

PROFESSOR BENJAMIN NOBLE (UCSD)

---

# WORDS IN SPACE

## It's Trump's platform, with a dose of DeSantis



Sophia Cai

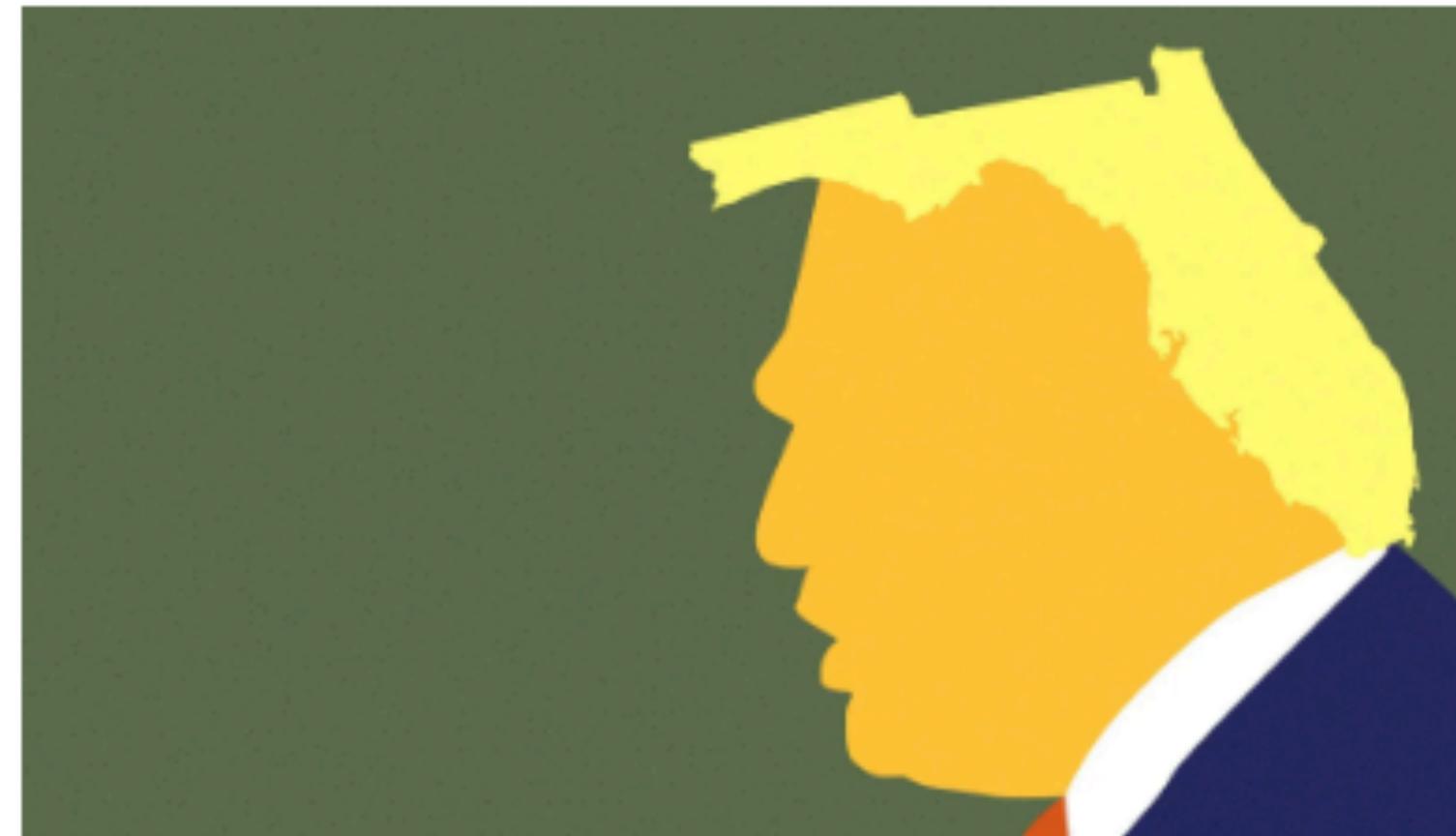


Illustration: Shoshana Gordon/Axios

[Donald Trump](#) used to trash Ron DeSantis. Now, he copies him.

**Trump gradually** has taken a few pages from DeSantis' playbook on issues that resonate with the far right.

- Last year he revealed an education plan with a proposal for a "Parental Bill of Rights." It calls for more federal influence in how public schools are run, and for cutting funds to schools that teach "gender ideology."
- Months earlier, DeSantis had signed a "Parental Rights in Education" bill with provisions that banned talk about gender and sexuality in schools.
- Today, "defund schools" is one of Trump's go-to phrases when he talks about his plans for a second term, [an Axios analysis](#) found.

via Axios

### SOME PROPERTIES WE MIGHT WANT FROM A MEASURE OF SIMILARITY

- ▶ Consider two documents, **a** and **b**...
- ▶ Perfect/maximum similarity should occur when comparing **a** to **a**.
- ▶ Symmetry: document **a** is always as close to **b** as **b** is to **a**.
- ▶ If **a** and **b** share no words, similarity should be at a minimum.
- ▶ As **a** and **b** share more words, then similarity should increase.

### ENTER: THE DOT PRODUCT

- ▶ Recall: cat, dog, fish.
- ▶ The document “cat dog” can be represented as  $(1, 1, 0)$ .
- ▶ The document “dog fish” can be represented as  $(0, 1, 1)$ .
- ▶ The dot product of two vectors:  $a \cdot b = a^T b = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$
- ▶  $\text{sim}(d_1, d_2) = 1 \cdot 0 + 1 \cdot 1 + 0 \cdot 1 = 1$

	“by”	“man”	“upon”	Vector	Word count
Hamilton	859	102	374	(859,102,374)	1335
Jay	82	0	1	(82,0,1)	83
Madison	474	17	7	(474,17,7)	498
???	15	2	0	(15,2,0)	17

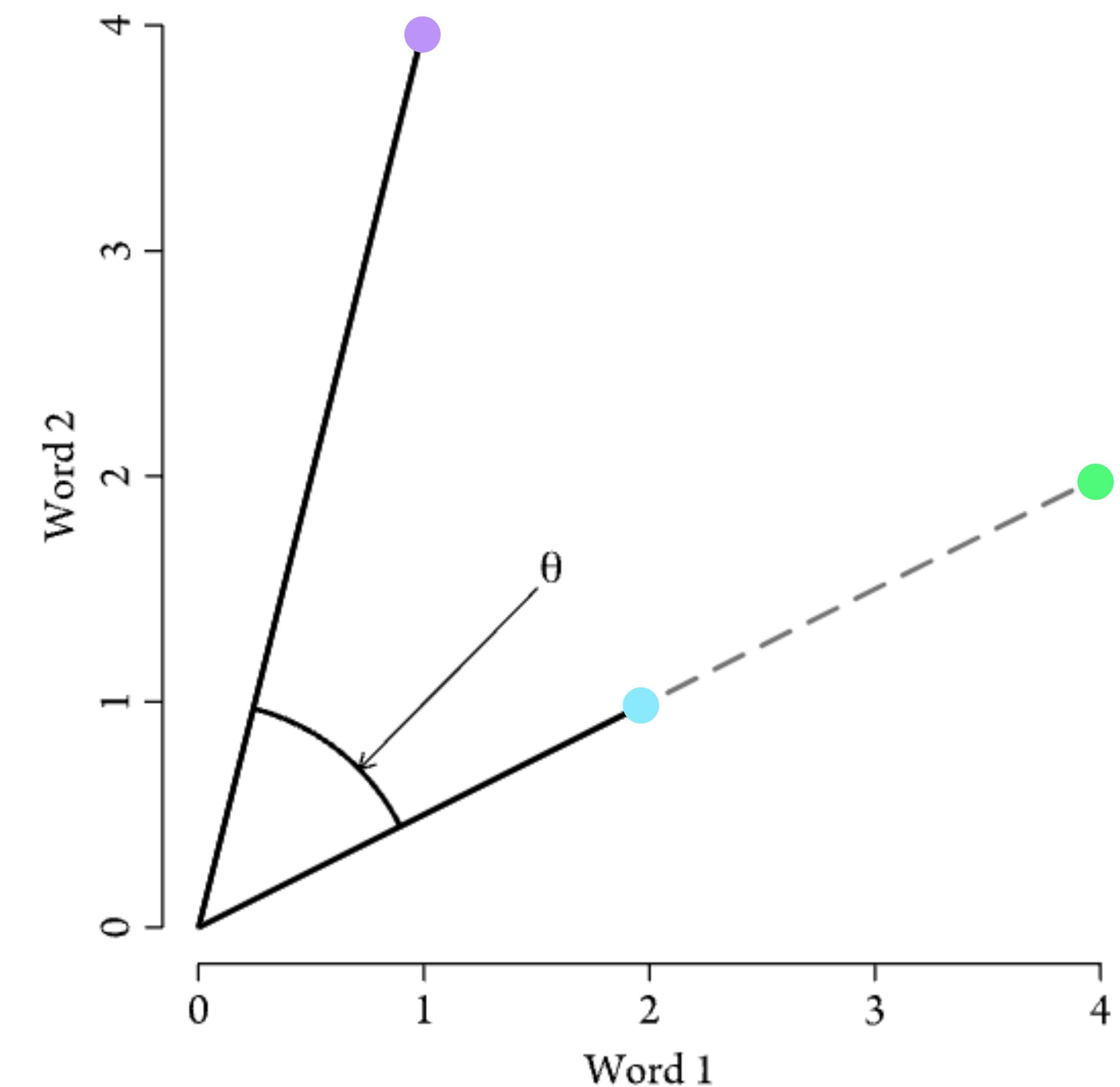
WORDS IN SPACE

### APPLYING THE DOT PRODUCT

- ▶  $W_H \cdot W_? = (859, 102, 374) \cdot (15, 2, 0) = 859 \cdot 15 + 102 \cdot 2 + 374 \cdot 0 = 13089$
- ▶  $W_M \cdot W_? = (474, 17, 7) \cdot (15, 2, 0) = 7144$
- ▶ Normalizing vectors:
  - ▶  $\|W_H\| = \sqrt{W_H \cdot W_H} = \sqrt{859^2 + 102^2 + 374^2} = 942.42$
  - ▶  $\|W_M\| = 474.36$

## COSINE SIMILARITY

- ▶ cosine similarity( $a, b$ ) =  $\cos \theta = \frac{a}{\|a\|} \cdot \frac{b}{\|b\|}$
- ▶  $\text{cs}(W_H, W_?) = \frac{W_H \cdot W_?}{\|W_H\| \cdot \|W_?\|} =$
- ▶  $\frac{(859,102,374) \cdot (15,2,0)}{\|(859,102,374)\| \|(15,2,0)\|} = 0.918$
- ▶  $\text{cs}(W_M, W_?) = 0.995$



# WORDS IN SPACE

question commission  
her duties number  
nations officers vessels  
years secretary means  
powers without well mexico  
duty laws authority  
through amount during between also system  
within service these now do he  
time much congress against  
action state they which but one my part  
long increase own if for upon most  
thus many general may by in for had when being  
claims report act made war far each since  
purpose territory some other on to and  
while who we as all would trade  
force before no as from law new revenue  
condition great not this with such interest  
union every public this will citizens just  
nation only have has so last interests  
lands national states its a it been than same could  
commerce were their is be shall  
what peace government our united into first session  
however consideration two under should his your first  
out american year can them was or people must navy  
among both present those more department rights  
army attention subject necessary effect  
important treasury power constitution  
you large work business  
legislation relations  
make proper

# Early State of the Union Addresses

citizens  
rights give use meet home  
reform jobs where still children  
had provide know education  
power into every administration shall needs  
during national was there first act business  
much work government system both  
against congress from must those were before  
come two own by have american energy  
while law tax also nations defense  
care economic year an in  
life free which a so is and  
past program them our  
military world this  
next out if for  
effort state or that  
president billion make be  
trade united more  
take some years on we  
part under these last who  
better than their health  
high peace us with has  
action great than but not  
even many its people  
without policy help other should when  
service right security country what increase  
here legislation public over only progress  
important support upon up tonight  
international efforts let freedom  
problems hope

# Late State of the Union Addresses

### WEIGHTING

- ▶ Tf-idf (term frequency inverse document frequency) weighting:

$$W_{ij}^{tf-idf} = \underbrace{\frac{W_{ij}}{W_i}}_{\text{tf}} \times \overbrace{\log \left( \frac{N}{n_j} \right)}^{\text{idf}}$$

- ▶ The word “jobs” appears 612 times in the later SOTUs and 0 times in the early SOTUs.  
To determine tf-idf for “jobs” in the late period, we...

$$\text{▶ } W_{\text{jobs}}^{tf-idf} = \frac{612}{762087} \times \log \left( \frac{2}{1} \right) = 0.000803 \times 0.693 = 0.000557$$

# WORDS IN SPACE

electors  
1824  
decree  
1889  
1847  
exhibited  
depredations  
steamer  
1879  
7th  
specie  
1898  
slaves  
1885  
1869  
1833  
bullion  
1874 1875  
cession

1909 1859  
1868 annexion  
1860 emperor  
1858 mails  
1891  
exposition  
coinage  
fur  
1893  
1864  
1888  
1867  
1895  
1892  
liable  
1818  
steamer  
1884  
inferio  
frauds  
1891  
1893  
1897  
hon  
1890  
1890  
excite  
1900  
interoceanic  
ceded  
doubted  
postage  
fortifications  
whilst  
certained  
ratifications  
proofs  
d'affaires  
kindly  
prussia

# Early State of the Union Addresses

# Late State of the Union Addresses

### WHERE NAIVE WORD SIMILARITY FAILS...

- ▶ “The president spoke tonight” and “Biden delivered the State of the Union Address” have low similarity.



**Senator Laphonza Butler**  
@Senlaphonza

Extremists have launched frivolous legal battles to rob women of their right to safe, effective abortion pills. While this decision is an important step forward, we cannot stop fighting until these attacks on our reproductive freedom end.

[www.nbcnews.com/politics/supreme-court/supreme-court-rejects-bid-restrict-access-abortion-pill-rcna151308](http://www.nbcnews.com/politics/supreme-court/supreme-court-rejects-bid-restrict-access-abortion-pill-rcna151308)

𝕏 Posted on X • 11:38 AM



**Sen. Lisa Murkowski**  
@lisamurkowski

I was pleased to see that today the Supreme Court unanimously provided certainty for Alaskans who seek access to mifepristone, which has proven to be a safe and effective drug.

[apnews.com/article/supreme-court-abortion-mifepristone-fda-4073b9a7b1cbb1c3641025290c22be2a](http://apnews.com/article/supreme-court-abortion-mifepristone-fda-4073b9a7b1cbb1c3641025290c22be2a)

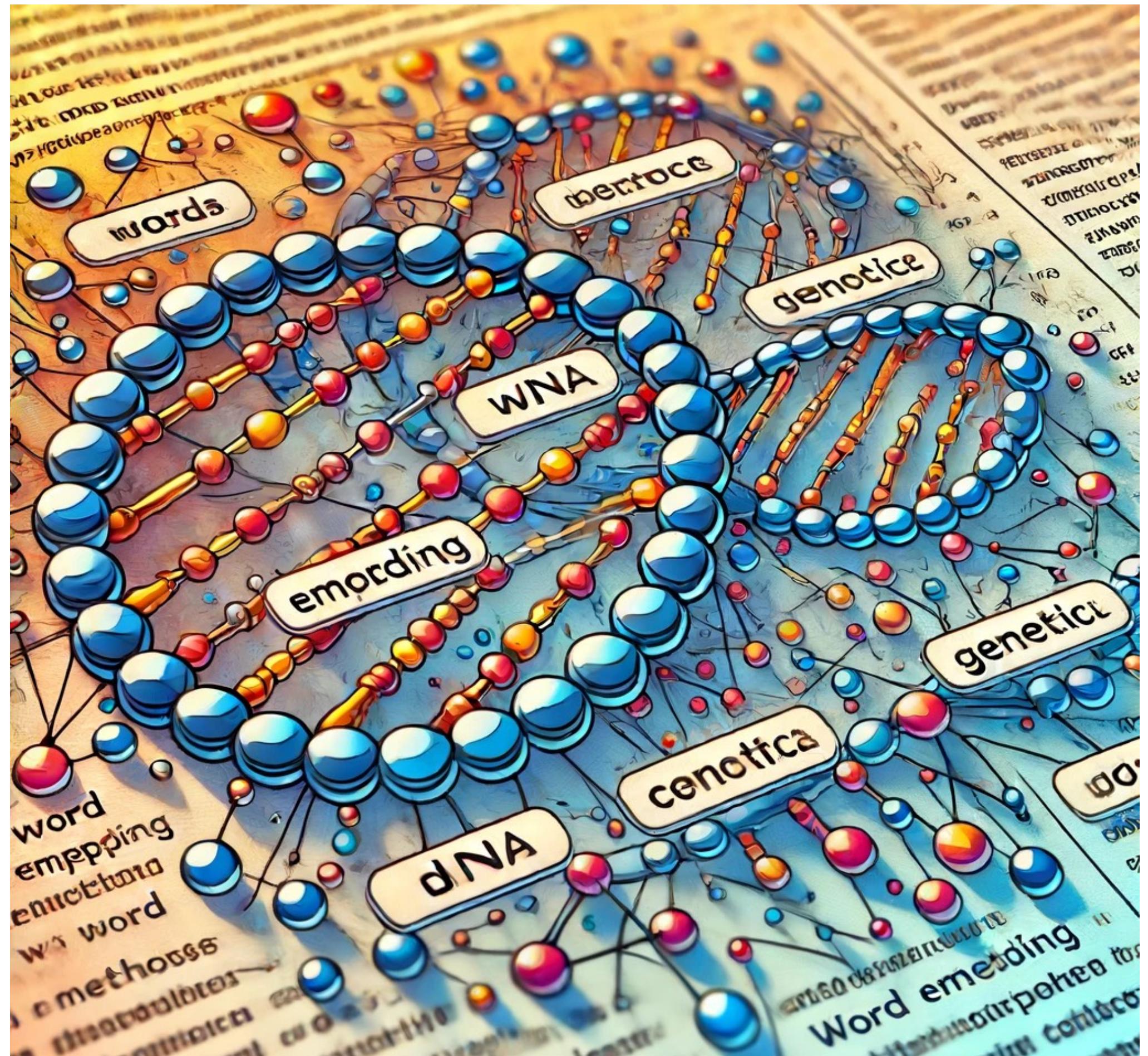
𝕏 Posted on X • 2:39 PM

## WORD EMBEDDINGS

- ▶ Up to this point, we have treated words as **sparse vectors**:  $\text{cat} = (1, 0, 0)$ .
- ▶ Now, we will think of words as **dense vectors**:
- ▶  $\text{cat} = (0.9, 0.1)$ ,  $\text{dog} = (0.8, 0.2)$ ,  $\text{fish} = (0.1, 0.9)$ .
- ▶ 
$$\text{cs(cat,dog)} = \frac{0.9 \times 0.8 + 0.1 \times 0.2}{\sqrt{0.9^2 + 0.1^2} \times \sqrt{0.8^2 + 0.2^2}} = 0.99$$
- ▶ 
$$\text{cs(cat,fish)} = \frac{0.9 \times 0.1 + 0.1 \times 0.9}{\sqrt{0.9^2 + 0.1^2} \times \sqrt{0.1^2 + 0.9^2}} = 0.22$$

### KEY INSIGHT OF WORD EMBEDDING METHODS

- ▶ We use a large corpus (e.g., wikipedia) to learn which words are similar.
- ▶ The distributional hypothesis: “you shall know a word by the company it keeps” (Firth 1957).
- ▶ We use those embeddings from our source corpus to learn about our target corpus (**transfer learning**).
- ▶ Huge advantage: totally unsupervised, just add text!



via DALL-E

### ADVANTAGES OF WORD EMBEDDINGS

- ▶ Encoding similarity.

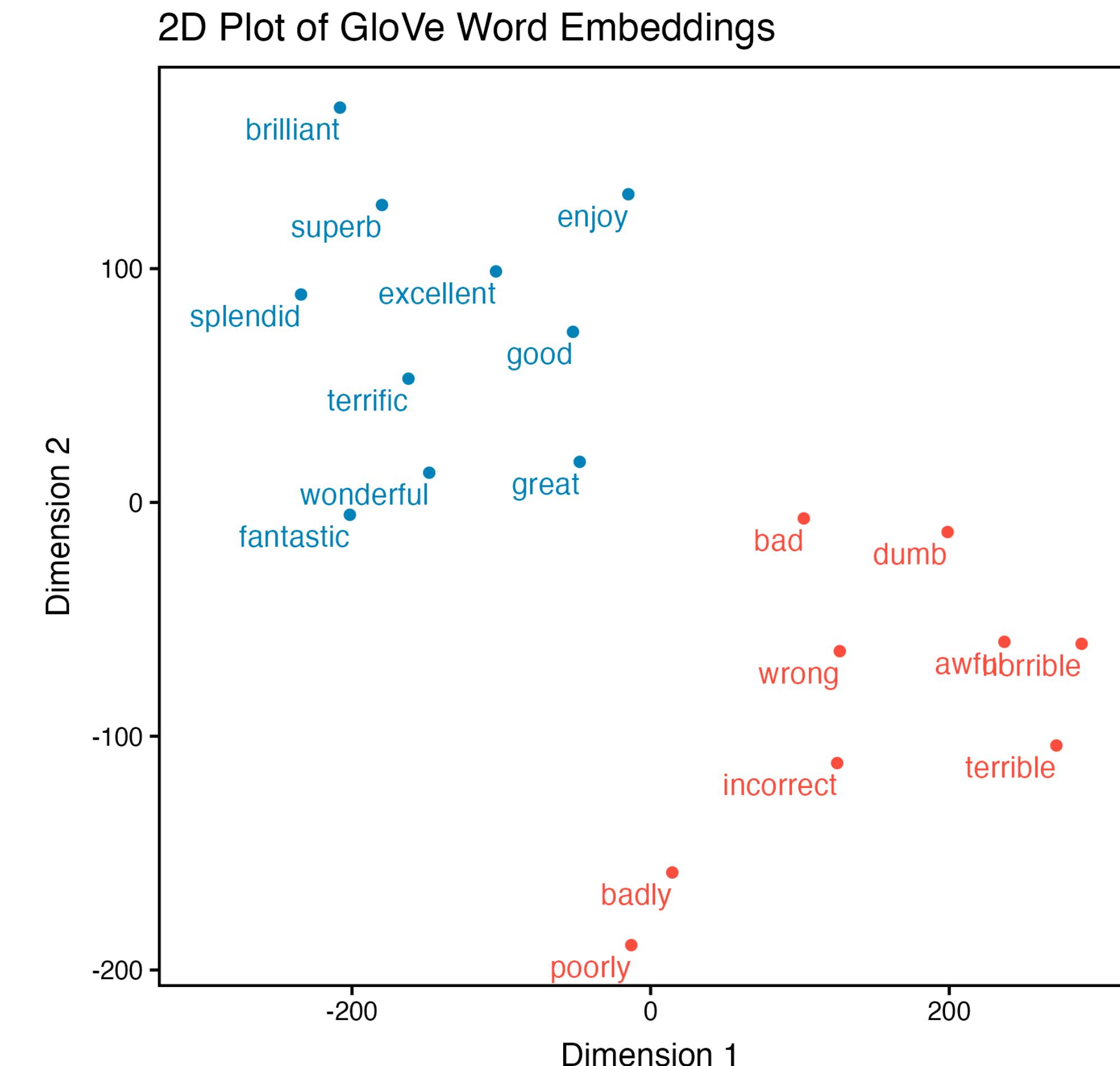
good = (-0.136, -0.116,...,0.246)

wonderful = (0.051,-0.349,...,0.095)

$\cos(\text{good}, \text{wonderful}) = 0.61$

### ADVANTAGES OF WORD EMBEDDINGS

- ▶ Encoding similarity.
- ▶ Automatic generalization.



### ADVANTAGES OF WORD EMBEDDINGS

- ▶ Encoding similarity.
- ▶ Automatic generalization.
- ▶ Measuring meaning.
- ▶ Rep. Mike Thompson (D-CA): "Make no mistake, **women's access to reproductive care** remains at risk."
- ▶ Rep. Mark Green (R-TN): "We must continue to fight against **mail-order abortion pills** and fight for **the lives of the unborn**."

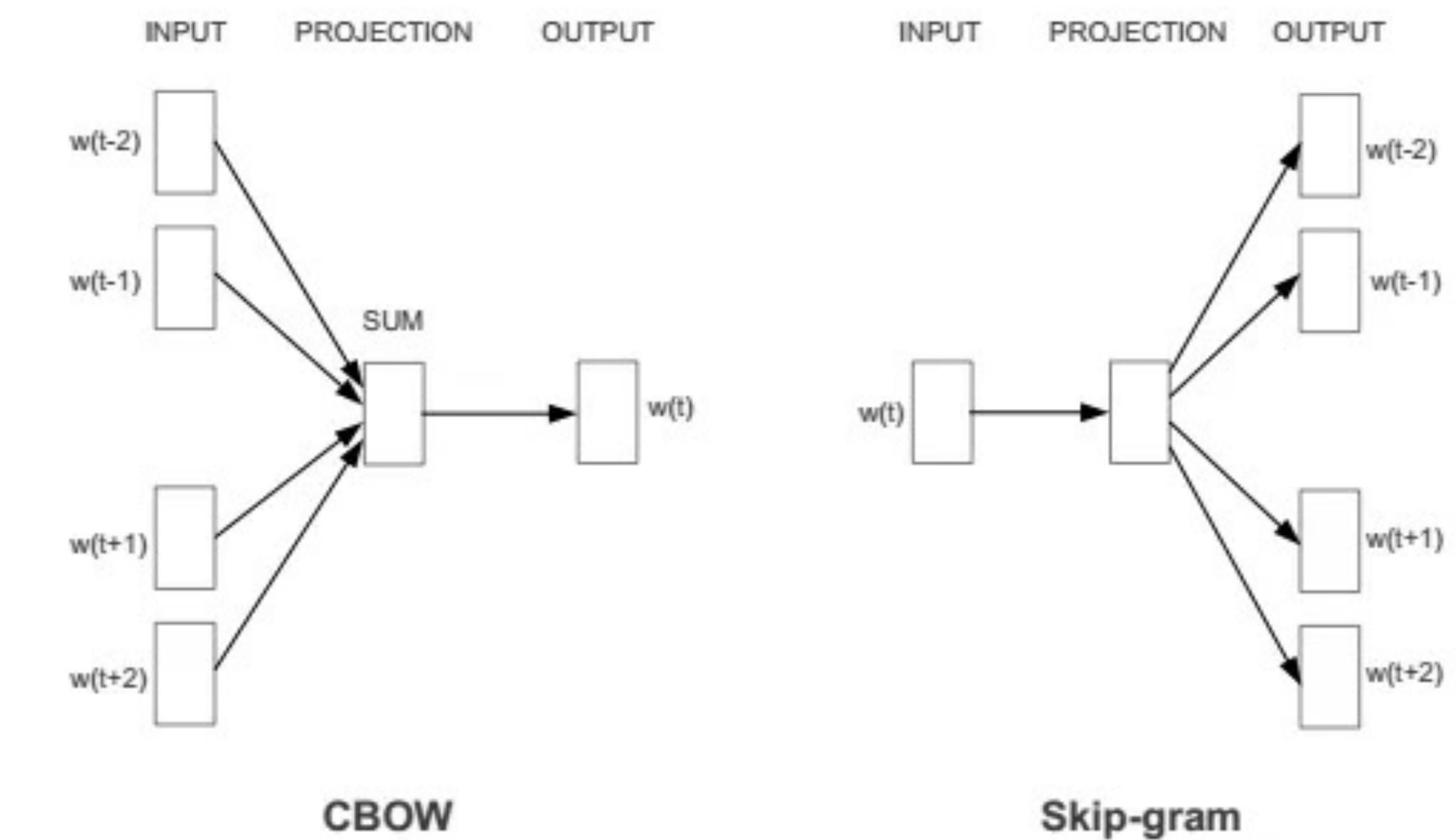
via Jamie Dupree

### SELF SUPERVISION

- ▶ Self-supervision: labels are provided by the data.

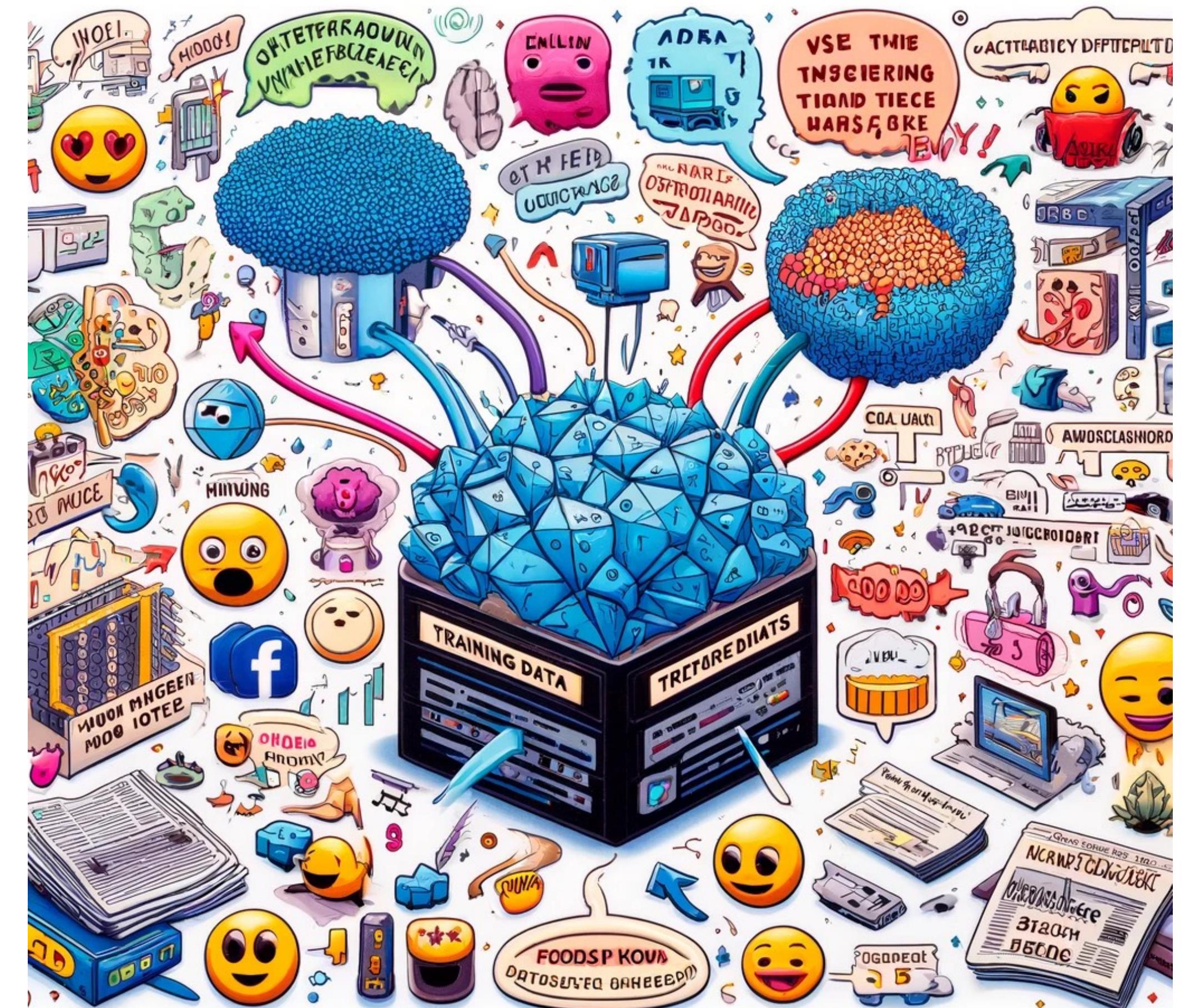
- ▶ An example: “Oranges are the color \_\_\_\_\_.”

- ▶ Orange: 60%, red: 30%, president: 0.0000000000001%.



### WHAT YOU NEED TO DECIDE AS THE RESEARCHER

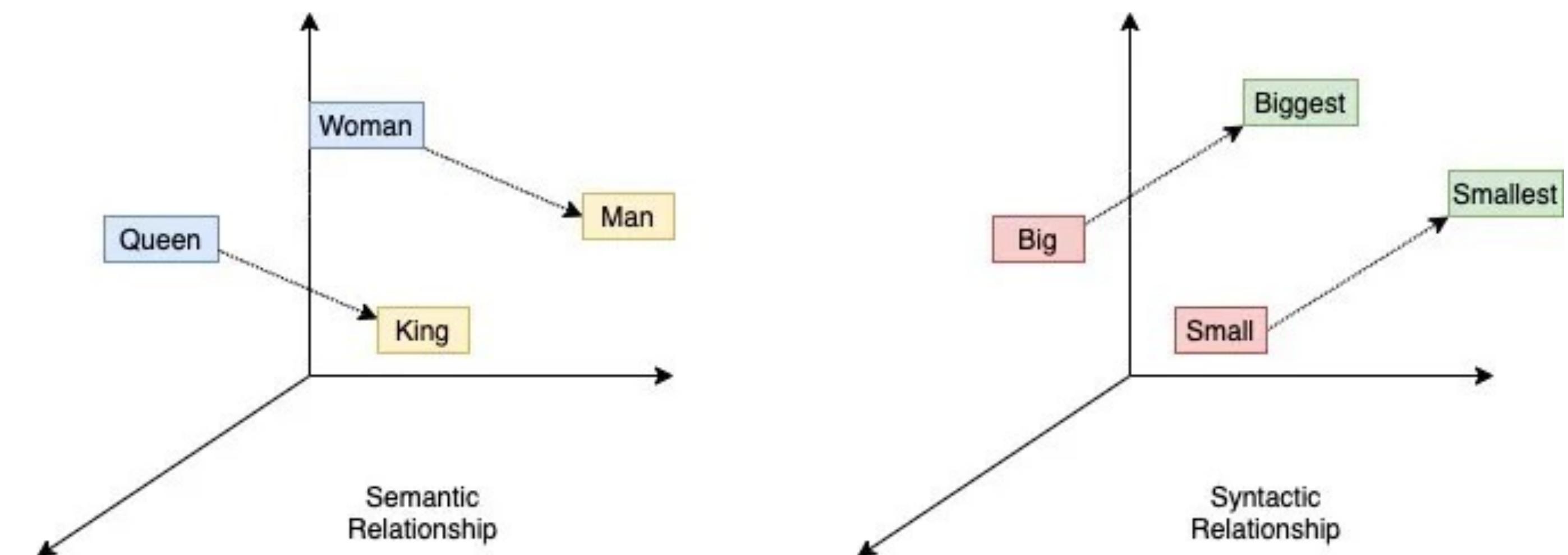
- ▶ The training data.



via DALL-E

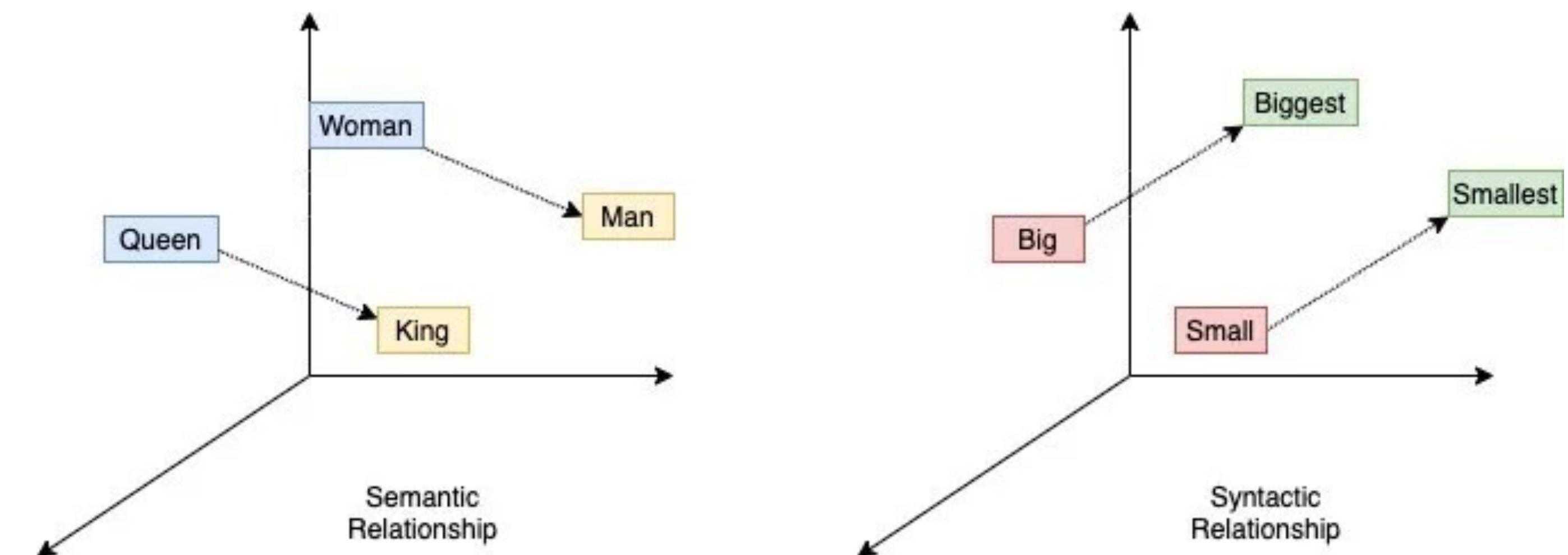
### WHAT YOU NEED TO DECIDE AS THE RESEARCHER

- ▶ The training data.
- ▶ The context window.



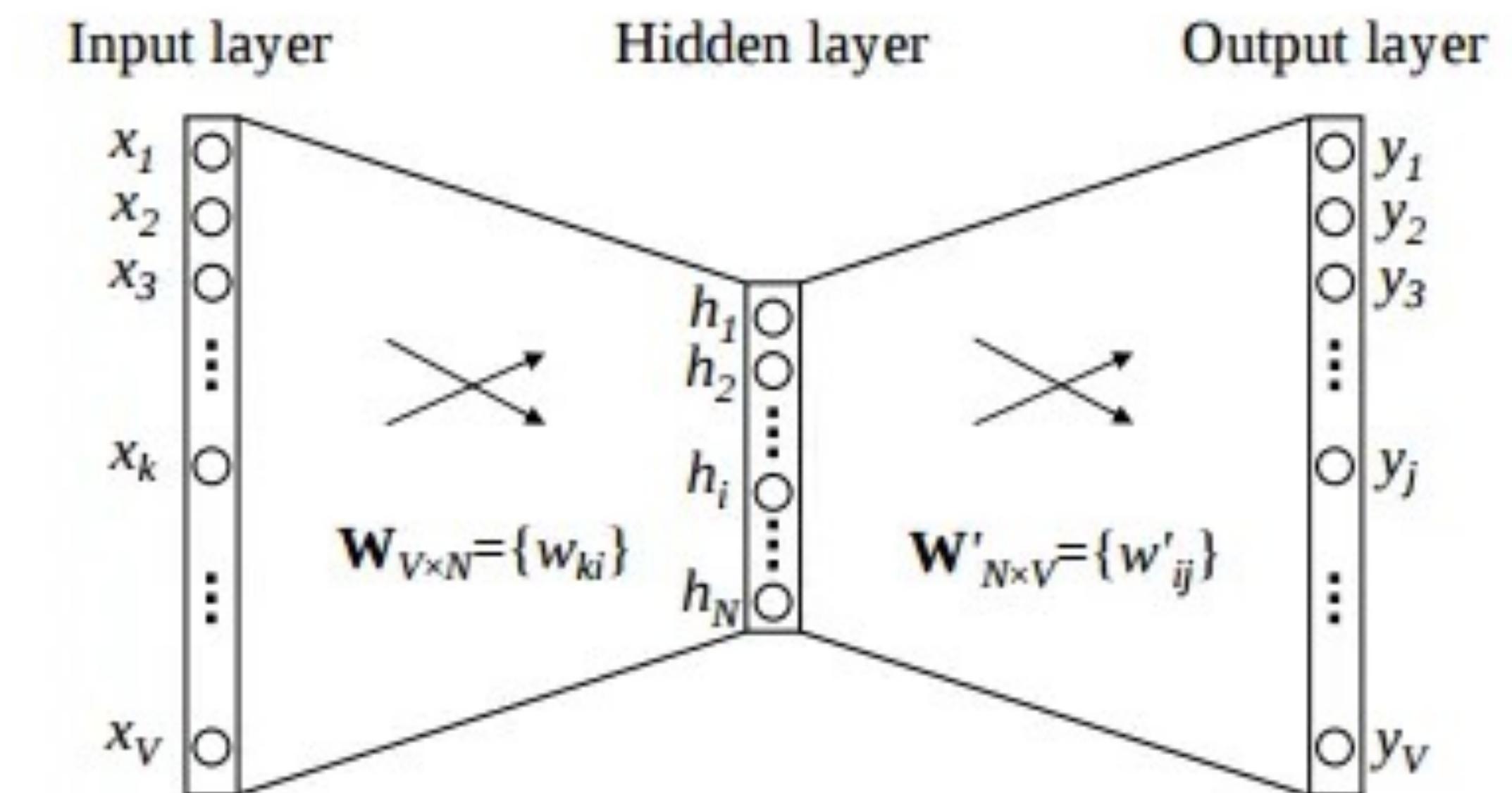
### WHAT YOU NEED TO DECIDE AS THE RESEARCHER

- ▶ The training data.
- ▶ The context window.
- ▶ The embedding dimensions.



### WHAT YOU NEED TO DECIDE AS THE RESEARCHER

- ▶ The training data.
- ▶ The context window.
- ▶ The embedding dimensions.
- ▶ The algorithm.



word2vec model architecture

### THE WORD2VEC CBOW MODEL

- ▶ CBOW: continuous bag of words.
- ▶ Consider the sentence: “the quick brown fox \_\_\_ over the lazy dog.”

$$\operatorname{argmax}_{\mu_{jumps}} = \frac{\exp(\mu_{jumps} \cdot \bar{v})}{\sum_j \exp(\mu_j \cdot \bar{v})}$$

- ▶ Where  $\bar{v}$  is the average embedding vector of all context words.

## THE GLOVE MODEL

- ▶ Tries to account for word co-occurrence across the corpus.
- ▶ Minimize  $J = \sum_{i,j=1}^V f(X_{ij}) \left( w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$ 
  - ▶ Where  $w_i^T \tilde{w}_j$  is the dot product of the target word and a context word.
  - ▶ Where  $b_i + \tilde{b}_j$  are bias terms that help with model fitting.
  - ▶ And  $\log X_{ij}$  is the log of the co-occurrence counts for the context and target.

### WORDS AS THE OBJECT OF STUDY

- ▶ Research question: do Republicans and Democrats mean something different when they say “immigration.”
- ▶ Hypothesis: Republicans use words about enforcement and crime; Democrats use words about reform and humanitarianism.

**TABLE 4. Top 10 Nearest Neighbors for the Target Term “Immigration”**

<b>Democrats</b>	enact, overhauling, reform, legislation, enacting, overhaul, reforming, revamp, entitlement, bipartisan
<b>Republicans</b>	enforce, laws, enact, enacting, legislate, legislations, enforcing, regularize, immigration, legislation

**TABLE 5. Subset of Top Nearest Contexts For The Target Term “Immigration”**

<b>Democrats</b>	this congress to take on comprehensive immigration reform and fix our broken immigration should get to work on comprehensive immigration reform the immigration system we have
<b>Republicans</b>	administration wants to ignore our nation's immigration laws and immigration process the problem broken is the enforcement of our immigration laws and we have seen that

### FROM WORDS TO DOCUMENTS

- ▶ Word embedding models output a matrix where each row is a word and the columns account for (unknown) attributes.
- ▶ One way to embed documents: average all of the word vectors.
- ▶ Determine document similarity through cosine similarity of document averages.



Fig. 4 Nearest neighbors of the LIWC positive emotions dictionary

Garten et al. (2018)

### EVEN MORE CONTEXT

- ▶ word2vec and GloVe are not context aware.
- ▶ e.g., only one embedding for “bat,” but are we talking about baseball or flying mammals?
- ▶ Models like BERT and GPT account for this context, but complicated and costly.



via DALL-E