D R A F T
9/23/13

Education Data for All: Forging Policies, Developing Structures

Una-May O'Reilly, Ph.D., ALFA Group, CSAIL, MIT
Lori Breslow, Ph.D., TLL, MIT

At least one of the promises of massive open online courses (MOOCs) has already been fulfilled: their potential to generate huge amounts of data about learners and the process of learning. edX and Coursera, which, together, offer a total of over 500 courses and count close to six million participants, have the ability to record every "click" of each of those users in every course. The extent to which we can "peer" into the experience of learning for a set of extraordinarily diverse learners is unparalleled.

As one example, researchers studying the first edX MOOC, "Circuits and Electronics" (6.002x), had access to 230 million interactions with course components, almost 100,000 posts on a discussion forum, and an end-of-course survey from over 7,000 students.[1] This research provided a sophisticated picture of user demographics, the course resources they accessed, and the pathways they took as they moved through the class. The data also allowed researchers to begin to understand what contributes to persistence and achievement in MOOCs, as well as what gets in the way of those goals. With the data generated by MOOCs, we have the potential to answer the fundamental question posed by Duke University Professor Cathy Davidson at the dawn of the MOOC era: "What modes of learning work in what situations and for whom?"[2]

But this sheer amount of data also poses enormous challenges. In the case of 6.002x, it took computer scientists several months to organize the data before educational researchers could begin to undertake analyses.[3] Even with the data in usable form, researchers had to wait until questions about privacy and the institution's responsibilities under FERPA were resolved. As of this writing, the three major MOOC providers—edX, Coursera, and Udacity—are built on separate course platforms, with no methods of sharing data, let alone explicit incentives to do so. Policies still do not exist to guide stakeholders through well-defined protocols to access data; instead those data are available on a case-by-case basis only.

Although the potential exists to use MOOC data to benefit students, instructors, and institutions of higher education, as well as those who have a stake in producing college graduates who will flourish in a global economy, as yet there is no "ecosystem"[4] with technology, protocols, policies, structures and community commitment in place to support those aims. MOOCs challenge higher education to grapple with issues of data interoperability, structure, and organization, as well as to create policies for access that do not hinder legitimate and productive uses of those data.

Proposed Project: A Dartmouth Conference for Educational Data

We suggest an organized effort similar to the Dartmouth Conference in 1956, which is generally acknowledged to have launched the field of artificial intelligence. Specifically, we propose organizing a week-long conference that will bring together educational researchers and computer scientists with system design, machine learning, database and/or data mining expertise to begin an

---

[1] Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom," *Journal of Research and Practice in Assessment,* http://www.rpajournal.com.
[2] Davidson, C. (2012). What can MOOCs teaching us about learning? Retrieved 10/1/12, from http://hastac.org/blogs/cathy-davidson/2012/10/01/what-can-moocs-teach-us-about-learning.
[3] This work was done by Dr. Daniel Seaton and Dr. Yoav Bergner, MIT's RELATE (Research in Learning, Assessing, and Tutoring Effectively) group, led by Professor David Pritchard, and continued by Dr. Jennifer DeBoer, MIT Teaching and Learning Laboratory.
[4] President's Council of Advisors on Science and Technology (2010). Report to the President realizing potential of health information technology to improve healthcare for Americans: The path forward. Retrieved 9/22/13, from http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf.

effort to address the challenges posed by MOOC data as well as the opportunities they present. We believe it is imperative this effort be undertaken as soon as possible for data is amassing at an ever-expanding rate. Processes and policies need to take into account the current technology and methods of online learning while anticipating what will happen in the future. We expect conference participants will focus on the tasks identified below, outlining a process for making progress on each:

- Outlining the principles of consistent local policies or a single global policy for data access that take into account issues of privacy and legal constraints while striving to be as open as possible. Identifying enabling technology and providing a roadmap for its development and deployment where it will serve the entire community.
- Developing a process of identifying required protocols, reaching consensus on them and updating them. Protocols may be necessary for data archiving and sharing, extracting research study data, citing data provided for others, and following specific steps from raw collection up to open access.
- Outlining necessary common data organization standards and practices for data consistency and provenance. This will entail identifying questions that educational researchers, policy makers, institutions, and commercial providers will ask of the data.
- Ensuring the education research community has adequate access to technology that allows straightforward data access and provides commonly required analytics.
- Identifying methods to incentivize constituents to adopt protocols, structures, and policies.

Deliverables

The deliverable of the week-long conference will be a paper that is a blueprint for a way forward to tackle each of these issues. An immediate benefit will be that experts from computer science and educational will be together in one place at one time discussing their particular perspectives and the contribution that each field can make to the effort. We hope fruitful long-term collaborations will be a by-product of the meeting.

The final deliverable will be a report comparable to "Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward" (2010) written by members of the President's Council of Advisors on Science and Technology (PCAST).

The problems facing higher education have become a national priority, and, in turn, MOOCs have catalyzed higher education. While they are certainly not the answer to all the challenges higher education faces, their ability to collect massive amounts of data about learners and the process of learning provides an opportunity never available before. By meeting the challenges brought to the forefront by MOOCs, we can create processes, structures, and policies that will allow the potential of educational data mining to be met.