D3: A Massive Dataset of Scholarly Metadata

for Analyzing the State of Computer Science Research

Jan Philip Wahle, Terry Ruas, Saif M. Mohammad, Bela Gipp

University of Wuppertal
Germany

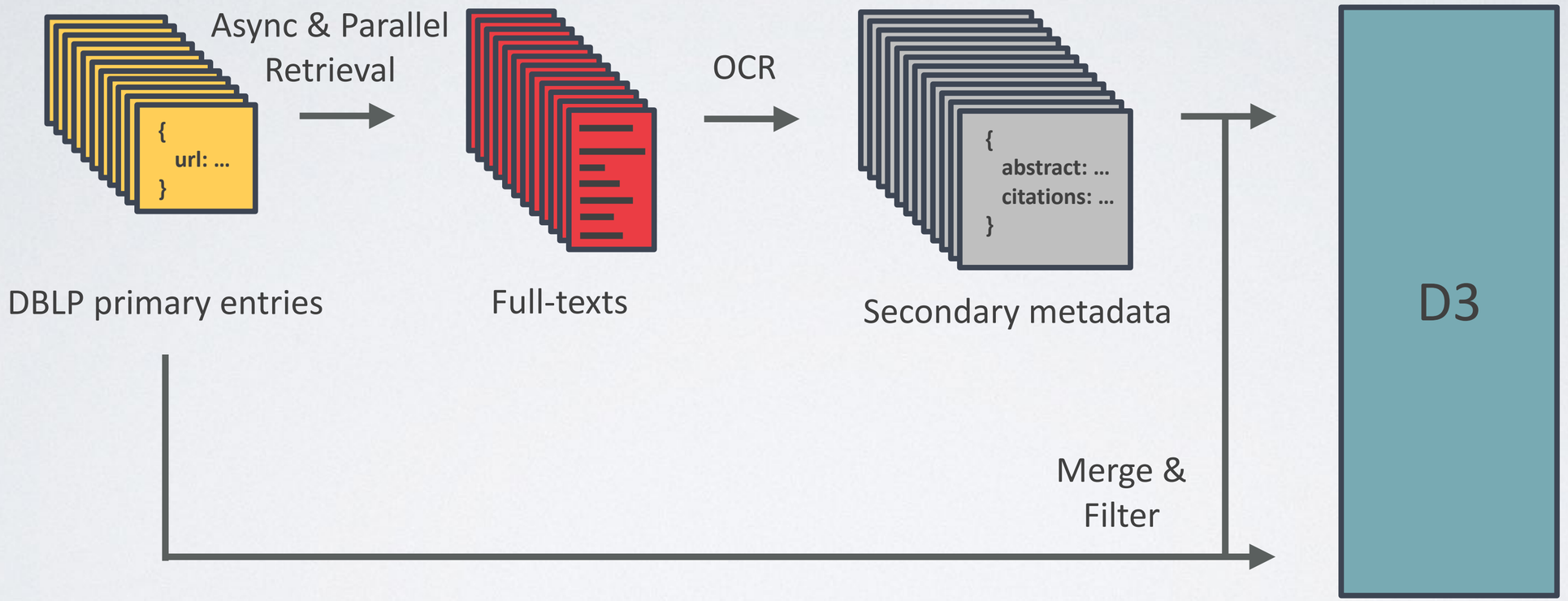
National Research Council
Canada

LREC 2022

MOTIVATION

- Computer Science (CS) is one of the most influential research fields
- DBLP is the largest repository of scientific articles focused on CS
- DBLP misses semantic information (e.g., abstracts, bibliographies) from full-texts to understand trends in impact, content, or bias of CS research
- The DBLP discovery dataset (D3) combines metadata from DBLP with sentient information from publications

DATASET COLLECTION



FEATURES OF D3

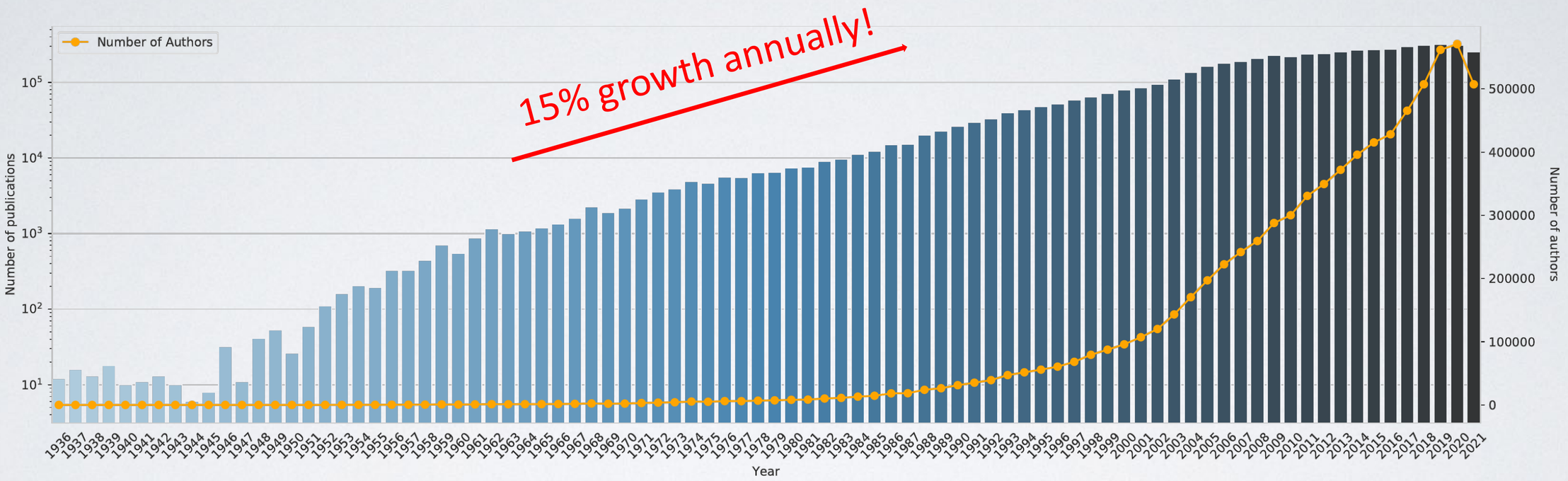
- Publications (title, abstract, access, doi, publisher, keywords, …)
- Authors (names, affiliation(s), webpage(s), …)
- Venues (acronyms, type, …)
- Citations (incoming, outgoing, …)

SUMMARY

- 6 million publications focused on CS journals and conferences
- 32 features: venues, citations, affiliations, ...
- Lightweight: 11GB compressed
- Abstracts and titles enable semantic analysis

ANALYSIS

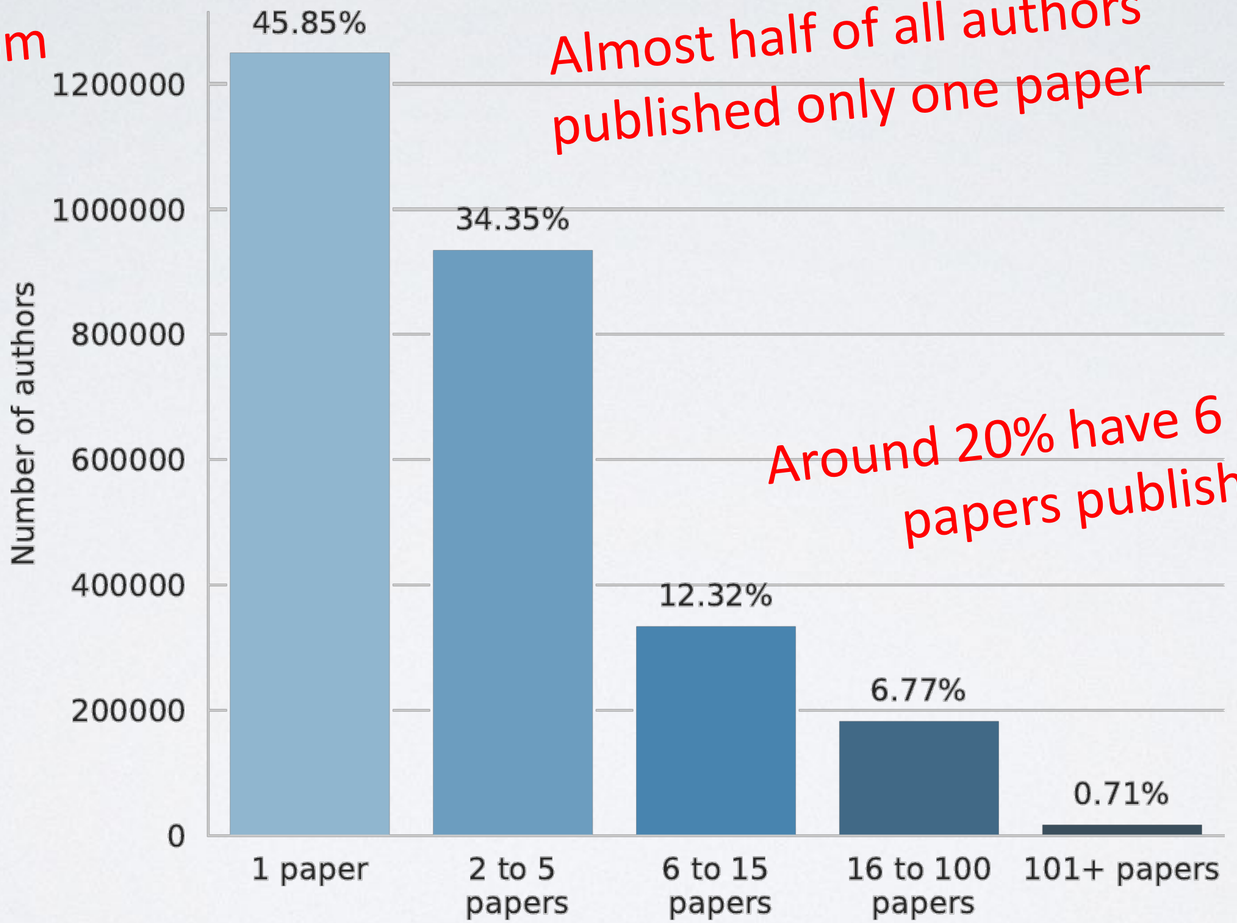
VOLUME



Wahle et al. "D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research" LREC 2022

RESEARCH ACTIVITY

Similar pattern compared to NLP publications from the ACL Anthology[1]



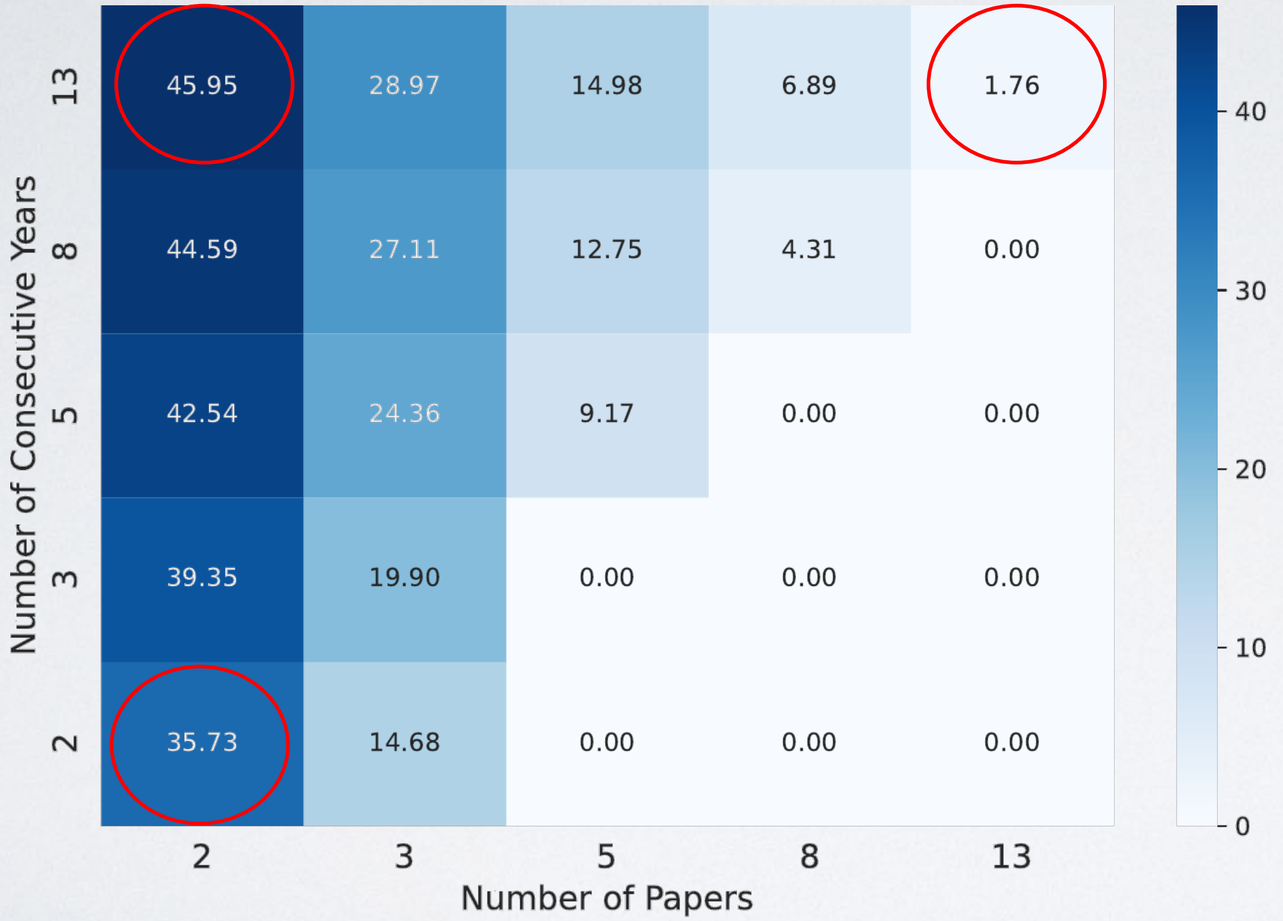
Almost half of all authors published only one paper

Around 20% have 6 or more papers published

[1]Saif M. Mohammad "NLP Scholar: A Dataset for Examining the State of NLP Research" LREC 2020

Wahle et al. "D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research" LREC 2022

RESEARCH ACTIVITY



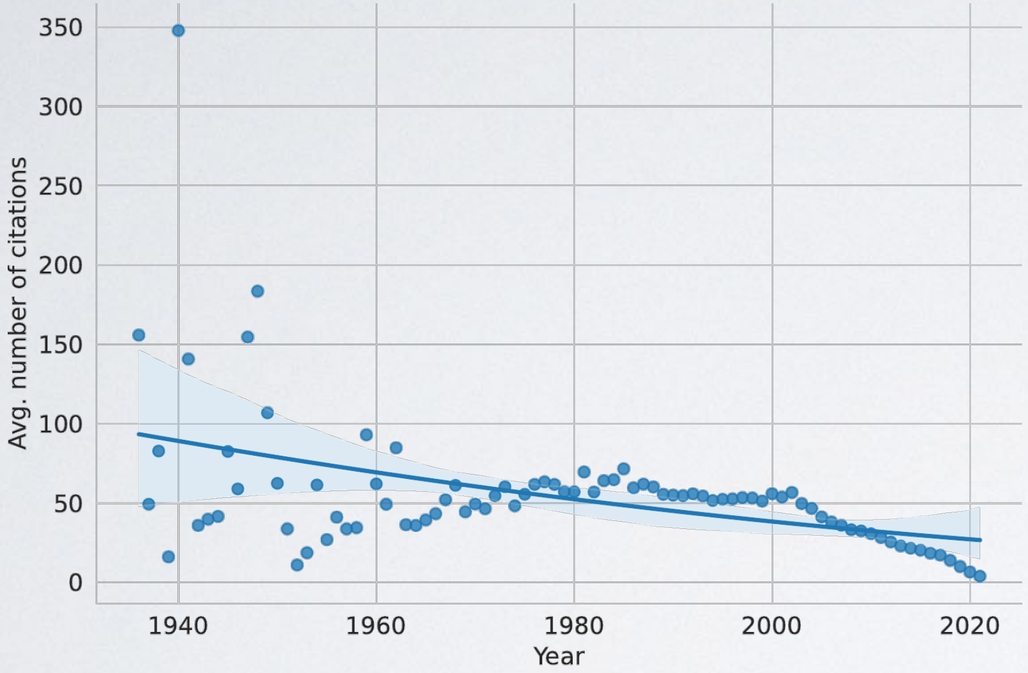
Persistence and output are indicators for research activity.

Most researchers publish few papers over a short time window.

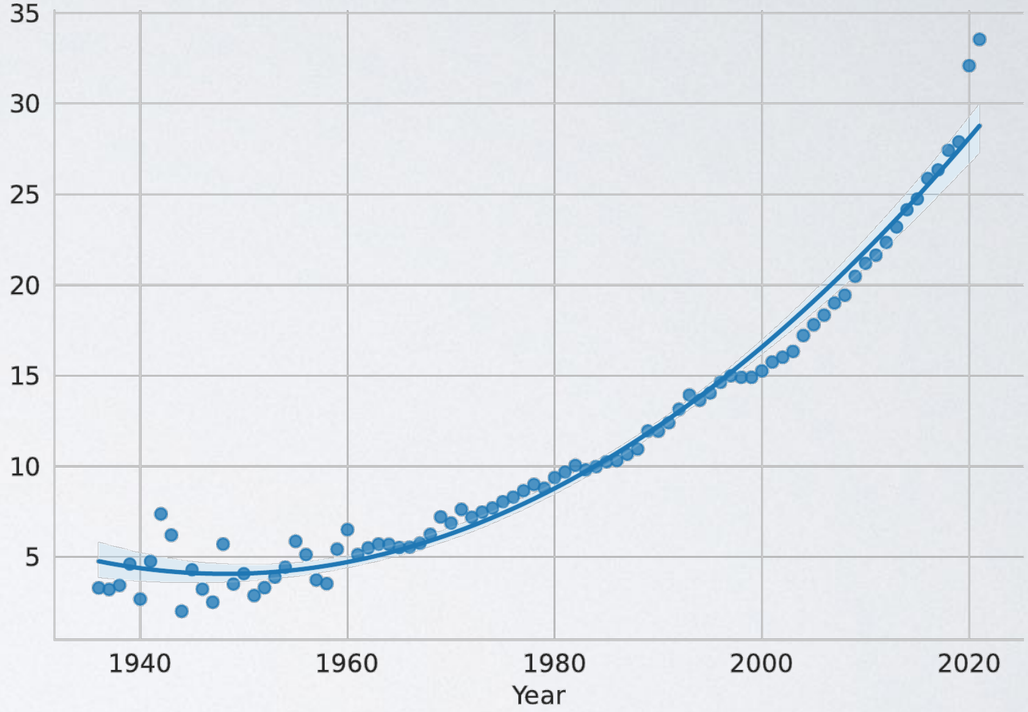
Wahle et al. "D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research" LREC 2022

TOPIC TRENDS

CITATIONS



Avg. incoming citations

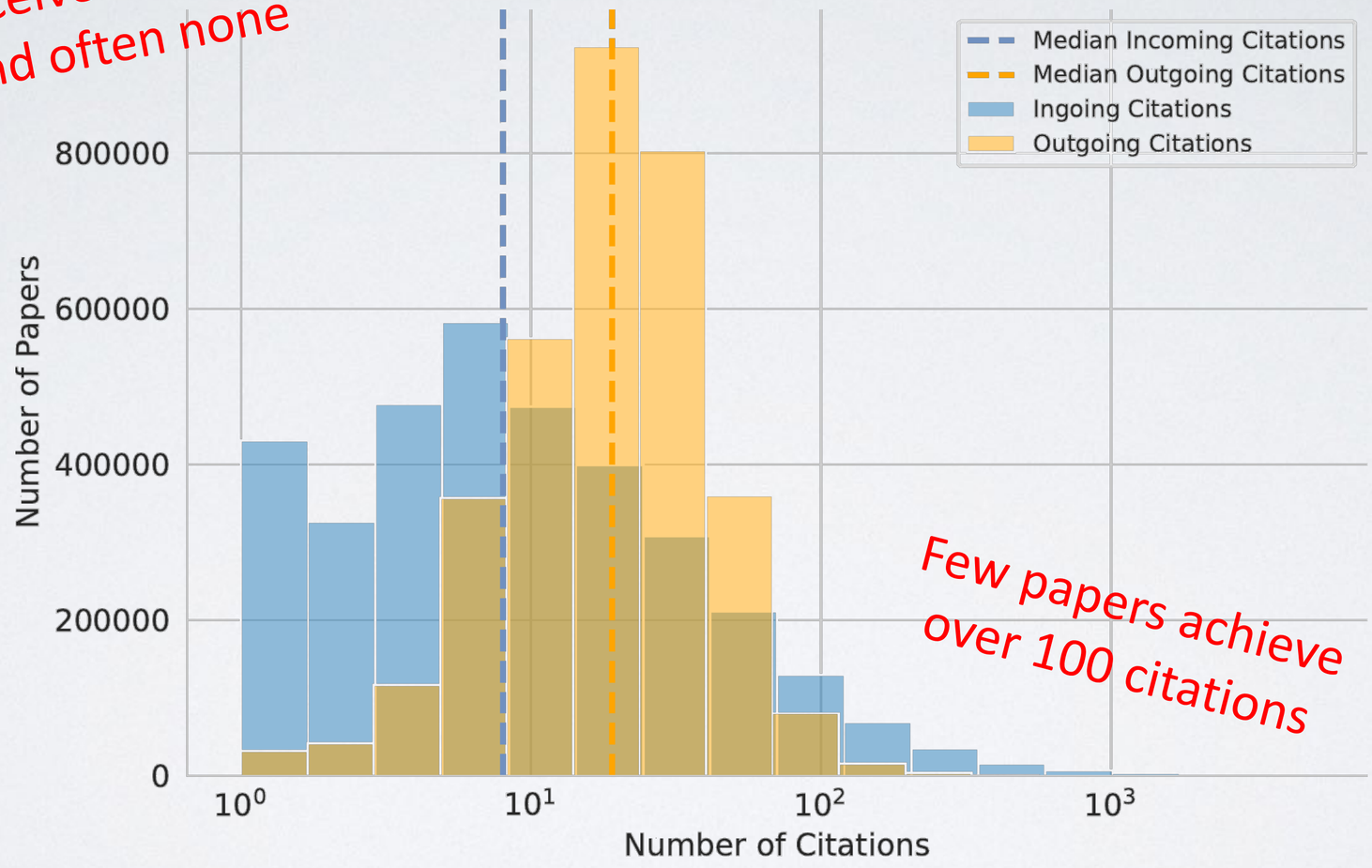


Avg. outgoing citations

Wahle et al. "D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research" LREC 2022

CITATIONS

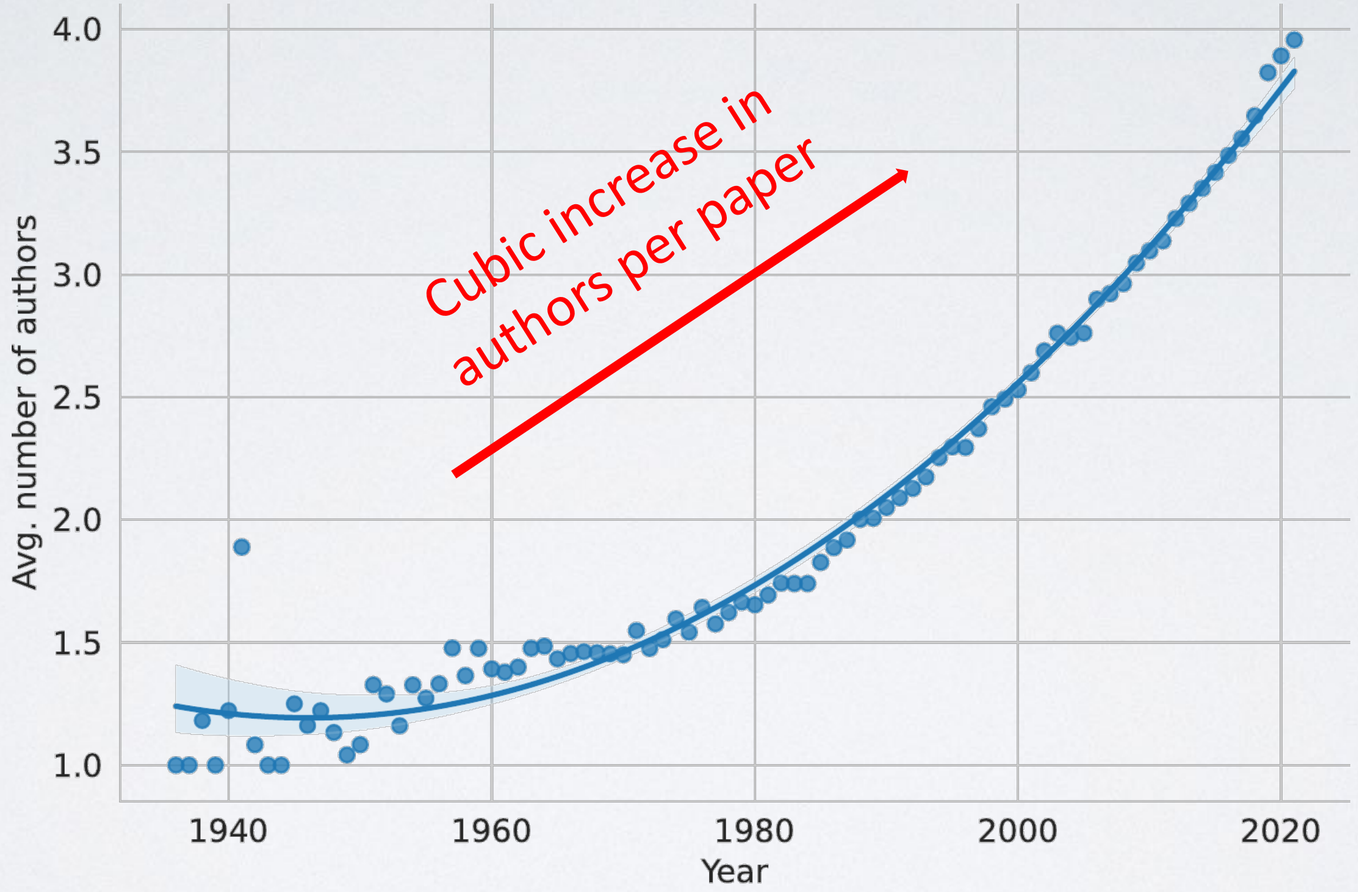
Most papers receive less than 10 citations and often none



Few papers achieve over 100 citations

Wahle et al. "D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research" LREC 2022

COLLABORATION

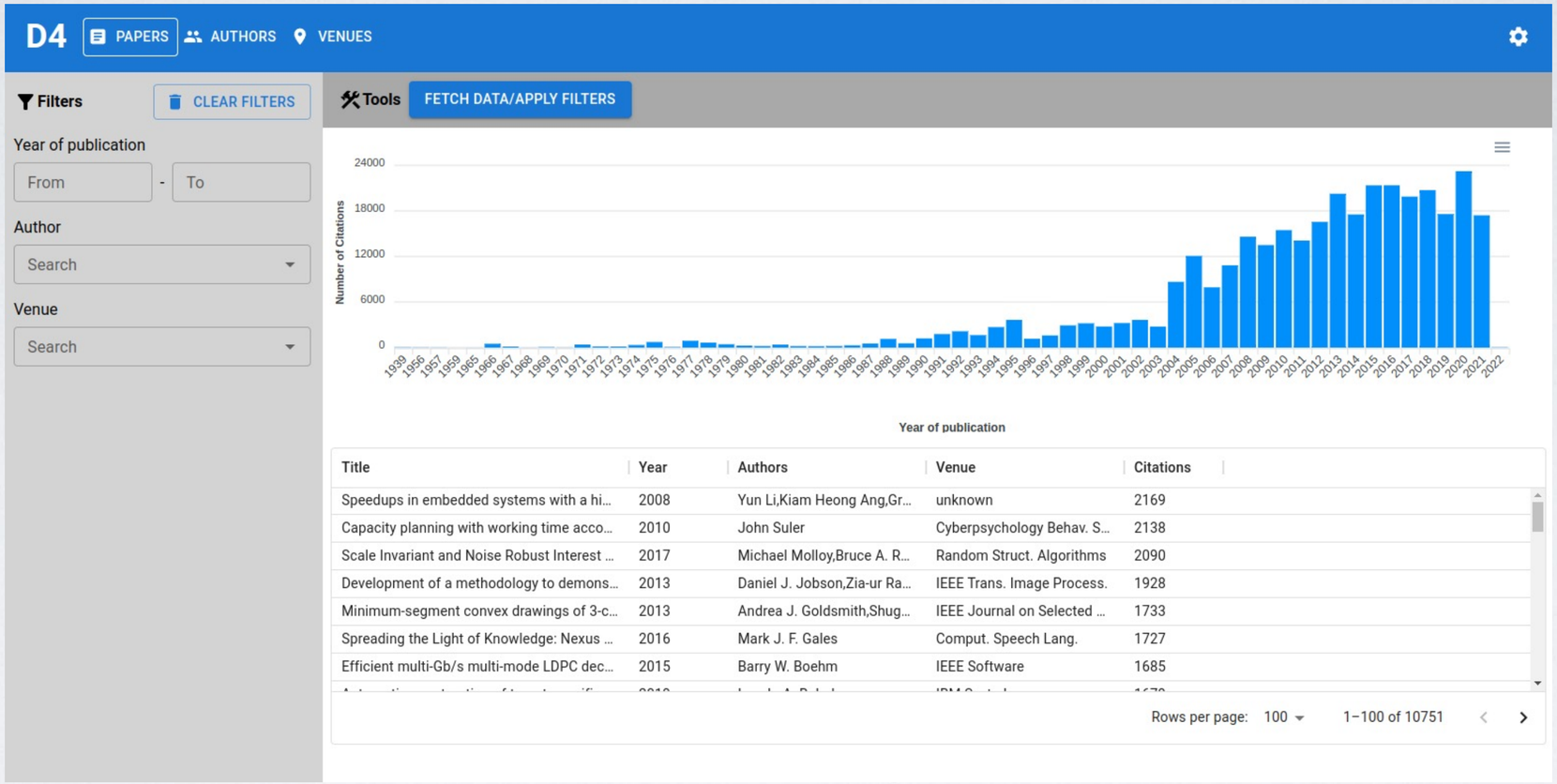


Wahle et al. "D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research" LREC 2022

FUTURE WORK

- Open and interactive webapp to explore and analyze CS papers
- Topic analysis to understand influence of research fields in CS
- Analyzing impact, success, and productivity
- Gender gap and fairness

SNEAK PEEK



Interactive webapp to explore and analyze CS papers

EXPLORE D3

https://github.com/ag-gipp/NLP-Land-backend

https://zenodo.org/record/6477785

First release!

wahle@uni-wuppertal.de