in student learning rates emerge in this course due to their resource choices and/or strategies? Do these results generalize to other online courses?

With the help of NSF funding (Data Infrastructure Building Blocks), a team of researchers at Carnegie Mellon University, MIT, Stanford University, and University of Memphis are building LearnSphere (learnsphere.org) to help data researchers address these questions.

GUMMER: THE EWING MARION KAUFFMAN FOUNDATIONS EDWISE

In contrast to the two preceding cases on micro-level learning data, EdWise centers on the meso- and macro-levels of educational data and the potential for integration across these data levels to inform research and policy studies. The NSF recently funded a proposal for researchers at SRI who are examining the ways in which teachers make use of data from an online learning platform that includes instructional resources and content assessments that serve as the central structure in the students' learning environments. The intent of the research is to examine the key challenges facing practitioners in their use of information that comes from data-intensive research methods and to identify what partnership activities best support evidence-based practices. The findings from this study will lead to an understanding of the utility and feasibility of a teacher's use of the volumes of data that come from virtual learning environments, effectively bridging the micro- and meso-level data categories.

[148] Liu Rand Koedinger KR. 2015. Variations in learning rate: Student classification based on systematic residual error patterns across practice opportunities. In *Proceedings of 8th International Conference on Educational Data Mining*. Madrid, Spain.

[149] Koedinger KR, Kim J, Jia J, McLaughlin EAand Bier NL. 2015. Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC. In Proceedings of the Second (2015) ACM Conference on Learning at Scale, 111-120.

The collection and use of data collected at the meso-level has lagged well behind the development of rich data archives at both the micro- and macro-level. The Race to the Top initiative of the U.S. Department of Education has supported a number of states to develop and implement Instructional Improvement Systems (IISs) that are currently being investigated. An IIS model frequently includes systems that support curriculum, formative and interim assessment, and instructional (lesson-planning) management. They also facilitate the use of electronic grade books and may support a professional development management component. The IIS frequently includes a daily import of data from the state's Student Information System (SIS) that includes attendance and disciplinary data. Usually constructed by a vendor identified through a competitive bidding process, these data systems include standards-aligned lesson plans developed by teachers and externally developed resources that are linked to grade books, enabling researchers to examine not only student achievement but also opportunity to learn. Data from these systems are also used to support the determination of early warning systems that inform districts and schools about at-risk students.

Much of the meso-level education data are collected through school and district-level systems that include student demographics, attendance, disciplinary behavior, course-taking, grades, local assessments (formative, benchmark and interim), state assessments, and SAT and/or ACT testing data. Student information systems are frequently linked to human resource data systems that facilitate connecting information about teachers to student data. Educators at the school and district level are provided data dashboards that facilitate the display of data in formats that are intended to be easy to interpret. Increasingly, educator use of these data systems has been a focus of research at the school level where the data do not necessarily correspond to big data. A study by Brunner, Fasca, Heinze, Honey, Light, Mandinach and Wexler (2005) documented the ways in which teachers used the paper and web-based data that were provided to them through the Grow Network in the New York public schools[150]. Findings from this study emphasized the focus on "bubble" students, those who are on the cusp of meeting proficiency on the high-stakes testing. Other researchers have examined the interpretive processes and social and organizational conditions under which data use is conducted[151]. But if we begin to study populations of teachers in districts using data, the scale increases significantly. These meso-level data sets connect to the macro-level data in that much of the data included in them are reported to the state longitudinal data systems.

The U.S. Department of Education's State Longitudinal Data Systems (SLDS) have supported the development of P-20 data systems that frequently are attached to workforce data. These data represent the macro level of data and they contain data at a much larger grain size than micro- and meso-level systems. While many states are at varying levels of interoperability of the data in these systems, some states have developed systems that allow for quite sophisticated research and policy questions to be addressed. For instance, the State of Washington Education Research and Data Center (WA-ERDC), housed in the state's Office of Financial Management, was created in 2007 to assemble, link, and analyze education and workforce data and to support research focusing on student transitions. The WA-ERDC includes data from the following agencies:

- Department of Social and Health Services (social service program participants);
- Department of Early Learning, Office (early learning and child-care providers);
- Office of Superintendent of Public Instruction (P-12 student state assessment, attendance, course-taking patterns, graduation, and information about teachers);

[150] Brunner C, Fasca, C,Heinze J, Honey M, Light D, Mardinach E,Wexler D. 2005. Linking data and learning: The Grow Network study Journal of Education for Students Placed At Risk 10 (3), 241-267.

[151] Coburn C Eand Turner EO. 2011. Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research and Perspectives, 9*, 173- 206.

- Washing Student Achievement Council (financial aid information);
- State Board for Community and Technical Colleges (students, courses, degrees, and majors);
- Public Centralized Higher Education Enrollment System (students, courses, degrees, and majors);
- Workforce Training and Education Coordinating Board (career schools and non-credit workforce programs);
- Labor and Industries (state apprenticeships); and
- Employment Security (industry, hours, and earnings).

From the integration of these data, the WA-ERDC can produce information for parents, teachers, administrators, policy makers, and researchers. The center routinely provides data sets to researchers that contain de-identified data that can still be linked longitudinally under specific Memoranda of Understanding that protect student privacy.

Seven of the first two years of awards from the NSF Building Community and Capacity for Data Intensive Research program focused on building the education and social science research community to use integrated systems that included education data at their core. Northwestern and Duke universities were funded to begin to develop a national interdisciplinary network of scholars that would use new data sets that linked K-12 data to birth and medical records, information from Medicaid and welfare programs, preschool and early childhood interventions, marriage and criminal records, and other workforce data. These linked data sets facilitate research on early childhood investments and interventions and their effect on school performance. They also provide the opportunity to focus on salient long-run adult outcomes rather than just test scores. The Minnesota Linking Information for Kids (Minn-LInK) project expanded the focus of cross-linked data to support a more complete understanding of child well-being with a special focus on at-risk children and youth. In Ohio, the Ohio Longitudinal Data Archive seeks to examine the effects of educational processes from pre-school through graduate study on economic development in that state. In Virginia, researchers working with Project Child HANDS are designing the data interface and analytic tools and determining the data governance structure and processes to facilitate the use of social services, childcare quality, and educational data.

HOW EDWISE ADDRESSES THIS SITUATION

In response to the Digital Accountability and Transparency Act (DATA Act), the Data Quality Campaign has increased calls for "data that are accessible, understandable, and actionable so they can make informed decisions… data collection and public reporting efforts should move away from simply complying with state and federal regulations and toward answering stakeholders questions[152]."

The Ewing Marion Kauffman Foundation has developed a data tool that is intended for such public use by multiple stakeholders. Called EdWise, it is scheduled for use during the early summer of 2015.

EdWise came out of the need for data to inform both the identification of schools that would benefit from foundation support and the need to be good stewards of EMKF funds by providing evidence of the potential influence or impact of the funding actions. The state of Missouri provides spreadsheets of data on its website that include thousands of lines of data. We have combined 14 million records of Missouri K-12 education data into a single easy-to-use online tool to help parents, educators, school districts, policy makers, and the public better understand the educational landscape and make informed education decisions. With these macro-level and aggregate data, parents can identify schools and districts in which to enroll their children. More importantly, school districts can better identify other districts that have similar characteristics and they might provide either more targeted examples to query for

[152] McClafferty S. 2014. Public reporting in the DATA Act. Retrieved August 3, 2015 from http://dataqualitycampaign.org/blog/2014/10/public-reporting-in-the-data-act/.

assistance, or to use as comparisons when newspapers report annual achievement rates. EdWise contains hundreds of variables that extend over two decades to understand trends over time. EdWise does not contain student or teacher data from Kansas as these data are currently embargoed under Kansas legislation. We are currently working with the departments of higher education in both Kansas and Missouri to connect aggregate data from postsecondary institutions with K-12 information. But what the higher education data users want is the connection of higher education data with that from work force so that they can demonstrate the importance of postsecondary education. In our experience, each level of the system wants to look both behind and ahead of its own level of data.

Integrating these multiple levels of data presents serious technical and system-level problems. Data are frequently still sequestered in silos within and across different levels. Figuring out how to address issues around identifiers is another technical problem. Privacy issues are also a barrier to integrating data sources. But what kinds of real-world educational questions might we answer if we solved these problems and developed the ability to truly track students across educational contexts and systems?

The most successful scientific data repositories employ a mixture of funding strategies to sustain the staff and infrastructure associated with them. We believe that core support from federal funding sources will be critical, but we also believe that a combination of host institutional support, non-host institutional subscription, user fees, new research-oriented grants, professional society partnerships, and journal and publisher support can sustain repositories over the long term.

Some Possible Implications of Data-Intensive Research for Education

George Siemens (University of Texas-Arlington) and Jere Confrey (North Carolina State University)

As we learn how to better conduct data-intensive research in education, how may this change schooling and research?

SIEMENS: A FOUNDATION FOR A NEW SCIENCE OF LEARNING[153]

Learning analytics (LA) have to date primarily imported concepts from big data, computer science, machine learning, and related fields. As a result, many of the methods of experimentation and research are not native to the learning space, but rather applications from sociology (social network analysis), language studies (discourse analysis), computer science (data mining, artificial intelligence, and machine learning), and statistics (analytic methods). While this has enabled LA to develop in influence and impact, it has not produced the types of insights that can be expected from a new knowledge domain that synthesizes and integrates insights from numerous fields while developing its own methodologies.

With the broad aim of redefining educational research–where we move from "dead data" to "live data"–two critical needs exist:

1. Development of personal learning graphs (PLeGs) to capture learner profile, knowledge, learning patterns, and learning history;
2. Creation of an open learning analytics architecture to enable academics to collaboratively develop analytics products and evaluate LA algorithms and test claims made by researchers and corporate providers

PERSONAL LEARNING GRAPH

Educators require a better profile of what a learner knows than currently exists. The previous experiences and knowledge of individual learners are inconsistently acknowledged in educational settings. Courses focus on what has been determined to be important for learners to know rather than personalizing to what an individual learner already knows. As a result, limited progress has been made around personalized and adaptive learning. Initiatives such as CMU and Stanford's OLI[154] and several corporate providers have gained attention, but are largely confined to courses with a clear right or wrong answers, such as statistics and math courses. In these instances, the learner's knowledge profile is kept within an existing software system or within a corporate platform.

In education, a PLeG is needed where a profile of what a learner knows exists. Where the learner has come to know particular concepts is irrelevant; this may be via work, volunteering, hobbies, personal interest, formal schooling, or massive open online courses (MOOCs). What matters is that all members involved in an educational process, including learners, faculty, and administrators, are aware of what a learner knows and how this is related to the course content, concepts, or curriculum in a particular knowledge space. Four specific elements are included in the multipartite graphs that comprise PLeG:

- Social learning activities and networks;
- Cognitive development and concept mastery;
- Affective and engagement;
- Process and strategy (meta-cognition)

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 (Figure 18).

[153] Thanks to previous publications: Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., Ferguson, R., ... & Baker, R. S. J. D. 2011. *Open Learning Analytics: an integrated & modularized platform* (Society for Learning Analytics Research).

[154] "Open Learning Initiative | Open Learning Initiative." Open Learning Initiative. Carnegie Mellon University, n.d. Web. 15 Sept. 2015. <http://oli.cmu.edu/>.

As education systems continue to diversify, offering a greater set of educational products and engaging with a broader range of students than in the past, the transition to personal learning graphs, instead of focusing on the content within a program, can enable the system to be far more intelligent than it currently is. For example, in a learning system based on a learner knowledge graph, one's career path would be greatly enhanced; a learner could know where she is in relation to a variety of other fields based on the totality of her learning (i.e., "this is your progress toward a range of careers"). (Figure 19) A student returning to university would have a range of course options, each personalized to her knowledge and skills, rather than being pushed through a pre-established curriculum without regard for existing knowledge. With PLeG, returning to a university to up-skill and enter new fields–an increasing requirement as entire fields of work are at risk from automation–will create a transition from a learner having a four-year relationship with a university to one where a learner has a forty-year relationship with a university. In this model, learners continue to learn in online or blended settings while employed and move to intensive on-campus learning when transitioning to a new career.

KNOWLEDGE DOMAINS	*PERSONAL KNOWLEDGE GRAPH*
NURSING	*82%*
COMPUTER SCIENCE	*65%*
LANDSCAPING	*38%*

Figure 19. *Returning and advancing degrees*

Pedagogically, PLeG affords new opportunities for individuals to take personal control of their learning (Figure 20). In this model, a learner can simultaneously engage with structured course content and create networked and connective knowledge structures[155]. As mobiles and wearable computing develop as critical

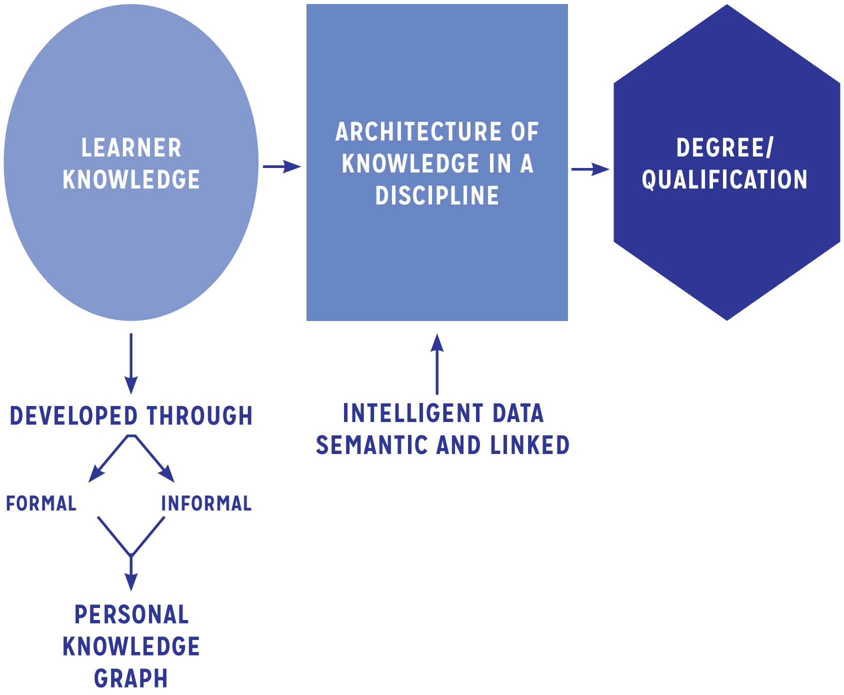


Figure 18: *Matching knowledge domains to learner knowledge*

[155] Siemens G. 2007. "Connectivism: Creating a Learning Ecology in Distributed Environment," in Didactics of Microlearning: Concepts, discourses, and examples, in T. Hug, (ed.),Waxmann Verlag, New York, pp. 53-68.

technologies for knowledge work, this approach is reflective of the networked world of learning and the personal lives of individuals. In addition to algorithmically guided personalized learning, socially navigated personal learning provides opportunities for serendipity and creative learning. Learning pathways, within PLeG, are established by machine learning and algorithmic models and by personal learning networks and social interactions.

To be effective, PLeG needs to be developed as an open platform where learners are able to share knowledge, personal profiles, and learning practices with universities and corporations. The model is envisioned to be similar to the IMS Learning Tools Interoperability[156] protocol where API access to certain types of information is brokered in a trusted environment. Essentially, learners would own their PLeG, and standards would be established that permits trusted sharing with education providers.

OPEN LEARNING ANALYTICS PLATFORM

The open learning analytics (OLA) platform addresses the need for integrated toolsets through the development of four specific tools and resources (see Figure 21 for visual representation):

1. Learning analytics engine
2. Adaption and personalization engine
3. Intervention engine: recommendations, automated support
4. Dashboard, reporting, and visualization tools

LEARNING ANALYTICS ENGINE

The analytics engine is the central component in the OLA system. Leveraging best practices from both the learning analytics and educational data-mining communities, the analytics engine incorporates data from learning management systems, social web, physical

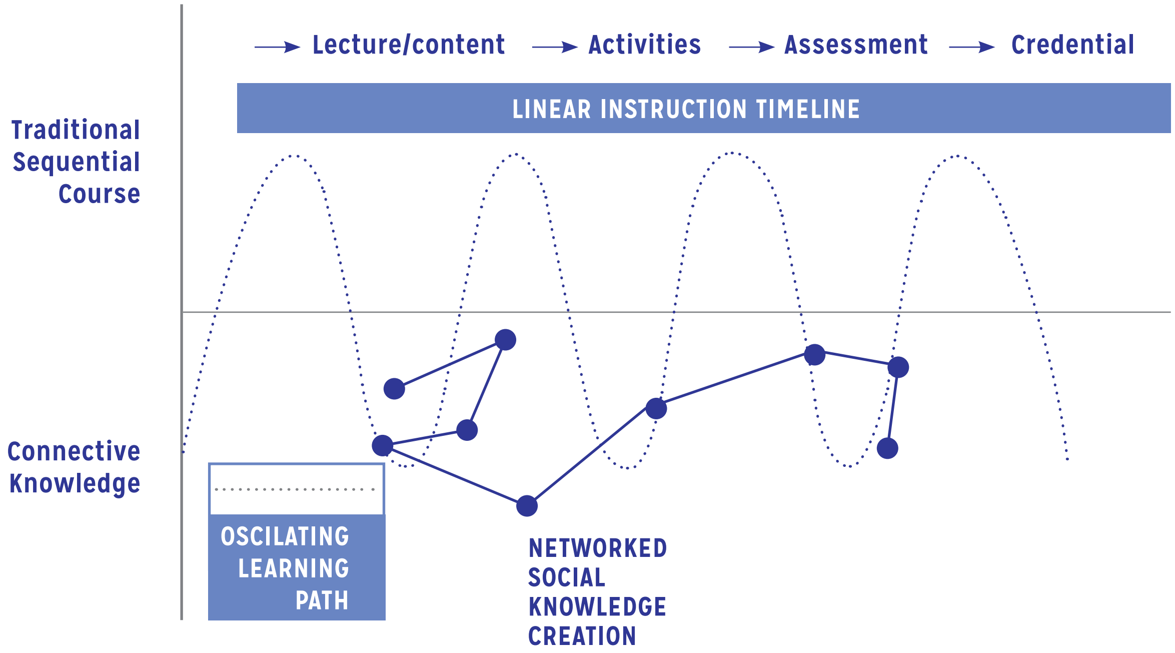

Lecture/content
Activities
Assessment
Credential
LINEAR INSTRUCTION TIMELINE
Traditional Sequential Course
Connective Knowledge
OSCILATING LEARNING PATH
NETWORKED SOCIAL KNOWLEDGE CREATION

Figure 20. *Flexible and variable learning pathways*

[156] "More Information." IMS Global: Learning Tools Interoperability. N.p., n.d. Web. 15 Sept. 2015. <http://www.imsglobal.org/toolsinteroperability2.cfm>.

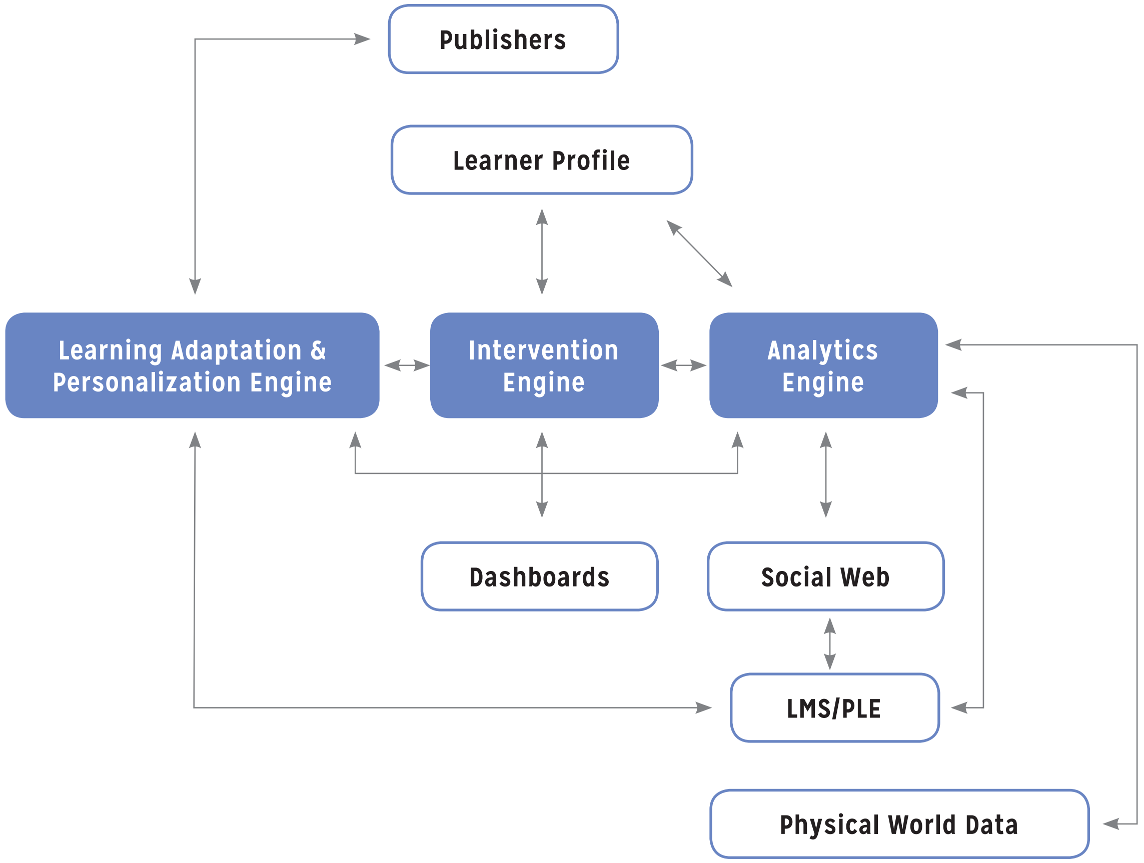


Figure 21. *Conceptual framework for open learning analytics platform*

world-data (such as classroom attendance, use of university resources, and GPS-data when completing activities such as surveying), and mobile and wearable technologies. This is essentially the "Apache" of learning analytics–an open platform where researchers can build their products and share as plugins with other researchers. Rather than engaging with a range of different tools, each with a distinct interface, the analytics engine provides a consistent space for interaction with data and various types of analysis. This is similar to libraries within Python or plugins in WordPress. The platform stays the same, but the functionality is extended by plugins.

The analytics engine then serves as a framework for identifying and then processing data based on various analysis modules (Figure 22). For example, the analysis of a discussion forum in an LMS would involve identifying and detailing the scope of the forums and then applying various techniques, such as natural language processing, social network analysis, process mining (to consider the degree of compliance between instructional design and the log data of learner activities), trace analysis of self-regulated learning, the development of prediction models based on human assessment of interactions, identification of at-risk students, and the process of concept development in small peer groups.

As learning analytics develops as a field, plugins created by other researchers or software vendors can be added as modules for additional analysis. Having a global research community creating modules and toolsets, each compatible with the analytics engine, will prevent the fragmentation that makes research difficult in numerous academic fields. If researchers share data, algorithms, and toolsets in a central environment like the open learning analytics platform, we expect to see the rapid growth of educational data-mining and learning analytics. This growth will, in turn, contribute to the formation and development of a new science of learning research that provides rapid feedback on real-time data to learners, academics, and institutions.

ADAPTION AND PERSONALIZATION ENGINE

The learning adaptation and personalization will include adaptively of the learning process, instructional design, and learning content. For example, this adaptation engine could connect the analytics engine with content developers. Developers could include existing publishers such as Pearson or McGraw-Hill, as well as institutional developers and any implemented curriculum documentation processes and tools. When learning materials are designed to reflect the knowledge architecture of a domain, the content delivered to individual learners can be customized and personalized. The personalization and adaptation engine draws from the learner's profile, when permitted by the learner, as defined in the learning management system and social media sources.

INTERVENTION ENGINE

The intervention engine will track learner progress and provide various automated and educator interventions using prediction models developed in the analytics engine. For example, a learner will receive recommendations for different content, learning paths, tutors, or learning partners. These soft interventions are nudges toward learner success by providing learners with resources, social connections, or strategies that have been predictively modeled to assist others. Recommendations have become an important part of finding resources online, as exemplified by Amazon, Spotify, and Google. In education, recommendations can help learners discover related, but important, learning resources. Additionally, the intervention engine can assist learners by tracking progress toward learning goals.

Automated interventions also include emails and reminders about course work or encouragement to log back in to the system when learners have been absent

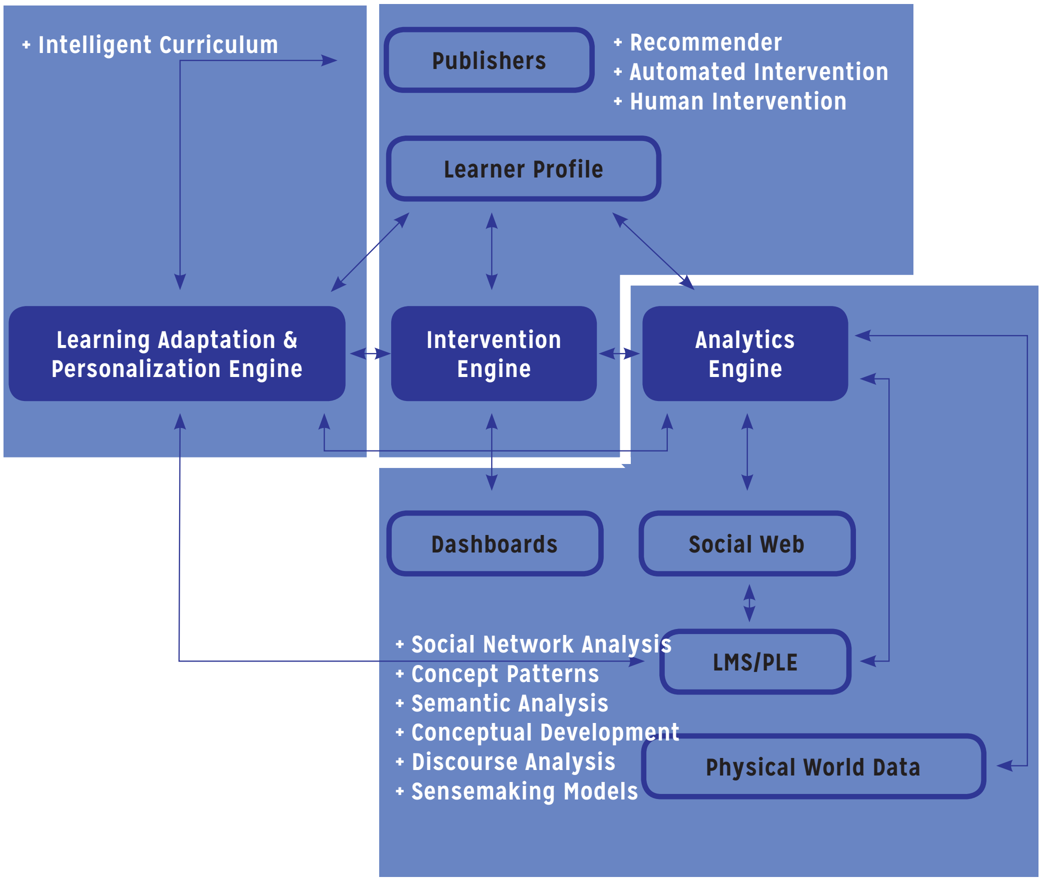


Figure 22. OLA function areas

for a period of time that might indicate "risky behavior." Interventions will also be triggered for educators and tutors. When a learner has been notified by automated email but has failed to respond, the intervention engine will escalate the situation by sending notices to educators and tutors to directly contact the student. The value of direct intervention by a teacher as a motivating condition for return to learning tasks is well documented by existing education research.

DASHBOARD REPORTING

The dashboard is the sense-making component of the LA system, presenting visualized data to assist individuals in making decisions about teaching and learning. The dashboard consists of four views: learner, educator, researcher, and institutional. Learners will be able to see their progress against that of their peers (names will be excluded where appropriate), against learners who have previously taken the course, against what they themselves have done in the past, or against the goals that the teacher or the learner herself has defined. Educators will be able to see various representations of learner activity, including conceptual development of individual learners, progress toward mastering the course's core concepts, and social networks to identify learners who are not well connected with others. Analytics for educators will be generated in real-time, as well as hourly or daily snapshots, depending on the context. The dashboard will provide institution-level analytics for senior administrators to track a learner's success and progress. When combined with academic analytics, this module will be a valuable business intelligence tool for analyzing institutional activities.

Based on criteria established through research of the learning analytics system (such as the impact of social connectivity on course completion, predictive modeling, and warning signals such as changes in attendance), automated and human interventions will be activated to provide early assistance to learners demonstrating a) difficulty with course materials, b) strong competence and needing more complex or different challenges, and c) risk of dropping out.

CONCLUSION

All stakeholders in the education system today have access to more data than they can possibly make sense of or manage. In spite of this abundance, however, learners, educators, administrators, and policy makers are essentially driving blind, borrowing heavily from techniques in other disciplines rather than creating research models and algorithms native to the unique needs of education. New technologies and methods are required to gain insight into the complex abundant data encountered on a daily basis. The development of personal learning knowledge graphs and an open learning analytics platform are critically needed innovations to contribute to and foster a new culture of learning sciences research. The proposed integrated learning analytics platform attempts to circumvent the piecemeal process of educational innovation by providing an open infrastructure for researchers, educators, and learners to develop new technologies and methods. Today's educational climate, which includes greater accountability in a climate of reduced funds, suggests new thinking and new approaches to change are required. Analytics hold the prospect of serving as a sense-making agent in navigating uncertain change by offering leaders with insightful data and analysis, displayed through user-controlled visualizations.

CONFREY: THE VALUE OF LEARNING MAPS AND EVIDENCE-CENTERED DESIGN OF ASSESSMENT TO EDUCATIONAL DATA MINING

Numerous efforts are underway to build digital learning systems, and the designs of such systems vary in critical aspects: components, organization, extensibility, adaptability, data intensity, and use. With support from the Bill and Melinda Gates Foundation, one set of researchers is linking learning maps to systems of assessment and analytics in order to define and examine progress in learning across large numbers of students (projects include: Next Generation Schools, Glass Labs, Enlearn, CRESST, and Dynamic Maps). One of the projects, Scaling Up Digital Design Studies (SUDDS) at North Carolina State University, is highlighting how structure and design can inform efforts for applying big data techniques in mathematics education. Articulating an explicit theory of student-centered learning can help in leveraging big data to improve the depth of learning, as opposed to simply leveraging performance from users of digital learning systems at a possible cost to understanding. It is a conjecture that remains to be tested.

A STUDENT-CENTERED DIGITAL LEARNING SYSTEM

A representation of a digital learning system (DLS) is shown below (Figure 23). It consists of a learning map, which delineates the topics to be learned as big ideas and their underlying learning structure. The constructs in the map are connected to a set of Internet resources that can be deployed as curricular materials, and a means of lesson delivery combined with a workspace and access to a set of math-specific tools. The map is also linked to multiple forms of assessments and reporting. The whole system will be undergirded with an analytic system to monitor, study, and modify the use of the DLS. Supports for teaching refer to activities around professional development materials and means for teachers to manage the system. The arrows along the bottom indicate from where feedback comes and to where it is delivered.

A DLS is student-centered when each of these components is designed to strengthen students' movement within the digitally enabled space. A student-centered DLS:

- Increases students' ability to understand what they are learning,
- Supports appropriate levels of choice in sequencing or making decisions about materials (with guidance of teachers or knowledgeable adults)
- Supports genuine mathematical work including an authentic use of the tools (not just filling in answers),
- Affords peer collaboration, including discussing and sharing results,
- Allows students to create, store and curate products, and
- Provides students' diagnostic feedback allowing them to self-monitor and set goals

Student-centeredness does not imply individualization, working largely alone at one's own speed, but it does

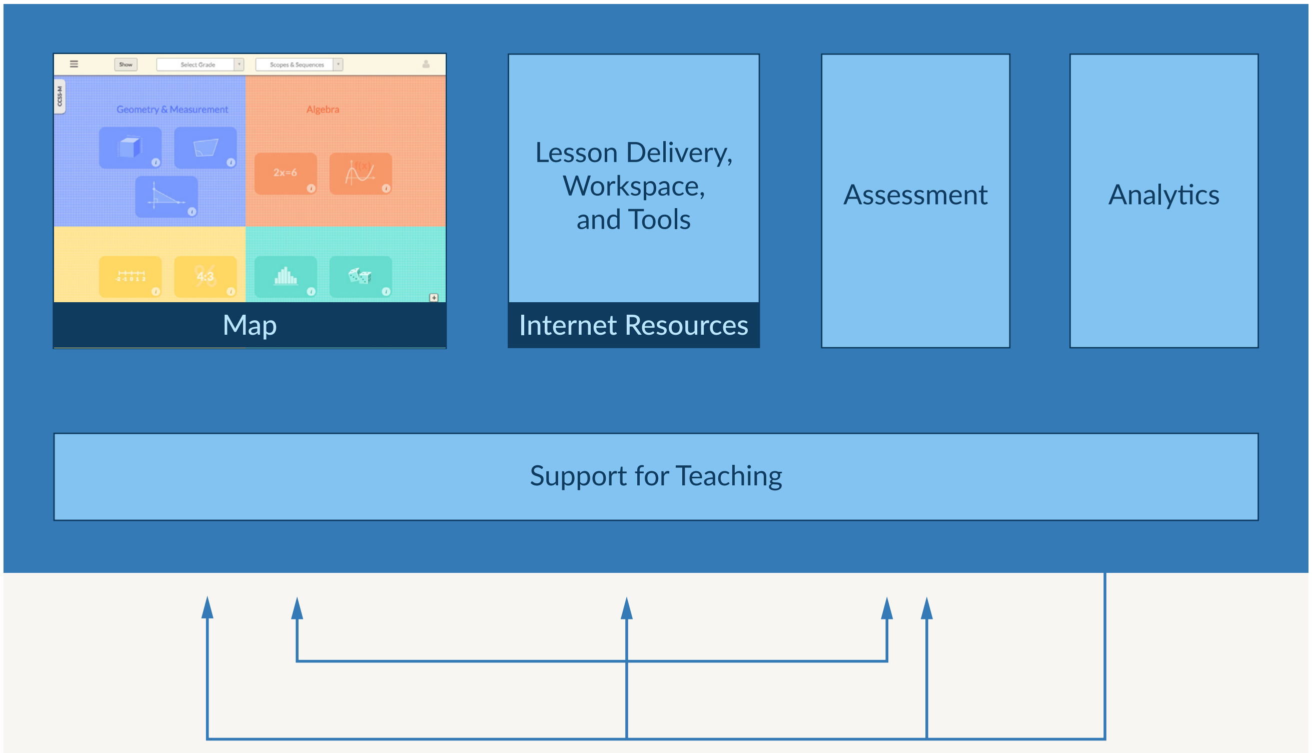


Figure 23. Components of a digital learning system

support personalization, making choices and self-regulation[157]. The DLS can be used by classes using predefined scope and sequences to coordinate activities.

INTRODUCING THE SUDDS GRADES 6-8 LEARNING MAP

A learning map is a configuration in space of the primary concepts to be learned. Our middle school version is organized hierarchically to show the major fields in mathematics (number, statistics, and probability; measurement and geometry; and algebra) and nine big ideas from across those fields. These nine big ideas, called regions, span all grades (6-8) and include such topics as "Compare quantities to operate and compose with ratio, rate, and percent" or "display data and use statistics to measure center and variation in distributions." Big ideas, rather than relying on individual standards, have the advantage of providing focus both at and across grades. Too many systems attempt to map learning standard by standard, which is problematic since standards vary in size and often apply to multiple big ideas.

The level below the regions is comprised of related learning clusters (RLCs). At this level, the research on student learning has a significant impact on the map. RLCs are sets of constructs that are learned in relation to each other, and their spatial configurations on the map convey to the user information about those relationships. For instance, within the big idea of ratio, rate, and percent, there are five RLCs: 1) key ratio relationships, 2) comparing and finding missing values in proportions, 3) percents, 4) calculating with percents, and 5) rational number operators. In a region or big idea, the RLCs' organization from bottom left to upper right conveys to the users to address the first cluster, key

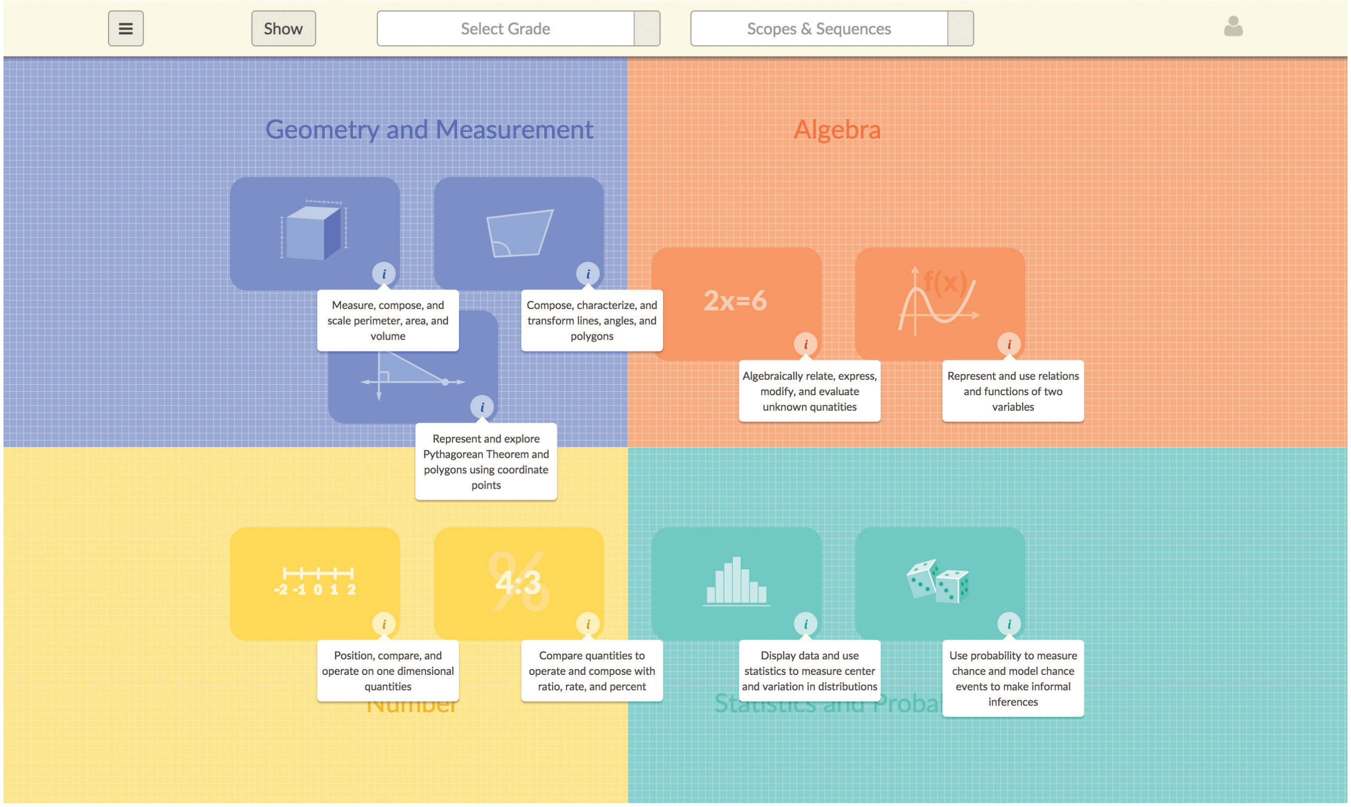

Show
Select Grade
Scopes & Sequences
Geometry and Measurement
Algebra
Measure, compose, and scale perimeter, area, and volume
Compose, characterize, and transform lines, angles, and polygons
2x=6
f(x)
Algebraically relate, express, modify, and evaluate unknown qunatities
Represent and use relations and functions of two variables
Represent and explore Pythagorean Theorem and polygons using coordinate points
-2 -1 0 1 2
4:3
Position, compare, and operate on one dimensional quantities
Compare quantities to operate and compose with ratio, rate, and percent
Display data and use statistics to measure center and variation in distributions
Use probability to measure chance and model chance events to make informal inferences
Number
Statistics and Proba

Figure 24. Nine big ideas (regions) within four mathematical fields for middle grades

[157] Confrey J. 2015. *Designing curriculum for digital middle grades mathematics: Personalized learning ecologies.* Plenary presentation for Mathematics Curriculum Development, Delivery, and Enactment in a Digital World, the Third International Conference of the Center for the Study of Mathematics Curriculum. Chicago, IL. Nov 7.

ratio relationships, before trying to compare ratios or build up to meeting values. The shape of a particular RLC–for example, key ratio relationships–also conveys suggested sequencing. Its shape as an inverted triangle conveys that users should begin with what it means for the ratio of two quantities to be equal, when there is more or less or both quantities. The parallel structure of the upper two vertices of the inverted triangle conveys that base ratio (lowest pair of relatively prime whole numbers) and unit ratio (where one of the values of the quantities equals one) can be learned in either order. By learning the ideas of equivalence, base ratio, and unit ratio before moving to comparing and finding missing values ensures more success as students learn to build up in a table or graph to find a missing value using the base or unit ratio, and eventually they learn to find a missing value directly through the application of a combination of multiplication and division. With this example of organization, one can see how the student-centered design of the map differs from a solely content-based logical analysis of mathematics, in that it is based on leveraging the research on student learning patterns.

At the next level of detail, the construct level, a user has access to what is called the "learning trajectory stack." The stack details the typical behaviors, conceptions, and language of children as they learn and revise ideas from naive to more sophisticated. In Figure 26, the first levels of the stack for unit ratio are shown. For a ratio where one number is a multiple of the other (12:3), one of the two unit ratios is (4:1), and these are the easiest for students to understand; for instance, in a recipe, 4 cups of flour per 1 cup of milk. At the second level, from a 4:1 ratio, they can reason to find the other ratio of (1:¼) or one cup of flour per one quarter cup of milk. At the next level, students can find unit ratios from base ratios, such as going from (2:3) to (1, 3/2) or (2/3:1). It is important for students to realize that either quantity in the ratio can become the "per one quantity." The next level is finding a unit ratio for a decimal or fractional value of one of the quantities (2.3:5) to become (23:5) and then, for instance (1, 5/23). Finally (non-visible in picture), a student can find the unit ratio for any ratio (a:b) as (a/b:1) or (1:b/a).

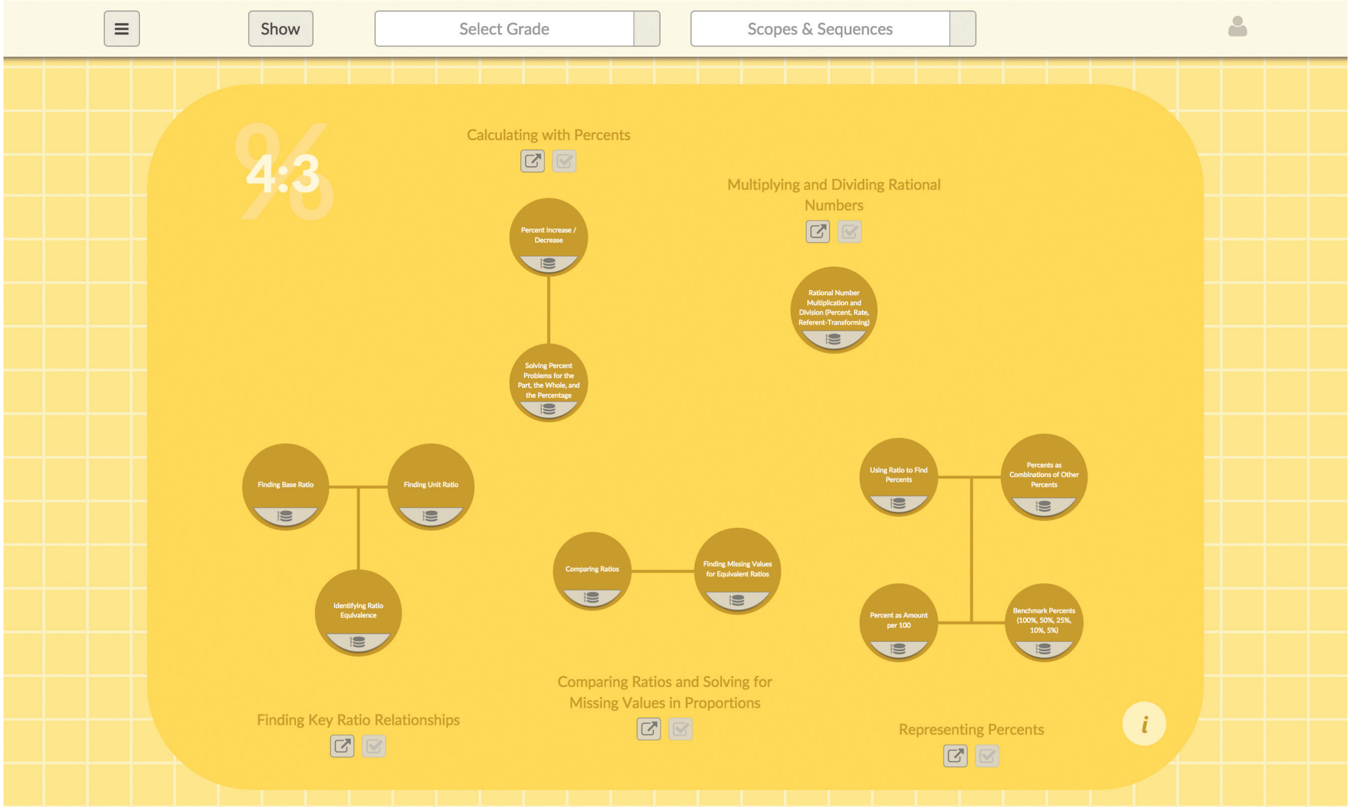
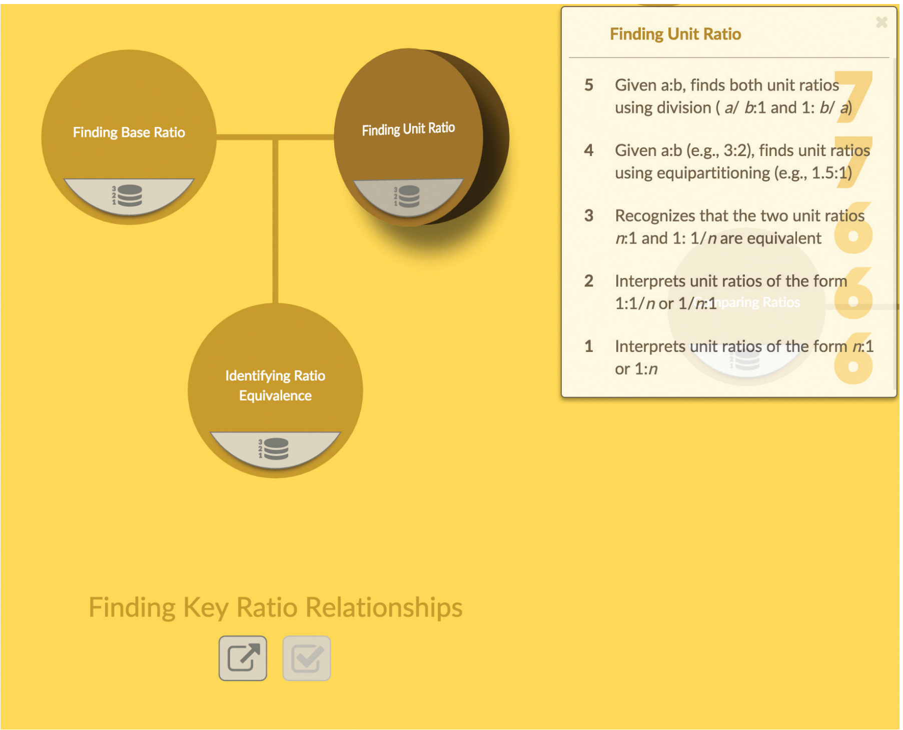

Show
Select Grade
Scopes & Sequences
4:3
Calculating with Percents
Percent Increase / Decrease
Solving Percent Problems for the Part, the Whole, and the Percentage
Multiplying and Dividing Rational Numbers
Rational Number Multiplication and Division (Percent, Rate, Referent-Transforming)
Finding Base Ratio
Finding Unit Ratio
Identifying Ratio Equivalence
Finding Key Ratio Relationships
Comparing Ratios
Finding Missing Values for Equivalent Ratios
Comparing Ratios and Solving for Missing Values in Proportions
Using Ratio to Find Percents
Percents as Combinations of Other Percents
Percent as Amount per 100
Benchmark Percents (100%, 50%, 25%, 10%, 5%)
Representing Percents
i

Figure 25. Five Related Learning Clusters (RLCs) for "Compare quantities to operate and compose with ratio, rate and percent"

Tapping on the symbol CCSS-M shows the standards that are related to this construct. Associating the standards with the constructs assures teachers that they are addressing the proper material, but as one can clearly see, the learning trajectory information is far more informative in terms of pedagogical content knowledge than are the standards, as should be expected.

Also choosing any particular standard, one sees flags in the places in the map where it plays a major role (in ratio, rate, and percent) and a minor role (in functions and relations). This way of creating a learning map, using a hierarchical structure and tying into big ideas permits a user, whether a teacher or a student, to comprehend the structure within which they are learning. It is in sharp contrast to learning systems that attempt to connect materials standard-by-standard. We claim that a standard-by-standard approach is ineffective due to the fact that standards can be mapped to multiple places on the map and they vary in grain size. Their systematic


Finding Base Ratio
Finding Unit Ratio
Identifying Ratio Equivalence
Finding Key Ratio Relationships
7
7
6
6
6
paring Ratios

Finding Unit Ratio

5 Given a:b, finds both unit ratios using division (*a*/ *b*:1 and 1: *b*/ *a*)

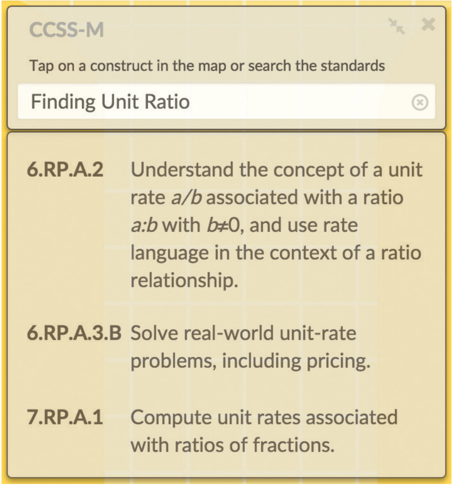
4 Given a:b (e.g., 3:2), finds unit ratios using equipartitioning (e.g., 1.5:1)

3 Recognizes that the two unit ratios *n*:1 and 1: 1/*n* are equivalent

2 Interprets unit ratios of the form 1:1/*n* or 1/*n*:1

1 Interprets unit ratios of the form *n*:1 or 1:*n*

Figure 26. A learning trajectories stack for the construct "Finding Unit Ratio"



CCSS-M

Tap on a construct in the map or search the standards

Finding Unit Ratio

6.RP.A.2	Understand the concept of a unit rate *a/b* associated with a ratio *a:b* with *b*≠0, and use rate language in the context of a ratio relationship.
6.RP.A.3.B	Solve real-world unit-rate problems, including pricing.
7.RP.A.1	Compute unit rates associated with ratios of fractions.

Figure 27. The Common Core state standards for the unit ratio construct

connection to big ideas is not sufficiently articulated in that the progressions remain implicit. We seek to make those relationships explicit and the basis of our assessment models.

The map offers other features to users. First of all, single or multiple grade levels can be selected to allow its flexible use across different grade configurations. Also, a scope and sequence generator is being constructed, so a school can map the regions to the days of instruction and can sequence down to the cluster level. Finally, also under construction, is a means to take short journeys from one cluster or construct to another, so that topics across the fields can be connected. The map provides insight into the underlying theory of learning and, as such, provides critical features that could be leveraged in the process of data mining.

DIAGNOSTIC ASSESSMENT AND REPORTING

Our map connects to a diagnostic assessment system at the level of the RLCs. When students complete the study of materials assigned at each of the constructs in a cluster, they take a diagnostic assessment of the RLC. By assessing at the cluster level, testing is given periodically but with sufficient frequency to provide useful diagnostic information and check connections and retention.

A unique quality of our assessment system is that the assessments are often designed to show the process of solving the problem and are evaluated to reveal the students' preferred method of solving a problem. For example, if a student represents univariate data in sixth grade statistics using the increasingly sophisticated elements of ordering, grouping, scale, and intervals, their progress to proficiency on this skill can be mapped. Many of our items use "item generation environments" that allow the systematic variation of the item parameters to show variations in processes across task classes[158].

The assessments are constructed through a process of "evidence-centered design"[159]. The value of a clear description of the student model in EDC was described as:

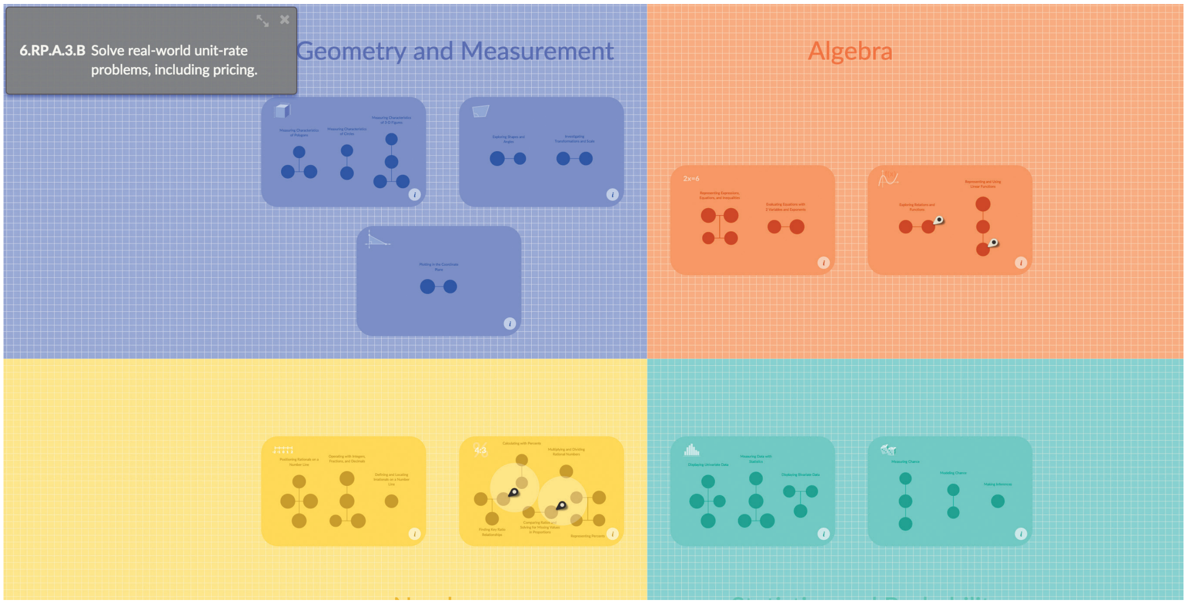

6.RP.A.3.B Solve real-world unit-rate problems, including pricing.

Figure 28. Showing the mapping of CCSS-M standards to clusters and constructs

[158] Confrey Jand Maloney AP. 2012. Next generation digital classroom assessment based on learning trajectories in mathematics. In Dede, C. & Richards, J. (Eds.) *Steps toward a digital teaching platform*. New York: Teachers College Press. (pp. 134-152).

[159] Mislevy R and Riconscente M. 2005. Evidence-Centered Assessment Design: Layers, Structures, and Terminology PADI Technical Report http://padi.sri.com/downloads/TR9_ECD.pdf

> We note that the patterns in data transcend the particular in which they were gathered in ways that we can talk about in terms of students' capabilities, which we implement as student model variables and organize in ways tuned to their purpose. Having the latent variables in student model as the organizing framework allows us to carry out coherent interpretations of evidence from a task with one set of surface features to other task that may be quite different on the surface. The machinery of probability-based inference in the evidence accumulation process is used to synthesize information from diverse tasks in the form of evidence about student capabilities, and quantifies the strength of that evidence. Psychometric models can do these things to the extent that the different situations display the pervasive patterns at a more fundamental level because they reflect fundamental aspects of the ways students think, learn, and interact with the world[160].

In our application of ECD, the assessment is tied to the levels in the learning trajectory stacks using an adaptive model. For each learning trajectory stack, the learning scientist marks levels that introduce a qualitatively different aspect of the big idea and writes an item to test for student understanding. When the levels below it are encapsulated in the tested level, an adaptive protocol ensures that test takers are correctly associated with levels. Items are carefully designed to generate diagnostic information. For instance, they capture evidence of commonly held misconceptions and/or strategies used to solve problems. This ensures that students and teachers are provided with appropriate feedback to inform next steps.

A strength and challenge of the system is that a diagnostic assessment can span across multiple grades of proficiency levels of the learning trajectories. Thus it can allow students to move more rapidly or slowly and signal users, including students and teachers, whether they are on track or above or below grade level in their learning levels.

Assessments can be used for pre-testing, practice testing, or as a diagnostic assessment level. Results of the assessments, with the exception of justifications, can be accessed immediately following the testing. Results are shown on the map to display both information on where a student has worked on activities and where they have shown proficiency with the materials. Students receive their individual data and can review their progress over time. Teachers can review either individual or class results, including analyzing those results by subgroups. While the current system is limited to diagnostic assessments at the level of RLCs, a standardized reporting system affords the use of other types of assessment including, for instance, perceptions of learning success and satisfaction.

An important characteristic of the assessment and reporting system in this DLS is that it provides a variety of types of feedback to students and teachers in a timely fashion. Feedback can be delivered as praise, as correctness, or as detailed evidence on process. It can be delivered immediately or delayed. Researchers have distinguished two broad classes: person-oriented and task-oriented[161]. It appears that while both can be important, the task-oriented feedback tends to show improved effects on the performance on a cognitive task. However, person-oriented feedback can support self-efficacy and improve a student's perception of themself as a motivated learner. An assessment system can deploy feedback in a variety of ways in order to permit experimentation on what produces the greatest gains in understanding.

LINKS TO CURRICULUM USE

The map can be linked to a curriculum by one of two methods. At the level of the RLC, one can select the relevant construct or constructs and then have a set of possible links addressing those topics become

[160] Mislevy R, Behrens J, Dicerbo Kand Levy R. 2012. Design and discovery in educational assessments: Evidence-centered design, psychometrics and educational data mining. *Journal of Educational Data Mining 4*(1): 1-48.

[161] Lipnevich Aand Smith J. 2008. Feedback: The effects of grades, praise, and source of information. June ETS RR-08-30.

visible. Teachers and schools can add links locally, but the overall map has links that are curated by the team. Contributions to the general map can be made on the basis of enough internal support via a teacher-to-teacher rating system. In addition, materials can be tagged based on a taxonomy of curricular features including whether the materials are problem-, project-, or practice-oriented; involve problem solving, group work, individualized activity and include or don't include formative assessment, etc.

Another means of accessing curricula is through a tool that permits a district to develop a scope and sequence. There are restrictions on those scopes and sequences to avoid over-fragmentation of the curricula. It requires the curricula designer to work across the year, sequencing first at the regional or "big idea" level and then within that, to sequence at the RLC level. Within a particular cluster, a curriculum designer can assign web resources and a student can work at the cluster level among those resources. In this scenario, a student can sign into the DLS by name and class and receive information about their assignments, expectations, and results.

A major challenge in the current instantiation of the digital learning system is how to get more substantive information from the students' experiences with the curricular materials. At this time, it is relatively easy to measure "time on task" and sequence, but to know whether the student completed the assignment and how well is beyond the current system's capability. One way to approach this problem is to set up a standardized means that designers of curricular materials could formatively assess student performance on their materials and pass these data back to the DLS's assessment system in a standardized way.

TOOLS AND WORKSPACE

Some DLS are comprised of only lesson tasks with problems asked and solutions submitted. However, to become a proficient mathematician, the CCSS-M recognizes the importance of developing a set of practices that describe how mathematics is done. One element of a sophisticated DLS is to offer a workspace with a variety of tools that can be a performance space for students, a canvas on which they can carry out and share their mathematical pieces of work, and then store and curate the resources from those experiences. To date, a number of exceptional tools exist, including DESMOS, Geometer's Sketchpad, Cabri, Fathom, Geogebra, and Tinker Plots. In addition, some tools exist for carrying out mathematical work and even for creating a screen capture of it. Few integrate the elements of a collaborative workspace, a tool set, and a means to create a portfolio or notebook, much less link them successfully to access to a database of tasks[162].

ANALYTICS

An analytic engine for our DLS will capture all the data about system use including, but not limited to, where a student has gone in the map, how long he or she has spent there during a session, what links were accessed and in what order, how many times a DA was taken and when, correct and incorrect percents, strategies used, and results on an item-by-item basis. Users can also see at what level of the stacks a learner is on and how quickly she or he is progressing relative to the time in the system. Because our current design does not capture the actual work a student does in a linked set of materials and we do not have the workspace or tools embedded in the system, limited information can be obtained on students' use of materials. Two variables that will be of prime importance will be those of time on task (ToT) and opportunity to learn (OTL). In the future, we hope to gather richer data on student activity either from what is done using the digital materials or from adding more opportunities for the capture of samples of student work or from behavior from teacher observations of classroom activity. Until these are available, connecting ToT and OTL measures with performance on the diagnostic assessments may prove insightful, especially concerning the navigational elements of the system. The harder problems of providing expert advice to the user of what to do

[162] Maloney APand Corley A K. 2014. Learning trajectories: A framework for connecting standards with curriculum. *ZDM*, 1–15.

following particular results on the assessments will likely be the most significant and essential challenge.

Mislevy, et al., warned that educational data mining (EDM) would benefit from considering how it links to the underlying cognitive models of the system it is mining, stating "It is easy to amass rich and voluminous bodies of low-level data, mouse clicks, cursor moves, sense-pad movements, and so on, and choices and actions in simulated environments. Each of these bits of data, however, is bound to the conditions under which it was produced, and does not by itself convey its meaning in any larger sense. We seek relevance to knowledge, skill, strategy, reaction to a situation, or some other situatively and psychologically relevant understanding of the action. We want to be able to identify data patterns that recur across unique situations, as they arise from patterns of thinking or acting that students assemble to act in situations. It is this level of patterns of thinking and acting we want to address in instruction and evaluation, and therefore want to express in terms of student model variables."[163]

With respect to the cognitive student model underlying the SUDDS DLS, our analytic model would be helpful if it could inform us of the degree to which we are able to achieve student-centered instruction. While the primary purpose of our work is to see students make progress on learning the big ideas successfully, as demonstrated by successful movement in the learning trajectory stacks and within the RLCs, a secondary purpose is for students to become self-regulating learners who are aware of their progress and able to make successful choices and collaborations towards learning and pursuing mathematics.

With this interpretation of their challenge set in the context of our work, I hope to have provided an example of future learning environments and how they can be understood as more than a delineation of a domain to be learned. Such student-centered models can, I hope, be considered and discussed at the upcoming conference. The iterative nature of the work supports the ability to get smarter as the system is built, but like an iterative function, converging to robust solutions also depends on beginning with a strong "seed." A student-centered DLS may provide such a seed. The question is: How can the empirical techniques of mining large-scale data provide insights into digital learning systems, and, in particular, how can they inform models of those systems with specific student models and an explicit purpose of strengthening student-centered learning?

IMPLICATIONS OF THESE LEARNING MODELS FOR FUTURE WORK IN ANALYTICS

The conference provided some insight into ways in which big data could inform, accelerate, and transform research on teaching and learning. A number of targets emerged that offer interesting contrasts to the current model of educational practice, largely driven by assumptions about classroom practice based on standards, ongoing classroom activities within school walls, and periodic forms of traditional assessment leading to grades and performance on high stakes testing. In contrast, foreshadowing new forms of learning and teaching, the conference participants offered numerous calls for:

- Increasing and diversifying the sources and types of assessment, including topics around identity, persistence, and socio-emotional states, or perhaps data from out of school settings such as makerspaces, instead of documenting only cognitive or academic accomplishments,

- Creating new ways to analyze students' acquisition of competencies for work and career and to document their continuous acquisition by students over time and place, and

- Responding to a ramped-up call for data-driven decision-making in real-time, not just as summaries and implications of assessment, but relative to patterns of activity and participation and with predictions

[163] Mislevy R, Behrens J, Dicerbo Kand Levy R. 2012. Design and discovery in educational assessments: Evidence-centered design, psychometrics and educational data mining. *Journal of Educational Data Mining 4*(1): 1-48.

Although most of these measures require new types of assessment, other examples presented at the conference pointed to coordinating existing data sources that can be assembled in new ways to yield insights, such as attendance, patterns of course taking, demographics, and records of participation in external activities. One participant, Kenneth Koedinger, captured this difference by describing his view of the dimensions of big data as "tall in participants," "wise in observations," "fine in frequency," "long in time span," and "deep in theory-driven relevance."

In addition to offering ways to leverage commonplace data to offer atypical solutions, the conference provided important emerging distinctions about learning. For some, online learning consisted largely of individual pursuit leading to ample data points. From these, opportunities emerge to make conjectures about learning rates, levels of participation, and content coverage. For others, this view of online learning was constrained and worrisome. Those participants argued that gaining insight into systems comprised of weak learning goals and a lack of interaction and collaboration would not substantially move the field forward, and would make new robust models of learning and teaching unlikely to emerge. For these second group members, digital learning could foster new forms of learning, where the richness of the possible environments would drive the records of data into new territories and challenges. Using big data with a solid innovative "seed" for teaching and learning and rich curricula would accelerate progress.

A major emerging issue is the question of how progress should be catalyzed and sequenced. Should the field continue to put the most of its efforts in a slow and gradual understanding of how to improve learning using digital resources and approaches reaching for the inclusion of big data as usage increases? Or should educational designers cease the steady slow progress to shift to a process of rapid prototyping of new data-rich tools that could revolutionize and transform education disruptively and discontinuously in time? Overall, this audience, having already utilized big data, argued that attention to providing adequate infrastructure, soliciting advice from experienced investigators, and sharing resources and protocols should begin as soon as possible.

It was clear across the days of the meeting that there are a number of shared commitments in infrastructure, which include providing open learning opportunities, experimenting with strong and weak models of explicit learning or bootstrapping empirically, working to provide rich personalization of paths and profiles of users, and creating dashboards that permit quick and flexible displays of different data representations and their comparisons. Other targets for creating increased resources for using big data more quickly and flexibly include providing incentives for sharing data, specifying the work flow of design and development teams that would help researchers organize, interpret, extract, model and visualize data, and incentives and opportunities to build new collaborations with high levels of trust around shared data to be parsed and diced into a variety of uses.

Privacy, Security, and Ethics

Elizabeth Buchanan (University of Wisconsin-Stout), Ari Gesher (Palantir), and Patricia Hammer (PK Legal)

BUCHANAN: THE ETHICAL REALITIES OF BIG DATA

Ethics is about what's possible and what's good or just. However, in our professional literatures and educational discourses, there tends to be more focus on compliance and restriction: What are the legal, technological, and economic constraints to our actions and decisions? Compliance is not ethics, and the goal of this thought paper is to encourage readers to move away from a prescribed and regulatory way of thinking about ethics and toward a more humanistic understanding of the ethics of technologies or, more specifically, the ethics of big data and the ethics of algorithms. I am concerned with the larger issue of harms that may result from algorithmic manipulation and the uses of big data. Redefining and appreciating the depth and variety of emotive harms is critical to the fields of big data science and analytics. Focusing on emotive harms allows us to talk about such complex issues as technological determinism, values in design, anticipatory ethics, and predictive design, among other ethical concerns.

As a relatively new field, data science is still in its infancy in terms of its values and ethical stances. As a profession matures, its values become more solidified for its professionals and evident to others influenced by the profession. Thus, a simple question arises: Where do data scientists, or those responsible for the creation, analysis, use, and disposal of big data, learn their professional ethics? The first Code of Conduct (not ethics) for Data Scientists[164] was released in 2013. It is unclear how many data science programs include any reference to the code of conduct, but a cursory review of the major big data analytics programs reveals few include ethics content.[165] Big data education focuses more on the technical, statistical, and analytic processes over the emotive, contextual, or values-based considerations with data. When do we consider the neutrality or bias of data? In the act of algorithmic processing, or manipulation, do data lose their neutrality and take on bias? And can data or an algorithm ultimately do harm?

Technology and information ethics considerations have long included such topics as access to information, ownership of information, copyright protections, intellectual freedom, accountability, anonymity, confidentiality, privacy, and security of information and data. Fields such as information studies, computer science, and engineering have grappled with these ethical concerns, and data science is now experiencing its own cadre of ethical concerns. Gradually, more attention is being paid to explicit and implicit bias embedded in big data and algorithms and the subsequent harms that arise. To this end, big data analytics should include methodologies of:

- Values-sensitive design,
- Community-based participatory design,
- Anticipatory ethics,
- Ethical algorithms, and
- Action research

These approaches situate our participants, actors, and users as central and informed, as empowered decision

[164] "DATA SCIENCE CODE OF PROFESSIONAL CONDUCT." Code of Conduct. Data Science Association, n.d. Web. 15 Sept. 2015. <http://www.datascienceassn.org/code-of-conduct.html>.

[165] Using two sources, "23 Great Schools with Master's Programs in Data Science." Top Schools for Master's Degrees in Data Science. N.p., n.d. Web. 15 Sept. 2015. <http://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science/>. and "Big Data Analytics Master's Degrees: 20 Top Programs - InformationWeek." InformationWeek. N.p., n.d. Web. 15 Sept. 2015. <http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-masters-degrees-20-top-programs/d/d-id/1108042>. curricular offerings were reviewed. Some, for example, Carnegie Mellon, includes an Ethics and Management course, or Maryland offers Business Ethics, while UC-Berkeley's unique in its Legal, Policy, and Ethical Considerations for Data Scientists course. The overwhelming majority of programs had no ethics content.

makers. Friedman states that "central to a value sensitive design approach are analyses of both direct and indirect stakeholders; distinctions among designer values, values explicitly supported by the technology, and stakeholder values; individual, group, and societal levels of analysis; the integrative and iterative conceptual, technical, and empirical investigations; and a commitment to progress (not perfection)."[166]

These approaches allow us to stimulate our moral imaginations and experience ethical opportunities in big data work while pushing the boundaries of our computational powers. The era of big data has been upon us for a number of years, and we've accepted the core characteristics of big data—velocity, veracity, volume, and variety—as the norm. We've accepted the ways in which we are targeted and identified through our big data streams and the ways algorithms silently (or, in many cases, not so silently) operate in the background of our daily technology-mediated experiences. Within these newfound strengths, algorithms, those processes or sets of rules followed in calculations or other problem-solving operations, seem smarter, faster, and more intentional. Big data and algorithms now tell us who is eligible for welfare, our political affiliations, and where our children will attend college. Today, "an algorithm is a set of instructions designed to produce an output: a recipe for decision-making, for finding solutions. In computerized form, algorithms are increasingly important to our political lives.... Algorithms...*become primary decision-makers in public policy*."[167]

Are we confident with big data and the ways in which algorithms make decisions? Are there decisions we would not defer to them? Recall the uproar about the Facebook Emotional Contagion study, when algorithms manipulated what news was seen by individuals on their news feeds. Using that experiment as an example, we can consider the differences between machine- and human-based decision making. "Our brains appear wired in ways that enable us, often unconsciously, to make the best decisions possible with the information we're given. In simplest terms, the process is organized like a court trial. Sights, sounds, and other sensory evidence are entered and registered in sensory circuits in the brain. Other brain cells act as the brain's "jury," compiling and weighing each piece of evidence. When the accumulated evidence reaches a critical threshold, a judgment—a decision—is made."[168] Consideration of risks and harms are part of the decision-making process, and we have an ability to readjust and change our decision if the risk-benefit ratio is out of alignment. "Scientists have found that when a decision goes wrong and things turn out differently than expected, the orbitofrontal cortex, located at the front of the brain behind the eyes, responds to the mistake and helps us alter our behavior."[169] But our human decisions are also affected by implicit and explicit biases and, to a great degree, "We are ruined by our own biases. When making decisions, we see what we want, ignore probabilities, and minimize risks that uproot our hopes."[170] When we consider big data analytics, we rely on probabilities, and we correlate data. The ethics of correlation and causation must be addressed in big data analytics. We can make the best and the worst out of data; algorithms can solve problems, just as they can cause them: "You probably hate the idea that human judgment can be improved or even replaced by machines, but you probably hate hurricanes and earthquakes, too. The rise of machines is just as inevitable and just as indifferent to your hatred."[171]

[166] "VSD: Home." Value Sensitive Design. University of Washington, n.d. Web. 15 Sept. 2015. <http://www.vsdesign.org/index.shtml>.

[167] Eubanks V. "The Policy Machine." Web log post. Slate. Slate, n.d. Web. 15 Sept. 2015. <http://www.slate.com/articles/technology/future_tense/2015/04/the_dangers_of_letting_algorithms_enforce_policy.html?wpsrc=sh_all_tab_tw_top&utm_content=bufferbab23&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer>.

[168] "Decision-Making." Web log post. BrainFacts.org. Society for Neuroscience, 9 Apr. 2013. Web. 15 Sept. 2015. <http://www.brainfacts.org/sensing-thinking-behaving/awareness-and-attention/articles/2009/decision-making/>.

[169] Ibid.

[170] Berman D K. "So, What's Your Algorithm?" Business Technology. Wall Street Journal, 4 Jan. 2012. Web. 15 Sept. 2015. <http://www.wsj.com/articles/SB10001424052970203462304577138961342097348>.

[171] Ibid.

To return to the concepts of harms generated from big data analytics, here are a few examples: A widow is continually reminded of her deceased spouse on birthdays, anniversaries, and special occasions, but she does not want to change the late partner's Facebook status as it will disrupt their shared experiences on Facebook. A young man is greeted by pictures of his apartment burning down, as one of the features in his "Year in Review." And, perhaps most well quoted, Erik Meyer has described his response to an algorithmically generated experience, calling it "inadvertent algorithmic cruelty":

> A picture of my daughter, who is dead. Who died this year.
>
> Yes, my year looked like that. True enough. My year looked like the now-absent face of my little girl. It was still unkind to remind me so forcefully.
>
> And I know, of course, that this is not a deliberate assault. This inadvertent algorithmic cruelty is the result of code that works in the overwhelming majority of cases, reminding people of the awesomeness of their years, showing them selfies at a party or whale spouts from sailing boats or the marina outside their vacation house.
>
> But for those of us who lived through the death of loved ones, or spent extended time in the hospital, or were hit by divorce or losing a job or any one of a hundred crises, we might not want another look at this past year.
>
> To show me Rebecca's face and say "Here's what your year looked like!" is jarring. It feels wrong, and coming from an actual person, it would be wrong. Coming from code, it's just unfortunate. These are hard, hard problems. It isn't easy to programmatically figure out if a picture has a ton of Likes because it's hilarious, astounding, or heartbreaking.
>
> Algorithms are essentially thoughtless. They model certain decision flows, but once you run them, no more thought occurs. To call a person "thoughtless" is usually considered a slight, or an outright insult; and yet, we unleash so many literally thoughtless processes on our users, on our lives, on ourselves.[172]

Reputational harms, or informational harms, are often touted as the only real risks in big data analytics. These examples are related to privacy invasions, but are different. These experiences are not of that quality. "These abstract formulas have real, material impacts."[173] They are emotive harms, and recognition of these types of harms must occur at the design and implementation stage of analytics and big data.

What would ethical algorithms do differently? How can we ensure our work with big data is ethically informed? Jeremy Pitt of Imperial College is working on ethical algorithms: "One is about resource allocation, finding a way an algorithm can allocate scare resources to individuals fairly, based on what's happened in the past, what's happening now and what we might envisage for the future.... Another aspect is around alternative dispute resolution, trying to find ways of automating the mediation process.... A third is in what we have called design contractualism, the idea that we make social, moral, legal and ethical judgments, then try to encode it in the software to make sure those judgments are visually perceptive to anyone who has to use our software."[174]

From a harms perspective, the lack of transparency in big data analytics is concerning. "Computer algorithms can create distortions. They can become the ultimate hiding place for mischief, bias, and corruption. If an

[172] Meyer E. "Inadvertent Algorithmic Cruelty." Thoughts From Eric Inadvertent Algorithmic Cruelty Comments. N.p., 24 Dec. 2014. Web. 15 Sept. 2015. <http://meyerweb.com/eric/thoughts/2014/12/24/inadvertent-algorithmic-cruelty/>.

[173] Eubanks V. "The Policy Machine." Web log post. Slate. Slate, n.d. Web. 15 Sept. 2015. <http://www.slate.com/articles/technology/future_tense/2015/04/the_dangers_of_letting_algorithms_enforce_policy.html?wpsrc=sh_all_tab_tw_top&utm_content=bufferbab23&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer>.

[174] Say M. "The Quest for Ethical Algorithms." CIO UK. CIO, 23 June 2014. Web. 15 Sept. 2015. <http://www.cio.co.uk/insight/compliance/quest-for-ethical-algorithms/>.

algorithm is so complicated that it can be subtly influenced without detection, then it can silently serve someone's agenda while appearing unbiased and trusted.... Whether well or ill intentioned, simple computer algorithms create a tyranny of the majority because they always favor the middle of the bell curve. Only the most sophisticated algorithms work well in the tails."[175]

To an ethical end, Eubank[176] recently recommended four strategies:

1. We need to learn more about how policy algorithms work.
2. We need to address the political context of algorithms.
3. We need to address how cumulative disadvantage sediments in algorithms.
4. We need to respect constitutional principles, enforce legal rights, and strengthen due process procedures.

As we continue to explore the potential and boundlessness of big data, and increase our analytical and computation powers, ethics must be at the fore of our advances, not an inadvertent afterthought.

HAMMER: IMPLICATIONS OF AND APPROACHES TO PRIVACY IN EDUCATIONAL RESEARCH

Changes in research opportunities and independent review board (IRB) oversight have changed the way that social sciences research in general, and education research in specific, are being implemented. Fears about privacy is leading to the stifling of public education research and pushing research into corporate hands where transparency is less required and compliance is easier. This change poses risks to educational research, which is often the source of new ideas implemented in public institutions and across socioeconomic levels.

Improved technologies now allow educators to conduct research in ways never before possible, such as the use of big data, in-home or in-classroom audiovisual recordings, or biometric stress indicators. Also, data can be collected simultaneously from across the world and analyzed across hundreds of potential variables. IRBs, parents, and subjects may be concerned with the risks posed by these technologies, but there are security approaches to address each of the risks posed. By identifying privacy concerns and risks, researchers can build safeguards into their research to minimize risks and more easily have research approved. If the combination of research and safeguarding can be pre-approved by a reputable, knowledgeable, and accountable institution, such as the U.S. Department of Education (DOE), this will benefit IRBs as they will have clear guidelines to follow and standards they can rely upon. Therefore, researchers will have less difficulty getting projects approved by IRBs. Society will also benefit by increasing the amount of research done in educational and research facilities as opposed through non-transparent commercial processes.

Educational research is noninvasive, and the greatest perceived risk is often a privacy risk. IRBs generally do not include a privacy or technology expert, and the IRB may perceive a risk to be greater than it is because no member possesses the relevant expertise. Often this leads to delay, indecision, or overly conservative restrictions being placed on the researchers. One often proposed solution is to include privacy experts and security technology experts as part of the IRB. This solution can be difficult, based on the limited number of available experts and, in my belief, dilutes the purpose of the IRB, which is to evaluate the risk posed by the research. Instead of having one or two members be experts on privacy and/or technology, DOE or another third-party organization could develop a set of baseline standards for privacy and system protection in educational and/or other types of research. A project

[175] Ibid.

[176] h Eubanks V. "The Policy Machine." Web log post. Slate. Slate, n.d. Web. 15 Sept. 2015. <http://www.slate.com/articles/technology/future_tense/2015/04/the_dangers_of_letting_algorithms_enforce_policy.html?wpsrc=sh_all_tab_tw_top&utm_content=bufferbab23&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer>.

could demonstrate that it met the minimum standards before IRB review, which would inherently expedite the process and limit or eliminate the institution's liability in the event of a privacy breach.

Each 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, controlling 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 U.S. 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. A risk approach allows for changes in the types of research being done and the range of safeguarding solutions that could be applied. This would provide a framework to allow the newest research into privacy practices, security approaches, and research methodologies to be evaluated for how they mitigate risk, and to reuse those evaluations across the research community. Standardization and reuse would minimize the cost of evaluation while increasing the quality of evaluation. The IRB could still be the organization's voice in determining acceptable risk, but would be addressing these questions from a position of knowledge.

EVOLVING RESEARCH CAPABILITIES AND PRIVACY ISSUES

It is critical that any standard be developed with an ongoing evaluation function. This function must allow for new research in privacy and new approaches to research. In the field of privacy, continued research is needed in identifying new threats, vulnerabilities, and mitigation approaches.

DATA AGGREGATION AND MAINTAINING LARGE DATA SETS

People's ubiquitous use of the Internet has led to an explosion in the amount of commercially available data concerning individuals. Although this data is available for a price, many research organizations have shied away from maintaining large, aggregating data sets. The concerns about maintaining the data are often not weighed against the benefits to research that might be possible through maintaining long-term, large-population data sets (e.g., quicker access to the data for other related studies, and the ability to execute longitudinal studies). In the commercial environment, the cost benefit is often easier to understand and document than the societal benefits produced by the public researcher.

BIG DATA

Using and compiling big data can allow researchers to see trends or anomalies across a wide spectrum of individuals. Researchers can take data trends from persons they have never met and analyze the data to find trends based on age, race, income, location, level of education, time of day, physical activity, physical traits, and so on. In education, researchers may be able to correlate math scores with scores in other subjects, such as science and music, to identify a possible causation. Or make determinations of how someone best learns in order to develop a more personalized learning plan.

Big data also brings in the possibility of found data. In contrast to researcher-designed data, which are data sets of information collected according to a defined protocol by private and government sector agencies, the big data collectors are not research organizations. They usually collect the data as an auxiliary function of their core business. They use the data to improve business processes and to document organization activities. Social scientists have become interested in these data because it is a) timely, often real-time documentation of behavior, b) collected on large sets of individuals, yielding massive data sets, c) relatively inexpensive to acquire, and d) relevant to behaviors that are of common interest to social scientists. This data is growing due to social media, wearable technology, and Internet-connected sensors that collect and store data. The Internet has spawned new businesses that actively collect detailed attributes about their customers. Indeed, for many of these businesses, the personal data resource is their business.

However, these data are often limited in their attributes. In order to enrich the set of attributes to be studied, education research often uses these massive but lean data sets in combination with another source of data (e.g., demographic data on geographical units based on census and other measurements). Indeed, companies that assemble these data sources into unified data sets are popular sources of marketing data on individuals and households.

MINIMIZATION AND DE-ANONYMIZATION

The trend in data privacy is to minimize the amount of data collected, which would reduce the risks of de-anonymization to which subjects are exposed. However, with the growth of big data and associated analysis techniques, the validity of anonymization is being questioned. In light of this, there must be careful consideration, lest the data set suffer over-minimization, which could expose more subjects than necessary to privacy risks. Using big data as an example, hundreds of variables could be collected in one study over three years that tracks a student's progress. If the researcher is focused on math performance by geographic location, a later researcher may want to use the same data to correlate performance trends over the three years, or performance by gender, or math performance with music performance. By being able to use the same data set with anonymized data, researchers have limited the risk to the 10,000 students involved. If an IRB told the researchers to only collect data absolutely necessary for the study, subsequent researchers may need to conduct the same type of testing using different subjects to obtain a variable not collected the first time. This approach would expose 20,000 subjects to a possible risk instead of 10,000. By not collecting a variable, there could be a trend or correlation the researchers are missing that could otherwise innovate education.

Minimization of data collection should exclude personally identifiable information not necessary to a study, while including information that may be helpful to that or future studies. Educational research probably does not need a student's Social Security number, street address, or fingerprint, but ethnicity, age, and native language may be generally extremely useful. By reducing the number of times that similar studies must be conducted, researchers can limit the overall risk to any group of students, not just the students in the original study. Additionally, if the same research can be used repeatedly but analyzed in different ways, then subsequent studies do not need IRB approval because the data collected does not affect any new subjects in a new way.

Overall, there are risks to society from making it difficult to study educational impact. The delay in research, students opting out of research that may not be approved, and studies that never take place diminish our knowledge. We lose the opportunity to obtain great strides in education that a more personalized learning program may allow. We also push research into the hands of commercial entities that need not be transparent and compliant in their testing. By developing a risk-based standard approach to privacy and information security in the social sciences, we create a community that can better leverage the available data, seize research opportunities, and share knowledge.

GESHER: PRIVACY RISK (PERCEIVED AND ACTUAL) AS AN IMPEDIMENT TO DATA-INTENSIVE RESEARCH IN EDUCATION

We are in an era when more and more of our instructional materials and student metrics are becoming digitized–the data exhaust of schooling has grown to be substantial. It's reasonable to believe that significant useful insights can be gleaned in properly composed educational data sets. However, educational data, due to its use of children as subjects, is fraught with worries about its misuse, abuse, and theft. These persistent anxieties have created an atmosphere of paralysis, with many data owners preferring not to share data rather than incur the risk of their data being abused.

At the same time, our world is awash in anecdotes–some public, many private–of educational data being shared with seemingly zero thought about the security and privacy of the data subjects therein. Lost unencrypted laptops, emailed spreadsheets, default passwords, and generally lackadaisical security practices are not uncommon in the educational arena.

The main thrust of objections looks something like this:

1. Here is a theoretical privacy harm that, even given careful anonymization in shared data, could be achieved by a determined attacker.
2. Once data is shared, there is no control over its use or transfer.
3. Given the first two arguments, no assurances can be made about protection from privacy harms.

So, what would it take to create an atmosphere where the anxiety about privacy risks is reduced to the point where data can be shared among institutions and researchers for the betterment of education while, at the same time, increasing the overall safety and privacy of the students about whom the data is recorded?

What's needed is a set of agreed-upon set of best practices: both policy prescriptions and technical architectures. Best practices that can assure the data owners that data can be shared with a reasonable expectation of safety and privacy. Simultaneously, adherence to the same set of standards will raise the overall level of data safety across the field of the education.

The recipe looks something like this:

1. Research: enumeration of risks and in-depth threat modeling to create a shared taxonomy of reasonably likely potential harms that could result from the sharing and combining of educational data sets
2. Policy and standards: the matching of those enumerated risks to a set of existing data-handling practices designed to mitigate those harms
3. Technical implementation: the technical piece can come in two distinct forms: the first is the creation of cloud-based data-sharing environments that have carefully implemented safe data-sharing environments. The second is boilerplate technical specifications that product and service vendors in the educational space can be compelled to adopt

These three steps should create an environment that gives data owners the assurances they need to be comfortable making their data sets available for inclusion into data-intensive research efforts.

THREE KEY TECHNIQUES TO REMOVING PRIVACY PARALYSIS

THREAT MODELING

The reasons for paralysis around data sharing in education have much to do with a focus on the possible harms rather than the likely harms. In general, risk is composed of a combination of the probability of some harm occurring and the cost incurred if and when that harm does occur. There needs to be a shift away from focusing on the stakes and a better comprehension of the probability to understand the what and how of safely sharing educational data. This sort of risk analysis is known as threat modeling and comes to questions of data privacy via the field of computer security.

Computer security and privacy protections are related but distinct practices. Security is concerned with stopping unauthorized access to systems and data, and it relies on both technical access controls and active oversight to accomplish that goal. By contrast, privacy controls are about stopping unauthorized use of data by authorized users. Effective privacy controls also require a mix of technical access controls and active oversight to prevent abuse. In fact, privacy controls can often be thought of security mechanisms applied against a different outcome. And in practice, privacy's foundation is effective security; it's a simple thought exercise to realize that absent or easily circumventable security controls make the addition of privacy controls a moot point.

The field of computer security has evolved significantly in the past two decades, and as more and more systems were connected to the Internet, the criminal and political value of compromising security increased in lockstep with the modern world's growing dependence on information systems, and the level of sophistication of both attackers and the systems themselves have increased.

In the nascent days of computer security, securing a system was largely viewed as a black-and-white affair. From a design perspective, the practice was concerned with creating Maginot Line-like fortifications, with systems being designed to be unbreachable. This philosophy suffered the same fate as the Maginot Line during World War II as compromise techniques were perfected, often routing around the hardened parts of systems to find side-channel attacks that would subvert a system's lower layers or weak points. As computer security professionals learned more about the real-world problems encountered in their work, the philosophy of security moved from one of a *fortification* to one of *mitigation*. It was no longer assumed that a system could not be breached because of security measure x, y, and z, but rather that systems should be designed to quickly detect failure and enable a rapid response.

This sea change is probably most visible in Bruce Schneier's *Secrets & Lies*. In it he tells the tale of starting his career as a cryptographer (the epitome of hard, technical access controls) and detailing his journey into the world of security in general. The book contains a meditation on physical security and the design of safes and vaults. There's a simple design aesthetic in the world of safe design: make the safe more expensive to crack than the value of what it protects. This relativism lies in stark contrast to the absolutism of building "uncrackable" computer systems. And so the idea of threat modeling was introduced into computer security: namely, that building effective security requires an understanding of the value of what's being protected and to whom it is valuable. Understanding the potential attacker and their motivations greatly informs the process designing effective and usable security.

PRIVACY CATCHES UP

The early days of privacy engineering were similarly focused on seemingly unbreakable technical controls, using techniques like anonymization, aggregation, and deresolution to create anonymized data sets that could be shared with researchers. In a world of barely networked, expensive, and slow computers, these were often very effective controls. But the rise of low-cost, high-performance computing and proliferation of data science techniques around inference have moved each of these techniques from the tried, true, and trusted column into the doesn't-work column.

So in a world where none of the anonymization techniques are foolproof, has the ability to share data sets been destroyed? Instead of closing up shop, the privacy world has begun its own evolution into the world of threat modeling, moving from a model of *potential privacy harm* to one of *likely privacy harm*. Paul Ohm lays out this shift in his seminal paper on the subject, "Sensitive Information" *(Ohm, Paul, Sensitive Information (September 24, 2014). Southern California Law Review, Vol. 88, 2015, Forthcoming):*

> Computer security experts build threat models to enumerate and prioritize security risks to a computer system. Although "threat modeling" can mean different things, one good working definition is "the activity of systematically identifying who might try to attack the system, what they would seek to accomplish, and how they might carry out their attacks." Threat modeling is brainstorming about a system trying to find ways to subvert the goals of the system designers.
>
> Everybody builds threat models, even if only informally. Adam Shostack gives as an example the highway driver, working out how fast and recklessly to drive, factoring in the "threats" of the police, deer, or rain. But computer security experts have developed tools and formal models that make their threat modeling seem very different from everyday threat modeling, deploying tools and complex methodologies, building things called "attack trees" and murmuring acronyms like STRIDE and DREAD. But the formal veneer should not obscure the fact that

threat modeling is just brainstorming for pessimists; when done well, it is a formal and comprehensive game of "what-if" focused on worst-case scenarios.

Computer experts have used threat modeling techniques primarily to assess and improve security, not privacy. They build threat models to identify and prioritize the steps needed to secure systems against hackers, scammers, virus writers, spies, and thieves, for example. Recently, scholars from opposite sides of the law-technology divide have begun to adapt threat modeling for privacy too. From the law side, scholars, most prominently Felix Wu, have talked about building threat models for privacy law. Wu has constructed the unstated, implied threat models of existing privacy law, trying to study statutory text or common law court opinions to reveal the implicit threat lawmakers and judges held in mind when they created the laws. He also tries to find "mismatches," where the implicit threat model of a privacy law does not seem to address real world concerns.

From the technology side, computer scientists such as Mina Deng and Adam Shostack have asked whether the rigorous and well-documented threat models for security might be extended to privacy. Just as we build attack trees to decide where to prioritize scarce computer programming resources to shore up security, so too can we build attack trees to decide how best to tackle possible privacy harms.

ACTIVE OVERSIGHT

In the privacy domain, it's well understood that any access to data represents some level of privacy risk. At the limit, the only safe data is data that no one can access or use in any way–

clearly an extreme and absurd stance to take around the utility of data, but one that is still very common in the world today. In the educational domain, some of the potential harms cannot be mitigated through policy and technical controls alone, so addressing those concerns will require the adoption of active oversight as a core tenet of data-intensive educational research. It's only through the use of active oversight that any assurances can be made about the integrity of a system designed to preserve privacy. Active oversight acts as the final bulwark against abuse, the way to protect against the harms that, users could potentially perpetrate.

Technical measures like secure systems, encryption-at-rest, and anonymization are used to reduce the window of possible harm coming from the access and sharing of data. Contrast that with active oversight, which consists of teams of auditors looking, ex post facto, for particular patterns of use by those with access to data, which indicate attempts to circumvent privacy policies. Some examples:

- In a system designed to prohibit the wholesale export of data, a pattern of queries indicating a user is attempting to methodically return every record in the system.
- In a system designed to mask individual identities through the use of aggregates, a pattern of carefully constructed aggregate calculations that can be intersected to disaggregate individual identities.
- A pattern of queries that lie far outside the declared domain of interest for a particular user.
- The import and integration of identified data sets against anonymized data sets that could easily be used for the unmasking of identities.

While the creation of active oversight requires instrumentation, policy, procedures, and staffing around the handling of educational data, it may be the only way to mitigate real harms that can arise from the sharing of data for research purposes. Furthermore, active monitoring for privacy abuse may be the necessary bar to assure data owners that it is safe to share with research efforts.

MANAGED CLOUD ENVIRONMENTS FOR DATA ANALYSIS

The simplest method of sharing data is to provide a wholesale copy of the data set. It's possible to filter the data set to make it harder to identify individuals in the data set. However, modern privacy researchers have shown that most forms of data set anonymization can be pierced by integrating identified data sets and then applying careful data science across the two to unmask the individuals in the data set.

However, if the work of researchers with the data can be observed and monitored by a privacy oversight team, the use of these techniques is easy to detect. Building that sort of surveillance and monitoring infrastructure, as well as the expertise to adjudicate behavior using that data, is a non-trivial but tractable problem.

If a safe data-sharing arrangement requires that work takes place inside of such an environment, it makes sense to centralize access to the data in a managed environment. Modern cloud-hosting environments are a cost-effective way of creating such environments. It's imaginable that a data-sharing and analysis environment would be owned and operated by a central trusted authority. An environment like this would not only include locales to place data sets, but also the analysis tools, instrumented for auditing, for researchers to work with the data.

Therefore, data sharing would become a two-step process:

1. Data owners would no longer grant access to or copies of data sets to researchers. Instead, data owners would make data sets available to a cloud data-sharing and analysis environment.
2. Researchers would apply for access to the data sets they need to do their research, including importing their own private data into the cloud environment.
3. All data manipulation and analysis would take place in a managed and audited environment.

The final upshot to this architecture is that the confluence of various researchers working in the same environment could lead to greater collaboration and cross-pollination among research teams.

Summary of Insights

Elizabeth Burrows (AAAS S&T Policy fellow at NSF), Lida Beninson (AAAS S&T Policy fellow at NSF), and Chris Dede (Harvard University)

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.

Numerous challenges must be overcome to realize this potential. The introduction summarizes some necessary steps, among many: developing theories on what various types of data reveal about learning; resolving issues of privacy, security, and ethics; and building computational infrastructures, tools, and human capacity to enable educational data science. Other parts of the report describe re-conceptualizing data generation, collection, storage, and representation processes; developing new types of analytic methods; building human capacity; and making advances in privacy, security, and ethics.

INSIGHTS ABOUT DATA-INTENSIVE RESEARCH FROM THE SCIENCES AND ENGINEERING

The sciences and engineering have made progress in developing models for how to apply data science in their fields; this report documents that insights from STEM fields can offer guidance for the evolution of data-intensive research in education. In particular, the section on insights from the sciences and engineering, describes models that the sciences and engineering are using to conduct data-intensive research. Building on these initiatives outside of education can empower improvements in data-intensive research in teaching, learning, and schooling.

Insights of particular relevance to education from data-intensive research in STEM fields include:

- **Collaborate.** 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 the common language among research groups provided by ontologies, the development of standards and shared infrastructure for data storage and data analysis is key to interoperability. 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 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.** As discussed on pages 13-22, data-intensive research strategies effective in the five STEM cases in the first workshop provide insights for educational researchers who face similar challenges in the nature of the data they collect and analyze.

As documented in sciences and engineering cases previously discussed, 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 interoperatibity; 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. The section below describes 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

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. As an example, in the section on using predictive models in higher education, on pages 24-25 Yaskin provided an example for Hayes' concept of digital engagement (2014): "the use of technology and channels to find and mobilize a community around an issue." Yaskin writes,

> To facilitate a digital engagement strategy, higher education institutions can: (1) leverage an enterprise success platform (e.g., the Starfish platform) to analyze student performance data using predictive analytics, (2) solicit feedback from key members of a student's success network, (3) deliver information to the right people who can help the student, and (4) collectively keep track of their efforts along the way—all of which leads to a continuous, data-informed process-improvement cycle.

For each type of data discussed in the report (e.g., MOOCs, games and simulations, tutoring systems, and assessments), briefing paper authors and workshop participants identified important educational issues for which 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.

A breakout group at the second workshop centered on building producer/consumer partnerships as a method of mobilizing communities. 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, as seen in the Zooniverse example presented by Lucy Fortson during the first workshop, when citizen scientists are employed, 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. For example, when data are produced via crowdsourcing, there is the benefit of "free" and ample data production, but–in order for the findings to be trustworthy–many more replicates are required, as seen in Galaxy Zoo (galaxyzoo.org), where the public was invited to classify galaxies, and each galaxy was classified 35 times. As another illustration, when data are reused for a purpose entirely different from the one for which they were produced, this optimizes the value of the data and most efficiently increases knowledge, but there is a greater chance that the data could be misunderstood and thus misused, either due to insufficient metadata or insufficiently trained researchers.

There are many examples of widely used public databases in the sciences and engineering, such as OpenTopography for LiDAR topography data, iPlant for several types of plant genomics data, and, on the horizon, the data produced by the Large Synoptic Survey Telescope, which will constitute a publicly available 20–40 petabyte database catalog. In each of these examples, the type of producer/consumer relationship is similar; the consumer accesses a repository that also provides analytic tools and uses the data in the way that pertains to their application.

In education, there are public service repositories and analytic services as well (e.g., EdWise, an online tool that accesses 14 million K-12 education records from Missouri and allows them to be easily analyzed; for more details, see "Integrating Data Repositories"). However, as with other disciplines, several different types of data producers and consumers exist. For example:

- *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," using resources such as the Data Quality Campaign (see "Privacy, Security, and Ethics"). Central to this discussion is the level of data that should be shared. Data scientists need to be trained to be highly technical *as well as* highly ethical.

  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 such different goals that they essentially speak different languages. 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. To begin bridging that gap, the Department of Education published an Ed Tech Developers Guide (http://tech.ed.gov/files/2015/04/Developer-Toolkit.pdf) to help funders choose what to support. Further work needs to be done on determining the breadth of applicability of certain research findings, because each set of students, from each unique zip code, may respond differently, based on factors yet to be determined. Actual applicability can only be determined when models elucidate the true causal relationships, such that the process is understood, not simply the outcome. Also, it is important that entire platforms are not created from just a few trends. For the response variables, such as ROI or student success, to be optimized, they must be carefully chosen and optimized in an ethical fashion, for the right reasons.

- *Faculty as producers*. Another complex type of educational data produced are teacher evaluations, which are mainly "consumed" by deans and their administration. 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 are changing over time; this is analogous to anticipating changes in wheel technology while simultaneously building a hover craft.

- *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.

  Even with consumers' needs defined, constant interaction between producers and consumers (or the blurring of the lines between producers and consumers) is important in co-developing data "products," since consumers may not know the extent of what is possible to produce. One example where common definitions are needed is in the Department of Education's Innovation Clusters, where assessment of the current clusters is necessary to identify gaps to be filled by new clusters. Another aspect to consider when creating common definitions in data-intensive education research is including not only traditional primary, secondary, and post-secondary education, but also the education of technicians and individuals in jobs that require specific onsite training. Common definitions need to address many types of non-traditional education environments, such as part-time students, online learners, and onsite learners.

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. If not done this way, there could be a tragedy of the commons situation for the types of data-intensive research that would be most beneficial. Anticipating this 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. Briefing paper authors and workshop participants frequently noted the importance of infusing new forms of evidence throughout educational decision-making systems. As an illustration, in the section on using predictive models in higher education, on pages 23-24, Wagner discussed:

> ...augmenting and expanding [institutional research] beyond descriptive reporting and clusters of small n experimental and quasi-experimental designs. Instead, the great opportunity for data-intensive research is to help educational stakeholders make better decisions, obtain deeper and better insights, and to find new patterns that can help provide new understandings about learning and cognition through predictive and prescriptive analytics, actively seeking out risk and actively mitigating risks through targeted treatments, interventions, and supports. Predictions are of greater institutional value when tied to treatments and interventions for improvement and to intervention measurements to make sure results are being delivered.

The importance of making the outputs from these systems more helpful for their users through better visualizations and embedded workflows is also discussed, as is connecting back findings to instructional methods.

In the section on MOOCs, on page 32 Mitros delineated the types of big data in education that potentially provide a variety of opportunities:

I. Individualizing a student's path to content mastery, through adaptive learning or competency-based education.

II. Better 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.

III. Targeting interventions to improve students' success and reduce overall costs to students and institutions.

IV. Using game-based environments for learning and assessment, where learning is situated in complex information and decision-making situations.

V. A new credentialing paradigm for the digital ecosystem, integrating micro-credentials, diplomas, and informal learning in ways that serve individuals and employers.

VI. 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.

As articulated various workshop discussions, each of these types 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.

Later in the MOOCs section, on page 36, Ho discussed the effects of predictive models on decision-making:

> In any formative educational process, the criterion for prediction is not accuracy, as measured by the distance between predictions and outcomes. Instead, it is impact, as measured by the distance between student learning with the predictive algorithm in place, and student learning had it not been in place. I find the emphasis on technically

> sophisticated predictive models and intricate learning pathways to be disproportionate, and I think there is too little attention to rigorous experimental designs to ascertain whether students and instructors can use these tools to increase learning. In short, we want educational predictions to be wrong. If our predictive model can tell that a student is going to drop out, we want that to be true in the absence of intervention, but if the student does in fact drop out, then that should be seen as a failure of the system. A predictive model should be part of a prediction-and-response system that a) makes predictions that would be accurate in the absence of a response and b) enables a response that renders the prediction incorrect.

This is a substantial shift from current practices in using data for educational decision-making.

As an illustration of this theme, a breakout group at the second workshop posited that most forms of teaching could benefit from data analytics. Data analytics can be used on the small scale, to provide real-time feedback within one classroom, or on the large scale with studies that show, for example, that two lectures and one lab per week generally provides better outcomes than three lectures (a study by the National Center for Academic Transformation, www.thencat.org). This study received wide adoption, and it would be useful to determine what was done to diffuse this work, because a main goal for data-intensive research and the production of new tools is to increase their implementation.

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 professors 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, such as Continuous Formative Assessment (CFA), benefits both students and teachers, with continuous data-based improvement for classroom management and teaching, often coupled with a decrease in workload for the professors. One successful example is **A**ssessment and **LE**arning in **K**nowledge **S**paces (ALEKS), which are web-based adaptive-learning math and sciences courses. Publishers are getting $1 billion-a-year for the use of ALEKS to supplement K-12 and college math courses. One risk with the personalization of education is a loss of discourse, so it is important that humans are in the loop; for example, to confirm software-suggested student pairings.

This diffusion of innovation needs to be both top-down (new technologies produced and implemented) and bottom-up (professors 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. In one example, four recent graduates were paired with middle school teachers to research evidence-based content and provide it to the teachers to use, and 75-85 percent of the teachers participated, simply because they had a connection with the recent graduate who was helping them. Another example is a case where schools hired counselors to keep students engaged rather than hiring more teachers. Alternative to these examples, CourseSource is a successful online resource for professors teaching biological sciences, but it does not have the direct human component, although teachers generally take recommendations on new tools from other teachers or individuals they trust. Therefore, increasing interactions to share best practices among teachers, researchers, and practitioners would be highly beneficial.

Networked improvement communities in education (Bryk et al. 2011), proposed after examining successful large networks in other disciplines, may have the potential to bring about large-scale change. For example, Tony Bryk at the Carnegie Foundation for the Advancement of Teaching has shown a three-fold success rate in developmental math completion from using networked improvement. But with 1 million students per year in developmental math, there may need to be a business model for funding the large interdisciplinary teams necessary to take on the task of bringing that innovation to scale.

Another method to increase adoption of evidence-based techniques in higher education in STEM fields is to focus on Ph.D. students who are interested in becoming professors. Often these students want more teaching skills, but are advised against diverting time from their dissertation research. The Ph.D. students in STEM fields who are interested in becoming professors are essentially pre-service college teachers, and they certainly have the quantitative skills and interest to implement data-analytics techniques and tools. Even though only about 10 percent of Ph.D. recipients will become professors, and the average age at which one receives tenure is currently about 53, the skills offered in this type of training would be broadly applicable. The core competencies required for teaching largely overlap with those required for effective leadership, so a course for Ph.D. students could focus on both simultaneously, while also keeping a focus on the use of data analytics and tools.

In order to determine and thus further increase the adoption of evidence-based education, a common set of assessments must be developed. Assessing the degree-completion rate is insufficient because many professions, such as welders or smart builders, require certain skills rather than a degree, so a different metric for competency is necessary. A common set of assessments would also allow for straightforward aggregation and comparison across experiments, and thus stronger conclusions from data-intensive research in education.

DEVELOP NEW FORMS OF EDUCATIONAL ASSESSMENT

Two of the briefing papers focused on developing new forms of educational assessment, and other papers and conference discussions highlighted this vision as an important opportunity. On page 59, Shute depicted an aspirational vision for assessment:

> Imagine an educational system where high-stakes tests are no longer used. Instead, students would progress through their school years engaged in different learning contexts, all of which capture, measure, and support growth in valuable cognitive and noncognitive skills. This is conceivable because, in our complex, interconnected, digital world, we're all producing numerous digital footprints daily. This vision thus involves 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).

Shute goes on to delineate what advances are needed in educational data science to achieve this vision.

Similarly, on pages 80-81 Confrey described an infrastructure to embed assessment into learning:

A DLS is student-centered when each of these components is designed to strengthen students' movement within the digitally-enabled space.

A student-centered DLS:

> Increases students' ability to understand what they are learning.
>
> Supports appropriate levels of choice in sequencing or making decisions about materials (with guidance of teachers or knowledgeable adults).
>
> Supports genuine mathematical work including an authentic use of the tools (not just filling in answers).
>
> Affords peer collaboration, including discussing and sharing results.

> Allows students to create, store and curate products, and provides students' diagnostic feedback allowing them to self-monitor and set goals.

Student-centeredness does not imply individualization, working largely alone at one's own speed, but it does support personalization, making choices and self-regulation.

As with Shute's vision and other models for new forms of assessment discussed at the workshops, Confrey's application of data science in education would dramatically change both learning and assessment by providing new forms of evidence for decision-making to students, teachers, and other stakeholders. Realizing this type of educational model was seen as a grand challenge for the field.

RECONCEPTUALIZE DATA GENERATION, COLLECTION, STORAGE, AND REPRESENTATION PROCESSES

Numerous people at the workshops indicated that an opportunity for data science in education is to extend the range of student learning data that is both generated and collected. On page 32, Mitros discussed:

> There are many types of big data that can be collected in learning environments. Large amounts of data can be gathered across many learners (broad between-learner data), but also within individual learners (deep within-learner data). Data in 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

On page 34, Ho pushed for an emphasis on:

> …"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. If I create a massive open online course (MOOC), or an online educational game, or a learning management system, or an online assessment, I am less enabling the collection of data, than creating data in a manner that happens to enable its collection. I am arguing that one should be critical of any line of work that touts its "data intensive" or big data orientation without describing the contexts and processes that generate the data. When the context and process are particular, as they often are in big data educational research, applicants that promise general contributions to "how we learn" are likely to damage or at least muddy a field already overpopulated with mixed findings.

Several briefing papers discussed Evidence Centered Design (ECD) as an approach for determining what data to generate. For example, on page 48 Klopfer indicated:

ECD defines four relevant models:

- the *student* model (what the student knows or can do);
- the *evidence* model (what a student can demonstrate and we can collect to show what they know);
- the *task* model (the designed experience from which we can collect data); and
- 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 was seen by authors and participants as an important next step. For example, on page 63 Gilmore discussed the importance of repositories and open data sharing:

> Open data sharing can help to translate insights from scientific research into applications serving essential human needs. Open data sharing bolsters 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. Data sharing and reuse increases the impact of public investments in research and leads to more effective public policy. Although many researchers in the developmental, learning, and education sciences collect video as raw research data, most research on human learning and development remains shrouded in a culture of isolation[177]. 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.

On page 70, Gummer highlighted different types of data that should be collected and then stored in a way that facilitates analysis and sharing:

> In contrast to the two preceding cases on micro-level learning data, EdWise centers on the meso- and macro-levels of educational data and the potential for integration across these data levels to inform research and policy studies. The NSF recently funded a proposal for researchers at SRI who are examining the ways in which teachers make use of data from an online learning platform that includes instructional resources and content assessments that serve as the central structure in the students' learning environments. The intent of the research is to examine the key challenges facing practitioners in their use of information that comes from data-intensive research methods and to identify what partnership activities best support evidence-based practices. The findings from this study will lead to an understanding of the utility and feasibility of a teacher's use of the volumes of data that come from virtual learning environments, effectively bridging the micro- and meso-level data categories.

An early model of a repository that has successfully advanced data science in education is described by Koedinger on pages 68-69:

> A major output of LearnLab has been the creation of DataShop, the world's largest open and free repository of educational technology data and analytic methods[178]. One of the many insights that can be drawn from the vast amount of data we have collected in DataShop is evidence on the rate at which learning occurs (Figure 17). We see across many data sets that each opportunity to practice or learn a skill in the context of problem-solving reveals a rather small average improvement in student success (or, equivalently, drop in error rate, as shown in Figure 17). These changes in student success across opportunities to practice or learn (get as-needed feedback or instruction on a skill) can be modeled as learning curves.

On page 74, Siemens presented a vision of how data might be generated, collected, and shared to create:

> In education, a PLeG is needed where a profile of what a learner knows exists. Where the learner has come to know particular concepts is irrelevant; this may be via work, volunteering, hobbies, personal interest, formal schooling, or massive open online courses (MOOCs). What matters is that all members involved in an educational process, including learners, faculty, and administrators, are aware of what a learner knows and how this is related to the course content, concepts, or curriculum in a particular knowledge space. Four specific elements

[177] Adolph KE, Gilmore R O, Freeman C, Sanderson Pand Millman D. 2012. Toward open behavioral science. *Psychological Inquiry, 23* (3), 244–247.doi:10.1080/1047840X.2012.705133 .

[178] Koedinger KR, Baker R, Cunningham K, Skogsholm A, Leber Band Stamper J. 2011. A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, R.S.J.d. Baker (Eds.). *Handbook of Educational Data Mining* (pp. 43-55). Boca Raton, FL: CRC Press.

> are included in the multipartite graphs that comprise PLeG:
>
> - Social learning activities and networks;
> - Cognitive development and concept mastery;
> - Affective and engagement;
> - Process and strategy (meta-cognition)
>
> 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.

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, on page 37 Mitros described the shortcomings of current analytic approaches:

> Numerical techniques, which presume that assessments are designed based on principles which optimize for measurement, often fail when applied to the much broader set of classroom assessments. There is an inherent friction between:
>
> - Having a sufficient number of problems for statistical significance vs. long-form assessments which allow students to exercise complex problem solving and mathematical maturity.
> - Measuring individual students vs. group work.
> - Standardized assessments vs. diversity in education. The US economy benefits from a diverse workforce, and the educational system, especially at a tertiary level, is designed to create one. There are over ten thousand distinct university-level courses.
> - Aiming for 50% of questions correct (maximize measurement) vs. 100% of concepts mastered (mastery learning)
>
> Many traditional psychometric techniques rely on a relatively uniform dataset generated with relatively unbiased sampling. For example, to measure learning gains, we would typically run a pretest and a posttest on the same set of students. In most at-scale learning settings, students drop out of classes and take different sets of classes; indeed, the set of classes taken often correlates with student experience in previous classes. We see tremendous sampling bias. For example, a poor educational resource may cause more students to drop out, or to take a more basic class in the future. This shifts demographics in future assessments to stronger students taking weaker courses, giving a perceived gain on post- assessment unless such effects are taken into account.
>
> Likewise, 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.
>
> Big data is succeeding largely due to the availability of open source tools like Hadoop that provide common platforms for many researchers and industry players to collaborate around. That kind of collaboration is essential to foster, especially in education where the issues are not yet well understood. Also, open source tools allow people to replicate each other's work: one can take another's algorithms, run them on data, and see whether the results are consistent. Replicability is important in general, but it is especially important in contexts where student confidential data cannot be shared, but code can. Further, open source tools allow people to build on each other's work. One can take another's student model, and use it in an adaptive system. This also is key to progress in big data analytics.

On page 56, Dede discussed the inadequacies of current methodological approaches for analyzing data from games and simulations:

Quellmalz, Timms, and Schneider (2009) examined issues of embedding assessments into games and simulations in science education[179]. Their analysis included both tightly structured and open-ended learning experiences. After studying several immersive games and simulations related to learning science, including River City, they noted that the complex tasks in simulations and games cannot be adequately modeled using only classical test theory and item response theory. This shortfall arises because these complex tasks have four characteristics[180]:

1. Completion of the task requires the student to undergo multiple, nontrivial, domain-relevant steps and/or cognitive processes.
2. Multiple elements, or features, of each task performance are captured and considered in the determination of summaries of ability and/or diagnostic feedback.
3. The data vectors for each task have a high degree of potential variability, reflecting relatively unconstrained work product production.
4. Evaluation of the adequacy of task solutions requires the task features to be considered as an interdependent set, for which assumptions of conditional independence do not hold.

Quellmalz et al. (2009) concluded that, given the challenges of complex tasks, more appropriate measurement models for simulations and games—particularly those that are open ended—include Bayes nets, artificial neural networks, and model tracing. They added that new psychometric methods beyond these will likely be needed.

While a number of authors and participants described work on new analytic methods, breakthroughs in this area are clearly a necessary advance for data science in education.

BUILD HUMAN CAPACITY TO DO DATA SCIENCE AND USE ITS PRODUCTS

Authors and participants frequently discussed the 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. On page 45, Oblinger indicated the importance of building capacity in data science:

Data-intensive environments demand a new type of professional that some call data scientists. No matter what the name, higher education needs to develop the skills of these professionals as well as a pipeline into the profession. Data science is a blend of fields, including statistics, applied mathematics, and computer science. Qualities of data scientists who can address data-intensive challenges include:

- Technical skills: Mathematics, statistics, and computer science skills to work with data and analyze it.
- Tool mastery: Complex software tools are critical to analyzing massive amounts of data.
- Teamwork skills: Almost all of the data science roles are cross-disciplinary and team-based; hence, teamwork skills are critical.
- Communication skills: Deriving insights from data, communicating the value of a data insight, and communicating in a way that decision makers can trust what they're being told.
- Business skills: Understanding the business and bringing value from contextual understanding to the data analysis.[181]

[179] Quellmalz E S, Timms M Jand Schneider SA. 2009. *Assessment of student learning in science, simulations, and games*. Paper prepared for the National Research Council Workshop on Gaming and Simulations. Washington, DC: National Research Council.

[180] Williamson D M, Bejar I Iand Mislevy R J. 2006. *Automated scoring of complex tasks in computer-based testing. Mahwah, NJ: Erlbaum.*

[181] Woods D. 2012, March. What Is a Data Scientist?: Michael Rappa, Institute for Advanced Analytics. *Forbes Magazine*. http://www.forbes.com/sites/danwoods/2012/03/05/what-is-a-data-scientist-michael-rappa-north-carolina-state-university/3/ .

Developing an understanding of the skills essential in data scientists and others who support big data systems will be important so that institutions can develop the appropriate training and education programs, as well as attract students.

Berland, on page 58, described insights he has gained in preparing creative data scientists:

> The group has learned several factors of successful data-driven learning: students love analyzing data about themselves; teachers understand better than we do when data would be helpful for teaching; and using advanced data analytics on constructive, creative learning environments is both possible and not nearly as hard as we had thought. In short, we learned that training novice data scientists through real constructive work–as researchers on my team, designers on my team, teachers we work with, and students themselves–is not only possible, but can enjoyable for all parties. We have found that people become deeply engaged and understand complex data analytic content more fully when they are deeply connected to that content. From there, it is possible for both learners and researchers to think differently about data by connecting and visualizing many different modes of those data, such as transcripts, game play, pre- and post-tests, and more longitudinal data. Those connections to both the data and across different types and modes of data seem essential to more deeply understanding learning trajectories.

Building human capacity was the theme of a breakout group at the second workshop. When discussing almost any aspect of data-intensive research, the overarching issue of how to educate vast numbers of people with varying backgrounds to be data-savvy is surprisingly difficult to avoid. In both the public and private sectors, many researchers and organizations have access to more data than they can manage or process. Multiple discussions at the workshop emphasized the need to increase overall human capacity in data science (defined in the executive summary as the large-scale capture of data and the transformation of those data into insights) in response to this data deluge, whether discussing skills for education researchers, educators, or students.

In the context of this workshop, access to educational and student data is rapidly advancing the understanding of the science of learning, but advancement is limited to the data literacy of researchers in the education sciences. Participants at both workshops discussed that the field of educational research needs to help scholars work with, understand, and appreciate the culture of data-intensive research. In addition to supporting and growing a community of data-savvy education researchers, professional societies in education research could develop data literacy standards, online tools, and training programs to supplement researchers with data literacy skills.

That said, a major challenge for the field to expand its data science capacities is that few data science education programs currently exist, and most educational research programs do not require data literacy beyond a graduate statistics course. Participants at the first workshop discussed the lack of consensus concerning what skills comprise data literacy, beyond computational literacy and statistical analysis. To address this lack of consensus, in section XI of this report George Siemens from the University of Texas at Arlington suggests building international data and learning analytics communities. Also, in section VII of this report, Matthew Berland from the University of Wisconsin-Madison describes his team's work in educating data scientists.

Building human capacity in data-driven educational research requires that the current research culture emphasize cross-disciplinary study and group projects more than is currently the case. Infusing educational research with data science training or providing an education "track" for data scientists could provide these cross-disciplinary opportunities. Section VI of this report on massively open online courses suggests that the NSF provide targeted investments to support data research training for graduate students in education. This program could grant students access to real data sets and train them in hands-on analytic methods that include skills in statistics, computer science, tool mastery, teamwork, communication, and business. The authors noted that the Institute of Education Science Research Training Program could serve as a model for this NSF-targeted investment, which could, in turn, develop a bigger pipeline into the profession of data science.

As the field of data-driven educational research expands, two sections of this report on *Games and Simulations and Privacy, Security, and Ethics* respectively, point out that researchers will need to be sensitive to privacy, confidentiality, and ethical issues in their analyses. Ethics should be included in every step of data science training, and not be treated as an afterthought, to reduce the unintentional emotional harm that could result from misusing various analyses. Utilizing big data in educational research has enormous potential to advance teaching and learning, but it is imperative that students' and educators' identities remain protected at all stages of research.

Overall, the need to build capacity in both producers and consumers of big data in education was discussed repeatedly in both workshops; clearly, this is a major issue for the field.

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. On pages 89-90, Buchanan discussed the complexities of ethics, particularly algorithmic processing of data sets:

> Gradually, more attention is being paid to explicit and implicit bias embedded in big data and algorithms and the subsequent harms that arise. To this end, big data analytics should include methodologies of:
>
> ◗ Values-sensitive design,
>
> ◗ Community-based participatory design,
>
> ◗ Anticipatory ethics,
>
> ◗ Ethical algorithms, and
>
> ◗ Action research
>
> These approaches situate our participants, actors, and users as central and informed, as empowered decision makers. Friedman states that "central to a value sensitive design approach are analyses of both direct and indirect stakeholders; distinctions among designer values, values explicitly supported by the technology, and stakeholder values; individual, group, and societal levels of analysis; the integrative and iterative conceptual, technical, and empirical investigations; and a commitment to progress (not perfection)."[182]

On page 93, Hammer suggested approaches Institutional Review Boards (IRBs) and other regulatory groups could follow:

> Each 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, controlling 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 U.S. 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. A risk approach allows for changes in the types of research being done and the range of safeguarding solutions that could be applied. This would provide a framework to allow the newest research into privacy practices, security approaches, and research methodologies to be evaluated for how they mitigate risk, and to reuse those evaluations across the research community. Standardization and

[182] "VSD: Home." Value Sensitive Design. University of Washington, n.d. Web. 15 Sept. 2015. <http://www.vsdesign.org/index.shtml>.

> reuse would minimize the cost of evaluation while increasing the quality of evaluation.

On page 95, Gesher addressed issues of security for educational data:

> The reasons for paralysis around data sharing in education have much to do with a focus on the possible harms rather than the likely harms. In general, risk is composed of a combination of the probability of some harm occurring and the cost incurred if and when that harm does occur. There needs to be a shift away from focusing on the stakes and a better comprehension of the probability to understand the what and how of safely sharing educational data. This sort of risk analysis is known as threat modeling and comes to questions of data privacy via the field of computer security… the privacy world has begun its own evolution into the world of threat modeling, moving a model of *potential privacy harm* to *likely privacy harm*… In the educational domain, some of the potential harms can not be mitigated through policy and technical controls alone, so addressing those concerns will require the adoption of active oversight as a core tenet of data-intensive educational research.

On page 42 O'Reilly described the concept of differential privacy:

> While research in differential privacy is largely theoretical, advances in practical aspects could address how to support the content and platform providers who transmit the data when they want to choose a tradeoff between risk of re-identification and utility. Subsequently, effort would be required to mature the demonstrations for regular use by the development of prototypes that have user-friendly interfaces to inform controller decisions. Controller acceptance will require a set of technology demonstrations that, in turn, require major effort and resources. Demonstrations would be feasible if a "safety zone" could be set up where technology can be explored and validated against friendly re-identification adversaries who try to crack identities without causing any threat of real harm to the learners' data. Data scientists in the MOOC analytics sphere who develop variables and analytic models should be encouraged and supported to explore differential privacy mechanisms and bring them to practice.

Workshop authors and participants agreed that, unless stakeholders in education feel comfortable about privacy, security, and ethical considerations, data-intensive research in education will be unable to realize its potential.

THE EDITOR'S CONCLUDING THOUGHTS

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.

Building on the foundation of models for data-intensive research in the sciences and engineering through adapting and evolving these for studying learning, teaching, and schooling is a golden opportunity for educators. To rapidly advance in realizing the potential of this approach, the many strategies described in this report will be most effective if applied together rather than piecemeal. Furthermore, 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 by one or two of these groups.

The goal of this report is to inspire and foster continuing dialogues about how best to move forward with data-intensive research in education. If, in several years, its ideas and approaches have been supplanted by rapidly emerging, next-generation strategies, this document will have been a great success.

Notes:

CRA

Computing Research
Association

1828 L Street, NW, Suite 800
Washington, DC 20036
P: 202 234 2111 F: 202 667 1066
www.cra.org cccinfo@cra.org