Word Vectors

Natural Language Processing Tasks

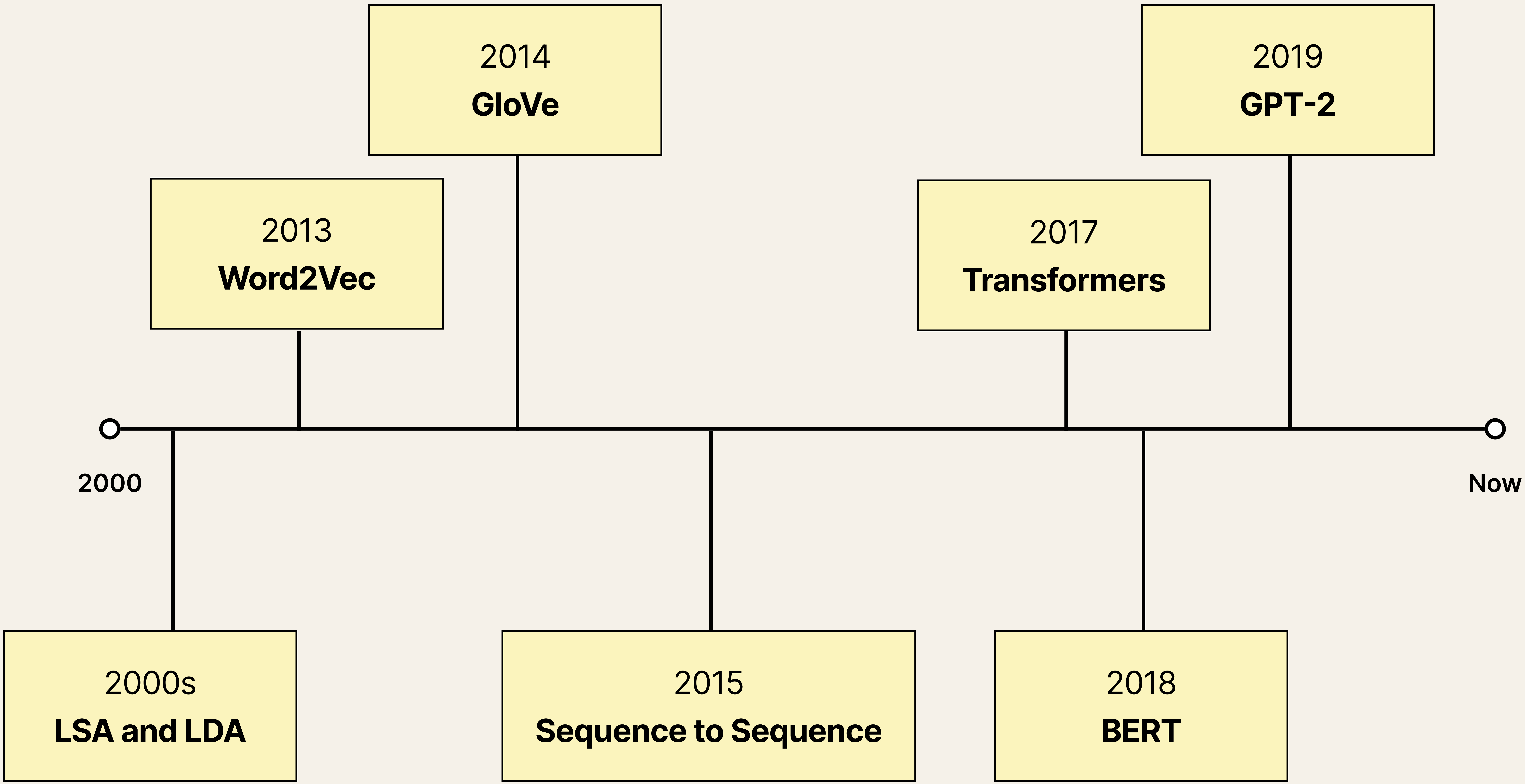
- Machine translation
- Question answering and information retrieval
- Summarization and analysis of text
- Note: speech-to-text

Natural Language Processing Tasks

Word Representation

Signifier
Sign

Signified
Meaning


2014
GloVe
2019
GPT-2
2013
Word2Vec
2017
Transformers
2000
Now
2000s
LSA and LDA
2015
Sequence to Sequence
2018
BERT

Word meaning is endlessly complex

Zuko makes the tea for his uncle.

Zuko makes the coffee for his uncle.

Zuko makes the drink for his uncle.

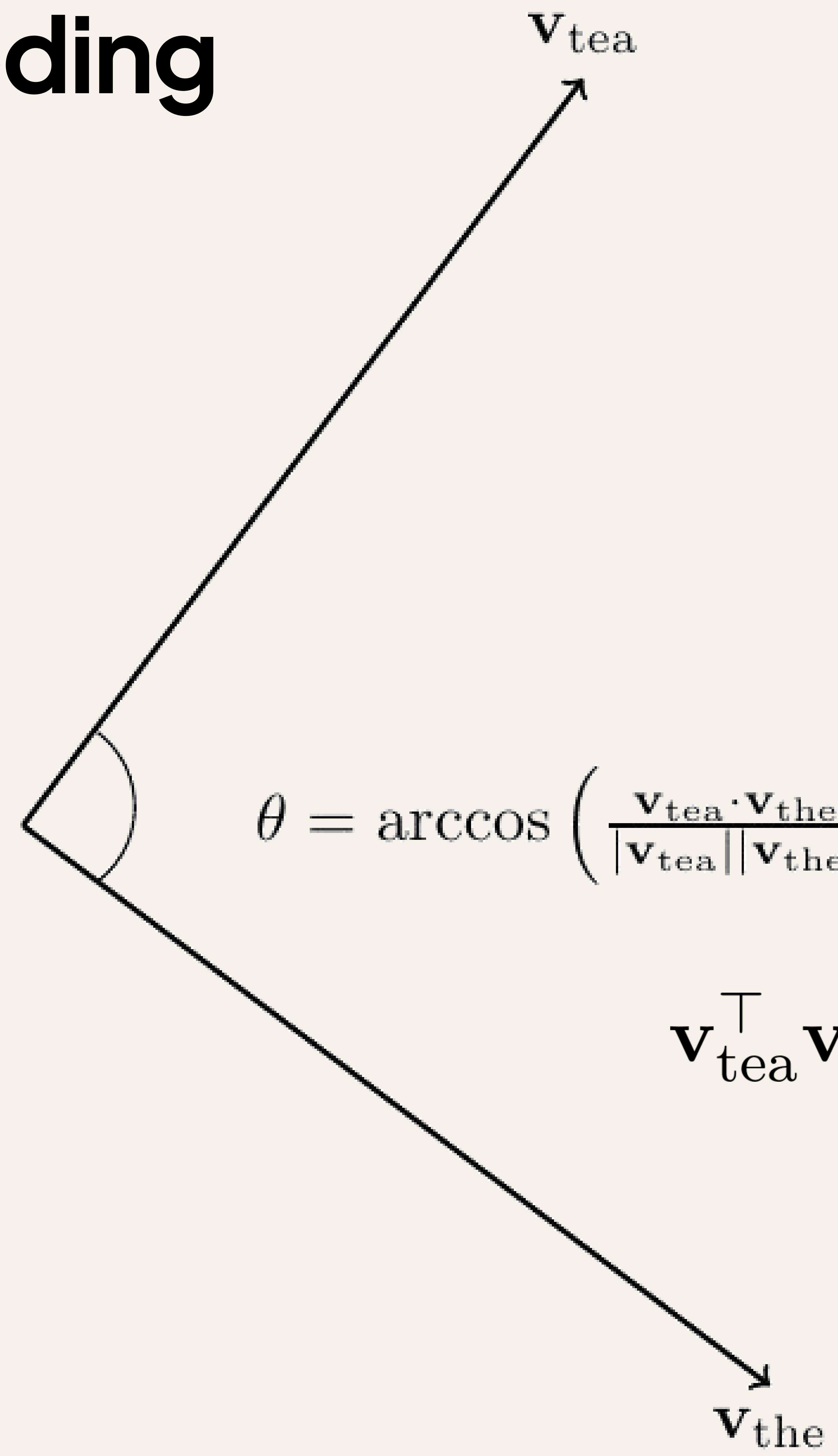
1-hot encoding

$$\mathbf{v}_{\text{tea}} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \qquad \mathbf{v}_{\text{coffee}} = \begin{bmatrix} \vdots \\ 0 \\ 0 \\ 1 \\ \vdots \end{bmatrix}$$

No notion of similarity

$$\mathbf{v}_{\text{tea}}^{\top}\mathbf{v}_{\text{coffee}} = \mathbf{v}_{\text{tea}}^{\top}\mathbf{v}_{\text{the}} = 0$$

1-hot encoding



$$\theta = \arccos\left(\frac{\mathbf{v}_{\text{tea}} \cdot \mathbf{v}_{\text{the}}}{|\mathbf{v}_{\text{tea}}||\mathbf{v}_{\text{the}}|}\right)$$

No notion of similarity

$$\mathbf{v}_{\text{tea}}^{\top}\mathbf{v}_{\text{coffee}} = \mathbf{v}_{\text{tea}}^{\top}\mathbf{v}_{\text{the}} = 0$$

Vectors from annotated discrete properties

$$\mathbf{v}_{\text{tea}} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{array}{l} \text{(plural noun)} \\ \text{(3rd singular verb)} \\ \text{(hyponym-of-beverage)} \\ \vdots \\ \text{(synonym-of-chai)} \end{array}$$

Reduced vocabulary
Always incomplete

Sparse vector

Neural Networks prefer dense vectors

[Miller, 1995] Miller, G. A. (1995). Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.

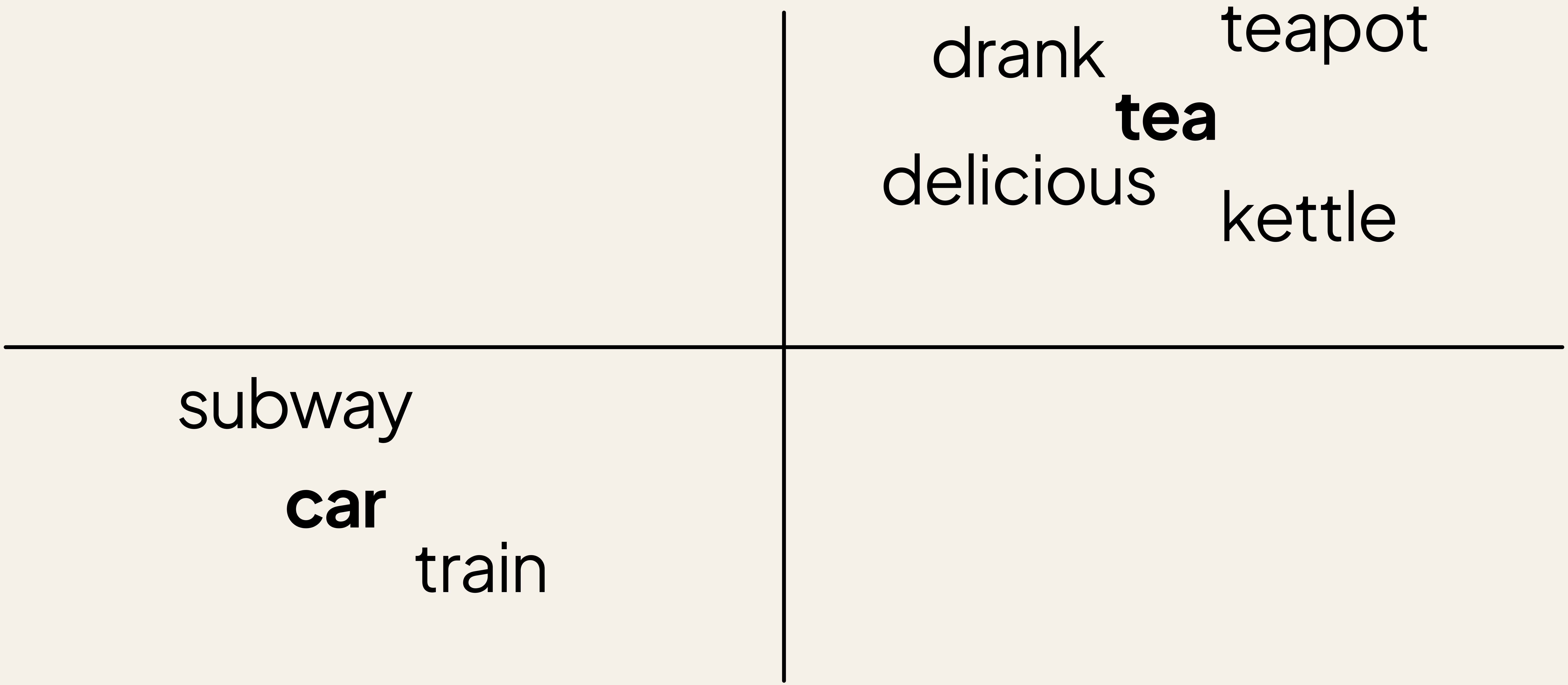
Deep Learning promises to learn

Self learning

Rich Representations

of complex objects

What should a good word distribution be like?



Co-occurrence matrices

More useful that 1-hot

$$X = \begin{array}{c|cccccc} & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 \\ \hline w_1 & 0 & 0.4 & 0.3 & 0.1 & 0.1 & 0.1 \\ w_2 & 0.2 & 0 & 0.3 & 0.2 & 0.2 & 0.1 \\ w_3 & 0.25 & 0.25 & 0 & 0.25 & 0.15 & 0.1 \\ w_4 & 0.3 & 0.1 & 0.2 & 0 & 0.3 & 0.1 \\ w_5 & 0.1 & 0.4 & 0.2 & 0.1 & 0 & 0.2 \\ w_6 & 0.2 & 0.2 & 0.2 & 0.1 & 0.3 & 0 \end{array}$$

How count co-occurrence

Whole document

Window with fixed size

$$[\text{It's hot and delicious. } [\text{I poured } [\text{the } \underbrace{\text{tea}}_{\text{center word}} \text{ for}]_1 \text{ my uncle}]_3.]_{\text{document}}$$

Feedforward Neural Net Language Model (NNLM)

Efficient Estimation of Word Representations in Vector Space (original word2vec paper)

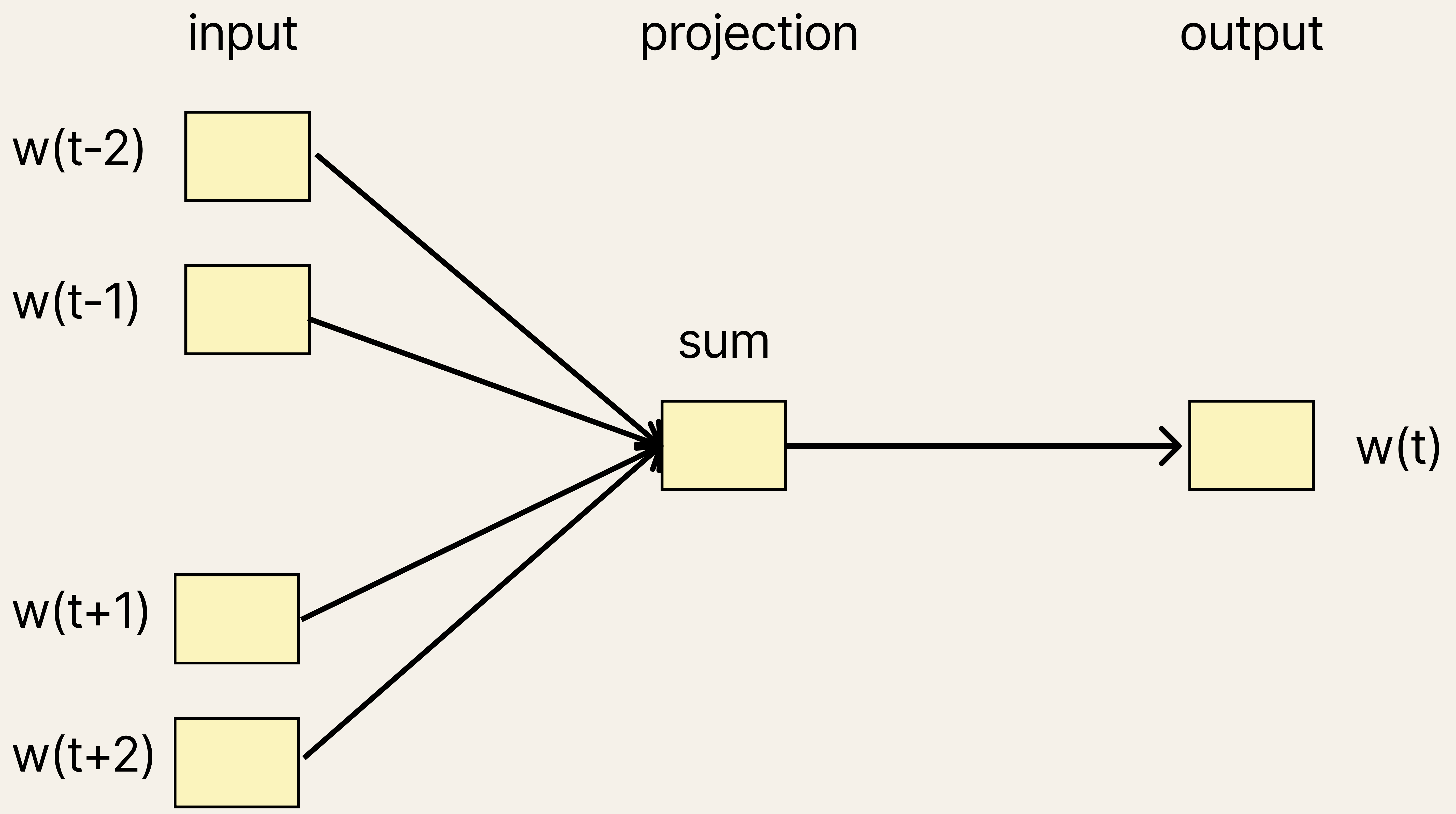
notes 1: https://web.stanford.edu/class/cs224n/readings/cs224n-2019-notes02-wordvecs2.pdf

Recurrent Neural Net Language Model (RNNLM)

Efficient Estimation of Word Representations in Vector Space (original word2vec paper)

notes 1: https://web.stanford.edu/class/cs224n/readings/cs224n-2019-notes02-wordvecs2.pdf

Continuous Bag-of-Words Model (CBOW)



Efficient Estimation of Word Representations in Vector Space (original word2vec paper)
notes 1: https://web.stanford.edu/class/cs224n/readings/cs224n-2019-notes02-wordvecs2.pdf

Continuous Skip-gram Model

Efficient Estimation of Word Representations in Vector Space (original word2vec paper)

GloVe

GloVe: Global Vectors for Word Representation (original GloVe paper)

notes: https://web.stanford.edu/class/cs224n/readings/cs224n-2019-notes02-wordvecs2.pdf

CS224N Assignment 1: Exploring Word Vectors