Natural Language Processing

01: Introduction

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: 2022-04-01

Technology
Arts Sciences
TH Köln

Information Retrieval Research Group

Prof. Dr. Philipp Schaer

- IR, evaluation of IR systems, digital libraries

Timo Breuer, M.Sc.

- Living Labs, reproducibility project STELLA

Fabian Haak, M.Sc.

- Bias in query suggestions, project ESUPOL

Björn Engelmann, M.Sc.

- Journalistic Information Extraction, project JoIE

Dr. Christin Kreutz

- Bibliographic metadata, scientometrics

Projects, jobs, theses: **https://ir.web.th-koeln.de**

Information Retrieval Research Group

Prof. Dr. Philipp Schaer

- IR, evaluation of IR systems, digital libraries

Timo Breuer, M.Sc.

- Living Labs, reproducibility project STELLA

Fabian Haak, M.Sc.

- Bias in query suggestions, project ESUPOL

Björn Engelmann, M.Sc.

- Journalistic Information Extraction, project JoIE

Dr. Christin Kreutz

- Bibliographic metadata, scientometrics

Projects, jobs, theses: **https://ir.web.th-koeln.de**

Syllabus – Exams and grades – Bachelor

The final grade is based on a **group project** with following parts:

- **Presentation** on concept and 1st milestone (40%)
- Final workshop-grade **paper** incl. code and data (60%)

- Topics and groups will **be assigned in the second half** of the semester – Details to follow!

- Grading will be based on a Rubic table to ensure transparency

Syllabus – Exams and grades – Master

The final grade will be based on an **oral exam**!

- One short **presentation on course-related materials** (50%)
- Open questions on topics related to the talks and the material of the lecture and tutorials (50%)

- Grading will be based on a Rubic table to ensure transparency

Syllabus – Course organization

This course is split in two parts:

- **1st half of semester: Lectures and tutorials (B.Sc. and M.Sc.)**
- **2nd half of semester: Lectures and group project (B.Sc.)**

The lecture will be online (afaik for 3 weeks) and later @Südstadt

- Zoom + GitHub
- I will give you access to slides and material as early as possible.

Each topic will be complemented with a tutorial

- Please **prepare @ home**
- In the tutorial sessions we will assume that you are familiar with the basics.

Schedule for Summer Term 2022

Date	#	Slot 13:30	Slot 15:15
01.04.22	1	Introduction and Overview (L)	Basic Text Processing (L)
08.04.22	2	Standard NLP Pipelines (T)	Common Toolkits: Spacy, NLTK (T)
15.04.22	no lecture (Good Friday)		
22.04.22	3	WordNet (L)	Vector Semantics (L)
29.04.22	4	WordNet, GermaNet (T)	Vector Semantics (T)
06.05.22	5	Information Extraction (L)	Sentiment Analysis (L)
13.05.22	Project week		
20.05.22	6	Language Models/ Ethics in NLP (L)	Group assignment (P)
27.05.22	7	Group work (P)	Group work (P)
03.06.22	8	Data Programming for IE (L)	Group work (P) / **Oral Exam (M.Sc.)**
10.06.22	9	Guest Lecture (L)	Group work (P)
17.06.22	10	Group work (P)	Group work (P)
24.06.22	11	***Student talks – Milestone presentation (B.Sc.)***	
31.08.22	**Submit term paper (B.Sc.)**		

NLP is interdisciplinary

- Artificial intelligence
- Machine learning (ca. 2000—today), statistical models, neural networks
- Linguistics (representation of language)
- Social sciences/humanities (models of language at use in culture/society)

NLP = processing language with computers

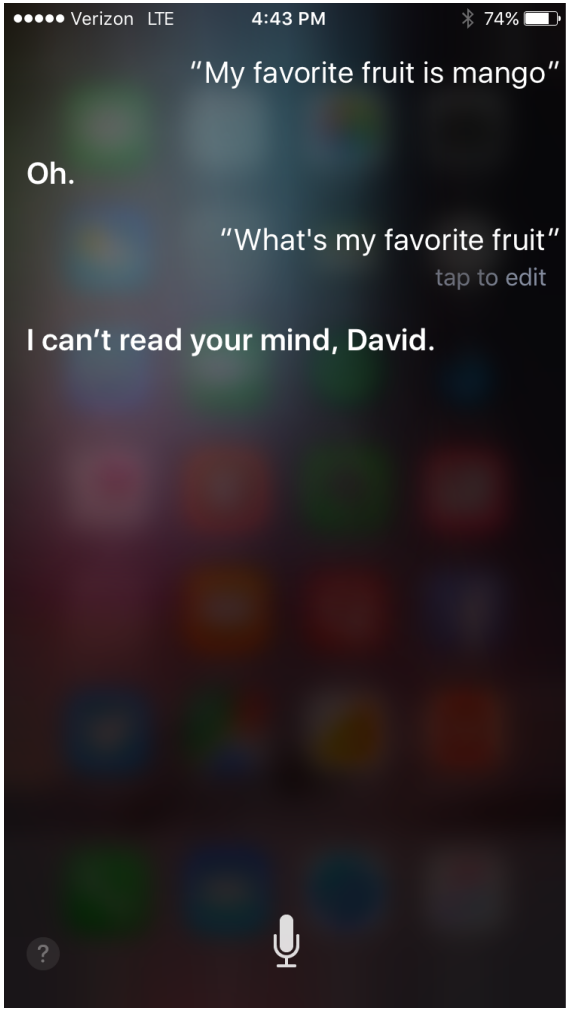
Turing Test

- Distinguishing human vs. computer only through written language (Turing 1950)

- Today:
  completely automated public Turing test to tell computers and humans apart

Images: https://en.wikipedia.org/wiki/CAPTCHA, public domain, https://www.bankofengland.co.uk/banknotes/polymer-50-pound-note

Where are we now…



Where are we now…

Q: What is your favorite animal?

A: My favorite animal is a dog.

Q: Why?

A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?

A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many bonks are in a quoit?

A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii to seventeen?

A: It takes two rainbows to jump from Hawaii to seventeen.

Lacker (2020), "Giving GPT-3 a Turing Test" (https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html); Brown et al. (2020), "Language Models are Few-Shot Learners" (NeurIPS)

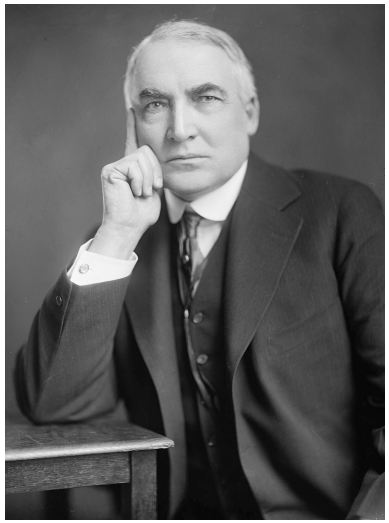
What makes language hard?

- Language is a complex social process
- Tremendous ambiguity at every level of representation
- Modeling it is AI-complete (requires first solving general AI)

What makes language hard?

- Speech acts – “can you pass the salt?” (Austin 1962, Searle 1969)
- Conversational implicature (“The opera singer was amazing; she sang all of the notes”). (Grice 1975)
- Shared **knowledge** (“Warren ran for president”)

Elizabeth Warren
2020



Warren G. Harding
1920

Ambiguity

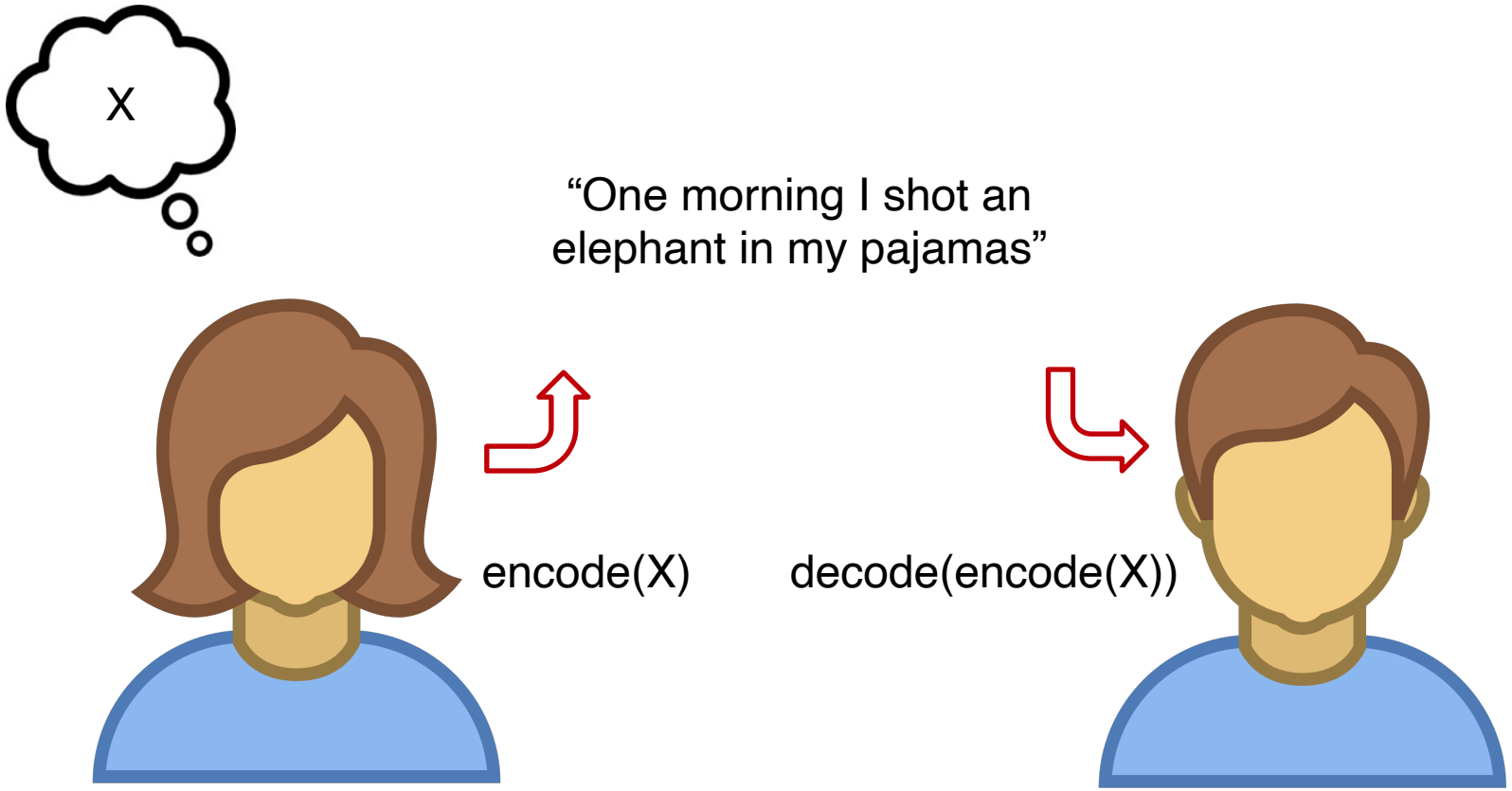
- “One morning I shot an elephant in my pajamas”

Animal Crackers

Processing as Representation

- NLP generally involves representing language for some end, e.g.:
  - dialogue
  - translation
  - speech recognition
  - text analysis

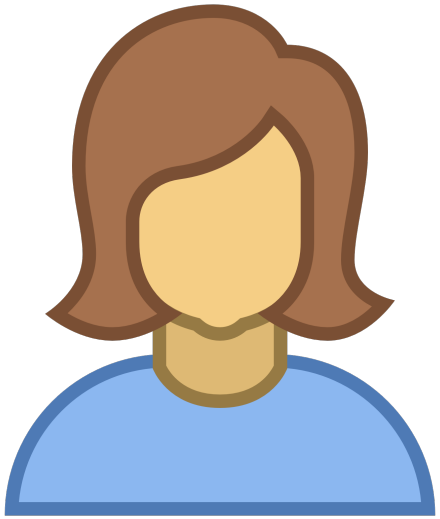
Information theoretic view



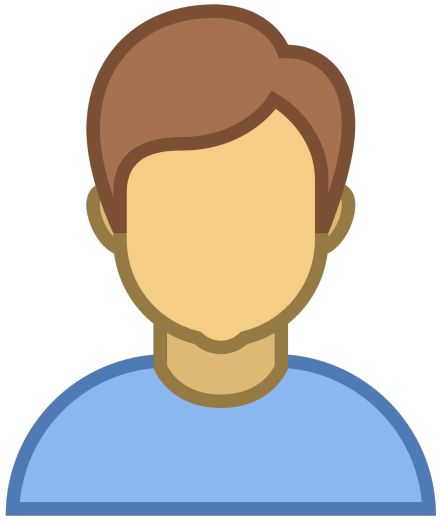
Shannon 1948

Rational speech act view

"One morning I shot an elephant in my pajamas"



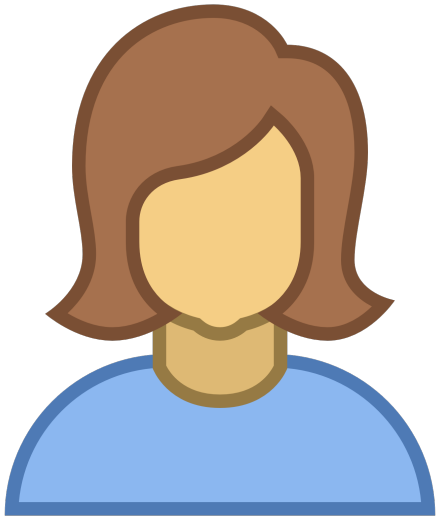
Communication involves recursive reasoning: how can X choose words to maximize understanding by Y?



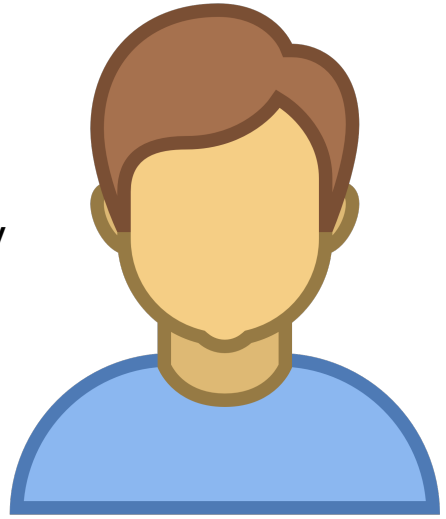
Frank and Goodman 2012

Pragmatic view

"One morning I shot an elephant in my pajamas"

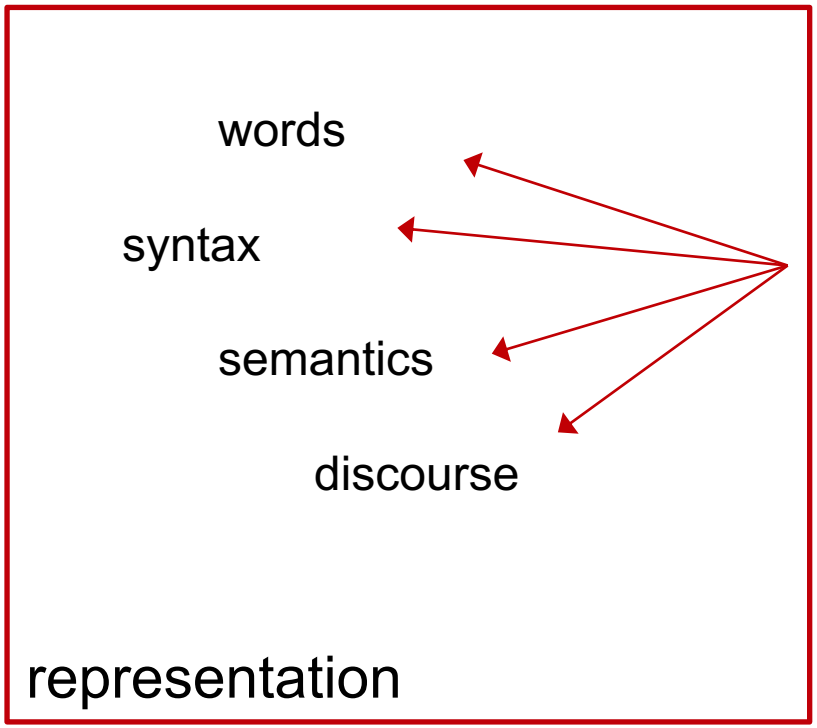


Meaning is co-constructed by the conversational partners and the context of the utterance

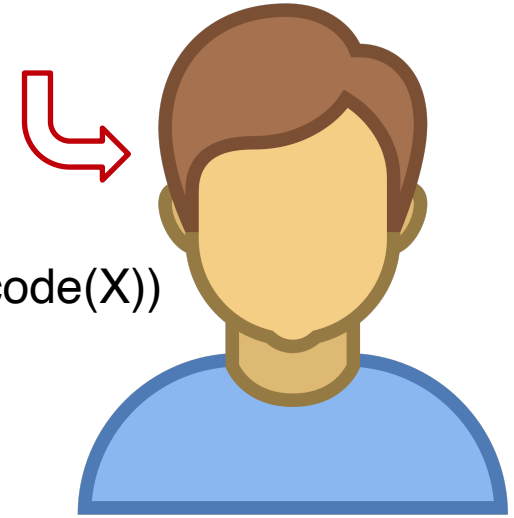


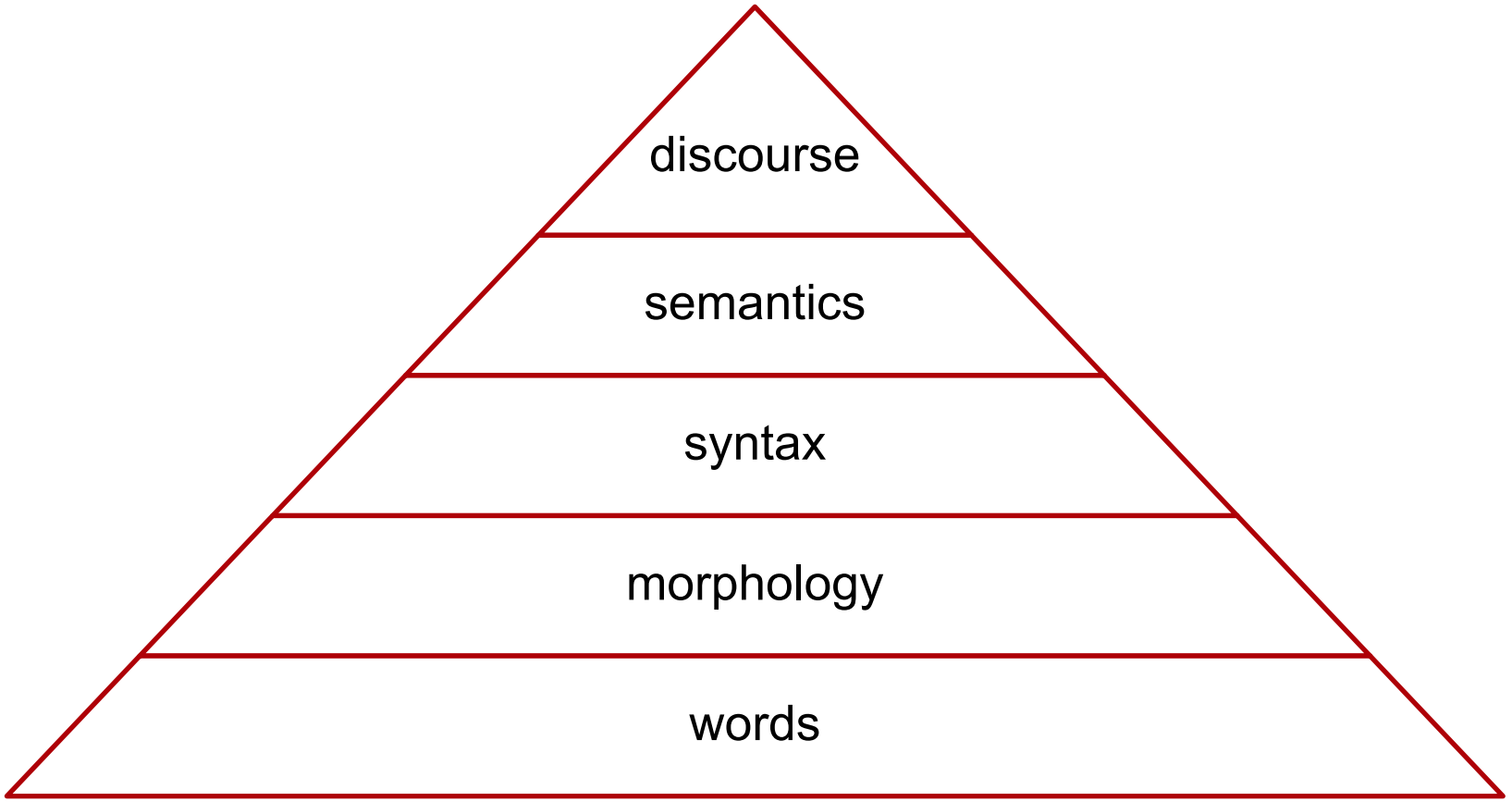
Decoding

“One morning I shot an elephant in my pajamas”



decode(encode(X))




discourse
semantics
syntax
morphology
words

Words

- One morning I shot an elephant in my pajamas
- I didn’t shoot an elephant
- Imma let you finish but Beyonce had one of the best videos of all time
- 一天早上我穿着睡衣射了一只大象

Parts of speech

noun verb noun noun

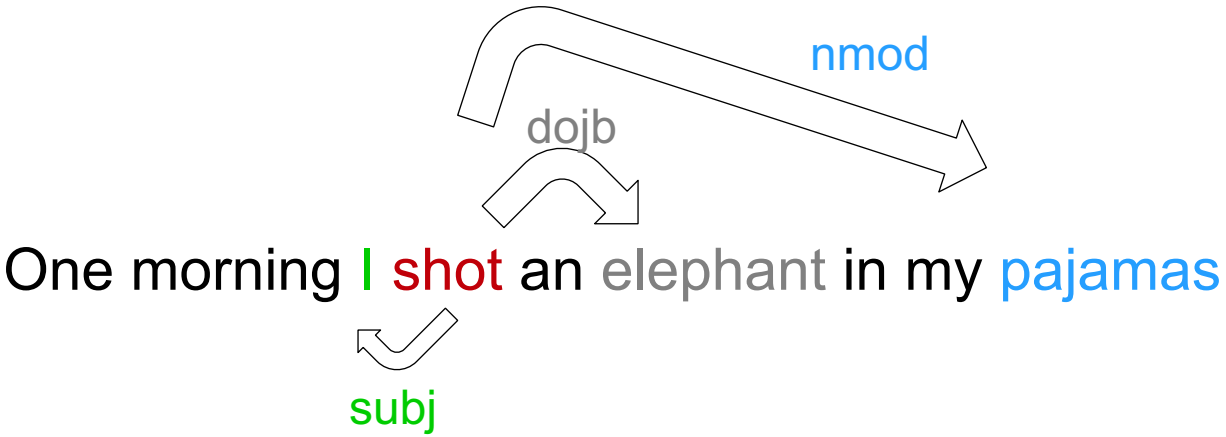
One morning I shot an elephant in my pajamas

Names entities

person

Imma let you finish but Beyonce had one of the best videos of all time

Syntax



Sentiment analysis

"Unfortunately I already had this exact picture tattooed on my chest, but this shirt is very useful in colder weather." (Overlook 1977)

Question answering

- What did Barack Obama teach?

Barack Hussein Obama II (born August 4, 1961) is an American attorney and politician who served as the 44th President of the United States from January 20, 2009, to January 20, 2017. A member of the Democratic Party, he was the first African American to serve as president. He was previously a United States Senator from Illinois and a member of the Illinois State Senate.

Obama was born in 1961 in Honolulu, Hawaii, two years after the territory was admitted to the Union as the 50th state. Raised largely in Hawaii, he also spent one year of his childhood in Washington state and four years in Indonesia. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago. In 1988, he enrolled in Harvard Law School, where he was the first black president of the *Harvard Law Review*. After graduating, he became a civil rights attorney and a professor, teaching constitutational law; at the University of Chicago Law School from 1992 to 2004.

Barack Obama

44th President of the United States

In office

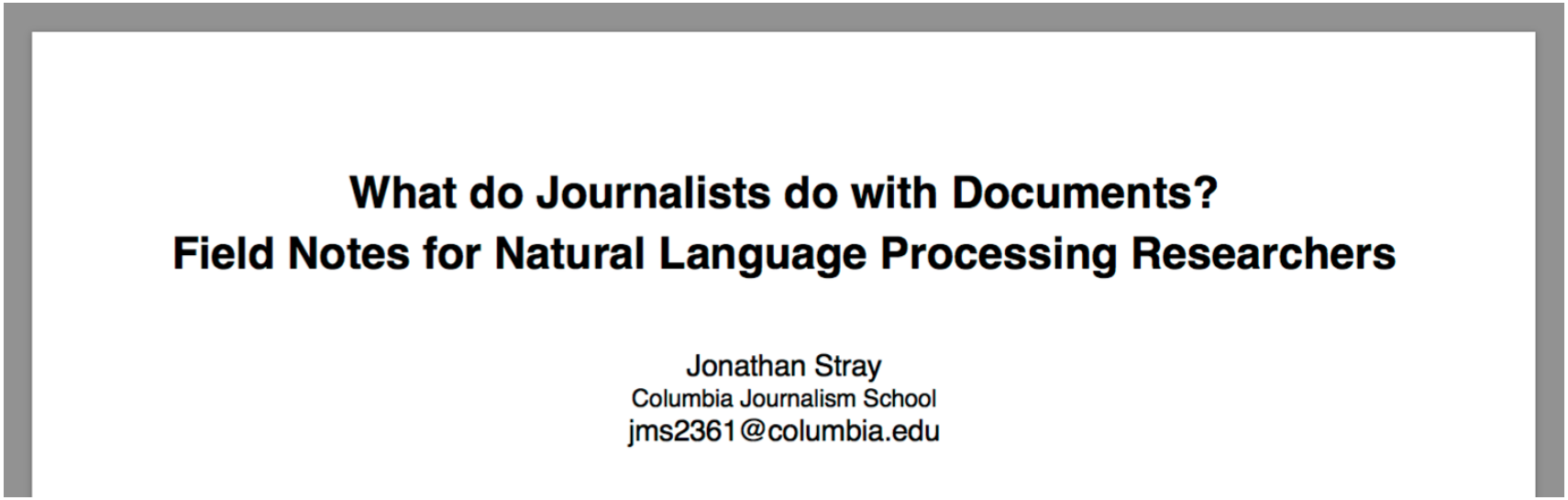
NLP

- Machine translation
- Question answering
- Information extraction
- Conversational agents
- Summerization
- … NLP + X

Google

IBM WATSON

Computation Journalism

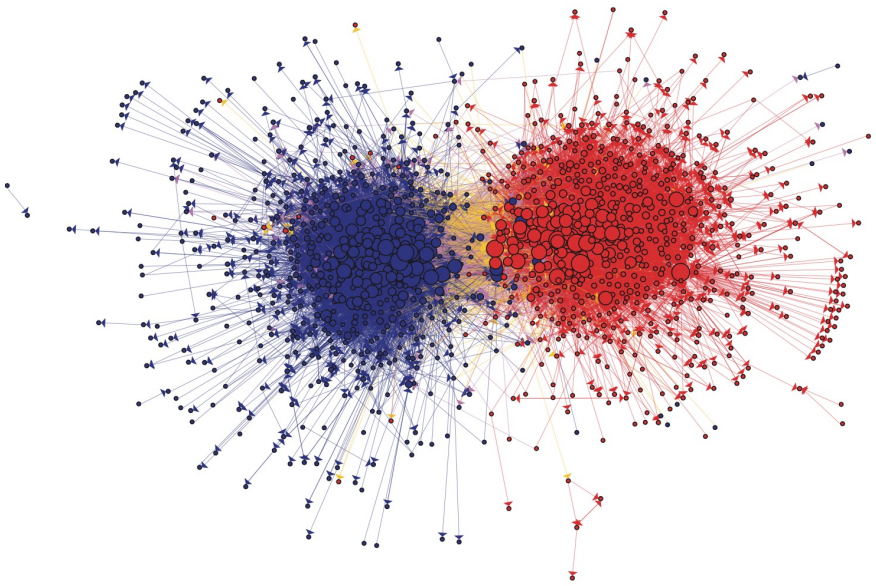


- Robust import
- Robust analysis
- Search, not exploration
- Quantitative summaries
- Interactive methods
- Clarity and Accuracy

Project JoIE: https://ir.web.th-koeln.de/projects/joie/

Computational Social Science

- Inferring ideal points of politicians based on voting behavior, speeches
- Detecting the triggers of censorship in blogs/social media
- Inferring power differentials in language use



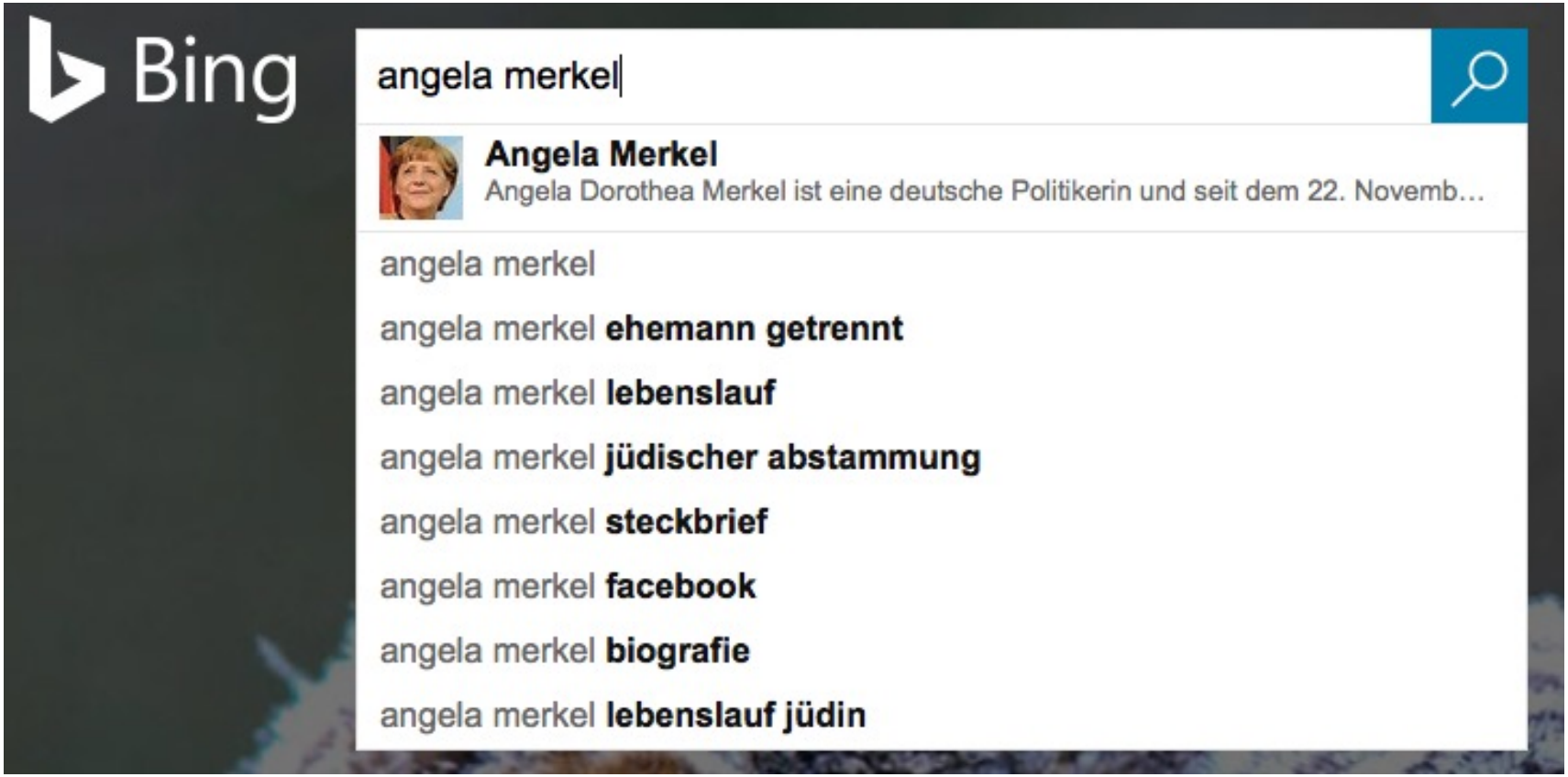
Link structure in political blogs in Adamic and Glance 2005

Project ESUPOL: https://ir.web.th-koeln.de/projects/esupol/

Topical course focus: Ethics and Bias

https://twitter.com/ilyasut/status/1359586696682160128

A long time ago... Query Expansion



http://www.unwomen.org/en/news/stories/2013/10/women-should-ads

http://www.unwomen.org/en/news/stories/2013/10/women-should-ads

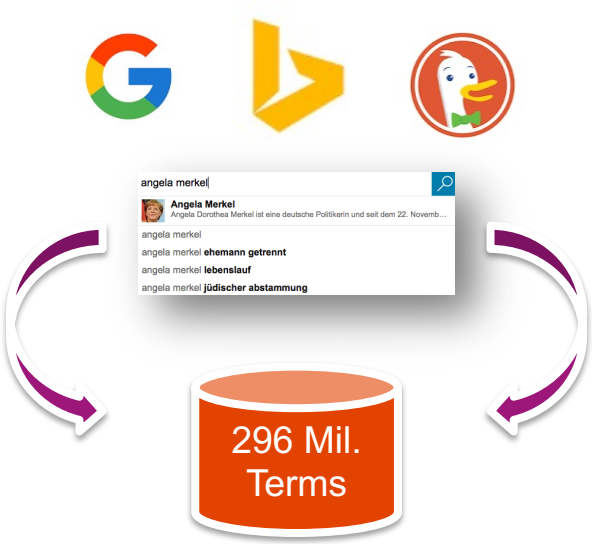
So, we crawled some data…

Since February 2017 we crawl Google, Bing, DuckDuckGo on a daily basis and gathered

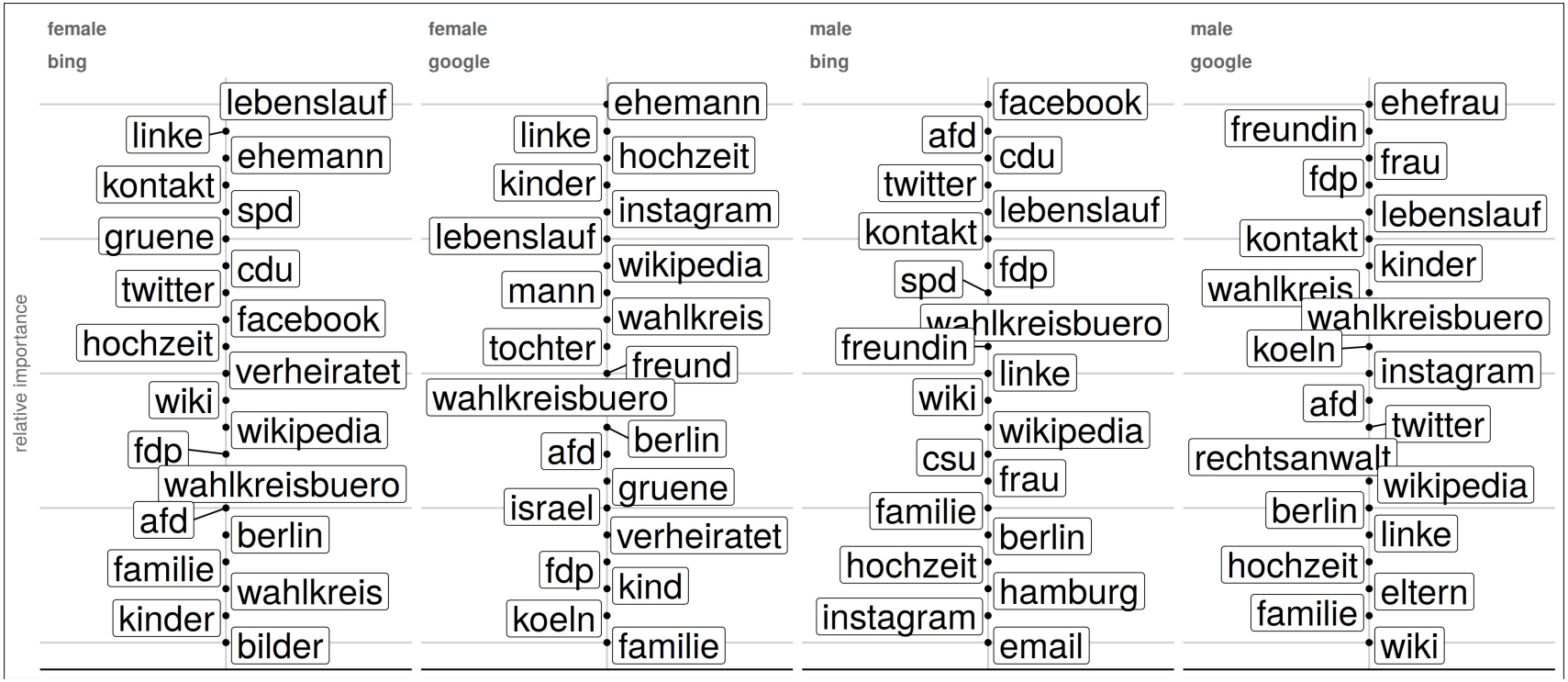
- **43.077.293 query suggestions** with
- **296.321.947 single suggestion terms**
- ~ 6,8 terms per query
- **~ 3.100 names of politicians**

Split along

- genders
- parties
- city/rural areas
- migration background (turkish parents, …)
- ...



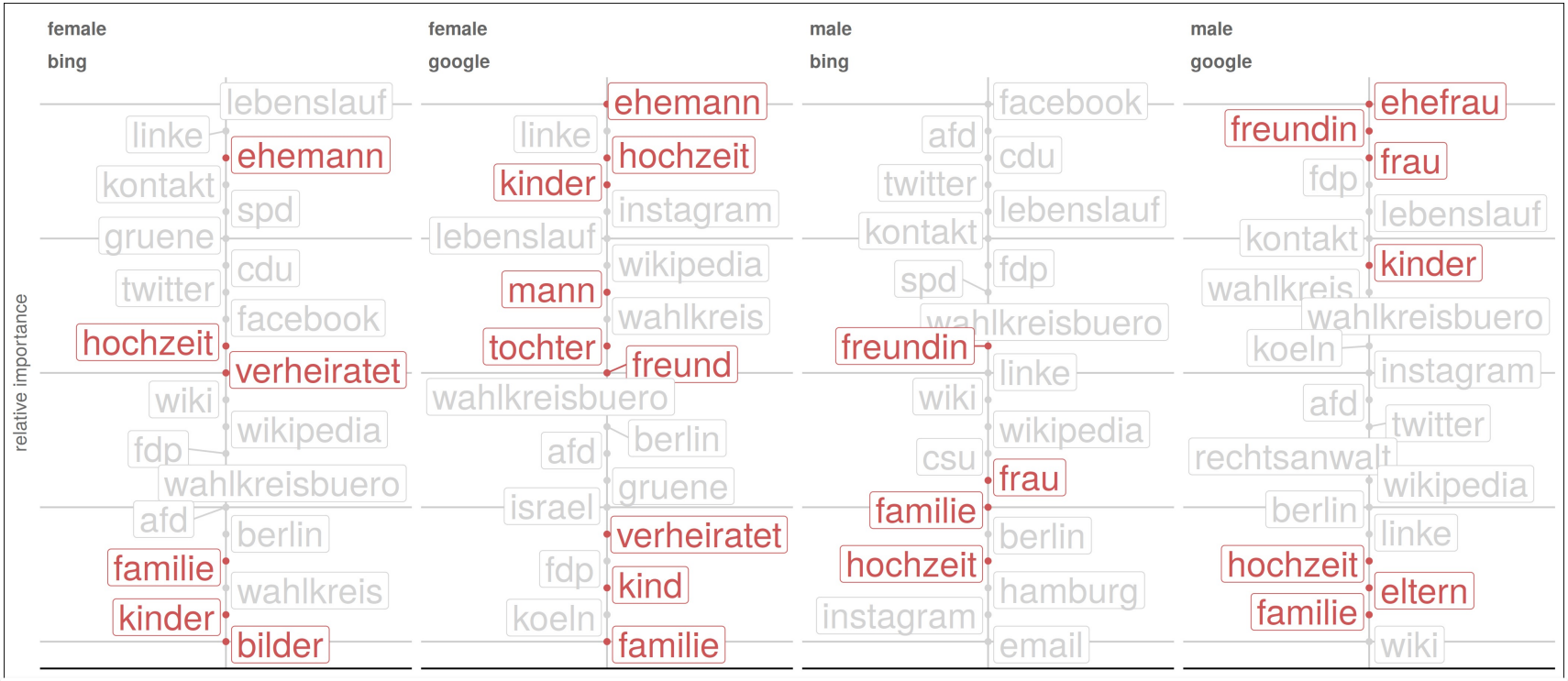
What did we find so far?



Different groups of terms emerge

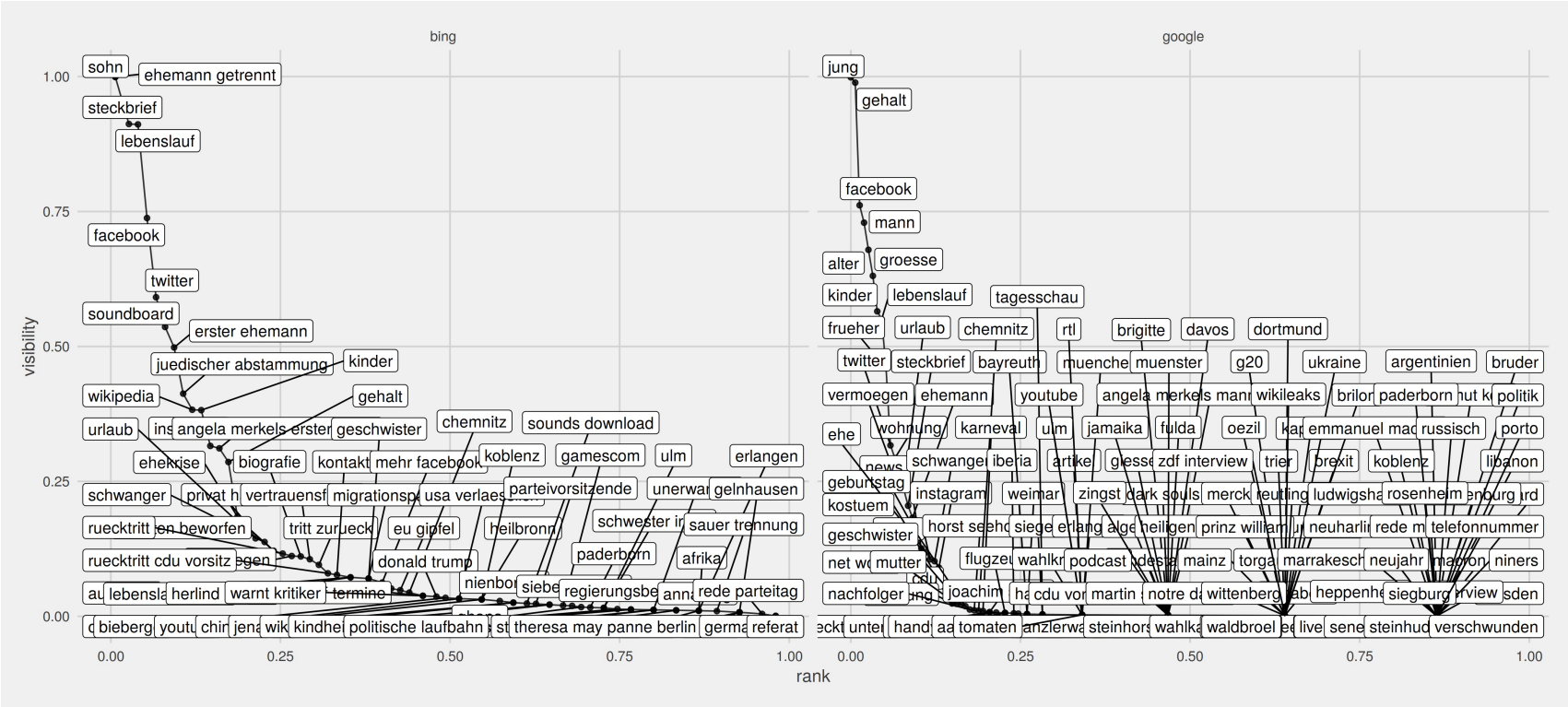
- organizational terms (party, companies, etc.)
- locations (birth town, etc.)
- private-life-related terms...

Let‘s have a closer look…

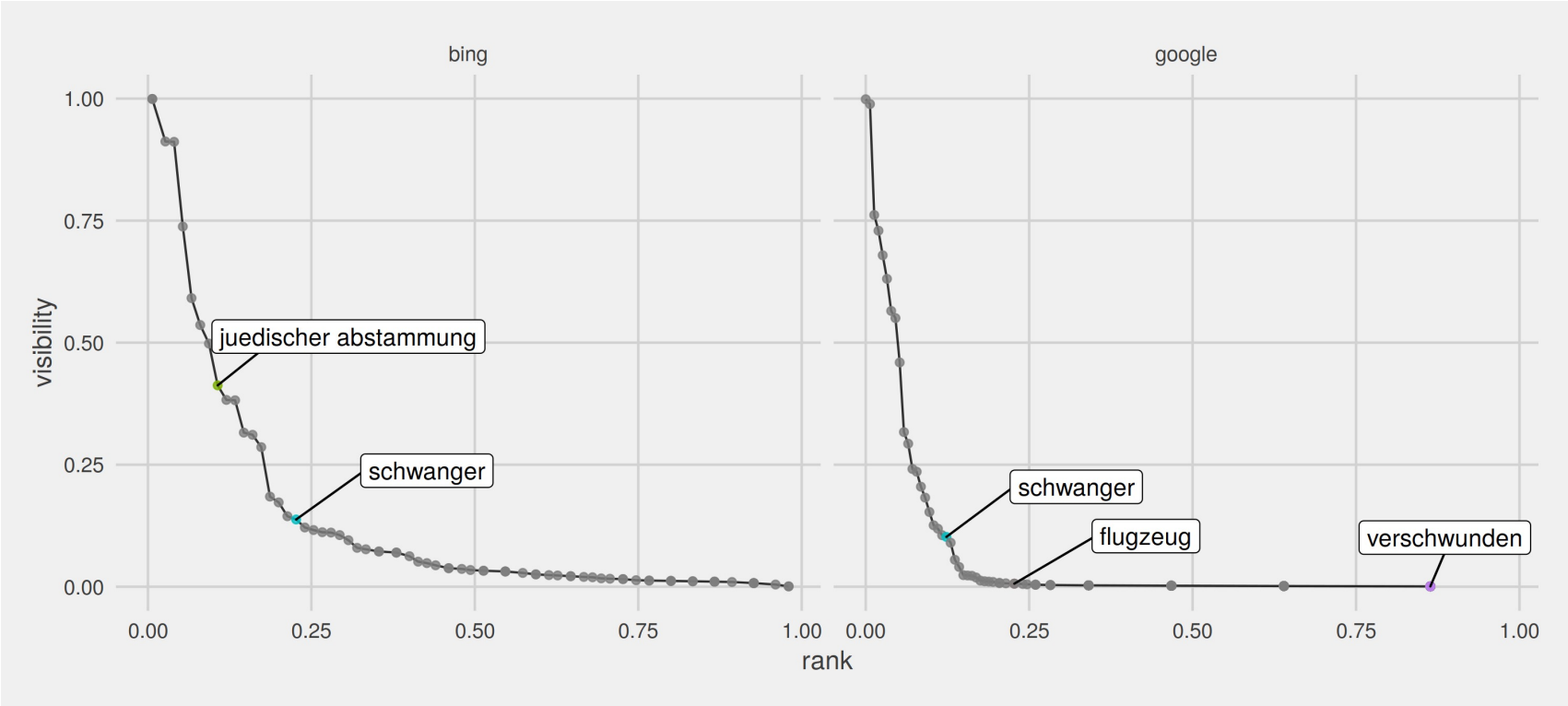


We found that
- private terms are shown statistically fewer for men
- private terms are shown statistically fewer for older people

Let‘s keep up with Angela Merkel



Better representation



We detected a set of stable terms that re-appear were frequent but others that are very rare…

Methods

- Finite state automata/transducers (tokenization, morphological analysis)
- Rule-based systems
- Probabilistic models
- Naive Bayes, Logistic regression, HMM, MEMM, CRF, language models
- Dynamic programming (combining solutions to subproblems)
- Dense representations for features/labels (generally: inputs and outputs)
- Neural networks: multiple, highly parameterized layers of interactions mediating the input/output

Course material and sources

Dan Jurafsky and **James H. Martin** (2021): Speech and Language Processing (3rd ed. draft)

- https://web.stanford.edu/~jurafsky/slp3/

Jacob Eisenstein (2019): Natural Language Processing

- https://mitpress.mit.edu/books/introduction-natural-language-processing

(some) Slides and code from **David Bamman** (Berkeley)

- https://people.ischool.berkeley.edu/~dbamman
- https://github.com/dbamman/anlp19