Information Retrieval Evaluation

Emine Yilmaz
emine.yilmaz@ucl.ac.uk

Why is Evaluation so Important?

"What you can't measure you can't improve"

Lord Kelvin

Most retrieval systems are tuned to optimize for an objective evaluation metric

Measures for an IR system

- How fast does it index
  - Number of documents/hour
  - (Average document size)
- How fast does it search
  - Latency as a function of index size
- Expressiveness of query language
  - Ability to express complex information needs
  - Speed on complex queries
- Uncluttered UI
- Is it free?

Measures for an IR system

- All of the preceding criteria are *measurable*: we can quantify speed/size
- The key measure: user happiness
  - What is this?
  - Speed of response/size of index are factors
  - But blindingly fast, useless answers won't make a user happy
- Need a way of quantifying user happiness

Information Retrieval Systems

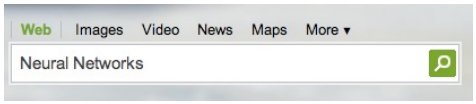
Match information seekers with
the information they seek

I want to learn about Neural Nets…

Results

User's Request

Search Engine

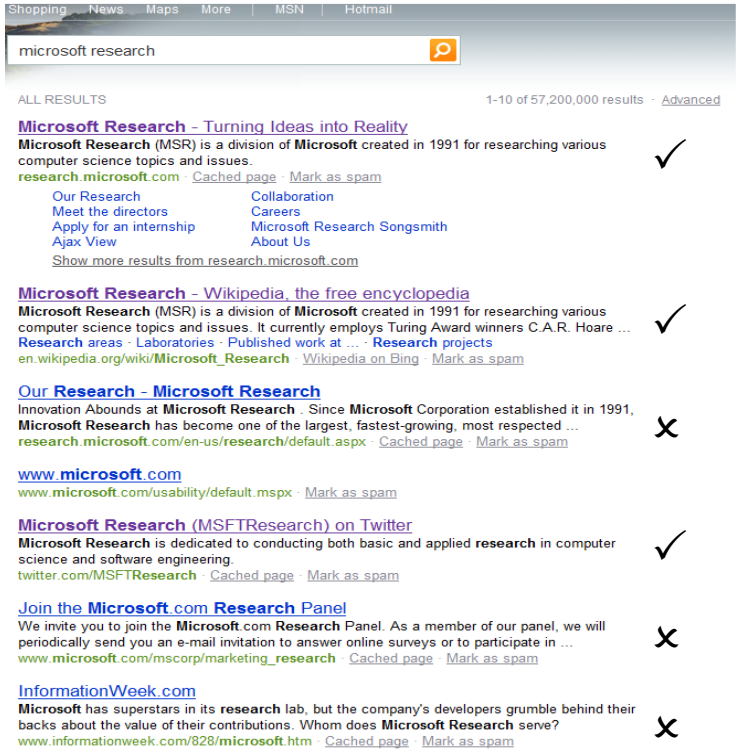


Collection

Different Approaches to Evaluation

- Online Evaluation
  - Design interactive experiments
  - Use users' actions to evaluate the quality
- Offline Evaluation
  - Controlled laboratory experiments
  - The users' interaction with the engine is only simulated

Online Evaluation



✓

✓

✗

✓

✗

✗

Click/Noclick

Evaluate

Online Evaluation

- Standard click metrics
  - Clickthrough rate
  - Improved version of clickthrough rate focusing clicks with high dwell time
    - Dwell time: Time spent on a clicked document
  - Queries per user
  - Probability user skips over results they have considered (pSkip)

- Current approach: Result interleaving

What is result interleaving?

- A way to compare rankers online
  - Given the two rankings produced by two methods
  - Present a combination of the rankings to users
  - Team Draft Interleaving (Radlinski et al., 2008)
    - Interleaving two rankings
      - Input: Two rankings ("can be seen as teams who pick players")
      - Repeat:
        - » Toss a coin to see which team (ranking) picks next
        - » Winner picks their best remaining player (document)
        - » Loser picks their best remaining player (document)
      - Output: One ranking (2 teams of 5)
    - Credit assignment
      - Ranking providing more of the clicked results wins

Ranking A

1. Napa Valley – The authority for lodging...
   www.napavalley.com
2. Napa Valley Wineries - Plan your wine...
   www.napavalley.com/wineries
3. Napa Valley College
   www.napavalley.edu/homex.asp
4. Been There | Tips | Napa Valley
   www.ivebeenthere.co.u
5. Napa Valley Wineries an
   www.napavintners.com
6. Napa Country, California
   en.wikipedia.org/wiki/N

Ranking B

1. Napa Country, California – Wikipedia
   en.wikipedia.org/wiki/Napa_Valley
2. Napa Valley – The authority for lodging...
   www.napavalley.com
3. Napa: The Story of an American Eden...
   books.google.co.uk/books?isbn=...
4. Napa Valley Hotels – Bed and Breakfast...
   .com

   y.org
   y Marathon
   ymarathon.org

Presented Ranking

Ranking A

1. Napa Valley – The authority for lodging...
   www.napavalley.com
2. Napa Valley Wineries - Plan your wine...
   www.napavalley.com/wineries
3. Napa Valley College
   www.napavalley.edu/homex.asp
4. Been There | Tips | Napa Valley
   www.ivebeenthere.co.u
5. Napa Valley Wineries an
   www.napavintners.com
6. Napa Country, California
   en.wikipedia.org/wiki/N

Ranking B

1. Napa Country, California – Wikipedia
   en.wikipedia.org/wiki/Napa_Valley
2. Napa Valley – The authority for lodging...
   www.napavalley.com
3. Napa: The Story of an American Eden...
   books.google.co.uk/books?isbn=...
4. Napa Valley Hotels – Bed and Breakfast...
   .com

Presented Ranking

1. Napa Valley – The authority for lodging...
   www.napavalley.com
2. Napa Country, California – Wikipedia
   en.wikipedia.org/wiki/Napa_Valley
3. Napa: The Story of an American Eden...
   books.google.co.uk/books?isbn=...
4. Napa Valley Wineries – Plan your wine...
   www.napavalley.com/wineries
5. Napa Valley Hotels – Bed and Breakfast...
   www.napalinks.com
6. Napa Valley College
   www.napavalley.edu/homex.asp
7. NapaValley.org
   www.napavalley.org

Click

Click

B wins!

Ranking A

1. Napa Valley – The authority for lodging...
   www.napavalley.com
2. Napa Valley Wineries - Plan your wine...
   www.napavalley.com/wineries
3. Napa Valley College
   www.napavalley.edu/homex.a
4. Been There | Tips | Napa Valley
   www.ivebeenthere.co.u
5. Napa Valley Wineries an
   www.napavintners.com
6. Napa Country, California
   en.wikipedia.org/wiki/N

Repeat Over Many Different Queries!

Ranking B

1. Napa Country, California – Wikipedia
   en.wikipedia.org/wiki/Napa_Valley
2. Napa Valley – The authority for lodging...
   www.napavalley.com
3. he Story of an American Eden...
   google.co.uk/books?isbn=...
4. Napa Valley Hotels – Bed and Breakfast...
   .com

y.org
on
tion.org

Presented Ranking

1. Napa Valley – The authority for lodging...
   www.napavalley.com
2. Napa Country, California – Wikipedia
   en.wikipedia.org/wiki/Napa_Valley
3. Napa: The Story of an American Eden...
   books.google.co.uk/books?isbn=...
4. Napa Valley Wineries – Plan your wine...
   www.napavalley.com/wineries
5. Napa Valley Hotels – Bed and Breakfast...
   www.napalinks.com
6. Napa Valley College
   www.napavalley.edu/homex.asp
7. NapaValley.org
   www.napavalley.org

Click

Click

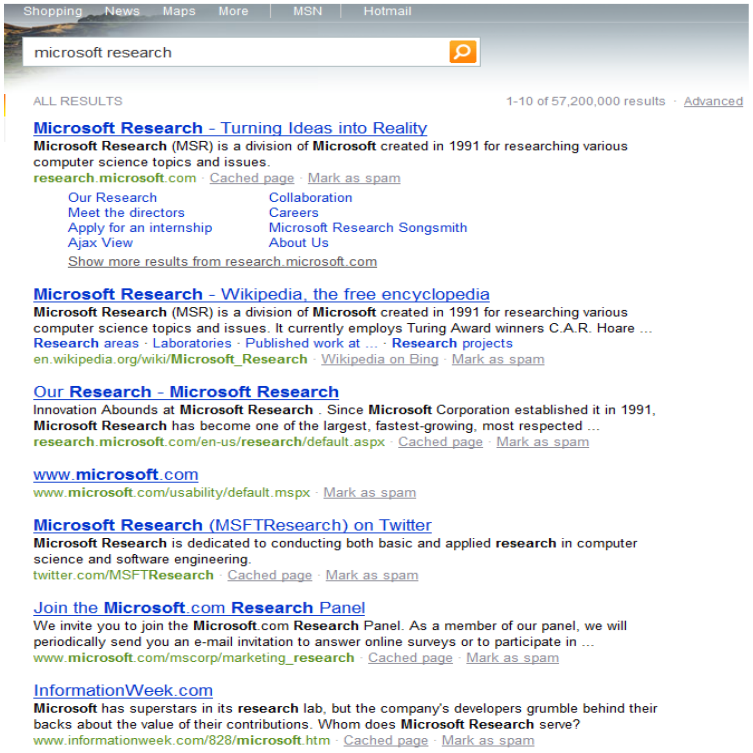
B wins!

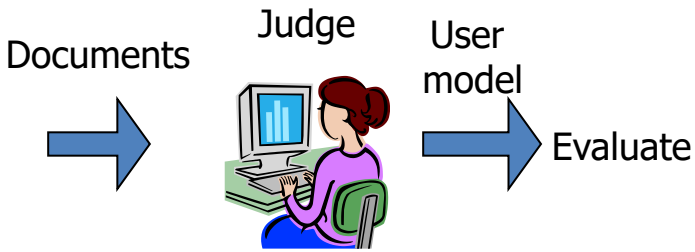
Offline Evaluation

- Controlled laboratory experiments

- The user's interaction with the engine is only simulated
  - Ask experts to judge each query result
  - Predict how users behave when they search
  - Aggregate judgments to evaluate

Offline Evaluation





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

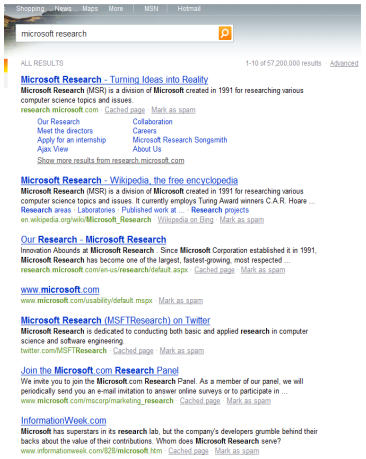
Online vs. Offline Evaluation

	Online	Offline
Pros	Cheap Measure actual user reactions	Fast Easy to try new ideas Amortized Cost Portable
Cons	Need to go live Noisy Slow Not duplicable	Needs ground truth Slow at the beginning "Expensive" Inconsistent Difficult to model how users behave

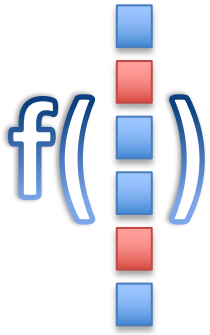
Offline Evaluation: Traditional Experiment

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

Results

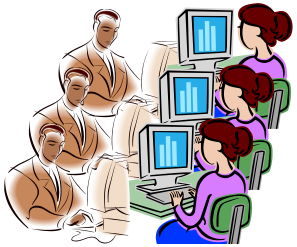


Search Engines



How many good docs have I missed/found?

Judges



Offline Evaluation

- Given a *test collection* consisting of
  - a collection of documents
  - a set of queries
  - relevance judgments on each document for each query
- Evaluate the quality of a retrieval system

Evaluation in Early IR systems

- Many early IR systems were 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

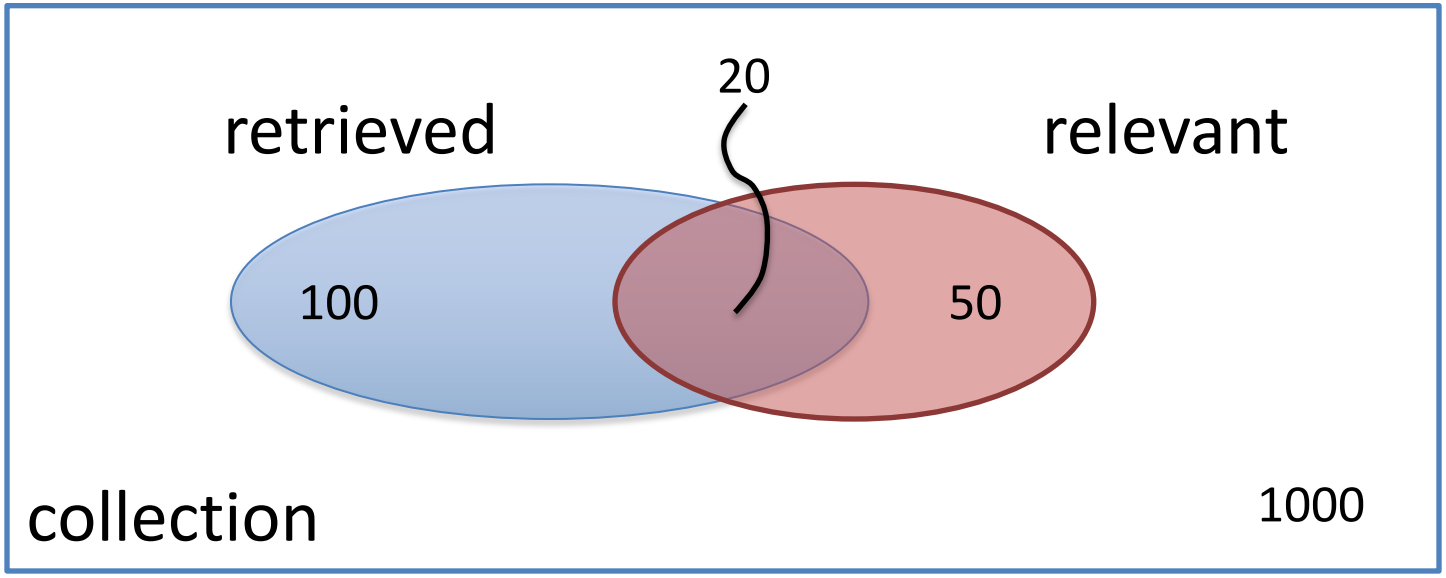
Measuring Boolean Output

- Contingency table

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

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

Measuring Boolean Output



Precision = 20/100 = 0.2

Recall = 20/50 = 0.4

Trading off Between Precision and Recall

- Precision tends to decrease as recall increases

- We usually care about both precision and recall
  - Van Rijsbergen's F: weighted harmonic mean

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

F measure: Example

	relevant	not relevant	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

- P = 20/(20 + 40) = 1/3
- R = 20/(20 + 60) = 1/4
- when α=0.5,

$$F_1 = \frac{1}{\frac{1}{2} \times \frac{1}{\frac{1}{3}} + \frac{1}{2} \times \frac{1}{\frac{1}{4}}}$$

Evaluating best match systems (Ranking)

- Modern IR systems: Best Match
  - Return a ranked list of documents instead of a set of documents

- A good system returns relevant documents before non-relevant

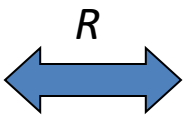
Precision-recall curve

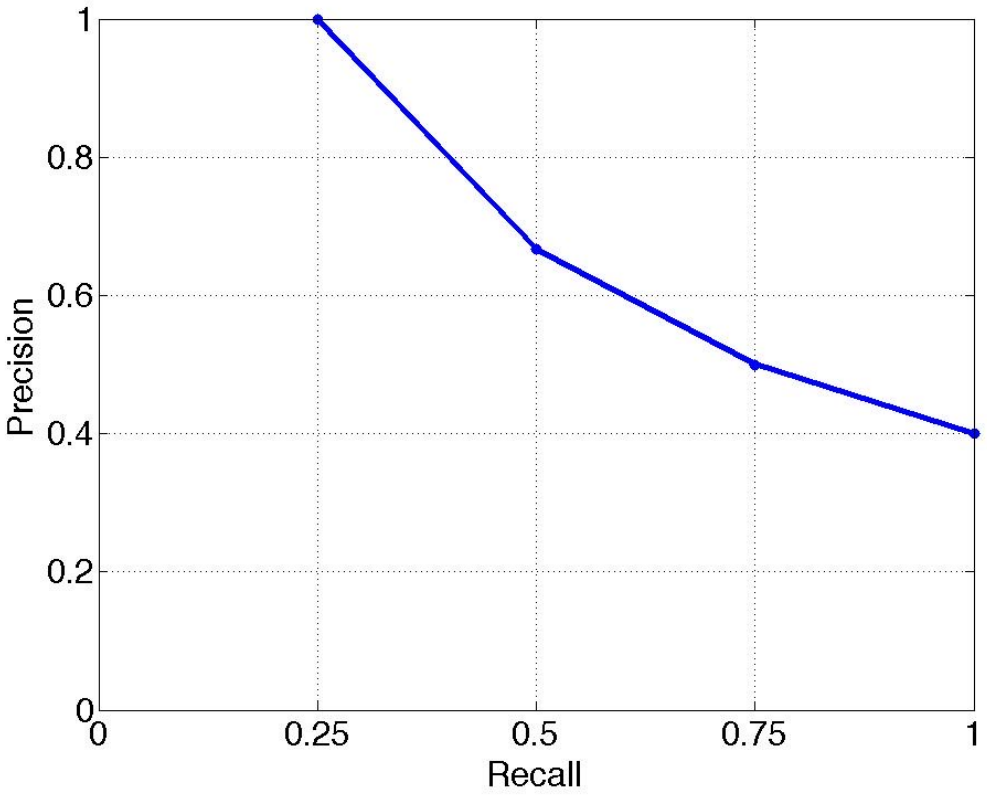
- Precision/recall/F are measures for unranked sets.
  - We can easily turn set measures into measures of ranked lists.
- Just compute the set measure for each “prefix”: the top 1, top 2, top 3, top 4 etc. results
- Doing this for precision and recall gives you a precision-recall curve.

Visualizing Retrieval Performance: Precision-Recall Curves

Retrieved list

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





Average Precision

- Average of precisions at relevant documents
- If a relevant document is not retrieved by a system precision is assumed to be zero!
- Piecewise-linear approximation to the area under the precision-recall curve

Average Precision

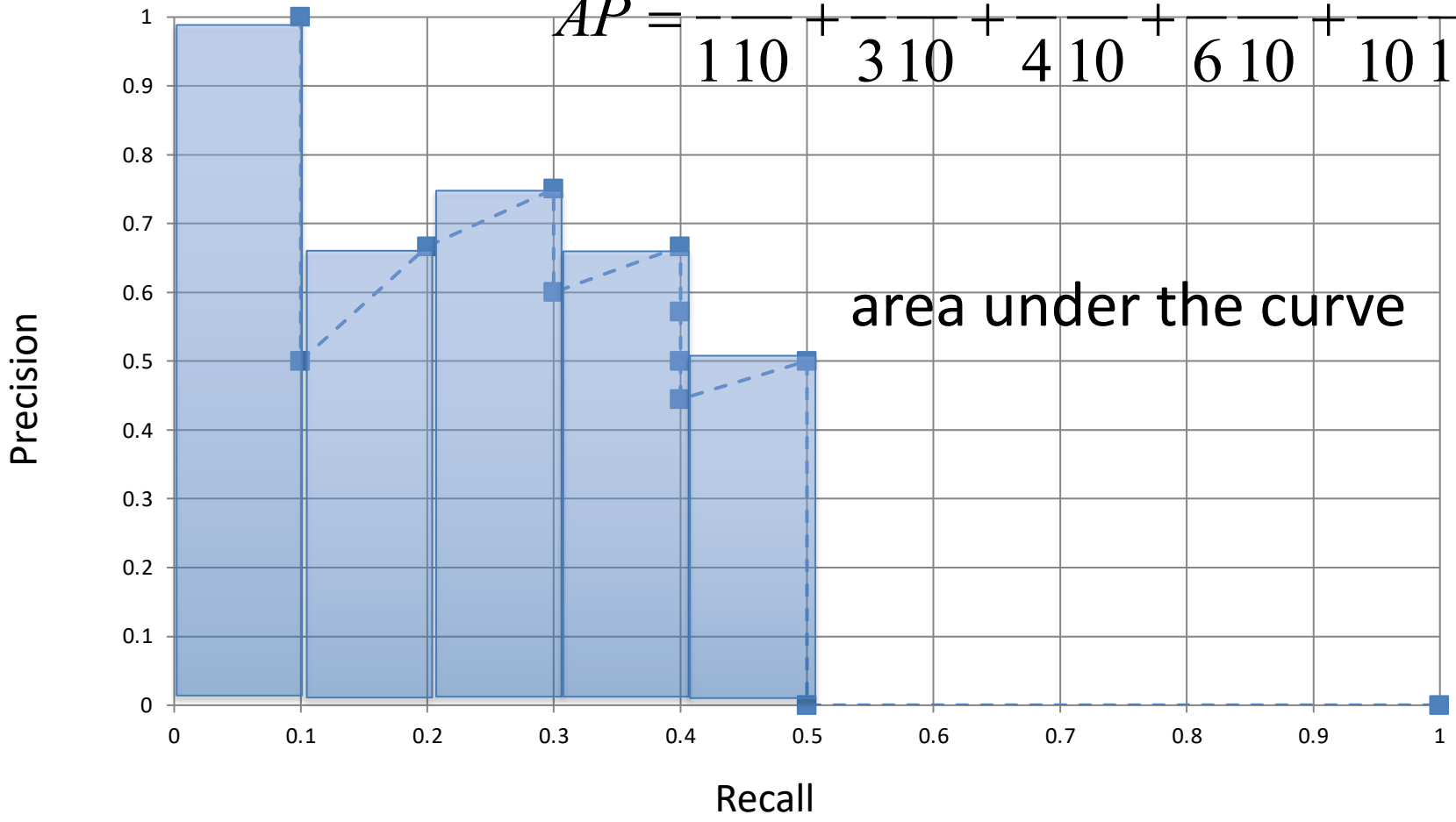
Topic 1

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

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

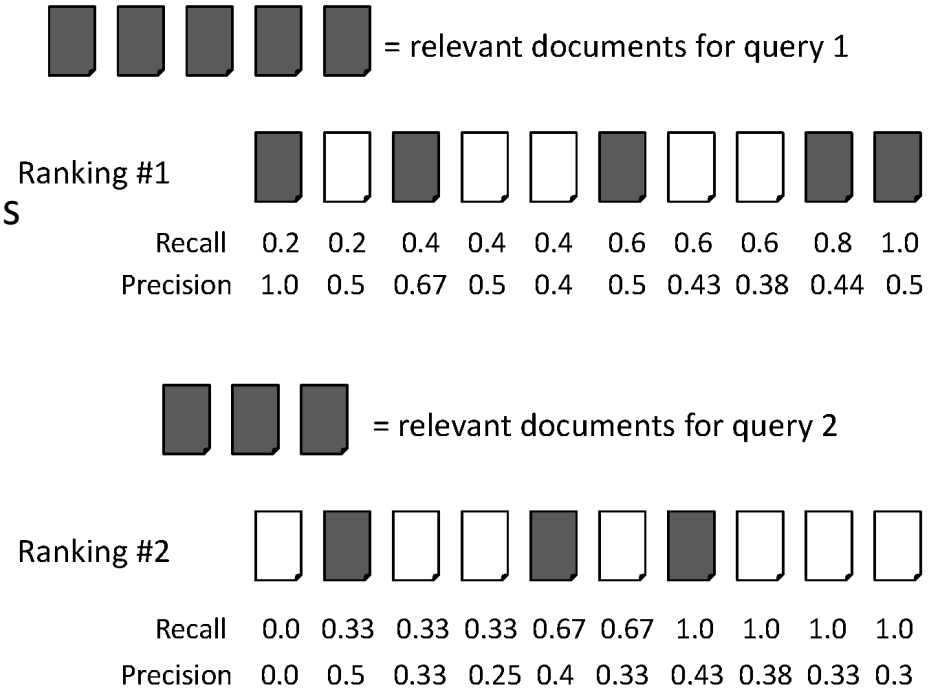
Average Precision

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



Mean Average Precision

- Mean average precision of a system:
  - Average AP values over all queries



$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$mean\ average\ precision = (0.62 + 0.44)/2 = 0.53$$

Metrics that Focus on Top Documents

- ***Precision at Rank k***
  - k is typically 5, 10, 20
  - easy to compute, average, understand
  - not sensitive to rank positions less than k

- ***R-Precision*** - Precision at rank ***R***, where ***R*** is number of relevant documents in the query

- ***Reciprocal Rank***
  - reciprocal of the rank at which the first relevant document is retrieved
  - very sensitive to rank position

- ***Expected Search Length*** - Number of non-relevant documents ranked before seeing the 1st relevant document

Discounted Cumulative Gain (DCG)

- Popular measure for evaluating web search and related tasks
  - emphasis on retrieving highly relevant documents
    - Graded relevance judgments as opposed to binary
  - used by many web search companies

- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant document
  - the lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined

Discounted Cumulative Gain (DCG)

- Uses *graded relevance* as a measure of the usefulness, or *gain,* from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks
- Typical discount is 1/*log (rank)*
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Discounted Cumulative Gain (DCG)

- *DCG* is the total gain accumulated at a particular rank *p*:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

gain

discount

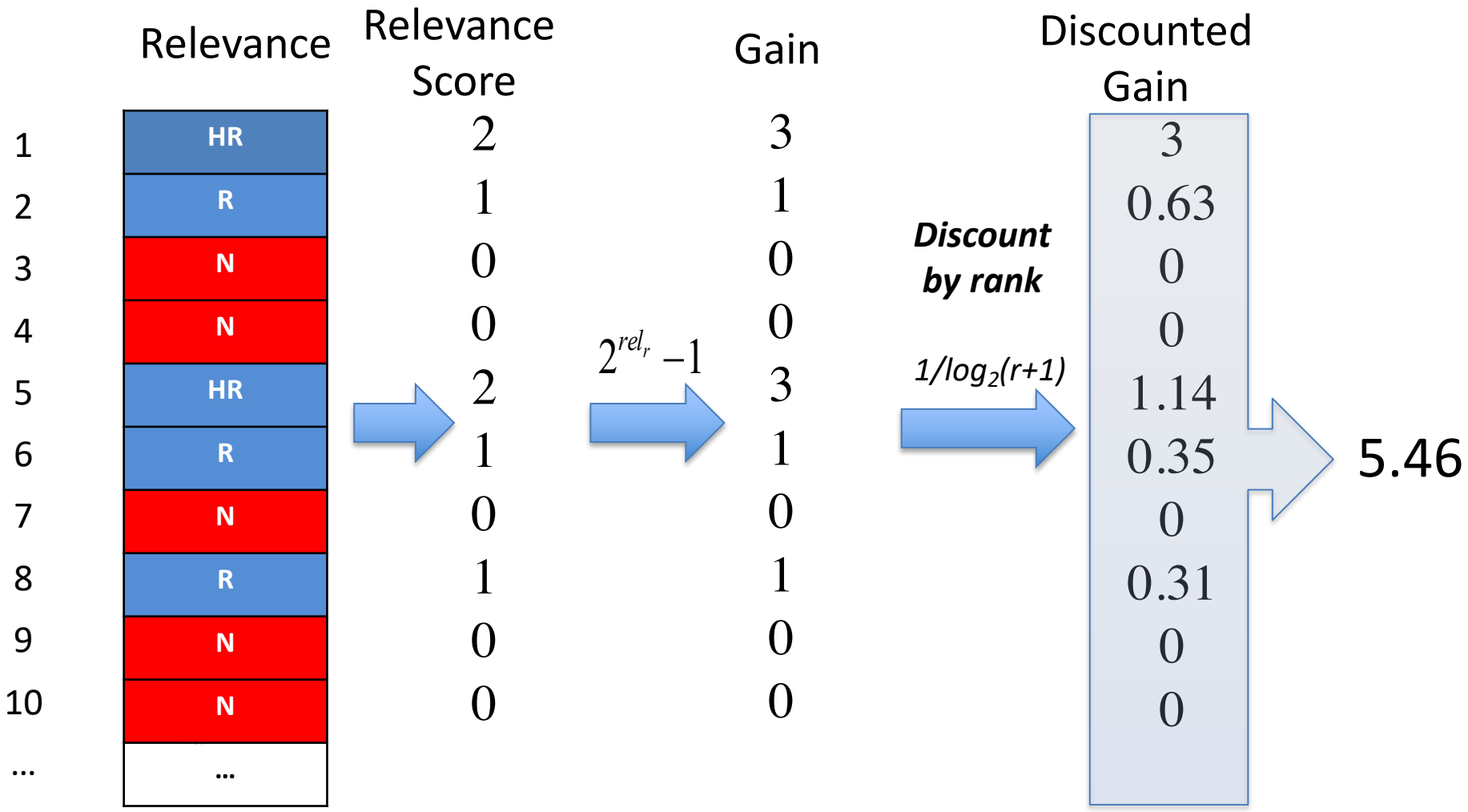
- Alternative formulation (adding nonlinear gain function):

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

gain

discount

DCG Example

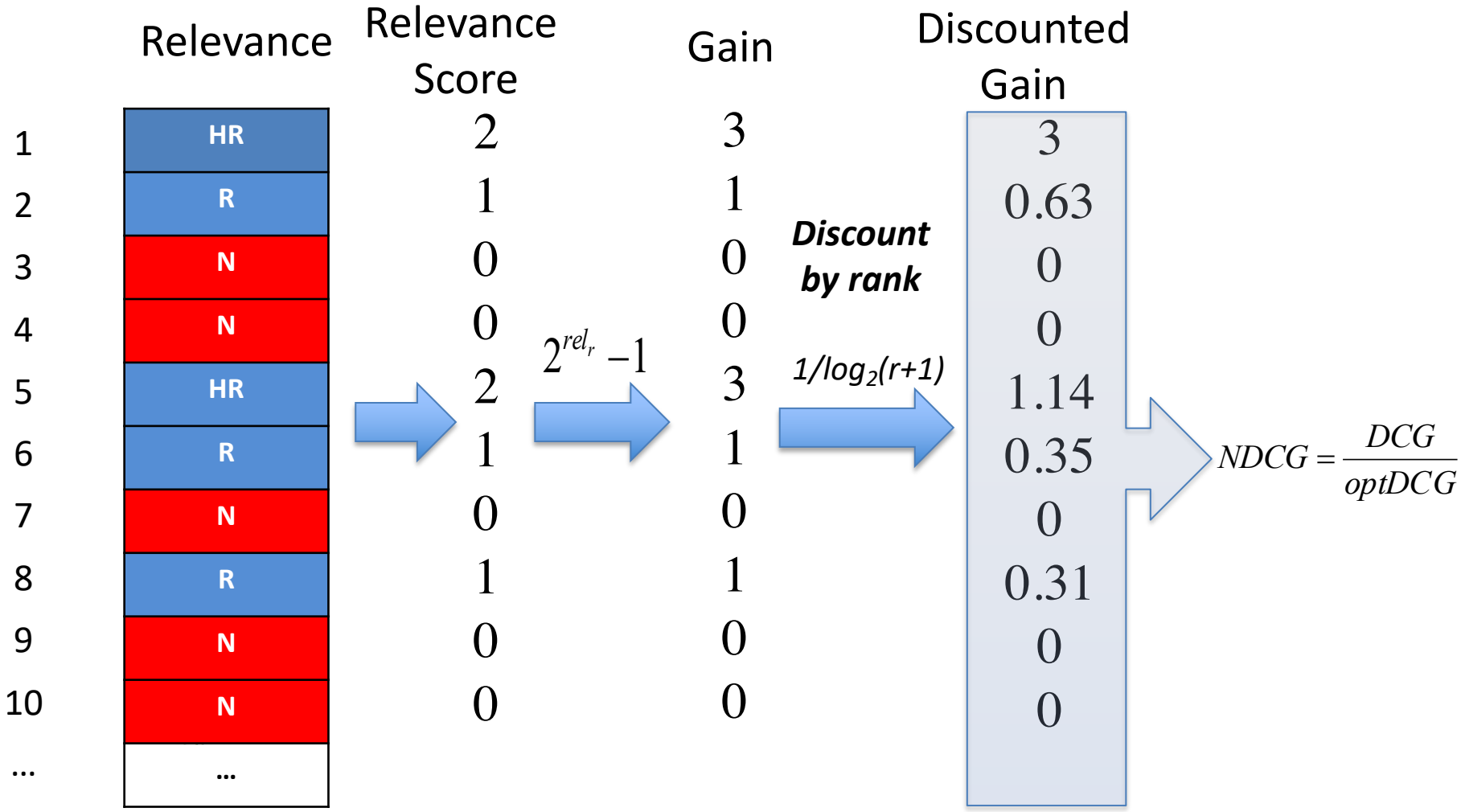

Relevance
Relevance Score
Gain
Discounted Gain
1
2
3
4
5
6
7
8
9
10
...
HR
R
N
N
HR
R
N
R
N
N
...
2
1
0
0
2
1
0
1
0
0
$2^{rel_r} - 1$
3
1
0
0
3
1
0
1
0
0
Discount by rank
$1/\log_2(r+1)$
3
0.63
0
0
1.14
0.35
0
0.31
0
0
5.46

Normalized Discounted Cumulative Gain (NDCG)

- Normalizes DCG against the best possible DCG result (the *perfect ranking*) for the query
  - 0<=NDCG<=1 (DCG values can go much larger)
  - makes averaging easier for queries with different numbers of relevant documents

- Then the NDCG can be averaged over all queries

Normalized Discounted Cumulative Gain



Evaluation Using Preferences

- Two rankings described using preferences can be compared using the Kendall's tau (τ) coefficient

$$\tau = \frac{P - Q}{P + Q}$$

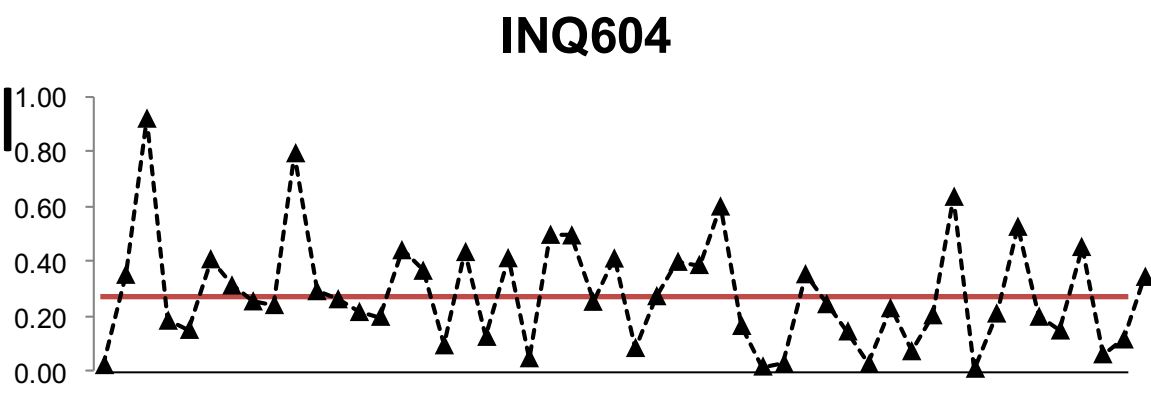
- P is the number of preferences that agree and Q is the number that disagree

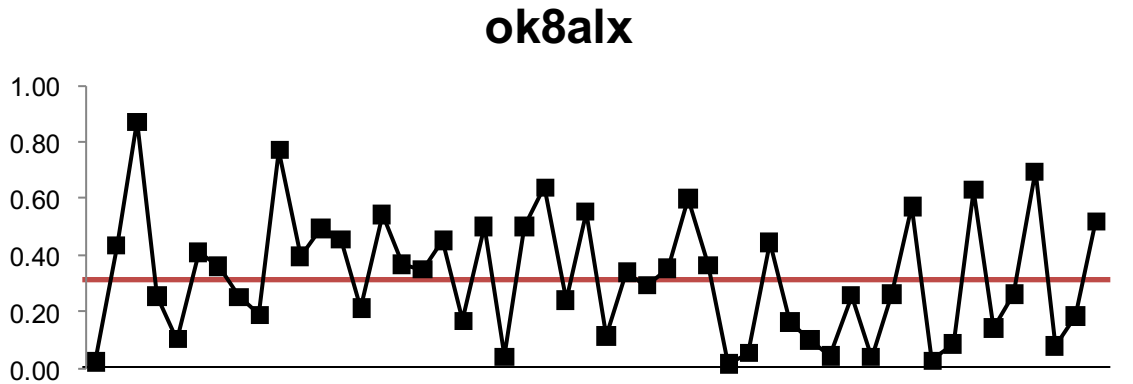
Categories of Evaluation Measures

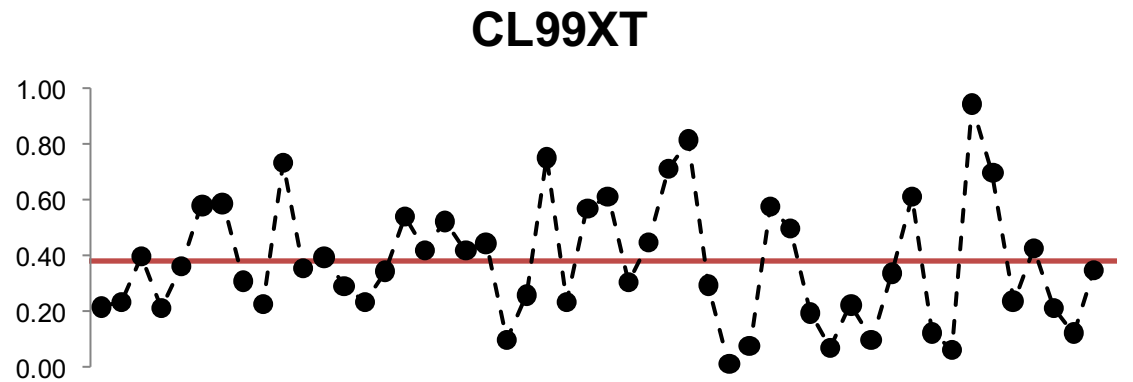
- User-oriented
  - Underlying user model to simulate user's interaction
    - **Precision@k**, **nDCG@k**, **MRR**, RBP, ERR, …
- System-oriented
  - Also often based on a user model [Robertson SIGIR08]
  - Capture the overall effectiveness of retrieval system
    - **Average Precision, R-Precision**, **nDCG**, normalized Recall, PRES [Magdy and Jones SIGIR10], GAP [Robertson et al SIGIR10], …

Comparing Retrieval Systems

- INQ604 - 0.281
- ok8alx - 0.324
- CL99XT - 0.373



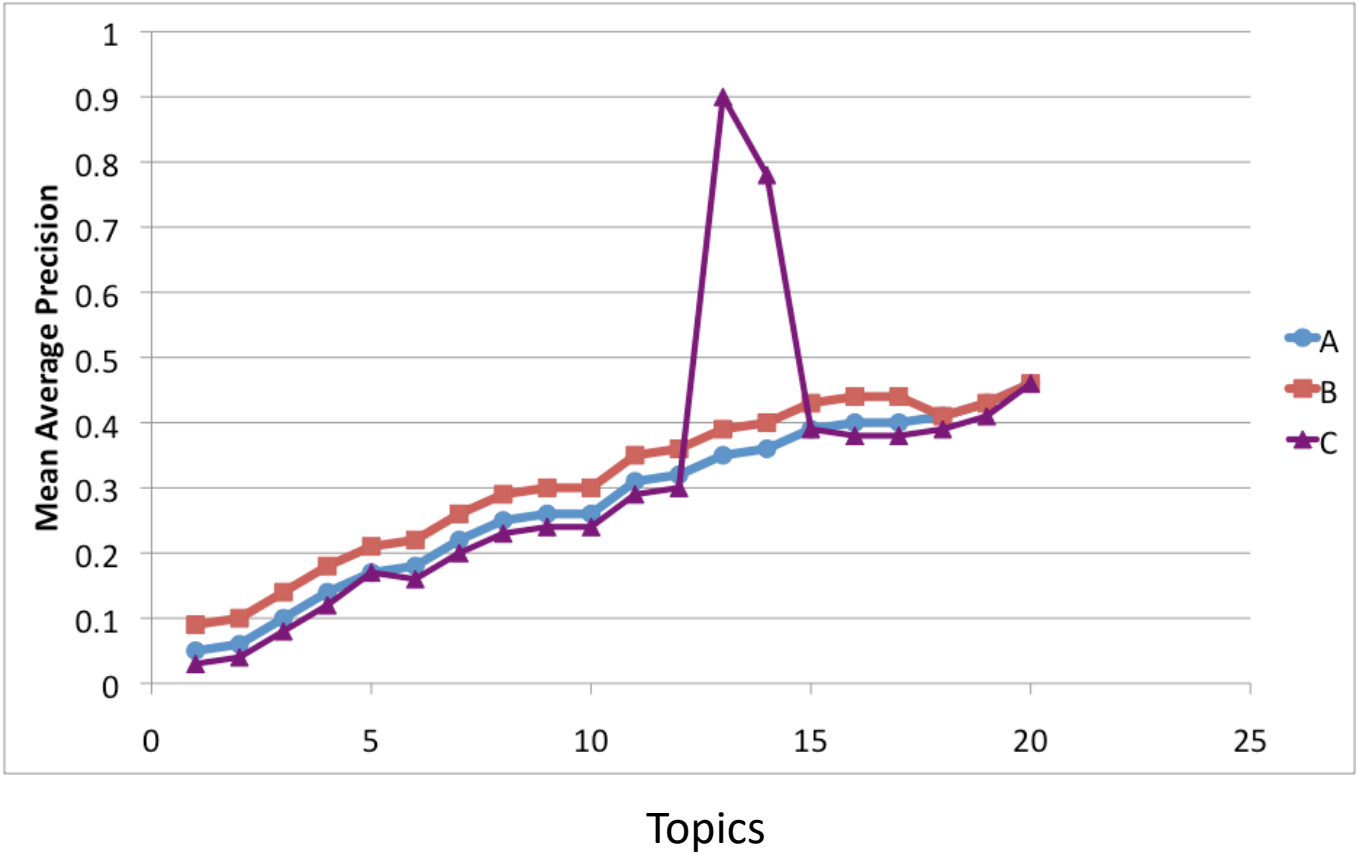




How much is the mean

- A product of
  - your IR system
  - chance?
- What if we had slightly different set of topics?
  - Would the average change?

Is C really better than A?



Comparison and Significance

- Variability in effectiveness scores across different queries
- When observing a difference in effectiveness scores across two retrieval systems
  - Does this difference occur by **random chance**?
- Significance testing
  - Estimates the probability *p* of observing a certain difference in effectiveness given that null hypothesis H_0 is true.
  - In IR evaluation
    - H_0 : the two systems are in effect the same and any difference in scores is by random chance.

Significance Testing

- Significance testing framework:
  - Two hypotheses, e.g.

    H_0: $\mu = 0$

    H_a: $\mu \neq 0$
  - System performance measurements over a sample of topics
  - A test statistic t computed from those measurements
  - A p-value, which is the probability of sampling t from a distribution obtained by assuming H_0 is true

Significance Testing in IR

- Calculate effectiveness measure for each query for each engine
- Use those values to compute a *test statistic* that has a certain distribution when the null hypothesis is true
  - Null hypothesis H_0: There is no difference in system performance
- Obtain a *p-value* from that distribution
  - The value of the test statistic gives us a p value
  - p-value: probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true

Significance Testing in IR

- If the *p*-value is less than a given **α** value, conclude that the null hypothesis is false
  - i.e., there is a difference between the systems

- **α:** probability of type I error
  - Probability of rejecting the null hypothesis when it is true

- Typical values of **α=0.05**, 0.1

Significance Testing: Different Approaches

- Parametric tests
  - Data comes form a type of probability distribution
  - Makes inferences about the parameters of the distribution

- Non-parametric tests
  - No assumption about a probability distribution
  - Based on ranks of observations

Student's t-test

- Parametric test
- Assumptions
  1. effectiveness score differences are meaningful
  2. effectiveness score differences follow normal distribution

- Statistic :

$$t = \frac{\overline{B-A}}{\sigma_{B-A}} . \sqrt{N}$$

Student's t-test

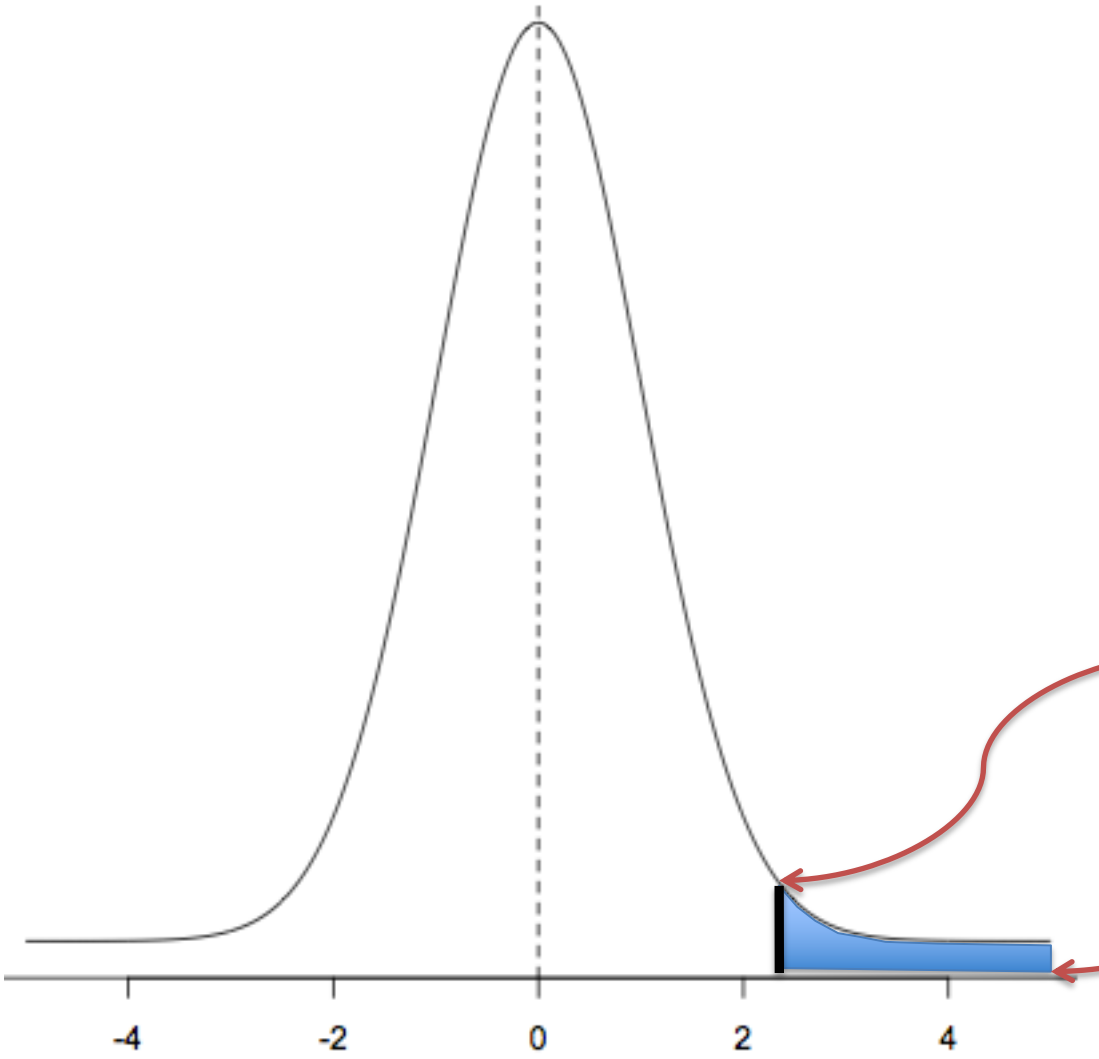
Query	A	B	B-A
1	.25	.35	+.10
2	.43	.84	+.41
3	.39	.15	-.24
4	.75	.75	0
5	.43	.68	+.25
6	.15	.85	+.70
7	.20	.80	+.60
8	.52	.50	-.02
9	.49	.58	+.09
10	.50	.75	+.25

$$\overline{B-A} = 0.214$$

$$\sigma_{B-A} = 0.291$$

$$t = \frac{B-A}{\sigma_{B-A}}\sqrt{n} = 2.33$$

Student's t-test



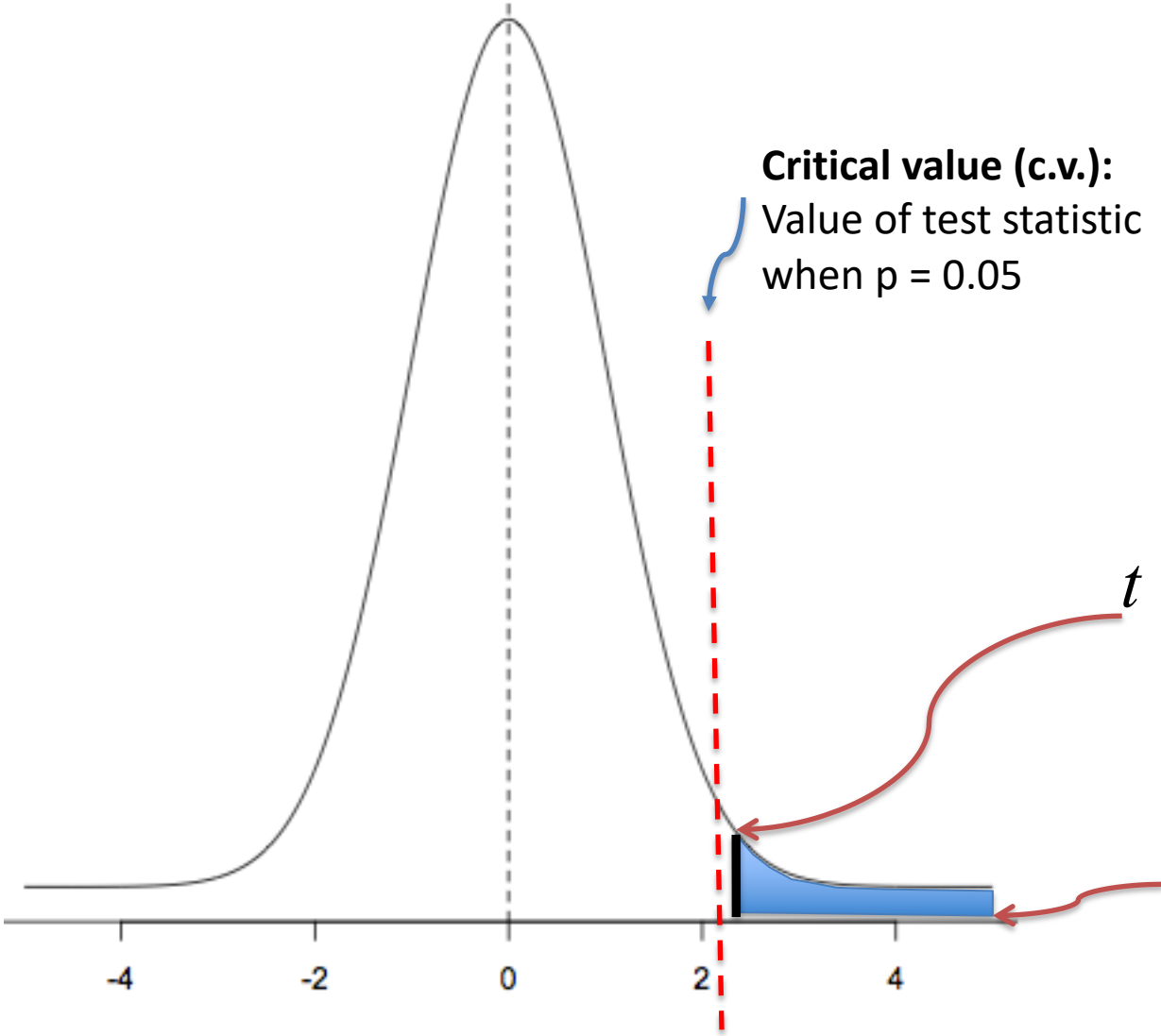
$$\overline{B-A} = 0.214$$

$$\sigma_{B-A} = 0.291$$

$$t = \frac{B-A}{\sigma_{B-A}}\sqrt{n} = 2.33$$

$$p - value = 0.02$$

Student's t-test



Critical value (c.v.):
Value of test statistic when p = 0.05

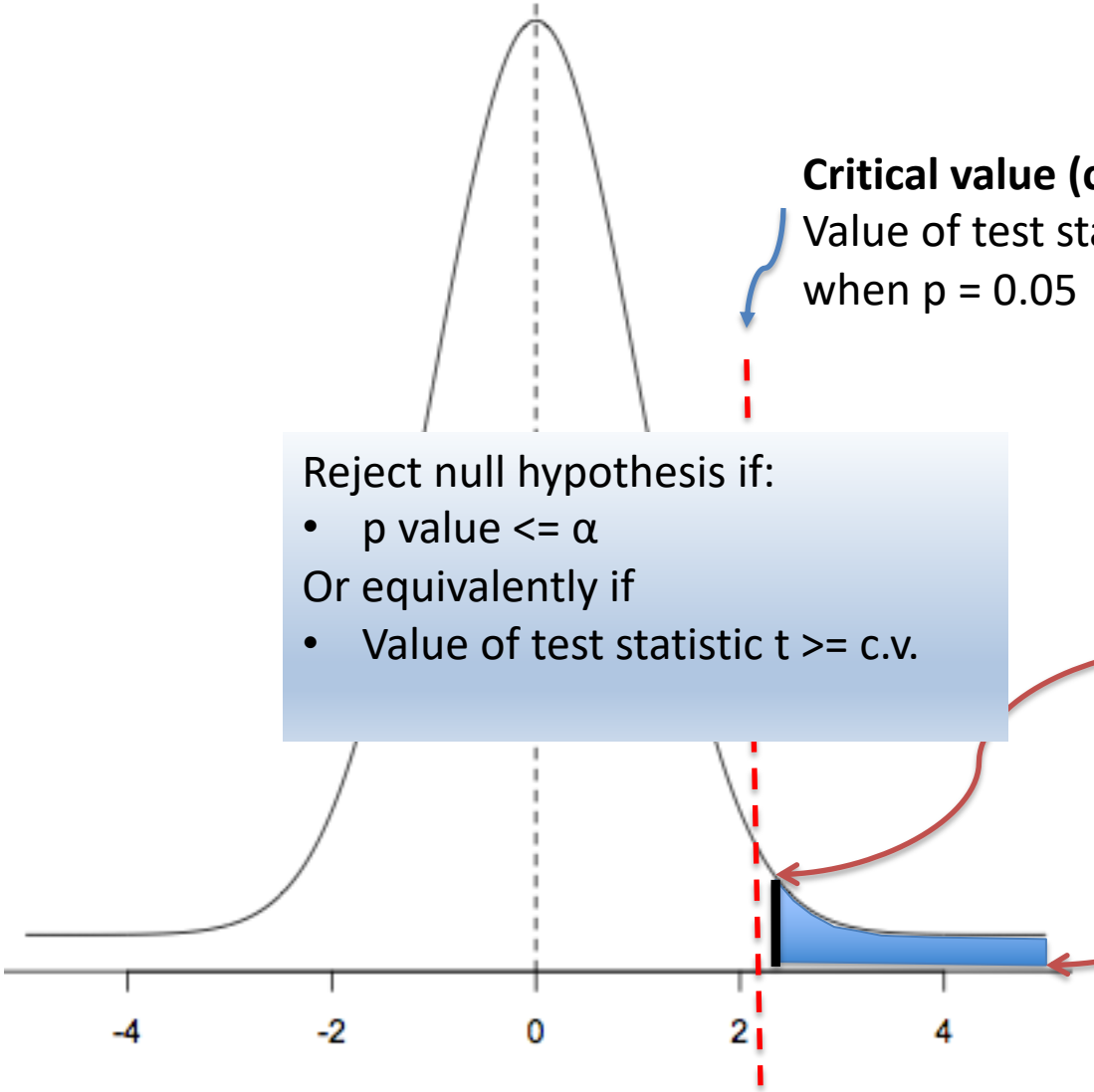
$$\overline{B-A} = 0.214$$

$$\sigma_{B-A} = 0.291$$

$$t = \frac{\overline{B-A}}{\sigma_{B-A}}\sqrt{n} = 2.33$$

$$p - value = 0.02$$

Student's t-test



Critical value (c.v.):
Value of test statistic when p = 0.05

Reject null hypothesis if:
- p value <= α

Or equivalently if
- Value of test statistic t >= c.v.

$$\overline{B-A} = 0.214$$

$$\sigma_{B-A} = 0.291$$

$$t = \frac{\overline{B-A}}{\sigma_{B-A}}\sqrt{n} = 2.33$$

$$p-value = 0.02$$

Sign Test

- Non-parametric test

- Ignores magnitude of differences
- Null hypothesis for this test is that
  - $P(B > A) = P(A > B) = ½$

- Statistic : number of pairs where B>A

Wilcoxon Signed-Ranks Test

- Non-parametric test

- Statistic: $$w = \sum_{i=1}^{N} R_i$$

  - R_i is a signed rank of absolute differences
  - N is the number of differences ≠ 0

Wilcoxon Signed-Ranks Test

Query	A	B	B-A
1	.25	.35	+.10
2	.43	.84	+.41
3	.39	.15	-.24
4	.75	.75	0
5	.43	.68	+.25
6	.15	.85	+.70
7	.20	.80	+.60
8	.52	.50	-.02
9	.49	.58	+.09
10	.50	.75	+.25

Sorted
-.02
+.09
+.10
-.24
+.25
+.25
+.41
+.60
+.70

Signed-rank
-1
+2
+3
-4
+5
+6
+7
+8
+9

$$w = \sum_{i=1}^{N} R_i$$

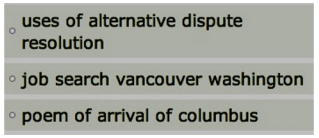
w = 35 =>

p = .025

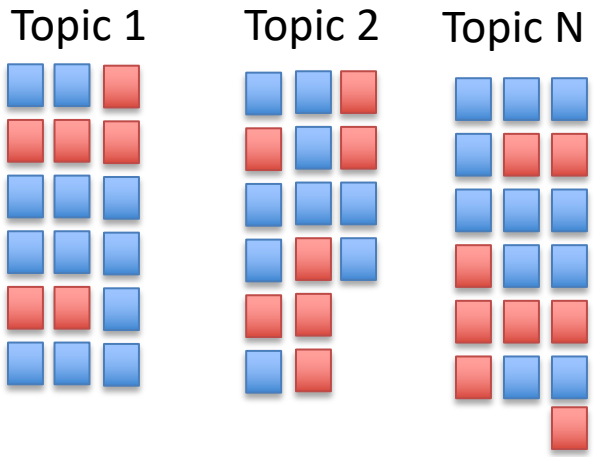
Test Collections

- Document Corpus

- Topics



- Relevance Judgments

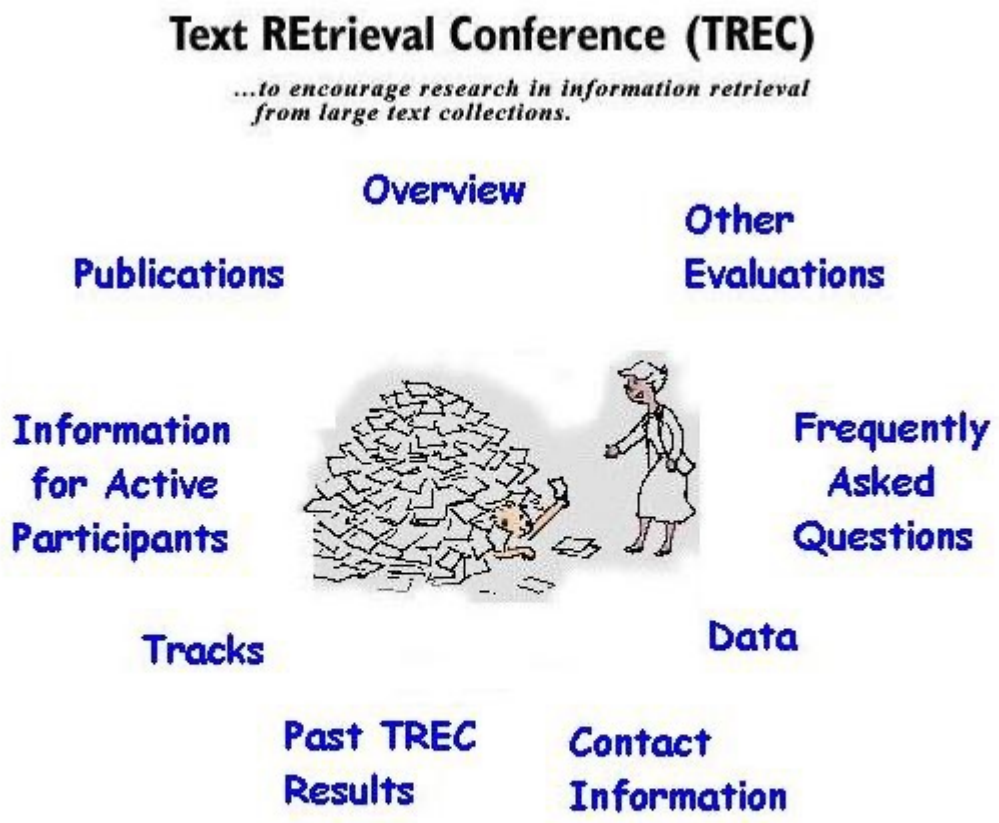


Test Collections Simulate

Retrieval Scenarios

TREC 1992

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



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

Test Collection Design

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

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

- ClueWeb12 collection
  - about 733,019,372 pages
  - 6 TB, compressed (~25 TB, uncompressed)
  - collected between February 10, 2012 and May 10, 2012
- 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?

TREC topics

- Early TREC collections
  - Searched collection for potential topics
  - Removed topics that returned too many relevant documents
  - Removed topics that returned too few relevant documents
  - Removed topics that appeared ambiguous

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 (very few)

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

Test Collection Design

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

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 Track 2010

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 Track 2010

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-Rel*

   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 documents to judge?

- Many measures are based on
  - recall : "*out of all good docs in the collection how many did the algorithm find?*"
  - all good documents in the collection need to be identified
    - Impossible for most current collections, e.g. the Web

Judgment Effort

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

Results

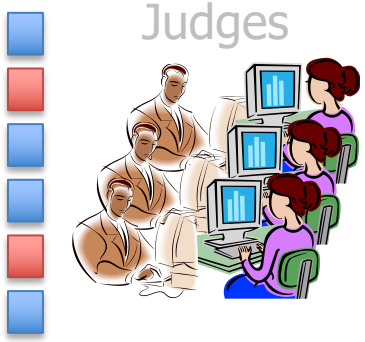
Search Engines

Cost

vs.

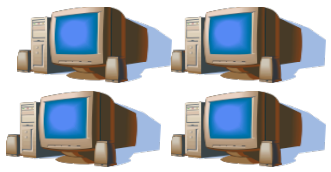
Reliability

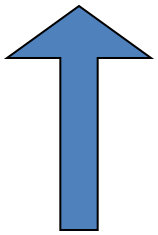
Judges



TREC Evaluation Setup

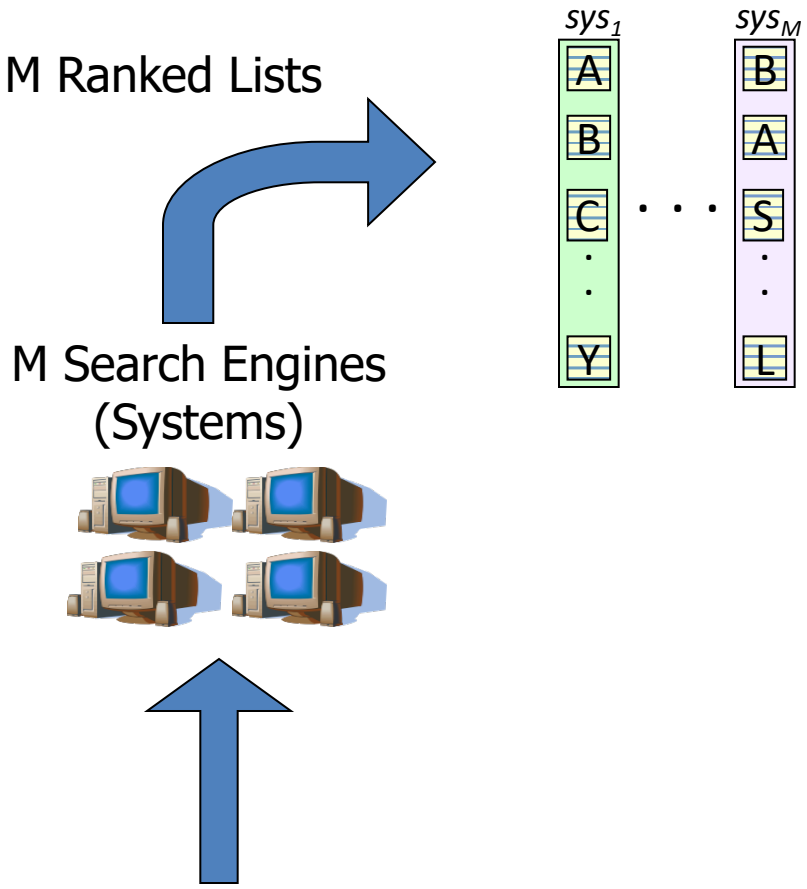
M Search Engines
(Systems)



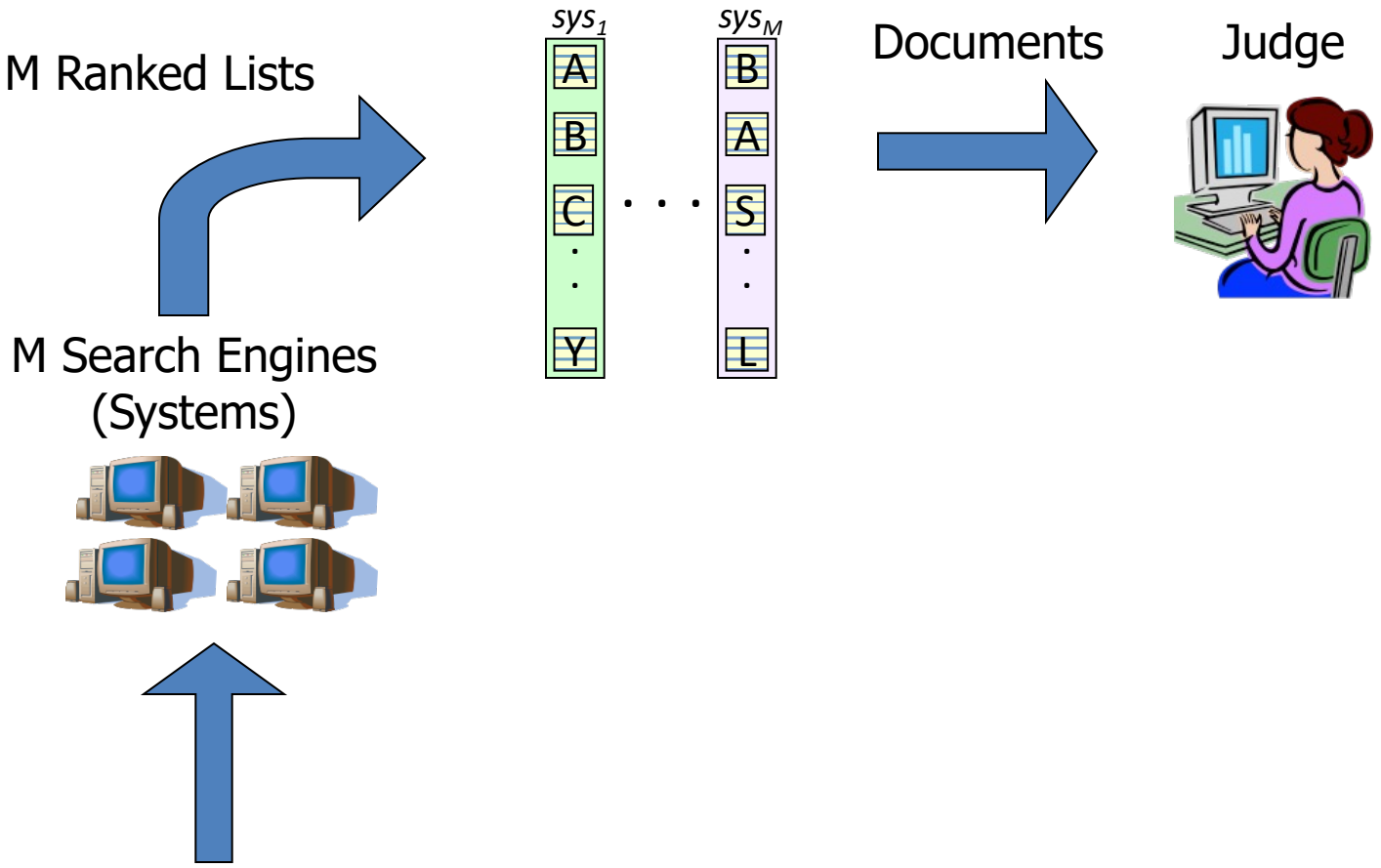


TREC Query (Q1)

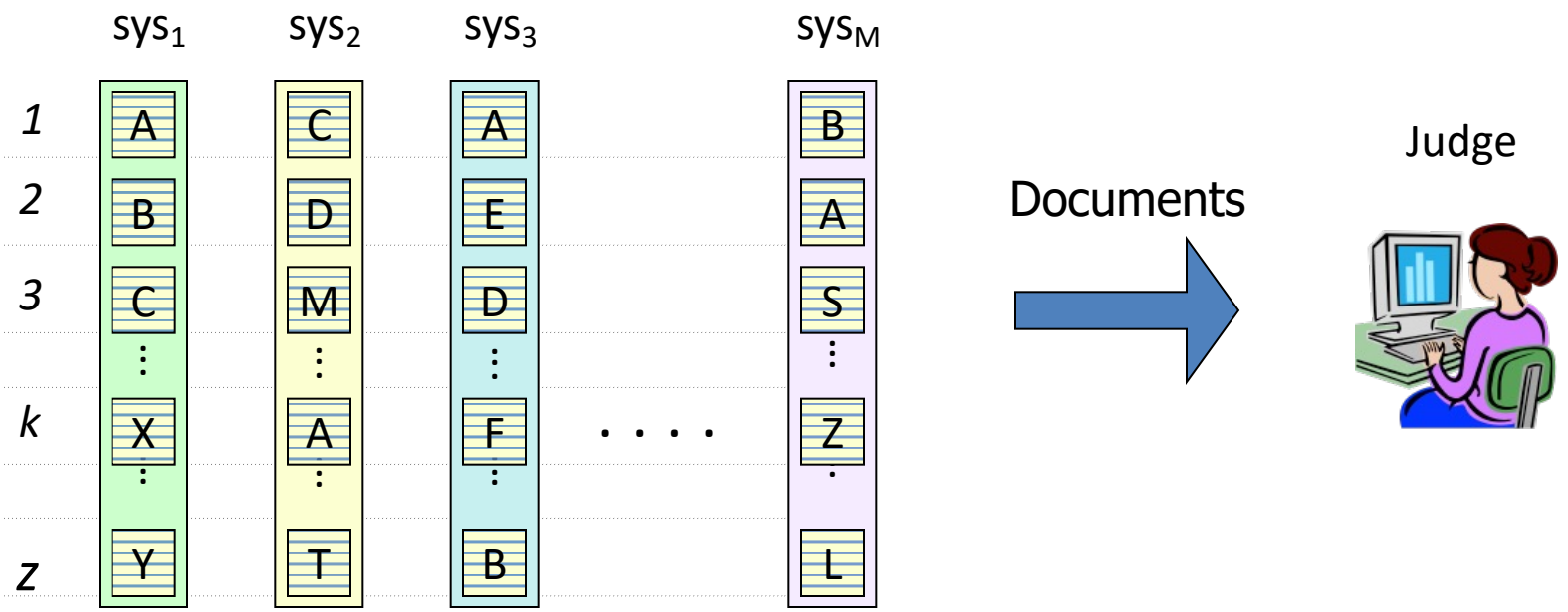
TREC Evaluation Setup



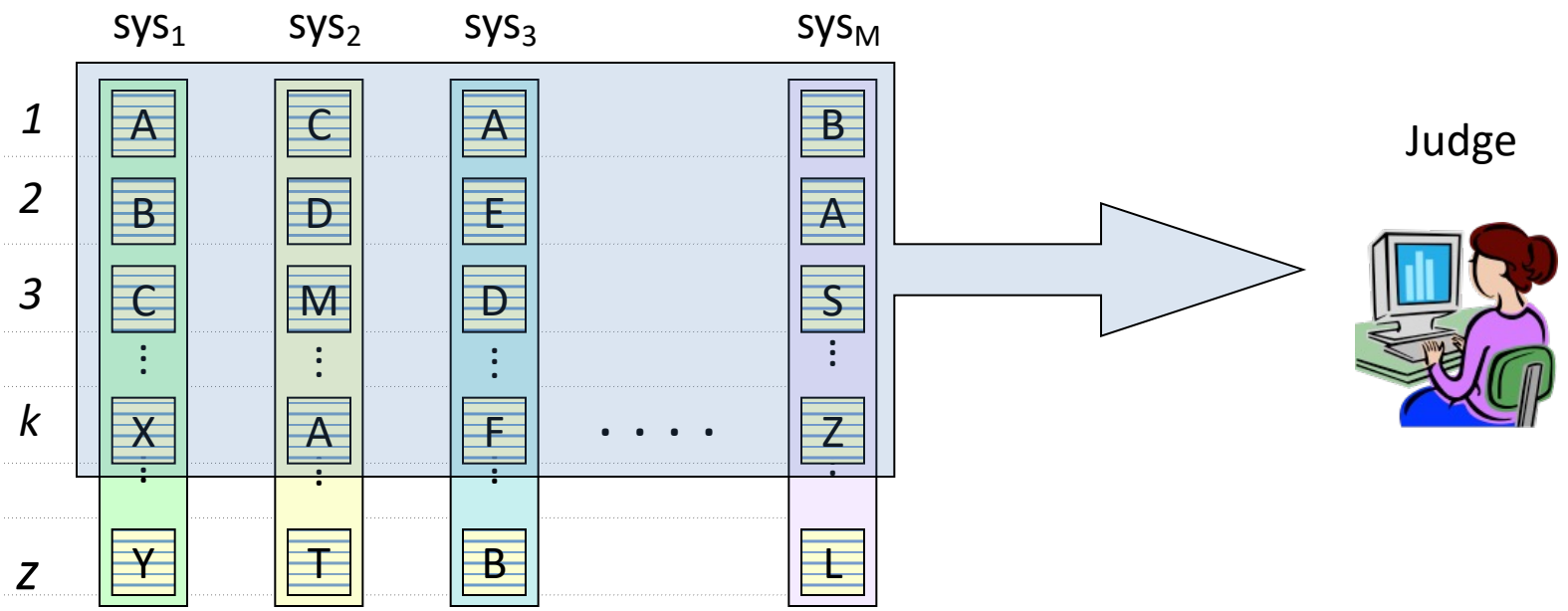
TREC Evaluation Setup



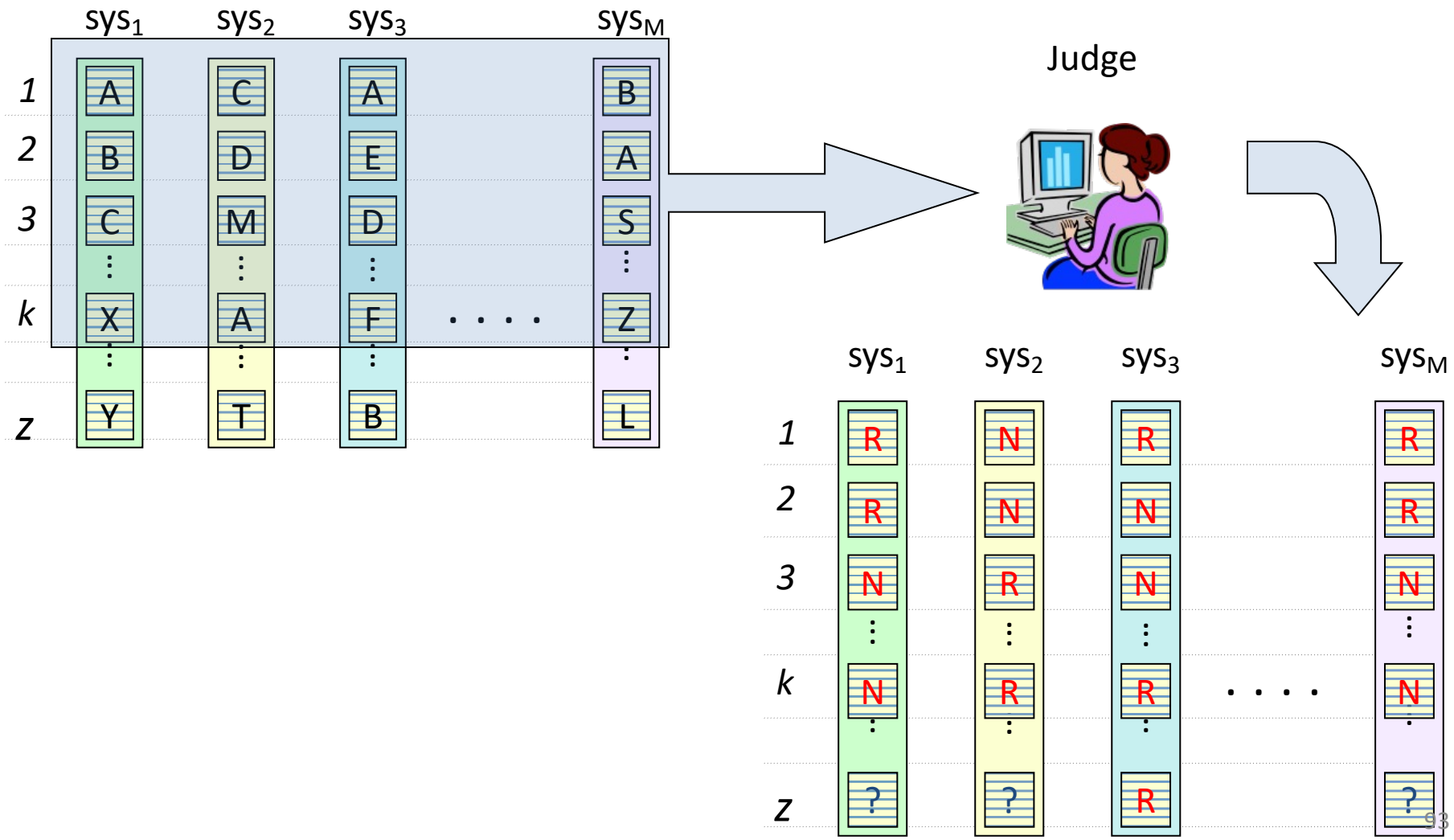
Techniques to Reduce Judgment Effort: Depth-k Pooling



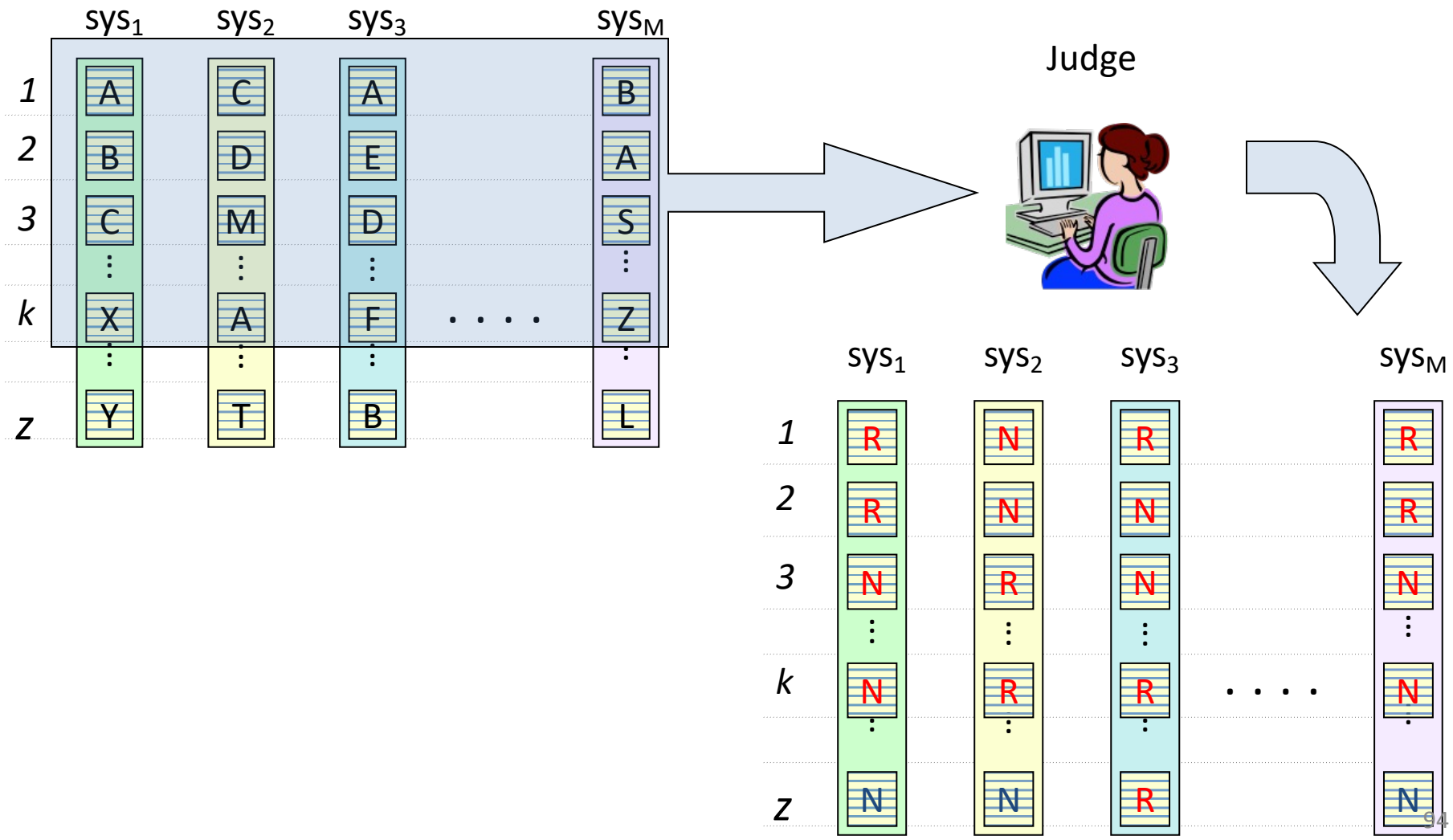
Depth-k Pooling



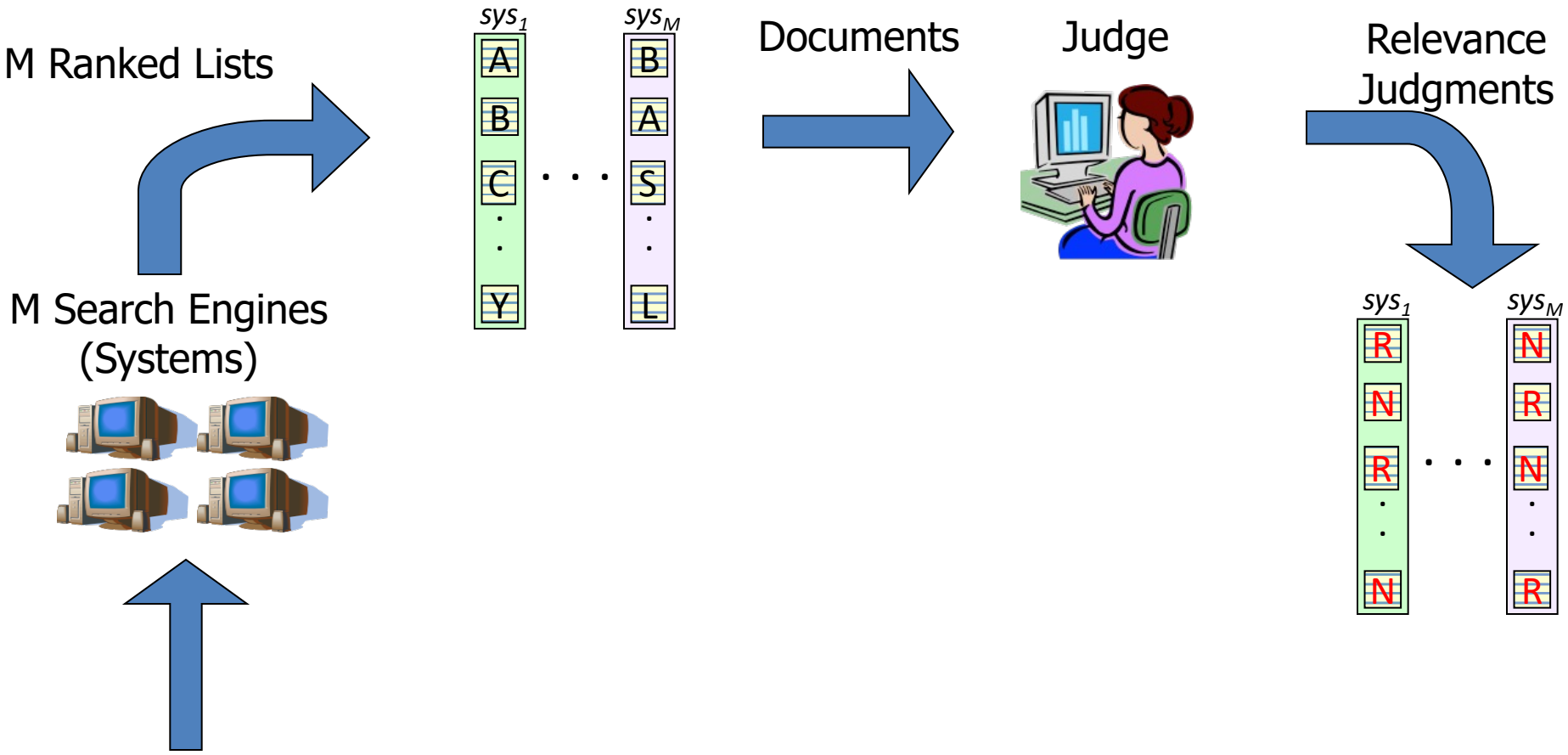
Depth-k Pooling



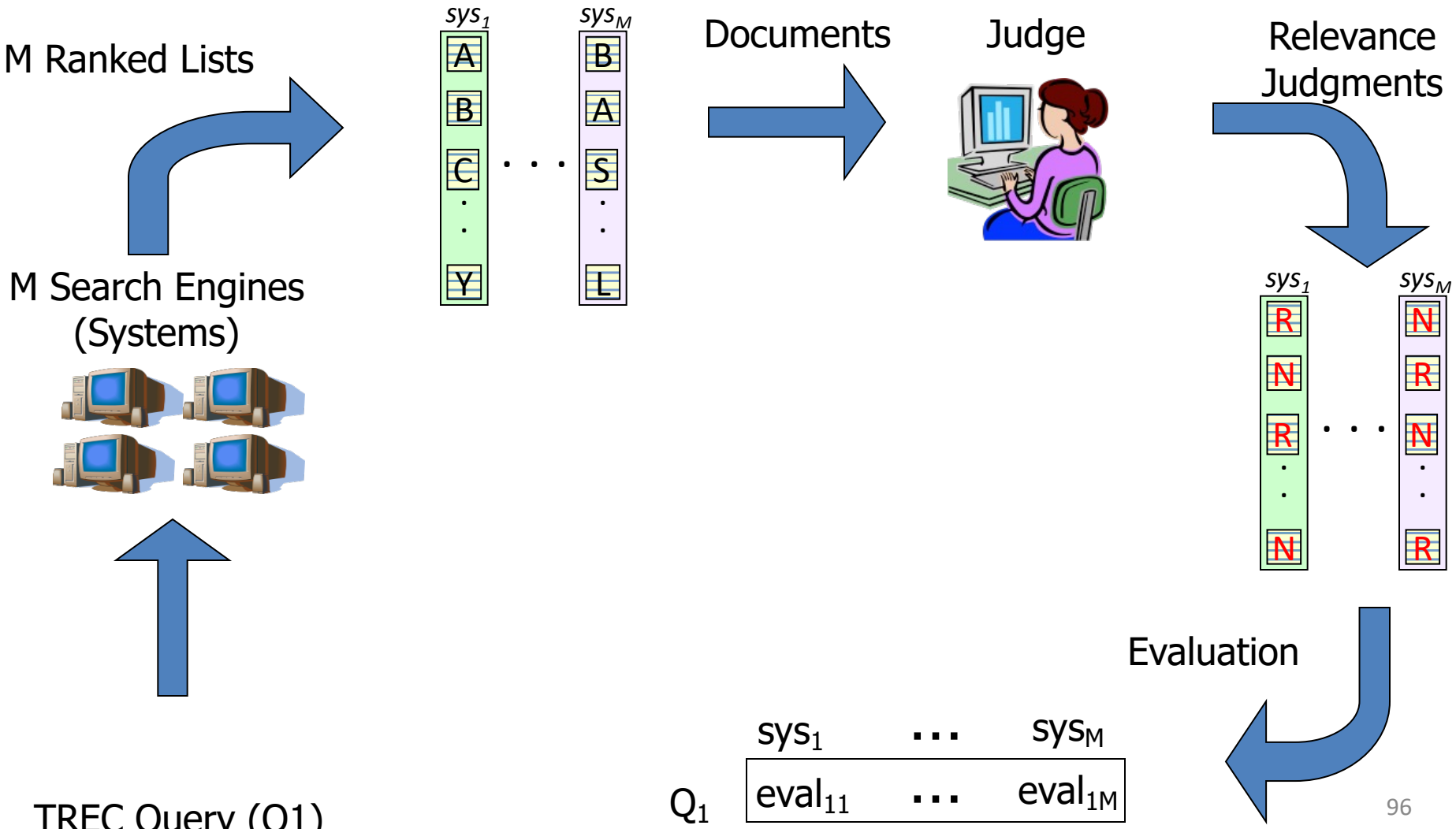
Depth-k Pooling



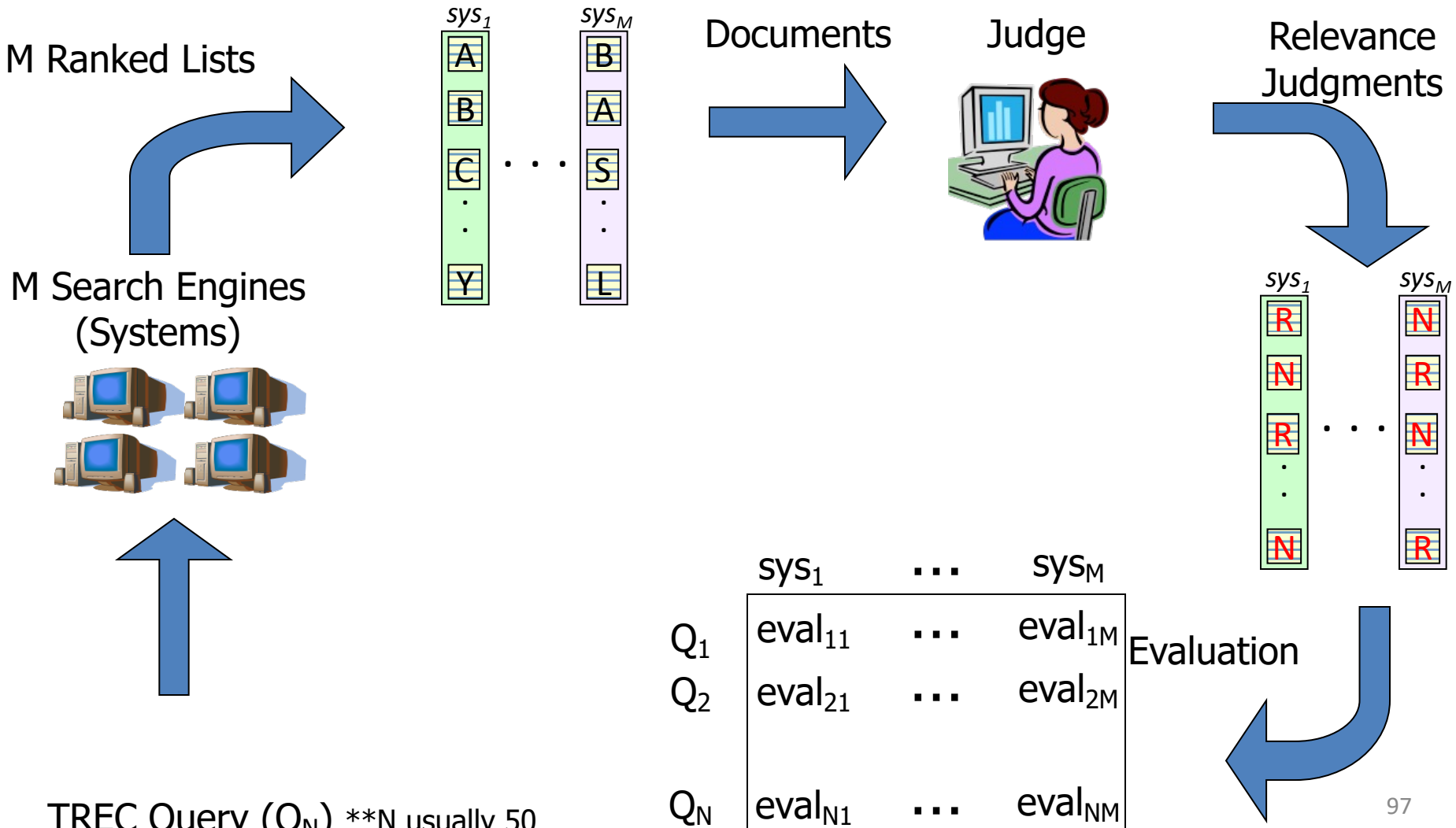
TREC Evaluation Setup



TREC Evaluation Setup



TREC Evaluation Setup



Conclusions

- Evaluation
  - Online/offline evaluation
  - Significance testing
- Test collection construction
- Reducing judgment effort