Approaches to Scalable Personal Guidance in MOOCs and On Campus

Zachary A. Pardos
Assistant Professor

Graduate School of Education
School of Information

zachpardos.com
zp@berkeley.edu

BAYLAN
Bay Area Learning Analytics Network

CAHL

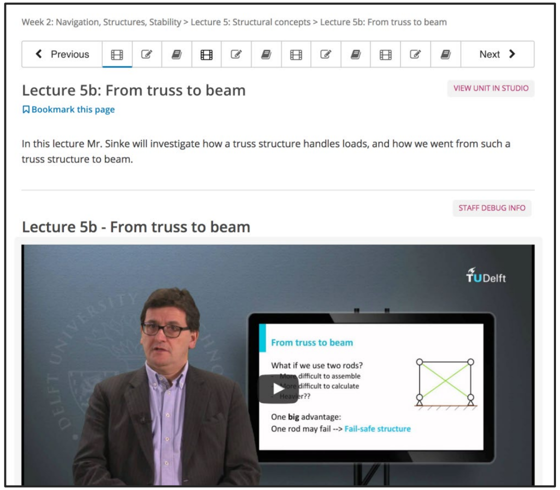
Computational Approaches to
Human Learning (CAHL) research lab

GRADUATE SCHOOL OF EDUCATION

UC Berkeley School of Information

Contexts

moocs



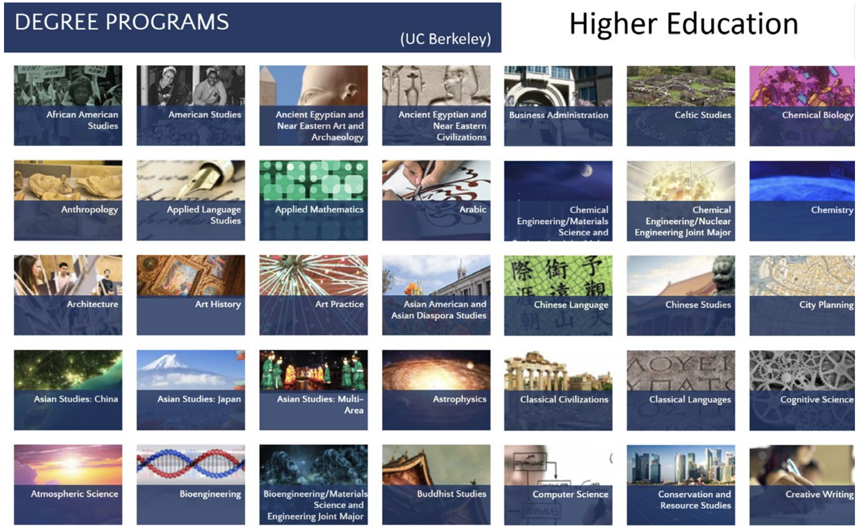
Week 2: Navigation, Structures, Stability > Lecture 5: Structural concepts > Lecture 5b: From truss to beam

Lecture 5b: From truss to beam

In this lecture Mr. Sinke will investigate how a truss structure handles loads, and how we went from such a truss structure to beam.

Lecture 5b - From truss to beam

TUDelft
From truss to beam
What if we use two rods?
- More difficult to assemble
- More difficult to calculate
- Heavier??
One big advantage:
One rod may fail --> Fail-safe structure



DEGREE PROGRAMS (UC Berkeley)

Higher Education

African American Studies	American Studies	Ancient Egyptian and Near Eastern Art and Archaeology	Ancient Egyptian and Near Eastern Civilizations	Business Administration	Celtic Studies	Chemical Biology
Anthropology	Applied Language Studies	Applied Mathematics	Arabic	Chemical Engineering/Materials Science and	Chemical Engineering/Nuclear Engineering Joint Major	Chemistry
Architecture	Art History	Art Practice	Asian American and Asian Diaspora Studies	Chinese Language	Chinese Studies	City Planning
Asian Studies: China	Asian Studies: Japan	Asian Studies: Multi-Area	Astrophysics	Classical Civilizations	Classical Languages	Cognitive Science
Atmospheric Science	Bioengineering	Bioengineering/Materials Science and Engineering Joint Major	Buddhist Studies	Computer Science	Conservation and Resource Studies	Creative Writing

- Large scale enrollment
- Evolution of the textbook
- Access ≠ Success
- “Low touch”

- High degree requirement complexity
- Many course options (~2,500 / semester)
- 40% 4-year graduation rate (U.S.)
- 1:400 student:adviser ratio nationwide

Can analytics help scale guidance in these contexts?

Scaling personalized guidance using…

online course data (MOOC clickstream sequences)

play_video_1, pause_video_1, answer_Q2_correct, load_page2, play_video_2 pause_video2

Scaling Instructor Personalization in a MOOC

Paper link: http://tiny.cc/**aied_communication_paper**

Christopher Vu Le
Zachary A. Pardos
Samuel D. Meyer
Rachel Thorp

University of California at Berkeley

THE UNIVERSITY OF CALIFORNIA · BERKELEY · LET THERE BE LIGHT · 1868

AIED 2018

Berkeley | EECS
Electrical Engineering and Computer Sciences

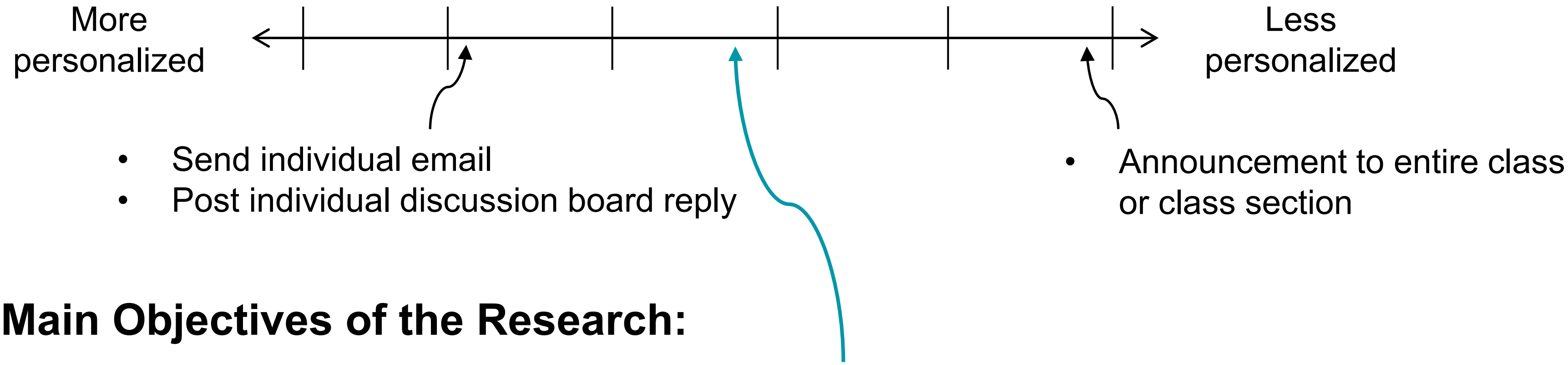
CAHL Computational Approaches to Human Learning (CAHL) research lab

GRADUATE SCHOOL OF EDUCATION

UC Berkeley School of Information

One-on-one instructor communication is scarce in “at scale” classrooms

Communication options for online instructors:



More personalized

Less personalized

- Send individual email
- Post individual discussion board reply

- Announcement to entire class or class section

Main Objectives of the Research:

1. Provide instructors an intermediary level of personalized communication based on learners’ engagement analytics

2. Deploy a working instructor communications interface in an edX course with daily updated analytics as proof-of-concept

Related work on engagement (drop-out)

Drop-out interventions

- Drop-out survey as unintentional intervention (Whitehill et al., 2015)
- Peer social chat within a course (Ferschke, 2015)
- Early warning course drop-out system on-campus (Jayaprakash, 2014)

Drop-out prediction models

- Hidden Markov Models (B
- Support Vector Machines
- Logistic regression (Jiang
- Recurrent Neural Networks (Mi & Yeung, 2015) hand-engineered features
- Ensembles (Boyer & Veeramachaneni, 2016)

Additional Note on Motivation

“What good is prediction?”

Making predictive models useable in real-world contexts is as valuable an endeavor for the community as is discovery and data mining with those models

Drop-out model frameworks for evaluation/replication

- Drop-out prediction replication frameworks (Andres et al., 2017;Gardner & Brooks, 2018)

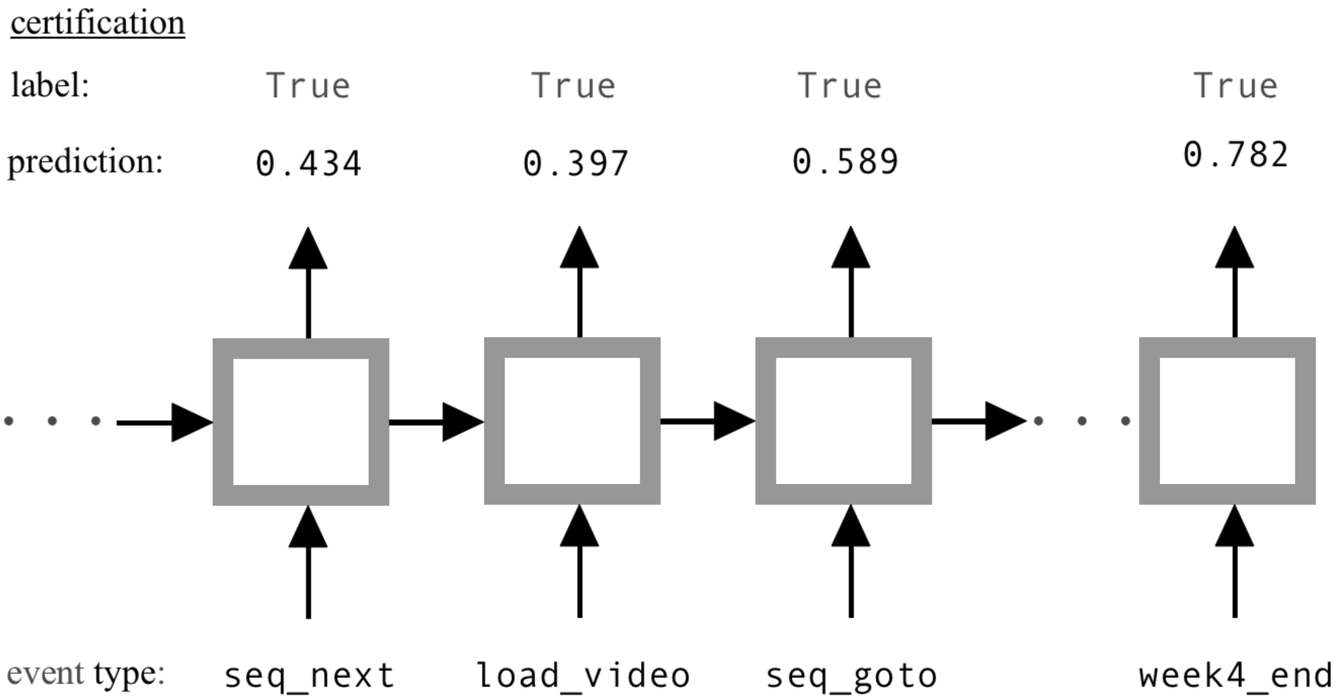
Our Methodology

1. **Evaluate past predictive models** + RNNs on large MOOC datasets

2. **Build an analytics back-end and front-end interface** in edX to surface predictions to instructors

3. **Allow email communications to be sent** based on these analytics

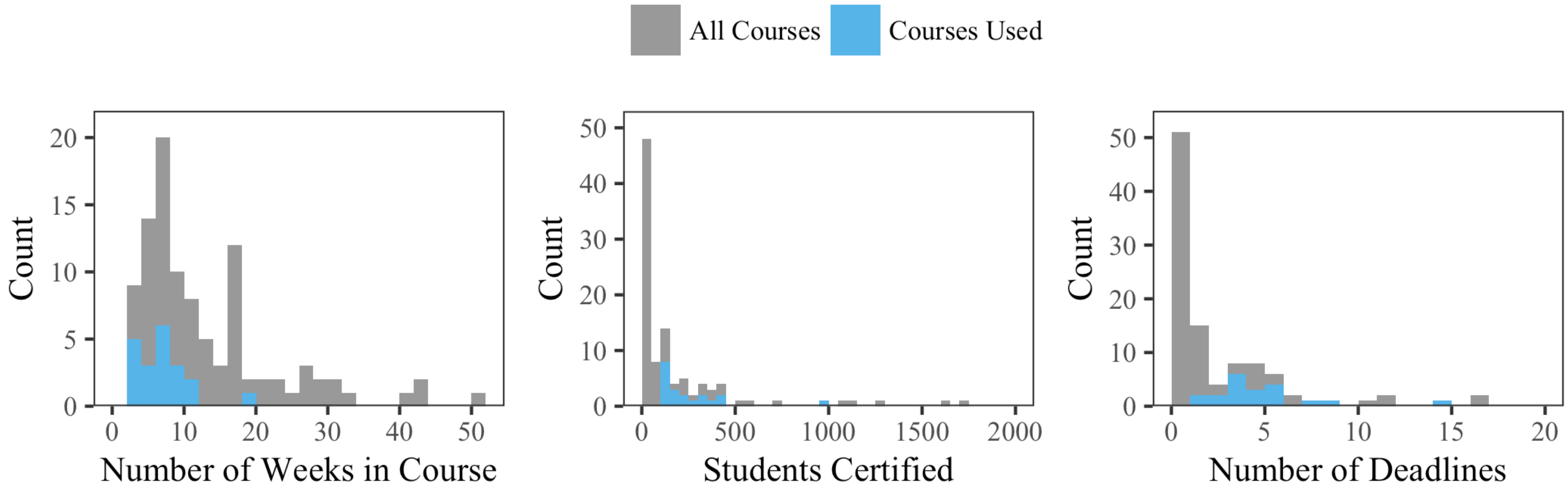
Model Inputs and Outputs
(neural network version)



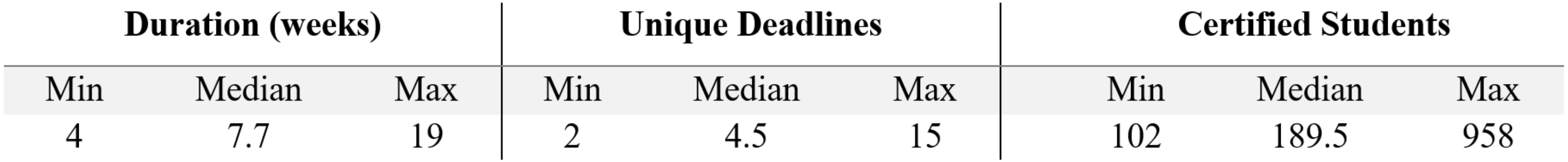
Dataset

Final set was 20 courses with 13.6 million clickstream events total

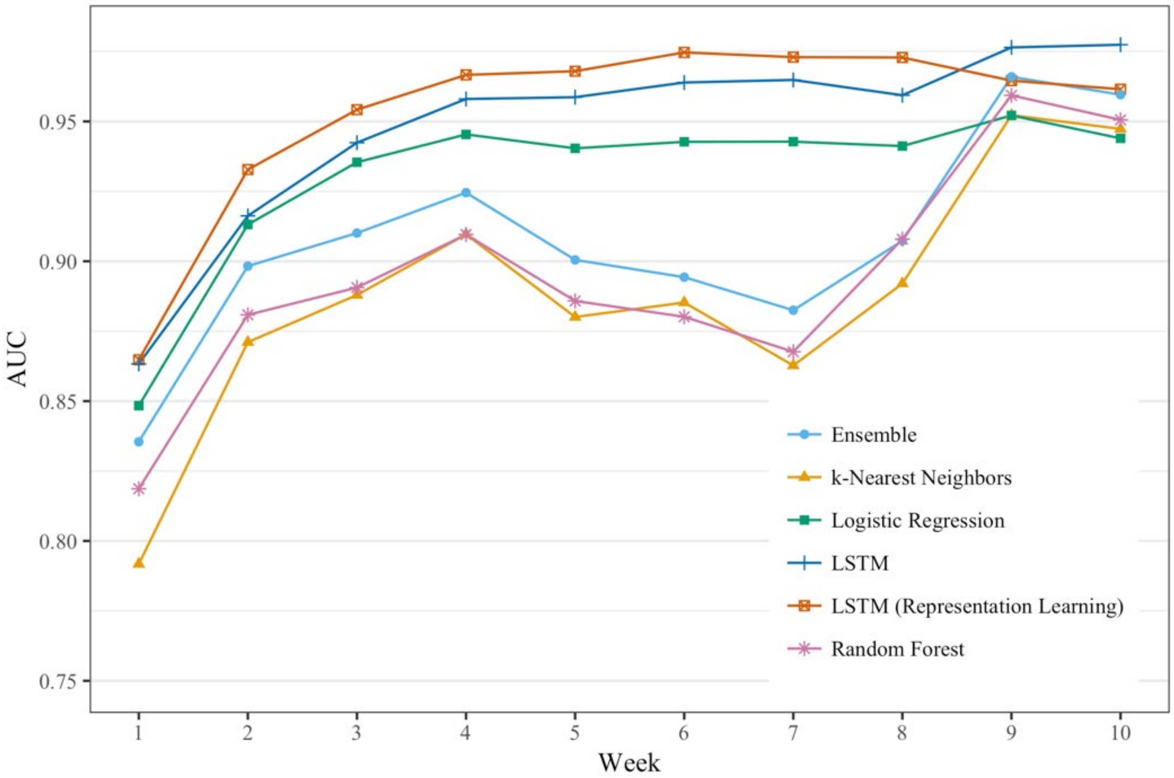
Comparison of distributions between the original 102 courses and the selected 20

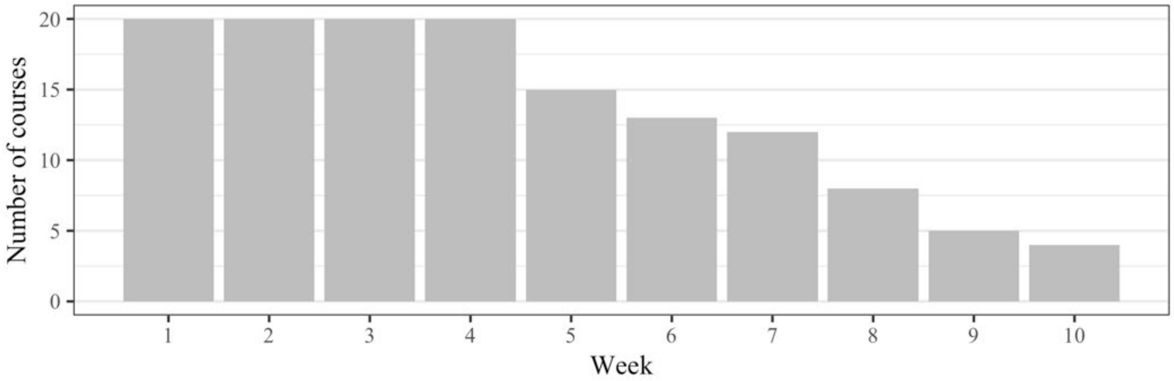


Descriptive statistics for the selected 20 courses



Duration (weeks)			Unique Deadlines			Certified Students		
Min	Median	Max	Min	Median	Max	Min	Median	Max
4	7.7	19	2	4.5	15	102	189.5	958



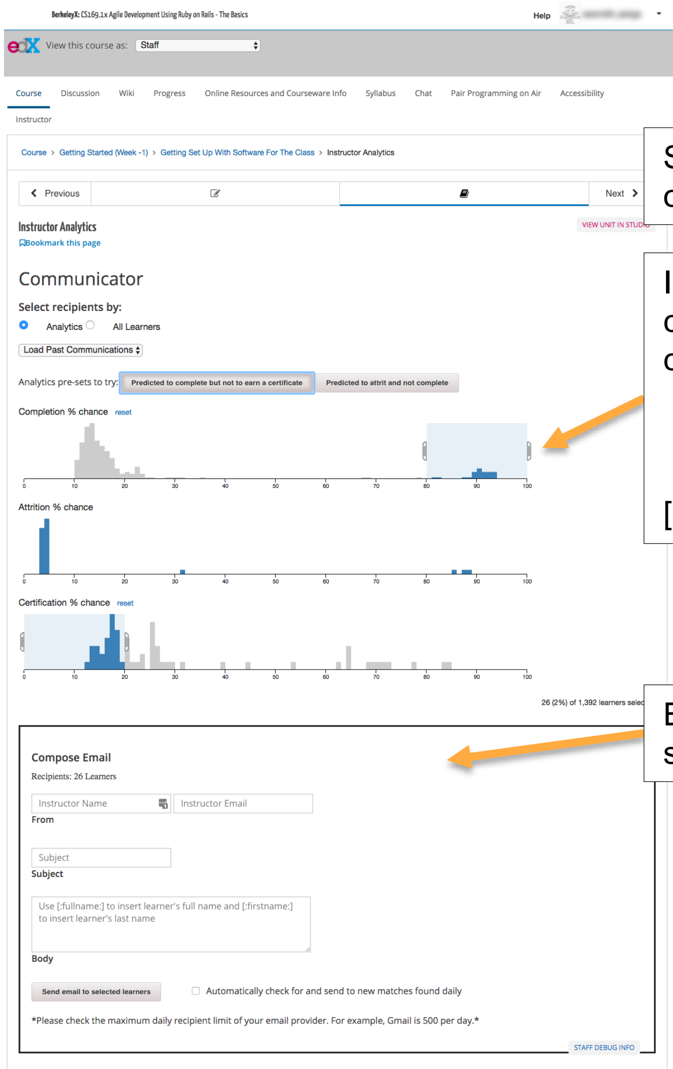


Prediction Results

(certification)

- 5-fold cross-validation
  (16 courses training, 4 testing)
- LSTM with representation learning outperformed all other approaches except for last two weeks ($p < 0.05$)
- Logistic regression better than non-RNN methods (including Ensemble)
- LSTM (representation learning) used for additional drop-out and completion outcome prediction models

Dashboard (front-end) Design



BerkeleyX: CS169.1x Agile Development Using Ruby on Rails - The Basics

Help

View this course as: Staff

Course | Discussion | Wiki | Progress | Online Resources and Courseware Info | Syllabus | Chat | Pair Programming on Air | Accessibility

Instructor

Course > Getting Started (Week -1) > Getting Set Up With Software For The Class > Instructor Analytics

< Previous | Next >

Instructor Analytics

Bookmark this page

VIEW UNIT IN STUDIO

Communicator

Select recipients by:

● Analytics ○ All Learners

Load Past Communications

Analytics pre-sets to try: Predicted to complete but not to earn a certificate | Predicted to attrit and not complete

Completion % chance reset

Attrition % chance

Certification % chance reset

26 (2%) of 1,392 learners selec

Compose Email

Recipients: 26 Learners

Instructor Name | Instructor Email

From

Subject

Subject

Use [:fullname:] to insert learner's full name and [:firstname:] to insert learner's last name

Body

Send email to selected learners

□ Automatically check for and send to new matches found daily

Please check the maximum daily recipient limit of your email provider. For example, Gmail is 500 per day.

STAFF DEBUG INFO

Student engagement analytics displayed on staff only viewable dashboard

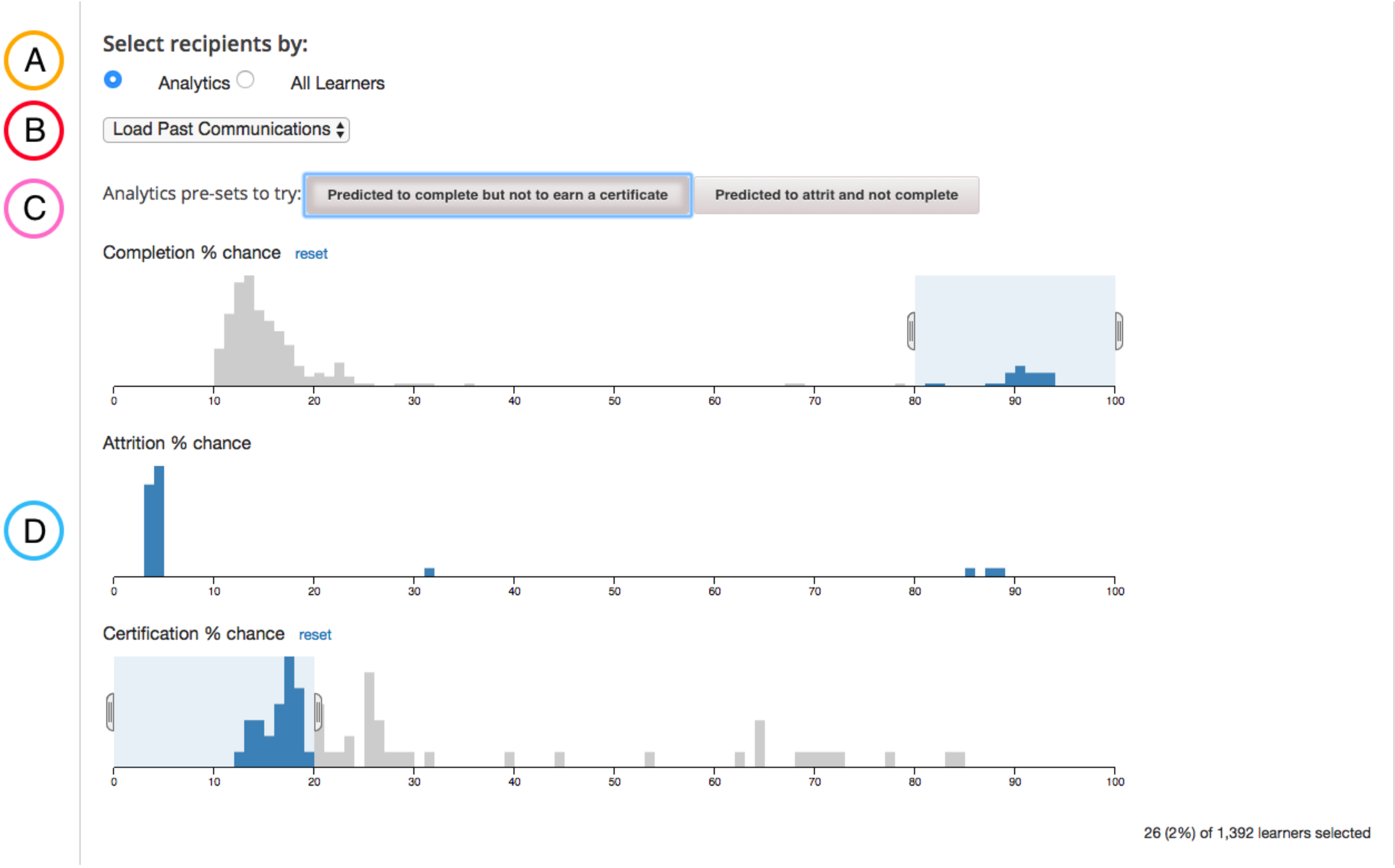
Instructor selects learners to communicate with based on analytics consisting of per-student predictions of:

- Completion
- Attrition
- Passing/Certification

[*generated from daily edX event logs*]

Email composed and sent to selected learners

Selection of recipients based on engagement analytics



Select recipients by:

● Analytics ○ All Learners

Load Past Communications ⬍

Analytics pre-sets to try: **Predicted to complete but not to earn a certificate** **Predicted to attrit and not complete**

Completion % chance reset

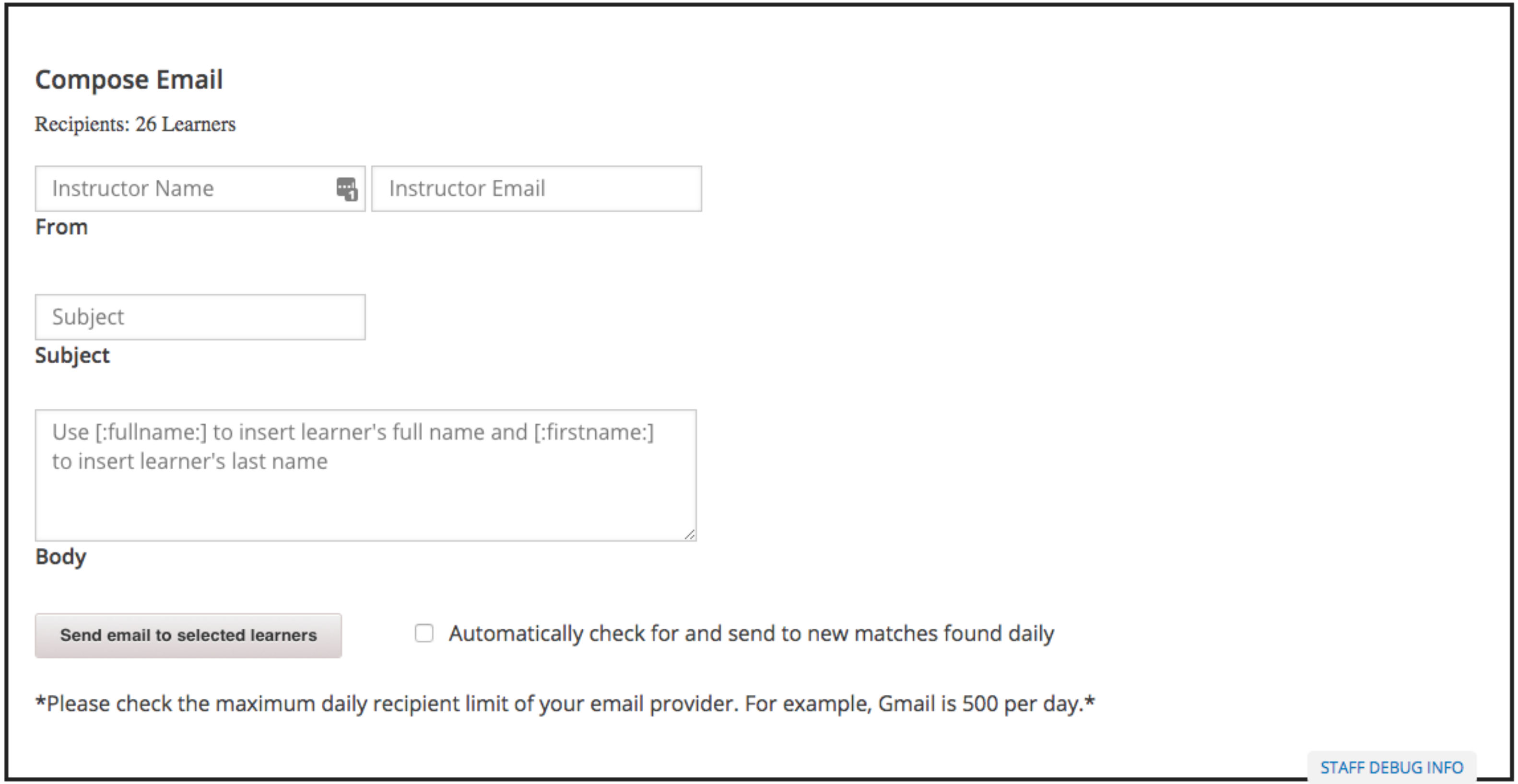
Attrition % chance

Certification % chance reset

26 (2%) of 1,392 learners selected

Composition of email to selected recipients

E



Compose Email

Recipients: 26 Learners

Instructor Name
Instructor Email

From

Subject

Use [:fullname:] to insert learner's full name and [:firstname:] to insert learner's last name

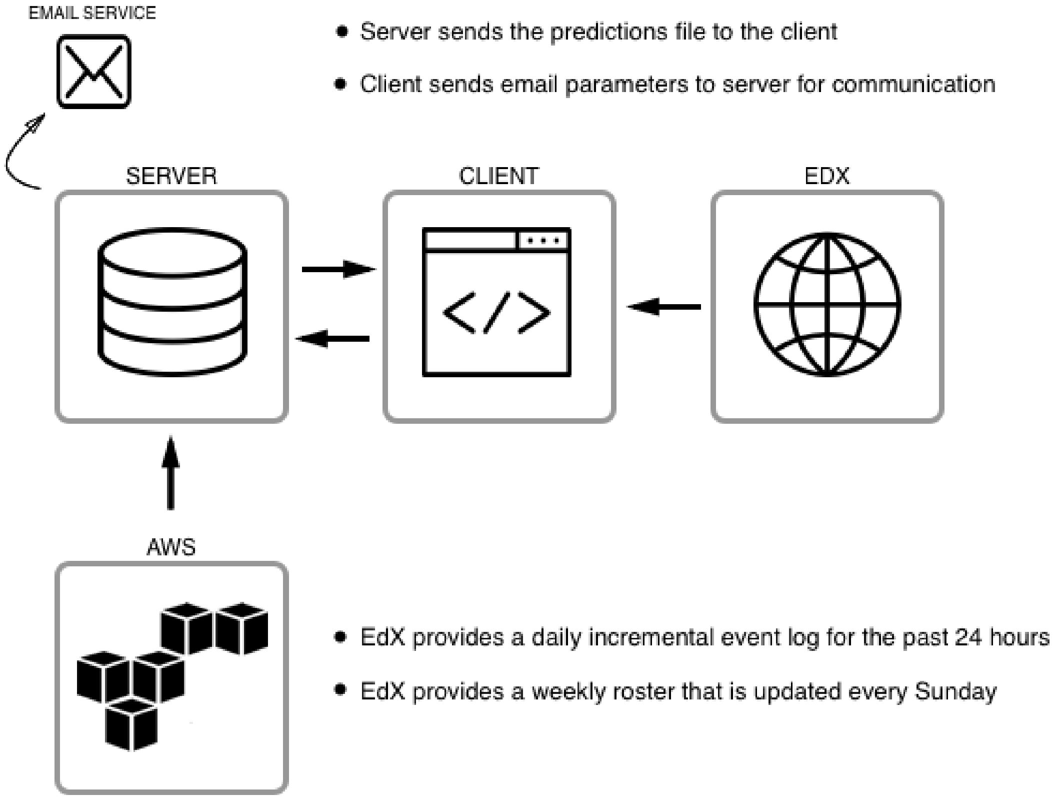
Body

Send email to selected learners

□ Automatically check for and send to new matches found daily

Please check the maximum daily recipient limit of your email provider. For example, Gmail is 500 per day.

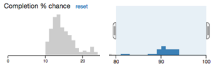
Engagement Analytics (back-end) API



- Server sends the predictions file to the client
- Client sends email parameters to server for communication

- EdX provides a daily incremental event log for the past 24 hours
- EdX provides a weekly roster that is updated every Sunday

Replication requirements

edX data assets	COMMUNICATOR 
Staff course access to edX studio to insert dashboard html into vertical	X
Daily event log from deployment course e.g. *berkeleyx-events-2018-06-05.log.gz*	X
Weekly roster from deployment course e.g. *BerkeleyX-CS169.2x-1T2018-auth_user-prod-analytics.sql*	X

https://github.com/CAHLR/Communicator

Interested in joining the open-edx pilot?

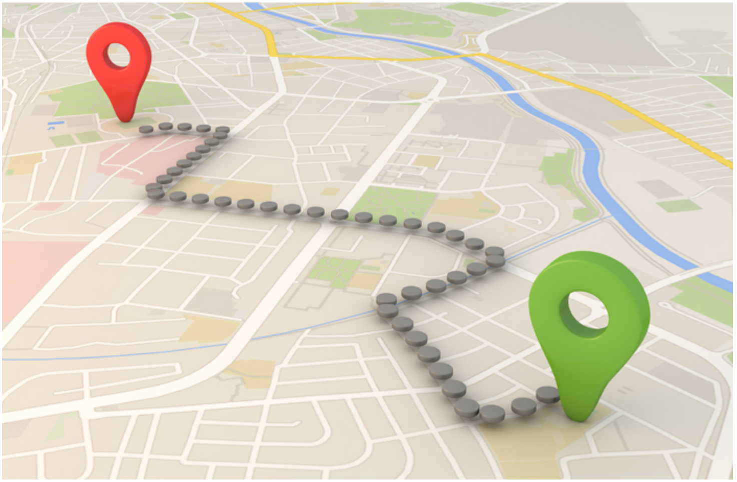
Send me a calendar invite: tiny.cc/zpUCB

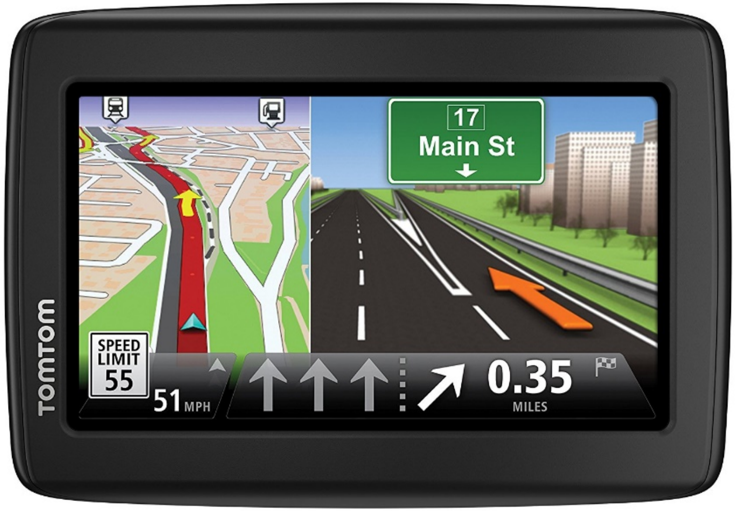
Scaling personalized guidance using…

University course selection (sequences of course enrollment)

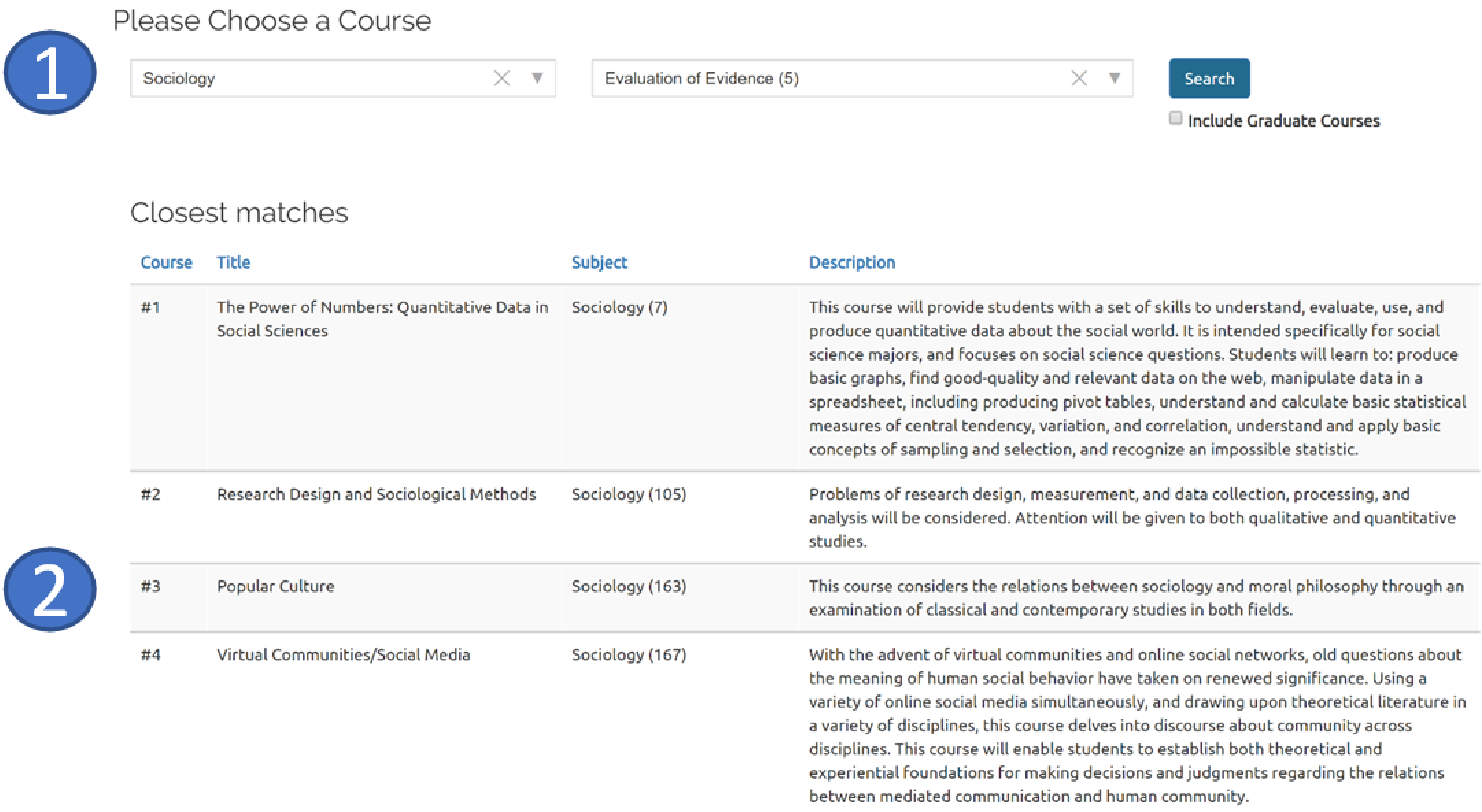
CS61A, MATH1B, SPA12, STAT200B, CUE100A, CS188, CS267, CS268, ENN1B

Information vs. Guidance





Facilitating Exploration



Please Choose a Course

Sociology ▾ | Evaluation of Evidence (5) ▾ | Search

□ Include Graduate Courses

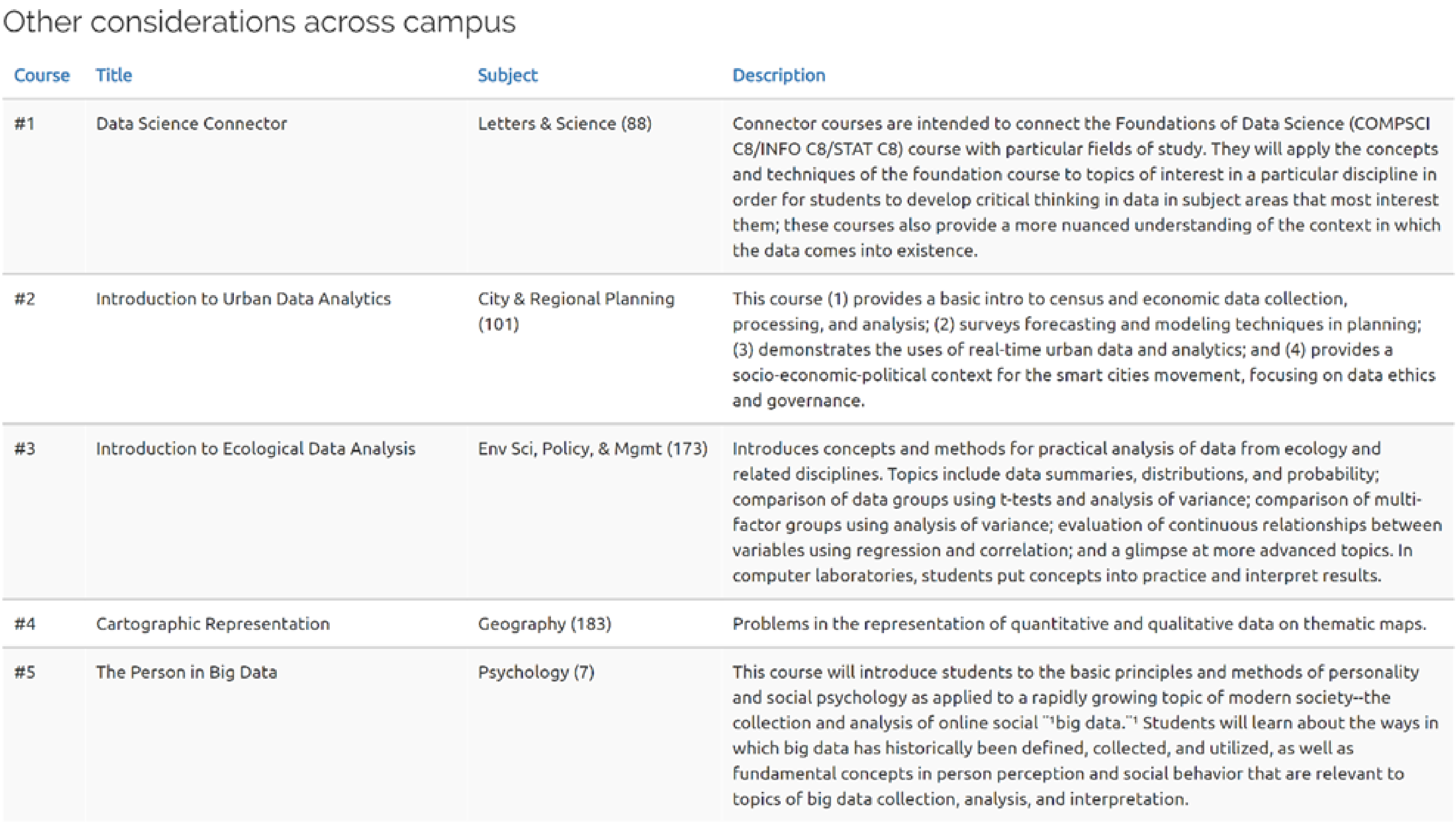
Closest matches

Course	Title	Subject	Description
#1	The Power of Numbers: Quantitative Data in Social Sciences	Sociology (7)	This course will provide students with a set of skills to understand, evaluate, use, and produce quantitative data about the social world. It is intended specifically for social science majors, and focuses on social science questions. Students will learn to: produce basic graphs, find good-quality and relevant data on the web, manipulate data in a spreadsheet, including producing pivot tables, understand and calculate basic statistical measures of central tendency, variation, and correlation, understand and apply basic concepts of sampling and selection, and recognize an impossible statistic.
#2	Research Design and Sociological Methods	Sociology (105)	Problems of research design, measurement, and data collection, processing, and analysis will be considered. Attention will be given to both qualitative and quantitative studies.
#3	Popular Culture	Sociology (163)	This course considers the relations between sociology and moral philosophy through an examination of classical and contemporary studies in both fields.
#4	Virtual Communities/Social Media	Sociology (167)	With the advent of virtual communities and online social networks, old questions about the meaning of human social behavior have taken on renewed significance. Using a variety of online social media simultaneously, and drawing upon theoretical literature in a variety of disciplines, this course delves into discourse about community across disciplines. This course will enable students to establish both theoretical and experiential foundations for making decisions and judgments regarding the relations between mediated communication and human community.

1. Student selects “Sociology 5: Evaluation of Evidence” as a favorite course
2. First, close course description matches to the selected course are shown

Facilitating Exploration

3



Other considerations across campus

Course	Title	Subject	Description
#1	Data Science Connector	Letters & Science (88)	Connector courses are intended to connect the Foundations of Data Science (COMPSCI C8/INFO C8/STAT C8) course with particular fields of study. They will apply the concepts and techniques of the foundation course to topics of interest in a particular discipline in order for students to develop critical thinking in data in subject areas that most interest them; these courses also provide a more nuanced understanding of the context in which the data comes into existence.
#2	Introduction to Urban Data Analytics	City & Regional Planning (101)	This course (1) provides a basic intro to census and economic data collection, processing, and analysis; (2) surveys forecasting and modeling techniques in planning; (3) demonstrates the uses of real-time urban data and analytics; and (4) provides a socio-economic-political context for the smart cities movement, focusing on data ethics and governance.
#3	Introduction to Ecological Data Analysis	Env Sci, Policy, & Mgmt (173)	Introduces concepts and methods for practical analysis of data from ecology and related disciplines. Topics include data summaries, distributions, and probability; comparison of data groups using t-tests and analysis of variance; comparison of multi-factor groups using analysis of variance; evaluation of continuous relationships between variables using regression and correlation; and a glimpse at more advanced topics. In computer laboratories, students put concepts into practice and interpret results.
#4	Cartographic Representation	Geography (183)	Problems in the representation of quantitative and qualitative data on thematic maps.
#5	The Person in Big Data	Psychology (7)	This course will introduce students to the basic principles and methods of personality and social psychology as applied to a rapidly growing topic of modern society--the collection and analysis of online social "¹big data."¹ Students will learn about the ways in which big data has historically been defined, collected, and utilized, as well as fundamental concepts in person perception and social behavior that are relevant to topics of big data collection, analysis, and interpretation.

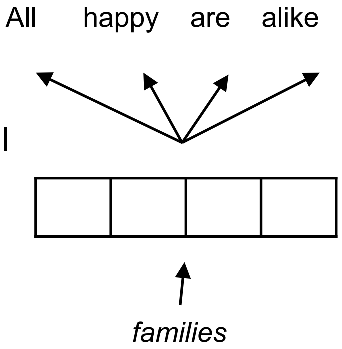
3. The vector representation model is used to surface similar courses across campus that may not share catalog description terms

Inspirations from computational text

(From Language)

"All happy families are alike; each unhappy family is unhappy in its own way."

(Tolstoy)

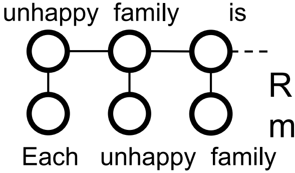


All happy are alike

families

Skip-gram model

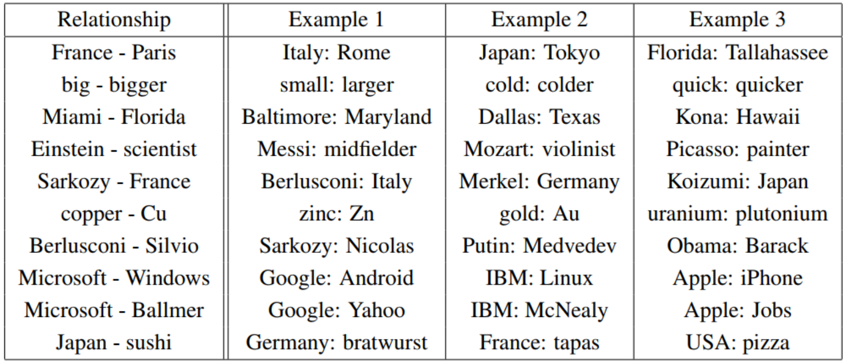
Mikolov et al. (2013)



unhappy family is

Each unhappy family

RNN language model Mikolov et al. (2010)



Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Mikolov, Chen, Corrado, & Dean (2013)

Distributed representation of "royalty" (emergent semantics)
(KING[vec] – MAN[vec] +)WOMAN[vec] ≈ QUEEN[vec]

Mikolov, Yih, & Zweig (2013)

Learning Emergent Semantics

(From Language)

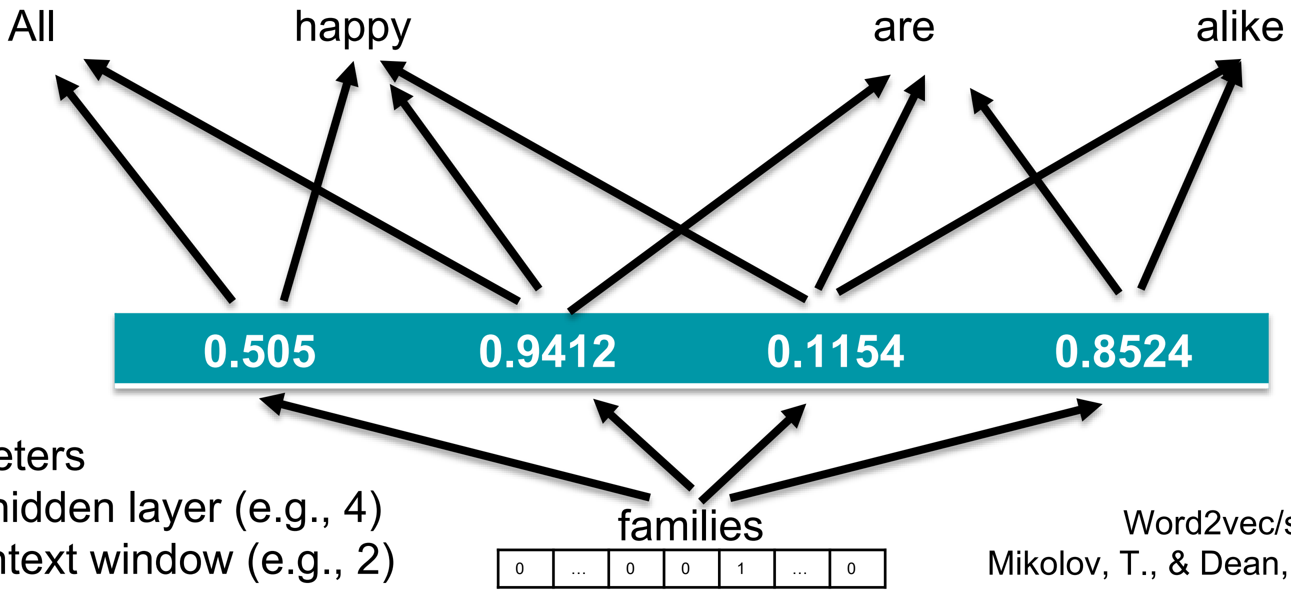
context

"All happy families are alike; each unhappy family is unhappy in its own way."

input

(e.g., Google News archive)

This training process, using SGD, is run on a corpus of 1b words to learn vector representations of each word in the vocabulary



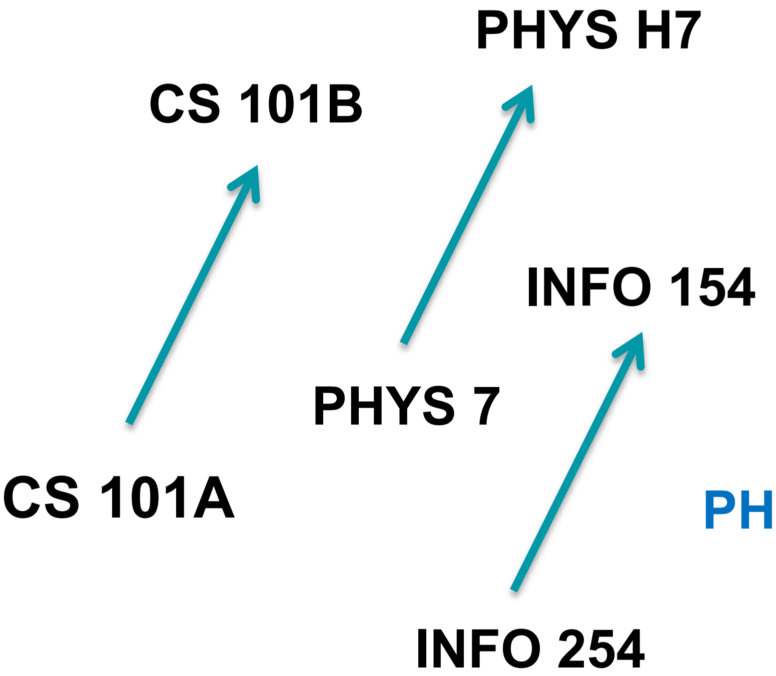
Hyper parameters

- Length of hidden layer (e.g., 4)
- Size of context window (e.g., 2)

Word2vec/skip-gram
Mikolov, T., & Dean, J. (2013)

Methodology

- Skip-gram (word2vec) algorithm applied to enrollment sequences



PHYS H7

CS 101B

INFO 154

PHYS 7

CS 101A

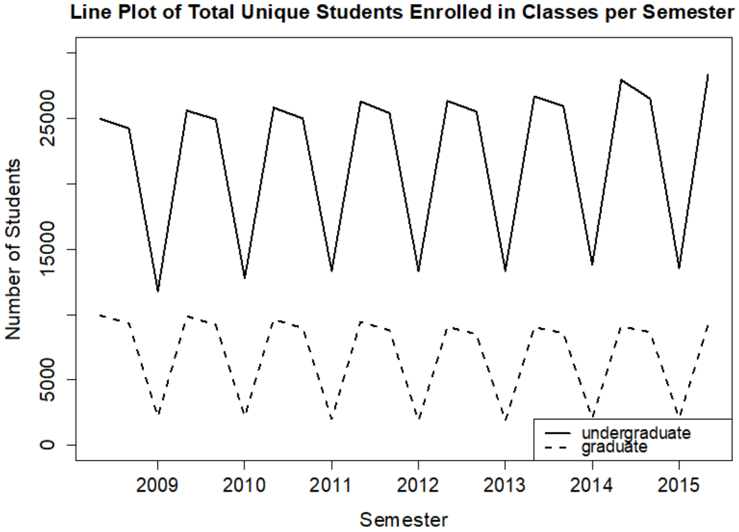
INFO 254

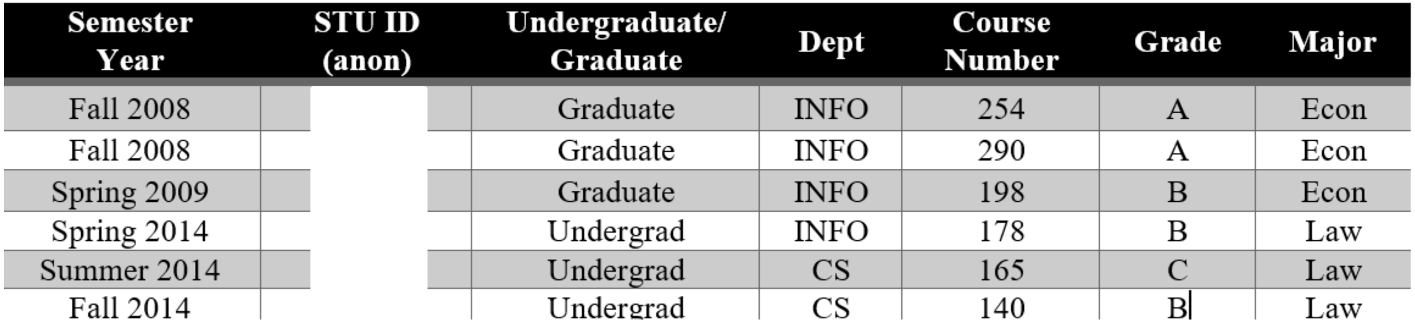
Can this approach embed courses into a “concept space”?

PHYSH7, MATH1B, SPA12, STAT200B…

Dataset

- 3.6M enrollments at UCB from Fall ‘08 through Fall ‘15
- 110,335 undergraduates
- 38,147 graduates
- 9,038 unique lectures courses
  - across 17 colleges
  - 124 departments



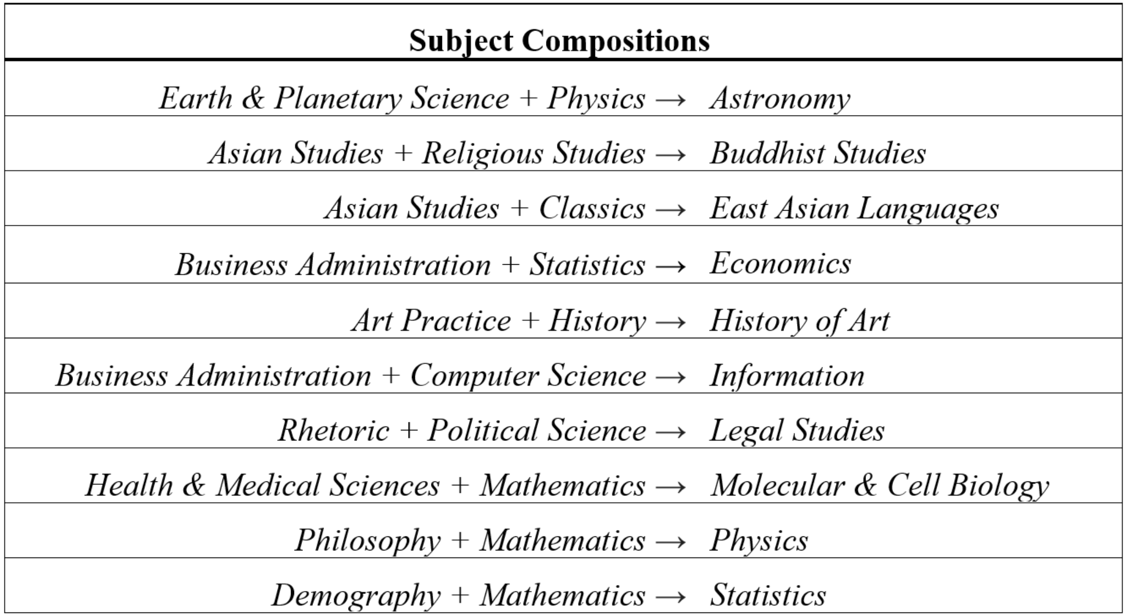


Semester Year	STU ID (anon)	Undergraduate/ Graduate	Dept	Course Number	Grade	Major
Fall 2008		Graduate	INFO	254	A	Econ
Fall 2008		Graduate	INFO	290	A	Econ
Spring 2009		Graduate	INFO	198	B	Econ
Spring 2014		Undergrad	INFO	178	B	Law
Summer 2014		Undergrad	CS	165	C	Law
Fall 2014		Undergrad	CS	140	B	Law

Access to anonymized student data granted by the UCB Registrar & Committee for the Protection of Human Subjects

Exploring the arithmetic properties of the space

A vector space theoretically possesses arithmetic and scalar closure properties. This was tested by adding department centroids together and observing the nearest neighbor department centroid that resulted.

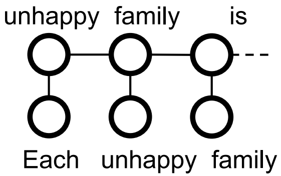


Subject Compositions
Earth & Planetary Science + Physics → *Astronomy*
Asian Studies + Religious Studies → *Buddhist Studies*
Asian Studies + Classics → *East Asian Languages*
Business Administration + Statistics → *Economics*
Art Practice + History → *History of Art*
Business Administration + Computer Science → *Information*
Rhetoric + Political Science → *Legal Studies*
Health & Medical Sciences + Mathematics → *Molecular & Cell Biology*
Philosophy + Mathematics → *Physics*
Demography + Mathematics → *Statistics*

Pardos & Nam (2018)

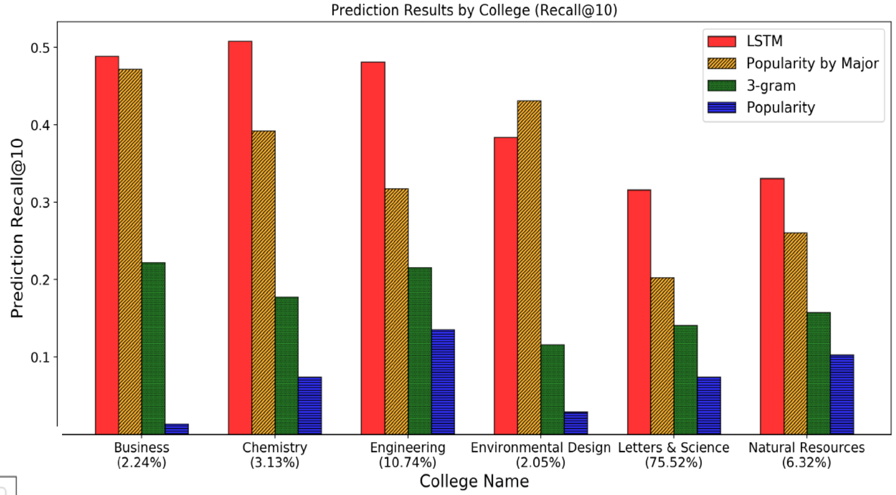
Next course prediction (normative)

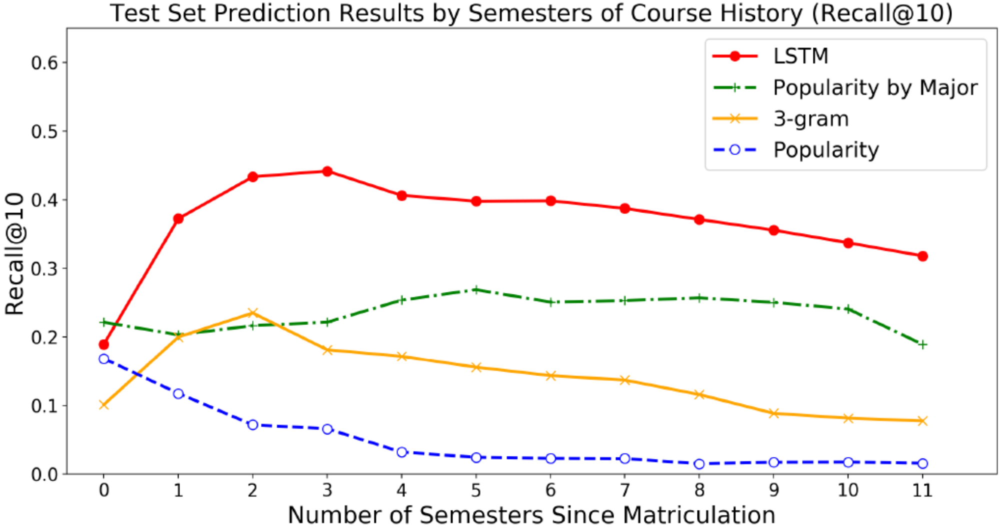
- Trained predictive models of course selection



RNN language model

Mikolov et al. (2010)

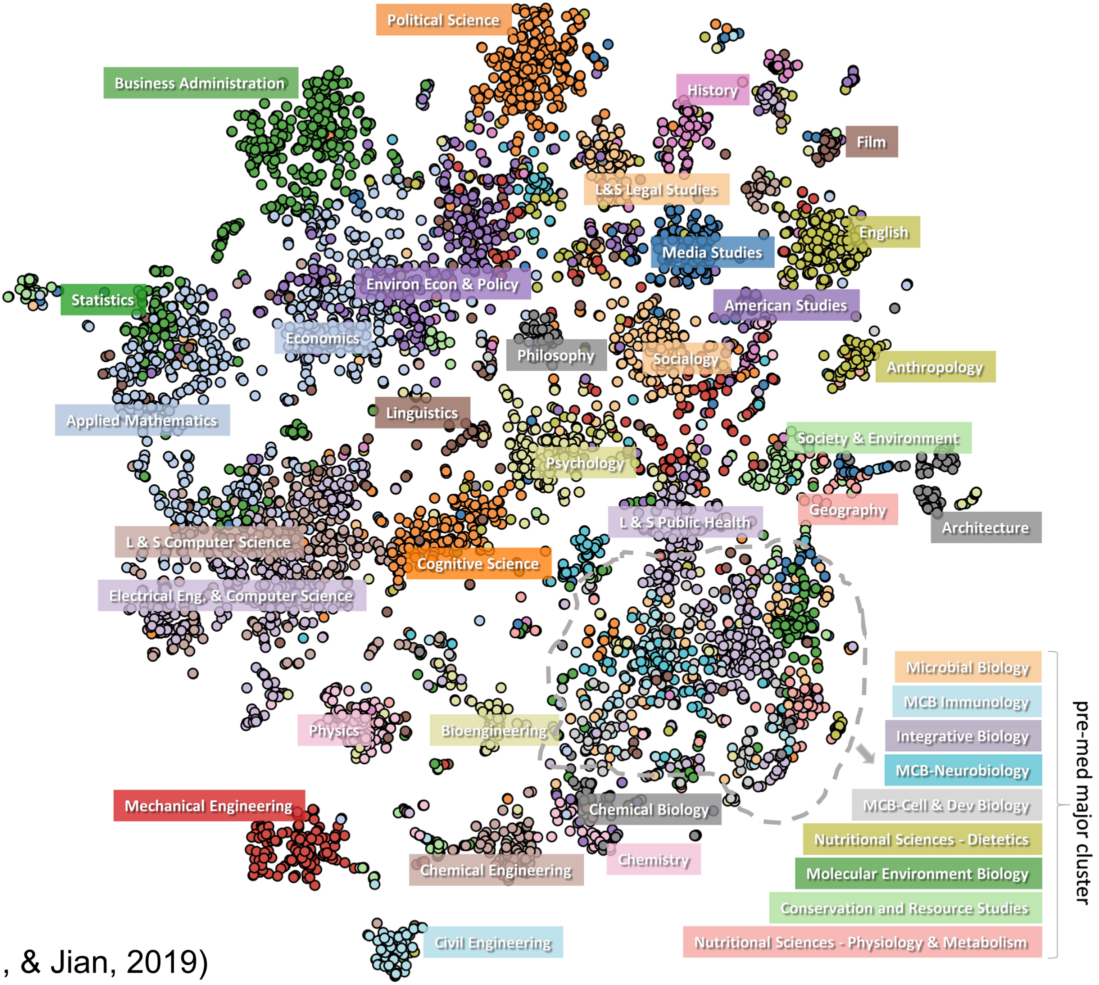




Pardos, Z.A., Fan, Z., Jiang, W. (2019) **Connectionist Recommendation in the Wild: On the utility and scrutability of neural networks for personalized course guidance.** *User Modeling and User-Adapted Interaction*.

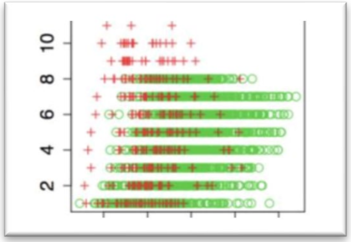
Acknowledgement: We thanks Andrew Eppig (OPA), Mark Chiang (EDW), Johanna Metzgar (ex-OR), Jen Stringer (ETS), Aswan Movv (EDW), Daniel Grieb (EDW), Anji Gannavarapu (EDW), Max Michel (EDW), Larry Conrad (CIO), and Walter Wong (Registrar) for their support in developing the system.

Visualization of all undergrad students the semester before they graduate

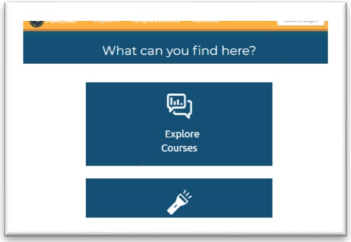


(Pardos, Fan, & Jian, 2019)

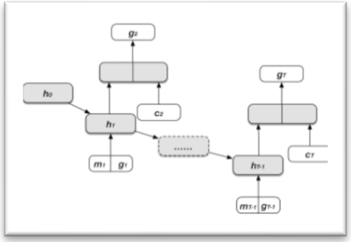
Applications of the enrollment vector space



Predicting on-time graduation: a case study of Integrative Biology students (Luo & Pardos, AAAI EAAI 2018)



Developing the vector-based course information system at UCB (Pardos, Fan, & Jiang, UMUAI 2019)



Inferring and personalizing course prerequisite relationships (Jiang, Pardos, & Wei, LAK 2019)

References

- Pardos, Z. A., Fan, Z., Jiang, W. (2019) Connectionist Recommendation in the Wild: On the utility and scrutability of neural networks for personalized course guidance. *User Modeling and User-Adapted Interaction*. https://doi.org/10.1007/s11257-019-09218-7
- Jiang, W., Pardos, Z.A., Wei, Q. (2019) Goal-based Course Recommendation. In C. Brooks, R. Ferguson & U. Hoppe (Eds.) *Proceedings of the 9th International Conference on Learning Analytics and Knowledge* (LAK). ACM. Tempe, Arizona. Pages 36-45.
- Pardos, Z. A., & Nam, A. J. H. (2018) A Map of Knowledge. *CoRR preprint*, abs/1811.07974. https://arxiv.org/abs/1811.07974
- Luo, Y., Pardos, Z. A. (2018) Diagnosing University Student Subject Proficiency and Predicting Degree Completion in Vector Space. In E. Eaton & M. Wollowski (Eds.) *Proceedings of the Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (EAAI). New Orleans, LA. AAAI Press. Pages 7920-7927.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).
- Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Approaches to Scalable Personal Guidance in MOOCs and On Campus

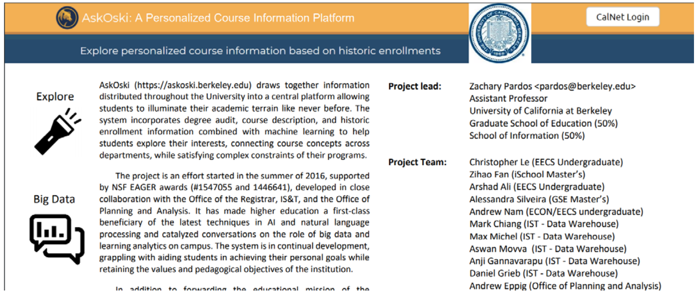
BAYLAN
Bay Area Learning Analytics Network

Thank You!

Questions?

Zachary A. Pardos
University of California at Berkeley

zachpardos.com
zp@berkeley.edu


AskOski: A Personalized Course Information Platform
Explore personalized course information based on historic enrollments
CalNet Login
Explore
Big Data

AskOski (https://askoski.berkeley.edu) draws together information distributed throughout the University into a central platform allowing students to illuminate their academic terrain like never before. The system incorporates degree audit, course description, and historic enrollment information combined with machine learning to help students explore their interests, connecting course concepts across departments, while satisfying complex constraints of their programs.

The project is an effort started in the summer of 2016, supported by NSF EAGER awards (#1547055 and 1446641), developed in close collaboration with the Office of the Registrar, IS&T, and the Office of Planning and Analysis. It has made higher education a first-class beneficiary of the latest techniques in AI and natural language processing and catalyzed conversations on the role of big data and learning analytics on campus. The system is in continual development, grappling with aiding students in achieving their personal goals while retaining the values and pedagogical objectives of the institution.

In addition to forwarding the educational mission of the

Project lead: Zachary Pardos <pardos@berkeley.edu>
Assistant Professor
University of California at Berkeley
Graduate School of Education (50%)
School of Information (50%)

Project Team: Christopher Le (EECS Undergraduate)
Zihao Fan (iSchool Master's)
Arshad Ali (EECS Undergraduate)
Alessandra Silveira (GSE Master's)
Andrew Nam (ECON/EECS undergraduate)
Mark Chiang (IST - Data Warehouse)
Max Michel (IST - Data Warehouse)
Aswan Movva (IST - Data Warehouse)
Anji Gannavarapu (IST - Data Warehouse)
Daniel Grieb (IST - Data Warehouse)
Andrew Eppig (Office of Planning and Analysis)

One-page recommender system synopsis: tiny.cc/askoski

Acknowledgement: This work was funded by the National Science Foundation (Awards #1547055, #1446641)