Word Embeddings

What is meaning??

What is meaning?

the idea that is represented by the text

the idea that a person wants to convey by using words

the idea that is expressed in a work of writing, art

How can a computer interpret language?

Synonyms

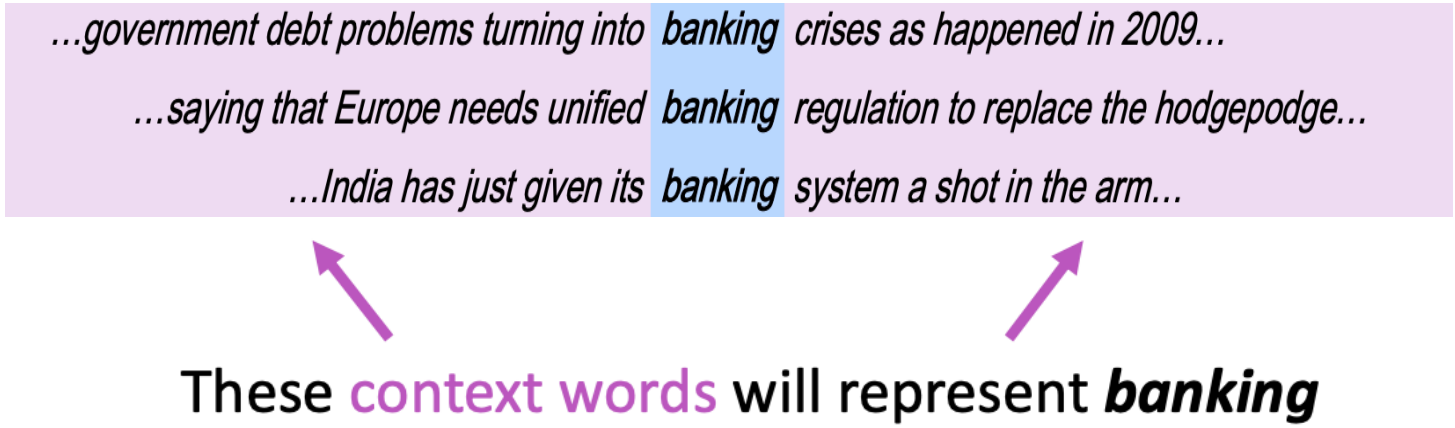
- only true in certain sentence contexts
  Example: “proficient” is listed as a synonym for “good”
- Subjective
- hard to update new words and slang as well as manual upkeep of a thesaurus

CountVectorizer/TF-IDF

- No notion of similarity or meaning of individual words
- Only gives context of importance in corpus
  - Ex: Plaza Hotel and Waldorf Astoria should have a high similarity

words representations by context

- **Distributional semantics**
- A word's meaning is given by the words that frequently appear around it
- When a word appears in a document its context is the set of words that appear nearby (within a fixed-size window)



Word2Vec

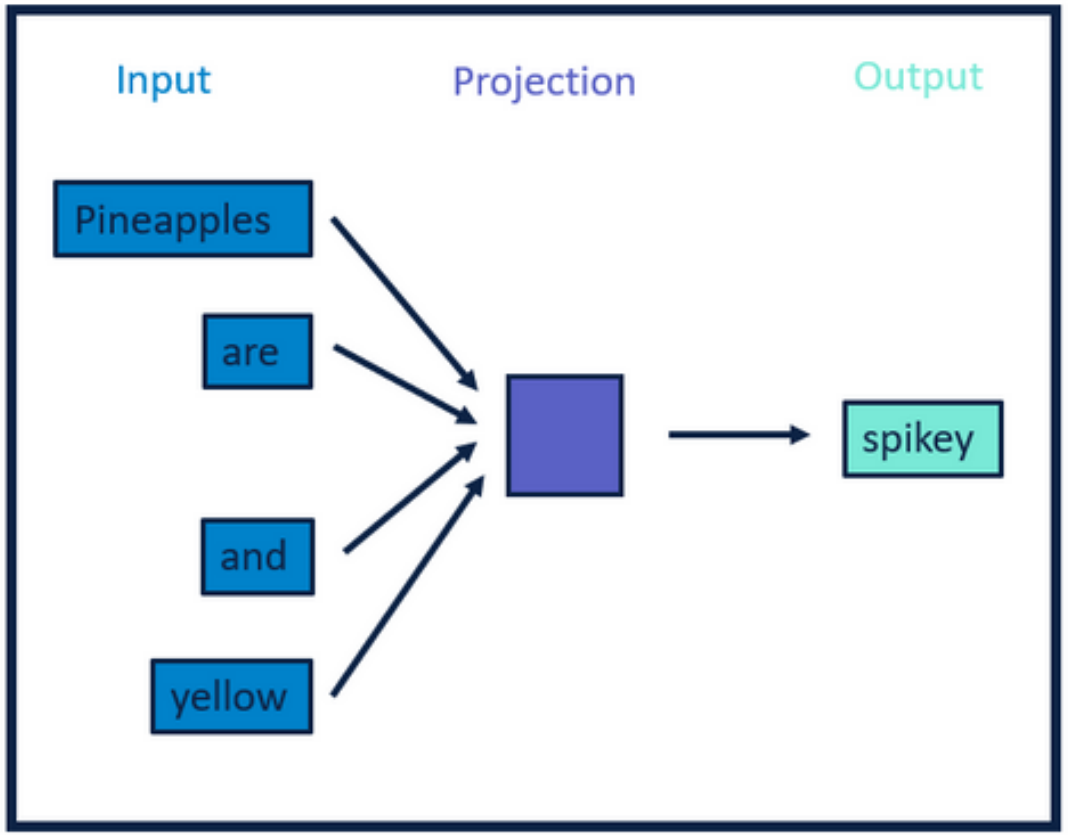
Uses a shallow (1 hidden layer) neural net to compute dense vector representation for each word

Captures the meaning of word in a corpus using predefined windows of words

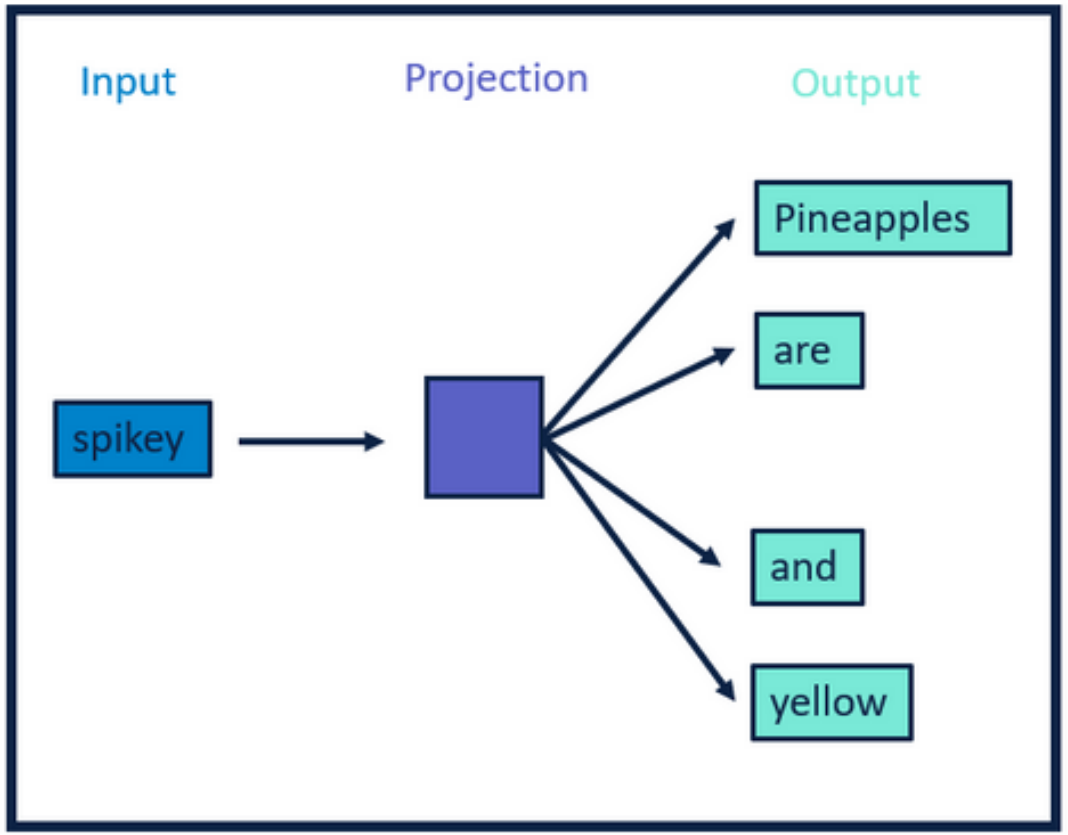
2 different implantations (Skip-gram, CBOW)

Readily available pretrained embeddings on different corpora such as Wikipedia and Google News

Continuous Bag of Words vs Skip-gram

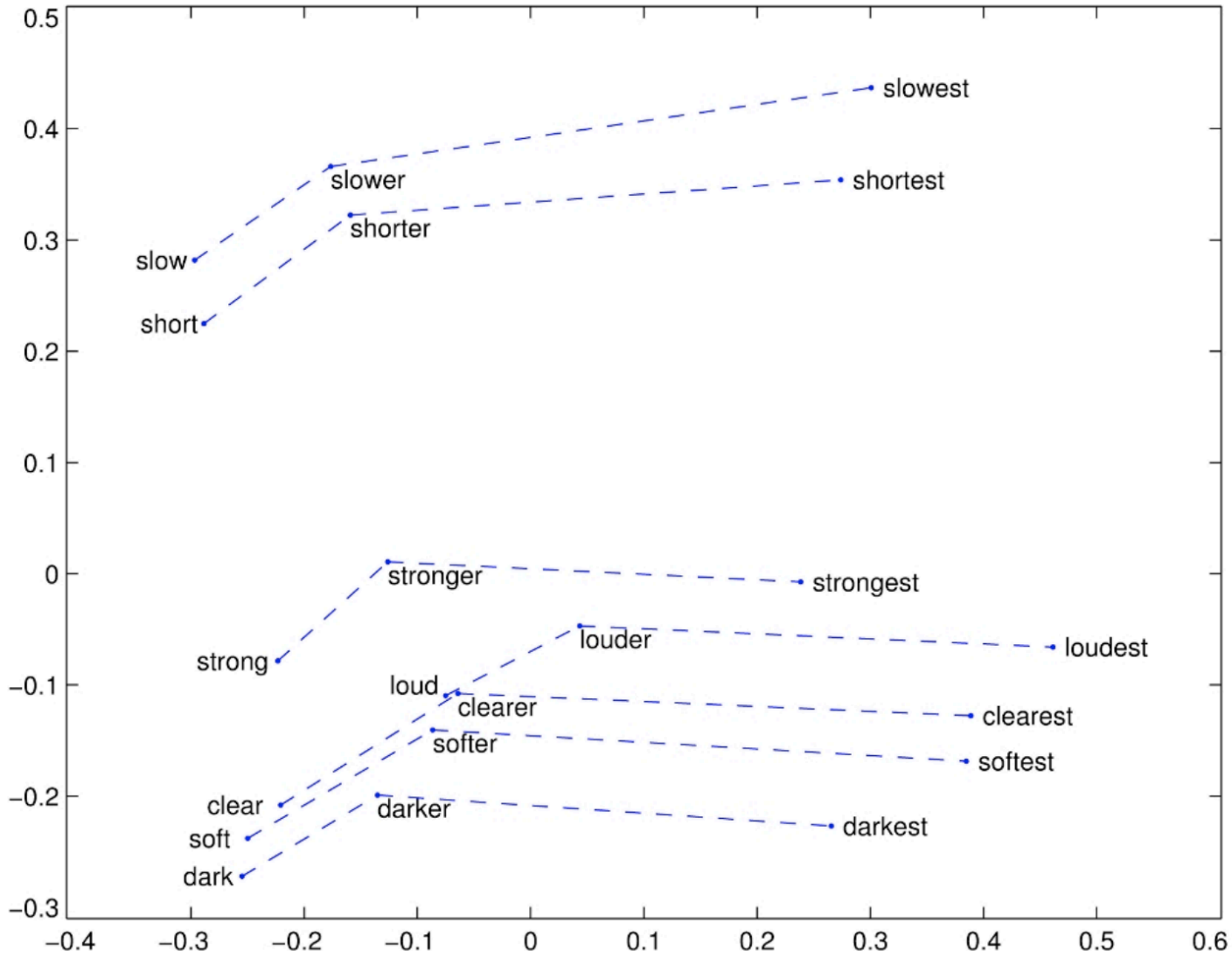


CBOW

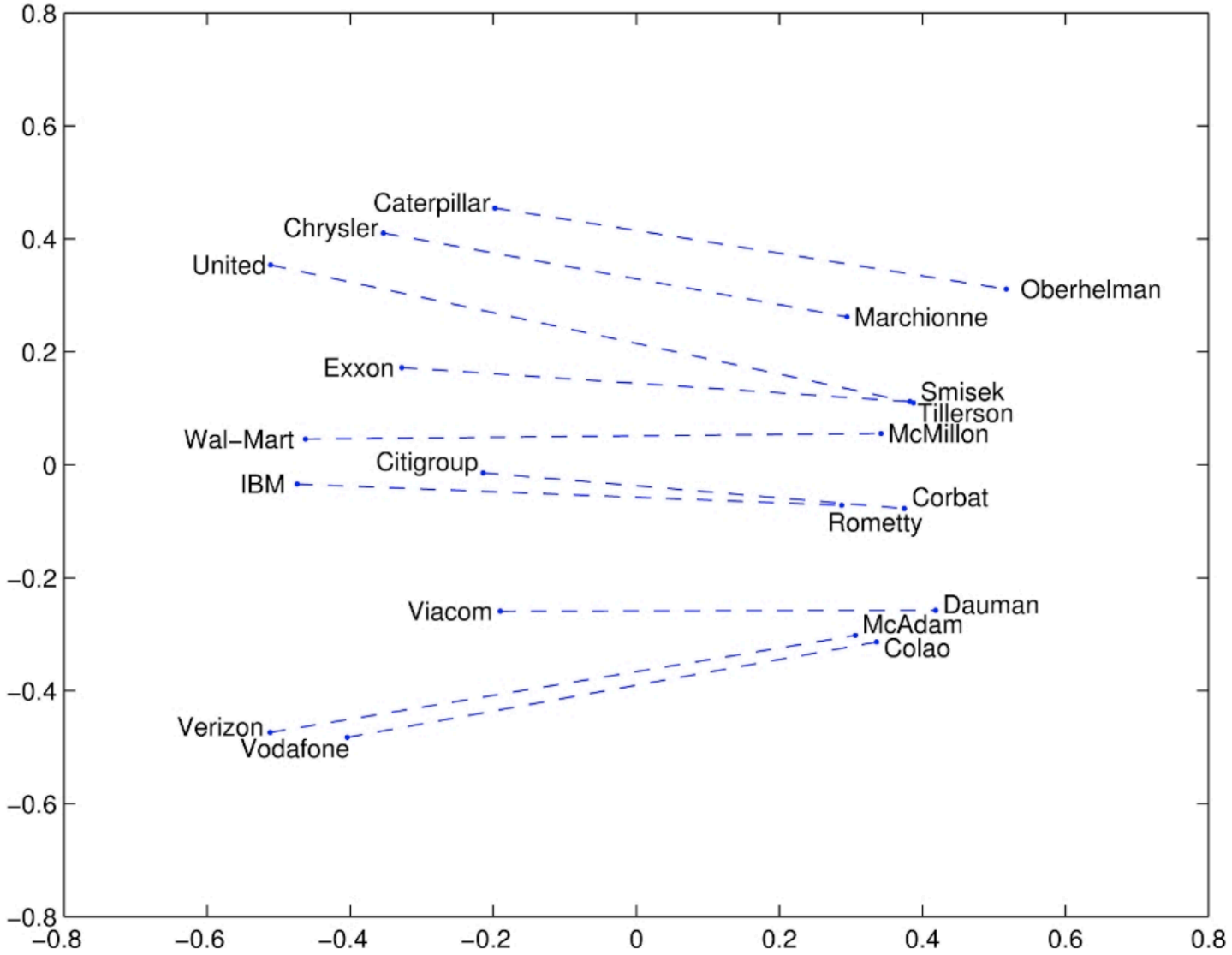


Skip-gram

Superlatives



Company → CEO


0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
Caterpillar
Chrysler
United
Exxon
Wal−Mart
Citigroup
IBM
Viacom
Verizon
Vodafone
Oberhelman
Marchionne
Smisek
Tillerson
McMillon
Corbat
Rometty
Dauman
McAdam
Colao