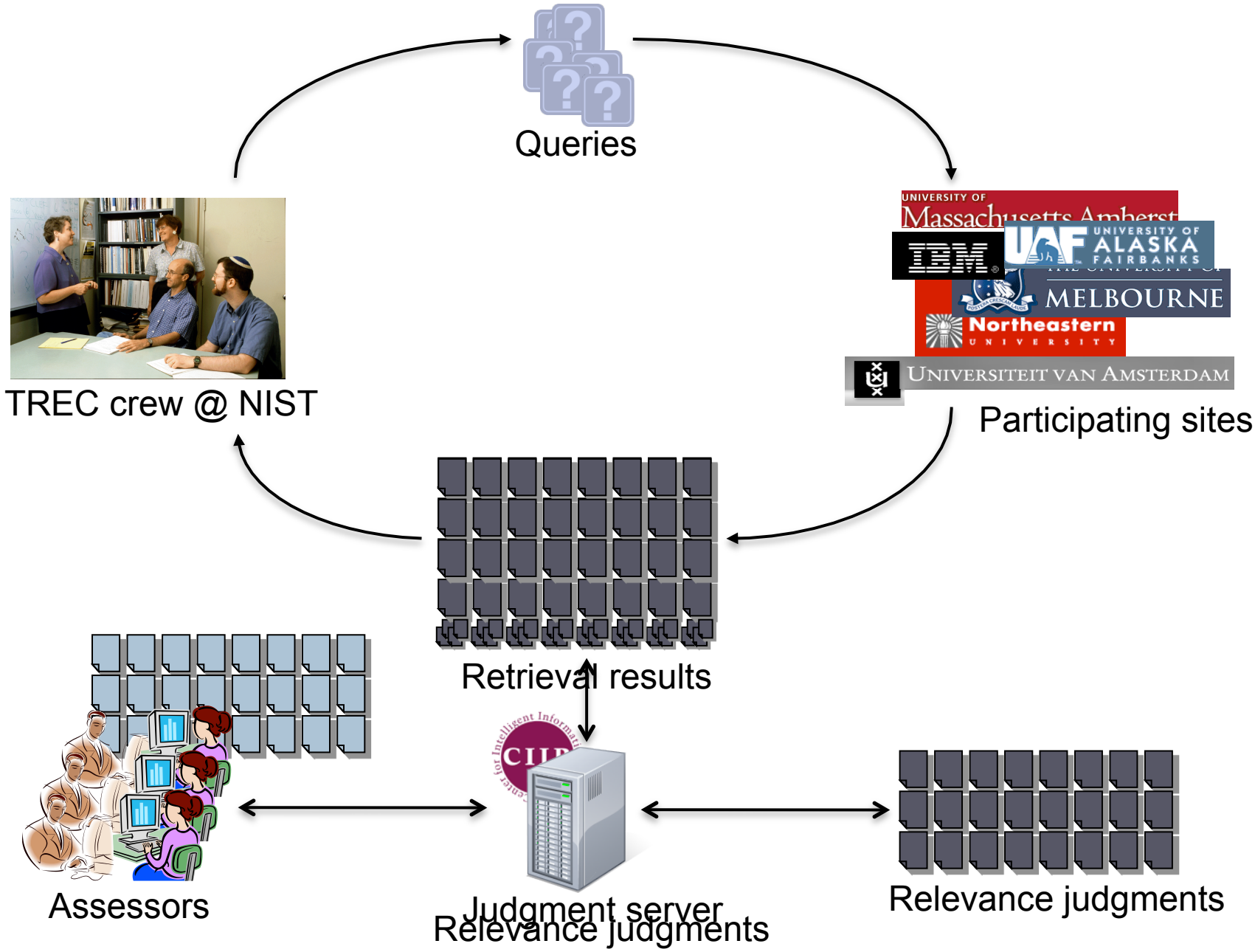
Evaluation Over Thousands of Queries

Ben Carterette
James Allan

Virgil Pavlu
Evangelos Kanoulas
Javed Aslam

TREC 2007 Million Query Track

- Questions:
  - Can low-cost methods reliably evaluate retrieval systems?
  - Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?
- Experiment overview:
  - Retrieval task: ad hoc.
  - Corpus: GOV2 (25M web pages).
  - Queries: 10,000 queries sampled from logs of a search engine.
  - Evaluate 24 retrieval runs from 10 participating sites.

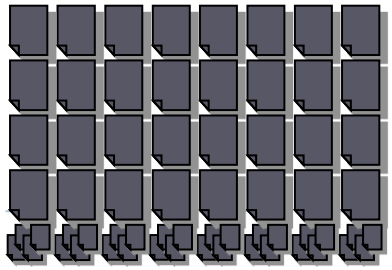

Queries
TREC crew @ NIST
Participating sites
Retrieval results
Assessors
Judgment server
Relevance judgments
Relevance judgments

Queries

- 10,000 queries sampled from logs of a search engine.
- Each had at least one click on a web page in the .gov domain.
  - Assumption: at least one relevant web page in corpus.

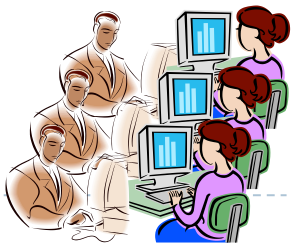
- Example queries:
  - arnold shwartzenegger
  - health care facility stress
  - fairfax county va divorce
  - crown vetch seed
  - ayanna

Retrieval Runs



- 24 runs from 10 sites.
- Different retrieval engines:
  - Lemur, Indri, Lucene, Zettair, among others.
- Different retrieval models:
  - Vector space, language modeling, inference networks, dependence models.
  - Pseudo-relevance feedback, external expansion, network-link models, HTML structure.
- Different stemmers:
  - Porter, Krovetz.
- Different stop lists.

Assessors



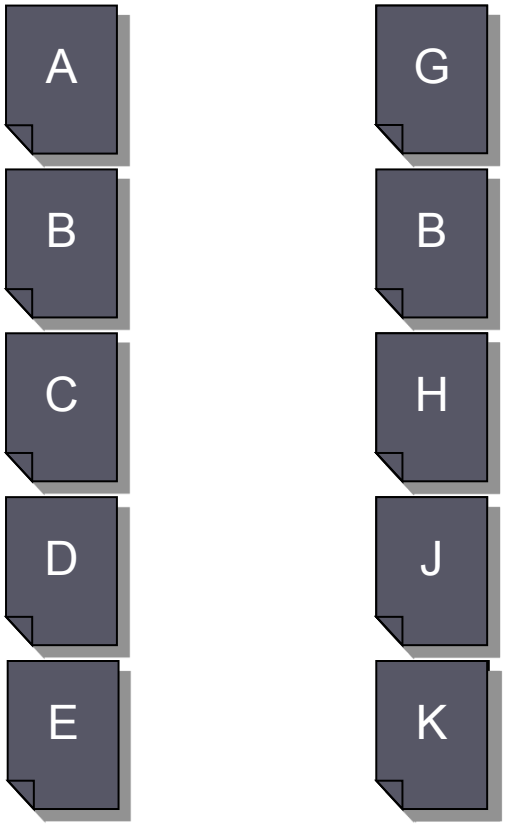
- Three groups of assessors:
  - NIST, participating sites, UMass undergrads.
- Given instructions and trained on a query.
- Given a list of 10 queries, picked one to judge.
- Develop query into topic by "back-fitting":
  - Imagine what information need might presage selected query.
  - Write full description of information need.
  - Explain what information on a page would make it relevant, and notable types of related information that are not relevant.

Judgment Server

- Implemented two low-cost algorithms.
  - "MTC" – UMass' algorithmic selection method.
    - Carterette, Allan, & Sitaraman, 2006.
  - "statAP" – NEU's statistical sampling method.
    - Aslam & Pavlu, 2008.
- Each query served by either MTC, statAP, or an alternation of the two.
- Required at least 40 judgments for each query.

MTC – Algorithmic Document Selection

- Given two ranked lists, how few documents do we need to judge to discriminate them?



Limiting case: ranked lists are identical; no judgments needed.

If two documents swap, they become most interesting.

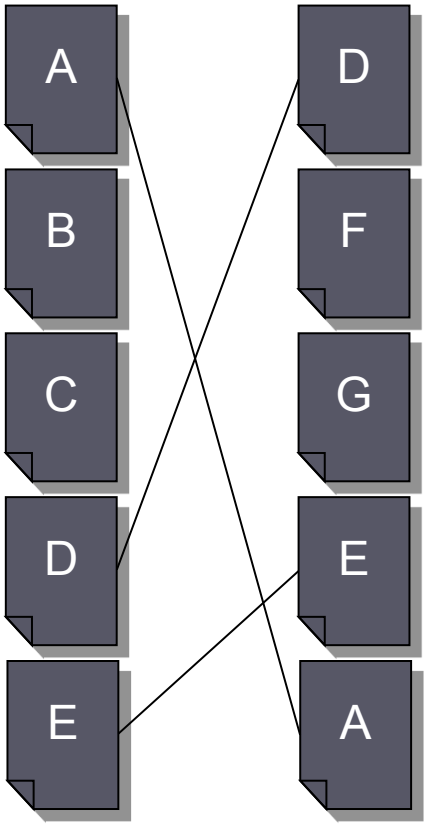
A document ranked by one system but not the other is interesting.

Limiting case: ranked lists are completely different, but relevance is the same at every rank.

$$\sum_{i=1}^{R}\sum_{j\geq i}$$

MTC – Algorithmic Document Selection

- Assign each document a weight according to its potential contribution to understanding the difference in AP.



- Judge top-weighted document.
- Update weights to reflect new info.

$$AP_1 = \frac{1}{R}\left(x_A + x_B\frac{(x_A+x_B)}{2} + x_C\frac{(x_A+x_B+x_C)}{3} + \cdots\right)$$

$$AP_2 = \frac{1}{R}\left(x_D + x_F\frac{(x_D+x_F)}{2} + x_G\frac{(x_D+x_F+x_G)}{3} + \cdots\right)$$

$$\Delta AP = AP_1 - AP_2$$

Greatest-weight documents generally at a high rank in one system and a low rank in the other.

Expected Mean Average Precision

- Let X_i be a random variable representing the relevance of document i.
- Let $p_i = P(X_i = 1)$.
- Then:

$$E[AP] = \frac{1}{\sum p_i}\left(\sum_{i=1}^{n}\frac{1}{i}p_i + \sum_{i=1}^{n}\sum_{j>i}\frac{1}{j}p_i p_j\right)$$

$$EMAP = \frac{1}{T}\sum E[AP]$$

- Probabilities p_i estimated using expert aggregation (Carterette 2007).

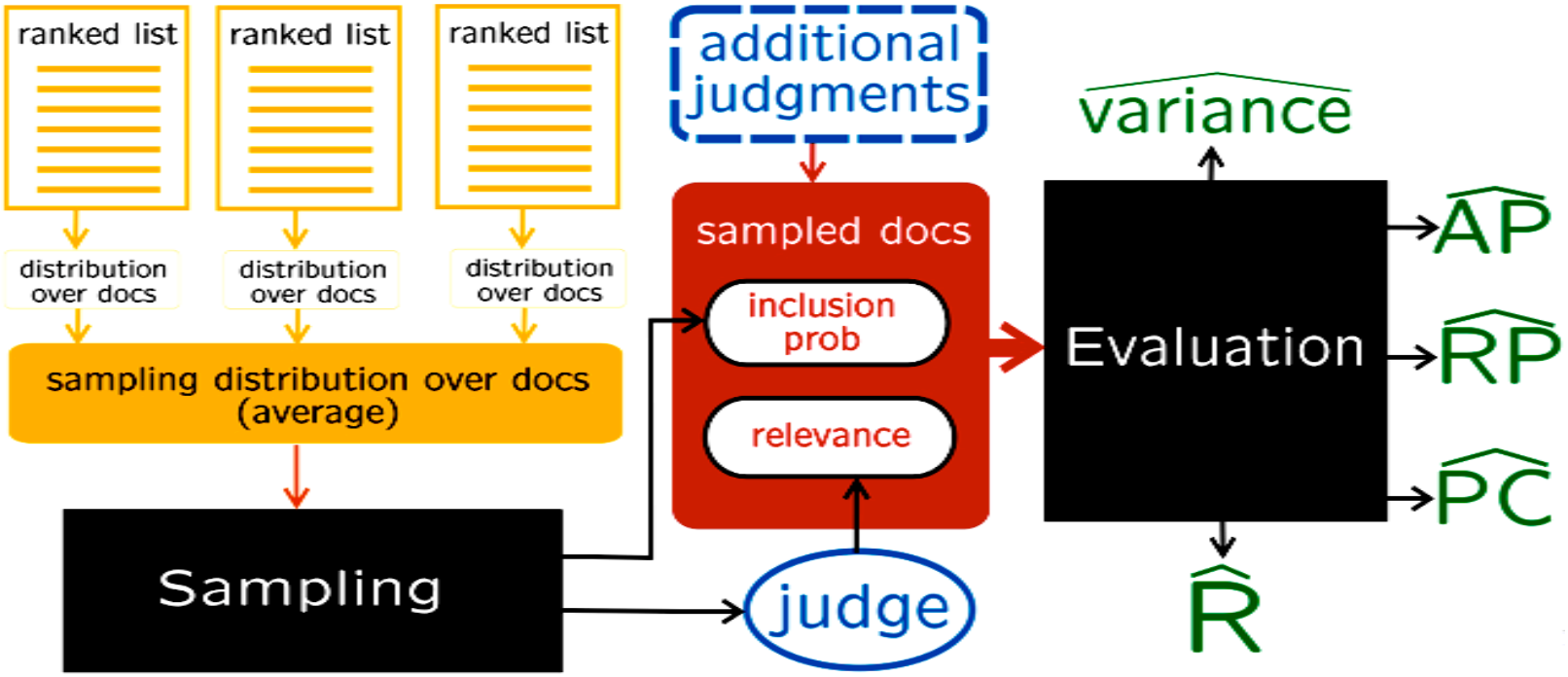
NEU statAP Method

- *Goal*: unbiased, low variance estimates of AP, ...
- *Method*: statistical sampling and evaluation
  - survey theory, market research, medical studies, ...
- *Analogy*: election forecasting
  - implicit evaluation distribution
    - often uniform
  - explicit sampling distribution
    - designed for accuracy (low variance)
    - inclusion probability measures "sampling bias"
  - estimator
    - given sample and inc. prob., produces unbiased estimates

NEU statAP Method

- three independent modules
  - each of them can be chosen in many ways
  - central: the sample (relevance + incl prob) a.k.a. probabilistic qrel

- 1: prior
- 2: sampling
- 3: evaluation

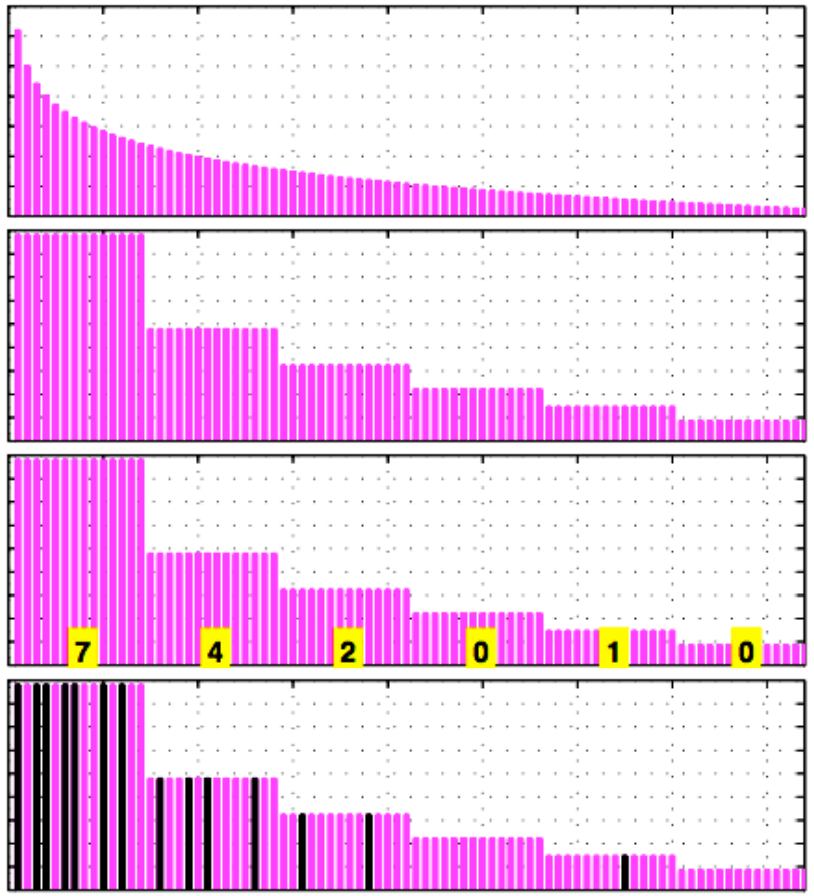


NEU statAP Sampling

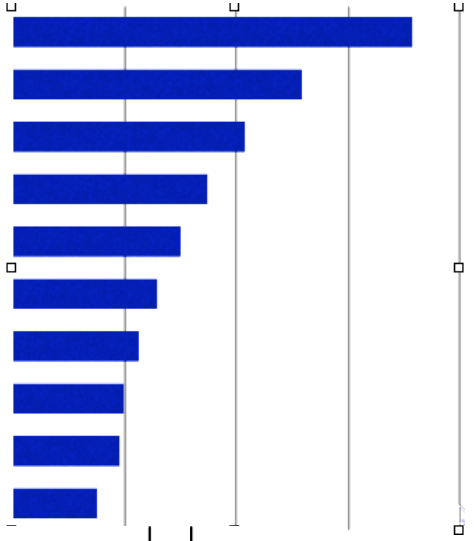
- given a set of ranked lists, choose a prior of relevance over documents considering ranks

- sample in 3 stages:

- group the docs in buckets of size m= sample size desired (m=14 in the example)
- sample the buckets with repetition m times according with cumulative bucket weight (register the hits)
- randomly pick in each bucket a number of docs equal with the number of hits registered at step two. The inclusion probability of each doc is the cumulative weight of the bucket containing that doc.



Sampling Prior



- Define a weight associated with a rank in a list (|s|=length of list s).

- Prior at rank r is the sum of weights accumulated by a document over all ranked lists:

$$w_s(r) = \frac{1}{2|s|}\left(1 + \frac{1}{r} + \frac{1}{r+1} + ... + \frac{1}{|s|}\right) \approx \frac{1}{2|s|}\log\frac{|s|}{rank(d,s)}$$

- Document prior is then:

$$Prior(d) \approx \sum_s w_s(rank(d,s))$$

NEU statAP Evaluation

- Given a sample of docs and associated relevance and inclusion probabilities $\{rel_k, \pi_k\}$, we apply survey theory to estimate:
  - Precision at rank r:

$$\widehat{PC}(r) = \frac{1}{r} \sum_{rank(k) \leq r} \frac{rel_k}{\pi_k}$$

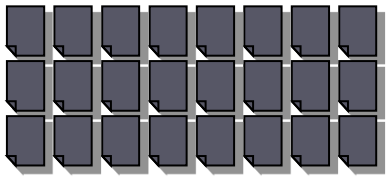
  - Number of relevant docs (in collection):

$$\widehat{R} = \sum_{rel_k=1} \frac{1}{\pi_k}$$

  - AP:

$$\widehat{AP} = \frac{\sum_{r_k=1} \widehat{PC}(rank(k))/\pi_k}{\sum_{rel_k=1} 1/\pi_k}$$

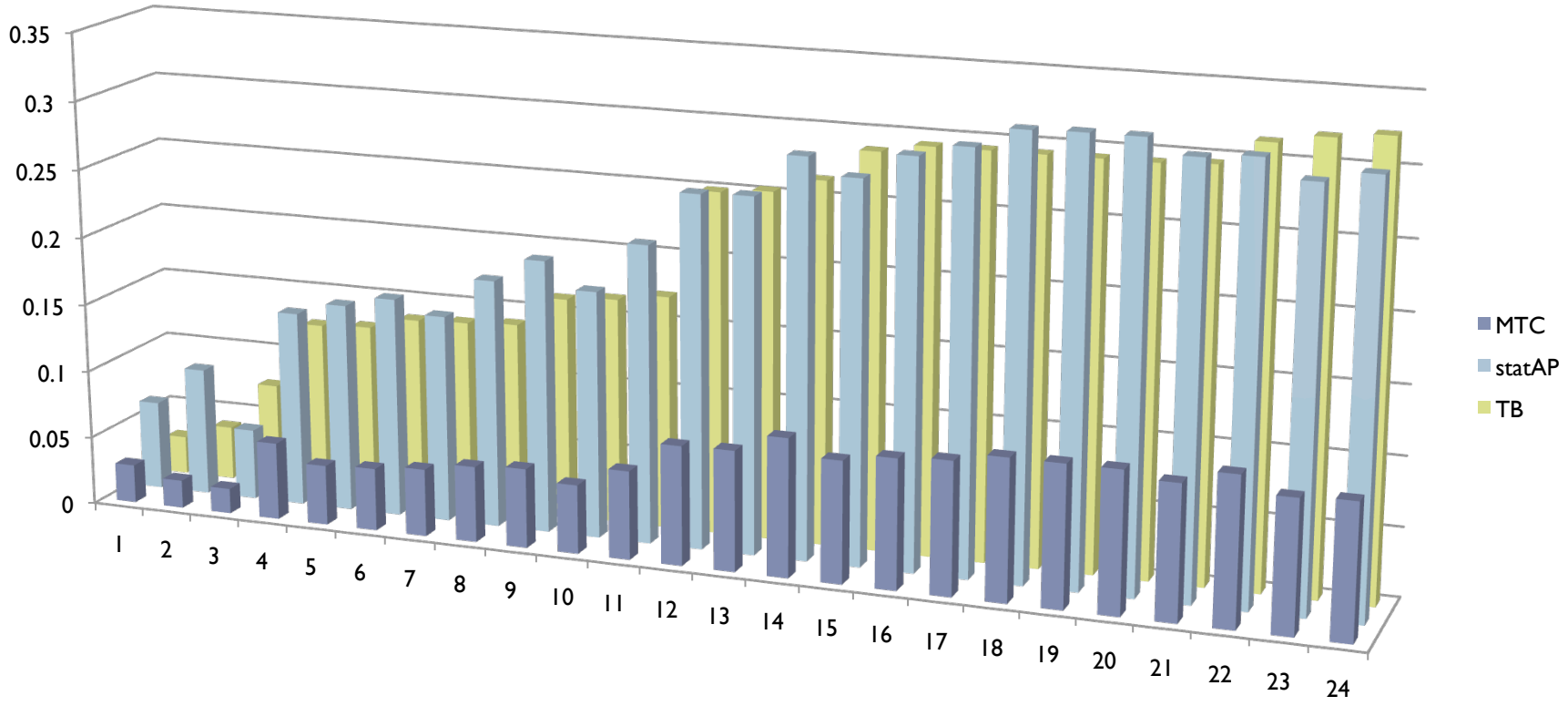
Relevance Judgments



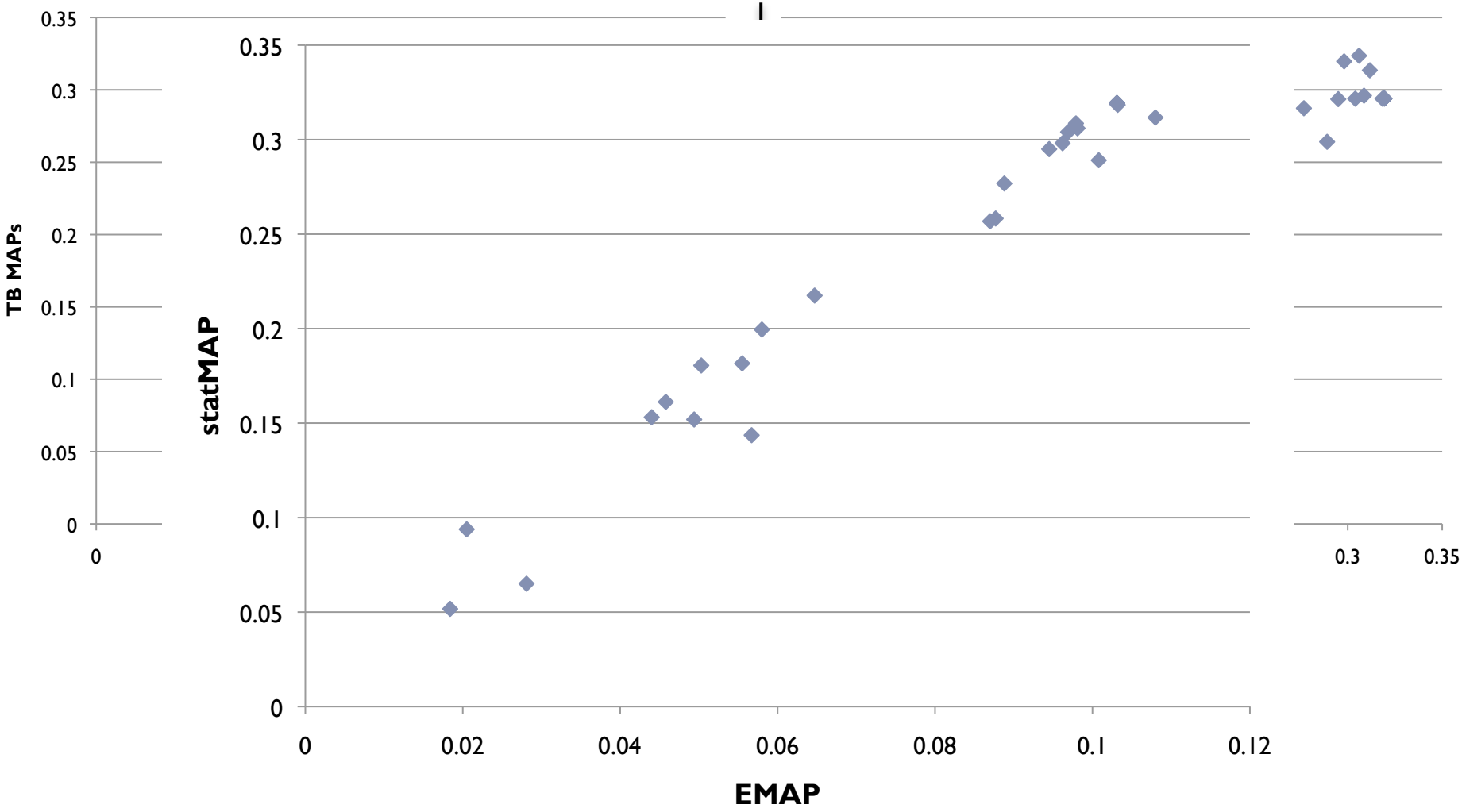
- 1,692 of the 10,000 queries judged.
  - 429 by MTC (UMass).
  - 443 by statAP (NEU).
  - 801 by alternation.
- 69,730 total judgments, roughly 40 per query.
  - Comparable to past years' totals with 50 queries and pooling.
- 10.62 relevant documents per query on average.
  - 25% relevant.
  - Greater percentage than usual.
- Assessors judged 40 documents in about 14 minutes.
  - About 21 seconds per judgment.

Results

- "Baseline": TREC queries 701-850.
  - "Full" judgments.
  - Seeded into 10,000 sampled queries.



Comparison of Mean Scores



Analysis

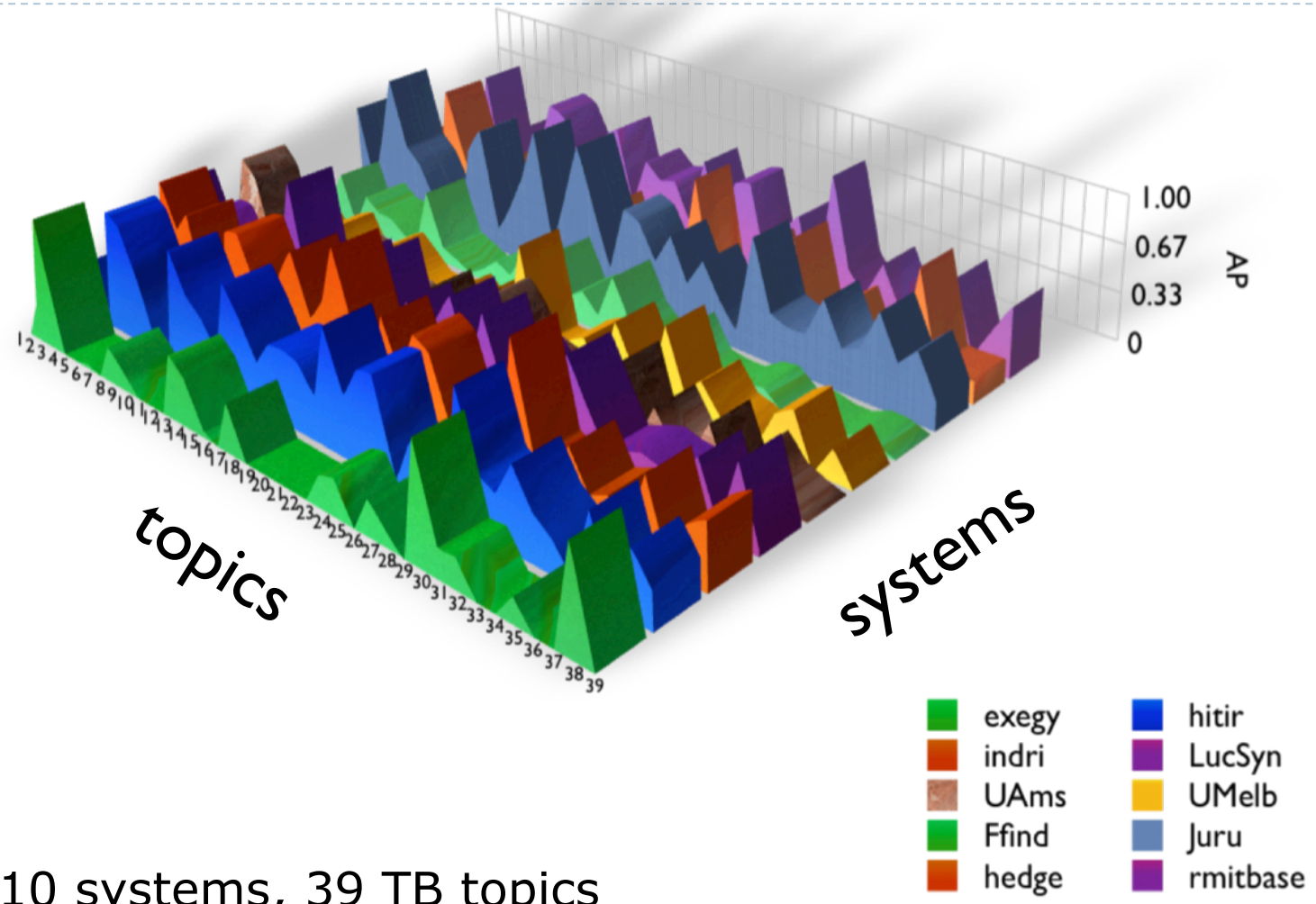
- Do we need thousands of queries to reach the same conclusions?

- Analysis of variance (ANOVA):
  - How much of the variance in MAP is due to the topics?
  - How many topics are needed to keep that variance low?
- Cost analysis:
  - How few queries and how few judgments per query are needed to reach a stable conclusion?

Efficiency Studies

- Systems run on a specific *set of topics*
  - Performance of each system measured by Mean Average Precision
- Systems run on a second *set of topics*
- How many queries are necessary so as
  - Ranking of systems is the same for both sets
  - Mean Average Precision values are the same for both sets
- How quickly in terms of queries one can arrive at accurate evaluation results

Variance in Average Precision values



10 systems, 39 TB topics

Average Precision Variance Components

$AP(s_k, q_t)$ $\sigma^2(AP(s,q))$

due to the system $\sigma^2(v_s)$

due to the topic $\sigma^2(v_q)$

due to the interaction between the system and the topic $\sigma^2(v_{sq})$

$E_s E_q[AP(s,q)] = \mu$

Experimental Setup

- Analysis of Variance
  - 429 topics exclusively selected by MTC with 40 relevance judgments per topic
  - 459 topics exclusively selected by statAP with 40 relevance judgments per topic
- The ratio of variance due to system and the total variance
- The ratio of variance due to system and the variance that affect the ranking of systems

Average Precision Variance Components

- statAP
  - $\sigma^2(v_s) = 0.0069$ or 11% of the total variance
  - $\sigma^2(v_q) = 0.0247$ or 40% of the total variance
  - $\sigma^2(v_{sq}) = 0.0311$ or 49% of the total variance

- MTC
  - $\sigma^2(v_s) = 0.0007$ or 9% of the total variance
  - $\sigma^2(v_q) = 0.0054$ or 69% of the total variance
  - $\sigma^2(v_{sq}) = 0.0017$ or 22% of the total variance

MAP Variance Components

$MAP(s_k, Q_l)$ $\sigma^2(MAP(s,Q))$

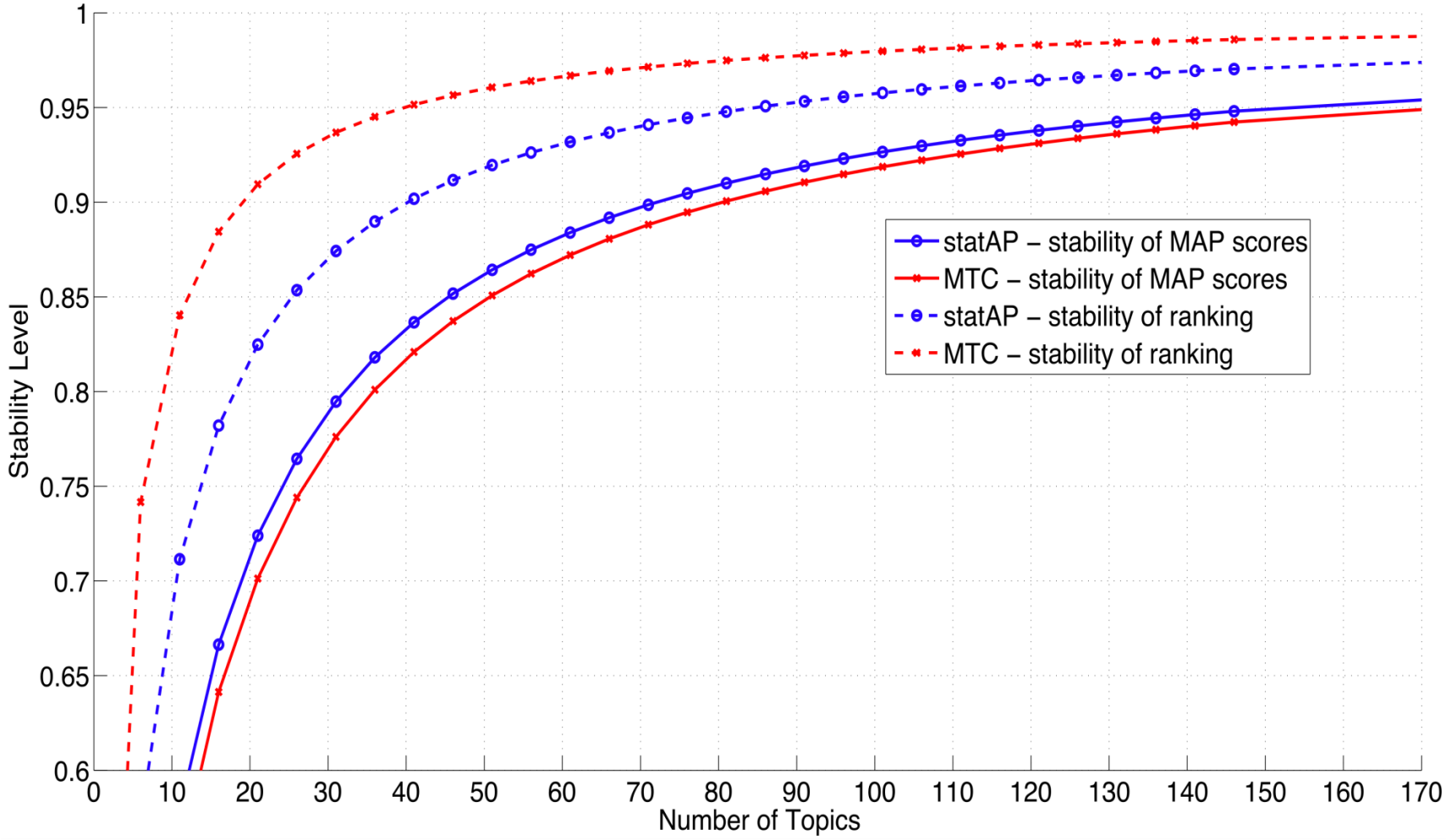
due to the system $\sigma^2(v_s)$

due to the *set of topics* $\sigma^2(v_Q) = \frac{\sigma^2(v_q)}{n_q}$

due to the interaction between the system and the *set of topics* $\sigma^2(v_{sQ}) = \frac{\sigma^2(v_{sq})}{n_q}$

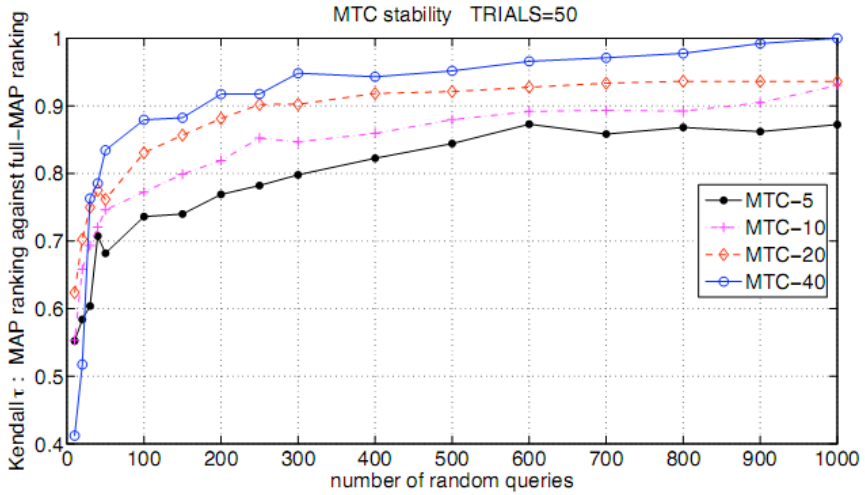
$E_s E_Q[MAP(s,Q)] = \mathrm{M}$

MAP Variance Components



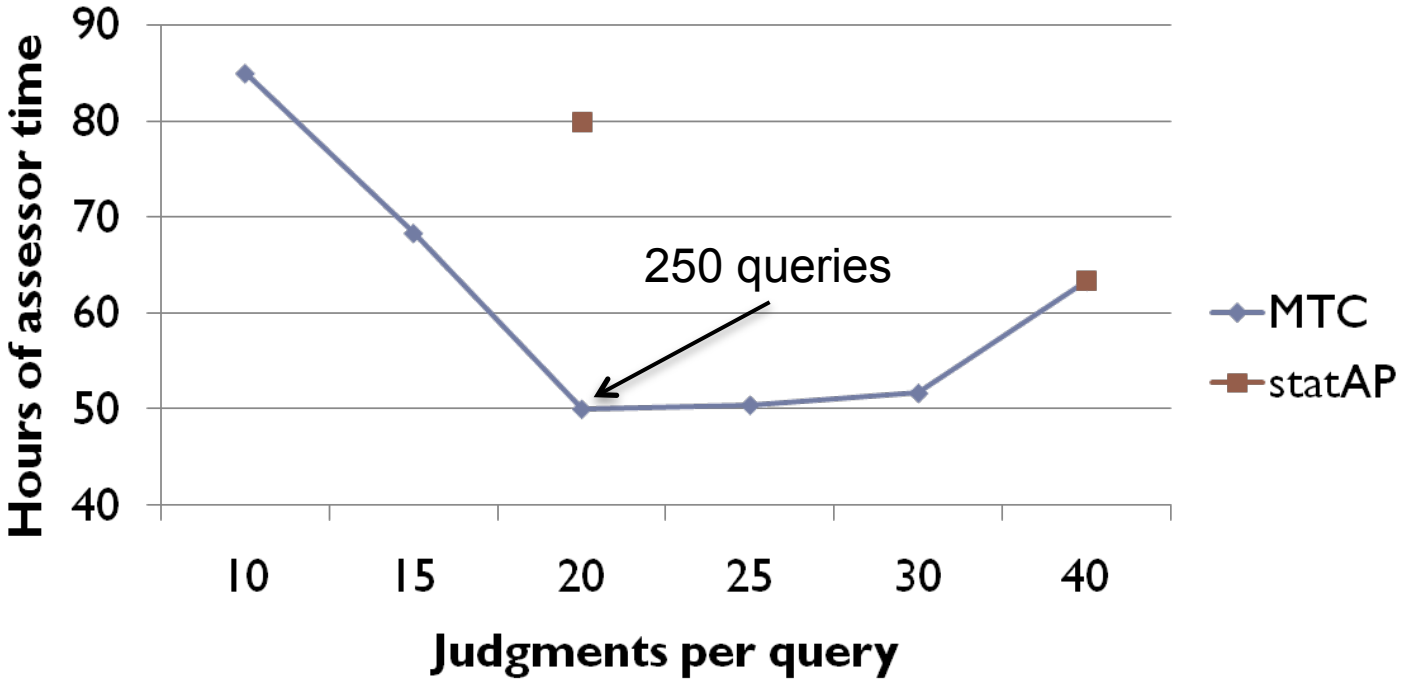
Cost Analysis

- What is the minimum cost needed to reach final result?
  - Or Kendall's tau = 0.9 with final result.
- Simulate judging with increasing numbers of queries and increasing numbers of judgments per query.
  - MTC can be stopped at any point.
  - statAP can use 20 judgments or 40 judgments per query.



Cost Analysis

- Estimate assessor time:
  - Time ≈ 5 min to develop query * # of queries
  - + 21s to judge a document * total # of judgments



Conclusion

- Low-cost methods reliably evaluate retrieval systems with very few judgments.
- Both methods accomplish their respective goals:
  - statAP more successfully estimates MAP.
  - MTC more successfully converges on a correct ranking.
- Both methods work with only a few hundred topics and a few dozen judgments per topic.