Chapter 03

Evaluation of Information Retrieval Systems

Evaluation

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

Why System Evaluation?

- There are **many retrieval models/ algorithms/ systems**, which one is the best?
- What is the best component for:
  - Ranking function (dot-product, cosine, …)
  - Term selection (stopword removal, stemming…)
  - Term weighting (TF, TF-IDF,…)

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 inviting the system is.

Relevance

- In what ways can a document be relevant to a query?
  - Answer precise question precisely.
  - Partially answer question.
  - Suggest a source for more information.
  - Give background information.
  - Remind the user of other knowledge.
  - Others ...

Which is the Best Rank Order?

a	b	c	d	e	f	g	h
R			R	R	R		
	R		R	R			R
R				R	R	R	R
	R					R	R
R			R	R		R	R
	R	R			R	R	
R		R				R	
	R	R					R
R		R	R				
	R	R					
		R					
			R	R		R	R

Relevance

- How relevant is the document
  - for this user for this information need.
- Subjective, but
- Measurable to some extent
  - How often do people agree a document is relevant to a query
- How well does it answer the question?
  - Complete answer? Partial?
  - Background Information?
  - Hints for further exploration?

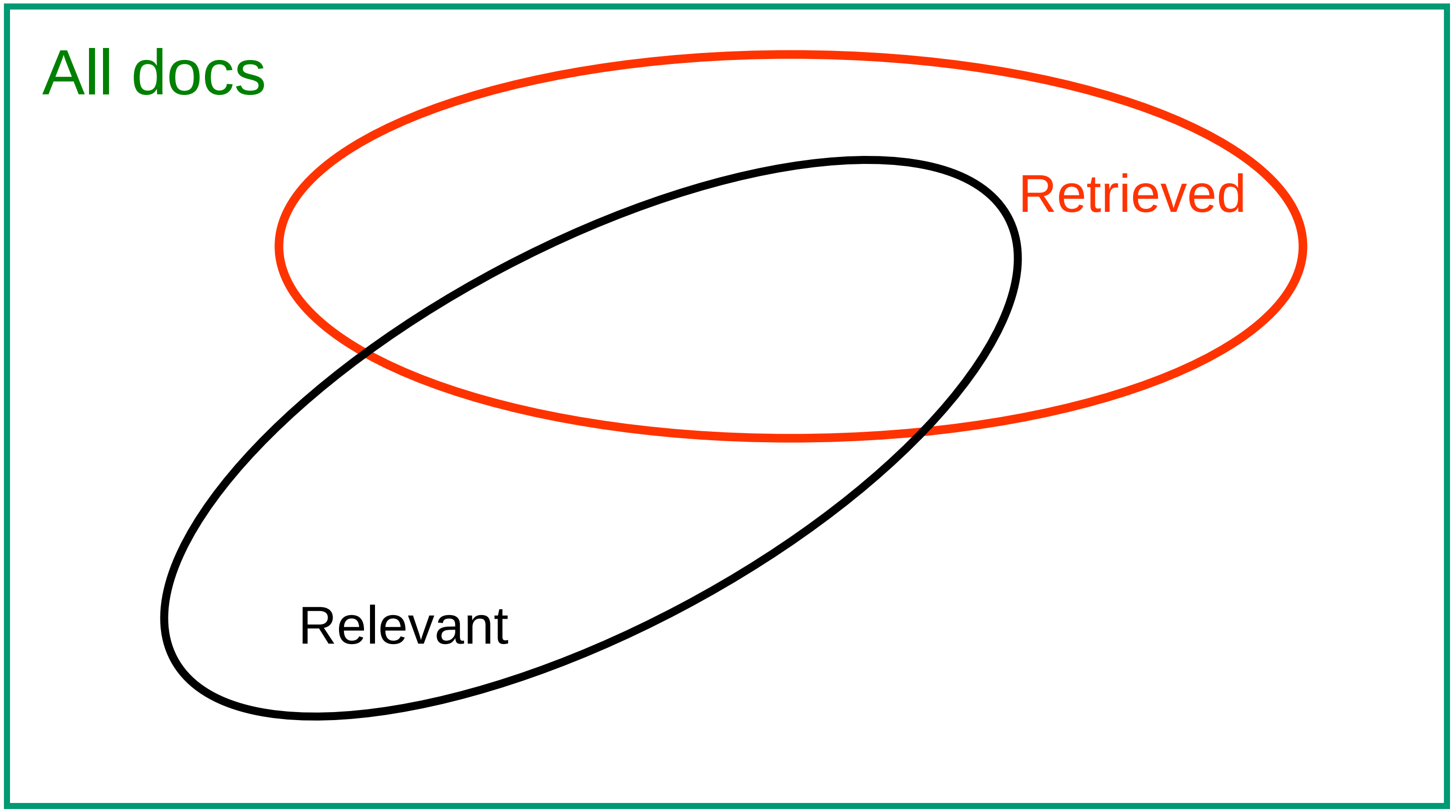
What to Evaluate?

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

effectiveness

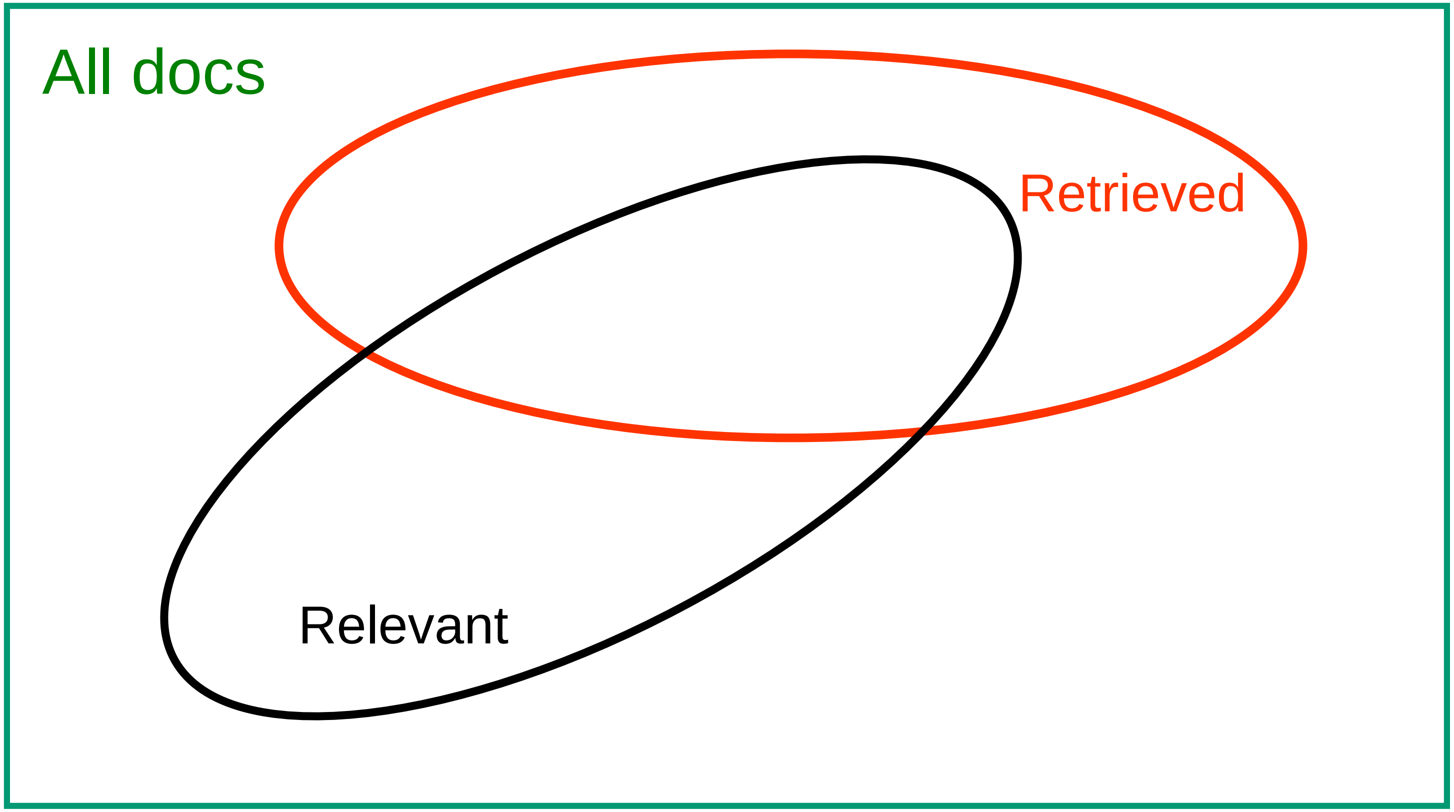
Relevant vs. Retrieved



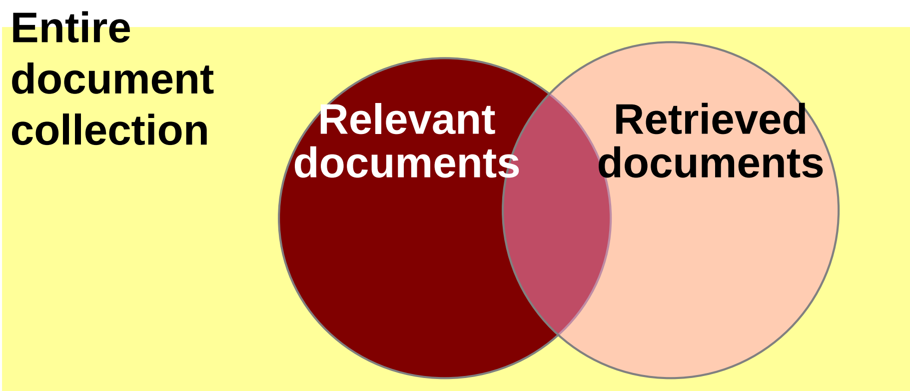
Precision vs. Recall

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

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



Precision and Recall



	retrieved	not retrieved
irrelevant	retrieved & irrelevant	Not retrieved & irrelevant
relevant	retrieved & relevant	not retrieved but relevant

$$recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}$$

$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved}$$

Another common representation

	Relevant	Not relevant
Retrieved	A	B
Not retrieved	C	D

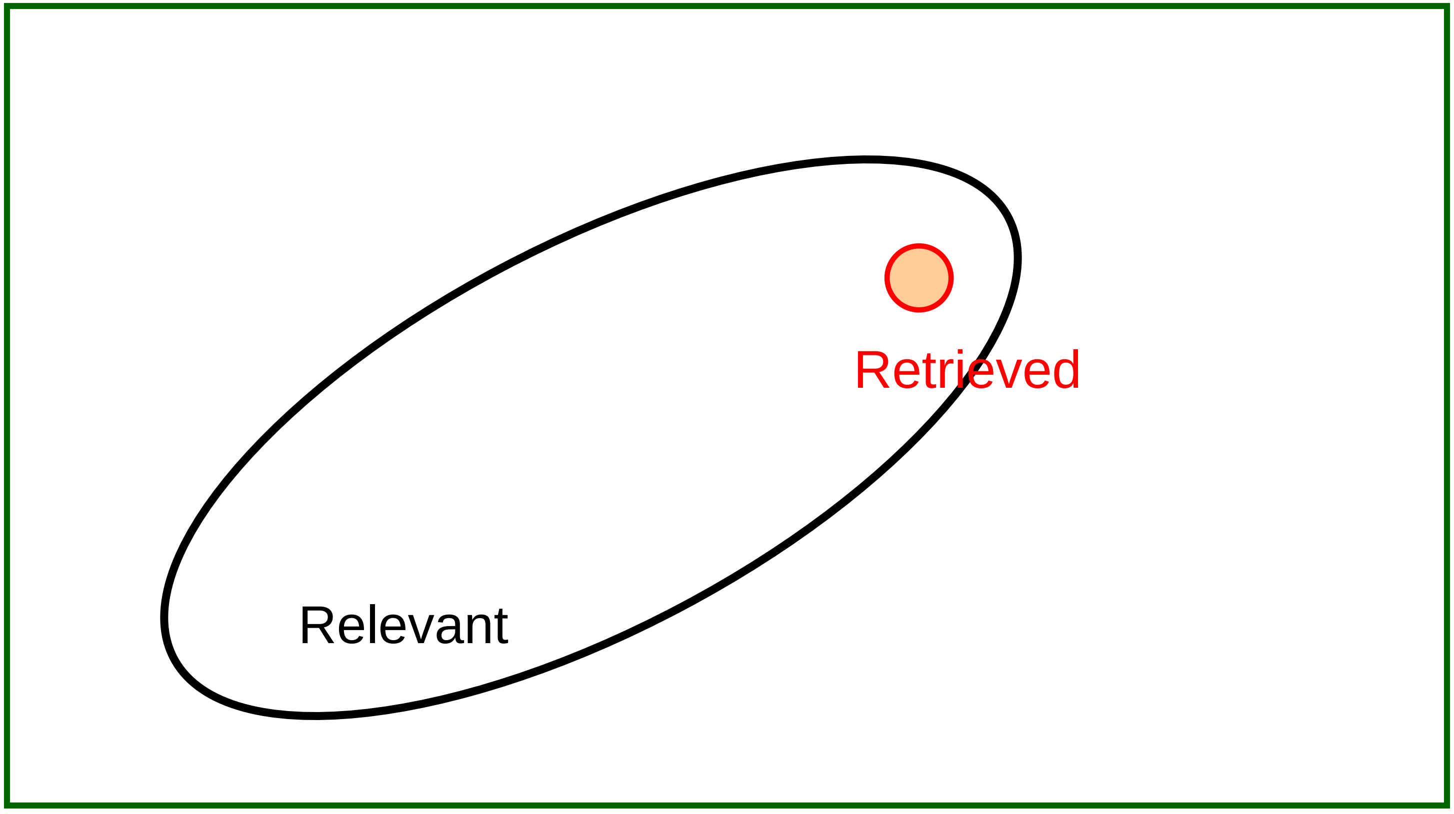
- Relevant = A+C
- Retrieved = A+B
- Collection size = A+B+C+D

- Precision = A ÷ (A+B)
- Recall = A ÷ (A+C)

- Miss = C ÷ (A+C)
- False alarm (fallout) = B ÷ (B+D)

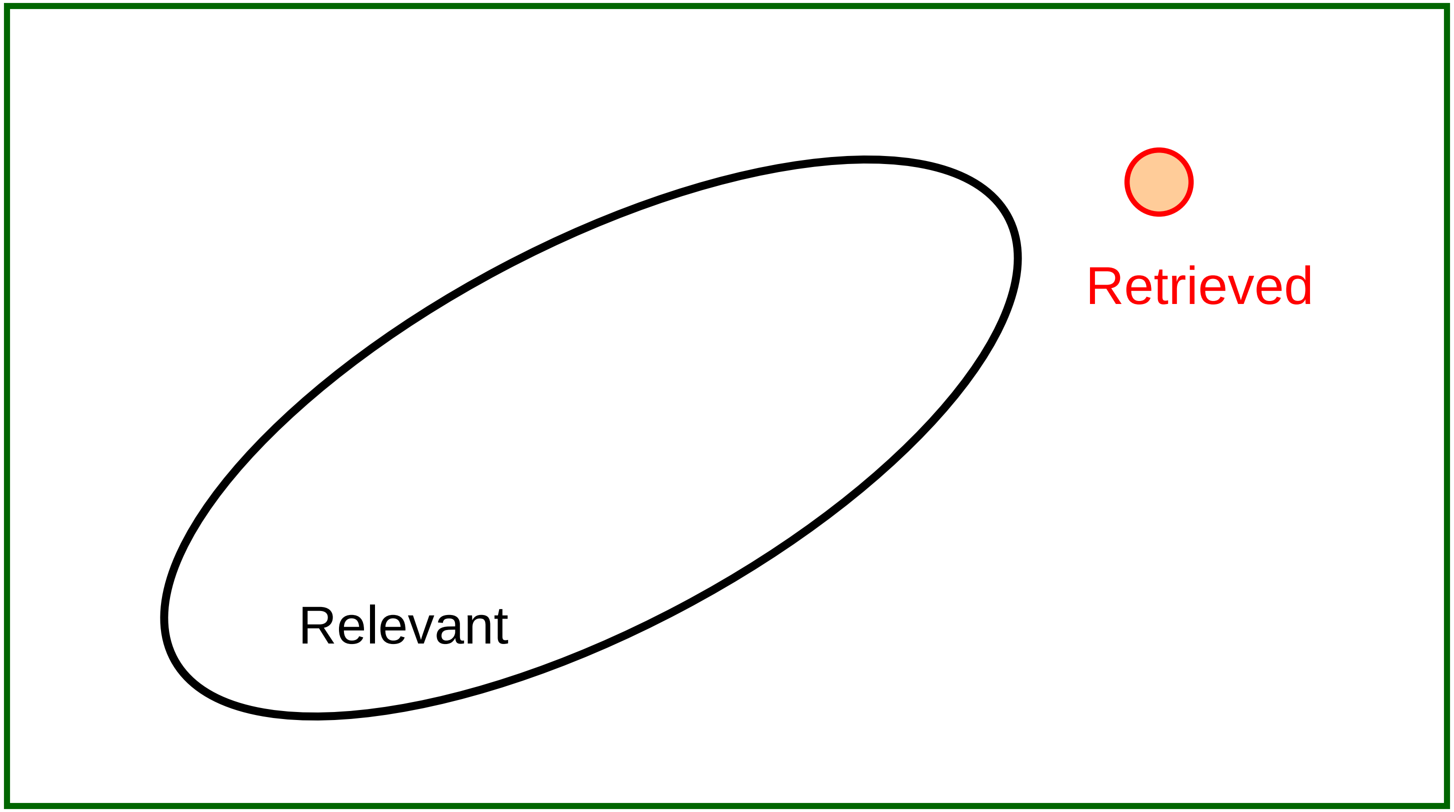
Retrieved vs. Relevant Documents

Very high precision, very low recall



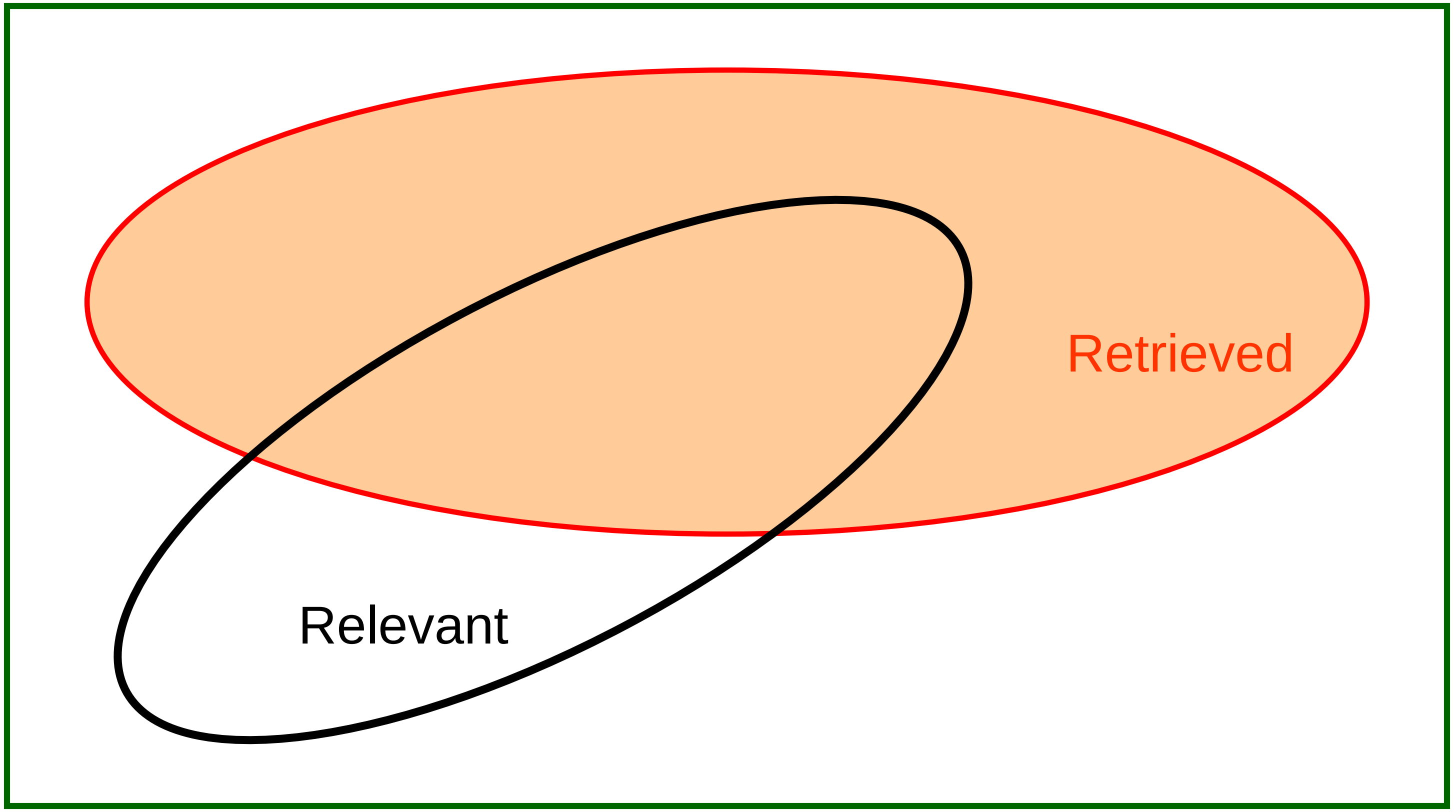
Retrieved vs. Relevant Documents

Very low precision, very low recall (0 in fact)



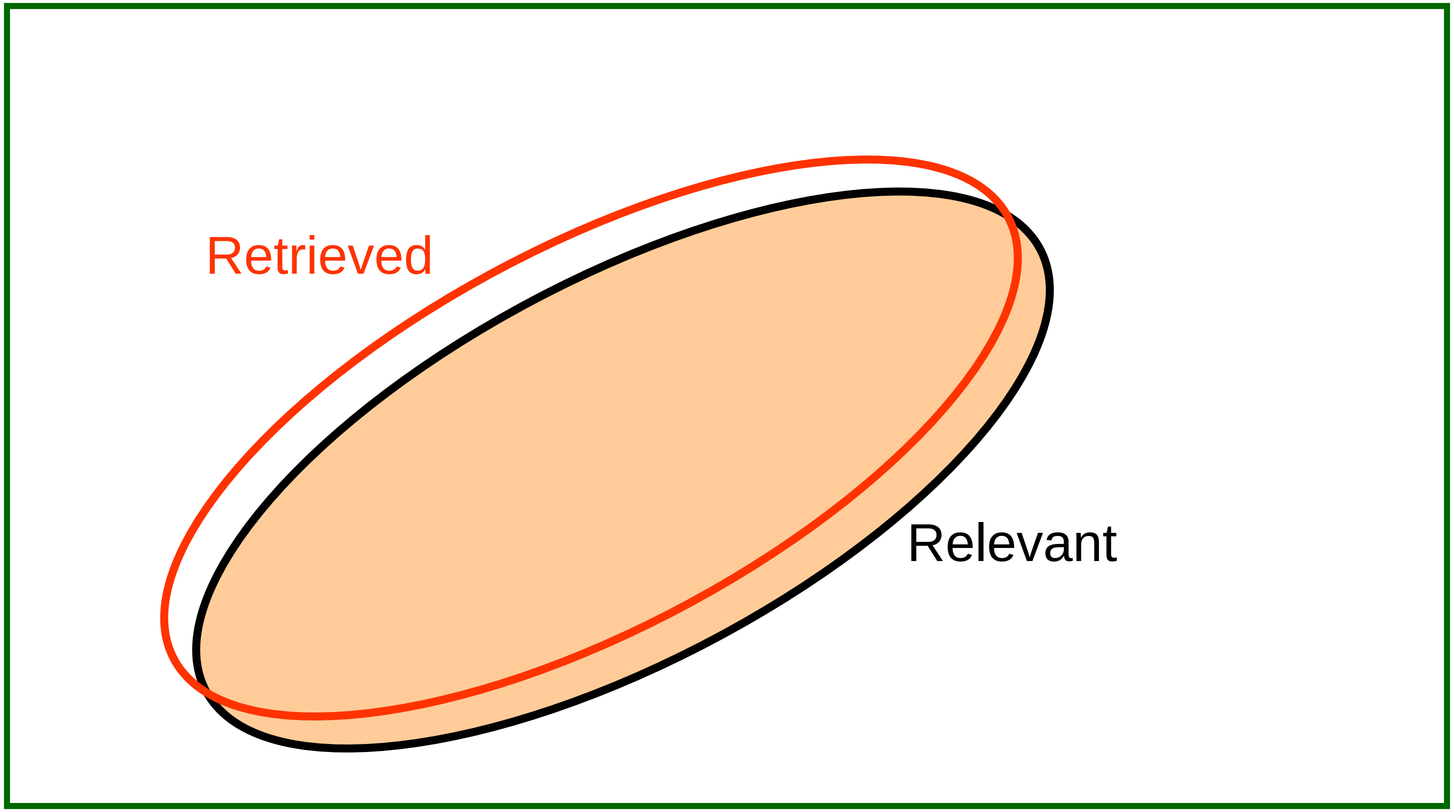
Retrieved vs. Relevant Documents

High recall, but low precision



Retrieved vs. Relevant Documents

High precision, high recall (at last!)

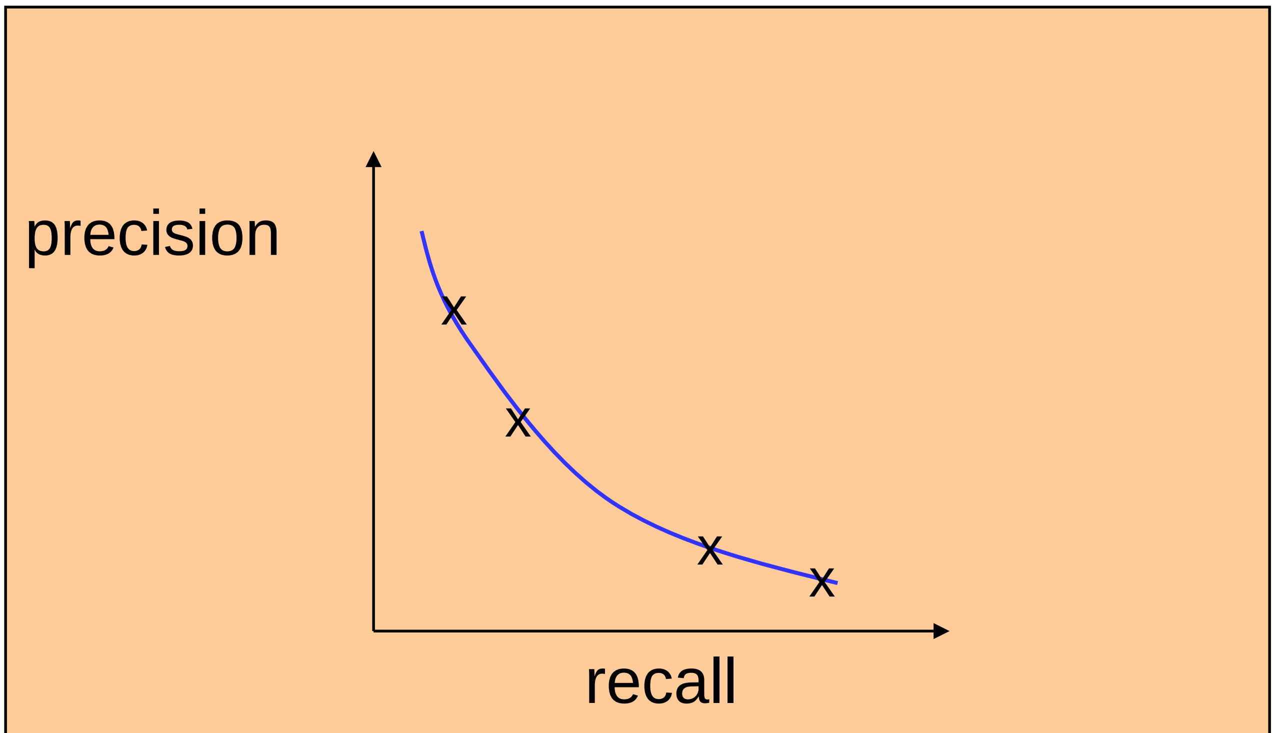


Average Recall/Precision Curve

- Typically average performance over a large ***set*** of queries.
- Compute average precision at each standard recall level across all queries.
- Plot average precision/recall curves to evaluate overall system performance on a document/query corpus.

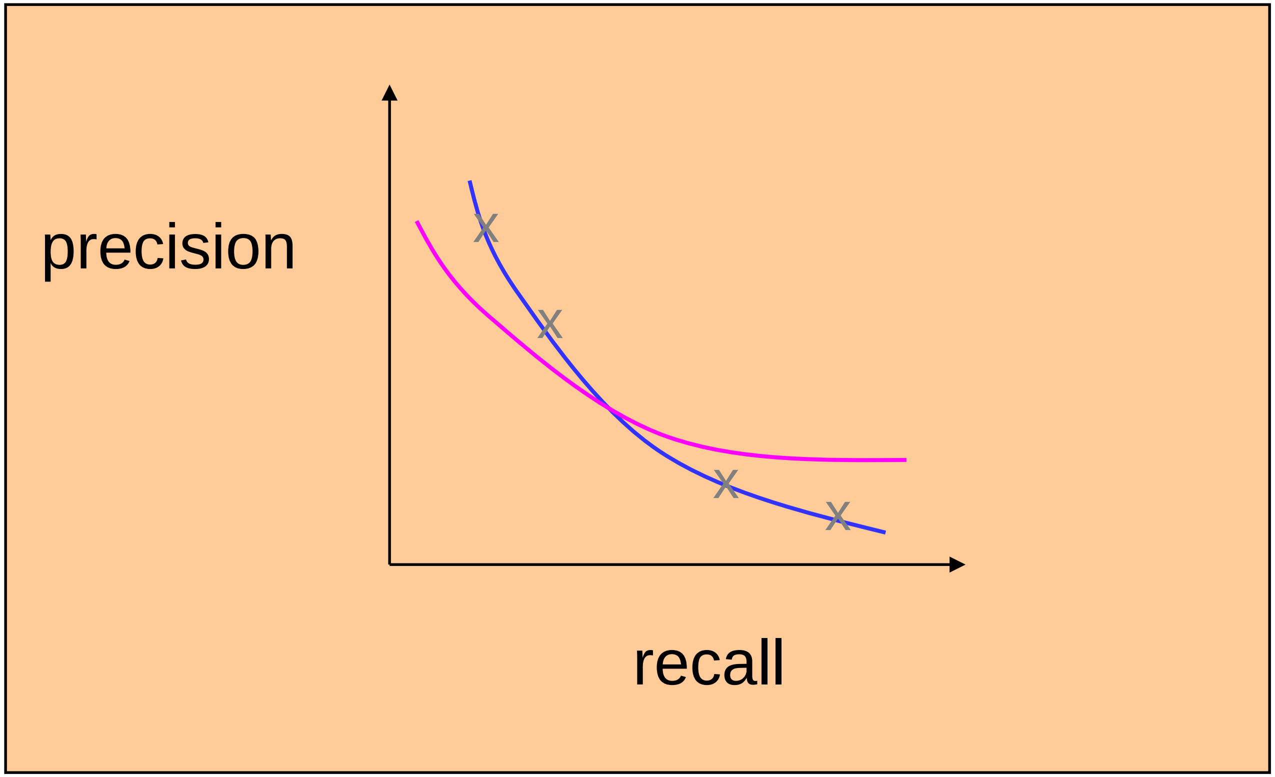
Precision/Recall Curves

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



Precision/Recall Curves

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



5 เอกสาร ที่ตรงประเด็น

การจัดลำดับ #1

#เอกสารที่ออกมา	1	2	3	4	5	6	7	8	9	10
P	1/1	1/2	2/3	2/4	2/5	3/6	3/7	3/8	4/9	5/10
R	1/5	1/5	2/5	2/5	2/5	3/5	3/5	3/5	4/5	5/5

เปรียบเทียบผลลัพธ์จาก Search Engine A และ B

A 

P	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5
R	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0

$AVGPrec_A = 62.2\%$

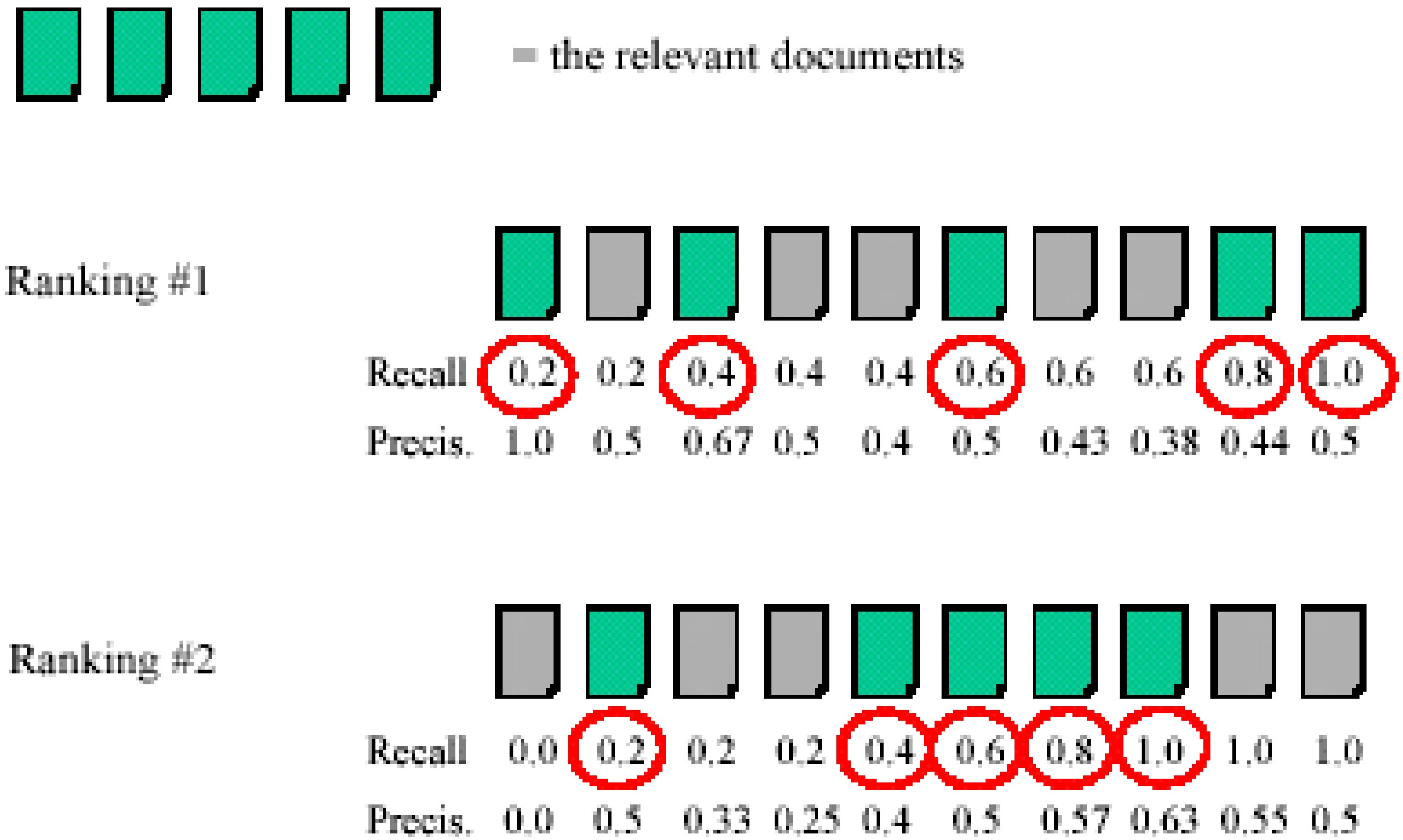
B 

P	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.63	0.55	0.5
R	0.0	0.2	0.2	0.2	0.4	0.6	0.8	1.0	1.0	1.0

$AVGPrec_B = 52\%$

$AVGPrec_A > AVGPrec_B$ $\therefore$ A ดีกว่า B

Precision and Recall example



Computing Recall/Precision Points: An Example

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Let total # of relevant docs = 5
Check each new recall point:

R=1/5=0.2; P=1/1=1

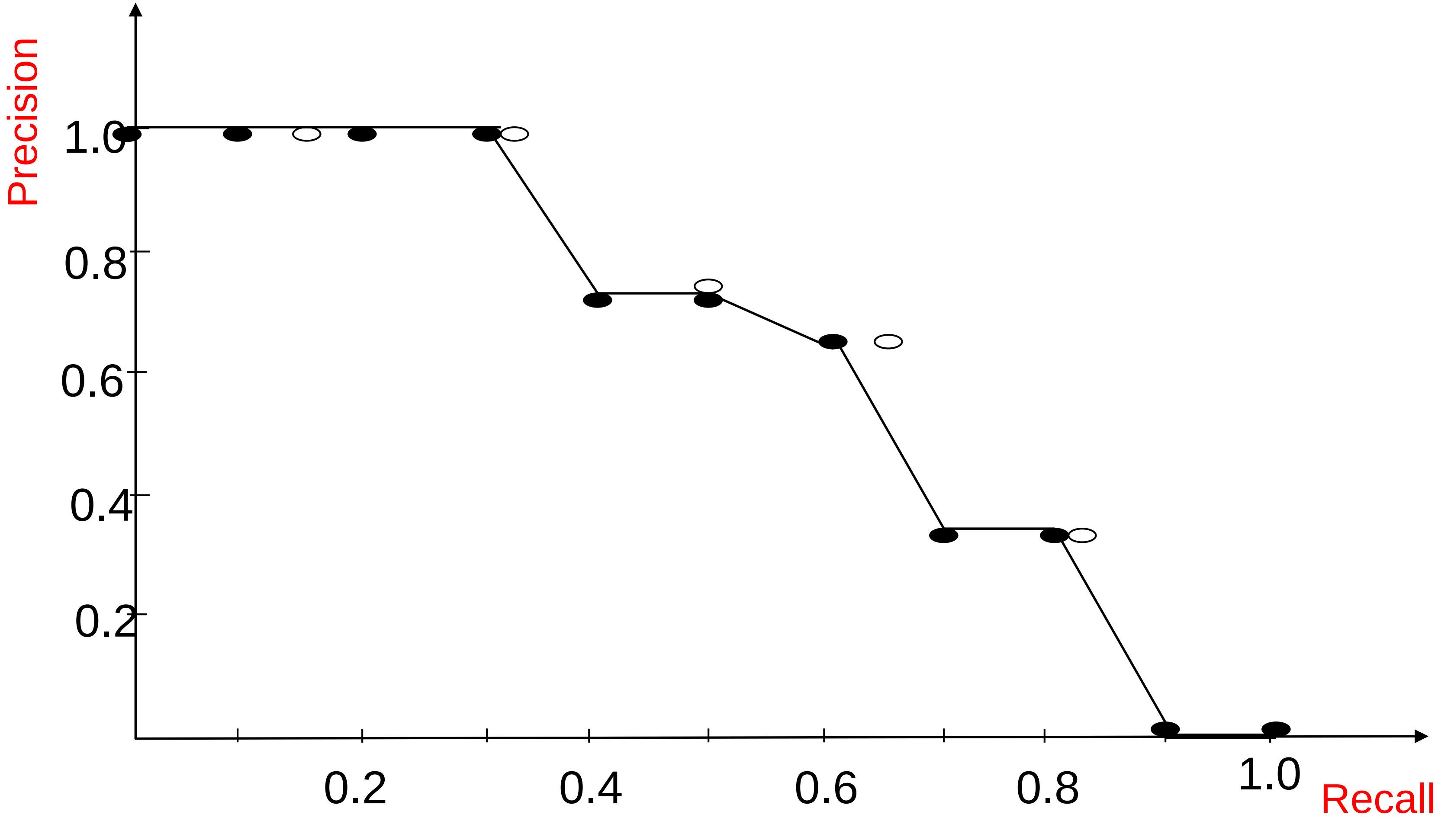
R=2/5=0.4; P=2/2=1

R=3/5=0.6; P=3/4=0.75

R=4/5=0.8; P=4/6=0.67

R=5/5=1.0; p=5/13=0.38

Interpolating a Recall/Precision Curve: An Example



Precision versus recall curve

- $R_q=\{d_3,d_5,d_9,d_{25},d_{39},d_{44},d_{56}, d_{71},d_{89},d_{123}\}$

Ranking for query q:

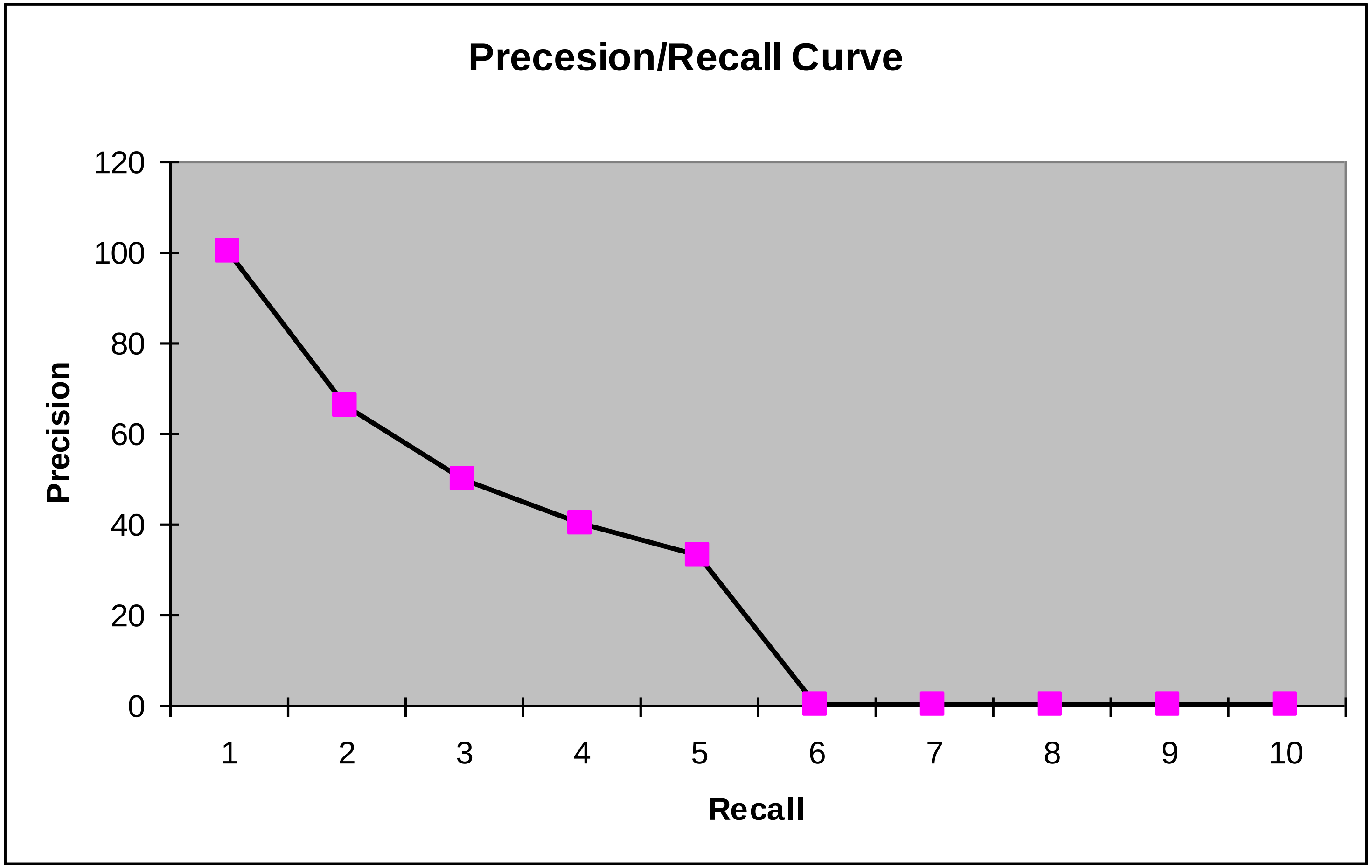
1.d_{123}*	6.d_9*	11.d_{38}
2.d_{84}	7.d_{511}	12.d_{48}
3.d_{56}*	8.d_{129}	13.d_{250}
4.d_6	9.d_{187}	14.d_{11}
5.d_8	10.d_{25}*	15.d_3*

- **P = 1 at R = 0.1**
- **P = 0.66 at R = 0.2**
- **P = 0.5 at R = 0.3**
- **P = 0.4 at R = 0.4**
- **P = 0.33 at R = 0.5**

Usually based on 11 standard recall levels: 0%, 10%, ..., 100%

Precision versus recall curve

- For a single query



Precision and Recall

- Precision
  - The ability to retrieve top-ranked documents that are mostly relevant.
- Recall
  - The ability of the search to find ***all*** of the relevant items in the corpus.

Determining Recall is Difficult

- Total number of relevant items is sometimes not available:
  - Sample across the database and perform relevance judgment on these items.
  - Apply different retrieval algorithms to the same database for the same query. The aggregate of relevant items is taken as the total relevant set.

Average Precision

- Often want a single-number effectiveness measure
  - E.g., for a machine-learning algorithm to detect improvement
- Average precision is widely used in IR
- Calculate by averaging precision when recall increases

Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precis.	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

AvgPrec= 62.2%

Recall	0.0	0.2	0.2	0.2	0.4	0.6	0.8	1.0	1.0	1.0
Precis.	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.63	0.55	0.5

AvgPrec= 52.0%

Precision and Recall second example

= the relevant documents (as before)

Ranking #1

Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precis.	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

AvgPrec= 62.2%

= different query's relevant documents

Ranking #3

Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precis.	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

AvgPrec= 44.3%

R- Precision

- Precision at the R-th position in the ranking of results for a query that has R relevant documents.

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

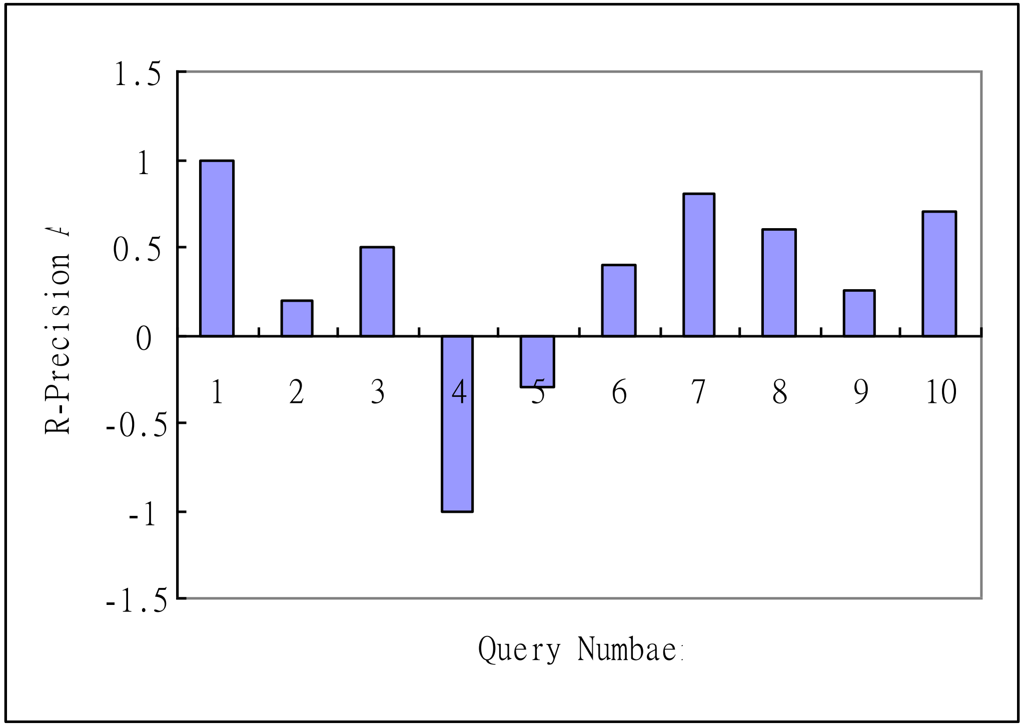
R = # of relevant docs = 5

R-Precision = 3/5 = 0.60

Precision Histograms

- Use R-precision measures to compare the retrieval history of two algorithms through visual inspection

$$RP_{A/B}(i) = RP_A(i) - RP_B(i)$$



F-Measure

- One measure of performance that takes into account both recall and precision.
- Harmonic mean of recall and precision:

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

- Compared to arithmetic mean, both need to be high for harmonic mean to be high.

E Measure (parameterized F Measure)

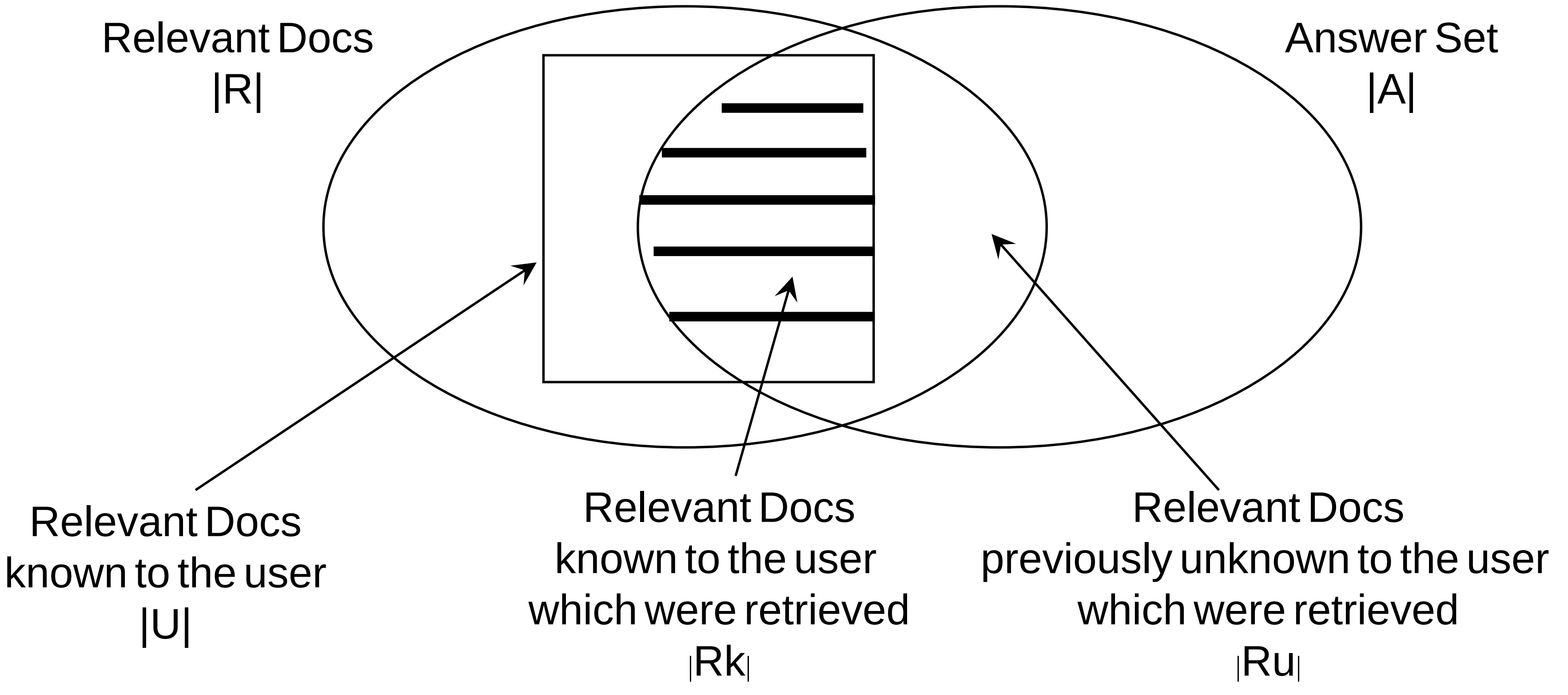
- A variant of F measure that allows weighting emphasis on precision over recall:

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

- Value of β controls trade-off:
  - β = 1: Equally weight precision and recall (E=F).
  - β > 1: Weight precision more.
  - β < 1: Weight recall more.

User-Oriented Measure

- Coverage=$|R_k|/|U|$
- Novelty=$|R_u|/(|R_u|+|R_k|)$



Fallout Rate

- Problems with both precision and recall:
  - Number of irrelevant documents in the collection is not taken into account.
  - Recall is undefined when there is no relevant document in the collection.
  - Precision is undefined when no document is retrieved.

$$Fallout = \frac{no.of\ nonrelevant\ items\ retrieved}{total\ no.of\ nonrelevant\ items\ in\ the\ collection}$$