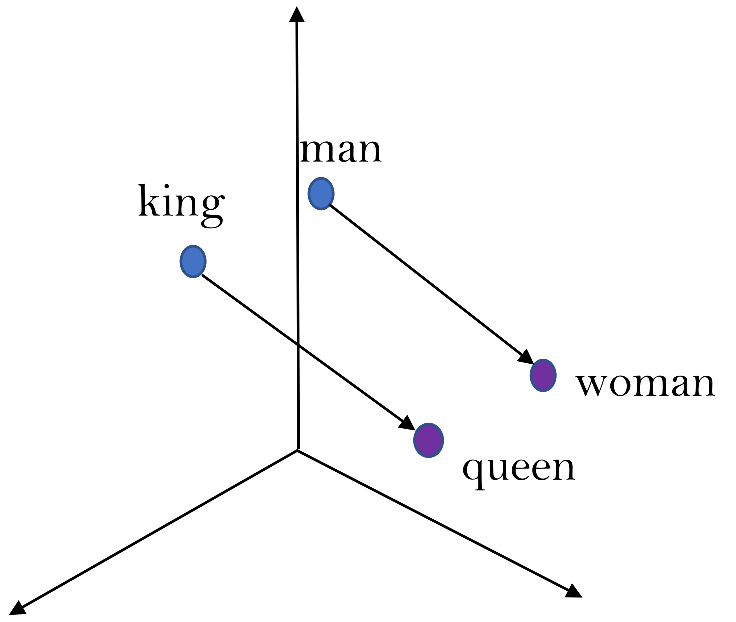
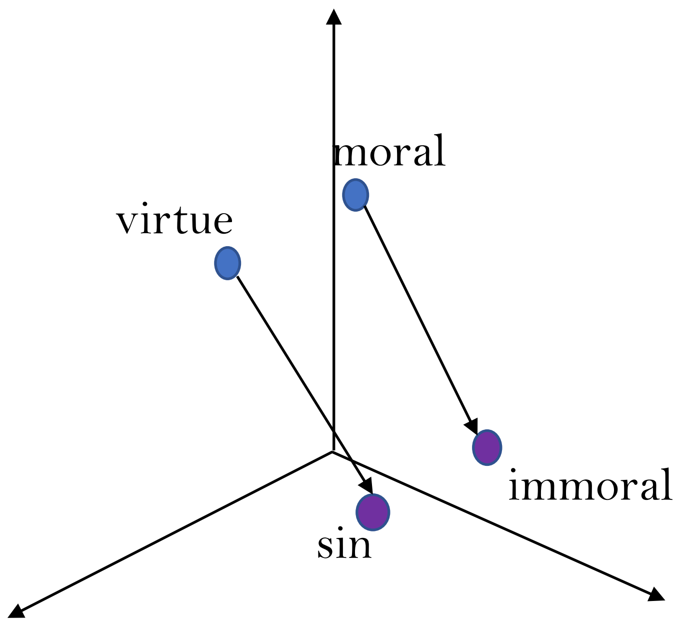
Learning meaning with neural word embeddings



Alina Arseniev-Koehler
UCLA Sociology



Outline

1. Why word embeddings?
2. What are word embeddings?
3. Models to “learn” word embeddings
4. Surprising features of word embeddings
5. Research applications of word embeddings

1. Why word embeddings?

from human-readable text → computer readable numbers

Representing text data with co-occurrences

"the increasing prevalence of obesity is like a hundred car freight train going downhill with no brakes"

"a national epidemic of childhood obesity"

"obesity is on the increase"

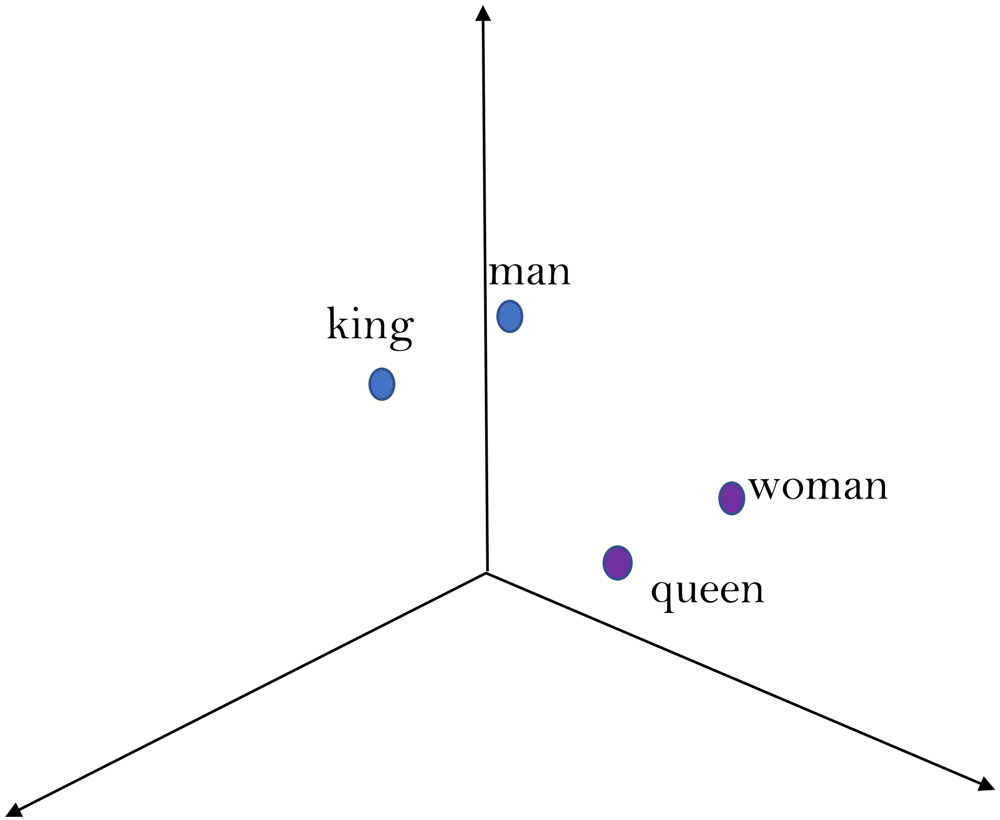
Vocabulary size: 24

	A	Brakes	Childhood	Downhill	Epidemic	…
A	0	1	1	1	1	…
Brakes	1	0	0	1	0	…
Childhood	1	0	0	0	1	…
Downhill	1	1	0	0	0	…
Epidemic	1	0	1	0	0	…
…	…	…	…	…	…	…

- each word represented as a (sparse) vector
- meaning is relational; distributional hypothesis
- measure similarity

2. What are word embeddings?

Representing text with embeddings

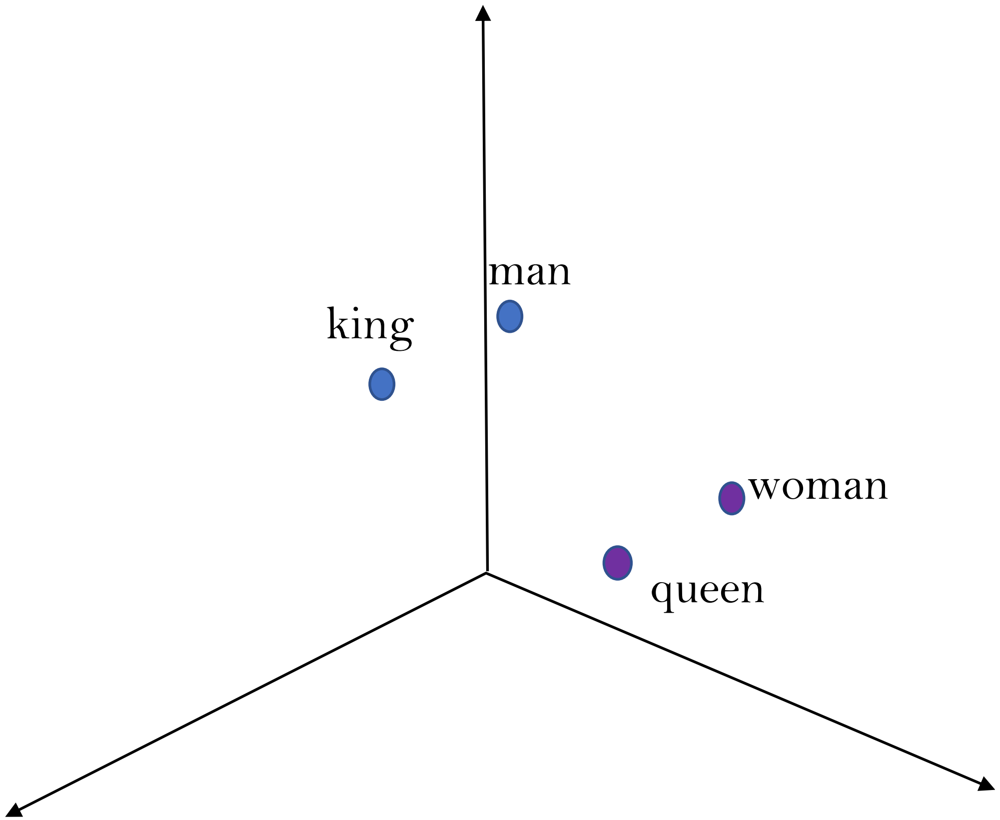


Meaning of a word is distributed across N dimensions

Vocabulary Word	Dimension_1	Dimension_2	Dimension_3
King	.07284	.383918	.0694749
Queen	0.2203	0.03286	0.032342
Man	0.027485	0.4286	0.103234
Woman	.28933	.11193	.11947
Womanly	.9284	.0535	.10324
…	…	…	…

- “learn” a vector representation from text data
- not so sparse anymore…

Representing text with embeddings



Meaning of a word is distributed across N dimensions

Vocabulary Word	Dimension_1	Dimension_2	Dimension_3
King	.07284	.383918	.0694749
Queen	0.2203	0.03286	0.032342
Man	0.027485	0.4286	0.103234
Woman	.28933	.11193	.11947
Womanly	.9284	.0535	.10324
…	…	…	…

```
currentmodel.wv.most_similar('woman', topn=5)
```

```
[('young_woman', 0.7615835070610046),
 ('girl', 0.6503051519393921),
 ('young_girl', 0.6443690061569214),
 ('housewife', 0.6167056560516357),
 ('man', 0.6113157868385315)]
```

```
currentmodel.most_similar('obesity', topn=10)
```

```
[('childhood_obesity', 0.8337880373001099),
 ('obesity_epidemic', 0.7985647916793823),
 ('Obesity', 0.773392915725708),
 ('childhood_obesity_epidemic', 0.7247533798217773),
 ('Childhood_obesity', 0.7202274799346924),
 ('obese', 0.6735087633132935),
 ('diabetes', 0.6472741365432739),
 ('unhealthy_diets', 0.6450413465499878),
 ('Physical_inactivity', 0.6439779996871948),
 ('heart_disease', 0.6417500972747803)]
```

- “learn” a vector representation
- not so sparse anymore…
- closeness = cosine similarity (crude)

3. Models to “learn” word embeddings

- Word2Vec (2 variants: SkipGram and CBOW)
- GlovE
- FastText
- BERT and ELMO
- ….

How does Word2Vec learn word-vectors?

Can you guess the missing word?

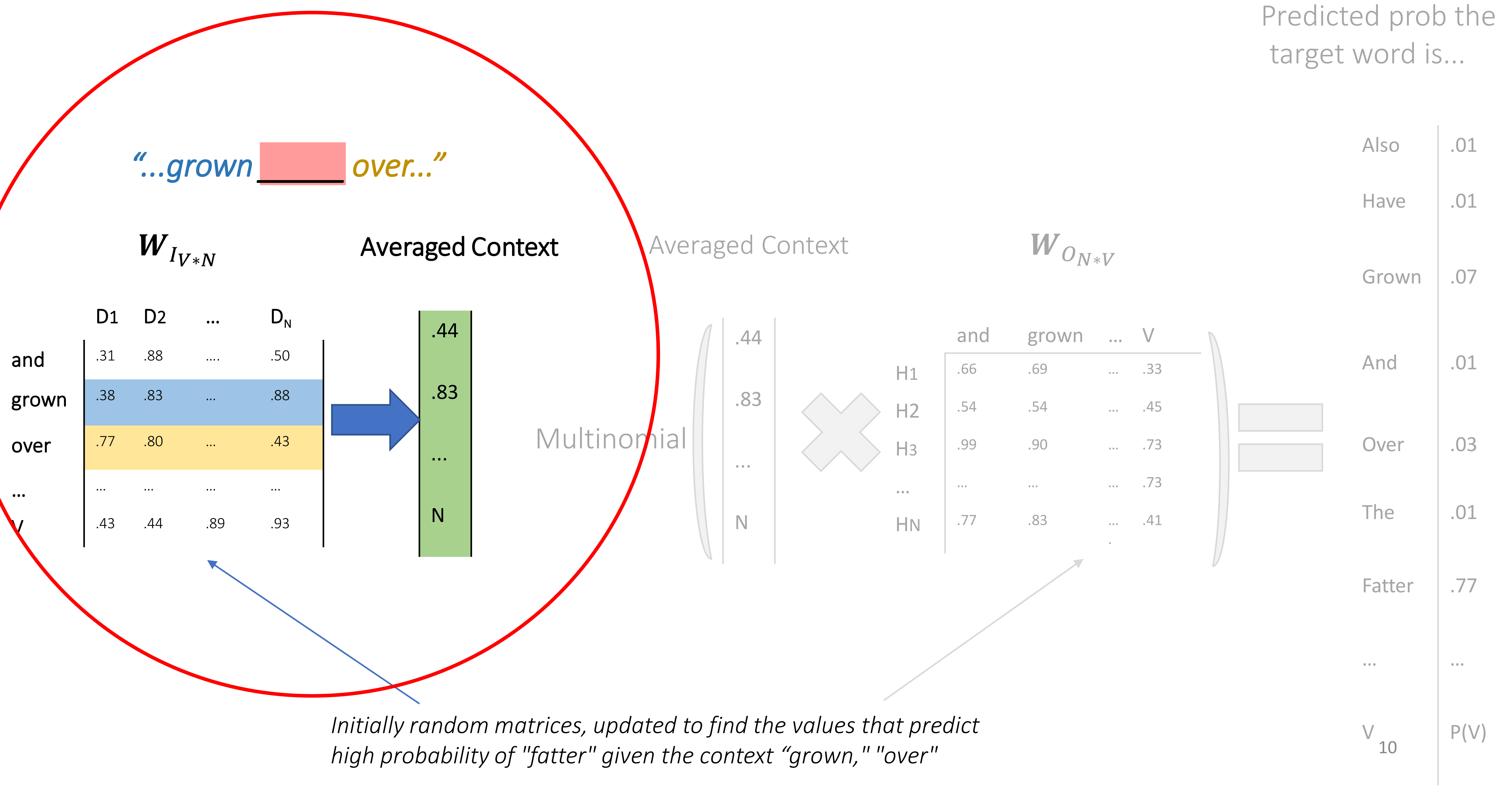
"…Americans have grown ______ over the last generation, inviting more heart disease, diabetes and premature deaths…"

What word has the highest *cosine similarity* to the context words?

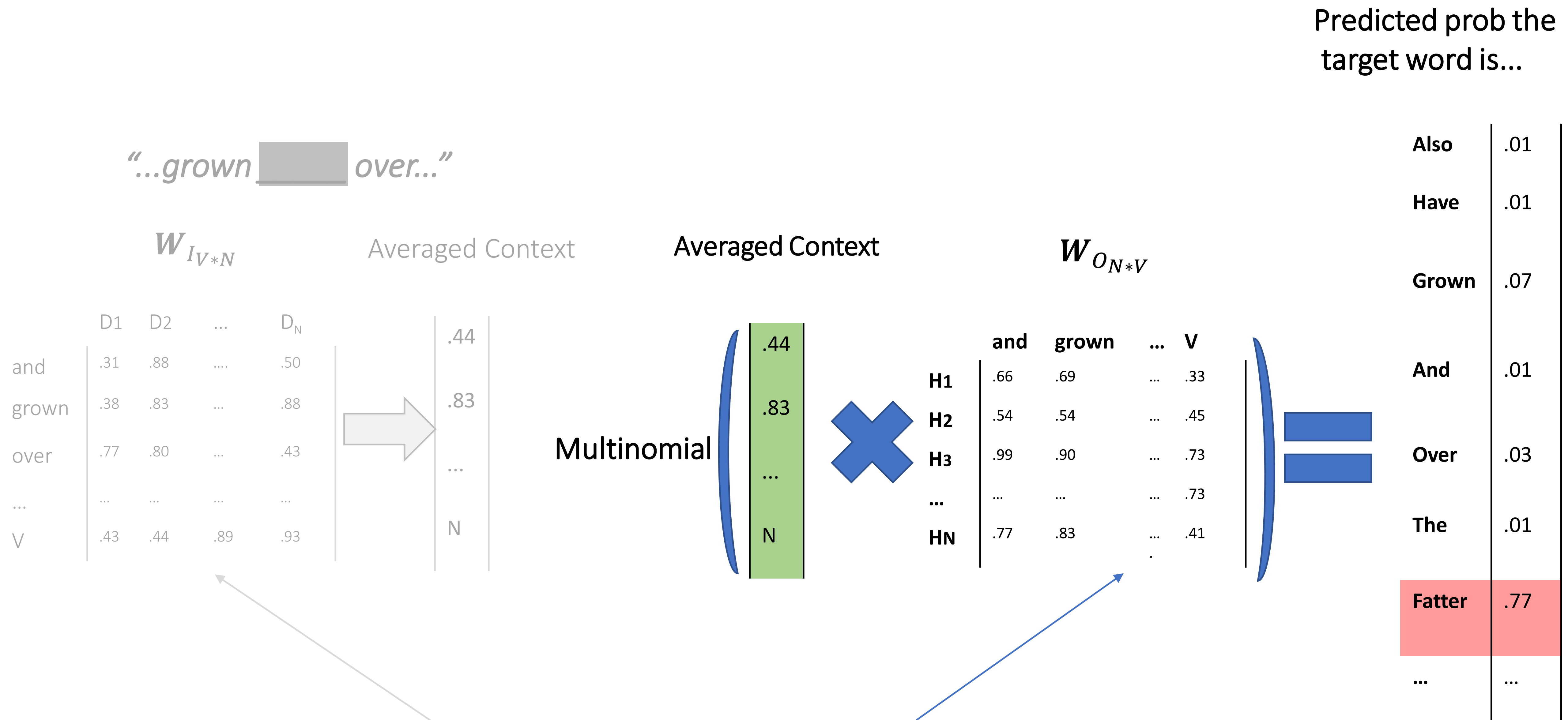
→We know what the missing word actually is in the NYT ("fatter")

1. Word2Vec gives correct answer? Then we have good word-vectors
2. Wrong answer? Tweak the word-vectors

How does Word2Vec learn word-vectors? (The details!)

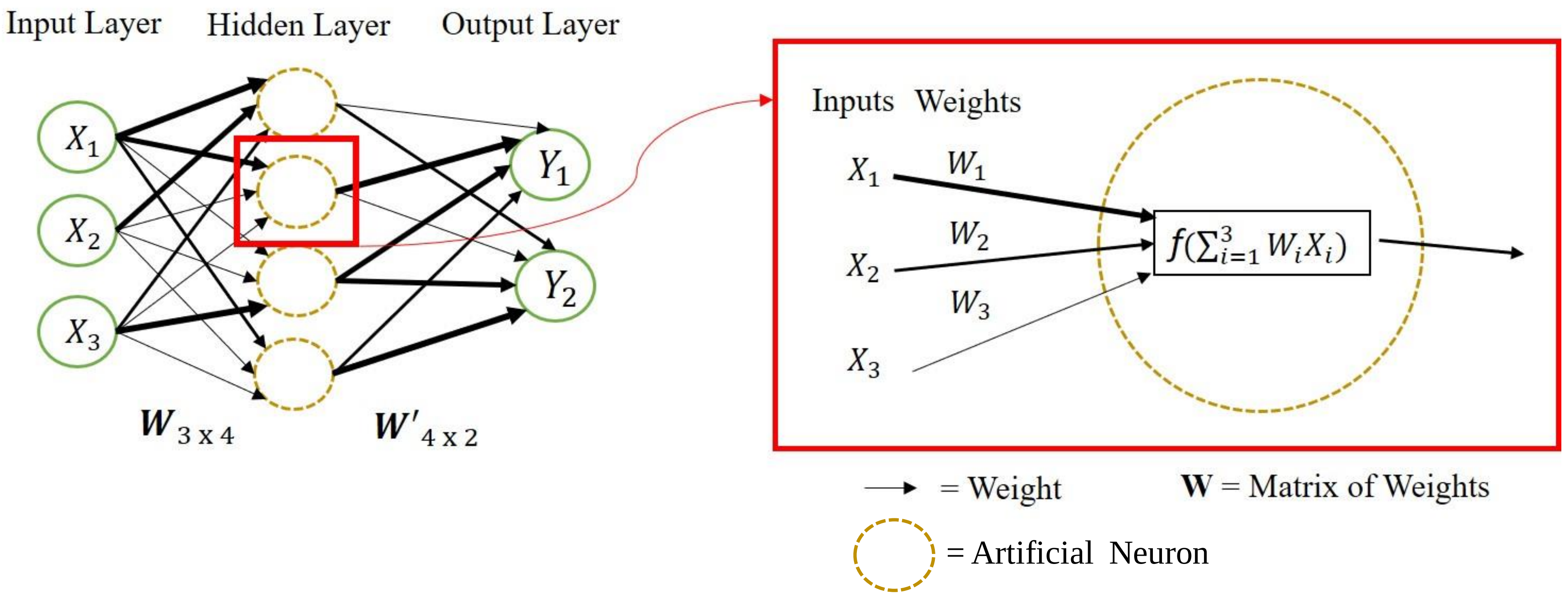


How does Word2Vec learn word-vectors? (The details!)

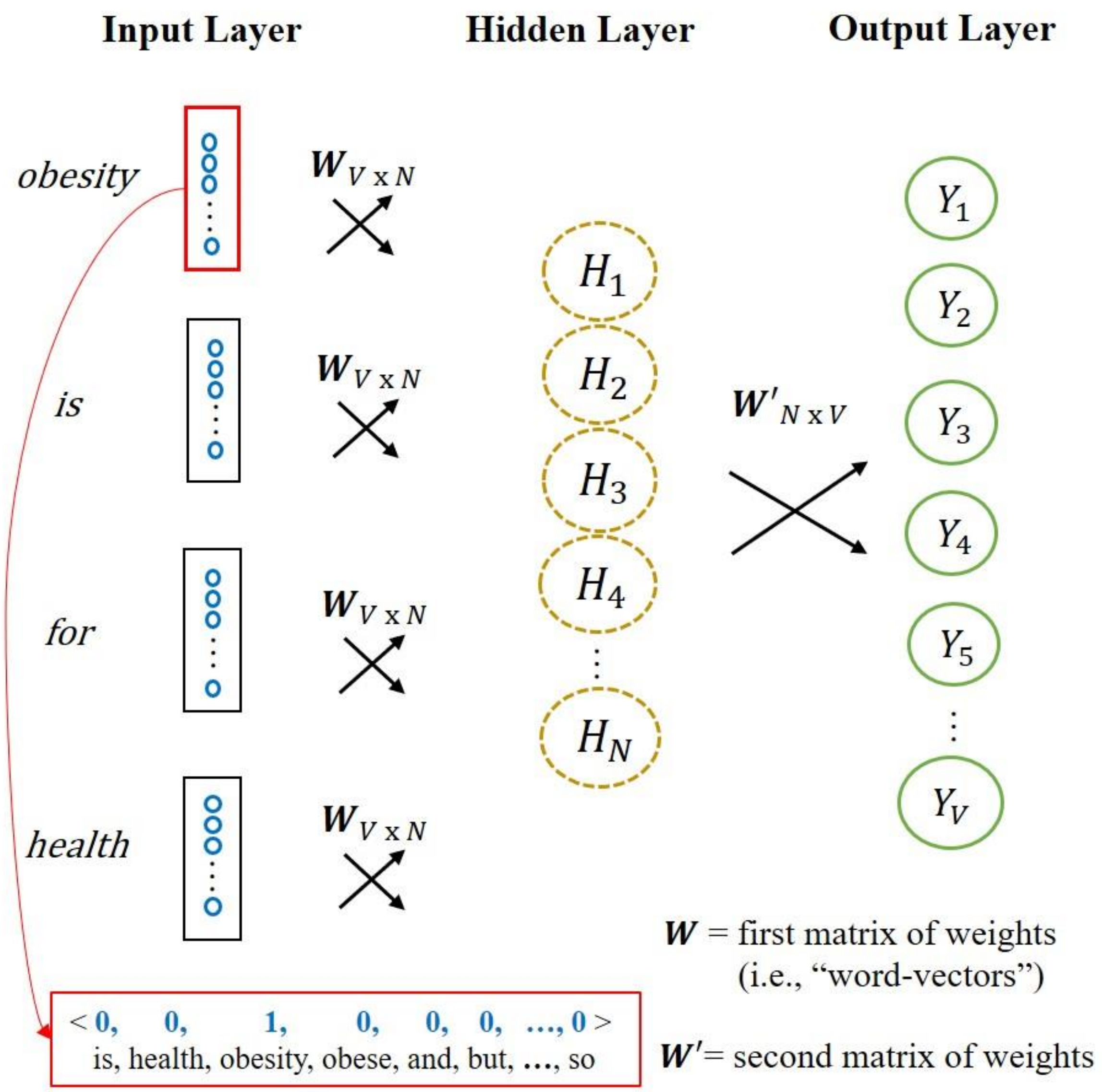


Initially random matrices, updated to find the values that predict high probability of "fatter" given the context "grown," "over"

Simple Artificial Neural Network (ANN)

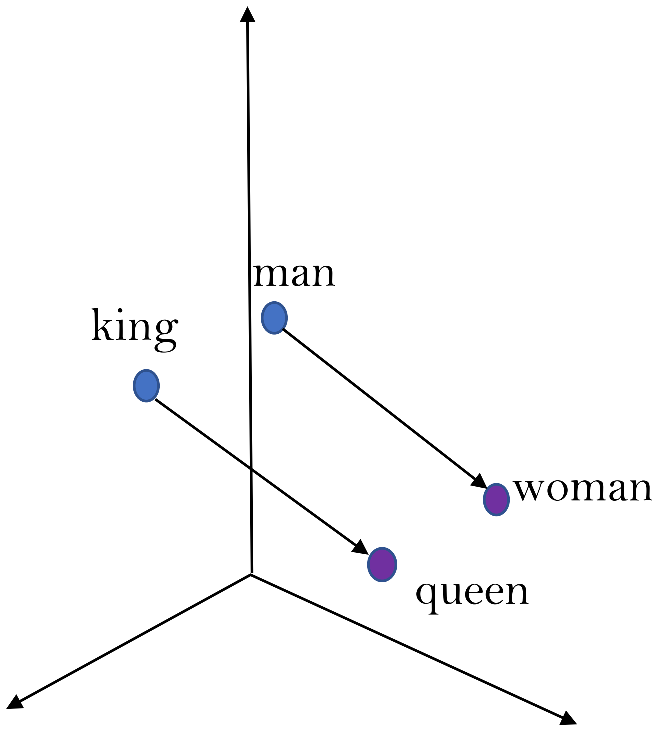


How does Word2Vec learn word-vectors? ANN explanation☺

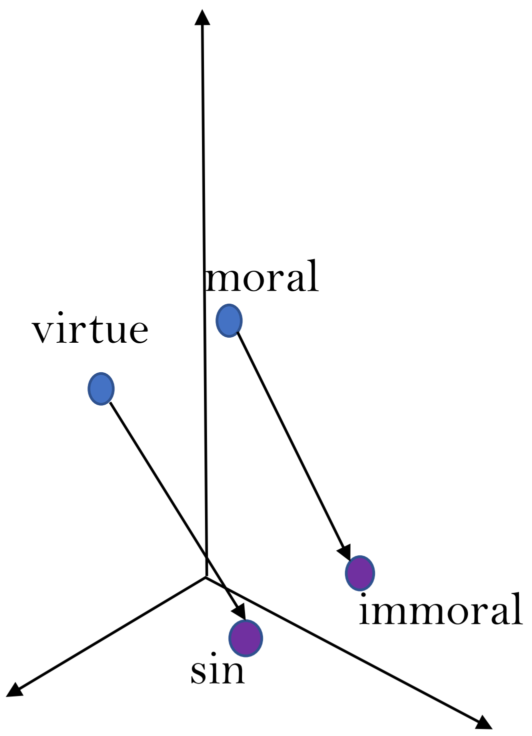


4. Surprising features of word embeddings

Latent Dimensions in Embeddings

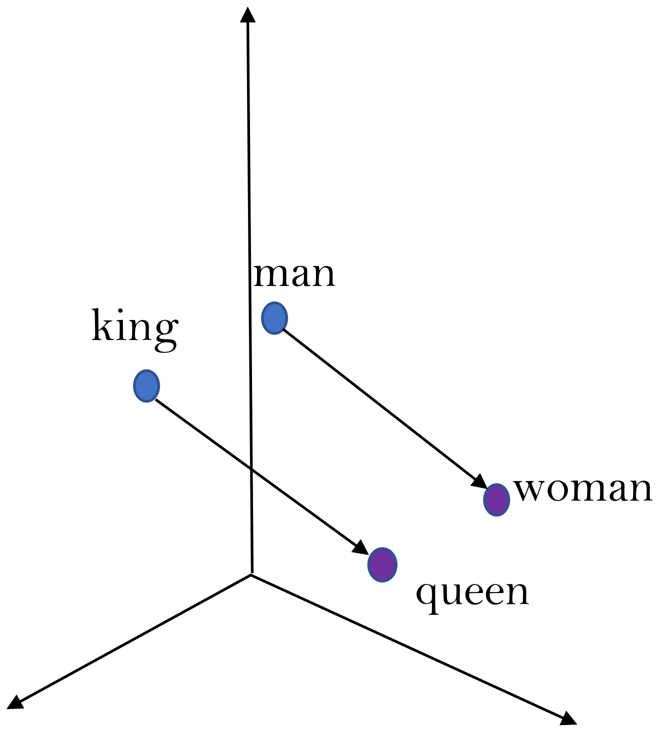


a "gender" dimension



a "moral" dimension

Latent Dimensions and Analogies



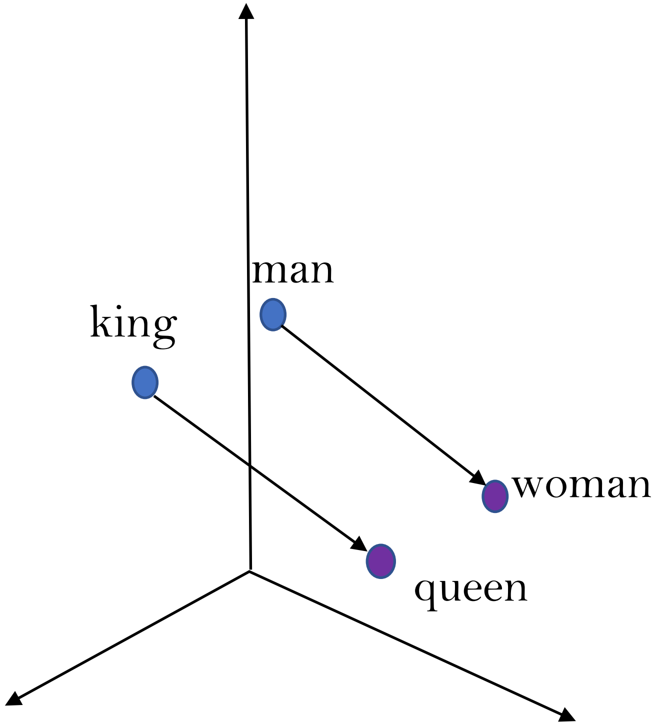
a “gender” dimension

“man” is to “woman” as “king” is to _____?

If:
woman-man= queen – king
Then:
(woman-man) + king= queen

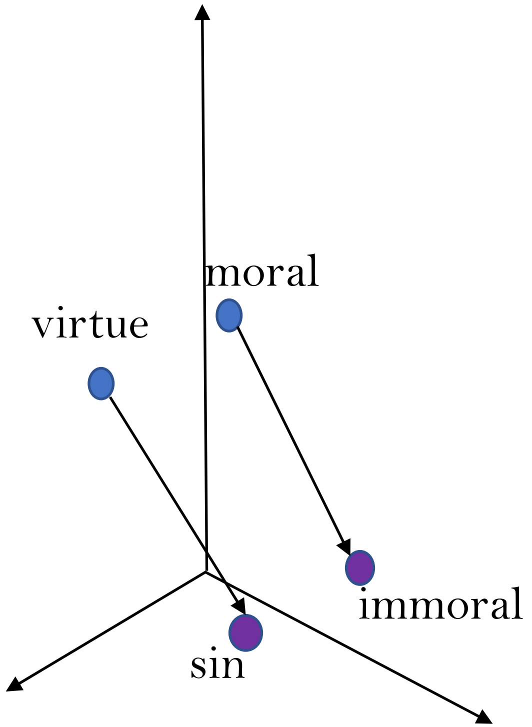
Now, try to solve with word-vectors:
What is the closest word-vector to (woman-man) + king ?

How to Extract Latent Dimensions



a "gender" dimension

=AVG(feminine words) – AVG(masculine words)



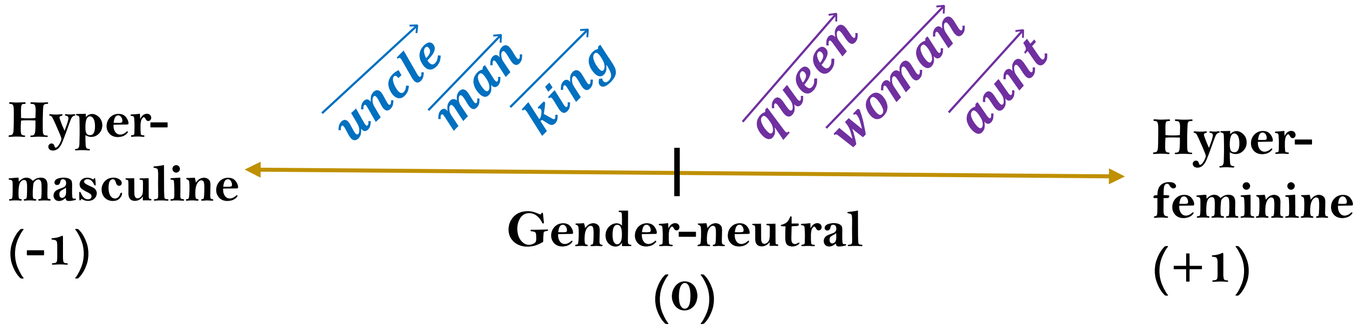
a "moral" dimension

=AVG(moral words) – AVG(immoral words)

*other methods, too
*fewer words may be better

E.g., how is a word gendered?

- Cosine similarity between this *gender dimension* and some *new word*
  - Tells us gender as masc (-) or fem (+), and strength



Cosine sim $(\overrightarrow{gender}, \overrightarrow{woman}) = .2$

Cosine sim $(\overrightarrow{gender}, \overrightarrow{man}) = -.2$

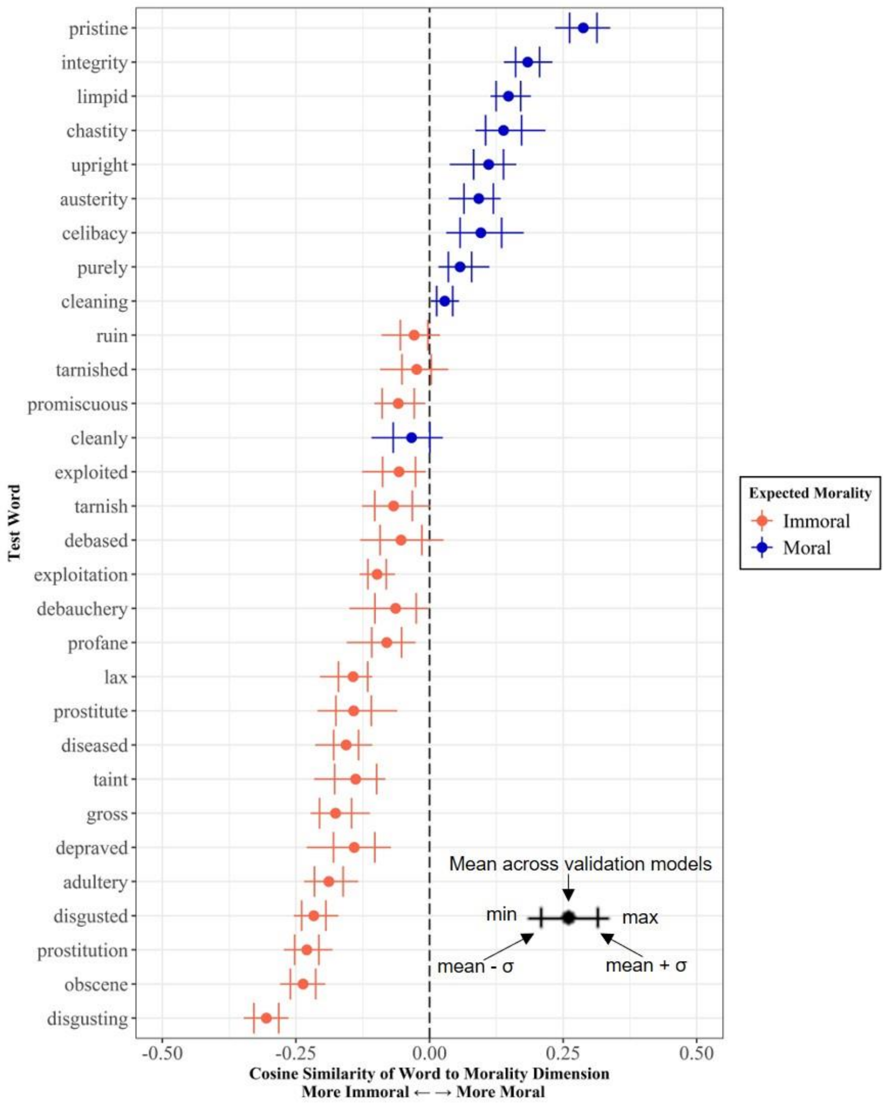
- Generalizable to morality, health, social class, etc.
- Meaning is relational; binary opposition

Kozlowski, Taddy, Evans 2018
Levi-Strauss 2008

Questions on Dimension Extraction?

See Extra Slides on Latent Dimensions at theend

Validating an extracted dimension



Embeddings learned from the *New York Times*

Predicted moral purity of a word =
cosine similarity between purity dimension and word

Most test words classified as expected!

Ideally, compare to survey data from diverse pool of raters, or IAT data, e.g., Caliskan et al (2017)

Validating an extracted dimension (binary or continuous)

	Testing Words Correctly Classified N (%)	Total Testing Words
Gender	57 (95%)	60
Morality	59 (98%)	60
Health	55 (92%)	60
Social Class	59 (98%)	60

Binary: Moral (-) or Immoral (+) classification
Continuous: *How* moral or immoral

5. Research applications of word embeddings

Implementation

1. Collect text data, clean it, train a model (e.g., Gensim in Python)
   - Decide hyperparameters (e.g., dimensionality)
   - Validate with Google Analogy Test
2. OR, use a pre-trained model

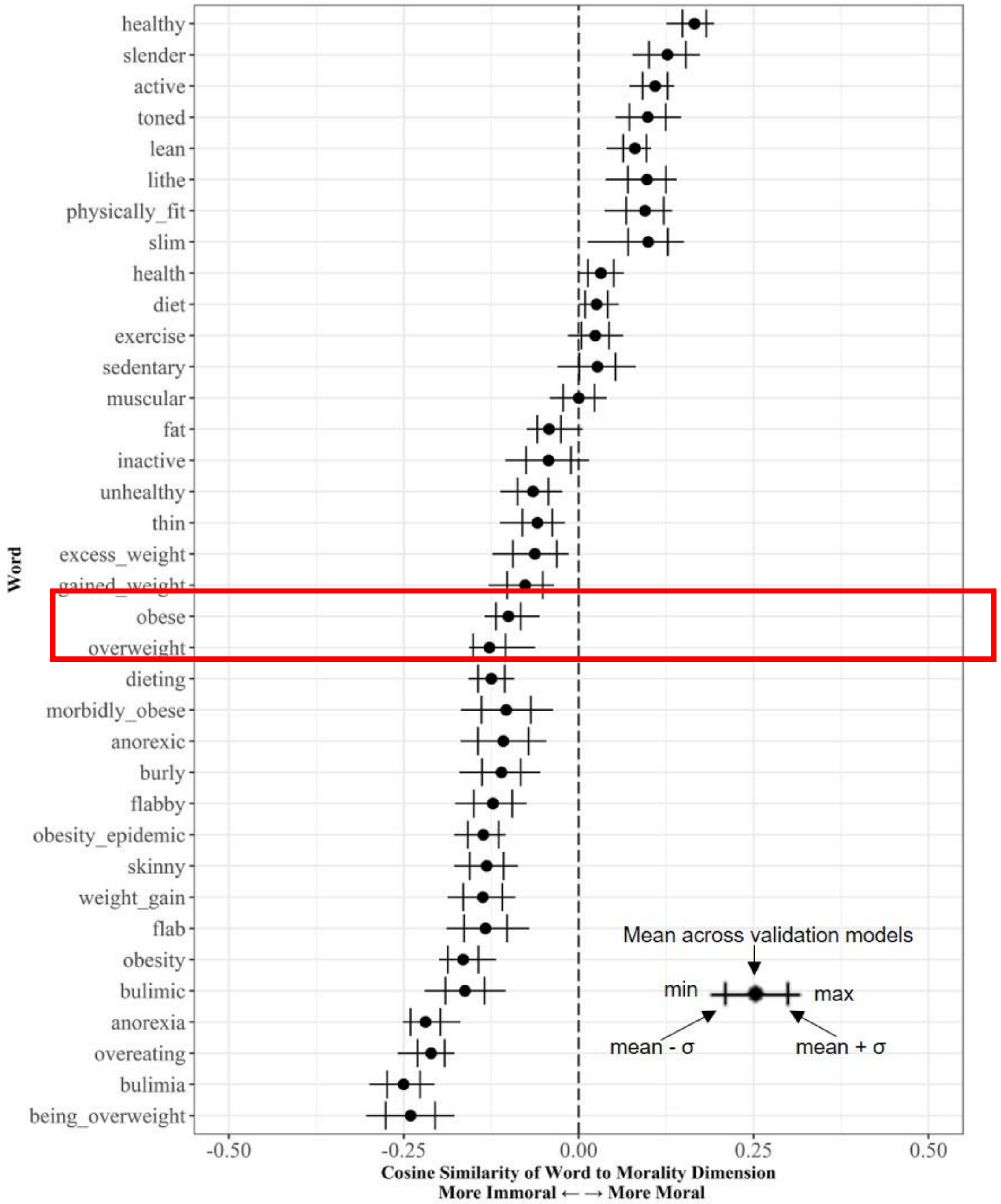
- Pros/cons of each?

Obesity in News Discourse

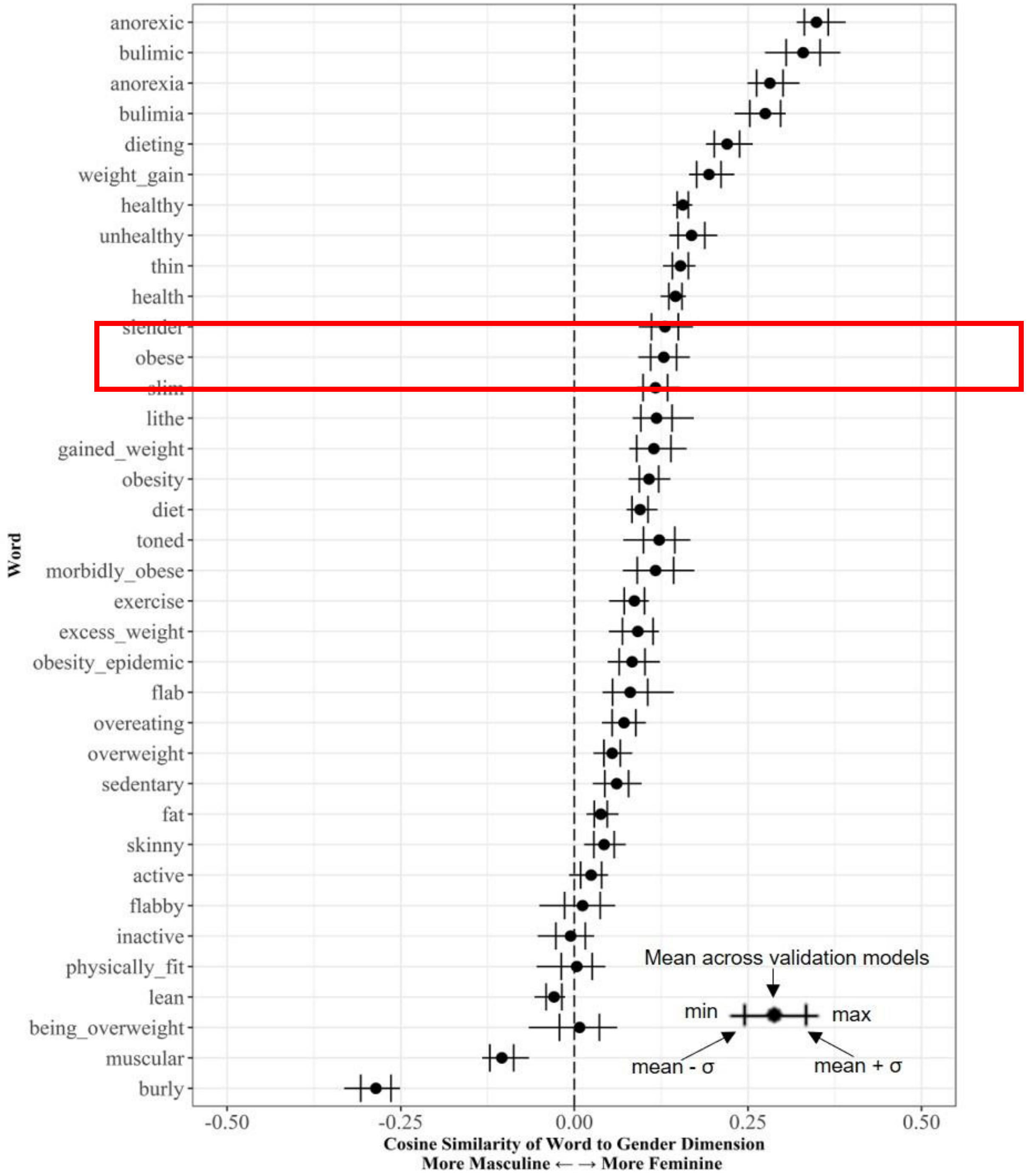
- Word2Vec models trained on 100k New York Times articles on obesity and health 1980-2016

- Qualitative literature: obesity connotes immorality, low class, and illness, and is often discussed in context of women

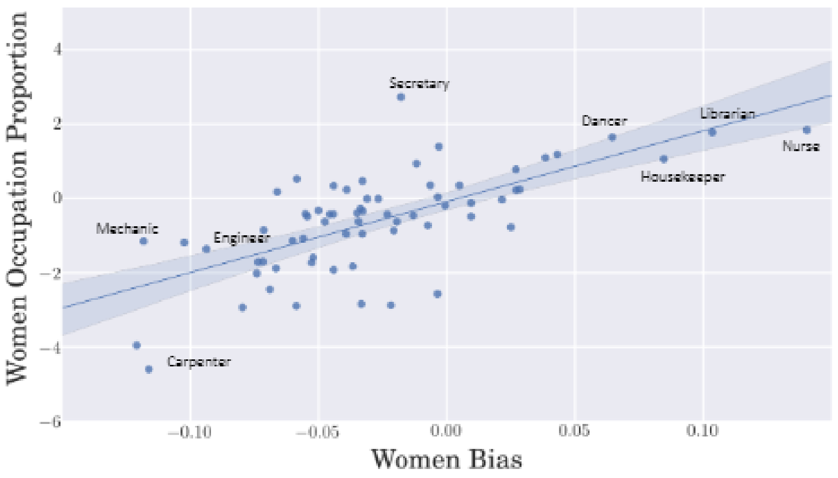
1. Extract 4 dimensions: morality, social class, health, and gender
2. Test how these dimensions make up the meaning of keywords around body weight (“obese,” “overweight,” “slender,” etc.)



Obesity is Immoral in News Discourse



Obesity is Gendered in News Discourse

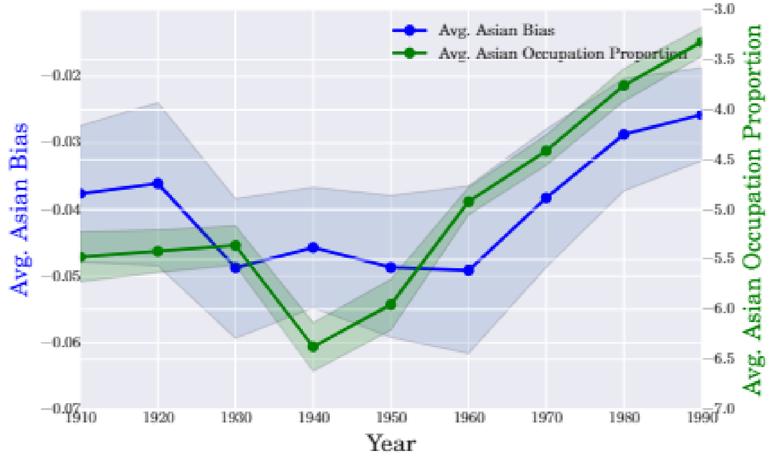


(a) Woman occupation proportion vs embedding bias in Google News vectors. More positive indicates more women biased on both axes. $p < 10^{-9}$, r-squared = .462.

(b) Average gender bias score over time in COHA embeddings in occupations vs the average log proportion. In blue is relative women bias in the embeddings, and in green is the average log proportion of women in the same occupations.

Hispanic	Asian	White
housekeeper	professor	smith
mason	official	blacksmith
artist	secretary	surveyor
janitor	conductor	sheriff
dancer	physicist	weaver
mechanic	scientist	administrator
photographer	chemist	mason
baker	tailor	statistician
cashier	accountant	clergy
driver	engineer	photographer

(c) The top ten occupations most closely associated with each ethnic group in the Google News embedding.



(d) Average ethnic (Asian vs White) bias score over time for occupations in COHA (blue) vs the average conditional log proportion (green).

Garg, Nikhil, et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes."

Analyses with word embeddings

- Look at most similar words to each other, get to know the model
- Look at specific dimensions
- Clustering among words, or relationships between a specific word-set
  - e.g., obesity words, or occupational words
- Use word embeddings to *predict* an outcome
  - e.g., predict racism from tweets

- Other embedding models:
  - Glove
  - FastText (sub-word vectors)
  - ELMO, BERT
  - Doc2Vec, Sentence Embeddings

5. Assumptions about the nature of meaning

“Dimensions” in Semantic Space

- Examples:
    - masculine/feminine
    - moral/immoral
    - low class/high class
    - health/illness
    - attractive/unattractive
    - positive sentiment / negative sentiment
    - safe / dangerous

- Key structure of meaning: binary opposition

- Meaning is relational, defined by contrasts

Kozlowski, Taddy, Evans 2018
Levi-Strauss

Binary Oppositions

- two concepts that are defined by each other, such as "masculinity" and "femininity" align in many aspects of meaning (such as that both are human and animate) but differ on one specific aspect (here, gender).

Componential Analysis				
	Human	Animacy	Age (Adult)	Gender (Feminine)
"Woman"	+	+	+	**+**
"Man"	+	+	+	**-**

Culture as binary oppositions

- A lot of meaning takes on this binary form

- But does *all* meaning takes this form?
  - What about "gender-neutral" words?
  - Why are some meanings (i.e., gender) easier to extract than others?
    - Safe/danger; strength/weakness; high prestige/low prestige

Still many unknowns…

- Sometimes, these methods to extract oppositions don't work so well…why not?
  - E.g., safe/danger, strength/weakness
  - "Purity" of the opposition
  - Different types of binary oppositions (e.g., mutually exclusive, exhaustive, gradable, binary, continuous)

- How much do training words matter?

- How much does the training corpus matter?

- How do we select training words?

- How should we validate extracted dimensions?

Are we just picking up “good” vs “bad?”

- Cosine similarity between sentiment and …..
  - Morality: .65
  - Gender: .05
  - Health: .57
  - Social Class: .28

- Meaning is interrelated:
  - CosSim(Moral, Social Class) = .24
  - CosSim(Moral, Health) = .53
  - CosSim(Social Class, Health) = .23

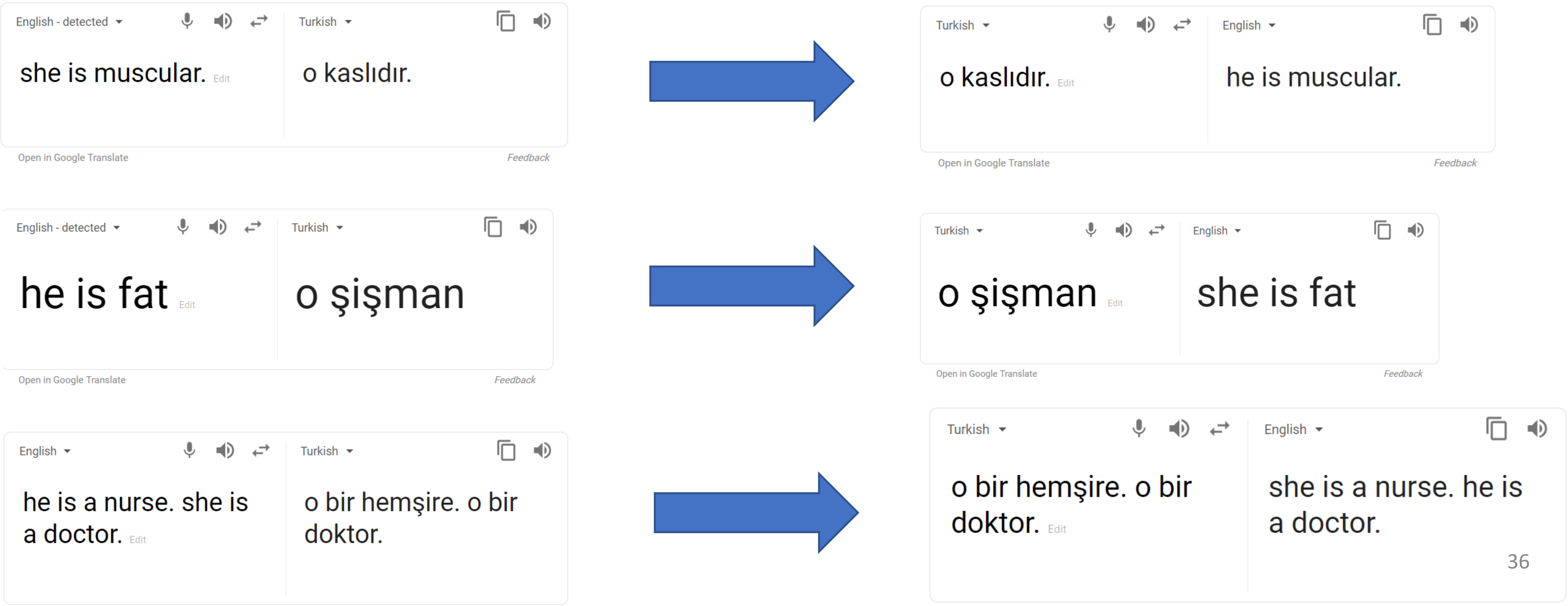
Culture, or "Machine-Learned Bias"?

- Machine-learned biases
  - Bolukbasi et al (2016) extract gender to show how occupations are gendered, and how to remove this gender bias
  - Caliskan et al (2017) examine gender biases in occupations

- Social Science and Culture
  - Garg et al (2017) examine gender and race of occupations across time
  - Kozlowski et al (2018) extract gender, wealth, political views and race
  - Arseniev-Koehler and Foster (in preparation) extract gender, morality, social class, and health in obesity news

Google translate infers gender

Machine-learned model of language
machine-learned **bias**? meanings of gender in our **language**?



Takeaways from Word Embeddings Applications

- **Cultural dimensions are encoded in word embeddings**

- Method and theory
  - Meaning is relational, structured
  - assuming vs discovering structure

- Next directions:
  - validation?
  - other structures of meaning?
  - how do structures of meanings change and vary?
  - How does meaning affect (and get affected by) social outcomes? (e.g., occupational stereotypes)
  - static vs contextualized

Questions?

Contact: arsena@g.ucla.edu

Code and Tutorials: https://github.com/arsena-k/Word2Vec-bias-extraction

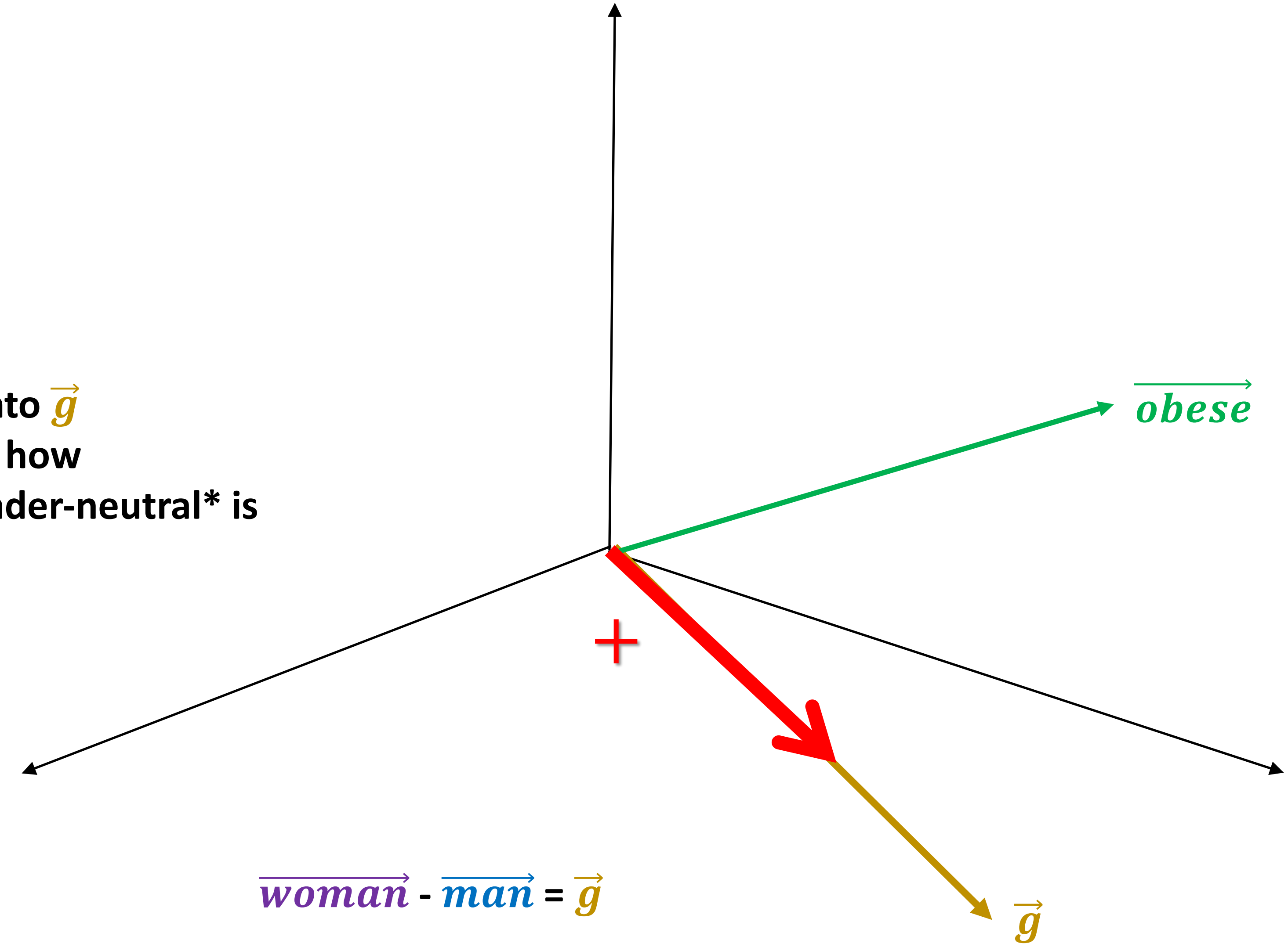
Next: Download repo for code:

https://github.com/arsena-k/Word2Vec-bias-extraction

Will also need to download, at the least, a trained embedding model (directions in code)

**Extra Slides on Latent Dimensions

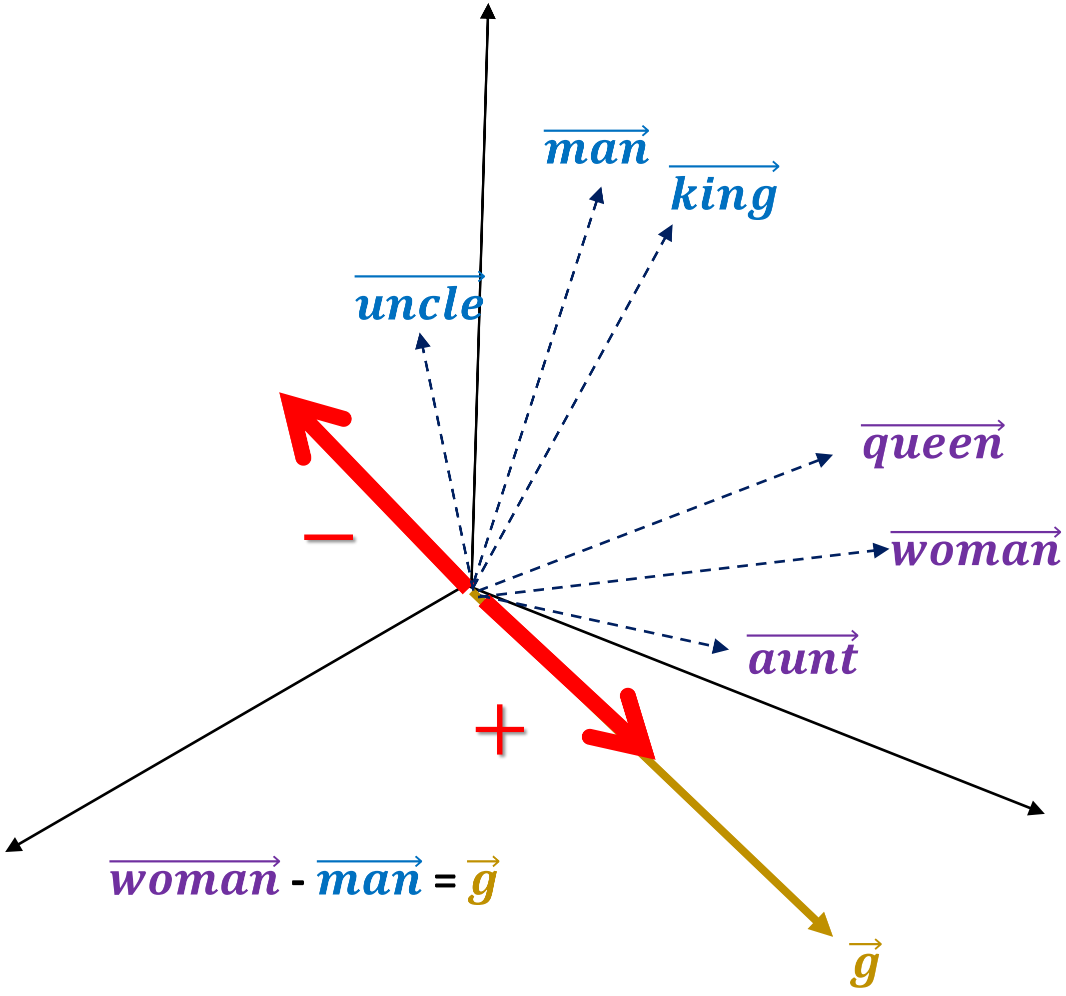
Project $\overrightarrow{obese}$ onto $\vec{g}$
To get a value for how
masculine, feminine, or gender-neutral* is
$\overrightarrow{obese}$

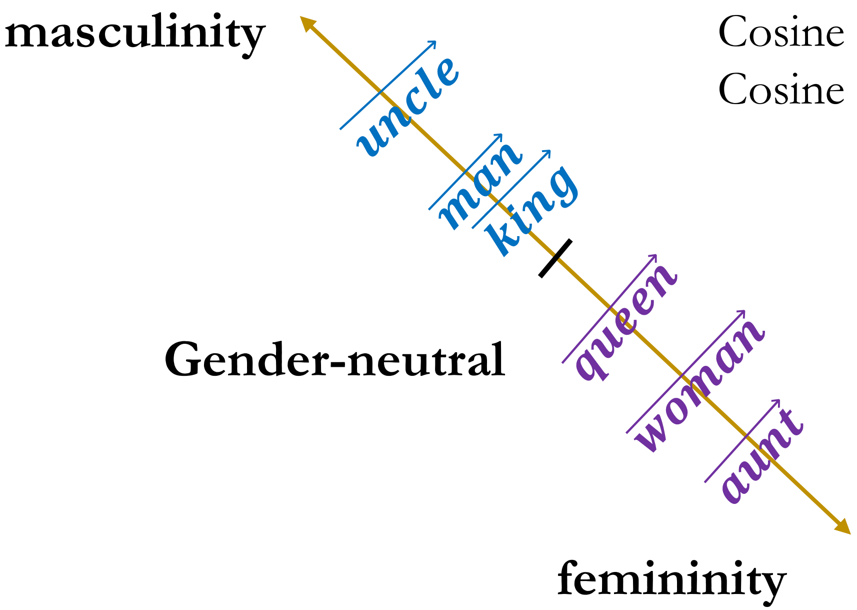


$\overrightarrow{woman}$ - $\overrightarrow{man}$ = $\vec{g}$

$\vec{g}$ is a vector representing the dimension of gender

Gender in Word Embeddings

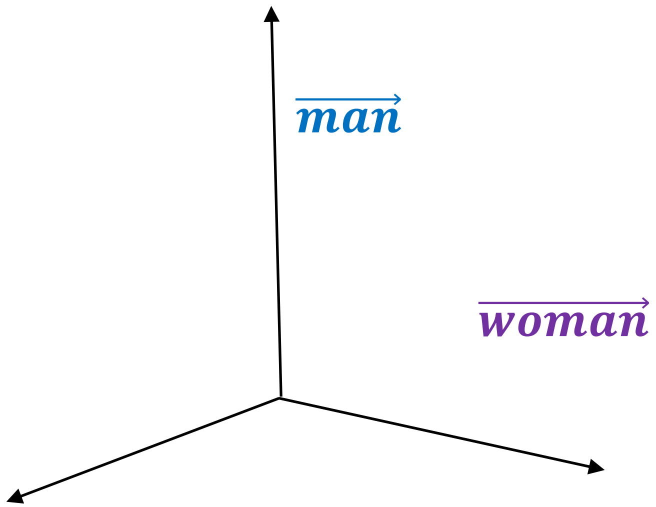




Cosine sim ($\vec{g}$, $\overrightarrow{woman}$) = .2
Cosine sim ($\vec{g}$, $\overrightarrow{man}$) = - .2

$\vec{g}$ is a vector representing the dimension of gender

Extracting a Gender Dimension



“woman” and “man” share a lot of (latent) meaning
but the biggest difference is gender

woman – man =

= (adult, femininity, mammal) – (adult, masculinity, mammal)

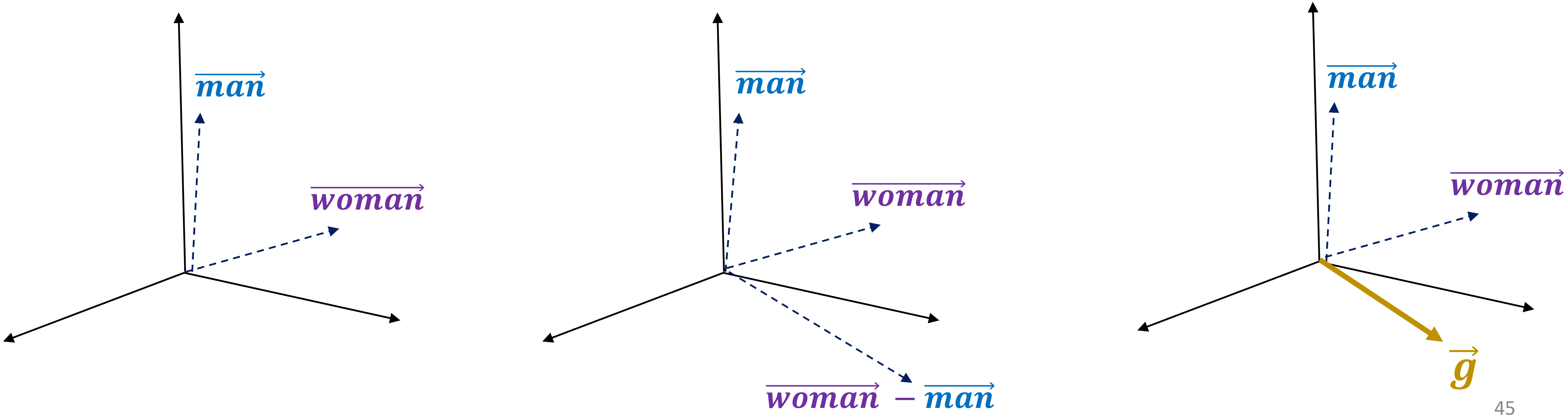
= (~~adult~~, femininity, ~~mammal~~) – (~~adult~~, masculinity, ~~mammal~~)

= gender

“man” and “woman” share a lot of latent meaning - both adults, humans- but the biggest difference is gender

$\overrightarrow{woman}$ - $\overrightarrow{man}$ = $\vec{g}$

$\vec{g}$ is a vector representing the dimension of gender



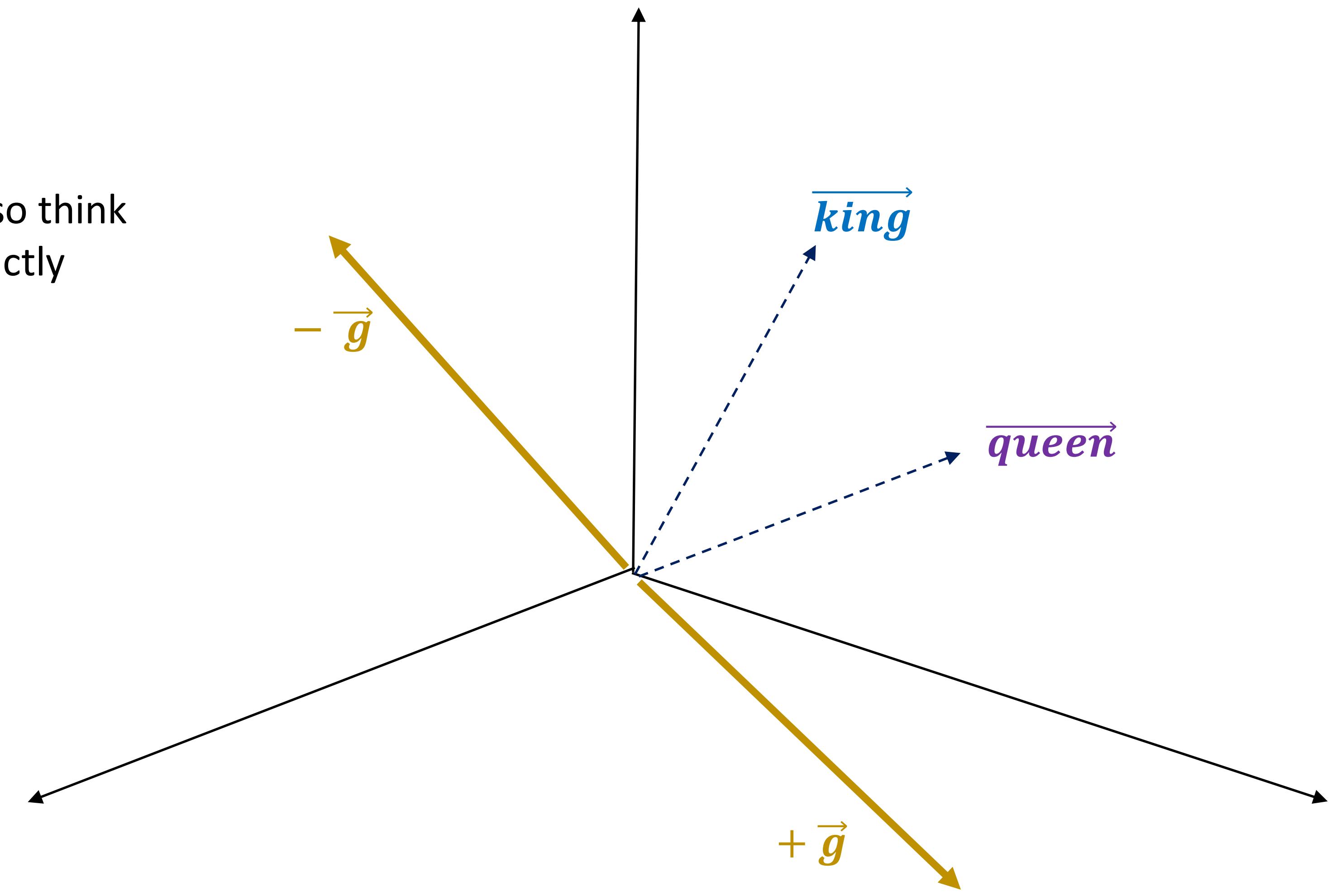
$\vec{g}$ is a direction, but we can also think about $-\vec{g}$, which points in exactly the opposite direction as $\vec{g}$

$\overrightarrow{woman} - \overrightarrow{man} = \vec{g}$

$-(\overrightarrow{woman} - \overrightarrow{man}) = -\vec{g}$

$-\overrightarrow{woman} + \overrightarrow{man} = -\vec{g}$

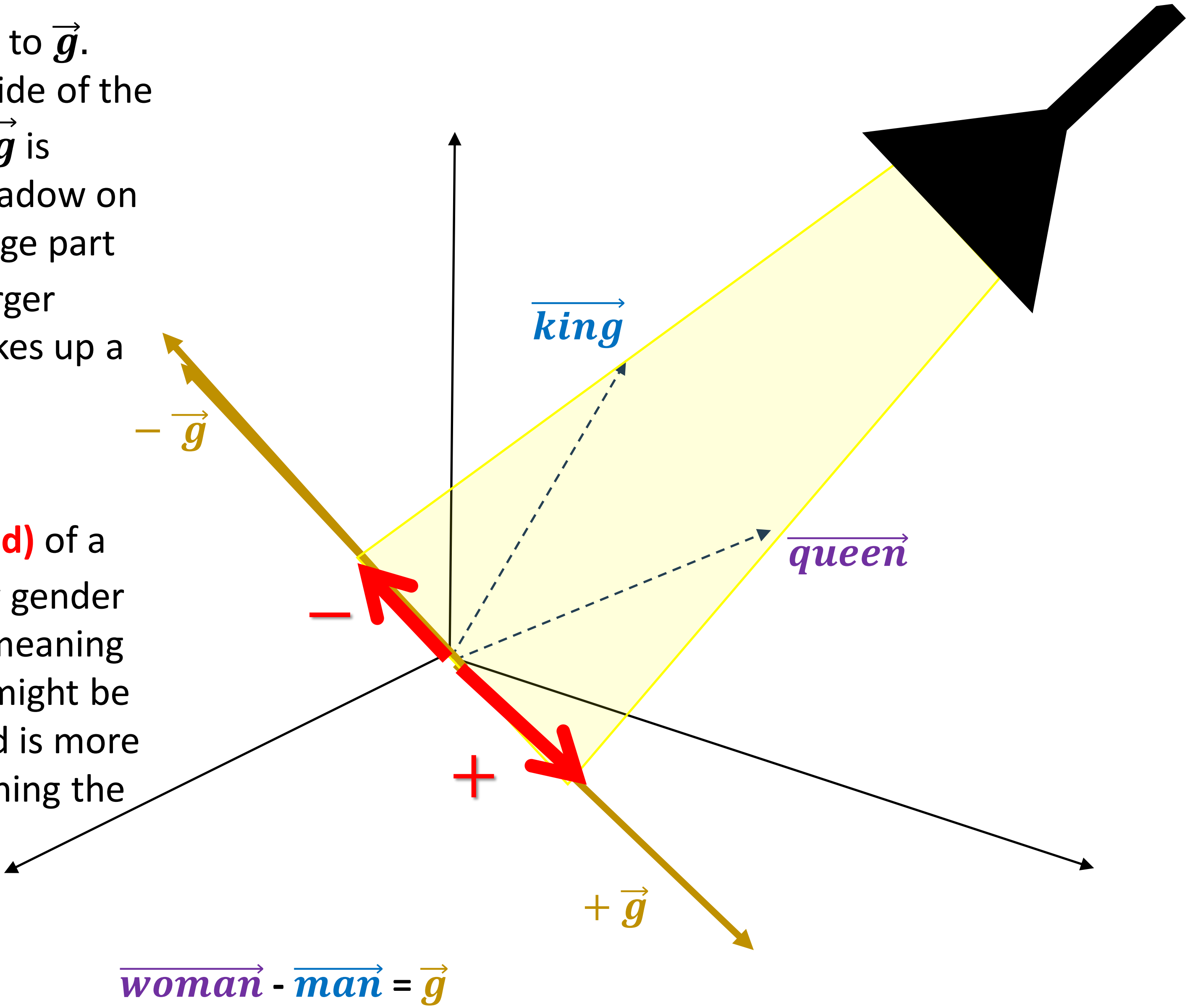
$\overrightarrow{man} - \overrightarrow{woman} = -\vec{g}$



$\vec{g}$ is a vector representing the dimension of gender

Imagine shining a flashlight perpendicular to $\vec{g}$. The size of the **shadow (red)**, and which side of the flashlight the shadow is, tells us how $\overrightarrow{king}$ is gendered. In this vector space, a larger shadow on - $\vec{g}$ means that masculinity makes up a large part of the meaning of $\overrightarrow{king}$. Meanwhile, a larger shadow on + $\vec{g}$ means that femininity makes up a large part of the word-vector.

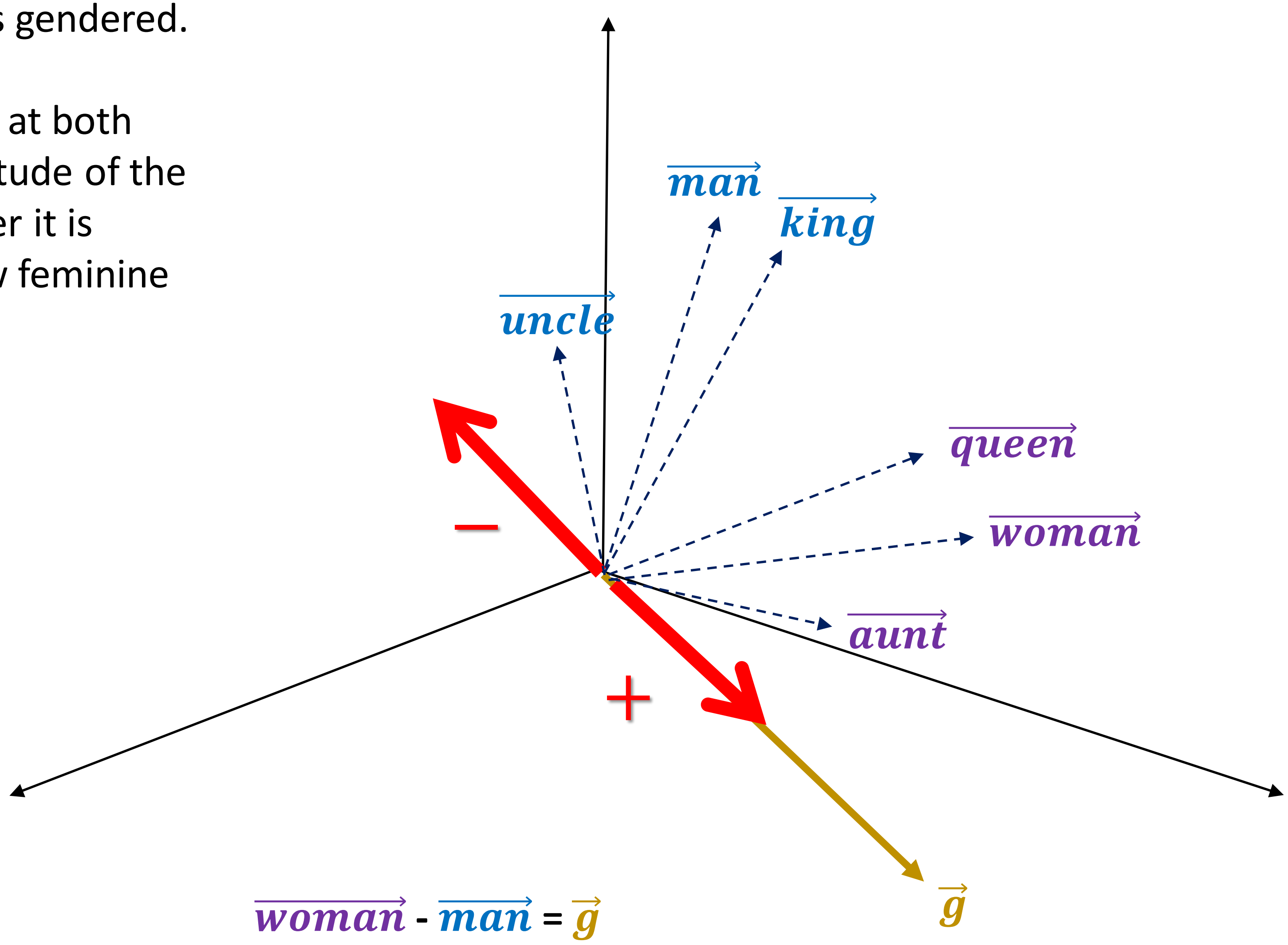
More technically, the scalar **projection (red)** of a word-vector like $\overrightarrow{king}$ onto $\vec{g}$ tell us how gender is a component of the word, or, how the meaning of king is made up by gender. The result might be a large, negative scalar (meaning the word is more masculine) or a large positive scalar (meaning the word is more feminine).



$\vec{g}$ is a vector representing the dimension of gender

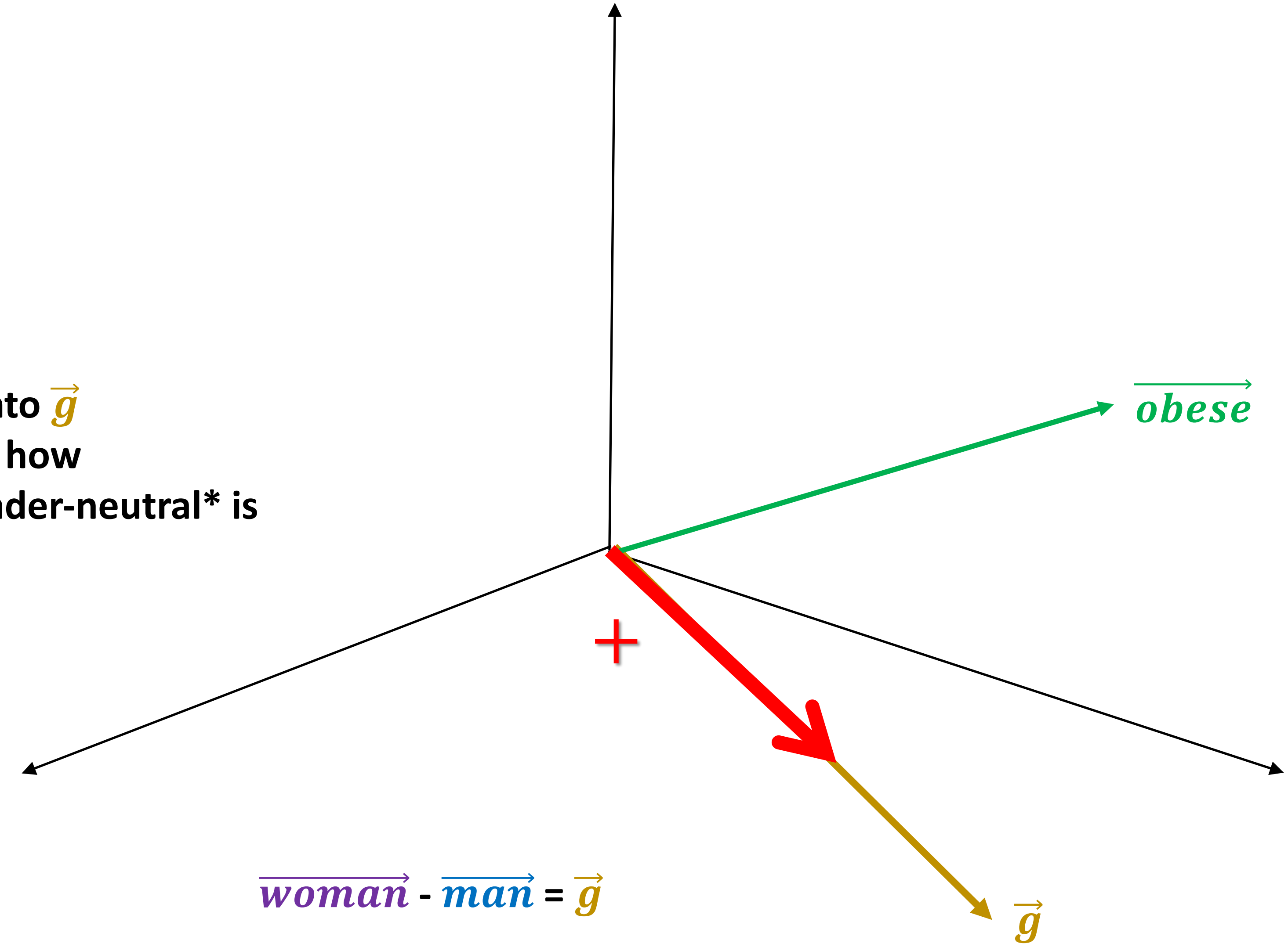
Now, we can look at the **projection (red)** of any word-vector to see how it is gendered.

As mentioned earlier, we'll look at both direction (+ or -) and the magnitude of the projection to determine whether it is feminine or masculine, and how feminine or masculine.



$\vec{g}$ is a vector representing the dimension of gender

Project $\overrightarrow{obese}$ onto $\vec{g}$
To get a value for how
masculine, feminine, or gender-neutral* is
$\overrightarrow{obese}$



$\vec{g}$ is a vector representing the dimension of gender