

Learning Word Representations

Vivek Kulkarni



Understanding Textual Content

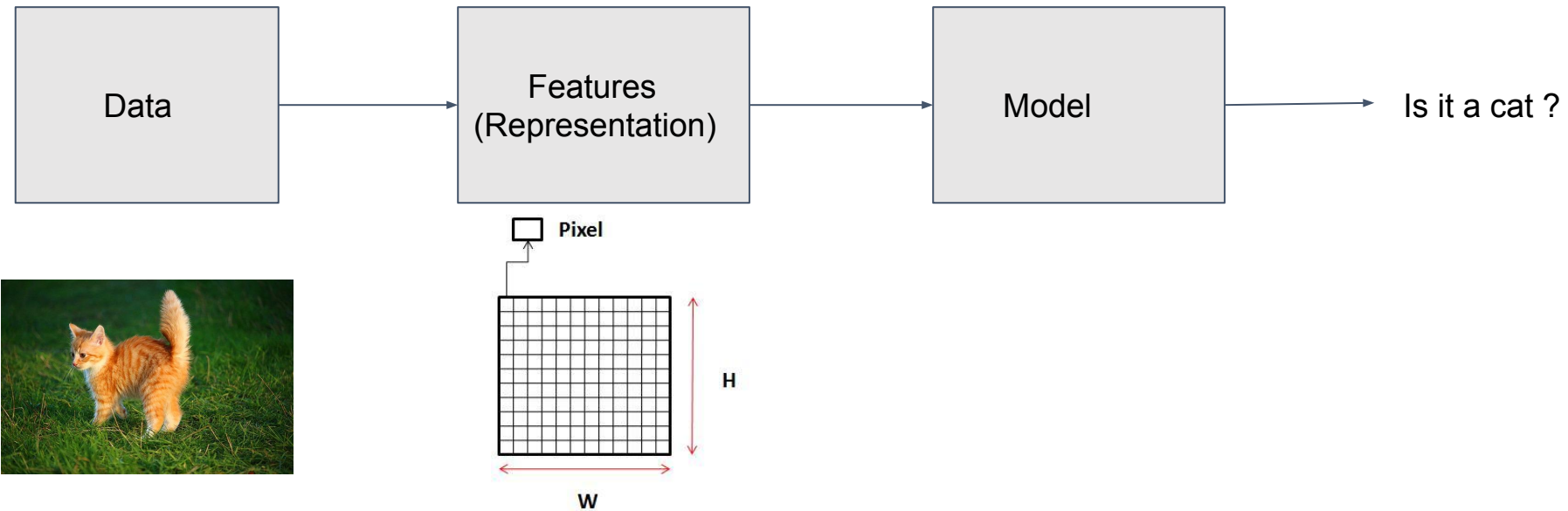
- Computationally analyze textual content to understand text

Sitting here, all alone Watching the snow fall Looking back at the days We threw them snow balls

VERB ADV. DET ADV VERB DET NOUN NOUN VERB ADV ADP DET NOUN PRON VERB PRON ADJ NOUN

- Dominant approach to analyzing/understanding text is Statistical Learning
 - Learn the appropriate input to output transformation from data!
 - Pro: No need to laboriously design complex rule based systems
 - Better Generalization

The learning from data paradigm



Representing Text

- How to represent text?
- Choose what granularity is used for representation
 - Document
 - Sentence/Phrases
 - Words
 - Characters
- Properties of a good representation
 - Useful for the task
 - Allow the model to efficiently use it for the task
 - Bonus: Useful for several tasks and not just a specific task

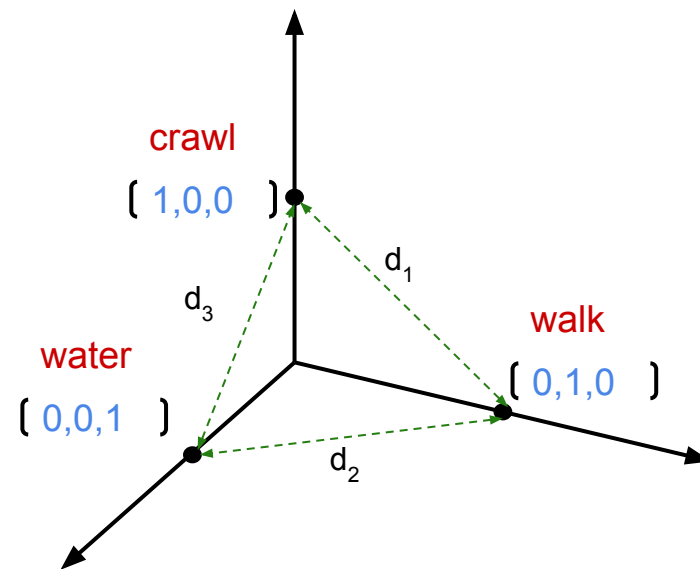
Representing Words

- A 1-hot representation

tiger = [0,0,0,0,0,0,0,1,0,0,0]

lion = [0,0,0,0,0,0,0,1,0,0,0]

- Vector with a single non-zero dimension
- Representation does not capture similarity between words!




Distributional Method

A word is known by the company it keeps – John Rupert Firth

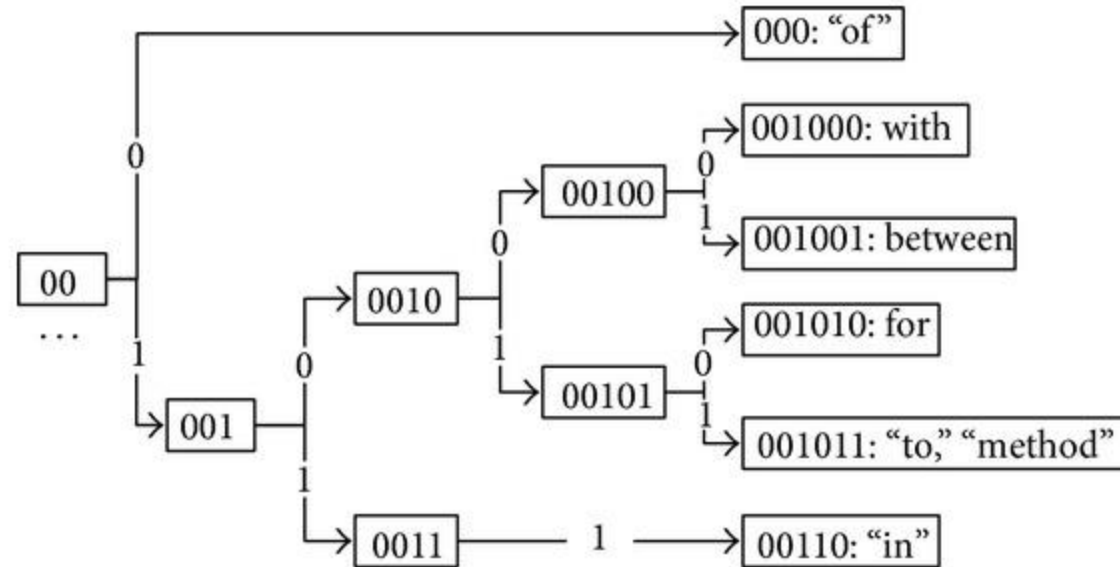
stains open and the moon shining in on the
and the cold , close moon " . And neither o
the night with the moon shining so bright
in the light of the moon . It all boils do
ly under a crescent moon , thrilled by ice
the seasons of the moon ? Home , alone ,
dazzling snow , the moon has risen full an
d the temple of the moon , driving out of

Co-Occurrence Matrix



	planet	night	full	shadow	shine	crescent
moon	10	22	43	16	29	12
sun	14	10	4	15	45	0
dog	0	4	2	10	0	0

Representing Words - Brown Clusters



- Hierarchical clustering of words (based on classes)
- Discrete representation
- Very competitive and popular
- Useful for variety of tasks like NER, POS tagging etc

[Image from: https://www.researchgate.net/figure/261610872_fig1_A-hierarchical-structure-fragment-generated-by-Brown-clustering-for-7-words-from-the]

Distributional Method-Fundamentals

A word is known by the company it keeps – John Rupert Firth

1. I like Deep Learning
2. I enjoy flying
3. I like NLP

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

[Courtesy:Socher]

Distributional Method-Problems with Raw Co-occurrence Matrices

Very high dimensional. Increases with vocabulary size

Less robust models due to data sparsity.

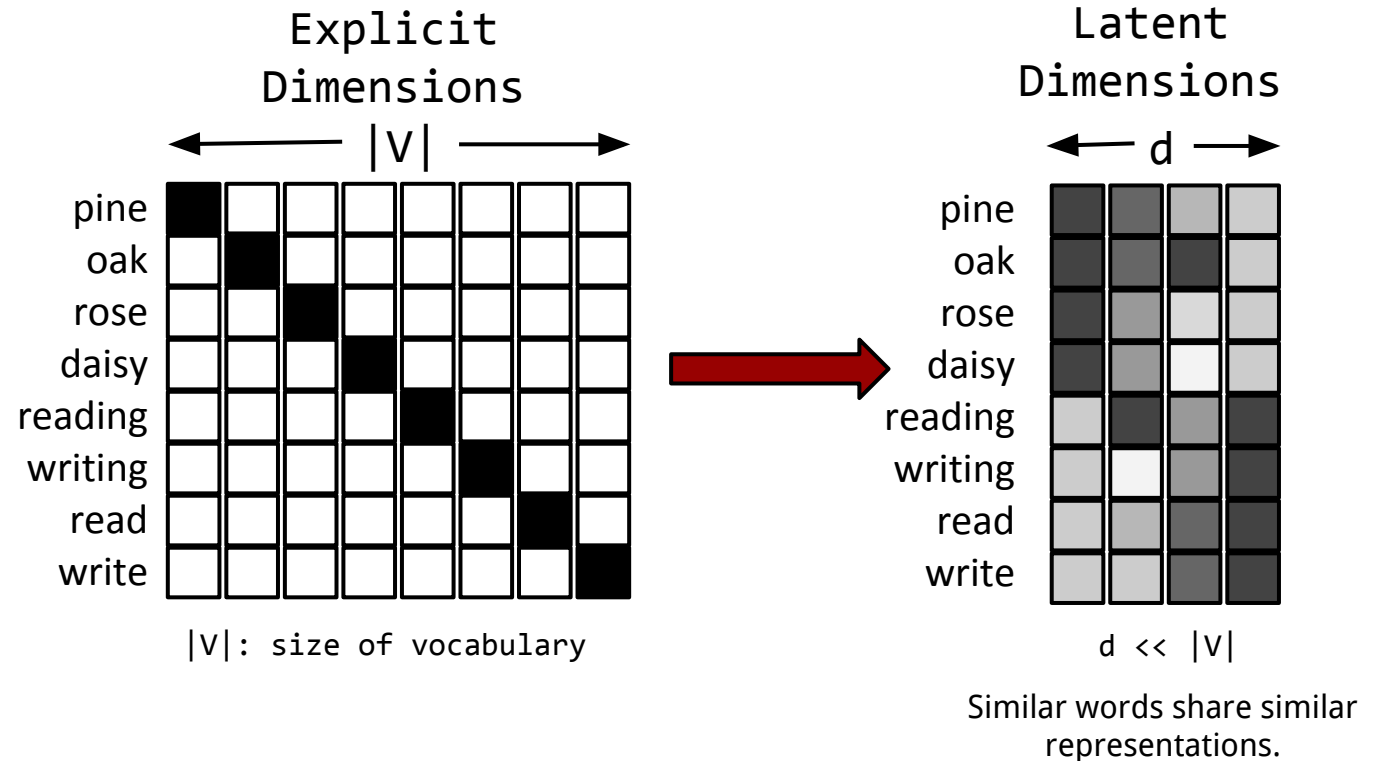
Store important information in a fixed dimension dense vector.

Distributed Word Representations

Word Embeddings are latent representations of words.

We factorize the co-occurrence matrix

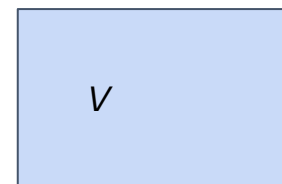
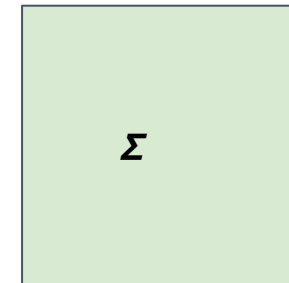
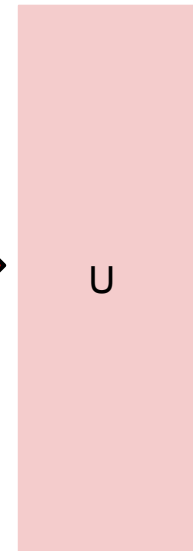
Can be viewed as an online implicit factorizing method and thus scalable



SVD Word Embeddings

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

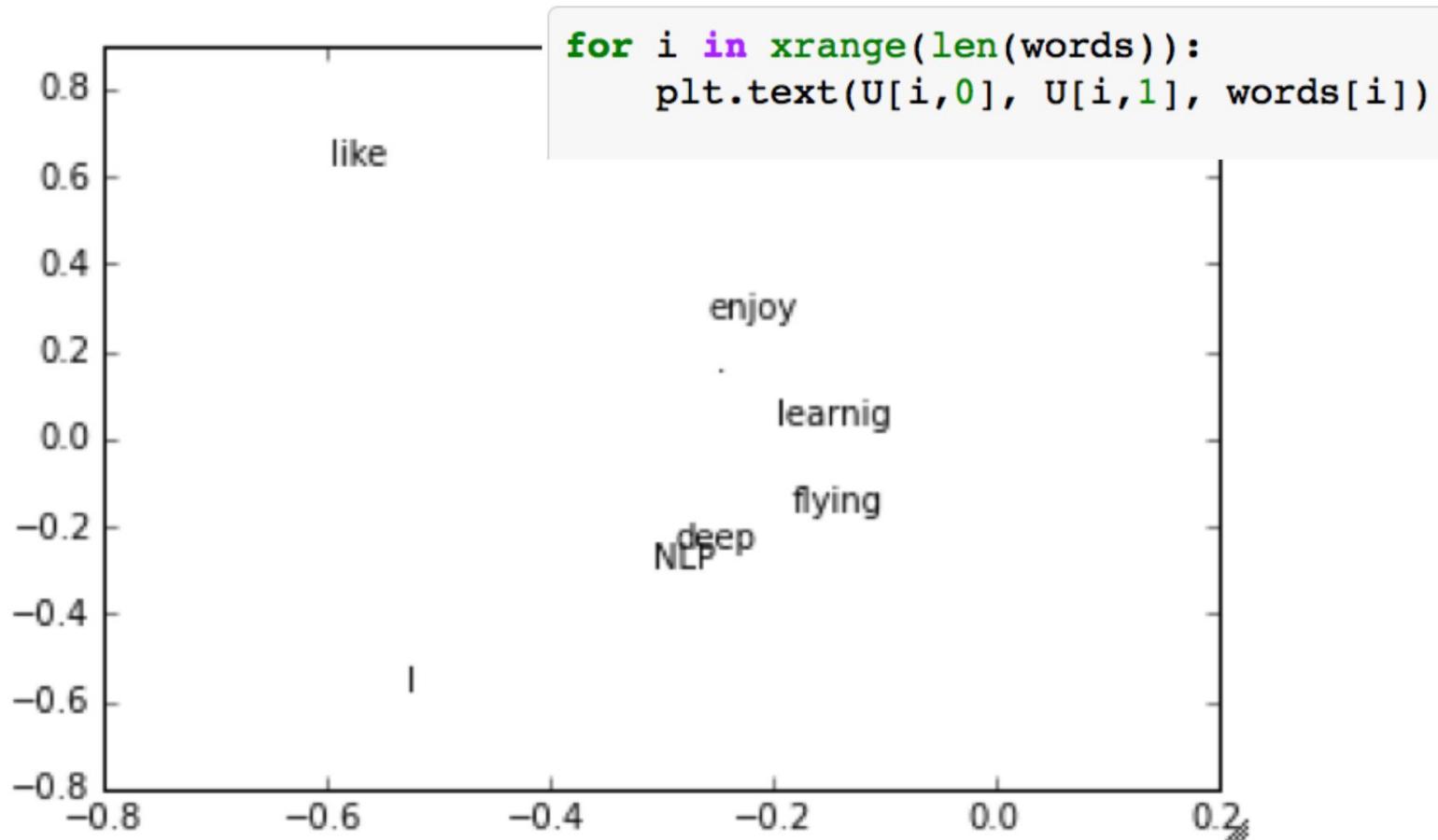
SVD
→



SVD Word Embeddings (Visualization)

Corpus: I like deep learning. I like NLP. I enjoy flying.

Printing first two columns of U corresponding to the 2 biggest singular values



Issues with Word Embeddings

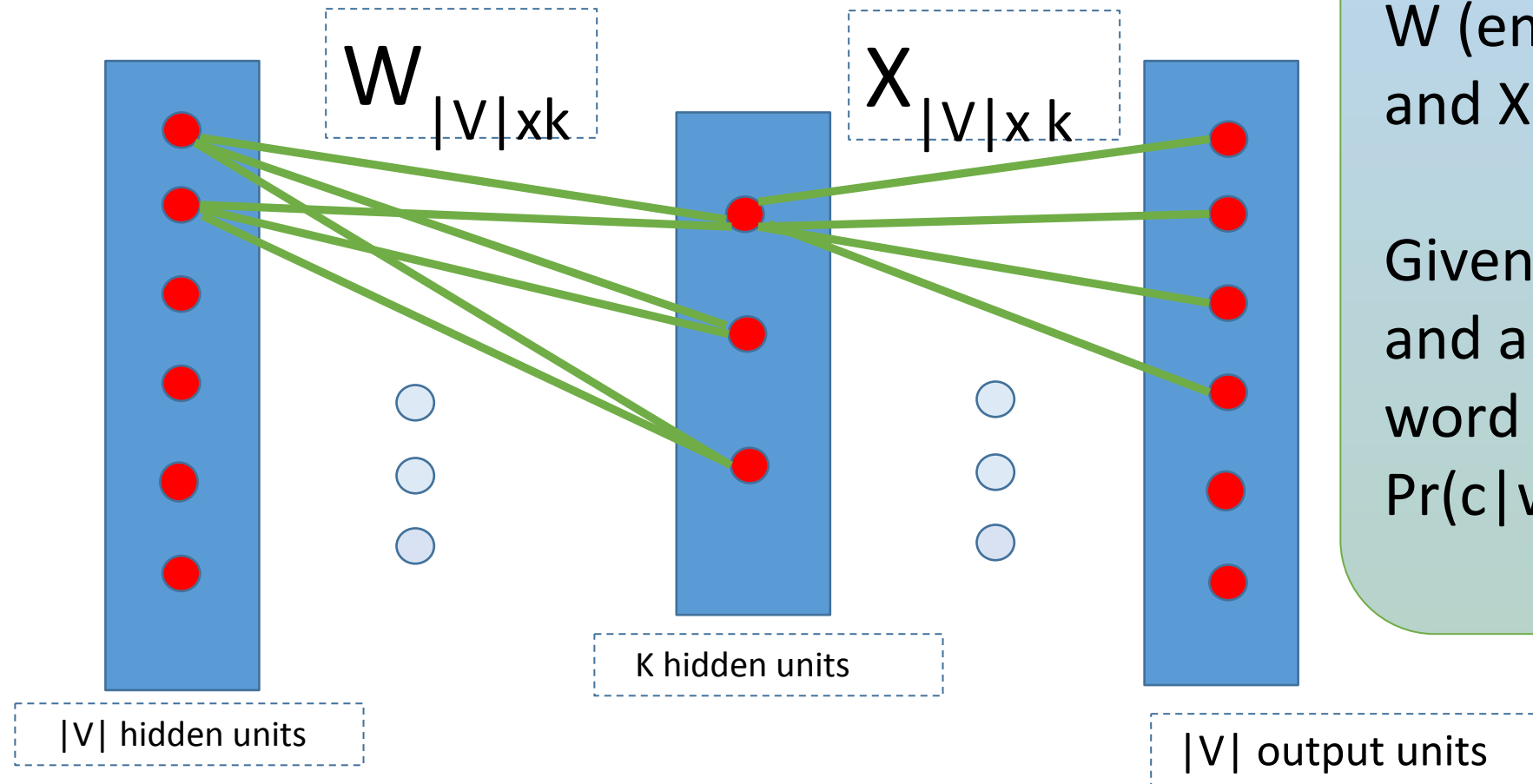
- Computational Scalability: $O(n^3)$
- Does not scale well when we have millions of words.
- Might need to apply transformations on raw co-occurrence matrices (PPMI etc) to obtain high quality embeddings

An alternative approach: Neural Word Embeddings

- Learn word embeddings directly from data
- Use a neural network based architecture
- Online, scalable to large data sets
- Implicitly factorizes the co-occurrence matrix

Skipgram model – Learning Word Embeddings

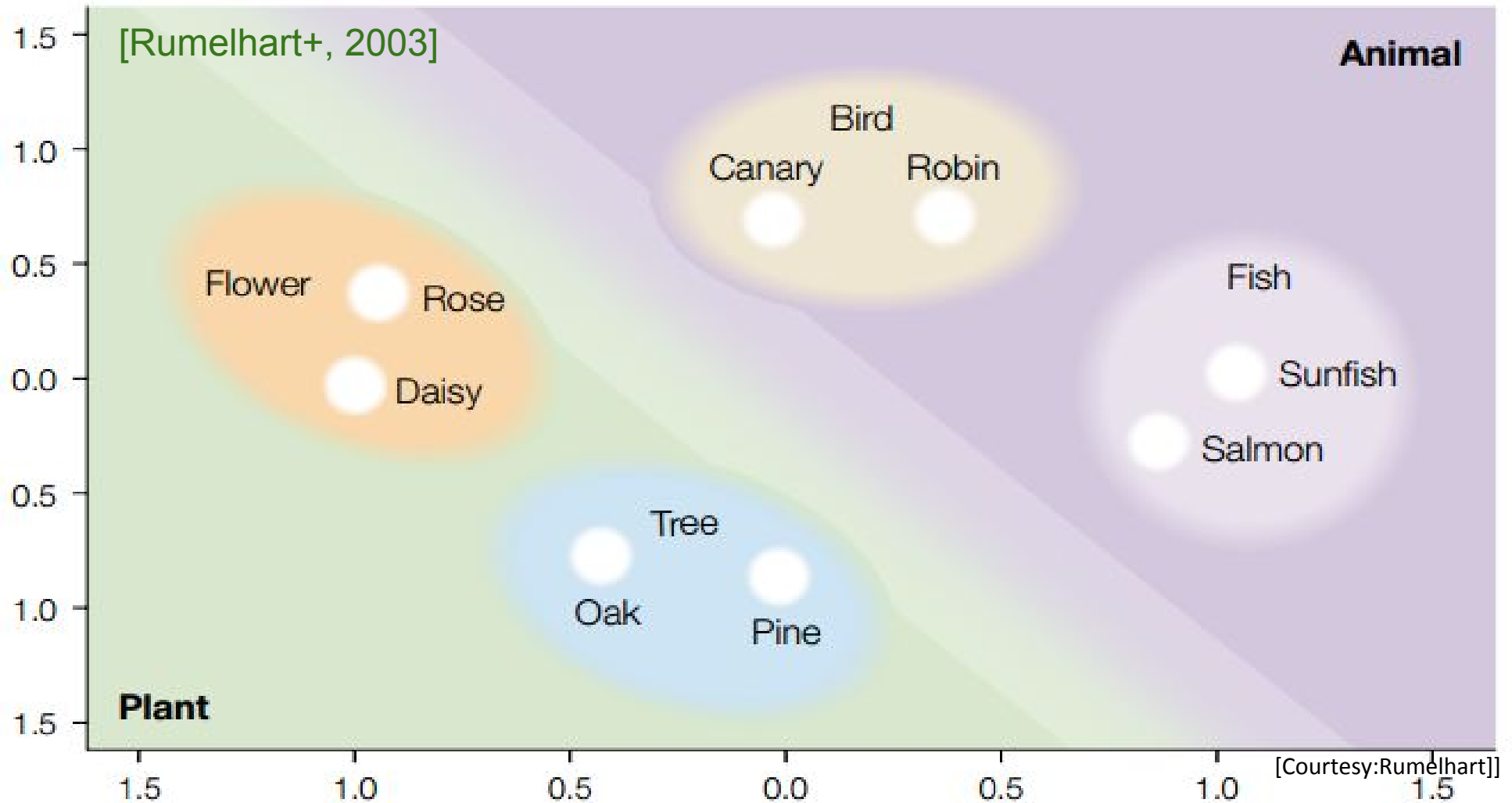
V: Vocabulary , k: Embedding size



Learn parameters W (embeddings) and X .

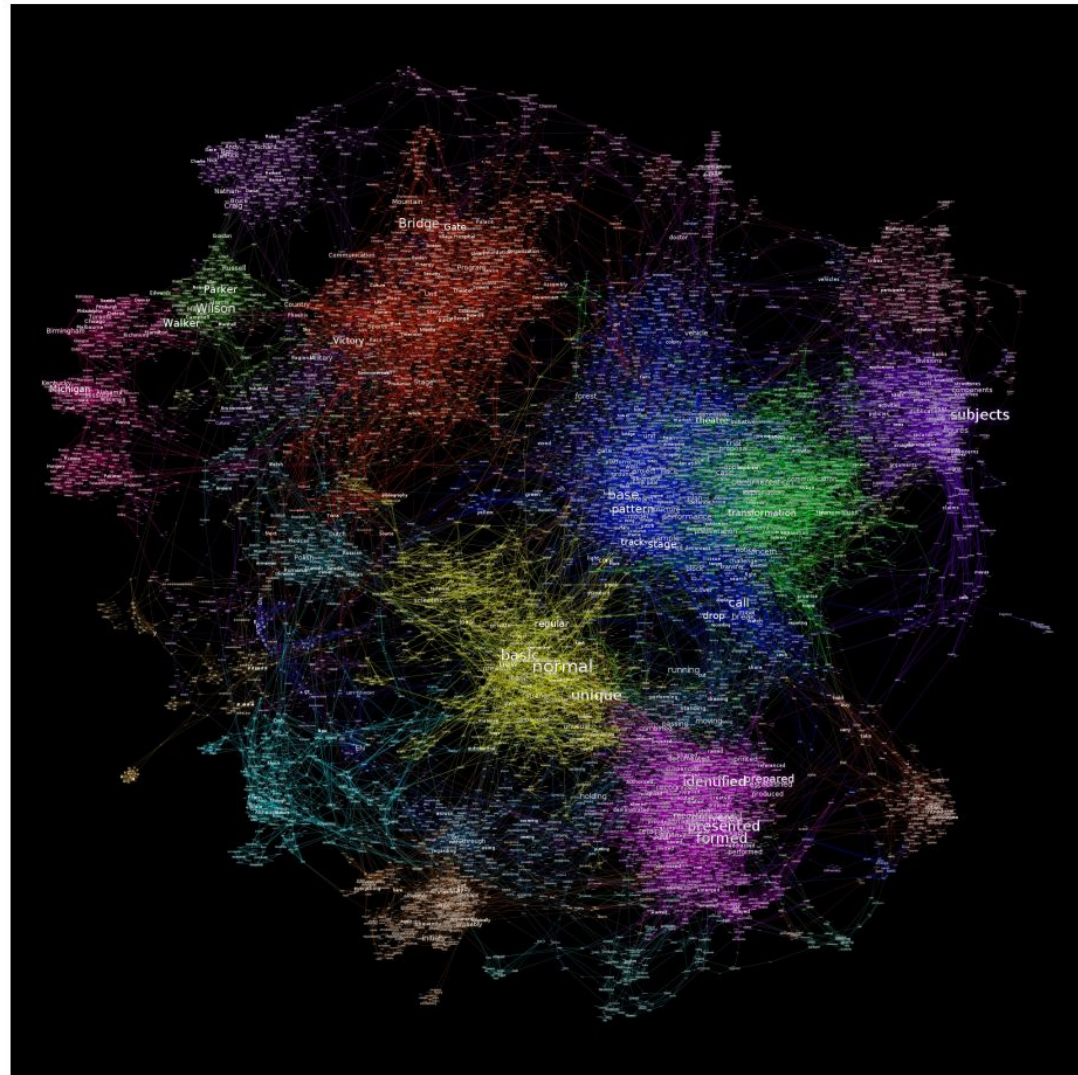
Given a word w and a context word c , maximize $\Pr(c|w)$.

Visualizing word embeddings

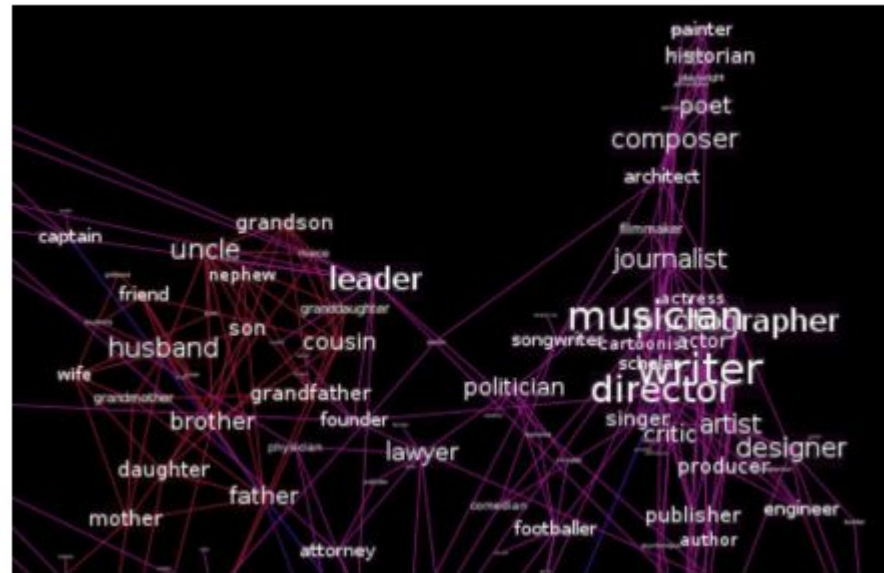
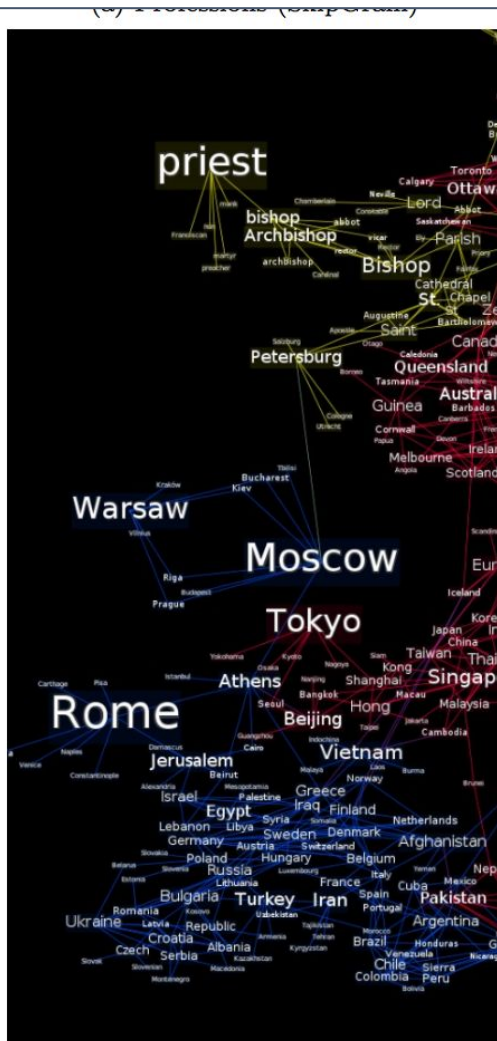


Learn a mapping from words to a continuous space.

Visualizing Word Embeddings - Word Network



Interesting clusters



Summary

- Word Embeddings are learned directly from data
- Represent words in a low dimensional space capturing similarity in meaning
- Shown to be useful features for several NLP Tasks
- Scale well to large data

THANK YOU