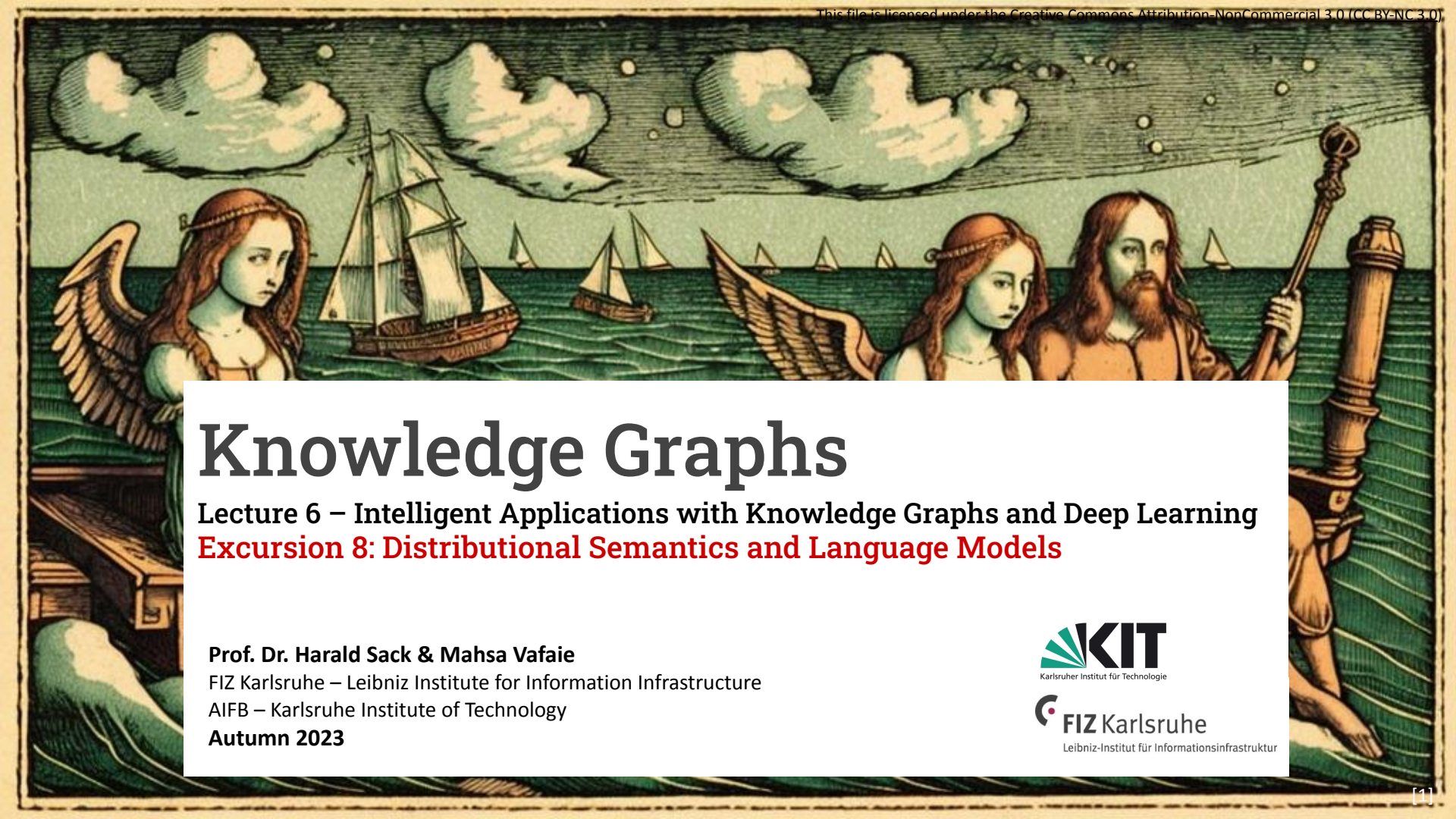


Knowledge Graphs

Lecture 6 – Intelligent Applications with Knowledge Graphs and Deep Learning

Excursion 8: Distributional Semantics and Language Models

Prof. Dr. Harald Sack & Mahsa Vafaie
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
AIFB – Karlsruhe Institute of Technology
Autumn 2023

Karlsruher Institut für Technologie

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Knowledge Graphs

Lecture 6: Intelligent Applications with Knowledge Graphs and Deep Learning

Karlsruher Institut für Technologie

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

[1]

How to represent Natural Language Text?

- For the sake of simplicity we are focussing on the question
  How to represent words in the computer?
- Traditional solution:
  represent words as **unique integers** associated with words:

```
{1=movie, 2=hotel, 3=apple, 4=movies, 5=art}
```

- Equivalent solution: **1-Hot Encoding**

```
movie = [1, 0, 0, 0, 0]
hotel = [0, 1, 0, 0, 0]
. . .
art   = [0, 0, 0, 0, 1]
```

1-Hot Encoding

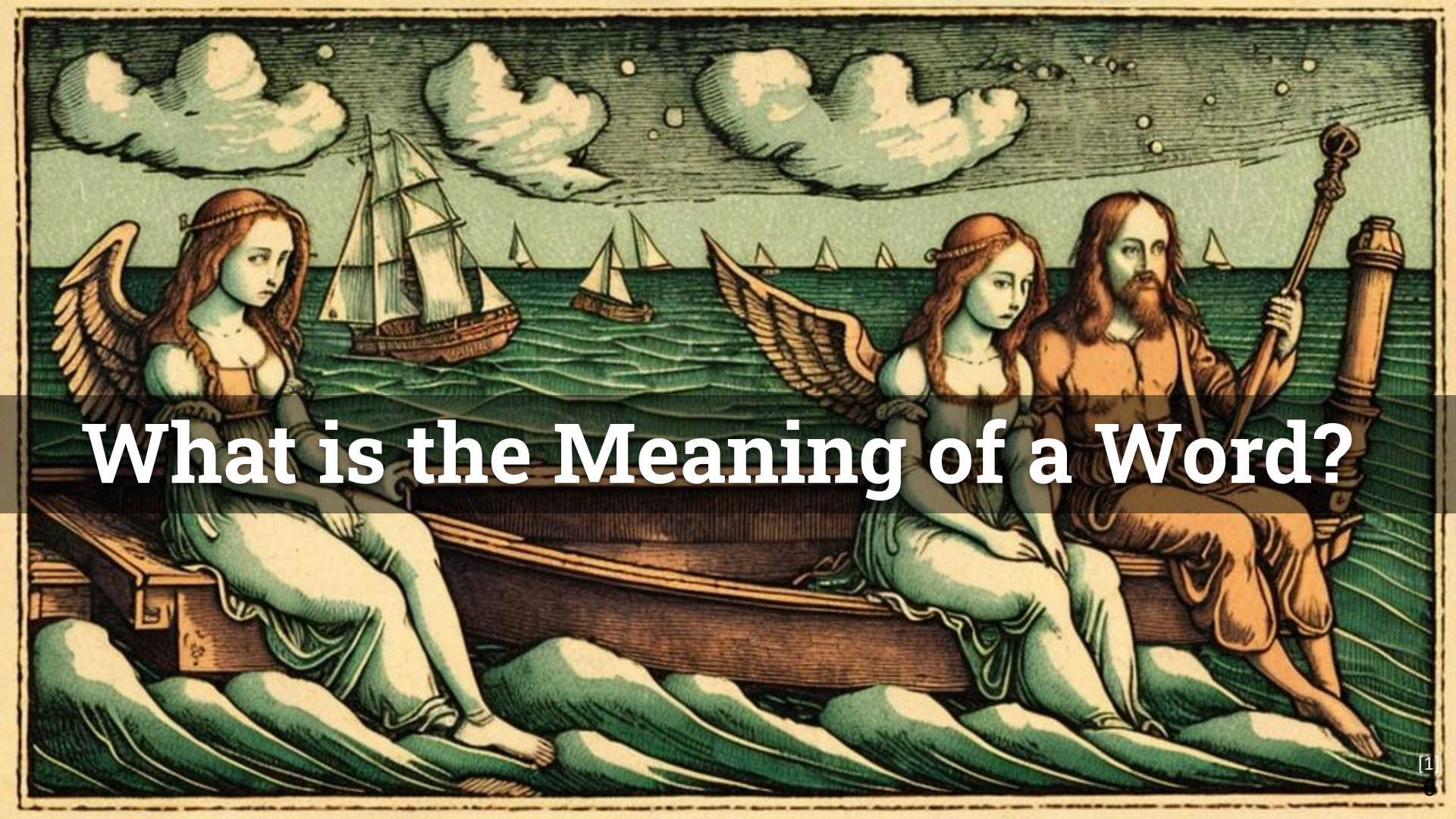
- Most basic representation of any textual unit
- **Vector space**: word vectors constitute an orthogonal base
  - orthogonal ($x^T y = 0$)
  - normalized ($x^T x = 1$)
- **Problem 1:** No relation to semantics
  - E.g. *car* and *automobile* are different (orthogonal) vectors.
  - All words are equidistant:
    $\|cat - dog\| = \|proton - carrier\|$
- **Problem 2**: polysemy
  Should *jaguar (the cat)* have the same vector as *jaguar (the car)*?

Feature Based Representation of Words

- Words can also be represented with handcrafted **features and relations**
- Potential features:
  - Morphological features: *prefix, suffix, stem, lemma, …*
  - Grammatical features: *part-of-speech, gender, number, …*
  - Structural features: *capitalization, hyphen, digit(s), …*
- Potential relations:

  Synonyms, antonyms, hyper- and hyponyms, meronyms and holonyms, ...
- **Problems**:

  Annotation requires high manual effort, annotator disagreement, accuracy, scalability, ...

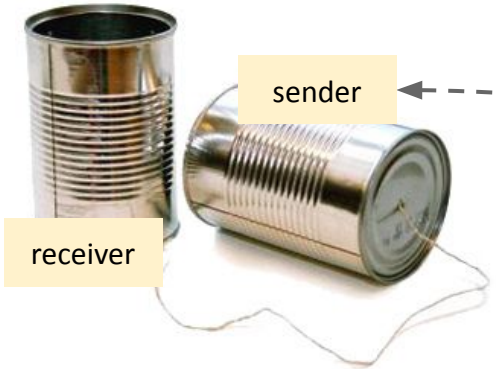


[1]

What is the Meaning of a Word?

Karlsruher Institut für Technologie
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

[2,3,4,5]



Experience

Concept

symbolizes

refers to

Context

Experience

Symbol

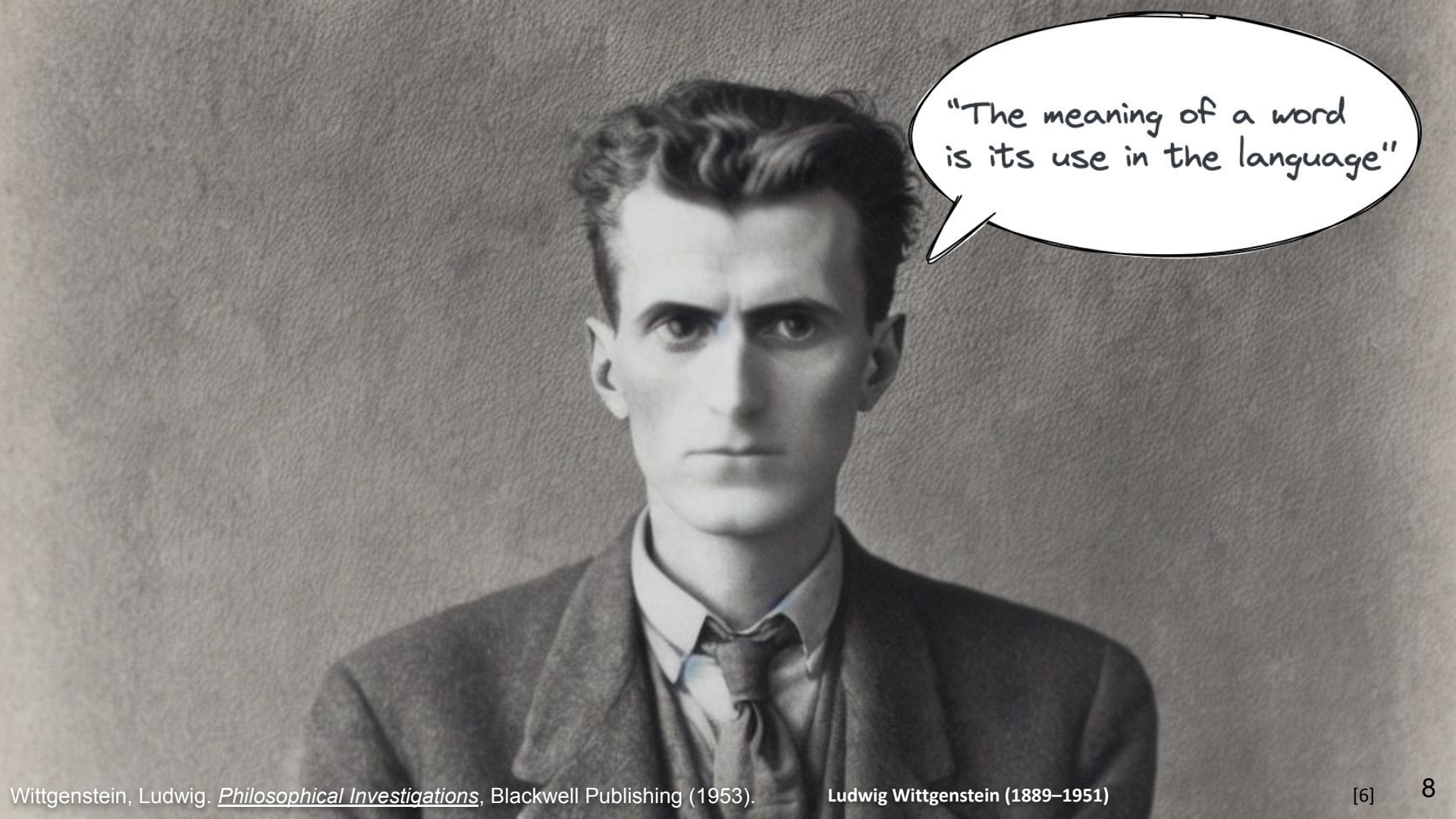
stands for

Object

"Jaguar"

Pragmatics

Ogden, Richards: The **Meaning of Meaning**: A Study of the Influence of Language upon Thought and of the Science of Symbolism, 1923.



Wittgenstein, Ludwig. *Philosophical Investigations*, Blackwell Publishing (1953).

Ludwig Wittgenstein (1889–1951)

[6]

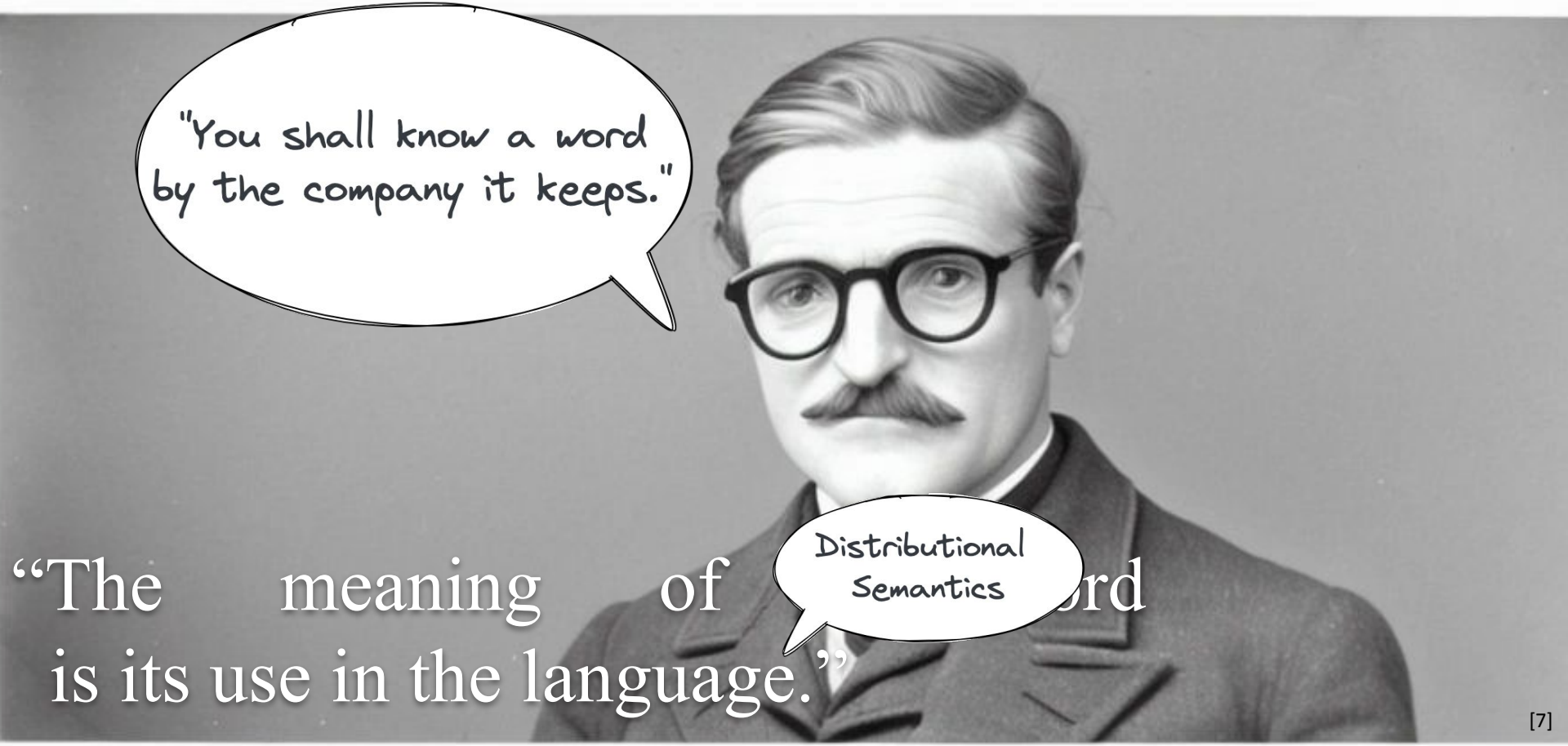
Let's Define Words by their Usage

- In particular, words are defined by their environments (i.e. the words around them).

- *"If [words] A and B have almost* ***identical environments*** *[…] we say that they are* ***synonyms****."*
  Zellig S. Harris (1954)

- Thereby: semantic representations for words can be derived through analysis of patterns of lexical co-occurrence in large language corpora.

Zellig S. Harris (1954) *Distributional Structure*, WORD, 10:2-3, 146-162, DOI: 10.1080/00437956.1954.11659520



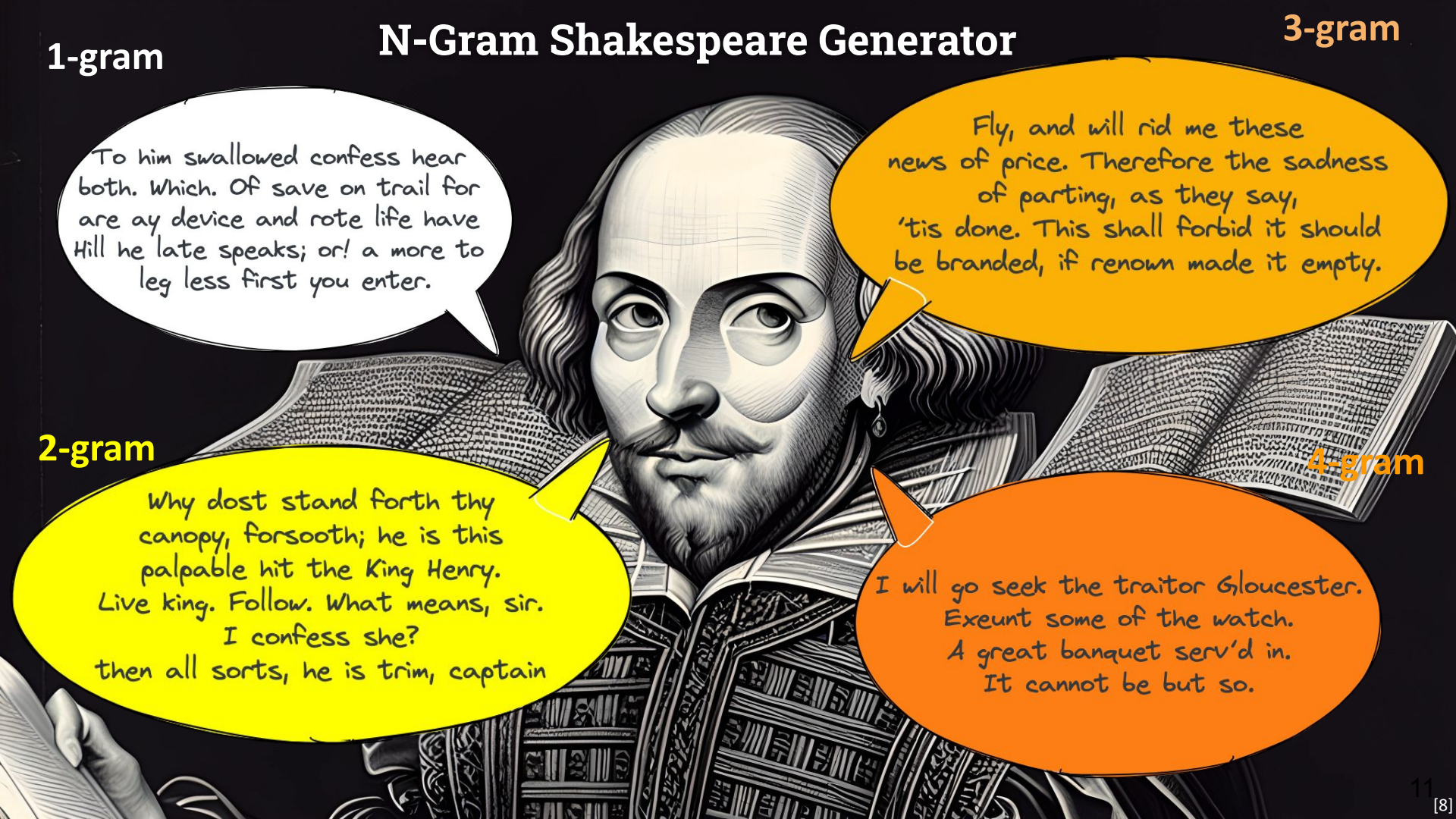
"The meaning of [illegible]ord is its use in the language."

[7]

J.R. Firth (1957) *A synopsis of linguistic theory, Studies in linguistic analysis*, Blackwell, Oxford.

John Rupert Firth (1890–1960)

N-Gram Shakespeare Generator



[8]

Write a dialogue between Caliban and Puck from an unknown Shakespeare play.

Caliban: Puck, wherefore art thou here on this island?

Puck: I am but a messenger, Caliban, sent by the fairy queen to bring magic and mischief to this place.

Caliban: And what manner of magic do you bring?

Puck: Oh, all sorts! I can change shapes, cast spells, and make illusions. And I have a particular talent for causing trouble for those who cross me.

Caliban: I see. And what do you know of me?

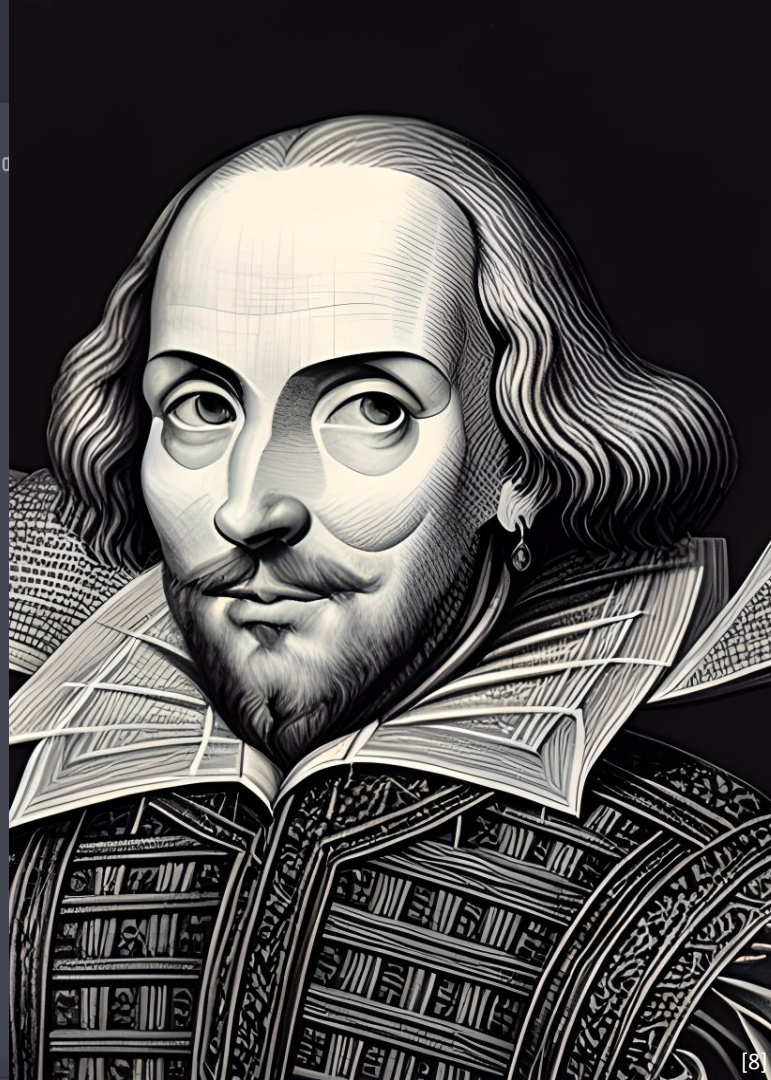
Puck: I know that thou art Prospero's servant, Caliban, and that thou art unhappy with thy lot in life.

Caliban: Aye, that is true. I am a slave to Prospero, forced to do his bidding and denied my freedom.

Puck: Well, perhaps I can help thee with that. I can cause mischief for Prospero and make it easier for you to escape his grasp.

Caliban: Really? You would do that for me?

Puck: I would indeed. But thou must be willing to trust in me and do as I say



https://chat.openai.com/chat

What does “Ong Choi” mean?

- Suppose you see these sentences:
  - **Ong choi** is delicious sautéed with garlic.
  - **Ong choi** is superb over rice.
  - **Ong choi** leaves with salty sauces...
- And you've also seen these:
  - **…spinach** sautéed with garlic over rice.
  - **Chard** stems and leaves are delicious.
  - **Collard greens** and other salty leafy greens...
- Conclusion:

  Ong choi is a **leafy green** like **spinach**, **chard**, or **collard greens**.

Ong choi: *Ipomoea aquatica* “Water Spinach”

https://commons.wikimedia.org/wiki/File:Ipomoea_aquatica.jpg

Distributional Semantics

- A word's meaning is given by the words that frequently appear close-by.
- When a word w appears in a text, its **context** is the set of words that appear nearby (within a fixed-size window).
- Use the different contexts of w to build up a representation of w.

Though quite agile on land, **capybaras** are equally at home in the water.
A giant cavy rodent native to South America, the **capybara** actually is the largest living rodent.

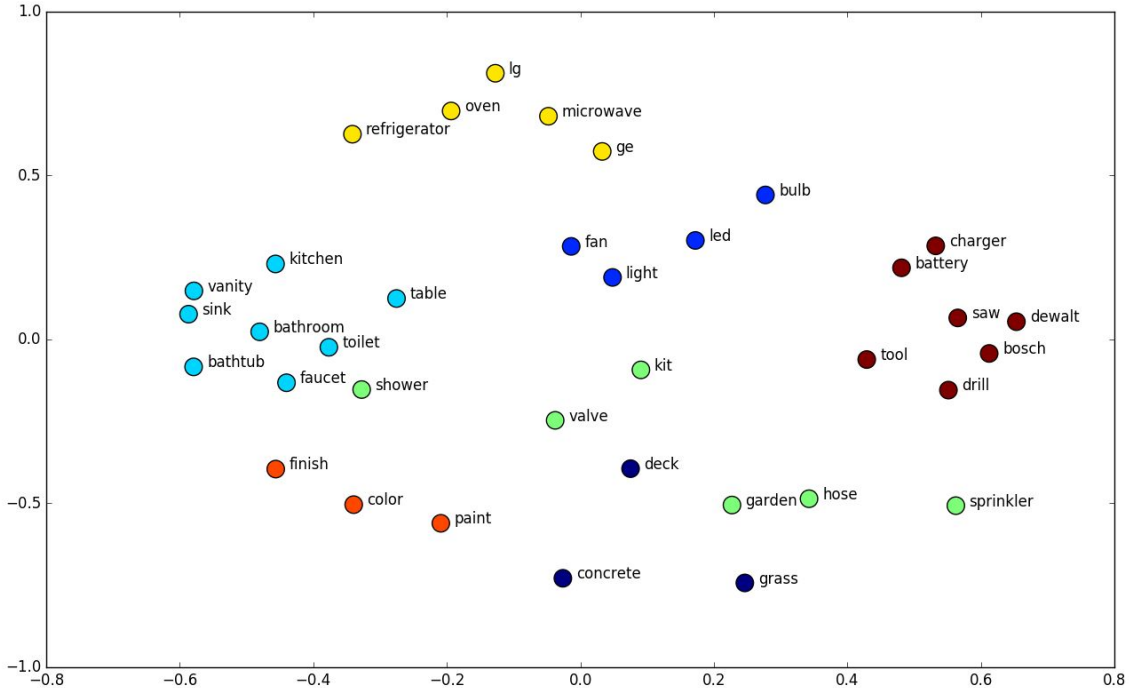
Word Vectors

- We will build a **dense vector** for each word, so that it is similar to vectors of words that appear in similar contexts.

$$\text{capybara} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

- **Word vectors** are a distributed representation. They are also referred to as **word embeddings** or word representations.

Word Vectors



- Combines **distributional semantics** (statistical language model) and **vector intuition.**
- Semantically similar words are nearby in a vector space.
- Called an "**embedding**" because it's embedded into a vector space.
- The standard way to represent meaning in NLP.

Word2Vec

Overview

- **Word2Vec** (Mikolov et al. 2013) is a framework for learning word vectors.

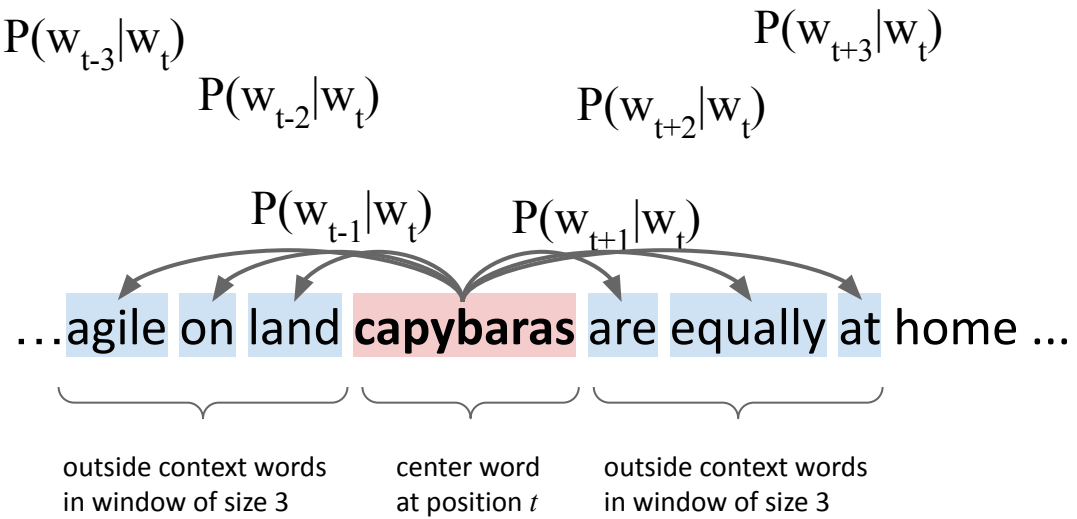
- Operating Principle:
  - We need to have a large corpus of text.
  - Every word in a fixed vocabulary is represented by a vector.
  - Go through **each position t** in the text, which has a **center word c** and **context ("outside") words o**.
  - Use the **similarity** of the word vectors for c and o to **calculate the probability of o given c** (or vice versa).
  - Keep adjusting the word vectors to maximize this probability.

Mikolov, Tomas; et al. (2013). "*Efficient Estimation of Word Representations in Vector Space*". arXiv:1301.3781

Word2Vec

Overview

Example windows and process for computing $P(w_{t+j}|w_t)$

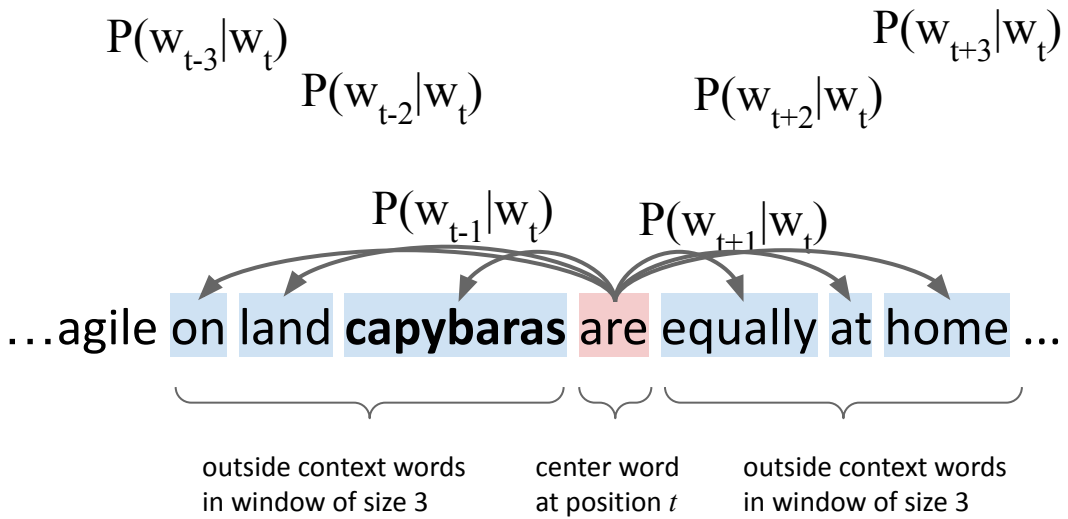


Mikolov, Tomas; et al. (2013). "*Efficient Estimation of Word Representations in Vector Space*". arXiv:1301.3781

Word2Vec

Overview

Example windows and process for computing $P(w_{t+j}|w_t)$

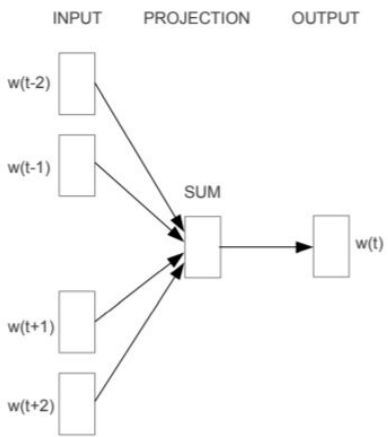


Mikolov, Tomas; et al. (2013). "*Efficient Estimation of Word Representations in Vector Space*". arXiv:1301.3781

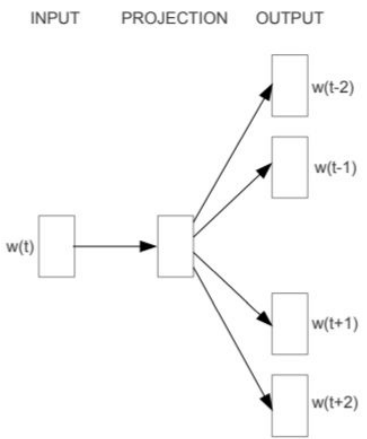
Word2Vec

Variants

- Word2Vec maximizes the objective function by putting similar words nearby in vector space.
- Two model variants:
  - **Skip-gram (SG):**
    Predict (sequence independent) context words given the center word.
  - **Continuous Bag of Words (CBOW):**
    Predict center word from (bag of) context words.



CBOW



Skip-gram

Mikolov, Tomas; et al. (2013). "*Efficient Estimation of Word Representations in Vector Space*". arXiv:1301.3781

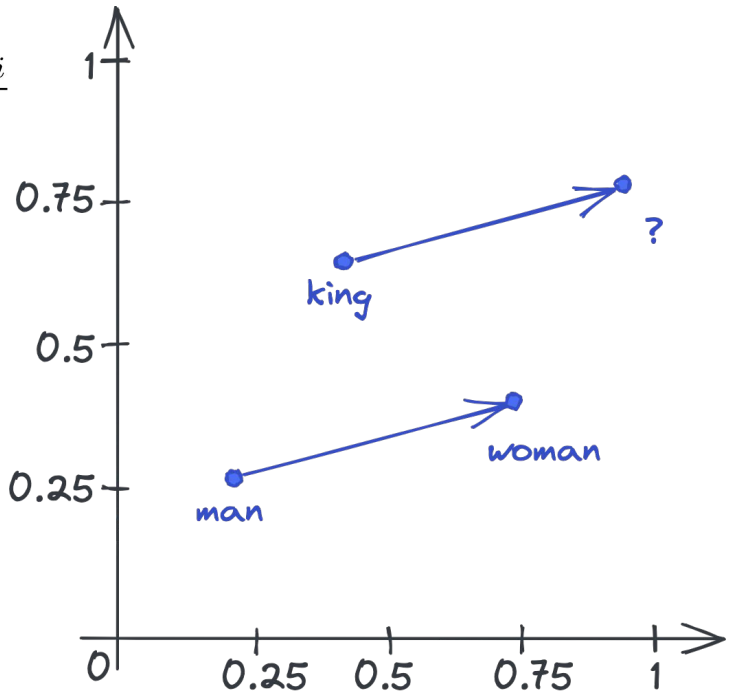
Word2Vec

Intrinsic Evaluation of Word Vectors

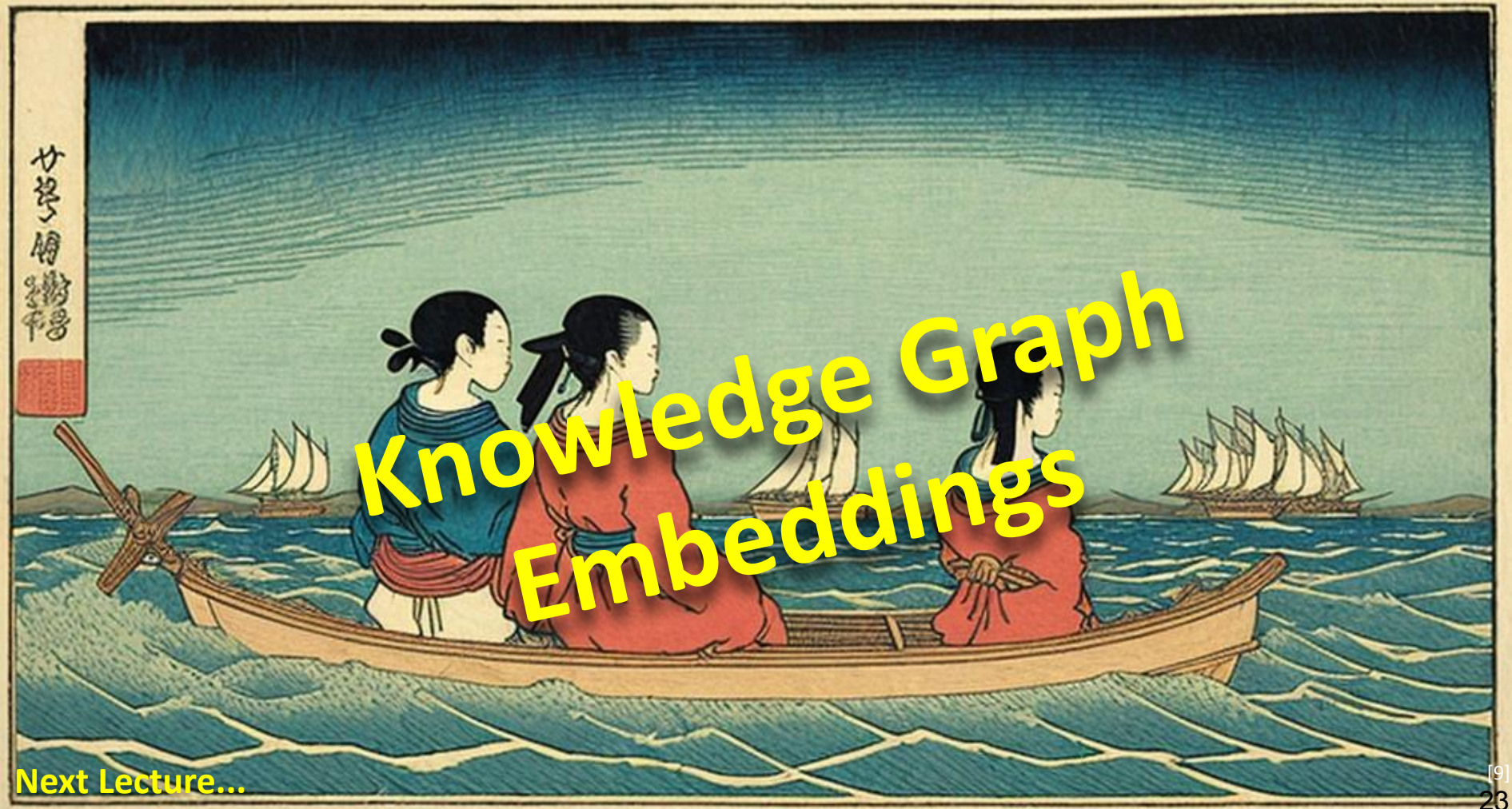
- Word Vector Analogies

$$a{:}b :: c{:}d \longrightarrow d = \arg\max_i \frac{(x_b - x_a + x_c)^T x_i}{||x_b - x_a + x_c||}$$

man : woman = king : ?



- Evaluate word vectors by how well their cosine distance after addition captures **intuitive semantic** and **syntactic analogy questions**.
- Discarding the input words from the search!
- Problem: What if the information is there but not linear?



Next Lecture...

[9]

Bibliographic References:

- Ogden, Charles Kay, & Richards, Ivor Armstrong (1923). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*, K. Paul, Trench, Trubner & co., ltd, Harcourt, Brace & company, inc.
- Wittgenstein, Ludwig (1953). *Philosophical Investigations*, Blackwell Publishing.
- Zellig S. Harris (1954) *Distributional Structure*, WORD, 10:2-3, 146-162, DOI: 10.1080/00437956.1954.11659520
- J.R. Firth (1957) *A synopsis of linguistic theory, Studies in linguistic analysis*, Blackwell, Oxford.
- Mikolov, Tomas; et al. (2013). "*Efficient Estimation of Word Representations in Vector Space*". arXiv:1301.3781

Picture References:

[1] "On this colorized woodcut in the style of Albrecht Dürer we see a pensive cupid together with a beautiful female angel, bothe are melancholically watching two sailing ships on the vast ocean …, created via ArtBot, Deliberate, 2023, [CC-BY-4.0], https://tinybots.net/artbot

[2] Ch. Meinel, H. Sack: *Digital Communication - Communication Multimedia, Security*, Springer, 2014.

[3] Phone Conversation Indicating Chit Chat And Discussion, [CC0], https://www.pigsels.com/en/public-domain-photo-jrdox

[4] Car Jaguar Vehicle, OpenClipart-Vectors (pixabay.com), https://www.needpix.com/photo/101821/car-jaguar-vehicle-automobile-transportation

[5] Felis onca, Geoffroy-Saint-Hilaire & Cuvier, Histoire naturelle des mammifères, pl. 170, [public domain], https://commons.wikimedia.org/wiki/File:Felis_onca_-_1818-1842_-_Print_-_Iconographia_Zoologica_-_Special_Collections_University_of_Amsterdam_-_(white_background).jpg

[6] "Historical portrait photography of philosopher Ludwig Wittgenstein as young man"., created via ArtBot, ProtoGen, 2022, [CC-BY-4.0], https://tinybots.net/artbot

[7] "A 1950s photography of J. R. Firth aged 55, with glasses and a white toothbrush mustache, English linguist and leading figure in British linguistics"", created via ArtBot, ProtoGen, 2022, [CC-BY-4.0], https://tinybots.net/artbot

[8] "Create an image in the style of a renaissance engraving with a portrait of William Shakespeare", created via ArtBot, ProtoGen, 2022, [CC-BY-4.0], https://tinybots.net/artbot

[9] On this colorized woodcut in the style of Hiroshige we see a pensive cupid together with a beautiful female angel, bothe are melancholically watching two sailing ships ….", created via ArtBot, Deliberate, 2023, [CC-BY-4.0], https://tinybots.net/artbot