Briefing Book

Workshop on Data-Intensive Research in Education

Arlington, Virginia; May 31-June 2, 2015

Computing Research Association

Sponsored by the National Science Foundation

Organized by Chris Dede, Harvard University

CRA

Computing Research Association

Preface

Chris Dede, Harvard University

We've all attended workshops where all the right people were in the room to have really interesting and important discussions, but the meeting was disappointing because presentations used up almost the time for dialogue. This workshop is designed to avoid this problem, in part through this briefing book.

All participants are asked to read these think-pieces in advance. This provides each presenter with the opportunity to get their ideas on the table without using time at the workshop to do that. The collective think-pieces also provide a way to sketch the multiple dimensions of data-intensive research in education. "Big data" reminds me of the fable about the blind men and the elephant. Each of us has part of the beast: the trunk, an ear, a leg. At the workshop, we'll put the puzzle together, and the briefing book is the first step in the process of "seeing the elephant."

Presenters have been asked to "start in the middle," to assume that participants have read their thought-piece so they can build on those ideas. In this way, we can get deeper in our dialogues. Plus, even before their session you can approach presenters to discuss the ideas in their think-piece.

So, please read these in advance. Enjoy…

Advancing Data-Intensive Research in Education

Waterview Conference Center
1919 North Lynn St.
Arlington, VA, 22209
http://www.executiveboard.com/exbd/waterview/local-area/directions/index.page

Sunday, May 31, 2015

5:30-7:00 Opening reception at Le Meridian Hotel, Arlington

Monday, June 1, 2015

7:30-8:30 Breakfast

8:30-8:50 Welcome from *Joan Ferrini-Mundy, Assistant Director, Directorate for Education and Human Resources (EHR)*

8:50-9:00 Purposes and Processes of the Workshop
Susan Singer, Division Director, Division of Undergraduate Education (EHR/DUE)
Chris Dede, Timothy E. Wirth Professor in Learning Technologies, Harvard University

9:00-10:00 NSF's Role in Advancing Data Science
Susan Singer (DUE); Taylor Martin, Program Officer, Division of Research on Learning (EHR/DRL); Gül Kremer, Program Officer, Division of Undergraduate Education (EHR/DUE); Elizabeth Burrows, NSF AAAS Fellow, Division of Mathematical Sciences (MPS/DMS)

10:00-10:15 Break

10:15-11:15 Predictive Models based on Behavioral Patterns in Higher Education
Ellen Wagner (PAR Framework); David Yaskin (Hobsons)
Chair: Chris Dede, Harvard

11:15-12:30 Dialogue on Privacy, Security, and Ethics
Elizabeth Buchanan (U.W. Stout); Ari Gesher (Palantir); Patricia Hammer (PK Legal); Una May O'Reilly (MIT)
Chair: Anthony E. Kelly (Office of the Assistant Director, EHR/OAD)

12:30-1:30 Working Lunch
("birds of a feather" groups to discuss analytics, infrastructure, data sharing, data standards/interoperability, privacy/security/ethics, producer/consumer relationships, building human capacity, visualization…)

1:30-2:45 Integrating Data Repositories
Ken Koedinger (LearnLab Datashop); Rick Gilmore (Databrary); Edith Gummer (Kauffman Foundation)
Chair: Gül Kremer, Program Officer DUE

2:45-3:00 Break

3:00-4:15 MOOCs
Diana Oblinger (EDUCAUSE); Piotr Mitros (edX); Andrew Ho (Harvard)
Chair: John Cherniavsky, Senior Advisor, EHR

4:15-5:15 Plenary Discussion – Synthesis of the Day

Dinner on your own – please form groups around topics of interest

Tuesday, June 2, 2015

7:30-8:30 Breakfast

8:30-8:45 Summary of yesterday; framing of today

Susan Singer and Chris Dede

8:45-10:00 Games and simulations for training, informal post-secondary learning
Matthew Berland (UW); Eric Klopfer (MIT); Valerie Shute (Florida State)
Chair: *Dexter Fletcher (IDA)*

10:00-10:15 Break

10:30-12:00 Breakout Groups

a. New forms of teaching and learning based on data-rich environments, visualization, and analytics
Moderator: Susan Singer

b. Infrastructure
Moderator: Taylor Martin

c. Producer/consumer relationships and partnerships
Moderator: Chris Dede

d. Building human capacity
Moderators: Earnestine Easter, Program Officer, Division of Graduate Education (EHR/DGE) and Michelle Dunn, *Senior Advisor for Data Science Training, Diversity, and Outreach (NIH)*

12:00-1:00 Working Lunch – Sharing insights from Breakouts

1:00-2:15 The Way Forward: Integrating Insights
Vijay Kumar (MIT); Jere Confrey (NCSU); George Siemens (UT-Arlington)
Chair: Chris Dede

2:15-2:30 Funding Opportunities

2:30-3:00 Plenary Discussion: Closing Thoughts and Next Steps
Susan Singer and Chris Dede

How Big is "Big Data" Across Disciplines: A Preliminary Analysis of Workshop Presentations on Model Projects Funded by Five NSF Directorates

Elizabeth H. Burrows[1]

1. AAAS S&T Policy Fellow, Big Data track, placed at National Science Foundation, Division of Mathematical Sciences

Abstract

The increase in capacity of and cost reduction in computing technologies has enabled unprecedented efficiencies in scientific discovery through curation, analyses and interpretation of massive datasets. However, it is observed that the uptake level and concentration on "Big Data" opportunities for scientific purposes are varied across disciplines. We assert that this variation is caused by the nature of the data needed within disciplinary communities, and can be characterized using the Velocity-Variety-Veracity-Volume typology. Using National Science Foundation sponsored exemplary projects from geological, engineering, biological, computational, and atmospheric sciences, we plan to analyze data characteristics within these projects and compile salient lessons learned to inform the scientific community at large, with specific attention to education research. Presented within is an example analysis from a project in Biology, the National Plant Genome Initiative (NPGI).

Introduction

In January 2015, the first workshop in this series was held, titled, "Towards Big Steps Enabled by Big Data Science", and its focus was on case studies of effective partnerships outside of education between big data producers and consumers. The agenda, presentations, and rationale for the connection between the first workshop and the current one described in this briefing book, are available online at http://cra.org/events/big-data-initiative. *Prior to delving into the content of the second workshop, focused on education, it is worthwhile to conduct a comparative analysis of the status of "Big Data" in the projects presented in the first workshop.*

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 complexity of dealing with "big" data may not be as fruitful or important compared to other disciplines. Attesting to this, in their survey of the definitions of Big Data, Stuart and Barker (2013) indicated that the literature using the term Big Data came from many disciplines, and yielded "multiple, ambiguous and often contradictory definitions." (pg. 1). They then compiled definitions, ranging from more abstract in nature to the ones utilizing facets that induce complexity in handling and analyzing data. These definitions included industry's (e.g., Microsoft, Intel, Oracle) input as well as other organizations', such as National Institute of Standards and Technology (NIST).

Stuart and Barker's (2013) review of big data definitions converged on the criticality of the following: a) Size: the volume of the datasets; b) Complexity: the structure, behavior and permutations of the datasets; and c) Technologies: the tools and techniques that are used to process high volume and complex datasets. Indeed, these critical factors are reflected in one of the most recent definitions put

forth by NIST's Big Data Public Working Group (2015, pg. 5): "Big Data consists of extensive datasets primarily in the characteristics of volume, variety, velocity, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis." The volume, velocity and variety factors are similar to those presented as early as in 2001 within the Gartner Report, where complexities due to (1) the increasing size of data (volume), (2) the increasing rate at which it is produced (velocity), and (3) its increasing range of formats and representations (variety). Veracity is added to the factors describing complexities of big data, encompassing widely differing qualities of data sources, with significant differences in the coverage, accuracy and timeliness of data (Dong and Srivasta, 2013).

We opine that a preliminary analysis of sample big data implementations across disciplinary boundaries using the Velocity-Variety-Veracity-Volume typology will support the discussions on interdisciplinary research and development in this domain, and cross-fertilize further development with the benefit of lessons learned informing all disciplines. With this intent, below we first describe the methodology we have adopted for the preliminary analysis, and then present the results. Salient lessons learned from various disciplines are also summarized to further inform all disciplines, including education.

Methodology

Our goal is to conduct a preliminary analysis, using the Velocity-Variety-Veracity-Volume typology, 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. In order to choose model case studies, NSF program officers from different directorates were contacted to identify exemplary big data projects. Nominees were then contacted to attend a workshop where they introduced their projects discussing challenges and opportunities. Data for our analysis is comprised of the presentations by these speakers (project principle investigators), supplemented by the notes taken during the presentation and Q&A sessions and related literature. The remaining steps of the analysis methodology can be summarized as follows.

1. Mine PowerPoints, notes, and related publications from the workshop and fill out a table with each speaker in the rows and the 4 Vs in the columns
2. Cross-validate classifications
3. Ask speakers for corroboration
4. Seek additional literature to support/ reject conclusions
5. Conduct a broader literature review of the current status and evolution (including profile and trajectory) of data-intensive research in each discipline

Results and Discussion

Preliminary assessments of the case studies are presented, followed by a deeper examination of a case study in biology. Further quantitative detail and a thorough literature review will be conducted for each project presented at the workshop.

High-level Reflections

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 LIDAR (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 capability to its 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, 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 massive amounts of learning data that could be collected from MOOCs, intelligent tutoring systems, and digital teaching platforms.*

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 the types of behavioral and learning data discussed above.*

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 for 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 discussed 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.

Biological Sciences Case Study

Research Challenges and Resource Needs in Cyberinfrastructure & Bioinformatics: BIG DATA in Plant Genomics, *Diane Okamuro*

Current Big Data Boundary. While the National Plant Genome Initiative (NPGI) is advancing capabilities in Big Data science with relation to all four V's, variability is perhaps their greatest challenge. Particularly with their 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 ranging from genomic and proteomic to phenotypic data. NPGI has over 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 comes in many different forms at different, but ever-increasing velocities.

Lessons Learned. Stressing the importance of standards and ontologies from the beginning is 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, it is highly beneficial when companies have incentive to make their data available and collaborate with academics. In genomics, this incentive came about 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 this field, Howe et al. (2008) 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 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 often report newly cloned genes without providing GenBank IDs, the human gene CDKN2A has ten 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.

Conclusion

Once all of the presentations from the first workshop are analyzed, conclusions will summarize salient common and uncommon lessons learned across disciplines.

Acknowledgements

This paper was completed with tremendous contribution and direction from Gul Kremer, Program Director at National Science Foundation, Division of Undergraduate Education, on leave from Pennsylvania State University, Department of Industrial and Manufacturing Engineering and Christopher Dede, Timothy E. Wirth Professor in Learning Technologies, Harvard University Graduate School of Education.

References

Dong, X.L. and Srivasta, D. (2013). "Big Data Integration", ICDE Conference, pg. 1245-1248.

Douglas, L. 3d data management: Controlling data volume, velocity and variety. *Gartner. Retrieved*, 6, 2001.

Howe, D., Rhee, S.Y. et al. (2008). "Big Data: The Future of Biocuration", Nature, Vol.455/4, pg. 47-50.

NIST Special Publication 1500-1 (2015). "Draft NIST Big Data Interoperability Framework: Volume 1, Definitions", http://bigdatawg.nist.gov/_uploadfiles/M0392_v1_3022325181.pdf

Stuart, J. and Barker, A. (2013). "Undefined By Data: A Survey of Big Data Definitions", 1-2.

Strategies for Scaling Student Success: The PAR (Predictive Analytics Reporting) Framework

Ellen Wagner
PAR Framework

Metrics currently used to describe and compare the performance of higher education institutions in the United States do not reflect the post-traditional students, instructional methods, business models, and data resources that distinguish contemporary higher education. This paper describes the evolution of a massive data research project using predictive analytics to gain a multi-institutional perspective on patterns of student loss and momentum for all types of students in the US post-secondary system. This project, named the Predictive Analytics Reporting Framework, and known as PAR, is now informing the development of institutionally specific predictive models and national outcomes benchmarks for the postsecondary community, providing insight into the performance of for-profit and alternative delivery models, including online learning. PAR has also started to identify potential improvements to federal data collections, statutory disclosure and reporting requirements, especially with regards to transfer students and adult learners. Perhaps of greatest potential value is PAR's current work on intervention measurement.

.

PAR began as big audacious idea, 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 US post-secondary education. PAR originally intended to pay attention to improving the retention and completion rates of online students. Despite much investment and myriad solutions for improving student success, postsecondary education completion rates have generally remained unchanged for the past forty years. Of all students who enroll in postsecondary education, less than half (46.1 percent) complete a degree within 150 percent of "normal time" to degree. (Knapp, Kelly-Reid and Ginder, 2012)[1] While online learning offers

[1] [1] Knapp, L.G.; Kelly-Reid, J.E. and Ginder, S.A. (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)

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,[2] with even lower rates of completion and retention than in their on-the-ground counterpart courses and programs.

As described by Ice et al (2012)[3], PAR commenced by working 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 worked together to determine 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 sixteen colleges and universities to the collaborative and an additional 44 variables in the dataset in the three years that followed. From these 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.

After receiving four research grants from the Bill & Melinda Gates Foundation between 2011 and 2014 to conduct rigorous testing and evaluation of the predictive models, benchmarks and intervention ROI tools, PAR launched as a non-profit provider of analytics-as-a-service in

---

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

[3] 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

January 2015. PAR has further 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 differentiates itself by connecting predictions of risk to solutions that mitigate risk as measured by improved 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, (2014)[4] 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 specific point of need.

PAR Framework Current Status

PAR currently holds over 2,600,000 anonymized student records and 24,000,000 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 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. PAR is included among the Institute for Higher Education Policy (IHEP)'s PostsecData Collaborative national Voluntary Data Projects[5]. Gartner Research[6] notes that PAR is distinguished among the many data analytics solutions emerging in the education domain by its common, openly published data definitions and student success frameworks.

---

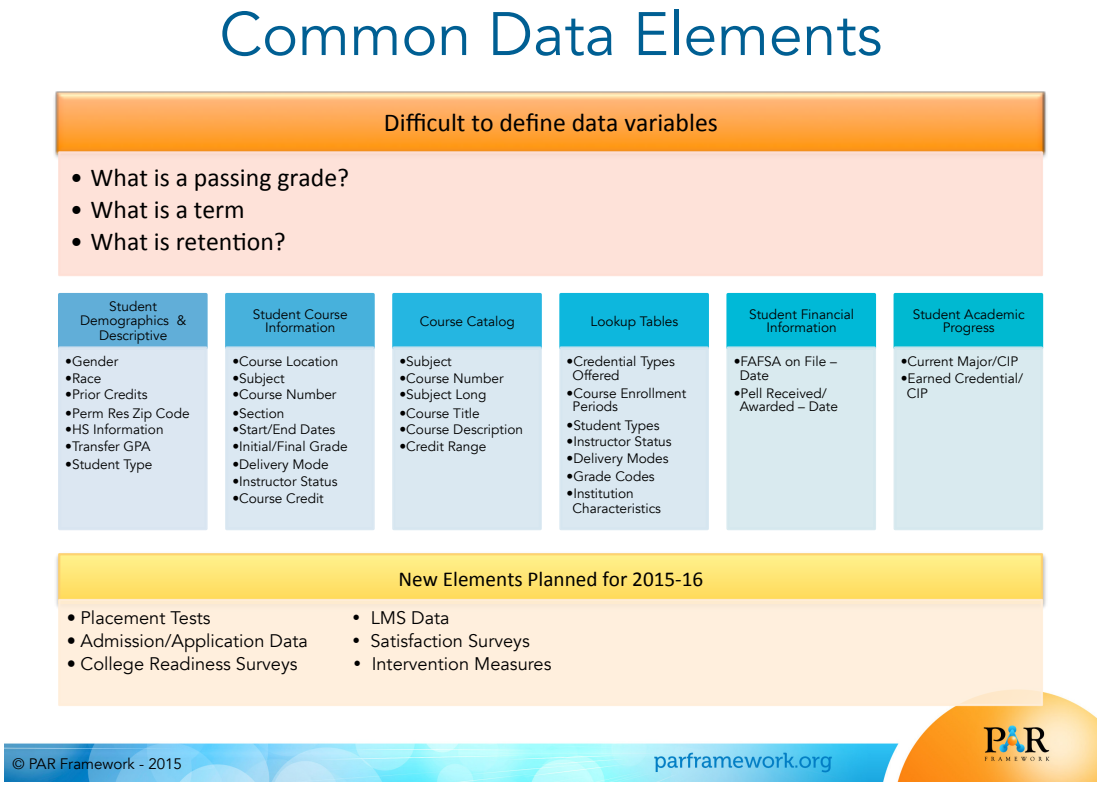
[4]Bloemer, B., Swan, K., Cook, V., Wagner, E.D., Davis, B. (2014) The Predictive Analytics Reporting Framework: Mitigating Academic Risk Through Predictive Modeling, Benchmarking, and Intervention Tracking, Illinois Education Research Conference, Bloomington IL, Oct 7, 2014.

[5] http://www.ihep.org/research/initiatives/postsecondary-data-collaborative-postsecdata

[6] Lowendahl, J.M (2014). The Education Hype Cycle, 2014. Gartner Research. https://www.gartner.com/doc/2806424/hype-cycle-education-

How PAR Works

The PAR Framework identifies factors that affect success and loss among 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 common data definitions provides insights that are not available when looking at records from a single institution.



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

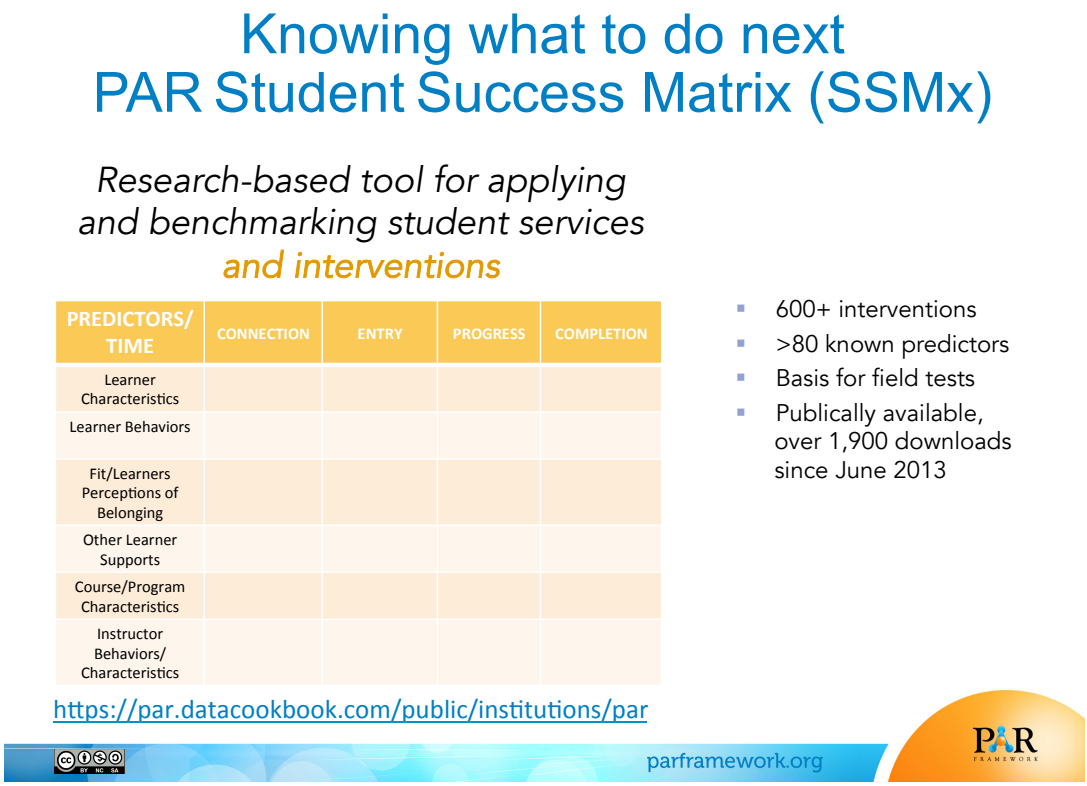
ID which is used to easily re-identify those students after the data has been analyzed by PAR. PAR puts the data through more than 600 quality assurance tests as it is prepared for inclusion in PAR's Amazon Web Services-hosted data warehouse. Data are then analyzed to develop institutional-, program-, 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/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 data experts work hand-in-hand with member institutions, providing individualized support for understanding, gathering and delivering longitudinal student-level data from across institutional systems using a well validated and market-adopted set of data definitions and file specifications. The PAR processes and support, combined with scalable, automated Quality Assurance tools, guide member institutions in crafting and delivering accurate and meaningful student level record sets. Throughout the process PAR data representatives help college and university staff diagnose and correct cumbersome and potentially costly institutional data issues that can impede correct reporting, insight, and availability of funds based on performance funding, student financial and veteran aid. 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.
- 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's student success framework and Student Success Matrix (SSMx) application use a validated mechanism to inventory those student success activities across the institution. PAR SSMx gives users the tools to capture, measure and compare ROI at the individual intervention level.



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

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

Reflections after Four Years in the Data Trenches

- 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 project to talk to one another between and within projects.

- Common data definitions are a game changer for scalable,

generalizable, repeatable learner analytics.

- Predictions are of greater institutional value when tied to treatments and interventions for improvement, and intervention measurement to make sure 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 come to being fluent with data-driven decision-making, 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 IPAS systems extends value and utility of tech investments.

- It takes all of us working together toward the same goal in our own unique ways to make the difference.

Using Predictive Analytics to Drive Student Success

David Yaskin, Senior Vice President for Student Success, Hobsons
John Plunkett, Vice President of Policy and Advocacy, Hobsons
Mark Wicks, Ph.D., Director of Data Science, Hobsons
Amanda Mason-Singh, Research Analyst, Hobsons

Although 90 percent of students enter college with the intention of completing a degree or certificate (Ruffalo Noel-Levitz, 2013), 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 (National Center for Education Statistics, 2014). Thus, it is not surprising that higher education institutions are being pressured, either by regulation or law, to submit "student success[1]" data to state, regional, or federal agencies in order to receive funding (Hayes, 2014). Currently, 34 states are either using or in the process of implementing performance-based funding (National Conference of State Legislatures, 2015). In addition, the Federal Postsecondary Institutional Rating System seeks to link college performance, via retention, completion, and loan default rates, to financial aid (U.S. Department of Education, 2013).

Habley and Randy (2004) 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. Even so, student completion rates have not significantly changed, leading Tinto and Pusser (2006) 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.

Creating a Digital Engagement Strategy for Student Success

Hayes (2014) argues that digital engagement—defined as the "use of technology and channels to find and mobilize a community around an issue and take action" (Visser &

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

Richardson, 2013)—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.

We will discuss a specific example of such an enterprise success platform that uses predictive analytics later in this paper. However, first a discussion of system-centered versus student-centered data is warranted.

Using Student-Centered Data Versus System-Centered Data

Institutions collect vast amounts of data about their students, and the most important aspect of an enterprise success platform is making better use of the data they already have. The 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 centers, 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 (2014) recommends a *student-centered* approach for institutional data use. By focusing on the student instead of the systems that generated the data, stakeholders see a

comprehensive view of a student's experience. According to Conrad and colleagues (Conrad, Gaston, Lundberg, Commodore, & Samayoa, 2013), a student-centered approach facilitates greater understanding of the relations between students, their college experiences, and outcomes. 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.

Valuable Drivers for Employing Predictive Analytics

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. Staff need 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.

We believe 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. To illustrate this point, we will provide an example of how we are using predictive analytics within the Starfish Enterprise Success Platform.

Example: Predictive Analytics and the Starfish Platform

Starfish Retention Solutions, which became part of Hobsons in early 2015, began developing its predictive analytics solution in mid-2014. During development, Starfish worked with two existing clients to build models from their data and provide predictive success scores for their students. To extend this work, in December 2014 at the White House College Opportunity Day of Action, Starfish committed to offer complimentary predictive analytics services to Davidson County Community College, Northeast Wisconsin Technical College, and Morgan State University in Baltimore.

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 appear to advisors in the Starfish Platform. Each individual student gets a score (e.g., 80% chance of continuing).

Starfish employs machine-learning techniques and random forest models, a type of nonlinear, nonparametric regression model, which are known for their versatility, performance, and ability to scale to large amounts of data (Breiman, 2001). Because these models are nonlinear, they find patterns such as discontinuities, threshold effects, break points in predictor variables, and interaction effects. These effects are nonlinear and therefore cannot be discovered automatically by generalized linear models (GLMs) such as linear regression or logistic regression.

The predictive model is built from data contained within the Starfish database, which includes data from the institution's SIS, the LMS, and the Starfish application itself. For new clients who do not have historical Starfish data, an initial model is constructed from an initial load of historical SIS data. Data available from the SIS includes admissions data, GPAs (term

and cumulative), credit hours attempted, credit hours earned (term and cumulative), credit hours attempted but not completed (term and cumulative), 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 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 within the Starfish Platform.

As the Starfish Platform is used to advise and monitor students, it records additional behavioral data that can define or refine future models. These data can include appointment types, reasons for making appointments (e.g., tutoring or advisement), topics discussed in meetings (as documented by Starfish "speed notes"), instructor-raised flags for attendance or other concerns, and system-raised flags (e.g., low assignment grades in LMS). We have begun to incorporate some of these behavioral data into the 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, we do not assume that appointments of type X will necessarily have predictive value at another institution. As we go forward, we will continue to explore the use of additional behavioral data.

In addition to providing predictive scores, we are working to provide 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 with an above-average age) 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.

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 (Vuong & Hairston, 2012).

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 (Center for Community College Student Engagement, 2013; Kuh et al., 2006; Tinto & Pusser, 2006; Vuong & Hairston, 2012).

We argue that the use of an enterprise success platform combined with predictive analytics that are based on historical student data can make institutional staff more effective at helping students succeed.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5-32.

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

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: University of Pennsylvania Graduate School of Education. Retrieved from http://www.gse.upenn.edu/pdf/cmsi/using_educational_data.pdf

Habley, W., & McClanahan, R. (2004). *What works in student retention?* Iowa City, IA: ACT, Inc.

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

Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & 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

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

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

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

Taylor, L., & 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

Tinto, V., & 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

U.S. Department of Education. (2013, December 17). Request for information to gather technical expertise pertaining to data elements, metrics, data collection, weighting, scoring, and presentation of a postsecondary institution ratings system. *Federal Register*. Retrieved from https://www.federalregister.gov/articles/2013/12/17/2013-30011/request-for-information-to-gather-technical-expertise-pertaining-to-data-elements-metrics-data

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

Vuong, B., & 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

Elizabeth A. Buchanan, Ph.D.
Endowed Chair in Ethics
University of Wisconsin-Stout
PO Box 790
Menomonie, Wisconsin, USA 54729
715.232.5184
buchanane@uwstout.edu

National Science Foundation Workshop Thought Paper: Ethics, Big Data, and Algorithms[1]

In this brief report, I'd like to focus on the relationships between and among privacy, big data, algorithms and the concept of harms. I will use a context of applied ethics and professional ethics to frame issues and controversies, and hope to encourage reflection and reasoned debate about the ethical realities of big data.

Ethics is about what's possible and what's "good," what's "just." However, in our professional literatures and educational discourses, there tends to be more focus on compliance and restriction: What are the legal, technological, 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 towards 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[2] 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.[3] 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 ultimately, can data, or an algorithm, 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,
- Action research.

These approaches situate our participants, actors, 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)."[4]

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 not so silently in many cases) 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 and faster, and more intentional. Big data and algorithms now tell us who is eligible for welfare, what political affiliations we have, and where are children will go to 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*"[5].

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 over 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."[6] 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 ration 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."[7] 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."[8] 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."[9]

To return to the concepts of harms generated out of big data analytics, take a few examples: A widow is continually reminded of her deceased spouse, on birthdays, anniversaries, special occasions; she does not want to change his Facebook status as it will disrupt past their shared experiences on Facebook. A young man is greeted by pictures of his burning 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.[10]

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."[11] 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, 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 judgements, then try to encode it in the software to make sure those judgements are visually perceptive to anyone who has to use our software."[12]

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 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 favour the middle of the bell curve. Only the most sophisticated algorithms work well in the tails."[13]

To an ethical end, Eubank[14] 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.

---

[1] This briefing is a revised version of my keynote address, "Living in a Time of (Un) Ethical Algorithms, Information Architecture Summit, 25 April 2015, Minneapolis, Minnesota.
[2] http://www.datascienceassn.org/code-of-conduct.html
[3] Using two sources, http://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science/ and 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.
[4] http://www.vsdesign.org/index.shtml
[5] The Policy Machine,
http://www.slate.com/articles/technology/future_tense/2015/04/the_dangers_of_letting_algorith

ms_enforce_policy.html?wpsrc=sh_all_tab_tw_top&utm_content=bufferbab23&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

[6] http://www.brainfacts.org/sensing-thinking-behaving/awareness-and-attention/articles/2009/decision-making/

[7] http://www.brainfacts.org/sensing-thinking-behaving/awareness-and-attention/articles/2009/decision-making/

[8] http://www.wsj.com/articles/SB10001424052970203462304577138961342097348

[9] http://www.wsj.com/articles/SB10001424052970203462304577138961342097348

[10] http://meyerweb.com/eric/thoughts/2014/12/24/inadvertent-algorithmic-cruelty/

[11] http://www.slate.com/articles/technology/future_tense/2015/04/the_dangers_of_letting_algorithms_enforce_policy.html

[12] http://www.cio.co.uk/insight/compliance/quest-for-ethical-algorithms/

[13] http://www.cio.co.uk/insight/compliance/quest-for-ethical-algorithms/

[14] http://www.slate.com/articles/technology/future_tense/2015/04/the_dangers_of_letting_algorithms_enforce_policy.html

Ari Gesher, Palantir Technologies

Privacy Risk (Perceived and Actual) as a 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 there are significant useful insights to be gleaned in properly composed educational data sets. However, due to its proximity to children, educational data is fraught with worries around its misuse, abuse, and theft. These persistent anxieties have created an atmosphere of paralysis, with many data owners preferring not to share data at all 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 being put into 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 around privacy risks is reduced to the point where data can shared amongst 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 world 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 various educational data sets.
2. **Policy & 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 give 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 the stakes and 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 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 easily circumventable (or absent) security controls make the addition of privacy controls a moot point.

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

In the nascent days of computer security, securing a system was viewed, more or less, as a black-and-white affair. From a design perspective, the practice was concerned with creating *Maginot Line*-like fortifications: systems designed to be unbreachable. This philosophy suffered the same fate as the real Maginot Line as compromise techniques were perfected, often routing around the hardened parts of the systems to find side-channel attacks that would subvert lower layers or weak points in systems. As computer security professionals learned more about the real world problems encountered in their work, the view of security moved from a *fortification* philosophy to one of *mitigation*. It was no longer assumed that 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 rapid-response. This sea change is probably most visible in Bruce Schneier's Secrets & Lies (https://www.schneier.com/books/secrets_and_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, aggregration, and deresolution to created anonymized data sets that could be shared with researches.

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 world where none of the anonymization techniques are foolproof has the ability to share datasets been destroyed? Instead of closing up shop, 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*. Paul Ohm lays out this shift in his seminal paper on the subject, "Sensitive Information (http://ssrn.com/abstract=2501002)" *(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.274 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 Shostack278 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 still very common in the world today. 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.

It's only through a 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 bulkwark against abuse, the way to protect against the harms that, through granting the access they need to do their work, 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 an integration of identified datasets against anonymized datasets 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 only way mitigate real harms that can come 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 dataset. It's possible to filter the dataset to make it harder to identify individuals in the dataset. However, modern privacy researchers have shown that most forms of dataset anonymization can be pierced by integrating identified data sets and then applying careful data science across the two to unmask the individuals in the dataset.

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 some central trusted authority. An environment like this would not only include places to place datasets, but also the analysis tools, instrumented for auditing, for researchers to work with the data.

So data-sharing would be 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 datasets available to cloud data-sharing and analysis environment.
2. Researchers would apply for access to the datasets they need to do their research, including importing their own private data into the environment.
3. All 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 amongst research teams.

PATRICIA 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. Fear over privacy is leading to the stifling of public educations research and pushing research into corporate hands where transparency is less required and compliance is easier. This poses risks to educational research, which is often the font 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 use of big data, in-home or in-classroom audiovisual recordings, or biometric stress indicators, and 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 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 Department of Education (DOE), will benefit IRBs by developing , clear guidelines to follow and standards they can rely upon. Researchers will therefore have less difficulty having projects approved by IRBs. Society also benefits by increasing the amount of research done in educational and research facilities vice 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 has expertise in the field. Often this leads to delay, indecision, or overly conservative restrictions being placed on the researchers. One solution, often proposed, 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 experts available 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 baselines standards for privacy and system protection in educational and/or other types of research. A project could demonstrate that it met the minimum guidelines before IRB review, which would inherently expedite the process and limit or eliminate the institution's liability in the case 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, 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. 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 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 organizations voice in determining acceptable risk but would be addressing these questions from a position of knowledge.

Evolving Research Capabilities/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 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 cost, many research organizations have shied away from maintaining large, aggregating data sets. The concerns over maintaining the data are often not weighed against the benefit to research that might be available 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 for 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 data to find trends based on age, race, income, geographic location, level of education, time of day, physical activity, physical traits, etc. 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 learns best 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 sector and government sector agencies, the big data collectors are not research organizations. They usually collect the data as an auxiliary function to 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 they are 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 based on social media, wearable technology, and other internet 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 the attributes they describe. Education research often uses these massive data, but lean, sets in combination with some other source of data (e.g., demographic data on geographical units based on census and other measurements), in order to enrich the set of attributes to be studied. Indeed, companies that assemble these data sources into unified data sets are popular sources of marketing data on individuals and households.

Minimization/De-anonymization

The trend in data privacy is to minimize the amount of data collected, which would then 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 if this, there must be careful consideration, lest the data set suffer over-minimization, which could actually 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 only to collect data absolutely necessary for the study, subsequent researchers may need to conduct the same type of testing using different subjects to collect a variable not collected the first time. This would expose 20,000 subjects to a risk instead of 10,000. By not collecting a variable there could be a trend or correlation the researchers are missing which 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 similar studies must be conducted researchers can limit the overall risk to any group of students, not just the students involved 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.

Conclusion

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.

NSF workshop on data-intensive research in education
Thought Paper

Una-May O'Reilly, unamay@csail.mit.edu
ALFA Group, MIT CSAIL

CASES:

The kinds of big data/data-intensive research in education you have experience with and the specific types of data collected

I founded and currently partner, with Dr Kalyan Veeramachaneni, the AnyScale Learning For All (ALFA) Group at MIT's Computer Science and Artificial Intelligence Lab (CSAIL). 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 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 crowd sourced 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.

Specific Types of Data Collected:

A short description of the multiple raw data streams of MOOC edX platform data that are supplied for data science/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 very nice (friendly) summary.

We create data: 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 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.

Technologies, infrastructures, and tools you use, including mechanisms for collection, storage, analysis, and sharing

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. Machine learning is, perhaps less frequently known outside academia

Issues with standards and interoperability
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 MOOCdb documentation. We develop software that we intend to share once it is release-ready.

Methods and analytic approaches you use, including visualization
Machine learning: we largely use open source libraries situated within our own research frameworks that allow rapid scaling and result comparison.

REFLECTIONS
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 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 their internal business, gets them "right" and then offers them externally where they really help save development time.

Another way to answer this question is to list explicit examples of producers and consumers. In the MOOC-sphere the producers are the platform providers: edX and Coursera. The consumers of data are stakeholders: students, instructors, education technologists, institutional registrars, learning scientists. In the MOOC-sphere, relationship building has been driven by the platform providers because they have 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 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 return for using the site. They are briefed of 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.[1]

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.

In its entirety, the process is based upon trust that is granted based on direct verification of people and institutions. The process culminates, indirectly, in trust that the best efforts of the parties involved to honor their commitments will be sufficient. This endpoint exposes vulnerability: it assumes the data won't fall into the wrong hands inadvertently, when, in fact, it may. This problem could happen when the data is held by any of the data controllers. This 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. It is the HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset. It holds aggregate records, one per individual per single edX course for 5 MOOCs offered by Harvard X and 8 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" (see [Daries 2014] for more details). 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 dataset, 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 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 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.

Una-May O'Reilly's Author Bio:
I founded and currently partner the AnyScale Learning For All (ALFA) group at CSAIL. ALFA focuses on scalable machine learning, evolutionary algorithms, and frameworks for large-scale knowledge mining, prediction and analytics. The group has data science projects in MOOC technology: MoocDB, student persistence and resource usage analysis, data privacy protection technology; clinical medicine knowledge discovery:

arterial blood pressure forecasting and pattern recognition, diuretics in the ICU; wind energy: turbine layout optimization, resource prediction, cable layout.
My research is in the design of scalable data science systems that execute on a range of hardware systems: clouds, GPUs, grids, clusters, and volunteer compute networks. I am interested in agile and rapid intelligent data analytics capability and apply unsupervised, semi-supervised, and supervised learning algorithms for similarity search, classification, non-linear regression, and forecasting. I consider end-to-end systems, i.e. ones that start with raw data, move to data organization and information transformation, next on to inferential analysis on conditioned exemplars, and finally to the deployment and evaluation of learned "algorithmic machines" in the original application context.

[Daries 2014] Daries, Reich, Waldo, Young, Whittinghill, Seaton, Ho, and Chuang] Daries, Jon P, Reich, Justin, Waldo, Jim, Young, Elise M, Whittinghill, Jonathan, Seaton, Daniel Thomas, Ho, Andrew Dean, and Chuang, Isaac. Quality social science research and the privacy of human subjects requires trust. acmqueue, 2014.

[1] I can provide detailed descriptions of how MOOC data is released at Stanford U and MIT if it would be helpful.

What Causes Changes in Learning Rate? Data Intensive Research Opportunities

Ken Koedinger, Director of LearnLab, Professor of Human-Computer Interaction and Psychology, Carnegie Mellon University
koedinger@cmu.edu

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; Ericcsson et al., 1993) indicating that a substantial amount of deliberate practice is needed to acquire expertise, Hambrick et al. (2014) quote Gardner (1995) 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 (see learnlab.org; Koedinger, Perfetti, & Corbett, 2013). 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 (Koedinger, Baker, Cunningham, Skogsholm, Leber, Stamper, 2011). 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 (see Figure 1). 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 1). 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 which uses a logistic regression generalization of item-response theory, called AFM (cf., Koedinger, McLaughlin & Stamper, 2012). The predictions of AFM are shown in blue (dotted) lines in Figure 1, with the actual data (average error) shown in red (solid) lines.

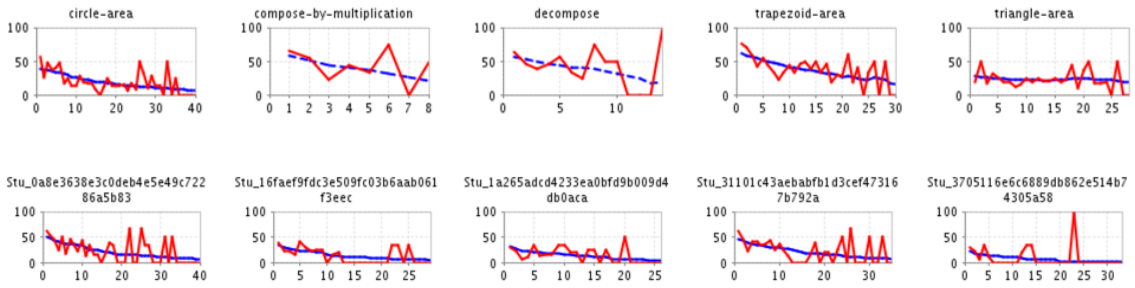


Figure 1. Learning curves showing a decrease in error rate (y-axis) for each successive opportunity (x-axis) to demonstrate or learn a skill, averaged across students for different skills in the first row and averaged across skills for different students in the second row. The variations in learning 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 .03.]

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

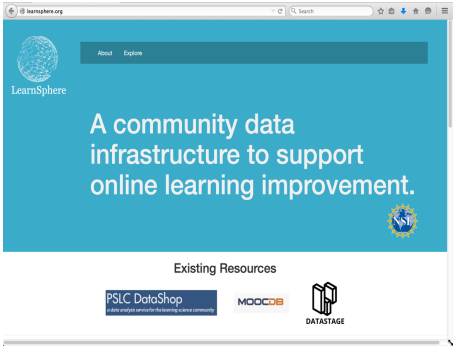
at about 50% correct on a skill, after one opportunity of practice they will now be at about 54% correct. This 4% increase diminishes as correctness increases toward 100%. Thus, to get from 50% correct to 95% correct requires about 20 practice opportunities.

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 (e.g., Ericsson et al., 1993) that suggests it takes about 10,000 hours to become an expert. 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 1). 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 a irregular shape is .15 logits. However, there is a relatively small variation across students (Liu & Koedinger, 2015). 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 other things and these skill difficulty variations are common across all students. We do find some variation in learning rate across students (Liu & Koedinger, 2015) 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 sudents 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/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 (Koedinger, Booth, & Klahr, 2013) 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 (Koedinger et al., 2015). 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, and University of Memphis are building LearnSphere (see learnsphere.org) to help data researchers address these questions.

References

Duckworth, A. L., Kirby, T. A., Tsukayama, E., Berstein, H., & Ericsson, K. A. (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.

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

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.

Hambrick, D.Z., Oswald, F.L, Altmann, E.M., Meinz, E.J., Gobet, F. & Campitelli, G. (2014). Deliberate practice: Is that all it takes to become an expert? *Intelligence*, 45, 34-45.

Koedinger, K.R., Baker, R., Cunningham, K., Skogsholm, A., Leber, B., 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.

Koedinger, K.R., Booth, J.L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science, 342*, 935-937.

Koedinger, K.R., Corbett, A.C., & Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36 (5)*, 757-798. ISSN: 0364-0213 print / 1551-6709 online DOI: 10.1111/j.1551-6709.2012.01245.x

Koedinger, K.R., Kim, J., Jia, J., McLaughlin, E.A., & Bier, N.L. (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.

Koedinger, K. R., McLaughlin, E. A., & Stamper, J. C. (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.**

Liu, R. & Koedinger, K.R. (2015). Variations i 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.

Open Video Data Sharing Can Transform Education Research

Rick O. Gilmore
The Pennsylvania State University, The Databrary Project

Karen E. Adolph, David S. Millman
New York University, The Databrary Project

Author Note

Rick O. Gilmore is in the Department of Psychology, The Pennsylvania State University, University Park, PA 16802, rogilmore@psu.edu. Karen E. Adolph is in the Department of Psychology at New York University. David S. Millman is with the NYU Library. Databrary is based on work supported by the National Science Foundation under Grant No. BCS-1238599, the Eunice Kennedy Shriver National Institute of Child Health and Human Development under Cooperative Agreement U01-HD-076595, and the Society for Research in Child Development. Any opinions, findings, and conclusions or recommendations expressed in the material contributed here are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Eunice Kennedy Shriver National Institute of Child Health and Human Development, or the Society for Research in Child Development.

Abstract

Video captures the complexity, richness, and diversity of behavior unlike any other measure. As a result, large numbers of people who study teaching and learning employ video. Video documents itself to a large degree. This presents significant potential for reuse by others. The potential remains largely unrealized because videos are rarely shared. Video contains information about personal identities. This poses challenges to sharing. The large size of video files, diversity of formats, and incompatible software tools pose technical challenges. We describe how the Databrary data library has overcome the most significant barriers to sharing video within the developmental sciences community. Databrary has developed solutions to maintaining participant privacy, storing, streaming, and sharing video, and for managing video datasets and associated metadata. The Databrary experience suggests ways that video and other identifiable data collected in the context of education research might be shared. We envision a data intensive science of teaching and learning, with video as its core, that allows educational experiences to be tailored to students in ways that big data promises to personalize medicine. The creation and support of repositories that enable the open sharing of dense, richly informative, high value, and high impact data about teaching and learning will help realize this ambitious vision.

Open Video Data Sharing Can Transform Education Research

Introduction

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 (Adolph, Gilmore, Freeman, Sanderson, & Millman, 2012). Researchers share interpretations of distilled, not raw data, almost exclusively through publications and presentations. The path from raw video 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. This paper describes how the Databrary data library has overcome the most significant barriers to sharing video within the developmental sciences community. It highlights how open video data sharing might improve scientific practice and advance research on learning and development.

The Promise and Challenge of Video

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 (Karasik, Tamis-LeMonda, & Adolph, 2014). Researchers record videos of students in classrooms (Alibali & Nathan, 2012) to understand what teachers do and how students respond. 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 success 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 is how to capitalize on video's potential to improve teaching and learning.

The answer requires overcoming 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/week (Gilmore & Adolph, 2012). 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 (Amso, Haas, Tenenbaum, Markant, & Sheinkopf, 2014; Yu & Smith, 2013; Fathi, Hodgins, & Rehg, 2012; Google Research, 2014;

Raudies & Gilmore, 2014).

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, 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 (Ascoli, 2006a). 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 (Ascoli, 2006b; Ferguson, 2014). 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 datasets 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, featuring more than 1,800 infant, child, and adult participants. Databrary has more than 100 authorized researchers representing more than 60 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 (databrary.org) 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 whose data are preserved indefinitely in a secure storage facility at NYU. 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, .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 user communities within Databrary 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 of 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.

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 restricting access to 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 (Databrary Project, 2015) with standard language we recommended for use with study participants. This language helps participants to understand what is involved in sharing video data, with whom the data will be shared, and the potential risks of releasing video and other identifiable data to other researchers.

Managing Data for Sharing

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 datasets 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.

Building a Community

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.

Conclusion

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 dataset. 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 involving identifiable information. We empower these learning scientists to extract from the data meaningful insights 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.

As Gesell once noted, cameras can record behavior in ways that make it "...as tangible as tissue" (Scott, 2011). 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 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.

References

Adolph, K. E., Gilmore, R. O., Freeman, C., Sanderson, P., & Millman, D. (2012). Toward open behavioral science. *Psychological Inquiry*, *23*(3), 244–247. doi: 10.1080/1047840X.2012.705133

Alibali, M. W., & 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

Amso, D., Haas, S., Tenenbaum, E., Markant, J., & 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

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

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

Databrary Project. (2015). *Databrary Release.* Retrieved from `http://databrary.org/access/policies/release-template.html`

Fathi, A., Hodgins, J., & 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

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

Gilmore, R. O., & Adolph, K. E. (2012). *Video Use Survey of ICIS and CDS listserv members.*

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`

Karasik, L. B., Tamis-LeMonda, C. S., & Adolph, K. E. (2014). Crawling and walking infants elicit different verbal responses from mothers. *Developmental Science*, *17*(3),

388–395. doi: 10.1111/desc.12129

Raudies, F., & Gilmore, R. O. (2014). Visual motion priors differ for infants and mothers. *Neural Computation*, *26*(11), 2652–2668.

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

Yu, C., & Smith, L. B. (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.0079659

Integrating Data: Imagining the Possibilities

Edith Gummer
Education Research and Policy Director
Ewing Marion Kauffman Foundation

The National Science Foundation *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 teaching and learning in STEM. 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. (NSF, 2013)

Both Ken Koedinger and Rick Gilmore reporting in the *Integrating Data Repositories* panel will discuss the micro level data. What I want to focus on is the meso and macro levels of 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.

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 Department of Education has supported a number of states to develop and implement Instructional Improvement Systems (IIS) 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 students at risk.

Much of the meso-level education data are collected through school and district level systems that include student demographics, attendance, disciplinary, course-taking, grade, local assessments (formative, benchmark and interim), state assessment and SAT/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. 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 (Coburn & Turner, 2011). 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 Department of Education State Longitudinal Data Systems (SLDS) have supported the development of P-20 data systems that frequently are attached to workforce data as well to the tune of over half a billion dollars. 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 levels of interoperability of the data in these systems, a number 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 states Office of Financial Management, was created in 2007 to assemble, link and analyze education and workforce data and 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;
- 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, 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 datasets 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 datasets 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, child care quality and educational data.

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. States' data collection and public reporting efforts should move away from simply complying with state and federal regulations and toward answering stakeholders' questions (DQC, no date).

The Ewing Marion Kauffman Foundation has developed a data tool that is intended for such public use by multiple stakeholders. Called *EdWise*, the tool 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 their website that include thousands of lines of data. We have combined fourteen million records of Missouri K-12 education data into a single easy-t-use online tool to help parents, educators, school districts, policymakers, and the general 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 the might provide either more targeted examples to query for 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 their 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. Ken and Rick identify even more technical problems. 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?

Diana Oblinger

Introduction

This document touches on many types of "big data" applications. Large amounts of data can be gathered across many learners (broad between-learner data), but also within individual learners (deep within-learner data). 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.[1] Possible applications range from game-based learning environments to analytics to MOOCs, to integrated advising systems, to competency-based systems, and more.

Big data in education provides many 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 diagnosis of learning needs or course trouble spots.
- Targeted interventions to improve student success and reduce overall costs to students and institutions.
- Deeper learning and better transfer of knowledge by using game-based environments for learning and assessment, where learning is situated in complex information and decision-making situations, using games as an architecture for engagement and assessment of skills such as systems thinking, collaboration, problem solving in the context of subject-area knowledge.
- 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 DFW rates, eliminating bottleneck courses, aligning course capacity with changing student demand, etc.

While there is tremendous potential, many questions remain unasked and unanswered. Below are some of the challenges that might be addressed through additional research. Note that several items are not discussed here because they are likely to be addressed in other papers (e.g., analytics or game-based environments).

Challenges

There are a number of challenges associated with data-intensive environments. Below is a sampling of issues. One illustrates the challenges of complex systems (integrated competency management system for students, higher education and employers); another focuses on technical infrastructure (next generation digital learning environment). Two illustrate challenges in human capacity, specifically awareness/adoption and workforce development. The final area illustrated deals with policy. Note that

[1] Thille, C., Schneider, D. E., Kizilcec, R. F., Piech, C., Halawa, S. A., & 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

some (e.g., policy issues) 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

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."[2]

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% 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.[3]

"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."[4]

There are at least 5 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, etc.
- Promote: Marking the assembled evidence to link candidates and opportunities, which may include social media analytics, behavioral assessment, and other data-mining techniques.
- Align: Using feedback loops to constantly evaluate performance and make improvements and 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

[2] Newman, Adam. (2015, February). Evidence of Learning: The Case for an Integrated Competency Management System. http://tytonpartners.com/library/evidence-learning-case-integrated-competency-management-system/
[3] Ibid.
[4] Ibid, page 6

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

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." EDUCAUSE research suggests that the next generation digital learning environment (NGDLE) will be an ecosystem of sorts, characterized by:

- 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.

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. [5]

[5] 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

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

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%) said they are either conceptually or experientially familiar with MOOCs; compare this to only one in four undergraduates (24%) who say they know what a MOOC is. Though few faculty reported having actually taught a MOOC (3%), they are much more likely than students to know about this alternative model for online learning.

Part-time faculty (53%) expressed more support than full-time faculty (38%); furthermore, non-tenure-track faculty (46%) were more supportive than tenured (34%) or tenure-track (39%) faculty. About two in five faculty (43%) with less than 10 years of teaching experience were supportive, whereas somewhat fewer faculty (37%) with 10 or more years of experience were supportive. Not surprisingly, the picture painted here is that newer (less experienced) faculty have more positive perceptions of MOOCs adding value to higher education.[6]

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%. While they are still only 5% of the learners on average, the increase may be tied to teachers.[7] Recent research from edX and HarvardX illustrated that a major audience for MOOCs are teachers (28% of enrollees in 11 different MOOCs were former or active teachers).[8] 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 audience, experience, 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.

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

[7] 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

[8] 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/

- 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 of the business, bringing value from contextual understanding to the data analysis.[9]

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

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. Appropriate policies must address privacy, security, and data sharing. Federal regulations, such as FERPA, are often misunderstood.

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 education and awareness activities that balance the need to use information to support institutional missions with the need to protect it from internal and external threats and ensure the privacy of the campus community. These practices constantly evolve as the threat landscape evolves.

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 to conduct activities without 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.[10]

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

- Legal and compliance issues: The consequences of compliance failure may be significant in analytics systems. Regulatory compliance (e.g., FERPA, HIPAA), e-discovery rules, open records laws, student privacy expectations (confidentiality), and the role of the institutional review board may all come into play.

[9] Dan Woods (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/

[10] Ho, Lisa. (2015) Privacy vs. Privacy. http://www.educause.edu/blogs/lisaho/privacy-vs-privacy

- Unintended consequences of third-party data access/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 inappropriate use of the data presented in dashboards or reports, or misunderstand their limits.[11]
- 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.[12]

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.

Conclusion

The five sections included in this thought paper are illustrative of the opportunities, challenges and research needed to advance data-intensive areas in education.

Diana Oblinger
President and CEO, EDUCAUSE
May, 2015

[11] 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

[12] Jones, K. M. L., 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

Big Data and Assessment of Complex Skills

Piotr Mitros
edX

Historically, assessment in classrooms was limited to instructor grading, or problems that lend themselves well to relatively simple automation, such as multiple-choice questions. Progress in educational technology, combined with economies of scale, has allowed us to 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 our measurements of what students learn, allowing us to tailor instruction to specific students needs and giving individualized feedback for an increasing range of issues. In addition, social interactions have increasingly moved on-line. We now have traces of a substantial portion of student-student interactions. By integrating these and other sources of data, we have data with which we can estimate complex skills, such as mathematical maturity, complex problem solving, and teamwork for large numbers of students. This paper looks at the potential information found in the data we now collect, some of the challenges with making sense of that data, and some early successes in analyzing that data. The data is complex. Actually extracting useful high-level metrics has proven difficult. The next grand challenge in big data in education will be finding ways to analyze complex data from heterogeneous sources to extract such measurements.

Keywords: educational datamining, assessment

Twenty years ago, most digital assessments consisted of multiple choice questions and most social interactions happened in person. Data was spread out over multiple systems with no practical means of integration. Over the past two decades, we have seen fundamental progress in educational technology, combined with broad-based adoption of such technology at scale[1]. Digital assessment has increasingly moved towards 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, and as a result, they either contain a large number of small problems (which are statistically significant, but generally fail to capture skills which require more than a minute or two to measure), or a small number of large problems (which, on a per-student basis lack statistical significance). Today, we are increasingly collecting data for students doing a large numbers of complex problems as part of their regular coursework. For example, the first edX/MITx course[2], 6.002x (Mitros et al., 2013) 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). Since these types of questions have a near-infinite number of possible solutions, answers cannot be guessed. Students could attempt to 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. We see similarly rich assessments in courses such as chemistry, biology, physics, digital electronics, and many others. Such complex assessments, taken together across many courses, give rich data about problem solving skills, creativity, and mathematical maturity.

Furthermore, we now collect microscopic data about individual student actions. We can see not only which problems students answered correctly, but how they got there. Extensive literature on expert-novice shows differences in problem solving strategy between novices and experts. For example, experts can chunk information (Schneider, Gruber, Gold, & Opwis, 1993) – an expert looking at an analog circuit will be able to remember that circuit, whereas a novice will not (Egan & Schwartz, 1979). In our data sets, we can see actions which reflect such differences. 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 which are proxies for expertise.

Next, social interactions are increasingly moving on-line. As we introduce increased amounts of digital group work to

[1] 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 SPOCs.

[2] Used both in a pure on-line format, as well as in a blended format in a number of schools

courses, we start to see traces of social activity in our logs. We can begin to look for students who under-perform or over-perform in group tasks, and directly measure students' contributions to groups. We have enough data to begin to look for specific actions and patterns that lead to good overall group performance, and hopefully we will be able to use such patterns to provide feedback to students. 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 analysis of social activities, such as e-mails and forum posts, as well. We believe this will begin to give insights into soft skills, writing processes (Southavilay, Yacef, Reimann, & Calvo, 2013), communications styles, and group dynamics.

Finally, aside from just looking within individual courses, we can perform longitudinal analysis across a student's educational career. 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 (Siemens et al., 2011) such as edX Insights (Mitros, 2013) or Tin Can provide a common data repository for all of a student's digital learning activities.

However, going from data to measurement is a complex problem. In the next few sections of this paper, we will discuss some of the challenges, as well as early successes.

Challenges – Pedagogical Design

There is substantial friction between the design for different educational purposes, of which, measurement is just one. 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 (Sadler, 1989).
- **The principal means by which student learn new information**. In many subjects, most student learning happens through assignments where they manipulate, derive, or construct knowledge (Chi, 2011) – not lectures, videos, or readings.
- **A key components 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 different 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 (Ericsson, Krampe, & Tesch-Römer, 1993), as well as principles such as rapid feedback, active learning, and constructive learning. Educational psychology (Bloom, 1984) and gamification emphasize mastery learning (where students eventually get all questions right).

Numerical techniques which presume that assessments are developed 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[3].
- 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)

To give an example of how friction comes into play, the MIT RELATE group applied item response theory (Embretson & Reise, 2000), a traditional psychometric technique, to calibrate the difficulty of problems in 6.002x, the first MITx/edX course. 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 – any sufficiently dedicated student could answer all questions correctly. To apply IRT in this context, RELATE had to substantially adapt the technique (Champaign et al., 2014).

Challenges – Diversity and Sample Bias

Many traditional psychometric techniques rely on a relatively uniform dataset generated with relatively unbiased sampling. For example, to measure learning gains, we would

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

typically run a pre-test and a post-test on the same set of students. In most at-scale learning settings, students drop out of classes, take different sets of classes, and 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 a future assessments to a stronger students taking weaker courses, giving a perceived gain on post-assessment if such effects were not controlled for.

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 moves from traditional statistics increasingly into machine learning, and calls for very different techniques from those developed in traditional psychometrics.

Challenges – Data Size and Researcher Skillset

Traditionally, big data educational research was conducted by statisticians in schools of education with tools such as spreadsheets, and numerical packages such as R. This worked well when data sets were reasonably small. A typical data set from a MOOC is several gigabytes. 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. With continuing moves towards technologies such as teleconferencing, we expect datasets to grow manyfold.

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 replicating well-established decades-old results from classical educational research.

Meaningful research requires skillsets from both backgrounds. There are few researchers with such skillsets, and collaborations are sometimes challenging due to substantial cultural differences between schools of education and schools of computer science.

Early Successes

An early set of high-profile successes in this sort of data integration came from systems which analyzed data across multiple courses in order to predict student success in future courses. This includes systems such as Purdue Course Signals (Arnold & Pistilli, 2012), Marist Open Academic Analytics Initiative (Lauría, Moody, Jayaprakash, Jonnalagadda, & Baron, 2013), and Desire2Learn Student Success System (Essa & Ayad, 2012).

There have been early successes with system which 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 assignments submitted by other students.
- **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 (Mitros & Paruchuri, 2013), each of the four grading systems contributes a different type and amount of information. 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.

Conclusion

While many of the goals of an educational experience cannot be easily measured, it is 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 actually cause 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 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 in outside of the classroom. Instructors cannot effectively critique longer formats of communications, such as e-mail threads, meetings, and similar without extreme student:faculty ratios – 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 (VanLehn, 2011) (National Research Council, 2000) (Patterson, Gavrin, & Christian, 1999). 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 (Basu, Jacobs, & Vanderwende, 2013), where data is integrate from very diverse sources, and where data is collected longitudinally.

With this shift, for the first time, we have data about virtually all aspects of students skills – including complex ones that are, ultimately, more important than simple factual knowledge (Sternberg, 2013). We have the potential to provide new means to assess students in ways which can improve the depth, frequency, and response time, potentially dramatically 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 tool 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.

References

Arnold, K. E. & Pistilli, M. D. (2012). Course signals at purdue: using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267–270). ACM.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391–402.

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.

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

Chi, M. T. H. (2011). Differentiating four levels of engagement with learning materials: the icap hypothesis. *International Conference on Computers in Education*.

Egan, D. E. & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory & Cognition*, *7*(2), 149–158.

Embretson, S. & Reise, S. (2000). *Item response theory*. Psychology Press.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, *100*(3), 363.

Essa, A. & Ayad, H. (2012). Student success system: risk analytics and data visualization using ensembles of predictive models. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 158–161). ACM.

Lauría, E. J., Moody, E. W., Jayaprakash, S. M., Jonnalagadda, N., & Baron, J. D. (2013). Open academic analytics initiative: initial research findings. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 150–154). ACM.

Mitros, P. (2013). Open platforms for pedagogical innovation. In *Learning analytics summer institute*. LAK.

Mitros, P., Affidi, K., Sussman, G., Terman, C., White, J., Fischer, L., & Agarwal, A. (2013). Teaching electronic circuits online: lessons from MITx's 6.002x on edX. In *Iscas* (pp. 2763–2766). IEEE.

Mitros, P. & Paruchuri, V. (2013). An integrated framework for the grading of freeform responses. *The Sixth Conference of MIT's Learning International Networks Consortium*.

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

Patterson, N. G., Gavrin, A., & Christian, W. (1999). Just-in-time teaching: blending active learning with web technology. Upper Saddle River NJ,: Prentice Hall.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, *18*(2), 119–144.

Schneider, W., Gruber, H., Gold, A., & Opwis, K. (1993). Chess expertise and memory for chess positions in children and adults. *Journal of Experimental Child Psychology*, *56*(3), 328–349.

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., Ferguson, R., … Baker, R. (2011). *Open learning analytics: an integrated & modularized platform* (Doctoral dissertation, Open University Press).

Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013). Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 38–47). ACM.

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

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221.

Four variations on a theme of data-intensive research in education

Andrew Ho
Harvard Graduate School of Education
Thought Paper – National Science Foundation Workshop on Data-Intensive Research in Education

In this brief thought paper, I offer four loosely related perspectives and arguments on the present and possible future of data-intensive research in education:

1) Before "data collection" comes "data creation"

Where do data come from? The phrases, "data collection," and, "data mining," both suggest that data simply exist 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 new contexts, where conventional notions of enrollment, participation, curriculum, and achievement required reconceptualization (DeBoer et al., 2014). We tempered early optimism around 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 (Reich, 2014). 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 NSF 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.

2) 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 refocus research productively 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, Harvard and MIT open online courses (Ho et al., 2014; 2015), 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 (see also, Seaton et al., 2015) and varying levels of initial commitment (see also, Reich, 2014). We and others have argued that this makes evaluating MOOCs extremely difficult (Reich & Ho, 2014), with the uncritical use of "completion rates" as an outcome variable being particularly problematic. 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 provides 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% 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.

3) NSF 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. Especially as “big data” in education are attracting those with little background in causal inference, assessment, or educational research, inferential and analytic errors remain common: 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 nails.

I think that a targeted investment by NSF 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.

4) 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 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 demonstrate this empirically is to exploit random variation in the assignment of the system, as in random assignment of the prediction-and-response system to some but not all students.

References

DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "course": Reconceptualizing educational variables for Massive Open Online Courses. *Educational Researcher, 43*, 74-84.

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

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

Reich, J. (2014). MOOC completion and retention in the context of student intent. *Educause Review Online*.

Reich, J., & Ho, A. D. (2014, January 24). Op-Ed; The tricky task of figuring out what makes a MOOC successful. *The Atlantic.*

Seaton, D.T., Coleman, C., Daries, J., and Chuang, I. (2015). Enrollment in MITx MOOCs: Are We Educating Educators? *Educause Review Online*.

Creating creative learning data scientists

Matthew Berland (University of Wisconsin–Madison), mberland@wisc.edu

I am an assistant professor of Digital Media in the Dept. of Curriculum & Instruction at University of Wisconsin–Madison, part of the Games+Learning+Society group, with appointments in both Computer Science and Library & Information Studies. My background is in both learning sciences and computer science, and my first paper on "big data" (or “very large corpora; Berland & Charniak, 1999) is taught in computer science classes globally. I am also the director of the Learning Games Play Data Consortium (PDC). The PDC is made of up 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 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.

As director of the PDC, I work with a wide range of people who are using data in new ways to understand learning, education, games, play, and creativity. We have currently identified five 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. In this piece, I will cover one core aspect that underlies many of the others: training new creative learning data scientists.

Training a diverse and wide-ranging set of designers and researchers to think about new possibilities with data is hard. Training people to think creatively about the possibilities for data in education turns out to be really hard, 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, 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 I look at 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 (myself 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 towards creating educational learning environments. Learning sciences - my home academic field and a place in which novel work is being done - is quite small and only modestly funded, but computer scientists (of which I am also one) 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 and learning sciences. In short, there are very few people who 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, and it may be the most likely one. 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 (boyd, 2007). 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" - furthermore, 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 both making people happy, helping them learn more quickly, and creating the next generation of data scientists.

Our group at UW–Madison (together with our many wonderful colleagues across the US) 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 multimodally through our PDC Dashboard. Our view is that learning data look fundamentally different from the kinds 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 (Berland, Baker, & Blikstein, 2014). 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 (e.g., Berland, Martin, Benton, Petrick Smith, & Davis, 2013; Berland, Smith, & Davis, 2013; Berland & Wilensky, 2015); 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 target students themselves - is not only possible but that it 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 researcher to think differently about that data by connecting and visualizing many different modes of those data, such as transcript, game play, pre-/post-tests, and more longitudinal data. Those connections both to the data and across different types and modes of data seem essential to understanding the data more fully.

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 come 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.

References

ADAGE. (2015). Retrieved May 15, 2015, from http://adageapi.org/

Berland, M., Baker, R., & 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

Berland, M., & Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 57–64). Association for Computational Linguistics.

Berland, M., Martin, T., Benton, T., Petrick Smith, C., & 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

Berland, M., Smith, C. P., & Davis, D. (2013). Visualizing Live Collaboration in the Classroom with AMOEBA. In *Proceedings of the International Conference on Computer-Supported Collaborative Learning*.

Berland, M., & 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

boyd, danah. (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.

Play Data Consortium. (2015). Retrieved May 15, 2015, from http://playdataconsortium.org/

Stenerson, M. E., Salmon, A., Berland, M., & Squire, K. (2014). Adage: an open API for data collection in educational games. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play* (pp. 437–438). Toronto, Ontario, Canada: ACM.

Eric Klopfer

Game-Based Learning Ecosystem

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 application of concepts. This differentiates what has come to be known (via Jim Gee and others) as the **g**ame (the digital distributed experience) from the **G**ame (the entire experience including what happens on and off screen—including interactions with peers and mentors, and 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 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 scaffolded 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?*

Effectively addressing this challenge actually 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/learners?
- What kind of data do we provide to teachers that is actionable?

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. Thus, games 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. Games may also play a role as a means for summative assessment, as they provide rich and complex problem spaces. But for the purpose of this paper I will focus on formative assessment.

The data generated by games, and in games, creates a tremendous opportunity for supporting better learning experiences. As is so 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

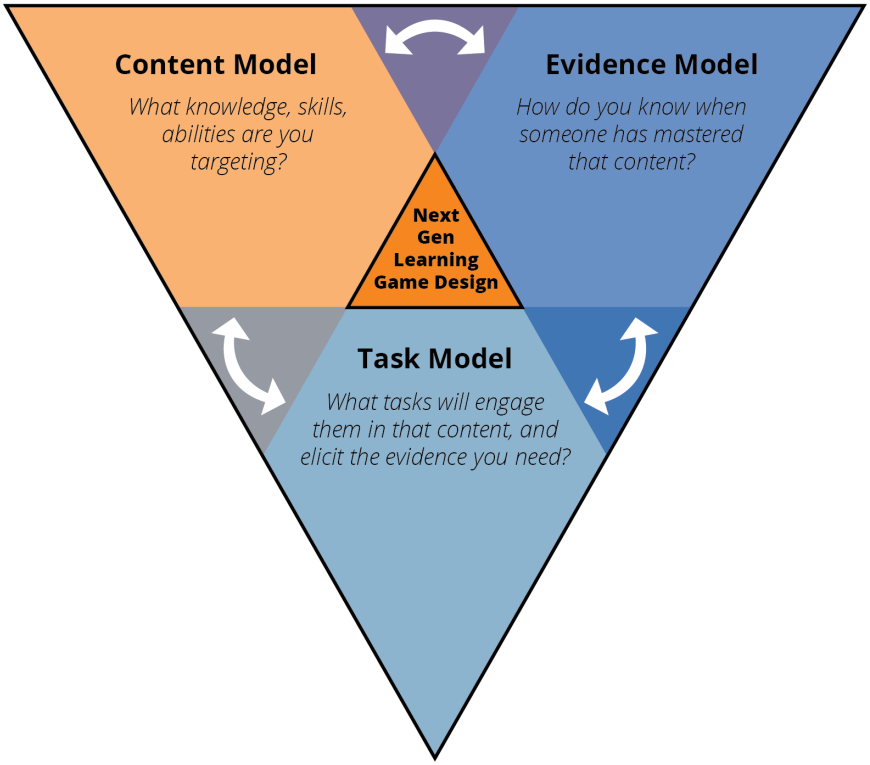
Reich says) of the keystrokes and data trails of games to recognize successful patterns and differentiate them from those of players who struggle, that methodology is not yet sufficient for realizing this potential of games. Instead, we must design for the learning experiences of games, the data they 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 then differentiate success from failure, but to 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 (ECD – Mislevy, Almond & Lukas, 2003) is one useful – and thus far highly popular amongst 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 amongst 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 that not only embodies the potential to create useful data, but the ECD framework serves as a design lens that can also provide engagement and challenge that draws players into the game.

In reality, the 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 this initial thoughtful and intentional design. Variations on ECD for design of educational games may make this methodology easier and more effective to follow. Groff et al. (2015) have proposed a simplified version that reduces this to a Content Model (the relevant knowledge and skills), Evidence Model (the information needed to know if someone has that knowledge) and the Task Model (what the person is engaged in doing to elicit that data) as a lens for instructional design that aligns both content and assessment data in games. This model further links each of these 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.



A simplified version of ECD known as XCD (Groff et al. 2015)

Similarly we (Conrad et al. 2014) 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 as well as math-based games, in which players conduct experiments that both models the practices of those disciplines and also provides 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 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, gets back an outcome and performs another experiment based upon that outcome. They may in fact need to perform a fairly extensive sequence of these experiments, based both upon the complexity of the task, and the random variation that may occur within those experiments.

This is a methodology that we apply 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 (tasks with proximate goals) along the way. Quests range from fixing buildings using geometry skills to diagnosing disease based on understanding of body systems. One of the quest lines is around 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 you just have a sample too small to show diversity. After that outcome, it is important to see what the player does next. Do they do 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 are also important in complex spaces to allow for exploration, that is times when players are simply orienting themselves and pursuing their own interests, which may not be targeting a particular learning outcome. This is often seen as a desirable outcome. In Radix, if players are exploring and conducting additional experiments on their 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, or simply don't know 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 players who are truly confused back 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 players on a pathway to success or failure, even what that action leads to an outcome which may be perceived as negative (death of the player). But such testing of the boundaries is important for the player understanding how the world in which they exist works. Longer sequences of actions – what the player does after that event – may provide for 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, where 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 pathway, 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 directly inform students about their progress, directly (providing information about what they are doing wrong when that is identified) 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, red do not provide teachers with enough information to be useful on a case by case basis, other than knowing whether the class is "getting it" or not. The other end of the spectrum, which shows the outcome of every game action for every player is too much information to be useful. A teacher with 100 or more students cannot use such information to address individual (or even classwide) issues.

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

This all means that using games to know 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 well. 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.

References

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

Groff, J., Clarke-Midura, J., Owens, E., Rosenheck, L, & 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.*

Conrad, S., Clarke-Midura, J., & 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.

Big Data in Education: Opportunities, Challenges, and Future Research

Valerie Shute, Florida State University

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 (e.g., Black & Wiliam, 1998; Shute, 2009), particularly learning outcomes and processes necessary for students to succeed in the 21st century. Most current approaches to assessment/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 (2001, p. 284). 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 information obtained 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 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 (see Bergmann & Sams, 2012). 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 (Strayer, 2012).

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.

1. *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 used, 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, Mislevy et al., 2003; Assessment Engineering, Lucht, 2013) 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. Towards 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).

2. *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 (Shavelson & Kurpius, 2012). 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 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.

3. *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 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.

4. *Privacy/Security*. The fourth hurdle involves figuring out a way to resolve privacy, security, and ownership issues regarding students' information. The privacy/security issue relates to the accumulation of student data from disparate sources. The recent failure of the $100 million inBloom initiative (see McCambridge, 2014) showcases the problem. 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 (Mislevy et al., 2003), it is able to accumulate evidence across disparate sources (e.g., homework assignment, in-class quiz on an iPad, 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 coming in 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) (see Almond, 2010 for more on this process). 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 going over 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 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 to shape the vision for a richer, deeper, more authentic assessment (to support learning) of students in the future. How many current businesses would elect to return to pre-barcode days?

References

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.

Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. International Society for Technology in Education (ISTE).

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice, 5*(1), 7-74.

Luecht, R. M. (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.

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

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

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Shavelson, R. J., & 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.

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

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

On the Value of Post-Secondary School Training

J. D. Fletcher

Institute for Defense Analyses

These comments concern the assessment of training and training innovation, mostly in the Department of Defense (DoD). The may or may not be relevant to this group, but, undaunted, I'll carry on.

By 'training' I mean preparation to perform specific jobs and tasks. It is a means to an end that is at least somewhat predictable. The requirements of jobs and tasks change, of course, and transfer of learning is still as critical for training as for education. But training stands in contrast to education, which is an end in its own right and preparation for life. Training seems particularly useful in assessing the effectiveness of instructional approaches, old and new, because we can determine relatively quickly how successful they are in producing necessary learning.

Educators may assume that training is not relevant to education, but I contend that instructional approaches are much the same for education and training. Both endeavors have much to learn from each other, and the list of training approaches and techniques, especially those of the DoD, that have made their way into K-12 education is long. These comments focus on DoD results and data from post-secondary training, but they may be relevant to most if not all training and education.

'Value' in the above title is central to these comments. Budget battles for training in DoD, especially training of individuals in residence (i.e., schoolhouses), fare as poorly as they often do in K-12 education. That may be partly due to our focus on training effectiveness and neglect of the "so what" question. To say that we find superior learning from a training approach is only part of the issue for decision makers and check writers in Defense and industry. They need to know what effectiveness means for success of the missions they must pursue. What, for instance, is the priority for training compared to other approaches that contribute to mission success? What is a pound of training worth? The same balance and priority determination must be found for public service expenditures where the question is to determine what a pound of education is worth. Over the years, I have been concerned with answering this question in both venues.

The issue is not cost -- for either civilian or military budgets. If it were, we might arm the U.S. Ari Force with Piper Cubs. Concluding that an instructional approach is unaffordable because of cost is insufficient. A full answer to the question must address both cost and what we get for it – i.e., return on investment. This issue is rarely addressed in either education or training

assessments as we compete for funds with other, perfectly respectable alternatives. In the military, the issue can be whether to buy jet propulsion fuel or improved training. We know how many sorties the fuel will bring and how that affects mission success. How does that compete with training outcomes? Similar balances must be determined for profit and loss in business and for public funds used to support local and national well-being. What, again, is a pound of training worth? There are ROI studies for education and training, but I suggest they are too few and too rare.

Measuring the cost of an investment is easier than measuring its return. Assessing operational return (e.g., mission effectiveness) is far too messy to discuss in this short note. It is difficult and wildly uncertain. However, three assessments of monetary return can be briefly mentioned here. In all three cases the investment is in computer technology for training.

As a first example, we find from data that the rate at which students in a classroom (civilian or military, children or adults) is at least 4:1. Other data have found that the use of computer technology to individualize (people are saying 'personalize' these days) instruction so that the slowest students can receive the time they need to reach basic learning levels and the fastest students can be all they can be (to coin a phrase). This capability has been found to reduce overall time to learn by at least 30 percent. Suppose we were to reduce the time for 60% of military technicians to complete entry level specialized training, which cost about $9B in 2014, by 30%. The savings would amount to about $1.8B. Cost of the computer software and hardware to accomplish this needs to be determined, but it is likely to be considerably less than $1.8B.

Second, we have found a recently developed, digital tutoring system for training Information Systems Technology (IT) technicians can, after 16 weeks, produce US Navy sailors who outscore in IT knowledge and troubleshooting skill other sailors with more than 9 years of Fleet IT experience. The effect sizes in this assessment exceed 3 standard deviations for the Fleet ITs (and for newly graduated sailors who received 35 weeks of training). Comparing this ability to accelerate the development of expertise to that required by 9 years of on-job experience and training and assuming the current training pipeline of 2,000 Navy ITs a year suggests annual savings to the Fleet of about $300M per year.

Third, we found that the same system used to train post-Gulf War veterans, most of whom were unemployed, for 18 weeks provided them civilian job offers with a median salary of $73,000, which is roughly equivalent to that earned by industry network administrators with 3-5 years of experience. Aside from the impact on these veterans' lives, the monetary return to the government for supporting veterans in this course (tuition, lodging, and meals) was about twice the return in revenue (i.e., taxes paid) received from other, more typical approaches providing similar support for two- and four-year academic institutions.

These examples approach cost-effectiveness and return on investment in different ways. The first holds effectiveness/return constant and minimizes cost/investment by releasing students as soon as they reach a threshold of learning. The second and third examples hold costs/investment constant while maximizing effectiveness/return. The latter approach may be much preferred because it is more compatible with established personnel practices, which have difficulty dealing with different learners finishing courses of instruction at different times.

The data reported in these examples may change with additional scrutiny and replication, although substantial change in their overall findings seems unlikely. However, the point of these examples is to suggest how the value – the "so what" – question might be addressed in similar assessments and research. It is to further suggest that return on investment can and should become a routine element in education and training assessments. Researchers may complain (as they have) about becoming accountants, but the nature of our business is changing and so should we if we are to defend investments in education and training and compete more successfully with other demands for public support and budget allocation.

The Data Dividend for Network enabled Open Education

M. S. Vijay Kumar Ed. D
Massachusetts Institute of Technology
(This "thought-paper" was done with contributions from Claudia Urrea, researcher from the Online Education Policy Initiative at the MIT Office of Digital Learning and Jeff Merriman from the MIT Office of Digital Learning)

The Data deluge is affecting all disciplines requiring new computational capabilities (tools, technologies, and platforms) and skills to capture, manipulate, visualize, integrate and manage (including preservation) large amounts of data. At the same time, it is presenting great opportunities in the field of educational research (Koedinger, McLaughlin & Stamper, 2014[1], Breslow et al. 2013[2]). Massive Open Online Courses(MOOCs) such as edX, have added legitimacy and even, perhaps, urgency to the field of educational research at MIT and elsewhere.

HarvardX and MITx research units are also following the recipe for advancing MOOC research recommended by Reich (2015): improving assessments, conducting experiments, and sharing data" (Ho et. al. 2015)[3].

Indeed, there are high expectations all around for massive open online courses to not only bring the best education in the world to the most remote corners of the planet, but also that the large data sets generated by users numbering in the hundreds of thousands will provide insights into education. These insights will inform faculty on how to use technology in their teaching, and will enhance the experience of learners everywhere. Large data sets will help legitimize some of the existing theories and methods of learning, and help establish new insights and theories of how people learn online, how communities emerge and interact, etc.

This "thought-paper" sets out to identify opportunities for data-intensive research in education to improve practice and policy. It draws from a set of research themes and associated issues/questions that were identified as part of an MIT exercise to frame a research agenda on MOOCs.

The paper also draws on activities underway at MIT with which the author and his associates are engaged that address issues related to developing technological considerations and

---

[1] Koedinger, Kenneth R., Elizabeth A. McLaughlin, and John C. Stamper. "Data-Driven Learner Modeling to Understand and Improve Online (2014).
[2] Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., and Seaton, D. T. Studying Learning in the Worldwide Classroom Research into edX's First MOOC. Research and Practice in Assessment 8, 19 (2013), 13–25.
[3] Andrew Dean, et al. "HarvardX and MITx: Two Years of Open Online Courses Fall 2012-Summer 2014." Available at SSRN 2586847 (2015)

capabilities for delivering quality learning opportunities at scale. The activities are in the area of Distributed Assessments and Mapping Courses to Skills development[4]. Interoperability considerations are central to the approaches.

The Data Dividend

The document identifies 3 Opportunity Areas for Big Data/ data-intensive research in education to improve practice and policy:

1: **Distributed, Embedded Assessment**
2: **Competency and Skill based Learning**
3: **Strategic Scaffolding**

1. Distributed, Embedded Assessment

Creating Repositories and Recommenders
A substantial body of research has been compiled on the best practices in structuring and using feedback in traditional educational settings. One of the most exciting features of online learning, as demonstrated amply in edX, is the platform's ability to ask students to *apply* the concepts they have just encountered and test them to provide timely feedback leading to mastery or different pathways (Embedding assessments frequently for formative testing). An enormous investment is currently being made at MIT and elsewhere to design and develop effective assessments, for the rapidly expanding roster of MOOC offerings from colleges and universities worldwide.

How can we make this process scalable and cost effective? If the power of technology can be pressed into service for direct assessment, education would benefit enormously.

To date, no technology exists to effectively manage and share this content or to support re-use of assessment items across courses, departments, institutions—or more importantly, across educational platforms and technologies. In addition, no approach to managing and authoring assessments has yet to effectively map them to learning objectives or track item response analytics and other valuable use [and user?] data. As we move forward in developing a robust digital learning infrastructure, the need for a new set of tools to facilitate and improve assessments is critical.

To address this need, we propose the creation of the **Digital Learning Assessment Bank**, a global federation of assessment tools that will facilitate the availability of online assessments. A secure and interoperable Federated Assessment Service would support the creation of such an assessment bank where users will be able to create and update assessment offerings and perform assessment authorizing, reporting, learning objectives mapping and analytics.

[4] Core Concept Catalog: MC3 (http://oeit.mit.edu/gallery/projects/core-concept-catalog-mc3)

Big Data capability can allow us to study the effectiveness of assessments drawn from these assessments banks for different learning outcomes/ in different contexts and make this information available along with these assessments. We can imagine Assessment Recommenders that facilitate the identification and selection of assessments from the digital learning assessment bank for use by course authors. As such Digital Learning/MOOC environments can provide more opportunity to personalize learning (through analytics), more opportunity to provide interactivity with content and more opportunity for asynchronous interaction.

Finally, research into digital assessment—its use and its results—can feed into our understanding of learning itself. As students interact with online materials, the data generated will give researchers insight into how learners struggle to master concepts, how they deal with misconceptions and skills, and how they ultimately succeed.

Questions:

- What are the alternative methods for understanding and making learning visible?
- Would Digital Learning Assessment Banks result in better learning objectives? Or vice versa?

2. Competency and Skill based Learning

Keys measures of educational effectiveness, from the perspective of the community colleges, Department of Labor (DOL), Department of Education (DOE) and the Trade Adjustment Assistance Community College Career Training (TAACCCT) grant program, are data that indicates the numbers of graduates of degree or certificate programs that go on to get jobs, and most importantly, the “growth in the wage premium associated with higher education and cognitive ability"(Autor, 2014)[5]. The data required for these measures are captured in various institutional and state systems, educational SIS, labor and wage reporting, departments of revenue, and cross-organizational agreements. The techniques for sharing this data for the purposes of research and reporting grant effectiveness are just beginning to emerge, largely, from what we can tell, driven by the requirements for these organizations to report results of recent federal grants. A robust, well design system that integrates real data would inform policy makers and at the same time, help create well-informed policy.

A missing component of this data train is where the mapping occurs between educational courses/programs and job skills. As we know, from our own design efforts around learning objective models, we cannot get an adequate end-to-end picture across the entire path from educational program to job placement and make sense of it without modeling educational goals related to course and programs. Without goal modeling, this mapping is done by hand, by

[5] Autor, David. "Skills, education, and the rise of earnings inequality among the" other 99 percent" 2014

people who make educated decisions regarding which programs of study map to a particular job classification[6].

We can imagine a number of activities that could help with this:

One, of course, being the learning objective cataloging systems we are building at MIT. In fact, for the TAACCCT Round 4 Data Bus project, we are planning to leverage this work, standing up services to manage and map data on educational goals. We are hopeful that one or more of the Mass Community Colleges might be willing to run an experiment for curricular mapping and linkages to job skills.

Another is semantic analysis. Starting with knowledge models, developed by domain experts, we are currently exploring ways to auto-generate learning outcomes from available data. For instance, given what we know about the content that our faculty are relating to their authored learning goals, we can infer other related content or requisite relationships or even identify missing learning goals that are evidenced by the data. In the same way we could consider analysis of expert-authored crosswalks to begin automatically identifying additional educational opportunities available for a particular job code, or conversely, we can better identify job opportunities that may be available to students that were not initially conceived.

Question:

- Which models can be implemented under current conditions (e.g. social, political, technological, and structural within education) and which will require change?

3. Strategic Scaffolding (Help for the Networked Learner/)

Scaffolding, an instructional strategy designed to promote deeper level of understanding and autonomous learning, is of particular relevance for online learning. It recognizes diverse pathways and forms of knowledge and expertise, and it takes into account learning experience, concept, and abilities. In an ideal MOOC, students should be presented with a great variety of content and activities, as well as feedback and support strategies that:

- illustrate concepts, problems, and processes in multiple ways to ensure understanding
- model a process before students are asked to complete their own
- allow connections to previous knowledge and experience
- provide instant feedback about level of understanding
- enable deeper level of absorption, understanding and application of knowledge
- offer a network of support comprised of peers and experts.

MOOCs bring together a diversity of participants with different levels of preparation and backgrounds and a variety of motivations, interests and needs. Understanding how these conditions inform the different pathways and levels of success is of critical to the goal of

[6] Standard Occupational Classification (SOC) systeml (http://www.bls.gov/soc/).

increasing equitable access to high-quality online learning opportunities at scale. Big Data driven research can help us understand the interaction among students, pedagogy, curricular material, support networks and the circumstances under which successful learning occur. More importantly, it would help create automatic support and scaffolding strategies for a wide variety of learners. It would provide the "temporary" learning structures that can enhance students' performance during a particular learning situation, gradually increasing the level of complexity needed to achieve mastery and higher levels of sophistication.

Both the **MOOC learning experience** as well as the characteristics/needs of the **networked learner** [7](such as their help seeking behavior) present added dimensionality that impact how and when help (timely, appropriate) can be provided or for that matter how best key aspects of quality such as "personalization" can be realized. They also suggest moving beyond the expert driven model of identifying misconception to a data driven model of understanding the learning strategies and behaviors of Networked Open Learners.

For example, an important consideration in the MOOC/online environment revolves around the possibility that scale is an essential "input" (i.e. more students participating in a course would actually improve the experience) which is the opposite of what we believe for face-to-face settings. The quality and diversity of interactions might actually improve in a large online course due to the level of participation in forums. Related is the fact that the MOOC learning experience – implies the ability to navigate an online experience – is one that involves forming and interacting with communities.

Questions:

- What models of digital scaffolding exist already? How is their success measured?
- In what ways do underserved communities currently benefit from access to online education? What are the conditions in which that happens?
- How can we create successful communities of practice? Is a "critical mass" of learners prerequisite for community engagement?
- How can we ensure the privacy of students and other participants?

[7] Drexler, Wendy. "The networked student model for construction of personal learning environments: Balancing teacher control and student autonomy." Australasian Journal of Educational Technology 26.3 (2010). http://www.ascilite.org.au/ajet/ajet26/drexler.html

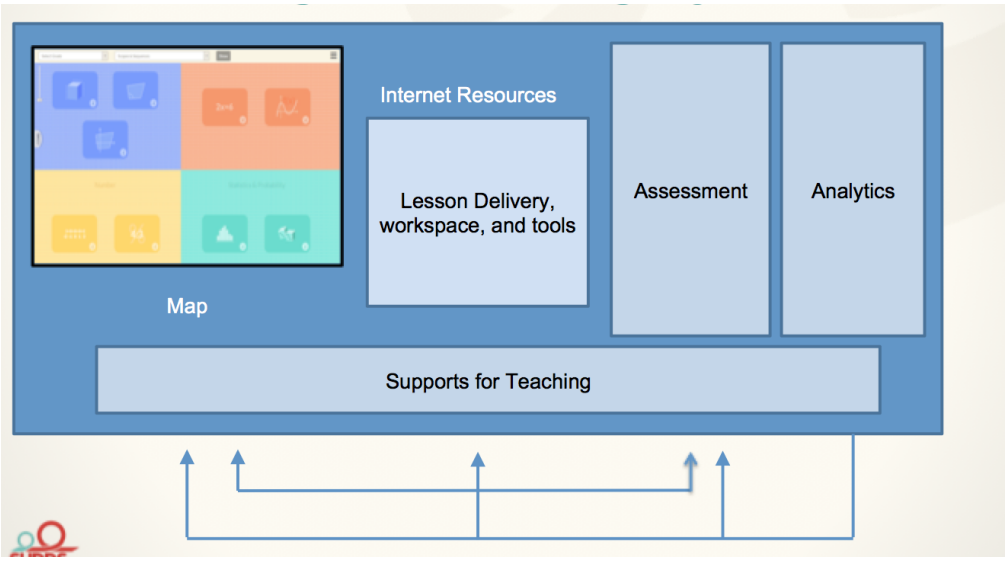
The Value of Learning Maps and Evidence-Centered Design of Assessment to Educational Data Mining (A Pre-Conference Thought Piece)

Jere Confrey (jere_confrey@ncsu.edu)
STEM Department
College of Education
North Carolina State University

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, Dynamic Maps, SUDDS). In this paper, I describe one of the projects, Scaling Up Digital Design Studies(SUDDS) at North Carolina State University, highlighting how its structure and design can inform efforts at applying big data in mathematics education. I propose that articulating *an explicit theory of student-centered learning* can help in leveraging "big data" to improve the *depth of learning* and not simply leverage 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 (DLS)

A representation of a digital learning system is shown below. 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.



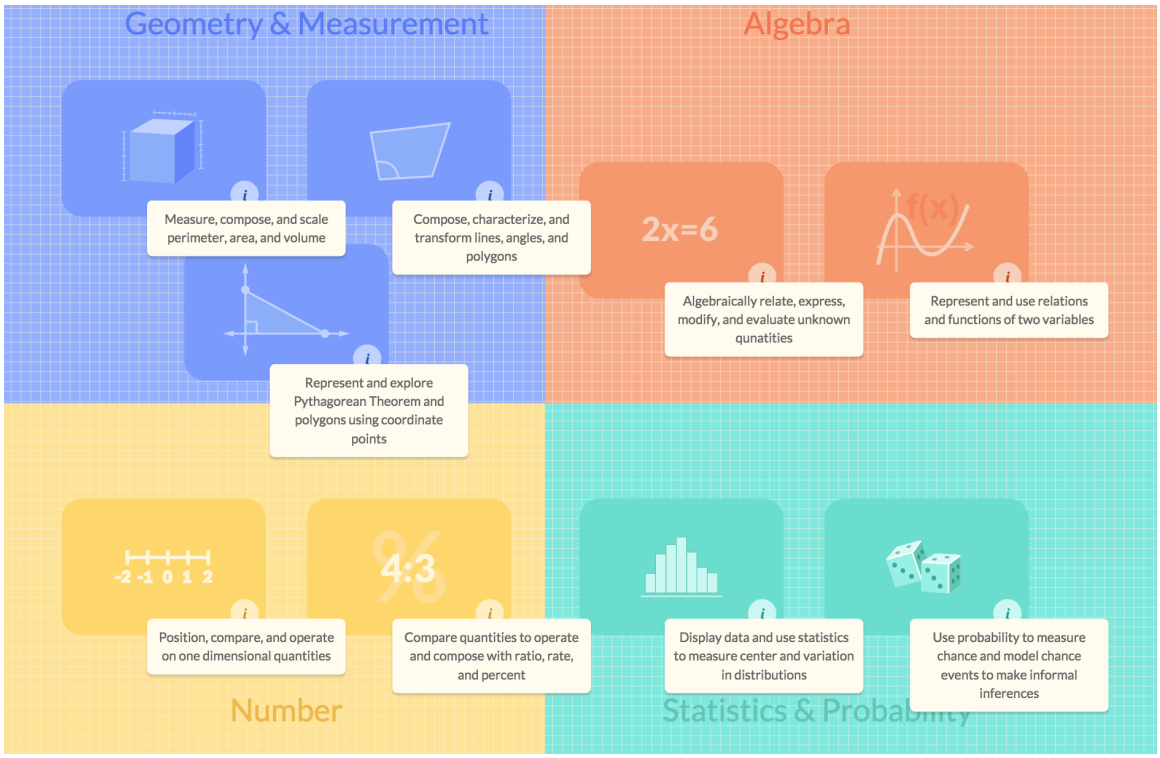
A DLS is student-centered when each of these components is designed to strengthen students' movement within the digitally-enabled space. A student-centered DLS:

- 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, 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.

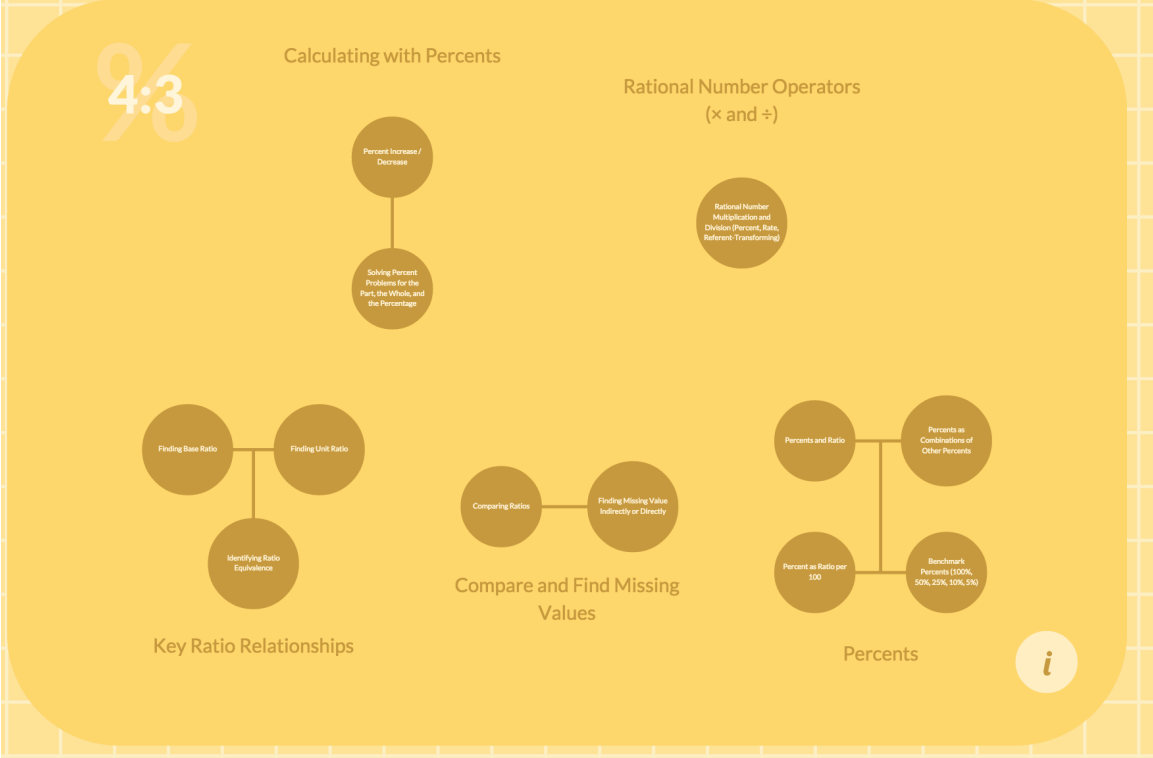
Student-centeredness does not imply individualization, working largely alone at one's own speed, but it does support personalization, making choices and self-regulation (Confrey, 2014). The DLS can be used by classes using predefined scope and sequences to coordinate activities.

Introducing the SUDDS Grades 6-8 Learning Map

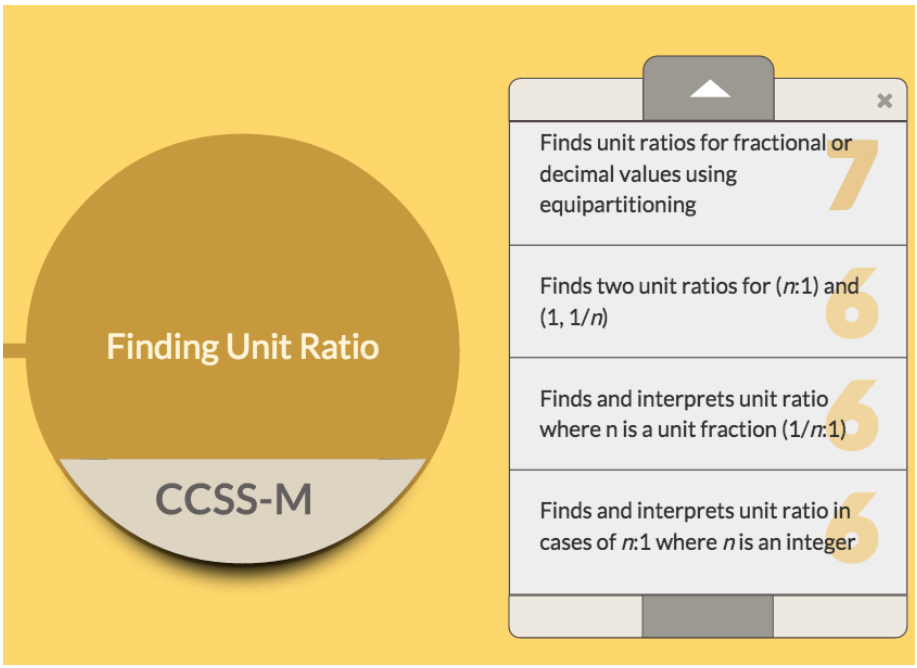
A learning map is a configuration in space of the primary concepts to be learned. Our middle school version is organized hierarchically to show the major fields in mathematics (number, statistics and probability, measurement and geometry, and algebra) and nine big ideas from across those fields. These 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 standard by standard which is problematic since standards vary in size and often apply to multiple big ideas.



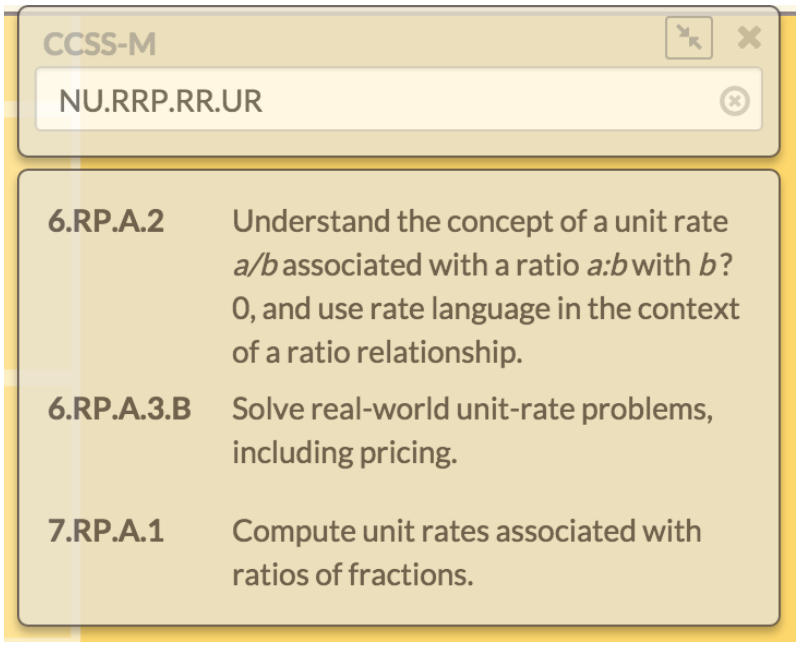
The level below the regions is comprised of *related learning clusters (RLCs*). At this level, the research on student learning has a significant impact on the map. RLCs are sets of *constructs* that are learned in relation to each other and their spatial configurations on the map convey to the user 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, 3) percents, 4) calculating with percents and 5) rational number operators. In a region or big idea, the RLC's organization from bottom left to upper right conveys to the users to address the first cluster, key 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 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 student 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.



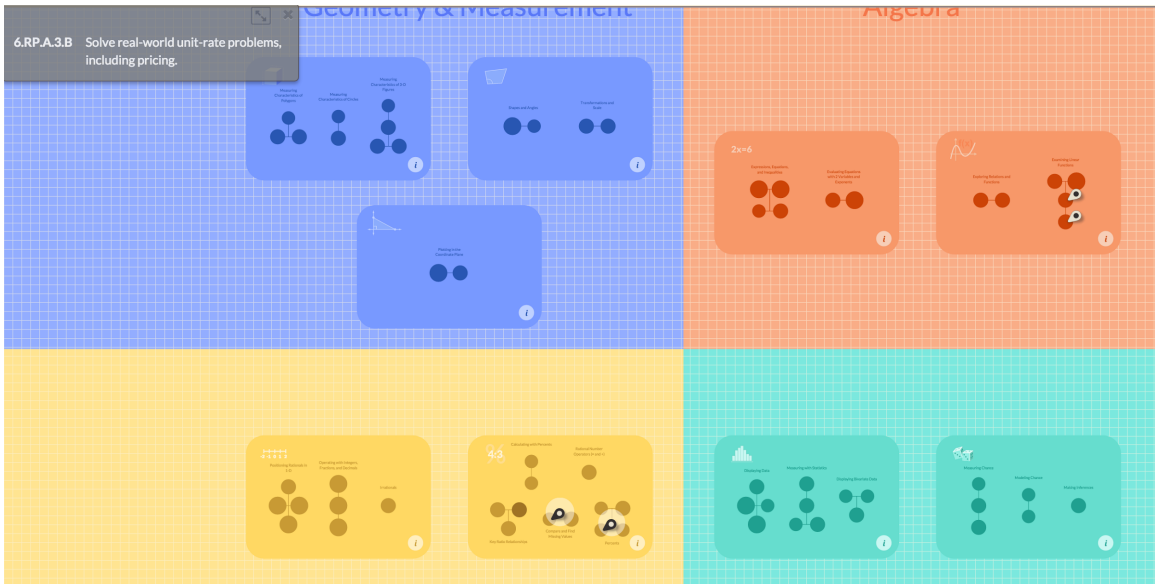
At the next level of detail, the construct level, a user have access to what is called the "learning trajectory stack." The stack details the typical behaviors, conceptions, and language of children as they learn and revise ideas from naive to more sophisticated. In the figure below, the first levels 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 one 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).



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 it is 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 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, a single or multiple grade levels can be selected to allow its flexible use across different grade configurations. Also, in construction, is a scope and sequence generator, 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 yet 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 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" which allow the systematic variation of the item parameters to show variations in processes across task classes (Confrey and Maloney,).

The assessments are constructed through a process of "evidence-centered design" (Mislevy and Riconscente, 2005) The value of a clear description of the student model in EDC was described by Mislevy, Behrens, Dicerbo, and Levy (2012)as:

> 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. (p. 30).

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 scientists 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.t Iems are carefully designed to be 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 (students and teachers) whether they are on track, 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 level 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 kinds 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 (Lipnevich A., & Smith, J. 2008). It appears that while both can be important, the task-oriented feedback tends to show improved effects on performance on cognitive task. However, person-oriented feedback can support self-efficacy and improve a student's perception of themselves 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 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-oriented, practice-oriented, involve problem solving, group work, individualized activity, 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 scope 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 then assign web resources and students 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' experience 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 capability of the current system. 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 recognize the importance of developing a set of practices that describe how mathematics is done. One element of a sophisticated DLS is then 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 (DESMOS, Geometer's Sketchpad, Cabri, Fathom, Geogebra, Tinker Plots). In addition, some tools exist for carrying out mathematical work and even for creating 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 (see Confrey, 2014 for a description of these design elements.).

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, when a DA was taken, how many times, percent correct and incorrect, 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 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). We hope in the future to gather richer data on student activity either from what is done using the digital materials or adding more opportunities for the capture of samples of student work or 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 as concerns the navigational elements of the system. The harder problems of providing expert advice to the user of what to do following particular results on the assessments will likely be the most significant and essential challenge.

Mislevy et al, warn that educational data mining (EDM) would benefit from considering how it links to the underlying cognitive models of the system it is mining, as they write, "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." P35-6

With respect to the cognitive student model underlying the SUDDS DLS, our analytic model would be helpful if it could inform us on 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 hopefully 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?

References

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.

Confrey, J., Hasse, E., Maloney, A., Nguyen, K. H., & Varela, S. (2011). *Designing Technology-Enabled Diagnostic Assessments for K-12 Mathematics*. A summary report from the conference designing technology-enabled diagnostic assessments for K-12 mathematics. Raleigh, NC.

Confrey, J., & Maloney, A. P. (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).

Lipnevich A., & Smith, J. (2008). Feedback: The effects of grades, praise, and source of information. June ETS RR-08-30

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

Mislevy, R., Behrens, J., Dicerbo, K., & 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.

A Foundation for a New Science of Learning[1]

George Siemens, University of Texas, Arlington

Overview

While learning analytics as a field has developed quickly, it has largely borrowed expertise from other disciplines and has failed to develop analytics products and platforms specifically for the education sector. Toolsets are being developed piecemeal and often have little interoperability with other tools or datasets. To create a new science of learning, personal knowledge graphs and open learning analytics platforms are required. This paper introduces these concepts and details the structure of both and the importance of funding to support their advancement.

Introduction

Learning analytics have to date primarily imported concepts from big data, computer science, some 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 insight that can be expected from a new knowledge domain that synthesizes and integrates insight 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 knowledge graphs (PLKG) 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 Knowledge 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/Stanford's OLI[2] and several corporate providers have gained attention, but are largely confined to courses with a clear right/wrong answers (such as statistics and math courses). In

---

[1] 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).

[2] http://oli.cmu.edu/

these instances, the learner's knowledge profile is kept within an existing software system or within a corporate platform. In education a Personal Learning/Knowledge Graph (PLKG) is needed where a profile of what a learner knows exists. Where the learner has come to know particular concepts is irrelevant – work, volunteering, hobbies, personal interest, formal schooling, or MOOCs. What matters is that all members involved in an educational process, including learners, faculty, 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. PLKG 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 (see figure 1).

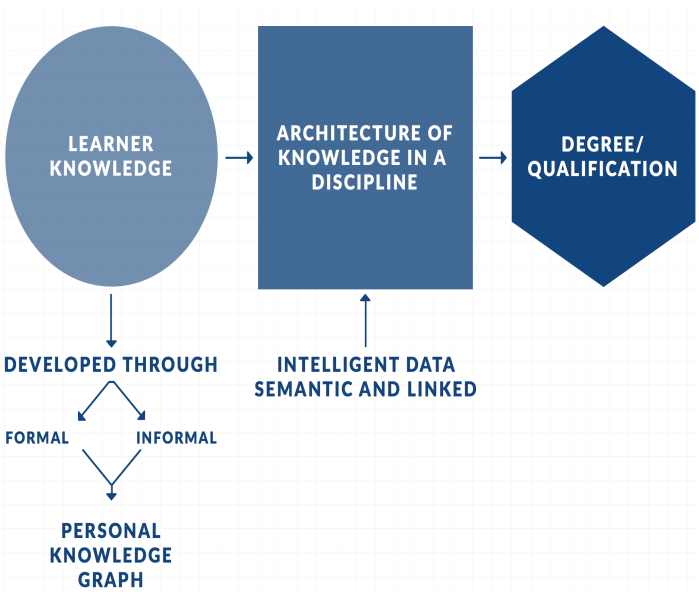


Figure 1: Matching Knowledge Domains to Learner Knowledge

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 learner knowledge 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, the career path alone 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" – see Figure 2. A student returning to university would have a range of course options, each personalized to her knowledge and skills, rather than be pushed through a pre-established curriculum without regard for existing knowledge. With PLKG, returning to university to up-skill and enter new fields – an increasing requirement as entire fields of work risk 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.

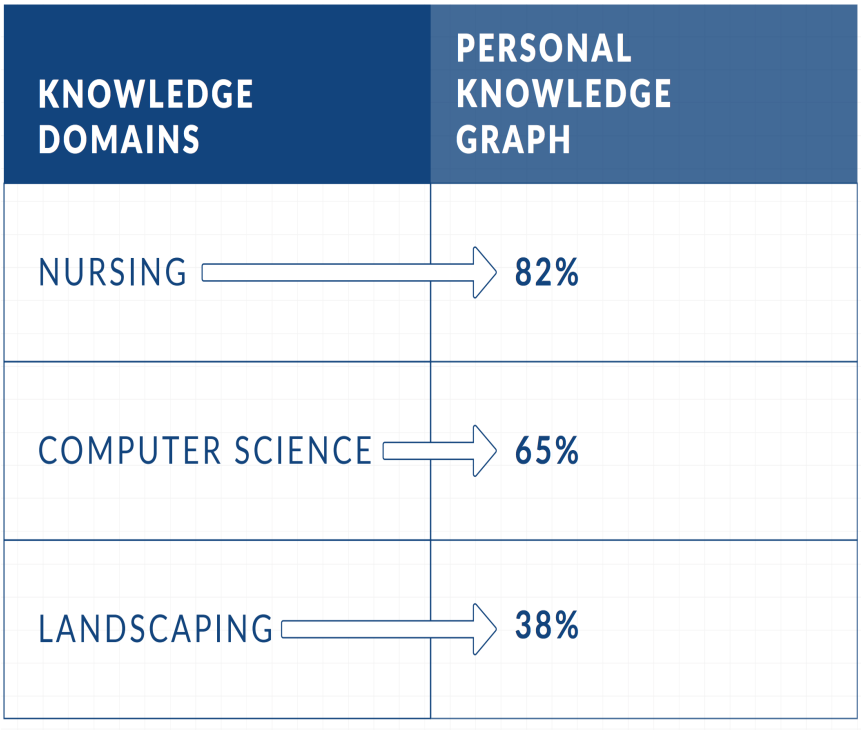


Figure 2: Returning and Advancing Degrees

Pedagogically, PLKG affords new opportunities for individuals to take personal control of their learning (see Figure 3). In this model, a learner can simultaneously engage with structured course content and create networked, connective, knowledge structures[3]. This approach is reflective of the networked world of learning and the personal lives of individuals as mobiles and wearable computing develop as critical technologies for knowledge work. In addition to algorithmically guided personalized learning, socially navigated personal learning provides opportunities for serendipity and creative learning. Learning pathways, within PLKG, are established by machine learning/algorithmic models and by personal learning networks and social interactions.

In order for PLKG to be effective, it 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[4] protocol where API access to certain types of information are brokered in a trusted environment. Essentially, learners

[3] 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

[4] http://www.imsglobal.org/toolsinteroperability2.cfm

would own their PLKG and standards would be established that permits trusted sharing with education providers.

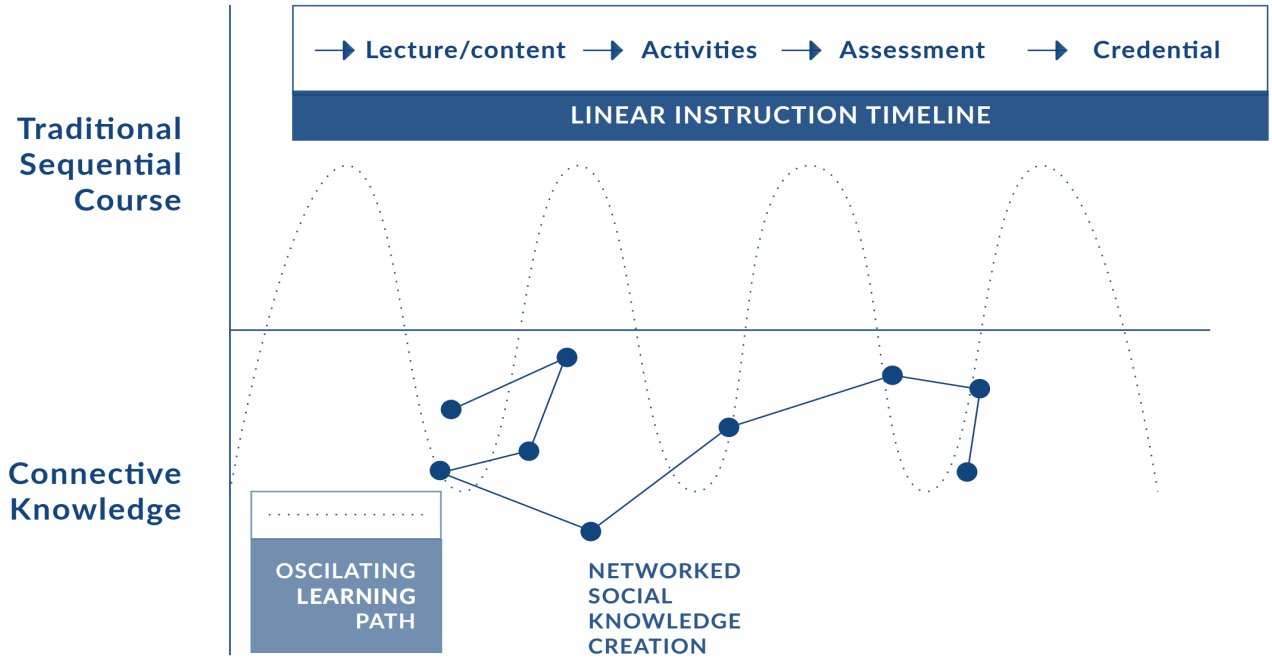


Figure 3: Flexible and Variable learning Pathways

Open Learning Analytics Platform

The open learning analytics platform addresses the need for integrated toolsets through the development of four specific tools and resources (see figure 4 for visual representation):

1. Learning analytics engine
2. Adaptive/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. The analytics engine incorporates data from learning management systems, social web, and physical world-data (such as classroom attendance, use of university resources, GPS-data when completing activities such as surveying), mobile and wearable technologies, and will leverage best practices from both the learning analytics and educational data mining communities. 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 as plugins in WordPress. The platform stays the same, but the functionality is extended by plugins. This then serves as a framework for identifying

and then processing data based on various analysis modules (see Figure 5). 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, or the process of concept development in small peer groups.

As LA develops as a field, plugins developed 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 (Open Learning Analytics Platform), we expect to see the rapid growth of educational dataming 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 realtime data to learners, academics, and institutions.

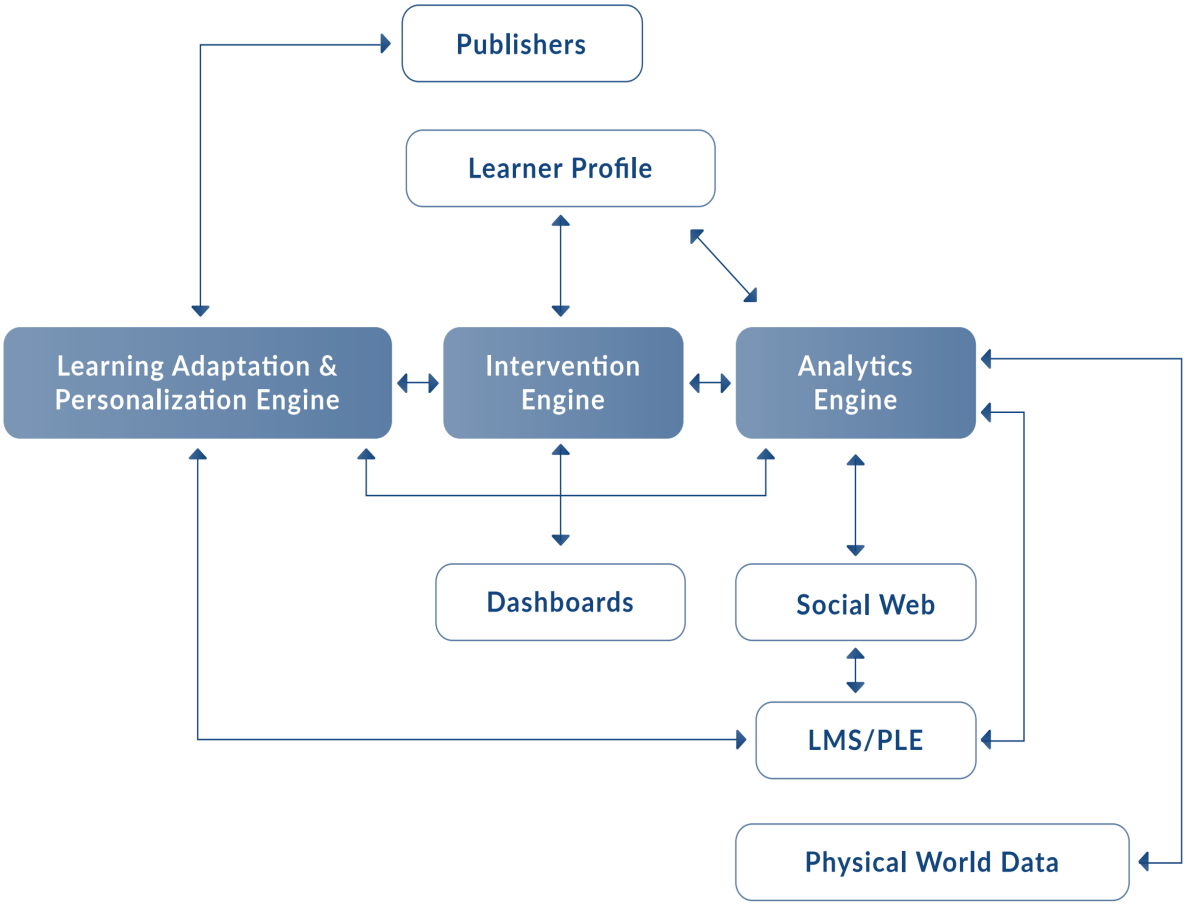


Figure 4: Conceptual Framework for Open Learning Analytics Platform

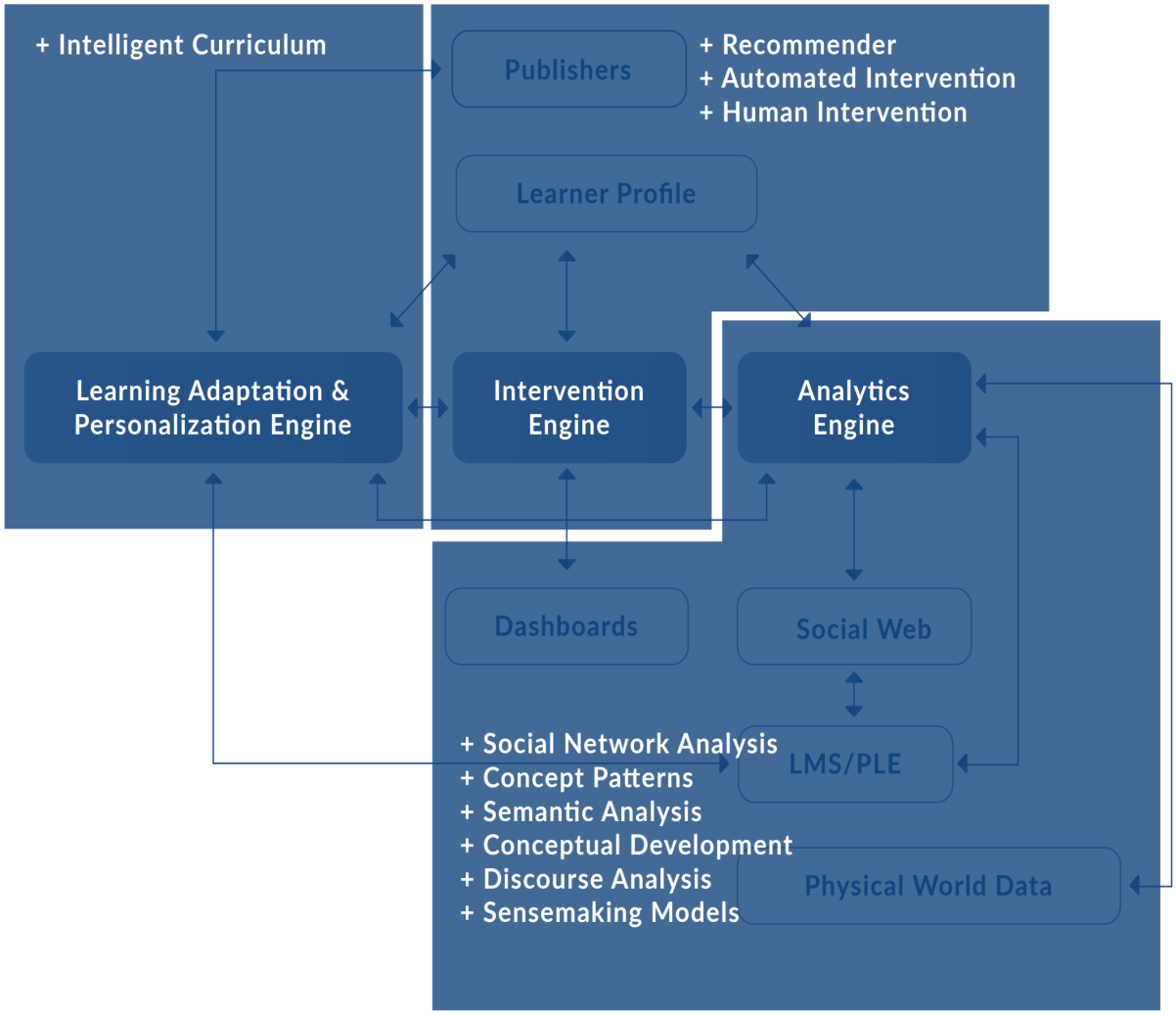


Figure 5: OLA Function areas

Adaptive/Personalization Engine
The learning adaptation and personalization will include adaptivity 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 such as instructional designers 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 as defined in the learning management system and social media sources (when permitted by the learner).

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 (books), Spotify (music), and Bing or Google (search). 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 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 educators and tutors notices 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 sensemaking 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 core concepts of the course, and social networks to identify learners who are not well connected with others. Analytics for educators will, depending on the context, be generated real time as well as hourly or daily snapshots. The dashboard will provide institution-level analytics for senior administrators to track learner success and progress. When combined with academic analytics, this module will be valuable for analyzing institutional activities (business intelligence).

Based on criteria established through research of the learning analytics system (such as the impact of social connectivity on course completion, warning signals such as changes in attendance patterns, predictive modeling), 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) at risk for drop 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. This paper proposes the development of Personal Learning Knowledge Graphs and an Open Learning Analytics Platform as 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 provided an open infrastructure for researchers, educators, and learners to develop new technologies and methods. In today's educational climate – 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 sensemaking agent in navigating uncertain change by offering leaders with insightful data and analysis, displayed through user-controlled visualizations.

Data-Intensive Research on Immersive Simulations for Learning

Chris Dede, Harvard University

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 (Dede, 2014). By immersing students in authentic simulations, MUVEs and AR can promote two deeper-learning strategies, apprenticeship-based learning and learning for transfer, that are very important in developing cognitive, intrapersonal, and interpersonal skills for the 21st century (National Research Council, 2012). However, complex tasks in open-ended simulations and games cannot be adequately modeled using only classical test theory and item response theory (Quellmalz, Timms, & Schneider. 2009). 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 be needed.

Illustrative Cases

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 scientific inquiry (both collaborative and individual) 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 ten 45-minute lessons and includes a complex scenario in which ecological change is caused by the interplay of multiple factors (Metcalf et al., 2013). 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 1, 2). 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 understand the complex causality of the pond ecosystem.

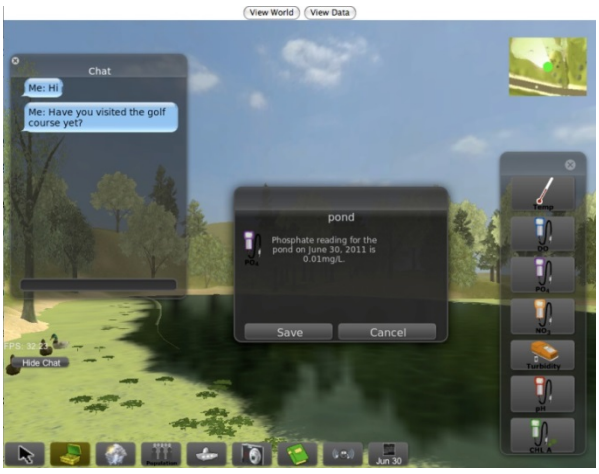


Figure 1. Students can collect pond and weather data

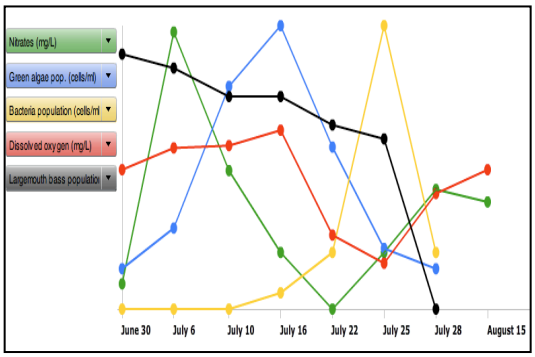


Figure 2. Summarizing and interpreting data

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 teams of four, students are given roles that embody specific areas of expertise (naturalist, microscopic specialist, water chemist, 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 inter-connected 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, 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, that document their group's assertions about its systemic causal relationships, with adduced supporting evidence. Using Go-Pro 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 teacher, 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 (Ketelhut, Nelson, Clarke, & Dede, 2010), we used log file data to generate event paths (Figure 3) 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 (2009) 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 location and time bins. Figure 4 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).

Figure 3. Event paths in RC for a three-person team

Figure 4. A heat map showing high-performing and low-performing students in RC.

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, although choosing the best way to display student paths through a learning environment is a complex type of visualization not well understood at present (Dukas, 2009). 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 (2007) 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. 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. As an 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 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 (Nelson, 2007). In addition to using the log files to personalize the guidance provided to each student, we conducted analyses of guidance use. We knew when and if 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 subsequent to viewing a given guidance message. This 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" (Johnson, Rickel, & Lester, 2000, p. 47). Beyond engaging students and providing a limited form of mentoring, APAs have advantages for interwoven diagnostic assessment in immersive authentic simulations in two respects: 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. Both 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. 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 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 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 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, 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 science education (http://ecomobile.gse.harvard.edu). The EcoMOBILE curriculum is a blend of the EcoMUVE learning experiences with the use of geo-located digital experiences that enhance students' real-world activities (Kamarainen et al., 2013). 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 on-line 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 5, 6).

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 field trip. 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.

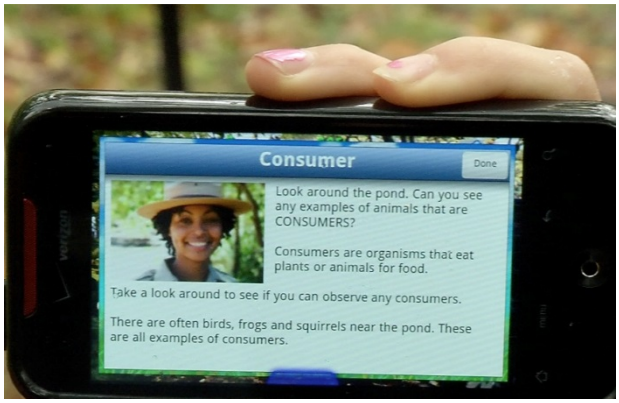


Figure 5. Handheld device delivering information

Figure 6. Collecting water data on turbidity

Parallel to EcoMUVE, EcoMOBILE devices capture and store "big data" about motivation and learning for each student that includes time-stamped logfiles of paths through the real world and data collected in that ecosystem (e.g., images, sound-files, probeware), as well as geo-located interactions with digital augmentations (e.g., simulations, guidance systems, assessments). Using Go-Pro 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, these data combined could support rich types of feedback to students, teachers, and researchers.

The Challenge

Quellmalz, Timms, and Schneider (2009) examined issues of embedding assessments into games and simulations in science education. 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 (Williamson, Bejar, & Mislevy, 2006). First, completion of the task requires the student to undergo multiple, nontrivial, domain-relevant steps and/or cognitive processes. Second, multiple elements, or features, of each task performance are captured and considered in the determination of summaries of ability and/or diagnostic feedback. Third, the data vectors for each task have a high degree of potential variability, reflecting relatively unconstrained work product production. Fourth and finally, 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. Bennett, Persky, Weiss, and Jenkins (2010) described both progress in applying probabilistic models and the very difficult challenges involved. 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.

In summary, 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 log files—make it easy 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: 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 Go-Pro cameras to record learners' visual perspectives and verbal utterances as their team interacts can provide another resource for analysis. Given the engagement, evocation, and evidence immersive learning provides, these media are among the most powerful and valid instructional/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.

Acknowledgments

The EcoMUVE and EcoMOBILE projects are supported by research grants from the Institute of Education Sciences (IES) of the U.S. Department of Education; the Qualcomm, Inc. Wireless Reach initiative; and the National Science Foundation. We also thank Texas Instruments and MoGo, Mobile, Inc. for resources and support.

References

Beal, C. R., & Stevens, R. H. (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.

Behrens, J. T., Frezzo, D., Mislevy, R., Kroopnick, M., & 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.

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

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

Dede, C. (2012). *Interweaving assessments into immersive authentic simulations: Design strategies for diagnostic and instructional insights* (Commissioned White Paper for the ETS Invitational Research Symposium on Technology Enhanced Assessments). Princeton, NJ: Educational Testing Service. http://www.k12center.org/rsc/pdf/session4-dede-paper-tea2012.pdf

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.

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11, 47–78.

Ketelhut, D. J., Nelson, B. C., Clarke, J. E., & 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.

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

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.

Quellmalz, E. S., Timms, M. J., & Schneider, S. A. (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.

Shute, V. J., Ventura, M., Bauer, M. I., & 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. J. Cody, & P. Vorderer (Eds.), *The social science of serious games: Theories and applications* (pp. 295–321). Philadelphia, PA: Routledge.

Williamson, D. M., Bejar, I. I., & Mislevy, R. J. (2006). *Automated scoring of complex tasks in computer-based testing. Mahwah, NJ: Erlbaum.*