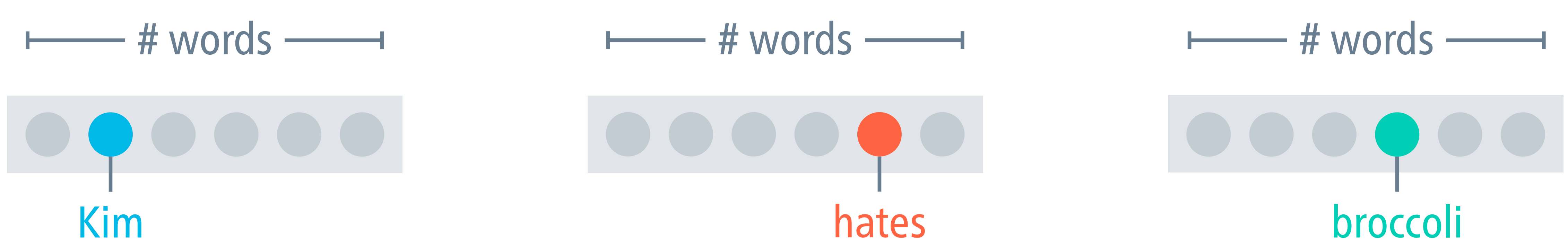
Introduction to word representations

Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

One-hot vectors

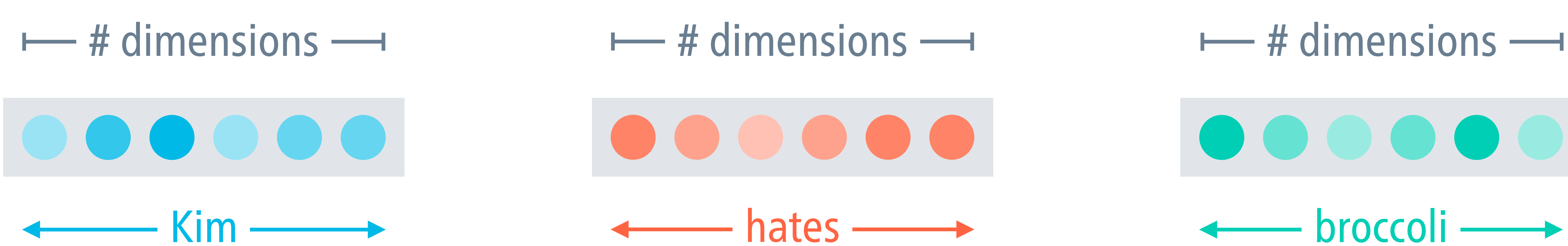
- To process words using neural networks, we need to represent them as vectors of numerical values.
- The classical way to do this is to use **one-hot vectors** – vectors in which all components but one are zero.



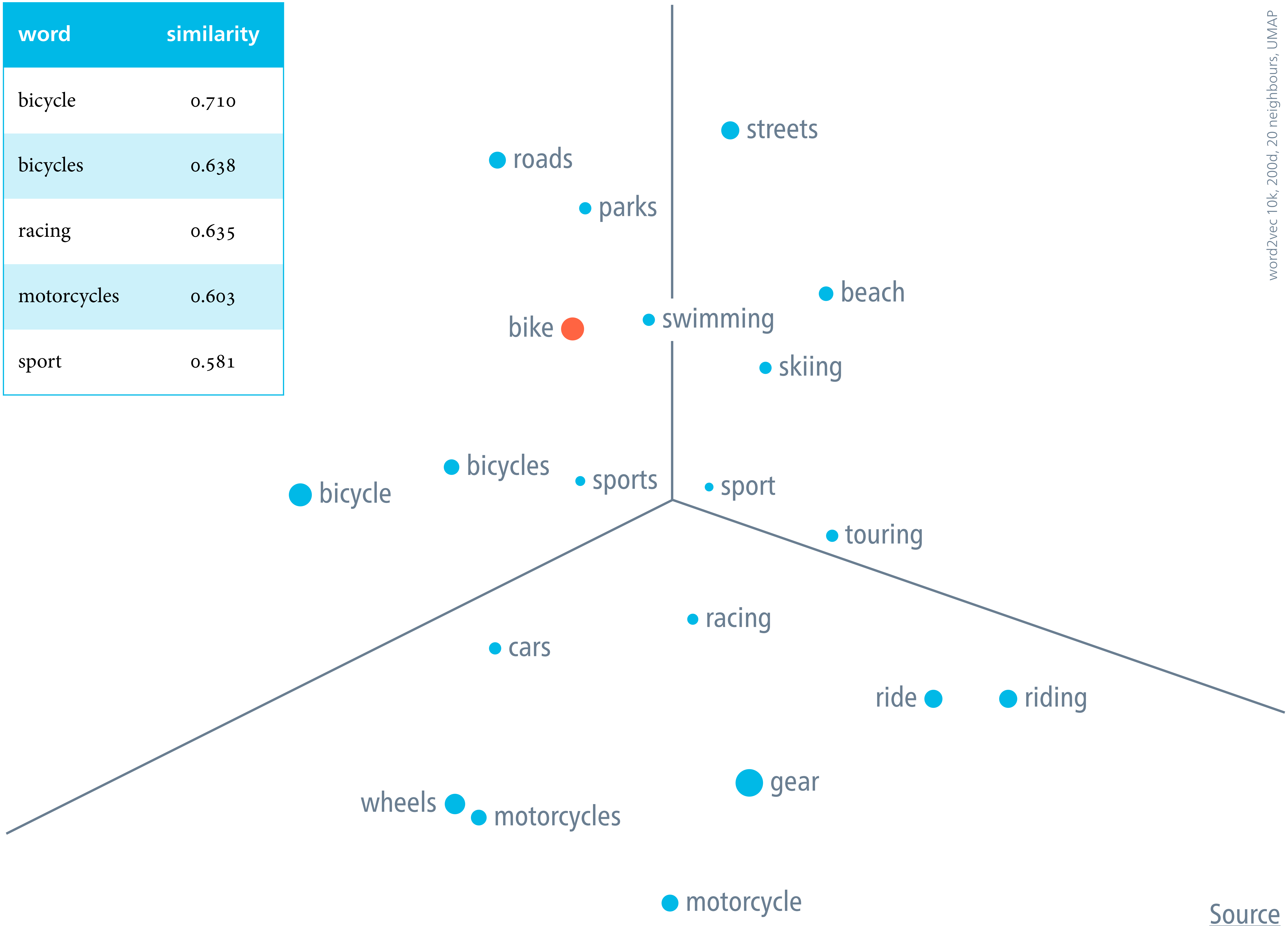
Word embeddings

Compared to one-hot vectors, **word embeddings**

- are shorter but dense
- support a useful notion of similarity
- can be learned from data



word	similarity
bicycle	0.710
bicycles	0.638
racing	0.635
motorcycles	0.603
sport	0.581



Source

You shall know a word by the company it keeps

What do the following sentences tell us about *Garrotxa*?

- *Garrotxa* is made from milk.
- *Garrotxa* pairs well with crusty country bread.
- *Garrotxa* is aged in caves to enhance mould development.

Sentences taken from the English Wikipedia

The distributional hypothesis

Eisenstein § 14.1

- The **distributional hypothesis** states that words with similar distributions have similar meanings.

  with similar distributions = are used and occur in the same contexts

- This suggests that we can learn word representations from co-occurrence statistics.

  similar co-occurrence distributions = similar meanings

Co-occurrence matrix

	cheese	bread	goat	sheep
cheese				
bread				
goat				
sheep				

as olives cheese or bread

Co-occurrence matrix

	cheese	bread	goat	sheep
cheese		1		
bread				
goat				
sheep				

as olives cheese or bread

of sheep cheese and milk

Co-occurrence matrix

	cheese	bread	goat	sheep
cheese		1		1
bread				
goat				
sheep				

as olives cheese or bread

of sheep cheese and milk

goat milk cheese can be

Co-occurrence matrix

	cheese	bread	goat	sheep
cheese		1	1	1
bread				
goat				
sheep				

as olives cheese or bread

of sheep cheese and milk

goat milk cheese can be

bread and cheese for breakfast

Co-occurrence matrix

	cheese	bread	goat	sheep
cheese		2	1	1
bread				
goat				
sheep				

as olives cheese or bread

of sheep cheese and milk

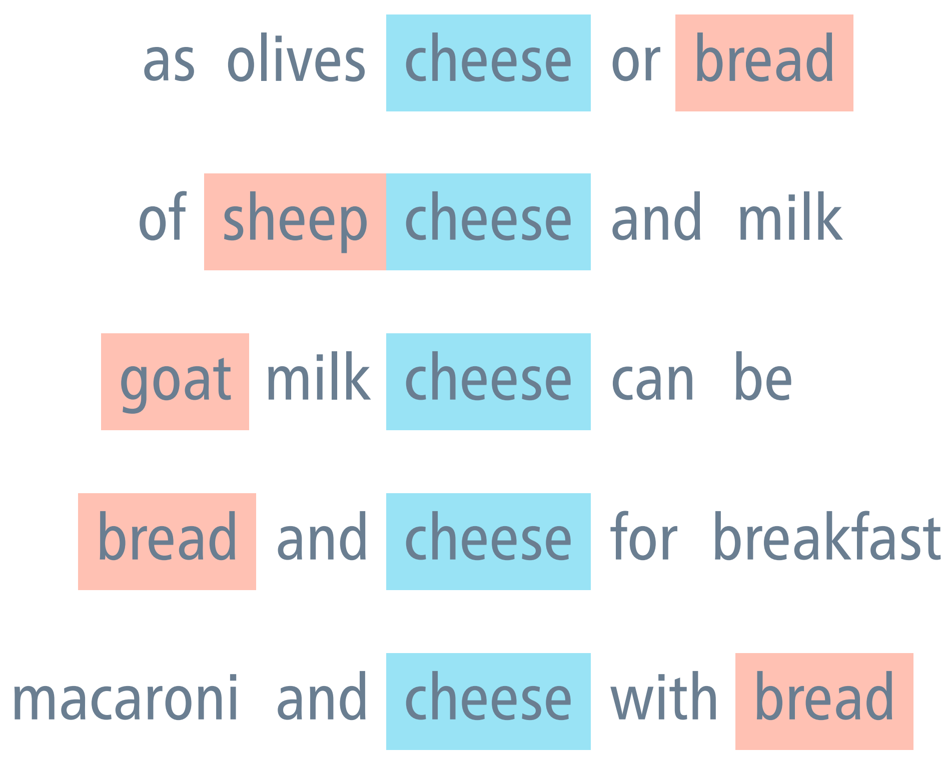
goat milk cheese can be

bread and cheese for breakfast

macaroni and cheese with bread

Co-occurrence matrix

	cheese	bread	goat	sheep
cheese		3	1	1
bread				
goat				
sheep				

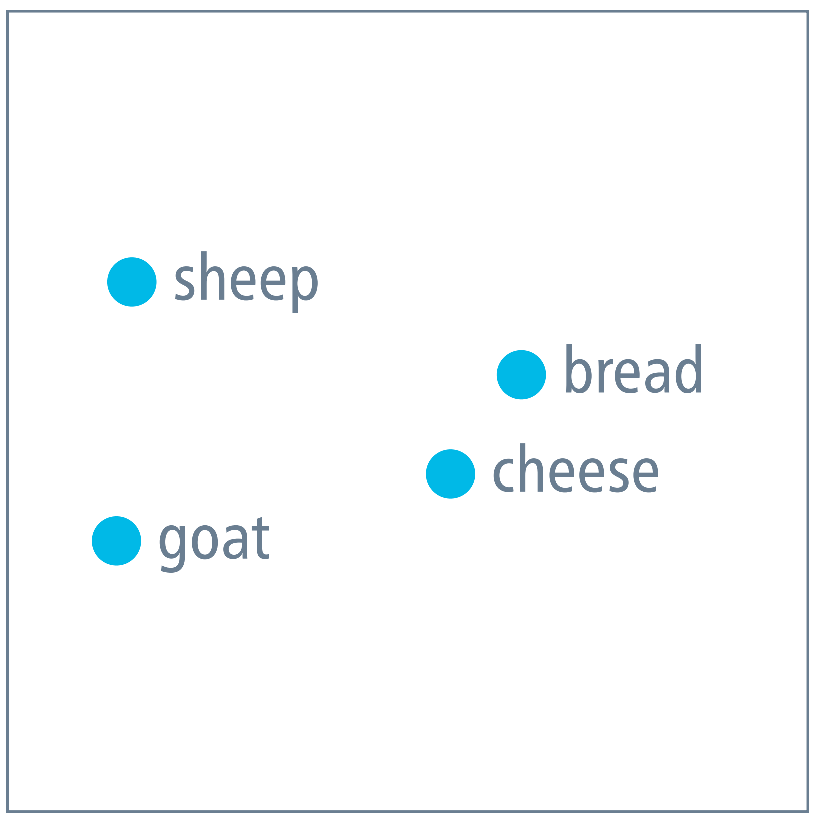


Co-occurrence matrix

	cheese	bread	goat	sheep
cheese	14	7	5	1
bread	7	12	0	0
goat	5	0	8	12
sheep	1	0	12	2

word vector
for *cheese*

Vector similarity = meaning similarity



vector space (PCA)

	cheese	bread	goat	sheep
cheese	1.00	0.80	0.49	0.38
bread	0.80	1.00	0.17	0.04
goat	0.49	0.17	1.00	0.67
sheep	0.38	0.04	0.67	1.00

cosine similarities

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$$

Learning word embeddings

- **Count-based methods: Matrix factorisation**

  Minimise the difference between the co-occurrence matrix and an approximate reconstruction of it from word embeddings.

- **Prediction-based methods: Neural networks**

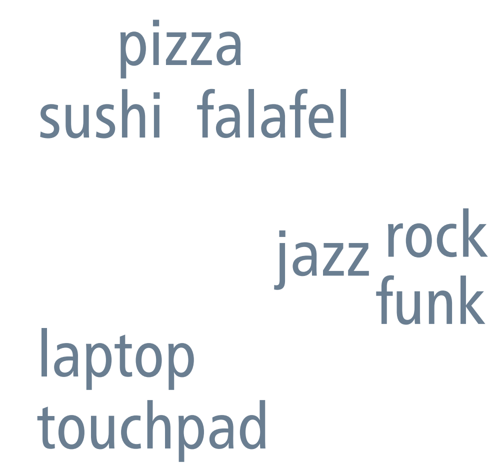
  Maximise the likelihood of a corpus under a probability model that is conditioned on the word embeddings.

Evaluation of word embeddings

Eisenstein § 14.6

- visualisation of the embedding space

  Requires dimensionality reduction (PCA, t-SNE, UMAP)



- computing relative similarities

  cosine similarity, Euclidean distance

- similarity benchmarks

  Example: odd one out – *breakfast lunch dinner* *<u>surgery</u>*

- analogy benchmarks

  Example: *woman* is to *man* as *sister* is to ?