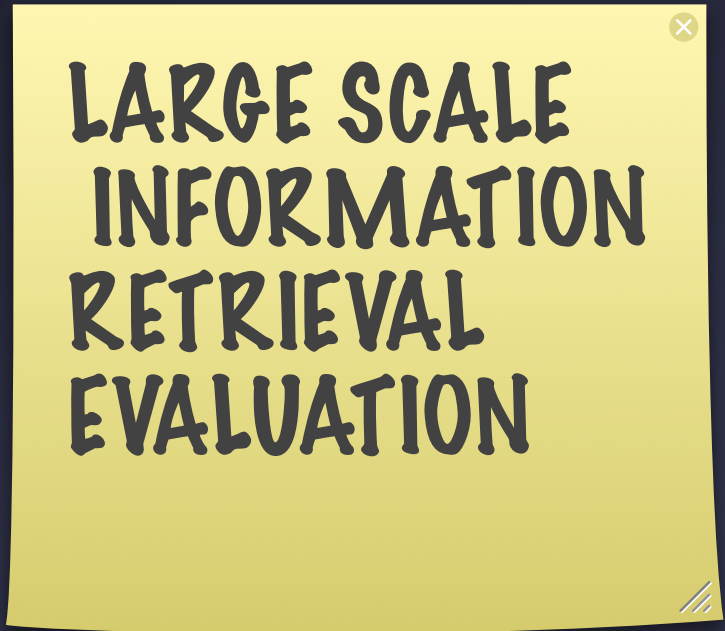


# Large Scale IR Evaluation

---

Virgil Pavlu



**LARGE SCALE  
INFORMATION  
RETRIEVAL  
EVALUATION**

# Large Scale IR Evaluation

---

# Large Scale Evaluation

---

# Large Scale IR Evaluation

---

- People are not well organized
  - nor consistent with each other

# Large Scale IR Evaluation

---

- People are not well organized
  - nor consistent with each other
- People are publishing everything

# Large Scale **IR** Evaluation

---

- People are not well organized
  - nor consistent with each other
- People are publishing everything
- Today, Culture, Commerce, Science and Military depend on ability to store and make use of information

# Large Scale IR Evaluation

---

- People are not well organized
  - nor consistent with each other
- People are publishing everything
- Today, Culture, Commerce, Science and Military depend on ability to store and make use of information
- IR field is dedicated to organization of information
  - search is the principal component

# Large Scale **IR** Evaluation

## ● IR vs Databases

	Databases	IR
Data	<b>Structured</b>	<b>Unstructured</b>
Fields	<b>Clear semantics</b> (SSN, age)	<b>No fields</b> (other than text)
Queries	<b>Defined</b> (relational algebra, SQL)	<b>Free text</b> ("natural language"), Boolean
Recoverability	<b>Critical</b> (concurrency control, recovery, atomic operations)	<b>Downplayed</b> , though still an issue
Matching	<b>Exact</b> (results are <i>always</i> "correct")	<b>Imprecise</b> (need to measure effectiveness)

# Large Scale IR Evaluation

---

# Large Scale IR Evaluation

---

- Critical for research
- Commercially used for optimization
- Can measure many aspects of returned results
- Text REtrieval Conference
  - queries, collections, search engines, performance
  - many tracks every year
  - judges several 100K documents

# Large Scale IR Evaluation

---

INTRODUCE

2 METHODS  
FOR EVALUATION

THAT DEALS WITH LARGE

# Large Scale IR Evaluation

- Current size of datasets make traditional methods less applicable
  - Terabyte GOV2 collection has 25M documents
  - On the web, focus is on top of the list

INTRODUCE

2 METHODS  
FOR EVALUATION

THAT DEALS WITH LARGE

# Large Scale IR Evaluation

- Current size of datasets make traditional methods less applicable
  - Terabyte GOV2 collection has 25M documents
  - On the web, focus is on top of the list
- Assuming a list of results is summary-able, we should be able to interpret measurements as statistics
  - what is the point of retrieving 30000 documents

INTRODUCE

2 METHODS  
FOR EVALUATION

THAT DEALS WITH LARGE

# overview

---

## ◆ Introduction

- Relevance Prior
- Hedge
- Sampling
- Future work



# Precision and Recall

---

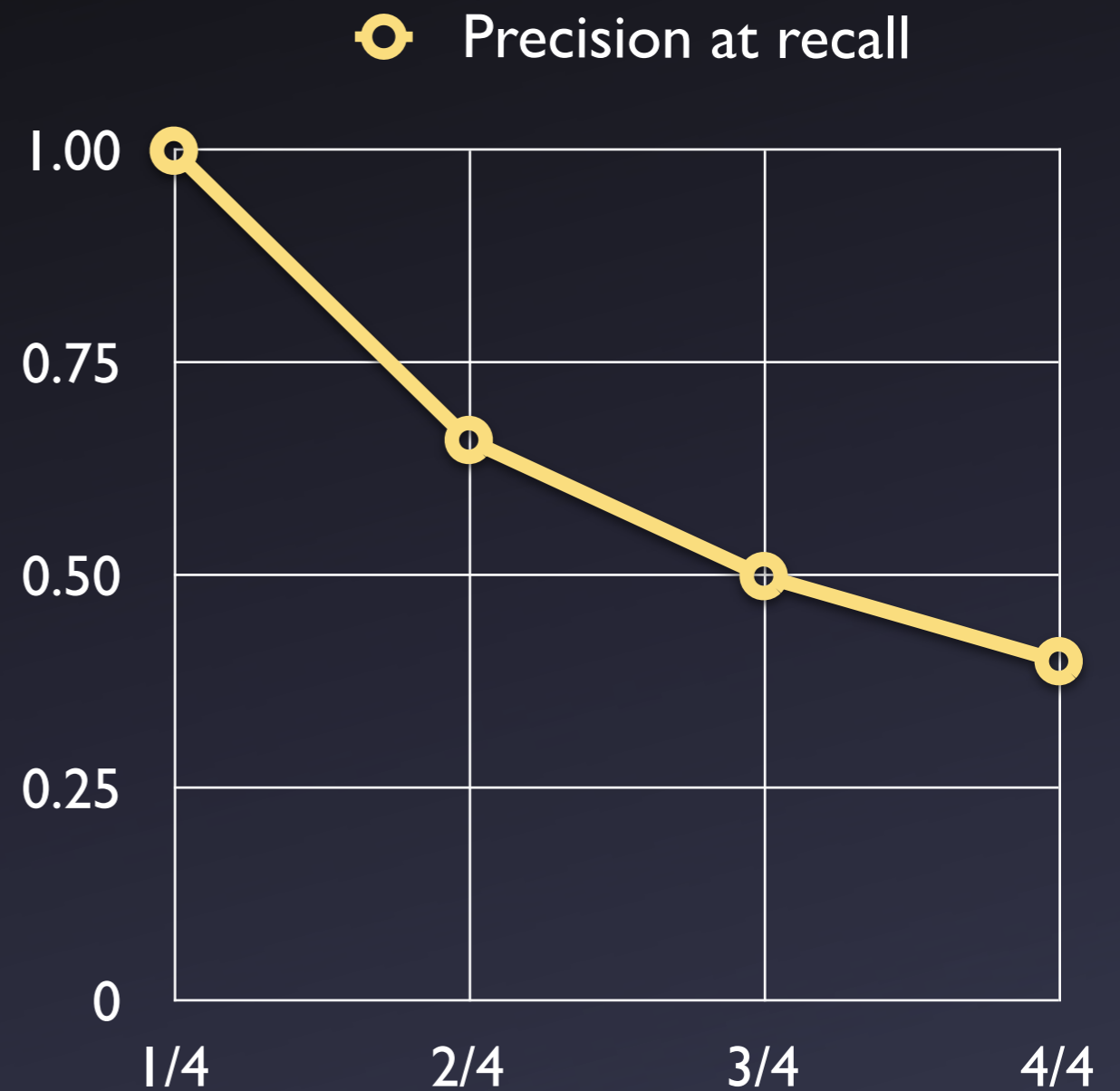
List:

	precision	recall
R	1/1	1/4
N	1/2	1/4
R	2/3	2/4
N	2/4	2/4
N	2/5	2/4
R	3/6	3/4
N	3/7	3/4
N	3/8	3/4
N	3/9	3/4
R	4/10	4/4

# Precision and Recall

List:

	precision	recall
R	1/1	1/4
N	1/2	1/4
R	2/3	2/4
N	2/4	2/4
N	2/5	2/4
R	3/6	3/4
N	3/7	3/4
N	3/8	3/4
N	3/9	3/4
R	4/10	4/4



# Average Precision

List:

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

- AP = average of precisions at relevant ranks

- use 0 for relevant documents not returned

change in AP for 1 rank

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

# relevance prior

---

- Assess the relative importance of ranks
  - should be monotonic decreasing
  - can we use search engines scores ?
- Many developed
  - zipfian, logarithmic, logistic regression etc

# AP relevance prior

- Sum-Precision is the numerator of AP
  - denominator is query-constant

$$\begin{aligned} SP &= \sum_{i: rel(i)=1} Prec(i) = \sum_{i=1}^Z rel(i) \cdot Prec(i) \\ &= \sum_{i=1}^Z rel(i) \sum_{j=1}^i rel(j) / i = \sum_{1 \leq j \leq i \leq Z} \frac{1}{i} \cdot rel(i) \cdot rel(j) \end{aligned}$$

rel = relevant (0,1)

# AP relevance prior

- Sum-Precision is the numerator of AP

— denominator is query-constant

$$SP = \sum_{i: rel(i)=1} Prec(i) = \sum_{i=1}^Z rel(i) \cdot Prec(i)$$

$$= \sum_{i=1}^Z rel(i) \sum_{j=1}^i rel(j)/i = \sum_{1 \leq j \leq i \leq Z} \frac{1}{i} \cdot rel(i) \cdot rel(j)$$

rel = relevant (0,1)

- Infer a weighting scheme for ranks

	1	2	3	...	Z
1	1				
2	$\frac{1}{2}$	$\frac{1}{2}$			
3	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$		
⋮					
Z	$\frac{1}{Z}$	$\frac{1}{Z}$	$\frac{1}{Z}$	...	$\frac{1}{Z}$

	1	2	3	...	Z
1	2	$\frac{1}{2}$	$\frac{1}{3}$	...	$\frac{1}{Z}$
2	$\frac{1}{2}$	1	$\frac{1}{3}$	...	$\frac{1}{Z}$
3	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	...	$\frac{1}{Z}$
⋮					
Z	$\frac{1}{Z}$	$\frac{1}{Z}$	$\frac{1}{Z}$	...	$\frac{2}{Z}$

# AP relevance prior

---

- After marginalization

$$W(r) = \frac{1}{2Z} \left( 1 + \frac{1}{r} + \frac{1}{r+1} + \cdots + \frac{1}{Z} \right)$$

– future work : raised at 3/2 for variance reduction

- If we have many systems we can derive an average prior over documents

$$M(i) = \frac{1}{N} \sum_s W_s(\text{rank}(i, s))$$

**CIKM 2003 Homepage**

... 03, Hotel Inter-Continental, Novem  
Conference on Information and Know  
Description: ACM Conference on Info  
Category: Reference > Knowledge M  
bit.csc.lsu.edu/~cikm2003/ - 9k - Ca

**ACM CIKM 2003 Call For**

... Management (CIKM03). A  
highest caliber papers submit  
bit.csc.lsu.edu/~cikm2003/CI  
[ More results from bit.csc.lsu

**CIKM-2003 Registration**

... All Amounts in USD. Attendee Sta  
one author of an accepted paper mus  
www.cikmreg.org/ - 13k - Cached - S

**ACM CIKM 2003 Call For Pap**

ACM CIKM 2003 -Call for Industry T  
Sponsored by ACM SIGIR and ACM  
www.cs.wisc.edu/dbworld/messages

**CIKM 2002 Homepage**

... 03), Las Vegas, Nevada, June 23-  
Conference on Information and Know  
Description: SAIC Headquarters, Mc  
Category: Business > Management  
www.cikm.org/2002/ - 16k - Cached -

**Conference on Informatio**

... The CIKM 2003 web page i  
devoted to emerging areas of  
www.cikm.org/ - 6k - Cached -

**[Asis-I] CIKM 2003**

[Asis-I] CIKM 2003. Padmini Srinivas  
Mon, 29 Sep 2003 12:59:36 -0500: P  
mail.asis.org/pipermail/asis-I/ 2003-S

**(DBWORLD) CIKM 2003 (Pad**

(DBWORLD) CIKM 2003 (Padmini S  
CIKM 2003; From: "Padmini Srinivas  
wwwiti.cs.uni-magdeburg.de/MailArc

**bridge-cikm-2003**

bit.csc.lsu.edu/~cikm2003.  
www.logistics-2000-versailles.net/site

**Collaborative Filtering Mailing**

[collab@sims] CFP: CIKM 2003. ...

# relevance prior

$$W(r) = \frac{1}{2Z} \left( 1 + \frac{1}{r} + \frac{1}{r+1} + \dots + \frac{1}{Z} \right)$$

# overview

---

- Introduction
- Relevance Prior
- ◆ Hedge
- Sampling
- Future work

# Hedge online learning

---

# Hedge online learning

---

- Say I want to invest in stocks
  - quick return : sell after a week or less
  - repeat over a long period of time
  - invest a fixed amount every day

# Hedge online learning

---

- Say I want to invest in stocks
  - quick return : sell after a week or less
  - repeat over a long period of time
  - invest a fixed amount every day
- I have 4 friends brokers
  - Alice, Bob, Carlos and Daniel
  - each recommends a ranked list
  - I decide to ask all 4 for advice every time

# combining experts

---

- Day 1 : which stock should I buy ?

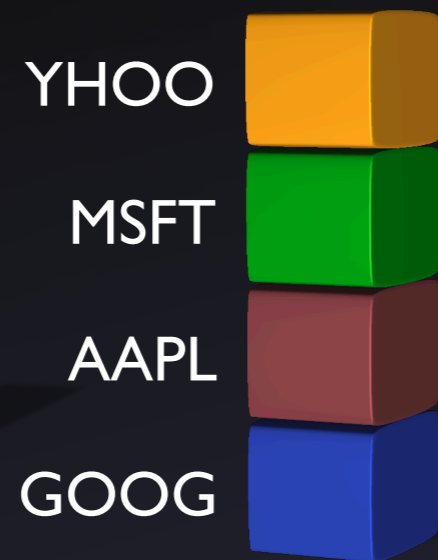
# combining experts

---

- Day 1 : which stock should I buy ?



Alice



Bob



Carlos



Daniel

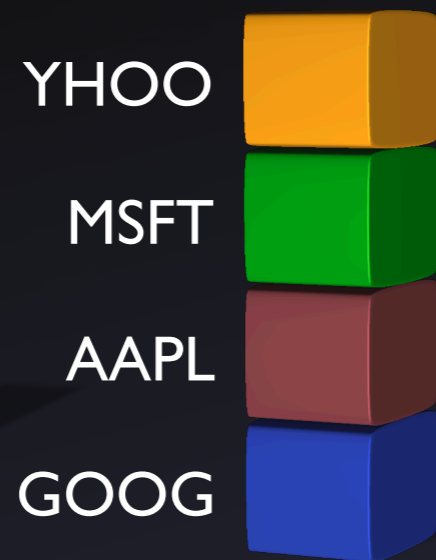
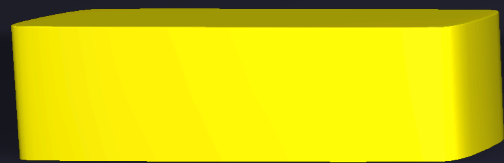
# combining experts

---

- Day 1 : which stock should I buy ?



Alice



Bob



Carlos



Daniel



- Maybe I will buy MSFT

# episodic loss

---

- Next day : MSFT announces buying YHOO
  - on expectation that they will ruin YMail, Ymessenger, Flickr etc MSFT stock falls
- I will reflect the loss into the confidence I have in each broker, based on the recommended list

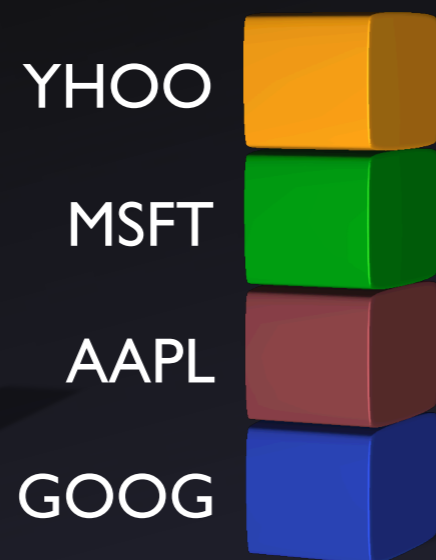
# episodic loss

---

- day 1 recommendations:



Alice



Bob



Carlos



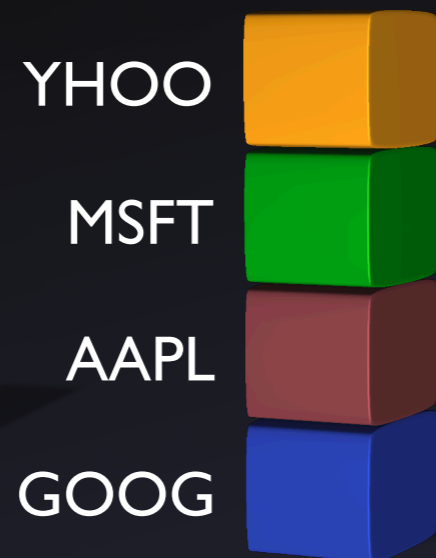
Daniel

# episodic loss

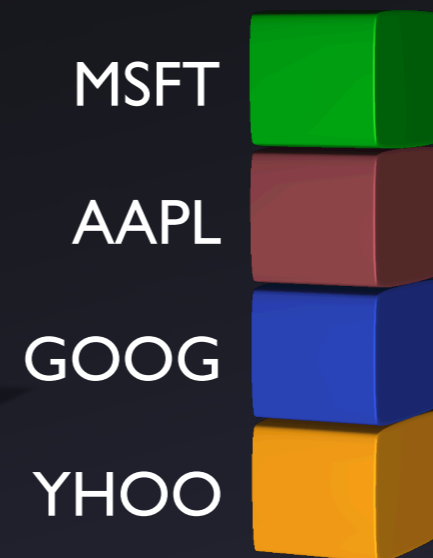
- day 1 recommendations:



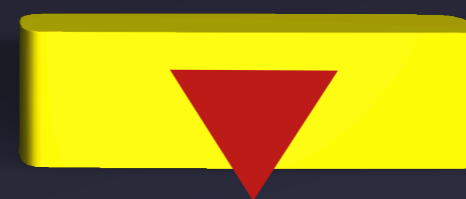
Alice



Bob



Carlos



Daniel



# update confidence

---

● after day 1:



Alice



Bob



Carlos



Daniel

# update confidence

---

● after day 1:

Alice



Bob



Carlos



Daniel



# combining experts

---

- Day 2 : which stock should I buy ?

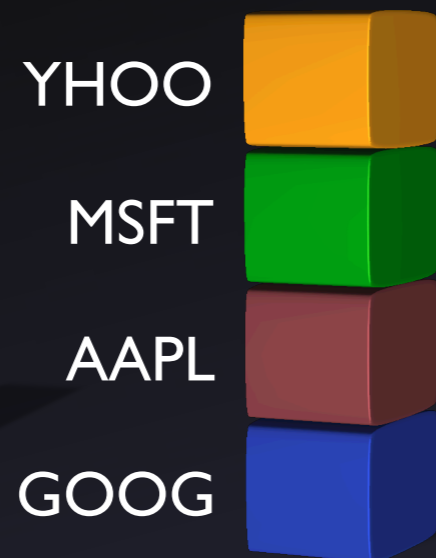
# combining experts

---

● Day 2 : which stock should I buy ?



Alice



Bob



Carlos



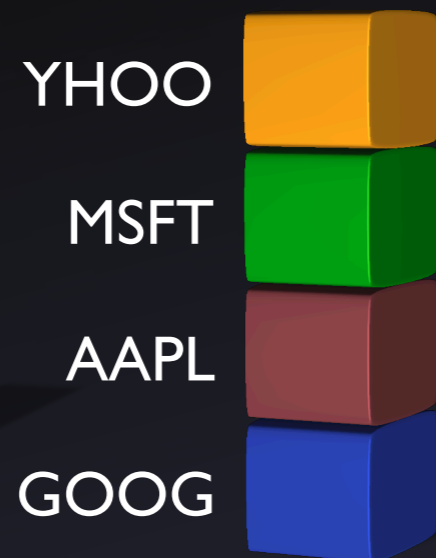
Daniel

# combining experts

● Day 2 : which stock should I buy ?



Alice



Bob



Carlos



Daniel



● Maybe I will buy AAPL

# cumulative performance

---

- Next : AAPL most-expected iPhone gets delayed
  - I lose money, update confidence etc

# cumulative performance

---

- Next : AAPL most-expected iPhone gets delayed
  - I lose money, update confidence etc
- But if at least one broker is any good (and consistent) I will make money overall

# cumulative performance

---

- Next : AAPL most-expected iPhone gets delayed
  - I lose money, update confidence etc
- But if at least one broker is any good (and consistent) I will make money overall
- At the end of the year I compare my cumulative loss/gain with the overall best-broker

$$\text{LOSS}_{Hedge} \leq \frac{\min_b \{L_b\} \cdot \ln(1/\beta) + \ln(N)}{1-\beta}$$

- future work on  $\beta$

# adaptation to IR setup

---

## HEDGE

- given : experts
- incoming : loses
- reweight: experts

## IR

- given : search engines
- incoming: documents
  - compute losses
- reweight : search engines

# metasearch

---

- Given several lists returned on the same query, combine them in a single list
- Goal: perform as good as possible
  - usually compares with best system
- Rank-based or Score-based
- Several known techniques
  - Comb MNZ
  - Borda
  - Condorcet

**Google™** Advanced Search Preferences Language

cikm 2003

Web Images Groups Directories

Searched the web for **cikm 2003**

**CIKM 2003 Homepage**

... 03, Hotel Inter-Continental, Novem  
Conference on Information and Know  
Description: ACM Conference on Info  
Category: [Reference > Knowledge M](#)  
[bit.csc.lsu.edu/~cikm2003/](http://bit.csc.lsu.edu/~cikm2003/) - 9k - [Ca](#)

**ACM CIKM 2003 Call For**

... Management (CIKM03). AC  
highest caliber papers submitte  
[bit.csc.lsu.edu/~cikm2003/CI](http://bit.csc.lsu.edu/~cikm2003/CI)  
[ [More results from bit.csc.lsu](#)

**CIKM-2003 Registration**

... All Amounts in USD. Attendee Sta  
one author of an accepted paper mus  
[www.cikmreg.org/](http://www.cikmreg.org/) - 13k - [Cached](#) - [S](#)

**ACM CIKM 2003 Call For Pap**

ACM **CIKM 2003** -Call for Industry Te  
Sponsored by ACM SIGIR and ACM  
[www.cs.wisc.edu/dbworld/messages](http://www.cs.wisc.edu/dbworld/messages)

**CIKM 2002 Homepage**

... 03), Las Vegas, Nevada, June 23-  
Conference on Information and Know  
Description: SAIC Headquarters, Mc  
Category: [Business > Management S](#)  
[www.cikm.org/2002/](http://www.cikm.org/2002/) - 16k - [Cached](#) -

**Conference on Information**

... The **CIKM 2003** web page i  
devoted to emerging areas of c  
[www.cikm.org/](http://www.cikm.org/) - 6k - [Cached](#) -

**[Asis-] CIKM 2003**

[Asis-] **CIKM 2003**. Padmini Srinivas  
Mon, 29 Sep 2003 12:59:36 -0500: P  
[mail.asis.org/pipermail/asis-l/2003-S](mailto:mail.asis.org/pipermail/asis-l/2003-S)

**(DBWORLD) CIKM 2003 (Pag**

(DBWORLD) **CIKM 2003** (Padmini S  
**CIKM 2003**; From: "Padmini Srinivas  
[www.witi.cs.uni-magdeburg.de/MailArc](http://www.witi.cs.uni-magdeburg.de/MailArc)

**bridge-cikm-2003**

[bit.csc.lsu.edu/~cikm2003](http://bit.csc.lsu.edu/~cikm2003).  
[www.logistics-2000-versailles.net/site](http://www.logistics-2000-versailles.net/site)

**Collaborative Filtering Mailing**

[collab@simsl] CFP: **CIKM 2003**. ...

**altavista™** Web Images

cikm 2003

SEARCH: ☐ Worldwide ☐ U.S.

Did you mean: *ci km 2003*

**AltaVista found 5,186 results** [About](#)

**CIKM 2003 Homepage**

12th international conference on info  
[bit.csc.lsu.edu/~cikm2003](http://bit.csc.lsu.edu/~cikm2003) • Refresh  
[More pages from bit.csc.lsu.edu](#)

**ACM CIKM 2003 Call For**

... Management (CIKM03) AC  
caliber papers submitted to C  
[bit.csc.lsu.edu/~cikm2003/CI](http://bit.csc.lsu.edu/~cikm2003/CI)  
[More pages from bit.csc.lsu.e](#)

**CIKM 2002 Homepage**

... 2002 Advance Technical Program  
Treasurer, Dr. Nicholas, the **CIKM02**  
[www.cikm.org/2002](http://www.cikm.org/2002) • [Related Pages](#)

**Conference on Information**

**CIKM CIKM 2003 CIKM Topi**  
Knowledge Management The  
[www.cikm.org](http://www.cikm.org) • [Related Page](#)  
[More pages from www.cikm.o](#)

**CIKM**

... Information and Knowledge Mana  
**CIKM 2003 Home Page 11. CIKM 2**  
[www.informatik.uni-trier.de/~ley/db/c](http://www.informatik.uni-trier.de/~ley/db/c)  
[More pages from www.informatik.uni](#)

**7. CIKM 1998: Bethesda,**

7. **CIKM 1998: Bethesda, Ma**  
Copyright © by Michael Ley  
[www.informatik.uni-trier.de/~l](http://www.informatik.uni-trier.de/~l)  
[More pages from www.inform](#)

**SIGIR Information Server**

... award. Upcoming SIGIR Spons  
**2003. JCDL 2004 (Tucson, AZ) - Jun**  
[www.acm.org/sigir](http://www.acm.org/sigir) • Refreshed in p  
[More pages from www.acm.org](#)

**Web Caching Publications/Venu**

... 2000 Books 2001 Books 2002 Books **2003** net.  
2004 ... 2003 ... 2002 ... 2001 ...

**Web Caching Publications/Venu**

... 2000 Books 2001 Books 2002 Books **2003** net.  
2004 ... 2003 ... 2002 ... 2001 ...

**alltheweb™** find it all

advanced search

cikm 2003

Results in: C

**Web** News Pictures Video Audio

**1 - 10 of 6,293 Results for cikm 2003**

**CIKM 2003 Homepage**

... Information and Knowledge Management **CIKM**  
an international ...  
**Description:** The ACM Conference on Information  
and knowledge management, as well as recent ad  
[more hits from:](#) <http://bit.csc.lsu.edu/~cikm2003/>

**IVML Call For Papers Archive: ACM CI**

Fo ACM **CIKM 2003** Preliminary Call For Papers  
message ... Digital Games Conference **2003"** Prev  
**Description:** Previous message: Kostas Karpouzis  
Conferences - SPIE Web  
<http://www.image.ntua.gr/cfp/archive/2003/0234.htm>

**CIKM 2002 Homepage**

... Information and Knowledge Management (CIKM  
been ... Information and Knowledge Management  
**Description:** SAIC Headquarters, McLean, Virginia  
[more hits from:](#) <http://www.cikm.org/2002/> - 15 k

**ACM WIDM'2003**

... **CIKM 2003** ... Information and Knowledge Man  
of the workshop ...  
**Description:** WIDM03 is the fifth in a series of wor  
Information and Knowledge Management (CIKM),  
[more hits from:](#) <http://www.cais.ntu.edu.sg/widm2>

**MMDB'03**

... Programs ACM - MMDB **2003** The First ACM In  
Knowledge Management (ACM **CIKM 2003**). One  
**Description:** The First ACM International Workshop  
university researchers, scientists, industry profess  
[more hits from:](#) <http://www.cs.fiu.edu/mmdb03/> -

**D-Lib Workshops and Conferences: 20**

... proceedings, and other documents. 2004 **2003**  
November **2003**, Sicily, Italy ACM **CIKM 2003**, 12  
**Description:** Below is a listing of digital library mee  
Click on a year to find additional events with links  
[more hits from:](#) <http://www.dlib.org/groups.html> -

**Invited paper to be presented in CIKM (**

Invited paper to be presented in **CIKM** (Conference  
Privacy Policy - Copyright © 2000-2003 NEC and  
**Description:** Abstract: The main objective of a virtu  
resources contributed by the organizations toward  
[more hits from:](#) <http://gunther.smeal.psu.edu/199>

**Internet Conferences 2003**

... 2000 Books 2001 Books 2002 Books **2003** net.  
2004 ... 2003 ... 2002 ... 2001 ...

**Internet Conferences 2003**

... 2000 Books 2001 Books 2002 Books **2003** net.  
2004 ... 2003 ... 2002 ... 2001 ...

# Hedge adaptation to IR setup

The image displays three side-by-side screenshots of search engine results for the query "cikm 2003".

- Google (Left):** Shows search results for "cikm 2003". The first result is "CIKM 2003 Homepage" with a description: "... 03, Hotel Inter-Continental, November 2003, Conference on Information and Knowledge Management (CIKM) ...". Other results include "ACM CIKM 2003 Call For Papers" and "CIKM-2003 Registration".
- Altavista (Middle):** Shows search results for "cikm 2003". The first result is "CIKM 2003 Homepage" with a description: "... 12th international conference on information and knowledge management, as well as recent advances in information and knowledge management ...". Other results include "ACM CIKM 2003 Call For Papers" and "CIKM 2002 Homepage".
- alltheweb (Right):** Shows search results for "cikm 2003". The first result is "CIKM 2003 Homepage" with a description: "... Information and Knowledge Management (CIKM) ...". Other results include "CIKM 2002 Homepage" and "ACM WIDM'2003".

[illegible]

# Hedge adaptation to IR setup



The image displays three side-by-side screenshots of search engine results for the query "cikm 2003".

- Google:** Shows search results for "cikm 2003" with a list of links including "CIKM 2003 Homepage", "ACM CIKM 2003 Call For Papers", "CIKM-2003 Registration", "ACM CIKM 2003 Call For Papers", "CIKM 2002 Homepage", "[Asis-I] CIKM 2003", "(DBWORLD) CIKM 2003 (Padmini Srinivas)", "bridge-cikm-2003", and "Collaborative Filtering Mailing".
- altavista:** Shows search results for "cikm 2003" with a list of links including "CIKM 2003 Homepage", "ACM CIKM 2003 Call For Papers", "CIKM 2002 Homepage", "Conference on Information and Knowledge Management (CIKM)", "CIKM", "SIGIR Information Server", and "Web Caching Publications/Venue".
- alltheweb:** Shows search results for "cikm 2003" with a list of links including "CIKM 2003 Homepage", "ACM CIKM 2003 Call For Papers", "CIKM 2002 Homepage", "ACM WIDM'2003", "MMDB'03", "D-Lib Workshops and Conferences: 2003", "Invited paper to be presented in CIKM", and "Internet Conferences 2003".

## metasearch

### CIKM 2003 Homepage

ACM CIKM 2003 Call For Paper  
CIKM-2003 Registration  
ACM CIKM 2003 Call For Paper  
CIKM 2002 Homepage  
Conference on Information and Knowledge Management (CIKM)  
[Asis-I] CIKM 2003  
(DBWORLD) CIKM 2003 (Padmini Srinivas)  
bridge-cikm-2003  
Collaborative Filtering Mailing List  
CIKM Home Page ACM DL: CIKM  
Yahoo! Groups : webir Message Board  
dbforums - Cfp: Cikm 03  
Mailing List ARL-ERESERVE@arl.mil  
Received: from cni.org by b.cri.org  
ACM WIDM 2003  
ACM - MMDB 2003  
ACM CIKM 2003 PRELIMINARY  
Selected Publications  
Web Caching Publications/Venue  
cikm '03  
Call for papers  
Mario A. Nascimento - Person  
Iceved  
Conferences and Journals on Information and Knowledge Management  
Conferences On Information and Knowledge Management  
Calendrier des manifestations  
(DBWORLD) Final Call for ACM CIKM 2003  
Collaborative Filtering Mailing

The figure displays three side-by-side screenshots of search engine results for the query "cikm 2003".

- Left Screenshot (Google):** Shows the Google search interface with the query "cikm 2003". The first result is "CIKM 2003 Homepage" with a description of the conference and a link to the ACM website. Other results include "ACM CIKM 2003 Call For Papers", "CIKM 2003 Registration", "ACM CIKM 2003 Call For Papers", "CIKM 2002 Homepage", and "CIKM 2003".
- Middle Screenshot (Altavista):** Shows the Altavista search interface with the query "cikm 2003". The first result is "CIKM 2003 Homepage" with a description of the conference and a link to the ACM website. Other results include "ACM CIKM 2003 Call For Papers", "CIKM 2002 Homepage", "CIKM 2003", "CIKM 1998: Bethesda, Maryland", "SIGIR Information Server", and "Web Caching Publications/Venue".
- Right Screenshot (alltheweb):** Shows the alltheweb search interface with the query "cikm 2003". The first result is "CIKM 2003 Homepage" with a description of the conference and a link to the ACM website. Other results include "IVML Call For Papers Archive: ACM CIKM 2003", "CIKM 2002 Homepage", "ACM WIDM'2003", "MMDB'03", "D-Lib Workshops and Conferences: 2003", and "Invited paper to be presented in CIKM (Conference on Information and Knowledge Management)".

## 23

The figure displays three side-by-side screenshots of search engine results for the query "cikm 2003".

- Left Screenshot (Google):** Shows the Google search interface with the query "cikm 2003". The results include:
  - CIKM 2003 Homepage:** ... 03, Hotel Inter-Continental, Novem... Conference on Information and Know... Description: ACM Conference on Info... Category: Reference > Knowledge M... bit.csc.lsu.edu/~cikm2003/ - 9k - Ca...
  - ACM CIKM 2003 Call Fo...** ... Management (CIKM03). AC... highest caliber papers submitte... bit.csc.lsu.edu/~cikm2003/CI... [ More results from bit.csc.lsu...
  - CIKM-2003 Registration** ... All Amounts in USD. Attendee Sta... one author of an accepted paper mus... www.cikmreg.org/ - 13k - Cached - S...
  - ACM CIKM 2003 Call For Pap...** ACM CIKM 2003 -Call for Industry Te... Sponsored by ACM SIGIR and ACM... www.cs.wisc.edu/dbworld/messages...
  - CIKM 2002 Homepage** ... 03), Las Vegas, Nevada, June 23-... Conference on Information and Know... Description: SAIC Headquarters, Mc... Category: Business > Management >... www.cikm.org/2002/ - 16k - Cached - S...
  - Conference on Information...** ... The CIKM 2003 web page i... devoted to emerging areas of... www.cikm.org/ - 6k - Cached - S...
  - [Asis-] CIKM 2003** [Asis-] CIKM 2003. Padmini Srinivas... Mon, 29 Sep 2003 12:59:36 -0500: P... mail.asis.org/pipermail/asis-l/ 2003-S...
  - (DBWORLD) CIKM 2003 (Pag...** (DBWORLD) CIKM 2003 (Padmini S... CIKM 2003; From: "Padmini Srinivas... www.witi.cs.uni-magdeburg.de/MailArc...
  - bridge-cikm-2003** bit.csc.lsu.edu/~cikm2003. www.logistics-2000-versailles.net/site...
  - Collaborative Filtering Mailing** [collab@siml] CFP: CIKM 2003. ...
- Middle Screenshot (Altavista):** Shows the Altavista search interface with the query "cikm 2003". The results include:
  - CIKM 2003 Homepage** 12th international conference on info... bit.csc.lsu.edu/~cikm2003 • Refresh... More pages from bit.csc.lsu.edu
  - ACM CIKM 2003 Call For...** ... Management (CIKM03) AC... caliber papers submitted to C... bit.csc.lsu.edu/~cikm2003/CI... More pages from bit.csc.lsu.e...
  - CIKM 2002 Homepage** ... 2002 Advance Technical Program... Treasurer, Dr. Nicholas, the CIKM02... www.cikm.org/2002 • Related Pages...
  - Conference on Information...** CIKM CIKM 2003 CIKM Topi... Knowledge Management The... www.cikm.org • Related Page... More pages from www.cikm.o...
  - CIKM** ... Information and Knowledge Mana... CIKM 2003 Home Page 11. CIKM 2... www.informatik.uni-trier.de/~ley/db/c... More pages from www.informatik.uni...
  - 7. CIKM 1998: Bethesda,** 7. CIKM 1998: Bethesda, Ma... Copyright © by Michael Ley C... www.informatik.uni-trier.de/~l... More pages from www.inform...
  - SIGIR Information Server** ... award. Upcoming SIGIR Sponsore... 2003. JCDL 2004 (Tucson, AZ) - Jun... www.acm.org/sigir • Refreshed in pa... More pages from www.acm.org
  - Web Caching Publications/Venu...** ...
- Right Screenshot (alltheweb):** Shows the alltheweb search interface with the query "cikm 2003". The results include:
  - CIKM 2003 Homepage** ... Information and Knowledge Management CIKM... an international ... Description: The ACM Conference on Information... and knowledge management, as well as recent ac... more hits from: http://bit.csc.lsu.edu/~cikm2003/...
  - IVML Call For Papers Archive: ACM CI...** Fo ACM CIKM 2003 Preliminary Call For Papers... message ... Digital Games Conference 2003" Prev... Description: Previous message: Kostas Karpouzis... Conferences - SPIE Web... http://www.image.ntua.gr/cfp/archive/2003/0234.ht...
  - CIKM 2002 Homepage** ... Information and Knowledge Management (CIKM... been ... Information and Knowledge Management... Description: SAIC Headquarters, McLean, Virginia... more hits from: http://www.cikm.org/2002/ - 15 k...
  - ACM WIDM'2003** ... CIKM 2003 ... Information and Knowledge Mana... of the workshop ... Description: WIDM03 is the fifth in a series of wor... Information and Knowledge Management (CIKM),... more hits from: http://www.cais.ntu.edu.sg/widm2...
  - MMDB'03** ... Programs ACM - MMDB 2003 The First ACM In... Knowledge Management (ACM CIKM 2003). One... Description: The First ACM International Worksho... university researchers, scientists, industry profess... more hits from: http://www.cs.fiu.edu/mmdb03/ -...
  - D-Lib Workshops and Conferences: 20...** ... proceedings, and other documents. 2004 2003... November 2003, Sicily, Italy ACM CIKM 2003, 12... Description: Below is a listing of digital library mee... Click on a year to find additional events with links... more hits from: http://www.dlib.org/groups.html -...
  - Invited paper to be presented in CIKM (** Invited paper to be presented in CIKM (Conference... Privacy Policy - Copyright © 2000-2003 NEC and... Description: Abstract: The main objective of a virtu... resources contributed by the organizations toward... more hits from: http://gunther.smeal.psu.edu/199...
  - Internet Conferences 2003** ... 2000 Books 2001 Books 2002 Books 2003 net... 2003... 2003... 2003... 2003...

[CIKM 2003 Homepage](#)  
[ACM CIKM 2003 Call For Paper](#)  
[CIKM-2003 Registration](#)  
[ACM CIKM 2003 Call For Paper](#)  
[CIKM 2002 Homepage](#)  
[Conference on Information and Knowledge Management](#)  
[\[Asia-1\] CIKM 2003](#)  
[\(DBWORLD\) CIKM 2003 \(Padm](#)  
[bridge-cikm-2003](#)  
[Collaborative Filtering Mailing](#)  
[CIKM Home Page ACM DL: CIK](#)  
[Yahoo! Groups : webir Messag](#)  
[dbforums - Cfp: Cikm 03](#)  
[Mailing List ARL-ERESERVE@ar](#)  
[Received: from cni.org by b.cr](#)  
[ACM WIDM 2003](#)  
[ACM - MMDB 2003](#)  
[ACM CIKM 2003 PRELIMINARY](#)  
[Selected Publications](#)  
[Web Caching Publications/Ver](#)  
[cikm '03](#)  
[Call for papers](#)  
[Mario A. Nascimento - Person](#)  
[Iceved](#)  
[Conferences and Journals on T](#)  
[Conferences On Information V](#)  
[Calendrier des manifestations](#)  
[\(DBWORLD\) Final Call for ACM](#)  
[Collaborative Filtering Mailing](#)

[CIKM 2003 Homepage](#)  
[CIKM-2003 Registration](#)  
[CIKM 2002 Homepage](#)  
[ACM CIKM 2003 Call For Paper](#)  
[ACM CIKM 2003 Call For Paper](#)  
[CIKM Home Page ACM DL: CIKM](#)  
[Conference on Information and Knowledge Management](#)  
[ACM CIKM 2003 PRELIMINARY PROGRAM](#)  
[CIKM Home Page ACM DL: CIKM](#)  
[ACM CIKM 2003 PRELIMINARY PROGRAM](#)  
[\[Asis-I\] CIKM 2003](#)  
[\(DBWORLD\) CIKM 2003 \(Padmabridge-cikm-2003\)](#)  
[Collaborative Filtering Mailing List](#)  
[Yahoo! Groups : webinfo Message Board](#)  
[dbforums - Cfp: Cikm 03](#)  
[Mailing List ARL-ERESERVE@arl](#)  
[Received: from cni.org by b.cri](#)  
[ACM WIDM 2003](#)  
[ACM - MMDB 2003](#)  
[Selected Publications](#)  
[Web Caching Publications/Version](#)  
[cikm '03](#)  
[Call for papers](#)  
[Mario A. Nascimento - Personal](#)  
[Received](#)  
[Conferences and Journals on Information](#)  
[Conferences On Information and Knowledge](#)  
[Calendrier des manifestations](#)

# Hedge adaptation to IR setup



The image displays three side-by-side screenshots of web search engines, each showing results for the query "cikm 2003".

- Google:** The top result is "CIKM 2003 Homepage" with a description of the conference. Other results include "ACM CIKM 2003 Call For Papers", "CIKM-2003 Registration", and "CIKM 2002 Homepage".
- Altavista:** The top result is "CIKM 2003 Homepage" with a description of the conference. Other results include "ACM CIKM 2003 Call For Papers", "CIKM 2002 Homepage", and "CIKM 1998: Bethesda".
- AlltheWeb:** The top result is "CIKM 2003 Homepage" with a description of the conference. Other results include "ACM CIKM 2003 Call For Papers", "CIKM 2002 Homepage", and "CIKM 1998: Bethesda".

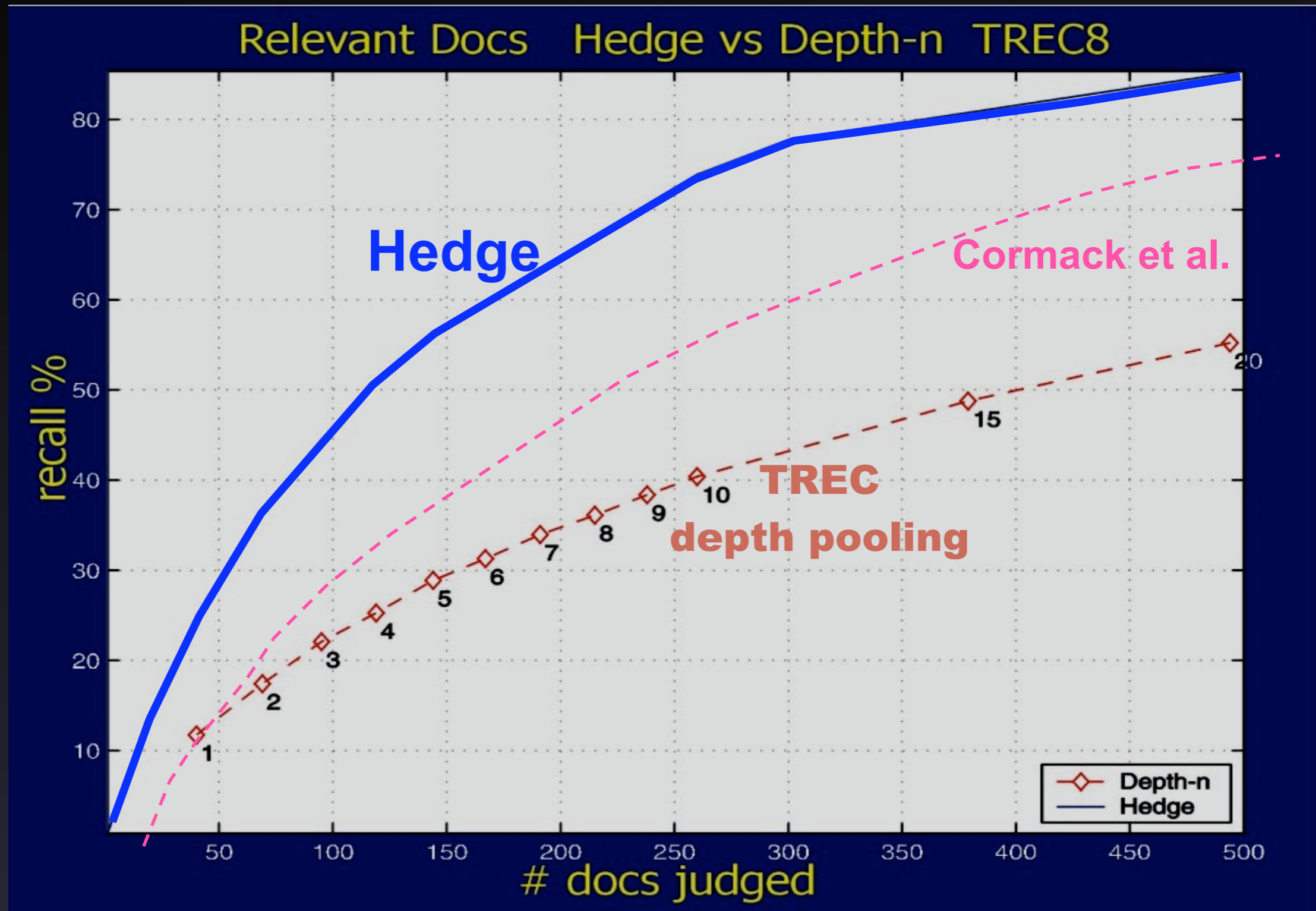
## metasearch

CIKM 2003 Homepage  
ACM CIKM 2003 Call For Paper  
CIKM-2003 Registration  
ACM CIKM 2003 Call For Paper  
CIKM 2002 Homepage  
Conference on Information and  
[Asis-I] CIKM 2003  
(DBWORLD) CIKM 2003 (Padm  
bridge-cikm-2003  
Collaborative Filtering Mailing  
CIKM Home Page ACM DL: CIK  
Yahoo! Groups : webir Messag  
dBforums - Cfp: Cikm 03  
Mailing List ARL-ERESERVE@a  
Received: from cni.org by b.cr  
ACM WIDM 2003  
ACM - MMDB 2003  
ACM CIKM 2003 PRELIMINARY  
Selected Publications  
Web Caching Publications/Ver  
cikm '03  
Call for papers  
Mario A. Nascimento - Person  
Iceved  
Conferences and Journals on T  
Conferences On Information V  
Calendrier des manifestations  
(DBWORLD) Final Call for ACM  
Collaborative Filtering Mailing

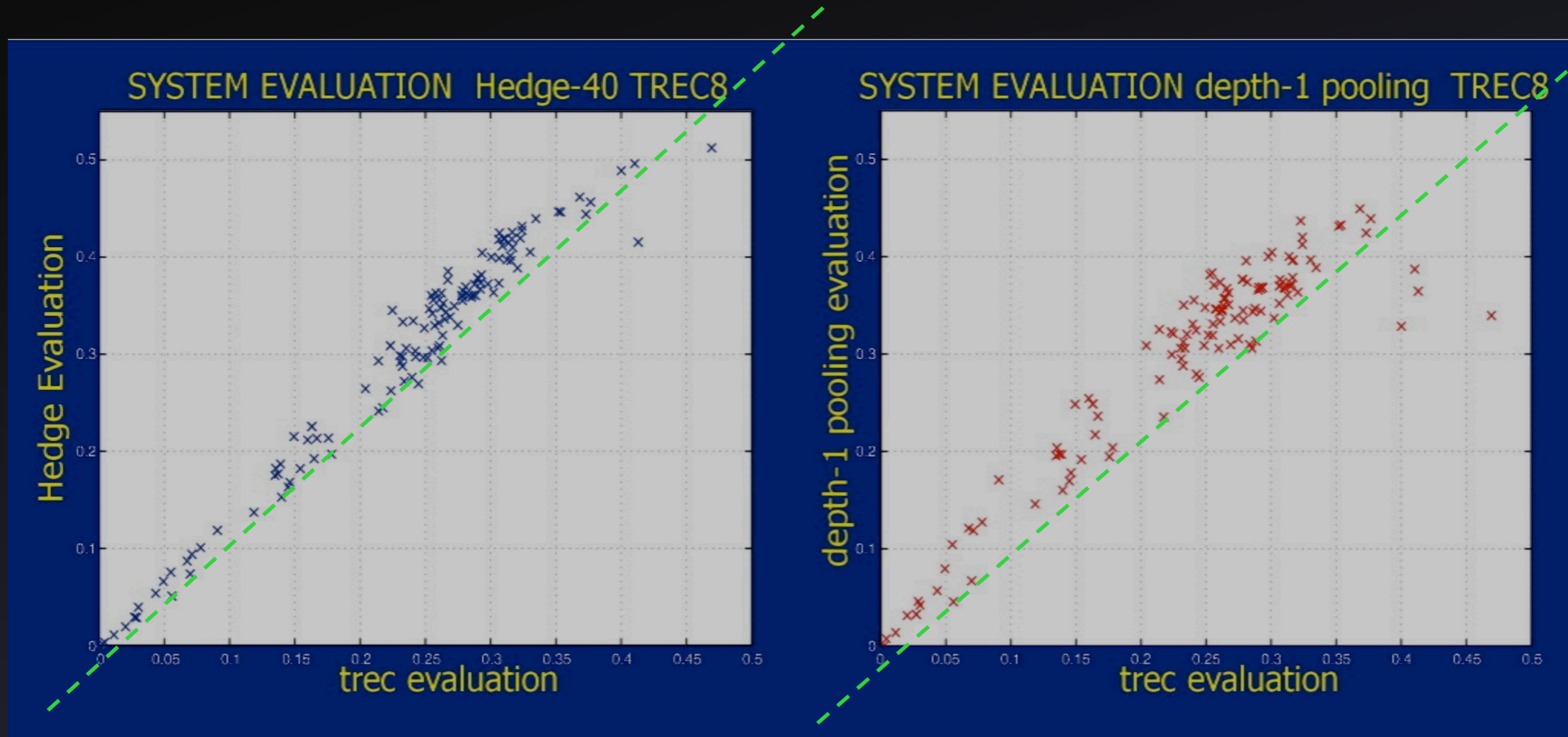
## metasearch

CIKM 2003 Homepage  
CIKM-2003 Registration  
CIKM 2002 Homepage  
ACM CIKM 2003 Call For Paper  
ACM CIKM 2003 Call For Paper  
CIKM Home Page ACM DL: CIK  
Conference on Information and  
ACM CIKM 2003 PRELIMINARY  
CIKM Home Page ACM DL: CIK  
ACM CIKM 2003 PRELIMINARY  
[Asis-I] CIKM 2003  
(DBWORLD) CIKM 2003 (Padm  
bridge-cikm-2003  
Collaborative Filtering Mailing  
Yahoo! Groups : webir Messag  
dBforums - Cfp: Cikm 03  
Mailing List ARL-ERESERVE@a  
Received: from cni.org by b.cr  
ACM WIDM 2003  
ACM - MMDB 2003  
Selected Publications  
Web Caching Publications/Ver  
cikm '03  
Call for papers  
Mario A. Nascimento - Person  
Iceved  
Conferences and Journals on T  
Conferences On Information V  
Calendrier des manifestations

# Hedge results - recall rate



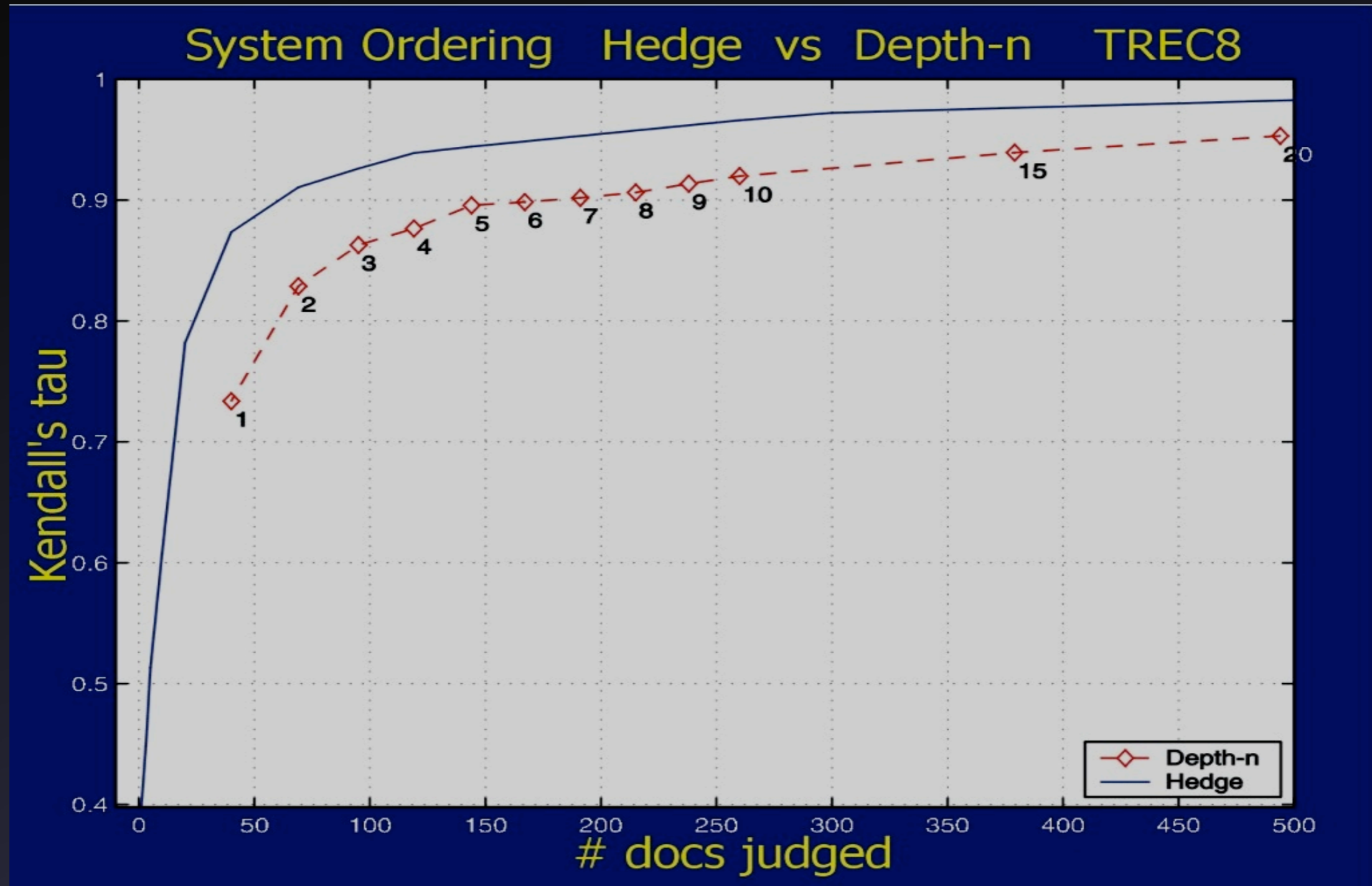
# Hedge results - evaluation



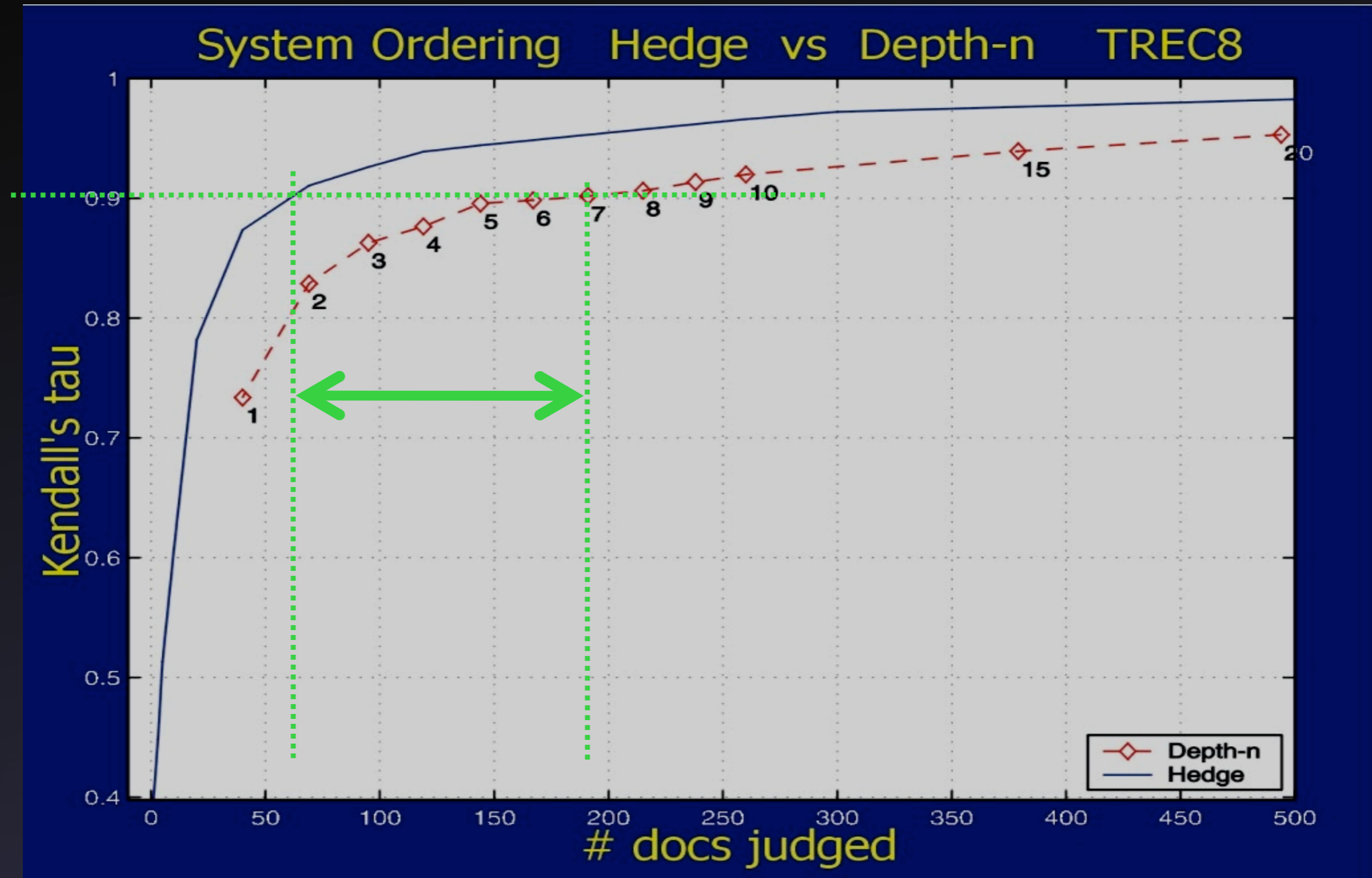
# Hedge results - evaluation

---

# Hedge Evaluation - Kendall's $\tau$



# Hedge Evaluation - Kendall's $\tau$



# Hedge results - metasearch

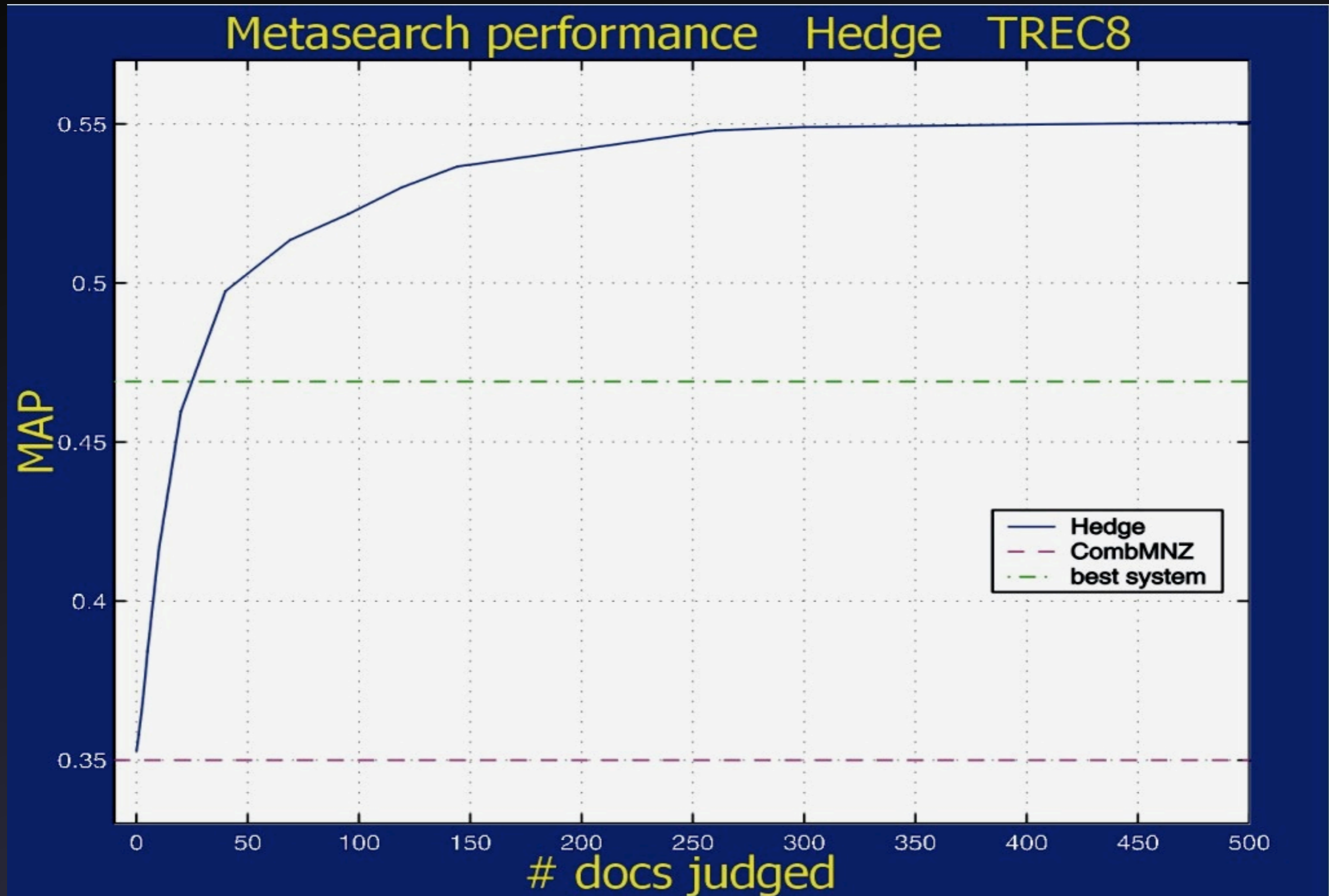
no relevant judgments  
(uniform system weights)

TREC	MNZ	COND	Hedge-0	% <i>MNZ</i>	% <i>COND</i>
3	0.423	0.403	0.418	-1.2	+3.7
5	0.294	0.307	0.309	+5.1	+0.6
6	0.341	0.315	0.345	+1.2	+9.5
7	0.320	0.308	0.323	+0.9	+4.9
8	0.350	0.343	0.352	+1.4	+2.6

MNZ=CombMNZ(Fox,Shaw,Lee et al)

COND=Condorcet(Aslam,Montague)

# Hedge results - metasearch



# overview

---

- Introduction
- Relevance Prior
- Hedge
- ◆ Sampling
- Future work

## SAMPLING TRICKS:

- non-uniform
- without replacement
- ratio estimator

# sampling example

---

- Say I have 1000 animals
- I want to find percentage of sick animals
- Obvious solution:
  - examine all 1000
  - return  $\text{\#sick}/1000$

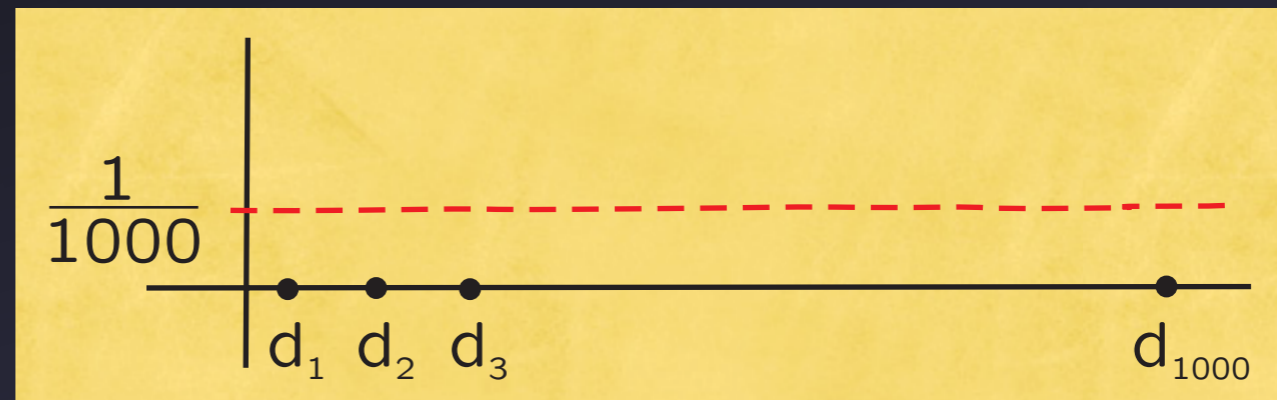
# sampling example

- Alternate solution:

- uniformly sample animals
- examine the sampled ones
- return  $\text{\#sick-seen}/\text{\#samples}$

- Distribution: uniform over 1000

- $p_i = 1/1000$



- Random variable:  $X = \text{sick}$

- 1 if sick, 0 if not

# theory

---

- LLN: Avg of random sample converges to mean

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

- CLT: But how fast ?

- Average of i.i.d. r. v. rapidly becomes Gaussian
- Mean is preserved
- Variance decreases linearly in  $n$
- SD decreases by root  $n$



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

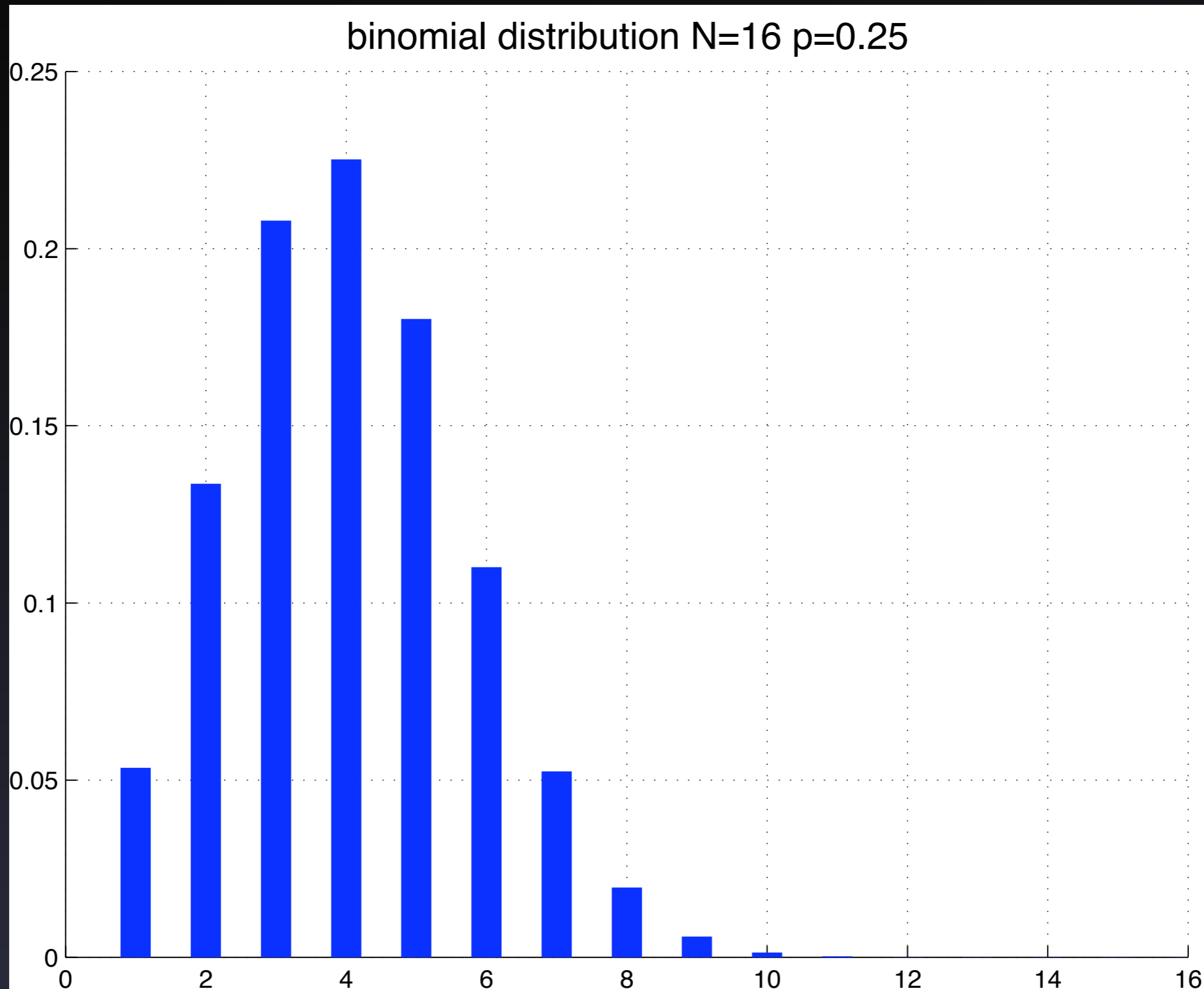
# estimator accuracy

---

- Suppose that, unknown to us,  
Percentage\_sick =  $p = 0.25$ 
  - i.e., there were 250 sick animals
- How many samples until we could estimate that accurately, say  $\pm 0.03$ ?
- LLN & CLT

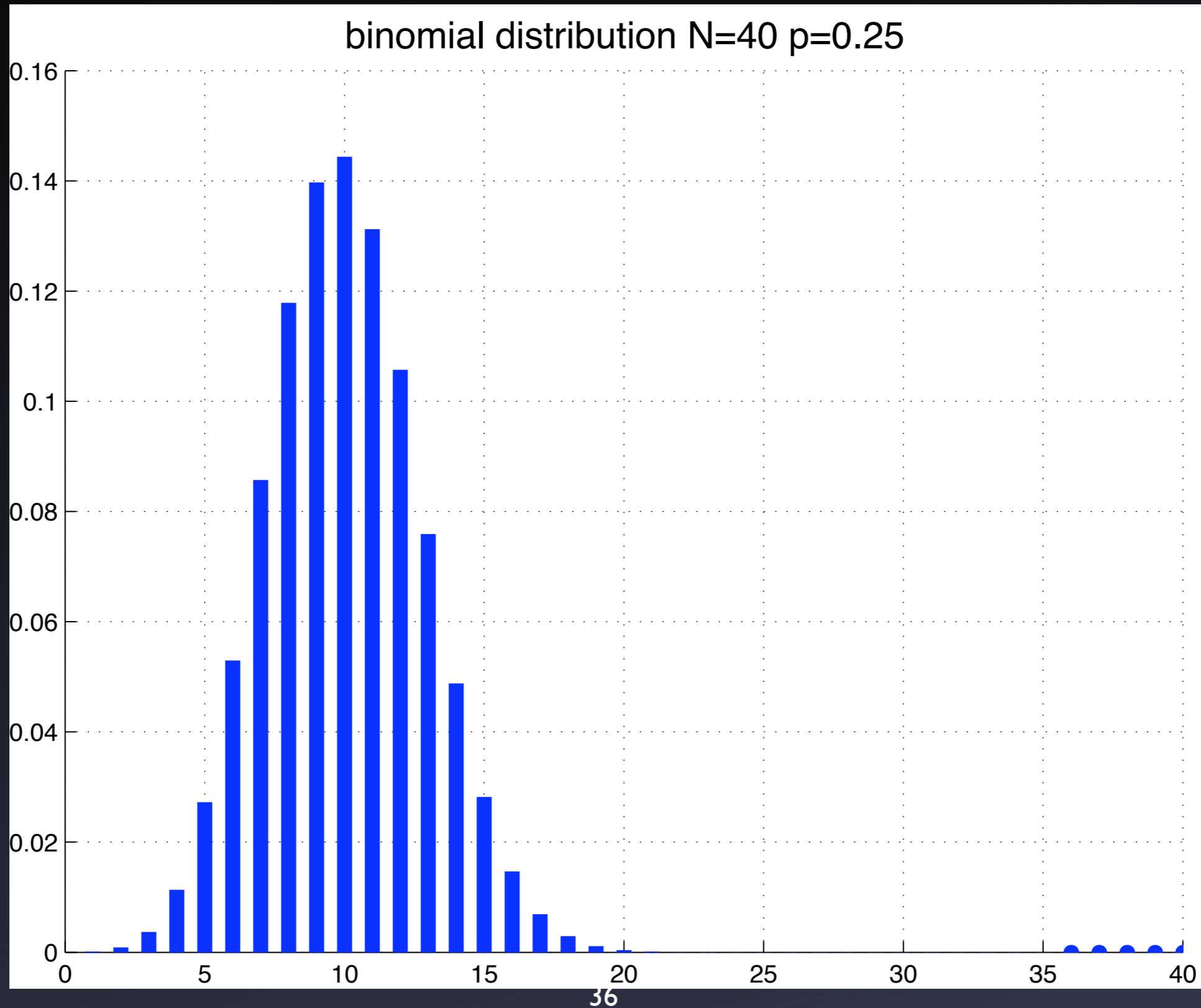
# LLN and CLT $N=16$

---



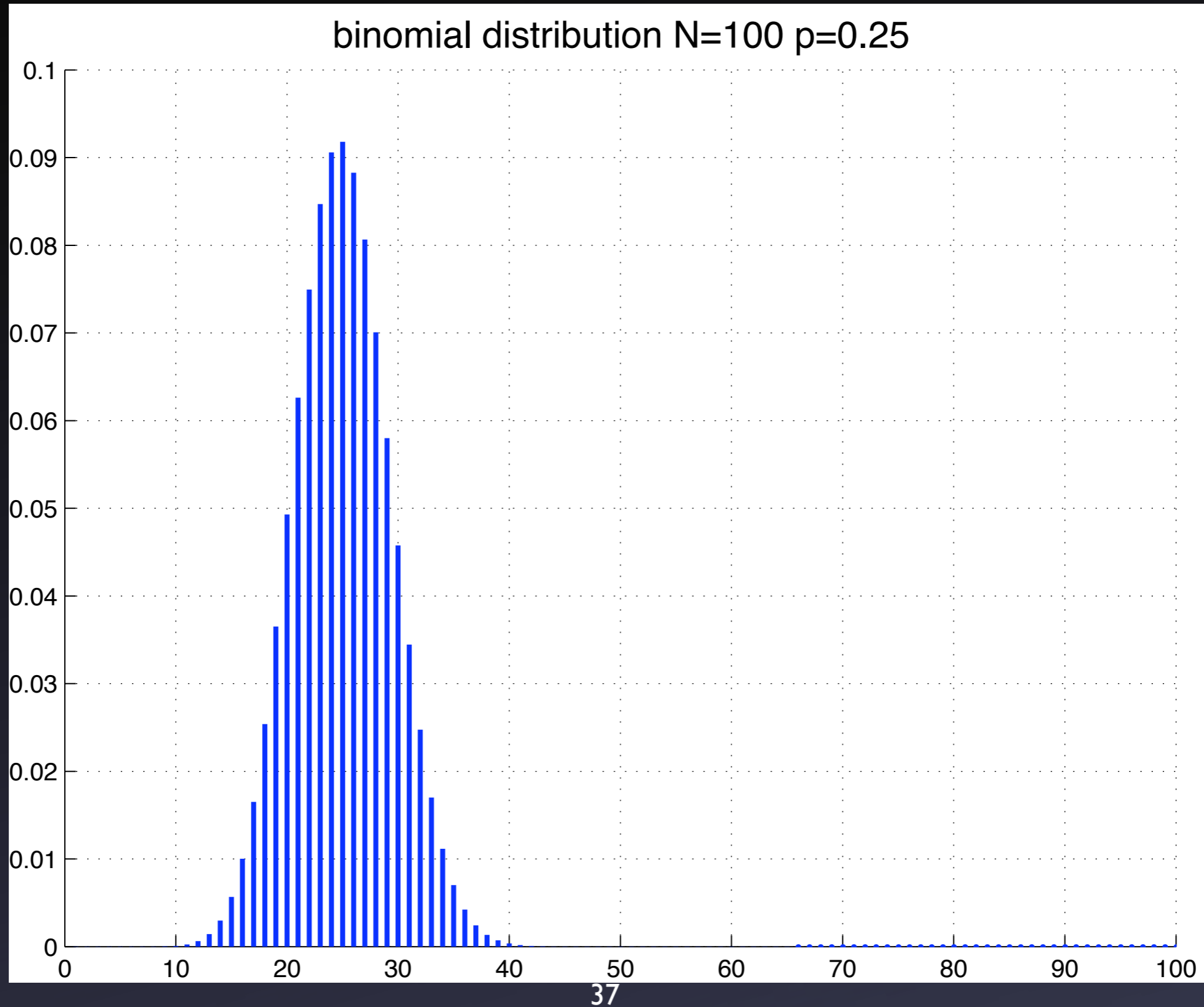
# LLN and CLT $N=40$

---

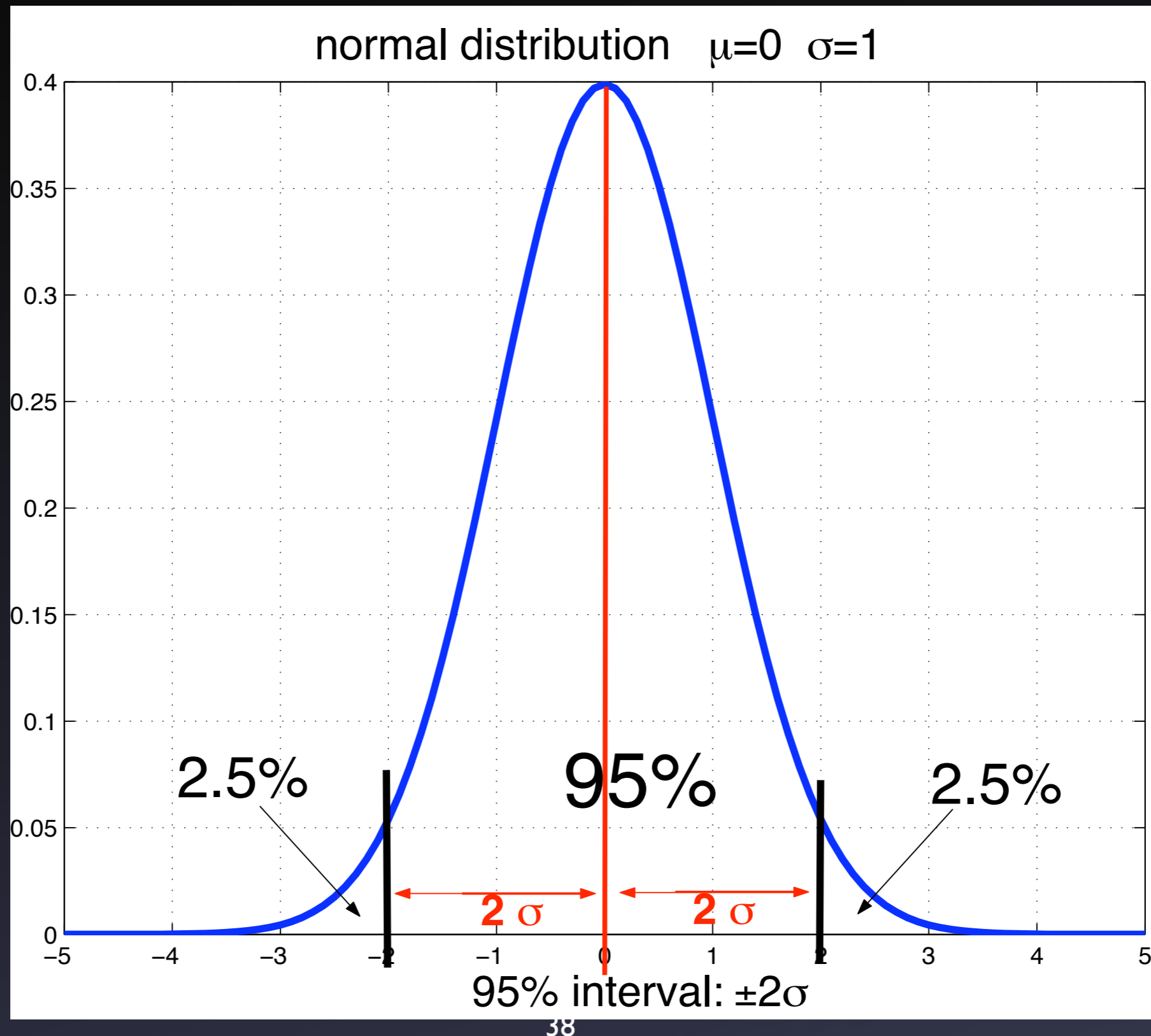


# LLN and CLT $N=100$

---



# Normal Distribution 95% CI



# uniform sample size

---

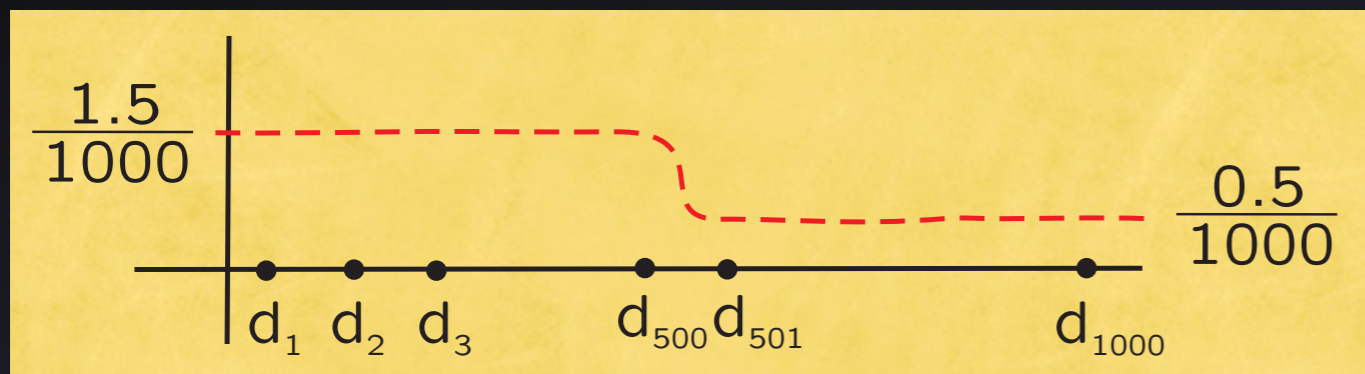
$$\text{Var}[X] = p(1 - p) = 3/16 \quad 2\sigma = 2\sqrt{p(1 - p)/n} = 0.03$$

$$\text{Var}[\bar{X}] = p(1 - p)/n \quad n = 4p(1 - p)/0.03^2 \approx 833$$

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

# importance sampling

- Sample “more” where sick animals are
  - for example categorize/order them by age:
    - 1–500 old; 501–1000 young



$$p_i = \begin{cases} 1.5/1000 & i \leq 500 \\ 0.5/1000 & i > 500 \end{cases}$$

- How to correct for estimated mean ?
  - scaling factors

$$x_i = \begin{cases} sick(i) \cdot 2/3 & i \leq 500 \\ sick(i) \cdot 2 & i > 500 \end{cases}$$

# 2 extreme cases

all sick in top half

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

sample size = 463

all sick in bottom half

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

sample size = 1944

● Comparison with uniform sampling:

— var = 1.87;      sample size required = 833

# sampling and evaluation

---

- non-uniform

- pps is ideal
- we are going to use the prior (avg over systems)

- without replacement

- $\pi_k$  = inclusion probabilities must be computed
- stratified sampling

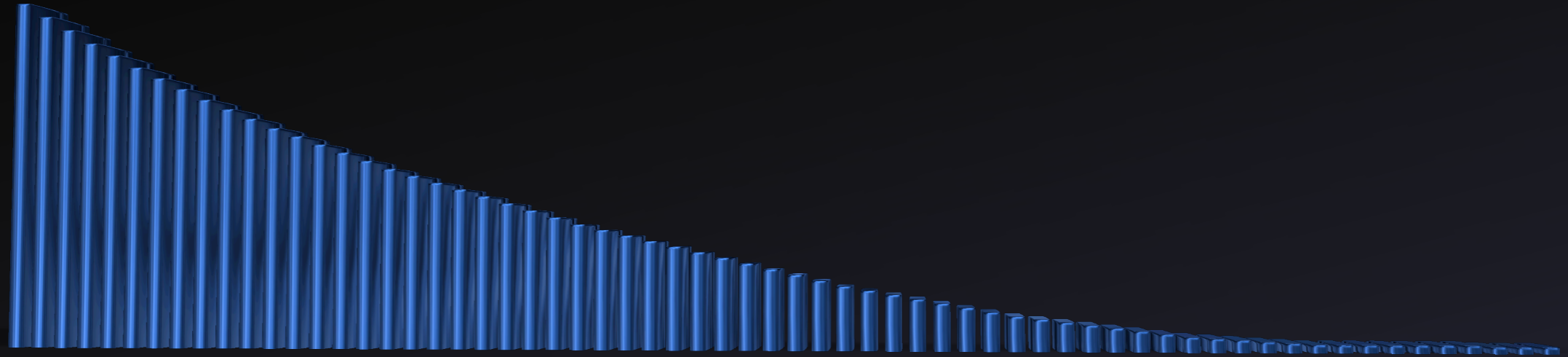
- use a ratio estimator

$$\hat{A}P = \frac{\sum_{k \in S} p_k / \pi_k}{\sum_{k \in S} 1 / \pi_k}$$

- prior, sampling and estimation independent

# stratified sampling

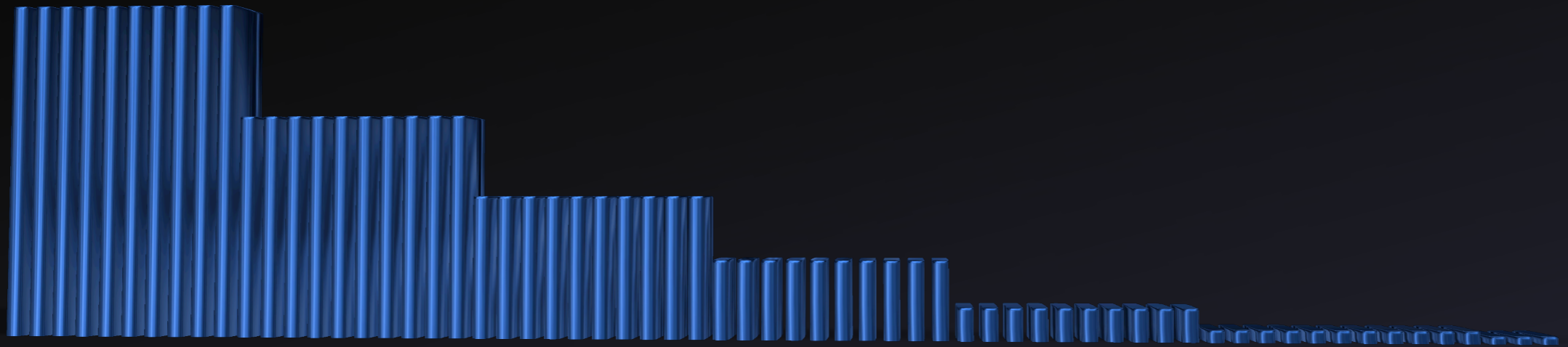
---



- non-uniform distribution; goal sample size =  $m$

# stratified sampling

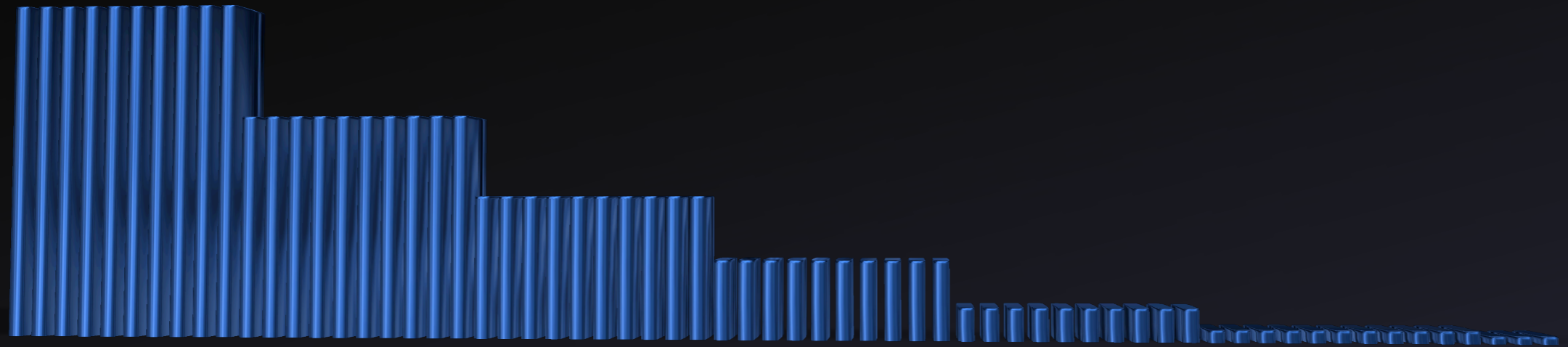
---



- non-uniform distribution; goal sample size =  $m$

# stratified sampling

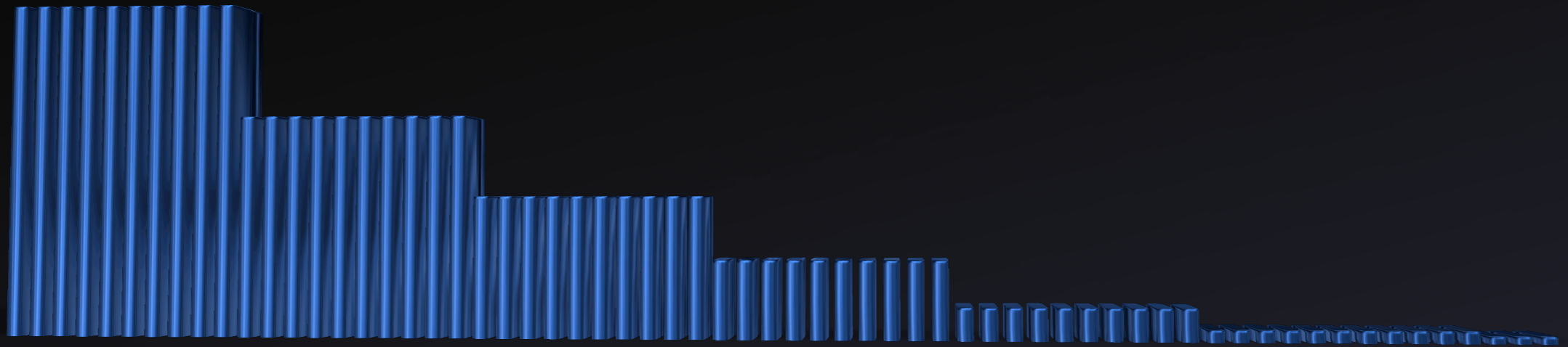
---



- non-uniform distribution; goal sample size =  $m$
- partition docs in buckets of size  $m$  each

# stratified sampling

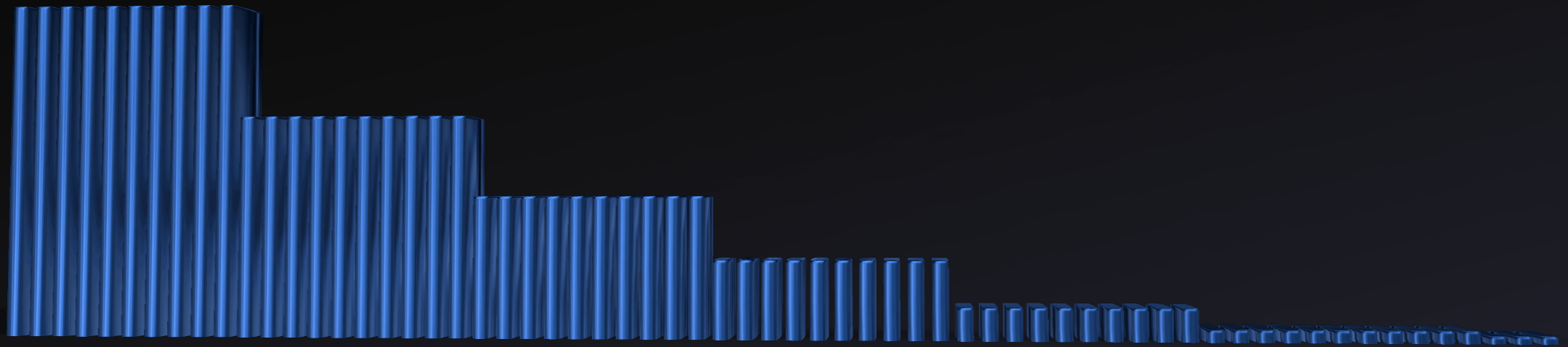
---



- non-uniform distribution; goal sample size =  $m$
- partition docs in buckets of size  $m$  each
- sample the buckets with replacement  $m$  times
  - based on the cumulative weight for each bucket

# stratified sampling

---



- non-uniform distribution; goal sample size =  $m$
- partition docs in buckets of size  $m$  each
- sample the buckets with replacement  $m$  times
  - based on the cumulative weight for each bucket
- for each bucket, if picked  $k$  times, sample uniform without replacement  $k$  docs in it

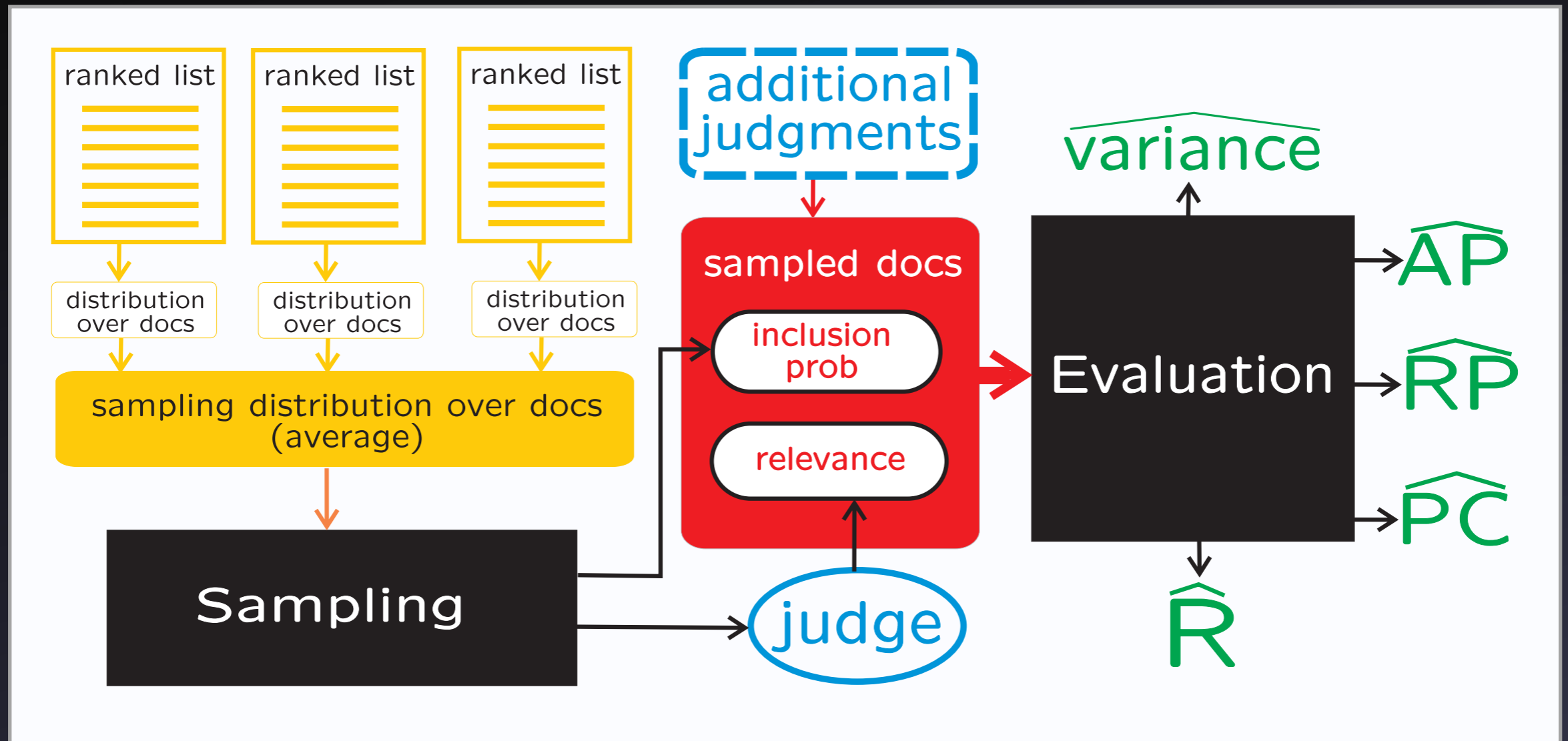
# sampling for AP estimation

---

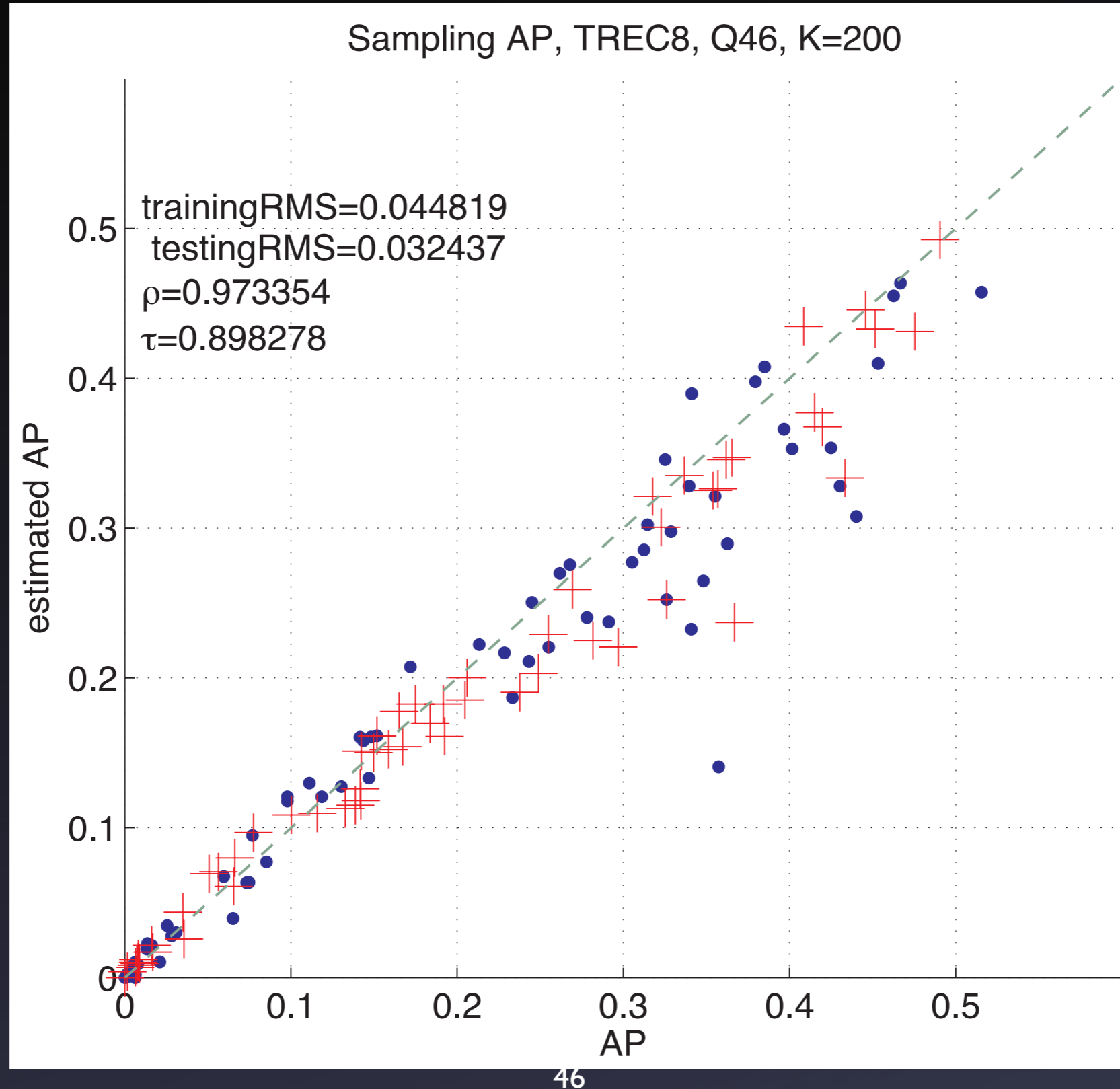
- AP is the average of precisions at relevant ranks
  - population is precision values at relevant ranks
  - those need to be estimated too
- Estimate performance on other systems
- We can also estimate other measures/quantities
- We can use additional judgments

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

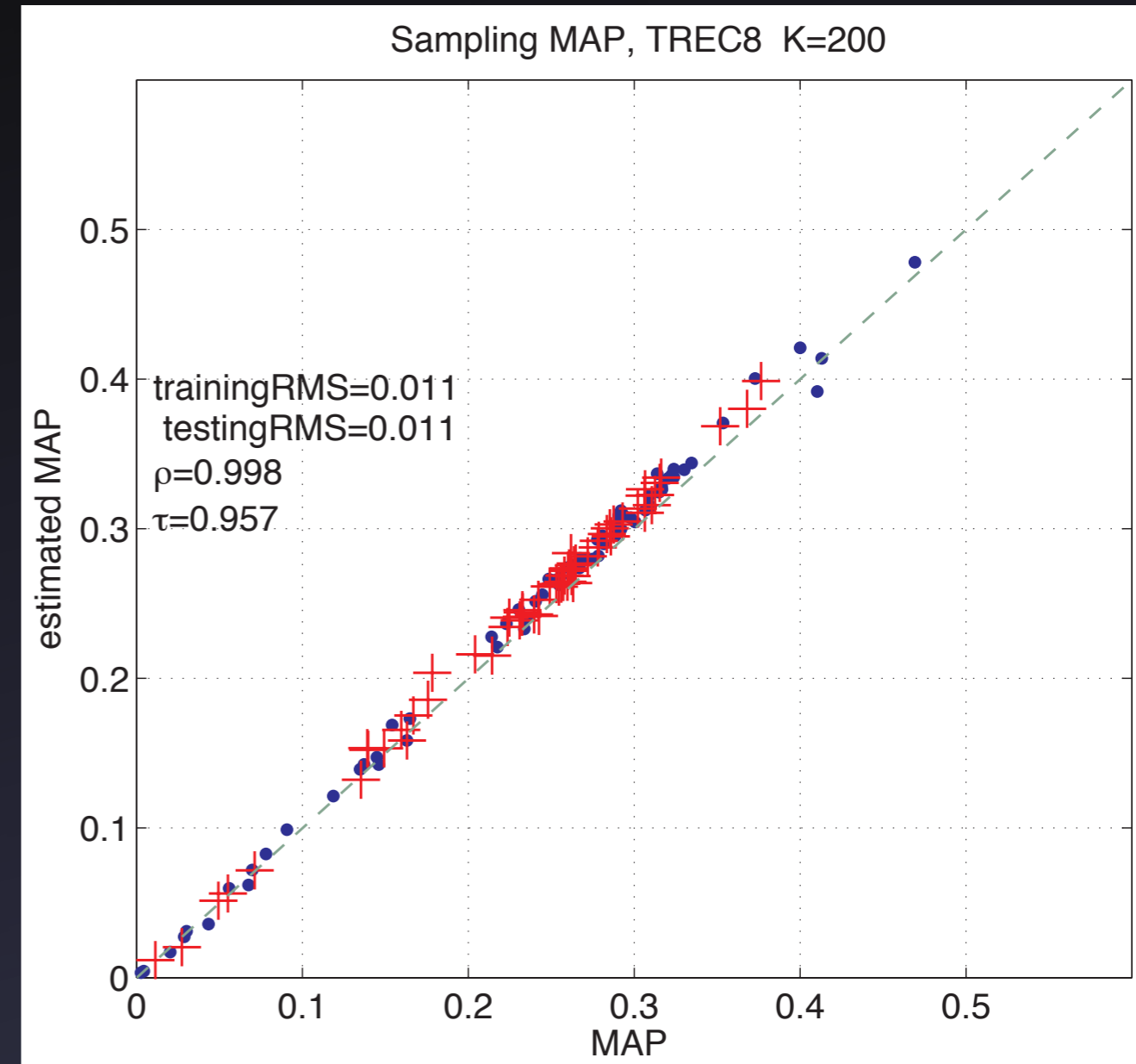
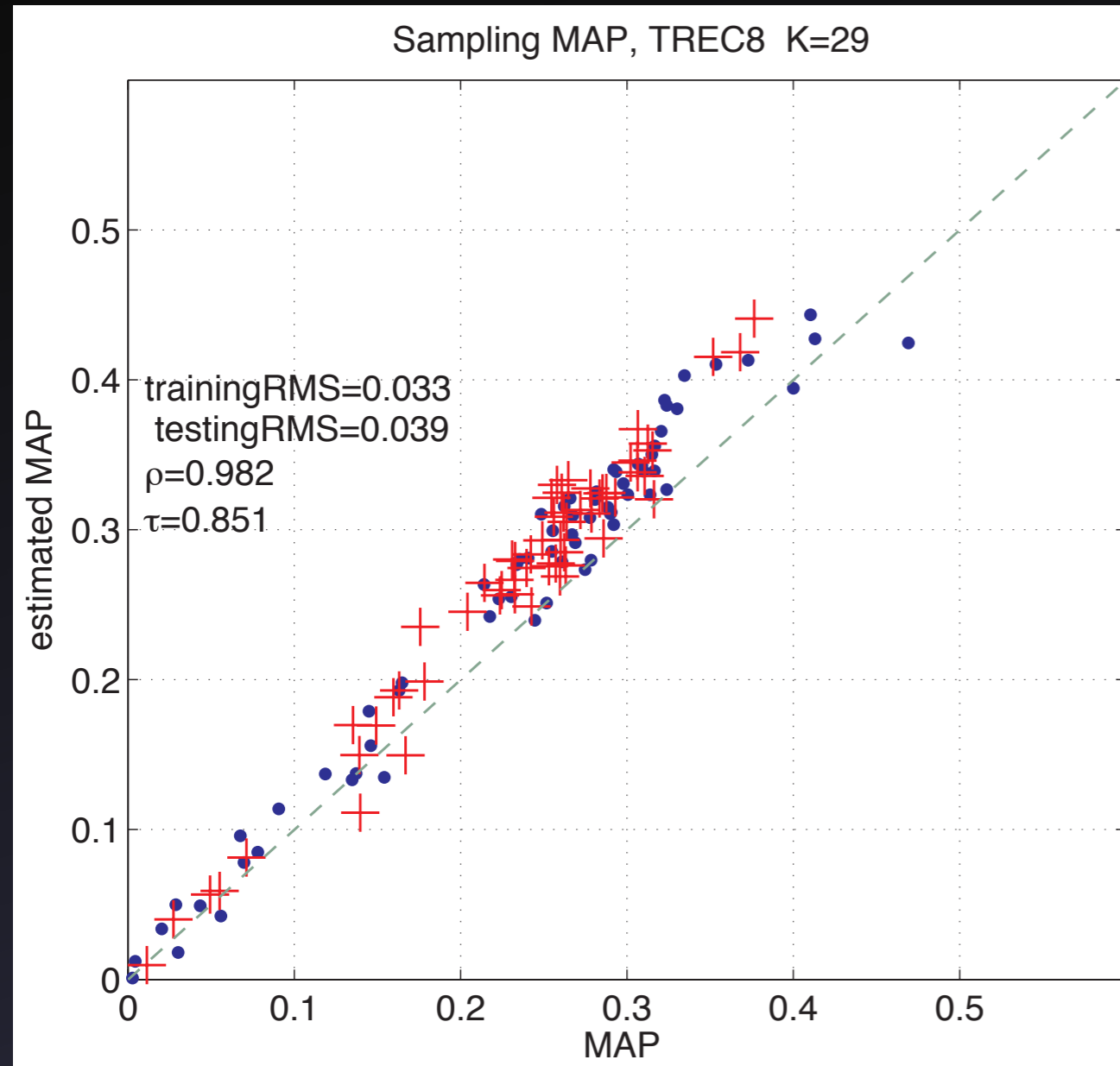
# sampling for AP estimation



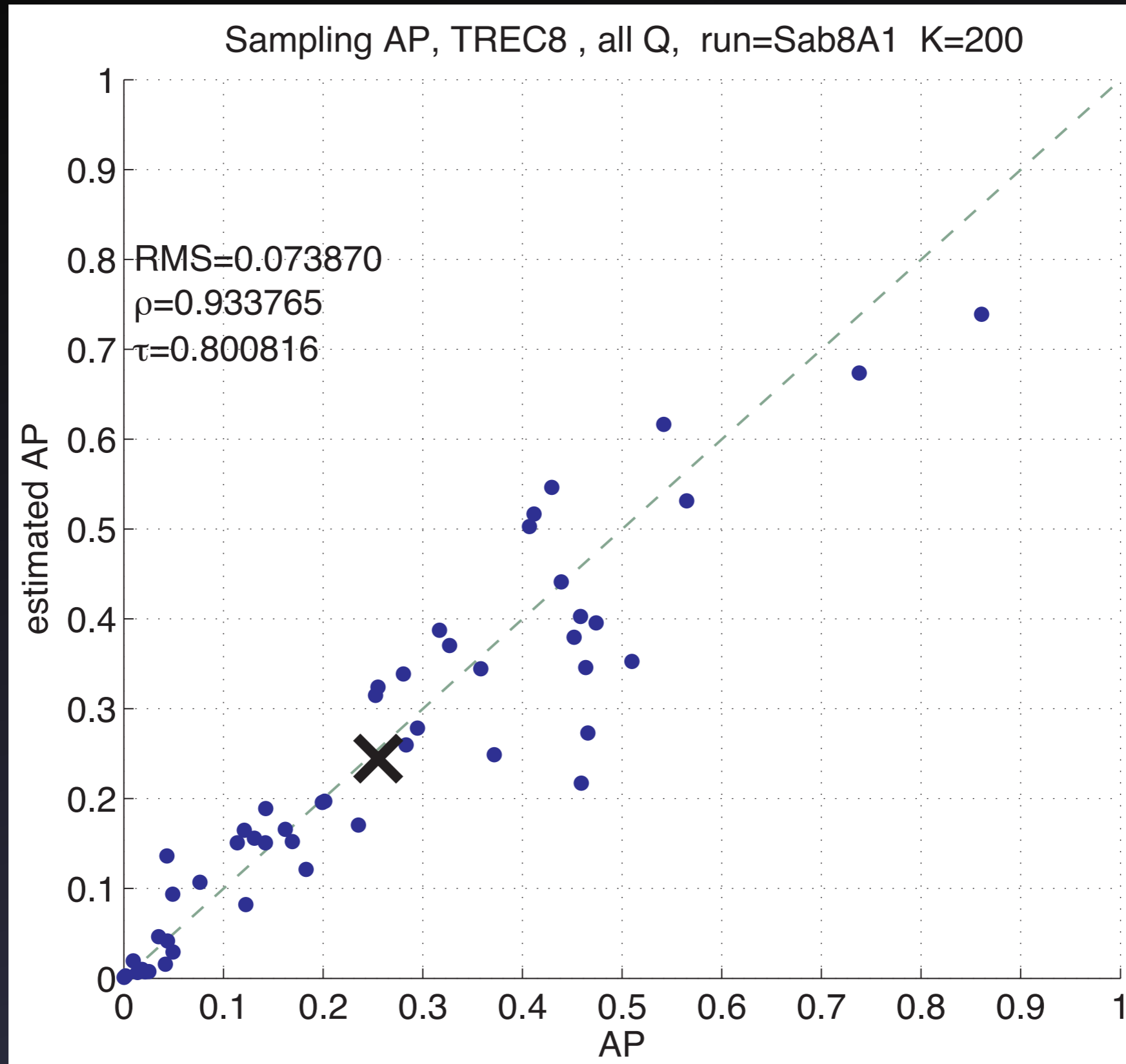
# sampling results: one query



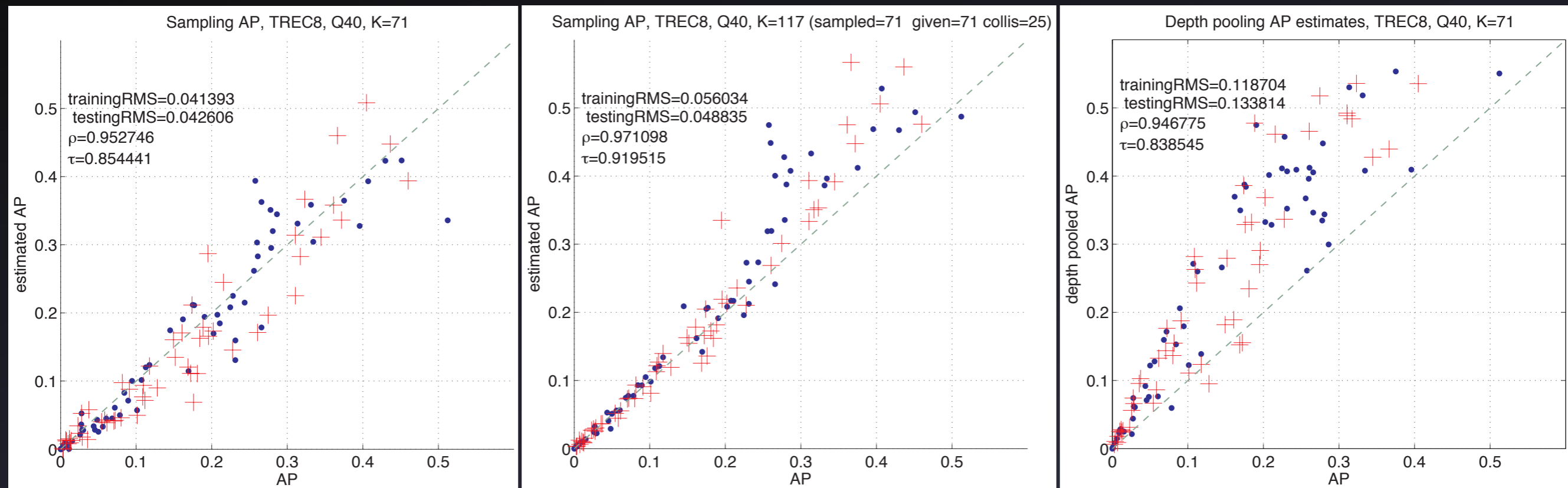
# sampling results: all queries



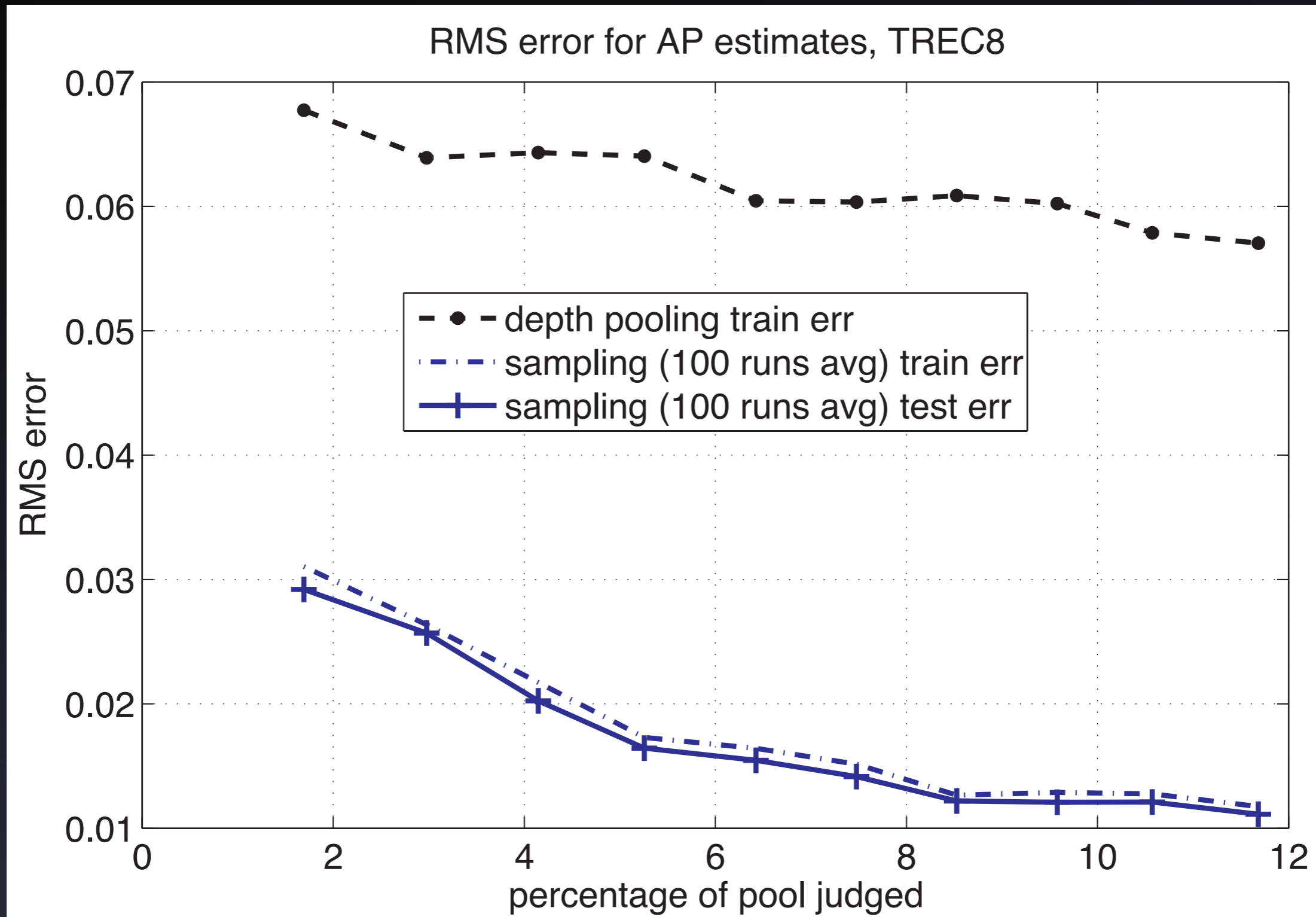
# sampling results: one system



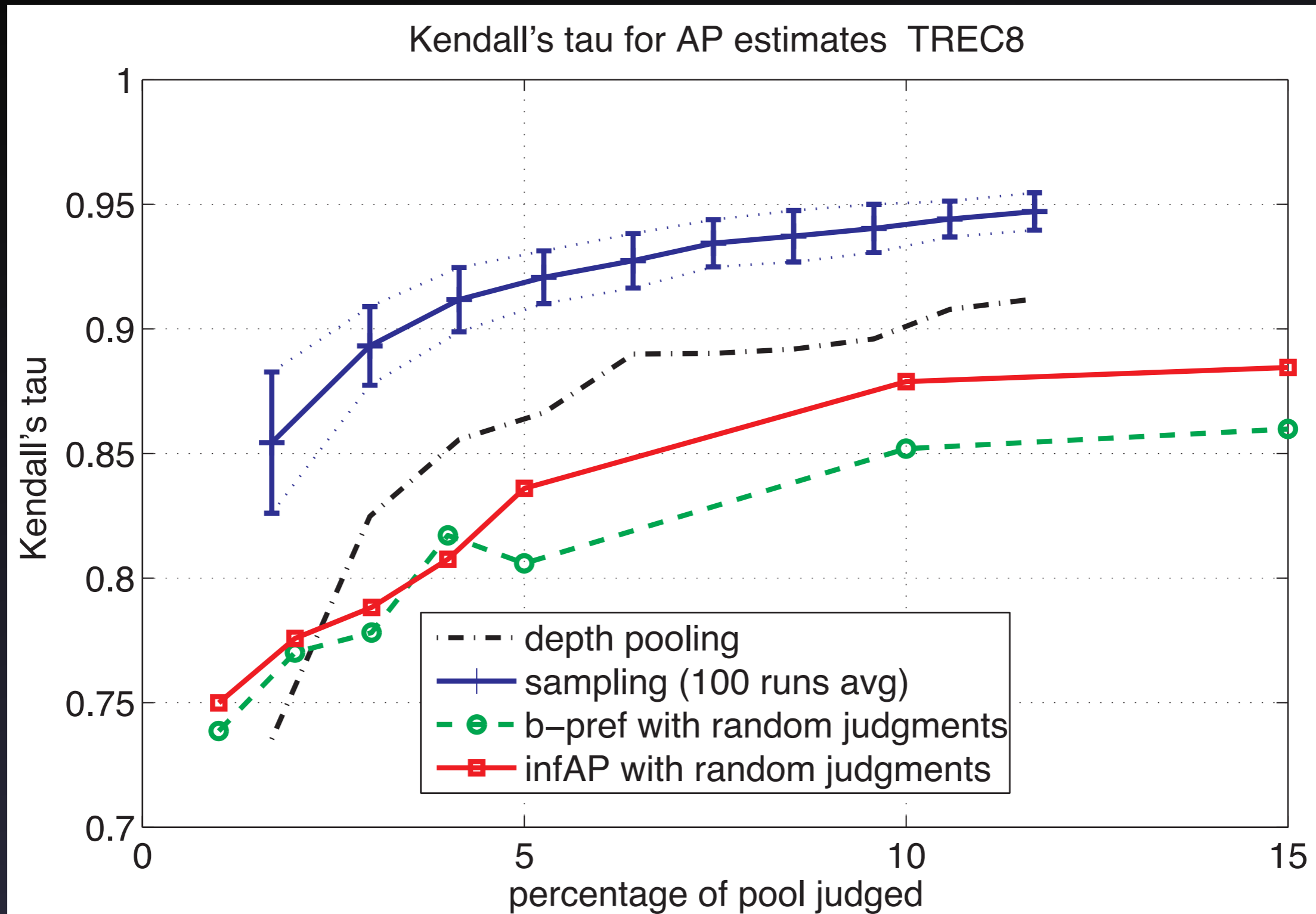
# add deterministic judgments



# sampling results : trends(RMS)



# sampling results : trends( $\tau$ )



# overview

---

- Introduction
- Relevance Prior
- Hedge
- Sampling
- ◆ Future work

# future work: Hedge optimization

---

- Bound known to be optimal for that class of algorithms

- but performance much better than the bound

$$\text{LOSS}_{Hedge} \leq \frac{\min_b \{L_b\} \cdot \ln(1/\beta) + \ln(N)}{1-\beta}$$

- Use a variable  $\beta$

- can we get a better bound ?

- Experiment on machine learning datasets

# future work: sampling bias

---

- the estimator used is not unbiased
  - but bias very small as sample grows

$$\hat{A}P = \frac{\sum_{k \in S} p_k / \pi_k}{\sum_{k \in S} 1 / \pi_k}$$

- Idea: for small sample sizes use a correction

$$\widehat{AP} = \frac{\widehat{SP}}{\widehat{R}} \cdot \frac{1}{1 + \frac{\sigma_R^2}{\widehat{R}^2}}$$

# future work: sampling variance

---

- pair-inclusion probabilities can be computed

$$\pi_{df} = \begin{cases} \pi_d \cdot \pi_f & \text{if } \pi_d = \pi_f \\ \frac{|S|-1}{|S|} \pi_d \cdot \pi_f & \text{if } \pi_d \neq \pi_f \end{cases}$$

- the variance of the estimator

$$\widehat{var}(\widehat{AP}) = \frac{1}{\widehat{R}^2} \left( \sum_{d \in S, rel} \frac{1 - \pi_d}{\pi_d^2} p_d^2 - \frac{1}{|S| - 1} \sum_{d, f \in S, rel} \frac{p_d \cdot y_f}{\pi_d \cdot \pi_f} \right)$$

- derive confidence intervals

assuming  
we can fix  
the bias

# future work: sampling + active learning

---

- experiment with mixing the samples
  - we know how to mix a random sample with a deterministic one
  - try to combine two samples obtained random
- try an active learning strategy
  - mixed within sampling process

# future work: sampling + active learning

---

- experiment with mixing the samples
  - we know how to mix a random sample with a deterministic one
  - try to combine two samples obtained random
- try an active learning strategy
  - mixed within sampling process

	online	batch
random	?	sampling
deterministic	Hedge	depth-pool

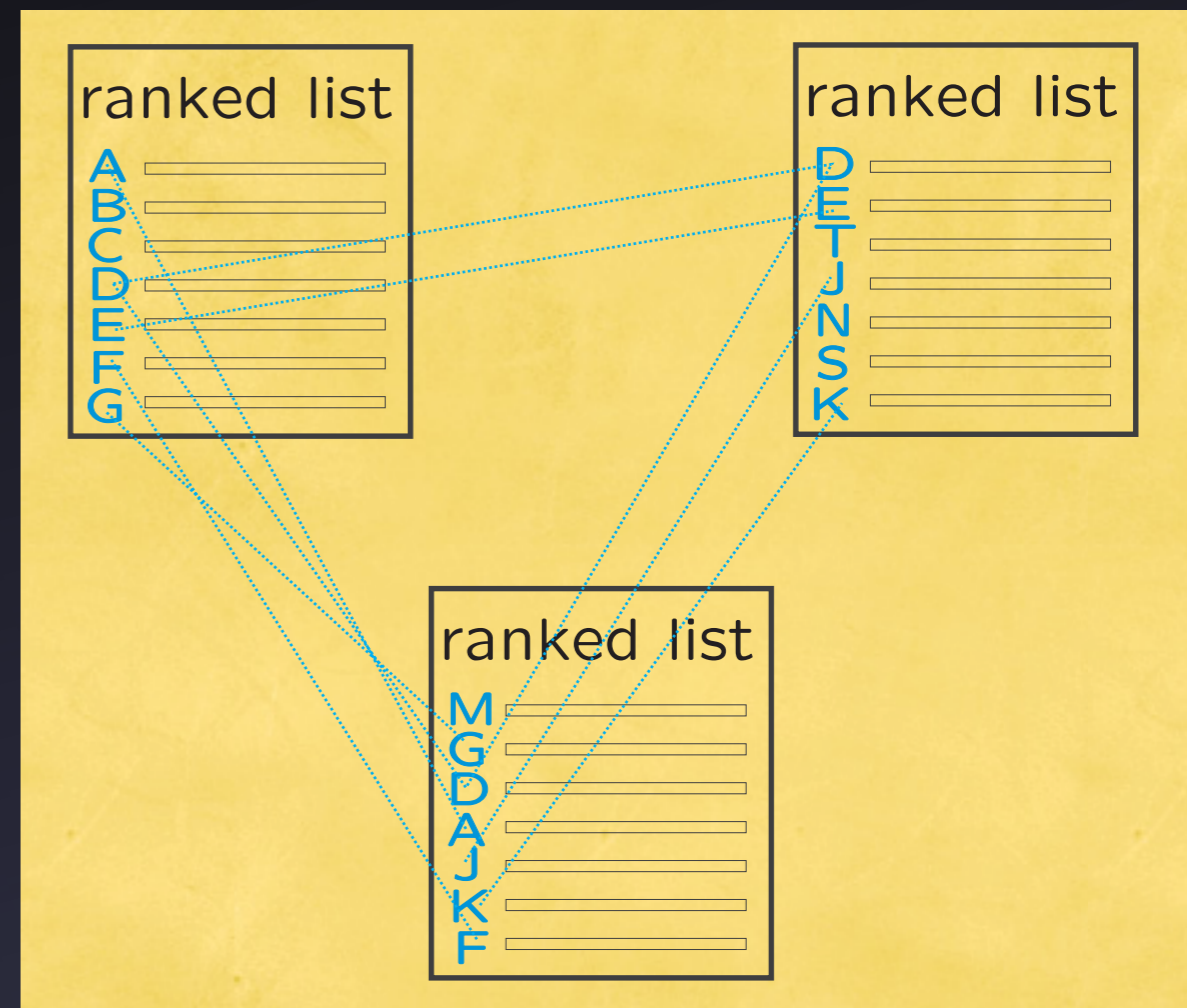
# future work: maximum entropy

---

- How much information can be in one number ?
  - given AP of a list, what can we infer about it?
  - particular interest in Precision-Recall curve
- Maximum entropy method
  - given some constraints and an objective model
  - compute the instance of the model that has the highest entropy and satisfies the constraints
  - usually involves constrained nonlinear optimization

# future work : query hardness

- Say we have access to many lists in response to the same query
- Query Hardness Hyp:
  - hard if the lists are diverse
  - easy if the lists are similar
- How to measure diversity?
  - transform lists in distributions
  - compare the distributions



# future work: prec-recall models

---

- Task : construct prec-recall models based on search engine type
- Possible leads:
  - explore connection between AP, RP and MedianP
  - known distributions for rel/nonrel documents, also derived based on engine type [Manmatha et al, Robertson et al]
  - work by S.Robertson on fall-out curves
  - ROC curves
  - $p(r) = (1 - r)/(1 + \alpha * r)$

# time line

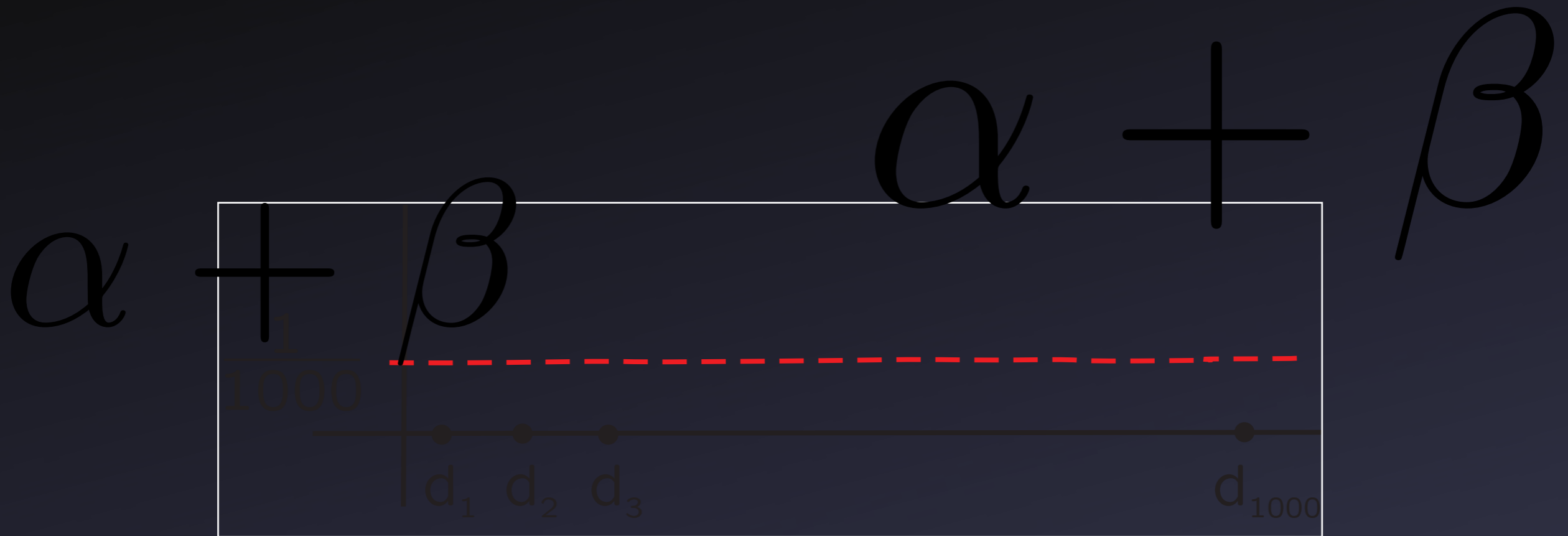
---

- Query Hardness Estimation      June 07
- Max Entropy Application      June 07
- Priors, Prec-Recall Models      October 07
- Sampling: bias      January 08
- Sampling: variance, CI      January 08
- Active learning      April 08
- Defend thesis      Summer 08

# Thank You

---

$$\alpha + \beta_{\text{Text}}$$



$$n = 4p \cdot (1 - p) / (0.03)^2 \approx 833$$