HarvardX Research: enrollment worldmap data specification

The interactive visualization appearing at nesterko.com/blog/2013/08/22/data-from-harvardx-research-worldwide-student-enrollment/ and at harvardx.harvard.edu/blog/interactive-visualization-worldwide-enrollment-harvardx is based on data on HarvardX enrollment for courses offered through the edX platform. Enrollment is defined as individuals on HarvardX course lists. This document aims to describe (1) the way data was prepared and its possible sources of error, (2) questions the visualization helps answer, (3) possible misinterpretations of the data.

Preparation and possible sources of error

The data was prepared by parsing self-reported mailing address provided at registration by current HarvardX students as of August 18, 2013, and then scaling up the counts by country for each course using total course student count and inferred country proportions under the assumption that any missing data is Missing At Random. Lists of countries, states, and provinces used in parsing, can be found at nesterko.com/visuals/worldmap-harvardx/country_names_and_code_elements.txt, nesterko.com/visuals/worldmap-harvardx/states.csv, and nesterko.com/visuals/worldmap-harvardx/provinces.csv. The Python code for the parser can be found at nesterko.com/visuals/worldmap-harvardx/enrollment_geolocation_self-rep.tgz. Resulting dataset is at nesterko.com/visuals/worldmap-harvardx/enrollment_geolocation_sr.csv.

Specification of possible sources of error:
1. The mailing address data is self-reported and may not be fully correct.
2. The data is time-invariant, so it actually records the country where a person resided at the time of their registration, not currently.
3. The data is inferred from a parser developed in Python using regular expressions to match country, state and province names from self-reported addresses. It does not work well with accents such as é. We estimate the error level due to parsing errors to be up to 10% of the inferred total student count for each country, but it could be much higher or lower on average and specific countries might be much higher.
4. Mailing address data were not collected for the first few months of edX, so people who were early members of edX are not included
5. Around 46% of students provided a mailing address from which a country could be parsed. The numbers of students for each country for each course was inferred using the Missing At Random (MAR) assumption for missing data and has error. It is possible that MAR assumption is not fully satisfied as, for example, people from certain countries might be less likely to choose to share their address, and therefore could be underrepresented.

Possible misinterpretations

1. It is important to note that this figure represents all registered students, without any kind of

classification filter based on activity in the course. It may be that the distribution of active students by country is different than the distribution of registered students by country.
2. As noted above, we cannot verify the accuracy of the location data, and we know there are several possible sources of error. For analyses where precise residence data is important, it might be better to use multiple sources such as course registration surveys, IP lookups and other methods.