Neural Vector Representations beyond Words: Sentence and Document Embeddings

Gerard de Melo
http://gerard.demelo.org

Rutgers University

Gerard de Melo

Assistant Professor at Rutgers University

Head of Deep Data Lab

RUTGERS

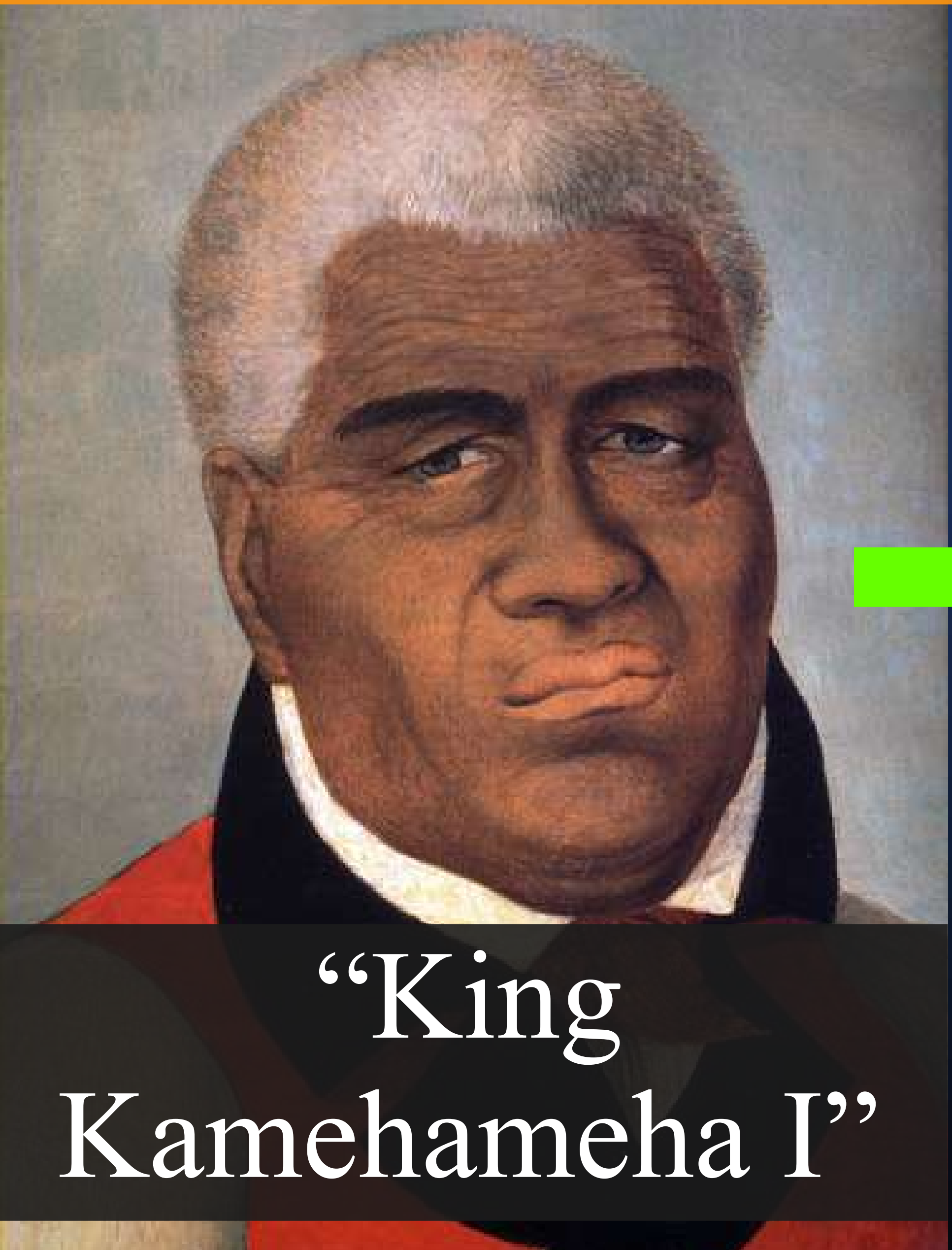
Image: https://commons.wikimedia.org/wiki/File:Rutgers_building_on_stilts_Livingston_campus.JPG

Vector Representations

$\vec{x}_i$
0.23
1.21
0.76
0.00
0.12

Image: https://en.wikipedia.org/wiki/File:Fire_ants_01.jpg

Vector Representations



$\vec{x}_i$
0.33
0.53
0.37
1.00
0.73

Image: https://commons.wikimedia.org/wiki/File:Kamehameha_I.png

Vector Representations

"Octel said the purchase was expected."

$\vec{x}_i$
0.33
0.53
0.37
1.00
0.73

Vector Representations

Honolulu (/ˌhɒnəˈluːluː/; Hawaiian: [honoˈlulu]) is the capital and largest city of the U.S. state of Hawaiʻi. It is an unincorporated part of and the county seat of the City and County of Honolulu along the southeast coast of the island of Oʻahu. The city is the main gateway to Hawaiʻi and a major portal into the United States. The city is also a major hub for international business, military defense, as well as famously being host to a diverse variety of east-west and Pacific culture, cuisine, and traditions.

Honolulu is the most remote city of its size in the world and is the westernmost major U.S. city. For statistical purposes, the United States Census Bureau recognizes the approximate area commonly referred to as "City of Honolulu" (not to be confused with the "City and County") as a census county division (CCD). Honolulu is a major financial center of the islands and of the Pacific Ocean. The population of the Honolulu census designated place (CDP) was 359,870 as of the 2017 population estimate, while the Honolulu CCD was 390,738 and the population of the consolidated city and county was 953,207.

Honolulu means "sheltered harbor" or "calm port". The old name is Kou, a district roughly encompassing the area from Nuʻuanu Avenue to Alakea Street and from Hotel Street to Queen Street which is the heart of the present downtown district. The city has been the capital of the Hawaiian Islands since 1845 and gained historical recognition following the attack on Pearl Harbor by Japan near the city on December 7, 1941.

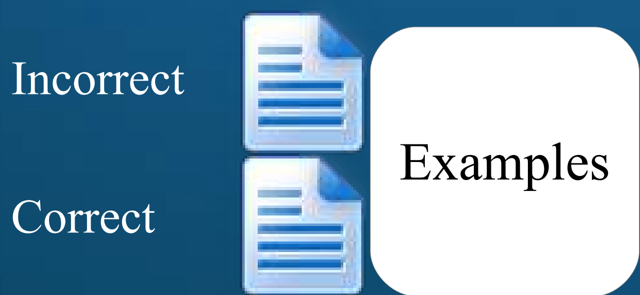
As of 2015, Honolulu was ranked high on world livability rankings, and was also ranked as the 2nd safest city in the U.S. It is also the most populated Oceanian city outside Australasia and ranks second to Auckland as the most-populous city in Polynesia.

$\vec{x}_i$
0.33
0.53
0.37
1.00
0.73

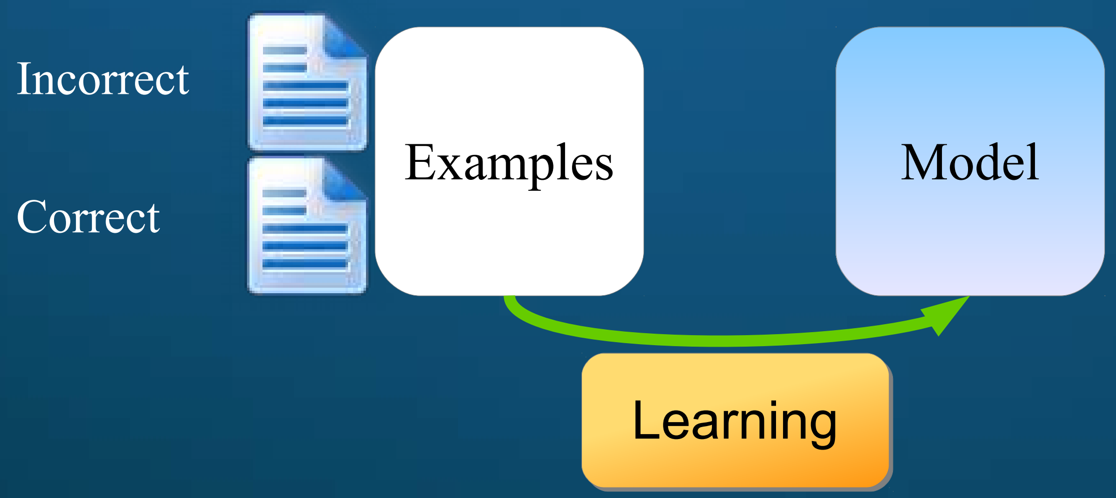
Image: https://en.wikipedia.org/wiki/Honolulu

Why Vectors Representations?

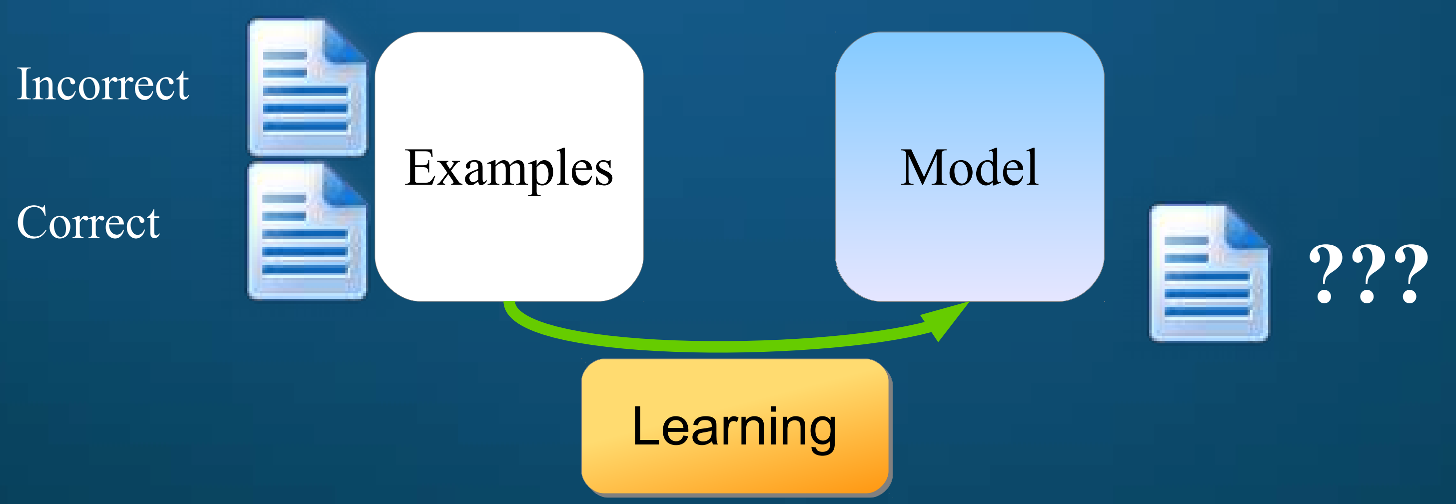
1. Facilitate Machine Learning



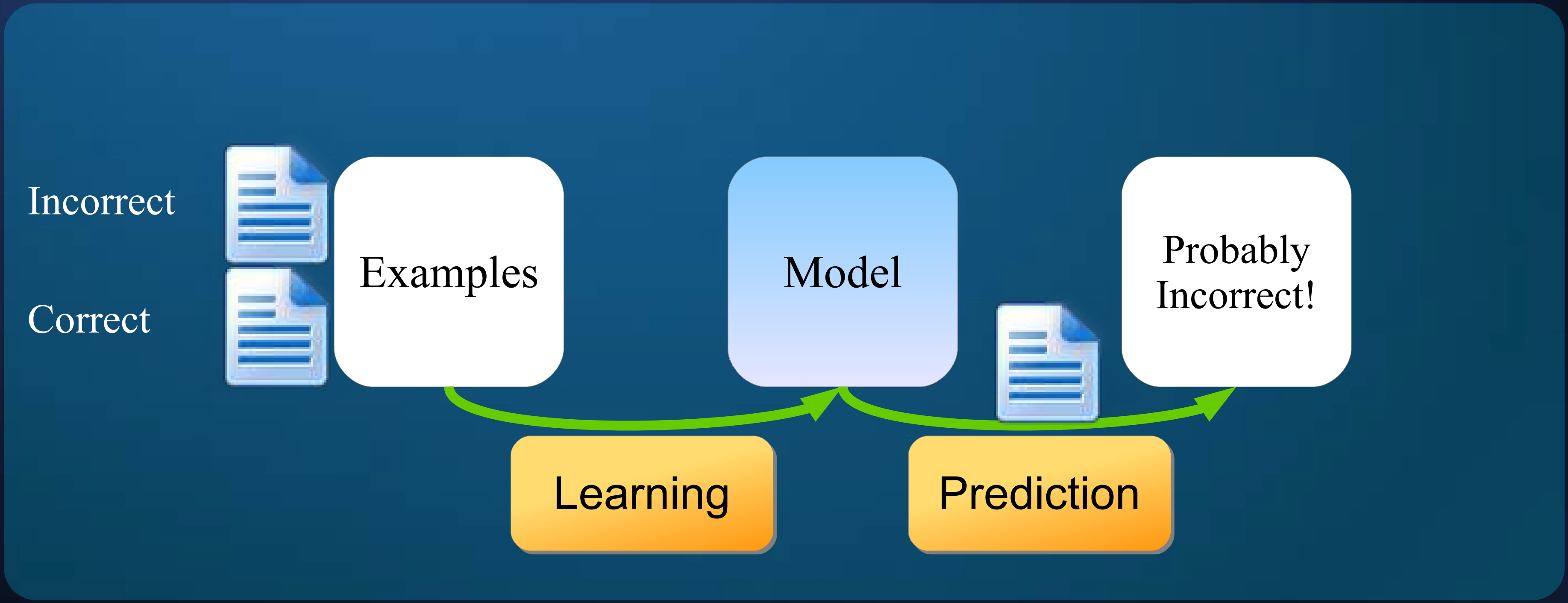
1. Facilitate Machine Learning



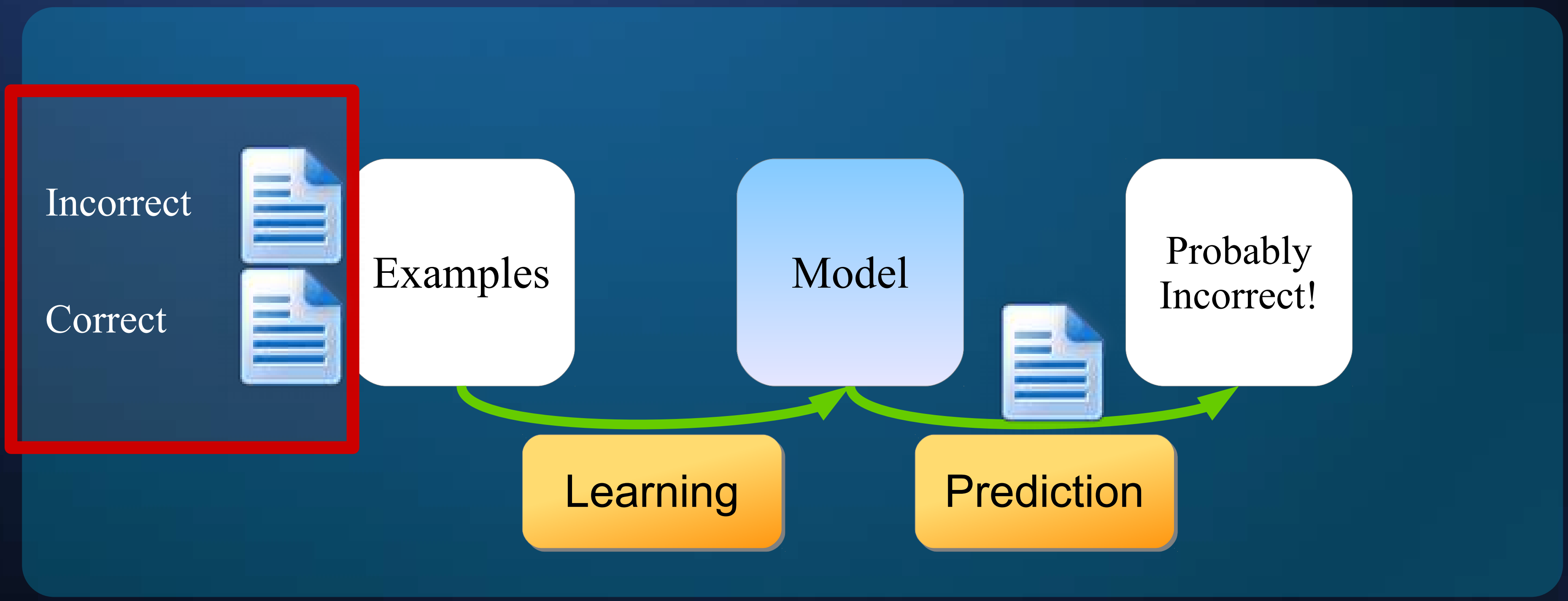
1. Facilitate Machine Learning



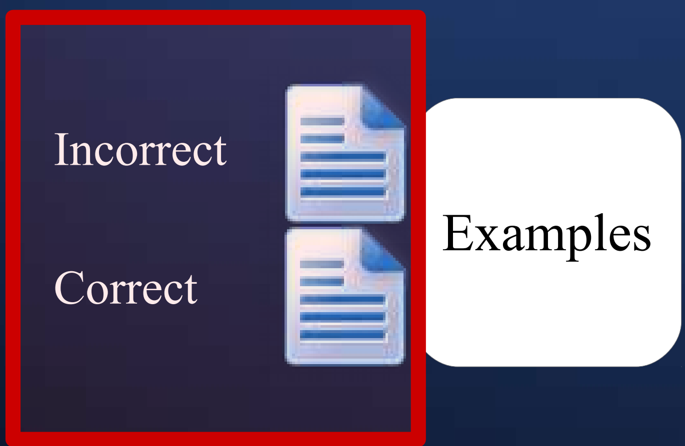
1. Facilitate Machine Learning



1. Facilitate Machine Learning



1. Facilitate Machine Learning



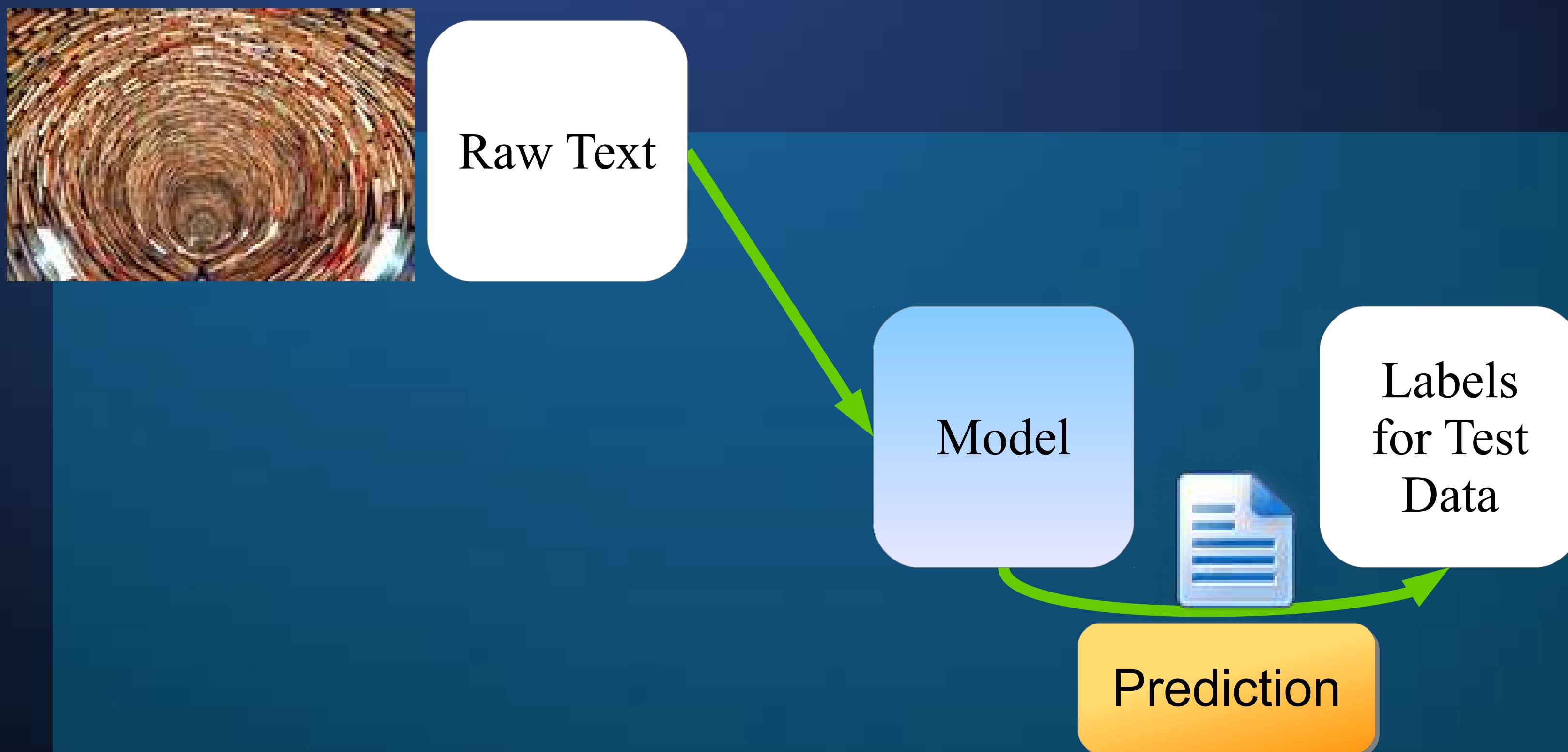
Problem:
Training Data usually requires human work, which is slow!

Example:
Penn Chinese TreeBank

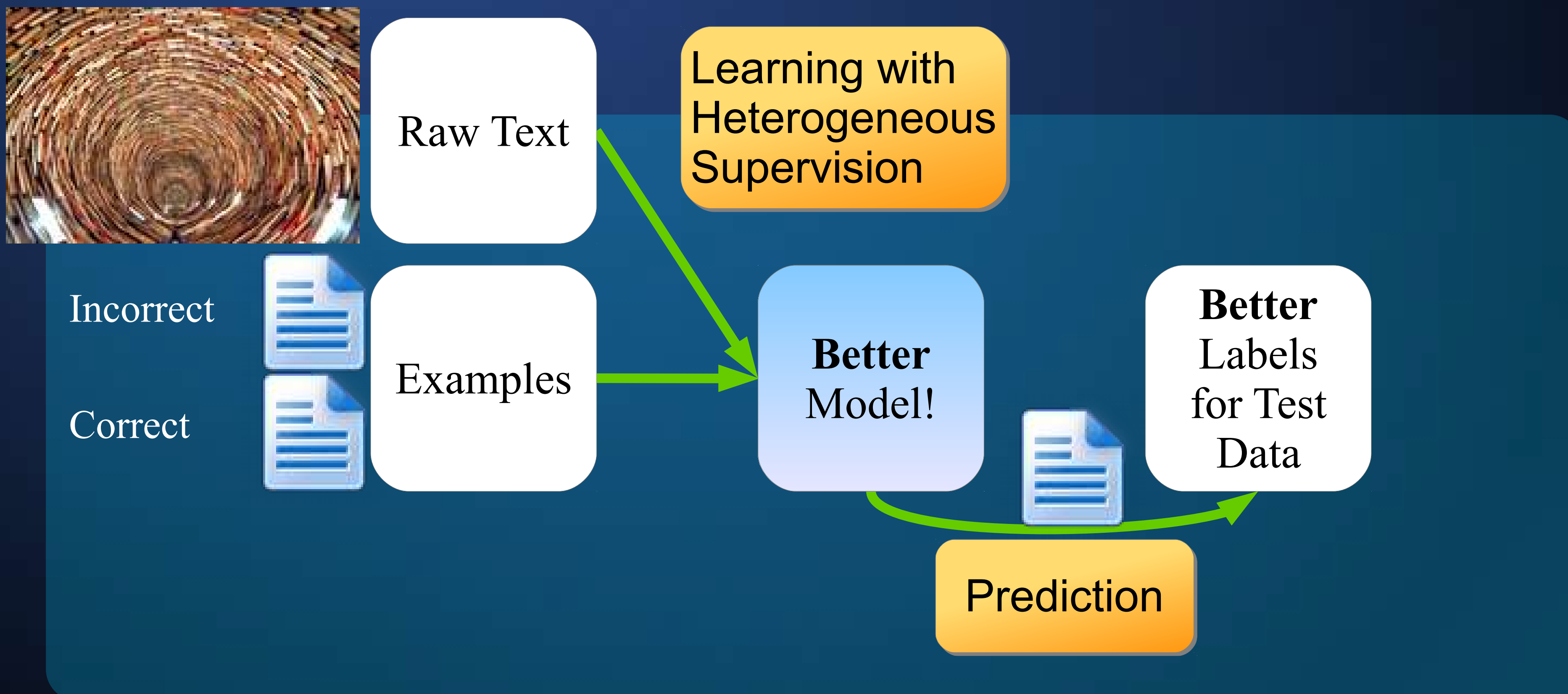
2 years for
4000 sentences

New effort required for each new **language** and **domain**
(e.g. news vs. tweets vs. biomedical text)

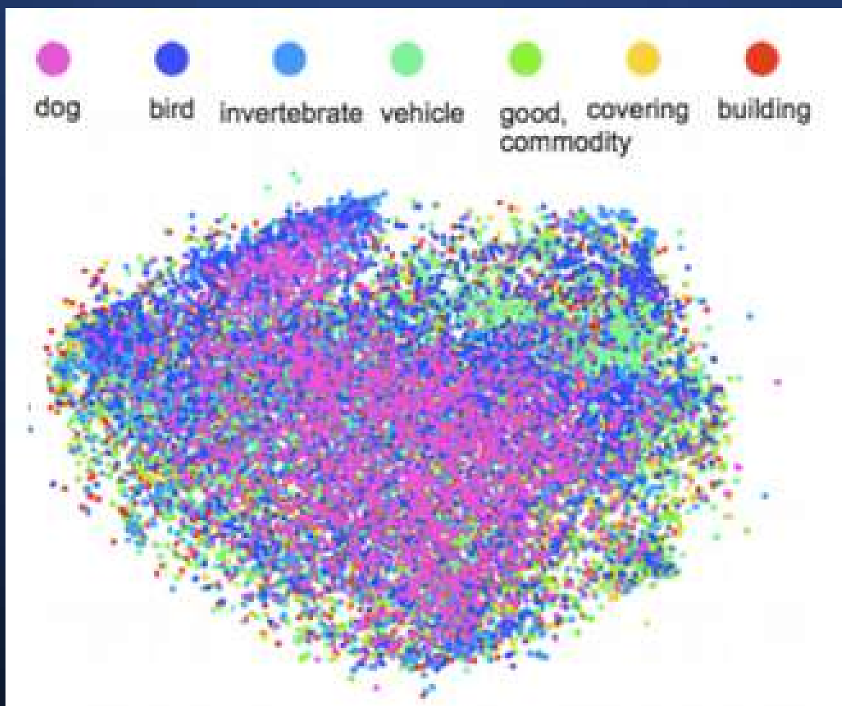
1. Facilitate Machine Learning



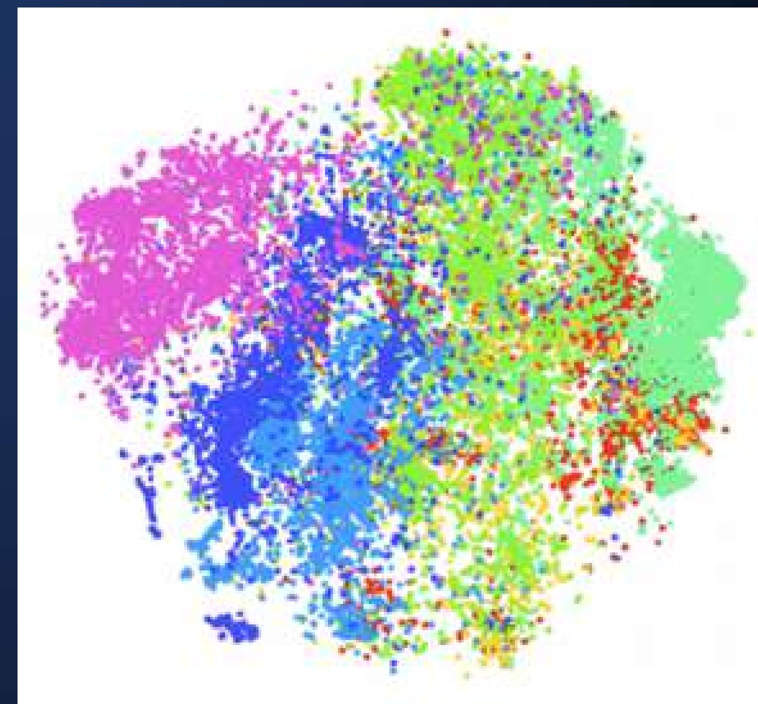
1. Facilitate Machine Learning



1. Facilitate Machine Learning



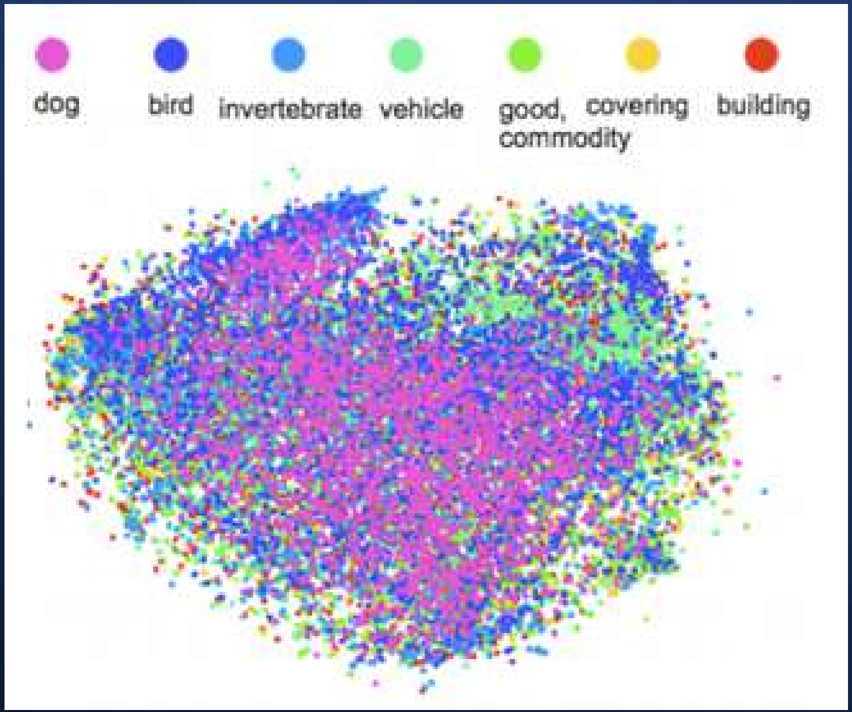
Vectors with raw/low-level data (e.g. pixels)

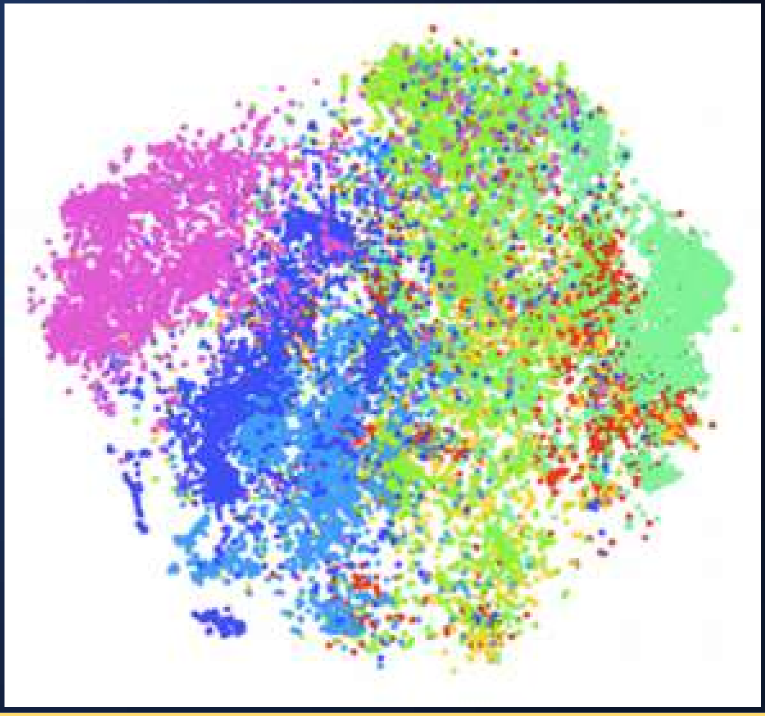


Vectors with higher-level representations (DeCAF)

Images: Donahue et al. 2014

1. Facilitate Machine Learning





Vectors with raw/low-level data
(e.g. pixels)

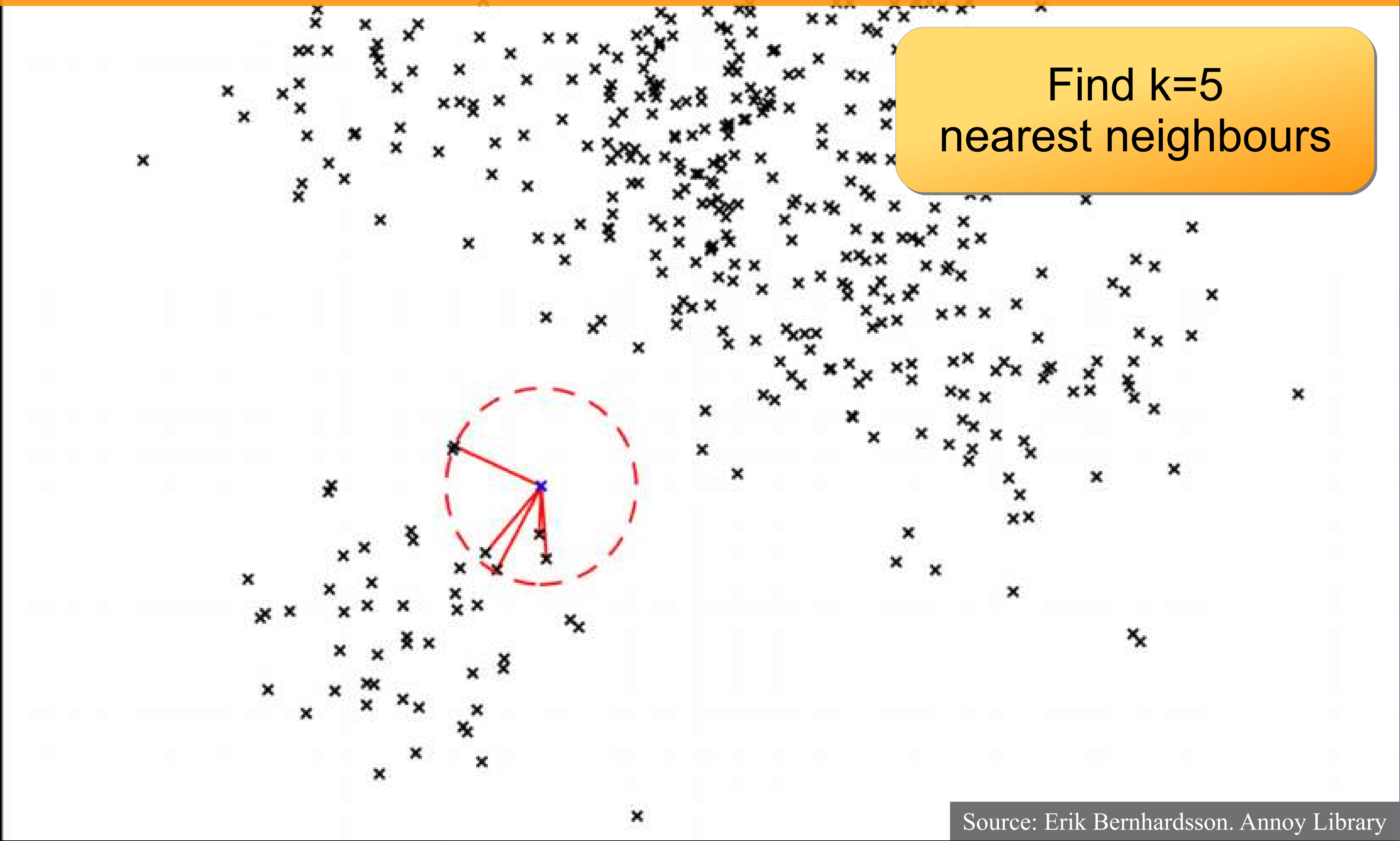
With good representations, machine learning becomes much easier

Images: Donahue et al. 2014

2. Find Nearest Neighbours

Image: https://www.teepublic.com/t-shirt/1279333-please-direct-me-to-the-nearest-pub

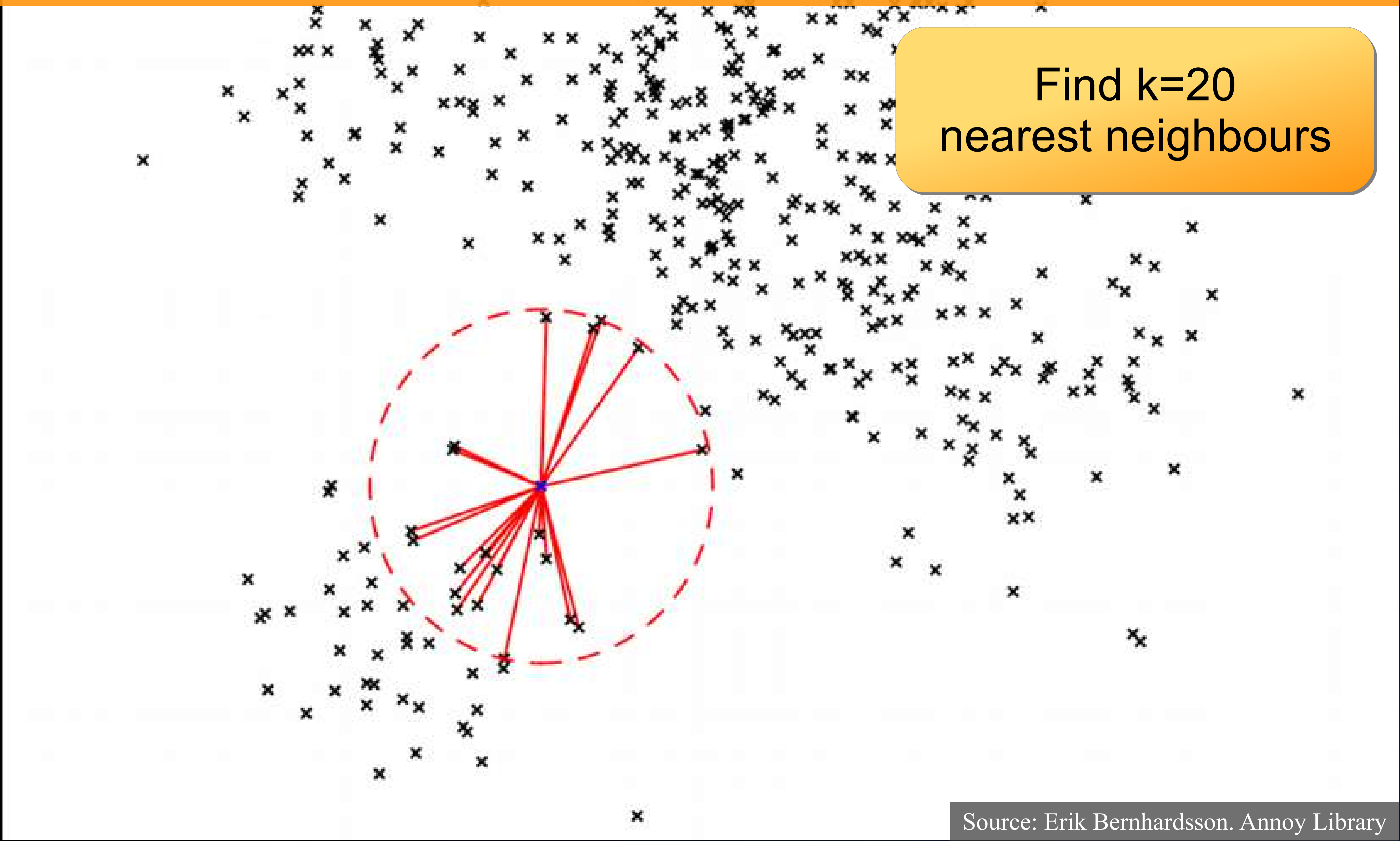
2. Find Nearest Neighbours



Find k=5
nearest neighbours

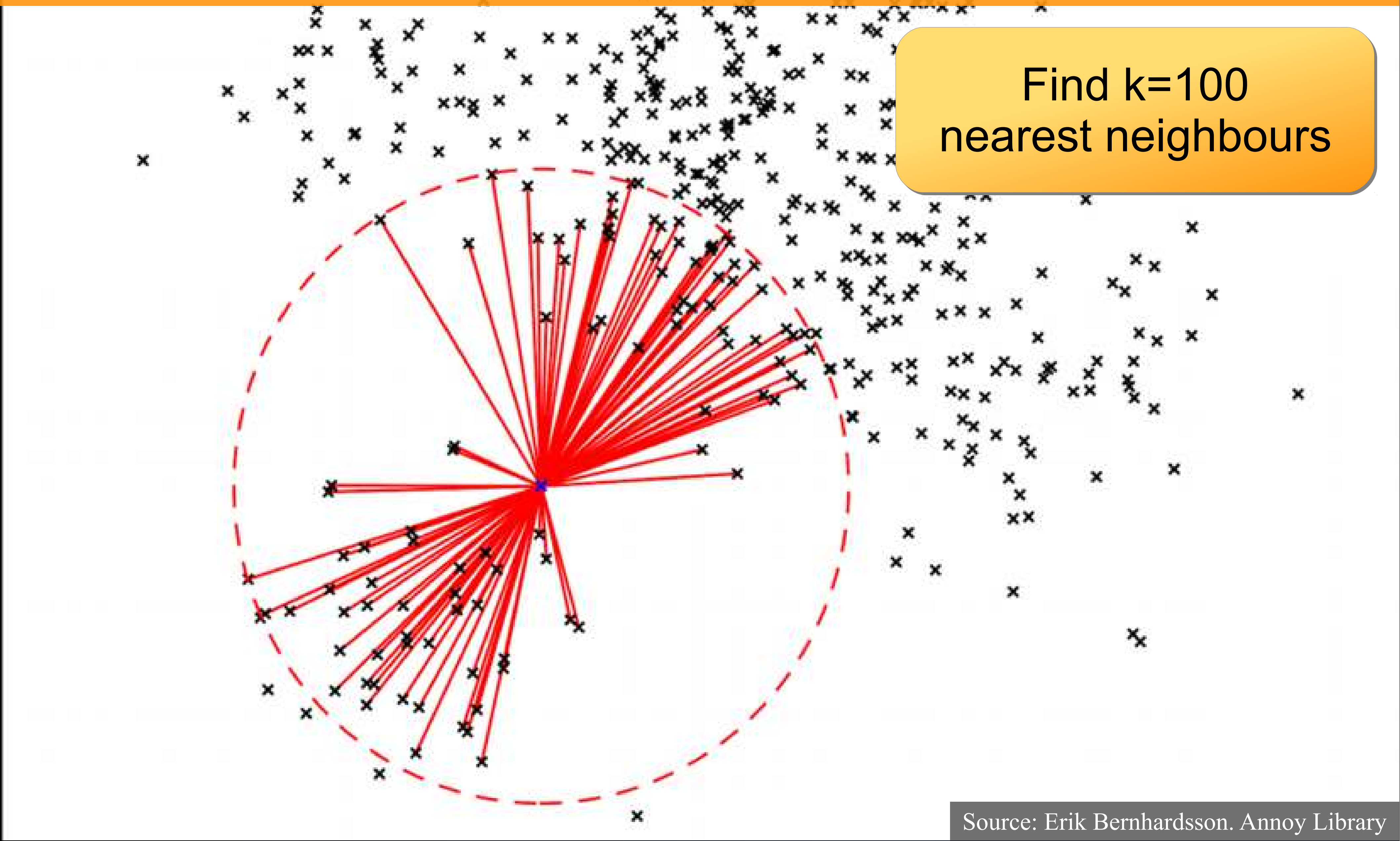
Source: Erik Bernhardsson. Annoy Library

2. Find Nearest Neighbours



Source: Erik Bernhardsson. Annoy Library

2. Find Nearest Neighbours



Find k=100 nearest neighbours

Source: Erik Bernhardsson. Annoy Library

2. Find Nearest Neighbours

Enter word or sentence (EXIT to break): Chinese river

Word	Cosine distance
Yangtze_River	0.667376
Yangtze	0.644091
Qiantang_River	0.632979
Yangtze_tributary	0.623527
Xiangjiang_River	0.615482
Huangpu_River	0.604726
Hanjiang_River	0.598110
Yangtze_river	0.597621
Hongze_Lake	0.594108
Yangtse	0.593442

Source: Google word2vec

Outline

- Word Representations
- Phrase Representations
- Sentence Representations
- Document Representations
- Applications and Outlook

Outline

- Word Representations
- Phrase Representations
- Sentence Representations
- Document Representations
- Applications and Outlook

Public Domain Image from https://pixabay.com/en/book-text-read-paper-education-451067/

Word Representations

“dry” ≠ “arid”

Word Representations

“dry”

0
0
0
1
0
0
0
…
0
0

≠

“arid”

0
0
0
0
1
0
0
…
0
0

Even when you have Bag-of-Words representations

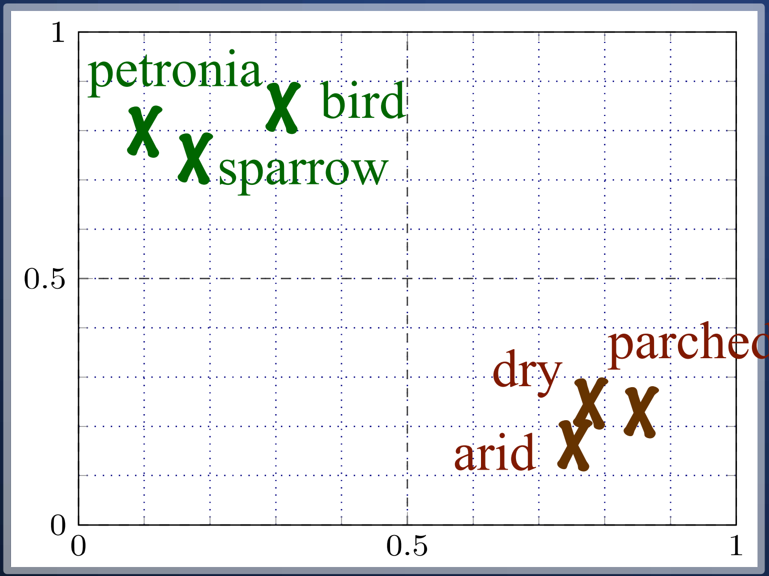
Word Representations

"dry"	≈	"arid"
0.00		0.01
0.23		0.23
0.03		0.01
0.31		0.29
0.01		0.00
0.03		0.03
0.91		0.92
...		...
0.31		0.30
0.50		0.51

Distributed Vector Representation

(e.g. with d=300 dimensions)

Word Representations

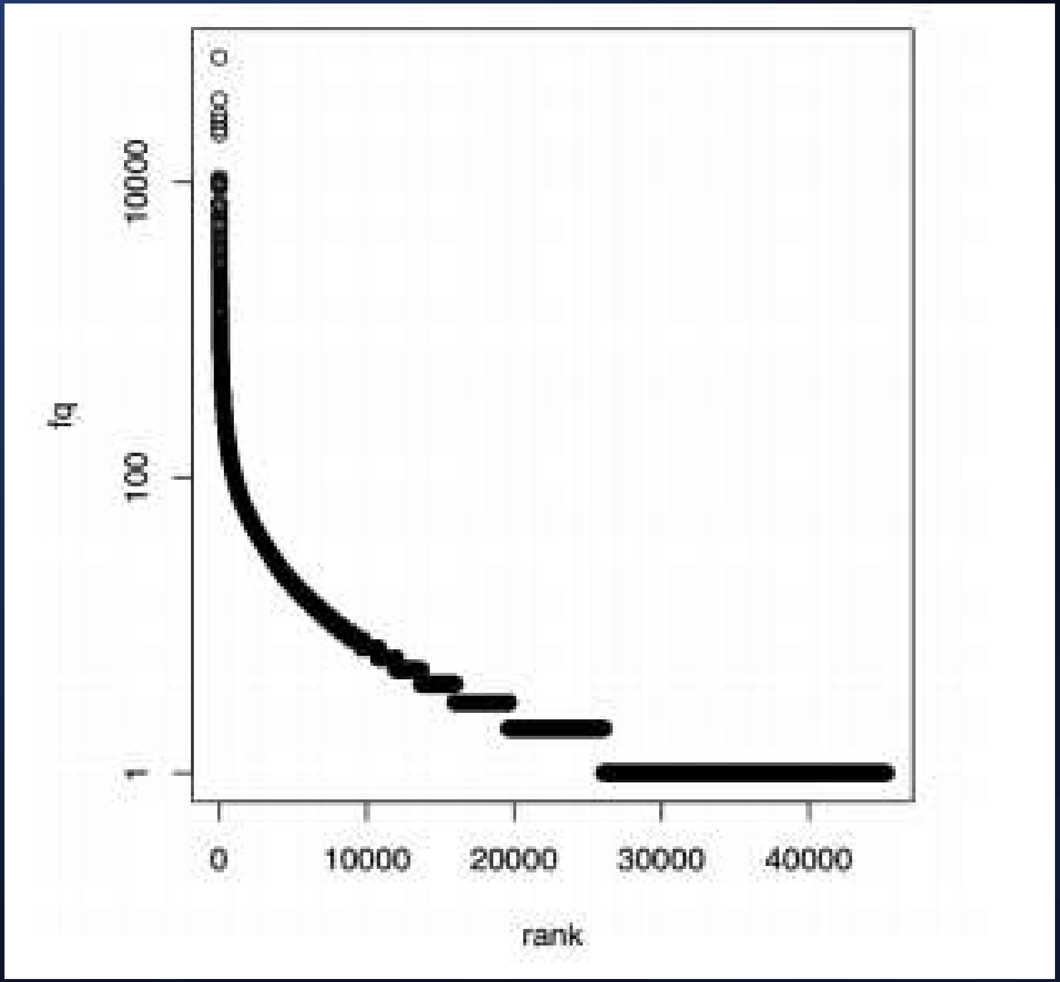


Why Generalization is Important

Data Sparsity:
Features may be rare.

→ Missing in training data

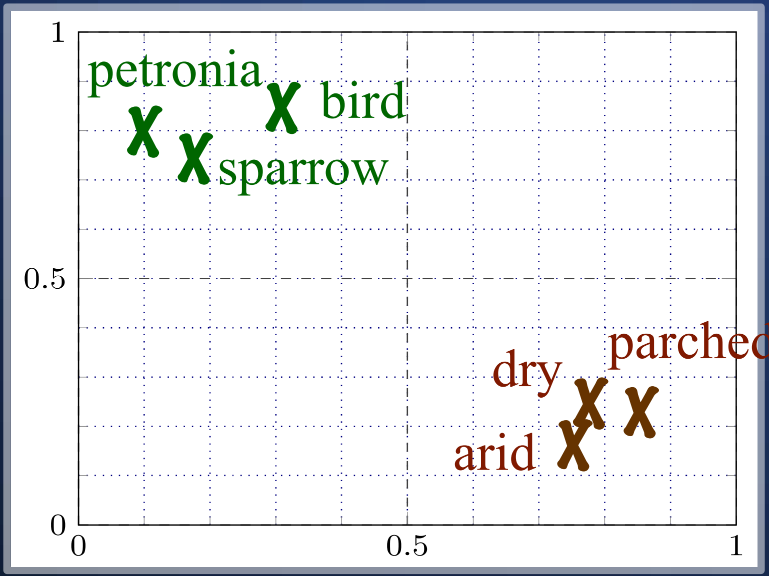
E.g. most words have a low frequency (in the “long tail”)



Frequencies of Words in Brown Corpus

Image: Baroni & Evert

Word Representations



How can we obtain such vector representations?

Distributional Semantics

We found a little, hairy
wampimuk
sleeping behind the tree

cf. McDonald & Ramscar (2001)

Image: https://commons.wikimedia.org/wiki/File:Mahogany_glider.jpg

Distributional Semantics

He filled the **wampimuk**,
passed it around
and we all drunk some.

cf. McDonald & Ramscar (2001)

Term-Term Cooccurrence Matrix

dogs are animals

cats are animals

orchids are plants

roses are plants

	Animals	Are	Cats	Dogs	Orchids	Plants	Roses
Animals		X	X	X			
Are	X		X	X	X	X	X
Cats	X	X					
Dogs	X	X					
Orchids		X				X	
Plants		X			X		X
Roses		X				X	

Image: Rafael Banchs

Term-Term Cooccurrence Matrix

dogs are animals

cats are animals

orchids are plants

roses are plants

	Animals	Are	Cats	Dogs	Orchids	Plants	Roses
Animals		X	X	X			
Are	X		X	X	X	X	X
Cats	X	X					
Dogs	X	X					
Orchids		X				X	
Plants		X			X		X
Roses		X				X	

Image: Rafael Banchs

Term-Document Matrix

D1: dogs are animals

D2: cats are animals

D3: orchids are plants

D4: roses are plants

	D1	D2	D3	D4
Animals	X	X		
Are	X	X	X	X
Cats		X		
Dogs	X			
Orchids			X	
Plants			X	X
Roses				X

Large context scope.
Typically, more topic-oriented similarity.

Classic methods: Dimensionality Reduction on matrix (SVD)

Image: Rafael Banchs

Learning Representations

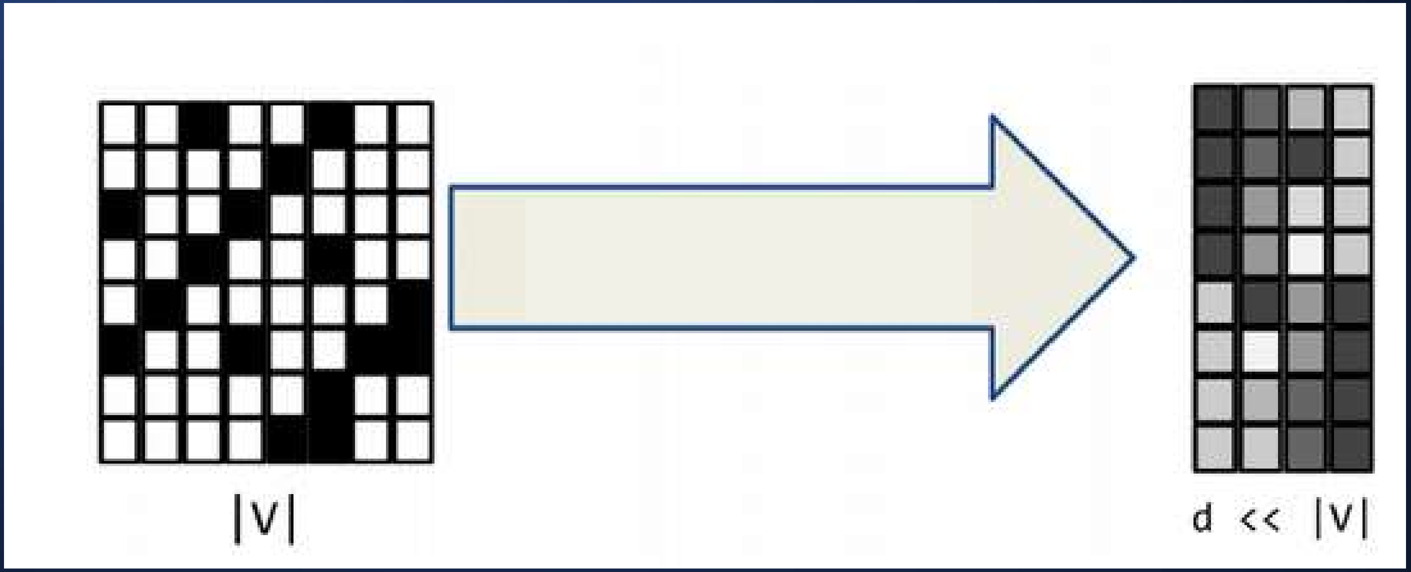


Image: Adapted from Perozzi et al. 2014. DeepWalk: Online Learning of Social Representations

SVD

E.g.
Cooccurrence Matrix

Image: Adapted from Tim Roughgarden & Gregory Valiant

SVD

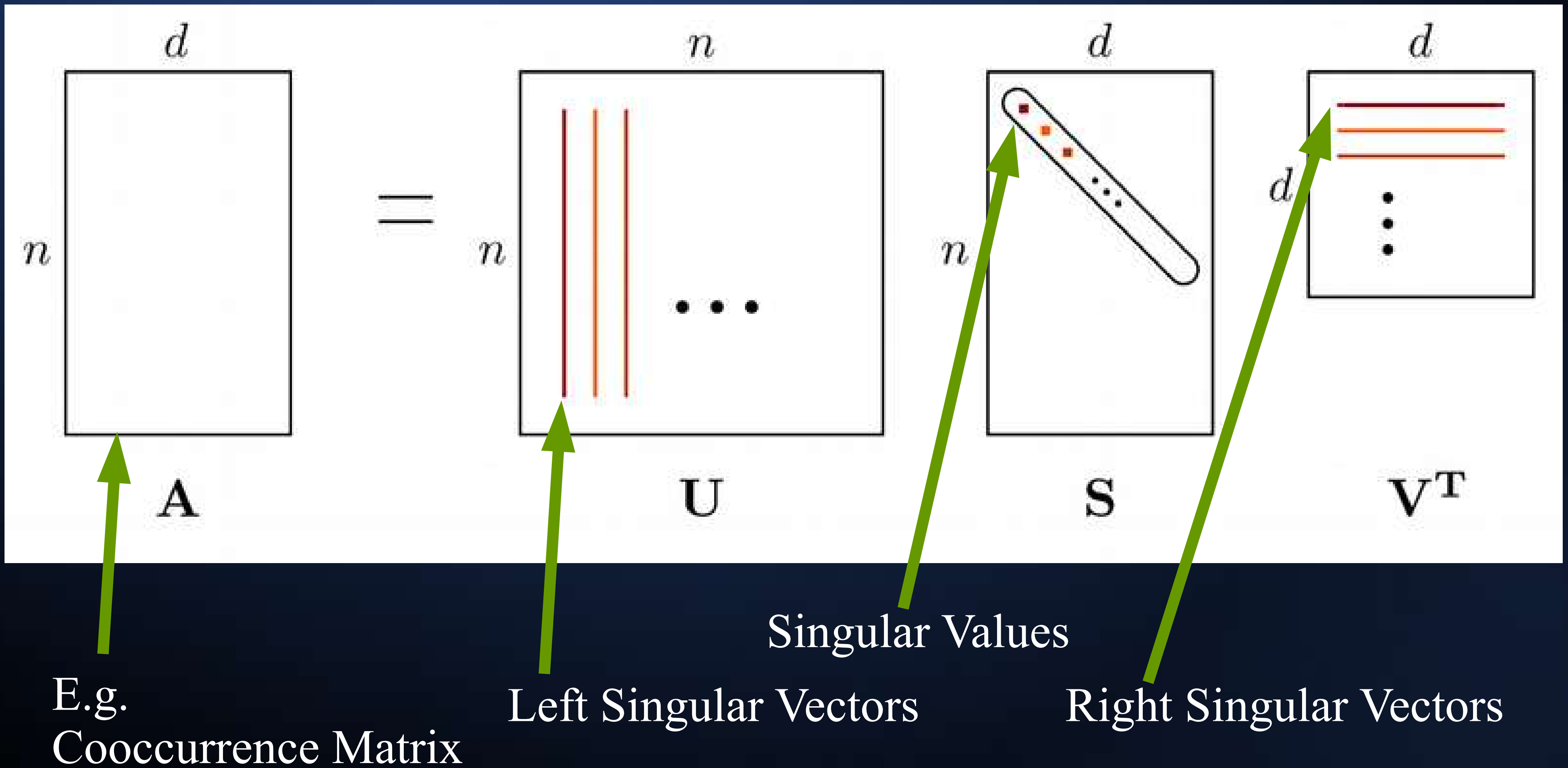
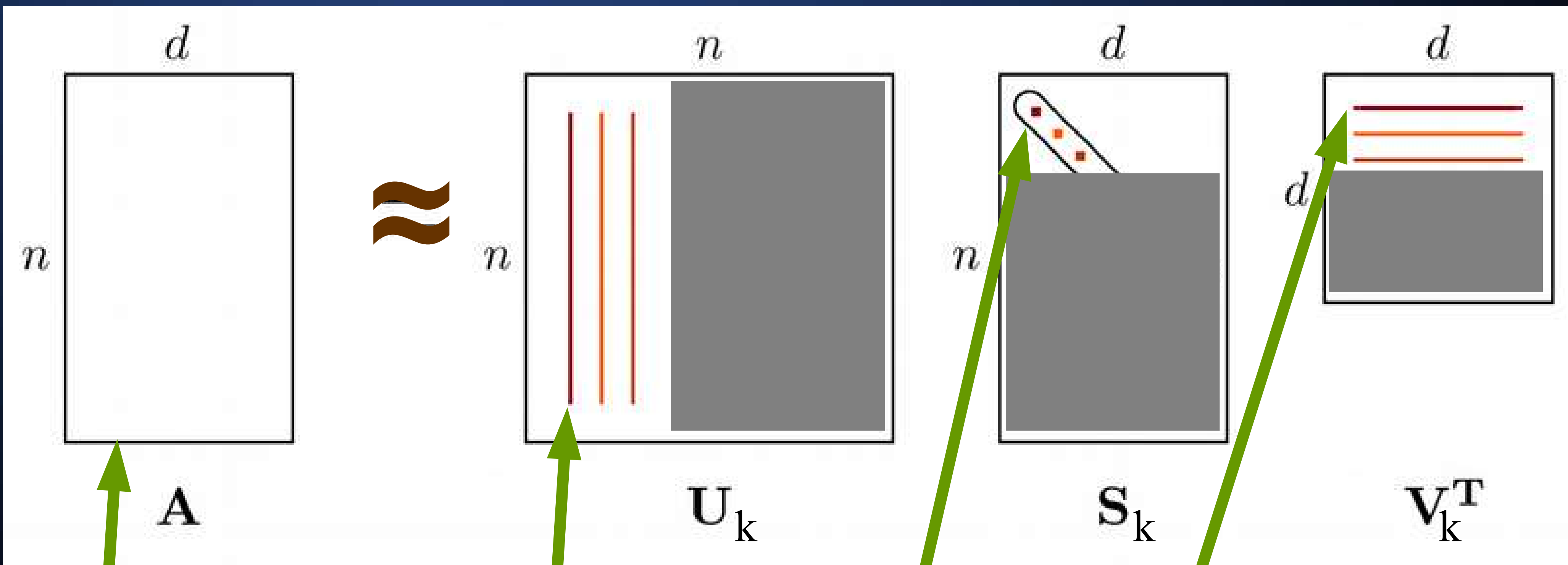


Image: Adapted from Tim Roughgarden & Gregory Valiant

Low-Rank Approximation via SVD



E.g.
Cooccurrence Matrix

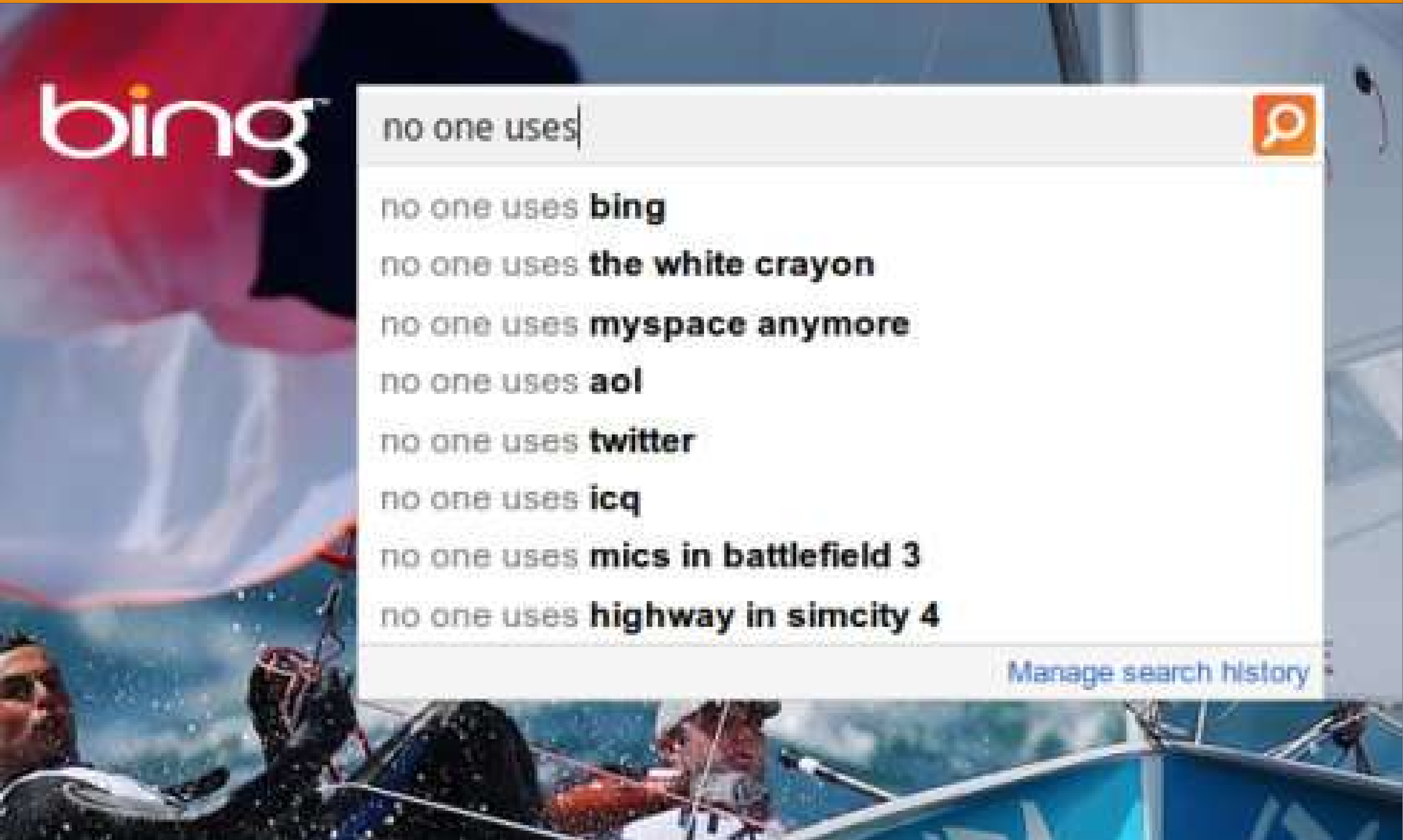
k
Left Singular Vectors

k Singular
Values

k
Right Singular Vectors

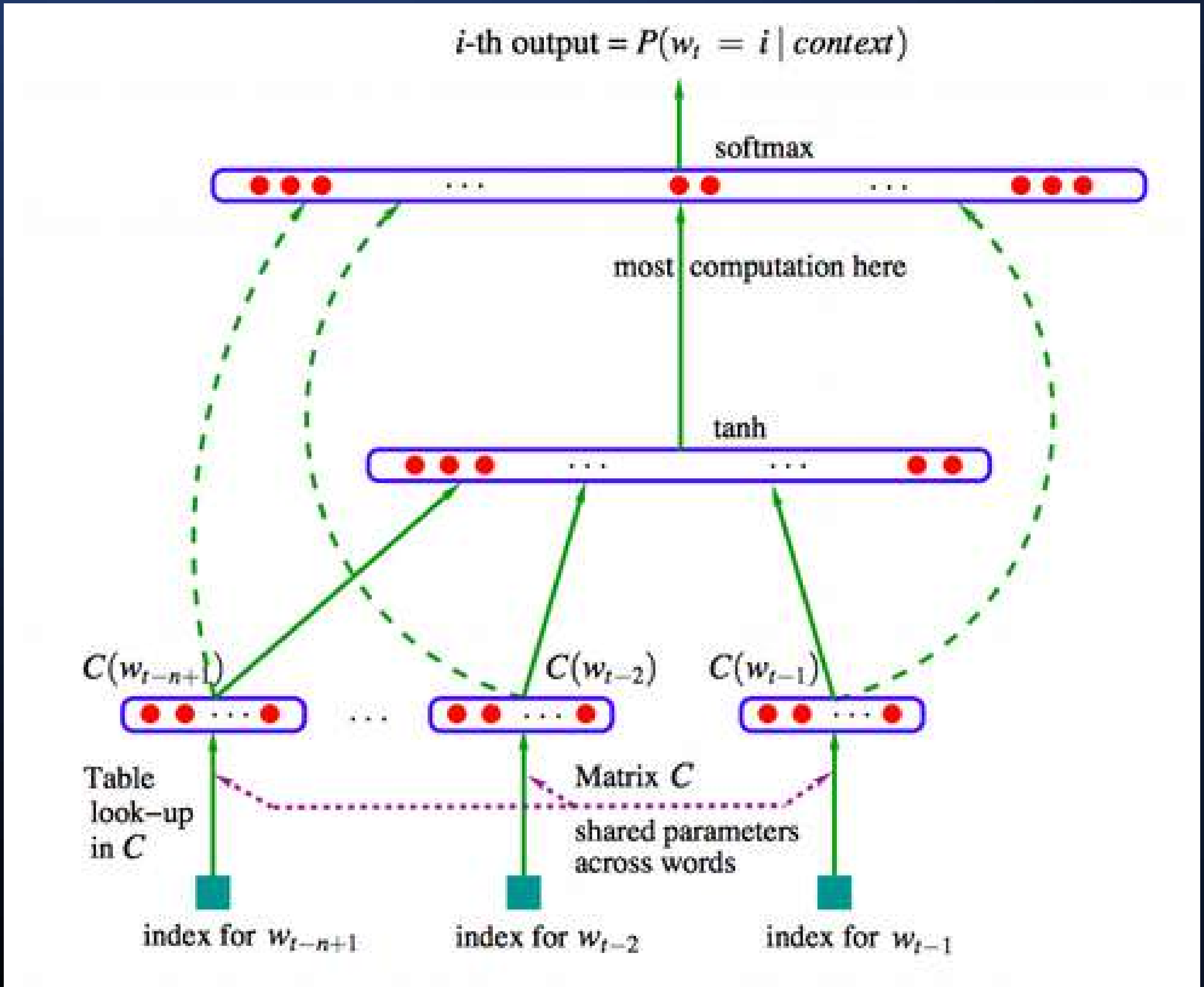
Image: Adapted from Tim Roughgarden & Gregory Valiant

Language Models



Language Models for Neural Word Vectors

Bengio et al. (2003). A Neural Probabilistic Language Model



SENNA Embeddings (2008, 2011)

Collobert et al. (2008, 2011):

Natural Language Processing (Almost) From Scratch

One architecture for part-of-speech tagging, chunking, and named entity recognition

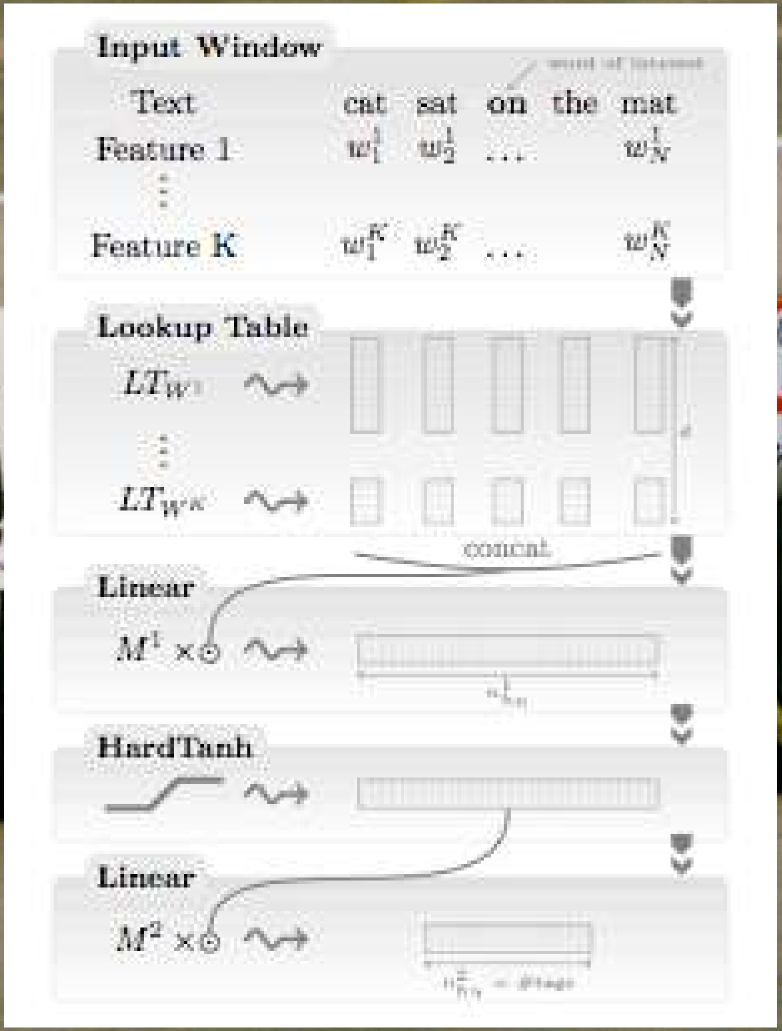
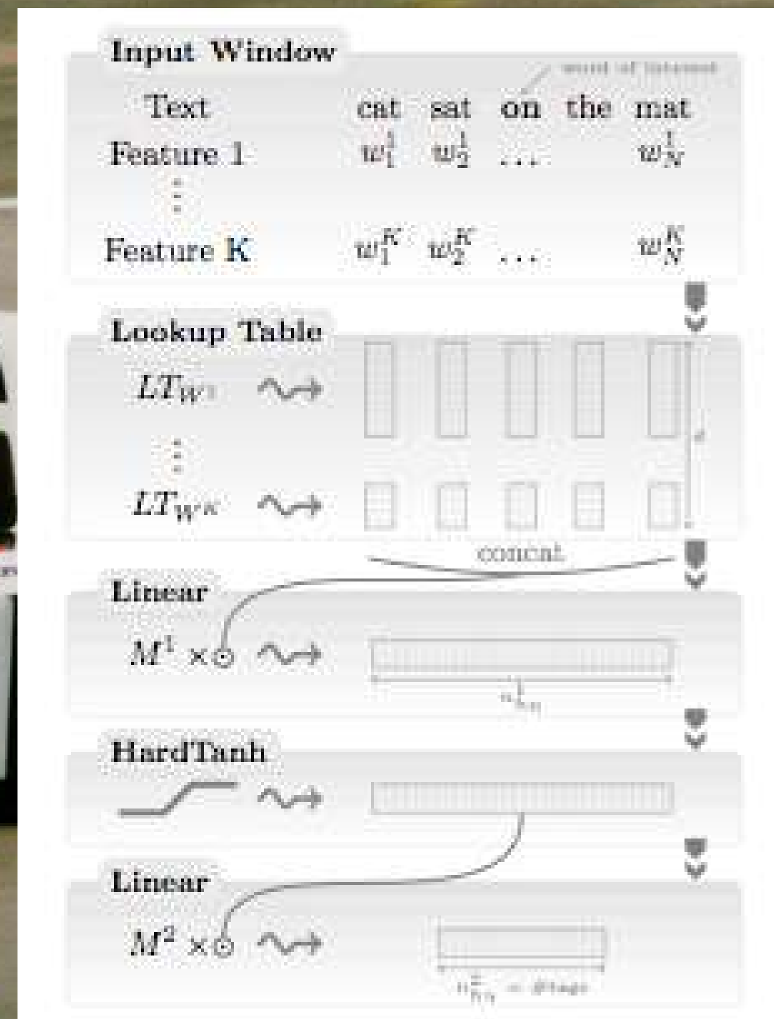


Image: https://commons.wikimedia.org/wiki/File:Senna%27s_McLaren_MP4-8.jpg

SENNA Embeddings (2008, 2011)

Collobert et al.
(2008, 2011):

Natural Language Processing
(Almost) From Scratch



Looked at both sides
of context to train
embeddings.

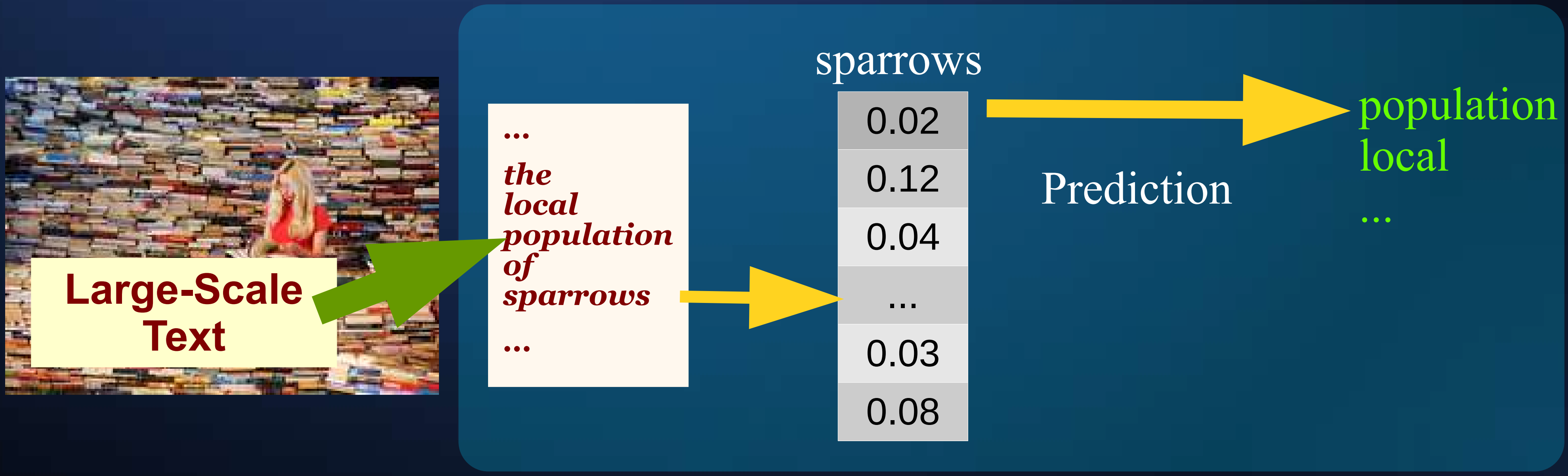
Training Time (2011):
4 weeks + 3 weeks

Image: https://commons.wikimedia.org/wiki/File:Senna%27s_McLaren_MP4-8.jpg

Word Vector Representations: word2vec

word2vec

Tool for computing continuous distributed representations of words.



word2vec Skip-Gram Model

Word Vector Representations: word2vec

word2vec

Tool for computing continuous distributed representations of words.

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec

word2vec

Tool for computing continuous distributed representations of words.

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec

word2vec

Tool for computing continuous distributed representations of words.

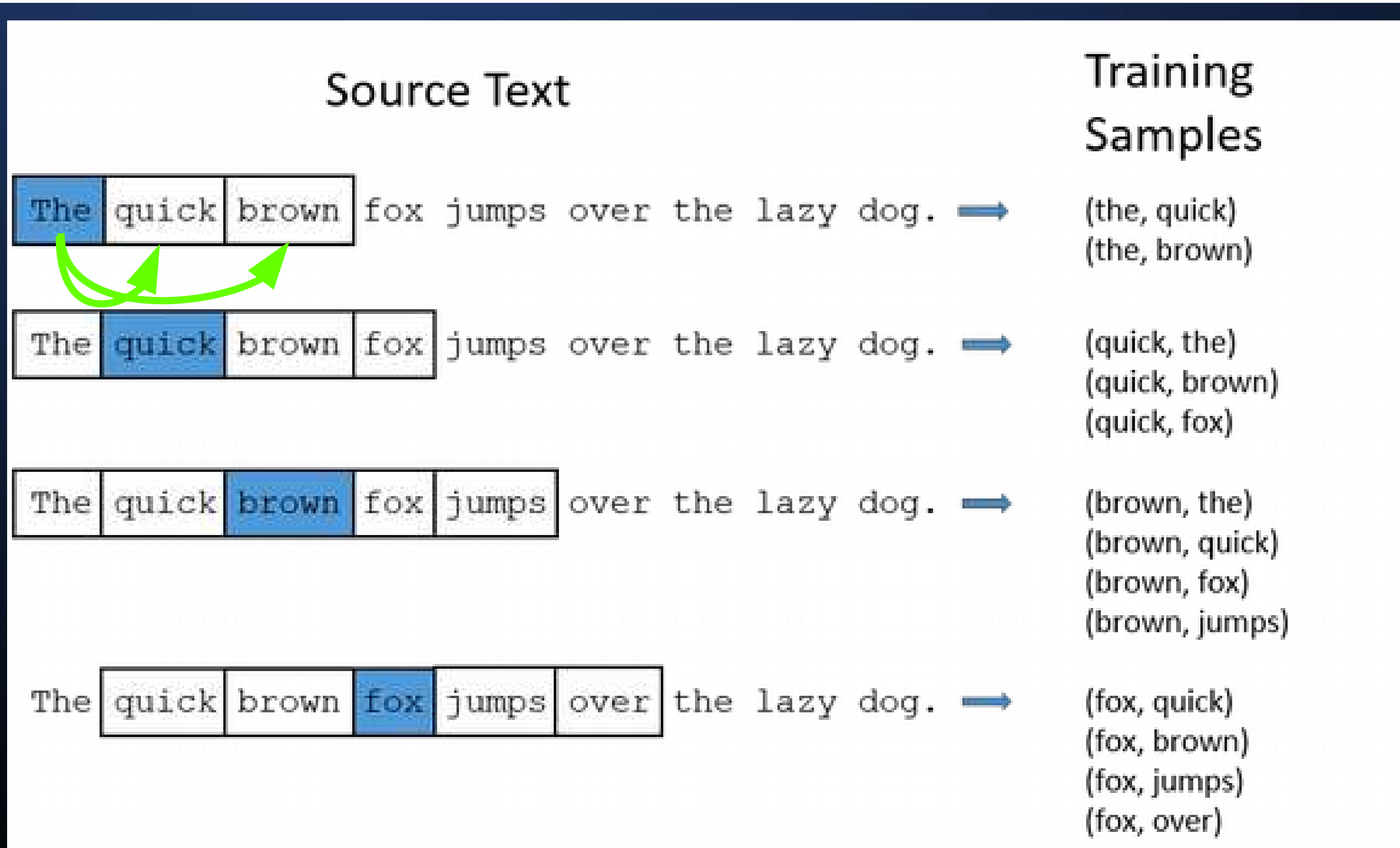
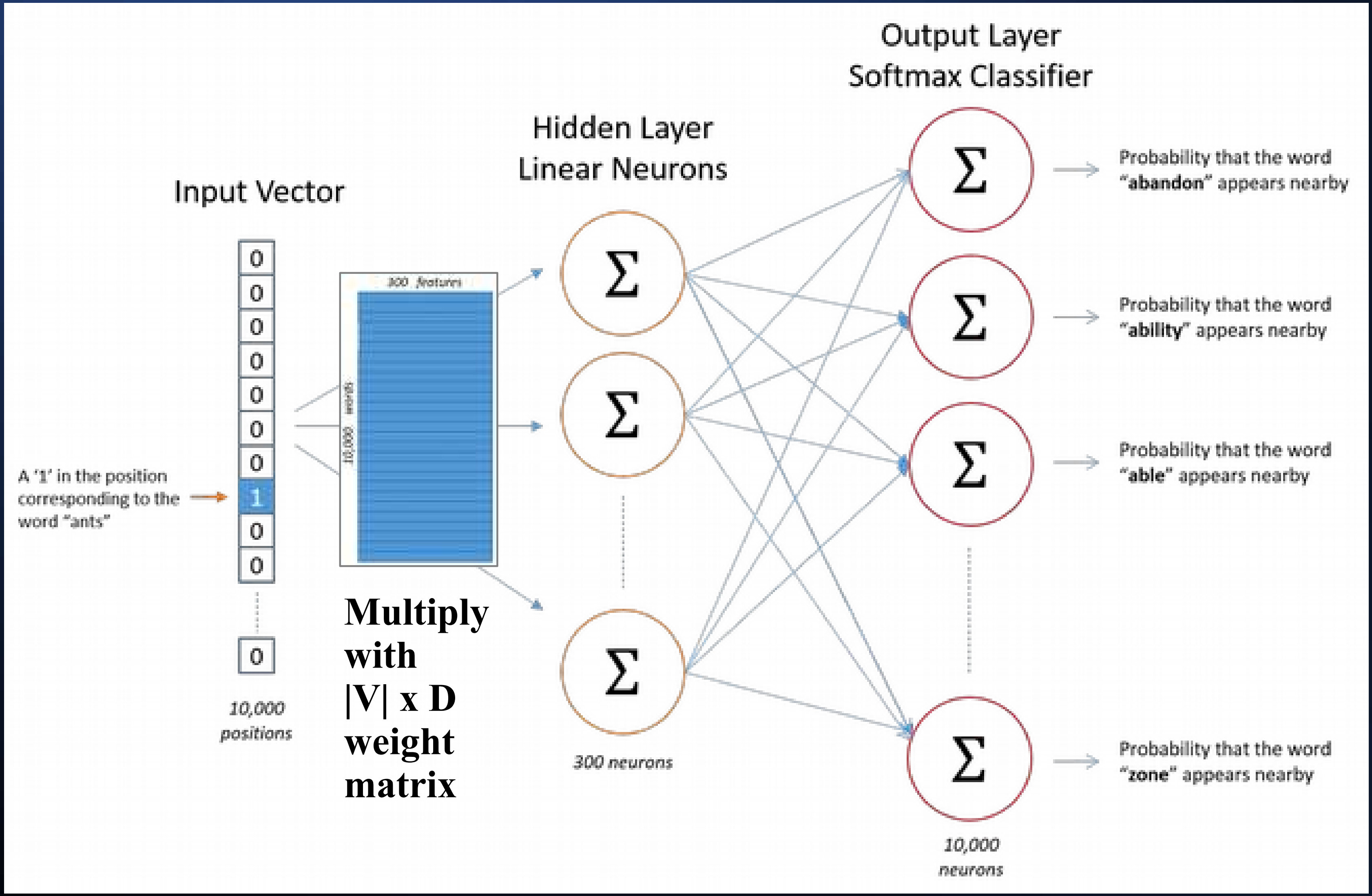


Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

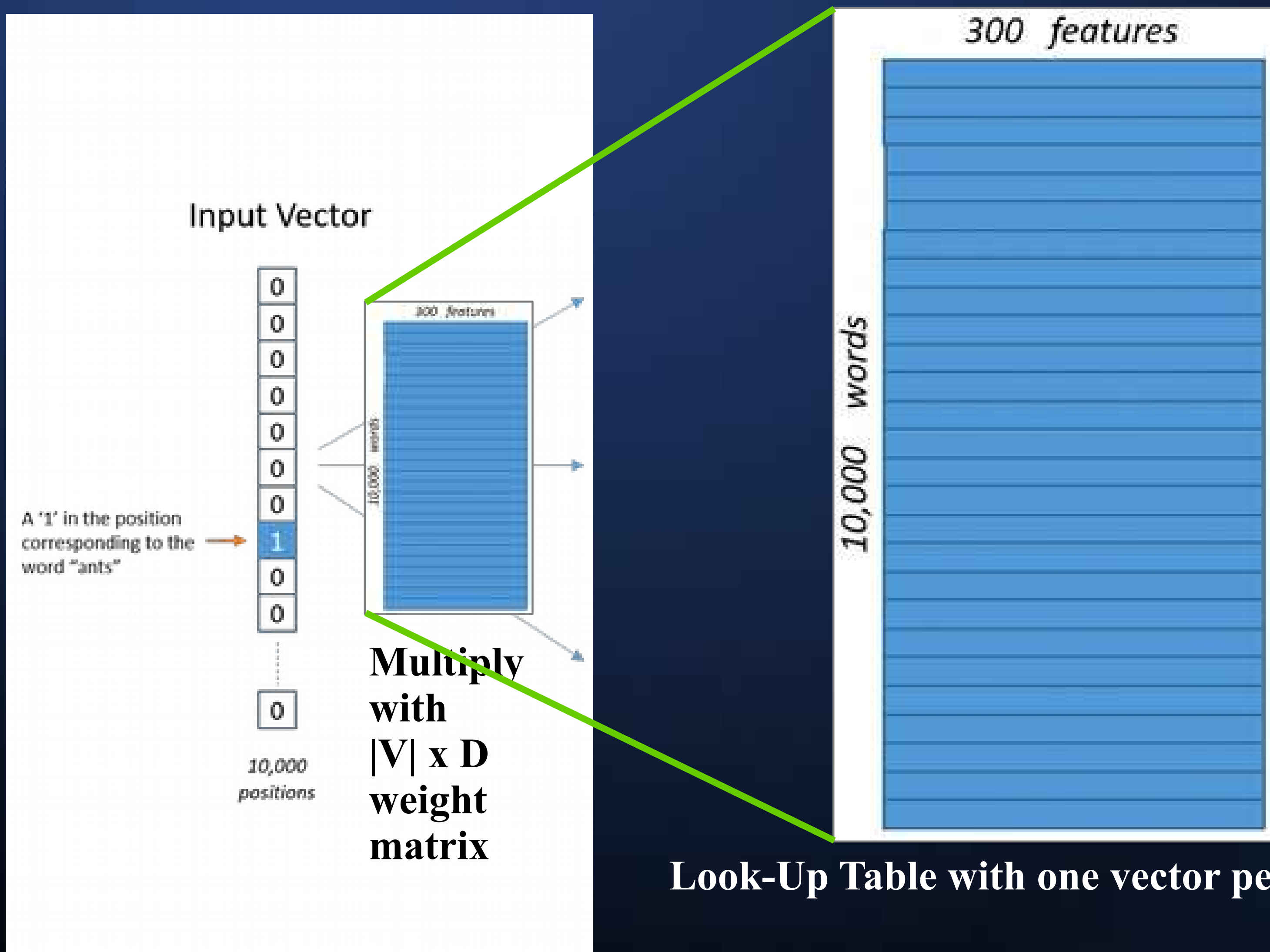
Word Vector Representations: word2vec



Multiply with |V| x D weight matrix

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

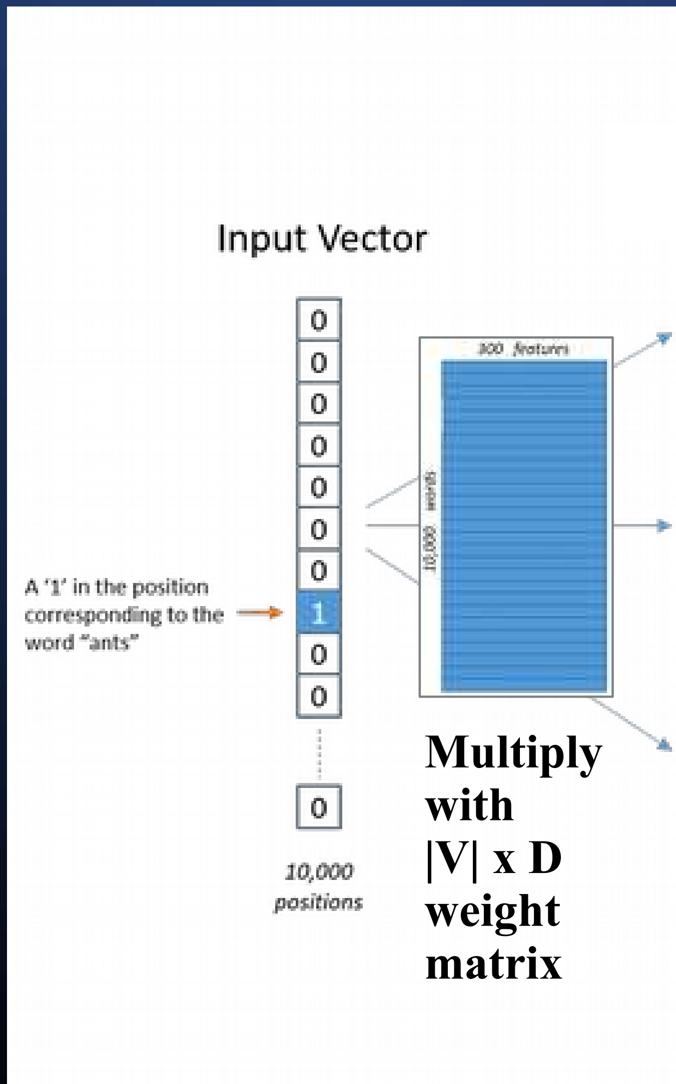
Word Vector Representations: word2vec



Look-Up Table with one vector per word

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec



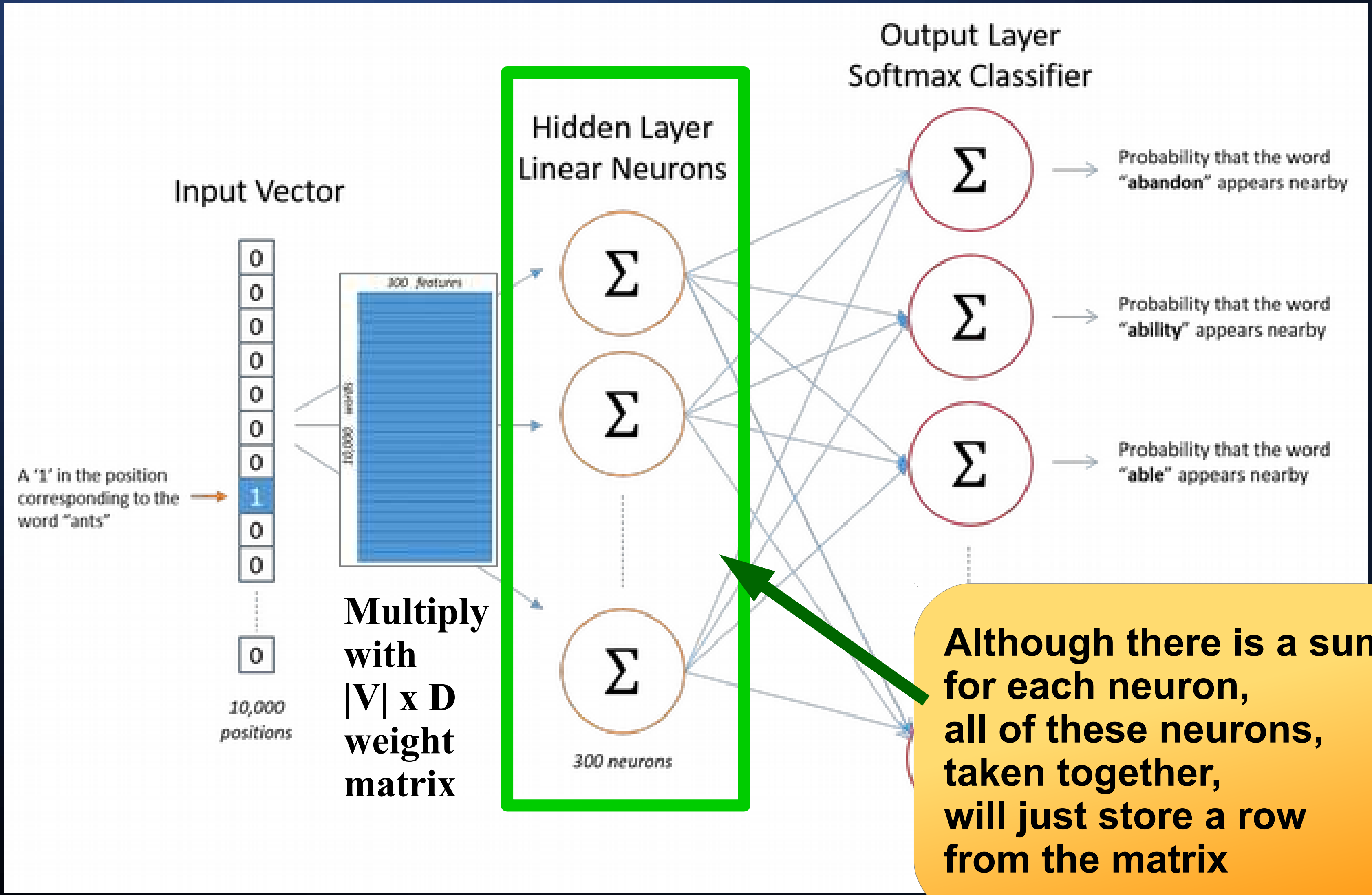
Multiply with |V| x D weight matrix

$$[0 \quad 0 \quad 0 \quad \boxed{1} \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ \boxed{10} & \boxed{12} & \boxed{19} \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

Multiplying a one-hot vector with this matrix will simply look-up a row in the matrix!

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec



Multiply with |V| x D weight matrix

Although there is a sum for each neuron, all of these neurons, taken together, will just store a row from the matrix

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec

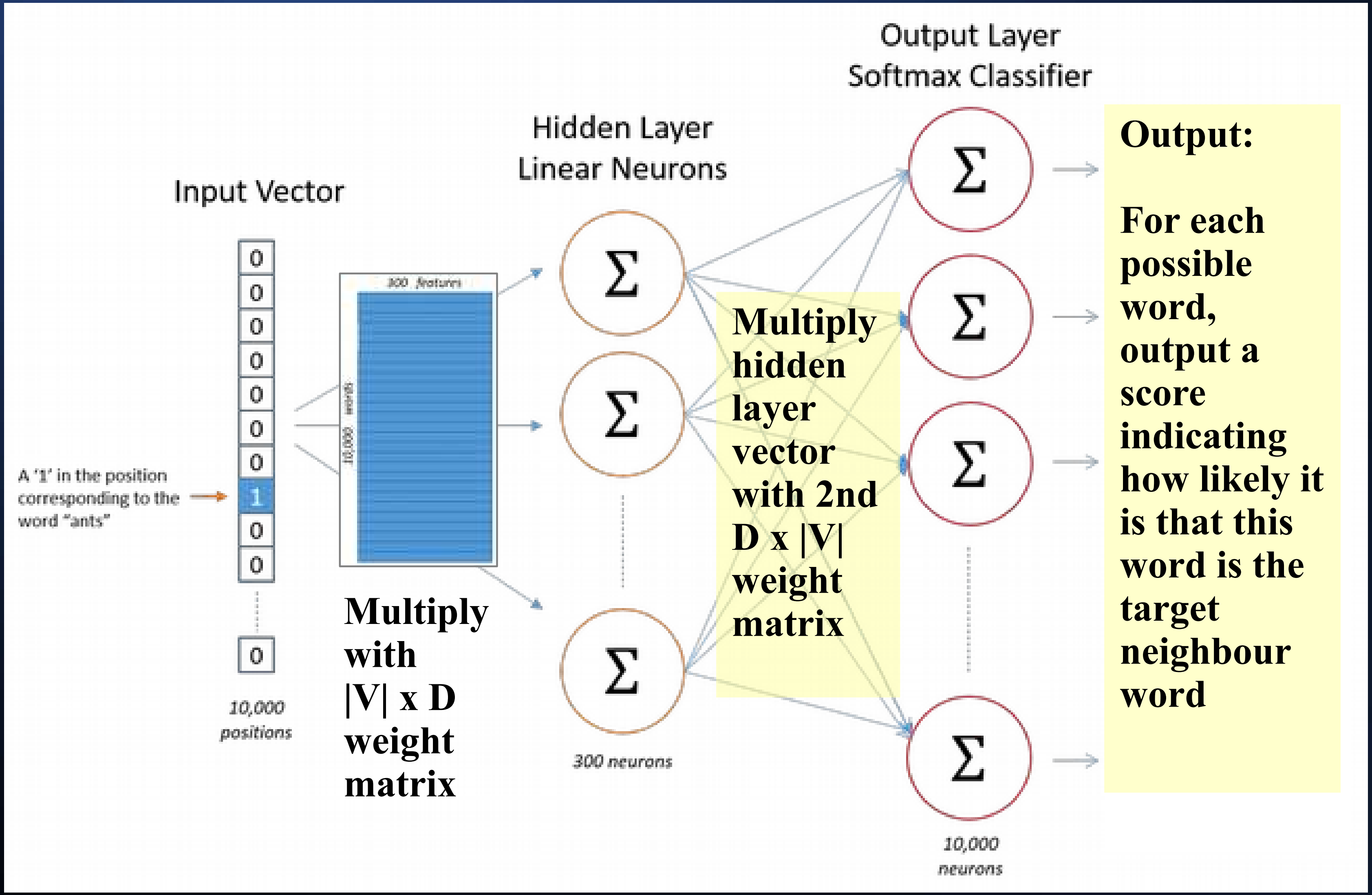
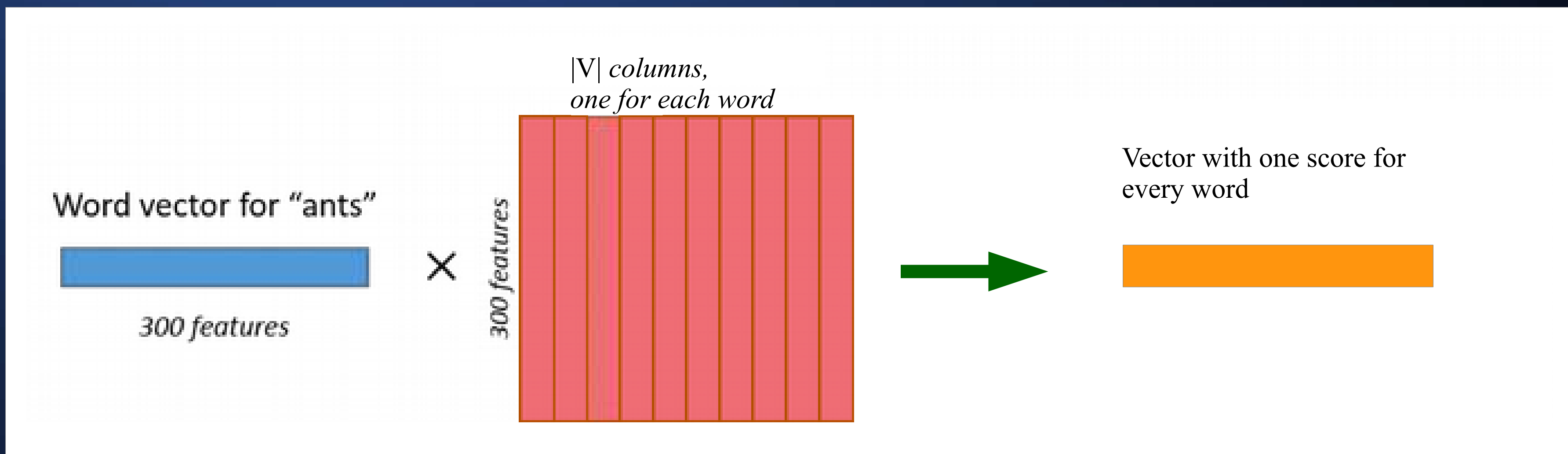


Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

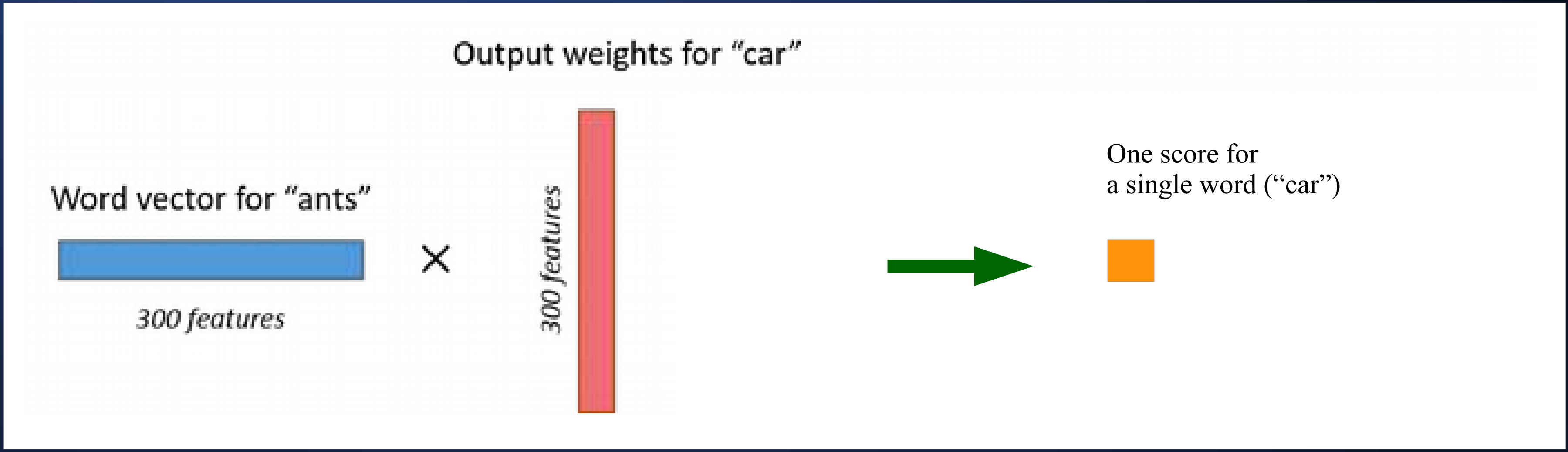
Word Vector Representations: word2vec



"The <u>ant</u> trail at the car ..."

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec



"The ant trail at the car ..."

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec

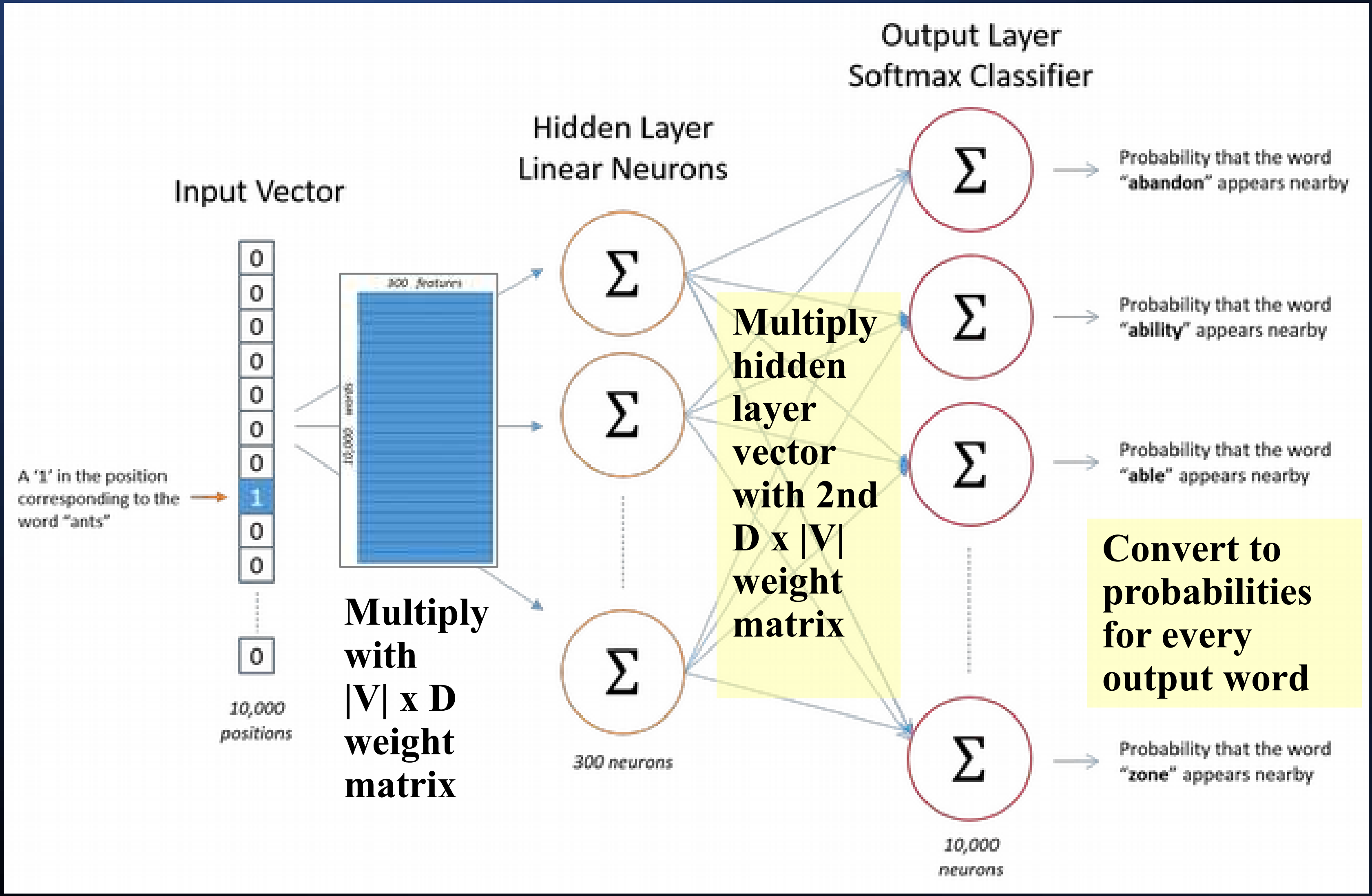
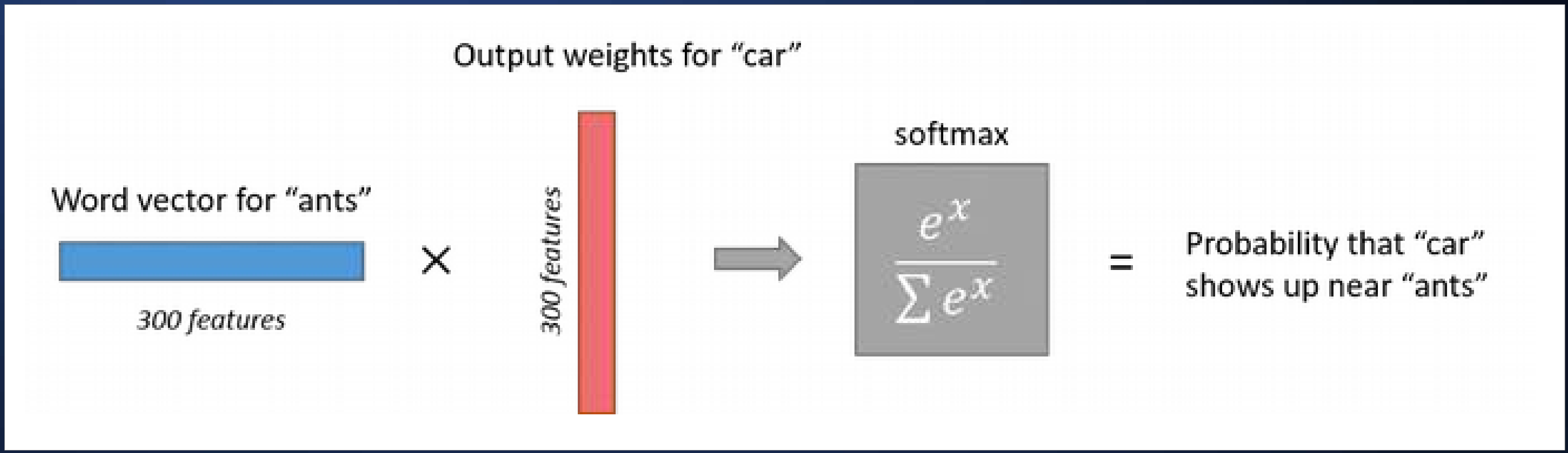


Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec

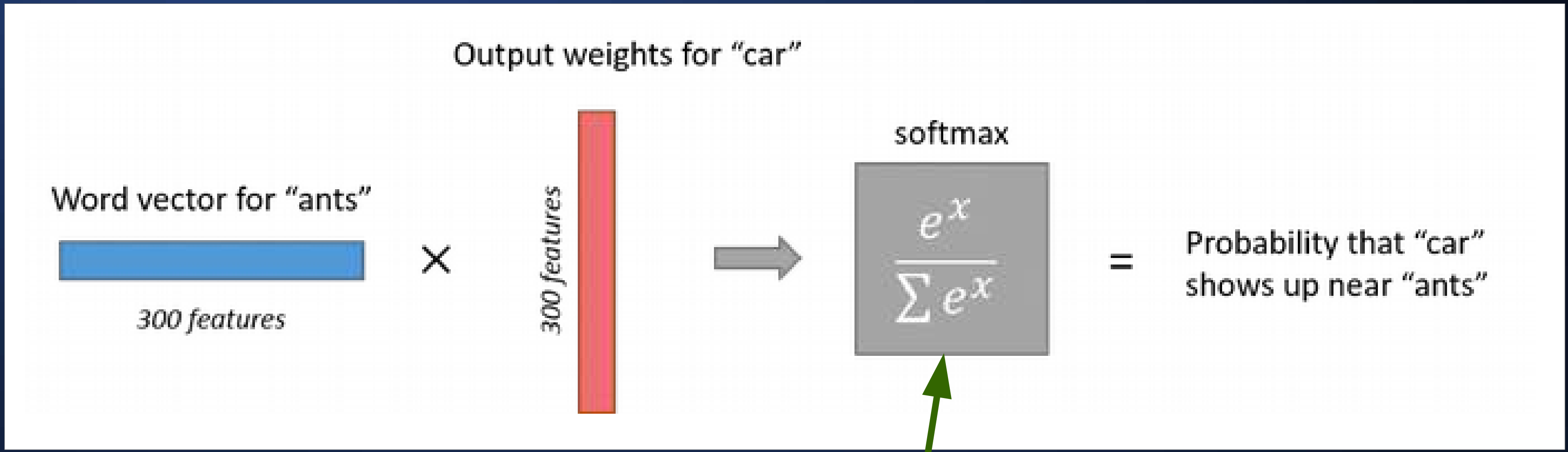


“The <u>ant</u> trail at the car ...”

Probability should be highest for the current target word we are considering.

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec



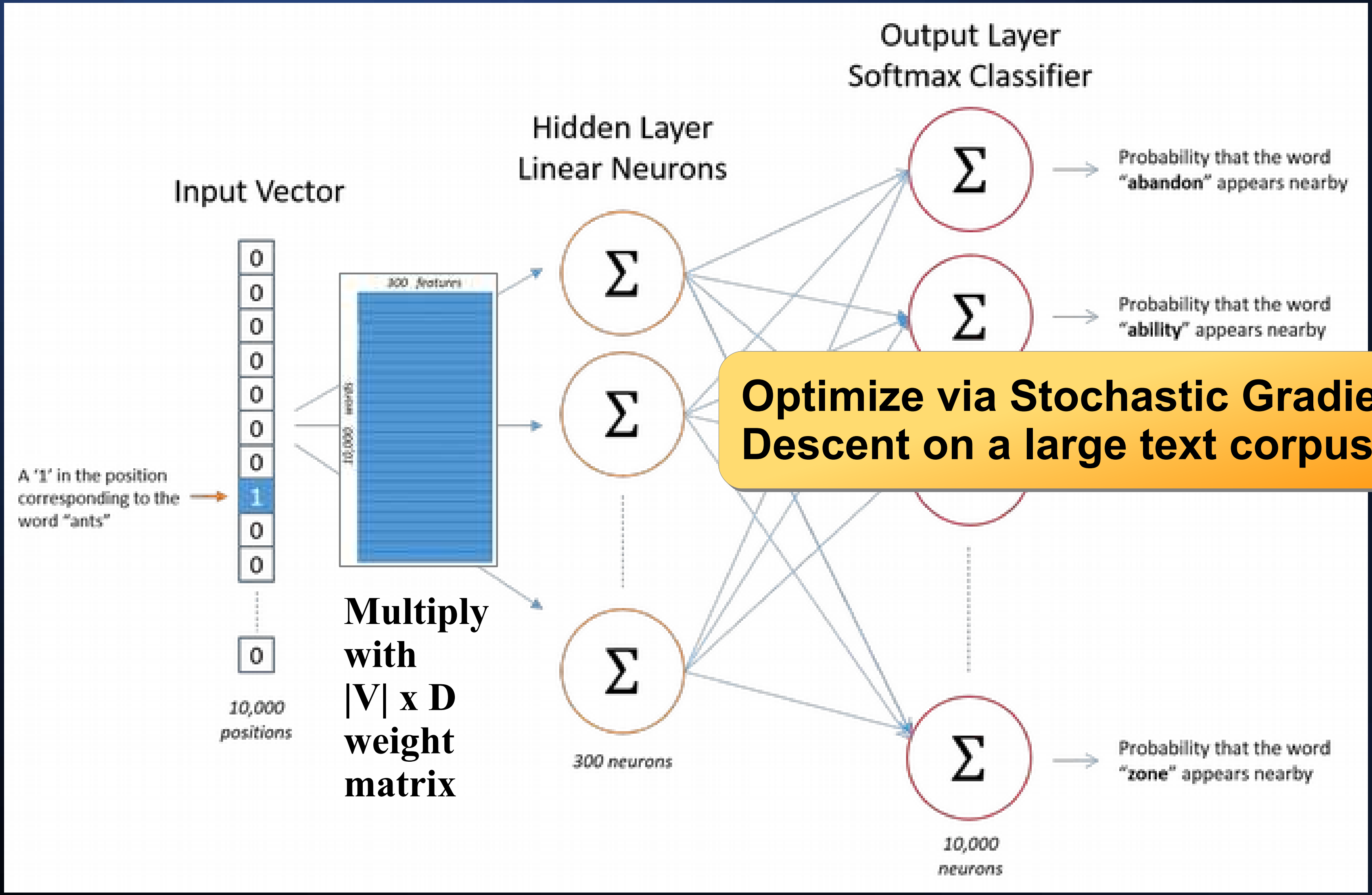
"The ant trail at the car ..."

Probability should be highest for the current target word we are considering.

Optimization: Instead of normalizing over all alternative words, just consider k negative example words in addition to current positive example target word

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec



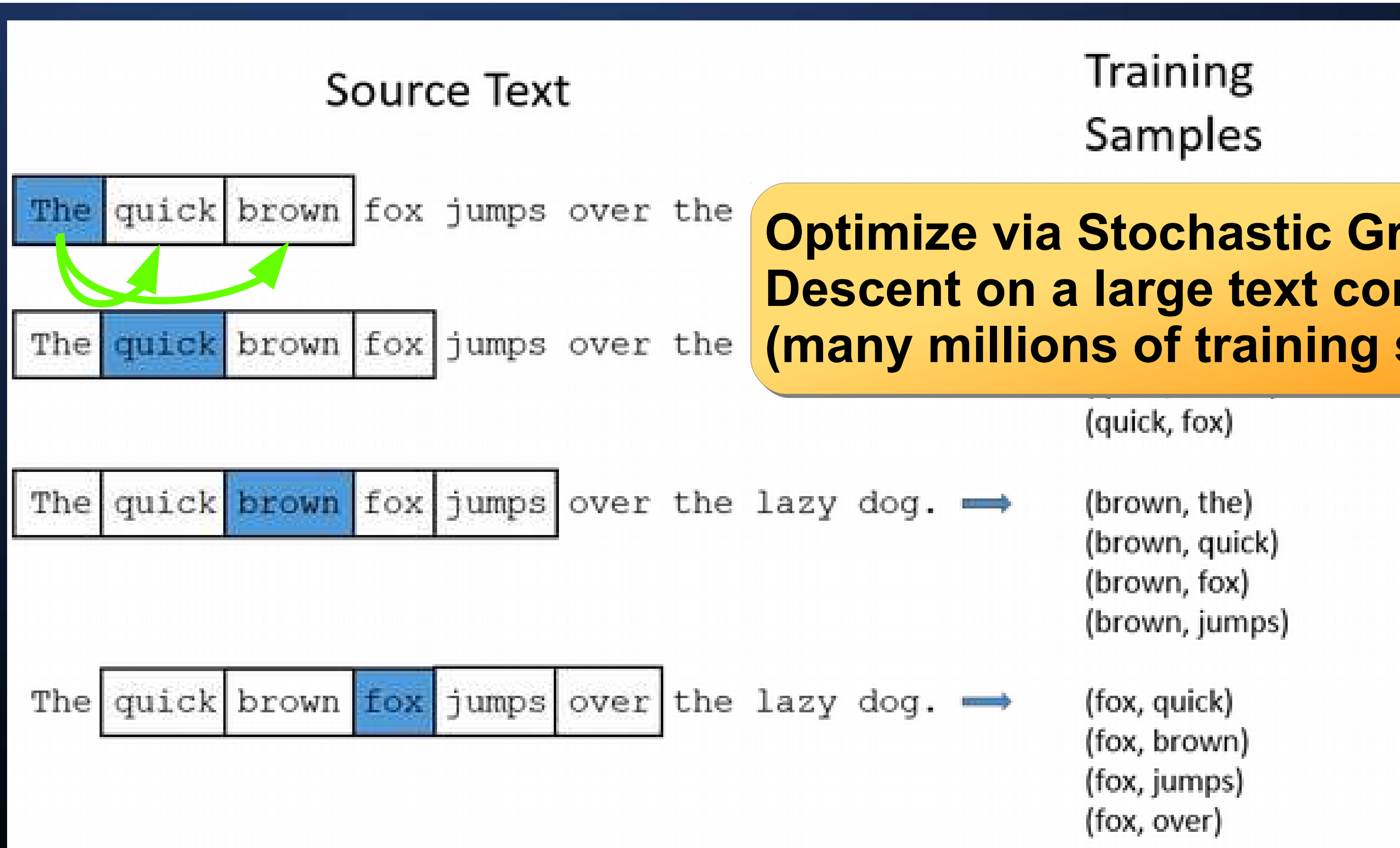
Optimize via Stochastic Gradient Descent on a large text corpus

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec

word2vec

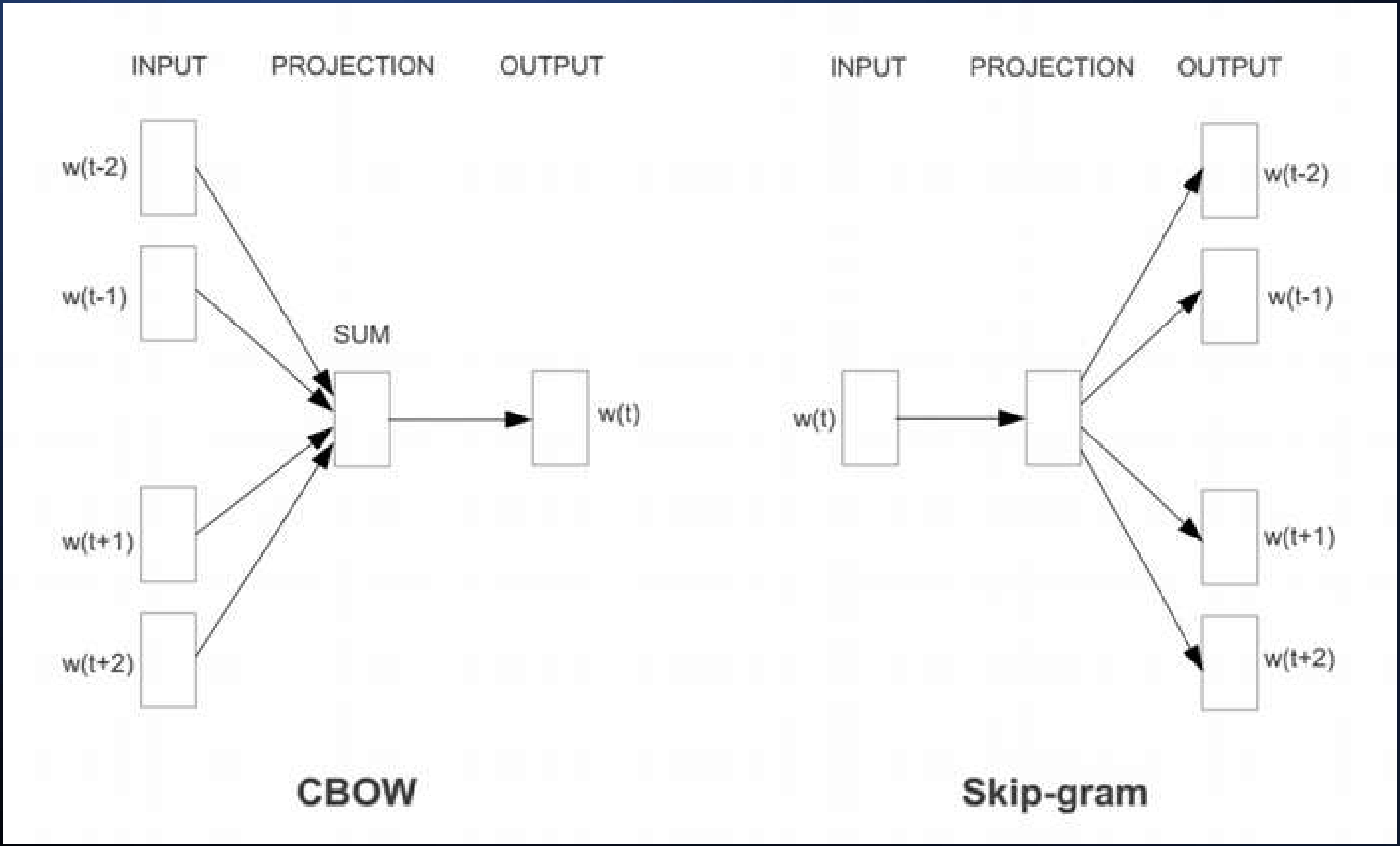
Tool for computing continuous distributed representations of words.



Optimize via Stochastic Gradient Descent on a large text corpus (many millions of training samples)

Image: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Vector Representations: word2vec



Mikolov et al. Distributed Representations of Words and Phrases and their Compositionality. Proc. NIPS 2013

Questions?

Image: Stefan Baudy. https://www.flickr.com/photos/-bast-/349497988

Rare Words

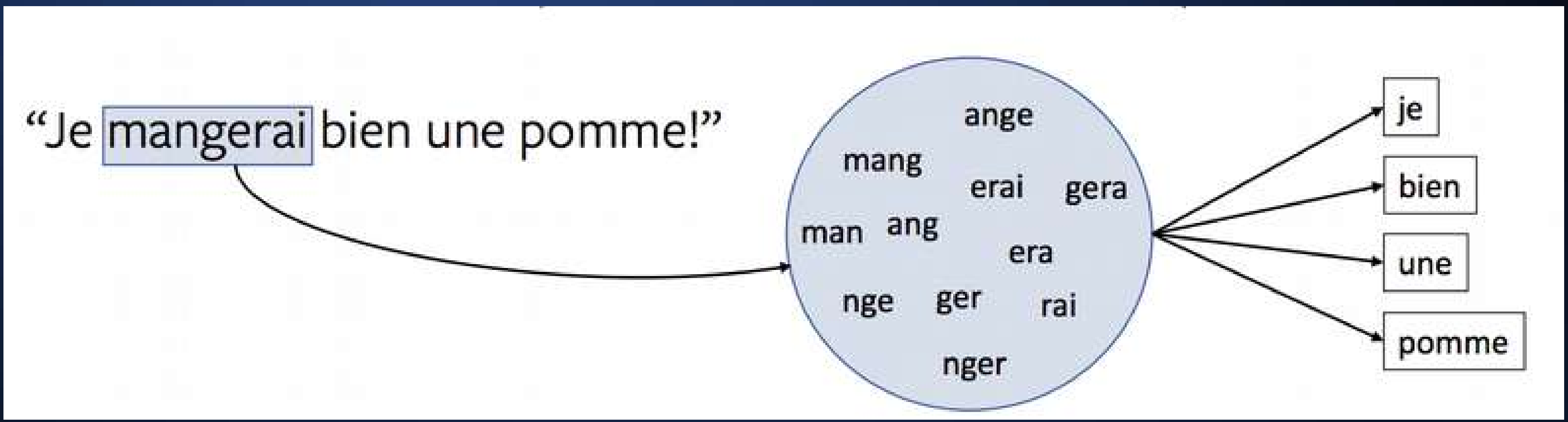
Case 1: Rare Forms

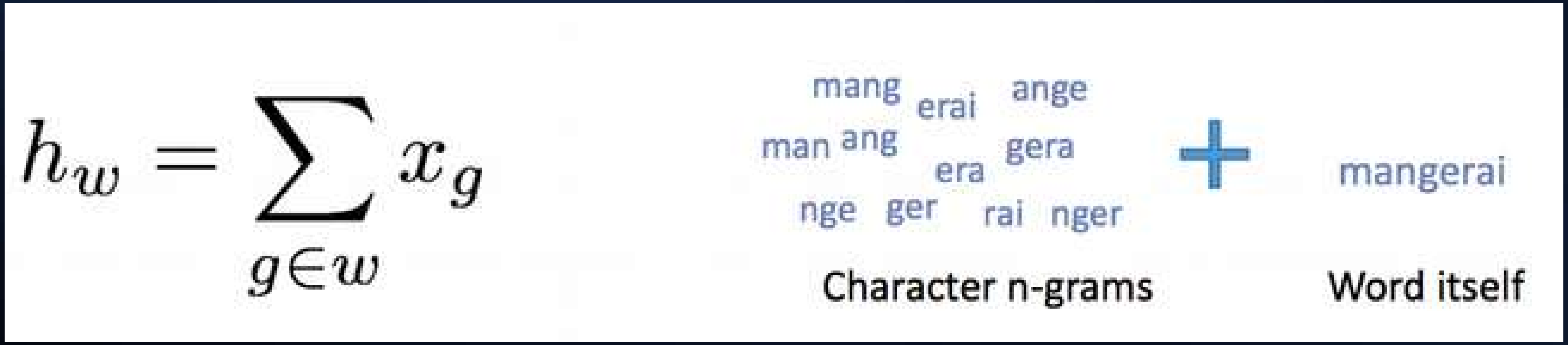
	Singular	Plural
Nominative	uniwersytet	uniwersytety
Genetive	uniwersytetu	uniwersytetów
Dative	uniwersytetowi	uniwersytetom
Accusative	uniwersytet	uniwersytety
Instrumental	uniwersytetem	uniwersytetami
Locative	uniwersytecie	uniwersytetach
Vocative	uniwersytecie	uniwersytety

Polish declension

Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. Enriching word vectors with subword information. TACL 2017

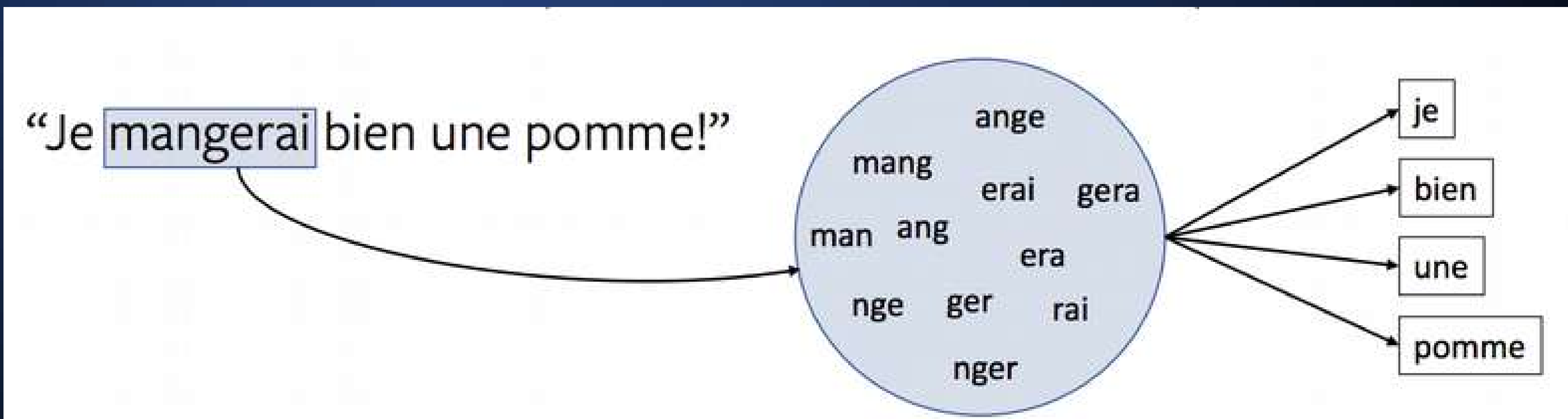
Character N-Gram Approach as in fastText



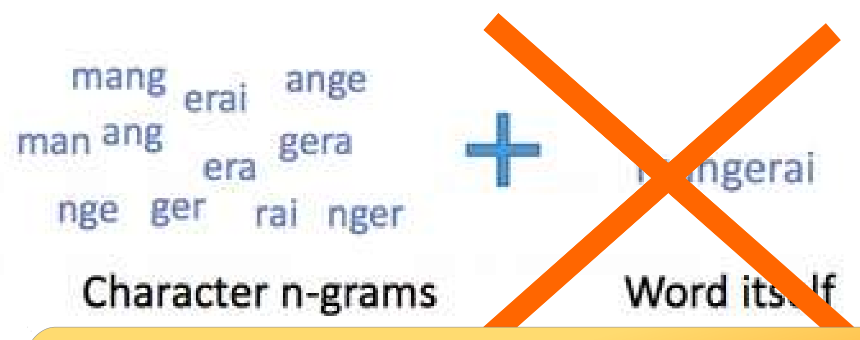


Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. Enriching word vectors with subword information. TACL 2017

Character N-Gram Approach as in fastText



$$h_w = \sum_{g \in w} x_g$$



Can also omit word itself and support out-of-vocabulary forms

Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. Enriching word vectors with subword information. TACL 2017

Character N-Gram Approach as in fastText

fastText

a library for efficient text classification and word representation

fastText is not the first work to consider morphology, but one of the easiest tools to use

Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. Enriching word vectors with subword information. TACL 2017

Case 2: Rare Words/Names

Does anyone know what

“Mangalia”

is?

Case 2: Rare Words/Names

Mangalia:
Romanian port city

Image: https://commons.wikimedia.org/wiki/File:Port_mangalia.jpeg

Zipf's Law

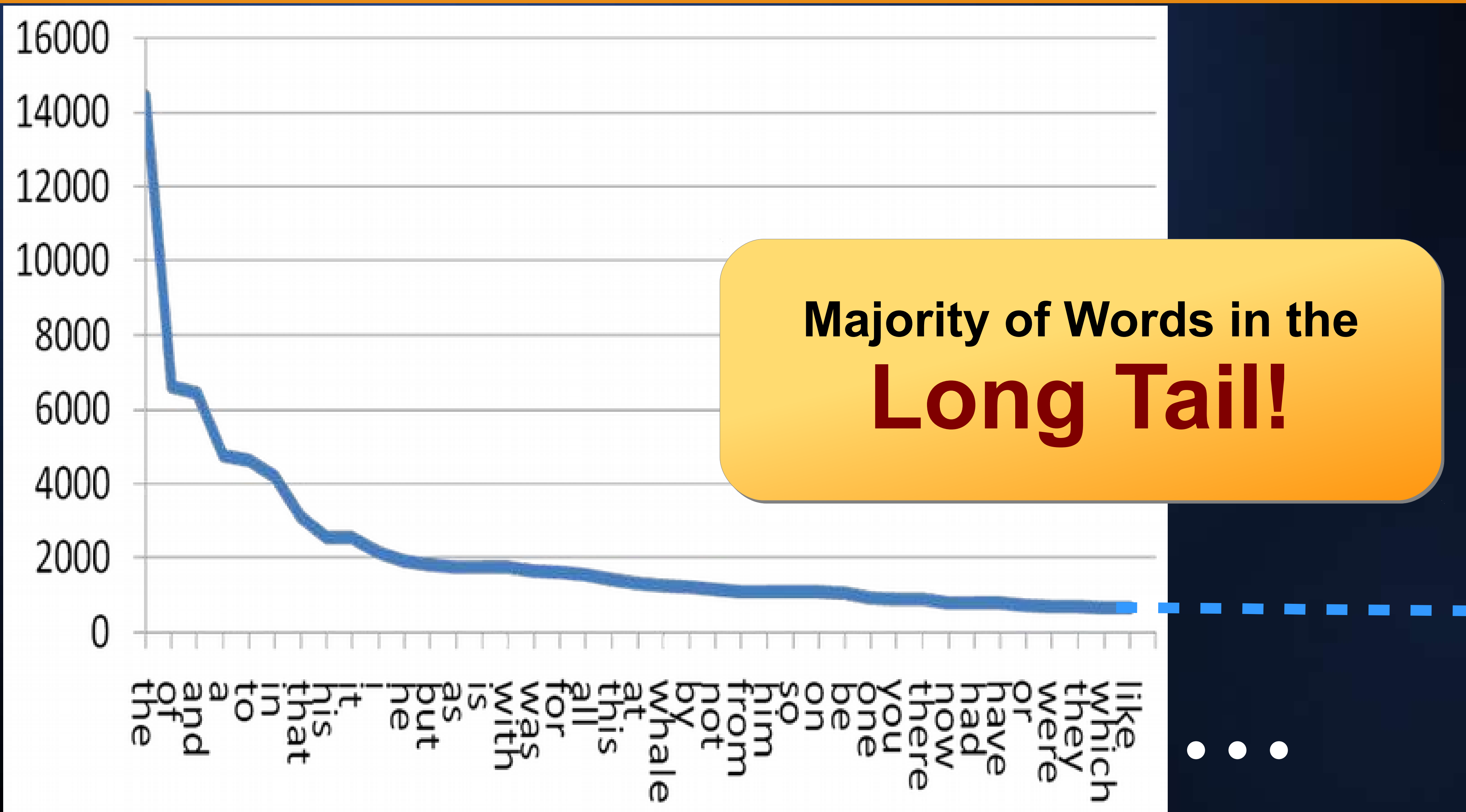


Image: https://www.watzthis.com/articles/zipfs-law-and-seo/

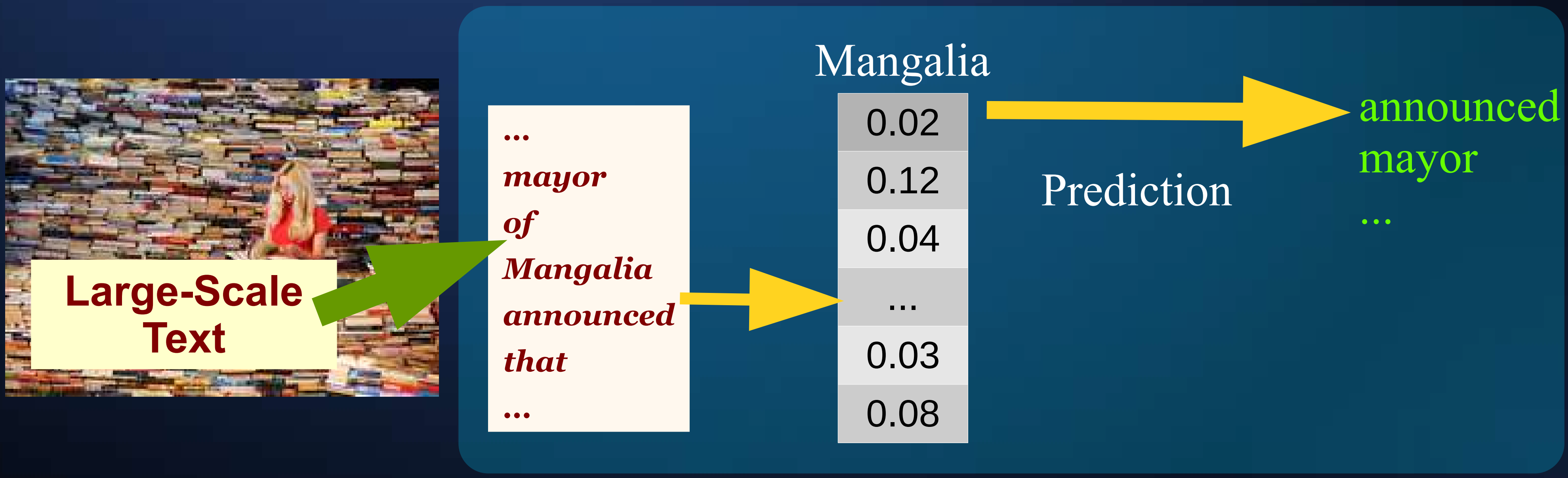
Out-of-Vocabulary Problem

The mayor of ***UNK*** traveled to ***UNK***.

Word Vector Representations: word2vec

word2vec

Tool for computing continuous distributed representations of words.



word2vec Skip-Gram Model

Word Vector Representations: word2vec

word2vec

Tool for computing continuous distributed representations of words.

...
mayor
of
Mangalia
announced
that
...

Mangalia

0.02
0.12
0.04
...
0.03
0.08

Prediction

announced
mayor
...

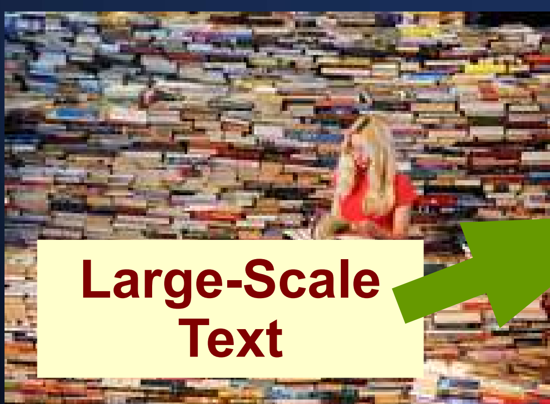
Problem:
Not very discriminative.
E.g. "*of mayor Laszkovic who announced that*"

word2vec Skip-Gram Model

Context-Based Few-Shot Word Representation Learning

Idea 1

1. **Predict n-grams** (e.g. “New York”, “how much”)
2. **Predict position-specific patterns**



...
mayor
of
Mangalia
announced
that
...

Mangalia

0.02
0.12
0.04
...
0.03
0.08

Prediction

announced
mayor
...
mayor of <x>
of <x>
<x> announced
<x> announced that
...

Problem: Combinatorial explosion in no. of prediction targets!

Chenxing Li, Gerui Wang, Gerard de Melo. Context-Based Few-Shot Word Representation Learning

Context-Based Few-Shot Word Representation Learning

<u>Idea 2</u>

Constrain # prediction targets by binning into feature buckets

... mayor of Mangalia announced that ...

Mangalia

0.02
0.12
0.04
...
0.03
0.08

Prediction

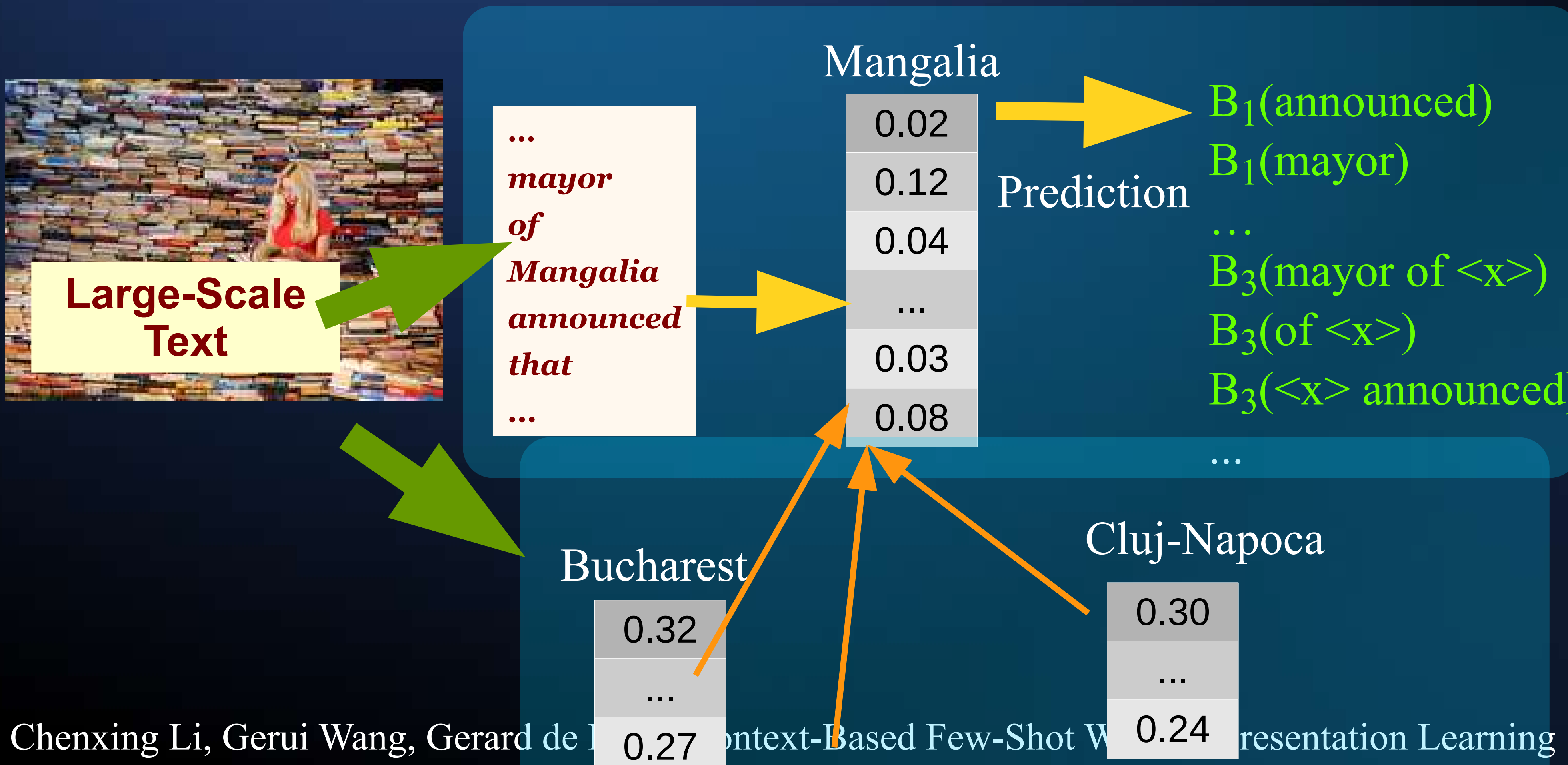
B_1(announced)
B_1(mayor)
…
B_3(mayor of <x>)
B_3(of <x>)
B_3(<x> announced)
...

$$f_i = \{f_0 \mid h(f_0) \bmod |F| = i\}$$

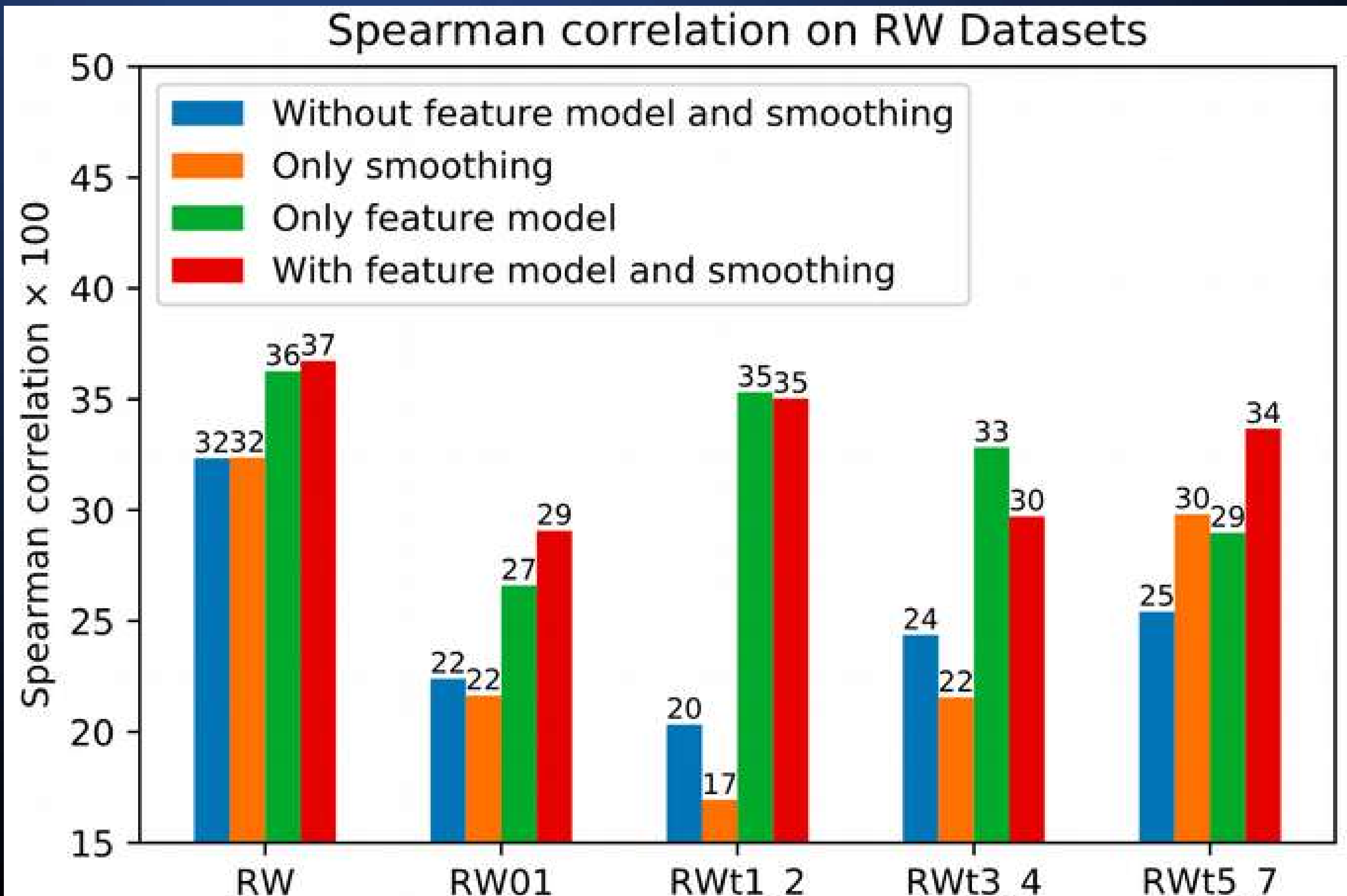
Context-Based Few-Shot Word Representation Learning

Idea 3

Smoothen vectors of rare words via vectors of more frequent words occurring in similar contexts

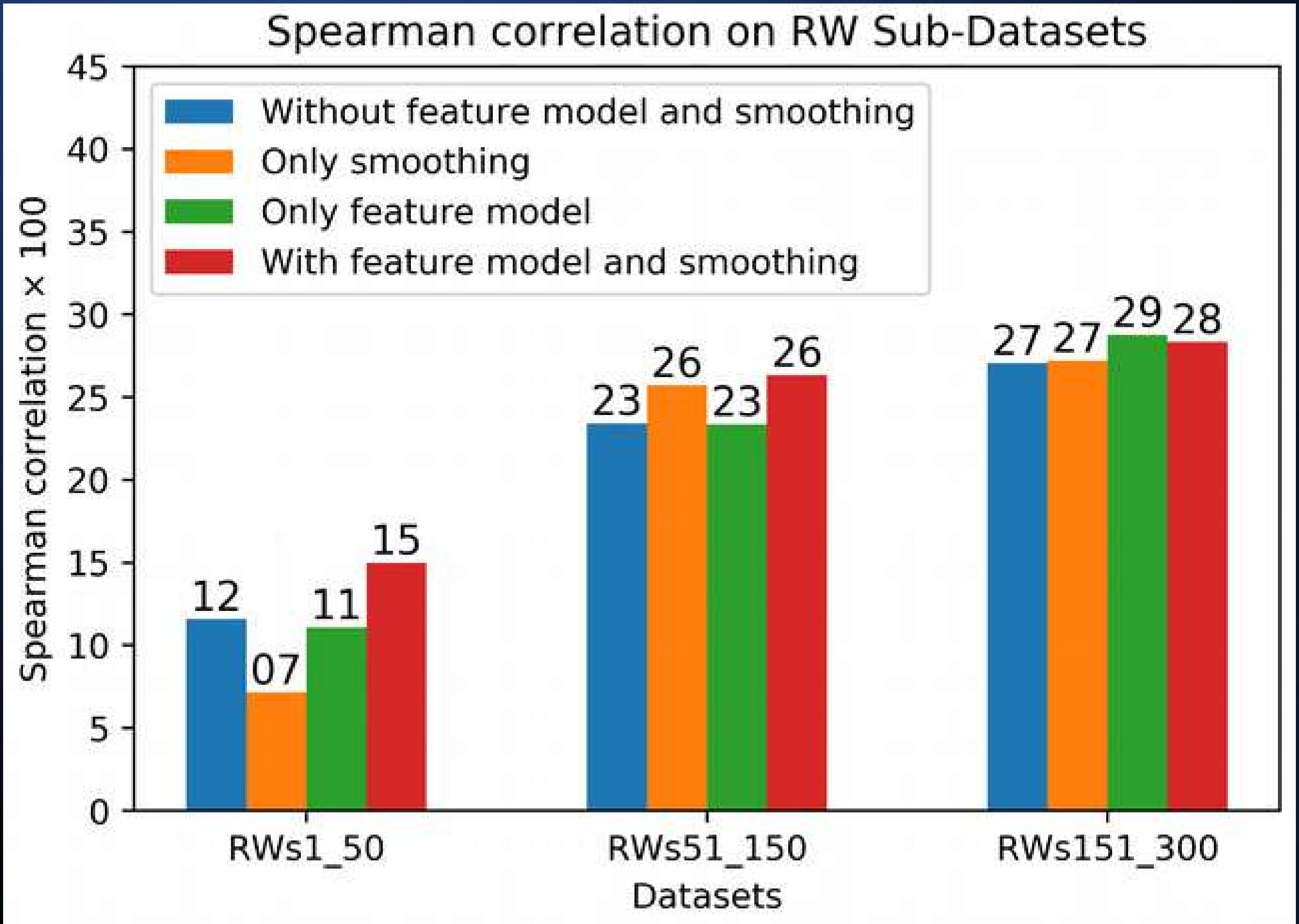


Context-Based Few-Shot Word Representation Learning



Chenxing Li, Gerui Wang, Gerard de Melo. Context-Based Few-Shot Word Representation Learning

Context-Based Few-Shot Word Representation Learning



Chenxing Li, Gerui Wang, Gerard de Melo. Context-Based Few-Shot Word Representation Learning

Questions?

Image: https://www.flickr.com/photos/opensourceway/5556249000