



# Natural Language Processing

## 01: Introduction

---

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: 2021-04-08

# Information Retrieval Research Group



## Prof. Dr. Philipp Schaer

- Information retrieval, evaluation of IR systems, digital libraries



## Timo Breuer, M.Sc.

- Living Labs Infrastructure for Information Retrieval, project STELLA



## Fabian Haak, M.Sc.

- Sentiment Analysis, Query Expansion, Political Retrieval, project ESUPOL



## Björn Engelmann, M.Sc.

- Information Extraction, Machine Learning, Scientific Journalism, project JoIE

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

# Information Retrieval Research Group



## Prof. Dr. Philipp Schaer

- Information retrieval, evaluation of IR systems, digital libraries



## Timo Breuer, M.Sc.

- Living Labs Infrastructure for Information Retrieval, project STELLA



## Fabian Haak, M.Sc.

- Sentiment Analysis, Query Expansion, Political Retrieval, project ESUPOL



## Björn Engelmann, M.Sc.

- Information Extraction, Machine Learning, Scientific Journalism, project JoIE

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

# Syllabus – Exams and grades

These are the parts for your final grade

- One **presentation on course-related materials** (25%)
- **Course project**
  - **Presentation** on concept and 1<sup>st</sup> milestone (25%)
  - Final workshop-grade **paper** incl. code and data (50%)
- Grading will be based on a Rubic table to ensure transparency
- I use the following grading scale:

<50	50	55	60	65	70	75	80	85	90	95
fail	4,0	3,7	3,3	3,0	2,7	2,3	2,0	1,7	1,3	1,0

# Syllabus – Course organization

This course is split in two parts:

- **Lecture and tutorial**

The lecture will be online, once a week

- Zoom + GitHub
- I will try to give you access to slides and material as early as possible, but I most likely will not be able to pre-record all lectures.

Each topic will be complemented with a tutorial

- Zoom + GitHub
- Please **prepare @ home**
- In the tutorial sessions we will assume that you are familiar with the basics. You **have to** work in groups of 3 to 4.

# Schedule for Summer Term 2021

Date	Lecture	Topics	Tutorial
09.04.21	1	Introduction, overview	
16.04.21	2	<b>Basic text processing</b>	nltk, numpy, scikit-learn
23.04.21	3	<b>WordNet</b>	WordNet, GermaNet
30.04.21	4	<b>Vector semantics</b>	word2vec, glove
07.05.21	5	<b>Ethics and truth in NLP</b>	Paper reading
14.05.21	6	<i>Homework: Video Lectures about <b>Bias in NLP</b> / group work</i>	
21.05.21		Project week	
28.05.21	7	<i>Student talks – Paper presentation</i>	
04.06.21	8	<b>Sentiment analysis</b>	Working on project
11.06.21	9	<b>Information Extraction</b>	Working on project
18.06.21	10	<b>Evaluation of IE</b>	Working on project
25.06.21	11	<b>Annotations (Guest)</b>	Working on project
02.07.21	12	<i>Student talks – Milestone presentation</i>	
31.08.21		Submit term paper	

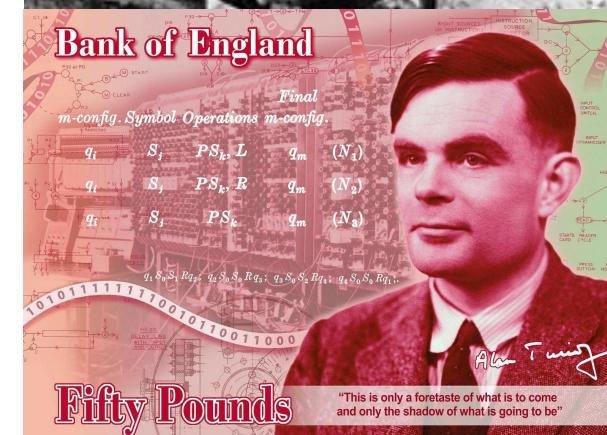
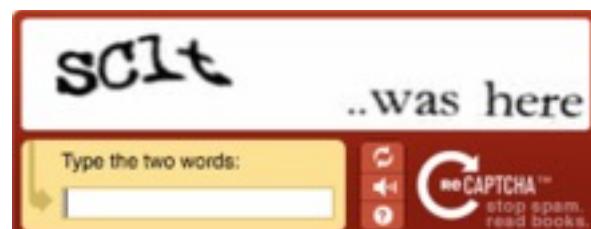
# NLP is interdisciplinary

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

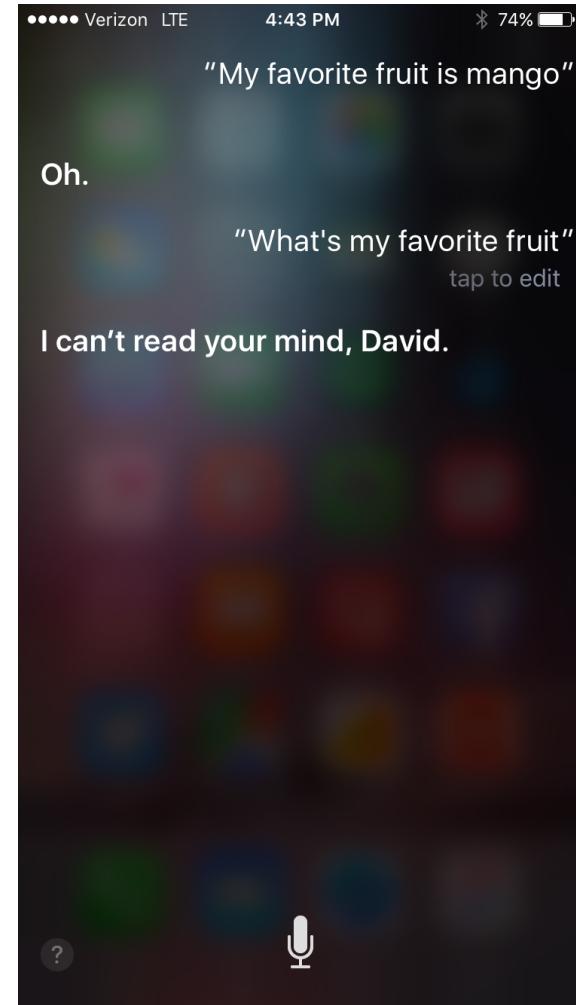
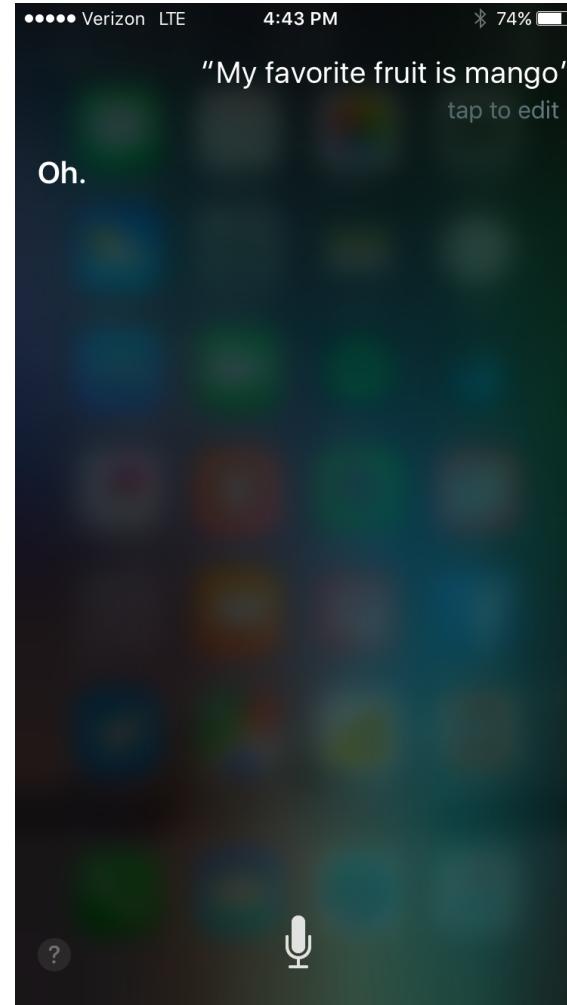
NLP = processing language with computers

# Turing Test

- Distinguishing human vs. computer only through written language (Turing 1950)
- Today:  
*completely automated public Turing test to tell computers and humans apart*



# 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.

# What makes language hard?

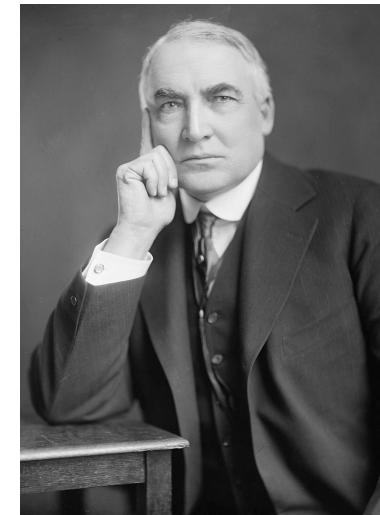
- Language is a complex social process
- Tremendous ambiguity at every level of representation
- Modeling it is AI-complete (requires first solving general AI)

# 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”

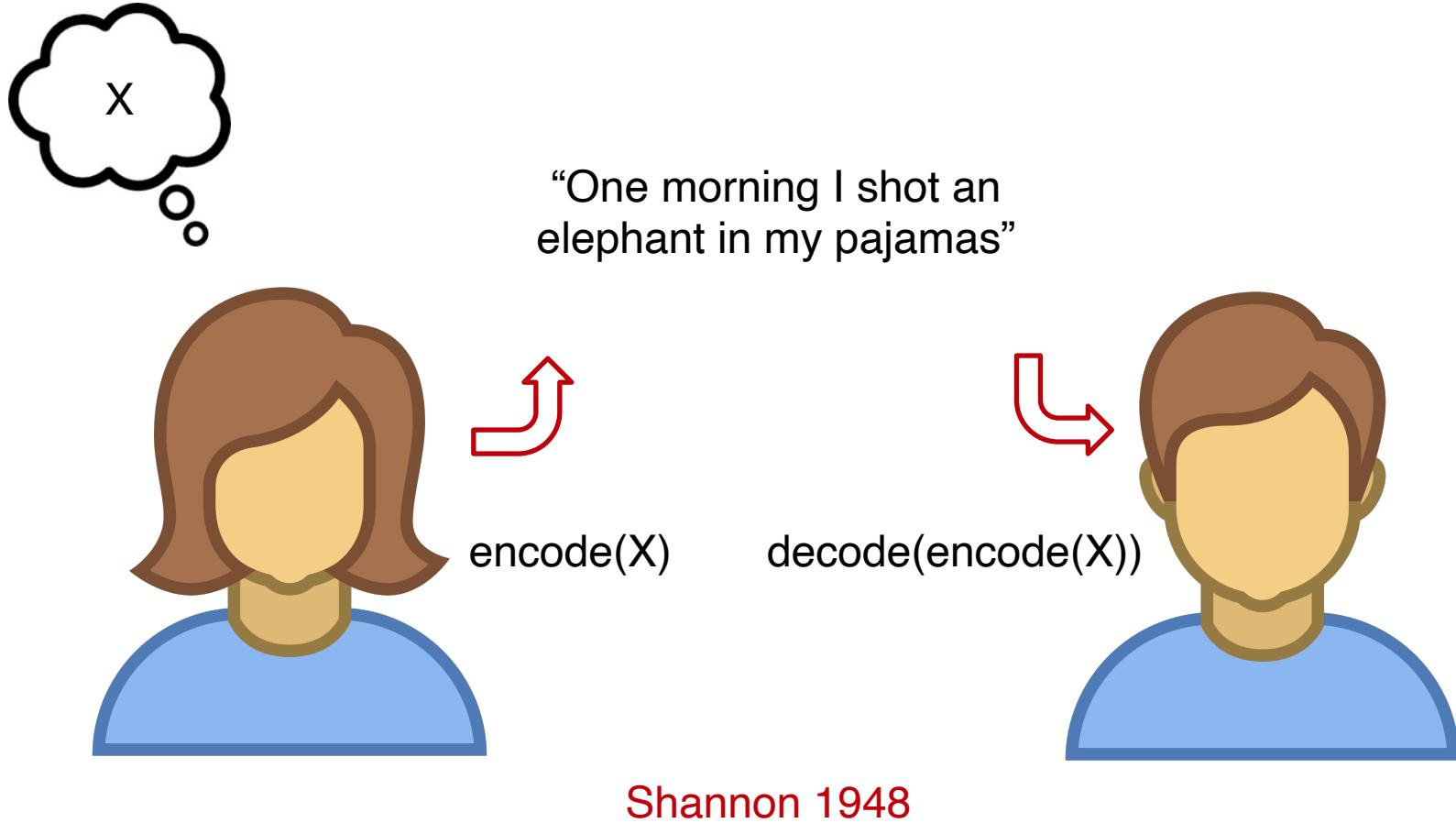


*Animal Crackers*

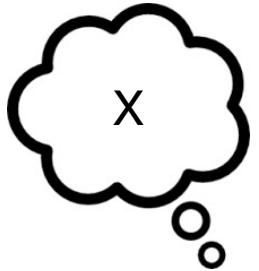
# Processing as Representation

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

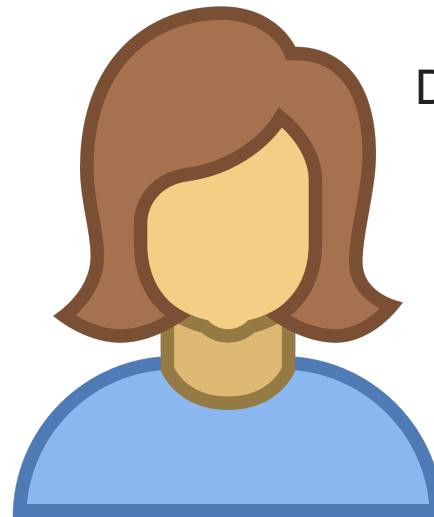
# Information theoretic view



# Rational speech act view

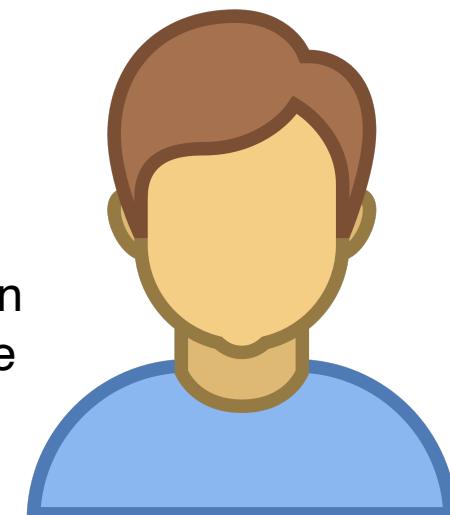


“One morning I shot an elephant in my pajamas”



Dionysios (Dennis) Tsichritzis

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



Frank and Goodman 2012

# Pragmatic view



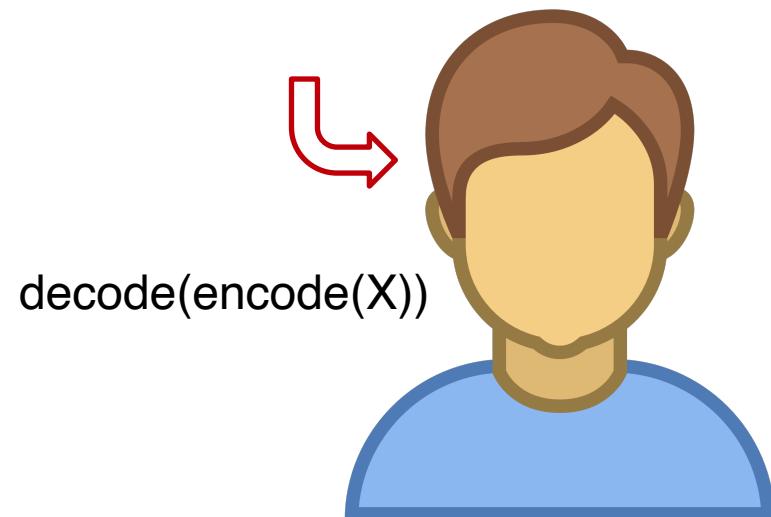
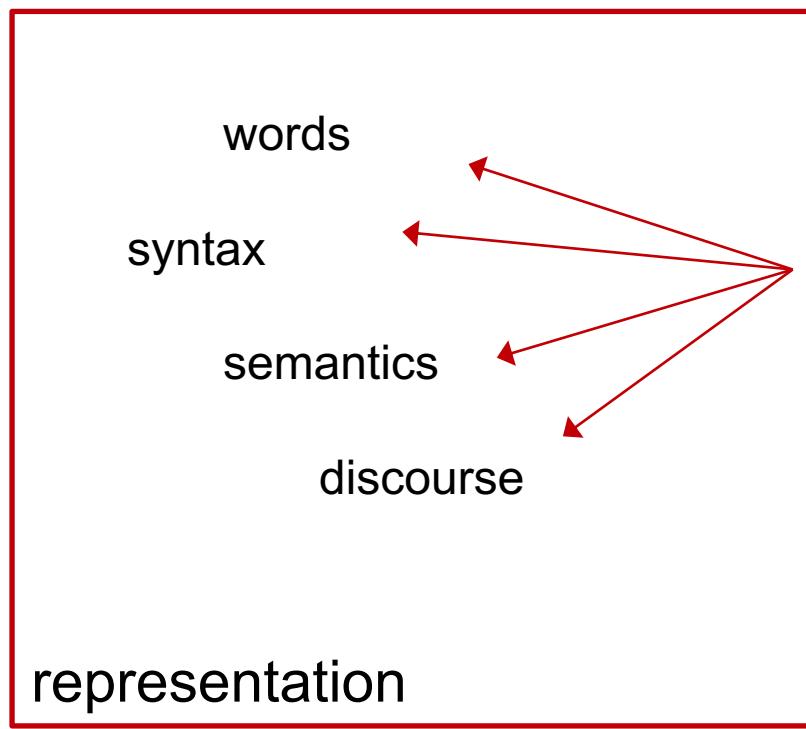
“One morning I shot an elephant in my pajamas”

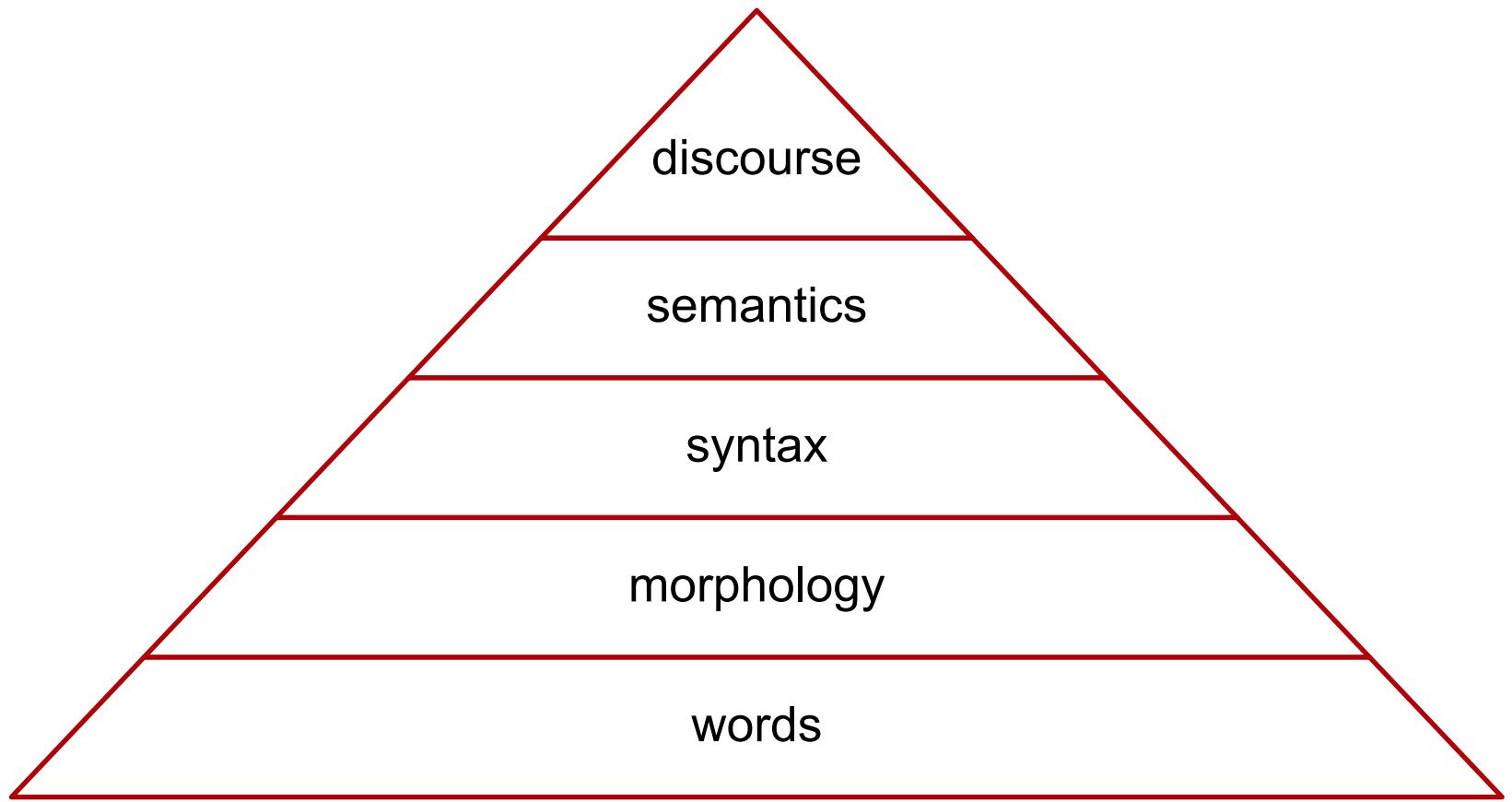
Meaning is co-constructed by  
the interlocutors and the  
**context** of the utterance



# Decoding

“One morning I shot an elephant in my pajamas”





# 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

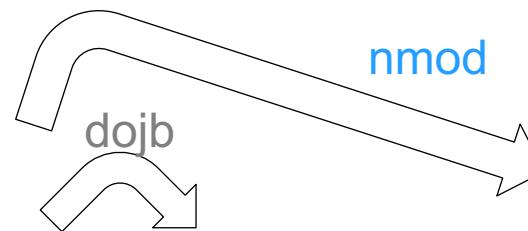
One morning I shot an elephant in my pajamas

# Names entities

person

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

# Syntax



One morning I shot an elephant in my pajamas



# 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 [constitutional law](#); at the [University of Chicago Law School](#) from 1992 to 2004.

**Barack Obama**



44th [President of the United States](#)

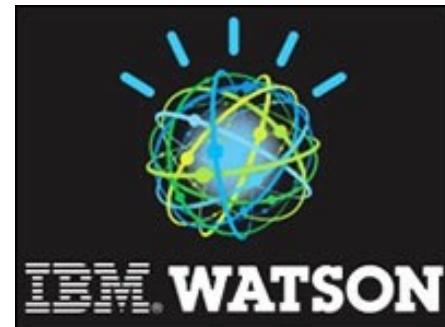
In office

# NLP

- Machine translation
- Question answering
- Information extraction
- Conversational agents
- Summarization
- ... NLP + X



Google



# Computation Journalism

## **What do Journalists do with Documents? Field Notes for Natural Language Processing Researchers**

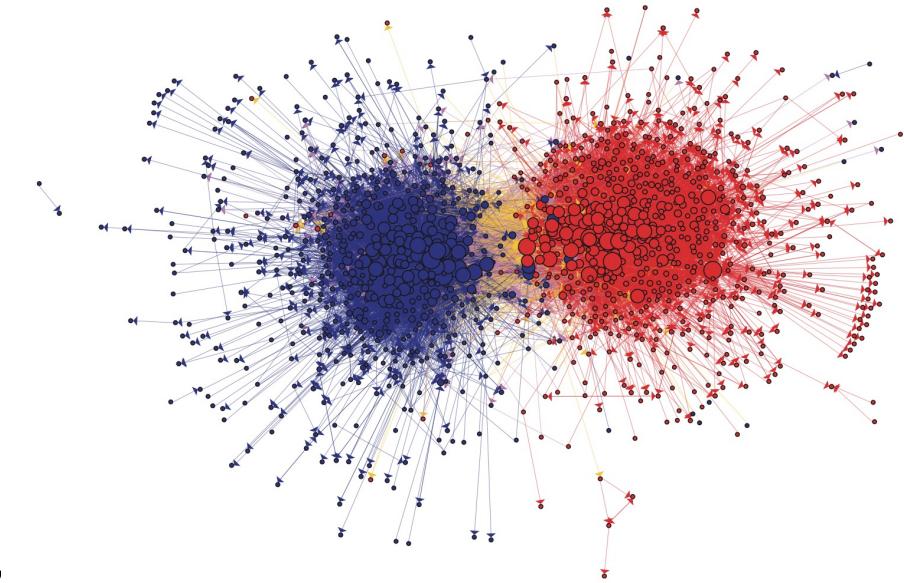
Jonathan Stray  
Columbia Journalism School  
jms2361@columbia.edu

- 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



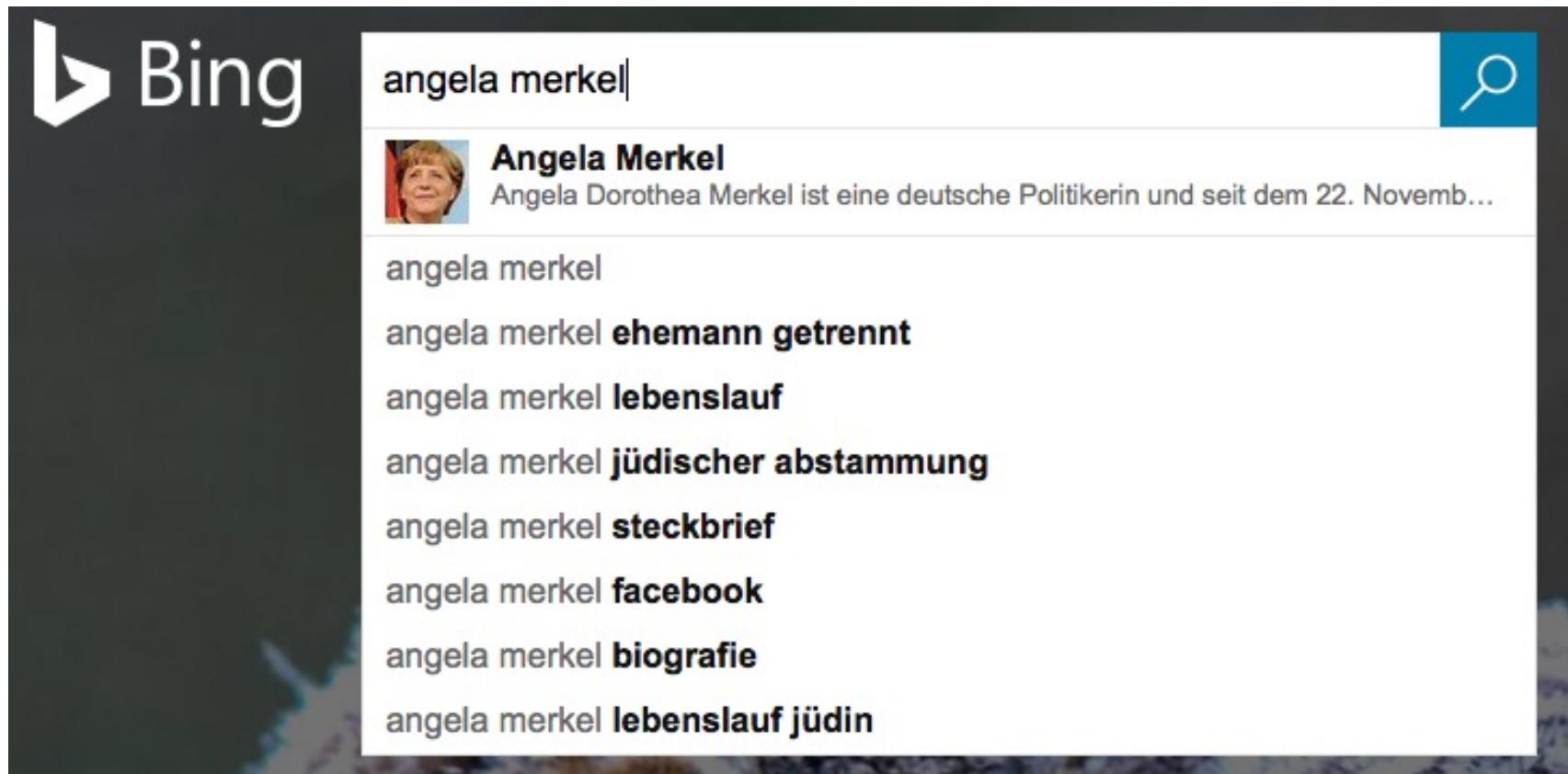
Ilya Sutskever  
@ilyasut

...

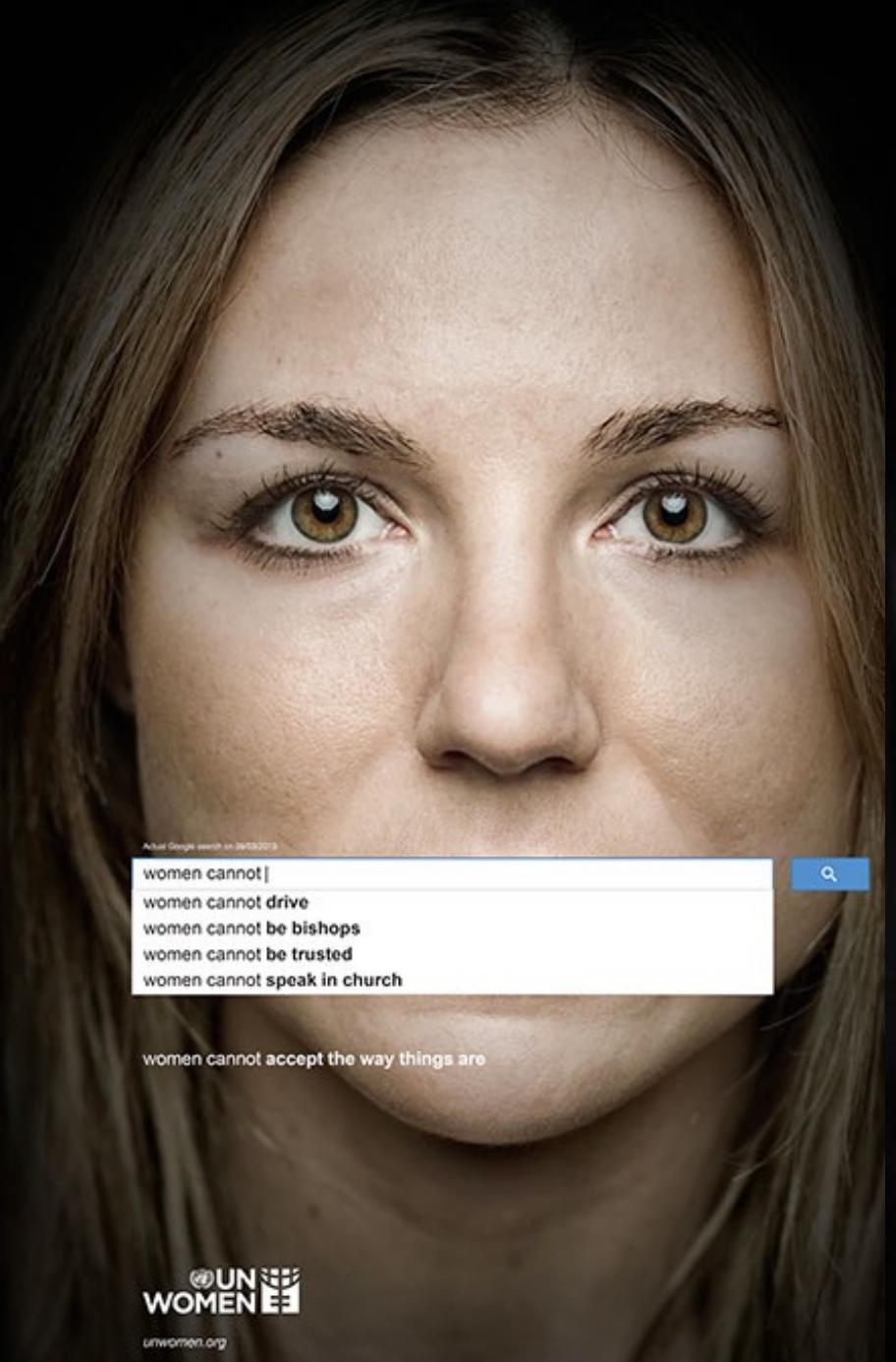
Real progress in AI can only be  
achieved through 🔥 very intense  
~~work ethic~~ S

11:35 AM · Feb 10, 2021 · Twitter Web App

# A long time ago... Query Expansion



The image shows a screenshot of a Bing search results page. The search query "angela merkel" is entered into the search bar. Below the search bar, there is a snippet for "Angela Merkel" featuring a small portrait photo and a brief description: "Angela Dorothea Merkel ist eine deutsche Politikerin und seit dem 22. Novemb...". To the right of the search bar is a blue magnifying glass icon. Below the snippet, a list of suggested or related search terms is displayed, each preceded by "angela merkel": "angela merkel", "angela merkel **ehemann getrennt**", "angela merkel **lebenslauf**", "angela merkel **jüdischer abstammung**", "angela merkel **steckbrief**", "angela merkel **facebook**", "angela merkel **biografie**", and "angela merkel **lebenslauf jüdin**". The background of the screenshot is dark, and the overall interface is characteristic of the Bing search engine.



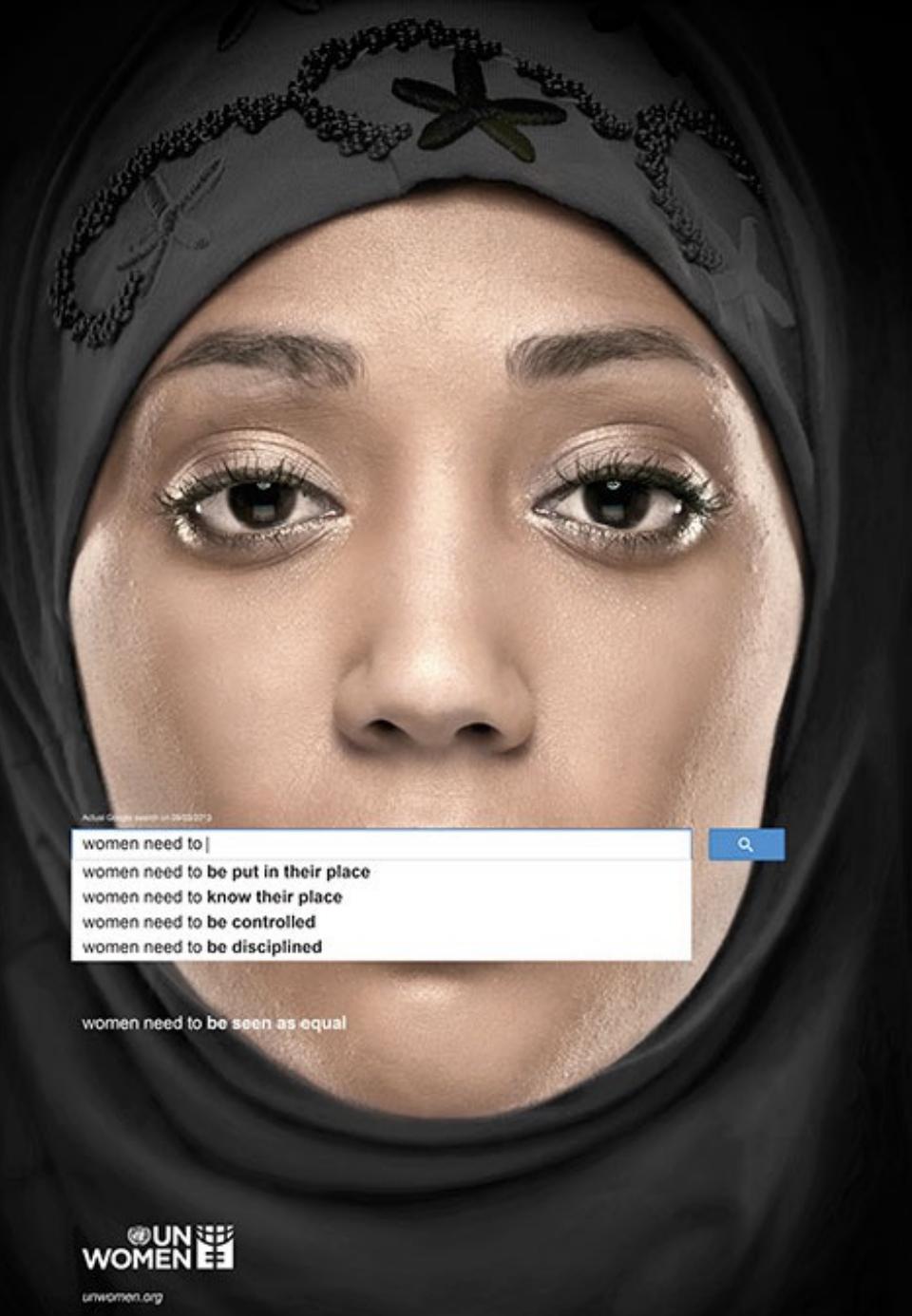
Actual Google search on 26/03/2013

women cannot |  
women cannot drive  
women cannot be bishops  
women cannot be trusted  
women cannot speak in church

women cannot accept the way things are



unwomen.org



Actual Google search on 26/03/2013

women need to |  
women need to be put in their place  
women need to know their place  
women need to be controlled  
women need to be disciplined

women need to be seen as equal



unwomen.org



unwomen.org



unwoman.org

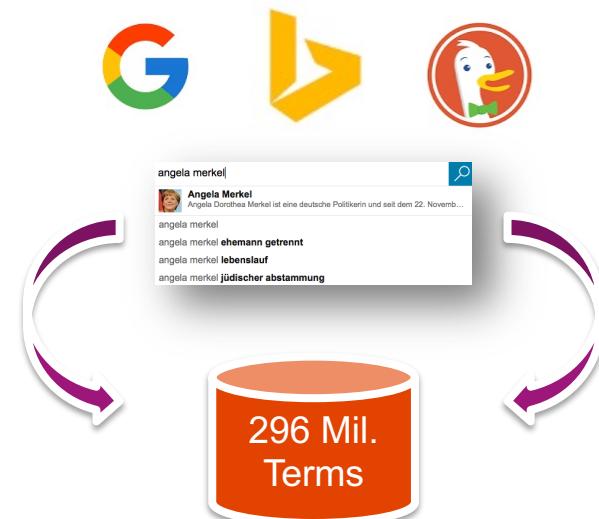
# So, we crawled some data...

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

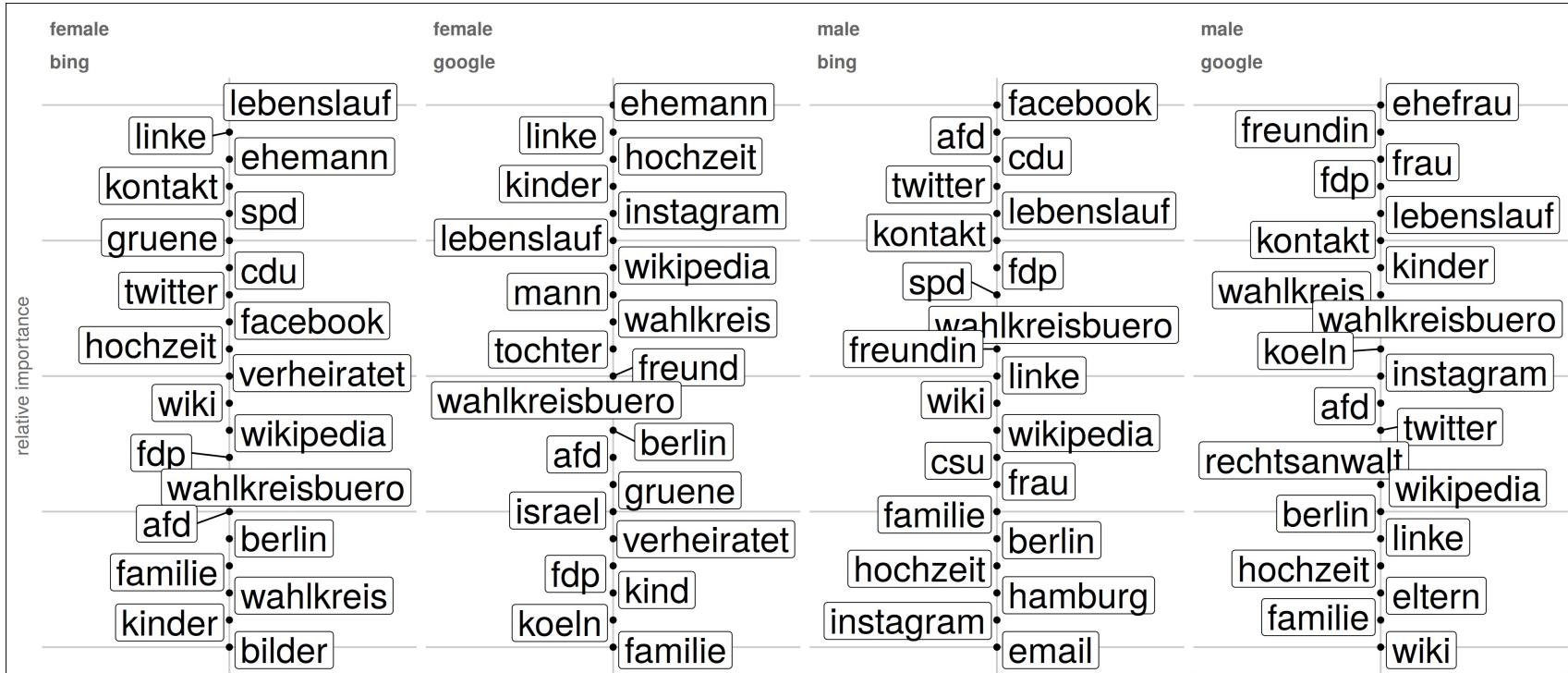
- **43.077.293 query suggestions** with
- **296.321.947 single suggestion terms**
- **~ 6,8 terms per query**
- **~ 3.100 names of politicians**

Split along

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



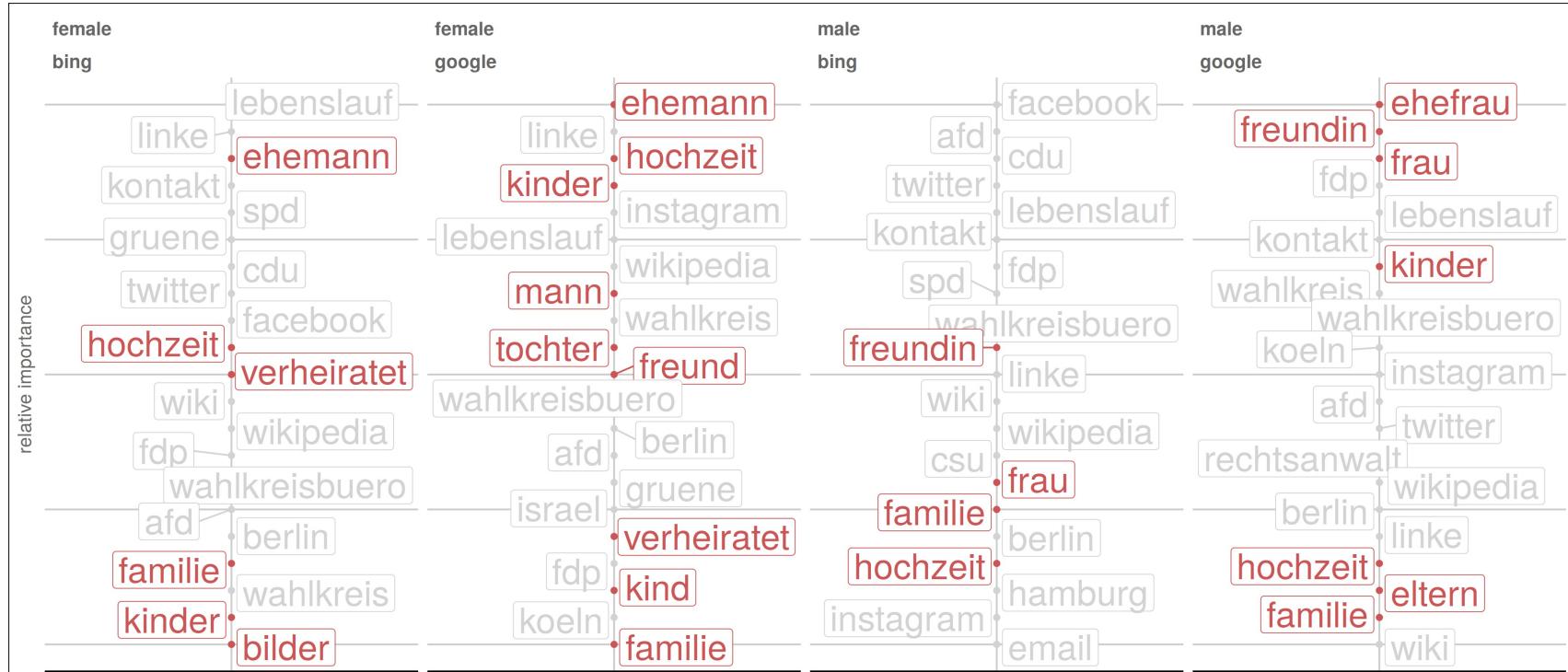
# What did we find so far?



Different groups of terms emerge

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

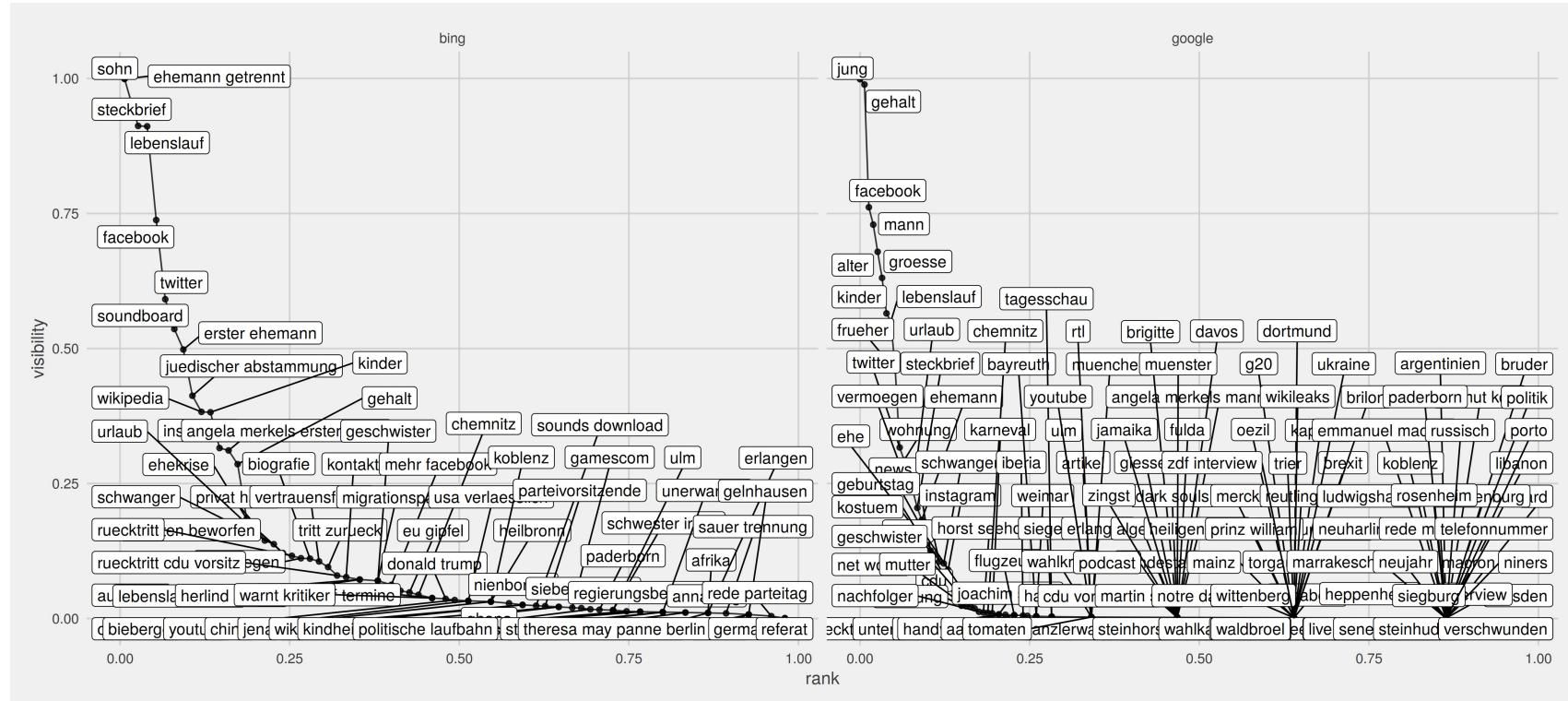
# Let's have a closer look...



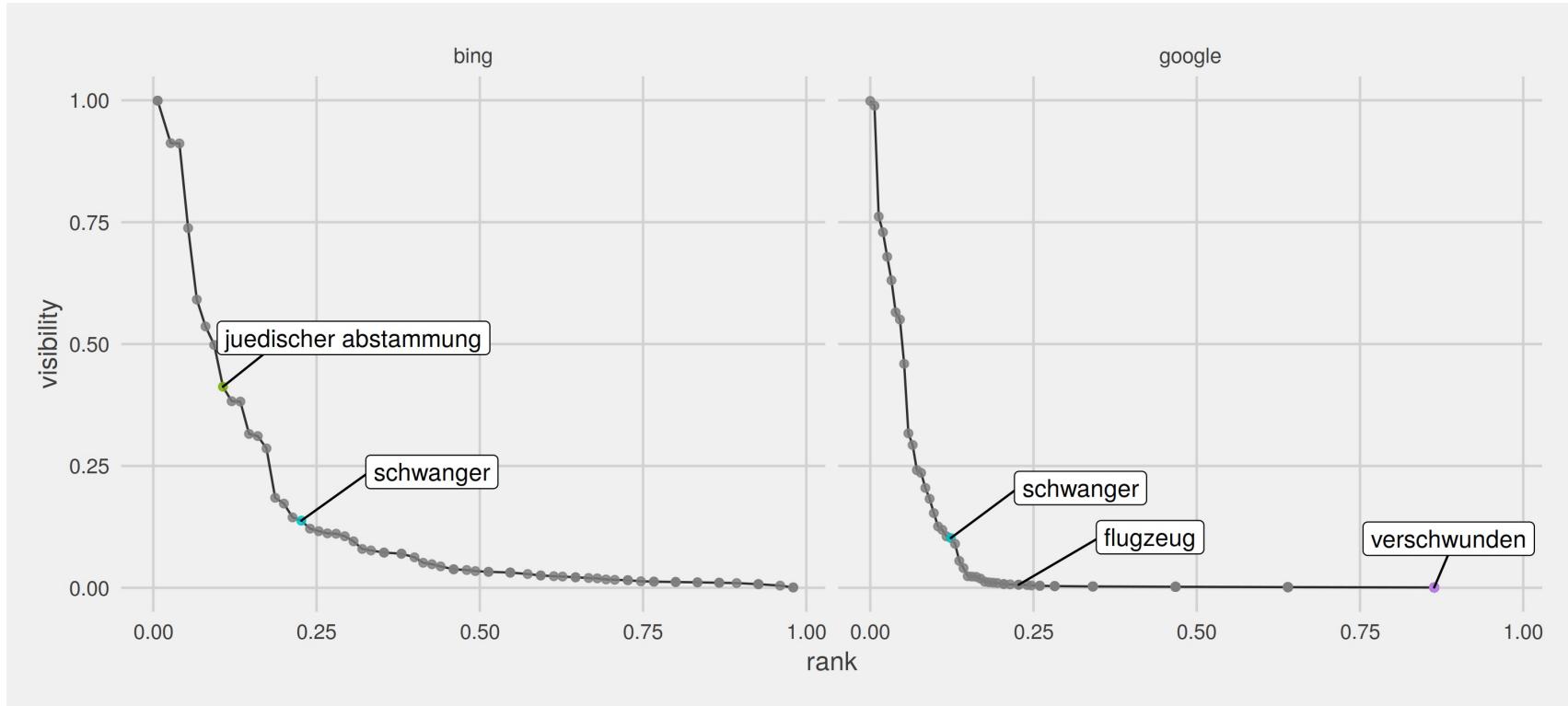
We found that

- 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>