

Today's Goals

- Understand the difference between **sparse** and **dense** representations
- Learn about **bag of words** (BOW) representations
- Learn about **word2vec** and **doc2vec**

Ham or Spam?

From: offr4u@rsph.com
Subject: Unique wealth offerings
To: dirk.hovy@unibocconi.it

Greetings dear friend

We have an amazing offer 4U: Click here to get access to a free consultation for serious wealth benefits! Urgent: offer expires soon.

Works guaranteed! Triple your income.

Ham or Spam?

From: offr4u@rsph.com
Subject: Unique wealth offerings
To: dirk.hovy@unibocconi.it

Greetings dear friend

We have an amazing offer 4U: Click here to get access to a free consultation for serious wealth benefits! Urgent: offer expires soon.

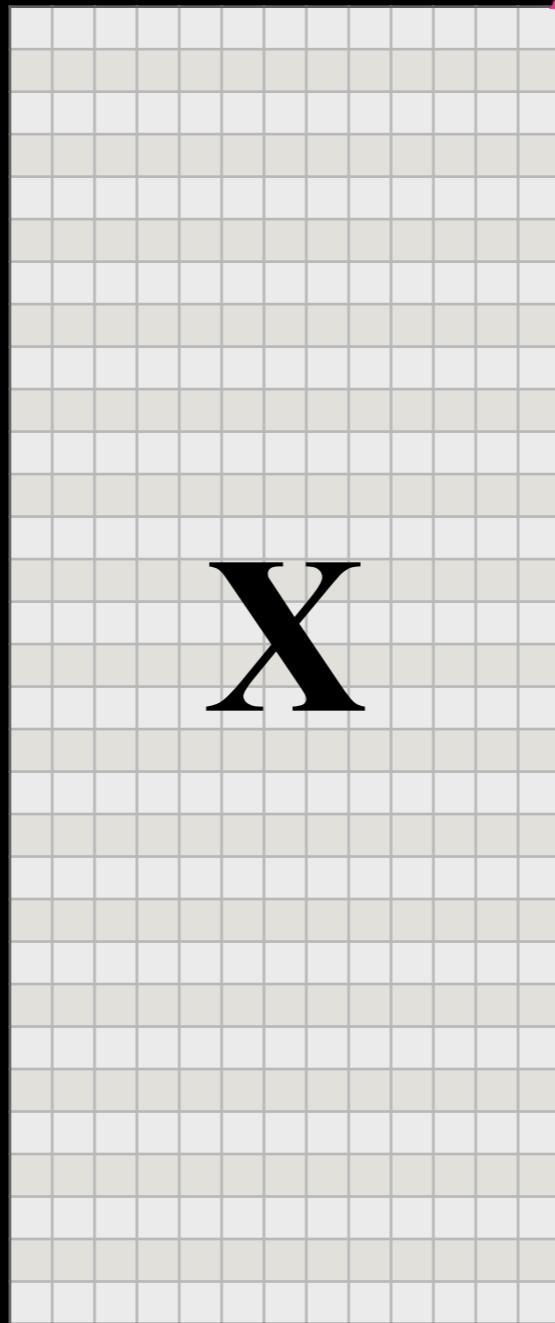
Works guaranteed! Triple your income.

Spam terms:

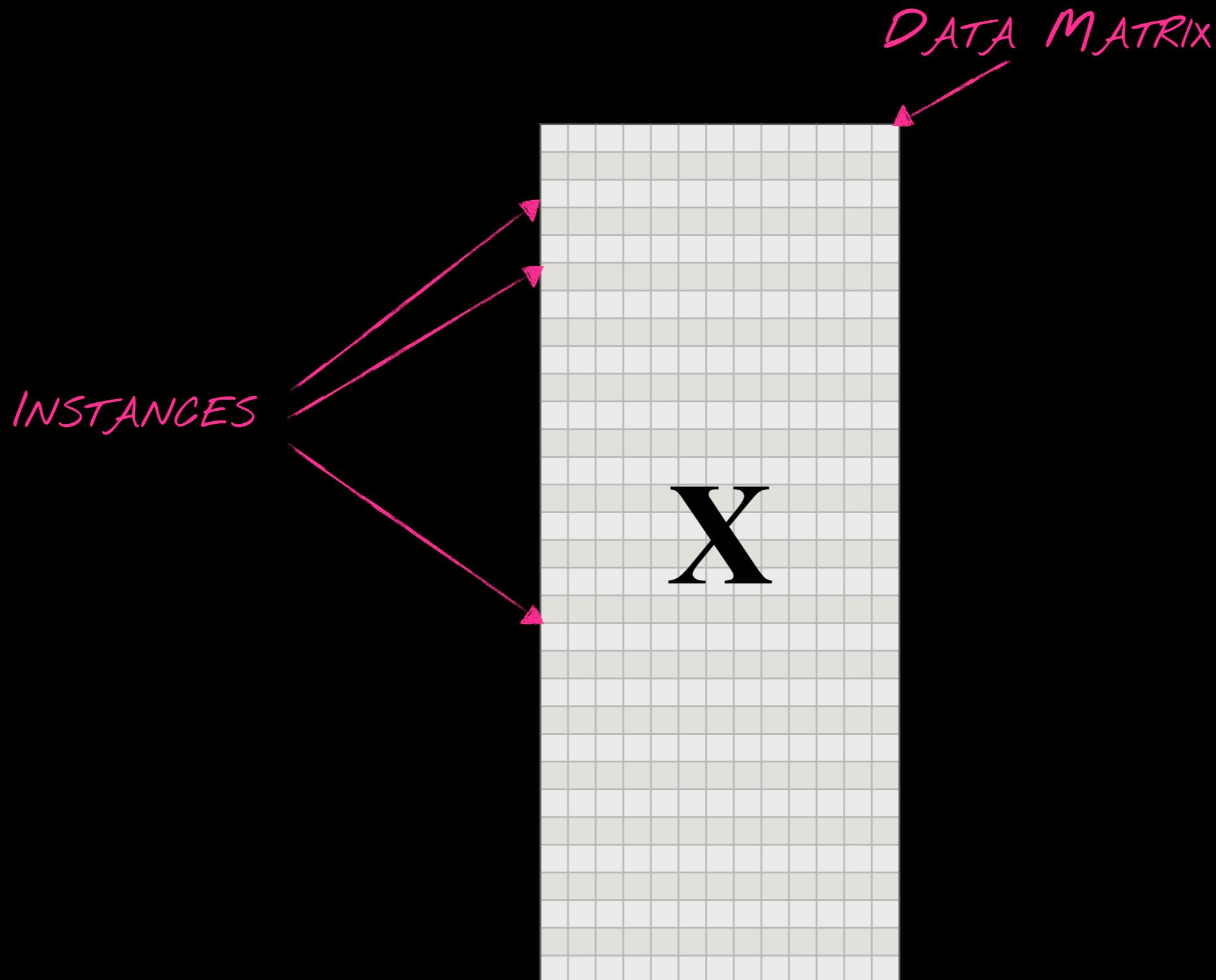
- 4U
- click
- amazing
- free
- guarantee
- offer
- urgent
- dear friend
- income
- serious

Terminology

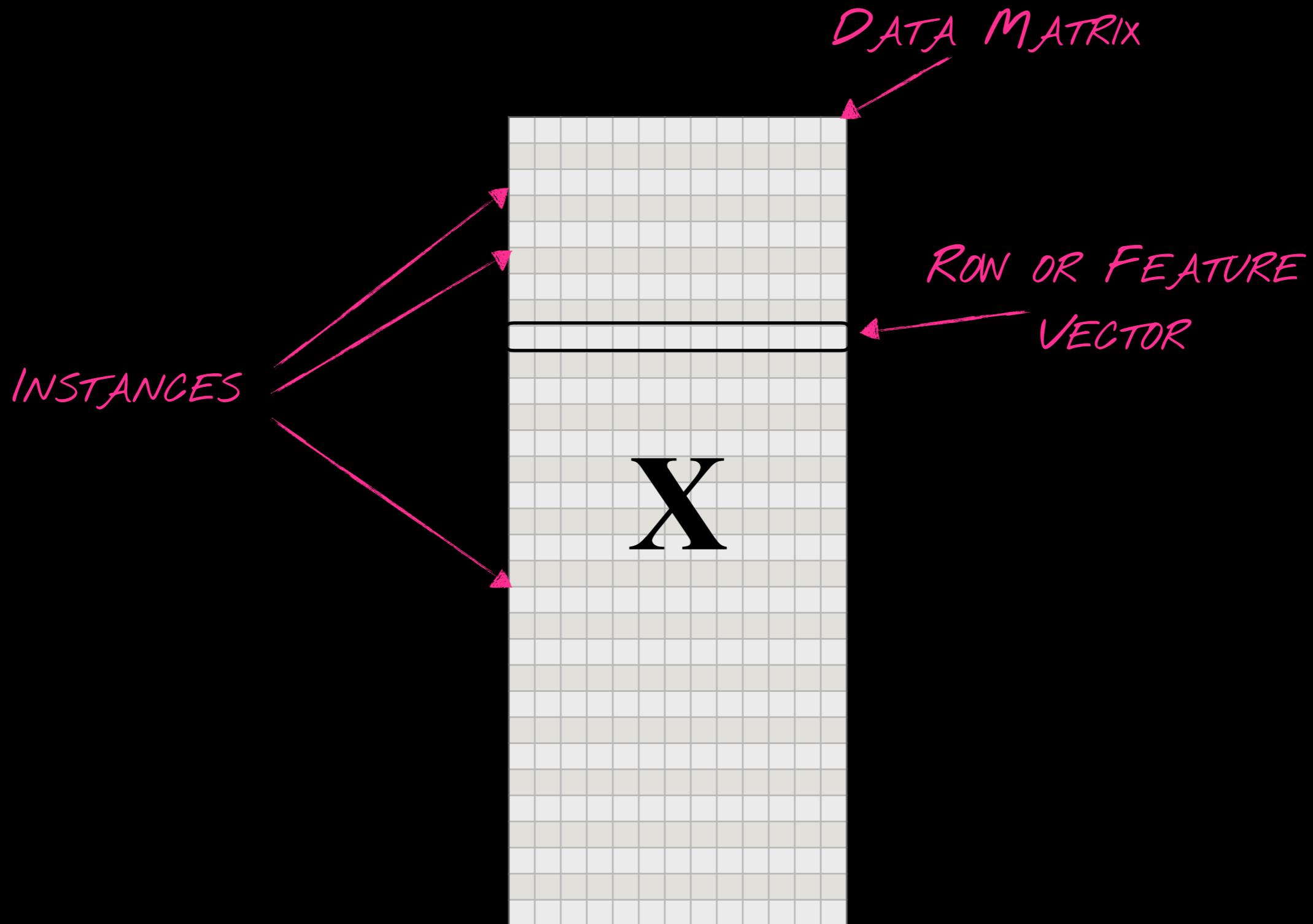
DATA MATRIX



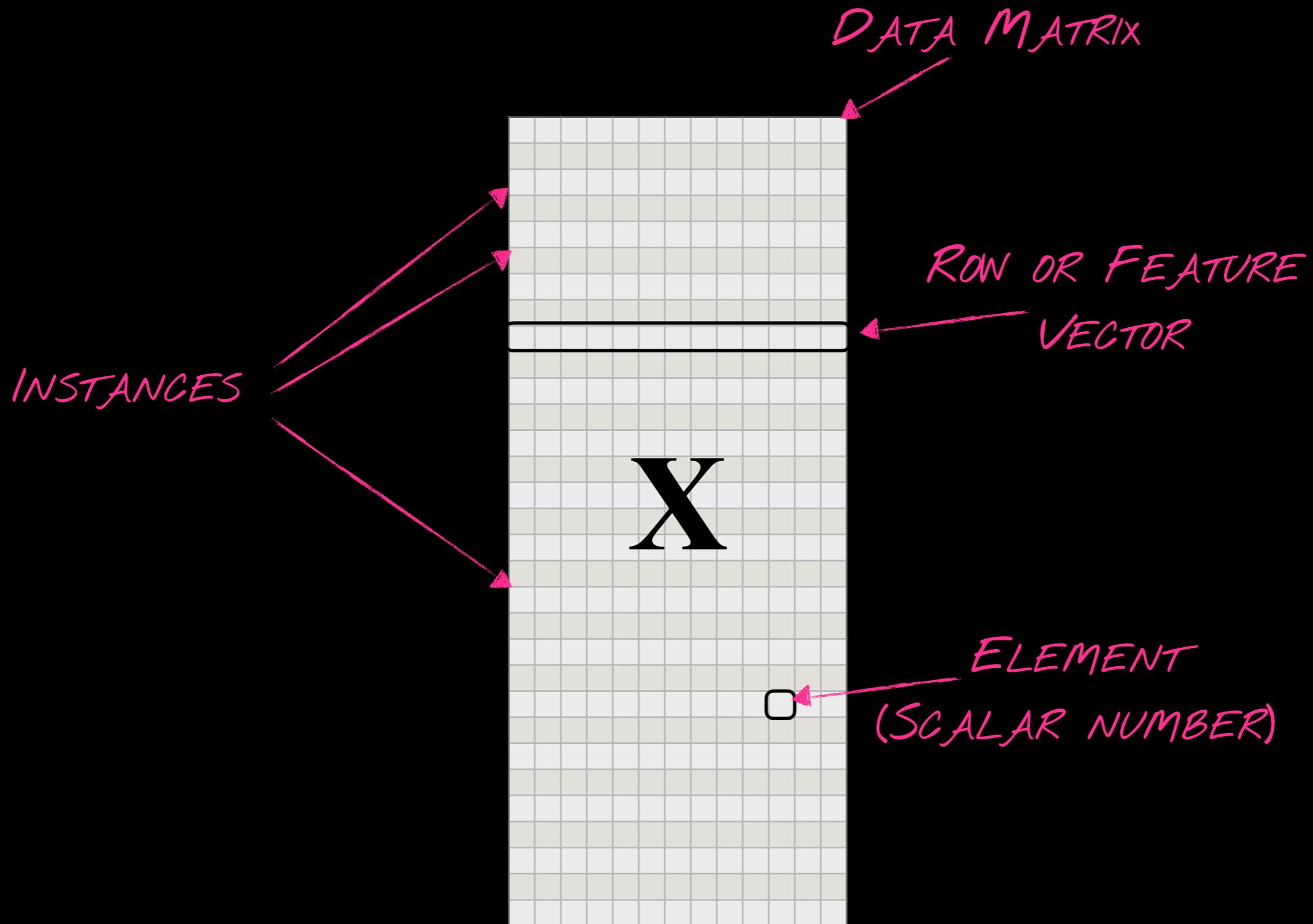
Terminology



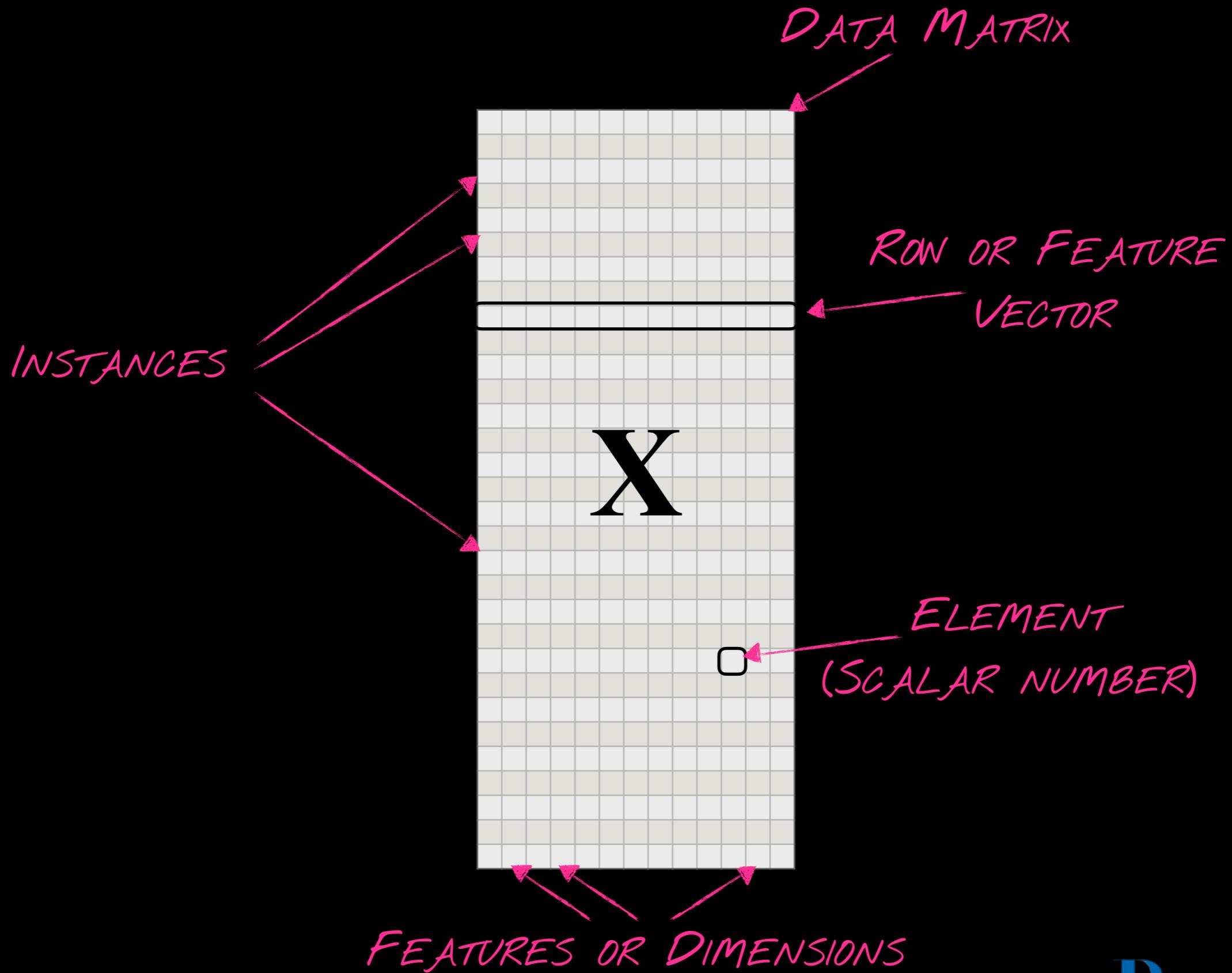
Terminology



Terminology



Terminology



Discrete Representations

Bags of words (BOW)

COUNT WORDS

```
{  
    'shakespeare': 6,  
    'in': 20,  
    'love': 6,  
    'is': ...  
}
```

shakespeare
in
...
love
beer

6	...	20	0	...	0	6	0	...	0
---	-----	----	---	-----	---	---	---	-----	---

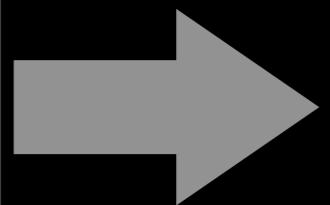
X

VECTORIZE FEATURES

Bags of words (BOW)

COUNT WORDS

```
{  
    'shakespeare': 6,  
    'in': 20,  
    'love': 6,  
    'is': ...  
}
```



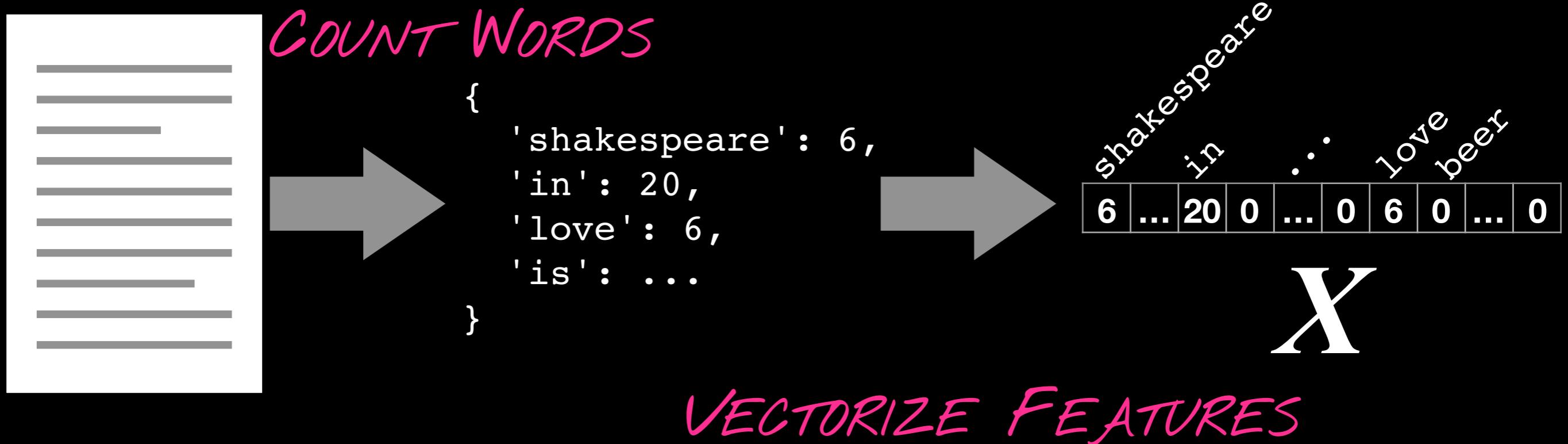
VECTORIZE FEATURES

shakespeare
in
...
love
beer

6	...	20	0	...	0	6	0	...	0
---	-----	----	---	-----	---	---	---	-----	---

X

Bags of words (BOW)



Quiz!

What happens if we allow *every possible word* to constitute a feature?

Quiz!

What happens if we allow *every possible word* to constitute a feature?

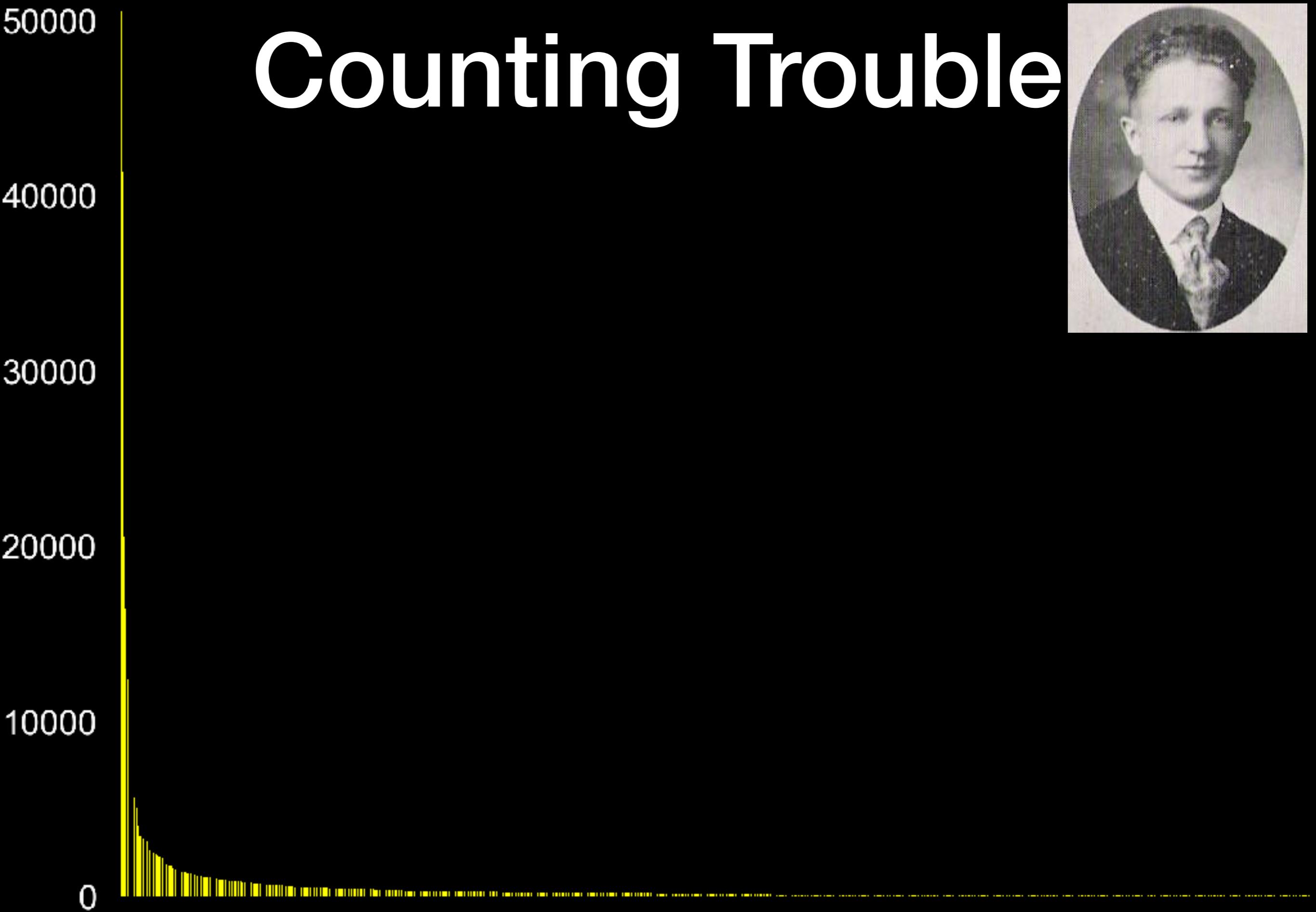
Expensive computation, and vectors have too many zeros.

Quiz!

What happens if we allow *every possible word* to constitute a feature?

Expensive computation, and vectors have too many zeros.
Limit to most frequent/informative words!

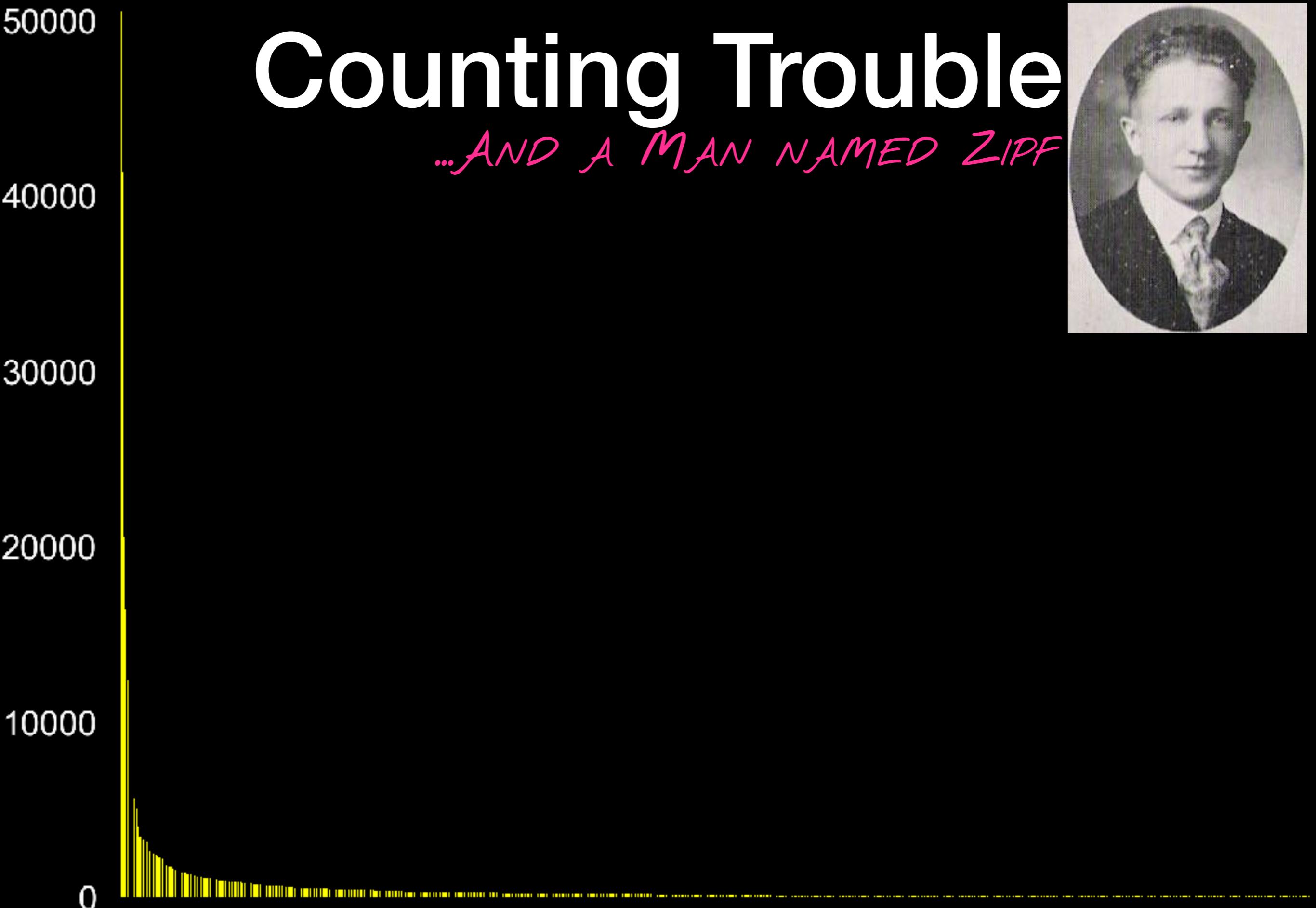
Counting Trouble



Bocconi

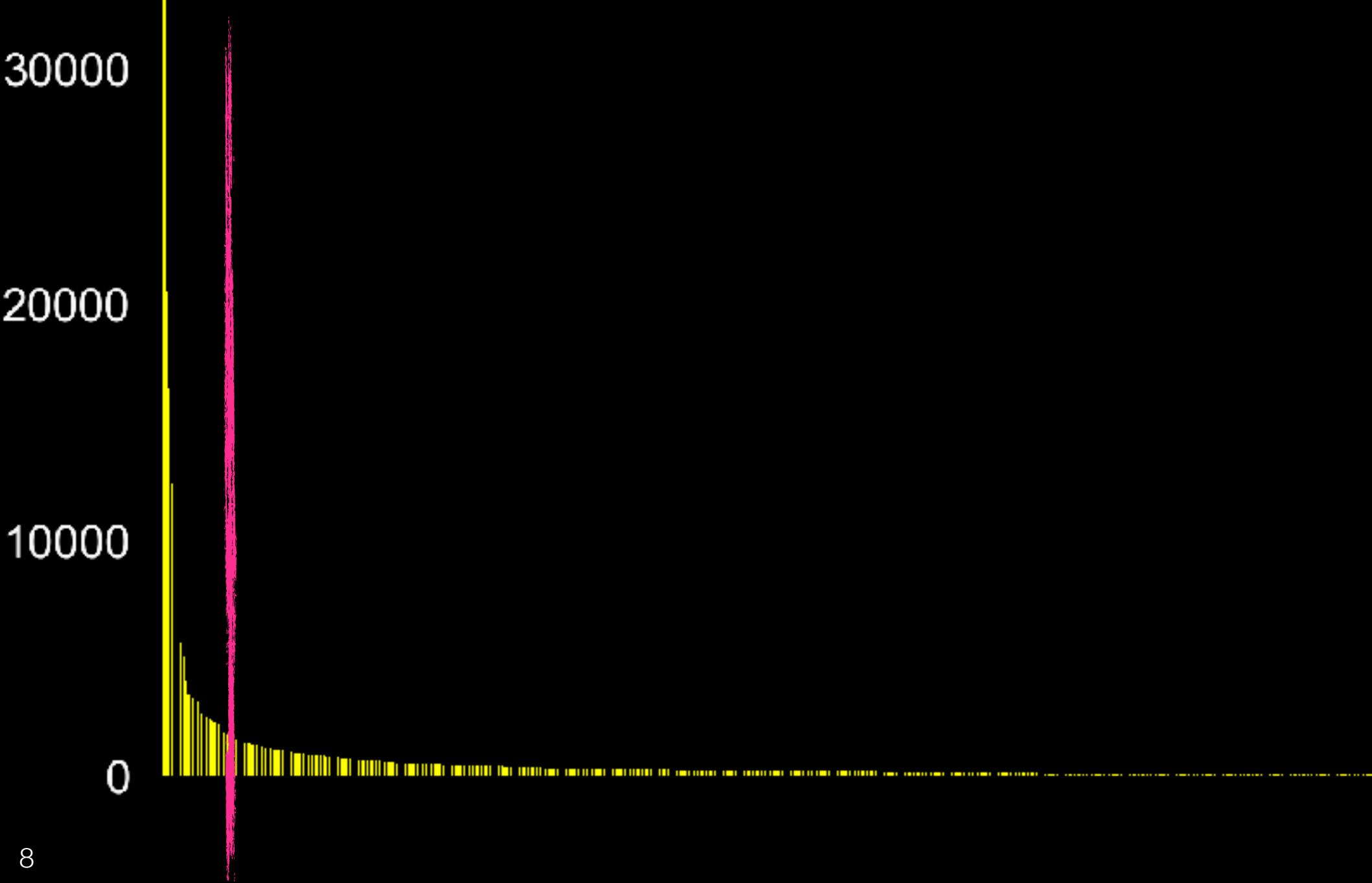
Counting Trouble

...AND A MAN NAMED ZIPF



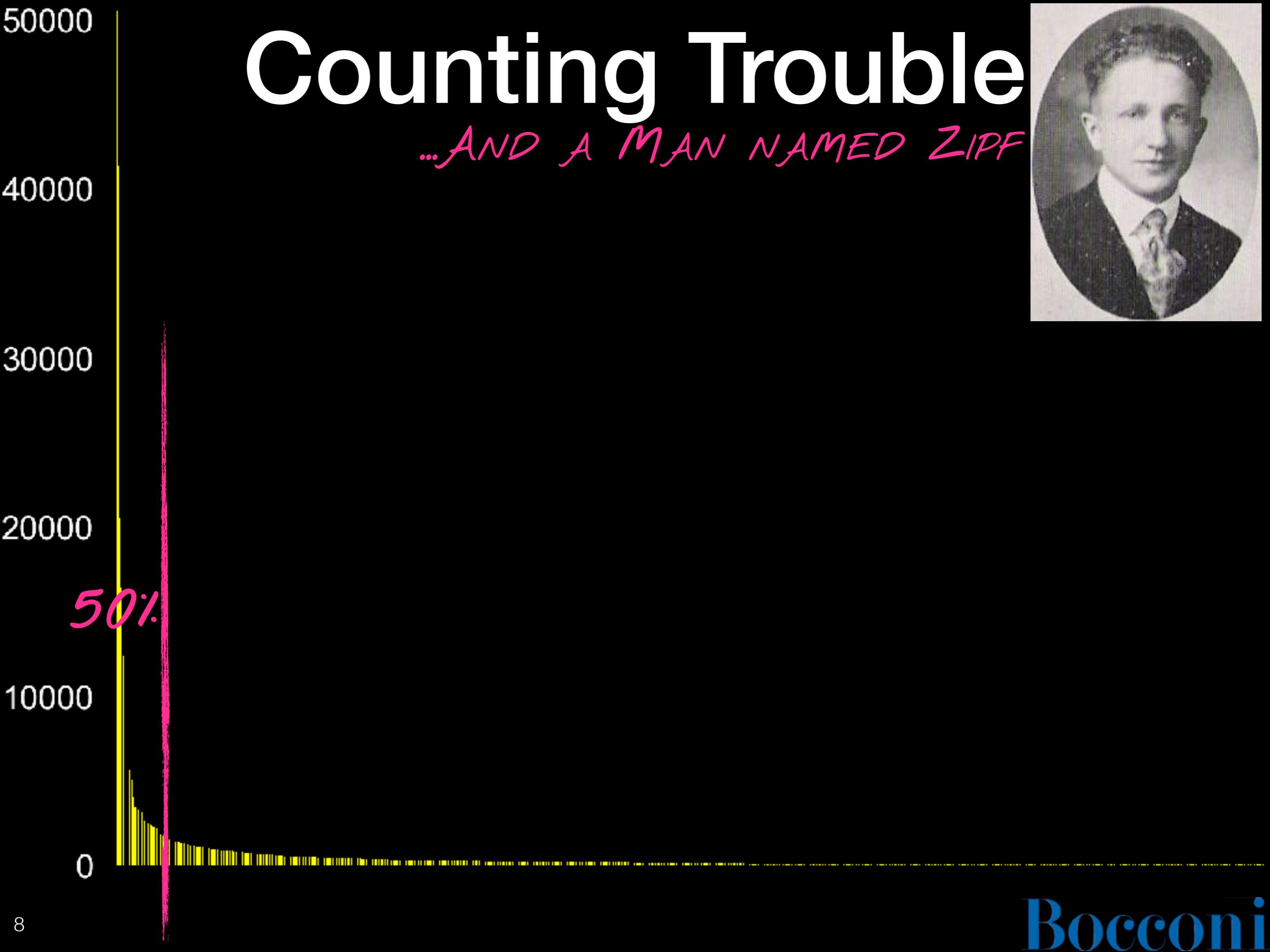
Counting Trouble

...AND A MAN NAMED ZIPF



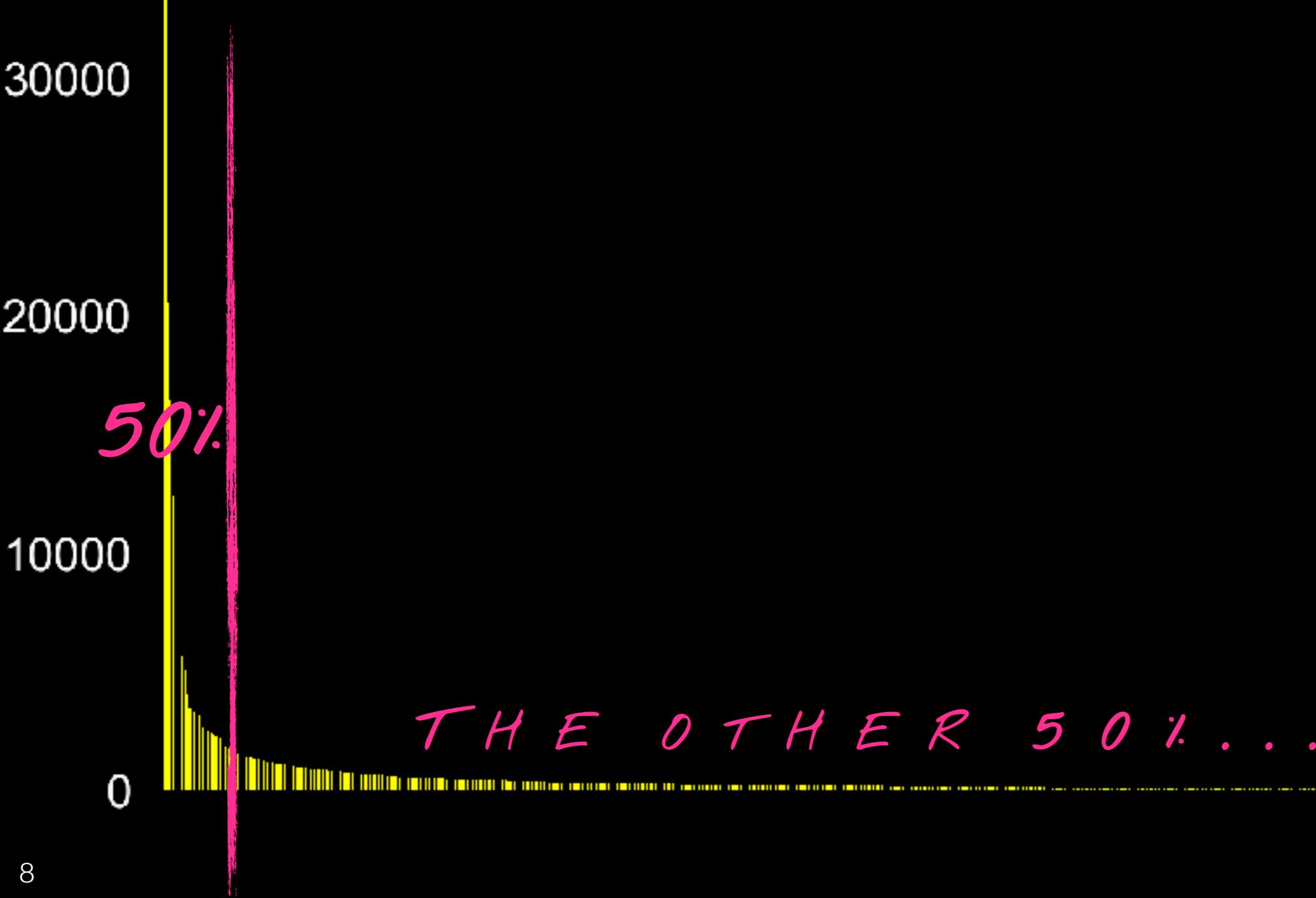
Counting Trouble

...AND A MAN NAMED ZIPF



Counting Trouble

...AND A MAN NAMED ZIPF



N-grams

"As Gregor Samsa awoke one morning from uneasy dreams, he found himself transformed in his bed into a gigantic insect-like creature."

N-grams

"As Gregor Samsa awoke one morning from uneasy dreams, he found himself transformed in his bed into a gigantic insect-like creature."

Unigrams As, Gregor, Samsa, awoke, one, morning, from,
uneasy, dreams, ...

N-grams

"As Gregor Samsa awoke one morning from uneasy dreams, he found himself transformed in his bed into a gigantic insect-like creature."

Unigrams **As, Gregor, Samsa, awoke, one, morning, from,**
uneasy, dreams, ...

Bigrams **As_Gregor, Gregor_Samsa, Samsa_awoke, awoke_one,**
one_morning, ...

N-grams

"As Gregor Samsa awoke one morning from uneasy dreams, he found himself transformed in his bed into a gigantic insect-like creature."

Unigrams As, Gregor, Samsa, awoke, one, morning, from,
uneasy, dreams, ...

Bigrams As_Gregor, Gregor_Samsa, Samsa_awoke, awoke_one,
one_morning, ...

Trigrams As_Gregor_Samsa, Gregor_Samsa_awoke,
Samsa_awoke_one, awoke_one_morning, ...

N-grams

"As Gregor Samsa awoke one morning from uneasy dreams, he found himself transformed in his bed into a gigantic insect-like creature."

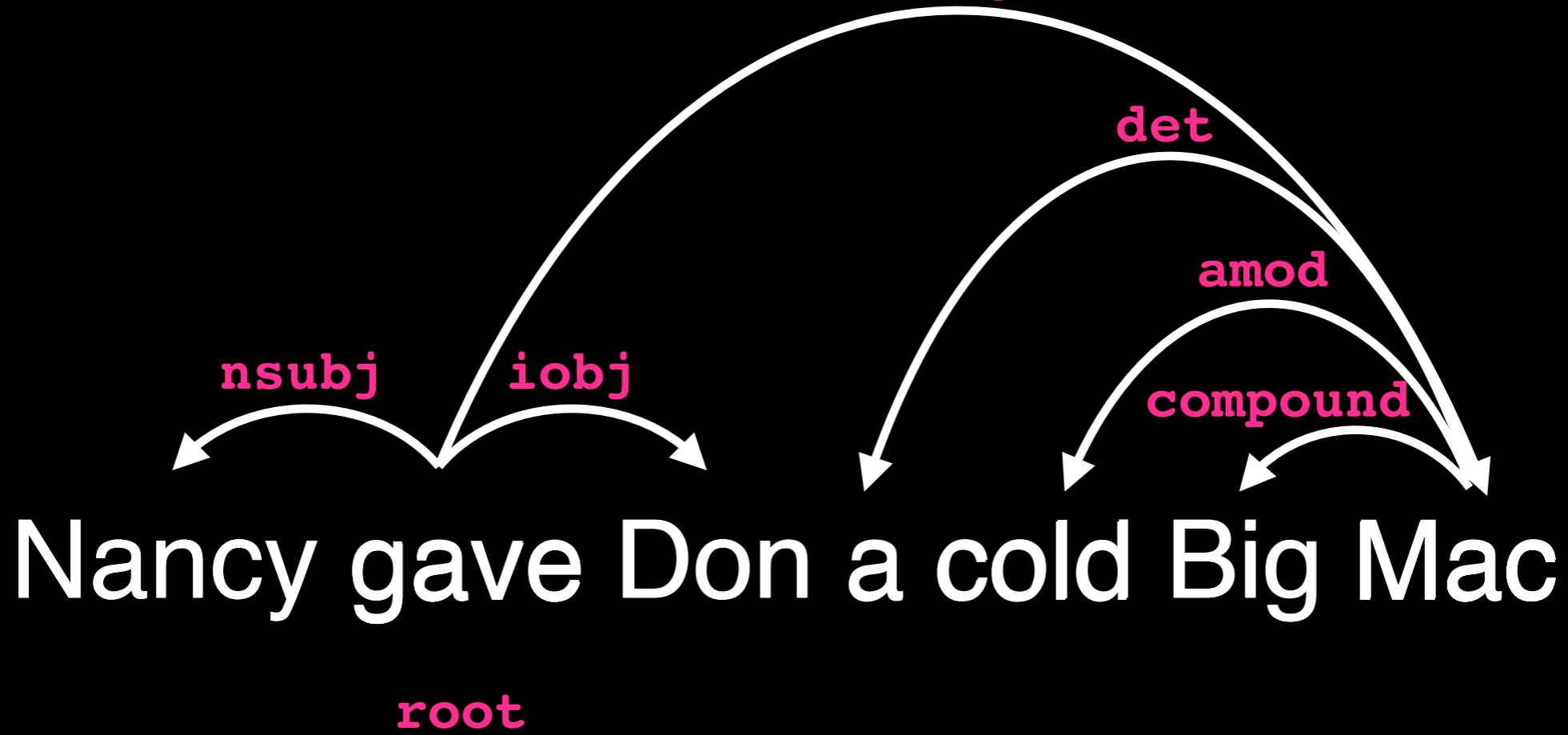
Unigrams As, Gregor, Samsa, awoke, one, morning, from,
uneasy, dreams, ...

Bigrams As_Gregor, Gregor_Samsa, Samsa_awoke, awoke_one,
one_morning, ...

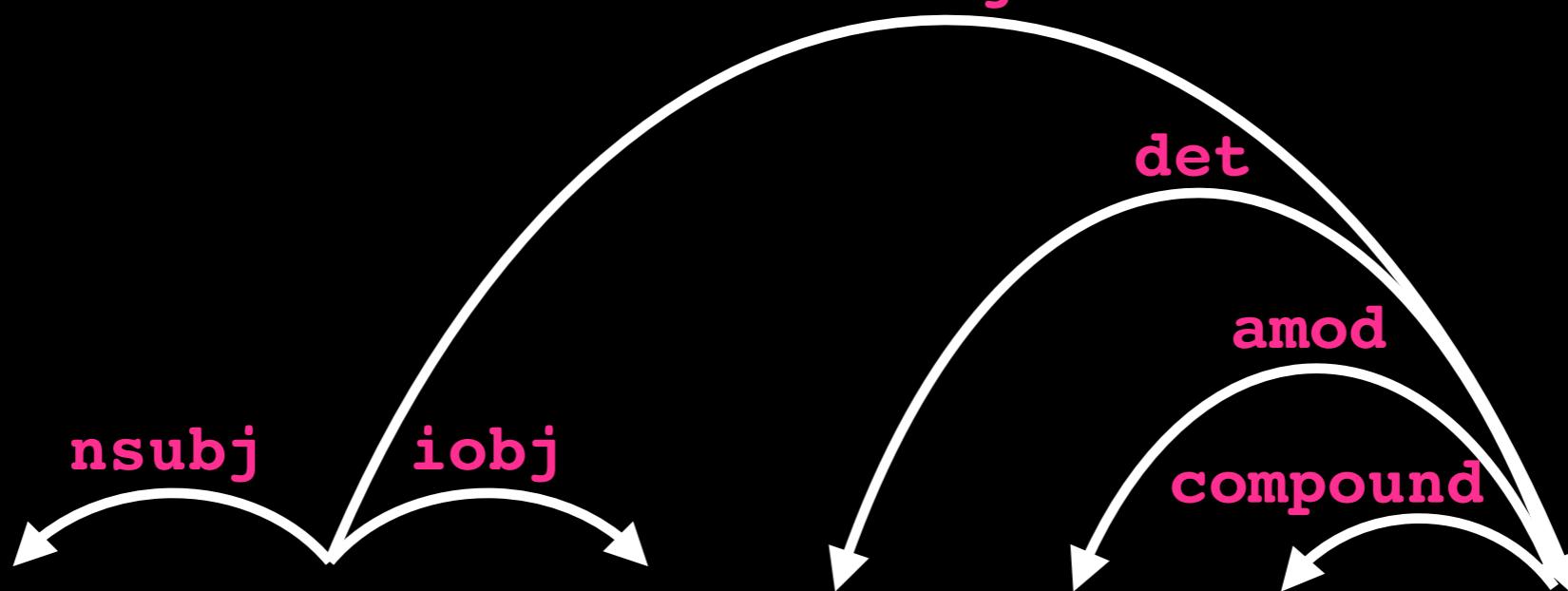
Trigrams As_Gregor_Samsa, Gregor_Samsa_awoke,
Samsa_awoke_one, awoke_one_morning, ...

4-grams As_Gregor_Samsa_awoke, Gregor_Samsa_awoke_one,
Samsa_awoke_one_morning, ...

Dependency n -grams



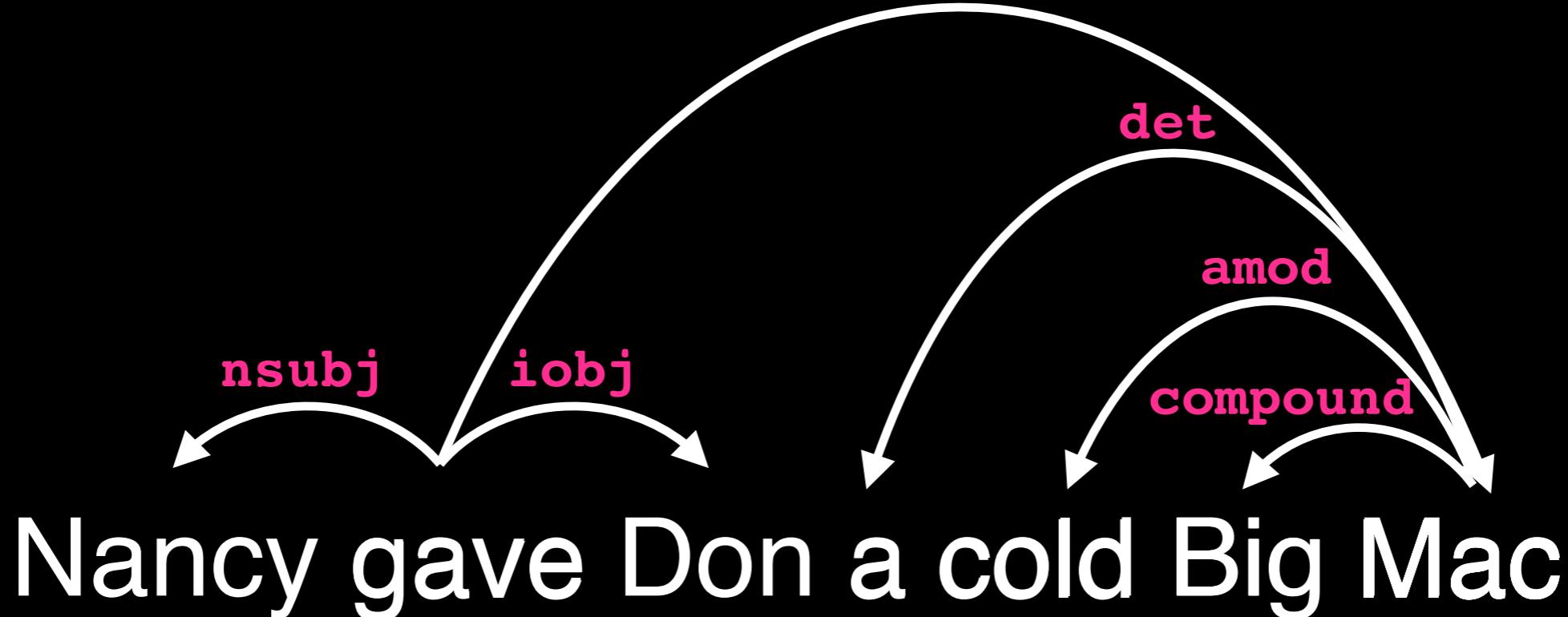
Dependency n -grams



Nancy gave Don a cold Big Mac

Nancy gave

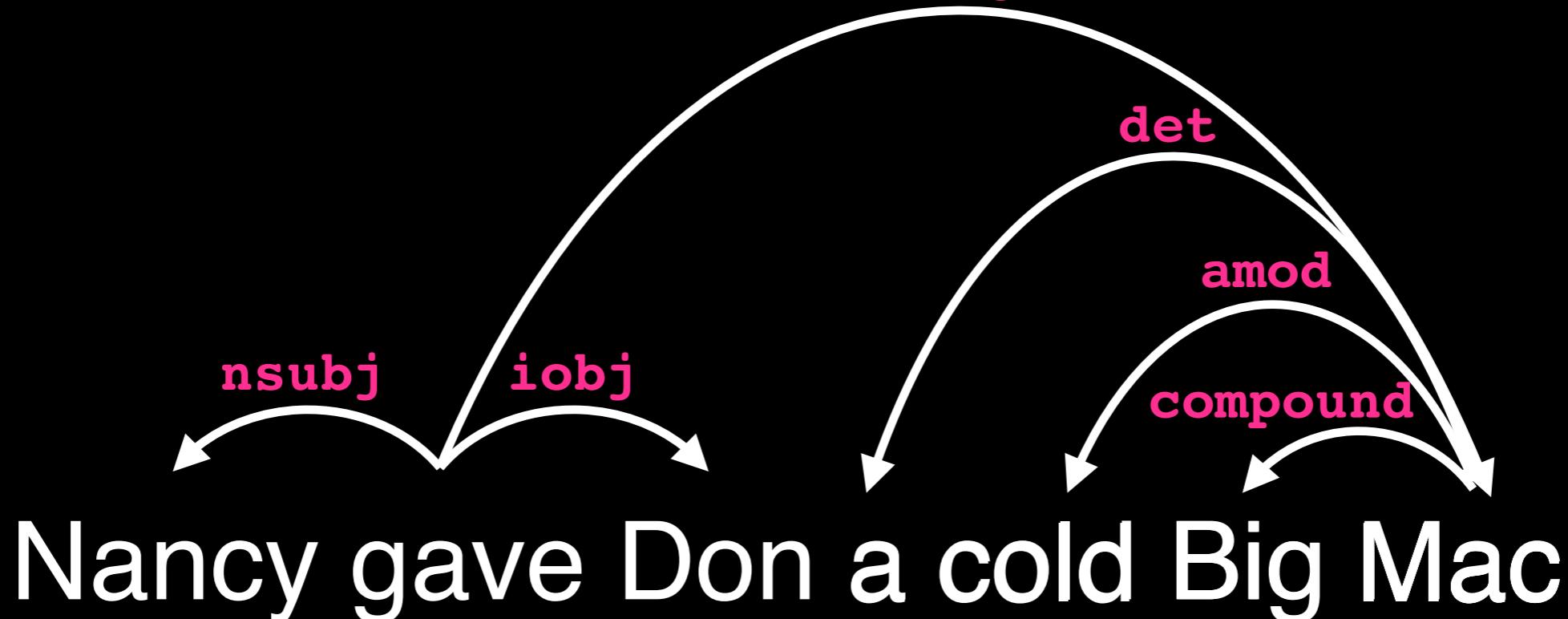
Dependency n -grams



Nancy gave
Don gave

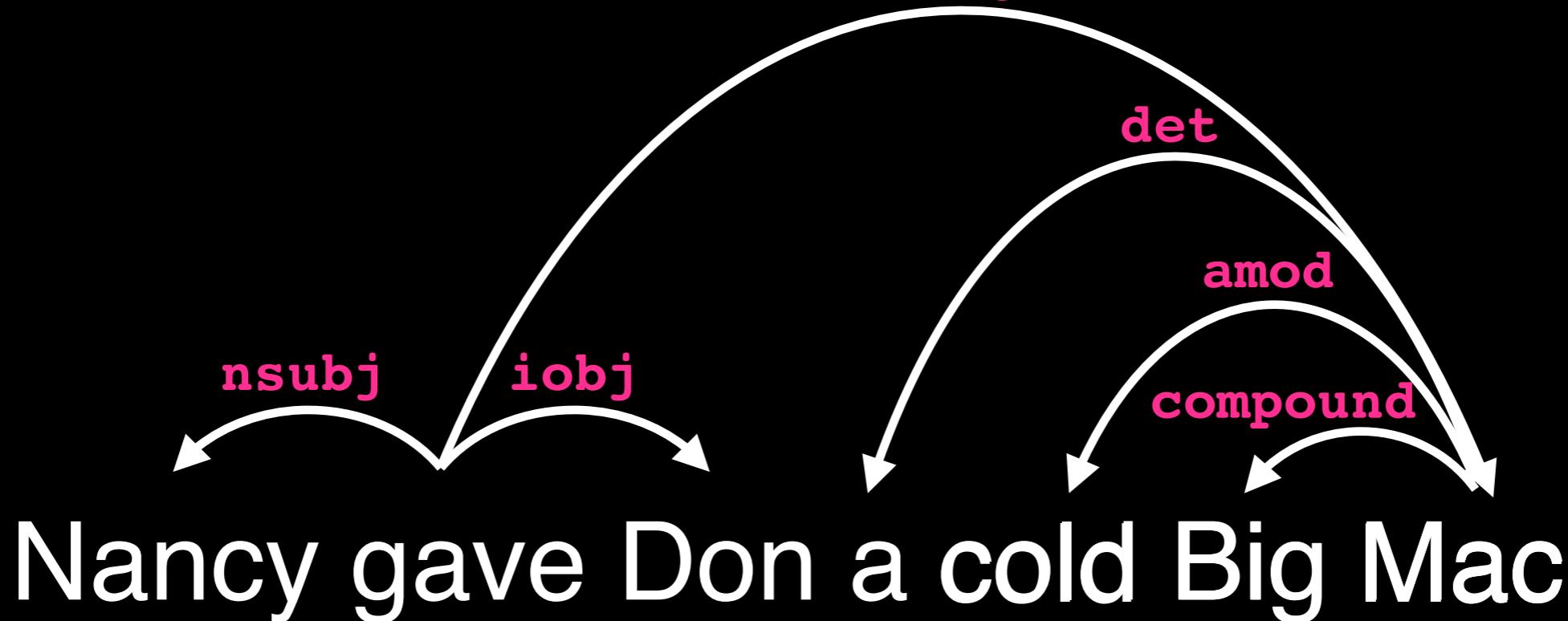
root

Dependency n -grams



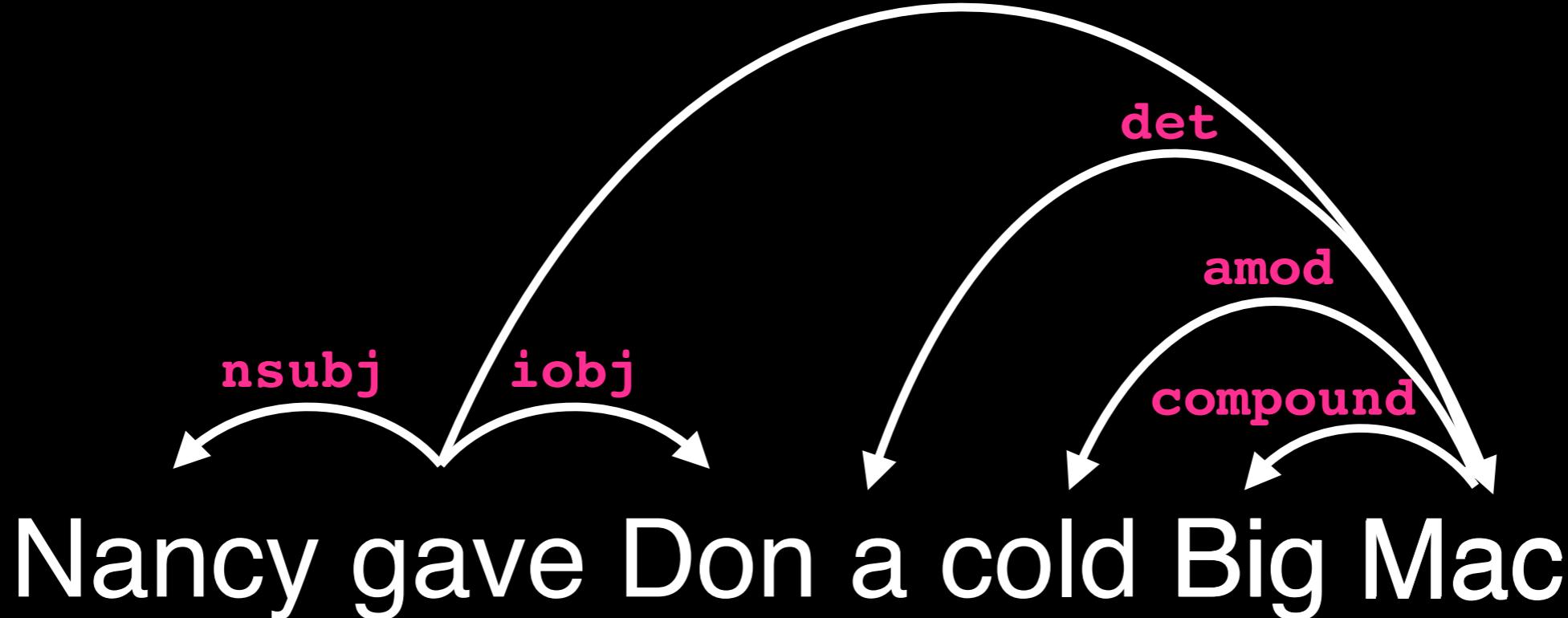
Nancy gave
Don gave
Mac gave

Dependency n -grams



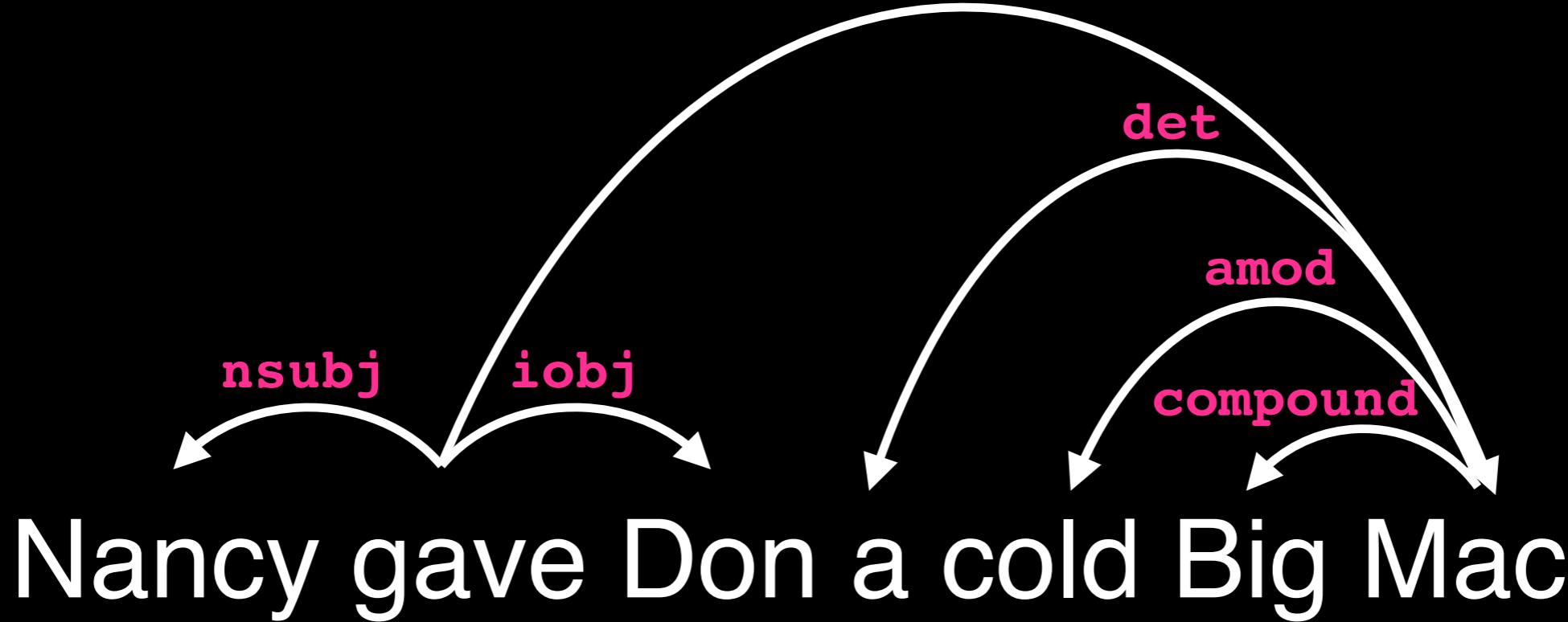
Nancy gave
Don gave
Mac gave
a Mac

Dependency n -grams

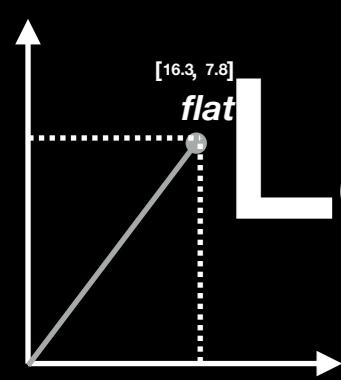


Nancy	gave
Don	gave
Mac	gave
a	Mac
cold	Mac

Dependency n -grams



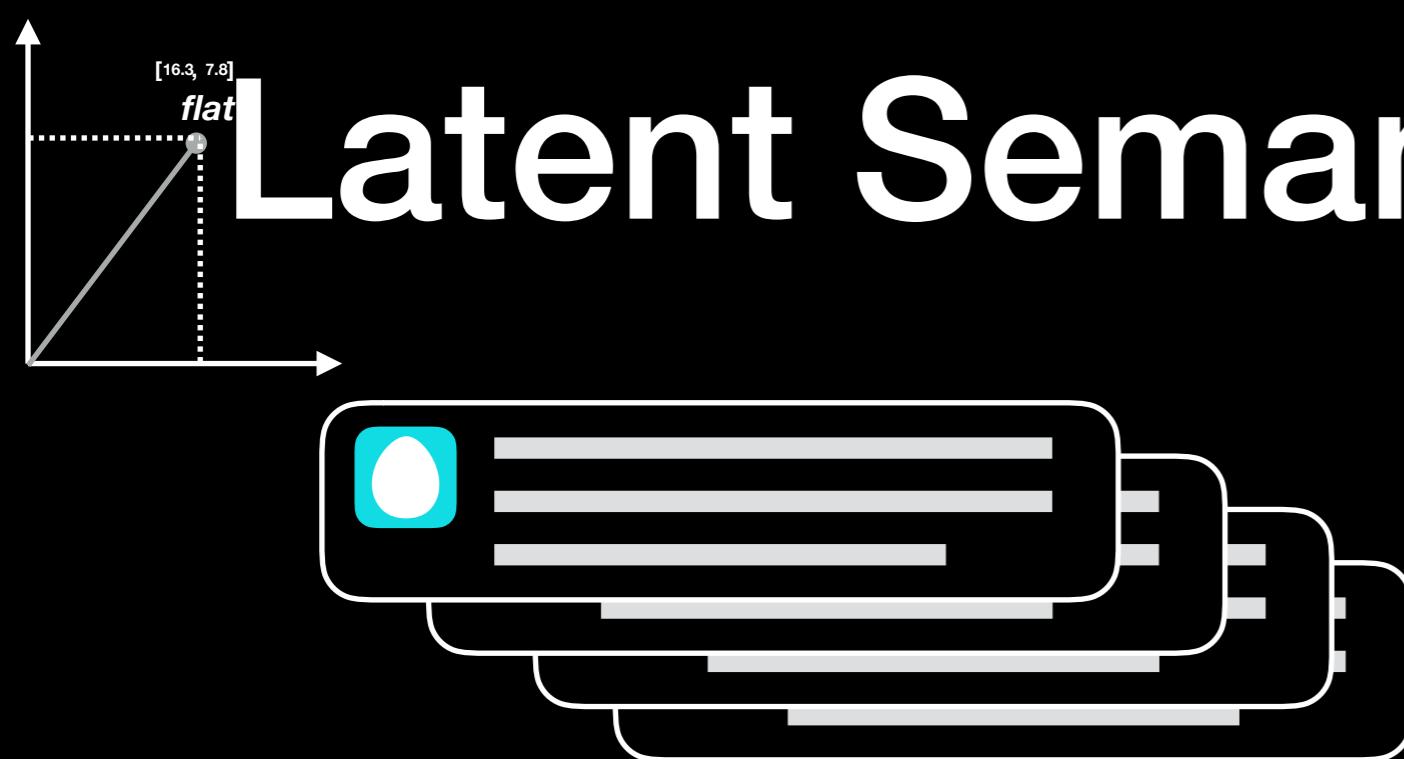
Nancy gave
Don gave
Mac gave
a Mac
cold Mac
Big Mac



Latent Semantic Analysis

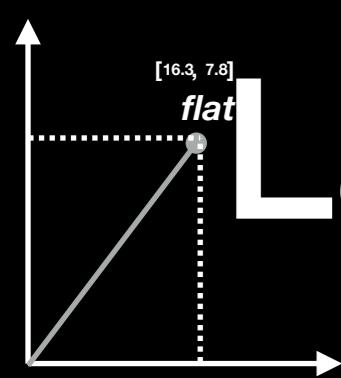
[16.3, 7.8]

flat

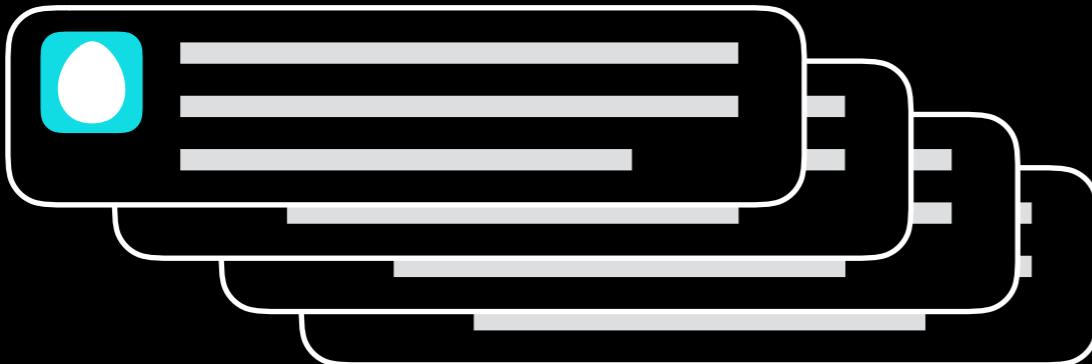


[16.3, 7.8]

flat



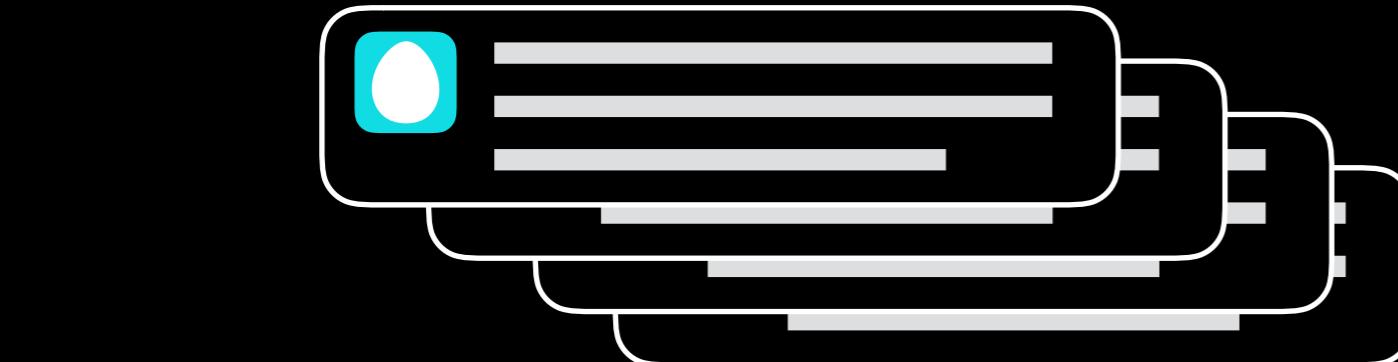
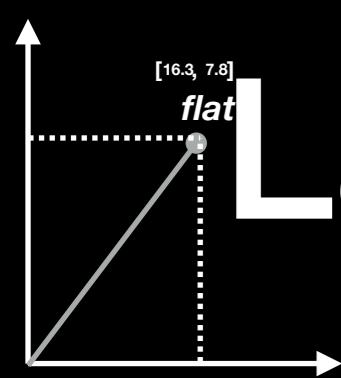
Latent Semantic Analysis



	rent	location	fairy	rainbow	prince	sleep
flat	87	73	14	11	7	45
apartment	83	87	12	23	11	32
unicorn	27	46	79	92	54	16
toad	4	37	73	55	67	73
bed	34	42	21	15	62	97

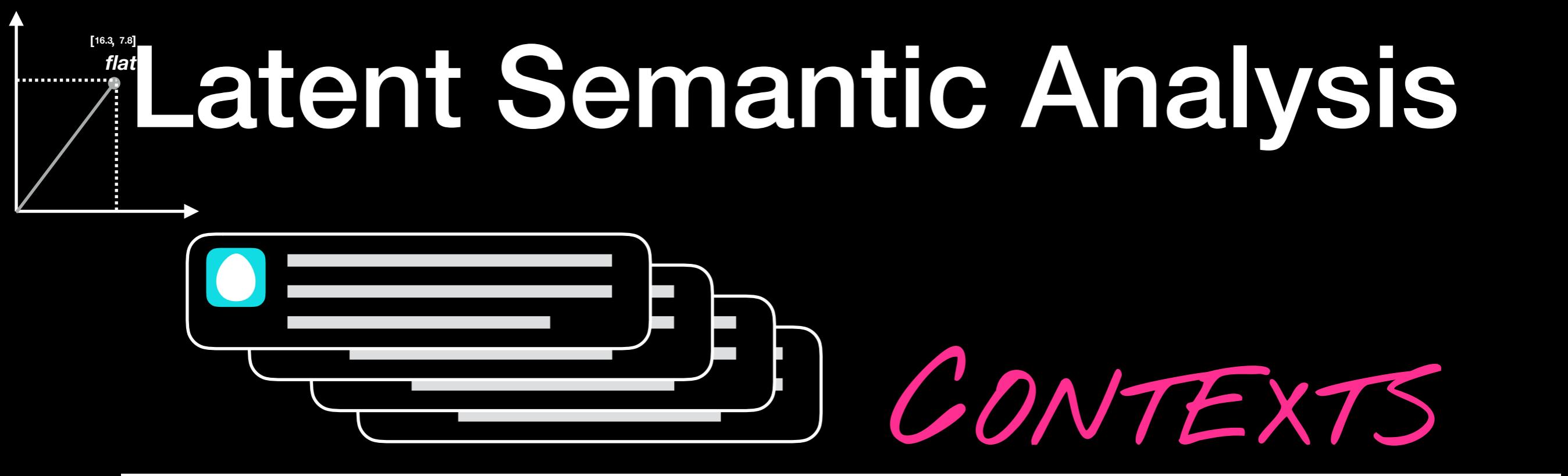
[16.3, 7.8]

flat



CONTEXTS

	rent	location	fairy	rainbow	prince	sleep
flat	87	73	14	11	7	45
apartment	83	87	12	23	11	32
unicorn	27	46	79	92	54	16
toad	4	37	73	55	67	73
bed	34	42	21	15	62	97

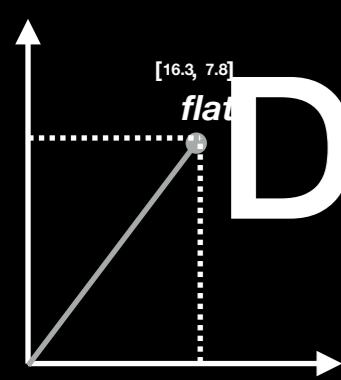


TARGETS

	rent	location	fairy	rainbow	prince	sleep
flat	87	73	14	11	7	45
apartment	83	87	12	23	11	32
unicorn	27	46	79	92	54	16
toad	4	37	73	55	67	73
bed	34	42	21	15	62	97

11

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

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

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

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

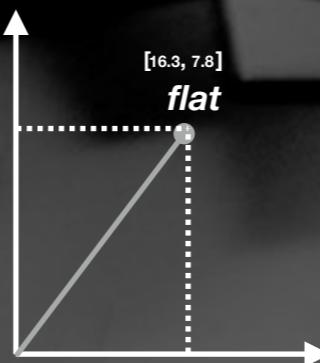
Treehouses from \$39.00/day ZZZs in the Trees

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

Part 1

Representing Words as Vectors



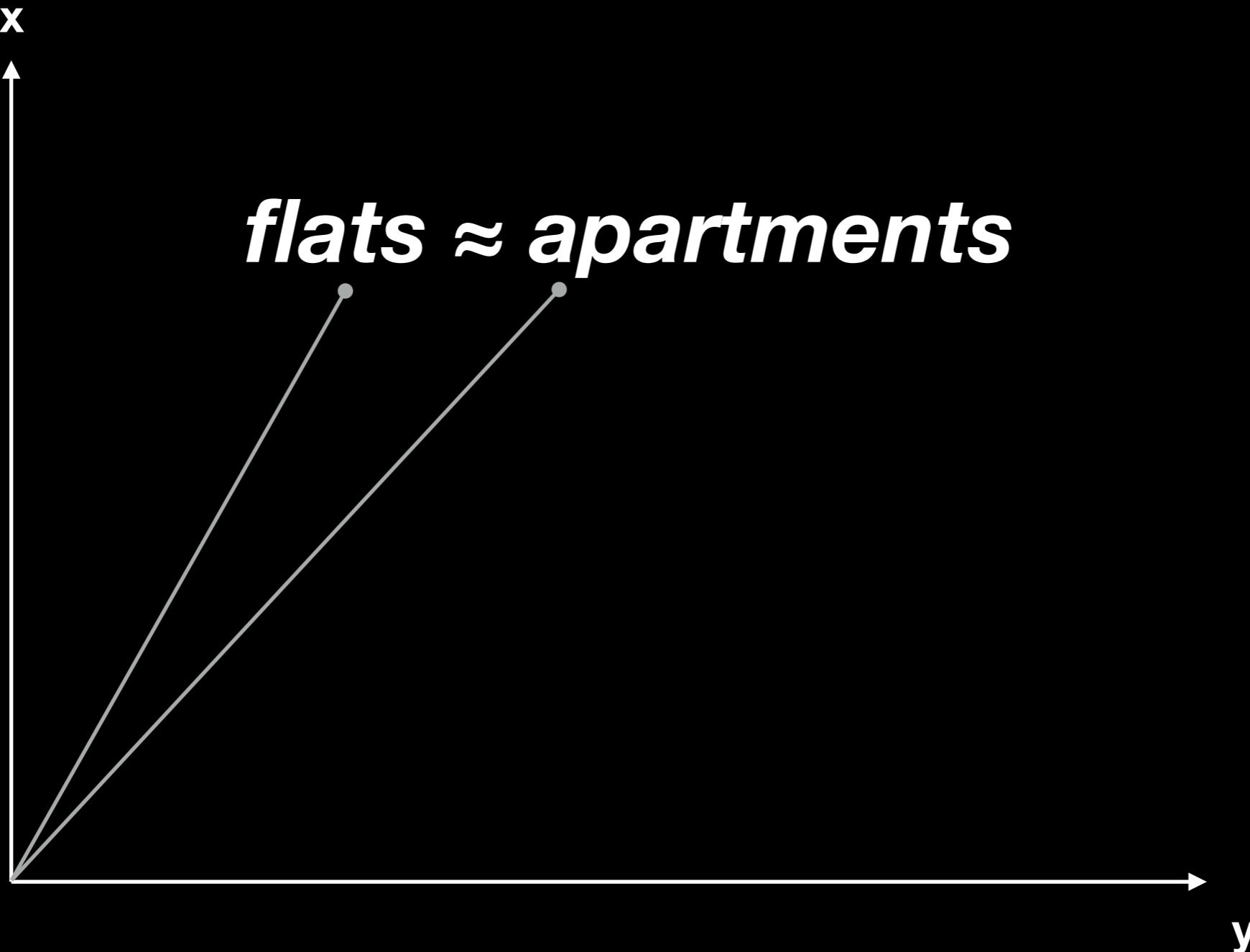
Semantic Similarity

flats ≈ apartments

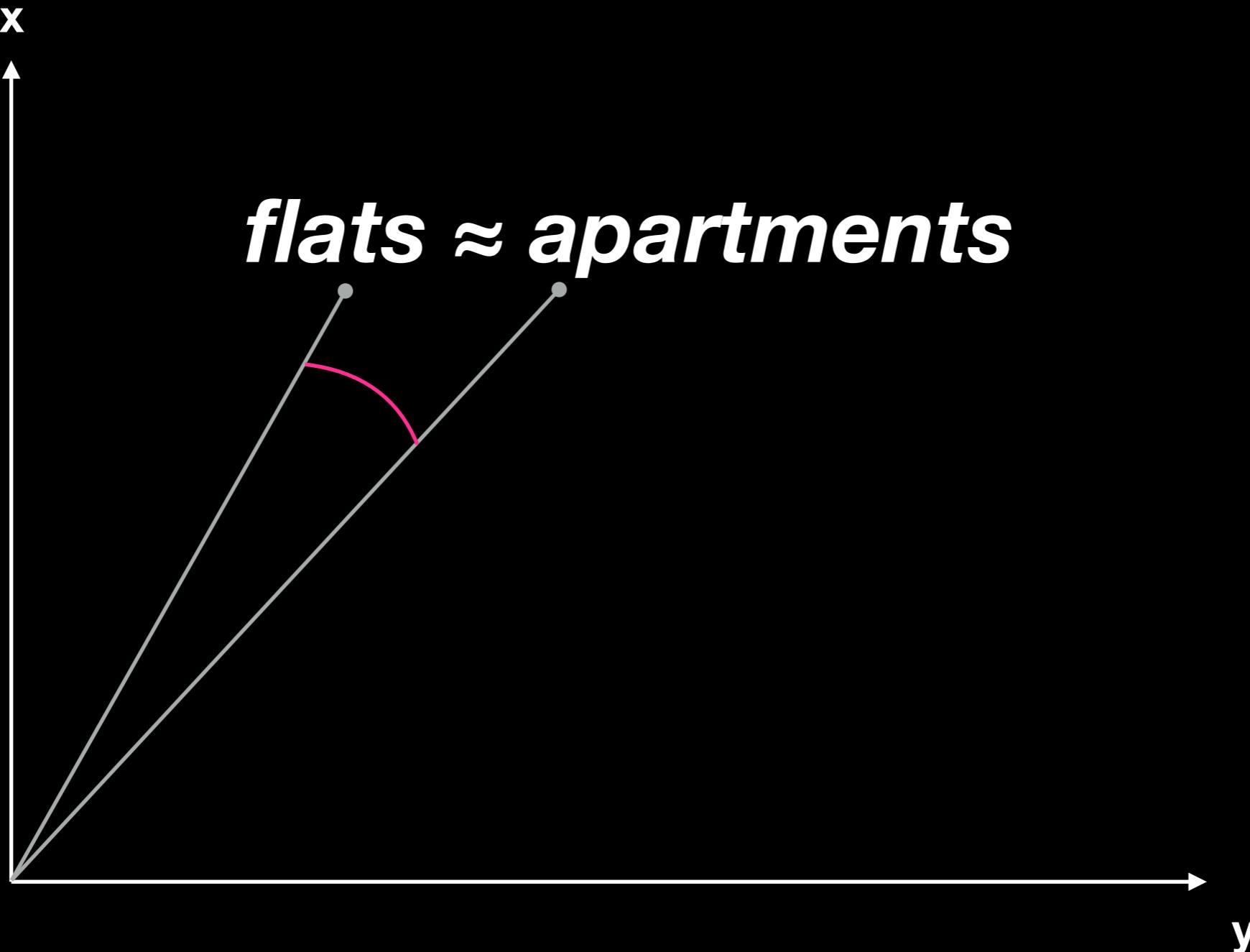
Semantic Similarity



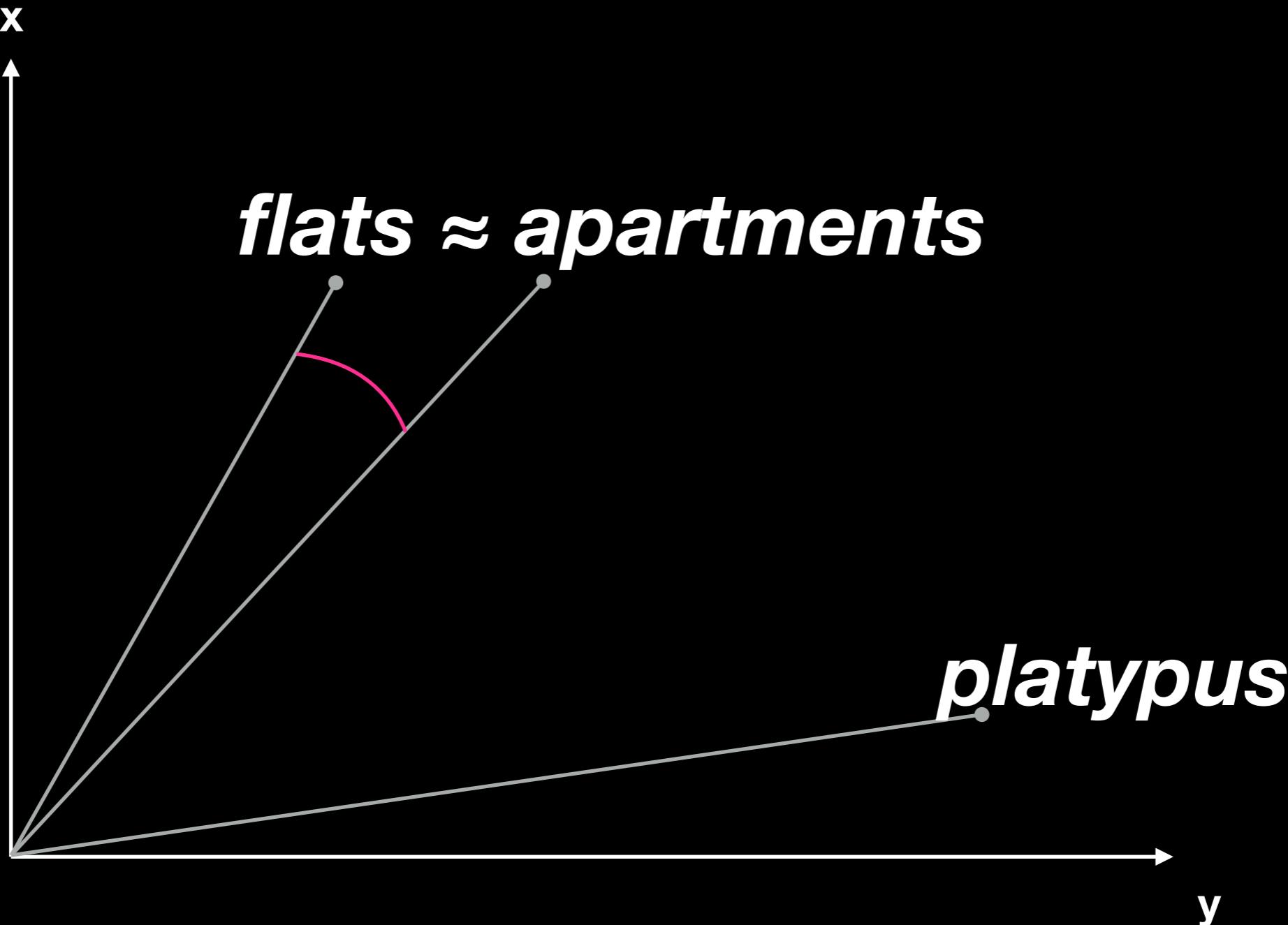
Semantic Similarity



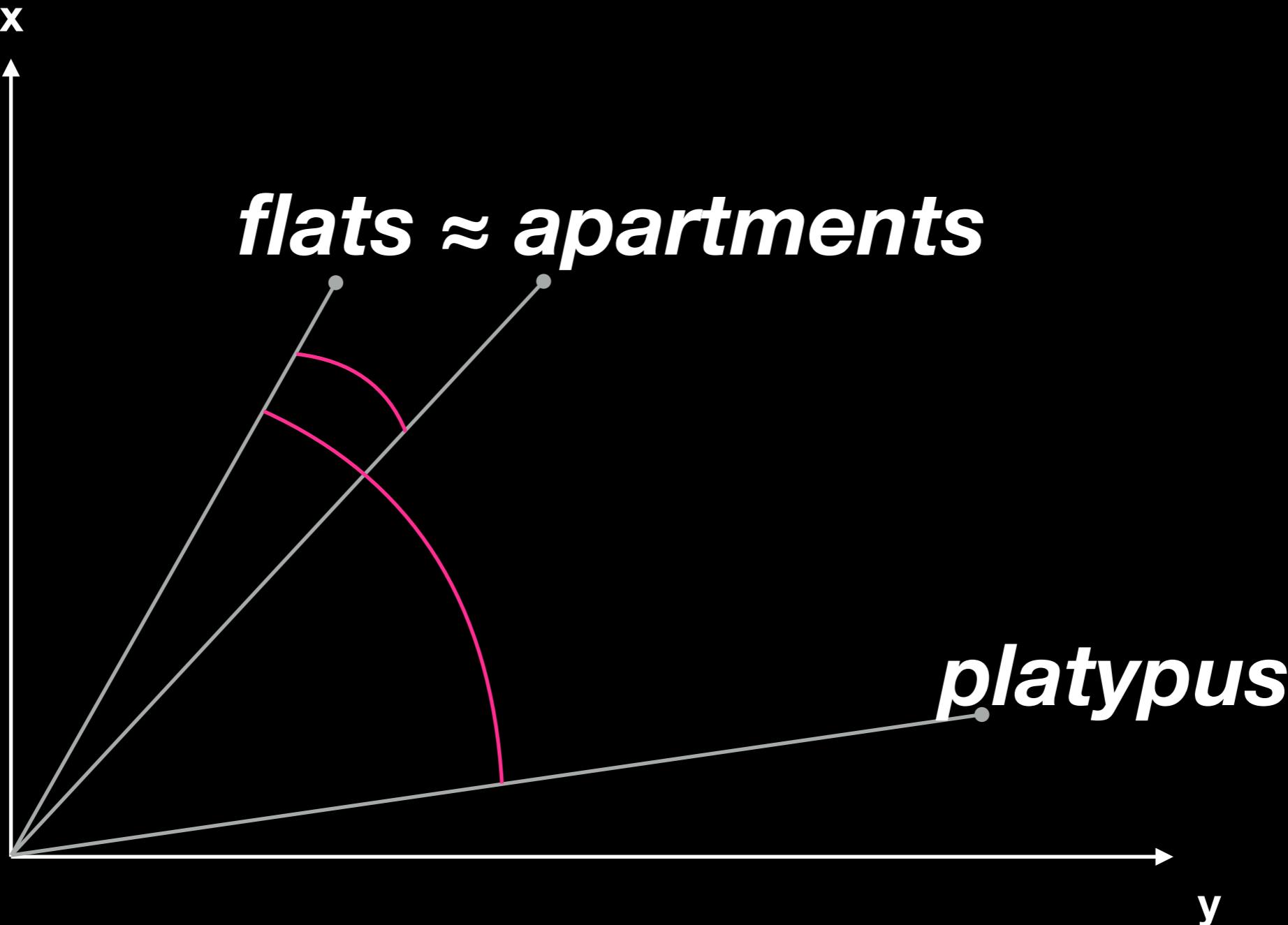
Semantic Similarity



Semantic Similarity



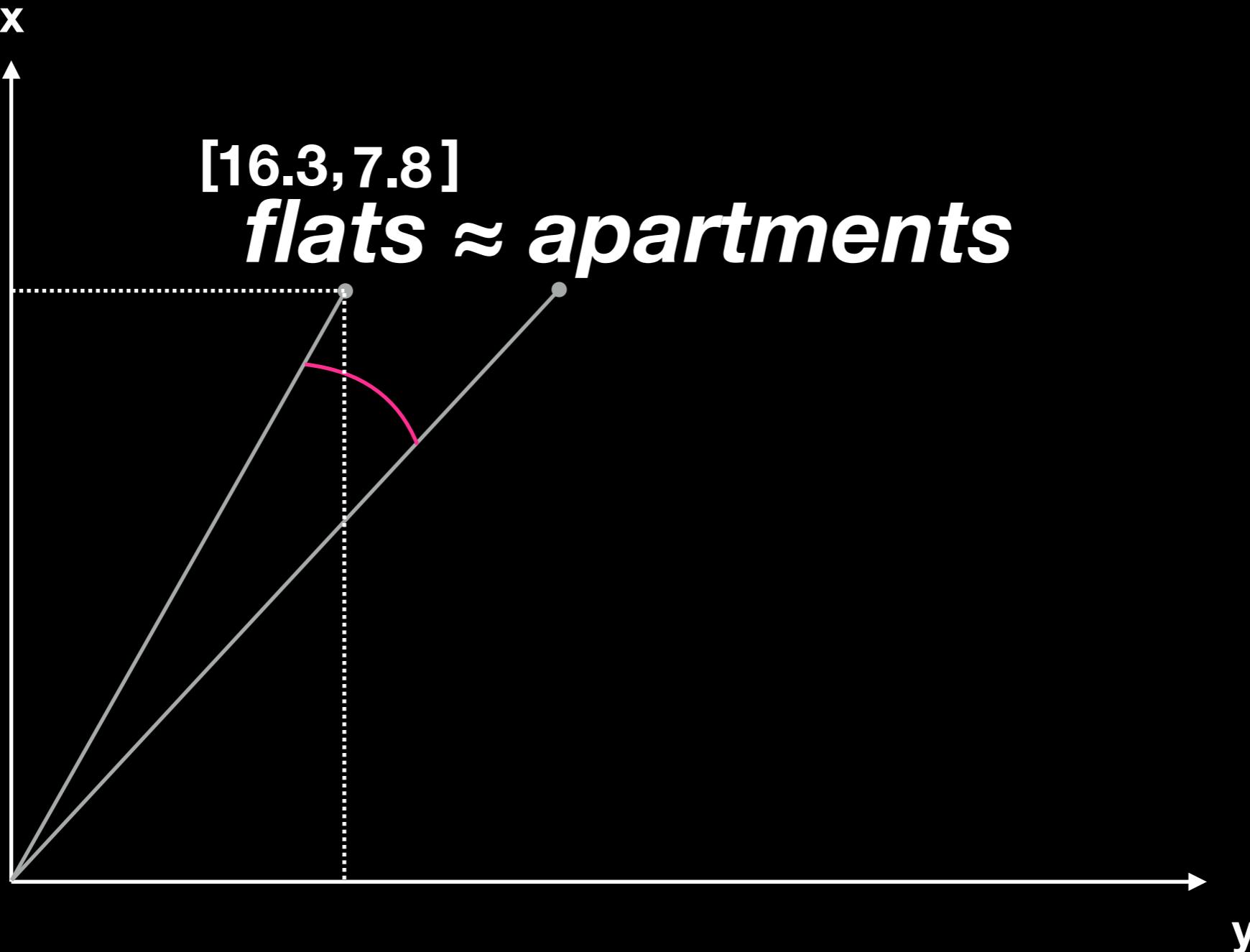
Semantic Similarity



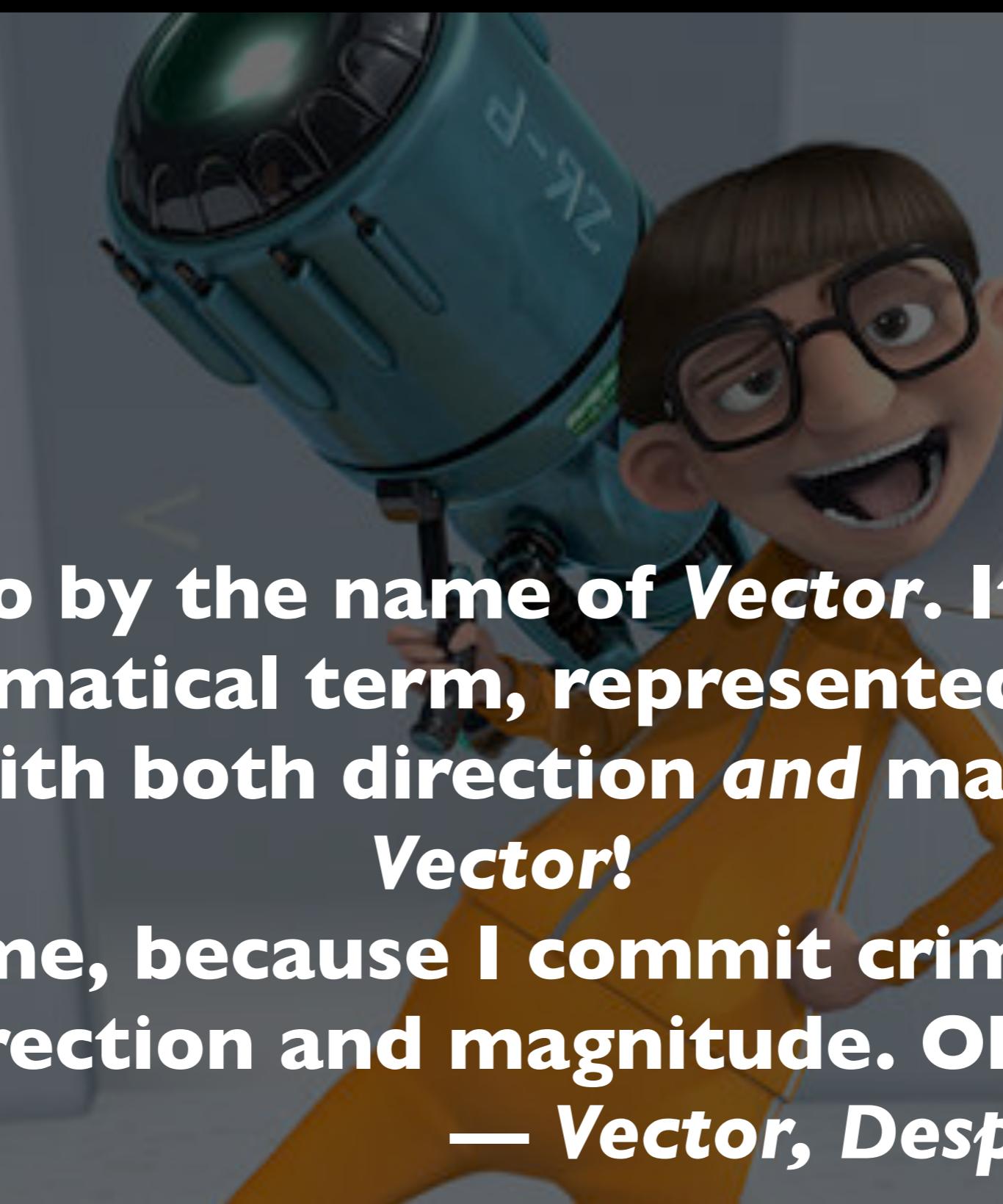
Semantic Similarity



Semantic Similarity



Vectors!

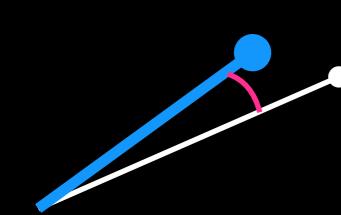


“I go by the name of Vector. It's a mathematical term, represented by an arrow with both direction and magnitude.

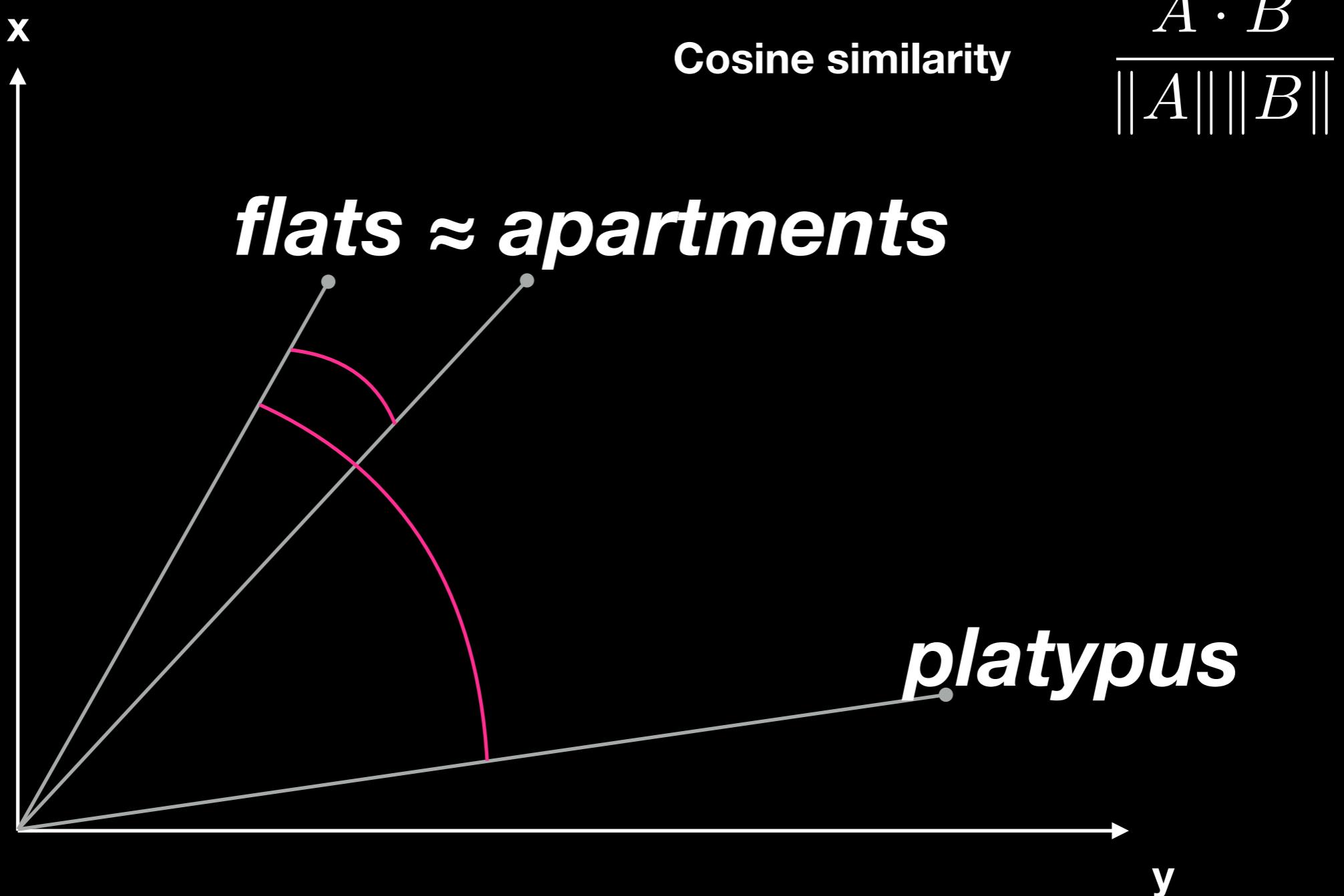
Vector!

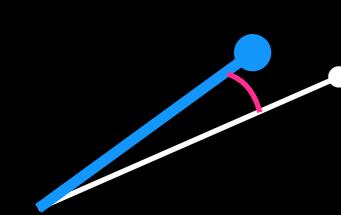
That's me, because I commit crimes with both direction and magnitude. Oh yeah!”

— Vector, Despicable Me

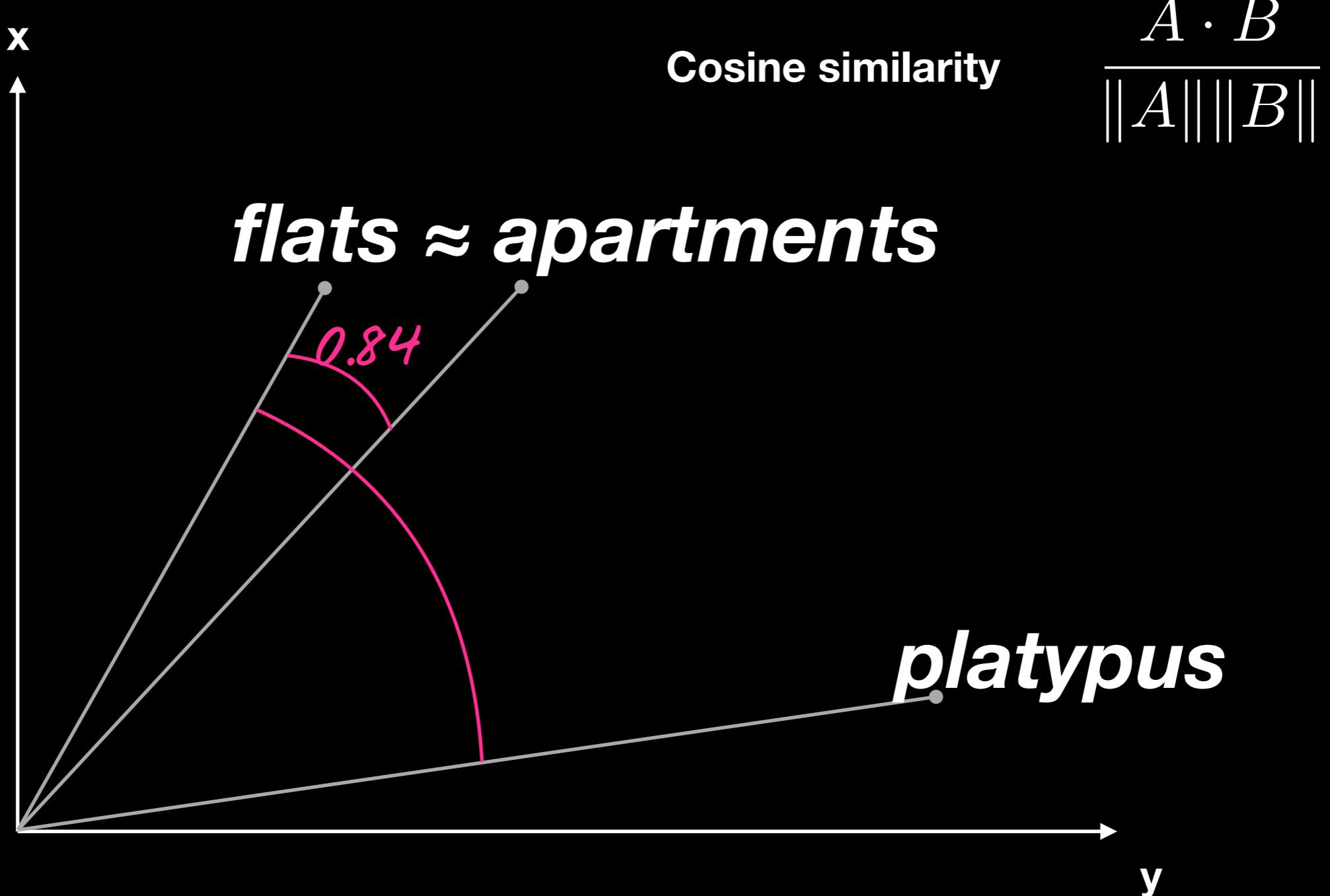


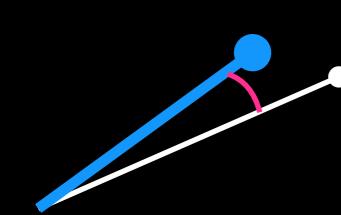
Similarity Measures



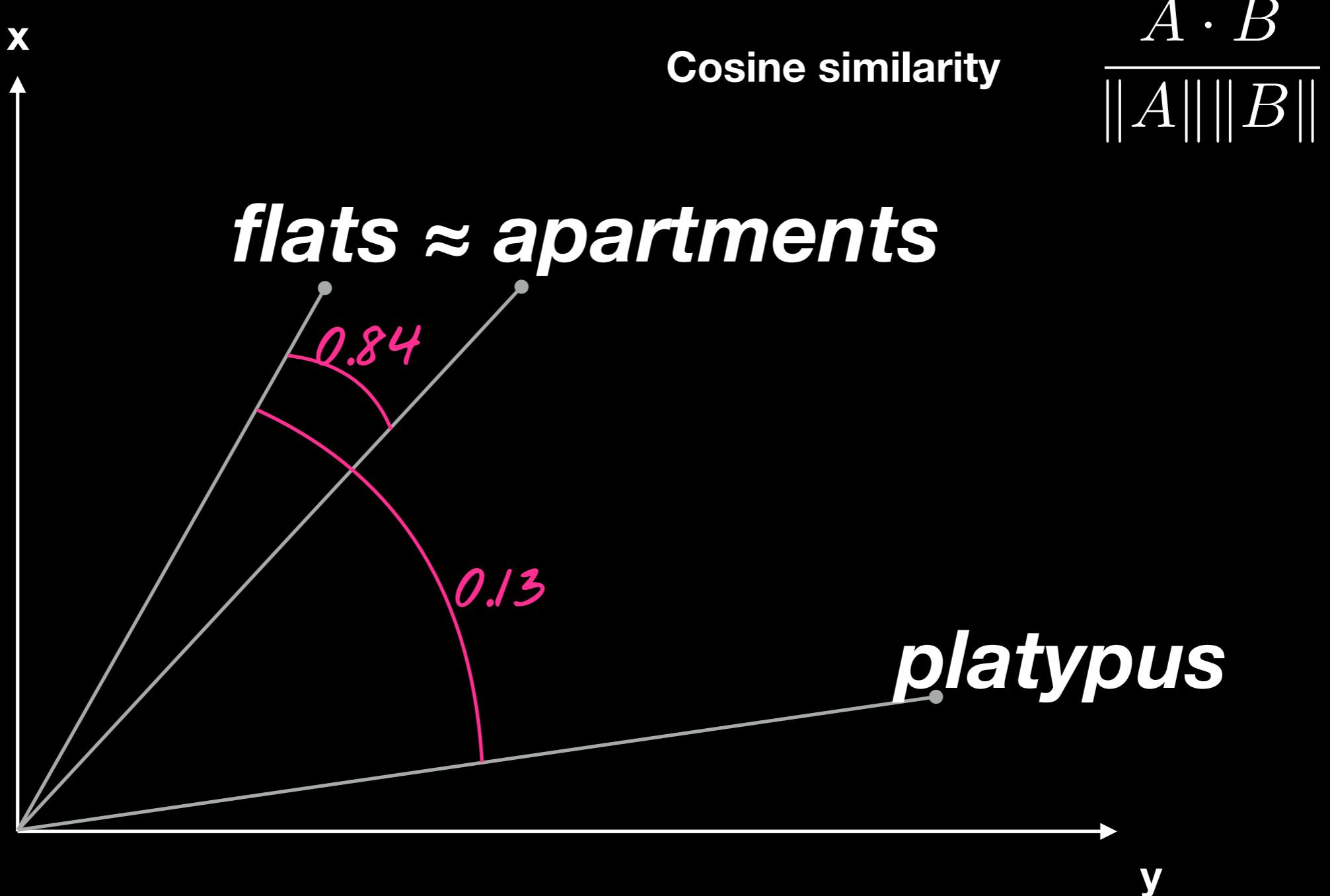


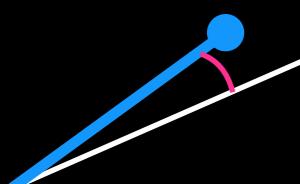
Similarity Measures



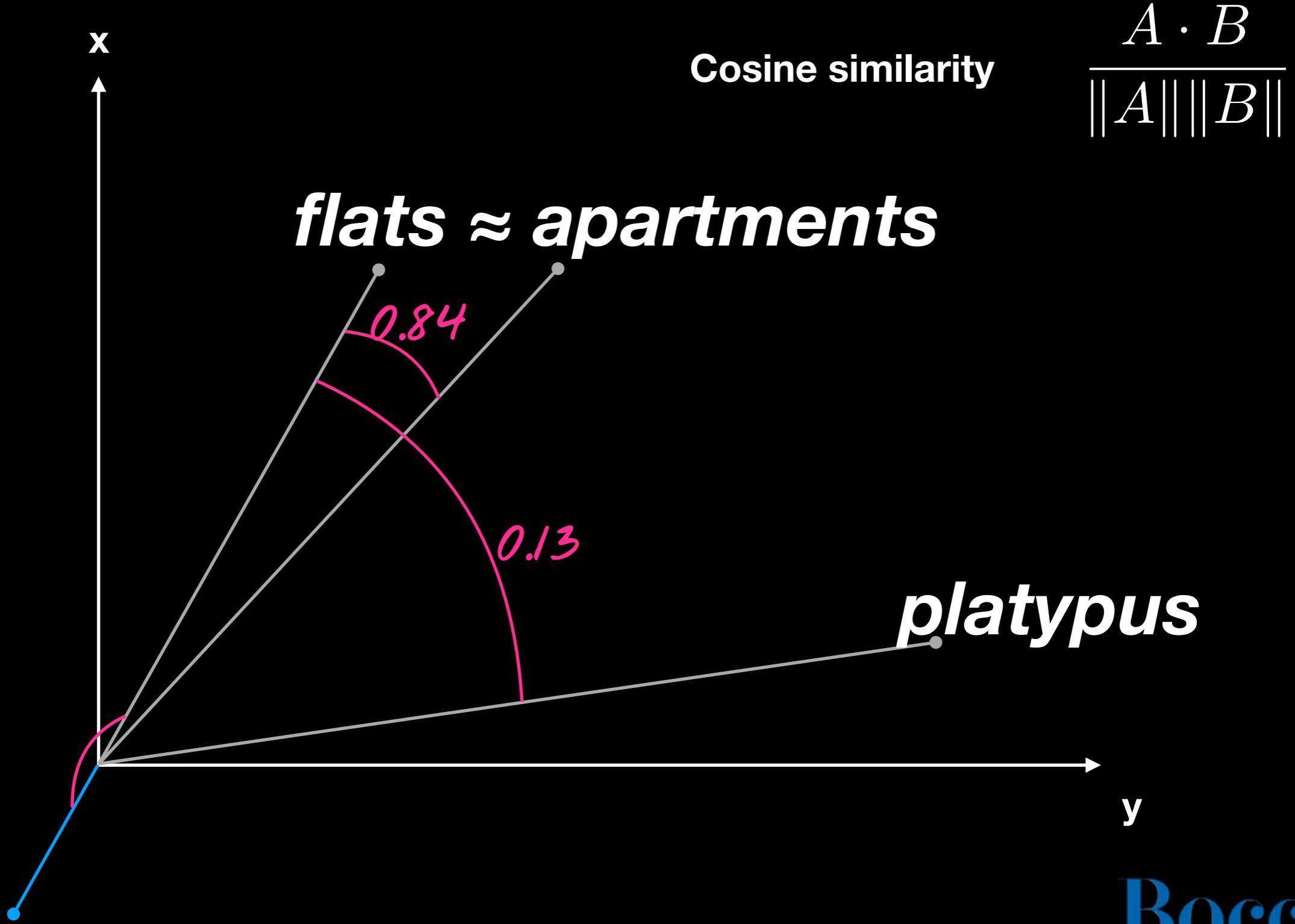


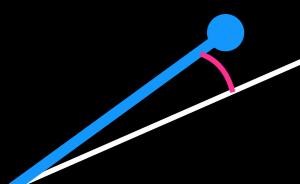
Similarity Measures



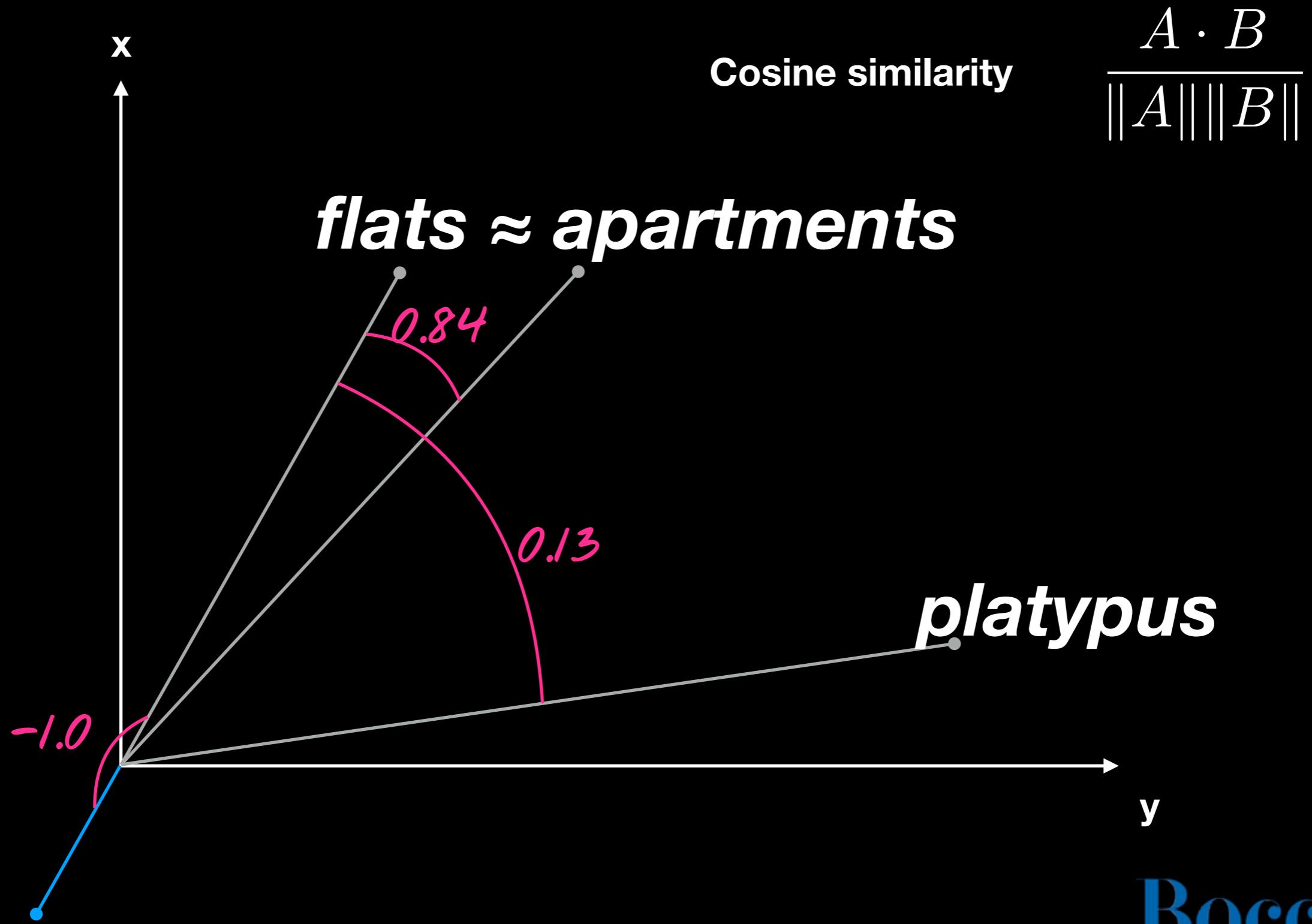


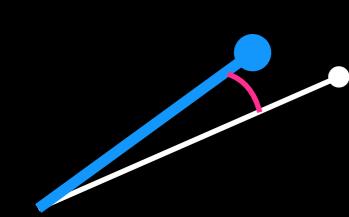
Similarity Measures





Similarity Measures



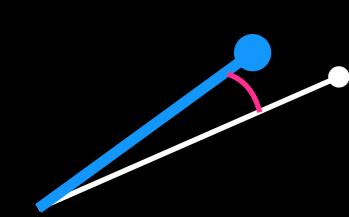


Dot Product

- “combine” vectors to a scalar

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

$$\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 6 \end{bmatrix} =$$



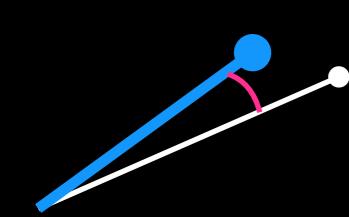
Dot Product

- “combine” vectors to a scalar

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

MULTIPLY

$$\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 6 \end{bmatrix} =$$



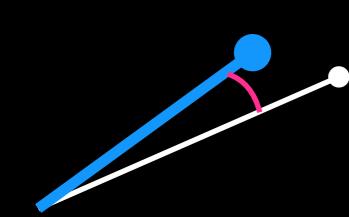
Dot Product

- “combine” vectors to a scalar

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

MULTIPLY

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



Dot Product

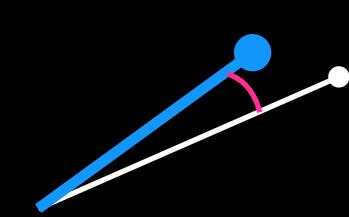
- “combine” vectors to a scalar

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

MULTIPLY

A handwritten-style pink arrow points from the word "MULTIPLY" to the term $x_i y_i$ in the summation formula.

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



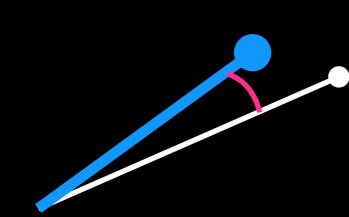
Dot Product

- “combine” vectors to a scalar

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

SUM *MULTIPLY*

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



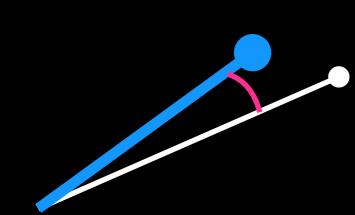
Dot Product

- “combine” vectors to a scalar

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

SUM *MULTIPLY*

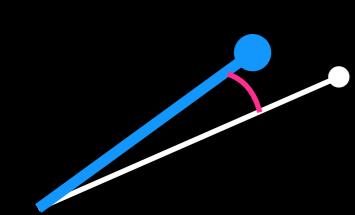
$$\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 6 \end{bmatrix} = 4$$



Vector Norm

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

$$\begin{bmatrix} 2 \\ 6 \end{bmatrix}$$

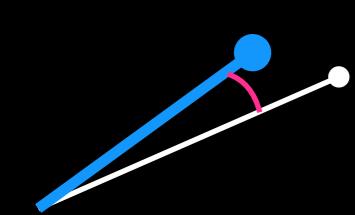


Vector Norm

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

$$\begin{bmatrix} 2 \\ 6 \end{bmatrix}$$

$$= \sqrt{2^2 + 6^2}$$



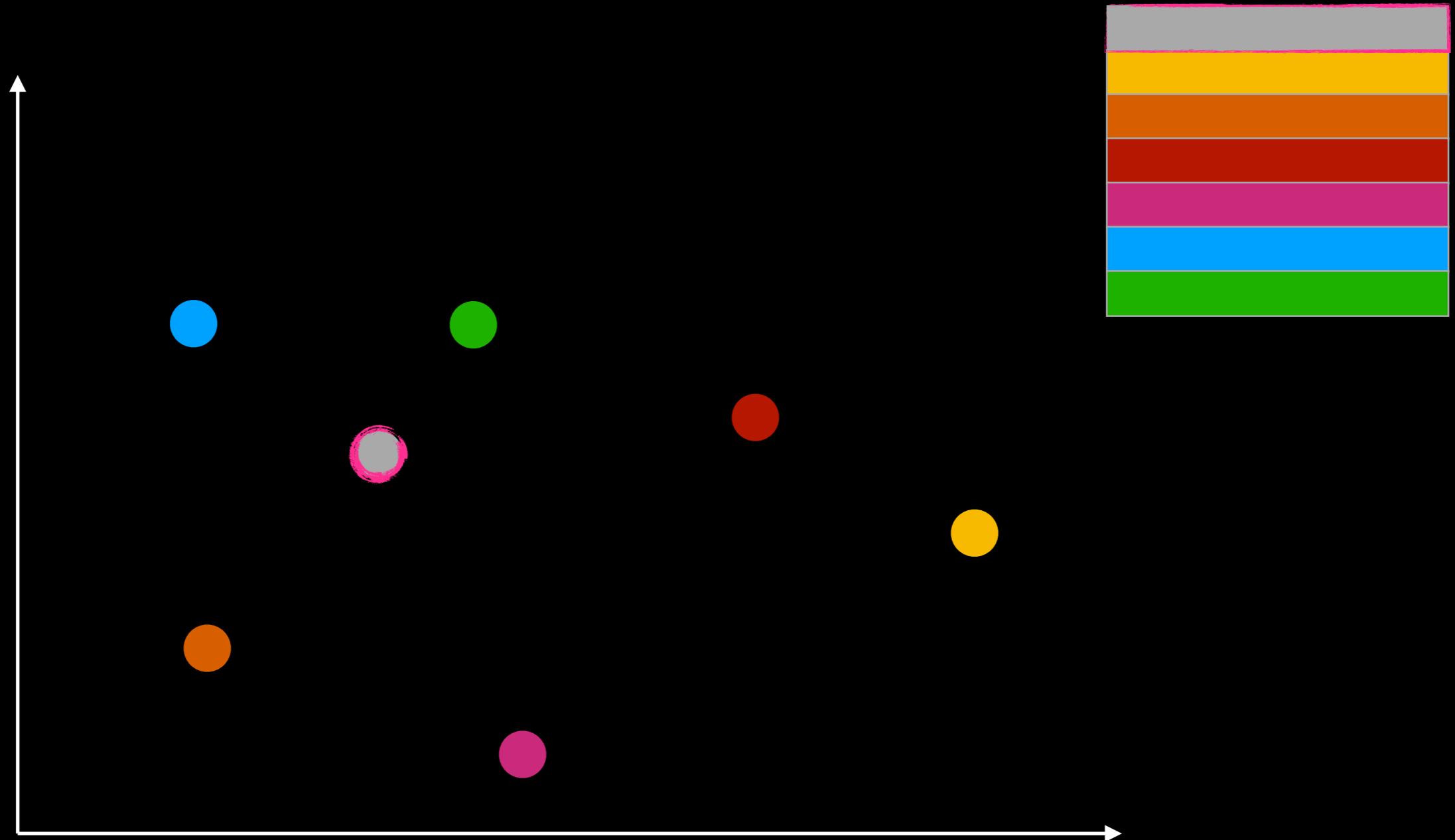
Vector Norm

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

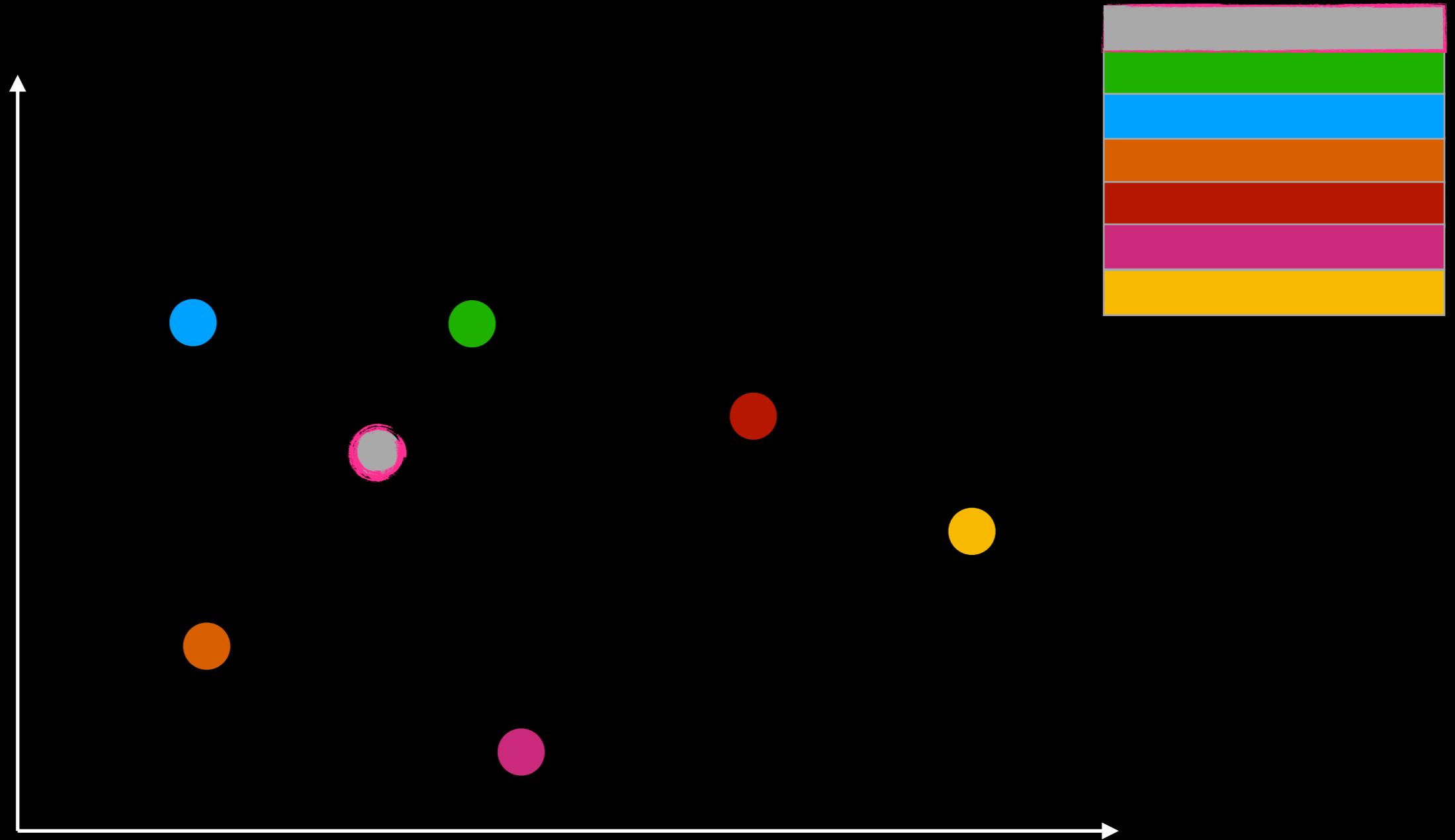
$$\begin{bmatrix} 2 \\ 6 \end{bmatrix}$$

$$= \sqrt{2^2 + 6^2} = 6.324$$

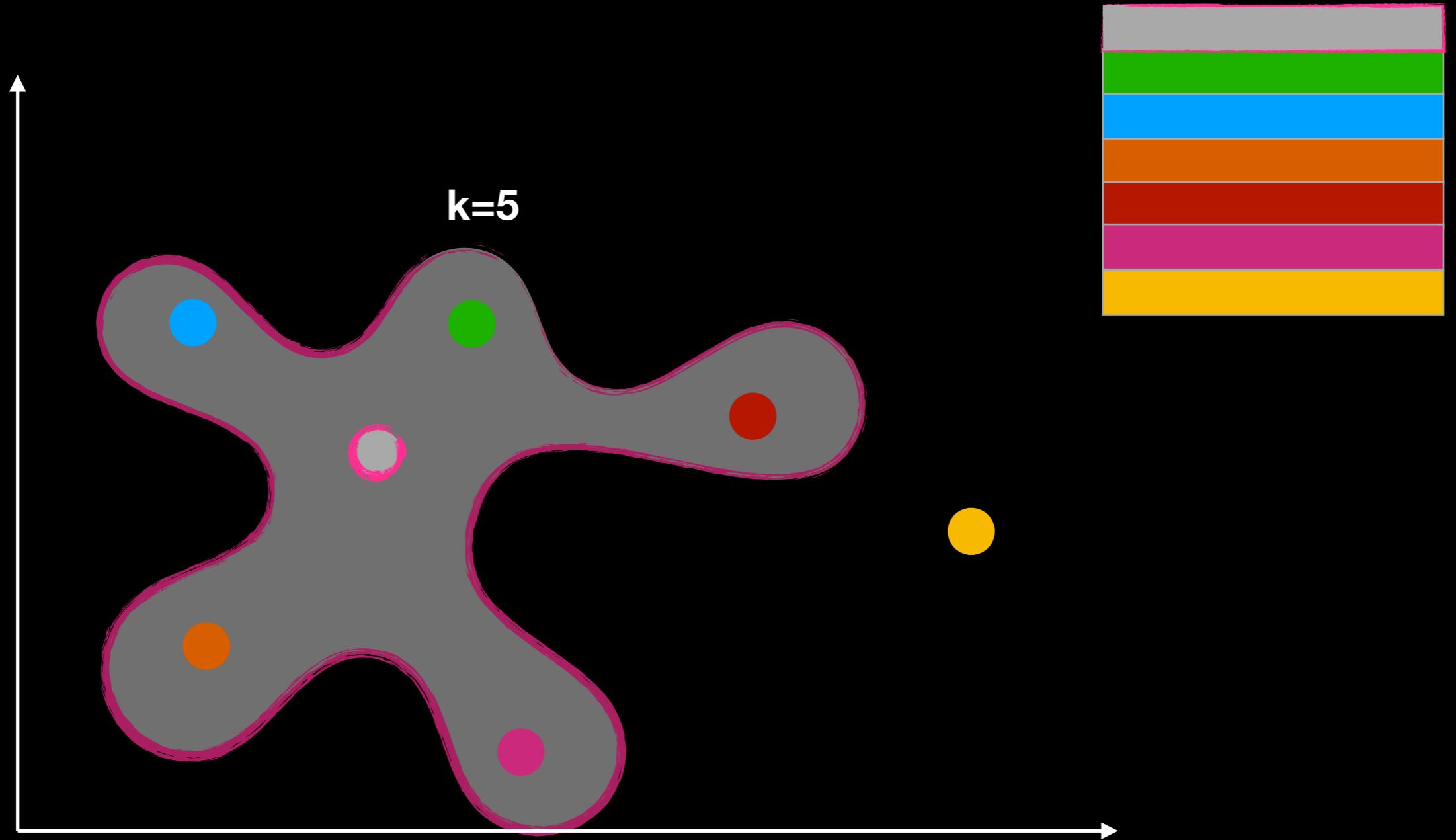
Nearest neighbors



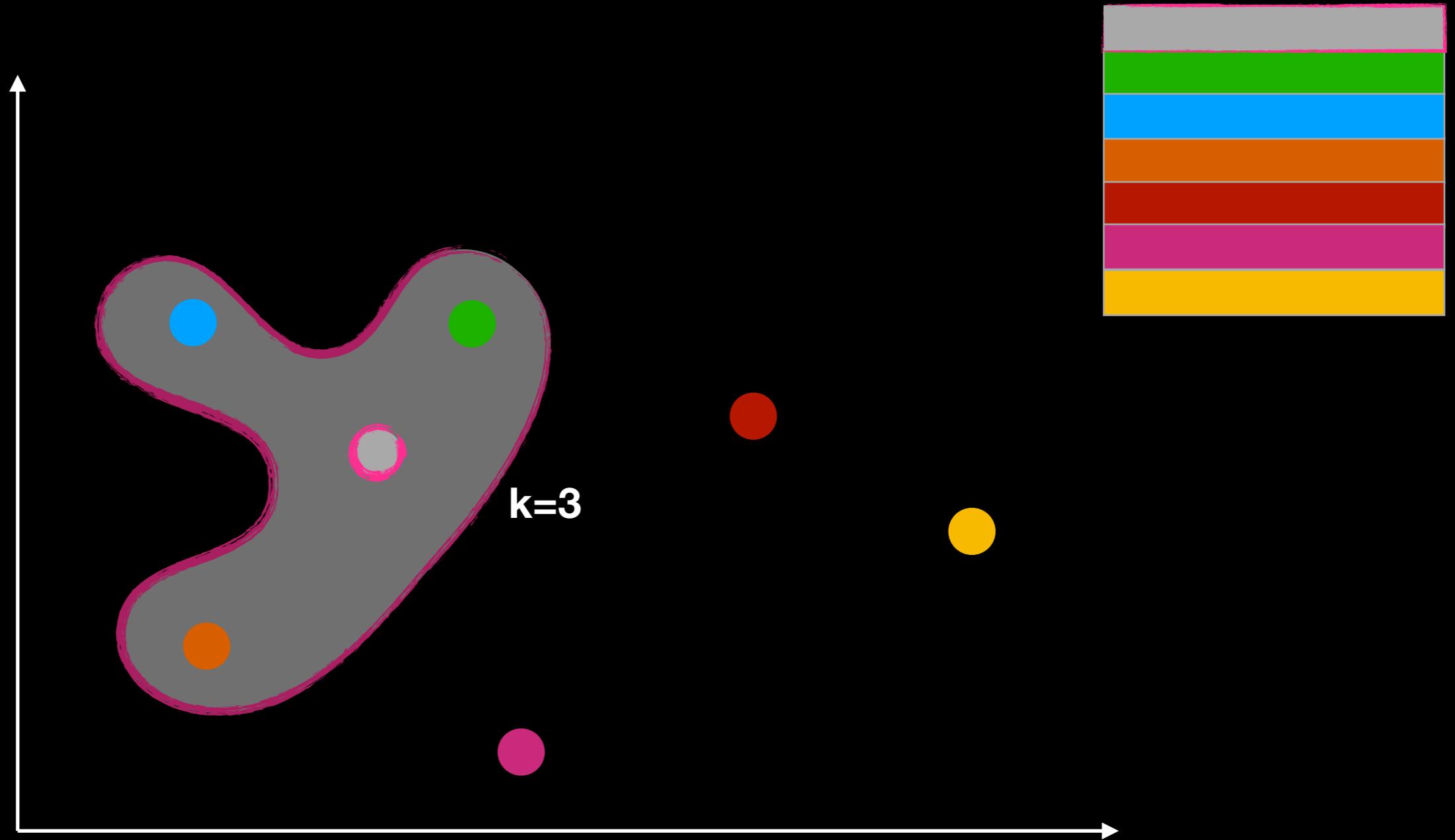
Nearest neighbors



Nearest neighbors



Nearest neighbors



Word2Vec – Intuitively

place all words randomly on fridge

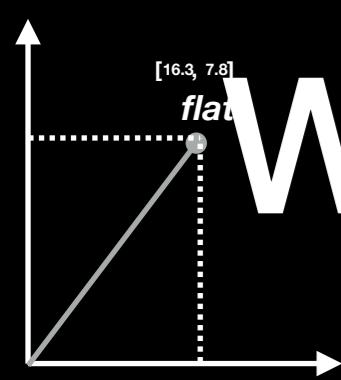
for each pair of words:

if in same sentence:

move closer together

else:

move further apart



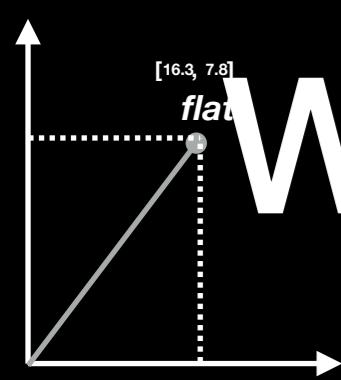
Word2Vec – CBOW Model

MATRIX OF
TARGET WORDS

MATRIX OF
CONTEXT WORDS

Renting out large *apartment* in great location

Bocconi



Word2Vec – CBOW Model

MATRIX OF

OUTPUT

apartment

TARGET WORDS

INPUT

rent

large

great

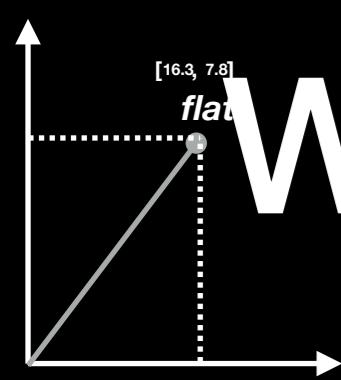
location

MATRIX OF

Renting out large *apartment* in great location

CONTEXT WORDS

Bocconi



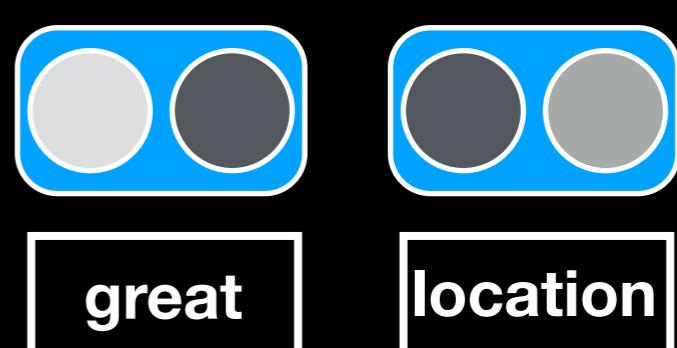
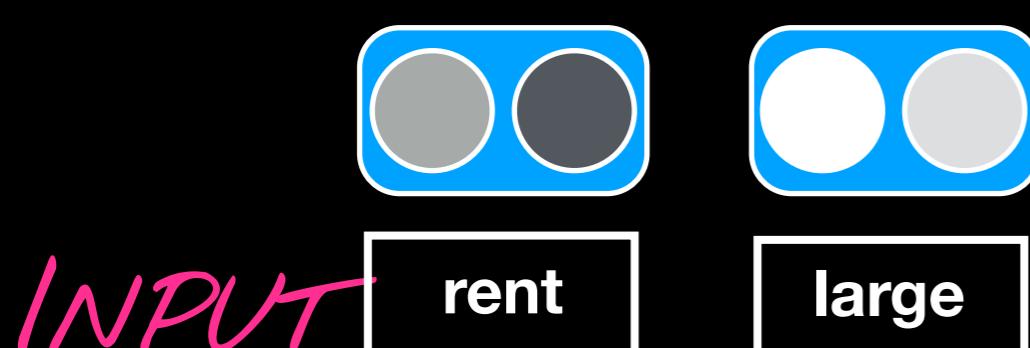
Word2Vec – CBOW Model

MATRIX OF

OUTPUT

apartment

TARGET WORDS



Renting out large *apartment* in great location

MATRIX OF
CONTEXT WORDS

Bocconi

Word2Vec – CBOW Model

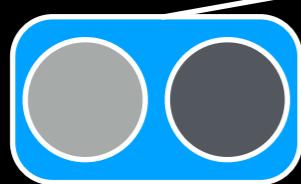
MATRIX OF

TARGET WORDS

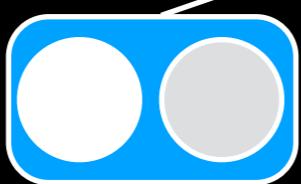
OUTPUT

apartment

INPUT

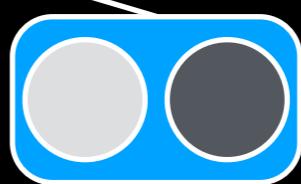


rent

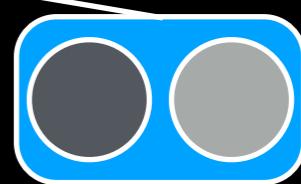


large

SUM



great

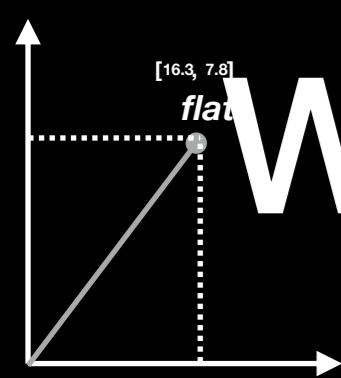


location

MATRIX OF
CONTEXT WORDS

Renting out large *apartment* in great location

Bocconi



Word2Vec – CBOW Model

MATRIX OF

TARGET WORDS

OUTPUT

garden

SUM

INPUT

rent

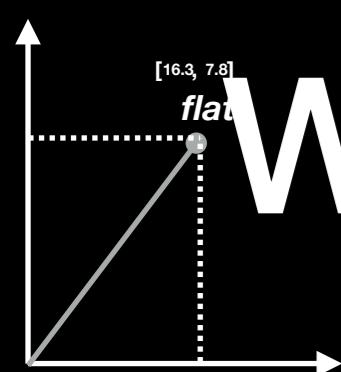
large

MATRIX OF

CONTEXT WORDS

Renting out large *apartment* in great location

Bocconi



Word2Vec – CBOW Model

MATRIX OF

TARGET WORDS

OUTPUT

apartment

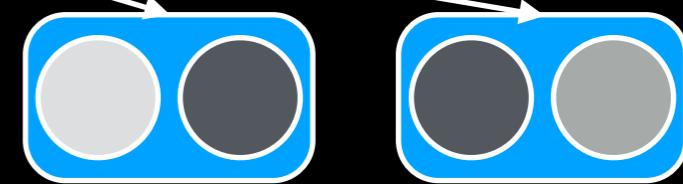
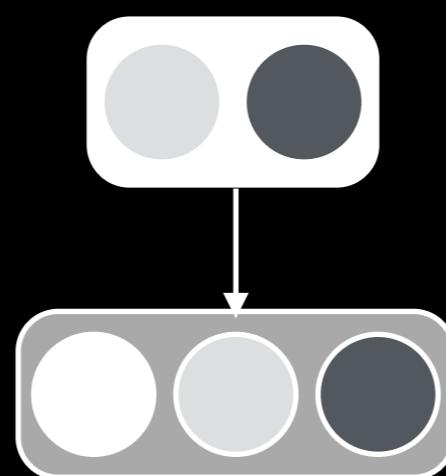
ERROR

BACKPROPAGATION

INPUT

rent

large



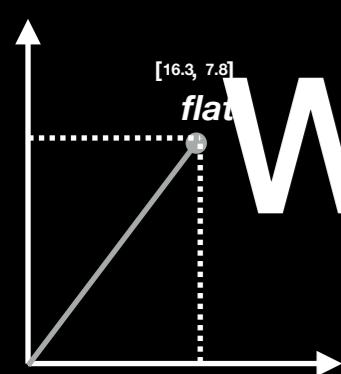
great
location

MATRIX OF

CONTEXT WORDS

Renting out large *apartment* in great location

Bocconi



Word2Vec – CBOW Model

MATRIX OF

TARGET WORDS

OUTPUT

apartment

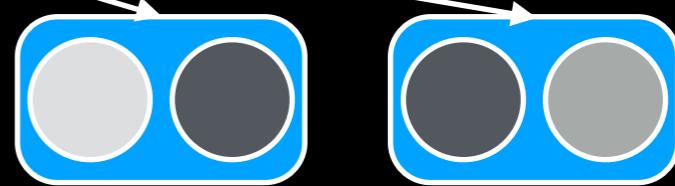
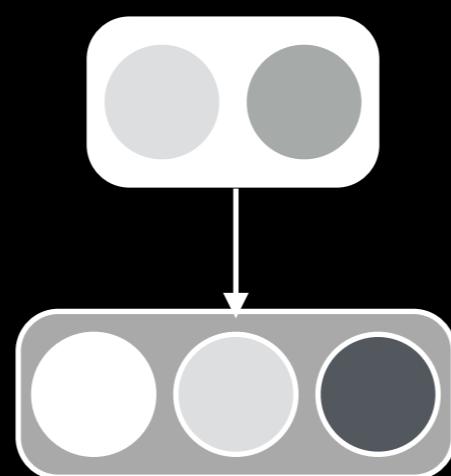
ERROR

BACKPROPAGATION

INPUT

rent

large



great
location

MATRIX OF

CONTEXT WORDS

Renting out large *apartment* in great location

Bocconi

Word2Vec – CBOW Model

MATRIX OF

TARGET WORDS

OUTPUT

apartment

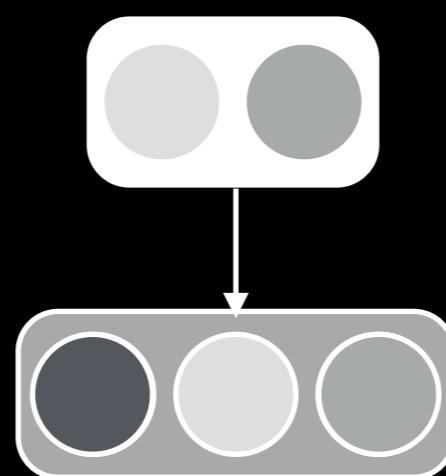
ERROR

BACKPROPAGATION

INPUT

rent

large



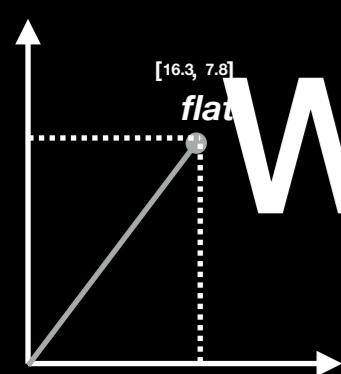
great
location

MATRIX OF

CONTEXT WORDS

Renting out large *apartment* in great location

Bocconi



Word2Vec – CBOW Model

MATRIX OF

TARGET WORDS

OUTPUT

apartment

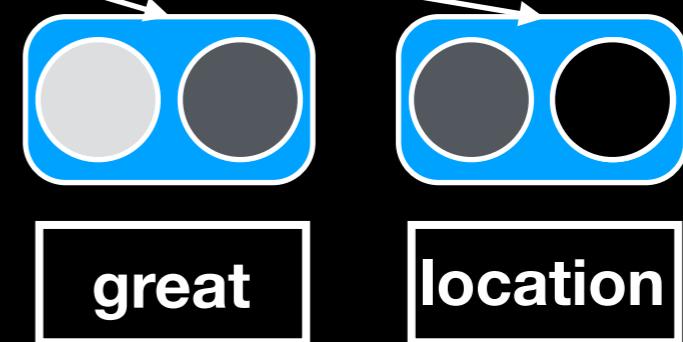
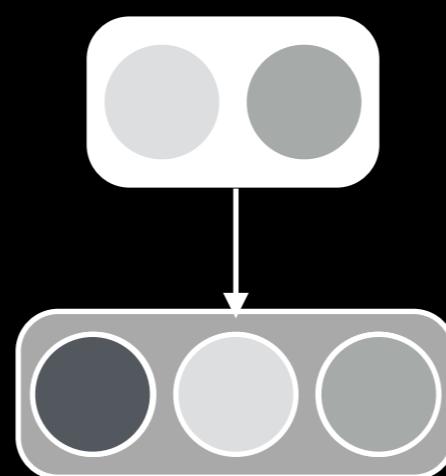
ERROR

BACKPROPAGATION

INPUT

rent

large



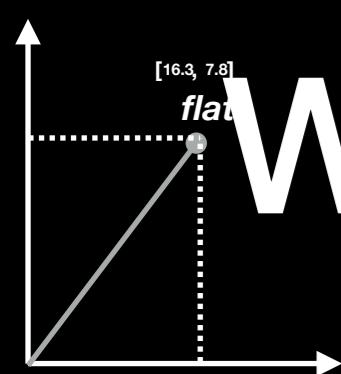
great
location

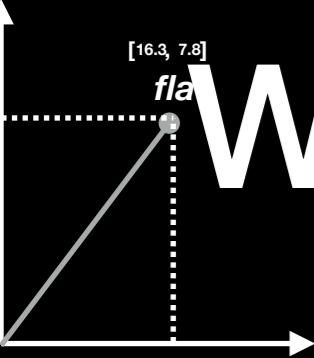
MATRIX OF

CONTEXT WORDS

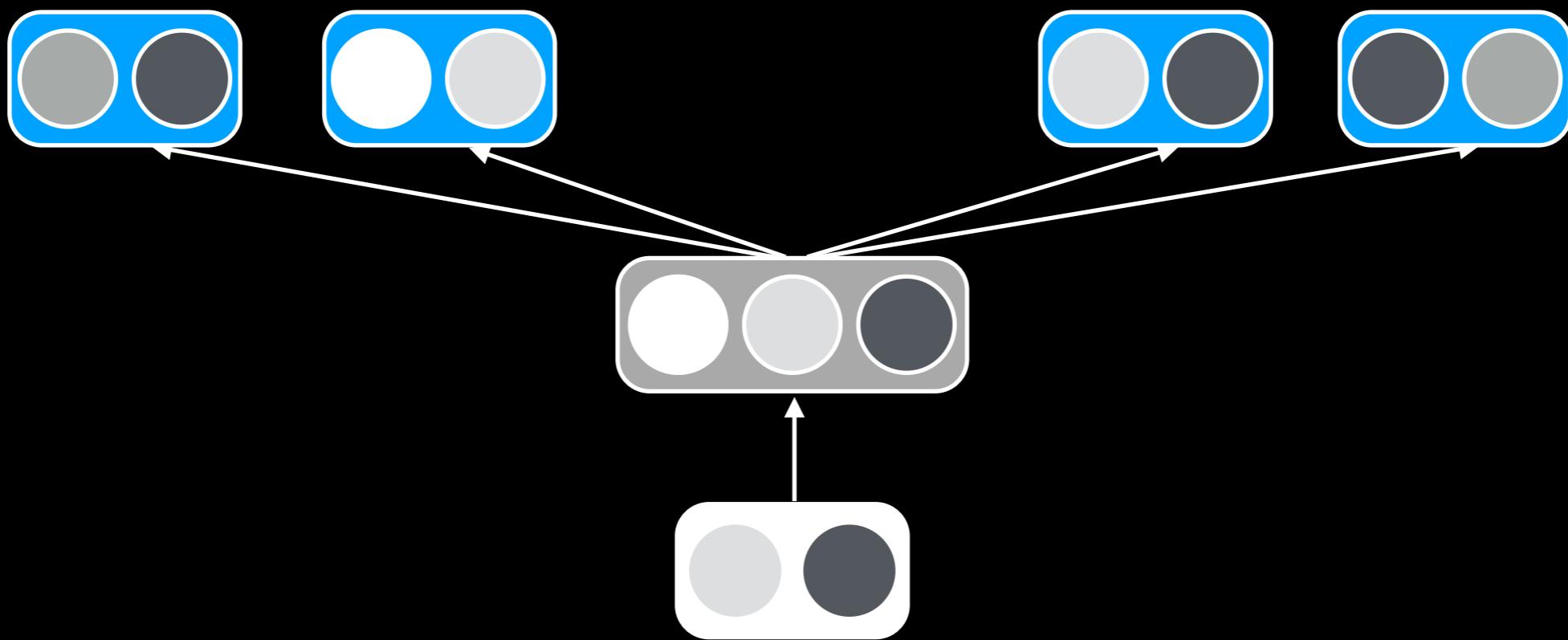
Renting out large *apartment* in great location

Bocconi

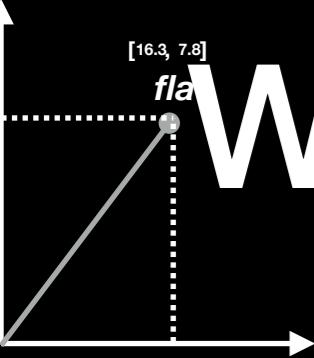




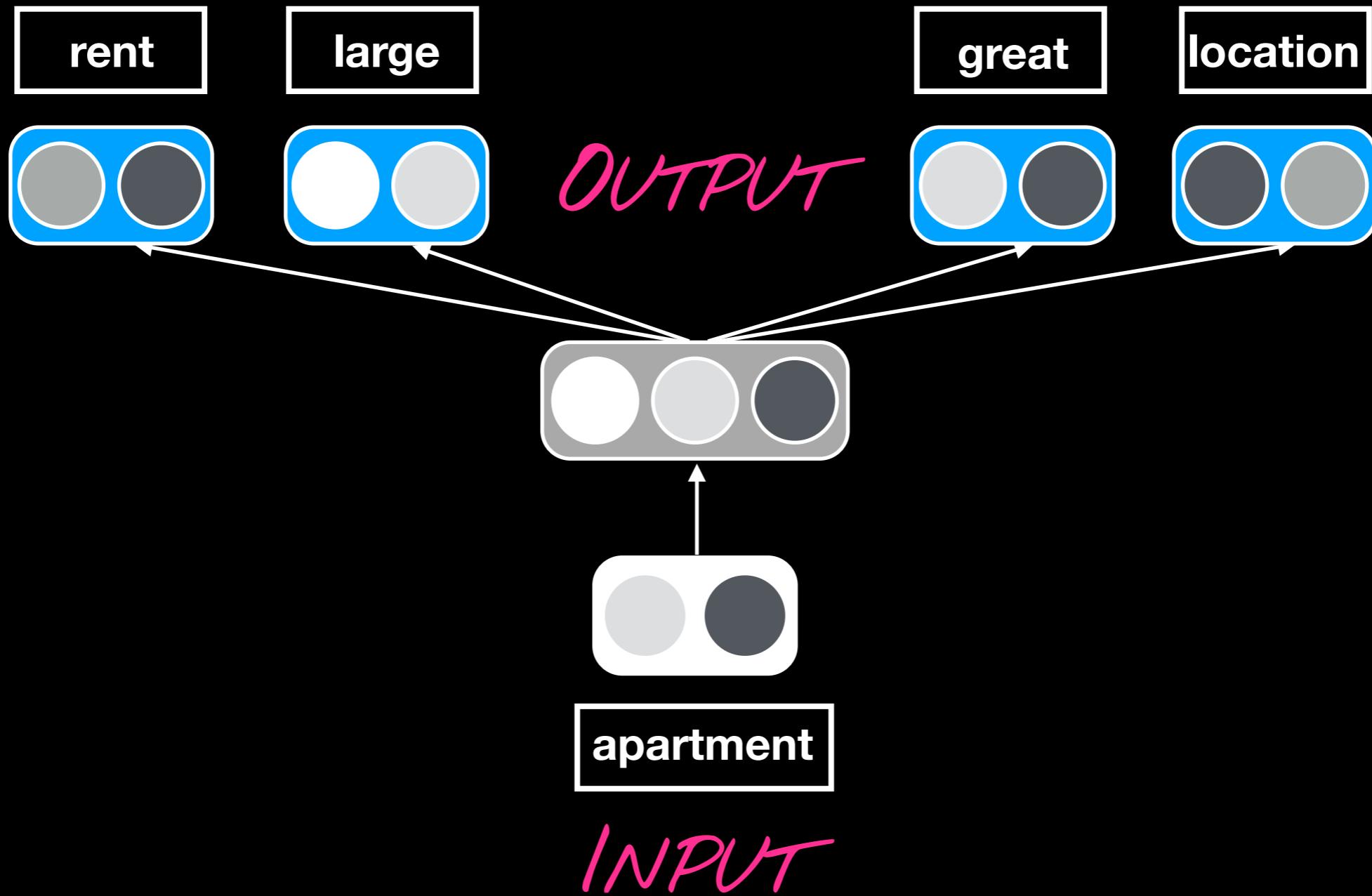
Word2Vec – Skipgram Model



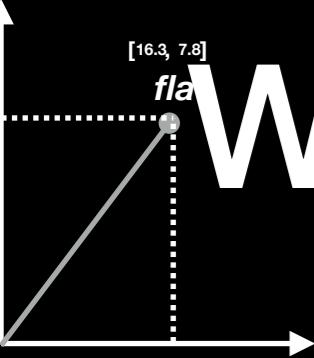
Renting out large *apartment* in great location



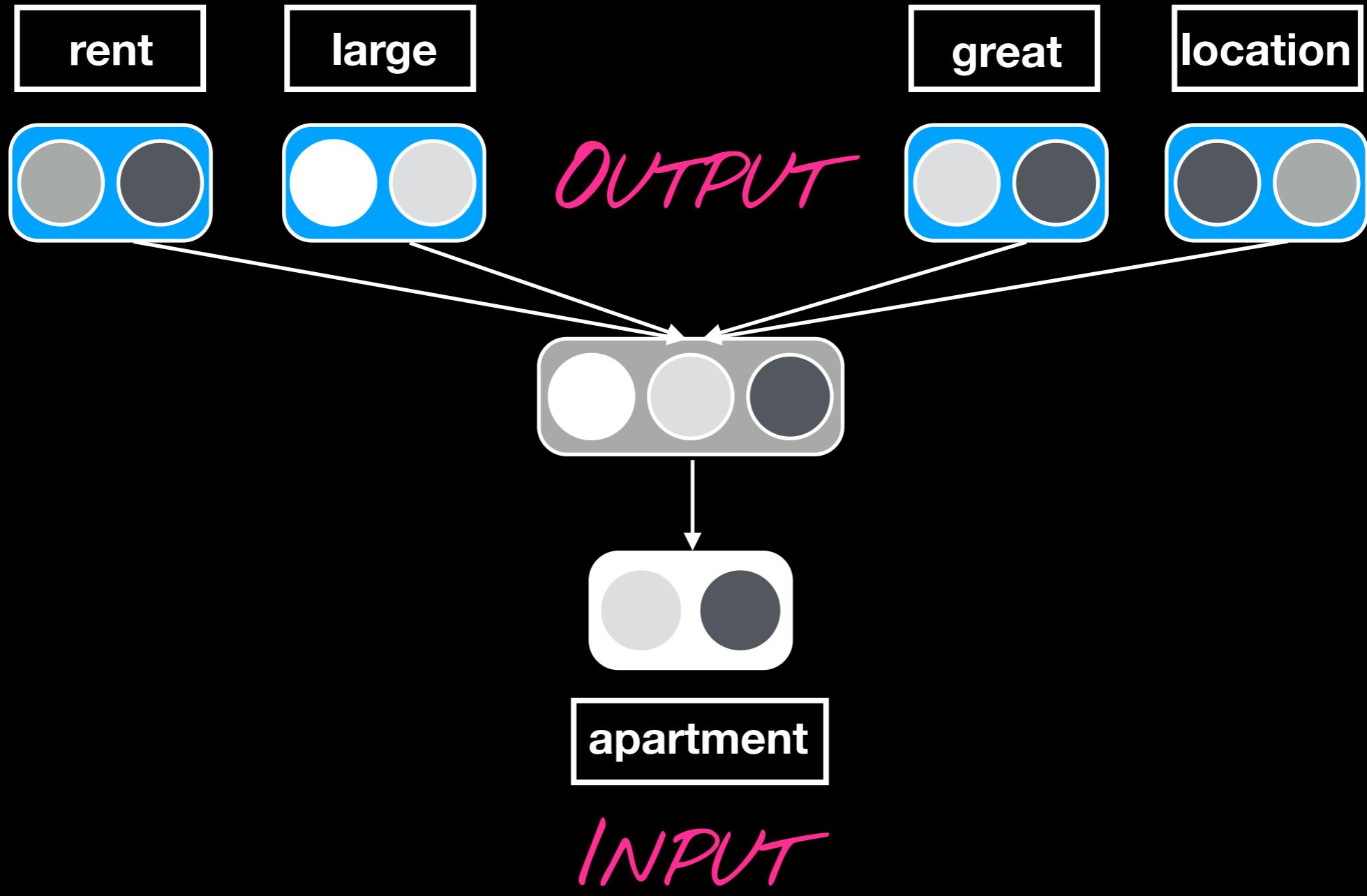
Word2Vec – Skipgram Model

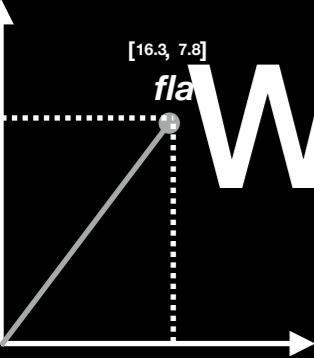


Renting out large *apartment* in great location

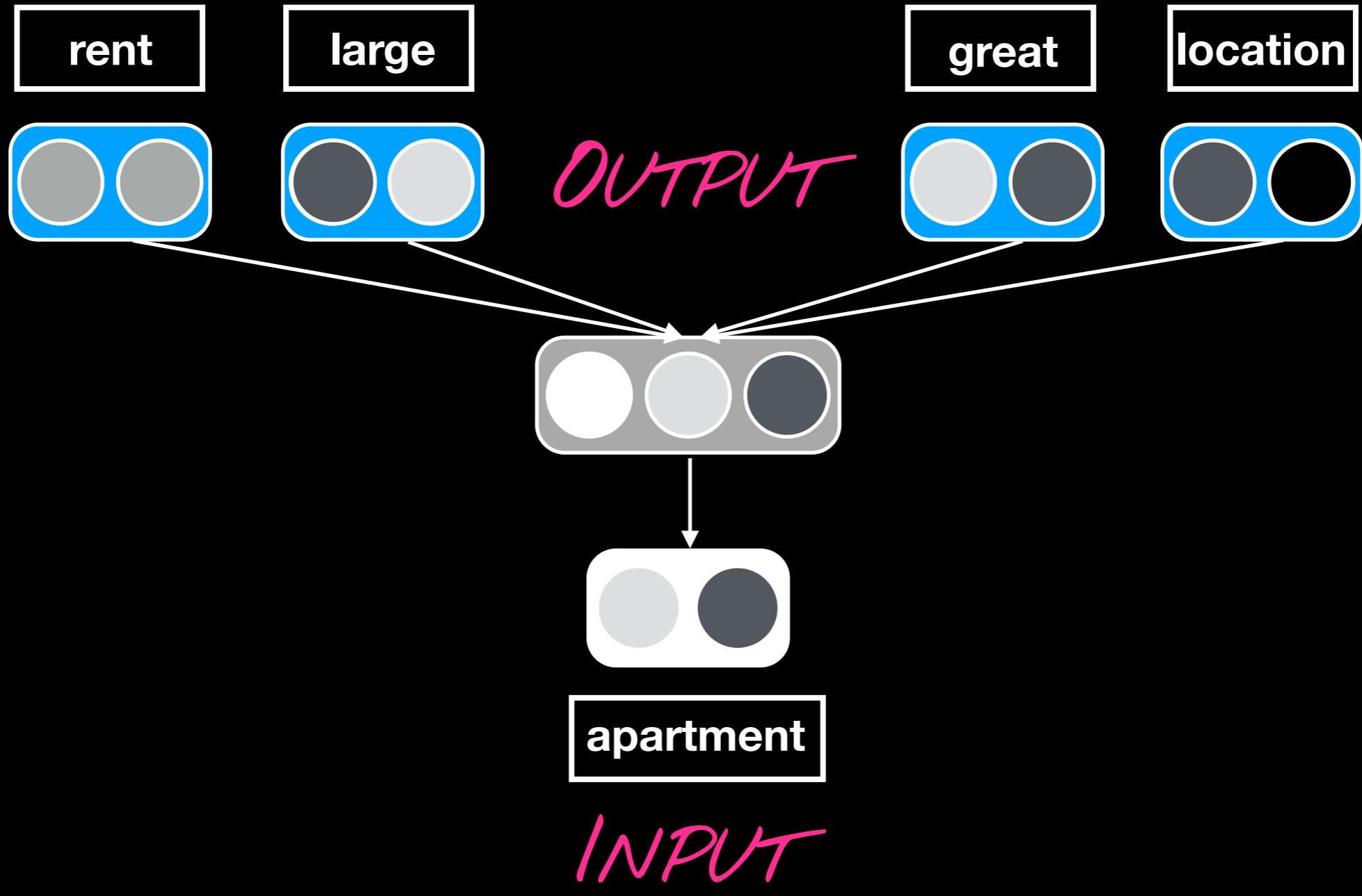


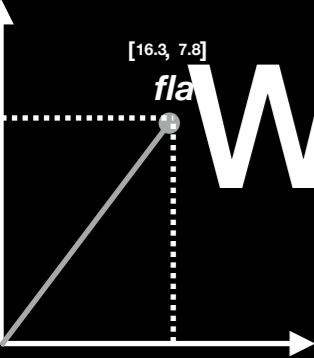
Word2Vec – Skipgram Model



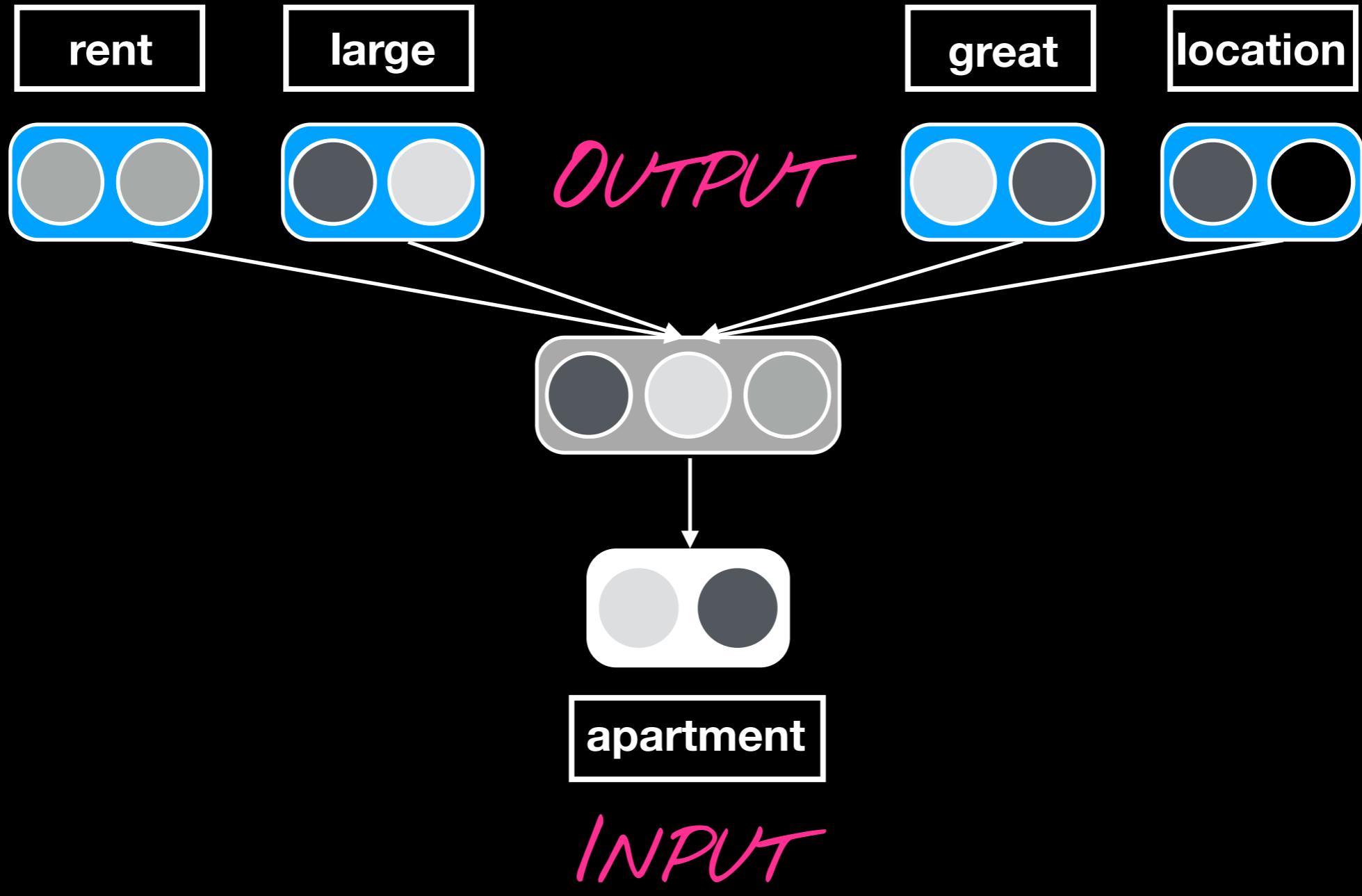


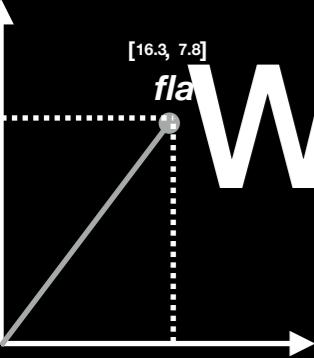
Word2Vec – Skipgram Model



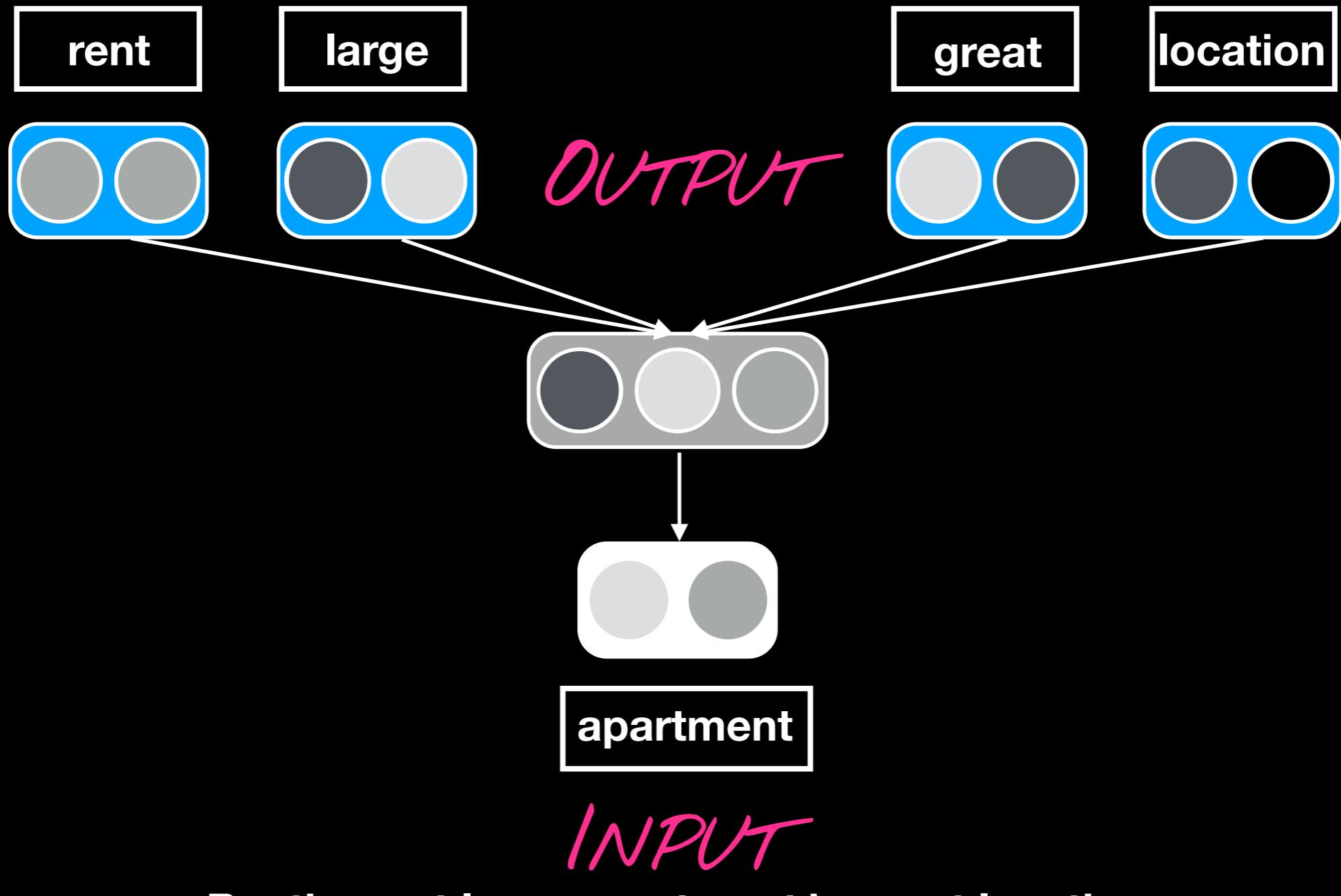


Word2Vec – Skipgram Model



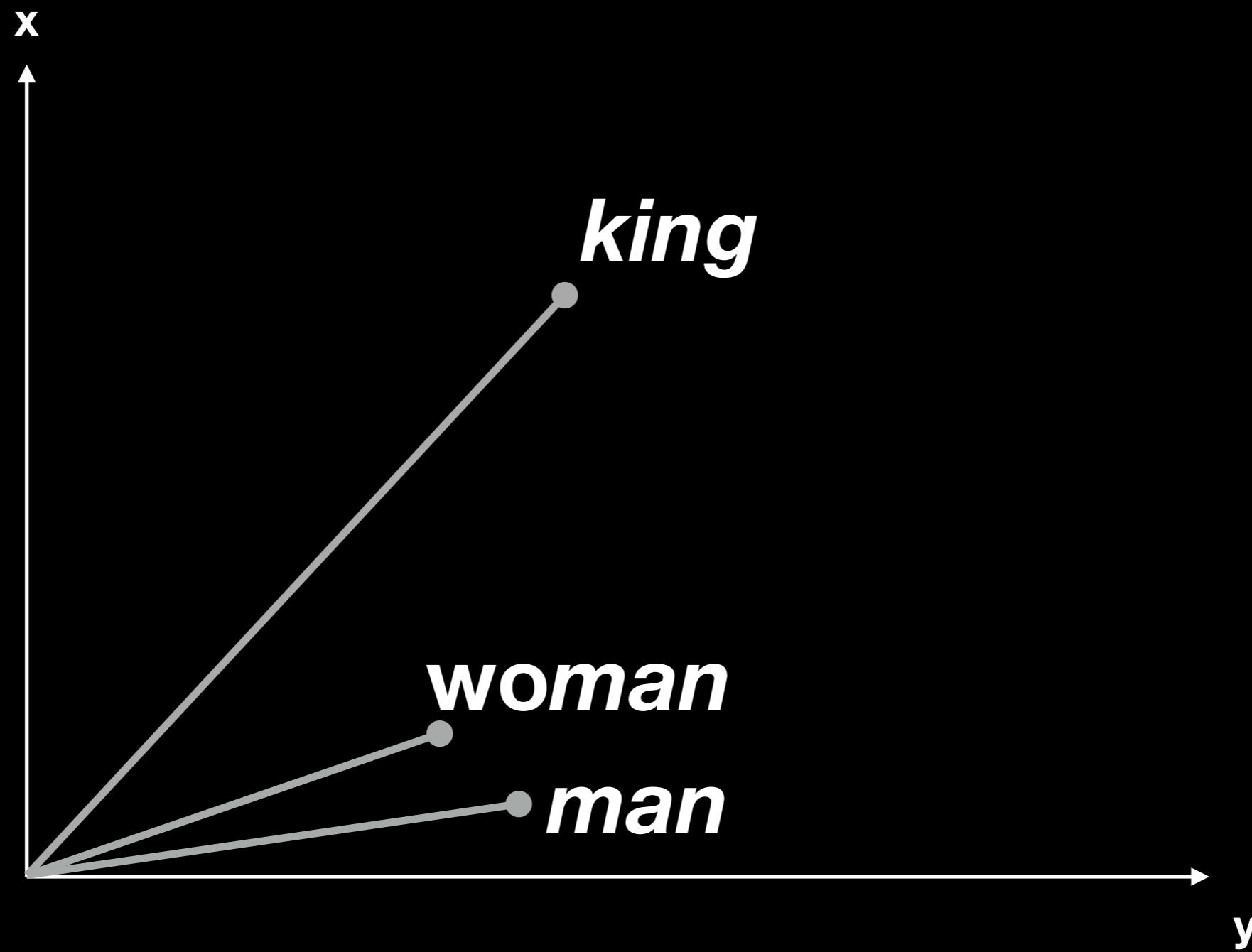


Word2Vec – Skipgram Model



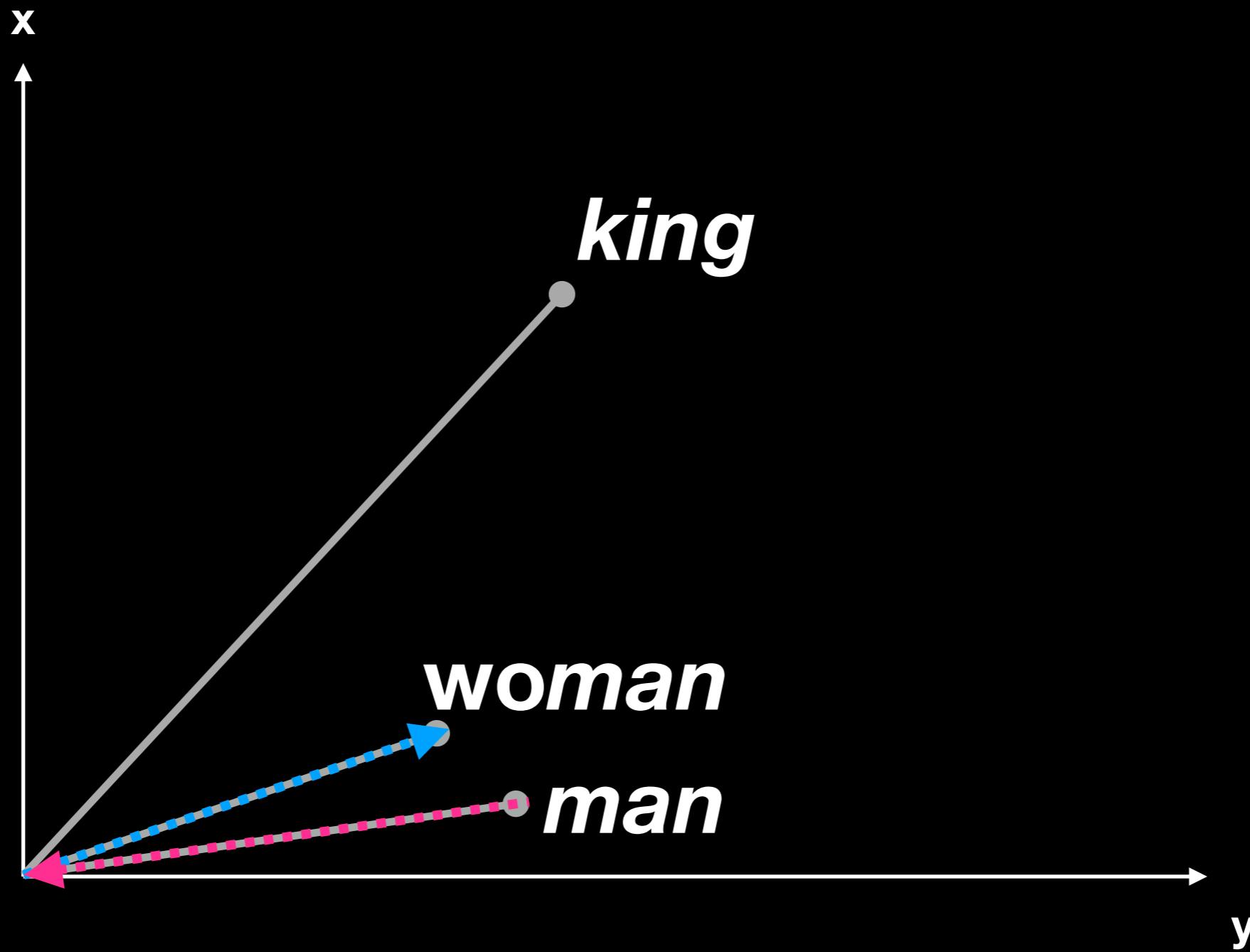
Vector Space Semantics

$$king - man + woman \approx ???$$



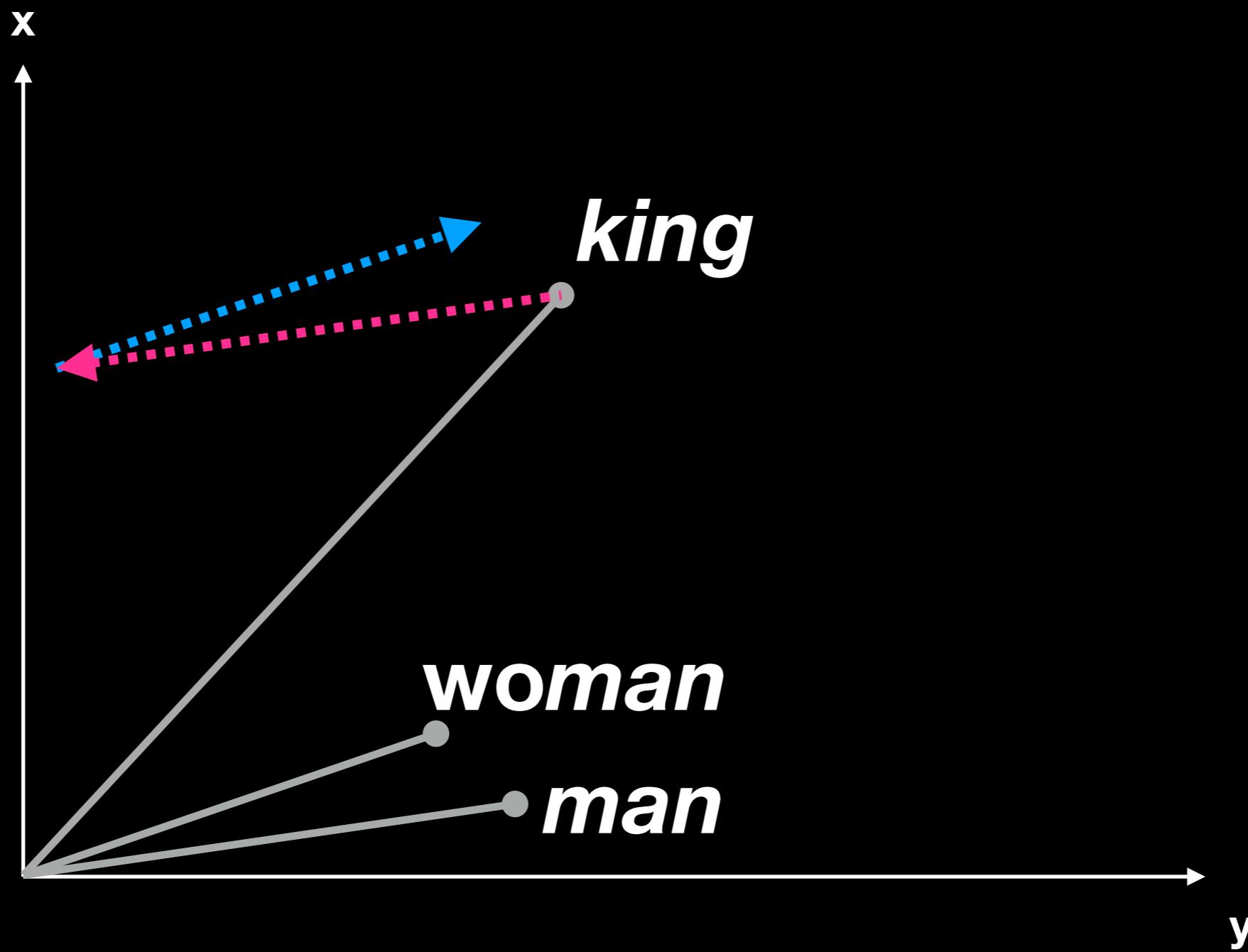
Vector Space Semantics

$$king - man + woman \approx ???$$



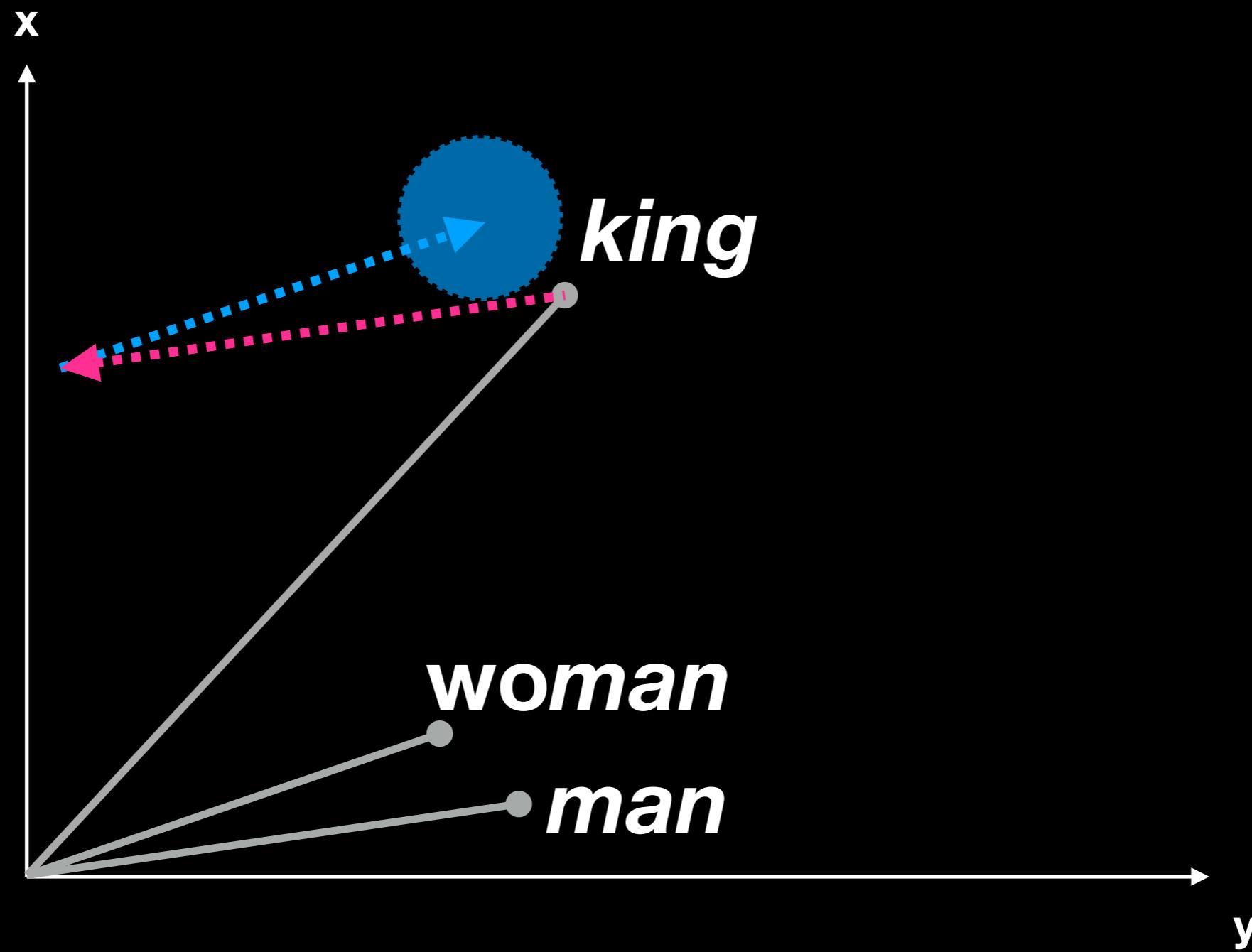
Vector Space Semantics

$$king - man + woman \approx ???$$



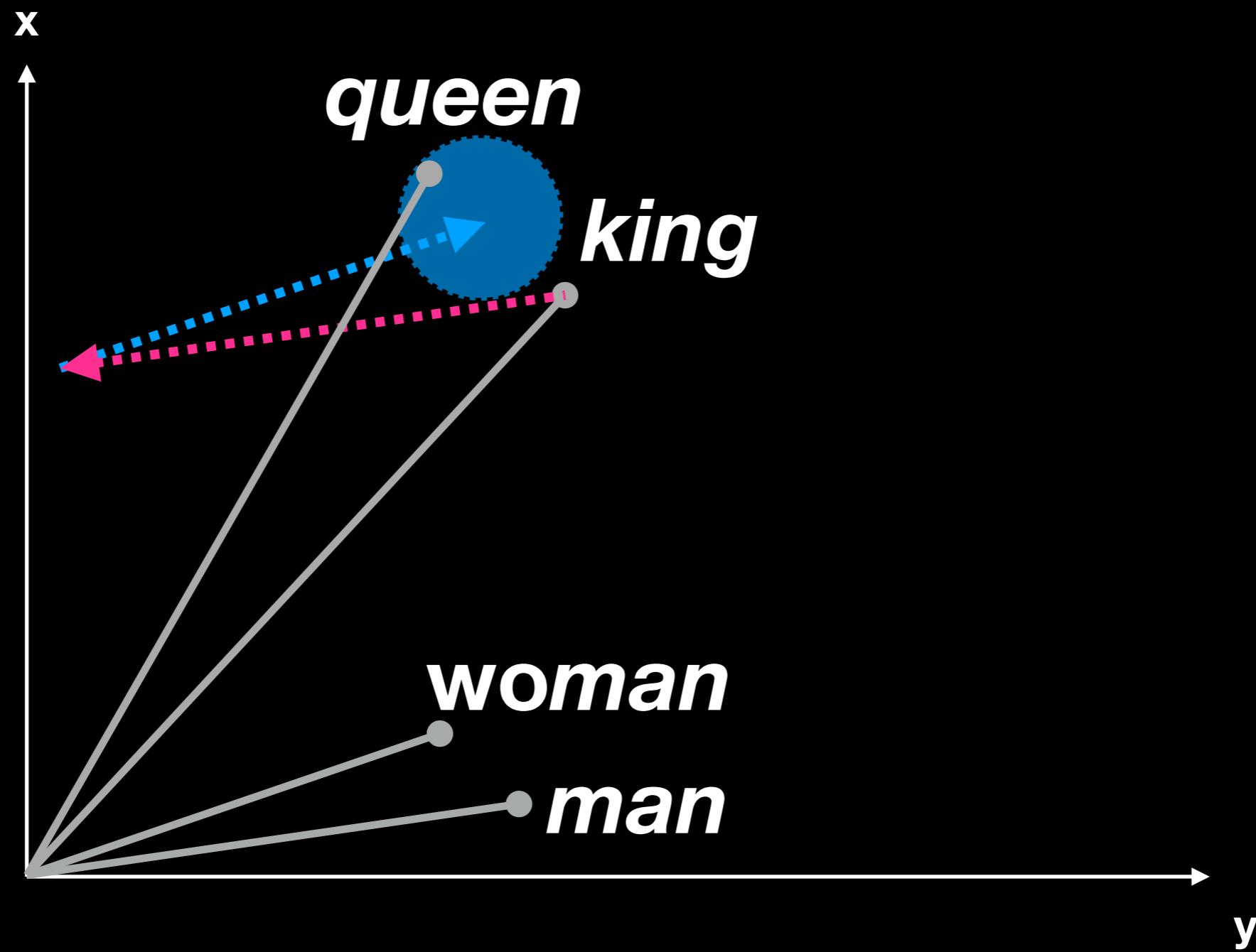
Vector Space Semantics

$$king - man + woman \approx ???$$



Vector Space Semantics

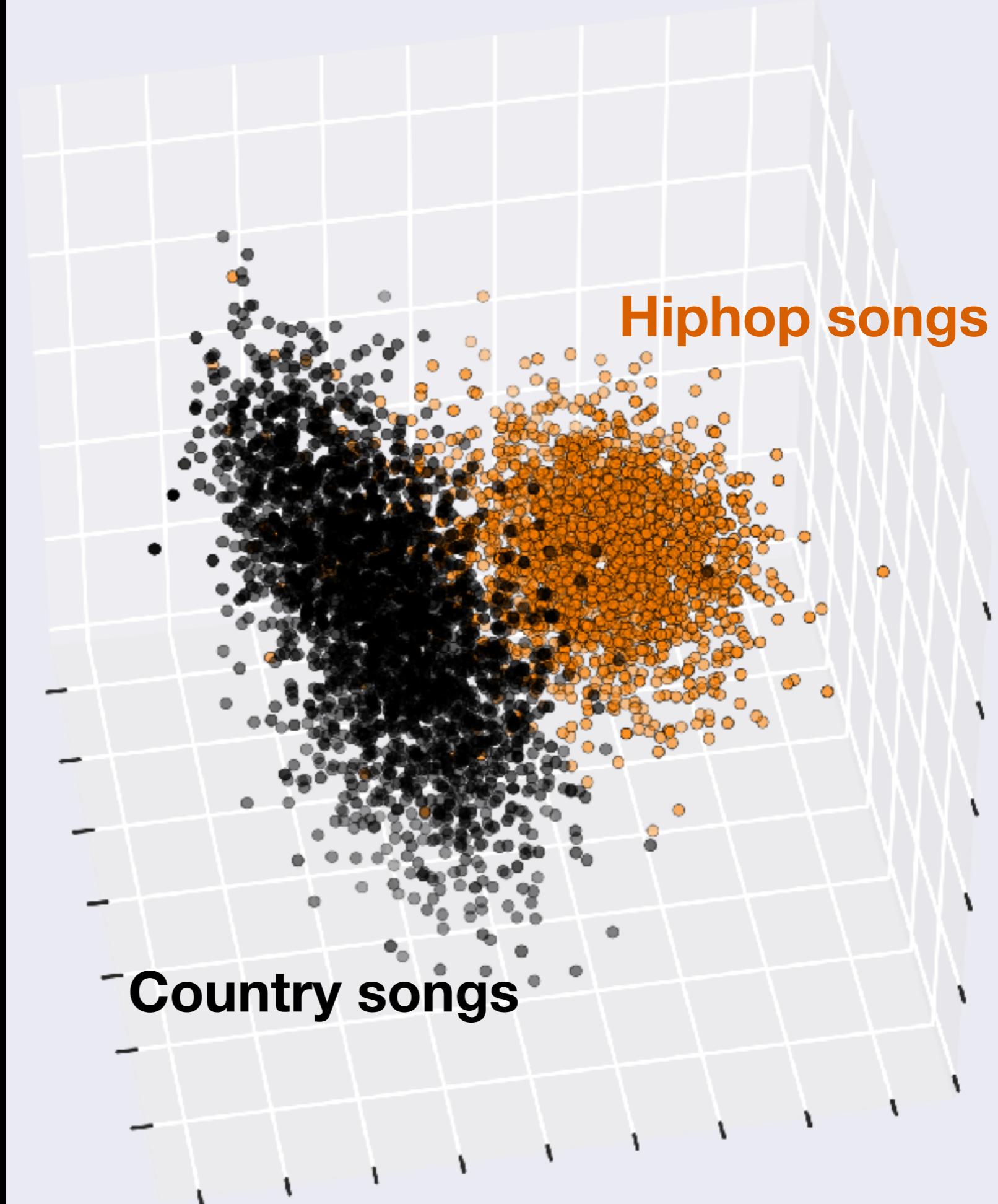
$$king - man + woman \approx queen$$



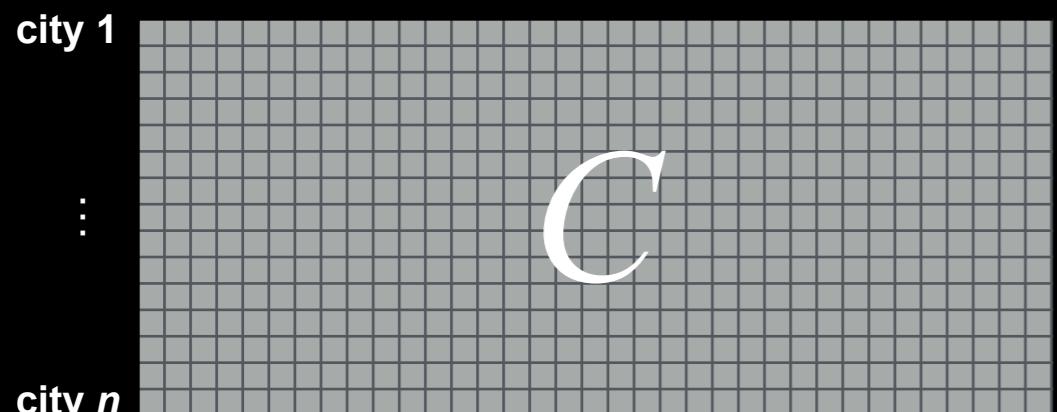
Part 2

Representing Documents as Vectors

Example 1: Songs

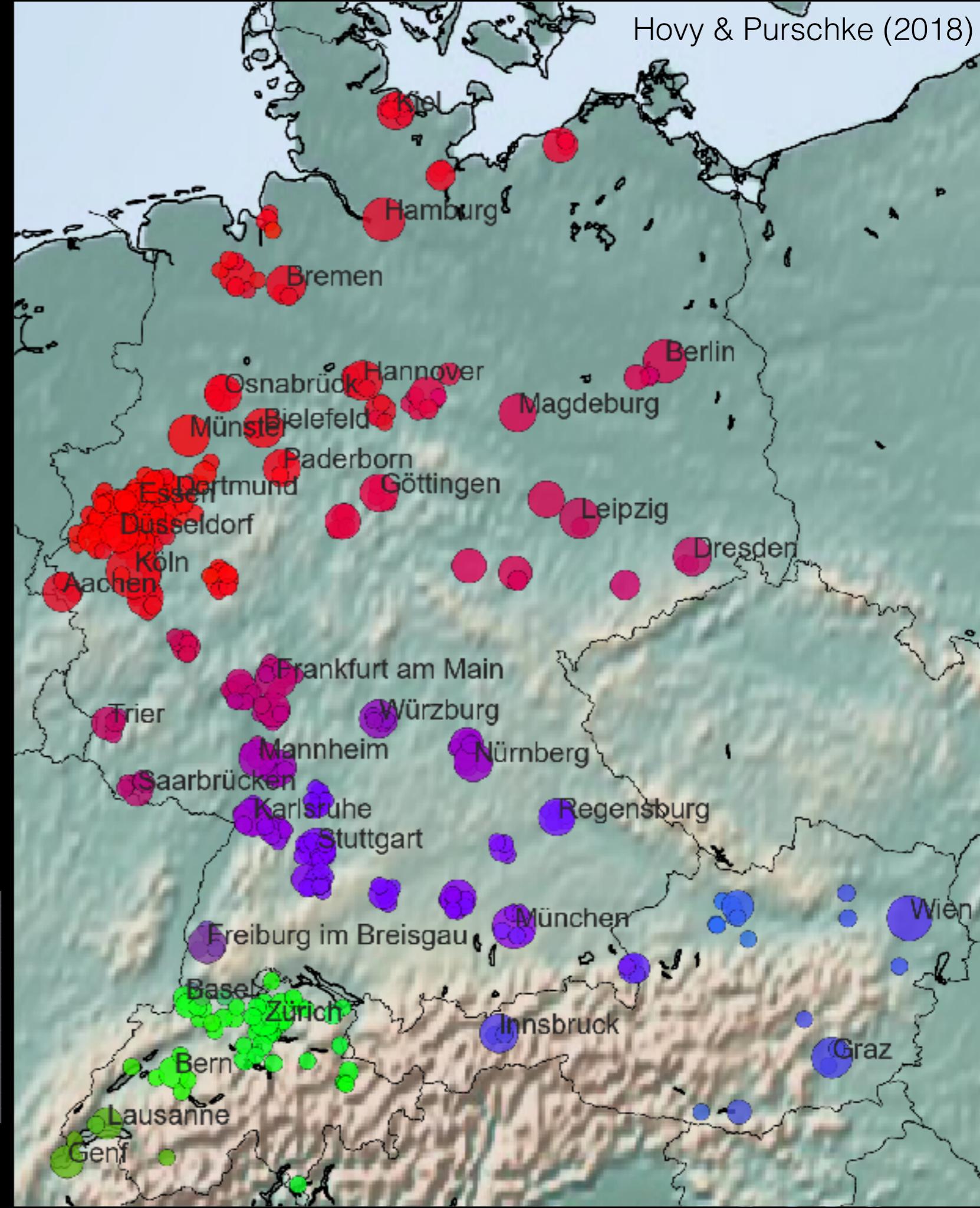


Example 2: Cities



Example 2: Cities

Hovy & Purschke (2018)



city 1

city *n*

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
```

Doc2Vec – Model

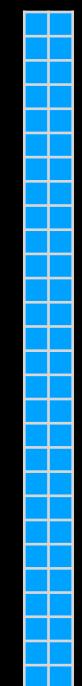
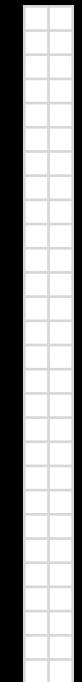
heute

echt

mal

beschweren

Hamburg



Doc2Vec – Model

heute

echt

mal

beschweren

C

Hamburg

Doc2Vec – Model

heute

echt

mal

beschweren

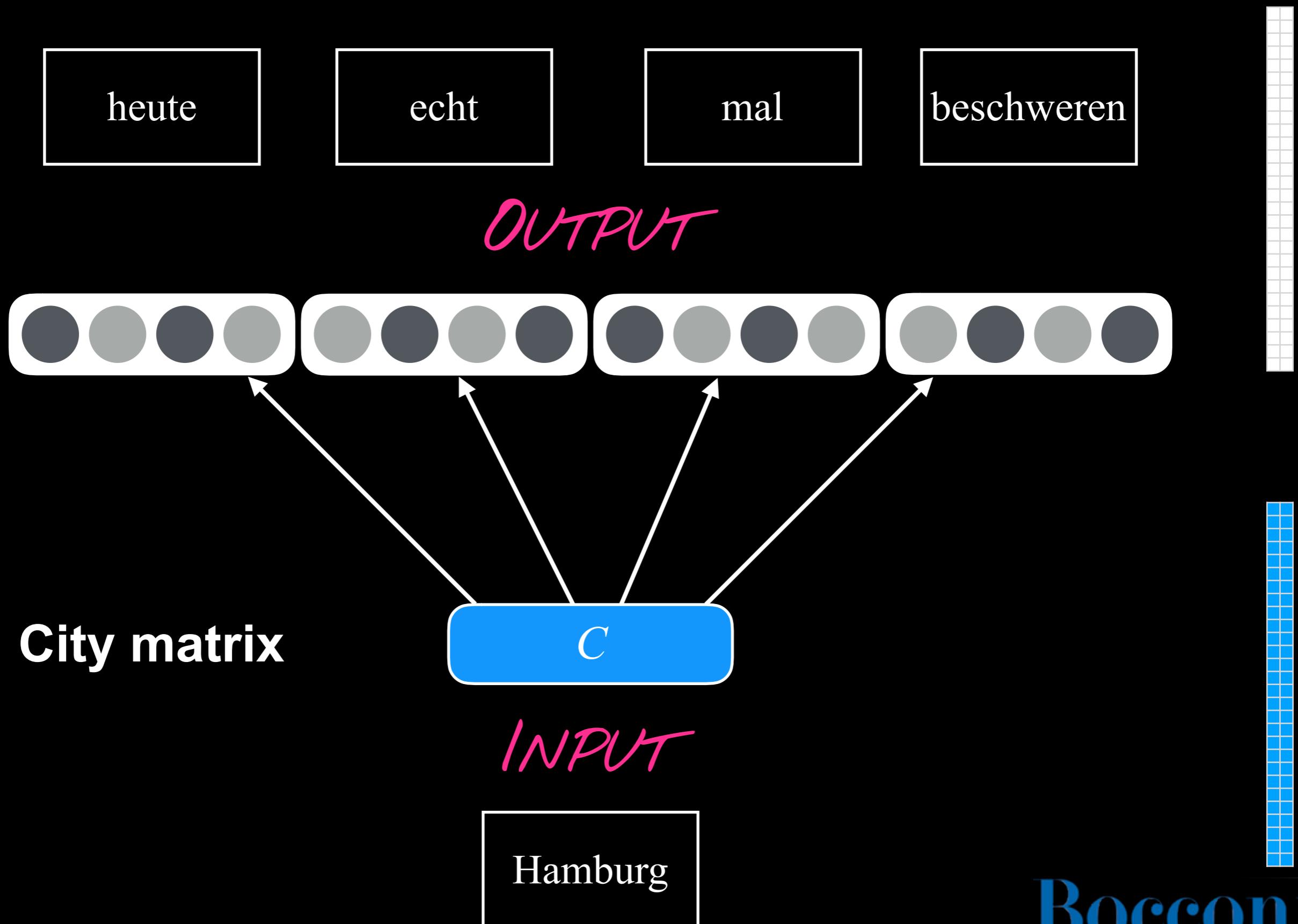
City matrix

C

Hamburg

Bocconi

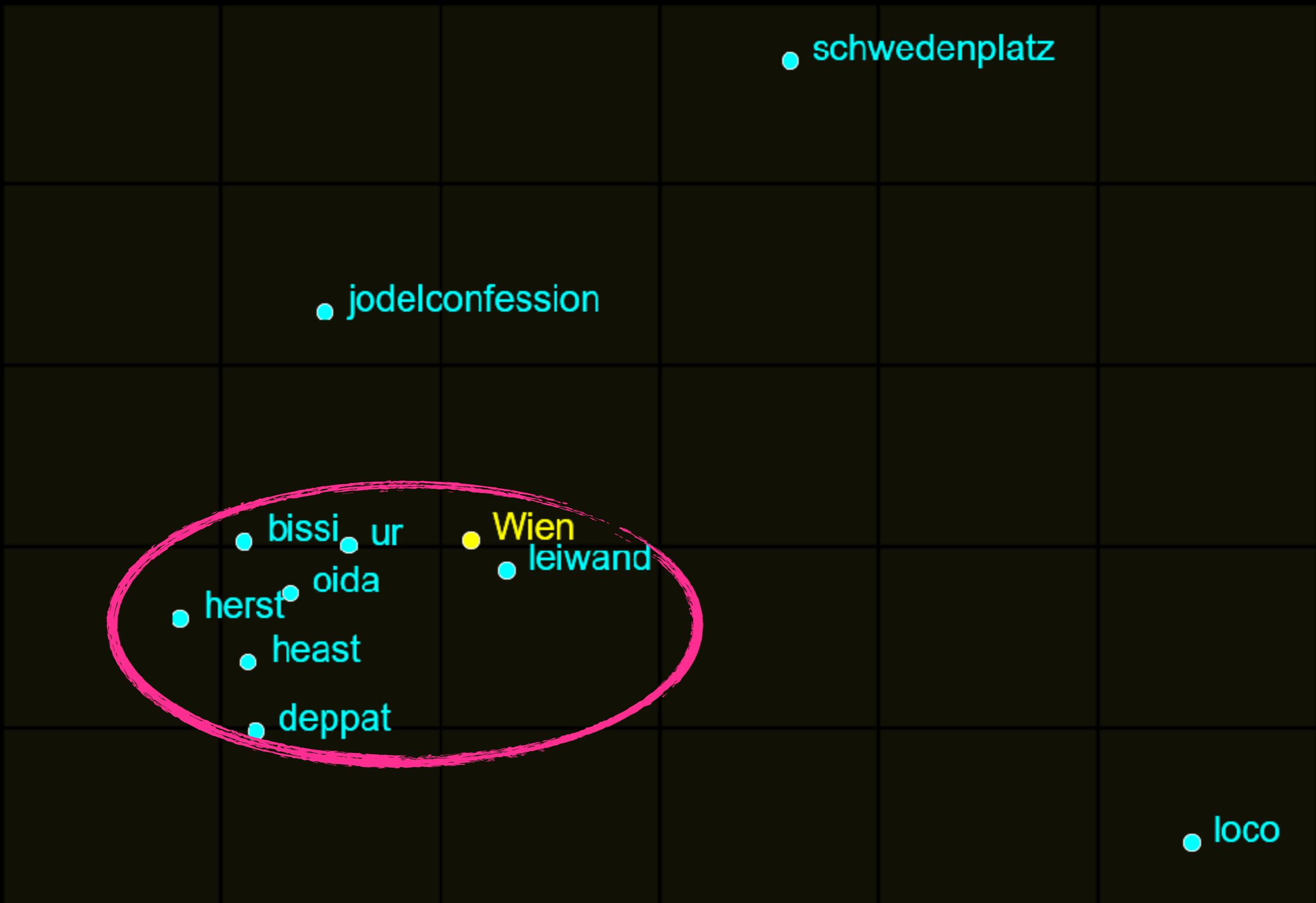
Doc2Vec – Model



Words and Documents



Words and Documents



Wrapping up...

Comparison

	Discrete	Distributed
Dimensions	Interpretable	Arbitrary
Content	Count-based	Coefficients
Density	Sparse	Dense
Strength	Interpretability	Similarity

Take home points

- Words and texts can be represented in two ways:
 - As **sparse, discrete** feature vectors over counts
 - As **dense, continuous** embedding vectors
- Discrete features allow us to **interpret** the input
- Word and document embeddings reflect **semantic similarity** in high-dimensional space