Ad Hoc Committee on Privacy of Student Records

Final Report

May 31, 2014

Hal Abelson, Bradley Abruzzi, Mary Callahan, Isaac Chuang (chair),
Michelle Hanlon, Eamon Kearns, Paul Lagace, Krishna Rajagopal,
Susan Silbey, Lydia Snover, Danny Weitzner

Executive Summary

With the rapid growth in data from online courses such as those offered through MITx or edX, both internally at MIT and externally to learners elsewhere, and with the growing interest of researchers to use these data to better understand how teaching and learning occur, there is a need to review and possibly to clarify MIT's policies on access to and use of such data. While the Committee on Student Information Policy has oversight of traditional student academic records, such as those maintained by the Registrar, the growth of more detailed interactions of students through a subject and the use of MIT subject material by learners not registered at MIT raise additional issues concerning privacy expectations, legal constraints on access to data, and appropriate data curation.

The Charge: The Ad Hoc Committee on Privacy of Student Records was charged by the Chancellor to consider these issues and issue a set of recommended policies and procedures for curation, maintenance, and access to learner data acquired through online delivery of subject material.

The Process: The Committee studied a number of actual and potential scenarios, noting that MITx courses on edX currently have over 1.2 million registrants, who have generated over 997 million data records. MIT has received dozens of requests for these data, from sources at MIT and elsewhere. This intense interest in the data has generated tremendous pressure for immediate action to fill something of a policy vacuum, to settle questions of institutional interest, such as how compliance with applicable privacy laws — namely, the federal Family Education Rights and Privacy Act ("FERPA") — should be certified and monitored, who should be responsible for online learner data de-identification, and how to resolve MIT student privacy issues related to use of MITx and other online learning systems.

The Findings: Overall, the Committee finds that MIT should act now to adopt and promote online learner data policies and procedures. Broadly across the Institute there is a general lack of awareness of the issues, and recognition of responsibilities, related to online learner data privacy. Research and policy implementation are

evidently being slowed by the lack of a clear point of responsibility for online learner data handling. Moreover, the current MIT Student Information Policy, conceived before MITx with only on-campus students in mind, is missing consideration of non-MIT learners. De-identification of data reduces the risks associated with using and disclosing online learner data for research purposes, but it is widely acknowledged that complete and irreversible de-identification of data is not possible. Expert attention therefore must be given to the question of balancing data de-identification with data utility. Finally, it is evident that with respect to privacy of MIT student data generated or maintained by third-party online learning tools, the advancing wave of adoption has far outpaced policy. By way of example, we understand from its advertising that Piazza, an online forum tool, has been used by over 360 classes at MIT, involving over 4,800 students. The privacy issues raised by such third-party tool adoption go beyond the scope of this Committee's work and call for serious consideration of complex risk management issues, in a discussion that should include student involvement.

The recommendations: The Committee recommends immediate action to:

- **Construct and curate as a public trust learner data from MITx** and other online courseware systems in use by MIT, following a general policy based on existing principles of faculty governance, for data access and management, including:
  - Establishment of a **Learner Data Trustee** (the Director of Institutional Research in the Office of the Provost);
  - Designation of a standing **faculty committee for Learner Data Access**; and
  - Formalization of data access procedures and regulations according to the **Learner Data Access Policy and Procedures** (draft included herein).

The Committee further recommends, as secondary actions, that MIT:

- Revisit **definitions of categories of student data.**
- Base **learner-data de-identification** on "expert determination" and new approaches such as differential privacy, in lieu of checklist approaches like "Safe Harbor."
- Establish a new **Information Security Policy** for handling student and institutional records, with methodology and appropriate exceptions to be established (*e.g.*, by the Information Technology Governance Committee).
- Charge a committee, with a broad mandate and student representation, to **explore the issue of on-campus use of third-party online educational tools**.

I. Introduction

The Ad Hoc Committee on Privacy of Student Records was charged by the Chancellor's office to:

- Consider the privacy interests that online learners may have, the relevance and importance of meeting these expectations to the success of the MITx and edX enterprises, and ways to accommodate both the interests of researchers and the interests of learners;

- Investigate best practices of peer institutions in dealing with data access, storage, and security;

- Identify the range of data types that should be covered by any proposed policies, and conditions under which this data or portions thereof could be released or otherwise made accessible to researchers;

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

- Recommend policies and procedures for storage of data, for protection of personally identifiable data prior to release to researchers, for identifying appropriate personnel entitled to access data in various stages of de-identification, conditions under which such data may be published, and any other factors the committee believes are important to the protection of the data. Such recommendations should be consistent with 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 using such data for publication or research purposes would still be subject to recommended guidelines.

This final report summarizes the findings of the Committee, in its work between January 7, 2014, when it was first convened, and the date of this report, May 31, 2014.

We begin (in Section II) with a description of the current state of affairs, including six illustrative scenarios. Findings are then presented, considering (Section III) the definitions of student data generally, including the issues associated with "de-identifying" data; (Section IV) the best practices in industry and among our peer institutions with regard to online learner data; (Section V) proposed data security policies for online learner data; and (Section VI) a proposal for a formal process to enable the construction and curation of online learner data as a public trust, via a draft Student Data Access Policy and Procedures document.

II. Current state and scope of issue

How large an issue for MIT is student data privacy, given MIT's recent commitment to online learning initiatives as an institutional priority? Whom does this issue concern? What are the current policies?

We establish the basic dimensions of the privacy issue below, by providing a sense of the volume of online data under discussion and describing where MIT stands with respect to student data privacy today. We then consider six hypothetical scenarios that draw out the scope and complexity of the issue and the needs raised.

II.1 The flood of online learner data

II.1a. Data from open online educational systems

MITx launched in the spring of 2012 with a single course, 6.002x Circuits & Electronics, that drew roughly 150 thousand registrants and generated 26 million records. In the summer of 2012, Harvard University joined MIT to found edX as a separate, non-for-profit corporation dedicated to online learning. In the time since (see Figure 1), MIT has offered 25 MITx courses on edX, drawing over 1.5 million registrants from over 195 countries and producing over 997 million data records.[1]

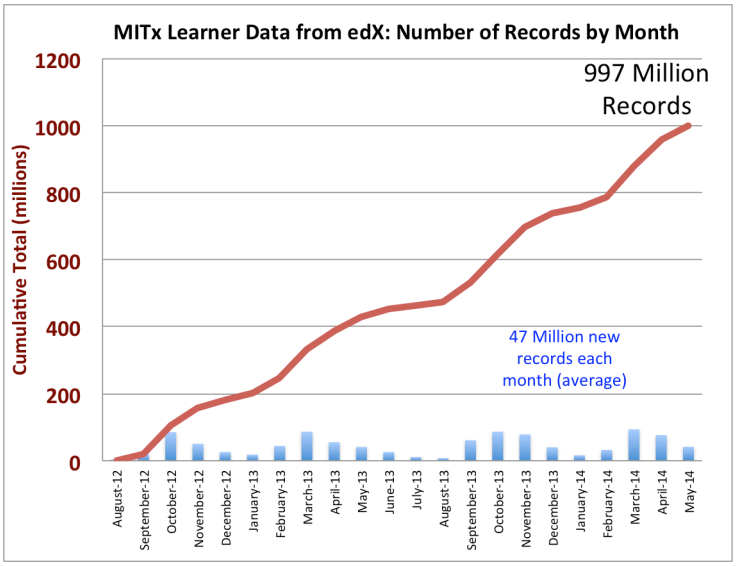

Figure 1: Volume of data from MITx courses on edX, with time, as of May 30, 2014.

It is important to point out here that in order to access MITx, a person must fill out a registration form and initiate a "sign up" button. Thus, MITx course content on the edX.org site is not accessible to casual browsers. Any further discussion of MITx

[1] See Ho, A. D., Reich, J., Nesterko, S., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). "HarvardX and MITx: The first year of open online courses" (HarvardX and MITx Working Paper No. 1).

registrants, or "learners," refers to people who have initiated this process via the registration form and sign-up button.

These MITx records from edX capture virtually every interaction of a learner with the edX servers, including not just timestamps of when video and text resources are accessed, but also detailed information about interactions with assessment problems and simulations; the data record numbers of attempted answers, responses tried as answers, and buttons clicked by the MITx learner, indicating, for example, when the learner requested "show answer." In addition, the data include detailed logs of online forum conversations, which include open-text messages authored by learners, along with photographs and other media the learners may choose to upload. It is widely believed that such rich data will make possible new insights and understandings about problem solving and teaching method efficacy[2].

The volume, intricacy, and richness of these data are unprecedented for MIT and are a reflection of the current moment of digital technology. It is estimated that as of Spring 2014, "massive open online courses" have drawn over 20 million registrants across all platforms, including edX, Coursera, Udacity, and others.[3] Other online courseware providers are also drawing increasingly huge audiences and generating large datasets. Khan Academy[4], a site founded by Salman Khan (MIT '98), reportedly draws 10 million (largely K–12) students each month; its YouTube channel boasts over 400 million views (versus MIT OpenCourseWare's 58 million). Quizlet[5], a site founded by Andrew Sutherland (MIT undergraduate, on leave since 2011), has reached over 100 million users and is the 95th most-visited website in the U.S.

II.1b. Data from "traditional" online learners

This growth of online learning is not merely a feature of the open web; it is a trend changing the nature of education at residential colleges. At MIT, starting in Fall 2012, three courses used the residential MITx system for a substantial fraction of material used in the courses. The number of courses using residential MITx

***MITx* @ MIT**

Semster	# Courses	# unique students	Courses
Fall 2012	3	~600	8.01rq, CC.801, ES.801
Spring 2013	7	994	5.11x, 6.041,6.s064, 8.011, 18.05, CC.801, ES.802
Fall 2013	13	1689	2.01, 2.03, 3.012, 3.091, 5.37, 6.341, 6.042, 8.21, 8.033, 8.021, 8.13, 18.03, CC.801
Spring 2014	27	2318	2.01, 2.003, 5.03, 3.039, 3.086, 3.091, 5.11x, 7.013, 7.014, 7.06, 7.28, 8.011, 8.02, 8.13, 8.421, 6.00, 6.002, 6.003, 6.042, 6.s076, 6.874, 16.003, 16.90,, 18.03, 18.05, 18.06, CC.802

Figure 2: Adoption of the residential MITx online learning system on campus, from Fall 2012 to Spring 2014.

[2] U.S. Department of Education "Expanding Evidence Approaches for Learning in a Digital World."

[3] Peter Shea, CGA Conference on Geospatial Technology and Online Education, Harvard University, May 2, 2014.

[4] http://www.khanacademy.org/; http://en.wikipedia.org/wiki/Khan_Academy

[5] http://quizlet.com/; http://en.wikipedia.org/wiki/Quizlet

in this way has roughly doubled each semester since (Figure 2), to 27 courses in Spring 2014. To date, over 65% of current undergraduates have taken an MIT class using the residential MITx system to deliver a nontrivial amount of course content. The residential MITx system is principally used to deliver interactive, instant-assessment, auto-graded problem set questions. These interactive problems generate a prodigious amount of information, essentially all of which is generated by MIT students; the Spring 2014 courses alone have resulted in over 15 million data records, to date.

More broadly across the U.S., the number of online students enrolled in diverse programs is rising. These students, like traditional students, are typically registered at a college or university and pay tuition; the student body includes not just residential college students, but also students enrolled in extension schools and continuing and professional education programs. It is estimated that 32% of the total U.S. college population of students are currently such online students[6]. The Harvard and U.C. Berkeley Extension Schools are both examples of extension school programs that have made substantial commitments to provide online curricula. Georgia Tech's offering of an online master's degree in computer science, in a joint program[7] with AT&T and Udacity, provides another example.

II.1c. Blurring of boundaries: third-party online tools in use at MIT

Online and on-campus learning environments are becoming increasingly mixed, not just within institutions, but also across multiple institutions. Specifically, students at MIT are increasingly using online learning tools provided by parties outside of MIT — and in many cases, unaffiliated with MIT.

A prominent example is Piazza[8], an online forum that specializes in "facilitating interactions among students and instructors in an efficient and intuitive manner." This product (currently offered for use at no charge) is very popular at MIT; the company advertises[9] that over 362 MIT classes, 4857 MIT students, and 307 MIT instructors have used Piazza. This includes classes from EECS, Biology, Math, the Sloan School, Mechanical Engineering, Biological

Figure 3: Piazza advertisement.

[6] Peter Shea, ibid; I. Elaine Allen and Jeff Seaman, "Going the distance: Online Education in the United States 2011," Babson Survey Research Group, 2011.

[7] http://www.omscs.gatech.edu/

[8] https://piazza.com/

[9] https://piazza.com/school/mit/l3

Engineering, and Political Science. Classes include large core MIT classes like 8.01, 18.02, 18.03, and 6.01.

Piazza positions itself as "a centralized place for instructors to conduct all class related communication," and its popularity arguably reflects its success, in the eyes of instructors looking to provide the best tools for their students. Indeed, we are advised[10] that U.C. Berkeley has recently executed an agreement with Piazza to provide campus-wide online forum services for its students.

Consider, however, the vast amount of data being gathered — about MIT students — by Piazza, and by numerous other third-party online tools in use at MIT, including Panda Grader[11], P2PU[12], DropBox, and Google Docs, to name just a few. MIT has agreements with some of these vendors, but for many vendors, MIT has not secured independent agreements. Who has access to student data being gathered by the third-party provider? How can MIT uphold its commitments to student privacy, given the accumulation of data about MIT students outside of MIT's authority and control, and given the growing use of these attractive third-party tools on campus?

While these are complex issues, it is clear that MIT student data privacy issues and online learner data privacy issues are intertwined and will likely grow even more interdependent, as boundaries continue to blur.

II.2 MIT Student Information Policy

II.2a. Policy statement

The definitive statement of MIT's position on the privacy of student data is given by MIT's Student Information Policy ("SIP"), which resides at section 11.3 of MIT's *Policies & Procedures* document. This policy applies to information held by MIT relating to enrolled MIT students. The SIP parses student information into three defined categories:

- Directory Information: data about students that, although FERPA protected, can be released for any reason without the student's prior consent (though subject to student's right to opt out of disclosure), and without a record being made of these disclosures. These data include:
    - Name
    - Home address, MIT address, email address,
    - Date of birth

[10] Personal communication, from U.C. Berkeley Professor Armando Fox

[11] https://www.pandagrader.com/courses

[12] http://learn.media.mit.edu/ ; https://p2pu.org/en/

  - Degrees received

- Other Student Information subject to FERPA (Family Education Rights and Privacy Act of 1974), which cannot be released without student prior consent, including:
  - Date of birth
  - Grades
  - Admissions information
  - Biographical information, including place of birth, nationality, race, ethnicity, photographs
  - Coursework, including papers and exams, communications between students and teaching staff, and between students and other students in the class
  - UROP and internship records
  - Disciplinary records
  - Financial records

- "Information subject to other or additional provisions," exempted from FERPA protection, including:
  - Personal files of institute faculty and staff
  - Campus police records
  - Medical records
  - Records of students as employees

The examples listed here are illustrative and not exhaustive. The thrust of the SIP is understood to restate and build upon the requirements of FERPA[13], which applies generally to student's "education records," a broadly defined term in that law. The SIP's definition of these three subsets of student information reflects provisions in FERPA that treat information in these categories differently.

[13] "FERPA" refers to the Family Educational Rights & Privacy Act, a federal statute that, along with supporting regulations of the Department of Education, (1) provides students with a limited right to review their "education records," as the law defines that term, and (2) imposes limited restrictions on institutions' use and further disclosure of education records maintained by the school. Consulting with Harvard and edX, MIT has concluded that FERPA's requirements apply with equal force to online learner data, notwithstanding that an MITx learner's relationship to MIT differs considerably in nature from an MIT student's. See Appendix E (walking through the legal analysis). Some institutions have concluded that FERPA does not apply, but we have not found that their positions withstand close scrutiny. The Department of Education, which administers the FERPA regulations, has not taken an official position on the matter, and we will continue to monitor and, as appropriate, participate in any regulatory developments relating to the privacy of online learner data. As it happens, FERPA's requirements provide a useful framework for balancing online learners' privacy interests against MIT's institutional research and instruction interests. Our recommendations in this draft report are undertaken on the assumption that FERPA applies.

It is important to note, however, that the SIP goes beyond FERPA in several areas, upholding a higher level of privacy than the law requires. Specifically, MIT restricts access to the following data, which are largely not covered by FERPA:

- Personal files of institute faculty and staff
- Campus police records
- Medical records
- Records of students as employees
- Parent's financial records
- Library circulation records
- Alumni records

II.2a. Policy intent and implications

The object of the SIP is to provide a way for a student to have some control over the disclosure of information from his or her educational records, as unchecked disclosure of personally identifiable information derived from education records could leave students in a vulnerable position.

At MIT, we have seen cases in which students have been stalked or harassed by other students or by people off campus. Although these situations arise infrequently, disclosure of student information could have serious consequences under these circumstances, and we believe that the Institute has responsibility to protect its students.

While a well-meaning faculty member, for example, might elect to use social media or a third-party service vendor such as Piazza to enhance students' class experience, there is a risk that doing so might result in the loss of control of data that are protected under FERPA and especially sensitive in this context. Thus, if a stalker were able to access unprotected data that indicate when and where a particular student may be in class, the stalker could wait outside the room for the student.

We recall the incident several years ago wherein a stalker threw a chemical concoction on a student after her class. The risk of incidents like this may increase in proportion to the amount of shared online data about our students, as well as the increased interactions among anonymous strangers.

MIT has historically chosen to use reasonable methods to reduce risk and retain a level of trust with its students, and we propose to continue in that same vein as we manage questions raised by online/ digital learner data.

II.3 Six Illustrative Scenarios

This short survey of the scope of online learning and the state of MIT privacy policies brings to the fore a number of concerns about privacy and online learning. We present six illustrative scenarios here to bring those concerns into sharper focus. Each of these scenarios is fictitious in the particularity, but not unrealistic, as several cases are drawn from actual events that have transpired at MIT.

II.3a. Scenario 1: MIT Student + MITx Course

Consider the scenario presented in Figure 4: an MIT student takes an MITx course on edX, possibly at the suggestion of a UROP supervisor, for example. Though edX was co-founded by MIT and MIT participates in its governance, it is an independent company that operates by its own rules. Notably, FERPA allows certain exemptions for access to educational data. Moreover, and quite importantly, FERPA's only mechanism for penalty for violations is to withhold federal funding from the educational agencies and institutions subject to its provisions. edX does not receive such funding; moreover, the researchers in India requesting the identified data may present a legitimate basis for their request: for example, they may have licensed 7.00x course materials and are seeking to study their efficacy. In such a case, edX would certainly be motivated to provide the data.

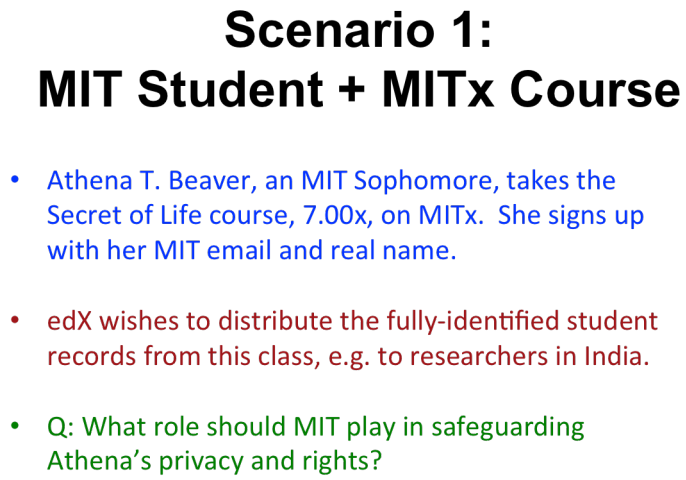

Figure 4: Scenario 1.

However, it happens that MIT does have a contractual agreement with edX, under which edX is bound to follow MIT's policies, with respect to privacy of learner data for MITx courses on edX. On the other hand, MIT currently has no established, formal policies about MITx learner data, other than that the data fall under FERPA. As an MIT student, should Athena T. Beaver's records on MITx receive greater protection, through the SIP, than other non-MIT-student learners taking MITx courses? Or is she just another online learner?

II.3b. Scenario 2: Professor wants de-identified student data

Consider the scenario presented in Figure 5: a professor requests data that are de-identified for FERPA purposes, and therefore not subject to FERPA protection, for students in his course to study.

The fact remains that, as we discuss in detail in Section III.2, no matter how carefully data are de-identified, they typically retain correlations ("quasi-information") with outside contextual information that can permit re-identification of the data, as this scenario relates.

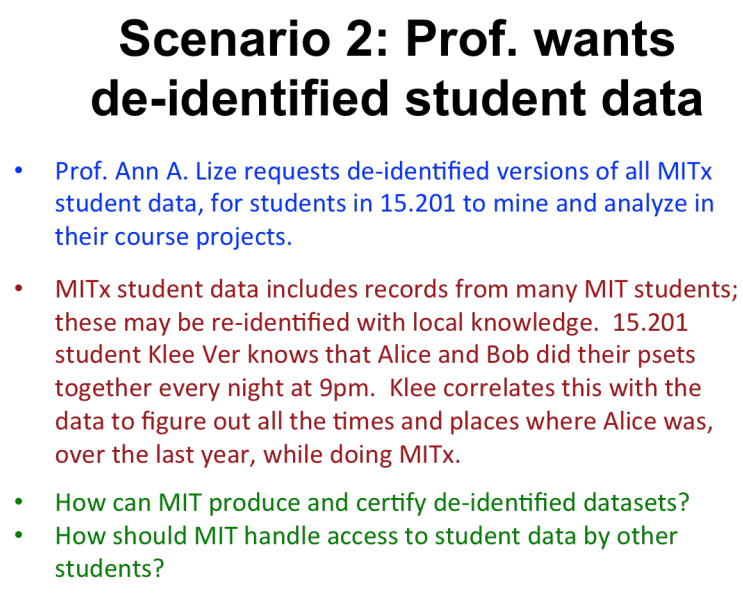

Figure 5: Scenario 2.

There are several ways to handle the questions posed by this particular scenario. For example, MIT might choose to impose a 4-year embargo on the use of MITx datasets by undergraduates, to mitigate the effects of quasi-information leakages. Another option would be to require all students using such data to undergo training in privacy rights and the ethical conduct of research.

This example thus points out the need for a clear point of responsibility for data requests, expertly executed de-identification procedures, and policies for data access that are adaptable to circumstances and will accommodate exceptional cases.

II.3c. Scenario 3: Researcher requests MITx learner data

Consider the scenario presented in Figure 6: a research scientist requests all learner data from all MITx courses, to be used in an experimental new privacy-protected open data access system.

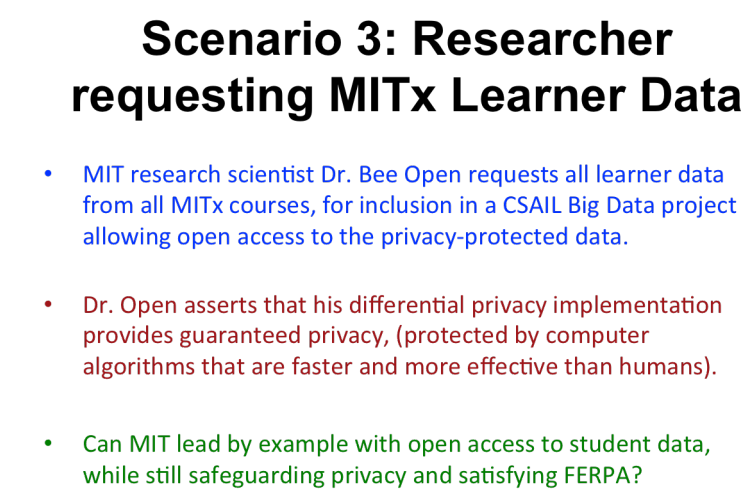

Figure 6: Scenario 3.

Clearly, MIT wants to help pave the way for research to push boundaries. On the other hand, MIT is obligated to uphold student privacy. This scenario illustrates the tension between these two mandates. In fact, *differential privacy*[14] is a very real and new mechanism — an alternative to de-identification — that the computer science community believes may hold great promise, but which has not yet gained widespread acceptance in practice.

Again, MIT might respond in this scenario in a number of ways. The Institute might deny access to Dr. Bee Open until the privacy implementation has been proven and used for non-student data for a period of time. MIT might make student records available if Dr. Bee Open signs documents accepting liability for breaches of security, though of course her ability to take on such liability is limited. Or Dr. Bee Open might offer his or her algorithm to the MIT Learner Data Trustee (defined in Section VI) to use in de-identifying data.

This scenario illustrates the point that data de-identification is a matter of institutional concern, and not something we believe that individuals or individual units can undertake on their own, even within MIT. MIT currently has data de-identification expertise in the Institutional Research section of the Office of the Provost, and it would make sense for IT to serve as the central team responsible for learner and student data de-identification.

II.3d. Scenario 4: Two professors want each other's educational data

Consider the scenario presented in Figure 7: two professors in rival departments each challenge the other's conclusions from an educational experiment and request the other's raw data for analysis.

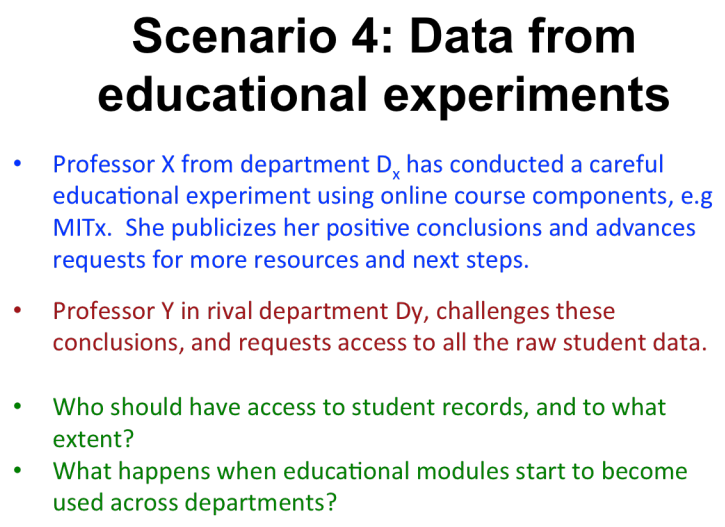

Figure 7: Scenario 4.

The principle that science is organized for the production and evaluation of knowledge claims is fundamental to MIT's educational and research mission.

[14] Dwork, C., "Differential Privacy: A Survey of Results," Lecture notes in *Computer Science*, vol. 4978, p. 1, 2008.

Thus, scientific knowledge results from transparent processes that make available to the community of observers the grounds (evidence, logic, interpretations) of the knowledge claims to be assessed, critiqued and improved with additional research. As a matter of academic freedom, Professor X can say what she believes to be the results of her experiment, but she cannot do so without making the details of the experiment and the results accessible to an audience for evaluation and critique.

On the other hand, scenarios like this one are complicated by the need to respect student privacy, and by a growing trend of re-use of educational materials. For example, the course content at issue in this scenario might be one or more thermodynamics modules, used by instructors in multiple departments. These modules might have authors across departments, and even across institutions. Each of the faculty may then quite reasonably assert rights to access student data, in order to study the efficacy of teaching materials they have developed or assembled.

This scenario illustrates the need for faculty oversight in the process of deciding who gets access to what data. Section VI presents a proposal for such oversight, which we would propose to assign to a Faculty Committee for Learner Data.

II.3e. Scenario 5: MIT Student + MIT course with online component

Consider the scenario presented in Figure 8: an MIT student takes an MIT course with required online components that are provided by a third-party vendor. This vendor does not have an existing agreement with MIT setting terms, conditions, and restrictions regarding use and disclosure of the student data and in fact is known to have a practice of selling student profile data.

This example raises two issues. First, MIT holds a responsibility to uphold the Student Information Privacy policy for its students. However, the student data gathered by the third-party vendor in this scenario are inaccessible to MIT and outside its control.

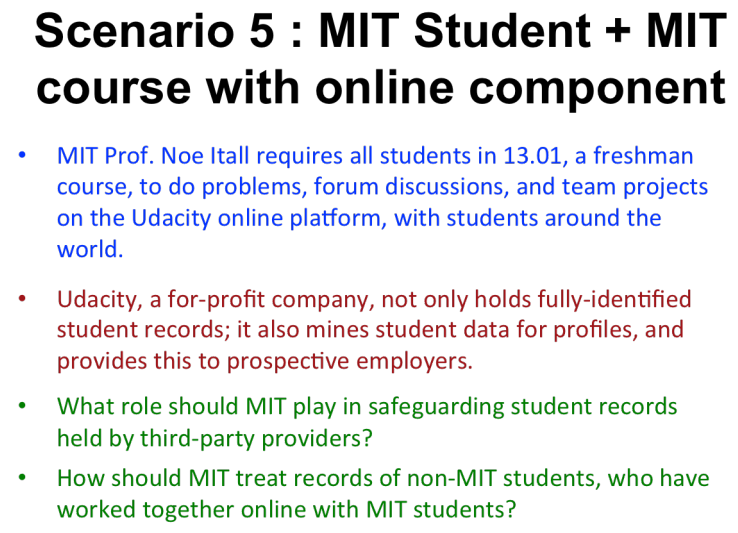

Figure 8: Scenario 5.

Second, the MIT students in this example are asked to interact with a worldwide audience of students, who are also using the same online platform. Data records of these non-MIT students thus may contain considerable data correlated with that of the MIT students. For example, there may be joint projects between both student populations. There may also be records of conversations by non-MIT students, about MIT students, possibly revealing personal information, including names.

We can see a number of ways to manage a scenario like this one, but none of them jumps out to us as optimal. MIT faculty might decide by vote that student assignments worked on third-party platforms cannot be a required part of an MIT course, unless MIT is able to ensure students' privacy. Or MIT might ask students to waive their privacy rights for work done through third-party platforms (we note that a student's waiver of privacy rights, required as a condition of registering in a course, is not a valid consent to disclosure of his or her education records under FERPA). A third option might be that MIT faculty can use third-party platforms only if the Institute secures agreement from the provider that it will abide by MIT's SIP. Alternatively, MIT might be satisfied with agreements whereby the third-party provider keeps only aggregate data that do not identify any particular participant and further agrees that no information identifying particular participants will be sold for use by for-profit firms. Last, MIT might inform faculty and students that the Institute cannot protect data in online third-party platforms, pushing the risk to individuals.

Facts on the ground have rapidly overtaken existing policy in this space (*e.g.*, the widespread use of Piazza at MIT today, discussed in II.1.c). Such is the breadth and intricacy of the issues presented by third-party service providers that this Committee believes MIT should charge a committee with a broad mandate to explore the issue of on-campus use of third-party online educational tools. This committee should have student representation. A natural place for this charge would be the existing Committee on Student Information Policy.

Meanwhile, MIT should move forward expeditiously to educate its faculty and staff — *i.e.*, the persons at MIT electing to use third-party tools —about student data privacy issues. MIT should also endeavor to make it clear to students what privacy they may be giving up, and where MIT may be unable to uphold the covenant of the SIP.

II.3f. Scenario 6: Beyond MIT

Consider the scenario presented in Figure 9: a prominent peer institution requests that MIT join its initiative to openly release learner data to the public for the good of the public.

The challenge is that an open online course platform such as edX's includes quasi-public components like online discussion fora. Data from these are known to allow re-identification of datasets purportedly de-identified via traditional procedures.

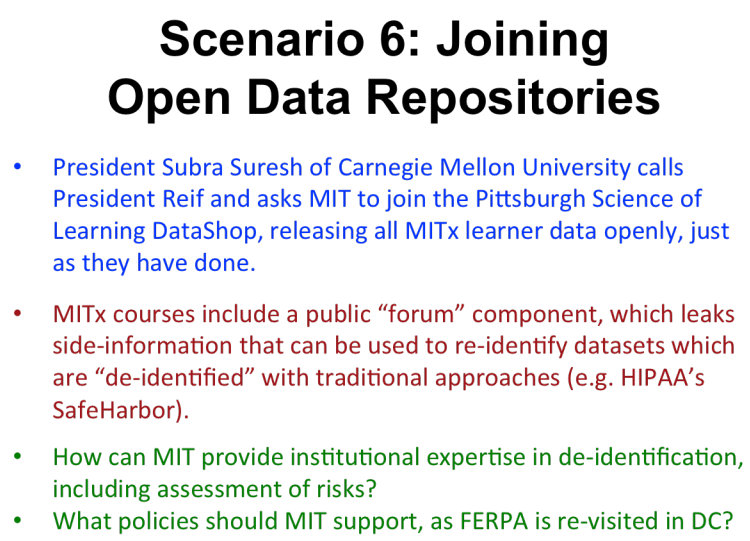

Figure 9: Scenario 6.

This scenario illustrates the growing tension between the desire to exploit patterns in large-scale educational data, and the desire to protect the privacy of individuals. MIT joins this conversation with an opportunity to provide leadership, given the role it has taken with MITx and edX, and provided that it articulates clear and well thought-out policies in a timely manner.

MIT should commit to the concept of constructing and curating, as a public trust, data from MITx and other online courseware systems in use by MIT.

Achieving this in practice will require clarity, such as given by the Learner Data Access Policy & Procedures, and points of responsibility, *e.g.*, the Learner Data Trustee and the Faculty Committee for Learner Data Access, ideas for both of which are proposed herein (Section VI).

Beyond policy, procedures, and governance, MIT should also strive to advance the boundaries of knowledge in this arena, for example, by exploring the use of novel approaches to guaranteeing student data privacy while opening doors to research — *e.g.*, via mechanisms such as differential privacy (Section III.2).

III. Student Data: Definition and De-Identification

We considered the following questions, in view of the Chancellor's charge to the committee, and with respect to existing policies and definitions:

- What levels of privacy are needed for the different kinds of student data that should be recognized by MIT?
- How might student data be "de-identified," in order to open up research opportunities by relaxing privacy-imposed constraints?

III.1 Definitions of Student Data

We begin by reviewing the state of student data definitions at MIT, then, after describing new institutional interests in using data arising from online course activities, we propose a framework for considering data policies that defines four categories of student information.

III.1b. Need for new data categories due to online activities

	Campus	MITx (on edX)	Non-MITx
MIT Student	SIP chapter 11	SIP 11.7 ?	?
Non-MIT Student		?	?

Figure 10: Coverage of existing student information policies.

Current MIT Student Information Policy (Section II.2) leaves unsettled issues arising due to the rise of online courses and course components at MIT, provided by MIT to the world via MITx or edX, or via combinations of on- and off-campus students and interactions. For example, as illustrated in Figure 10, data from non-MIT students taking MITx courses on edX are not addressed by the existing student information policy. Also, we recognize that the huge volume of "click-level" data from online courses can be of tremendous benefit to advancing education; these include data from online course components used by on-campus MIT students. But the desire to unlock this potential through research can be at odds with policy, particularly when constraints are unspecified and categories of data are unclear.

III.1c. Recommendations for definition of four categories of student information

We believe that an effective way to define categories of data is to approach the question through operational definitions, which address three questions:

- Who approves data access?
- Who is responsible for data curation?
- What are access and security procedures?

Given these questions, we suggest four explicit, but non-exclusive, categories of student data at MIT defined by how the data are operationally handled (and illustrated in Fig. 11):

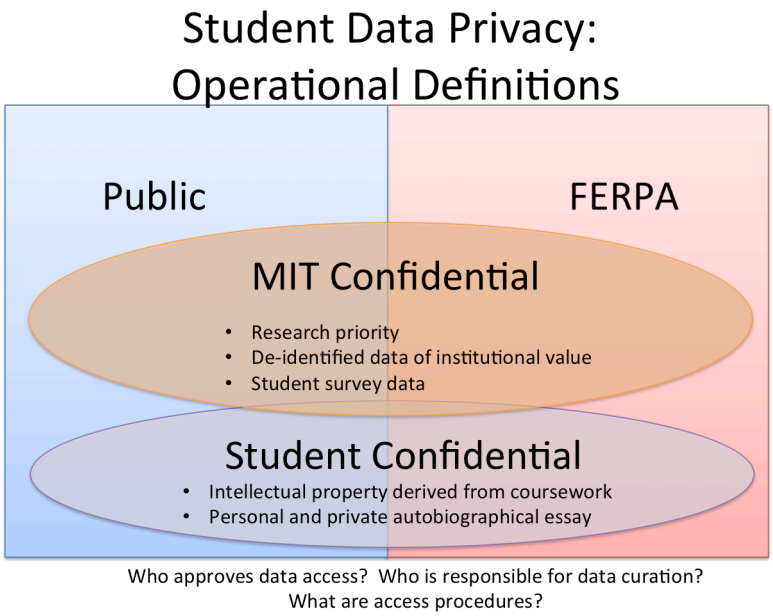

Figure 11: Suggested categories of student data.

- **FERPA-governed**: personally identifying student information subject to the access requirements and use and disclosure restrictions established by federal law.

- **Public**: student information not subject to FERPA restrictions that MIT freely publishes, *e.g.*, aggregate statistics on enrollment and gender across campus students and MITx registrants.

- **MIT Confidential**: student information that is not necessarily governed by FERPA, but that MIT desires to keep confidential for institutional reasons, *e.g.*:

    - Data of significant potential value for research;
    - De-identified data that could be public under FERPA but could facilitate undesired re-identification of released data;
    - Survey data of institutional sensitivity.

- **Student Confidential**: student information that is not necessarily governed by FERPA, nor sensitive for institutional reasons, but that MIT desires to keep confidential in order to safeguard student privacy and opportunity, *e.g.*:

    - Intellectual property derived from coursework;
    - A personal and private autobiographical essay submitted as coursework.

Example scenarios for MIT Confidential data:

- Research team working with course instructors designs A/B experiment in connection with a course, intends to write paper on the results. MIT may be concerned that other researchers who do not fully understand the study design may draw mistaken conclusions based solely on accessing the data.
- Every student in MIT.xx fails in a particular term. MIT may not want the grade average for that course to be released, because it would catalyze undesired revelation of individual student performance.
- MIT surveys its students on a sensitive topic, and desires to keep the results — including anonymized results — non-public.

Example scenarios for Student Confidential data:

- Essay written for class on ethics revealingly describes an embarrassing instance in which a student made an "un-thoughtful action," disclosure of which could subject the student to persecution or retribution.
- Final project in course (*e.g.*, 6.111) discloses innovation by student team that has commercial potential and might compete with an idea developed by grad student TAs of the class.

With regard to the three operational questions (data access approval, data curation responsibility, and access procedures), we suggest that existing institutional units at MIT be employed where possible. Specifically:

- Data access approval: Beyond what is covered in existing policy, we suggest that research access requests for student records from online learning initiatives require evaluation and approval by a data "Trustee," via a process that we describe in section VI and that will be regularly reviewed by a faculty committee. The Trustee may also in some instances consult the faculty committee about particular requests for access to the data.

- Data curation responsibility: The creation of labeled datasets of student records from online educational systems should be the responsibility of units at MIT that have already traditionally held this responsibility. New sources of data, *e.g.*, arising from MITx and edX programs, should be curated by a unit of the Provost's office, where the MITx program originated, and from which oversight and funding continues.

- Data access and security procedures: We recommend that in addition to adopting the data security policy proposed below, all distributions of,

manipulations of, and accesses[15] to FERPA, MIT Confidential, and Student Confidential data be logged. This will enable detailed audits, which is more in keeping with our needs at MIT, than restrictive access policies.[16]

III.2 De-identification of Student Data

We begin by reviewing the role of de-identification and the challenges inherent in attempting to create de-identified datasets, particularly for student data from online systems. This challenge is central to many fields, and we illustrate how the medical field deals with de-identification needs. We then present recommendations for MIT.

III.2a. The role of de-identification and its challenges

De-identification of data is a process that transforms data from one category of legal protection to another (*e.g.*, FERPA protected data to Public access), by aggregation of records and fields and removal of person identities associated with the data. This process (Fig. 12), which is sometimes also known as "anonymization," is central to enabling research on student information, while respecting student privacy. De-identification is a challenge, however, because it has been shown that nominally de-identified datasets may often be re-identified by drawing correlations between released data and data from public sources and/or previously released data sets. Noteworthy recent examples of deficiencies in de-identification include the re-identification of Governor William Weld from anonymized health information[17], and the re-identification of 21% of the participants in a database of the Public Genome Project. These two examples, which resulted from work by Prof. Latanya Sweeney of Harvard University (who did her Ph.D. work with Prof. Abelson at MIT), were possible because the "de-identified" data sets gave zip codes and approximate birth dates for individuals, which together with public records enabled re-identification of the data.

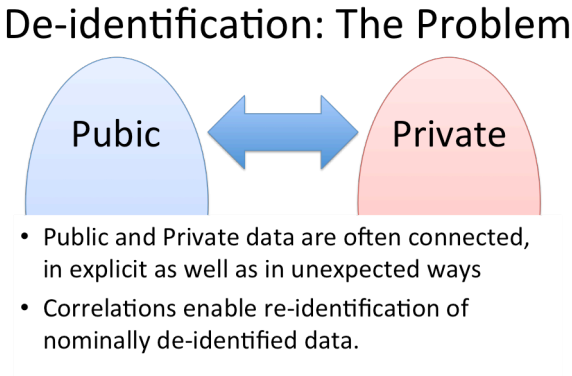

Figure 12: The challenge of de-identification.

[15] There is an ongoing discussion as to whether logging all data accesses is feasible.

[16] See Prof. John Guttag's presentation at the MIT Big Data Privacy Workshop, co-hosted with the White House Office of Science and Technology Policy, March 3, 2014.

[17] See account of Sweeney's work, chronicled in Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," August 13, 2009 (SSRN #1450006).

This risk of allowing re-identification of nominally "de-identified" data is exacerbated by data connecting student activity to publicly accessible information. This is a fundamental issue for MITx courses on edX, which have public fora. Student "click-stream" data that align with forum usage are thus "quasi-identifiers," to re-identify students.

Fundamentally, we believe there is no failsafe automatic de-identification procedure, though there are a number of widely employed approaches to address the de-identification challenge (Fig. 13). Traditional approaches rely on two methods: (a) generalizing/"fuzzing" data to quantitatively increase the level of de-identification of solo datasets, and (b) selective removal of "quasi-identifying" information from data records. The extent to which these methods are successful is then judged by expert determination, or by compliance with legally mandated levels of data redaction.

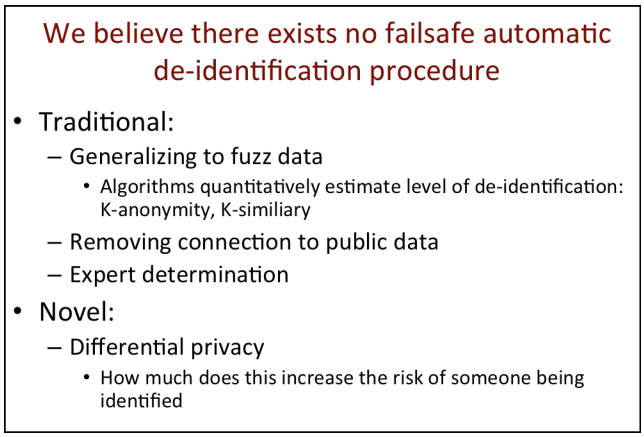

Figure 13: Approaches to de-identification

Novel approaches to this issue are arising; one of the most interesting is the notion of "differential privacy," which, strictly speaking, does not produce *per se* de-identified data sets. Instead, differential privacy provides an automated mechanism by which users query source data and are presented with query results that maintain a certain level of privacy for individuals within the data set. This usage scenario is quite different from the traditional approach, because no data set is actually released. Also, the fundamental ideas behind differential privacy rest on provable mathematical theorems, given certain reasonable assumptions. While this approach remains a very active and promising area of research, few differential privacy systems have actually been implemented, to date.

III.2b. Best-practices in de-identification

The Health Insurance Portability and Accountability Act (HIPAA) privacy rule gives very specific guidance for de-identification of medical records, providing an informative example of one approach. HIPAA specifies two approaches to de-identification (Fig. 14), known as "Safe Harbor" and "Expert Determination."

We note here that of course student records are not subject to HIPAA requirements, and for that matter, the access, use and disclosure restrictions that FERPA imposes on student information are less significant, as a legal matter, than those in HIPAA.

We consider the HIPAA regulations here because they describe two government-endorsed pathways to establishing de-identified data sets.

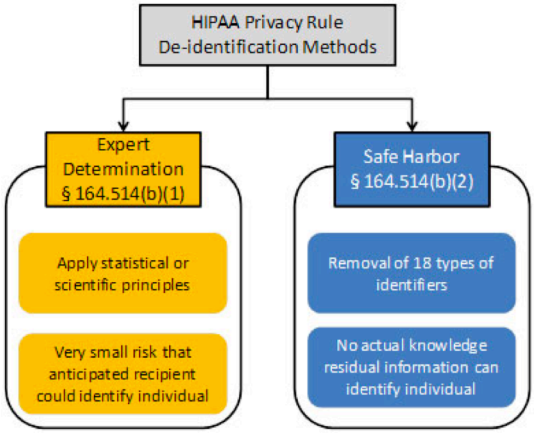

Figure 14: HIPPA de-identification methods.

The **Safe Harbor** approach specifies that a data set may be sufficiently "de-identified," for purposes of HIPAA, by removing 18 types of identifiers, and from there generalizing/"fuzzing" certain of the data that remain. The HIPAA safe harbor approach would require removal of the following identifiers of the individual or his or her relatives, employers, or household members: names; all geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code; all elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older; telephone numbers; vehicle identifiers and serial numbers, including license plate numbers; fax numbers; device identifiers and serial numbers; email addresses; web Universal Resource Locators; social security number; Internet Protocol addresses; medical record numbers; biometric identifiers; health plan beneficiary numbers; full-face photographs and any comparable images; account numbers; any other unique identifying number, characteristic, or code.

HIPAA de-identification by **Expert Determination**, by contrast, stipulates that health information is not individually identifiable only if "a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable: (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and (ii) Documents the methods and results of the analysis that justify such determination."

By contrast, FERPA provides no legal standard for the de-identification of education records. In fact, the regulations are remarkably vague, stipulating simply that education records are de-identified if they "[do]not allow a reasonable person in the school community to identify the student with reasonable certainty." The Department of Education has cautioned, in a non-binding guidance document accompanying recent regulations, that institutions should take care not to release

iterations of data sets that, taken together, could enable the data to be re-identified. The Privacy Technical Assistance Center of the U.S. Department of Education publishes a number of reference reports about approaches to de-identification and compliance with FERPA. These include a case study on a state educational agency that suggests redacting, blurring (to a level that is apparently synonymous with 5-anonymity), and applying "a statistical algorithm . . . to swap data elements for a small number of individuals" for additional disclosure avoidance. This is consistent with what we have heard from James Waldo, Professor of the Practice of Computer Science and Chief Technical Officer, Harvard University, and his team studying de-identification practices, in collaboration with researchers from the Harvard Law School.

III.2c. Recommendations for MIT

MIT does not currently define standards or procedures for de-identification of student information. We recommend that MIT continue with this practice, and that MIT not adopt any specific methodology akin to HIPAA's Safe Harbor, or a checklist of tests, because of the inherent impossibility of failsafe automatic de-identification and the manifest fragility of adopting any rigid list of information to be redacted.

Instead of endorsing a specific de-identification methodology, we recommend two avenues to address the needs of opening up access for researchers to student information, while maintaining privacy values:

- Focus on expert determination. The Institutional Research section of the Office of the Provost provides analytical and research support to the Provost, academic departments, research laboratories, and centers. This IR team has traditionally provided expertise on de-identification, *e.g.*, for anonymized data requests drawn from the Registrar's student records. This institutional expertise in de-identification should be maintained and strengthened, as an impartial focal point for generating and certifying de-identified datasets from MITx and other online educational activities.

- Invest in cutting-edge advances such as differential privacy. President Reif has made it clear that MITx data are to be constructed and curated as a public trust. Enabling this will require effective research access to the datasets, while respecting privacy needs. New avenues of enabling such research, which go beyond de-identification, *e.g.*, by providing the ability to execute queries on the unadulterated data, while mathematically guaranteeing a specified level of individual privacy, are promising and deserve investment.

We note as well that FERPA permits disclosure of *identified* student information under certain conditions.[18]

[18] FERPA permits disclosure of identified student information to persons within the institution (MIT) who have a legitimate educational interest in the information, and to researchers within and without

IV. Best Practices Regarding Online Learner Data

The Committee gathered information[19] from other peer institutions on a number of issues related to online learning in order to identify best practices and policies. We queried other institutions using the Association of American Universities (AAU) data exchange (AAUDE) listserv, the IVY+ IT Auditors listserv, and personal contacts. The group also visited institutional websites to gather public policies.

Our findings are summarized as follows:

- All of the universities treat data from online learners similarly to other student data, *i.e.*, subject to FERPA, when the online learner is taking the online course for credit.
- However, there is substantial variation in opinion and policies when the on-line learners are not earning credit.
- Online providers, such as Coursera, often provide an option in a long agreement by which the student can click accept/agree, which then allows the data to be used for research studies.
- All of the institutions agree that Institutional Review Board (IRB) approval is required by educational researchers in order to use the data from online learners.
- As to which office on campus has the responsibility for fulfilling requests from education researchers, answers vary, but the most common were the registrar, the counterpart offices to MIT's Office of Digital Learning, and the school's office of institutional research.
- Many of the institutions we consulted stated that their policies were still in development.
- Some of the universities have policies and guidelines related to the distribution of data for research, specifically, but many said policies and guidelines are still in development.

In reviewing best practices in the handling of online learners' data, we did not see any one practice that we should consider for adoption at MIT. Privacy policies, beyond what currently exists at schools that have active MOOC endeavors, appear to

---

MIT whose work is directed at (a) improving instruction; (b) developing, validating, or administering predictive tests; or (c) administering student aid programs.

[19] See also the following:

- Appendix A: Information from University of North Carolina's Business School, which has a full online MBA program.
- Appendix B: Listing of websites for some of the universities.
- Appendix C: Responses from the AAU Data Exchange Query
- Appendix D: Responses from the IVY+ IT Auditors query

be at a formulation stage. The current state of affairs elsewhere suggests that MIT is in a position to provide leadership to the academic community on issues regarding access to and privacy of online student data.

IV.A Recommendation: Information Security Policy

Several of the schools that we reviewed have Information Security Policies that go beyond what currently exists at MIT. MIT's Student Information Policy, discussed above, prescribes who should or should not have access to personally identifiable information and under what circumstances. The Institute has also adopted a Security Policy (P&P 13.2) that describes the policy on use of information technology resources.

However, an effective **Information Security** policy prescribes technological and procedural safeguards that promote and ensure compliance with the privacy policy. At such a dynamic time, it can be a challenge to implement sound, consistent data management and stewardship. Creating an Information Security policy and a data governance framework is a large and complicated task that extends well beyond the charge of the Ad Hoc Committee on Privacy of Student Records. But it is an important one, given the many operational, legal, and privacy risks entailed in the Institute's storage and management of data.

We therefore recommend that the ITGC (Information Technology Governance Committee) take up the review and development of data governance programs with respect to data access, storage and security. A starting point, with regard to the specific context of online learner data, is proposed in Section VI.

V. Proposed Information Security Policy Elements

The specific context of the charge to this Committee relates to online learner data, from MITx and other online educational systems employed by MIT. Within this context, the Committee proposes that MIT establish and adopt an Information Security Policy governing the storage and transmission of online learner data at various points in the lifecycle of the data. This policy should include the following elements:

- When the data have been prepped for release by the Trustee (see section VI) and are at rest on the hosting server:
    - If the data are hosted in the MIT data center, then they do not need to be encrypted.
    - Data should be encrypted if hosted on a cloud-based service. This will be done by encrypting the raw data before uploading. Encryption

ensures the data are not useful in the event of a security breach. It should also be clear in the agreement with the cloud provider that the data are retrievable in the event the contract is terminated for any reason.

- When the data are being accessed and retrieved from the hosting server.
    - When a researcher has been approved to access a data set, an account will be created for him/her if one doesn't exist.
    - The researcher account will be added to an Access Control List (ACL) for that data set.
    - Authenticated web-based access will be provided.
    - All access and any changes to data made will be audited
- When the data are at rest on a researcher's device. In addition to Usage and Disclosure restrictions (see Section VI of this report) the researcher agrees to abide by the following security requirements:
    - Researcher will encrypt the data on any device he or she is using to store data.
    - Researcher will require authentication to access the device storing the data.
    - Device will be regularly monitored to ensure it is virus-free.
    - Data will be deleted from all devices when period of use is complete.
    - The researcher commits to informing MIT of any potential security breach involving a local device hosting online learner data.

These elements represent considerations relevant to online learner data; they do not form a complete information security policy. The Committee recommends that a full information security policy be determined and established, for example, via the ITGC.

VI. Learner Data Access Policy and Procedures

A central charge to this Committee is to consider appropriate policy and procedures for storing online learner data and administering access to and distribution of data sets in furtherance of research and other legitimate uses of the data. On these questions, we recommend the following policy and procedures for adoption by MIT.

Note at the outset that we intend the draft policy and procedures below to apply to *all* online learner data, whether or not the learner is an enrolled MIT student. In cases where an MIT student may be engaging in online learning as part of his or her studies for MIT credit, the resulting learner data is subject to the further protection of the Student Information Policy. In most every respect the policy and procedures specified below are consistent with the SIP (and with record retention requirements imposed on federally funded research[20]). There is one exception, which is that under the SIP, the Chancellor must approve release of student information for use in research. We believe we can meet this requirement, and therefore treat all online learner data the same (whether or not the learner is an MIT student pursuing credit), if the Chancellor grants a general "blanket" approval for release of student information consisting of online learner data, pursuant to the policy and procedures below.

Learner Data Access Policy and Procedures (Draft)

This Committee strongly recommends that to the extent possible, student data from online learning initiatives (*e.g.*, MITx) be constructed and curated as a public trust, and that MIT adopt these specific policies and actions for data arising from MIT online learning initiatives:

1. The data generated in connection with MIT online learning initiatives, including MITx data, shall be treated no differently than any other organizational data that contains information about unique individuals. Thus,

[20] Federally funded researchers accessing MIT's online learning data may be subject to record retention requirements under generally applicable federal conditions on funded research (typically three years from closeout of the grant). The policy and procedures below, which require researchers to return or destroy online learner data when they no longer need them, do not conflict with the retention obligation — MIT can permit researchers to retain the data for an extended period, so long as the data are returned or deleted at the close of that period. Likewise, we do not see that the restrictions we would impose on researchers' downstream distribution of the data will conflict with the open data access policies under development by federal agencies, pursuant to the Obama Administration's open research data initiative. Those policies relate to data that result from research, rather than data that investigators receive for research use, and we expect they will make allowances for sensitive, private, or confidential data under study.

the data should be governed in a manner consistent with existing systems for protecting the privacy of students and human subjects of research.

2. The governance of the data shall be managed within existing principles of academic faculty governance by delegation to a Learner Data Trustee (the Director of Institutional Research in the Office of the Provost. "Trustee" hereafter) with advisory and periodic oversight[21] by a specifically designated faculty committee (the "Committee") with voting members from 5 schools at MIT and one representative each from the Office of Digital Learning and the Office of the General Counsel attending and advising.

3. As used in this Policy Statement,

    a. "Learner Data" means any information that relates to a specific Learner and is acquired by MIT or generated in connection with the Learner's participation in an MIT Online Learning Initiative.

    b. "De-identified Data Set" means a data set that includes Learner Data and is established through procedures adopted by MIT for de-identification (see Section III of this report). A data set that includes any information relating to MITx forum posts, whether it be metadata or the content of the posts, is not a De-identified Data Set, due to the presence of quasi-identifying information which may lead to re-identification.

    c. "Identified Data Set" includes any data set that includes Learner Data and is not a De-identified Data Set.

    d. "Learner" means any person, whether or not a student enrolled in an MIT degree program, who participates in an MIT Online Learning Initiative.

    e. "MIT Online Learning Initiative" or "Initiative" means an MITx course or any other interactive online educational programming offered by MIT and designated as such an Initiative by the Office of Digital Learning.

    f. "Instructor" means the person or persons designated by the sponsoring MIT department or program responsible for content and/or management of a particular Initiative.

[21] There is ongoing discussion in the Committee about how to phrase this such that it is clear the oversight is not a managerial role; in addition, it should be clear that "within existing principles of academic faculty governance" implies a chain of reporting, which need not be made more explicit here.

4. Categories of data sets available from MITx for instructional and other educational and research purposes may include:

    a. Course Reports (summary data on specific subjects taught);

    b. Certain De-identified Data Sets periodically made available on open access online, as per #6 below;

    c. All other De-identified Data Sets, with notice to Instructors; and

    d. Identified Data Sets (with or without forum data), with notice to Instructors.

5. MIT departments or programs sponsoring MIT Online Learning Initiatives shall have access to all Learner Data from those Initiatives, immediately and in perpetuity, for analysis for improvement of department curriculum and pedagogy, without need to apply through procedures described below. Instructors will have access immediately and in perpetuity to the Learner Data from their Initiatives, which they may use for the purposes specified above. The Director of Institutional Research will supply guidelines to departments and Instructors regarding data management and security. In the event an Instructor seeks to use Learner Data from his or her own Initiative for research purposes, he or she shall submit an application to the Trustee, pursuant to bullet 8, below, for approval of such research use[22].

6. The Committee constituted to advise and oversee Trustee's responsible management of Learner Data will ensure that no subject-specific or subject-identifying Learner Data will be posted for open access online without first notifying[23] the applicable sponsoring department or program.

7. No Learner Data will be released to any researcher, following application to the Director of Institutional Research, without first notifying the subject Instructor and the sponsoring department or program. Learner Data can be released only after six months from the close of the course, during which time the Instructors and/or their departments or programs shall have exclusive use of the Learner Data, except in special circumstances determined by the Trustee. This period of Instructor exclusivity shall not limit the access entitlements to Learner Data described above in Paragraph 5.

8. Researchers seeking access to Learner Data for research to improve teaching and curriculum or contribute to scholarship on teaching and learning shall

[22] See guidelines given by FERPA.

[23] It remains here to specify who makes this notification; presumably the Trustee would do so.

submit an application to the Director of Institutional Research. Researchers shall provide information on the application (see form, Appendix F) concerning the following requirements for access to MITx data[24]:

a. Researcher's name, institution, department, address, contact information.

b. Source of funding for research.

c. Documentation of approval by department head or equivalent.

d. Indication whether the proposal has undergone peer review and by whom.

e. List of data sets requested, including whether the requested sets are De-Identified Data Sets or Identified Data Sets, and all collaborating researchers who will need access to the Learner Data (together, the "Study Group").

f. In the case of MIT researchers, documentation of approval or exemption by the MIT Committee on the Use of Humans as Experimental Subjects ("COUHES").

g. In the case of non-MIT researchers, (1) documentation of approval or exemption by the institutional review board for human subjects research ("IRB") at the researcher's home institution; and (2) documentation of approval or exemption by COUHES or of its election to rely on determinations made by the home institution's IRB.

h. Documentation or certification of completion of the CITI human subjects training program by all members of the Study Group. This requirement shall apply whether or not COUHES or a non-MIT researcher's IRB has determined that the study is exempt from review or poses "minimal risk" to subjects.

i. A research proposal of 2–5 pages shall be included with the application. This proposal is normally included in the application for IRB and COUHES approval, required in (f.) and (g.) above. The Trustee shall consider the substance of the research proposal only as necessary to reach a determination that disclosure of the Learner Data to the research is consistent with applicable law.

[24] These requests for information apply both to researchers from research institutions, as well as to researchers from other institutional frameworks.

Such proposals should include the following information:

What is your research question? How do you plan to answer it with data from MITx? What forms of data are requested and why this form rather than another (*e.g.*, de-identified data, identified data)? What methods of analysis will you use that will exploit characteristics of this data? What theory, device, or simulation will be amended or developed, and how will this analysis contribute? What form will the results take (*e.g.*, course project; thesis; peer reviewed publication; TLO)?

Each MIT principal investigator using MITx data will sign or otherwise assent to Terms of Use ("TOU") specifying that the researchers understand and agree to the conditions of access to and handling of Learner Data, and requiring the researchers to follow specified procedures for securing the anonymity of the subject participants and protecting access to and confidentiality of the data. A non-MIT investigator's home institution or organization will enter into a Data Use Agreement with MIT setting forth the conditions of access to and handling of Learner Data, and requiring the researchers to follow specified procedures for securing the anonymity of the subject participants and protecting access to and confidentiality of the data.

The Director of Institutional Research may charge a fee[25] for making data available.

9. MITx will provide a catalog of available Learner Data, specifying their format and codes.

10. Requests for Learner Data for purposes other than academic, scholarly research with expectations to publish in peer-reviewed journals or presses may be considered[26] on a case-by-case basis with reference to whether meeting the request is lawful and in the interest of the Institute.

11. TOU and DUAs issued under this policy shall include a provision by which any publication prepared using Identified Learner Data must be submitted to the Trustee for review at the same time it is submitted for any public availability, solely in order to ensure that the publication does not disclose Identified Learner Data.

[25] The Committee has discussed how to frame this fee, but in lieu of making a specific recommendation, it is felt that MIT should wait and see what costs are incurred, to see what expenses should be recovered.

[26] Some thought should be given to who decides, in these cases.

Appendices

Appendix A: Information from University of North Carolina's Business School

University of North Carolina Business School

(Discussion with: Susan E. Cates ▪ President, Executive Development ▪ Executive Director, MBA@UNC ▪ UNC Kenan-Flagler Business School)

UNC has an online MBA – a full degree-granting program. Online MBA student is granted the same MBA degree as an on campus MBA student. In this program, FERPA applies to the online data of the participants. They take FERPA very seriously and take a conservative approach. They treat the online student exactly the same as the on campus student with regard to FERPA. They work with independent company for online MBA. This independent company also applies FERPA to online student data (and even asked if perspective students visiting classes pose a problem). Ms. Cates was not sure if they would treat students taking classes not for credit would be treated any differently.

UNC's Business School also has an executive education program, mostly custom (UNC Business Essentials). There really has not been any discussion that she knows of about the data for these online learners and the application of FERPA.

To her knowledge, no data from the UNC Business School has been made available to educational researchers. They have discussed how educational researchers would probably find the data useful and interesting. To date, they have not released any data for this purpose to her knowledge. If this would occur, she would work closely with university counsel to make sure in line with FERPA. They maintain data in several places. They are careful with the data just like with student data generally.

The main difference she noted was the data the faculty have on an ongoing basis. They can see which students (identifiable) watch the videos, which do the pre-class assignments, etc. To her knowledge there are no policies about how to store and access that data. Again, they have discussed how researchers would probably find it useful but to date to their knowledge no researcher has asked for it or used it.

Appendix B: Listing of websites for some of the universities

Useful websites for from our survey of best practices include:

- http://security.harvard.edu/research-data-security-policy
  http://www.security.harvard.edu/
- http://vpol.stanford.edu/
- http://www.upenn.edu/almanac/volumes/v56/n25/confidentiality.html
- http://www.mais.umich.edu/access/download/data_stewards.pdf
- http://www.it.cornell.edu/security/data/types.cfm
- http://www.dfa.cornell.edu/dfa/treasurer/policyoffice/policies/volumes/informationtech/infosecurity .cfm

Appendix C: Responses from the AAU Data Exchange Query

The Association of American Universities Data Exchange (AAUDE) is a public service organization whose purpose is to improve the quality and usability of information about higher education. MIT is a member of AAUDE (through the Institutional Research section of the Office of the Provost). The Director of IR, Lydia Snover, coordinated a query to the AAUDE members, as part of this Committee's survey of best practices. Below is the query, and the responses.

Query to the AAUDE:

> Background and Description
>
> With the rapid growth in data from online courses such as those offered through MITx or edX, both internally at MIT and externally to learners elsewhere, and with the growing interest of researchers to use that data to better understand how teaching and learning occur, MIT is in the process of reviewing and possibly to clarifying our policies on access and use of such data. While the Committee on Student Information Policy has oversight of traditional student academic records, such as those maintained by the registrar, the growth of more detailed interactions of students through a subject and the use of MIT subject material by learners not registered at MIT raise additional issues concerning privacy expectations, legal constraints on access to data, and appropriate data curation. MIT has set up an *ad hoc* committee to recommend policies and procedures for curation, maintenance, and access to learner data acquired through online delivery of subject material.
>
> We are interested in best practices of peer institutions in dealing with data access, storage, and security.
>
> **Request/Questions**: Does your institution offer online courses and does your institution have data from these courses? If so,
>
> 1. Has your made a determination of whether FERPA governs data collected from on line learners?
>
> 2. Have you or do you plan to make any data available to educational researchers?
>
> 3. In what format do you or will you provide data to educational researchers?
>
>   a. Raw identifiable data
>
>   b. Identifiable aggregate data

c. De-identified data set

d. Other, please describe

5. Which office at your institution is responsible for the distribution of data to educational researchers (i.e., Registrar, IR, Libraries, Office dedicated to on line learning, etc.)?

6. Do you require IRB approval prior for researchers using identifiable data?

7. What guidelines, if any, do you provide to researchers about the use of this data?

8. Does your institution have established policies on access, storage, and security with regard to on line learning data? If so, would you share them?

Any documentation you might have on any of these issues will be very much appreciated.

Responses from AAUDE institutions:

Institution	Online courses	Data from online courses	1. FERPA?	2. Data available to educational researchers
University of Arizona	Yes. While open online courses (MOOC's) are not being monitored centrally at Arizona, we know of one course offered through Udemy and another instance where an associate professor of chemistry is developing a MOOC with a $50,000 grant from online giant Google. There may be others, but these are two examples.		The Registrar's office has not discussed this with the Office of the General Counsel. Our Registrar's take on this would be that MOOC students are not covered under FERPA. Our interpretation of when a student is officially covered by FERPA is when they are first enrolled. MOOC students do not exist in our official transactional systems and are not officially enrolled. If Arizona signs a contract with Udemy, then that would make us liable for student data.	
University of Rochester	Yes		Our associate vice president of online learning does not believe our compliance with FERPA is dependent on modality of instruction.	No specific plan at this time.
Ohio State University	Yes		Yes	Yes, to those in the institution
Rutgers	Yes. Ongoing without any final			
UW-Madison	Yes we have some data from these courses. Answers to these questions depend on whether we are talking about 1) regular courses delivered online to matriculated UW-Madison students or 2) "courses" delivered via a 3rd party platform (i.e. MOOC) to participants who are not enrolled for credit at UW-Madison.		Yes, FERPA covers matriculated students taking courses for credit at UW-Madison. It does not cover participants in a MOOC (Coursera platform)	Yes. Researchers interested in administrative data related to matriculated UW-Madison students work with the Registrar and IRB to determine whether their request can be met within the confines of FERPA and human subjects consent. While FERPA does not pertain to our MOOC participants, IRB and human subjects consent issues do. We anticipate dealing with requests from researchers on a case by case basis as they obtain IRB approval. We currently have one approved protocol and are working through these issues right now.

Institution	3. Format of data	4. Which office is responsible for the distribution of data to educational researchers	5. IRB approval	6. Guidelines for researchers
University of Arizona		Arizona does not have a centralized distributions point. However, Institutional Research supplies data and reports to administrators on non-MOOC online courses.	Yes	We require a standard FERPA agreement for those studies that come to the attention of the Registrar's Office.
University of Rochester	To be determined as appropriate.	No specific plan at this time.	IRB continues to provide oversight for research as appropriate.	Default to IRB, but everyone must follow IT Policy: http://www.rochester.edu/it/policy/assets/pdf/INFORMATION_TECHNOLOGY_POLICY.pdf
Ohio State University	Raw identifiable data Identifiable aggregate data Most typically De-identified data set	Office dedicated to online learning is the data steward after researcher gains any necessary approvals	Yes	Office of Responsible Research Practices information, Institution research guidelines
Rutgers	Likely provide all four			
UW-Madison	a. Raw identifiable data. Yes, if consent standards are met through the IRB process. Note, our MOOC data is raising issues that are new to IRBs. For example, some MOOC participants are from countries whose citizens do not have the right to give informed consent. Another example is how to get consent from parents of participants under age 18 where there is to mechanism to contact them and/or verify authenticity of communication. b. Identifiable aggregate data. Yes, if consent standards are met through the IRB process. c. De-identified data set. Yes.	The Registrar is the custodian of student records. Requests that deal with data related to matriculated students are coordinated by the Registrar. Occasionally, we assist with these requests if IR has expertise with data requested that is not typically under the purview of the Registrar. For example, a research request that deals with enrollments records as well as financial aid attributes would generally involve IR as well. For MOOC data, the Coursera platform requires each participating institution to have a named Data Coordinator. Currently, I am fulfilling this role. The Data Coordinator is responsible for all data requests from the institution to Coursera and is responsible for providing data to researchers in a legal manner. Although your questions are mainly related to data security and IRB issues, we also have ongoing questions related to the workload associated with fulfilling data requests for research purposes.	Yes, federal law requires this.	Requests for data on enrolled students fulfilled by the Registrar are provided though a data use agreement that specifies requirements for data use, security, etc. I anticipate doing the same thing with MOOC data but have not yet gotten far enough along to be able to provide anything.

Institution	7. Established policies on access, storage and security	Confidential
University of Arizona	No	No
University of Rochester	The University of Rochester has the IT Policy which includes Data Classification and Access Restrictions (PDF link above, in #6) and a Record and Retention Policy: http://www.rochester.edu/adminfinance/records.html	No
Ohio State University	Institution data security guidelines	Yes
Rutgers		No
UW-Madison	Our policies are not specific to on line learning. We would use a similar data use agreement with on-line learning data requests that we would for other research requests.	No

Institution	Online courses	Data from online courses	1. FERPA?	2. Data available to educational researchers
University of Kansas	Yes The University of Kansas does offer online courses and we do have data.		Yes. We have determined that FERPA applies	Yes. We do so on a case by case basis utilizing the same philosophy and practices that apply to our campus-based learners.
University of Iowa	Yes	Yes	YES, any registered student at the University of Iowa is governed under FERPA	Except for rare cases it would only be available for departmental/collegiate program assessment. Not student or external researchers.
University of Michigan	Yes	Yes	Yes, FERPA, as well as other regulations, are being taken into consideration	yes
Stanford University	Yes	Yes	We do not consider participants in our MOOCs who are not enrolled Stanford students to be covered by FERPA.	Yes

Institution	Online courses	Data from online courses	1. FERPA?	2. Data available to educational researchers
University of Kansas	Yes The University of Kansas does offer online courses and we do have data.		Yes. We have determined that FERPA applies	Yes. We do so on a case by case basis utilizing the same philosophy and practices that apply to our campus-based learners.
University of Iowa	Yes	Yes	YES, any registered student at the University of Iowa is governed under FERPA	Except for rare cases it would only be available for departmental/collegiate program assessment. Not student or external researchers.
University of Michigan	Yes	Yes	Yes, FERPA, as well as other regulations, are being taken into consideration	yes
Stanford University	Yes	Yes	We do not consider participants in our MOOCs who are not enrolled Stanford students to be covered by FERPA.	Yes

Institution	7. Established policies on access, storage and security	Confidential
University of Kansas	We do not have established policies on this specific to online learning data	no
University of Iowa	They are treated the same as a regular face-to-face course	no
University of Michigan	We are currently drafting our policy on access to the data; however, we do have policies on security see: IT Policy and Security Resources • U-M IT Policies - http://cio.umich.edu/policy/ • U-M Safe Computing Sensitive Data Policies and Regulatory Compliance - http://safecomputing.umich.edu/protect-um-data/laws.php • U-M Safe Computing Protect University Data - http://safecomputing.umich.edu/protect-um-data/ • U-M Sensitive Data Guide to IT Services - http://safecomputing.umich.edu/dataguide/	no
Stanford University	Yes, we have a draft for such a document: http://infolab.stanford.edu/~paepcke/VPOL/analyticsSecurity.html. We also coordinated our anonymization techniques with our IRB. Applications for data are submitted online, and reviewed by at least two members of a data governance committee. The application is work in progress; the current version is at http://vpol.stanford.edu/research.	No

Institution	Online courses	Data from online courses	1. FERPA?	2. Data available to educational researchers
University of California Berkeley	Yes, BerkeleyX (MOOCs) uses the Edx platform; all other on-line courses use Canvas or Angel platforms.		We have determined that FERPA applies to data held by UC Berkeley that is collected from any participant in a UC Berkeley academic program, including online courses	Yes
MIT	yes	yes	Yes	Yes
University of Pittsburgh	Our institution offers for-credit online courses to matriculated students as well as a handful of open courses (MOOCs) to non-matriculated students. Re: MOOCs, we have partnered with Coursera and will follow their data sharing procedures.		FERPA applies to our matriculated online students but its applicability has not been determined for MOOC students.	Decisions on data availability will be made through an IRB process.
University of Maryland	Yes	Yes	We believe that FERPA does not apply to students taking MOOCs. Presumably it does apply to regular matriculated students that are taking courses online.	We do not have a different and specific plan for regular matriculated students that take online courses. For MOOCs, we make data available on request to instructors of their own courses just as they would have access to any other student data for their own courses. For other instructor's (UMD) courses, we give access if they have explicit permission from those other instructors and if they have IRB approval

Institution	3. Format of data	4. Which office is responsible for the distribution of data to educational researchers	5. IRB approval	6. Guidelines for researchers
University of California Berkeley	a. Raw identifiable data b. Identifiable aggregate data c. De-identified data set	Registrar, although online course data not yet finalized.	yes	Under development
MIT	a. Raw identifiable data b. Identifiable aggregate data c. De-identified data set	undecided	Yes	Under development
University of Pittsburgh	For MOOCs, that decision would be made on a case-by-case basis using Coursera's policies as guidelines and individual IRB recommendations. For for-credit, online courses taken by matriculated students, the format in which the data are made available would be guided by the individual research design and subsequent IRB approval.	For MOOCs, that has not been decided. For for-credit, online courses taken by matriculated students, it would depend on the type of data required by the individual research design.	Yes	For MOOCs, we have not yet established any guidelines beyond what Coursera provides. For for-credit, online courses taken by matriculated students, guidelines are provided by our institution's data management and FERPA policies.
University of Maryland	For MOOCs, we distribute data using Coursera's format which offers the data in 3 subsets, each with their own encoded student IDs. Then there is a "linking" database that ties those encoded student IDs to the student's specified name and email address.			

Institution	7. Established policies on access, storage and security	Confidential
University of California Berkeley	No, under development, but happy to share once finalized	No
MIT	under development	No
University of Pittsburgh	For MOOCs, we have not yet established policies. For for-credit, online courses taken by matriculated students, the same policies that apply to any personal data that exist on University computing systems also apply to data that exist on the University's learning management system.	No
University of Maryland		

Appendix D: Responses from the IVY+ IT Auditors query

Institution	**School Privacy Officer**	**Institute-wide PO or Dedicated to Student Data**	**Privacy Policies Specific to Student Data**
Univ. of Rochester	Chief Security Officer	Registrar manages record privacy	FERPA
			https://www.esm.rochester.edu/registrar/?id=02.07.01
Harvard	No	NA	FERPA
			http://security.harvard.edu/book/31-student-information-and-ferpa-overview
Univ. of Penn.	Yes.	Institute-wide	Based on FERPA
			http://www.upenn.edu/almanac/volumes/v56/n25/confidentiality.html

	Student Data Protection Standards for:		
Institution	**Granting and Removing Access to Student Data**	**Storage of Student Data**	**Transmission of Student Data**
Univ. of Rochester	These would be student system specific and based on departmental policies. General need to know security standards. (Access to data, however, is not open and must be requested and vetted.)	Policy on Retention of University Records - Electronic Records	Unsure.
Harvard	Yes	Yes	transmission of records must be encrypted
	http://security.harvard.edu/book/27-limit-user-access-confidential-information	http://security.harvard.edu/book/28-confidential-information-harvard-computing-devices	
Univ. of Penn.	DO not have any specific policies for student data protection standards, we define data as confidential, sensitive, or public and the computer security policy indicates the standards that must be used.	We are working on a data-centric policy, again by the type of data and further precautions that will be required	
		http://www.net.isc.upenn.edu/policy/approved/20100308-computersecurity.html	

Institution	**Change Handling of Student Data obtained through Online Learning**
Univ. of Rochester	Very, very limited on line learning.
Harvard	NA
Univ. of Penn.	No policies on student data obtained through online learning.

Institution	**School Privacy Officer**	**Institute-wide PO or Dedicated to Student Data**	**Privacy Policies Specific to Student Data**
Northwestern	NA	NA	Yes, based on FERPA
			http://www.registrar.northwestern.edu/academic_records/FERPA_policy.html

	Student Data Protection Standards for:		
Institution	**Granting and Removing Access to Student Data**	**Storage of Student Data**	**Transmission of Student Data**
Northwestern	Access to the Student Enterprise System http://ses.northwestern.edu/access.htm	Nothing specific to student data	Nothing specific to student data
	http://ses.northwestern.edu/access.htm		

Institution	**Change Handling of Student Data obtained through Online Learning**
Northwestern	None

Appendix E: Application of FERPA – Legal Analysis

A threshold question for this Committee is one of scope: to what data should the Committee's policy proposals apply? The Chancellor's charge directs the Committee to examine the particular privacy considerations surfaced by online learning initiatives. MIT is subject to privacy laws that confer particular obligations under MIT with respect to certain data — principally, the federal FERPA law. Accordingly, the Committee should scope its definition of data to include, at a minimum, all data as to which MIT has legal compliance obligations.

FERPA

The Family Educational Rights and Privacy Act ("FERPA") , 20 U.S.C. § 1232g, and the supporting regulations of the Department Education's Family Policy Compliance Office, 34 C.F.R. pt. 99, set forth limited access and privacy rights for students in their "education records." The law defines "education record" extremely broadly, to include essentially every scrap of information that a school that receives federal funds maintains about a student. Written in 1974, the law clearly never considered the possibility that institutions would acquire and maintain records on the scale they do now, and even the most recent updates to the regulations do not take account of the emergence of MOOCs.

We understand that the Department of Education is considering amending its regulations to take account of MOOCs and other recent developments of significance in higher education. Until it does, we remain subject to the extremely broad definitions of the current regulatory scheme, by which FERPA's requirements apply with equal force to digital data about MITx learners (MITx Learner Data). Although not every edX consortium partner university agrees, we do not see any way to read the current regulation to avoid this result.

MIT's basis for reaching this conclusion, according to the Office of General Counsel, is set forth below.

FERPA's requirements apply to an "educational agency or institution" [27] in its management of "education records."

EDUCATIONAL INSTITUTION

FERPA regulations define "educational institution" to mean "an educational agency or institution to which funds have been made available under any program administered by the Secretary [of Education], if [t] he educational institution

[27] Underlined terms in this appendix are defined terms in the FERPA regulations.

provides educational services or instruction, or both, to students." 34 C.F.R. § 99.1(a)(1). Funds are "made available by the Secretary" to an institution if they (1) "[a]re provided to the agency or institution by grant, cooperative agreement, contract, subgrant, or subcontract; or (2) [a]re provided to students attending the agency or institution and the funds may be paid to the agency or institution by those students for educational purposes, such as under the Pell Grant Program and the Guaranteed Student Loan Program." *Id.* § 99.1(c). By virtue of its participation in Title IV federal financial aid programs and its receipt of federal research grants, MIT is an educational institution subject to FERPA. The regulations go on to establish that a college or university must comply with FERPA in all its programs — even those that do not themselves receive federal funding. *Id.* § 99.1(d) ("If an educational . . . institution receives funds under one or more of the programs covered by this section, the regulations in this part apply to the recipient as a whole, including each of its components (such as a department within a university).").

EDUCATION RECORDS

The FERPA regulations define "education records" to mean "those records that are (1) [d]irectly related to a student; and (2) [m]aintained by an educational . . . institution or by a party acting for the . . . institution." *Id.* § 99.3. "Student" means "any individual who is or has been in attendance at an . . . institution and regarding whom the . . . institution maintains education records." *Id.* The regulations do not define the limits of the term "attendance," but they do make clear that the term includes "[a]ttendance in person or by paper correspondence, videoconference, satellite, Internet, or other electronic information and telecommunications technologies for students who are not physically present in the classroom." *Id.*

Because MITx courses are officially sanctioned MIT educational programming, participation in those courses must constitute "attendance" at MIT under the FERPA regulations, which do not admit any exceptions for programming that, say, is not offered in connection with a certificate or degree program, is offered free of charge, or does not confer academic credit upon completion. MITx Learners are therefore "students" of MITx for FERPA purposes. MITx Learner Data consist of records that relate directly to these students and are maintained either by edX on MIT's behalf or, after their transfer to the Office of Digital Learning, by MIT itself. MITx Learner Data held by edX or MIT are therefore "education records" of MIT.

The Committee has raised the question whether the scope of its charge should include all other online learning initiatives conducted at MIT or by members of the MIT community. The extent to which FERPA would apply to student data generated in connection with these initiatives depends on whether the online learner can be said to be "in attendance at" MIT by virtue of his or her participation in the

program.[28] A fair way to mark the limits of FERPA's applicability to online learning initiatives, absent any real guidance from the government, might be to ask whether the online programming is really "offered" by MIT. Indicia of "offering" might include (1) use of the MIT name in connection with the offering (other than simply to identify the credentials of an instructor); (2) institutional backing and support for the enterprise; (3) whether the initiative, if hosted or enabled by a third party, is so hosted or enabled under a contract between the third party and MIT; (4) whether MIT's Office of Digital Learning is administering the program; (5) whether institutional officials other than the faculty member certify completion of the course; and (6) whether faculty are providing the programming within the scope of their employment at MIT or as an outside professional activity.

[28] To be clear, if the online learner is otherwise "in attendance at" MIT — for example, because he or she is also enrolled in an MIT degree program or has been an MITx Learner, then records generated in connection with the online learning initiative are "education records" of a "student," subject to protection whether or not participation in the particular initiative establishes "attendance" at MIT. That fact does not counsel in favor of searching the records of every online learning initiative not otherwise covered by FERPA for names of MIT students and MITx Learners. There comes a point at which it is entirely appropriate to overlay a reasonableness filter on one's FERPA compliance. But we make note of it as an illustration of the law's extreme breadth.

Appendix F: Sample Data Request Application Form

MITx Data Requests

Managed by Institutional Research, Office of the Provost at MIT.

Project name *
Please provide a brief title for your project. If you have a COUHES application, please provide the application title.

Project Description *
Please describe your data request in terms of research intent, data requested, and plans for dissemination of results.

Point of Contact *
Should be person responsible for receiving data transfer.

email for Point of Contact *

Affiliation (University/Company) *

Course ID *

Do you currently have an MOU (or DUA)?

Do you currently have COUHES approval?

If you answered yes to the COUHES or MOU question, please upload ALL relevant documents.
Choose File No file chosen

All Involved Researchers
Please put each researcher on a new line

When do you expect to need the requested data?

Send me a copy of my responses

Email address