

# Replicability Measures for Longitudinal Information Retrieval Evaluation

**CLEF 2024 – Best of LongEval Lab**  
Jüri Keller, Timo Breuer, Philipp Schaer

24-12-11 – Grenoble, France  
<https://ir.web.th-koeln.de>



# The LongEval Lab

- *Longitudinal Evaluation of Model Performance*
- Two tasks: retrieval and classification
- Classic web search
- Two languages: French and English
- Over time!

## LongEval CLEF 2024 Lab



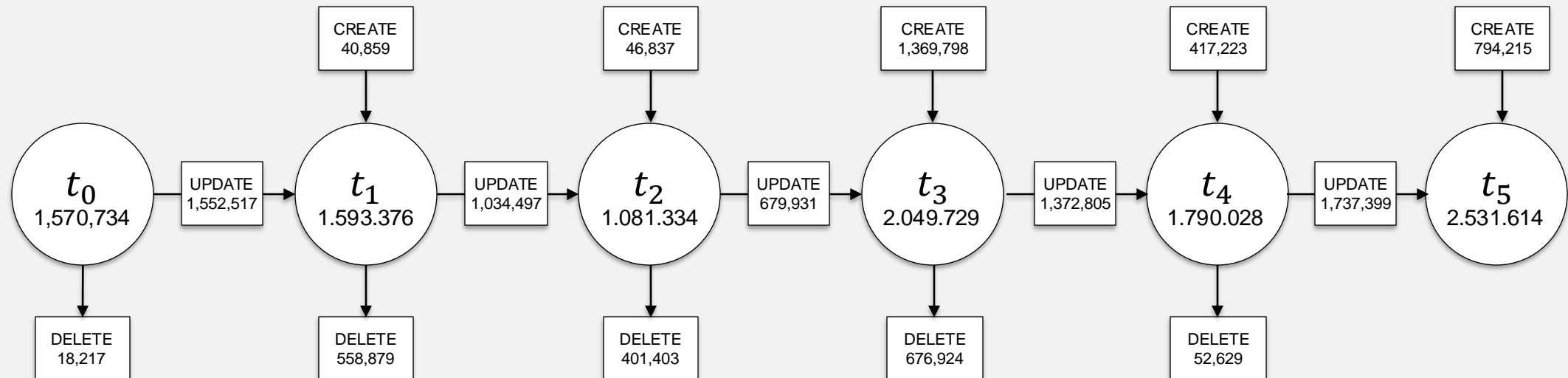
## LongEval CLEF 2023 Lab



[Description](#) [Dates](#) [Organizers](#) [Tasks](#) [Data](#) [Submissions](#)

# Introduction

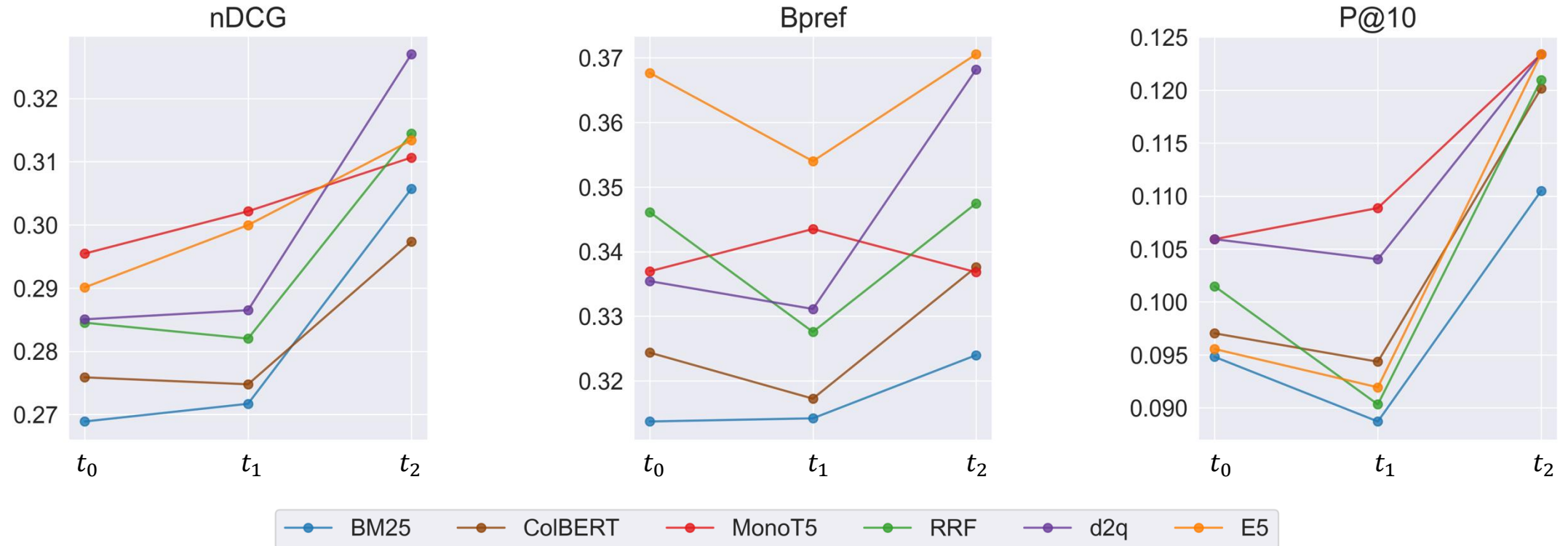
- Six sub-collections
- That evolved over time
- With overlap



# Problem

- IR systems are exposed to constant change
- Conventional evaluations abstract these changes
- Results and effectiveness changes
- **No direct comparison is possible**
- *How can we compare the effectiveness across time?*

# Problem



- Effectiveness changes over time
- The ranking of systems changes as well

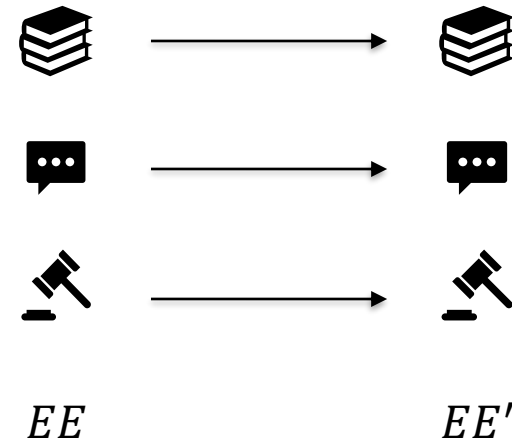
# Replicability

- Investigate temporal change as a replicability problem
- ACM: Same systems but different experimental setup

## Fixed systems

- BM25
- + CoBERT
- + monoT5
- + D2Q
- RRF
- E5




## Evolved test collection



# Approach

**Differentiate between changes**




**Comparison Strategy**

	CREATE	UPDATE	DELETE
	Extension of document collection	Document content changed (e.g., online news articles, or websites)	Documents removed (e.g., due to licensing issues)
	New queries / topics (like current topics of interest)	Changed (head) queries from user logs (e.g., changed popularity)	Removed topics (due to missing interest or inappropriateness)
	Added new relevance labels (from old or new assessors)	Assessors changed their mind; new judgment guidelines	Relevance labels removed (due to low inter-rater agreement)

# Approach

**Differentiate between changes**

**Comparison Strategy**

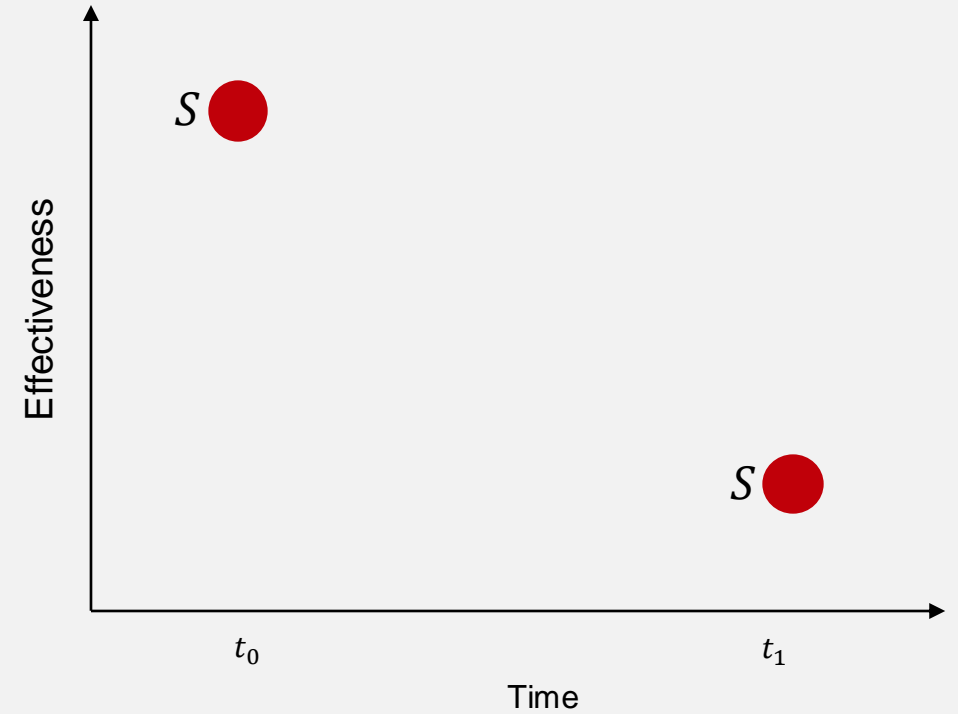
	CREATE	UPDATE	DELETE
	Extension of document collection	Document content changed (e.g., online news articles, or websites)	Documents removed (e.g., due to licensing issues)
	New queries / topics (like current topics of interest)	Changed (head) queries from user logs (e.g., changed popularity)	Removed topics (due to missing interest or inappropriateness)
	Added new relevance labels (from old or new assessors)	Assessors changed their mind; new judgment guidelines	Relevance labels removed (due to low inter-rater agreement)



# Approach

**Differentiate between changes**

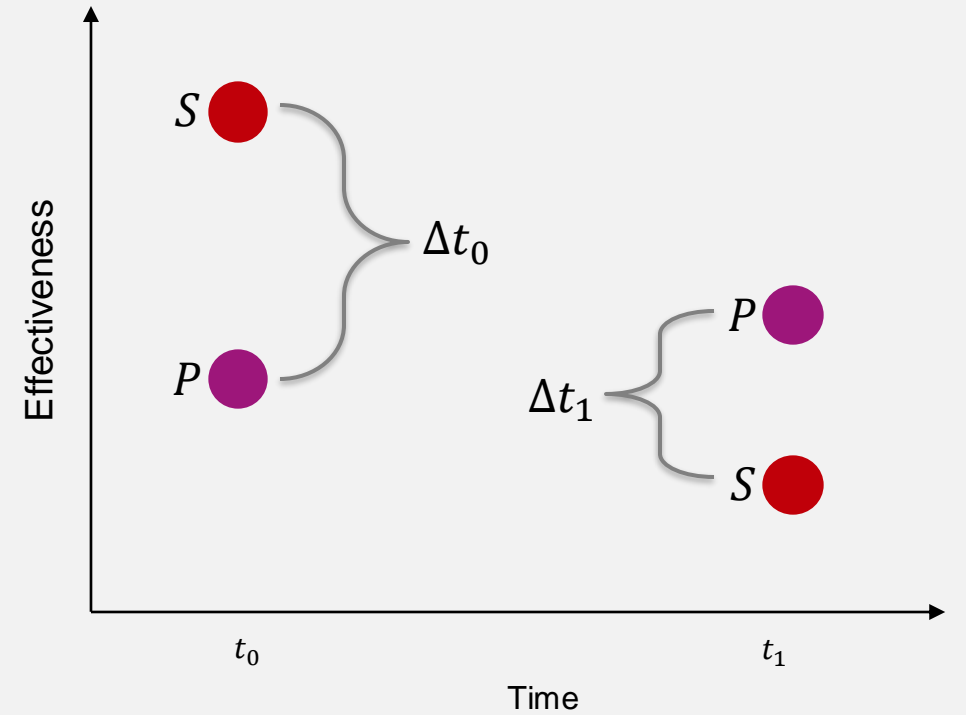
**Comparison Strategy**



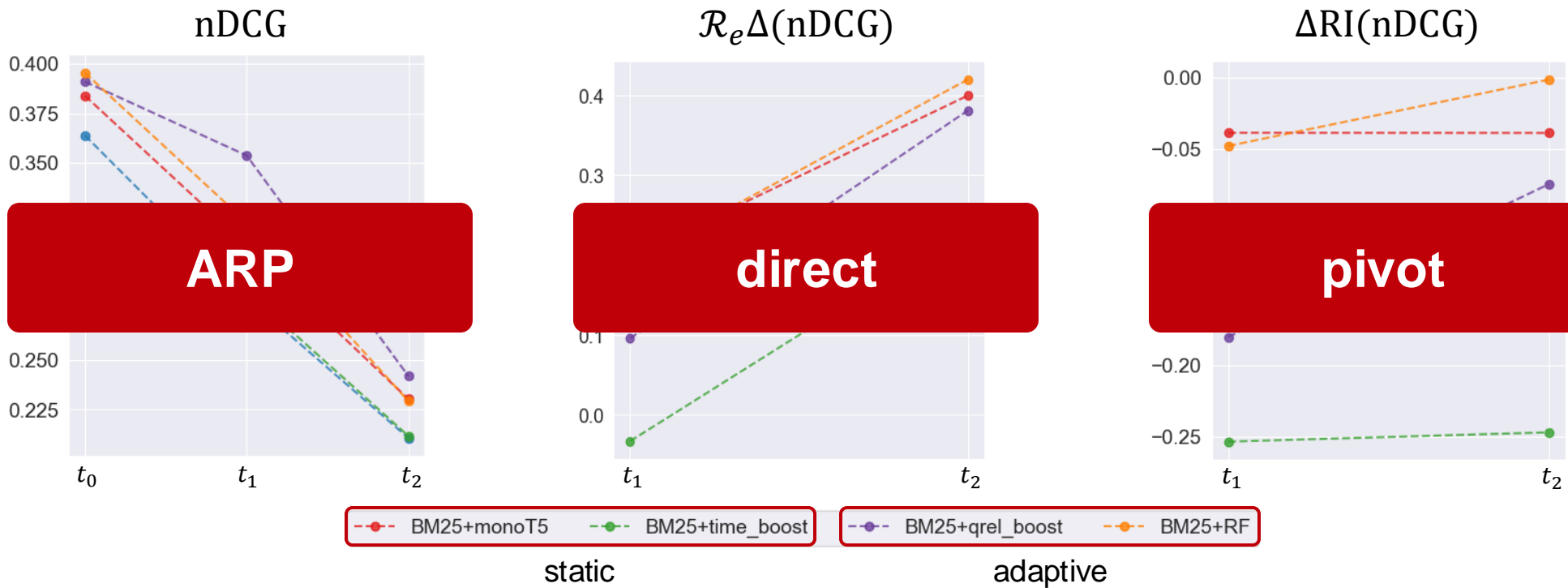
# Approach

Differentiate between changes

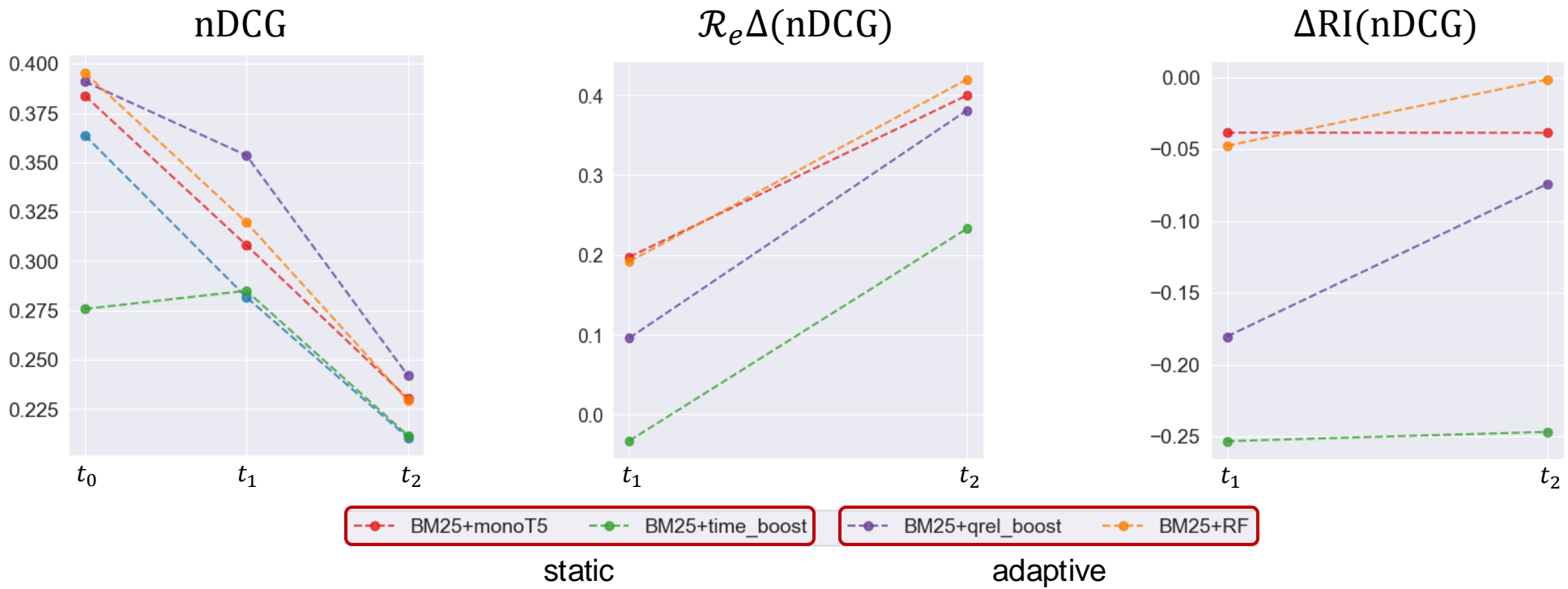
Comparison Strategy



# Results



# Results



# Discussion

- Comparing results over time is more difficult than expected
  - Attribution is unclear, direct comparison not necessarily possible
- Comparison strategy is needed
- Changes overlap, isolation is difficult
- Only little agreement across:
  - Topics, time, measure, *robustness*



# Works

## ICTIR: *Evaluation of Temporal Change in IR Test Collections*

- Different retrieval scenarios
- More test collections
- More measures beyond the effectiveness



CEUR  
Workshop  
Proceedings

ceur-ws.org  
ISBN 1613-073

## LongEval: *Leveraging Prior Relevance Signals in Web Search*

- Exploit old relevance labels to boost effectiveness

# Conclusion

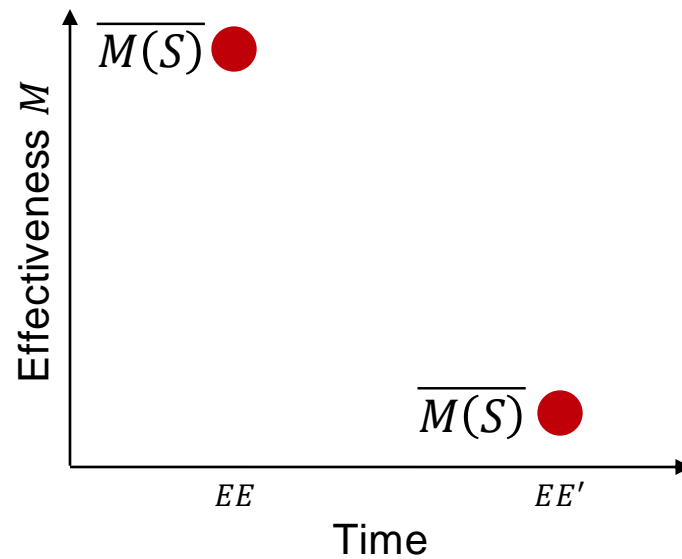
- The experimental setup strongly influences the result
- Effectiveness depend on the point in time
- We can not directly compare evaluation results across time
  - $\mathcal{R}_e\Delta$  extracts the influence of the experimental setup
  - $\Delta RI$  and  $ER$  extract the influence of the system

# Thank You!



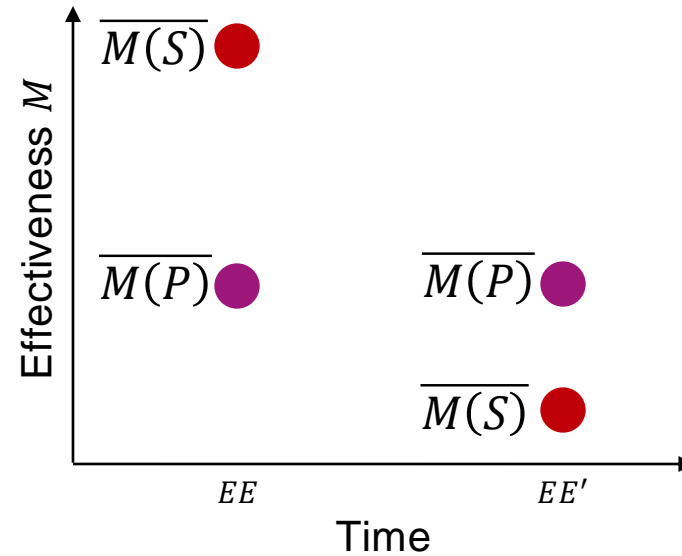


# Result Delta $\mathcal{R}_e\Delta$



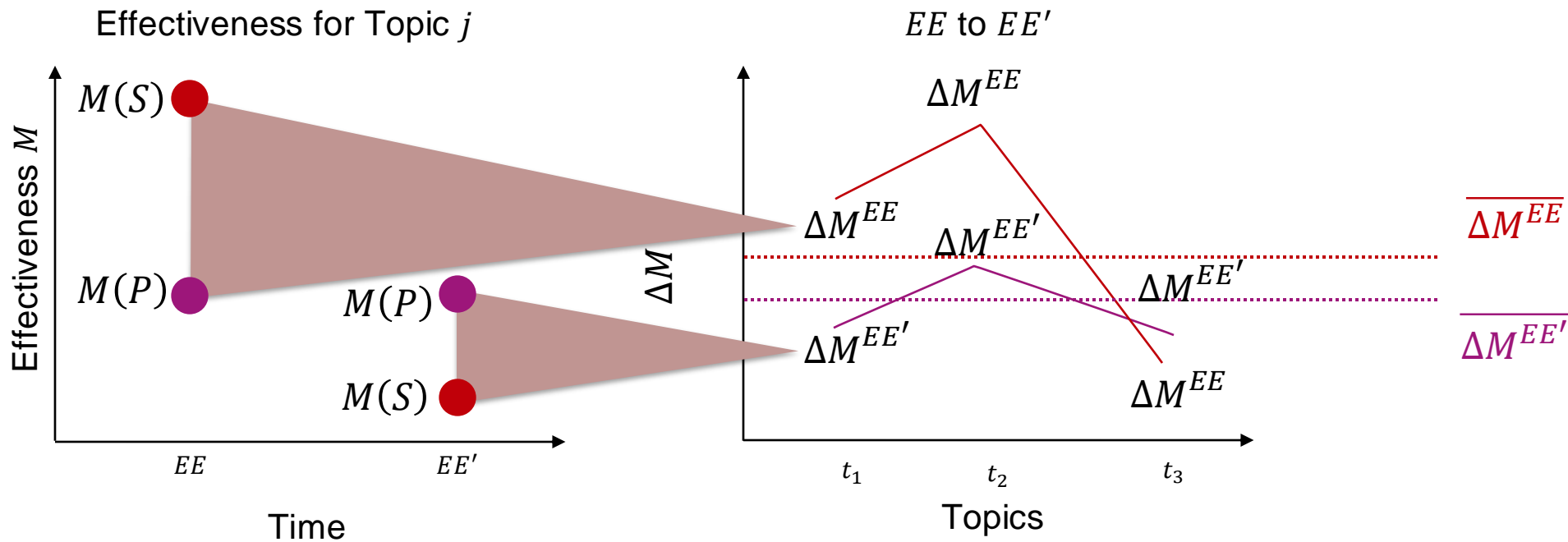
$$\mathcal{R}_e\Delta = \frac{\overline{M^{EE}(S)} - \overline{M^{EE'}(S)}}{\overline{M^{EE'}(S)}}$$

# Delta Relative Improvement $\Delta RI$



$$RI = \frac{\overline{M^{EE}(S)} - \overline{M^{EE}(P)}}{\overline{M^{EE}(P)}}, \quad RI' = \frac{\overline{M^{EE'}(S)} - \overline{M^{EE'}(P)}}{\overline{M^{EE'}(P)}}, \quad \Delta RI = RI - RI'$$

# Effect Ratio $ER$



$$\Delta M_j^{EE} = M_j^{EE}(S) - M_j^{EE}(P)$$

$$\Delta M_j^{EE'} = M_j^{EE'}(S) - M_j^{EE'}(P)$$

$$ER(\Delta M_j^{EE'}, \Delta M_j^{EE}) = \frac{\overline{\Delta M_j^{EE'}}}{\overline{\Delta M_j^{EE}}} = \frac{\frac{1}{n^{EE'}} \sum_{j=1}^{n^{EE'}} \Delta M_j^{EE'}}{\frac{1}{n^{EE}} \sum_{j=1}^{n^{EE}} \Delta M_j^{EE}}$$