Interpreting Word Vectors

Goals

1. Appreciating the scope of application of word embeddings

Caveat

The choice regarding the linguistic context's length affects word2vec similarities

Window size = 2

The nearest words to **Hogwarts** are:

- Sunnydale
- Evernight

Window size = 5

The nearest words to **Hogwarts** are:

- Dumbledore
- Malfoy
- Half-blood

Word embeddings' scope of application

Traversing vectors to find related words

- The Euclidean distance (or cosine similarity) between two-word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words.
- Sometimes, the nearest neighbours, according to this metric, reveal rare but relevant words that lie outside an average human's vocabulary.

For example, here are the closest words to the target word frog:

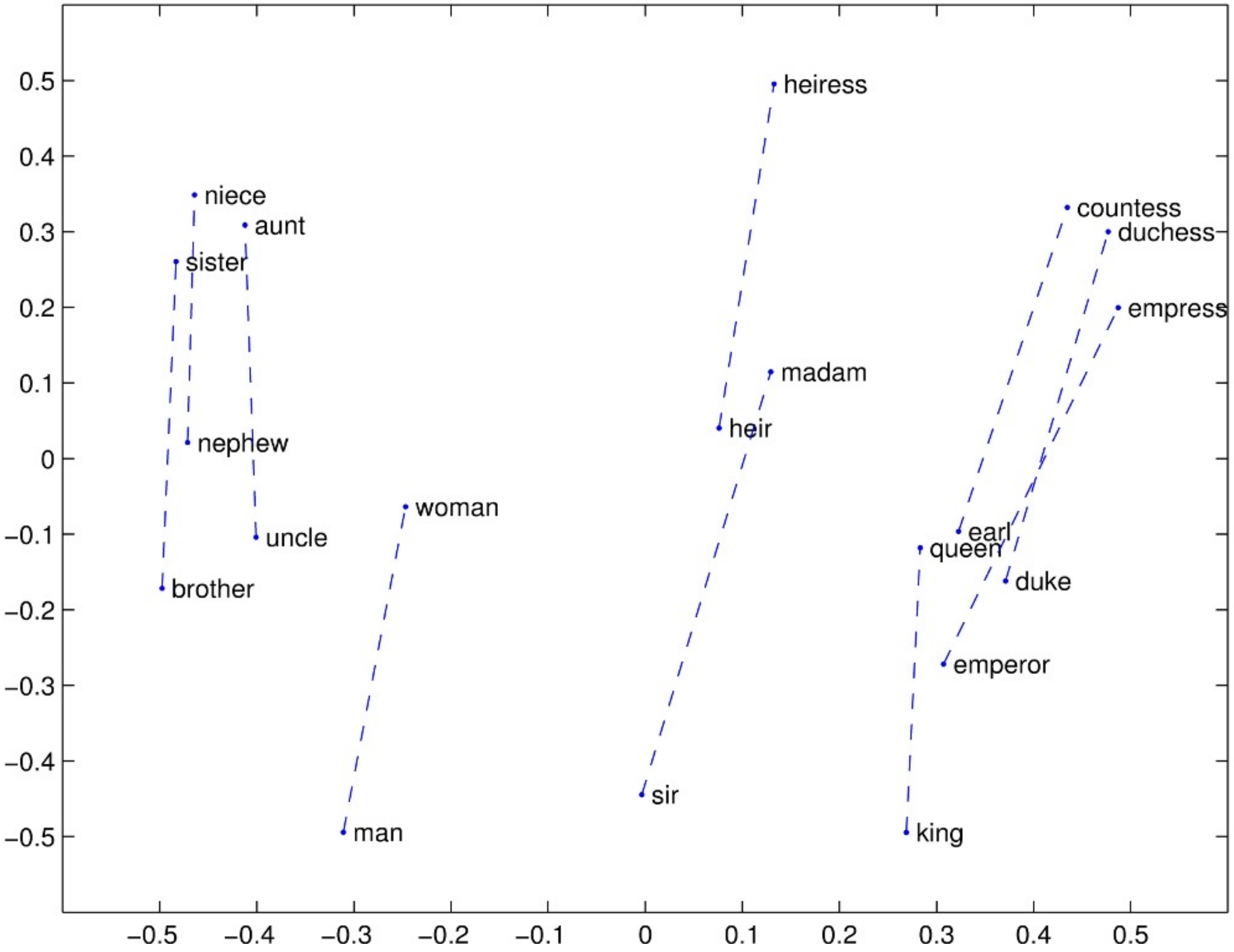
- frog
- frogs
- toad
- litoria
- leptodactylidae
- rana
- lizard
- eleutherodactylus

Embedding capture relational meanings

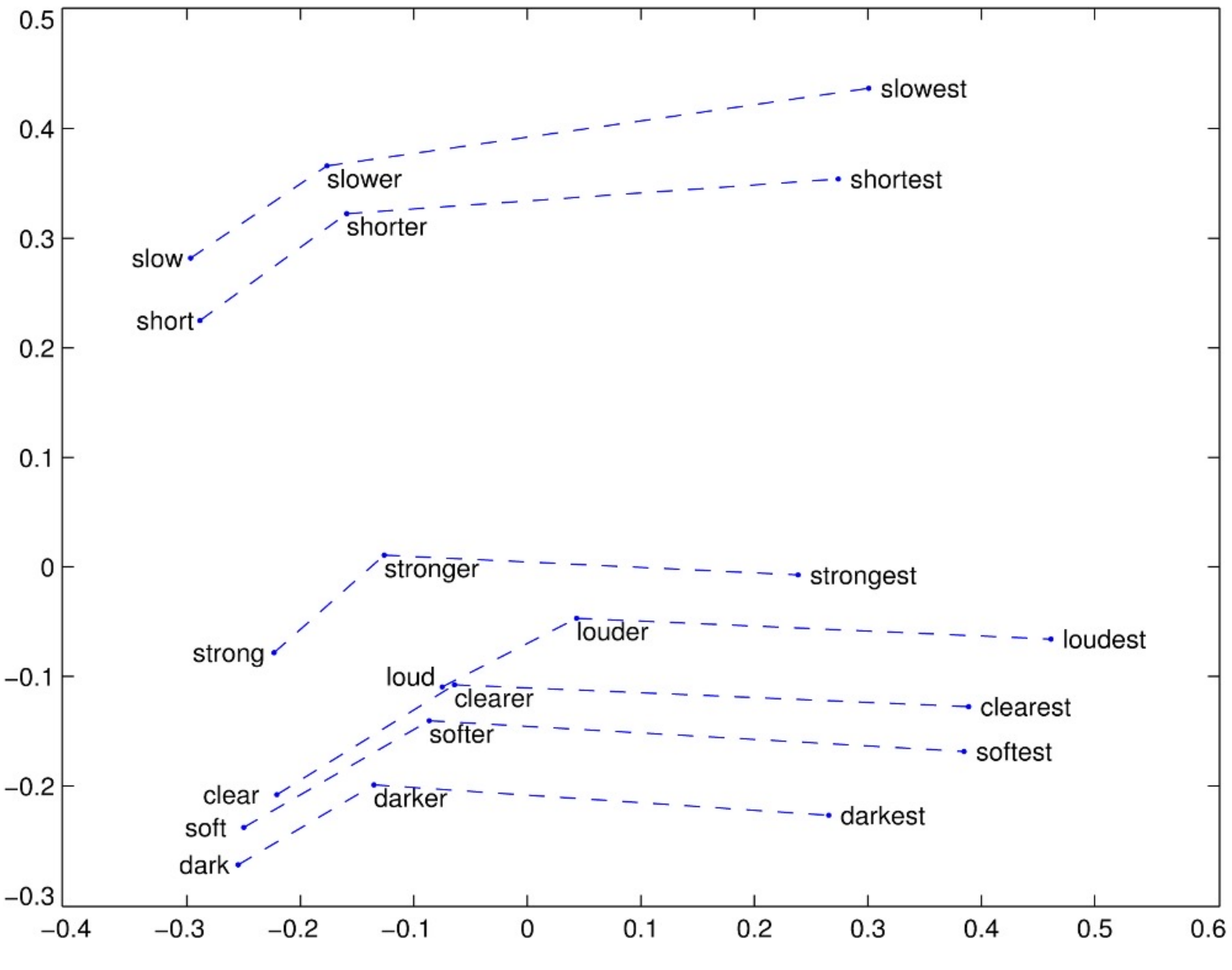
```
vector(‘king’) - vector(‘man’) + vector(‘woman’) = vector(‘queen’)

vector(‘Paris’) - vector(‘France’) + vector(‘Italy’) = vector(‘Rome’)
```

Visual inspection of embeddings (1/2)



Visual inspection of embeddings (2/2)

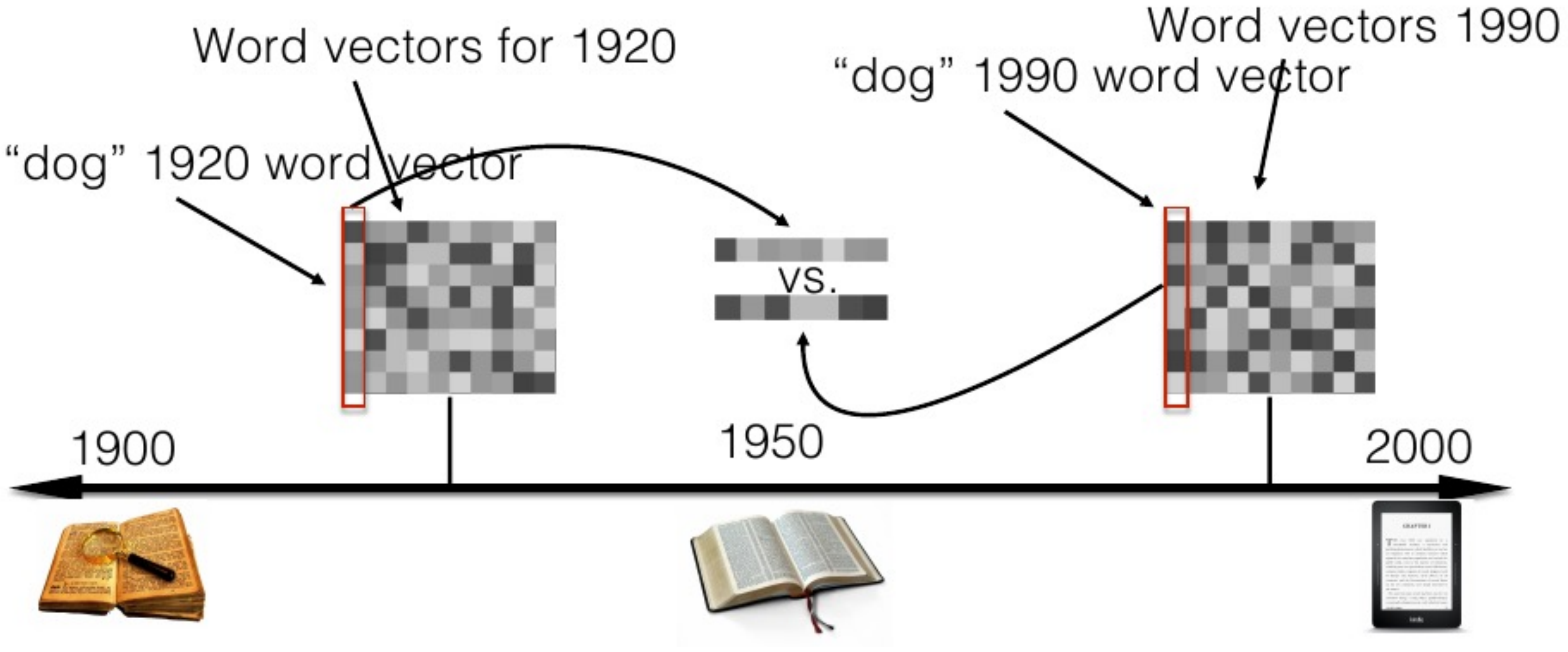


Word embeddings' application examples

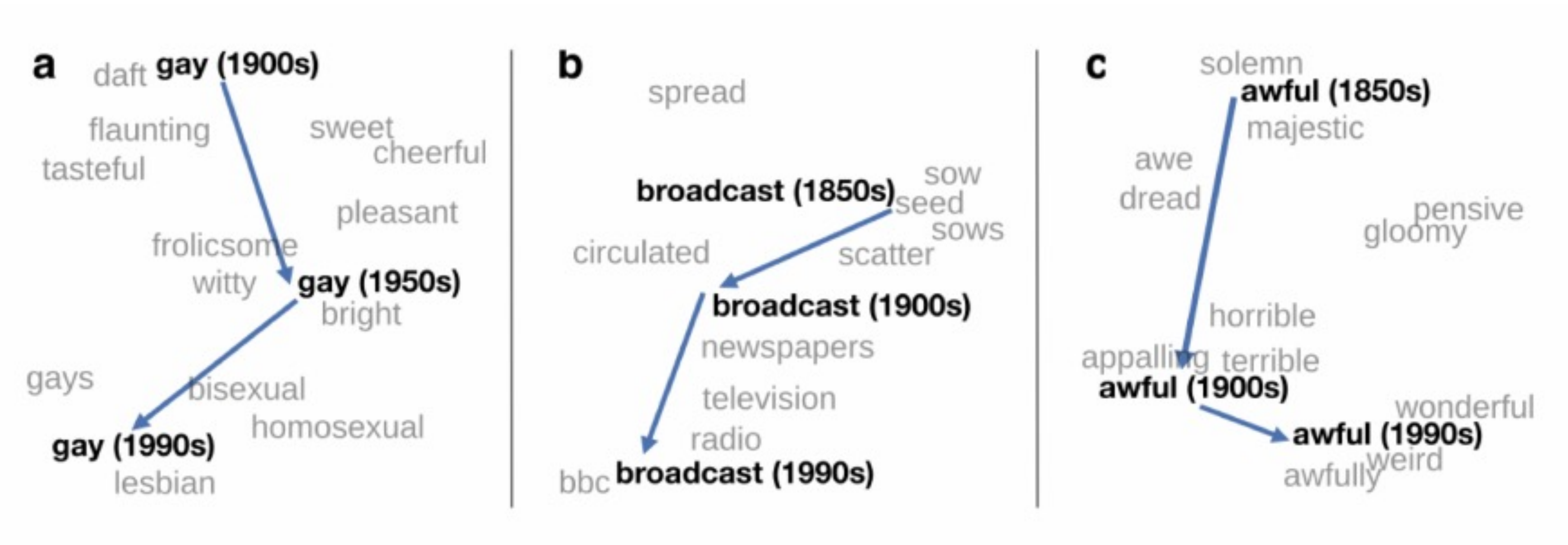
Embeddings: Fields of application

- Business and economic analysis (a few examples)
- Cultural analysis (many examples)

Studying changes in meanings with Google Books data (1/2)



Studying changes in meanings with Google Books data (2/2)

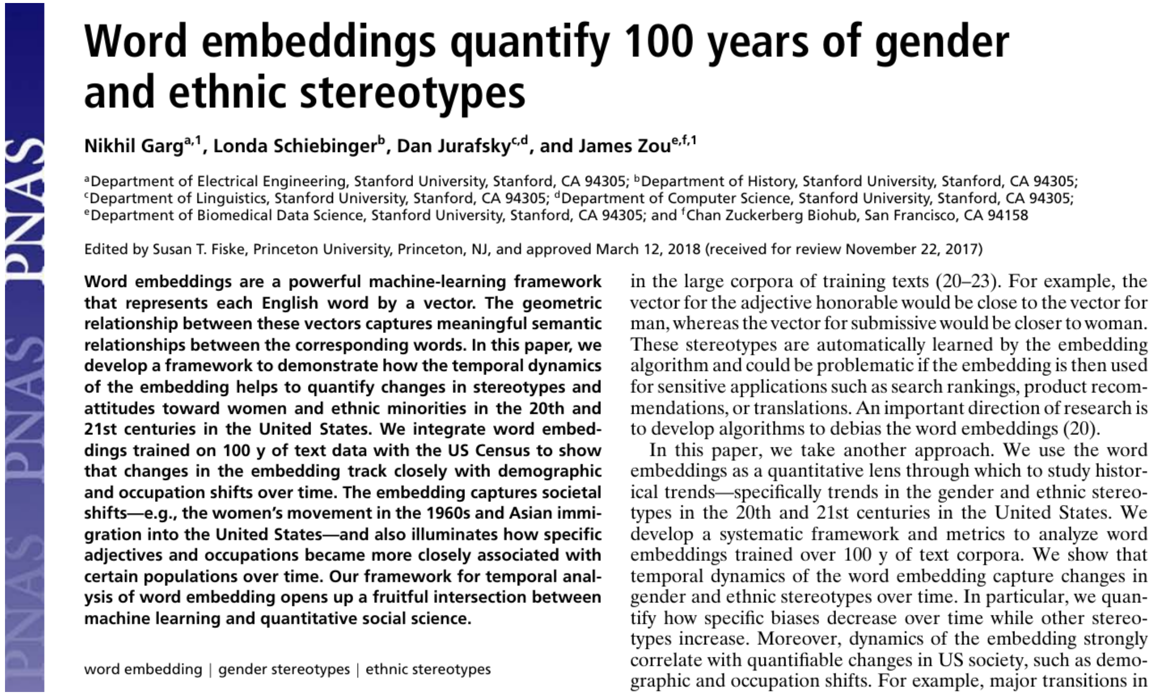


~30 million books, 1850-1990, Google Books data

Using embeddings as a historical tool to study bias

The paper in a nutshell:

- Embeddings for competence adjectives are biased toward men.
- Smart, wise, brilliant, intelligent, resourceful, thoughtful, logical, etc.
- This bias is slowly decreasing over time



Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg[a,1], Londa Schiebinger[b], Dan Jurafsky[c,d], and James Zou[e,f,1]

[a]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; [b]Department of History, Stanford University, Stanford, CA 94305; [c]Department of Linguistics, Stanford University, Stanford, CA 94305; [d]Department of Computer Science, Stanford University, Stanford, CA 94305; [e]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and [f]Chan Zuckerberg Biohub, San Francisco, CA 94158

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 12, 2018 (received for review November 22, 2017)

Word embeddings are a powerful machine-learning framework that represents each English word by a vector. The geometric relationship between these vectors captures meaningful semantic relationships between the corresponding words. In this paper, we develop a framework to demonstrate how the temporal dynamics of the embedding helps to quantify changes in stereotypes and attitudes toward women and ethnic minorities in the 20th and 21st centuries in the United States. We integrate word embeddings trained on 100 y of text data with the US Census to show that changes in the embedding track closely with demographic and occupation shifts over time. The embedding captures societal shifts—e.g., the women's movement in the 1960s and Asian immigration into the United States—and also illuminates how specific adjectives and occupations became more closely associated with certain populations over time. Our framework for temporal analysis of word embedding opens up a fruitful intersection between machine learning and quantitative social science.

word embedding | gender stereotypes | ethnic stereotypes

in the large corpora of training texts (20–23). For example, the vector for the adjective honorable would be close to the vector for man, whereas the vector for submissive would be closer to woman. These stereotypes are automatically learned by the embedding algorithm and could be problematic if the embedding is then used for sensitive applications such as search rankings, product recommendations, or translations. An important direction of research is to develop algorithms to debias the word embeddings (20).

In this paper, we take another approach. We use the word embeddings as a quantitative lens through which to study historical trends—specifically trends in the gender and ethnic stereotypes in the 20th and 21st centuries in the United States. We develop a systematic framework and metrics to analyze word embeddings trained over 100 y of text corpora. We show that temporal dynamics of the word embedding capture changes in gender and ethnic stereotypes over time. In particular, we quantify how specific biases decrease over time while other stereotypes increase. Moreover, dynamics of the embedding strongly correlate with quantifiable changes in US society, such as demographic and occupation shifts. For example, major transitions in

Wrap-up

Main points

- Navigating word vectors allows analysts to discover latent semantic associations — i.e., connections that are difficult to articulate in one’s mind.
- Word vectors also reveal the cultural stereotypes that affect the behaviour of individuals, groups and more giant economic and social formations such as markets.