# MOOC Prediction Analysis and Pattern Discovery

Zhouxiang Cai
Co Tran
Rachel Witner

Mentor:
Prof. Huzefa Rangwala

# Overview

- Motivation/background
- The data
- Co Tran’s work
- Rachel’s work
- Zhouxiang’s work


canvas
Offer a Course | About Us | FAQs | Institutions | Login
ENGAGE IN MORE THAN A COURSE:
ENGAGE IN A LEARNING EXPERIENCE.
Enroll in open, online courses from colleges, universities, and organizations worldwide.

# Motivation

- How to define a successful MOOC?
- How to define course completion?
    - 38% of user-course pairs were active for at least one week
    - ~9% of students were “active” for at least half of the weeks of a course
    - 0.08%  of student enrollment logs had a date of completion
    - 37% of computed final scores were missing
    - 45% of non-missing computed final scores were 0
- How to define engagement?
- Are there recurring patterns of interaction across courses, users, and time?

# The Data

- Canvas Network open courses released by Harvard Dataverse
- January 2013 to July 2016
- ~380 courses
- ~400,000 students enrolled
- User page views (requests)

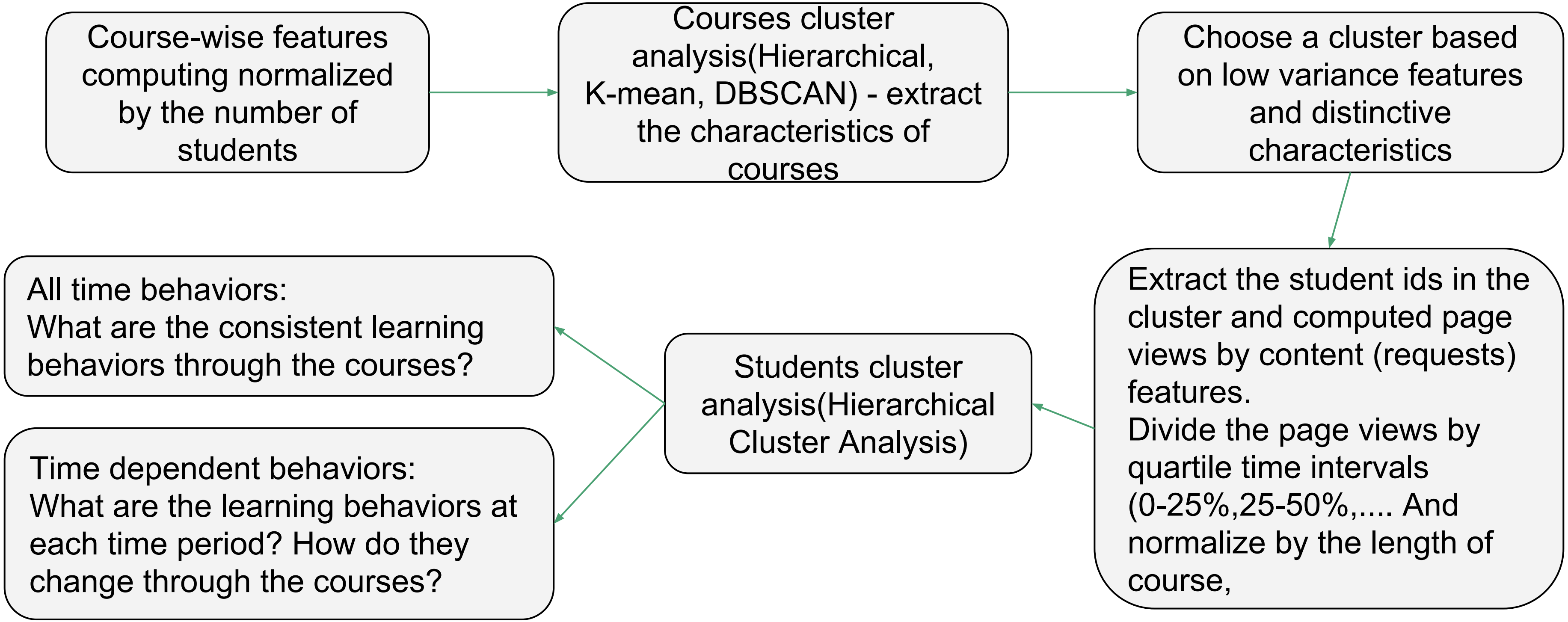

A sample of the star schema structure

# Typology of learning behaviors of students in online courses - Co Tran

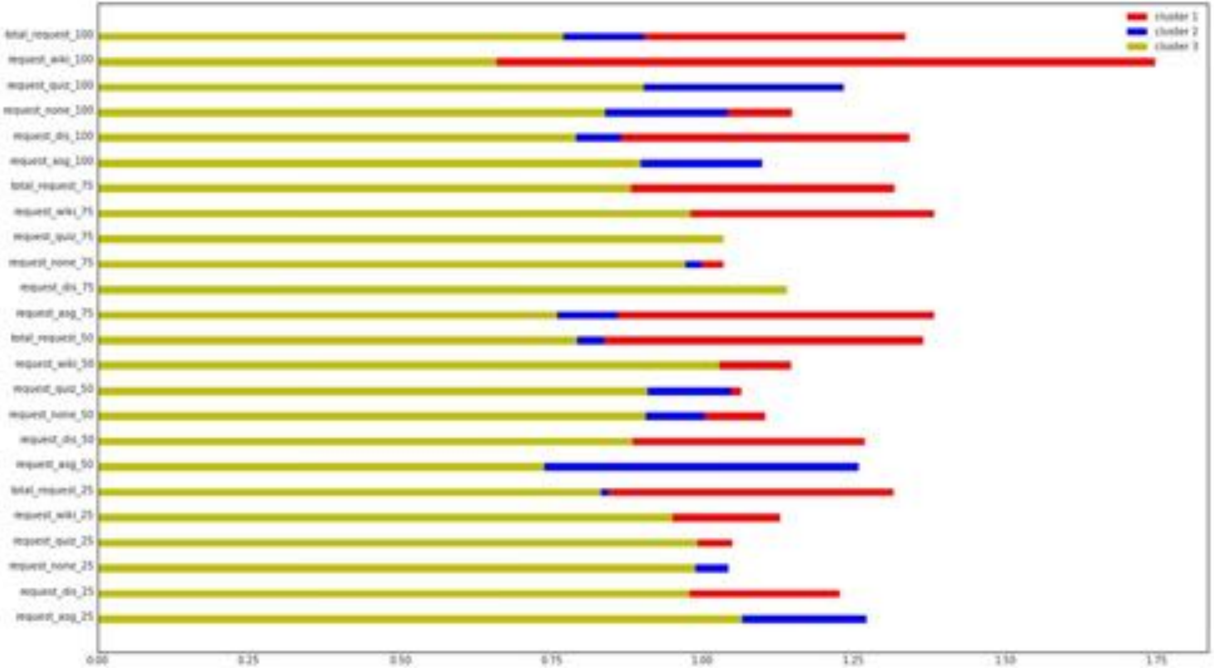
Motivation : - Previous studies of student learning pattern researched on the sample size of 1 or 2 courses.

Objective : - Studying the learning behaviors of students in the scale of multi-course using cluster analysis.

# Method


Course-wise features computing normalized by the number of students
Courses cluster analysis(Hierarchical, K-mean, DBSCAN) - extract the characteristics of courses
Choose a cluster based on low variance features and distinctive characteristics
Extract the student ids in the cluster and computed page views by content (requests) features.
Divide the page views by quartile time intervals (0-25%,25-50%,.... And normalize by the length of course,
Students cluster analysis(Hierarchical Cluster Analysis)
All time behaviors:
What are the consistent learning behaviors through the courses?
Time dependent behaviors:
What are the learning behaviors at each time period? How do they change through the courses?

# All time behaviors approach - Characteristics and student outcomes explained

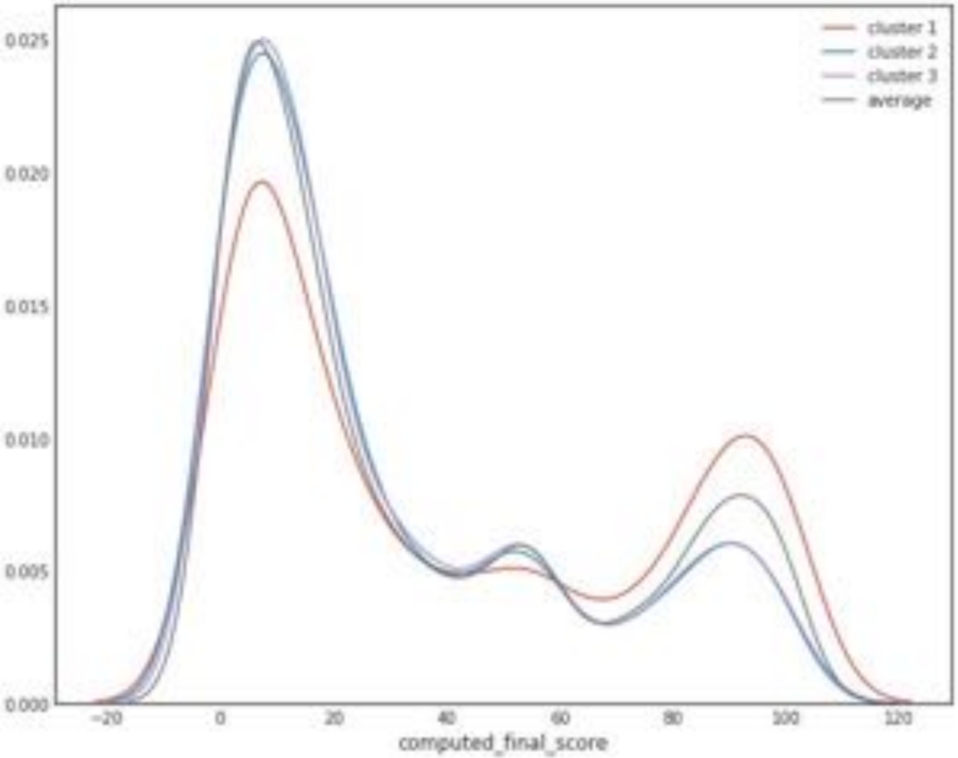


Cluster 1: high engagement in discussion and reading wiki pages and higher average score.

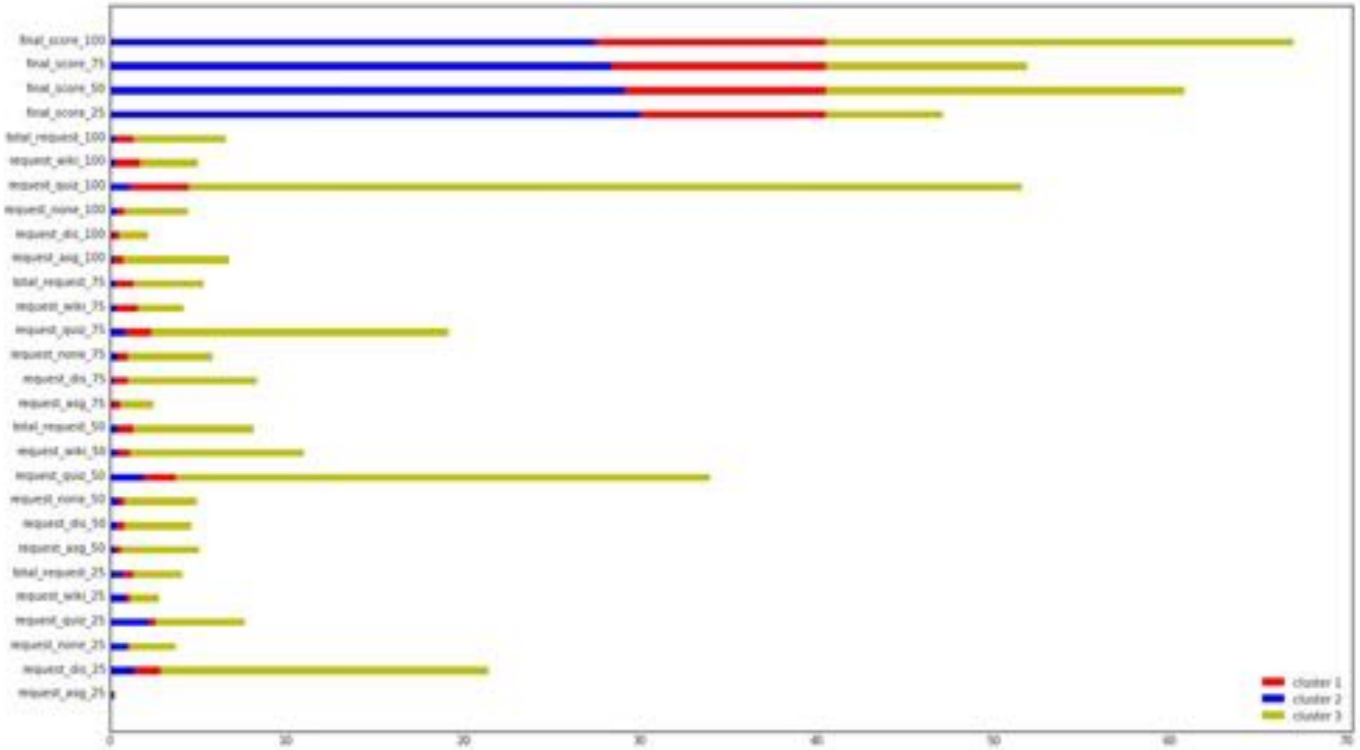
Cluster 2: low engagement in discussion and reading wiki pages, high engagement in assignment and low average score.

Cluster 3: has low engaging in every activity especially in assignment, discussion, and reading wiki pages.

# All time behaviors approach - Characteristics and student outcomes explained



# Results - Time dependent behaviors approach




Cluser 1: normal engagement in all activities.
Cluster 2: low engagement in all activities.
Cluster 3: high engagement in all activities

# Changes in the memberships of clusters

**Jaccard similarity coefficient**: The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

| group | cluster_1 | cluster_2 | cluster_3 |
|---|---|---|---|
| group_25_50 | 1.0 | 0.928501193988 | 0.0468164794007 |
| group_50_75 | 0.999194414608 | 0.887533683166 | 0.1053227633070 |
| group_75_100 | 0.999194414608 | 0.866238401142 | 0.1274418604651 |
| group_100_25 | 1.0 | 0.892426367461 | 0.0179257362356 |
| group_25 | 3721 | 6906 | 238 |
| group_50 | 3721 | 6823 | 321 |
| group_75 | 3724 | 6486 | 655 |
| group_100 | 3721 | 6587 | 557 |

**Table 7: Jaccard coefficient and number of students in each cluster**

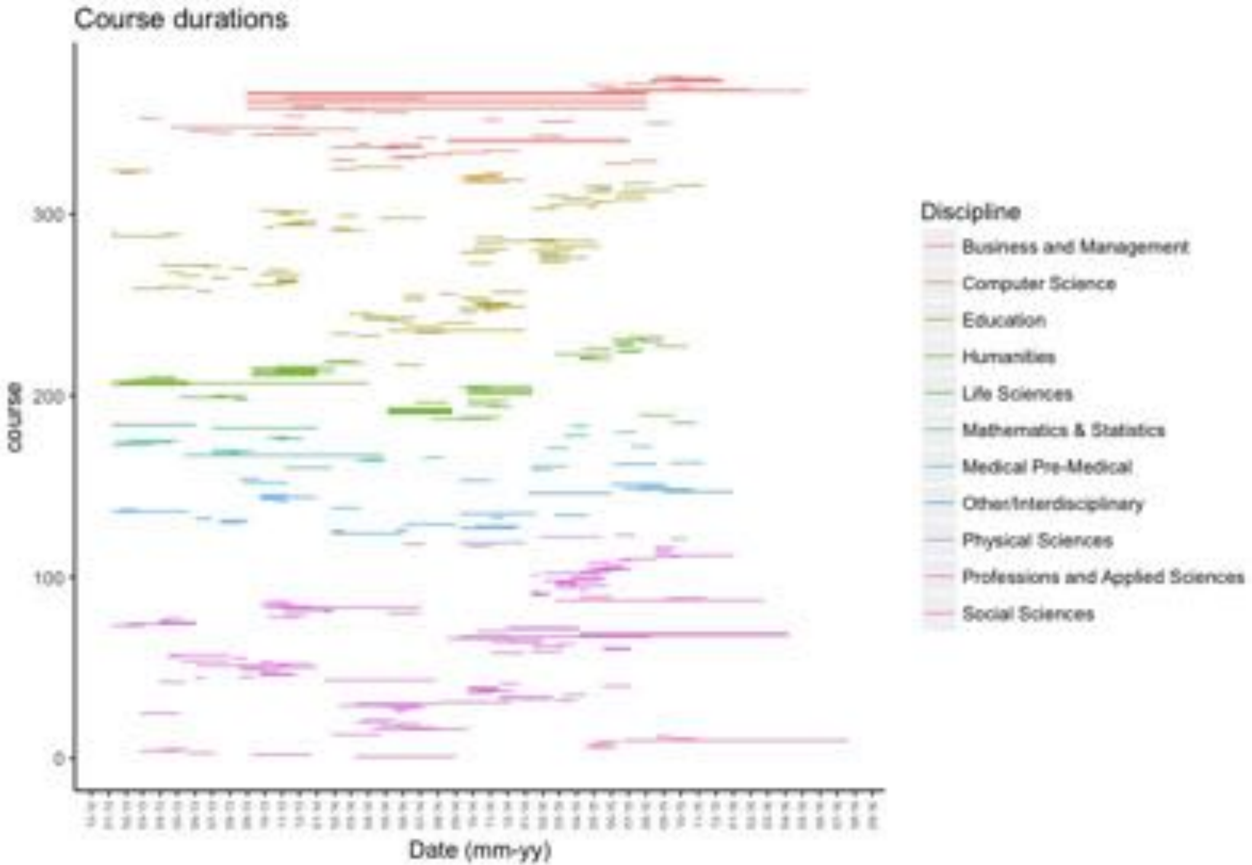
# Interesting findings

- Cluster 1 in both approaches has the same memberships

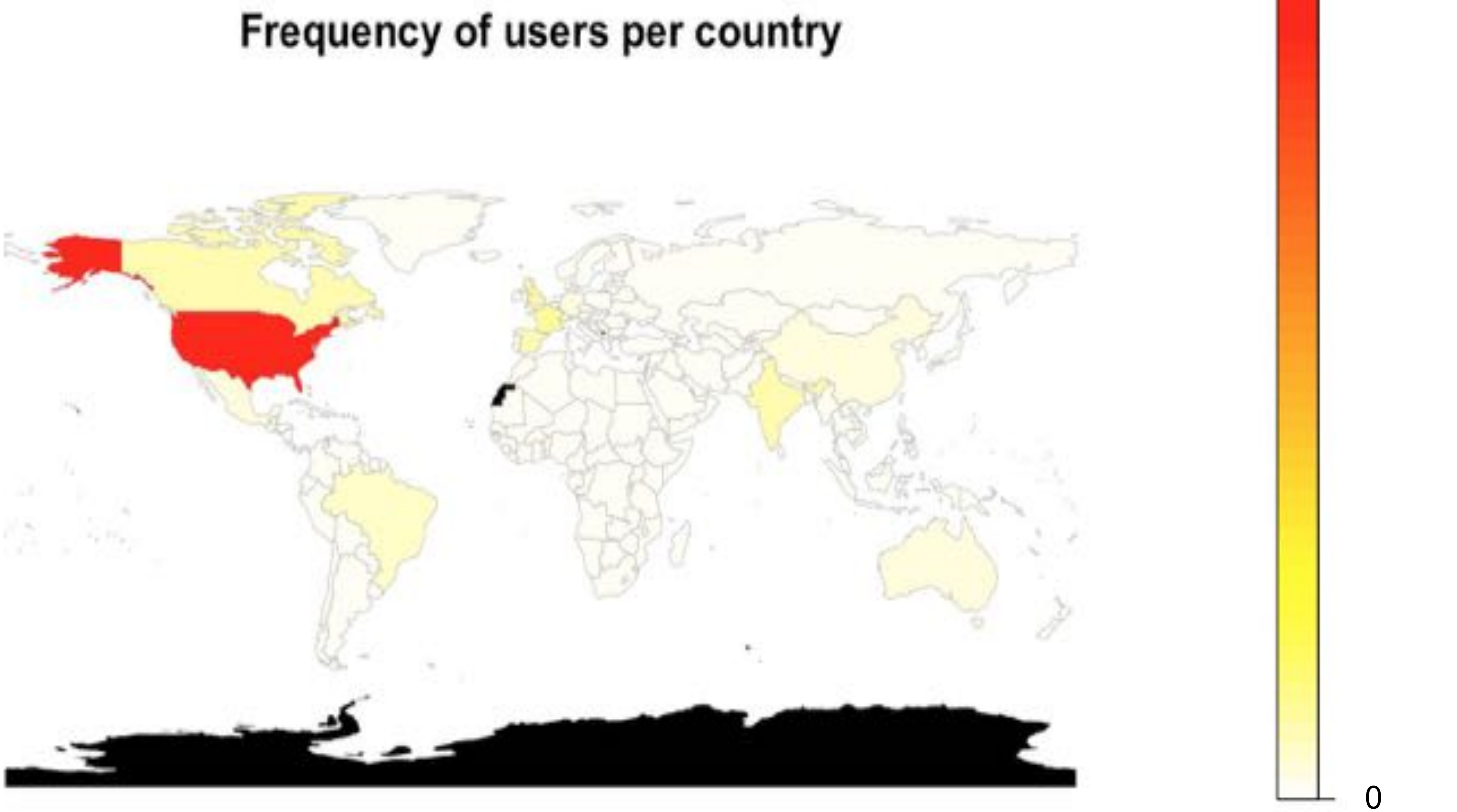
→ The learning behaviors of students in cluster 1 are mostly the same in each intervals of time as well as throughout the courses.

- The exchanges in memberships of time dependent behaviors approach mostly appear in cluster 2 (low engagement) and cluster 3 (high engagement).

# Data Visualizations

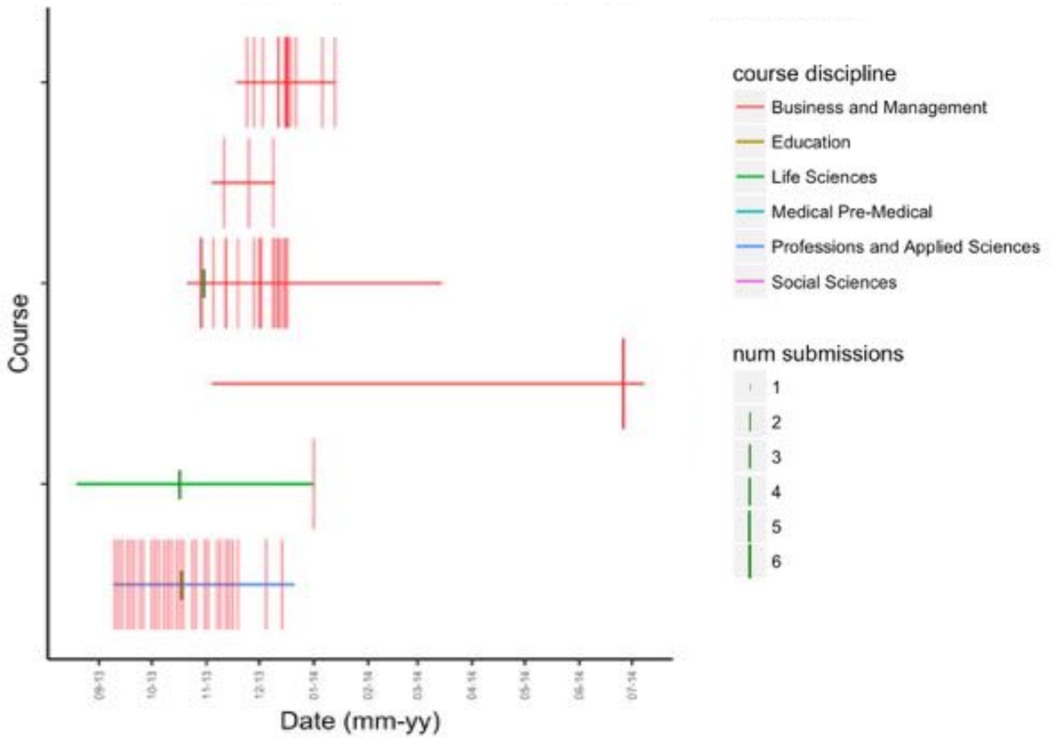


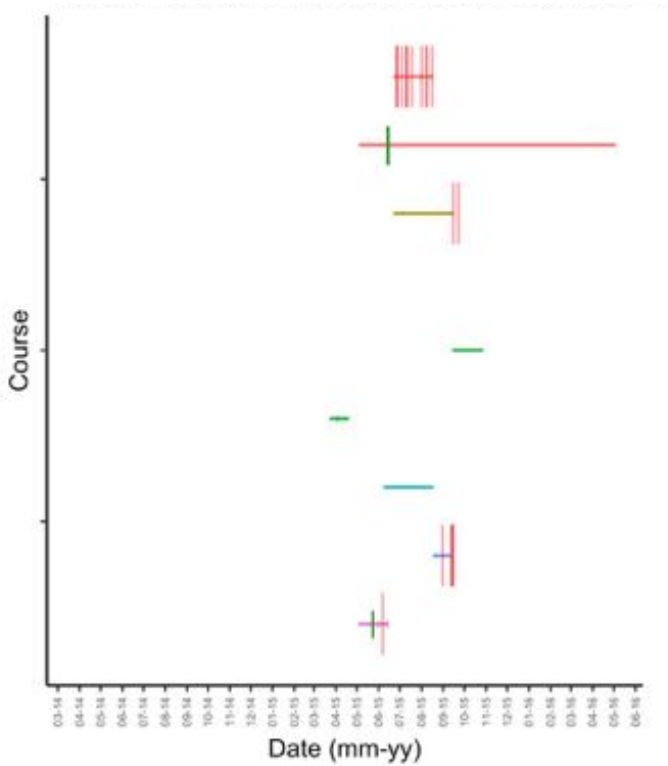
# Data Visualizations



# Data Visualizations

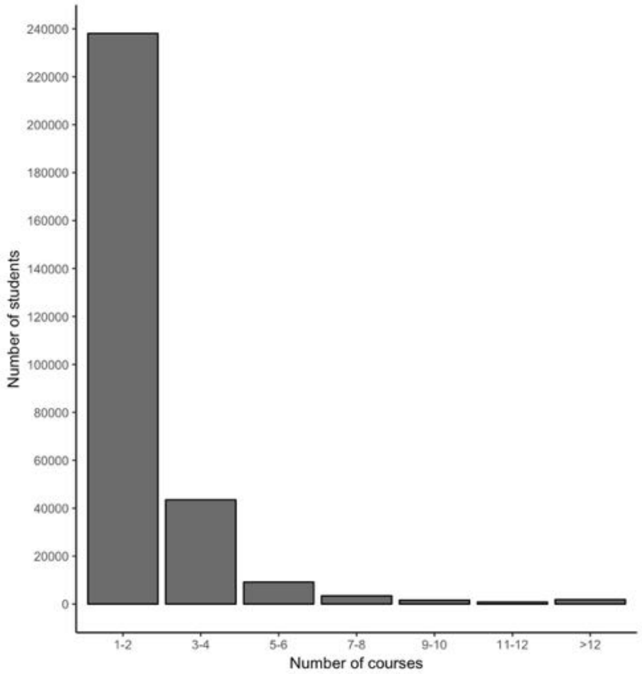
Submissions (green) and due dates (red) for 2 randomly selected individuals



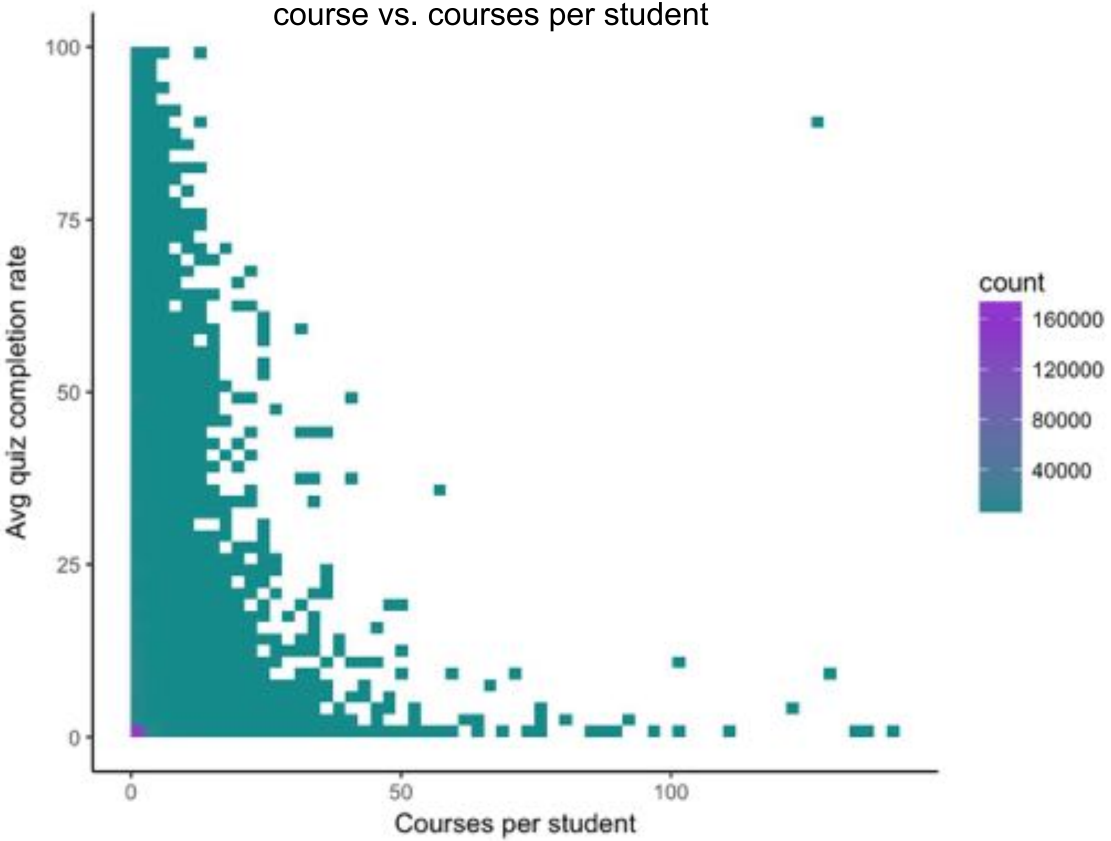



# Data Visualizations

Distribution of Courses per Student

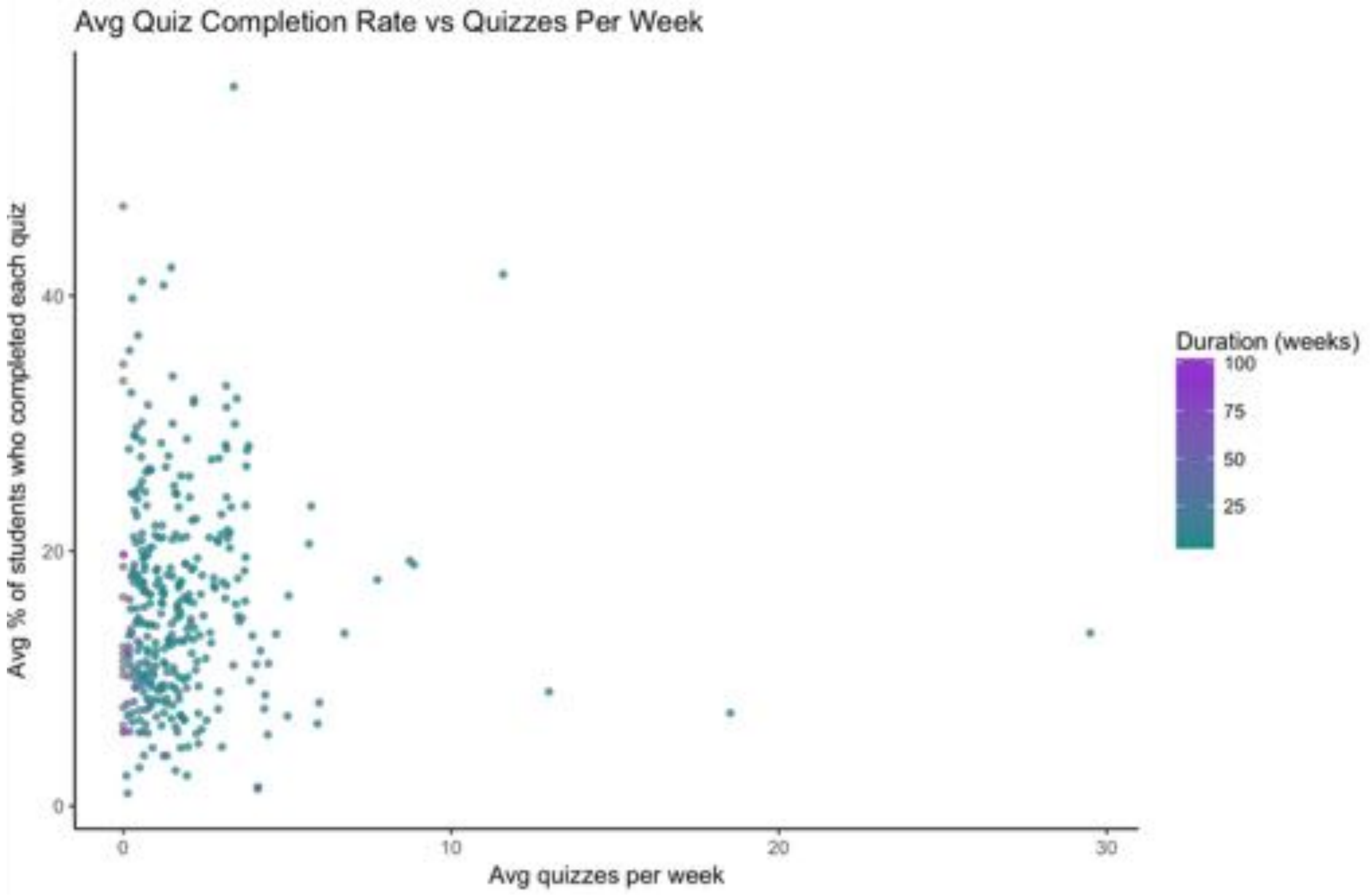



Avg % of quizzes a student completed per course vs. courses per student



# Data Visualizations

# Data Visualizations



# Weekly Interaction Clustering

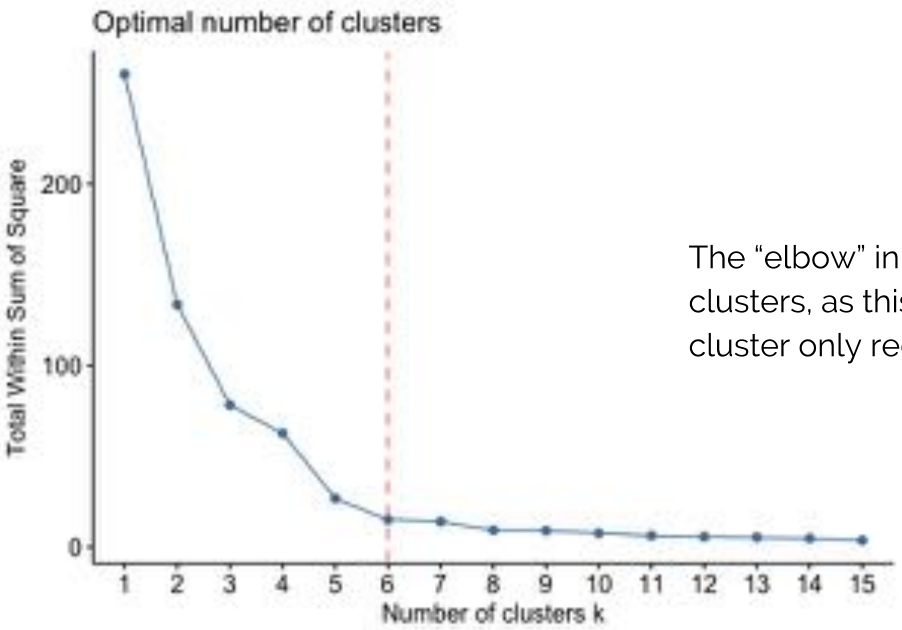
| user id | course id | week | count social | count quiz subs | count other | label |
|---|---|---|---|---|---|---|
| 876763763 | 3425142 | 23 | 0 | 8 | 1 | ? |
| 876763763 | 3425142 | 25 | 2 | 0 | 0 | ? |
| 876763763 | 9812343 | 5 | 4 | 2 | 0 | ? |
| 892332345 | 1434241 | 57 | 3 | 5 | 2 | ? |
| ... | ... | ... | ... | ... | ... | ... |

n = 480,000

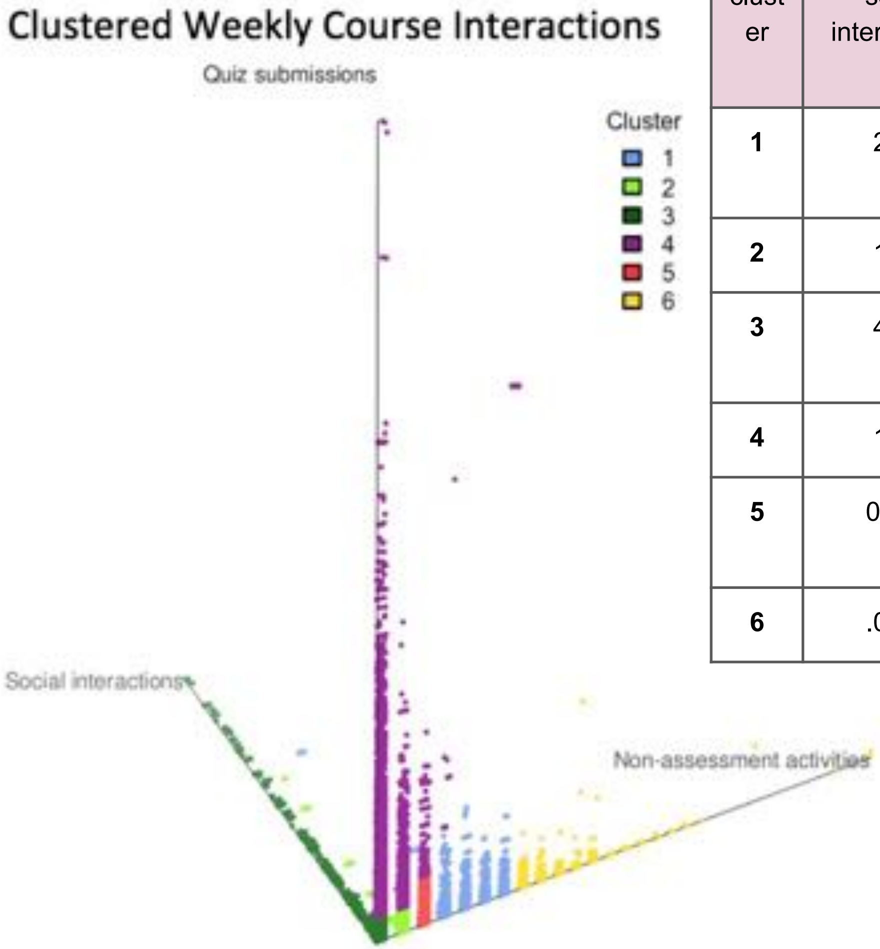
CLARA (**C**lustering **Lar**ge **A**pplications)

- Draw a random sample D' from the original dataset D
- Apply PAM (partitioning around medoids) algorithm to D' to find the k medoids
- Use these k medoids and the dataset D to calculate the current dissimilarity
- If it is smaller than the one you get in the previous iteration, then these k medoids are kept as the best k medoids
- The whole process is performed a specified number of times
- In this case, I used 5,000 samples of size 10,000

# Weekly Interaction Clustering




The “elbow” in the plot suggests an optimal number of clusters, as this is the point where each additional cluster only reduces SSE by a small amount.

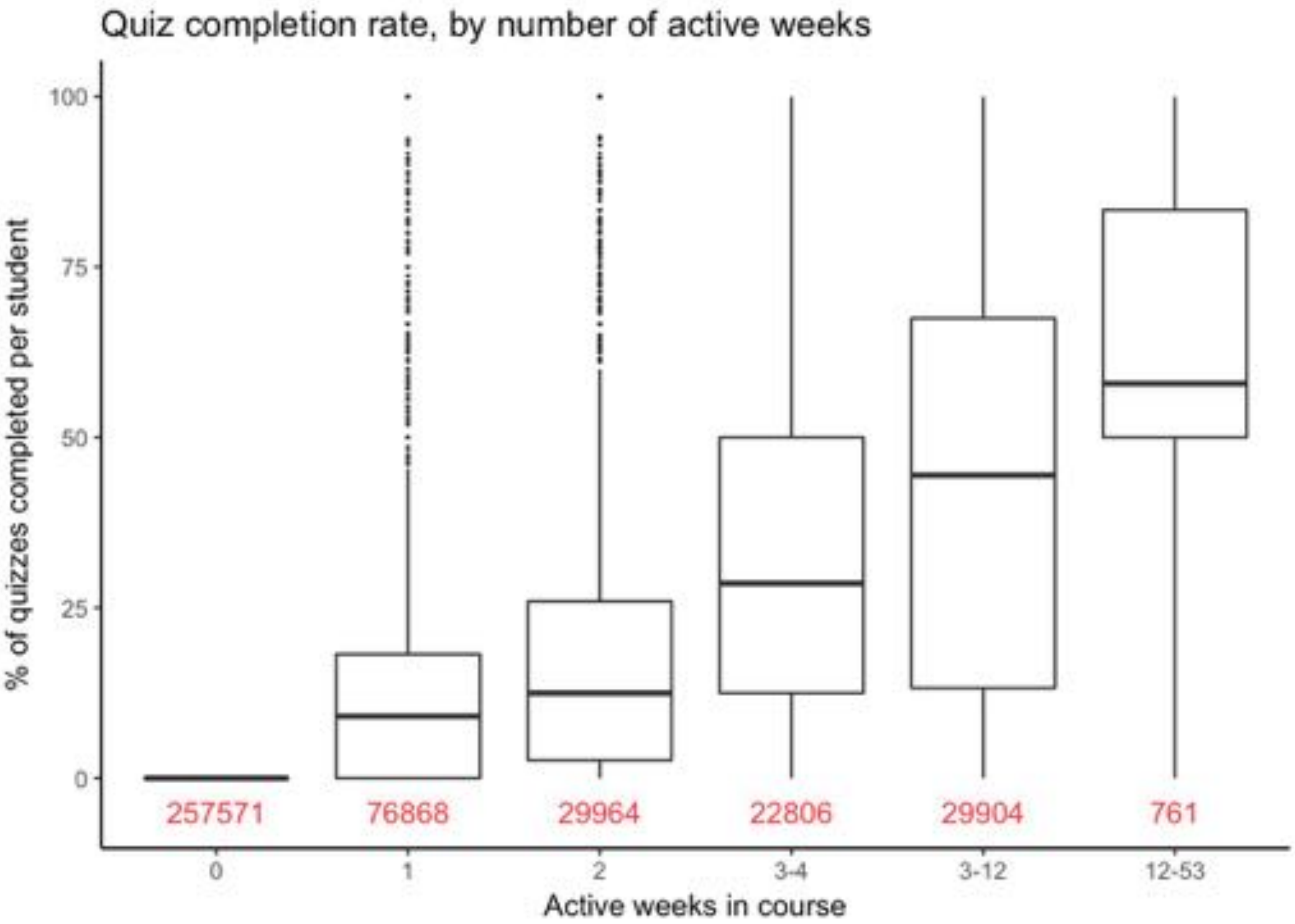



| cluster | social interactions | quiz submissions | non-assessment activities | label + interpretation |
|---|---|---|---|---|
| **1** | 2.99 | 0.217 | 3.72 | M – moderate activity in all three features |
| **2** | 1.13 | 0.440 | 1.00 | L – low activity in all three features |
| **3** | 4.15 | 1.94 | 0.00 | S – mostly social interaction; no non-assessment activity |
| **4** | 1.50 | 10.7 | 0.057 | Q – most quiz submissions |
| **5** | 0.286 | 0.252 | 2.00 | A – moderate non-assessment; low quiz and social |
| **6** | .0146 | 0.239 | 8.74 | N – mostly non-assessment activity |

| user id | course id | week | count social | count quiz subs | count other | label |
|---|---|---|---|---|---|---|
| 876763763 | 3425142 | 23 | 0 | 8 | 1 | Q |
| 876763763 | 3425142 | 25 | 2 | 0 | 0 | S |
| 876763763 | 9812343 | 5 | 4 | 2 | 0 | S |
| 892332345 | 1434241 | 57 | 3 | 5 | 2 | M |

Creation of interaction string for each user-course. For weeks with no interaction, E represents 'enrolled in course but didn't interact with it' and O represents weeks before or after the course's official start/end dates.

| user id | course id | engagement string |
|---|---|---|
| 876763763 | 3425142 | OOOEEEEEEMLQSEAENOOOO… |
| 876763763 | 9812343 | OOOQEEANEEOOOOOOOOOO…. |
| 892332345 | 1434241 | OOOOOSSEEEELLEEOOOOOO… |


Quiz completion rate, by number of active weeks
% of quizzes completed per student
100
75
50
25
0
257571
76868
29964
22806
29904
761
0
1
2
3-4
3-12
12-53
Active weeks in course

# Early Warning Approach

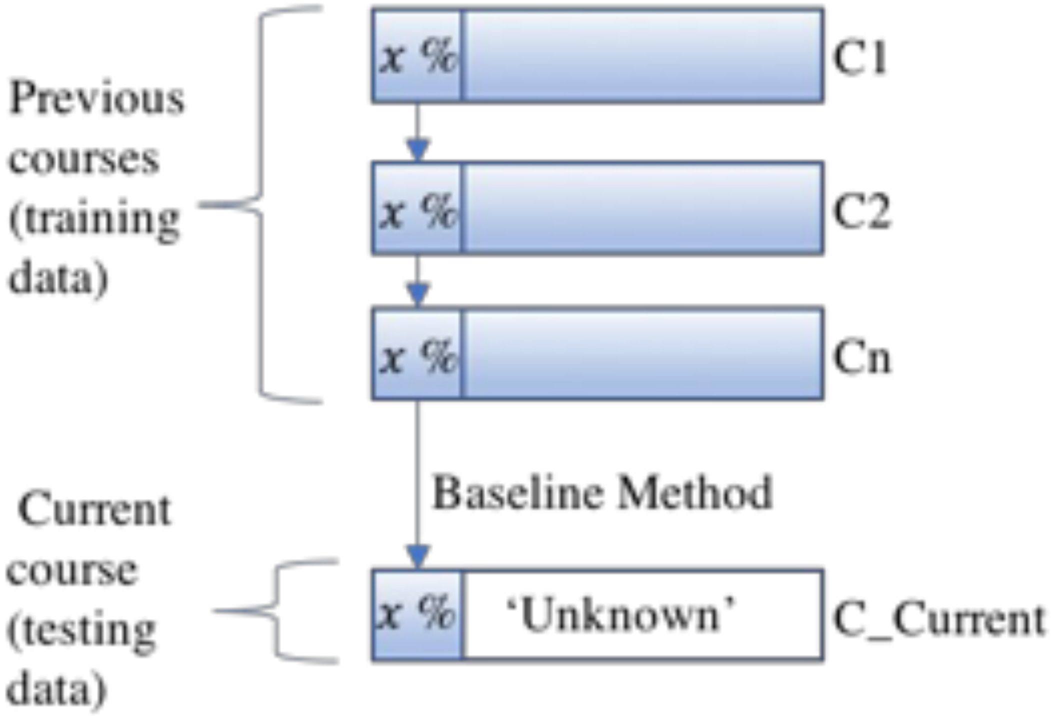
1: Nationally, the average 6-year graduation rate is 60\%.

2: In universities or online courses with high enrollment, faculty and advisors are unaware of the challenges faced by students until the end of the semester.

3: Students without up-to-date help would fail in classes and can't graduate on time.

4: An early warning approach is a tool that can help instructors to identify students at-risk of receiving poor grades


100%
X%
10%
Request_Times
Asst_Submission
Quiz_Submission
300
250
200
150
100
50
80%
70%
100%
20%
100%
100%
Apr 5
2015
Apr 12
Apr 19
Apr 26
May 3
May 10
May 17
May 24
May 31


Previous courses (training data)
x %
C1
x %
C2
x %
Cn
Baseline Method
Current course (testing data)
x %
'Unknown'
C_Current

# Feature Description (Course Feature)

CourseLen:  How long a course is.

Type: There have 12 different discipline courses in database.

Size: denoted how many students register for this course.

#Q: The total number of quizzes of a course.

#A: The total number of assignment of a course

# Feature Description (Student Feature)

QSubmission: How many quiz submissions of a student made before a specific timing.

QScore: How many scores student earned based on the submitted quiz and normalized the value by comparing the average quiz score of the class.

QAttempt: The average attempts times of the submitted quiz made by one student.

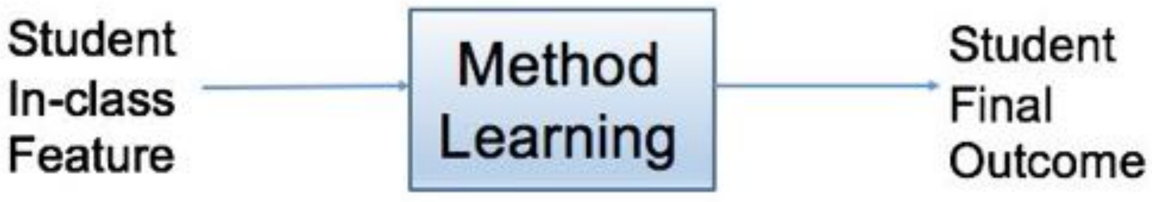
QTime: The average spending time of the summited quiz made by one student.

ASubmission: Same with QSubmission

AScore: Same with QScore

Acperday: How many times a student access to course management system

# Basic Framework



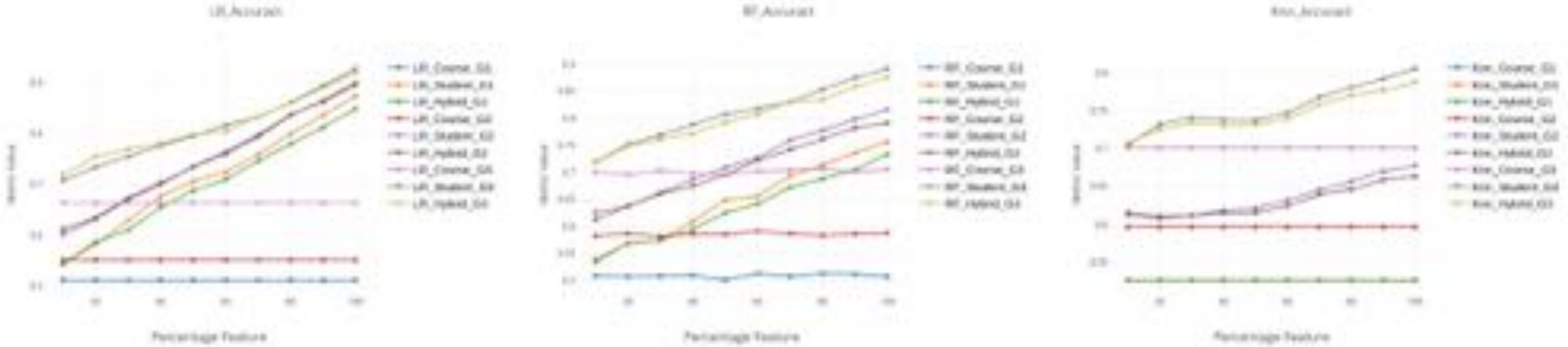
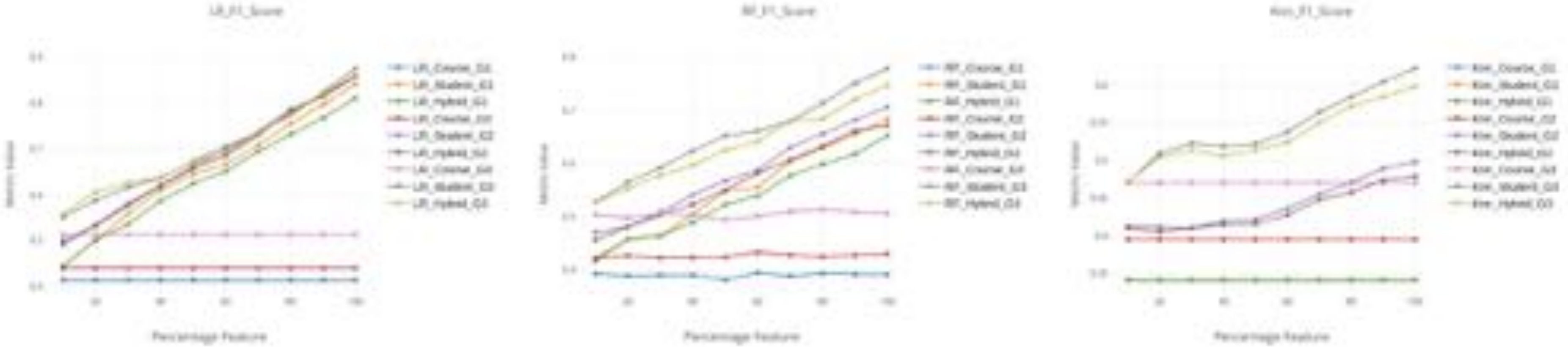


**Figure 10: Avaerage accuracy using course, student, hybrid features respectively for three different classification method.**



| Time-Stamp | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| **LR_C** | 0.44 | 0.44 | 0.44 | 0.44 | 0.44 | 0.44 | 0.44 | 0.44 | 0.44 | 0.44 |
| **LR_S** | 0.48 | 0.527 | **0.576** | **0.619** | **0.657** | **0.685** | **0.724** | **0.771** | **0.809** | **0.853** |
| **LR_H** | **0.485** | **0.53** | 0.567 | 0.608 | 0.645 | 0.672 | 0.716 | 0.759 | 0.796 | 0.84 |
| **KNN_C** | 0.384 | 0.384 | 0.384 | 0.384 | 0.384 | 0.384 | 0.384 | 0.384 | 0.384 | 0.384 |
| **KNN_S** | 0.391 | 0.396 | 0.398 | 0.401 | 0.403 | 0.411 | 0.423 | 0.433 | 0.444 | 0.45 |
| **KNN_H** | 0.39 | 0.393 | 0.396 | 0.397 | 0.399 | 0.405 | 0.418 | 0.425 | 0.434 | 0.438 |
| **RF_C** | 0.425 | 0.424 | 0.426 | 0.424 | 0.422 | 0.424 | 0.428 | 0.427 | 0.423 | 0.424 |
| **RF_S** | 0.456 | 0.485 | 0.508 | 0.538 | 0.565 | 0.592 | 0.624 | 0.656 | 0.68 | 0.707 |
| **RF_A** | 0.455 | 0.48 | 0.499 | 0.514 | 0.545 | 0.567 | 0.603 | 0.621 | 0.648 | 0.667 |

**Table 3: Average F1 Score of 586 students**

| Time-Stamp | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| **LR_C** | 0.551 | 0.551 | 0.551 | 0.551 | 0.551 | 0.551 | 0.551 | 0.551 | 0.551 | 0.551 |
| **LR_S** | 0.592 | 0.626 | **0.665** | **0.701** | **0.731** | **0.755** | **0.786** | **0.824** | **0.856** | **0.891** |
| **LR_H** | **0.599** | **0.631** | 0.66 | 0.694 | 0.724 | 0.745 | 0.779 | 0.815 | 0.846 | 0.88 |
| **KNN_C** | 0.583 | 0.583 | 0.583 | 0.583 | 0.583 | 0.583 | 0.583 | 0.583 | 0.583 | 0.583 |
| **KNN_S** | 0.591 | 0.594 | 0.595 | 0.597 | 0.598 | 0.605 | 0.613 | 0.62 | 0.627 | 0.632 |
| **KNN_H** | 0.59 | 0.591 | 0.594 | 0.595 | 0.595 | 0.6 | 0.609 | 0.614 | 0.62 | 0.624 |
| **RF_C** | 0.572 | 0.572 | 0.572 | 0.57 | 0.569 | 0.572 | 0.574 | 0.574 | 0.568 | 0.569 |
| **RF_S** | 0.601 | 0.627 | 0.646 | 0.67 | 0.693 | 0.715 | 0.738 | 0.765 | 0.783 | 0.802 |
| **RF_H** | 0.599 | 0.622 | 0.64 | 0.652 | 0.676 | 0.697 | 0.725 | 0.739 | 0.76 | 0.772 |

**Table 4: Average F1 Score of 586 students**


90% student (training data)
x %
TR_S 1
x %
TR_S2
x %
TR_Sn
Baseline Method
10% student (testing data)
x %
'Unknown'
TE_S1…..TE_Si

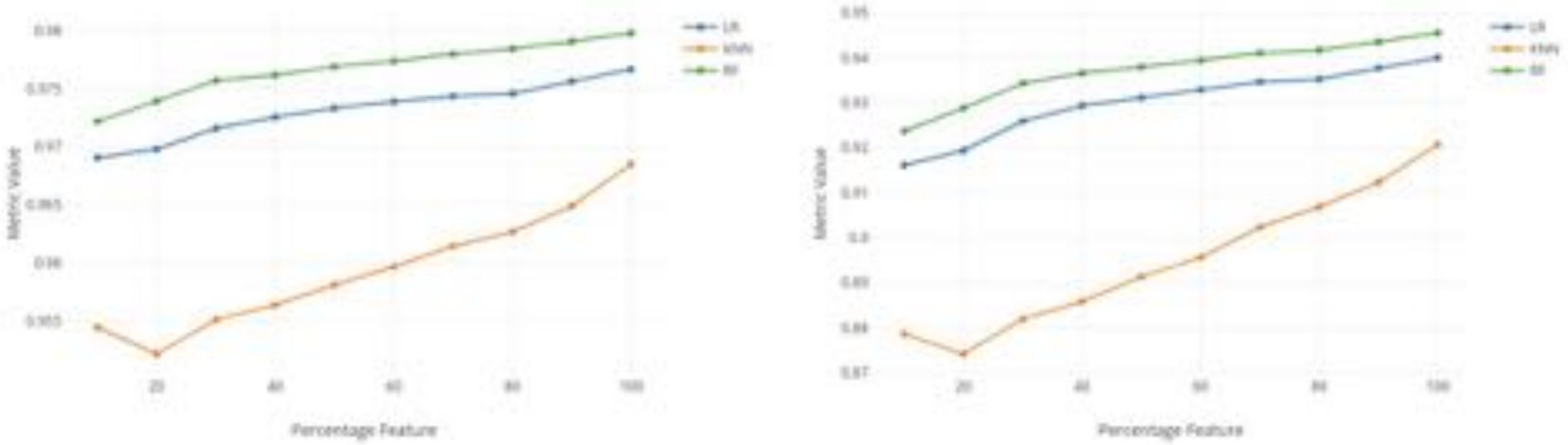



**Figure 12: Average accuracy and F1 score result for Course-Specific-Approach**

# Full Paper Available

# Designing Early Warning Approach using Student's Early In-class Study Behavior

Zhouxiang Cai
George Mason University
Fairfax, Virginia
zcai2@gmu.edu

Huzefa Rangwala
George Mason University
Fairfax, Virginia
rangwala@cs.gmu.edu

## ABSTRACT

Nationally, the average 6-year graduation rate is 60%. In universities or online courses with high enrollment, faculty and advisors are unaware of the challenges faced by students until the end of the semester. Students without up-to-date help would fail in classes and can't graduate on time. It is essential to find an approach that detects at-risk students before issues worsen in their college life. An early warning approach is a tool that can help instructors to identify students at-risk of receiving poor grades analyzing student study features recorded in course management system (CMS) such as Blackboard and Canvas. We used several machine learning methods such as logistic regression (LR), random forests (RF), and k-nearest neighbors algorithm (KNN) to achieve our goal. We perform our comprehensive evaluation of de-identified data obtained from Canvas Network open courses, which have sufficient classes to solve the issue that the absence of training data happened in other scholar's studies. This study introduces two early warning approaches to support advisors and university administrators to classify at-risk students. Our experimental results show that we are able to predict the student final learning outcomes with high accuracy based on two early warning approaches. We also help identically essential features within a course found on the different stages of the course.

## CCS CONCEPTS

•**Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; •**Networks** → Network reliability;

## KEYWORDS

Early Warning, Learning Analytics, Regression, Classification, Early Feature, Student Behavior

**ACM Reference format:**
Zhouxiang Cai and Huzefa Rangwala. 2019. Designing Early Warning Approach using Student's Early In-class Study Behavior. In *Proceedings of LAK'19 International Conference on Learning Analytics and Knowledge, Temple, Arizona, USA, March 7-9 (LAK'19)*, 10 pages.
DOI: 10.475/123_4



## 1 INTRODUCTION

The environment of educational institutions today is more complicated. As student enrollments in higher education expansion, class size along with the heterogeneity of students, such as virtual or face to face, traditional or non-traditional [10]. With such striking growing enrollment numbers, educators must ensure the learning outcomes are not affected. However, according to the National Center for Education Statistics [14], more than 41% of students who attended a four-year undergraduate program in Fall 2009 failed to graduate within six years. Schneider and Yin [18] calculated the hidden cost for college dropouts from just a single cohort of entering students lost is $3.8 billion [21].

CMS provided students and instructors a convenient way to overcome the limitation of space and time. Educators can assess overall learning performances, and determine how considerably students are learning from the course and what academic challenges they might be facing [3, 8, 13]. There have various engagement features associated with the course offering such as the time of studying chapters, completing quizzes and assignments, etc. [16]. By evaluating these student interactions as well as course information, the educational researcher can find latent student study behavior features. An educator can reach these data to determine the student learning outcome and then to provide timely feedback and interventions [11]. So it is significant to understand what variables features would affect the outcome of university courses. Researchers usually use data mining to interpret these data in CMS. Data mining is referred to as knowledge discovery involves methods that search for hidden data relationship or classification dataset [4]. Educational Data Mining (EDM) has been applied to understand latent student learning behavior better and help them to succeed, and there are increasing research interests in using (EDM) [1].

We extract multiple features to identify the learning behavior of different students in individual courses. These features can show students' engagement effort as well as course feature and feature detailed description writes in Section 4.2. In Section 6.4, we also use lasso regression to evaluate the feature importance. In this paper, we implemented two early warning approaches to identify at-risk students. Specifically, we named **Student-Specific approach** and **Course-Specific approach**. In the first method, we looked into a student's course history and based on these course history features to predict the current course's final grade. We implemented our Course-Specific approach that we randomly pick a course and try to find out the earlier offered version of that course. We used already ended course as the training data to predict current semester performance of the student.



## 2 LITERATURE REVIEW

Predicting students' learning outcomes is more and more popular in EDM. Several papers have focused on the analysis and predict student's in-class performance based on student's social and learning features. Romero et al. [17] evaluate different data mining approaches to classify students based on their CMS usage data. Ren et al. [16] predicated on the assessment performance using multi-regression models and it provided a way to collect features about student's interaction between Massive Open Online Courses (MOOCs). Devasia et al. [5] predicted students' performance by analyzing the students' social features such as gender and living habits. Instead of focuses on graded learning features such as assignments and quizzes, Sahebi and Brusilovsky [19] took advantage of students' non-graded activities features, and their approach could reduce the error of student performance prediction.

Besides thinking about the student in-class performance characteristic, an understanding of suitable approaches or theories of learning analytics is also necessary for examining learning behavior [12]. Pittman [15] have compared data mining techniques used to predict student retention and concluded that logistic regression would be an optimal tool. Boroujeni and Dillenbourg [2] discovered the possible study pattern based on the MOOC interaction sequence and found the study pattern transition probabilities for learners. Zhang and Rangwala [21] developed an Iterative Logistic Regression (ILR) method to address the challenge of early prediction and got a much better result than standard logistic regression.

In this paper, we study the application of early warning technology to student grade prediction. Similar performance warning techniques have been explored. Jiang et al. [9] used a combination of students' first-week assignment performance and social interaction within the MOOC to predict their final performance by logistic regression. He et al. [6] investigated the early and accurate prediction of students at risk of failing a MOOC by evaluating on multiple offerings and under potentially non-stationary data. They build predictive models weekly based on the numerous offerings of a course. Jokhan et al. [10] designed an early warning system based on the students feature such as gender, age, student status and engagement feature and achieve 60.8% accuracy based on that model. Due to the absence of data from previous classes, Hlosta et al. [7] developed a 'self-learner' method that used current course data as the training set to identify at-risk students.

Our study is going to make up for the two shortcomings in other scholars' studies. (1): few studies provide a flexible way for educators to set a modifiable timing parameter for collecting student features. Previous studies only focused on the first few weeks' features to predict students learning outcomes; however, in real universities, instructors might need to give students feedback in any time they want such as withdraw deadline or mid-term report period. Our approaches provide a dynamic way to meet this requirement. (2): the studies not always have large enough course dataset as training data to make more persuasive experiments. As a result, these have little research focused on the early student feature and based on their course history to identify at-risk students. In MOOCs, studies on predicting student performance often seek to predict grade for homework and quizzes in a single course [20]. Even though scholars explored deep and various features such as video, quiz, homework engagement from one single course, experiments were done in a single course is not too comprehensive and reliable. The previously taken courses have the complete features, and they are excellent training data to predict the final learning outcomes of a student's current course. Our dataset has more than 300 courses to provide more compelling results.

## 3 PROBLEM SPECIFICATION

Student's engagement features could present as a time-series data. Figure 1 shows a typically student's engagement time-series data, and in Figure 1, we can view students' various activity with specific timestamps. The dots in time-series means the submission of quizzes and assignments made by this student and the percentage value means the score he/she earned. Based on time-series data characteristics, we can set up a specific timing and only extract the feature before that timing. In other words, we can set the parameter $X$ to whatever number we want. It could be the first two weeks, first 24 hours or a drop deadline for a course. For example, for a course C1, if we set the parameter $X$ to 0.1, which means the first 10% of the feature for class $C1$ and shows as 10% in Figure 1. In our study, we set X to a small value to catch student's feature at the beginning of the course, and we defined as **Early Feature**.

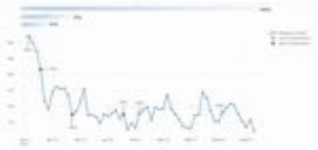


**Figure 1: A sample student engagement time-series data**

Given a database about the history course of a student, our object is to identify if the student will perform well in the current class. More formally, we have a student with N courses marked as $\{C_1,...,C_n\}$ and sets of these previous course marked as $C_{previous}$ Each session associated with some features. We noted features for each class as $\{F_1,...,F_n\}$ and each element is a 1-D array. We know the complete study feature as well as the final grade for each course since it's a previous course. The current class marked as $C_{current}$ and marked its feature as $F_current$ We have the limited information for $C_{current}$ because $F_current$ is an on-going class. To keep the training and testing data consistent, we should limit $C_{previous}$ features to the same timeline as $C_{current}$. By applying timestamp parameter X to $C_{previous}$, we could get an early feature defined as $\{X^*F_1,...,X^*F_n\}$. We regard these feature as a training set and $F_{current}$ as a testing set. Then we put both training set and testing set into the machine learning model to get a binary output, which means 0 is passing, and one is failing.

Educational institution will offer the same course in different semesters. We also consider identifying possible risking student in

Thanks for this summer