

The ABC of Computational Text Analysis

#9 ETHICS AND THE EVOLUTION OF NLP

Alex Flückiger
Faculty of Humanities and Social Sciences
University of Lucerne

01 May 2025

Recap last lecture

- 2nd assignment accomplished 
- reminder: [share mini-project idea](#) by today !
- perform first real-world (data) analysis 
analyse discourse on wokeness in Swiss media

Outline

- evaluate this course 
- ethics is everywhere 
... and your responsibility
- understand the development of modern NLP 
... or how to put words into computers



Course Evaluation

Tell me... 

Please follow the QR code or link in the email received



Thanks for any constructive feedback,
be it sweet or sour! 

Ethics is more than an academic subject.

It is **everywhere**.

Apply for a job at a big company



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 Engineer

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 4 subjects)

Activities: Math Society, Physics Society, Tennis Club



Skills

- LANGUAGES

German

Native

English

Full

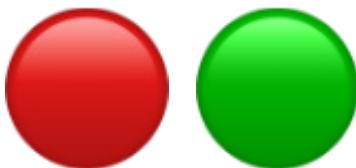
French

Limited

Chinese

Limited

Does your CV pass the automatic
pre-filtering?



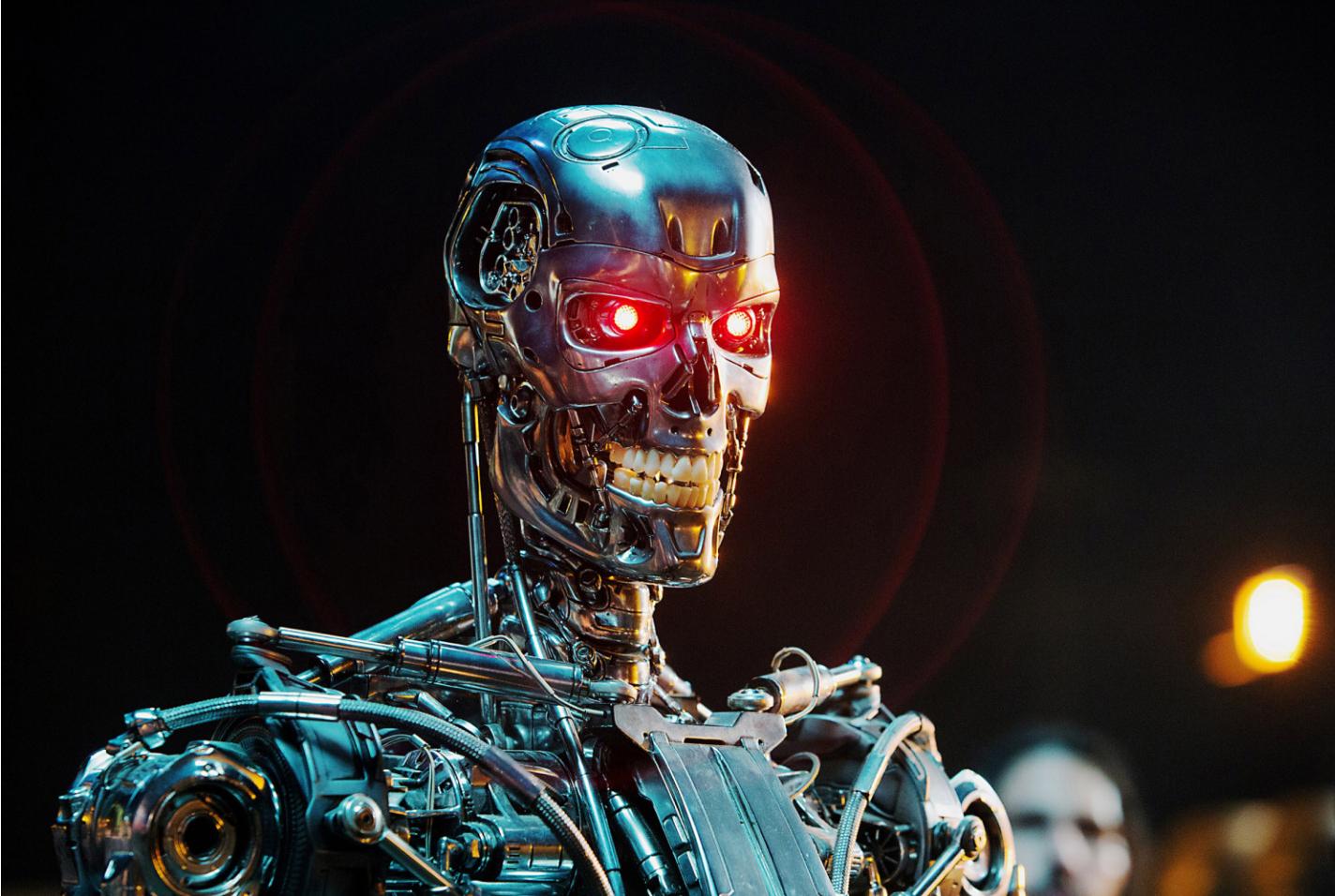
Your interview is recorded. 😎汗

What personality is inferred automatically?



Facial expressions as perceived by a model by (Peterson et al. 2022)

Don't worry about the future...



The narrative of autonomous and evil-minded robots is an illusion.

...worry about the present 😞

- AI is persuasive in everyday life
 - assessing risks and performances (credit, job, crime, terrorism etc.) ([Hofmann et al. 2024](#))
- AI is extremely capable
 - increasingly difficult to assess where it fails
- AI has data-driven bias
 - systems are often evaluated poorly



An (r)evolution of NLP

What is a word?

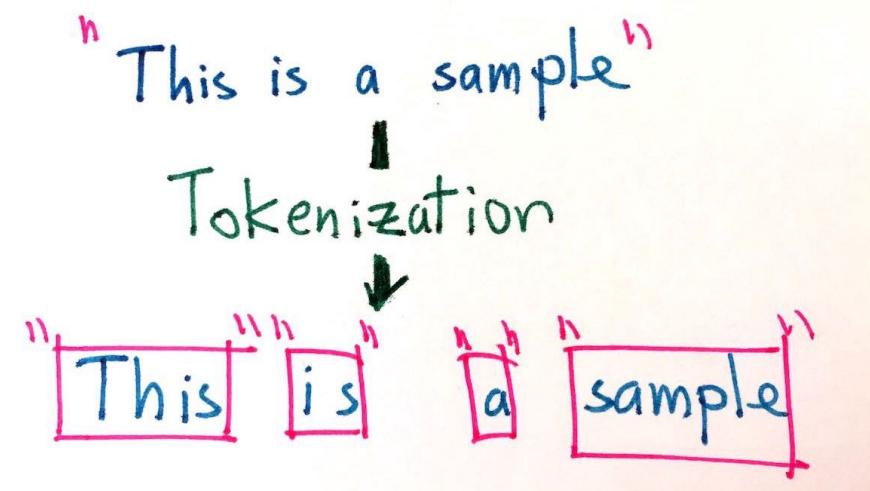
- words ~ segments between whitespace
- yet, there are ...

contractions: U.S., don't

collocations: New York

Token

- token ~ computational unit
representation of words
- lemma ~ base form of a word
texts → *text*
goes → *go*
- stop words ~ functional words
lacking deeper meaning
the, a, on, and ...



Segmenting a text into tokens

Let's tokenize this sentence! Isn't it easy? 😎

Classic processing steps in NLP

1. Tokenizing

segmenting text into words, punctuation etc.

2. Tagging part-of-speech (POS)

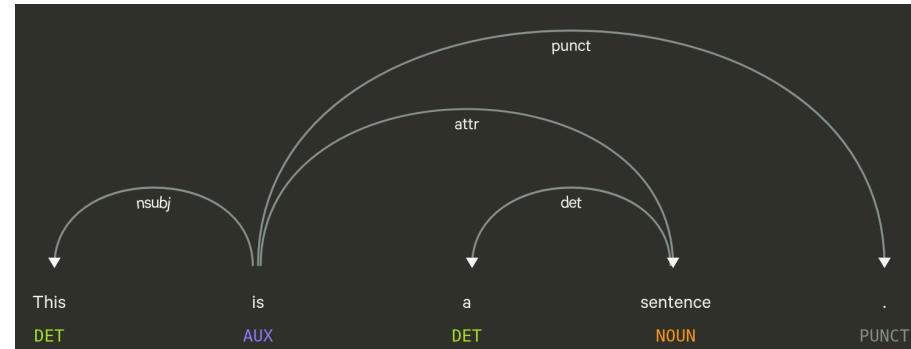
assigning word types (e.g. verb, noun)

3. Parsing

describing syntactic relations

4. Named Entity Recognition (NER)

organizations, persons, locations, time etc.



Automatically inferred information of a sentence

👉 Catch up on NLP with

Jurafsky and Martin (forthcoming)

How to represent words in computers?

From bag of words to embeddings

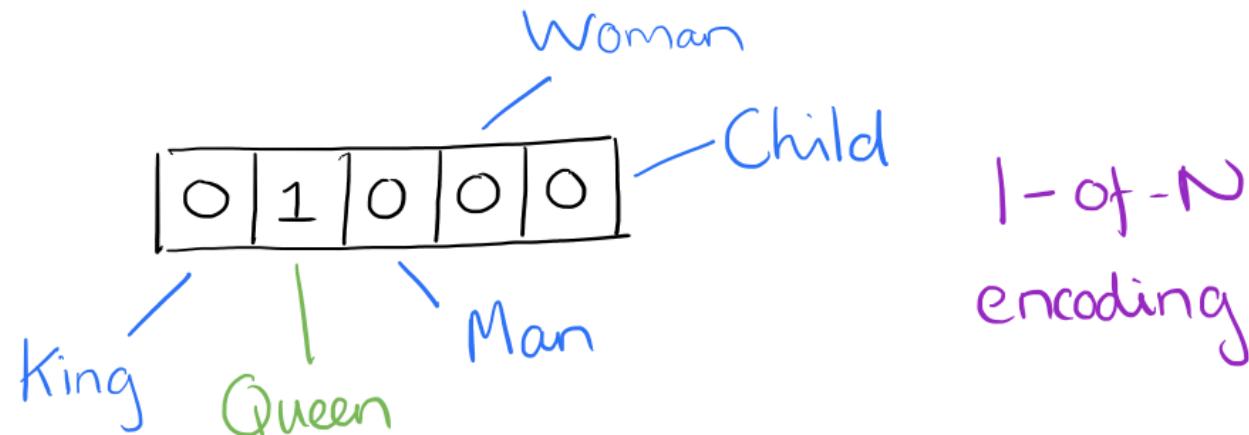
Putting words into computers

(Smith 2020; Church and Liberman 2021; Manning 2022)

- how to measure similarity of words and documents?
- from **counting** to rich **semantic 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

A 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 ____ tonight”.

“The pizza was ____.”

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

Firth (1957)

Formalize the linguistic intuition

Train a model by

1. masking words
2. letting the computer predict the masked words using context

Word embeddings

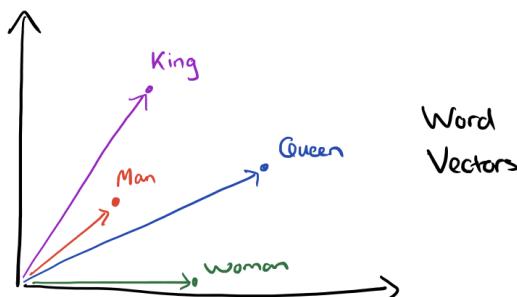
word2vec (Mikolov et al. 2013)

- words as continuous vectors
accounting for similarity between words
- semantic similarity

$$\text{King} - \text{Man} + \text{Woman} = \text{Queen}$$

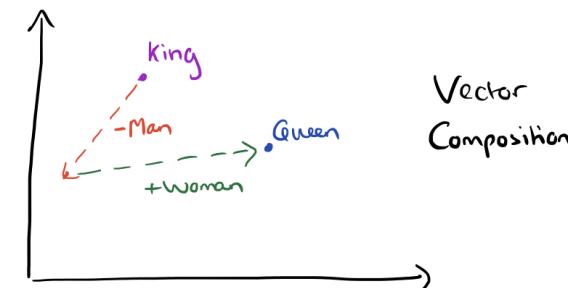
$$\text{France} / \text{Paris} = \text{Switzerland} / \text{Bern}$$

King	Queen	Woman	Princess
0.99	0.99	0.02	0.48
0.99	0.05	0.01	0.02
0.05	0.93	0.999	0.94
0.7	0.6	0.5	0.1
:			



Single continuous vector per word (Colyer 2016)

Words as points in a semantic space



Doing arithmetics with words

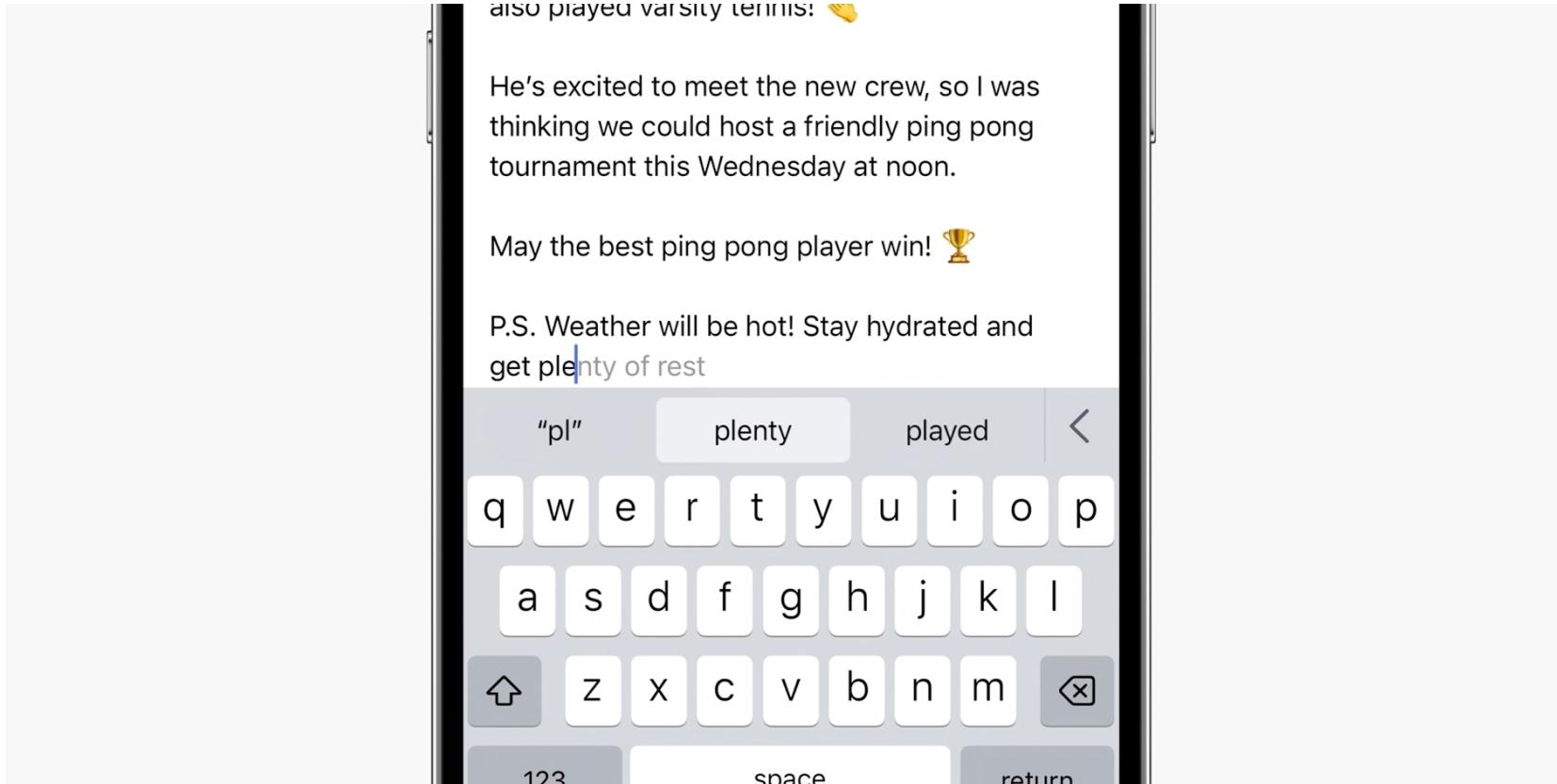
Contextualized word embeddings

BERT (Devlin et al. 2019)

- recontextualize static word embedding
 - word gets different embedding in different contexts
 - accounting for ambiguity (e.g., bank)
- acquire linguistic knowledge from loads of data
 - mask random words/phrases in sentences

From embeddings to generation

Instead of masking, train the model to predict the next word



Autocompletion on iPhone

Next word prediction is
extremely powerful! 💪

The birth: Universal problem solver

Any task can be modeled as Text-to-Text

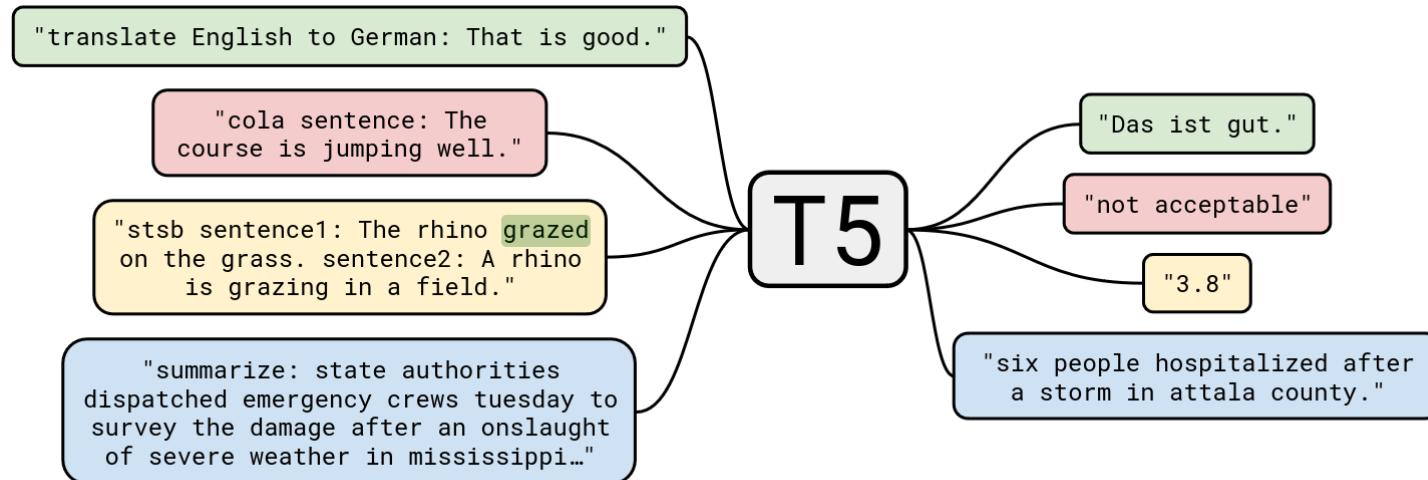


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. “T5” refers to our model, which we dub the “**Text-to-Text Transfer Transformer**”.

(Raffel et al. 2020)

Large Language Models (LLM)

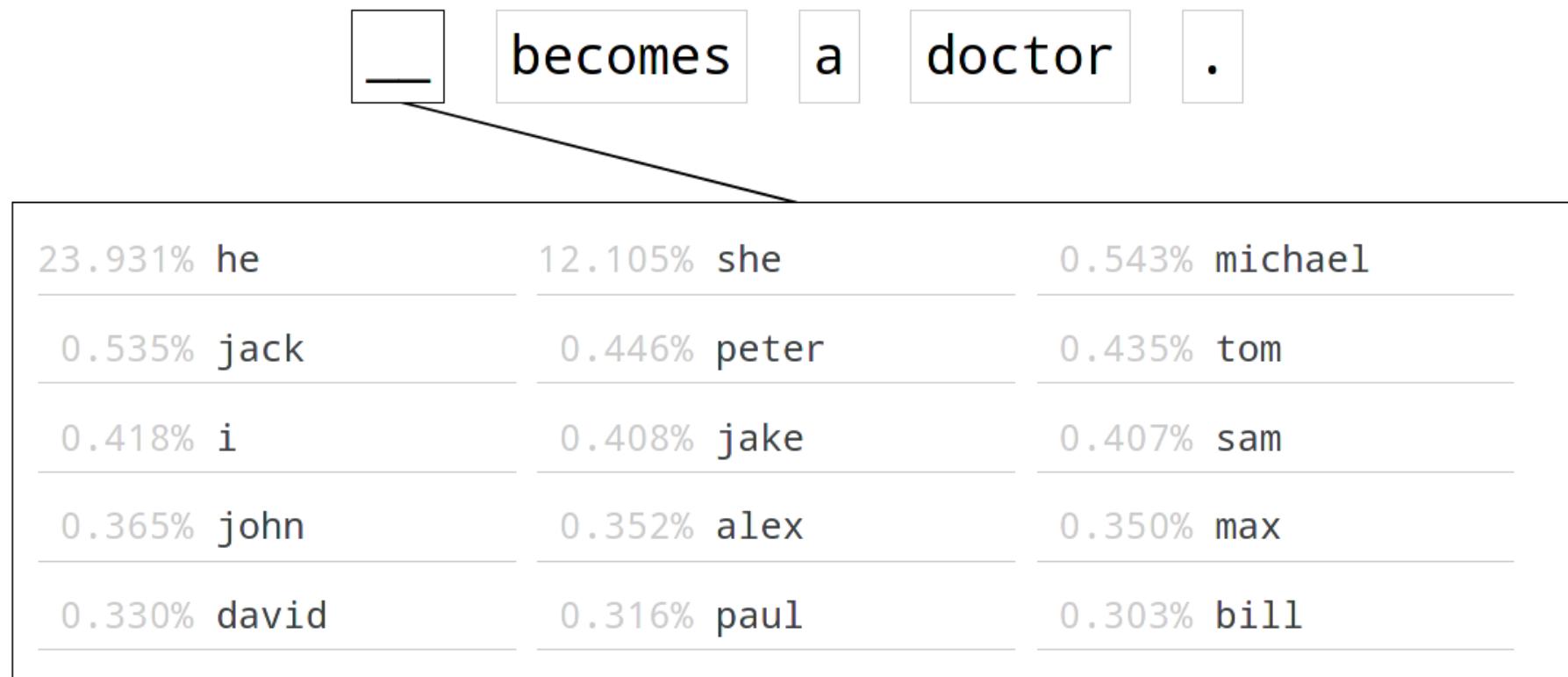
- scale up approach of predicting the next word
 - bigger models and more data
- optimize for dialogue instead of prose text
 - instruction-tuning on many [prompt + answer pairs](#) (summarize, translate, explain etc.)
 - Reinforcement Learning (RL)

Modern NLP is propelled by data

Associations in data

« becomes a doctor.»

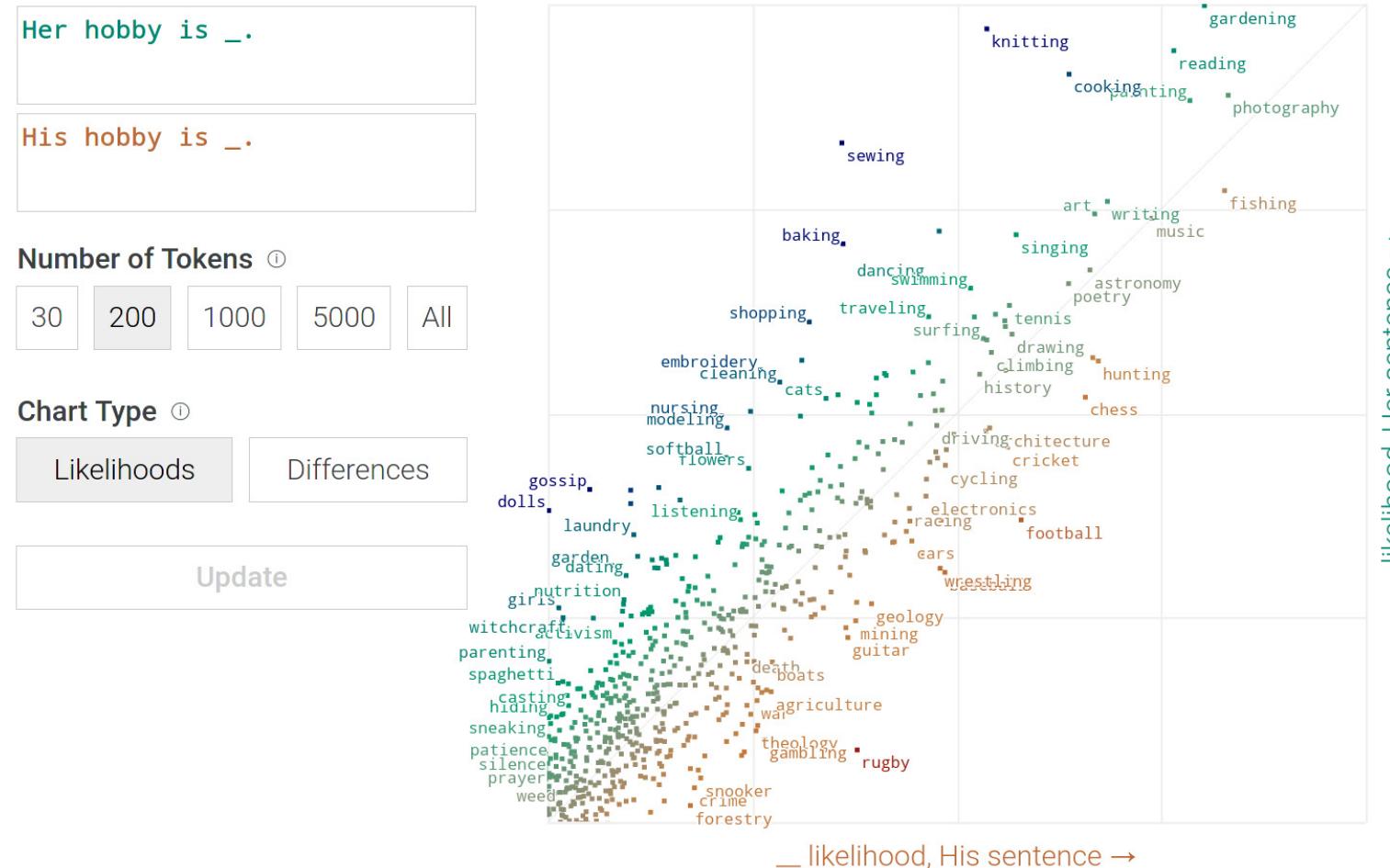
Learning patterns from data



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)

Bias with practical implications

The screenshot shows the Google Translate interface comparing English and German text. The English input on the left includes sentences about flatmates being smart and beautiful, and child carers getting promotions. The German output on the right translates these sentences, showing a clear bias towards male subjects (Mitbewohner, Ingenieur) compared to female subjects (Mitbewohnerin, Kinderbetreuerin).

Text input (English):

- Your flatmate is smart.
- Your flatmate is beautiful.
- The engineer gets a promotion.
- The child carer goes to the zoo with the kids.
- The child carer gets a promotion.

Text output (German):

- Dein Mitbewohner ist schlau.
- Deine Mitbewohnerin ist wunderschön.
- Der Ingenieur wird befördert.
- Die Kinderbetreuerin geht mit den Kindern in den Zoo.
- Der Kinderbetreuer bekommt eine Beförderung.

Feedback icon: A star icon is present next to the German output, likely for reporting bias.

Page controls: 166 / 5,000, a settings gear icon, and a send feedback link.

Gender bias in Google Translate

In-class: Exercises I

1. Open the following website in your browser:

<https://pair.withgoogle.com/explorables/fill-in-the-blank/>

2. Read the article and play around with the interactive demo.

3. What works surprisingly well? What looks flawed by societal bias?

«Raw data is an oxymoron.»

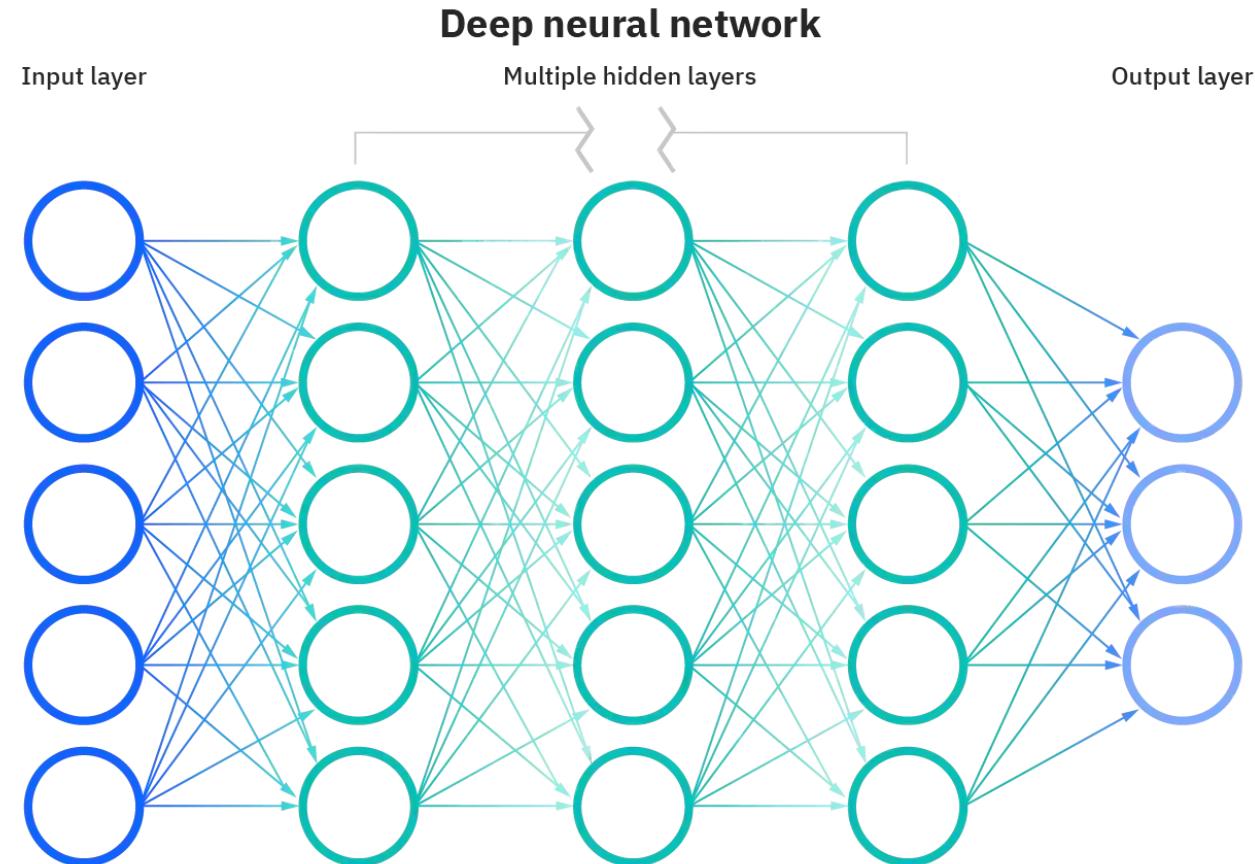
Gitelman (2013)

Fair is a fad

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

Modern AI = DL

How does deep learning model look like?



Simplified illustration of a neural network. Arrows are weights and bullets are (intermediate) states.

How does deep learning work?

Deep learning works like a huge bureaucracy 

1. start with random predictions
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 is still in its infancy
- requires tons of data due to inefficient learning
- generalizes poorly

👉 avoid anthropomorphizing



Questions?

Additional resources

- deepen your understanding of modern LLMs with
 - [this video](#)
 - [this animated explainer](#)
- racial bias in [CV screening](#) and [generated pictures](#) (interactive stories by Bloomberg)

Optional Exercise I: Limitations of DL

Consider the following translation:

“This sentence contains 37 characters.”

“Dieser Satz enthält 32 Buchstaben.”

Optional Exercise I: Limitations of DL

Reasoning improved, yet still wrong

AL

Translate this sentence into German: „This sentence contains 37 characters.“



„Dieser Satz enthält 37 Zeichen.“

AL

Sind es immer noch 37 Zeichen?



Ja, es sind immer noch 37 Zeichen. Die Übersetzung des Satzes hat keine Auswirkung auf die Anzahl der Zeichen im Originalsatz.

Test April 2024

Translate this sentence into German: “This sentence contains 37 characters.”

The German translation of “This sentence contains 37 characters.” is:

„Dieser Satz enthält 37 Zeichen.“

Sind es immer noch 37 Zeichen?

Gute Frage! Schauen wir uns das genau an:

„Dieser Satz enthält 37 Zeichen.“

Zählen wir die Zeichen (inklusive Leerzeichen und Punkt):

- „**Dieser**“ → 6
- **Satz** → 4
- **enthält** → 7
- **37** → 2
- **Zeichen** → 7
- „.“ → 1
- Leerzeichen zwischen den Wörtern → 4

Gesamt: $6 + 4 + 7 + 2 + 7 + 1 + 4 = 31$ Zeichen

☞ Nein, die deutsche Version enthält nur 31 Zeichen.

Möchtest du eine deutsche Version mit genau 37 Zeichen?

□ □ □ □ □ □ □

Test April 2025

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? 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.
- Hofmann, Valentin, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. “Dialect Prejudice Predicts AI Decisions about People’s Character, Employability, and Criminality.” March 1, 2024. <https://doi.org/10.48550/arXiv.2403.00742>.
- Jurafsky, Dan, and James H. Martin. forthcoming. *Speech and Language Processing*. 3rd (Feb 3, 2024 draft). London: Prentice Hall. <https://web.stanford.edu/~jurafsky/slp3/>.
- 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*