# Corpora Performance Prediction

## ECIR 2025 QPP++

**Andrew Parry[1], Jan Heinrich Merker[2], Simon Ruth[3], Maik Fröbe[2] and Harrisen Scells[4]**

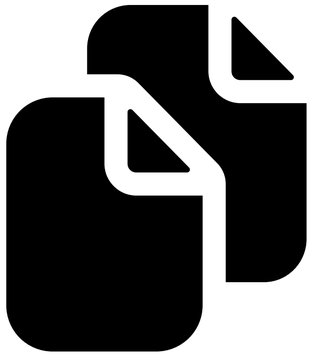
**[1]University of Glasgow, [2]Friedrich-Schiller-Universität Jena, [3]Universität Kassel, [4]University of Tübingen**

Most Likely

Least Likely

**Effectiveness**

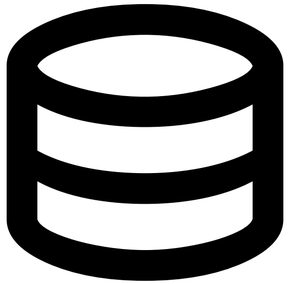





Most Likely

Least Likely

**Effectiveness**



Most Likely

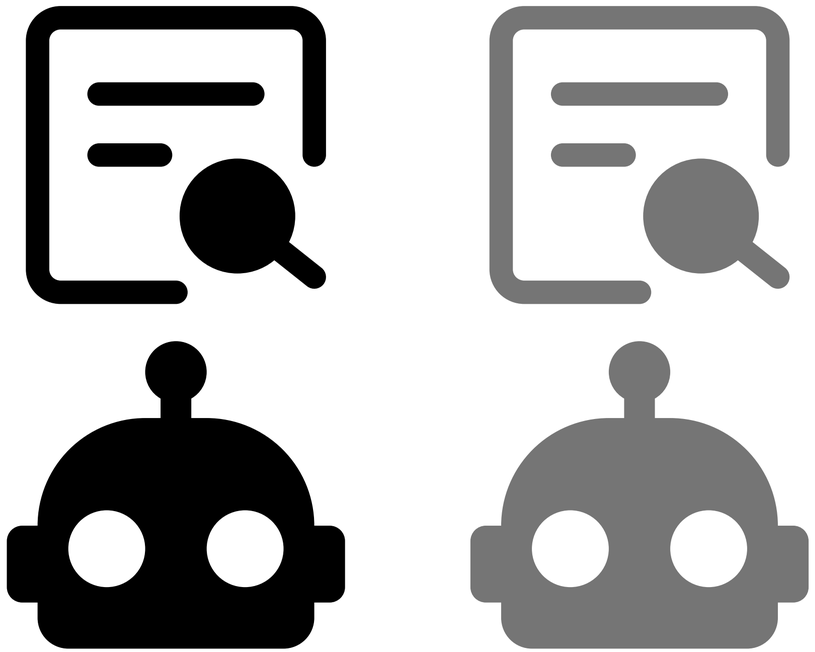
Least Likely

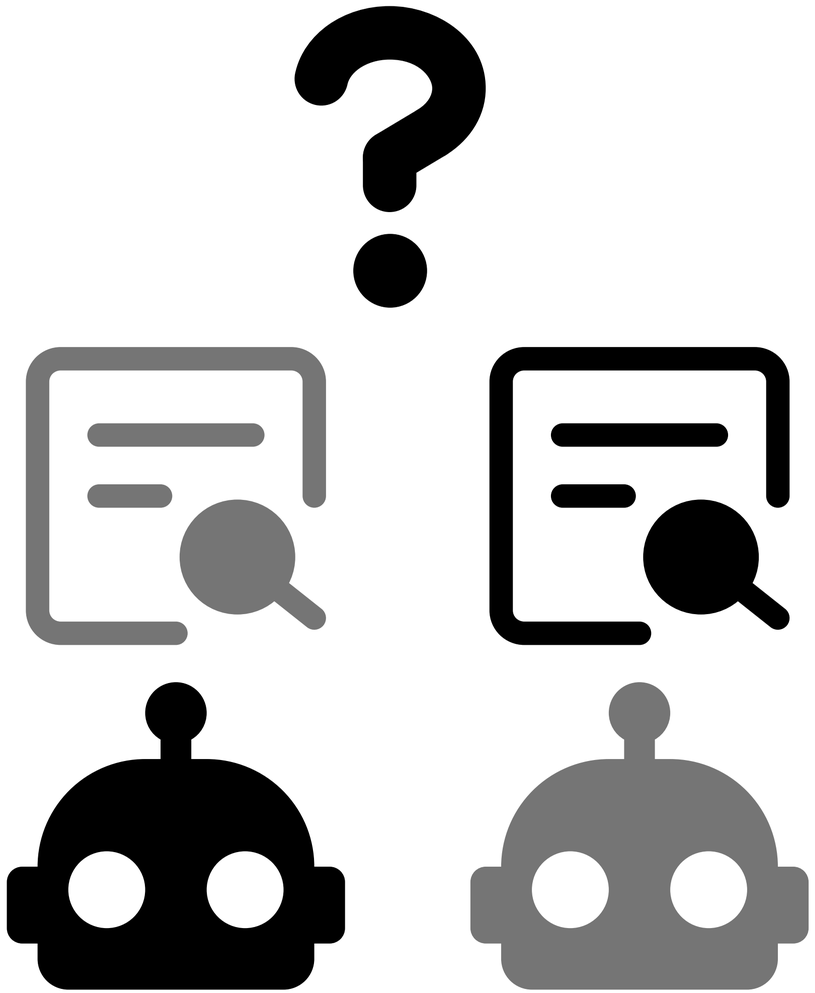
**Effectiveness**

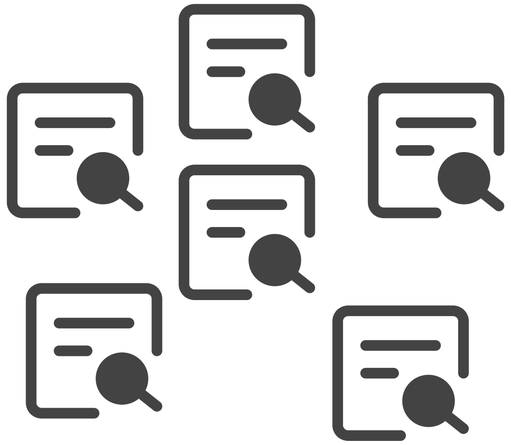
Most Likely

Least Likely

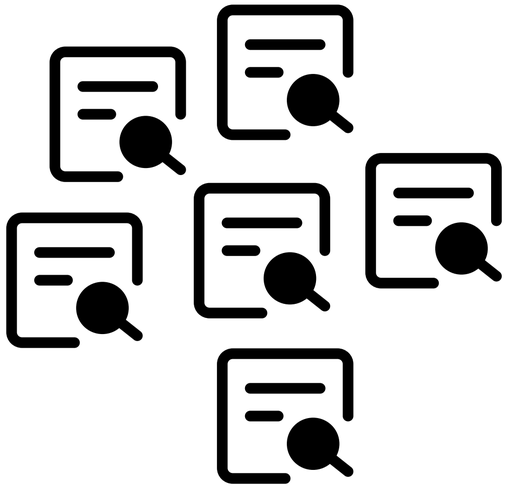
**Effectiveness**



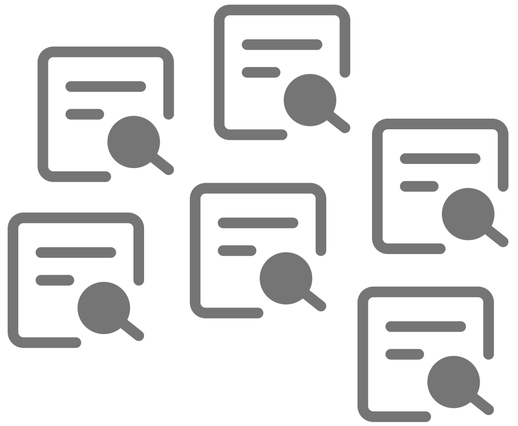




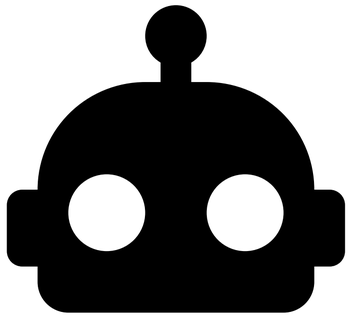
Medical Queries

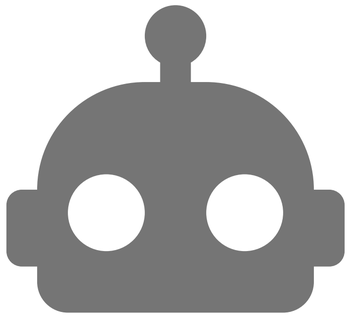


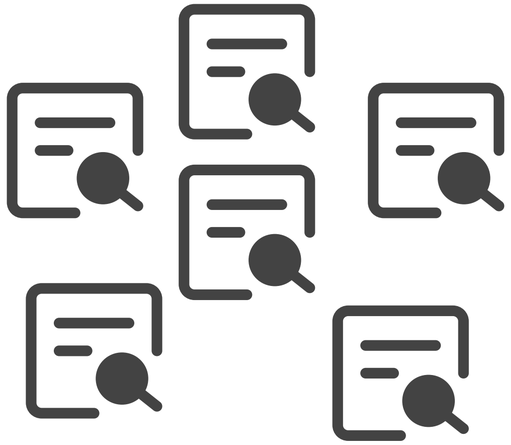
Arbitrary Questions



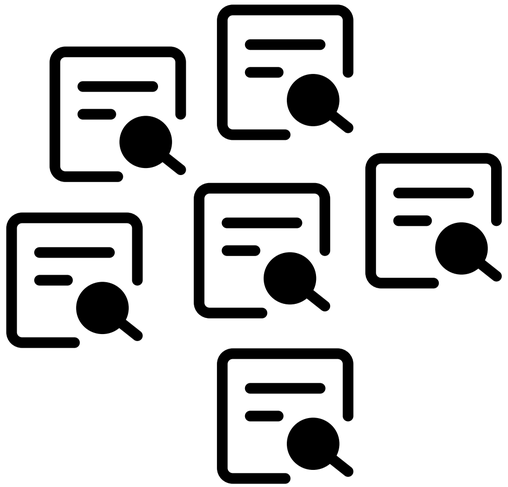
News Queries



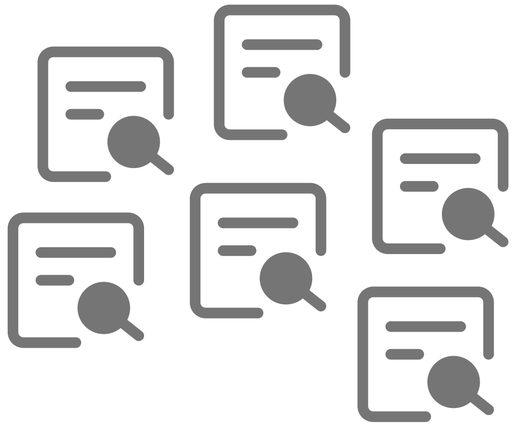




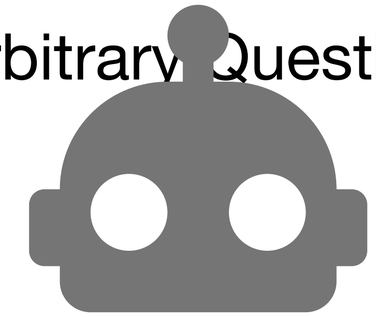
Medical Queries

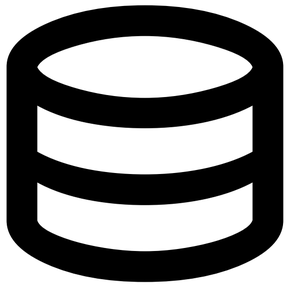


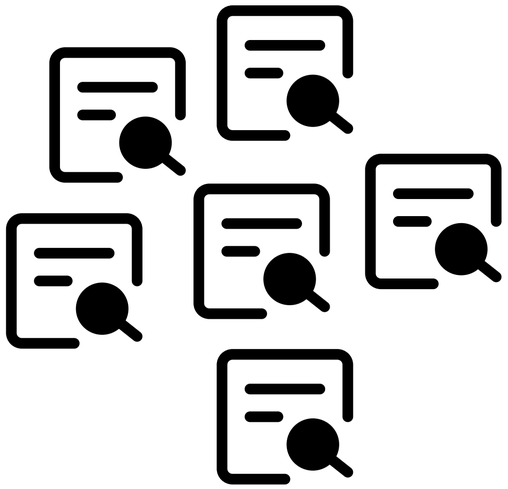
Arbitrary Questions



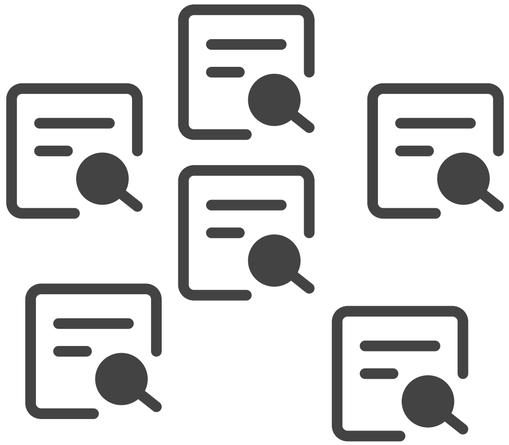
News Queries



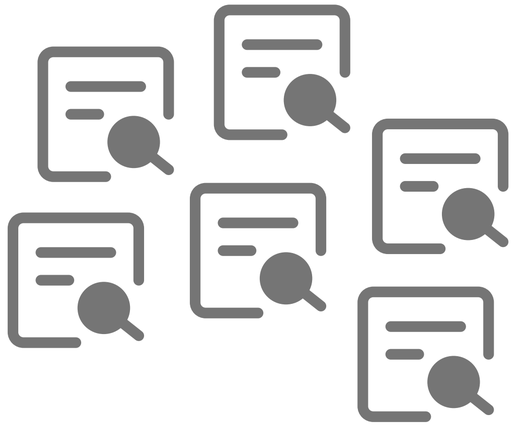




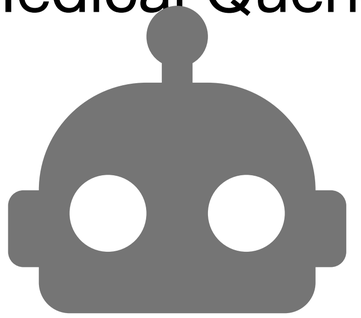
Arbitrary Questions

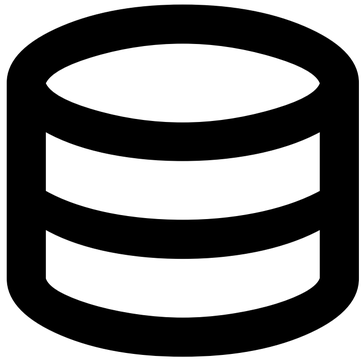


Medical Queries



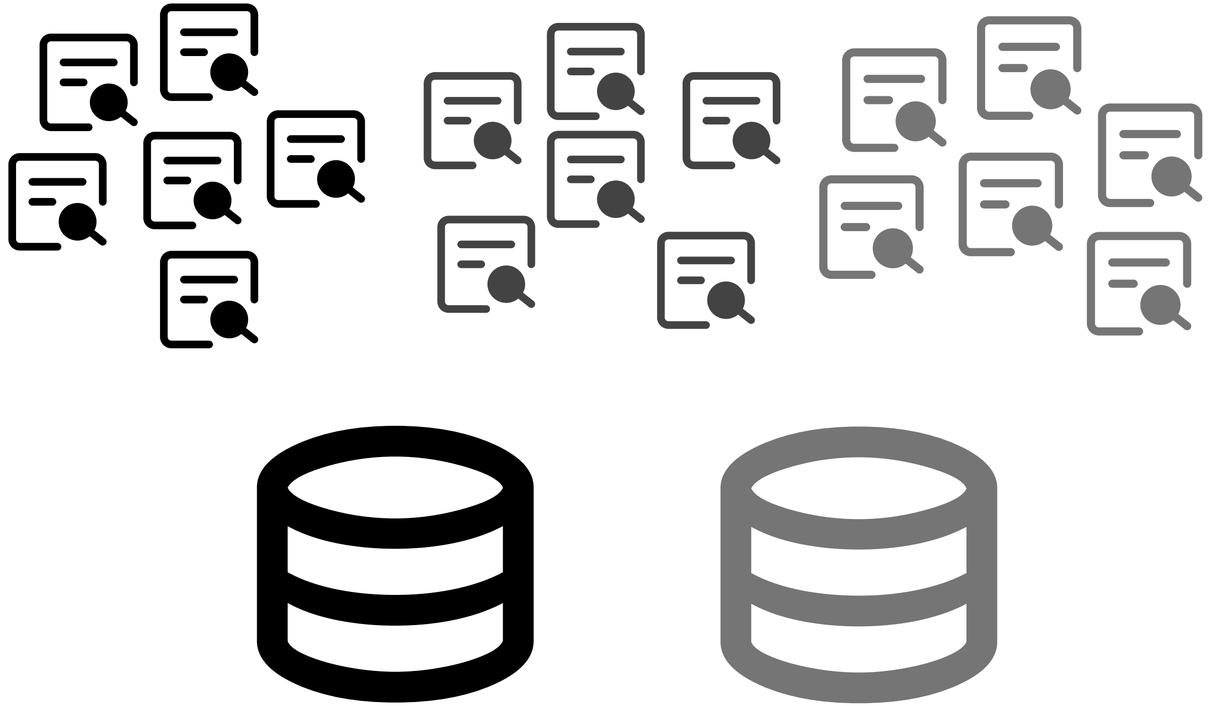
News Queries





# Corpora Performance Prediction

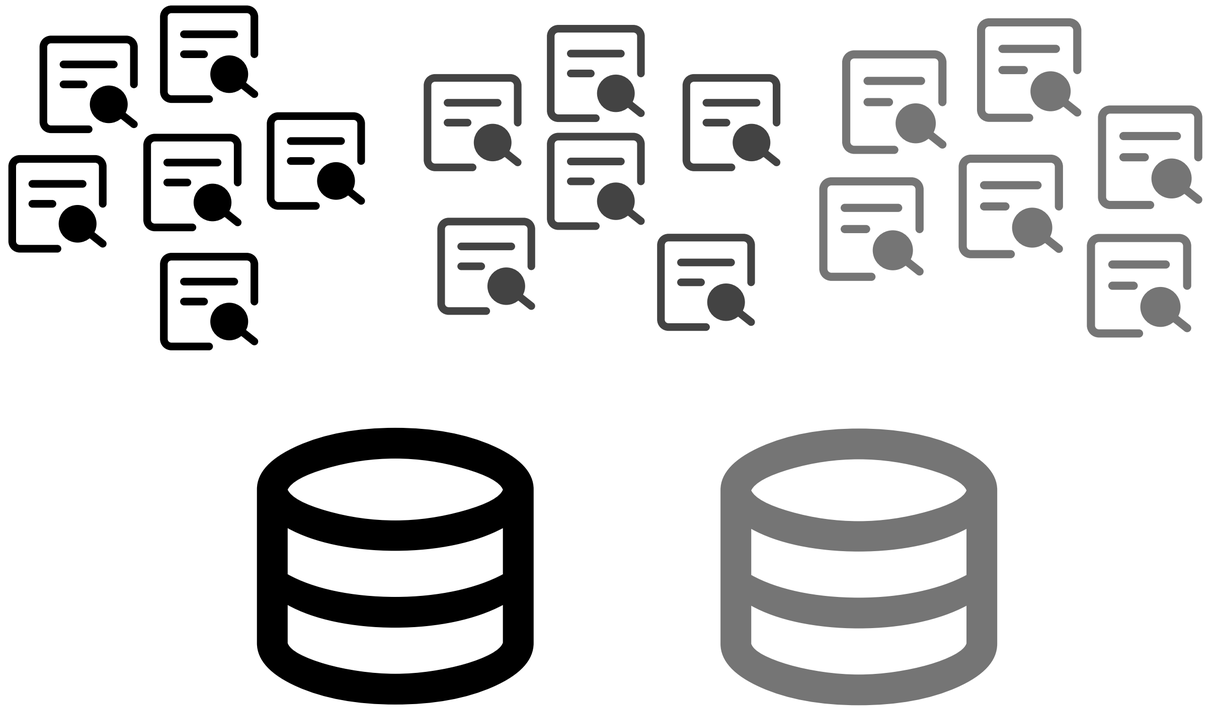
## Motivation



# Corpora Performance Prediction

## Motivation

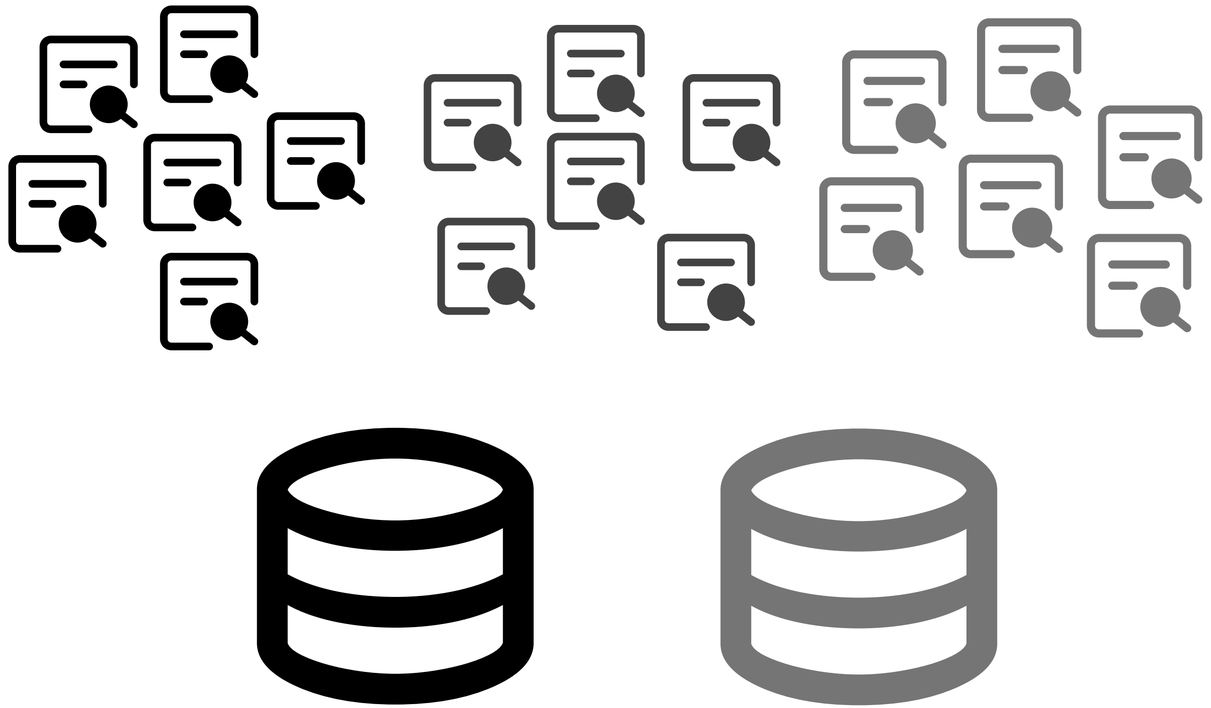
- Holistic evaluation of corpora



# Corpora Performance Prediction

## Motivation

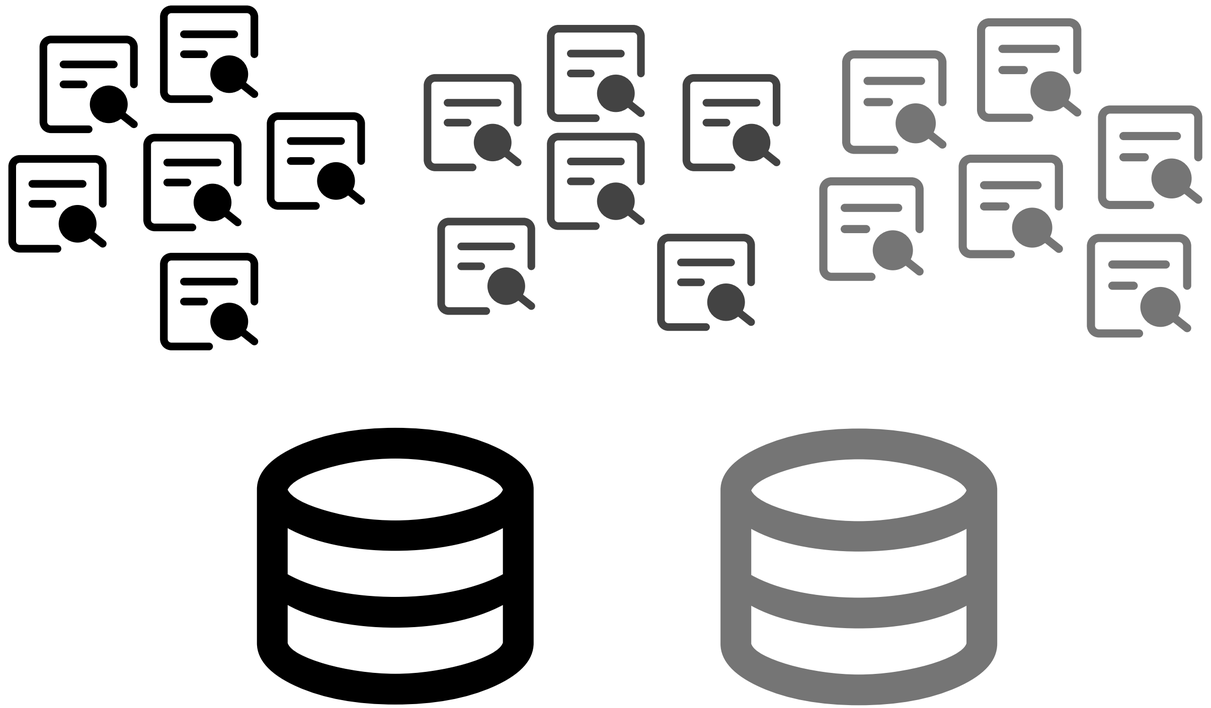
- Holistic evaluation of corpora
- Lightweight heuristics



# Corpora Performance Prediction

## Motivation

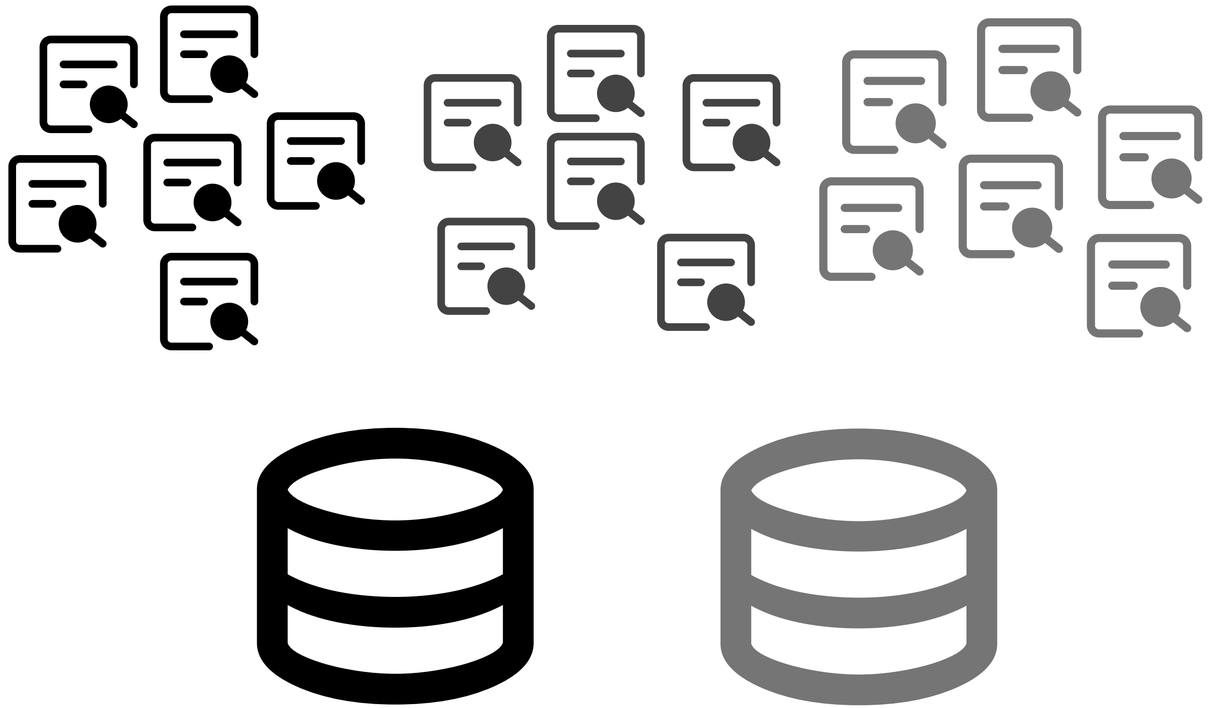
- Holistic evaluation of corpora
- Lightweight heuristics
- Comparisons in terms of the ability to *serve queries*



# Corpora Performance Prediction

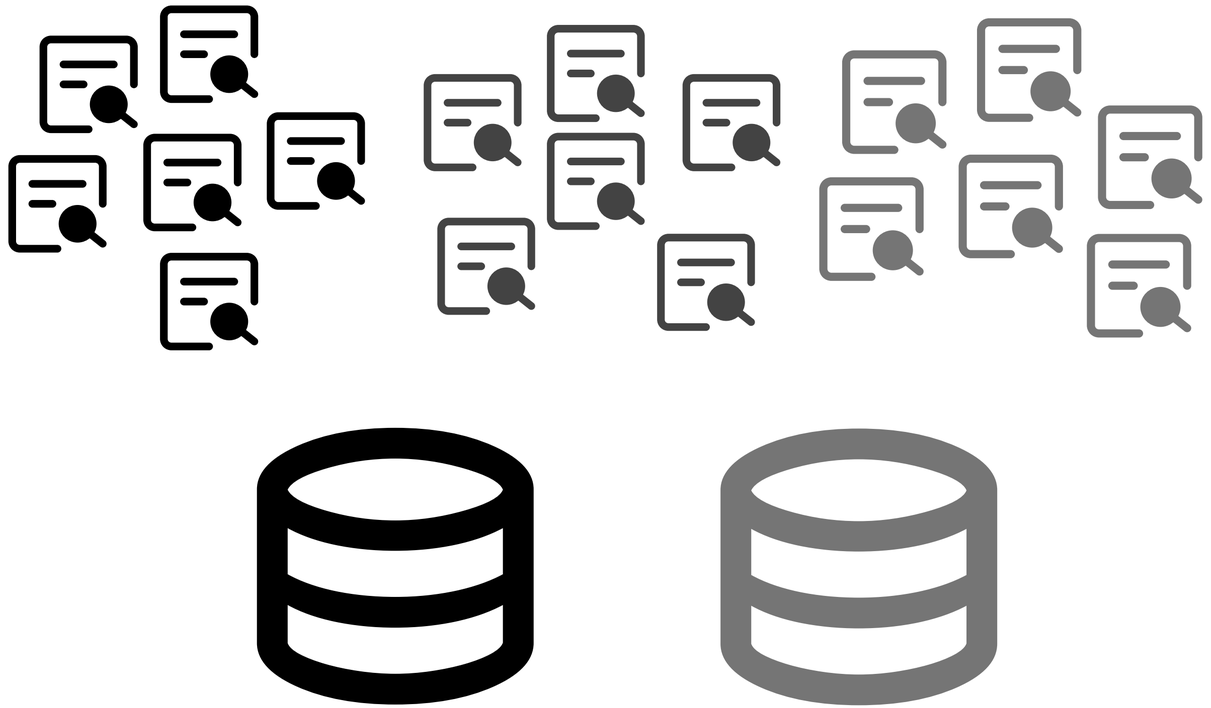
## Motivation

- Holistic evaluation of corpora
- Lightweight heuristics
- Comparisons in terms of the ability to *serve queries*
- Linked to retrievability



# Corpora Performance Prediction

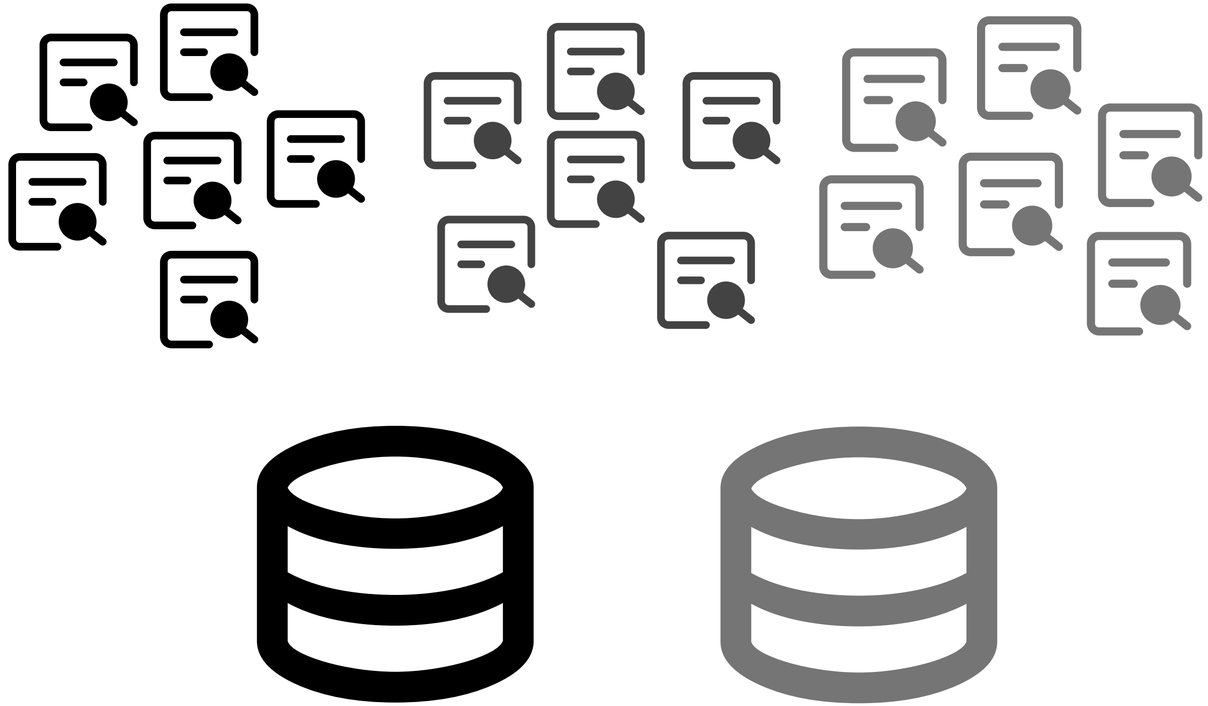
## Motivation



- Holistic evaluation of corpora
- Lightweight heuristics
- Comparisons in terms of the ability to *serve queries*
- Linked to retrievability
- Provides additional applications of QPP over *multiple queries*

# Corpora Performance Prediction

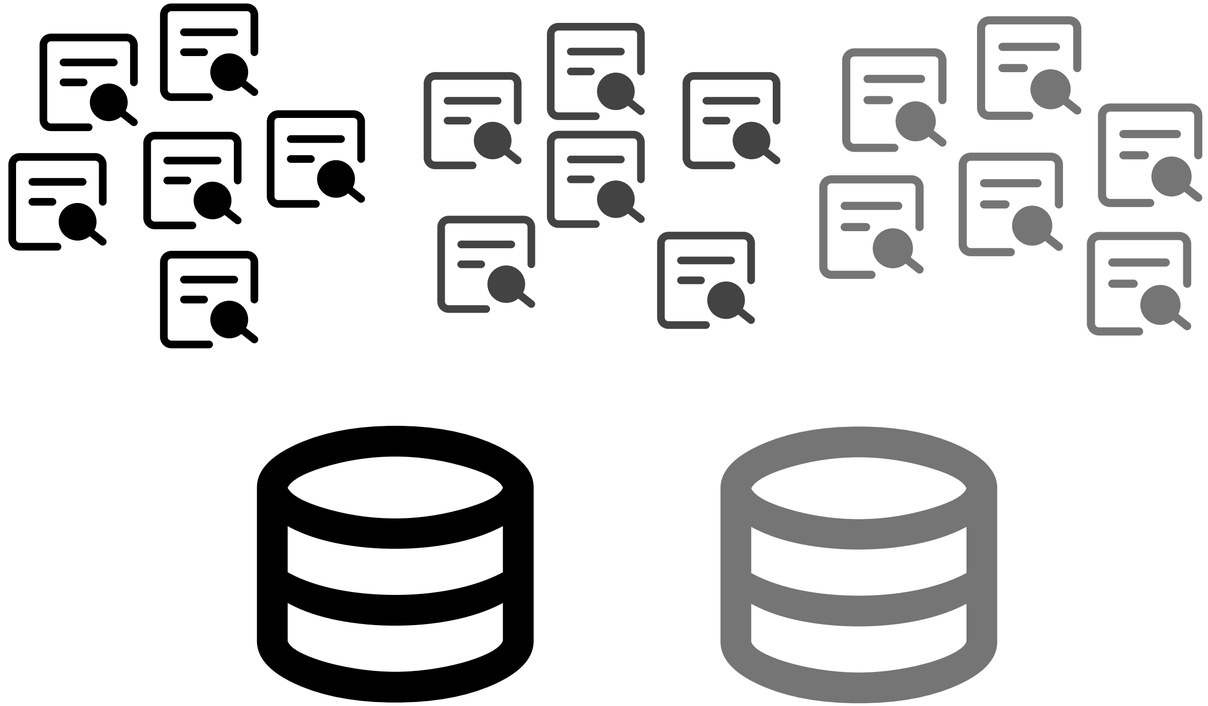
## Our Approach



# Corpora Performance Prediction

## Our Approach

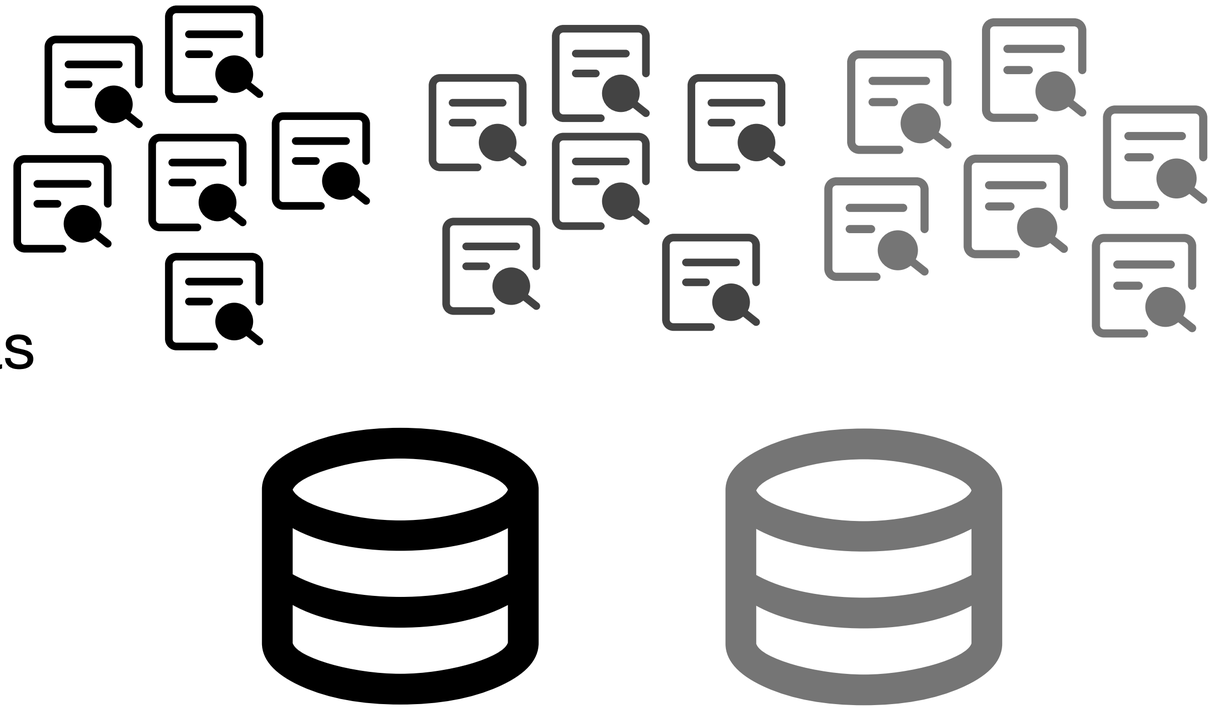
- Use QPP heuristics over *domains*



# Corpora Performance Prediction

## Our Approach

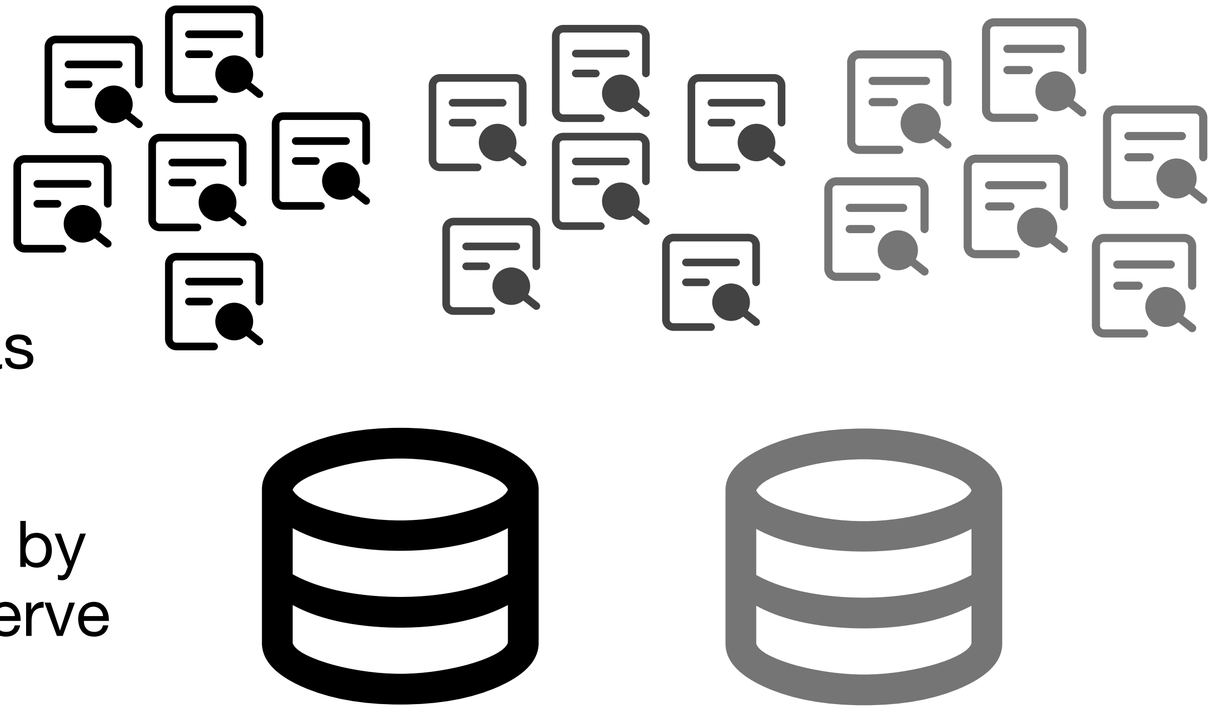
- Use QPP heuristics over *domains*
- *Aggregate* QPP measures are taken as CPP measures



# Corpora Performance Prediction

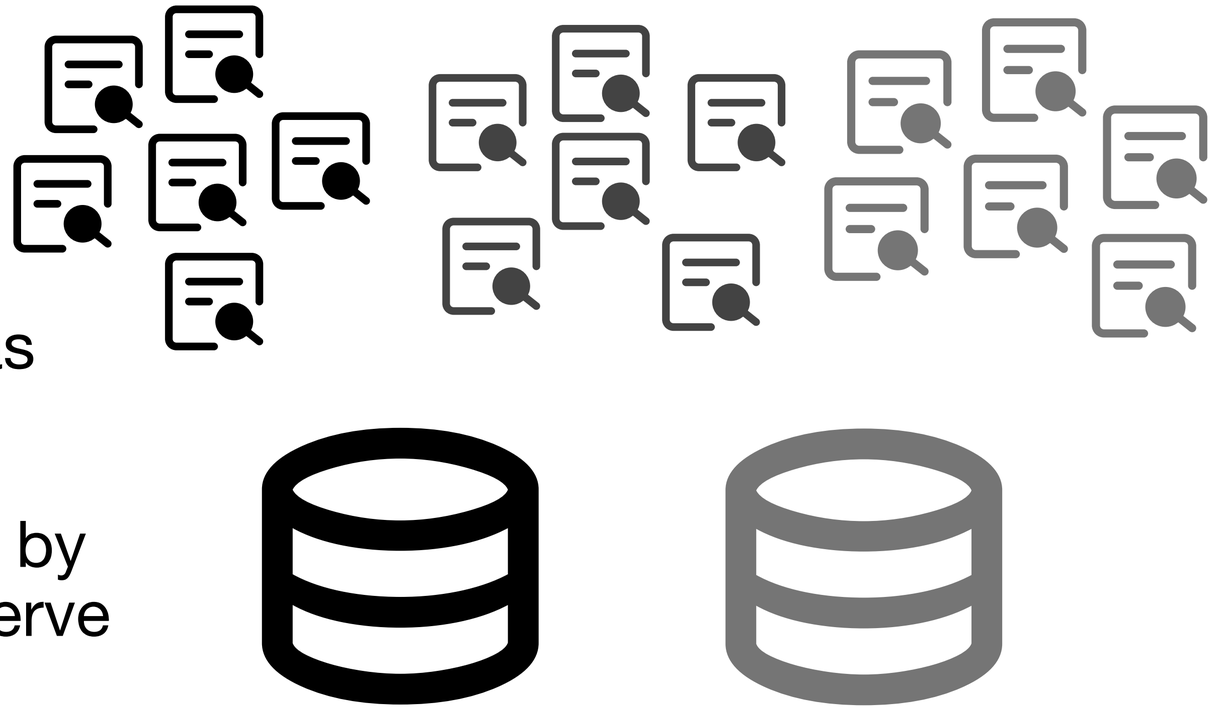
## Our Approach

- Use QPP heuristics over *domains*
- *Aggregate* QPP measures are taken as CPP measures
- In doing so we can compare corpora by the domains they are best suited to serve



# Corpora Performance Prediction

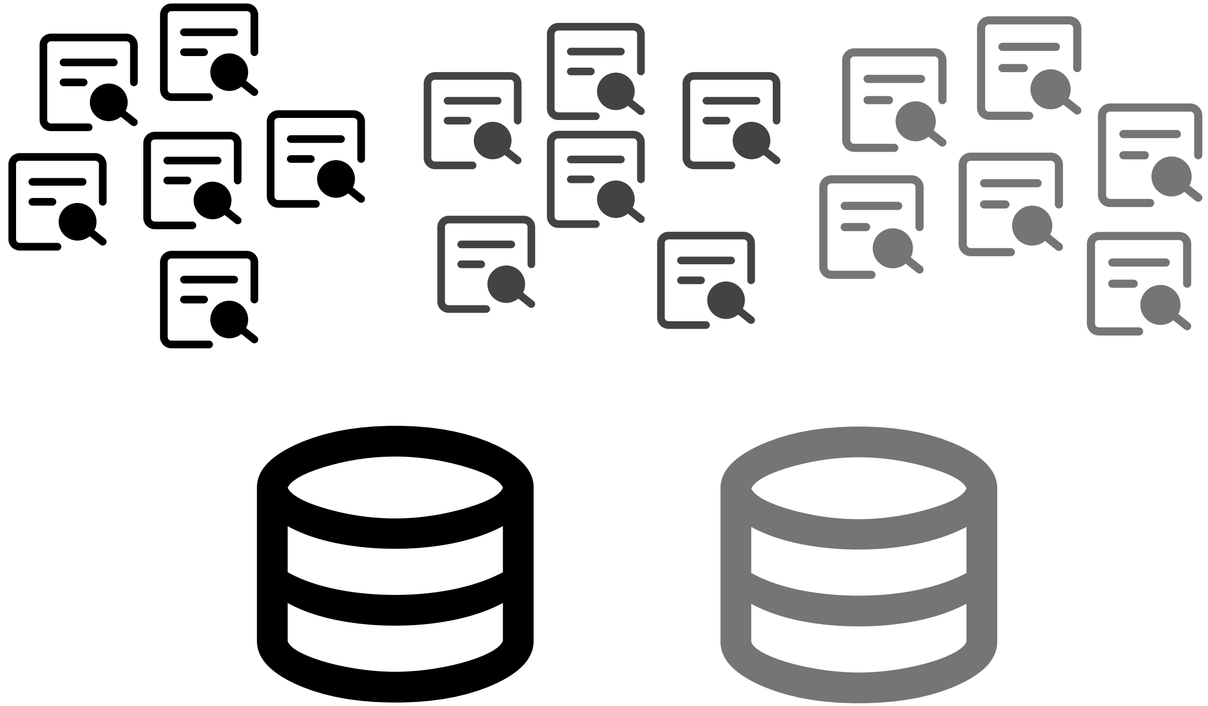
## Our Approach



- Use QPP heuristics over *domains*
- *Aggregate* QPP measures are taken as CPP measures
- In doing so we can compare corpora by the domains they are best suited to serve
- Conversely, akin to retrievability we can observe domains for which queries are difficult to serve across multiple corpora

# Corpora Performance Prediction

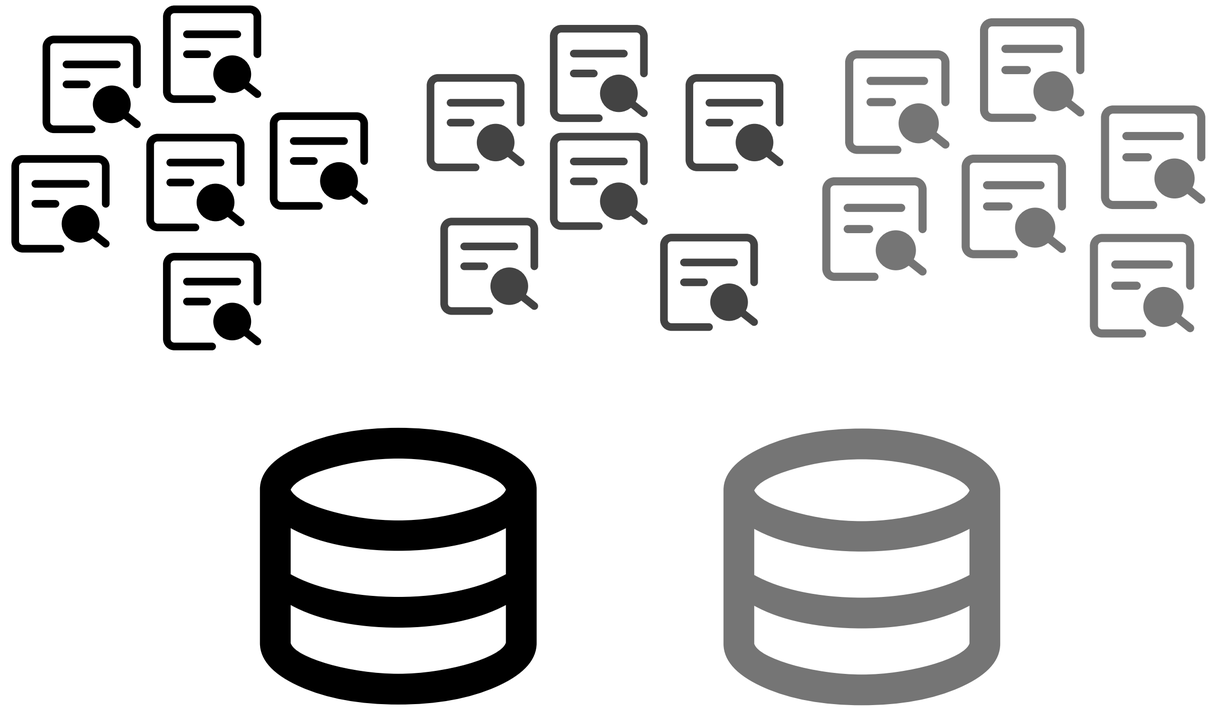
## Requirements



# Corpora Performance Prediction

## Requirements

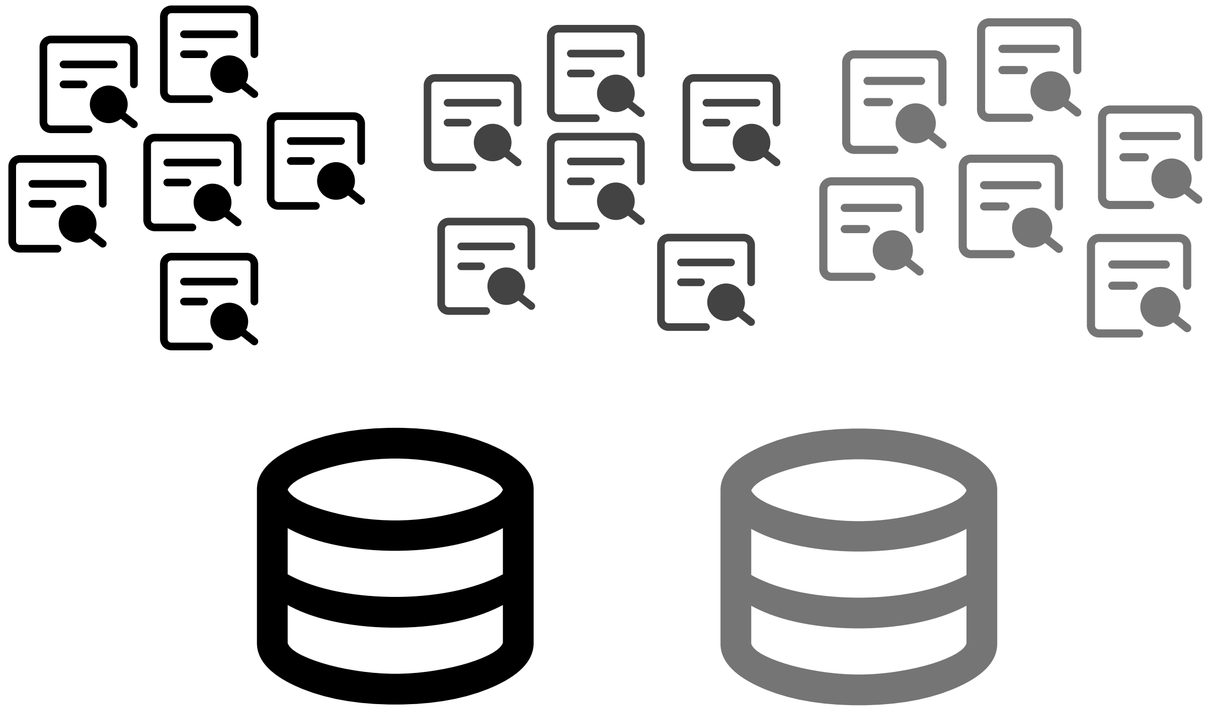
- Query Log



# Corpora Performance Prediction

## Requirements

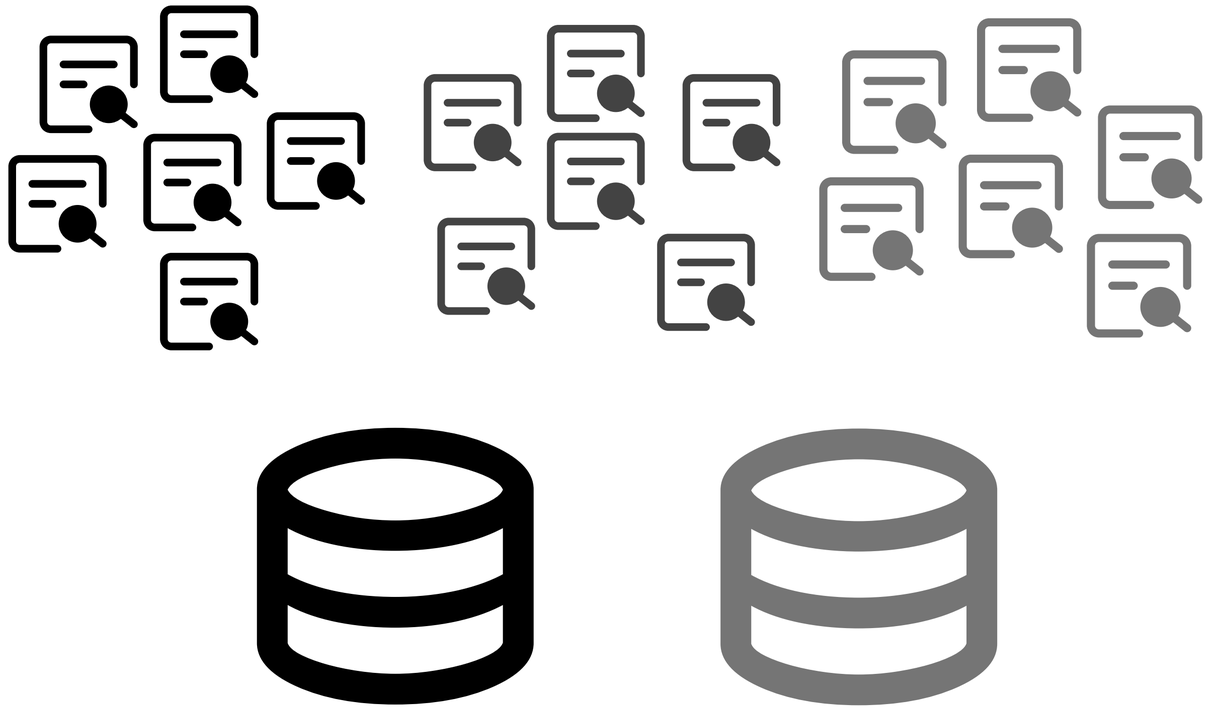
- Query Log
  - *Multiple Domains*



# Corpora Performance Prediction

## Requirements

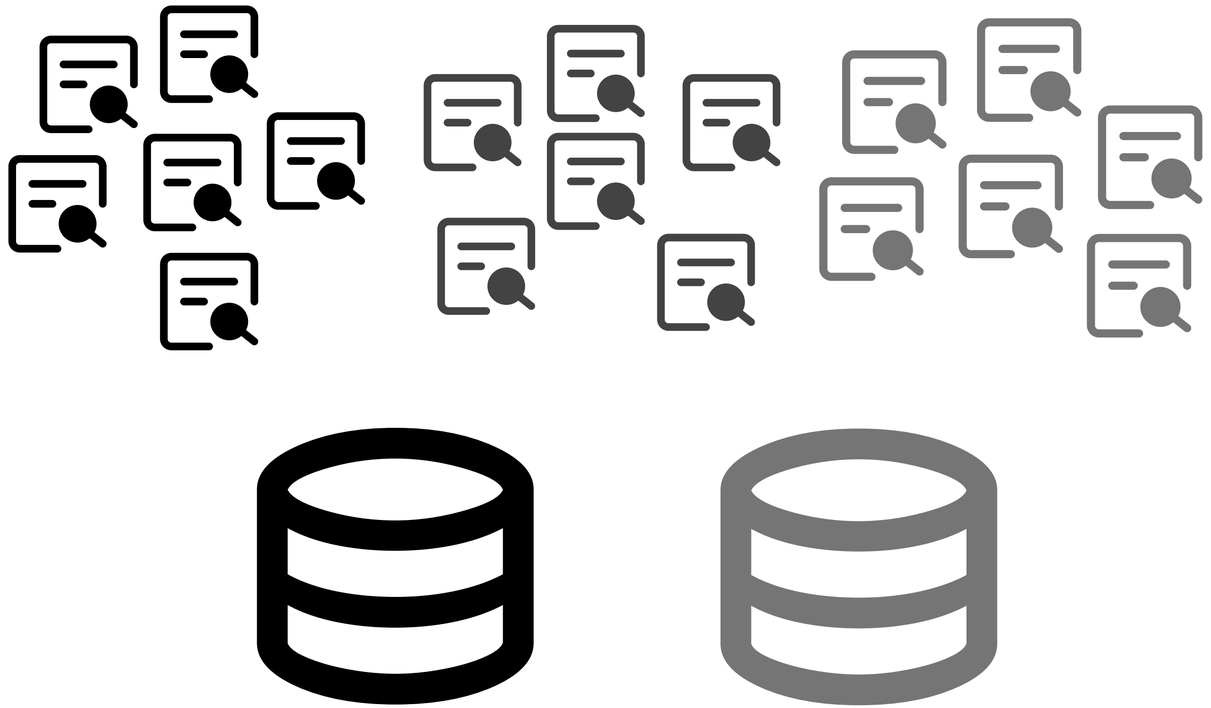
- Query Log
  - *Multiple Domains*
- Candidate Corpora



# Corpora Performance Prediction

## Requirements

- Query Log
  - *Multiple Domains*
- Candidate Corpora
- QPP Measures



# Corpora Performance Prediction

## Query Log

# Corpora Performance Prediction

## Query Log

- We have test corpora which may be mined from a particular domain

# Corpora Performance Prediction

## Query Log

- We have test corpora which may be mined from a particular domain
  - These corpora are *small*

# Corpora Performance Prediction

## Query Log

- We have test corpora which may be mined from a particular domain
  - These corpora are *small*
- We leverage the *archive query log*

# Corpora Performance Prediction

## Query Log

- We have test corpora which may be mined from a particular domain
  - These corpora are *small*
- We leverage the *archive query log*
  - 64 million queries, 550 providers (domains)

# Corpora Performance Prediction

## Query Log

- We have test corpora which may be mined from a particular domain
  - These corpora are *small*
- We leverage the *archive query log*
  - 64 million queries, 550 providers (domains)
  - Sample of 20000 queries over 15 providers

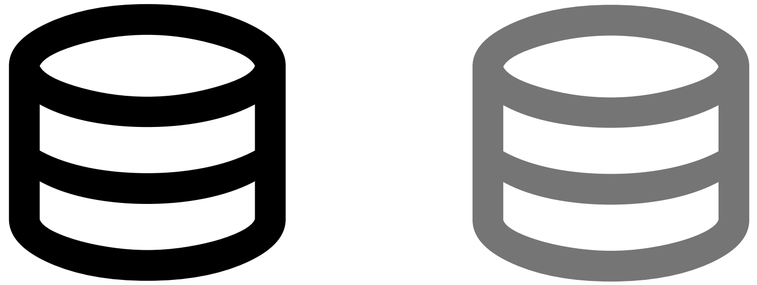
# Corpora Performance Prediction

## Query Log

| | |
|---|---|
| google | "Entoprocta"<br>"Darwin Deez" "Constellations"<br>"Csereüvegek" site:hu.wikipedia.org<br>"Samarqand Restaurant" -wikipedia<br>"Armas e equipamentos da Guerra Russo-Ucraniana" -wikipedia |
| naver | 바나다 알루미늄 블루투스 삼각대 셀카봉, WS-SQB641(화이트) 후기<br>hijrah<br>힐로 스테인레스 싱크롤 선반 20롤 대형, 블랙 후기<br>sumer<br>위성인터넷 |
| yahoo | ISSN "0340-1707"<br>payless All Size Waste Dumpsters Calgary<br>wichita craiglists<br>what causes vertigo in older adults<br>belvedere palace vienna |

# Corpora Performance Prediction

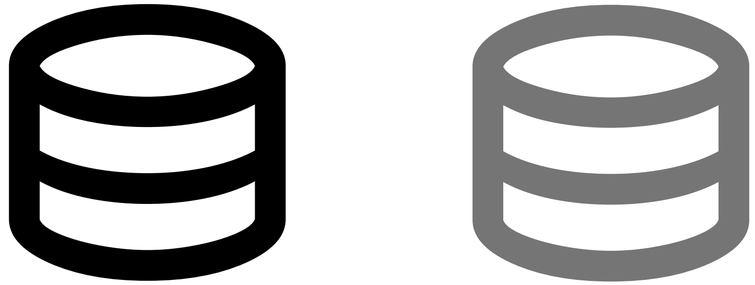
## Candidate Corpora



# Corpora Performance Prediction

## Candidate Corpora

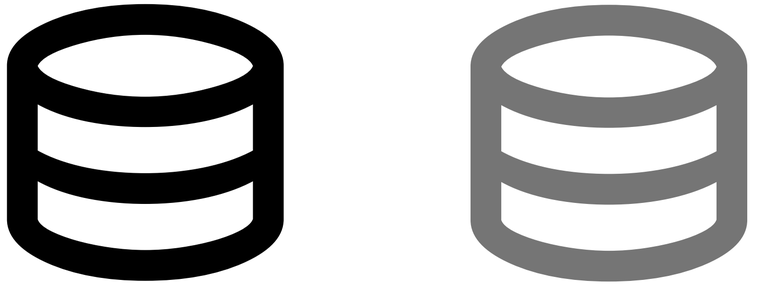
- MSMARCO Passage



# Corpora Performance Prediction

## Candidate Corpora

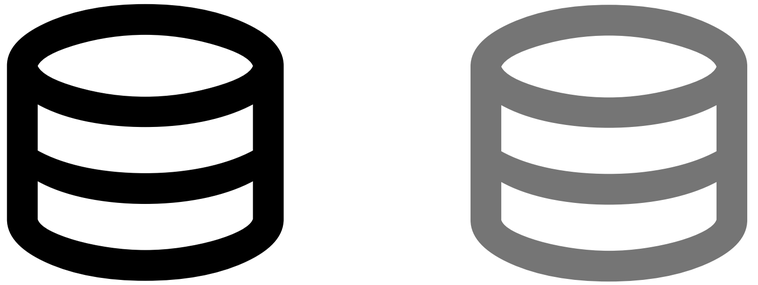
- MSMARCO Passage
  - Minimal Test Collection Subsample



# Corpora Performance Prediction

## Candidate Corpora

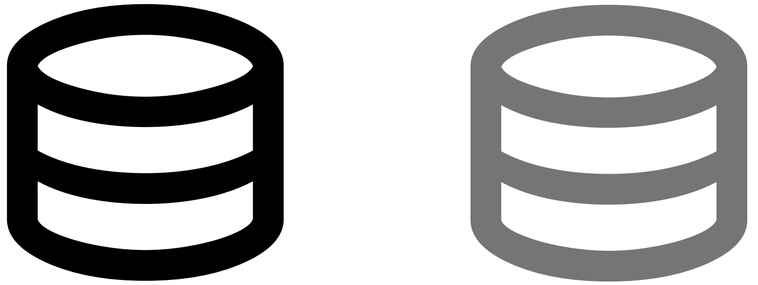
- MSMARCO Passage
  - Minimal Test Collection Subsample
- Touche Argument Retrieval



# Corpora Performance Prediction

## Candidate Corpora

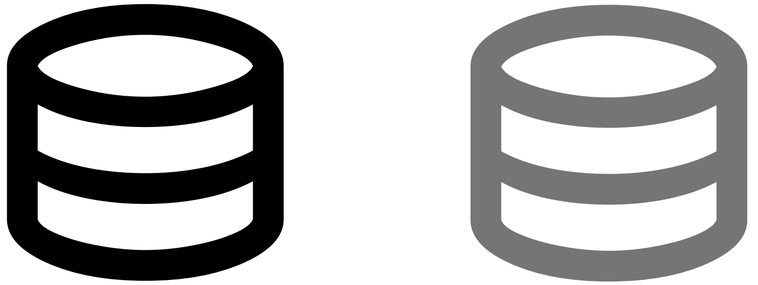
- MSMARCO Passage
  - Minimal Test Collection Subsample
- Touche Argument Retrieval
- NFCorpus



# Corpora Performance Prediction

## Candidate Corpora

- MSMARCO Passage
  - Minimal Test Collection Subsample
- Touche Argument Retrieval
- NFCorpus
- Cranfield



# Corpora Performance Prediction

## QPP Measures

# Corpora Performance Prediction

## QPP Measures

- 12 measures implemented in QPPTK (*dockerised in TIRA!*)

# Corpora Performance Prediction

## QPP Measures

- 12 measures implemented in QPPTK (*dockerised in TIRA!*)
- For this initial investigation we focus on WIG, SCQ, NQC and average-IDF

# Corpora Performance Prediction

## QPP Measures

- 12 measures implemented in QPPTK (*dockerised in TIRA!*)
- For this initial investigation we focus on WIG, SCQ, NQC and average-IDF
  - Our approach could in principle apply any pre- or post-retrieval heuristic

# Corpora Performance Prediction

## QPP Measures

- 12 measures implemented in QPPTK (*dockerised in TIRA!*)
- For this initial investigation we focus on WIG, SCQ, NQC and average-IDF
  - Our approach could in principle apply any pre- or post-retrieval heuristic
- For post-retrieval in these preliminary findings we solely use BM25

# Corpora Performance Prediction

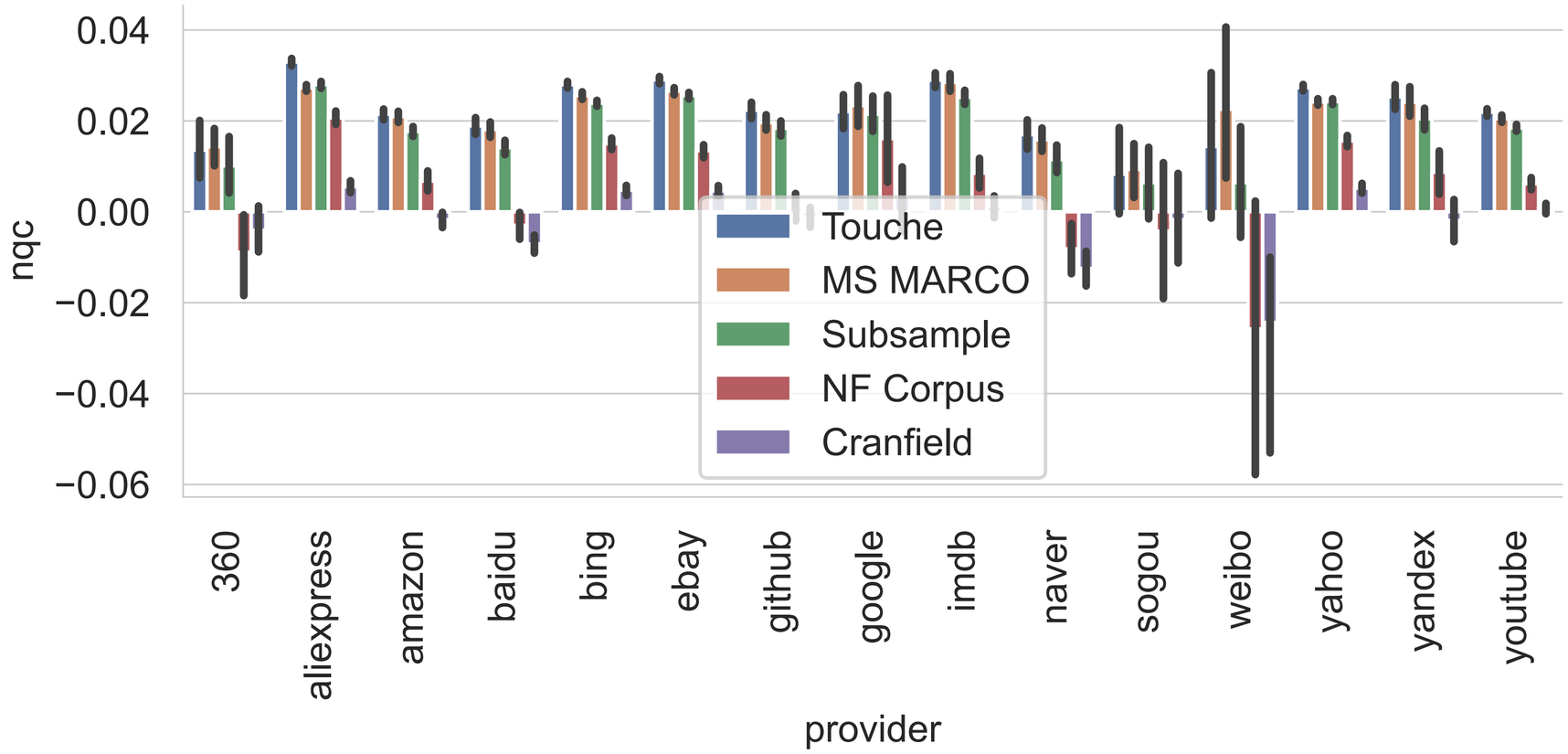
## QPP Measures

- 12 measures implemented in QPPTK (*dockerised in TIRA!*)
- For this initial investigation we focus on WIG, SCQ, NQC and average-IDF
  - Our approach could in principle apply any pre- or post-retrieval heuristic
- For post-retrieval in these preliminary findings we solely use BM25
- Our initial study is primarily concerned with the feasibility of comparing corpora by QPP measures, we make some assumptions about the faithfulness of the QPP measures

# Experiments

# Experiments

## Corpora by Class Performance



# Experiments

## Overall Similarity

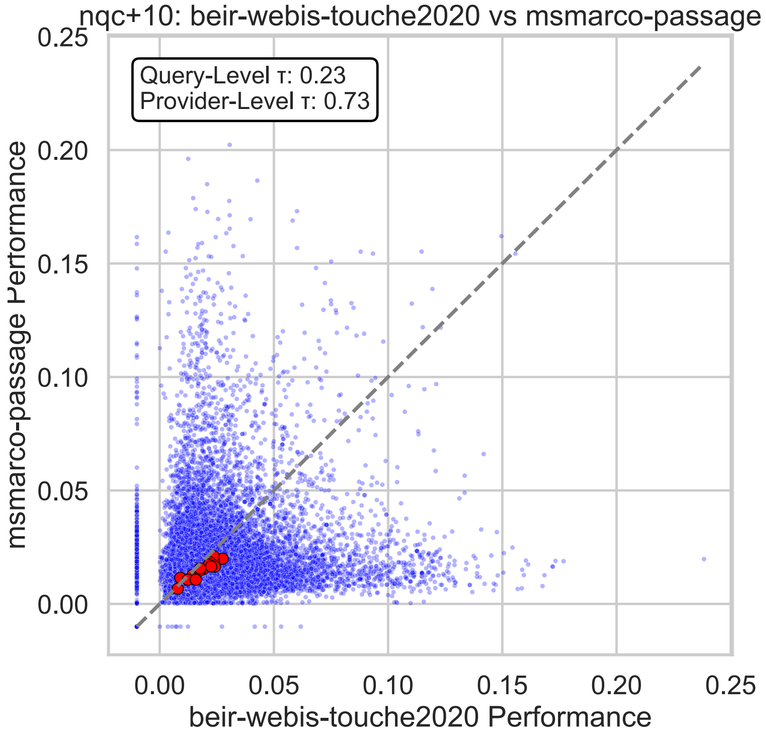
**Table 2**

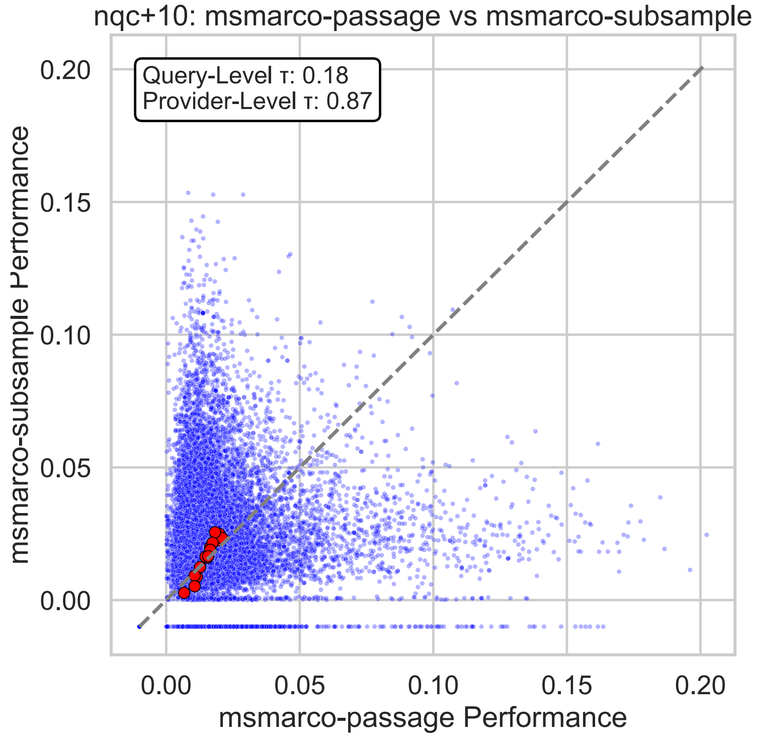
Kendall's $\tau$ correlation between the NQC values of different corpora, across all search providers.

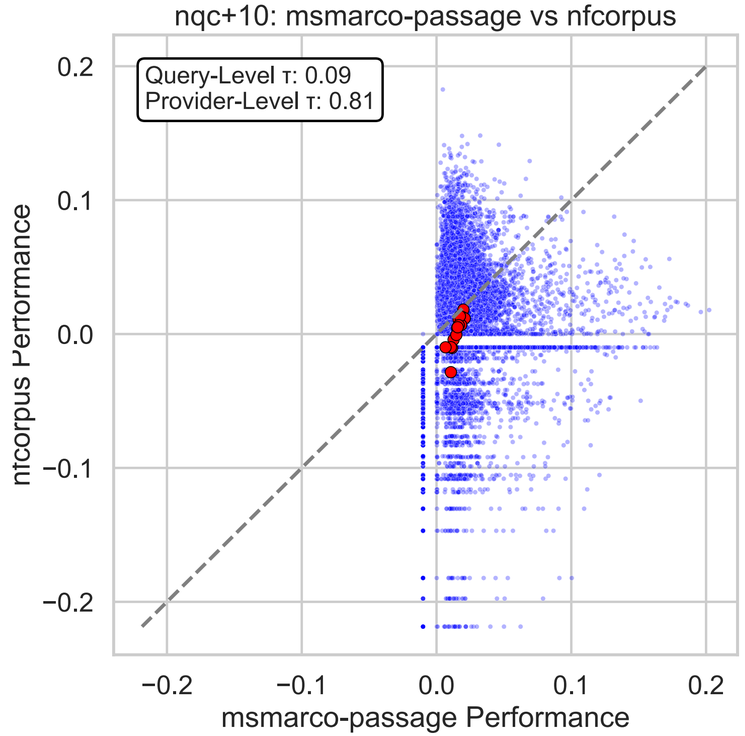
| | Touche | MS MARCO | Subsample | NF Corpus | Cranfield |
|---|---|---|---|---|---|
| Touche | - | 0.3225 | 0.4627 | 0.2313 | 0.0977 |
| MS MARCO | - | - | 0.2578 | 0.0845 | 0.0021 |
| Subsample | - | - | - | 0.2925 | 0.1361 |
| NF Corpus | - | - | - | - | 0.3574 |
| Cranfield | - | - | - | - | - |

# Experiments

## Granularity of Comparisons



# Experiments

## Per Domain Correlation

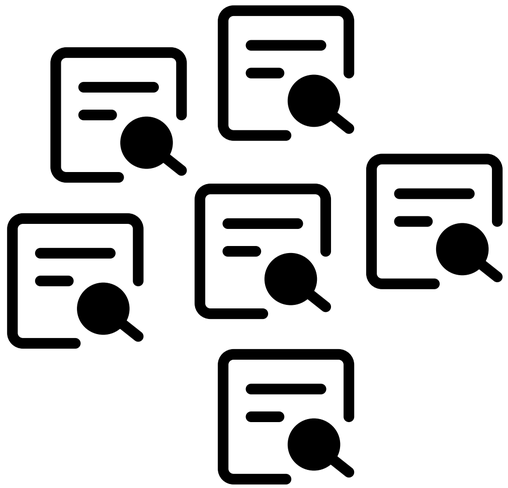
**Table 3**

Kendall's $\tau$ correlation between the NQC values of different corpora, grouped by search providers.

| provider | | Touche | MS MARCO | Subsample | NF Corpus | Cranfield |
|---|---|---|---|---|---|---|
| 360 | Touche | - | 0.3416 | 0.6867 | 0.4459 | 0.2210 |
| | MS MARCO | - | - | 0.2955 | 0.2265 | 0.0447 |
| | Subsample | - | - | - | 0.4661 | 0.2788 |
| | NF Corpus | - | - | - | - | 0.3989 |
| | Cranfield | - | - | - | - | - |
| aliexpress | Touche | - | 0.1249 | 0.3326 | 0.1270 | 0.0353 |
| | MS MARCO | - | - | 0.0853 | -0.0343 | -0.0470 |
| | Subsample | - | - | - | 0.1856 | 0.0660 |
| | NF Corpus | - | - | - | - | 0.2147 |
| | Cranfield | - | - | - | - | - |
| amazon | Touche | - | 0.4172 | 0.5539 | 0.3483 | 0.1679 |
| | MS MARCO | - | - | 0.3710 | 0.2361 | 0.1193 |
| | Subsample | - | - | - | 0.4318 | 0.2209 |
| | NF Corpus | - | - | - | - | 0.3921 |
| | Cranfield | - | - | - | - | - |
| baidu | Touche | - | 0.5011 | 0.5932 | 0.3641 | 0.1886 |
| | MS MARCO | - | - | 0.4314 | 0.2172 | 0.1293 |
| | Subsample | - | - | - | 0.4275 | 0.2552 |
| | NF Corpus | - | - | - | - | 0.3733 |
| | Cranfield | - | - | - | - | - |

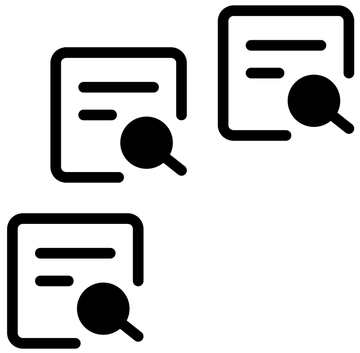
# Experiments

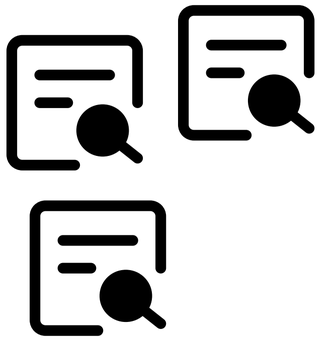
## Stability of Comparisons



# Experiments

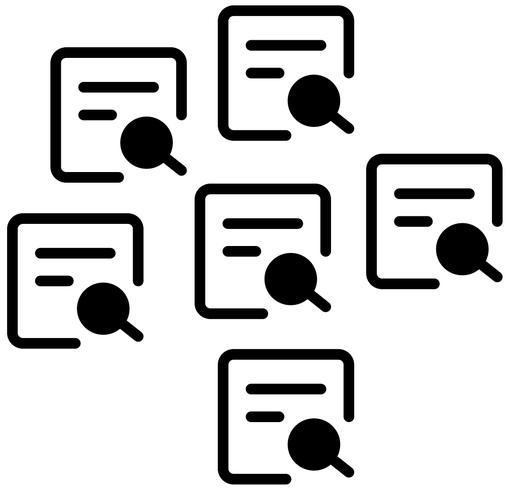
## Stability of Comparisons





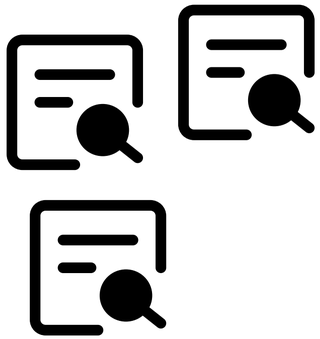
# Experiments

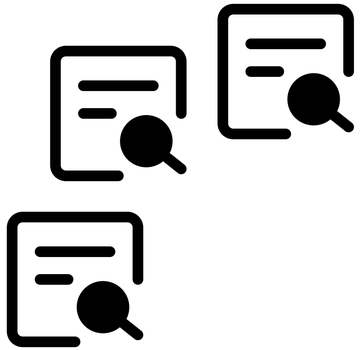
## Stability of Comparisons



# Experiments

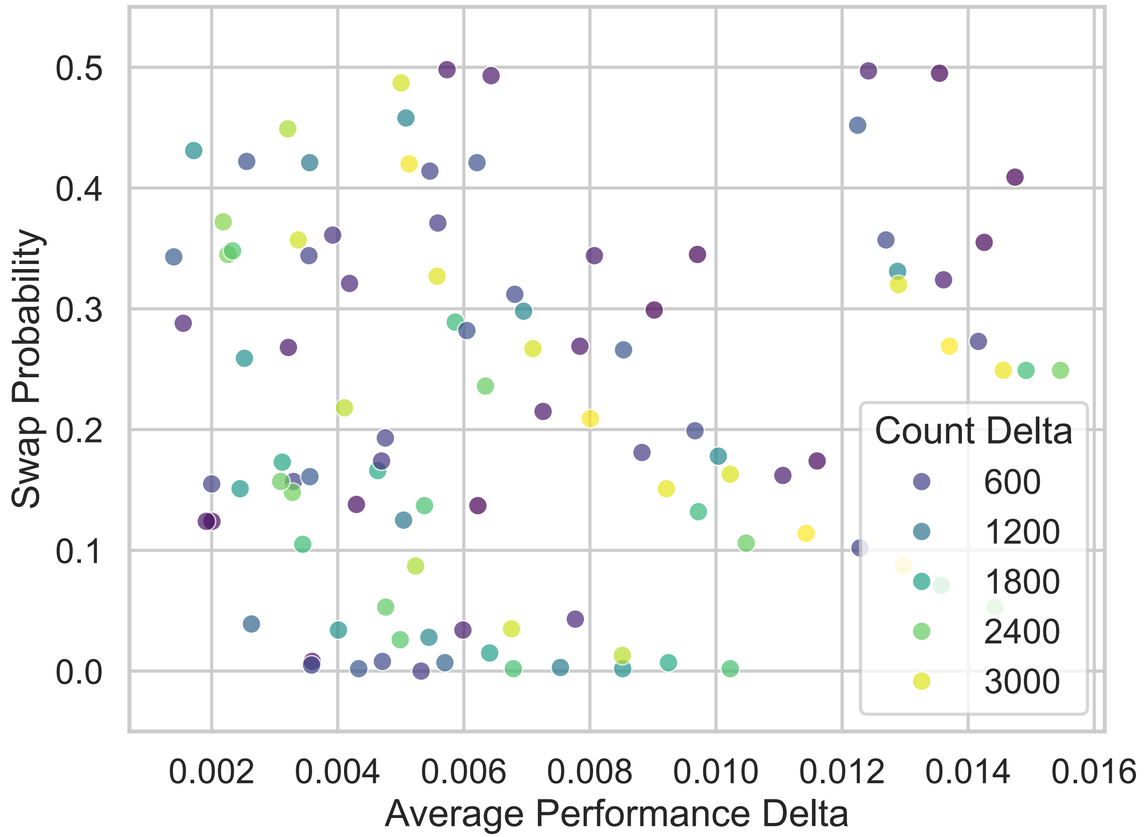
## Stability of Comparisons





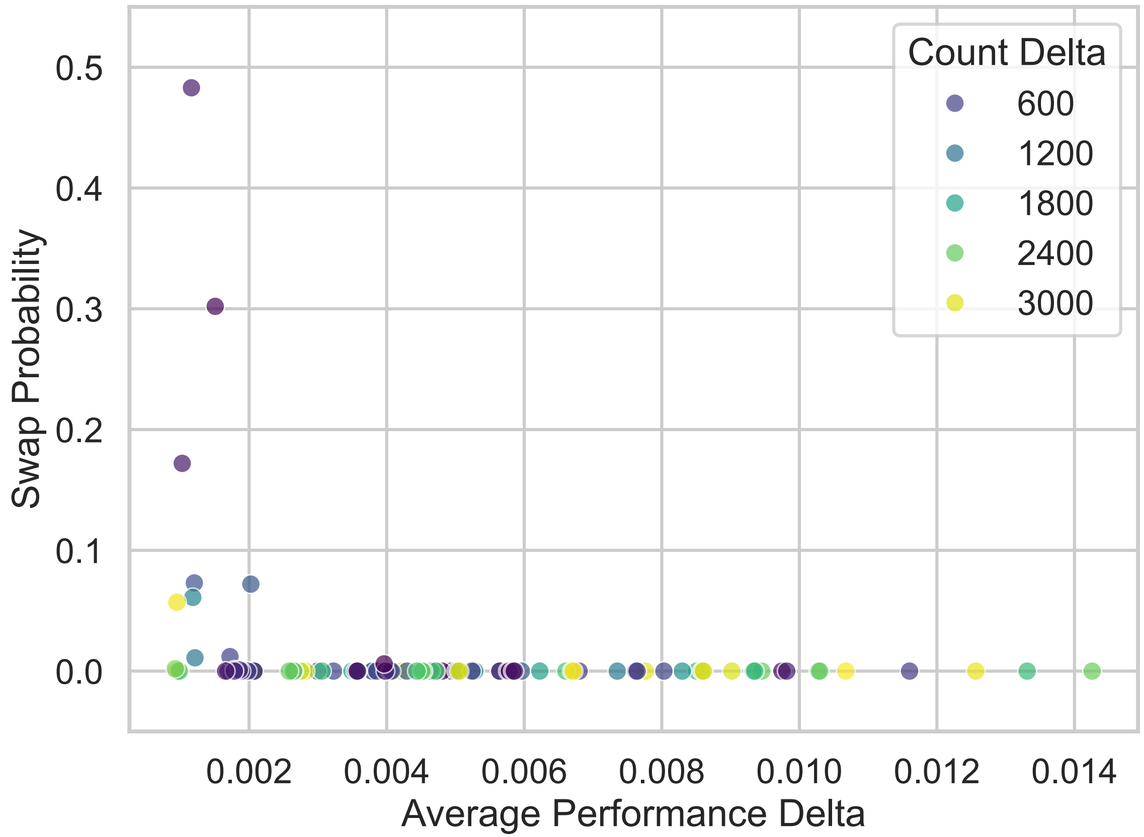
# Experiments

## Stability of Comparisons



# Experiments

## Stability of Comparisons



# Future / Continuing Work

# Future / Continuing Work

## What causes CPP similarities?

# Future / Continuing Work

## What causes CPP similarities?

- Domain?

# Future / Continuing Work

## What causes CPP similarities?

- Domain?
- Size?

# Future / Continuing Work

## What causes CPP similarities?

- Domain?
- Size?
- Entropy reduction by query difficulty?

# Future / Continuing Work

## What pitfalls from QPP may be reduced under aggregation

# Future / Continuing Work

## What pitfalls from QPP may be reduced under aggregation

- Our correlation is on aggregate akin to system order

# Future / Continuing Work

## What pitfalls from QPP may be reduced under aggregation

- Our correlation is on aggregate akin to system order
- Where precision at a query level is required for QPP, other methods may be more feasible at the domain level

# Future / Continuing Work

## What pitfalls from QPP may be reduced under aggregation

- Our correlation is on aggregate akin to system order
- Where precision at a query level is required for QPP, other methods may be more feasible at the domain level
- Broader correlation studies are required

# Future / Continuing Work

## What pitfalls from QPP may be reduced under aggregation

- Our correlation is on aggregate akin to system order
- Where precision at a query level is required for QPP, other methods may be more feasible at the domain level
- Broader correlation studies are required
  - LLM “aluminum judgements” come to mind to allow for validation of query difficulty

# Future / Continuing Work

## How fine-grained can our analysis be?

# Future / Continuing Work

## How fine-grained can our analysis be?

- In making minor changes to a system or a corpus update

# Future / Continuing Work

## How fine-grained can our analysis be?

- In making minor changes to a system or a corpus update
  - Are domains still stable?

# Future / Continuing Work

## How fine-grained can our analysis be?

- In making minor changes to a system or a corpus update
  - Are domains still stable?
  - Can we differentiate between a model better serving a domain and a domain simply now having better coverage?

# Future / Continuing Work

## How fine-grained can our analysis be?

- In making minor changes to a system or a corpus update
  - Are domains still stable?
  - Can we differentiate between a model better serving a domain and a domain simply now having better coverage?
- How do we do this without LLMs? Exciting but expensive and under small sample sizes comparisons are noisy

- CPP allows for holistic comparisons of corpora over a shared reference set of diverse queries

- CPP allows for holistic comparisons of corpora over a shared reference set of diverse queries
- Aggregation over topics leads to greater stability over weaker heuristics

- CPP allows for holistic comparisons of corpora over a shared reference set of diverse queries
- Aggregation over topics leads to greater stability over weaker heuristics
- Domain- and corpus-level effectiveness prediction can complement broader QPP evaluation

- CPP allows for holistic comparisons of corpora over a shared reference set of diverse queries
- Aggregation over topics leads to greater stability over weaker heuristics
- Domain- and corpus-level effectiveness prediction can complement broader QPP evaluation

**Thanks for your attention!**