Word Embeddings for Descriptive Corpus Analysis

Digging Deeper into Analogies, Polysemy, and Stability

Slides: bit.ly/embedding-tutorial-slides
Notebook: bit.ly/embedding-tutorial-notebook

Neha Kennard @nnkennard

May 9 2022

Overview

Goals

- Revisit mental models
- Learn about possible pitfalls

- Explore word embedding analogies
- Talk about how phenomena in lexical semantics affect word embeddings' behavior
- Look at factors that can affect the stability of word embeddings

Overview

Non-goals: things we won't discuss

- BERT
  - NLP+CSS: BERT for Computational Social Scientists
- How to train word embeddings (e.g. hyperparameter selection)
  - Spirling and Rodriguez
- Comparing word embedding models
  - NLP+CSS: Comparing Word Embedding Models
- Word embeddings in languages besides English

Recap

What are word embeddings?

- A representation of a word in text
- Takes the form of a vector in R^d
- Learned by a class of algorithms (including word2vec, GloVe, …)
- Represents the meaning of a word

Recap

What are word embeddings?

- A representation of a word in text
- Takes the form of a vector in R^d
- Learned by a class of algorithms (including word2vec, GloVe, ...)
- Represents the “meaning” of a word*

- * as it is used in a corpus

Recap

What are word embeddings?

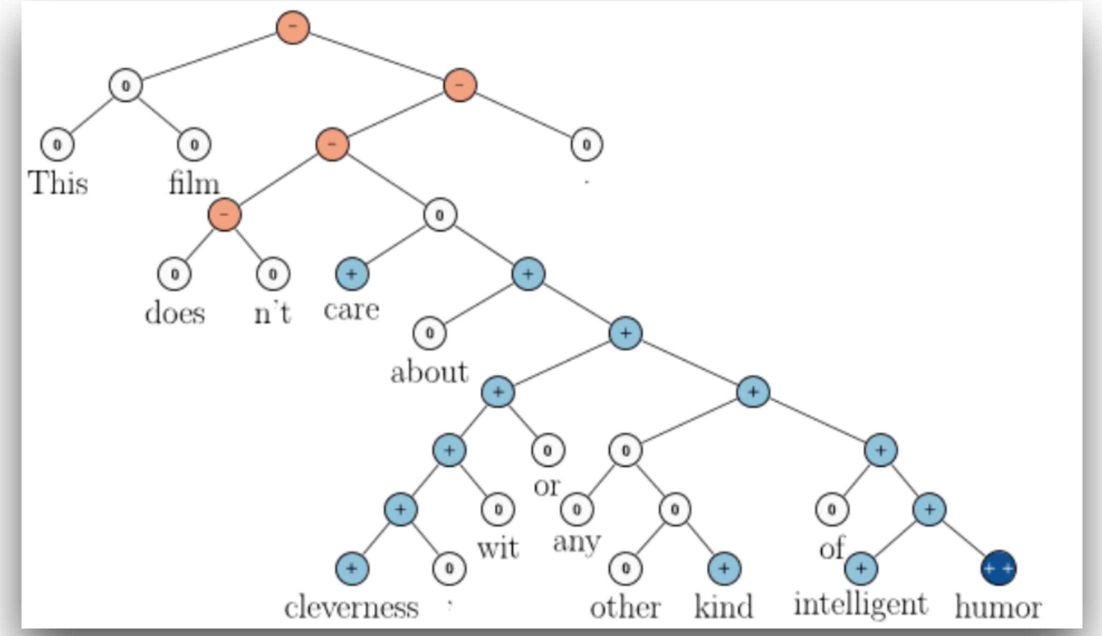
- A representation of a word in text
- Takes the form of a vector in R^d
- Learned by a class of algorithms (including word2vec, GloVe, …)
- Represents the "meaning"** of a word*

- * as it is used in a corpus
- ** if we accept that a description of a word's contexts constitute its meaning

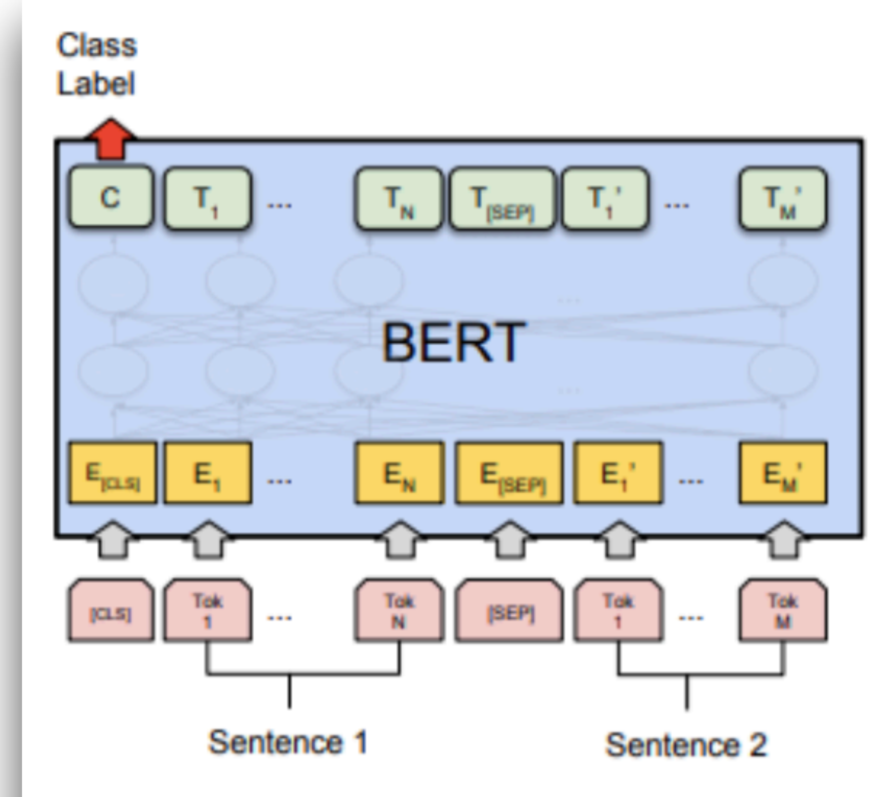
Recap

Two different uses of word embeddings

- **Downstream-centered***
  - Incorporated into deep learning models
- **Corpus-centered***
  - Directly studied as a representation of the mental models of the producers of corpus text



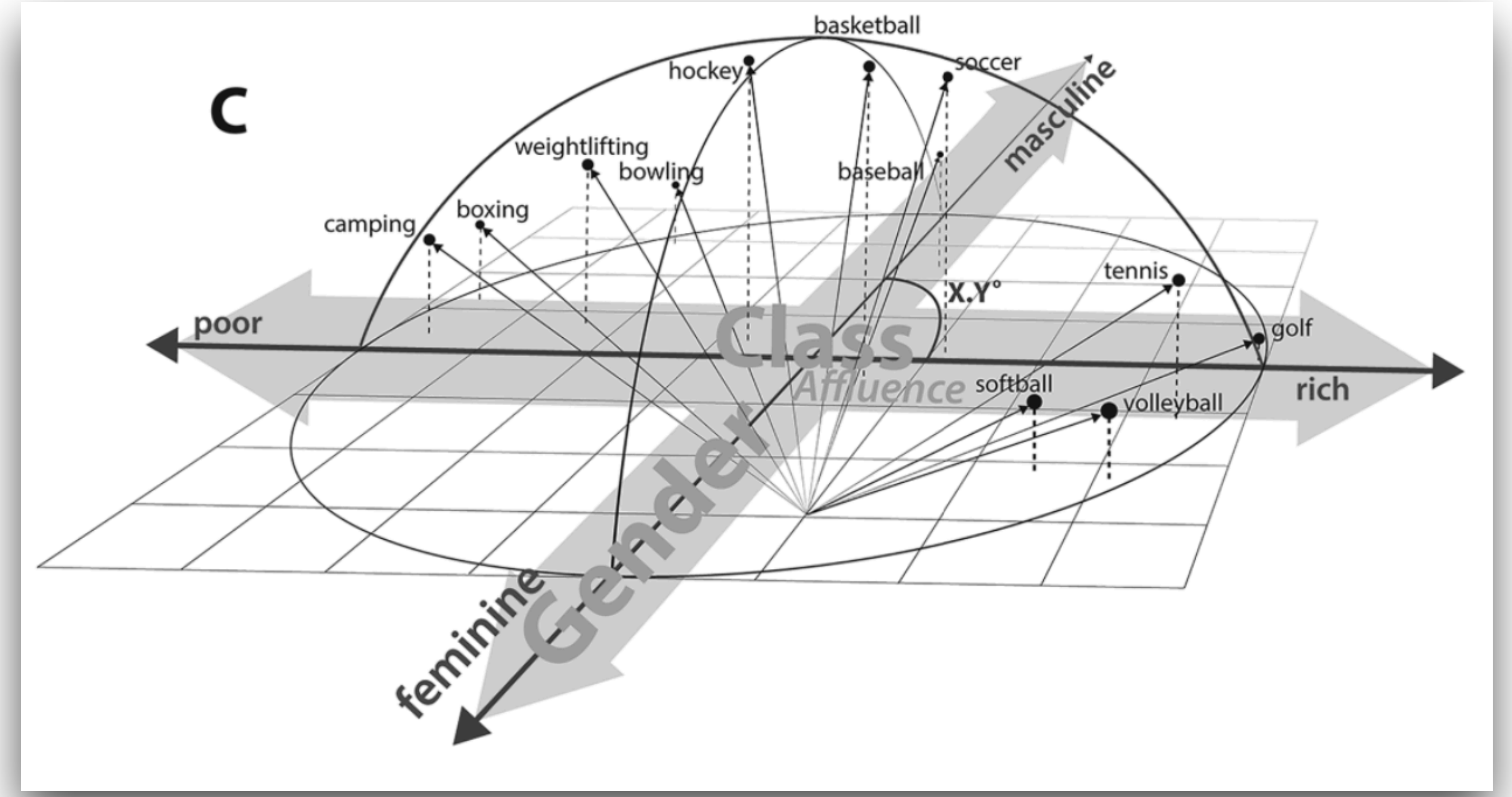
Sentiment analysis: Socher et al. 2013



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

BERT: Devlin et al 2019

Diachronic word embeddings: Hamilton et al. 2016



Semantic projection: Kozlowski et al. 2019

* terminology from Antoniak and Mimno 2018

Overview

Corpus-centered social science research with word embeddings

- Emotion and reason in political language (Gennaro and Ash 2022)
- Construct an emotion/reason dimension using vector representations for ‘affect’ and ‘cognition’ in US Congress speeches
- Findings
  - Emotionality spikes in times of war, and with patriotism
  - Emotionality is higher for:
    - Democrats, women, ethnic/religious minorities, the opposition party, members with ideologically extreme roll-call voting records.

Overview

Finding needles in haystacks

word2vec (among other pubications): Mikolov et al. 2013

TITLE	CITED BY	YEAR
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Neural information processing systems	33113	2013

GloVe: Pennington et al. 2014

TITLE	CITED BY	YEAR
Glove: Global vectors for word representation J Pennington, R Socher, CD Manning Proceedings of the 2014 conference on empirical methods in natural language …	27700	2014

Almost all of these citations are for downstream-centered tasks!

Today's focus is on corpus-centered uses of word embeddings.

The meaning of ‘meaning’

- Word embedding algorithms convert *co-occurrence probabilities* into appropriate *cosine similarities*
- Why is this encoding ‘meaning’?
- We’ll revisit the objective functions of word2vec and GloVe

The meaning of 'meaning'

- GloVe and word2vec both build two representations for each word
  - As context word: a vector c for each word, in the matrix C
  - As focus word: a vector v for each word, in the matrix V
- The final word embeddings are built by optimizing both sets of representations

The meaning of 'meaning'

- word2vec: represents conditional probability as

$$p(c|w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

context word vector focus word vector

$$\frac{e^{\vec{c} \cdot \vec{v}}}{\sum_{\vec{u} \in C} e^{\vec{u} \cdot \vec{v}}}$$

Maximize if cooccurring, minimize otherwise

- GloVe: objective

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

co-occurrence count of c and v

context word vector focus word vector

$$(\vec{c} \cdot \vec{v} + \cdots - \log X_{cv})^2$$

Minimize

The meaning of 'meaning'

- In both cases, we maximize *v*.*c* for all words *c* occurring in *v*'s context.
- If *v* and *v'* both appear near the same words c_1, c_2, ... c_n
- <=> We're optimizing distances to similar sets of *c*s
- <=> *v* and *v'* are optimized with similar constraints
- <=> *v* and *v'* are similar (in space)

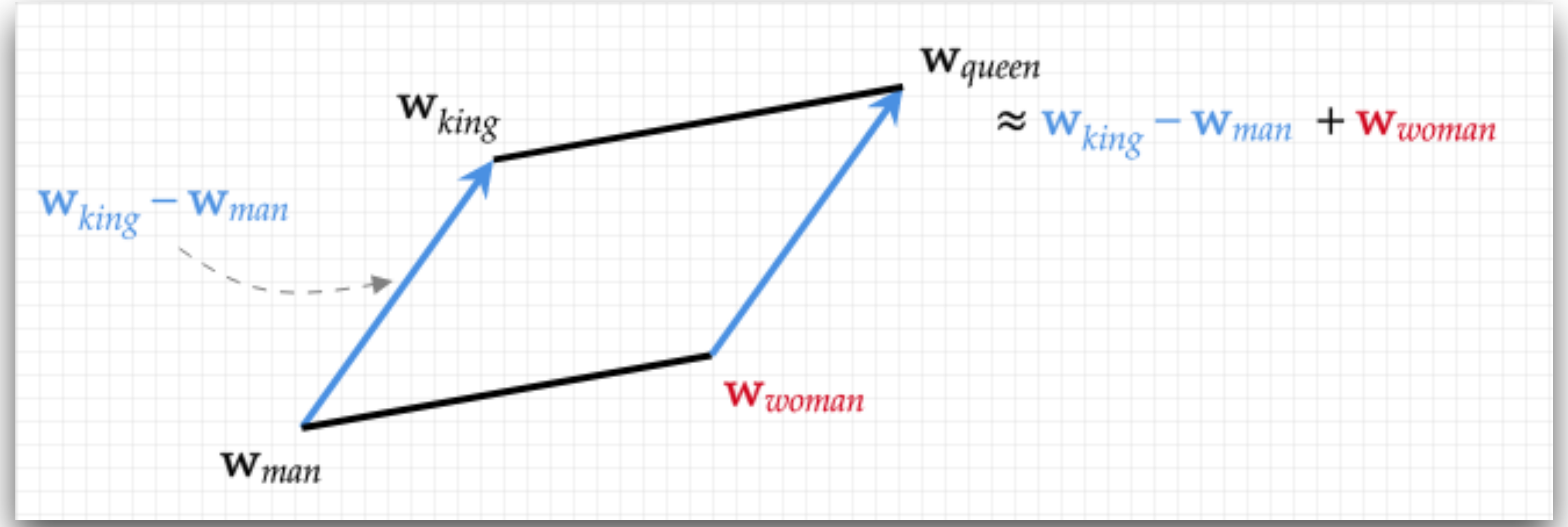
- So, words appearing in similar contexts have similar vectors

The meaning of ‘meaning’

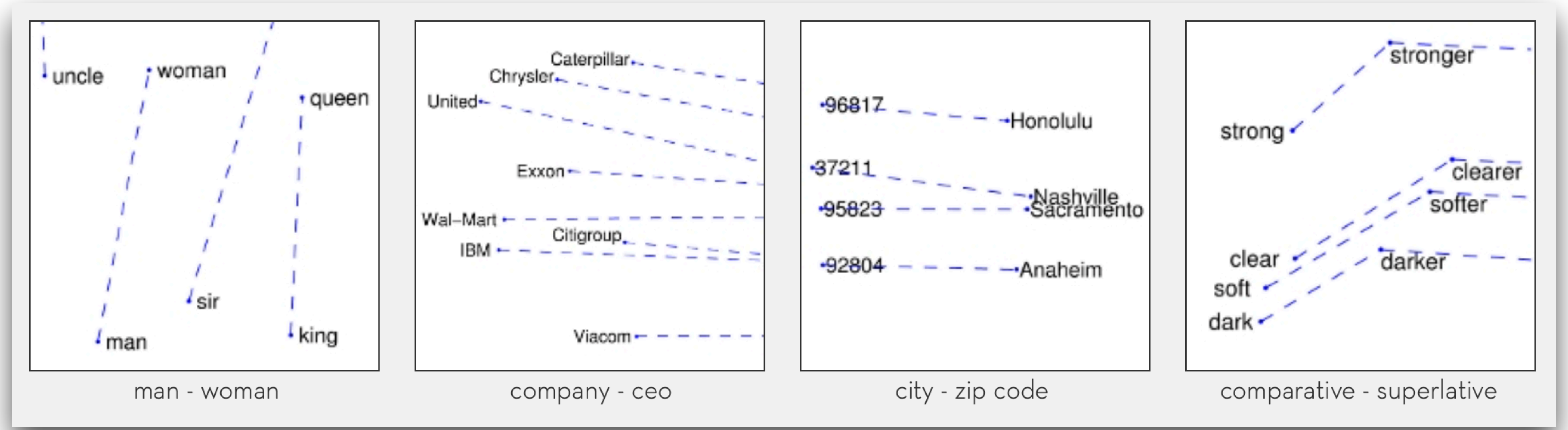
- Distributional hypothesis: *Words that occur in similar contexts tend to have similar meanings*
- Similar contexts <=> similar meanings ; similar contexts <=> similar vectors
- So, words with similar meanings have similar vectors.

Analogies

Analogies



Word embedding analogies. Source



Linear substrcture in GloVe. Source

Analogies

- 19,544 questions, of the format:
  - *Athens* : *Greece* :: *Madrid* : *Spain*
  - *policeman* : *policewoman* :: *groom* : *bride*
  - *flying* : *flew* :: *swimming* : *swam*

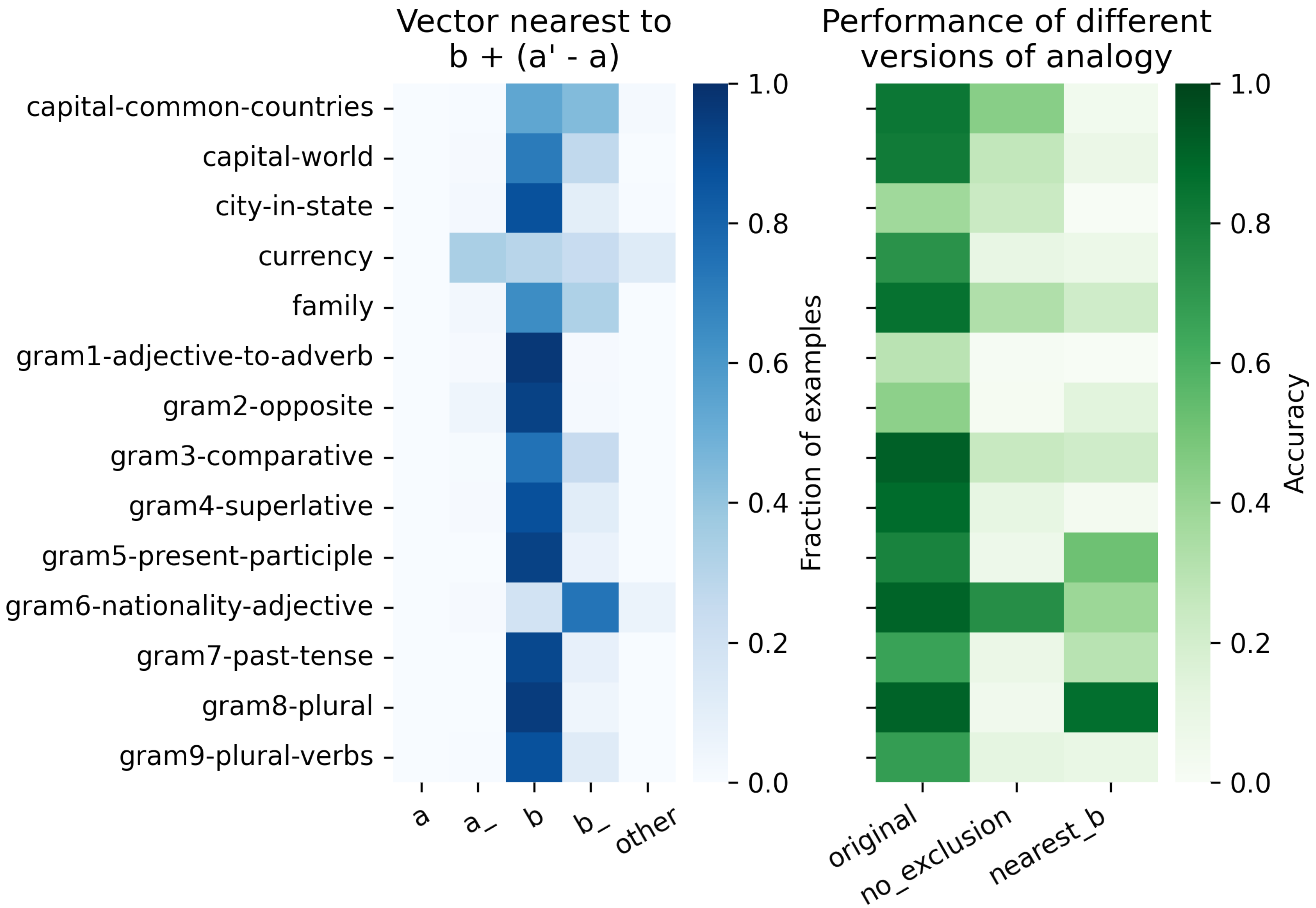
    a a’ b b’

- Predict the word *b’* by calculating $x = b + (a' - a)$, and finding its nearest neighbor
- *(a’ - a)* is supposed to encode a country-capital/masculine-feminine/past tense-present tense relation

Analogies

- *flying* and *flew* occur in a lot of similar contexts, and are close together in vector space
- *(a' - a)* in the formula $x = b + (a' - a)$ is often quite small
- Originally, $\{b, a', a\}$ were excluded as candidates
- But b is often the nearest neighbor of $b + (a' - a)$

Effects of proximity



Reproduction of results shown in Linzen 2016

Takeaway
Performance on the analogy task can be explained by other phenomena besides the existence of a constant offset

Analogies

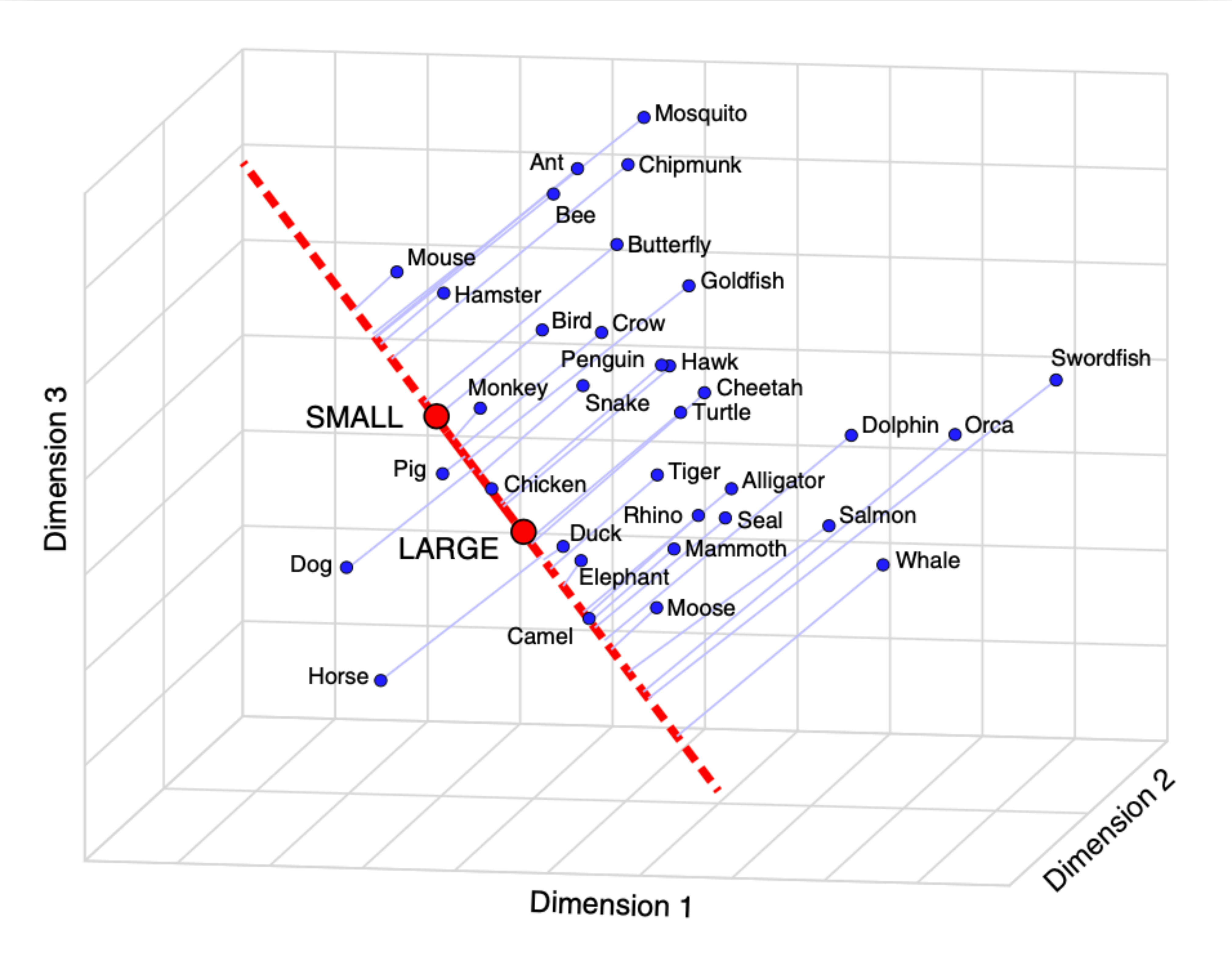
- The Google analogy dataset measured performance on 9 analogical relations
- BATS (Rogers et al. 2017) introduced 40 relations that are substantially more difficult than the original analogies
  - animal - typical sound (*fly - buzz*)
  - thing - color (*emerald - green*)
  - meronym - holonym (*star - galaxy*)
- Ethayarajh et al. 2019 was able to predict which sets of word pairs would be successful as analogies
  - Based on their co-occurrence and geometric properties of the space

Takeaway
Some analogy tasks are easier than others.
i. e. Some relations are better suited to vector analogies than others

Analogies

And semantic projection

- How is semantic projection affected by these findings?
- A semantic dimension is built based on antonyms like {*small, little, tiny*}, {*big, large, huge*}
- A word's position along this dimension corresponds to its qualities
- Don't need analogies to hold perfectly
- Can use a litmus test (Colab notebook)



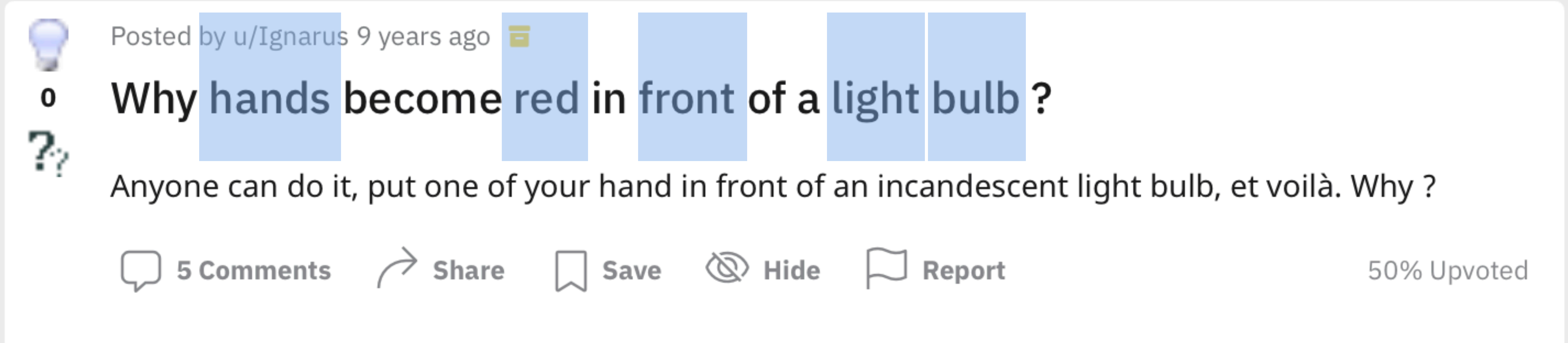
Semantic projection: Grand et al. 2022

Polysemy

And other shenanigans in lexical semantics

Polysemy

Prevalence



*… **bulb** of a syringe …*

*… tulip **bulb** …*

*… hired **hand** …*

*… try their **hand** …*

*… give me a **hand** …*

*… 20 **hands** tall …*

*… the minute **hand** …*

*… a cold **front** …*

*… the Western **front** …*

*… national liberation **front** …*

*… **front** man …*

*… a different **light** …*

*… **light** as a feather …*

*… do you have a **light** …*

*… **light** a fire …*

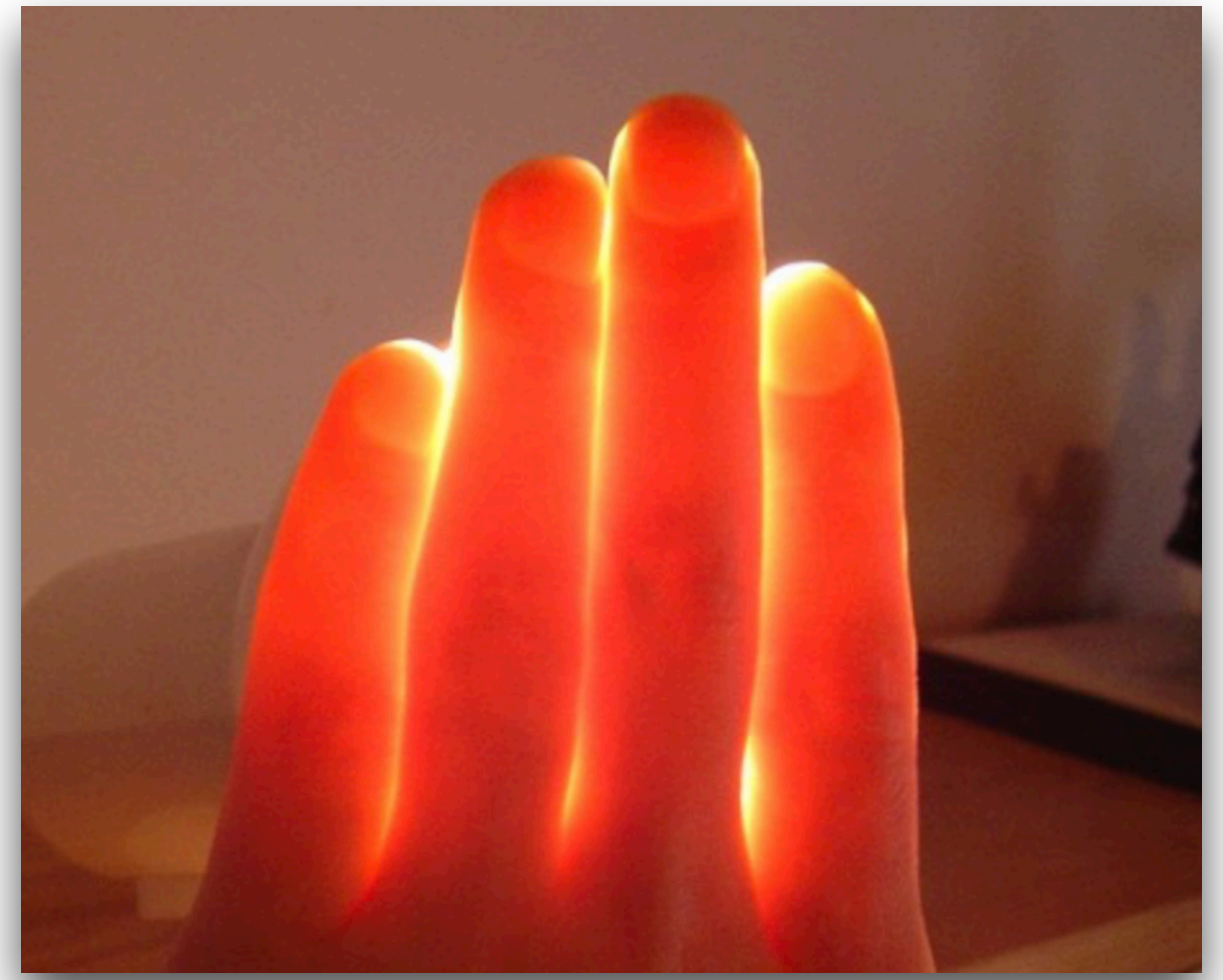
*… **light** load …*

*… a **light** diet …*

*… **red** state …*

*… **Red** Scare …*

*… in the **red** …*

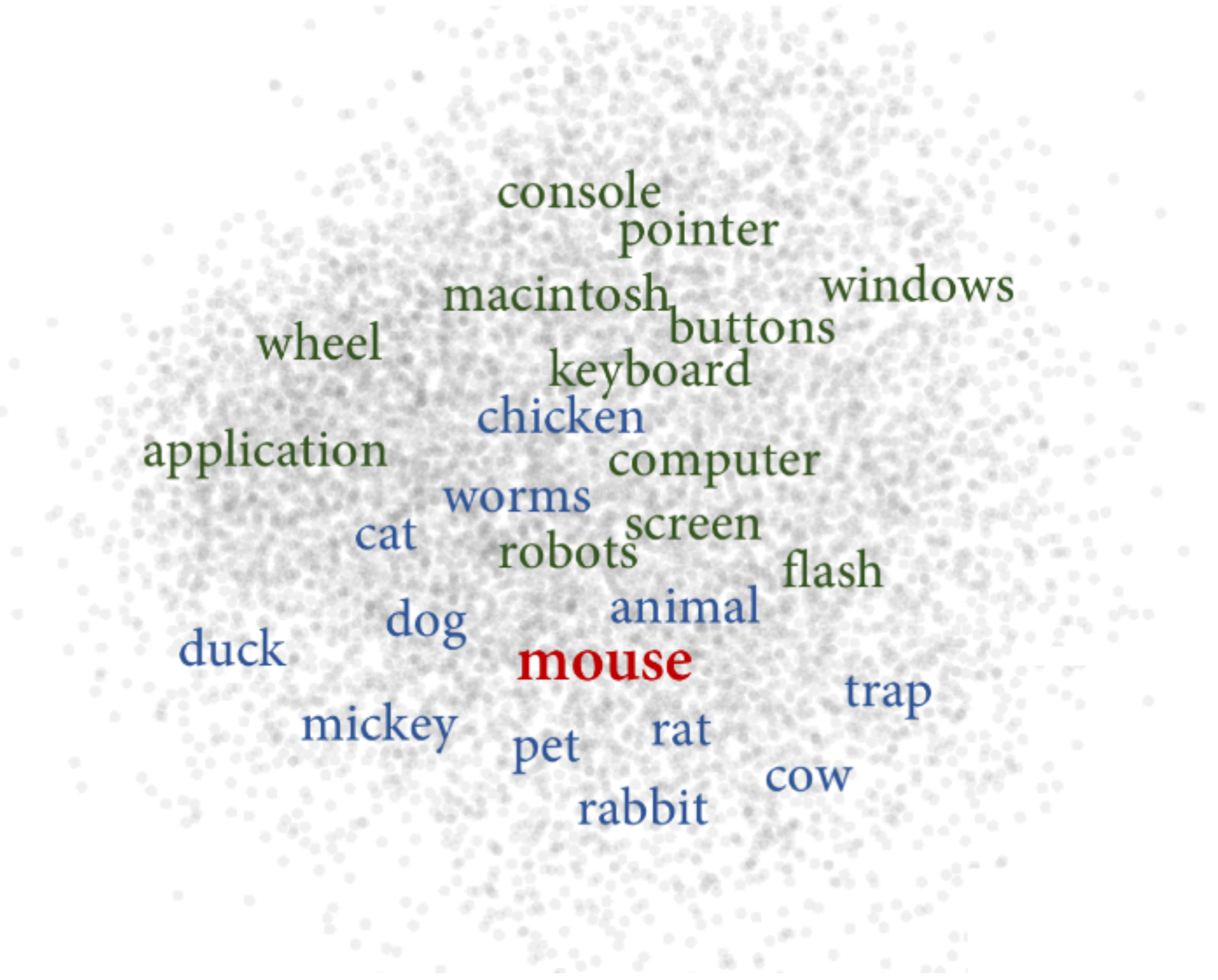


Senses from WordNet

Polysemy

Effect on word embedding models

- A word v (e.g. *mouse*) might have multiple sets of less-related context vectors:
  - c_1, c_2, c_3, (e.g. *console, printer, keyboard*)
  - c'_1, c'_2, c'_3 (e.g. *tail, pet, cow*)
- Two sets of constraints, pulling v in two different directions
- Consequently, pulling words like *keyboard* and *cow* together



Camacho-Collados et al. 2018

context word vector   focus word vector

$$\frac{e^{\vec{c}\cdot\vec{v}}}{\sum_{\vec{u}\in C} e^{\vec{u}\cdot\vec{v}}}$$

Polysemy

How this affects you

- If you are using a very particular corpus, multiple senses not present: so no worries!
- If you are interested in a primary sense, most of the contexts are relevant: so no worries!
- One possibility: you happen to care about a particular word (sense) which is overshadowed by a more prevalent sense
- e.g. you are looking for contexts of *rich*/*poor* in Wikipedia, and you get *vibrant*, *tangy*, *lively*, *fertile* as nearest neighbors

Polysemy

Mitigations

- Multi-prototype embeddings
- Dictionary-based methods
- Since 2018: BERT-based approaches
  - Contextual embeddings like BERT account for different senses by design

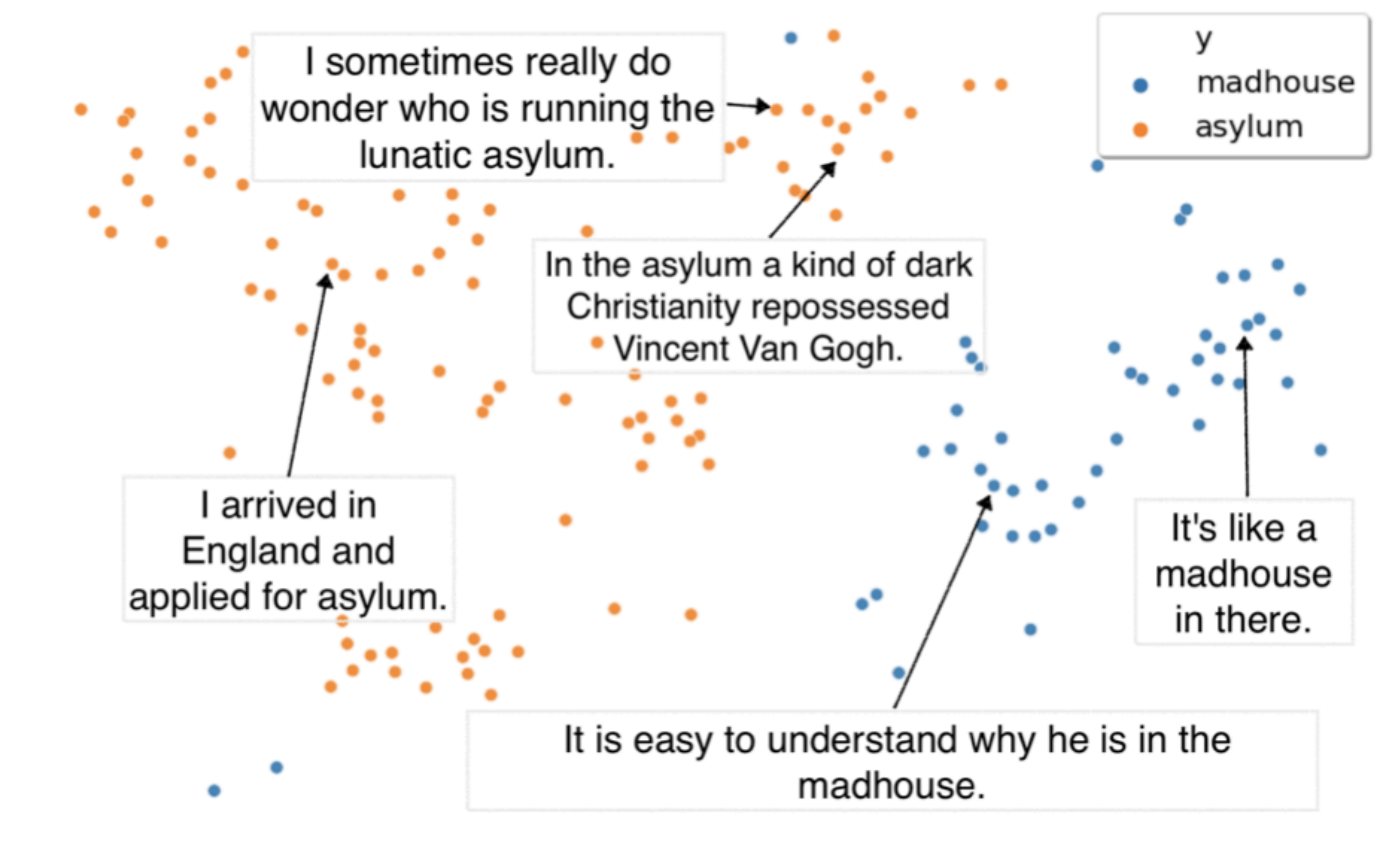


Figure 1: 2D t-SNE visualization of layer 8 vectors for tokens of *asylum* and *madhouse* sampled from the BNC.

Multi-prototype BERT embeddings: Chronis and Erk 2020

Takeaway

Polysemy is prevalent, but it may or may not affect you

If you find that it does, contextual embeddings are the way to go

Antonyms

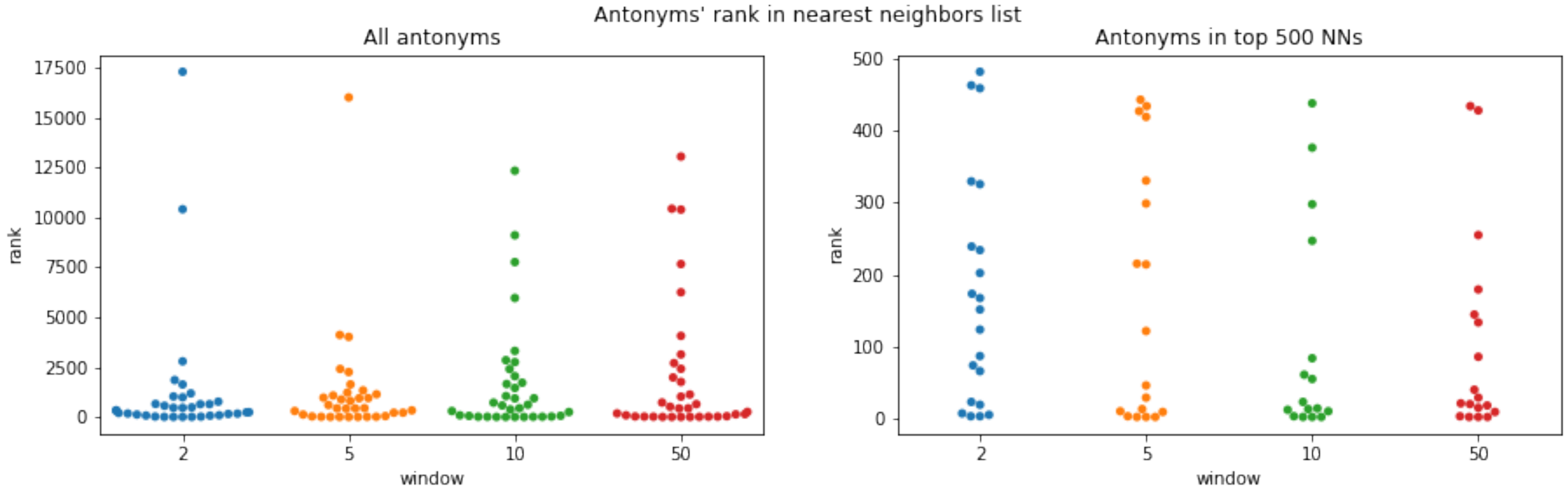
What even are they?

- Perfect antonyms aren’t easy to come by
- *Republican* is the opposite of *democrat,* but it is substantially more similar to *democrat* than most things
  - examples: *stapler, mitosis, carelessly, luminous*

Antonyms

Takeaway
Antonyms are very similar, and you should take their counterintuitive behavior into account if relying on them

- Antonyms are often nearest neighbors!
- To really get a sense of embedding space, feel free to play the game Semantle
  - It will make you cry
- It was found (Levy and Goldberg 2014) windows of size 5 or larger contain topical content, while smaller windows contain information about the focus word itself.



Antonyms

A definition

- They differ along one dimension of meaning (perhaps occupying opposing poles), but are identical in all other dimensions. (Cruse 1986)
- Is ‘progressive’ the opposite of ‘redneck’?
  - Dissimilar with respect to political orientation
  - Also dissimilar with respect to connotations

Stability of word embeddings

Results from Antoniak and Mimno 2018

Stability

Training is stochastic!

- The final values of the vectors depend on
  - The precise combination of documents used to train your model
  - The order in which documents appear

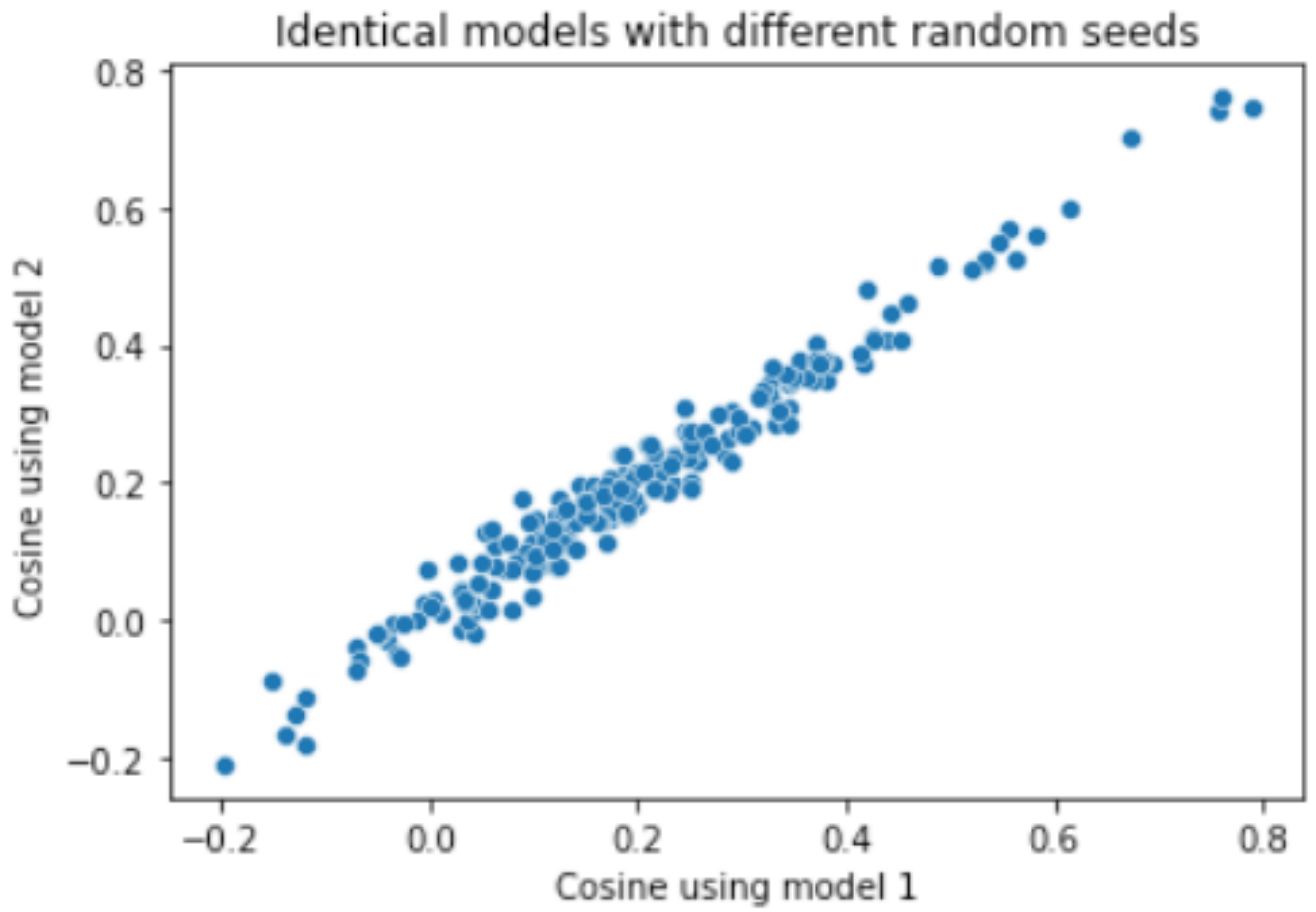
  - Preprocessing
  - The random seed used for training

  - The order in which threads are scheduled during training

Stability

Due to change in random seed

- Two models trained with the exact same corpus, but different random seeds
- Cosine similarities between pairs of 'topic' words in /r/AskScience dataset:
  - *bacteria, plant, species, brain, muscle, sleep, human, galaxy, space, planet, universe, electricity, light, magnetic, field, power, calorie, chemical, temperature, pressure*



Takeaway
The exact cosine value does not have any intrinsic meaning — it only has meaning with respect to other cosines in the same space

Stability

We can't rely on cosine values?

- If you have two training settings, A and B, what can you compare with cosine?
  - *cosine*(cat_A, dog_A) and *cosine*(cat_A, tree_A) ✅
  - *cosine*(cat_A, dog_A) and *cosine*(tree_A, car_A) ❓
  - *cosine*(cat_A, dog_A) and *cosine*(cat_B, dog_B) ❗
  - *cosine*(cat_A, dog_B) omg please no ❌

Stability

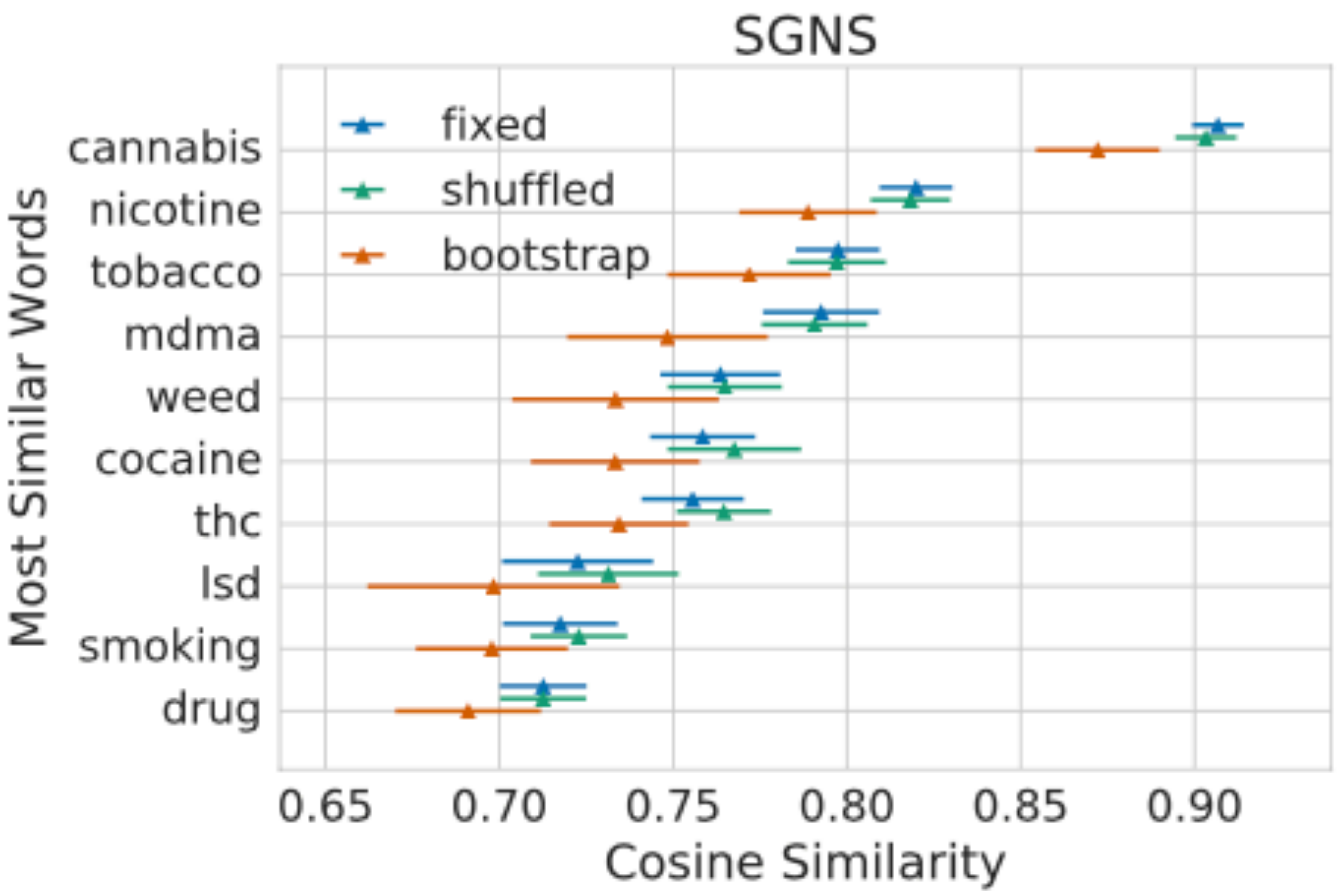
Mitigation?

- One way to mitigate instability is to run multiple runs and report averages
- Word embeddings are a representation of the author's mental model
  - Only a *sample* of the author's mental model
  - This can introduce error — but how much?

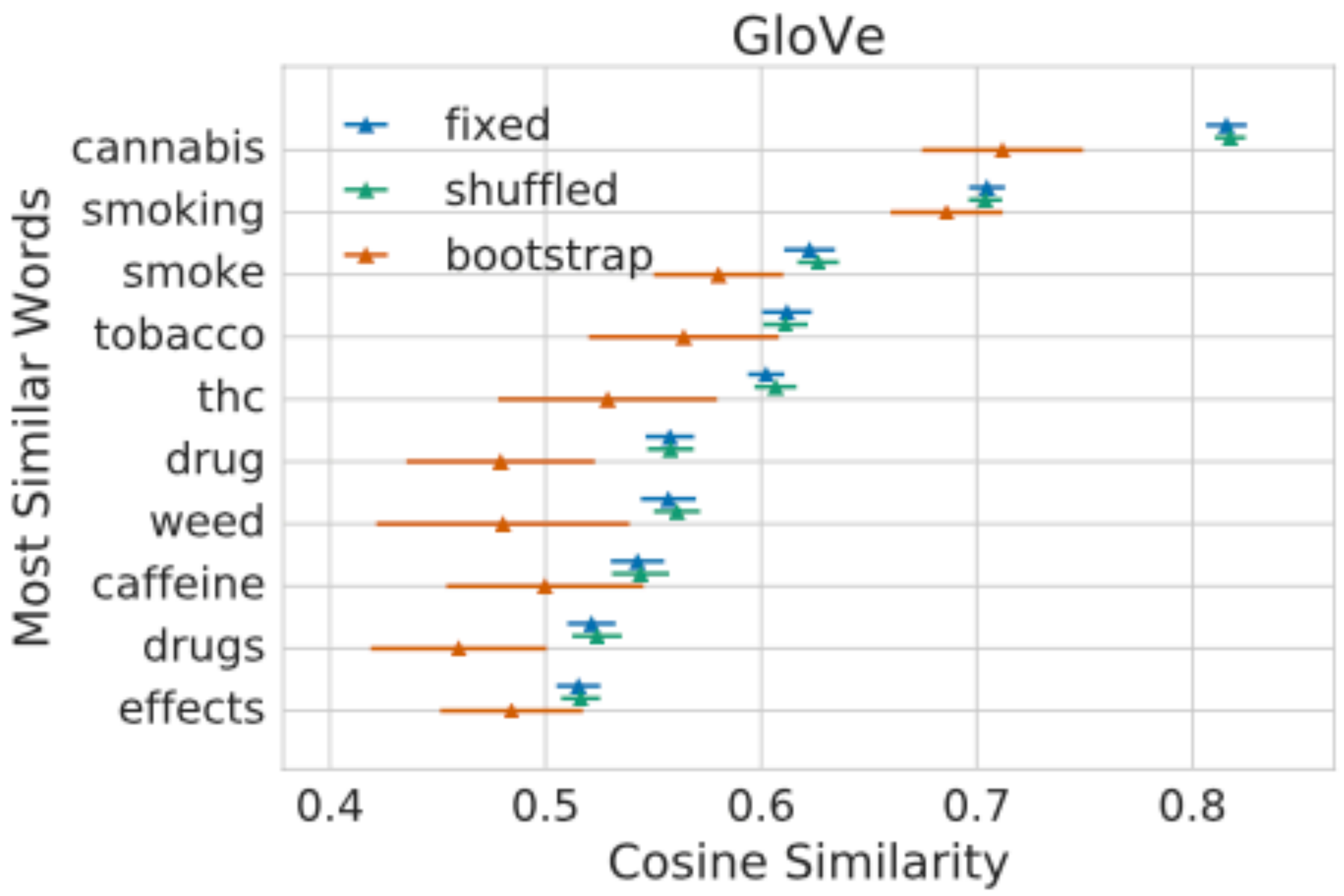
Stability

Document selection

- Results of 50 models trained on /r/AskScience data, showing cosines of *marijuana* and its most similar words



Source: Table 4, Antoniak and Mimno 2018



Source: Table 4, Antoniak and Mimno 2018

Stability

Mitigation!

- Run multiple instantiations of the same model, and report averages rather than raw values
  - Draw documents with replacement (Colab notebook)
- You can also report ranks of nearest neighbors, rather than their similarity scores

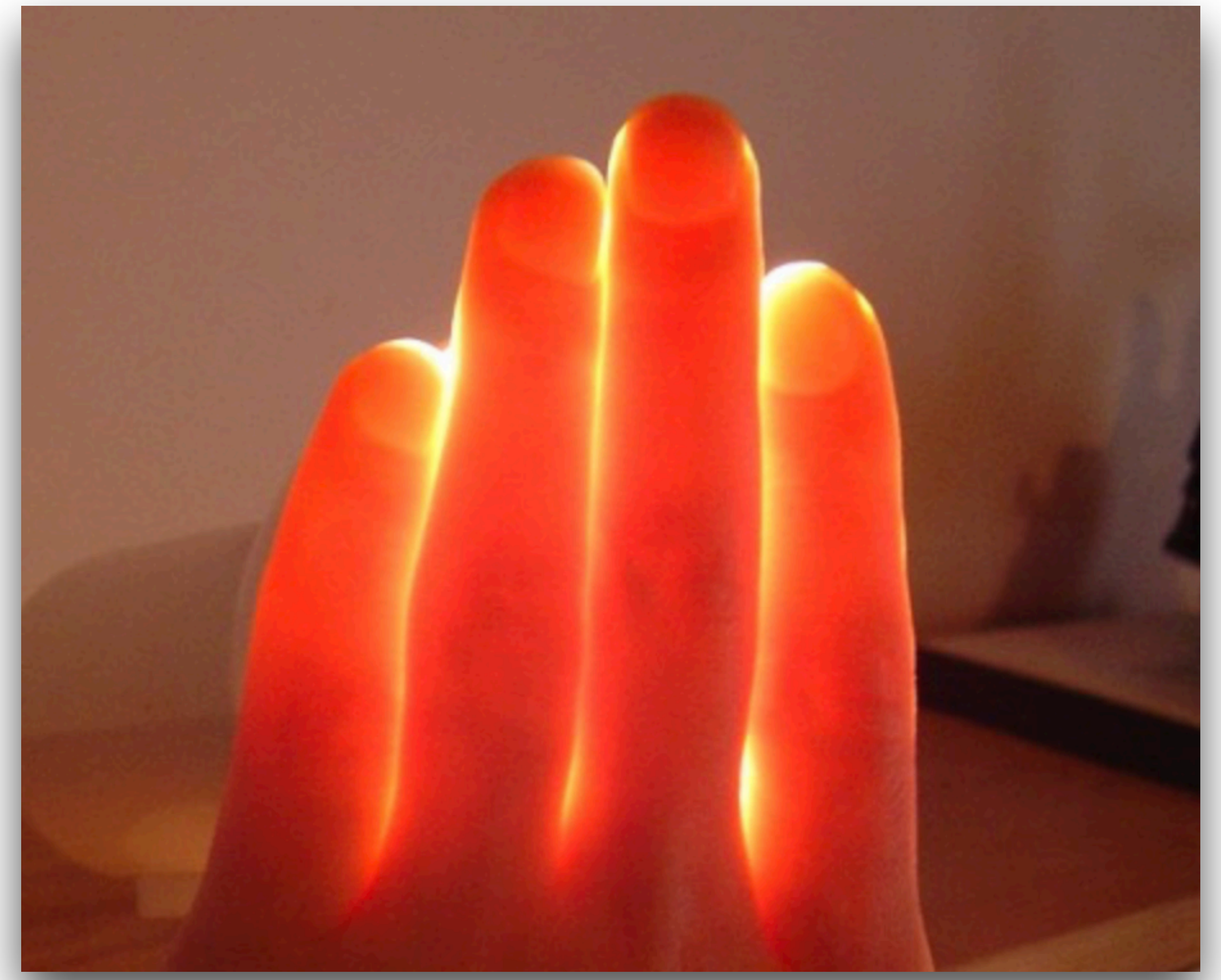
Instability

Caution

- Why draw documents with replacement instead of drawing sentences with replacement?
- A word's contexts in a document are likely to be similar to each other
- 'One sense per discourse' (Gale et al. 1992)

... *light* ***bulb*** ...

... *tulip* ***bulb*** ...



Parting thoughts

- For normalized vectors, the ranking of nearest neighbors using cosine or Euclidean distance is the same (Manning and Schütze 1999, Sec. 8.5)
- You can use a dot product to calculate a lot of cosine distances at the same time
  - Just make sure you remember normalize everything first!
- Magnitude is hard to interpret

Parting thoughts

- The flip side of corpora reflecting authors' mental models
- Word embeddings and other machine learning models reproduce hegemonic viewpoints (who produces the texts we are studying?)
- Machine learning models can reify these viewpoints
- References
  - Bolukbasi et al. 2016
  - Caliskan et al. 2017

Thank you!

http://www.github.com/nnkennard/embeddings-tutorial/