Grundlagen des Information Retrieval

Vorbesprechung - 188.977

GIR – 2019/20

Lecturers

Allan Hanbury

Sebastian Hofstätter

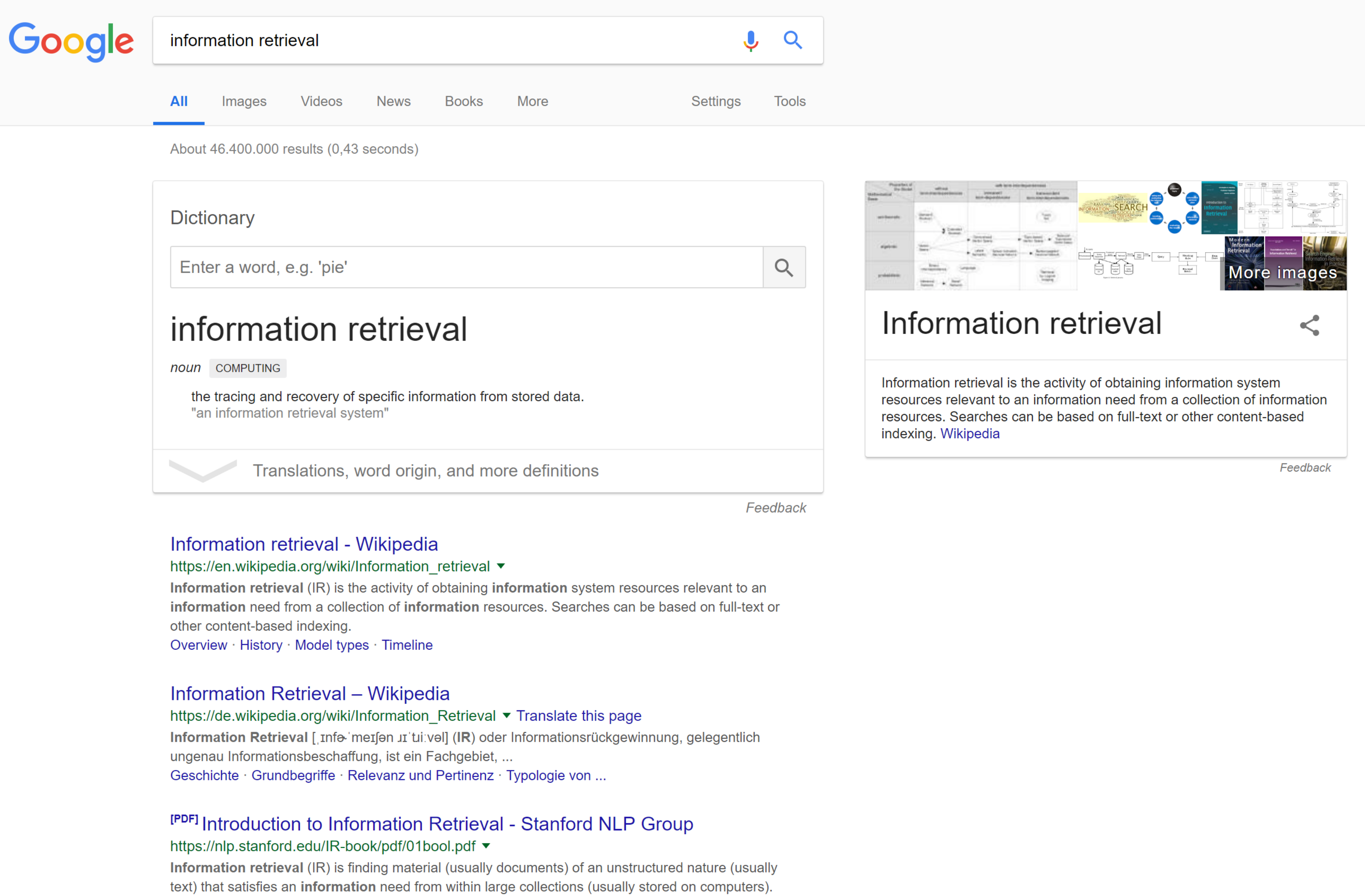
Mihai Lupu

Andreas Rauber

Contact / Questions

Use the TUWEL Forum

or write an email to sebastian.hofstaetter@tuwien.ac.at



About 46.400.000 results (0,43 seconds)

Dictionary

information retrieval

noun COMPUTING

the tracing and recovery of specific information from stored data.
"an information retrieval system"

Translations, word origin, and more definitions

Feedback

Information retrieval - Wikipedia
https://en.wikipedia.org/wiki/Information_retrieval ▾
Information retrieval (IR) is the activity of obtaining **information** system resources relevant to an **information** need from a collection of **information** resources. Searches can be based on full-text or other content-based indexing.
Overview · History · Model types · Timeline

Information Retrieval – Wikipedia
https://de.wikipedia.org/wiki/Information_Retrieval ▾ Translate this page
Information Retrieval [ˌɪnfɚˈmeɪʃən ɹɪˈtiːvəl] (**IR**) oder Informationsrückgewinnung, gelegentlich ungenau Informationsbeschaffung, ist ein Fachgebiet, ...
Geschichte · Grundbegriffe · Relevanz und Pertinenz · Typologie von ...

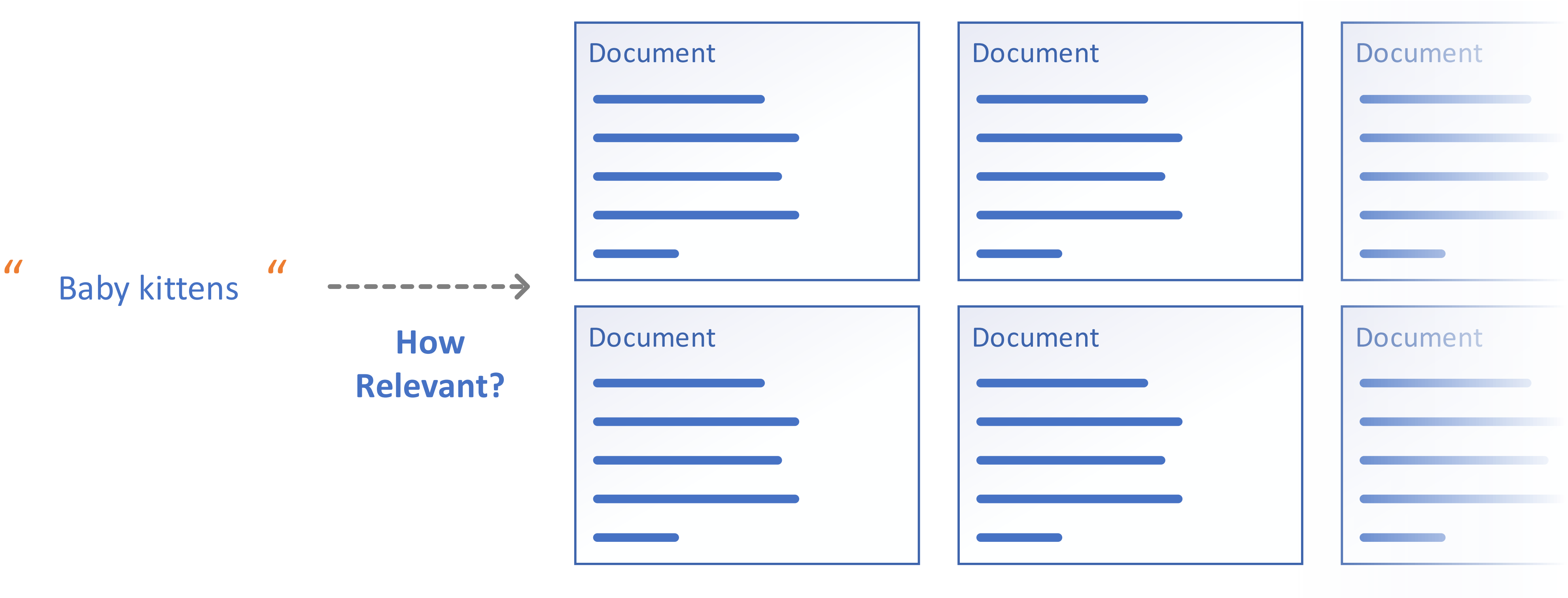
[PDF] Introduction to Information Retrieval - Stanford NLP Group
https://nlp.stanford.edu/IR-book/pdf/01bool.pdf ▾
Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an **information** need from within large collections (usually stored on computers).

Information retrieval

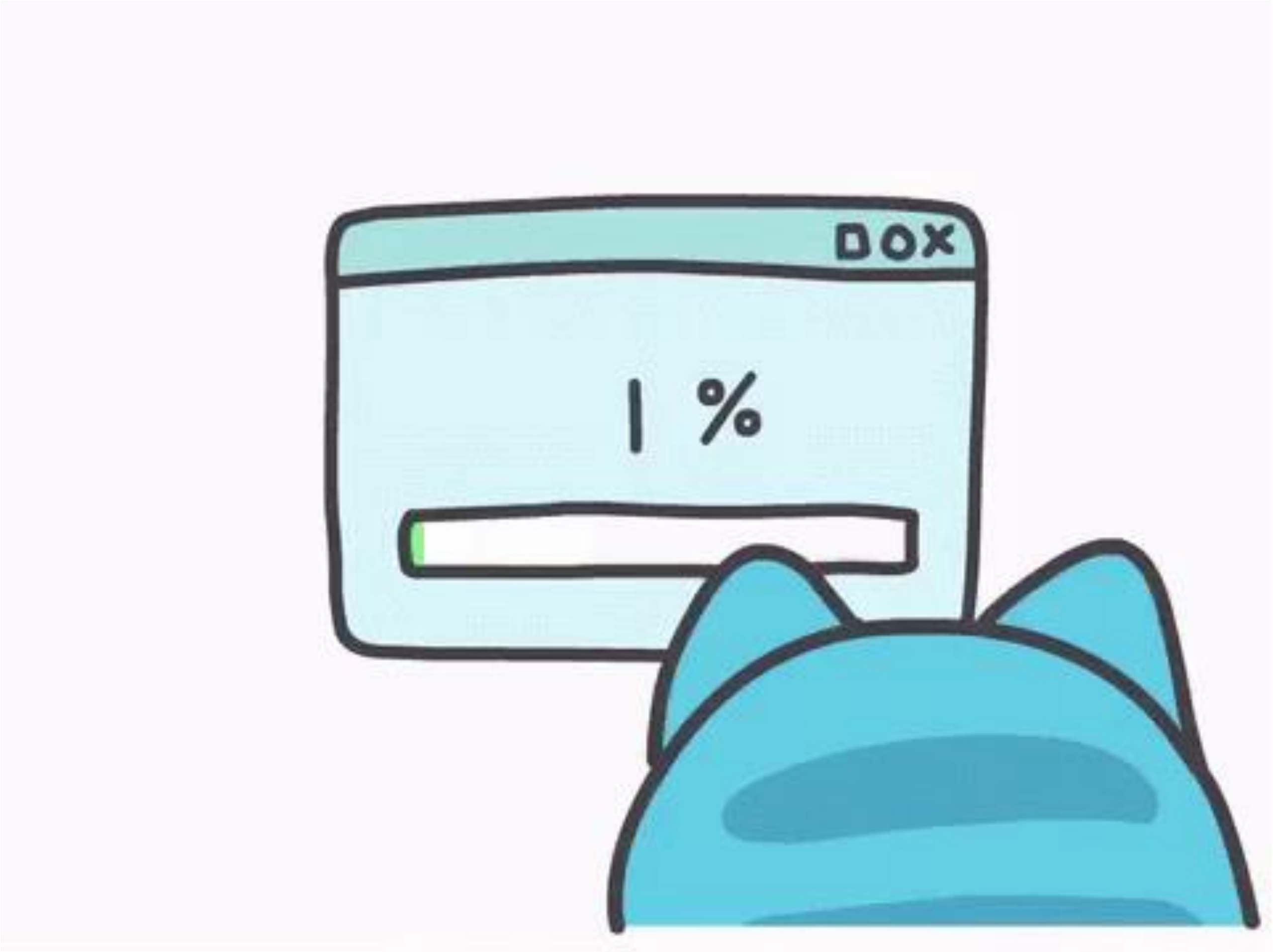
Information retrieval is the activity of obtaining information system resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing. Wikipedia

Feedback

Information Retrieval (Finding the needle in the haystack)



Reality …



From: https://twitter.com/SwiftOnSecurity/status/969213410490675201

Organization

Lectures, Exercises, Grading

Lectures / Content

Thursday, 16:00 – 18:00, EI 3A

3.10.	Vorbesprechung
10.10.	Foundations of Information Retrieval
17.10.	Efficient text processing + Exercise 1 Infos
24.10.	Scoring and search
31.10.	Evaluation
7.11.	Web search
14.11.	Search interfaces
TBD	Music and image retrieval

Exercises

- Two exercise in total (#1: search engine, #2: music retrieval)

- Shared group for all exercises
  - 2 persons per group (managed via TUWEL)
  - First exercise will be evaluated in an interview (Abgabegespräch)

- All exercises share 1 private GitHub repository (via GitHub classroom)

- Lot’s of bonus point opportunities in #1 🎉👏💯✨

Exercise #1

- Implement your own search engine!
  - With lots of freedom and lots of bonus points 👌💯

- Index a subset of Wikipedia and make it searchable
  - And evaluate with given queries
- Focus on efficiency and correctness

- **More on this in the lecture @ 17.10. 👍**
  - Practical how to: write efficient text processing code
  - Fixing teams
  - Presentation of the exercise and QA

Exam

- Style: open answer questions

- We provide 2 dates:

  - **12.12.19** 13:00 FAV 1 (Hörsaal Favoritenstraße)
  - **9.1.20** 10:00 FAV 1 (Hörsaal Favoritenstraße)

Grading

Exercise 1 (Search engine):	50%
Exercise 3 (Music IR):	10%
Exam:	40% (min 30% to pass)
Total	**100% (min 50% to pass)**

Sounds cool?

Advanced Information Retrieval

- Available next semester!
- Content (Very fancy machine learning ☺)
  - Word Embeddings
  - Neural Networks for NLP
  - Neural IR
  - State of the art developments
- Exercise:
  - Implement different Neural IR models in PyTorch & create creative visualizations

Responsibility

+ Classroom Discussion

Responsibility – Social Impact of Ranking

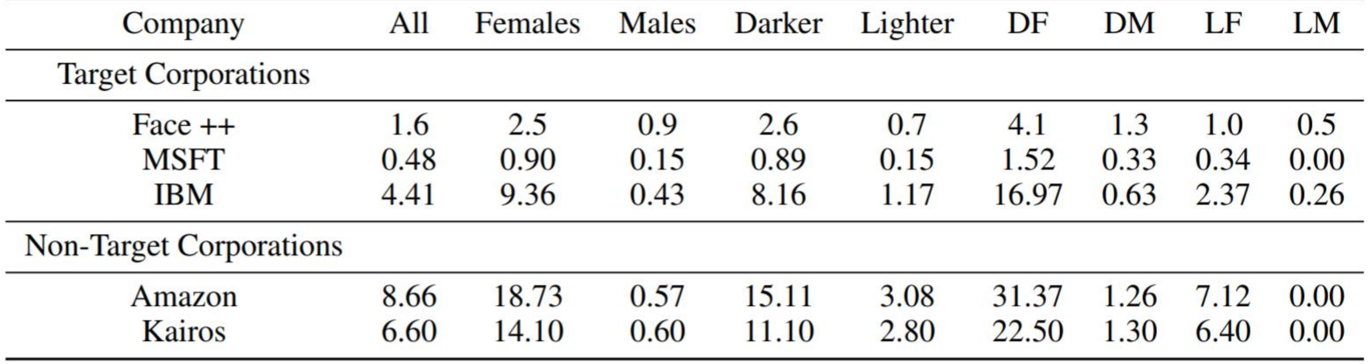
- Recommendations are optimized for time spent on a platform (= revenue)
- Easy to fall down the rabbit hole – because scandalous & “click here to find the truth” videos keeps you on the platform
- Multilingual problems: Manually blocking English content does not automatically translate to other languages

How YouTube Radicalized Brazil https://www.nytimes.com/2019/08/11/world/americas/youtube-brazil.html

Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, Ed Chi. 2019. Recommending What Video to Watch Next: A Multitask Ranking System. RecSys ’19

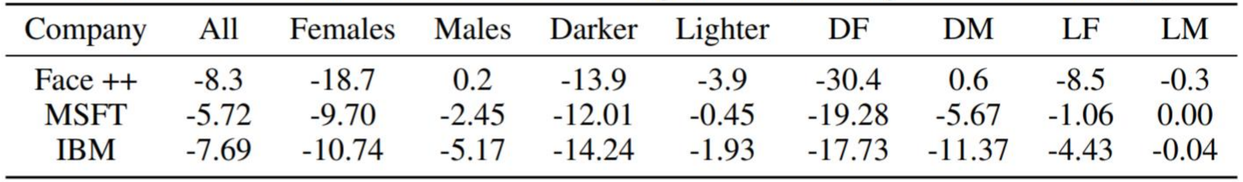
Responsibility – Facial Recognition Error

Table 1: Overall Error on Pilot Parliaments Benchmark, August 2018 (%)



Company	All	Females	Males	Darker	Lighter	DF	DM	LF	LM
Target Corporations									
Face ++	1.6	2.5	0.9	2.6	0.7	4.1	1.3	1.0	0.5
MSFT	0.48	0.90	0.15	0.89	0.15	1.52	0.33	0.34	0.00
IBM	4.41	9.36	0.43	8.16	1.17	16.97	0.63	2.37	0.26
Non-Target Corporations									
Amazon	8.66	18.73	0.57	15.11	3.08	31.37	1.26	7.12	0.00
Kairos	6.60	14.10	0.60	11.10	2.80	22.50	1.30	6.40	0.00

Table 2: Overall Error Difference Between August 2018 and May 2017 PPB Audit (%)



Company	All	Females	Males	Darker	Lighter	DF	DM	LF	LM
Face ++	-8.3	-18.7	0.2	-13.9	-3.9	-30.4	0.6	-8.5	-0.3
MSFT	-5.72	-9.70	-2.45	-12.01	-0.45	-19.28	-5.67	-1.06	0.00
IBM	-7.69	-10.74	-5.17	-14.24	-1.93	-17.73	-11.37	-4.43	-0.04

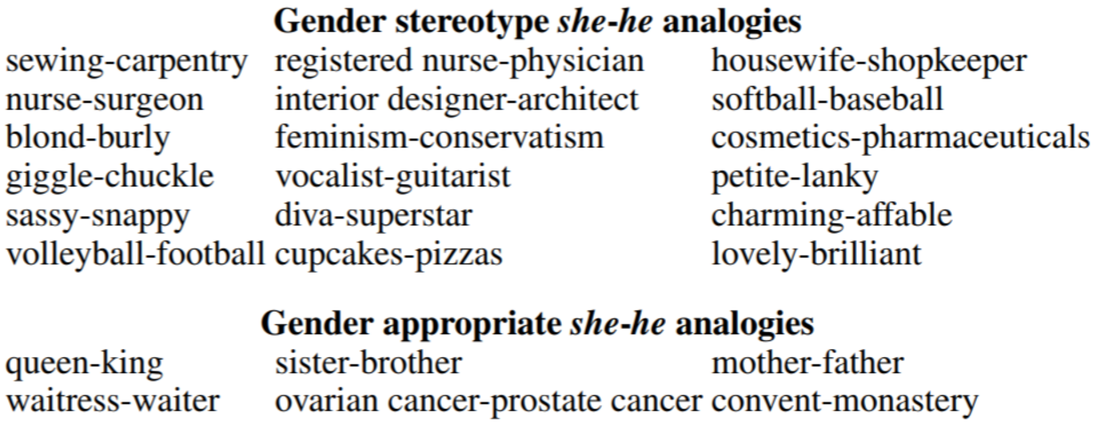
Buolamwini, J., Gebru, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Raji, I & Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. Conference on Artificial Intelligence, Ethics, and Society.

Responsibility – Word Embedding Bias

- Word embeddings are trained on large scale unlabeled text
  - For example: Wikipedia
- If training data is biased, vectors are biased as well

- Word2Vec trained on Wikipedia contains significant gender bias



Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." *Advances in neural information processing systems*. 2016.

Classroom Discussion

- What should we do as computer scientists?
  - Just measure existing data bias?
  - Pro-actively improving information systems?
- Do you have an example you encountered in your work?
- Who is responsible for fair machine learning / search engines?

See you next week :)