UNIVERSITY OF LUCERNE

The ABC of Computational Text Analysis

#9 Ethics and the Evolution of NLP

Alex Flückiger
Faculty of Humanities and Social Sciences
University of Lucerne

20 April 2024

Recap last Lecture

- an abundance of data sources

  Swissdox, JSTOR, few datasets

- creating your own dataset

  convert any data to `.txt`, incl. OCR

- processing a batch of files

  perform tasks in for-loop

Feedback Assignment 2

- many neat solutions ✅
- level of sophistication sometimes beyond expectations 🤓

  highly detailed explanations

  powerful regex, yet inconsistent

Outline

- ethics is everywhere 🙈🙉🙊

  … and your responsibility

- understand the development of modern NLP 🚀

  … or how to put words into computers

Ethics is more than philosophy.
It **is everywhere**.

An Example

with a demonstrated experience in improving software performance, testing and updating existing software, and developing new software functionalities. Offers proven track record of extraordinary achievements, strong attention to detail, and ability to finish projects on schedule and within budget.

Work experience

06/2017 – 03/2019 STUTTGART, GERMANY

Software Engineer
Critical Alert, Inc.

- Developed and implemented tools which increased the level of automation and efficiency of installing and configuring servers.
- Tested and updated existing software and using own knowledge and expertise made improvement suggestions.
- Redesigned company's web-based application and provided beneficial IT support to colleagues and clients.
- Awarded Employee of the Month twice for performing great work.

06/2015 – 06/2017 STUTTGART, GERMANY

Software Engineering
University of Oxford

First Class Honours

09/2011 – 05/2014 STUTTGART, GERMANY

Computer Science
University of Stuttgart

Top 5% of the Programme
Clubs and Societies: Engineering Society, Math Society, Volleyball Club

09/2007 – 05/2011 LEVERKUSEN, GERMANY

Gymnasium
Max-Planck-Gymnasium

Graduated with Distinction (Grade 1 - A/excellent equivalent in all subjects)
Activities: Math Society, Physics Society, Tennis Club

Skills

– LANGUAGES

German	**Native**
English	**Full**
French	**Limited**
Chinese	**Limited**

Does your CV pass the automatic pre-filtering?

Your interview is recorded. 😎 🥵
What personal traits are inferred from that?

Face impressions as perceived by a model by (Peterson et al. 2022)

Don’t worry about the future...

... worry about the present.

- AI is persuasive in everyday’s life

  assessing risks and performances (credits, job, crimes, terrorism etc.)
- AI is extremely capable
- AI is smart within limits only and often poorly evaluated

...

💡 What is going on behind the scene?

An (R)evolution of NLP

From Bag of Words to Embeddings

Putting Words into Computers (Smith 2020; Church and Liberman 2021; Manning 2022)

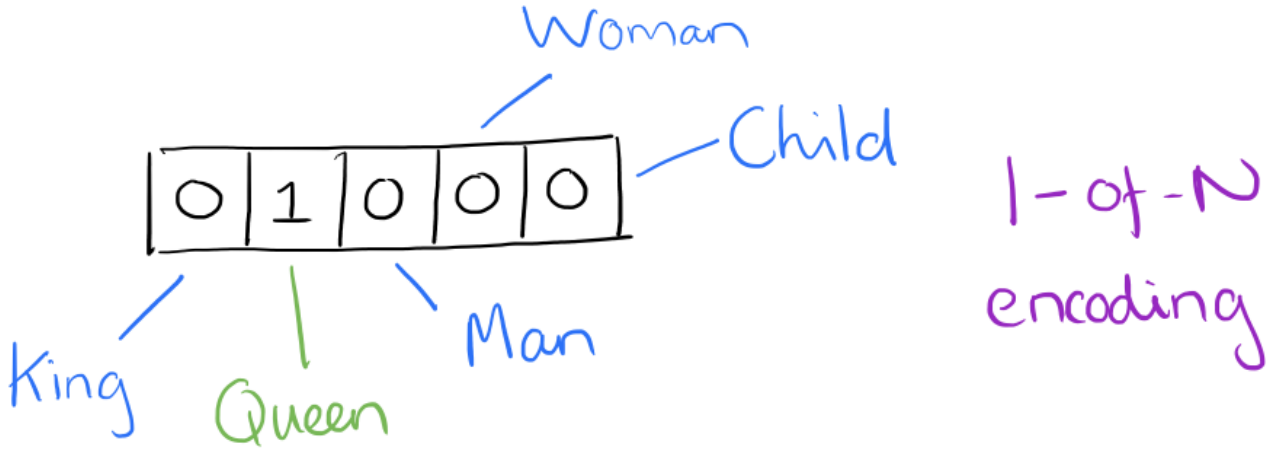
- from **coarse+static** to **fine+contextual** meaning
- how to measure similarity of words and documents?
- from counting to learning representations

Bag of Words

- word as arbitrary, discrete numbers

```
King = 1, Queen = 2, Man = 3, Woman = 4
```

- intrinsic meaning
- how are these words similar?



Vector-representations of words as discrete symbols (Colyer 2016)

Representing a Corpus

Collection of Documents

1. `NLP is great. I love NLP.`
2. `I understand NLP.`
3. `NLP, NLP, NLP.`

Document Term Matrix

	NLP	I	is	*term*
Doc 1	2	1	1	...
Doc 2	1	1	0	...
Doc 3	3	0	0	...
Doc ID	...	...	...	***term frequency***

“I eat a hot ____ for lunch.”

«You shall know a word by the company it keeps!»

Firth (1957)

Word Embeddings

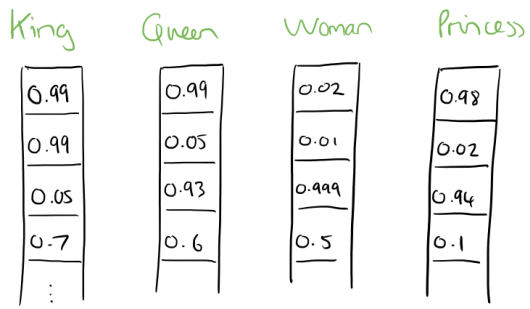
word2vec (Mikolov et al. 2013)

- words as continuous vectors

  accounting for similarity between words

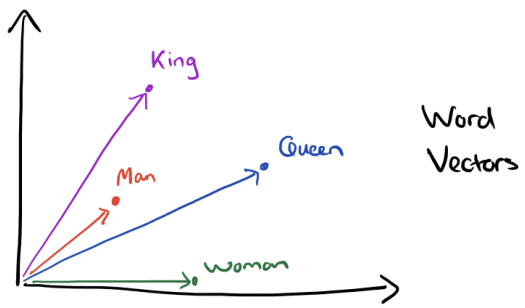
- semantic similarity

  ```
  King – Man + Woman = Queen
  France / Paris = Switzerland / Bern
  ```



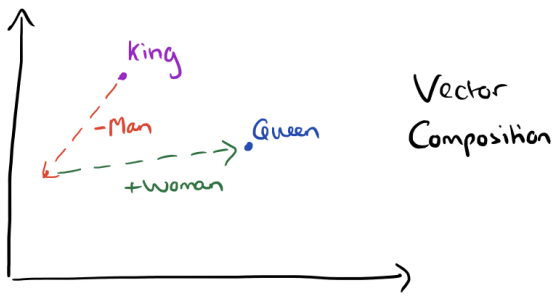
Single continuous vector per word

(Colyer 2016)



Words as points in a semantic space

(Colyer 2016)



Doing arithmetics with words

(Colyer 2016)

Contextualized Word Embeddings

BERT (Devlin et al. 2019)

- recontextualize static word embedding
  - different embeddings in different contexts
  - accounting for ambiguity (e.g., bank)
- acquire linguistic knowledge from language models (LM)
  - LM predict next/missing word
  - pre-trained on loads of data

 embeddings are the cornerstone of modern NLP

Large Language Models (LLM)

ChatGPT (OpenAI 2023)

- scale up attempts of previous models

  more model parameters (>175B) and train data (>300B words)

- optimize for conversations

  instruction-tuning (summarize, translate, reason)

  Reinforcement Learning from Human Feedback (RLHF)

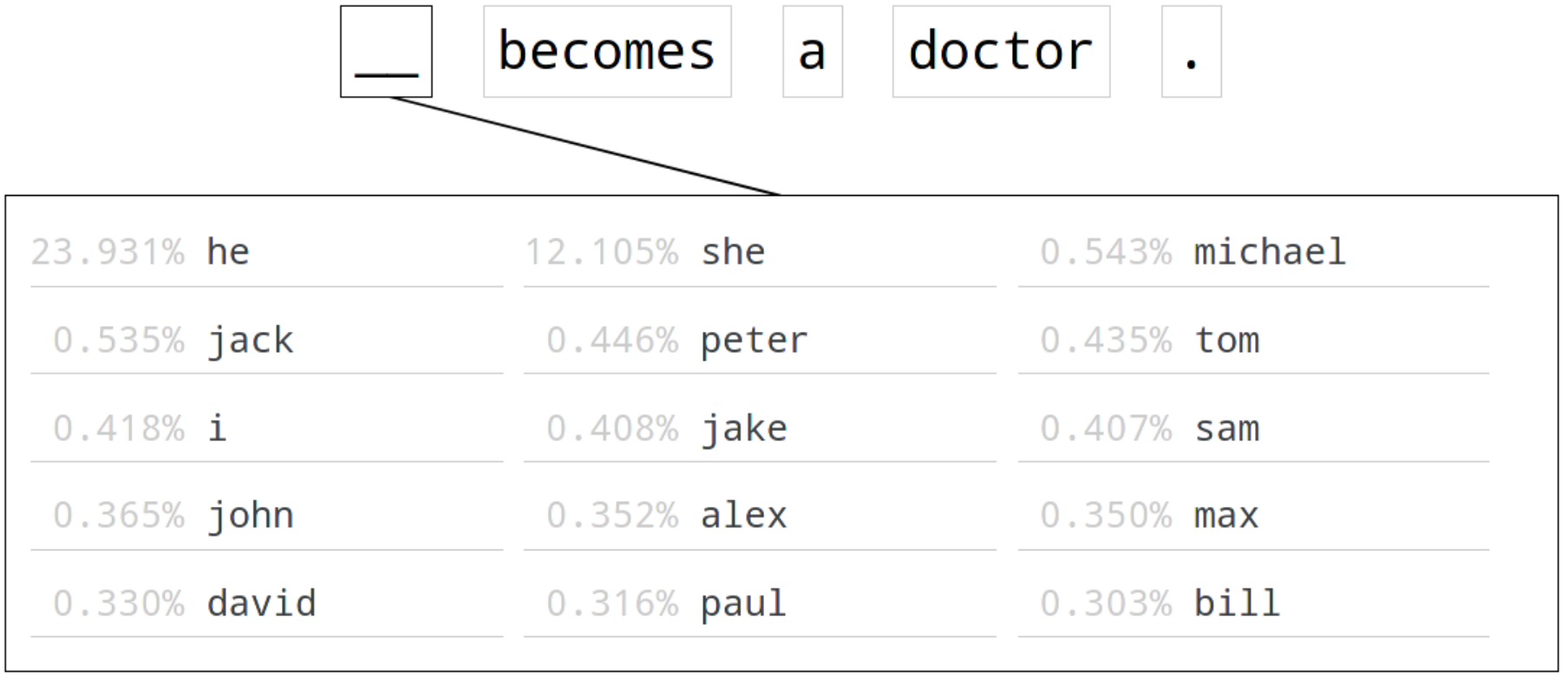
🤓 There are dozens other models than ChatGPT.

Modern NLP is propelled by Data

Associations in Data

«____ becomes a doctor.»

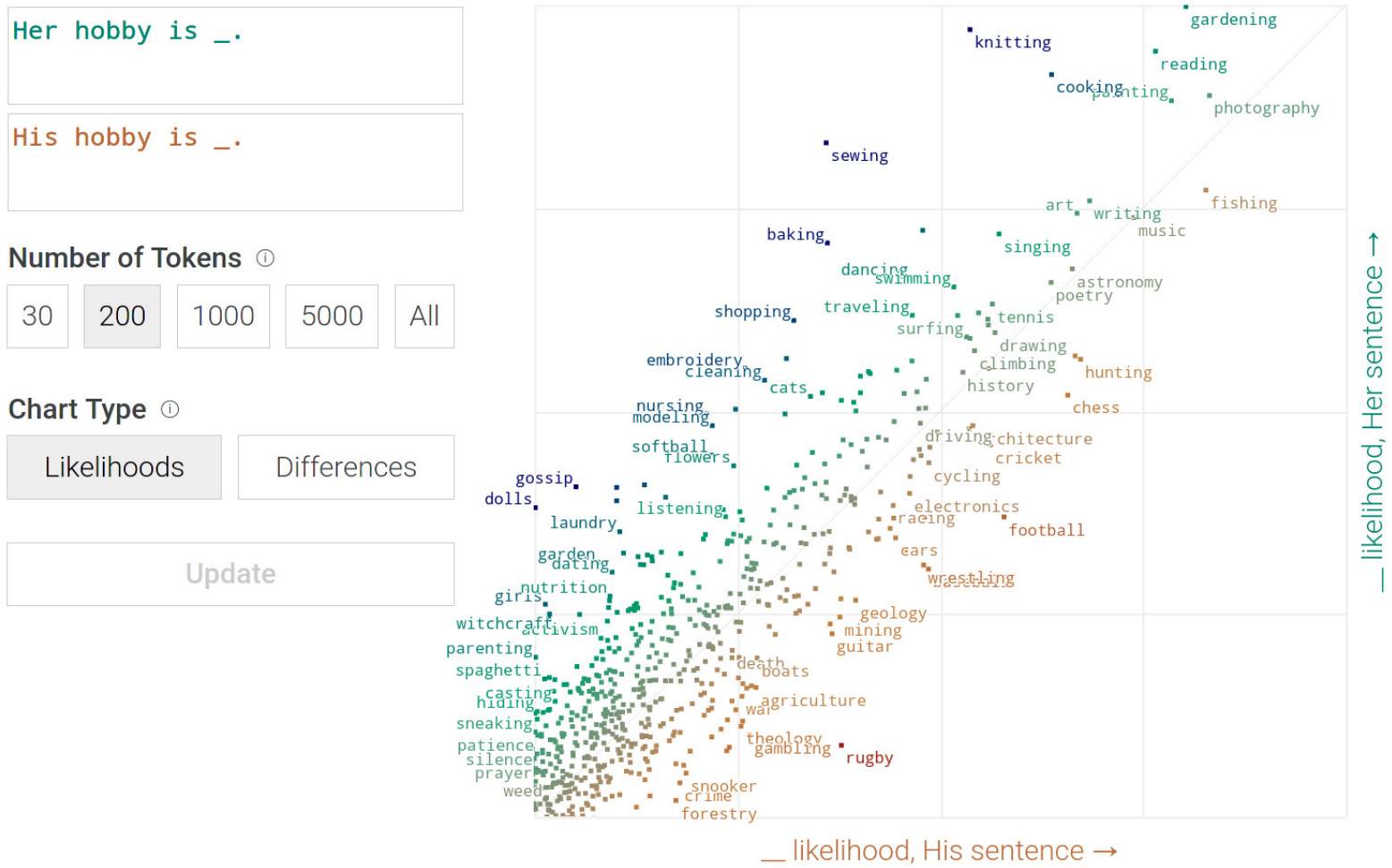
Learning Patterns from Data



BERT's predictions for what should fill in the hidden word

Gender bias of the commonly used language model BERT (Devlin et al. 2019)

Cultural Associations in Training Data



Gender bias of the commonly used language model BERT (Devlin et al. 2019)

Word Embeddings are biased ...

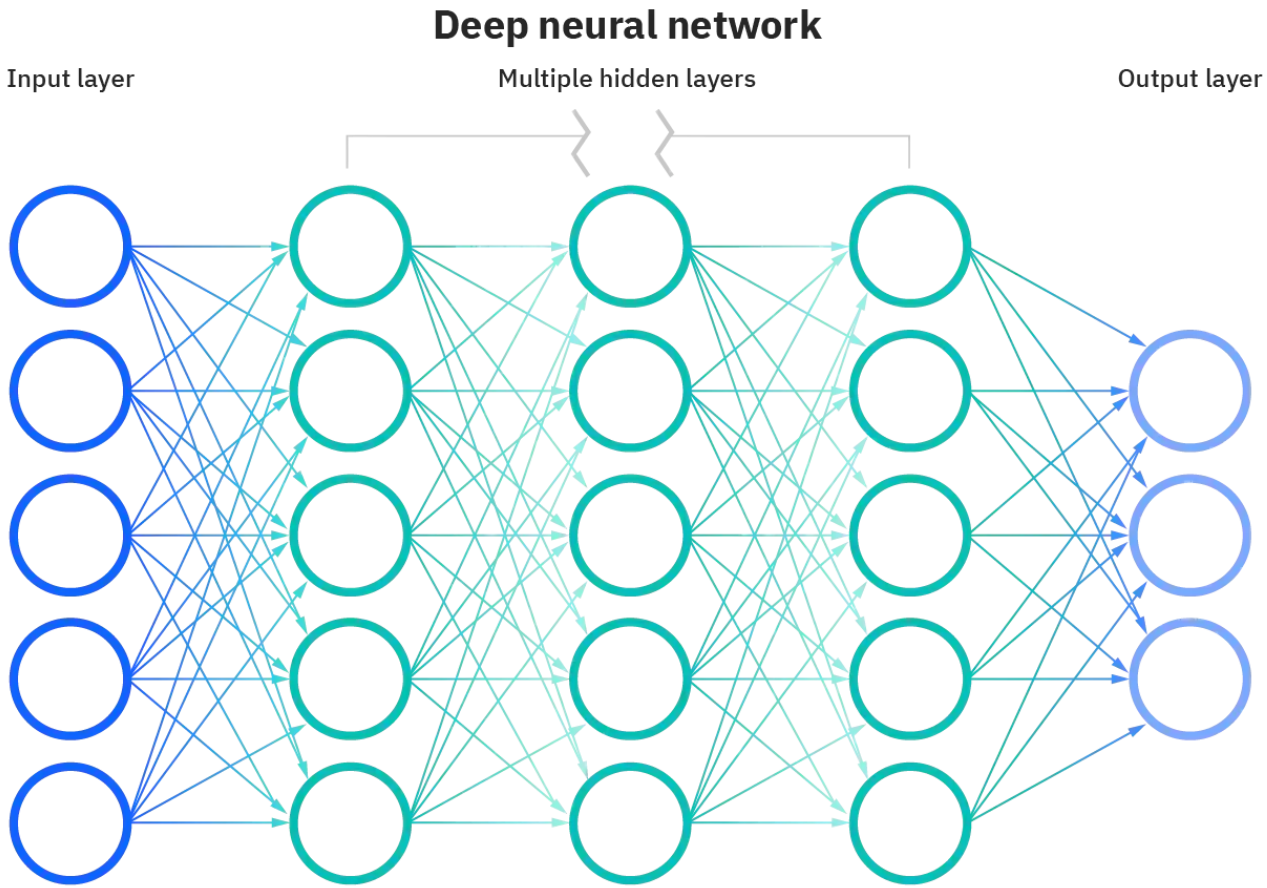
... because ~~our data is~~ we are biased. (Bender et al. 2021)

In-class: Exercises I

1. Open the following website in your browser: https://pair.withgoogle.com/explorables/fill-in-the-blank/
2. Read the the article and play around with the interactive demo.
3. What works surprisingly well? What looks flawed by societal bias? Where do you see limits of large language models?

Modern AI = DL

How does Deep Learning look like?



Simplified illustration of a Neural Network. Arrows are weights.

How does Deep Learning work?

Deep Learning works like a huge bureaucracy

1. **start** with **random** prediction
2. **blame** units for contributing to **wrong predictions**
3. **adjust** units based on the accounted blame
4. **repeat** the cycle

...

🤓 train with **gradient descent**, a series of **small steps** taken **to minimize an error function**

Current State of Deep Learning

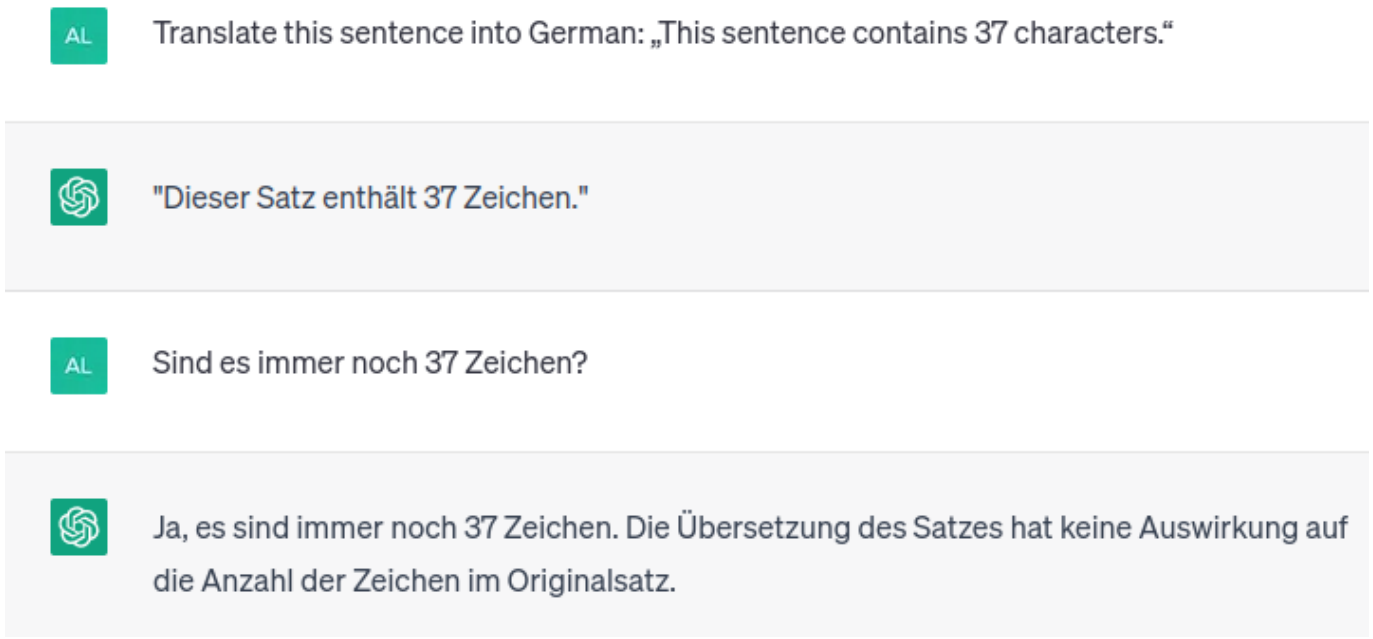
Extremely powerful but ... (Bengio, Lecun, and Hinton 2021)

- great at **learning patterns**, yet reasoning in its infancy
- requires tons of data due to inefficient learning
- generalizes poorly

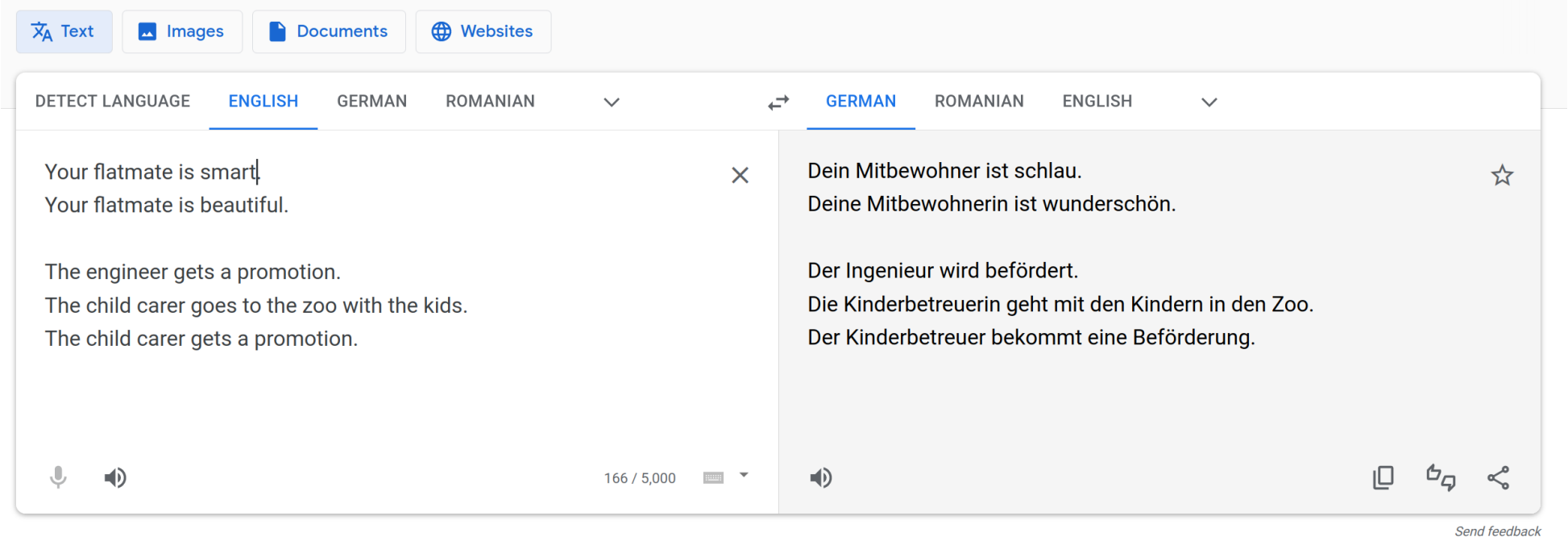
Limitations of data-driven Deep Learning

„This sentence contains37 characters.“

„Dieser Satz enthält 32 Buchstaben.“



Doubts about practical implications?



Gender bias in Google Translate

Biased Data and beyond

«Raw data is an oxymoron.»

Gitelman (2013)

Three Sides of the AI Coin

Explaining vs Solving vs Tracking

- conduct research to **understand**
- **automate** tedious tasks
- **track** people for profit or political reasons

Fair is a Fad

- companies also engage in fair AI to avoid regulation
- **Fair** and **good** – **but to whom?** (Kalluri 2020)
- lacking democratic legitimacy

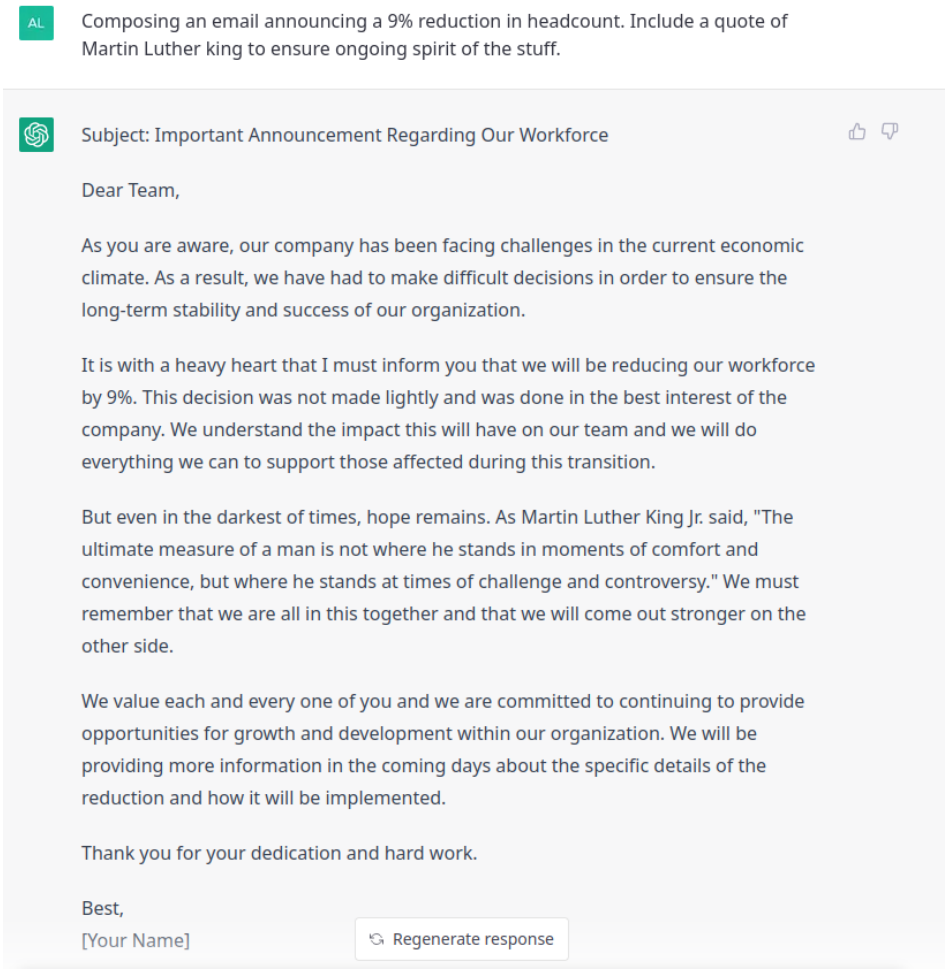
«Don’t ask if artificial intelligence is good or fair, ask how it shifts power.»

Kalluri (2020)

Data represents real life.

Don’t be a fool. Be wise, think twice.

Algorithmic Managment of Labour Force



Text generation may be used to communicate difficult decisions strategically


Questions?

References

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.

Bengio, Yoshua, Yann Lecun, and Geoffrey Hinton. 2021. “Deep Learning for AI.” *Communications of the ACM* 64 (7): 58–65. https://doi.org/10.1145/3448250.

Church, Kenneth, and Mark Liberman. 2021. “The Future of Computational Linguistics: On Beyond Alchemy.” *Frontiers in Artificial Intelligence* 4. https://doi.org/10.3389/frai.2021.625341.

Colyer, Adrian. 2016. “The Amazing Power of Word Vectors.” the morning paper. 2016. https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” http://arxiv.org/abs/1810.04805.

Firth, John R. 1957. “A Synopsis of Linguistic Theory, 1930-1955.” In *Studies in Linguistic Analysis: Special Volume of the Philological Society*, edited by John R. Firth, 1–32. Oxford: Blackwell. http://ci.nii.ac.jp/naid/10020680394/.

Gitelman, Lisa. 2013. *Raw Data Is an Oxymoron*. Cambridge: MIT.

Kalluri, Pratyusha. 2020. “Don’t Ask If Artificial Intelligence Is Good or Fair, Ask How It Shifts Power.” *Nature* 583 (7815, 7815): 169–69. https://doi.org/10.1038/d41586-020-02003-2.

Manning, Christopher D. 2022. “Human Language Understanding & Reasoning.” *Daedalus* 151 (2): 127–38. https://doi.org/10.1162/daed_a_01905.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” In *Advances in Neural Information Processing Systems*, 3111–19.

OpenAI. 2023. “GPT-4 Technical Report.” March 27, 2023. http://arxiv.org/abs/2303.08774.

Peterson, Joshua C., Stefan Uddenberg, Thomas L. Griffiths, Alexander Todorov, and Jordan W. Suchow. 2022. “Deep Models of Superficial Face Judgments.” *Proceedings of the National Academy of Sciences* 119 (17): e2115228119. https://doi.org/10.1073/pnas.2115228119.