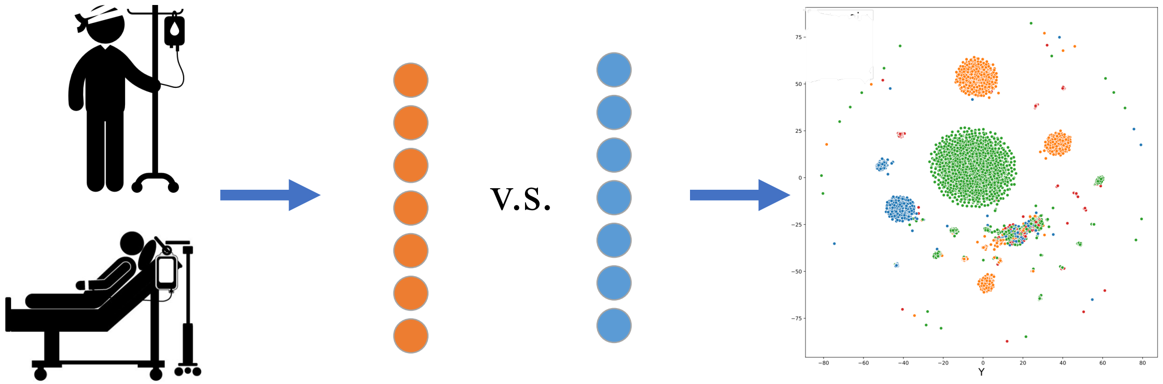
Enriching Unsupervised User Embedding via Medical Concepts

Xiaolei Huang[1], Franck Dernoncourt[2], Mark Dredze[3]

1. University of Memphis 2. Adobe Research 3. Johns Hopkins University

User Embedding

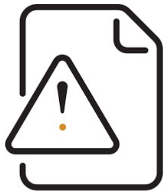
- *User embedding* models user behaviors by mapping all user info into a unified vector space.



An example for cohort selection using user embedding.

- Unsupervised user embedding.
  - No human supervision.

reduce *labor*



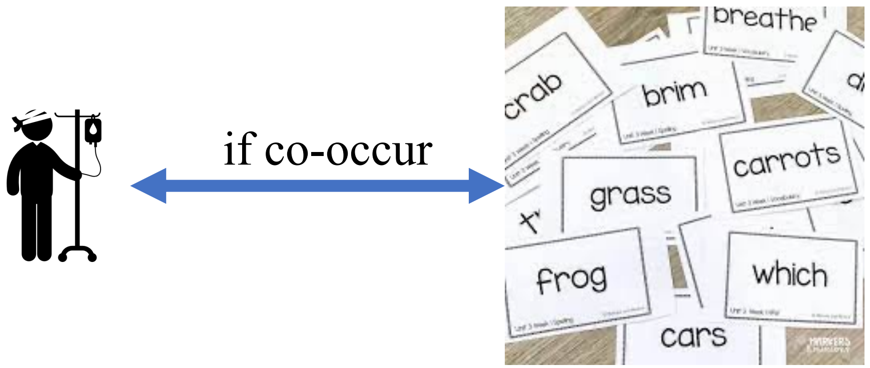
avoid *error-prone labels*[1]

save *money*

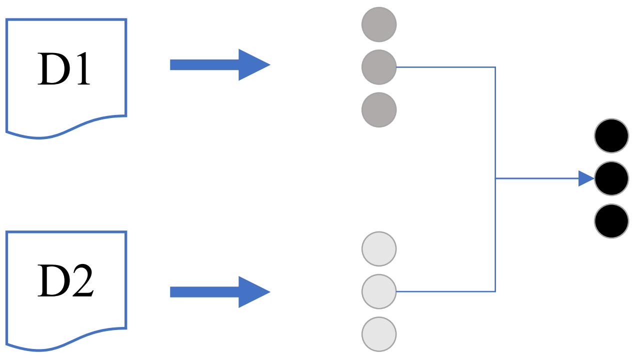
1. Birman-Deych, et al. Accuracy of icd-9-cm codes for identifying cardiovascular and stroke risk factors.

Existing Unsupervised Approaches

- usr2vec[1]
  - predict binary relation between users and single tokens.
- doc2vec[2]
  - merge feature vectors of all text documents of users.



Token - level



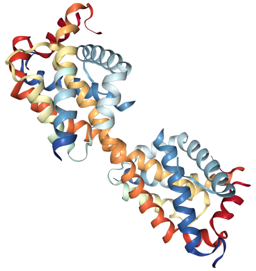
Doc - level

1. Single tokens may lose important information;
2. Clinical notes can be very long, neural networks can suffer long dependency.

1. Amir et al. Quantifying mental health from social media with neural user embeddings
2. Ding, et al. Predicting delay discounting from social media likes with unsupervised feature learning.

Medical Concept Matters

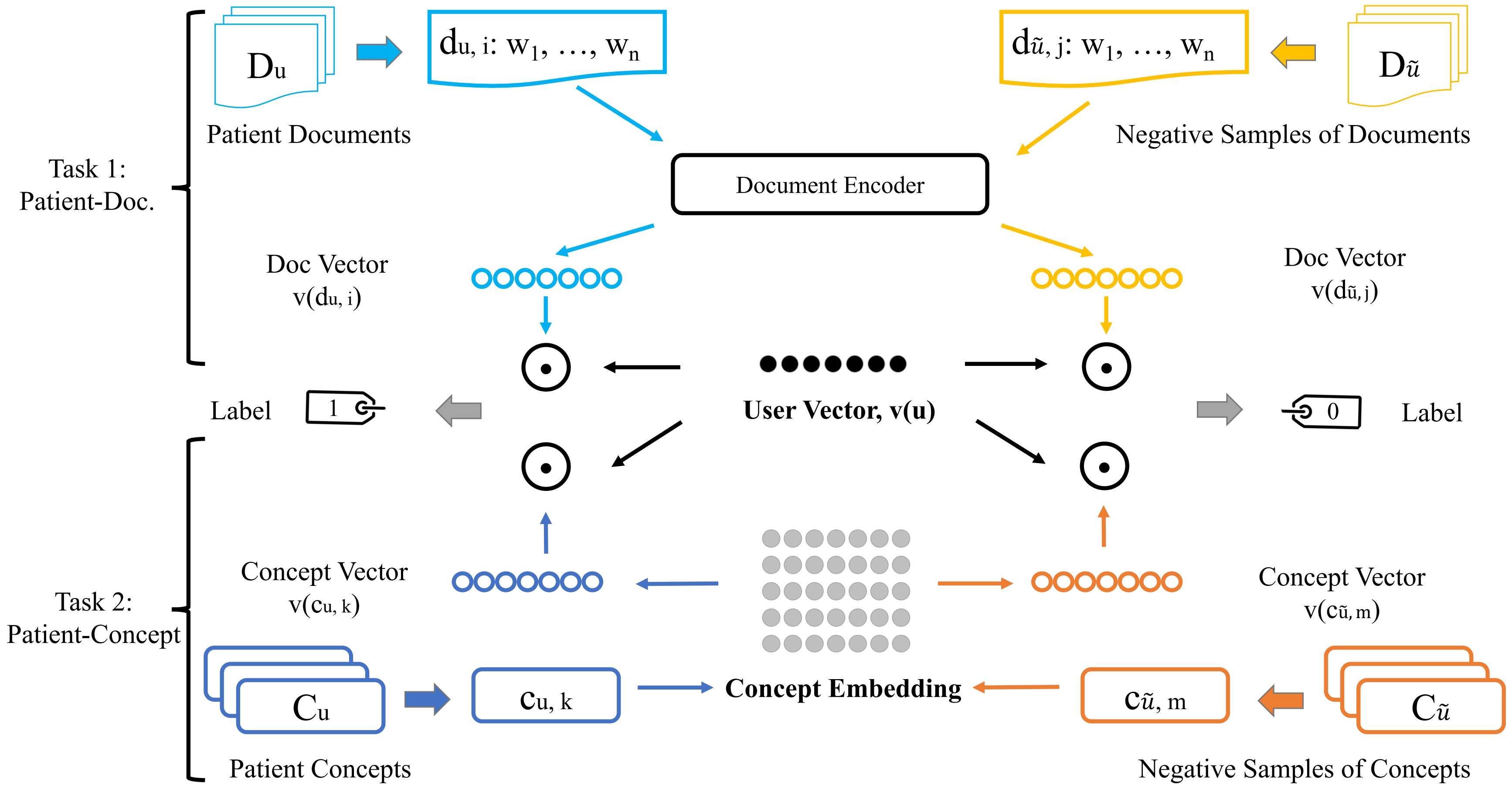
- *Medical concepts*: basic units for medical info, such as disease symptom and clinical drug.
  - Concepts can have more meaningful semantic indications.

bid + protein → 

heavy + drinker → alcohol-abuse

- How to incorporate medical concepts and clinical notes into a meaningful way?

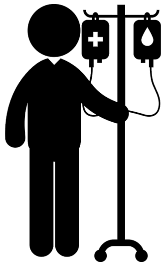
Concept-Aware User Embedding (*CAUE*)



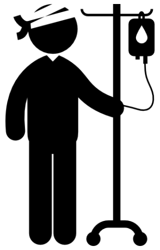
CAUE

<table>
<tr><th>Task 1: Patient-Document</th><th>Task 2: Patient-Concept</th></tr>
<tr><td>$$\mathcal{L}(u,d) = -log(\sigma(v(u)\cdot v(d_u))) - log(1-\sigma(v(u)\cdot v(d_{\tilde{u}})))$$</td><td>$$\mathcal{L}(u,c) = -log(\sigma(v(u)\cdot v(c_u))) - log(1-\sigma(v(u)\cdot v(c_{\tilde{u}})))$$</td></tr>
<tr><td colspan="2">simulate diagnosis process:
enforce models to recognize patients (u) of medical notes (d) / concepts (c).</td></tr>
<tr><td>Key Methods</td><td>1. contrastive learning[1]: generate counterfactuals → model robustness
2. negative sampling: convert to binary prediction (self-supervision)</td></tr>
</table>

Counterfactuals:



V.S.



1. Logeswaran and Lee. An efficient framework for learning sentence representations.

Clinical Data

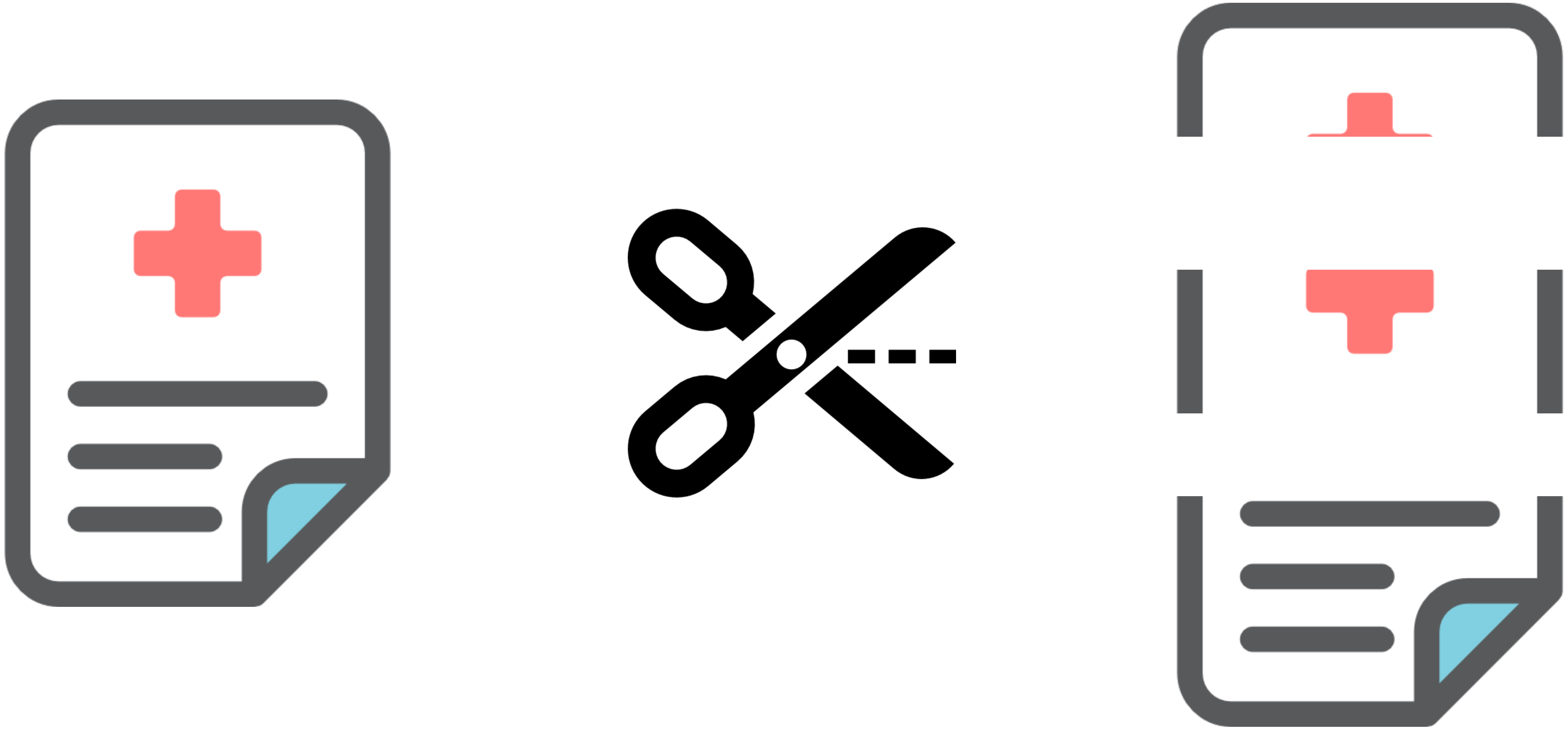
Dataset	Document Statistics			User Stats				Concept Stats	
	Doc	Vocab	Token-stats	User	Age	F	U-label	Concept	Type
Diabetes[1]	1265	34592	2426, 483, 42	288	63.13	0.45	10	68938	89
MIMIC-III[2]	54888	390237	7522, 1263, 50	48807	62.47	0.44	276	10761211	94

- Medical notes are too long to fit into neural model.
    - e.g., regular BERT models only fit for 512 tokens.

1. Stubbs et al. Cohort selection for clinical trials: n2c2 2018 shared task track 1.
2. Johnson, et al. Mimic-iii, a freely accessible critical care database.

Short Snippets by Random Split

- This simulates clinical settings in real-world that physicians can recognize their patients with partial symptom descriptions.



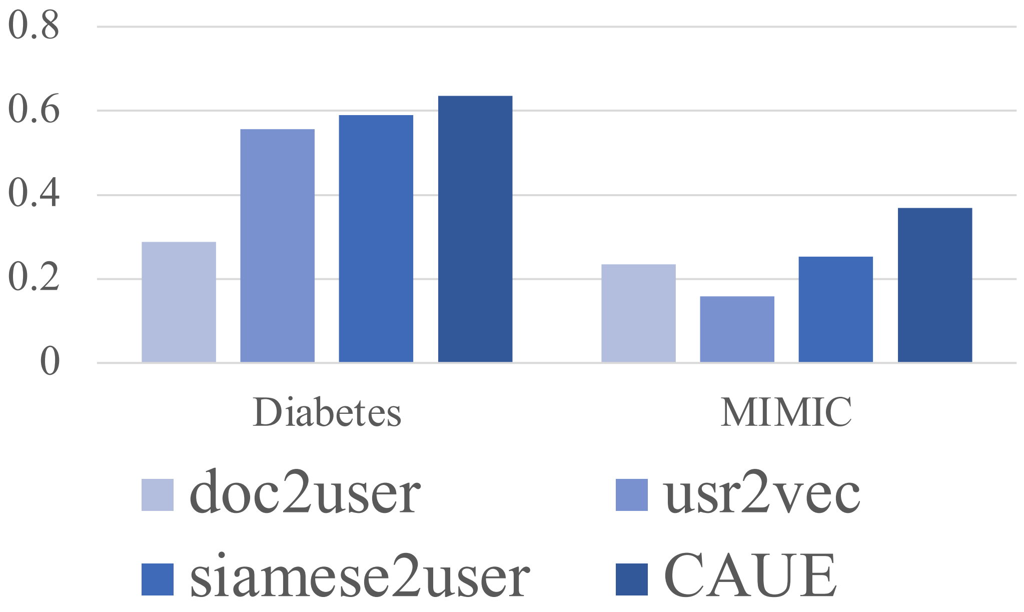
Evaluation

- Intrinsic Evaluation
  - Patient Retrieval
    - Potential application: Information Retrieval System
  - Patient Relatedness
    - Potential application: Cohort Selection
- Extrinsic Evaluation
  - Phenotype Inference
  - In-hospital Mortality Prediction

Evaluation

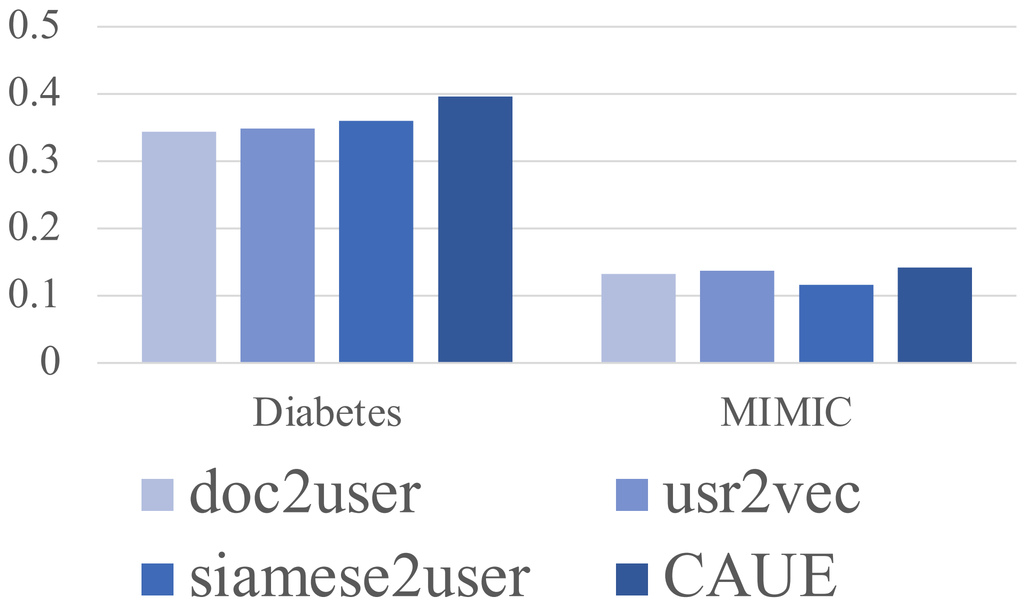
Phenotype Inference

MAP: Mean Average Precision



Patient Retrieval

$$Jaccard(u_1, u_2) = \frac{|l_1 \cap l_2|}{|l_1 \cup l_2|}$$



Effects of Medical Concepts

+ Concept	Phenotype Prediction		Mortality	Patient Relatedness		Retrieval	
	Diabetes	MIMIC-III	MIMIC-III	Diabetes	MIMIC-III	Diabetes	MIMIC-III
word2user	+7.6% (.044)	+40.4% (.130)	+4.8% (.041)	+3.4% (-.018)	+1.2% (-.010)	-0.5% (-.002)	+24.6% (.016)
doc2user	-43.4% (-.125)	-32.3% (-.076)	+2.8% (.012)	+1.1% (-.006)	+0.8% (-.02)	+6.1% (.021)	-3.0% (-.004)
dp2user	+22.6% (.094)	+52.1% (.164)	+3.8% (.033)	+2.0% (-.011)	+24.9% (-.207)	+17.6% (.056)	+18.6% (.021)
usr2vec	+9.4% (.052)	+249.0% (.394)	+111.1% (.470)	+6.2% (-.022)	-63.7% (.444)	+9.5% (.033)	-36.5% (-.050)
siamese2user	+7.6% (.045)	+45.8% (.116)	+45.8% (.116)	+28.7% (-.118)	+6.0% (-.013)	+7.2% (.026)	+19.8% (.023)
Average	+.8% (.013)	+61.84 (.146)	+33.66% (.134)	+8.28% (-.035)	-6.16% (.194)	+8.0% (.027)	+4.7 (.001)
Median	+9.4% (.052)	+45.8% (.116)	+4.8% (.041)	+3.4% (-.018)	+1.2% (-.010)	+7.2% (.026)	+18.6% (.021)

Performance gains of user embedding models combining with medical concepts (+Concept) comparing to standard non-concept information.

Summary

- Our proposed *unsupervised* user embedding is effective to capture patient patterns with broad real-world applications, such as diagnosis, retrieval and cohort selection.

- Medical concepts significantly boost patient modeling & potentially enhance interpretation.

- https://github.com/xiaoleihuang/UserEmb_Explainable