Evaluation of Information Retrieval Systems

Lecture by Shangsong Liang
Sun Yat-sen University

Thanks to Marti Hearst, Ray Larson, Chris Manning

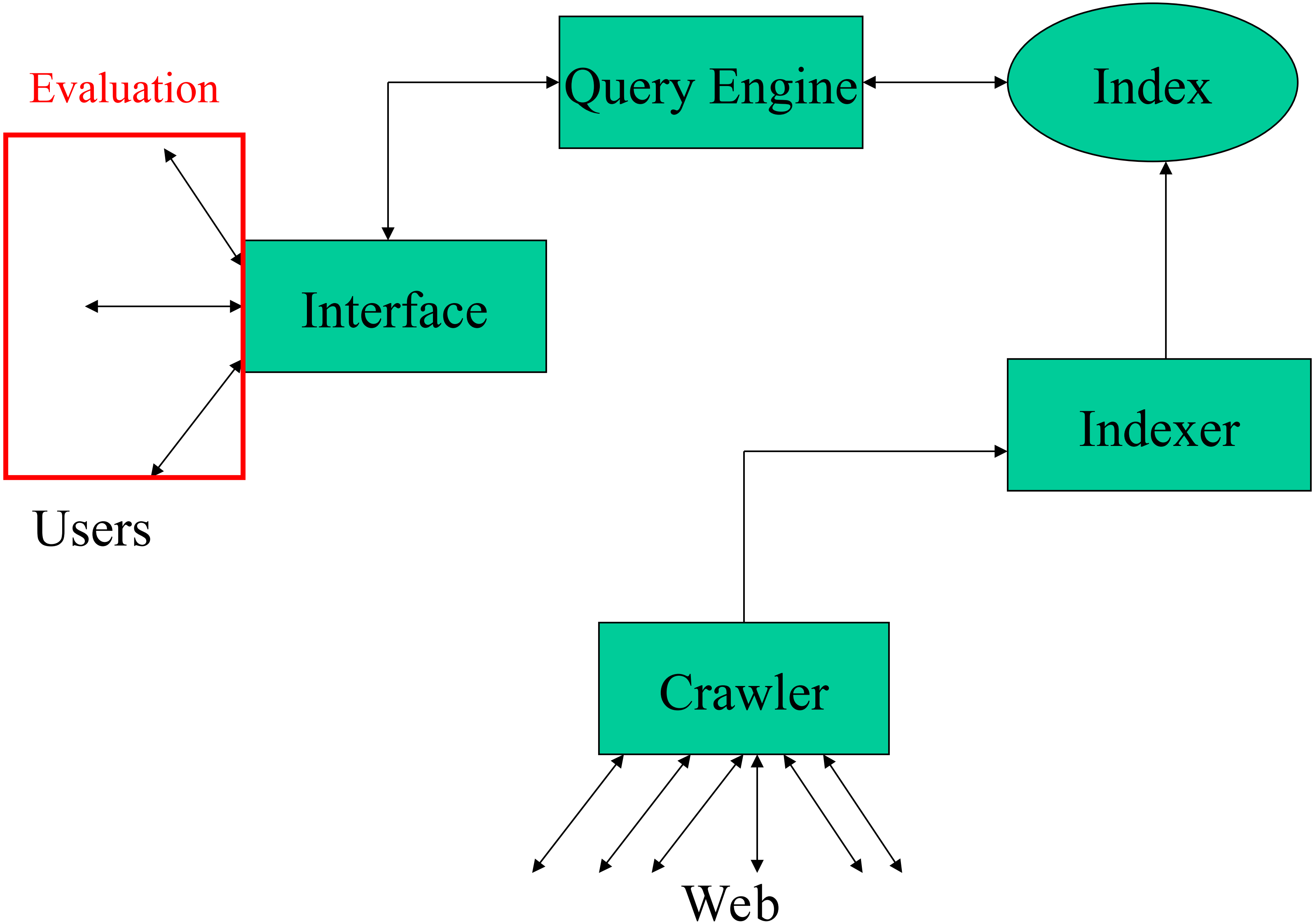
Today: Evaluation of IR Systems

- Performance evaluations
- Retrieval evaluation
- Quality of evaluation - Relevance
- Measurements of Evaluation
  - Precision vs recall
  - F number
  - others
- Test Collections/TREC

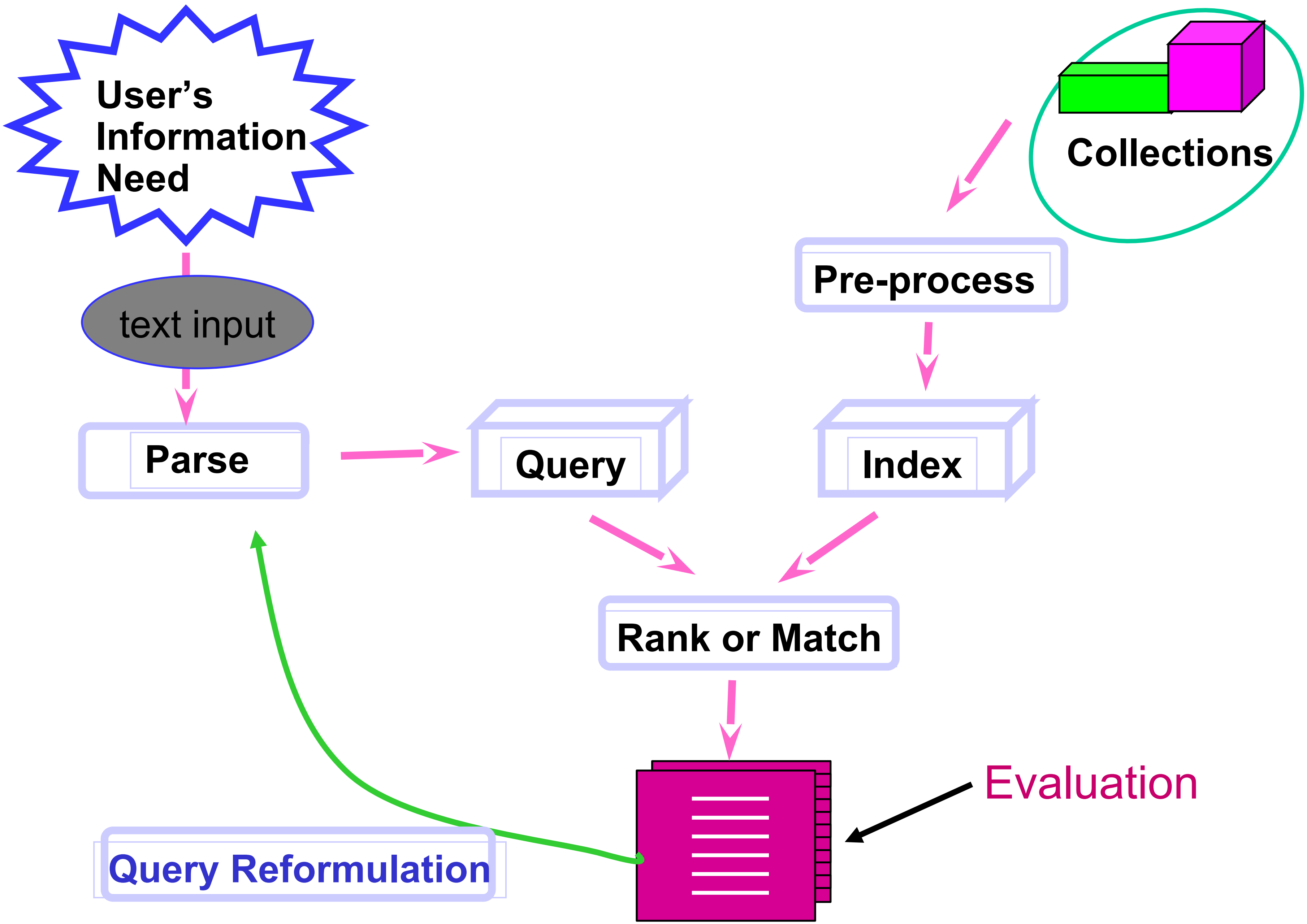
Performance of the IR or Search Engine

- Relevance
- Coverage
- Recency
- Functionality (e.g. query syntax)
- Speed
- Availability
- Usability
- Time/ability to satisfy user requests

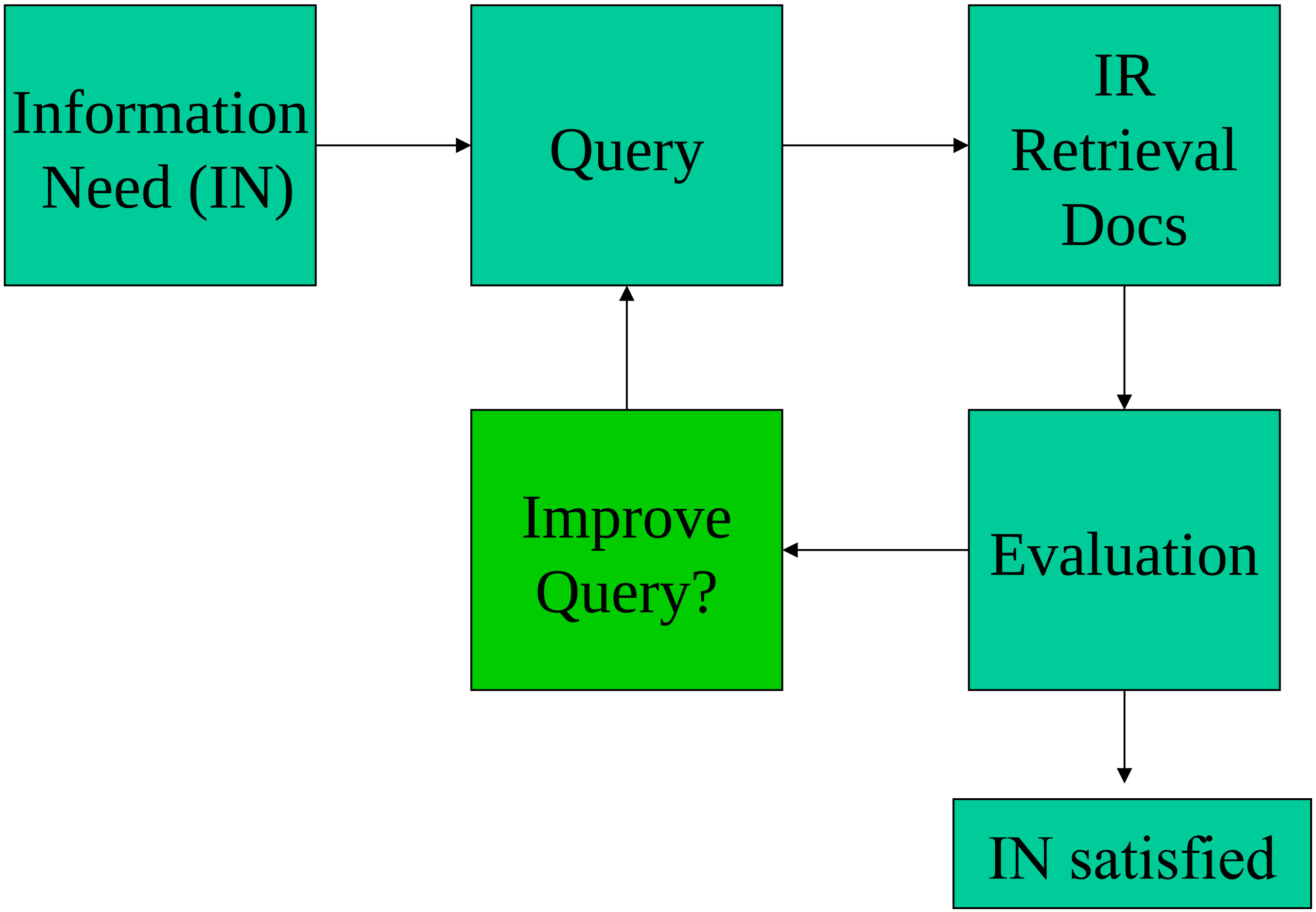
- *Basically "happiness"*



A Typical Web Search Engine


User's Information Need
Collections
text input
Pre-process
Parse
Query
Index
Rank or Match
Evaluation
Query Reformulation

Evaluation Workflow



What does the user want? Restaurant case

- The user wants to find a restaurant serving sashimi. User uses 2 IR systems. How we can say which one is better?

Evaluation

- Why Evaluate?
- What to Evaluate?
- How to Evaluate?

Why Evaluate?

- Determine if the system is useful
- Make comparative assessments with other methods/systems
  - Who’s the best?
- Test and improve systems
- Marketing
- Others?

What to Evaluate?

- How much of the information need is satisfied.
- How much was learned about a topic.
- Incidental learning:
  - How much was learned about the collection.
  - How much was learned about other topics.
- How easy the system is to use.
- *Usually based on what documents we retrieve*

Relevance as a Measure

Relevance is everything!

- How relevant is the document retrieved
  - for the user's information need.
- Subjective, but one assumes it's measurable
- Measurable to some extent
  - How often do people agree a document is relevant to a query
    - More often than expected
- How well does it answer the question?
  - Complete answer? Partial?
  - Background Information?
  - Hints for further exploration?

What to Evaluate?

What can be measured that reflects users' ability to use system? (Cleverdon 66)

- Coverage of Information
- Form of Presentation
- Effort required/Ease of Use
- Time and Space Efficiency
- Effectiveness

- Recall
  - proportion of **relevant** material actually retrieved
- Precision
  - proportion of **retrieved** material actually relevant

Effectiveness!

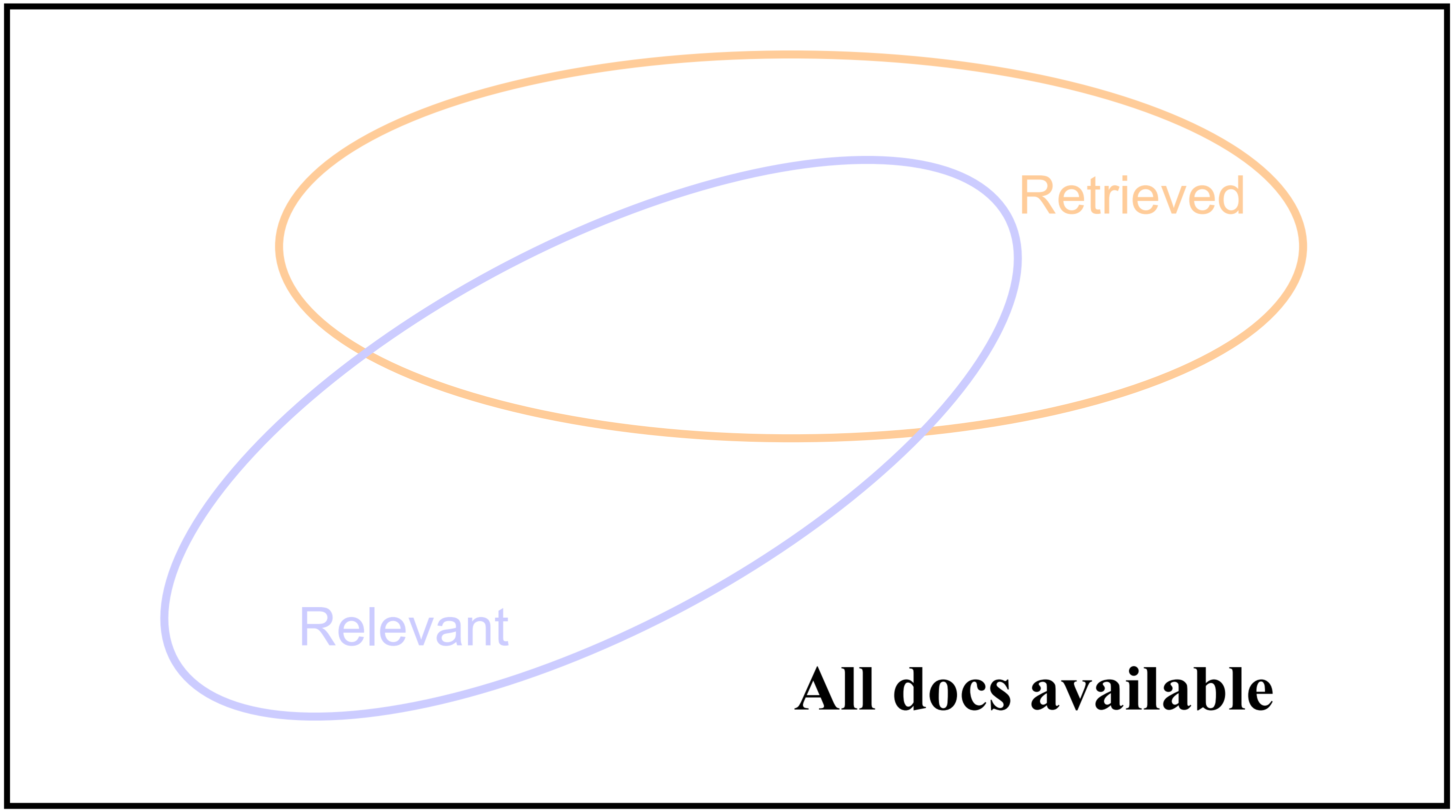
How do we measure relevance?

- Measures:
  - Binary measure
    - 1 relevant
    - 0 not relevant
  - N-nary measure
    - 3 very relevant
    - 2 relevant
    - 1 barely relevant
    - 0 not relevant
  - Negative values?
- N=? consistency vs. expressiveness tradeoff

Given: we have a relevance ranking of documents

- Have some known relevance evaluation
  - Query independent – based on information need
  - Experts (or you)
- Apply binary measure of relevance
  - 1 - relevant
  - 0 - not relevant
- Put in a query
  - Evaluate relevance of what is returned
- What comes back?
  - Example: lion

Relevant vs. Retrieved Documents



Set approach

Contingency table of relevant and retrieved documents

		relevant	
		Rel	NotRel
retrieved	Ret	Ret_{Rel}	Ret_{NotRel}
	NotRet	$NotRet_{Rel}$	$NotRet_{NotRel}$

$Ret = Ret_{Rel} + Ret_{NotRel}$

$NotRet = NotRet_{Rel} + NotRet_{NotRel}$

$Relevant = Ret_{Rel} + NotRet_{Rel}$ $\qquad$ $Not\ Relevant = Ret_{NotRel} + NotRet_{NotRel}$

Total # of documents available $N = Ret_{Rel} + NotRet_{Rel} + Ret_{NotRel} + NotRet_{NotRel}$

- Precision: $P = Ret_{Rel} / Retrieved$
- Recall: $R = Ret_{Rel} / Relevant$

$P = [0,1]$

$R = [0,1]$

Contingency table of classification of documents

Actual Condition

Test result		Present	Absent
	Positive	tp	fp type1
	Negative	fn type2	tn

fp type 1 error

fn type 2 error

present = tp + fn
positives = tp + fp
negatives = fn + tn

Total # of cases N = tp + fp + fn + tn

- False positive rate α = fp/(negatives)
- False negative rate β = fn/(positives)

Retrieval example

- Documents available:

  D1,D2,D3,D4,D5,D6, D7,D8,D9,D10
- Relevant: D1, D4, D5, D8, D10
- Query to search engine retrieves: D2, D4, D5, D6, D8, D9

	relevant	not relevant
retrieved		
not retrieved		

Example

- Documents available:

  D1,D2,D3,D4,D5,D6, D7,D8,D9,D10
- Relevant: D1, D4, D5, D8, D10
- Query to search engine retrieves: D2, D4, D5, D6, D8, D9

	relevant	not relevant
retrieved	D4,D5,D8	D2,D6,D9
not retrieved	D1,D10	D3,D7

Contingency table of relevant and retrieved documents

retrieved \ relevant	Rel	NotRel
Ret	$Ret_{Rel}=3$	$Ret_{NotRel}=3$
NotRet	$NotRet_{Rel}=2$	$NotRet_{NotRel}=2$

$Ret = Ret_{Rel} + Ret_{NotRel}$
$= 3 + 3 = 6$

$NotRet = NotRet_{Rel} + NotRet_{NotRe}$
$= 2 + 2 = 4$

$Relevant = Ret_{Rel} + NotRet_{Rel}$
$= 3 + 2 = 5$

$Not\ Relevant = Ret_{NotRel} + NotRet_{NotRel}$
$= 2 + 2 = 4$

Total # of docs $N = Ret_{Rel} + NotRet_{Rel} + Ret_{NotRel} + NotRet_{NotRel} = 10$

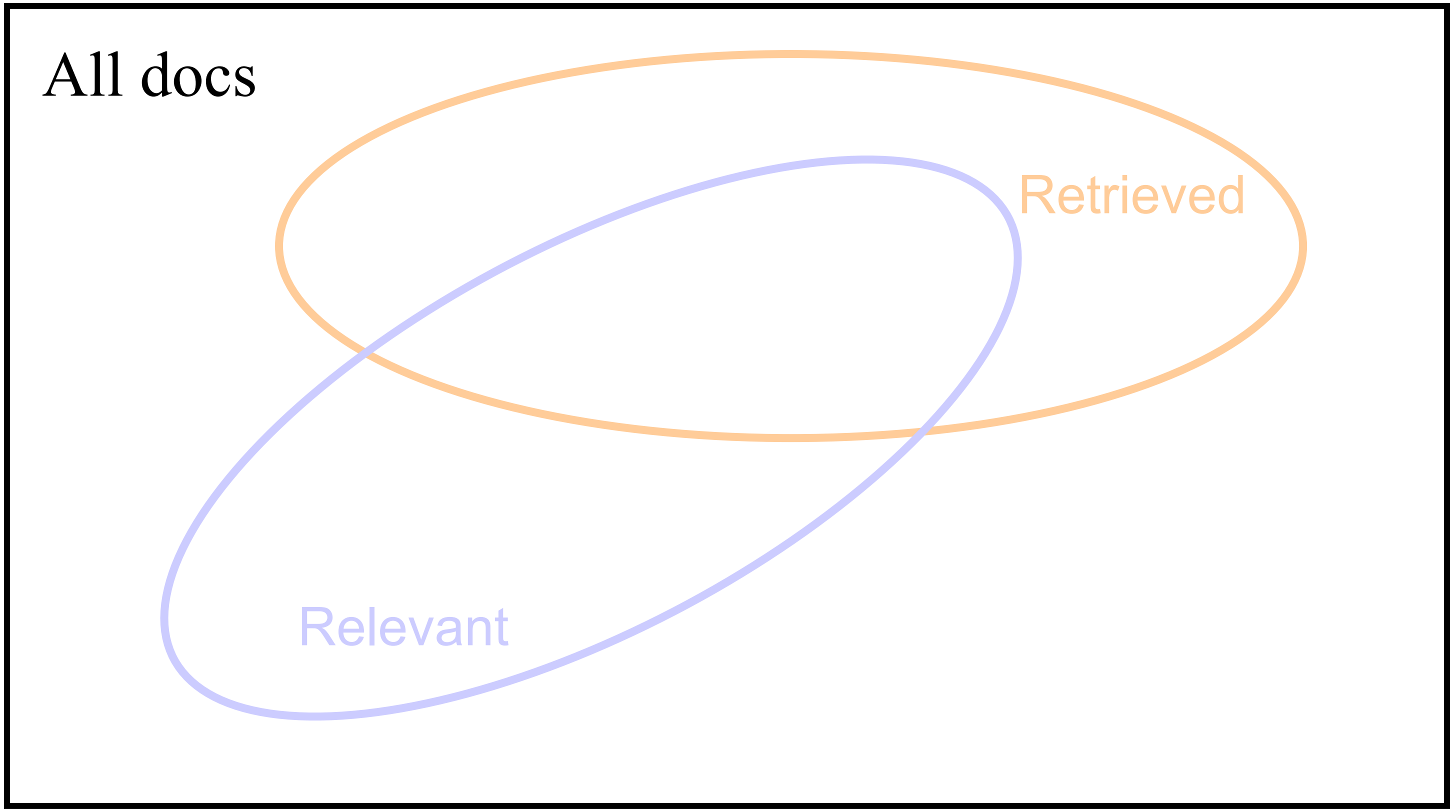
- Precision: $P = Ret_{Rel}$ / Retrieved = 3/6 = .5
- Recall: $R = Ret_{Rel}$ / Relevant = 3/5 = .6

$P = [0,1]$
$R = [0,1]$

What do we want

- Find everything relevant – high recall
- Only retrieve those – high precision

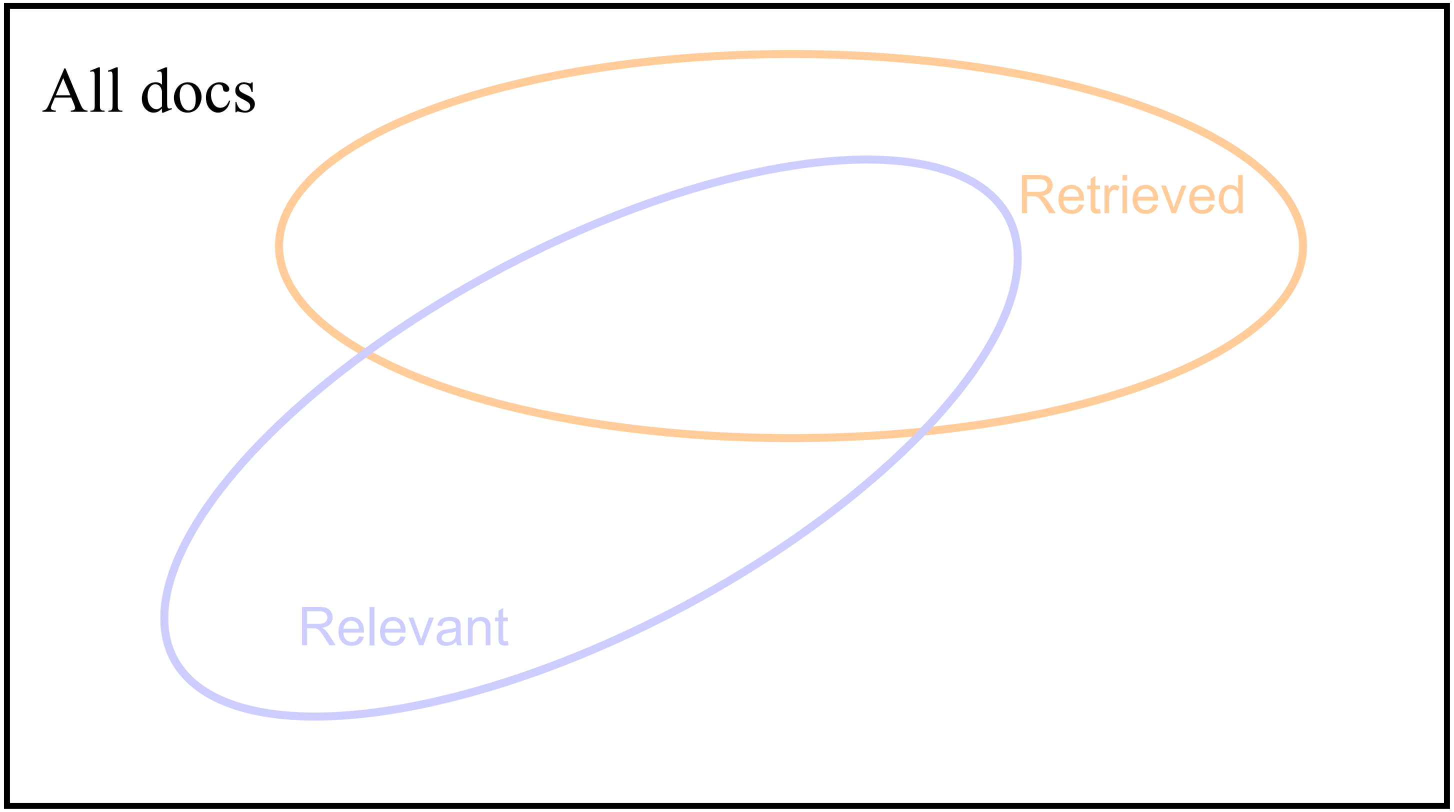
Relevant vs. Retrieved



Precision vs. Recall

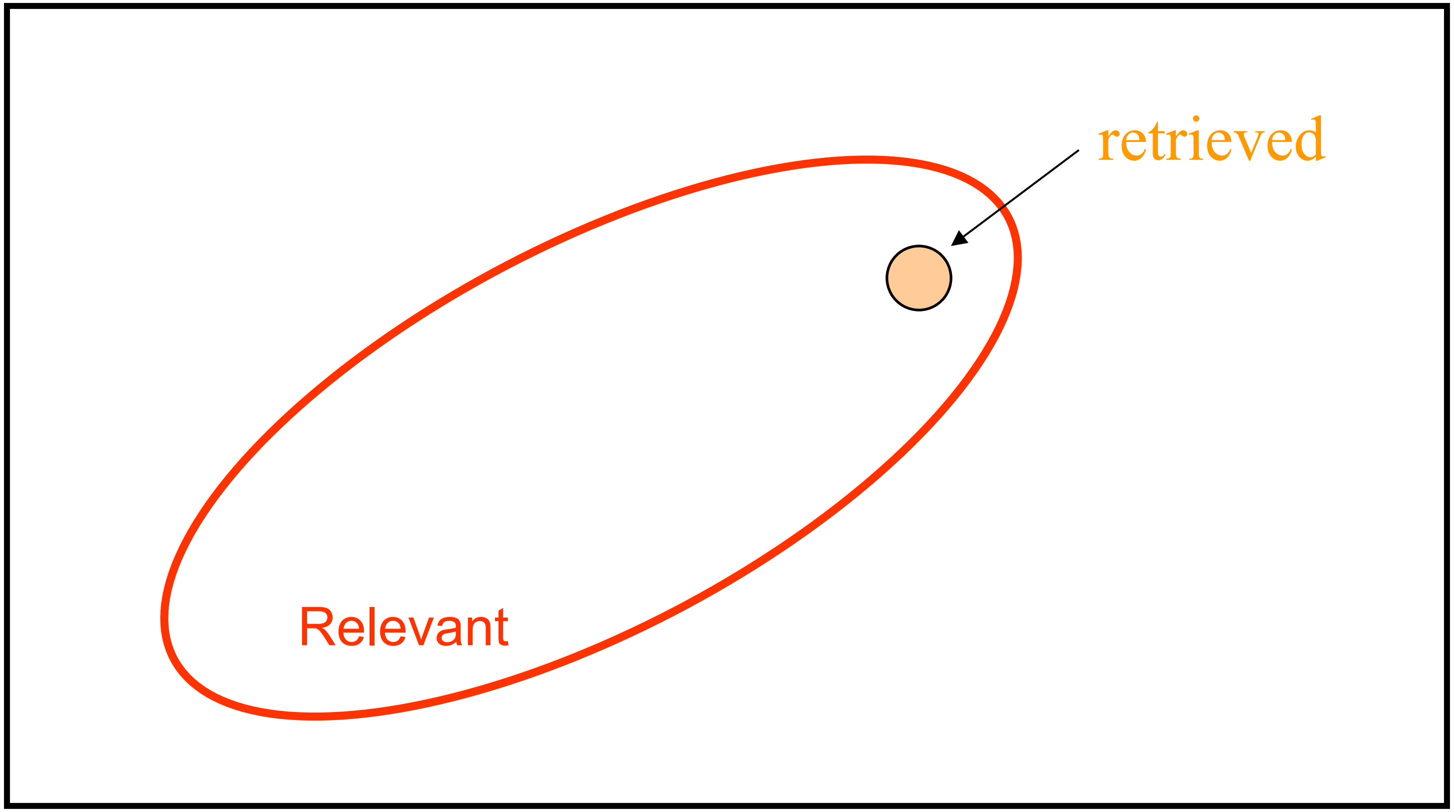
$$\text{Precision} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Retrieved}\,|}$$

$$\text{Recall} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Rel in Collection}\,|}$$



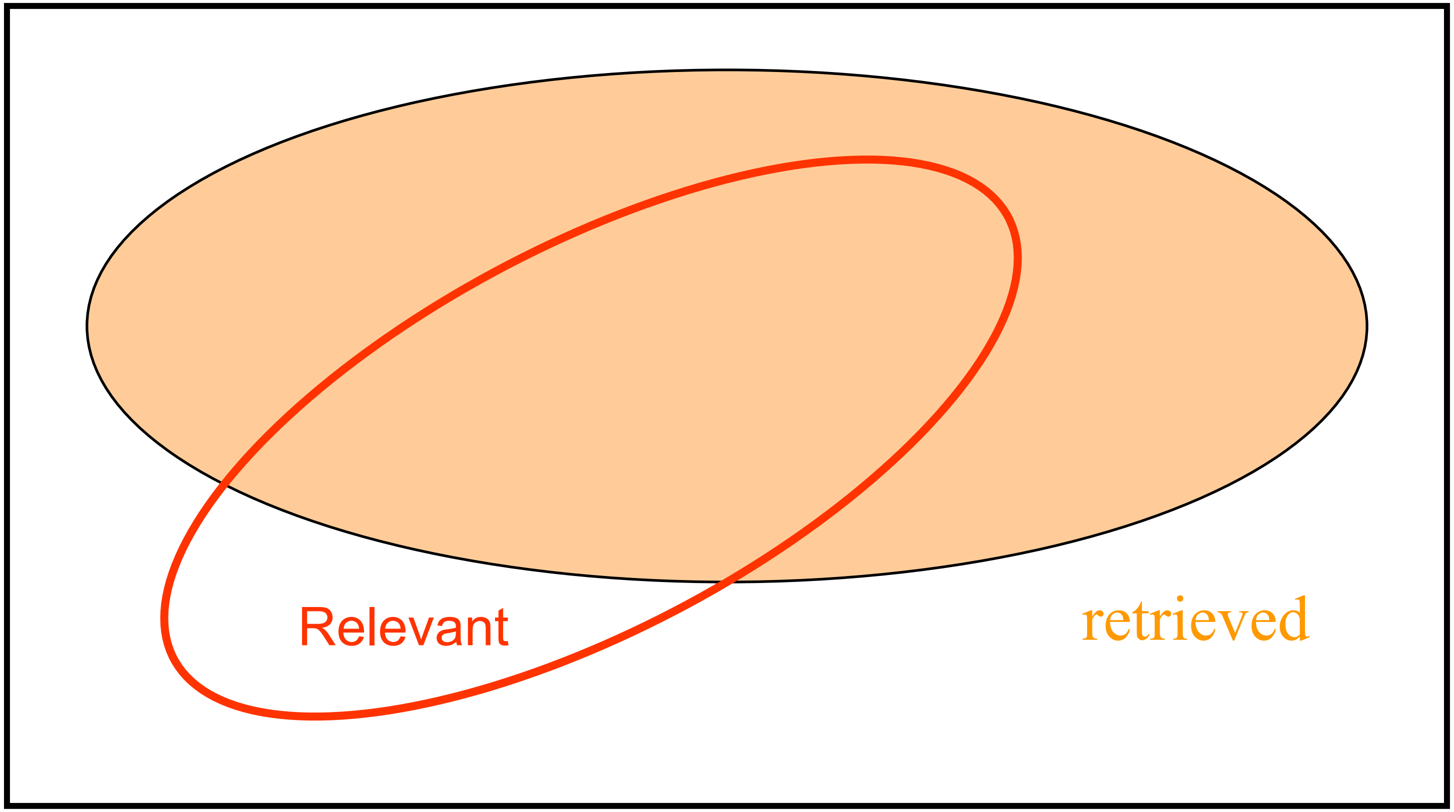
Retrieved vs. Relevant Documents

Very high precision, very low recall



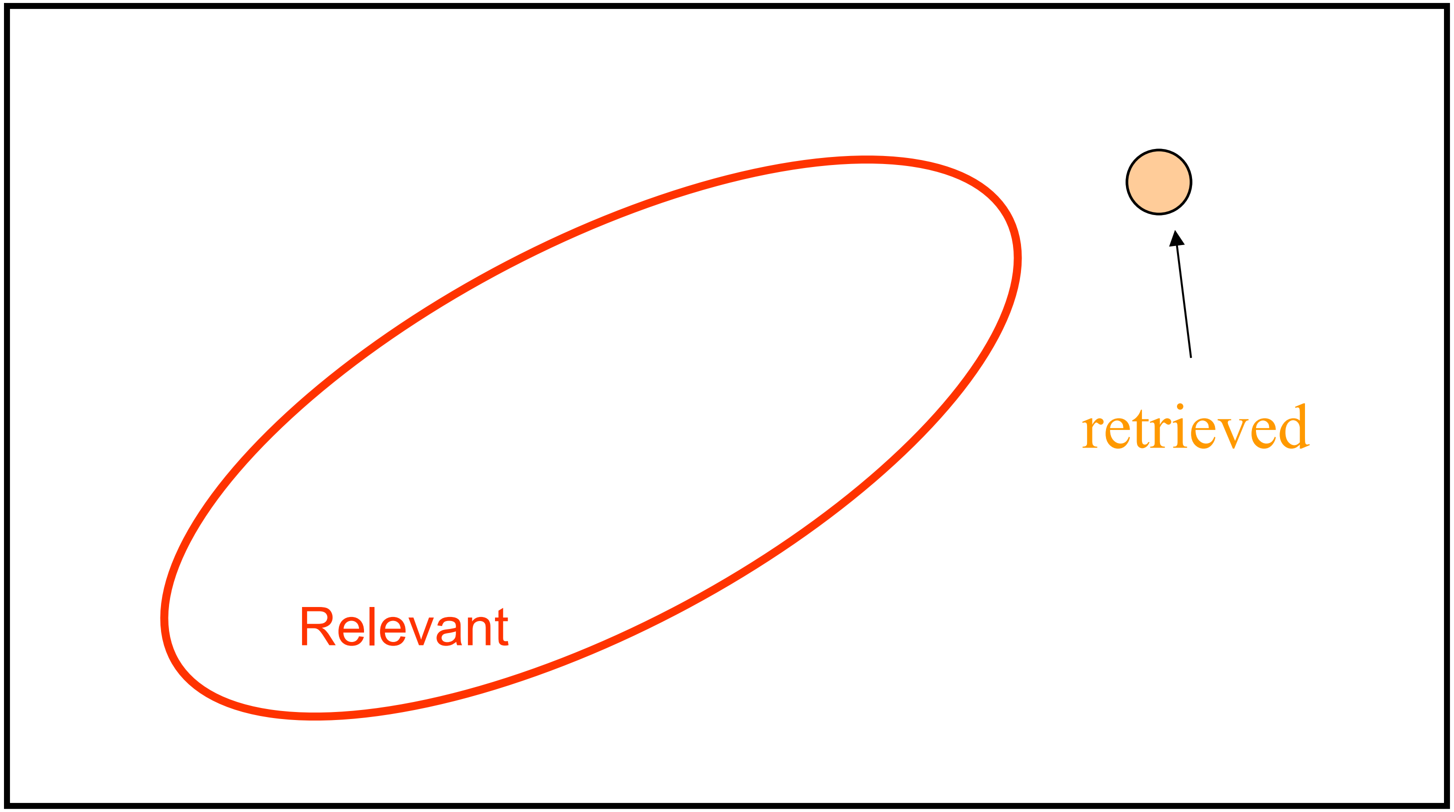
Retrieved vs. Relevant Documents

High recall, but low precision



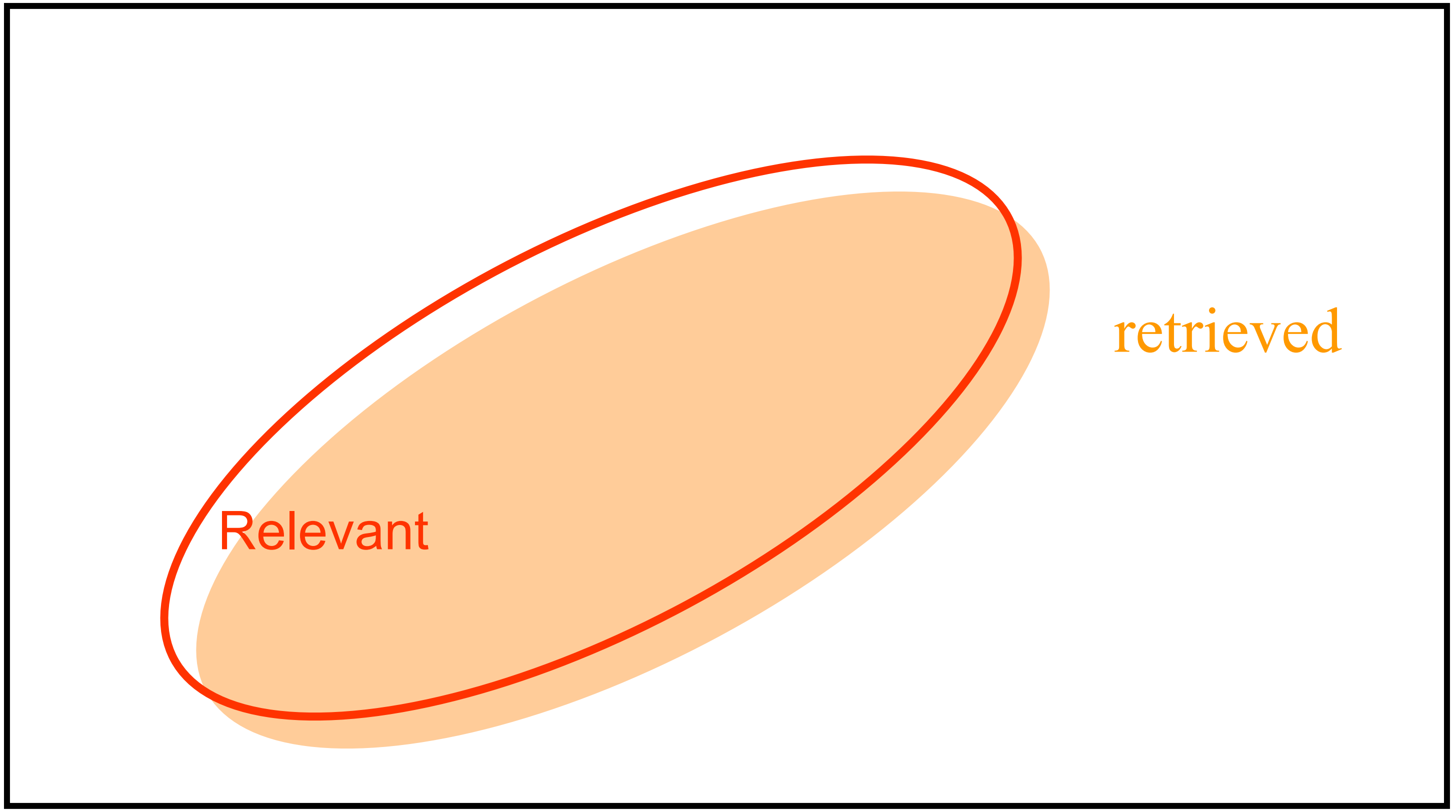
Retrieved vs. Relevant Documents

Very low precision, very low recall (0 for both)



Retrieved vs. Relevant Documents

High precision, high recall (at last!)



Why Precision and Recall?

Get as much of what we want while at the same time getting as little junk as possible.

Recall is the percentage of relevant documents returned compared to everything that is available!

Precision is the percentage of relevant documents compared to what is returned!

What different situations of recall and precision can we have?

Experimental Results

- Much of IR is experimental!
- Formal methods are lacking
  - Role of artificial intelligence
- Derive much insight from these results

Rec- recall NRel - # relevant
Prec - precision

Retrieve one document at a time without replacement and in order.

Given: **only** 25 documents of which 5 are relevant (D1, D2, D4, D15, D25)

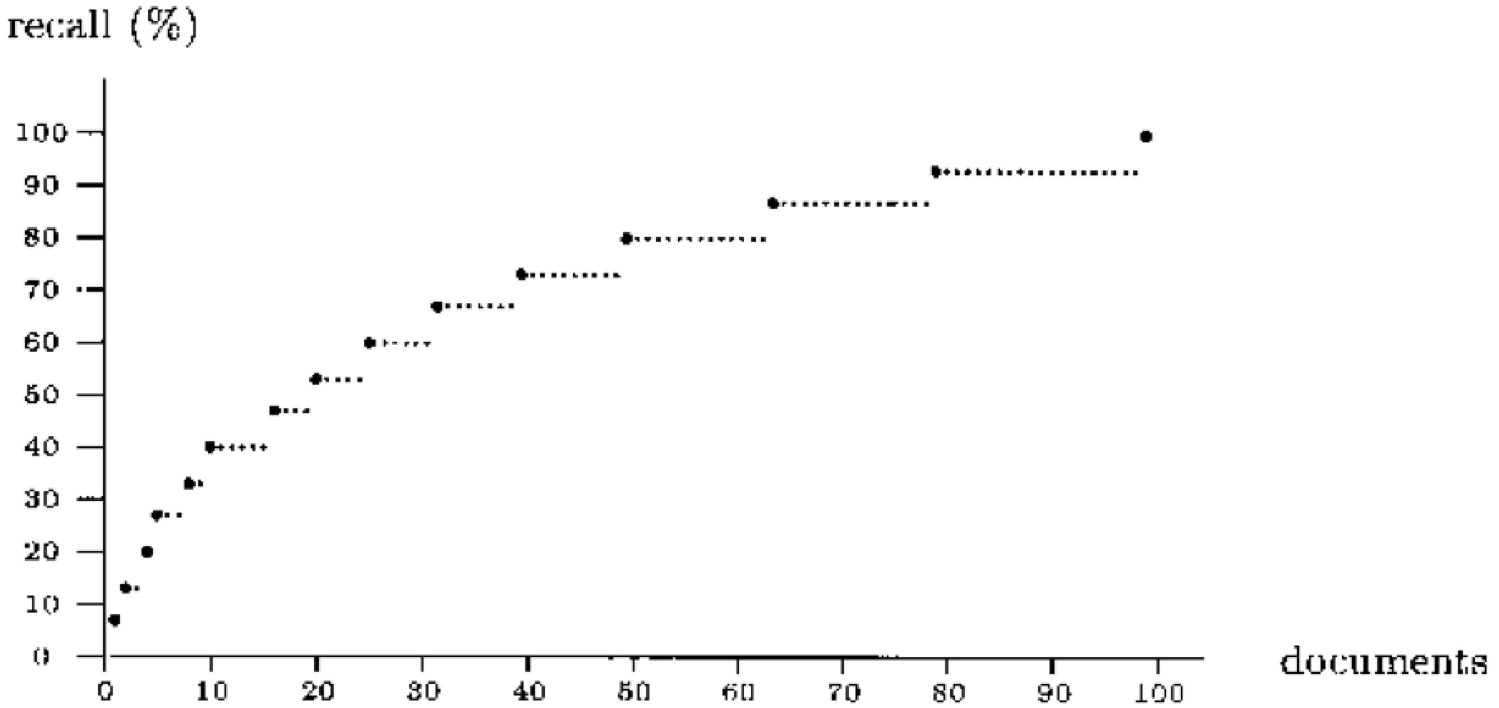
Calculate precision and recall after each document retrieved

Retrieve D1
    Have D1
Retrieve D2
    Have D1, D2
Retrieve D3
    Now have D1, D2, D3

	Rel?	NRel	Rec	Prec
1	1	1	0.20	1.00
2	1	2	0.40	1.00
3	0	2	0.40	0.67
4	1	3	0.60	0.75
5	0	3	0.60	0.60
6	0	3	0.60	0.50
7	0	3	0.60	0.43
8	0	3	0.60	0.38
9	0	3	0.60	0.33
10	0	3	0.60	0.30
11	0	3	0.60	0.27
12	0	3	0.60	0.25
13	0	3	0.60	0.23
14	0	3	0.60	0.21
15	1	4	0.80	0.27
16	0	4	0.80	0.25
17	0	4	0.80	0.24
18	0	4	0.80	0.22
19	0	4	0.80	0.21
20	0	4	0.80	0.20
21	0	4	0.80	0.19
22	0	4	0.80	0.18
23	0	4	0.80	0.17
24	0	4	0.80	0.17
25	1	5	1.00	0.20

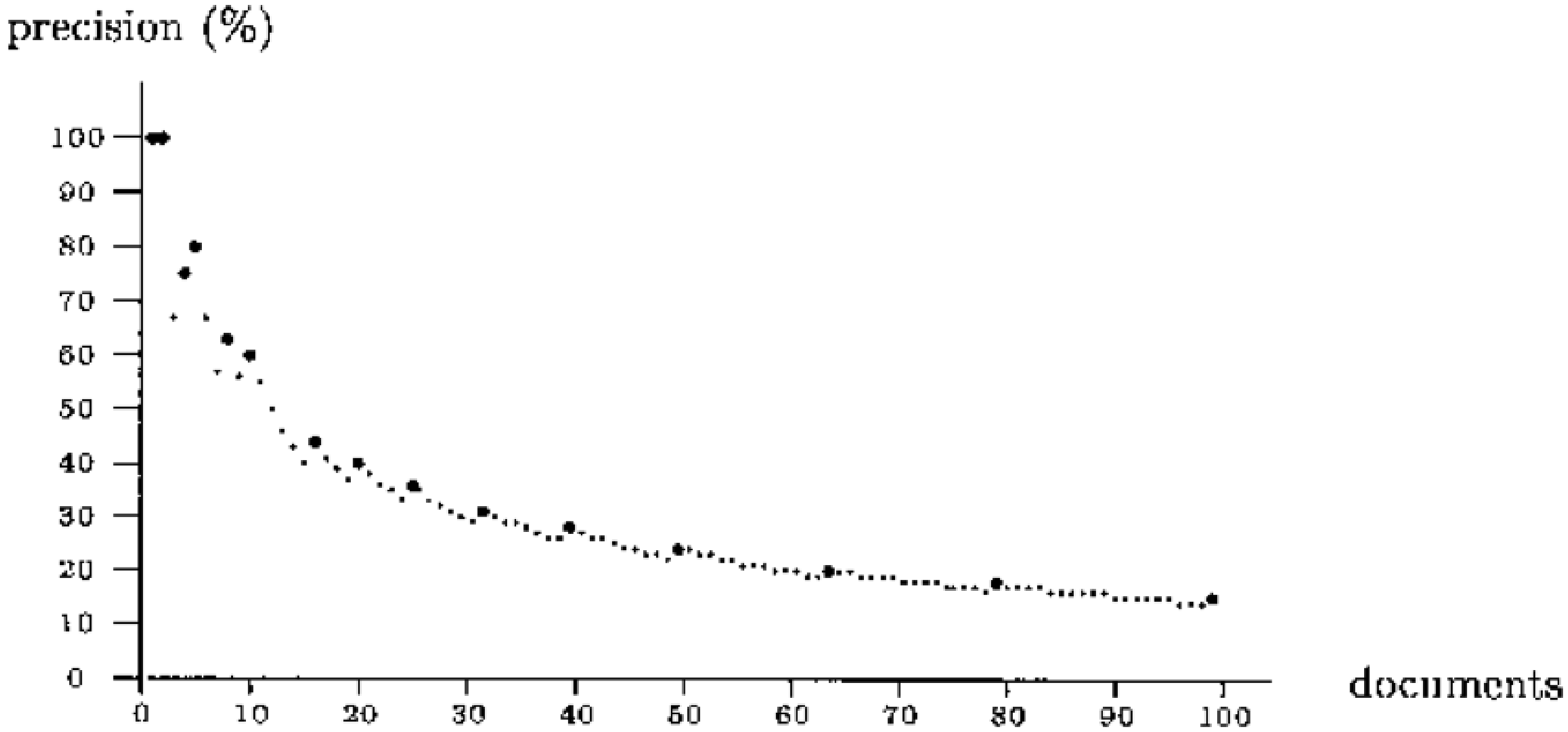
Recall Plot

- Recall when more and more documents are retrieved.
- Why this shape?



Precision Plot

- Precision when more and more documents are retrieved.
- Note shape!

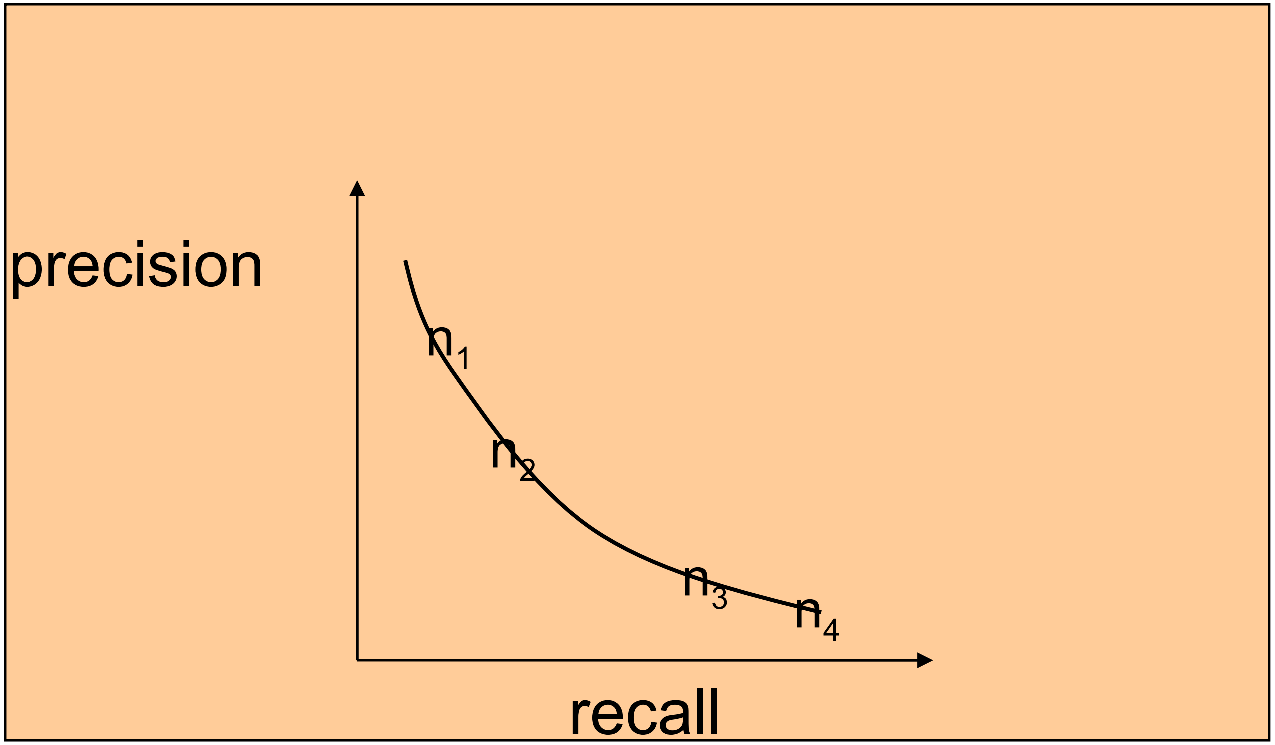


Precision/recall plot

- Sequences of points (p, r)
- Similar to y = 1 / x:
  - Inversely proportional!
  - Sawtooth shape - use smoothed graphs
- How we can compare systems?

Recall/Precision Curves

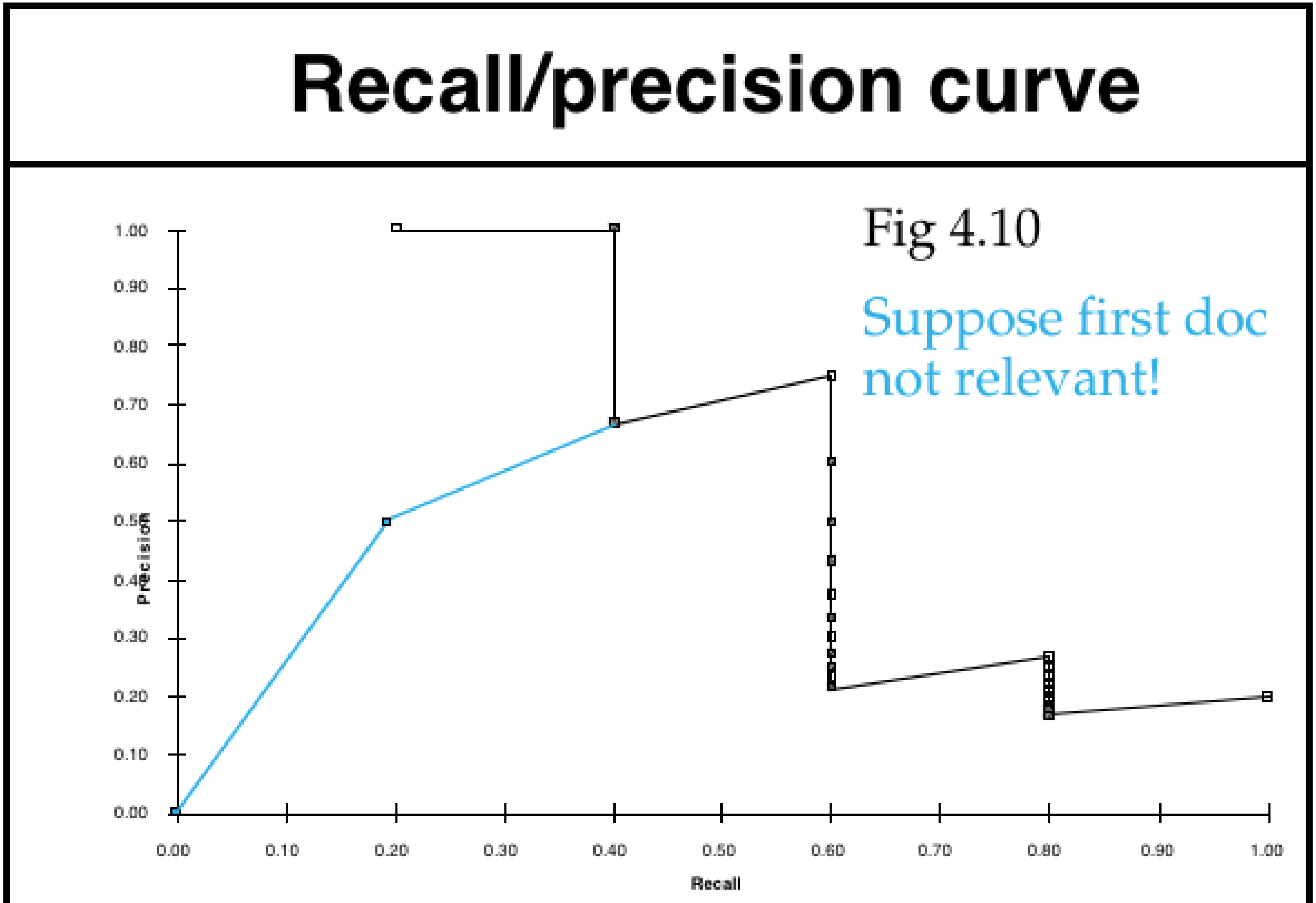
- There is a tradeoff between Precision and Recall
  - So measure Precision at different levels of Recall
- Note: this is usually an AVERAGE over MANY queries



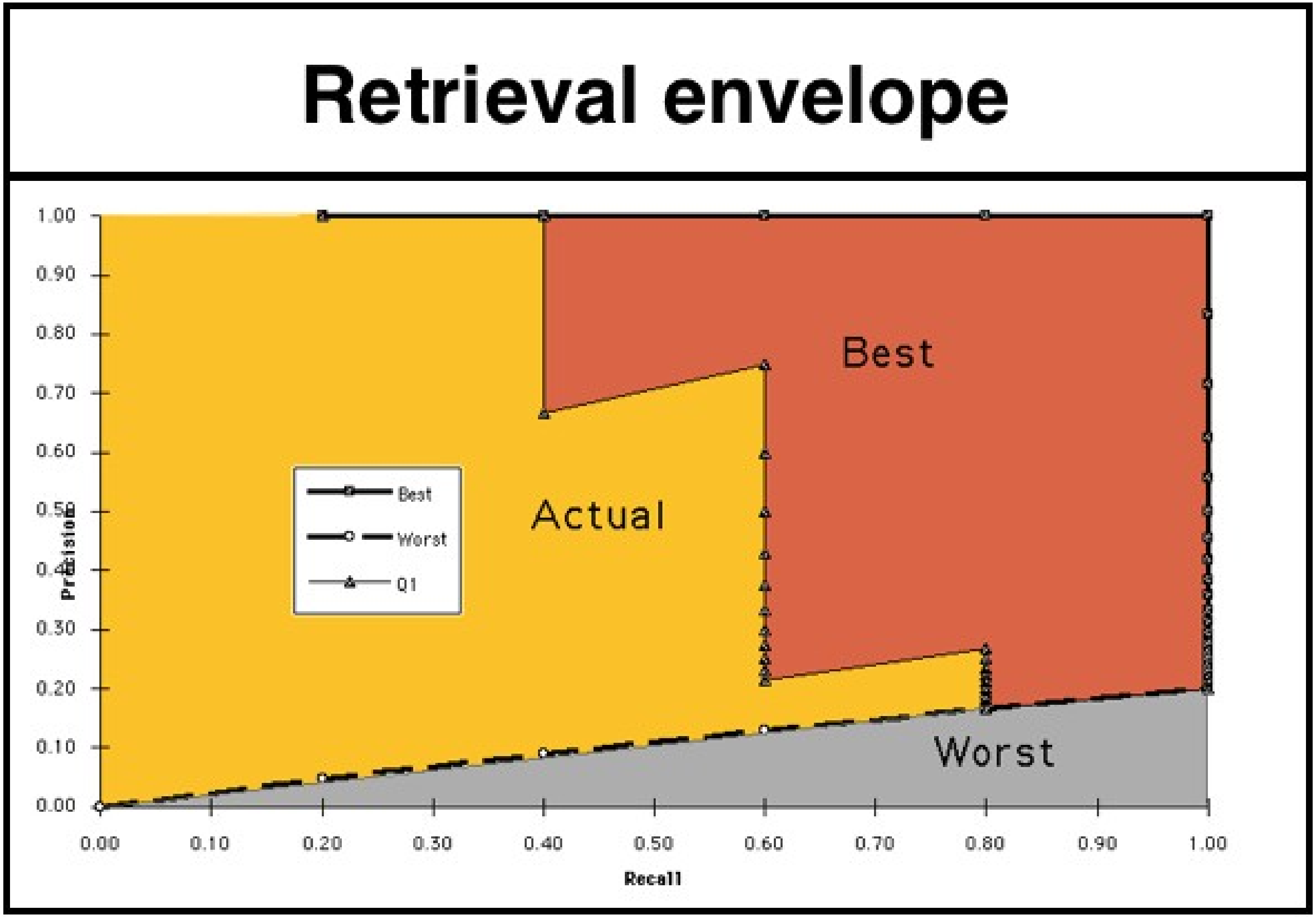
Note that there are two separate entities plotted on the x axis, recall and numbers of Documents.

n_i is number of documents retrieved, with $n_i < n_{i+1}$

Actual recall/precision curve for one query

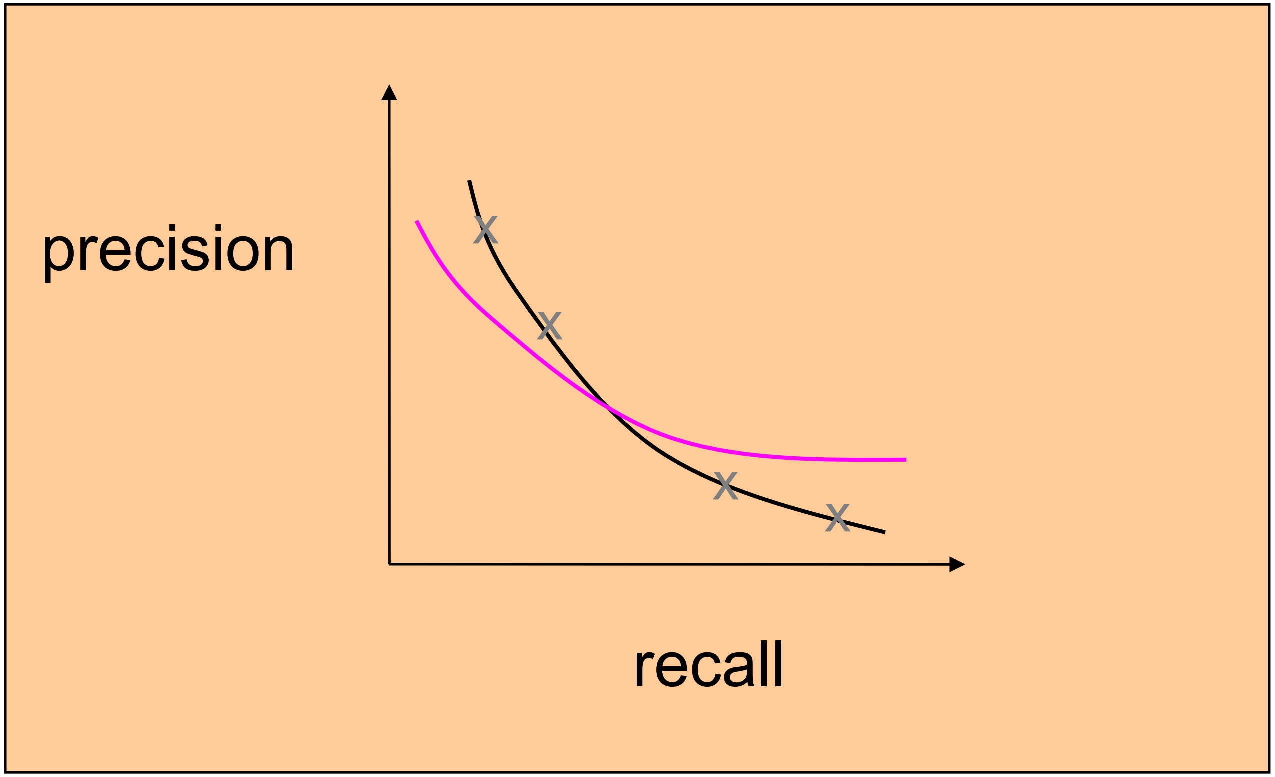


Best versus worst retrieval

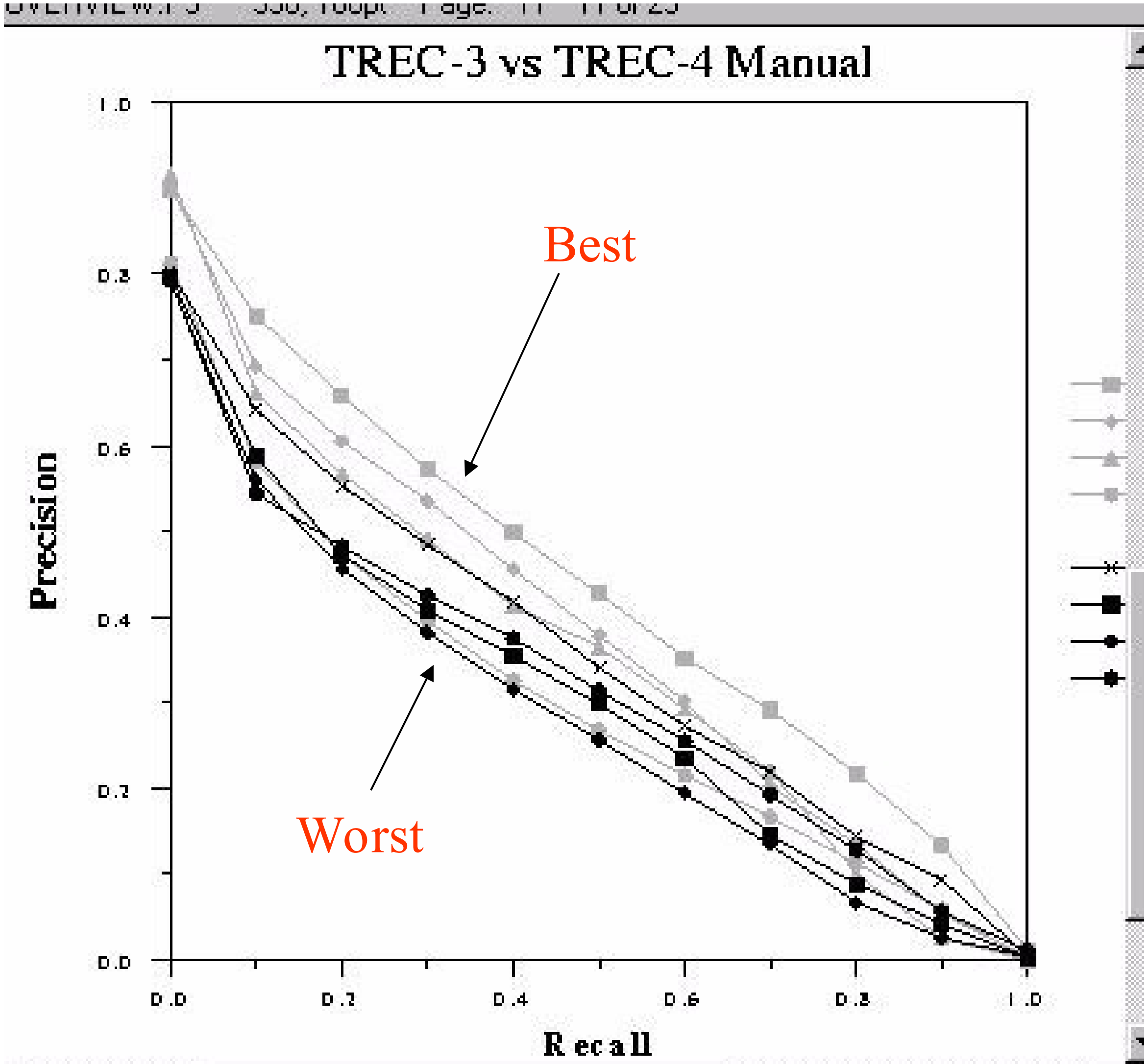


Precision/Recall Curves

- Sometimes difficult to determine which of these two hypothetical results is better:



Precision/Recall Curve Comparison



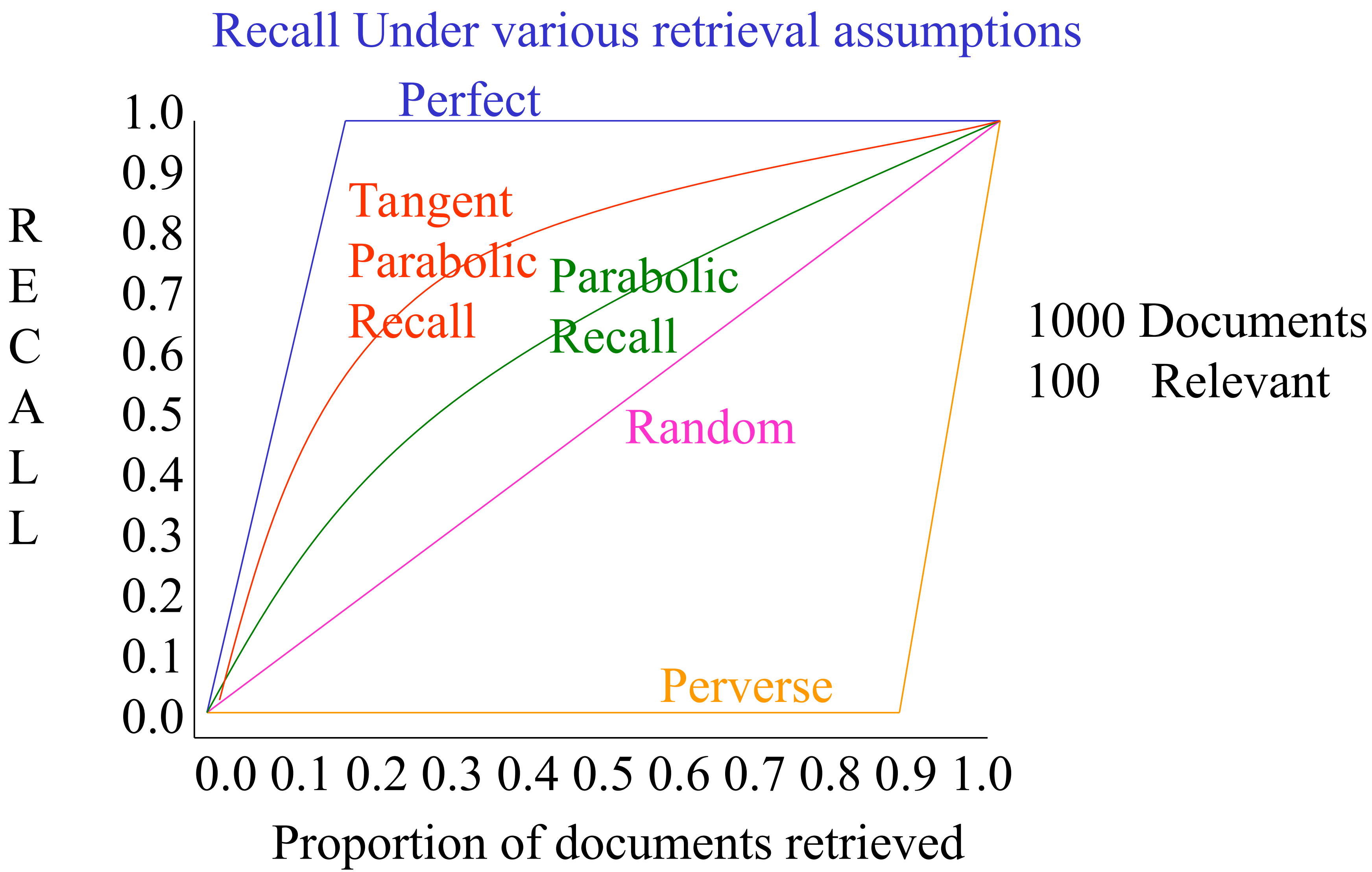
Document Cutoff Levels

- Another way to evaluate:
  - Fix the number of documents retrieved at several levels:
    - top 5
    - top 10
    - top 20
    - top 50
    - top 100
    - top 500
  - Measure precision at each of these levels
  - Take (weighted) average over results
- This is a way to focus on how well the system ranks the first k documents.

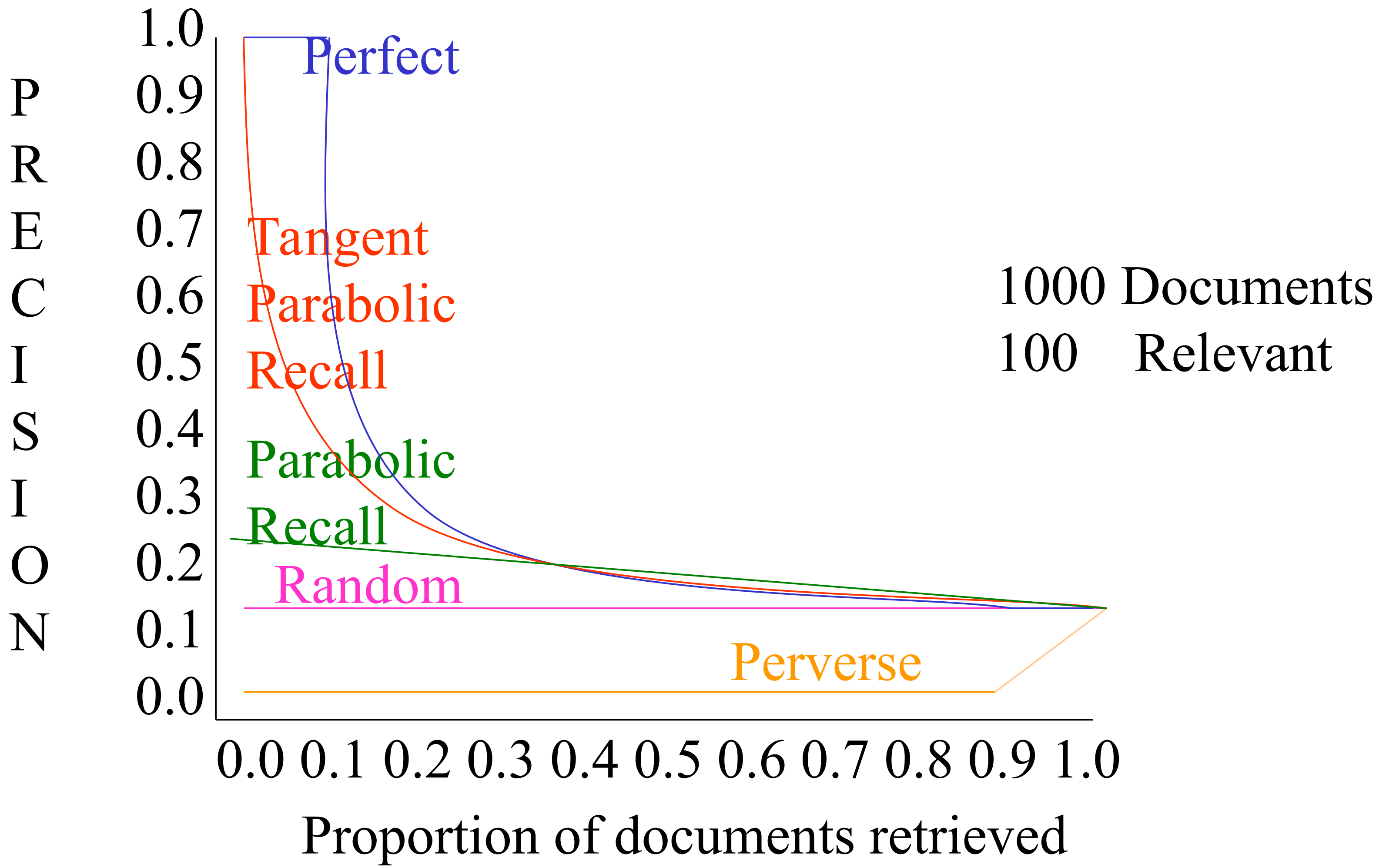
Problems with Precision/Recall

- Can't know true recall value (recall for the web?)
  - except in small collections
- Precision/Recall are related
  - A combined measure sometimes more appropriate
- Assumes batch mode
  - Interactive IR is important and has different criteria for successful searches
- Assumes a strict rank ordering matters.

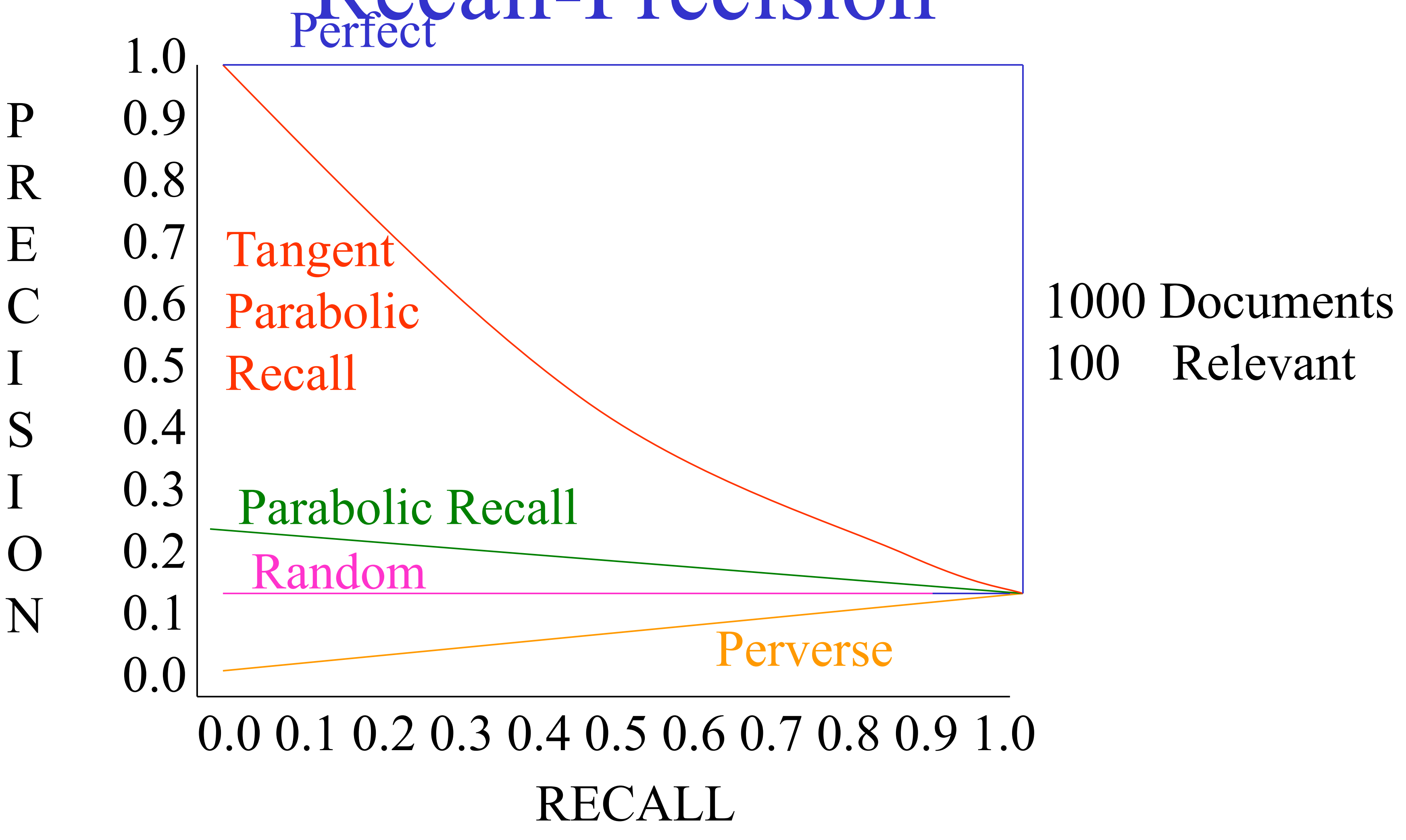
Buckland & Gey, JASIS: Jan 1994



Precision under various assumptions



Recall-Precision



Relation to Contingency Table

	Doc is Relevant	Doc is NOT relevant
Doc is retrieved	a	b
Doc is NOT retrieved	c	d

- Accuracy: (a+d) / (a+b+c+d)
- Precision: a/(a+b)
- Recall: a/(a+c)
- Why don't we use Accuracy for IR?
  - (Assuming a large collection)
    - Most docs aren't relevant
    - Most docs aren't retrieved
    - Inflates the accuracy value

The F-Measure

Combine Precision and Recall into one number

$$F = \frac{2}{1/R + 1/P} = 2\frac{RP}{R+P}$$

$P = precision$
$R = recall$

F = [0,1]
F = 1; when all ranked documents are relevant
F = 0; no relevant documents have been retrieved
Harmonic mean – average of rates

AKA
F_1 measure,
F-score

The E-Measure

Other ways to combine Precision and Recall into one number (van Rijsbergen 79)

$$E = 1 - \frac{1 + b^2}{\dfrac{b^2}{R} + \dfrac{1}{P}}$$

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

$$\alpha = 1/(\beta^2 + 1)$$

P = precision
R = recall
b = measure of relative importance of P or R

For example,
b = 0.5 means user is twice as interested in precision as recall

Interpret precision and recall

- Precision can be seen as a measure of exactness or fidelity
- Recall is a measure of completeness
- Inverse relationship between Precision and Recall, where it is possible to increase one at the cost of reducing the other.
  - For example, an information retrieval system (such as a search engine) can often increase its Recall by retrieving more documents, at the cost of increasing number of irrelevant documents retrieved (decreasing Precision).
  - Similarly, a classification system for deciding whether or not, say, a fruit is an orange, can achieve high Precision by only classifying fruits with the exact right shape and color as oranges, but at the cost of low Recall due to the number of false negatives from oranges that did not quite match the specification.

Measures for Large-Scale Eval

- Typical user behavior in web search systems has shown a preference for high precision
- Also graded scales of relevance seem more useful than just “yes/no”
- Measures have been devised to help evaluate situations taking these into account

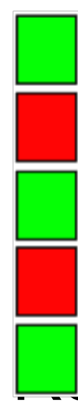
Rank-Based Measures

- Binary relevance
  - Precision@K (P@K)
  - Mean Average Precision (MAP)
  - Mean Reciprocal Rank (MRR)

- Multiple levels of relevance
  - Normalized Discounted Cumulative Gain (NDCG)

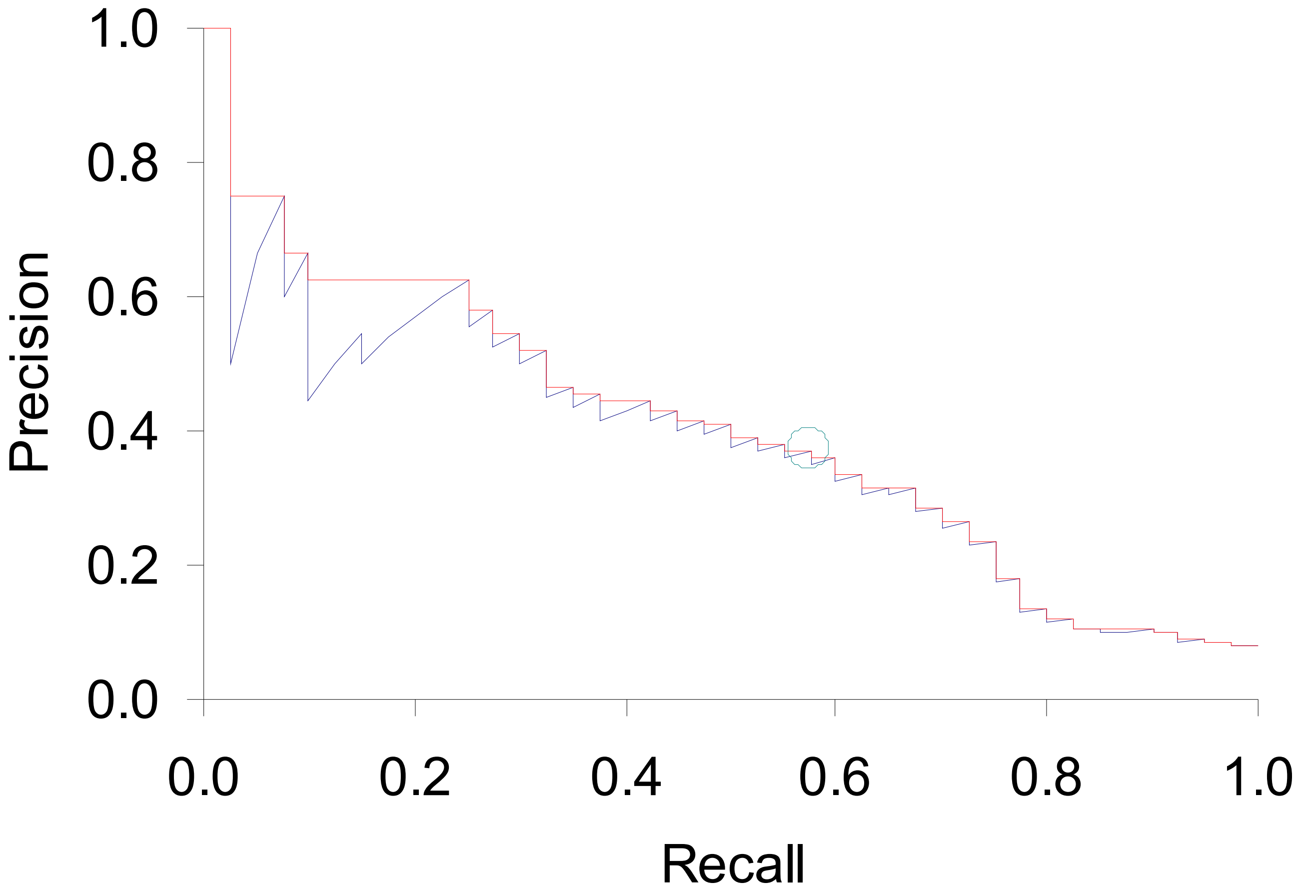
Precision@K

- Set a rank threshold K
- Compute % relevant in top K
- Ignores documents ranked lower than K
- Ex:
  – Prec@3 of 2/3
  – Prec@4 of 2/4
  – Prec@5 of 3/5



- In similar fashion we have Recall@K

A precision-recall curve

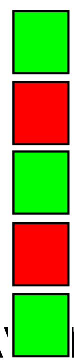


Mean Average Precision (MAP)

- Consider rank position of each ***relevant*** doc
  - K_1, K_2, … K_R

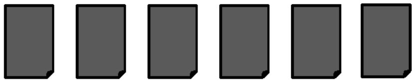
- Compute Precision@K for each K_1, K_2, … K_R

- Average precision = average of P@K

- Ex:  has AvgPrec of $$\frac{1}{3} \cdot \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$$

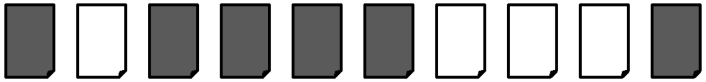
- MAP is Average Precision across multiple queries/rankings

Average Precision



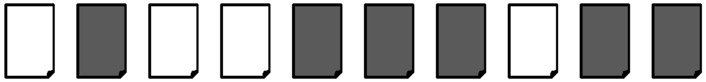
= the relevant documents

Ranking #1



Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2

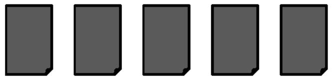


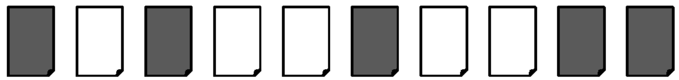
Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

$$\text{Ranking \#1: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

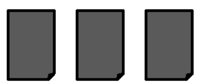
$$\text{Ranking \#2: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

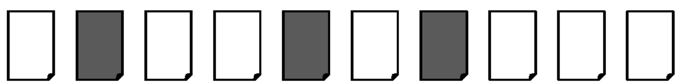
MAP


= relevant documents for query 1

Ranking #1


Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5


= relevant documents for query 2

Ranking #2


Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

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

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

Mean average precision

- If a relevant document never gets retrieved, we assume the precision corresponding to that relevant doc to be zero
- MAP is macro-averaging: each query counts equally
- Now perhaps most commonly used measure in research papers
- Good for web search?
- MAP assumes user is interested in finding many relevant documents for each query
- MAP requires many relevance judgments in text collection

Beyond binary relevance

YAHOO!

Web Images Video Local Shopping More

Toyota safety

Search

Options

Search Pad

SearchScan - On

108,000,000 results for
Toyota safety:

- **Show All**
- Toyota
- Motor Trend
- CarsDirect
- Shopping Sites

Also try: **toyota safety** ratings, **toyota safety** recall, More...

Sponsored Results

Toyota Recall
Toyota Takes Care of its Customers. Read the FAQs at **Toyota**.com.
www.**Toyota.com**/Recall

Toyota Safety
& Latest Prices. Free Info. **Toyota** Research, Reviews.
www.**Toyota.Edmunds.com**

fair

TOYOTA | Car **Safety** Innovation and Technology
Toyota home page for car **safety** and car technology Prius model.
www.**safetytoyota.com** - Cached

fair

Toyota home page for car **safety** and car technology ...
We are presenting **Toyota's safety** technologies for cars. We clearly explain about car **safety** and car technology using movies and more.
www.**safetytoyota.com**/en-gb - Cached

Good

Toyota Safety Ratings - **Toyota Safety** Features - Motor Trend **...**
MotorTrend offers **Toyota safety** ratings, comprehensive auto **safety** reports, and more. View a all of the standard **Toyota safety** features. **...**
motortrend.com/new_cars/07/**toyota**/**safety**_ratings/index.html - 149k - Cached

Toyota Motor Europe Corporate Site **Safety**
Our approach. **Toyota** believes that all stakeholders in the road **safety** equation share a responsibility to reduce the frequency of road accidents. **...**
www.**toyota.eu**/**Safety** - Cached

[PDF] pdf European **Safety** Brochure 2005
4047k - Adobe PDF - View as html
not guarantee that all accidents or injuries will be avoided when driving a **Toyota** and/or Lexus brand motor vehicle equipped with the **safety** systems **...**
www.**toyota.no**/Images/**Safety**_Brochure_tcm308-344461.pdf

Toyota - Star **Safety** System
Star **Safety** System **...** **Toyota** Mobility Program. Careers. Contact Us. Home. contact us. site map. your privacy rights. legal terms. **Toyota** Newsroom. sign up for info **...**
www.**toyota.com**/vehicles/demos/star-**safety**.html - 58k - Cached

Toyota Prius **Safety** Ratings - CarsDirect
Get overall **safety** ratings and NHTSA crash test results for the **Toyota** Prius at CarsDirect.

Sponsored Results

Safety for a **Toyota**
Research **Safety** Ratings and Reviews For New Car at Kelley Blue Book.
www.**kbb.com**

Toyota Safety
Find **Toyota Safety** dealers, new cars, prices, and photos.
www.**NewCars.org**

Toyota Safety
Toyota safety Discount Prices Save Money Shopping Online Today.
www.**smarter.com**

Saftey Toyoto
Explore 5,000+ Pro Sports Choices. Save On Saftey Toyoto.
BaseballGear.Shopzilla.com

See your message here...

Discounted Cumulative Gain (DCG)

- Popular measure for evaluating web search and related tasks

- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant documents
  - 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

- Uses *graded relevance* as a measure of 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

Summarize a Ranking: DCG

- What if relevance judgments are in a scale of [0,r]? r>2
- Cumulative Gain (CG) at rank n
  - Let the ratings of the n documents be r_1, r_2, …r_n (in ranked order)
  - CG = $r_1+r_2+\ldots r_n$
- Discounted Cumulative Gain (DCG) at rank n
  - DCG = $r_1 + r_2/\log_2 2 + r_3/\log_2 3 + \ldots\ r_n/\log_2 n$
    - We may use any base for the logarithm

Discounted Cumulative Gain

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

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

- Alternative formulation:

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

  - used [illegible] npanies
  - emphasis on retrieving *highly* relevant documents

DCG Example

- 10 ranked documents judged on 0-3 relevance scale:

  3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- discounted gain:

  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- DCG:

  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

Summarize a Ranking: NDCG

- Normalized Discounted Cumulative Gain (NDCG) at rank *n*
  - Normalize DCG at rank *n* by the DCG value at rank *n* of the ideal ranking
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc
- Normalization useful for contrasting queries with varying numbers of relevant results

- NDCG is now quite popular in evaluating Web search

NDCG – Example 1

4 documents: d_1, d_2, d_3, d_4

i	Ground Truth		Ranking Function$_1$		Ranking Function$_2$	
	Document Order	r_i	Document Order	r_i	Document Order	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	$NDCG_{GT}$=1.00		$NDCG_{RF1}$=1.00		$NDCG_{RF2}$=0.9203	

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

$$MaxDCG = DCG_{GT} = 4.6309$$

NDCG – Example 2

- For the documents ordered by the ranking algorithm as

$$D_1, D_2, D_3, D_4, D_5, D_6$$

- The user provides the following relevance scores:
  - 3, 2, 3, 0, 1, 2
- The cumulative Gain of this search result list is:

$$\mathrm{CG}_6 = \sum_{i=1}^{6} rel_i = 3 + 2 + 3 + 0 + 1 + 2 = 11$$

NDCG – Example 2

i	rel_i	$\log_2(i+1)$	$\frac{rel_i}{\log_2(i+1)}$
1	3	1	3
2	2	1.585	1.262
3	3	2	1.5
4	0	2.322	0
5	1	2.585	0.387
6	2	2.807	0.712

So the DCG_6 of this ranking is:

$$\mathrm{DCG_6} = \sum_{i=1}^{6} \frac{rel_i}{\log_2(i+1)} = 3 + 1.262 + 1.5 + 0 + 0.387 + 0.712 = 6.861$$

NDCG – Example 2

- The ideal ordering is:
  - 3, 3, 3, 2, 2, 2, 1, 0

The DCG of this ideal ordering, or *IDCG (Ideal DCG)* , is computed to rank 6:

$$\mathrm{IDCG_6} = 8.740$$

And so the nDCG for this query is given as:

$$\mathrm{nDCG_6} = \frac{DCG_6}{IDCG_6} = \frac{6.861}{8.740} = 0.785$$

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- ROUGE-N: Overlap of N-grams[3] between the system and reference summaries.
  - ROUGE-1 refers to the overlap of ***1-gram*** *(each word)* between the system and reference summaries.
  - ROUGE-2 refers to the overlap of ***bigrams*** between the system and reference summaries.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Let's take the example from above. Let us say we want to compute the **ROUGE-2 precision and recall** scores.

System Summary:

```
the cat was found under the bed
```

Reference Summary:

```
the cat was under the bed
```

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

System Summary Bigrams:

```
the cat, cat was, was found, found under, under the, the bed
```

Reference Summary Bigrams:

```
the cat, cat was, was under, under the, the bed
```

Based on the bigrams above, the ROUGE-2 recall is as follows:

$$ROUGE2_{Recall} = \frac{4}{5} = 0.8$$

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Based on the bigrams above, the ROUGE-2 recall is as follows:

$$ROUGE2_{Recall} = \frac{4}{5} = 0.8$$

Essentially, the system summary has recovered 4 bigrams out of 5 bigrams from the reference summary, which is pretty good! Now the ROUGE-2 precision is as follows:

$$ROUGE2_{Precision} = \frac{4}{6} = 0.67$$

What if the results are not in a list?

- Suppose there's only one Relevant Document
- Scenarios:
  - known-item search
  - navigational queries
  - looking for a fact
- Search duration ~ Rank of the answer
  - measures a user's effort

Mean Reciprocal Rank（平均倒数排名）

- Consider rank position, K, of first relevant doc
  - Could be – only clicked doc

- Reciprocal Rank score = $\frac{1}{K}$

- MRR is the mean RR across multiple queries

How to Evaluate IR Systems?
Test Collections

Test Collections

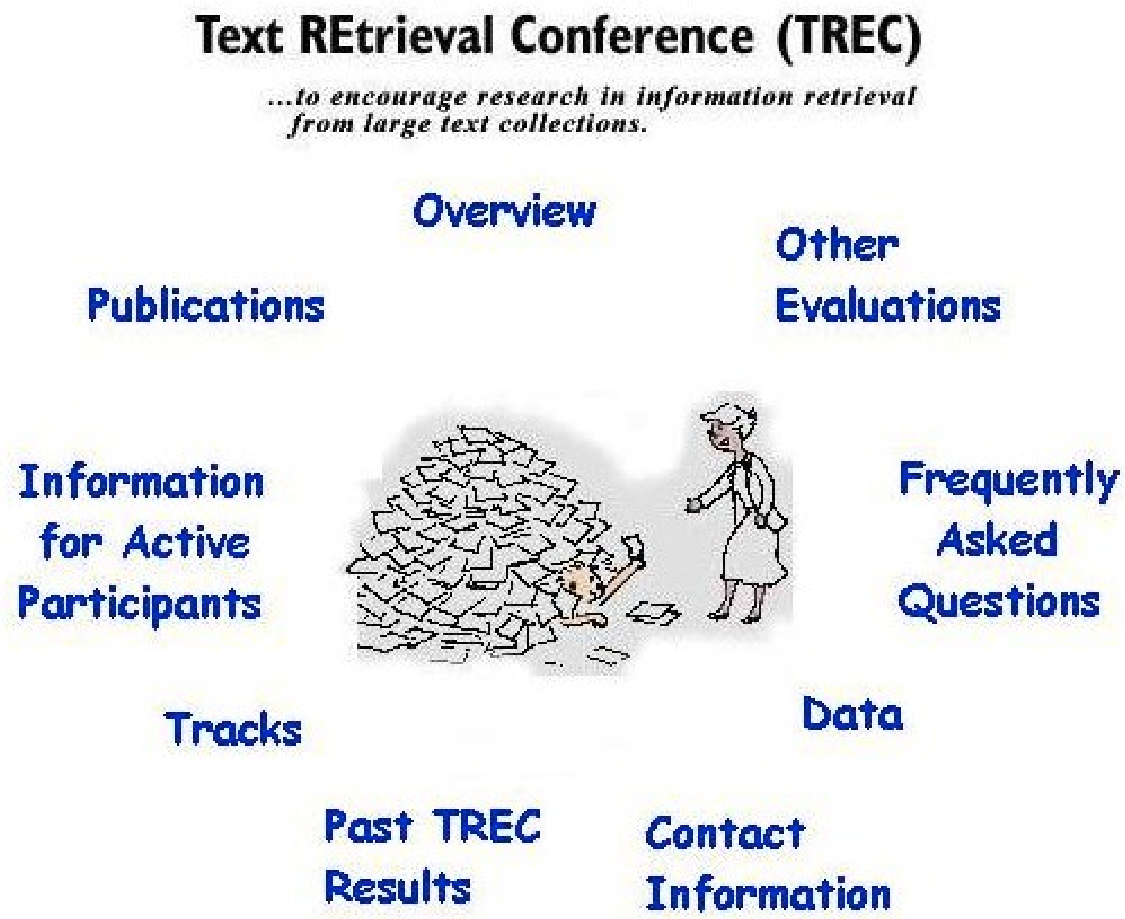
Old Test Collections

- Cranfield 2 –
  - 1400 Documents, 221 Queries
  - 200 Documents, 42 Queries
- INSPEC – 542 Documents, 97 Queries
- UKCIS -- > 10000 Documents, multiple sets, 193 Queries
- ADI – 82 Document, 35 Queries
- CACM – 3204 Documents, 50 Queries
- CISI – 1460 Documents, 35 Queries
- MEDLARS (Salton) 273 Documents, 18 Queries
- Somewhat simple

Modern Well Used Test Collections

- Text Retrieval Conference (TREC) .
  - The U.S. National Institute of Standards and Technology (NIST) has run a large IR test bed evaluation series since 1992. In more recent years, NIST has done evaluations on larger document collections, including the 25 million page GOV2 web page collection. From the beginning, the NIST test document collections were orders of magnitude larger than anything available to researchers previously and GOV2 is now the largest Web collection easily available for research purposes. Nevertheless, the size of GOV2 is still more than 2 orders of magnitude smaller than the current size of the document collections indexed by the large web search companies.
- NII Test Collections for IR Systems (NTCIR).
  - The NTCIR project has built various test collections of similar sizes to the TREC collections, focusing on East Asian language and cross-language information retrieval , where queries are made in one language over a document collection containing documents in one or more other languages. NTCIR
- Cross Language Evaluation Forum (CLEF).
  - Concentrated on European languages and cross-language information retrieval. CLEF
- Reuters-RCV1.
  - For text classification, the most used test collection has been the Reuters-21578 collection of 21578 newswire articles; see Chapter 13 , page 13.6 . More recently, Reuters released the much larger Reuters Corpus Volume 1 (RCV1), consisting of 806,791 documents. Its scale and rich annotation makes it a better basis for future research.
- 20 Newsgroups .
  - This is another widely used text classification collection, collected by Ken Lang. It consists of 1000 articles from each of 20 Usenet newsgroups (the newsgroup name being regarded as the category). After the removal of duplicate articles, as it is usually used, it contains 18941 articles.

TREC



- Text REtrieval Conference/Competition
  - http://trec.nist.gov
  - Run by NIST (National Institute of Standards & Technology)
- Collections: > Terabytes,
- Millions of entities
  - Newswire & full text news (AP, WSJ, Ziff, FT)
  - Government documents (federal register, Congressional Record)
  - Radio Transcripts (FBIS)
  - Web "subsets"

Text REtrieval Conference (TREC)

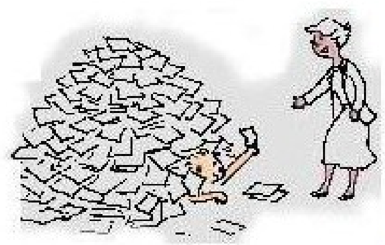
...to encourage research in information retrieval from large text collections.

Overview

Other Evaluations

Publications

Information for Active Participants



Frequently Asked Questions

Tracks

Data

Past TREC Results

Contact Information

TREC 2018 Call for Participation

Celebration of the 25th TREC: November 15, 2016

TREC Economic Impact Study

TREC Statement on Product Testing and Advertising

The TREC Conference series is co-sponsored by the NIST Information Technology Laboratory's (ITL) Retrieval Group of the Information Access Division (IAD)
Contact us at: trec (at) nist.gov

Tracks change from year to year

2010 TREC Tracks

- **Blog Track**
  The purpose of the blog track is to explore information seeking behavior in the blogosphere.
  Track coordinators: Craig Macdonald, Iadh Ounis, Ian Soboroff:
  trecblog-organisers (at) dcs.gla.ac.uk
  Mailing list: send a mail message to listproc (at) nist.gov such that the body consists of the line
  subscribe trec-blog <FirstName> <LastName>

- **Chemical IR Track**
  The goal of the chemical IR track is to develop and evaluate technology for large scale search in chemical documents including academic papers and patents to better meet the needs of professional searchers: specifically patent searchers and chemists.
  Track coordinators: John Tait, john.tait (at) ir-facility.org
  Jimmy Huang, jhuang (at) yorku.ca
  Jianhan Zhu, j.zhu (at) adastral.ucl.ac.uk
  Mhai Lupu, m.lupu (at) ir-facility.org
  Track Web Page: http://www.ir-facility.org/the_irf/trec_chem.htm
  Mailing list: follow the link on the web page to join the list

- **Entity Track**
  The overall aim of this new track is to perform entity-related search on Web data. These search tasks (such as finding entities and properties of entities) address common information needs that are not that well modeled as ad hoc document search.
  Track coordinators: Krisztian Balog, k.balog (at) uva.nl
  Paul Thomas, Paul.Thomas (at) csiro.au
  Arjen P. de Vries, arjen (at) acm.org
  Thijs Westerveld, thijs.westerveld (at) teezir.nl
  Track Web Page: http://ilps.science.uva.nl/trec-entity/
  Mailing list: visit http://groups.google.com/group/trec-entity to apply for membership.

- **Legal Track**
  The goal of the legal track is to develop search technology that meets the needs of lawyers to engage in effective discovery in digital document collections.
  Track coordinators: Gord Cormack, gvcormac (at) uwaterloo.ca
  Maura Grossman, MRGrossman (at) wlrk.com
  Bruce Hedin, bhedin (at) h5.com
  Doug Oard, oard (at) umd.edu
  Track Web Page: http://trec-legal.umiacs.umd.edu
  Mailing list: Contact oard (at) umd.edu to be added to the list.

[Tracks](#) change from year to year

TREC (cont.)

- Queries + Relevance Judgments
  - Queries devised and judged by "Information Specialists"
  - Relevance judgments done only for those documents retrieved -- not entire collection!
- Competition
  - Various research and commercial groups compete (TREC 6 had 51, TREC 7 had 56, TREC 8 had 66)
  - Results judged on precision and recall, going up to a recall level of 1000 documents

Sample TREC queries (topics)

<num> Number: 168
<title> Topic: Financing AMTRAK

<desc> Description:
A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK)

<narr> Narrative: A relevant document must provide information on the government's responsibility to make AMTRAK an economically viable entity. It could also discuss the privatization of AMTRAK as an alternative to continuing government subsidies. Documents comparing government subsidies given to air and bus transportation with those provided to aMTRAK would also be relevant.

TREC

- Benefits:
  - made research systems scale to large collections (pre-WWW)
  - allows for somewhat controlled comparisons
- Drawbacks:
  - emphasis on high recall, which may be unrealistic for what most users want
  - very long queries, also unrealistic
  - comparisons still difficult to make, because systems are quite different on many dimensions
  - focus on batch ranking rather than interaction
  - no focus on the WWW until recently

TREC evolution

- Emphasis on specialized “tracks”
  - Interactive track
  - Natural Language Processing (NLP) track
  - Multilingual tracks (Chinese, Spanish)
  - Filtering track
  - High-Precision
  - High-Performance
  - Topics
- http://trec.nist.gov/

TREC Results

- Differ each year
- For the main (ad hoc) track:
  - Best systems not statistically significantly different
  - Small differences sometimes have big effects
    - how good was the hyphenation model
    - how was document length taken into account
  - Systems were optimized for longer queries and all performed worse for shorter, more realistic queries

Evaluating search engine retrieval performance

- Recall?
- Precision?
- Order of ranking?

Evaluation Issues

To place information retrieval on a systematic basis, we need **repeatable** criteria to **evaluate** how **effective** a system is in **meeting the information needs of the user** of the system.

This proves to be very difficult with a human in the loop. It proves hard to define:

- the task that the human is attempting
- the criteria to measure success

Evaluation of Matching: Recall and Precision

If information retrieval were perfect ...

Every hit would be relevant to the original query, and every relevant item in the body of information would be found.

Precision: percentage (or fraction) of the hits that are relevant, i.e., the extent to which the set of hits retrieved by a query satisfies the requirement that generated the query.

Recall: percentage (or fraction) of the relevant items that are found by the query, i.e., the extent to which the query found all the items that satisfy the requirement.

Recall and Precision with Exact Matching: Example

- Collection of 10,000 documents, 50 on a specific topic
- Ideal search finds these 50 documents and reject all others
- Actual search identifies 25 documents; 20 are relevant but 5 were on other topics
- Precision: 20/ 25 = 0.8 *(80% of hits were relevant)*
- Recall: 20/50 = 0.4 *(40% of relevant were found)*

Measuring Precision and Recall

Precision is easy to measure:

- A knowledgeable person looks at each document that is identified and decides whether it is relevant.
- In the example, only the 25 documents that are found need to be examined.

Recall is difficult to measure:

- To know all relevant items, a knowledgeable person must go through the entire collection, looking at every object to decide if it fits the criteria.
- In the example, all 10,000 documents must be examined.

Evaluation: Precision and Recall

Precision and recall measure the results of a **single query** using a **specific search system** applied to a **specific set of documents**.

Matching methods:

Precision and recall are single numbers.

Ranking methods:

Precision and recall are functions of the rank order.

Evaluating Ranking: Recall and Precision

If information retrieval were perfect ...

Every document relevant to the original information need would be **ranked** above every other document.

With ranking, precision and recall are **functions of the rank** order.

Precision(n): fraction (or percentage) of the *n* most highly ranked documents that are relevant.

Recall(n) : fraction (or percentage) of the relevant items that are in the *n* most highly ranked documents.

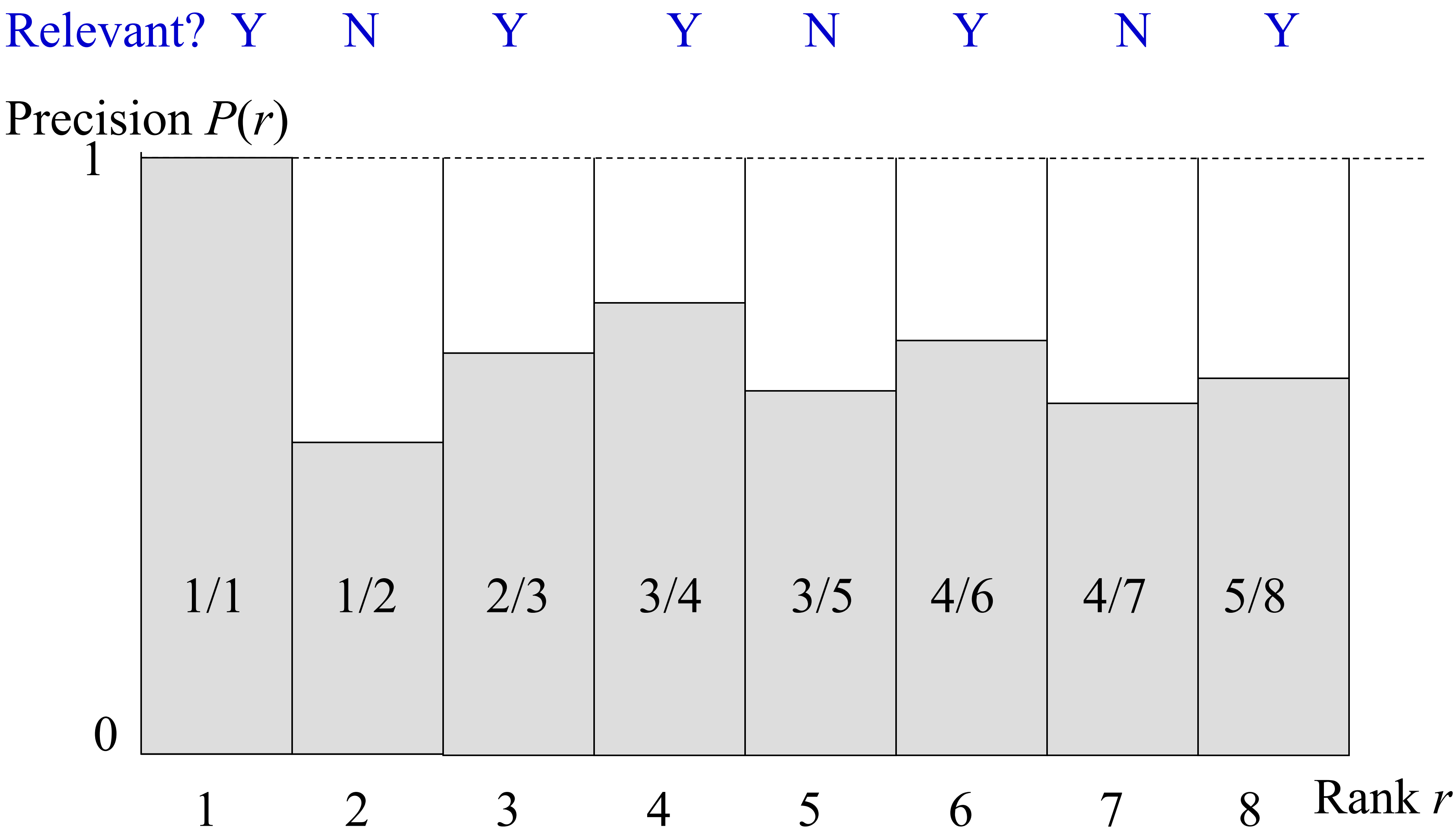
Precision and Recall with Ranking

Example

"Your query found 349,871 possibly relevant documents. Here are the first eight."

Examination of the first 8 finds that 5 of them are relevant.

Graph of Precision with Ranking: $P(r)$ as we retrieve the 8 documents.



What does the user want? Restaurant case

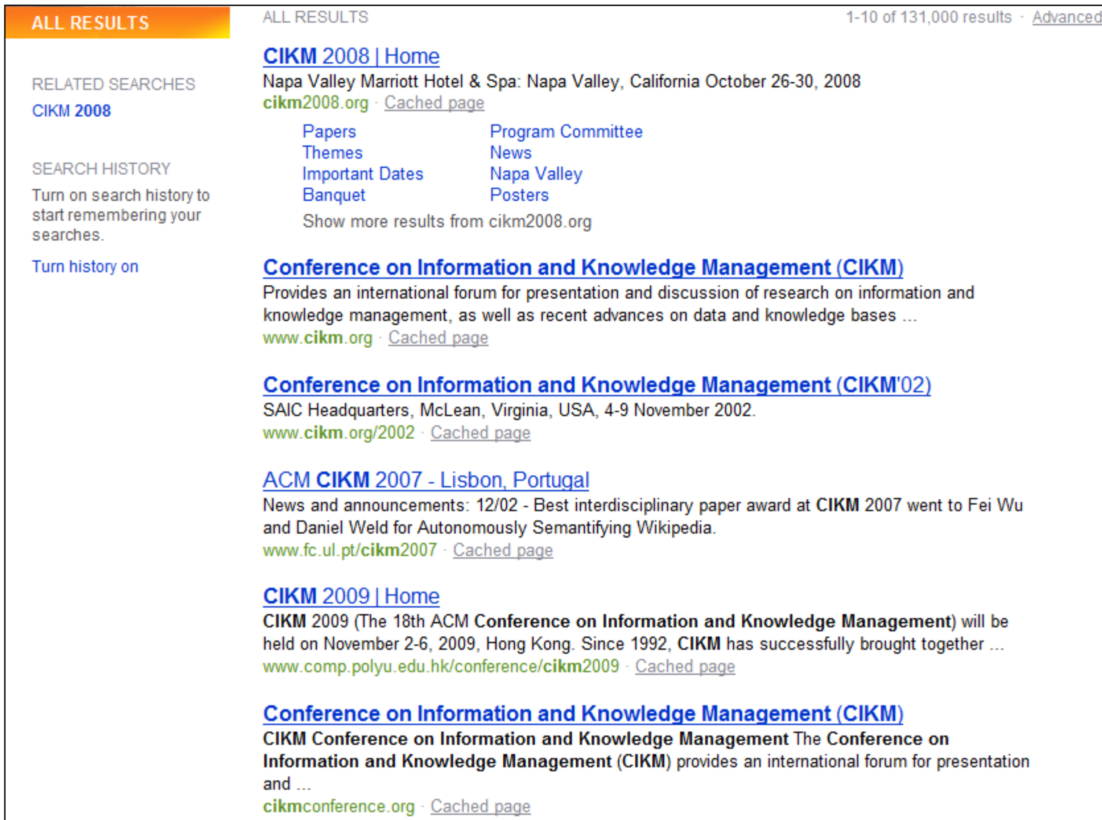
- The user wants to find a restaurant serving Sashimi. User uses 2 IR systems. How we can say which one is better?

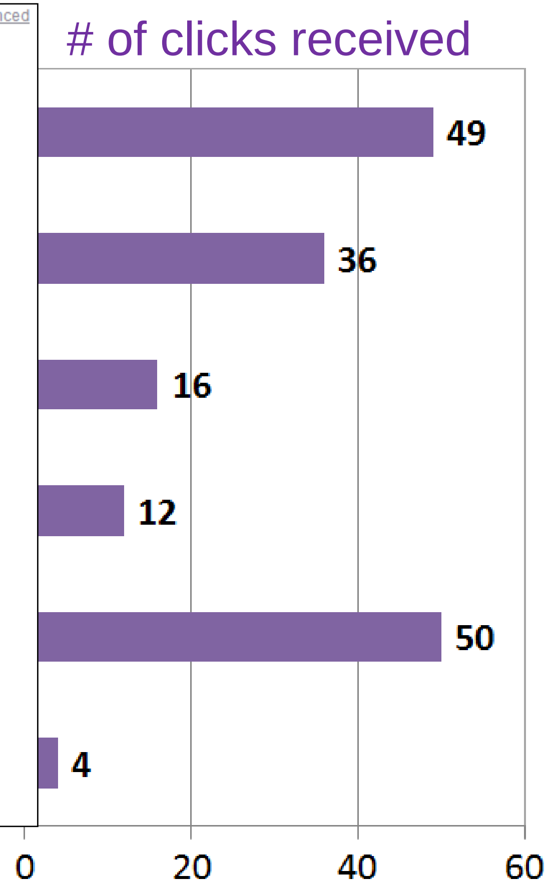
Human judgments are

- Expensive
- Inconsistent
  - Between raters
  - Over time
- Decay in value as documents/query mix evolves
- Not always representative of “real users”
  - Rating vis-à-vis query, vs underlying need
- So – what alternatives do we have?

Using user Clicks

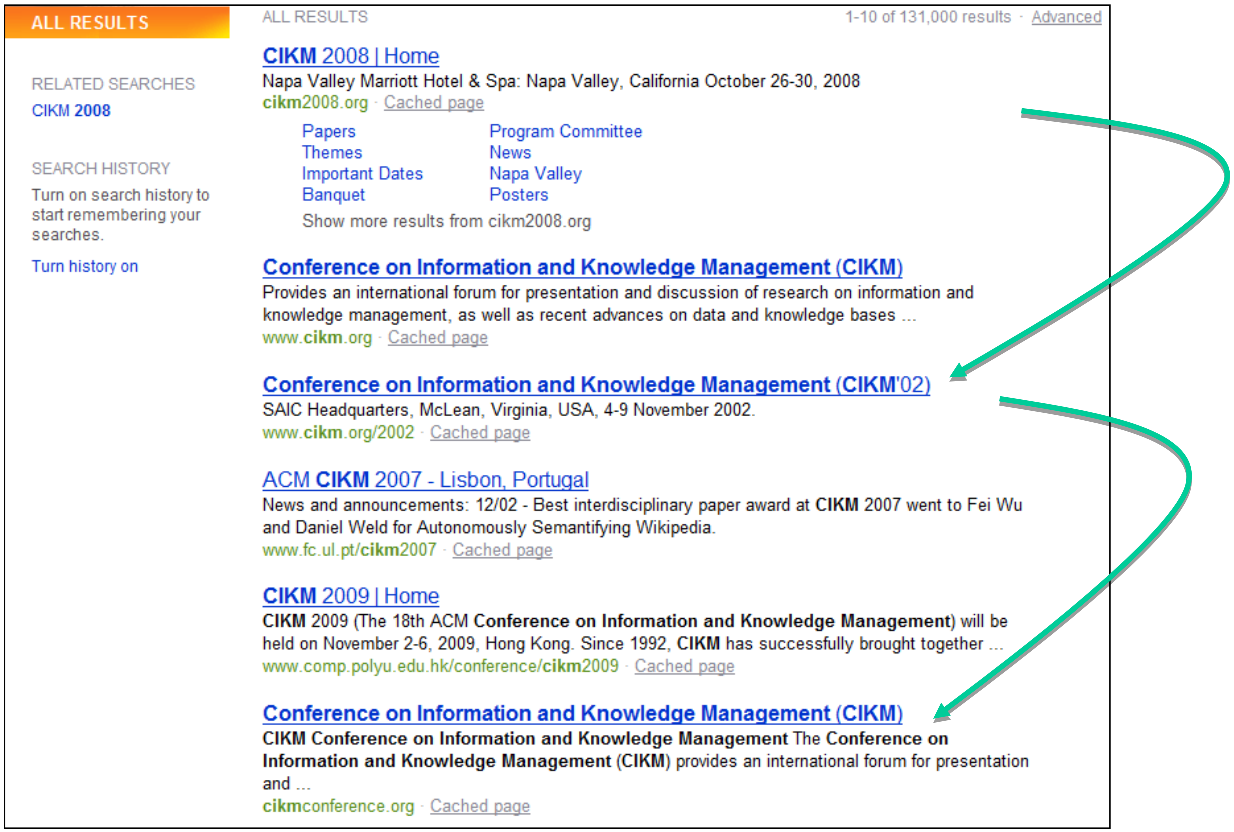
What do clicks tell us?





Strong position bias, so absolute click rates unreliable

Relative vs absolute ratings



User's click sequence

Hard to conclude <u>Result1 > Result3</u>
Probably can conclude <u>Result3 > Result2</u>

Pairwise relative ratings

- Pairs of the form: DocA <u>better than</u> DocB for a query
  - Doesn't mean that DocA <u>relevant</u> to query
- Now, rather than assess a rank-ordering wrt per-doc relevance assessments
- Assess in terms of conformance with historical pairwise preferences recorded from user clicks
- BUT!
- Don't learn and test on the same ranking algorithm
  - I.e., if you learn historical clicks from nozama and compare Sergey vs nozama on this history …

Comparing two rankings via clicks (Joachims 2002)

Query: [support vector machines]

Ranking A	Ranking B
Kernel machines	Kernel machines
SVM-light	SVMs
Lucent SVM demo	Intro to SVMs
Royal Holl. SVM	Archives of SVM
SVM software	SVM-light
SVM tutorial	SVM software

Interleave the two rankings

This interleaving starts with B

Kernel machines
Kernel machines
SVMs
SVM-light
Intro to SVMs
Lucent SVM demo
Archives of SVM
Royal Holl. SVM
SVM-light

…

Remove duplicate results

Kernel machines
Kernel machines
SVMs
SVM-light
Intro to SVMs
Lucent SVM demo
Archives of SVM
Royal Holl. SVM
SVM-light

…

Count user clicks

	Clicks
Kernel machines	A, B
Kernel machines	
SVMs	
SVM-light	A
Intro to SVMs	
Lucent SVM demo	A
Archives of SVM	
Royal Holl. SVM	
SVM-light	
…	

Ranking A: 3
Ranking B: 1

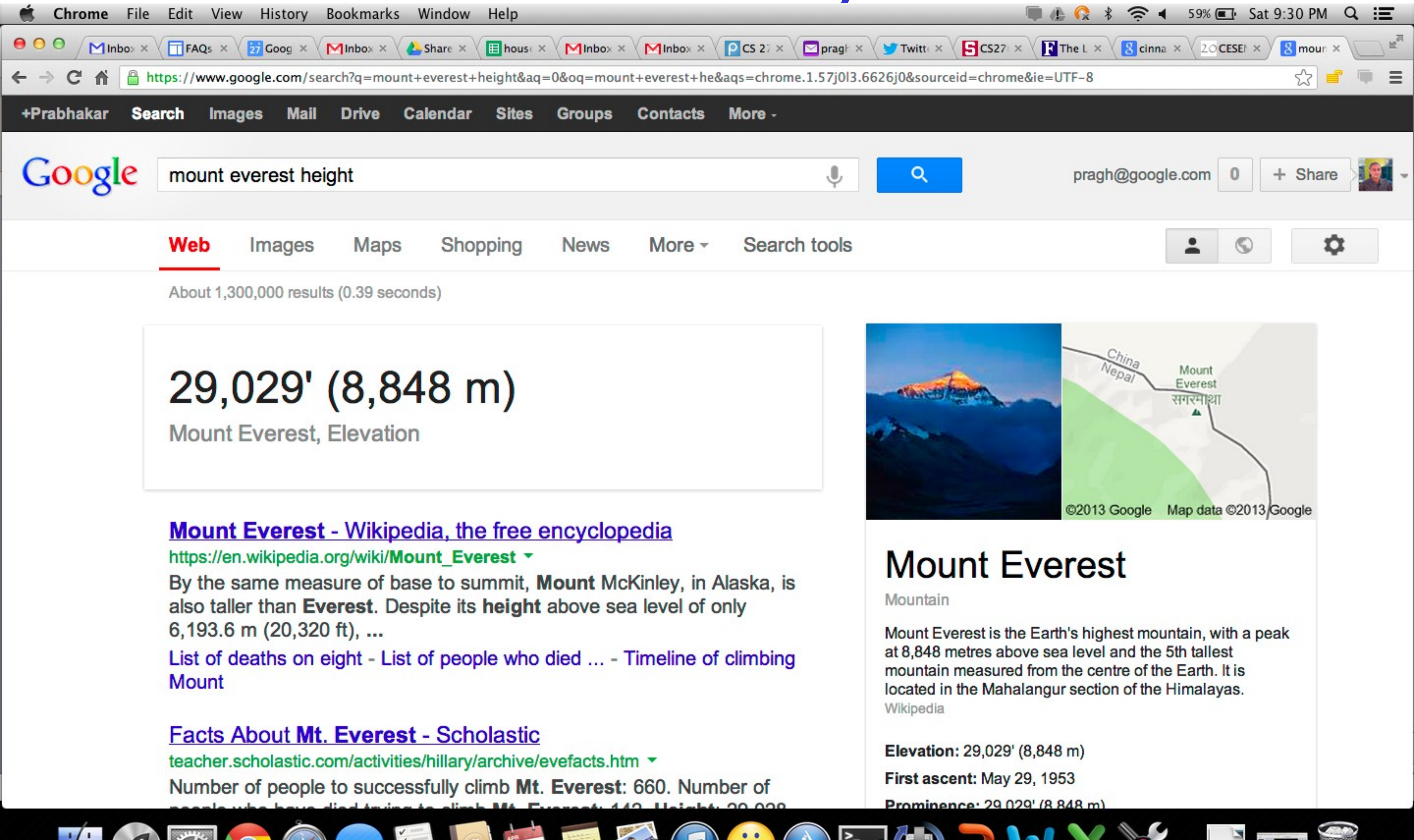
Interleaved ranking

- Present interleaved ranking to users
  - Start randomly with ranking A or ranking B to evens out presentation bias
- Count clicks on results from A versus results from B
- Better ranking will (on average) get more clicks

A/B testing at web search engines

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to an experiment to evaluate an innovation
  - Interleaved experiment
  - Full page experiment

Facts/entities (what happens to clicks?)



Recap

For ad hoc IR evaluation, need:

1. A document collection
2. A test suite of information needs, expressible as queries
3. A set of relevance judgments, standardly a binary assessment of either relevant or nonrelevant for each query-document pair.

Precision/Recall

- You can get high recall (but low precision) by retrieving all docs for all queries!
- Recall is a non-decreasing function of the number of docs retrieved

- In a good system, precision decreases as either number of docs retrieved or recall increases
  - A fact with strong empirical confirmation

Difficulties in using precision/recall

- Should average over large corpus/query ensembles
- Need human relevance assessments
  - People aren't reliable assessors
- Assessments have to be binary
  - Nuanced assessments?
- Heavily skewed by corpus/authorship
  - Results may not translate from one domain to another

What to Evaluate?

- Want an effective system
- But what is effectiveness
  - Difficult to measure
  - Recall and Precision are standard measures
    - F measure frequently used
    - Google stressed precision!

Evaluation of IR Systems

- Performance evaluations
- Retrieval evaluation
- Quality of evaluation - Relevance
- Measurements of Evaluation
  - Precision vs recall
- Test Collections/TREC