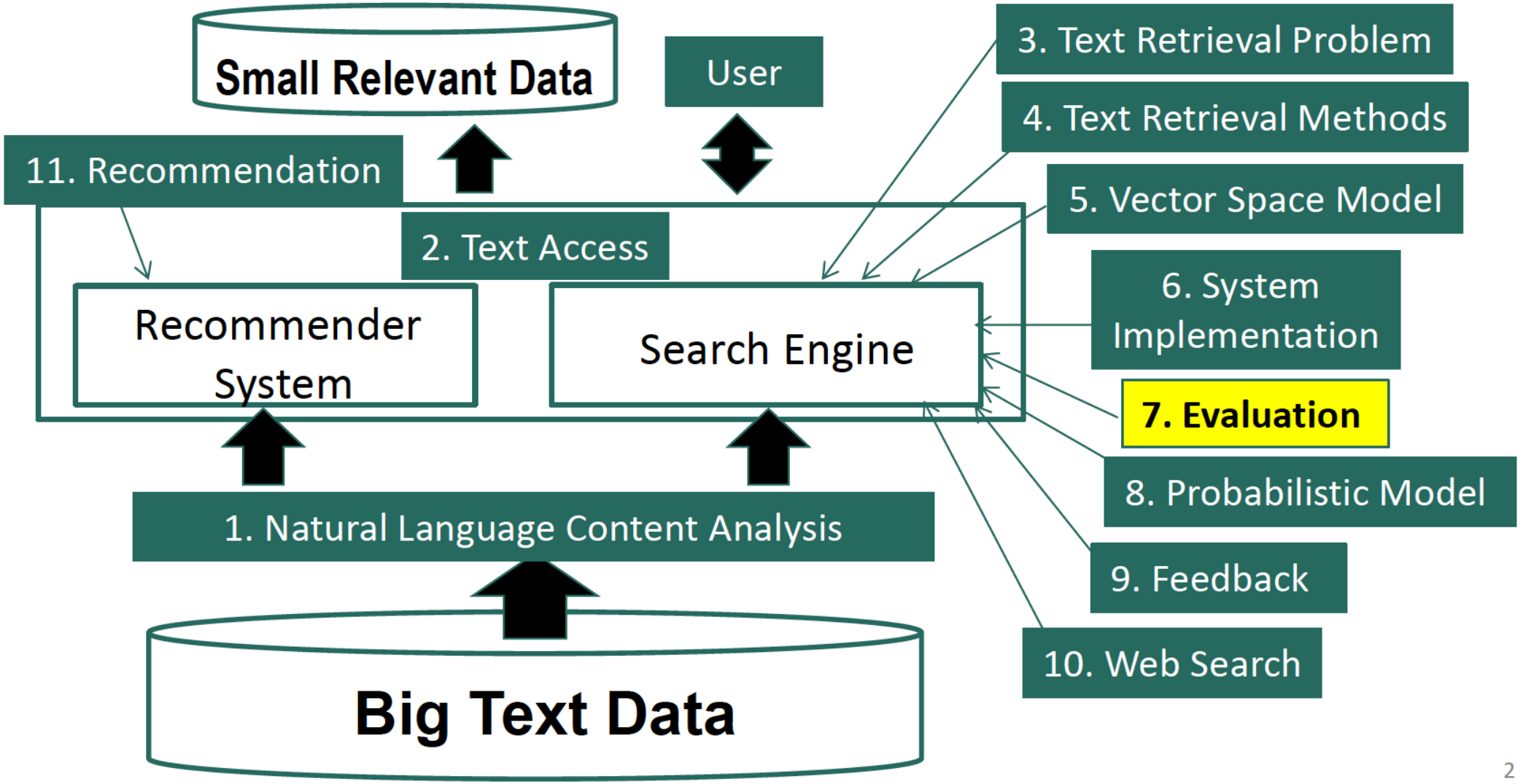
Information Retrieval & Text Mining

Evaluation of Text Retrieval Systems:

Evaluating a Ranked List

Dr. Saeed Ul Hassan
Information Technology University

Evaluation of Text Retrieval Systems



Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +		
D3 –		
D4 –		
D5 +		
D6 –		
D7 –		
D8 +		
D9 –		
D10 –		

Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +		
D6 –		
D7 –		
D8 +		
D9 –		
D10 –		

Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +		
D9 –		
D10 –		

Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –		

Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

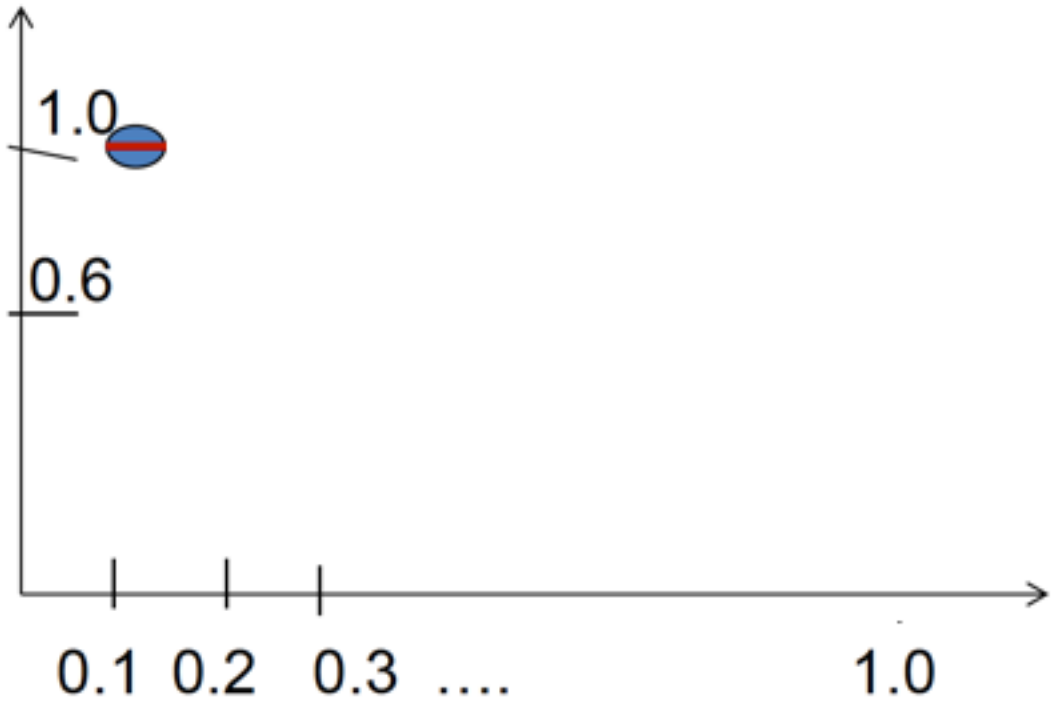
	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	10/10

Assume Precision=0?

Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	10/10

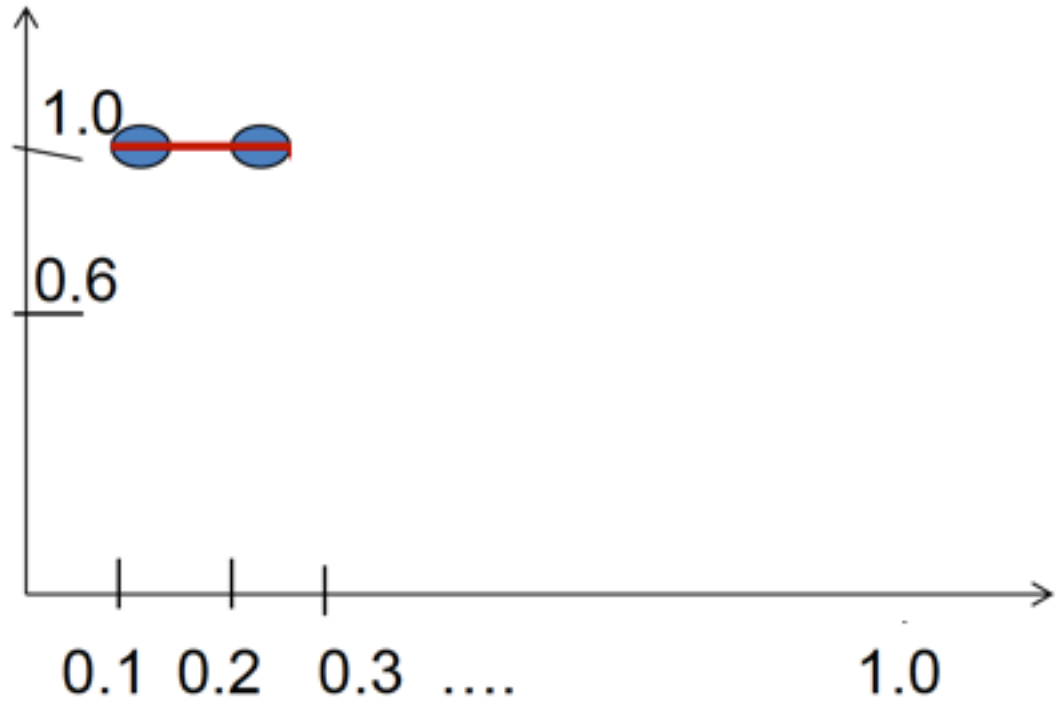


Assume Precision=0?

Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	10/10

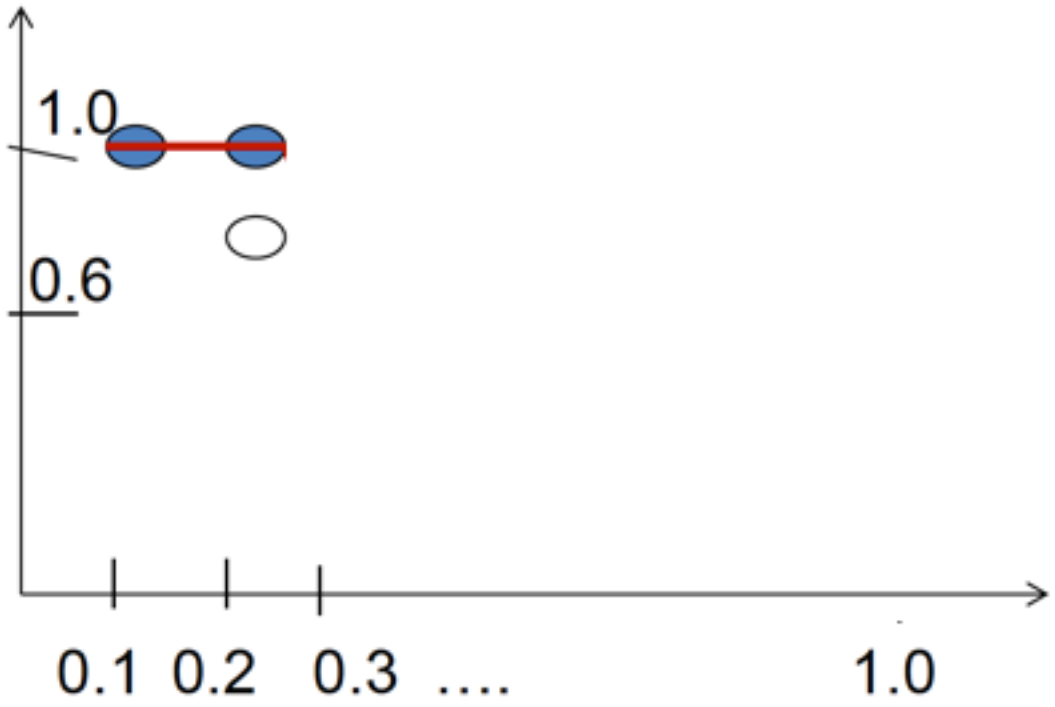


Assume Precision=0?

Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	10/10

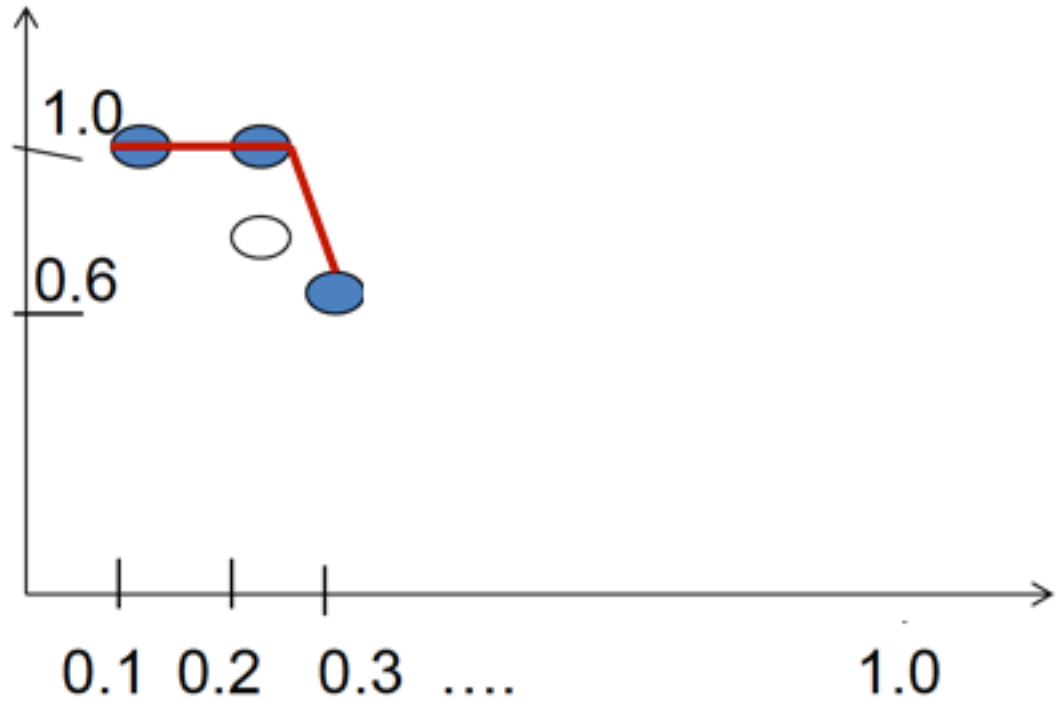


Assume Precision=0?

Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	10/10

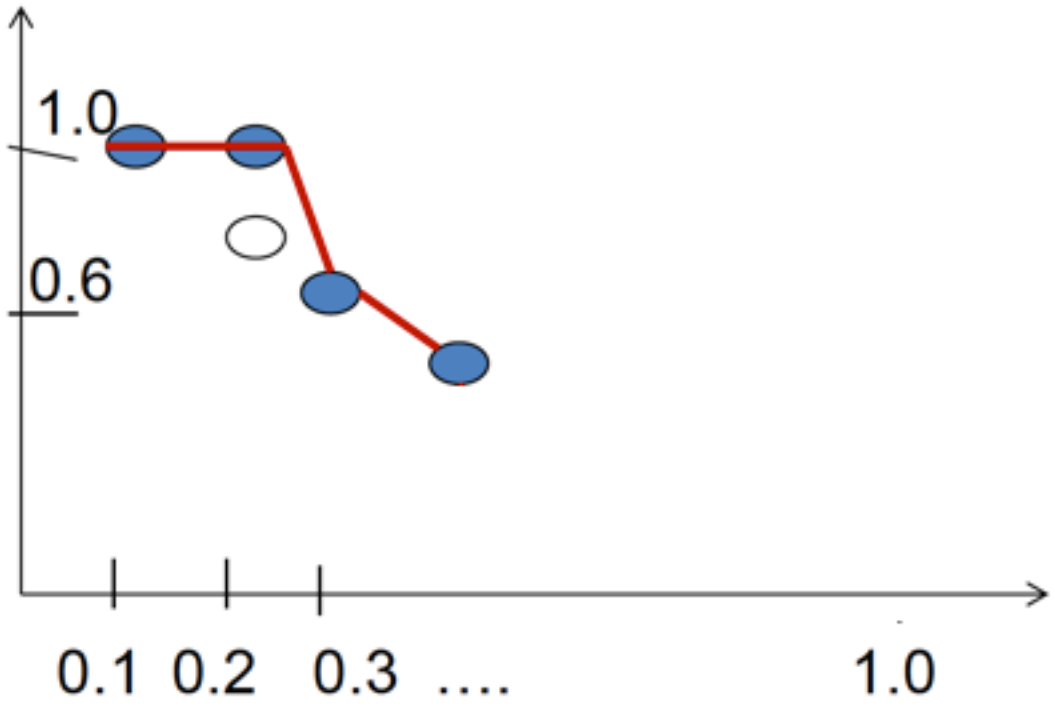


Assume Precision=0?

Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	10/10

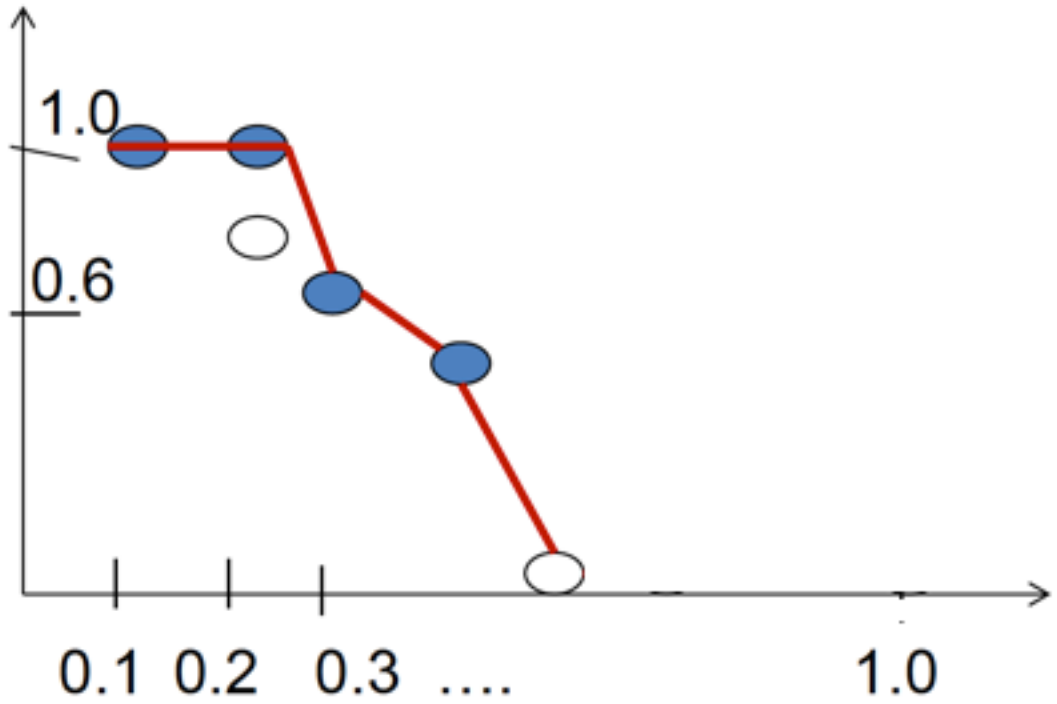


Assume Precision=0?

Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	10/10

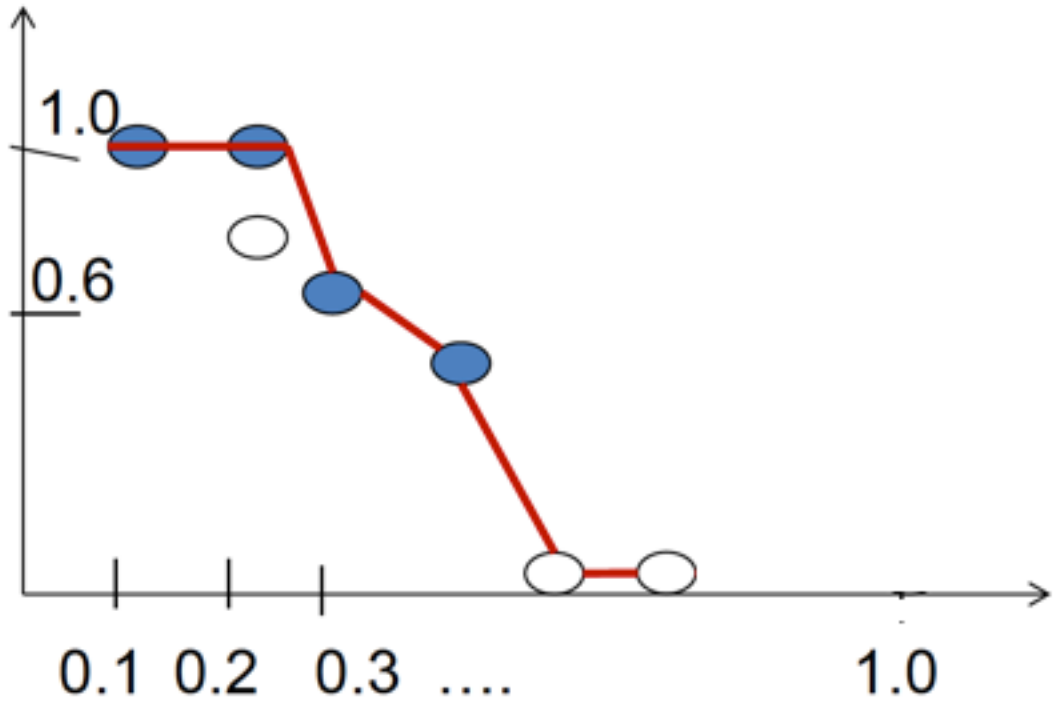


Assume Precision=0?

Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	10/10

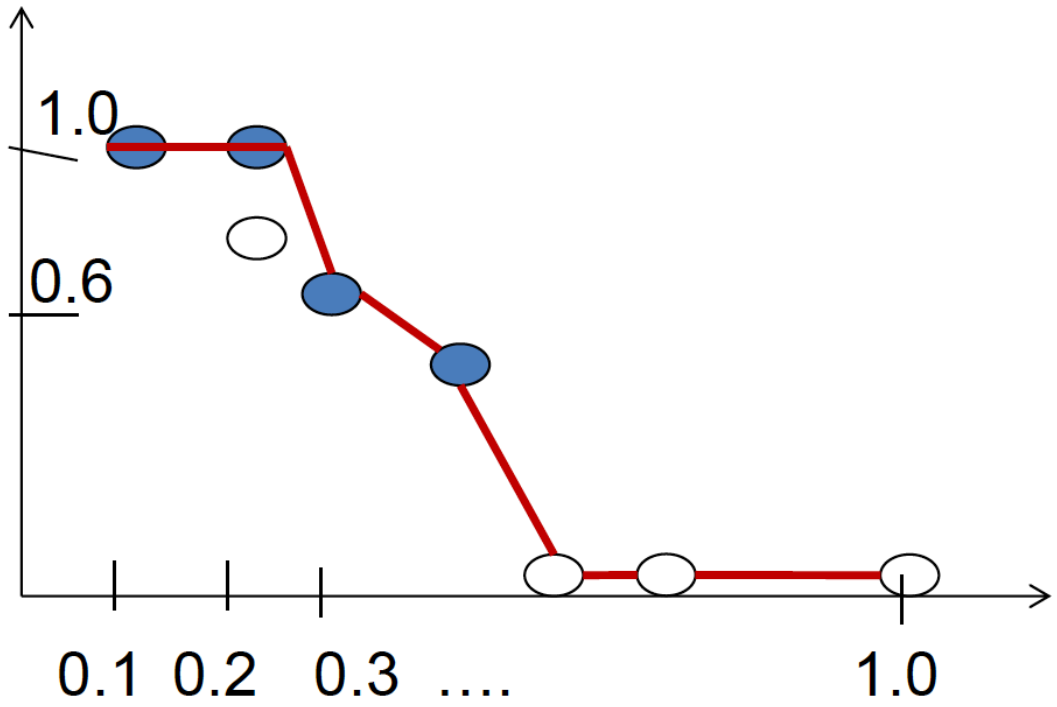


Assume Precision=0?

Evaluating Ranking: Precision-Recall (PR) Curve

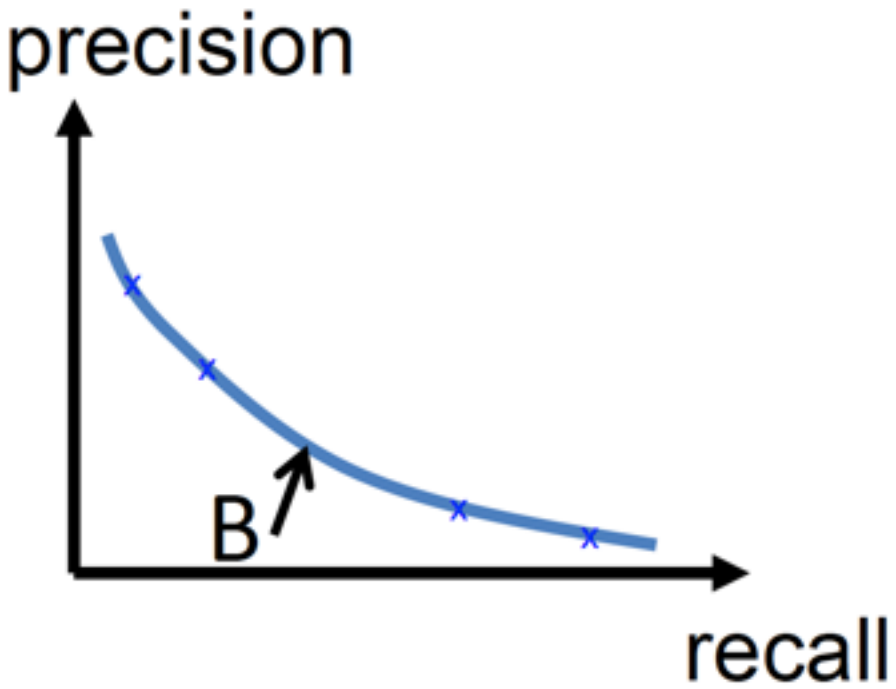
Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	10/10

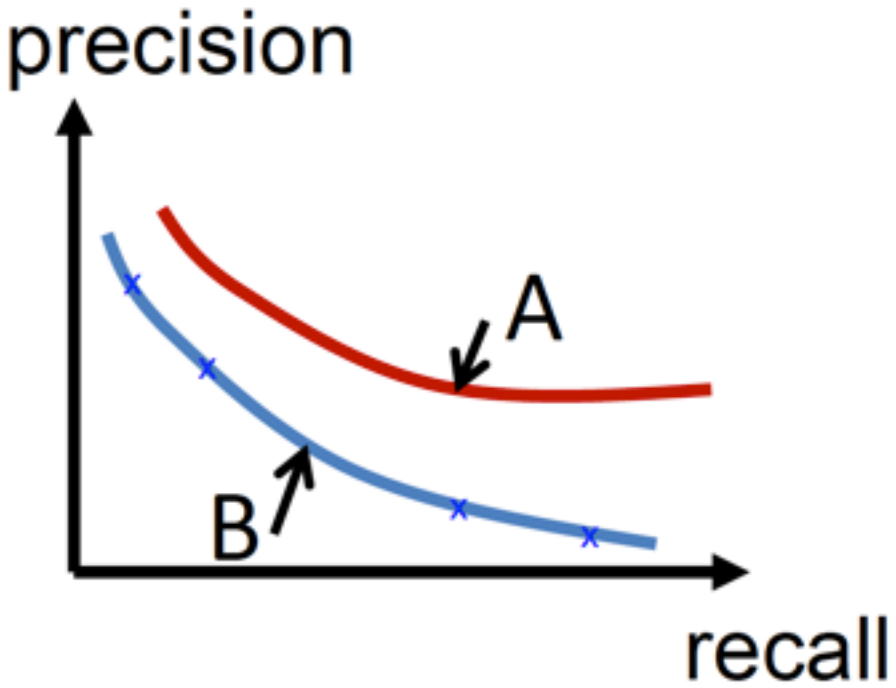


Assume Precision=0?

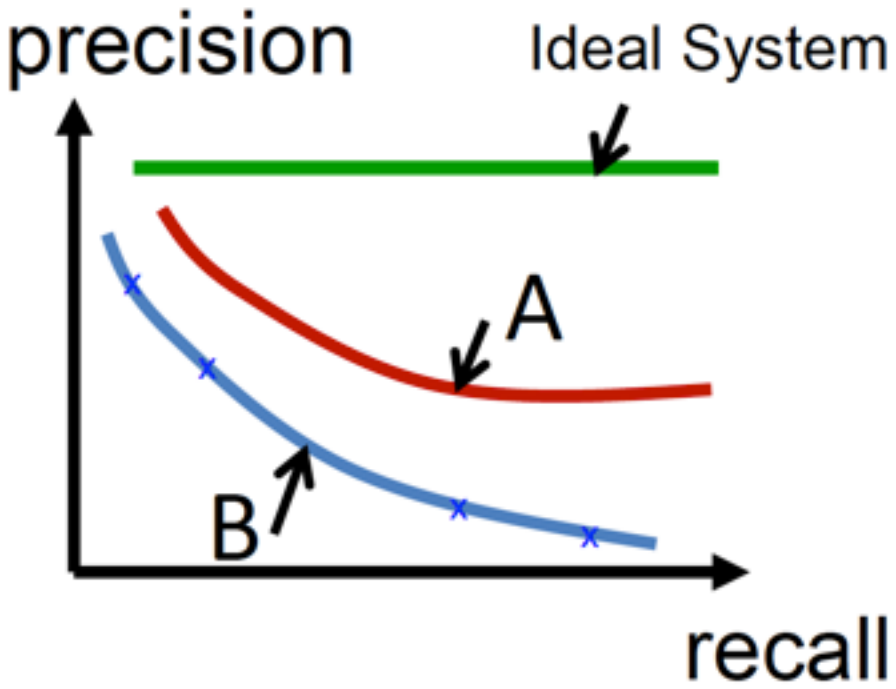
Comparing PR Curves



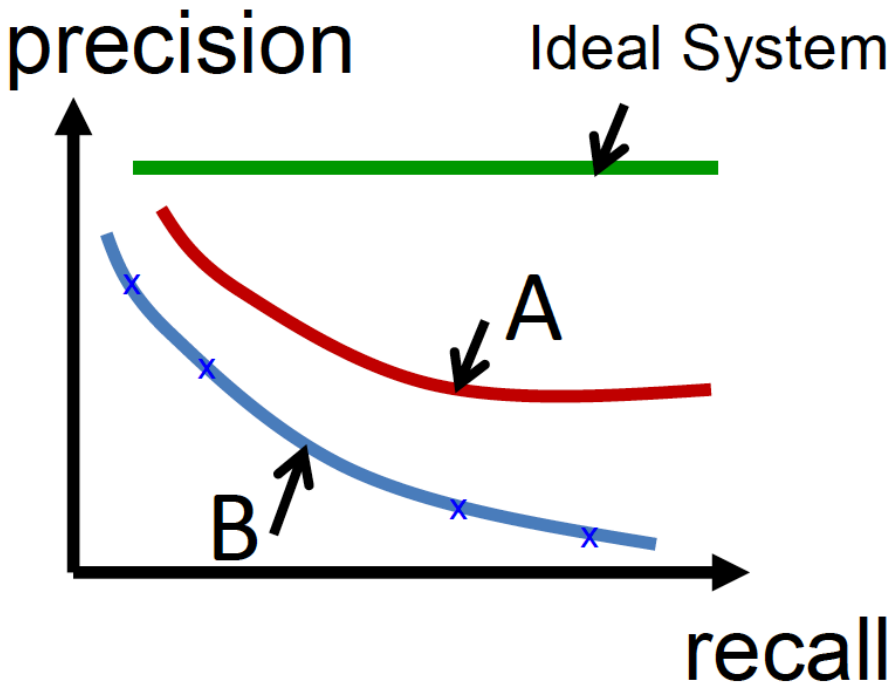
Comparing PR Curves

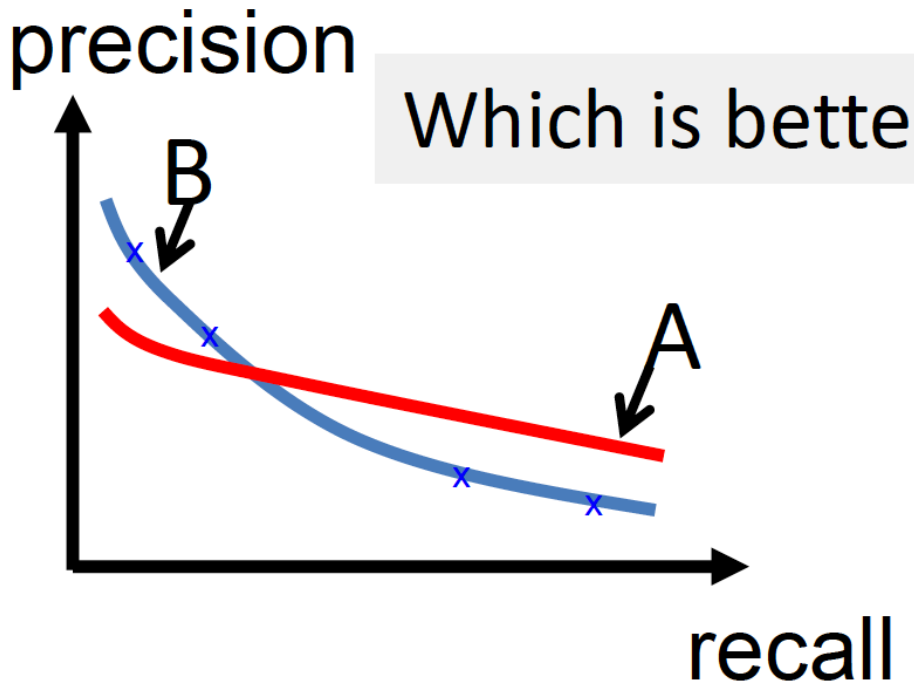


Comparing PR Curves



Comparing PR Curves





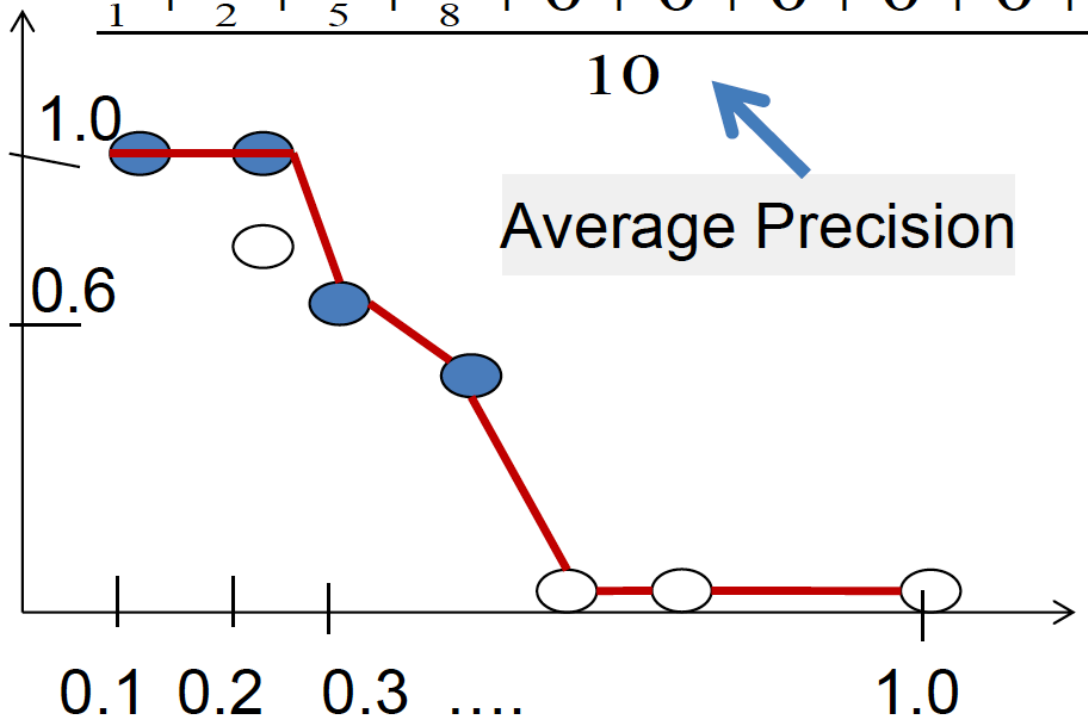
Which is better?

How to Summarize a Ranking

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	0	10/10

$$\frac{\frac{1}{1}+\frac{2}{2}+\frac{3}{5}+\frac{4}{8}+0+0+0+0+0+0}{10}$$



Average Precision

Mean Average Precision (MAP)

- Average Precision:
  - The average of precision at every cutoff where a new relevant document is retrieved
  - Normalizer = the total # of relevant docs in collection
  - Sensitive to the rank of each relevant document
- Mean Average Precisions (MAP)
  - MAP = arithmetic mean of average precision over a set of queries

Special Case: Mean Reciprocal Rank

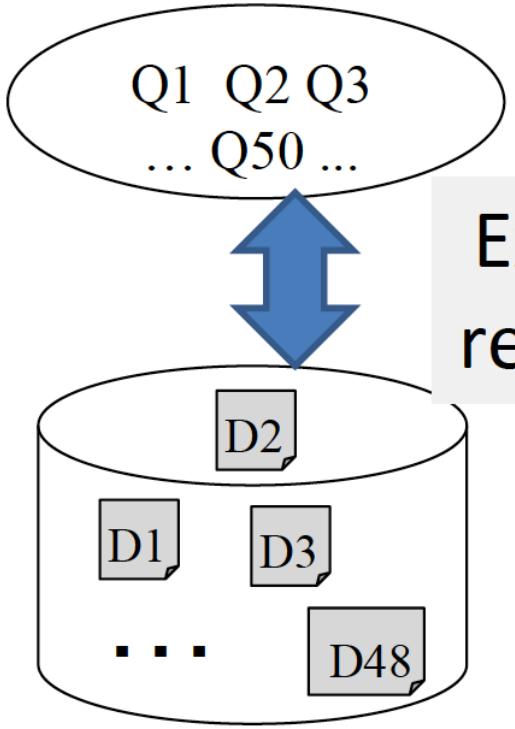
- When there's only one relevant document in the collection (e.g., known item search)
  - Average Precision = Reciprocal Rank = 1/r, where r is the rank position of the single relevant doc
  - Mean Average Precision → Mean Reciprocal Rank
  - Why not simply use r?

Summary

- Precision-Recall curve characterizes the overall accuracy of a ranked list
- The **actual** utility of a ranked list depends on how many top-ranked results a user would examine
- Average Precision is the standard measure for comparing two ranking methods
  - Combines precision and recall
  - Sensitive to the rank of **every** relevant document

Challenges in Creating a Test Collection

Queries: representative & many



Existence of relevant docs

Measures: capture the perceived utility by users

Docs: representative & many

Relevance Judgments

Judgments: completeness vs. minimum human work

…
Q2 D1 –
Q2 D2 +
Q2 D3 +
Q2 D4 –
…
Q50 D1 –
Q50 D2 –
Q50 D3 +

Statistical Significance Tests

- How sure can you be that an observed difference doesn't simply result from the particular queries you chose?

Experiment 1

Query	System A	System B
1	0.20	0.40
2	0.21	0.41
3	0.22	0.42
4	0.19	0.39
5	0.17	0.37
6	0.20	0.40
7	0.21	0.41
Average	0.20	0.40

Experiment 2

Query	System A	System B
1	0.02	0.76
2	0.39	0.07
3	0.16	0.37
4	0.58	0.21
5	0.04	0.02
6	0.09	0.91
7	0.12	0.46
Average	0.20	0.40

Pooling: Avoid Judging all Documents

- If we can't afford judging all the documents in the collection, which subset should we judge?
- Pooling strategy
  - Choose a diverse set of ranking methods (TR systems)
  - Have each to return top-K documents
  - Combine all the top-K sets to form a pool for human assessors to judge

Summary of TR Evaluation

- Extremely important!
  - TR is an empirically defined problem
  - Inappropriate experiment design misguides research and applications
  - Make sure to get it right for your research or application
- Cranfield evaluation methodology is the main paradigm
  - Precision@10docs is easier to interpret from a user's perspective
- Not covered
  - A-B Test [Sanderson 10]
  - User studies [Kelly 09]

Additional Readings

- Donna Harman, Information Retrieval Evaluation. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers 2011

- Mark Sanderson, Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval 4(4): 247-375 (2010)

- Diane Kelly, Methods for Evaluating Interactive Information Retrieval Systems with Users. Foundations and Trends in Information Retrieval 3(1-2): 1-224 (2009)