Evaluating IR

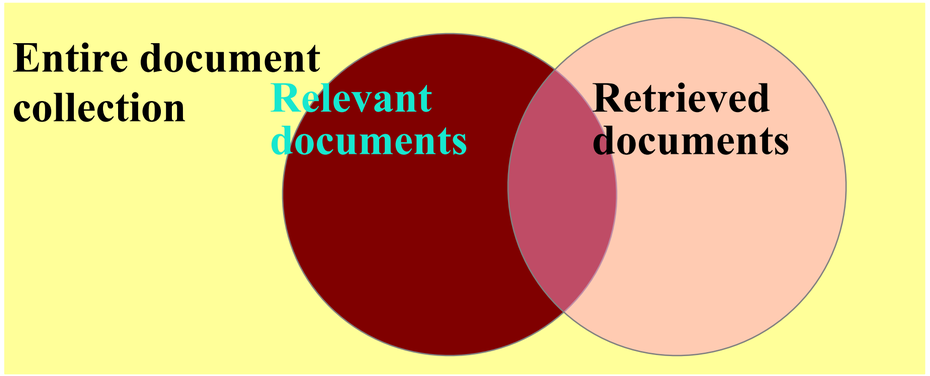
Difficulties in Evaluating IR Systems

- Effectiveness is related to the ***relevancy*** of retrieved items.
- Relevancy is not typically binary but continuous.
- Even if relevancy is binary, it can be a difficult judgment to make.
- Relevancy, from a human standpoint, is:
  - 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.

Precision and Recall

- Data is highly skewed to non-relevant
  - Max “accuracy” is almost always to NO.
  - But users want results so need stats that reward saying yes
- 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.

Precision and Recall



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

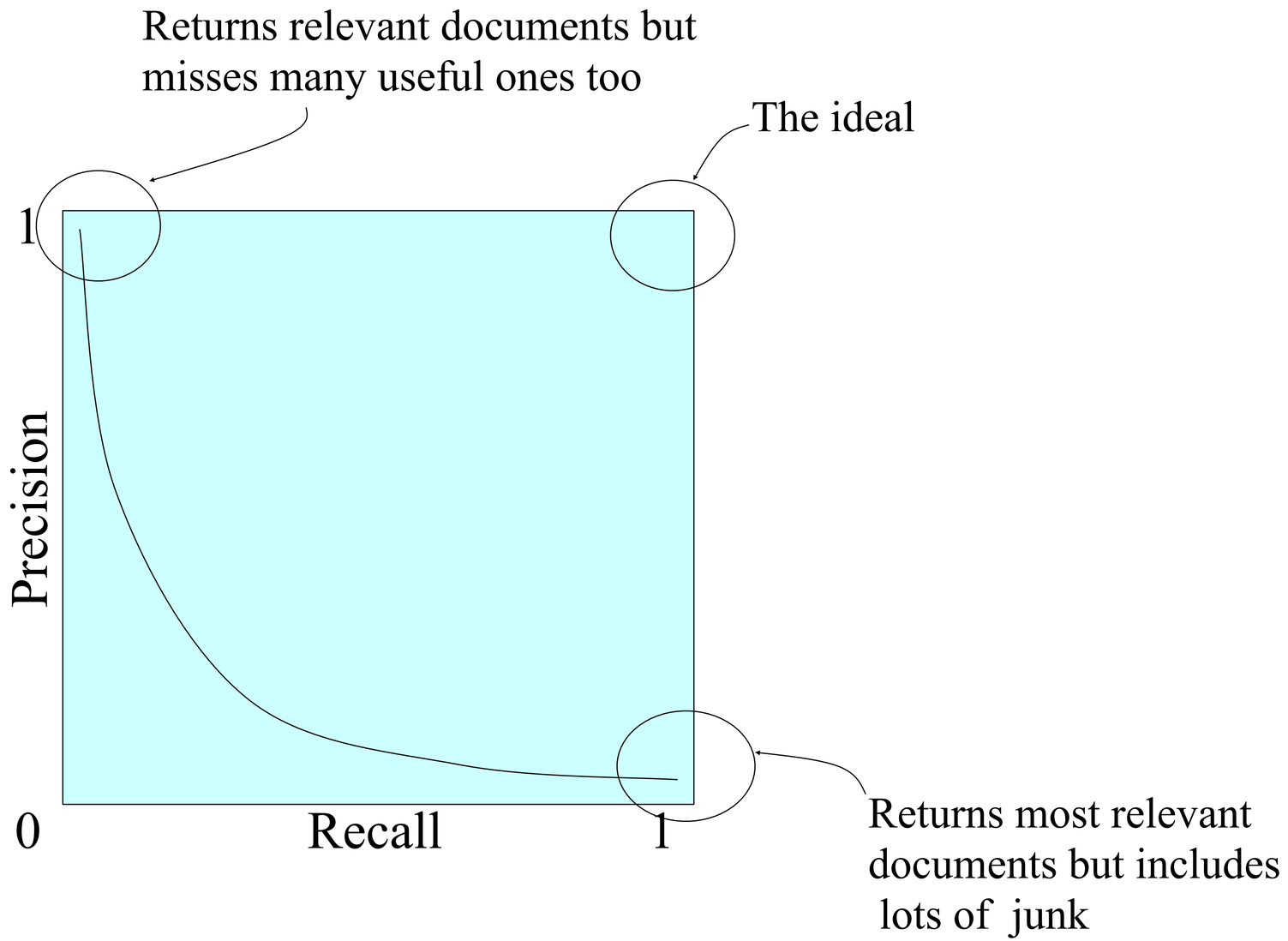
$$recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}$$

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

Precision:Easy Recall:Hard

- Total number of relevant items is sometimes not available:
  - Handle labelling by sampling: 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.
    - In the absence of hand-labelled data can you do better?

Trade-off between Recall and Precision



Computing Recall/Precision Points

- For a given query, produce the ranked list of retrievals.
- Mark each document in the ranked list that is relevant according to your gold standard.
- Compute a recall/precision pair for each position in the ranked list that contains a relevant document.
  - Or at least precision, since recall requires knowing the entire set of relevant documents

Computing Recall/Precision Points

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	

Suppose total # of relevant docs = 6
Check each new recall point:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; p=5/13=0.38

Missing one relevant document.

Computing Recall/Precision Points

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

Same data, same query, different System
total # of relevant docs = 6

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

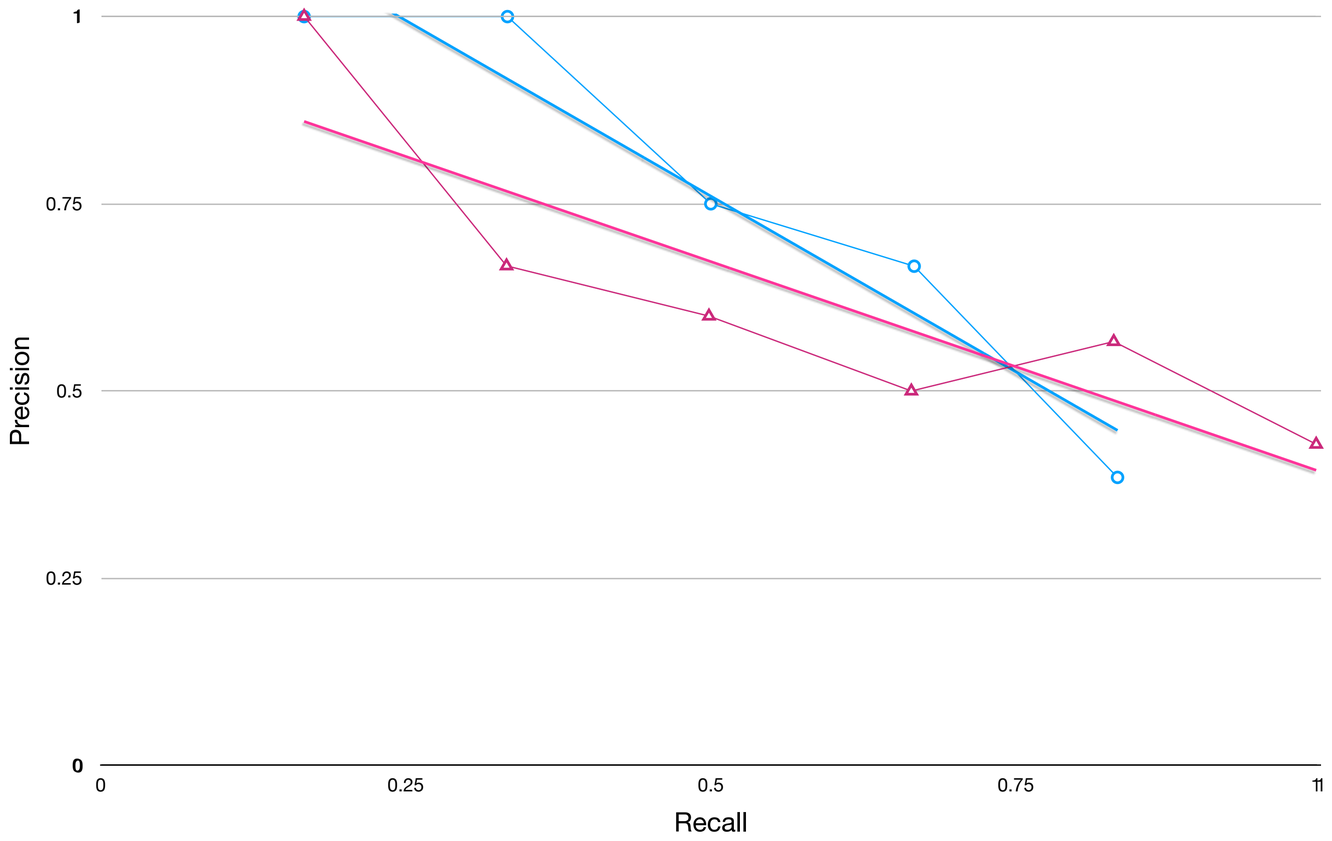
R=5/6=0.833; P=5/9=0.556

R=6/6=1.0; p=6/14=0.429

Q

- An IR system returns 8 relevant documents and 10 non-relevant. There are 20 relevant documents. What is the recall? Precision?

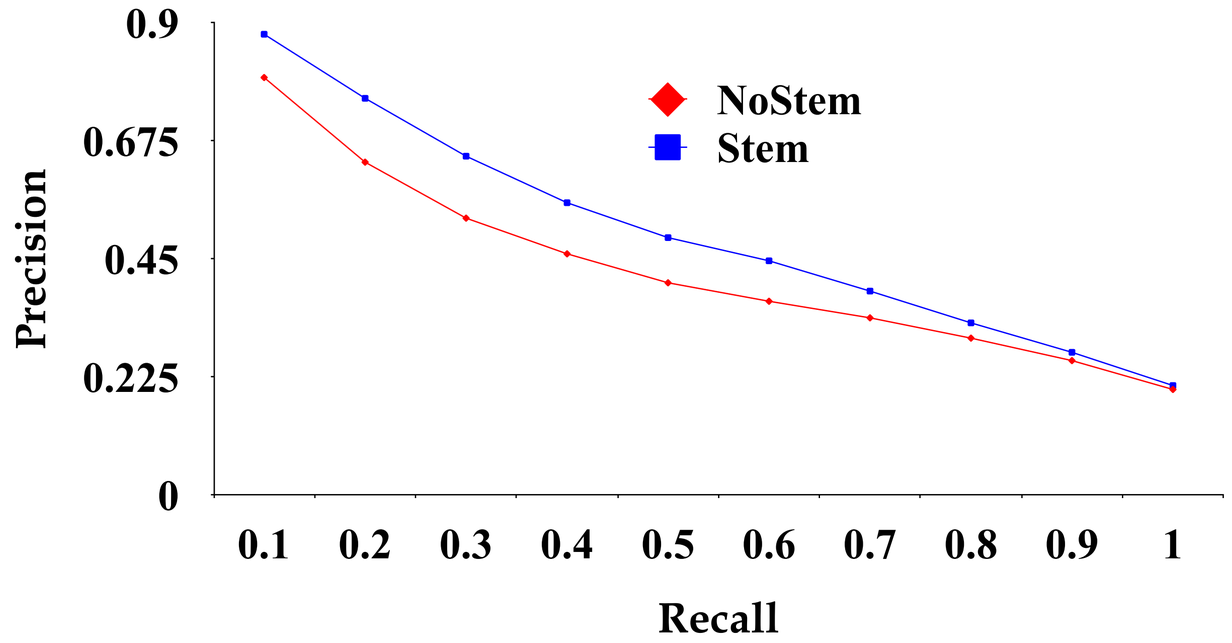
Precision/Recall Curve



For a single query
Which curve is better?

Compare Two or More Systems

- The curve closest to the upper right-hand corner of the graph indicates the best performance



Non-Binary Relevance

- Documents are rarely entirely relevant or non-relevant to a query
- Many sources of *graded relevance judgments*
  - Relevance judgments on a 5-point scale
  - Multiple judges
  - Click distribution and deviation from expected levels (but click-through != relevance judgments)

Cumulative Gain

- With graded relevance judgments, we can compute the *gain* at each rank.
- **Cumulative Gain** at rank n:

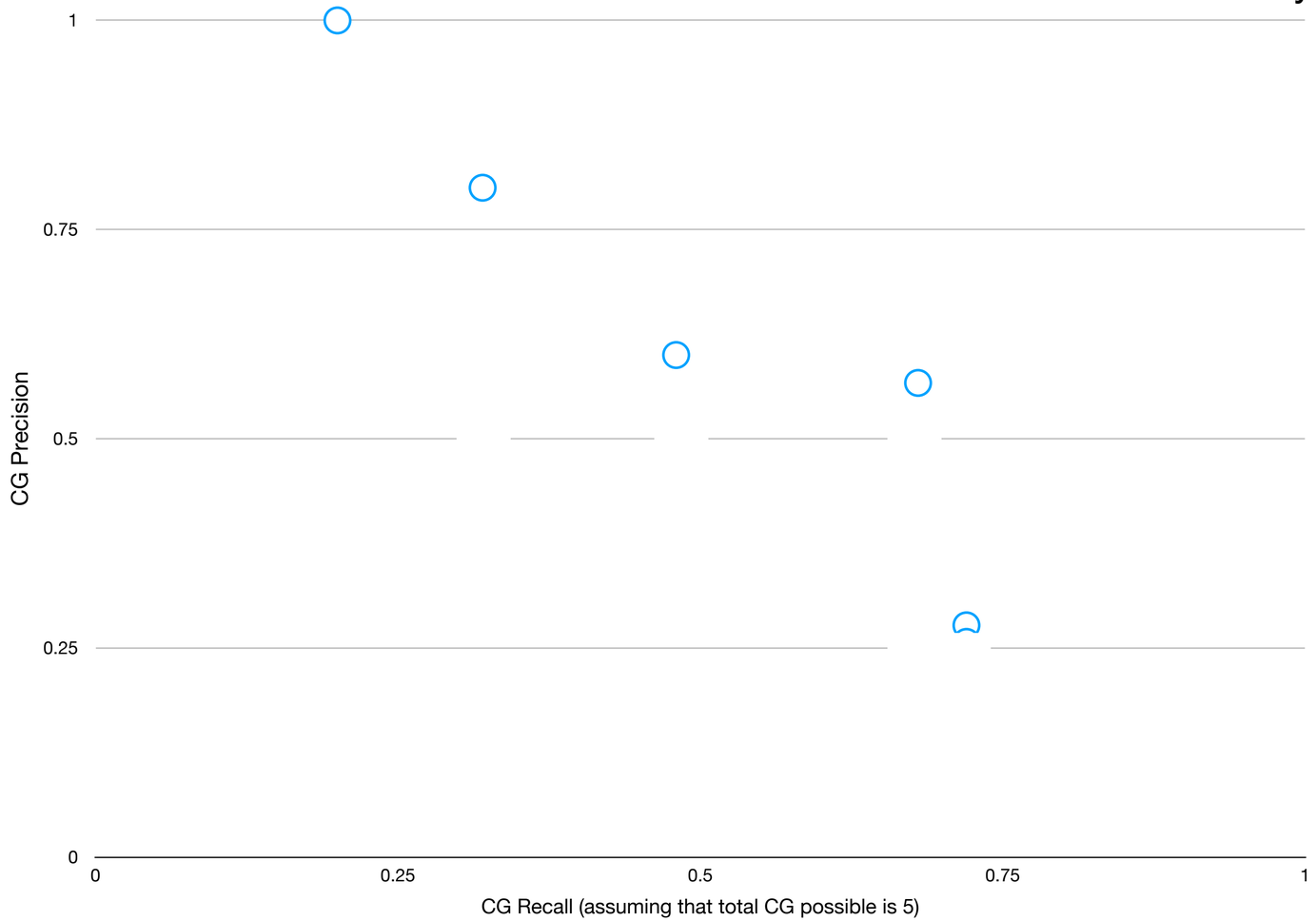
$$CG_n = \sum_{i=1}^{n} rel_i$$

(Where rel_i is the graded relevance of the document at position *i)*

n	doc #	relevance (gain)	CG_n
1	588	1.0	1.0
2	589	0.6	1.6
3	576	0.0	1.6
4	590	0.8	2.4
5	986	0.0	2.4
6	592	1.0	3.4
7	984	0.0	3.4
8	988	0.0	3.4
9	578	0.0	3.4
10	985	0.0	3.4
11	103	0.0	3.4
12	591	0.0	3.4
13	772	0.2	3.6
14	990	0.0	3.6

CG Precision Recall

(1,1) may not be even theoretically possible



Issues with Relevance

- ***Marginal Relevance:*** Do later documents in the ranking add new information beyond what is already given in higher documents.
  - Choice of retrieved set should encourage **diversity** and **novelty.**
- ***Coverage Ratio***: The proportion of relevant items retrieved out of the total relevant documents ***known*** to a user prior to the search.
  - Relevant when the user wants to locate documents which they have seen before (e.g., the budget report for Year 2000).

A/B Testing in a Deployed System

- Can exploit an existing user base to provide useful feedback.
- Randomly send a small fraction (1−10%) of incoming users to a variant of the system that includes a single change.
- Judge effectiveness by measuring change in ***clickthrough:*** The percentage of users that click on the top result (or any result on the first page).