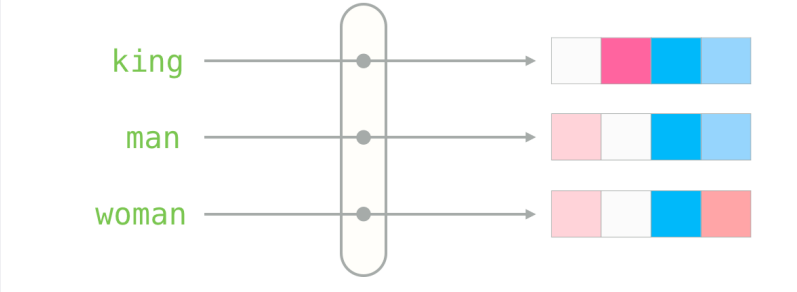
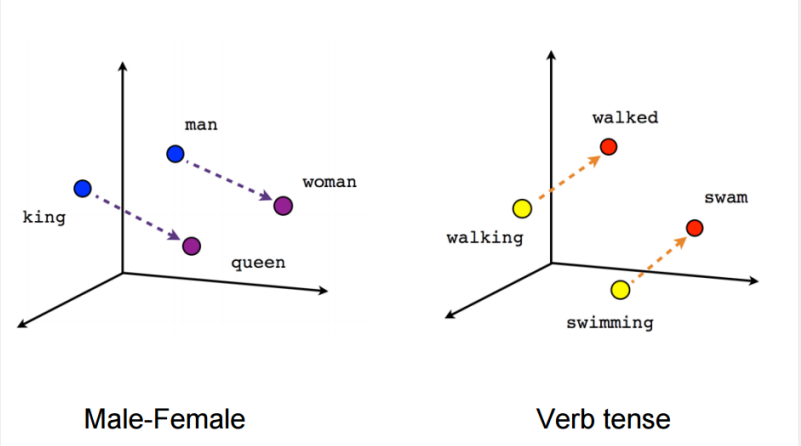
Removing Bias Example

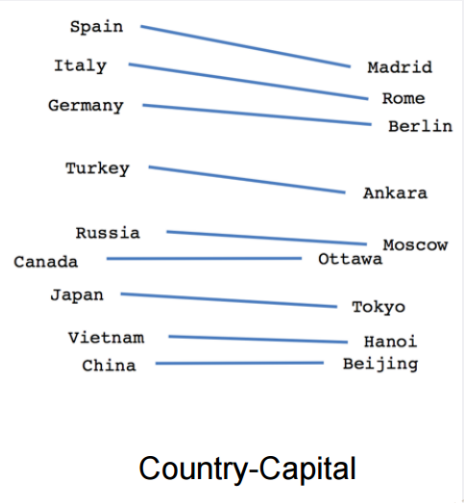
Natural Language Processing

Abuses **Bias** and Blessings of Data

Overview of Word2Vec







Vocabulary generation

How do you input a sentence like:

The quick brown fox jumps over the lazy dog

Into a learning model?

Vocabulary generation

The quick brown fox jumps over the lazy dog

One way could be to encode each character in the words to it's ascii code. For example:

quick = [113, 117, 105, 99, 107]

brown = [98, 114, 111, 119, 110]

Vocabulary generation

Word2Vec generates a vocabulary using the training set

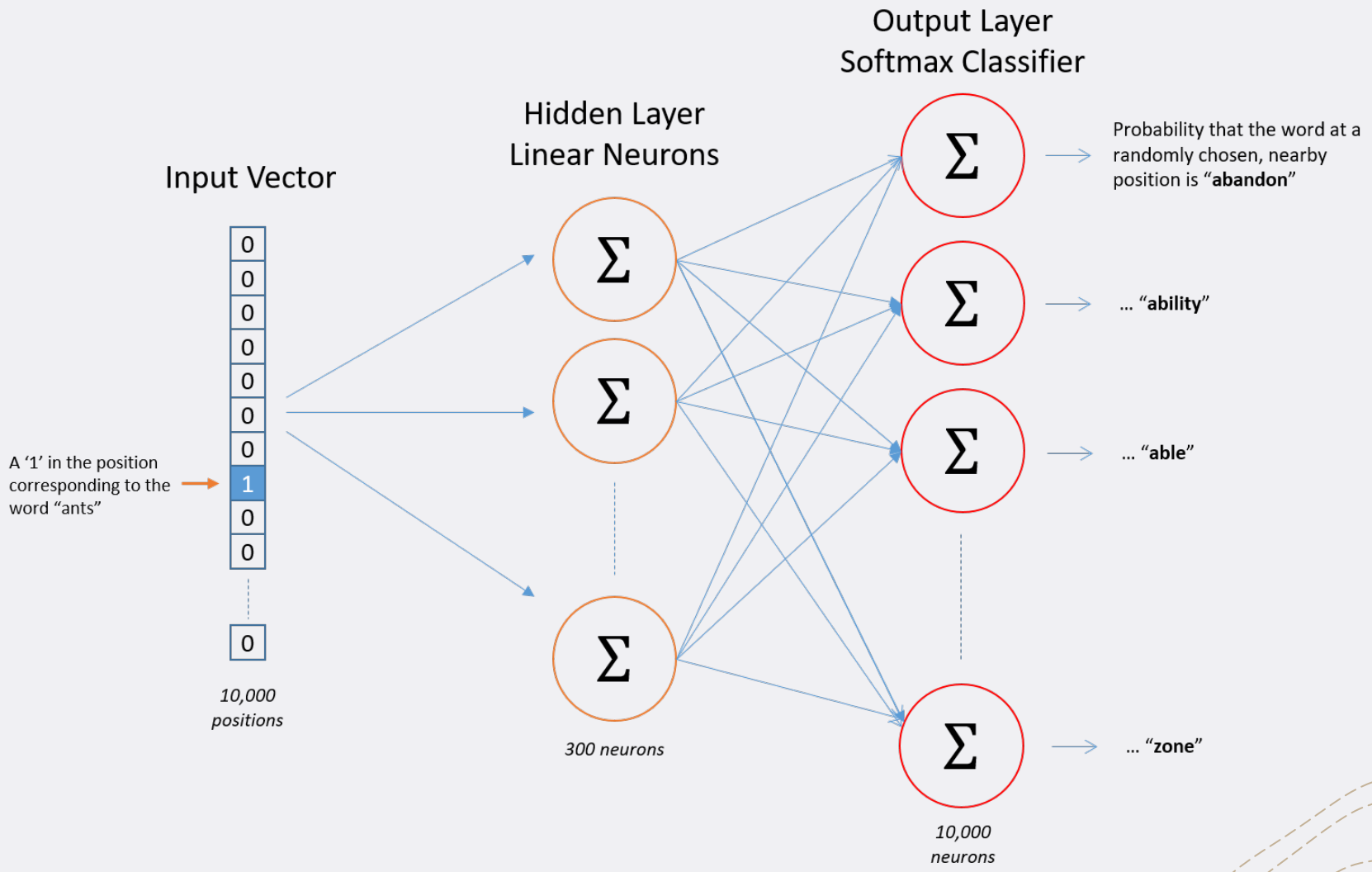
Every unique word is in the vocabulary.

Each word is assigned a unique number, ranging from 0 to the size of the vocabulary.

The input to the model matches the size of the vocabulary.

Some implementations also remove linking and common words (the, and, this etc.) from the vocabulary to reduce its size.

Word2Vec architecture

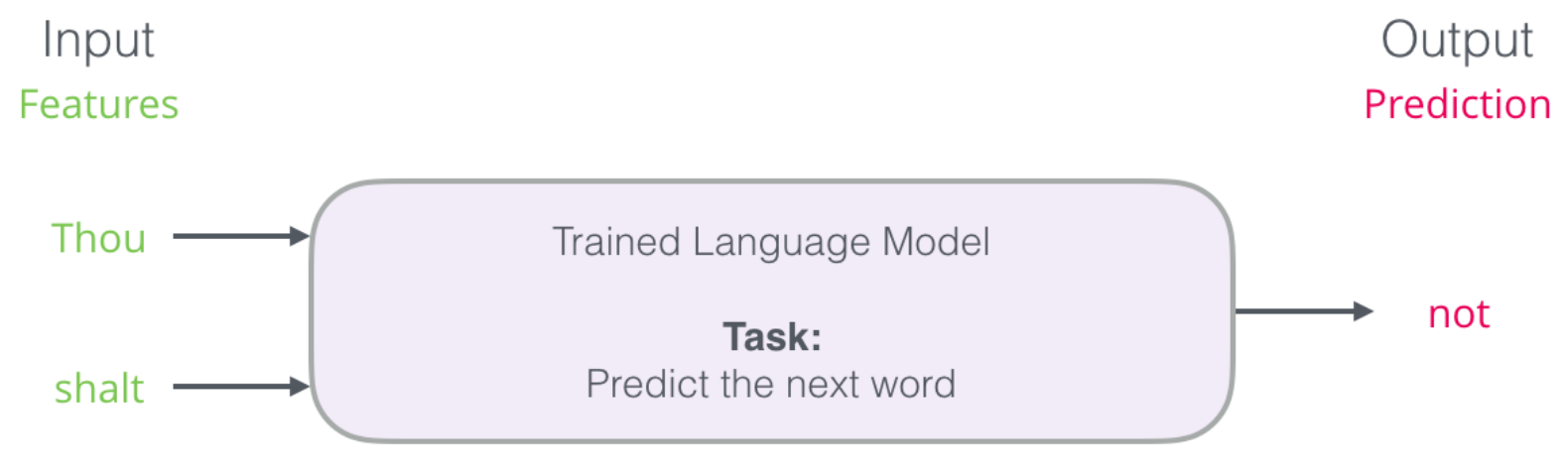


Training Word2Vec Model

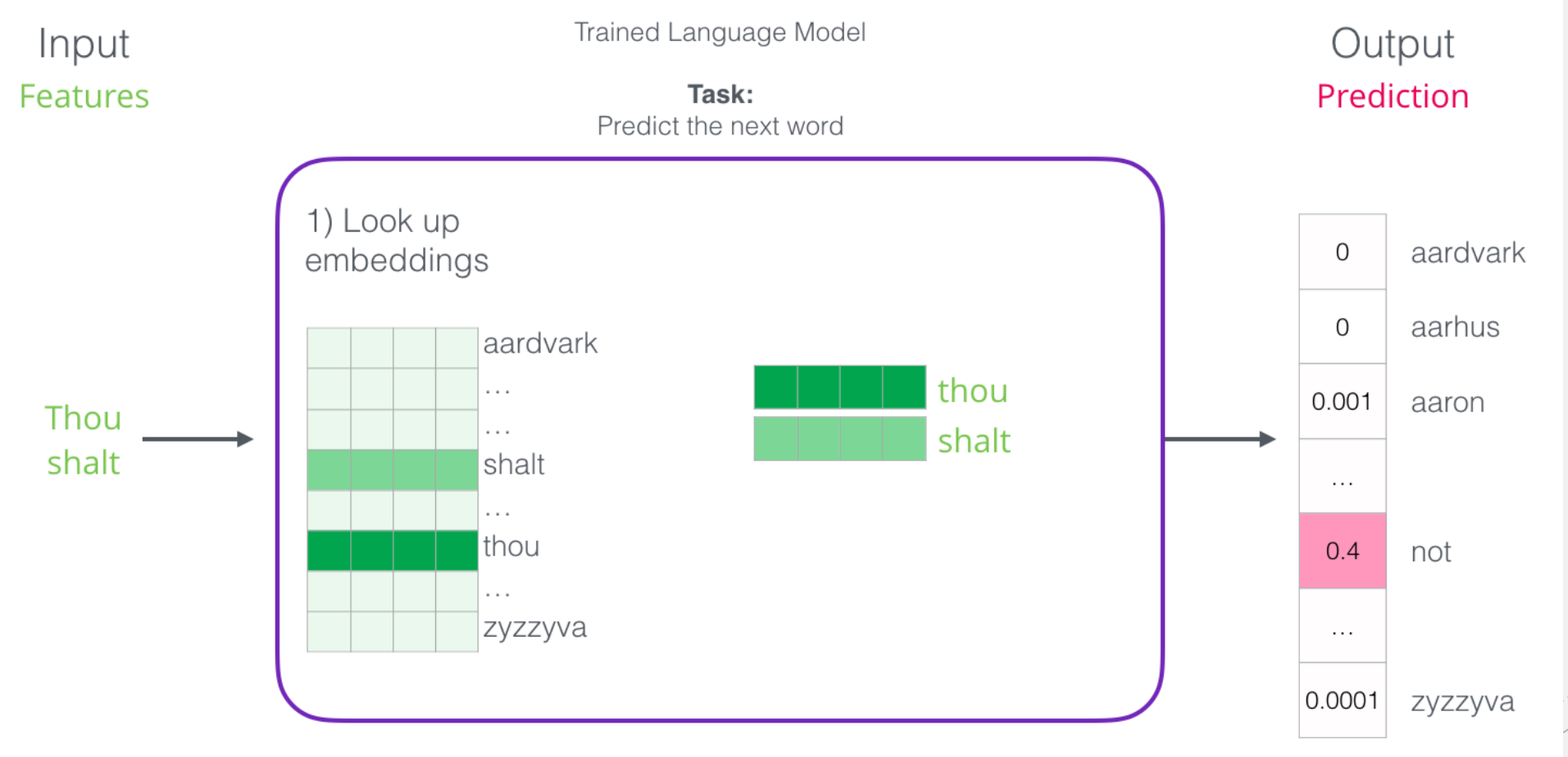
J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in EMNLP 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014.

Word2Vec training

input/feature #1	input/feature #2	output/label
Thou	shalt	____



Word2Vec training



Word2Vec training

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	

Dataset

input 1	input 2	output
thou	shalt	not
shalt	not	make
not	make	a
make	a	machine
a	machine	in

Word2Vec training

To get to university today I caught the _____

Word2Vec training

To get to university today I caught the _____ bus

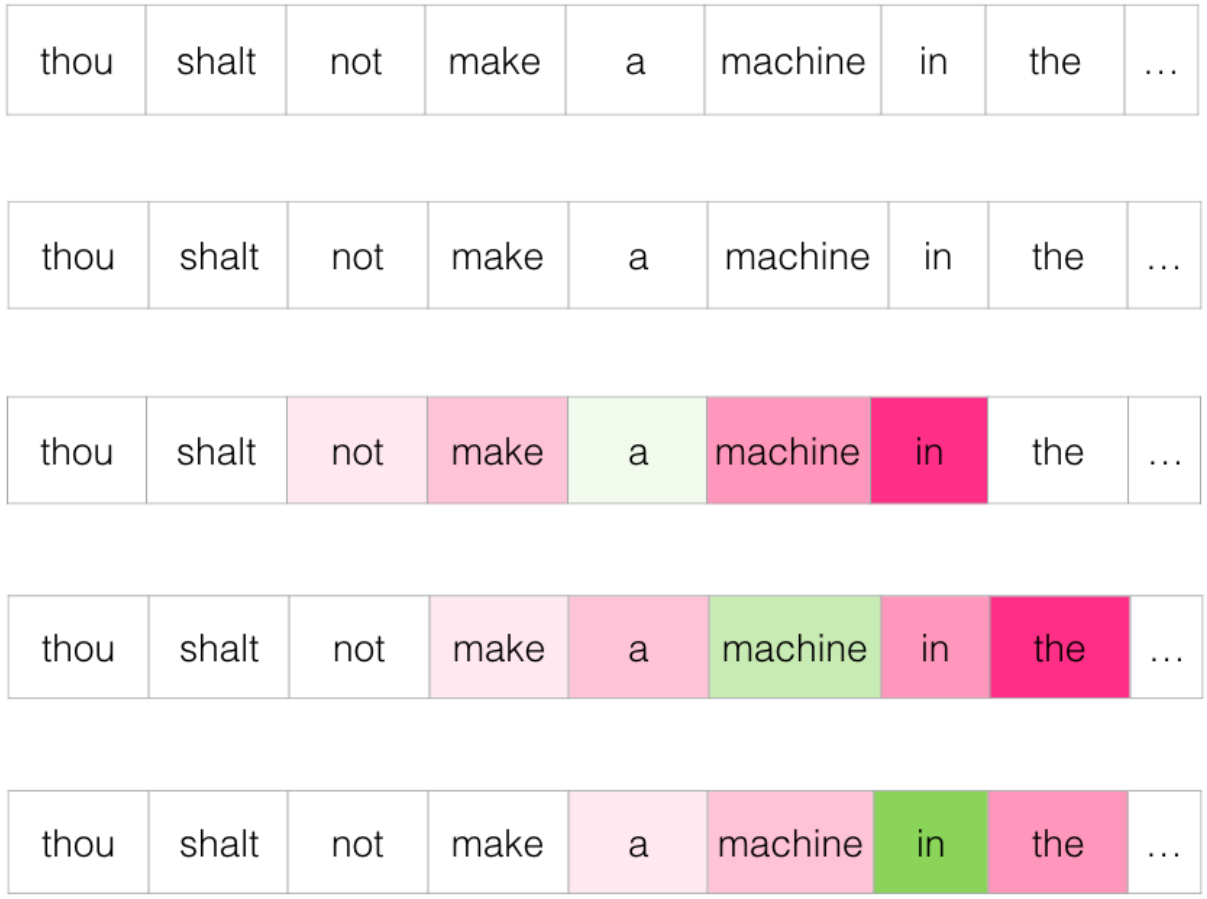
Word2Vec training: Skipgram

Thou shalt not make a machine in the likeness of a human mind

input word	target word
not	thou
not	shalt
not	make
not	a

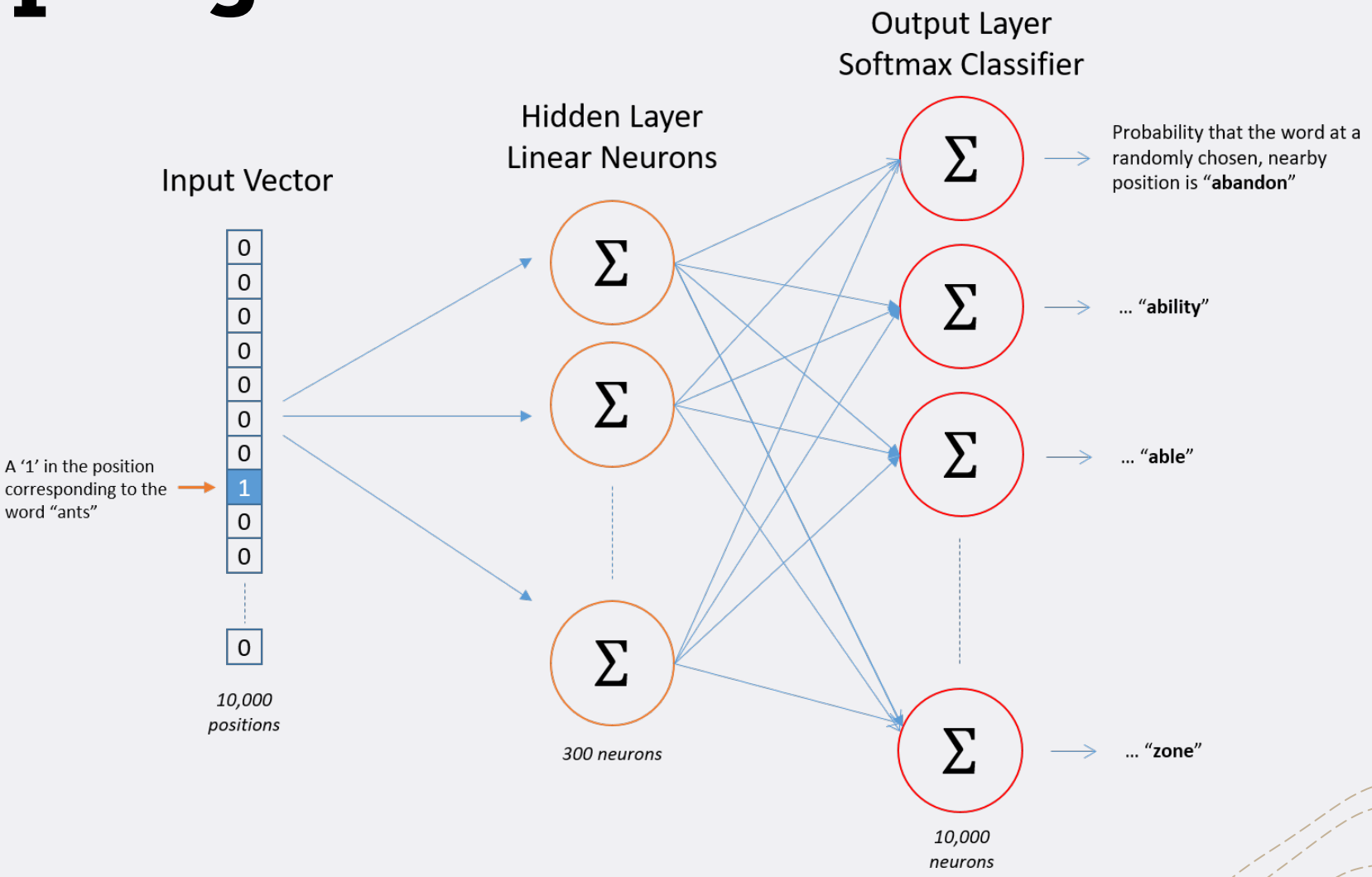
Word2Vec training: Skipgram

Thou shalt not make a machine in the likeness of a human mind

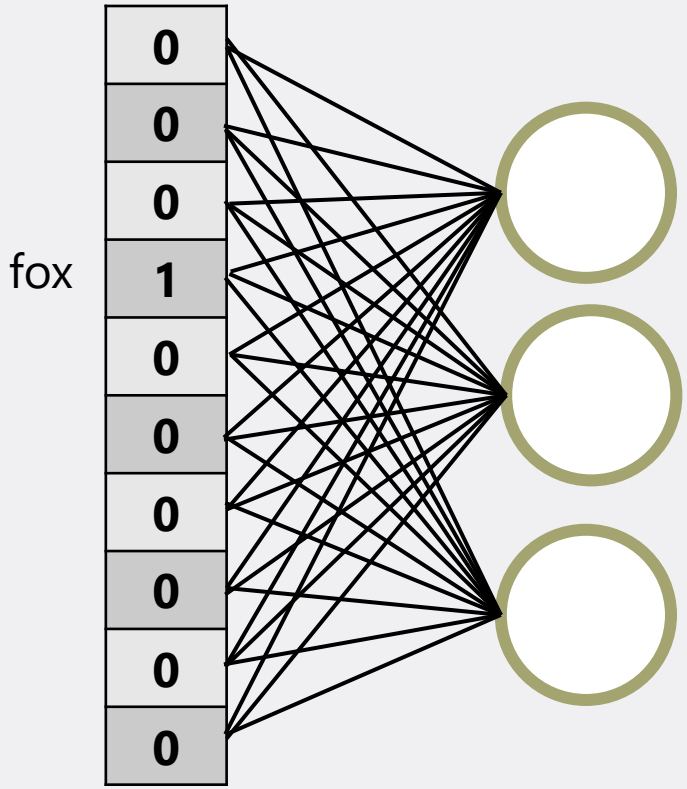


input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine
a	not
a	make
a	machine
a	in
machine	make
machine	a
machine	in
machine	the
in	a
in	machine
in	the
in	likeness

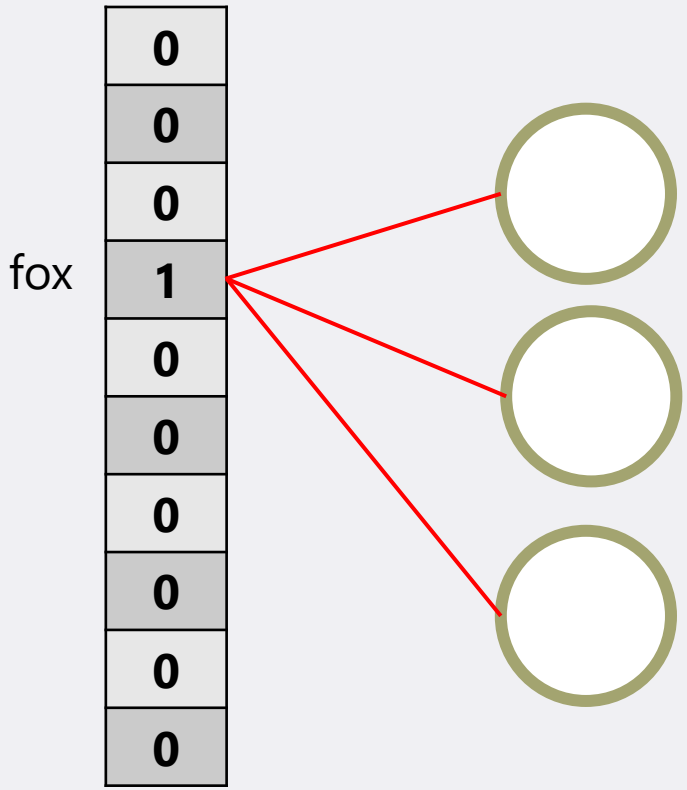
Word2Vec training: Negative Sampling



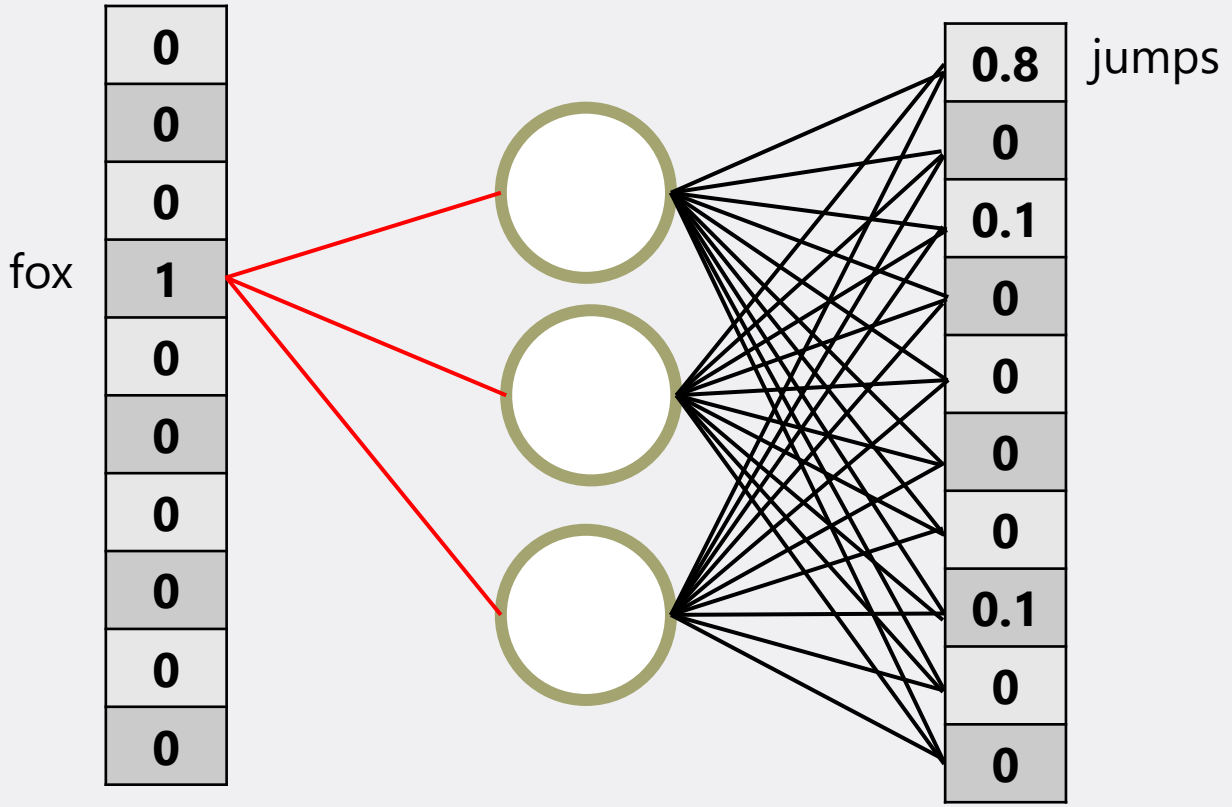
Word2Vec training: Negative Sampling



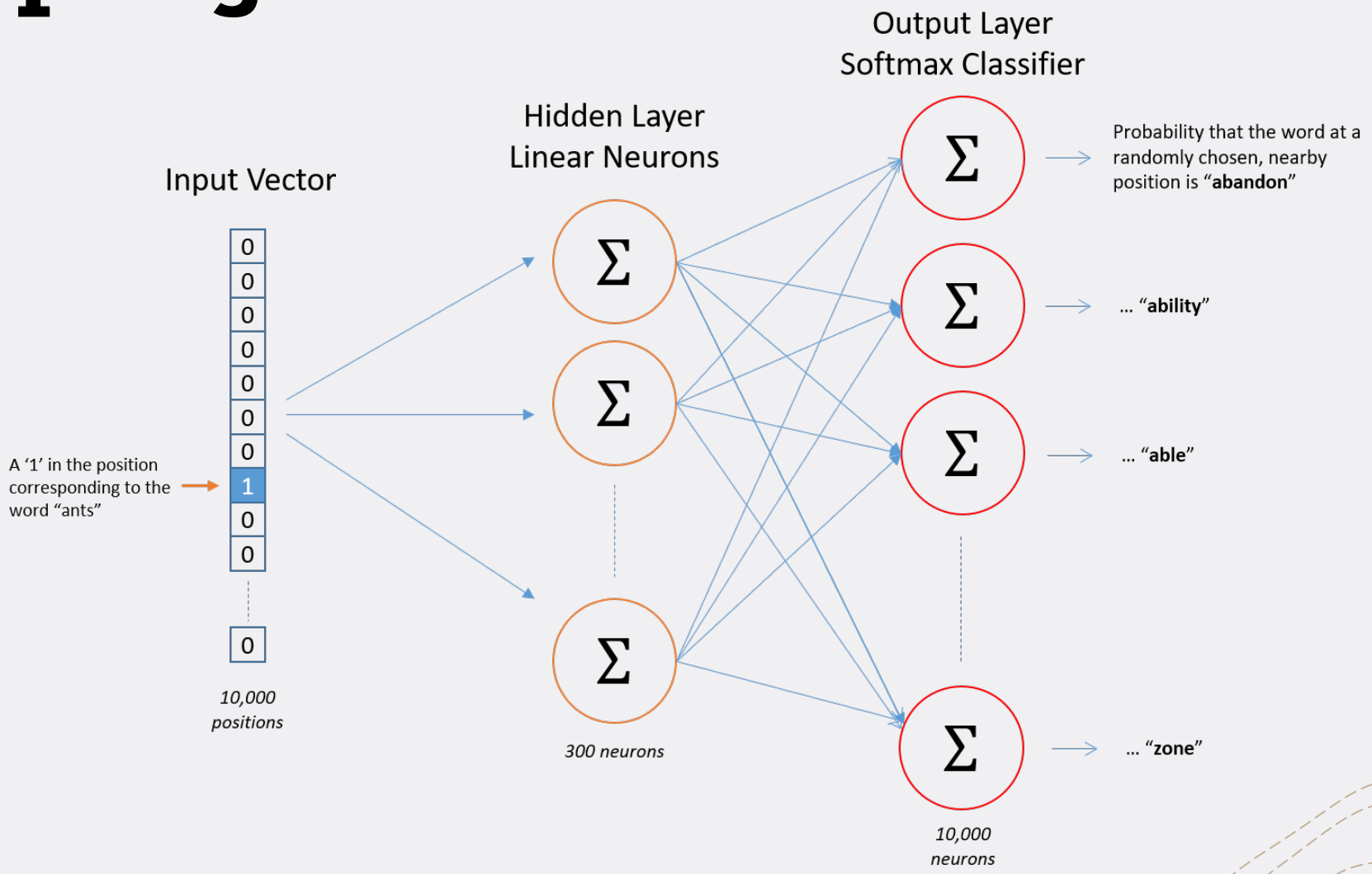
Word2Vec training: Negative Sampling



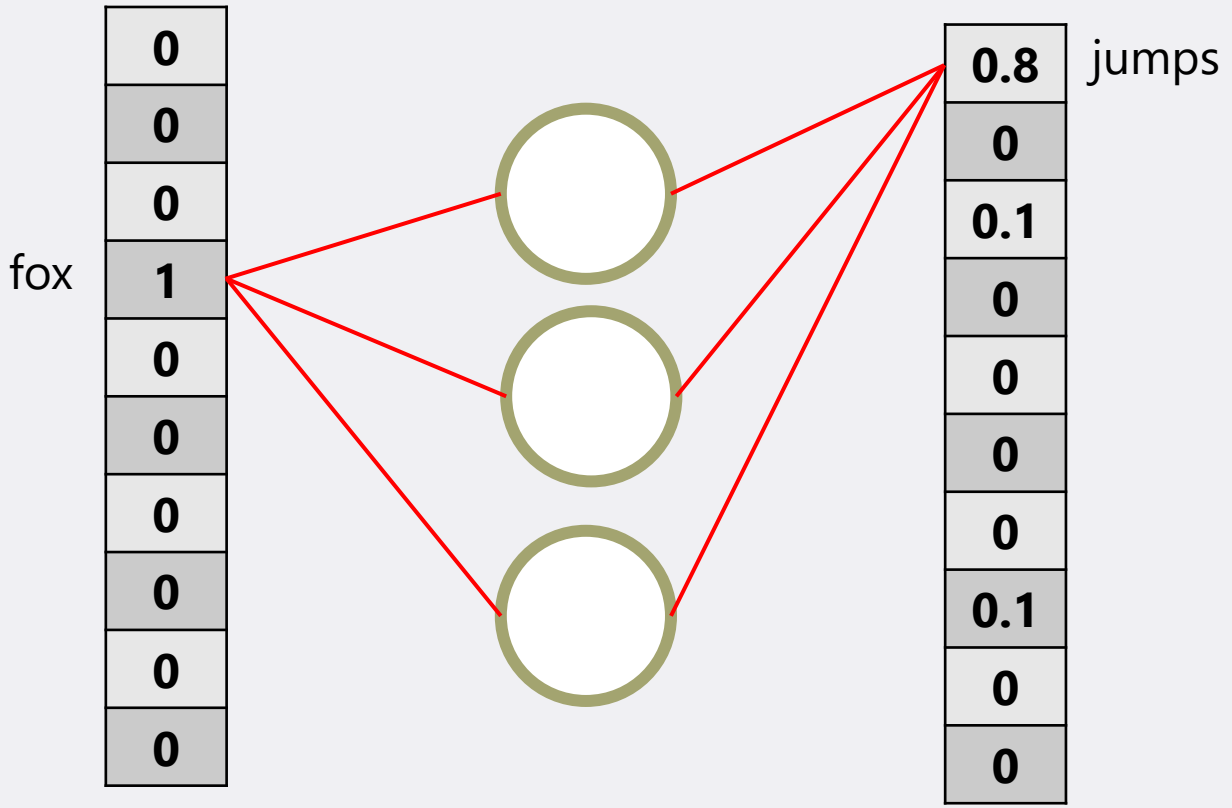
Word2Vec training: Negative Sampling



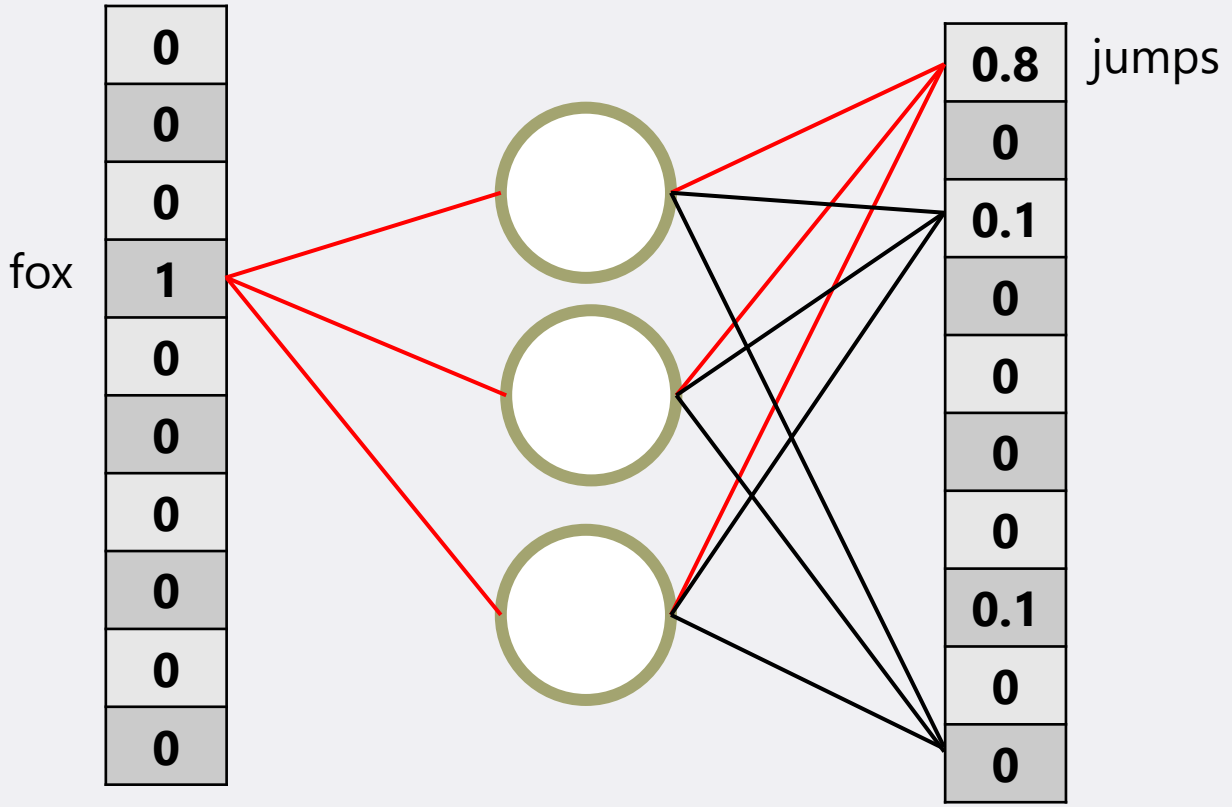
Word2Vec training: Negative Sampling



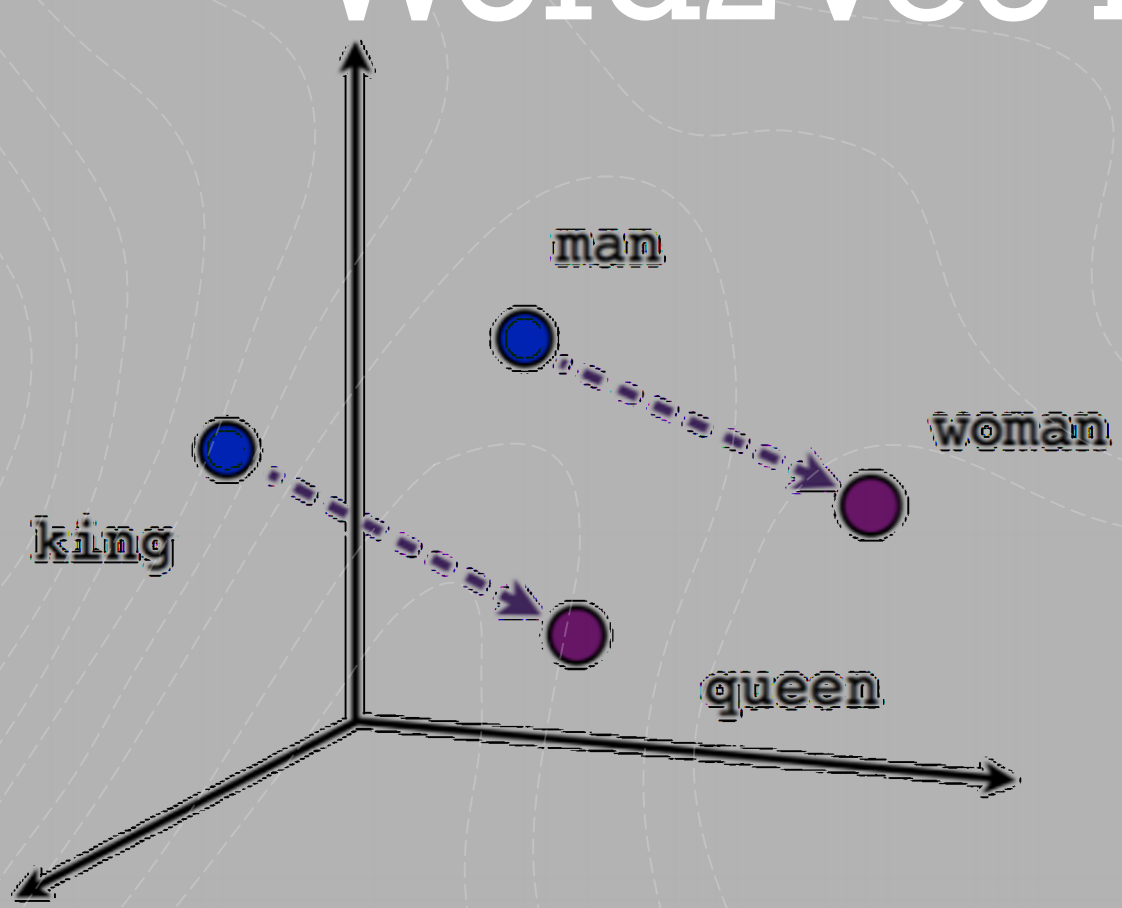
Word2Vec training: Negative Sampling



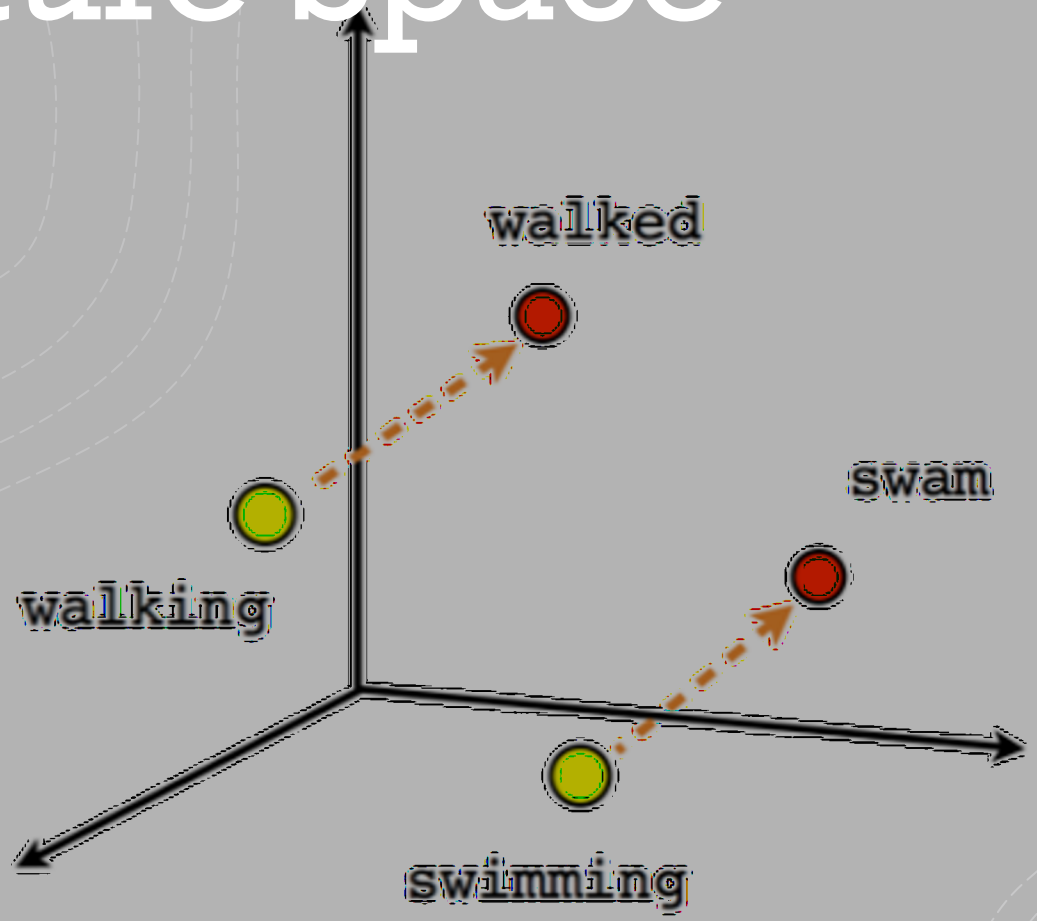
Word2Vec training: Negative Sampling



Word2Vec Feature Space



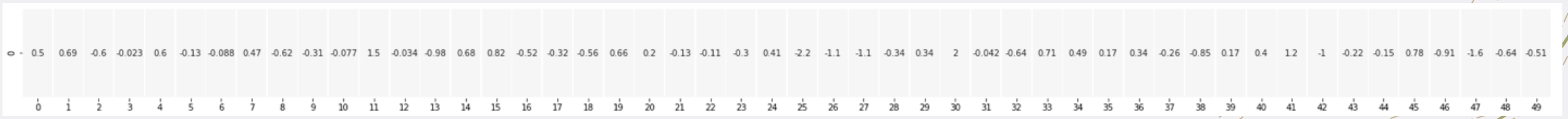
Male-Female



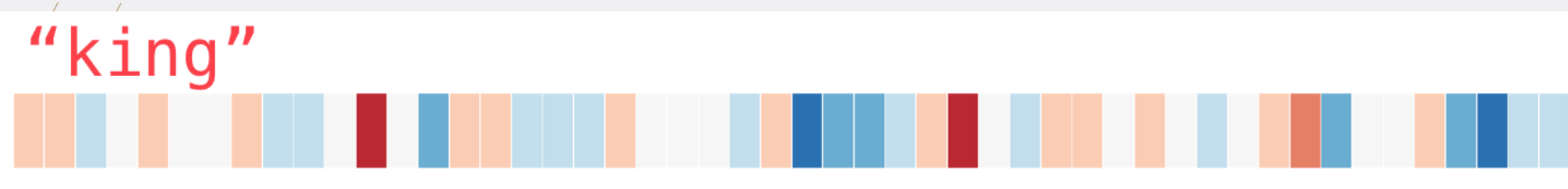
Verb tense

Word2Vec embeddings

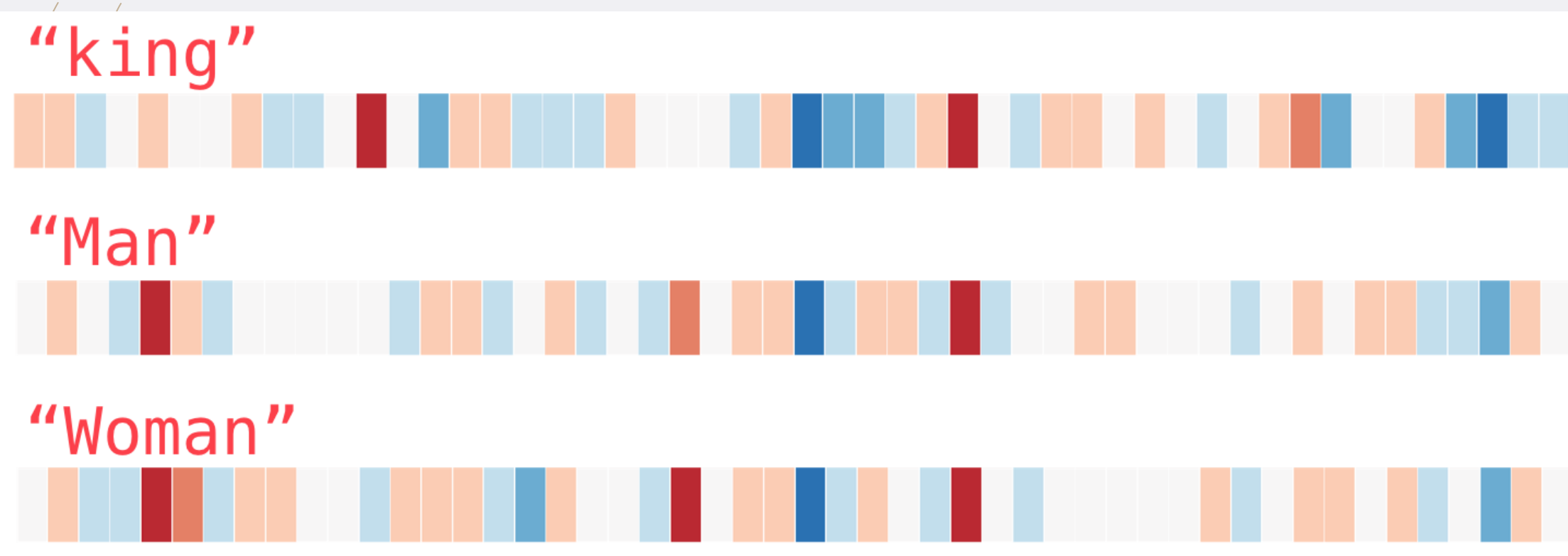
'king' = [0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -0.31012 , -0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961 , -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354 , 0.33505 , 1.9927 , -0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 , -1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042]



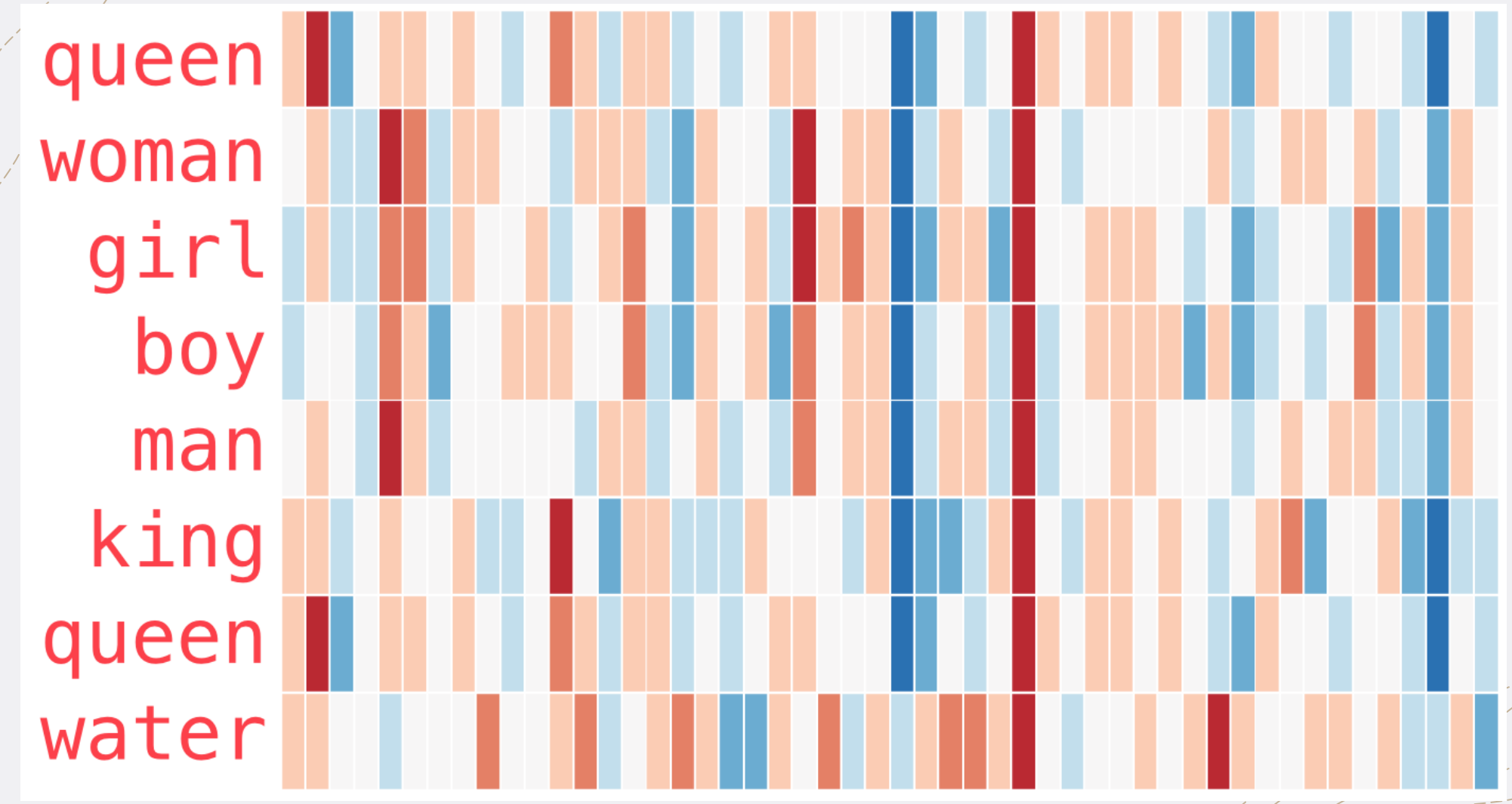
Word2Vec embeddings



Word2Vec embeddings



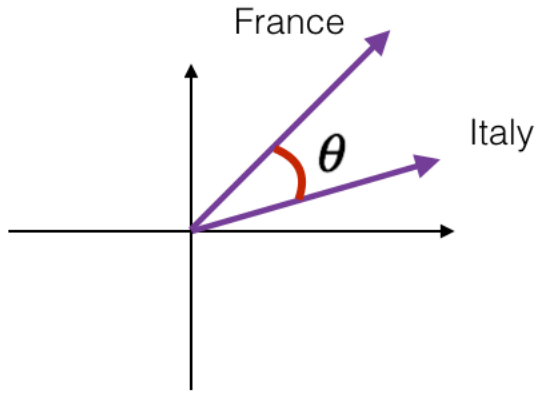
Word2Vec embeddings



Word2Vec embeddings

$$CosineSimilarity(u,v) = \frac{u.v}{||u||_2||v||_2} = \cos(\theta)$$

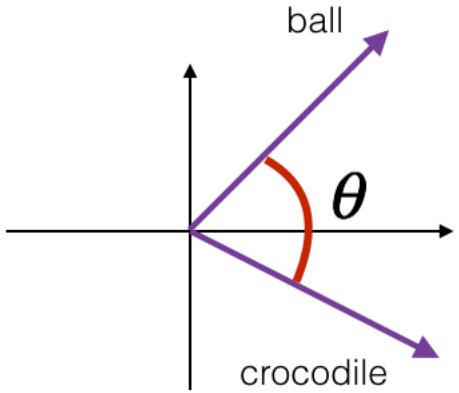
$$\|u\|_2 = \sqrt{\sum_{i=1}^{n} u_i^2}$$



France and Italy are quite similar

θ is close to 0°

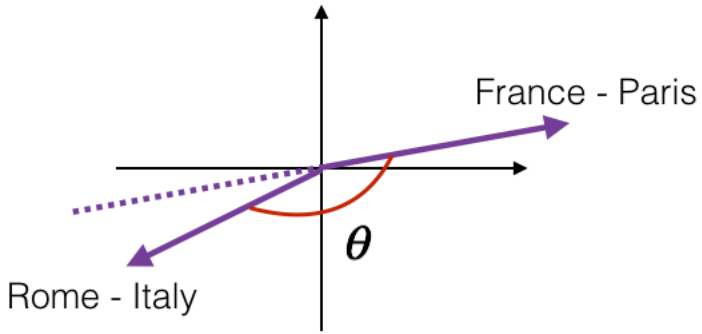
$\cos(\theta) \approx 1$



ball and crocodile are not similar

θ is close to 90°

$\cos(\theta) \approx 0$



the two vectors are similar but opposite
the first one encodes (city - country)
while the second one encodes (country - city)

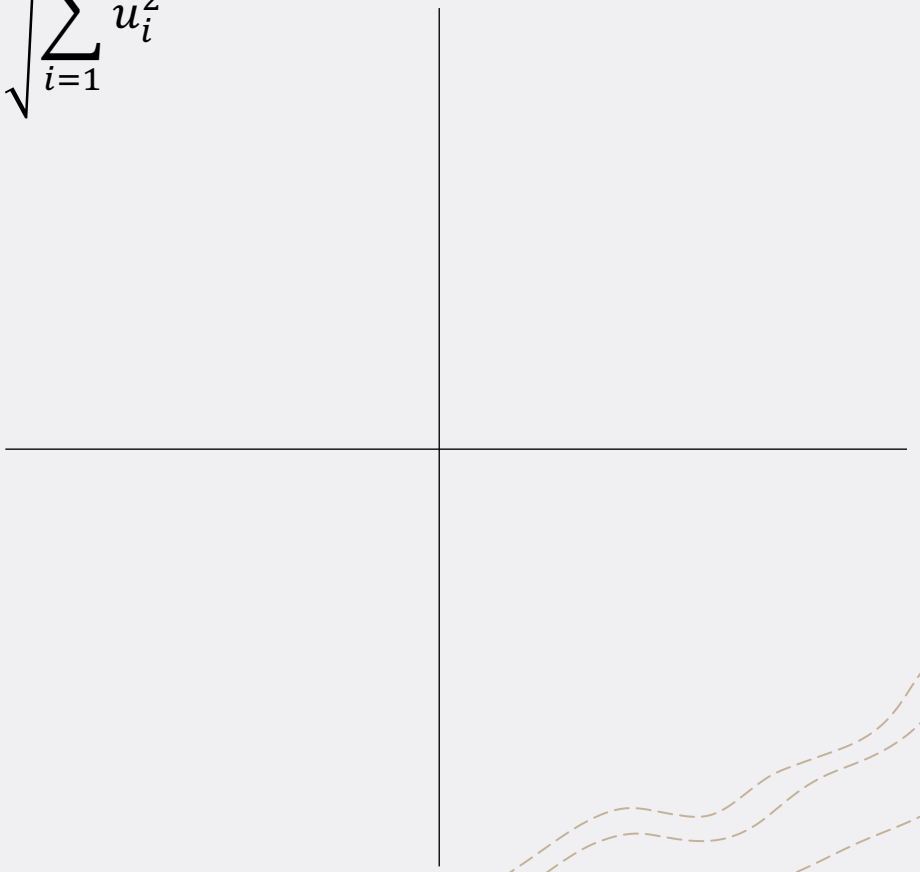
θ is close to 180°

$\cos(\theta) \approx -1$

Cosine Similarity: Worked example

$$CosineSimilarity(u, v) = \frac{u.v}{||u||_2||v||_2} = \cos(\theta)$$

$$\|u\|_2 = \sqrt{\sum_{i=1}^{n} u_i^2}$$



Word2Vec embeddings

```
# look up top 6 words similar to 'polite'
w1 = ["polite"]
model.wv.most_similar (positive=w1,topn=6)
```

```
[('courteous', 0.9174547791481018),
 ('friendly', 0.8309274911880493),
 ('cordial', 0.7990915179252625),
 ('professional', 0.7945970892906189),
 ('attentive', 0.7732747197151184),
 ('gracious', 0.7469891309738159)]
```

Word2Vec embeddings

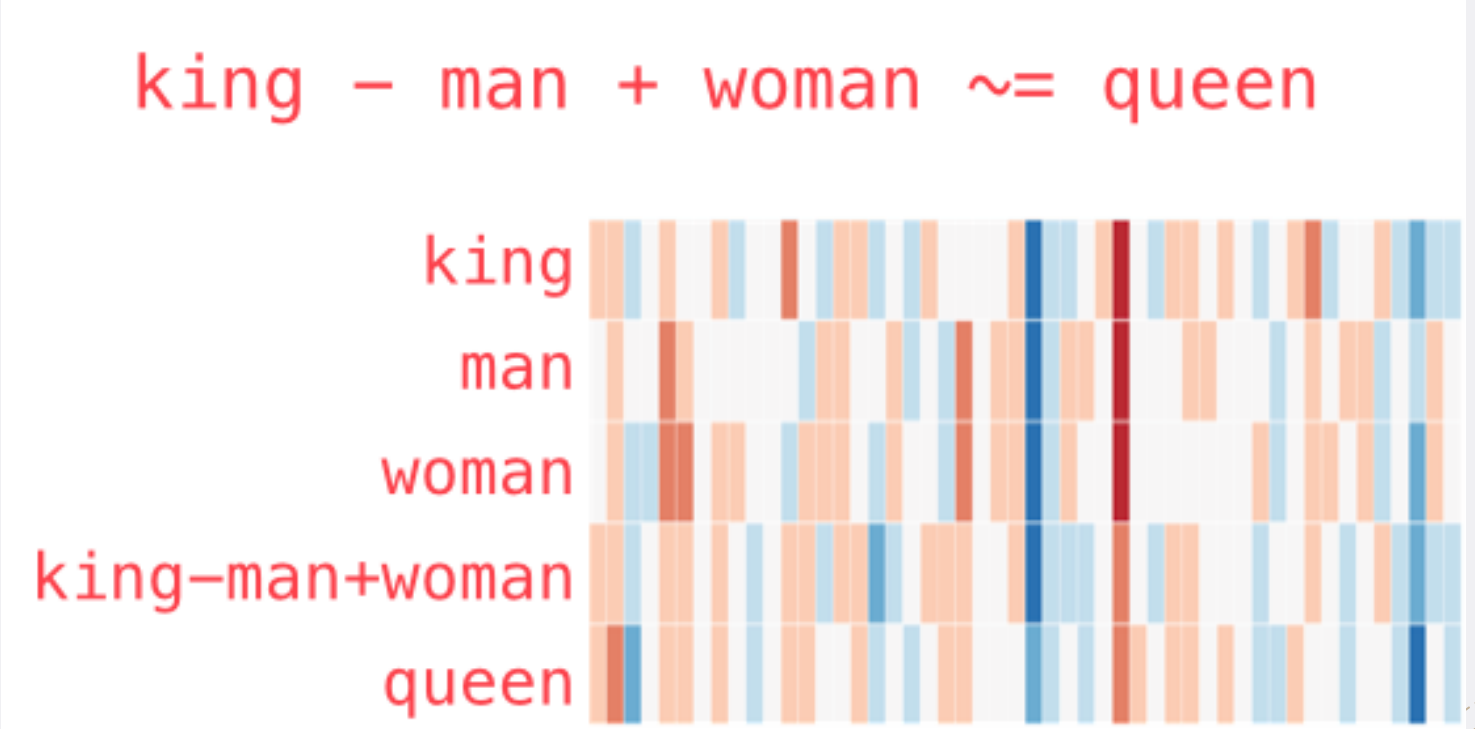
One fun thing we can try is arithmetic

What would the following equal?

```
king - man + woman ~= queen
```

Word2Vec embeddings

Word2Vec embeddings

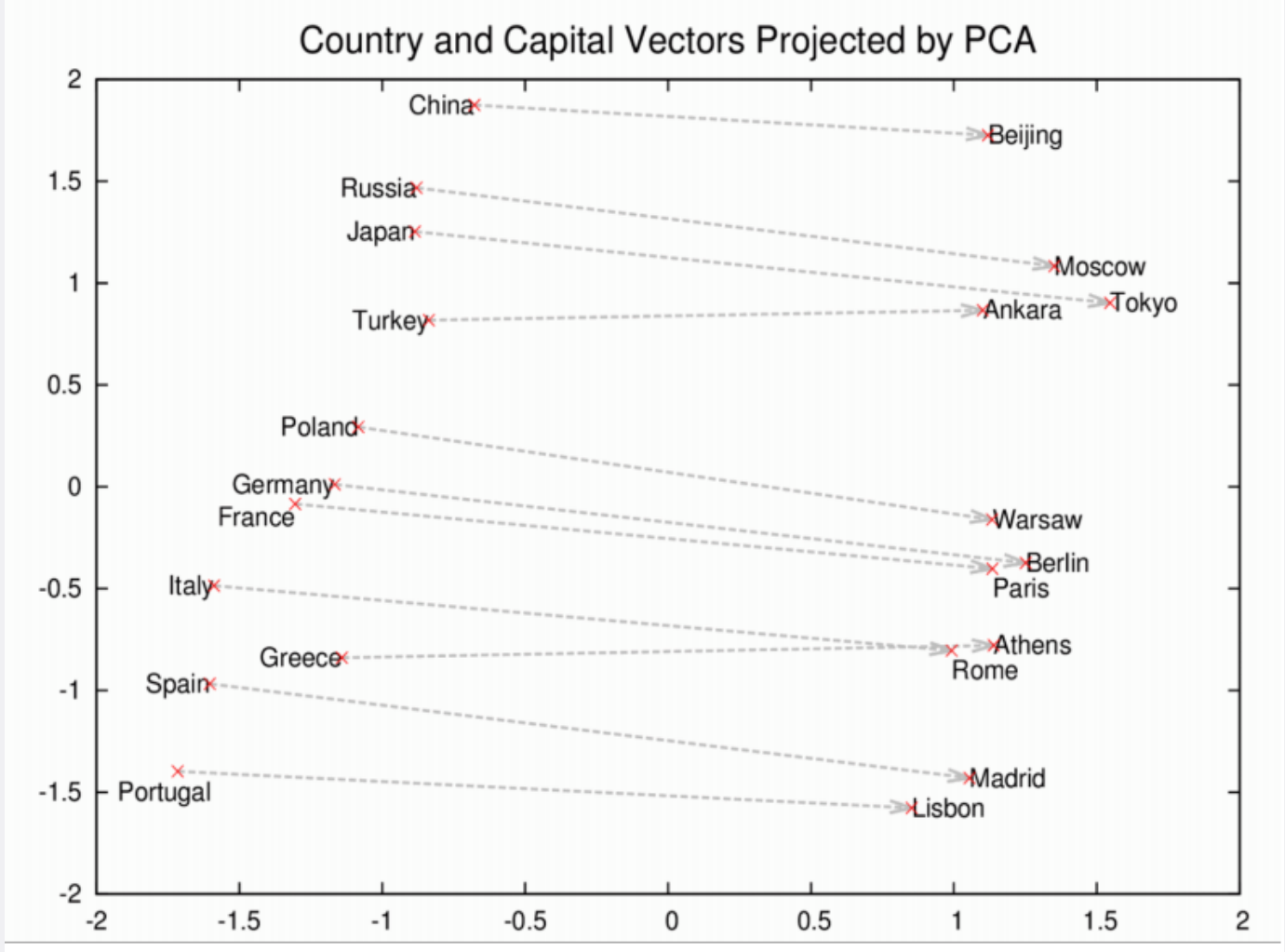


Word2Vec embeddings

```
model.most_similar(positive=["king","woman"], negative=["man"])
```

```
[('queen', 0.8523603677749634),
 ('throne', 0.7664333581924438),
 ('prince', 0.7592144012451172),
 ('daughter', 0.7473883032798767),
 ('elizabeth', 0.7460219860076904),
 ('princess', 0.7424570322036743),
 ('kingdom', 0.7337411642074585),
 ('monarch', 0.7214490771293640),
 ('eldest', 0.7184862494468689),
 ('widow', 0.7099430561065674)]
```

Word2Vec embeddings



Counter factuals and adverse results

- Man -> Computer Programmer
- Women -> ??
- Women + (Man – computer programmer) = ?
- Home maker

Bias in word2vec parameter space

T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, *"Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,"* Jul. 2016.Link to Paper.

Gender Vector

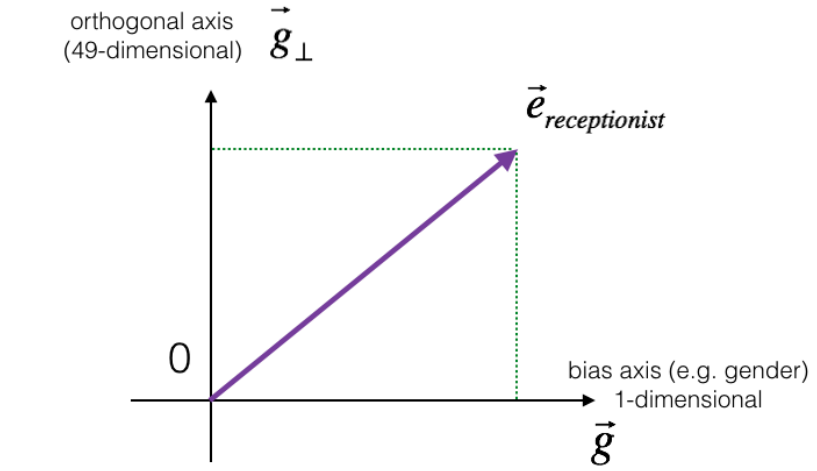
- How can we characterise Gender in the feature space?
  - Gender is the difference between woman and man
  - Repeat and average

$$g = e_{woman} - e_{man}$$

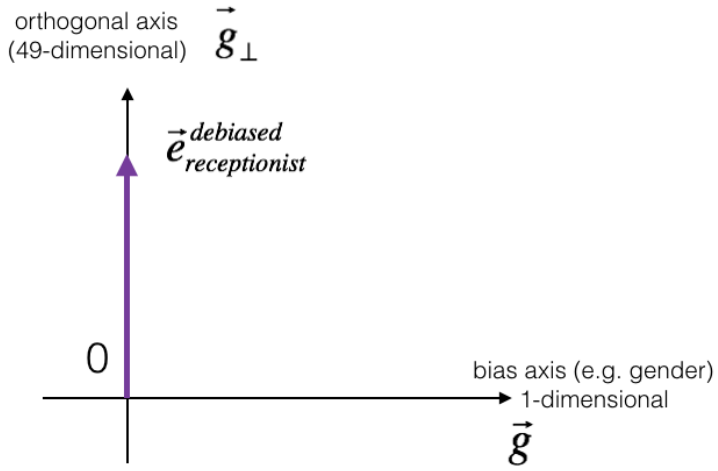
Neutralise

$$e^{bias\_component} = \frac{e.g}{||g||_2^2} * g$$

$$e^{debiased} = e - e^{bias\_componet}$$



before neutralizing,
"receptionist" is positively correlated with the bias axis



after neutralizing,
debased version, with the component
in the direction of the bias axis (g) zeroed out

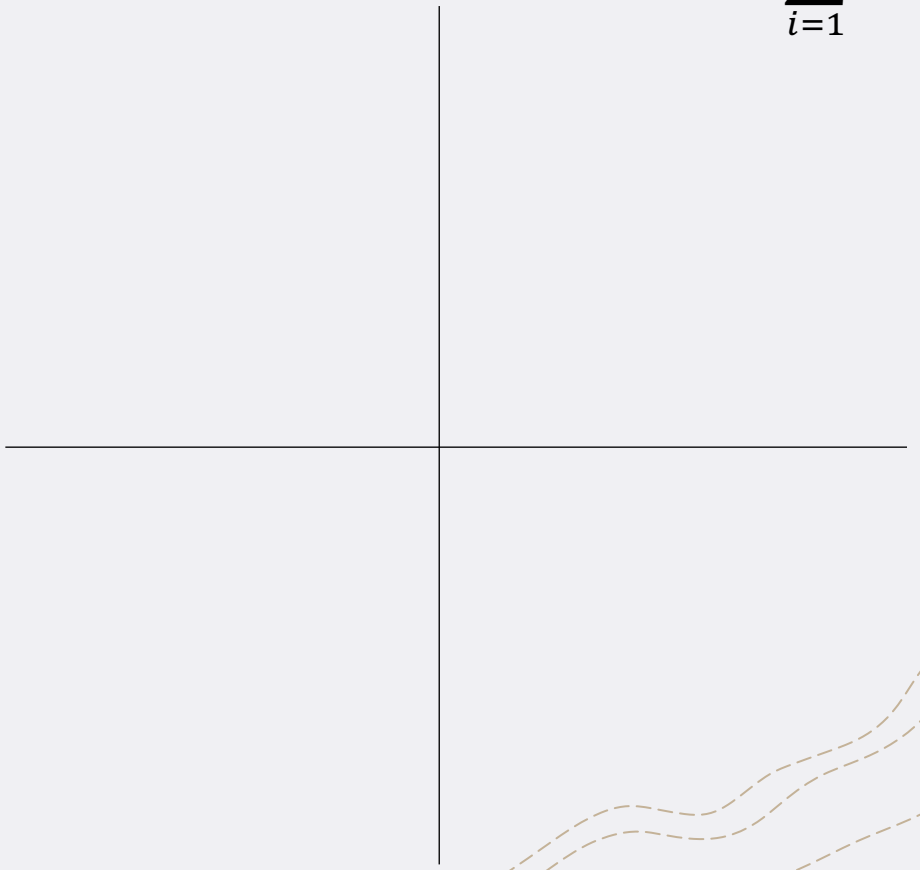
Neutralise Example

$$e^{bias\_component} = \frac{e.g}{||g||_2^2} * g$$

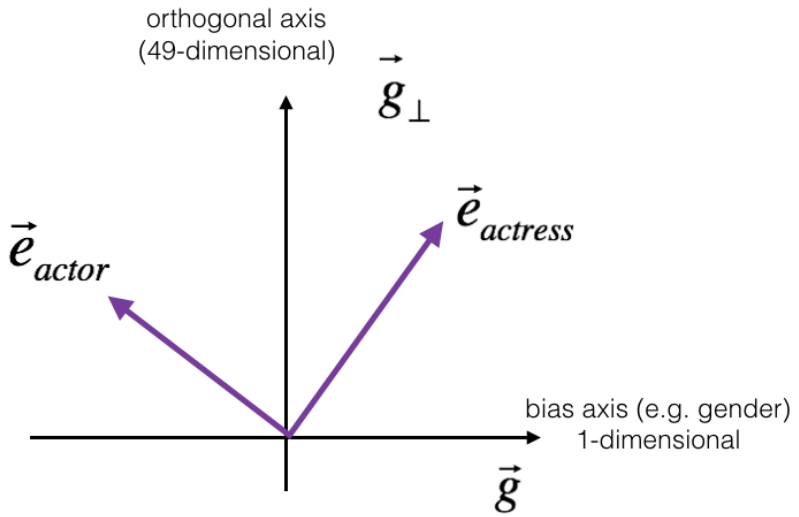
$$e^{debiased} = e - e^{bias\_componet}$$

$$\|g\|_2 = \sqrt{\sum_{i=1}^{n} u_i^2}$$

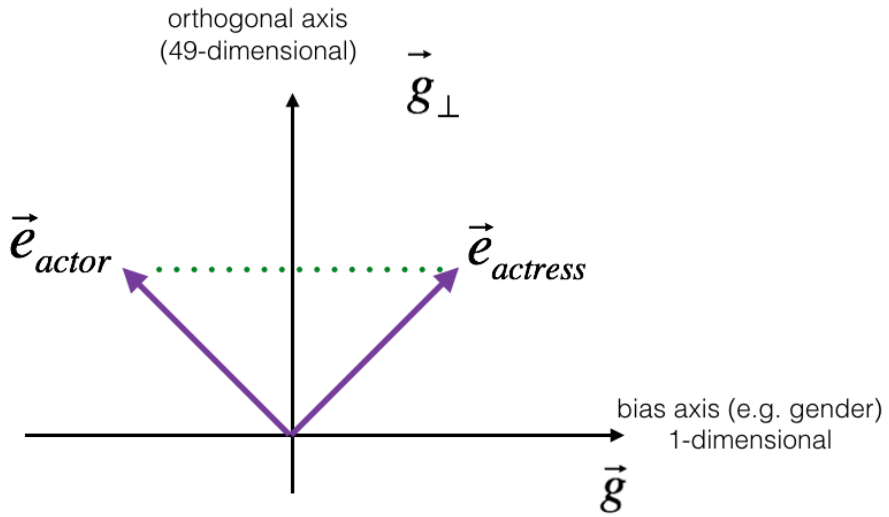
$$||g||_2^2 = \sum_{i=1}^{n} u_i^2$$



Equalise



before equalizing,
"actress" and "actor" differ in many ways beyond the direction of $\vec{g}$



after equalizing,
"actress" and "actor" differ only in the direction of $\vec{g}$, and further are equal in distance from $\vec{g}_\perp$

Equalise Explanation

$$\mu = \frac{e_{w1} + e_{w2}}{2}$$

$$\mu_B = \frac{\mu \cdot \text{bias\_axis}}{\|\text{bias\_axis}\|_2^2} * \text{bias\_axis}$$

$$\mu_\perp = \mu - \mu_B$$

$$e_{w1B} = \frac{e_{w1} \cdot \text{bias\_axis}}{\|\text{bias\_axis}\|_2^2} * \text{bias\_axis}$$

$$e_{w2B} = \frac{e_{w2} \cdot \text{bias\_axis}}{\|\text{bias\_axis}\|_2^2} * \text{bias\_axis}$$

$$e_{w1B}^{corrected} = \sqrt{|1 - \|\mu_\perp\|_2^2|} * \frac{e_{\text{w1B}} - \mu_B}{\|(e_{w1} - \mu_\perp) - \mu_B)\|_2}$$

$$e_{w2B}^{corrected} = \sqrt{|1 - \|\mu_\perp\|_2^2|} * \frac{e_{\text{w2B}} - \mu_B}{\|(e_{w2} - \mu_\perp) - \mu_B)\|_2}$$

$$e_1 = e_{w1B}^{corrected} + \mu_\perp$$
$$e_2 = e_{w2B}^{corrected} + \mu_\perp$$