A Data Analysis of Student Success and Motivations in the BJCx MOOC

Yifat Amir
Dan Garcia, Ed.

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2018-75
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-75.html

May 18, 2018

Acknowledgement

Thank you to Professor Armando Fox, Michael Ball, and Peter Sujan for helping me get access to the data. Thank you to Alvin Ghouas and Rita Guo for tips on data visualizations. Thank you to Aashna Mukerji for proofreading. Thank you to my dad, Arnon Amir, for talking through my ideas with me and giving me feedback on my writing. And a big thank-you to Professor Dan Garcia for the inspiration, support, and encouragement.

A Data Analysis of Student Success and Motivations in the BJCx MOOC

By

Yifat Amir

A technical report submitted in partial satisfaction of the
requirements for the degree of
Master of Science

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Teaching Professor Daniel Garcia (EECS), Chair
Assistant Professor Zachary Pardos (INFO/EDUC)

Spring 2018

The technical report of Yifat Amir, titled A Data Analysis of Student Success and Motivations in the BJCx MOOC, is approved:

Chair ____ Date ___________________

____ Date ___________________

University of California, Berkeley

Table of Contents

My involvement with The Beauty and Joy of Computing (BJC) began when I took CS10 at UC Berkeley in my second semester of undergrad. It was a transformational experience which eventually led to me switching into the Computer Science major (and now even completing a Master's degree in the field). Following my experience as a student in the course, I volunteered as a lab assistant and later went on to grade and TA and even become course instructor. Meanwhile, I also got involved in the development of BJCx and in the creation of the Snap! autograder. For almost a year, I was simultaneously on the course staff for CS10 and on the course staff for BJCx, so one can only imagine my passion for BJC.

Dedication

This work is dedicated to all those who self-select out of pursuing computer science because they feel like they would not fit in, as well as to those who were never given the exposure or the chance to try because of a society which had already made the selection for them.

A Poem

Producers[1]
By Yifat Amir

As the future comes closer
and shapeshifts to the present
each member of society, a composer
becomes responsible for the content

that we hear, that we read
that we view on the newsfeed
that we ingest and digest
during dinner, lunch, and breakfast.

Though the prevalence of all this technology
has its pros and its cons—it undeniably,
has an impact on our ideology,
our sociology, and our biology.

So as citizens of the community
we should care, care to learn
what is computing, and what does it do?
so that we're not only consuming it
but producing it, too.

[1] This poem was originally written for an extra credit assignment in CS10, Spring 2014.

Abstract

Computing is evolving more and more into a form of literacy, a powerful tool for effecting change. For this reason, it is important that all people have access to high quality, introductory computer science education. The Beauty and Joy of Computing (BJC) is an introductory computer science curriculum designed for students of all academic backgrounds and aimed at broadening participation in computing, especially for women and other underrepresented minorities. In addition to the traditional classroom setting, this curriculum has been offered as a free, publicly accessible, massive open online course (MOOC) called BJCx. In its first few years, BJCx has attracted tens of thousands of students from around the globe. However, like most other MOOCs, it has suffered from very low completion rates. Nonetheless, measuring student success in MOOCs cannot and should not be constrained purely to grades, since students sign up for online courses with a variety of motivations. For example, those who enroll purely to browse through content may very well achieve their goals without earning a passing score. In order to more accurately quantify success in an online course, student performance, behaviors, and satisfaction must be evaluated in proper context.

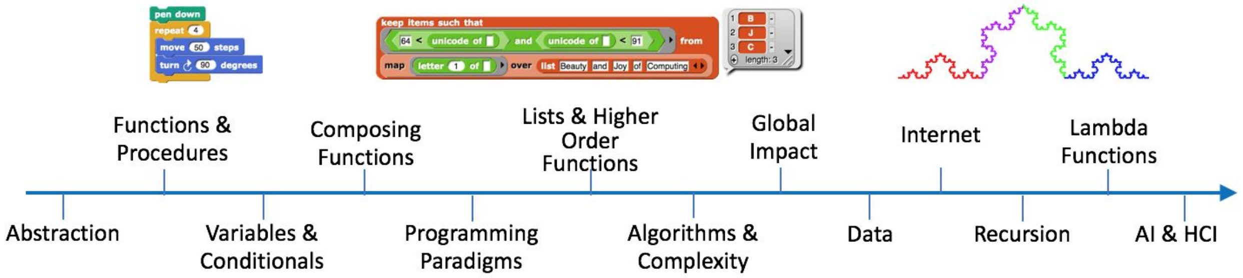
This technical report contains a detailed analysis of student motivations and success in the BJCx online course. Performance is quantified in the context of demographics, intentions, and motivations. Furthermore, a novel classification algorithm for categorizing students' written goals by their overarching motivations is proposed. It utilizes natural language processing (NLP) methods such as the Topic Model and distributed vector embeddings for words. The results of its application to the BJCx dataset are used for further analysis and contextualization of student performance. Results show that performance and engagement with course material vary with motivations as well as with demographics. These results, along with the goal classifier, could be used in the future for personalization of the online learning experience to students' goals or for the implementation of an automated intervention system to reduce student attrition. The analysis in this report is applied to BJCx, but the classification algorithm, methods of analysis, and many of the results can be generalized to other introductory computer science courses and to MOOCs in general.

Introduction

I. About BJC

The Beauty and Joy of Computing (BJC) is an introductory computer science course aimed at broadening participation in computing to all, especially underrepresented communities [1]. The curriculum approaches computer science from multiple perspectives, covering everything from computational problem solving to project design to social implications of technology. It has been offered as a undergraduate course (CS10 in Berkeley), a nationally-recognized high school AP course (AP CS Principles), and a massive open online course (MOOC) called BJCx. The material is welcoming to students of all ages and assumes no background in computer science. It is taught in a visual blocks-based programming language called Snap*!*, which allows beginners to focus on algorithmic design without getting stuck on syntax. For many students who otherwise would never have imagined studying computer science, anecdotally, the BJC curriculum is a gateway to pursuing a career in computing. The curriculum is summarized in the timeline diagram below.

A Summary of the BJC Curriculum



II. About BJCx

BJCx is one of BJC's initiatives to bring computer science education to all. It is a free, publicly-accessible course hosted on edX and spans about 32 weeks (approximately the length of a school year). It was first launched in August 2015, when it was split into four approximately 8-week-long "MOOClets": BJC.1x, BJC.2x, BJC.3x, and BJC.4x[2]. The full course names were:

- BJC.1x: Starting to Think Like a Computer Scientist and Develop Complex Programs
- BJC.2x: Lists, Algorithms, and Complexity
- BJC.3x: Data, Information, and the Internet
- BJC.4x: Recursion and Higher-Order Functions

Each MOOClet was administered as its own edX course, though the four were marketed as a sequence. Of the 19,168 students enrolled in BJC.1x, 1080 also enrolled in BJC.2x. Similarly, 416 from BJC.2x enrolled in BJC.3x and 253 from BJC.3x enrolled in BJC.4x.

Each course was broken down into weekly segments. A typical week would have 1-2 reading assignments with a reading quiz, discussion forum participation requirements, 3-4 lecture videos and quizzes, and lab exercises. The first three courses were designed with 5 weeks of curriculum and 3 weeks of project work, while the fourth and final course included only the 5 weeks of curriculum. Homework assignments whose scale was larger than lab exercises but smaller than projects were sprinkled throughout. Each course culminated in a multiple-choice final exam administered on edX. Students were expected to commit about 4-5 hours a week to the course and the threshold for passing was set at 75%.

Each of the projects that concluded a course was designed to be just open-ended enough to allow students to create something they were passionate about and do it from scratch. This is inspired by one of the main values of BJC: creativity. One of the goals of the course was for students to see the creativity in programming, to realize the beauty (and joy) of computer science. However, grading open-ended projects at MOOC scale is not a trivial

[2] In the 2016-2017 and 2017-2018 offerings, course administrators merged the first two and the last two segments, offering the curriculum in two 16-week-long MOOClets.

task. For this reason, students were assigned to peer-grade each other's work as a part of their project grades. They used a site called PeerStudio, which mediated the peer grading process.

But the projects were not the only opportunities for feedback; students also completed programming lab exercises and homework assignments. Feedback from these formative assessments was an integral component of the learning process. To support the scaling of BJC via MOOC, a team of mostly undergraduate UC Berkeley students implemented autograding capabilities for Snap*!* so that students could get feedback in real-time about their code. This would relieve the need for staff support to scale with the size of the class. Due to the technical challenges of grading code written in a blocks-based programming language, almost two years of development were put into building a suite of autograding tools for Snap*!* and creating autograders with specific feedback for all of the BJCx lab exercises and assignments. Once BJCx was launched, these autograders would facilitate code submissions, evaluate correctness, and give students detailed feedback on their work. This was integral to the success of the MOOC.

Related Work

In a time when internet access is largely commonplace, MOOCs represent the democratization of access to higher education. People who otherwise cannot access quality instruction due to financial burden, time commitment, or physical location can now learn virtually anything online in courses taught by professors from UC Berkeley, MIT, Harvard, and other world-renowned institutions. As Geoffrey Crowther, founder of the Open University in the UK, declared in the university's inaugural speech in 1969, "wherever there is an unprovided need for higher education... there is our constituency" [2]. With this spirit, the number of MOOCs and platforms which host them has skyrocketed over the recent decades. MOOCs have created a marketplace for a global audience of learners to come together, share ideas, and learn from the best. Unlike traditional classroom settings, MOOCs can scale to hundreds of thousands of students under the instruction of just one teacher.

Nonetheless, it is not a given that the students served by these online resources are those who need it most. In fact, some research has shown that MOOCs in general are being used primarily as enrichment for already well-educated individuals rather than as a gateway for socioeconomic mobility for disadvantaged or otherwise struggling students [3][4]. Other research has shown that the vast majority of those who succeed in these learning environments are those who are self-motivated and are looking to satisfy personal curiosities or advance in their career [5]. Although noble in their pursuits, these people are not the learners which MOOCs were originally intended to serve.

With the surge in popularity of online learning has come a great deal of research centered on data generated by MOOCs. Researchers have investigated learning behaviors, factors that lead to dropout, content personalization methods, the sociology of online social interaction, and much more. It has become known that students enroll in these courses for a variety of reasons and that their engagement patterns with the course content can take on many forms, depending on what they are seeking to gain from the course. Some work has found that students who enroll in order to earn a certificate of completion generally complete only as much of the course as is needed to pass, completely skipping even a fourth of the course content [6]. The same research also discovered that those students tend to engage in non-linear navigation of the course, jumping from homeworks to lectures and back more often than other students. This behavior does not necessarily lead to the best learning outcomes.

Other research has discovered latent categories of students by clustering their online behavioral patterns. One such clustering of engagement trajectories, which resulted in the four subpopulations Auditing, Completing, Disengaging, and Sampling, was analyzed along with student satisfaction and overall performance to find that Auditing learners and Completing learners usually reported similarly high levels of satisfaction at the end of the course, illustrating the need to define student success in MOOCs in context with student intentions [7]. Furthermore, the methodology proposed in that work could potentially be used to automatically detect disengagement before students drop out, providing an opportunity for intervention. One study has placed students' motivations in contrast with each other based on their status as a full-time student or a professional [8]. Another study, which was based on a series of interviews with MOOC users, has grouped students' motivations into four broad types: "fulfilling current needs," "preparing for the future," "satisfying curiosity," and "connecting with people" [9]. Similarly to other studies, this work also found that some students were satisfied with their learning even when they did not complete the course.

Although auditing students are an important subpopulation of learners, the majority of students who do not receive passing grades in the typical MOOC are students who completely stop engaging midway through the course. Many researchers have looked into methods of early automatic detection of attrition with hopes that these methods could be used to support preventative interventions at MOOC scale. One such work compared a variety of predictive models trained on time series clickstream data and found that the best model was a recurrent neural network (RNN) with long short-term memory (LSTM), which predicts future behavior as a function of recent behavior as well as historical behavior [10]. Other research has found that combining students' online behavior data with demographic information, which includes data about prior experience, can improve dropout prediction accuracy [11]. But clickstream data is not the only data which can be used to model students' behavior. One study used data from a course discussion forum in conjunction with social network theory to model social positioning and connectedness among students [12]. The results showed that social engagement which promotes commitment also prevents attrition. This insight may be useful for the future design of MOOCs. Another work that focused on discussion forum data rather than clickstream data found that a sentiment analysis of students' contributions in the forum helps to predict dropout [13]. These are just a few of many recent attempts to automatically detect disengagement.

In addition to preventing attrition, there has been recent interest in understanding how students' motivations for enrolling in MOOCs correlate with their success. Some studies have found that students whose motivations focus on the learning experience rather than on the course content, including motivations such as curiosity about online learning or desire for course credit from a prestigious university, are less likely to complete the course [14]. Others have found that students' motivations strongly predict their levels of participation, but in a manner that is dependent on the purpose of the course [15]. For example, motivations related to professional development were strong predictors of performance in an online course focused on career-based skills but not in a general interest online course. This suggests that passing rates and overall student performance depend on the alignment of a course's purpose with its perceived purpose.

In summary, the success of MOOCs cannot be accurately measured by a catch-all metric. Low rates of completion fail to take into account the success of auditing students; more generally, final grades do not necessarily reflect how students' cumulative learning experiences compare to their intentions. Nonetheless, predicting attrition and constructing personalized interventions to help struggling students stay engaged has the potential to impact thousands of people when implemented at MOOC scale. In the end, course administrators and MOOC researchers alike must not get lost in the numbers—they must be conscious of who it is they are working to serve and be intentional about designing their courses to encourage the success of those students.

Data

I. Overview

The data used in this analysis comes from multiple sites with which students interacted in the four MOOClets in the 2015-2016 school year: BJC.1x, BJC.2x, BJC.3x, and BJC.4x. Data about students and their participation has been sourced from edX for the four courses. In addition, students were encouraged but not required to register for BJCx's page on Piazza, a Q&A forum independent of edX. Aggregate statistics about each student's engagement in Piazza are also used in this report. Finally, students were asked to respond to surveys hosted on Qualtrics, which is also independent of edX. The data from these surveys includes students' responses to both closed-form and open-ended questions. The data used in this report is an aggregation and compilation of data sourced from edX, Piazza, and Qualtrics.

II. edX

Data from edX can be sorted into two tables: 1) information about each student and 2) information about each student's interaction with each page of the course (though not at clickstream granularity). Much of the data in the first table is optionally self-reported by the students, which leads to both an abundance of missing values as well as questionable integrity of the data. It is important to note that the analysis contained in this report is subject to any biases that may have skewed the reported data.

Below is a summary of some of the important fields contained in each of the two tables along with their respective descriptions and availability percentages. The availabilities provided represent the percentage of values which are not missing among all students who enrolled in BJC.1x, the very first iteration of the BJCx MOOC. In general, students who engaged with the course provided more information about themselves than those who enrolled but never "showed up," so the availability statistics provided serve as a lower bound for the availability of groups of students to be analyzed in this report.

edX Table 1: Data about each student

Field	**Example value**	**Availability (%) in BJC.1x**	**Description** (where appropriate)
student_id	12345678	100	A unique identifier for each student. Identifies them in all edX courses in which they enroll.
name	Joanne Smith	100	
email	jsmith_BJCx@gmail.com	100	
year_of_ birth	2000	84.16	
bio	I am an inquisitive soul!	2.71	A brief autobiography.
country	USA	60.23	
gender	Female	85.56	Either "Female", "Male", or "Other".
goals	To improve my analytical and problem-solving skills.	46.48	The student's motivations for enrolling in this course.
level_of_ education	Middle school	85.48	Highest level of education completed.
location	Berkeley, CA	3.08	Plain text input
grade[3]	0.85	80.51	Final cumulative grade in the course.
mode	honor	80.51	Either "honor", "verified", or "audit".

[3] Grade data is not available for students who enrolled in the MOOC after the end of its first live offering (Dec. 2015).

edX Table 2: Data about each student's interaction with each course page

Field	**Example value**	**Description** (where appropriate)
`student_id`	12345678	A unique identifier for each student which identifies them in all edX courses in which they enroll.
`module_type`	chapter	Either "chapter", "course", "sequential", "problem", "video", or "vertical".
`module_id`	i4x://BerkeleyX/BJC.1x/chapter/0da46fd9b51c4398ad1a3b968ce711c6	A unique identifier for the course page.
`state`	{correctness: "correct"}	A dictionary of values relevant to the particular page. Values may include correctness, number of attempts, hints, video pause markers, or other metadata.
`grade`	1.0	The number of points received by the student for their work on the given page, if it is graded.
`modified`	2015-10-03 20:41:33	Date on which the student last accessed the page.

III. Piazza

The data sourced from Piazza includes aggregate counts for each student's engagement with the Piazza forum. The fields are summarized below:

Piazza Table: Data about each student's forum engagement

Field	**Example value**	**Description** (if necessary)
`name`	Joanne Smith	
`email`	jsmith_BJCx@gmail.com	
`days online`	51	Number of days on which the student visited the course's Piazza page.
`views`	130	Number of posts which the student viewed.
`contributions`	6	Number of questions, notes, answers, and follow-up replies the student made.
`questions`	4	Number of questions the student posted.

IV. Surveys

The final source of data is from surveys created by course staff and hosted on Qualtrics. Although completing the surveys was optional, students were encouraged to submit them through the incentive of small amounts of course credit. Each MOOClet contained 3 surveys: "beginning", "middle", and "end" administered in the zeroth, fifth, and seventh (final) week of the course, respectively.

The work in this project concerning survey data primarily focuses on the data from the three surveys administered in BJC.1x, since they contain the most responses.The "beginning" survey asked students for demographic/background data such as age, gender, ethnicity, location of residence, educational background, employment status, and level of experience in programming and math. Students were also asked about their past experience in MOOCs and their intended participation in this course. All of these questions in the "beginning" survey were asked in multiple choice or select-all-that-apply format. The "middle" survey asked students to rate how helpful each component of the course was for them, including labs, discussions (on Piazza), videos, readings, and Piazza. It then asked students to rate the difficulty and pacing of the material thus far and asked how many hours per week students were spending on the course. These questions were all in multiple choice format. Additionally, in an open-response format, the survey asked students to describe the most and least useful things they had learned so far and elicited suggestions for improving the course. Finally, the "end" survey asked students to rate their level of agreement with the following statements:

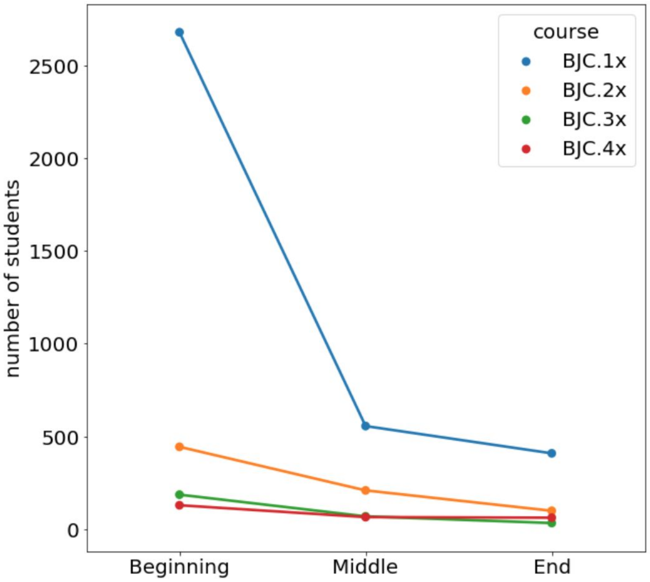
- "After reaching the end of the course, I was satisfied with my level of participation."
- "After finishing this course, I feel that I have gained a better understanding of computing."
- "I am content with the finished product of my final project."
- "There was sufficient support for me to effectively participate in the course."
- "After finishing this course, I want to recommend this course to a friend."

Each of these statements was followed by five multiple choice answer options: "Strongly Disagree," "Disagree," "Neutral," "Agree," and "Strongly Agree," as well as an open-ended question prompting "Why or why not?". Lastly, the survey concluded with asking students how many hours a week they spent on the course and eliciting suggestions for course improvement. The number of responses to each of the 12 surveys (3 in each

MOOClet) administered during the 2015-2016 school year is shown in the table and figure below.

Number of Student Responses to Surveys

Course	"Beginning" Survey	"Middle" Survey	"End" Survey
BJC.1x	2,680	556	409
BJC.2x	444	209	99
BJC.3x	186	69	33
BJC.4x	129	65	62



V. Data Availability

Due to the voluntary nature of enrolling in Piazza, completing the surveys, and providing personal information to edX, only select portions of students are represented by the analysis in this report. The table below details the number of students for which data is available for each data source and for each of the four MOOClets.

Number of Available Data Points

Course	Enrolled on edX	"Showed up" for the course[4]	Submitted the "beginning" Qualtrics survey	Enrolled in the Piazza page[5]
BJC.1x	19,168	9,923	2,680	1,567
BJC.2x	1,540	1,493	444	197
BJC.3x	909	876	186	77
BJC.4x	705	694	129	47

VI. Preprocessing/Cleaning

The data from all of these sources was pooled together and combined using student IDs and email addresses as student identifiers[6]. Rows and columns which were missing critical information, such as a student identifier, or which did not contain more than one unique value were dropped from the dataset. Data which overlapped across sources, such as highest level of education completed, was merged and used to fill in as many missing values as possible. A few manually encountered self-reported data which were obviously dishonest were put aside.[7] Furthermore, the following new variables were created:

- `passed` = True if `grade >= 0.75`, False otherwise.
- `age` = `2015 - year_of_birth`
- `recommend =` a discrete numerical variable representing students' level of agreement with the "end" survey statement "After finishing this course, I want to recommend this course to a friend."[8]

[4] This is (liberally) defined as having interacted with more than just the course welcome page on edX.

[5] Includes only students who enrolled in the Piazza page using the same email address as used in edX so that the data may be linked together across sources.

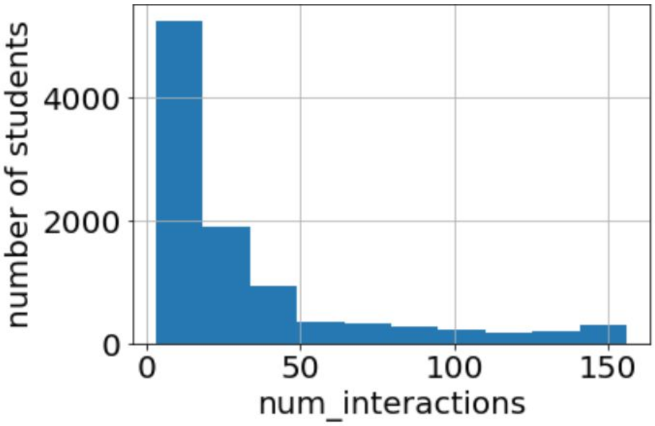
[6] Students whose email address on Piazza did not match their email address on edX were not linked with their Piazza data.

[7] E.g. one student self-identified their gender as "Chair," which was discarded yet valued as comic relief.

[8] Values were generated using the following mapping: {"Strongly Agree":1.0, "Agree":0.5, "Neutral":0.0, "Disagree": -0.5, "Strongly Disagree":-1.0}.

- `satisfied =` a discrete numerical variable representing students' level of agreement with the "end" survey statement "After reaching the end of the course, I was satisfied with my level of participation."[9]
- `num_interactions` = The number of pages of the course on edX that the student visited. For BJC.1x, the histogram of values for this variable is shown below, excluding `num_interactions` values below 2 (corresponding to students who only viewed the course welcome page).

Histogram of `num_interactions` in BJC.1x



The remainder of the report will often refer to these generated variables as well as to those which were part of the original dataset.

[9] Values were generated using the same mapping as for the `recommend` variable.

Chapter 1: How Demographics Correlate with Performance

I. Overview

In this section, distributions of student demographics and correlations between demographics and performance are visualized. These visualizations are meant to answer questions such as:

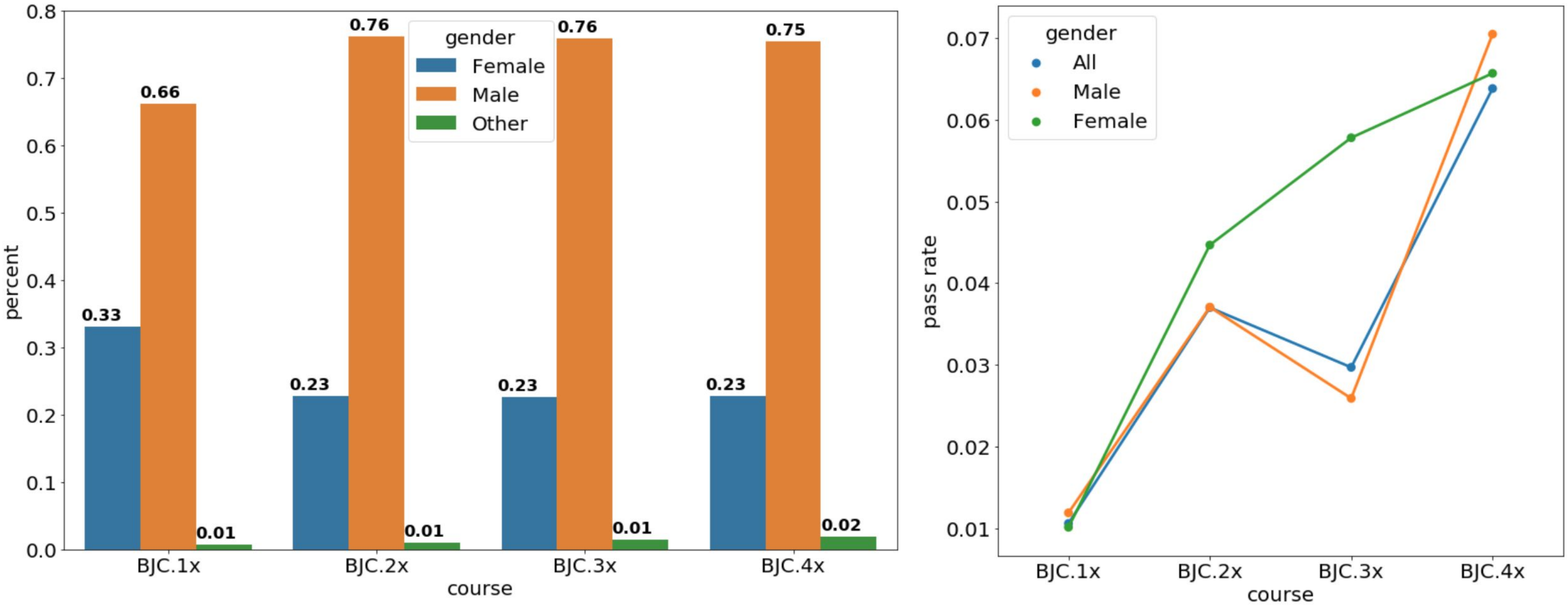
- Who is enrolling in BJCx?
- Who is receiving passing grades?
- How do gender, education level, employment status, and age correlate with the course?
- How does discussion forum participation correlate with passing?
- What kind of students engage with the course content but still do not pass?
- Who is this course serving most?

II. Gender

Diversity being one of the missions and values of BJC, it is interesting to investigate whether the gender makeup of BJCx reflects progress in the initiative towards 50/50. Compared to traditional computer science curricula, BJC has historically attracted more women and racial minorities due to its emphasis on collaboration and creativity, which stands in contrast to the programming-only nature of most introductory computer science courses. In the Spring 2018 semester, over 60% (record high!) of the 162 students enrolled in CS10 at UC Berkeley were women. In 2017, just a year after the launch of the AP CS Principles exam, the number of female, Hispanic/Latino, Black/African American, and rural students who took AP computer science courses more than doubled in comparison to 2016 numbers [16]. Taking BJCx was one way students across the US and elsewhere could access the BJC curriculum and prepare for the AP exam.

The figure below on the left shows proportions of students enrolled in each of the four MOOClets, labeled by gender. In BJC.1x, one-third of students were female. This is higher than the national average percentage of female students in AP CS Principles courses, which was 27% in 2017 [17]. However, it seems that proportionally, fewer women continued on to the second segment of the course. The chart shows a drop from 33% to 23% female between the first and second MOOClets, and a constant 23% female thereafter. The causes of this drop are yet to be determined.

Enrollment (left) and Passing Rates (right) by Gender across BJCx MOOClets

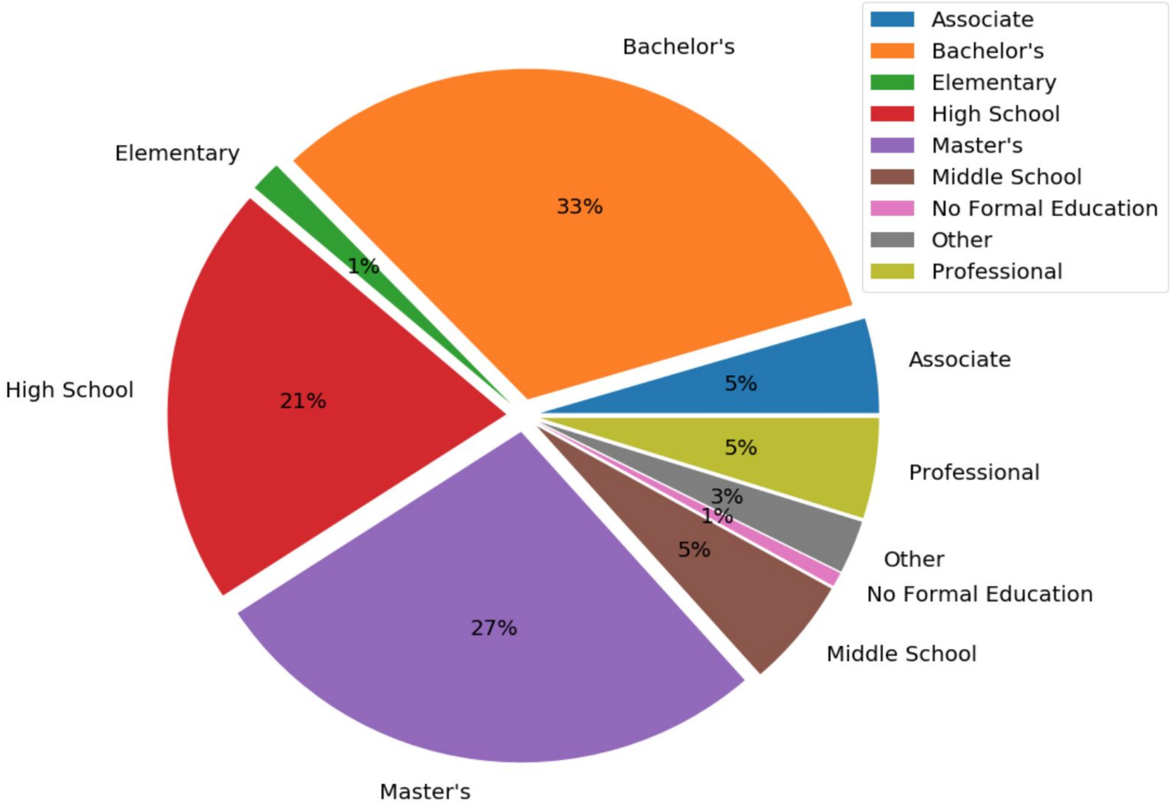


To add to the picture, the plot above on the right shows the passing rates of students broken down by gender. Although both genders generally exhibited about the same performance (note the magnitudes of the passing rates and their differences are small), which is expected, only the female students' passing rate seems to be upwards trending consistently across the MOOClets. This suggests that those women who did enroll for the later courses were more likely to also pass those courses.

III. Level of Education

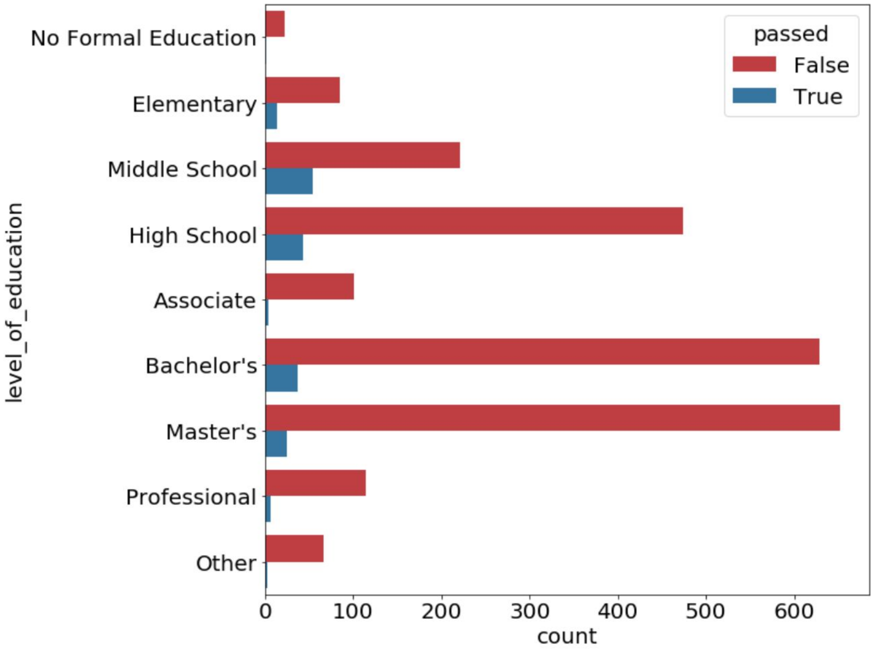
Although students of all ages and backgrounds are welcomed and encouraged to take the course, BJCx is targeted toward high school students. The pie chart below shows the enrollment breakdown for BJC.1x by *highest level of education completed*. Although the course was built for high school students, it shows significant enrollment by people with not only college but also graduate degrees.

Highest Level of Education Completed by Students Enrolled in BJC.1x



It is also valuable to see how educational background correlates with passing the course. The histogram below shows the number of students who passed and who failed the BJC.1x MOOClet, split by education level and ignoring students with a final grade of zero.

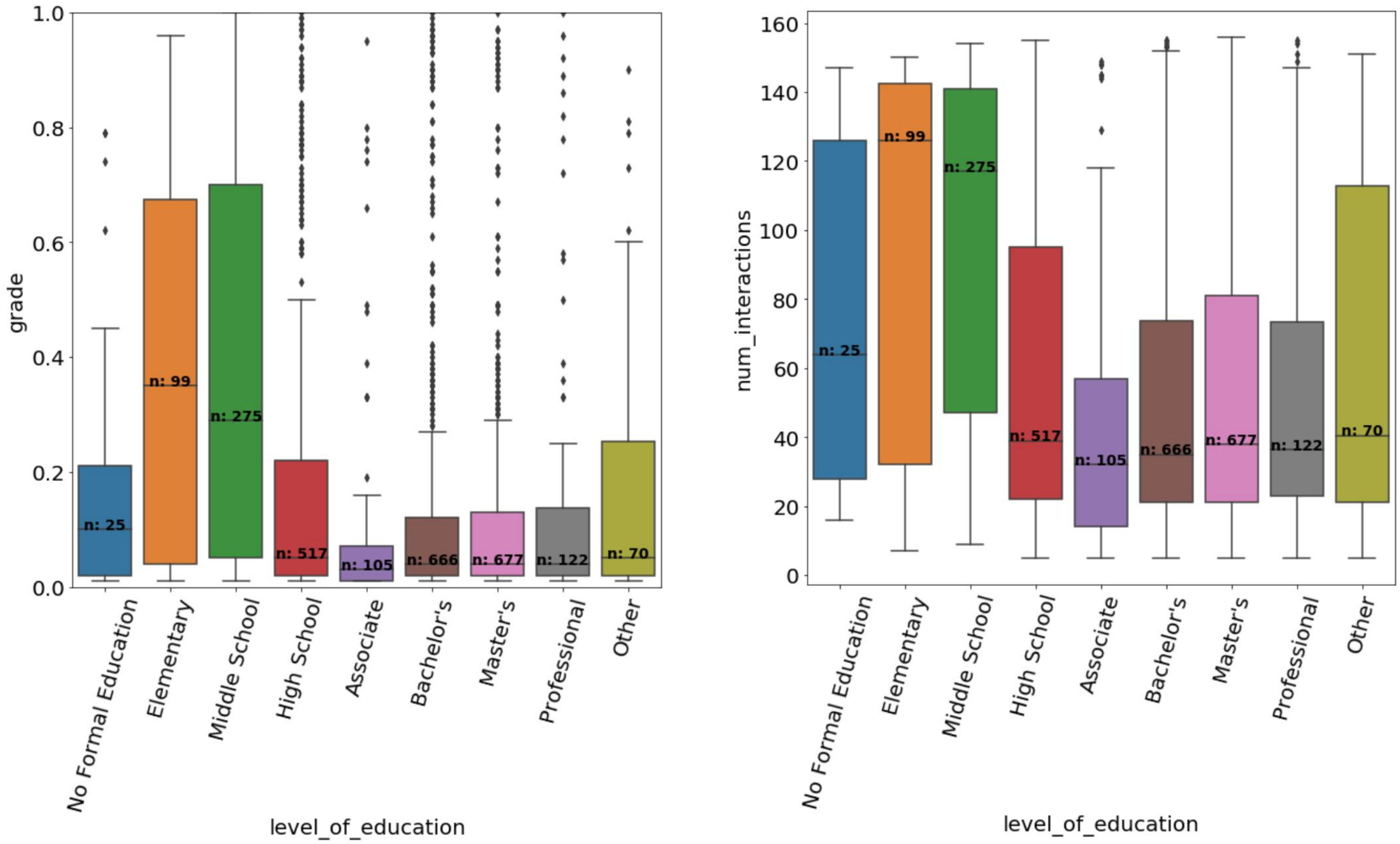
Number of Students Who Passed/Failed BJC.1x by Education Level



An eye-balled comparison of bar lengths lends itself to the understanding that students with educational backgrounds of only elementary, middle, or high school had the highest passing rates among all groups. So even though their raw enrollment numbers were lower than that of college-educated people, they seemed to be more likely to pass the course. This result supports the mission of targeting high school students with this introductory curriculum.

Next, it is interesting to see if there are any trends across student subpopulations which show contrasting levels of engagement and grades. The box-and-whisker plots below show grade distribution (left) and number of course pages visited (right) across students in BJC.1x with different levels of education, omitting students with final grades of zero. The plot on the left confirms the aforementioned result that those with primary or secondary schooling but not higher degrees performed best, contrary to the results of much past research which has found that more educated students do best. In fact, the median grades for those who have only completed elementary or middle school are higher than the 75th percentile grades for all other groups. Furthermore, it seems that the distributions of engagement with course content ordinally mirror that of grades (to be expected), but the mapping is nonlinear. A comparison of the two charts may suggest that students with no formal education were relatively active with the course content, yet proportionally did not score as high of grades.

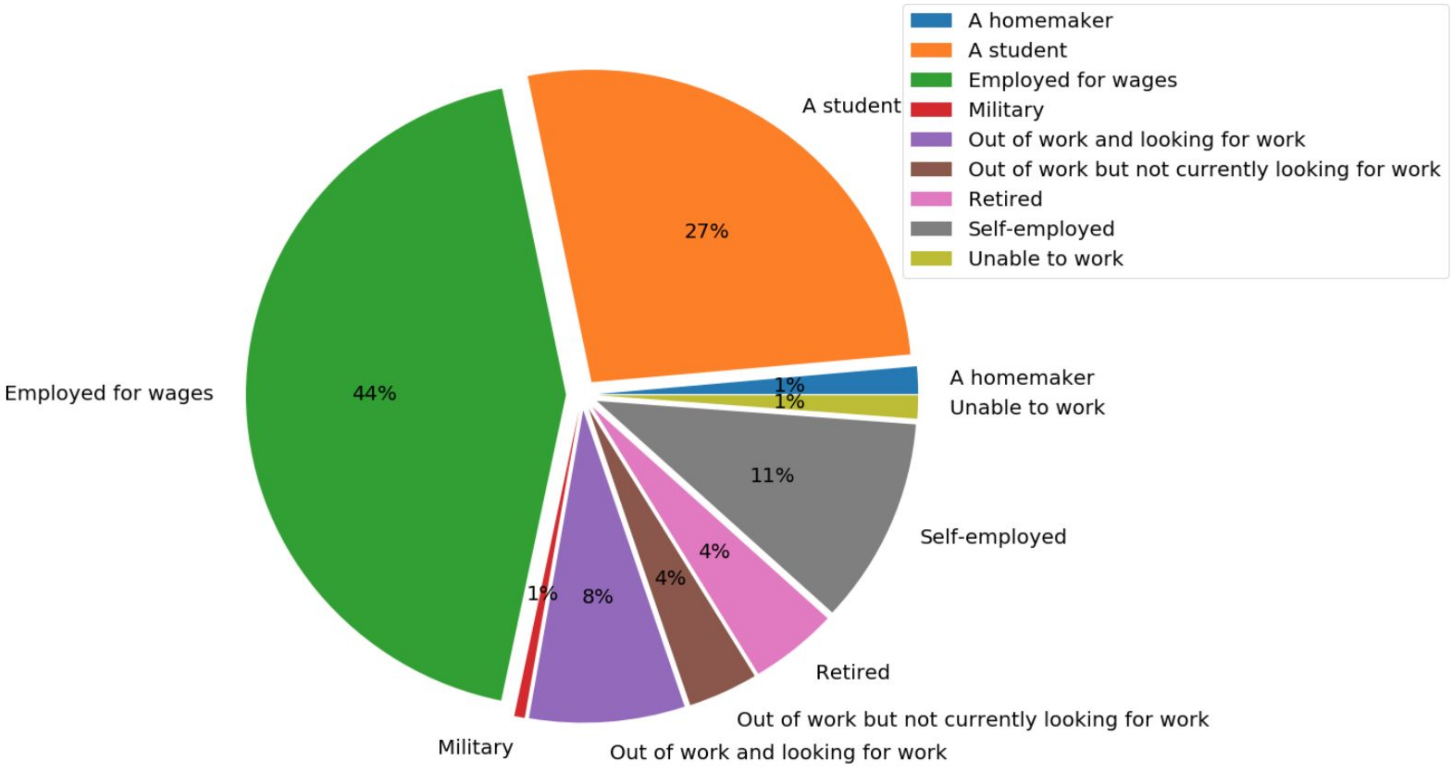
Grade (left) and Interactivity (right) Distributions by Education Level in BJC.1x



IV. Employment Status

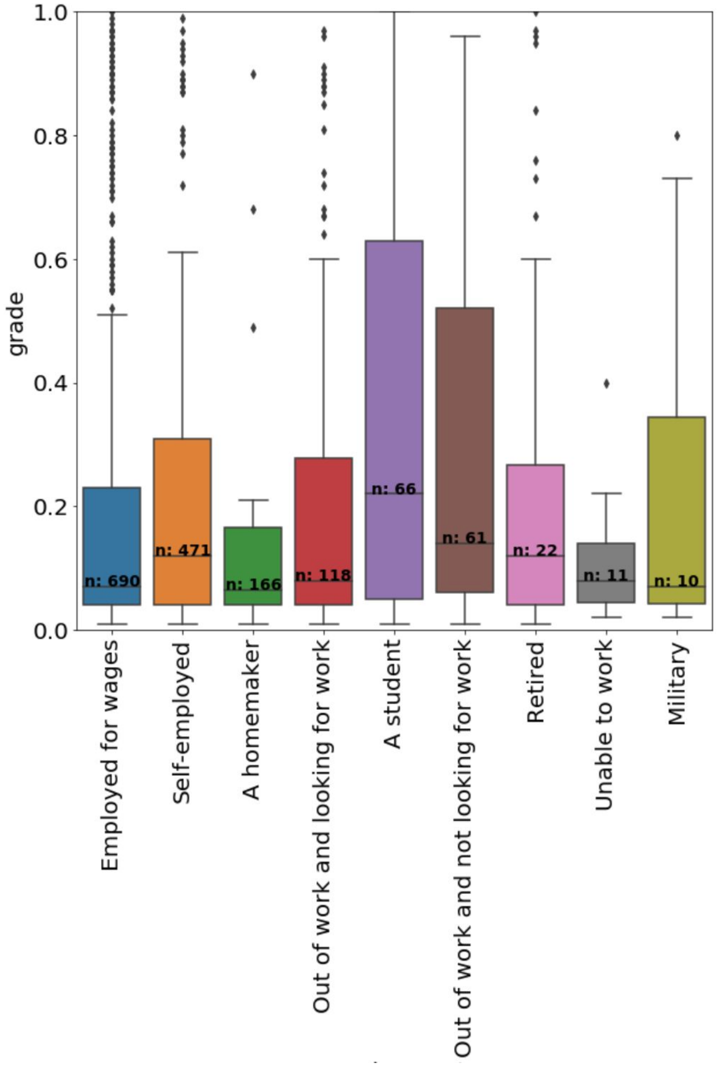
Although BJCx is targeted towards people who are full-time students, the available employment status data suggests that the majority (at least 55%) of people in the course were employed professionals. The pie chart below shows the makeup of enrolled students in BJC.1x in terms of employment status. This data was gathered in the “beginning” course survey. It reveals that only 27% of people enrolled were full-time students.

Status of Employment by Students Enrolled in BJC.1x



Although the targeted audience made up only a fraction of the class, it did turn out to be the group with the highest performance. The box-and-whiskers plot below displays the grade distributions across different statuses of employment for all students in BJC.1x with a final grade greater than zero. Full-time students proved to have performed best by a significant amount, in contrast to recent research which found that professionals performed better in MOOCs [3][5]. The plot also illustrates that those groups which follow full-time students in a ranking of median grades are 1) those who were out of work and not looking for work, 2) those who were self-employed, and 3) those who were retired. Potentially, those in the first and third group achieved higher grades because they have more time to dedicate to the course than working people have. Additionally, it may be that self-employed people are a self-selective group and are naturally self-motivated and disciplined, making them more likely to succeed in a free online course.

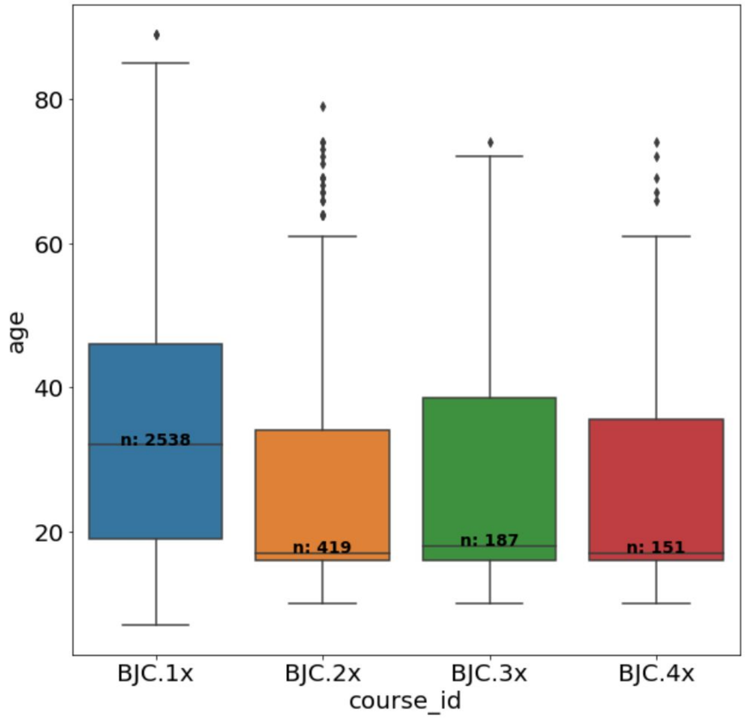
Grade Distribution by Employment Status in BJC.1x



V. Age

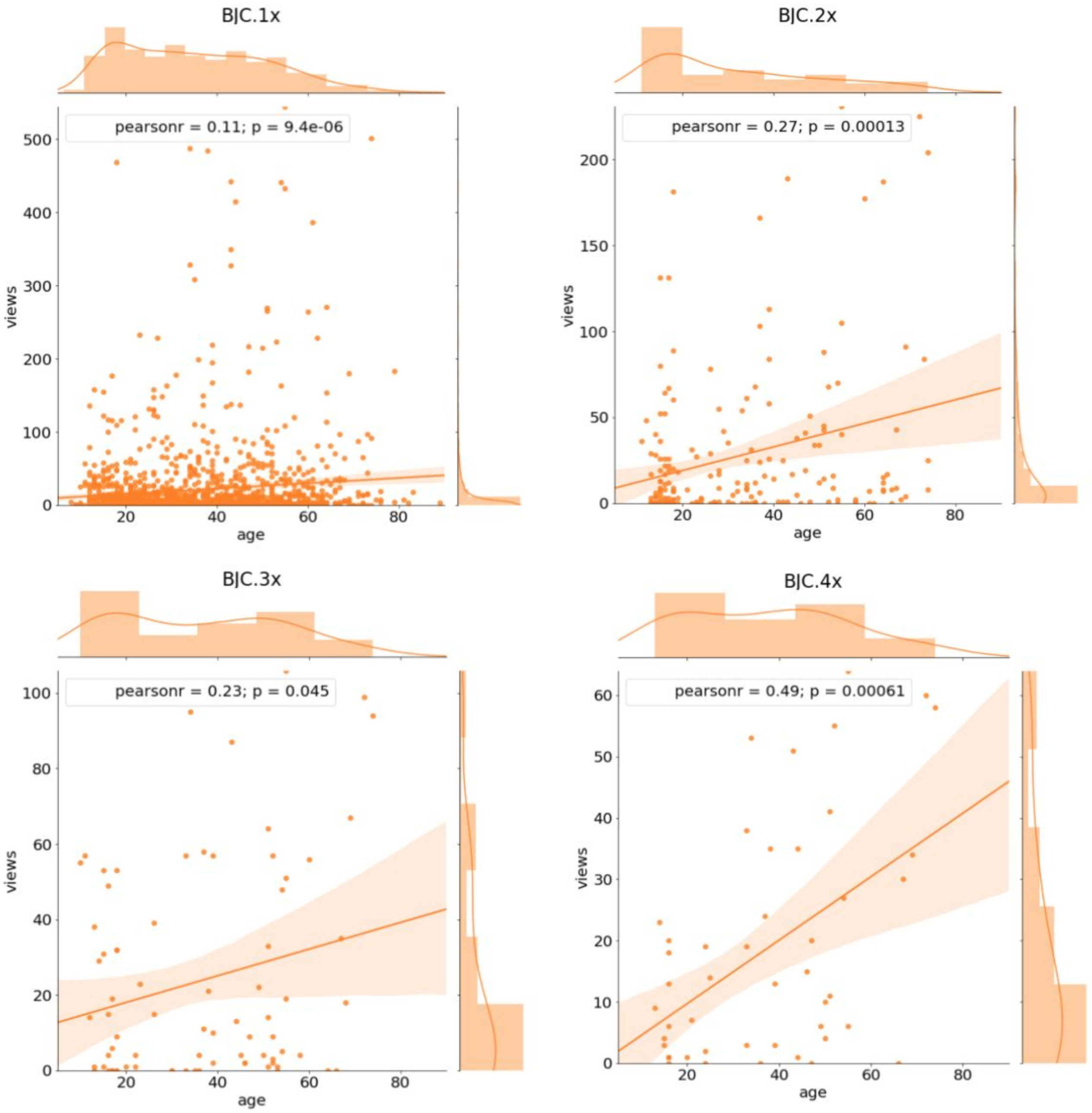
Below is a plot of student age distributions across the MOOClets, ignoring students with a final grade of zero. It shows a significant drop in the median age from the first course to the second, from early 30s to late teens. It seems that the median then stays about constant for the remaining two courses. This means that in general, older students were less likely to continue with the course than younger ones. Potentially this may be because younger students connect more with the course content and the way in which it is presented. It was, after all, designed to appeal to young people. This is only a potential explanation; the true cause is yet to be determined.

Age Distribution across BJCx MOOClets



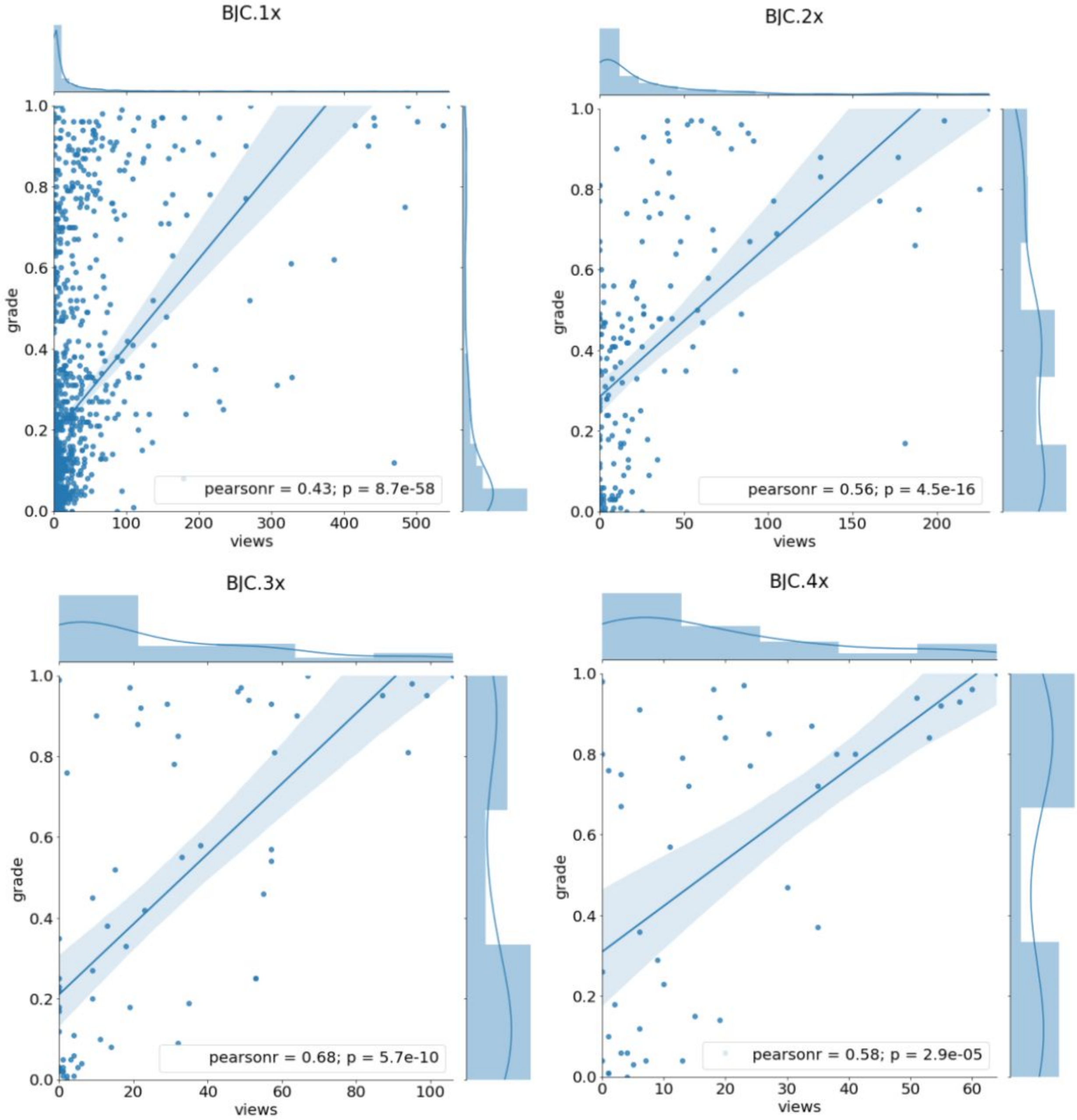
Next, it is curious to consider the relationship between age and level of engagement in the forum. In the four scatterplots below, the number of Piazza posts viewed by each student is plotted against the student's age for each of the MOOClets (chronologically from the top left plot to the bottom right plot). The data includes all students who enrolled in the Piazza page for each course, respectively. A line of best fit and the corresponding errors are overlaid to help show trends in the data. The axes are lined with histograms of their values to show their distributions. There are a few trends to observe from this figure. First, the correlation coefficient is positive in all four graphs. That means that consistently, older people were generally viewing more posts on the forum than younger people. Furthermore, the correlation coefficient is increasing almost monotonically across courses in time, suggesting that as the courses progressed, older students became more and more engaged with the forum in comparison with younger students. One potential interpretation is that younger students became less reliant on answers from the Q&A, or less engaged in reading discussions, or both, as time went on. Another explanation could be that as the age distribution shifted younger over time, only the very active older students remained in the course while the less active older students were those who left. There are many other possible interpretations of this trend.

Age vs. Number of Piazza Posts Viewed across BJCx MOOClets



Considering this correlation between age and forum interactivity is significant since engagement with the Q&A forum highly correlates with overall performance in the course. The following similar figure illustrates correlations between final grades and number of Piazza posts viewed. Across all four MOOClets, there is a very strong correlation. It is important to note that Piazza discussions accounted for 10% of students' grades. However, all of the points they received for forum participation relied on making one post each week about the readings. This sums up to only a handful of posts, and so it does not fully explain the trend. This strong positive correlation underscores the value of a discussion/Q&A forum in an online class.

Number of Piazza Posts Viewed vs. Final Grade across BJCx MOOClets

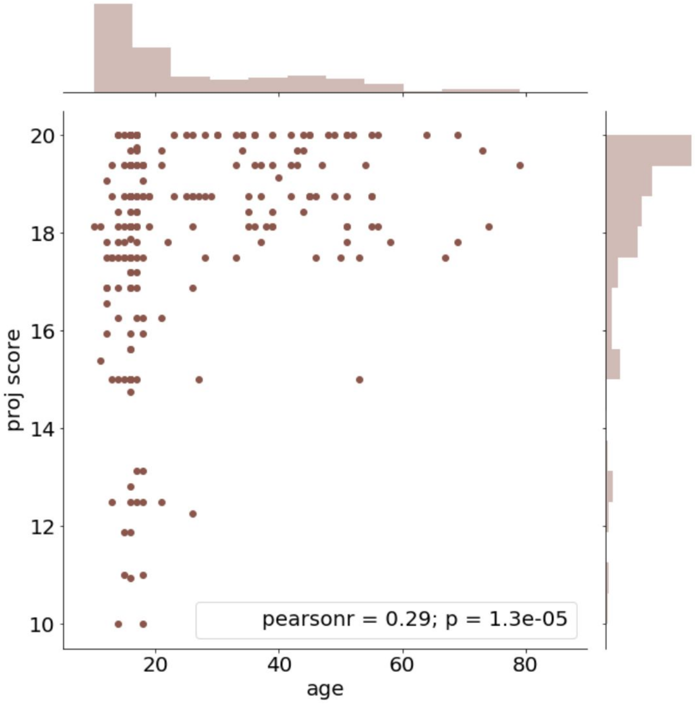


Finally, it is interesting to visualize the interaction between students' age and project scores. The scatterplot below shows each student's age plotted against the students' score on the "fun programming project" which concluded the BJC.1x course. The project was an open-ended opportunity for students to create whatever they'd like—potentially a game or some kind of interactive program—using the skills and concepts which they had learned in the first 5 weeks of the course. The data includes ages and scores of 222 students whose final project scores were greater than zero and whose ages were disclosed. Note that the projects were peer graded using a shared rubric created by the course staff.

Firstly, it is interesting to observe that the histogram of the ages of students who submitted the project skews very young. Note that this project was worth 20% of the final grade in BJC.1x. Since the grade cut-off for passing the course was 75%, it was possible to pass the course without completing the final project. In fact, 14.29% of students who passed the course （29/203） did not submit a final project.

Furthermore, there seems to be a positive correlation between age and scores on the final project, but not because more older students scored higher than younger students. Rather, it seems that the spread in project scores among young students is much wider than that among older students. The cause is unclear; it could be that some younger students struggled with creating something from scratch with little direction, since in school there are often strict instructions to follow. Or it could be that younger students have a wider range of expectations for themselves in creating a project, so some of them created programs that were too simple.

Age vs. Final Project Grade in BJC.1x



Chapter 2: How Experience and Intentions Correlate with Performance

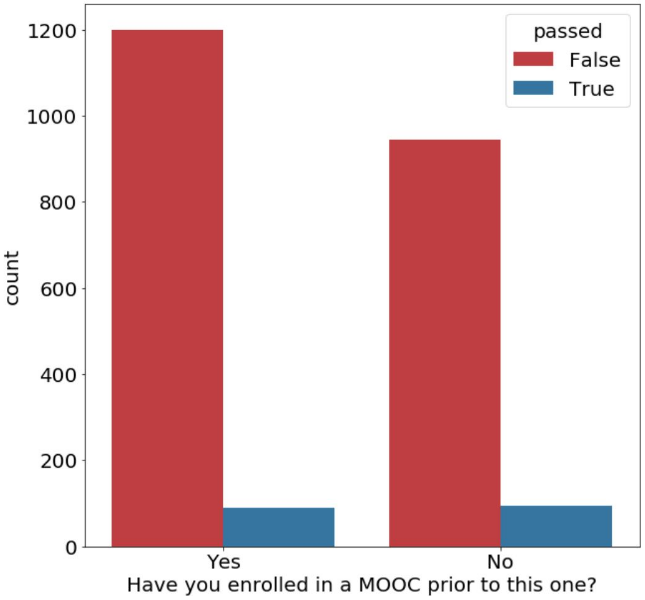
I. Overview

This section analyzes survey data from BJC.1x ranging from prior experience with MOOCs to intended level of participation in the course. In the "beginning" survey, students were asked if they had previously enrolled in a MOOC. An analysis of survey responses is performed in order to investigate whether prior MOOC experience correlates with performance in this course. Students were also asked in the "beginning" survey about their intended level participation and intended time commitment. In the "middle" and "end" surveys, students were asked about their actual time commitment. Visualizations show the interactions between intended participation/time commitment and final performance.

II. Past MOOC Experience

It is intriguing to see if students' performance in BJCx depends on past experience with MOOCs. In the "beginning" survey, students were asked if they had previously enrolled in a MOOC. The number of students who passed/failed BJC.1x is plotted below, split by their reported past experience. The plot shows that a bit over half of the students had prior experience in MOOCs, though the number of students who passed with prior experience is roughly the same as the number of students who passed without prior experience. This suggests that prior experience in MOOCs was not a predictor of passing the course.

Number of Students who Passed/Failed by Prior MOOC Experience



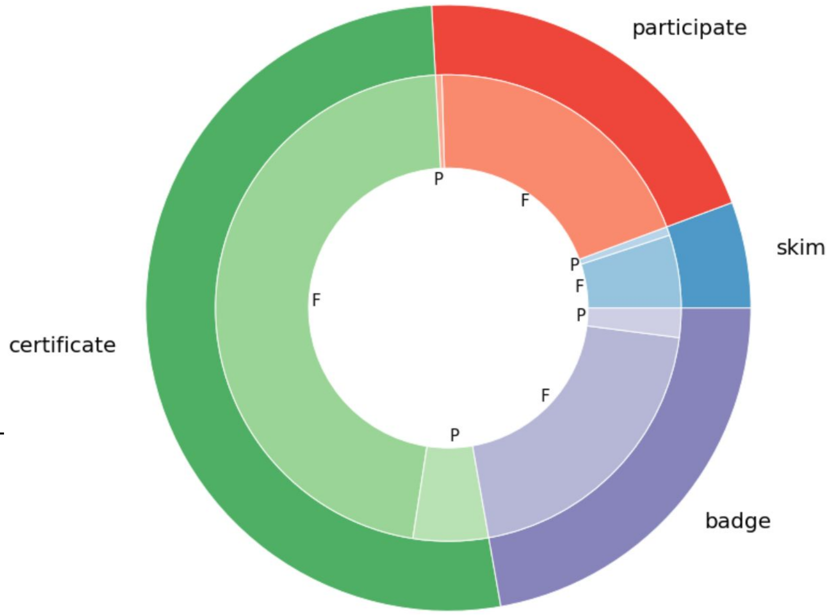
III. Intended Participation

Students came into the course with varying intended levels of participation. In the “beginning” survey, students were asked “How do you intend to participate in this course?” with the following multiple choice answer options：

- ❑ I plan to enter the course once or twice to skim material.
- ❑ I plan to participate in some course activities, but I will not complete all of the required course readings or assignments（assessment, discussion posts, etc.）
- ❑ I plan to complete all course requirements needed to earn a certificate of completion.
- ❑ I plan to complete all requirements needed to earn a digital badge[10].

The pie chart below visualizes the results of this survey question for BJC.1x. The outer ring of the pie chart shows the fractions of students who chose each of the answer options. It seems that just over half of the students intended to earn a certification of completion and only a small portion of students intended to skim the material. The inner ring of the pie chart shows the proportion of passing（P）and failing（F）grades received by students in each intended participation category. As expected, the majority of students who passed had intended to earn a certificate or a badge. Interestingly, some of the students who intended to only skim the material ended up passing the course.

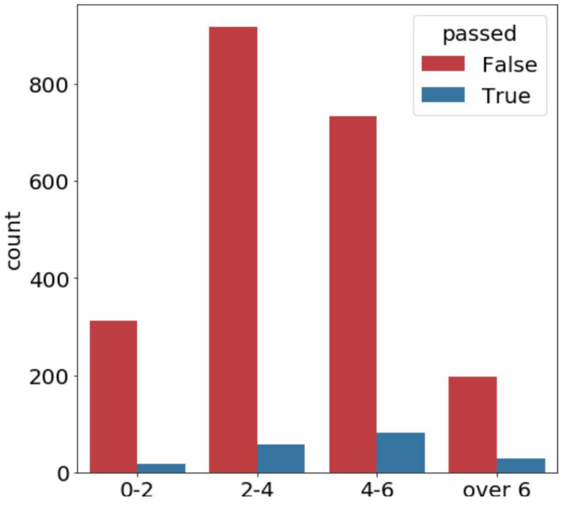
Number of Students who Passed/Failed by Prior MOOC Experience



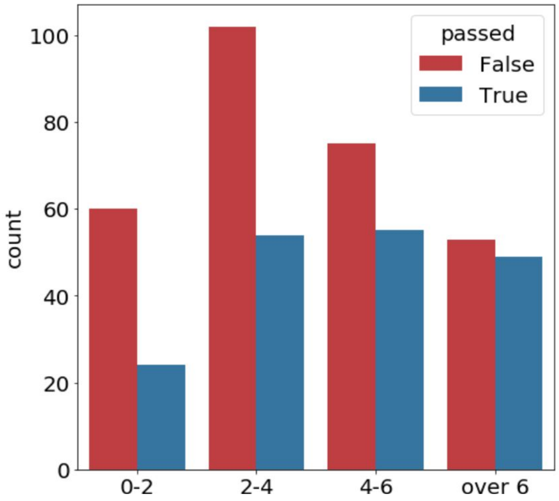
[10] For more information on digital badges, see https://open.edx.org/features/digital-badges.

Next is an analysis of time commitment in relation to performance. The three surveys asked students how many hours a week they intended to spend/were spending/had spent on the course. The plots below reveal how survey responses in BJC.1x differ between students who passed and who failed. When observing the two distributions in each plot, it becomes clear that not only did students who passed spend more time on the course than those who failed, but they also intended to spend more time on the course. A potential explanation is that intended time commitment is a proxy for level of dedication, and thus correlates highly with receiving a passing score. Also, it is surprising to observe the number of students who put in 6+ hrs/wk and failed and the number of students who put in 0-2 hrs/wk and passed. Lastly, it is interesting to compare these results with the course's suggested 4-5 hrs/wk commitment.

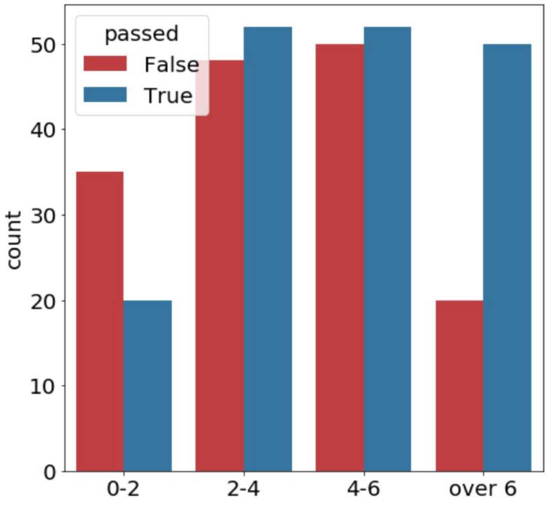
Histogram: Intended Number of Hours/Week



Histogram: Number of Hours/Week Spent ("Middle" Survey)



Histogram: Number of Hours/Week Spent ("End" Survey)



Chapter 3: A Classification Model for Student Motivations

I. Model Overview

In this section, an algorithm is proposed for classifying students' motivations for enrolling in the MOOC, as given by plain textual data. As discussed previously, people enroll in MOOCs for a variety of different reasons. These motivations are important to consider in defining student success, in designing interventions aimed at preventing attrition, in personalizing the content or workflow of the course, and arguably also in designing MOOCs in general. Given that students have the opportunity to describe their goals in plain text when enrolling in the course, this information is readily available to course administrators. However, manual inspection of the responses is not viable at MOOC scale, suggesting that an automated method for determining the most common motivations and for classifying each student's motivation could be valuable. In this section of the report, such a method is proposed and its results when applied to the BJC.1x MOOC are detailed. In the next section, this classification is used in order to analyze how motivations correlate with performance and satisfaction in the course.

In short, the algorithm takes in a list of students' written goals, uses a Topic Model to extract summaries of the most common distinct motivations present across students, and then uses distributed vector representations of words to classify each student's written goal as describing one of the previously extracted motivations.

II. Algorithm: Summary

Inputs:

- A list of **N** written goals (plain text data)
- Hyperparameters **n**, **m**, and **k**

Outputs:

- A list of **N** class labels, where each label $\in [1, n]$.

Summary:

1. Preprocess **N** written goals (plain text data) into bag-of-words (BOW) representations
2. Transform the BOW representations into tf-idf vectorized form
3. Perform matrix decomposition to extract the **n** most prevalent motivations, each described by **m** words (Topic Model)
4. Reduce BOW representations of goals to contain only the **k** most characterizing words
5. Compute a similarity score between each goal (**k** words) and each motivation (**m** words)
6. Classify each of the **N** goals into one of **n** classes.

III. Algorithm: Details

The algorithm begins by preprocessing the textual data, which contains **N** written goals. The preprocessing includes filtering punctuation, stop words, and words that are not in the English dictionary. The resulting text is treated as a bag-of-words (BOW), meaning that the ordering of the words, or the syntax, is ignored. In BOW representation, "*love to learn*" and "*learn to love*" are equivalent.

The filtered texts are then vectorized using the term frequency-inverse document frequency (tf-idf) statistic, which computes how important each term t is in each document d among a collection of documents. In short, terms which appear many times in a certain document but not in many other documents are considered important to that document. The statistic is computed using the following formulas:

$$tfidf(t,\ d) = tf(t,\ d) * idf(t)$$

$$tf(t,d) = \frac{\#\ occurences\ of\ t\ in\ d}{total\ \#\ of\ terms\ in\ d} \qquad idf(t) = log(\frac{total\ \#\ of\ documents}{\#\ of\ documents\ containing\ t})$$

When applied to a list of **N** documents which collectively contain **T** terms, this vectorization yields a **N**x**T** matrix in which rows represent documents and columns represent terms. The values in a given row represent how characteristic each term in the collective vocabulary is to a given document. Terms which are important to a given document which yield high values, while those which as less important will yield low values and those which do not appear will have value 0. In this scenario, each student's written goal is treated as a document.

Next, matrix decomposition is performed on the tf-idf matrix in order to extract the principle components. This can be done using a variety of decompositions, including principal component analysis (PCA) or non-negative matrix factorization (NMF). In this algorithm, since the tf-idf matrix is by definition non-negative, NMF is the chosen decomposition. The NMF decomposition factorizes a non-negative matrix **X** into two non-negative matrices **W** and **H** such that:

$$X \approx WH$$

In Python's scikit-learn implementation of NMF decomposition, the objective function which is minimized to yield this decomposition is:

$$min_{W,H} ||X - WH||^2_{Fro}$$

along with a series of L1 and L2 regularization terms. This objective function minimizes a squared Frobenius norm of the approximation error of the decomposition.

Since the algorithm seeks to extract the **n** most prevalent motivations described by students in their written goals, the decomposition of the tf-idf matrix is performed to yield **W** of dimension **N**x**n** and **H** of dimension **n**x**T**. The matrix **H** contains the **n** principal components of the tfidf matrix, where each component is a vector in dimension **T** containing weights corresponding to terms in the vocabulary. A topic, which in this case is a motivation, can be formed from each of the **n** principal components by concatenating the terms which correspond to the top **m** values in each component. This yields **n** topics, each described by **m** terms. This procedure is often referred to as a Topic Model.

In a similar fashion, each preprocessed written goal is reduced to only its **k** most characterizing terms (k-BOW), which are determined by the top **k** values in the row of the tf-idf matrix which corresponds to that written goal. This is done in order to simplify the representation of each written goal to a small bag-of-words. For preprocessed written goals that already contain **k** or fewer terms, nothing is done in this step.

Finally, a similarity score is produced for each (k-BOW, topic) pair and then each k-BOW is classified as belonging to the class of the topic with which it maximized the similarity score. In other words, each written goal is assigned to the motivation to which it is most similar:

$$class\ (goal_i)\ =\ argmax_j\ \ similarity(goal_i\ ,\ motivation_j\ )$$
$$\forall i \in [1,\ N],\ \ \forall j \in [1,\ n]$$

Now, computing the similarity between two BOWs is a non-trivial step. There are many ways this could be done. The chosen implementation in this work draws from word embeddings, which are dense, continuous, vector representations of words. A widely-used pre-trained embedding is word2vec, which was created by researchers at Google [18]. It was created using a recurrent neural network and a Google News corpus of 3 billion terms. It is widely used due to the size of its training data and its success in encoding nuanced semantic features of words. In fact, in this embedding, words with similar semantics are represented by vectors with high cosine similarity [19][20]. Cosine similarity is computed by:

$$sim\ (word_i,\ word_j)\ =\ cos(\theta_{ij})\ =\ \frac{vec(word_i)\cdot vec(word_j)}{||vec(word_i)||\ ||vec(word_j)||}$$
$$\forall word_i,\ word_j \in \{embedding\ vocabulary\}$$

This similarity function is used to compute the similarity score between two BOWs. More concretely, the similarity score of two BOWs is calculated using the n_similarity method of word2vec model. Because this word2vec model has been shown to encode many aspects of the semantics of words, it is valuable for computing how similar two words, or two BOWs, are to each other.

IV. Hyperparameter Selection

Hyperparameters **n**, **m**, and **k** influence the topics which result from the Topic Model as well as the values outputted by the similarity score function. These values can be chosen in a variety of ways. In this implementation, they are chosen using a grid-search aimed at maximizing the mean similarity score of the (k-BOW, motivation) pairs outputted by the classifier. A subset of the grid-search results from applying this algorithm to BJC.1x MOOC data is shown below for illustration.

Hyperparameter Selection Table： A Subset of Grid-Search Results

n	m	k	Mean of similarity scores of classified pairs	Std. of similarity scores of classified pairs
4	1	3	0.436	0.248
5	1	3	0.448	0.249
6	5	3	0.495	0.144
6	10	5	0.563	0.128
7	10	5	0.578	0.126

V. Limitations

This method comes with a few limitations that must be considered. For one, the Topic Model produces summaries of topics in a bag-of-words format, which can provide an idea of the relevant words for each topic, but disregards the context in which those words were used in the original text. This is because the tf-idf vectorization of the original textual data does not encode any syntactic information. To illustrate this, consider the following hypothetical goals written by Student A and Student B：

- A： “*I already know how to program, but I want to learn other information.*”
- B： “*I only want to learn how to program.*”

After preprocessing, the two written goals become：

- A： “*already know program want learn information*”
- B： “*want learn program*”

At this point, the two BOWs look quite similar to each other （shared words underlined）. When their vectorized tf-idf representations are used to extract only the **k** most characterizing terms from each BOW, it is plausible that the two results will be very similar, and potentially even the same. Because of this, the classifier would likely classify both written goals as the same topic. Any human reader would easily understand that student A and student

B did not have the same motivation for enrolling in the course, but the algorithm would likely fail to pick up on the semantic differences between the two students' written goals since it ignores syntax.

Another limitation of this method is that it assumes that each student only has one motivation. For each student's written goal, the algorithm classifies it with the topic which maximizes the similarity score. If a student's written goal describes more than one motivation, then the topic with which it is classified depends on the relative word frequencies used, which may be arbitrary in relation to the student's actual priorities. The model performs a significant simplification of written goals, which are already simplified representations of real, complex human motivations. It is also important to note that since the Topic Model assumes that each written goal describes only one motivation, resulting topics generated using goals which describe multiple motivations may not be informative or distinct.

Finally, a third limitation of this algorithm is the mis-classification of outliers. Since the algorithm classifies each written goal into one of **n** categories, goals which do not naturally fit into any of the categories are bound to be misclassified. It is neither practical nor valuable to try to represent all possible motivations in the topic model while still maintaining **n** << **N** at MOOC scale. Rather, it is valuable to reduce the amount of information contained by students responses to a mere summary of only the most common motivations.

VI. Classification Results

This motivation classification method was applied to data from the BJC.1x MOOC, which contains 8909 students' written goals. Below are five examples of students' written goals as taken from this dataset. It is apparent even just from these examples that students enroll in the course for a variety of reasons. This is both a challenge and a reward of designing MOOCs!

1. "*I teach Scratch to underprivileged children in Ukraine.*"
2. "*Life longer* [sic] *learner who recently exited Corporate America and now has the time to learn everything else that she never had time for before*"
3. "*I'm experimenting with on-line* [sic] *learning.*"
4. "*I want to expand my knowledge. I want to better understand English. I love to try new things.*"

5. "*I would like t* [sic] *pursue a career in computer programming or the computer science department*"

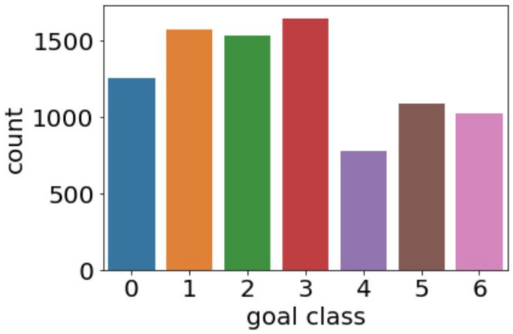
When vectorized in tf-idf form, the dataset yielded 2,479 terms. The Topic Model applied to the tf-idf matrix resulted in 7 distinct motivations, detailed in the table below.

Table of Goal Classes (**n** = 7, **m** = 10, **k** = 5)

Goal Class	**Motivation** (qualitatively chosen keywords in bold)
0	learn **programming** world **linux program** way topics want **technology coding**
1	learning **love life** interested long **lifelong experience fun** great learner
2	knowledge improve **career gain development** expand **professional** understanding enhance **skills**
3	want **online courses** know **good university better** time course try
4	education **free self personal** online **opportunity curiosity** continuing courses universities
5	like **science computer** interested programming course courses **school teacher teaching**
6	**new skills** things study interesting **improve** subjects explore **useful english**

It is quite remarkable that the algorithm yields these distinct motivations, many of which are known to be prevalent among MOOC students in general. Next, the algorithm performs classification using similarity scoring. The histogram below shows the distribution of students among the seven classes.

Distribution of Students Among Goal Classes



As a sanity-check, a random subset of the classification results were manually inspected. This manual check confirmed that similarity scoring is qualitatively performing reasonably well. A subset of the classification results is included in the table below. The first three results illustrate the success of the algorithm, while the last three results illustrate its limitations.

Example Inputs and Outputs of the Classifier

	Student's written goal	**Class assigned by the algorithm**
A	*"I teach Scratch to underprivileged children in Ukraine."*	5
B	*"I am interested in furthering my education in order to increase my employment opportunities."*	2
C	*"learning + joy + fun"*	1
D	*Use as a resource for my AP course that I teach*	6
E	*"I have deep interest in immunology. I want a research carrier* [sic] *in this feild* [sic] *in future."*	3
F	*"I have a deep interest in knowing if there are planets on which life may be feasible to exist and develop."*	6

Results A, B, and C are classifications which seem to be qualitatively correct. The classification of A makes sense since the motivation described by goal class 5 seems to be about teaching computer science. The classification of B is logical since the motivation described by goal class 2 seems to be about improving career opportunities. Lastly, the classification of C into goal class 1 is expected because the motivation described by class 1 seems to be the pleasure derived from learning. These three results illustrate the model's ability to correctly classify a subset of written goals.

However, results D, E, and F shed light on some of the algorithm's limitations. Qualitatively, result D seems like it should be classified in goal class 5 since it describes a motivation related to teaching computer science, though the algorithm has incorrectly placed it in class 6. Similarly, result E seems like it should have been classified in goal class 2, even though in reality it is an outlier which does not really fit any of the motivation categories. In this case, not only does the model ignore the misspelled instance "carrier," which should be "career," and thus fail to discover the similarity between this written goal and goal class 2, but it also fails to recognize that this written goal is an outlier in the first place. Similarly, result F is an outlier[11] which doesn't seem to belong to any of the categories, making it bound to be misclassified.

[11] It is interesting to note the existence of a sizeable minority of written goals which seem completely unrelated to the course content but related to other subjects. This may be a result of limited English fluency among the global audience of online learners.

Chapter 4: How Motivations Correlate with Demographics, Performance, and Satisfaction

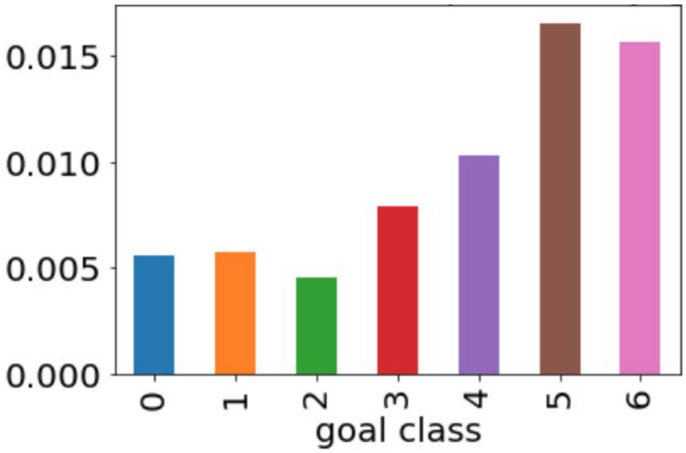
I. Overview

In this section, the results of the motivation classification algorithm applied to data from the BJC.1x MOOC are analyzed jointly with other data in order to see how motivations may correlate with student demographics, performance, and satisfaction. The data used to generate the plots in this section is limited to the subset of students who enrolled in the MOOC who also wrote something in the goals section of their enrollment form, which is 46.48% of all enrolled students.

II. Passing Rates

Below are passing rates of students in each goal class. These results suggest that students who enrolled with motivations related to teaching computer science or gaining useful skills/practicing English were more likely to pass the course than those who enrolled with motivations related to professional development or learning to code. This correlation can be intuitively justified by the nature of the BJC curriculum, which focuses more on teaching computer science principles and big ideas than on teaching practical programming skills to be used in industry.

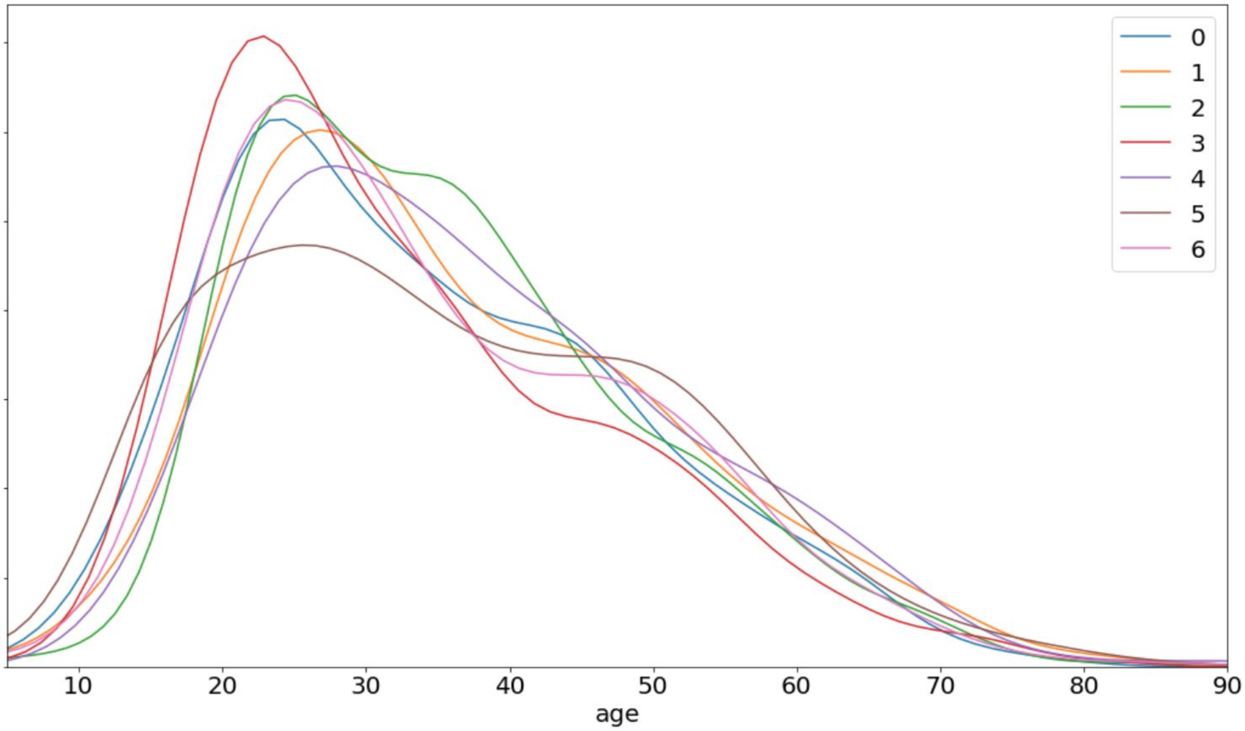
Passing Rates by Goal Class



III. Age Distributions

The figure below shows approximated (smoothed) age distributions for each of the seven goal classes of students enrolled in BJC.1x. This plot shows how different motivations correlate with age.

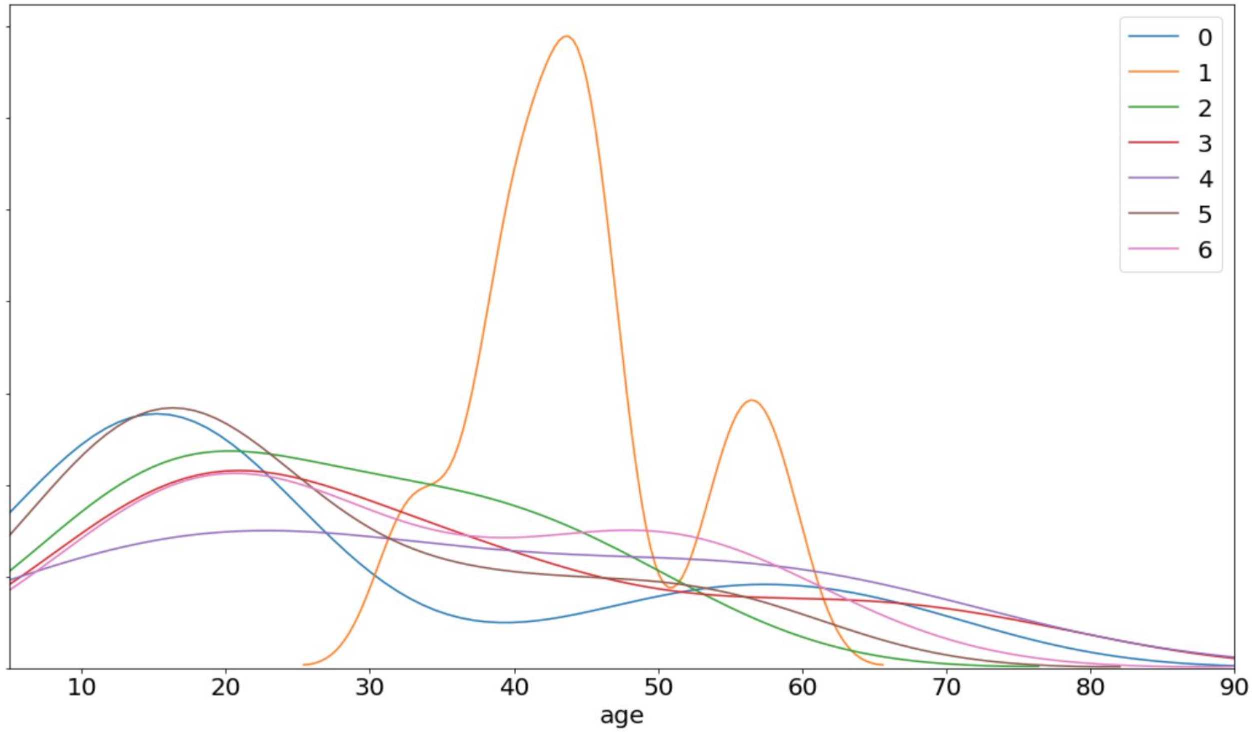
Age Distribution by Goal Class of Enrolled Students



There are a few interesting trends to interpret from this plot. Among students who enrolled, age distributions across goal classes are similar, with slight differences that can be intuitively justified. Several of the distributions are bimodal, with a peak age in the 20s and a point of inflection leading to a local max around the 40s or 50s. Those enrolled who were seeking coursework from a prestigious university (goal class 3) have the highest proportion of students in their teens and 20s. Additionally, students who were looking for career improvement and professional development (goal class 2) have the highest proportion of students in their 30s and early 40s. Furthermore, those whose motivations related to teaching computer science (goal class 5) have the lowest proportion of young people and the highest proportion of people in their 50s.

The following figure shows approximated (smoothed) age distributions for each of the seven goal classes of students who received passing scores in BJC.1x. It is important to note that the number of data points represented in this plot is quite small; only 78 students both passed the course and wrote about their goals in the enrollment form on edX.

Age Distribution by Goal Class of Students who Passed

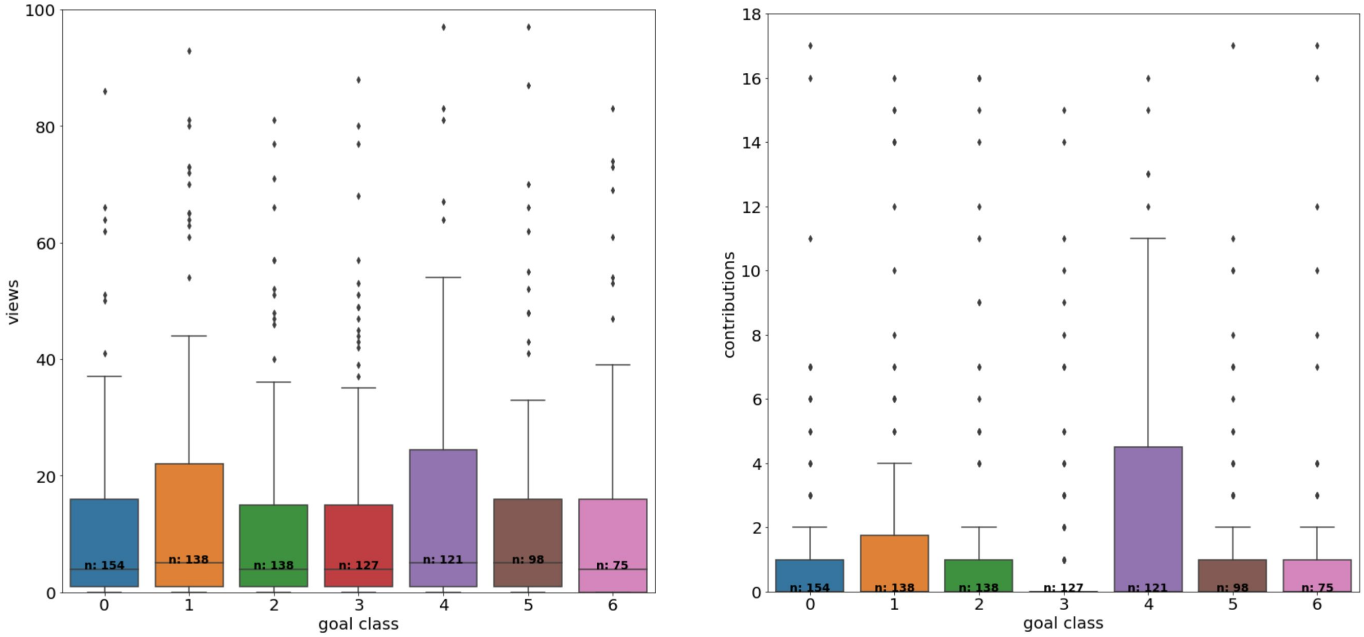


The age distribution of students who passed who were motivated by lifelong learning (goal class 1) clearly stands out in this figure. It is a bimodal distribution with peaks in the 40s and 50s. Another bimodal distribution in this figure is that of students motivated by learning how to code (goal class 0), which peaks in the teens and then again around age 60.

IV. Forum Participation

A point of interest is the correlation between different motivations and different levels of engagement on the Q&A forum. The data used for this analysis comes from 851 students who both registered for the course's Piazza page and described their goals on edX. The two boxplots below show the distributions of number of Piazza posts viewed and number of contributions made, respectively, by goal class. Major outliers have been removed from the plots for the sake of more interpretable visualizations.

Goal Class vs. Number of Piazza Posts Viewed (left) and Contributed (right)

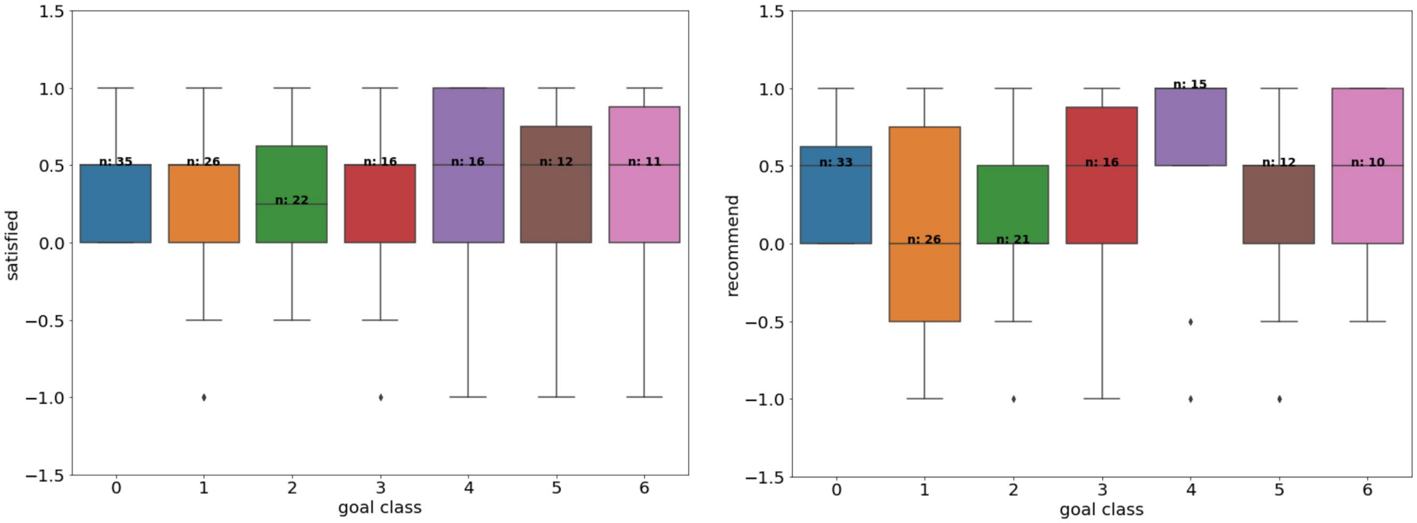


The 25th percentile, median, and 75th percentile number of posts viewed by each student (figure on the left) seems to be relatively constant across goal classes, with the exception of goal classes 1 and 4, which have higher 75th percentiles. These two classes include students motivated by lifelong learning or by personal curiosity. Both those groups also appear to be leading in terms of number of contributions to Piazza per student (figure on the right). These trends could be interpreted to mean that students with the most intrinsic motivations for taking the course are most likely to be engaged in asking/answering questions and discussing the course materials.

V. Satisfaction

Another trend to inspect is how students with different motivations may finish the course with different levels of satisfaction. There were 138 students who both described their goals on edX and completed the final course survey. The two plots below show their distributions, by goal class, of level of satisfaction with their participation in the course (on the left) and of likelihood to recommend the course to a friend (on the right). It is important to note that it is likely that this data is skewed by other factors which had led these students to get to the end of the course and to choose to complete the final survey. Nonetheless, the results show some relative differences in distributions across goal classes.

Goal Class vs. Level of Satisfaction （left） and Likelihood of Recommendation （right）

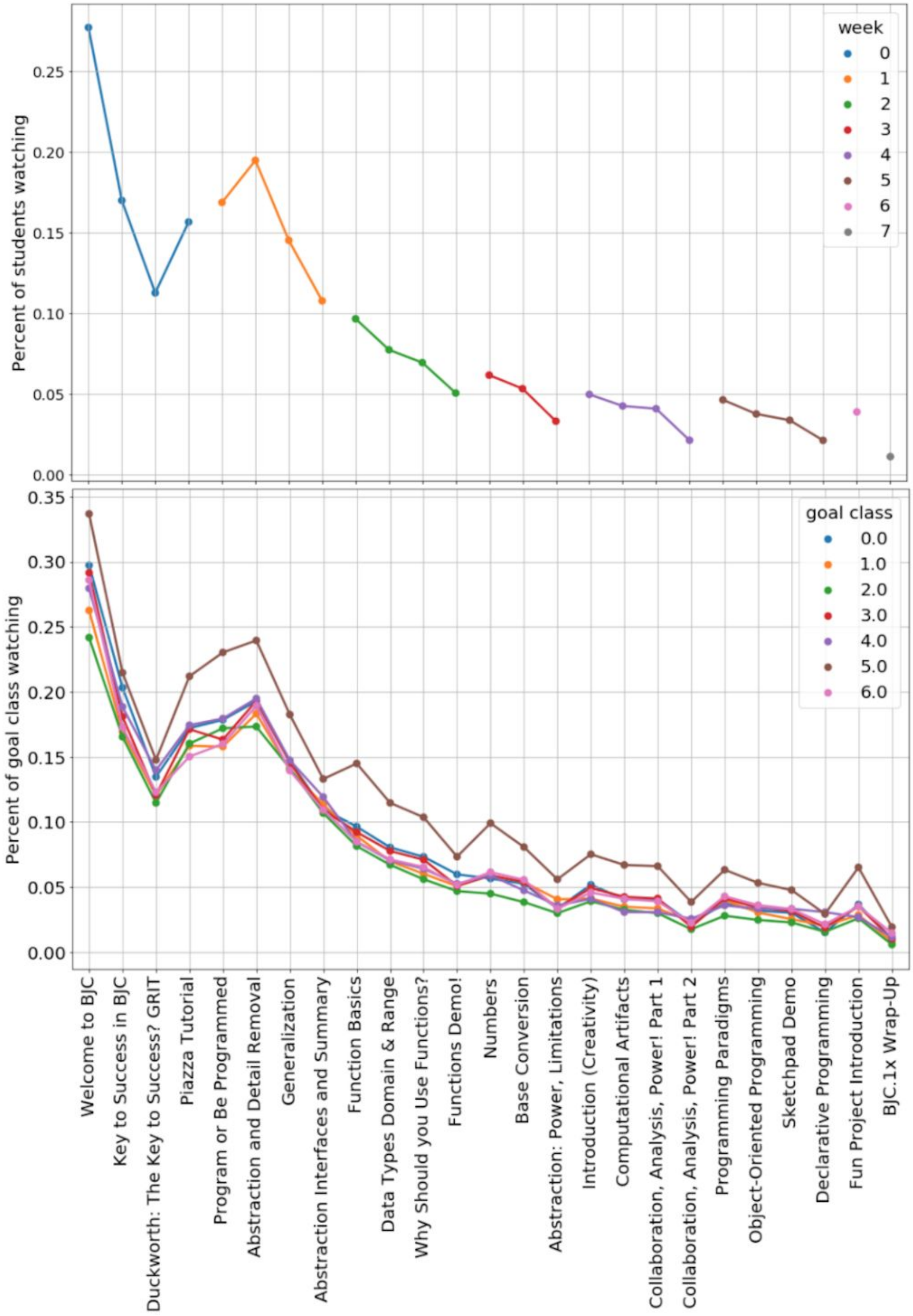


It is apparent that a majority of students who took the final survey were at least somewhat satisfied with their participation in the course and were at least somewhat likely to recommend the class to a friend, regardless of motivation category. Notably, those motivated by personal curiosity or by improving their English （goal classes 4 and 6） had the highest 75th percentile scores in both the satisfaction scale and the recommendation scale. In addition, students who were motivated by career improvement and professional development （goal class 2） had the lowest median satisfaction score and tied with the lifelong learners for lowest median recommendation score （which was still non-negative）.

VI. Attrition

Finally, a visualization of how motivations correlate with attrition can give insights into which students stay engaged in the course and at which points in the curriculum students drop out. In the two plots below, lecture videos are organized on the x-axis in chronological order and the percent of students watching each video is plotted.

Attrition: Percent of Enrolled Students Watching Lecture

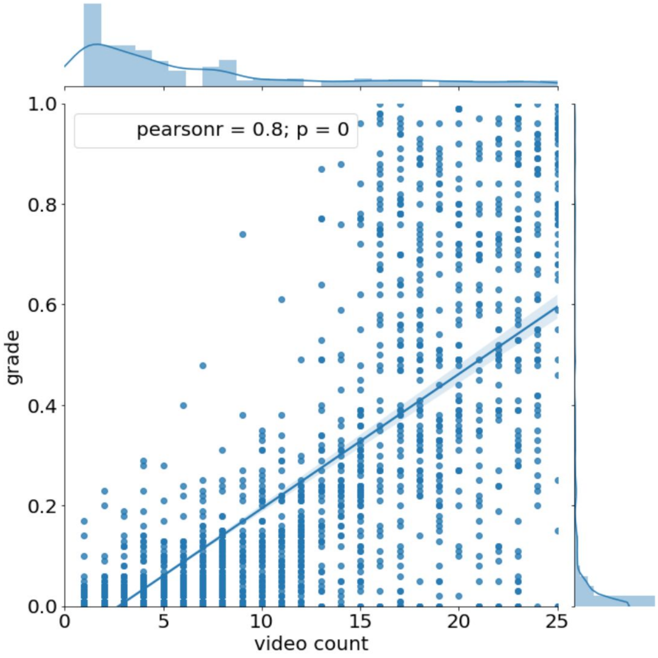


The first plot shows engagement percentages for the course as a whole (19,168 students) colored by week. It depicts that almost 30% of those who enrolled in the course watched the welcome video, while only about 20% watched the very first lecture video (titled “Abstraction and Detail Removal”). By the end of week 2, the vast majority of students who ended up disengaging from watching lectures had already disengaged; the percentages of students watching the lectures in weeks 3-5 are quite stable. Furthermore, it looks like the last video of each week would consistently have fewer views than the previous videos from the same week and than the first video of the following week. This was the case regardless of whether a week had 3 or 4 lecture videos. The cause of this is unclear, though it might be a case of students only watching as much of the lectures as they deem necessary for receiving a passing grade (>75%). Another hypothesis is that the last lecture of the week may have generally been more of a summary than an introduction to new material, and so students may have regarded it as optional in comparison to the others. It is also noteworthy to point out that about 4-5% of students watched a majority of the lectures, which is about 4x the percent of students who passed the course.

The second plot shows engagement percentages colored by goal class. The data used in this plot only includes the students for which there exist written goals (8,909 students). In general, students across the different goal classes exhibited very similar attrition patterns. Those who were motivated by teaching computer science (goal class 5) consistently had the highest lecture watch rate, while those motivated by career improvement and professional development almost always had the lowest watch rate.

Number of Lectures Viewed vs. Grade

This discussion is important because watching lecture videos highly correlates with performance; the lectures introduce material which is later reinforced in lab exercises and homeworks. The plot to the right shows the strong positive correlation between number of videos watched and final grade. It also depicts what seems like a latent threshold in which almost all students who passed (received a final grade over 0.75) watched at least 60% (15/25) of the lecture videos. Though the correlation is expected, the threshold is surprisingly stark.



Chapter 5: Student Reflections on BJCx

In the "middle" and "end" surveys, students were asked to reflect on their experience in the course. They were asked to rate the utility of each component of the curriculum and to quantify their learning takeaways and their levels of satisfaction with the course. The results of these survey questions in BJC.1x are illustrated in this section. In general, they suggest that students had very positive experiences in the course.

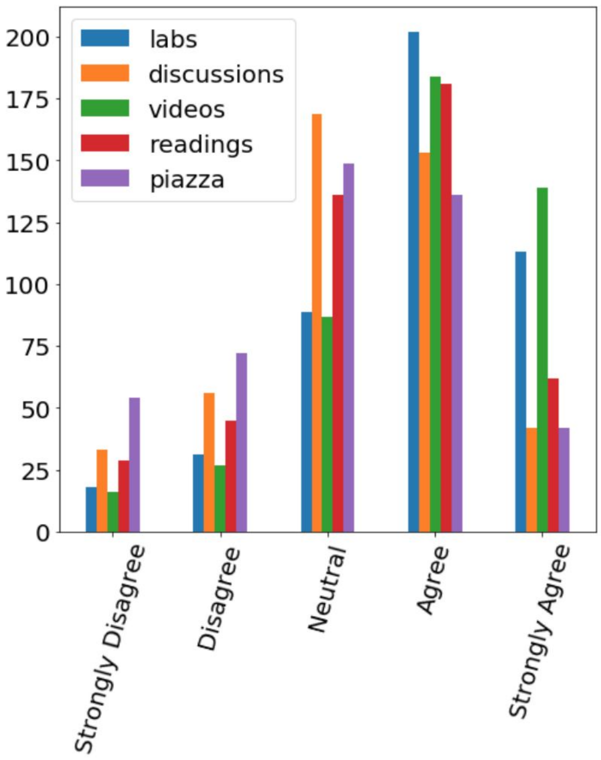
The plot below shows students' levels of agreement with the statement,

"The [course component] were helpful in my understanding of the material."

where *course component* is one of the following:

- Labs
- Discussions
- Videos
- Readings
- Piazza

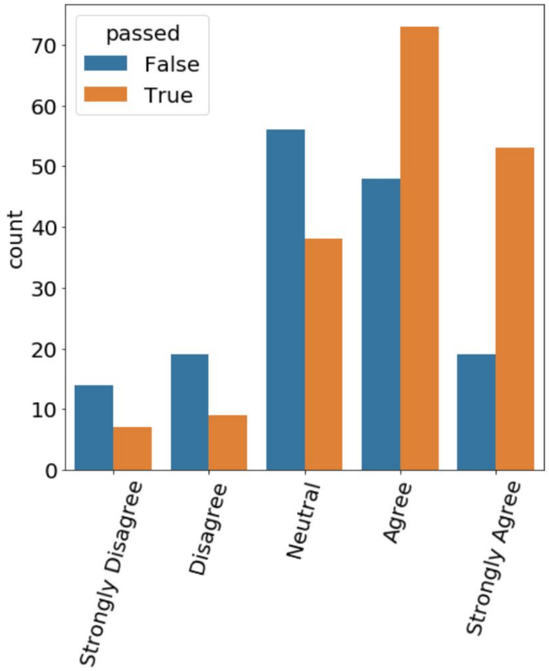
Helpfulness of Course Components



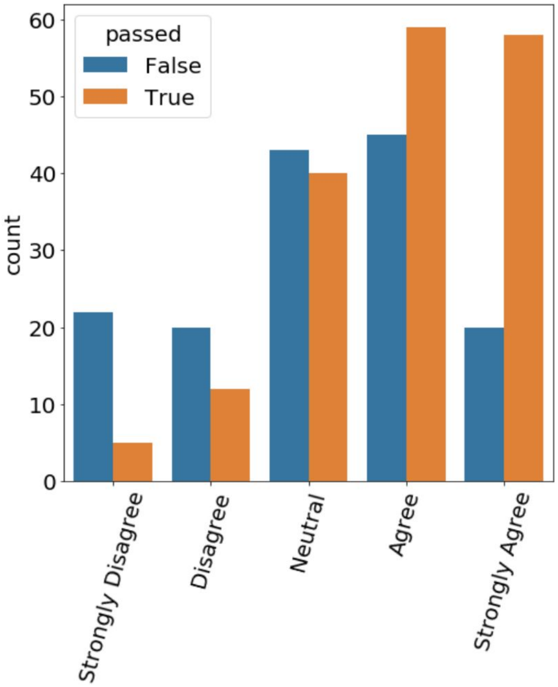
The plot above shows that in general, most students found each component of the course helpful. The distributions for discussions and Piazza are skewed to suggest that they were slightly less helpful for students than labs, videos, and readings. In all, labs and videos seem to be the most helpful components of the course.

Below are a set of plots which show students' level of satisfaction with their experience in the course, as measured by various survey questions, along with their overall performance in the course. The plots show how students who passed/failed the course reflected differently about their experiences. In general, it seems that students who passed were more satisfied with their experiences, as can be expected. However, it seems that many students who did not pass were satisfied with their experiences, too.

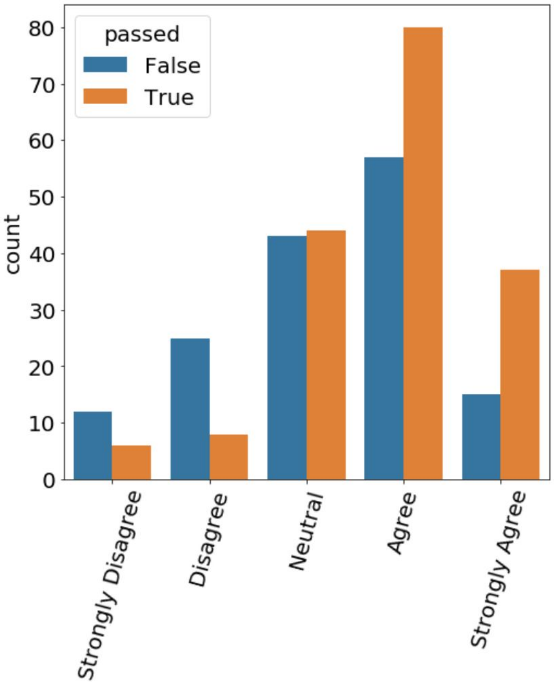
After reaching the end of the course,
I was satisfied with my level of participation.



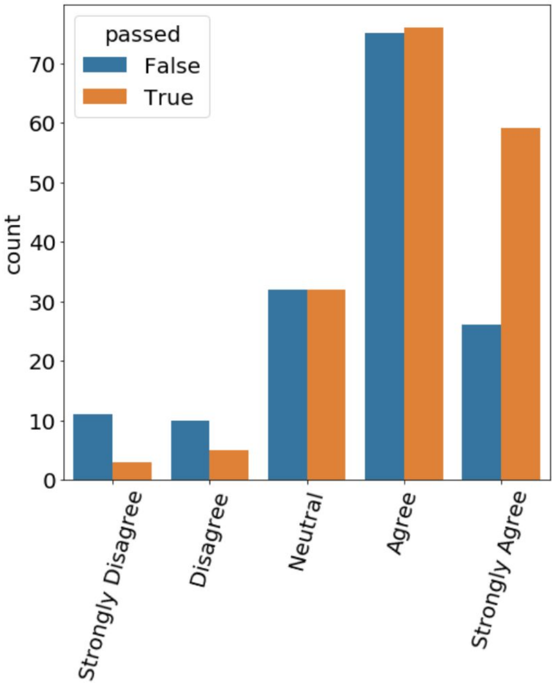
After finishing this course, I want to recommend this course to a friend.



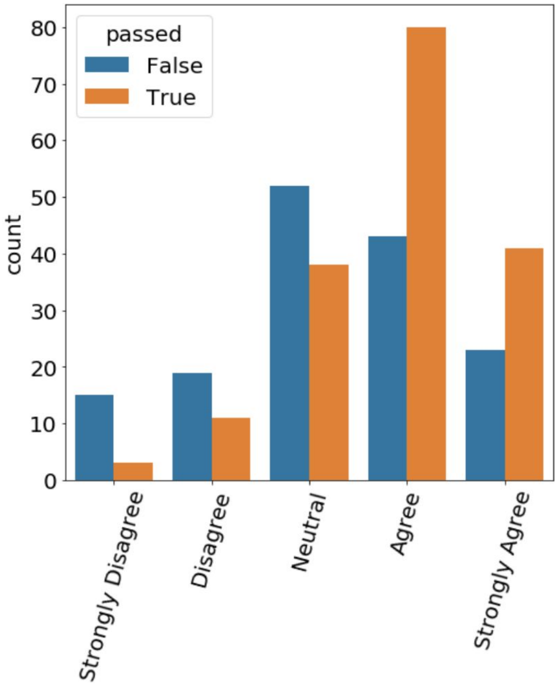
There was sufficient support for me to effectively participate in the course.



After finishing this course,
I feel that I have gained a better
understanding of computing.



I am content with the finished product of my final project.



Conclusion

This report contains a thorough analysis of students' demographic, interactivity, survey, and discussion forum data from the BJCx MOOC. The results of this analysis paint a more detailed picture of student enrollment, engagement, and performance. Also included in this work is the proposal of a novel algorithm for summarizing and classifying students' motivations as given in plain-text. The results of applying this algorithm to students' self-reported goals in BJCx gives way for further analysis and the contextualization of their performance and satisfaction levels in the course.

The goal-classifying algorithm can be used by MOOC administrators for a wide variety of purposes. Automatically, the algorithm extracts the most common goals students describe when writing about what they hope to get out of the course. This summary of common motivations can be used to direct a course's emphasis or workflow. Furthermore, the algorithm's classification of students by goal provides opportunity for the personalization of course content, communications, grading, or even interventions aimed at preventing attrition.

The results of this data analysis raise awareness of trends which course administrators, both of BJCx and of MOOCs in general, can work to address in future iterations of their classes. The proportional drop-off of female students after the first of four course segments is a surprising and disconcerting trend found in BJCx. To address this, specifically in the context of a computer science course, teachers should make sure that collaboration and creativity (which have been found to make computer science curricula more approachable and appealing to women, given their exclusion from the field in recent decades) are not just opportunities offered in the course but are accessible and emphasized aspects of the course experience. An online setting makes this challenging; collaboration between strangers can be hard to facilitate and creativity may be in tension with leaving assignments too open-ended. However, tackling these challenges has the potential to mitigate the gender imbalance that is particularly visible in computer science courses.

Another focus moving forward would be to ensure that the course description accurately depicts what students can expect to get out of completing the course. For BJCx, this could have helped filter enrollments by 1) encouraging students who want to satisfy personal curiosities or prepare for the AP CS Principles exam to enroll and 2) by discouraging students looking for professional development or career-based skills from enrolling. In general, a targeted description which aligns with the purpose of the course, detailing exactly what the course is and what it is not, may not only decrease attrition rates but may also attract more students from the target audience.

Although not a majority of those enrolled, the target audience of the BJC curriculum has been found to be actively engaged with the course material and to be achieving the highest level of performance out of all groups. Contrasting the results of past studies, which found that those who are most likely to succeed in MOOCs are people with college or graduate degrees, the data from BJCx depicts that full-time students and those with only primary or secondary school educations engage most and perform best.

Other results found in this analysis include that older students generally engage more with the Q&A/discussion forum and that past experience with MOOCs does not strongly correlate with future MOOC performance. Furthermore, this study reveals that students who pass not only commit more time to the course, but also *intend* to commit more time to the course from the very beginning.

The insights obtained through this data analysis, in conjunction with the use of a novel goal classification algorithm, can be used to help MOOC instructors better tailor their courses to the students which they aim to serve. Aided by the rapid advancement of technology, MOOCs have the potential to span all corners of the globe and to provide unparalleled personalized education at significantly lower costs. That, in essence, is the spirit with which MOOCs were created: to make higher education accessible to all. Educators in the online space must not get too distracted by the technology, the hype, and the numbers. They must remain focused on who it is they are working to serve and design their courses to best encourage the success of those students.

References

1. Garcia, Dan, Brian Harvey, and Tiffany Barnes. "The beauty and joy of computing." ACM Inroads 6.4 (2015): 71-79.
2. Crowther, Geoffrey. "Inaugural Address." Presentation of the Charter, July 1969, Open University, Speech transcript.
3. Dynarski, Susan. "Online Courses Are Harming the Students Who Need the Most Help." The New York Times, 19 Jan. 2018.
4. Zhenghao, Chen, et al. "Who’s Benefiting from MOOCs, and Why." Harvard Business Review, 22 Sept. 2015.
5. Konnikova, Maria. "Will MOOCs be Flukes?" The New Yorker, 7 Nov. 2014.
6. Guo, Philip J., and Katharina Reinecke. "Demographic differences in how students navigate through MOOCs." Proceedings of the first ACM conference on Learning@ scale conference. ACM, 2014.
7. Kizilcec, René F., Chris Piech, and Emily Schneider. "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses." Proceedings of the third international conference on learning analytics and knowledge. ACM, 2013.
8. Milligan, Colin, and Allison Littlejohn. "Why study on a MOOC? The motives of students and professionals." The International Review of Research in Open and Distributed Learning 18.2 (2017).
9. Zheng, Saijing, et al. "Understanding student motivation, behaviors and perceptions in MOOCs." Proceedings of the 18th ACM conference on computer supported cooperative work & social computing. ACM, 2015.
10. Fei, Mi, and Dit-Yan Yeung. "Temporal models for predicting student dropout in massive open online courses." Data Mining Workshop (ICDMW), 2015 IEEE International Conference on. IEEE, 2015.
11. DeBoer, Jennifer, et al. "Bringing student backgrounds online: MOOC user demographics, site usage, and online learning." Educational Data Mining 2013. 2013.
12. Yang, Diyi, et al. "Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses." Proceedings of the 2013 NIPS Data-driven education workshop. Vol. 11. 2013.

13. Chaplot, Devendra Singh, Eunhee Rhim, and Jihie Kim. "Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks." AIED Workshops. 2015.
14. Wang, Yuan, and Ryan Baker. "Content or platform: Why do students complete MOOCs?." Journal of Online Learning and Teaching 11.1 (2015): 17.
15. Brooker, Abi, et al. "A tale of two MOOCs: How student motivation and participation predict learning outcomes in different MOOCs." Australasian Journal of Educational Technology 34.1 (2018).
16. Alcon-Heraux, Maria. "7 Things to Know about the AP Program Results: Class of 2017." All Access: News for Members, College Board, 21 Feb. 2018, www.collegeboard.org/membership/all-access/academic/7-things-know-about-ap-program-results-class-2017.
17. Steinglass, Alice. "5 things to know about taking the AP CS Principles Exam." Code.org, Medium, 8 Feb. 2018, medium.com/@codeorg/5-things-to-know-about-taking-the-ap-cs-principles-exam-30c3770ee13.
18. Word2vec model: https://code.google.com/archive/p/word2vec.
19. Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations." Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013.
20. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Acknowledgements

Thank you to Professor Armando Fox, Michael Ball, and Peter Sujan for helping me get access to the data. Thank you to Alvin Ghouas and Rita Guo for tips on data visualizations. Thank you to Aashna Mukerji for proofreading. Thank you to my dad, Arnon Amir, for talking through my ideas with me and giving me feedback on my writing. And a big thank-you to Professor Dan Garcia for the inspiration, support, and encouragement.

Inspiration for the Reader

In a world directed by technology, we must have technical problem solvers from all walks of life creating the solutions which impact people from all walks of life.

You are not only welcomed, you are needed.
Let's make a computer science industry that is so diverse that there is no such thing as "fitting in."