Data-Assistive Course-to-Course Articulation Using Machine Translation

Zachary A. Pardos[1], Hung Chau[2], Haocheng Zhao[1]

Paper: tiny.cc/course-articulation

1

CAHL Computational Approaches to Human Learning (CAHL) research lab

GRADUATE SCHOOL OF EDUCATION

UC Berkeley School of Information

2

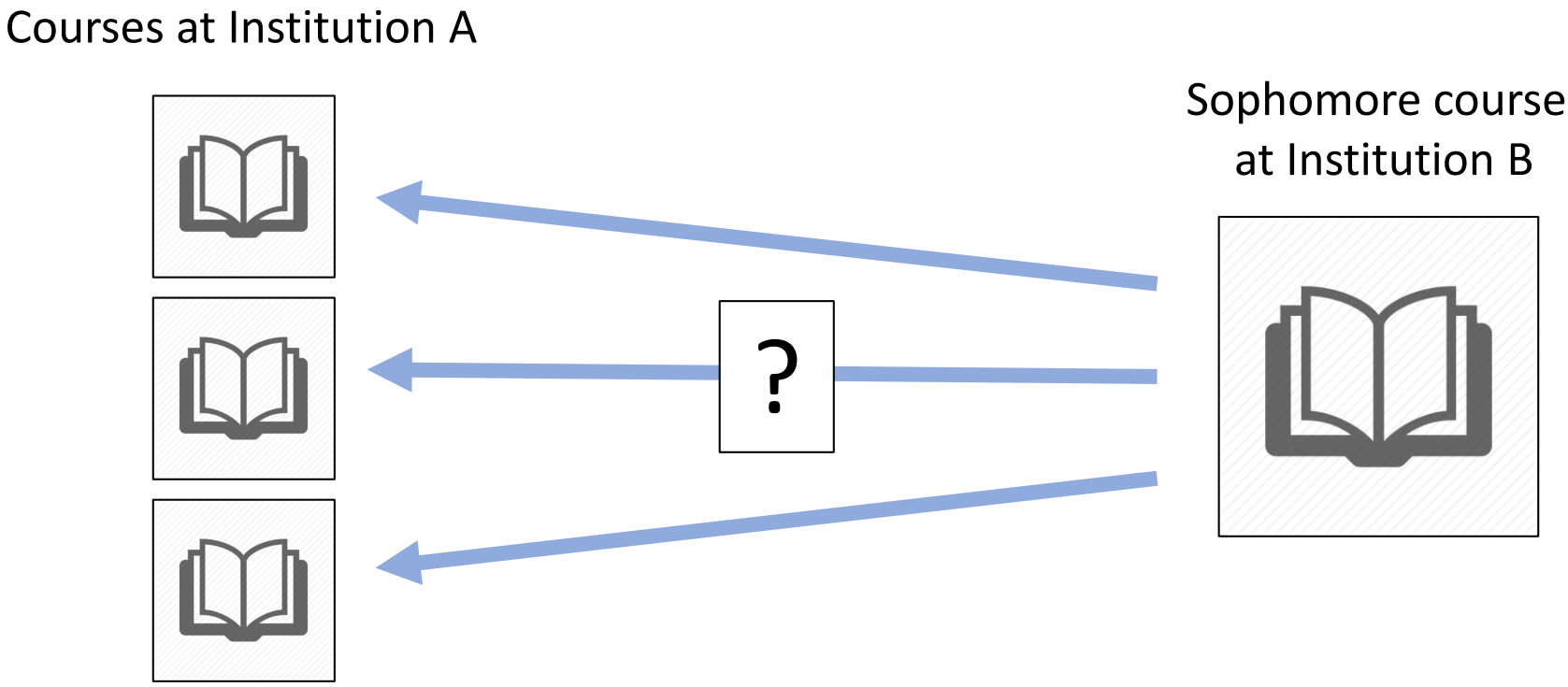
PAWS Personalized Adaptive Web Systems Lab

INFORMATICS AND NETWORKED SYSTEMS

University of Pittsburgh School of Computing and Information

Course-to-Course Articulation

Which course (if any) at Institution A is academically equivalent to a course at Institution B?



Taking these courses at Institution A is often required to qualify for transfer to Institution B

Background

20 million students enter higher education in the United States each year, 45% of these students begin at 2-year public institutions

(Hossler et al., 2012)

81.4% of surveyed (N=19,000) beginning 2-year students expressed an intent to transfer to a 4-year program

(Radford et al., 2010)

Six years after starting at the 2-year, only 13% of students had obtained a 4-year degree (N=852,439)

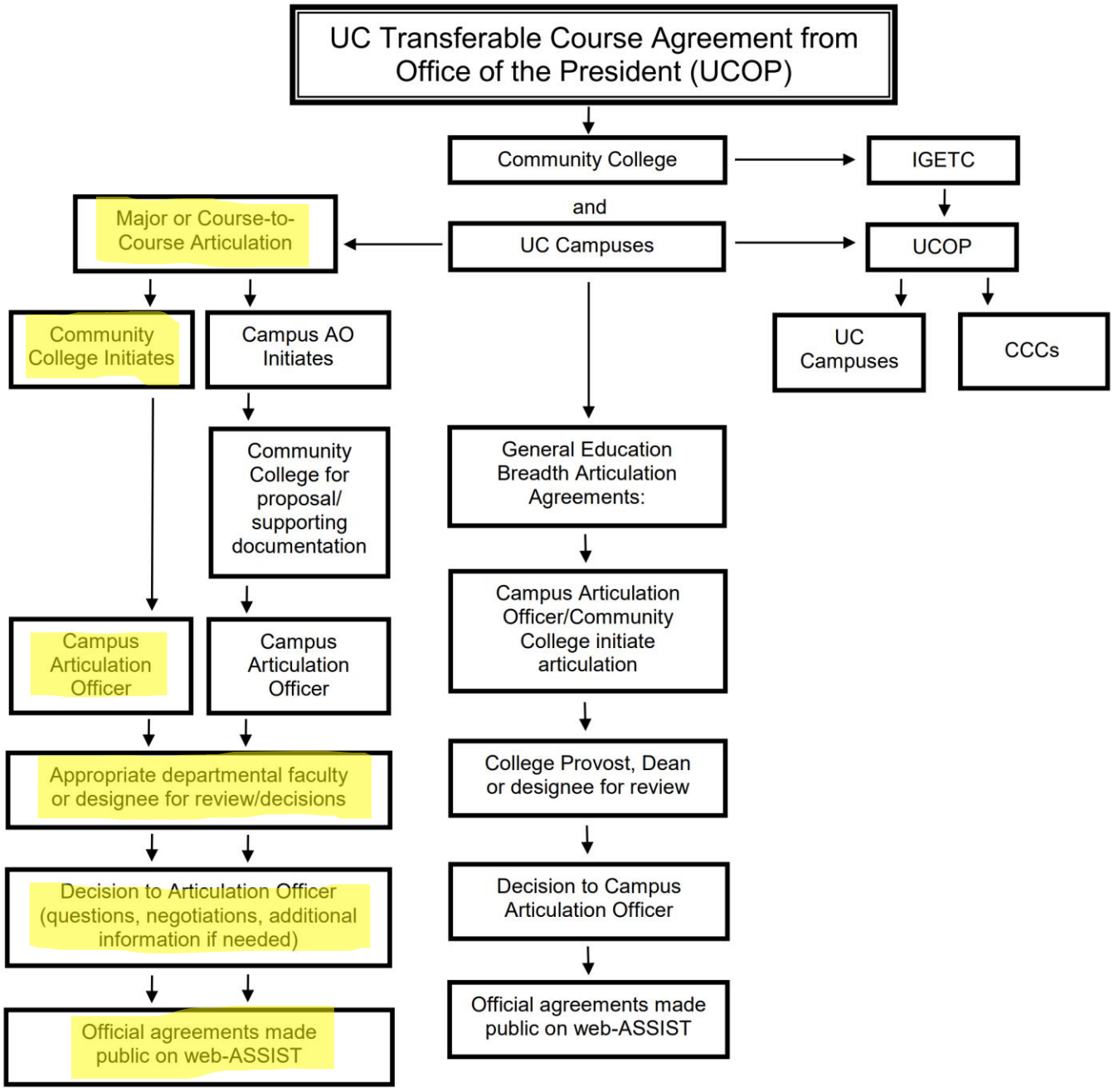
(Shapiro et al., 2017)

IES NCES National Center for Education Statistics

National Student Clearinghouse

Current Practice

- Based on demand
- Favors regional schools
- Favors public schools
  - 94% loss of credit transferring to a private (GAO, 2017)
  - 42% is lost when transferring to a public
- Slow process (~1yr)
- Open to instructor bias



(California Articulation Policies and Procedures Handbook)

The Challenge of Articulation

The California post-secondary system alone has:

- 115 2-year California Community Colleges (CC)
- 23 California State Universities (CSU)
- 9 University of California campuses (UC)

The number of articulations to consider between 1 CC and 1 UC:

- Without *department-to-department* mapping between institutions

63M pairs (1,000*7,000)

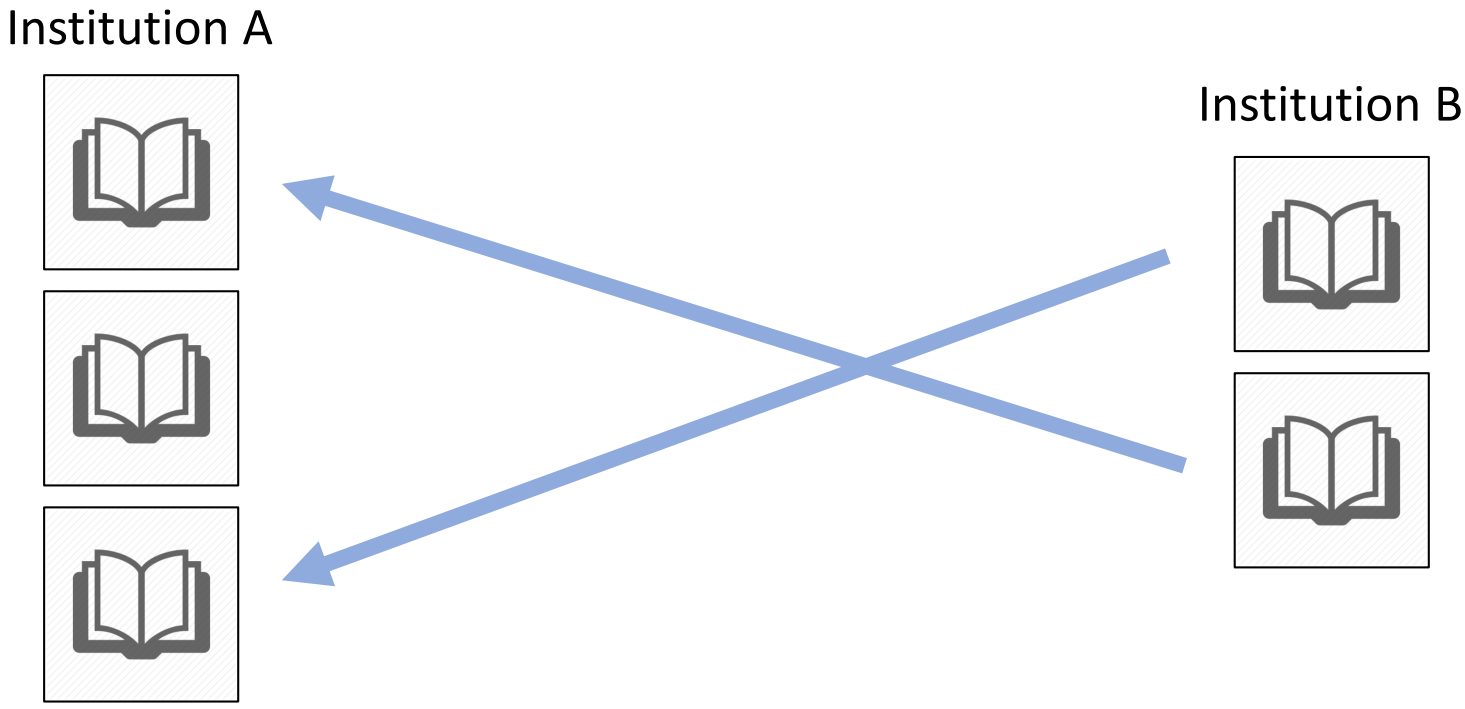
- With *department-to-department* mapping between institutions

35,000 (50*20*35)

<u>A methodological approach is needed that can scale to this magnitude</u>

We Propose

- A machine learning approach, using both course catalog descriptions and historic enrollment patterns to automatically surface academically equivalent courses across any two arbitrary institutions.



Potential Impact

- more (valid) articulations means more credit mobility which means students will have more opportunities to transfer.

	4- year University of California campus (UC1) (Fall 2008 - Fall 2016)	2-year California Community College (Fall 2013 – Fall 2018)
Enrollment records	4.8M	298,174
Number of students	164,196	58,716
Number of courses	7,487	1,000
Number of departments	179	53
Course descriptions	325 words per description 489 descriptions less than 10 words	27 words per description 62 descriptions less than 10 words 4 don't have descriptions

- Existing articulation pairs sourced from assist.org
  - 65 articulation pairs
  - 184 UC1 required degree courses that have no CC1 courses articulated

UC1's course	*CC1's course(s)*
AFRICAM5B	AFRAM_31
ASAMST20A	ASAME_45A; ASAME_45B
ASAMST20C	NO COURSE ARTICULATED

Models of Course Similarity

Catalog description / content-based:

Bag-of-words: Using a vocabulary frequency vector to represent a course description

Word	CS61A	MATH54	EDUC161
pointer	3	1	0
design	0	4	1
construction	0	2	0
sculpture	0	0	0
algorithm	2	0	5
regression	0	0	2
…			

Models of Course Similarity

Catalog description / content-based:

Bag-of-words: Using a vocabulary frequency vector to represent a course description

TF-IDF: Using term frequency – inverse document frequency to represent course description

Word	CS61A	MATH54	EDUC161
pointer	0.12	1.435	0
design	0	0.898	2.314
construction	0	1.123	0
sculpture	0	0	0
algorithm	2.97	0	0.67
regression	0	0	1.12
...			

$$w_{t,d} = (1 + log(tf_{t,d})) \times log_{10} \frac{N}{df_t}$$

Models of Course Similarity

Catalog description / content-based:

Bag-of-words: Using a vocabulary frequency vector to represent a course description

TF-IDF: Using term frequency – inverse document frequency to represent a course description

DescVec: Averaging Google's word2vec representation of all words in a course description

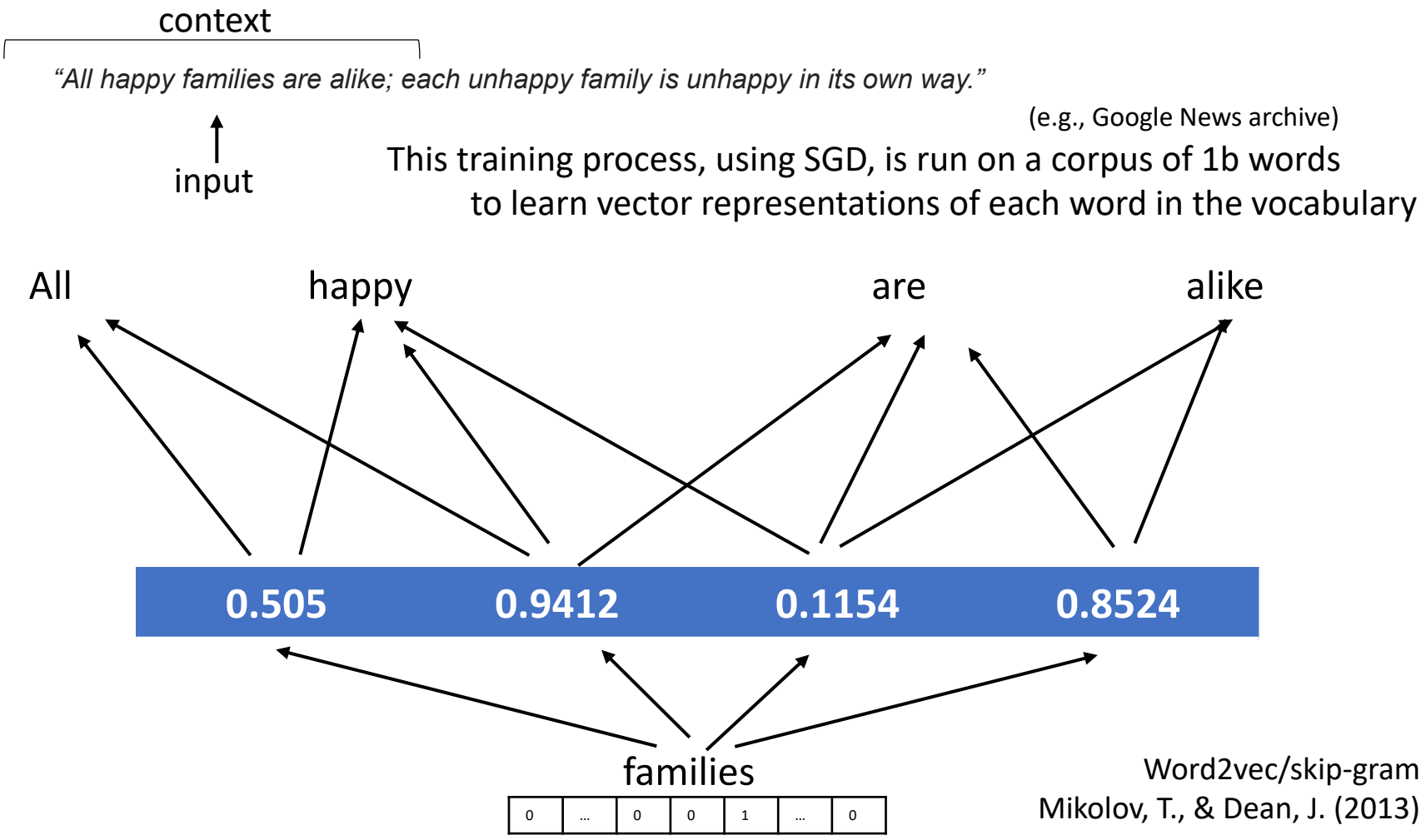
Embedding	CS61A	MATH54	EDUC161
Dim1	0.142	5.438	0.002
Dim2	-4.12	2.908	-0.005
Dim3	4.99	-1.108	3.020
Dim4	0.019	-5.109	1.021
Dim5	-1.194	2.232	-1.220
…	…	…	…
Dim300	-2.229	1.009	4.210

Models of Course Similarity

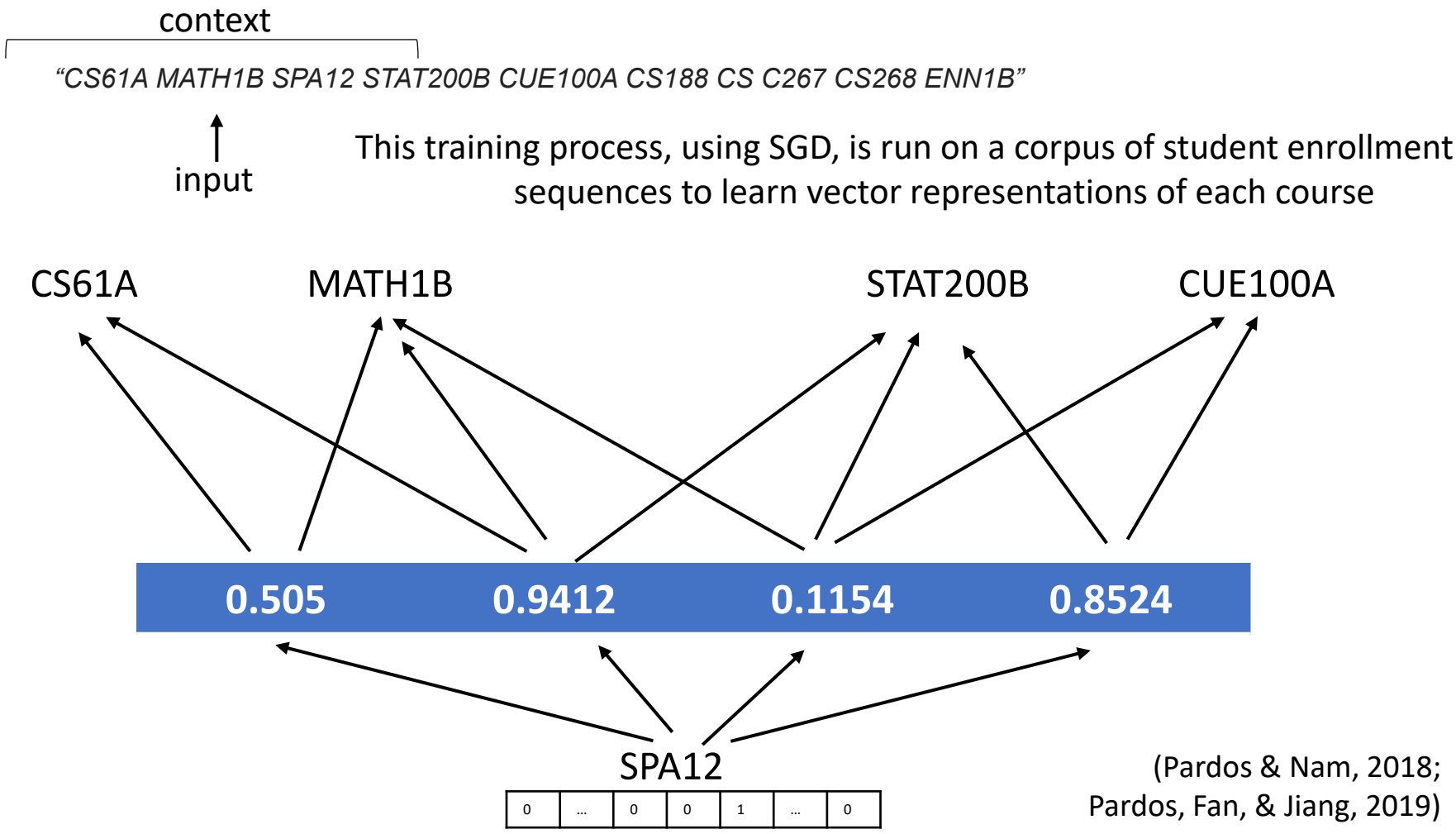
Course enrollment / collaborative-based:

Course2vec: Using continuous course vectors learned from enrollment histories

Learning Word Embeddings



Learning Course Embeddings



(Pardos & Nam, 2018;
Pardos, Fan, & Jiang, 2019)

Models of Course Similarity

Course enrollment / collaborative-based:

Course2vec: Using continuous course vectors learned from enrollment histories

Embedding	CS61A	MATH54	EDUC161
Dim1	1.152	2.438	-1.002
Dim2	2.16	-5.908	0.205
Dim3	-2.09	-1.708	-3.420
Dim4	1.059	2.109	2.521
Dim5	1.594	2.562	5.620
…	…	…	…
Dim229 (UC1) Dim20 (CC1)	2.229	-4.41	4.210

Models of Course Similarity

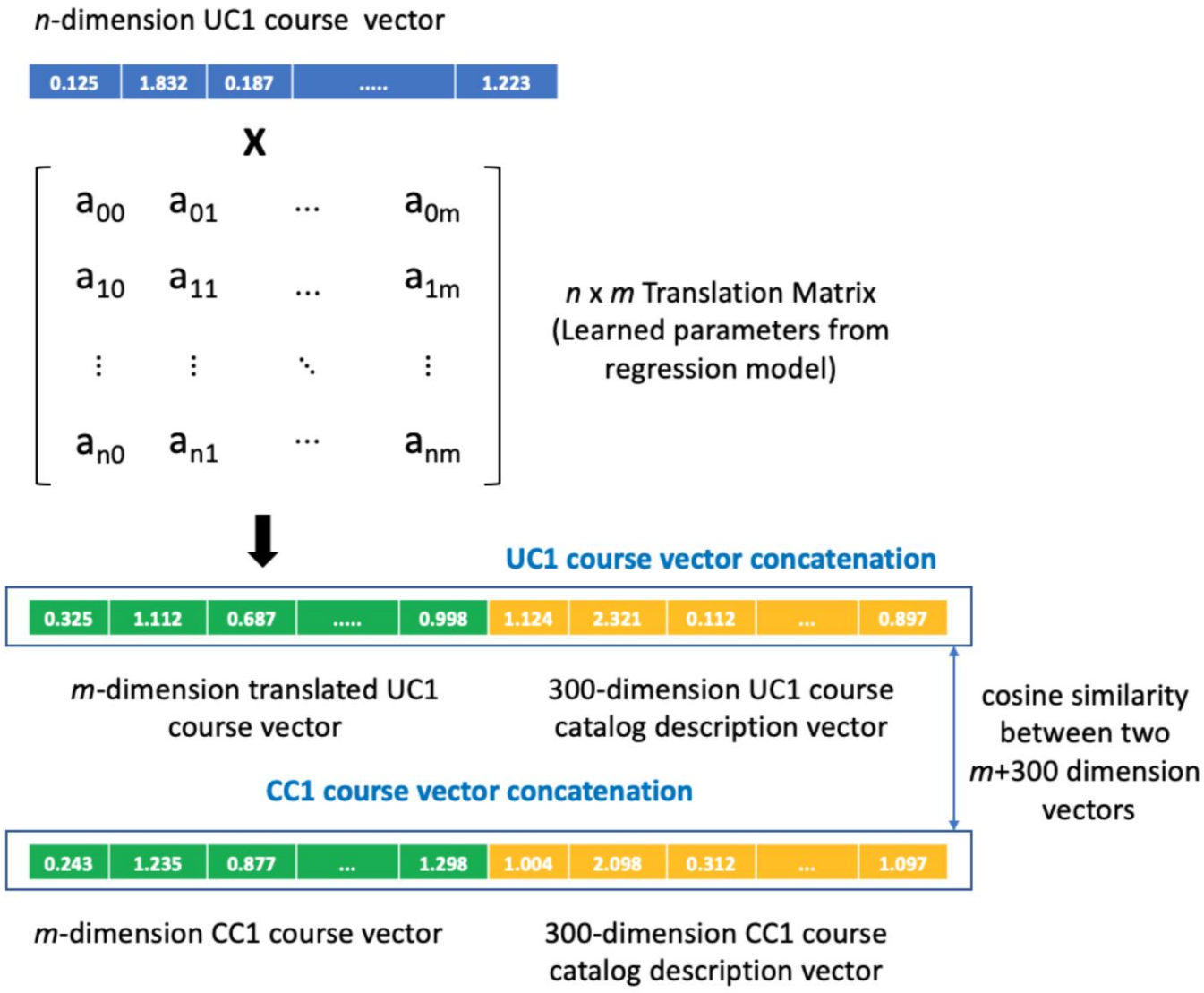
Course enrollment / collaborative-based:

Course2vec: Using continuous course vectors learned from enrollment histories
Course2vec+DescVec: Concatenating a course's course2vec vector with its DescVec vector

Models of Course Similarity

Machine Translation from UC1 embedding to CC1 embedding:

Course2vec+DescVec

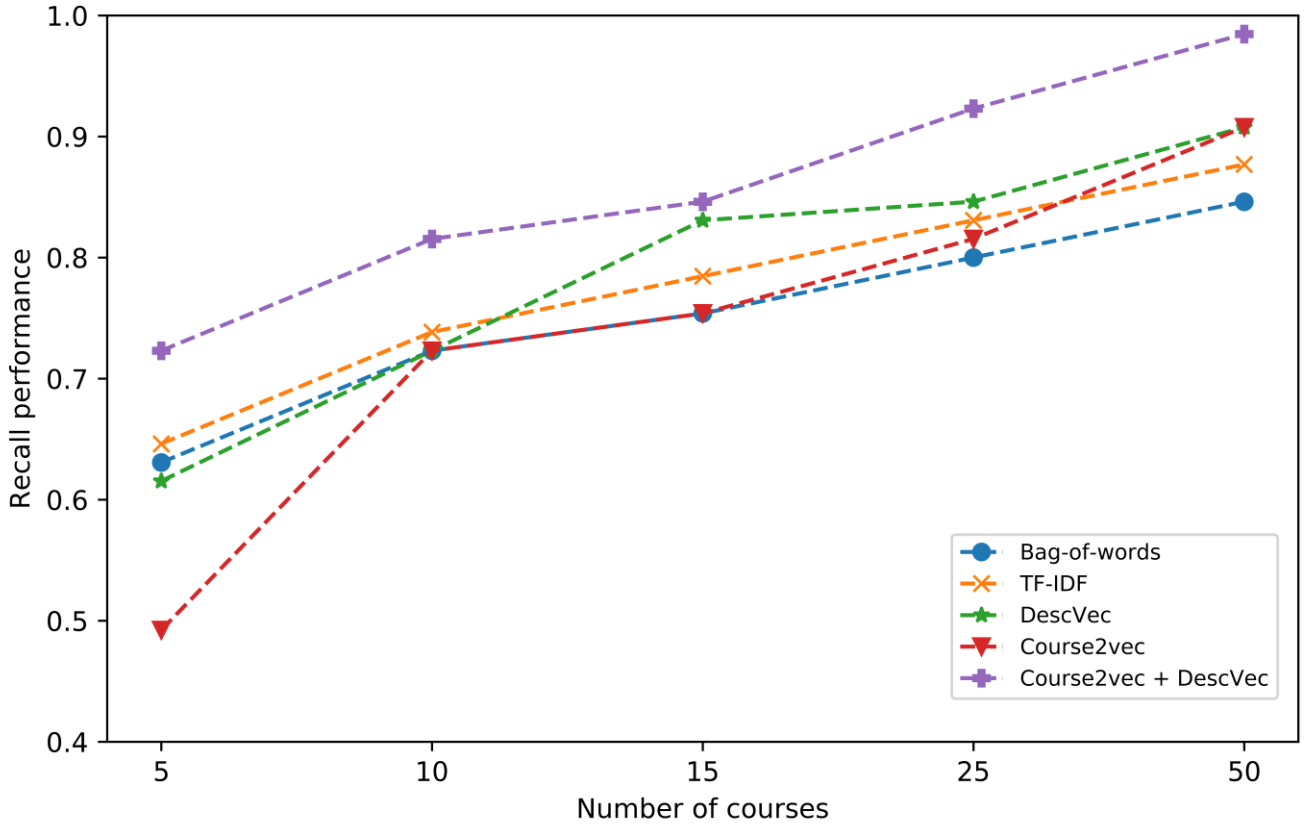


Evaluation

- Attempt to recover the 65 official articulation pairs
- For each method, suggest N candidate courses from CC1 to articulate to each UC1
- Report the percentage of suggestions containing the true course (Recall @ N)
- Use leave-one-out cross validation to train the translation

See paper for details on different loss functions, distance metrics, and hyper parameter tuning

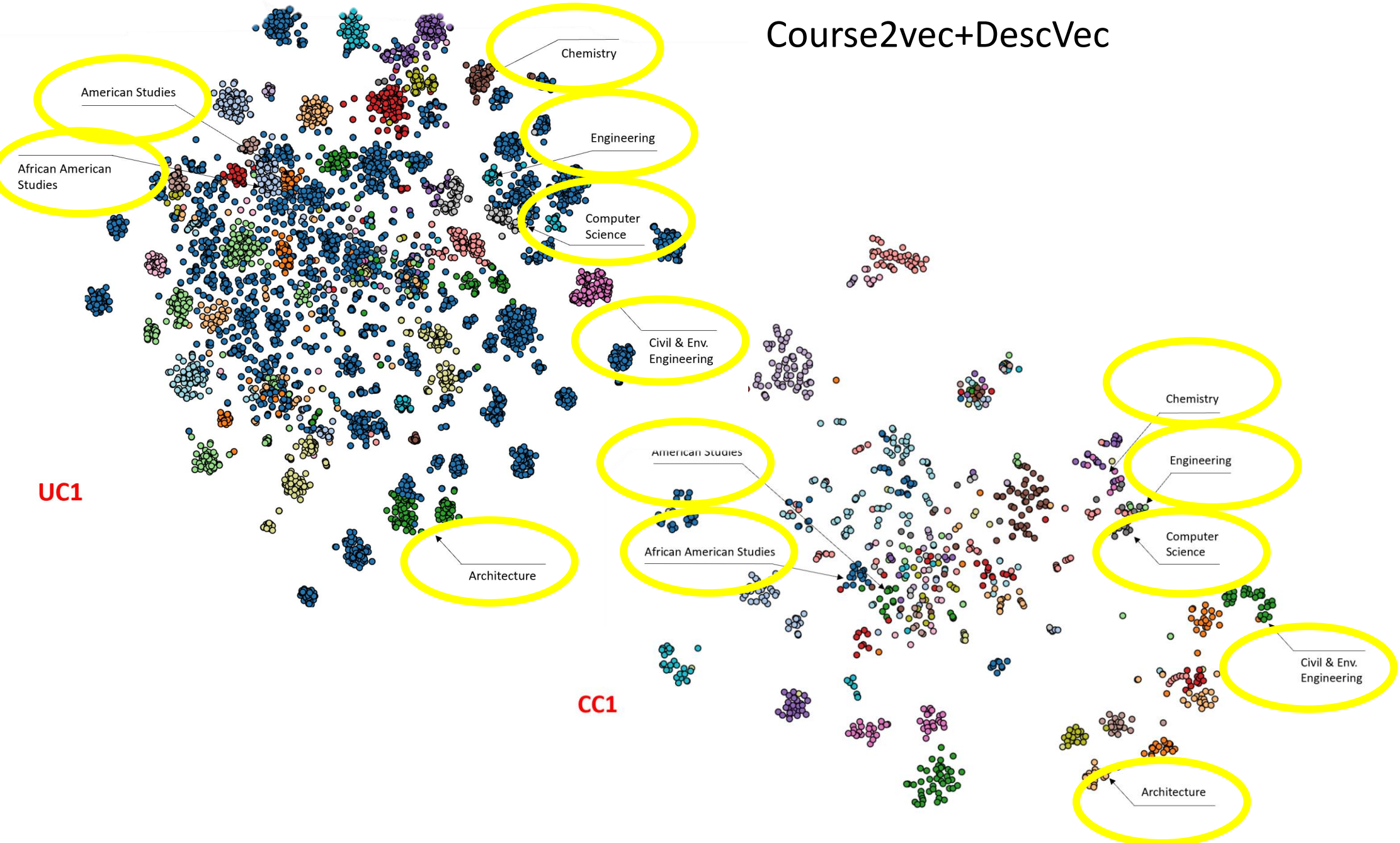
Results



Course Representation	*Median Rank*	*Mean Rank*	*Std of Rank*
Bag of words	3.0	59.12	173.28
TF-IDF	3.0	57.01	177.65
Doc2vec	3.0	21.06	57.94
Course2vec	6.0	17.74	33.65
Course2vec+doc2vec	**2.0**	**7.94**	**15.73**

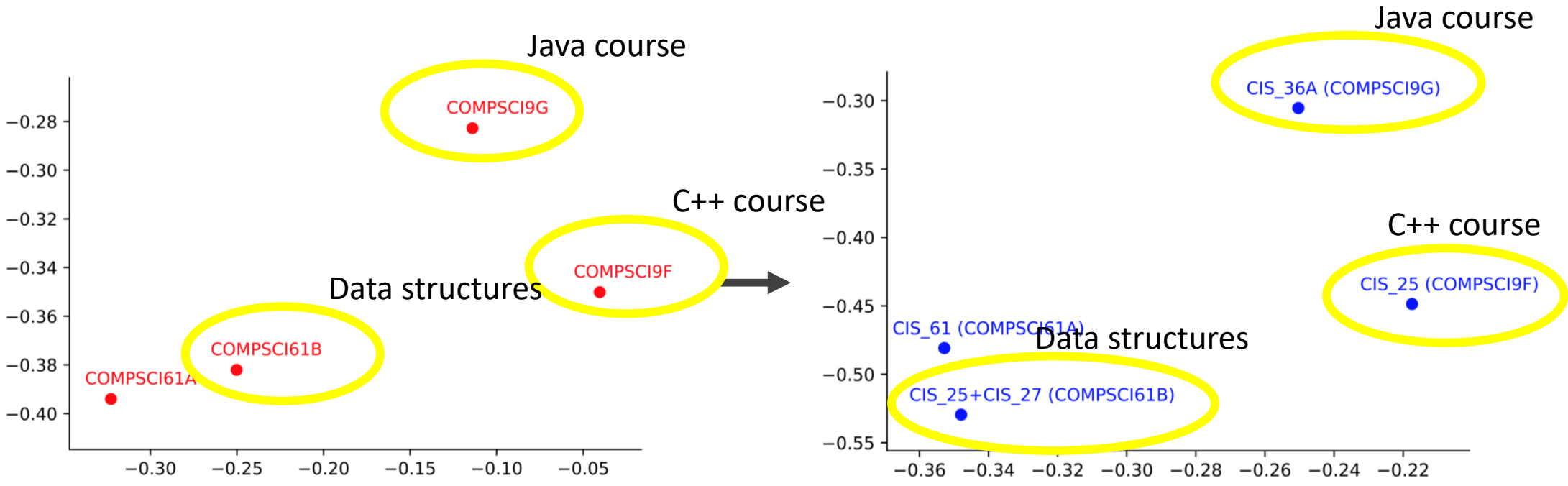
Model Inspection

t-SNE visualization of UC1 and CC1
Course2vec+DescVec



Model Inspection

PCA visualization of UC1 and CC1 Course2vec+DescVec,
Zoomed in on Comp Sci department courses



UC1 and CC1 exhibit similar topology within the Comp Sci department, suggesting a linear translation can be found between the two embeddings

Producing Articulation Candidates Report

Which of the 185 unarticulated UC1 courses have the most promising CC1 candidates?

Campus	*Course ID*	*Course Title*
UC1	ENGIN26	3D Modeling for Design
CC1	ENGIN_77	PROGRAMMING/MATLAB
CC1	ENGIN_22	ENGINEERING GRAPHICS
CC1	MATH_3E	LINEAR ALGEBRA
CC1	ENGIN_45	PROPERTIES/MATERIALS
CC1	MATH_3F	DIFFERENTIAL EQUATIONS
CC1	PHYS_4B	GEN PHYSICS W/CALCULUS
CC1	MATH_3C	CALCULUS III
CC1	ENGIN_35	ENGIN MECH-STATICS
CC1	PHYS_4A	GEN PHYSICS W/CALCULUS
CC1	MATH_11	DISCRETE MATHEMATICS
CC1	ENGIN_17	INTRO ELECT ENGIN
CC1	ENGIN_18	INTRO ELECTRICAL ENGIN
CC1	MATH_3B	CALCULUS II
CC1	CIS_61	STRUC/INTER COMP PRG
CC1	PHYS_4C	GEN PHYSICS W/CALCULUS

Syllabus: https://tbp.berkeley.edu/syllabi/758/download/

Campus	*Course ID*	*Course Title*
UC1	ECON2	Introduction to Economics
CC1	ECON_1	MACRO-ECONOMICS
CC1	ECON_2	MICRO-ECONOMICS
CC1	BUS_1A	FINANCIAL ACCOUNTING
CC1	BUS_1B	MANAGERIAL ACCTG
CC1	MATH_16A	CALCULUS-BUS/SOCSC
CC1	BUS_2	INTRO TO BUS LAW
CC1	MATH_1	PRE-CALCULUS
CC1	BUS_10	INTRO TO BUSINESS
CC1	MATH_13	INTRO TO STATISTICS
CC1	MATH_16B	CALCULUS-BUS/SOCSC
CC1	BUS_21	PAYROLL ACCOUNTING
CC1	MATH_3A	CALCULUS 1
CC1	BUS_4	COST ACCOUNTING
CC1	MUSIC_15A	JAZZ/BLUES/POP MUSIC
CC1	MUSIC_15B	JAZZ/BLUES/POP MUSIC

Syllabus: http://www.econ.berkeley.edu/.../Economics...pdf

See paper for details on heuristics used to narrow down the 185

Limitations

- Course2Vec not available without enrollment data (hard to get)
- Translation method requires existing articulations

Conclusions

- Enrollment data provides complementary information about courses
- Data-assistive methods can narrow down articulation candidates, making greater articulation tractable for institutions

Future directions

- Explore translation methods that do not require existing pairs
- Implement a student-facing articulation petition recommender

Data-Assistive Course-to-Course Articulation Using Machine Translation

Zachary A. Pardos, Hung Chau, Haocheng Zhao

Paper: tiny.cc/course-articulation

Thank You!