

Retrieval Effectiveness

- Evaluation of IR systems
- Relevance judgement
- Performance measures
 - Recall,
 - Precision,
 - Single-valued measures
 - etc.

Why System Evaluation?

- Any systems needs validation and verification
 - Check whether the system is right or not
 - Check whether it is the right system or not
- It provides the ability to measure the difference between IR systems
 - How well do our search engines work?
 - Is system A better than B?
 - Under what conditions?
- Evaluation drives what to study
 - Identify techniques that work well and do not work
 - There are many retrieval models/algorithms
 - which one is the best?
 - What is the best component for:
 - Similarity measures (dot-product, cosine, ...)
 - Index term selection (tokenization, stop-word removal, stemming...)
 - Term weighting (TF, TF-IDF,...)

Evaluation Criteria

What are the main evaluation measures to check the performance of an IR system?

- Efficiency

- Time and space complexity
- ❑ Speed in terms of retrieval time and indexing time
- ❑ Speed of query processing
- ❑ The space taken by corpus vs. index file
 - Index size: determine Index/corpus size ratio
 - Is there a need for compression

- Effectiveness

- How is a system capable of retrieving relevant documents from the collection?
- Is system X better than other systems?
- User satisfaction: How “good” are the documents that are returned as a response to user query?
- **Relevance** of results to meet information need of users

Types of Evaluation Strategies

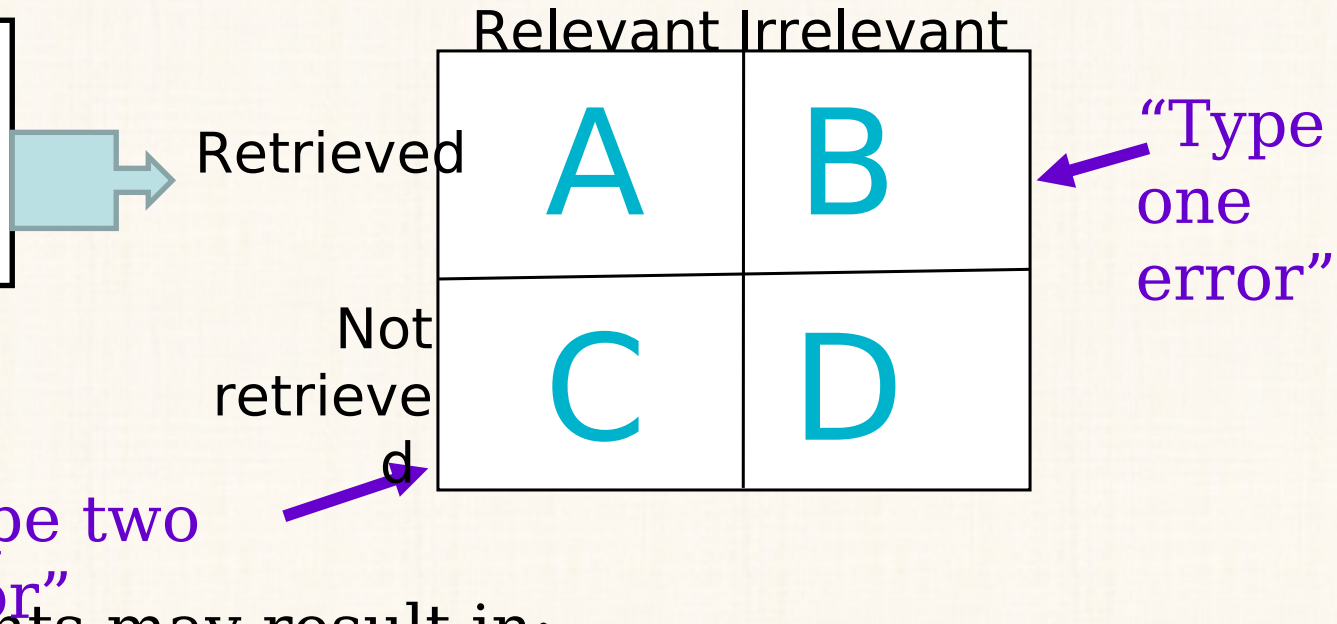
- System-centered evaluation
 - Given documents, queries, and relevance judgments
 - Try several variations of the system
 - Measure which system returns the “best” hit list
- User-centered evaluation
 - Given several users, and at least two retrieval systems
 - Have each user try the same task on both systems
 - Measure which system works the “best” for users information need
 - How to measure users satisfaction?

The Notion of Relevance Judgment

- Relevance is the measure of a correspondence existing between a document and query.
 - Construct document - query as determined by:
 - (i) the user who posed the retrieval problem;
 - (ii) an external judge;
 - (iii) information specialist
 - Is the relevance judgment made by users and external person the same ?
- Relevance judgment is usually:
 - **Subjective**: Depends upon a specific user's judgment.
 - **Situational**: Relates to user's current needs.
 - **Cognitive**: Depends on human perception and behavior.
 - **Dynamic**: Changes over time.

Measuring Retrieval Effectiveness

- Metrics often used to evaluate effectiveness of the system



- Retrieval of documents may result in:
 - False positive (Errors of omission)**: some irrelevant documents may be retrieved by the system as relevant.
 - False negative (False drop or Errors of commission)**: some relevant documents may not be retrieved by the system as irrelevant.
 - For many applications a good index should not permit any false drops, but may permit a few false positives.
 - True positive** and **True negative**

Measuring Retrieval Effectiveness

	Relevant	Not relevant
Retrieved	A	B
Not retrieved	C	D

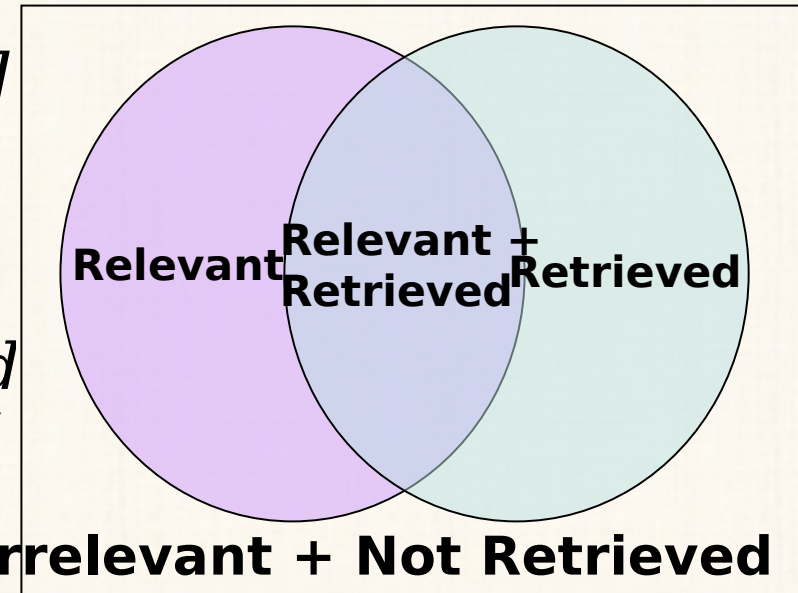
Collection size = $A+B+C+D$

Relevant = $A+C$

Retrieved = $A+B$

$$\text{Recall} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

$$\text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$



- When is precision important? When is recall important?

Example

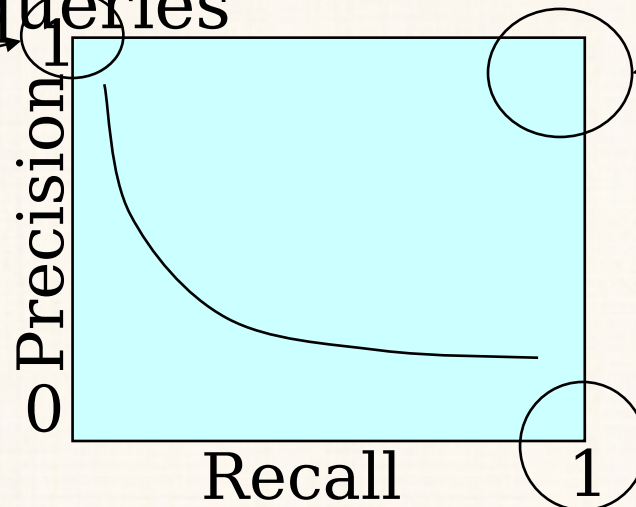
Assume that there are a total of 10 relevant

Ranking	document Relevance	Recall	Precision
1. Doc. 50	R	0.10	1.00
2. Doc. 34	NR	0.10	0.50
3. Doc. 45	R	0.20	0.67
4. Doc. 8	NR	0.20	0.50
5. Doc. 23	NR	0.20	0.40
6. Doc. 16	NR	0.20	0.33
7. Doc. 63	R	0.30	0.43
8. Doc 119	R	0.40	0.50
9. Doc 21	NR	0.40	0.44
10. Doc 80	R	0.50	0.50

Graphing Precision and Recall

- Plot each (recall, precision) point on a graph
- Recall is a non-decreasing function of the **number of documents retrieved**,
 - Precision usually decreases (in a good system)
- Precision/Recall tradeoff
 - Can increase recall by retrieving many documents (down to a low level of relevance ranking),
 - but many irrelevant documents would be fetched, reducing precision
 - Can get high recall (but low precision) by retrieving all documents for all queries

Returns
relevant
documents
but
misses many
useful ones

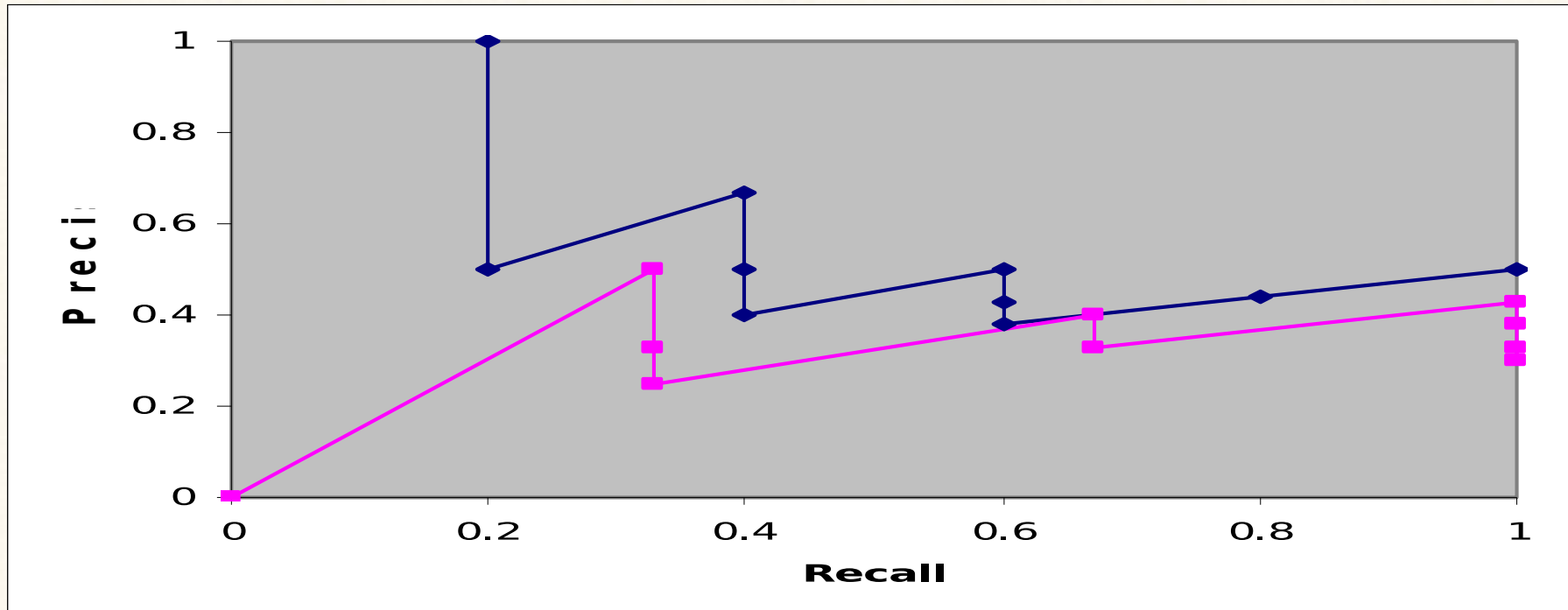


The ideal

Returns most relevant
documents but includes
lots of junk

Need for Interpolation

- Two issues:
 - How do you compare performance across queries?
 - Is the sawtooth shape intuitive of what's going on?



Solution: Interpolation!

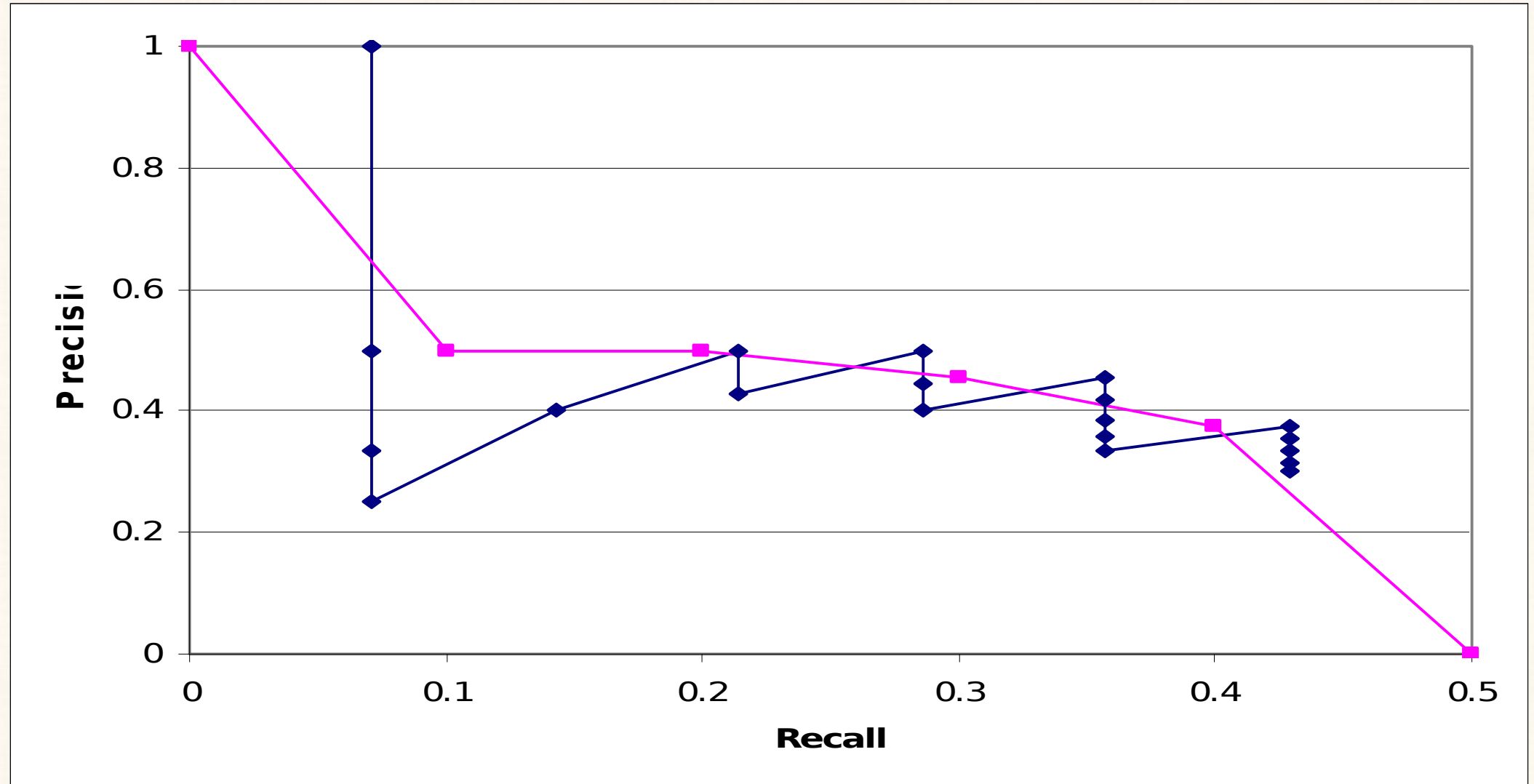
Interpolate a precision value for each *standard recall level*

Interpolation

- It is a general form of precision/recall calculation
- Precision change w.r.t. Recall (not a fixed point)
 - It is an empirical fact that *on average* as recall increases, precision decreases
- Interpolate precision at 11 standard recall levels:
 - $r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, where $j = 0$
.... 10
- The interpolated precision at the j -th standard recall level is the maximum known precision at any recall level between the j^{th} and $(j + 1)^{\text{th}}$ level:

$$P(r_j) = \max_{r_j \leq r \leq r_{j+1}} P(r)$$

Result of Interpolation



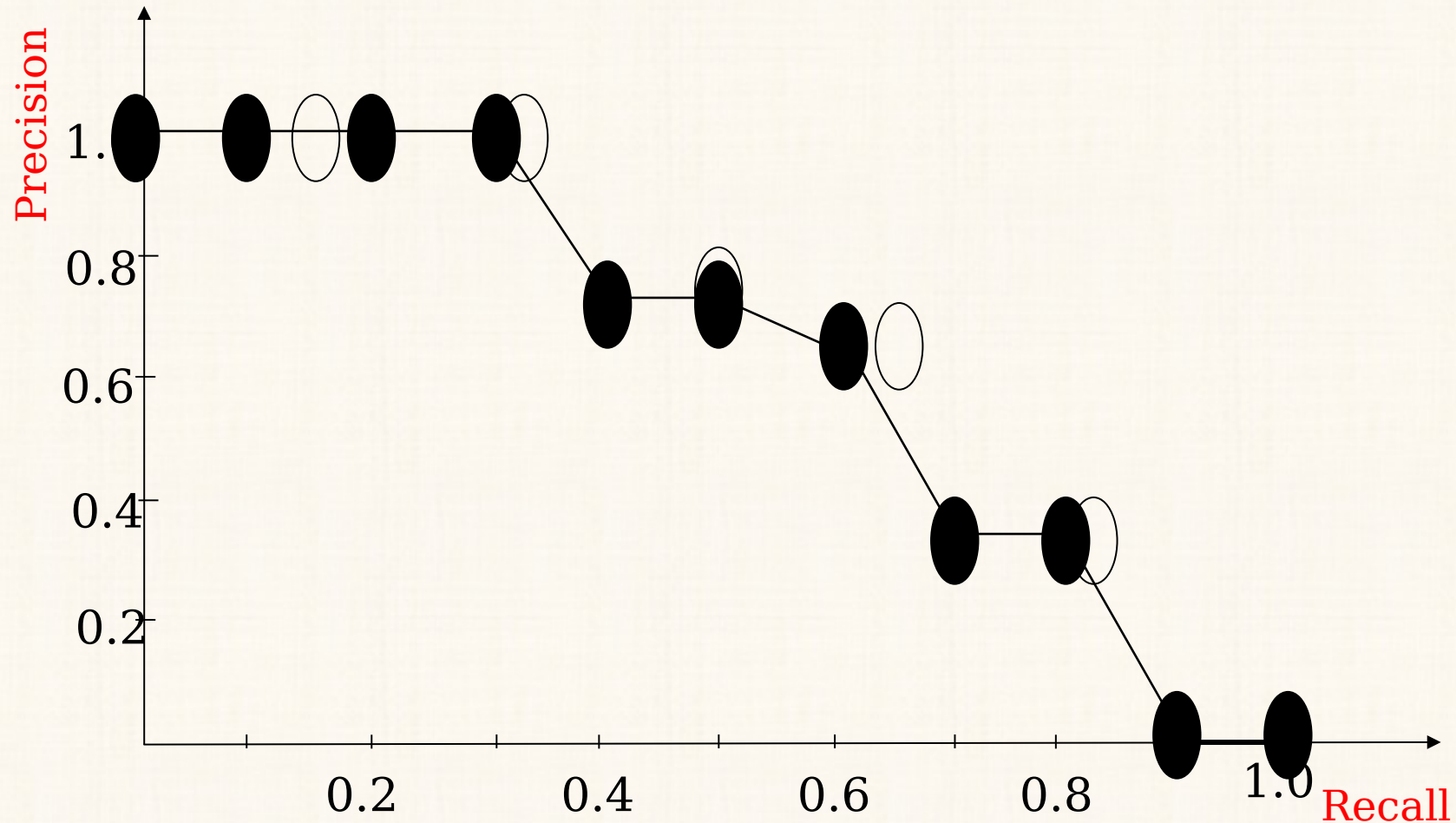
Exercise

- Let total number of relevant documents = 6, compute recall and precision for each cut off point n:

n	doc #	relevant	Recall	Precision
1	588	x	0.167	1
2	589	x	0.333	1
3	576			
4	590	x	0.5	0.75
5	986			
6	592	x	0.667	0.667
7	984			
8	988			
9	578			
10	985			
11	103			
12	591			
13	772	x	0.833	0.38
14	990			

Missing one relevant document. Never reach¹³

Interpolating a Recall/Precision Curve: Exercise



Computing Recall/Precision

Points:

Exercise 2

n	doc #	relevant
1	588	x
2	576	
3	589	x
4	342	
5	590	x
6	717	
7	984	
8	772	x
9	321	x
10	498	
11	113	
12	628	
13	772	
14	592	x

Let total # of relevant docs
= 6

Check each new recall

$R=1/6=0.167;$

$P=1/1=1$

$R=2/6=0.333;$

$P=2/3=0.667$

$R=3/6=0.5;$

$P=3/5=0.6$

$R=4/6=0.667;$

$P=4/8=0.5$

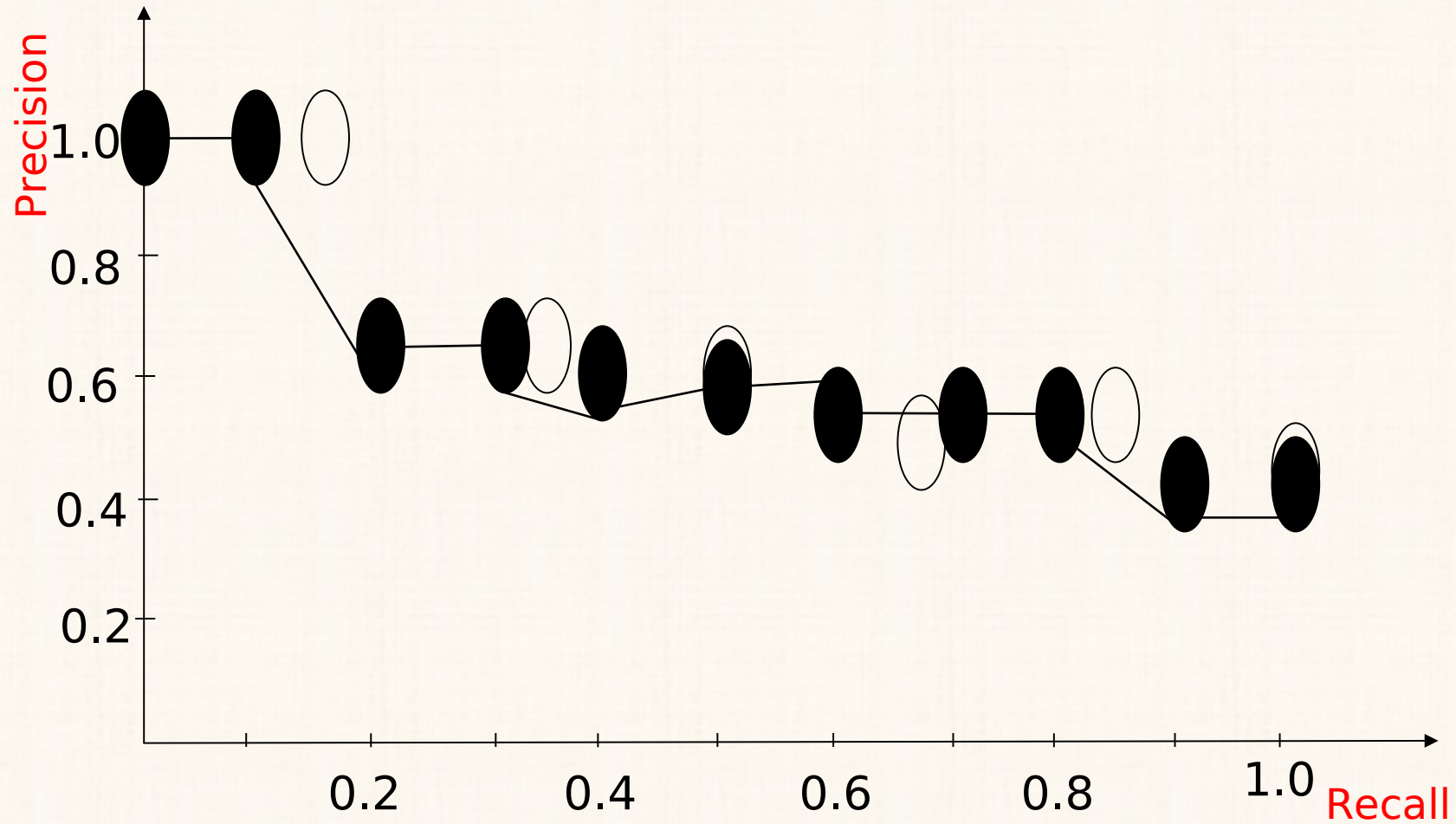
$R=5/6=0.833;$

$P=5/9=0.556$

$R=6/6=1.0;$

$p=6/14=0.429$

Interpolating a Recall/Precision Curve: Exercise 2



Interpolating across queries

- For each query, calculate precision at 11 standard recall levels
- 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.
- Average precision favors systems which produce relevant documents high in rankings

Single-valued measures

- Single value measures: calculate a single value for each query to evaluate performance
 - Mean Average Precision at seen relevant documents
 - Typically average performance over a large *set* of queries.
 - R-Precision
 - Precision at R^{th} relevant documents

MAP (Mean Average Precision)

- Computing mean average for more than one query
$$MAP = \frac{1}{n} \left[\sum_{Q_i} \frac{1}{|R_i|} \left(\sum_{D_j \in R_i} \frac{j}{r_{ij}} \right) \right]$$
 - r_{ij} = rank of the j^{th} relevant document for Q_i
 - $|R_i|$ = number of relevant documents for Q_i
 - n = number of test queries
- E.g. Assume there are 3 relevant documents for Query 1 and 2 for query 2. Calculate MAP?

Relevant Docs. retrieved	Query 1	Query 2
1 st rel. doc.	1	4
2 nd rel. doc.	5	8
3 rd rel. doc.	10	-

$$MAP = \frac{1}{2} \left[\frac{1}{3} \left(\frac{1}{1} + \frac{2}{5} + \frac{3}{10} \right) + \frac{1}{2} \left(\frac{1}{4} + \frac{2}{8} \right) \right]$$

R- Precision

- Precision at the R-th position in the ranking of results for a query, where R is the total number of relevant documents.
 - Calculate precision after R documents are seen
 - Can be averaged over all queries

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 = 6

R-Precision = $4/6 = 0.67$

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.
- What if no relevant documents exist?

E Measure

- Associated with Van Rijsbergen
- Allows user to specify importance of recall and precision
- It is 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:
 - $\beta = 1$: Equal weight for precision and recall ($E=F$).
 - $\beta > 1$: Weight recall more. It emphasizes recall.
 - $\beta < 1$: Weight precision more. It emphasizes precision.

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.

Other measures

- Noise = retrieved irrelevant docs / retrieved docs
- Silence/Miss = non-retrieved relevant docs / relevant docs
 - Noise = 1 - Precision; Silence = 1 - Recall

$$Miss = \frac{|\{Relevant\} \cap \{NotRetrieved\}|}{|\{Relevant\}|}$$

$$Fallout = \frac{|\{Retrieved\} \cap \{NotRelevant\}|}{|\{NotRelevant\}|}$$