Visualizing Data from Massive Online Open Courses

Katelin Bailey, Jialin Li, Naveen Sharma

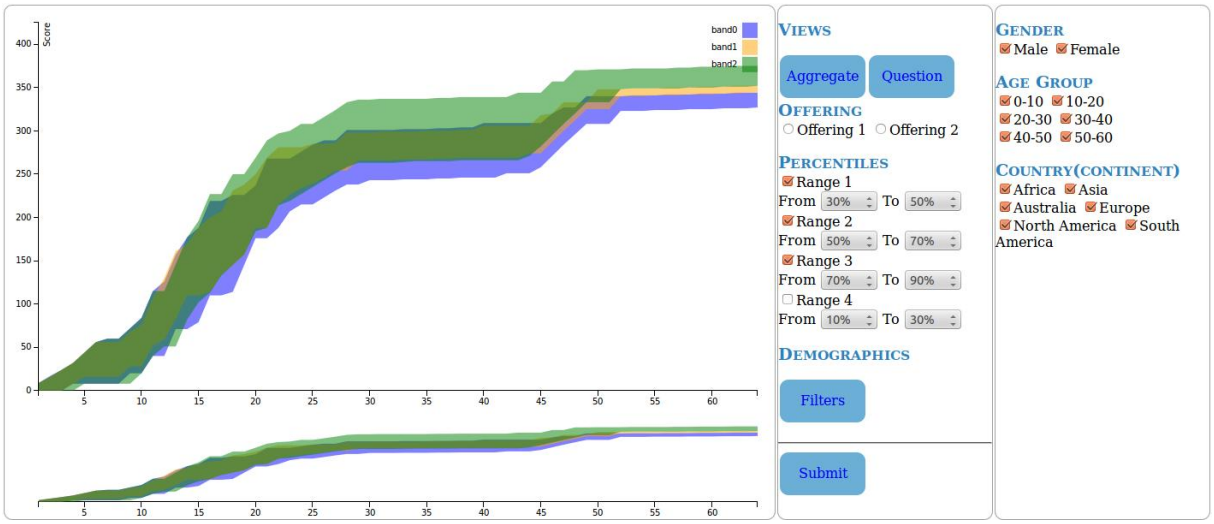

Fig. 1. Prototype: Initial storyboard image.

Abstract—This project addresses the massive amount of data available to instructors of MOOCs (massive open online courses). While some of the data is noise—students who never intend to participate, or fail to submit the majority of assignments —much of it is potentially valuable data on what methods and components in a course are effective. It is, however, massive in quantity. We intend to use the data readily available from Coursera to provide some exploratory visualizations for a generic MOOC class, generally tracking attrition and success rates.

More specifically, we will enable professors, who upload their own data, to add ress a variety of questions, including but not limited to the following: (1) Comparing two (or more) instances of the same course: was change X in assignment 3 effective? Were the overall statistics comparable? (2) Tracking characteristics throughout the course, based on intro demographic information. Do people who don't know recursion do significantly worse on this quiz? (3)racking the timeline of the course: when do people drop out? Can we tell why? For the purposes of this project, we will be prototyping from Dan Grossman's data, and working with him to determine desirable visualizations.

1 Introduction

Often, when course instructors seek information about their in-progress or completed courses, they turn to direct data manipulation in software such as Microsoft Excel, Numbers, or Google Docs [**?**, **?**, **?**]. However, with expansion of educational institutions into course styles such as Massive Open Online Courses (MOOCs), the straightforward manipulation of data to find outliers and students who need additional assistance becomes too overwhelming for the individual professor to manage in a reasonable time frame.

With courses that enroll tens of thousands of participants in a single offering of a single course, it is easy for instructors to be overwhelmed by the sheer amount of data gathered by such courses, both in terms of demographics as well as actual results on assignments. Sites such as Coursera [**?**] which facilitate the MOOC offerings through interfaces for posting videos as well as posting, submitting, and grading assignments provide limited interfaces for following a single student—or even a group of students—over the weeks of a course. Nor do they provide any satisfying visualization of the impact of the course elements, despite what is undoubtedly a surfeit of data.

This project comes in response to the questions of one such professor, who wanted to answer several questions about the programming languages class he had offered twice in the past. Starting with an exploration of the current Coursera data available and their visualizations, moving to working with data (modified for anonymization) generated by actual students, we sought to provide better visualization of the data with several main goals.

Accessibility By far the largest problem with the data in its initial form was accessibility and tractability for instructors: though they could get the data into .csv format, it was largely impractical to do calculations over the data without significant number churning. We provide visualizations with clustering and trends so that the data and emerging patterns can be more immediately accessible to the instructors.

Clustering and Comparison In addition to visualizing the data in entirety, we allow the instructors to select groups of students, through course offering, demographic information, or percentile groups, such that the trends between two different student groups can easily be visualized. Providing aggregate data over these groups removes the noise inherent in large datasets, and isolates the important trends.

Component Impact The last important feature we offer to instructors is the ability to isolate the impact of specific components of the course. Whether through the drop-off rate visualized on a time line or the exam questions viewed as a progression through course material, or a best-predictor type view for the final grades, we want

instructors to determine which components of their course have the largest and most accurate impact on student retention and success rate (as represented through grades).

With these goal in mind, we create a visual exploration tool for exploring MOOCs and their data. We go over the background and previous work in the next section, explaining the design and methodology in the following section, and concluding with results and feedback that lead to future work.

A note on student privacy. One of the most delicate aspects of a tool like this is that student data is inherently private. Though one professor did generously give us access to the data from two offerings of his course, much of the demographic data, for example, is not easily available from Coursera due to privacy concerns. Further, tracking students' performance of time may be identifiable in and of itself. For the purposes of this demonstration and writeup we have used the provided data that has been fully anonymized in terms of demographics and slightly tweaked, by the numbers, to ensure further privacy. We provide additional explanations of this in the methodology section.

2 Background

2.1 Coursera

2.2 D3

2.3 Related Research

We were unable to find previous work looking at courses on the scale of MOOC offerings.

Previous work seems isolated into a few different areas: analyzing student progress over years of standardized testing, and understanding the progress of students through a single course. The latter is the most similar to what we intend to provide, albeit with different approaches. For example, the work in [1] analyzes students and places them into states, creating a tree-like flow chart of where a student is likely to go next in the course. The overview of educational data mining (EDM) provided in [2] provides a very high-level description of where previous work has gone with respect to evaluating the effectiveness of courses, online material, advising, and so on. Much of this data is either not available to us or is outside the scope of our work, though the clustering techniques we implement are similar in nature to those discussed there. We intend to provide data specifically targeted at the world of MOOCs, where attrition is high, interaction with students is minimal, and the scale of data is too enormous to consider looking at some of the factors involved (like clicks on the course website). Moreover, we would like to provide an overview in a dashboard-like setting for the instructors of the course, to determine utility and effectiveness.

Outside the scope of EDM, work has gone on in several areas with respect to standardized ?high stakes? exams (e.g. state tests through elementary and high school). Bendinelli and Marder [3] model the data a flow problem, and provide some basis for analysis based on trends and demographic characteristics. However, the data in that paper is severely restricted in scope, and provides minimal analysis at tracking a single individual. Other work looks at the usage data, but not performance [4], and still others are restricted only to looking at progress over several courses [5].

3 Design and Methods

The data we chose to work with was largely data provided by default through Coursera. We have tables of data corresponding to assignments, each of which may be submitted multiple times by a single student. Each submission is tracked with a timestamp, submission number, and final score. For the midterm and final we have a single submission with timestamp and grade. This aggregate information is at the core of three of our views: **Aggregate View**, **Timescale View** and **Impact View**

In addition, for each exam we have a breakdown of questions, how many points each was worth, and what percentage of the students correctly answered the question. This information is used to generate **Exam View** and the comparisons for subviews.

Finally, we have demographic information for each student. Due to privacy concerns, we generated this data automatically, from the statistics and parameters provided to us based on the actual student data. This random demographic data (gender, age, continent/country and background) is sufficient for us to provide a prototype, though further work would require more accurate data. This data is summarized in the **Demographic View**

Though all pieces of data are maintained in separate .csv files, the (anonymized) student id allows tracking of student information from one piece of data to the next. Though 65,000 students enrolled in the course, fewer than 27,000 ever watched a video. In the end, we have fewer than 4,000 students who submitted assignments half that number actually completed the course with a passing grade.

3.1 Design

The core of the visualization provided are *views* which switch out the main *graph panel* to isolate different components of the course. To augment the abilities of the graph we provide *background computation* to isolate *clusters* of students and *impact* of components. As a fine-grained means of selection, we allow the instructor to superimpose *filters* over the data.

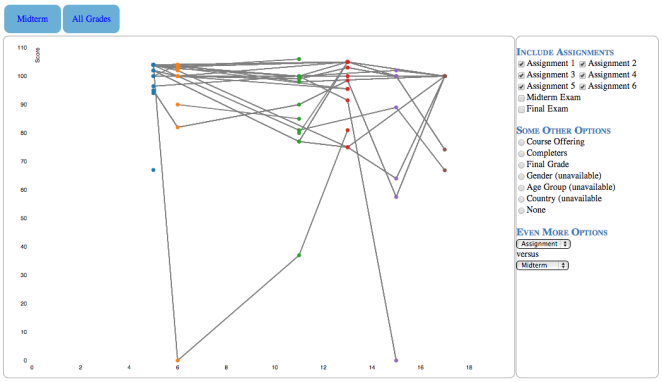

Fig. 2. Aggregate View.

Aggregate View (see Figure 2) This view provides a generalized timescale view of the course. Each assignment and exam is represented as a point on the x axis, with grades represented (in percent) on the y axis. This view highlights the retention rate of the class, particularly when clusters are enabled.

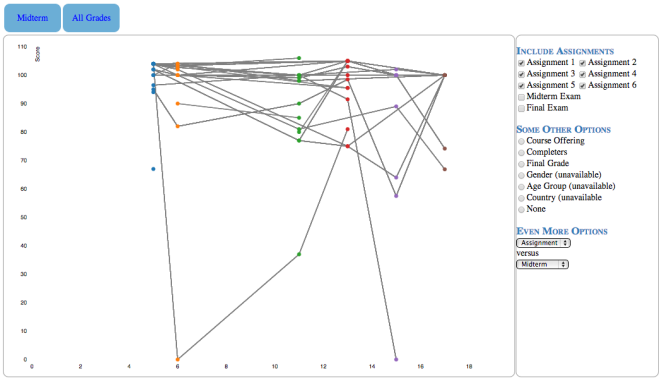

Fig. 3. Timescale View.

Timescale View (see Figure 3) This view provides a zoomed-in view of the assignments and exam scores in relation to the timestamp with which they were submitted. This view originated from a desire to answer the question *do late submitters overall score worse than early submitters?*

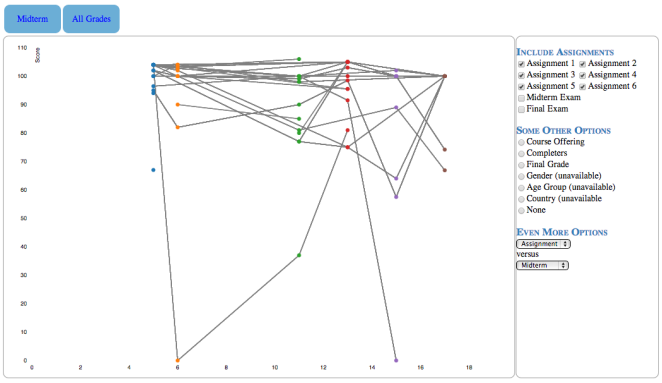

Fig. 4. Impact View.

Impact View (see Figure 4) This view provides a scatterplot reflecting the impact of an assignment or exam on the final grade. Each point is a single student, where the x value is the assignment grade and the y value is the final exam grade. In an ideal world, each assignment would be an equally good predictor of the final grade. When an assignment is a poor predictor, whether a student does well on that exam has little to no impact on the final grade, and we see a randomized scatter plot. When an assignment is a perfect predictor, we see a perfect diagonal formed by the scatter plot. Most assignments fall somewhere between the two.

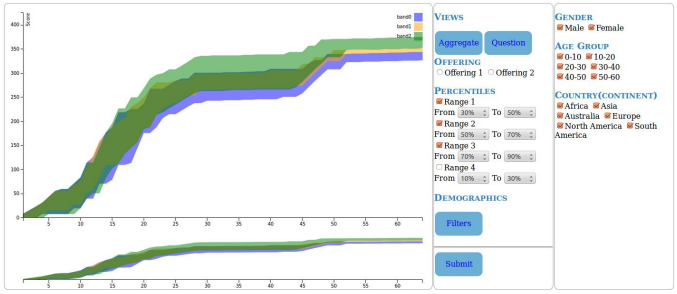

Fig. 5. Comparing Midterm Scores in Aggregate. The exam progresses to the right, and we see the fan-out of scores between percentiles as students diverge in point value.

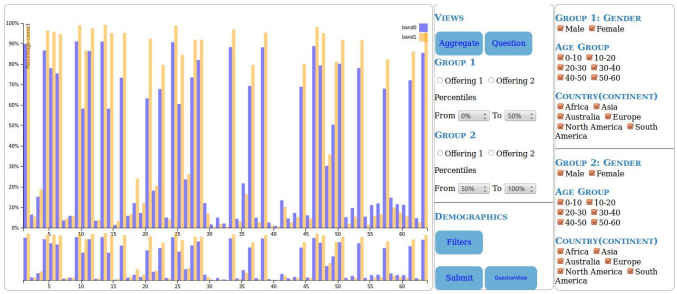

Fig. 6. Comparing Midterm Scores Across Offerings

Exam View (see Figure 5) TODO

Demographic View (see Figure 8) TODO

3.2 Background Computation

To save time in loading the actual visualization, some of the computation is done in the background.

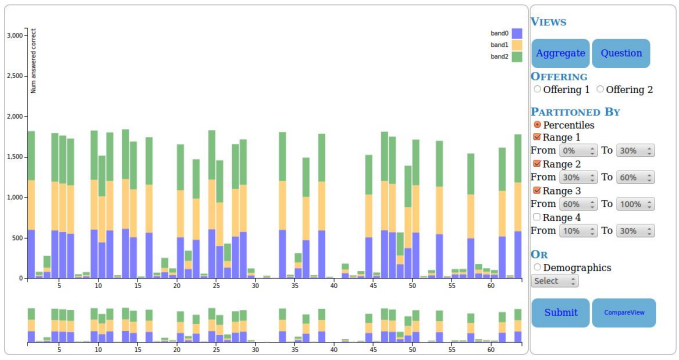

Fig. 7. Comparing Midterm Questions by Student Percentile.

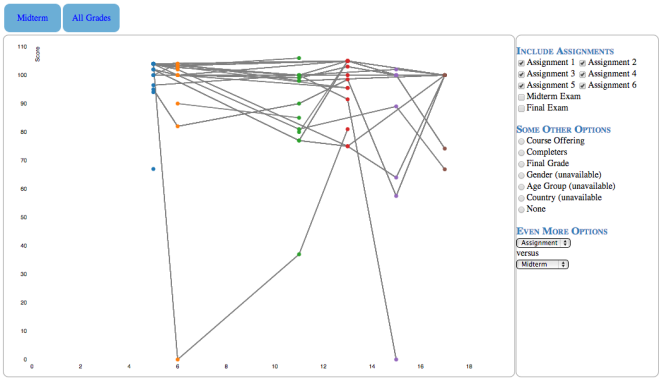

Fig. 8. Demographic View.

Clustering Clusters are precomputed with a script (which must be run prior to the site being launched) and stored in data files for loading when requested. Given the quantity of data being processed, this significantly improves the responsiveness for a given page. These clusters are computed where the grades of each student (for all time or an exam) are considered a vector, and kmeans clustering is applied over the vector, with the distance function being a simple function over each vector pair:

$$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2....(a_n-b_n)^2}$$

Impact The course element which most closely predicts the final grade is also precomputed and stored, using the same function above, this time with each assignment being a vector (of which each student is a dimension). However, for the actual display, though we default to the closest predictor (that with the closest distance and/or highest impact), each grade can be displayed without significant computation. Again, pre-computation of the distances saves significantly on load time.

4 Results

4.1 Performance

4.2 Insights and Best Practices

TODO: input instruction box text here:

5 Discussion

5.1 Design Insight

5.2 Instructor Feedback

We started the design with the goals of a single professor in mind. He was most interested in a few factors, which we sought to incorporate in our design:

1. Drop-off rate in student retention: how much is normal attrition and how much is caused by the course?

2. Cross-offering comparisons: was there significant difference between offerings?

3. Performance of female students in the course overall (programming languages is a male-dominated course, in large part. Back-of-the envelope calculations showed that his retention was worse for the female students: was there a reason why?

Though we targeted all three of these questions, we were most successful at answering the first two. Though we allow the instructors to explore trends within demographic groups, the depth of options in exploring the "why" of student retention coupled with demographic variables is not expressed. Further insight into this direction would be possible with stacked filters or pop-up boxes on hover that provide dominant information for a cluster or individual student highlighted. However, this is beyond the scope of the current implementation.

5.3 Student Feedback

Though we initially targeted our design at the instructors of MOOC courses, we found that students expressed an equal amount of interest in the data. Both our classmates and other graduate students who heard about the project were eager to discover whether similar visualizations, perhaps better aggregated, could be used to isolate the *difficult weeks* of a course, or which assignments and exam questions they were most likely to struggle on.

5.4 Additional Feedback

In presenting an early prototype of our project to a wider community base, via poster, we received plenty of additional feedback. Students and instructors were both intrigued by the retention rate and wanted to ask many more questions about the reasons students left. In addition, many professors were intrigued by the idea of isolating "problem" sections of their exams where they might improve, which drove the development of further exam views.

Our implementation of a timescale view was driven entirely by responses to the initial prototype, in which a professor confessed she'd always written down the times that assignments and exams were turned in, and had never bothered to plot them to find out if students turned in exams early because they were clueless or brilliant.

Additional views and data collection schemes were encouraged by nearly everyone, showing that the quantity and diversity of questions that can be asked in this arena is quite startling.

Finally, several people at the university level have expressed interest in incorporating a system such as this into their annual course analytics, incorporating student evaluations of the course and professor's teaching into the mix, but analyzing pure classroom data.

6 Future Work

Though we feel this project represents a solid prototype on the available data, there is a significant amount of work yet to be done in this area. We describe some of the improvements and extensions we would like to make below.

A Wealth of Information The amount of data available from the Coursera platform is mind-boggling. Though we used a simple set of grades, timestamps, and demographic data to be able to create this prototype, there is additional data that could provide more insight. Among other things, the use of forum participation, video viewings, and the optional in-video quizzes would provide a rich set of information from which to gauge engagement (frequency of participation) and understanding (frequency of questions and re-watched videos). Other pieces of data such as clicks, and total time to take an exam would be harder to gather, but perhaps more enlightening.

In addition to supplementing our data with additional information from a single offering or course, we would like to be able to add data from across multiple offerings or multiple courses. With appropriate learning algorithms, we might begin to answer the questions of whether exams are actually helpful, if certain length videos or courses encourage retention and so on. MOOCs are, after all, a large source of rich data to compute over, and Big Data mechanisms applied would provide insights into the utility of certain components.

Automation and Accessibility In addition to providing more data and insights to instructors, we would like to be able to provide easier access. Ideally, we would integrate our tool directly into platforms such as Coursera, though doing so would be a negotiation between development teams. If not that, than at least having a generic auto-upload feature that allows an instructor to upload .csv files and designate columns as demographic, assignment, or exam data would be useful in making the tool more widely available to instructors. As is, the tool is a bit too brittle to enable such usage.

Alternate Viewpoints Last of our long-term goals is to look at data sources other than MOOCs. Both students and professors had many suggestions for other uses of the data and resultant visualizations, both for classical classes in a classroom, and for exposing the difficult portions of a course. Exploring these options would provide richer platform.

7 Conclusion

In conclusion, we present a tool for exploring and visualizing the course data for performance and retention of students in MOOCs. Though the end results poses as many questions as it answers, we believe it is a useful and useable tool for beginning analysis of impact and improvement in course offerings.

Acknowledgments

The authors wish to heartily thank Dan Grossman, who provided an intro to the Coursera system, as well as a wealth of data to use in our prototype. We also thank the many students and professors who offered input during the development process.

References