Natural Language Processing

Lecture 07

Dirk Hovy

dirk.hovy@unibocconi.it

@dirk_hovy

Bocconi

Today's Goals

- Understand the difference between **sparse** and **dense** representations
- Learn about **word2vec** and **doc2vec**
- **Understand** the underlying algorithms

Dense Distributed Representations

Distributional Hypothesis

"You shall know the meaning of a word by the company it keeps"

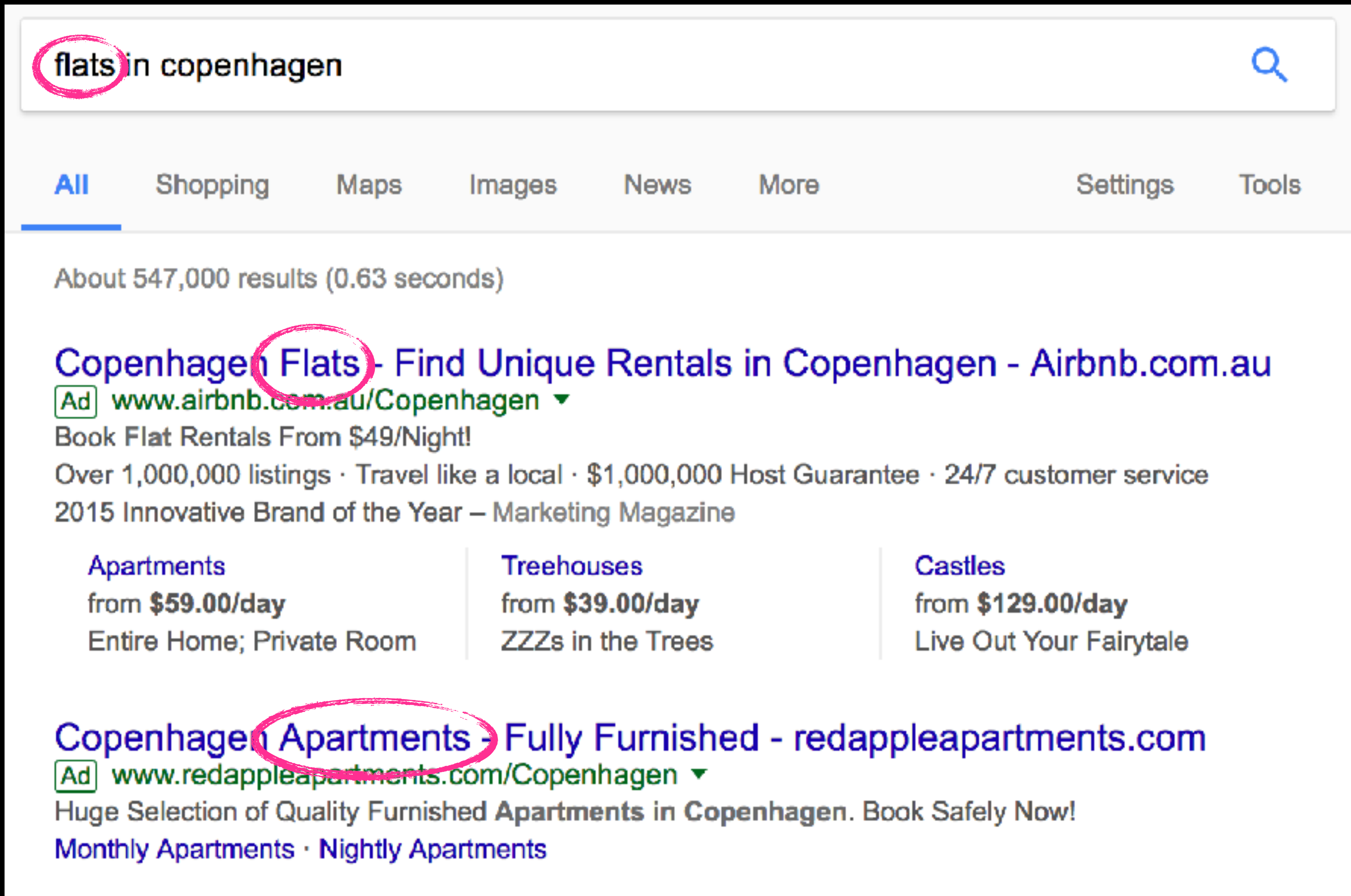
Firth (1957)

Similar words have similar **contexts**

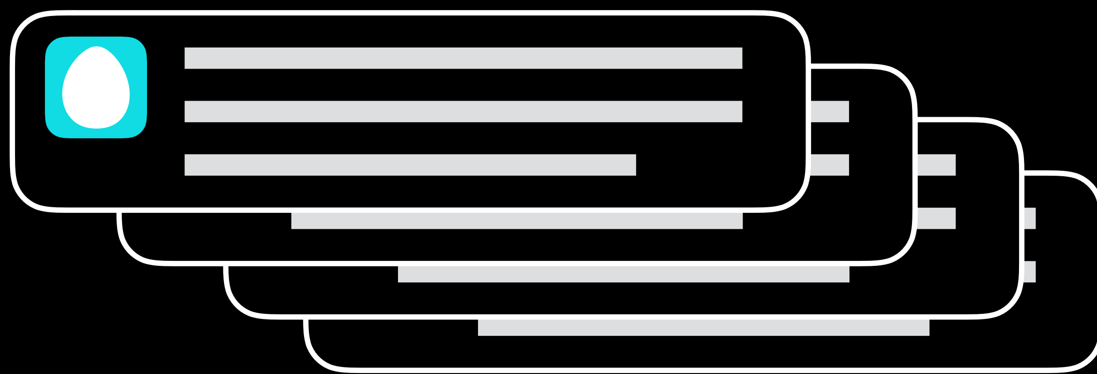
Represent **words** as **vectors**/points in space

Similar words have similar vectors

An Example


flats in copenhagen
All
Shopping
Maps
Images
News
More
Settings
Tools
About 547,000 results (0.63 seconds)
Copenhagen Flats - Find Unique Rentals in Copenhagen - Airbnb.com.au
Ad www.airbnb.com.au/Copenhagen
Book Flat Rentals From $49/Night!
Over 1,000,000 listings · Travel like a local · $1,000,000 Host Guarantee · 24/7 customer service
2015 Innovative Brand of the Year – Marketing Magazine
Apartments
from $59.00/day
Entire Home; Private Room
Treehouses
from $39.00/day
ZZZs in the Trees
Castles
from $129.00/day
Live Out Your Fairytale
Copenhagen Apartments - Fully Furnished - redappleapartments.com
Ad www.redappleapartments.com/Copenhagen
Huge Selection of Quality Furnished Apartments in Copenhagen. Book Safely Now!
Monthly Apartments · Nightly Apartments

Latent Semantic Analysis



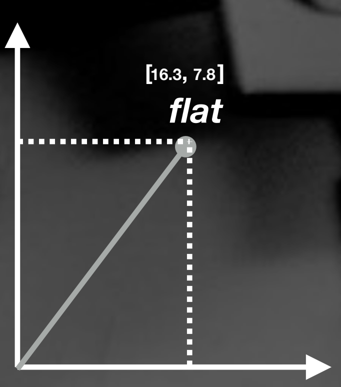
CONTEXTS

TARGETS

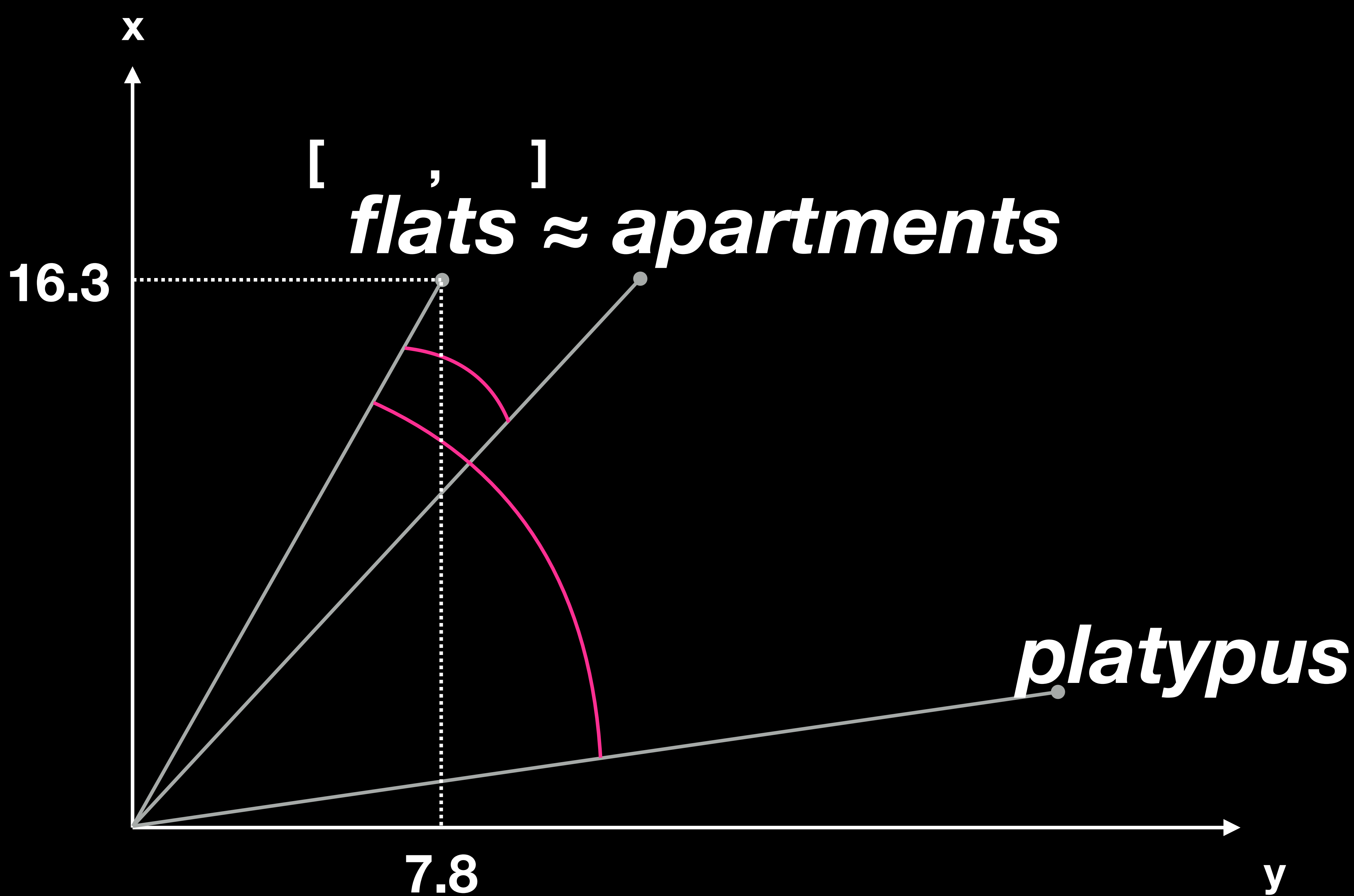
	rent	location	fairy	rainbow	prince	sleep
flat	87	73	14	11	7	45
apartment	[illegible]	[illegible]	12	[illegible]	11	32
unicorn	27	[illegible]	79	92	54	16
toad	4	[illegible]	73	55	[illegible]	[illegible]
bed	34	42	21	15	62	97

HAVE TO DEFINE CONTEXTS -> CHANGES NUMBERS CONSIDERABLY

Part 1
Representing Words as Vectors

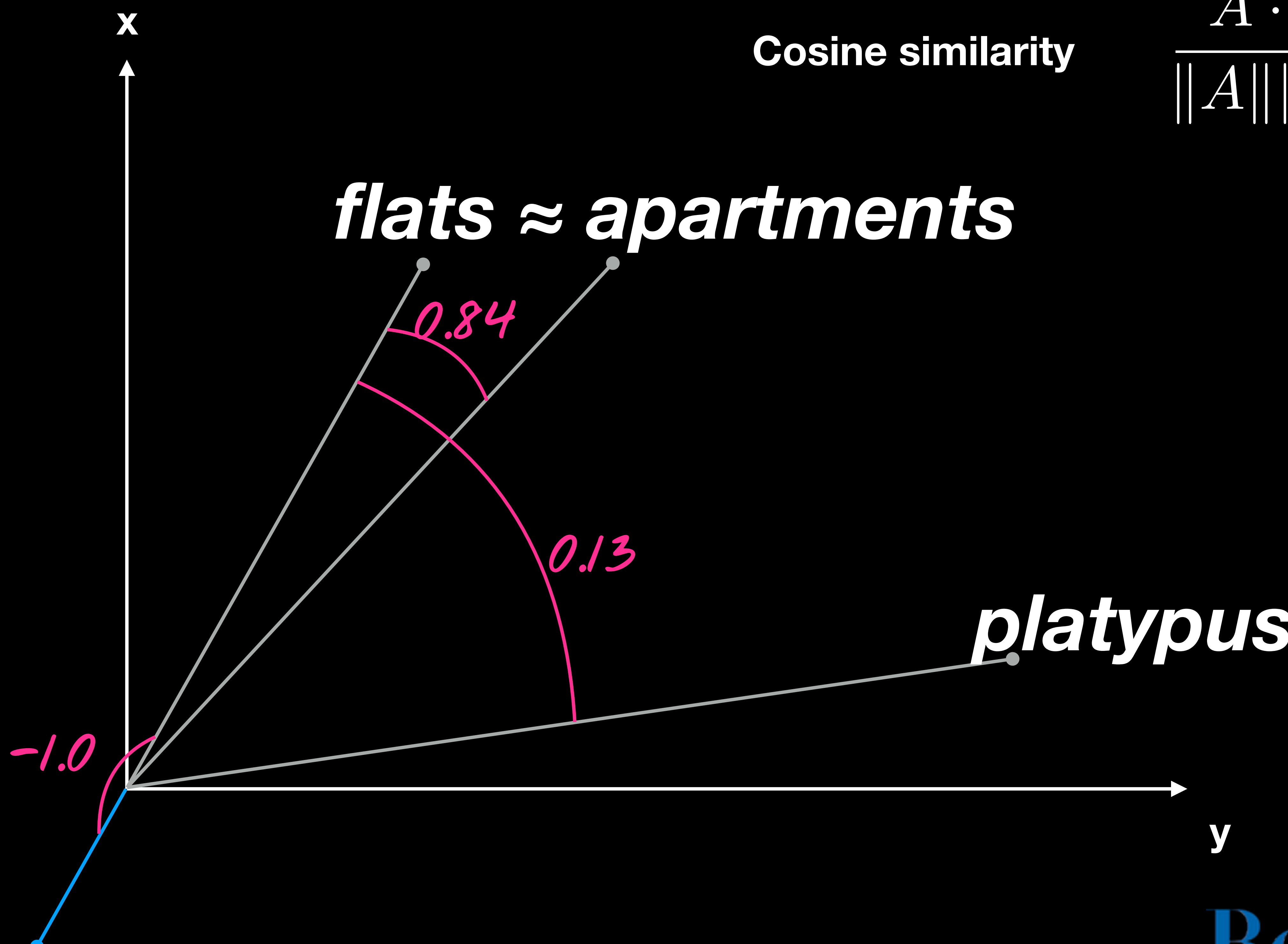


Semantic Similarity



Similarity Measures

Cosine similarity $$\frac{A \cdot B}{\|A\| \|B\|}$$



Dot Product

- “combine” vectors to a scalar

SUM

$$x \cdot y = \sum_{i=1}^{D} x_i y_i$$

MULTIPLY

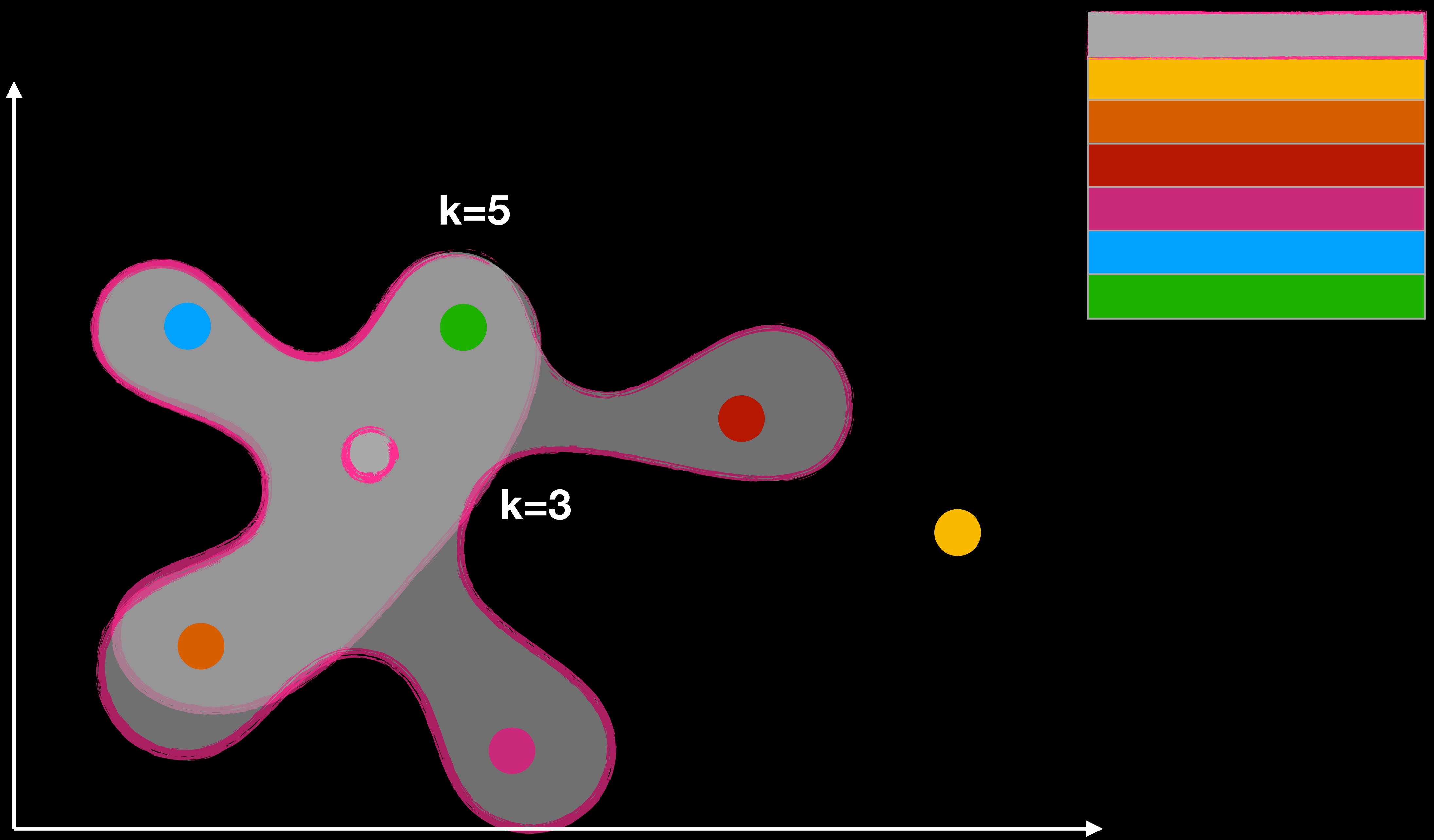
$$\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 6 \end{bmatrix} = \begin{matrix} 1 \\ 4 \\ 3 \end{matrix}$$

Vector Norm

- add up square of each element, take $\sqrt{}$

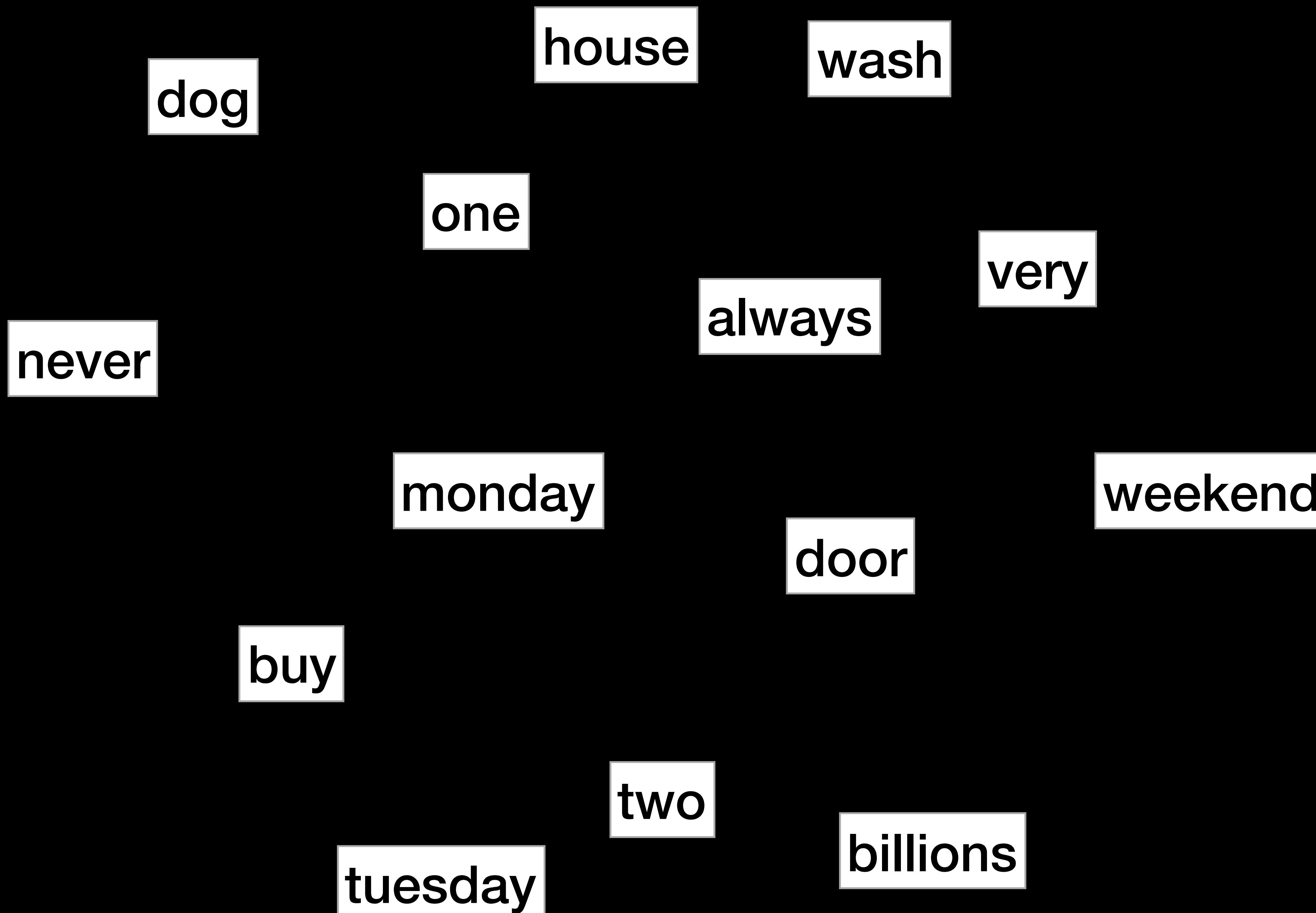
$$\begin{bmatrix} 2 \\ 6 \end{bmatrix} = \sqrt{2^2 + 6^2} = 6.324$$

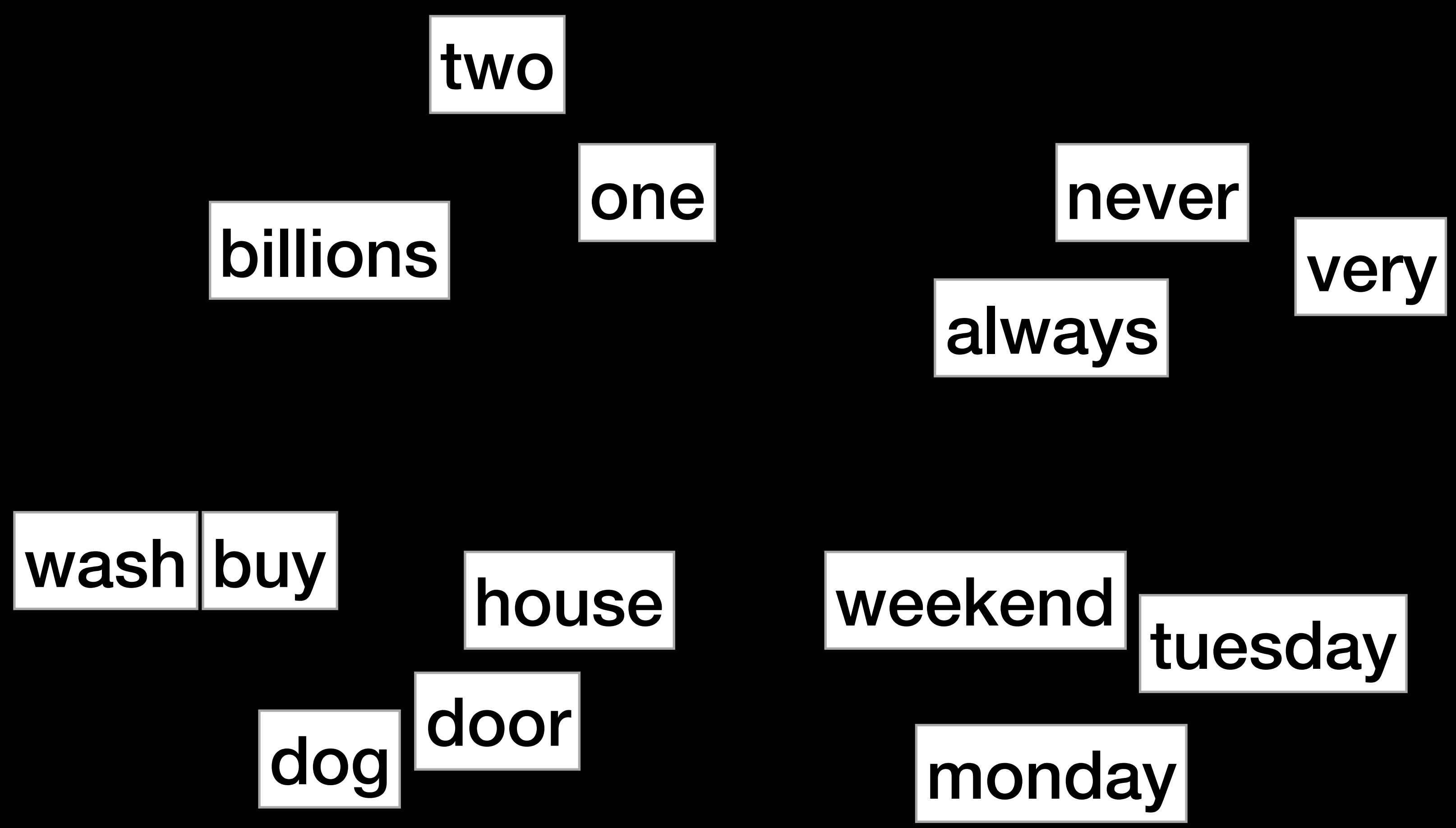
Nearest neighbors



Word2Vec – Intuitively

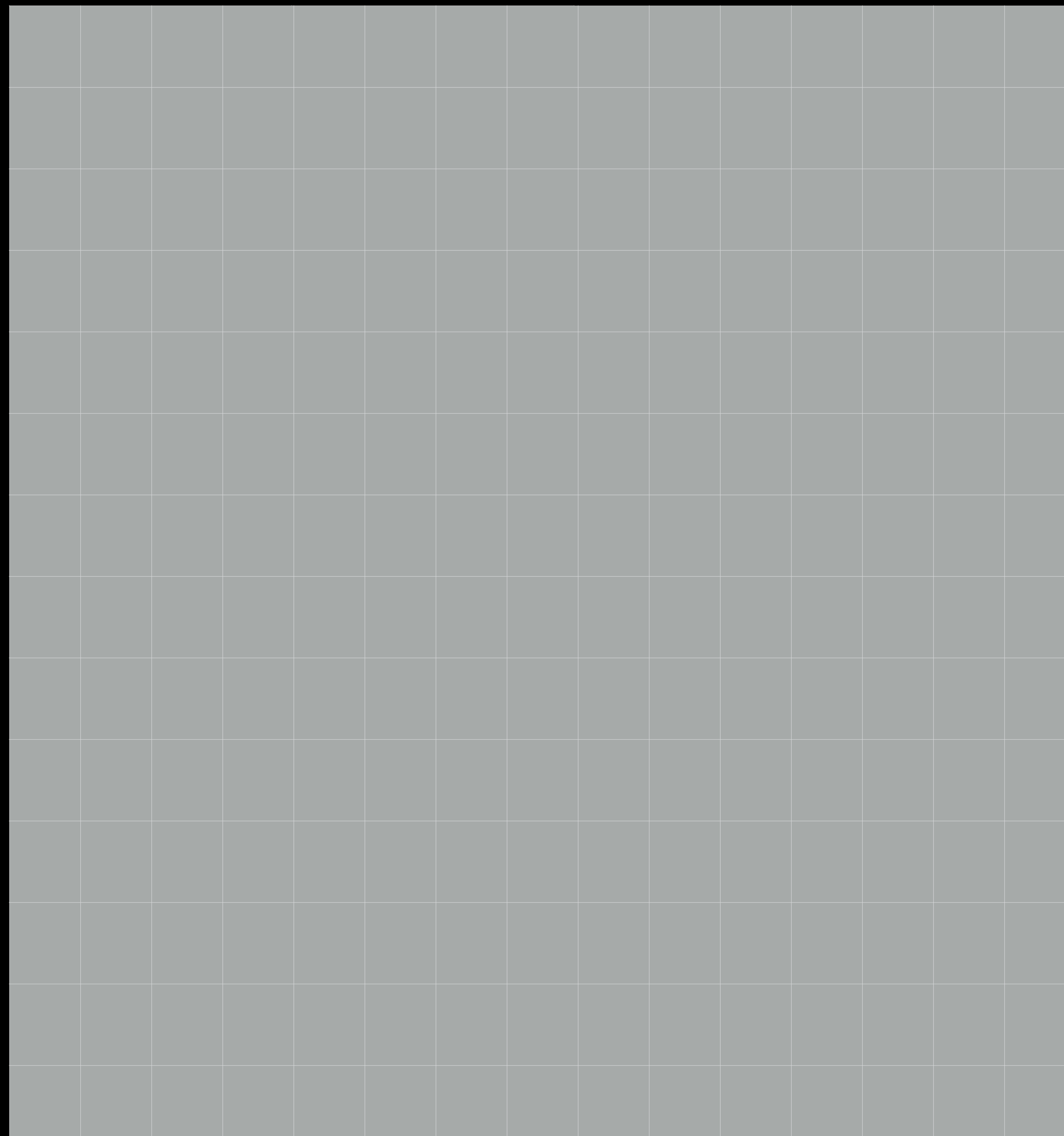
```
place all words randomly on fridge
for each pair of words:
    if in same sentence:
        move closer together
    else:
        move further apart
```


house
wash
dog
one
very
always
never
monday
weekend
door
buy
two
billions
tuesday


two
one
never
billions
very
always
wash
buy
house
weekend
tuesday
door
dog
monday

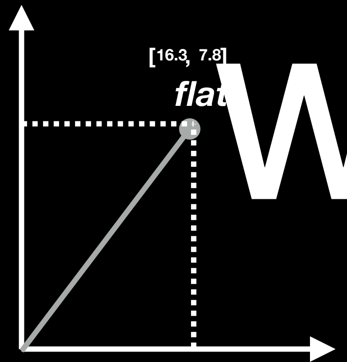
X-POS Y-POS

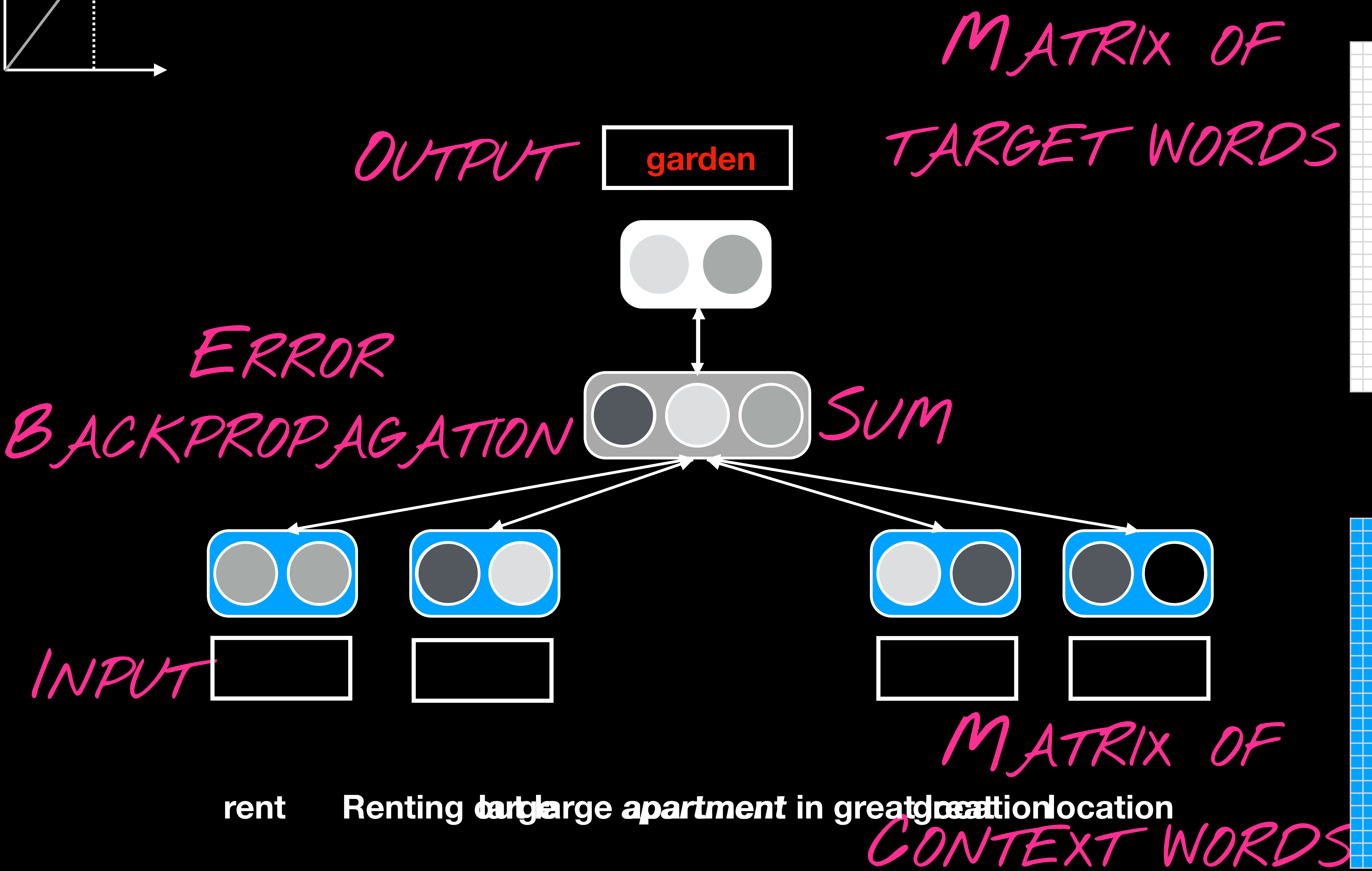
two
one
billions
never
very
always
wash
buy
house
door
dog
weekend
monday
tuesday



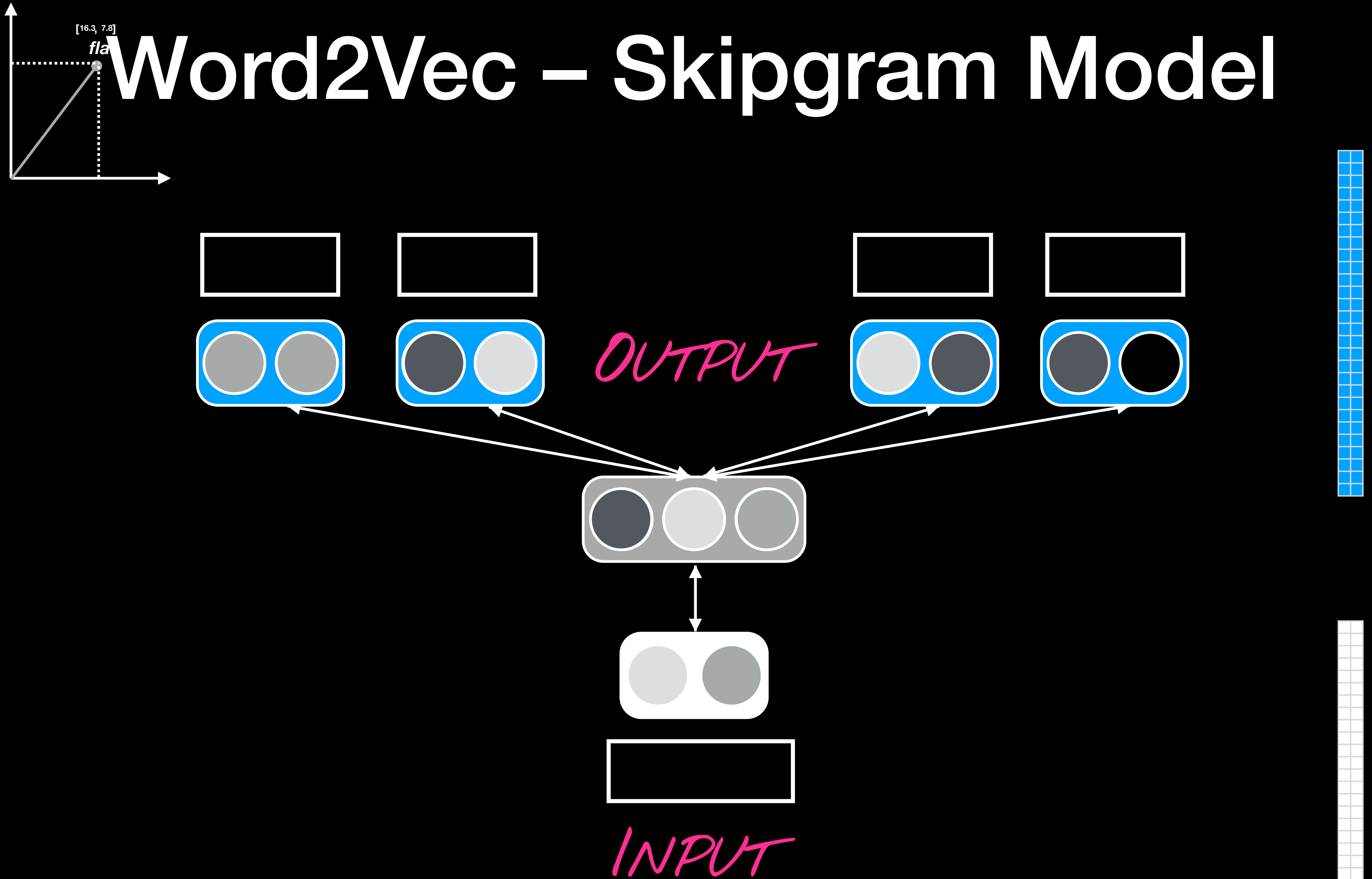
VECTORS

Word2Vec – CBOW Model





Word2Vec – Skipgram Model



Nuts and Bolts
and
Engineering Tricks

Bocconi

Problem?

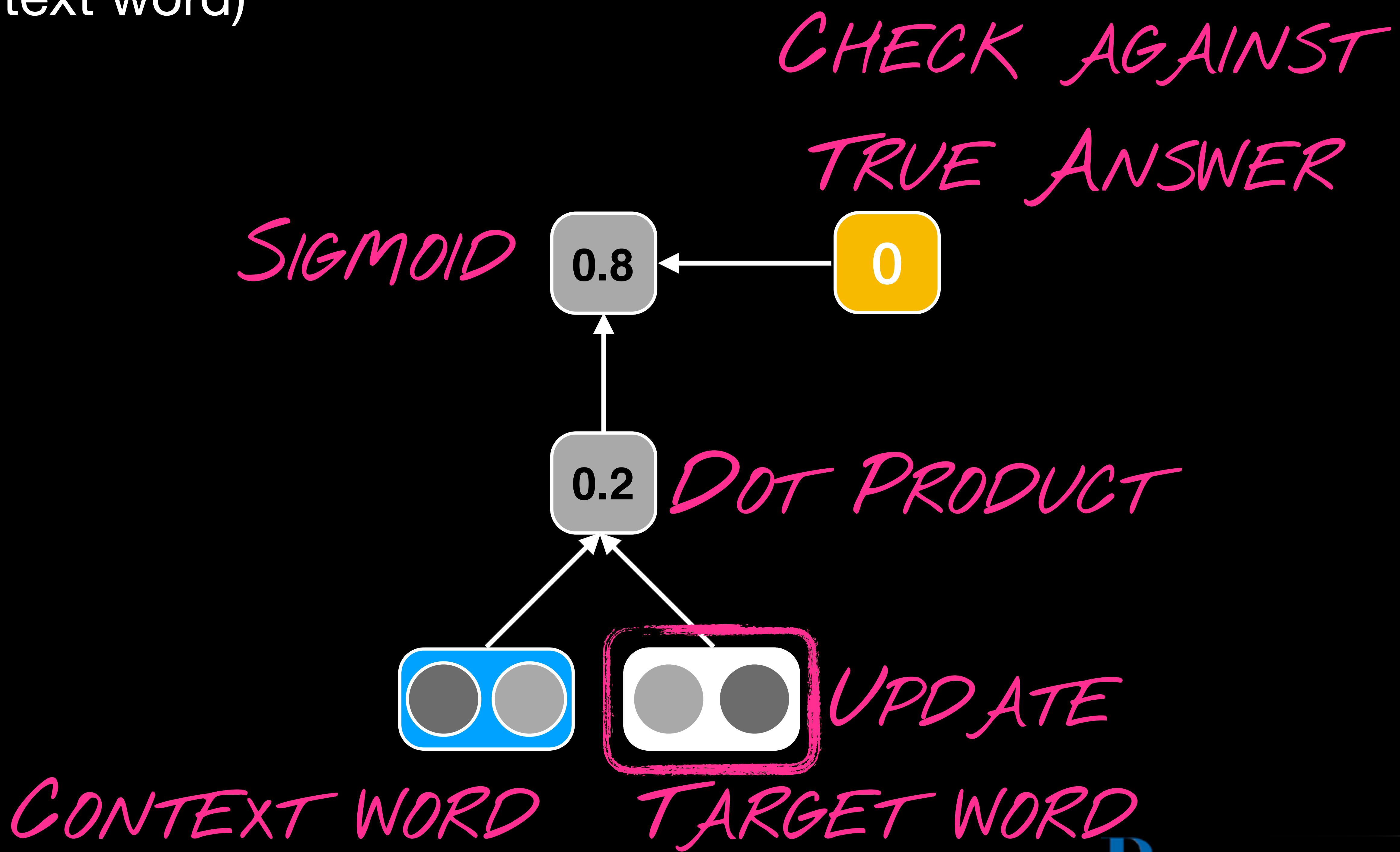
- We are trying to learn a conditional probability distribution over the vocabulary *for each word in the vocabulary*:

$$P(w_{out} | w_{in})$$

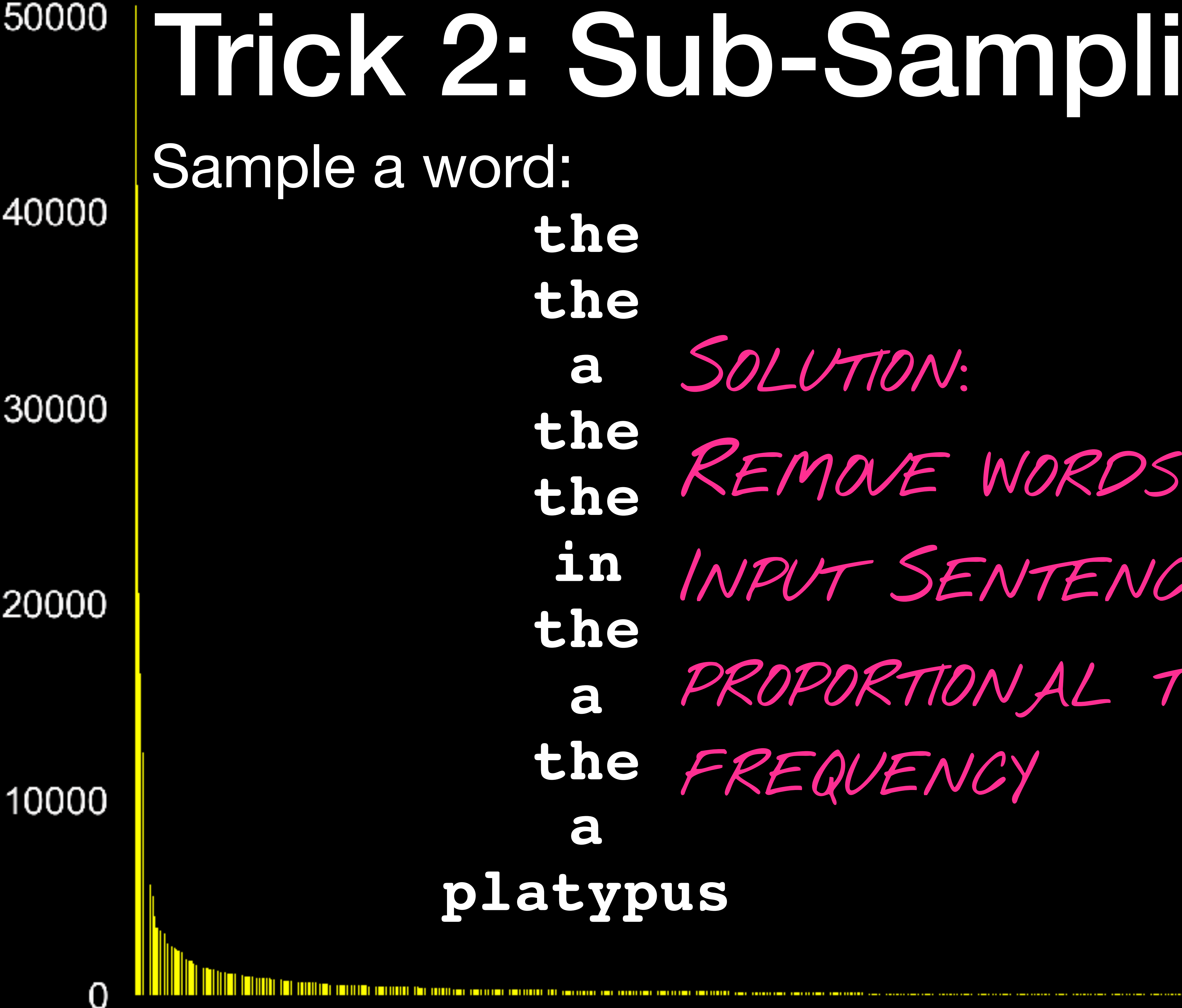
- With a large vocabulary comes large trouble…

Trick 1: Negative Sampling

Sample small set of words, labeled as 0 (not a context word) or 1 (is a context word)



Trick 2: Sub-Sampling



Sample a word:

the
the
a
the
the
in
the
a
the
a
platypus

SOLUTION:
REMOVE WORDS IN THE INPUT SENTENCE PROPORTIONAL TO THEIR FREQUENCY

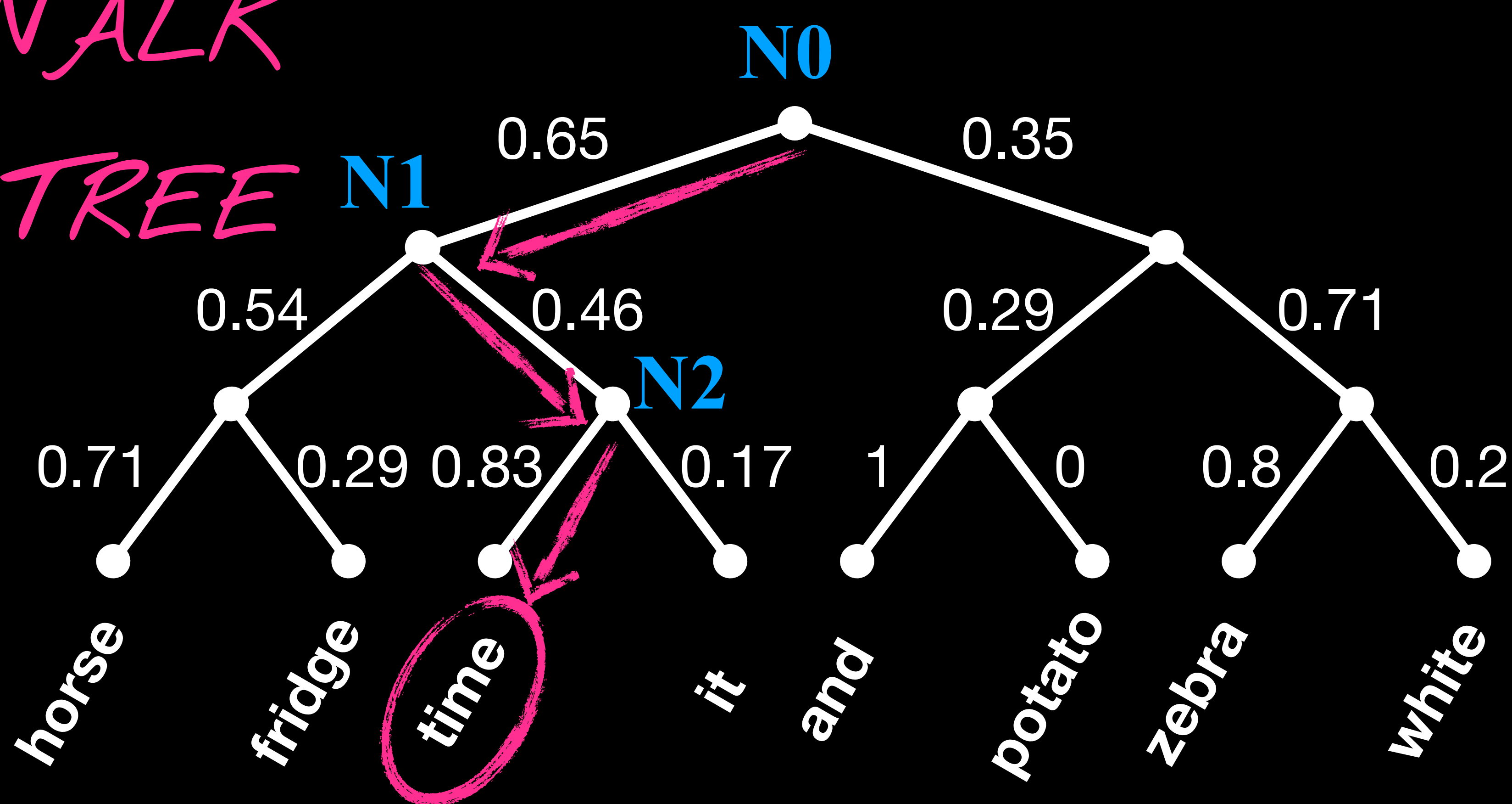
Bocconi

Trick 3: Hierarchical Softmax

Update to regular softmax: $O(|V|)$

Hierarchical softmax: $O(log|V|)$

RANDOM WALK
ALONG A TREE



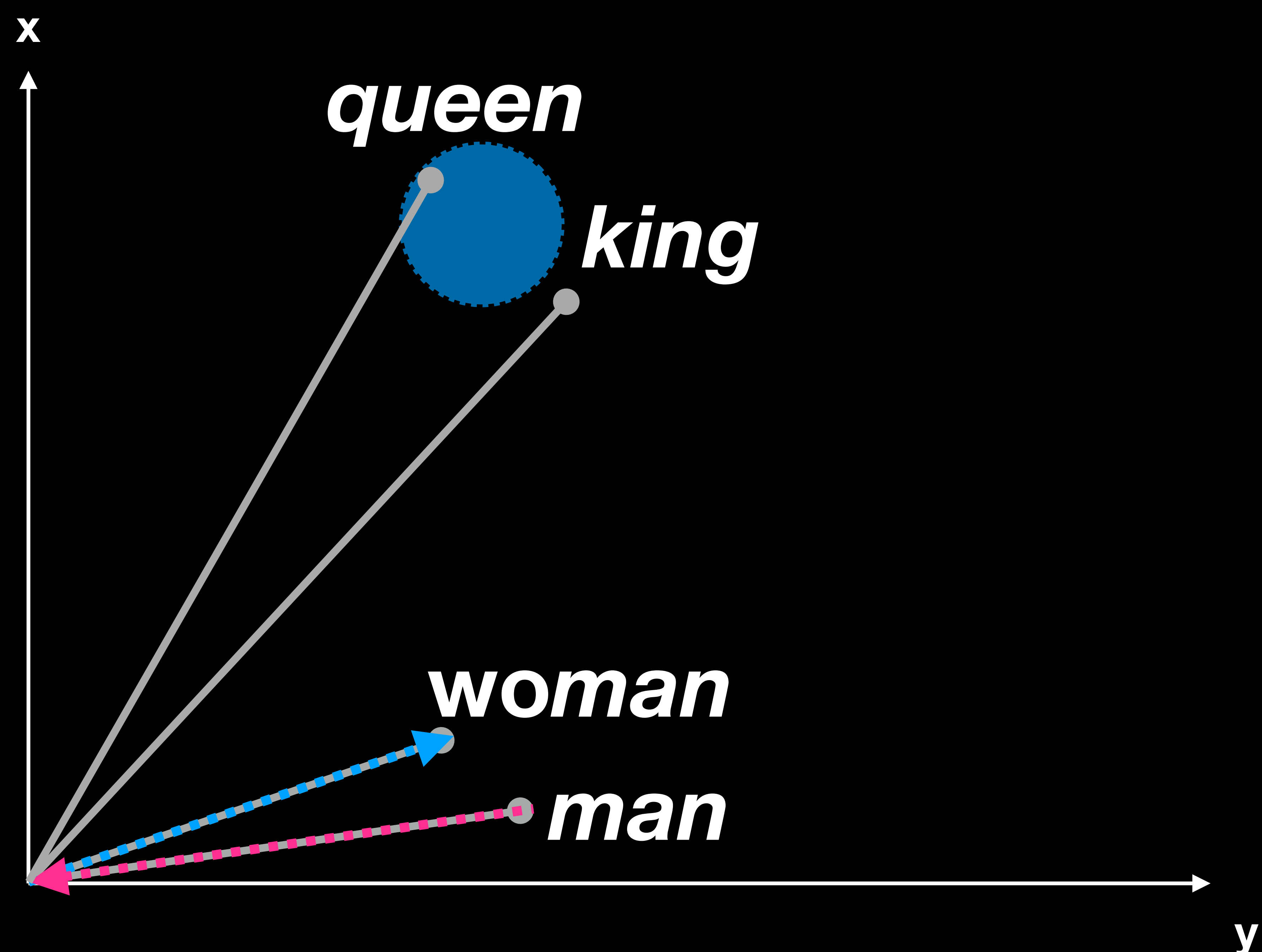
$$P(\texttt{time}\,|\,C) = P_{N0}(right\,|\,C)\cdot P_{N1}(left\,|\,C)\cdot P_{N2}(right\,|\,C) = 0.25$$

source: http://building-babylon.net/2017/08/01/hierarchical-softmax/

Bocconi

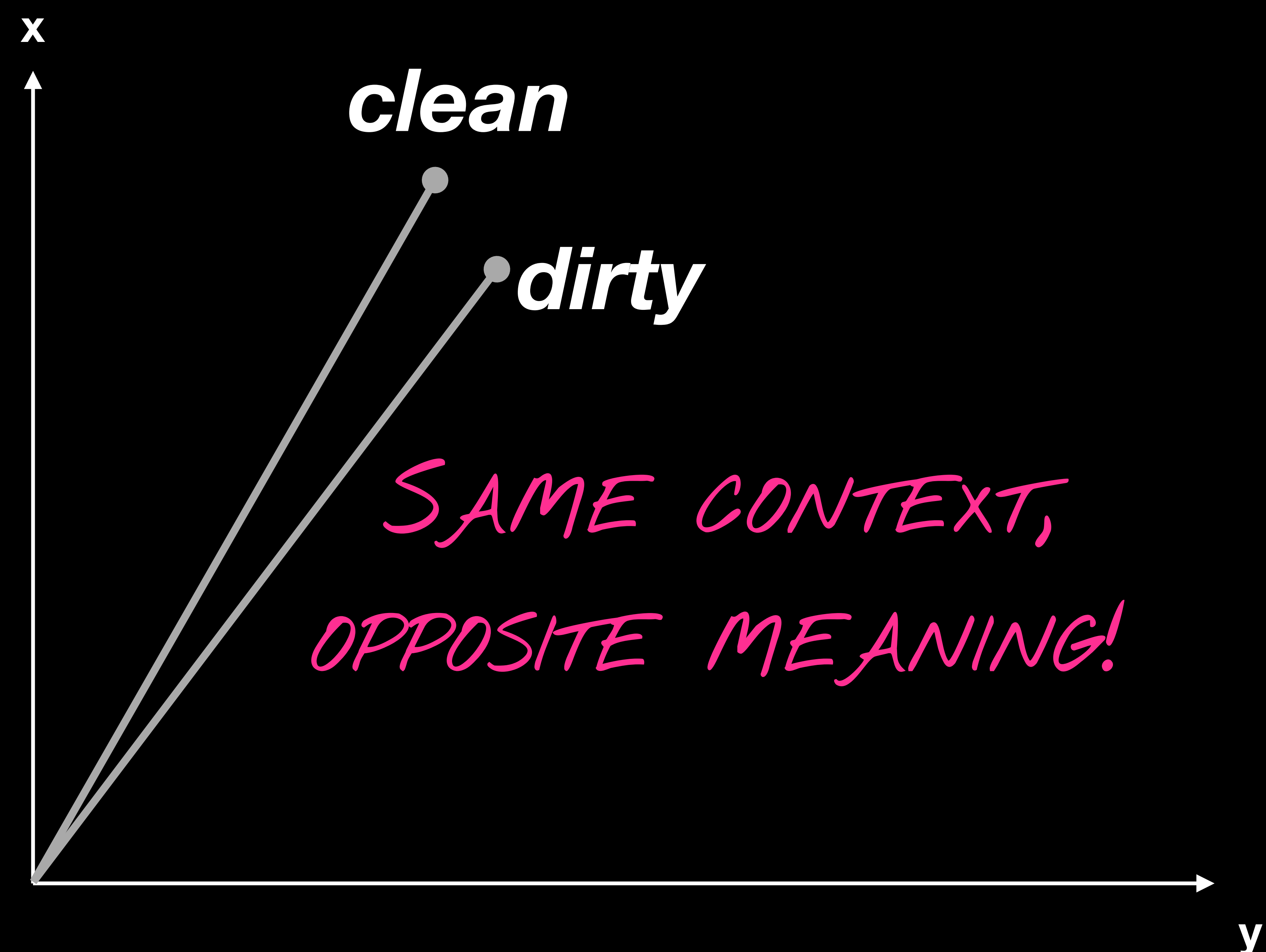
Vector Space Semantics

king – man + woman ≈ queen



Caveat: Antonyms

His kitchen was always very ___________

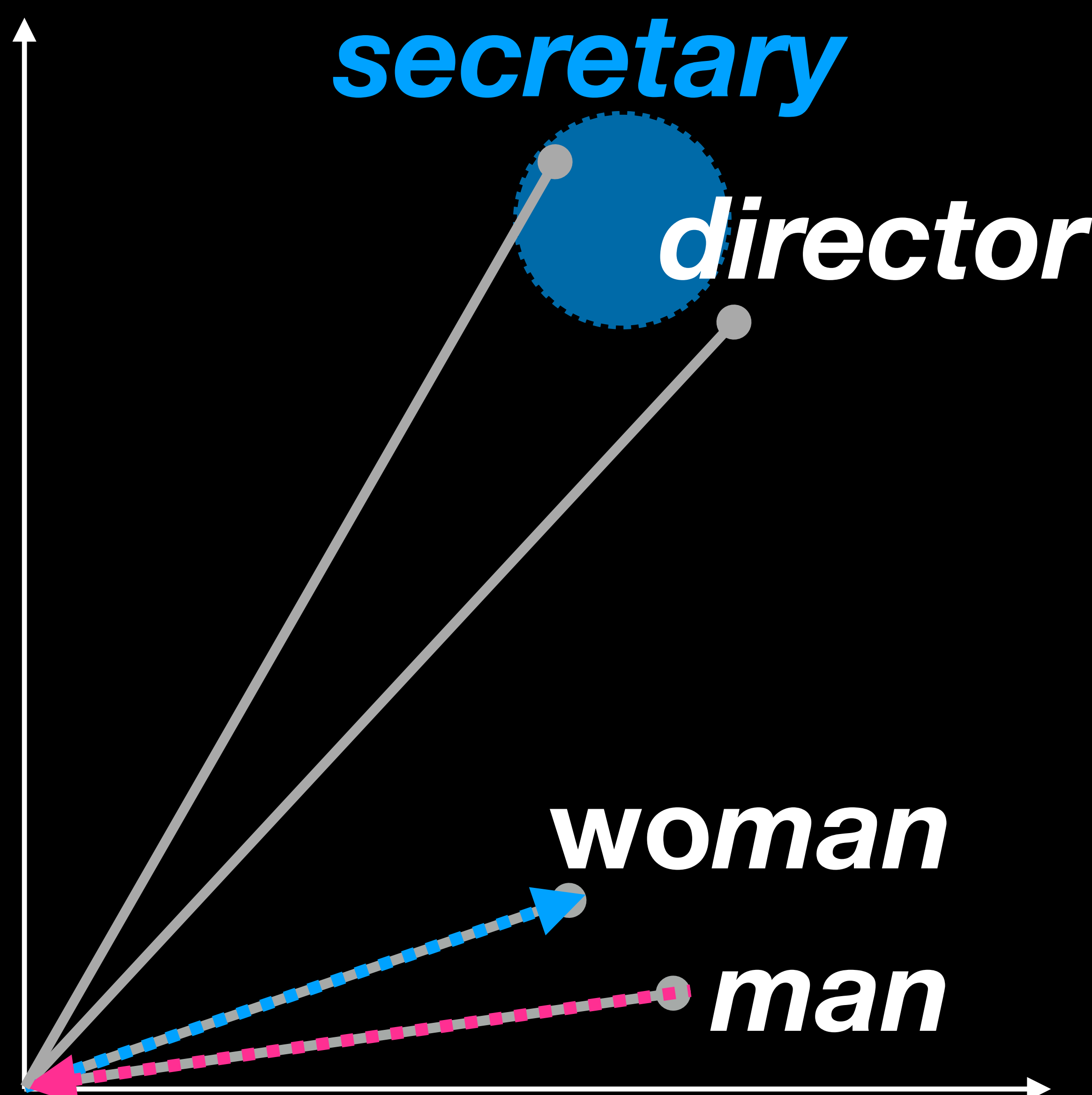


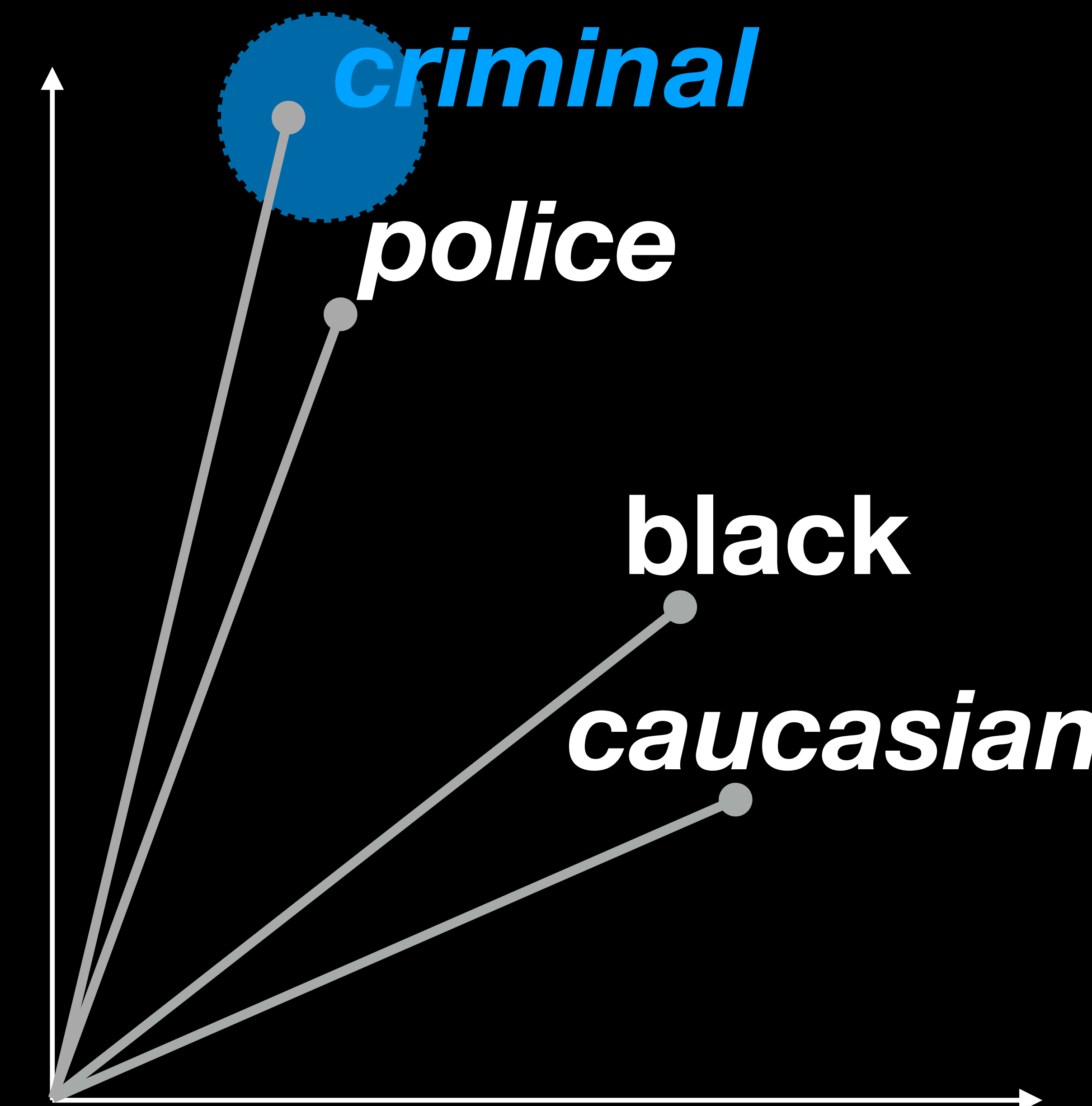
Bocconi

Bolukbasi et al. (2016)
Manzini et al. (2019)
Nissim et al. (2019)

Caveat: Bias

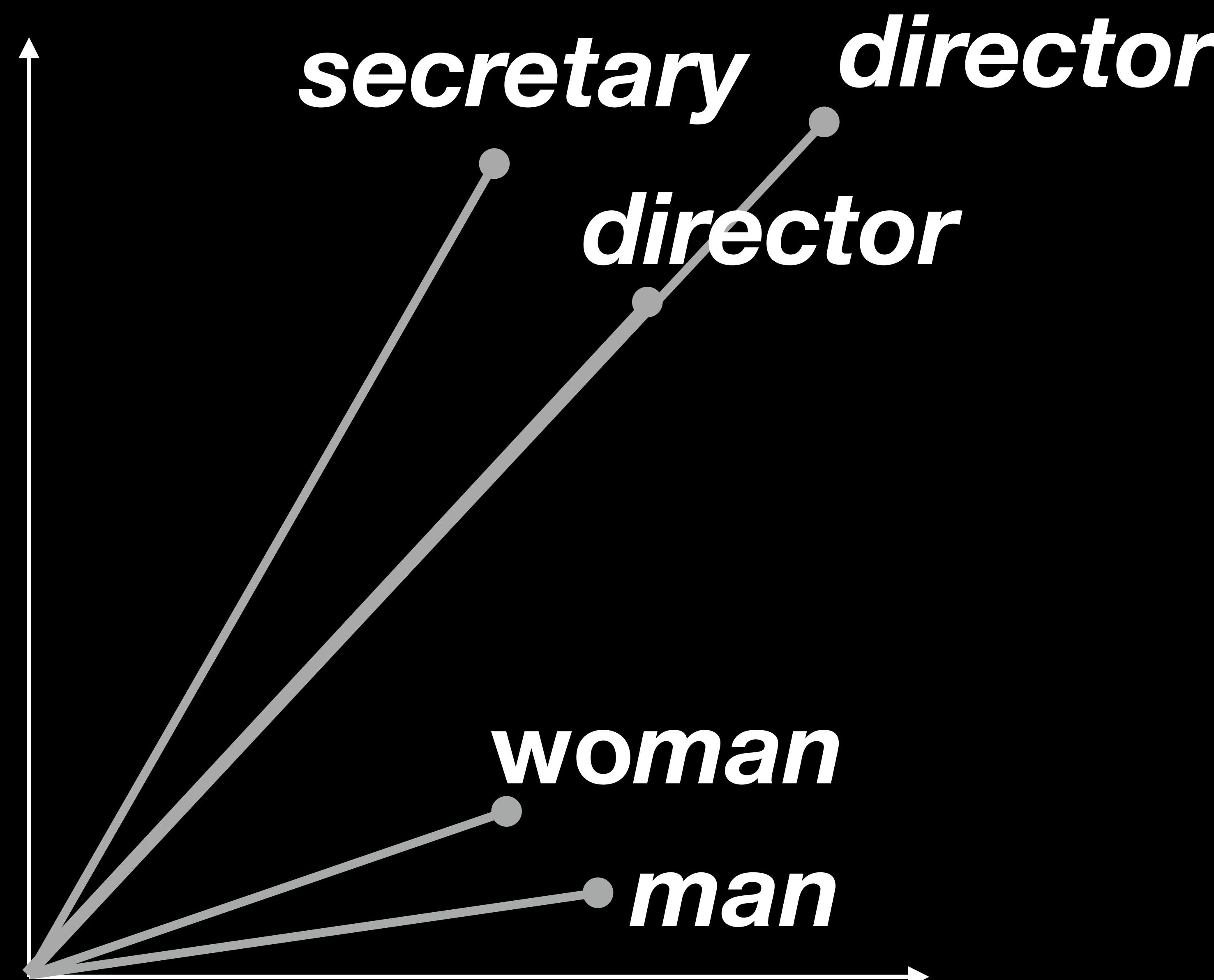
director – man + woman ≈ secretary
police – caucasian + black ≈ criminal





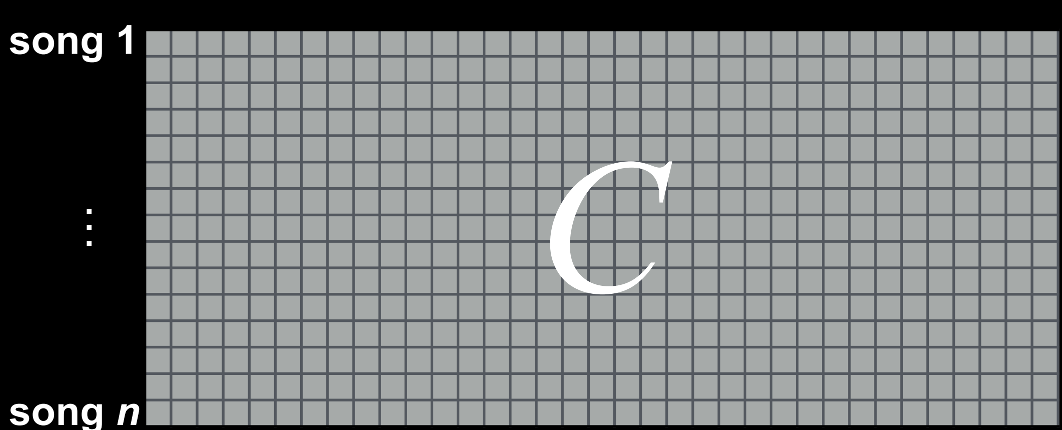
Bocconi

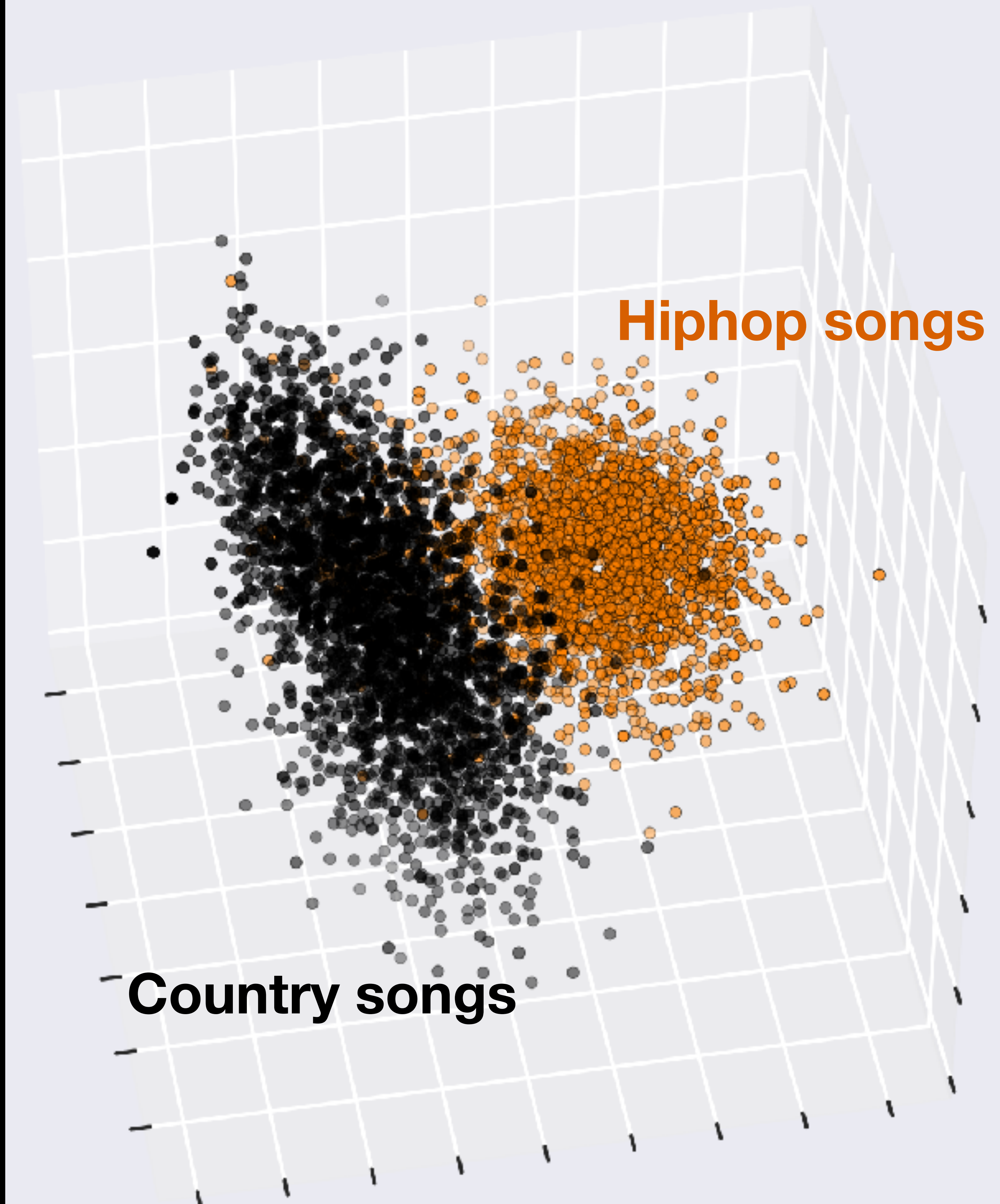
Debiasing Vectors



Part 2
Representing Documents as Vectors

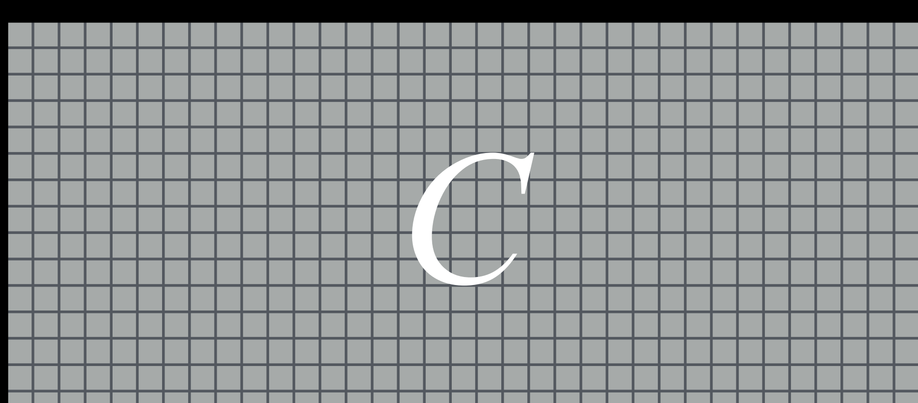
Example 1: Songs





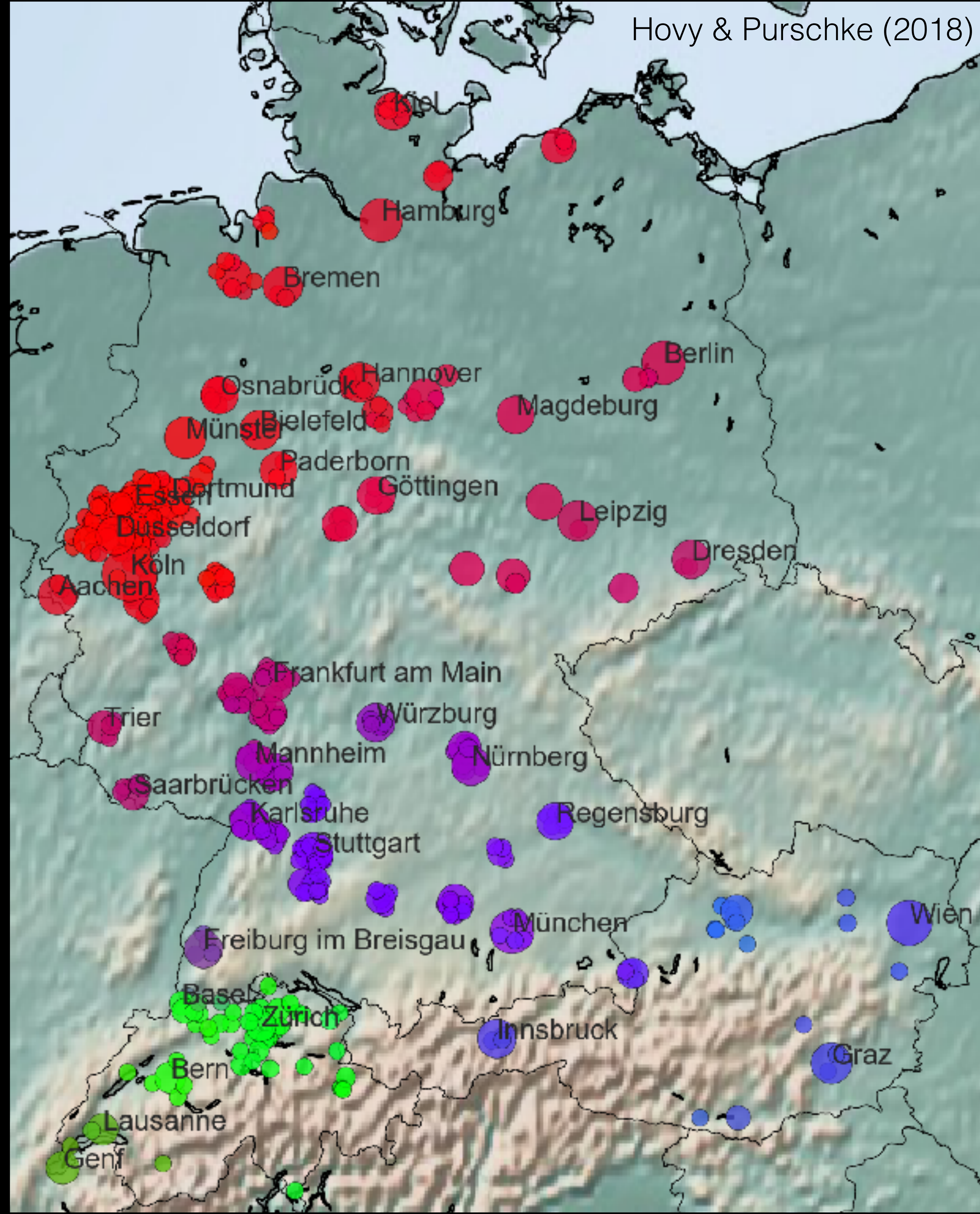
Example 2: Cities

city 1



⋮

city n



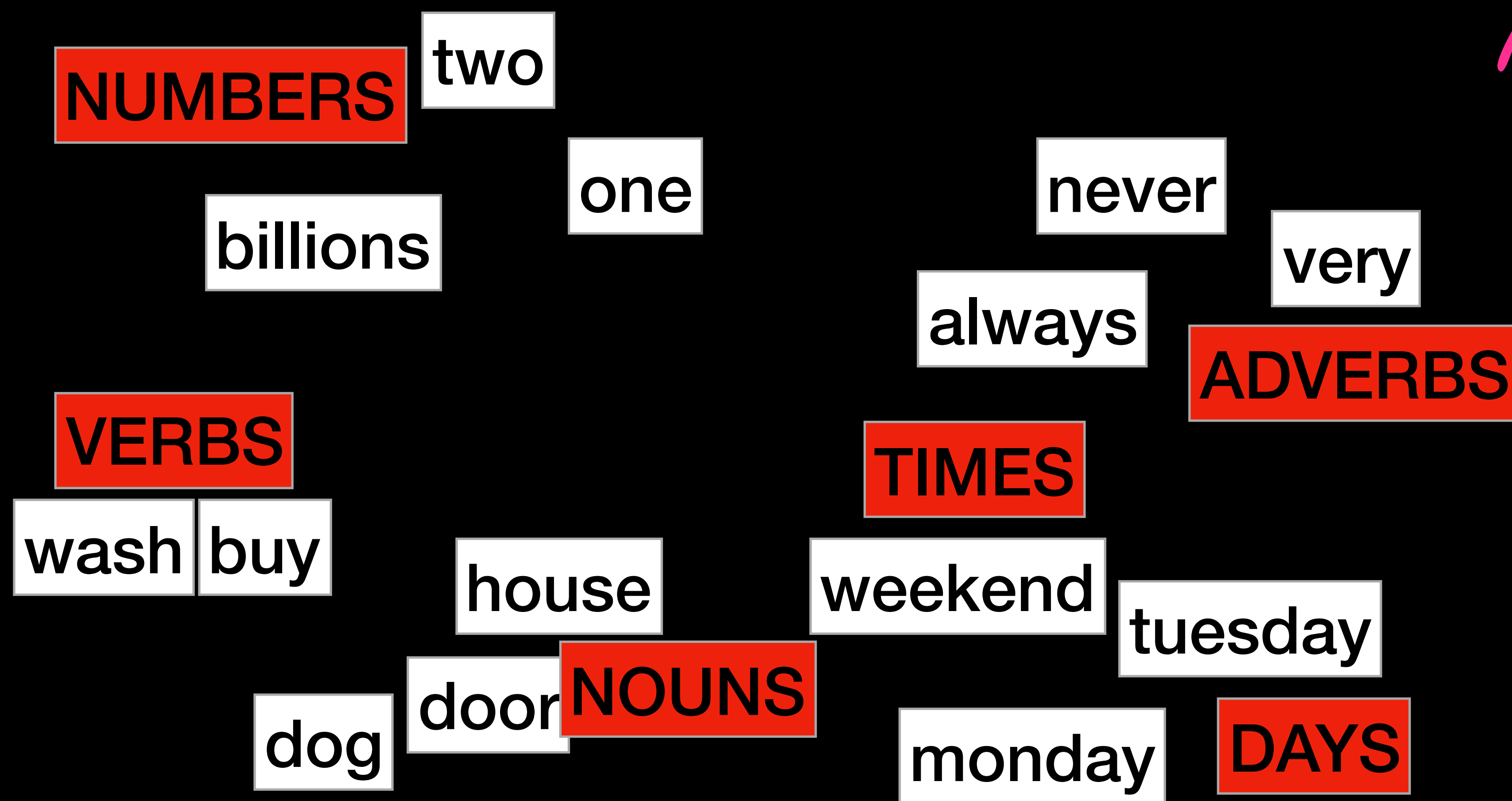
Hovy & Purschke (2018)

Doc2Vec – Intuitively

```
place words & cities randomly on fridge
for each pair of (word, city):
    if word seen in city:
        move closer together
    else:
        move further apart
```

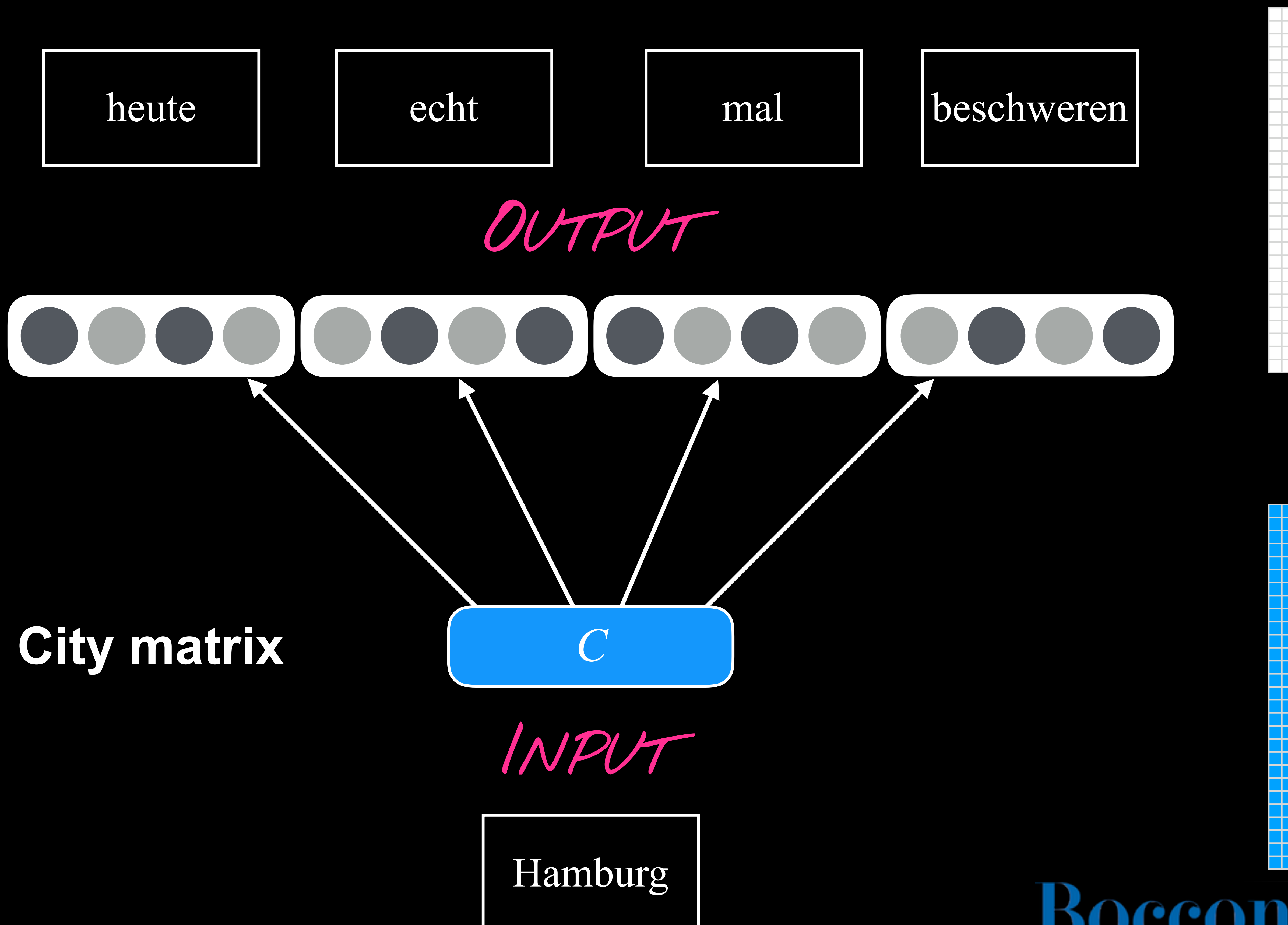
Adding Labels

I.E., CITIES, REGIONS, PEOPLE, ...

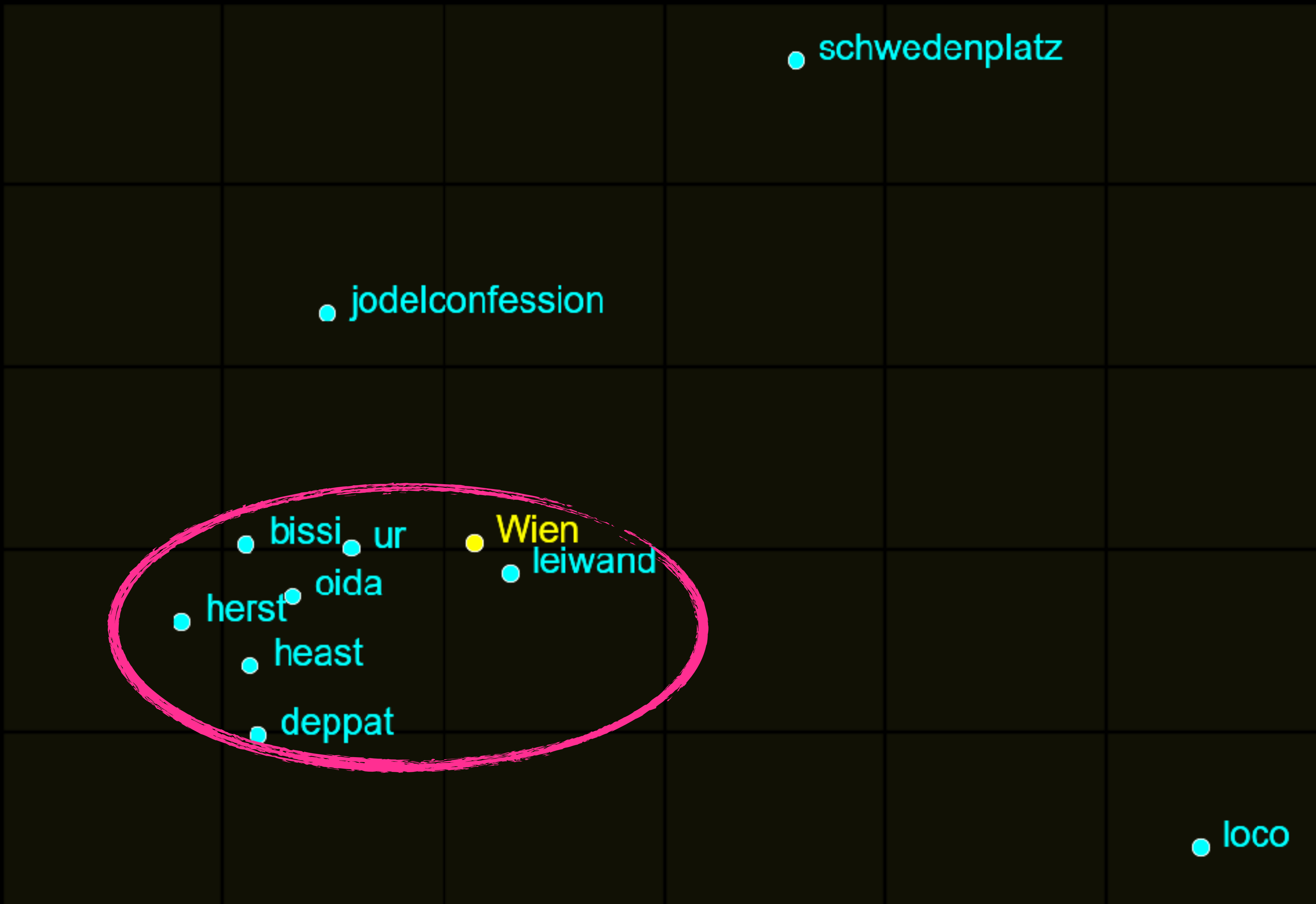


Le & Mikolov (2014)

Doc2Vec – Model



Words and Documents

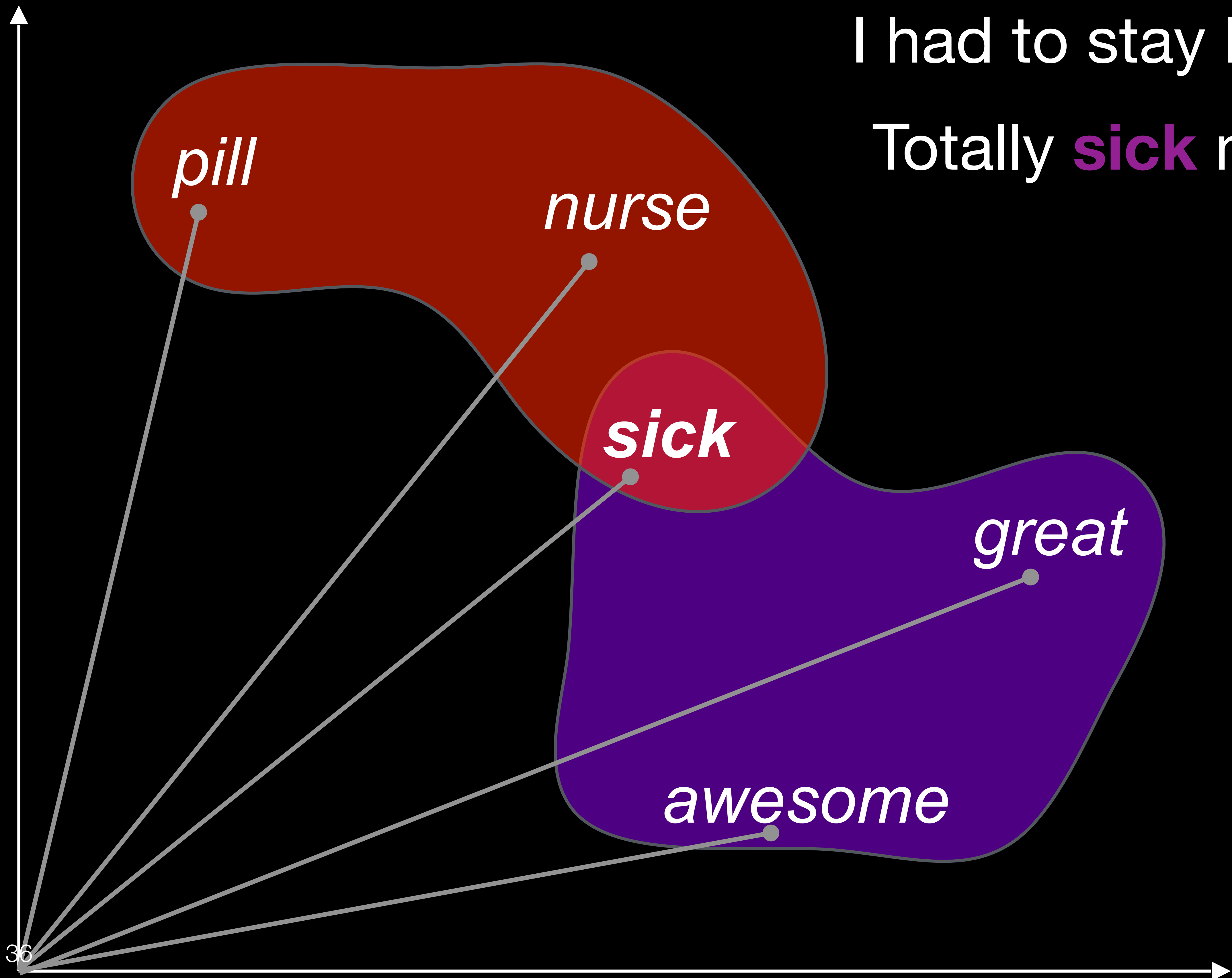


Preview:
Better, Contextualized Document Embeddings

Contextual Representations

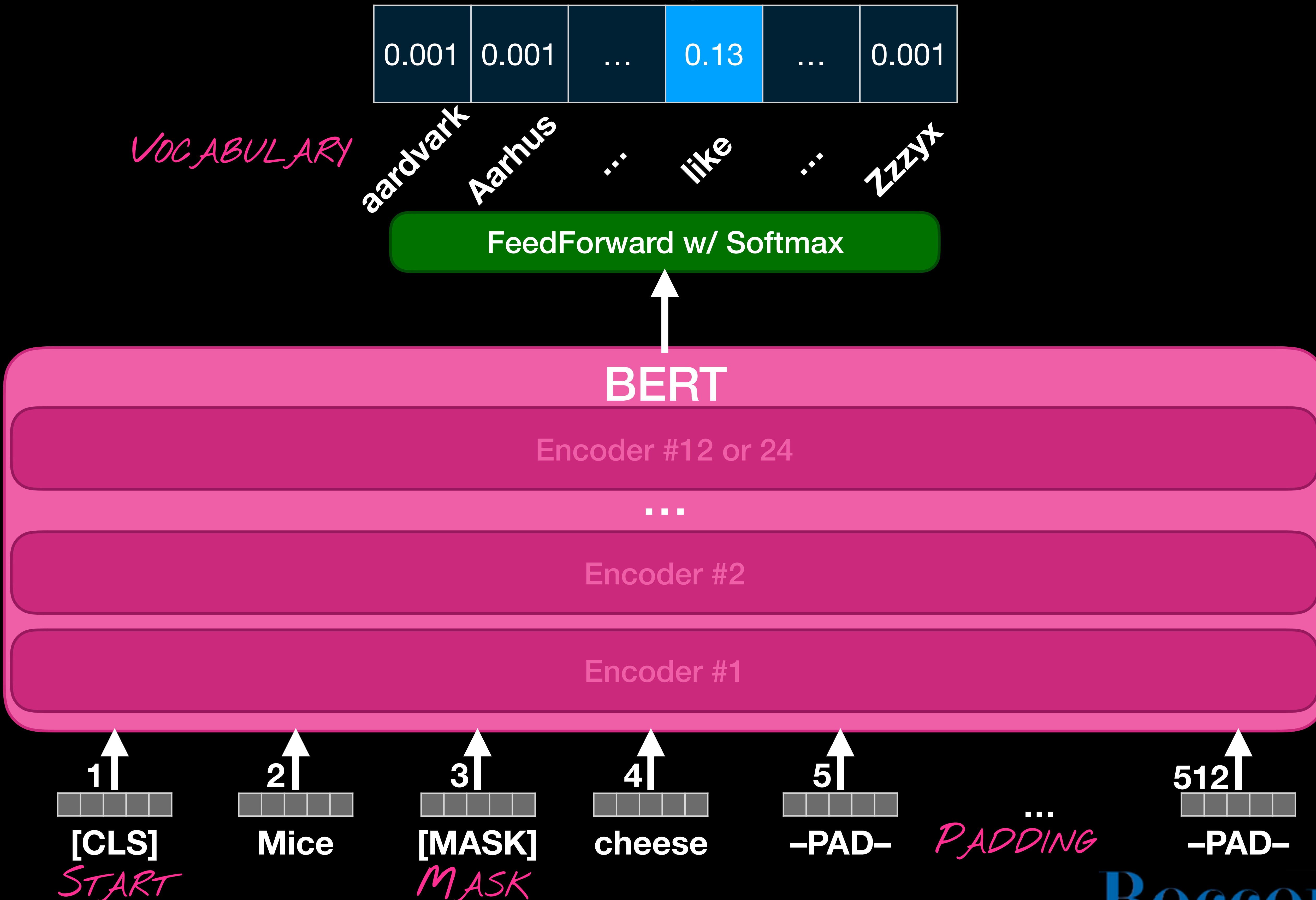
I had to stay home **sick**

Totally **sick** move, bro



Bocconi

Encoding Words



Wrapping up…

Representation Comparison

	Discrete	Distributed
#Dimensions	Data-dependent	Pre-defined
Content	Count-based	Coefficients
Density	Sparse	Dense
Strength	Interpretability	Similarity
Application	Understanding	Performance
School of thought	Rationalism	Empiricism

Bocconi

Take home points

- Text can be represented as dense, continuous embedding vectors
- Embedding models learn similarity via co-occurrence
- Word and document embeddings reflect semantic similarity in high-dimensional space
- Good for similarity, visualization, and classification, bad for analysis