A Statistical Method for System Evaluation Using Incomplete Judgments

Jay Aslam
Virgil Pavlu
Emine Yilmaz

Northeastern University

The Problem

The Problem

- Large-scale retrieval evaluation is expensive
  - 50k to 100k documents assessed per year in TREC

The Problem

- Large-scale retrieval evaluation is expensive
  - 50k to 100k documents assessed per year in TREC
- Q1: How to scale to larger collections?

The Problem

- Large-scale retrieval evaluation is expensive
  - 50k to 100k documents assessed per year in TREC
- Q1: How to scale to larger collections?
- Q2: How do other organizations cope?

The Problem

- Large-scale retrieval evaluation is expensive
  - 50k to 100k documents assessed per year in TREC
- Q1: How to scale to larger collections?
- Q2: How do other organizations cope?
- A: More efficient evaluation techniques...

Our Contributions: Ideas

- View performance measures as the outcome of a random experiment
  - PC(k), AP, RP, ...
- Apply sampling theory to efficiently estimate this outcome
  - Analogy: forecasting election results
- Many other benefits...

Our Contributions: Methodology

Our Contributions: Methodology

- An accurate and efficient method for estimating measures from a random sample
  - AP, PC(k), RP, R, ...
  - unbiased by statistical design
  - low variance via non-uniform sampling
  - scalable (many lists, many measures, simultaneously)
  - reusable (can estimate quality of unseen lists)

Outline

- A simple example in detail: estimating PC(k)
- Estimating AP
- Results on TREC data

Consider PC(1000)

Consider PC(1000)

- Obvious solution:
  - examine top 1000 documents
  - return #rel-seen/1000

Consider PC(1000)

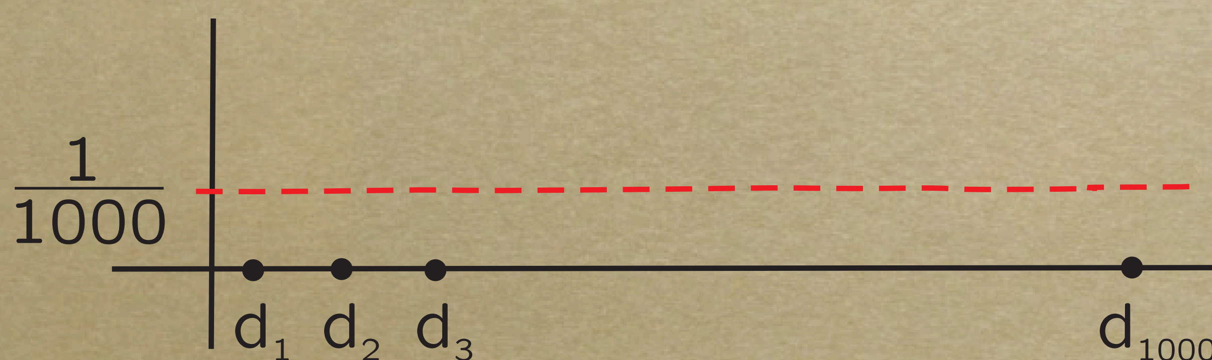
- Obvious solution:
  - examine top 1000 documents
  - return #rel-seen/1000
- Alternate solution:
  - uniformly sample documents from top 1000
  - judge random sample
  - return #rel-judged/sample-size

Carefully Define Random Experiment

- Sample space: top 1000 docs
  - $i \in \{1, \ldots, 1000\}$
- Distribution: uniform over top 1000 docs
  - $p_i = 1/1000$
- Random variable: X = relevance of doc
  - $x_i = rel(i) = \begin{cases} 0 & \text{if } d_i \text{ is non-relevant} \\ 1 & \text{if } d_i \text{ is relevant} \end{cases}$

Verify Expectation

$$
\begin{aligned}
E[X] &= \sum_{i=1}^{1000} p_i \cdot x_i \\
&= \frac{1}{1000} \sum_{i=1}^{1000} x_i \\
&= \frac{\#\text{rel}}{1000}
\end{aligned}
$$



Law of Large Numbers
Central Limit Theorem

- Average of random sample converges to mean

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j \to E[X]$$

- Average of random sample rapidly becomes Gaussian

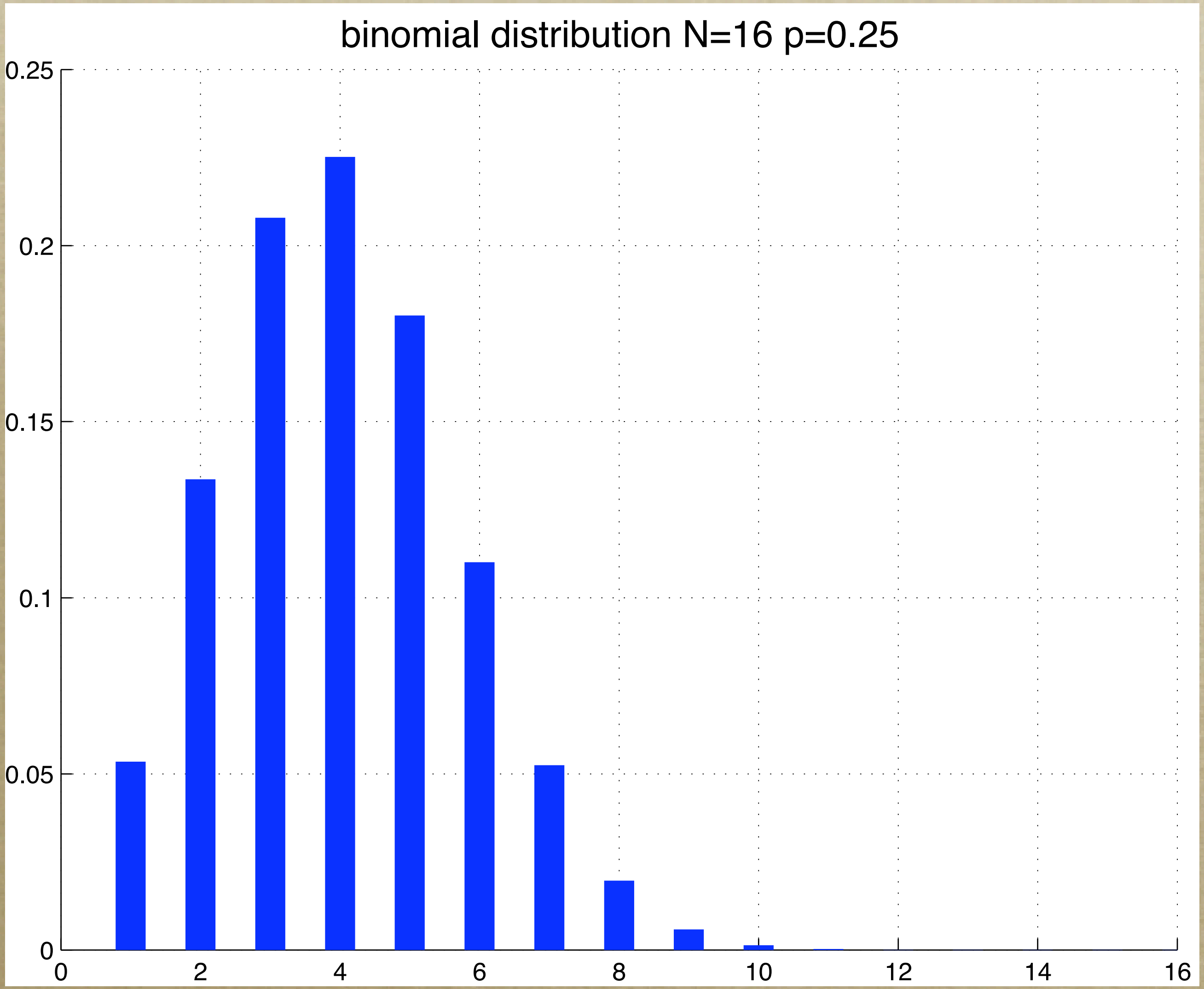
$$\begin{aligned} E[X] &= \mu \\ Var[X] &= \sigma^2 \end{aligned}$$

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j \quad \to \quad N(\mu, \sigma/\sqrt{n})$$

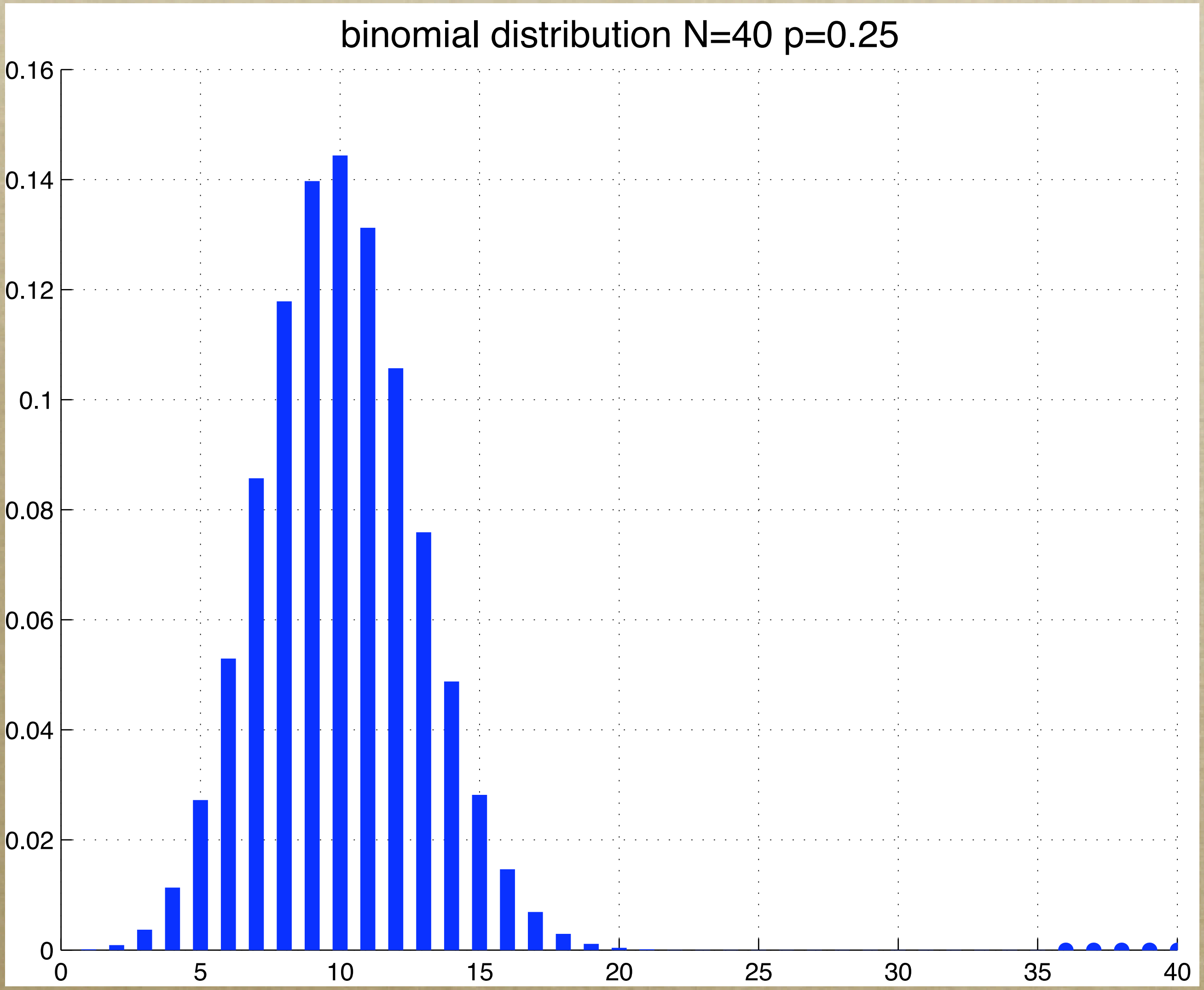
Back to PC(1000)

- Suppose that, unknown to us, PC(1000) = p = 0.25
  - i.e., there were 250 rel docs in top 1000
- How many samples until we could estimate that accurately, say +/- 0.03?
- LLN & CLT

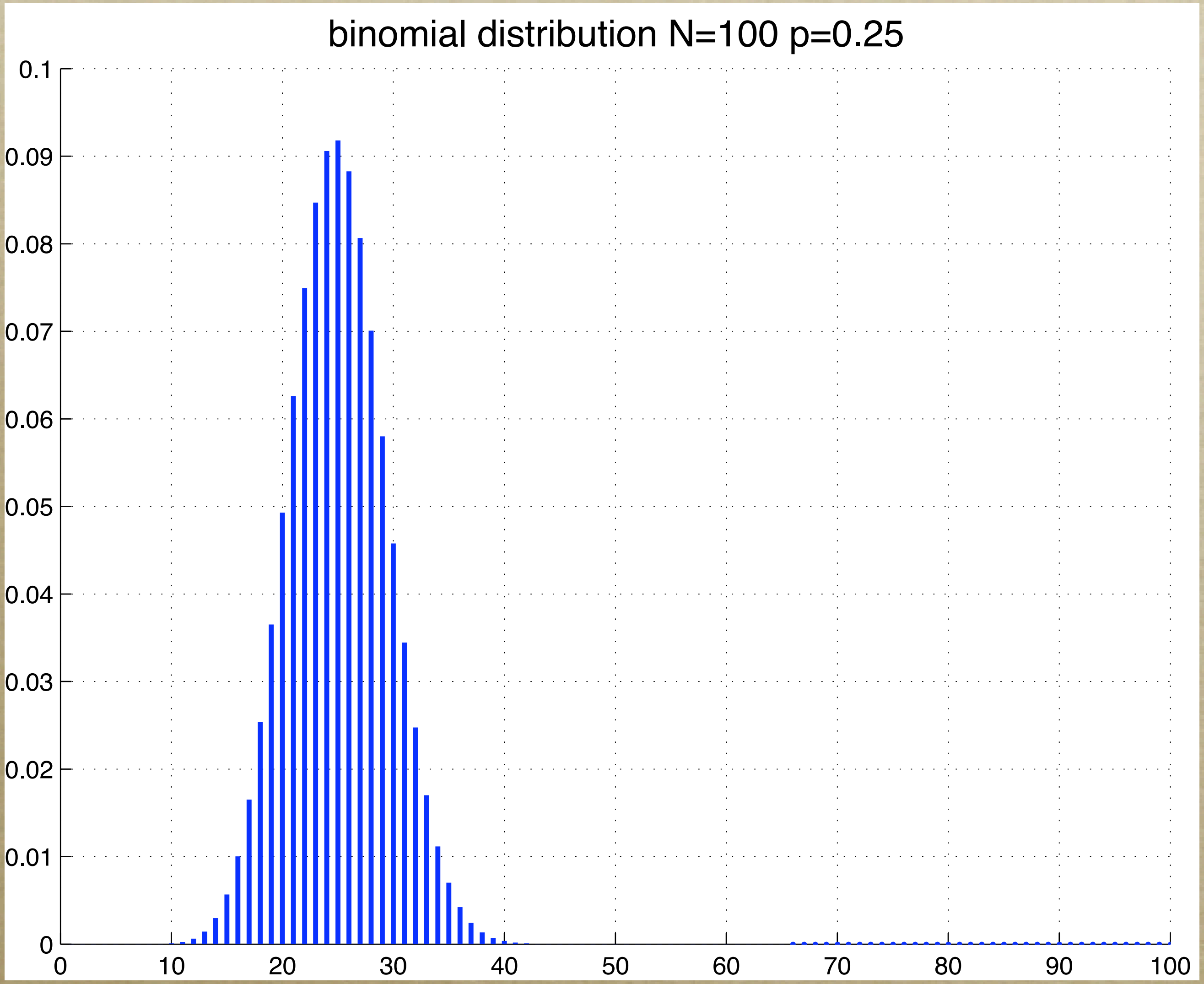
LLN and CLT Illustration: N=16



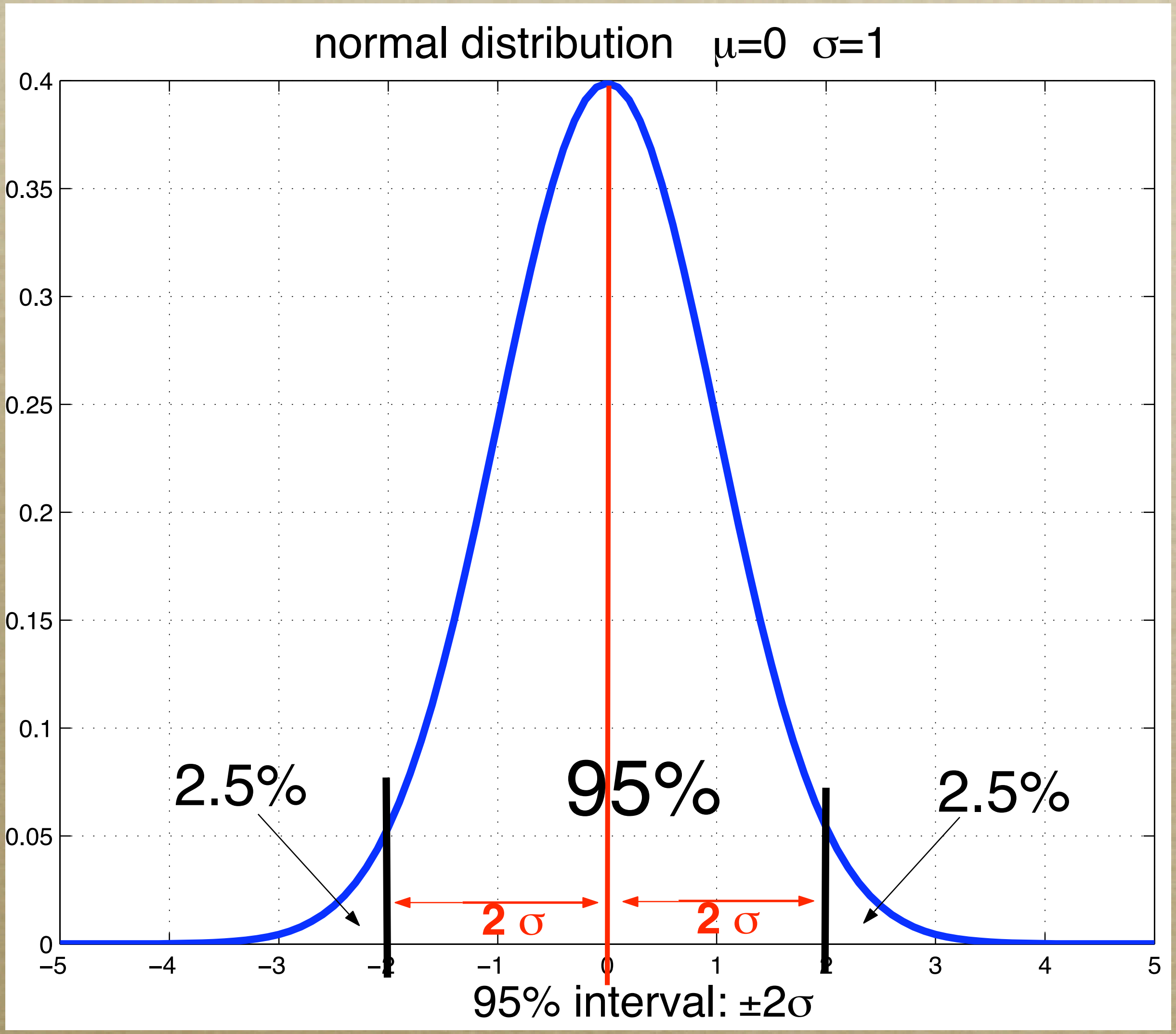
LLN and CLT Illustration: N=40



LLN and CLT Illustration: N=100



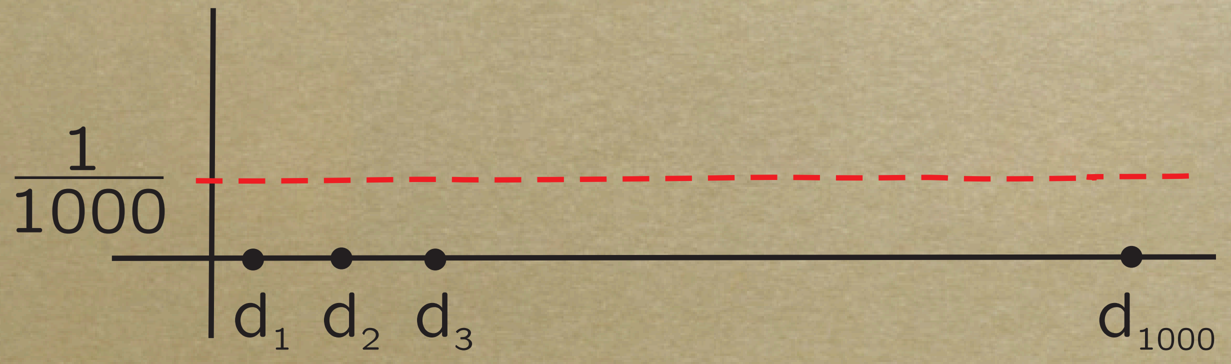
Normal Distribution: 95% CI



Sample Size Calculation

$$\begin{aligned} Var[X] &= p \cdot (1-p) \\ &= 3/16 \\ Var[\overline{X}] &= (3/16)/n \\ \sigma &= \sqrt{(3/16)/n} \end{aligned}$$

$$\begin{aligned} 2\sigma &= 2\sqrt{(3/16)/n} \\ &= 0.03 \\ n &\approx 833 \end{aligned}$$



Questions and Concerns

Questions and Concerns

- That's a lot of samples...

Questions and Concerns

- That's a lot of samples...
- Is there a "smarter" sampling strategy?
  - Yes: importance sampling

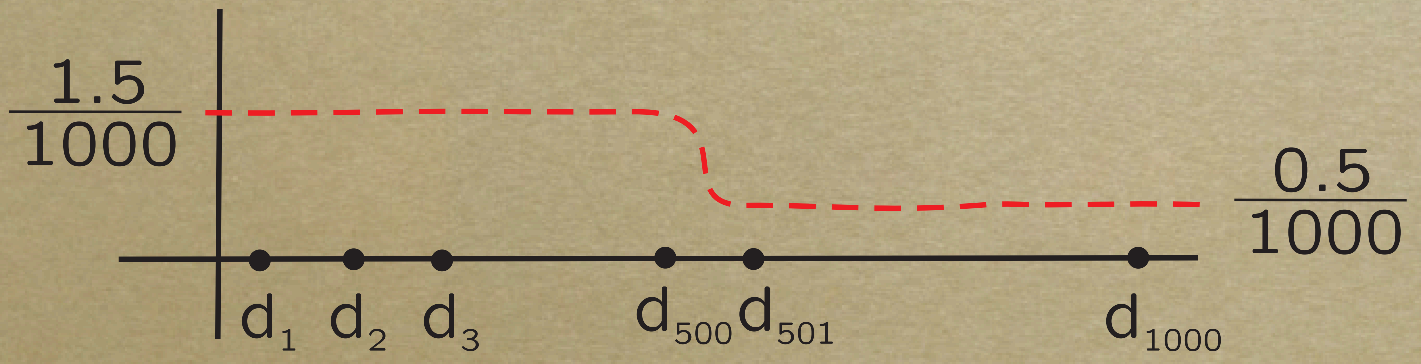
Questions and Concerns

- That's a lot of samples...
- Is there a "smarter" sampling strategy?
  - Yes: importance sampling
- What about multiple lists (multiple measures) with one sample?
  - Yes: scaling factors

PC(1000): Non-uniform Sampling

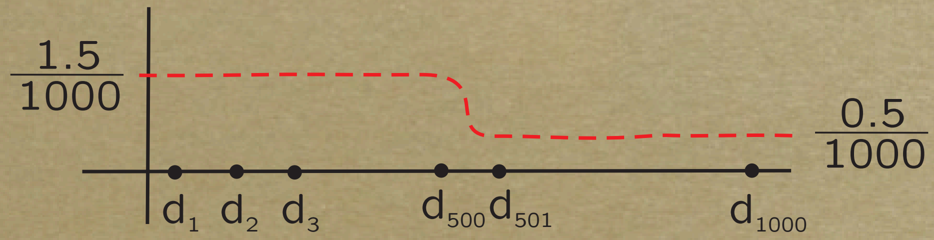
Q: How to correct for mean?

A: Scaling factors.



New Random Experiment

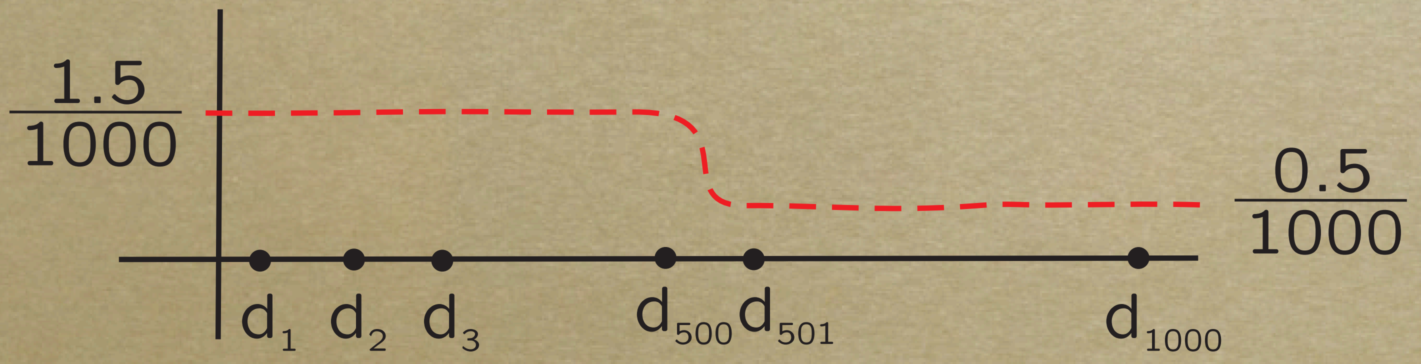
- Sample space: top 1000 docs
  - $i \in \{1, \ldots, 1000\}$
- Distribution: non-uniform
  - $p_i = \begin{cases} 1.5/1000 & 1 \leq i \leq 500 \\ 0.5/1000 & 501 \leq i \leq 1000 \end{cases}$
- Random variable: fix expectation (scaling factors)
  - $x_i = \begin{cases} rel(i) \cdot 2/3 & 1 \leq i \leq 500 \\ rel(i) \cdot 2 & 501 \leq i \leq 1000 \end{cases}$



E[X] as before...

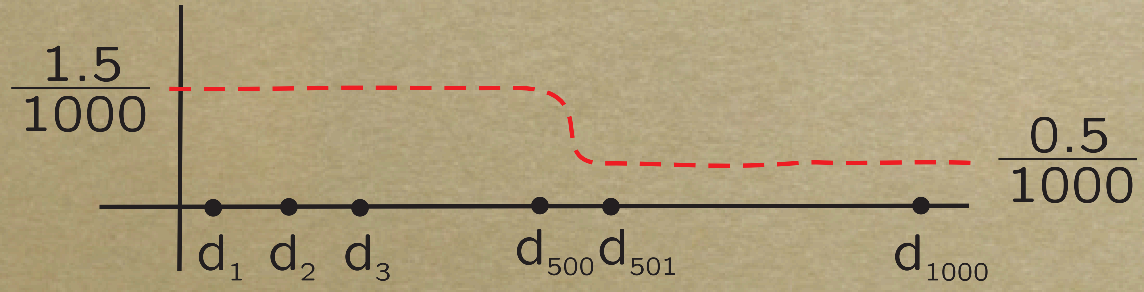
What About Variance?

- Consider two cases:
  - All 250 relevant docs in top half (good)
  - All 250 relevant docs in bottom half (bad)



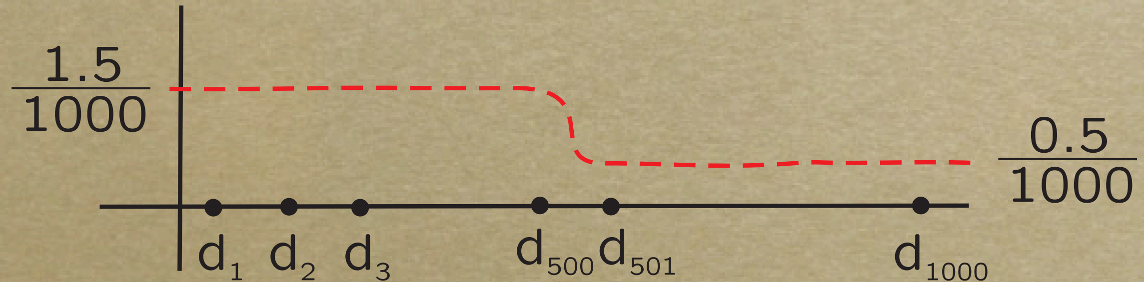
Case 1: Relevant Docs in Top Half

$$\begin{aligned}
Var[X] &= E[X^2] - E^2[X] \\
&= \sum_{i=1}^{1000} p_i \cdot {x_i}^2 - (1/4)^2 \\
&= \sum_{i=1}^{500} p_i \cdot {x_i}^2 + \sum_{i=501}^{1000} p_i \cdot {x_i}^2 - 1/16 \\
&= 250 \cdot \frac{3/2}{1000} \cdot (2/3)^2 + 0 - 1/16 \\
&= 1/6 - 1/16 \\
&= 5/48 \\
&\approx 0.1042
\end{aligned}$$



Case 2: Relevant Docs in Bottom Half

$$\begin{aligned}
Var[X] &= E[X^2] - E^2[X] \\
&= \sum_{i=1}^{1000} p_i \cdot {x_i}^2 - (1/4)^2 \\
&= \sum_{i=1}^{500} p_i \cdot {x_i}^2 + \sum_{i=501}^{1000} p_i \cdot {x_i}^2 - 1/16 \\
&= 0 + 250 \cdot \frac{1/2}{1000} \cdot 2^2 + 0 - 1/16 \\
&= 1/2 - 1/16 \\
&= 7/16 \\
&\approx 0.4375
\end{aligned}$$



Comparison

- Original variance: 3/16 = 0.1875
- Case 1 variance: 5/48 = 0.1042
- Case 2 variance: 7/16 = 0.4375
- Effect on sample size required...
  - 833 (original)
  - 463 (Case 1)
  - 1944 (Case 2)

Conclusion

Conclusion

- Sample where the relevant docs are!
  - importance sampling

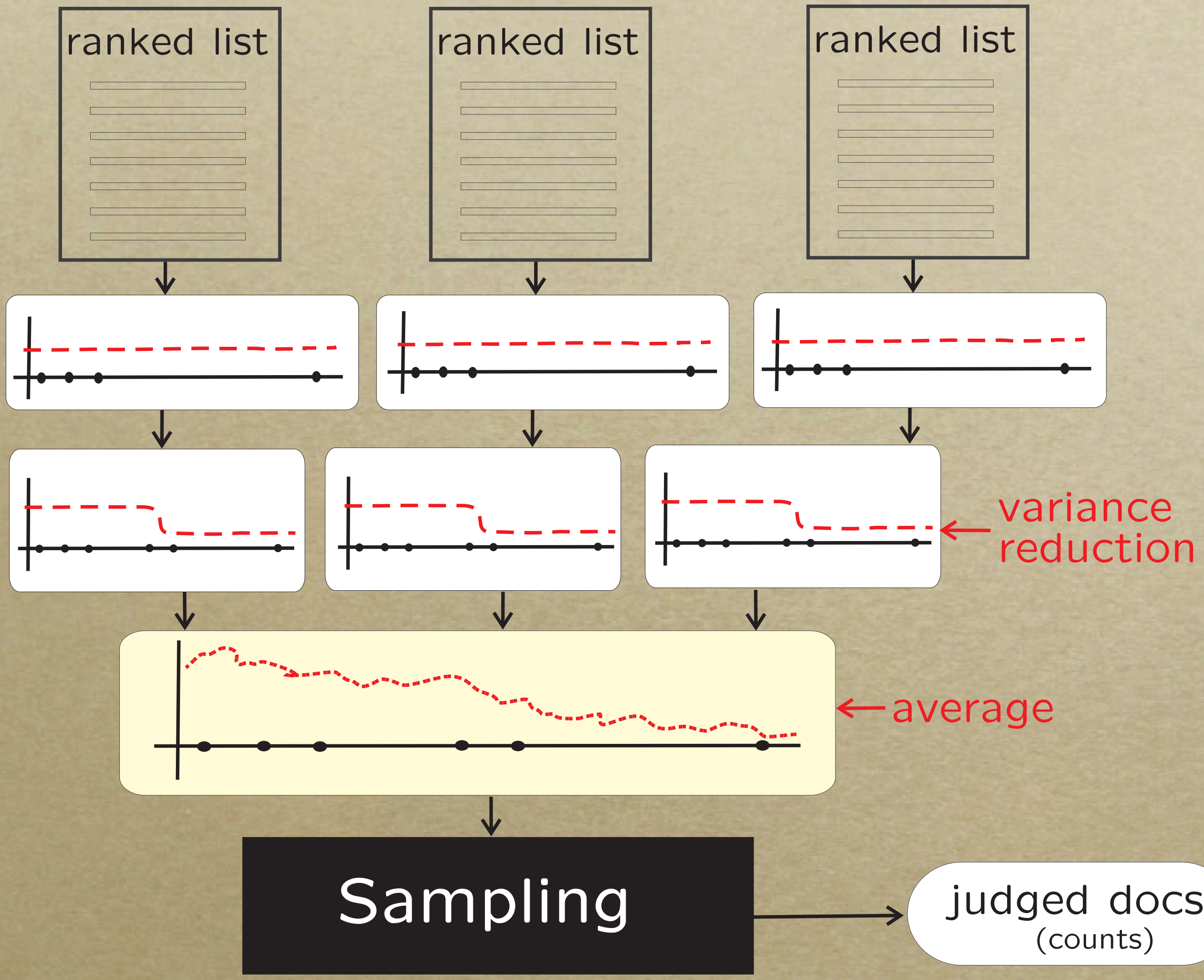
Conclusion

- Sample where the relevant docs are!
  - importance sampling
- Use scaling factors to correct expectation
  - ratio between "required" and actual distribution

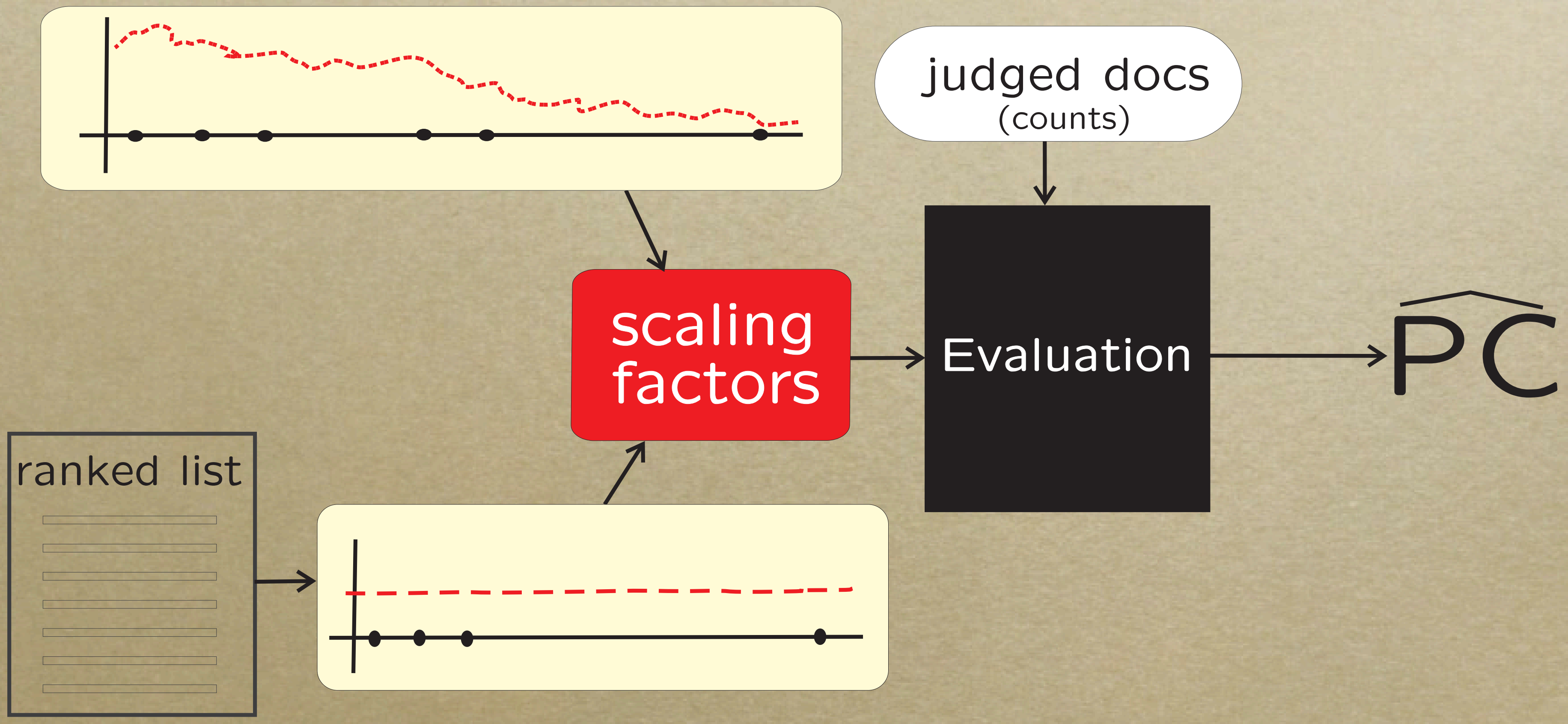
Conclusion

- Sample where the relevant docs are!
  - importance sampling
- Use scaling factors to correct expectation
  - ratio between “required” and actual distribution
- Many benefits...
  - variance reduction
  - many measures from one sample
  - many lists from one sample
  - reusability

PC(k): Sampling for Many Runs



Estimating PC



What About AP?

List:

R	1/1
N	
R	2/3
N	
N	
R	3/6
N	
N	
N	
R	4/10

$$\mathrm{AP} = \frac{1 + 2/3 + 3/6 + 4/10}{4} \approx 0.6417$$

Average Precision

$$\begin{aligned}
\text{AP} &= \frac{1}{R} \cdot \sum_{i\,:\,rel(i)=1} \text{PC}(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^{N} rel(i) \cdot \text{PC}(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^{N} rel(i) \sum_{j=1}^{i} rel(j)/i \\
&= \frac{1}{R} \cdot \sum_{1 \le j \le i \le N} \frac{1}{i} \cdot rel(i) \cdot rel(j)
\end{aligned}$$

We will sample for the summation (and R)...

Analogy to PC(1000)

$$\begin{aligned}
\text{AP} &= \frac{1}{R} \cdot \sum_{i\,:\,rel(i)=1} \text{PC}(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^{N} rel(i) \cdot \text{PC}(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^{N} rel(i) \sum_{j=1}^{i} rel(j)/i \\
&= \frac{1}{R} \cdot \sum_{1 \le j \le i \le N} \frac{1}{i} \cdot rel(i) \cdot rel(j)
\end{aligned}$$

$$\begin{aligned}
\text{PC}(1000) &= \\
&\vdots \\
&\vdots \\
&= \sum_{i=1}^{1000} \frac{1}{1000} \cdot rel(i)
\end{aligned}$$

Analogy to PC(1000)

$$
\begin{aligned}
\mathrm{AP} &= \frac{1}{R} \cdot \sum_{i\,:\,rel(i)=1} \mathrm{PC}(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^{N} rel(i) \cdot \mathrm{PC}(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^{N} rel(i) \sum_{j=1}^{i} rel(j)/i \\
&= \frac{1}{R} \cdot \sum_{1 \le j \le i \le N} \frac{1}{i} \cdot rel(i) \cdot rel(j)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{PC}(1000) &= \\
&\vdots \\
&\vdots \\
&= \sum_{i=1}^{1000} \frac{1}{1000} \cdot rel(i)
\end{aligned}
$$

Distribution

Analogy to PC(1000)

$$\begin{aligned}
\text{AP} &= \frac{1}{R} \cdot \sum_{i\,:\,rel(i)=1} \text{PC}(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^{N} rel(i) \cdot \text{PC}(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^{N} rel(i) \sum_{j=1}^{i} rel(j)/i \\
&= \frac{1}{R} \cdot \sum_{1 \leq j \leq i \leq N} \frac{1}{i}\, rel(i) \cdot rel(j)
\end{aligned}$$

$$\begin{aligned}
\text{PC}(1000) &= \\
&\vdots \\
&\vdots \\
&= \sum_{i=1}^{1000} \frac{1}{1000}\, rel(i)
\end{aligned}$$

Random Variable

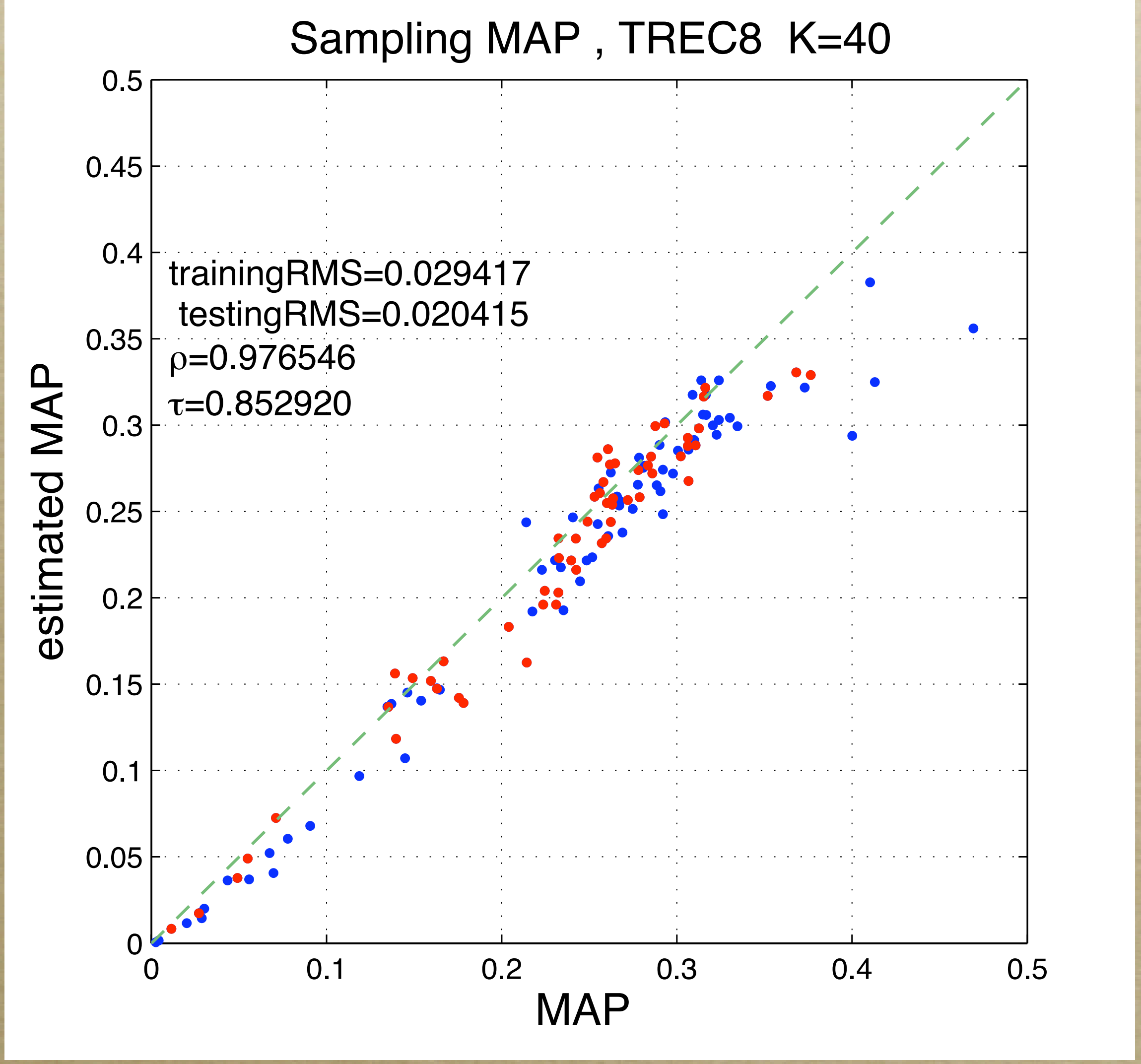
Define the AP Random Experiment

- Sample space:
  - all pairs of docs
- Distribution:
  - as defined above (1/i)
- Random variable:
  - product of relevances
- Many details...

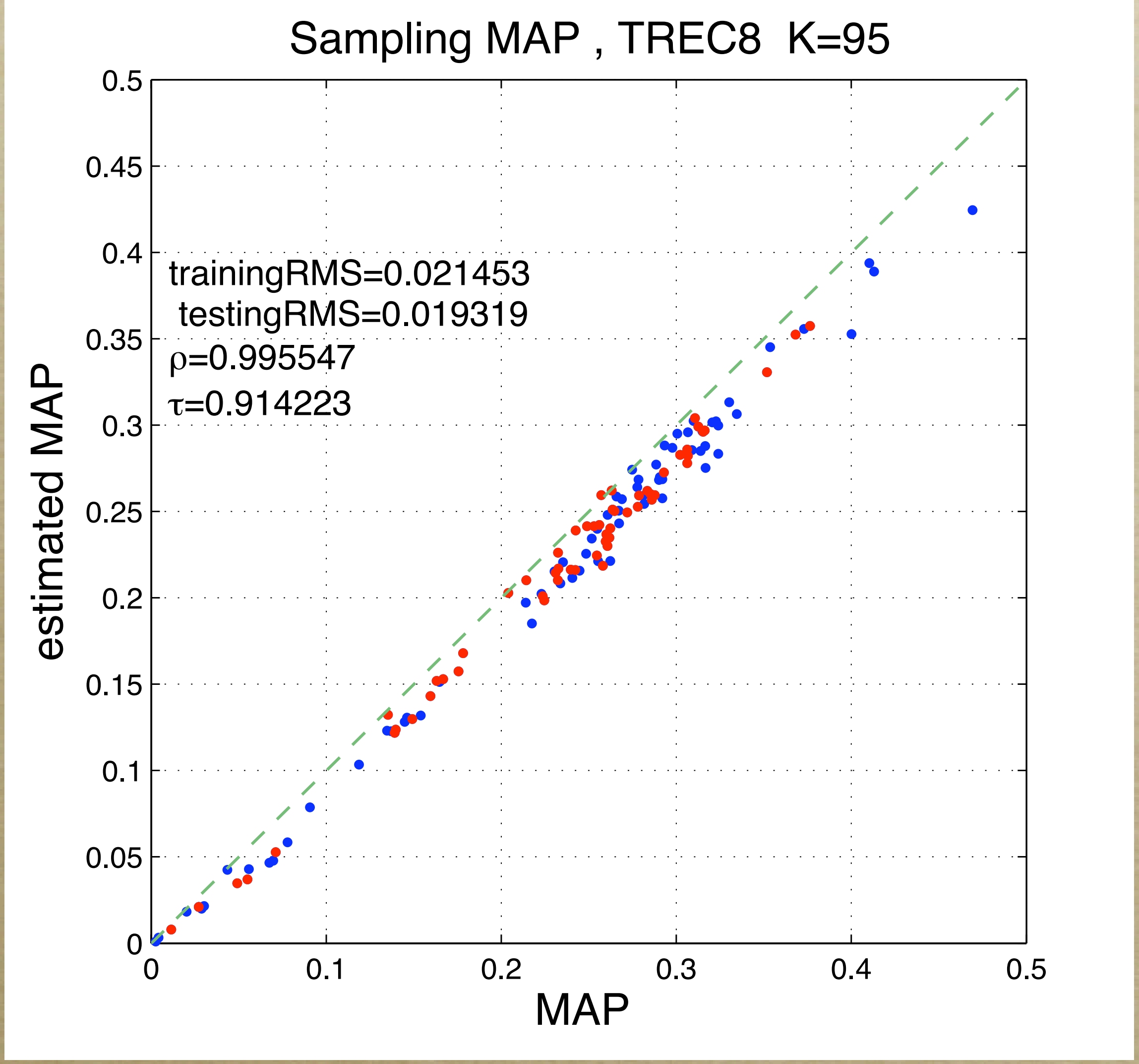
Experiments with TREC8 Data

- 50 queries
- 129 search engine runs per query
- Depth-100 pooling
  - 2/3 runs contributed to pool (training runs)
  - 1/3 did not (testing runs)
- Employ sampling methodology on training runs
  - single sample drawn per query
- Evaluate and compare
  - all runs, all measures

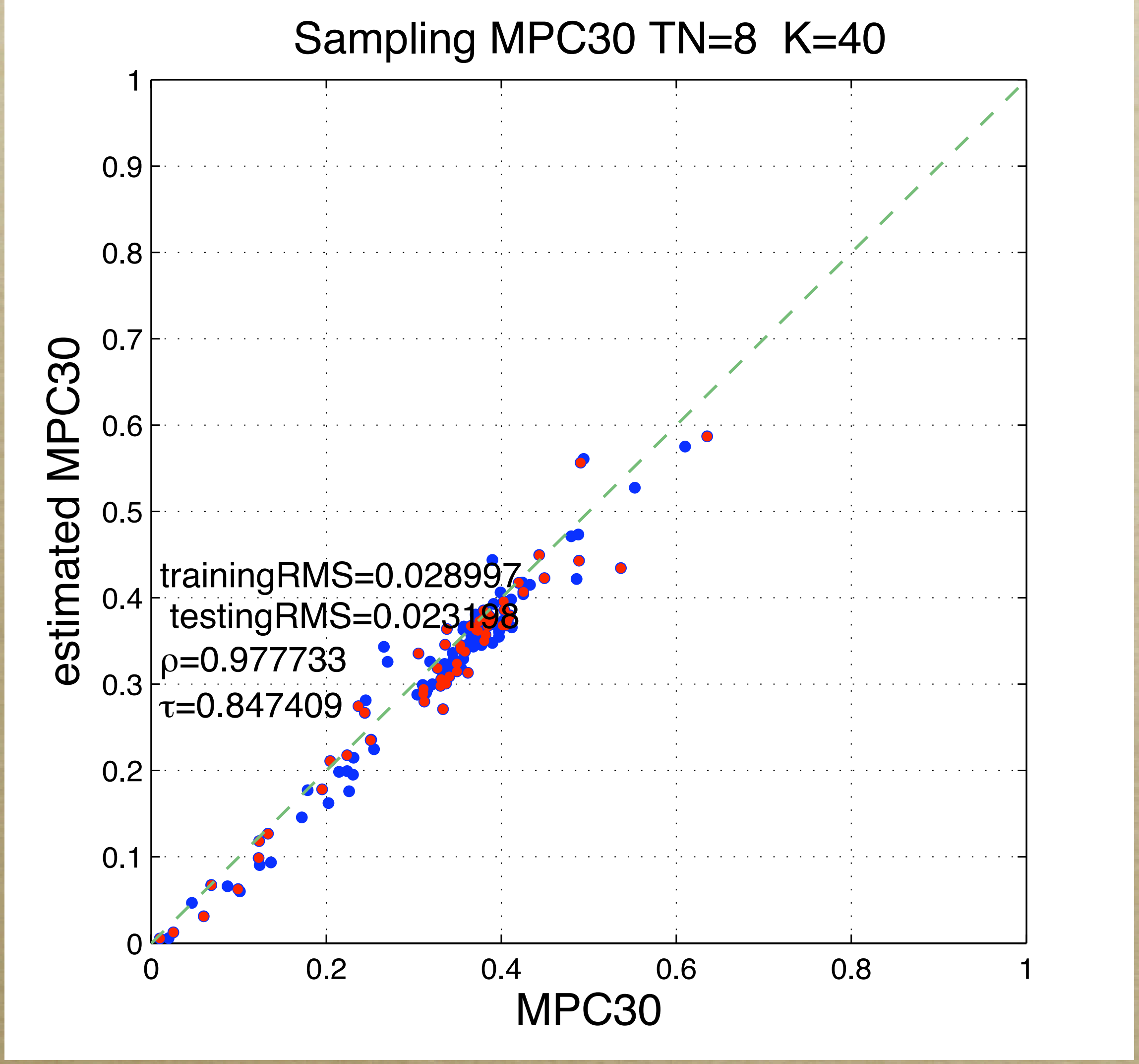
MAP Estimates: 40 JPQ



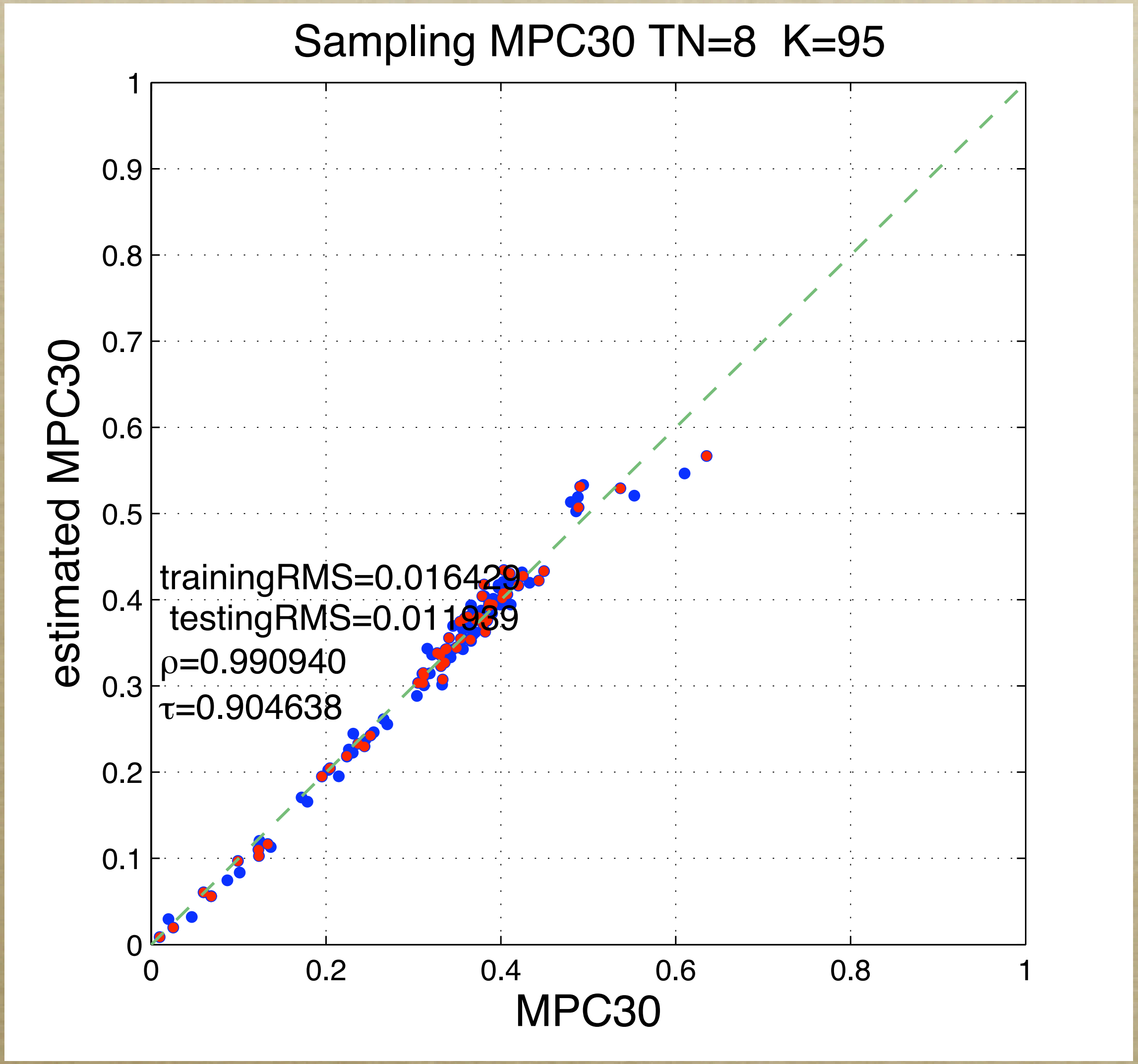
MAP Estimates: 95 JPQ



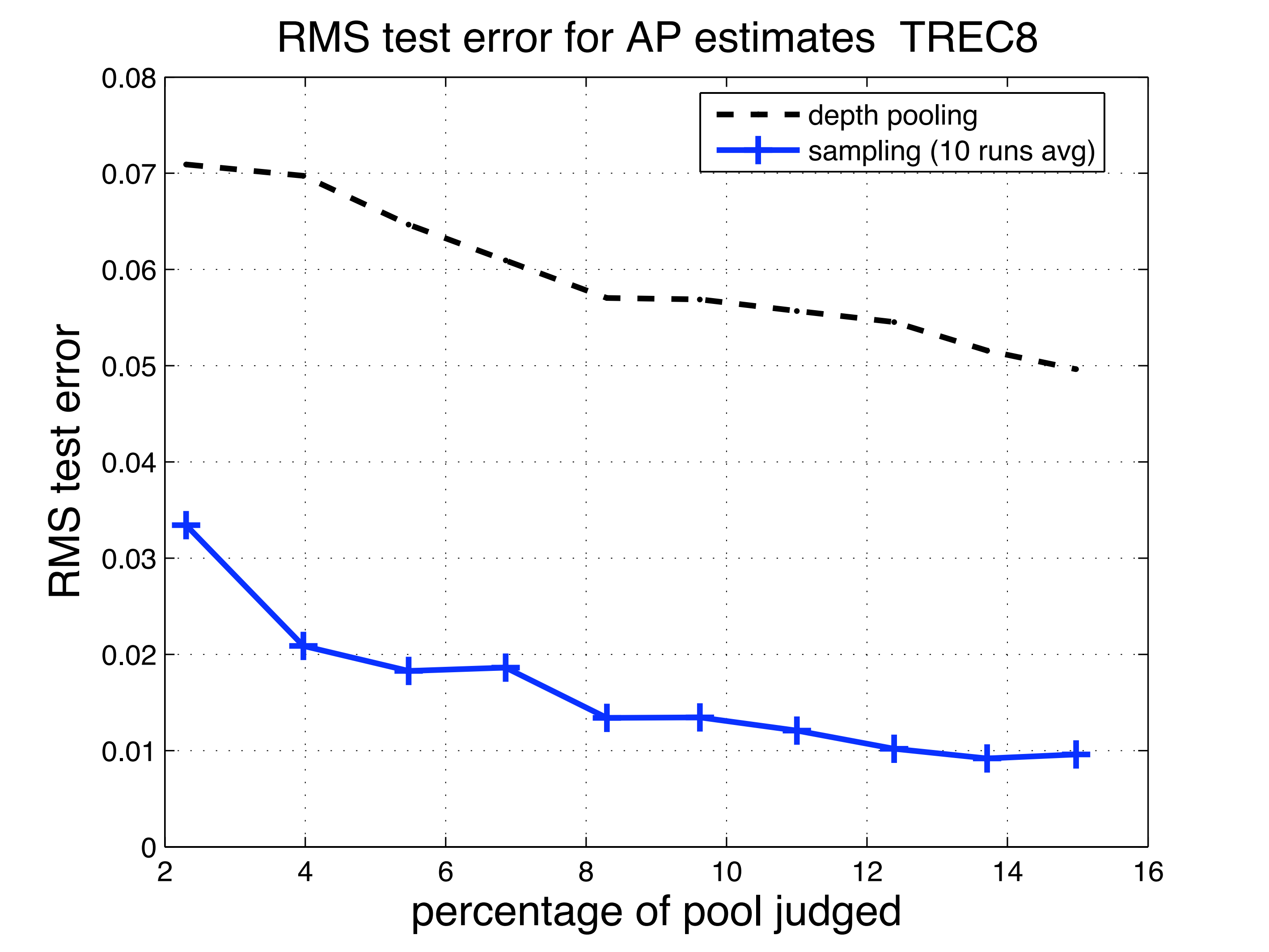
MPC(30) Estimates: 40 JPQ



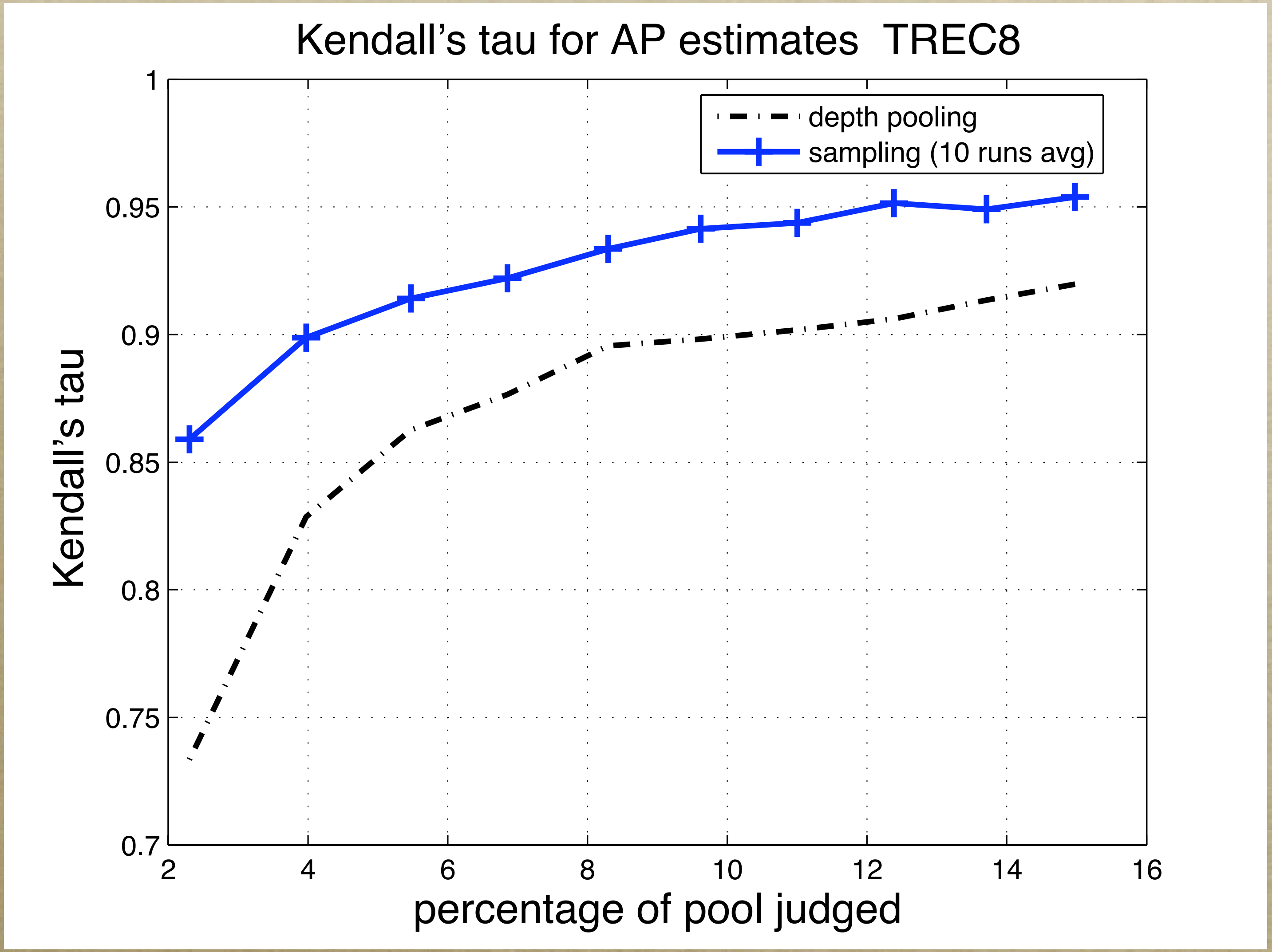
MPC(30) Estimates: 95 JPQ



MAP Summary: RMS



MAP Summary: Kendall’s tau



Summary and Future Work

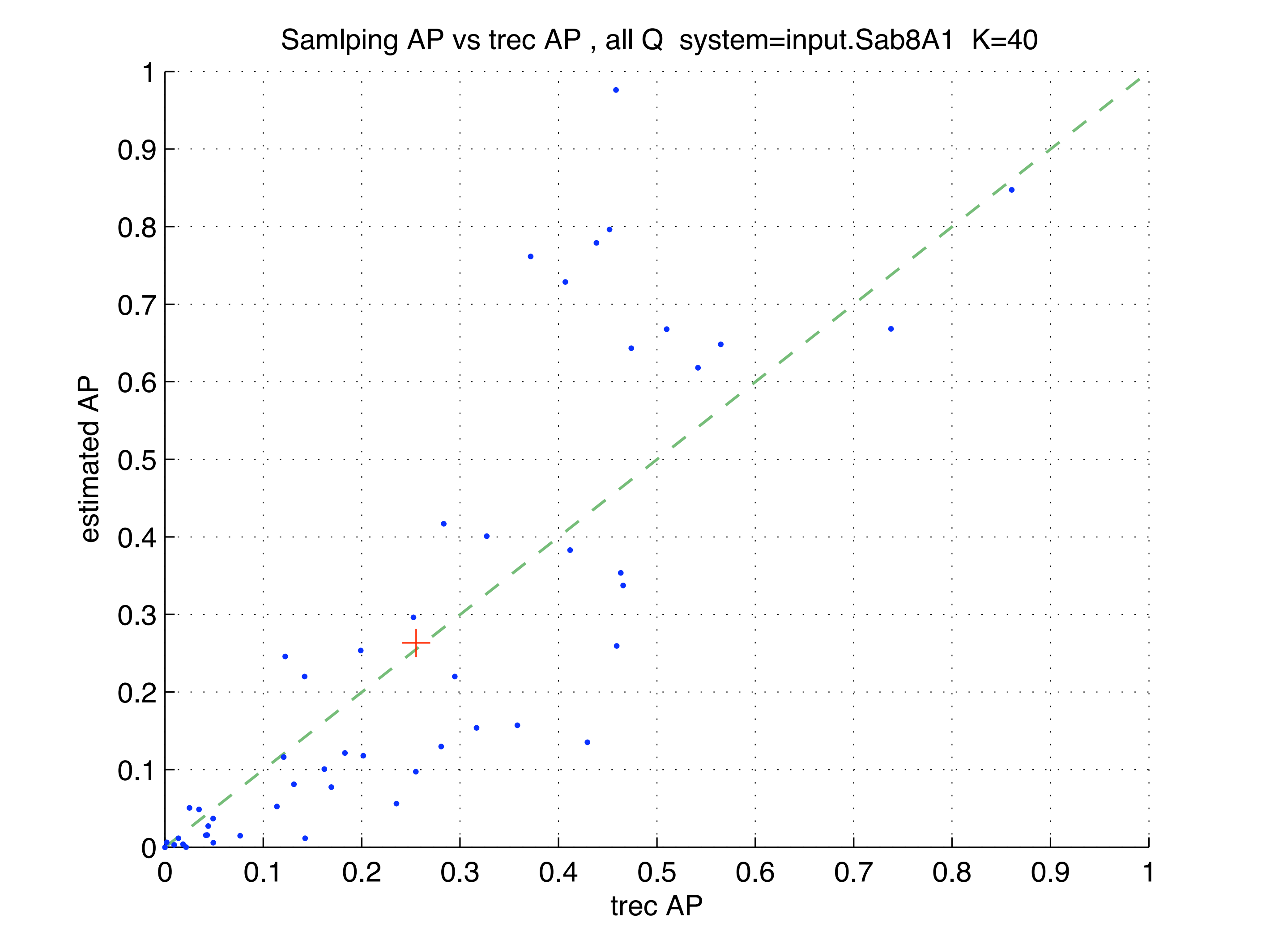
Summary and Future Work

- A statistical method for estimating performance
  - AP, RP, PC(k), R, ...
  - view measure as outcome of a random experiment
  - employ sampling theory
  - accurate, efficient estimates
  - reusable

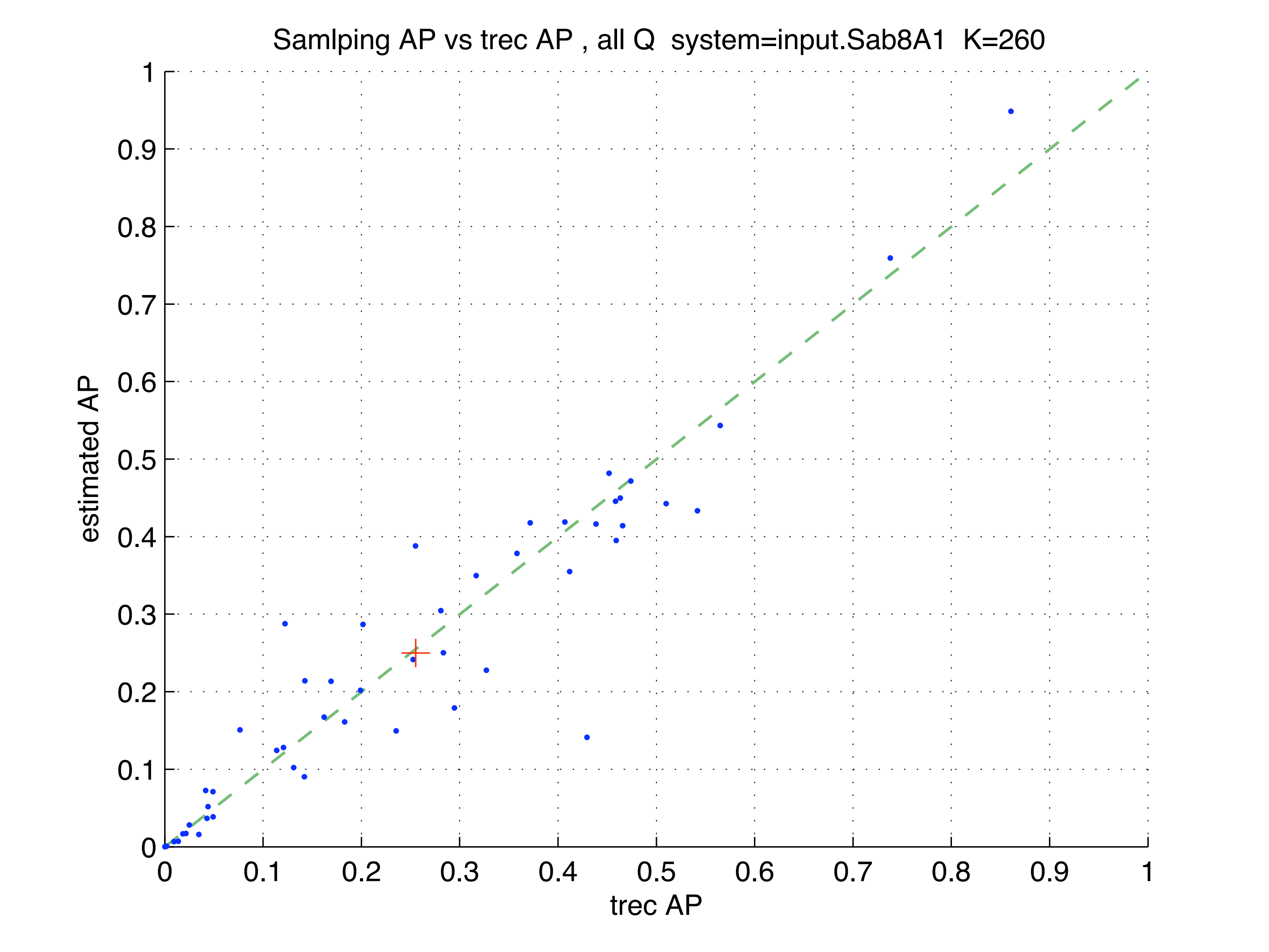
Summary and Future Work

- A statistical method for estimating performance
  - AP, RP, PC(k), R, ...
  - view measure as outcome of a random experiment
  - employ sampling theory
  - accurate, efficient estimates
  - reusable
- Future work:
  - optimal sampling distribution, lower bounds, etc.
  - explicit confidence intervals (estimate variance)
  - mixed strategies (greedy & statistical)

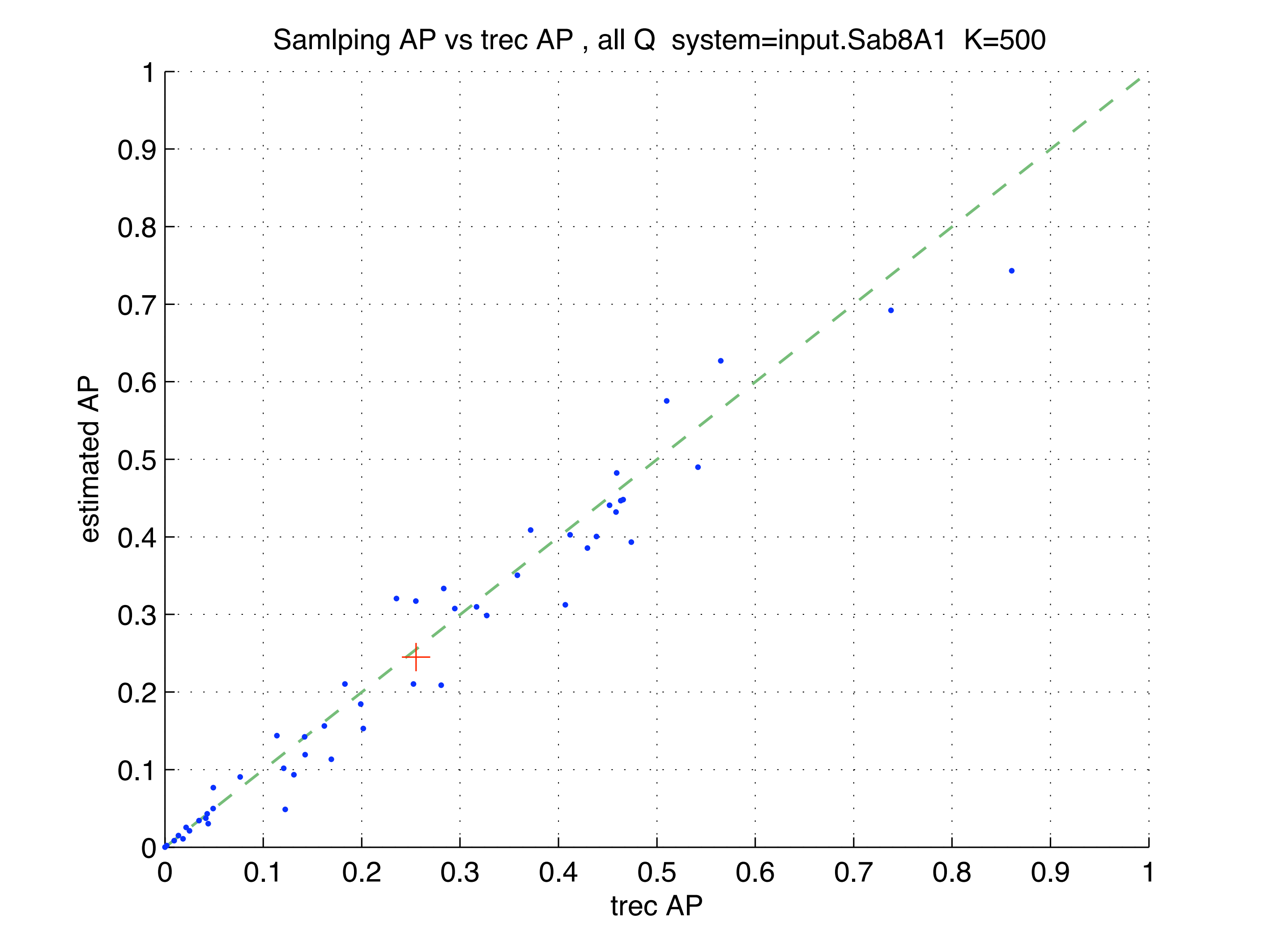
Query Variance: 40 Judgments



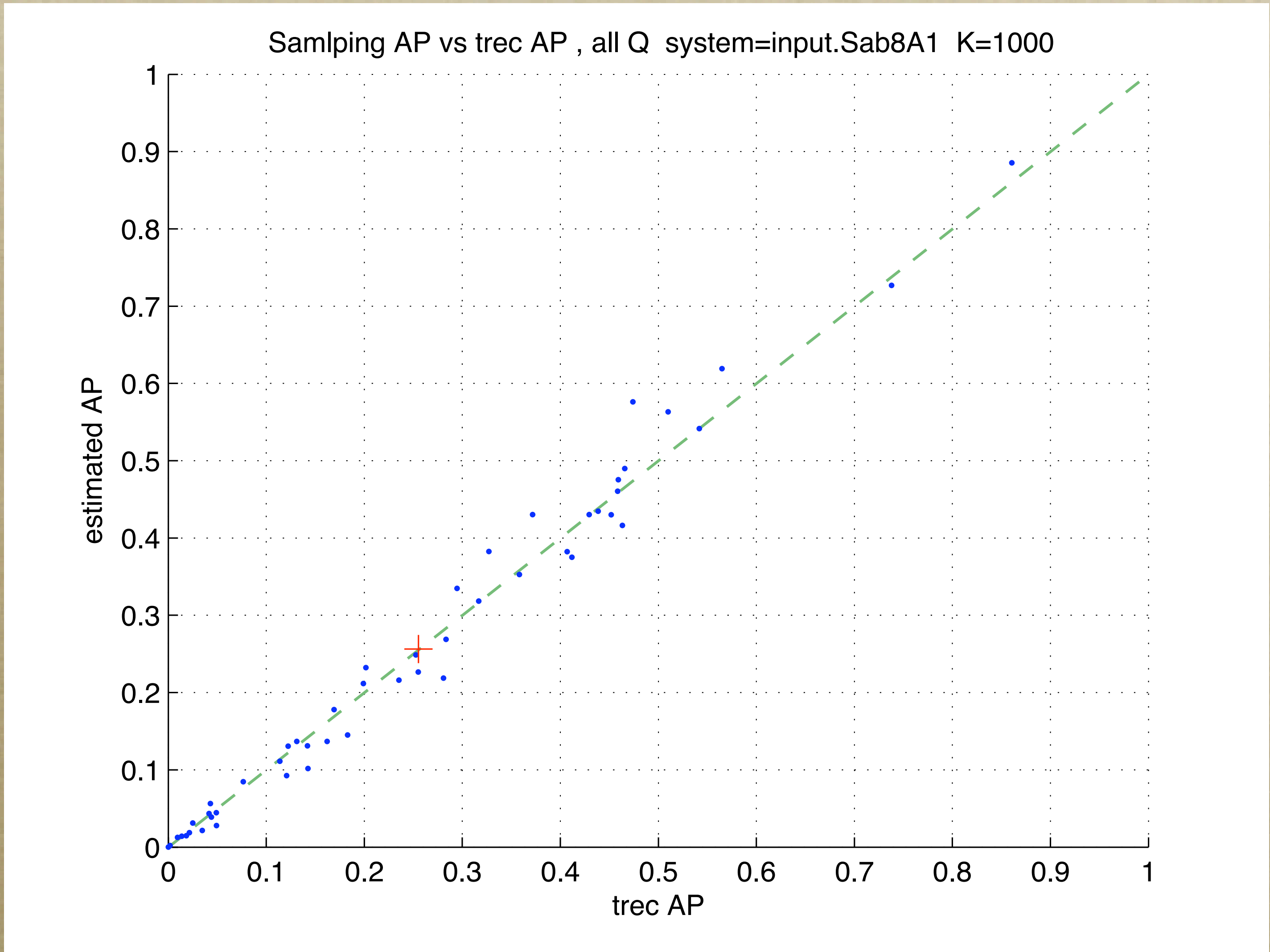
Query Variance: 260 Judgments



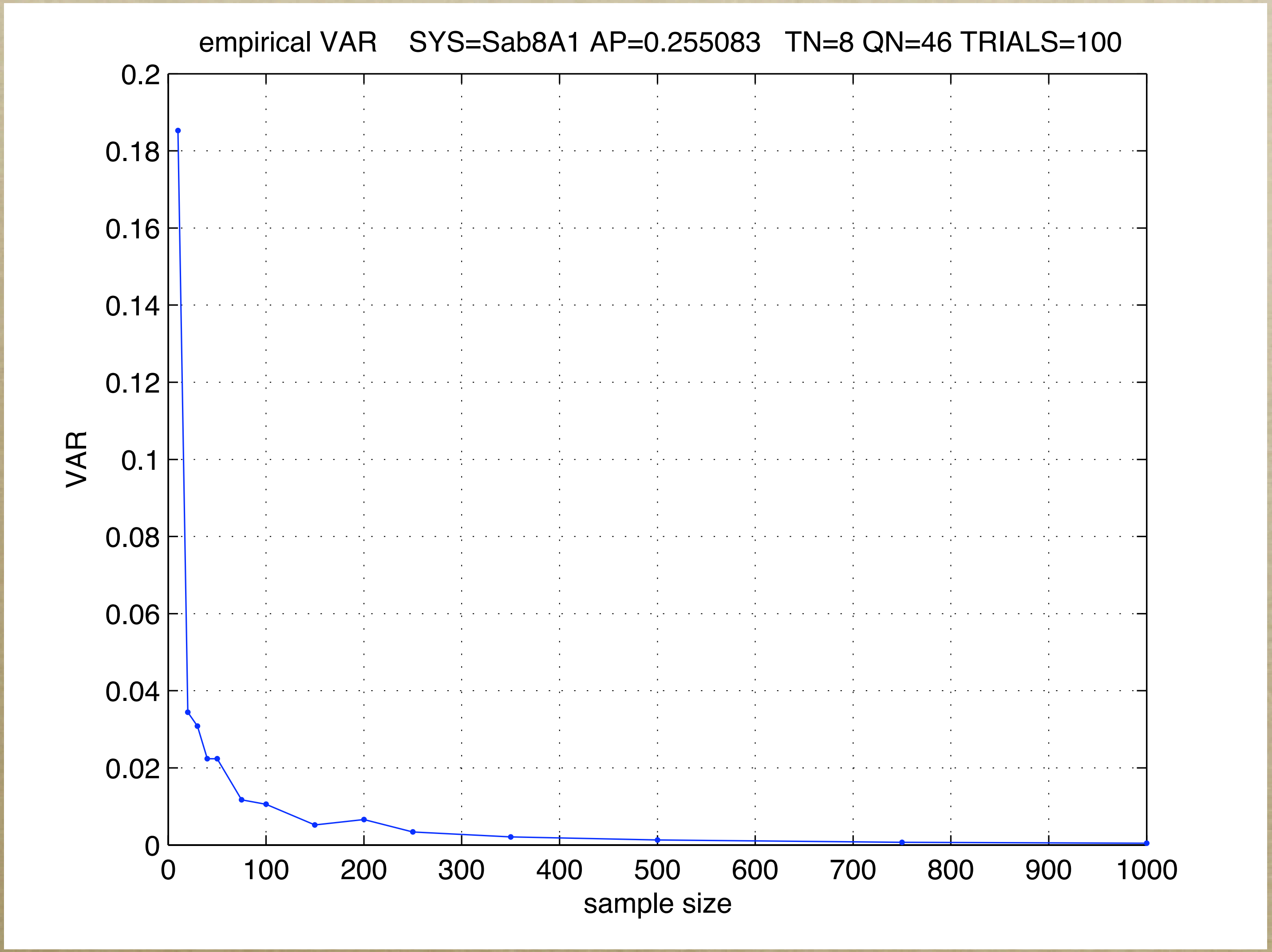
Query Variance: 500 Judgments



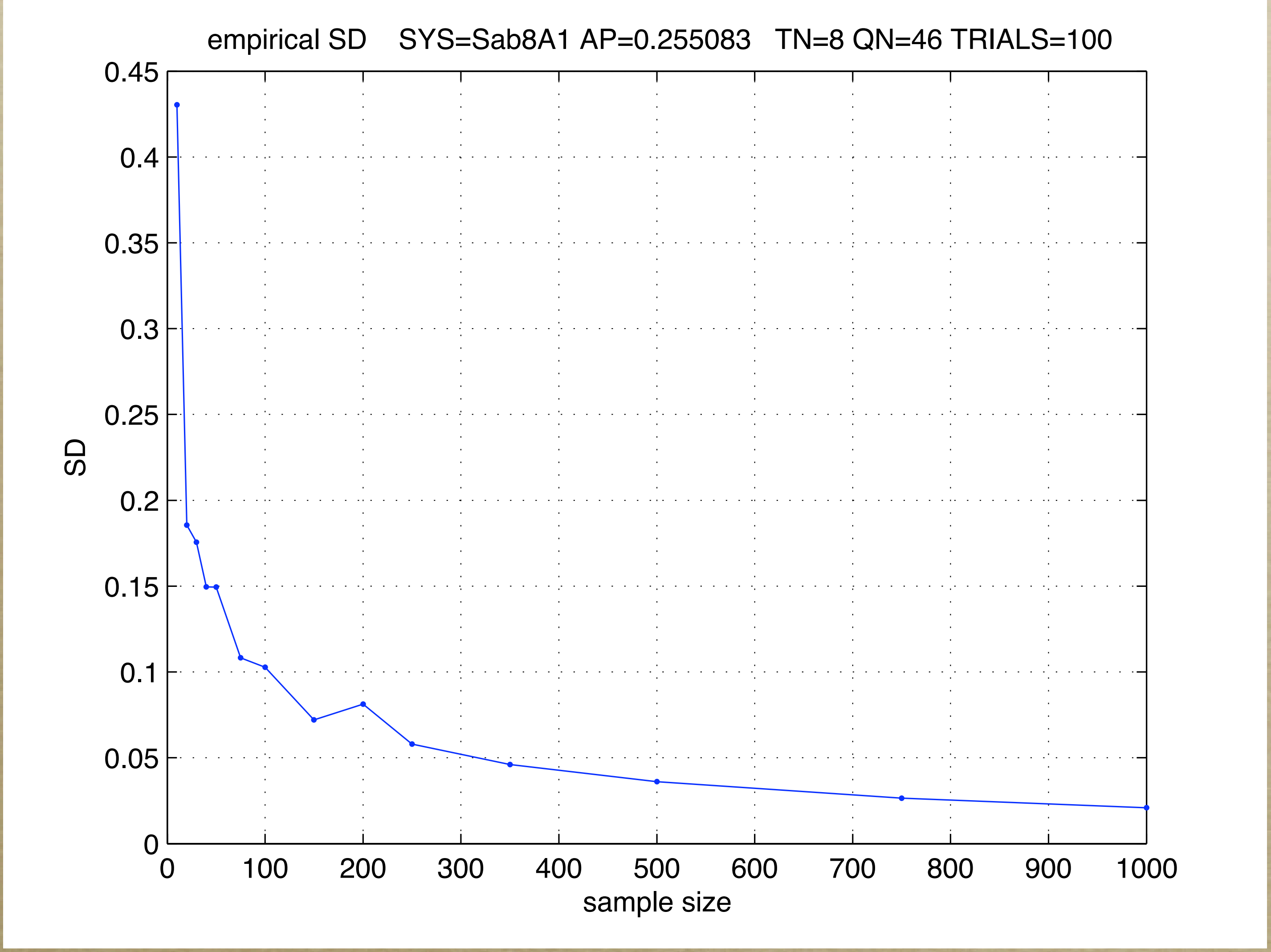
Query Variance: 1000 Judgments



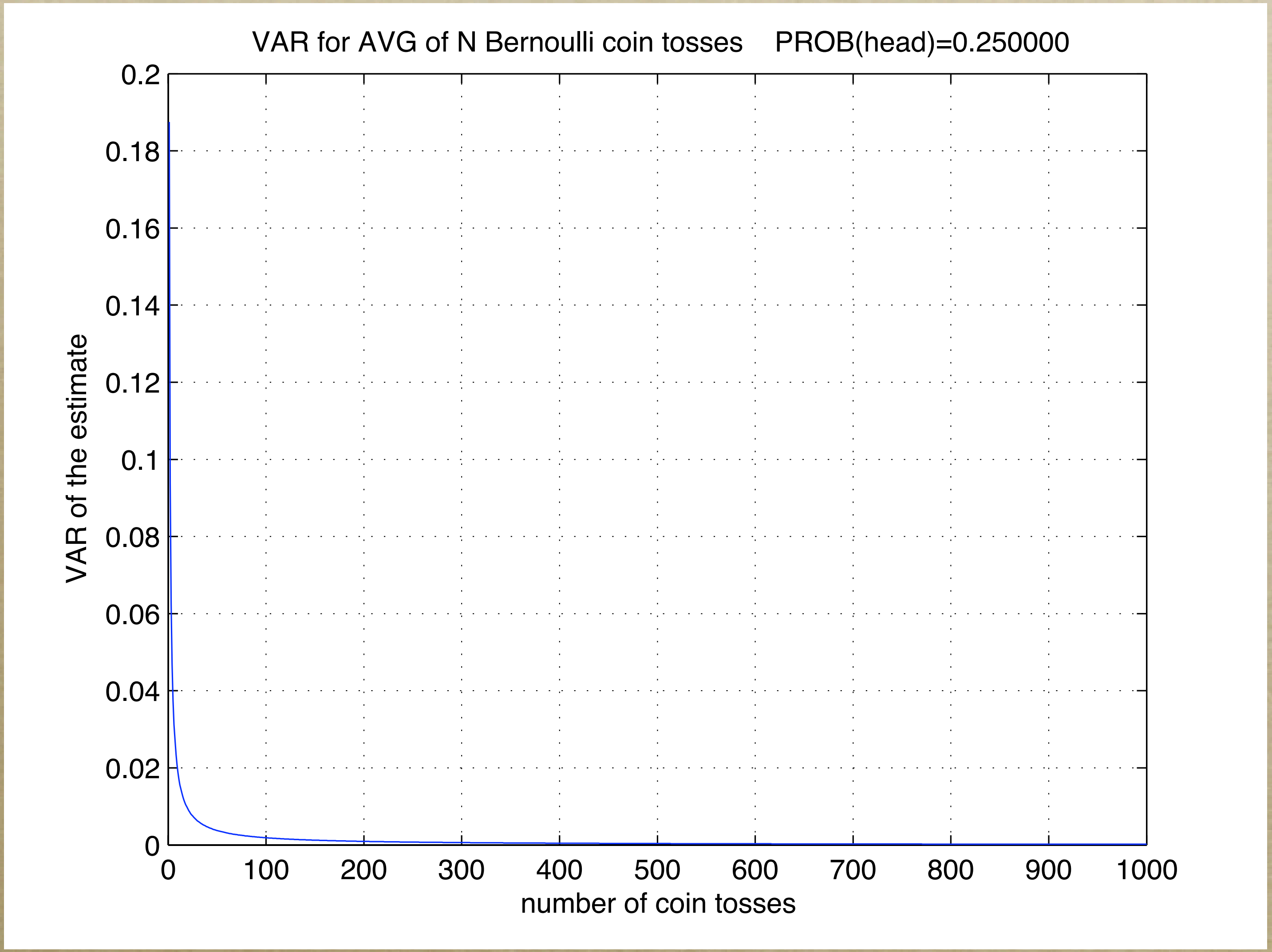
Variance Reduction via Sampling



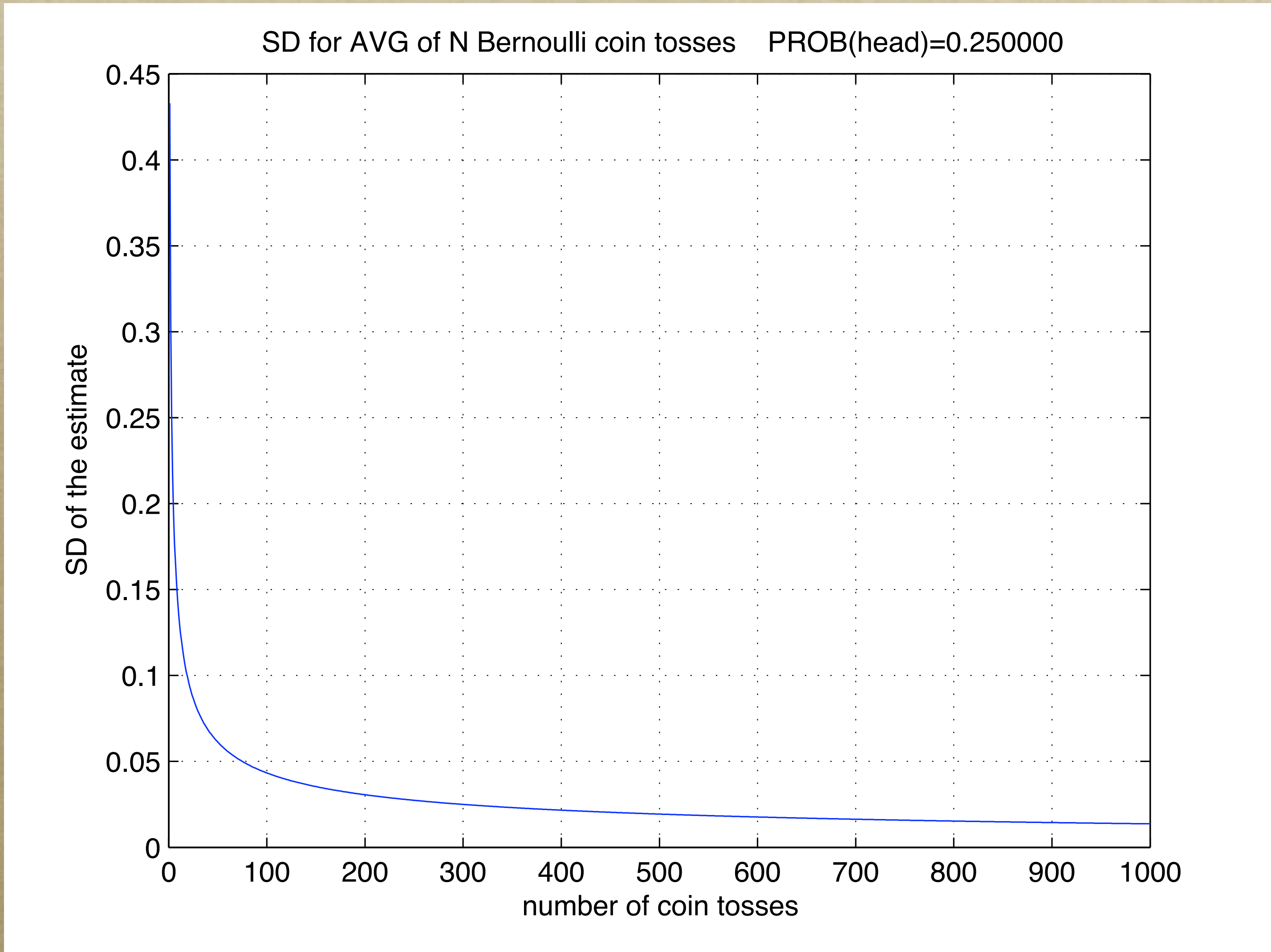
SD Reduction via Sampling



Variance Reduction: Sampling



SD Reduction: Sampling



What About AP?

List:

R	1/1
N	
R	2/3
N	
N	
R	3/6
N	
N	
N	
R	4/10

$$\mathrm{AP} = \frac{1 + 2/3 + 3/6 + 4/10}{4} \approx 0.6417$$

Average Precision

$$
\begin{aligned}
\mathrm{AP} &= \frac{1}{R} \cdot \sum_{i\,:\,rel(i)=1} \mathrm{PC}(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^{Z} rel(i) \cdot \mathrm{PC}(i) \\
&= \frac{1}{R} \cdot \sum_{i=1}^{Z} rel(i) \sum_{j=1}^{i} rel(j)/i \\
&= \frac{1}{R} \cdot \sum_{1 \le j \le i \le Z} \frac{1}{i} \cdot rel(i) \cdot rel(j)
\end{aligned}
$$

We will sample for the summation (and R)...

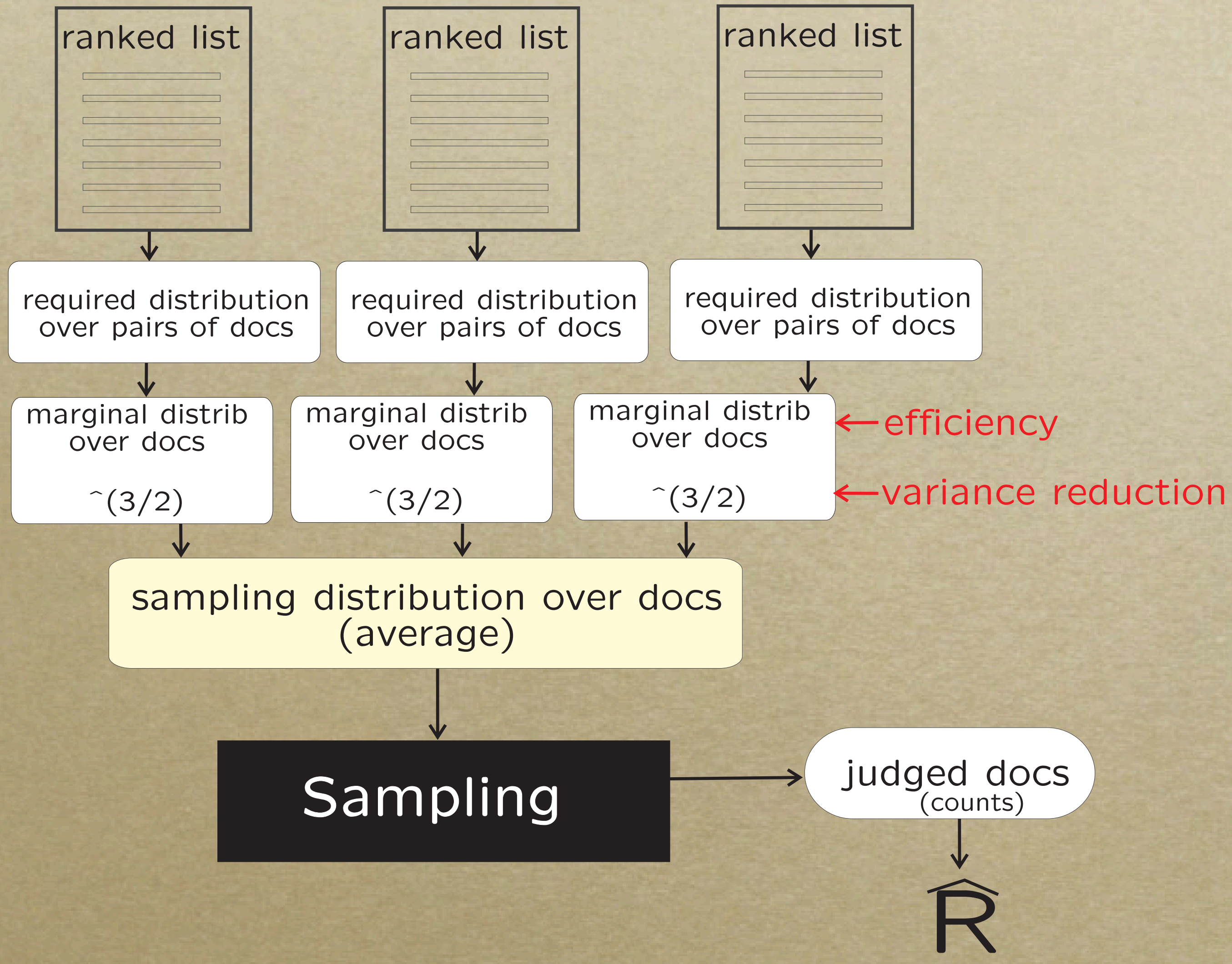
AP Sampling Distribution

	1	2	3	$\ldots$	Z
1	1				
2	$1/2$	$1/2$			
3	$1/3$	$1/3$	$1/3$		
$\vdots$					
Z	$1/Z$	$1/Z$	$1/Z$	$\ldots$	$1/Z$

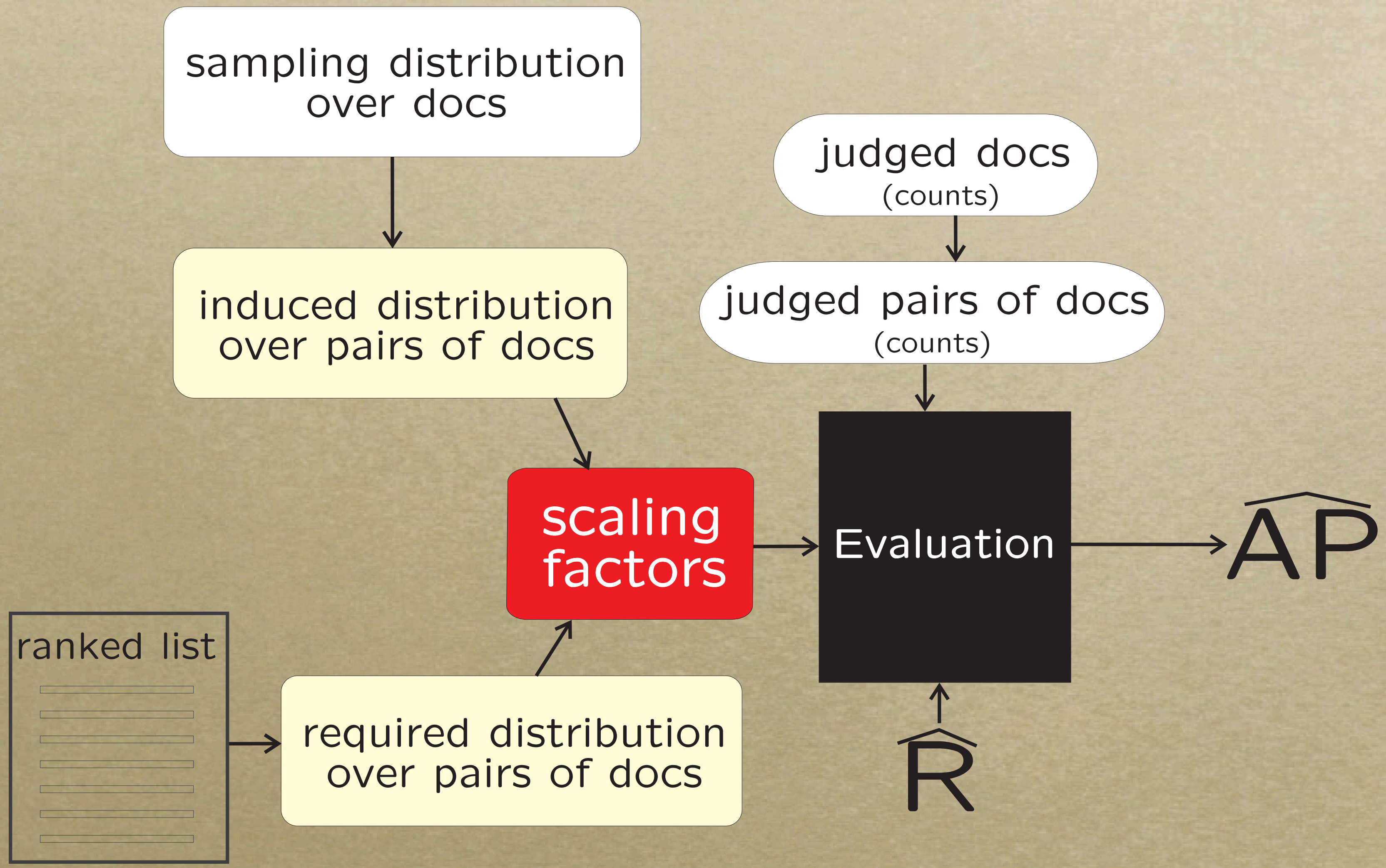
	1	2	3	$\ldots$	Z
1	2	$1/2$	$1/3$	$\ldots$	$1/Z$
2	$1/2$	1	$1/3$	$\ldots$	$1/Z$
3	$1/3$	$1/3$	$2/3$	$\ldots$	$1/Z$
$\vdots$					
Z	$1/Z$	$1/Z$	$1/Z$	$\ldots$	$2/Z$

Normalized, appropriately...

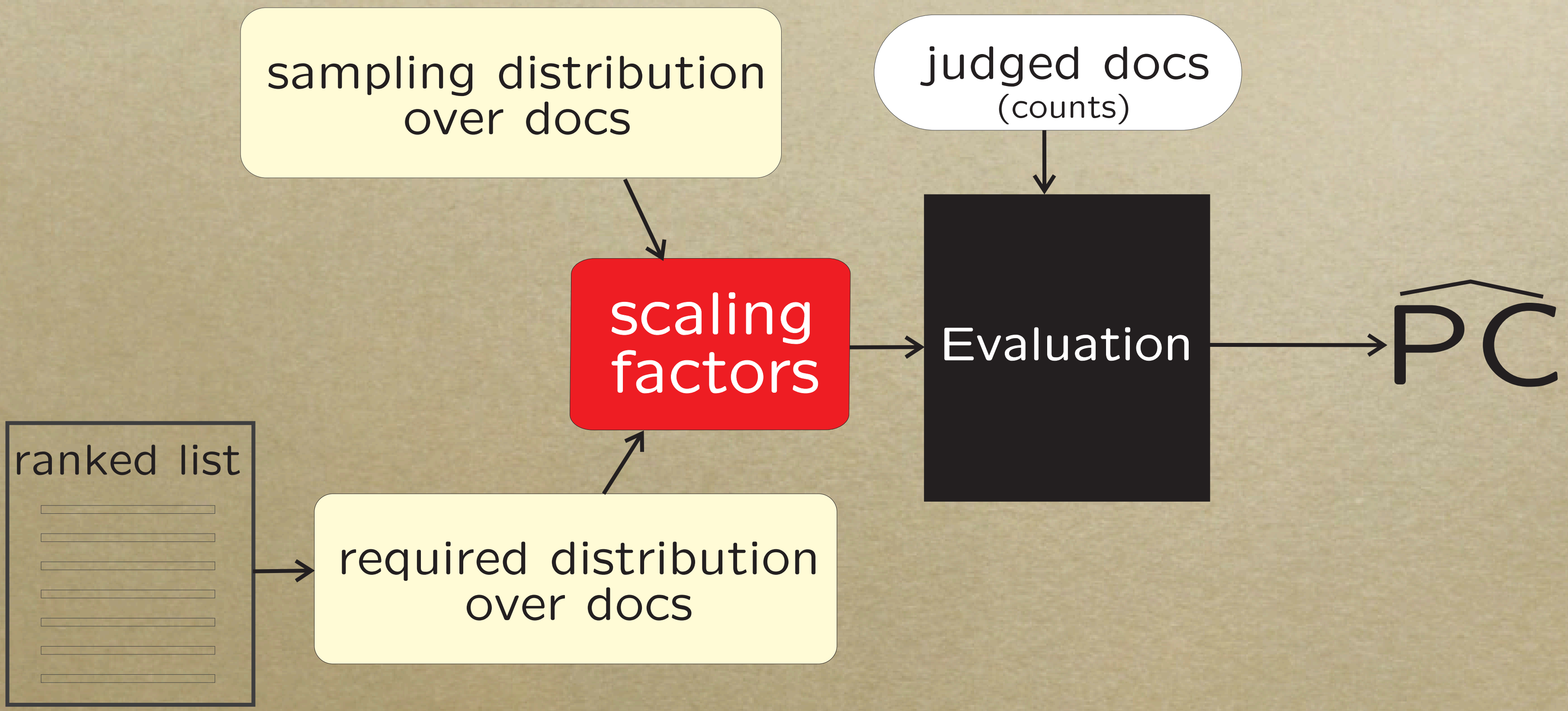
Sampling Strategy



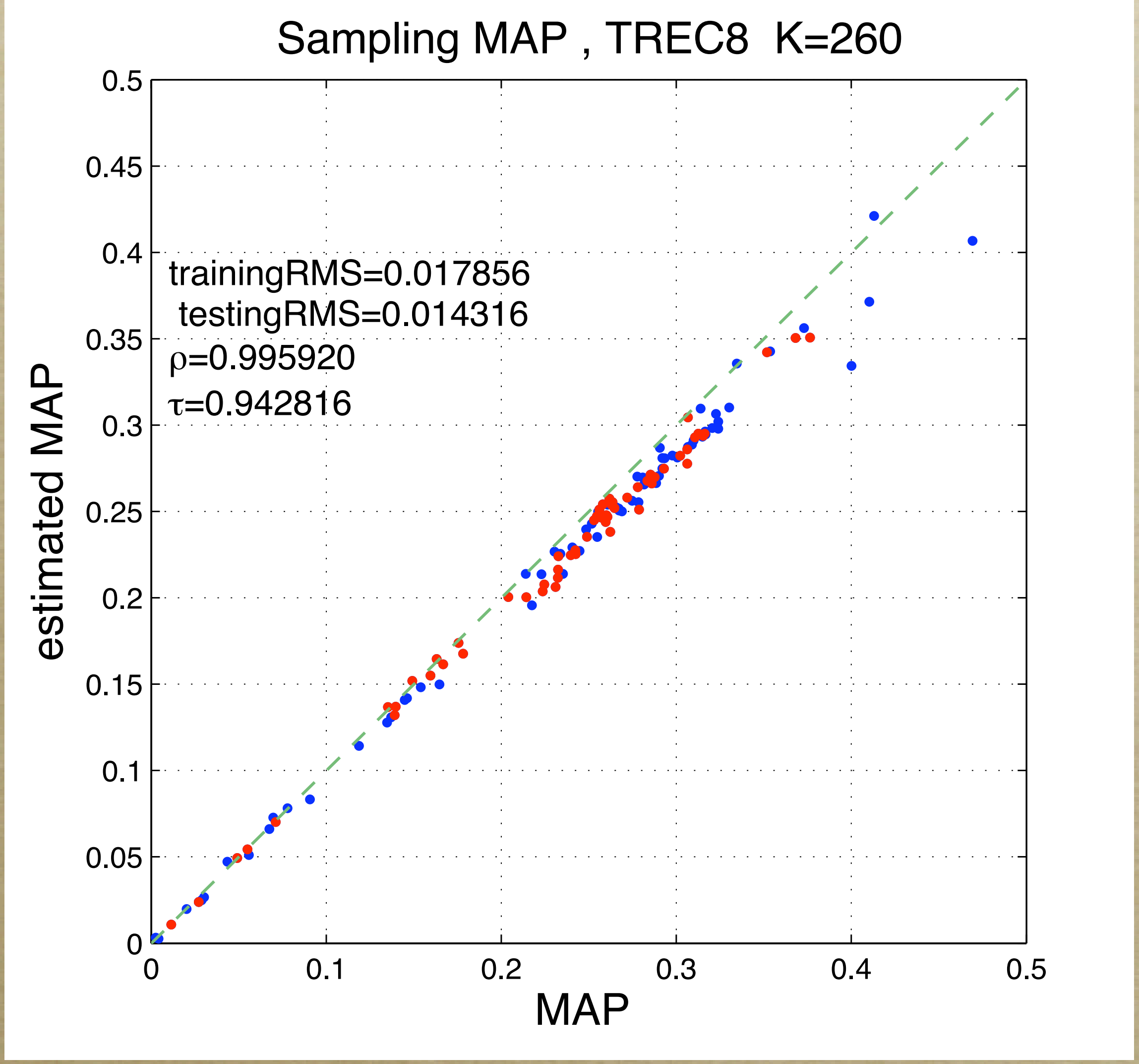
Estimating AP



Estimating PC (and RP)



MAP Estimates: 260 JPQ



MPC(30) Estimates: 260 JPQ