University of Amsterdam
Economics & Business

Data Science Minor - 2020/2021 - Sem. 1, Period 3

Text Retrieval Text Mining
Language Models, Word Embeddings

Julien ROSSI

University of Amsterdam
Economics & Business

Week 3

After this lecture, you will:

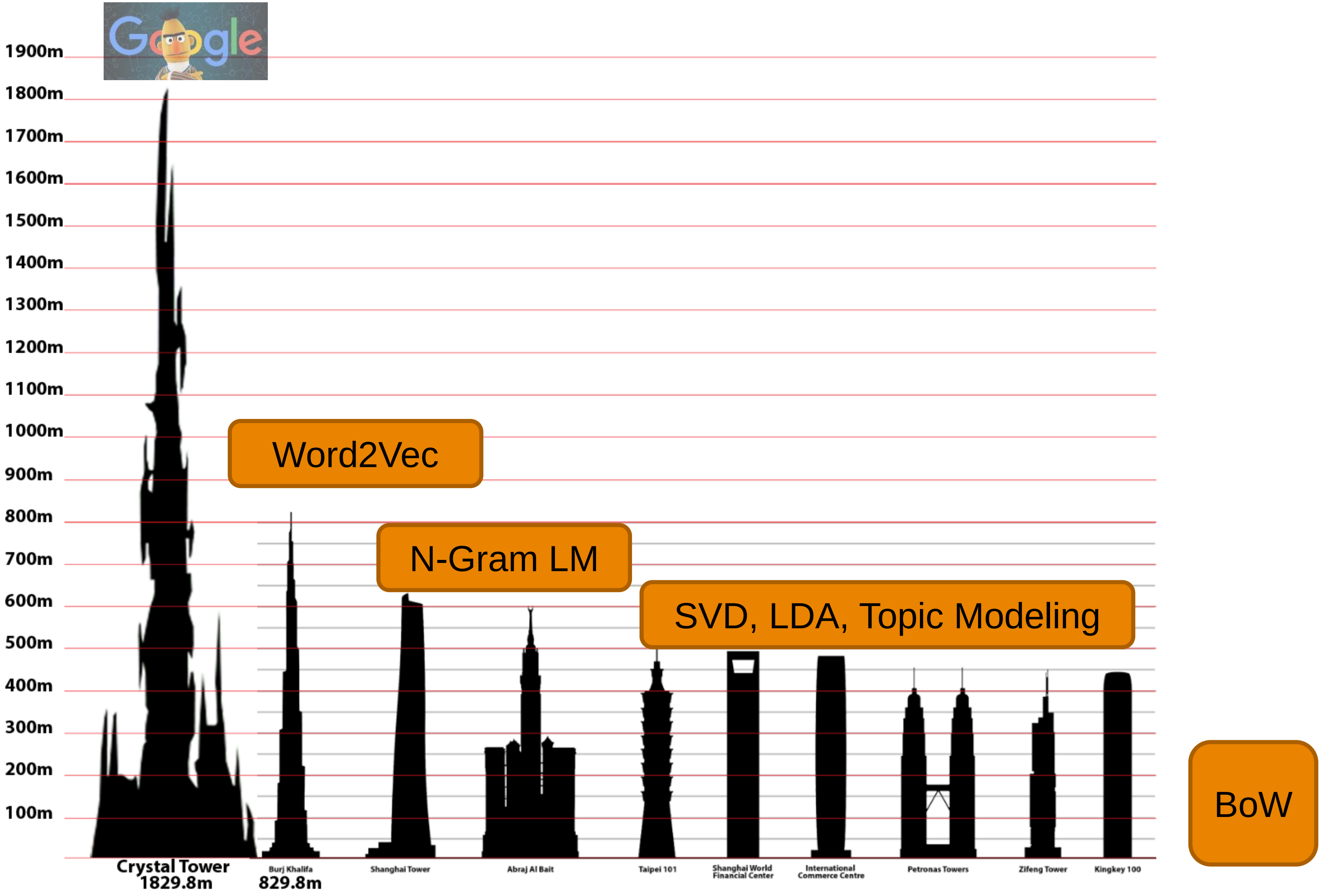
- Understand what a Language Model is
- Discover basics of Neural Networks

- Apply the concept to Word Embeddings (example: Word2Vec)

Previously in Text Representations

- **Lexical Representations**: based on words
  - **Bag of Words** raw count or TFIDF

- **Semantic Representations**: based on topics
  - **SVD** based on Term-Document matrix factorization
  - **LDA** based on random generative process



Tallest buildings chart from https://www.reddit.com/r/ffxiv/comments/e9iuqi/the_crystal_tower_compared_to_the_real_worlds/

Bert Google Image from https://www.seroundtable.com/google-bert-update-28427.html

The Journey

N-Gram Language Models

- The sentence starts with "*my cat jumps*", what could possibly be the 4th word? → Probability Distribution

Neural Networks

- Given an input, estimate the Probability Distribution of an output → Predict context words

Philosophy: what is the meaning of a word?

Distributional Semantics

- From a corpus, extract millions of (words - context words) samples,
- Predict context words with a Neural Network → Learn word embeddings (word2vec)

The Destination

- $\overrightarrow{\text{paris}}$ is the vector that represents the word “paris” (the name of a magnificent city)

- $\overrightarrow{\text{paris}} - \overrightarrow{\text{france}} + \overrightarrow{\text{netherlands}} \approx \overrightarrow{\text{amsterdam}}$
- $\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}} \approx \overrightarrow{\text{queen}}$

University of Amsterdam
Economics & Business

N-Gram Language Model

We learn a language by predicting which words to use.

N-Gram Language Model:

Probability Distribution of the next word in the sentence, based on the previous N-1.

Discrete Distribution over the Vocabulary.

3-Gram LM: What is likely to follow **“The cat”** in a text?

Pr[“hello”] = 0.0000000001

Pr[“is”] = 0.6

Pr[“traded”] = 0.001

N-Gram Language Model

- 3-Gram LM: Next words based on **2 (= 3 – 1)** previous words
  - Predict the next word after *"the cat"*
  - Know probability distribution over the vocabulary for the word following *"the cat"*
  - Get probability distribution from observing frequencies in a large corpus of texts

- N-Gram LM can be used for **text generation**

N-Gram Language Model

- N-Gram LM for **classification**

- Given 2 classes of document
  - For example POSITIVE or NEGATIVE review
  - 1 N-Gram LM for each class
- Given a text
  - Compute Prob[text] under each model
  - Attribute the text to the model that gave it the highest probability

University of Amsterdam
Economics & Business

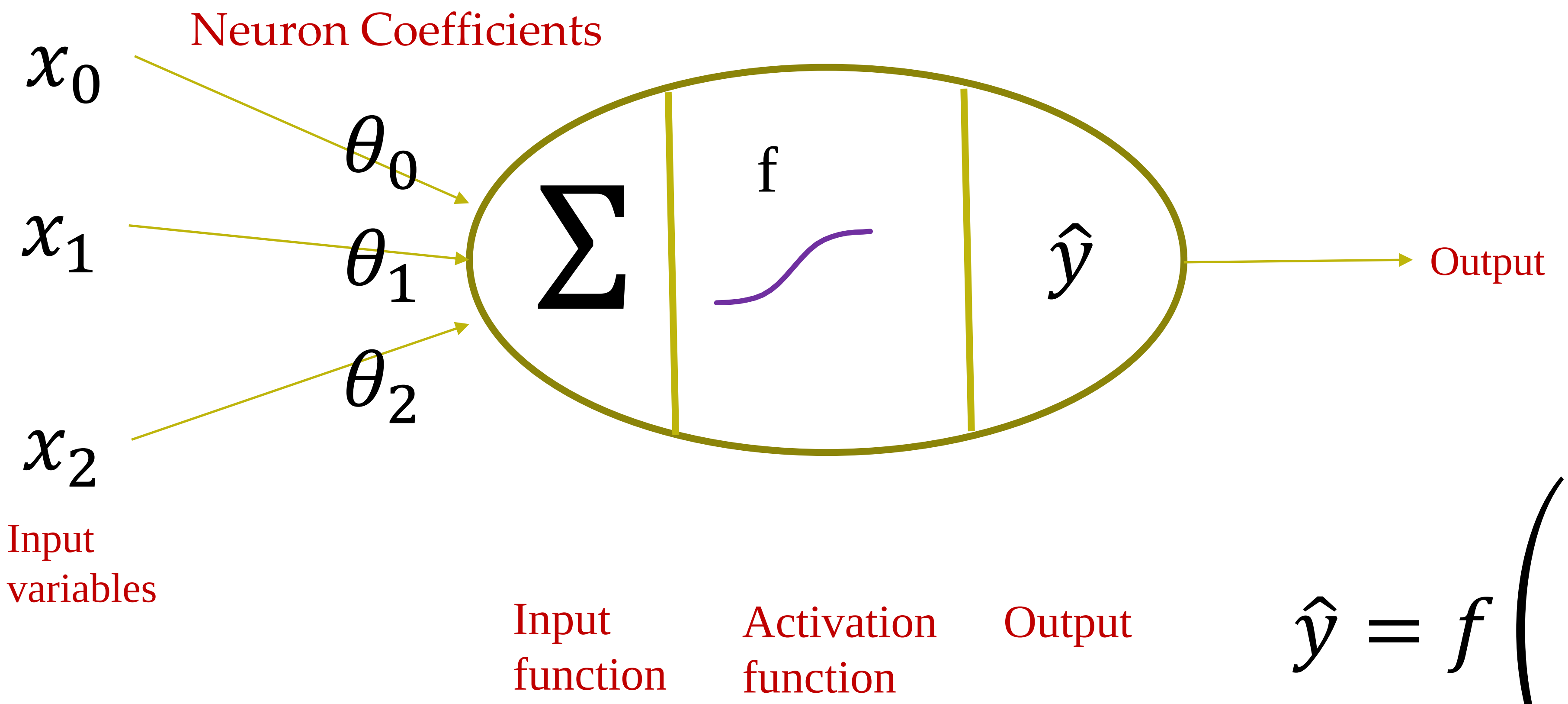
N-Gram Language Model

- See Notebook “N-Gram LM”

```
Generated Text - 47 words - Pr[Text]=8.3E-47
********************************************************************************
year - on - year agreement under which the sugar for export to third quarter
invisibles surplus was likely to have implications for the offer is scheduled to
be held up by subsidies and trade talks between the two companies said they cut
ours in half .
********************************************************************************
```

	precision	recall	f1-score	support
BROWN	0.98	0.94	0.96	5734
REUTERS	0.94	0.98	0.96	5472
accuracy			0.96	11206
macro avg	0.96	0.96	0.96	11206
weighted avg	0.96	0.96	0.96	11206

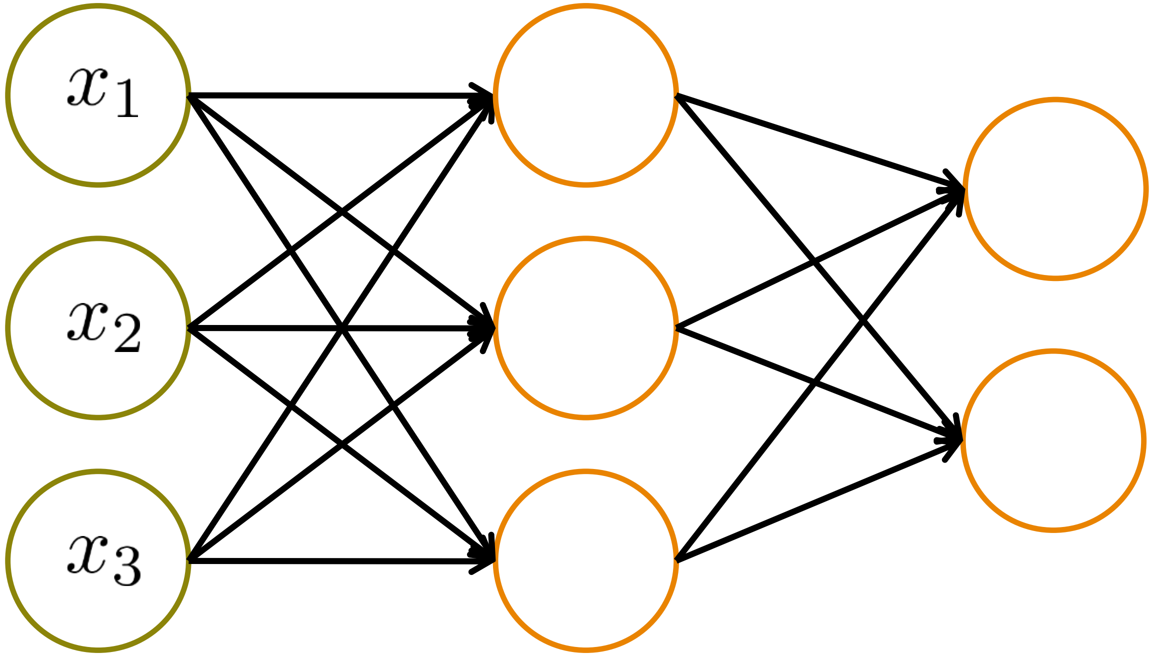
Neuron in a Neural Network



$$\hat{y} = f\left(\sum_{j=0}^{n} \theta_j x_j\right)$$

University of Amsterdam
Economics & Business

Neural Network



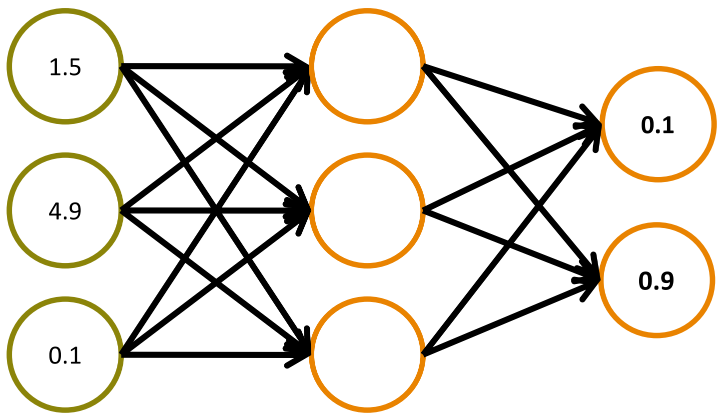
Input Layer

Hidden Layer

Output Layer

University of Amsterdam
Economics & Business

Neural Network



Input Layer | Hidden Layer | Output Layer

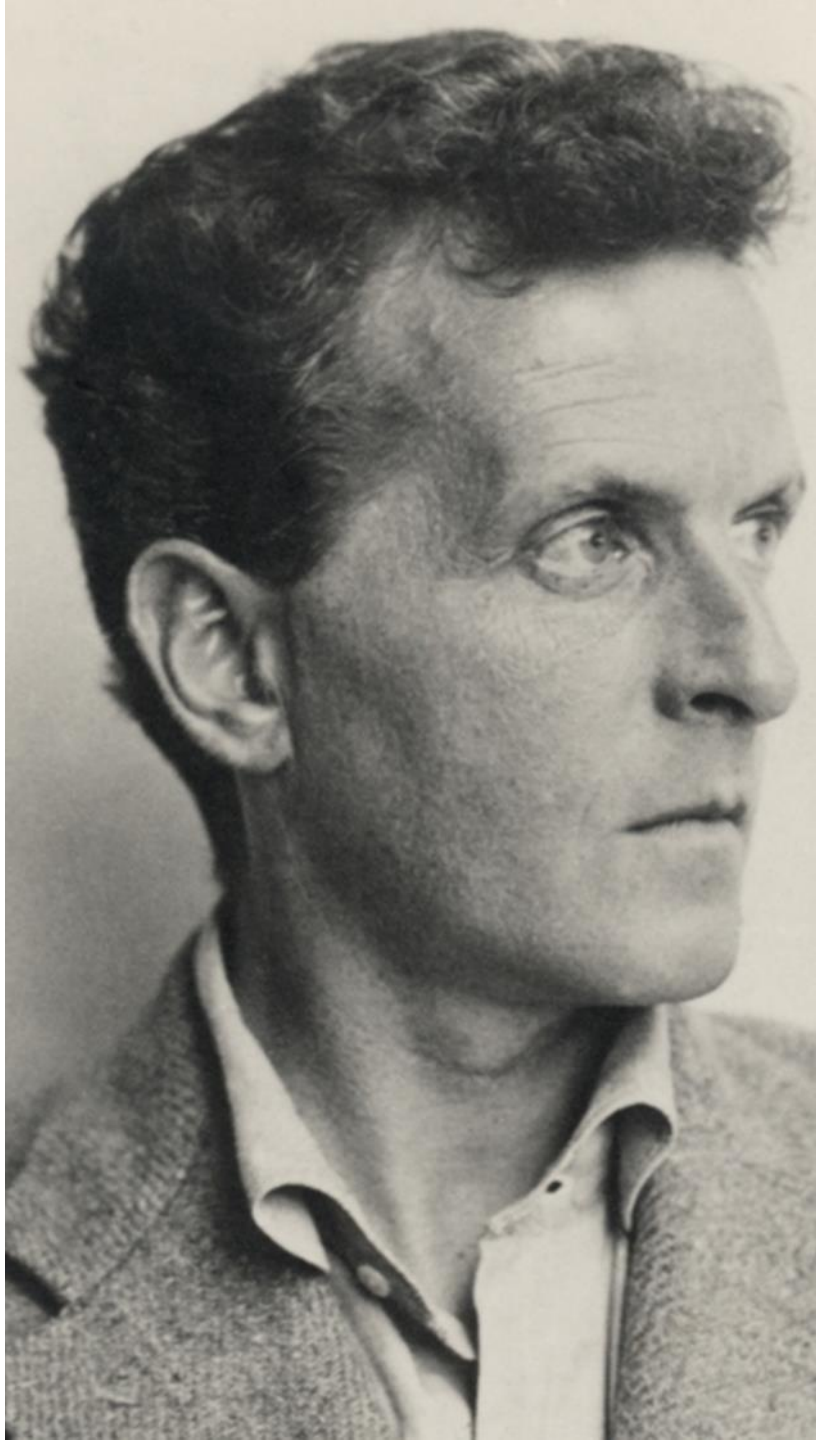
- Given the inputs
- Classification Probability:
  - 0.1 that it is of class A
  - 0.9 that it is of class B

Concept

*"In most cases, the **meaning** of a word is its **use**"*

Ludwig Wittgentein, in "Philosophical Investigations", 1953

- We can derive the meaning of a word by observing how it is used
- We can leverage collections of texts and extract meaning from the text
- As opposed to extract meaning from a linguistic lineage



Concept

"You shall know a word by the ***company*** *it keeps"*

John Rupert Firth, 1957

- We can derive the meaning of a word by observing which other words appear around it
- We can compare 2 words by comparing the words that appear around them

University of Amsterdam
Economics & Business

Concept

USE OF A WORD

Typical sentence including this word

USE OF A WORD

Words that occur around this word in existing texts

HOW DID WE LEARN WHEN TO USE WHICH WORD ?

UNIVERSITY OF AMSTERDAM
Economics & Business

Distributional Semantics

- Do not open a dictionary, or Google…

- What is **tesgüino** ?

Distributional Semantics

- Example of sentences with the word **tesgüino** in them:

A bottle of ***tesgüino*** is on the table
Everybody likes ***tesgüino***
Tesgüino makes you drunk
We make ***tesgüino*** out of corn.

- What is **tesgüino** ?

UNIVERSITY OF AMSTERDAM
Economics & Business

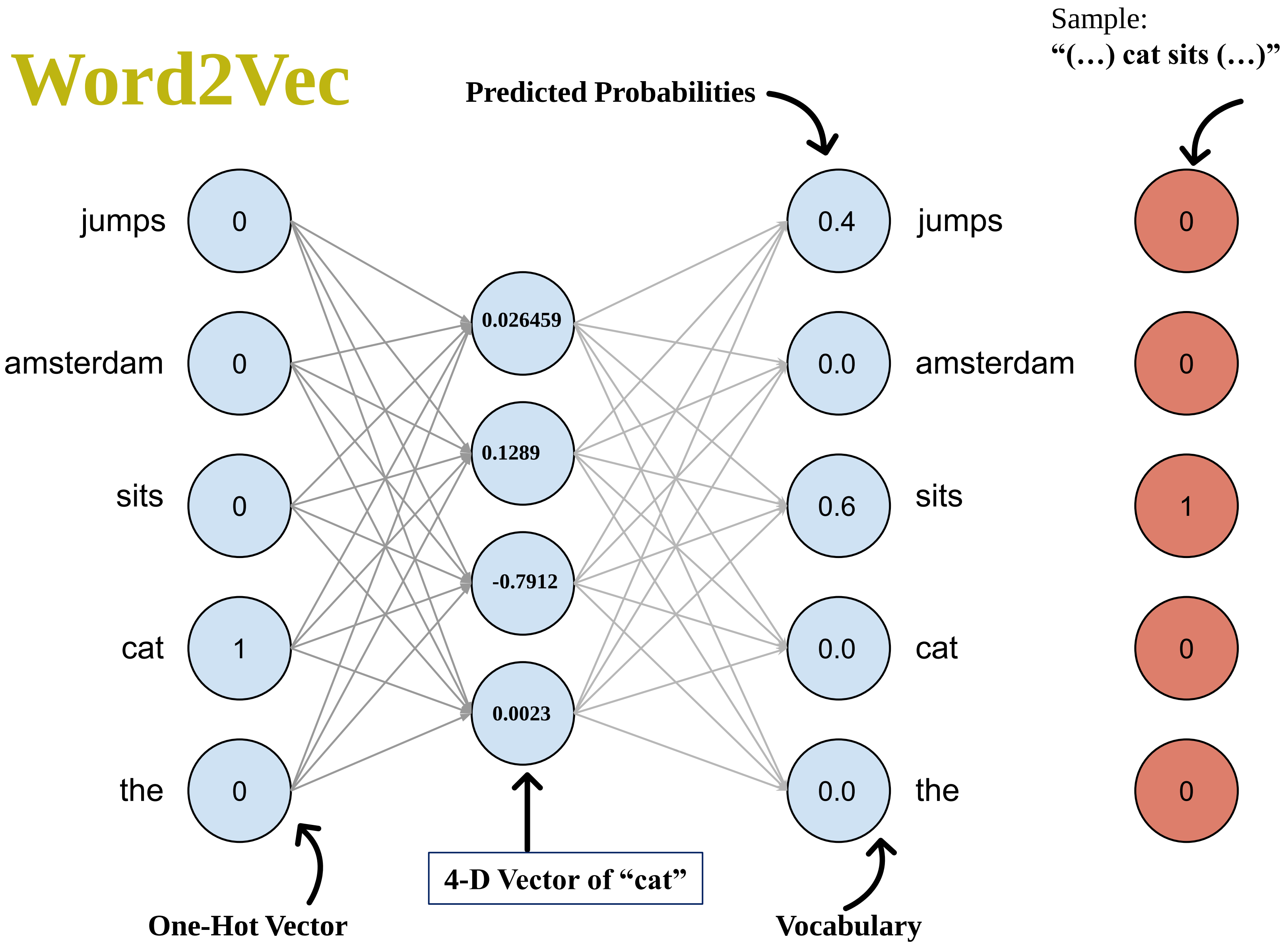
Word2Vec

<u>Mikolov & al (2013)</u>
"*Distributed representations of words and phrases and their compositionality*"

- Key concepts
  - Predict N words around a given word
  - Example <u>'cat'</u>
    - Evaluate a probability distribution over vocabulary
    - <u>'the'</u> and <u>'jumps'</u> are <u>very likely</u>
    - <u>'benefit'</u> and <u>'absorption'</u> are <u>unlikely</u>

- **Use available texts as training material**

Word2Vec


Sample:
"(…) cat sits (…)"
Predicted Probabilities
jumps
amsterdam
sits
cat
the
0
0
0
1
0
0.026459
0.1289
-0.7912
0.0023
0.4
0.0
0.6
0.0
0.0
jumps
amsterdam
sits
cat
the
0
0
1
0
0
One-Hot Vector
4-D Vector of "cat"
Vocabulary

University of Amsterdam
Economics & Business

Word2Vec

Present samples from existing texts

Minimize the cross-entropy between the predicted distribution and the ground truth

Word2vec

(Mikolov et al. 2013)

Word2Vec

- Given a word,
- We build a vector…
- … that builds the probability distribution of words around this word, as observed in our corpus
- We consider this vector to be a **word embedding**

University of Amsterdam
Economics & Business

Word2Vec

- Captures semantics, why ?

1. Words with similar meanings will appear around the same words
2. The probability distributions generated by their vectors will be similar
3. Their vectors will be similar

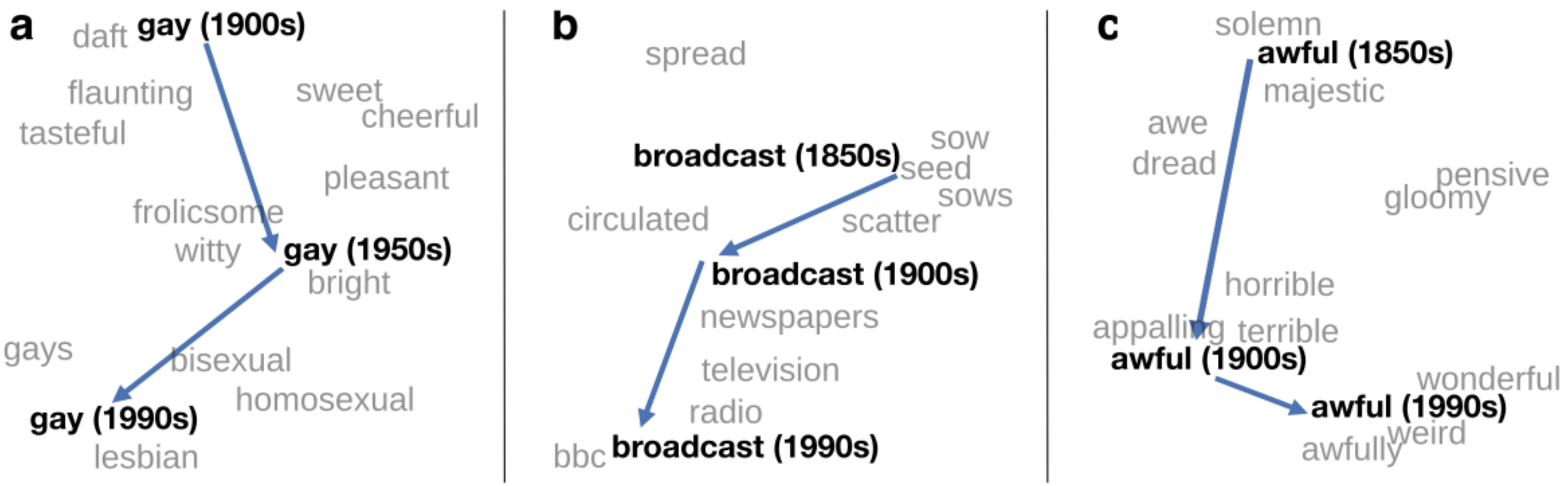
Word2Vec

- Composition captures semantics
- $\overrightarrow{\text{paris}} - \overrightarrow{\text{france}} + \overrightarrow{\text{netherlands}} \approx \overrightarrow{\text{amsterdam}}$
- In fact:
  - $\overrightarrow{\text{amsterdam}}$ is in the top-similar word embeddings
  - To the vector $\overrightarrow{\text{paris}} - \overrightarrow{\text{france}} + \overrightarrow{\text{netherlands}}$
  - With regards to **cosine-similarity**
  - When considering embeddings of words in the corpus

University of Amsterdam
Economics & Business

Word2Vec Applied

- How does word meaning evolve over time?



https://arxiv.org/abs/1411.3315 “Statistically Significant Detection of Linguistic Change”, Kulkarni et al. 2015

Document Embeddings

- From **word** embeddings to **document** embeddings
  - TF-IDF weighted sum
  - $\overrightarrow{\mathrm{doc}} = \sum \mathrm{tfidf}(\mathrm{term}, \mathrm{doc}) * \overrightarrow{\mathrm{term}}$

Notebook

- See the Notebook “Word2Vec”

Take-Away

- Word2Vec: 1 word → 1 vector
- **Corpus** dependent:
  - $\overrightarrow{wave}$ with corpus 1 $\neq$ $\overrightarrow{wave}$ with corpus 2
  - Corpus 1 “Maritime Trade”
  - Corpus 2 “Astronomy”

- Cosine similarity between vectors = meaning similarity between words

UNIVERSITY OF AMSTERDAM
Economics & Business

Next Week

- BERT (not exam material)
- Other State-of-the-Art Deep Learning Models (not exam material)

- **Recap = Q&A**

- Bring all your questions !!