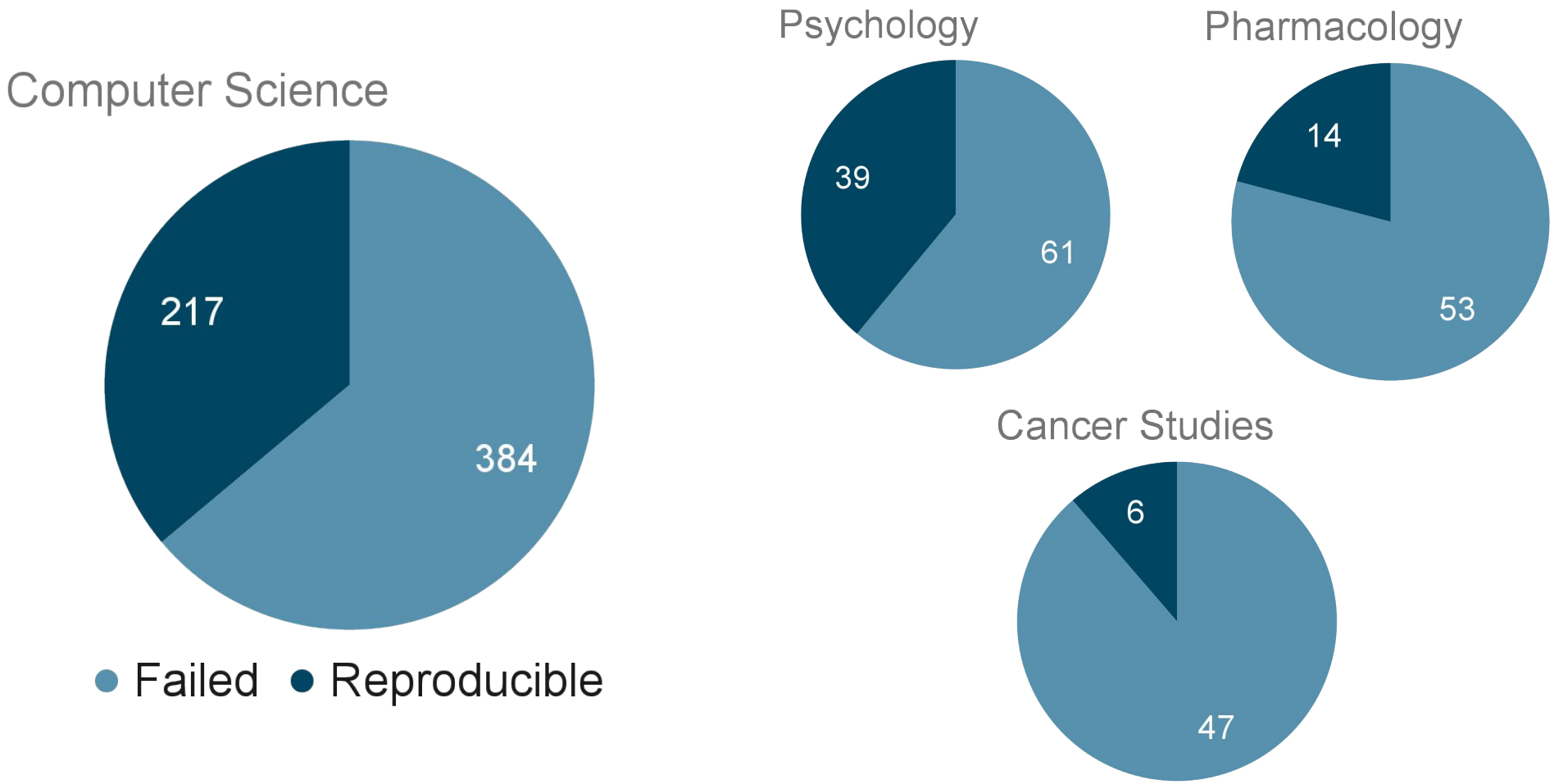
Reproduzierbarkeit(skrise) in den Daten-Wissenschaften

Forschungssymposium des Profilbereichs „Digitale Arbeit und Lebenswelten“
5. Mai 2023

Timo Breuer
Institut für Informationsmanagement,
Fakultät für Informations- und Kommunikationswissenschaften

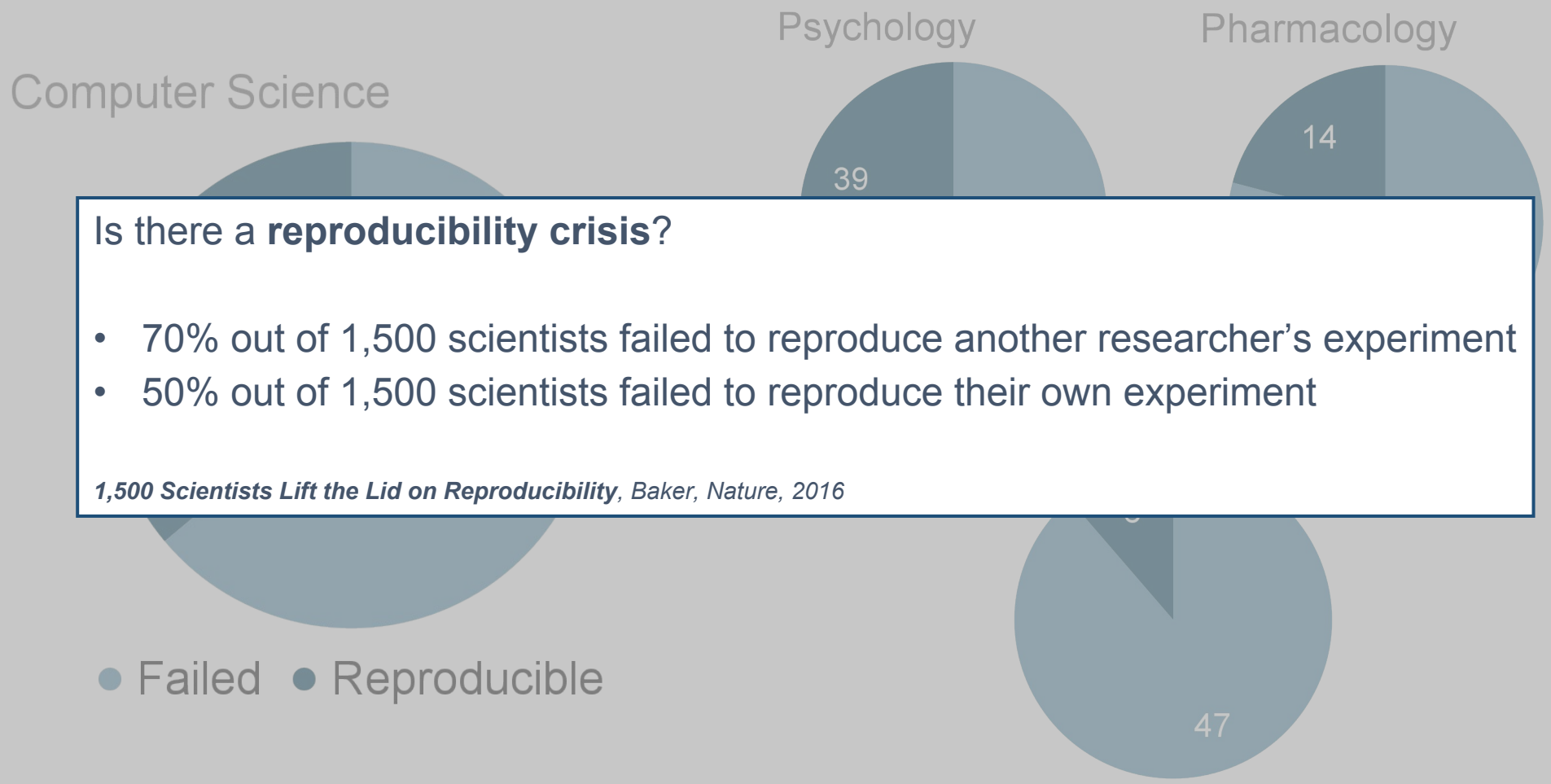
IR GROUP
Technology
Arts Sciences
TH Köln

Reproducibility of Published Research Articles



***Repeatability in Computer Systems Research**, Collberg and Proebsting; Commun. ACM; 2015*
***Estimating the Reproducibility of Psychological Science**, Open Science Collaboration; Science; 2015*
***Believe it or not: How Much can we Rely on Published Data on Potential Drug Targets?**; Prinz et al.; Nature Reviews on Drug Discovery, 2011*
***Raise Standards for Preclinical Cancer Research**, Begley and Ellis; Nature; 2012*

Reproducibility of Published Research Articles



Is there a **reproducibility crisis**?

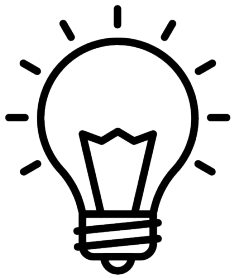
- 70% out of 1,500 scientists failed to reproduce another researcher's experiment
- 50% out of 1,500 scientists failed to reproduce their own experiment

***1,500 Scientists Lift the Lid on Reproducibility**, Baker, Nature, 2016*

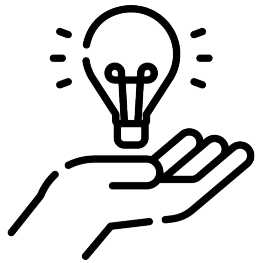
***Repeatability in Computer Systems Research**, Collberg and Proebsting; Commun. ACM; 2015*
***Estimating the Reproducibility of Psychological Science**, Open Science Collaboration; Science; 2015*
***Believe it or not: How Much can we Rely on Published Data on Potential Drug Targets?**; Prinz et al.; Nature Reviews on Drug Discovery, 2011*
***Raise Standards for Preclinical Cancer Research**, Begley and Ellis; Nature; 2012*

Overview

Common Issues

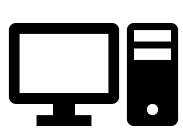


Countermeasures

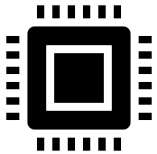


Our Contributions

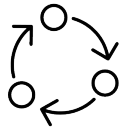
Common Issues and Countermeasures



Platform



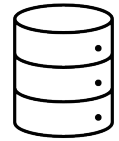
Implementation



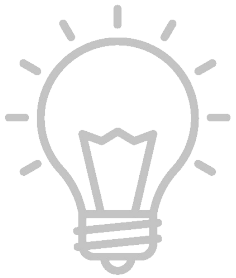
Method

Research goal

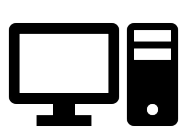
Actor



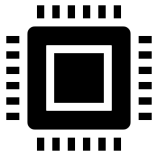
Data



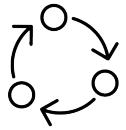
Common Issues and Countermeasures



Platform



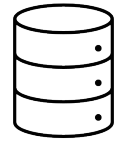
Implementation



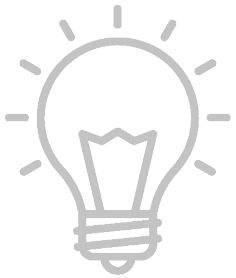
Method

Research goal

Actor

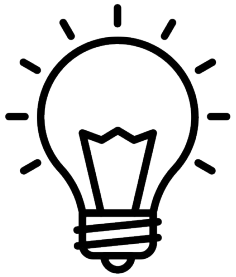


Data

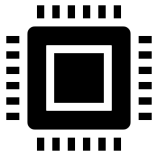


- Missing hardware, kernel, or operating system
- No source code, dependency management

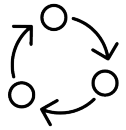
Common Issues and Countermeasures



Platform



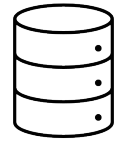
Implementation



Method

Research goal

Actor

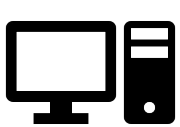


Data

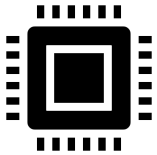
- Missing hardware, kernel, or operating system
- No source code, dependency management

- Virtualization or containerization
- Open-source code

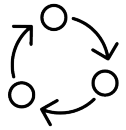
Common Issues and Countermeasures



Platform



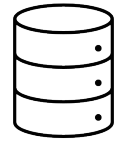
Implementation



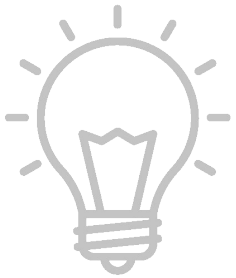
Method

Research goal

Actor



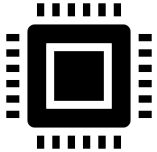
Data



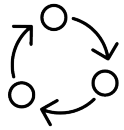
- Missing statistical evidence / low statistical power
- Leaderboard chasing (data dregding / p-hacking / fraud)
- Cognitive biases

Common Issues and Countermeasures

Platform



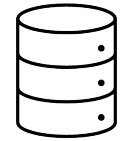
Implementation



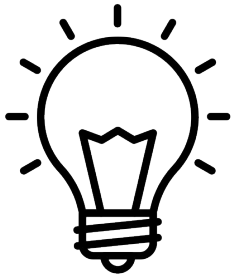
Method

Research goal

Actor



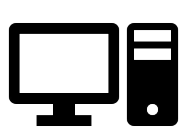
Data



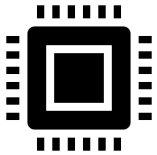
- Missing statistical evidence / low statistical power
- Leaderboard chasing (data dregding / p-hacking / fraud)
- Cognitive biases

- Research standards and guidelines, journal policies
- Common task framework / shared tasks
- Open reviews

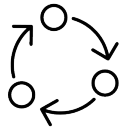
Common Issues and Countermeasures



Platform



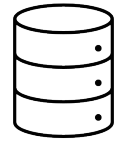
Implementation



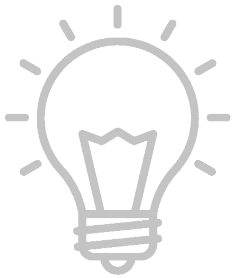
Method

Research goal

Actor

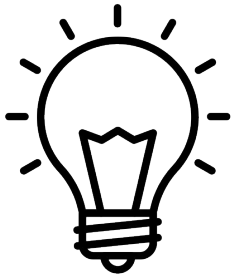


Data

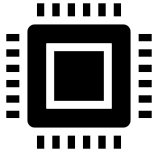


- Separation between code and data
- Private / closed / paywalled datasets
- Leakage
- Data bias

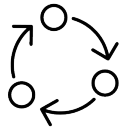
Common Issues and Countermeasures



Platform



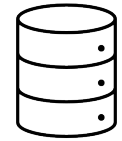
Implementation



Method

Research goal

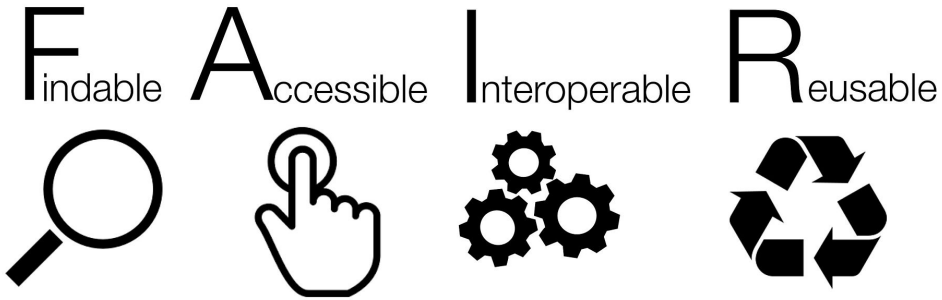
Actor



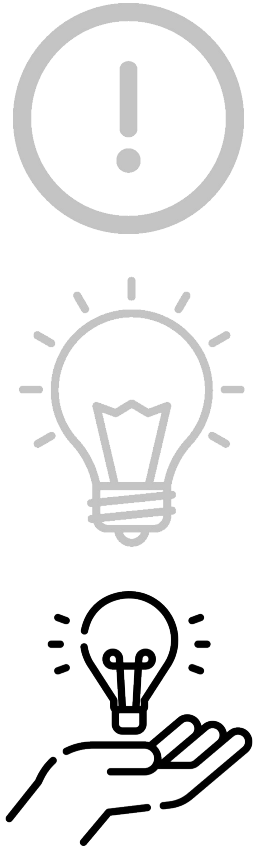
Data

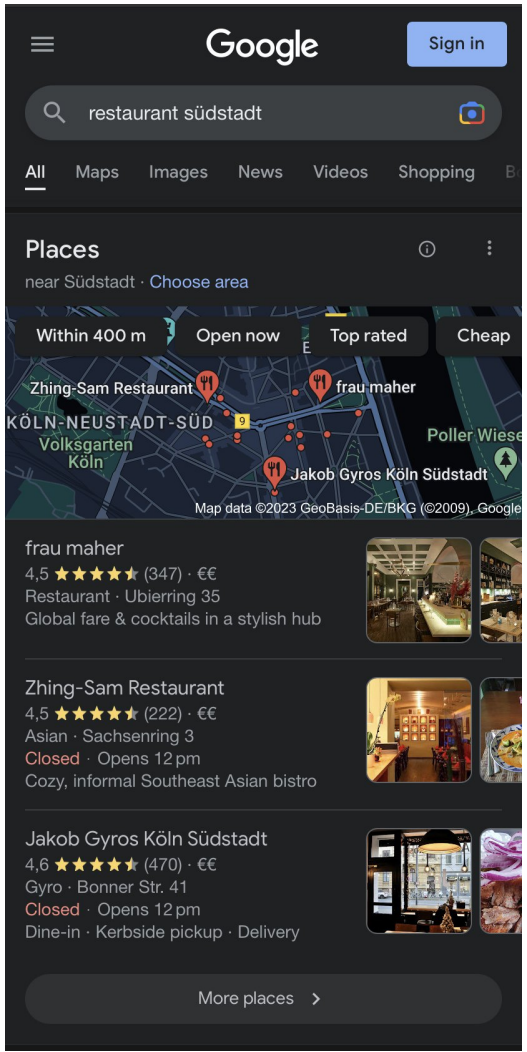
- Separation between code and data
- Private / closed / paywalled datasets
- Leakage
- Data bias

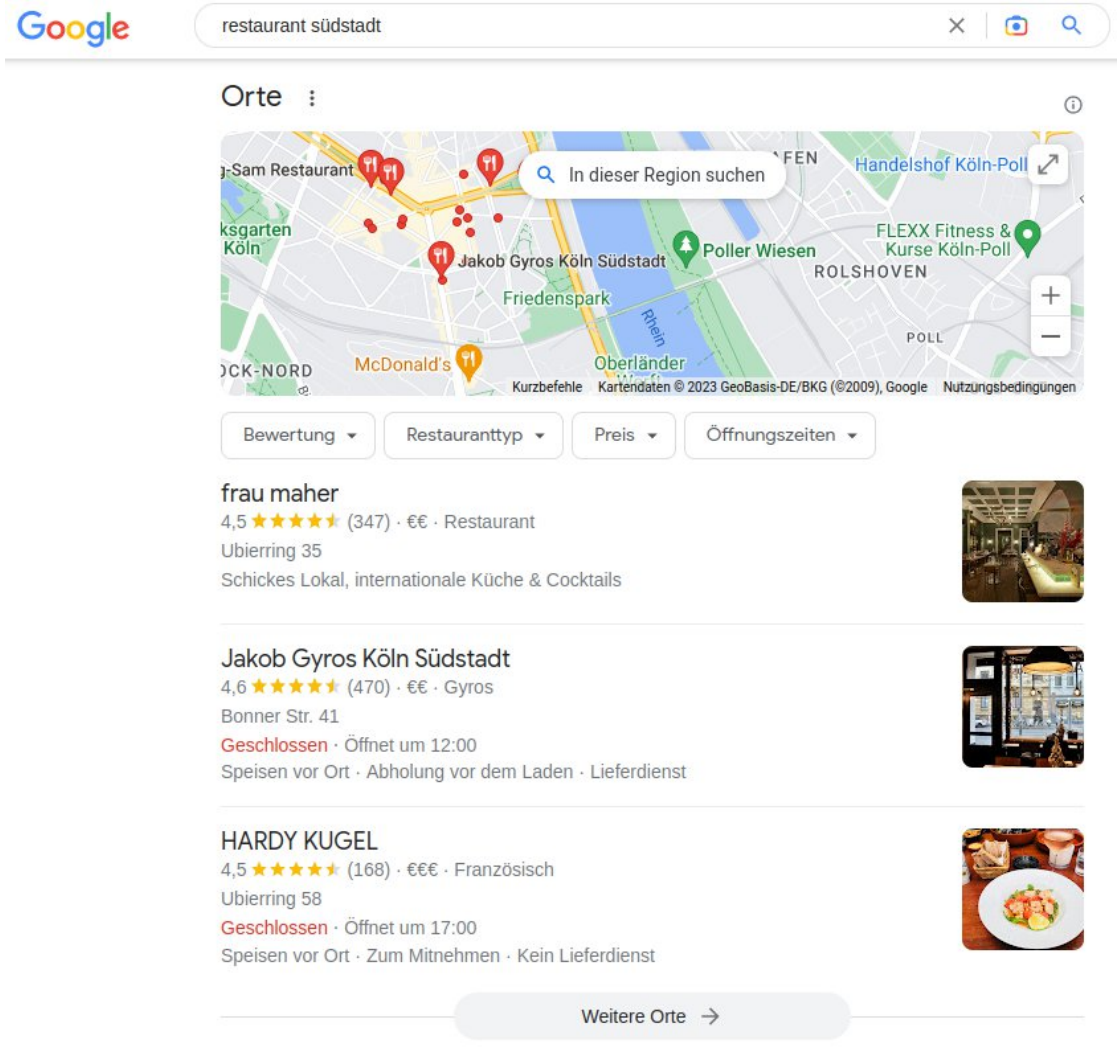
- Open data
- Data archival
- Data citation principles
- Datasheets



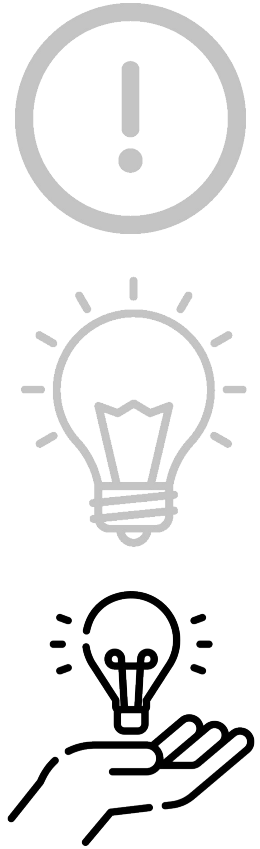
What About Information Retrieval?







What About Information Retrieval?

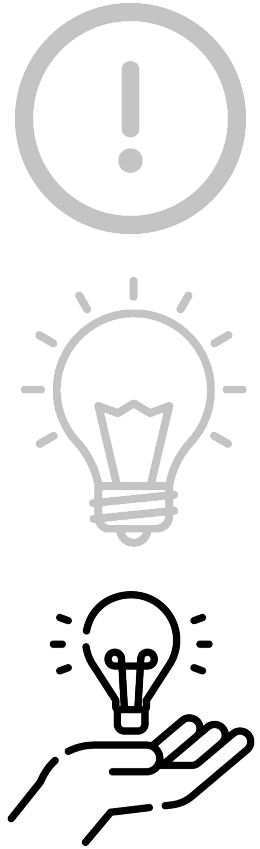


	System A			System B			System C		
	q1	q2	q3	q1	q2	q3	q1	q2	q3
	1	11	21	25	11	21	1	220	21
	2	12	22	23	12	326	402	219	326
	3	13	23	1	13	23	3	218	23
	4	14	24	24	14	24	404	217	524
	5	15	25	105	115	327	405	16	625
	6	16	26	106	116	329	406	15	426
	7	17	27	107	17	27	497	14	627
	8	18	28	108	18	28	438	13	728
	9	19	29	9	19	29	429	12	929
	10	20	30	10	20	30	410	11	430
P@10	0.4	0.6	0.5	0.5	0.4	0.6	0.2	1.0	0.3
Avg. P@10	0.5000			0.5000			0.5000		

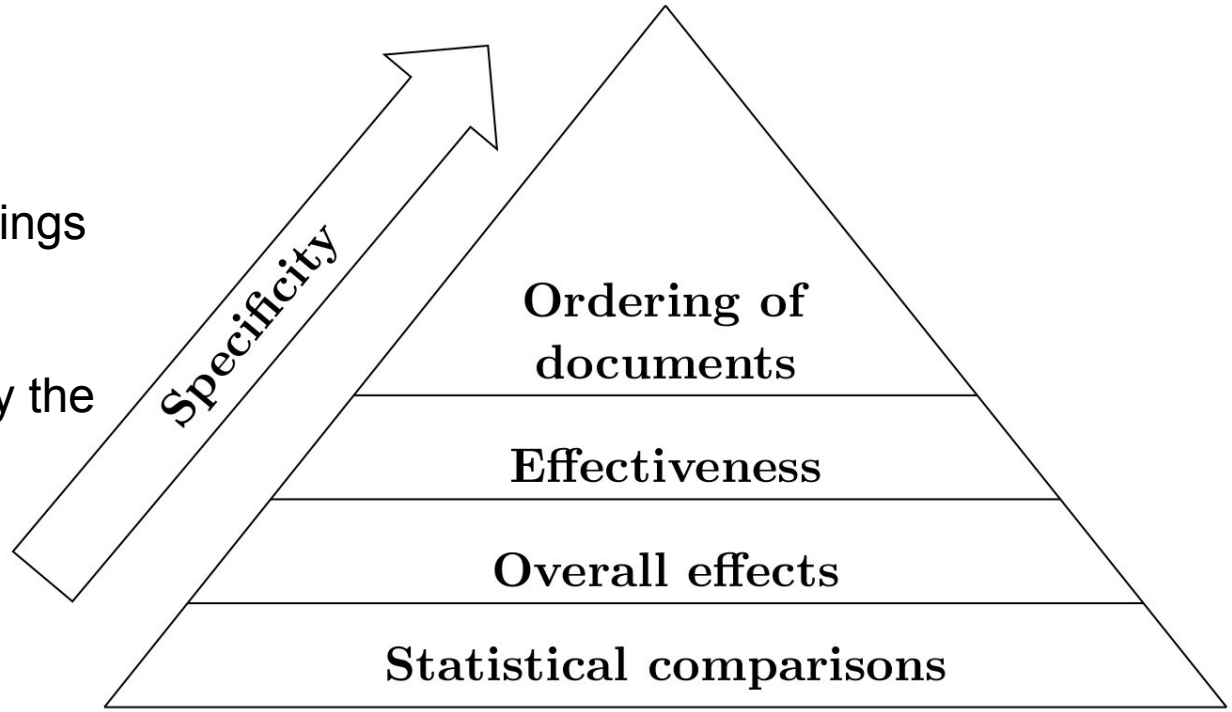
An in-depth investigation on the behavior of measures to quantify reproducibility; *Maistro, Breuer, Schaer, Ferro; IPM 2023*

Our Contributions

Reproducibility measures

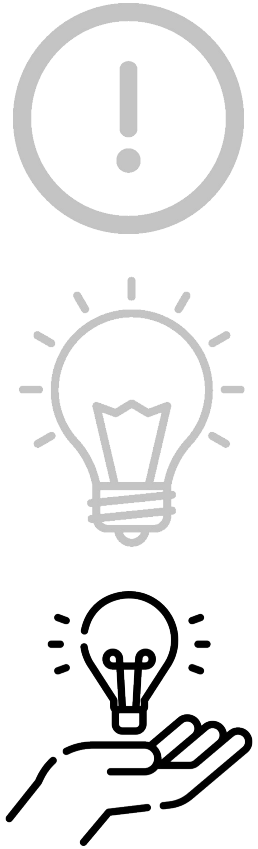


- Analysis of **reimplementations**
- **Comparison** of reimplemented rankings to the original results
- **Framework of measures** to quantify the degree of reproducibility
- Different **levels of fidelity**



***How to Measure the Reproducibility of System-Oriented IR Experiments**; Breuer, Ferro, Fuhr, Maistro, Sakai, Schaer, Soboroff; SIGIR 2020*
***repro_eval: A Python Interface to Reproducibility Measures of System-Oriented IR Experiments**; Breuer, Ferro, Maistro, Schaer; ECIR 2021*

Our Contributions

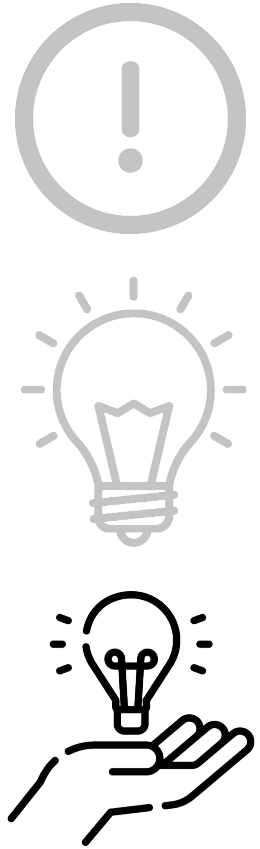


	System A			System B			System C		
	q1	q2	q3	q1	q2	q3	q1	q2	q3
	1	11	21	25	11	21	1	220	21
	2	12	22	23	12	326	402	219	326
	3	13	23	1	13	23	3	218	23
	4	14	24	24	14	24	404	217	524
	5	15	25	105	115	327	405	16	625
	6	16	26	106	116	329	406	15	426
	7	17	27	107	17	27	497	14	627
	8	18	28	108	18	28	438	13	728
	9	19	29	9	19	29	429	12	929
	10	20	30	10	20	30	410	11	430
P@10	**0.4**	**0.6**	**0.5**	**0.5**	**0.4**	**0.6**	**0.2**	**1.0**	**0.3**
Avg. P@10	**0.5000**			**0.5000**			**0.5000**		
RMSE	**0.0000**			**0.1414**			**0.2828**		
RBO	**1.0000**			**0.5663**			**0.3984**		
KTU	**1.000**			**0.4370**			**0.0815**		

An in-depth investigation on the behavior of measures to quantify reproducibility; *Maistro, Breuer, Schaer, Ferro; IPM 2023*

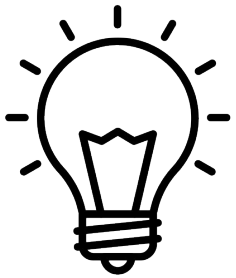
Our Contributions

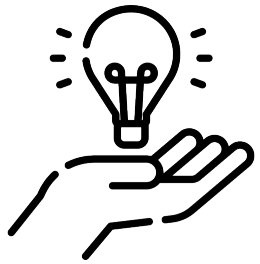
Metadata scheme



ir_metadata: An Extensible Metadata Schema for IR Experiments; Breuer, Keller, Schaer; SIGIR 2022

Wrap Up





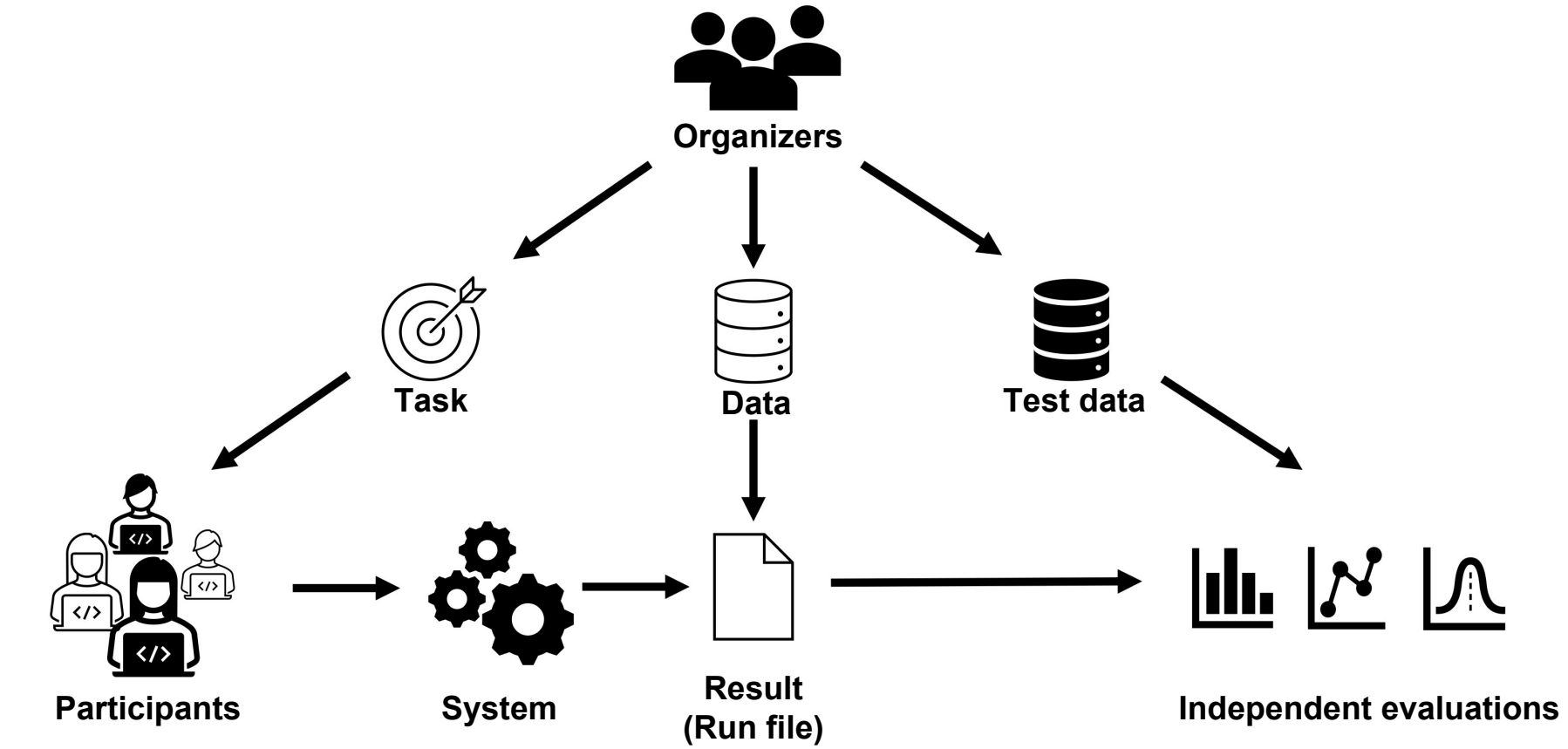
- **Reproducibility issues in data science are manifold!**
- **Countermeasures** and solutions are **diverse**
  - Technological, organizational, social, ethical
- Our **contributions** to reproducible information retrieval
  - **Reactive support** by reproducibility measures
  - **Proactive support** by metadata scheme

Wrap Up

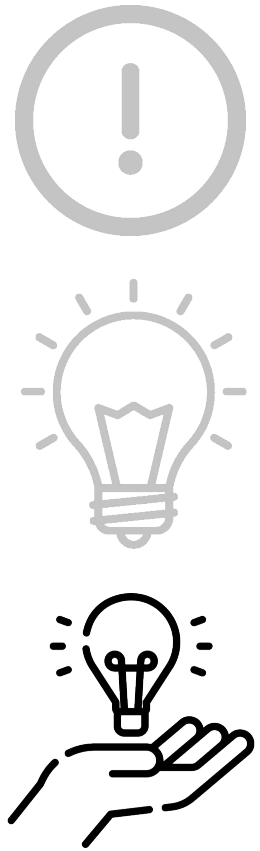
- **Re**
- **Co**
  -
- Ou retrieval
  -
  -

Thank you!

Common task framework



Common task framework



IR is deeply rooted in experimentation

- Common evaluation protocols (evaluation measures of relevance)
- Common test collections (reusable data)
- Common data outputs (run files as de-facto standard)

IR conferences host dedicated reproducibility tracks

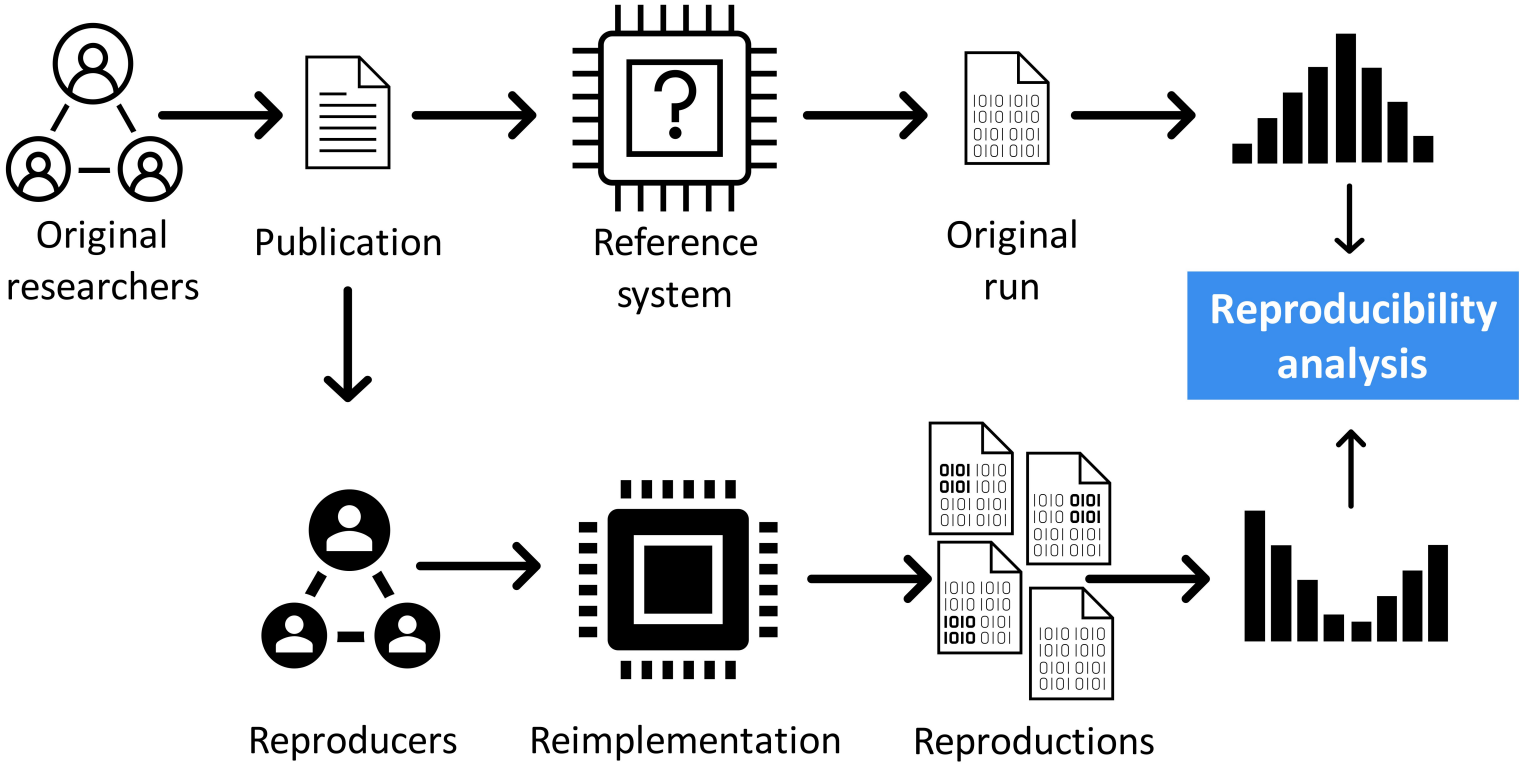
- ECIR since 2015
- SIGIR since 2022

IR community engages in several reproducibility initiatives

- Workshops (OSIRRC@SIGIR, CENTRE@CLEF)
- ACM SIGIR Artifact Badging
  https://sigir.org/general-information/acm-sigir-artifact-badging/

Our Contributions

Reproducibility analysis



Our Contributions

Metadata scheme

Supporting reproducibility by metadata annotations

- **PRIMAD** is the logical plan
- Annotation of **TREC** run files
- **YAML** formatted header as comment
- Focus on **extensibility**
- **Public resource** hosted on https://www.ir-metadata.org/