Reproducible Information Retrieval Research:
From Principled System-oriented Evaluations Towards User-oriented Experimentation

Disputation
Friday, March 31st, 2023

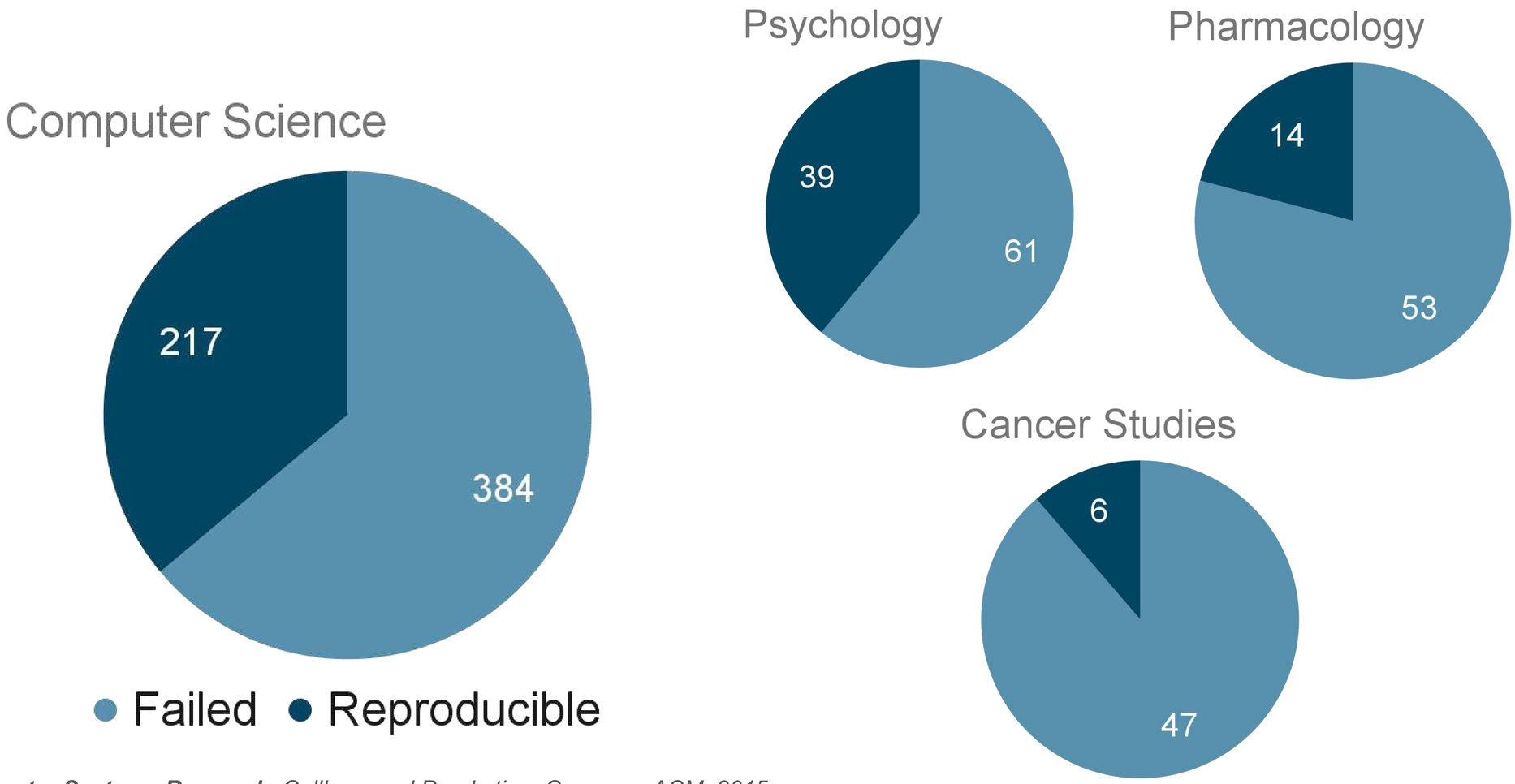
Timo Breuer

UNIVERSITÄT DUISBURG ESSEN

Technology Arts Sciences TH Köln

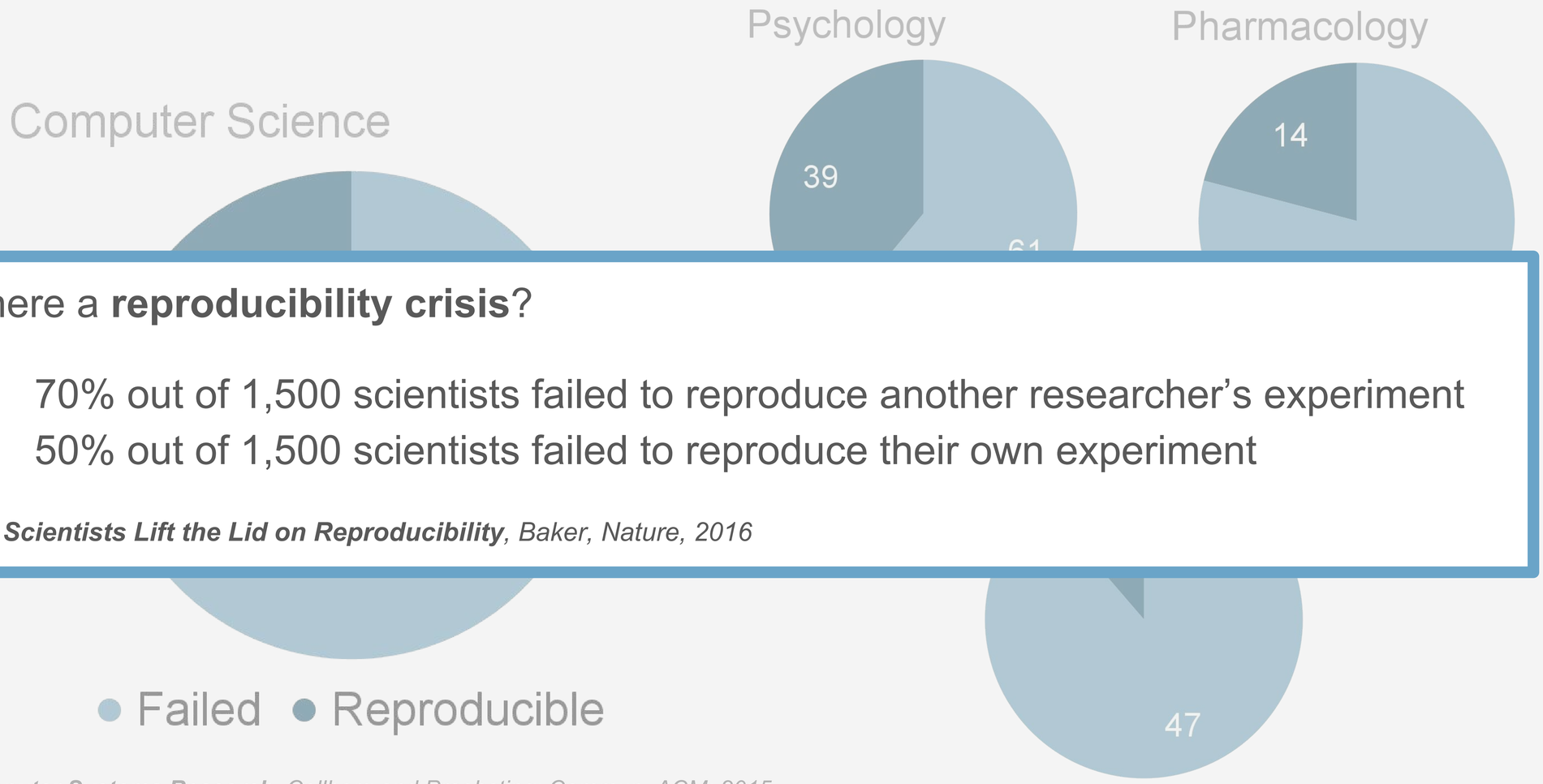
IR GROUP

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, Schlange, Asadullah; 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

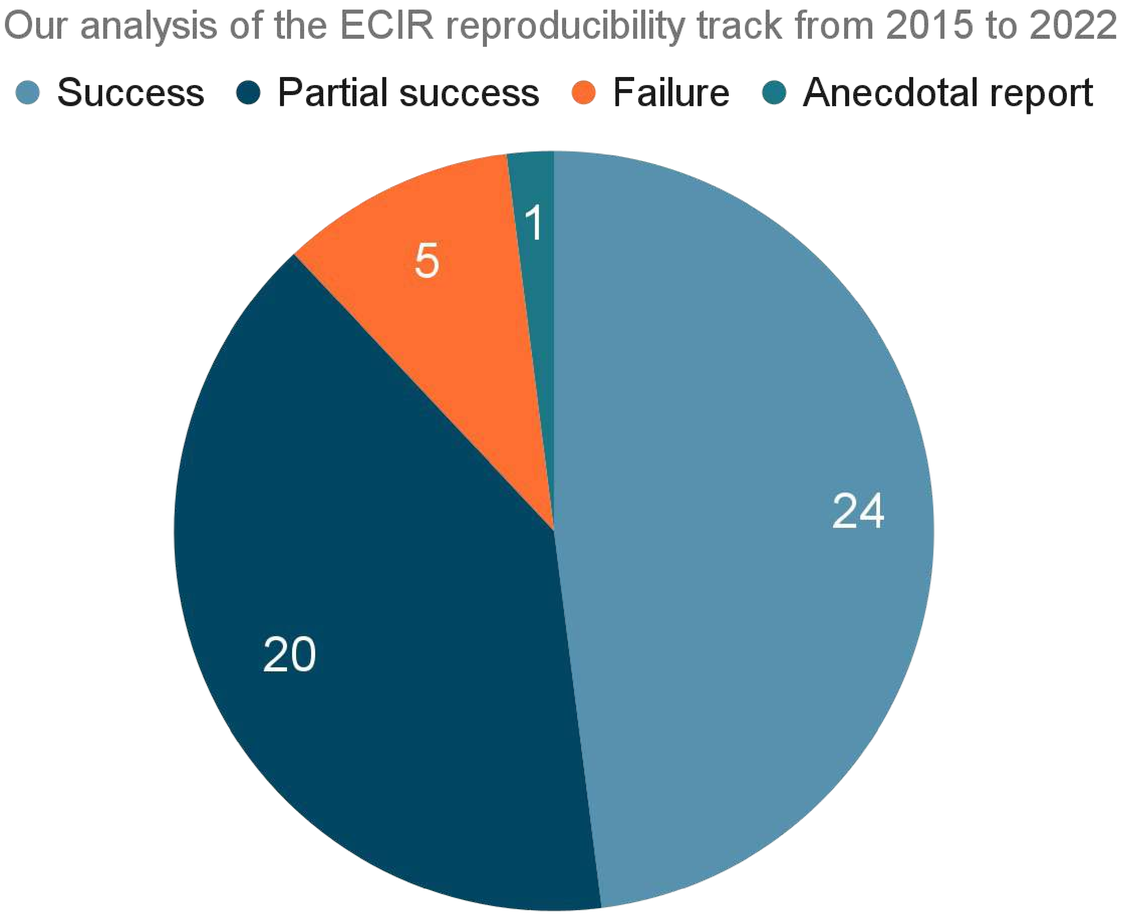
***1,500 Scientists Lift the Lid on Reproducibility**, Baker, Nature, 2016*

***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, Schlange, Asadullah; Nature Reviews on Drug Discovery, 2011*
***Raise Standards for Preclinical Cancer Research**, Begley and Ellis; Nature; 2012*

Reproducible Information Retrieval?

Open points:

- Inconsistent use of terminology
- Lack of evaluation standards
- User-oriented evaluations are underrepresented

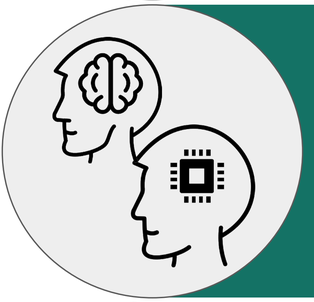


Outline and Contributions

INTERNAL VALIDITY

system-oriented experiments

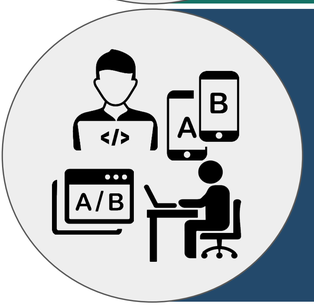
- PRIMAD extensions and metadata scheme
- Principled reproducibility evaluations



EXTERNAL VALIDITY

user simulations

- Query simulations and evaluation framework
- Click-based evaluations of system rankings



ECOLOGICAL VALIDITY

real user experiments

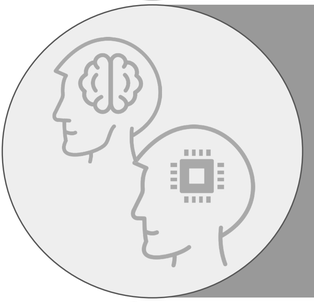
- Living lab infrastructure
- Shared task evaluations

Outline and Contributions

INTERNAL VALIDITY

system-oriented experiments

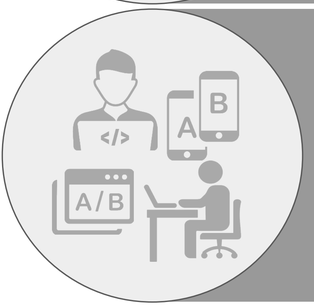
- PRIMAD extensions and metadata scheme
- Principled reproducibility evaluations



EXTERNAL VALIDITY

user simulations

- Query simulations and evaluation framework
- Click-based evaluations of system rankings

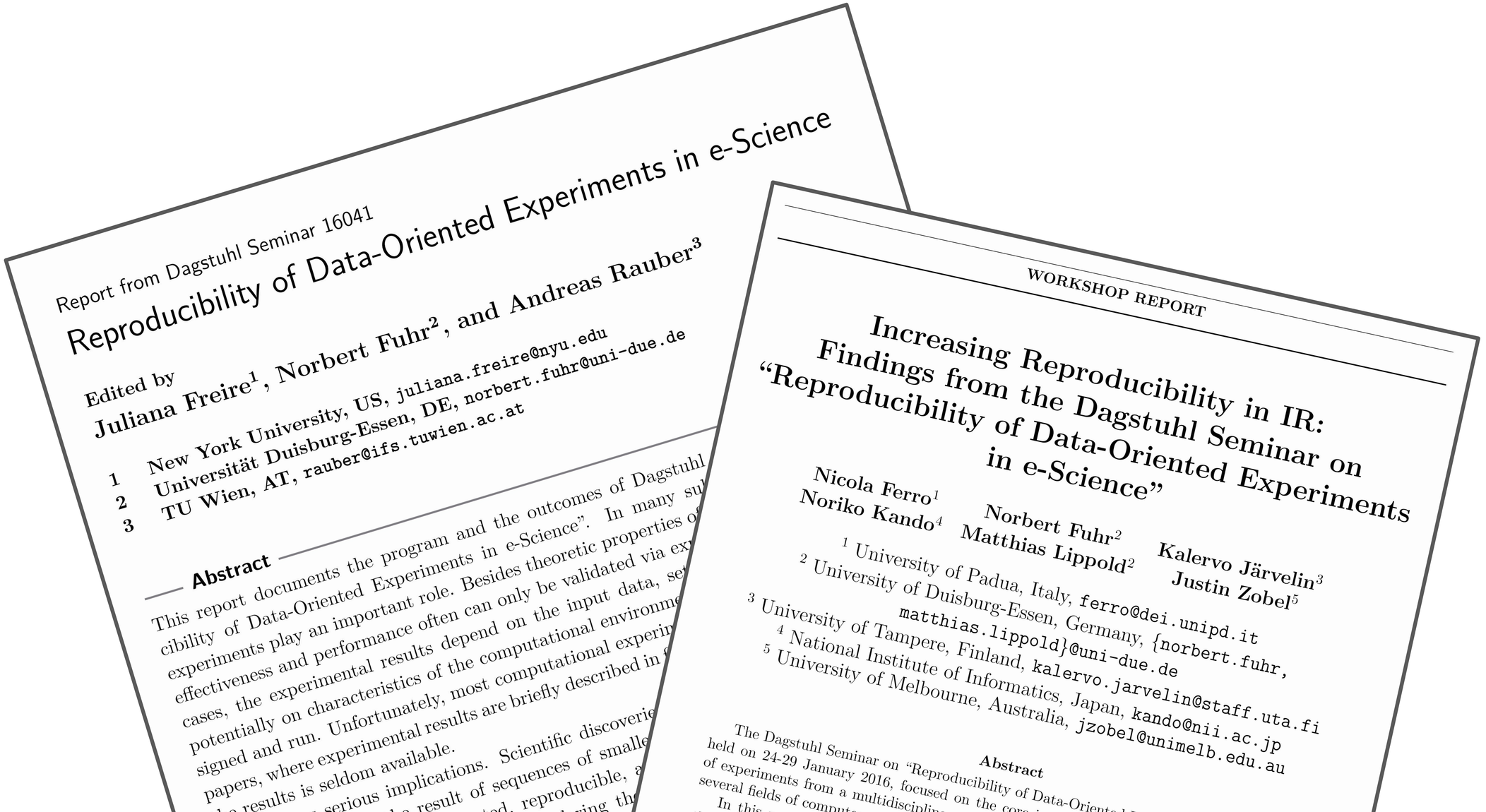


ECOLOGICAL VALIDITY

real user experiments

- Living lab infrastructure
- Shared task evaluations

PRIMAD - A Taxonomy for Reproducible IR Research



Report from Dagstuhl Seminar 16041

Reproducibility of Data-Oriented Experiments in e-Science

Edited by

Juliana Freire[1], Norbert Fuhr[2], and Andreas Rauber[3]

1 New York University, US, juliana.freire@nyu.edu
2 Universität Duisburg-Essen, DE, norbert.fuhr@uni-due.de
3 TU Wien, AT, rauber@ifs.tuwien.ac.at

Abstract

This report documents the program and the outcomes of Dagstuhl
cibility of Data-Oriented Experiments in e-Science". In many su
experiments play an important role. Besides theoretic properties o
effectiveness and performance often can only be validated via ex
cases, the experimental results depend on the input data, se
potentially on characteristics of the computational environme
signed and run. Unfortunately, most computational experim
papers, where experimental results are briefly described in
results is seldom available.
serious implications. Scientific discoveri
result of sequences of smalle
ed, reproducible, a
ring th

WORKSHOP REPORT

Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science"

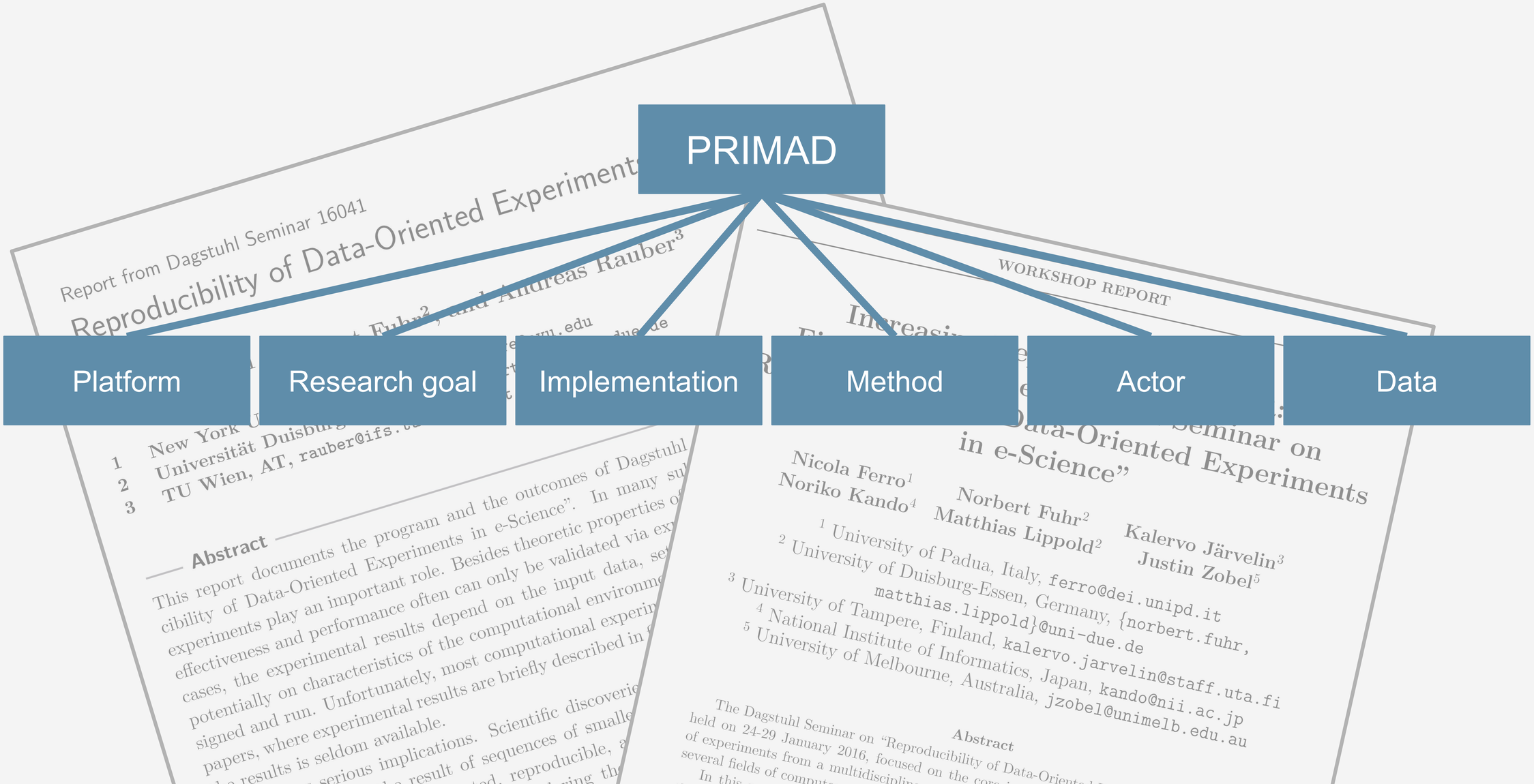
Nicola Ferro[1] Norbert Fuhr[2] Kalervo Järvelin[3]
Noriko Kando[4] Matthias Lippold[2] Justin Zobel[5]

[1] University of Padua, Italy, ferro@dei.unipd.it
[2] University of Duisburg-Essen, Germany, {norbert.fuhr, matthias.lippold}@uni-due.de
[3] University of Tampere, Finland, kalervo.jarvelin@staff.uta.fi
[4] National Institute of Informatics, Japan, kando@nii.ac.jp
[5] University of Melbourne, Australia, jzobel@unimelb.edu.au

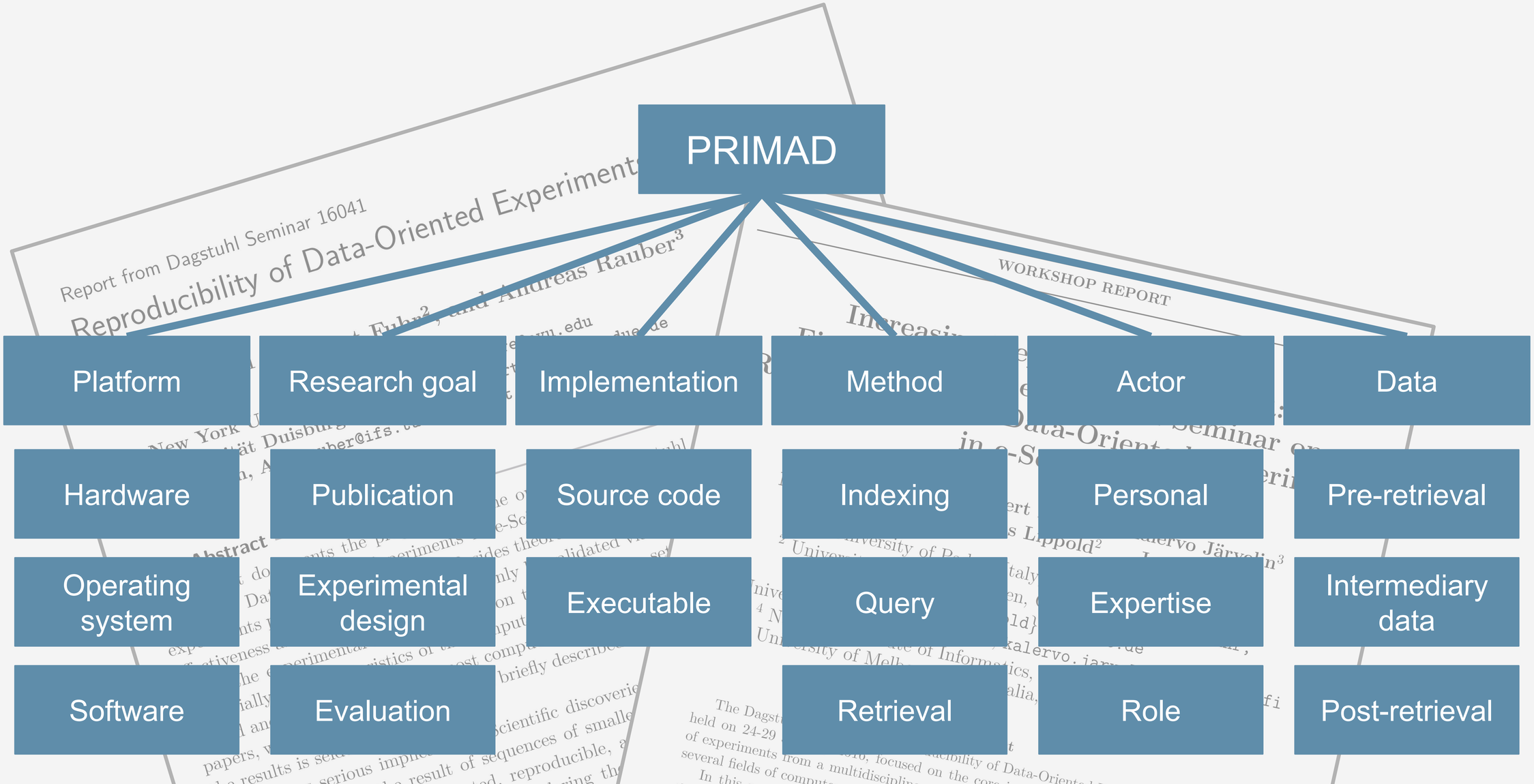
Abstract

The Dagstuhl Seminar on "Reproducibility of Data-Orient
held on 24-29 January 2016, focused on the cor
of experiments from a multidisciplin
several fields of comput
In this

PRIMAD - A Taxonomy for Reproducible IR Research



PRIMAD - A Taxonomy for Reproducible IR Research



Metadata Scheme

- **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/

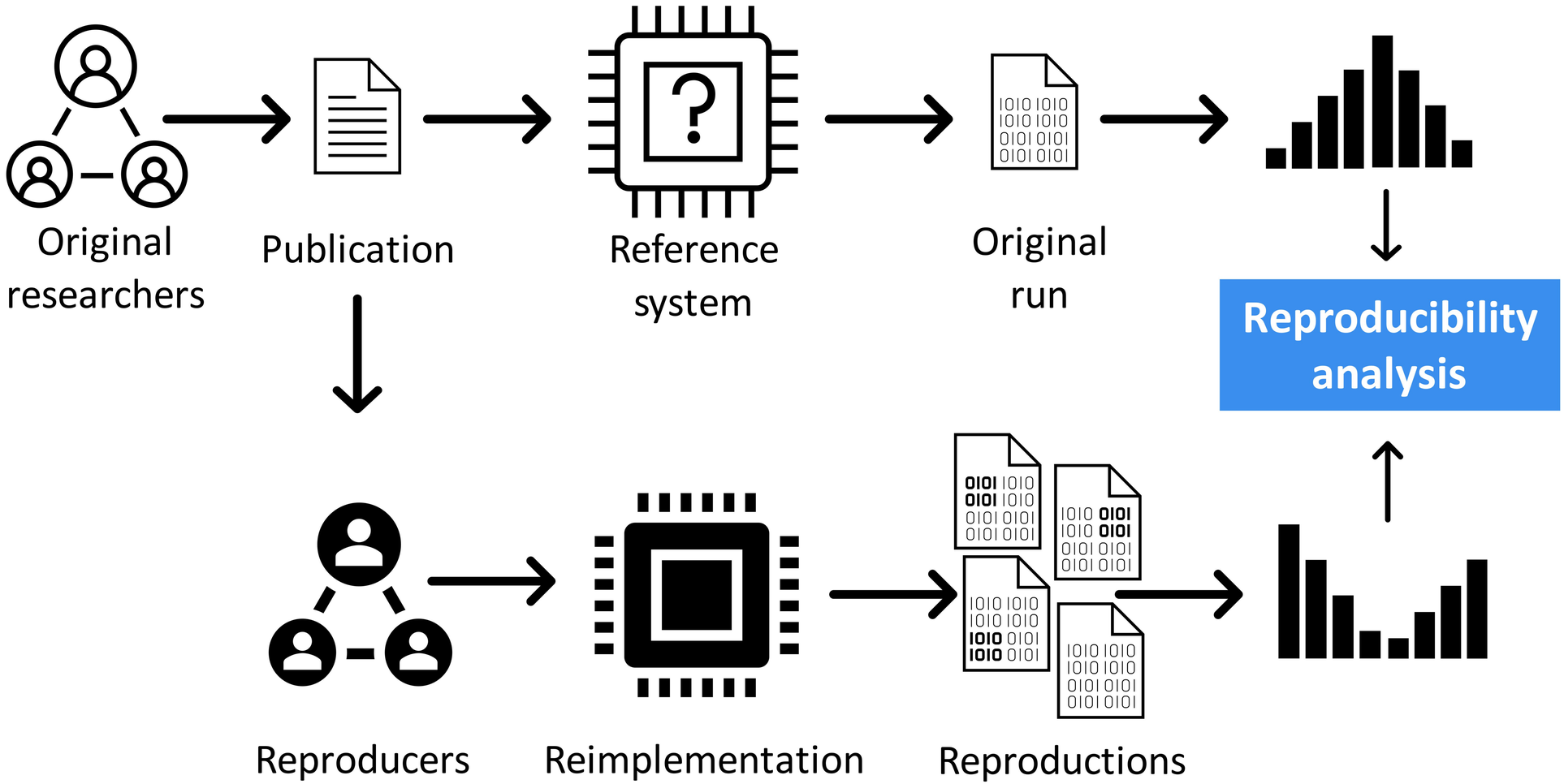
Metadata Annotations of Run Files

```
# ir_metadata.start
# platform:
#  ...
# research goal:
#  ...
# implementation:
#  ...
# method:
#  ...
# actor:
#  ...
# data:
#  ...
# ir_metadata.end
307          Q0          497476          1          0.9931          bm25
307          Q0          469928          2          0.9674          bm25
307          Q0          125806          3          0.9623          bm25
307          Q0          504815          4          0.9453          bm25
307          Q0          392547          5          0.9223          bm25
...
```

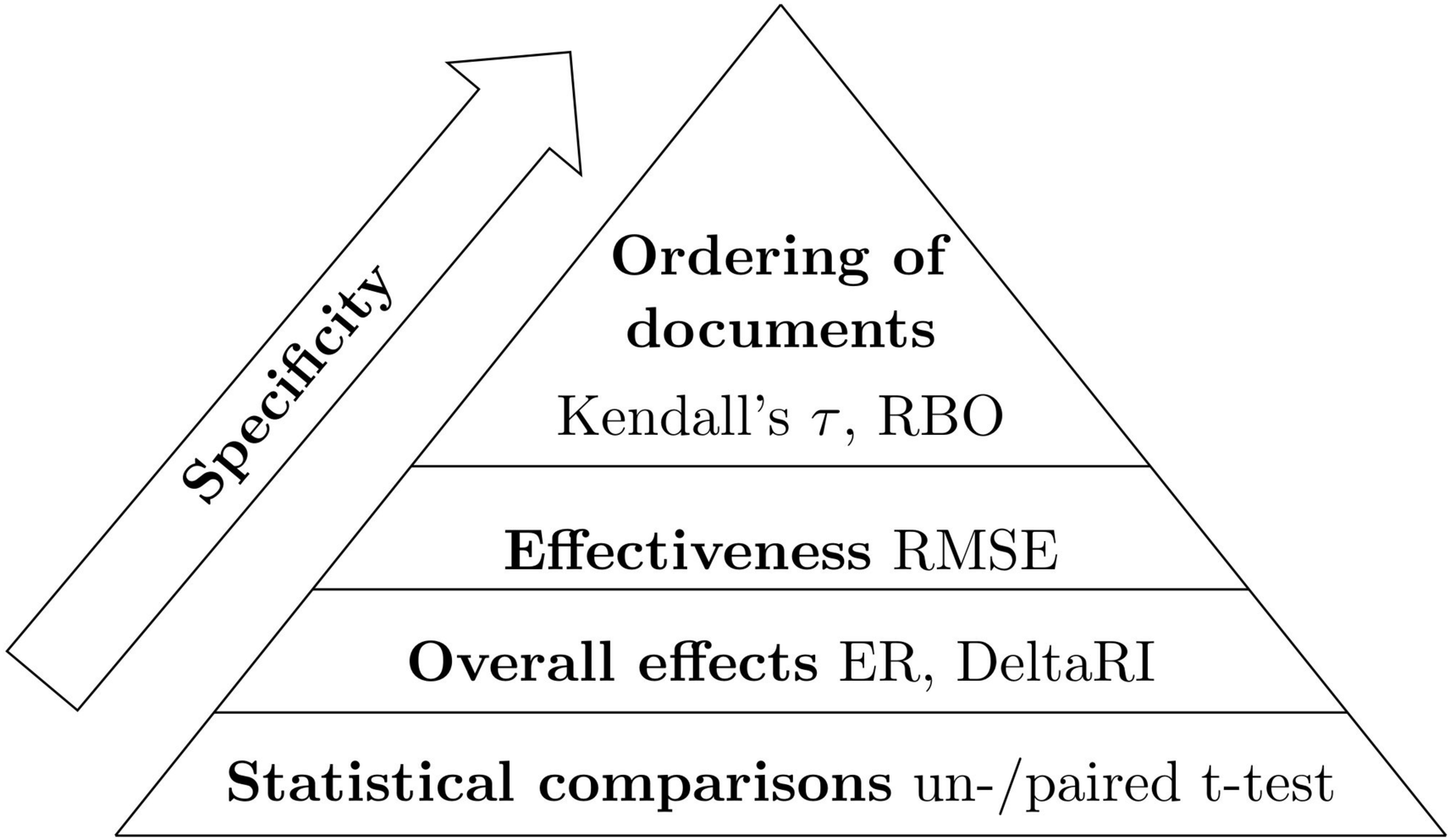
Metadata Annotations of Run Files

```
platform:
     hardware:
          cpu:
               model: 'Intel Xeon Gold 6144 CPU @ 3.50GHz'
               architecture: 'x86_64'
               operation mode: '64-bit'
               number of cores: 16
          ram: '64 GB'
     operating system:
          kernel: '5.4.0-90-generic'
          distribution: 'Ubuntu 20.04.3 LTS'
     software:
          libraries:
               python:
                  - 'scikit-learn==0.20.1'
                  - 'numpy==1.15.4'
               java:
                  - 'lucene==7.6'
          retrieval toolkit:
                  - 'anserini==0.3.0'
```

Reproducibility Analysis in IR



Reproducibility Measures

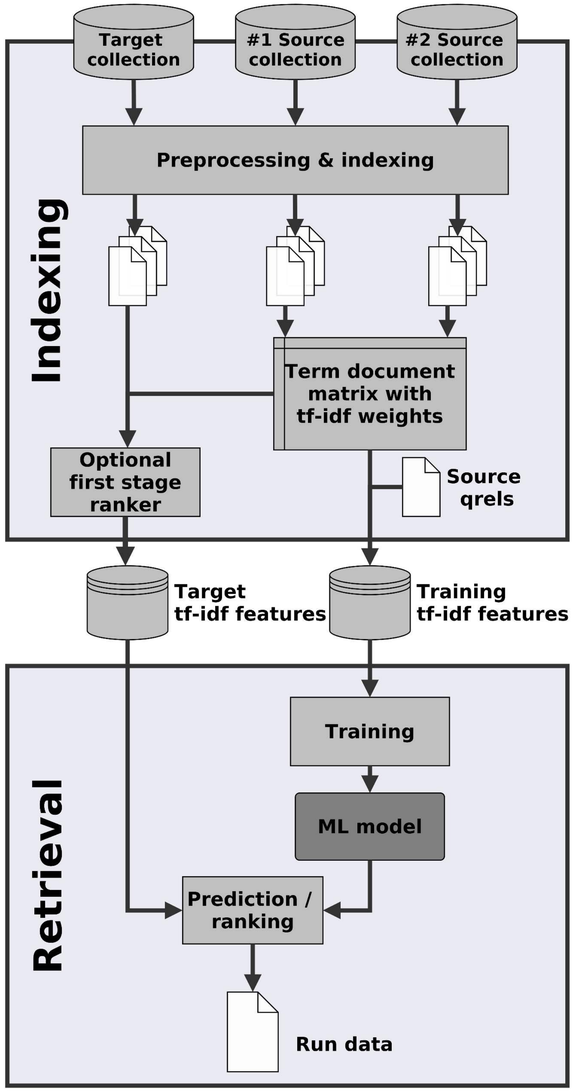


***How to Measure the Reproducibility of System-Oriented IR Experiments**; Breuer, Ferro, Fuhr, Maistro, Sakai, Schaer, Soboroff; SIGIR 2020*
***repro_eval: A Python Interface to Reproducibility Measures of System-Oriented IR Experiments**; Breuer, Ferro, Maistro, Schaer; ECIR 2021*

Meta-Evaluations / Reproducibility Experiments

Cross-collection relevance feedback
by Grossman and Cormack:

1. Derive tf-idf training samples from source collection(s)
2. Train topic-based relevance classifier
3. Rank target collection

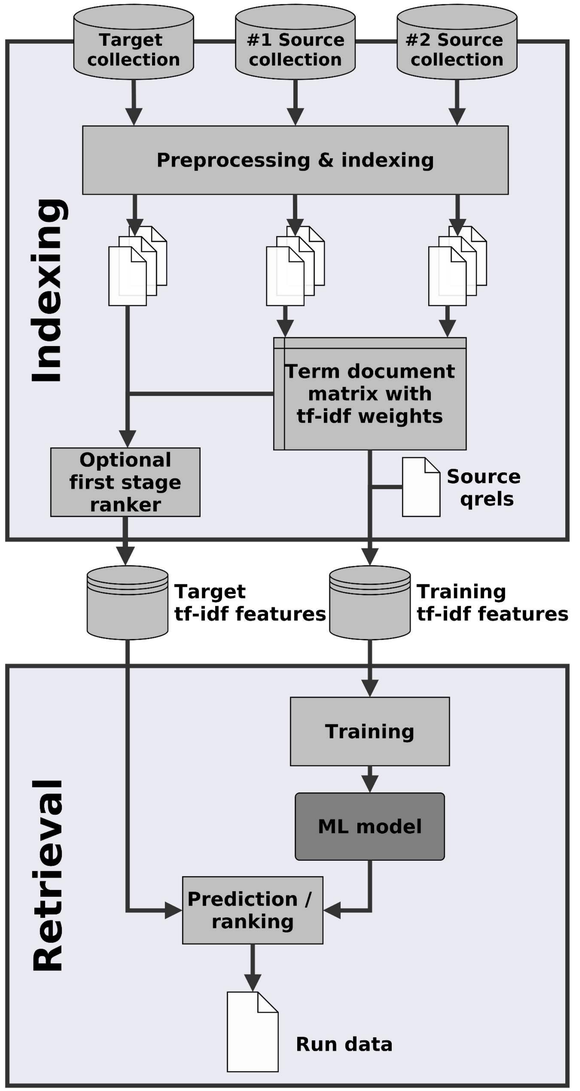


MRG_UWaterloo and WaterlooCormack Participation in the TREC 2017 Common Core Track;
Grossman and Cormack; TREC Common Core 2017
Simple Techniques for Cross-Collection Relevance Feedback;
Yu, Xie, and Lin; ECIR 2019
ir_metadata: An Extensible Metadata Schema for IR Experiments;
Breuer, Keller, Schaer; SIGIR 2022

Meta-Evaluations / Reproducibility Experiments

Dataset of runs

Researchers	Type	Venue
GC	Original experiment	TREC 2017
YXL	Reimplementation	ECIR 2019
BFFMSSS		SIGIR 2020
GC	Original experiment	TREC 2018
BPS	Reimplementation	CLEF 2021

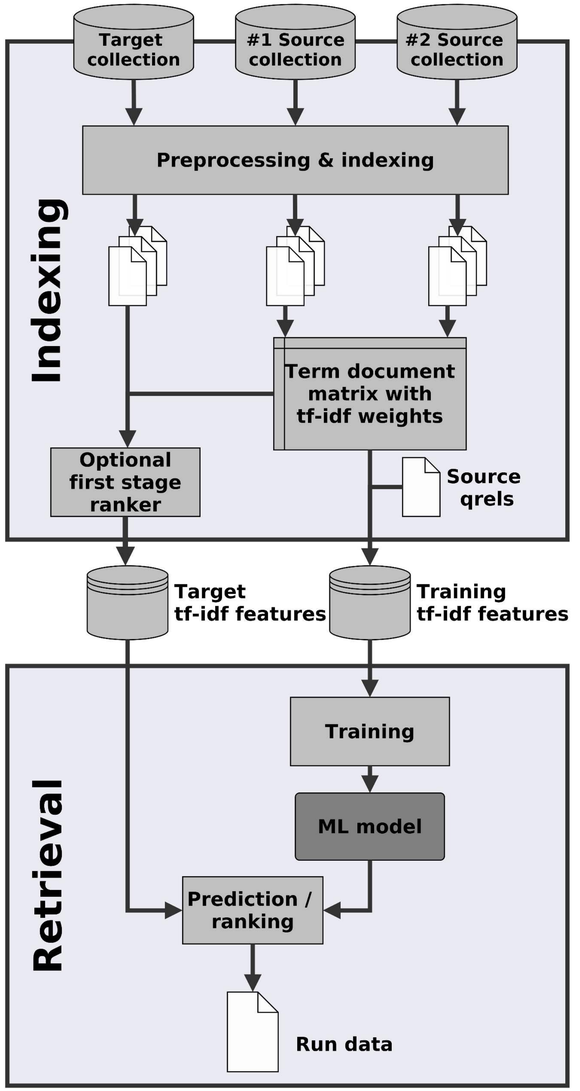


MRG_UWaterloo and WaterlooCormack Participation in the TREC 2017 Common Core Track;
Grossman and Cormack; TREC Common Core 2017
Simple Techniques for Cross-Collection Relevance Feedback;
Yu, Xie, and Lin; ECIR 2019
ir_metadata: An Extensible Metadata Schema for IR Experiments;
Breuer, Keller, Schaer; SIGIR 2022

Meta-Evaluations / Reproducibility Experiments

Experimental setups

Experiment	Type	Runs	Data
PRIM'AD	Parameter sweep	YXL	Core 17/18
P'R'I'M'A'D	Reproducibility	GC, YXL, BFFMSSS	Core 17
P'R'I'M'A'D'	Generalizability	GC, YXL, BPS	Core 17/18, Robust 04/05

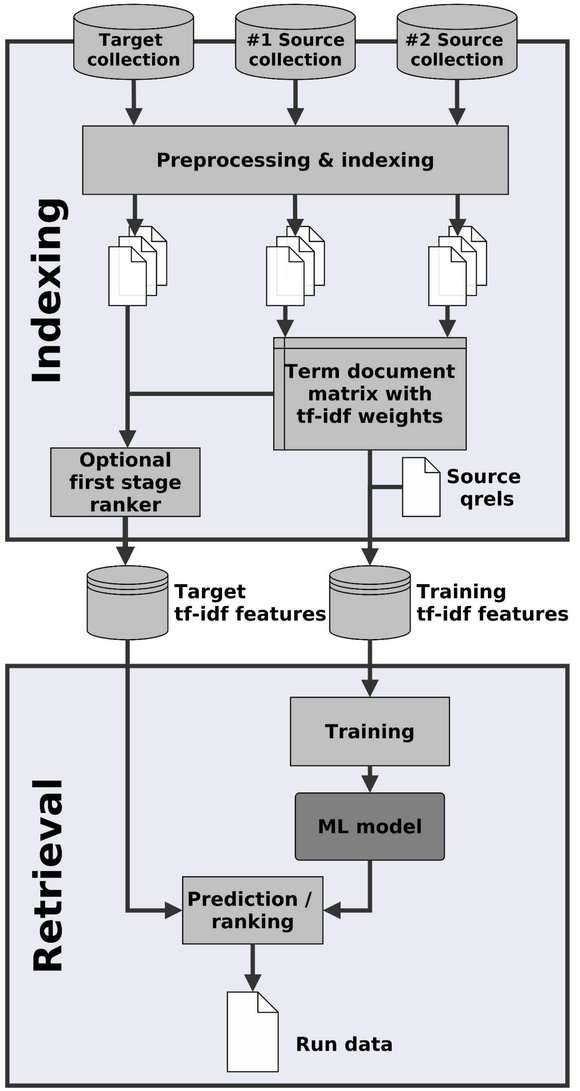


MRG_UWaterloo and WaterlooCormack Participation in the TREC 2017 Common Core Track;
Grossman and Cormack; TREC Common Core 2017
Simple Techniques for Cross-Collection Relevance Feedback;
Yu, Xie, and Lin; ECIR 2019
ir_metadata: An Extensible Metadata Schema for IR Experiments;
Breuer, Keller, Schaer; SIGIR 2022

Meta-Evaluations / Reproducibility Experiments

P’R’I’M’A’D

Measure	GC	YXL	BFFMSSS
Average Precision	0.3711	0.4018	0.3612
Kendall’s tau Union	1.0000	0.0086	0.0051
Rank-biased Overlap	1.0000	0.1630	0.5747
Root Mean Square Error	0.0000	0.1911	0.1071
p-value (paired t-test)	1.0000	0.1009	0.7885
Effect Ratio	1.0000	0.8267	1.0514
Delta Relative Improvement	0.0000	0.0362	-0.0123

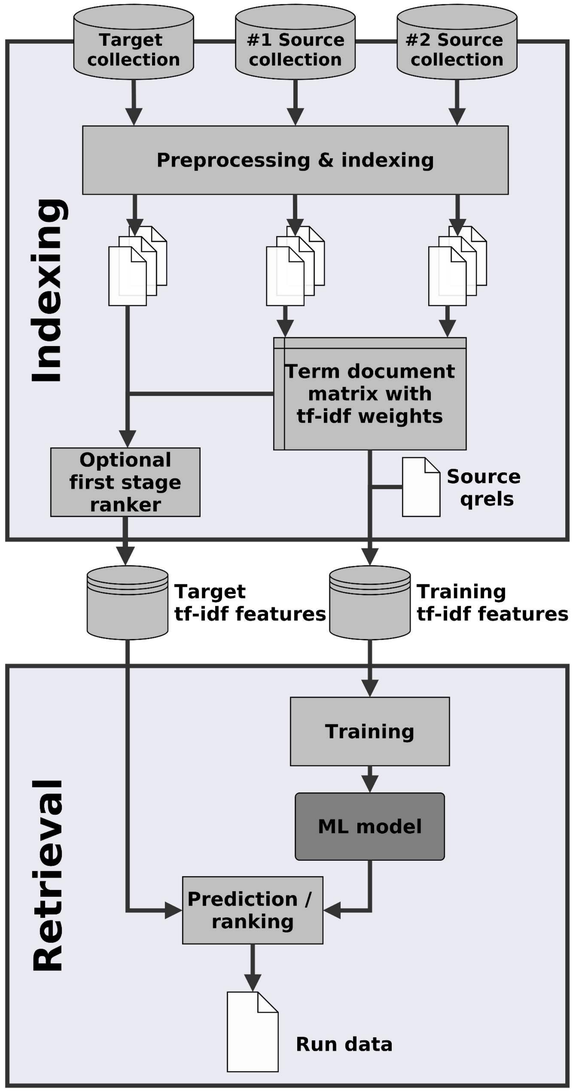


Test collection: TREC Common Core 2017 (The New York Times Annotated Corpus), 50 Topics

Meta-Evaluations / Reproducibility Experiments

P’R’I’M’A’D

Measure	GC	YXL	BFFMSSS
Average Precision	0.3711	0.4018	0.3612
Kendall’s tau Union	1.0000	0.0086	0.0051
Rank-biased Overlap	1.0000	0.1630	0.5747
Root Mean Square Error	0.0000	0.1911	0.1071
p-value (paired t-test)	1.0000	0.1009	0.7885
Effect Ratio	1.0000	0.8267	1.0514
Delta Relative Improvement	0.0000	0.0362	-0.0123

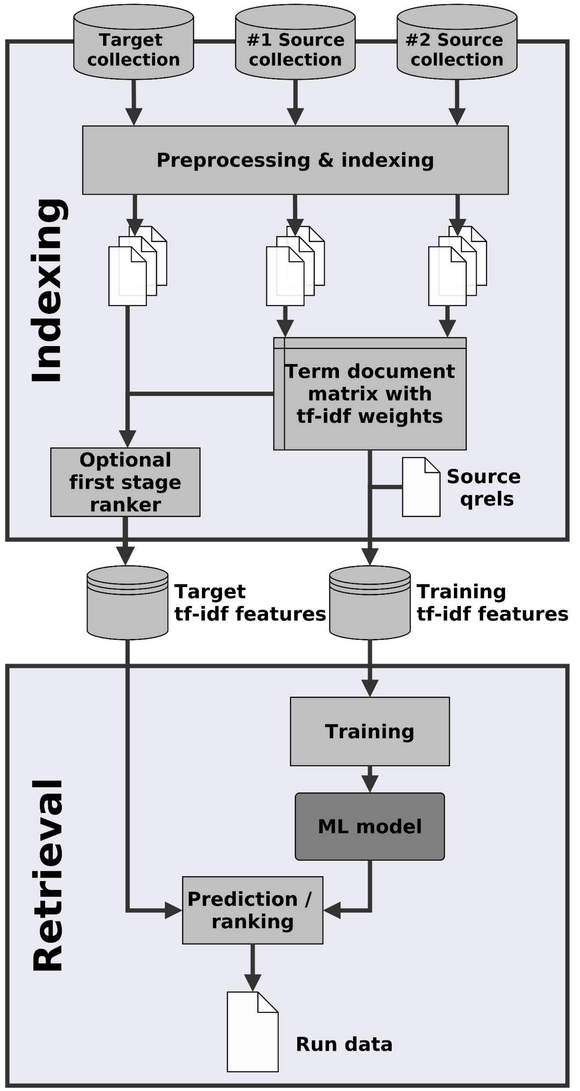


Test collection: TREC Common Core 2017 (The New York Times Annotated Corpus), 50 Topics

Meta-Evaluations / Reproducibility Experiments

P’R’I’M’A’D

Measure	GC	YXL	BFFMSSS
Average Precision	0.3711	0.4018	0.3612
Kendall’s tau Union	1.0000	0.0086	0.0051
Rank-biased Overlap	1.0000	0.1630	0.5747
Root Mean Square Error	0.0000	0.1911	0.1071
p-value (paired t-test)	1.0000	0.1009	0.7885
Effect Ratio	1.0000	0.8267	1.0514
Delta Relative Improvement	0.0000	0.0362	-0.0123

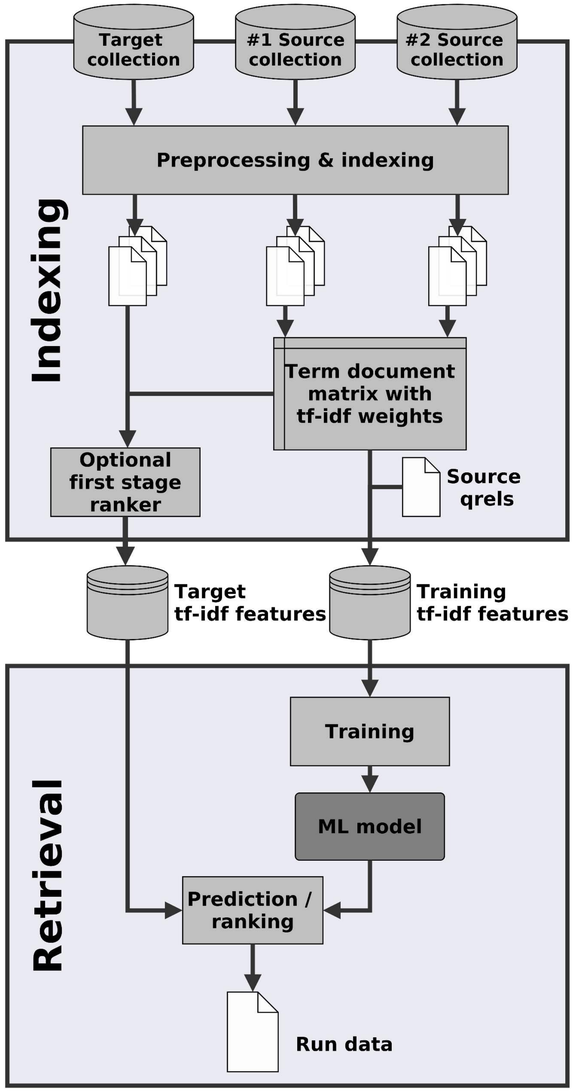


Test collection: TREC Common Core 2017 (The New York Times Annotated Corpus), 50 Topics

Meta-Evaluations / Reproducibility Experiments

P’R’I’M’A’D

Measure	GC	YXL	BFFMSSS
Average Precision	0.3711	0.4018	0.3612
Kendall’s tau Union	1.0000	0.0086	0.0051
Rank-biased Overlap	1.0000	0.1630	0.5747
Root Mean Square Error	0.0000	0.1911	0.1071
p-value (paired t-test)	1.0000	0.1009	0.7885
Effect Ratio	1.0000	0.8267	1.0514
Delta Relative Improvement	0.0000	0.0362	-0.0123

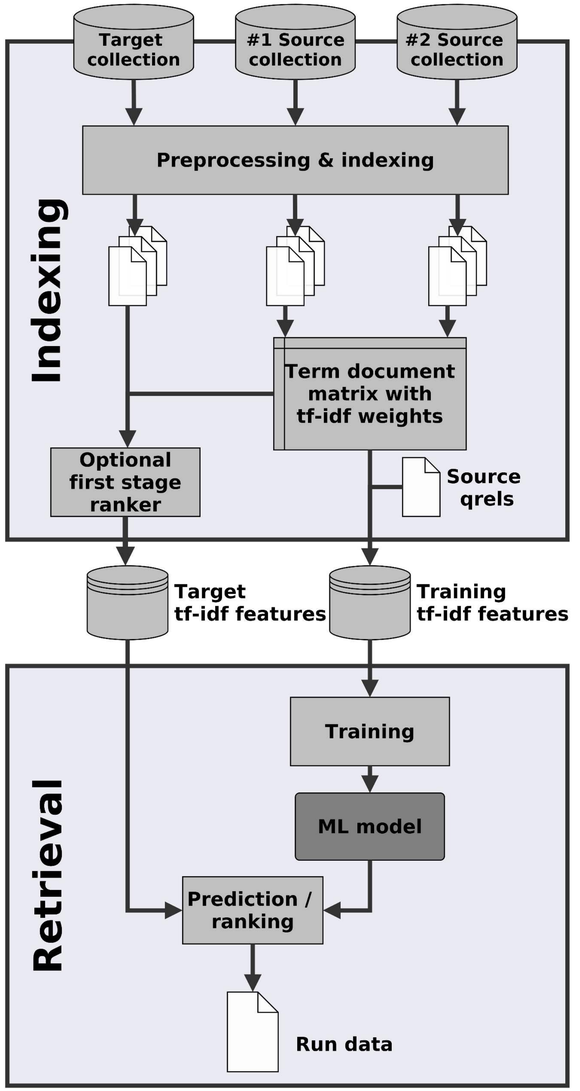


Test collection: TREC Common Core 2017 (The New York Times Annotated Corpus), 50 Topics

Meta-Evaluations / Reproducibility Experiments

P’R’I’M’A’D

Measure	GC	YXL	BFFMSSS
Average Precision	0.3711	0.4018	0.3612
Kendall’s tau Union	1.0000	0.0086	0.0051
Rank-biased Overlap	1.0000	0.1630	0.5747
Root Mean Square Error	0.0000	0.1911	0.1071
p-value (paired t-test)	1.0000	0.1009	0.7885
Effect Ratio	1.0000	0.8267	1.0514
Delta Relative Improvement	0.0000	0.0362	-0.0123



Test collection: TREC Common Core 2017 (The New York Times Annotated Corpus), 50 Topics

Meta-Evaluations / Reproducibility Experiments

P’R’I’M’A’D

Internal Validity

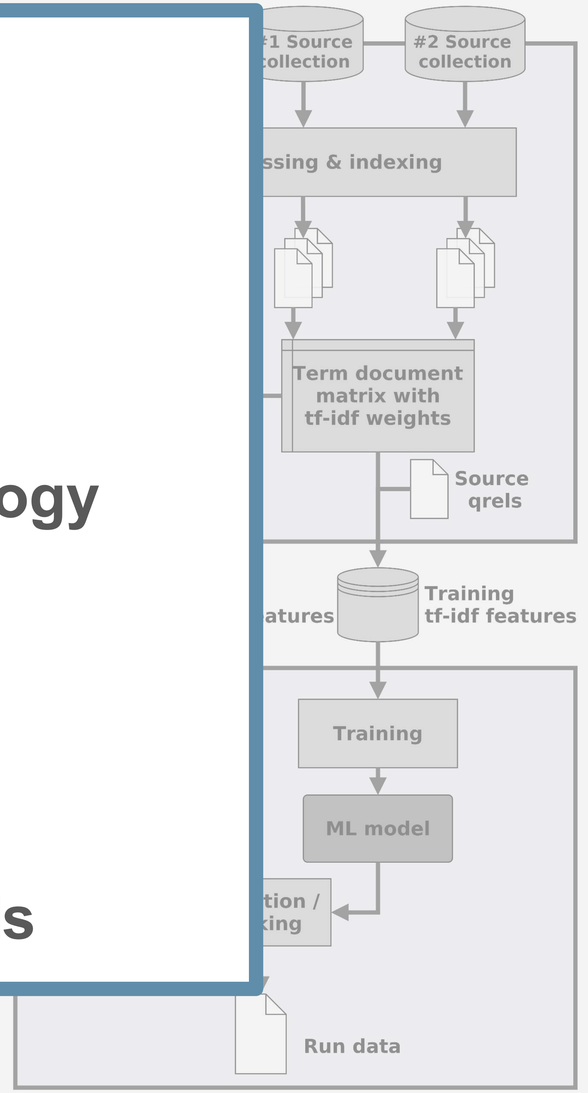
PRIMAD-based metadata annotations

- support proactive reproducibility
- **address inconsistent use of terminology**

Principled evaluations and measures

- support reactive reproducibility
- **address missing evaluation standards**

Measure			
Average Precision			
Kendall’s tau Union			
Rank-biased Overlap			
Root Mean Square Er			
p-value (paired t-test)			
Effect Ratio			
Delta Relative Improvement	0.0000	0.0362	-0.0123



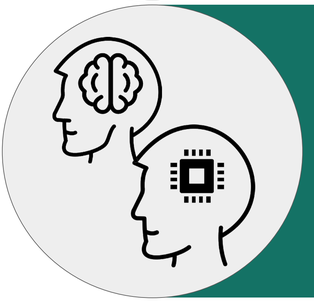
Test collection: TREC Common Core 2017 (The New York Times Annotated Corpus), 50 Topics

Outline and Contributions

INTERNAL VALIDITY

system-oriented experiments

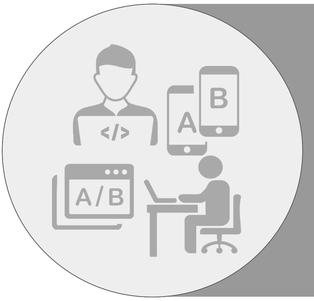
- PRIMAD extensions and metadata scheme
- Principled reproducibility evaluations



EXTERNAL VALIDITY

user simulations

- Query simulations and evaluation framework
- Click-based evaluations of system rankings



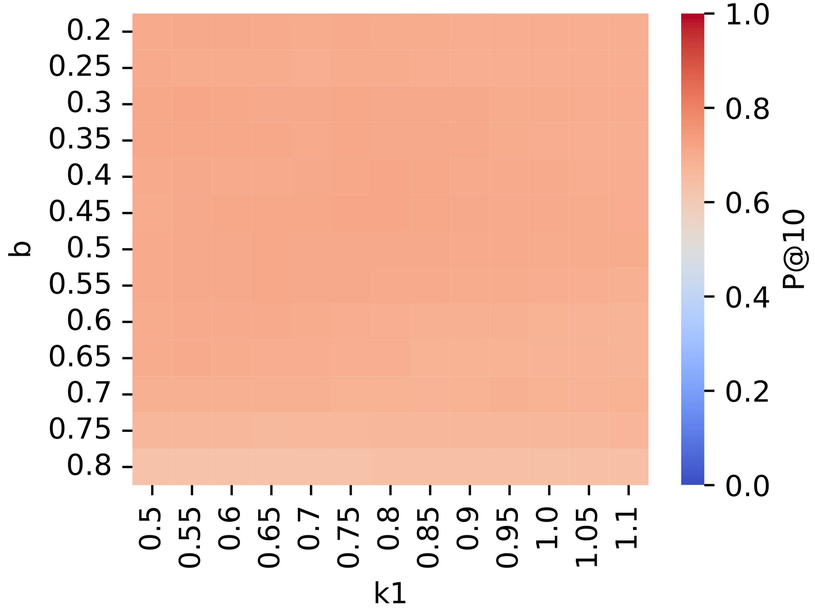
ECOLOGICAL VALIDITY

real user experiments

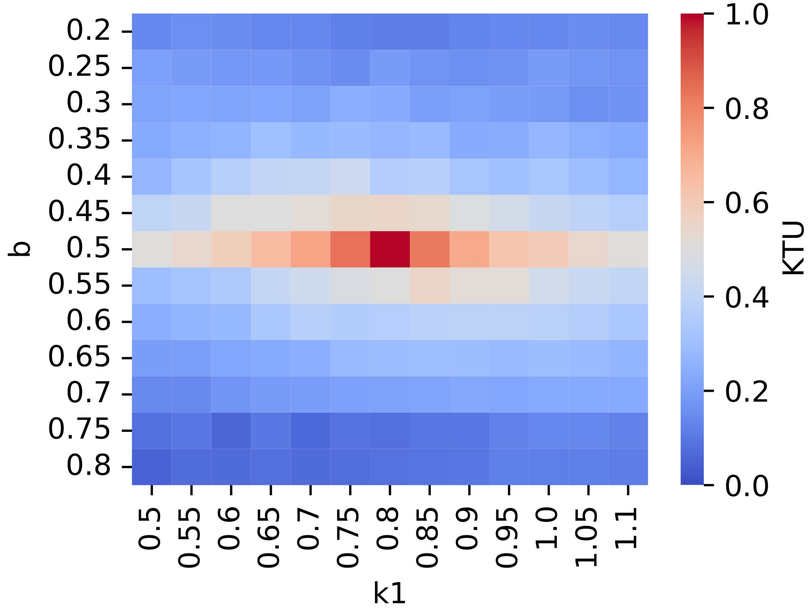
- Living lab infrastructure
- Shared task evaluations

What about the User?

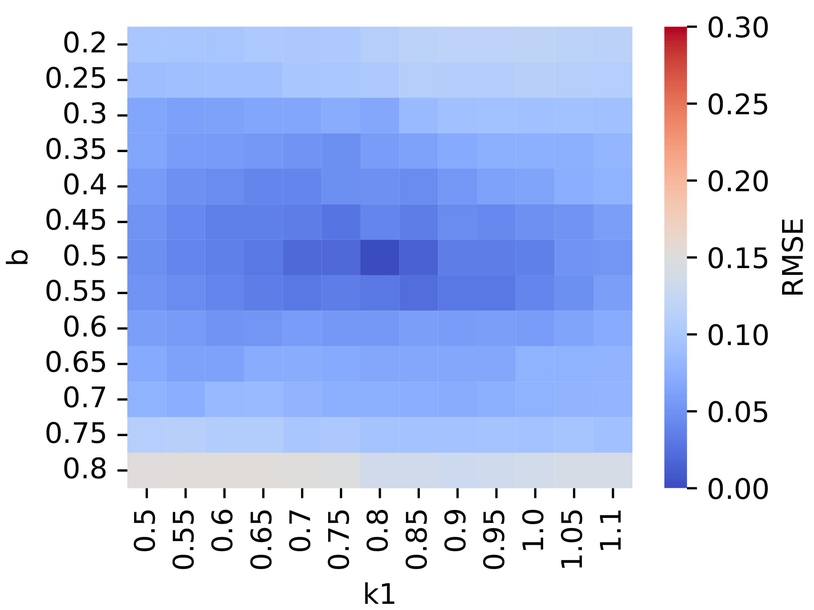
Same system effectiveness



Different document rankings

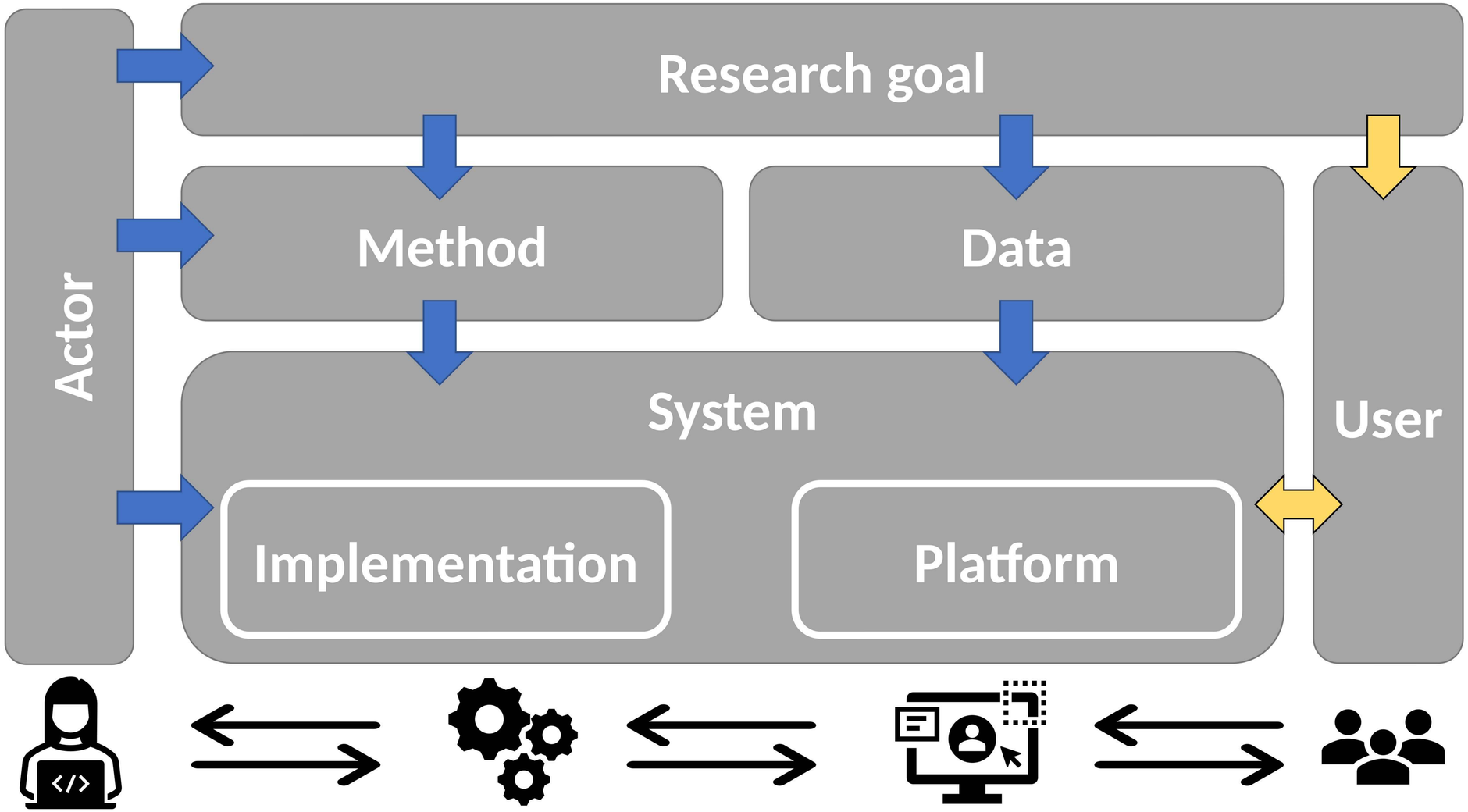


Slightly different topic scores

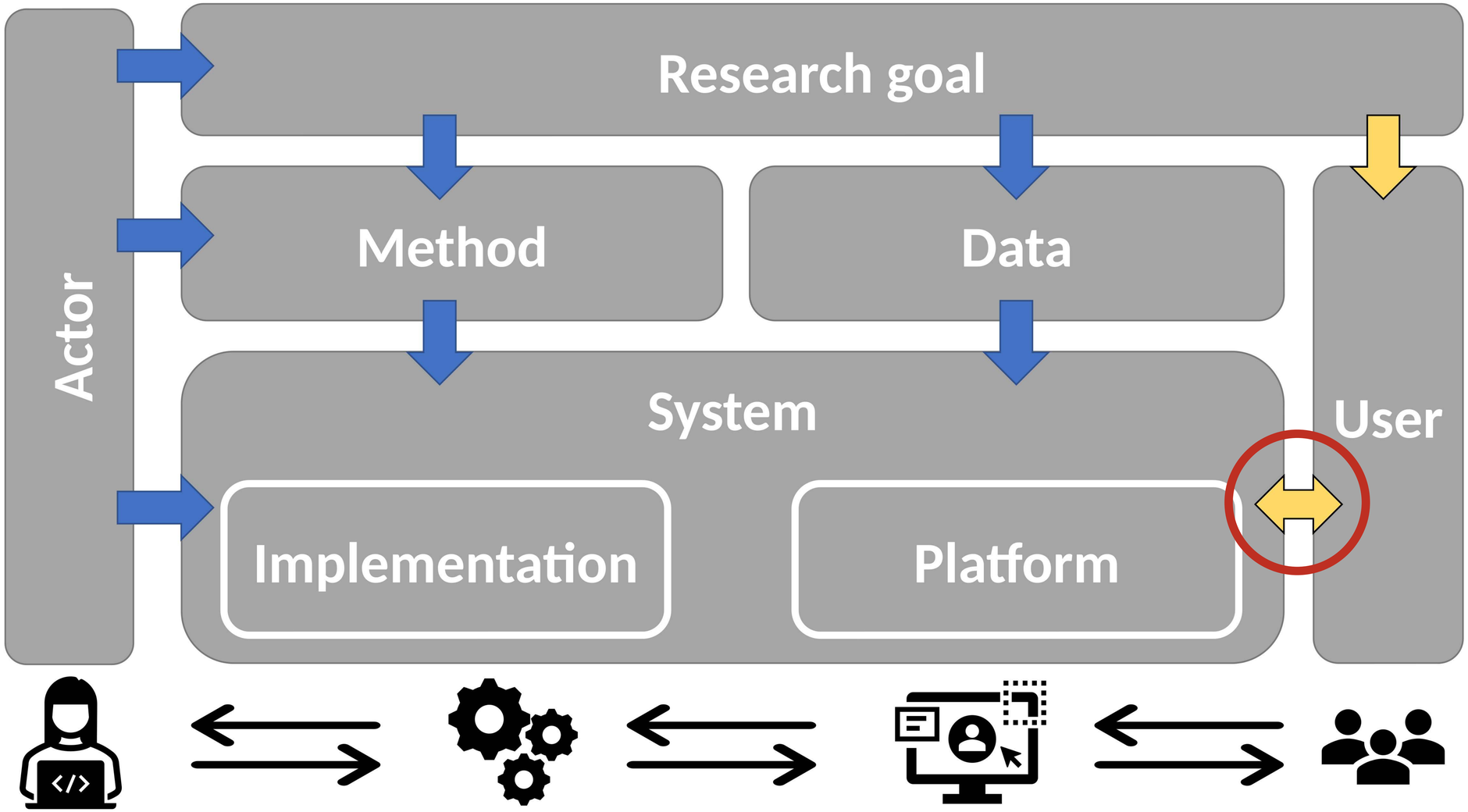


Test collection: *TREC COVID (Rnd. 5)* https://ir.nist.gov/trec-covid/

PRIMAD-U - A Holistic View on the IR Experiment



PRIMAD-U - A Holistic View on the IR Experiment

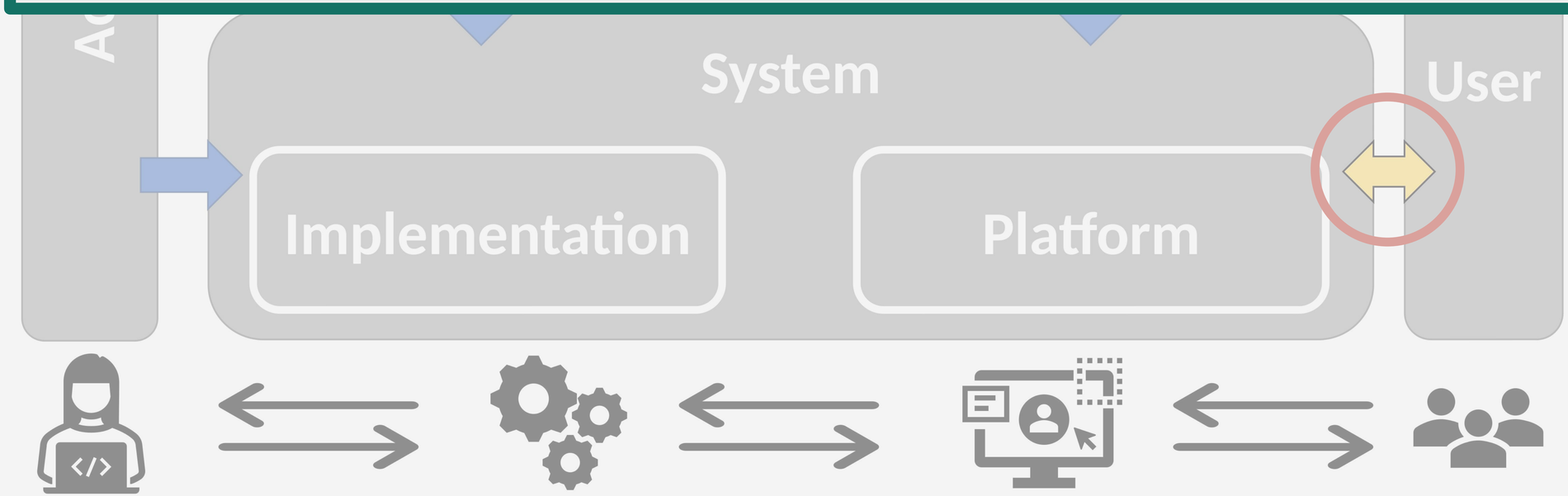


PRIMAD-U - A Holistic View on the IR Experiment

Query simulations and evaluation framework

- Simulating **system input**, i.e., user query variants (UQV)
- UQV simulation method and evaluation framework

Reproducibility of system-oriented outcomes based on UQV



PRIMAD-U - A Holistic View on the IR Experiment

Query simulations and evaluation framework

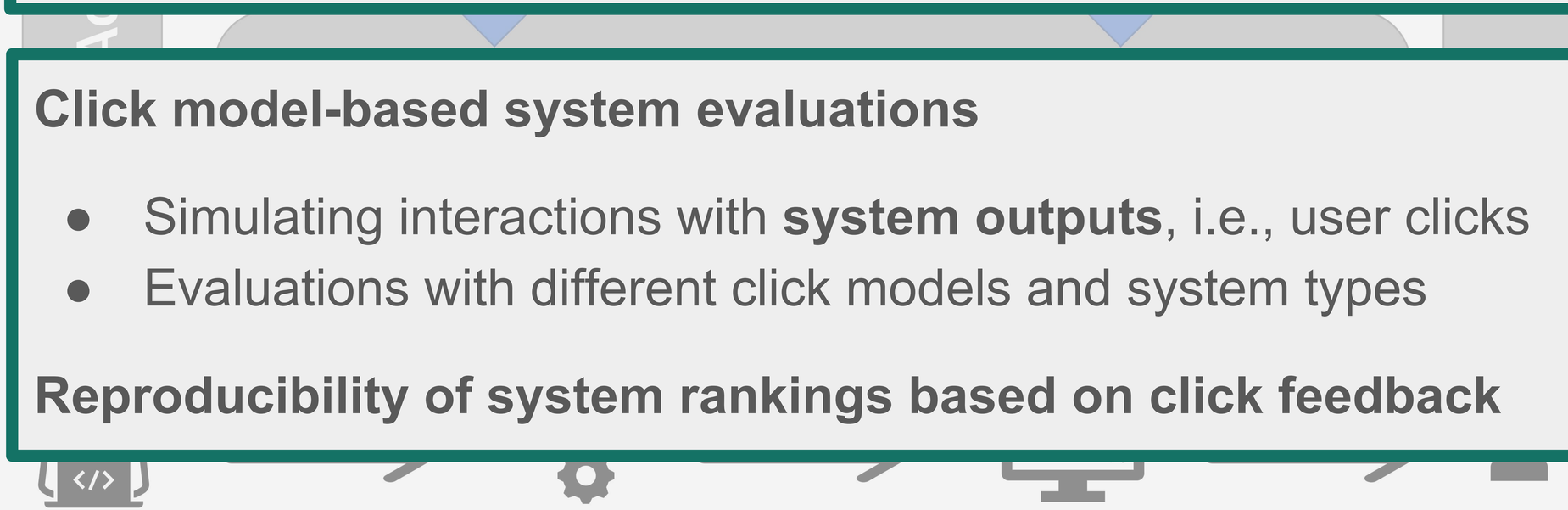
- Simulating **system input**, i.e., user query variants (UQV)
- UQV simulation method and evaluation framework

Reproducibility of system-oriented outcomes based on UQV

Click model-based system evaluations

- Simulating interactions with **system outputs**, i.e., user clicks
- Evaluations with different click models and system types

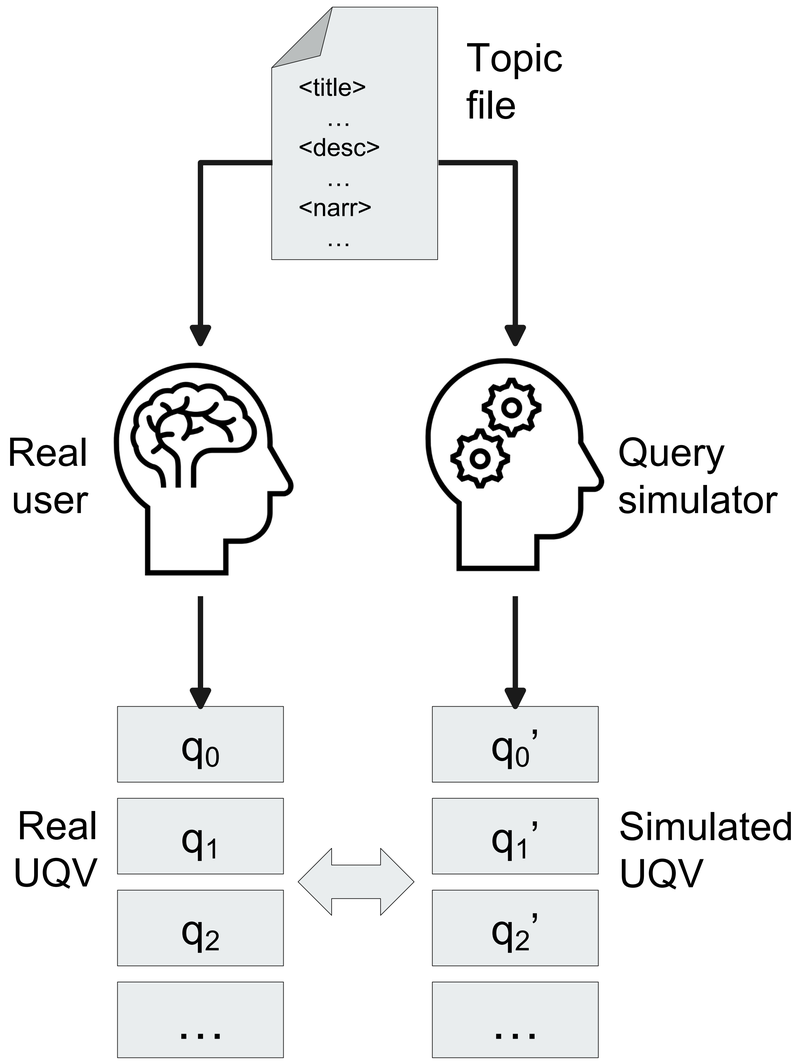
Reproducibility of system rankings based on click feedback



Simulating User Query Variants

- Simulations of **user query variants (UQV)** for a given topic of a TREC test collection

- Evaluation of different **query simulators**:
  - **TREC Topic Searcher**
  - **Known-item Searcher**
  - New method based on **Controlled Query Generation** and **Query Change Model**



***Validating Simulations of User Query Variants**; Breuer, Fuhr, Schaer; ECIR 2022*
***Using Controlled Query Generation to Evaluate Blind Relevance Feedback Algorithm**; Jordan, Watters, Gao; JCDL 2006*
***The Query Change Model: Modeling Session Search as a Markov Decision Process**; Yang, Guan, Zhang.; TOIS 2015*

UQV Evaluation Framework

Retrieval performance

- **Average retrieval performance**
- Root-Mean-Square-Error (**RMSE**)
- **p-values** of paired t-tests

Effort and effect

- **Session-discounted cumulative gain**
- Trade-off analysis of the **number of queries** and **browsing depth**

Shared task utility

- Relative system orderings compared by **Kendall's tau**

Query term similarity

- **Jaccard similarity** between the sets of query terms

***Validating Simulations of User Query Variants**; Breuer, Fuhr, Schaer; ECIR 2022*

UQV Evaluation Framework

Retrieval performance

- **Average retrieval performance**
- Root-Mean-Square-Error (**RMSE**)
- **p-values** of paired t-tests

Effort and effect

- **Session-discounted cumulative gain**
- Trade-off analysis of the **number of queries** and **browsing depth**

Shared task utility

- Relative system orderings compared by **Kendall's tau**

Query term similarity

- **Jaccard similarity** between the sets of query terms

***Validating Simulations of User Query Variants**; Breuer, Fuhr, Schaer; ECIR 2022*

Effort and Effect

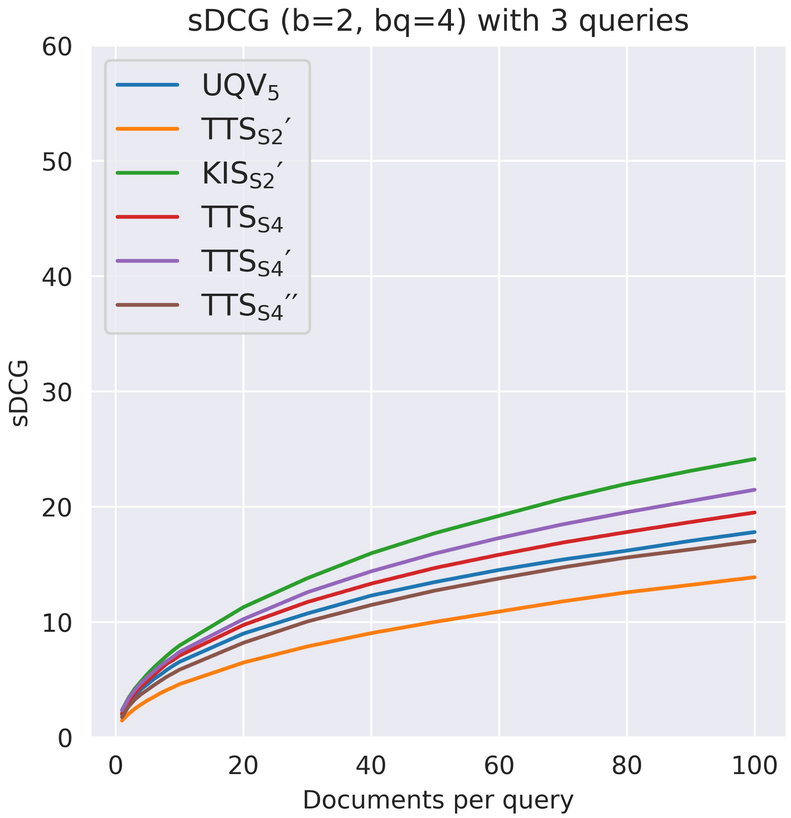
Session-based Discounted Cumulated Gain (sDCG)

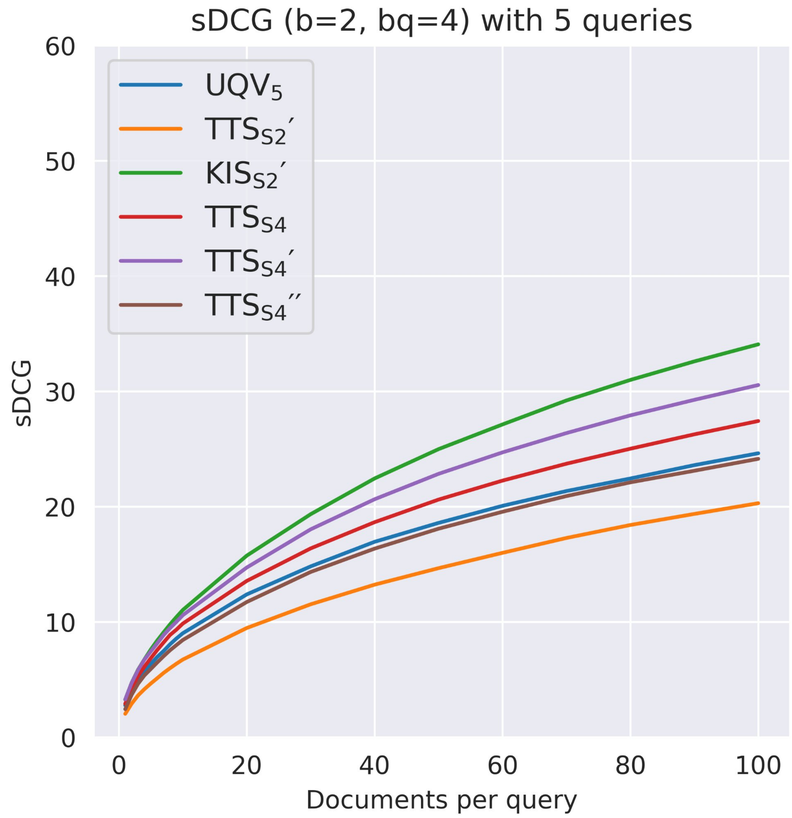
$$\mathrm{sDCG} = \sum_{i \in \{1,\ldots,n\}} \frac{\mathrm{DCG}_{q_i}}{1 + \log_{bq}(i)}$$

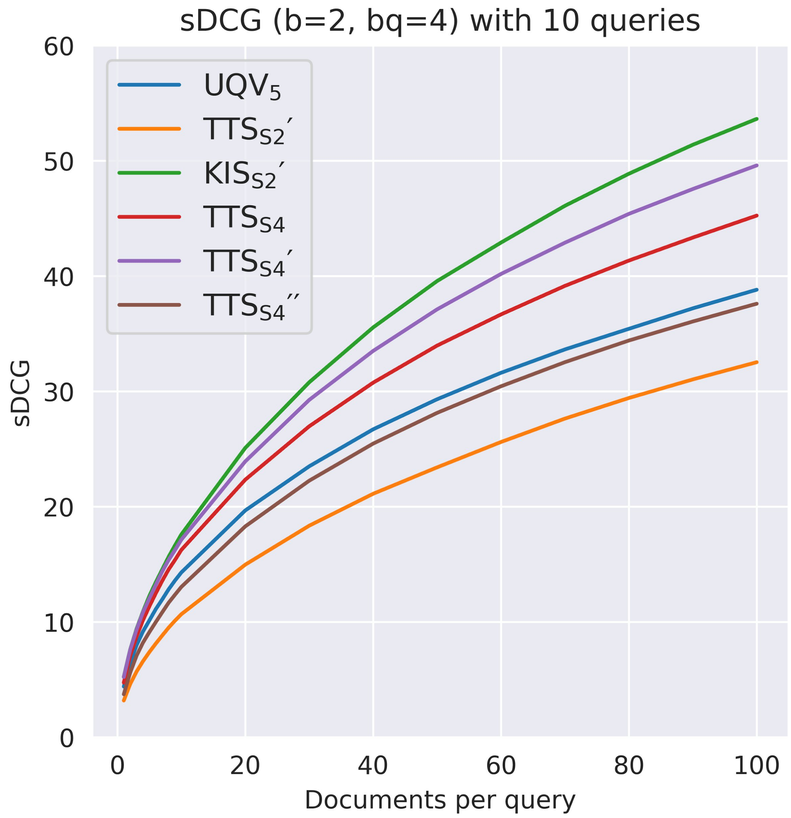
bq	logarithm base for the query discount
q_i	query at the i-th position in a session
DCG_{q_i}	discounted cumulative gain

***Discounted Cumulated Gain based Evaluation of Multiple-Query IR Sessions**; Järvelin, Price, Delcambre, Nielsen; ECIR 2008*

Effort and Effect







Test collection: TREC Common Core 2017 (The New York Times Annotated Corpus), 50 Topics
UQV dataset: Benham and Culpepper, https://culpepper.io/publications/robust-uqv.txt.gz

***Validating Simulations of User Query Variants**; Breuer, Fuhr, Schaer; ECIR 2022*

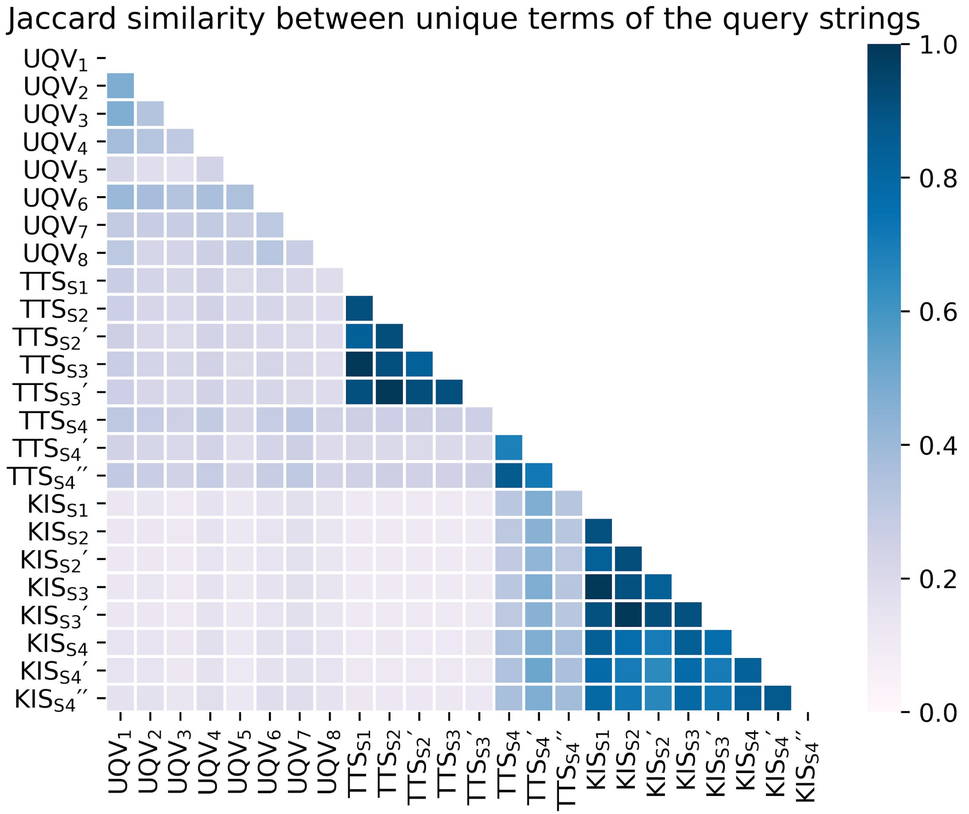
Query Term Similarity

Jaccard similarity as a measure of term variance

$$J(Q, Q') = \frac{|Q \cap Q'|}{|Q \cup Q'|}$$

Q reference query terms, e.g., real user query terms

Q' evaluated query terms, e.g., simulated query terms



***Validating Simulations of User Query Variants**; Breuer, Fuhr, Schaer; ECIR 2022*
***A Comparative Analysis of Human and Automatic Query Variants**; Liu, Craswell, Lu, Kurland, Culpepper; ICTIR 2019*

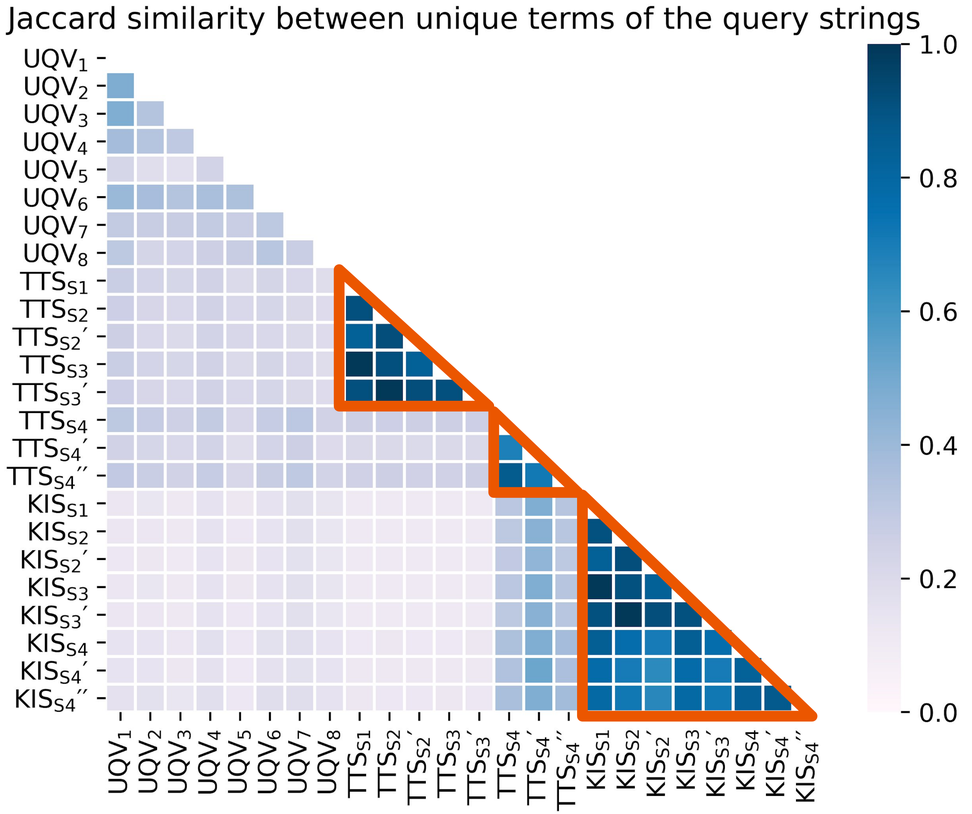
Query Term Similarity

Jaccard similarity as a measure of term variance

$$J(Q, Q') = \frac{|Q \cap Q'|}{|Q \cup Q'|}$$

Q reference query terms, e.g., real user query terms

Q' evaluated query terms, e.g., simulated query terms



***Validating Simulations of User Query Variants**; Breuer, Fuhr, Schaer; ECIR 2022*
***A Comparative Analysis of Human and Automatic Query Variants**; Liu, Craswell, Lu, Kurland, Culpepper; ICTIR 2019*

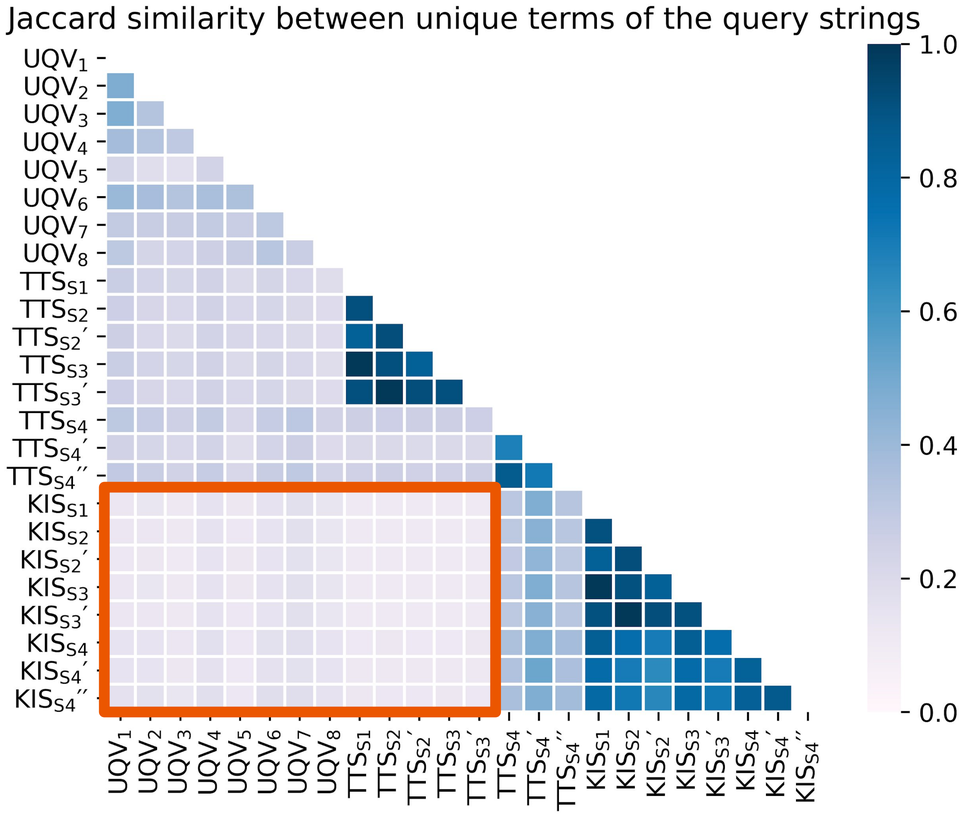
Query Term Similarity

Jaccard similarity as a measure of term variance

$$J(Q, Q') = \frac{|Q \cap Q'|}{|Q \cup Q'|}$$

Q reference query terms, e.g., real user query terms

Q' evaluated query terms, e.g., simulated query terms



***Validating Simulations of User Query Variants**; Breuer, Fuhr, Schaer; ECIR 2022*
***A Comparative Analysis of Human and Automatic Query Variants**; Liu, Craswell, Lu, Kurland, Culpepper; ICTIR 2019*

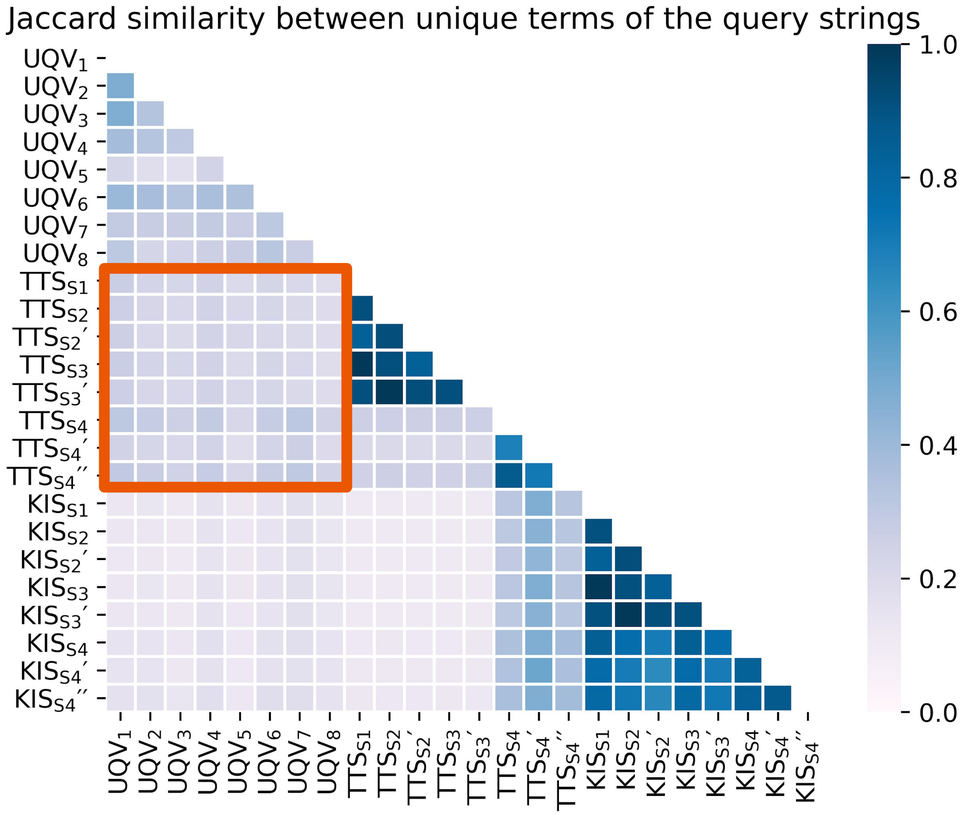
Query Term Similarity

Jaccard similarity as a measure of term variance

$$J(Q, Q') = \frac{|Q \cap Q'|}{|Q \cup Q'|}$$

Q reference query terms, e.g., real user query terms

Q' evaluated query terms, e.g., simulated query terms



***Validating Simulations of User Query Variants**; Breuer, Fuhr, Schaer; ECIR 2022*
***A Comparative Analysis of Human and Automatic Query Variants**; Liu, Craswell, Lu, Kurland, Culpepper; ICTIR 2019*

Query Term Similarity

Jaccard

ry strings

Q refe

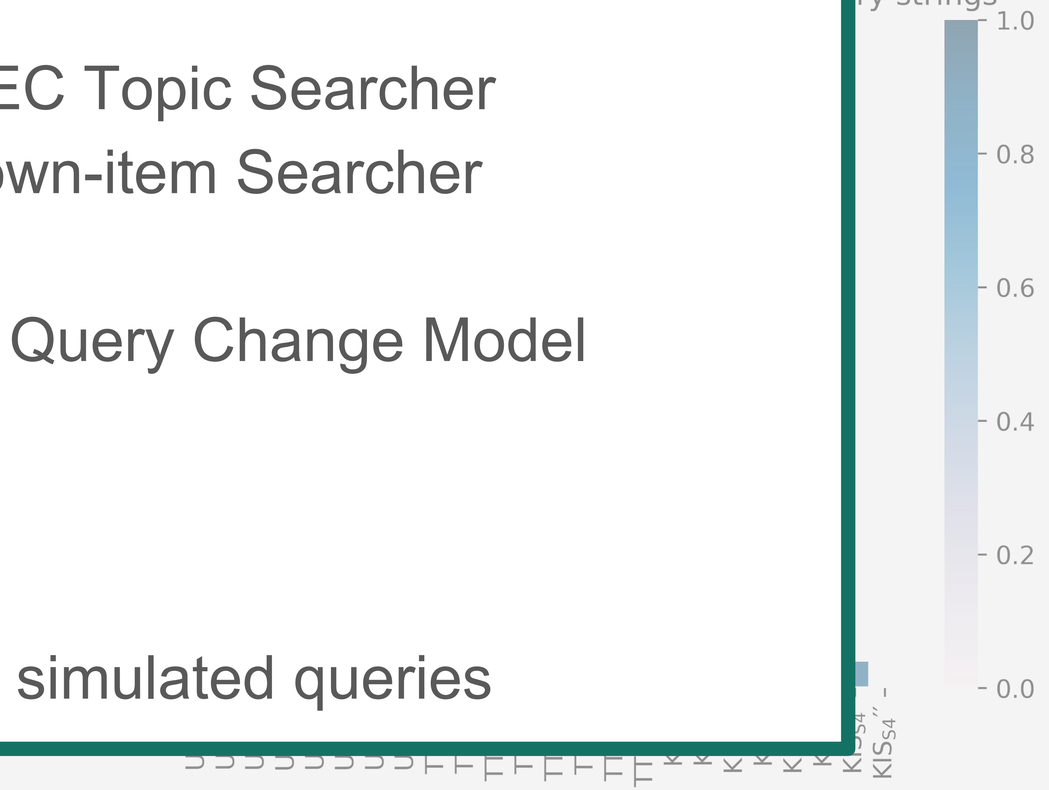
Q' eva

Effort and Effect

- **Lower bound performance:** TREC Topic Searcher
- **Upper bound performance:** Known-item Searcher
- **Better reproductions** based on
  Controlled Query Generation and Query Change Model

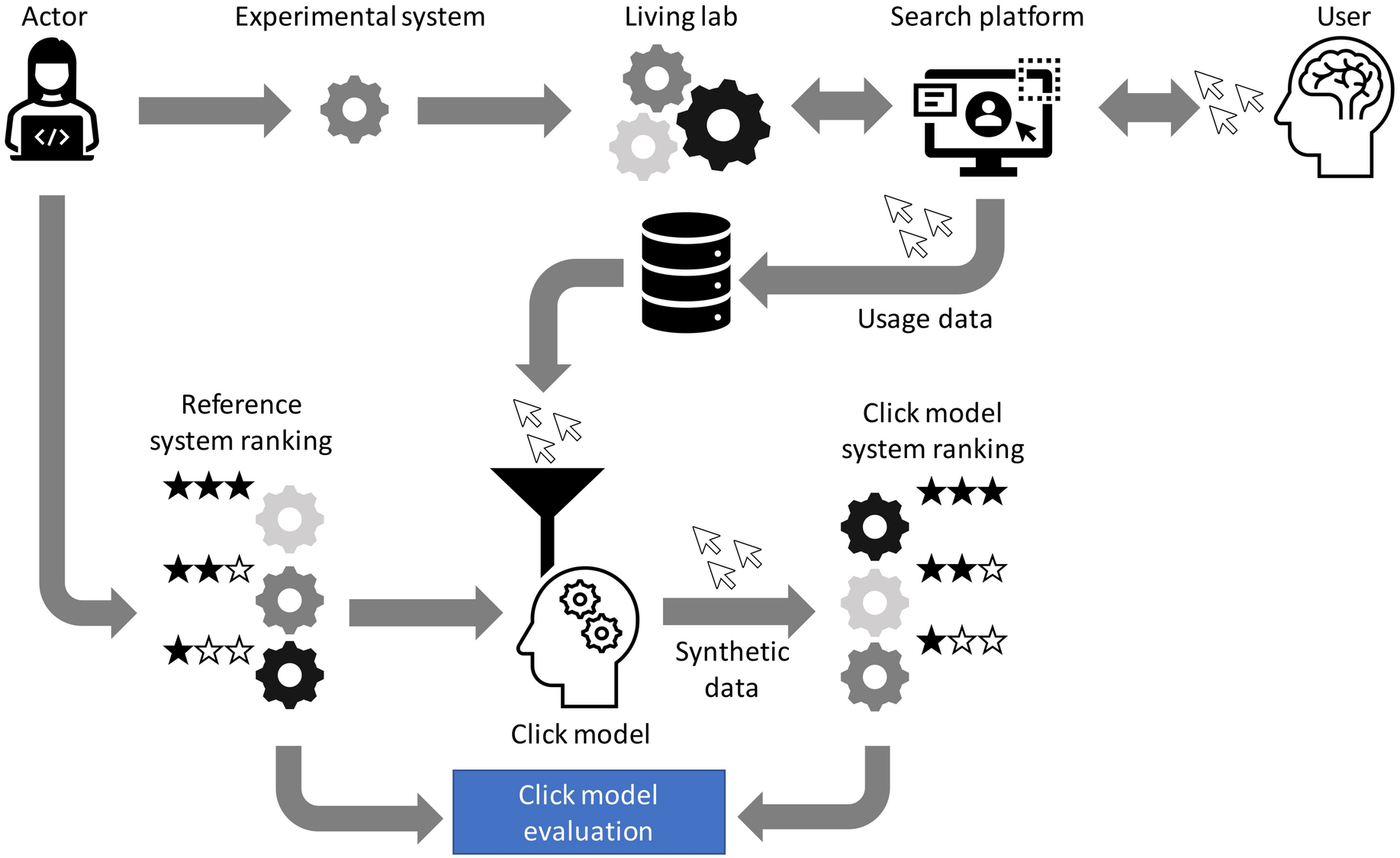
Query Term Similarity

- **Partial overlap** between real and simulated queries



__Validating Simulations of User Query Variants__; Breuer, Fuhr, Schaer; ECIR 2022
__A Comparative Analysis of Human and Automatic Query Variants__; Liu, Craswell, Lu, Kurland, Culpepper; ICTIR 2019

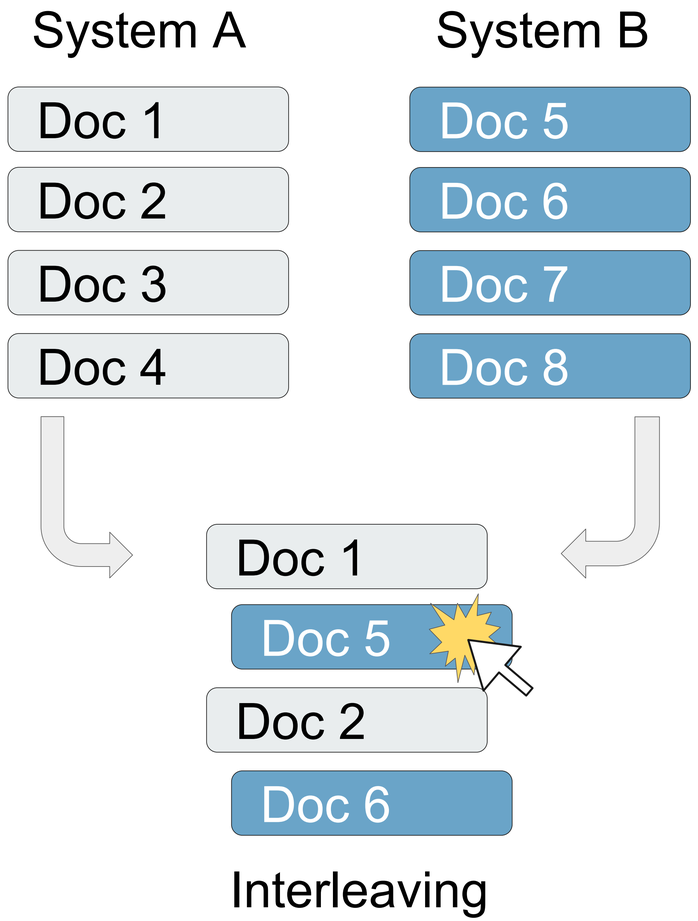
Can Click Models Reproduce System Rankings?



Can Click Models Reproduce System Rankings?

- Click models generate a click probability (P_{Click})
- System with highest P_{Click} **wins**, the other system **loses**
- **System ranking** is determined by:

$$Outcome = \frac{Wins}{Wins + Losses}$$



Experimental Setup

Two types of **system rankings**

- **LRM -** **L**exical **r**etrieval **m**ethods

$$\text{DFR}\chi^2 \succ \text{BM25} \succ \text{Tf} \succ \text{Dl} \succ \text{Null}$$

- **IRM -** **I**nterpolated **r**etrieval **m**ethods

$$\text{IRM}_{\rho=0.4} \succ \text{IRM}_{\rho=0.45} \succ \cdots \succ \text{IRM}_{\rho=1.0}$$

$$\text{score}(d,q) = \rho \cdot \text{score}_{\text{Dl}}(d,q) + (1-\rho) \cdot \text{score}_{\text{DFR}}(d,q)$$

Experimental Setup

Two types of **system rankings**

- **LRM -** **L**exical **r**etrieval **m**ethods

$$\text{DFR}\chi^2 \succ \text{BM25} \succ \text{Tf} \succ \text{Dl} \succ \text{Null}$$

- **IRM -** **I**nterpolated **r**etrieval **m**ethods

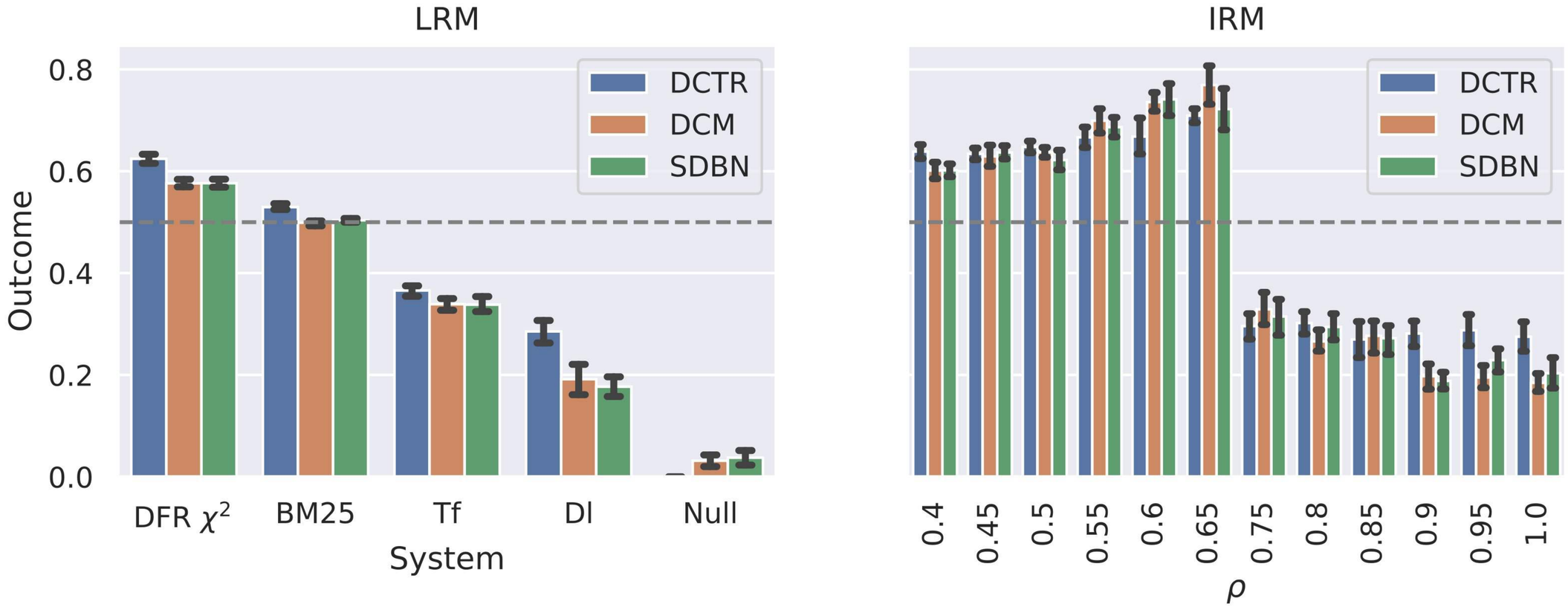
$$\text{IRM}_{\rho=0.4} \succ \text{IRM}_{\rho=0.45} \succ \cdots \succ \text{IRM}_{\rho=1.0}$$

$$\text{score}(d,q) = \rho \cdot \text{score}_{\text{Dl}}(d,q) + (1-\rho) \cdot \text{score}_{\text{DFR}}(d,q)$$

Three different **click models**

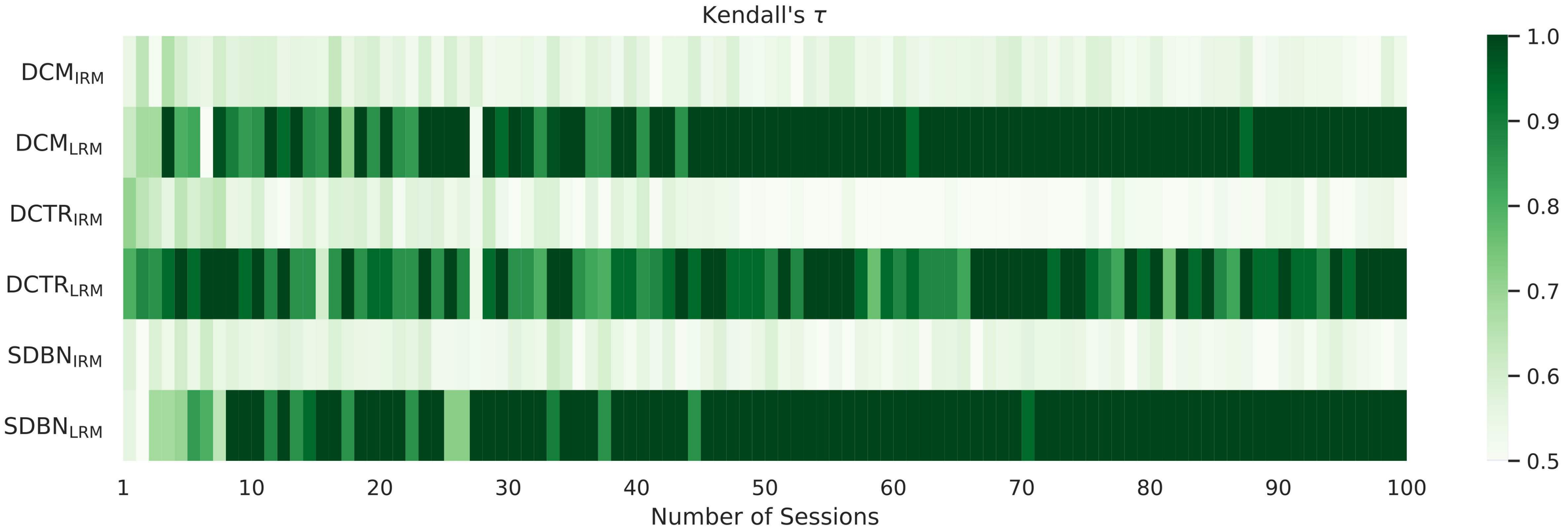
- **DCTR**: Document-based Click-Through Rate Model (based on **attractiveness assumption**)
- **DCM**: Dependent Click Model (**rank-based continuation probability**)
- **SDBN**: Simplified Dynamic Bayesian Network Model (**query-based satisfaction probability**)

Evaluations of Interleavings

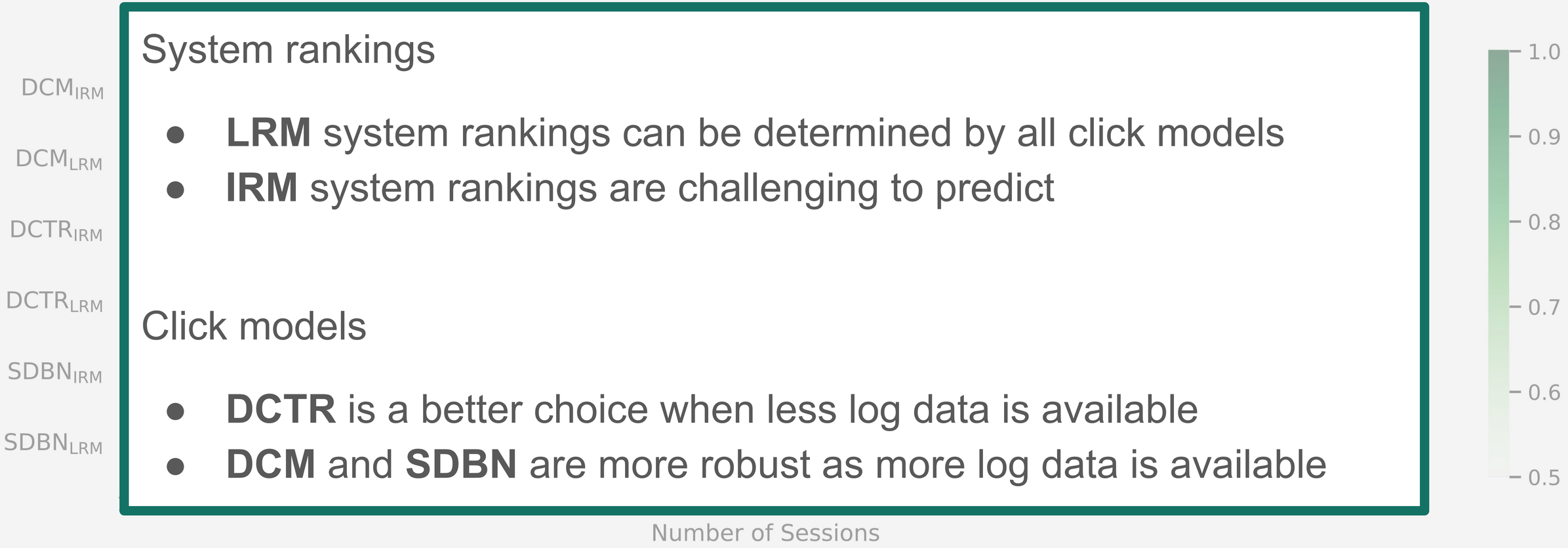


Test collection: TripClick https://tripdatabase.github.io/tripclick/; 50 queries, 100 sessions per query

Kendall’s tau of System Rankings vs. Sessions



Kendall’s tau of System Rankings vs. Sessions



System rankings

- **LRM** system rankings can be determined by all click models
- **IRM** system rankings are challenging to predict

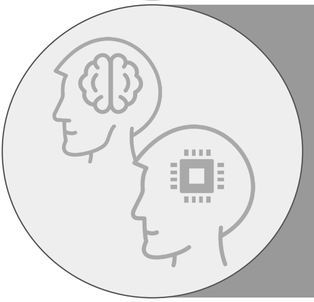
Click models

- **DCTR** is a better choice when less log data is available
- **DCM** and **SDBN** are more robust as more log data is available

Outline and Contributions

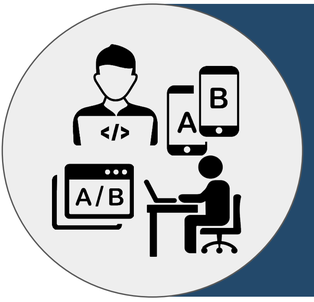
INTERNAL VALIDITY
system-oriented experiments

- PRIMAD extensions and metadata scheme
- Principled reproducibility evaluations



EXTERNAL VALIDITY
user simulations

- Query simulations and evaluation framework
- Click-based evaluations of system rankings



ECOLOGICAL VALIDITY
real user experiments

- Living lab infrastructure
- Shared task evaluations

Ecological Validation in Living Lab Environments

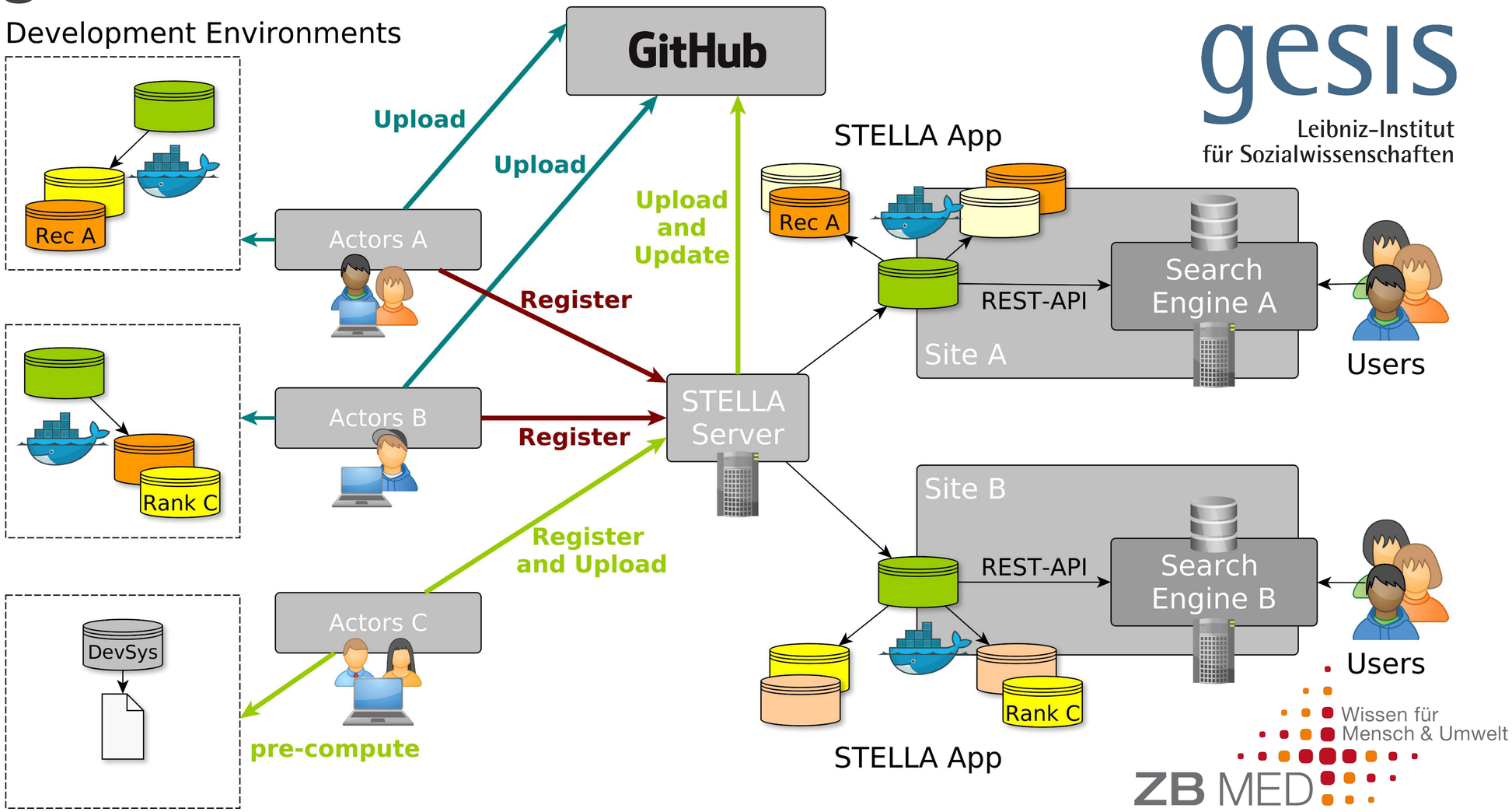
Living lab infrastructure

- **STELLA** - Infrastructures for Living Labs
- Docker-based **Evaluation-as-a-Service** platform

Shared task evaluations

- **CLEF‘21** “LiLAS - Living Labs for Academic Search”

Living Lab Infrastructure



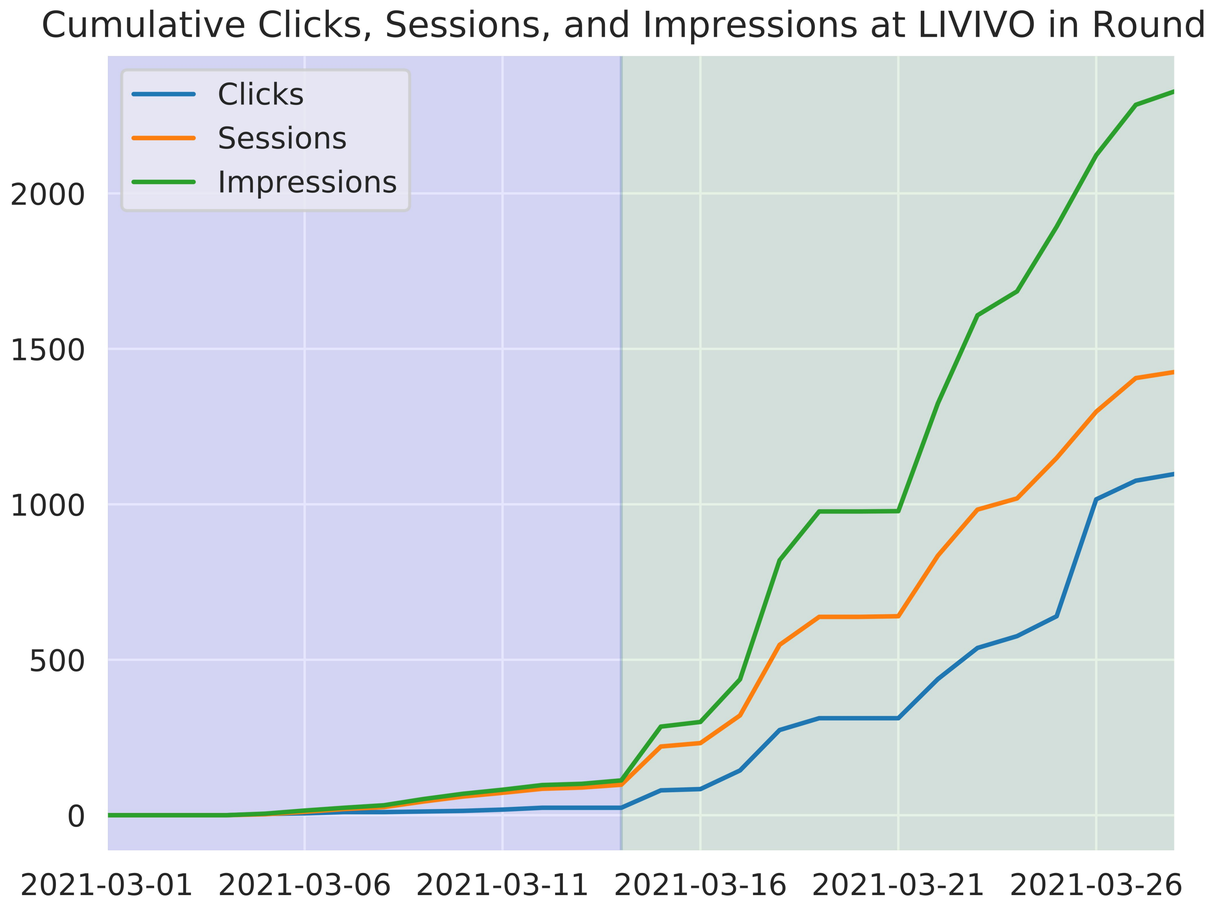
gesis
Leibniz-Institut für Sozialwissenschaften

Wissen für Mensch & Umwelt
ZB MED

***A Living Lab Architecture for Reproducible Shared Task Experimentation**; Breuer and Schaer; ISI 2021*

Pre-computed Results vs. Docker-based Systems

Pre-computed results
for top-k queries



Docker containers
with entire systems

***Overview of LiLAS 2021 - Living Labs for Academic Search**; Schaer, Breuer, Castro, Wolff, Schaible, Tavakolpoursaleh; CLEF 2021*

System Evaluations

- Evaluation based on clicks of **interleaving experiments**
- Relative user preferences determined by **wins, losses, ties**
- **GESIS recommendations** did not receive many clicks
- **LIVIVO baseline** (closed system) could not be outperformed

System	Win	Loss	Tie	Outcome	Sessions	Impressions	Clicks	CTR
Round 2								
$\text{GESIS}_{\text{Baseline}}$	51	68	2	0.43	3288	6034	53	0.0088
$\text{TFIDF}^{\text{Rec}}_{\text{Docker}}$	26	25	1	0.51	1529	2937	27	0.0092
$\text{BM25}^{\text{Rec}}_{\text{Precom}}$	42	26	1	0.62	1759	3097	45	0.0145
$\text{LIVIVO}_{\text{Baseline}}$	2447	1063	372	0.70	6481	12915	3791	0.2935
$\text{BM25}^{\text{Rank}}_{\text{Docker}}$	48	71	15	0.40	243	434	112	0.2581
$\text{DFR}^{\text{Rank}}_{\text{Docker}}$	707	1042	218	0.40*	3131	6274	1273	0.2029
$\text{DFR}^{+\text{Rank}}_{\dagger\text{Docker}}$	291	1308	135	0.18*	2948	6026	570	0.0946
$\text{LJM}^{\text{Rank}}_{\text{Precom}}$	6	13	0	0.32	61	69	10	0.1449
$\text{BM25}^{\text{Rank}\bullet}_{\text{Precom}}$	4	7	1	0.36	36	42	5	0.1190
$\text{BM25}^{\text{Rank}\blacklozenge}_{\text{Precom}}$	7	6	3	0.54	62	70	20	0.2857

***Overview of LiLAS 2021 - Living Labs for Academic Search**; Schaer, Breuer, Castro, Wolff, Schaible, Tavakolpoursaleh; CLEF 2021*

System Evaluations

- Evaluation based on clicks of **interleaving experiments**
- Relative user preferences determined by **wins, losses, ties**
- **GESIS recommendations** did not receive many clicks
- **LIVIVO baseline** (closed system) could not be outperformed

System	Win	Loss	Tie	Outcome	Sessions	Impressions	Clicks	CTR
Round 2								
$\text{GESIS}_{\text{Baseline}}$	51	68	2	0.43	3288	6034	53	0.0088
$\text{TFIDF}^{\text{Rec}}_{\text{Docker}}$	26	25	1	0.51	1529	2937	27	0.0092
$\text{BM25}^{\text{Rec}}_{\text{Precom}}$	42	26	1	0.62	1759	3097	45	0.0145
$\text{LIVIVO}_{\text{Baseline}}$	2447	1063	372	0.70	6481	12915	3791	0.2935
$\text{BM25}^{\text{Rank}}_{\text{Docker}}$	48	71	15	0.40	243	434	112	0.2581
$\text{DFR}^{\text{Rank}}_{\text{Docker}}$	707	1042	218	0.40*	3131	6274	1273	0.2029
$\text{DFR}^{+\text{Rank}}_{\text{Docker}}$	291	1308	135	0.18*	2948	6026	570	0.0946
$\text{LJM}^{\text{Rank}}_{\text{Precom}}$	6	13	0	0.32	61	69	10	0.1449
$\text{BM25}^{\text{Rank}\bullet}_{\text{Precom}}$	4	7	1	0.36	36	42	5	0.1190
$\text{BM25}^{\text{Rank}\blacklozenge}_{\text{Precom}}$	7	6	3	0.54	62	70	20	0.2857

***Overview of LiLAS 2021 - Living Labs for Academic Search**; Schaer, Breuer, Castro, Wolff, Schaible, Tavakolpoursaleh; CLEF 2021*

System Evaluations

- Evaluation based on clicks of **interleaving experiments**
- Relative user preferences determined by **wins, losses, ties**
- **GESIS recommendations** did not receive many clicks
- **LIVIVO baseline** (closed system) could not be outperformed

System	Win	Loss	Tie	Outcome	Sessions	Impressions	Clicks	CTR
Round 2								
$\text{GESIS}_{\text{Baseline}}$	51	68	2	0.43	3288	6034	53	0.0088
$\text{TFIDF}^{\text{Rec}}_{\text{Docker}}$	26	25	1	0.51	1529	2937	27	0.0092
$\text{BM25}^{\text{Rec}}_{\text{Precom}}$	42	26	1	0.62	1759	3097	45	0.0145
$\text{LIVIVO}_{\text{Baseline}}$	2447	1063	372	0.70	6481	12915	3791	0.2935
$\text{BM25}^{\text{Rank}}_{\text{Docker}}$	48	71	15	0.40	243	434	112	0.2581
$\text{DFR}^{\text{Rank}}_{\text{Docker}}$	707	1042	218	0.40*	3131	6274	1273	0.2029
$\text{DFR}^{+\text{Rank}}_{\dagger\text{Docker}}$	291	1308	135	0.18*	2948	6026	570	0.0946
$\text{LJM}^{\text{Rank}}_{\text{Precom}}$	6	13	0	0.32	61	69	10	0.1449
$\text{BM25}^{\text{Rank}\bullet}_{\text{Precom}}$	4	7	1	0.36	36	42	5	0.1190
$\text{BM25}^{\text{Rank}\blacklozenge}_{\text{Precom}}$	7	6	3	0.54	62	70	20	0.2857

***Overview of LiLAS 2021 - Living Labs for Academic Search**; Schaer, Breuer, Castro, Wolff, Schaible, Tavakolpoursaleh; CLEF 2021*

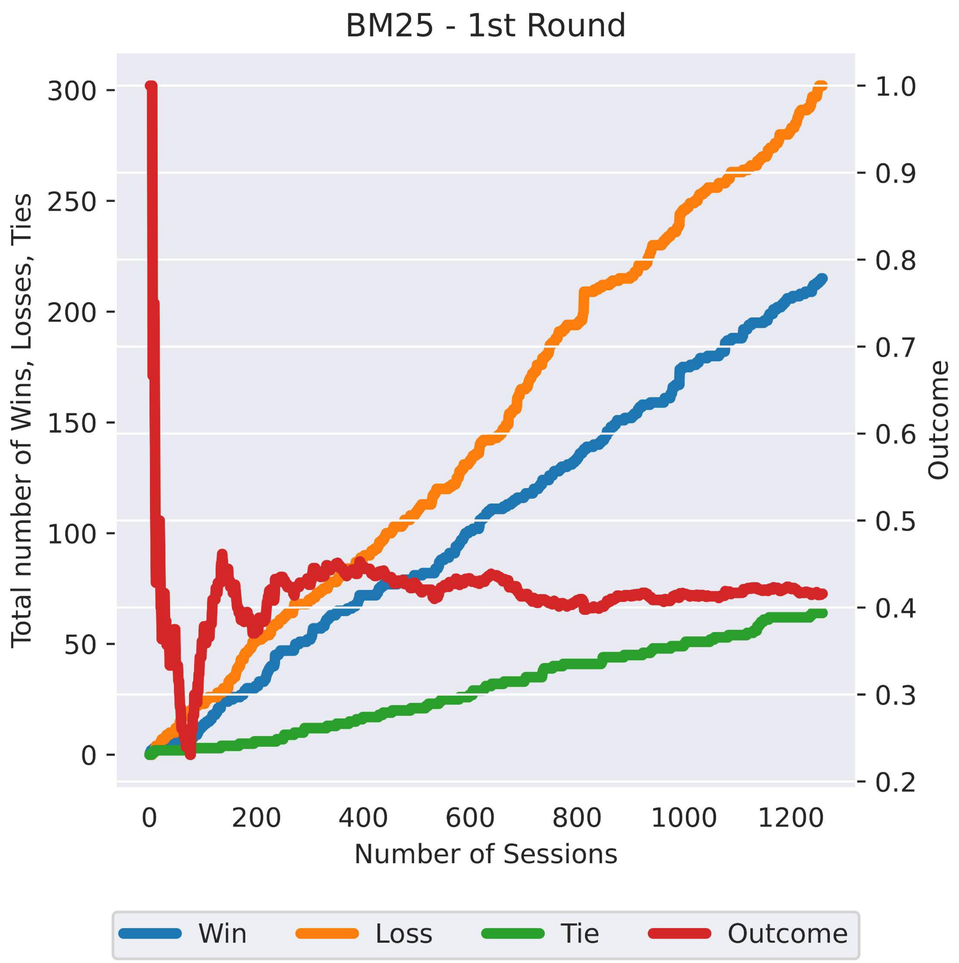
System Evaluations

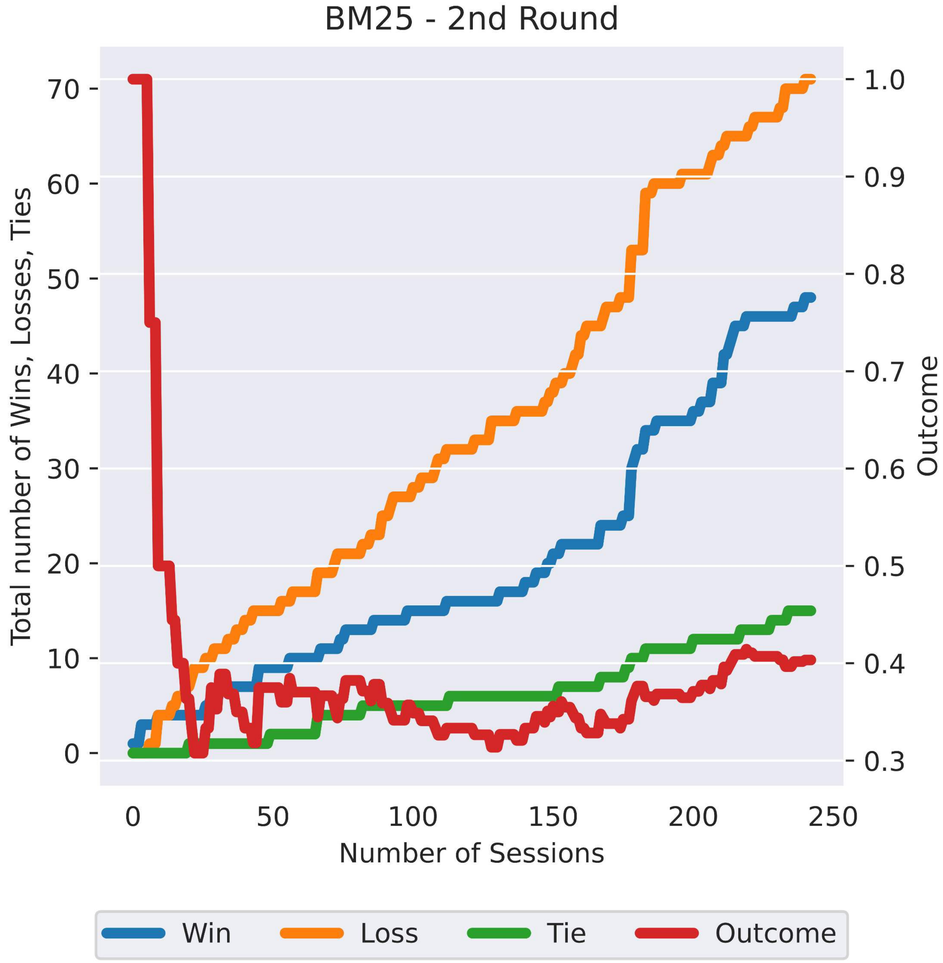
- Evaluation based on clicks of **interleaving experiments**
- Relative user preferences determined by **wins, losses, ties**
- **GESIS recommendations** did not receive many clicks
- **LIVIVO baseline** (closed system) could not be outperformed

System	Win	Loss	Tie	Outcome	Sessions	Impressions	Clicks	CTR
Round 2								
$\text{GESIS}_{\text{Baseline}}$	51	68	2	0.43	3288	6034	53	0.0088
$\text{TFIDF}^{\text{Rec}}_{\text{Docker}}$	26	25	1	0.51	1529	2937	27	0.0092
$\text{BM25}^{\text{Rec}}_{\text{Precom}}$	42	26	1	0.62	1759	3097	45	0.0145
$\text{LIVIVO}_{\text{Baseline}}$	2447	1063	372	0.70	6481	12915	3791	0.2935
$\text{BM25}^{\text{Rank}}_{\text{Docker}}$	48	71	15	0.40	243	434	112	0.2581
$\text{DFR}^{\text{Rank}}_{\text{Docker}}$	707	1042	218	0.40*	3131	6274	1273	0.2029
$\text{DFR}^{+\text{Rank}}_{\text{Docker}}$	291	1308	135	0.18*	2948	6026	570	0.0946
$\text{LJM}^{\text{Rank}}_{\text{Precom}}$	6	13	0	0.32	61	69	10	0.1449
$\text{BM25}^{\text{Rank}\bullet}_{\text{Precom}}$	4	7	1	0.36	36	42	5	0.1190
$\text{BM25}^{\text{Rank}\blacklozenge}_{\text{Precom}}$	7	6	3	0.54	62	70	20	0.2857

***Overview of LiLAS 2021 - Living Labs for Academic Search**; Schaer, Breuer, Castro, Wolff, Schaible, Tavakolpoursaleh; CLEF 2021*

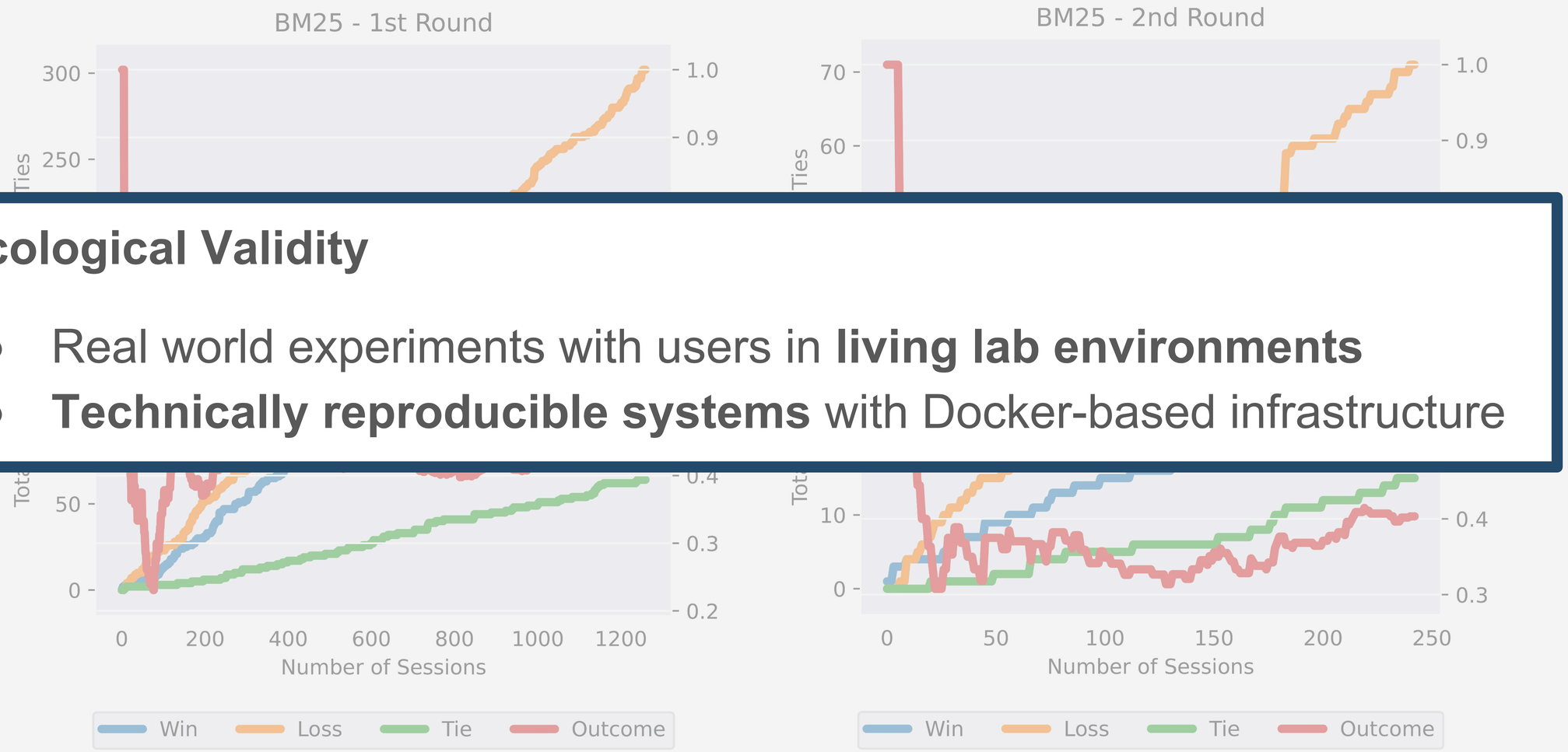
Reproducibility of Outcomes





***Overview of LiLAS 2021 - Living Labs for Academic Search**; Schaer, Breuer, Castro, Wolff, Schaible, Tavakolpoursaleh; CLEF 2021*

Reproducibility of Outcomes



Ecological Validity

- Real world experiments with users in **living lab environments**
- **Technically reproducible systems** with Docker-based infrastructure

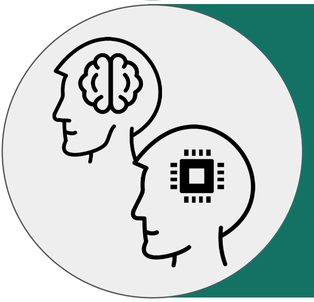
Overview of LiLAS 2021 - Living Labs for Academic Search; Schaer, Breuer, Castro, Wolff, Schaible, Tavakolpoursaleh; CLEF 2021

Future Work

INTERNAL VALIDITY

system-oriented experiments

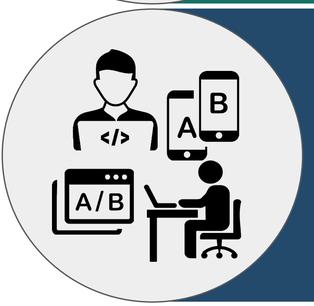
- Extension of the metadata scheme
- Evaluation of the reproducibility measures



EXTERNAL VALIDITY

user simulations

- Simulation of other user interactions
- Validation of user interaction sequences



ECOLOGICAL VALIDITY

real user experiments

- Development of a test collection
- Integrated evaluation life cycles

Norbert Fuhr - Philipp Schaer
Matthias Hagen - Maria Maistro
Nicola Ferro - Leyla Jael Castro
Daniel Hienert - Johann Schaible
Narges Tavakolpoursaleh - Benjamin Wolff
Zeljko Carevic - Jüri Keller - Anh Huy Tran
Melanie Pest - Dirk Tunger - Fabian Haak
Björn Engelmann - Christin Kreutz
Sven Wöhrle - Narjes Nikzad Khasmakhi
Malte Bonart - Mandy Neumann

Thank you!