Massachusetts Institute of Technology

6.UAP Final Report

MOOCdb:
An collaborative environment for MOOC data

December 11, 2013

Sherwin Wu

Contents

1 Introduction

With the current rise of massive open online courses (MOOCs) such as edX, Coursera, and Khan Academy, forward-looking educators have begun to collect massive amounts of data from their online students in the form of submissions, grades, and interactions with the online course content [1]. Hidden within the data are valuable, non-obvious insights about the nature of MOOCs and the students who use them. While each individual MOOC could operate individually to analyze their respective data sets, the Anyscale Learning For All (ALFA) group at MIT CSAIL believes that standardization of MOOC data would maximize the impact of any new insights for teachers and their students. The entire MOOC community is still at a very early developmental stage, and so we are in a unique position to lay the groundwork for a potentially large sector of the big-data community – particularly to avoid the mistakes made and technical debt accrued in other big-data sectors such as healthcare, transportation, and population data.

In the spirit of collaboration and open-source software, the ALFA lab envisions an environment where MOOC providers could work together by writing data analysis scripts to process the data, and creating visualizations to present the data. At the heart of this endeavor would be a unified database schema for all data stored by each of the MOOCs. We have proposed MOOCdb to be that unified schema for MOOCs. If all the different MOOCs store their data in the MOOCdb format, one script written by anyone could run on all the different MOOCdb instances to generate custom analytics across all data sets.

Using the MOOCdb schema, we created data pipelines to generate visualizations on a MOOCdb dataset. These pipelines, consisting of various data processing and visualizations scripts, have since been run at Stanford and other institutions on their MOOCdb instances, with great success. We have also created a working first iteration of MOOCvis, a web collaborative platform

for MOOCdb, which allows users to upload, share, and collaborate on analytics pipelines running on top of MOOCdb.

2 Motivation and Previous Work

The creation of the MOOCdb eco-system was motivated by a variety of factors. These include the observation of technical difficulties associated with non-standardization and divergent systems in other data sets, notably that of healthcare data. We also personally experienced, through other work done in the ALFA lab, the pain and loss in productivity when working with many disparate but similar data sets. Because the MOOC community is still developing, we thought it would be optimal to introduce a new precedent for this emerging data sector.

Creating a unified database schema for MOOC data, or any data set really, offers many benefits [1]:

Benefits of standardization: Standardization through a common database schema allows for cross-platform collaboration, sharing of data analysis pipelines and scripts, and definitions of variables that can be derived the same way across many MOOCdb databases. This project tries to capitalize mostly through this category of benefits.

Savings in productivity time: A unified schema eliminates the time needed to design a new schema. The collaborative element of the project allows for people to share database population scripts, so that piping data into a database can also be performed much more quickly.

Crowd source potential: Although not part of this project, we are using MOOCdb to run machine learning algorithms on the datasets. This frequently involves humans identifying features or variables to drive a response. With a unified database schema, we can intentionally consider the database

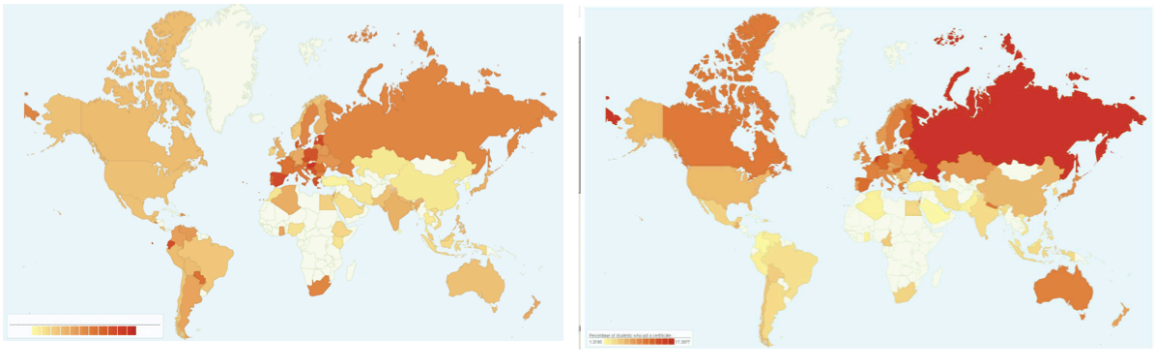
schema independent from the data, and allow crowd sourcing of feature identification without access to the data itself.

3 MOOCdb Visualizations

We have created several visualization scripts, which process and aggregate MOOCdb data from the database. While many visualizations have arisen from MOOCdb, to keep our report concise, we will present three different types of visualizations that we have made – a map based visualization, a stacked bar chart, and a force-directed nodes and links graph. Further, to demonstrate the benefits of the standardized MOOCdb schema, we try to present visualizations run on two instances of MOOCdb – our instance from 6.002X offered on MITx, and Stanford's instance from Crypto 1 offered on Coursera.

All three of the presented visualizations were generated by writing a series of data processing and visualization scripts that could be run on any MOOCdb instance. The map based visualization and stacked bar chart were indeed run on both instances of MOOCdb at MIT and Stanford, while the force directed graph visualization was only run on the MIT instance.

3.1 Map based visualization



(a) 6.002x: MIT/MITx (b) Crypto 1: Stanford/Coursera

Figure 1: Map coloring by country showing the ratio of certificate-winning students to the total number of registrants for (a) 6.002x offered at MITx and (b) Crypto 1 offered at Coursera. Darker colors represent higher ratios.

Figure 1 shows a colored map representing the ratio of the number of students awarded certificates in the course to the total number of registrants for the course. Darker shades of red represent higher ratios. The MOOCdb standardization allows both visualizations to be generated quickly using the same set of scripts. One can see that countries in Eastern Europe outperform the United States in this certificate to registrant ratio in both course offerings.

3.2 Stacked bar chart

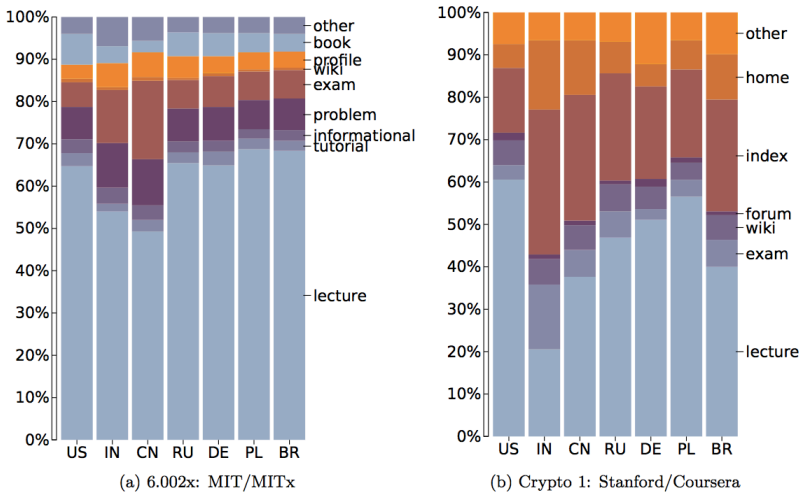
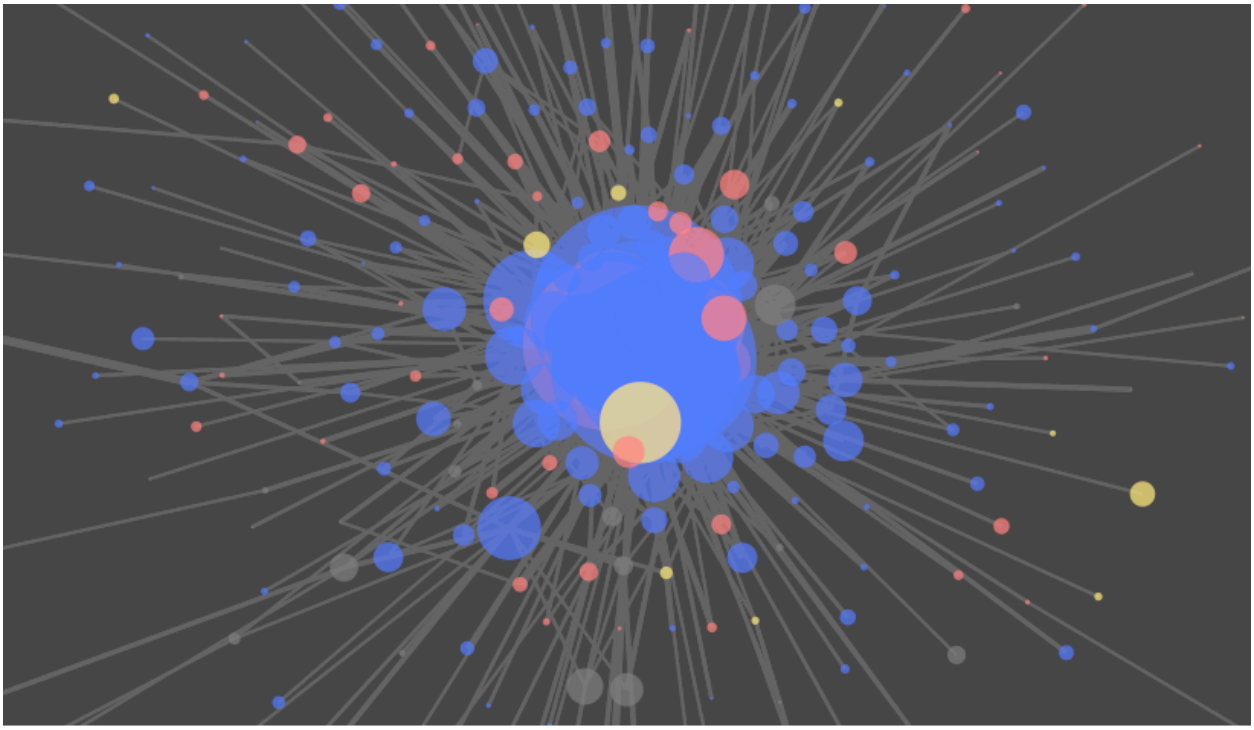


Figure 2: Stacked bar charts showing the relative time spent on different resources in the course by country for (a) 6.002x offered at MITx and (b) Crypto 1 offered at Coursera.

Figure 2 shows a stacked bar chart of the relative times spent on different resource types in the course, grouped by country. Again, both were easily generated by running the same exact data processing scripts and slightly modified visualization scripts. One can see interesting trends present in both courses. Students in India (IN) and China (CN) spend more time than students in the United States (US) on exams, but spend significantly less time on lectures – perhaps due to the language barrier.

3.3 Force Directed Graph



(a) 6.002x: MIT/MITx

Figure 3: A force directed graph showing the community on the 6.002x forum. Not pictured are a small number of 'detached' connected elements.

We also generated a force directed graph to visualize the interactivity of students on the 6.002x forum, shown in Figure 3. Nodes represent students who posted on the forum (blue nodes are 'A' students, red nodes are 'B' students, and yellow nodes are 'C' and below students), and edges between nodes were present when a student answered another student's question or commented on another student's answer. It is interesting to see how there is a very prominent 'core' to the community, which consists mainly of 'A' students (as evident by the largely blue center of the graph), and almost all other posts involved the core group of students. There was a small contingent of disconnected components to the graph (not pictured), but the vast majority of the posts on the forum were connected to the 'core' students.

4 MOOCvis

To build upon and expand the impact of the many visualizations that we made, we created MOOCvis, a web application to facilitate sharing data pipelines and scripts that could be run on any MOOCdb instance. MOOCvis would allow for scientists from different MOOCs and institutions to upload and share their visualizations, so work done by one person could benefit everyone in the MOOC community.

There are two core use cases at the heart of MOOCvis – one for viewing and exploring visualizations put up by other people, and secondly, uploading and sharing a visualization. We look at each functionality separately.

4.1 Visualization viewing and exploring

The home page of MOOCvis is a 'gallery view', where users see a grid view of all the different visualizations that have been shared on MOOCvis. The gallery displays an uploaded thumbnail of the visualization for a quick preview of what the visualization looks like. The gallery is organized by 'tags', which are high level categorizations of visualizations working on different sectors of the MOOCdb data space. Examples of tags may be 'resources', 'collaborations', or

Figure 4: The representation of a visualization pipeline and the display of a processing script's contents. Clicking on each step of the pipeline will load the contents of another file. There is an option to download all the files in the pipeline in a zip file, and also an option to download each individual file (not pictured).

'submissions'. If a visualization works primarily with the submissions of students, it should have a tag for 'submissions'.

Upon clicking on any visualization, you are taken to the view page for a specific visualization. On the view page, you can see all the metadata associated with a particular visualization (visualization name, description, and author username), as well as view the visualization itself as embedded HTML. We encourage visualizations to be made in d3.js (to maximize portability and cross-platform compatibility); consequently a large section of the view page for each visualization shows the embedded HTML from the visualization. This works extremely well with d3 visualizations in a manner similar to Mike Bostock's bl.ocks framework.

Below the embedded HTML visualization, we show a representation of the visualization pipeline, as well as the content of the files in each step of the pipeline (Figure 4). Clicking on each part of the pipeline will load the contents of that step's file, whether it is a script or a data file. There are also options to download a zip file of all the files in a given pipeline, and an option to download the current file you are viewing. So, from one page, a user can explore the entire pipeline that was used to generate a particular visualization.

A given visualization may also be run on multiple instances of MOOCdb (or different course 'offerings'), hence on the visualization view page, there is also an option to change offerings. If the offering identifiers are changed, the visualization for the newly selected offering may be loaded as well as the associated data files for that new offering.

4.2 Visualization sharing and uploading

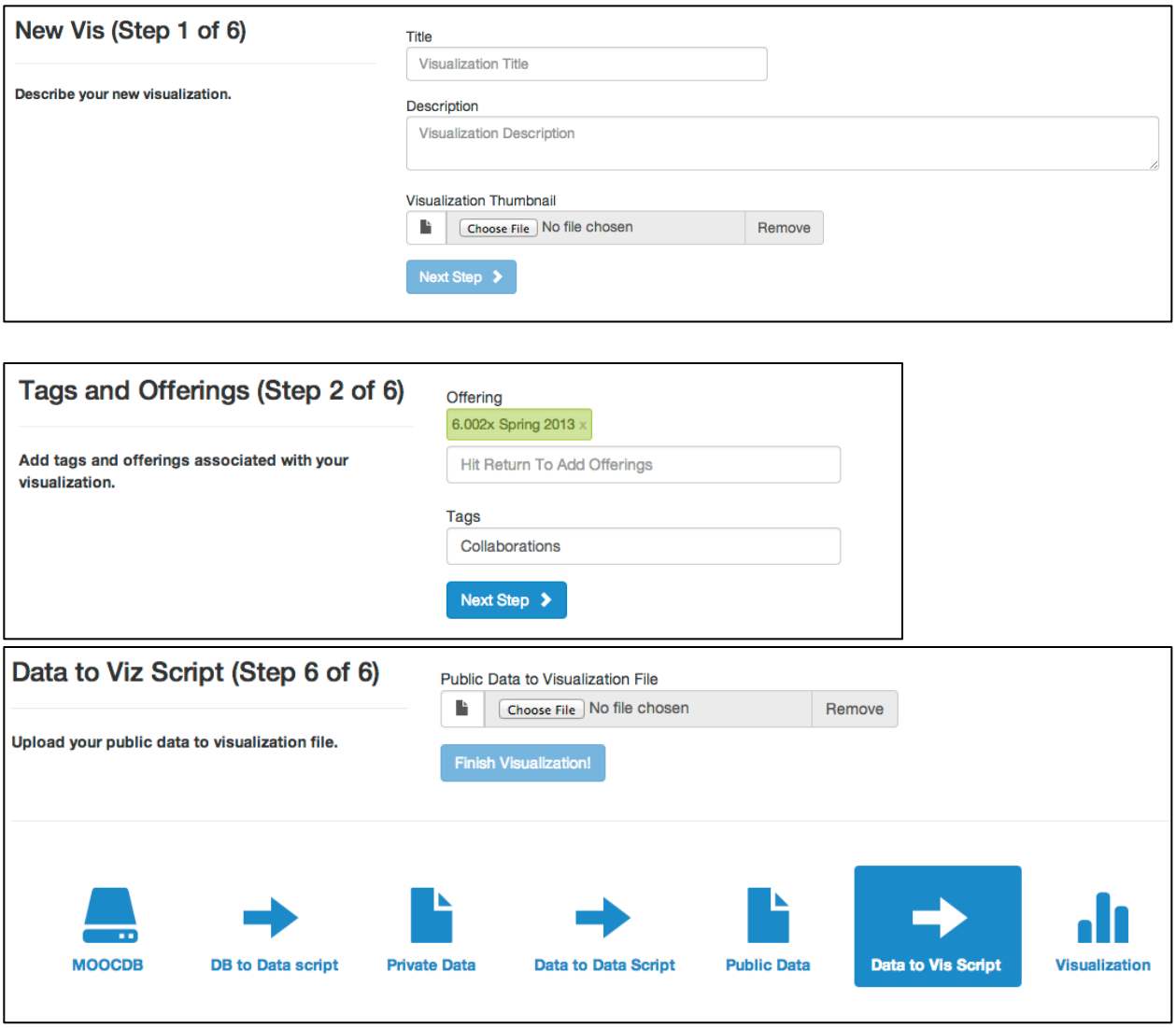


Figure 5: The visualization creation and upload page (we only included 3 of the 6 steps). Users can dynamically add offering labels and tags to the visualization, as well as easily upload their visualization files to MOOCvis in one page.

We also wanted to easily allow scientists with an existing visualization pipeline to share it on MOOCvis. The visualization creation and upload page allows users to upload all their files to MOOCvis with instant gallery display both on the upload and view pages.

On the upload page, users can easily and dynamically add or remove offerings and tags for the new visualization they are uploading. To reduce duplication of tags and offering names, when users type in an offering or tag name, we will autocomplete with the current offerings and tags on

MOOCvis already. On this page users can also easily add and upload their visualization pipeline files by adding script or data files to incrementally build up the pipeline.

5 Conclusion

The ALFA lab set out to create a generalizable foundation for the MOOC data eco-system, and started ambitiously by creating MOOCdb, a well-defined and robust schema for MOOC data. This project aimed to build upon that success and delve deeper into what could be done with the MOOCdb standard.

We have created many visualizations that can be run on any MOOCdb dataset, and presented three different types of those visualizations in this paper. These visualizations provided interesting insights into the courses and how the students interacted with them. Further, we were able to easily and conveniently show that the results applied across different courses by letting other institutions with MOOCdb instances (Stanford/Coursera) run our scripts. This shows the value in standardization, as well as the success of the MOOCdb schema.

We then aimed to help facilitate the collaborative environment further with the development of MOOCvis, a collaborative web platform for the MOOCdb eco-system. We showed that our application could allow scientists to easily upload and share their existing scripts and data processing pipelines, but could also allow anyone – with any technical background – to view the visualizations present on MOOCvis and download the scripts to run on their own instance of MOOCdb. By removing the friction between collaborators, we hope to see the collaborative setting that we originally envisioned come to fruition. Scientists from all over the world could download scripts and create visualizations on their own data sets, allowing for valuable cross-platform data to emerge,

Analysts could be inspired by someone else's scripts, and then proceed to download them and modify them slightly to provides new insights not available through the previous script.

We are optimistic about the MOOCdb endeavor, and we can foresee certain next steps to take in this project. One clear next step would be to iterate on MOOCvis and provide a cleaner release with additional features, such as support for custom data pipelines, including a 'run all' script inside the zip file of scripts, and integration with Git for better version control of uploads. Further, we could capitalize on the crowd-source potential of a standardized database schema to propel forward the machine-learning applications of the project. We are in the beginning steps of creating another web application which would allow for crowd-sourced feature identification on the MOOCdb dataset. Because of the standardized schema, the crowd could identify features based on the MOOCdb schema without access to the sensitive data itself.

6 References

[1] Veeramachaneni, et. all. "MOOCdb: Developing Data Standards for MOOC Data Science", Proceedings of the 1st Workshop on Massive Open Online Courses. "http://edf.stanford.edu/sites/default/files/Verramachaneni%20et%20al.%202013.pdf".