Distributional Hypothesis

CSCI 1460: Computational Linguistics

Lecture 5

Ellie Pavlick

Fall 2023

Announcements

- Assignment 2 (slight delay)

Quiz Recap

Quiz Recap

Candidate *H*

$$H_a = x \rightarrow \varnothing$$
$$H_b = x \rightarrow x + x$$
$$H_c = x \rightarrow x + x + 1$$
$$H_d = x \rightarrow x + 12^x - 22^x$$

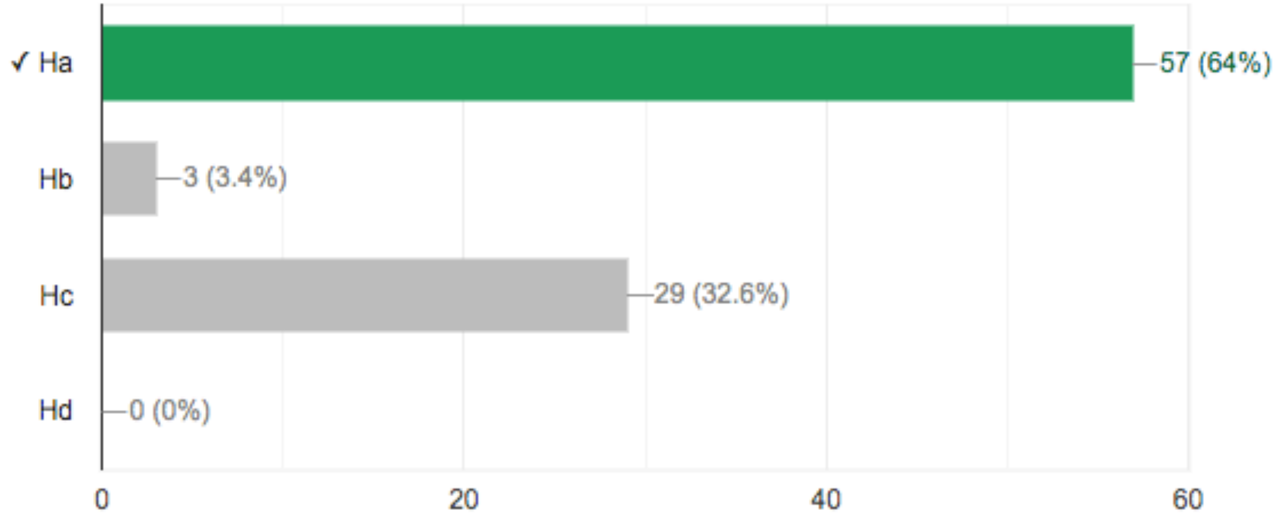
Observed *D*

Input	Output
4	9
2	5
1	3

In the context of bayesian language of thought, which hypothesis would we generally have the highest prior for?

Copy

57 / 89 correct responses



Quiz Recap

Candidate *H*

$$H_a = x \rightarrow \varnothing$$
$$H_b = x \rightarrow x + x$$
$$H_c = x \rightarrow x + x + 1$$
$$H_d = x \rightarrow x + 12^x - 22^x$$

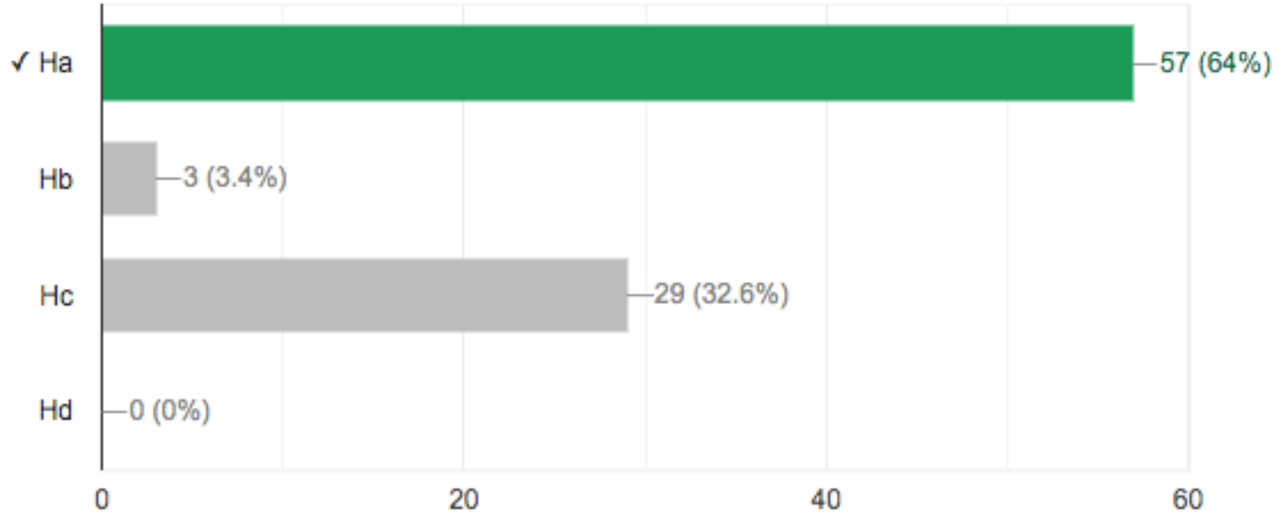
Observed *D*

Input	Output
4	9
2	5
1	3

In the context of bayesian language of thought, which hypothesis would we generally have the highest prior for?

Copy

57 / 89 correct responses



Hc better explains the data,
But is less likely as it has a
longer description length.

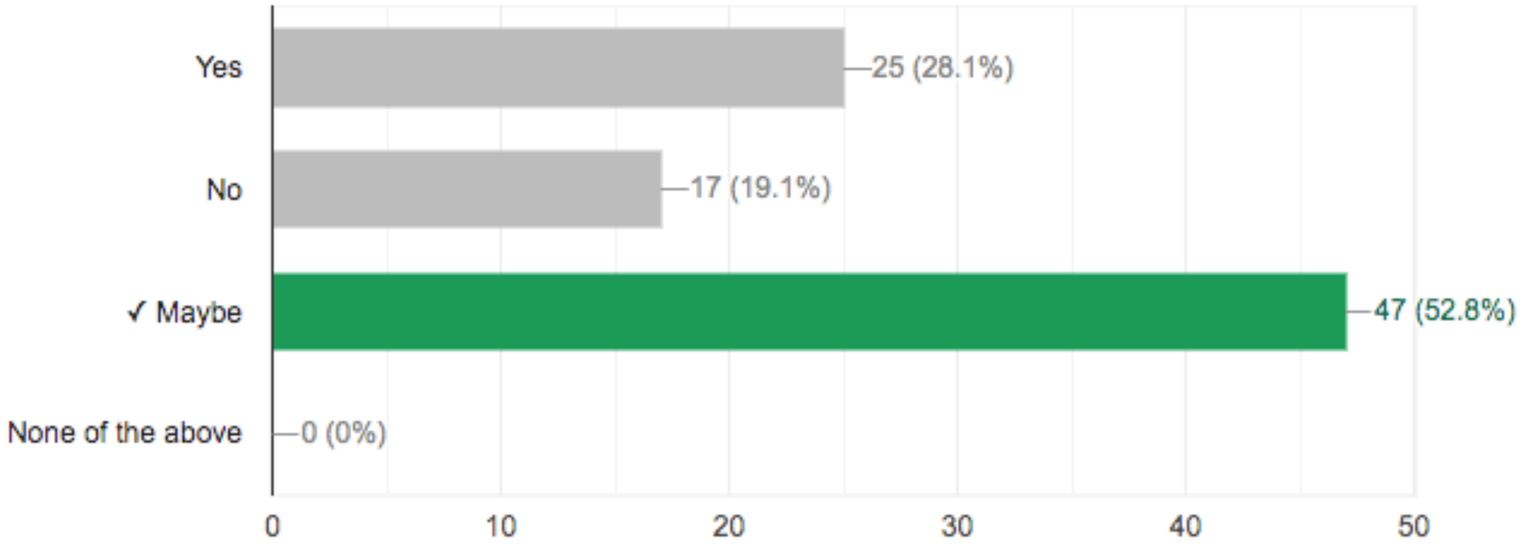
Quiz Recap

Copy

Assume:
P(Hb) = 0.04
P(Hc) = 0.01
P(D | Hb) = 0.4
P(D | Hc) = 0.99

If we were using Hastings Algorithm as we described it in class, and H = Hb and H' = Hc, would we accept the new H'?

47 / 89 correct responses



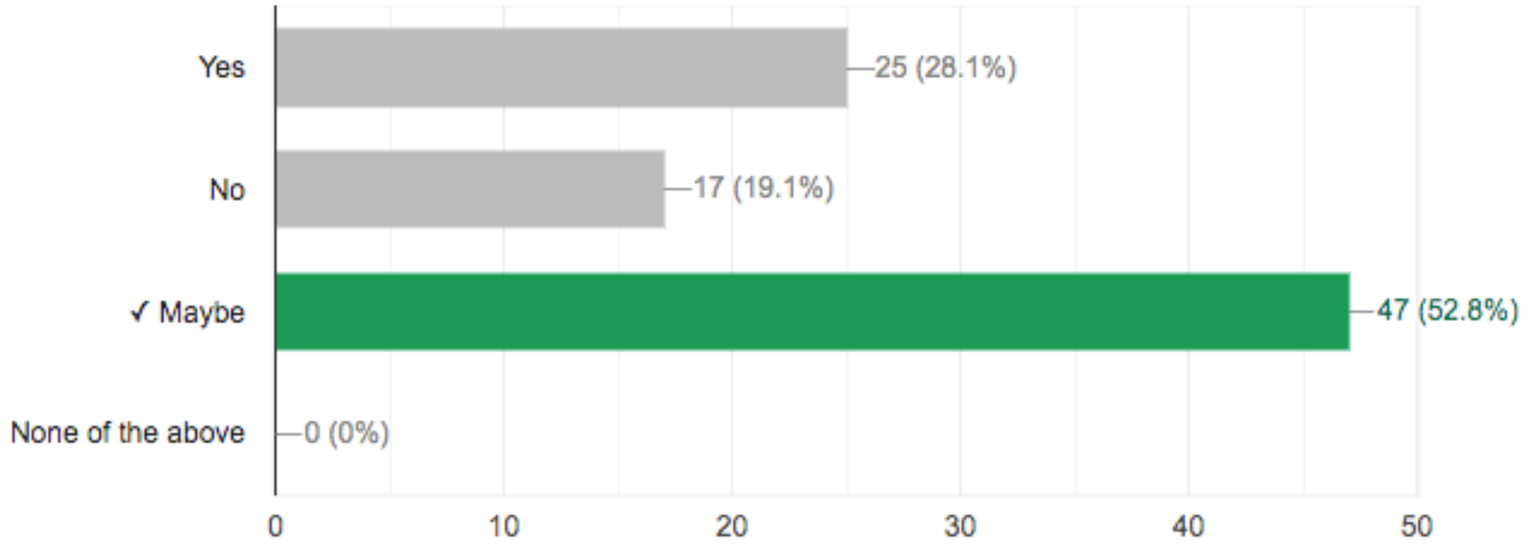
Quiz Recap

Copy

Assume:
P(Hb) = 0.04
P(Hc) = 0.01
P(D | Hb) = 0.4
P(D | Hc) = 0.99

If we were using Hastings Algorithm as we described it in class, and H = Hb and H' = Hc, would we accept the new H'?

47 / 89 correct responses



$$\frac{P(D|Hc)P(Hc)}{P(D|Hb)P(Hb)} = \frac{0.01 \times 0.99}{0.4 \times 0.04} \approx 0.6 < 1$$

Therefore, maybe.

Topics

- Theories of Word Meaning
- Vector Space Models (VSMs)
- Word Embeddings, Part 1

Topics

- **Theories of Word Meaning**
- Vector Space Models (VSMs)
- Word Embeddings, Part 1

What is the “meaning” of a word?

Naive BOW model has no notion of word meaning

	cat	kitten	cute	adorable	gradients
doc1	1	0	1	0	0
doc2	0	1	0	1	0
doc3	1	0	0	0	1
doc4	0	1	0	0	1
doc5	0	0	0	0	1

What is the “meaning” of a word?

Naive BOW model has no notion of word meaning

doc1 and doc2 are completely orthogonal

	cat	kitten	cute	adorable	gradients
doc1	1	0	1	0	0
doc2	0	1	0	1	0
doc3	1	0	0	0	1
doc4	0	1	0	0	1
doc5	0	0	0	0	1

What is the “meaning” of a word?

Naive BOW model has no notion of word meaning

treated as as different as doc1 and doc5

	cat	kitten	cute	adorable	gradients
doc1	1	0	1	0	0
doc2	0	1	0	1	0
doc3	1	0	0	0	1
doc4	0	1	0	0	1
doc5	0	0	0	0	1

What is the “meaning” of a word?

Naive BOW model has no notion of word meaning

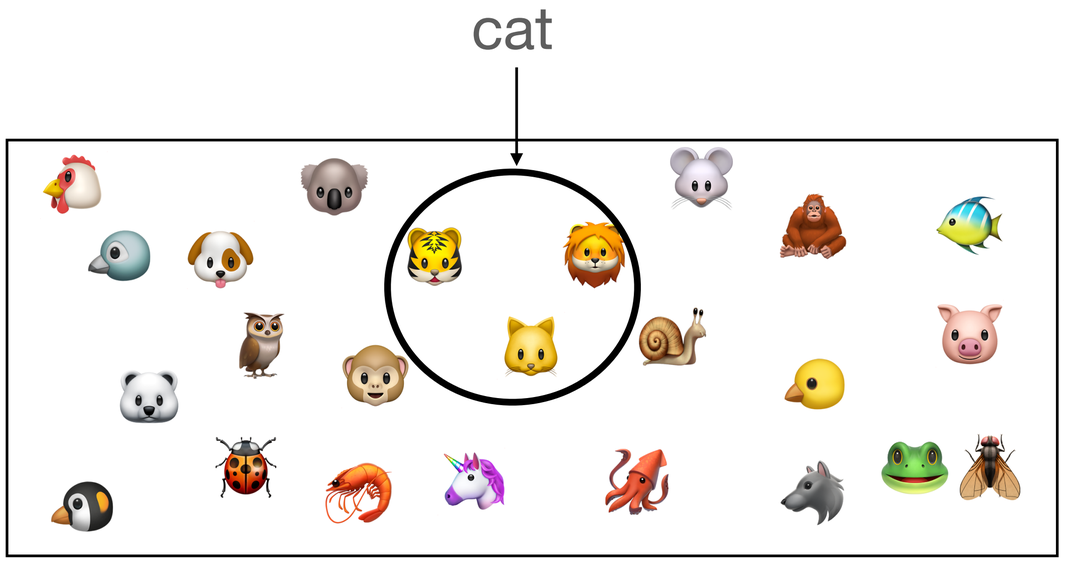
ideally, model should know that cat and kitten have similar meanings

	cat	kitten	cute	adorable	gradients
doc1	1	0	1	0	0
doc2	0	1	0	1	0
doc3	1	0	0	0	1
doc4	0	1	0	0	1
doc5	0	0	0	0	1

What is the meaning of a word?

Theories of word meaning

- Words refer to sets

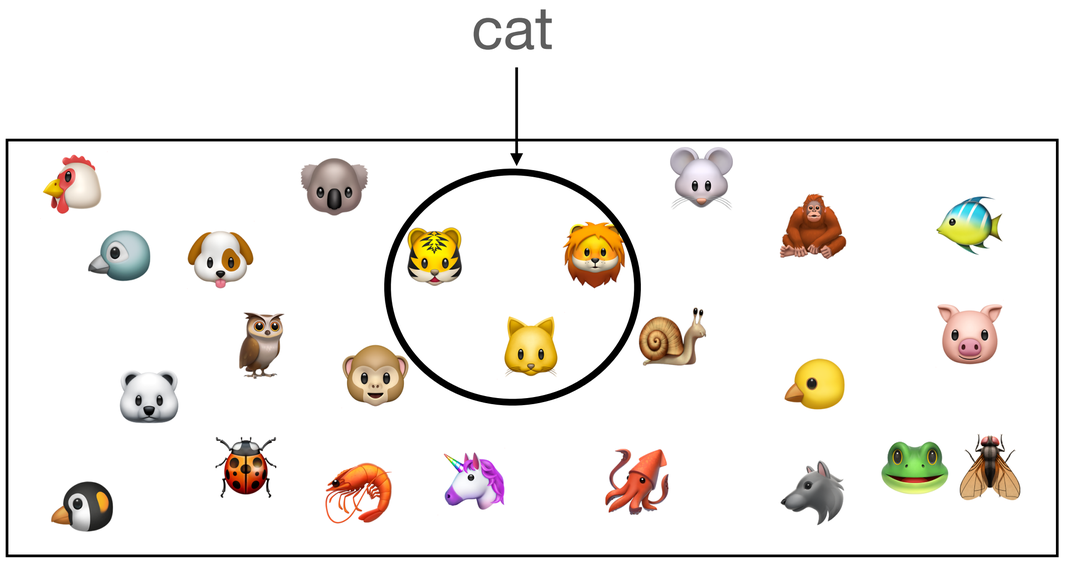


{x | x is a small domesticated carnivore, Felis domestica or F. catus, bred in a number of varieties.}

What is the meaning of a word?

Theories of word meaning

- Words refer to sets

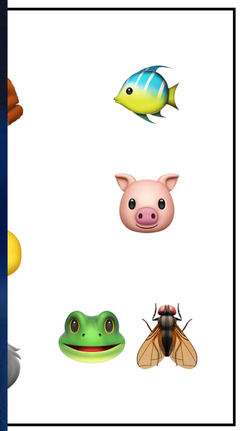


Problem: set boundaries are difficult (impossible?) to define…

What is the meaning of a word?

Theories of word meaning

- Words r



difficult

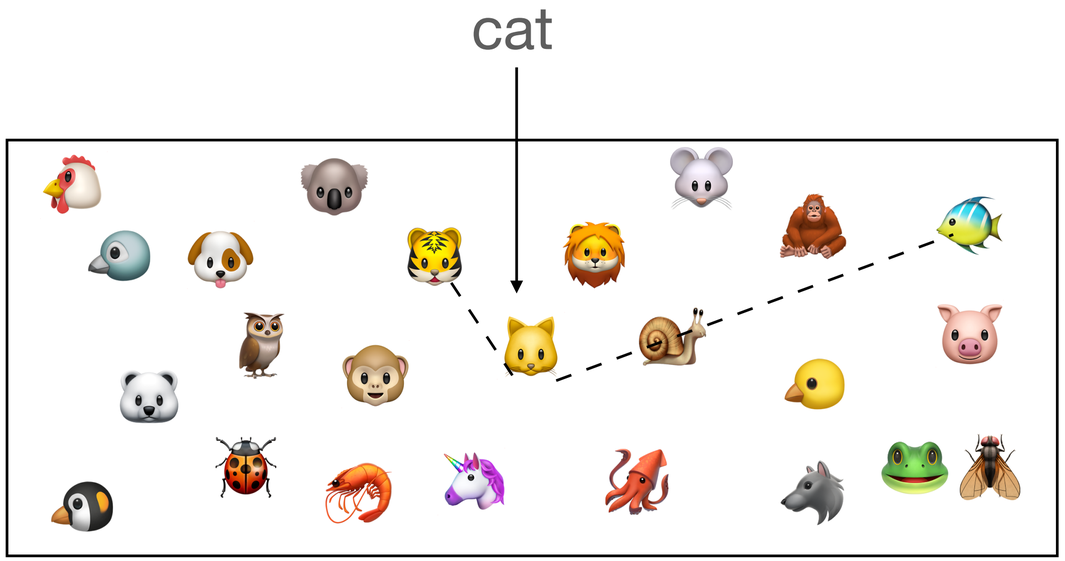
…

Thanks Roman Feiman for video reference! :-D

What is the meaning of a word?

Theories of word meaning

- Words refer to sets

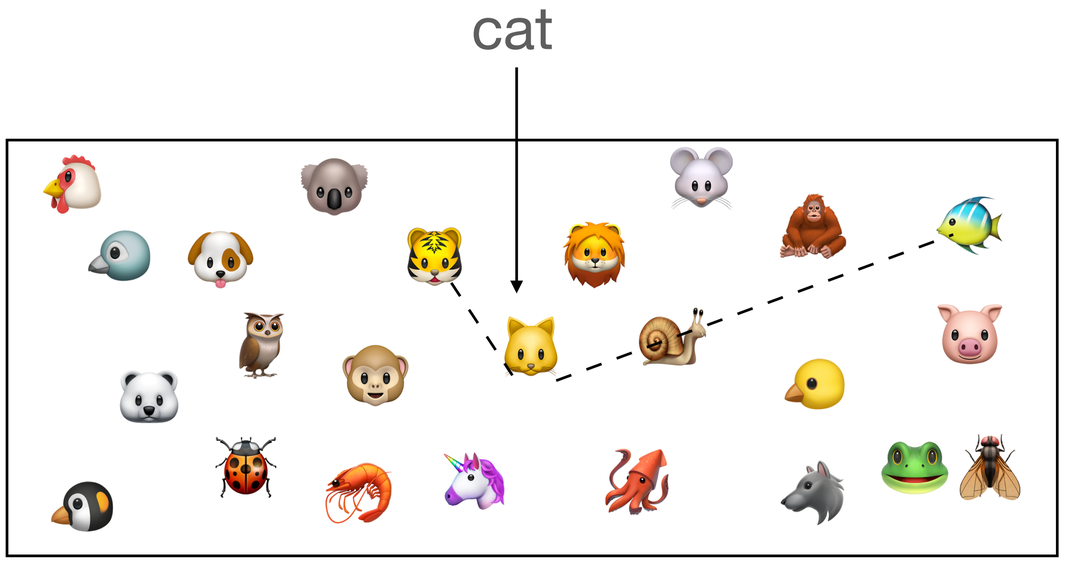


Problem: set boundaries are difficult (impossible?) to define…

What is the meaning of a word?

Theories of word meaning

- Words refer to sets
- Words refer to things (prototypes, exemplars)



cat = weighted combination([whiskers, ears, tail, fur….eats food, need oxygen])

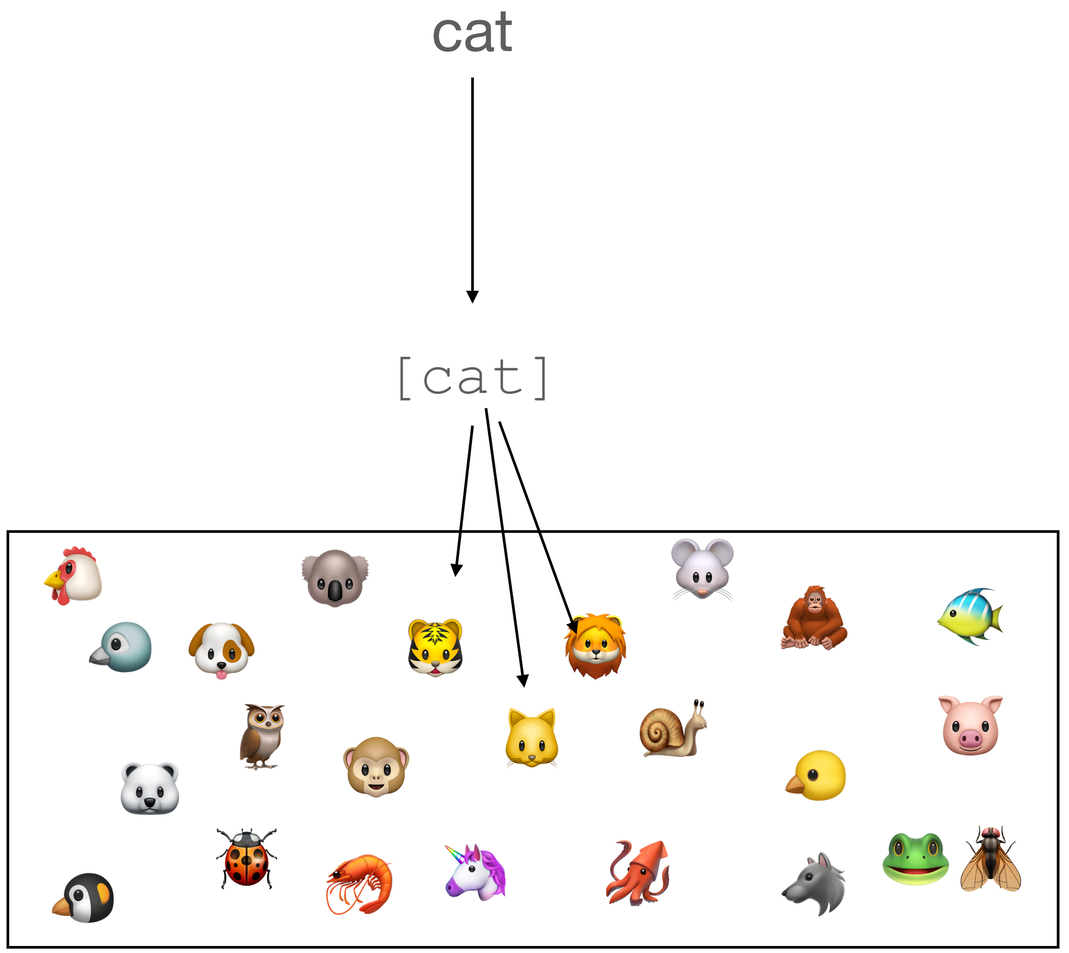
cat = 🐱

What is the meaning of a word?

Theories of word meaning

- Words refer to sets
- Words refer to things (prototypes, exemplars)
- Words refer to concepts (internal mental “symbols”)

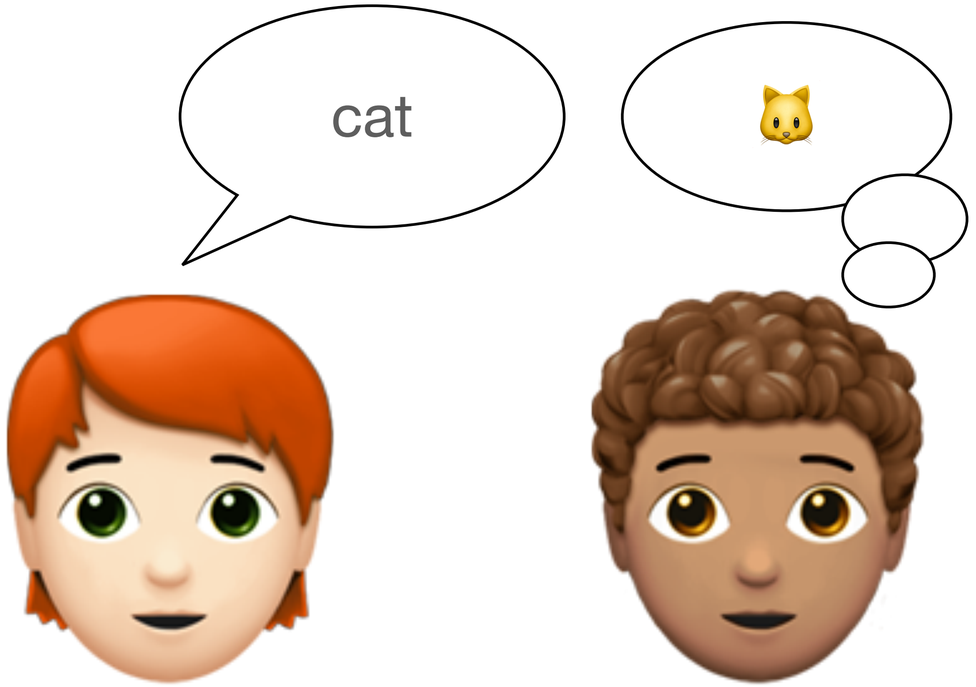
Tons of theory, philosophy, and debate on what exactly this symbol is. Happy to talk more in OH!



What is the meaning of a word?

Theories of word meaning

- Words refer to sets
- Words refer to things (prototypes, exemplars)
- Words refer to concepts (internal mental “symbols”)
- Words are defined by the contexts in which they are used



What is the meaning of a word?

The Distributional Hypothesis

- The meaning of a word is defined by its context
- “I know a word by the company it keeps”
- Usually credited to Harris and Firth
- Also Wittenstein
  - “Meaning in use”
  - “Language Games” —> words are optimized as tools for communication

What is the meaning of a word?

The Distributional Hypothesis

What is the meaning of a

The Distributional Hypothesis

the book?
the picture?
the sound?
our reactions?
the room?
the temperature in the room?
the thing that happened just before?
the thing that will happen just after?

What is the meaning of a word?

The Distributional Hypothesis

- Meaning depends on **“context”**, so the work of what meaning is is pushed into the definition of what “context” is
- “Context” can mean:
  - Perceptual (“grounded”) context or linguistic context
  - Symbolic features (fur=1, scales=0) or real-valued “impressions” ()
  - First order associations or higher-order abstraction?
    - If higher-order, where does this come from?

What is the meaning of a word?

The Distributional Hypothesis

- Strengths
  - (Computationally) elegant: clear how words are “learned” according to the theory
  - It works! Very well, in a lot of cases. Distributional models correlate well with lots of data on humans
- Weaknesses
  - Its “holistic” —> every context affects meaning; meaning is always changing
  - What about rare words, or words that have never been seen?
  - What about sentences??
  - (NLP might have some answers to these criticisms…)

Topics

- Theories of Word Meaning
- **Vector Space Models (VSMs)**
- Word Embeddings, Part 1

Vector Space Models

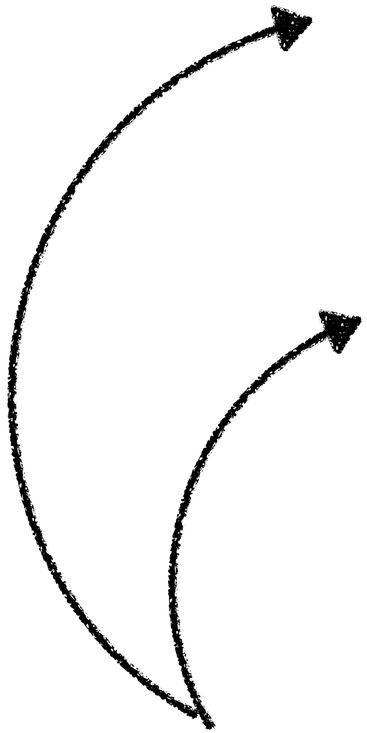
Definition

- Represent words as vectors, i.e., as points in space
- Words that have similar meaning are nearby in space

A Simple Distributional VSM

	cat	kitten	cute	adorable	gradients
doc1	1	0	1	1	0
doc2	0	1	0	1	0
doc3	1	0	1	1	1
doc4	0	1	0	0	1
doc5	0	0	0	0	1

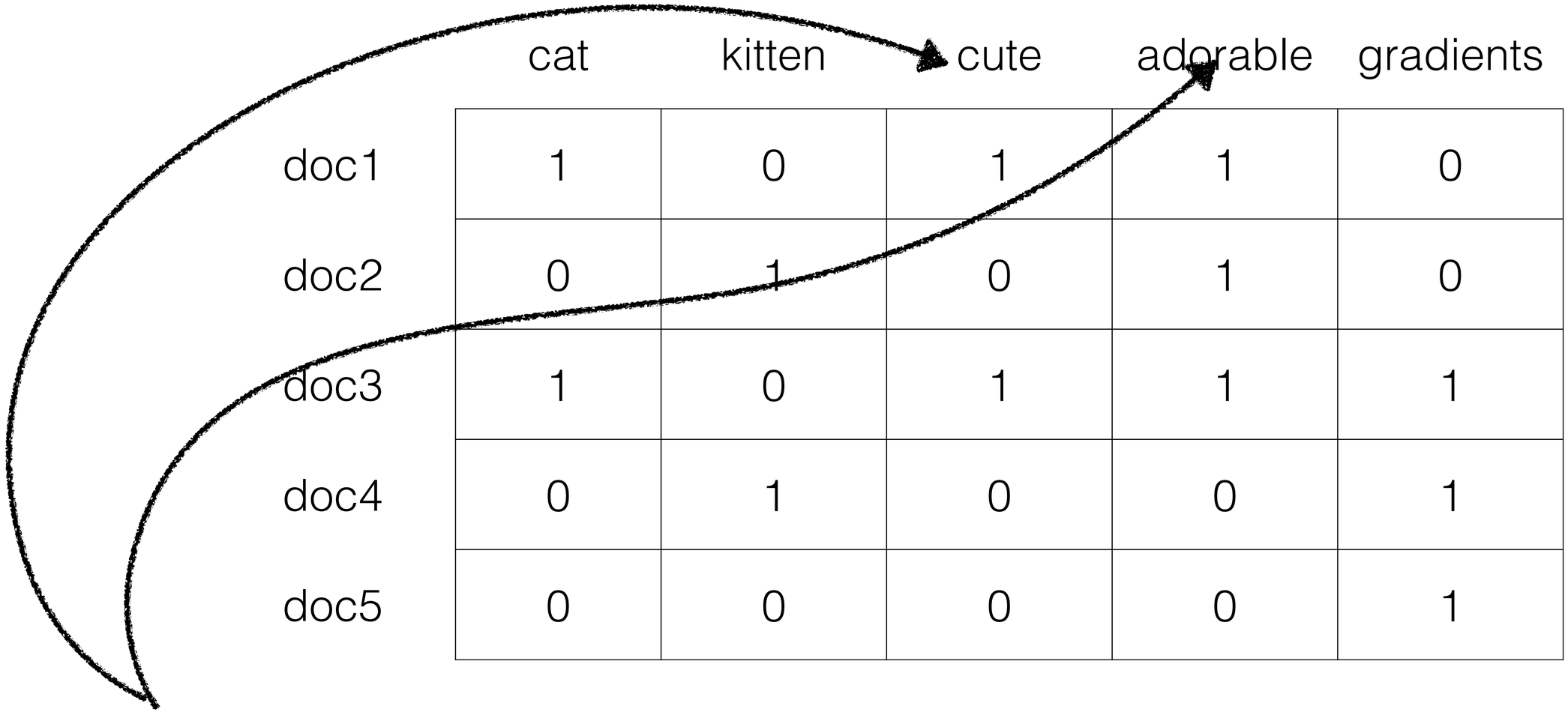
A Simple Distributional VSM



	cat	kitten	cute	adorable	gradients
doc1	1	0	1	1	0
doc2	0	1	0	1	0
doc3	1	0	1	1	1
doc4	0	1	0	0	1
doc5	0	0	0	0	1

similar documents tend to contain the same words....

A Simple Distributional VSM



	cat	kitten	cute	adorable	gradients
doc1	1	0	1	1	0
doc2	0	1	0	1	0
doc3	1	0	1	1	1
doc4	0	1	0	0	1
doc5	0	0	0	0	1

similar words tend to occur in the same documents....

A Simple Distributional VSM

	doc1	doc2	doc3	doc4	doc5
cat	1	0	1	0	0
kitten	0	1	0	1	0
cute	1	0	1	0	0
adorable	1	1	1	0	0
gradients	0	0	1	1	1

term-document matrix:
word meaning = set of documents in which it occurs

A Simple Distributional VSM

	doc1	doc2	doc3	doc4	doc5
cat	1	0	1	0	0
kitten	0	1	0	1	0
cute	1	0	1	0	0
adorable	1	1	1	0	0
gradients	0	0	1	1	1

term-document matrix

A Simple Distributional VSM

	doc1	doc2	doc3	doc4	doc5
cat	1	0	1	0	0
kitten	0	1	0	1	0
cute	1	0	1	0	0
adorable	1	1	1	0	0
gradients	0	0	1	1	1

can be binary indicators, real value counts, tf-idf values, etc.

A Simple Distributional VSM

	doc1	doc2	doc3	doc4	doc5
cat	1	0	1	0	0
kitten	0	1	0	1	0
cute	1	0	1	0	0
adorable	1	1	1	0	0
gradients	0	0	1	1	1

word meaning representation
= set of documents in which it occurs

A Simple Distributional VSM

Computing Similarity

	doc1	doc2	doc3	doc4	doc5
cat	1	0	1	0	0
kitten	0	1	0	1	0
cute	1	0	1	0	0
adorable	1	1	1	0	0
gradients	0	0	1	1	1

do cute and adorable have "similar" meanings?

A Simple Distributional VSM

Computing Similarity

A Simple Distributional VSM

Computing Similarity

- Option 1: Exact equivalence
  - w1 = w2 iff their vector representations are identical

A Simple Distributional VSM

Computing Similarity

	doc1	doc2	doc3	doc4	doc5
cute	1	0	1	0	0
adorable	1	1	1	0	0

A Simple Distributional VSM

Computing Similarity

- ~~Option 1: Exact equivalence~~
  - ~~w1 ≡ w2 iff their vector representations are identical~~

Real-world language is too variable! This would never work.

A Simple Distributional VSM

Computing Similarity

- ~~Option 1: Exact equivalence~~
  - ~~w1 = w2 iff their vector representations are identical~~
- Option 2: Jaccard Similarity
  - sim(w1,w2) = intersection(v1,v2)/union(v1,v2)

A Simple Distributional VSM

Computing Similarity

	doc1	doc2	doc3	doc4	doc5
cute	1	0	1	0	0
adorable	1	1	1	0	0

$$\text{Jaccard} = \frac{|\{doc1, doc3\}|}{|\{doc1, doc2, doc3\}|}$$

A Simple Distributional VSM

Computing Similarity

	doc1	doc2	doc3	doc4	doc5
cute	1	0	1	0	0
adorable	1	1	1	0	0

$$\text{Jaccard} = \frac{|\{\text{doc1, doc3}\}|}{|\{\text{doc1, doc2, doc3}\}|} = 0.67$$

A Simple Distributional VSM

Computing Similarity

- ~~Option 1: Exact equivalence~~
  - ~~w1 = w2 iff their vector representations are identical~~
- Option 2: Jaccard Similarity
  - sim(w1,w2) = intersection(v1,v2)/union(v1,v2)

Works well for binary vectors, but needs adjustments for real-valued dimensions

A Simple Distributional VSM

Computing Similarity

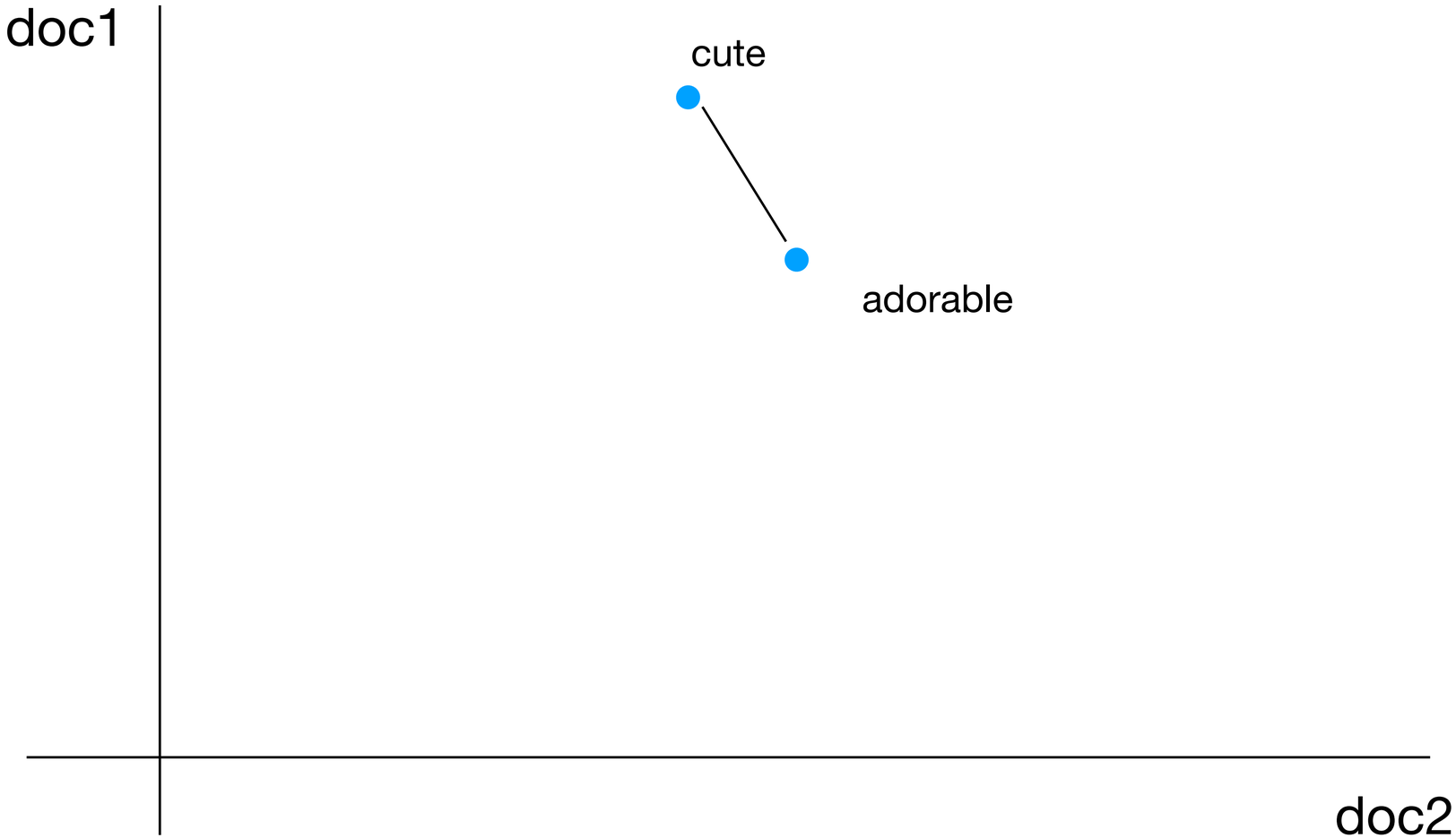
- ~~Option 1: Exact equivalence~~
  - ~~w1 = w2 iff their vector representations are identical~~
- Option 2: Jaccard Similarity
  - sim(w1,w2) = intersection(v1,v2)/union(v1,v2)
- Option 3: Euclidean Distance

A Simple Distributional VSM

Computing Similarity

A Simple Distributional VSM

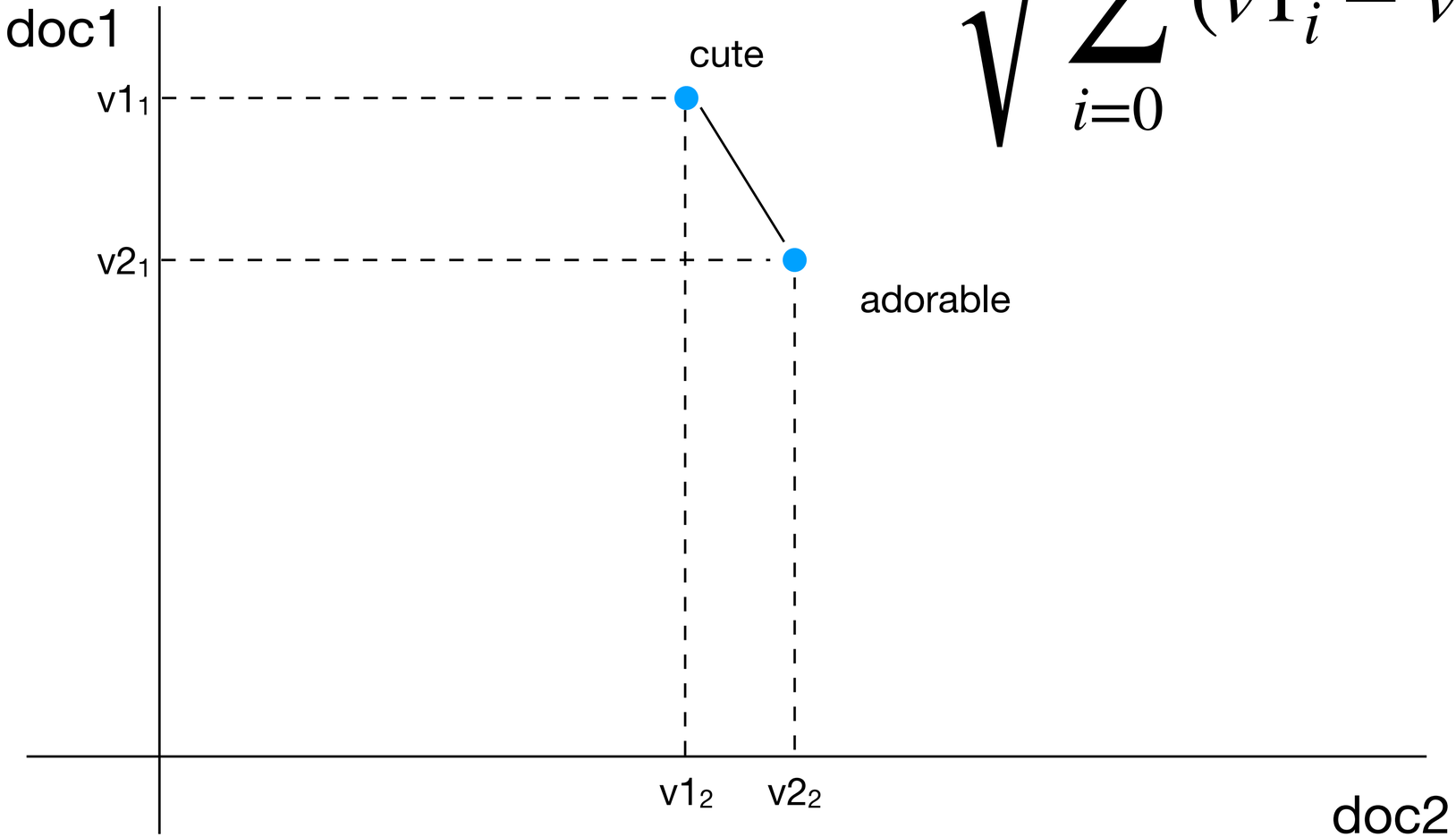
Computing Similarity



A Simple Distributional VSM

Computing Similarity

$$\sqrt{\sum_{i=0}^{n}(v1_i - v2_i)^2}$$



A Simple Distributional VSM

Computing Similarity

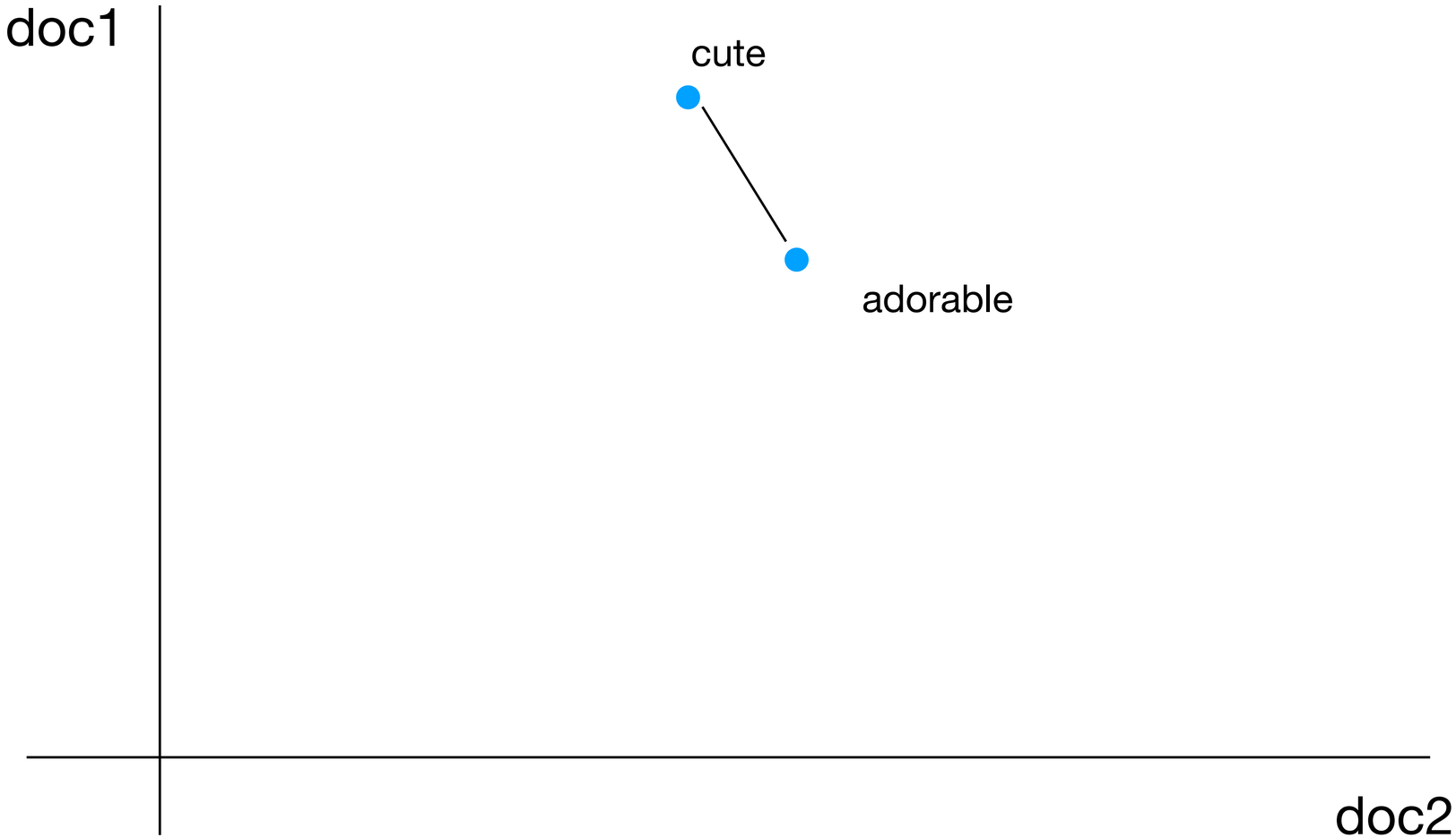
- ~~Option 1: Exact equivalence~~
  - ~~w1 = w2 iff their vector representations are identical~~
- Option 2: Jaccard Similarity
  - sim(w1,w2) = intersection(v1,v2)/union(v1,v2)
- Option 3: Euclidean Distance

$$\sqrt{\sum_{i=0}^{n}(v1_i - v2_i)^2}$$

Popular in a lot of applications. But, assumes similar words will be of similar magnitude (i.e., occur with similar frequency)

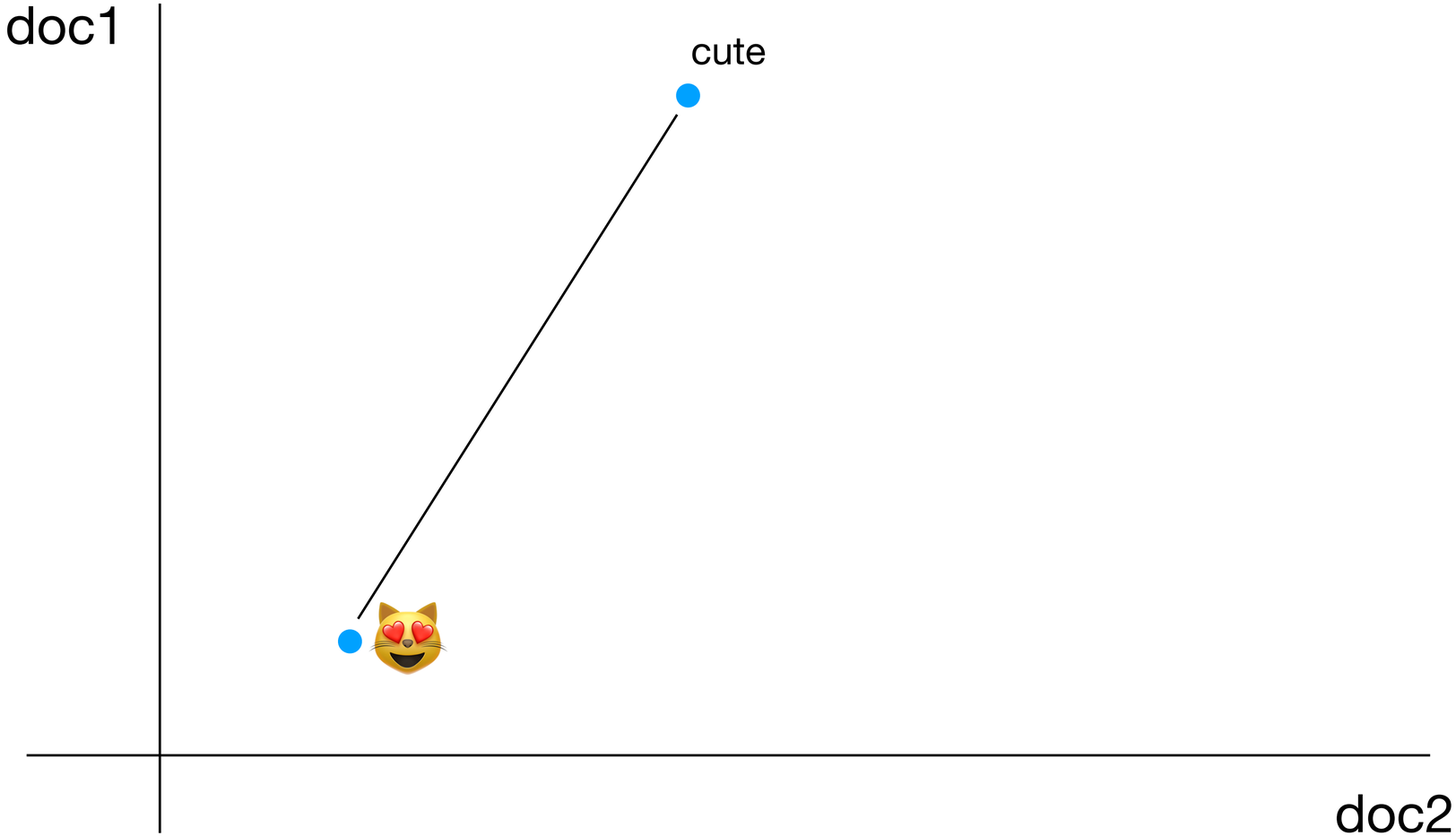
A Simple Distributional VSM

Computing Similarity



A Simple Distributional VSM

Computing Similarity



A Simple Distributional VSM

Computing Similarity

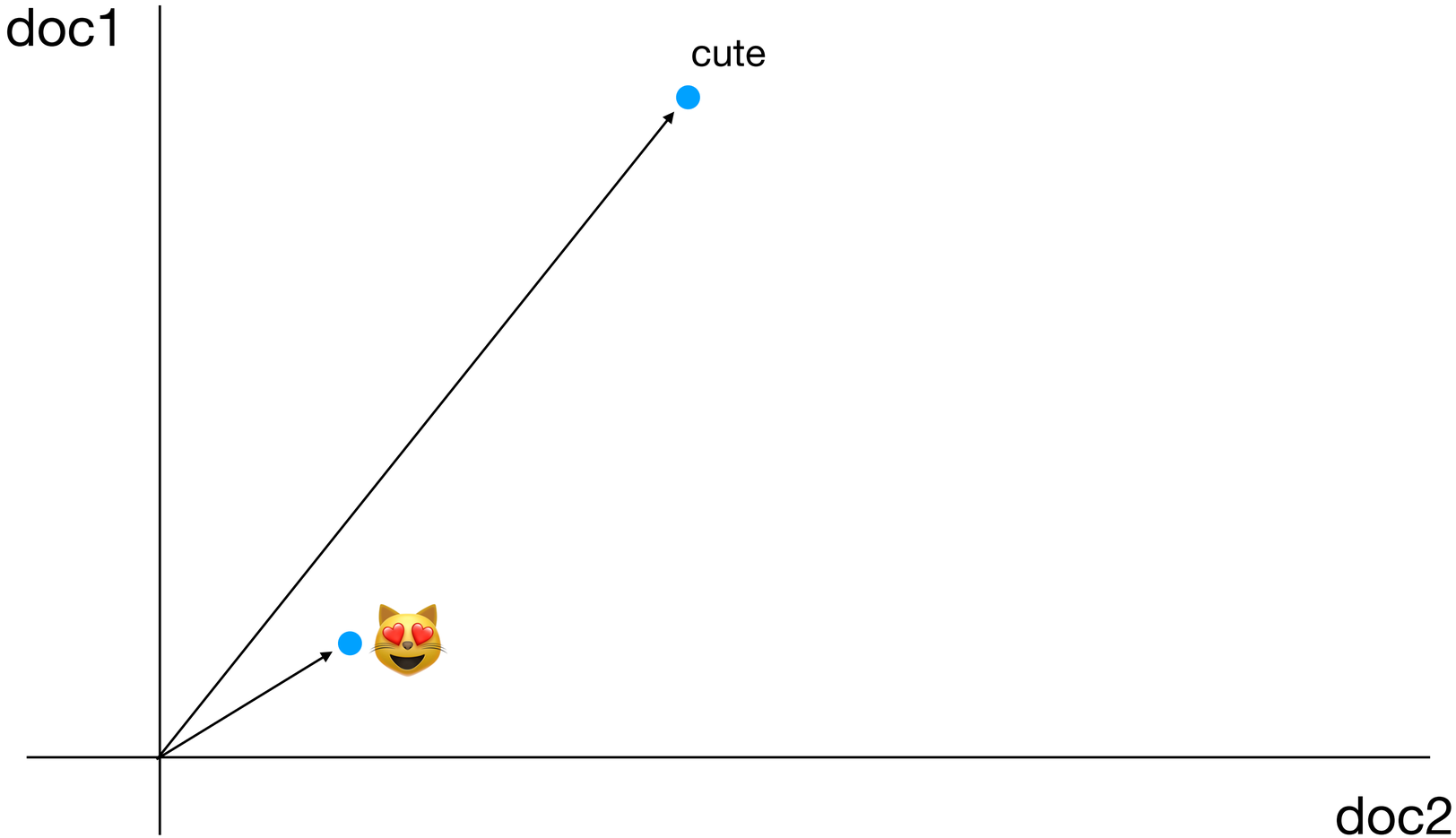
- ~~Option 1: Exact equivalence~~
  - ~~w1 = w2 iff their vector representations are identical~~
- Option 2: Jaccard Similarity
  - sim(w1,w2) = intersection(v1,v2)/union(v1,v2)
- Option 3: Euclidean Distance

$$\sqrt{\sum_{i=0}^{n}(v1_i - v2_i)^2}$$

- Option 4: Cosine Similarity

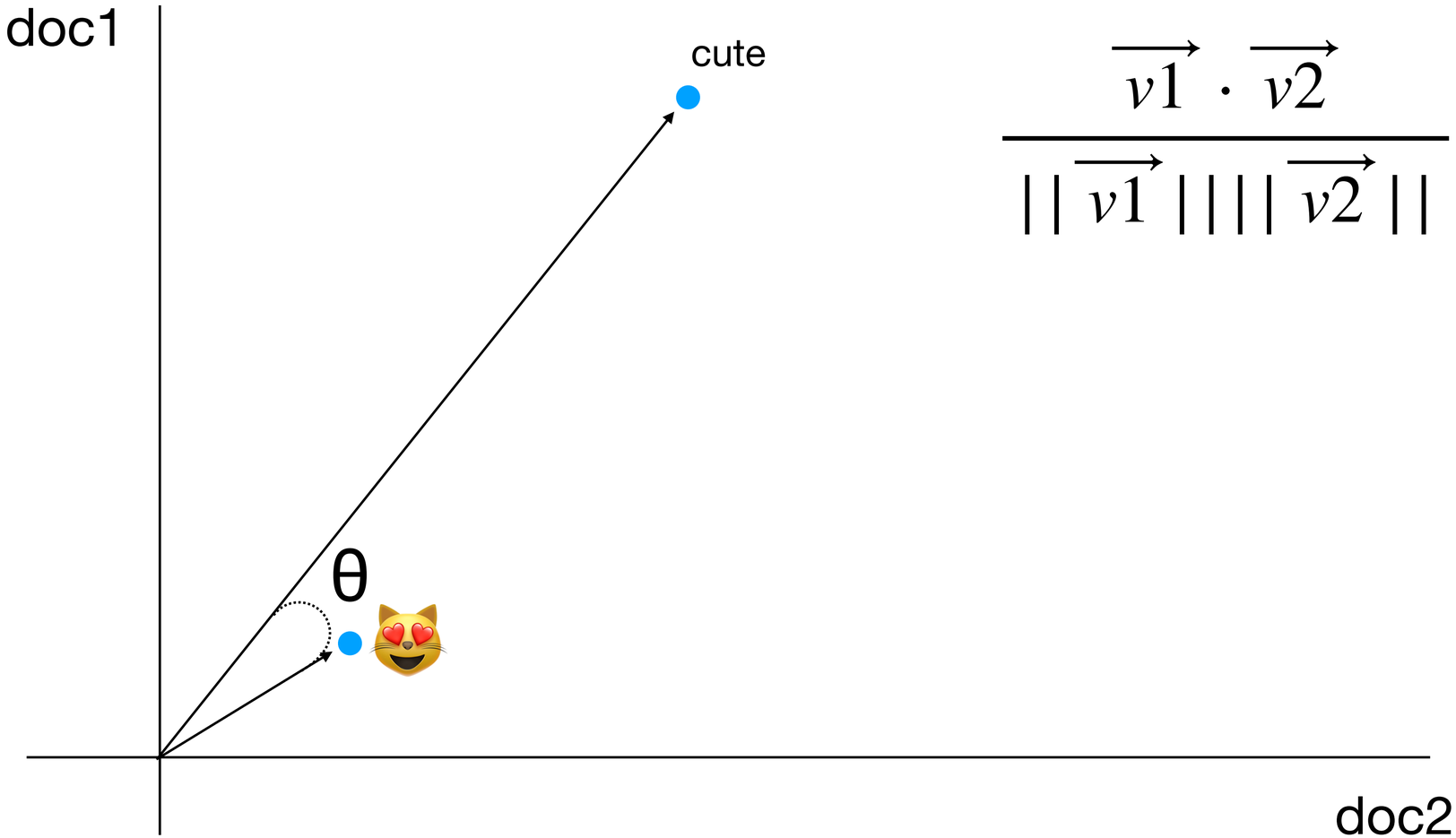
A Simple Distributional VSM

Computing Similarity



A Simple Distributional VSM

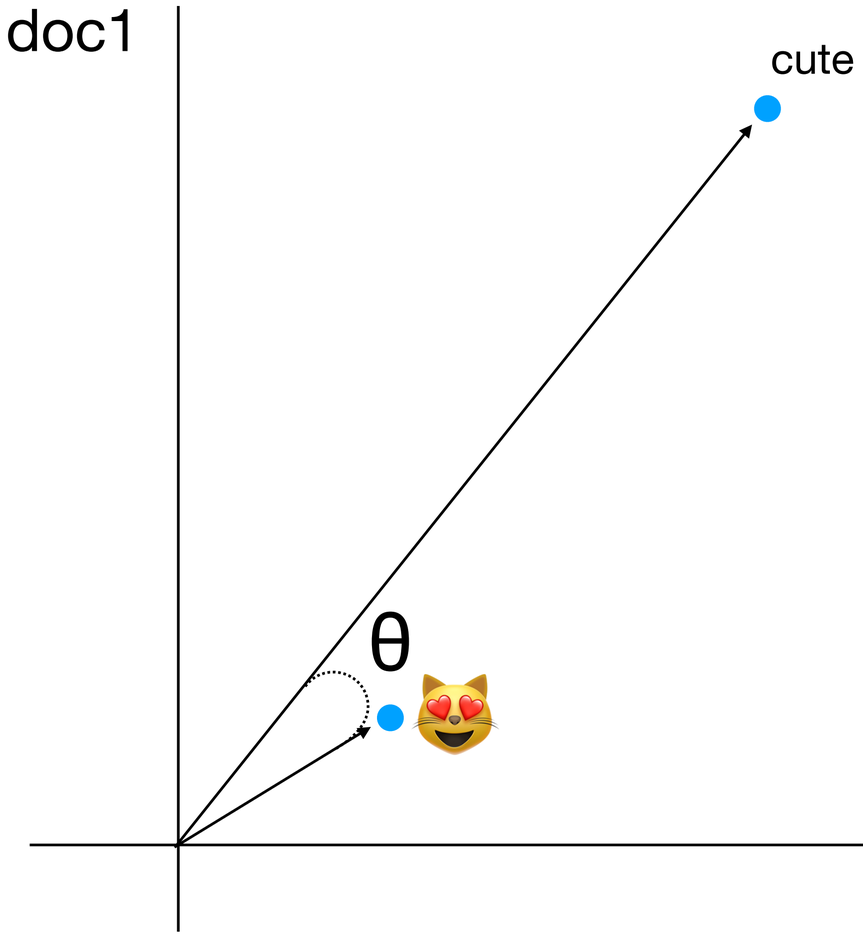
Computing Similarity



$$\frac{\overrightarrow{v1} \cdot \overrightarrow{v2}}{||\overrightarrow{v1}|| \, ||\overrightarrow{v2}||}$$

A Simple Distributional VSM

Computing Similarity



$$\frac{\overrightarrow{v1} \cdot \overrightarrow{v2}}{||\overrightarrow{v1}|| ||\overrightarrow{v2}||}$$

$$\frac{\sum_{i=0}^{n} v1_i v2_i}{\sqrt{\sum_{i=0}^{n} v1_i^2}\sqrt{\sum_{i=0}^{n} v2_i^2}}$$

A Simple Distributional VSM

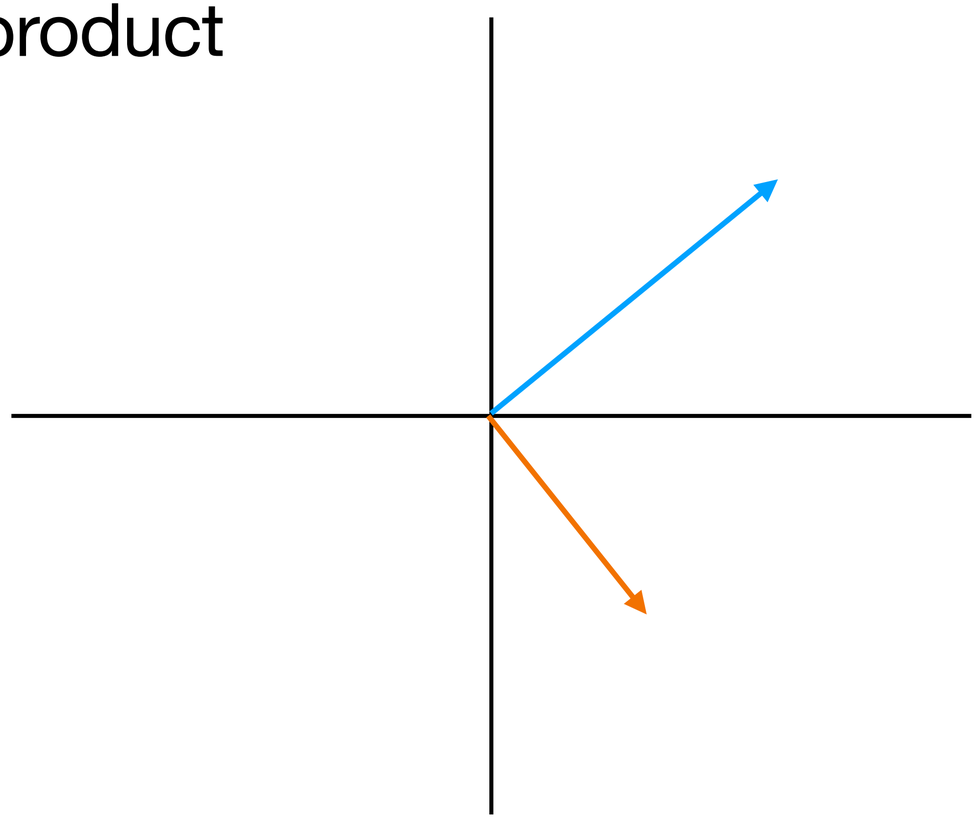
Detour: Dot Product

- dot product = scalar product = inner product
- $a \cdot b = \sum_{i=1}^{n} a_i b_i$
- $a \cdot b = \|a\| \|b\| cos\theta$

A Simple Distributional VSM

Detour: Dot Product

- dot product = scalar product = inner product
- $a \cdot b = \sum_{i=1}^{n} a_i b_i$
- $a \cdot b = \|a\| \|b\| cos\theta$

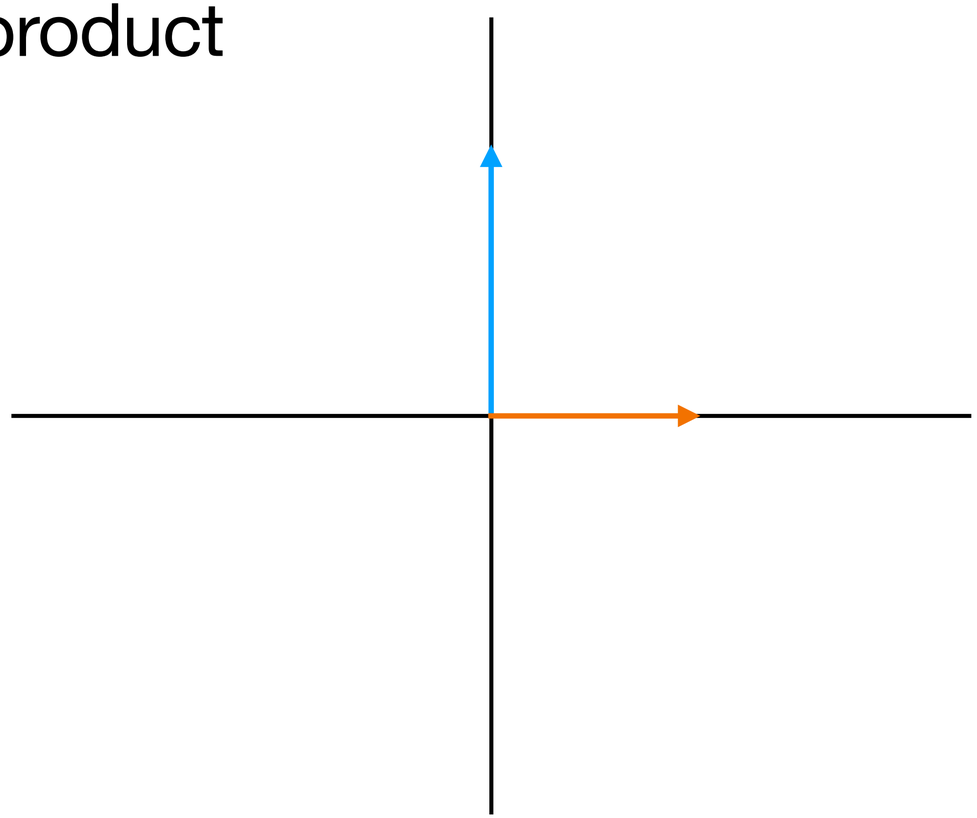


= 0 if vectors are orthogonal

A Simple Distributional VSM

Detour: Dot Product

- dot product = scalar product = inner product
- $a \cdot b = \sum_{i=1}^{n} a_i b_i$
- $a \cdot b = \|a\| \|b\| cos\theta$

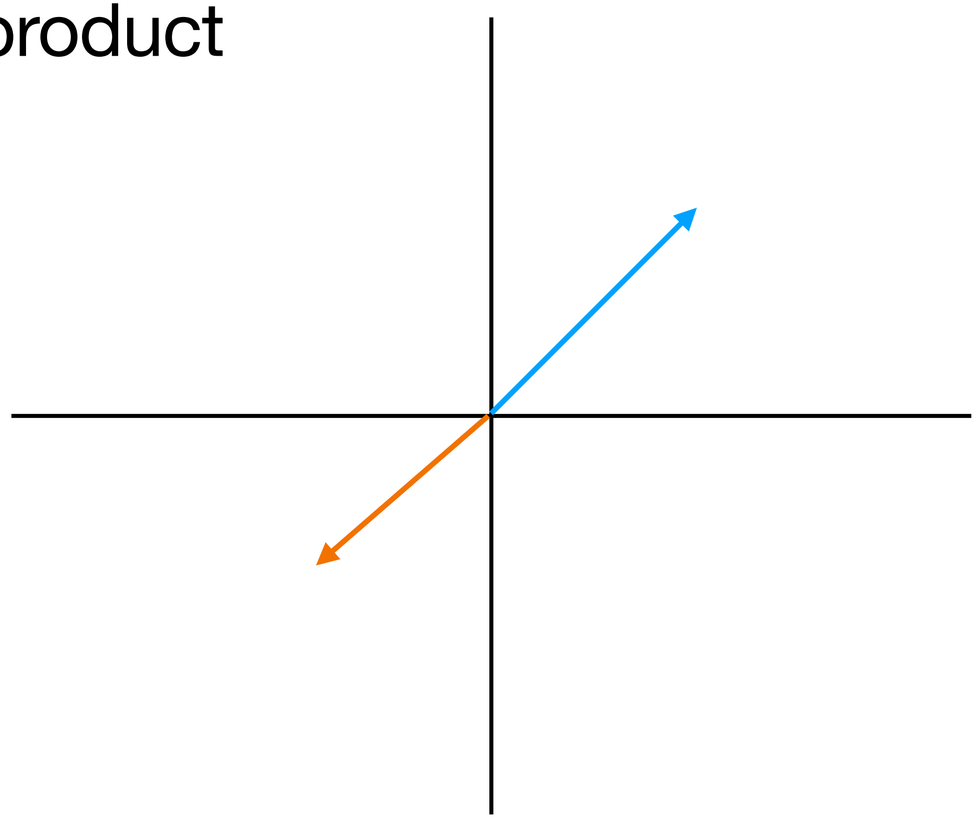


= 0 if vectors are orthogonal

A Simple Distributional VSM

Detour: Dot Product

- dot product = scalar product = inner product
- $a \cdot b = \sum_{i=1}^{n} a_i b_i$
- $a \cdot b = \|a\| \|b\| cos\theta$

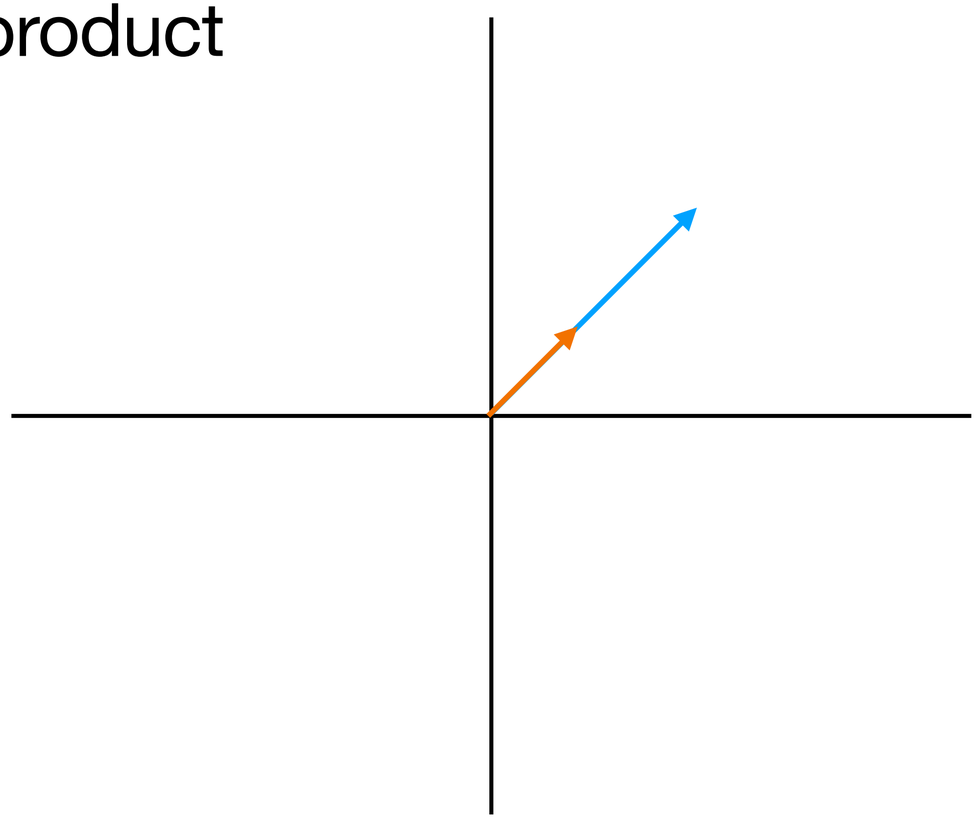


= -1 if vectors point in opposite directions

A Simple Distributional VSM

Detour: Dot Product

- dot product = scalar product = inner product
- $a \cdot b = \sum_{i=1}^{n} a_i b_i$
- $a \cdot b = \|a\| \|b\| cos\theta$

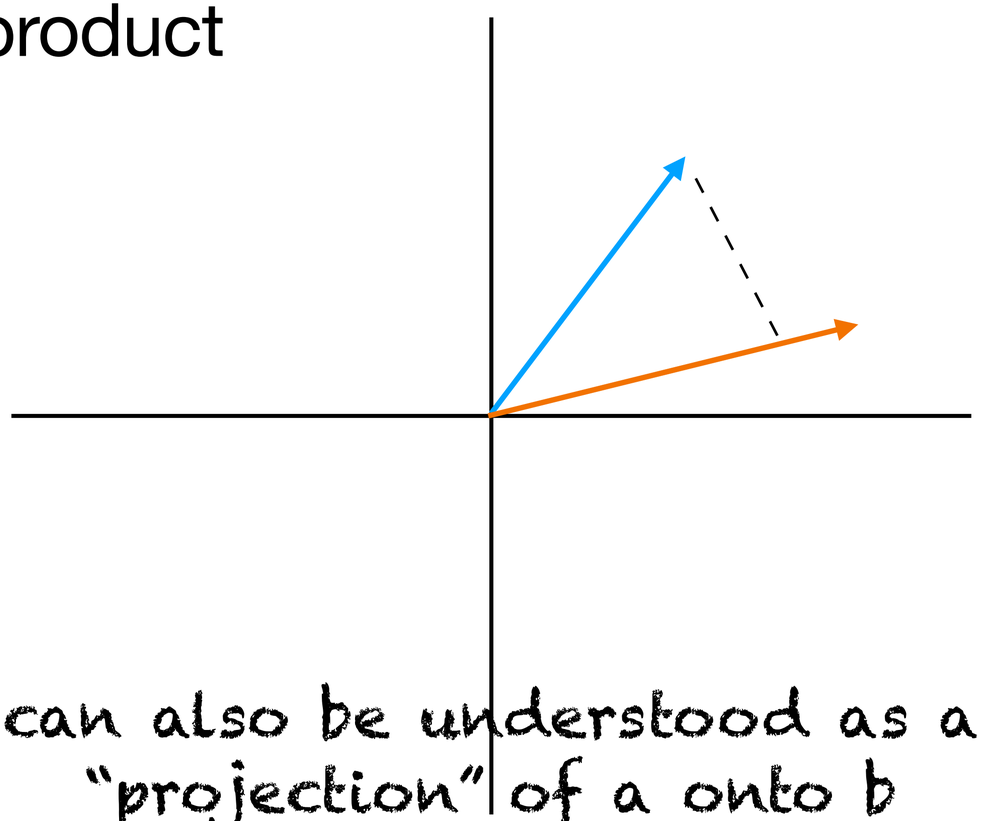


= 1 if vectors are parallel

A Simple Distributional VSM

Detour: Dot Product

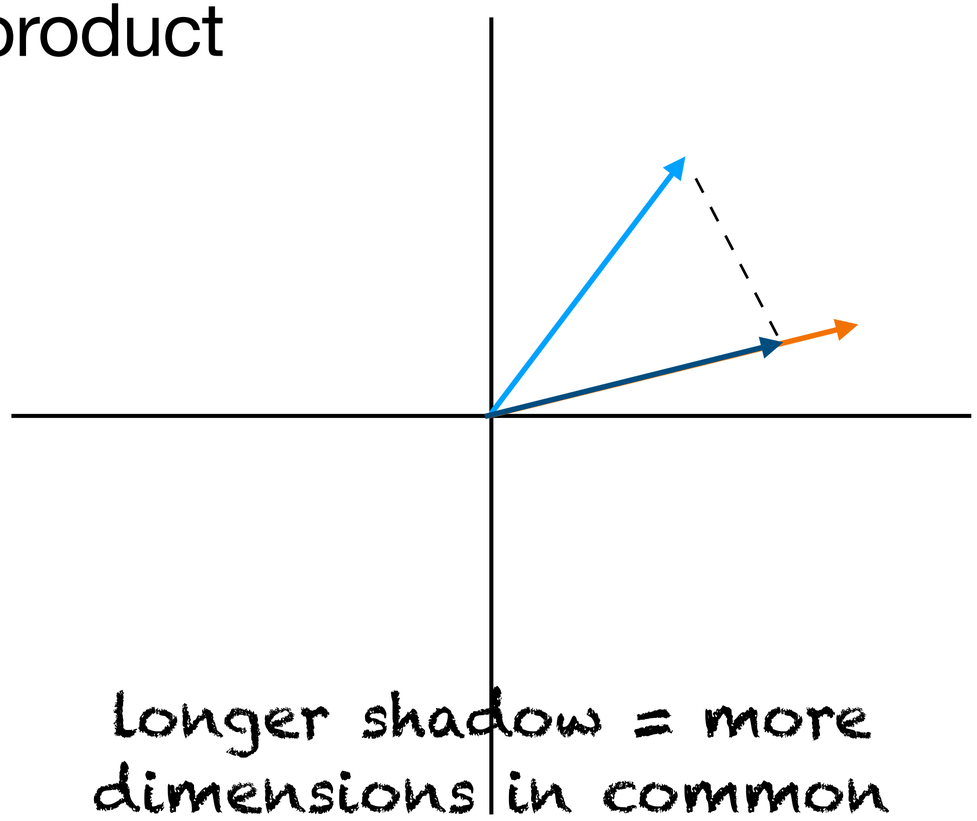
- dot product = scalar product = inner product
- $a \cdot b = \sum_{i=1}^{n} a_i b_i$
- $a \cdot b = \|a\| \|b\| cos\theta$



A Simple Distributional VSM

Detour: Dot Product

- dot product = scalar product = inner product
- $a \cdot b = \sum_{i=1}^{n} a_i b_i$
- $a \cdot b = \|a\| \|b\| cos\theta$



A Simple Distributional VSM

Computing Similarity

- ~~Option 1: Exact equivalence~~
  - ~~w1 ≡ w2 iff their vector representations are identical~~
- Option 2: Jaccard Similarity
  - sim(w1,w2) = intersection(v1,v2)/union(v1,v2)
- Option 3: Euclidean Distance

$$\sqrt{\sum_{i=0}^{n}(v1_i - v2_i)^2}$$

- Option 4: Cosine Similarity

$$\frac{\overrightarrow{v1} \cdot \overrightarrow{v2}}{||\overrightarrow{v1}||||\overrightarrow{v2}||}$$

By far the most popular/ standard way to compute word similarity

A Simple Distributional VSM

Practice Question!

Assume I have a term-document matrix built using New York Times articles. Which of the following pairs do you think will be most similar according to cosine similarity?

A. w1 = gas, w2 = economy

B. w1 = gas, w2 = petrol

C. w1 = gas, w2 = fuel

A Simple Distributional VSM

Practice Question!

Assume I have a term-document matrix built using New York Times articles. Which of the following pairs do you think will be most similar according to cosine similarity?

A. w1 = gas, w2 = economy

B. w1 = gas, w2 = petrol

C. w1 = gas, w2 = fuel

Term-Document Matrices will capture broad topical-similarity and co-occurrence. Not words with the "same meaning".

Word-Context Matrix

The domestic cat is a small, typically furry, carnivorous mammal.

Your cat's online owners manual, featuring articles about breed information, cat selection, training, grooming and care for cats and kittens.

Wish you had a secret decoder guide to cat behavior and cat language? Here's a primer to things your cat wishes you understood.

"The cat does not offer services," William Burroughs wrote. "The cat offers itself." But it does so with unapologetic ambivalence.

Welcome to the new WebMD Cat Health Center. WebMD veterinary experts provide comprehensive information about cat health care, offer nutrition and feeding …

Yes, they're independent and willful, but felines can be taught certain behaviors—to the benefit of both cat and human.

Word-Context Matrix

The domestic **cat** is a small, typically furry, carnivorous mammal.

Your **cat's** online owners manual, featuring articles about breed information, **cat** selection, training, grooming and care for **cats** and kittens.

Wish you had a secret decoder guide to **cat** behavior and **cat** language? Here's a primer to things your **cat** wishes you understood.

"The **cat** does not offer services," William Burroughs wrote. "The **cat** offers itself." But it does so with unapologetic ambivalence.

Welcome to the new WebMD **Cat** Health Center. WebMD veterinary experts provide comprehensive information about **cat** health care, offer nutrition and feeding …

Yes, they're independent and willful, but felines can be taught certain behaviors—to the benefit of both **cat** and human.

Word-Context Matrix

The domestic **cat** is a small, typically furry, carnivorous mammal.

Your **cat's** online owners manual, featuring articles about breed information, **cat** selection, training, grooming and care for **cats** and kittens.

Wish you had a secret decoder guide to **cat** behavior and **cat** language? Here's a primer to things your **cat** wishes you understood.

"The **cat** does not offer services," William Burroughs wrote. "The **cat** offers itself." But it does so with unapologetic ambivalence.

Welcome to the new WebMD **Cat** Health Center. WebMD veterinary experts provide comprehensive information about **cat** health care, offer nutrition and feeding …

Yes, they're independent and willful, but felines can be taught certain behaviors—to the benefit of both **cat** and human.

Word-Context Matrix

The domestic **cat** is a small, typically furry, carnivorous mammal.

Your **cat's** online owners manual, featuring articles about breed information, **cat** selection, training, grooming and care for **cats** and kittens.

Wish you had a secret decoder guide to **cat** behavior and **cat** language? Here's a primer to things your **cat** wishes you understood.

"The **cat** does not offer services," William Burroughs wrote. "The **cat** offers itself." But it does so with unapologetic ambivalence.

Welcome to the new WebMD **Cat** Health Center. WebMD veterinary experts provide comprehensive information about **cat** health care, offer nutrition and feeding …

Yes, they're independent and willful, but felines can be taught certain behaviors—to the benefit of both **cat** and human.

	the	domes-tic	is	a	your	online	owners	breed	informa-tion	selec-tion
cat	1000	40	500	700	400	3	80	100	15	6
dog	1050	50	400	950	500	1	105	160	4	2

Word-Context Matrix

	cat	kitten	cute	adorable	gradients
cat	0	0	1	1	0
kitten	0	0	1	1	0
cute	1	1	0	0	0
adorable	1	1	1	0	0
gradients	0	0	1	1	1

can be binary indicators, real value counts, tf-idf values, etc.

Word-Context Matrix

	cat	kitten	cute	adorable	gradients
cat	0	0	1	1	0
kitten	0	0	1	1	0
cute	1	1	0	0	0
adorable	1	1	1	0	0
gradients	0	0	1	1	1

similar words don't necessarily co-occur

Word-Context Matrix

	cat	kitten	cute	adorable	gradients
cat	0	0	1	1	0
kitten	0	0	1	1	0
cute	1	1	0	0	0
adorable	1	1	1	0	0
gradients	0	0	1	1	1

but they occur in similar contexts (i.e., can appear with the same types of words)

Term-Document vs. Word-Context

- Term Document:
  - More broad, topical similarity
  - Good for document classification tasks, retrieval
- Word-Context
  - More grammatical similarity (nouns vs. verbs)
  - More lexical similarity (synonyms)
- There are lots of other ways to define context!! Columns could include…
  - Syntactic context (e.g., using a dependency parse?)
  - Image features?
  - Audio/acoustic features?
  - Whatever else you can dream up!

VSMs vs. DSMs

- DSM = Distributional Semantics Model
  - Word meaning defined by its the contexts in which it is used
- VSM = Vector Space Models of Semantics
  - Represent words as points in space
- Most DSMs are VSMs, most VSMs are DSMs: but this doesn't have to be the case!
  - "Conceptual Spaces" by Peter Gardenfors: VSM that is not a DSM
    - dimensions of vector correspond to features such as "weight" or "temperature"
- In NLP, we use vector space DSMs
  - I.e., words are vectors, and the dimensions of those vectors reflect the contexts in which the word is used

Topics

- Theories of Word Meaning
- Vector Space Models (VSMs)
- **Word Embeddings, Part 1**

Traditional Word Vectors vs. Word Embeddings

- Word Vectors: sparse, very high-dimensional
- Word Embeddings: dense, low dimensional, dimensions are not directly interpretable
  - Still VSMs
  - Still DSMs

Word Embeddings

How do we get them

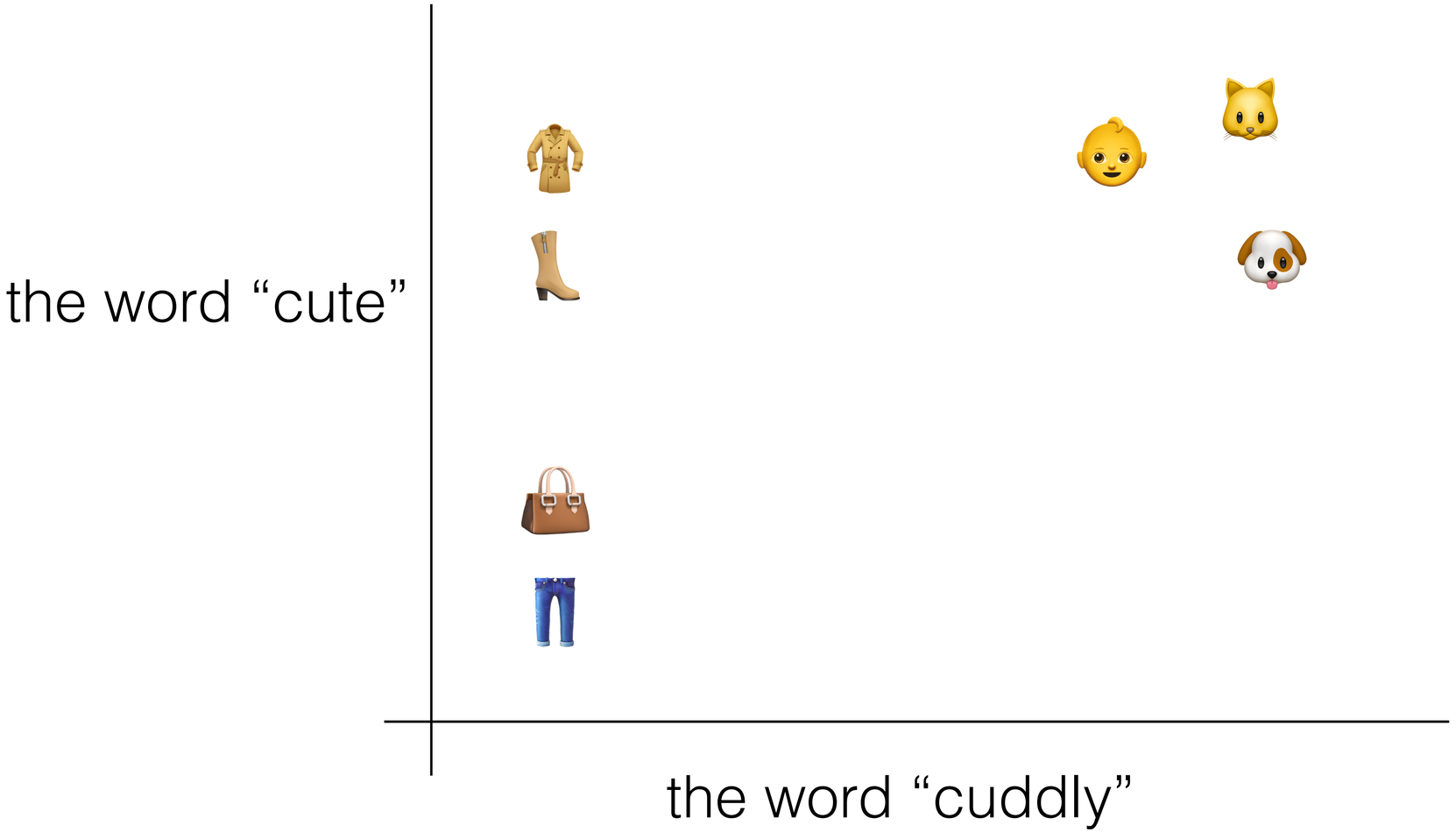
- Dimensionality Reduction
- Two main methods:
  - Matrix Factorization (today)
  - Neural Networks (later)

Word Embeddings

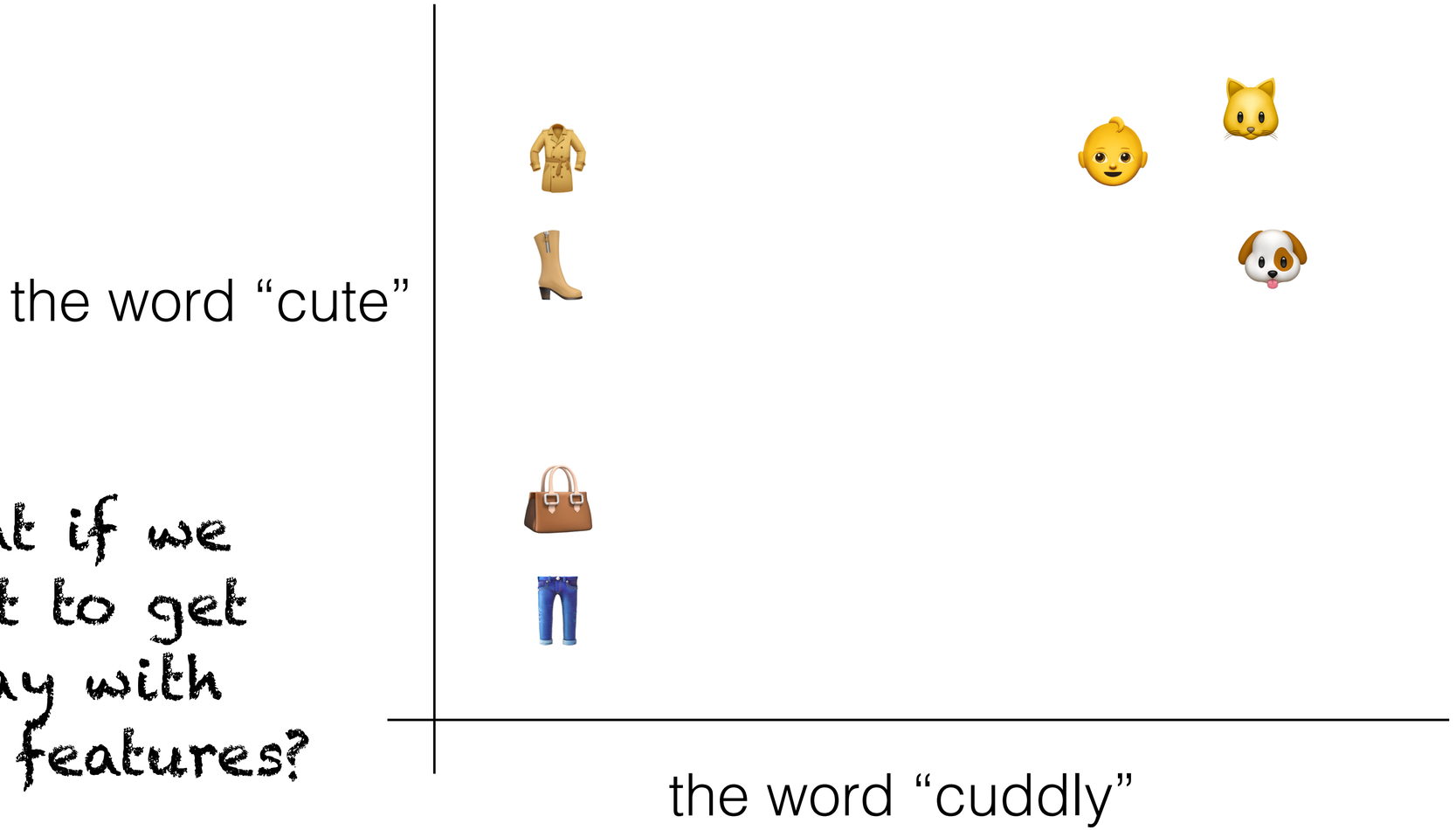
Dimensionality Reduction

- Linear algebraic method for transforming the feature matrix
- Goal is to represent each data point in a **new feature space**
- The new feature space is ideally:
  - **More informative for machine learning**, since it consolidates redundant feature and makes patterns clearer
  - Often **less interpretable to humans**, since the columns of feature matrix no longer refer to features that we have defined

Principle Component Analysis (PCA)

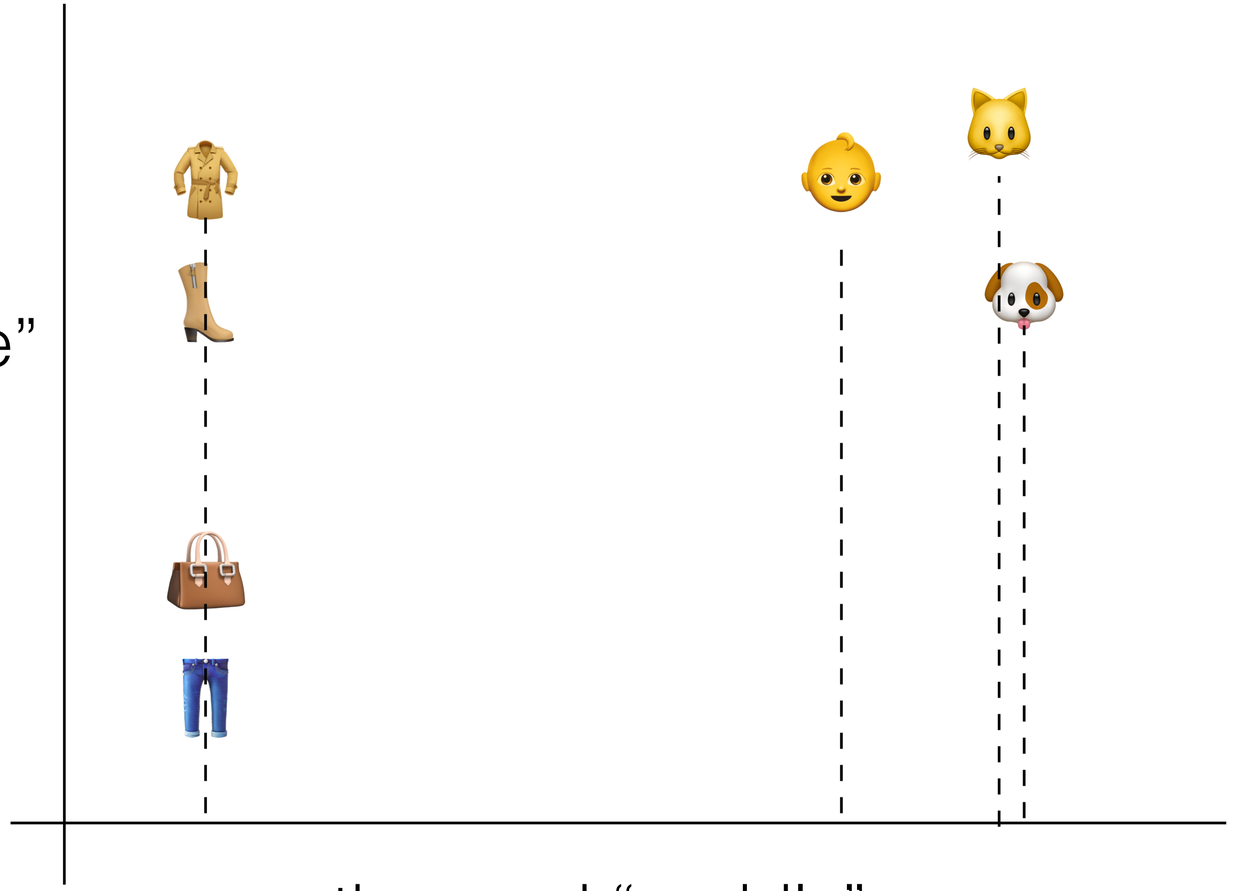


Principle Component Analysis (PCA)



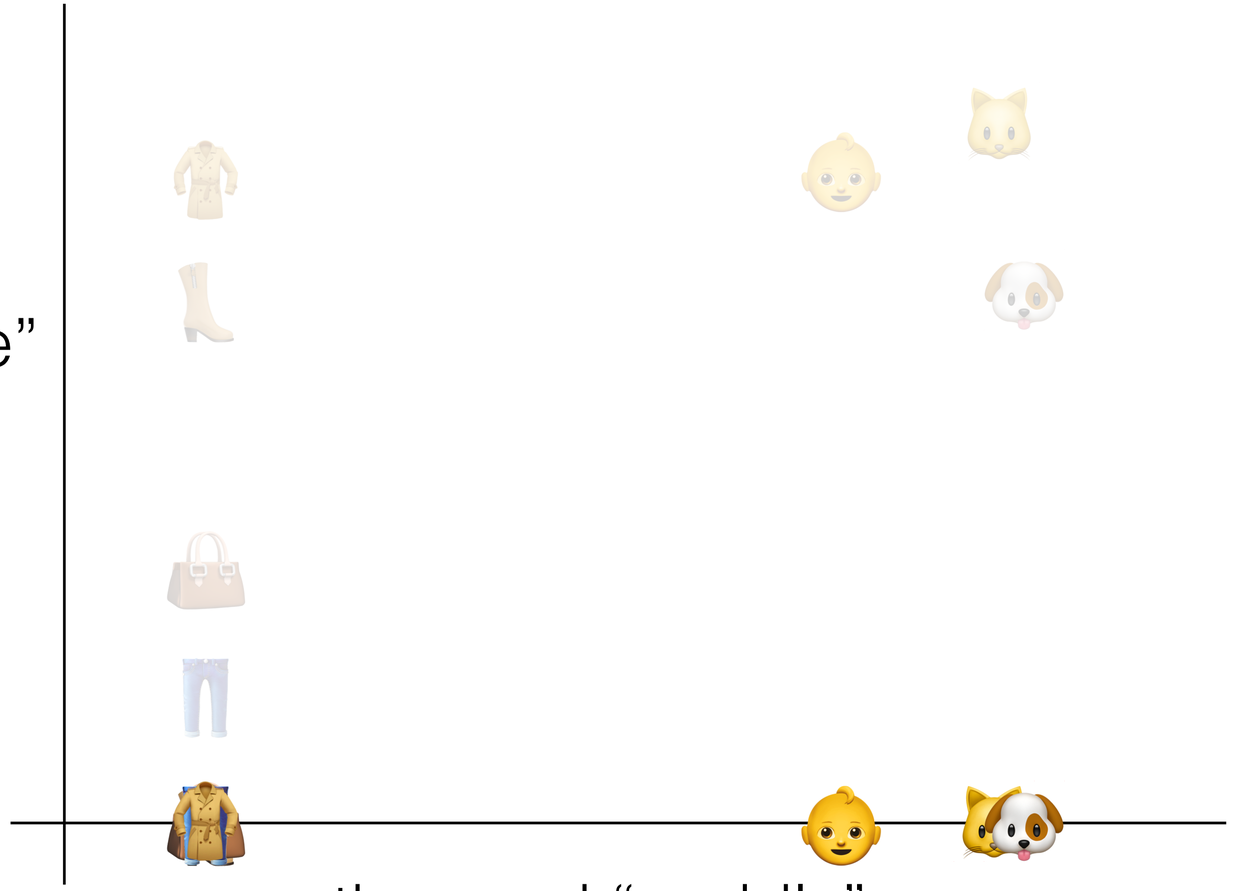
What if we want to get away with fewer features?

Principle Component Analysis (PCA)



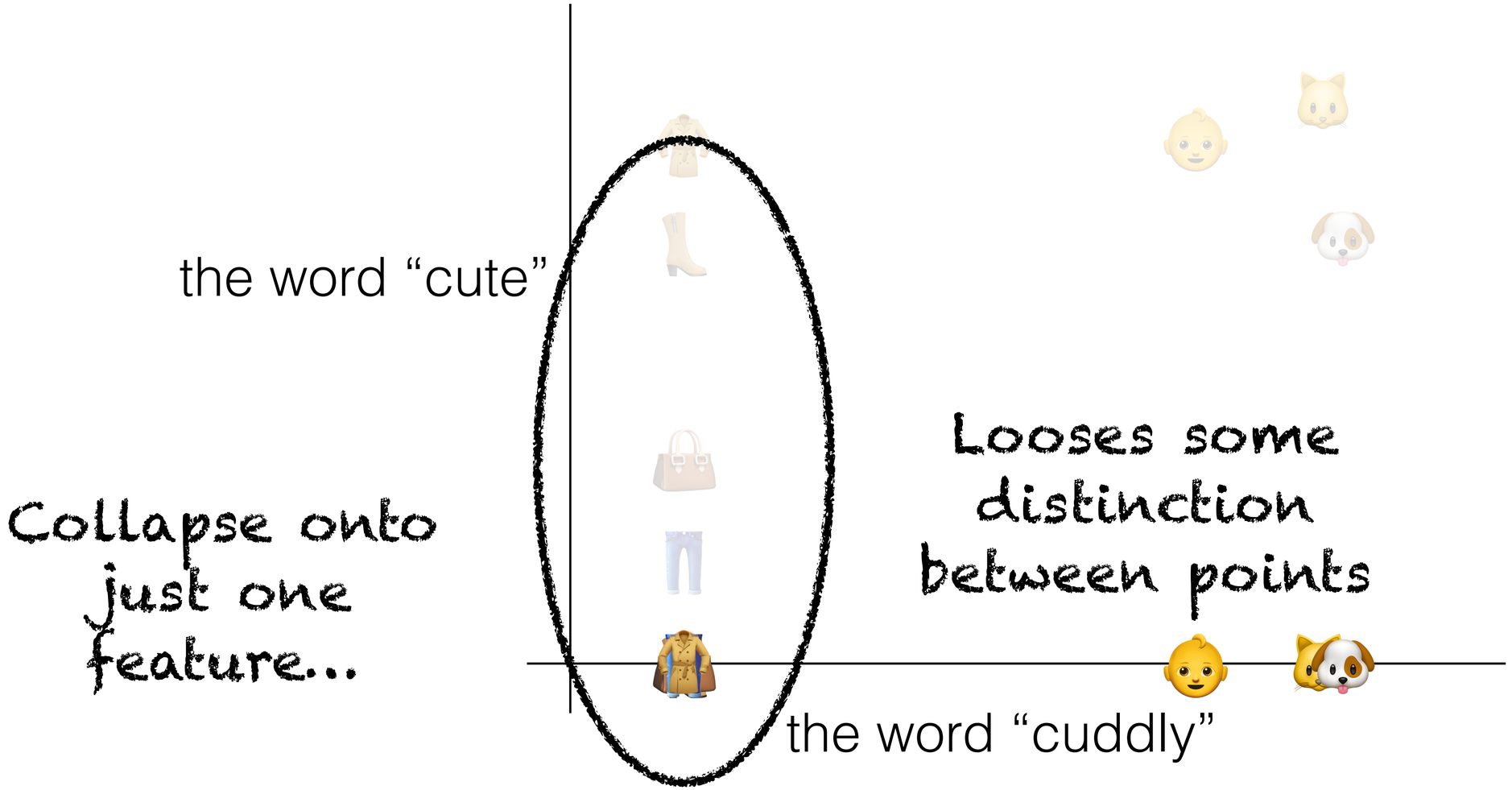
Collapse onto just one feature...

Principle Component Analysis (PCA)

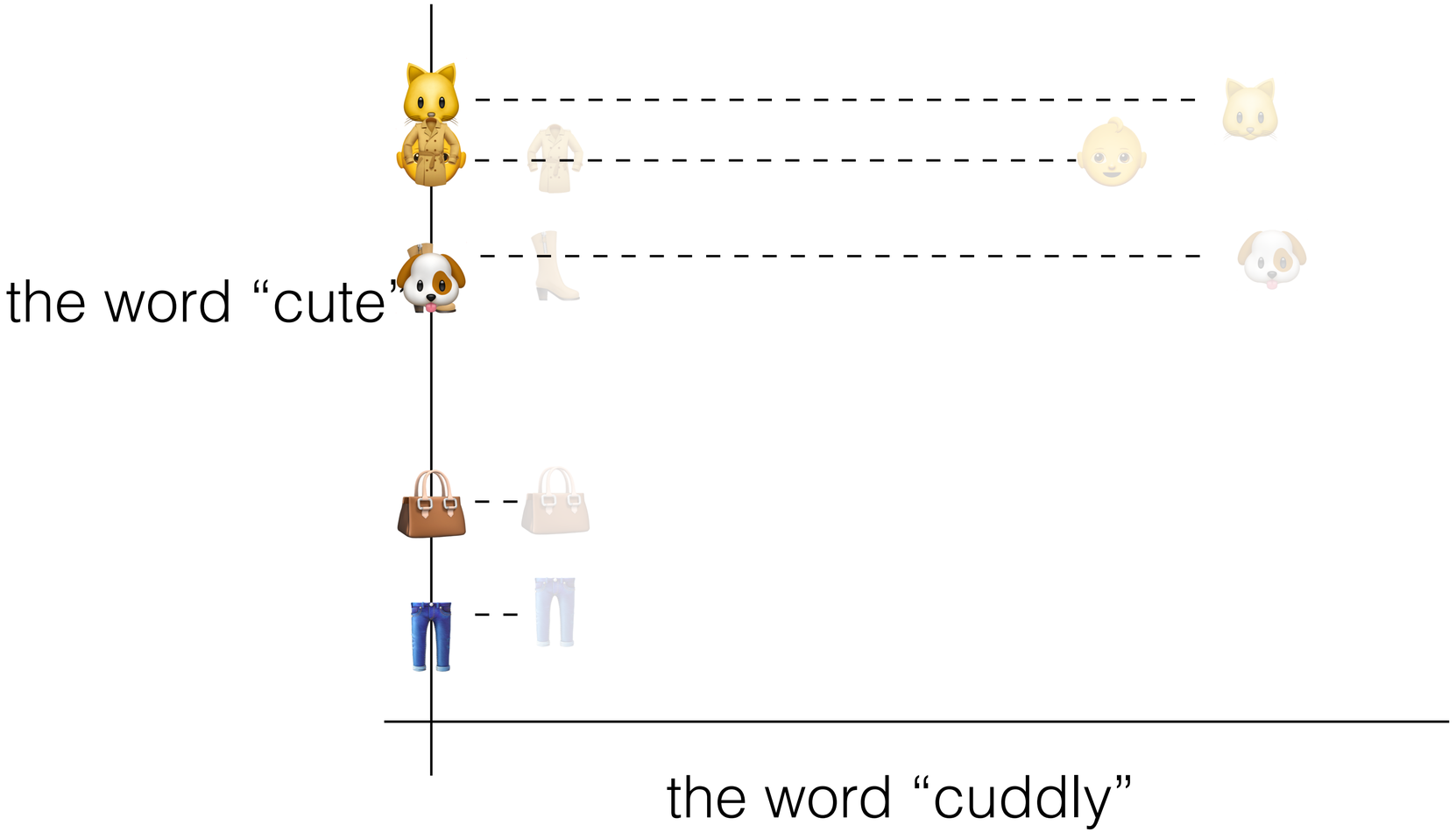


Collapse onto just one feature...

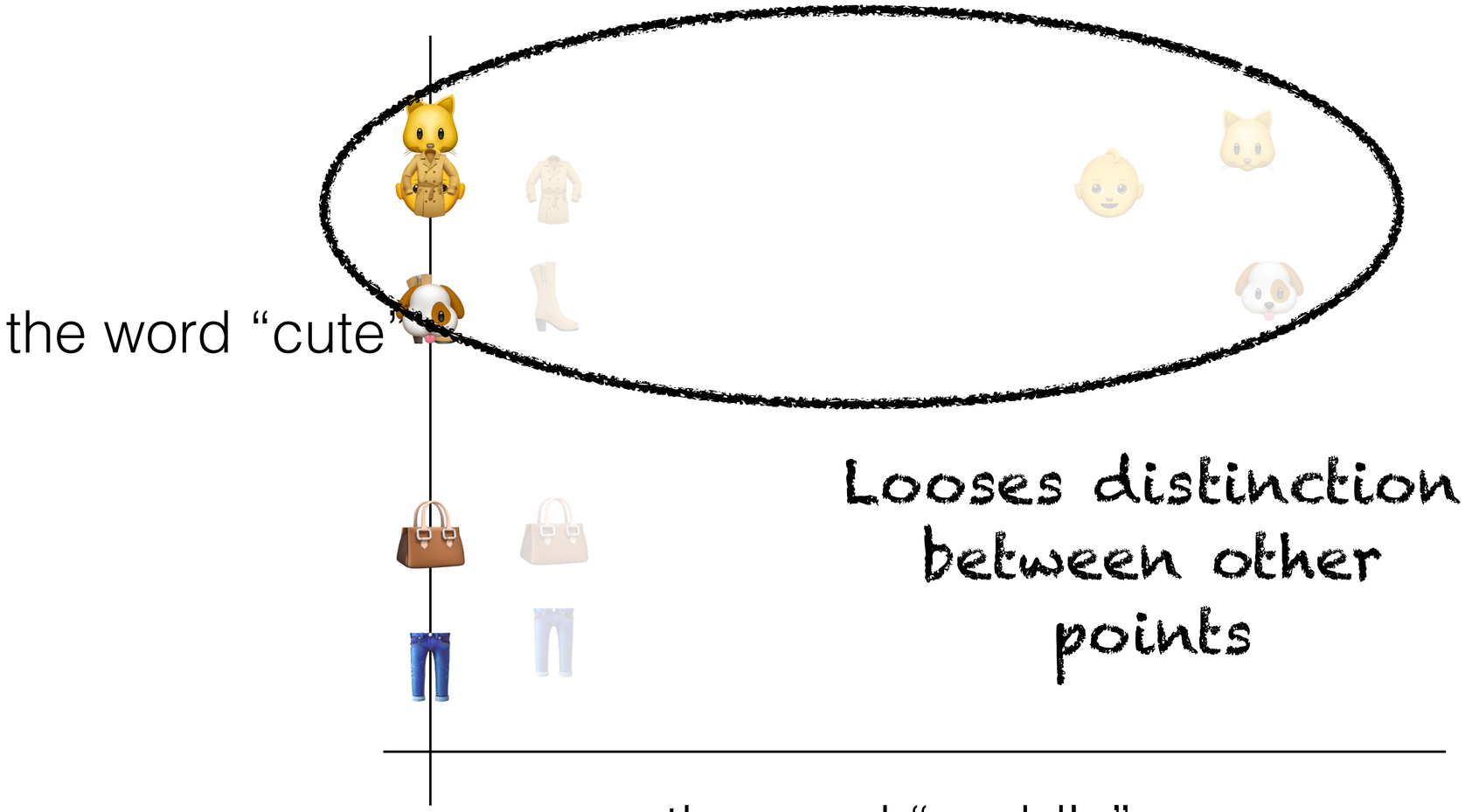
Principle Component Analysis (PCA)



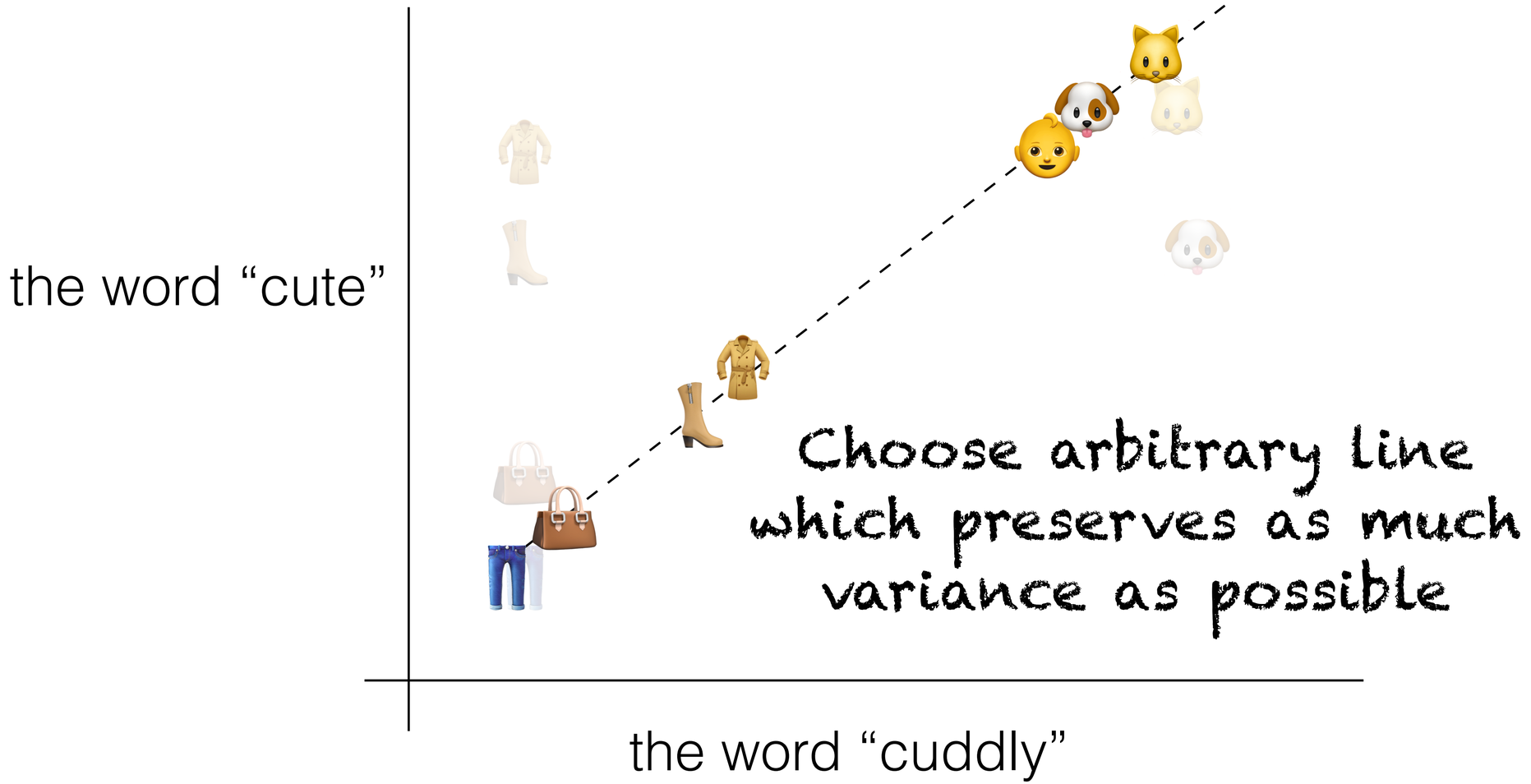
Principle Component Analysis (PCA)



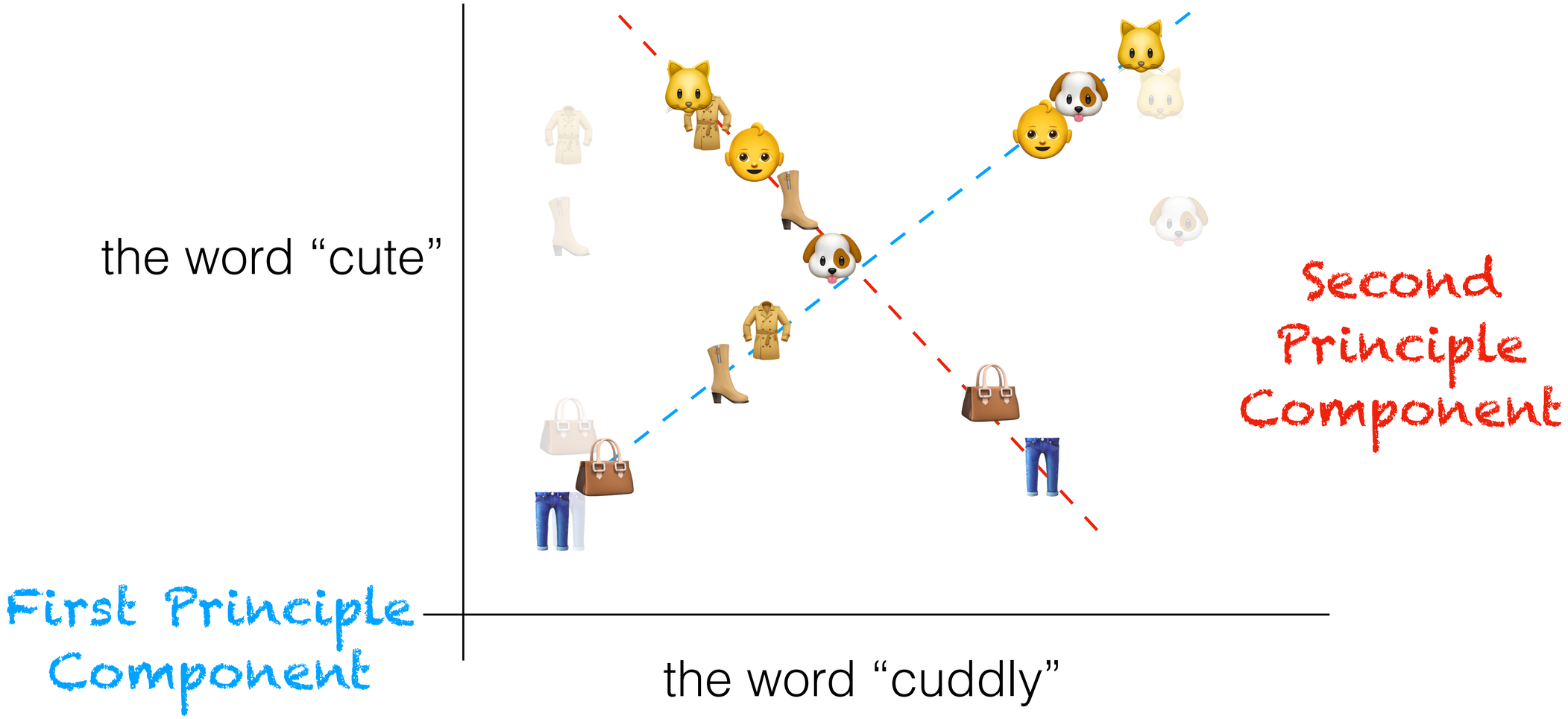
Principle Component Analysis (PCA)



Principle Component Analysis (PCA)



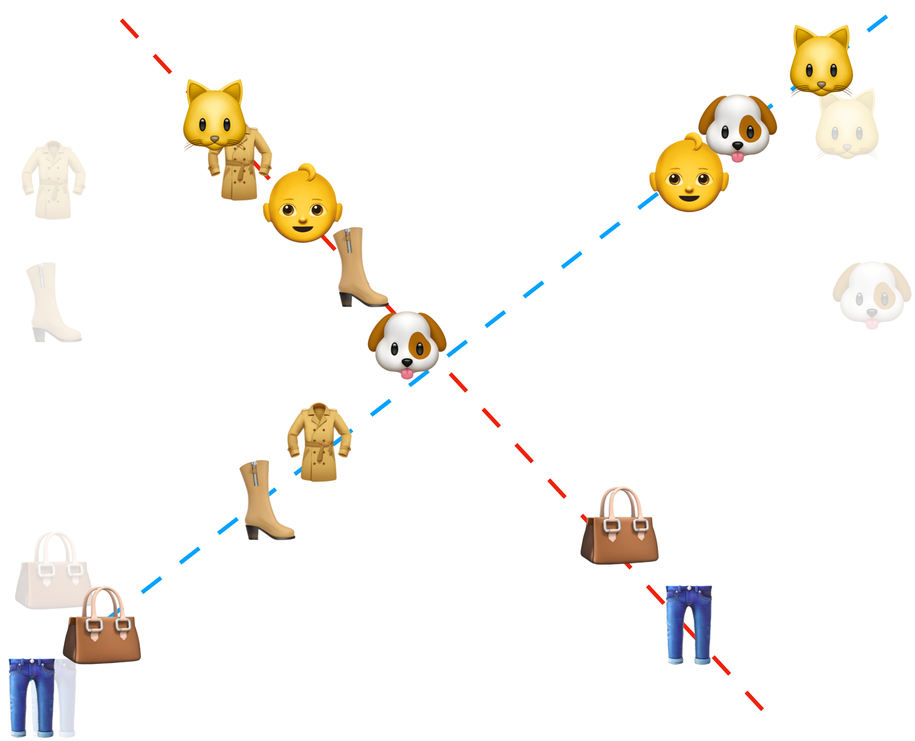
Principle Component Analysis (PCA)



Principle Component Analysis (PCA)

"change of basis"

the word "cute"

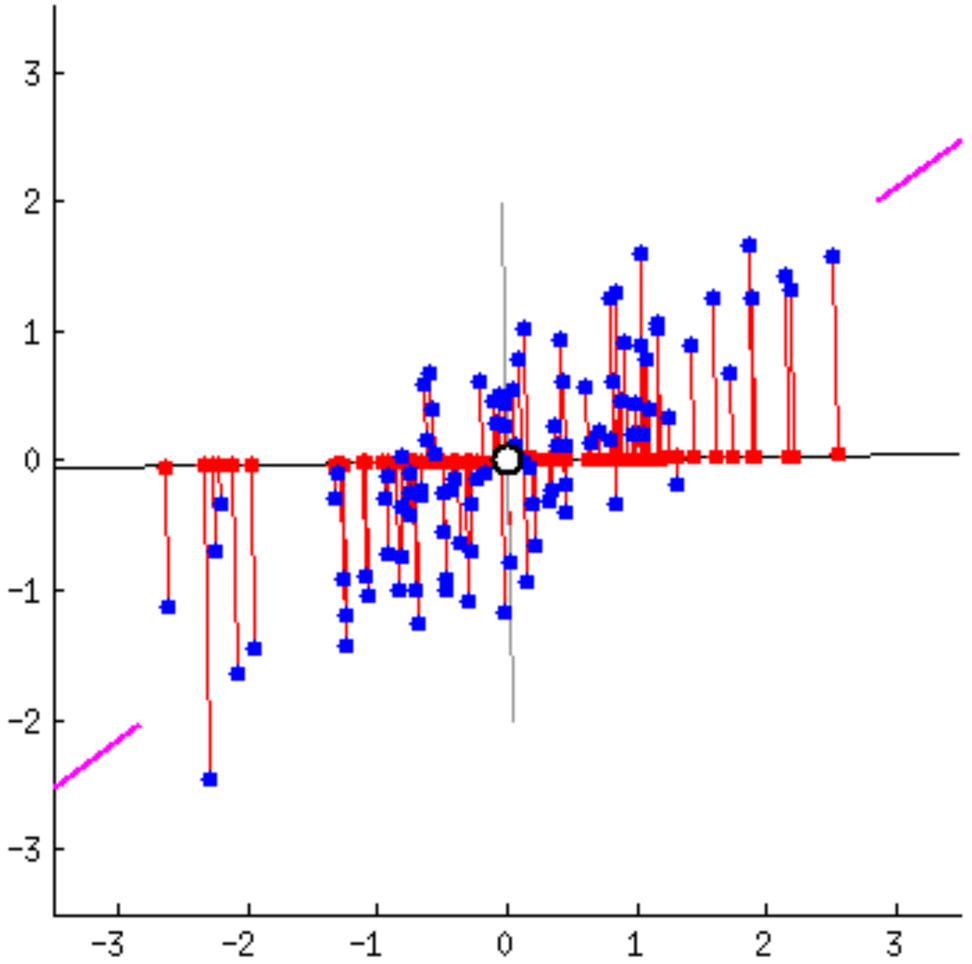


Second Principle Component

First Principle Component

the word "cuddly"

Principle Component Analysis (PCA)



Dimensionality Reduction

- Linear Algebra Prelims
- Singular Value Decomposition

Dimensionality Reduction

- **Linear Algebra Prelims**
- Singular Value Decomposition

Linear Algebra Prelims

Rank of a Matrix

2	1	1
4	3	1
2	0	2
8	4	4

Linear Algebra Prelims

Rank of a Matrix

2	=	1	+	1
4	=	3	+	1
2	=	0	+	2
8	=	4	+	4

Linear Algebra Prelims

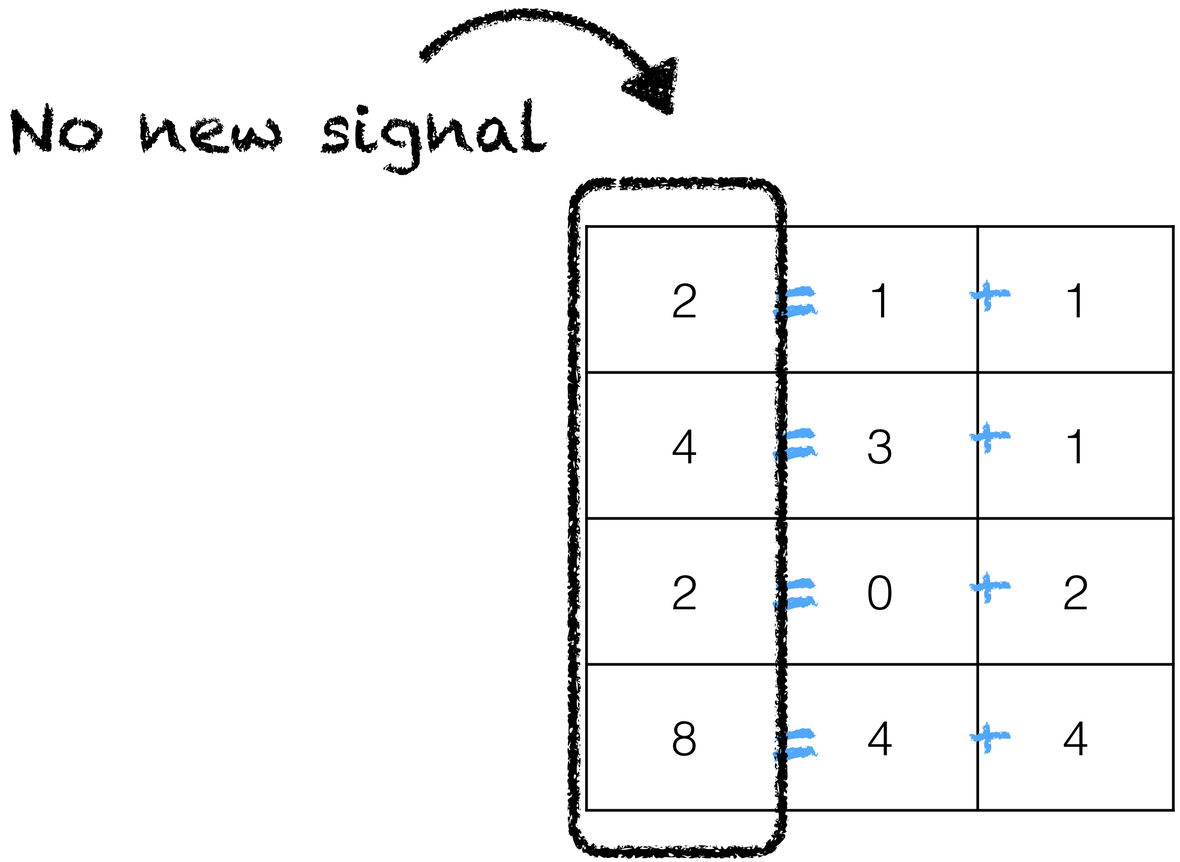
Rank of a Matrix

2	=	1	+	1
4	=	3	+	1
2	=	0	+	2
8	=	4	+	4

Rank = 2

Linear Algebra Prelims

Rank of a Matrix



Rank = 2

Linear Algebra Prelims

Rank of a Matrix

What is the rank of this matrix?

	the	congress	parliament	US	UK
doc1	1	1	1	1	0
doc2	1	0	1	0	1
doc3	1	1	0	1	0
doc4	1	0	1	0	1

a) 5
b) 4
c) 3
d) 2
e) 1

Linear Algebra Prelims

Rank of a Matrix

What is the rank of this matrix?

	the	**congress**	parliament	**US**	UK
doc1	1	**1**	1	**1**	0
doc2	1	**0**	1	**0**	1
doc3	1	**1**	0	**1**	0
doc4	1	**0**	1	**0**	1

a) 5
b) 4
c) 3
d) 2
e) 1

Linear Algebra Prelims

Rank of a Matrix

What is the rank of this matrix?

	the		parliament	US	UK
doc1	1		1	1	0
doc2	1		1	0	1
doc3	1		0	1	0
doc4	1		1	0	1

a) 5
b) 4
c) 3
d) 2
e) 1

Linear Algebra Prelims

Rank of a Matrix

What is the rank of this matrix?

	the		parliament	**US**	**UK**
doc1	**1**		1	**1**	**0**
doc2	**1**		1	**0**	**1**
doc3	**1**		0	**1**	**0**
doc4	**1**		1	**0**	**1**

a) 5
b) 4
c) 3
d) 2
e) 1

Linear Algebra Prelims

Rank of a Matrix

What is the rank of this matrix?

			parliament	US	UK
doc1			1	1	0
doc2			1	0	1
doc3			0	1	0
doc4			1	0	1

a) 5
b) 4
c) 3
d) 2
e) 1

Linear Algebra Prelims

Rank of a Matrix

What is the rank of this matrix?

			parliament	US	UK
doc1			1	1	0
doc2			1	0	1
doc3			0	1	0
doc4			1	0	1

a) 5
b) 4
c) 3
d) 2
e) 1

Linear Algebra Prelims

Rank of a Matrix

	the	congress	parliament	US	UK
doc1	1	1	1	1	0
doc2	1	0	1	0	1
doc3	1	1	0	1	0
doc4	1	0	1	0	1

"Low Rank Assumption": we typically assume that our features contain a large amount of redundant information

Linear Algebra Prelims

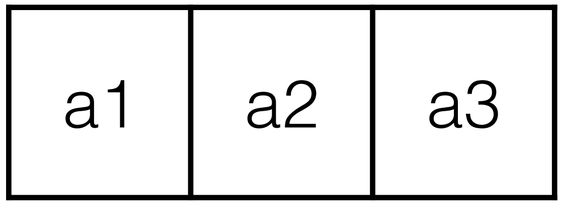
Rank of a Matrix

	the	congress	parliament	US	UK
doc1	1	1	1	1	0
doc2	1	0	1	0	1
doc3	1	1	0	1	0
doc4	1	0	1	0	1

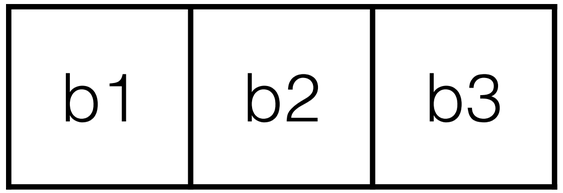
"Low Rank Assumption": we typically assume that our features contain a large amount of redundant information

Linear Algebra Prelims

Matrix Arithmetic



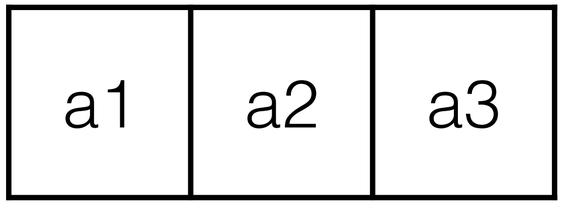
$\vec{a}$



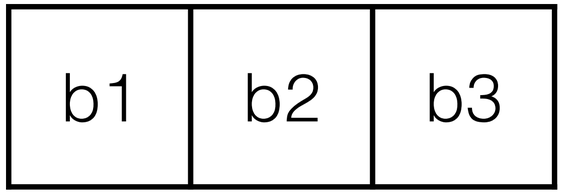
$\vec{b}$

Linear Algebra Prelims

Matrix Arithmetic



$\vec{a}$



$\vec{b}$

$$a \cdot \vec{b} = (a1xb1) + (a2xb2) + (a3xb3)$$

Linear Algebra Prelims

Matrix Arithmetic

a11	a12	a13
a21	a22	a23
a31	a32	a33

A

b11	b12
b21	b22
b31	b32

B

Linear Algebra Prelims

Matrix Arithmetic

a11	a12	a13
a21	a22	a23
a31	a32	a33

A

3x3

b11	b12
b21	b22
b31	b32

B

3x2

Linear Algebra Prelims

Matrix Arithmetic

a11	a12	a13
a21	a22	a23
a31	a32	a33

A

3x3

b11	b12
b21	b22
b31	b32

B

3x2

Linear Algebra Prelims

Matrix Arithmetic

a11	a12	a13
a21	a22	a23
a31	a32	a33

A

3x3

b11	b12
b21	b22
b31	b32

B

3x2

??	??
??	??
??	??

AB

3x2

Linear Algebra Prelims

Matrix Arithmetic

a11	a12	a13
a21	a22	a23
a31	a32	a33

A

mxk

b11	b12
b21	b22
b31	b32

B

kxn

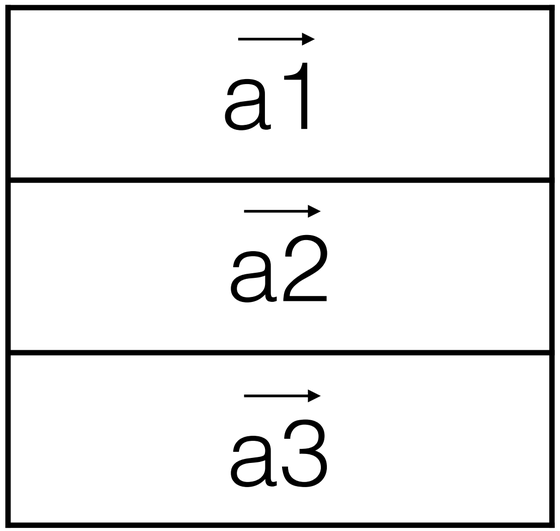
??	??
??	??
??	??

AB

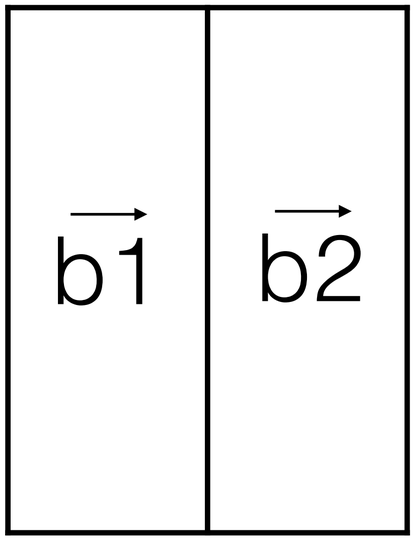
mxn

Linear Algebra Prelims

Matrix Arithmetic



A



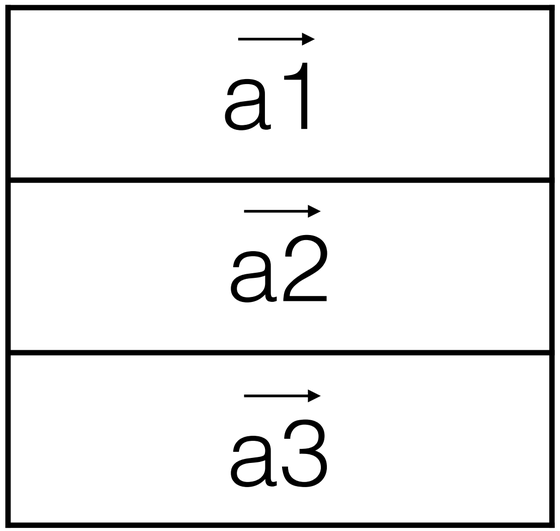
B

a1·b1	a1·b2
a2·b1	a2·b2
a3·b1	a3·b2

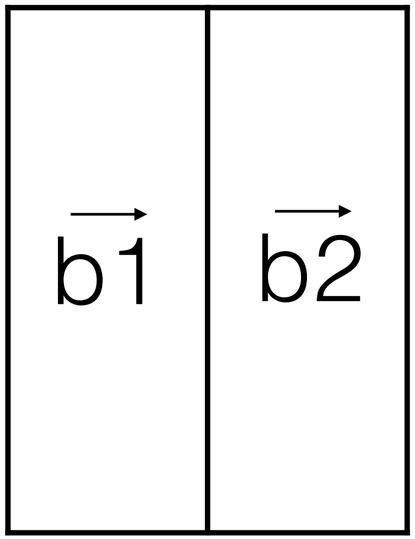
AB

Linear Algebra Prelims

Matrix Arithmetic



A



B

a1·b1	a1·b2
a2·b1	a2·b2
a3·b1	a3·b2

AB

$$AB[i][j] = ai \cdot bj$$

Linear Algebra Prelims

Matrix Arithmetic

1	2	3
3	4	5

X

2	4
1	2
3	1

(a)

13	25
13	6

(b)

14	7	6
20	10	10
26	13	14

(c)

13	11
25	25

Linear Algebra Prelims

Matrix Arithmetic

1	2	3
3	4	5

X

2	4
1	2
3	1

(a)

13	25
13	6

(b)

11	7	6
20	10	10
26	13	14

(c)

13	11
25	25

Linear Algebra Prelims

Matrix Arithmetic

1	2	3
3	4	5

X

2	4
1	2
3	1

(1x2)
+
(2x1)
+
(3x3)

(a)

13	25
13	6

(b)

11	7	6
20	10	10
26	13	14

(c)

13	11
25	25

Linear Algebra Prelims

Matrix Arithmetic

1	2	3
3	4	5

X

2	4
1	2
3	1

2 + 2 + 9

(a)

13	25
13	6

(b)

11	7	6
20	10	10
26	13	14

(c)

13	11
25	25

Linear Algebra Prelims

Matrix Arithmetic

1	2	3
3	4	5

X

2	4
1	2
3	1

4
+
4
+
3

(a)

13	25
13	6

(b)

11	7	6
20	10	10
26	13	14

(c)

13	11
25	25

Linear Algebra Prelims

Matrix Arithmetic

1	2	3
3	4	5

X

2	4
1	2
3	1

4 + 4 + 3

(a)

13	25
13	6

(b)

14	7	6
20	10	10
26	13	14

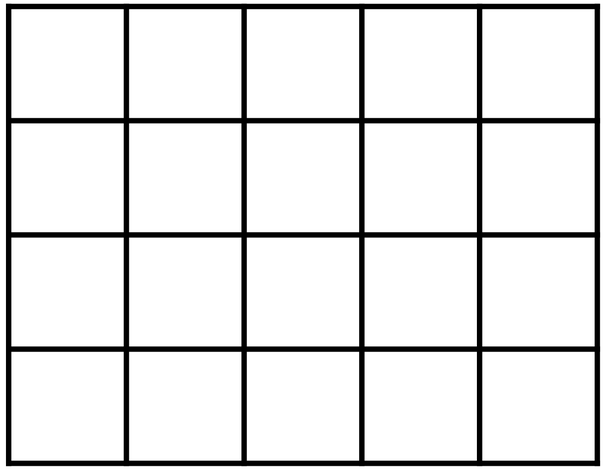
(c)

13	11
25	25

Dimensionality Reduction

- Linear Algebra Prelims
- **Singular Value Decomposition**

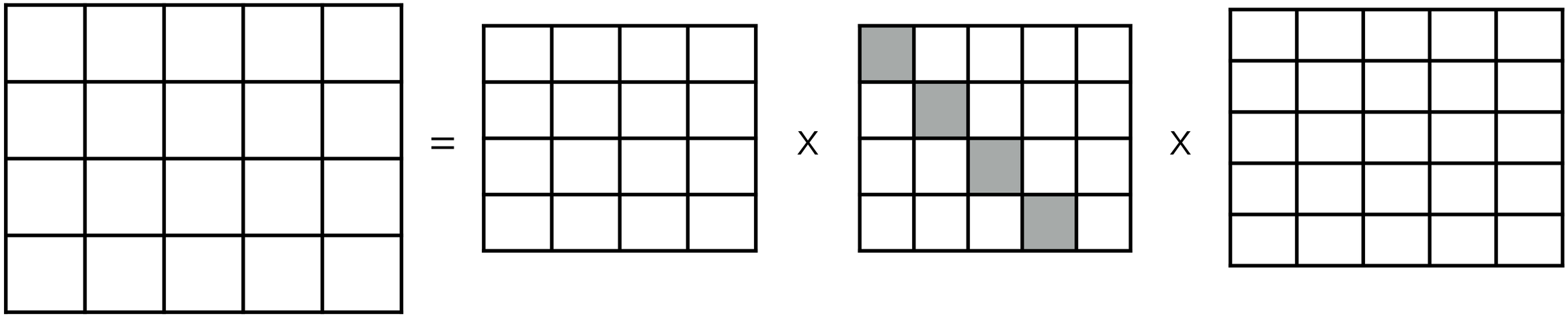
Singular Value Decomposition



M

m x n

Singular Value Decomposition



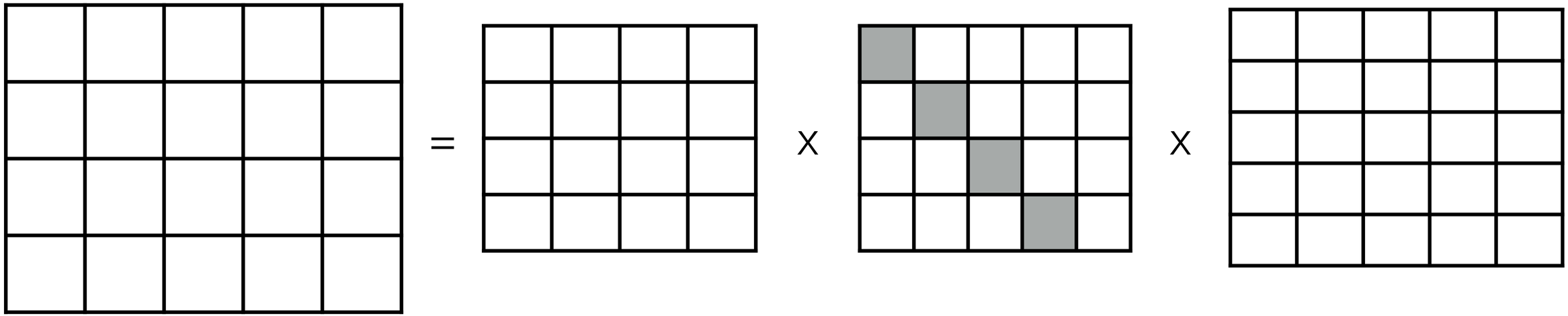
M
m x n

U
m x m

D
m x n

V
n x n

Singular Value Decomposition



M
m x n

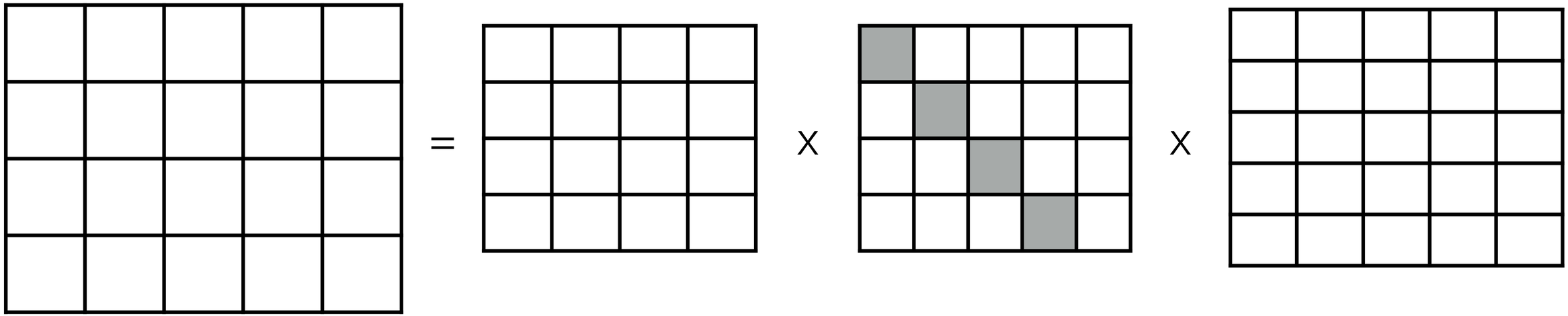
U
m x m

D
m x n

V
n x n

Data Matrix

Singular Value Decomposition



M
m x n

Data Matrix

=

U
m x m

x

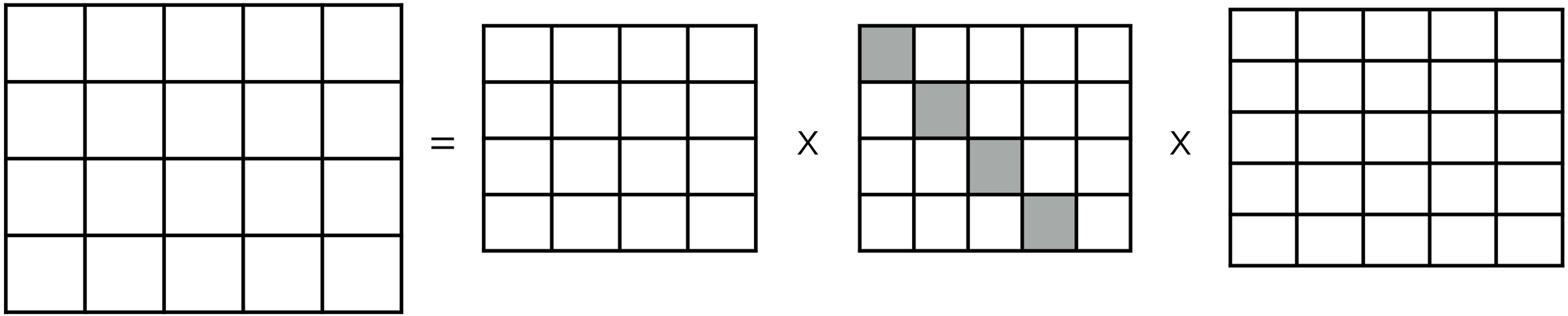
D
m x n

Singular Values of M

x

V
n x n

Singular Value Decomposition



M
m x n

U
m x m

D
m x n

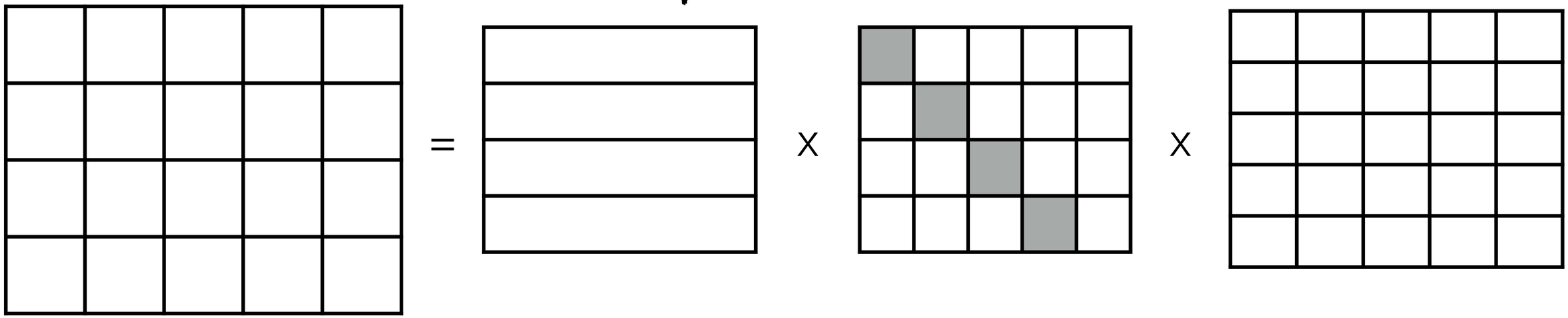
V
n x n

Data Matrix

Singular
Values of M
(#non-zero = rank M)

Singular Value Decomposition

Representation of rows of M in new feature space



M
m x n

U
m x m

D
m x n

V
n x n

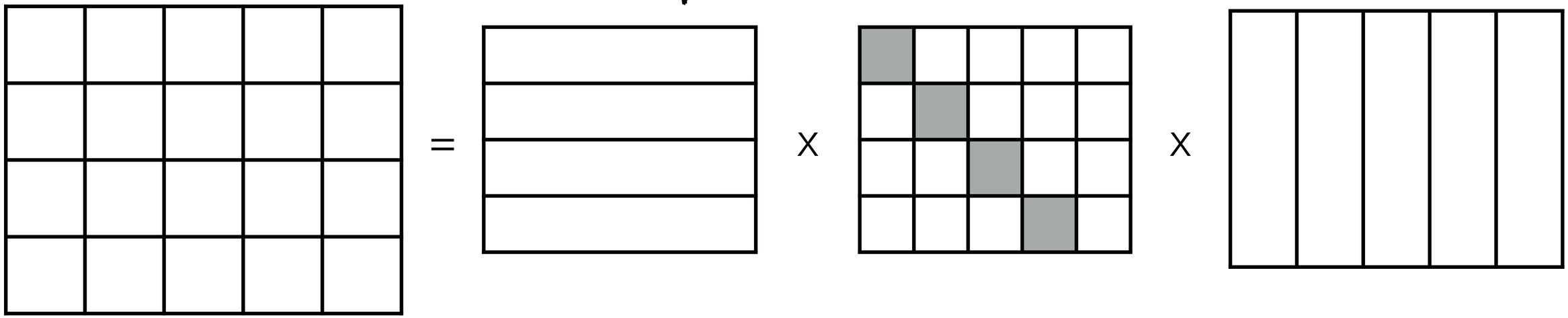
Data Matrix

Singular Values of M (#non-zero = rank M)

Singular Value Decomposition

Representation of rows of M in new feature space

Principle Components (new features)



M
m x n

U
m x m

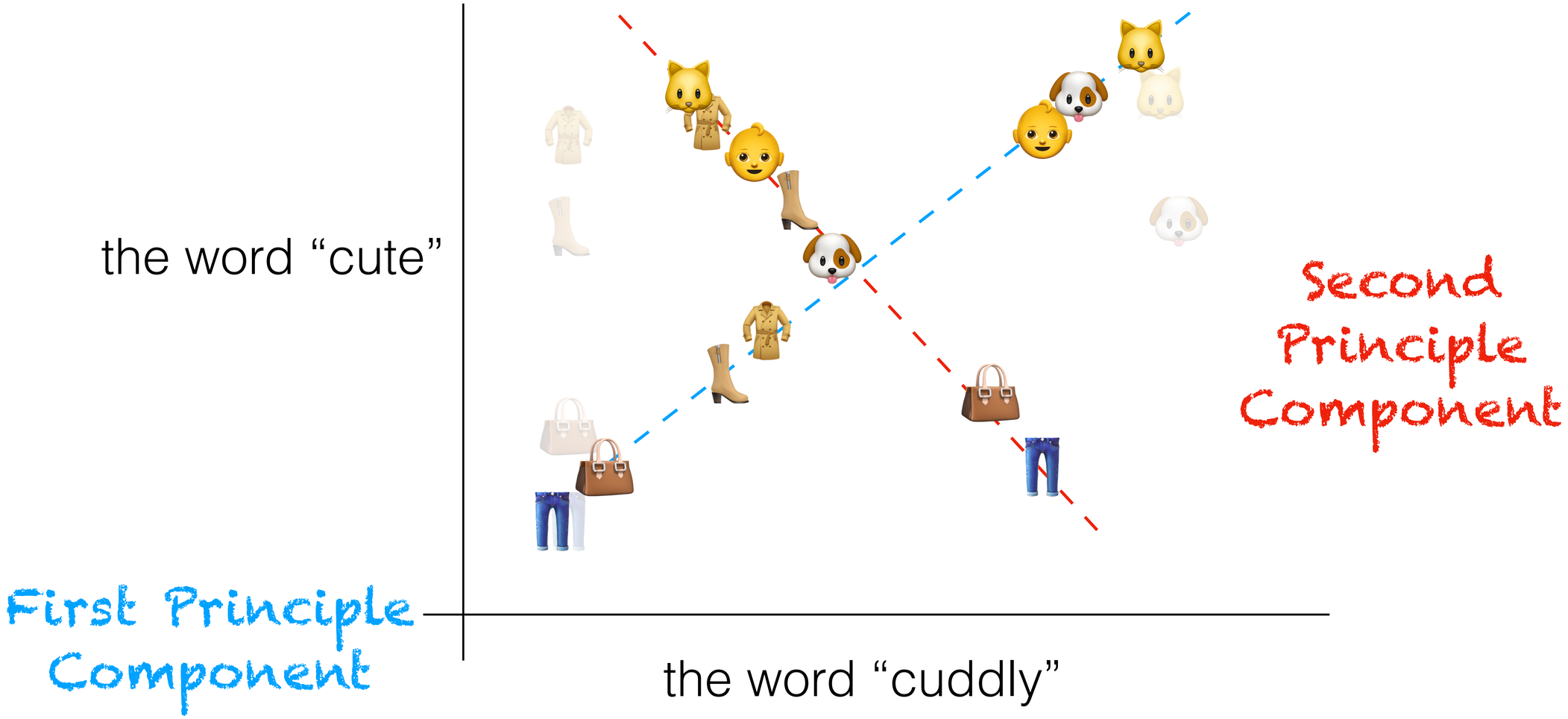
D
m x n

V
n x n

Data Matrix

Singular Values of M (#non-zero = rank M)

Singular Value Decomposition

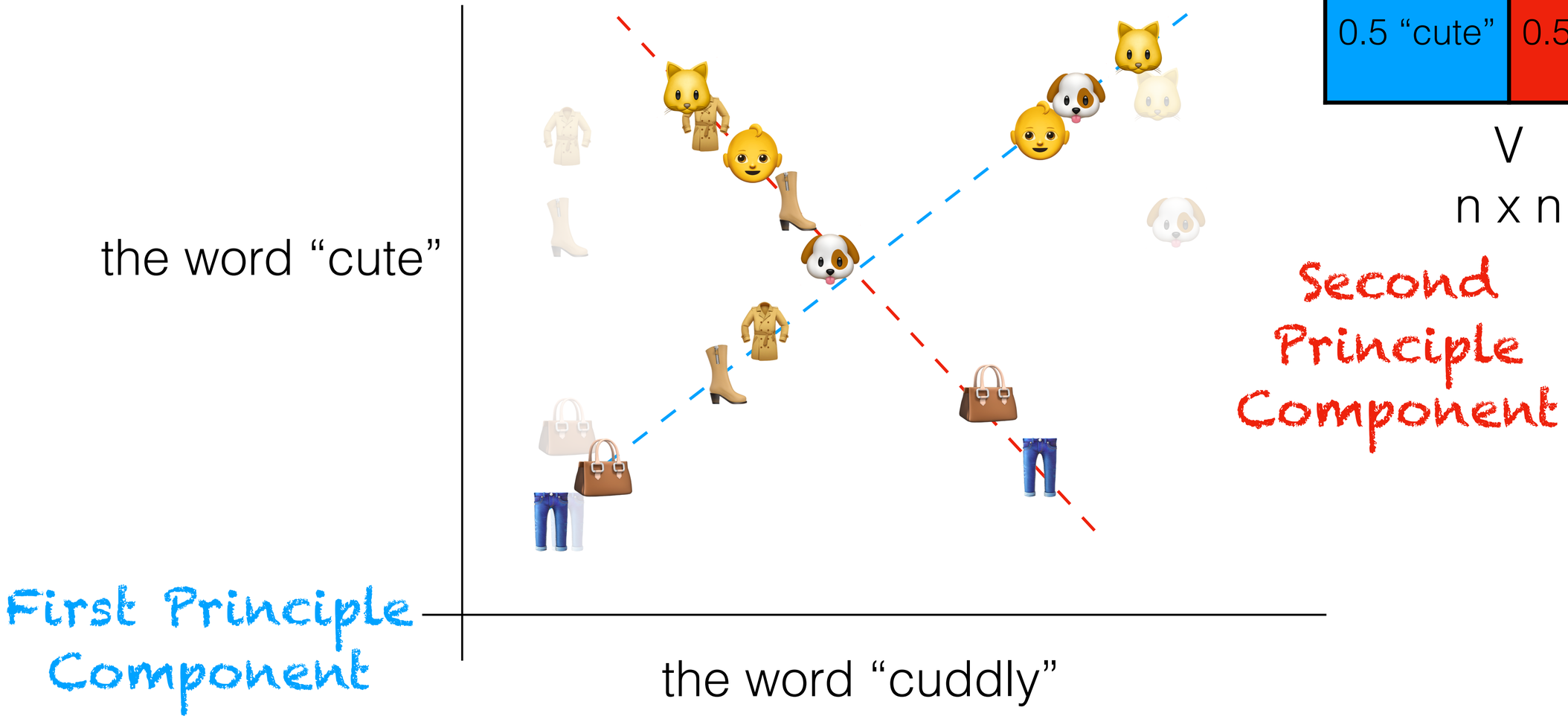


Singular Value Decomposition

0.5 “cuddly”	-0.5 “cuddly”
0.5 “cute”	0.5 “cute”

V

n x n

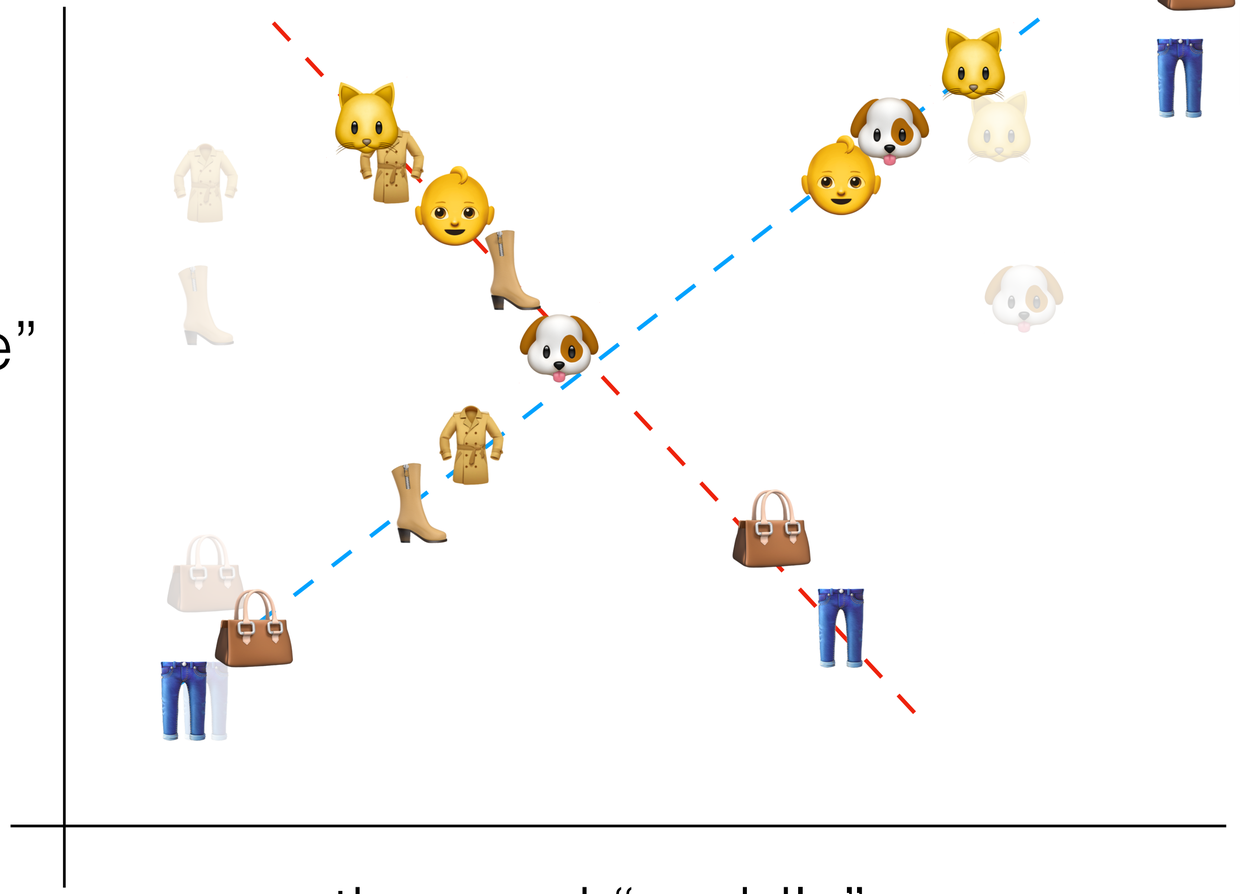


*warning: numbers are made up (not necessarily to scale)

Singular Value Decomposition

🐱	3	-3
🐶	2.5	-0.5
👜	-2	1.5
👖	-2.5	2

U
m x m

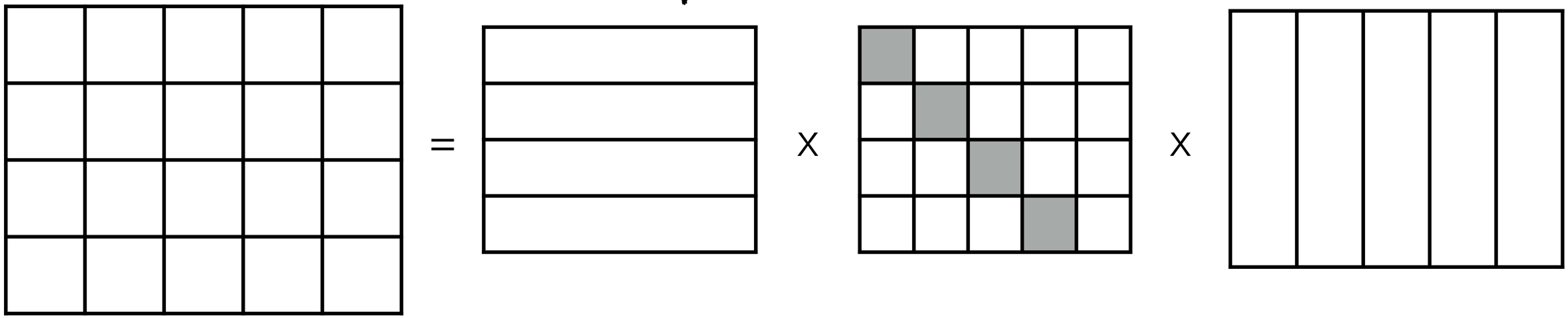


*warning: numbers are made up (not necessarily to scale)

Singular Value Decomposition

Representation of rows of M in new feature space

Principle Components (new features)



M
m x n

U
m x m

D
m x n

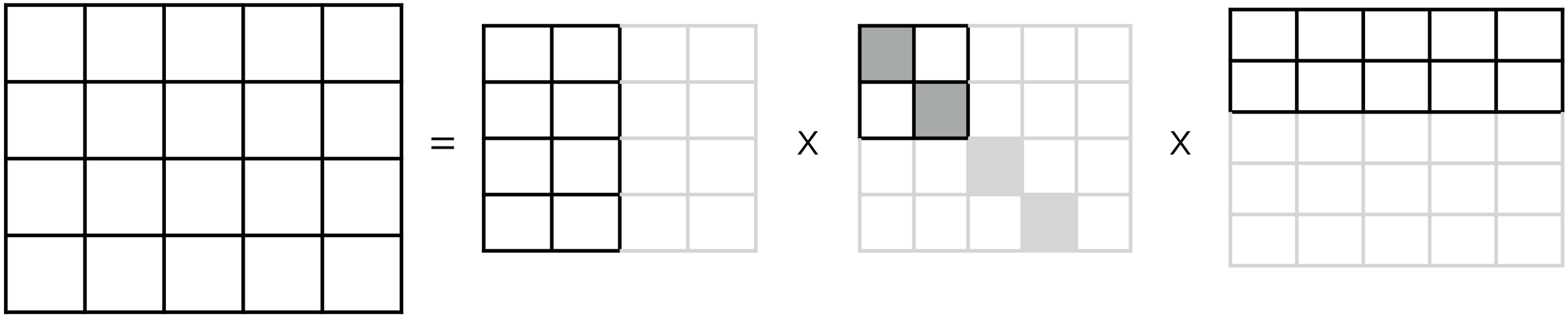
V
n x n

Data Matrix

Singular Values of M (#non-zero = rank M)

Truncated SVD

keep only first l components



M
m x n

U
m x l

D
l x l

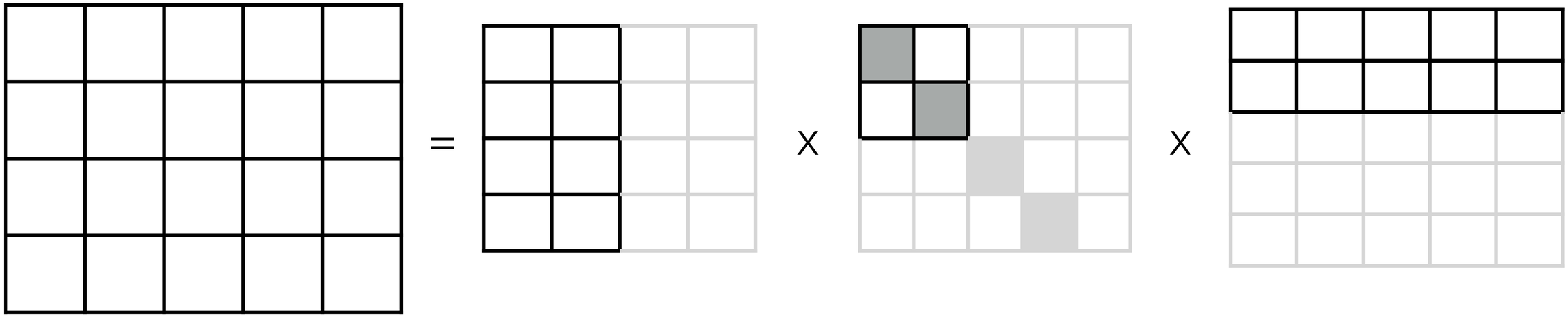
V
l x n

Data Matrix

Truncated SVD

keep only first l components

"best l-rank approximation of M"



M
m x n

U
m x l

D
l x l

V
l x n

Data Matrix

$||M - UDV||^2$ as small as possible

Truncated SVD

	cat	kitten	cute	adorable
cat	1	1	1	1
kitten	1	0	1	0
cute	1	1	0	1
adorable	1	0	1	1

M

cat	-0.60	-0.39	0.70	0.00
kitten	-0.48	0.50	-0.12	-0.71
cute	-0.43	-0.58	-0.69	0.00
adorable	-0.48	0.50	-0.12	0.71

U

3.06	0.00	0.00	0.00
0.00	1.81	0.00	0.00
0.00	0.00	0.57	0.00
0.00	0.00	0.00	0.00

D

cat	kitten	cute	adorable
-0.65	-0.34	-0.51	-0.34
0.02	-0.54	0.34	-0.54
-0.42	0.02	0.79	0.02
-0.63	0.27	0.00	0.37
-0.04	0.73	0.00	-0.68

V

Truncated SVD

Word-Context Matrix

	cat	kitten	cute	adorable
cat	1	1	1	1
kitten	1	0	1	0
cute	1	1	0	1
adorable	1	0	1	1

M

cat	-0.60	-0.39	0.70	0.00
kitten	-0.48	0.50	-0.12	-0.71
cute	-0.43	-0.58	-0.69	0.00
adorable	-0.48	0.50	-0.12	0.71

U

3.06	0.00	0.00	0.00
0.00	1.81	0.00	0.00
0.00	0.00	0.57	0.00
0.00	0.00	0.00	0.00

D

cat	kitten	cute	adorable
-0.65	-0.34	-0.51	-0.34
0.02	-0.54	0.34	-0.54
-0.42	0.02	0.79	0.02
-0.63	0.27	0.00	0.37
-0.04	0.73	0.00	-0.68

V

Truncated SVD

Word-Context Matrix

	cat	kitten	cute	adorable
cat	1	1	1	1
kitten	1	0	1	0
cute	1	1	0	1
adorable	1	0	1	1

M

Word Embeddings

cat	-0.60	-0.39	0.70	0.00
kitten	-0.48	0.50	-0.12	-0.71
cute	-0.43	-0.58	-0.69	0.00
adorable	-0.48	0.50	-0.12	0.71

U

3.06	0.00	0.00	0.00
0.00	1.81	0.00	0.00
0.00	0.00	0.57	0.00
0.00	0.00	0.00	0.00

D

cat	kitten	cute	adorable
-0.65	-0.34	-0.51	-0.34
0.02	-0.54	0.34	-0.54
-0.42	0.02	0.79	0.02
-0.63	0.27	0.00	0.37
-0.04	0.73	0.00	-0.68

V

Truncated SVD

Word-Context Matrix

	cat	kitten	cute	adorable
cat	1	1	1	1
kitten	1	0	1	0
cute	1	1	0	1
adorable	1	0	1	1

M

New Features ("Topics")

Word Embeddings

U

cat	-0.60	-0.39	0.70	0.00
kitten	-0.48	0.50	-0.12	-0.71
cute	-0.43	-0.58	-0.69	0.00
adorable	-0.48	0.50	-0.12	0.71

D

3.06	0.00	0.00	0.00
0.00	1.81	0.00	0.00
0.00	0.00	0.57	0.00
0.00	0.00	0.00	0.00

V

cat	kitten	cute	adorable
-0.65	-0.34	-0.51	-0.34
0.02	-0.54	0.34	-0.54
-0.42	0.02	0.79	0.02
-0.63	0.27	0.00	0.37
-0.04	0.73	0.00	-0.68

Word Embeddings

What's the point?

- Computational Reasons
  - Lower dimensional = less computationally intensive
- Better Generalization
  - Lower dimensional forces abstraction: two columns that capture effectively the same information can be combined
  - Lower dimensional removes noise: throw away columns that don't improve predictive power on held-out data
- Representational Richness
  - Dimensionality reduction can capture "second order" effects
  - E.g., w1 occurs with c1, w2 occurs with c2, c1 and c2 are similar. Thus, w1 and w2 are similar.

Word Embeddings

What's the point?

***Don't count, predict!* A systematic comparison of context-counting vs. context-predicting semantic vectors**

Marco Baroni and **Georgiana Dinu** and **Germán Kruszewski**
Center for Mind/Brain Sciences (University of Trento, Italy)
(marco.baroni|georgiana.dinu|german.kruszewski)@unitn.it

Word Embeddings

What's the point?

***Don't count, predict!* A systematic comparison of context-counting vs. context-predicting semantic vectors**

Marco Baroni and **Georgiana Dinu** and **Germán Kruszewski**
, Italy)
ki)@unitn.it

name	task	measure	source	soa
rg	relatedness	Pearson	Rubenstein and Goodenough (1965)	Hassan and Mihalcea (2011)
ws	relatedness	Spearman	Finkelstein et al. (2002)	Halawi et al. (2012)
wss	relatedness	Spearman	Agirre et al. (2009)	Agirre et al. (2009)
wsr	relatedness	Spearman	Agirre et al. (2009)	Agirre et al. (2009)
men	relatedness	Spearman	Bruni et al. (2013)	Bruni et al. (2013)
toefl	synonyms	accuracy	Landauer and Dumais (1997)	Bullinaria and Levy (2012)
ap	categorization	purity	Almuhareb (2006)	Rothenhäusler and Schütze (2009)
esslli	categorization	purity	Baroni et al. (2008)	Katrenko and Adriaans (2008)
battig	categorization	purity	Baroni et al. (2010)	Baroni and Lenci (2010)
up	sel pref	Spearman	Padó (2007)	Herdağdelen and Baroni (2009)
mcrae	sel pref	Spearman	McRae et al. (1998)	Baroni and Lenci (2010)
an	analogy	accuracy	Mikolov et al. (2013a)	Mikolov et al. (2013c)
ansyn	analogy	accuracy	Mikolov et al. (2013a)	Mikolov et al. (2013a)
ansem	analogy	accuracy	Mikolov et al. (2013a)	Mikolov et al. (2013c)

Word Embeddings

What's the point?

Don't count, predict! **A systematic comparison of context-counting vs. context-predicting semantic vectors**

Marco Baroni and **Georgiana Dinu** and **Germán Kruszewski**
Center for Mind/Brain Sciences (University of Trento, Italy)
(marco.baroni|georgiana.dinu|german.kruszewski)@unitn.it

	rg	ws	wss	wsr	men	toefl	ap	esslli	battig	up	mcrae	an	ansyn	ansem
best setup on each task														
cnt	74	62	70	59	72	76	66	84	98	41	27	49	43	60
pre	84	75	**80**	**70**	**80**	91	75	86	**99**	41	28	**68**	**71**	**66**
best setup across tasks														
cnt	70	62	70	57	72	76	64	84	98	37	27	43	41	44
pre	83	73	78	68	**80**	86	71	77	98	41	26	67	69	64
worst setup across tasks														
cnt	11	16	23	4	21	49	24	43	38	-6	-10	1	0	1
pre	74	60	73	48	68	71	65	82	88	33	20	27	40	10
best setup on rg														
cnt	(74)	59	66	52	71	64	64	84	98	37	20	35	42	26
pre	(84)	71	76	64	79	85	72	84	98	39	25	66	70	61

Word Embeddings

What's the point?

Don't count, predict! **A systematic comparison of context-counting vs. context-predicting semantic vectors**

Marco Baroni and **Georgiana Dinu** and **Germán Kruszewski**
Center for Mind/Brain Sciences (University of Trento, Italy)
(marco.baroni|georgiana.dinu|german.kruszewski)@unitn.it

	rg	ws	wss	wsr	men	toefl	ap	esslli	battig	up	mcrae	an	ansyn	ansem
best setup on each task														
cnt	74	62	70	59	72	76	66	84	98	41	27	49	43	60
pre	84	75	**80**	**70**	**80**	91	75	86	**99**	41	28	**68**	**71**	**66**
best setup across tasks														
cnt	70	62	70	57	72	76	64	84	98	37	27	43	41	44
pre	83	73	78	68	**80**	86	71	77	98	41	26	67	69	64
worst setup across tasks														
cnt	11	16	23	4	21	49	24	43	38	-6	-10	1	0	1
pre	74	60	73	48	68	71	65	82	88	33	20	27	40	10
best setup on rg														
cnt	(74)	59	66	52	71	64	64	84	98	37	20	35	42	26
pre	(84)	71	76	64	79	85	72	84	98	39	25	66	70	61