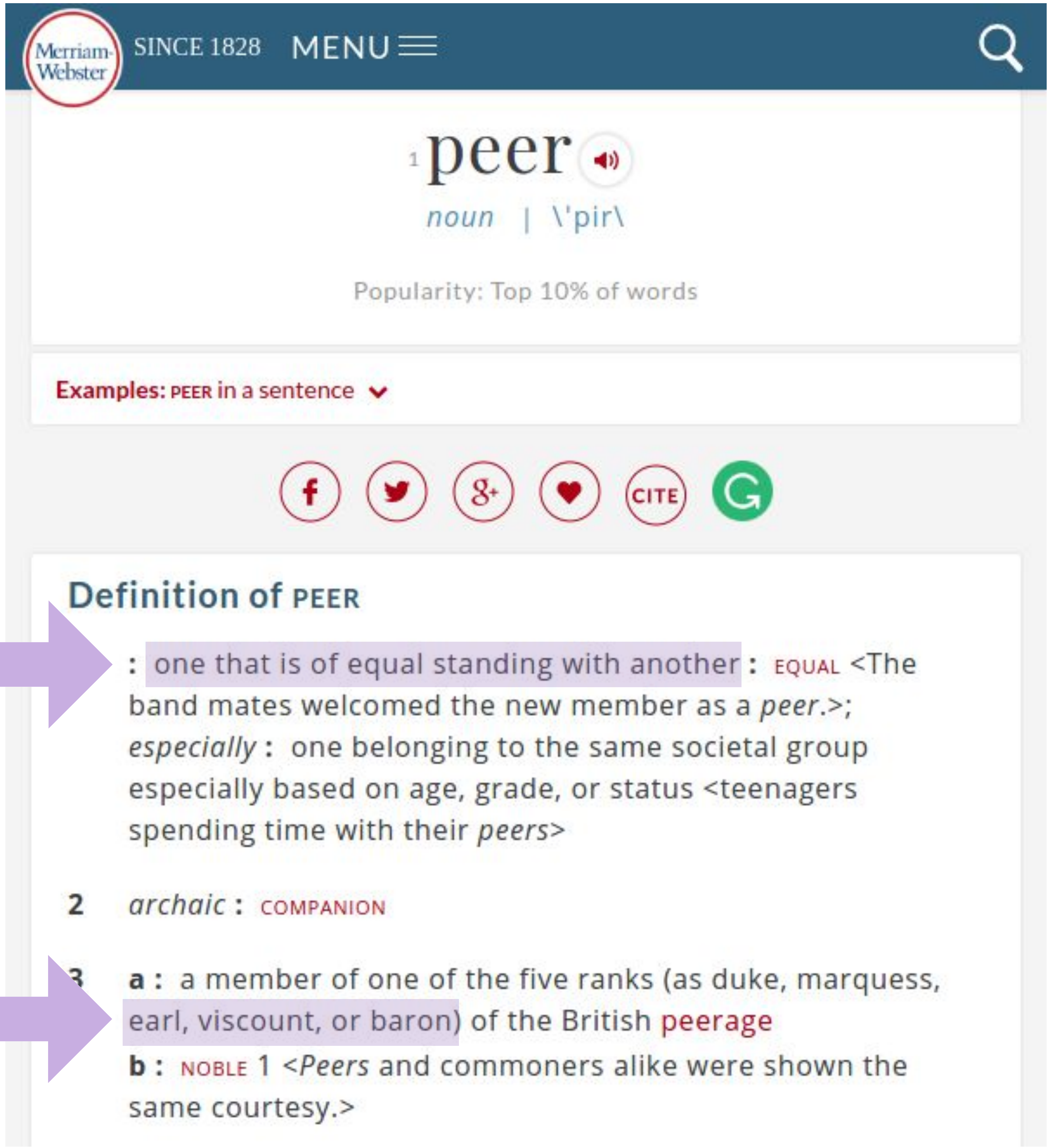
Probabilistic Embedding Models

Robert Bamler, UC Irvine

Internal Blockchain Meeting
7 October 2019

Who Are Your **Peers**?


Merriam-Webster SINCE 1828 MENU

[1] peer
noun | \'pir\

Popularity: Top 10% of words

Examples: PEER in a sentence

CITE

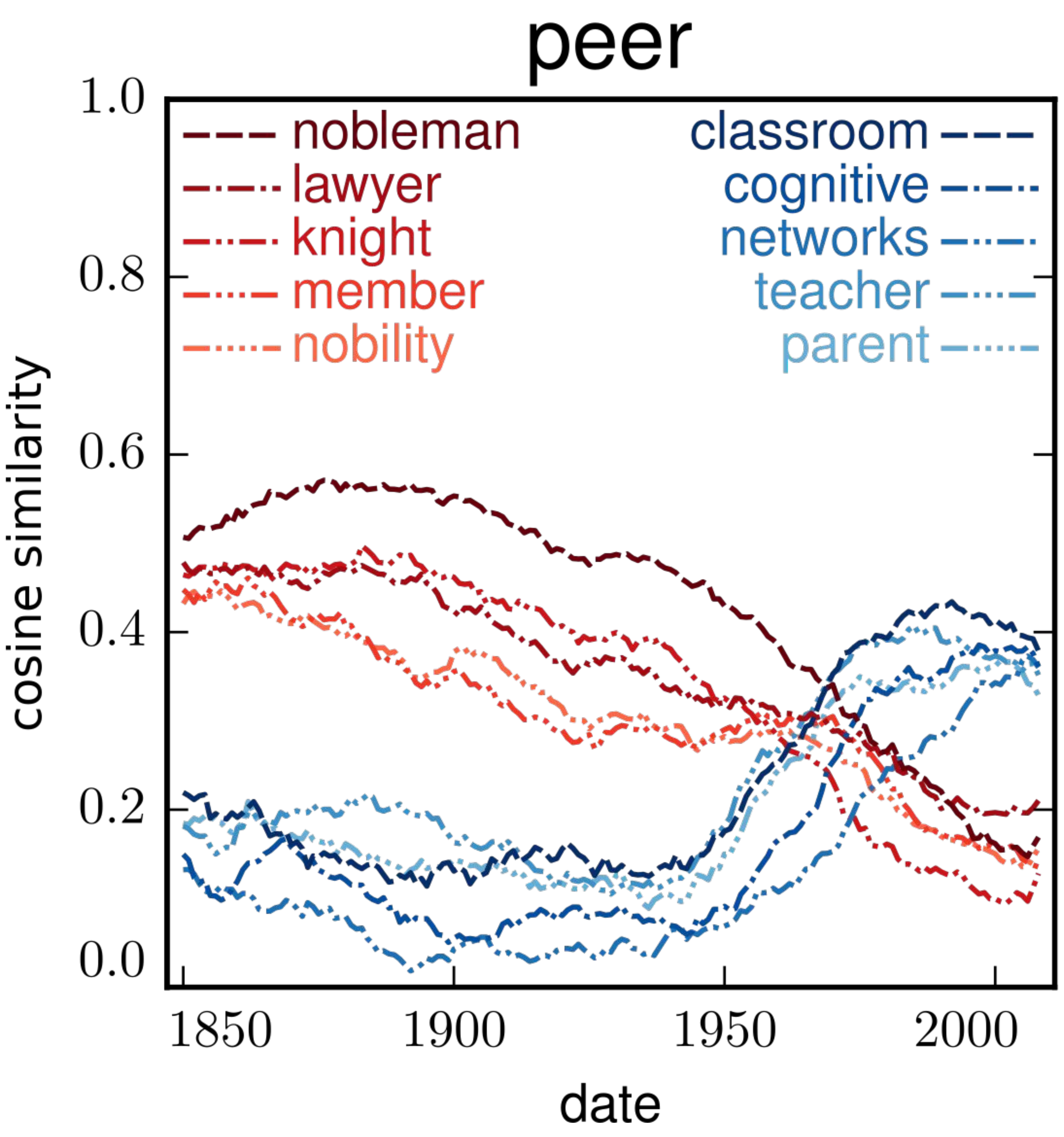
Definition of PEER

1 : one that is of equal standing with another : EQUAL <The band mates welcomed the new member as a *peer*.>; *especially* : one belonging to the same societal group especially based on age, grade, or status <teenagers spending time with their *peers*>

2 *archaic* : COMPANION

3 **a** : a member of one of the five ranks (as duke, marquess, earl, viscount, or baron) of the British peerage
b : NOBLE 1 <*Peers* and commoners alike were shown the same courtesy.>

https://www.merriam-webster.com/dictionary/peer



Word Embeddings

Word Embeddings

- Capture **semantic relations:**

"king"  − *"man"*  + *"woman"*  ≈ *"queen"* 

- Used for **transfer learning** in natural language processing, e.g., for **sentiment analysis:**

 *vs.* 

Image sources: Lego; Pixar; Joel Rouse (Ministry of Defence) & nagualdesign, OGL v3.0

Distributional Hypothesis

[Rubenstein & Goodenough, 1965; Sahlgren, 2008]

Assumption: Words that appear in similar contexts are similar in meaning.

≈ ≈

[...] [...]

[...] we can not **dedicate** — we can not **consecrate** — we can not **hallow** , this ground [...]

(A. Lincoln, 1863)

Semantic relations between words:

[...]

When **Seth** had lived 105 years , he became the father of **Enosh** . [...]

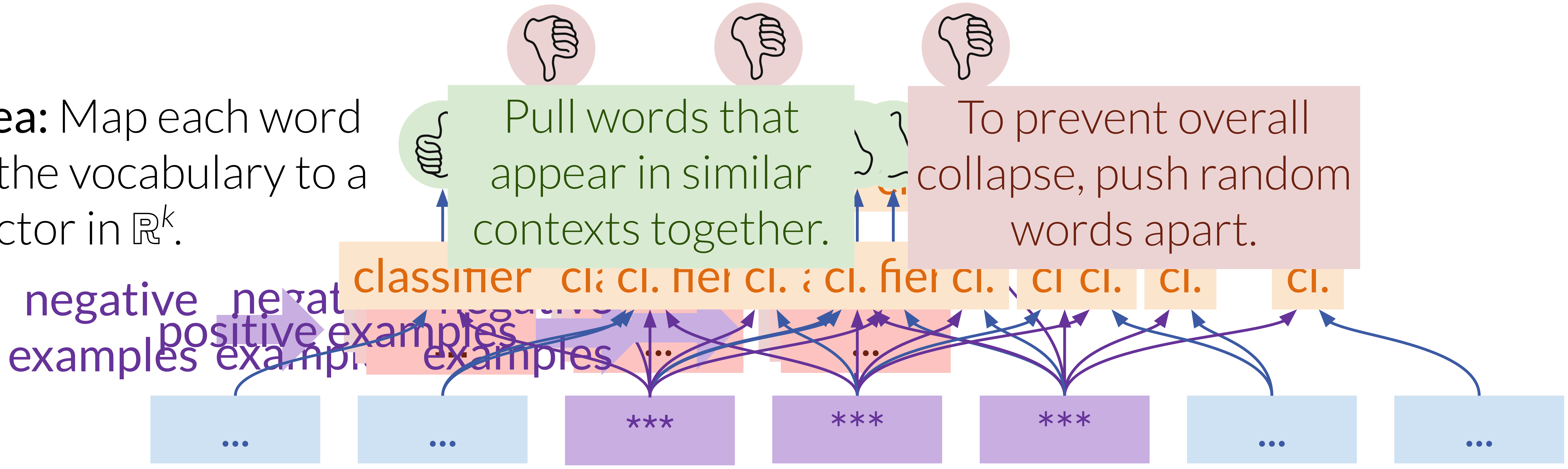
When **Enosh** had lived 90 years , he became the father of **Kenan** .

(Gen 5, 6 and Gen 5, 9)

"Neural" Word Embeddings: word2vec

[Mikolov et al., ICLR 2013 & NIPS 2013]

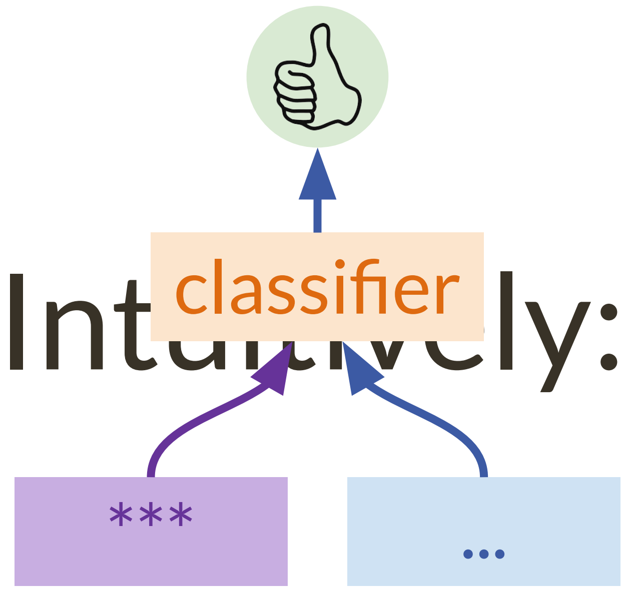
Idea: Map each word in the vocabulary to a vector in $\mathbb{R}^k$.

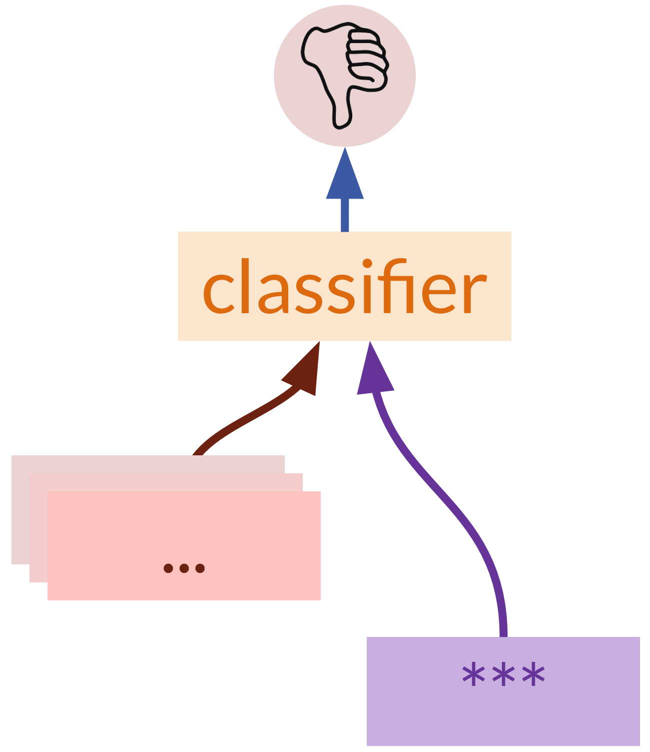


“Neural” Word Embeddings: word2vec

[Mikolov et al., ICLR 2013 & NIPS 2013]

Intuitively:





Minimize loss function: $$\ell = \overbrace{-\sum_{(i,j)\in\text{pos.}} \log \sigma(u_i^\top v_j)}^{\text{Pull words that appear in similar contexts together.}} \overbrace{-\sum_{(i,j)\in\text{neg.}} \log \sigma(-u_i^\top v_j)}^{\text{To prevent overall collapse, push random words apart.}}$$

u_i: **word embedding** (vector in $\mathbb{R}^k$)

v_j: **context embedding** (vector in $\mathbb{R}^k$)

Our Extension: Dynamic Word Embeddings

[Bamler & Mandt, ICML 2017]

"Computer" **in 1961**

© 20th century FOX

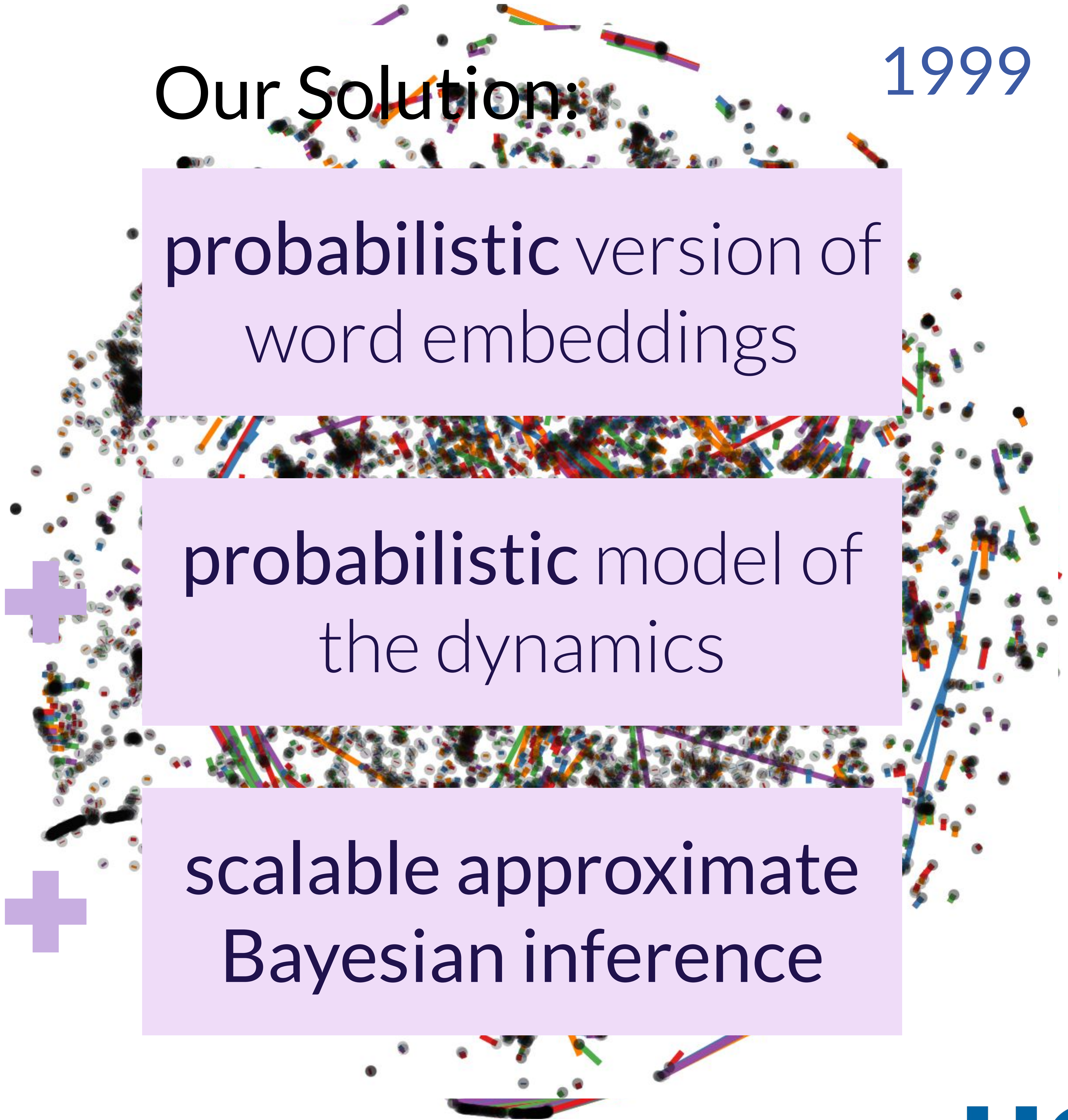
"Computer" **today**

Detecting **Subtle Changes** Over Time

[Bamler & Mandt, ICML 2017]

Naive idea:
Fit individual embedding vector for each word and each year.

Problem:
Only few data **per word & year**.
→ *small signal/noise ratio*



Our Solution:

probabilistic version of word embeddings

\+ **probabilistic** model of the dynamics

\+ **scalable approximate Bayesian inference**

UCI

1980

broadcast

radio

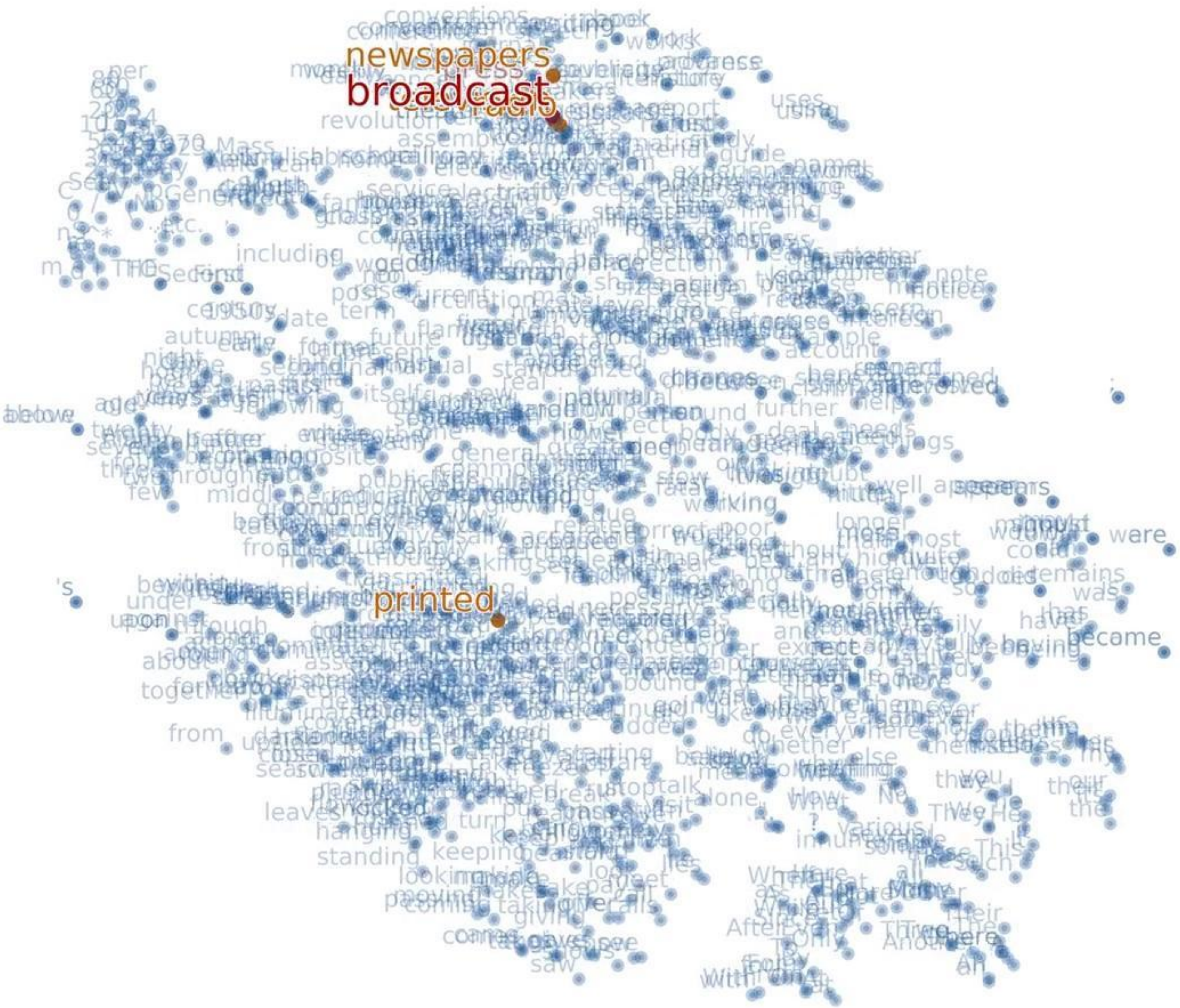
television

newspapers

TV

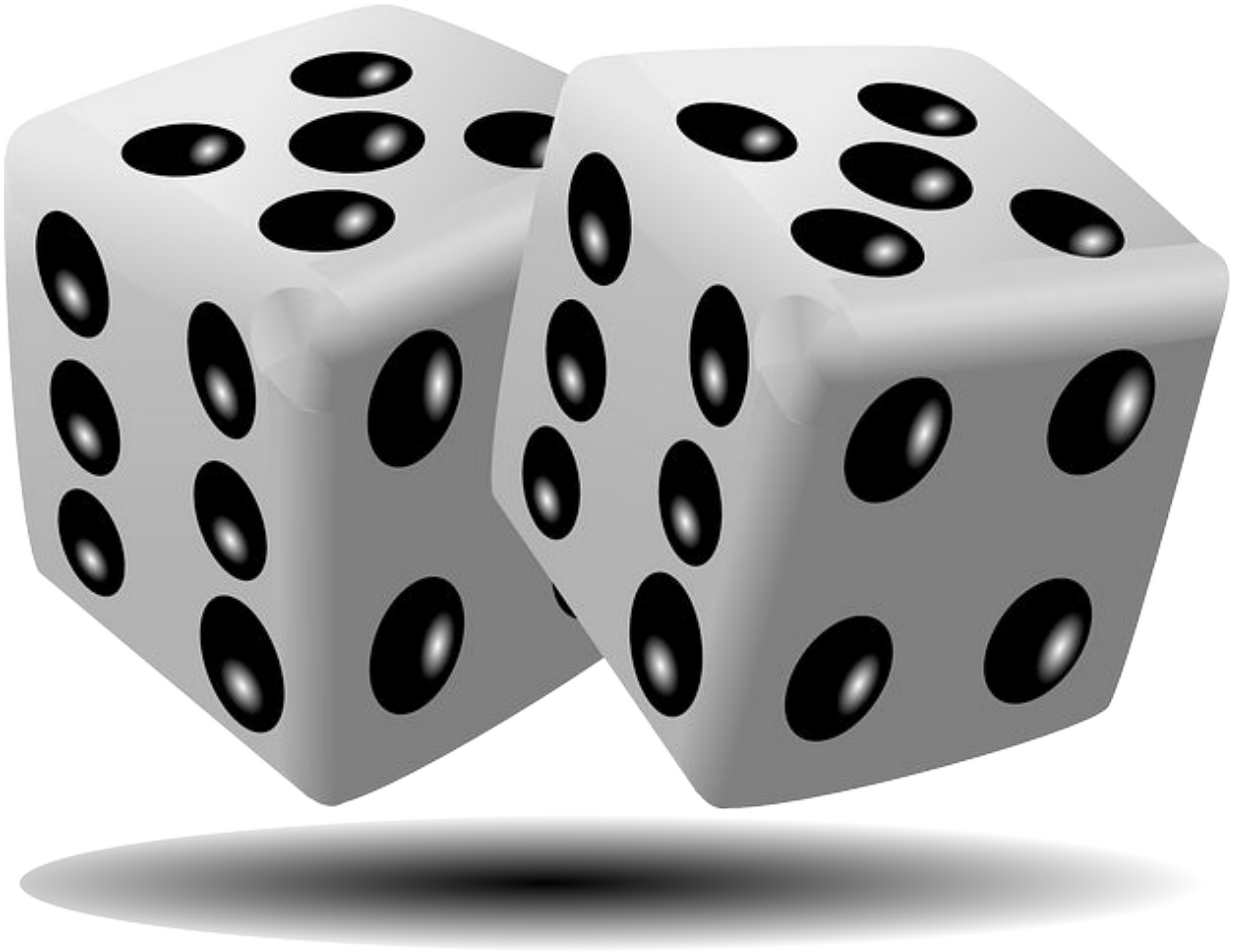
printed

press



Robert Bamler and Stephan Mandt, Dynamic Word Embeddings, ICML 2017.

Bear With Me: Probabilistic Models & Inference



Probabilistic Models & Bayesian Inference

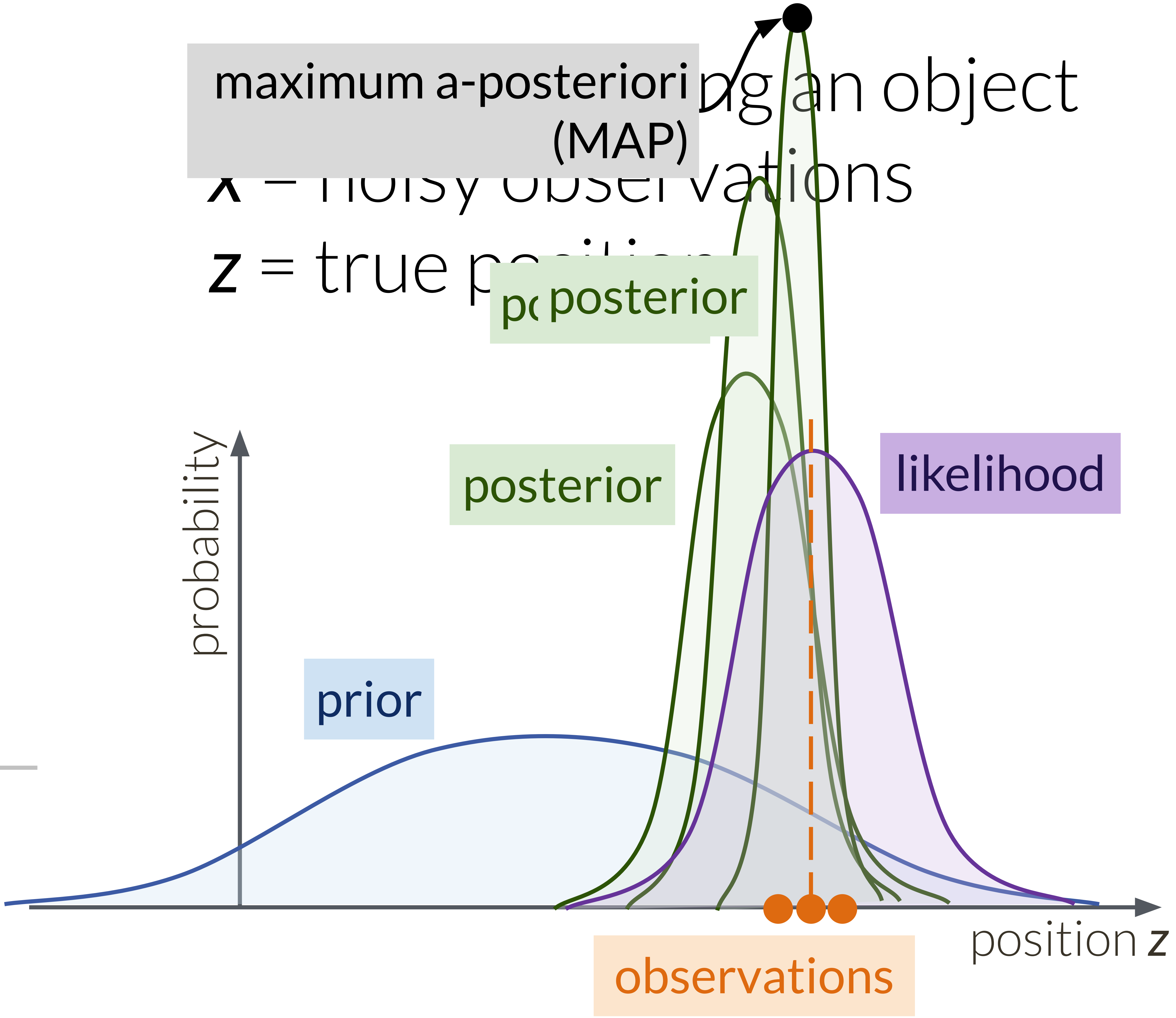
Notation: x = observations (data)
z = latent (i.e., unknown) variables that caused x

Probabilistic model: $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})\, p(\mathbf{x}|\mathbf{z})$

prior likelihood

Inference: find **posterior** $p(\mathbf{z}|\mathbf{x}) = \dfrac{p(\mathbf{x}, \mathbf{z})}{\int p(\mathbf{x}, \mathbf{z})\, dz}$

probability that latent variables $\mathbf{z}$ explain the observed data $\mathbf{x}$.



Example 1: Probabilistic Variant of word2vec

[Barkan AAAI 2017]

Reminder: word2vec minimizes loss $\ell = -\sum_{(i,j)\in\text{pos.}} \log \sigma(u_i^\top v_j) - \sum_{(i,j)\in\text{neg.}} \log \sigma(-u_i^\top v_j)$

word embedding
(vector in $\mathbb{R}^k$)

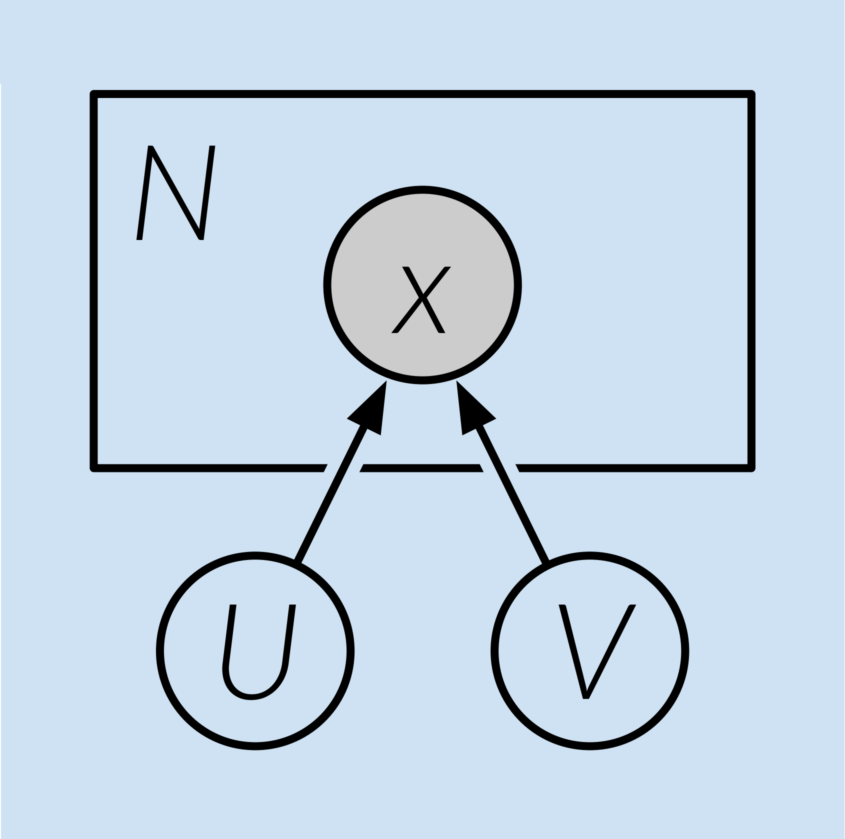
context embedding
(vector in $\mathbb{R}^k$)

Observation: $\ell = -\log p(x|z)$

Generative process:

repeat:

- Draw a random pair of words (i, j) from vocabulary.
- **With probability $\sigma(u_i^\top v_j)$:** mark (i, j) as a positive example;
  otherwise: mark (i, j) as a negative example.



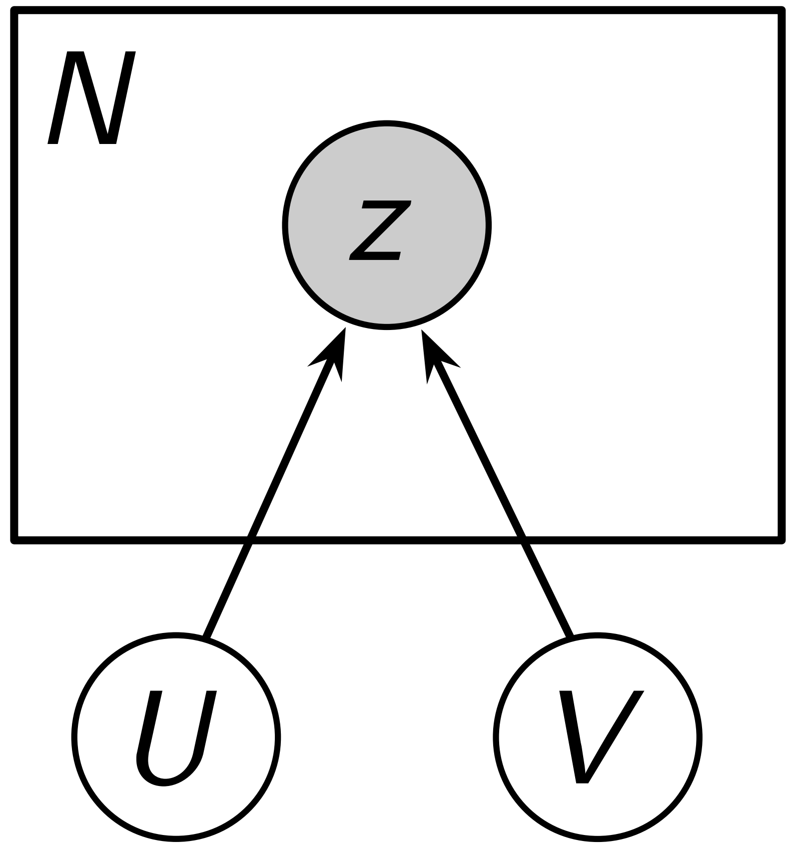
Enough Equations, **Back to Pretty Pictures**

Nymphenburg Park, Munich #nofilter

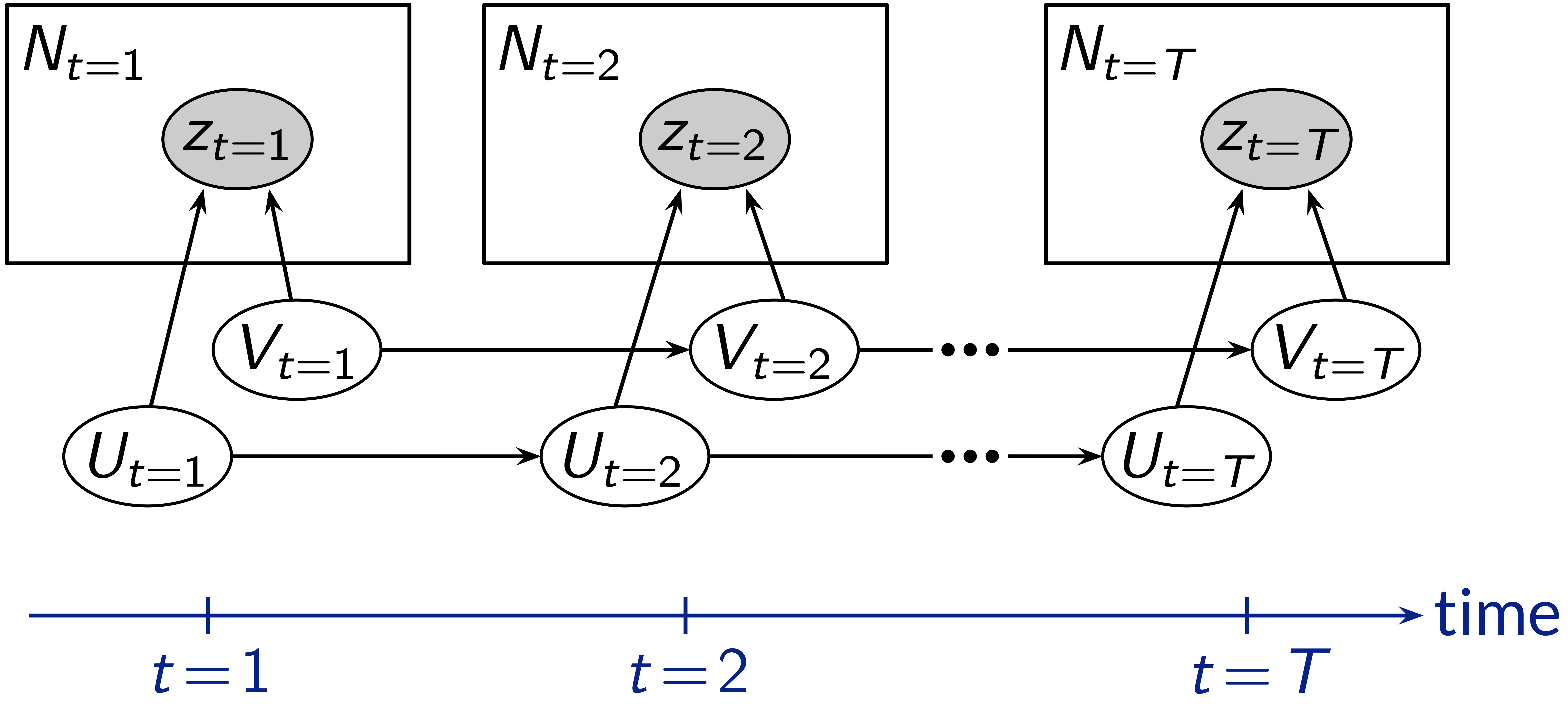
Dynamic Word Embeddings Model

[Bamler & Mandt, ICML 2017]

static model

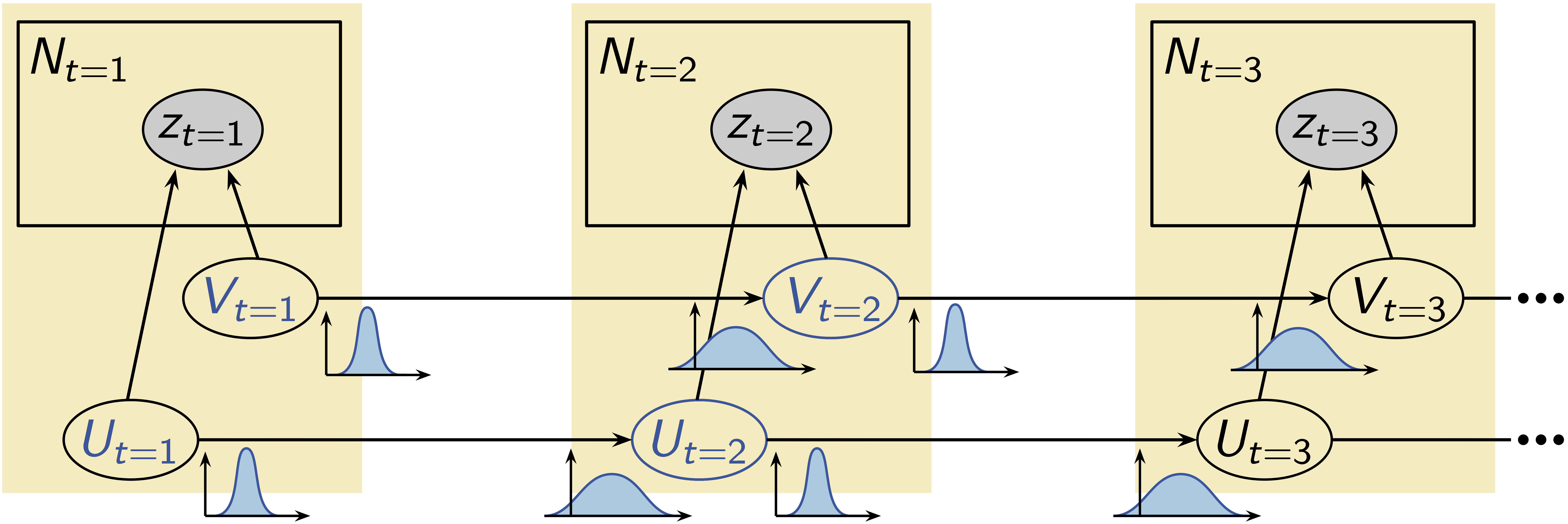


Dynamic Model



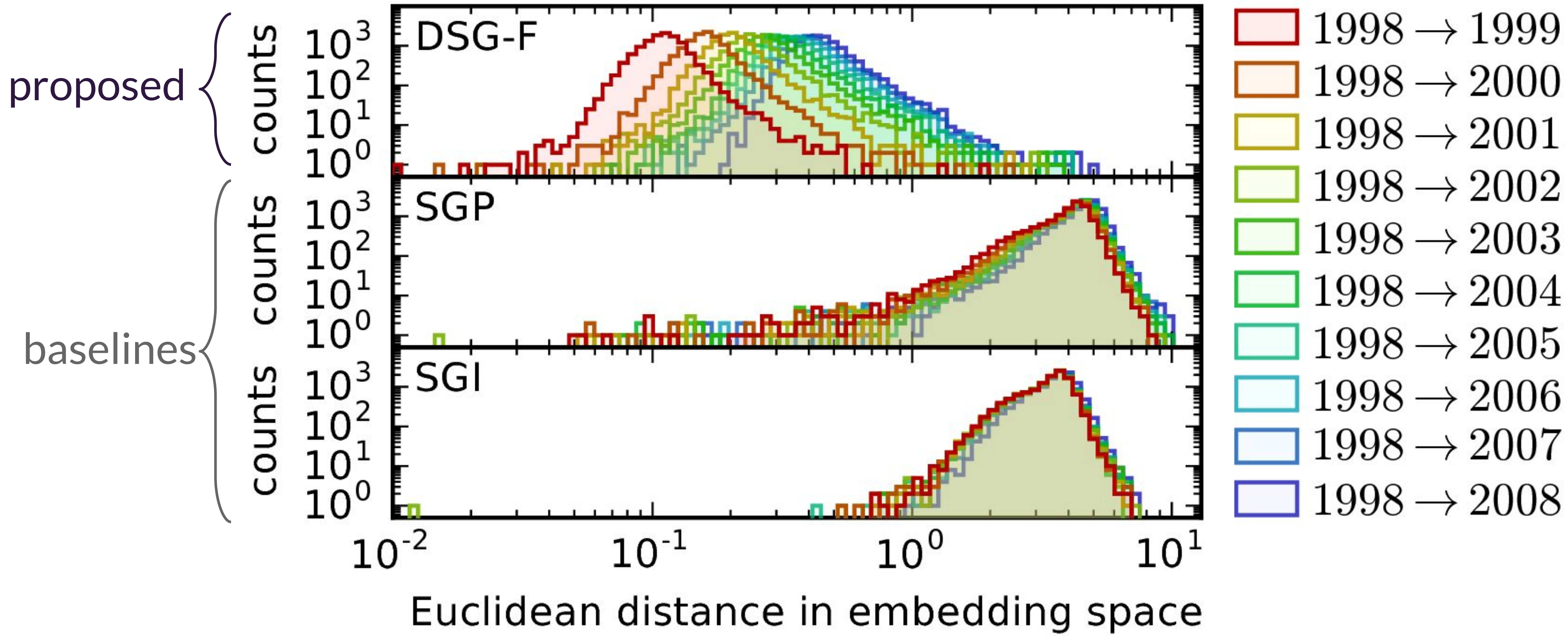
On-line Learning With Variational Inference

[see "Dynamic Word Embeddings", Bamler & Mandt, ICML 2017]



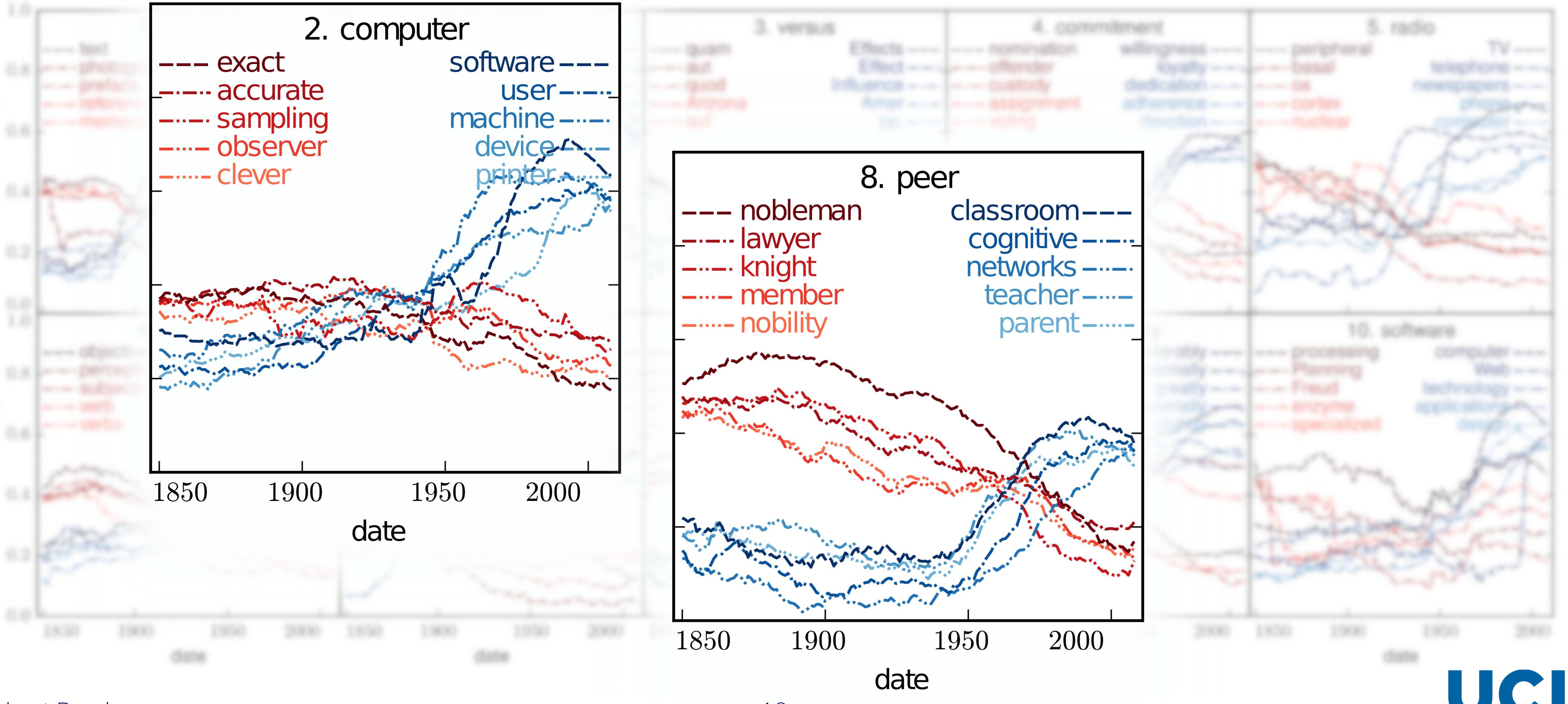
Results III: Smooth Trajectories in Embedding Space

(Training data: Google Books corpus [Michel et al., 2011])

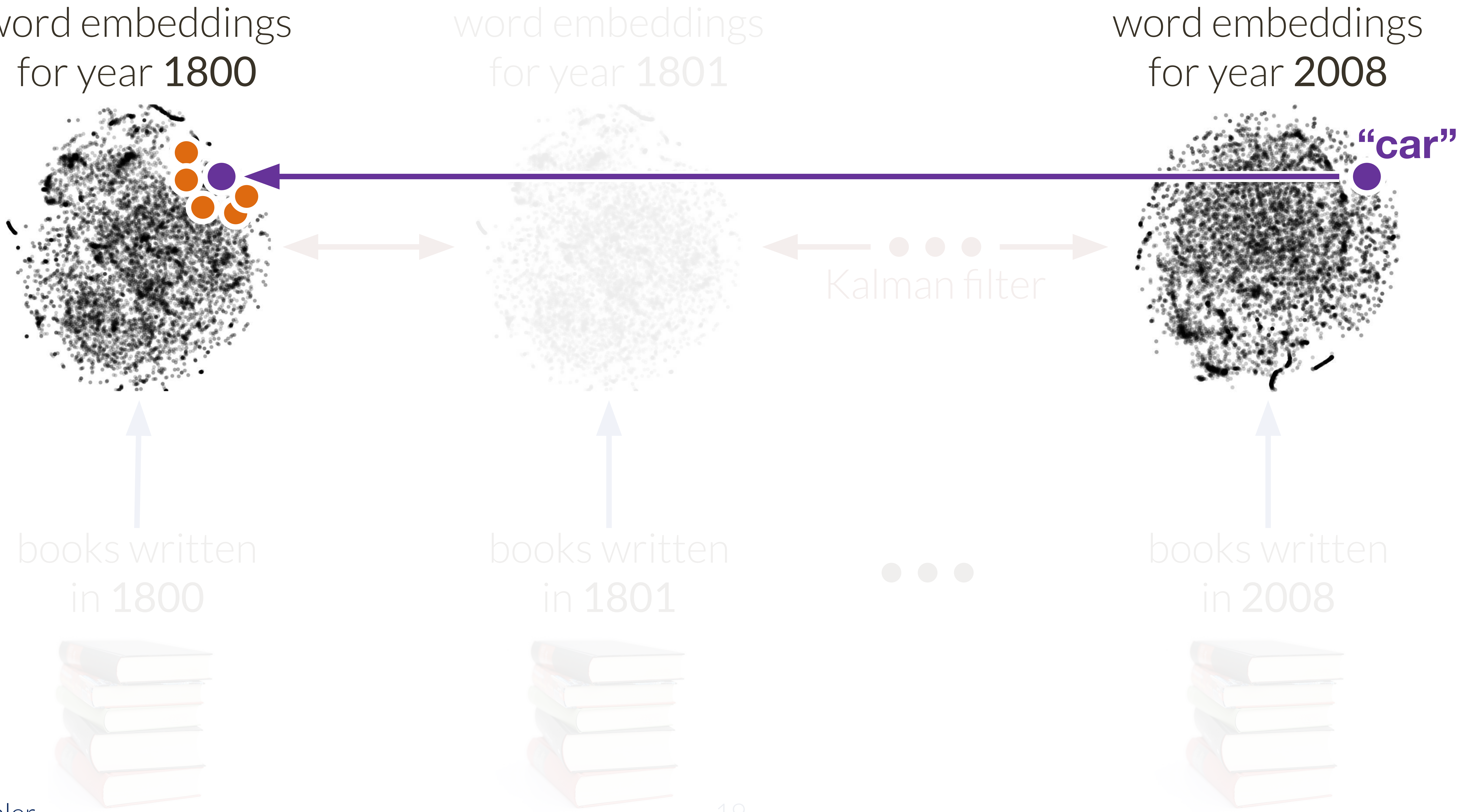


Results: Top 10 Most Mobile Words (1850→2008)

(Training data: Google Books corpus [Michel et al., 2011])



Time Travel


word embeddings
for year 1800
word embeddings
for year 1801
word embeddings
for year 2008
"car"
Kalman filter
books written
in 1800
books written
in 1801
books written
in 2008

Results: Word Aging with Goldstone GD

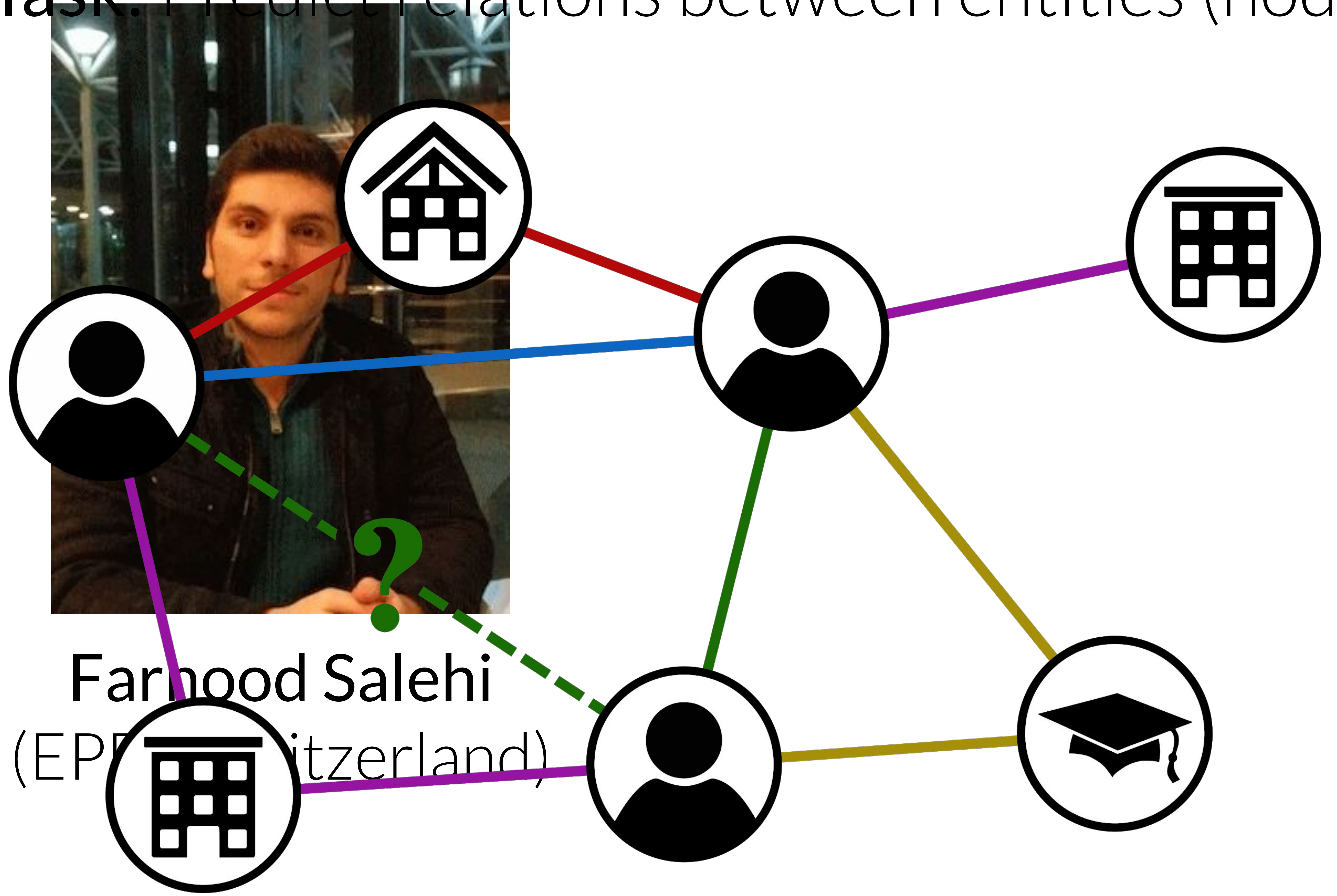
[Bamler & Mandt, ICML 2018]

2008	1800
car	boat, saddle, canoe, wagon, box
DNA	potassium, chemical, sodium, molecules, displacement
tuberculosis	chronic, paralysis, irritation, disease, vomiting

Example 2: Probabilistic Knowledge Graphs

[Bamler, Salehi & Mandt, UAI 2019]

Task: Predict relations between entities (nodes) in a knowledge graph.



Farhood Salehi
(EP[illegible]itzerland)

Example 2: Probabilistic Knowledge Graphs

[Bamler, Salehi & Mandt, UAI 2019]

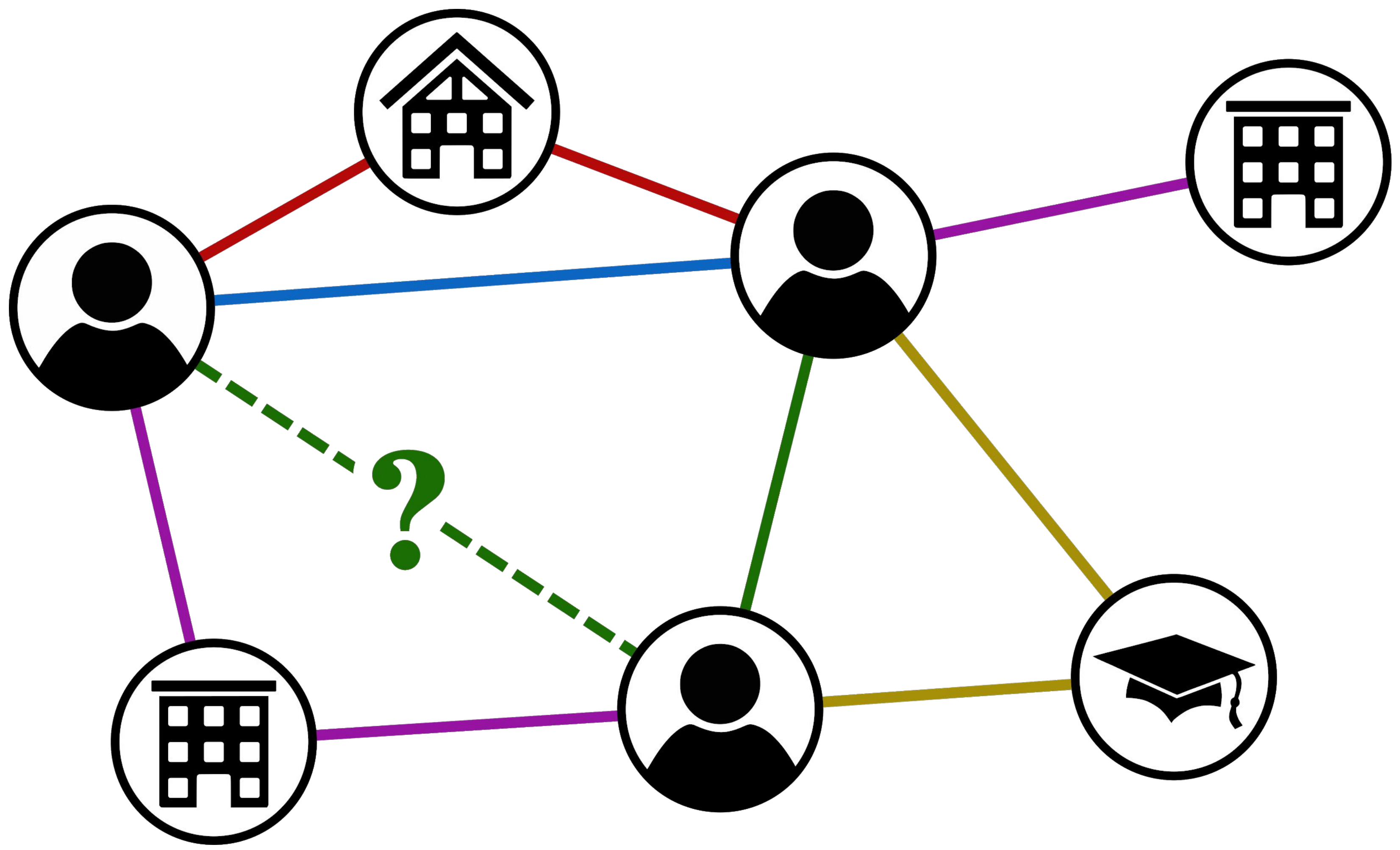
Task: Predict relations between entities (nodes) in a knowledge graph.

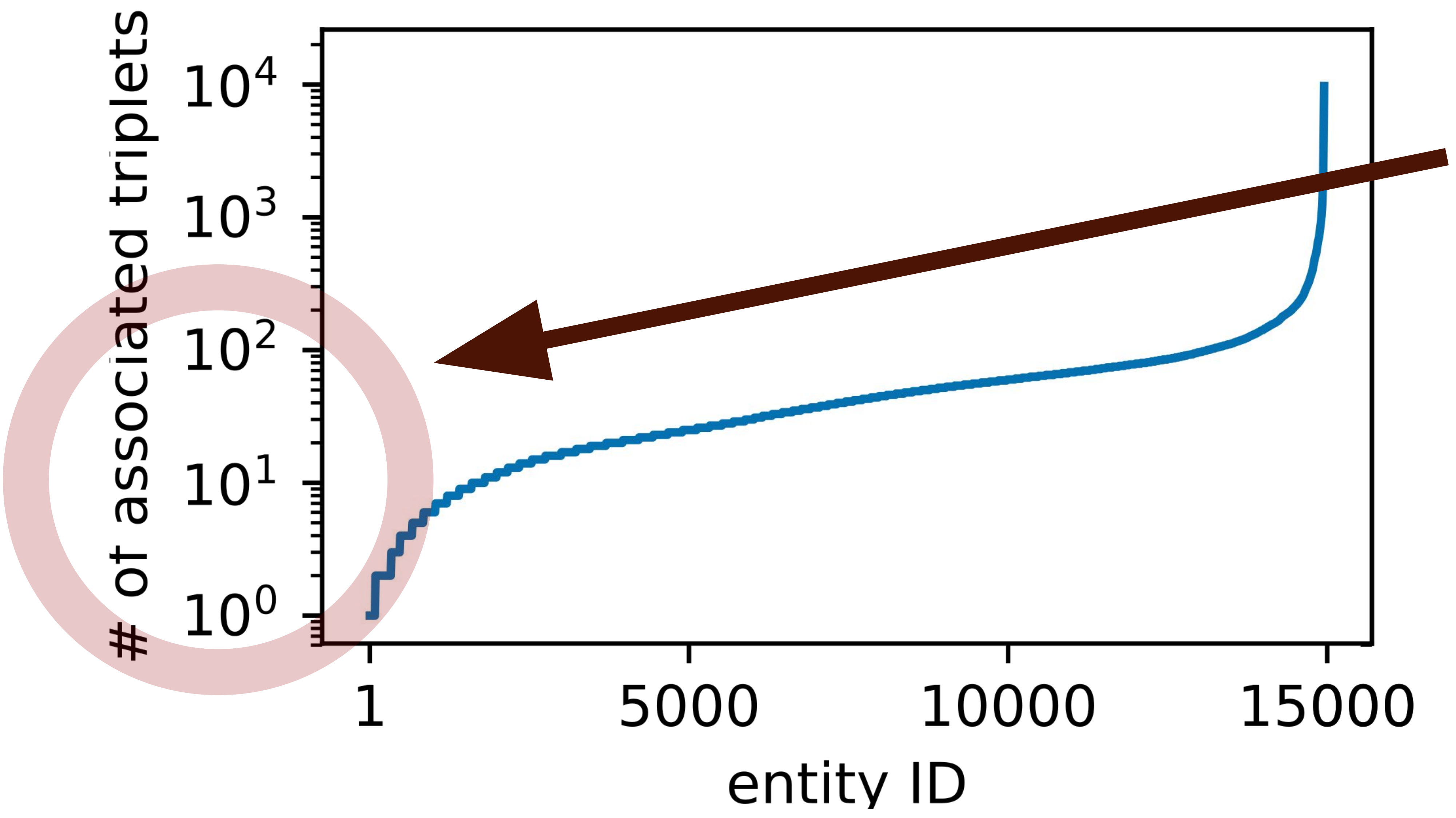
State of the Art:

Learn embeddings for entities and relation types.

Problem:

Highly sensitive to hyperparameters [Kadlec et al., 2017]





Reason:
Many entities are supported only by few data points.

Example 2: Probabilistic Knowledge Graphs

[Bamler, Salehi & Mandt, UAI 2019]

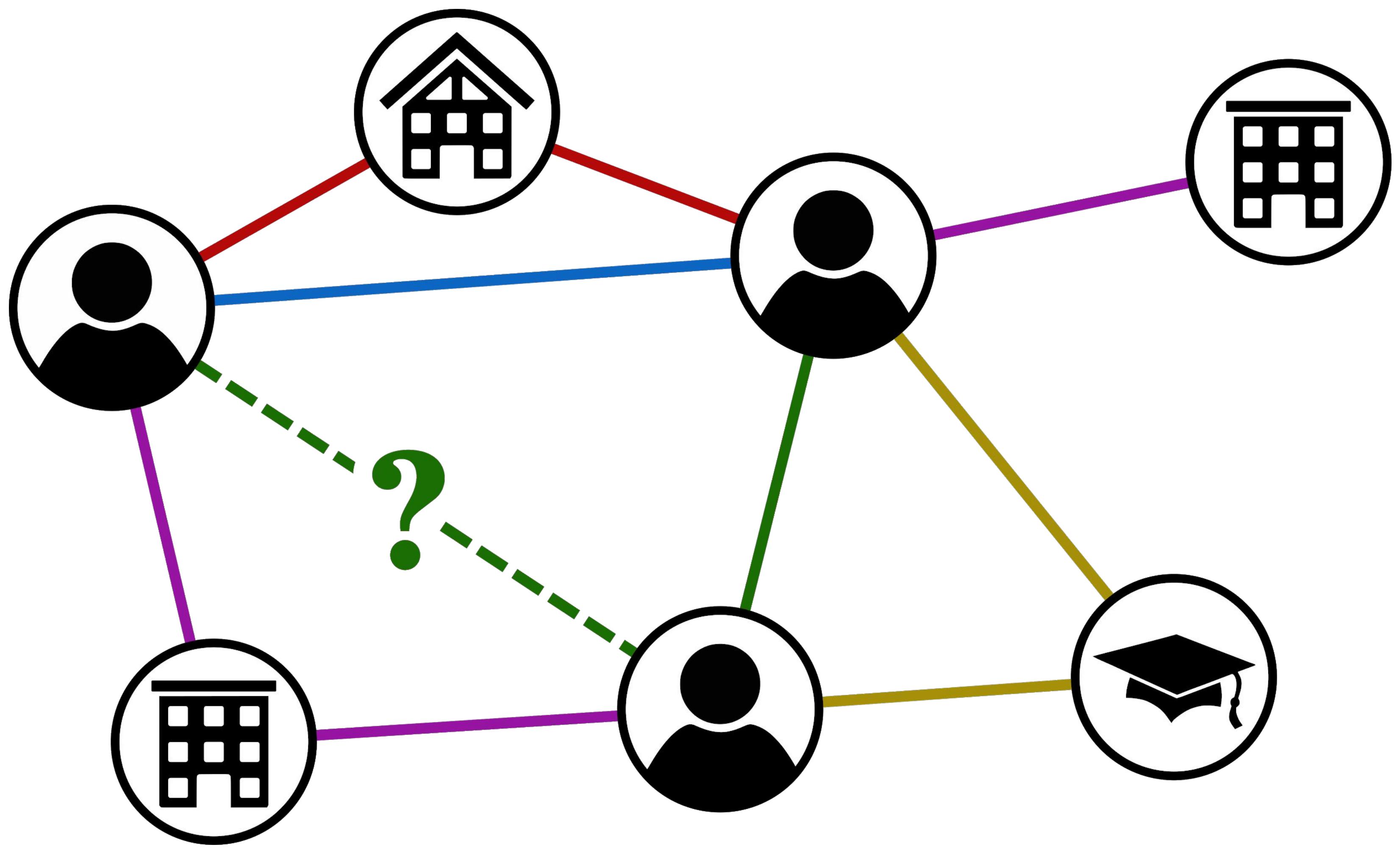
Task: Predict relations between entities (nodes) in a knowledge graph.

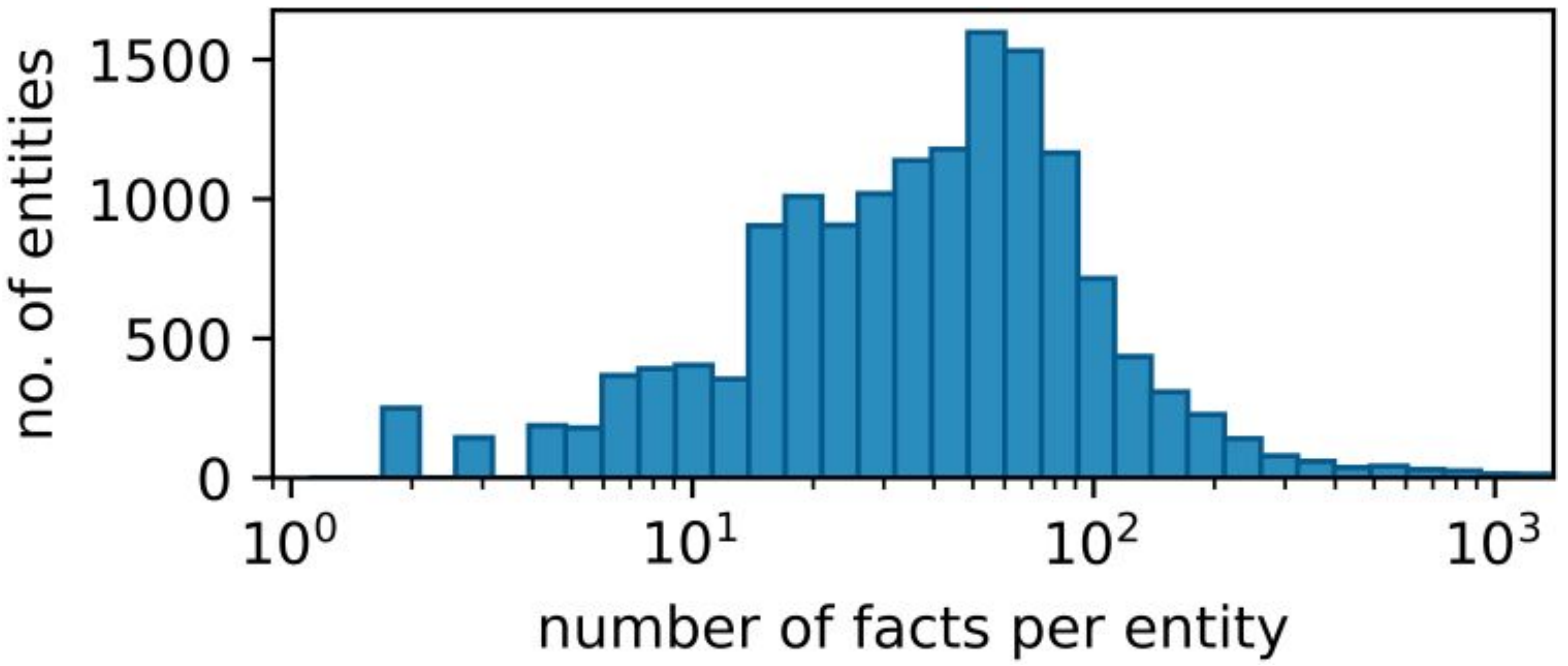
State of the Art:

Learn embeddings for entities and relation types.

Problem:

Highly sensitive to hyperparameters [Kadlec et al., 2017]





Example 2: Probabilistic Knowledge Graphs

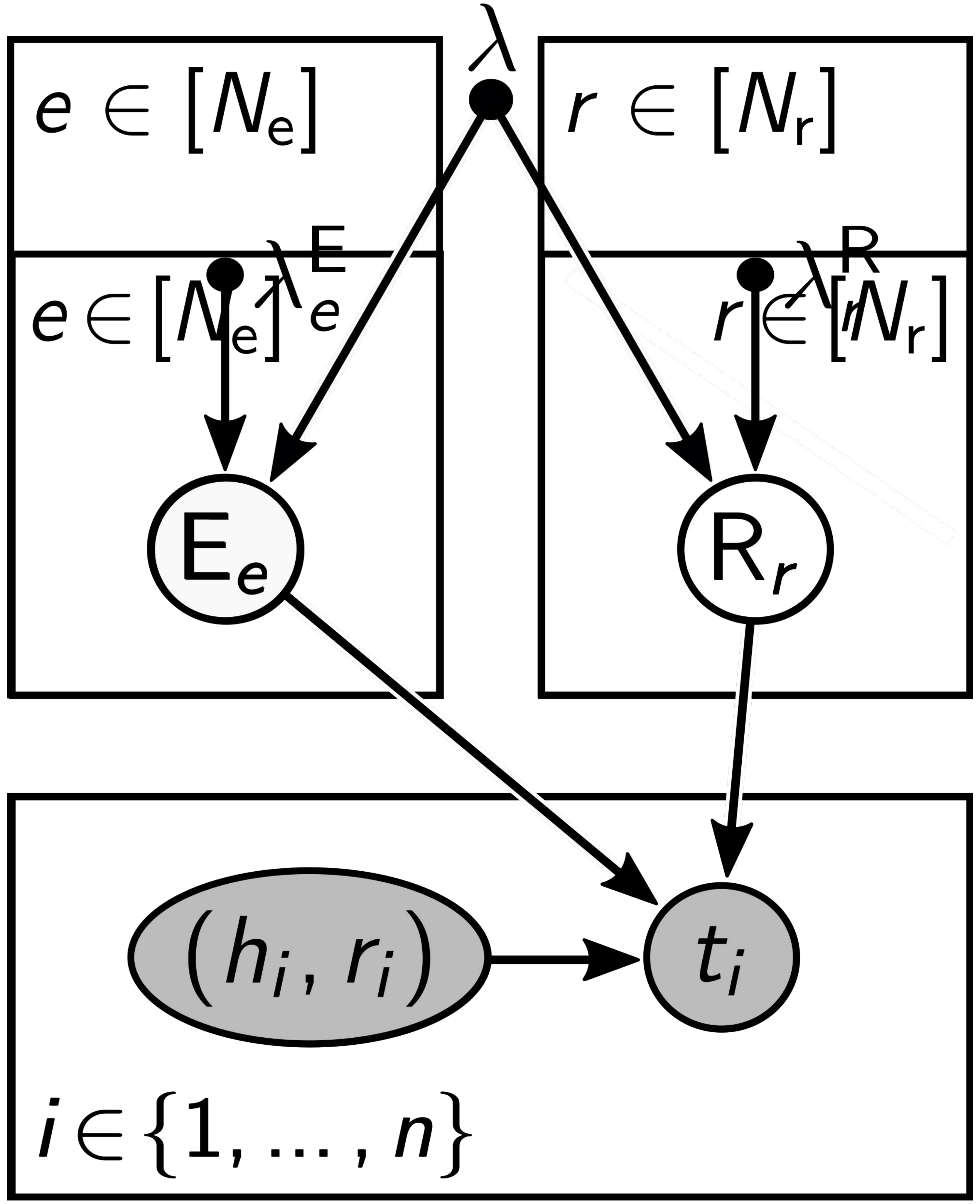
[Bamler, Salehi & Mandt, UAI 2019]

Our Solution:

→ Reinterpret existing models as **probabilistic generative models** of relational facts (***h**ead*, ***r**elation*, ***t**ail*).

→ Introduce macroscopic number of *local* hyperparameters **(> 10,000)**.

→ Tune hyperparameters efficiently with **variational expectation maximization**.



Results: Probabilistic Knowledge Graph Embeddings

[Bamler, Salehi & Mandt, UAI 2019]

Link prediction outperforms previous state of the art.

	data set →	WN18RR		WN18		FB15K-237		FB15K	
↓ model	↓ variant metric →	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10
DistMult	Yang et al. [2015] (orig.)	–	–	0.83	0.942	–	–	0.35	0.577
DistMult	Kadlec et al. [2017]	–	–	0.790	0.950	–	–	0.837	0.904
DistMult	Dettmers et al. [2018]	0.43	0.49	0.822	0.936	0.241	0.419	0.654	0.824
DistMult	Ours (after variational EM)	**0.455**	**0.544**	**0.911**	**0.961**	**0.357**	**0.548**	**0.841**	**0.914**
ComplEx	Trouillon et al. [2016] (orig.)	–	–	0.941	0.947	–	–	0.692	0.840
ComplEx	Lacroix et al. [2018]*	0.478	0.569	0.952	0.963	0.364	0.555	**0.857**	0.909
ComplEx	Ours (after variational EM)	**0.486**	**0.579**	**0.953**	**0.964**	**0.365**	**0.560**	0.854	**0.915**

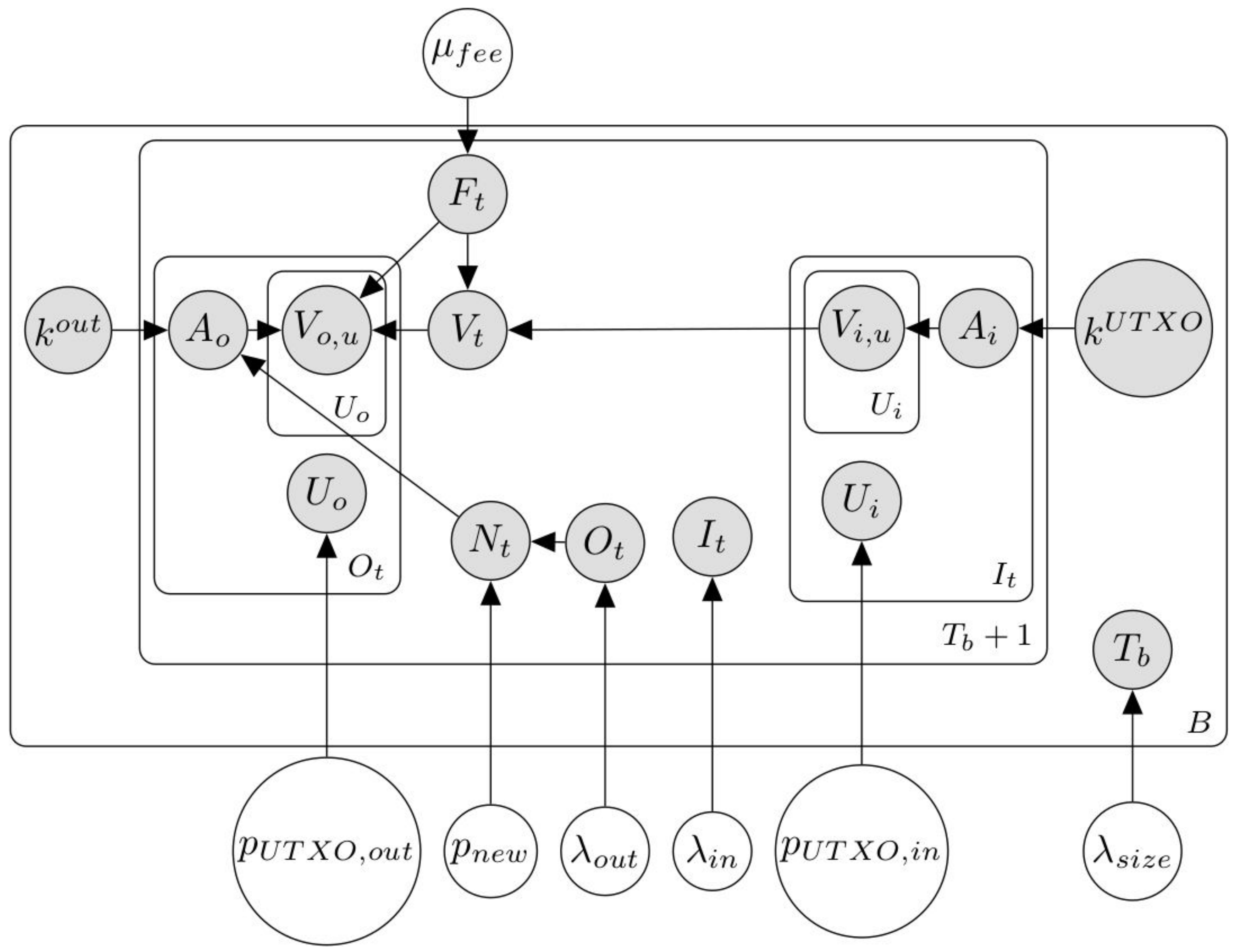
Discussion: Embedding the Blockchain

- **Analogy to knowledge graphs:**
  transaction ≈ relational fact; token ≈ relation

- **Temporal component:**
  ≈ Dynamic Word Embeddings, but more ephemeral.

- **Analysis (ideas):**
  - semantic analysis of users & tokens using embeddings
  - predict transactions?

- **Action Items:**
  - *Get a data set.*
  - Discuss existing literature.
  - Come up with more concrete analysis questions (maybe as we look at the data).



[Jourdan et al., 2019]