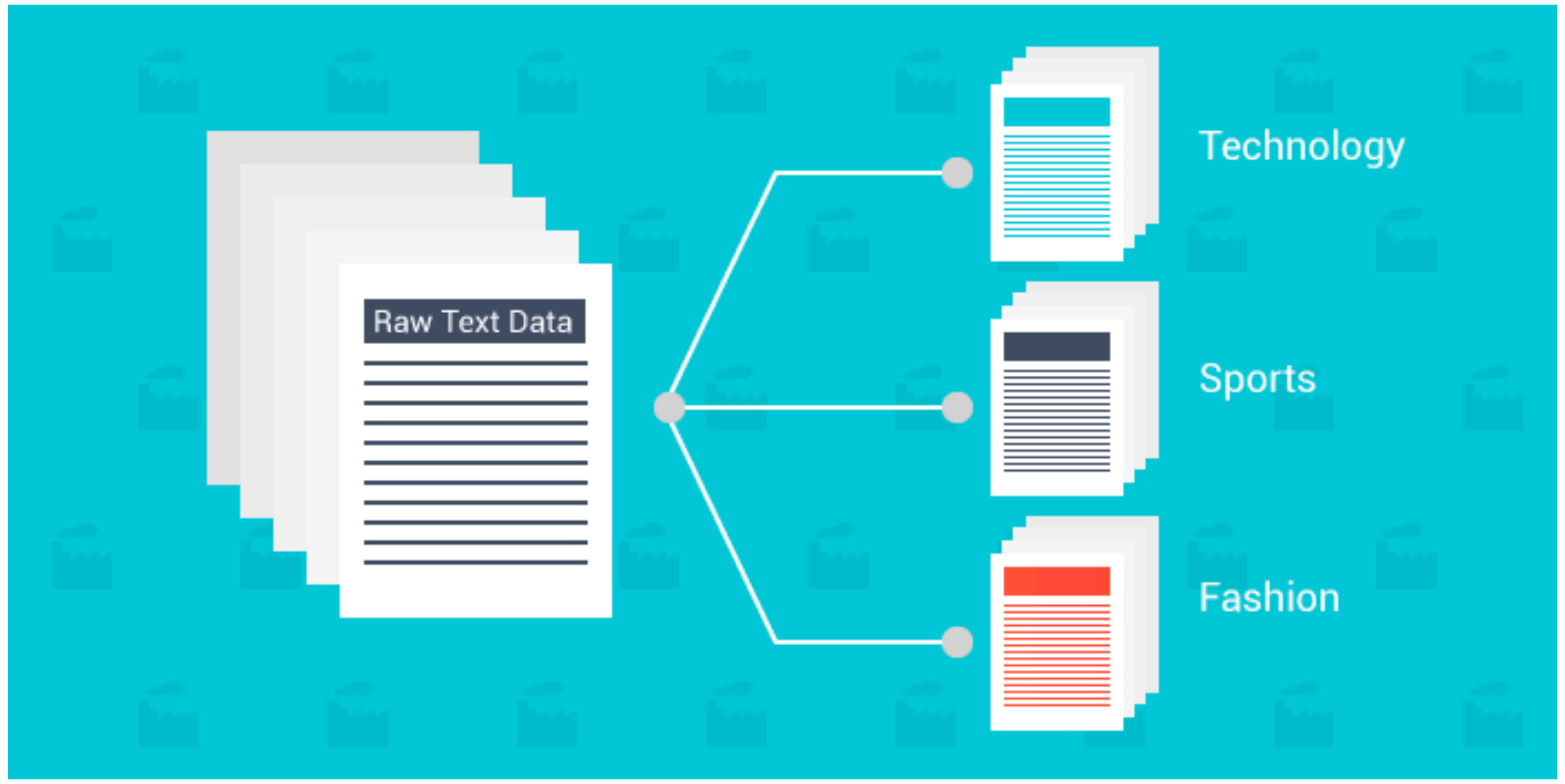
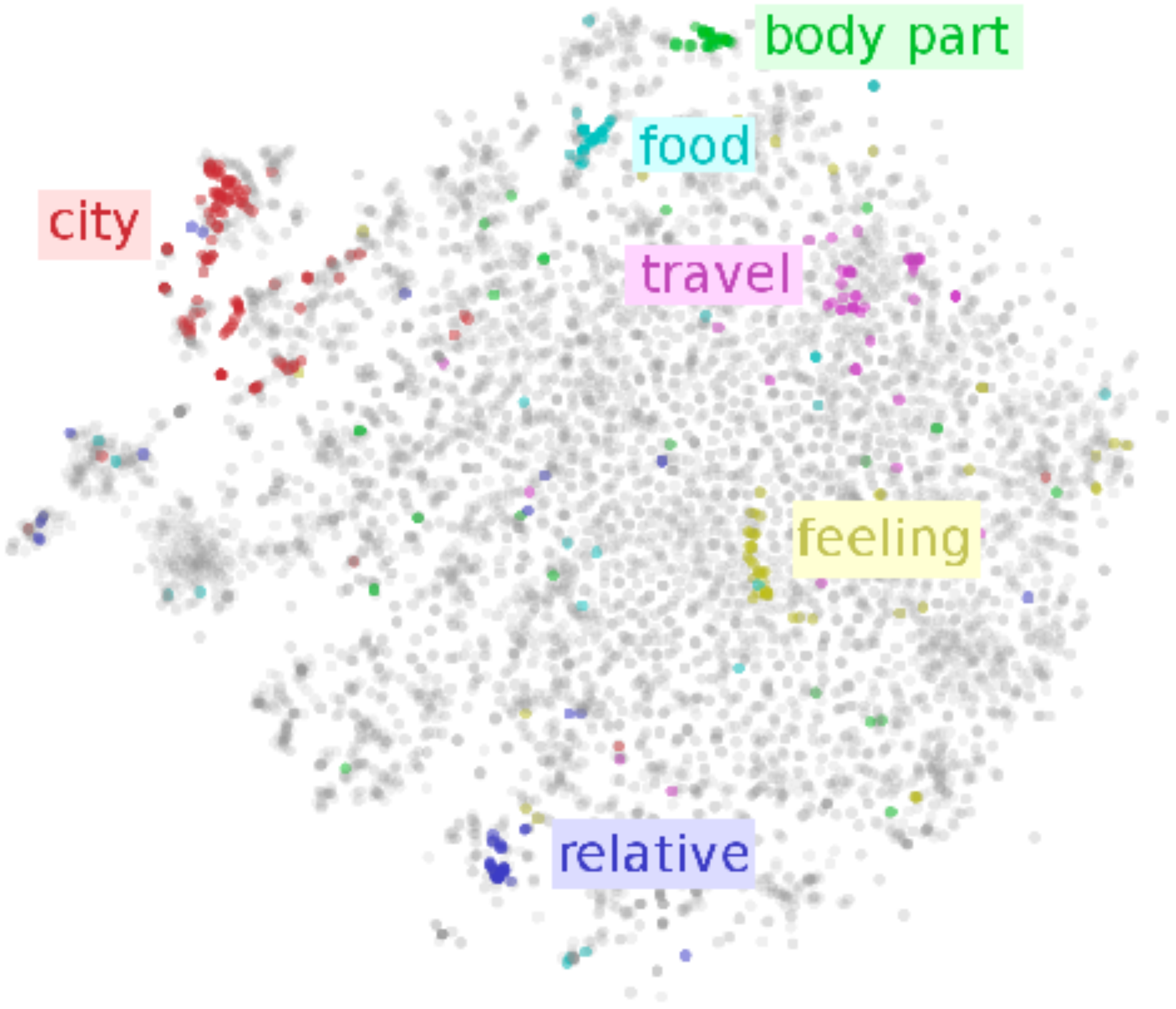
Machine Learning for text data

DIGHUM101
Tom van Nuenen

Why?



- Understand large collections of unstructured text bodies
- Classify documents
- Find relations between words
- Trace linguistic biases
- Etc.



Count Vectorizer

	1	**2**	**3**	**4**	**5**
Cat	0	0	1	0	0
Mouse	1	0	0	1	0
Dog	0	1	0	0	1
Food	0	0	1	0	0

Doesn’t tell you anything about word relationships!

TF-IDF

TF-IDF

Term Frequency-Inverse Document Frequency

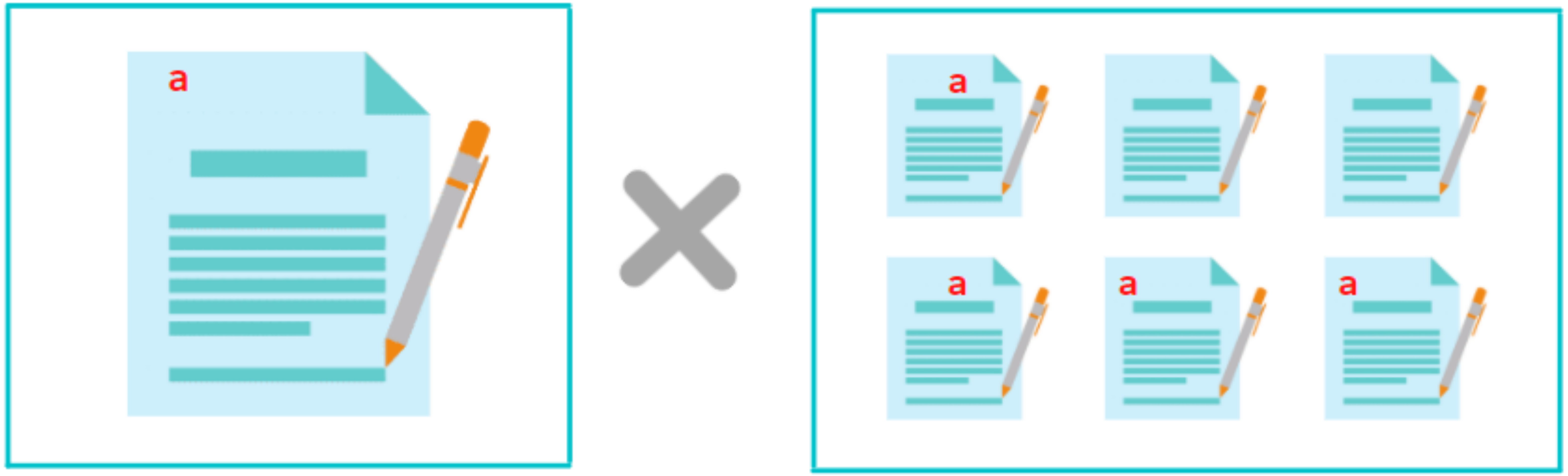
Term frequency: the amount of time a word shows up in a particular document, divided by the total number of words in the document.

Inverse document frequency: the inverse of the amount of documents that contain that term in your corpus.

```
IDF(t) = ln(Total # of documents / # of documents containing the term t)
```

TF-IDF

TF **IDF**



Frequency of a word within the document

Frequency of a word across the documents r/seduction

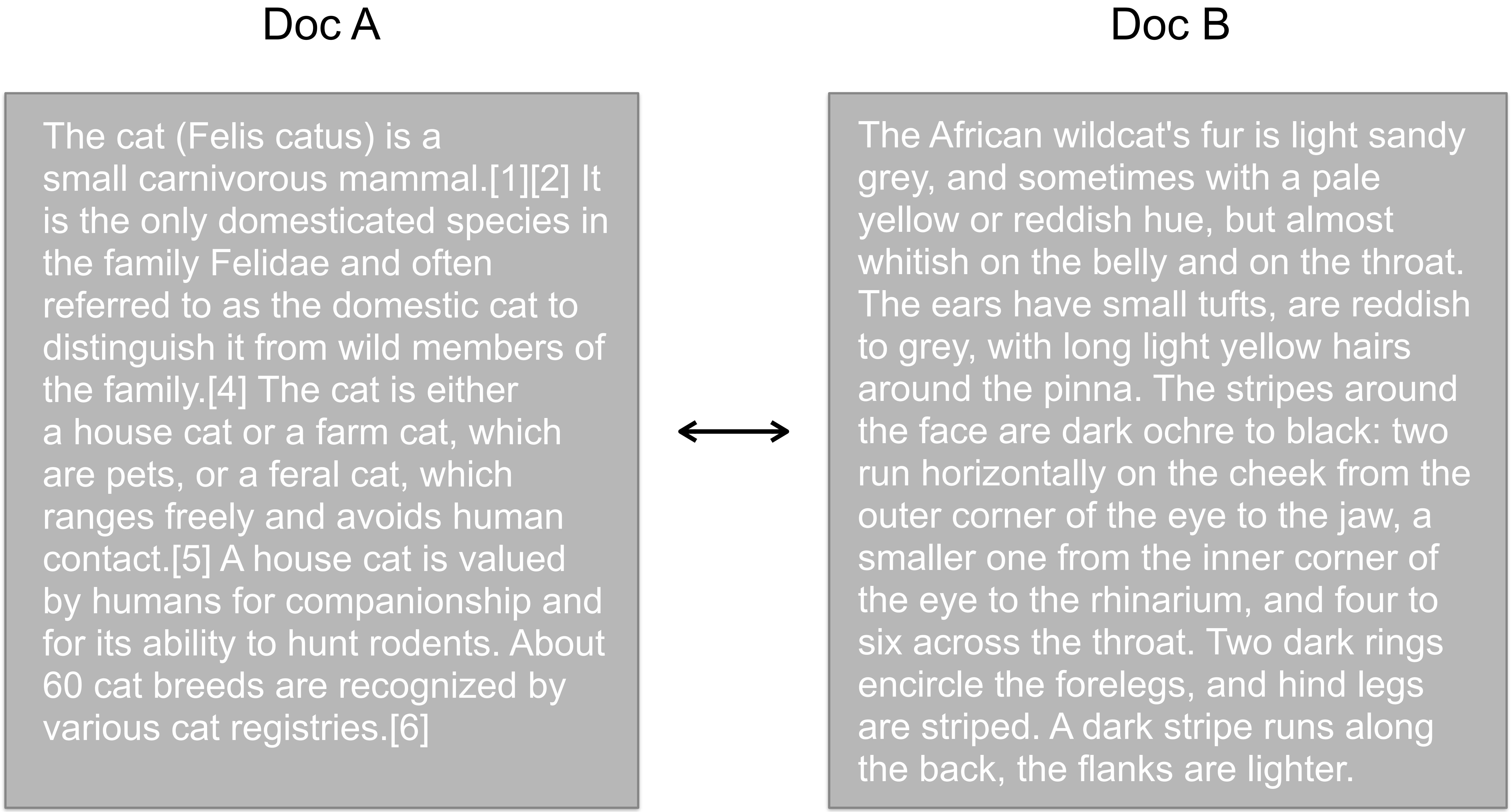
r/mensrights

r/theredpill

r/dating_advice

r/mgtow

“These distinctive words demonstrate a concern for [X]”



Doc A

The cat (Felis catus) is a small carnivorous mammal.[1][2] It is the only domesticated species in the family Felidae and often referred to as the domestic cat to distinguish it from wild members of the family.[4] The cat is either a house cat or a farm cat, which are pets, or a feral cat, which ranges freely and avoids human contact.[5] A house cat is valued by humans for companionship and for its ability to hunt rodents. About 60 cat breeds are recognized by various cat registries.[6]

Doc B

The African wildcat's fur is light sandy grey, and sometimes with a pale yellow or reddish hue, but almost whitish on the belly and on the throat. The ears have small tufts, are reddish to grey, with long light yellow hairs around the pinna. The stripes around the face are dark ochre to black: two run horizontally on the cheek from the outer corner of the eye to the jaw, a smaller one from the inner corner of the eye to the rhinarium, and four to six across the throat. Two dark rings encircle the forelegs, and hind legs are striped. A dark stripe runs along the back, the flanks are lighter.

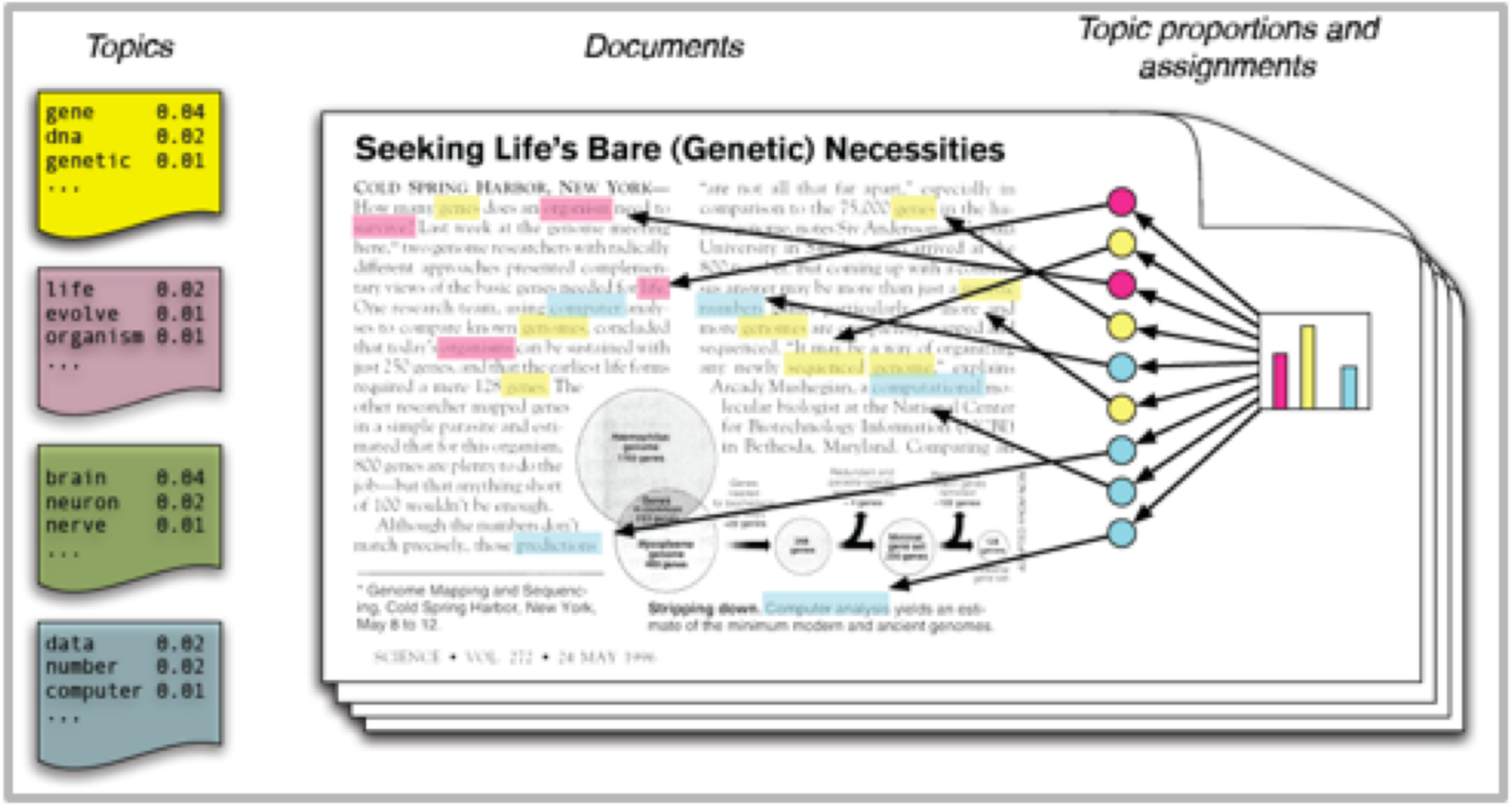
TFIDF: applications

- Finding distinctive words and documents in a corpus
- Summarizing & keyword extraction
- Creating classifiers
- Ranking search results based on relevance

Topic modeling

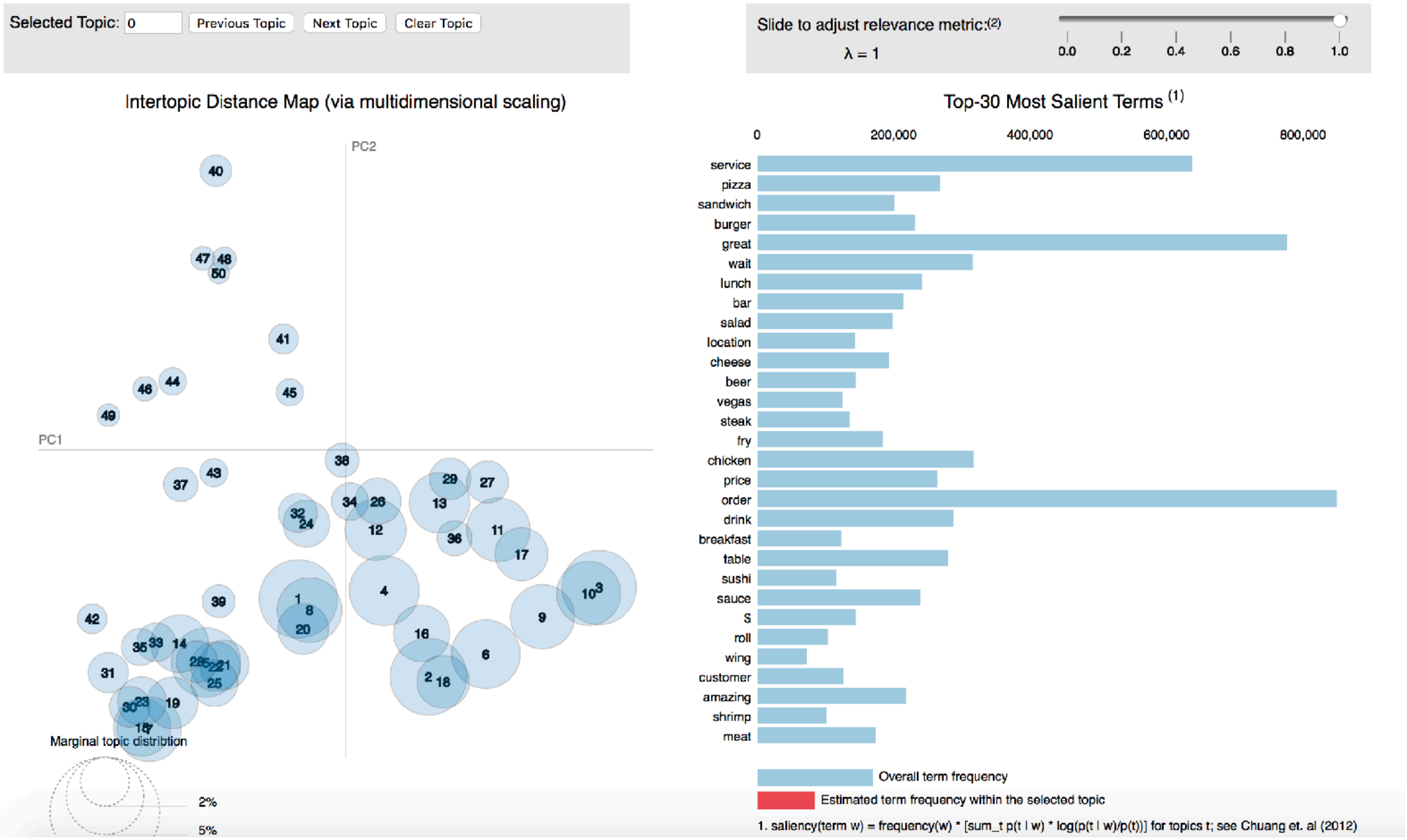
Topic modeling:
(1) unsupervised,
(2) probabilistic modeling
(3) using hidden variables

Topic modeling (LDA)



Topic: distribution over terms in vocabulary

Document: distribution over topic



Selected Topic: 0

Slide to adjust relevance metric:(2)

λ = 1

Intertopic Distance Map (via multidimensional scaling)

Marginal topic distribtion

Top-30 Most Salient Terms (1)

Overall term frequency

Estimated term frequency within the selected topic

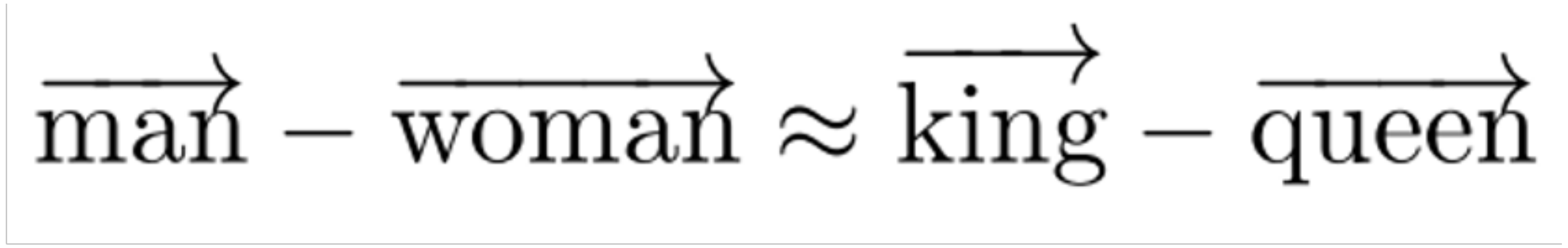
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

Topic modeling: applications

- Formulating new research questions
- Backing up qualitative research
- Legal discovery
- Recommender systems

Word embeddings

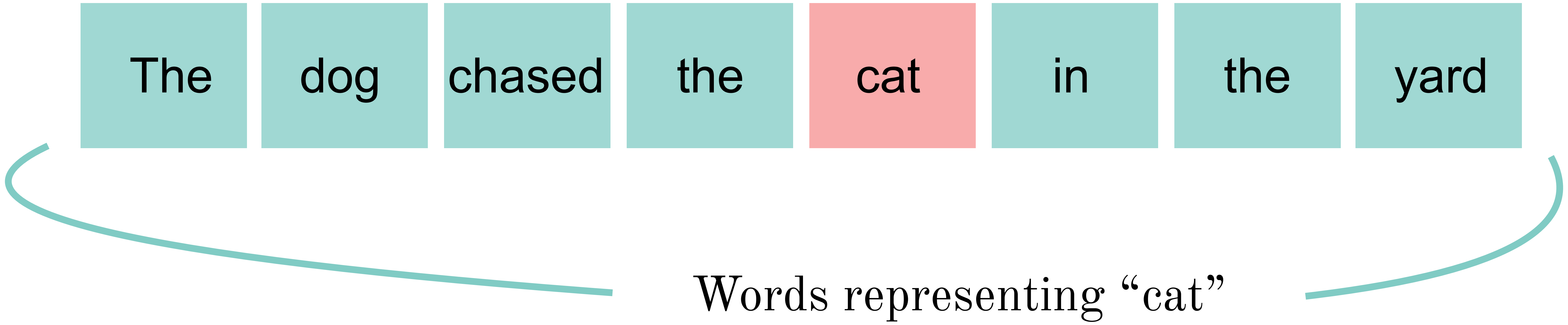
Word embeddings



$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

Distributional similarity

- “You shall know a word by the company it keeps” (JR Firth 1957)
- Looking at *neighboring words*

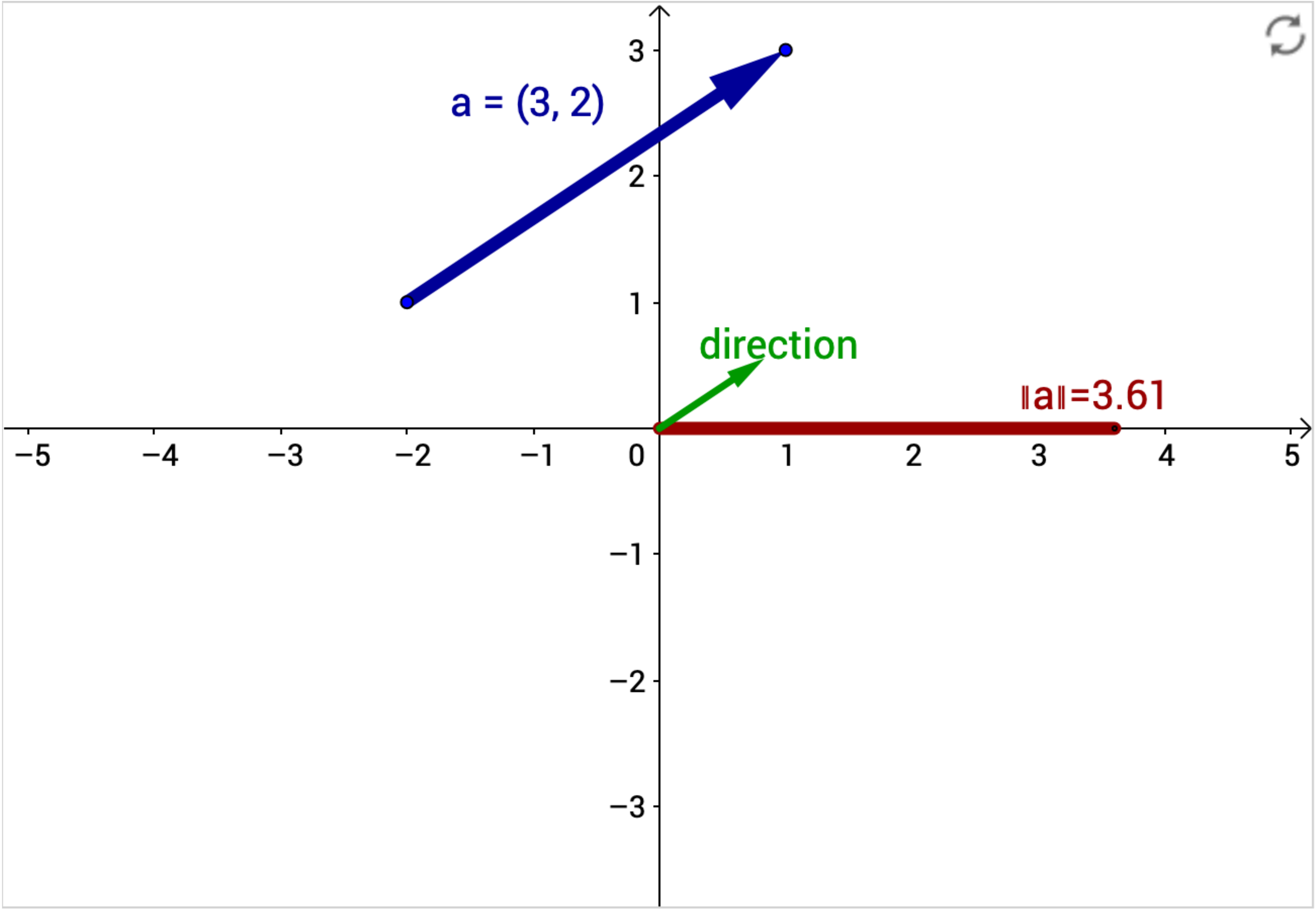


The dog chased the cat in the yard

Words representing “cat”

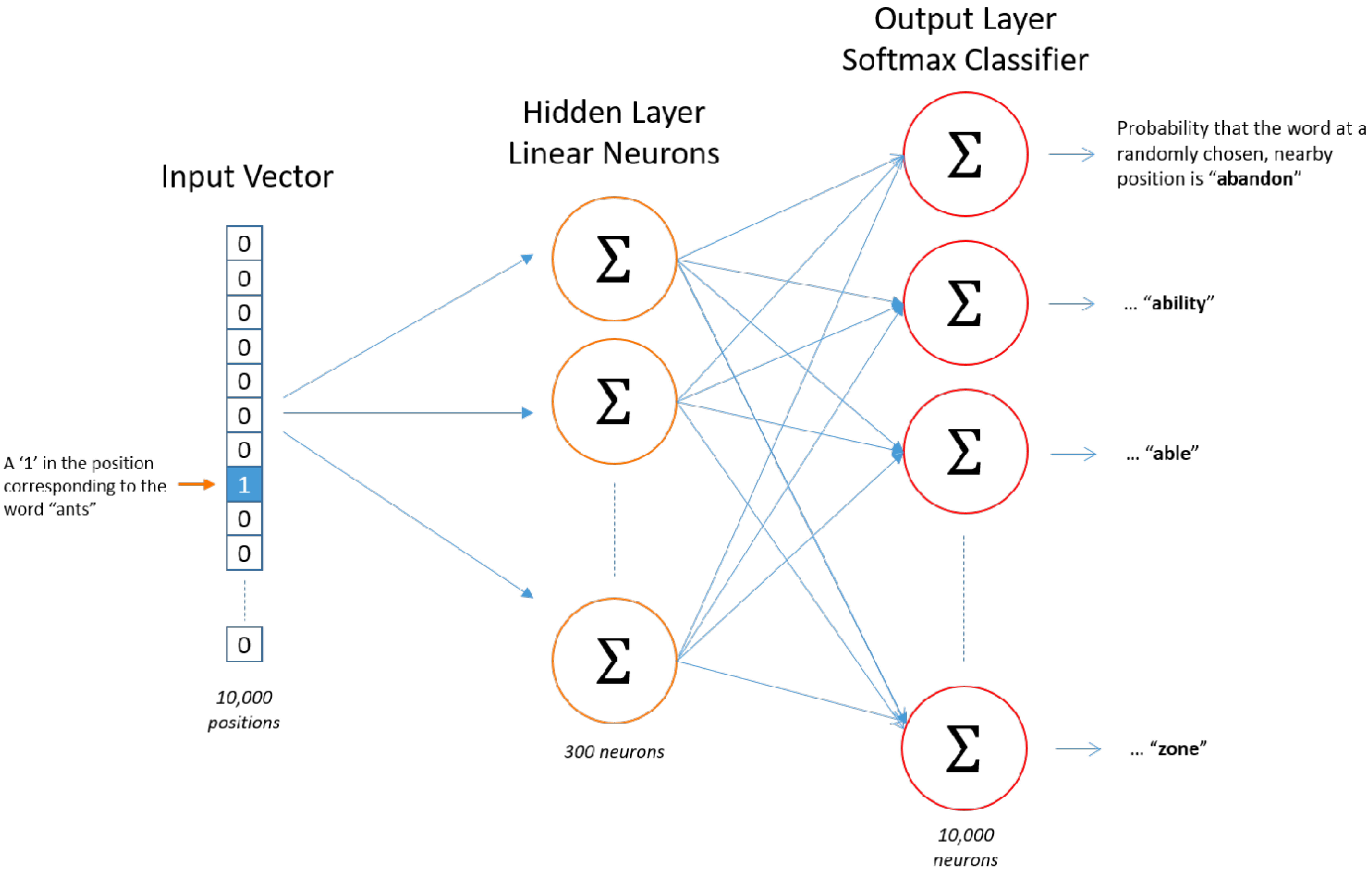
Word2vec: using word vectors that are good at predicting other words in its context

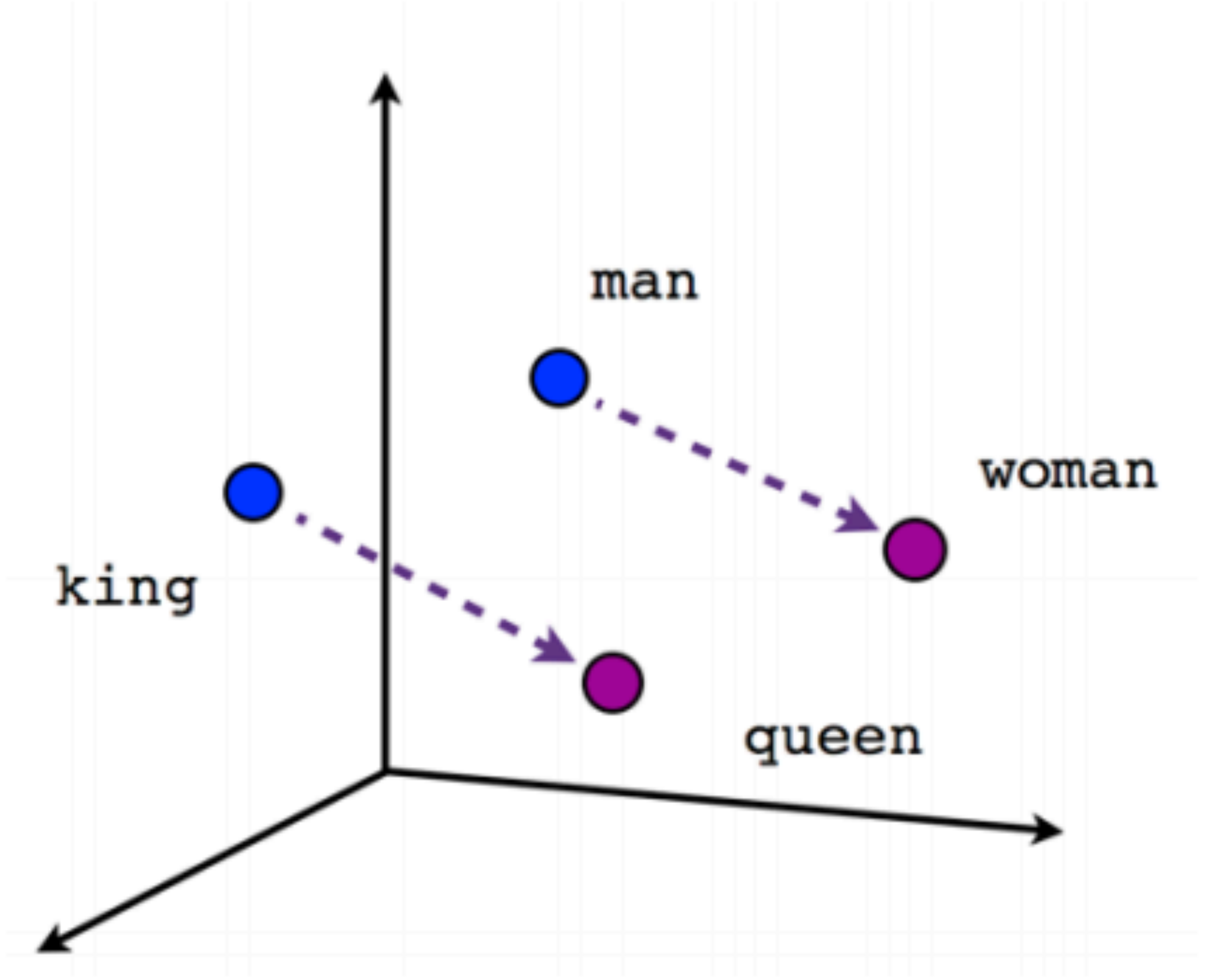
$$\text{Cat} = \begin{bmatrix} 0.275 \\ 0.819 \\ -0.187 \\ -0.203 \\ 0.374 \end{bmatrix}$$


a = (3, 2)
direction
‖a‖=3.61
−5
−4
−3
−2
−1
0
1
2
3
4
5
3
2
1
−1
−2
−3

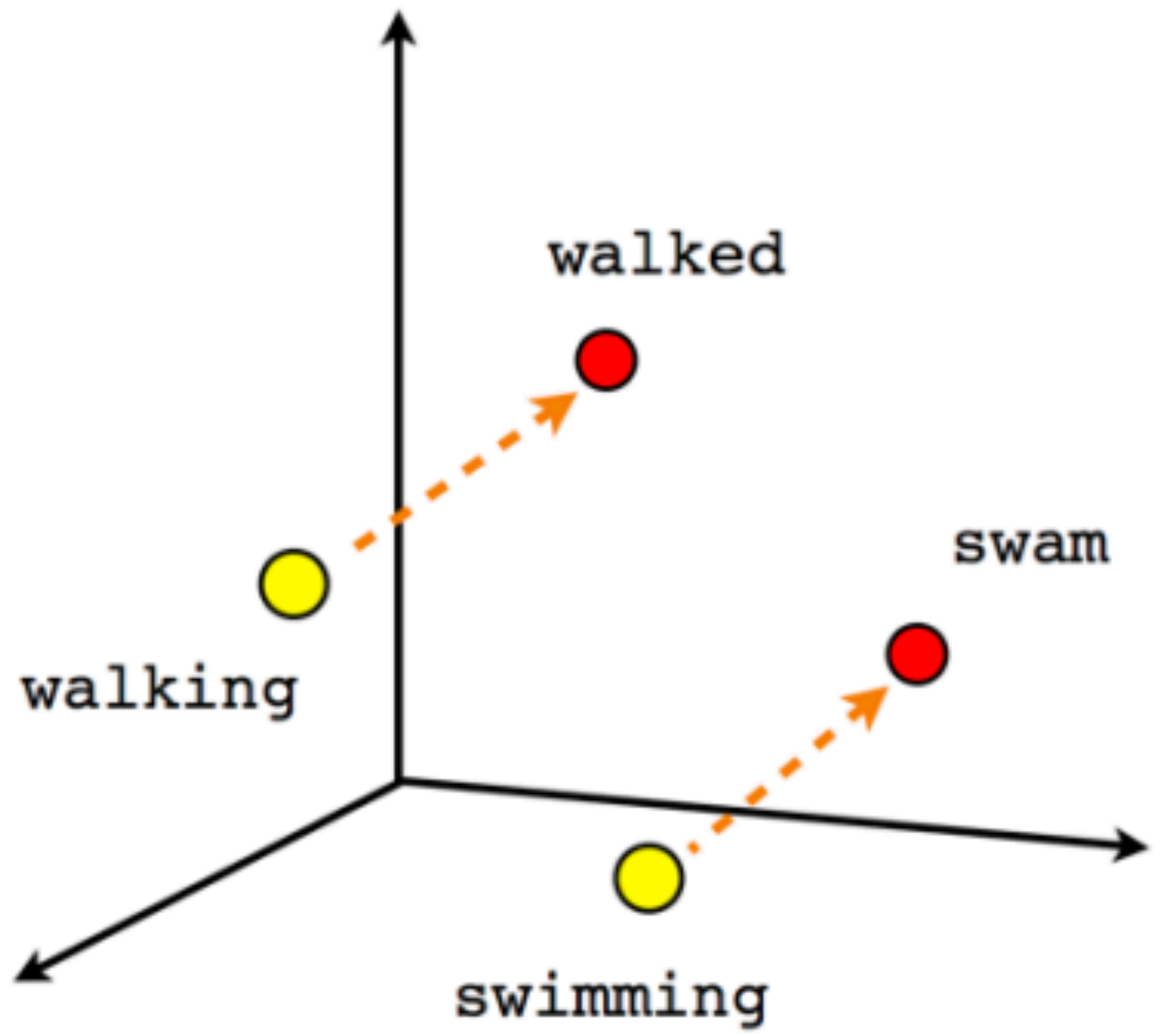
Word2vec

- Neural network word embeddings
- For each word type t=1, predict between surrounding words in a window of “radius” *m* of every word
- Objective (cost) function: maximize probability of any context word given the current center word

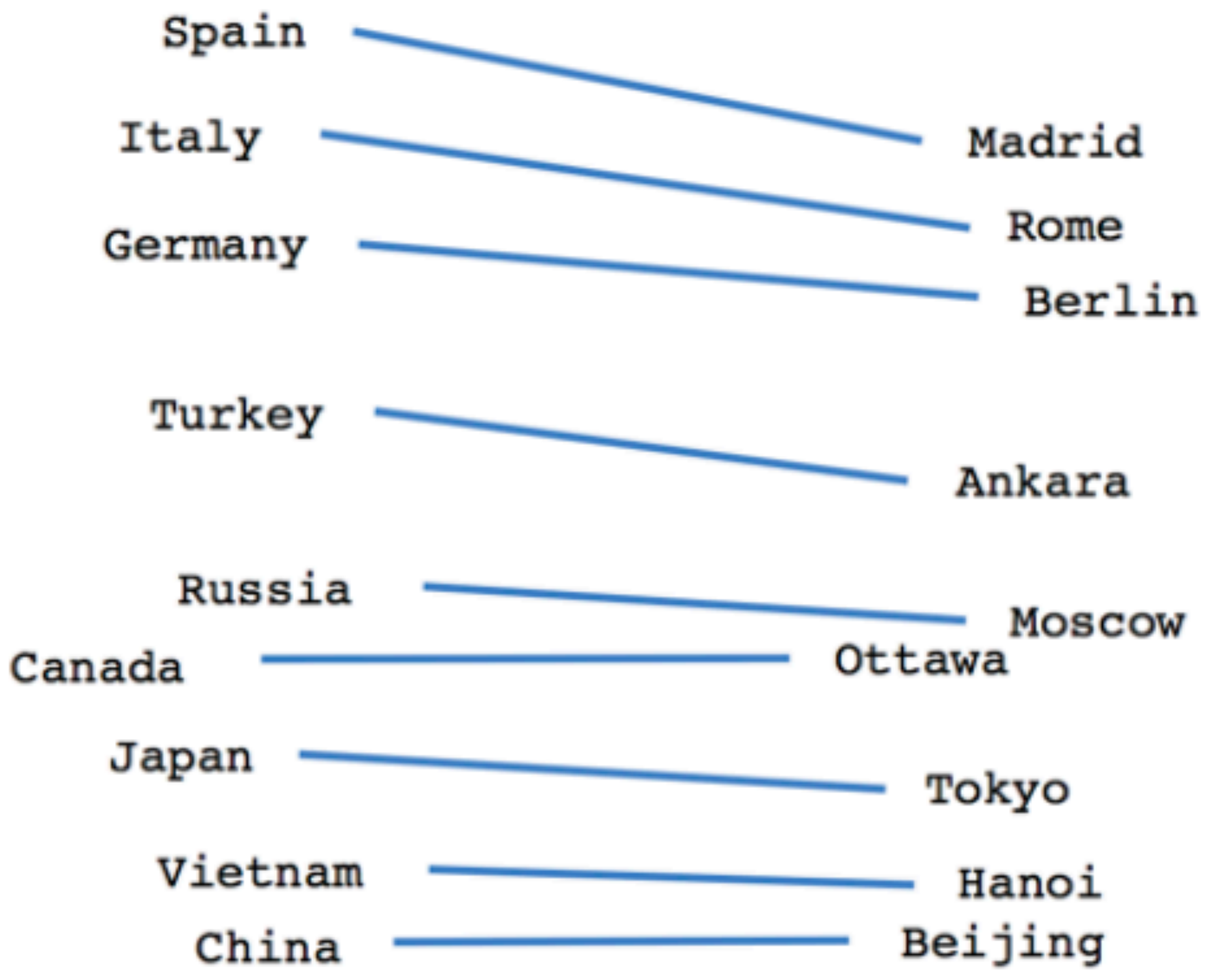

Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
0
0
0
0
0
0
0
1
0
0
0
A '1' in the position corresponding to the word "ants"
10,000 positions
Σ
300 neurons
Probability that the word at a randomly chosen, nearby position is "**abandon**"
... "**ability**"
... "**able**"
... "**zone**"
10,000 neurons



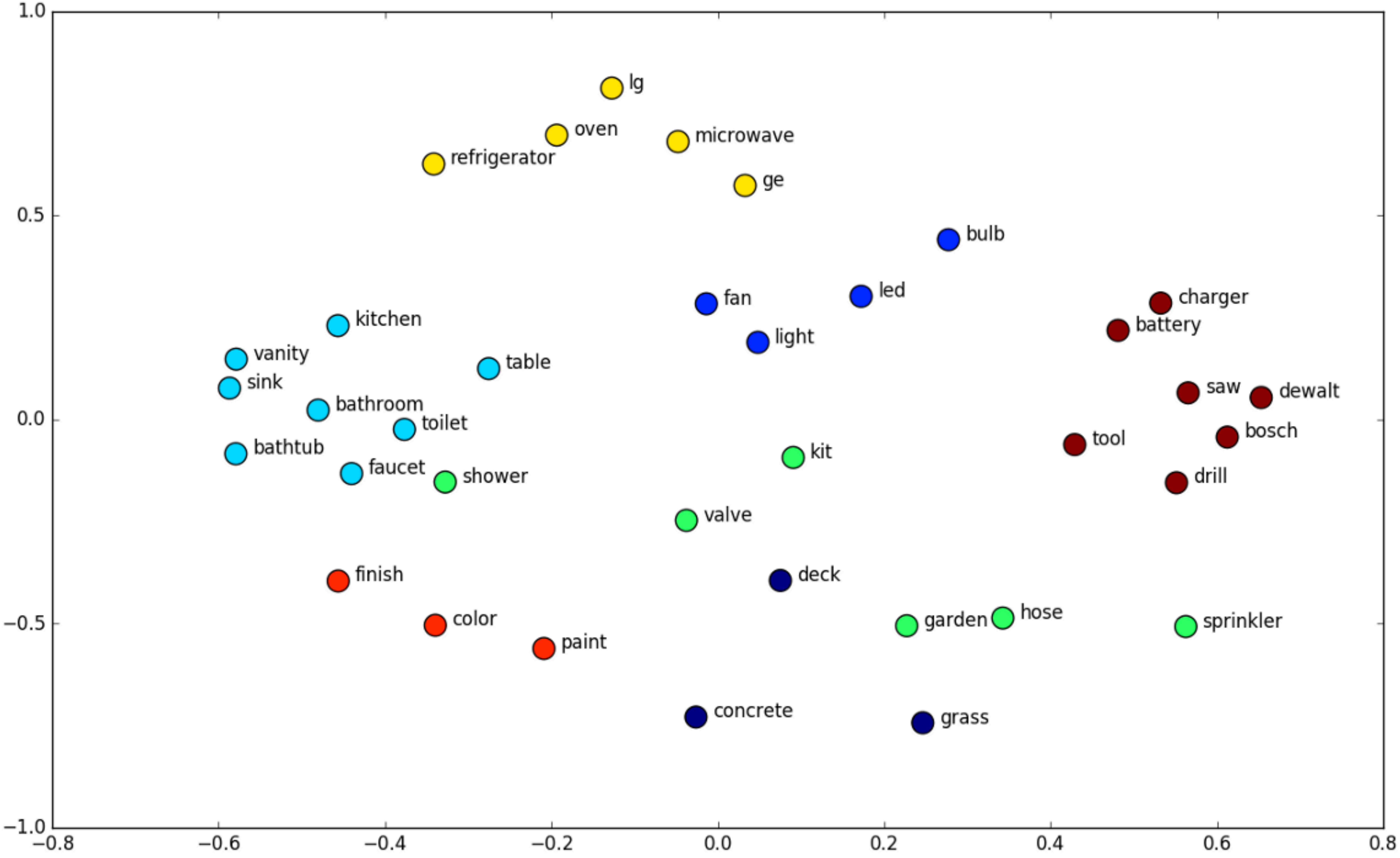
Male-Female



Verb tense



Country-Capital


lg
oven
microwave
refrigerator
ge
bulb
led
fan
charger
light
battery
kitchen
vanity
table
sink
saw
dewalt
bathroom
toilet
tool
bosch
bathtub
kit
faucet
shower
drill
valve
finish
deck
hose
garden
sprinkler
color
paint
concrete
grass
1.0
0.5
0.0
−0.5
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

↓

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

“Man is to computer programmer as woman is to homemaker”

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In D. D. Lee, M. Sugiyama, U. V Luxburg, I. Guyon, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 29 (pp. 4349–4357).

Word embeddings: applications

- Calculating word similarities
- Improving recommender systems
- Building chatbots
- Tracing language biases and stereotypes