The HarvardX-MITx Person-Course Dataset AY2013

Document Date: May 27, 2014

This document accompanies the release of de-identified data from the first year (Academic Year 2013: Fall 2012, Spring 2013, and Summer 2013) of MITx and HarvardX courses on the edX platform. These data are aggregate records, and each record represents one individual's activity in one edX course. For more information about the existing analyses of these data and the first year of HarvardX and MITx courses, please see the HarvardX and MITx working paper "*HarvardX and MITx: The first year of open online courses"* by Andrew Ho, Justin Reich, Sergiy Nesterko, Daniel Seaton, Tommy Mullaney, Jim Waldo, and Isaac Chuang (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263).

The first release of this dataset is the HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0, created on May 14, 2014.

File name: "HMXPC13_DI_v2_5-14-14.csv"

The md5sum for this release (HMXPC13_DI_v2_5-14-14.csv) is:
2b09c674af772d45dae429045cf7acfc

Overview

This dataset is at the level of one row per-person, per-course. So, for example, if one individual enrolled in three MITx or HarvardX courses during the period covered by the dataset (for this release, Fall 2012, Spring 2013, and Summer 2013), that person would have three rows associated with their userID.

The dataset includes both administrative variables and variables generated from user-provided data. For a detailed description of the variables included in the de-identified dataset, please see the variable list below.

Both institutions are motivated to make as much data freely available as possible. The key constraint limiting the release of data is that both Harvard and MIT consider data generated by learners who register for MITx and HarvardX courses on edX to be governed by the Family Educational Rights and Privacy Act (FERPA) (10 USC. § 1232g; 34 CFR Part 99), which protects the privacy of student records. In order to strike a balance between these competing interests, this dataset has been de-identified, according to the procedure described in another accompanying document (titled "Person-Course De-identification Process").

The courses included in this dataset are as follows (more detail can be found in the paper linked to above):

Institution	Course Code	Short Title	Full Title	Semester
HarvardX	CB22x	HeroesX	The Ancient Greek Hero	Spring-Summer 2013
HarvardX	CS50x	-	Introduction to Computer Science I	Fall 2012 – Spring 2013
HarvardX	ER22x	JusticeX	Justice	Spring-Summer 2013
HarvardX	PH207x	HealthStat	Health in Numbers: Quantitative Methods in Clinical & Public Health Research	Fall 2012
HarvardX	PH278x	HealthEnv	Human Health and Global Environmental Change	Summer 2013
MITx	14.73x	Poverty	The Challenges of Global Poverty	Spring 2013
MITx	2.01x	Structures	Elements of Structures	Spring-Summer 2013
MITx	3.091x	SSChem	Introduction to Solid State Chemistry	Offered twice: Fall 2012 and Spring 2013
MITx	6.002x	Circuits	Circuits and Electronics	Offered twice: Fall 2012 and Spring 2013
MITx	6.00x	CS	Introduction to Computer Science and Programming	Offered twice: Fall 2012 and Spring 2013
MITx	7.00x	Biology	Introduction to Biology – The Secret of Life	Spring 2013
MITx	8.02x	E&M	Electricity and Magnetism	Spring 2013
MITx	8.MReV	MechRev	Mechanics Review	Summer 2013

Variable sources and definitions

Notes:

- "administrative" indicates that the variable comes from the edX system or has been computed by the research team;

- "user-provided" indicates that the variable comes from questions asked by edX of the student at the time of registration with edX;
- values of "NA" in user-provided columns indicate that the student created an edX account before the corresponding student registration question was available;
- blank values in user-provided columns indicate that although the user created their edX account after the question was asked in the student registration process, the user declined to provide the information;
- "_DI" at the end of a variable name indicates that this variable was transformed during the de-identification process.

course_id: administrative, string, identifies institution (HarvardX or MITx), course name, and semester, e.g. "HarvardX/CB22x/2013_Spring"

userid_DI: administrative, string, first portion identifies dataset (MHxPC13 corresponds to MITx HarvardX Person-Course AY13), second portion is a random ID number. Example ID: "MHxPC130442623".

registered: administrative, 0/1; registered for course, =1 for all records in person-course.

viewed: administrative, 0/1; anyone who accessed the 'Courseware' tab (the home of the videos, problem sets, and exams) within the edX platform for the course. Note that there exist course materials outside of the 'Courseware' tab, such as the Syllabus or the Discussion forums.

explored: administrative, 0/1; anyone who accessed at least half of the chapters in the courseware (chapters are the highest level on the "courseware" menu housing course content).

certified: administrative, 0/1; anyone who earned a certificate. Certificates are based on course grades, and depending on the course, the cutoff for a certificate varies from 50% - 80%.

final_cc_cname_DI: mix of administrative (computed from IP address) and user-provided (filled in from student address if available when IP was indeterminate); during de-identification, some country names were replaced with the corresponding continent/region name. Examples: "Other South Asia" or "Russian Federation".

LoE: user-provided, highest level of education completed. Possible values: "Less than Secondary," "Secondary," "Bachelor's," "Master's," and "Doctorate."

YoB: user-provided, year of birth. Example: "1980".

gender: user-provided. Possible values: m (male), f (female) and o (other).

grade: administrative, final grade in the course, ranges from 0 to 1. Example: "0.87".

start_time_DI: administrative, date of course registration. Example: "12/19/12".

last_event_DI: administrative, date of last interaction with course, blank if no interactions beyond registration. Example "11/17/13".

nevents: administrative, number of interactions with the course, recorded in the tracking logs; blank if no interactions beyond registration. Example: "502".

ndays_act: administrative, number of unique days student interacted with course. Example: "16".

nplay_video: administrative, number of play video events within the course. Example: "52".

nchapters: administrative, number of chapters (within the Courseware) with which the student interacted. Example: "12".

nforum_posts: administrative, number of posts to the Discussion Forum. Example: "8".

roles: administrative, identifies staff and instructors, but blank as staff and instructors were removed from this release.

inconsistent_flag: administrative, identifies records that are internally inconsistent. Due to a variety of data issues, including missing tracking logs and one course (CS50x) which has virtually no logs because most of the course content is hosted outside of the edX platform, a portion of the records have null values for *nevents* but have non-null values for *ndays_act*, *nforum_posts*, or *nchapters*. The source for *nevents* and for *last_event_DI* is the tracking logs, whereas *ndays_act, nforum_posts,* and *nchapters* come from a data source known as the "Courseware Student Module"[1]. Due to the two different sources, if something is wrong with the Tracking Logs[2] for a class or a student, then records in Person Course can be internally inconsistent and have a value of '1' in this column.

---

[1] More information available at http://edx.readthedocs.org/projects/devdata/en/latest/internal_data_formats/sql_schema.html#courseware-progress-data

[2] More information available at http://edx.readthedocs.org/projects/devdata/en/latest/internal_data_formats/tracking_logs.html