Advances in IR Evaluation

Ben Carterette Evangelos Kanoulas Emine Yilmaz

Yesterday's Outline

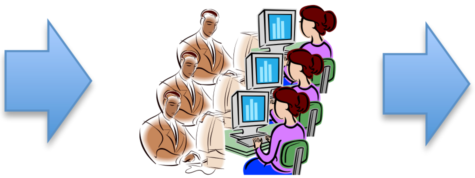
- Different evaluation methods
  - Interactive, on-line, off-line
- Off-line evaluation
- Basic measures of effectiveness
- Test collections
  - Judgment Effort

How many documents to judge?

- Many measures are based on
  - recall : "*out of all good docs in the collection how many did the algo find*"
  - all good documents in the collection need to be identified

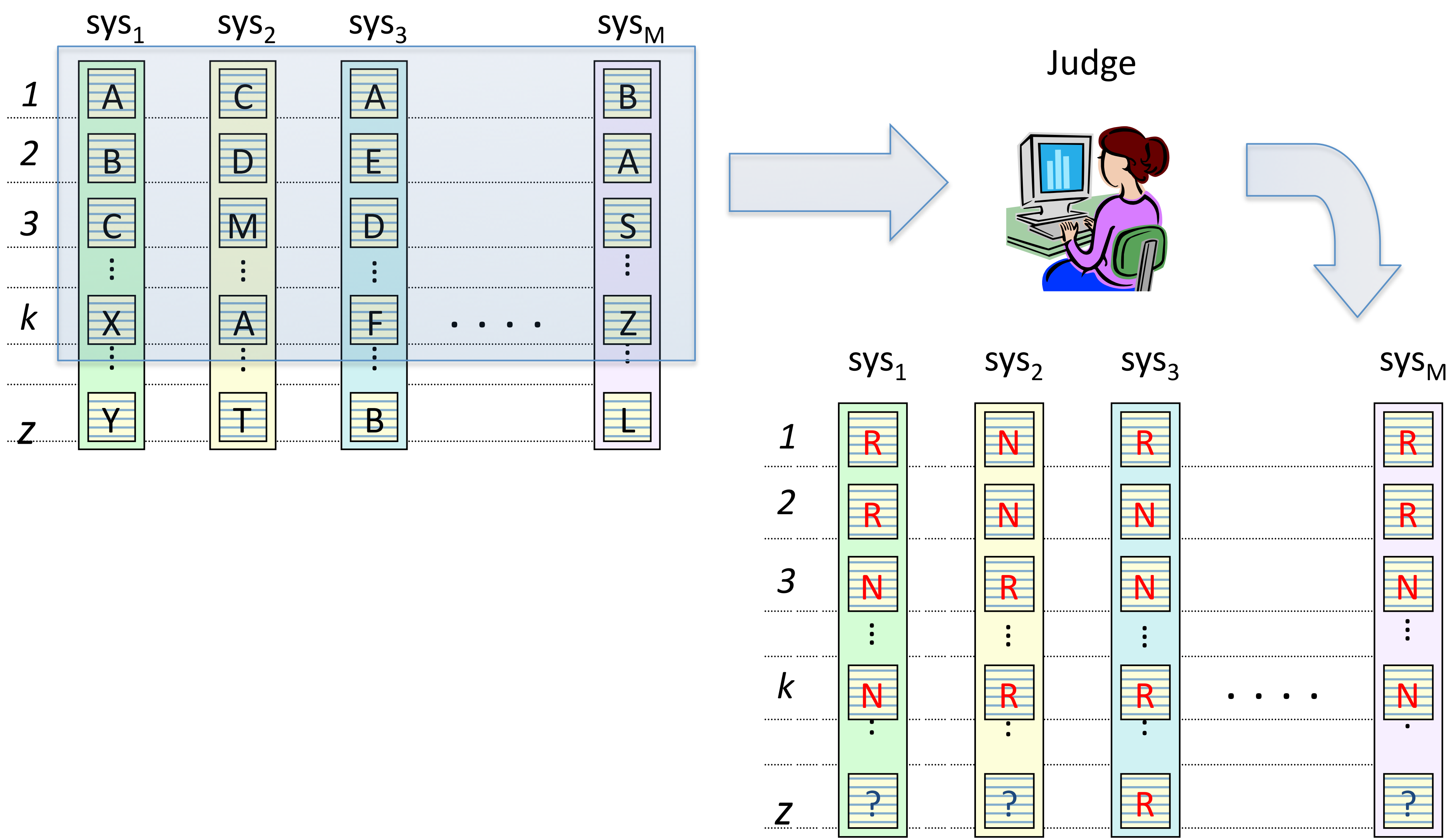
How many documents to judge?

- New measures are top-heavy
  - e.g. % of good docs in the first page of results



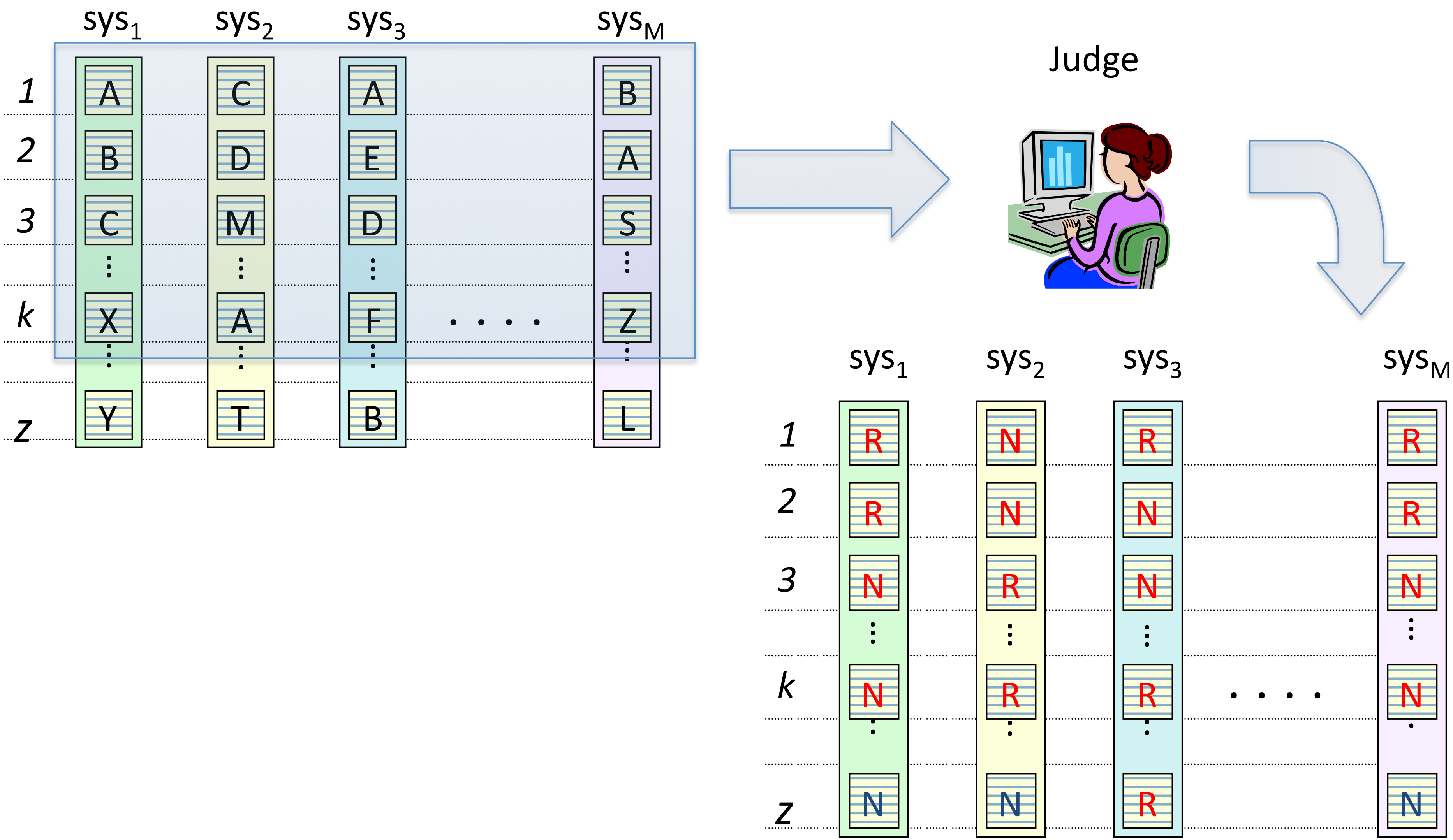
Retrieved List by SYS1		Retrieved List by FUTURE SYSTEM SYS2	
A	R	K	?
B	N	B	N
C	R	L	?
D	N	M	?
E	N	E	N
F	R	N	?
G	N	O	?
H	N	P	?
I	N	I	N
J	R	Q	?

Depth-k pooling
(TREC Standard Setup)



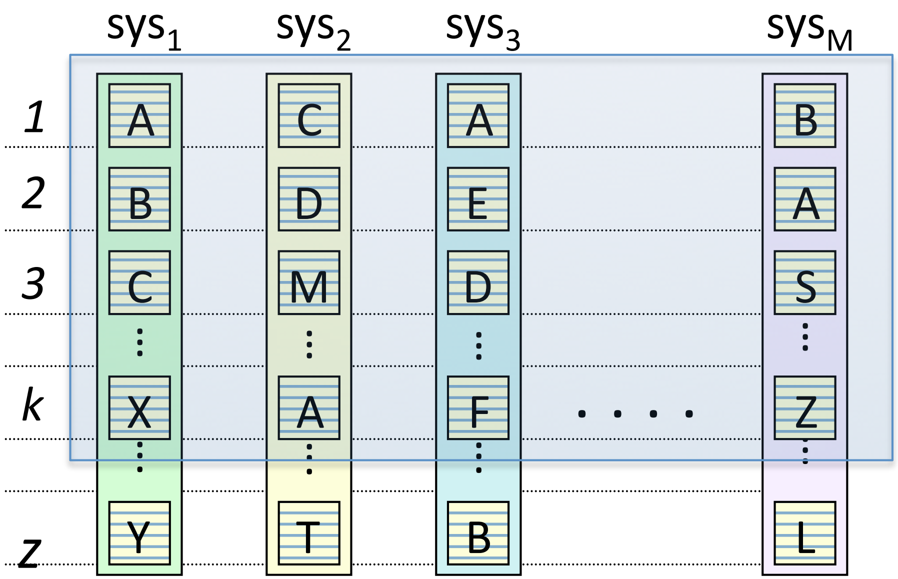
Depth-k pooling

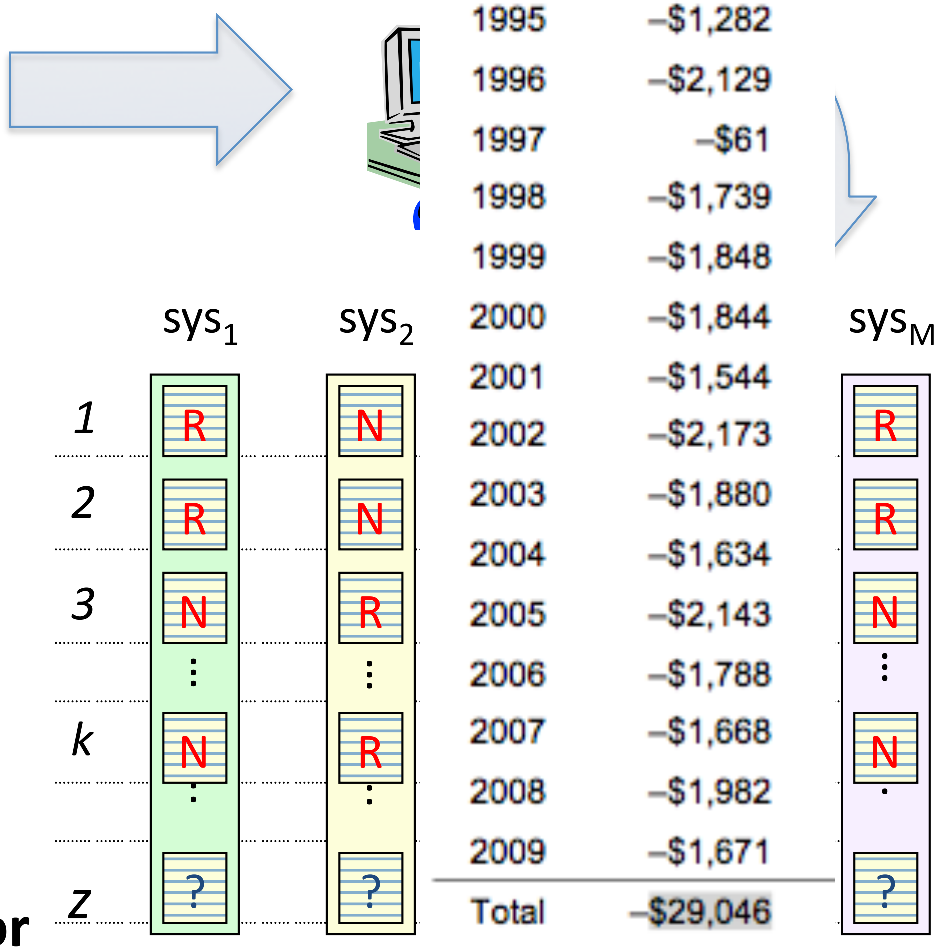
(TREC Standard Setup)



Depth-k pooling

(TREC Standard Setup





Year	Total TREC Investment Costs (thousands $2009)
1991	–$753
1992	–$713
1993	–$674
1994	–$1,522
1995	–$1,282
1996	–$2,129
1997	–$61
1998	–$1,739
1999	–$1,848
2000	–$1,844
2001	–$1,544
2002	–$2,173
2003	–$1,880
2004	–$1,634
2005	–$2,143
2006	–$1,788
2007	–$1,668
2008	–$1,982
2009	–$1,671
Total	–$29,046

TREC 8 test collection

- **50 topics, depth-100 pooling =>**
  86,830 judgments
- **30 sec per judgment =>**
  724 hours => 18 weeks of labor

Course Outline

- Intro to evaluation
  - Evaluation methods, test collections, measures, comparable evaluation
- Low cost evaluation
- Advanced user models
  - Web search models, novelty & diversity, sessions
- Reliability
  - Significance tests, reusability
- Other evaluation setups

Today's Outline

- Low cost evaluation

1. Depth-k pooling (standard method)

2. Evaluating without judgments (automatic eval)
3. Finding relevance documents as quickly as possible

4. Computing measures with incomplete judgments
5. Estimating measures
6. Inferring relevance judgments

Low-Cost Evaluation (1)

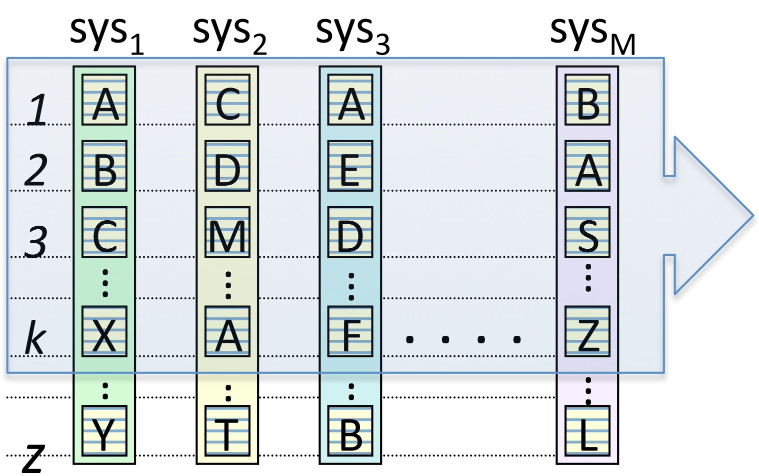
- Depth-k pooling

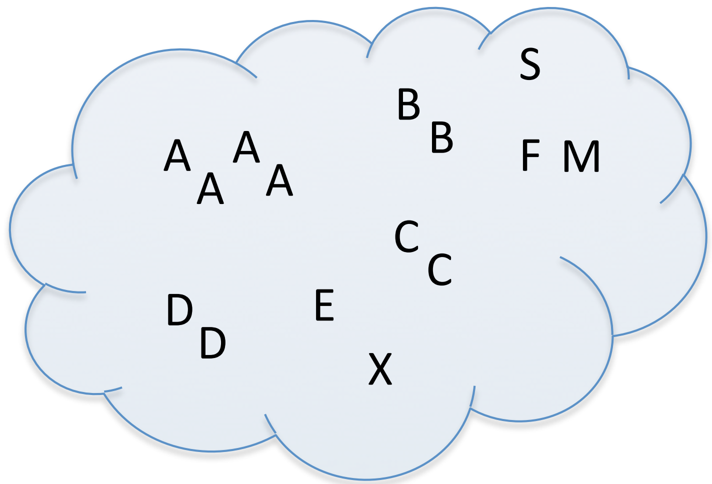
- Evaluation with no relevance judgments
  - Random relevance
    - Soboroff et al SIGIR01, Aslam and Savell SIGIR03, Wu and Crestani SAC03, Nuray and Can IPM06, Efron ECIR09, Hauff et al ECIR10, …

Low-Cost Evaluation (1)

- Depth-k pooling

- Evaluation with no relevance judgments
  - Random relevance
    - Soboroff et al SIGIR01, Aslam and Savell SIGIR03, Wu and Crestani SAC03, Nuray and Can IPM06, Efron ECIR09, Hauff et al ECIR10, …

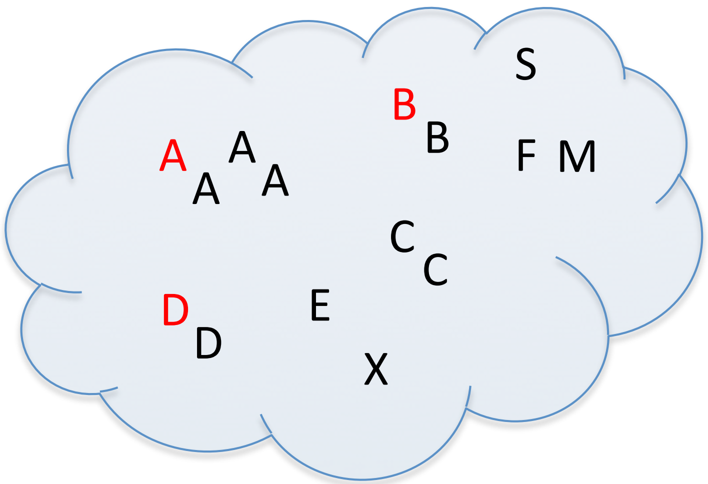


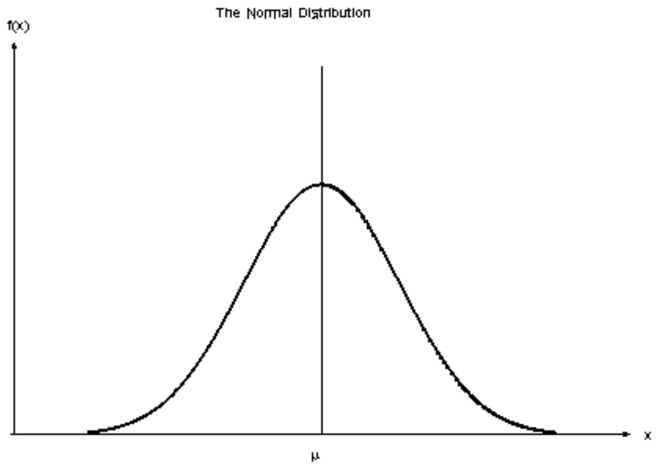


Low-Cost Evaluation (1)

- Depth-k pooling

- Evaluation with no relevance judgments
  - Random relevance
    - Soboroff et al SIGIR01, Aslam and Savell SIGIR03, Wu and Crestani SAC03, Nuray and Can IPM06, Efron ECIR09, Hauff et al ECIR10, …

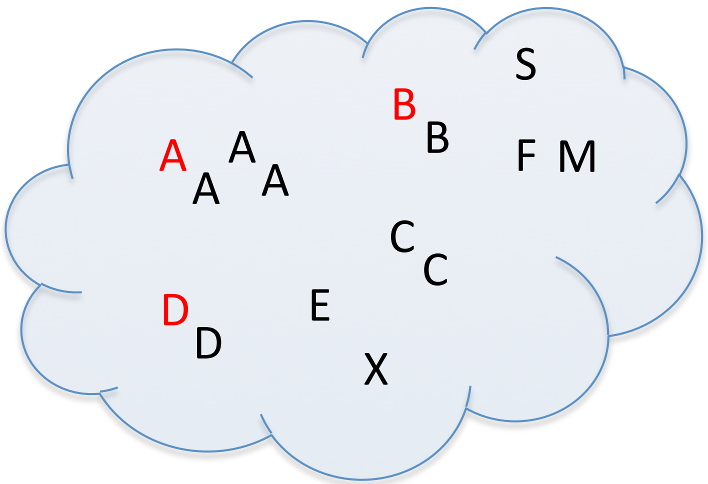


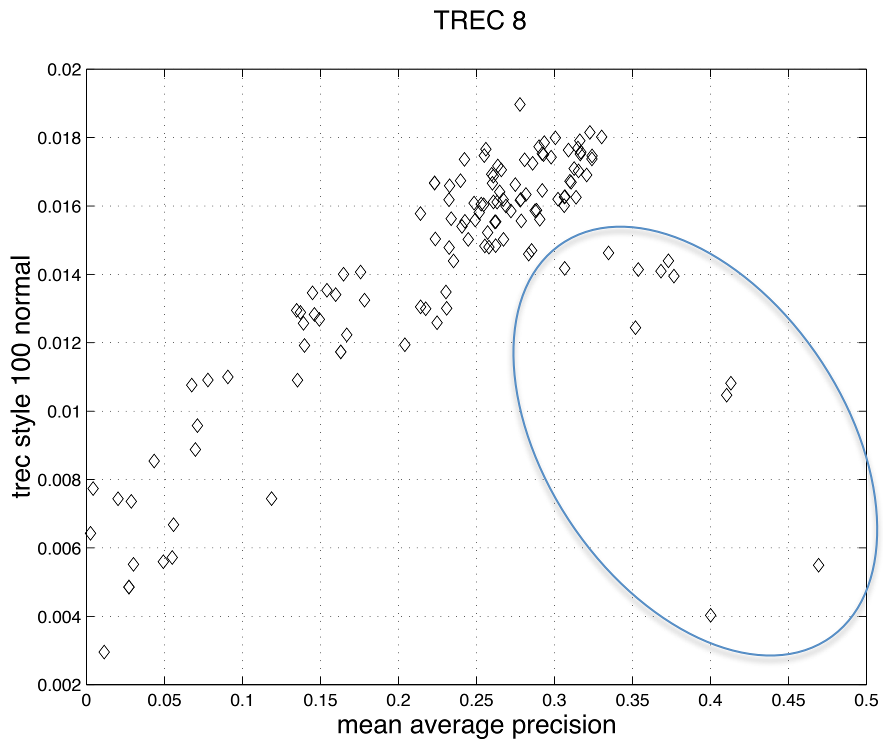


percentage of rel. docs.

Low-Cost Evaluation (1)

- Depth-k pooling
- Evaluation with no relevance judgments
  - Random relevance
    - Soboroff et al SIGIR01, Aslam and Savell SIGIR03, Wu and Crestani SAC03, Nuray and Can IPM06, Efron ECIR09, Hauff et al ECIR10, …

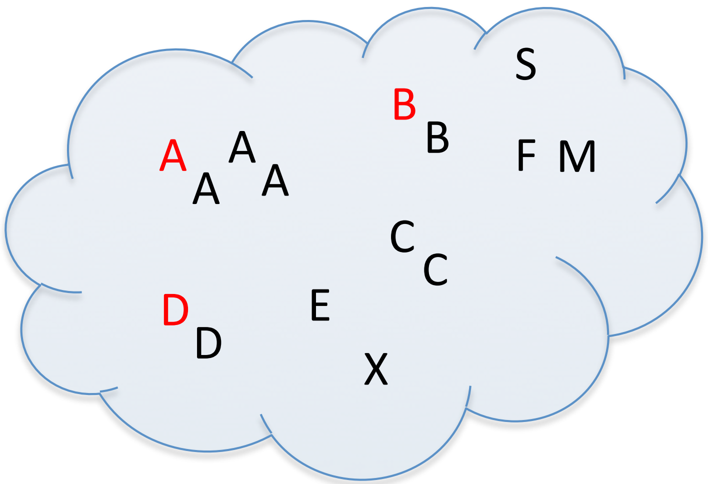




Low-Cost Evaluation (1)

- Depth-k pooling

- Evaluation with no relevance judgments
  - Random relevance
    - Soboroff et al SIGIR01, Aslam and Savell SIGIR03, Wu and Crestani SAC03, Nuray and Can IPM06, Efron ECIR09, Hauff et al ECIR10, …



"Tyranny of the masses"
[Aslam and Savell SIGIR03]

Low-Cost Evaluation (1)

- Depth-k pooling

- Evaluation with no relevance judgments
  [Wu and Crestani SAC03]
  - Rank systems by “reference count” : how many of the rest of the systems retrieved
    - the same documents
    - at similar ranks
    - with larger weight given towards the top of the list

Low-Cost Evaluation (1)

- Depth-k pooling

- Evaluation with no relevance judgments
  [Nuray and Can IPM06]
  - Good subset of p% of systems – the ones most different from the average
  - Merge documents by Condorcet voting
  - Consider top s% relevant.

Low-Cost Evaluation (1)

- Depth-k pooling

- Evaluation with no relevance judgments
  [Efron ECIR09, JASIST10]
  – Given a topic *t*
    - generate a small set of query aspects $\{a_i\}$
    - employ a single IR system *S*
    - run *S* over all aspects a_i
    - consider the union of the top *k* documents relevant

  – Better correlation with actual ranking than Soboroff et al.
    - Only automatic runs were tested [Hauff ECIR10, SIGIR10]

Today's Outline

- Low cost evaluation

1. Depth-k pooling (standard method)

2. Evaluating without judgments (automatic eval)
3. Finding relevance documents as quickly as possible

4. Computing measures with incomplete judgments
5. Estimating measures
6. Inferring relevance judgment

Low-Cost Evaluation (2)

- Alternatives to pooling
  - Zobel SIGIR98, Cormack et al SIGIR98, Aslam et al CIKM03, Moffat et al SIGIR07, …

Low-Cost Evaluation (2)

- Alternatives to pooling

  Interactive Searching and Judging [Cormack et al SIGIR98]

  – Assessor issue multiple searches per topic on a single IR system

  – Given a topic form and issue a query
  – Judge the results until the frequency of new relevant documents found drops to a certain level
  – Reformulate the query and repeat

Low-Cost Evaluation (2)

- Alternatives to pooling

  Interactive Searching and Judging [Cormack et al SIGIR98]

  - Implicitly implemented by TREC through *manual runs*
  - Explicitly used by some tracks in CLEF [Clough et al CLEF05] and NTCIR [Kuriyama et al IR02]
  - Used in Filtering Test Collection TREC 2002
    - Assessors issue a query over 4 IR systems (7 IR techniques/runs)
    - Judge the top 100 documents
    - Use relevance feedback and query expansion and reissue the query

  - Similar to Efron's query aspects [Efron ECIR09]

Low-Cost Evaluation (2)

- Alternatives to pooling
  [Zobel SIGIR98]
  – Some topics have more relevant documents than others
  – Focus assessor effort on those topics

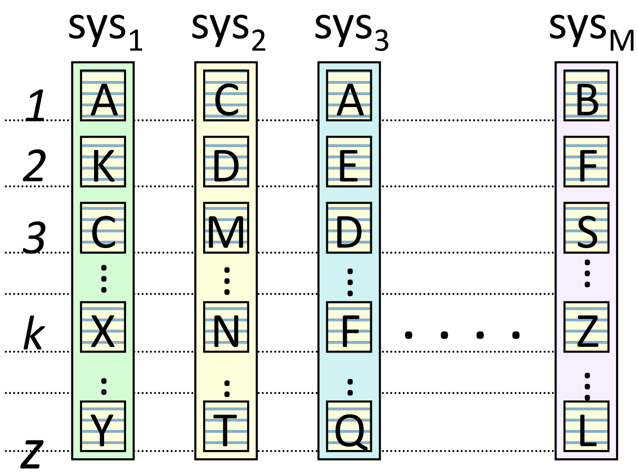
Low-Cost Evaluation (2)

- Alternatives to pooling

  Move-to-Front Pooling [Cormack et al SIGIR98]
  - Some systems retrieve more relevant documents than others
  - Focus assessor effort on those systems (*local* MTF)
  - Some topics have more relevant documents than others
  - Focus assessor effort both on "easy" topics and on "good" systems (*global* MTF)

Low-Cost Evaluation (2)

- Alternatives to pooling
  Move-to-Front Pooling [Cormack et al SIGIR98]

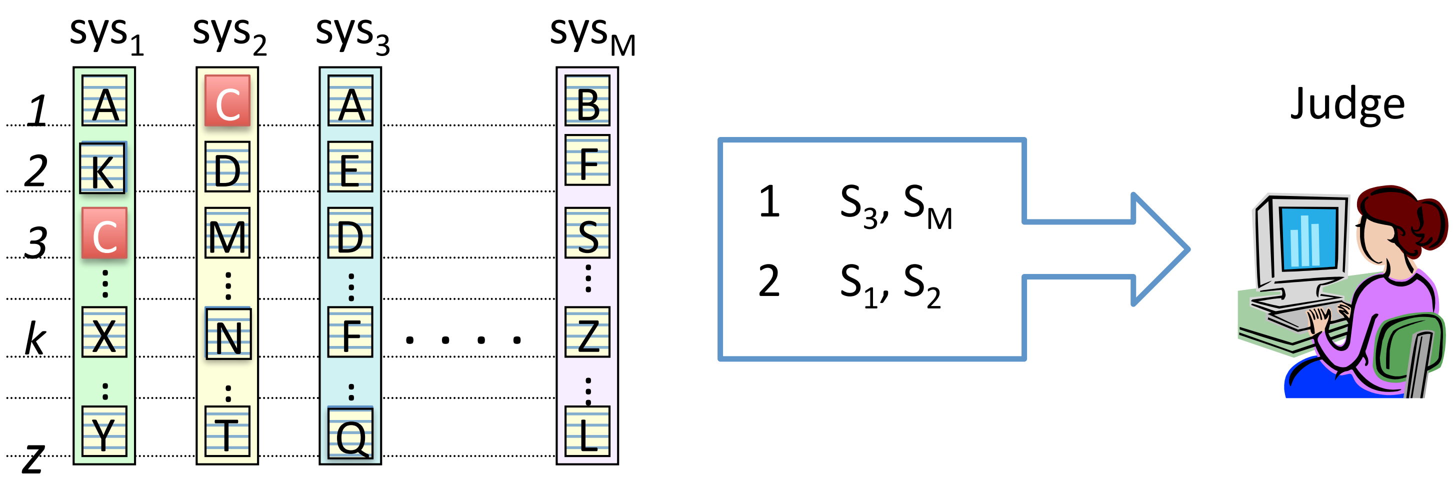


Judge

Low-Cost Evaluation (2)

- Alternatives to pooling

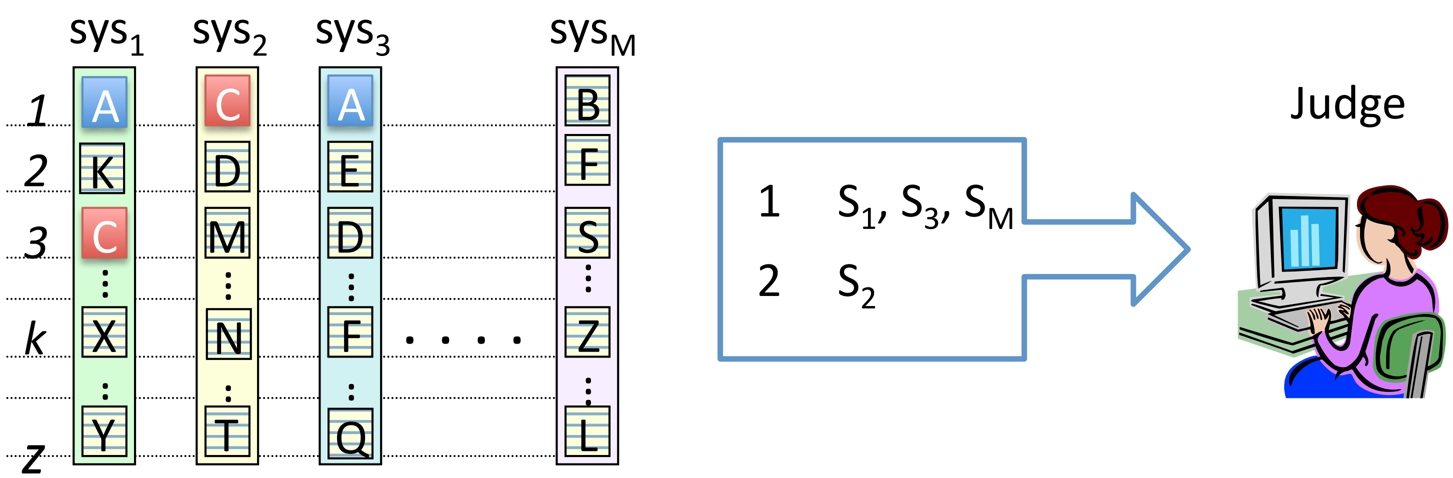
  Move-to-Front Pooling [Cormack et al SIGIR98]



Low-Cost Evaluation (2)

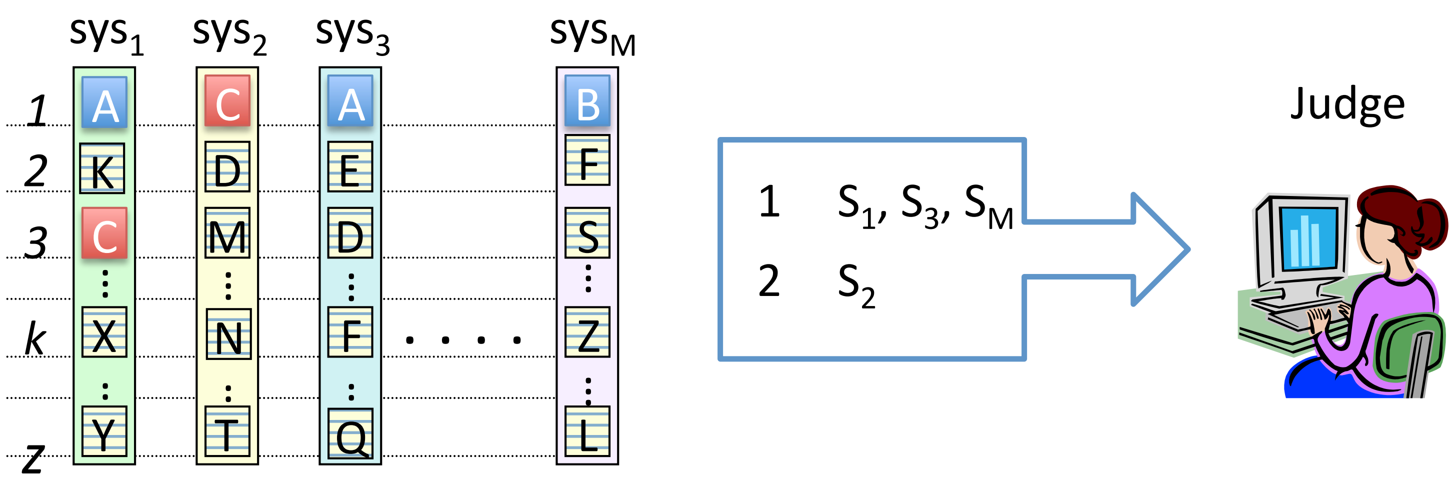
- Alternatives to pooling

  Move-to-Front Pooling [Cormack et al SIGIR98]



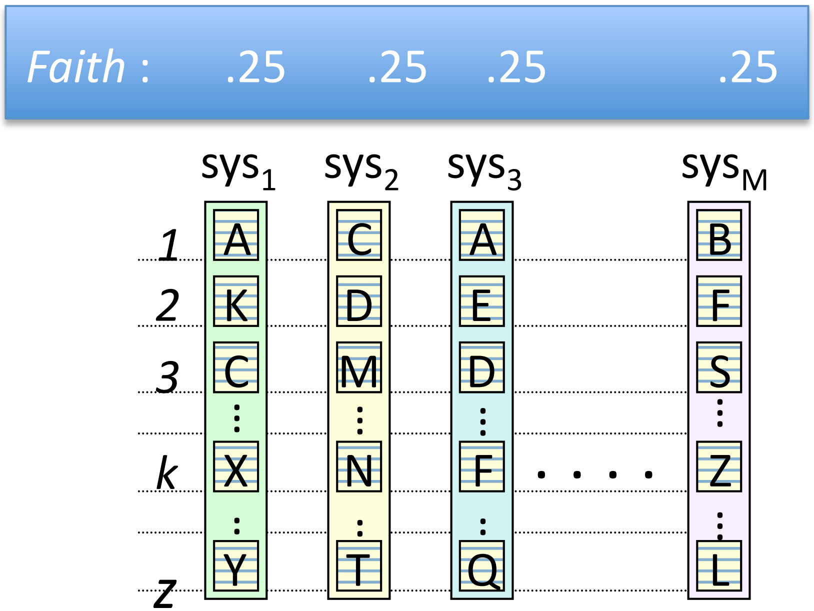
Low-Cost Evaluation (2)

- Alternatives to pooling
  Move-to-Front Pooling [Cormack et al SIGIR98]



Low-Cost Evaluation (2)

- Alternatives to pooling
  Hedge [Aslam et al CIKM03]
  - Each underlying IR system is an "expert" providing "advice" about the relevance

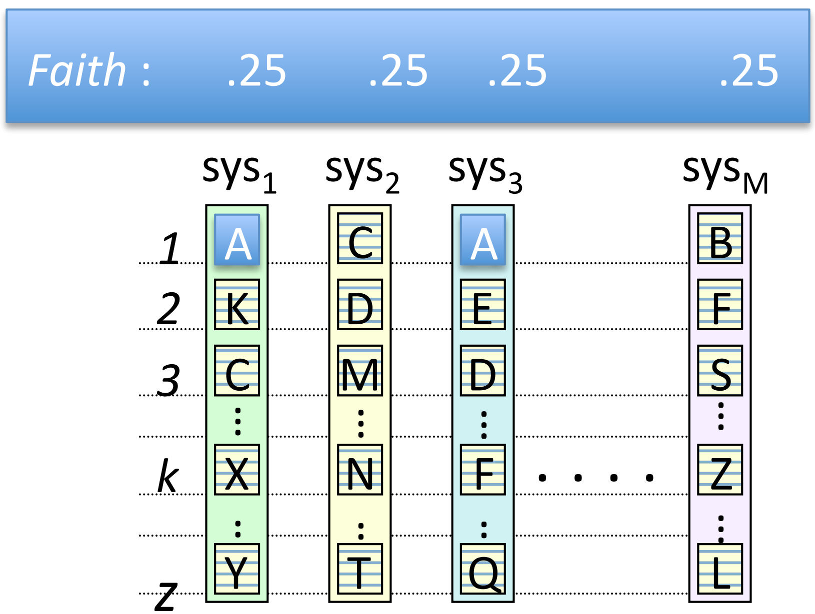


Low-Cost Evaluation (2)

- Alternatives to pooling

  Hedge [Aslam et al CIKM03]

  - Each underlying IR system is an "expert" providing "advice" about the relevance



- Consider total precision (sum of precisions at all documents)
- How much have we gained by A being relevant?

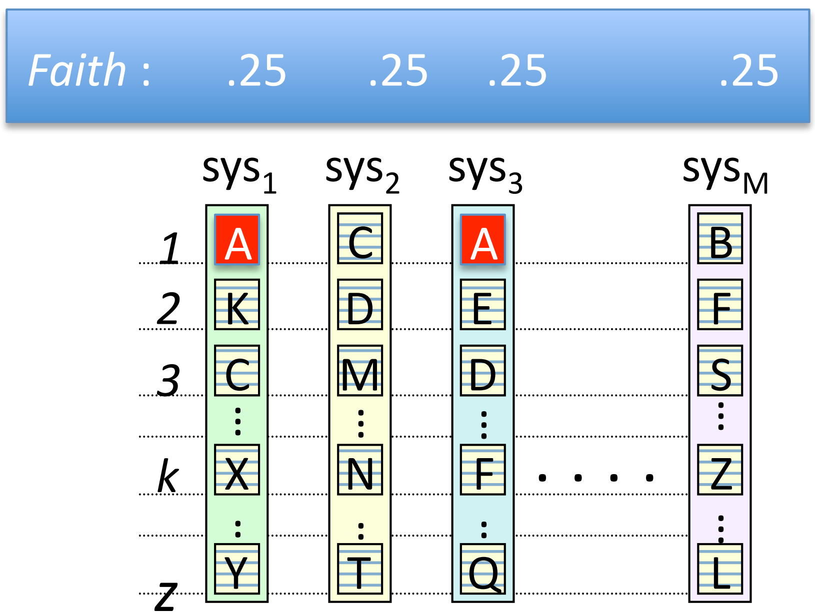
  GAIN = 1/1 + 1/2 + 1/3 + … + 1/N

- Update faith: w_1 to $w_0 * \beta^{-GAIN}$

Low-Cost Evaluation (2)

- Alternatives to pooling

  Hedge [Aslam et al CIKM03]
  - Each underlying IR system is an “expert” providing “advice” about the relevance



- Consider total precision (sum of precisions at all documents)
- How much have we gained by A being relevant?

  LOSS = 1/1 + 1/2 + 1/3 + … + 1/N

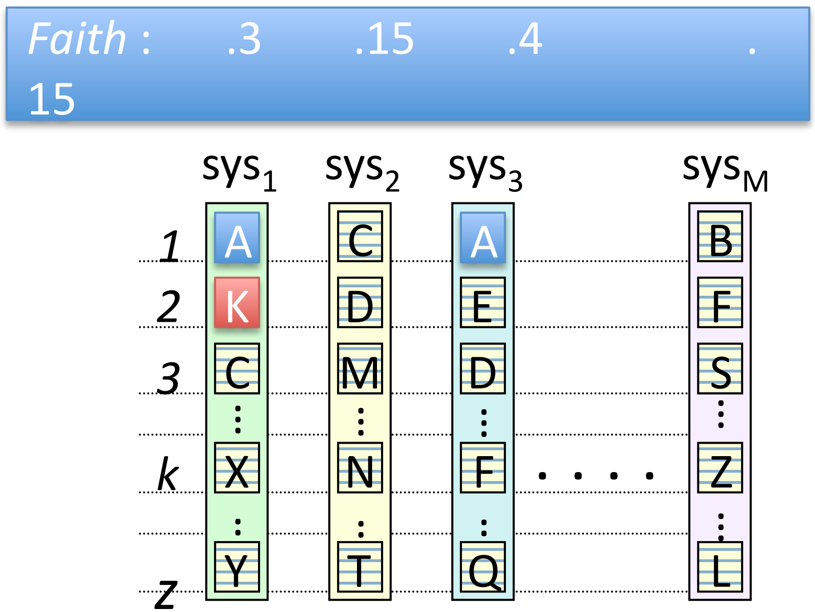
- Update faith: w_1 to $w_0 * \beta^{LOSS}$

Low-Cost Evaluation (2)

- Alternatives to pooling

  Hedge [Aslam et al CIKM03]

  - Each underlying IR system is an "expert" providing "advice" about the relevance



- Which document shall we pick next?

$$d = \underset{\text{d not labeled}}{\operatorname{argmax}} \left[\sum_{s=1}^{M} w_s^{t-1} \cdot GAIN(d,s \mid d = rel) \right]$$

Today's Outline

- Low cost evaluation

  1. Depth-k pooling (standard method)

  2. Evaluating without judgments (automatic eval)
  3. Finding relevance documents as quickly as possible

  4. Computing measures with incomplete judgments
  5. Estimating measures
  6. Inferring relevance judgments

Low-Cost Evaluation (3)

- Measures not robust to incomplete judgments
  - Buckley and Voorhees SIGIR06, Yilmaz and Aslam CIKM06, Bompada et al SIGIR07, Sakai SIGIR07

1.	R	R
2.	N	N
3.	R	N
4.	R	R
5.	N	N
6.	R	N
7.	N	N
8.	N	N
9.	R	R
10.	N	N

Low-Cost Evaluation (3)

- Standard evaluation measures not robust to incomplete judgments

  [Buckley and Voorhees SIGIR06, Bompada et al SIGIR07]

$$\textbf{bpref} = \frac{1}{R}\sum_{r}\left(1 - \frac{\textbf{number of } n \textbf{ above } r}{R}\right)$$

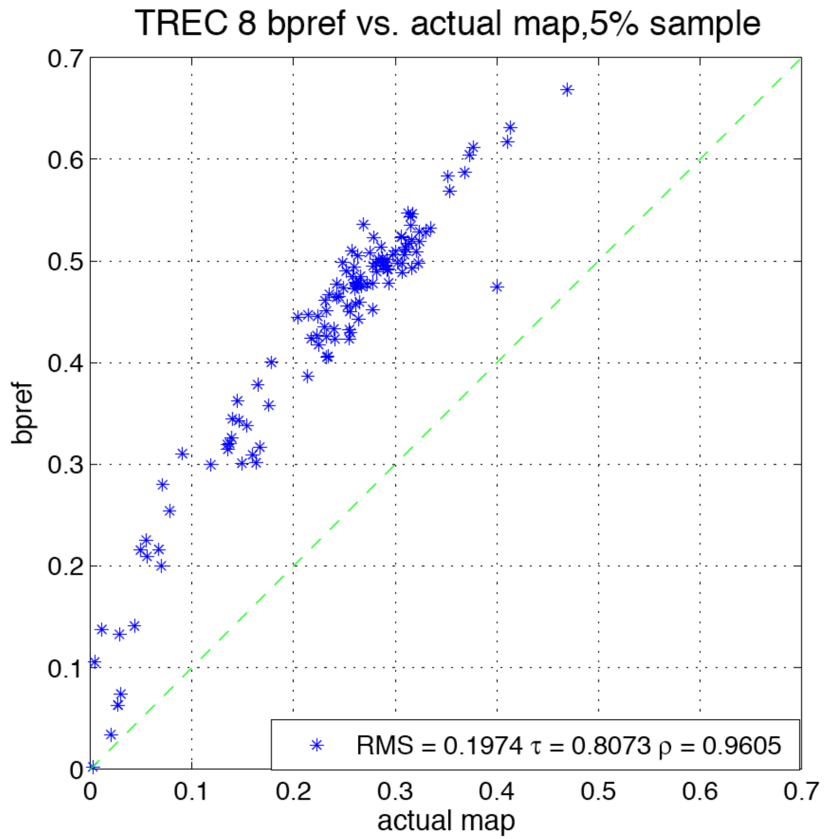
r : relevant document

R : number of judged relevant documents

n : member of top R judged nonrelevant documents

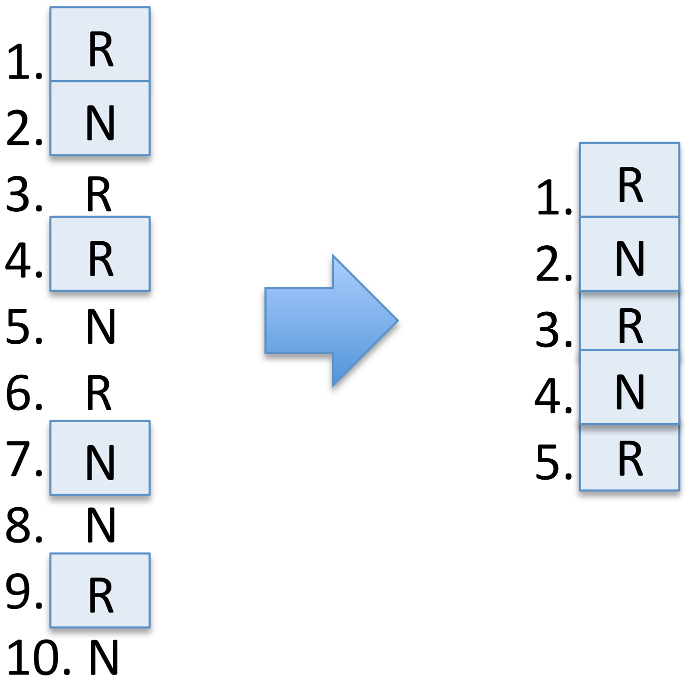
Low-Cost Evaluation (3)

- bpref :
  - More robust to incomplete relevance judgments than standard measures
  - Correlated with average precision when judgments are complete
  - Deviates from the value of AP when incomplete judgments



Low-Cost Evaluation (3)

- Induced measures
  – Yilmaz and Aslam CIKM06, Sakai SIGIR07



Low-Cost Evaluation (3)

- Induced measures
  - Yilmaz and Aslam CIKM06

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

1. R
2. N
3. R
4. N
5. R

$$\text{indAP} = \frac{1}{R} \sum_{r} \frac{\text{number } r \text{ upto } rank(r)}{rank(r)}$$

Low-Cost Evaluation (3)

- Induced measures
  - Yilmaz and Aslam CIKM06

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

→

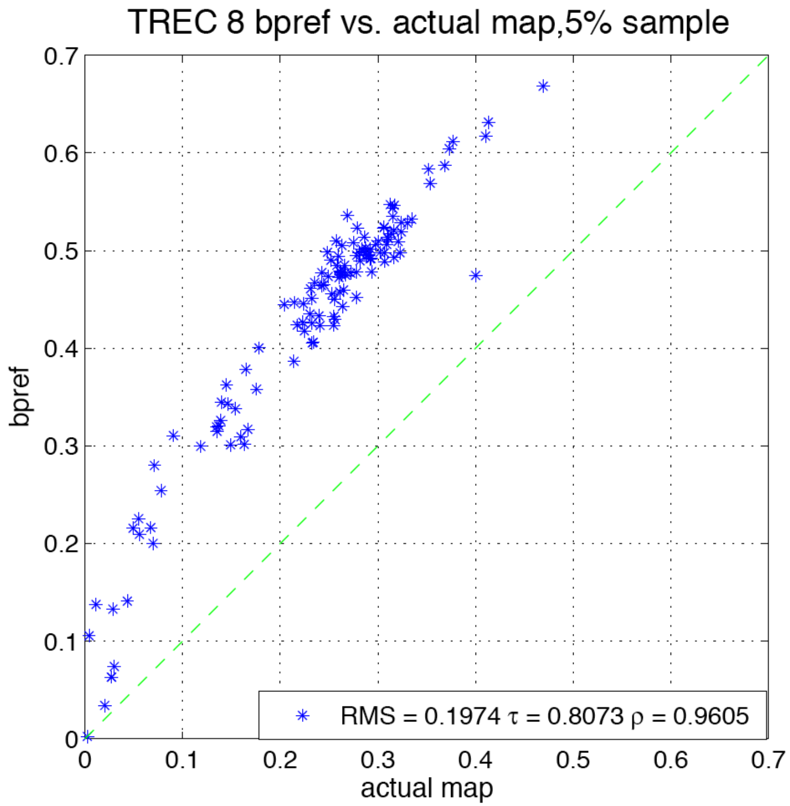
1. R
2. N
3. R
4. N
5. R

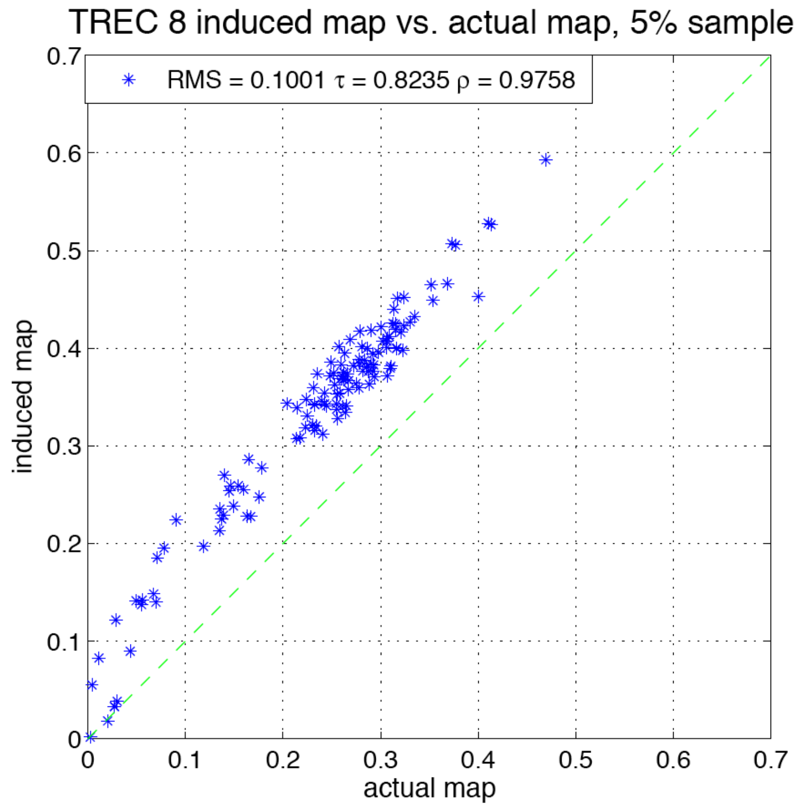
$$\text{indAP} = \frac{1}{R}\sum_{r}\left(1 - \frac{\text{number of } n \text{ above } r}{rank(r)}\right)$$

$$\text{bpref} = \frac{1}{R}\sum_{r}\left(1 - \frac{\text{number of } n \text{ above } r}{R}\right)$$

Low-Cost Evaluation (3)

- Induced measures
  - Yilmaz and Aslam CIKM06





Today’s Outline

- Low cost evaluation

1. Depth-k pooling (standard method)

2. Evaluating without judgments (automatic eval)
3. Finding relevance documents as quickly as possible

4. Computing measures with incomplete judgments
5. Estimating measures
6. Inferring relevance judgments

Low-Cost Evaluation (4)

- Estimating *measures* with less judgments
  - Aslam et al. SIGIR06, Yilmaz and Aslam CIKM06, Yilmaz et al SIGIR09

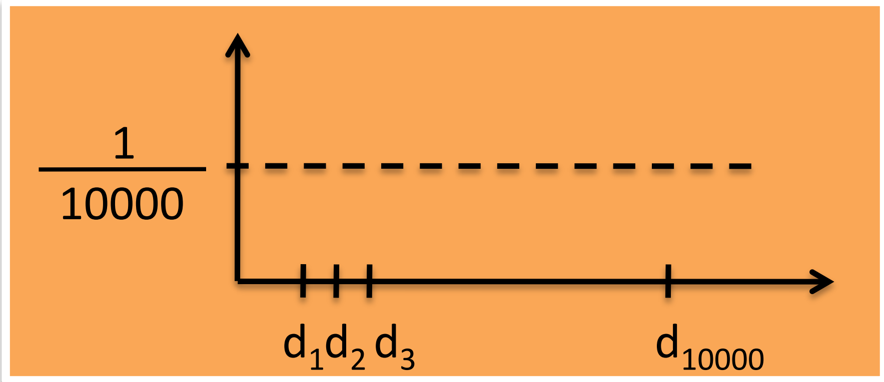
Sampling for Efficient Evaluation

- Sampling intuition:
- Consider a population of 10,000 animals
  - A percentage of which is sick
- I want to find the percentage of sick animals
  - Obvious solution : examine all 10,000
  - Return : #sick/10,000

Sampling for Efficient Evaluation

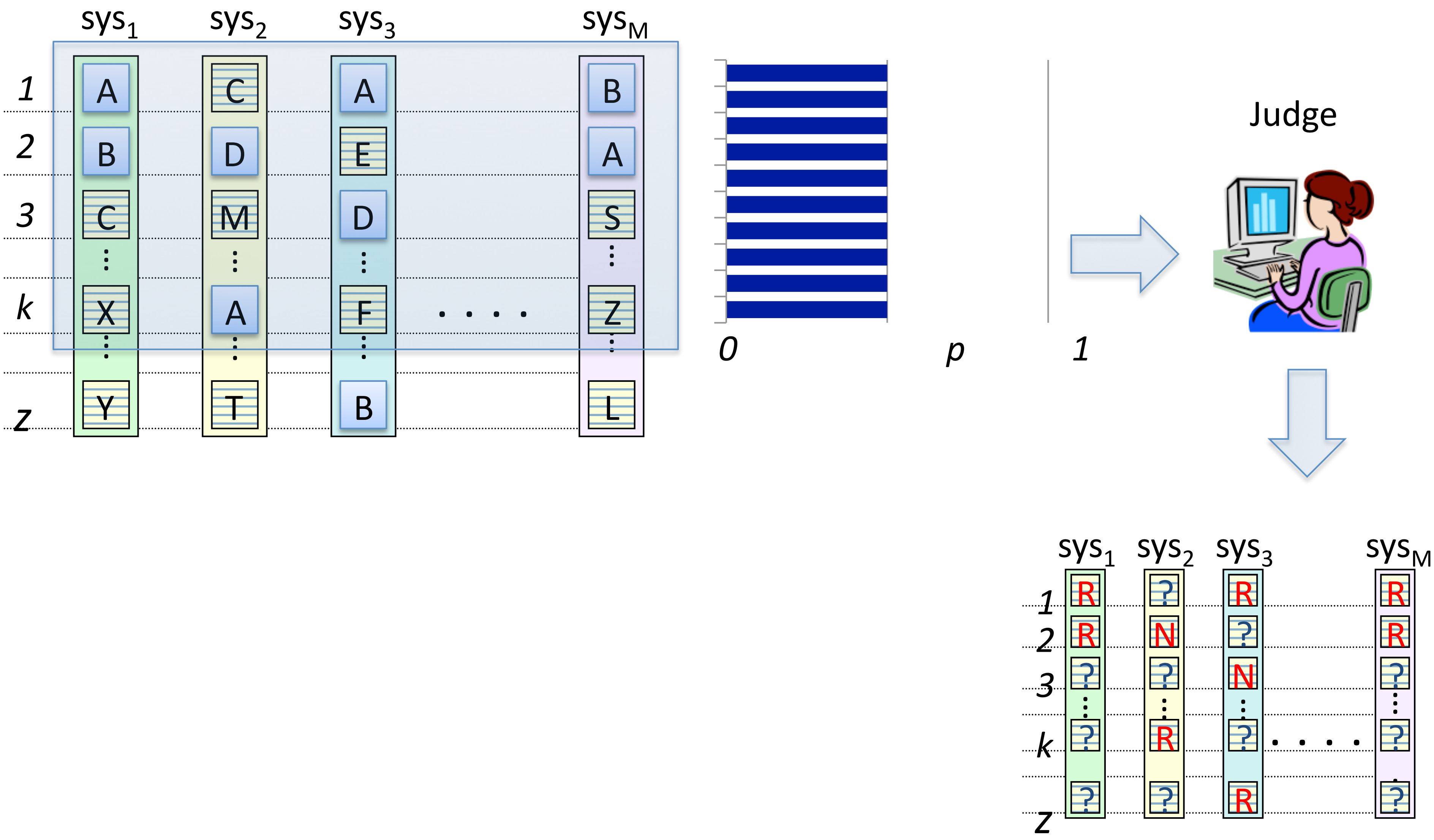
- Alternate solution:
  - uniformly sample animals
  - examine the sampled ones
  - return : #sick-seen/#samples
- Distribution: uniform over 10,000

$$p_i = \frac{1}{10{,}000}$$



- Random Variable: X = sick
  - 1 if sick, 0 otherwise

Uniform Random Sampling



Retrieval Evaluation with Incomplete Judgments

- Define a measure as outcome of a random experiment

- Estimate this outcome using random sampling
  - Incomplete judgments : a random sample drawn from the set of complete judgments

PC(k) as a Random Experiment

1. Select a rank at random from the set {1,….,k}

2. Output the binary relevance of document at this rank

PC(k) as a Random Experiment

1. Select a rank at random from the set {1,….,k}
2. Output the binary relevance of document at this rank.

- *PC(5) as an expectation of this random experiment*

R
R
N
R
N
N
N
R

PC(*k*) as a Random Experiment

1. Select a rank at random from the set {1,….,*k*}

2. Output the binary relevance of document at this rank.

- *PC(5) as an expectation of this random experiment*

1/5	R
1/5	R
1/5	N
1/5	R
1/5	N
	N
	N
	R

PC(k) as a Random Experiment

1. Select a rank at random from the set {1,….,k}

2. Output the binary relevance of document at this rank.

- *PC(5) as an expectation of this random experiment*

1/5	R	1
1/5	R	
1/5	N	
1/5	R	
1/5	N	
	N	
	N	
	R	

$$PC(5) = \frac{1}{5} \cdot 1 +$$

PC(*k*) as a Random Experiment

1. Select a rank at random from the set {1,….,*k*}

2. Output the binary relevance of document at this rank.

- *PC(5) as an expectation of this random experiment*

1/5	R	
1/5	R	1
1/5	N	
1/5	R	
1/5	N	
	N	
	N	
	R	

$$PC(5) = \frac{1}{5} \cdot 1 + \frac{1}{5} \cdot 1 +$$

PC(*k*) as a Random Experiment

1. Select a rank at random from the set {1,….,*k*}

2. Output the binary relevance of document at this rank.

- *PC(5) as an expectation of this random experiment*

1/5	R	
1/5	R	
1/5	N	0
1/5	R	
1/5	N	
	N	
	N	
	R	

$$PC(5) = \frac{1}{5}\cdot 1 + \frac{1}{5}\cdot 1 + \frac{1}{5}\cdot 0 +$$

PC(k) as a Random Experiment

1. Select a rank at random from the set {1,….,k}

2. Output the binary relevance of document at this rank.

- *PC(5) as an expectation of this random experiment*

1/5	R	
1/5	R	
1/5	N	
1/5	R	1
1/5	N	
	N	
	N	
	R	

$$PC(5) = \frac{1}{5} \cdot 1 + \frac{1}{5} \cdot 1 + \frac{1}{5} \cdot 0 + \frac{1}{5} \cdot 1 +$$

PC(*k*) as a Random Experiment

1. Select a rank at random from the set {1,….,*k*}

2. Output the binary relevance of document at this rank.

- *PC(5) as an expectation of this random experiment*

1/5	R	
1/5	R	
1/5	N	
1/5	R	
1/5	N	0
	N	
	N	
	R	

$$PC(5) = \frac{1}{5}\cdot 1 + \frac{1}{5}\cdot 1 + \frac{1}{5}\cdot 0 + \frac{1}{5}\cdot 1 + \frac{1}{5}\cdot 0$$

$$PC(5) = \frac{3}{5}$$

Average Precision as a Random Experiment

1. Select a relevant document at random
   – Rank of the document : *k*

2. Select a rank at random from the set {1,….,*k*}

3. Output the binary relevance of document at this rank.

- Average (step 1) of precisions at relevant documents (steps 2 and 3).

Average Precision as a Random Experiment

1. Select a relevant document at random
   – Rank of the document : k

2. Select a rank at random from the set $\{1,....,k\}$

3. Output the binary relevance of document at this rank.

R
R
N
R
N
N
N
R

Average Precision as a Random Experiment

1. Select a relevant document at random
   – Rank of the document : *k*

2. Select a rank at random from the set {1,….,*k*}

3. Output the binary relevance of document at this rank.

1/4	R
1/4	R
	N
1/4	R
	N
	N
	N
1/4	R

Average Precision as a Random Experiment

1. Select a relevant document at random
   - Rank of the document : *k*

2. Select a rank at random from the set {1,….,*k*}

3. Output the binary relevance of document at this rank.

1/4	R
1/4	R
	N
1/4	R
	N
	N
	N
1/4	R

$$AP = \frac{1}{4} \cdot 1 +$$

Average Precision as a Random Experiment

1. Select a relevant document at random
   - Rank of the document : *k*

2. Select a rank at random from the set {1,....,*k*}

3. Output the binary relevance of document at this rank.

1/4	R
1/4	R
	N
1/4	R
	N
	N
	N
1/4	R

$$AP = \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 1$$

Average Precision as a Random Experiment

1. Select a relevant document at random
   – Rank of the document : *k*

2. Select a rank at random from the set {1,….,*k*}

3. Output the binary relevance of document at this rank.

1/4	R
1/4	R
	N
1/4	R
	N
	N
	N
1/4	R

$$AP = \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot \frac{3}{4}$$

Average Precision as a Random Experiment

1. Select a relevant document at random
   – Rank of the document : *k*

2. Select a rank at random from the set {1,….,*k*}

3. Output the binary relevance of document at this rank.

1/4	R
1/4	R
	N
1/4	R
	N
	N
	N
1/4	R

$$AP = \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot \frac{3}{4} + \frac{1}{4} \cdot \frac{4}{8}$$

Average Precision as a Random Experiment

1. Select a relevant document at random
   - Rank of the document : *k*

2. Select a rank at random from the set {1,….,*k*}

3. Output the binary relevance of document at this rank.

1/4	R
1/4	R
	N
1/4	R
	N
	N
	N
1/4	R

$$AP = \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot\frac{3}{4} + \frac{1}{4}\cdot\frac{4}{8}$$

$$AP = \frac{1 + 1 + 3/4 + 4/8}{4}$$

Inferred AP [Yilmaz and Aslam, CIKM06]

(Adopted by TREC Terabyte, TREC VID)

- Select a relevant document at random
  - Uniformly sample from the complete judgments
  - Uniform distribution over the relevant documents

- Expected precision at a relevant document at rank *k*
  - Probability 1/k pick the current document
  - Probability (k-1)/k pick a document above

$$E[\text{prec at rank } k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} \cdot E[\text{prec above } k]$$

$$E[\text{prec above } k] = \frac{\text{judged rel above } k}{\text{judged rel above } k + \text{judged nonrel above } k}$$

Inferred AP

Search engine result:

R N R R N R N N R N

$$\text{actualAP} = \frac{1 + 2/3 + 3/4 + 4/6 + 5/9}{5} = 0.7278$$

Inferred AP

Search engine result:

R N ? R ? ? N ? R ?

$$\text{actualAP} = \frac{1 + 2/3 + 3/4 + 4/6 + 5/9}{5} = 0.7278$$

Inferred AP

Search engine result:

R N ? R ? ? N ? R ?

$E[prec] = 1$

$$\text{actualAP} = \frac{1 + 2/3 + 3/4 + 4/6 + 5/9}{5} = 0.7278$$

Inferred AP

Search engine result:

R N ? R ? ? N ? R ?

$E[prec] = 1$

$$E[prec] = \frac{1}{4} \cdot 1 + \frac{3}{4} \cdot \frac{1}{2} = \frac{5}{8}$$

$$\text{actualAP} = \frac{1 + 2/3 + 3/4 + 4/6 + 5/9}{5} = 0.7278$$

Inferred AP

Search engine result:

R N ? R ? ? N ? R ?

$$E[prec] = 1$$

$$E[prec] = \frac{1}{4} \cdot 1 + \frac{3}{4} \cdot \frac{1}{2} = \frac{5}{8}$$

$$E[prec] = \frac{1}{9} \cdot 1 + \frac{8}{9} \cdot \frac{2}{4} = \frac{5}{9}$$

$$\text{actualAP} = \frac{1 + 2/3 + 3/4 + 4/6 + 5/9}{5} = 0.7278$$

Inferred AP

Search engine result:

R N ? R ? ? N ? R ?

$$E[prec] = 1$$

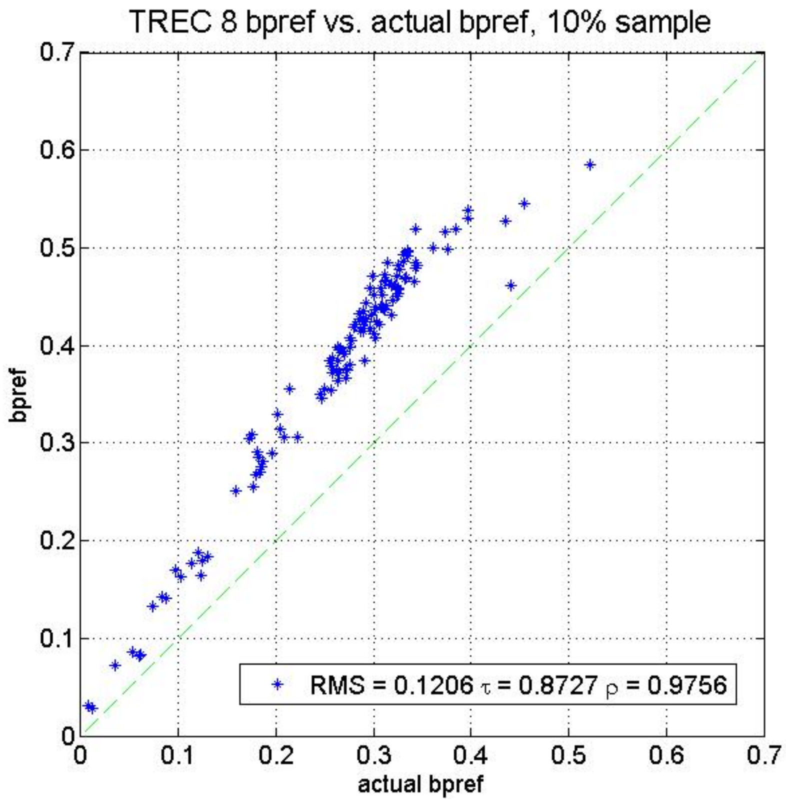
$$E[prec] = \frac{1}{4} \cdot 1 + \frac{3}{4} \cdot \frac{1}{2} = \frac{5}{8}$$

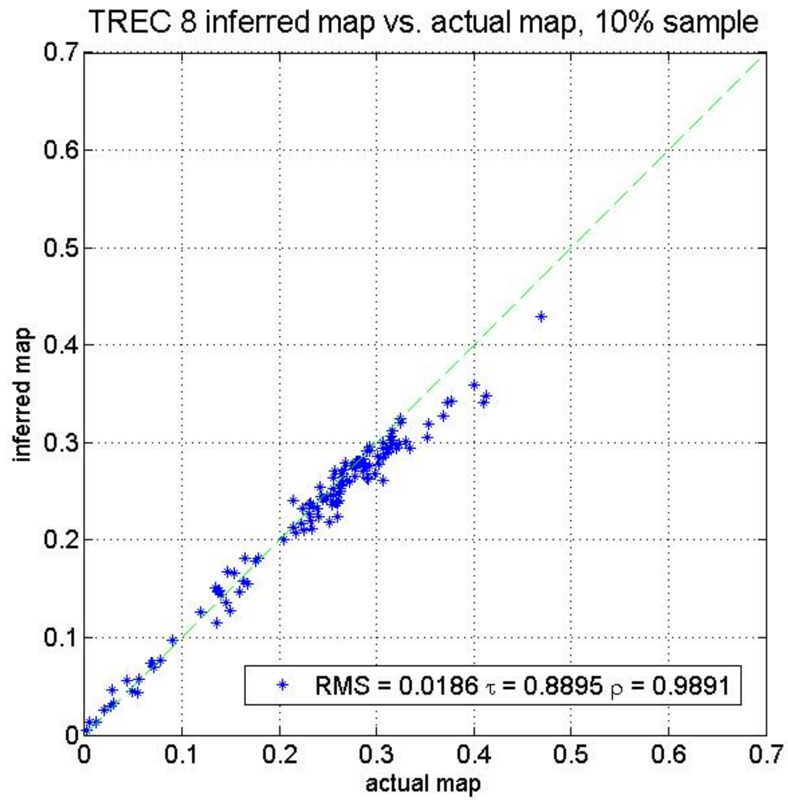
$$E[prec] = \frac{1}{9} \cdot 1 + \frac{8}{9} \cdot \frac{2}{4} = \frac{5}{9}$$

$$\text{inferredAP} = \frac{1 + 5/8 + 5/9}{3} = 0.7269$$

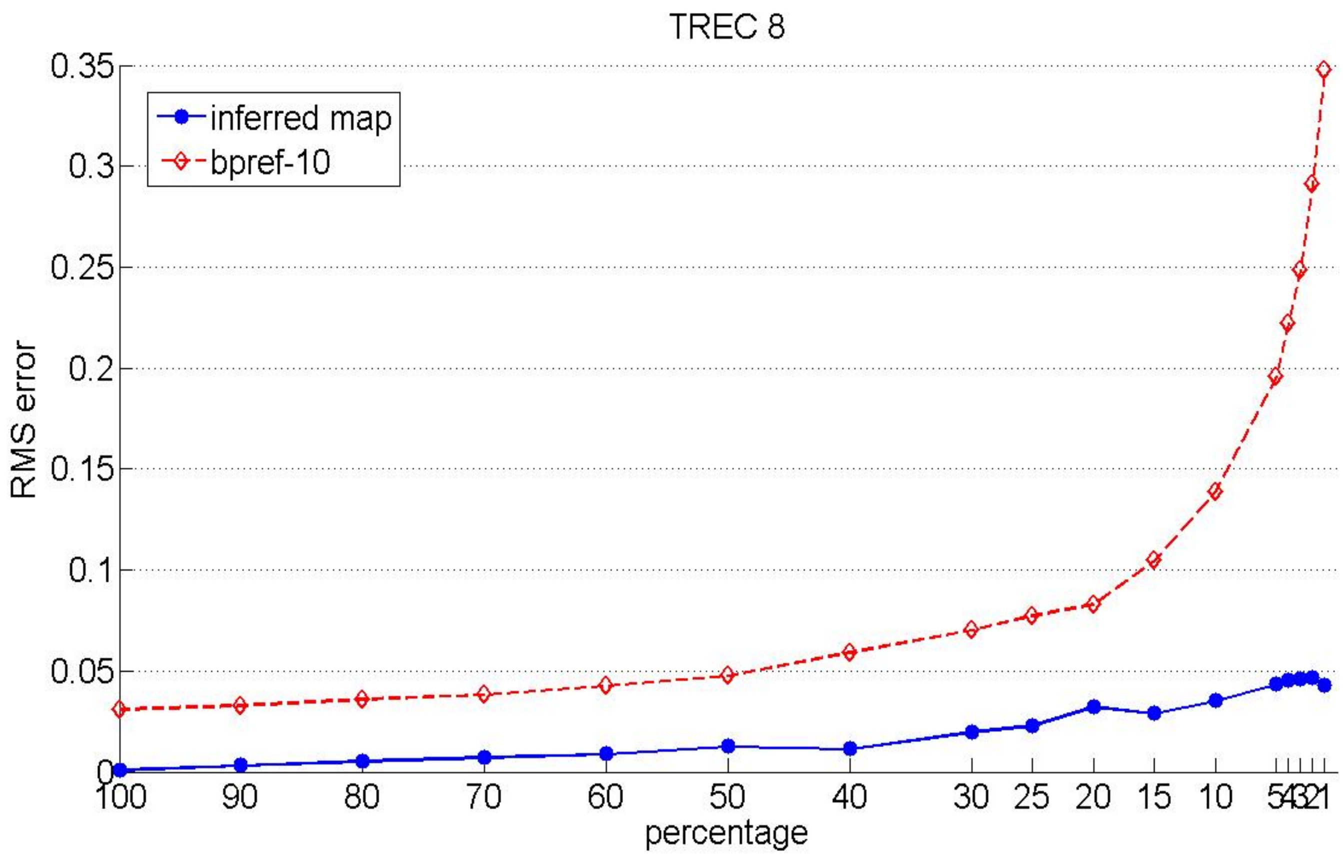
$$\text{actualAP} = \frac{1 + 2/3 + 3/4 + 4/6 + 5/9}{5} = 0.7278$$

Inferred AP, 10% Judgments

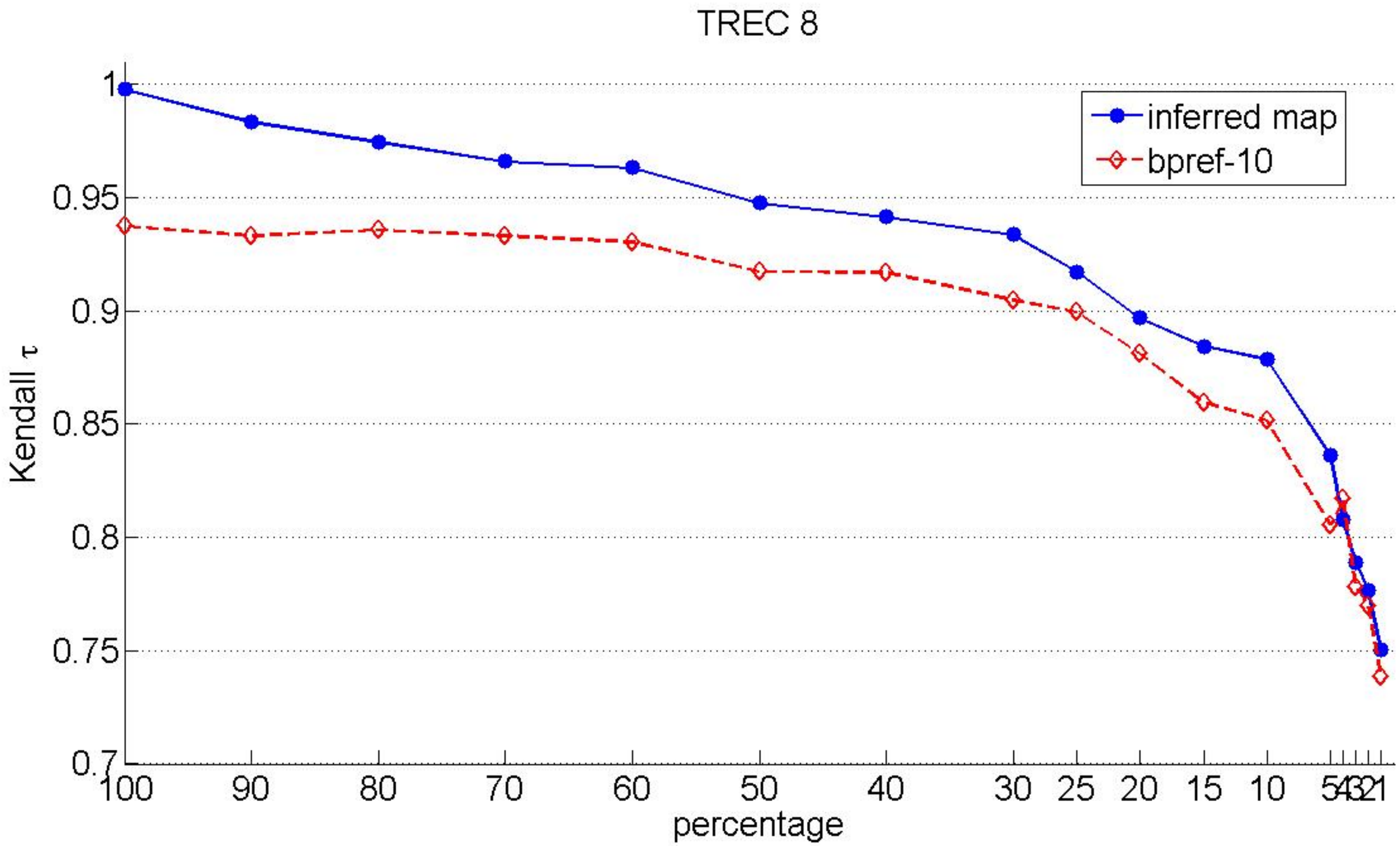




Comparison of the measures : RMS error



Comparison of the measures: Kendall's Tau



Variance in Inferred AP

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

- Inferred AP is unbiased in expectation
- Varies in practice
  - Variance and Confidence Intervals
- Random Experiment can be realized as two stage sampling

Variance in Inferred AP

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

- Two stages sampling
- Stage 1 : sample of *cut-off levels* (relevant documents) and average estimated precisions
  - 1st variance component

Variance in Inferred AP

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

- Two stages sampling
- Stage 2 : sample of *documents* above each selected cut-off level to compute precisions
  - 2nd variance component

Variance in Inferred AP

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

- Law of Total Variance
  - Total Variance in inferred AP =

    stage 1 variance + stage 2 variance

- Variance of Mean InfAP =

  Total Variance in InfAP / (# of Queries)^2

- Assign confidence intervals to Mean InfAP according to Central Limit Theorem

Variance in Inferred AP

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

- Law of Total Variance
  - Total Variance in inferred AP = stage 1 variance + stage 2 variance

$$\mathrm{var}[\mathrm{infAP}] = \mathrm{var}\big[E[\mathrm{infAP} \mid \mathrm{s_d}]\big] + E\big[\mathrm{var}[\mathrm{infAP} \mid \mathrm{s_d}]\big]$$

s_d : the sample of cut-off levels

Variance in Inferred AP

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

- Law of Total Variance
  - Total Variance in inferred AP =

    stage 1 variance + stage 2 variance

$$\mathrm{var}[\mathrm{infAP}] = \mathrm{var}\left[E\left[\mathrm{infAP} \mid \mathrm{s_d}\right]\right] + E\left[\mathrm{var}\left[\mathrm{infAP} \mid \mathrm{s_d}\right]\right]$$

$$E\left[\mathrm{infAP} \mid \mathrm{s_d}\right] = \frac{1}{r}\sum_{k \in \mathrm{s_d}} E\left[\widehat{\mathrm{PC(k)}} \mid \mathrm{s_d}\right] = \frac{1}{r}\sum_{k \in \mathrm{s_d}} \mathrm{PC(k)}$$

$$\mathrm{var}\left[E\left[\mathrm{infAP} \mid \mathrm{s_d}\right]\right] = \mathrm{var}\left[\frac{1}{r}\sum_{k \in \mathrm{s_d}} \mathrm{PC(k)}\right]$$

s_d : the sample of cut-off levels, r : number of relevant docs in s_d

Variance in Inferred AP

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

- Law of Total Variance
  - Total Variance in inferred AP =
    stage 1 variance + stage 2 variance

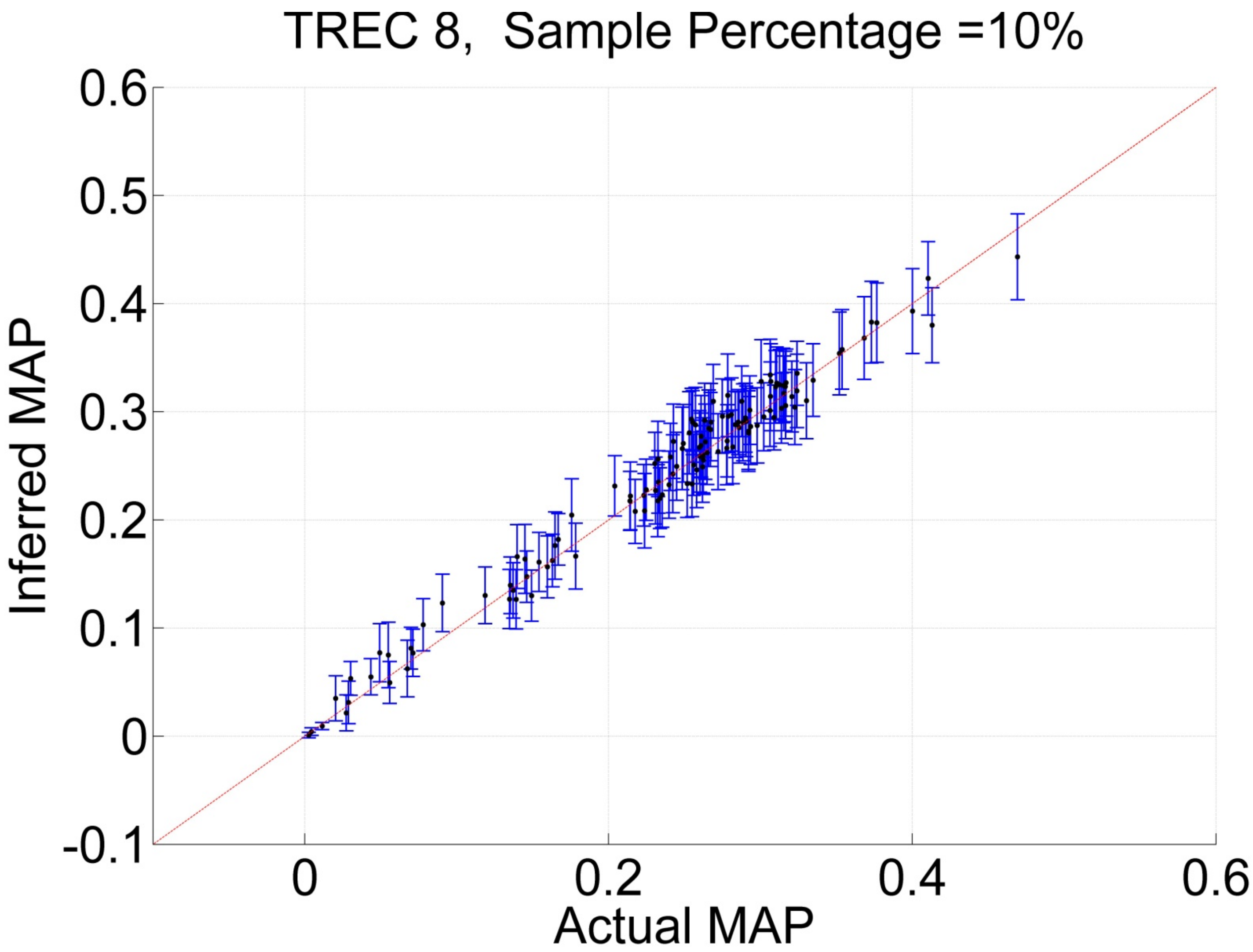
$$\mathrm{var}[\mathrm{infAP}] = \mathrm{var}\left[E\left[\mathrm{infAP} \mid \mathrm{s_d}\right]\right] + E\left[\mathrm{var}\left[\mathrm{infAP} \mid \mathrm{s_d}\right]\right]$$

$$\mathrm{var}\left[\mathrm{infAP} \mid \mathrm{s_d}\right] = \mathrm{var}\left[\frac{1}{r}\sum_{k\in \mathrm{s_d}} \widehat{\mathrm{PC(k)}}\right] = \frac{1}{r^2}\mathrm{var}\left[\sum_{k\in \mathrm{s_d}} \widehat{\mathrm{PC(k)}}\right]$$

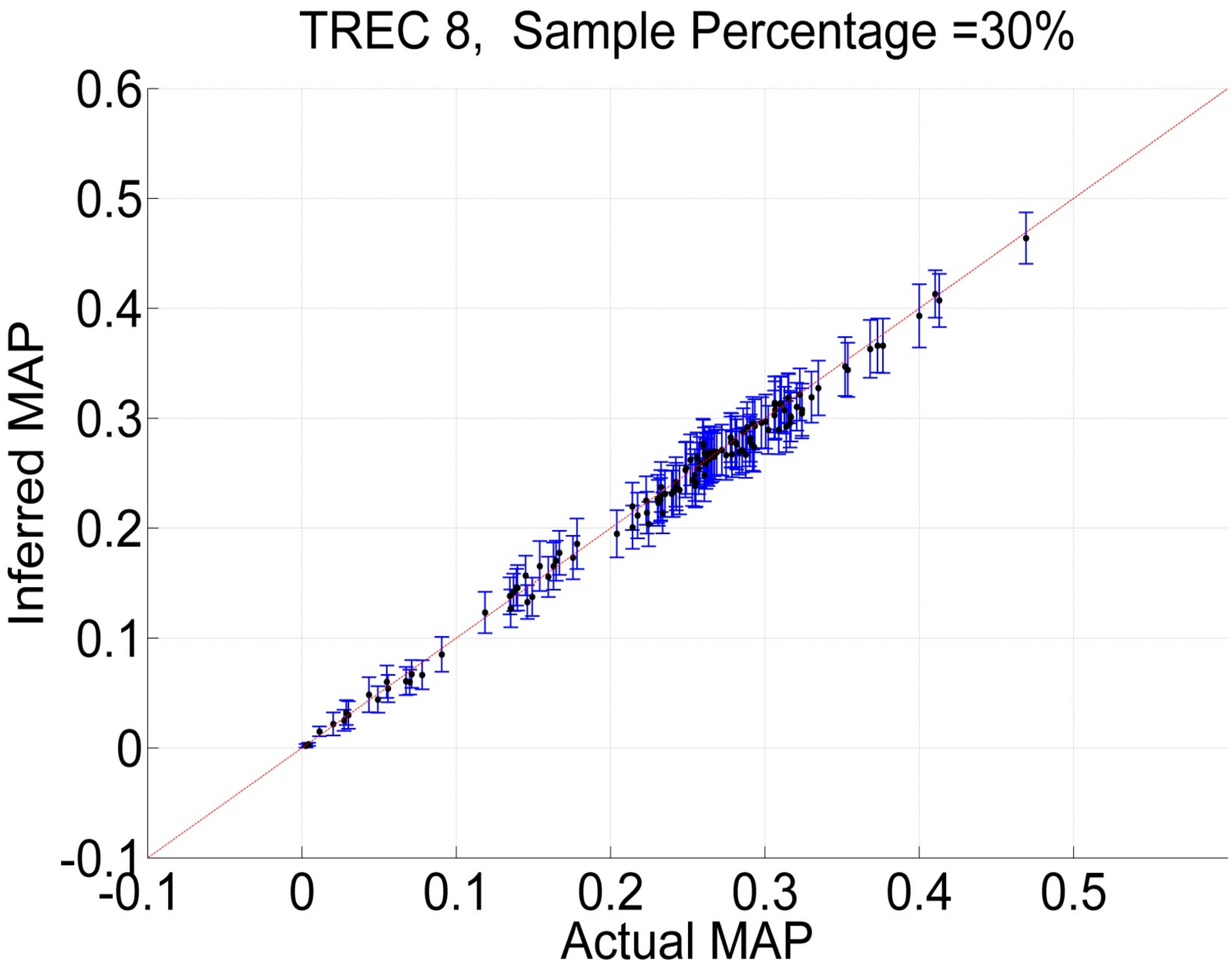
- If we consider precisions independent

$$= \frac{1}{r^2}\sum_{k\in \mathrm{s_d}} \mathrm{var}\left[\widehat{\mathrm{PC(k)}} \mid \mathrm{s_d}\right]$$

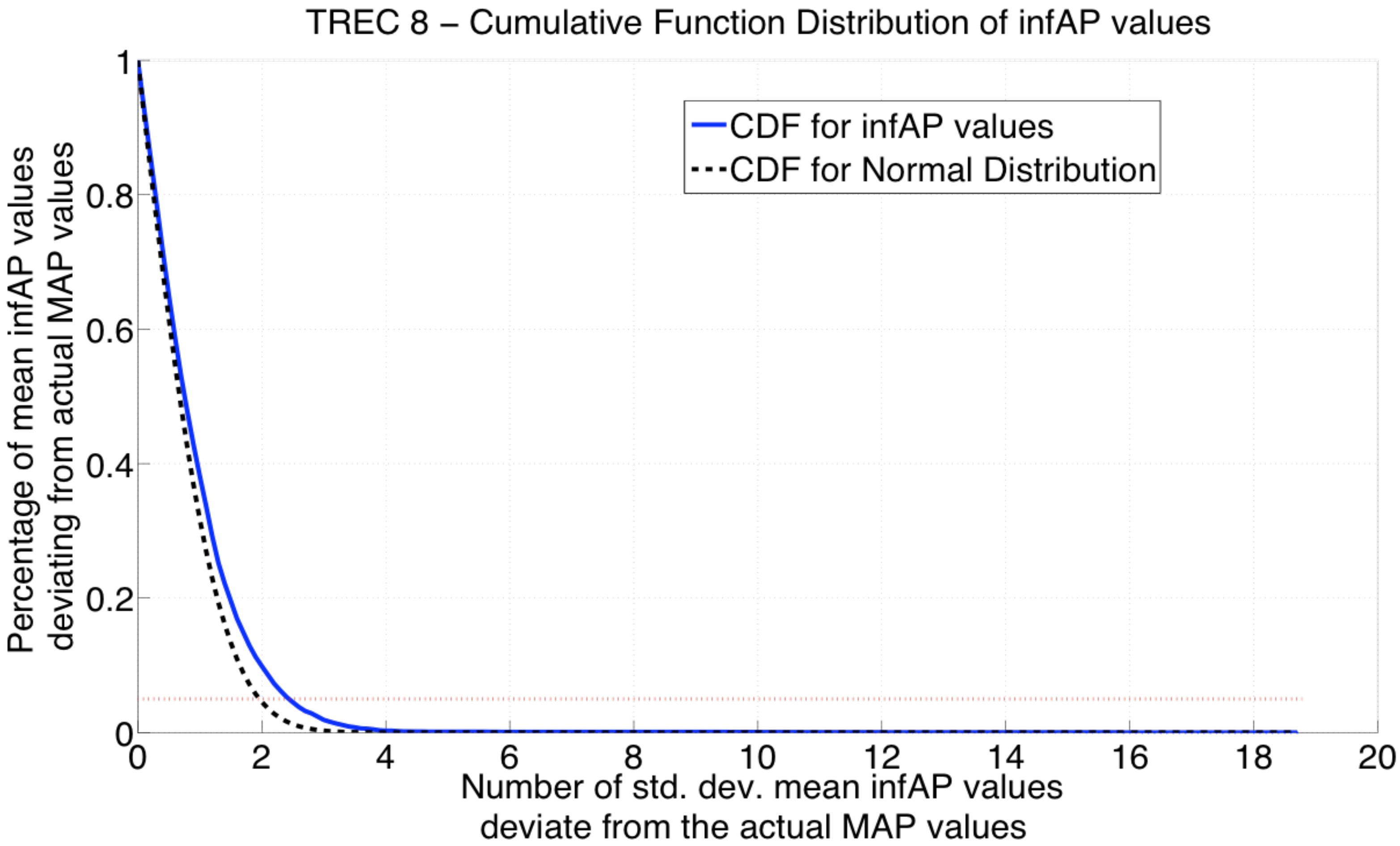
Confidence Intervals for Mean InfAP



Confidence Intervals for Mean InfAP

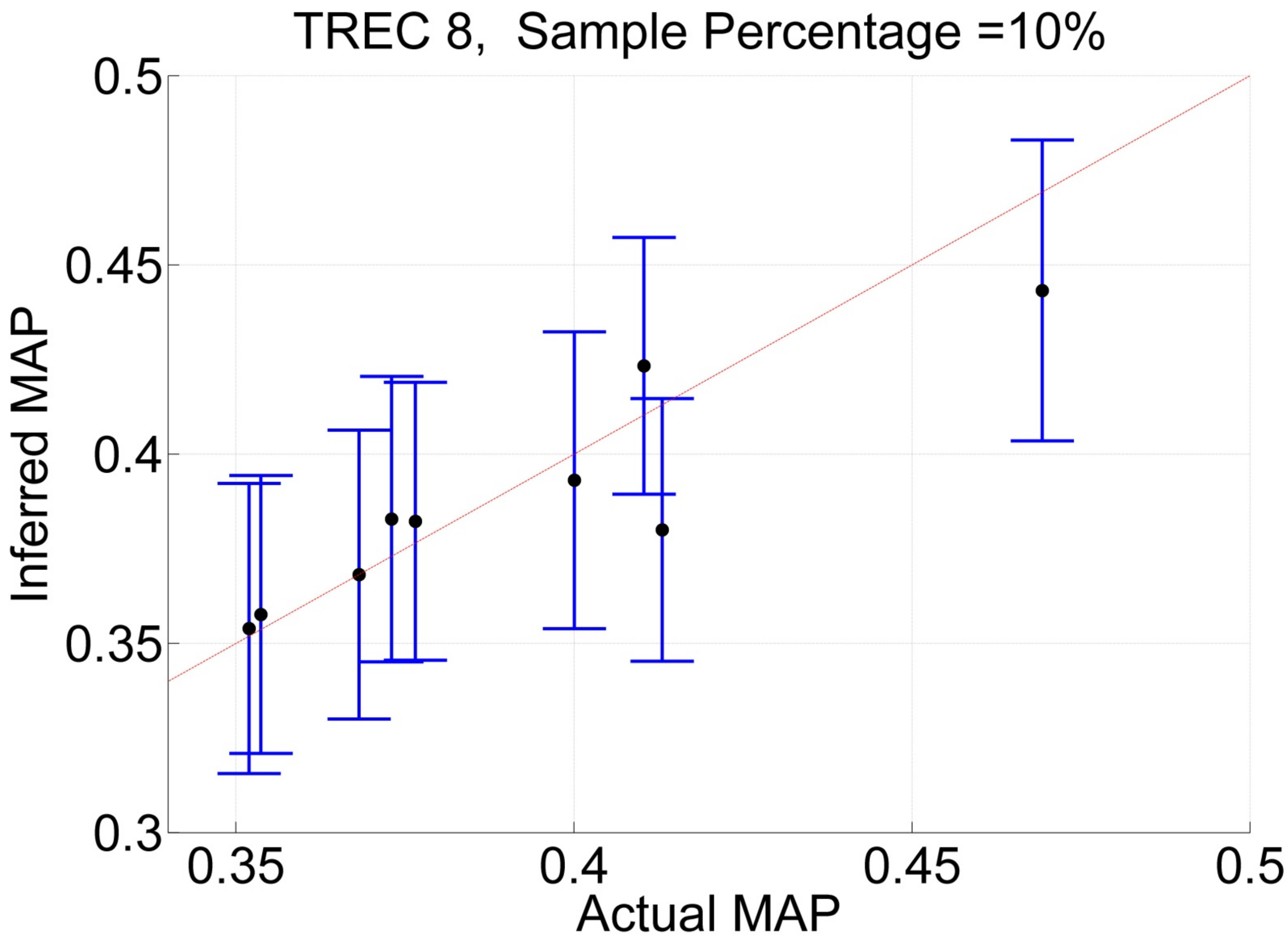


Confidence Intervals for Mean InfAP



- K-S test : for 90% of systems the hypothesis cannot be rejected ($\alpha = 0.05$)

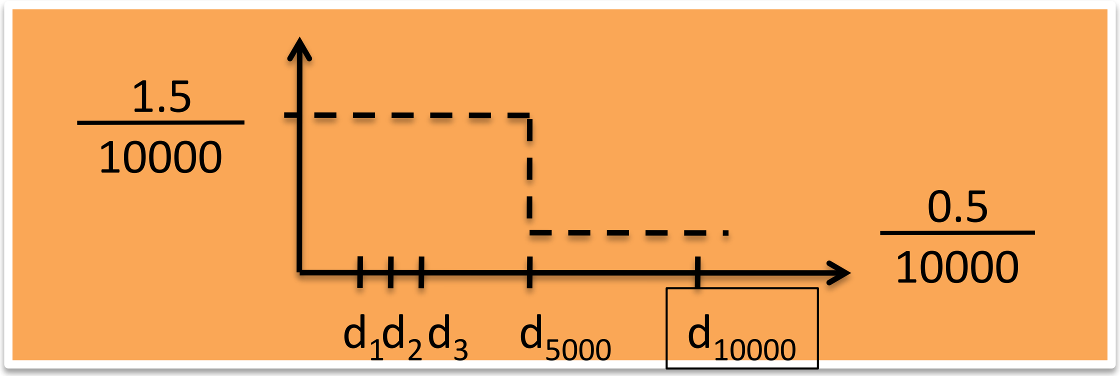
Confidence Intervals for Mean InfAP



Increasing the Certainty in Estimators

- Sample "more" where sick animals are
  - for example categorize/order them by age:
    - 1-5000 old;   5001-10000 young

- Distribution: stratified over 10,000

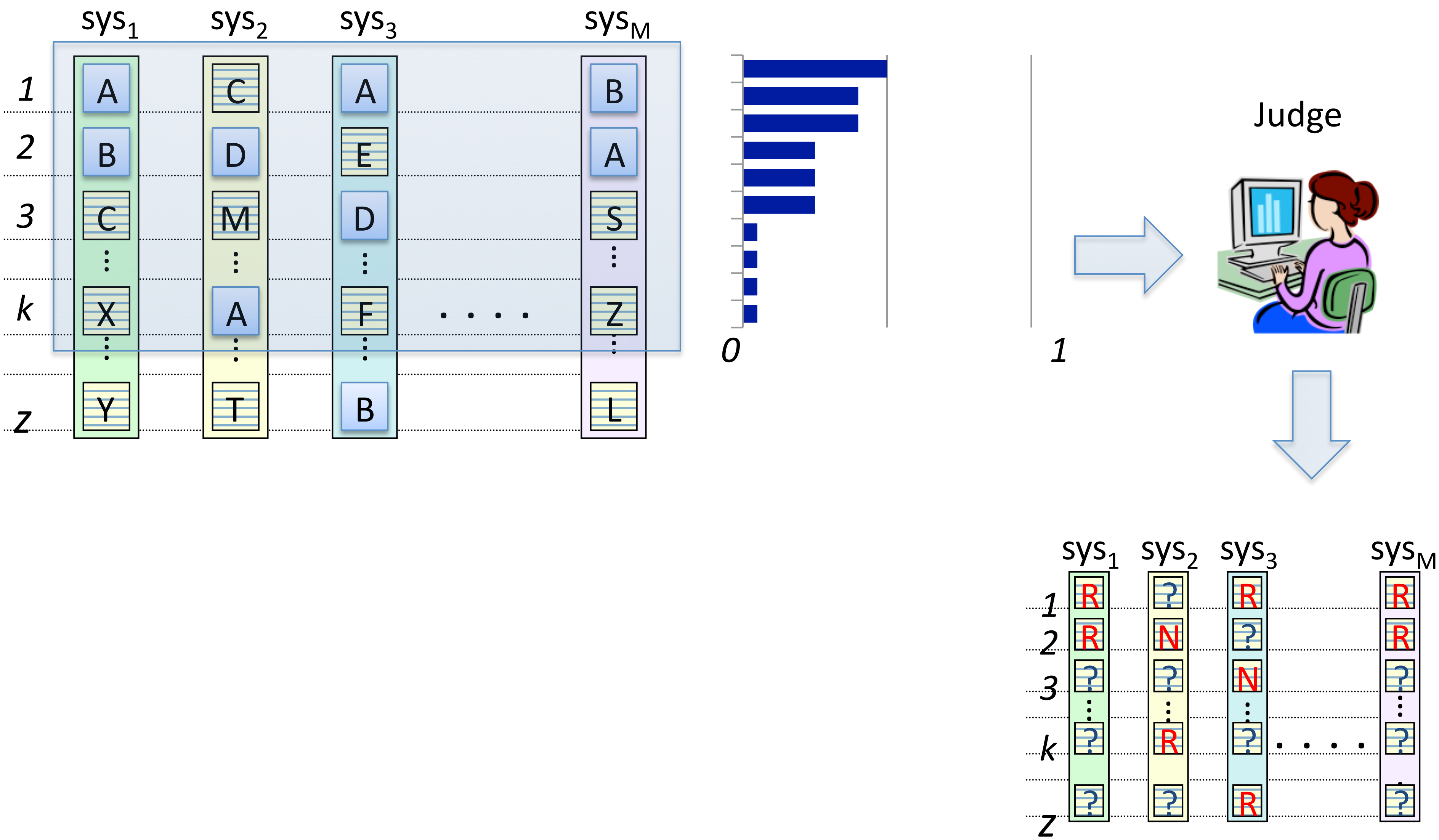


$$p_i = \begin{cases} 1.5/10{,}000 & i \le 5{,}000 \\ 0.5/10{,}000 & i > 5{,}000 \end{cases}$$

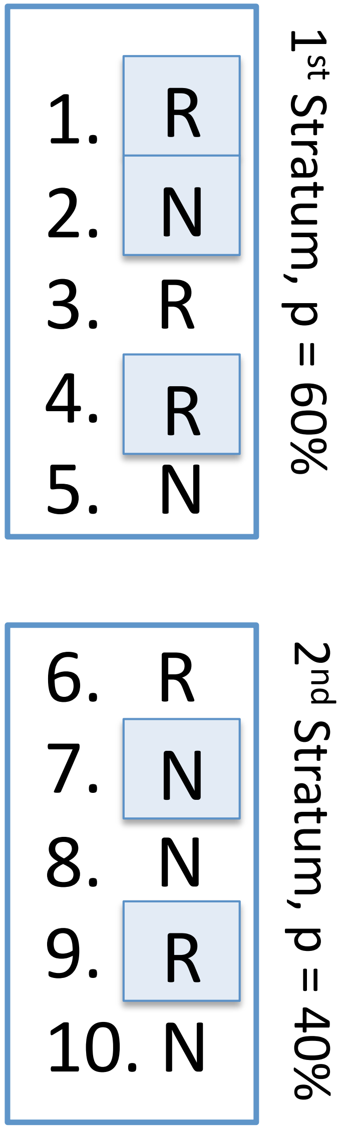
Stratified Random Sampling

- Goal : Decrease variance in the estimator
- Evaluation measures give more weight to documents towards the top of the list
- “Top-heavy” sampling strategy can reduce variance in evaluation measures

Stratified Random Sampling



Stratified Random Sampling



- Divide complete pool of judgments into strata (disjoint contiguous subsets)
- Randomly sample some documents from each stratum to be judged
- Sampling percentage within each stratum can be different
- Evaluate search engines with sampled documents

Extended infAP (xinfAP) [Yilmaz et al SIGIR08]

(Adopted by tracks in TREC, CLEF, INEX)

- Select a relevant document at random (1st step)
  - Selected relevant document can fall in any of the strata
  - By the definition of conditional expectation

$$\text{xinfAP} = E[AP] = \sum_{\forall s \in Strata} P_s \cdot E[AP_s]$$

P_s : Probability that a randomly picked rel docs falls into strata s

Extended infAP (xinfAP)

- Select a relevant document at random (1st step)
  - Probability of picking relevant document from stratum s

$$P_s = \frac{R_s}{R_Q}$$

R_s : Num rels within stratum s

R_Q : Num rels in query Q

Extended infAP (xinfAP)

- Select a relevant document at random (1st step)
  - Probability of picking relevant document from stratum s

$$P_s = \frac{R_s}{R_Q} \qquad R_s : \text{Num rels within stratum } s \qquad R_Q : \text{Num rels in query } Q$$

$$\hat{P}_s \sim \frac{E[R_s]}{E[R_Q]}$$

$$E[R_s] = \frac{|\text{rel docs sampled from } s|}{|\text{docs sampled from } s|} \cdot |\text{docs in } s| \qquad E[R_Q] = \sum_{\forall s} E[R_s]$$

Extended infAP (xinfAP)

1st Stratum, p = 60%

1. R
2. N
3. R
4. R
5. N

2nd Stratum, p = 40%

6. R
7. N
8. N
9. R
10. N

$$E[R_{s_1}] = \frac{2}{3} \cdot 5$$

$$E[R_{s_2}] = \frac{1}{2} \cdot 5$$

$$\hat{P}_{s_1} = \left(\frac{2}{3} \cdot 5\right) / \left(\frac{2}{3} \cdot 5 + \frac{1}{2} \cdot 5\right) = 0.57$$

Extended infAP (xinfAP)

$$\text{xinfAP} = E[AP] = \sum_{\forall s \in Strata} P_s \cdot E[AP_s]$$

- Select a relevant document at random (1st step)
  - Within each stratum:
    - Judged documents uniform random subset of all documents
    - Uniform distribution over the relevant documents
    - $E[AP_s]$ computed as average of precisions at judged relevant documents

Extended infAP (xinfAP)

- Precision at a relevant document at rank *k* (2nd and 3rd step)
  - Select a rank at random from the set {1,….,*k*}
  - Output the binary relevance of document at this rank.
  - Probability 1/k pick the current document

$$E[PC_k] = \frac{1}{k} \cdot 1$$

Extended infAP (xinfAP)

- Precision at a relevant document at rank *k* (2nd and 3rd step)
  - Select a rank at random from the set {1,….,*k*}
  - Output the binary relevance of document at this rank.

  - Probability 1/k pick the current document
  - Probability (k-1)/k pick a document above

$$E[PC_k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} E[PC \text{ above } k]$$

Extended infAP (xinfAP)

- Precision at a relevant document at rank *k* (2nd and 3rd step)
  - Select a rank at random from the set {1,….,*k*}
  - Output the binary relevance of document at this rank.

  - Probability 1/k pick the current document
  - Probability (k-1)/k pick a document above

$$E[PC_k] = \frac{1}{k}\cdot 1 + \frac{k-1}{k} E[PC \text{ above } k]$$

$$E[PC \text{ above } k] = \sum_{\forall s} \frac{N_s^{k-1}}{k-1} \cdot E_s[PC \text{ above } k]$$

Probability of picking a document (above k) from stratum s

Extended infAP (xinfAP)

- Precision at a relevant document at rank *k* (2nd and 3rd step)
  - Select a rank at random from the set {1,….,*k*}
  - Output the binary relevance of document at this rank.

  - Probability 1/k pick the current document
  - Probability (k-1)/k pick a document above

$$E[PC_k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} E[PC \text{ above } k]$$

$$E[PC \text{ above } k] = \sum_{\forall s} \frac{N_s^{k-1}}{k-1} \cdot E_s[PC \text{ above } k]$$

$$E_s[PC \text{ above } k] = \frac{\#\text{ judged rel above } k \text{ within } s}{\#\text{ judged above } k \text{ within } s}$$

Extended infAP (xinfAP)

- Precision at a relevant document at rank *k* (2nd and 3rd step)
  - Select a rank at random from the set {1,….,*k*}
  - Output the binary relevance of document at this rank.

  - Probability 1/k pick the current document
  - Probability (k-1)/k pick a document above

$$E[PC_k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} E[PC \text{ above } k]$$

$$E[PC \text{ above } k] = \sum_{\forall s} \frac{N_s^{k-1}}{k-1} \cdot E_s[PC \text{ above } k]$$

$$E_s[PC \text{ above } k] = \frac{\#\text{ judged rel above } k \text{ within } s + \varepsilon}{\#\text{ judged above } k \text{ within } s + 2\varepsilon}$$

Extended infAP (xinfAP)

1st Stratum, p = 60%

1. R
2. N
3. R
4. R
5. N

2nd Stratum, p = 40%

6. R
7. N
8. N
9. R
10. N

$$E[PC_k] = \frac{1}{k}\cdot 1 + \frac{k-1}{k} E[PC \text{ above } k]$$

$$E[PC_9] = \frac{1}{9}\cdot 1 + \frac{8}{9}\cdot\left(\frac{5}{8}\cdot\frac{2}{3} + \frac{3}{8}\cdot\frac{0}{1}\right) = 0.4815$$

Extended infAP (xinfAP)

1st Stratum, p = 60%

1. R
2. N
3. R
4. R
5. N

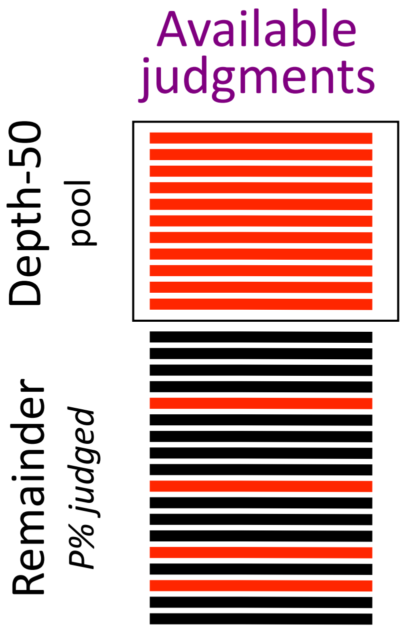
2nd Stratum, p =40%

6. R
7. N
8. N
9. R
10. N

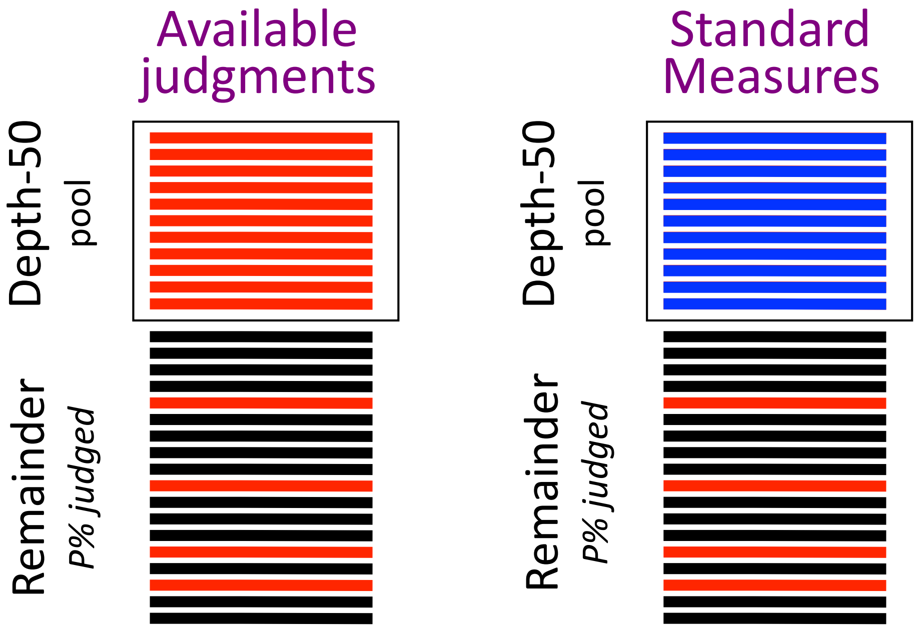
$$E[PC\ \text{above}\ k] = \sum_{\forall s} \frac{N_s^{k-1}}{k-1} \cdot E_s[PC\ \text{above}\ k]$$

$$E[PC_9] = \frac{1}{9} \cdot 1 + \frac{8}{9} \cdot \left(\frac{5}{8} \cdot \frac{2}{3} + \frac{3}{8} \cdot \frac{0}{1} \right) = 0.4815$$

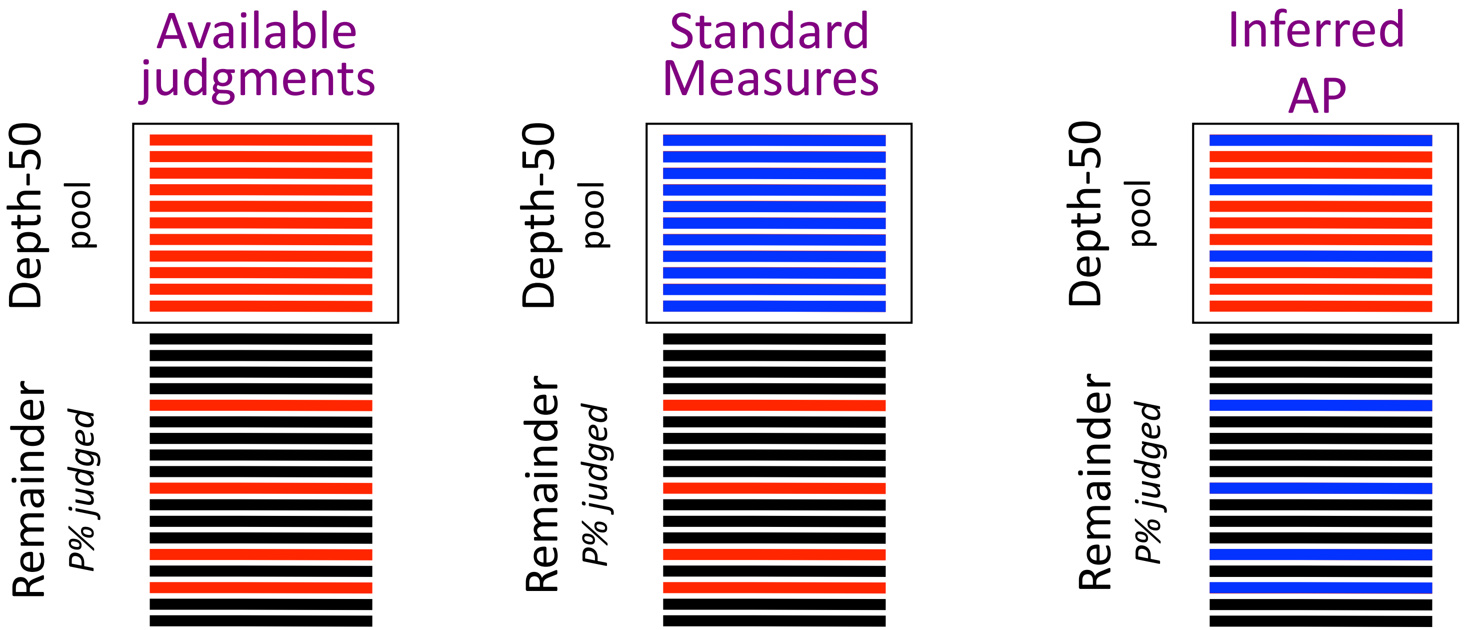
TREC Terabyte '06



TREC Terabyte ‘06



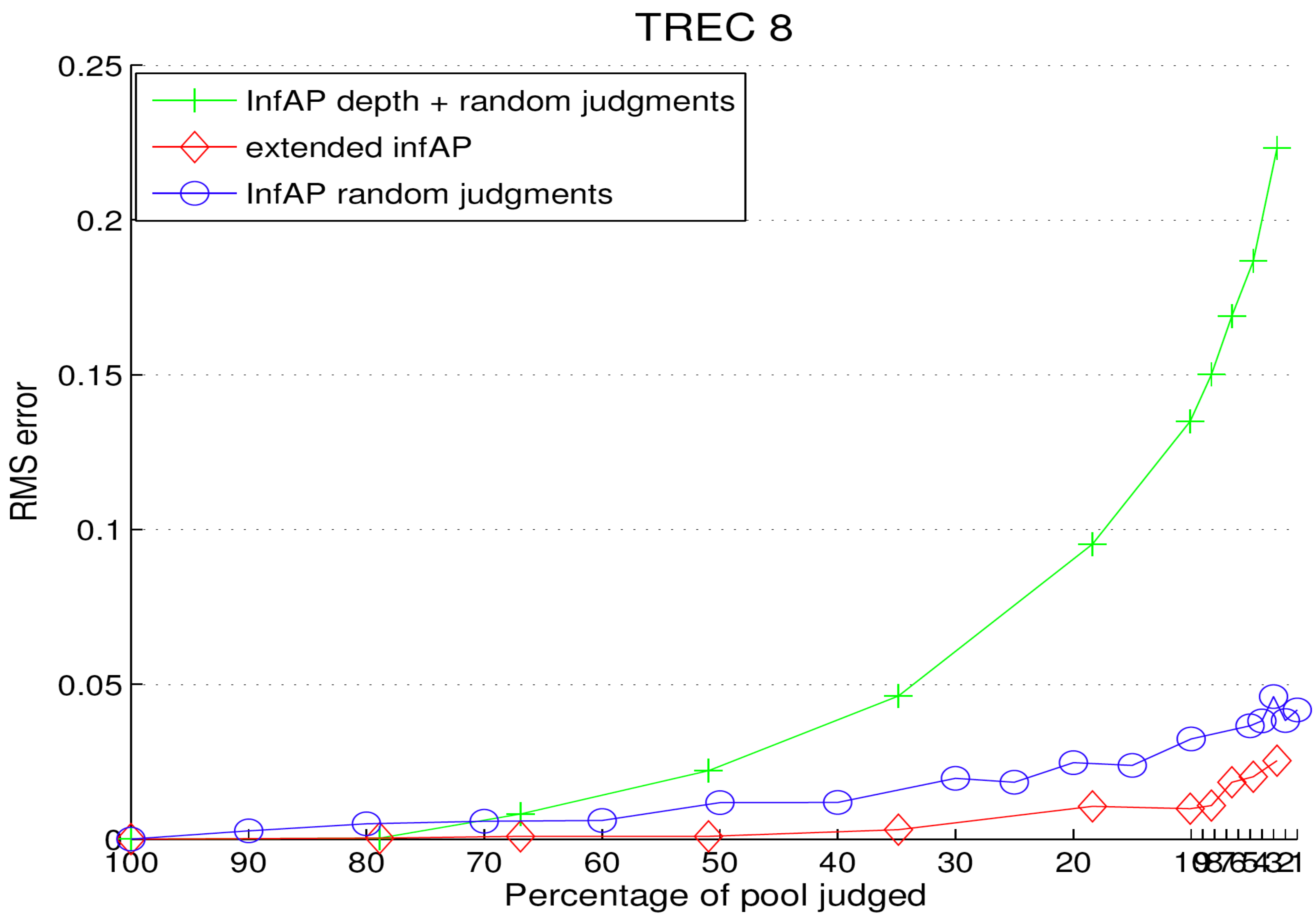
TREC Terabyte '06



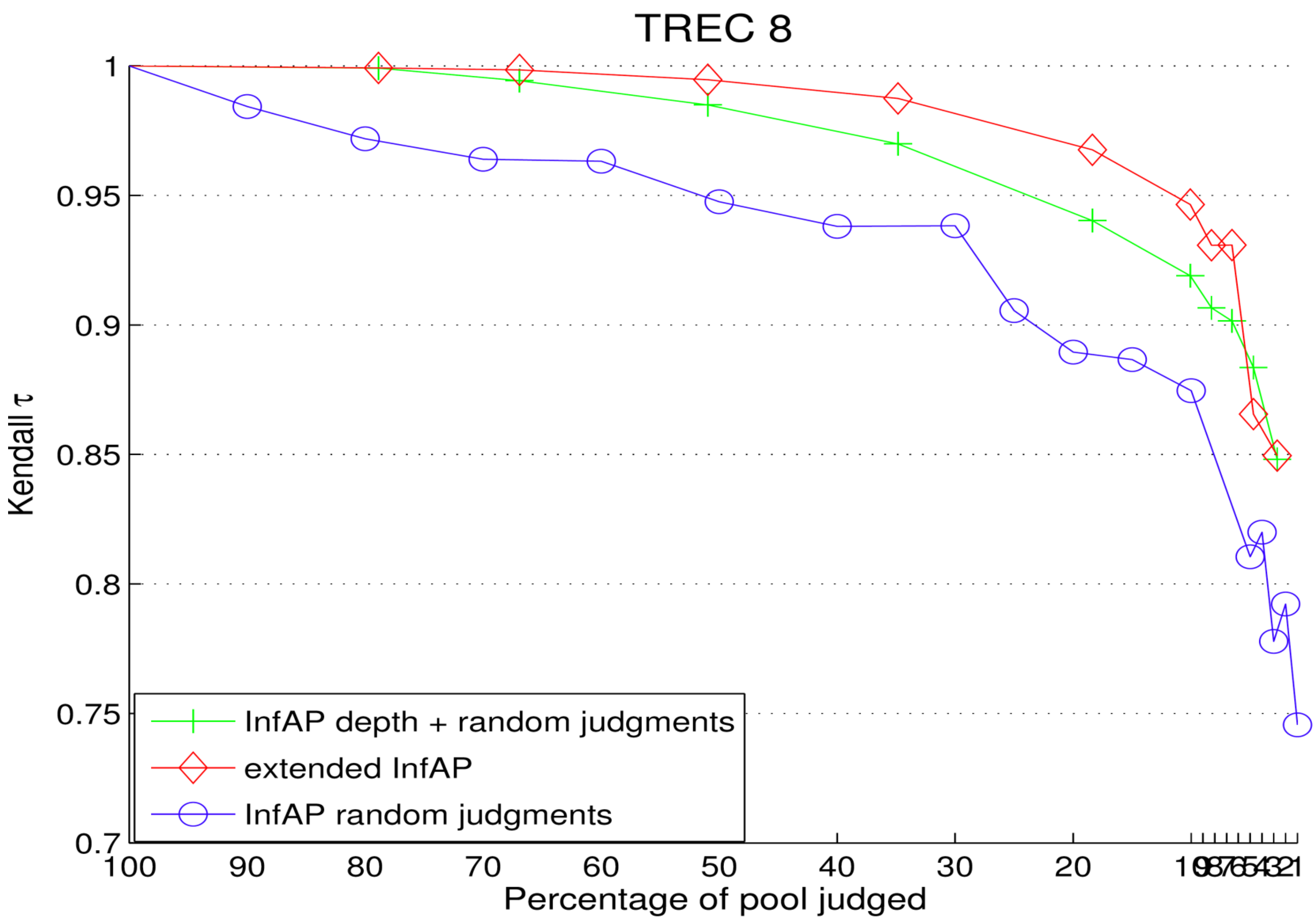
Simulate Terabyte Setup on TREC 8 data

- Assume complete judgments: depth-100 pool
- Form different depth-k pools
  - $k \in \{1,2,3,4,5,10,20,30,40,50\}$
- For each k compute the total number of documents in depth-k pool
- Randomly sample equal number of documents from the complete judgment set (excluding depth-k pool)
- Assume the remaining documents are unjudged
  - Evaluate search engines with sampled documents

Comparison of the measures : RMS error

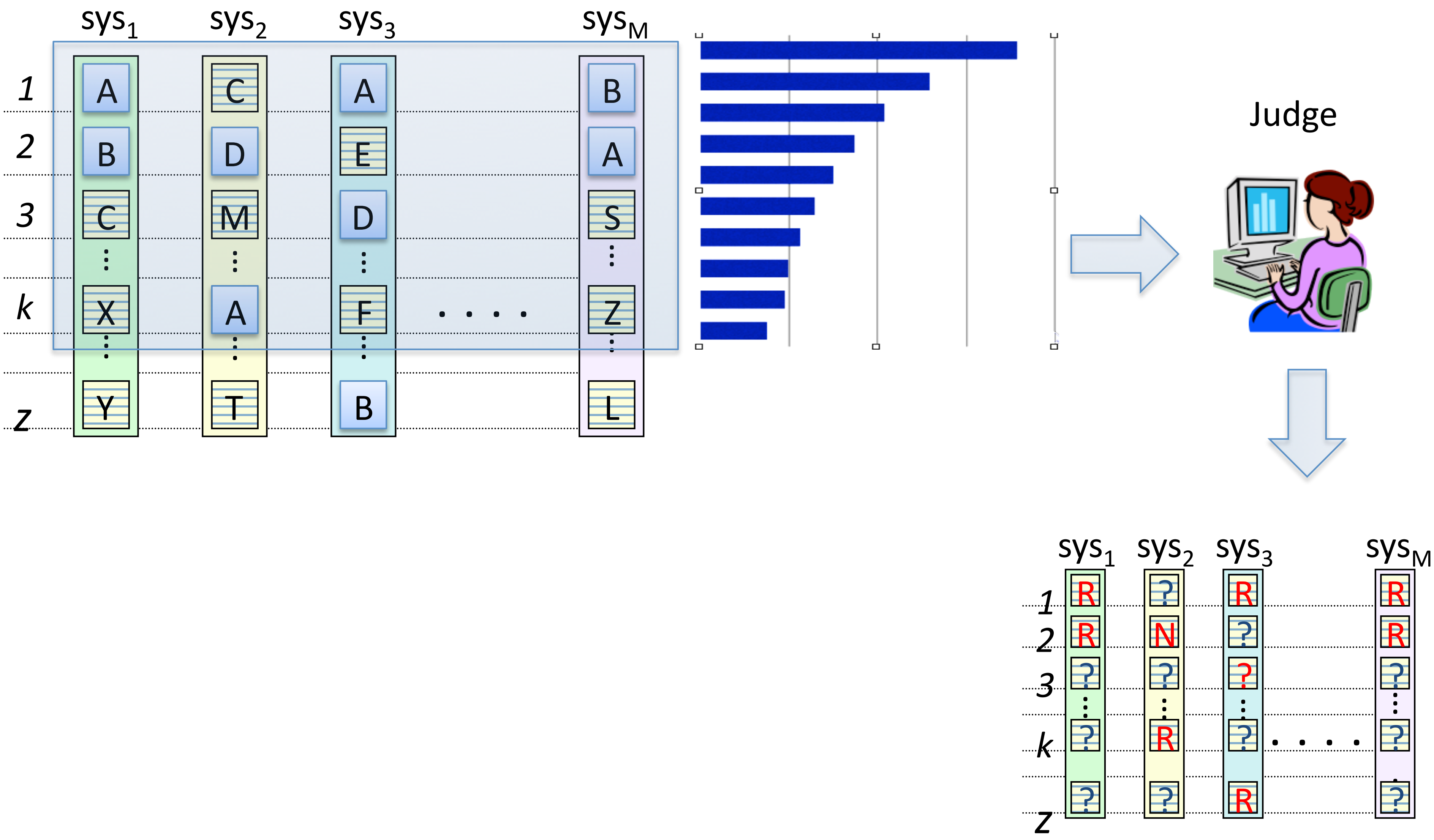


Comparison of the measures: Kendall's Tau

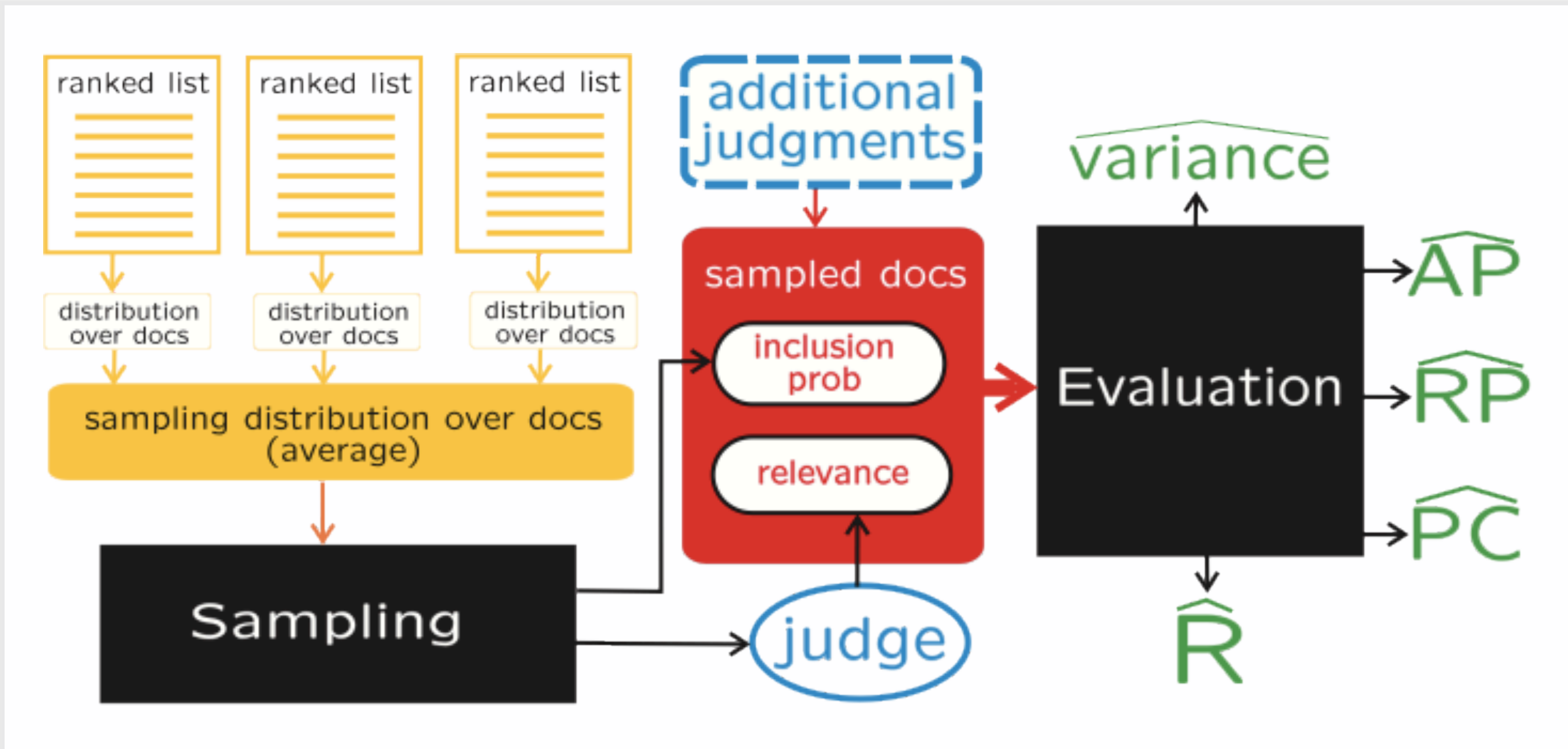


Importance Sampling

[Aslam and Pavlu, Tech. Report]



StatAP: Sampling w/out Replacement



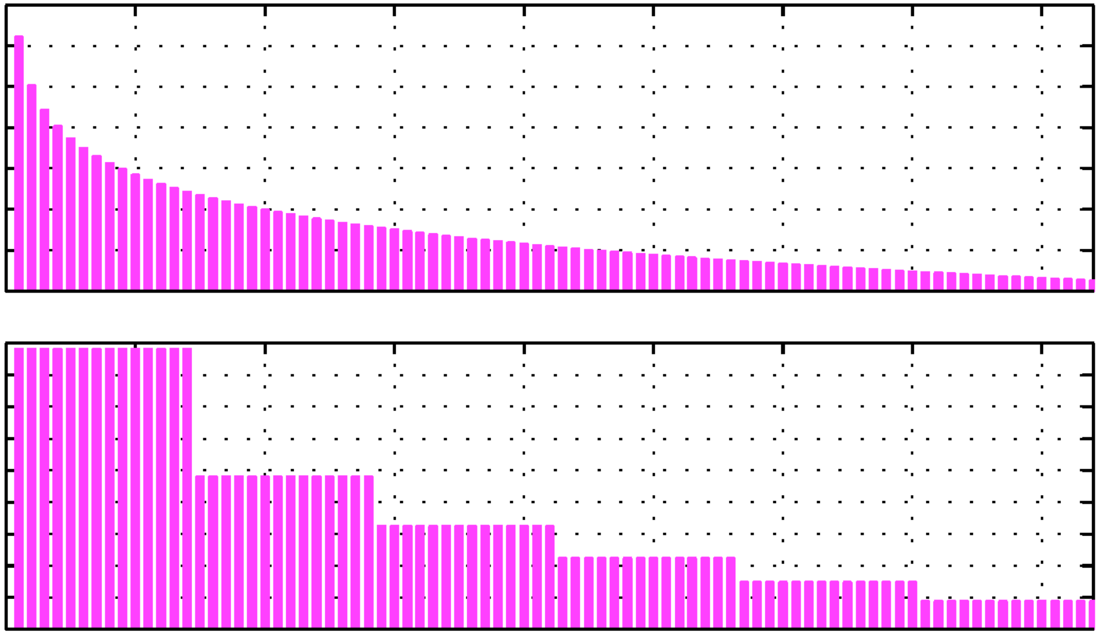
prior, sampling and estimation independent

StatAP

- Sampling without replacement
  - π_k : inclusion probabilities
  - stratified sampling
    - imagine using sequential sampling
- use a ratio estimator
  - estimate precision@rank
  - numerator: HT for sum-precision
  - denominator: HT for R

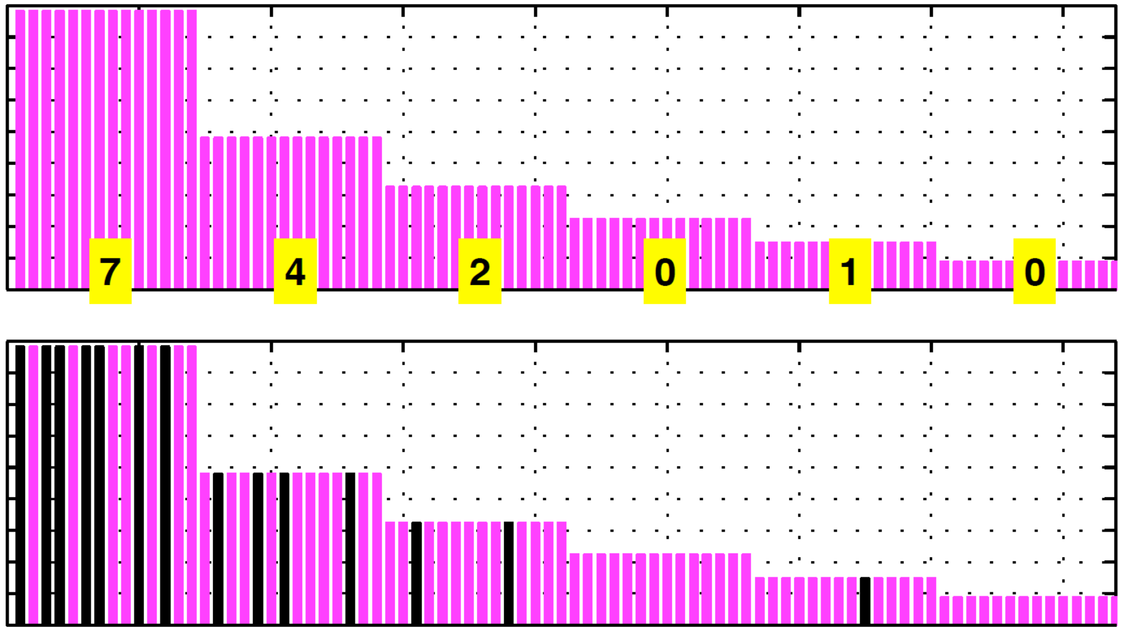
$$StatAP = \frac{\sum_{k \in S} p_k / \pi_k}{\sum_{k \in S} 1 / \pi_k}$$

Importance Sampling to Stratified Sampling



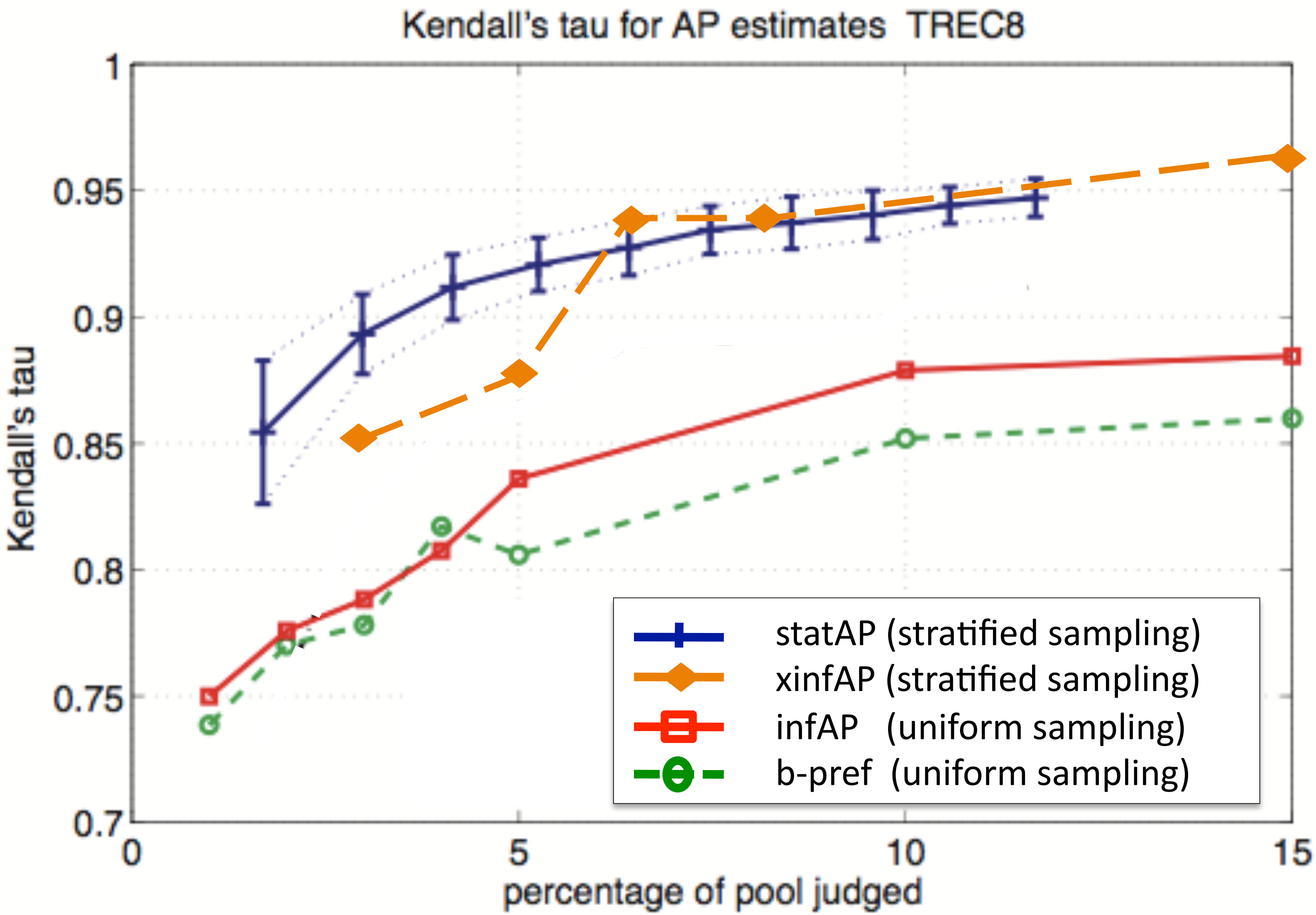
- non-uniform distribution; sample size = 14
- partition docs in buckets of size 14 each

Stratified sampling



- sample the buckets with replacement 14 times
  – based on the cumulative weight for each bucket
- for each bucket, if picked k times, sample uniformly without replacement k docs in it

Comparison of the measures: Kendall’s Tau



Today's Outline

- Low cost evaluation

1. Depth-k pooling (standard method)

2. Evaluating without judgments (automatic eval)
3. Finding relevance documents as quickly as possible

4. Computing measures with incomplete judgments
5. Estimating measures
6. Inferring relevance judgments

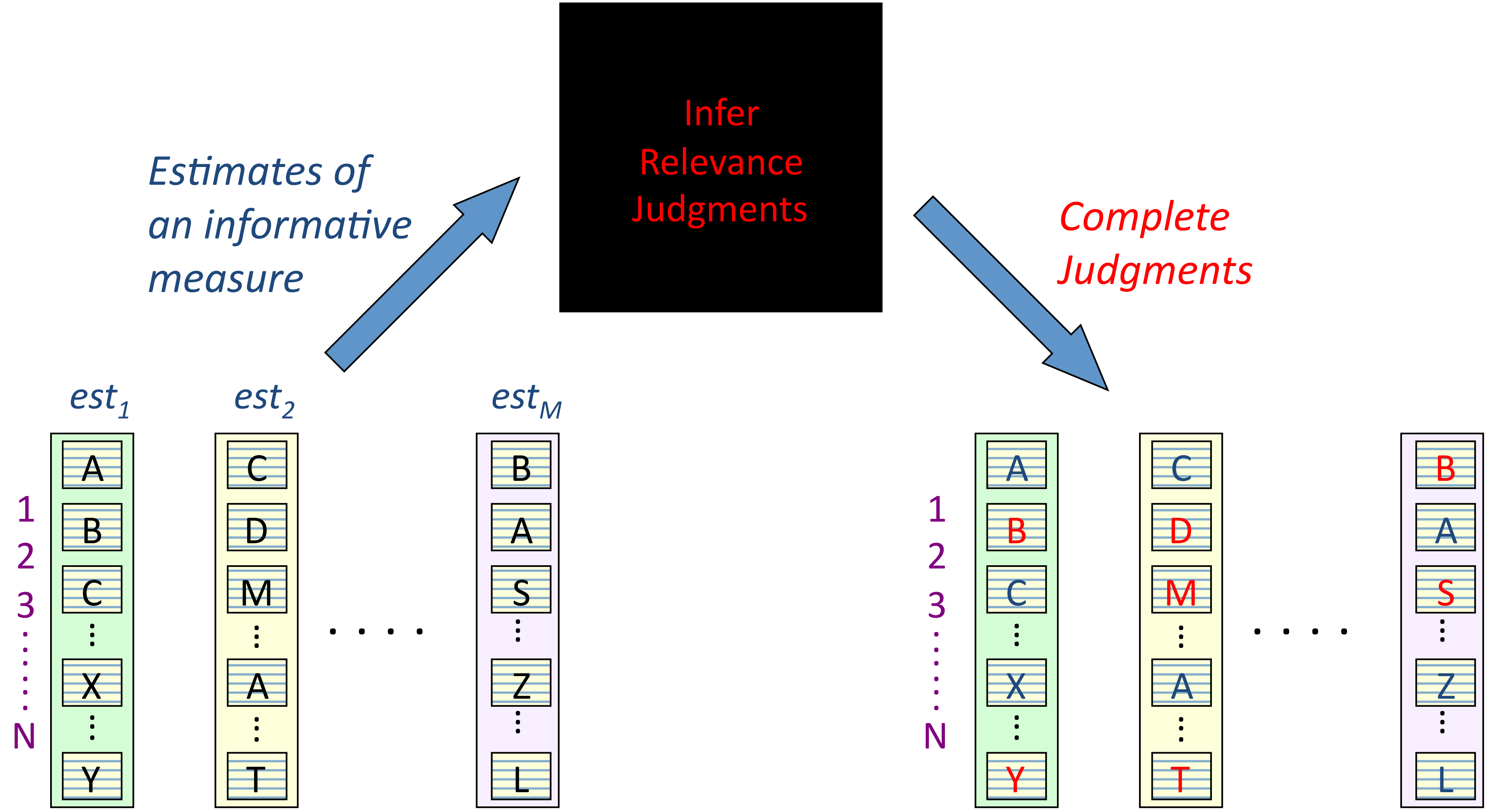
Low-Cost Evaluation (5)

- Inferring relevance judgments
  - Through Sampling (optimization approach)
    - Aslam and Yilmaz CIKM07
  - Document similarities/cluster hypothesis
    - Carterette and Allan CIKM07, Buttcher et al SIGIR07
  - Clicks and other user behavior features
    - Agrawal et al WSDM09, …

Inferring Relevance Judgments through Sampling

- Judge *some* documents
- *Estimate* the value of an *informative measure* using the judged documents
- Infer relevance of unjudged documents

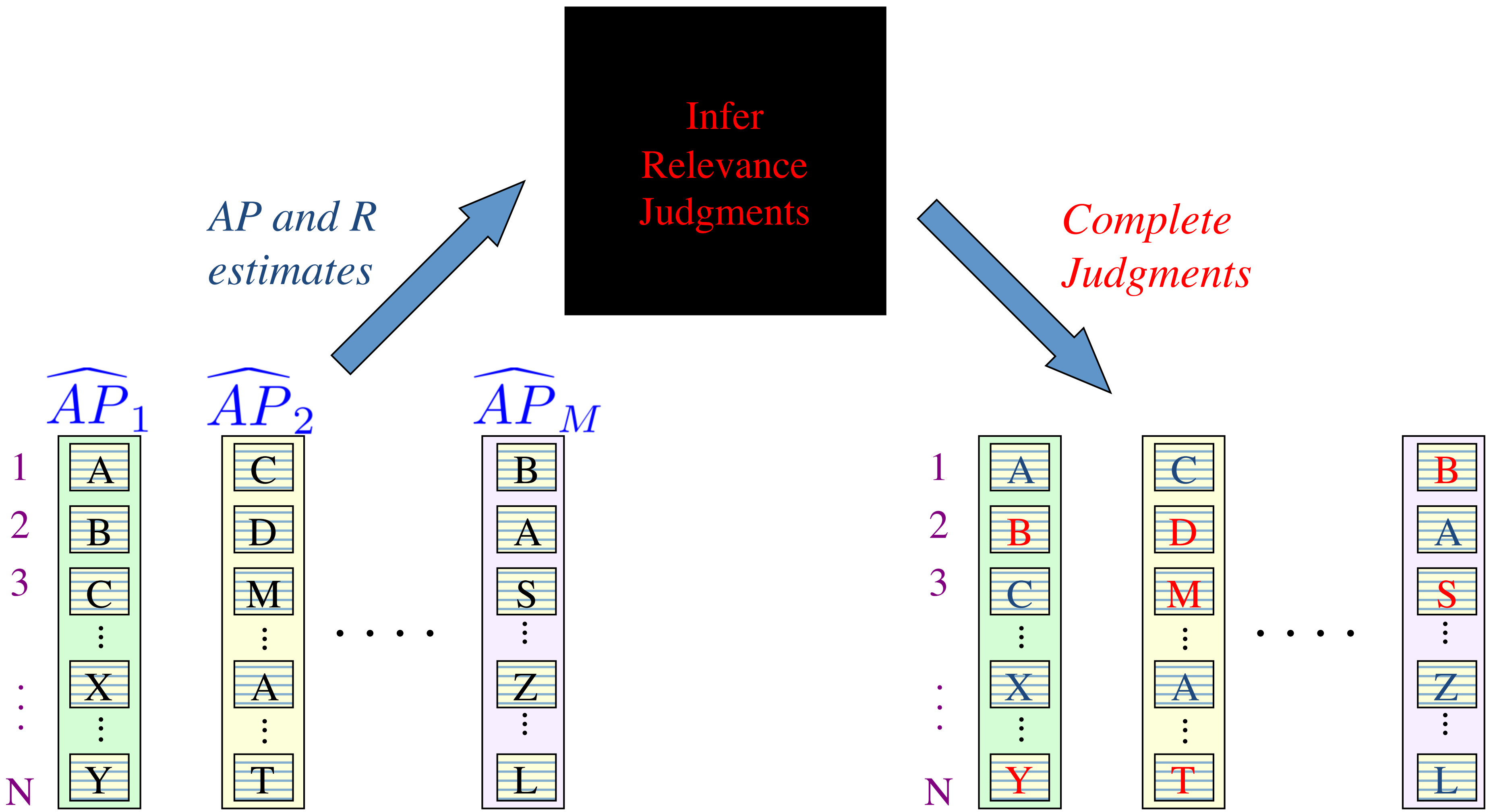
Proposed Solution: Inferring Relevance Judgments



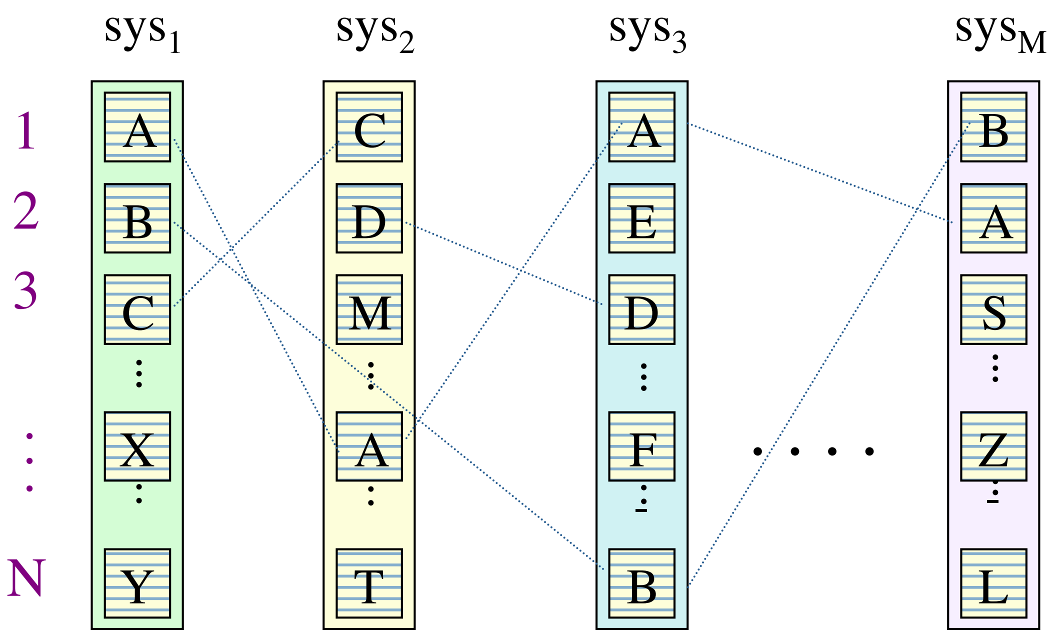
Inferring Relevance Judgments

- Average precision is highly informative [Aslam et al SIGIR05]
  - Given the value of AP of a system, accurately infer relevance of documents
- Given AP values of *multiple systems*, infer relevance of documents
- Given AP *estimates* of multiple systems, infer relevance of unjudged documents
  - E.g., statistical method to estimate AP

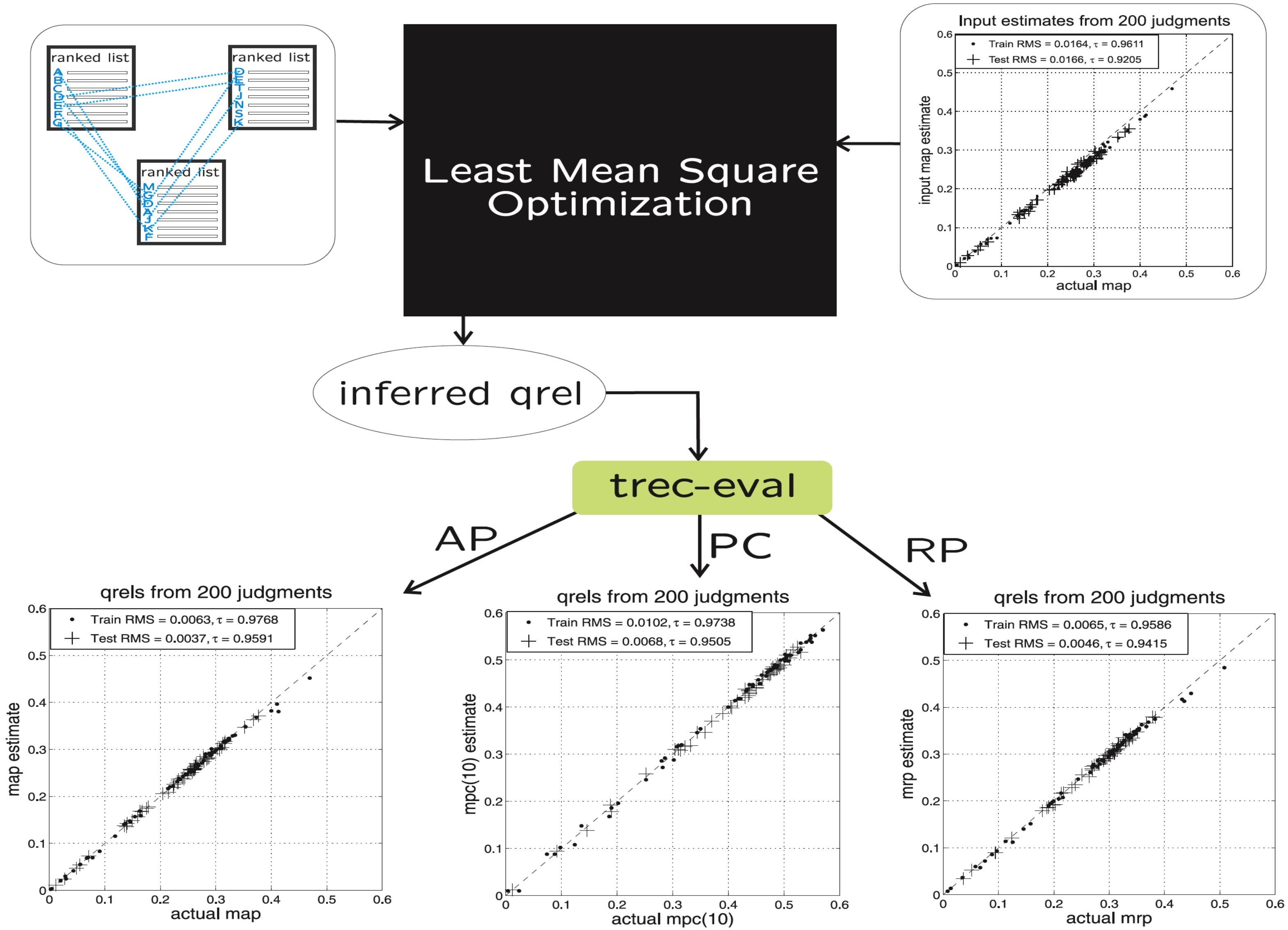
Inferring Relevance Judgments: Setup



Document Constraints



Inferring Relevance Judgments : Methodology



Inferring Relevance Judgments : Methodology

- Input :
  - Ranked list of documents
  - AP estimates associated with these lists
  - R estimate for the topic

- Goal : Assign *binary* relevance values to each document

- Optimization : Average precisions must be *close* to the given average precision estimates
  - *Minimize : Mean Squared Error*

- Constraints
  1. Total number of relevant documents is R_{est}
  2. Documents in multiple lists have the same relevance.

Inferring Relevance Judgments : Methodology

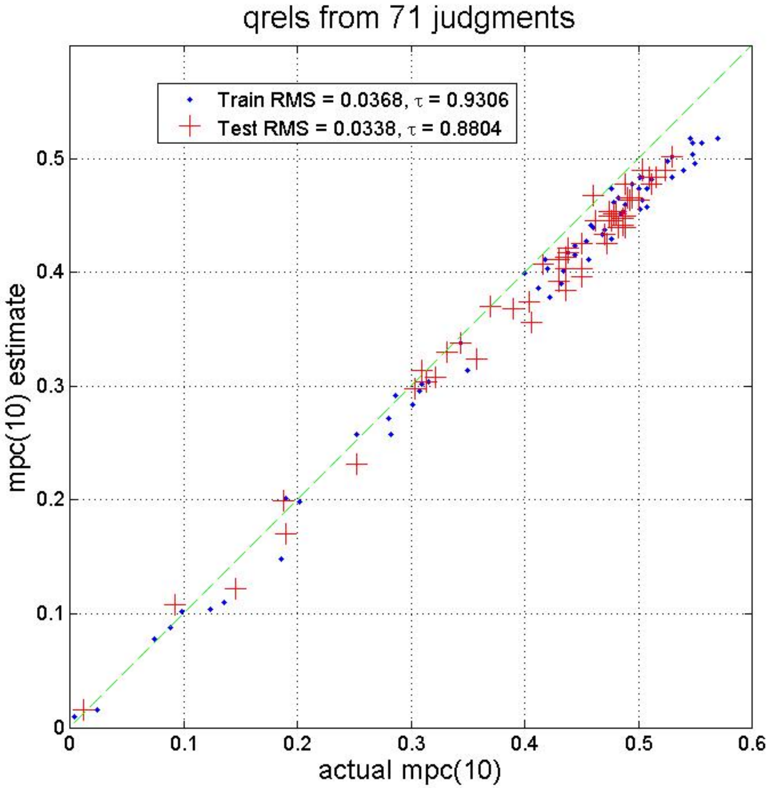
- Constrained integer optimization problem: INTRACTABLE!
- Allow probabilistic relevance assessments [Aslam et al SIGIR05]
  - p_i: probability that document at rank *i* is relevant

$$E[AP] = \frac{1}{R}\sum_{i=1}^{N}\left(\frac{p_i}{i}\left(1+\sum_{j=1}^{i-1}p_j\right)\right)$$

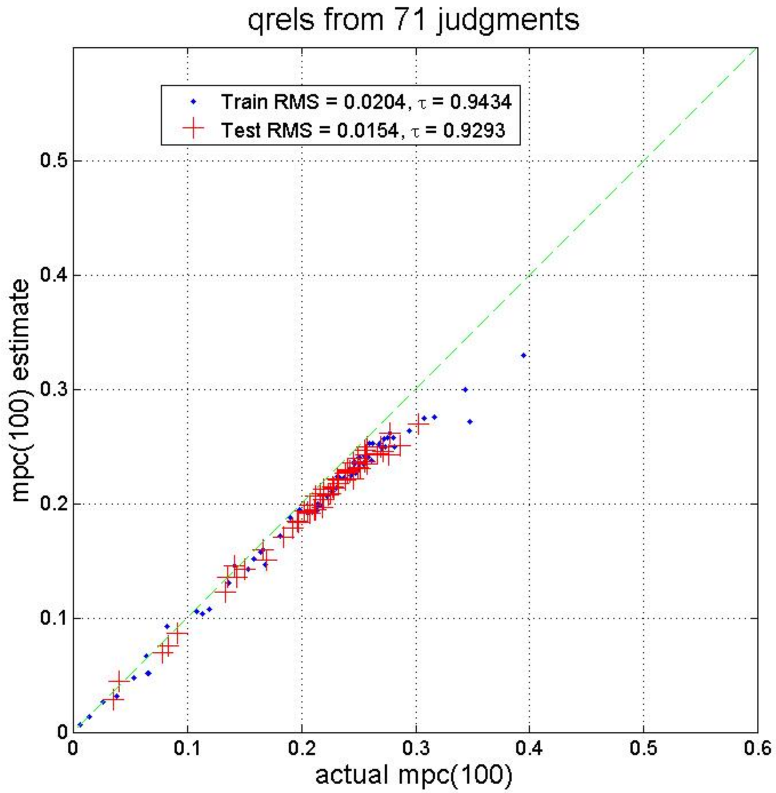
- *Randomized rounding* to convert probabilistic judgments to binary
  - Assign relevance score 1 with probability p_i and 0 otherwise.

How Good are the Inferred Qrels: 71 (4.1%) Judgments?

MPC(10)



MPC(100)



Difference of Inferred Qrels from Actual Qrels

Docs judged	Precision	Recall	F_1
1.7%	0.5562	0.3833	0.4171
4.1%	0.5919	0.5495	0.5332
6.3%	0.6243	0.6004	0.5880
11.7%	0.7068	0.6887	0.6906
21.8%	0.8101	0.7694	0.7835

Today's Outline

- Low cost evaluation

1. Depth-k pooling (standard method)

2. Evaluating without judgments (automatic eval)
3. Finding relevance documents as quickly as possible

4. Computing measures with incomplete judgments
5. Estimating measures
6. Inferring relevance judgments