Advances in IR Evaluation

Ben Carterette Evangelos Kanoulas Emine Yilmaz

The University Of Sheffield.

Microsoft Research

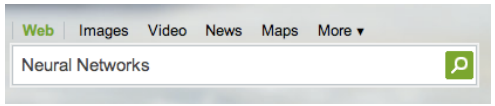
Information Retrieval Systems

Match information seekers with
the information they seek

I want to learn about Neural Nets…

User's Request

Search Engine



Results

Collection

Change the search algorithm.

How can we know whether we made the users happier?

“What you can’t measure
you can’t improve”
Lord Kelvin

By the end of this course…

You will be able to evaluate your retrieval algorithms

A. At low cost
B. Reliably
C. Effectively

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

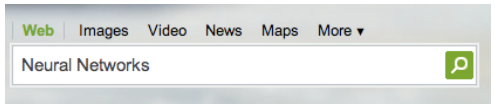
Today's Outline

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

I want to learn about Neural Nets…

User's Request

Search Engine



Results

Collection

Interactive Evaluation

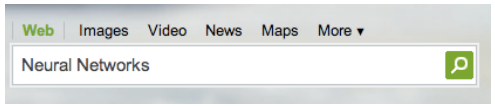
The user is part of the evaluation

"What is a neural net?"

"Pictures of white nights in St. Petersburg"

Participants' Request

Search Engine



Results

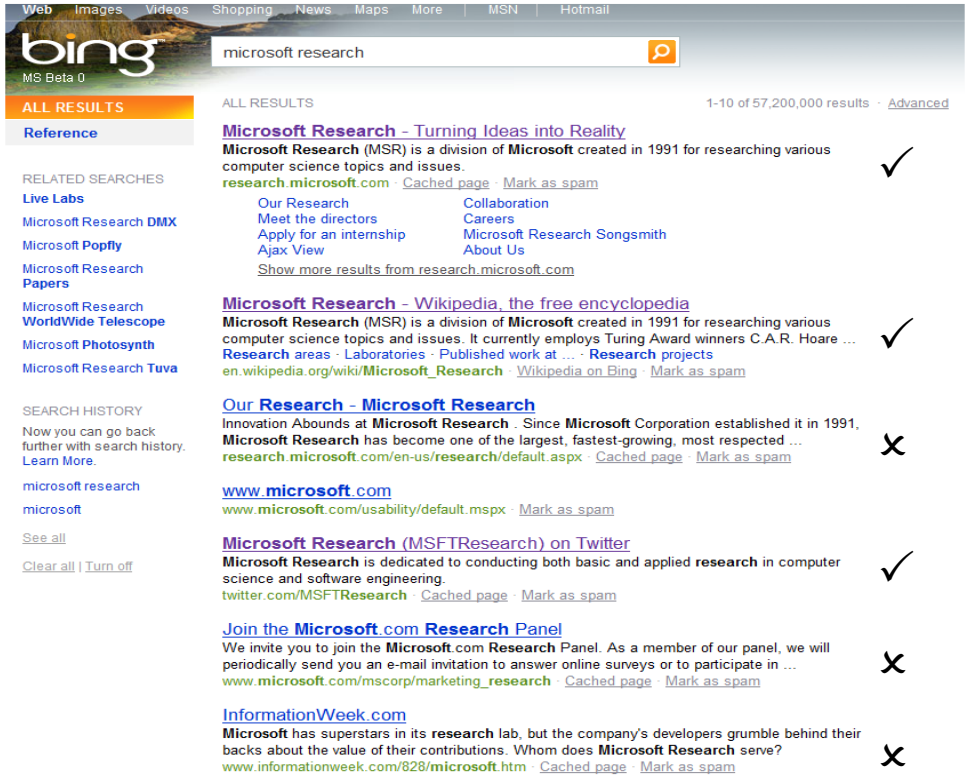
Collection

Online Evaluation

Go LIVE!!!

and observe the user's trails

Online Evaluation



✓
✓
✗
✓
✗
✗

Click/Noclick

Evaluate

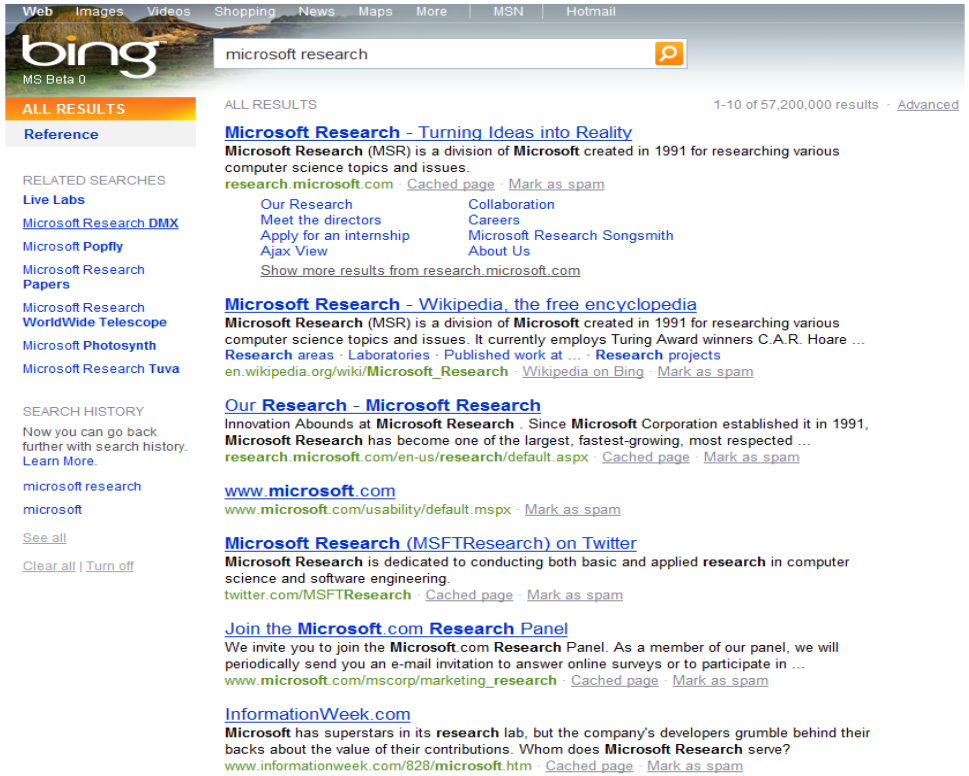
- Design interactive experiments
- Use users' actions to evaluate the quality

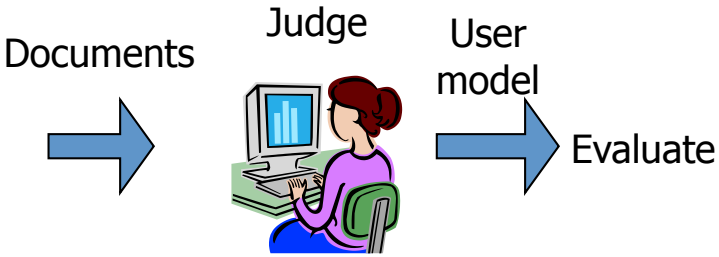
Offline Evaluation

Controlled laboratory experiments

The user's interaction with the engine is only simulated

Offline Evaluation





- Ask experts to judge each query result
- Predict how users behave when they search
- Aggregate judgments to evaluate

	Interactive	On-line	Off-line
Pros	User is part of the process Evaluate the overall experience and utility Learn “why’s”	Cheap Measure actual user reactions	Fast Easy to try new ideas Amortized Cost Portable
Cons	Noisy Not duplicable Expensive	Need to go live Noisy Slow Not duplicable	Needs ground truth Slow “Expensive” Inconsistent Difficult to model how users behave

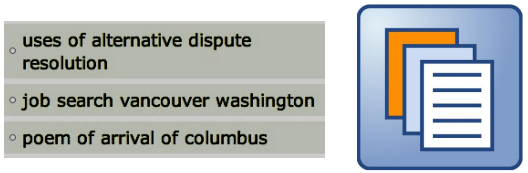
Offline vs. Online Evaluation

- Offline Evaluation
  - Used to optimize the engine
    - Most current engines optimize for an offline evaluation metric
- Online Evaluation
  - More correlated with user satisfaction
    - Directly derived from user behaviour (clicks)

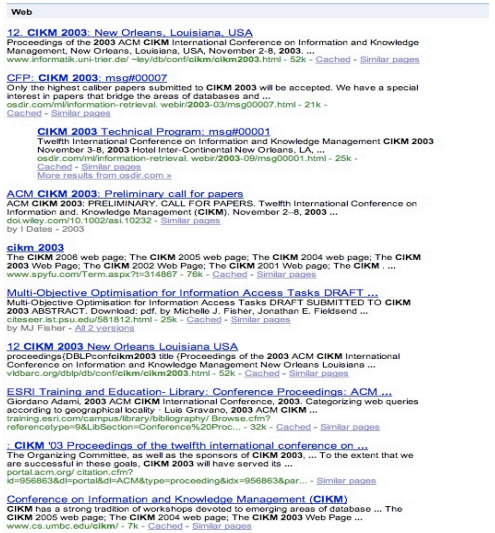
Today's Outline

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

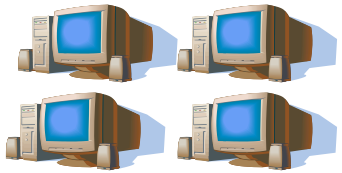
Traditional Experiment

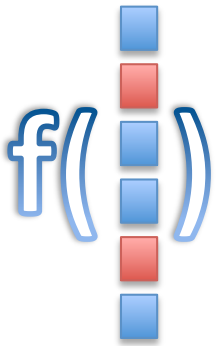


Results

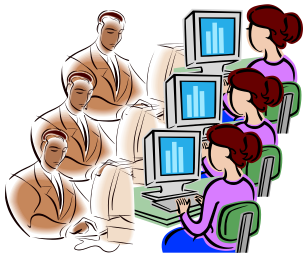


Search Engines





Judges

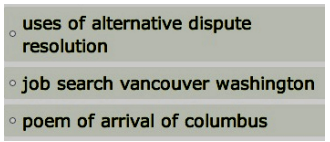


How many good docs have I missed/found?

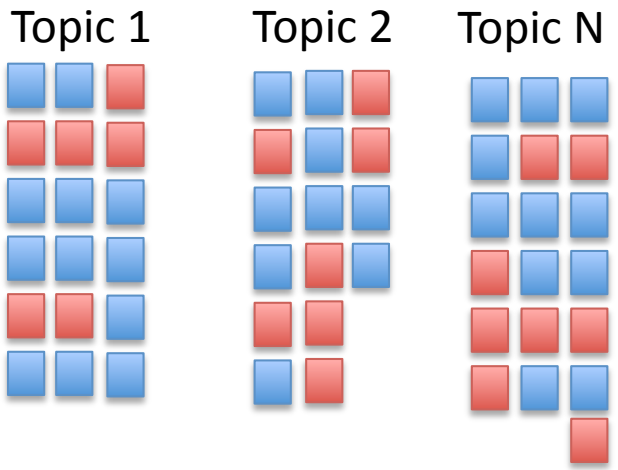
Portable Test Collection

- Document Corpus

- Topics



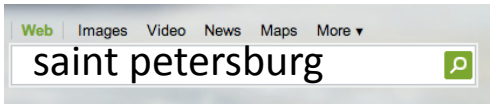
- Relevance Judgments (QRELs)



Simulation

Id	Topic	QRELS
1	Saint Petersburg	7623, 3256
2	Aurora	5425, 7654, 9582
3	Pitergov	3417, 6589
4	Dostoyevsky	6539, 8042
5	Peter the Great	4375, 5290
6	Russian Standard	9301, 7392
	…	…

Results (Run)

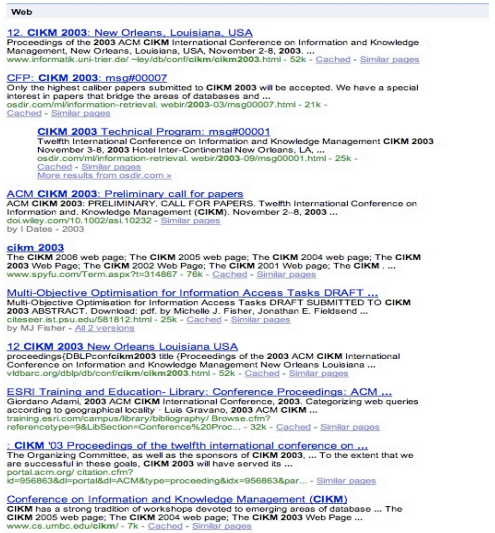


Id	Quality
1	…
2	
3	
…	

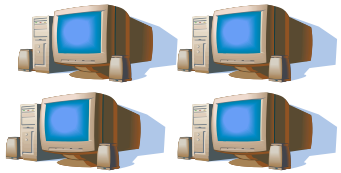
Traditional Experiment

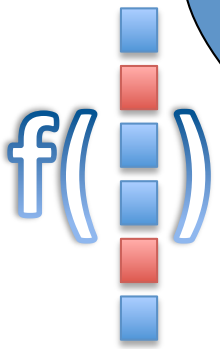
- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus

Results

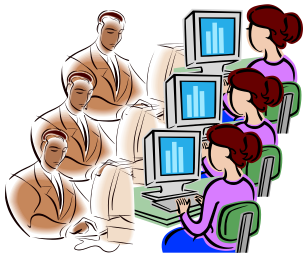


Search Engines

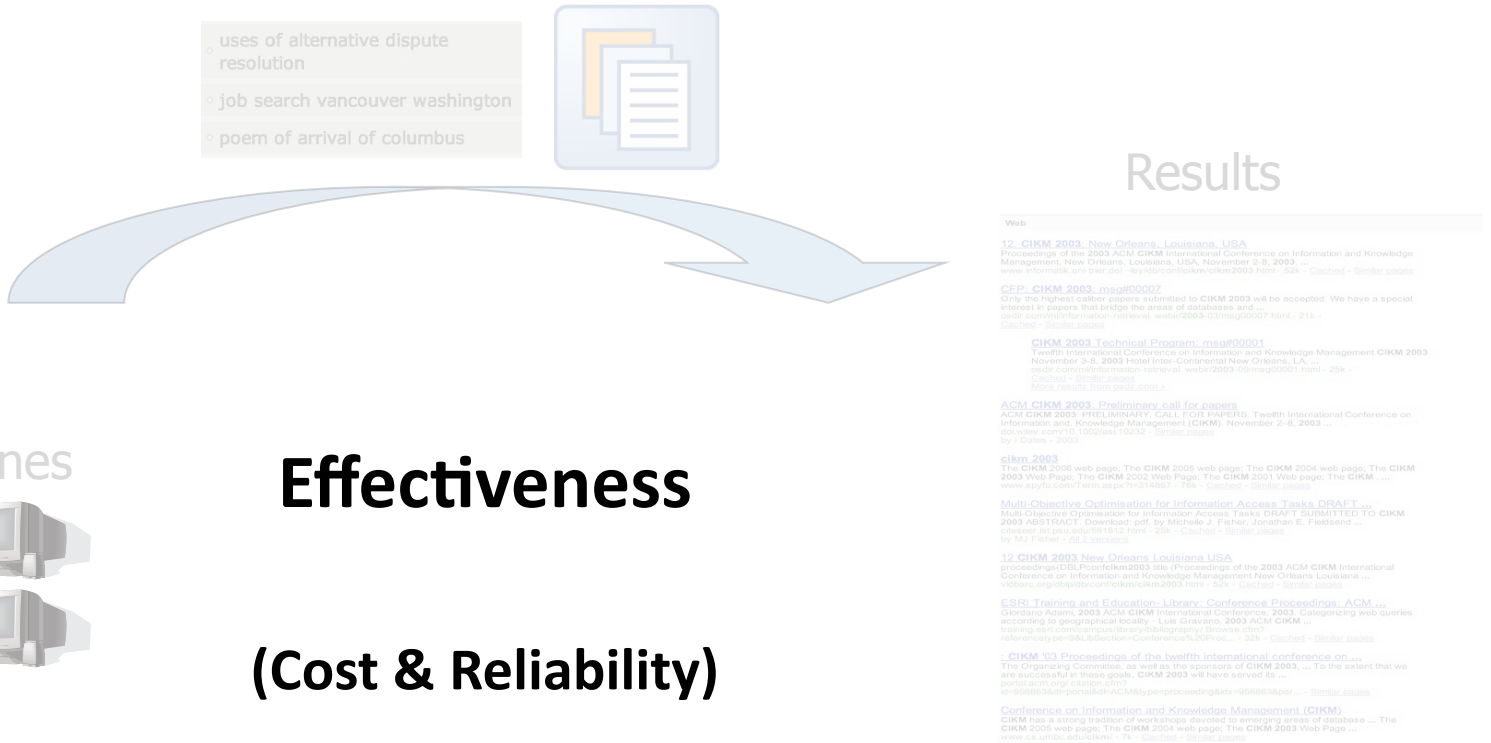




Judges



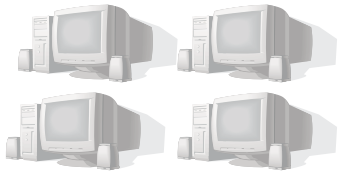
1 Evaluation Measures



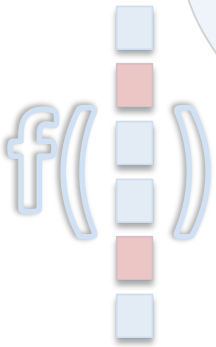
Judgment Effort

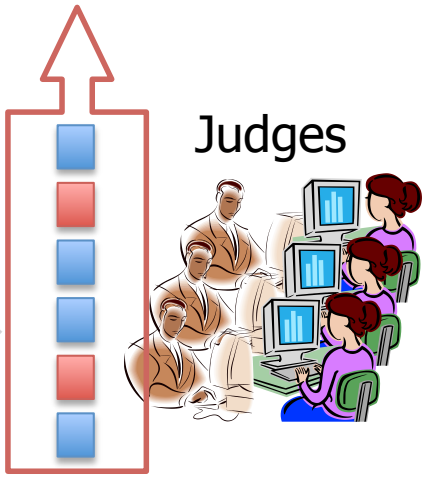
Results

Search Engines



Bottleneck
(human effort)





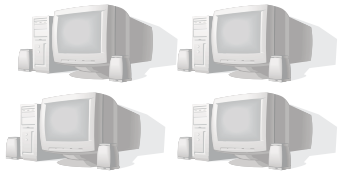
Judges

2

Judgment Effor

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

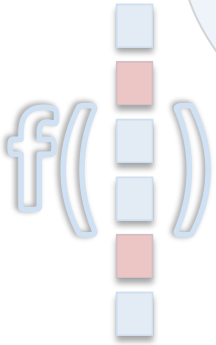
Search Engines



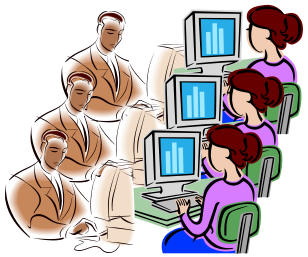
Cost

vs.

Reliability



Judges



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

Today's Outline

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

Evaluating early IR system

- Many early IR systems Boolean
  - Split collection in two: documents that
    - Match the query (Retrieved)
    - Don't match the query (Not retrieved)
  - Test collection: those documents that are
    - Relevant
    - Not Relevant

Measuring Boolean Output

- The goal of a retrieval system is

  A. To retrieve relevant documents

  B. Not to retrieve non-relevant documents

Measuring Boolean Output

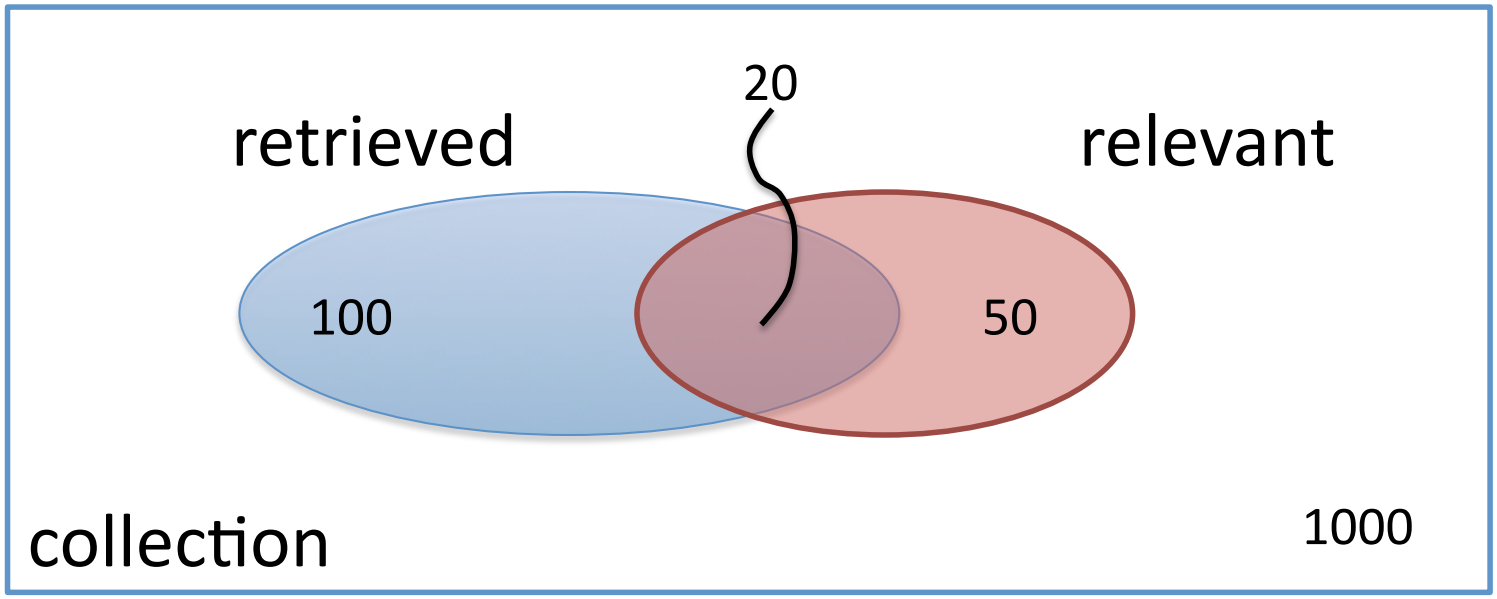
- Contingency table

	Relevant	Not-relevant	
Retrieved	a	b	a+b
Not retrieved	c	d	c+d
	a+c	b+d	a+b+c+d

$$\text{Precision} = \frac{a}{a+b} \qquad \text{Recall} = \frac{a}{a+c}$$

$$\text{Fallout} = \frac{b}{b+d}$$

Measuring Boolean Output



Precision = 20/100 = 0.2

Recall = 20/50 = 0.4

Fallout = (100-20)/(1000-50) = 0.08

Precision, Recall, Fall-out

- Contingency table

	Relevant	Not-relevant	
Retrieved	a	b	a+b
Not retrieved	c	d	c+d
	a+c	b+d	a+b+c+d = C

size of the collection

γ : generality

$$a + c = \gamma C$$

$$b + d = (1 - \gamma) C$$

$$\text{Precision} = \frac{\text{Recall}}{\text{Recall} + \left(\frac{1 - \gamma}{\gamma}\right) \text{Fallout}}$$

Summarising the two

- Isn't one measure better than two?
  - Van Rijsbergen's F: weighted harmonic mean

$$F = \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}$$

Summarising the two

- Isn't one measure better than two?
  - Van Rijsbergen's F: weighted harmonic mean
  - With $\alpha=1/(\beta^2+1)$

$$F_\beta = \frac{\left(\beta^2+1\right)PR}{R+\beta^2 P}$$

Evaluating best match systems

- Modern IR systems Best Match
  - Return a ranking of the collection instead of a set of documents

- A good system returns relevant documents before non-relevant

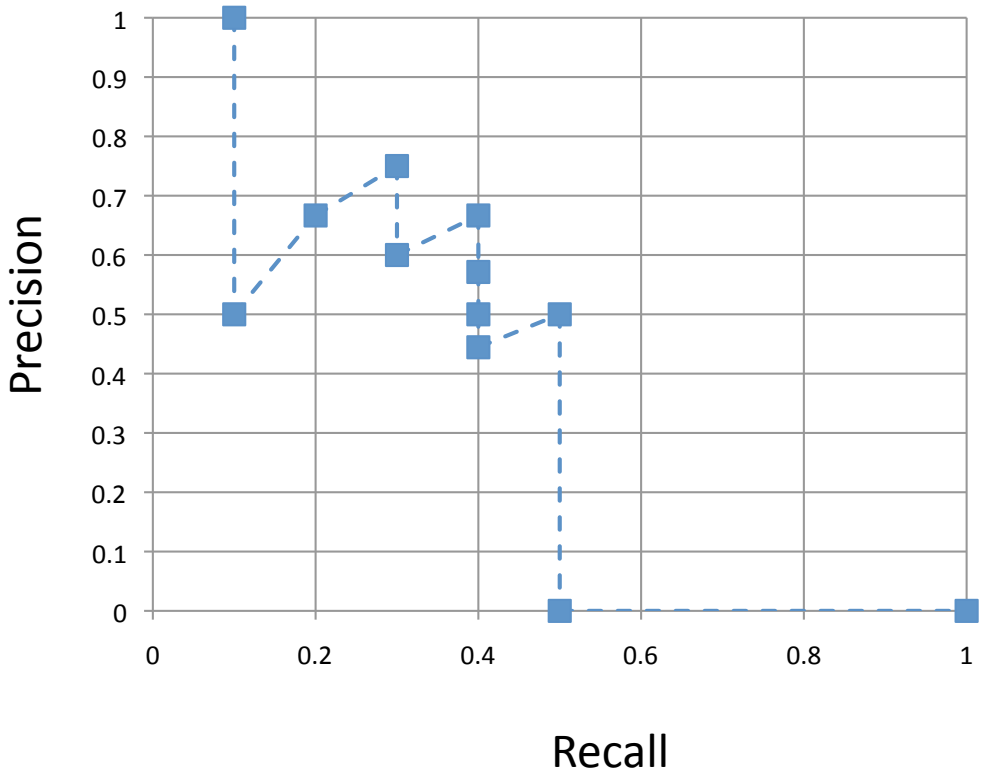
Ranking Measures

Topic 1

Rank	Rel.	
1	R	retrieved
2	N	
3	R	not retrieved
4	R	
5	N	
6	R	
7	N	
8	N	
9	N	
10	R	
…	…	
∞	R	

cut-off

Let R=10

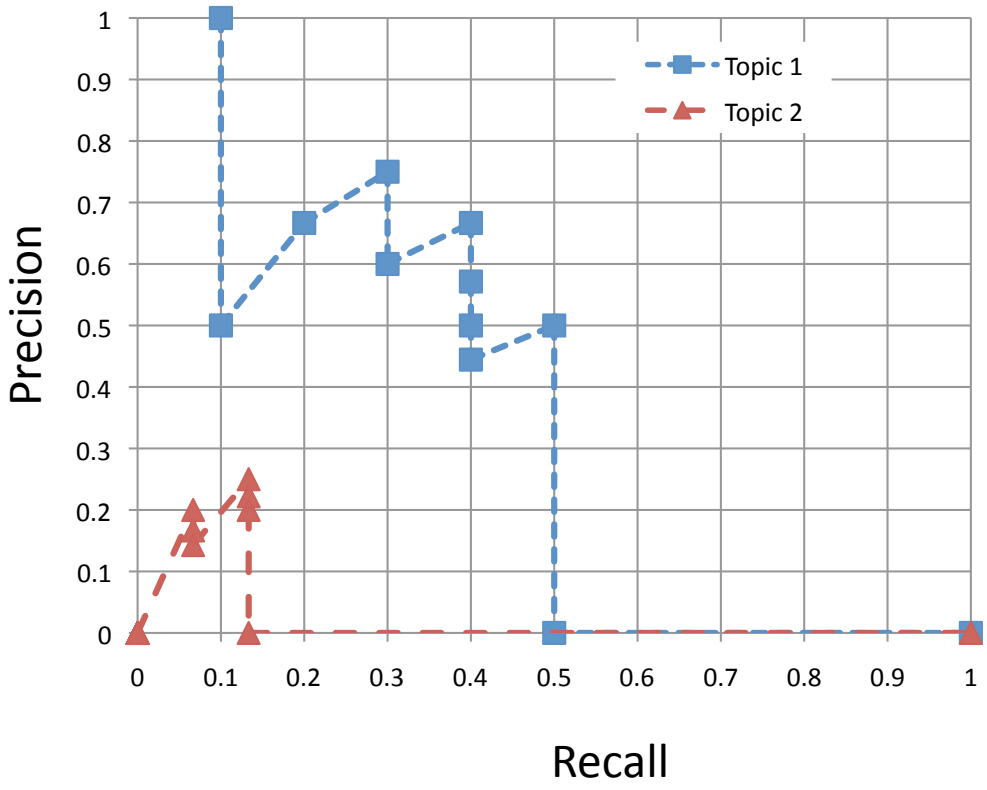


Ranking Measures

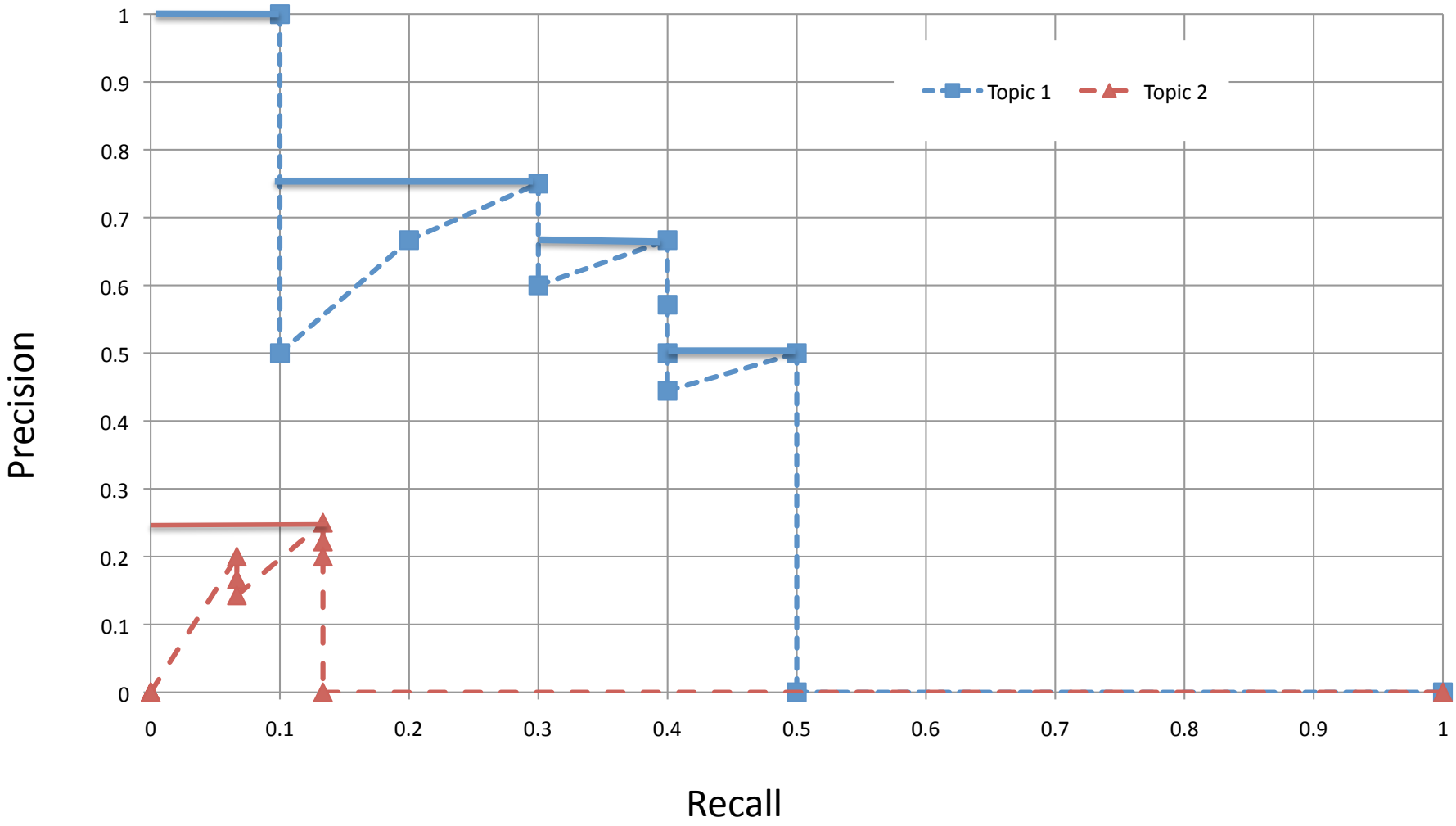
Topic 2

Rank	Rel.	Precision	Recall
1	N	0	0
2	N		
3	N	…	…
4	N		
5	R		
6	N	1/6	4/5
7	N		
8	R	…	…
9	N		
10	N	2/10	2/5
…	…	…	…
∞	R	0	10/10

Let R=15

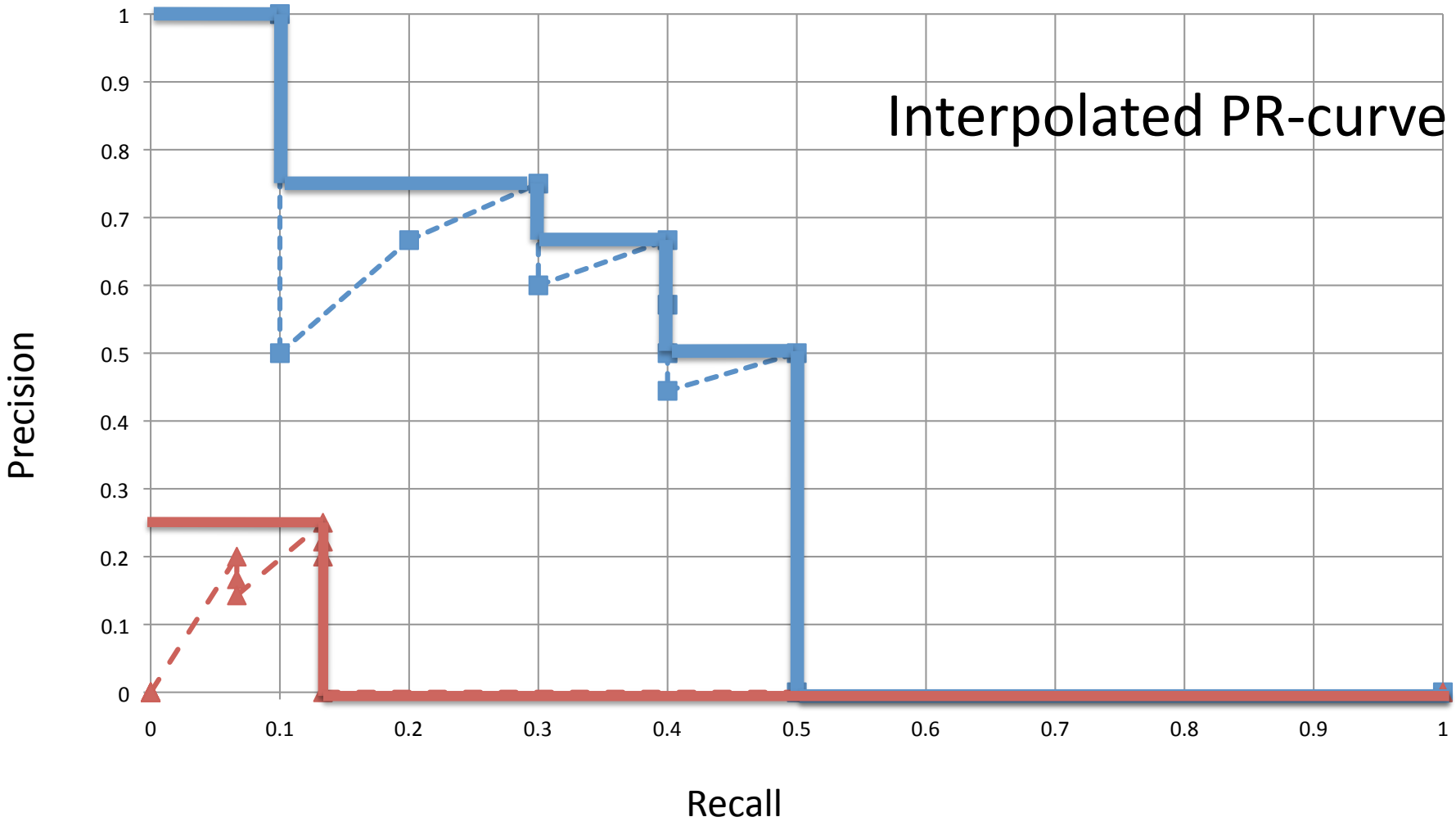


Recall/Precision Graph

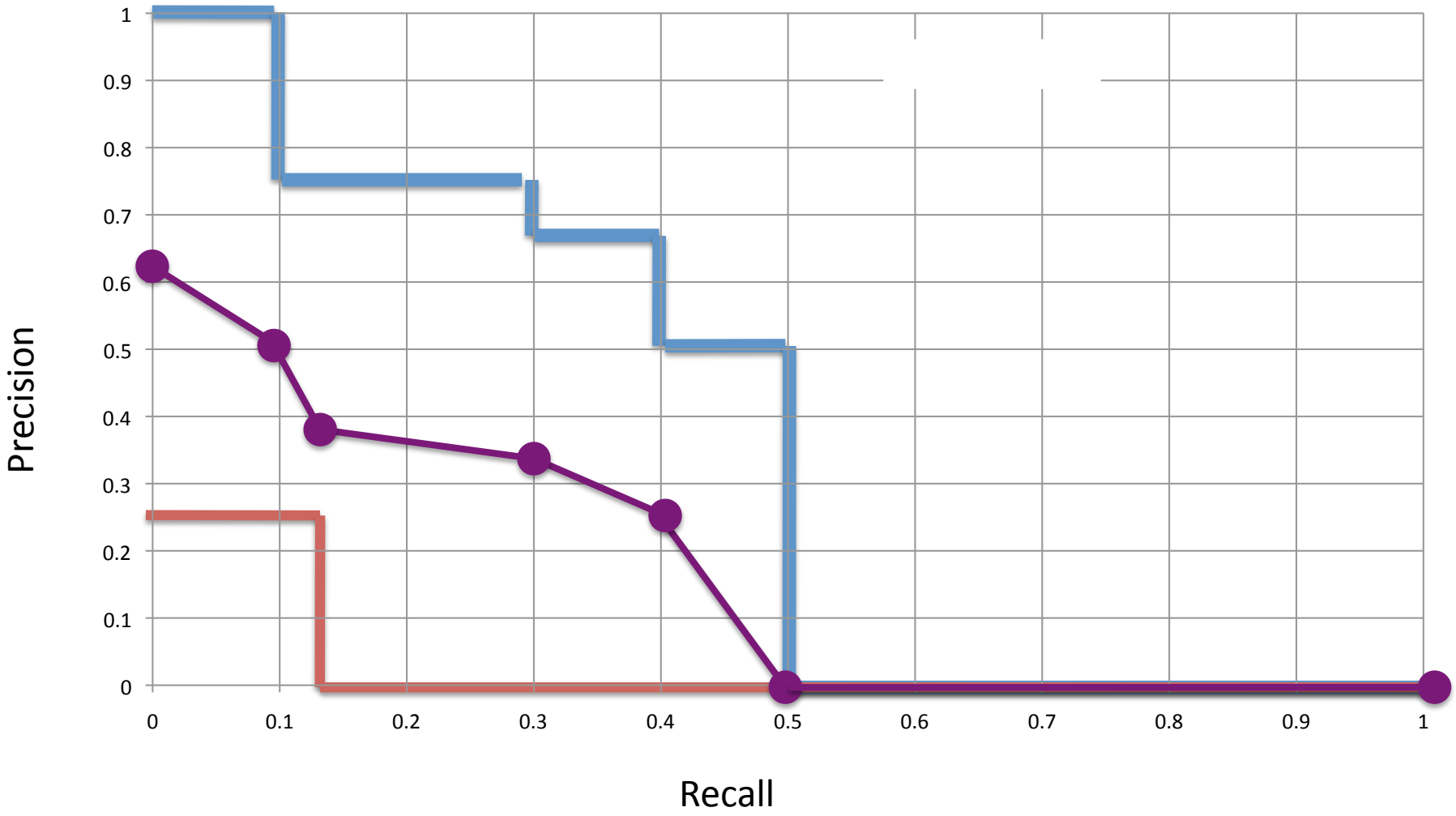


Average Graph

Recall/Precision Graph



Recall/Precision Graph



Average Precision

Topic 1

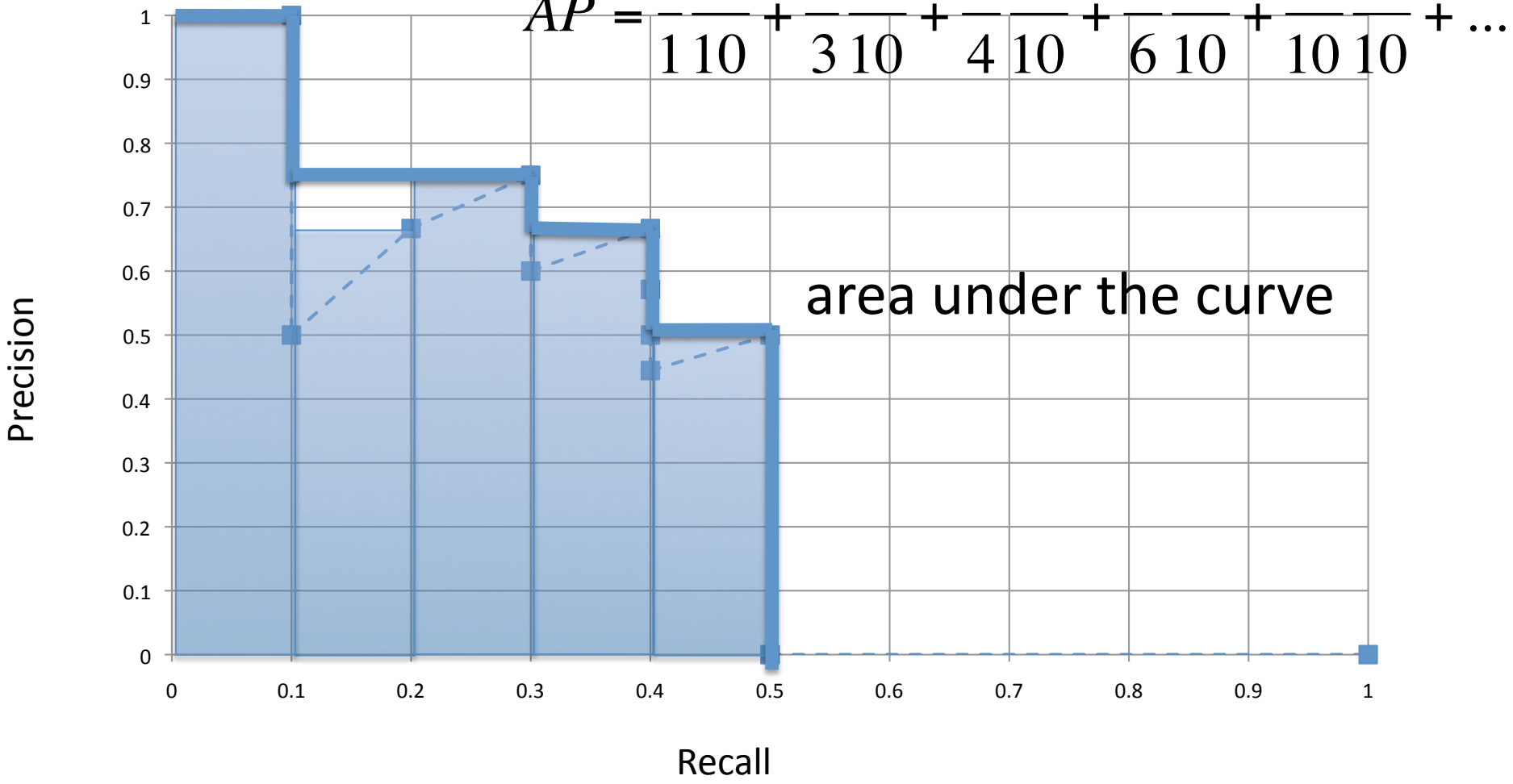
Rank	Rel.	Precision	Recall
1	R	1/1	1/10
2	N	1/2	1/10
3	R	2/3	2/10
4	R	3/4	3/10
5	N	…	…
6	R	4/6	4/10
7	N		
8	N	…	…
9	N		
10	R	5/10	5/10
…	…	…	…
∞	R	0	10/10

- Average Precision
  - Average of precisions at relevant documents

$$AP = \frac{\frac{1}{1} + \frac{2}{3} + \frac{3}{4} + \frac{4}{6} + \frac{5}{10} + \ldots}{10}$$

Average Precision

$$AP = \frac{1}{1}\frac{1}{10} + \frac{2}{3}\frac{1}{10} + \frac{3}{4}\frac{1}{10} + \frac{4}{6}\frac{1}{10} + \frac{5}{10}\frac{1}{10} + \ldots$$



R-precision

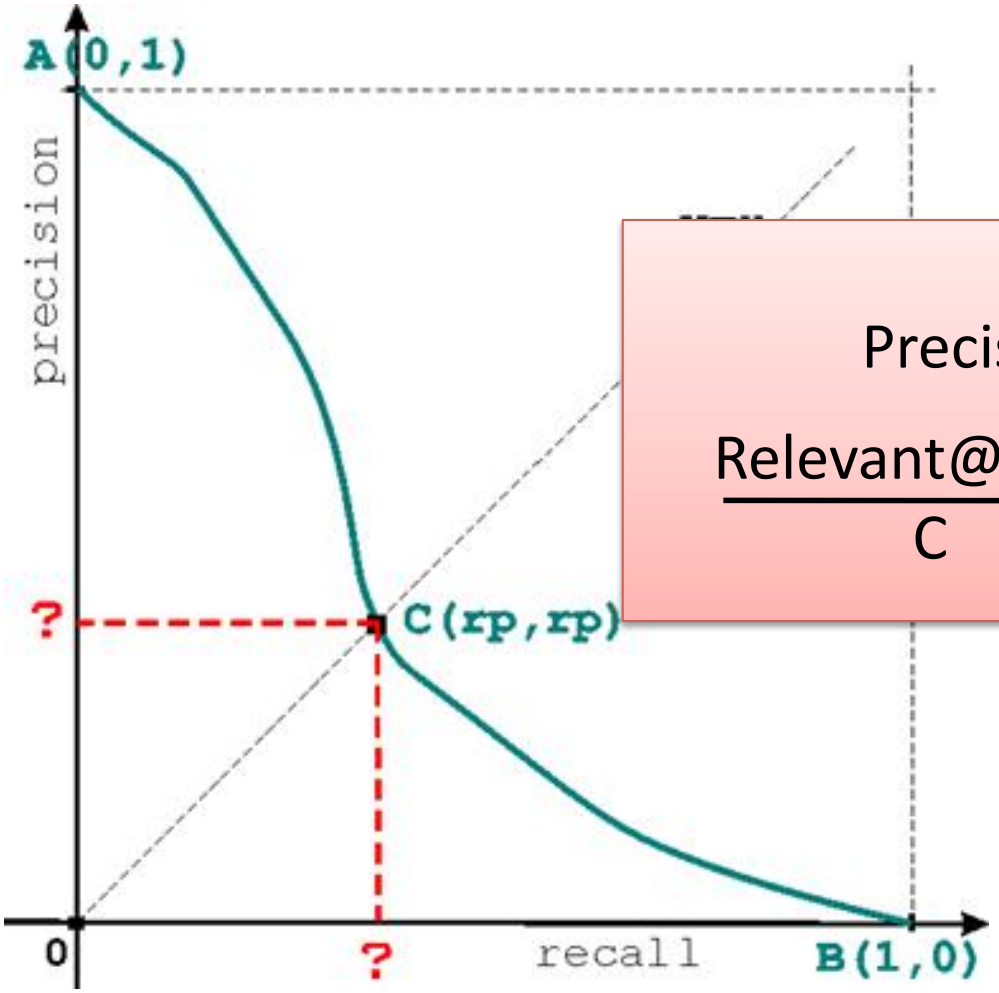
Topic 1

Rank	Rel.	Precision	Recall
1	R	1/1	1/10
2	N	1/2	1/10
3	R	2/3	2/10
4	R	3/4	3/10
5	N	…	…
6	R	4/6	4/10
7	N		
8	N	…	…
9	N		
10	R	5/10	5/10
…	…	…	…
∞	R	0	10/10

Let R=10

R-precision =
Precision@R = 5/10

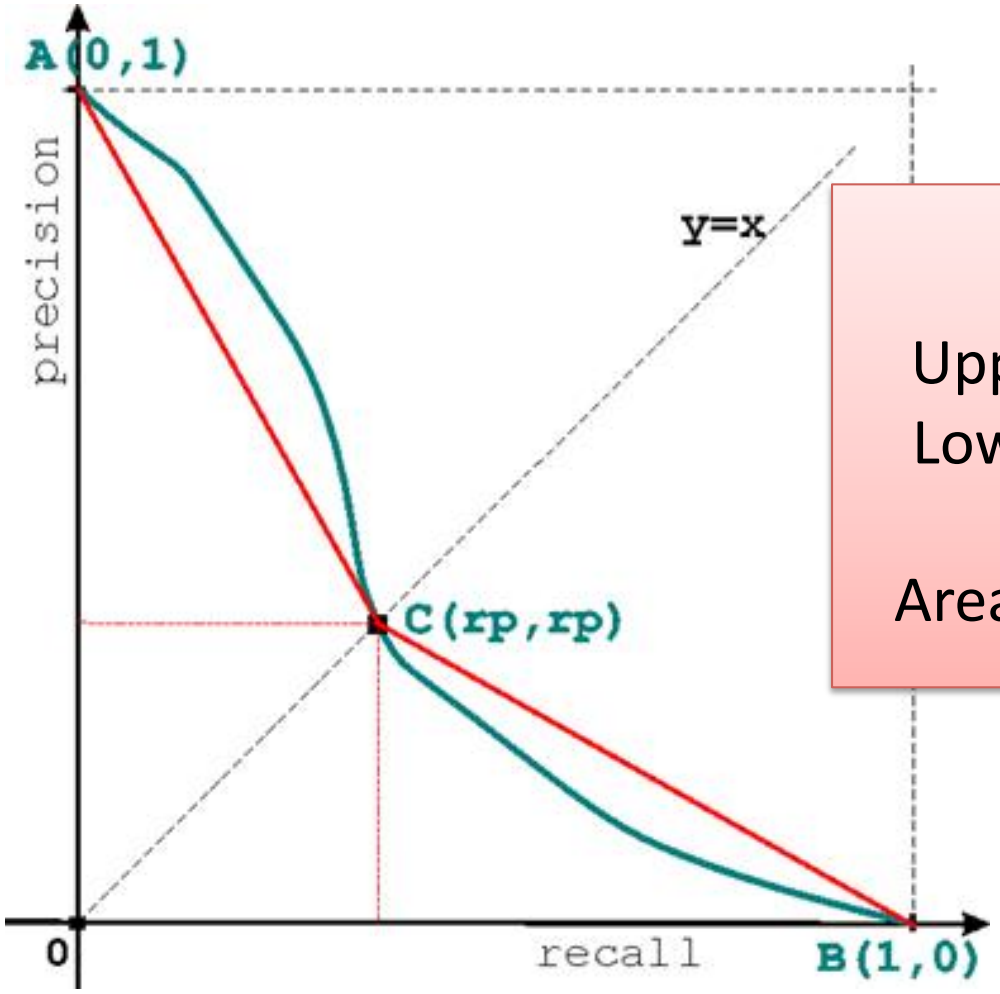
Recall/Precision



Precision@C = Recall@C

$$\frac{\text{Relevant@[1..C]}}{C} = \frac{\text{Relevant@[1..C]}}{R}$$

Recall/Precision



Square : rp × rp
Upper Triangle : (1-rp) × rp / 2
Lower Triangle : rp × (1-rp) / 2

Area = rp × rp + rp × (1-rp) = rp

Focusing on the top of the ranking

- Most users focus on the top ranked documents
- Many search tasks have a unique relevant document
  - e.g. home-page finding
- Recall-based measures are not appropriate

- Reciprocal Rank (RR) =

  1/rank of the (first) relevant document

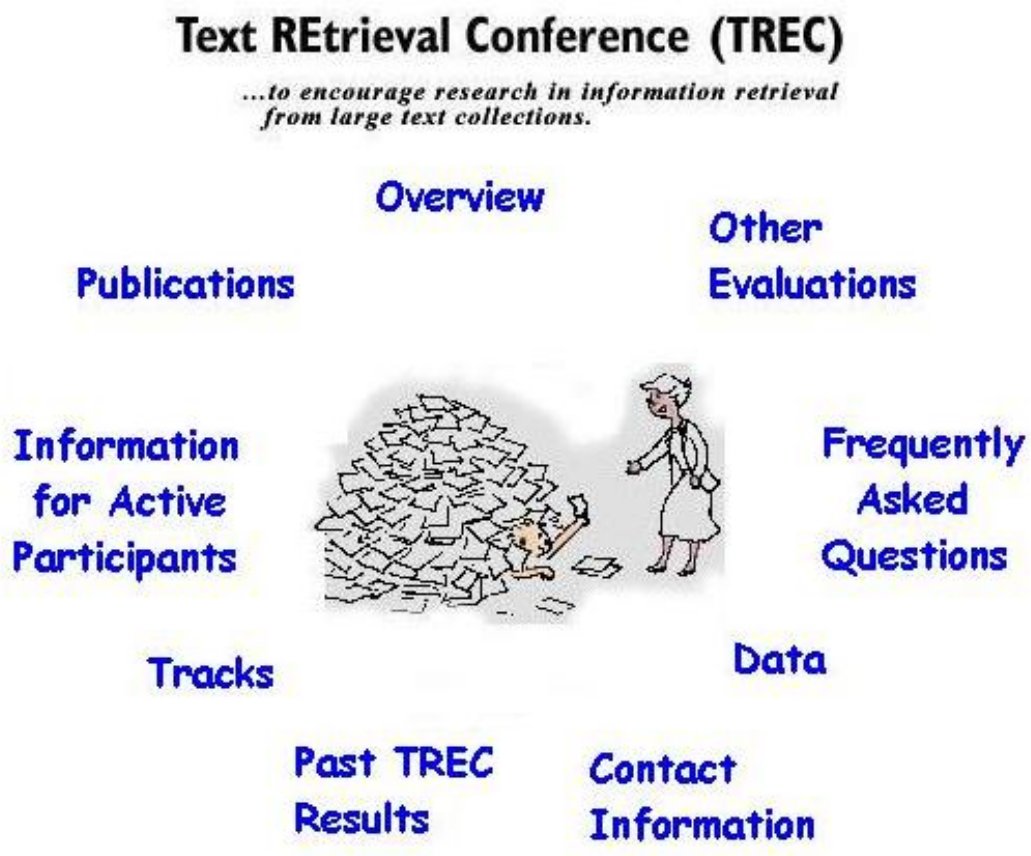
Evaluating Retrieval Systems

- Evaluation measures
  - User-oriented
    - Underlying user model to simulate user's interaction
    - **Precision@k**, nDCG@k, **MRR**, RBP,

      ERR [Chapelle et al CIKM09], EBU [Yilmaz et al CIKM10]…
  - System-oriented
    - Also often based on a user model [Robertson SIGIR08]
    - Capture the overall effectiveness of retrieval system
    - **Average Precision, R-Precision**, nDCG, normalized Recall, PRES [Magdy and Jones SIGIR10], GAP [Robertson et al SIGIR10], …

TREC 1992

- create test collections for a set of retrieval tasks
- standardize evaluation measures



http://trec.nist.gov/images/paper_3.jpg

trec_eval

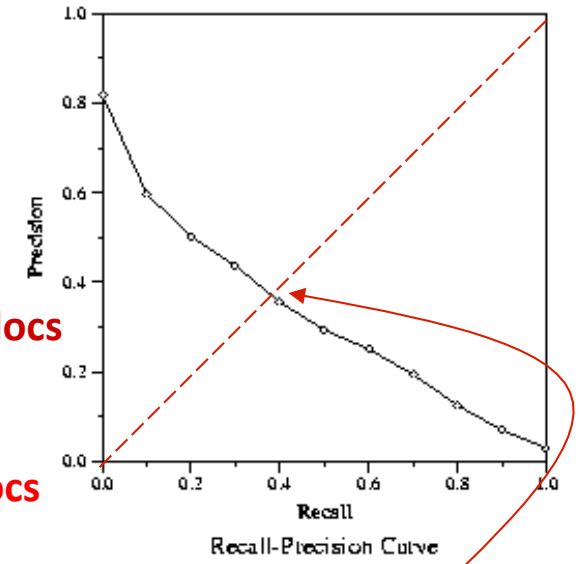
- TREC standardised evaluation code
  - http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz
- Given output from IR system searching over test collection
  - Produces all measures (and many more)
  - Most researchers use trec_eval to save time, avoid introducing bugs.
    - Look at some output…

Summary Statistics	
Run Number	ok8amxc
Run Description	Automatic, title + desc
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel-ret:	3212

Out of 4728 rel docs, we've got 3212

Recall=3212/4728

Precion-Recall Curve



Recall-Precision Curve

Recall Level Precision Averages	
Recall	Precision
0.00	0.8190
0.10	0.5975
0.20	0.5032
0.30	0.4372
0.40	0.3561
0.50	0.2936
0.60	0.2511
0.70	0.1941
0.80	0.1257
0.90	0.0696
1.00	0.0296
Average precision over all relevant docs	
non-interpolated	0.3169

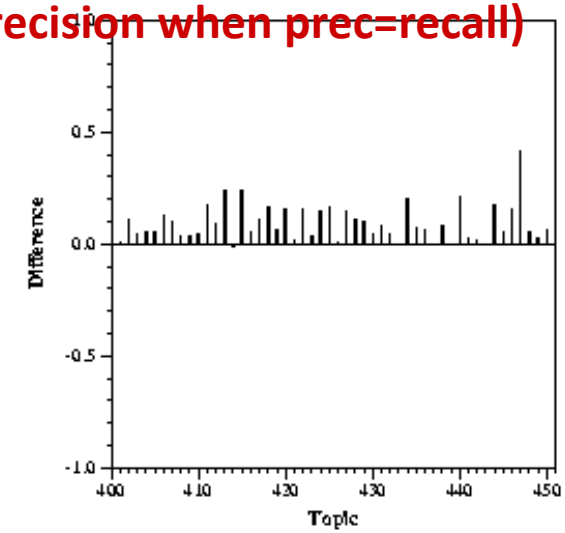
Document Level Averages	
	Precision
At 5 docs	0.5800
At 10 docs	0.5500
At 15 docs	0.4987
At 20 docs	0.4650
At 30 docs	0.4253
At 100 docs	0.2680
At 200 docs	0.1921
At 500 docs	0.1085
At 1000 docs	0.0642
R-Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.3470

Precision@10docs

about 5.5 docs in the top 10 docs are relevant

Breakeven Precision (precision when prec=recall)

Mean Avg. Precision (MAP)



Difference from Median in Average Precision per Topic

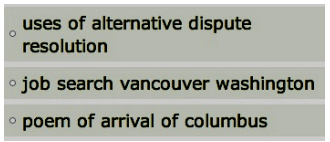
Today's Outline

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

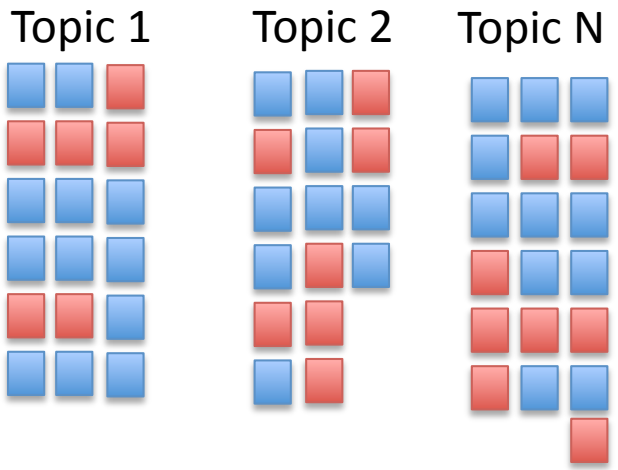
Test Collections

- Document Corpus

- Topics



- Relevance Judgments (QRELs)



Test Collections Simulate

Retrieval Scenarios

Test Collection Design

- What corpus?
- What queries?
- What judgments?
  - How many documents per topic?
  - How many assessors per document?

Early Test Collections

Name	Docs.	Qrys	Year	Size, Mb	Source document
Cranfield 2	1,400	225	1962	1.6	Title, authors, source, abstract of scientific papers from the aeronautic research field, largely ranging from 1945-1963.
ADI	82	35	1968	0.04	A set of short papers from the 1963 Annual Meeting of the American Documentation Institute.
IRE-3	780	34	1968	-	A set of abstracts of computer science documents, published in 1959-1961.
NPL	11,571	93	1970	3.1	Title, abstract of journal papers
MEDLARS	450	29	1973	-	The first page of a set of MEDLARS documents copied at the National Library of Medicine.
Time	425	83	1973	1.5	Full text articles from the 1963 edition of Time magazine.

http://ir.dcs.gla.ac.uk/resources/test_collections/

Problem with Scale

Name	Docs.	Qrys	Year	Size, Mb	Source document
Cranfield 2	1,400	225	1962	1.6	Title, authors, source, abstract of scientific papers from the aerodynamic research field, largely ranging from 1945-1963.
ADI	82	35	1968	0.04	A set of short papers from the 1963 Annual Meeting of the American Documentation Institute.
IRE-3	780	34	1968	-	A set of abstracts of computer science documents, published in 1959-1961.
NPL	11,571	93	1970	3.1	Title, abstract of journal papers
MEDLARS	450	29	1973	-	The first page of a set of MEDLARS documents copied at the National Library of Medicine.
Time	425	83	1973	1.5	Full text articles from the 1963 edition of Time magazine.

By mid 1970s, commercial IR systems searched hundreds of thousands of documents

Slightly bigger collections

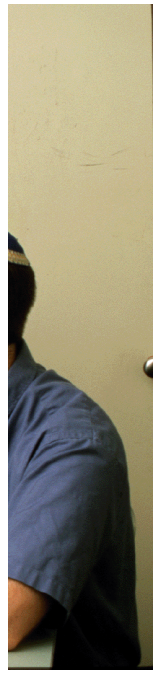
Name	Docs.	Qrys.	Year	Size, Mb	Source document
INSPEC	12,684	77	1981	-	Title, author, source, abstract and indexing information from Sep-Dec 1979 issues of Computer and Control Abstracts.
CACM	3,204	64	1983	2.2	Title, abstract, author, keywords and bibliographic information from articles of Communications of the ACM.
CISI	1,460	112	1983	2.2	Author, title/abstract, and co-citation data for the 1460 most highly cited articles and manuscripts in information science, 1969-1977.
LISA	6,004	35	1983	3.4	Taken from the Library and Information Science Abstracts database.

By 1990s, commercial IR systems searched millions of documents

Individual group

- Weren't able to produce test col
  sufficient scale
- Someone needed to coalesce th
  community
  - TREC
    - Donna Harman

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



http://www.itl.nist.gov/iad/photos/trec2001.gif

Early TREC collections

- Largely articles (news, journals, government)
- Long time to try web search
  - Assumption web wasn't different
    - Very wrong
  - Fixed now

Collection	Number of documents	Size	Average number of words/doc.
CACM	3,204	2.2 Mb	64
AP	242,918	0.7 Gb	474
GOV2	25,205,179	426 Gb	1073

Recent TREC collections

- ClueWeb09 collection
  - about 1 billion web pages in ten languages
  - 5 TB, compressed (25 TB, uncompressed)
  - collected by CMU in January and February 2009
- Other recent TREC collections
  - Collections from wide range of sources
    - Blogs, Twitter, Legal documents, Patents, …
- TREC model copied by others
  - CLEF, INEX, NTCIR, …

http://lemurproject.org/clueweb09.php/

Test Collection Design

- What corpus?
- What queries?
- What judgments?
  - How many documents per topic?
  - How many assessors per document?

TREC topics

- Early TREC collections
  - Searched collection for potential topics

TREC topics

```
<top>

<num> Number:  200

<title> Topic: Impact of foreign textile imports on U.S. textile industry

<desc> Description:  Document must report on how the importation of foreign
textiles or textile products has influenced or impacted on the U.S. textile
industry.

<narr>  Narrative:  The impact can be positive or negative or qualitative.
It may include the expansion or shrinkage of markets or manufacturing volume
or an influence on the methods or strategies of the U.S. textile industry.
"Textile industry" includes the production or purchase of raw materials;
basic processing techniques such as dyeing, spinning, knitting, or weaving;
the manufacture and marketing of finished goods; and also research in the
textile field.

</top>
```

TREC topics

- In the past : Some criticism
  - Not representative
  - No association with topic characteristics (?)
    - Topics with too few or too many relevant documents were avoided [Harman NIST93, Voorhees and Harman NIST99]
    - "Candidate topics were also rejected if they seemed ambiguous" [Voorhees and Harman NIST96]
  - 50~100 topics

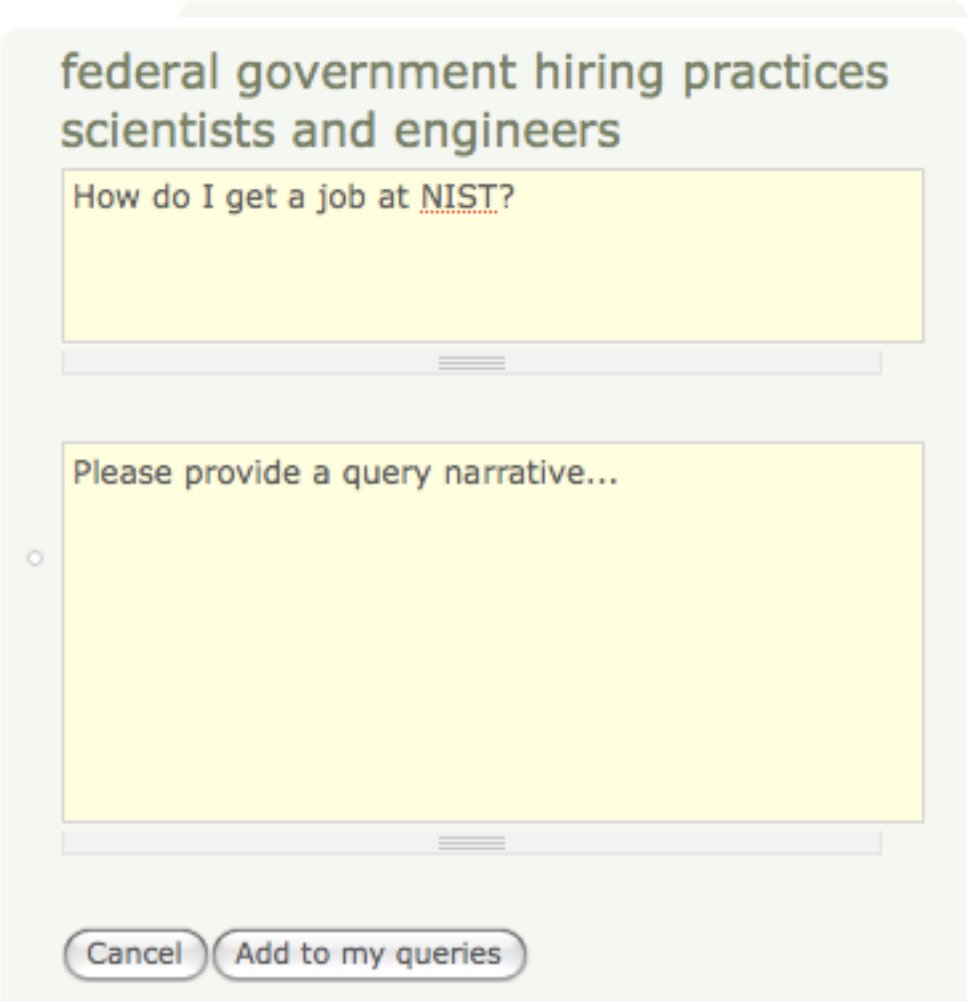
- Recently : Queries taken from query logs
  - "Torso queries" – neither too rare nor too popular
  - More representative
  - 50,000 queries in 2009 Million Query Track

TREC topics

- In early TREC collections
  - The creators of the topics (queries) were the judges of the documents
- How about queries taken from query logs?

Million Query Track

What is the user looking for ?



Test Collection Design

- What corpus?
- What queries?
- What judgments?
  - How many documents per topic?
  - How many assessors per document?

TREC relevance

- In early TREC collections documents judged either
  - Relevant
    - Even if just a single sentence was relevant
  - Not relevant
- In later collections (TREC 9 Web and on) ternary judgments
  - highly relevant / relevant / non-relevant

Web 2010 track

1. *Nav*

   This page represents a home page of an entity directly named by the query; the user may be searching for this specific page or site.

2. *Key*

   This page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine.

3. *Hrel*

   The content of this page provides substantial information on the topic.

http://plg.uwaterloo.ca/~trecweb/2010.html

Web 2010 track

4. *Rel*

   The content of this page provides some information on the topic, which may be minimal; the relevant information must be on that page, not just promising-looking anchor text pointing to a possibly useful page.

5. *Non*

   The content of this page does not provide useful information on the topic, but may provide useful information on other topics, including other interpretations of the same query.

6. *Junk*

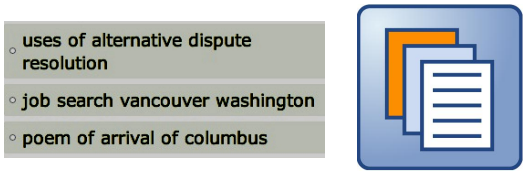
   This page does not appear to be useful for any reasonable purpose; it may be spam or junk.

http://plg.uwaterloo.ca/~trecweb/2010.html

Test Collection Design

- What corpus?
- What queries?
- What judgments?
  - How many documents per topic?
  - How many assessors per document?

Judgment Effort



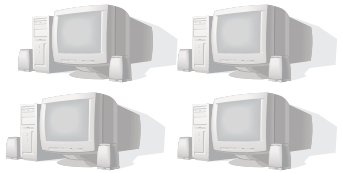
Results

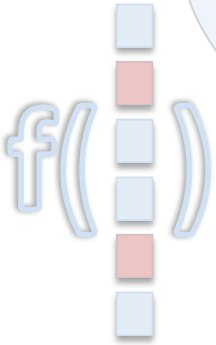
Cost

vs.

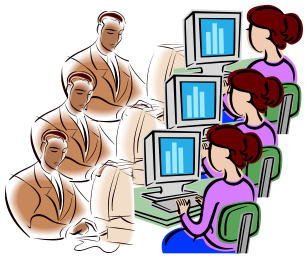
Reliability

Search Engines





Judges



Obtaining Relevance Judgments

- System-oriented measures have some recall component
  - AP defined on all relevant documents, optimal DCG requires optimal ranking of all relevant documents, etc.
  - All relevant documents in the collection need to be identified

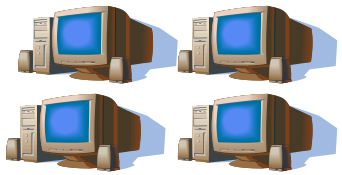
Obtaining Relevance Judgments

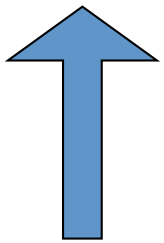
- User-oriented measures
  - Let's assume we are interested at Precision@10

Retrieved List by SYS1		Retrieved List by FUTURE 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	?

TREC Evaluation Setup

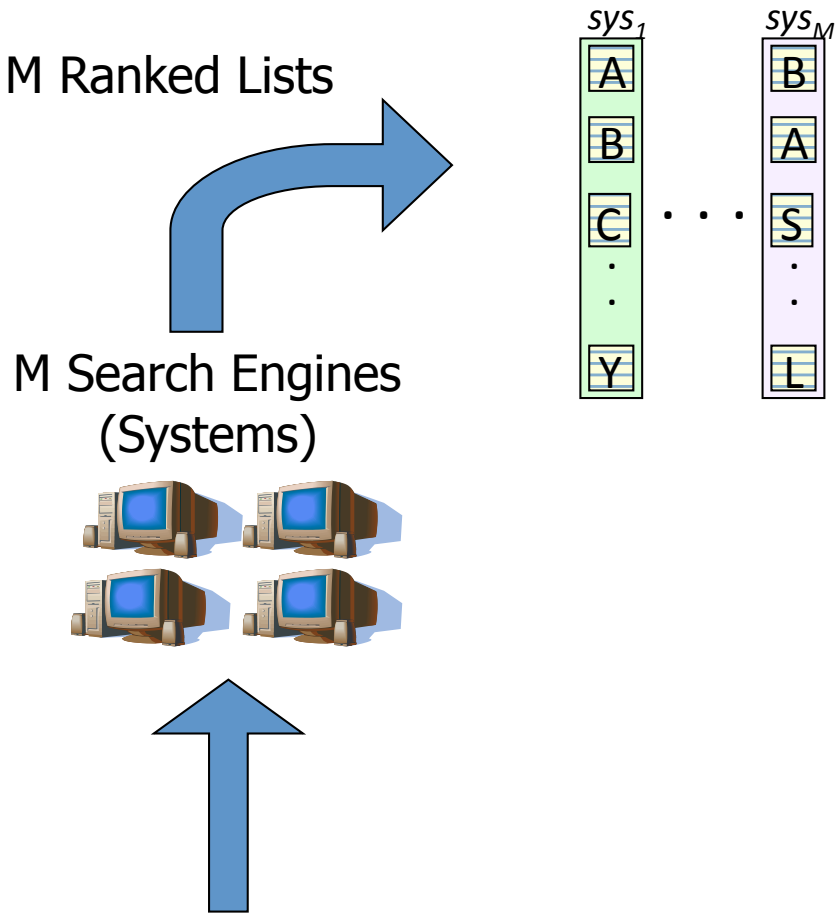
M Search Engines
(Systems)



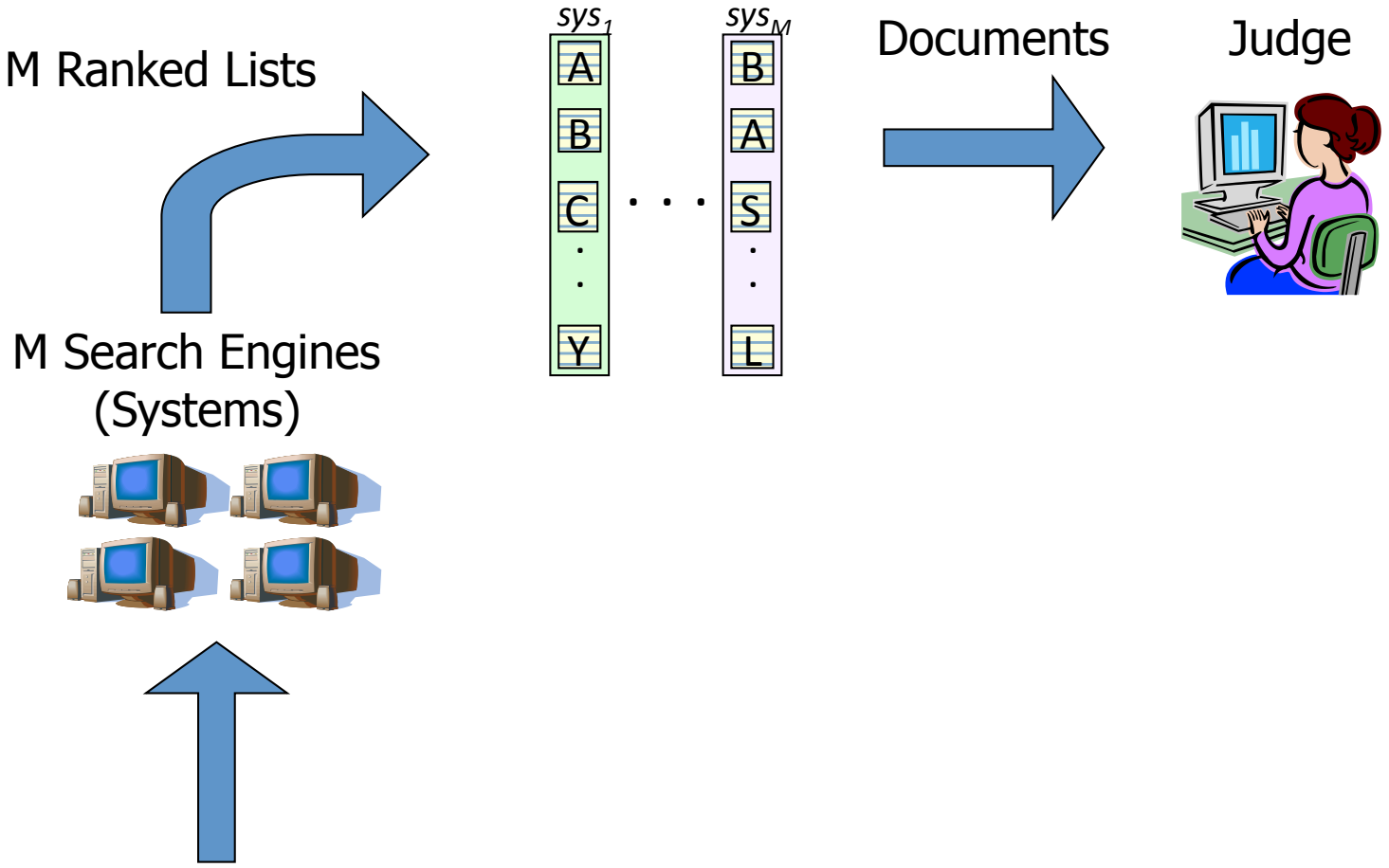


TREC Query (Q1)

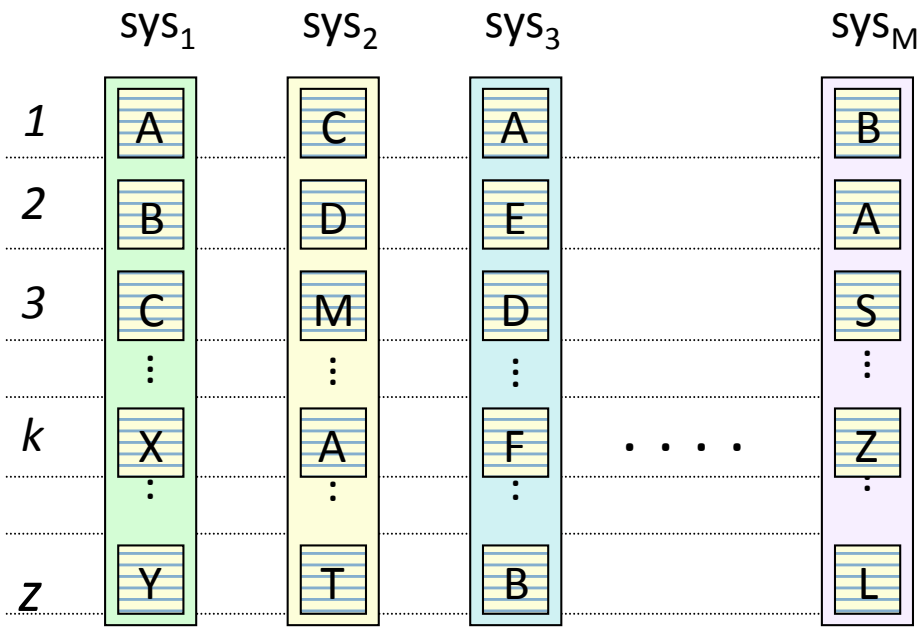
TREC Evaluation Setup

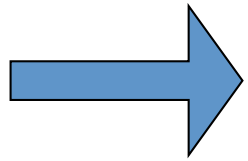


TREC Evaluation Setup

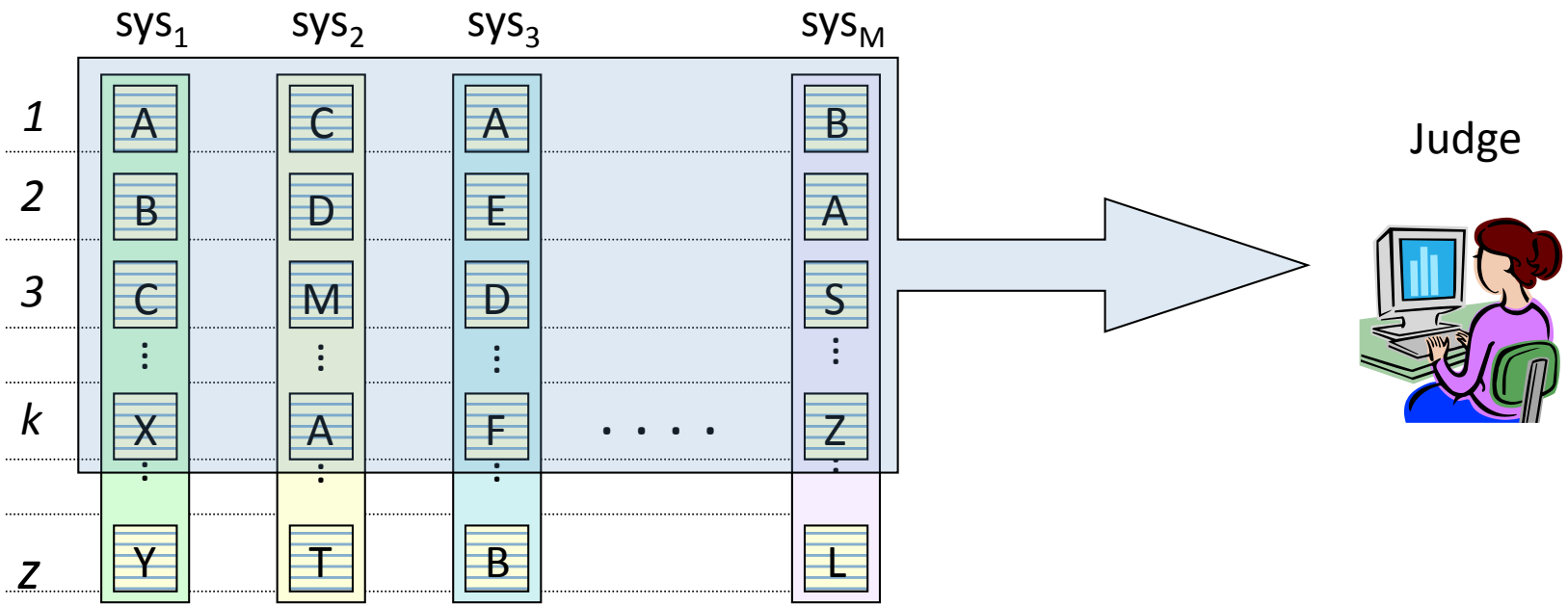


Depth-k Pooling

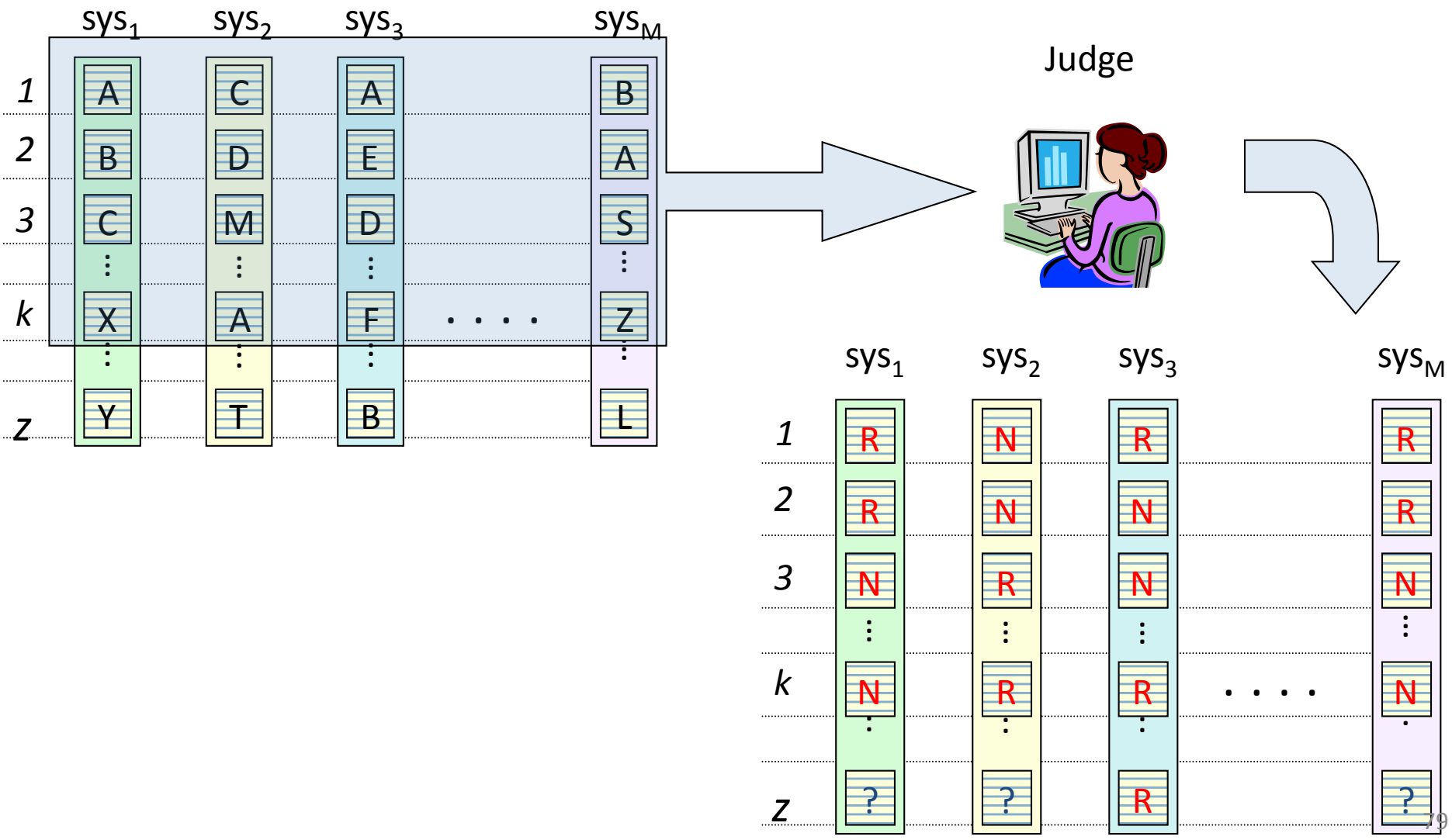




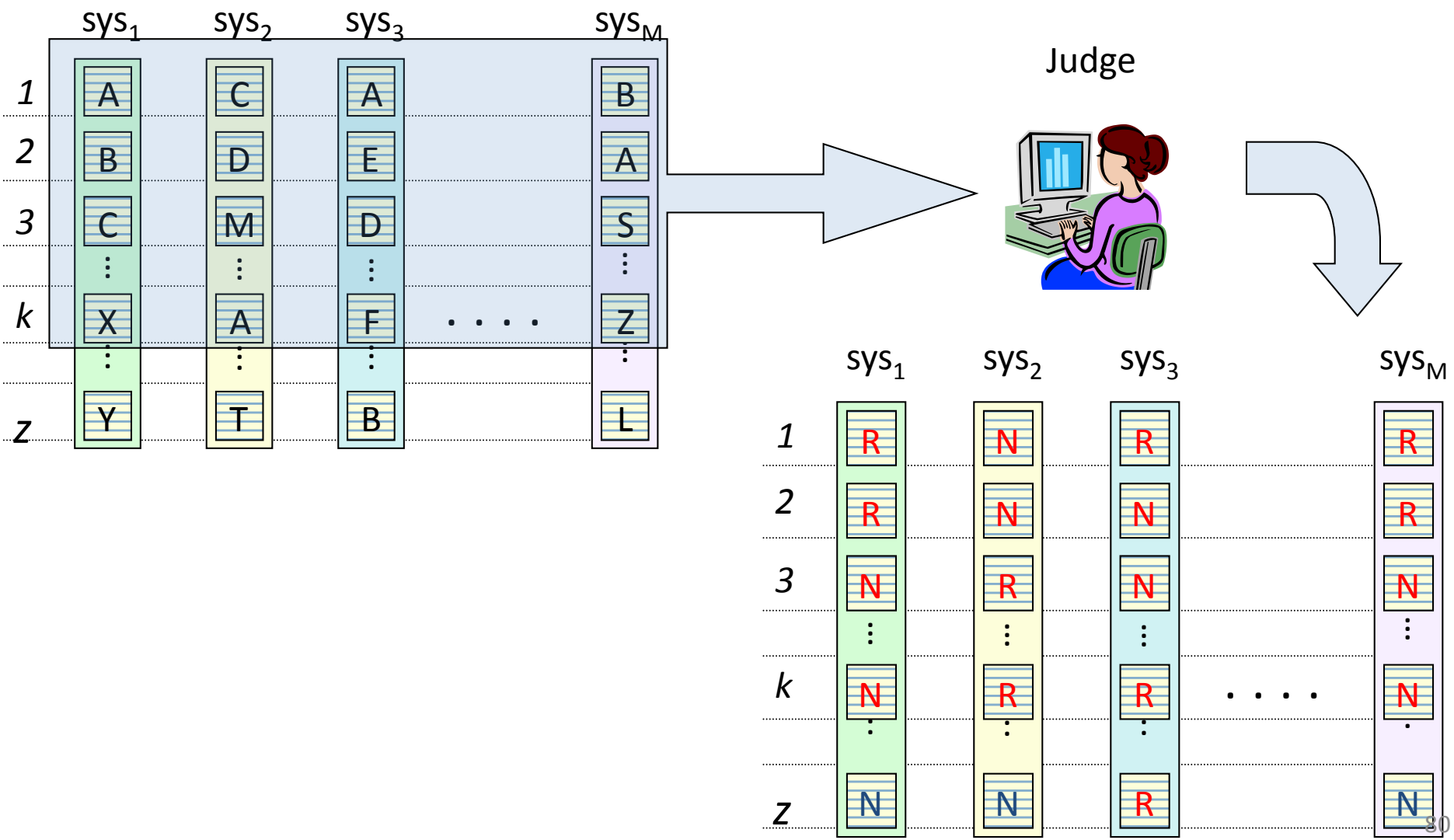
Depth-k Pooling



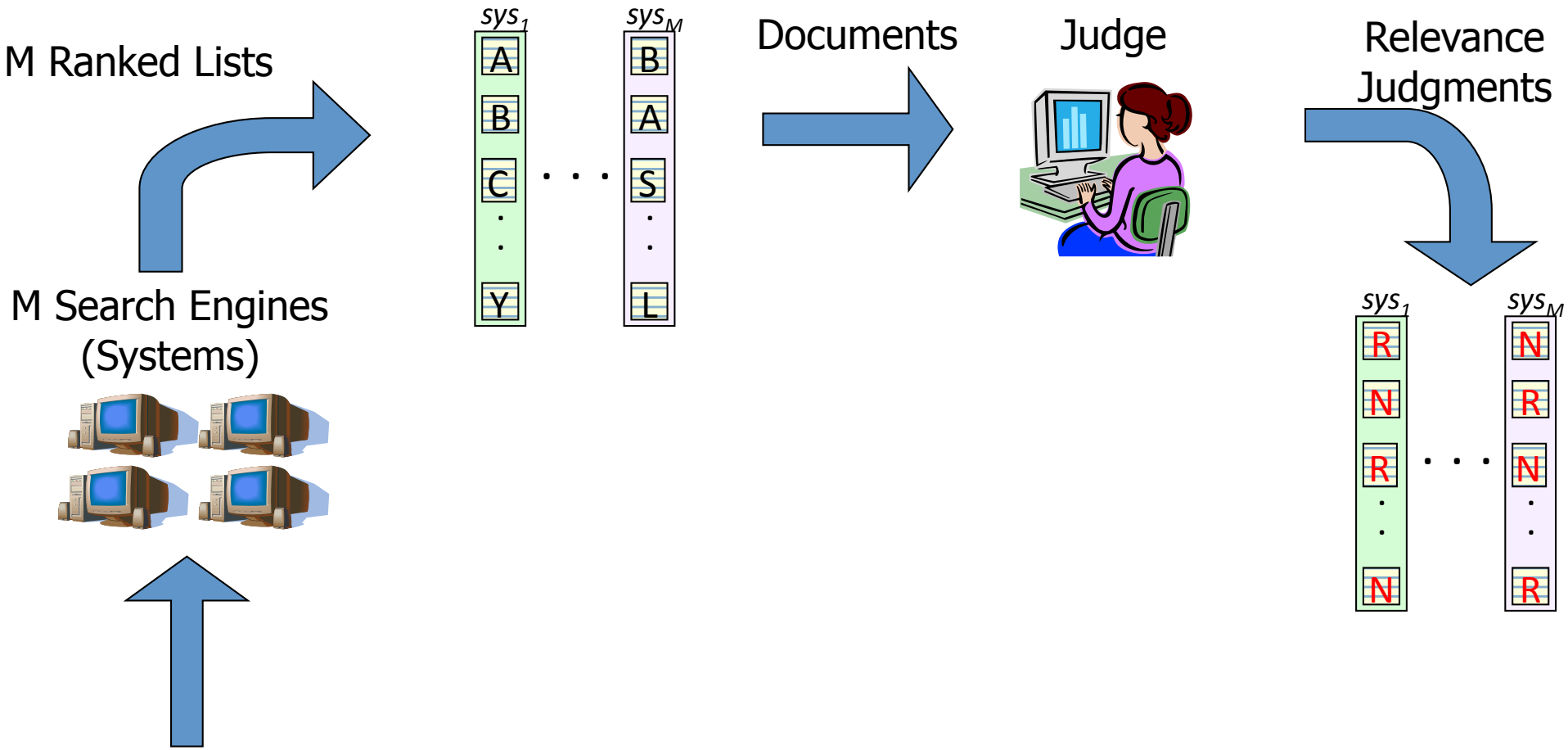
Depth-k Pooling



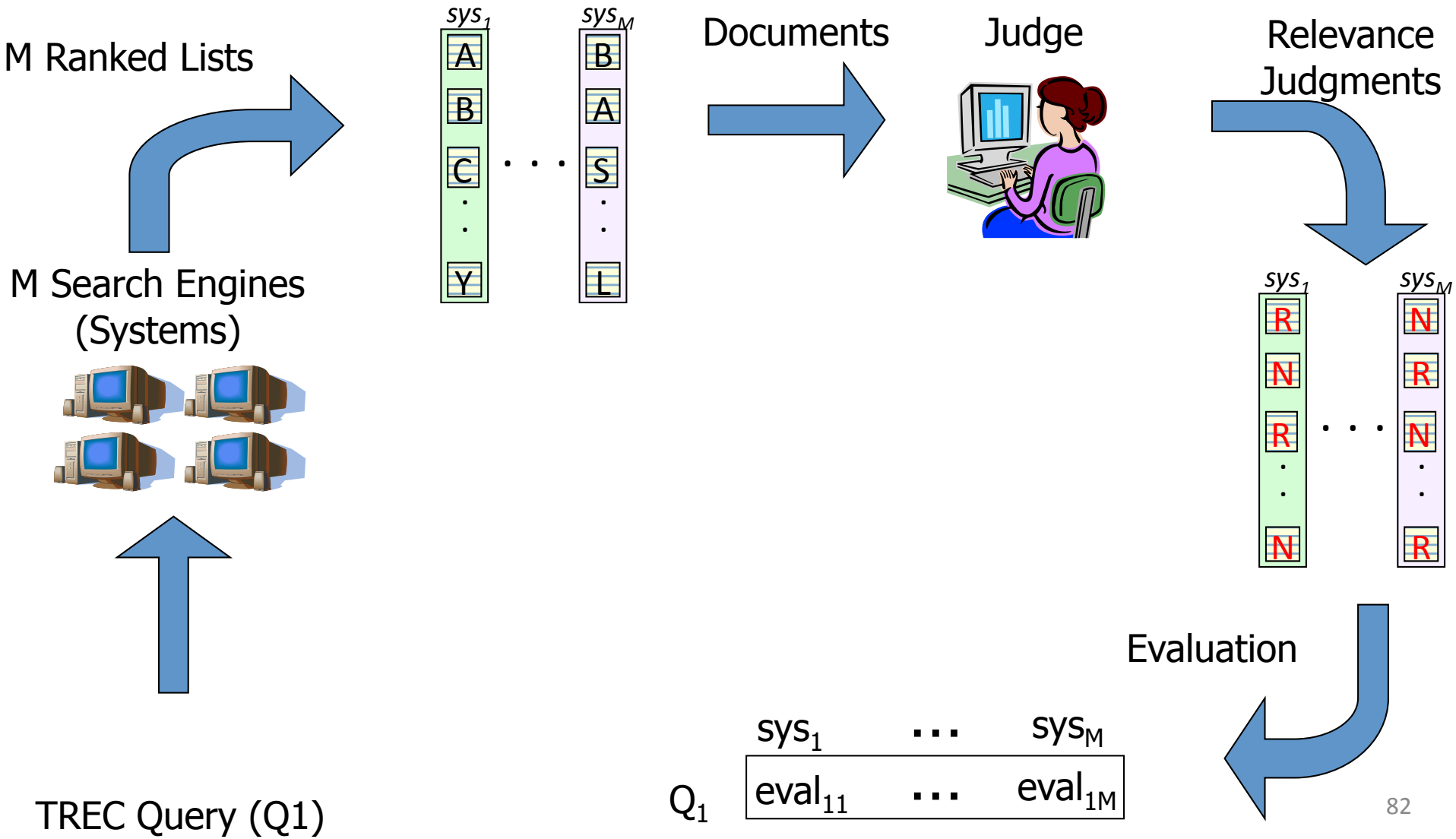
Depth-k Pooling



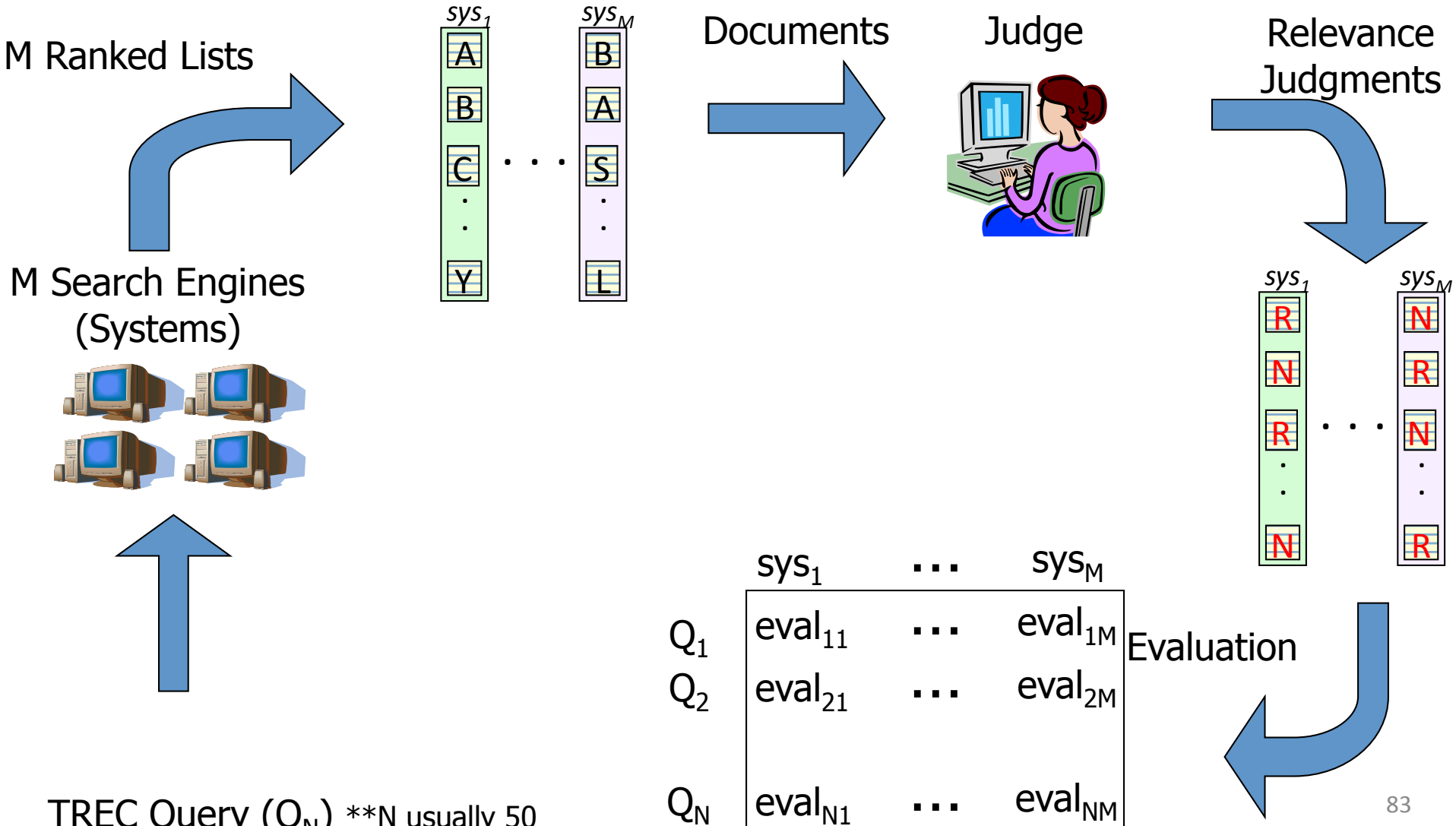
TREC Evaluation Setup



TREC Evaluation Setup



TREC Evaluation Setup



Depth-100 pooling

- Many missing relevant documents [Harman NIST96]
  - In small collection missing relevant docs do not affect the ranking of systems [Zobel SIGIR98]
  - In larger collections pooling is inadequate to identify "interesting" relevant docs [Buckley et al SIGIR06]
    - Most judged relevant docs simply contain query terms

- Still expensive
  - TREC 1 – 8 test collections [Voorhees and Harman NIST99]
    - 50 topics, depth-100 pooling => 73,000 judgments
    - 30 sec per judgment => 75 labor days
  - Preferably more than 50 queries (?)

Test Collection Design

- What corpus?
- What queries?
- What judgments?
  - How many documents per topic?
  - How many assessors per document?

Inconsistency among Assessors

- Evaluation robust to assessors' inconsistency
  - Variation in assessments although high did not affect the relative ranking of systems [Lesk and Salton JACM68, Cleverdon CLR70, Voorhees SIGIR98]
  - Consistency of judgments against the rank of documents [Lesk and Salton JACM68, Voorhees IPM00, Sanderson CIKM98]
- Evaluation not robust when assessors vary in task and topic expertise [Bailey et al. SIGIR08]

Today's Outline

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

TREC 1992

- create test collections for a set of retrieval tasks
- standardize evaluation measures

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

http://trec.nis _3.jpg

NIST Text REtrieval Conference (TREC)

Economic Impact Survey:

"…responsible for approximately ***one-third of an improvement*** of more than 200% in web search products that was observed between 1999 and 2009."

"…total extrapolated benefits were over ***$153 million*** for private, academic, and nonprofit organizations"

Company
Amazon
AOL
Apple
Ask
AT&T Bell Laboratories
Autonomy
Boeing Company
Cirrus Logic
CL Research
Clearwell Systems Inc.
Cleary Gottlieb Steen & Hamilton
Google
Harris Corporation
Hewlett-Packard
IBM
Language Computer Corporation
LCC International
Lucid Imagination
Microsoft
Omniture
Oracle Corp.
Progress Software Corp.
SABIR Research
Sapient Corporation
Sun Microsystems
Texas Instruments
The Echo Nest
Thomson Reuters Corporation
VIStology, Inc.
Xerox
Yahoo

Many thanks to Mark Sanderson @RMIT, for many of the introductory slides