6.00.2x: Introduction to Computational Thinking and Data Science

MITx Spring 2014 Course Report

Report prepared by Daniel T. Seaton, MIT Institutional Research, Office of the Provost. Questions should be addressed to irx@mit.edu.

August 1, 2014

The MIT Institutional Research group in the Office of the Provost, in conjunction with the MIT Office of Digital Learning, have established reporting mechanisms to inform course staff and the general public about activity within MITx open online courses on edX (sometimes referred to as Massive Open Online Courses - MOOCs). These reports are decidedly data driven, with text limited to a brief introduction, course description, and figure captions. The founding principles of these reports can be found in the inaugural MITx and HarvardX course reports; either the 2012-2013 cross-course synthesis [1] or one of the individual course reports from this same period (e.g., [3]; complete list available here http://odl.mit.edu/mitx-working-papers/). Questions regarding these reports, and the tools used to create them, can be directed to the Institutional Research group at MIT (irx@mit.edu).
Description of 6.00.2x taken from [www.edx.org](http://www.edx.org):

An introduction to using computation to understand real-world phenomena.

**About this Course**

6.00.2x is aimed at students with some prior programming experience in Python and a rudimentary knowledge of computational complexity. We have chosen to focus on breadth rather than depth. The goal is to provide students with a brief introduction to many topics, so that they will have an idea of what’s possible when the time comes later in their career to think about how to use computation to accomplish some goal. That said, it is not a “computation appreciation” course. Students will spend a considerable amount of time writing programs to implement the concepts covered in the course. Topics covered include plotting, stochastic programs, probability and statistics, random walks, Monte Carlo simulations, modeling data, optimization problems, and clustering.

What will I learn? If you successfully complete this course, you will have:

Developed some insight into the process of moving from an ambiguous problem statement to a computational formulation of a method for solving the problem, Learned a useful set of algorithmic and problem reduction techniques, Learned how to use simulations to shed light on problems that don’t easily succumb to closed form solutions, Learned how to use computational tools, including simple statistical, machine learning, and plotting tools, to model and understand data. All required readings are available within the courseware, courtesy of The MIT Press.

**Course staff for 6.00.2x:**

Eric Grimson, John Guttag, Ana Bell

**Certified versus Non-Certified:**

Open online courses lead to tremendous diversity in terms of enrollment and activity. Hence, basic grouping of participants helps distinguish behavior relevant to course teams and analysts. For example, the 2012-2013 course reports from HarvardX and MITx used four categories: registered, viewed, explored, and certified. We acknowledge these categories in Fig. 2.1c but opt for a pragmatic categorization in the bulk of this report, namely, “Non-Certified” versus “Certified”. Certification provides a behavioral reference to those students likely to interact with a significant amount of content, and its application in this report is not intended to focus discussion around certification rates. As shown in previous course reports, behavior of “Non-Certified” participants varies greatly, and many open questions remain.

**Course Metadata and participant Activity Highlights relative to certification:**

The following tables provide course metadata and summary stats of participant behavior. The lefthand table represents course information, enrollment, and the number certified. The righthand table provides total (summed) and mean metrics for participant behavior categorized by certification. Many of these metrics depend greatly on time of enrollment, and one should consider them only as highly generalized features of the course.

<table>
<thead>
<tr>
<th>Course Metadata</th>
<th>Activity Highlights</th>
<th>Non-Certified</th>
<th>Certified</th>
</tr>
</thead>
<tbody>
<tr>
<td>School</td>
<td>Summed Number of Clicks</td>
<td>3592314</td>
<td>8015051</td>
</tr>
<tr>
<td>Launch Date</td>
<td>Mean Clicks per User</td>
<td>197.9</td>
<td>5336.3</td>
</tr>
<tr>
<td>Course Length</td>
<td>Summed Total Time</td>
<td>32566.8 hrs</td>
<td>83317.4 hrs</td>
</tr>
<tr>
<td>Estimated Effort</td>
<td>Mean Total Time per User</td>
<td>2.4 hrs</td>
<td>55.5 hrs</td>
</tr>
<tr>
<td>N Enrolled</td>
<td>Summed Number of Forum Events</td>
<td>53797</td>
<td>328970</td>
</tr>
<tr>
<td>N Certified</td>
<td>Mean Forum Events per User</td>
<td>3</td>
<td>219</td>
</tr>
</tbody>
</table>

**Analysis Tools and Data Sources:**

The data sources for this report come solely from the weekly exports provided by the edX analytics team. All data is loaded into MongoDB, and some pre and post processing is needed to generate figures in this report. More information regarding the format of raw data can be found here: [code.edx.org](http://code.edx.org).
LIST OF FIGURES

1.1 Course structure visualization. Course structure visualization highlighting the order and density of resources in 6.00.2x. The y-axis represents the temporal order of resources, with chapters indicated by text. Bars are color coded by course component (see legend), while height and length represent density and weight toward final grade, respectively. 5

2.1 Enrollment and Last Event figures. Enrollment is based on registration dates, while Last Activity is based on the last action of any participant. If a participant only registered, there is no last activity measure, leading to potential discrepancies between cumulative enrollment and the last activity survival function. An additional figure indicates population sizes for categories defined in the original MITx and HarvardX course reports [1]. 6

3.1 Demographic variables. Distributions of demographic variables collected during edX registration for all participants reporting data, comparing non-certified versus certified participants: (top left) geolocation via IP look-up, (top-right) gender, (bottom left) level of education, i.e., highest degree attained, and (bottom right) age. Note: other than geolocation, all demographic variables are collected at registration and are self-reported. Hence, variables such as level of education are potentially out of date depending on when a participant registered, and bias issues dealing with self-reported data may also exist. 7

4.1 Distributions of activity metrics. Distributions of activity metrics collected from tracking-log data for each participant: number of events [4.1a], number of active days [4.1b], number of problem submissions [4.1c], number of progress checks [4.1d], and number of forum events [4.1e]. 8

5.1 Unique users accessing content visualized via course structure. Course structure visualization (lower image in Fig. 5.1a and Fig. 5.1b) integrated with the number of unique users accessing an individual resource in the course. 9

5.2 Unique video and problem accesses. Distribution of unique videos (left) and unique problems (right) accessed by non-certified and certified participants. 9

6.1 Daily time series. Daily time series for selected activity metrics: unique users (visitors), forum submissions. 10

7.1 Estimation of total time spent in course. Time spent is an “estimate” due to the application of a lower (10 seconds) and upper (1 hour) boundaries intended to account for navigation events and idle windows, respectively. For more information on the choice of these boundaries, please see [2], “Who does what in a Massive Open Online Course?”, Seaton et. al, CACM 2014. Figures within this section include distributions time spent and final grade in 6.00.2x, and a scatter plot whose bubble size represents the total number of forum interactions (any interaction, including navigation). 11

7.2 Unique accesses aggregated by chapter and week relative to course launch. Bubble chart of Relative Course Week versus Chapter relative to the Course Structure Index, where bubble size and color indicate the number of unique users accessing a chapter. Relative week is defined by 7-day increments relative to launch date. Bubbles are only drawn if more than 20 unique users access a chapter in a given week. The course structure index represents the integer order of resources as they appeared in the course; this integer index represents the same order of resources in Fig. 1.1. 12

REFERENCES


(a) Legend describing color coding of course components in the course axis visualization (left). Within the main course structure visualization, the height of each bar represents density of resources, while length is correlated with weight toward final grade.

(b) Counts of resource categories existing in 6.00.2x, where each category represents a fundamental element of any edX course. For more information on the kinds of resources and naming conventions, please see the edX documentation [http://docs.edx.org/].

<table>
<thead>
<tr>
<th>Resource Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>vertical</td>
<td>210</td>
</tr>
<tr>
<td>problem</td>
<td>208</td>
</tr>
<tr>
<td>html</td>
<td>93</td>
</tr>
<tr>
<td>video</td>
<td>75</td>
</tr>
<tr>
<td>sequential</td>
<td>41</td>
</tr>
<tr>
<td>chapter</td>
<td>16</td>
</tr>
</tbody>
</table>

Figure 1.1: **Course structure visualization.** Course structure visualization highlighting the order and density of resources in 6.00.2x. The y-axis represents the temporal order of resources, with chapters indicated by text. Bars are color coded by course component (see legend), while height and length represent density and weight toward final grade, respectively.
(a) Registration date for all participants; counts per day (silver - left scale) and cumulative enrollment (red - right scale). Dashed lines represent course start and end dates.

(b) Last event (click) for participants with at least one click within the course; counts per day (gray - left scale) and a last-event survival function (red - right scale). Dashed lines represent course start and end dates.

(c) Participant classification percentages and counts. The registered, viewed, explored, and certified categories are defined here [1].

Figure 2.1: Enrollment and Last Event figures. Enrollment is based on registration dates, while Last Activity is based on the last action of any participant. If a participant only registered, there is no last activity measure, leading to potential discrepancies between cumulative enrollment and the last activity survival function. An additional figure indicates population sizes for categories defined in the original MITx and HarvardX course reports [1].
Figure 3.1: **Demographic variables.** Distributions of demographic variables collected during edX registration for all participants reporting data, comparing non-certified versus certified participants: (top left) geolocation via IP look-up, (top-right) gender, (bottom left) level of education, i.e., highest degree attained, and (bottom right) age. Note: other than geolocation, all demographic variables are collected at registration and are self-reported. Hence, variables such as level of education are potentially out of date depending on when a participant registered, and bias issues dealing with self-reported data may also exist.
(a) Number of events (clicks). For clarity, only binning values > 1.

(b) Number of days active (at least one click in a given day).

(c) Number of problem submissions.

(d) Number of progress checks.

(e) Number of forum events (note, this is any activity in the forums, including navigation). For clarity, only binning values > 1.

Figure 4.1: Distributions of activity metrics. Distributions of activity metrics collected from tracking-log data for each participant: number of events (4.1a), number of active days (4.1b), number of problem submissions (4.1c), number of progress checks (4.1d), and number of forum events (4.1e).
(a) **Resource accesses for all participants** in terms of the course axis visualization (Fig. 1.1). The upper plot is a count of unique users, color coded by the course axis visualization.

(b) **Resource accesses for certificate earners** in terms of the course axis visualization (Fig. 1.1). Upper plot is a count of unique users, color coded by the course axis visualization in the lower plot.

Figure 5.1: **Unique users accessing content visualized via course structure.** Course structure visualization (lower image in Fig. 5.1a and Fig. 5.1b) integrated with the number of unique users accessing an individual resource in the course.

Figure 5.2: **Unique video and problem accesses.** Distribution of unique videos (left) and unique problems (right) accessed by non-certified and certified participants.
Figure 6.1: **Daily time series.** Daily time series for selected activity metrics: unique users (visitors), forum submissions.
(a) Scatter plot of grade versus total time spent in course (each point is one participant and has been jittered to highlight density); bubble radii (size) are proportional to total number of forum events for a given participant (any forum interaction, including navigation).

(b) Distribution of total time spent in 6.00.2x.

(c) Distribution of final grade in 6.00.2x. For clarity, only binning values > 0.10.

Figure 7.1: **Estimation of total time spent in course.** Time spent is an “estimate” due to the application of a lower (10 seconds) and upper (1 hour) boundaries intended to account for navigation events and idle windows, respectively. For more information on the choice of these boundaries, please see [2], “Who does what in a Massive Open Online Course”, Seaton et. al, CACM 2014. Figures within this section include distributions time spent and final grade in 6.00.2x, and a scatter plot whose bubble size represents the total number of forum interactions (any interaction, including navigation).
Figure 7.2: **Unique accesses aggregated by chapter and week relative to course launch.** Bubble chart of Relative Course Week versus Chapter relative to the Course Structure Index, where bubble size and color indicate the number of unique users accessing a chapter. Relative week is defined by 7-day increments relative to launch date. Bubbles are only drawn if more than 20 unique users access a chapter in a given week. The course structure index represents the integer order of resources as they appeared in the course; this integer index represents the same order of resources in Fig. 1.1.