If I Had a Million Queries

Ben Carterette, Virgil Pavlu, Evangelos Kanoulas,
Javed Aslam, James Allan

TREC 2008 Million Query Track

- Traditional TREC evaluation setup
  - Depth-100 pools judged
  - 50 queries
  - Infeasible (judgment effort) and insufficient
- Million Query evaluation setup
  - Reduce judgment effort by carefully selecting
    - Documents to judge
    - Types of queries to evaluate systems on

TREC 2008 Million Query Track

Questions:

1. Can low-cost methods reliably evaluate retrieval systems?

2. What is the minimum cost needed to reach reliable result?

3. Are some query types more informative than others?

4. Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?

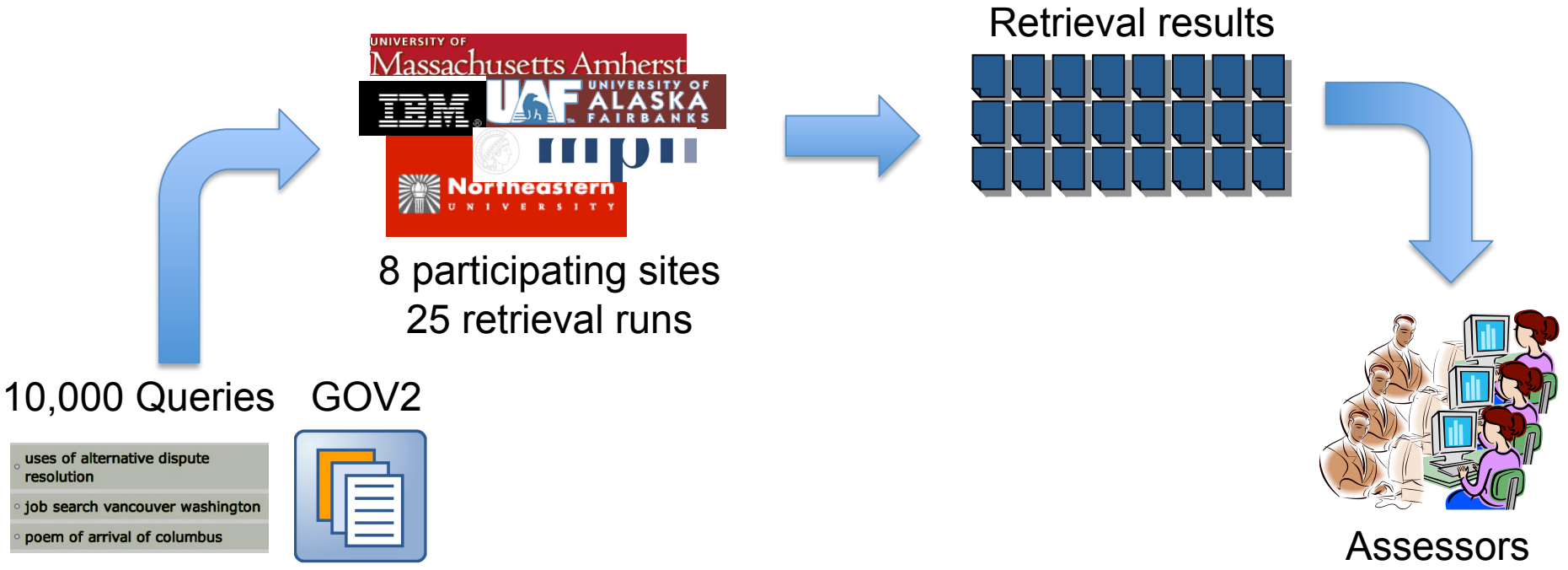
Million Query Track Setup

TREC crew @ NIST

Million Query Track Setup

TREC crew @ NIST

Million Query Track Setup



Million Query Track Setup

Retrieval results

8 participating sites
25 retrieval runs

10,000 Queries

GOV2

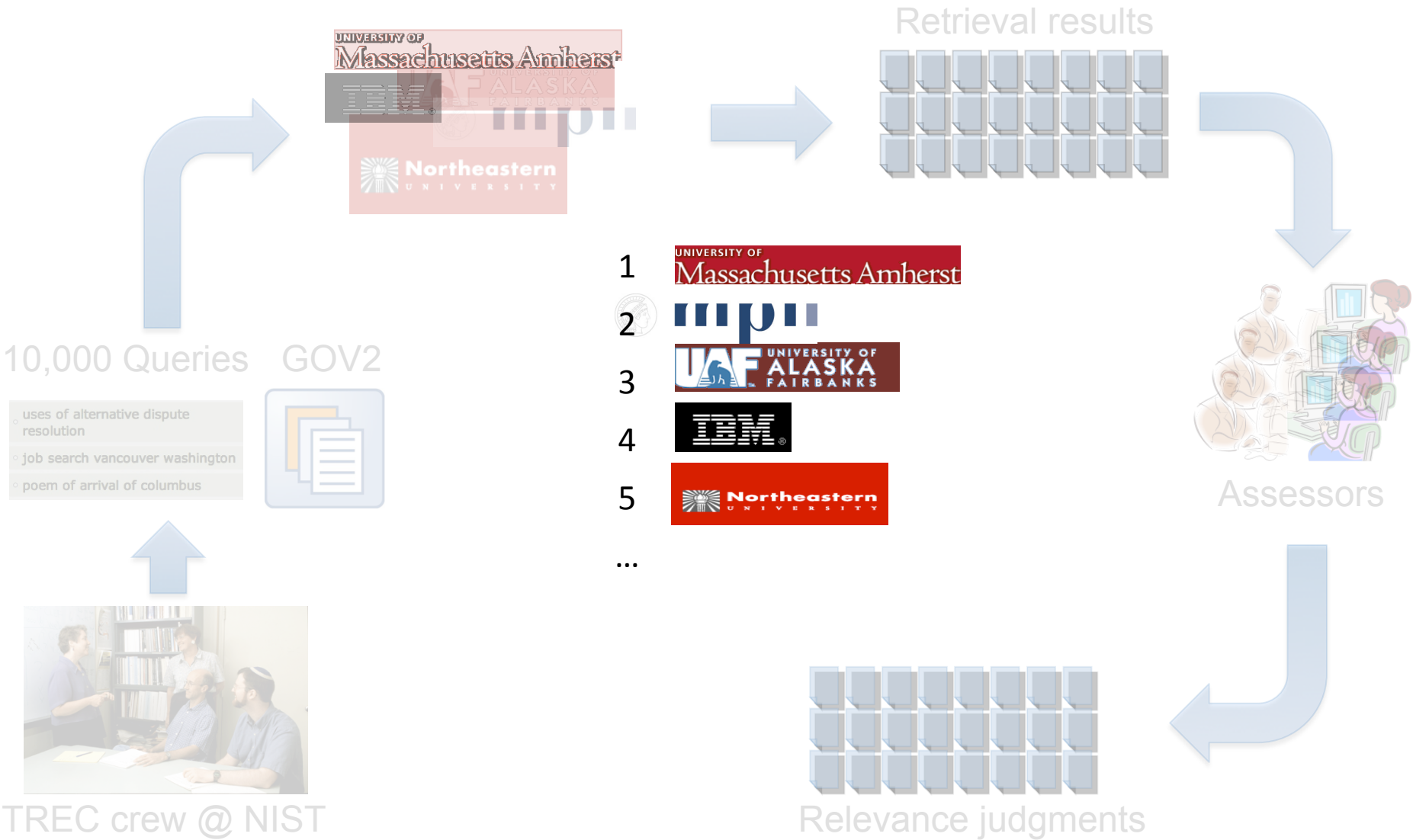
- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus

Assessors

TREC crew @ NIST

Relevance judgments

Million Query Track Setup



Document Selection and Evaluation

- Two low-cost algorithms
  - MTC (Carterette, Allan, & Sitaraman, 2006)

Document Selection

- Greedy on-line algorithm
- Selects most discriminative documents
- Targets at accurate ranking of systems

Evaluation

- Each document has a probability of relevance
- Measures as expectations over relevance distribution

Document Selection and Evaluation

- Two low-cost algorithms
  - statAP (Aslam & Pavlu, 2008)

    Document Selection
    - Stratified random sampling
    - Selects documents based on prior belief of relevance

    Evaluation
    - Apply well-established estimation techniques
    - Targets at accurate system scores

Queries

- 10,000 queries sampled from logs of a search engine.

- Queries were assigned categories
  - Long (more than 6 words) vs. Short
  - Gov-heavy (more than 3 clicks) vs. Gov-slant

	short	long
gov-slant	2,434	2,434
gov-heavy	2,434	2,434

Judgments per Query

- Five different targets for number of judgments
  - 8, 16, 32, 64 and 128 judgments targeted
  - Equal total number of judgments per target over all queries

Relevance Judgments

- 784 of the 10,000 queries judged
- 15,211 total judgments
  - ~75% less than in past years

Relevance Judgments

- Distribution of queries per category and judgment target

Category \ Judgments	8	16	32	64	128	Total
Short-govslant	95	55	29	13	4	196
Short-govheavy	118	40	26	10	3	197
Long-govslant	98	52	26	13	8	197
Long-govheavy	92	57	21	14	10	194
Total	403	204	102	50	25	784

Evaluation Measure

- Weighted MAP:

$$\text{wMAP} = \frac{1}{5}\sum_{j=1}^{5}\text{MAP}_j = \frac{1}{5}\sum_{j=1}^{5}\frac{1}{|\text{Q}_\text{j}|}\sum_{\text{q}\in\text{j}}\text{AP}_q$$

Judgments	8	16	32	64	128	Total
Total	403	204	102	50	25	784

TREC 2008 Million Query Track

Questions:

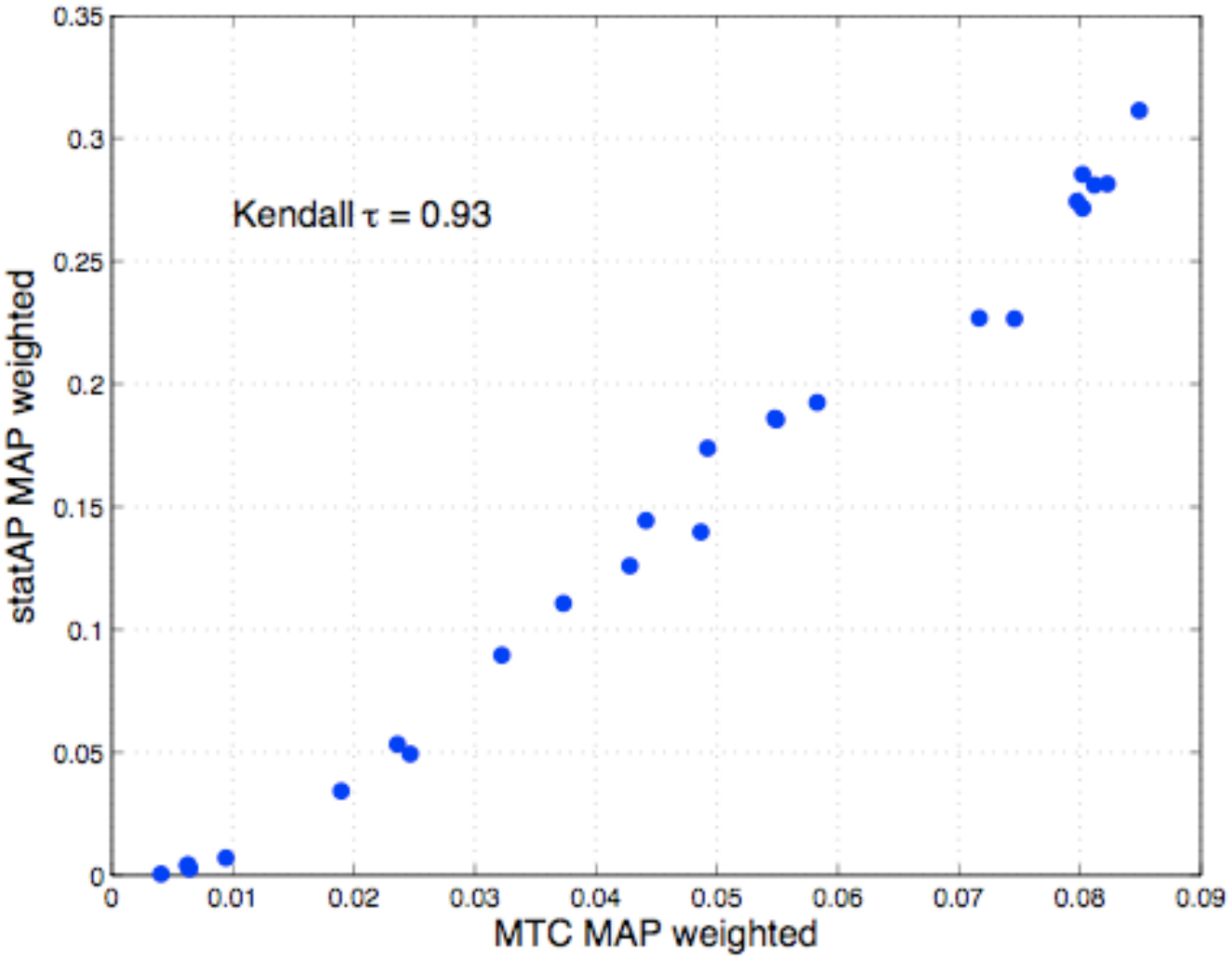
1. Can low-cost methods reliably evaluate retrieval systems?

2. What is the minimum cost needed to reach reliable result?

3. Are some query types more informative than others?

4. Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?

System Scores and Rankings



TREC 2008 Million Query Track

Questions:

1. Can low-cost methods reliably evaluate retrieval systems?

2. What is the minimum cost needed to reach reliable result?

3. Are some query types more informative than others?

4. Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?

Timing Info for Cost Analysis

- Query overhead

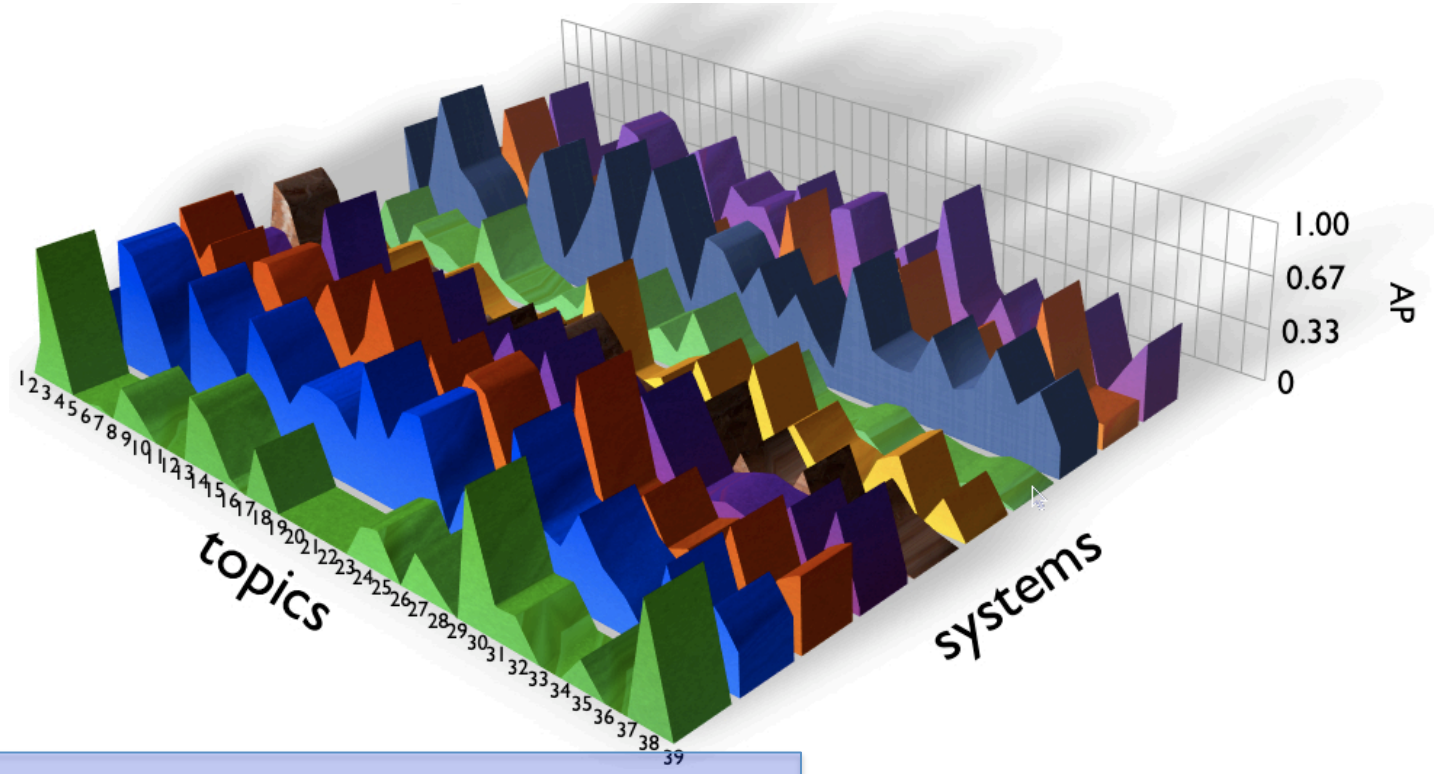
	refresh	view	last view	topic
short	2.34	18.0	25.5	67.6
long	2.54	24.5	31.0	86.5
gov-slant	2.22	22.5	29.0	76.0
gov-heavy	2.65	20.0	27.5	78.0
average	**2.41**	**22.0**	**29.0**	**76.0**

Timing Info for Cost Analysis

- Judging time per category and judgment target

	8	16	32	64	128	average
short	15.0	11.5	13.5	12.0	8.5	**12.5**
long	17.0	14.0	16.5	10.0	10.5	**13.0**
gov-slant	13.0	12.5	13.0	9.5	10.5	**12.0**
gov-heavy	19.0	13.0	17.0	12.5	8.5	**13.5**
average	**15.0**	**13.0**	**15.0**	**11.0**	**9.0**	**13.0**

Analysis of Variance



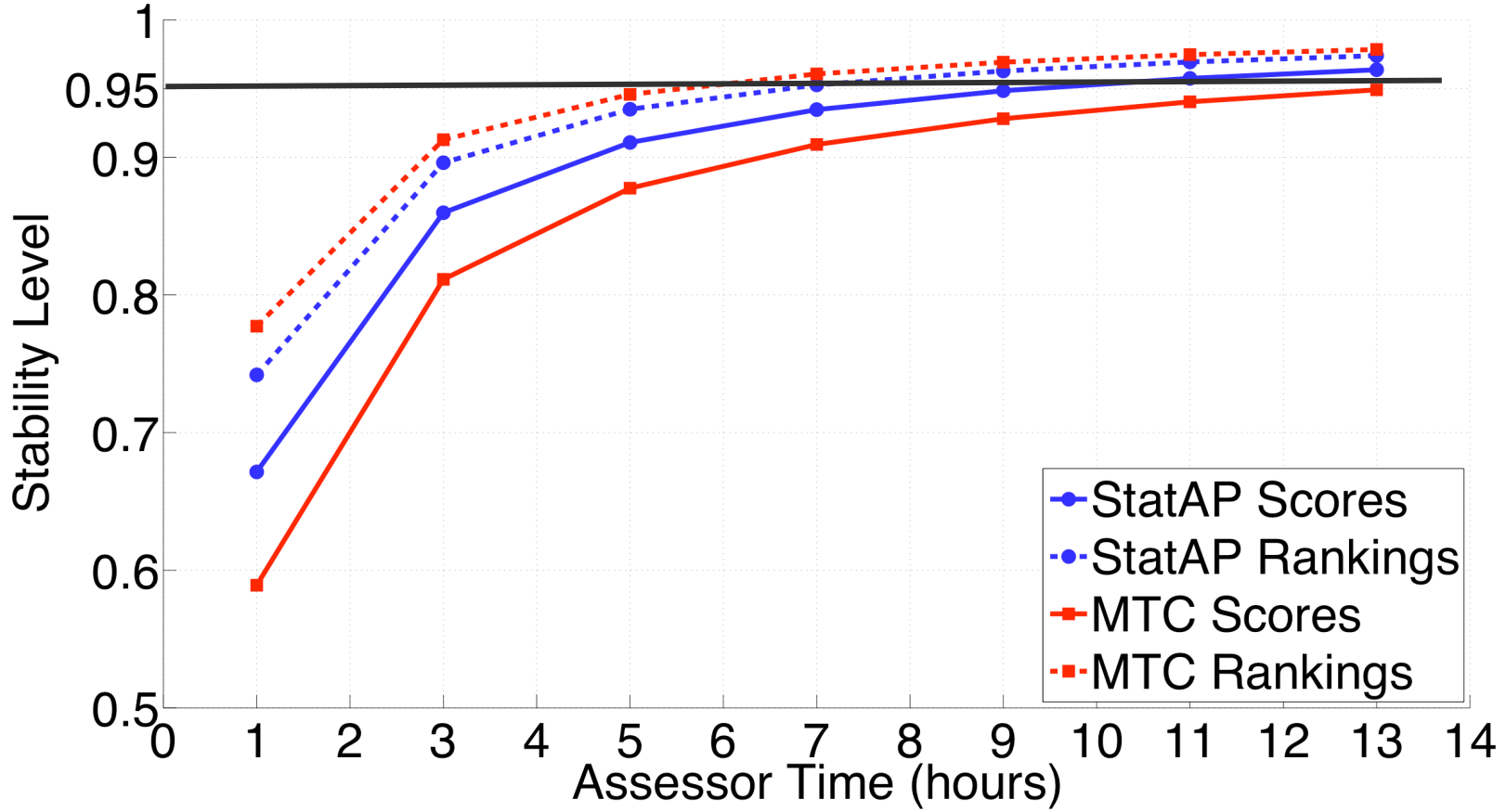
- σ_s = variance due to systems
- σ_q = variance due to queries
- σ_{sq} = variance due to query-system interaction

TREC 2008 Million Query Track

- Measure the stability of
  - Scores: $\frac{\text{Variance due to systems}}{\text{Total variance}}$
  - Rankings: $\frac{\text{Variance due to systems}}{\text{Var. due to systems + Var. due to query-system}}$

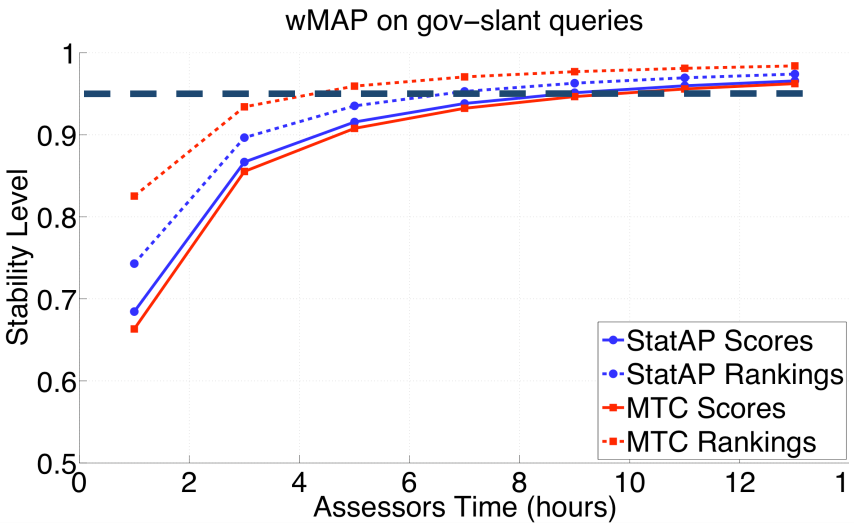
MAP Variance Components

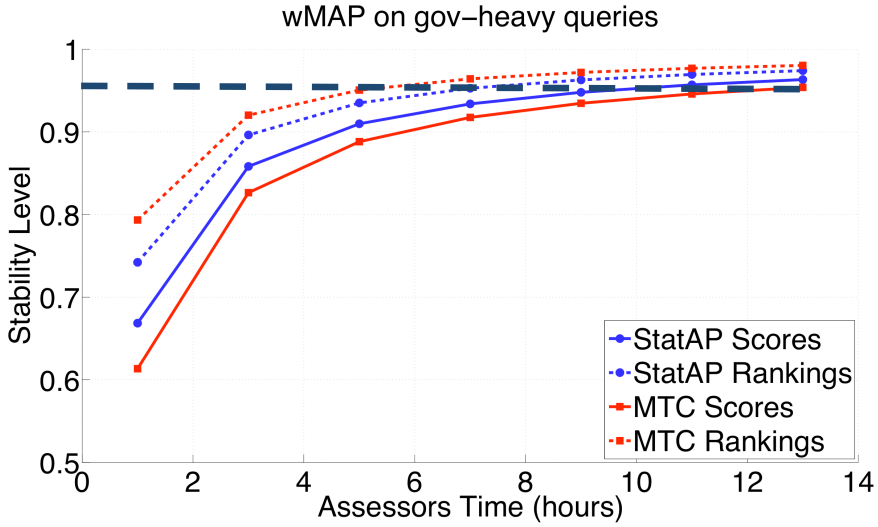
- What is the minimum cost needed to reach reliable result?



MAP Variance Components per Query Category

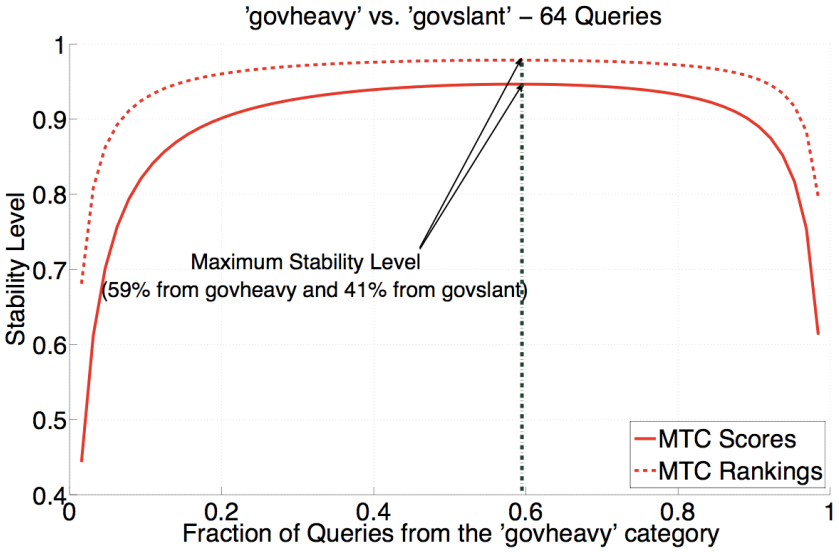
- Are some queries types more informative than others?

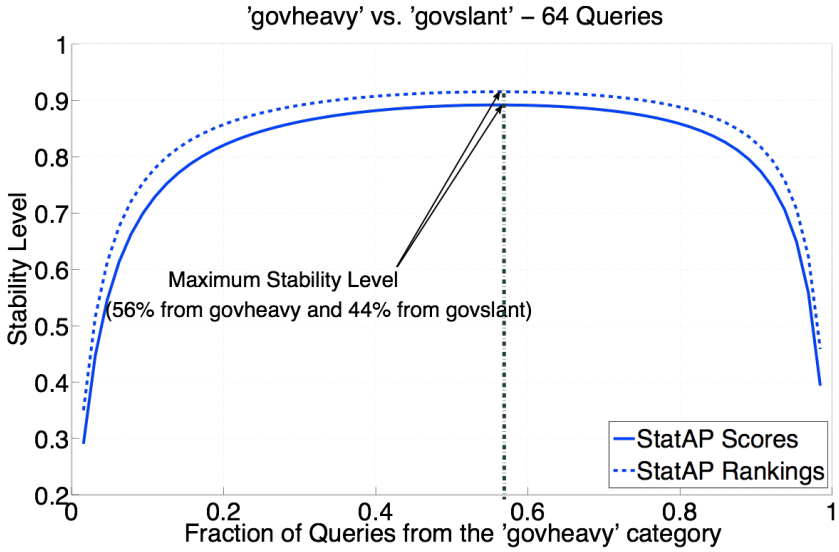




Query Selection

- Are some queries types more informative than others?





TREC 2008 Million Query Track

Questions:

1. Can low-cost methods reliably evaluate retrieval systems?

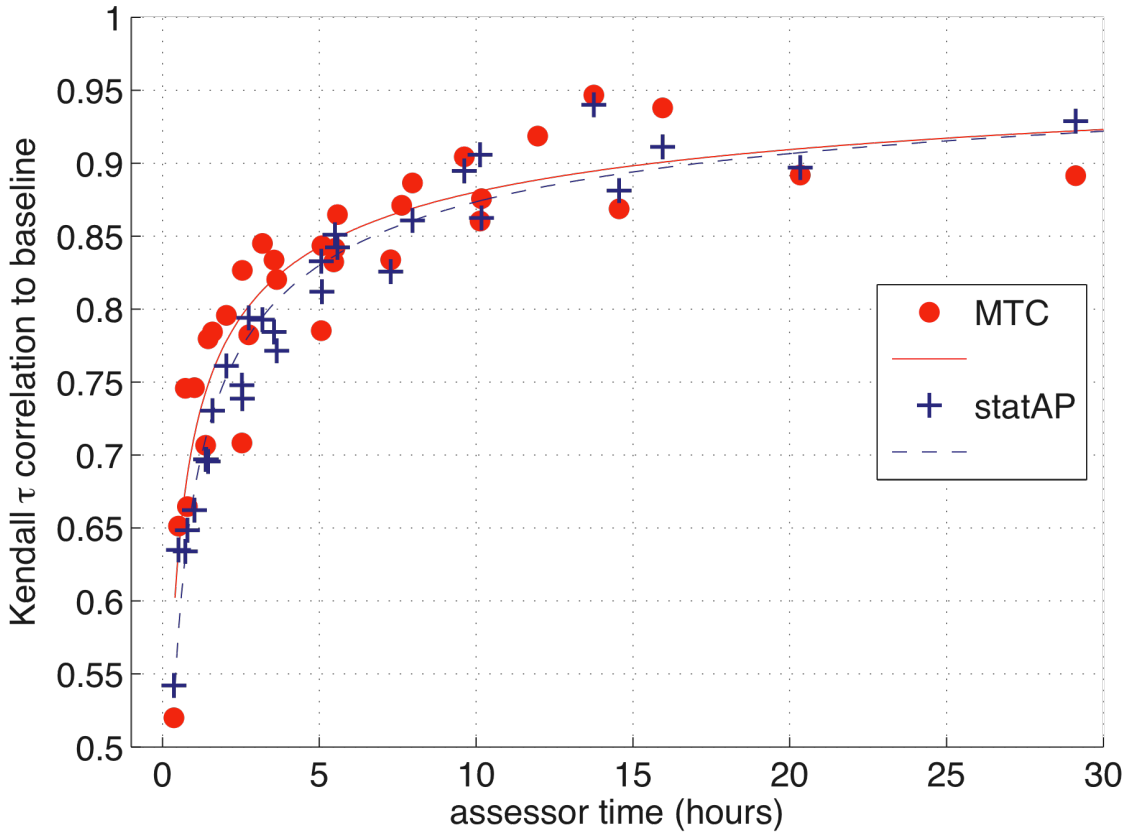
2. What is the minimum cost needed to reach reliable result?

3. Are some query types more informative than others?

4. Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?

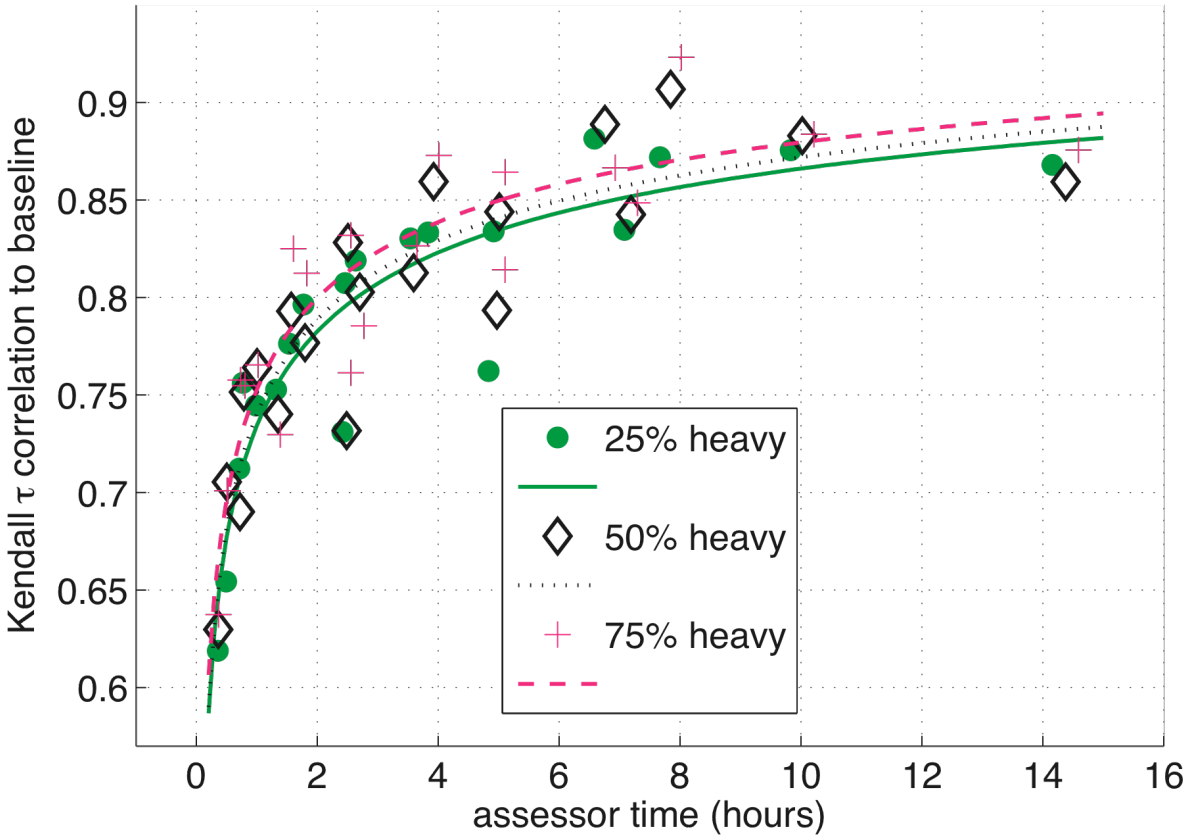
Kendall's tau Analysis

- What is the minimum cost needed to reach reliable result?



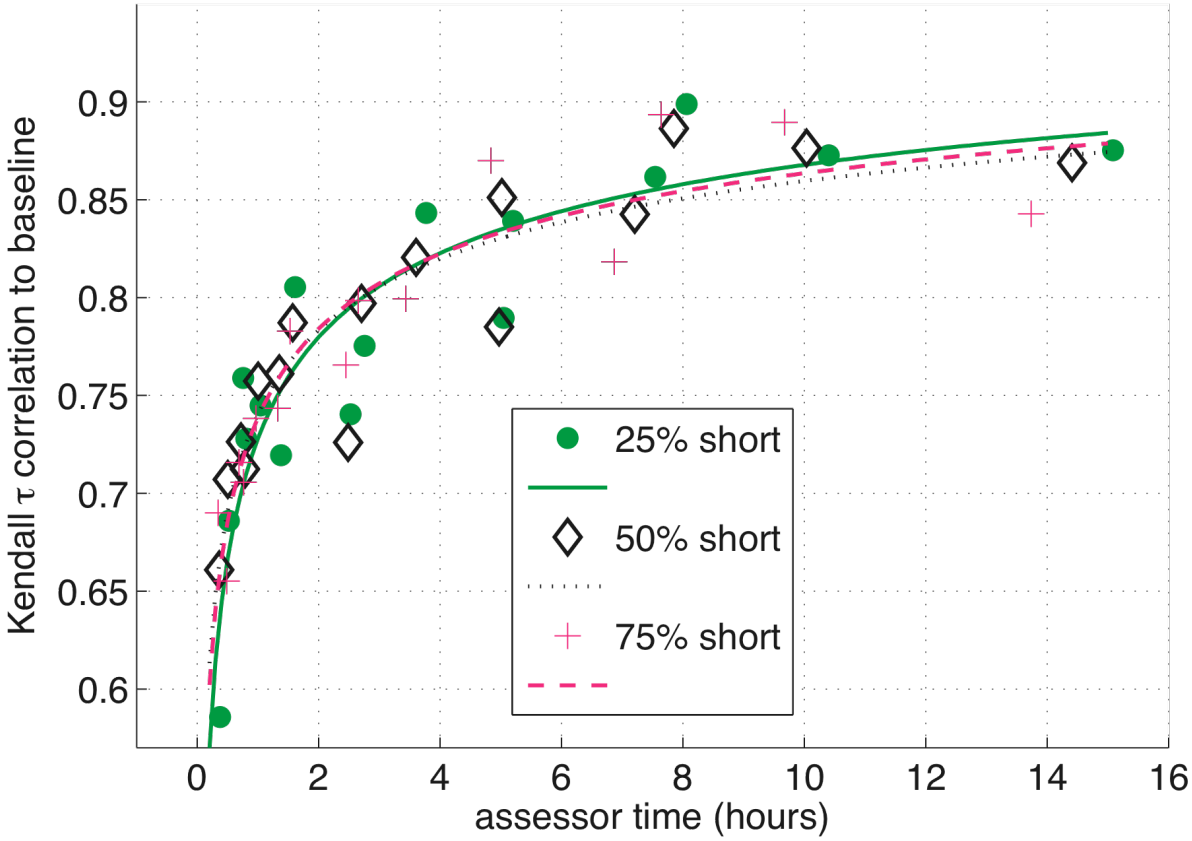
Kendall's tau Analysis

- Are some query types more informative than others?



Kendall's tau Analysis

- Are some query types more informative than others?



Relevance

- Percentage of relevant documents per query category and judgment target

Category \ Judgments	8	16	32	64	128	avg
Short-govslant	18.7	12.1	20.2	13.9	3.0	**14.6**
Long-govslant	20.2	17.0	17.3	12.0	13.7	**15.9**
Short-govheavy	24.6	30.8	30.4	23.4	37.4	**28.3**
Long-govheavy	28.8	20.4	22.3	13.6	16.0	**19.6**
avg	**23.1**	**19.3**	**22.5**	**15.2**	**15.7**	**19.3**

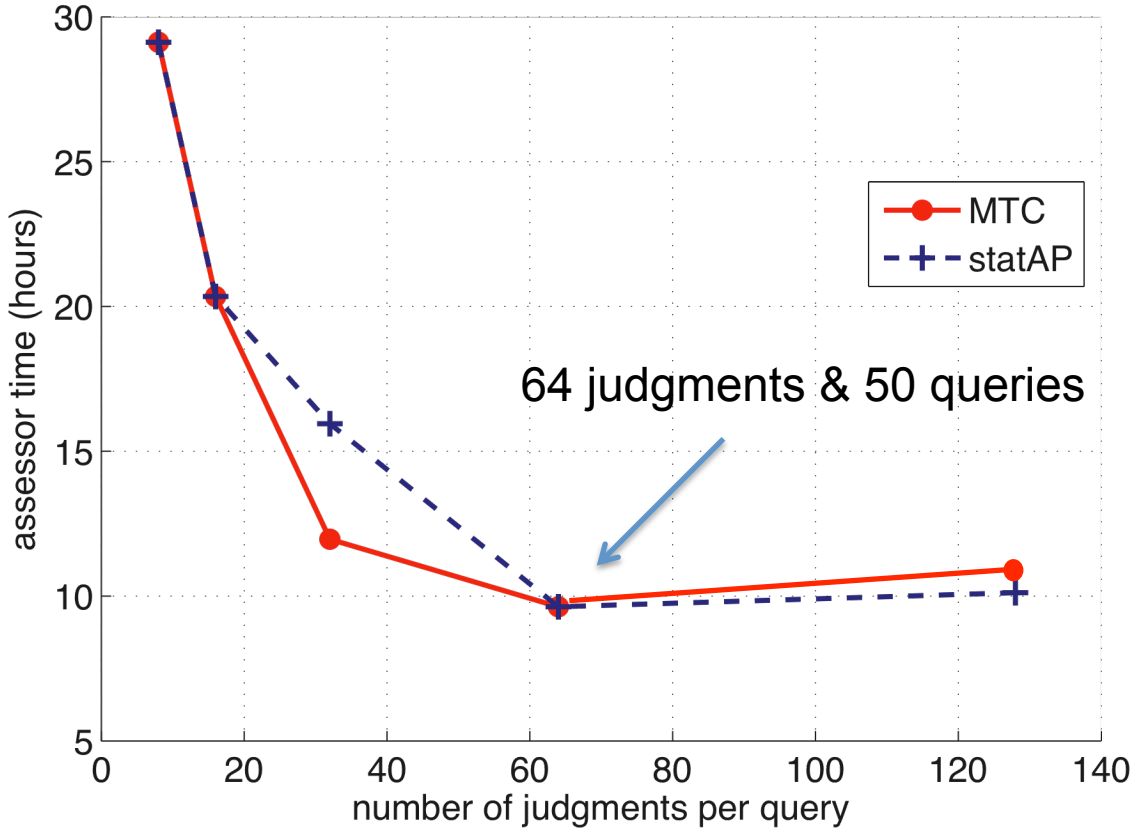
TREC 2008 Million Query Track

Questions:

1. Can low-cost methods reliably evaluate retrieval systems?
2. What is the minimum cost needed to reach reliable result?
3. Are some query types more informative than others?
4. Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?

Cost-Benefit Analysis

- Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?



Conclusion

- Low-cost methods reliably evaluate retrieval systems with very few judgments
- Minimum cost to reach reliable results
  - 10-15 hours of judgment time
- Some queries more informative than others
  - Gov-heavy more informative than gov-slant
- 64 judgments per query with around 50 queries is optimal for assessing systems' performance ranking