Understanding Embeddings for NLP

Mario Tormo Romero

Design IT.
Create Knowledge.

www.hpi.de

KISZ BB

The AI Service Center Berlin-Brandenburg (KISZ-BB) is a project of the Hasso-Plattner-Institute with the aim of lowering barriers to the use of AI in business and society through knowledge transfer and networking. The main research areas are operational research to investigate an AI data center with heterogeneous hardware and methodological research to adapt and optimize AI models. The KISZ-BB provides resources such as computing power, storage space, data and models for the development and use of AI applications. The KISZ-BB also offers educational and consulting services in the form of workshops, individual consultations and online courses. In this way, companies, start-ups and non-profit institutions are supported in successfully mastering the next steps towards the professionalization of AI applications.

KI Service Zentrum by Hasso-Plattner-Institut

HPI

SPONSORED BY THE

Federal Ministry of Education and Research

05.12.2023

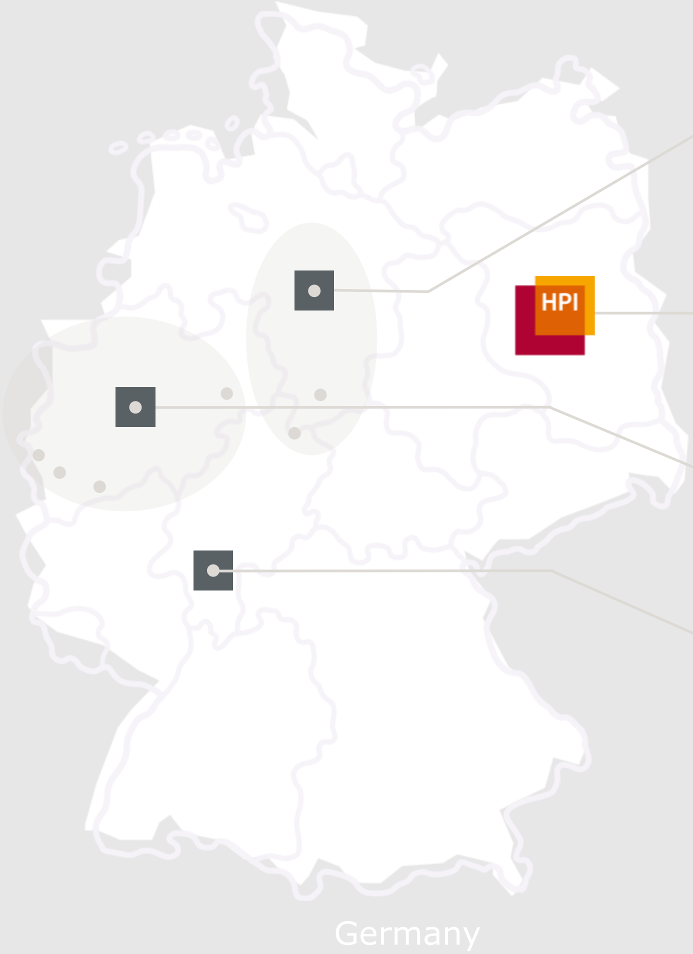
KISZ-BB

KI Service Zentrum
by Hasso-Plattner-Institut

SPONSORED BY THE
Federal Ministry of Education and Research

AI service centers

MOTIVATION, TASK & SERVICES



KISSKI
Hannover I Göttingen I Kassel
Sensitive and critical infrastructures
Focus: Medicine & Energy

KISZ-BB
Berlin I Brandenburg
Educational and advisory services
Use of AI in business and society

West AI
Dortmund I Bonn I Jülich I Aachen I Paderborn
Large and transferable AI models

hessian AI
Darmstadt
Explainability, generalizability and contextual adaptation

Service Center

KI Service Zentrum by Hasso-Plattner-Institut

HPI

SPONSORED BY THE Federal Ministry of Education and Research

Learning Goals

- Understand the challenges of converting unstructured text into numerical data for ML/DL/AI.
- Explore the evolution of solutions for representation learning, through a spectrum of embedding techniques, from historical approaches to modern algorithms.
- Recognize the significance of vector databases in storing and querying embeddings, and their advantages over traditional databases when dealing with embeddings.

KI Service Zentrum by Hasso-Plattner-Institut

HPI

SPONSORED BY THE Federal Ministry of Education and Research

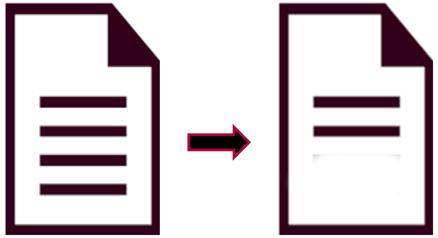
Agenda

1. Turning text into numbers
2. Improving the representations
   (with a small interlude)
3. Storing embeddings
4. Conclusion

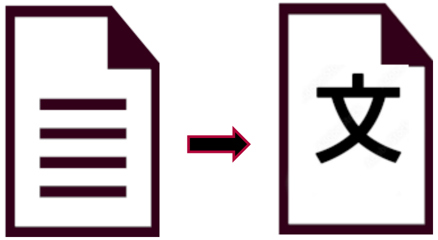
Part 1: Turning text into numbers

Why do we want to work with texts?

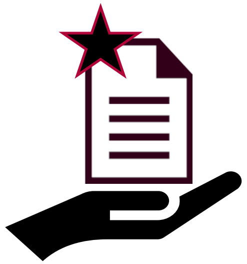
Information Retrieval



Document Summarization

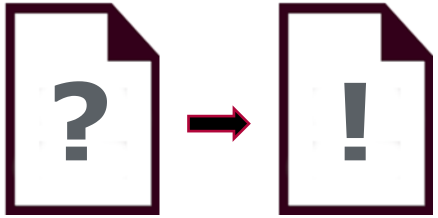


Language Translation



Content Recommendation

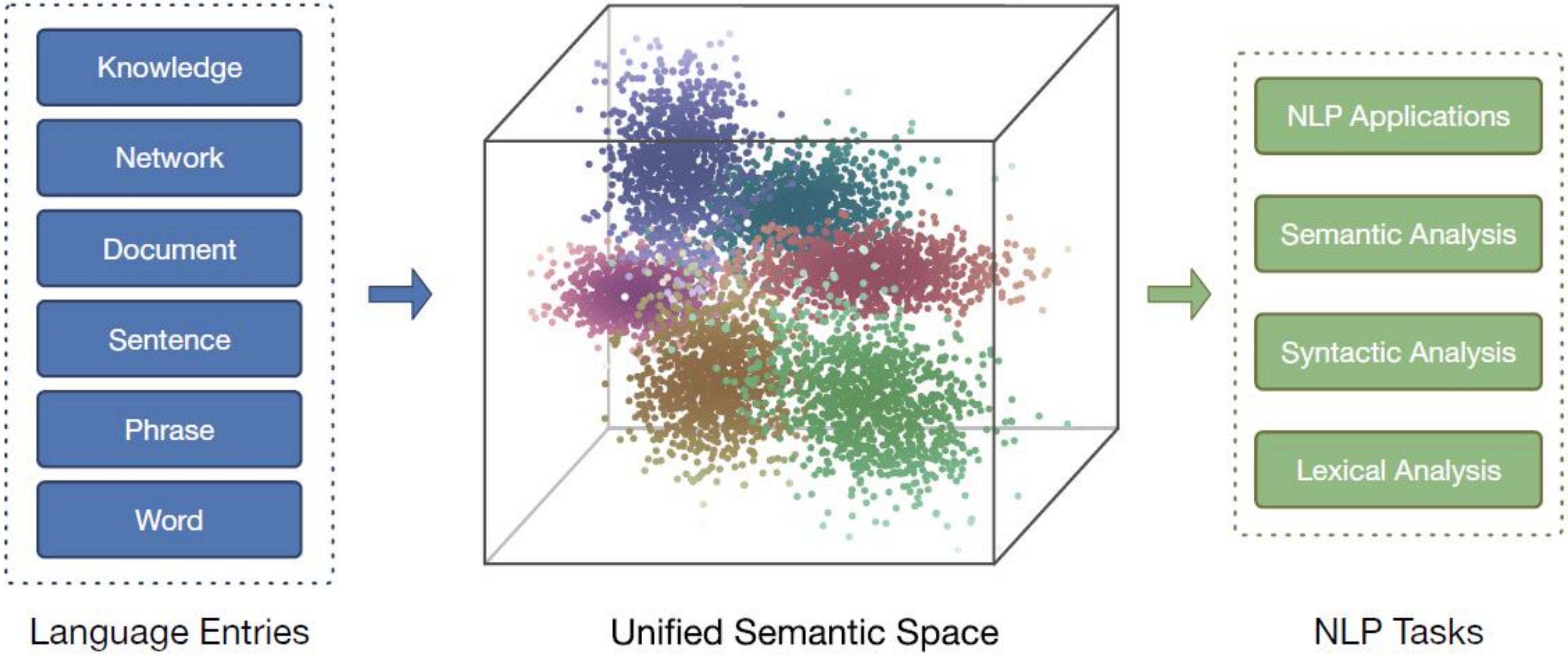
Sentiment Analysis



Question-Answering Systems

Data representation

with different depth levels | through different domains | for different tasks



KI Service Zentrum by Hasso-Plattner-Institut
SPONSORED BY THE Federal Ministry of Education and Research

Tokenization

In a hole in the ground there lived a hobbit.

Tokens are not always words. They can be

Bytes	Characters	Subwords or word pieces	Full words and their roots	Sentence pieces

Typical challenges during tokenization include

Contractions

I‘m, you‘ve, he‘s, Matt‘s

Stop words

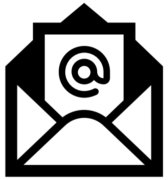
the, a, it, this, that

Languages with unclear word boundaries

姚明进入总决赛

“Yao Ming reaches the finals” in Chinese

Historical approaches: Rule-based systems



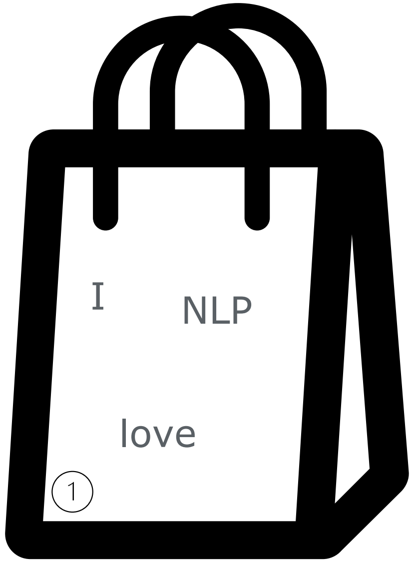
- Includes the words „free“, „money“, „winner“, …
- Includes words written completely with capital letters
- Includes excessive punctuation like „!!!“, “$$$” or “???”
- Has an inconsistent sender email address
- Includes an attachment of an .exe, .zip or .js file
- Has unusual character encoding or mixes multiple character sets

$$\cancel{\begin{Bmatrix}1\\1\\1\\0\\1\\0\\0\end{Bmatrix}} \Rightarrow \begin{Bmatrix}5\\3\\3\\0\\1\\0\\0\end{Bmatrix}$$

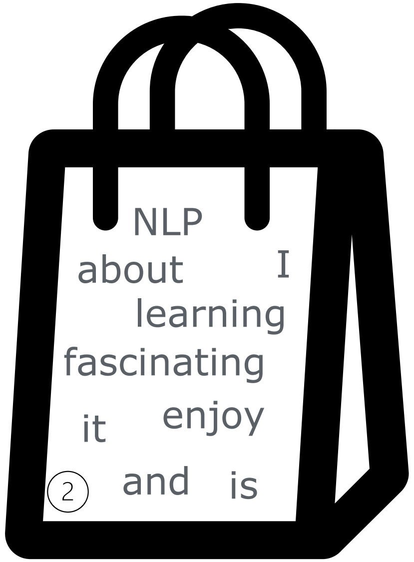
Limited generalization
No scalability / Constant Maintenance
Lack of contextual understanding
Subjectivity and Bias

Historical approaches: Bag of Words

Document 1: "I love NLP.“
Document 2: "NLP is fascinating, and I enjoy learning about it."



$$\begin{Bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{Bmatrix}$$



$$\begin{Bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{Bmatrix} \Rightarrow \begin{Bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{Bmatrix}$$

Vocabulary: ["I", "love", "NLP", "is", "fascinating", "and", "enjoy", "learning", "about", "it"]

Historical approaches: Bag of Words

Advantages

- Simplicity
- Efficiency
- Language agnostic
- Interpretability
- Useful for certain tasks

Disadvantages

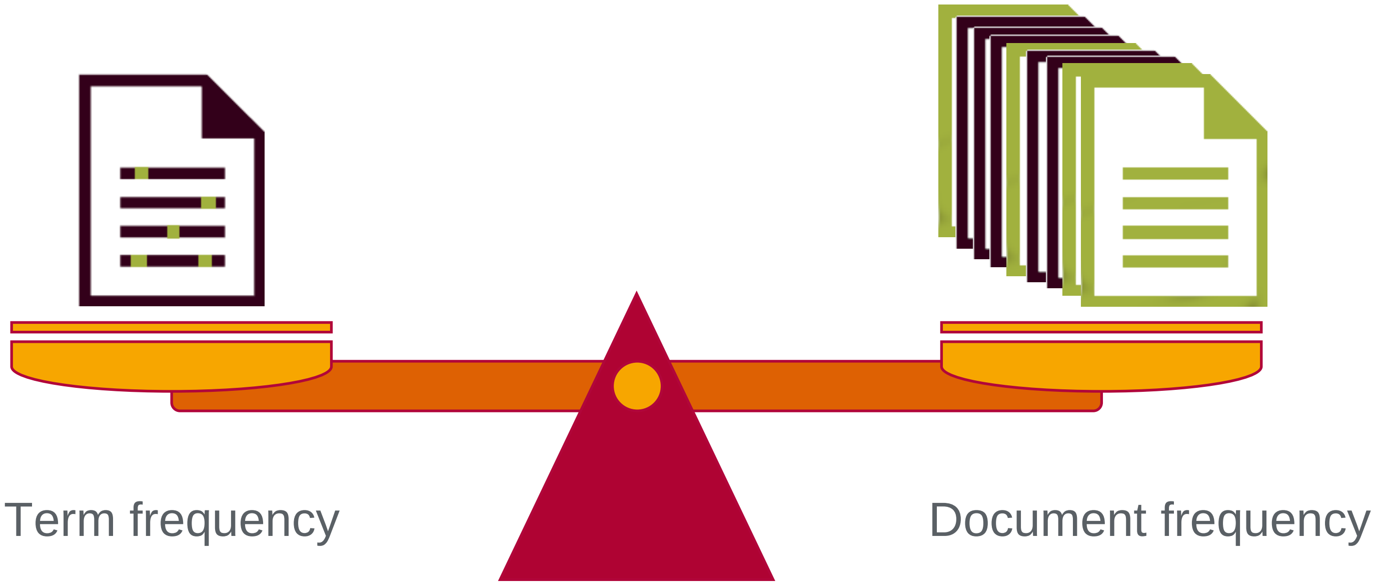
- Loss of sequence Information
- Fixed Vocabulary size
- Equal importance
- Inefficiency with large datasets
- Out of vocabulary words

KI Service Zentrum by Hasso-Plattner-Institut

HPI

SPONSORED BY THE Federal Ministry of Education and Research

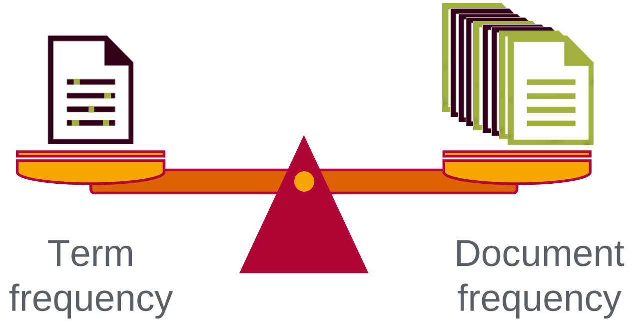
Historical approaches: tf-idf



$$\left\{\begin{array}{rrr} 0.14 & 0.51 & 0 \\ 0.02 & 0.01 & 0.02 \\ 0.21 & 0.13 & 0 \\ 0 & 0 & 0.65 \\ 0 & 0 & 0.22 \\ 0 & 0.59 & 0.27 \\ 0.05 & 0.03 & 0.03 \\ 0.62 & 0 & 0.27 \\ 0 & 0.14 & 0 \\ 0.77 & 0.30 & 0.14 \end{array}\right\}$$

Historical approaches: tf-idf

Advantages	Disadvantages
• Content relevance • Flexibility • Reduce common words • Language agnostic • Weighted representation	• Sparse vectors • Sensitivity to text length • Manual tuning required • Doesn‘t handle misspellings • Ignores semantic meaning



Part 2: Improving the representations

Lexical semantics

Multiple meanings (polysemy)

Synonyms

Word similarity

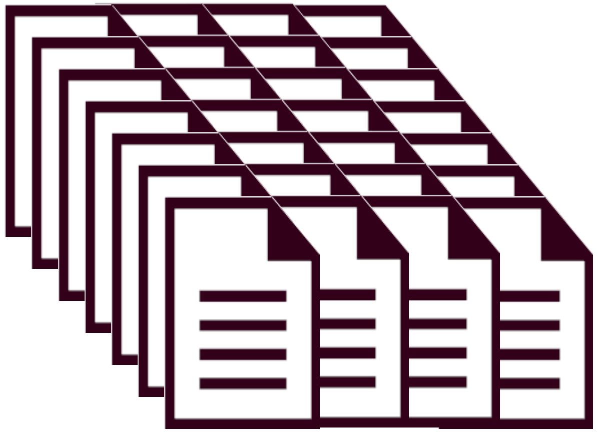
Word relatedness

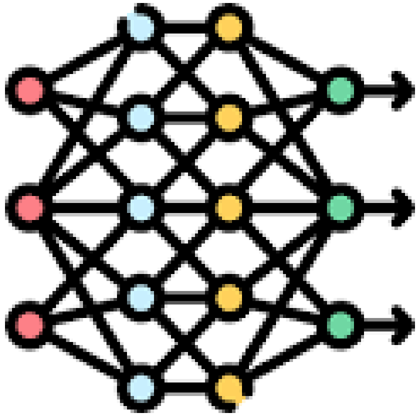
Semantic frames and roles

Connotations

The concept of embedding

$$\left\{\begin{matrix} 0.128 \\ 0.233 \\ 0.007 \\ 0.134 \\ 0.655 \\ 0.912 \\ 0.031 \\ 0.291 \\ 0.367 \\ 0.049 \end{matrix}\right\}$$



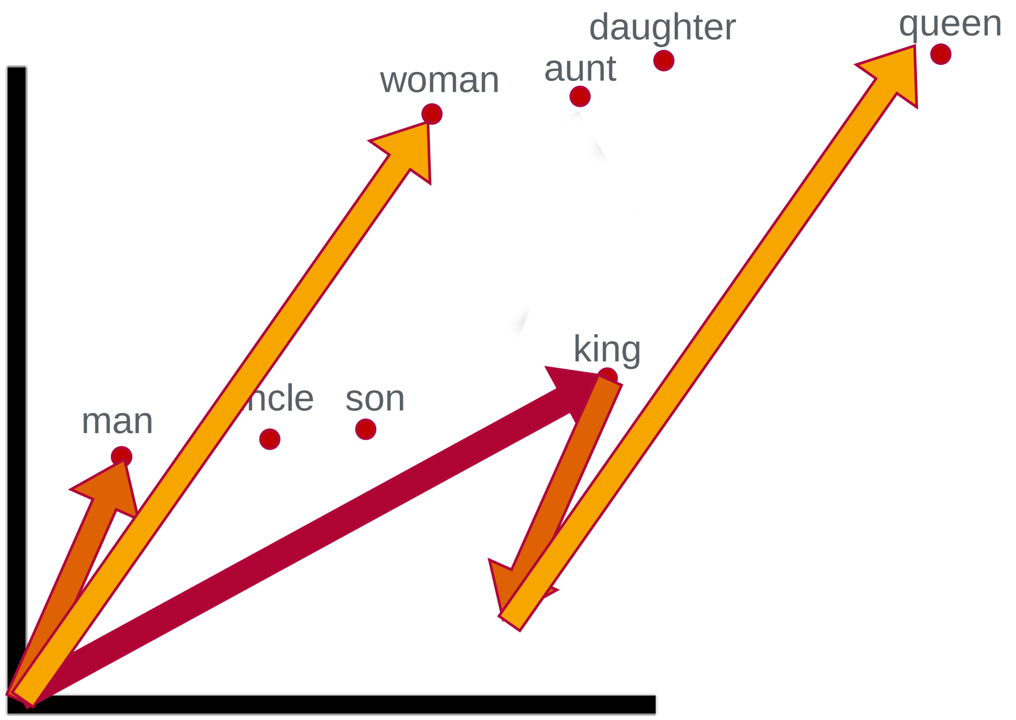


Analogy questions

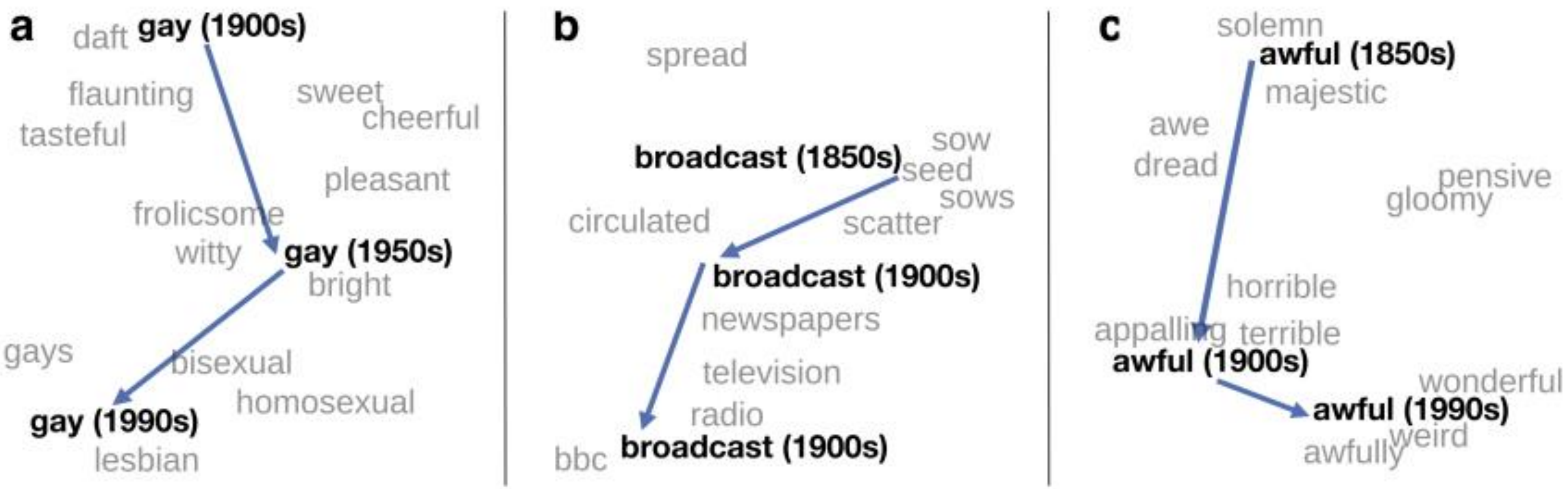
Who is to mathematics what Albert Einstein is to physics?

Which word is to woman what king is to man?

king – man + woman = …



Historical semantics



A small thing about the learned semantics

- the learned semantics don‘t necessarily correspond to the interpretation that we give to those words

- those semantics are learnt from millions of texts, mostly from the Internet

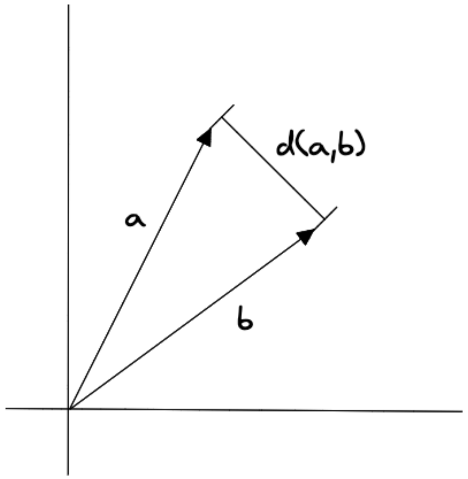
- they represent the average meaning of the texts that we have used for creating those representations

- the bias and prejudices present in the texts are also contained in our representations

Interlude: Metrics and Visualization

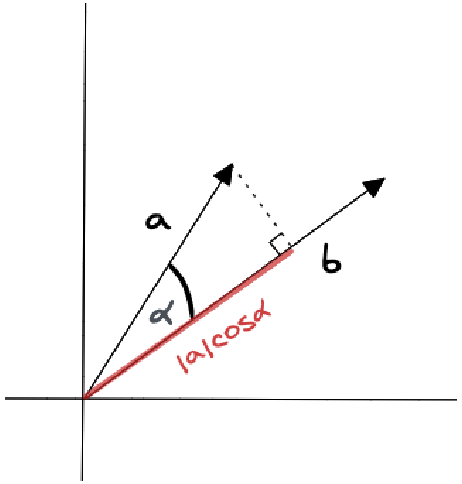
Vector comparison

Euclidean distance



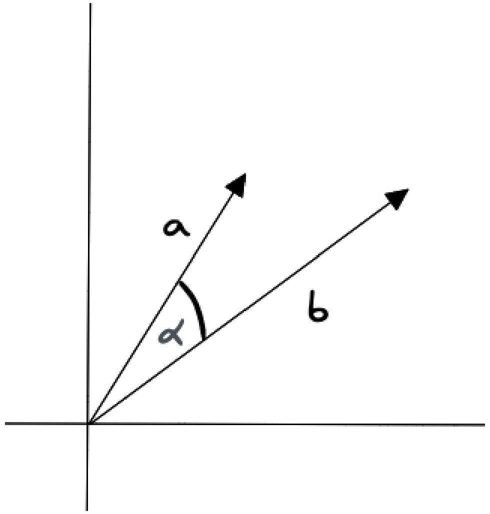
Distance between ends of vectors

Dot product similarity



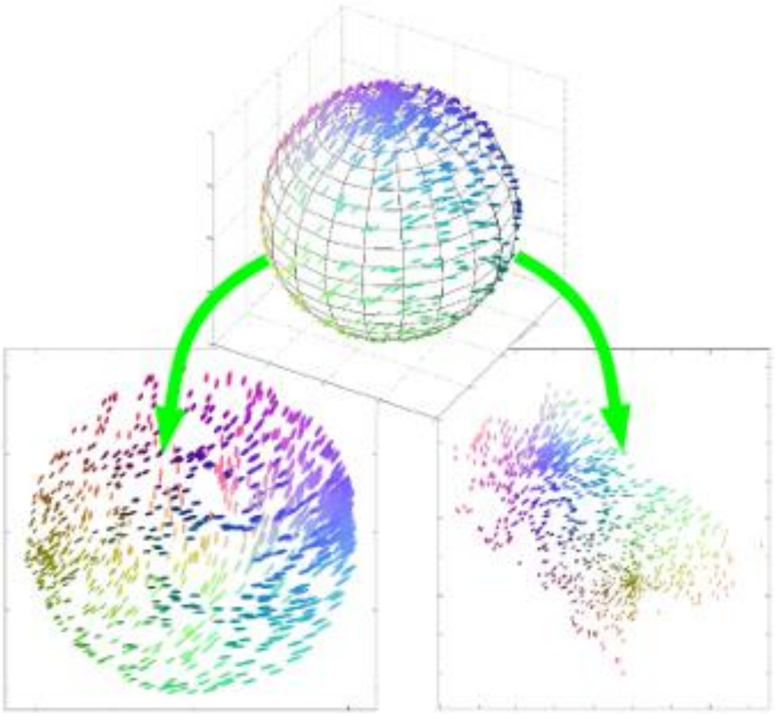
Product of the lengths of the projected vectors

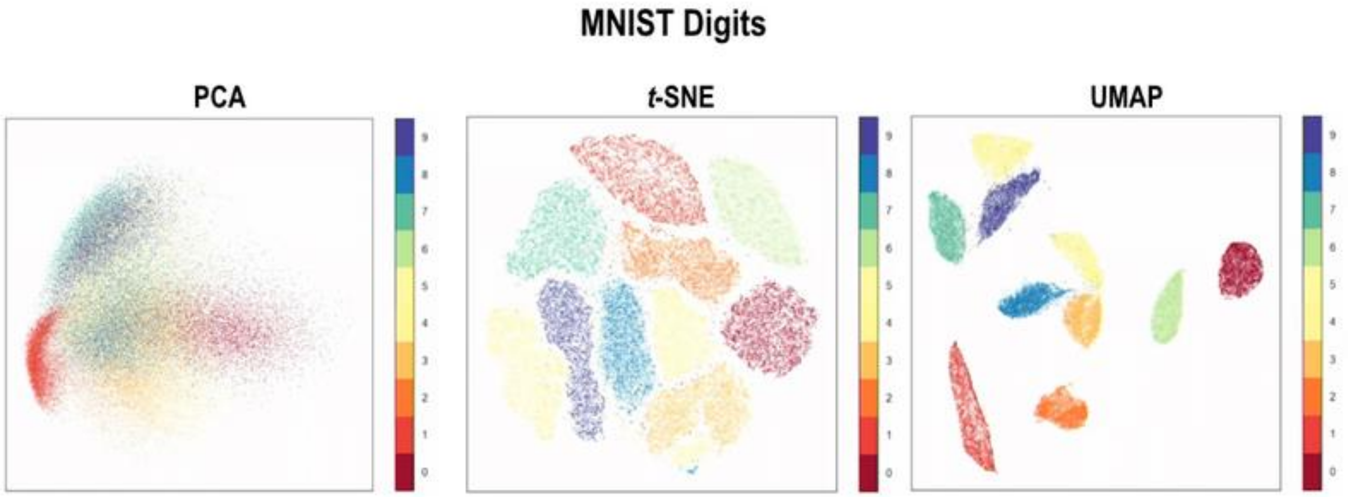
Cosine similarity



Cosine of angle θ between vectors

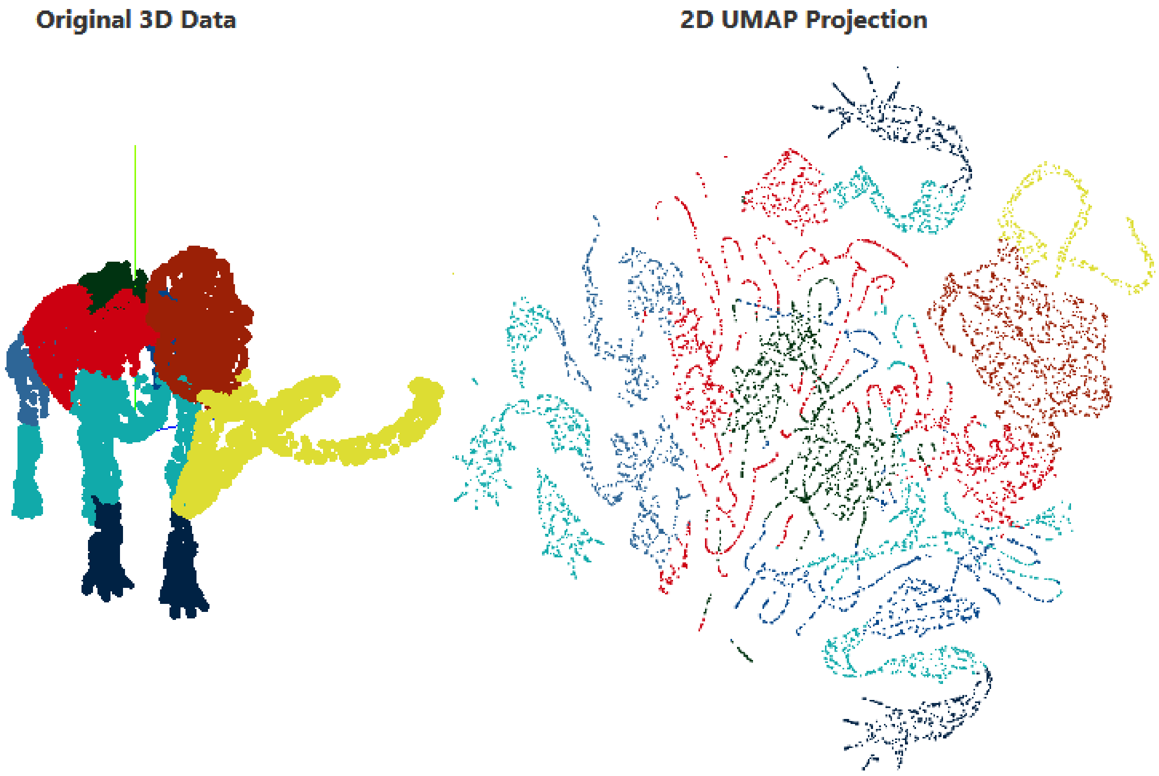
Vector visualization





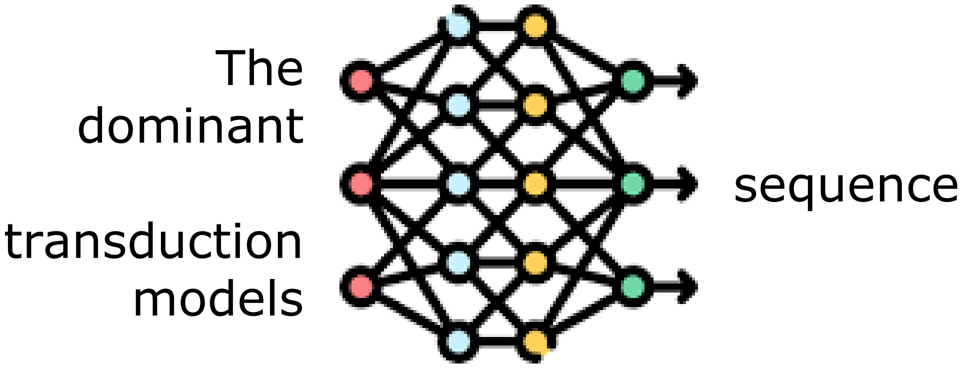
Limitations of these techniques

- **Information Loss**
- **Overcrowding and Clutter**
- **Interpretation Challenges**
- **Scalability Issues**
- **Subjectivity in Interpretation**
- **Algorithm Sensitivity**

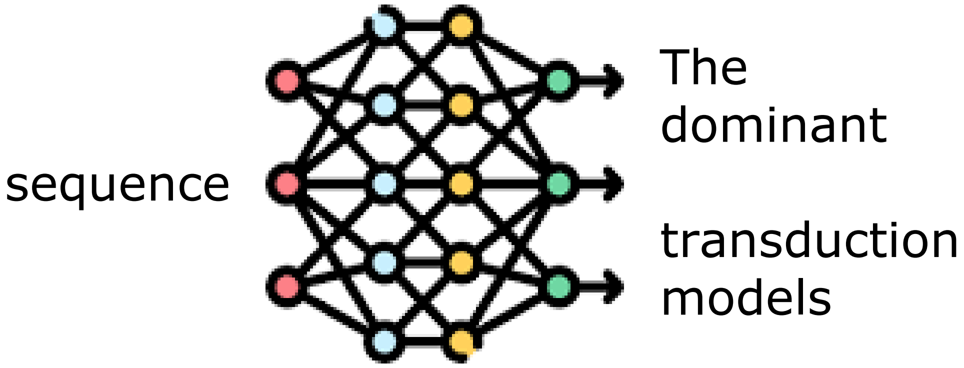


Part 2: Improving the representations (continued)

word2vec



Continuous Bag-of-Words (CBOW)



Skip-Gram

\+

Negative sampling

Sub Sampling

KI Service Zentrum
by Hasso-Plattner-Institut
HPI
SPONSORED BY THE
Federal Ministry of Education and Research

Other similar embeddings

GloVe

- Emphasizes Semantic Meaning
- Balances Global and Local Context

- Contextual Understanding in Large Corpora
- Information Retrieval and Search Engines

FastText

- Handling Out-of-Vocabulary Words
- Enhanced Understanding of Morphologically Rich Languages

- Misspelling Correction and Social Media Analysis
- Morphologically Complex Languages

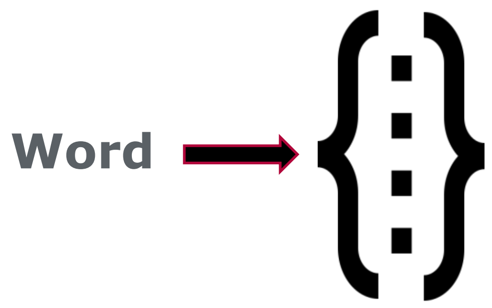
Doc2Vec

- Document-Level Representations
- Unsupervised Learning of Document Embeddings

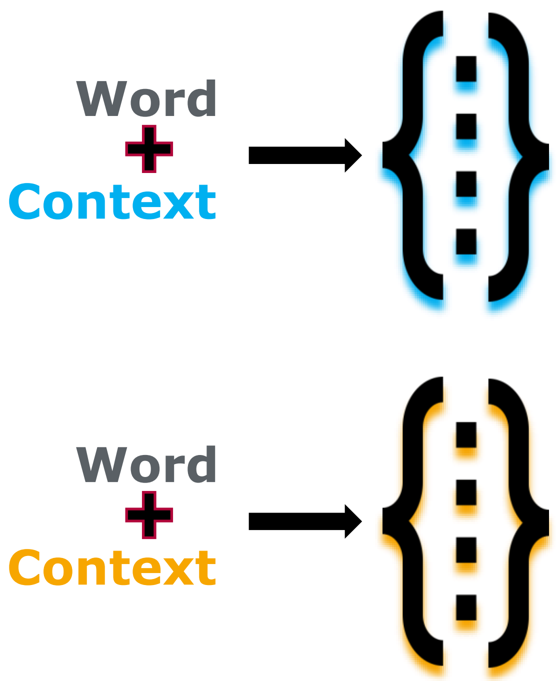
- Document Clustering and Information Retrieval
- Personalized Content Recommendation

Static vs contextual embeddings

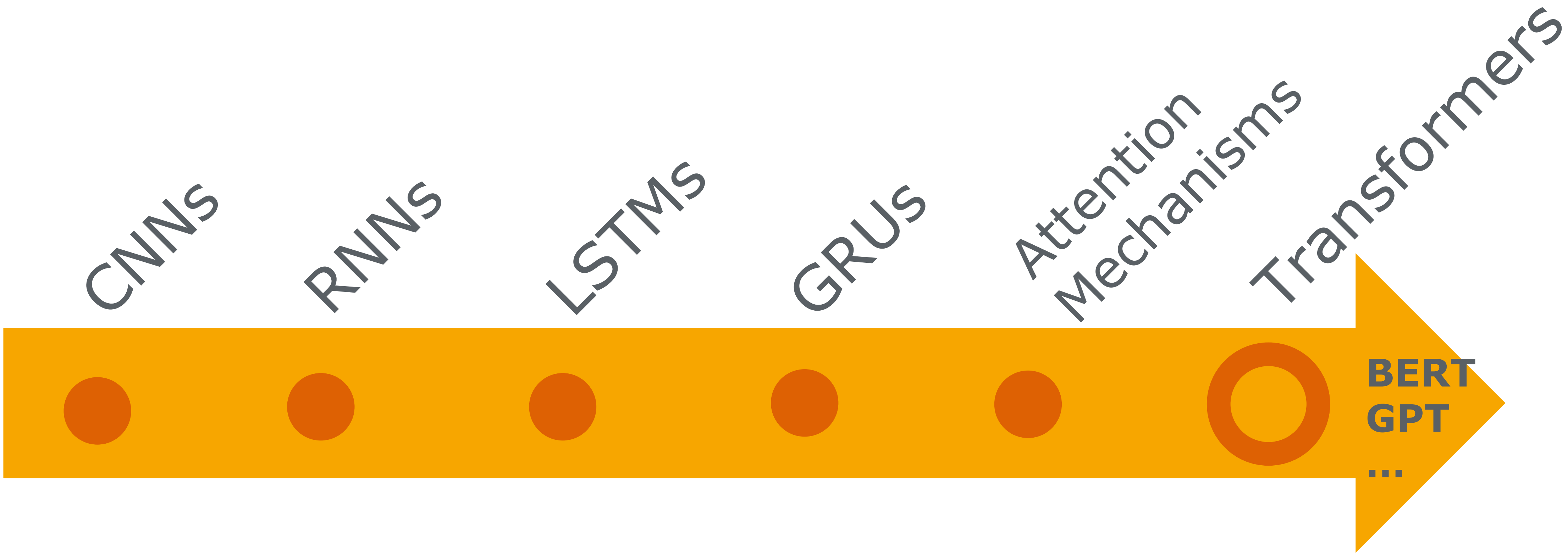
Static embeddings



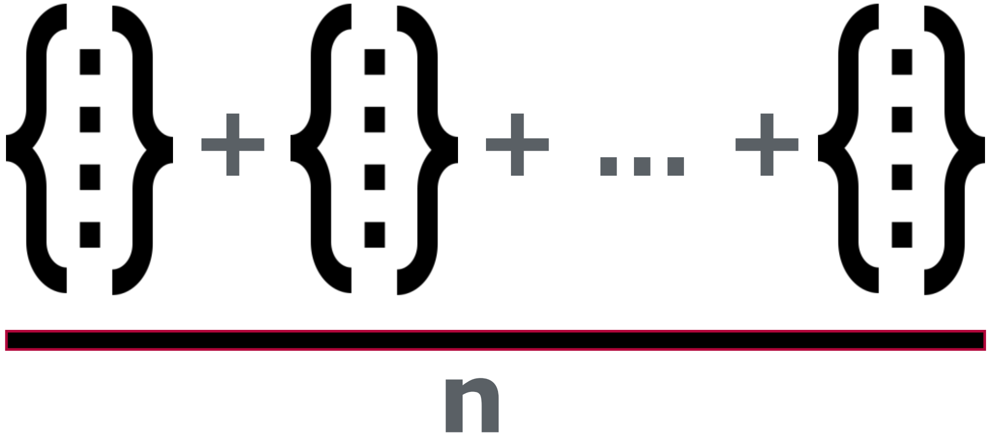
Contextualized embeddings



Development of more complex embeddings

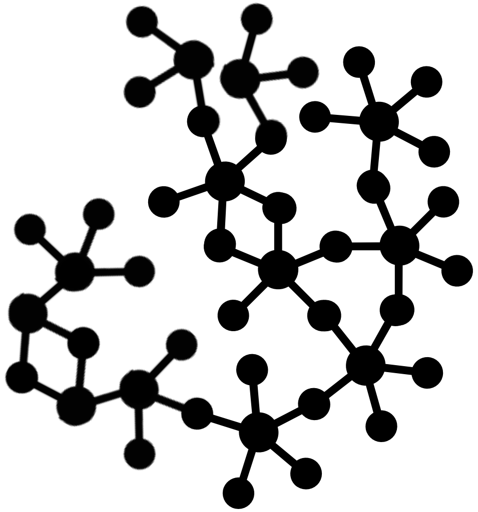


Sentence embeddings and sentence transformers



SBERT.net

Getting better embeddings

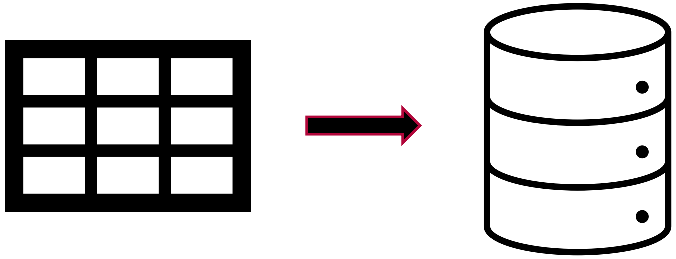


Knowledge graphs

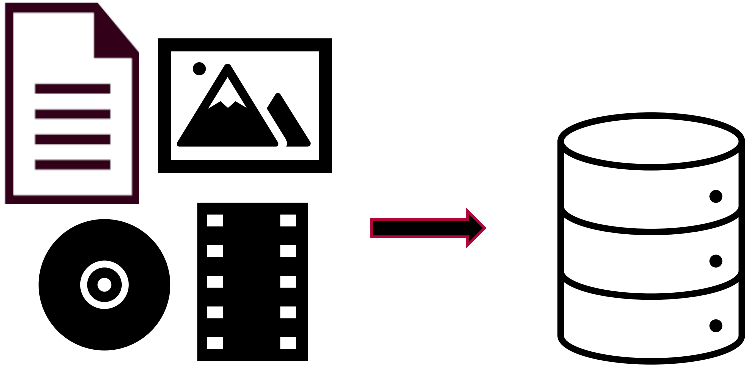
Multimodality

Part 3: Storing embeddings

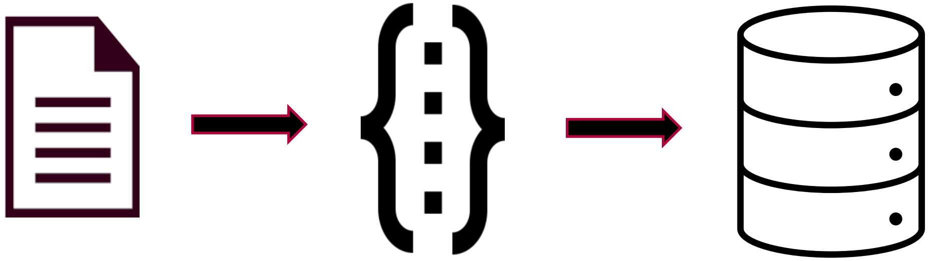
Vector databases



Relational Databases

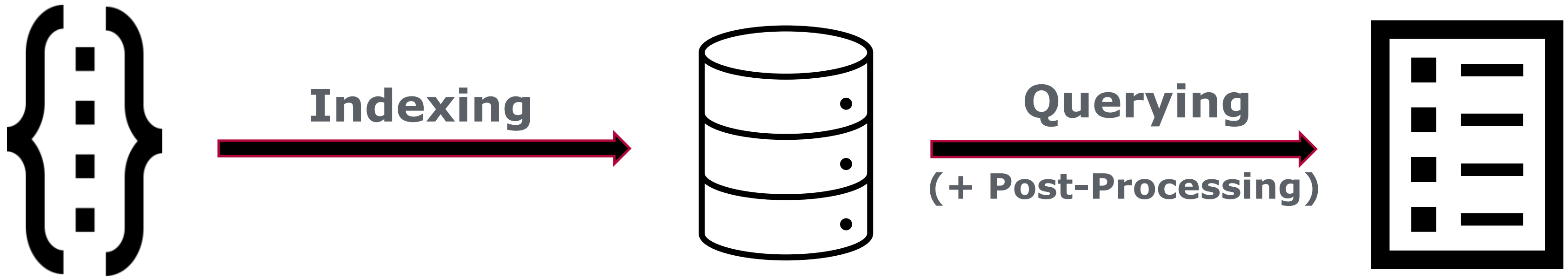


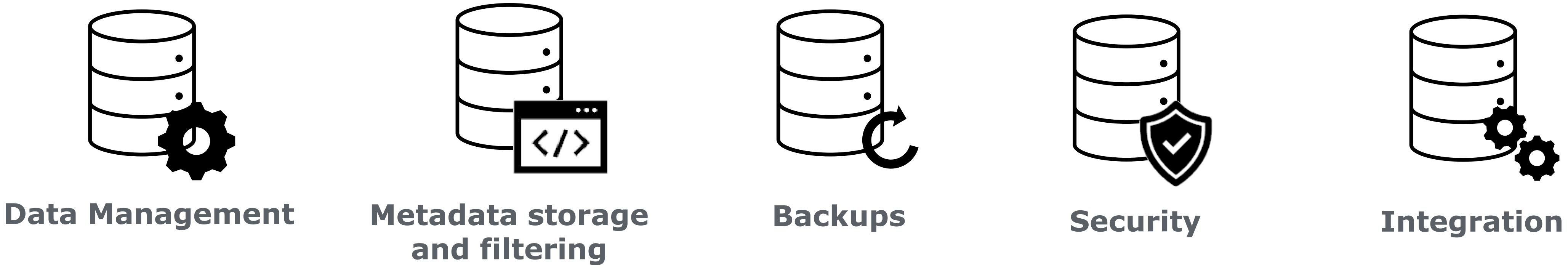
Non-Relational Databases



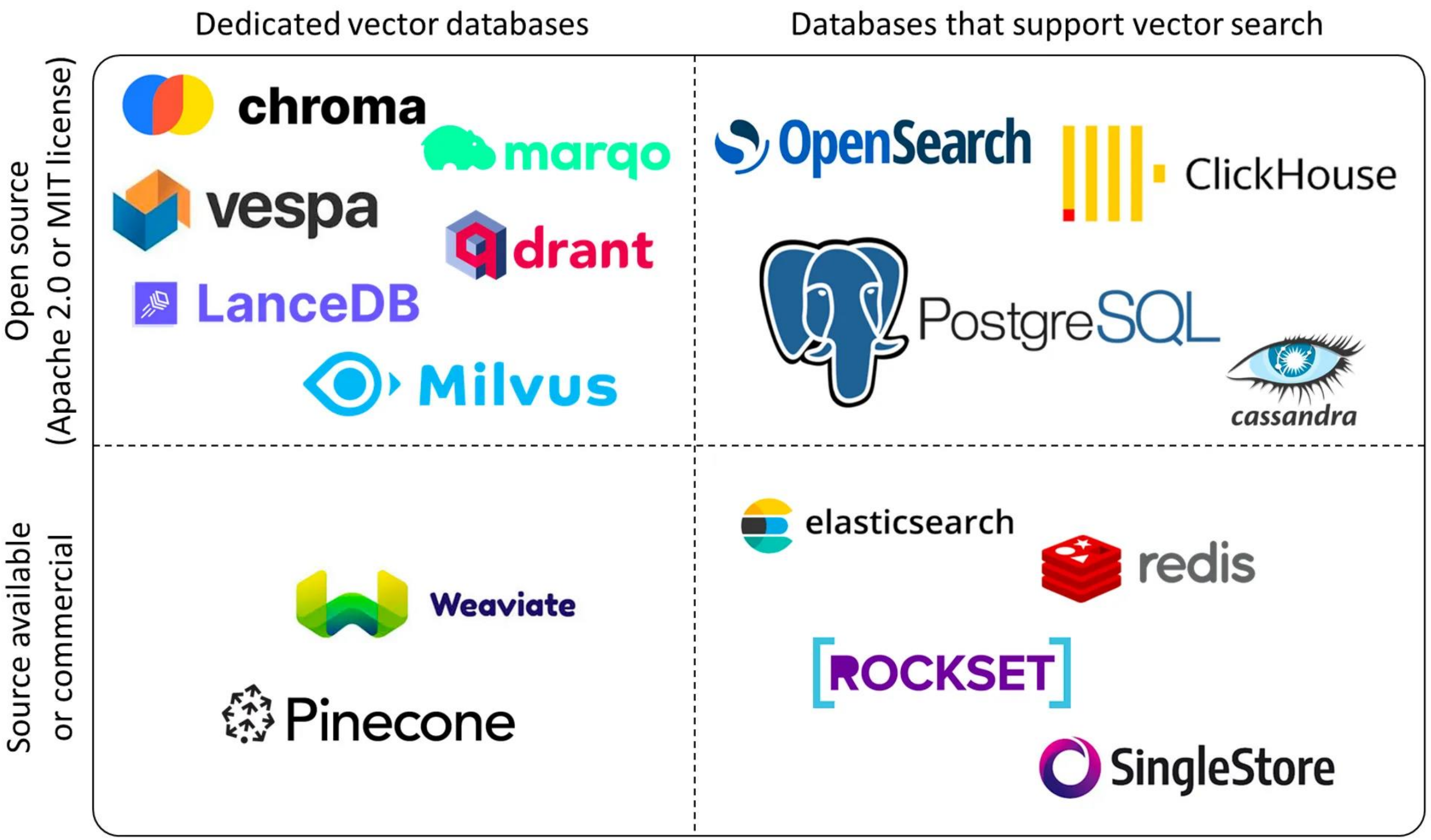
Vector Databases

Vector indices and vector databases





Examples of vector databases



Conclusion

Summing it up

Turning text into numbers
- Data representation
- Tokens
- Bag of Words
- TF-IDF

Improving the representations
- Lexical semantics
- Embeddings
- Analogy questions
- Historical semantics
- Bias and prejudices

Interlude
- Vector comparison
- Vector Visualization

Improving the representations
- Word2vec and other static embeddings
- Sentence embeddings and SBERT
- New directions in language representation

Storing embeddings
- Vector databases
- Vector indices and databases
- Examples of vector databases

KI Service Zentrum

Hasso Plattner Institut

Digital Engineering • Universität Potsdam

SPONSORED BY THE

Federal Ministry of Education and Research

Thank you for your attention

Design IT.
Create Knowledge.

www.hpi.de

Sources

[Dred00] *dredviz Documentation*
URL https://research.cs.aalto.fi//pml/software/dredviz/ - Accessed 2023-11-07

[HaLJ16] HAMILTON, WILLIAM L. ; LESKOVEC, JURE ; JURAFSKY, DAN: Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change, arXiv (2016), S. 2

[JuMa00] JURAFSKY, DANIEL ; MARTIN, JAMES H.: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. 3rd Edition (Draft 07-01-2023)

[LiLS20] LIU, ZHIYUAN ; LIN, YANKAI ; SUN, MAOSONG: *Representation Learning for Natural Language Processing*. Singapore : Springer Nature Singapore, 2020, S.4

[Schw00] SCHWABER-COHEN, ROIE: *Vector Similarity Explained | Pinecone*.
URL https://www.pinecone.io/learn/vector-similarity/ - Accessed 2023-11-07

[Umap00] *UMAP*. URL https://meta.caspershire.net/umap/ Caspershire Meta

[Unde00] COENEN, ANDY ; PEARCE, ADAM: *Understanding UMAP | Google Pair.*
URL https://pair-code.github.io/understanding-umap/ - Accessed 2023-11-07

[Wu23] WU, YINGJUN: *Why You Shouldn't Invest In Vector Databases?*
URL https://blog.det.life/why-you-shouldnt-invest-in-vector-databases-c0cd3f59d23c. - Accessed 2023-11-07