

Abstract

Because MOOCs bring big data to the forefront, they confront learning science with technology challenges. We describe an agenda for developing technology that enables MOOC analytics. Such an agenda needs to efficiently address the detailed, low level, high volume nature of MOOC data. It also needs to help exploit the data's capacity to reveal, in detail, how students behave and how learning takes place. We chart an agenda that starts with data standardization. It identifies crowd sourcing as a means to speed up data analysis of forum data or predictive analytics of student behavior. It also points to open source platforms that allow software to be shared and visualization analytics to be discussed.

AUTHORS

Una-May O'Reilly, Ph.D.
Massachusetts Institute
of Technology

Kalyan Veeramachaneni, Ph.D.
Massachusetts Institute
of Technology

Technology for Mining the Big Data of MOOCs

Massive Open Online Courses (MOOCs) are college courses offered on the Internet. Lectures are conveyed by videos, textbooks are digitized, and problem sets, quizzes and practice questions are web-based. Students communicate with one another and faculty via discussion forums. Grading, albeit constrained by somewhat restrictive assessment design, is automated.

The popularity of MOOCs has made a high volume of learner data available for analytic purposes. Some MOOC data is just like that which comes from the classroom. This can include teaching material, student demographics and background data, enrollment information, assessment scores and grades. But very important differences arise between MOOC and classroom in how behavioral data is collected and what is observable. The platform records, unobtrusively, through input, capture every mouse click, video player control use, and every submission to the platform such as problem solution choice selection, solution composition or text entry for a forum discussion. The level of recorded detail of behavior in a MOOC vastly surpasses that recorded in conventional settings.

Very directly, this data can provide a count of problem attempts and video replays. It can reveal how long a student stayed on a textbook page or the presence of very short, quick patterns of resource consultation. It can inform an individualized or aggregated portrait of how a student solves problems or accesses resources. It presents opportunities to identify and compare different cohorts of students in significant quantities, thus enabling us to personalize how content is delivered. It allows us to study learner activities not exclusive to problem-solving, such as forum interactions and video-watching habits (Thille et al., 2014). It also facilitates predictive analytics based on modeling and machine learning.

This data also contains large samples. Large sample sizes enable us to rigorously confirm or deny long held hypotheses about how learning takes place, whether there exist learning styles, whether there are effective ways to learn or teach types of material or whether there are effective concept correction strategies to help a student who has made an error.

CORRESPONDENCE

Email
unamay@csail.mit.edu

Beyond comparative studies, from a predictive modeling standpoint, we can build and validate predictive models at a scale never done before. For example, we can now build a reliable predictor for which students will exit the course before completion (Taylor, Veeramachaneni, & O'Reilly, 2014). In short, MOOC big data is a gold mine for analytics.

The enormous potential of MOOC big data prompts the questions: what are the appropriate ways to fully tap into it? What technology can be brought to practice to analyze it more efficiently and broadly? The process of answering these questions reveals challenges. The data is *high volume* and *low-level* in nature. Complete answers to any research question need to analyze the data from multiple entities, i.e., courses, platforms, institutions. The perspectives of multiple parties – students, instructors and education researchers – need to be explored.

We have decided to focus our research agenda on the challenges that arise from MOOC data characteristics and analytics needs. We have embraced increasing the number of contributors to MOOC analytics and accelerating analytics accomplishments as our central mission. We are focusing on developing community-oriented means of sharing software and analytic development efforts.

We start by proposing data standardization as a cornerstone. It will resolve the different formats of data resulting from different platforms. It will prevent MOOC data from following the path of healthcare data, which, even if privacy issues are completely resolved, is fragmented by different formats. It will also make the task of extracting variables for analyses more efficient, collaborative and sharable. We next propose easy-to-use, web-based platforms that democratize different aspects of data analytics:

- *MOOCviz* lets anyone share visualization software and their analytic renderings.
- *FeatureFactory* helps learning scientists enumerate possible variables for their models.
- *LabelMe-Text* helps learning scientists engage the crowd to get help tagging forum posts before they use machine learning to automate a labeler from the tagged examples.

MOOCdb – A Cornerstone for Shared Analytics

In order for a data oriented platform or framework to allow anyone to use it, it needs to either deal with many formats of data or be able to expect that all data is in a common format. The former proposition imposes a lot of extra work versus the latter. It leads to different versions of software. It bulks logic in software to dealing with format differences and it requires software updates every time a new format emerges. Thus, to make the latter proposition viable, we have pioneered a standardized schema for MOOC data (i.e., a data model) that is platform agnostic. It is called MOOCdb (Veeramachaneni, Halawa, et al., 2014).

The MOOCdb data model originally organized MITx data generated from the MITx platform that has now transitioned to edX. It offers advantages beyond what we emphasize here, among them removing the need to share data, independence from platform specifics and facilitating a data description that outsiders can refer to when contributing expertise in data privacy protection or database optimization. During the past year, we have adapted it to also capture the data subtleties and idiosyncrasies of both edX and Coursera platforms. A periodically updated technical report explains the data model, all the fields and how they are assembled for each platform. Complete documentation for MOOCdb and its data model will be perpetually updated via the wiki site <http://moocdb.csail.mit.edu/wiki>.

The MOOCdb data model is based on some basic core actions that students take on any online learning platform. Students usually interact with the platform in four different modes: *Observing*, *submitting*, *collaborating* and *giving feedback*. In observing mode students are simply browsing the online platform, watching videos, reading material, reading books or watching forums. In submitting mode, students submit information to the platform. This includes submissions towards quizzes, homework, or any assessment modules. In collaborating mode students interact with other students or instructors on

Large sample sizes enable us to rigorously confirm or deny long held hypotheses about how learning takes place, whether there exist learning styles, whether there are effective ways to learn or teach types of material or whether there are effective concept correction strategies to help a student who has made an error.

forums, collaboratively editing wiki or chatting on Google hangout or other hangout venues (Veeramachaneni, Halawa, et al., 2014).

To date, much of the analyses on MOOC data have been conducted with techniques transferred from conventional learning analytics or modestly adapted from them.¹ In the first three stages of their study, Breslow et al. (2013) followed a conventional methodology adapted for MOOC relevant questions. They worked with coarse-grained variables. That is, they studied the aggregate of certificate earners (choosing not to further subdivide students), they operationalized achievement to use the course grade (choosing not to consider specific problem set grades or time sequences of assessment grades) and they referenced factors such as age, gender and highest degree earned (choosing not to reference behavioral factors such as instructional component access). MOOCdb standardization will further leverage such work because it supports the extraction of quantities that can be composed into fine grained variables. It allows anyone to formulate (and answer) learning science research questions that are adaptations of conventional methods considering finely subdivided students, their achievements and their access of MOOC's instructional components.

Infrastructure for Sharing Data Visualizations

Transforming data into meaningful visualizations is a core part of any data science. In MOOC data science, different institutions, local research communities, user groups and other sorts of organizations, each have multiple stakeholders who have different needs that require data to be transformed in a different way and visualized. Ideally, they want to support each other as much as possible in this context by sharing software, demonstrations and opinions on design and interpretations of data.

Visualization infrastructure can provide one means of supporting this. HarvardX and MIT's Office of Digital Learning enable visualizations of their MOOC data^{2,3} via complementary website entitled *Insights*. These visualizations use world maps to show enrollment, certificate attainment by country, gender, education levels and age composition (Ho et al., 2014; Nesterko et al., 2013). Visualizations referencing clickstream or forum data are currently not available⁴ likely because plotting these streams is significantly more complicated. A streamlined workflow that reduces development time through software sharing and data standardization would reduce these complications.

The *Insights* website is also used as a distribution point and makes a modest attempt to encourage other visualizations that reference the data. For example, along with the data that populate visualizations, *Insights* makes source code and documentation available for download,⁵ though only as separate, non-integrated files. The website exemplifies a strong but minimal starting point for providing visualization infrastructure. Ideally, even beyond supporting better-integrated software sharing, an infrastructure needs to support the contribution of new visualizations. These should be able to come from others, i.e., not only the site's creators. Opening access to the community, so they can contribute, will allow many different questions to be answered by data visualizations expressed in multiple ways. It will address the reality that different people perceive different visualizations as useful.

People analyzing visualizations for their usefulness tend to zero in on either on the aesthetics of the visualization, e.g., a style choice like bar or pie chart, color or interaction

We start by proposing data standardization as a cornerstone. It will resolve the different formats of data resulting from different platforms. It will prevent MOOC data from following the path of healthcare data, which, even if privacy issues are completely resolved, is fragmented by different formats.

¹In the first paper in RPA on MOOCs, Breslow et al. (2013) note: Our first challenge has been choosing, or in some cases adapting, the methodological approaches that can be used to analyze the data. If educational researchers studying conventional brick and mortar classrooms struggle to operationalize variables like attrition and achievement, it is doubly difficult to do so for MOOCs (p. 14).

²MITx Insights is a collection of interactive data visualizations for all MITx offerings, updating at frequent, regular intervals. These visualizations are released along side a complementary set of visualizations from the HarvardX Research Committee. (url: <http://odl.mit.edu/insights/>)

³HarvardX Insights is a collection of interactive visualizations of learner data, which dynamically update at frequent, regular intervals. (url: <http://harvardx.harvard.edu/harvardx-insights>)

⁴In their reporting, the team notes: "The MITx and HarvardX Research teams intend for future interactive visualizations to include more nuanced descriptions of student participation and learning in our open online learning environments."

⁵It is highly structured and organized so whether it will support different visualizations is an open question (see e-literate for an opinion).

mode, or on the way the data was organized and aggregated before it was visualized. Such remarks motivate a fundamental goal for visualization infrastructure: to support a proliferation of many views of same data. This goal has driven us to develop a platform called MOOCviz that we now describe.

For example, we can now build a reliable predictor for which students will exit the course before completion.

MOOCviz – Sharing Software and Outcomes of Visualization

The MOOCviz platform (Figure 1) is designed to serve the diverse needs of a broad group of stakeholders and facilitates the sharing of software, demonstrations and opinions on design and interpretations of data. It enforces source code organization, allows source code to be contributed to a repository and it provides a means of web-based discussion around a visualization, all fundamental tenets for a community oriented infrastructure.

Transforming data to create visualization typically requires three steps: *source data extraction*, *variable formation (typically aggregation)* and *rendering*. Each of these steps is somewhat specialized according to each situation. They embed some assumptions and integrate some heuristics to transform and shape the data to create an interesting and informative visualization. Anyone with access to MOOC data in MOOCdb schema can develop a brand new visualization, modularize their software into the aforementioned three steps, extract, aggregate and render, and then upload the modules into MOOCviz’s software repository along with their first demonstration of the new visualization for other members to use and view.

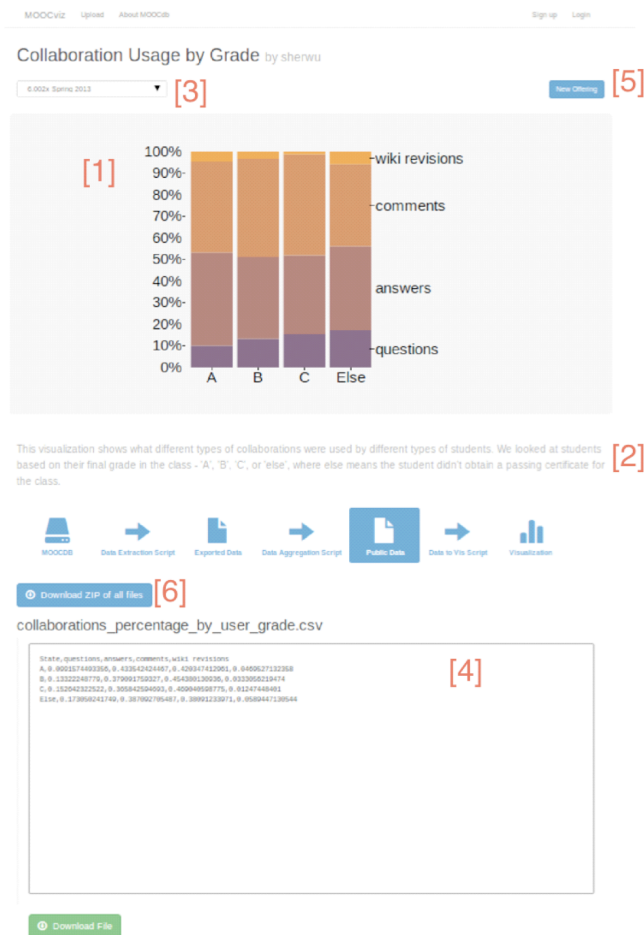


Figure 1. Current state of the MOOCviz platform. Users can select the course for which they would like to see the visualization (see [3]). The visualization is rendered in panel [1] and is described below the panel (see [2]). The workflow that generated the visualization from MOOCdb is shown below the description. users can click on any of the icons in the workflow and corresponding software or data is shown in panel parked as [4]. Users can upload the visualization for a new course by using the "New Offering" functionality (see [5]). [6] allows users to download the entire code from data extraction, aggregation to visualization.

⁶ In order to help a viewer choose between different visualizations, it will use popularity to rank multiple visualizations and only show the most popular one.

The *MOOCviz* platform software will eventually be shared under an open source license, and an organization or an instructor will be able to download and install it to create an independent instance, which they can populate with visualizations of their own data in MOOCdb format. Any member of the community will be able to enhance the platform's open source software and customize it to support specific use cases; e.g., cross-course comparisons or a single course report with multiple visualizations.

A *MOOCviz* platform offers:

- A central, shared gallery of participant-generated visualizations for a list of courses for which they have been rendered.
- The ability for the participants to download the software that generates visualizations and execute it over their own course data that is formatted in MOOCdb schema. They will also be able to automatically package the resulting rendered visualization and upload it to the gallery, adding to the list of courses.
- A means to contribute software for new visualizations to the gallery via the *MOOCviz* web-based interface.
- A means of commenting on any existing visualization by posting in the *comments* section underneath it. Discussions are free form. They likely will extend beyond the interpretation or thoughts provoked by the visualization to the ways that the data have been transformed in extraction and aggregate steps. We expect that discussions will stimulate ideas for new visualizations.

Infrastructure for Supporting Feature Engineering

Scaling feature engineering involves three processes: proliferation of an *ideation* process, the process in which candidate features are posited; support for an *operationalization* process, in which a mapping is formed between the data sources and the feature; and a feature *extraction* process, in which software is written to realize instances of these features.

The study of stopout, that is, predicting when students stop engaging course material before completion, provides an example (Taylor et al., 2014). If the outcome set is whether or not a student stops out, what predicts a stopout could include frequency of forum posts, grades to date, most recent problem set score, time spent watching videos, etc.

We have been formulating predictive and explanatory features for stopout. In the course of doing so, we have observed that the set of possible features for an outcome is likely much larger than we ourselves can propose (Veeramachaneni, O'Reilly, & Taylor, 2014). This is because our own experiences (or lack thereof), biases and intellectual context can go only so far and may be imposing limits on our investigations. This is a shortcoming not unique to us alone.

When working on stopout prediction (Taylor et al., 2014), we first tried to address this shortcoming by setting up meetings with students and instructors of a MOOC. At the meeting, we would solicit in person via a somewhat informal protocol, a group's input for predictors of stopout. We asked our helpers to fill out a form listing variables that would predict a student stopping out. We would then operationalize these variables via extraction and some modest arithmetic and add them to our predictor set (Veeramachaneni, O'Reilly, et al., 2014).). These exercises begged a general question: how can any MOOC data science group access a wider swath of the MOOC community to expand their feature/predictor list? As well, considering our mission to enable technology for MOOC analytics, how can we provide a general means of crowd access to the MOOC data science community at large?

FeatureFactory – Engaging the MOOC Crowd to Provide Hypotheses

To address both these questions, we are developing a second web-based collaborative platform called *FeatureFactory*. Our current version of this platform is shown in Figure 2. *FeatureFactory* offers two modes of engagement:

In order for a data oriented platform or framework to allow anyone to use it, it needs to either deal with many formats of data or be able to expect that all data is in a common format.

- The *solicit* mode is used by MOOC data science, education technology, or learning science research teams. A team describes the outcome it is currently studying or trying to predict. They give examples of what features or explanations are sought and it solicits help from the MOOC crowd.
- In the second mode, *helping*, the crowd proposes, explanations or variables, and suggests means to operationalize them. They provide comments on proposal or vote them up or down in popularity. The software savvy among them write and share software scripts written to operationalize the most popular or compelling proposals.

The MOOCdb data model is based on some basic core actions that students take on any online learning platform. Students usually interact with the platform in four different modes: Observing, submitting, collaborating and giving feedback.

Like *MOOCviz*, we intend to open source license and share the *FeatureFactory* platform software, so that an organization can create its own independent instance for local use. An organization can also customize their instance by modifying the platform source. They can use their platform in contexts when they need to garner assistance from the MOOC crowd.



Feature Factory MIT CSAIL ALFA Lab

Feature discovery is a challenging aspect of the data science and knowledge discovery. Creating an online interactive space where data scientists can benefit from each other's ideas on various features can significantly simplify and expedite the process. Feature Factory is an online platform where ALFA@CSAIL will present a prediction problem for which features are sought. For the prediction problem, the group will provide downloadable mock data so users can write their scripts and submit. Feature Factory seeks three kinds of contributions: ideas of new features, feature extraction code and comments on existing ones. [1]

Upon the submission of the feature extraction code, it will be validated on our online mock dataset and you will be notified of the result immediately. Upon validation, our team will execute the code on the real dataset to generate the features and insert the new feature into a number of machine learning models using discriminative (Decision trees, Neural networks, support vector Machines), generative (logistic regression, Gaussian process) and time series models. As a result, your features will be ranked against one another.

Current Focus Problem: Predict Student Stopouts on Massive Open Online Courses

In this problem, our goal is to predict when a student will stop engaging with the course. A student is assumed to have stopped out from a course when s/he stops to attempt problems/homeworks. We have data captured from students online behavior, which includes click stream data, their online forum interactions and their submissions for problems. We have a comprehensive data schema, called MOOCdb which captures the student activity data on a MOOC platform. The data schema is documented here. A small mock dataset that is in the form of the data schema can be downloaded in two formats: sql or csv.

We solicit participants for three distinct activities:

1. Propose a new feature by clicking on Add an idea
An example of a possible feature for this problem is: *Amount of time student spent on the course*
Below you can see a number of features already developed and extracted.
2. Write an SQL script for your idea or for an already existing idea
Below you can see a list of feature ideas. For some of them, extraction has not yet been performed.
3. Comment on an existing ideas Ideas develop when they are refined. So please feel free to comment or like the existing scripts.

Add an idea

Existing ideas and scripts

Average time (in days) the student takes to react when a new resource is posted. This pretends to... read more by Josep Marc Mengot	code ✓ comment 0 like 0
average time between problem submission time and problem due date by Rob Miller	code ✓ comment 0 like 0
Total time spent on each resource during the week by Franck	code ✓ comment 0 like 0
Number of forum posts by Franck	code ✓ comment 0 like 0
Number of Wiki edits by week by Franck	code ✓ comment 0 like 0

Figure 2. Current state of the *FeatureFactory* platform. In this illustration we show a screen shot of the website. First the rationale behind the *FeatureFactory* is described (see [1]), the current prediction problem of interest is described and the role participants can play is described (see [2]). Participants can submit a new idea using "Add an idea" (see [3]). Ideas collected so far are revealed under "Existing ideas and scripts" (see [4]). Participants can view the code (if available), comment on the idea and vote on the idea. All input from participants is collected in the back end in a database.

Infrastructure for Annotating Text

A central component of MOOC data is discussion forums. They are of great interest because they provide a lens on inter-student communication that, in turn, relates to learning science theories of engagement and achievement and self-efficacy. Most such language understanding tools rely on annotations of the content by humans (Gillani, 2013; Gillani & Eynon, 2014) and then employing machine learning to automatically annotate the text. The annotations range from qualifying the sentiment of the post, to tagging the posts by their types (content related, social affective, administrative, and other) to type of post (help seeking, help providing, neither) and many others. These tags help analyze the posts to understand the mood of the class, group posts by categories when presenting to the instructors, teaching assistants and others, categorizing students based on their post types so interventions can be designed, generating predictive variables for models on a per student basis and understanding the social discourse in the course (Rosé et al., 2014; Yang, Sinha, Adamson, & Rosé, 2013).

A working paper by Stump, DeBoer, Whittinghill, and Breslow (2013) provides a detailed account of how a protocol to annotate MOOC discussion forum posts was developed. The authors employed two students and used themselves to annotate the posts using a pre-determined set of labels derived from a categorization scheme. To facilitate their workflow they passed around an encrypted csv file that recorded labels. They then evaluated the quality of human annotations via a number of metrics that relate to inter-rater reliability. They finally filtered out ambiguously labeled posts. While they had over 90,000 forum posts, they found it impossible to examine and label all of them. They had to settle for referencing ~4,500 labeled posts. It is obvious that interpreting an entire set of posts would be preferable. But the process is slowed by the involvement of humans and hindered by the awkwardness of an ad hoc workflow. Concurrently, discussion arose outside the project arguing for an alternative annotation scheme (Gillani, 2013; Gillani & Eynon, 2014). This implies that annotation needs to become much easier because it will need to be done many ways by multiple research teams.

This context led us to consider what MOOC specific technology we could design to deal with such a large scale set of text and to support labeling according to the different annotation schemes of different studies. First, a web-based framework can support crowd based labeling for larger scale labeling. Second, the process and the workflow for processing labels can be streamlined. Third, much of the labeling can be automated. Machine learning can be used on the set of labeled posts to learn a rule for labeling the others, based upon features in the post. To address these needs, we are developing a web-based platform called *Label Me-Text*.

LabelMe-Text – Engaging the MOOC Crowd to Help with Forum Annotation

We developed an online platform where users would post their tagging projects and a crowd of helpers can participate in MOOC data science by selecting a project and tagging the content based on some instructions. We call the online collaborative platform that serves this purpose *LabelMe-Text's*.⁷ *LabelMe's* current incarnation is shown in Figure 3. It works in the following ways:

- Users requiring annotation of natural language can create an annotation project by providing a csv file for the content, instructions and examples for tagging.
- Taggers (*LabelMe-Text's* crowd) can participate by selecting a project, following the instructions and tagging the content.
- A database consisting of tags for the content for the project is initialized and populated as taggers work. A number of analytic services are provided around this database such as evaluation of inter rater reliability, summary of tags, and summary of activity for a project (how many taggers helped, time series of number of tags).

MOOCdb standardization will further leverage such work because it supports the extraction of quantities that can be composed into fine grained variables.

Transforming data into meaningful visualizations is a core part of any data science.

⁷ A framework called LabelMe already exists in the Computer Vision community (Russell, Torralba, Murphy, & Freeman, 2007). We used the same name, but identify it with suffix – text, by calling it LabelMe-Text.

- A service can be called upon to filter the tagged data based on the reliability measures just mentioned. It then uses methods based upon latent semantic analysis to learn a tagging model.
- Taggers (*LabelMe-Text's* crowd) are given credit for every tag they have provided and the number of their tags that pass the filters to be used in model learning.

Transforming data to create visualization typically requires three steps: source data extraction, variable formation (typically aggregation) and rendering.

Like *MOOCviz* and *FeatureFactory*, *LabelMe-Text* is open source software. Its eventual release will support organizations that wish to download and create a local version of it for internal use.

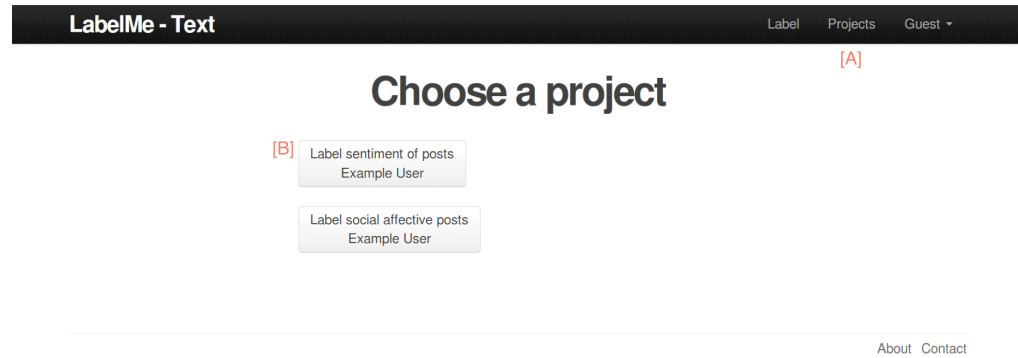


Figure 3. Crowd can select a project posted by a researcher by clicking on "Projects" marked using [B]. In this screen shot two such projects appear where it is marked as [A].

A central component of MOOC data is discussion forums. They are of great interest because they provide a lens on inter-student communication that, in turn, relates to learning science theories of engagement and achievement and self-efficacy.



Figure 4. Once users select the project, they then proceed to tagging/annotating a post/sentence dynamically selected by the platform from the pool of posts/sentences that need to be tagged. The sentence is displayed (see [A]), the choices for tags are displayed underneath it (see [B]) and instructions for tagging are presented as well (see [D]). The user can select the tag and hit "Submit Labels" (see [C]). All inputs from the participants/users are stored in a structured format in the back end in a database.

Conclusion

This paper considers the complexity MOOCs bring into learning science in view of the novel nature of the data they collect. It identifies certain technology challenges that need to be resolved before we can exploit the big data in MOOCs to its full potential. We call for enabling technology and for setting a course towards standardization and web-based platforms that help a large community of people to collaborate on developing analytics. We advocate frameworks that are deliberately open source so that, when they are released, everyone will be able to customize, refine and advance them.

AUTHORS NOTE:

We would like to thank the following student teams and students who, by contributing substantial efforts toward platform development, have been helping us fulfill our vision as described in this paper: Kiarash Adl (*FeatureFactory*), Preston Thompson, Colin Taylor, Brian Bell, Sherwin Wu (*MOOCviz*), Roy Wedge (*LabelMe*), Franck Dernoncourt (*MOOCdb*), Quentin Agren (*MOOCdb*), Sherif Halawa (*MOOCdb*). We would also like to acknowledge discussions with Juho Kim. We are grateful for funding from Quanta Computer. We also thank our reviewers.

References

- Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, 8(1), 13–25.
- Gillani, N. (2013). Learner communications in massively open online courses. *OxCHEPS Occasional Paper*, 53, 1–51.
- Gillani, N., & Eynon, R. (2014). Communication patterns in massively open online courses. *The Internet and Higher Education*, 23, 18–26.
- 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 Working Paper No. 1).
- Nesterko, S. O., Dotsenko, S., Han, Q., Seaton, D., Reich, J., Chuang, I., & Ho, A. D. (2013). Evaluating the geographic data in MOOCs. In *Neural Information Processing Systems*.
- Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P., & Sherer, J. (2014, March). Social factors that contribute to attrition in MOOCs. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 197–198). ACM.
- Russell, B., Torralba, A., Murphy, K., & Freeman, W. T. (2007). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1–3), 157–173.
- Stump, G. S., DeBoer, J., Whittinghill, J., & Breslow, L. (2013). Development of a framework to classify MOOC discussion forum posts: Methodology and challenges. TLL Working Paper, 2013.
- Taylor, C., Veeramachaneni, K., & O'Reilly, U. M. (2014). Likely to stop? Predicting stopout in massive open online courses. *arXiv preprint arXiv:1408.3382*.
- Thille, C., Schneider, E., Kizilcec, R. F., Piech, C., Halawa, S. A., & Greene, D. K. (2014). The future of data-enriched assessment. *Research & Practice in Assessment*, 9(2), 5-16.
- Veeramachaneni, K., Halawa, S., Dernoncourt, F., O'Reilly, U. M., Taylor, C., & Do, C. (2014). MOOCdb: Developing standards and systems to support MOOC data science. *arXiv preprint arXiv:1406.2015*.
- Veeramachaneni, K., O'Reilly, U. M., & Taylor, C. (2014). Towards feature engineering at scale for data from massive open online courses. *arXiv preprint arXiv:1407.5238*.
- Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2013). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop*.