7 March 2014 draft

Conference Rationale: Digital Learning Data as a Public Good:
Forging First Principles and Protocols for Scientific Collaboration

Asilomar Conference Grounds, Pacific Grove CA, 1-4 June 2014

Susan S. Silbey, ssilbey@mit.edu
Mitchell L. Stevens, mitchell.stevens@stanford.edu

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, Stanford, and Coursera alone offer hundreds of courses and count millions of participants. These are but three entities in a rapidly growing ecology of instructional production and consumption through digital media.

Data generated through online instruction offer unprecedented opportunities for building knowledge, but they come with substantial technological and ethical puzzles. Online learning tools are built on myriad separate platforms. There are few easy means of sharing data among providers and across platforms and many disincentives for doing so. Making these data accessible to researchers has complicated privacy, discretion, and equity implications. There is an urgent need to review and possibly clarify both federal and university policies on access and use of such data.

In February 1975 approximately 140 biologists, lawyers and physicians met at the Asilomar Conference Grounds in Pacific Grove, CA, to write voluntary guidelines designed to ensure the safety of recombinant DNA technology, which had only recently been invented. Due to potential safety hazards and public concern, scientists worldwide had halted experiments using recombinant DNA technology, which entailed combining DNAs from different organisms. The conference produced guidelines for safe handling of biological materials in laboratories and as waste, was eventually encoded in federal regulations, and is often credited with spurring the exceptional growth of biological research in subsequent decades. Applying a version of the precautionary principle, the conference also placed scientific research more fully into public discourse. Similarly, the Belmont Report for the Protection of Human Subjects of Research, which stipulates general principles governing research with human subjects, was also first drafted at a four-day conference of a National Commission at the Belmont Conference Center in Elkridge, MD in 1976. The products of these two conferences set the foundations of biological and social science research for almost forty years. We now face comparable research and human protection challenges emerging from the rapid proliferation of digitally mediated instruction.

Working in tandem with colleagues from our respective home institutions and with others nationwide, we are organizing a four-day workshop of educational and learning researchers, legal experts, sociologists and computer scientists with system design, machine learning, database and/or data mining expertise to address the data structure and data privacy challenges posed by online learning data. We believe it is imperative that this effort be undertaken as soon as possible because the number of parties producing and consuming these data grows almost daily. Processes and policies need to take into account the current technology and methods of online learning while also building a scientifically productive and ethically responsible future. We expect conference participants will focus on the tasks identified below, specifying clear steps for progress on each:

(1) Articulate guiding principles for consistent local policies regarding data access that account for issues of privacy and legal constraints while striving to be as open as possible. Imagine a socio-technical infrastructure comprised of technology and institutions and provide a roadmap for its development as a national and ultimately global public good.

- Consider the privacy interests that online learners may possess, and the importance of meeting privacy expectations to the success of online learning endeavors. Consider ways to simultaneously accommodate the interests of learners, instructors, and scientific researchers.

- Identify the range of data types that should be covered by any proposed principles or policies, and conditions under which these data or portions thereof can be made routinely accessible to the professional scientific community;

- Examine relevant legal requirements relating to access to, and use and distribution of, online learning data, including FERPA restrictions, NSF and other federal-agency data sharing and retention requirements, including open data initiatives currently under consideration by the federal government;

- Recommend policies and procedures for storage of data, for appropriate protection of personally identifiable information prior to release to researchers, for identifying appropriate personnel entitled to access data at different levels of de-identification, conditions under which such data may be published, and any other factors participants believe are important for protection. Such recommendations should be consistent with the currently standard practice that a faculty member teaching a class may use data from that class for guidance in improving teaching practice without requiring approval, but that using such data for publication or research purposes would still be subject to recommended guidelines.

(2) Outline broadly consensual protocols and conventions for data archiving and sharing, for citing data provided for others. Outline routines for the transit from from raw collection of data through to curated data access.

- Identify prior models of data standardization in education and in Web analytics. Some examples: IMS Global Caliper, PSLC DataShop, Advanced Distributed Learning Experience API, schema.org, Google Analytics, The MIT Core Concept Catalog (MC3).
- Specify the categories of stakeholders who are likely to want to use online learning data. Categories will certainly include academic learning researchers and other scientists, instructors (e.g., through instructor-facing analytics dashboards) students themselves (e.g., through student progress dashboards), and proprietary businesses of all kinds.
- Identify the units of data that might be used as input by researchers, instructors, educational organizations, and learners. These units might be keystrokes, events (e.g., a video view or answer to a question), or streams of keystrokes or events. Identify what contextual data might be connected to these units (e.g. demographic information about learners or organizational information about learning settings).
- Consider how instructors, educational organizations, and researchers may apply different pedagogical models – teacher-driven, student-driven, constructivist, cognitive, etc. – to data. The pedagogical model might be a particular analytic approach, and/or it could be in the form of an Application Program Interface (API).

(3) Identify methods to incentivize constituents to adopt shared data protocols and collaborative routines.

Outcomes

An immediate benefit of this conference will be that experts from the wide range of fields relevant to this new data science will be assembled simultaneously. We trust that fruitful long-term discussion and collaboration will evolve from the assembly.

The first written product of this conference will be a draft paper offering a blueprint for a way forward regarding the above issues. This draft document will be made available soon after the conference for distributed review nationally and worldwide. The final deliverable will be a report comparable to the Asilomar and Belmont Reports described above.

The problems facing higher education have become a national priority, and in turn have catalyzed awareness of the vast potential for improving educational knowledge and practice through digital media. While these media are certainly not the answer to all of the challenges facing higher education today, their capacity to produce massive amounts of empirical information provides opportunities never before available to scientists, practitioners, or policy makers. There is good reason for optimism.