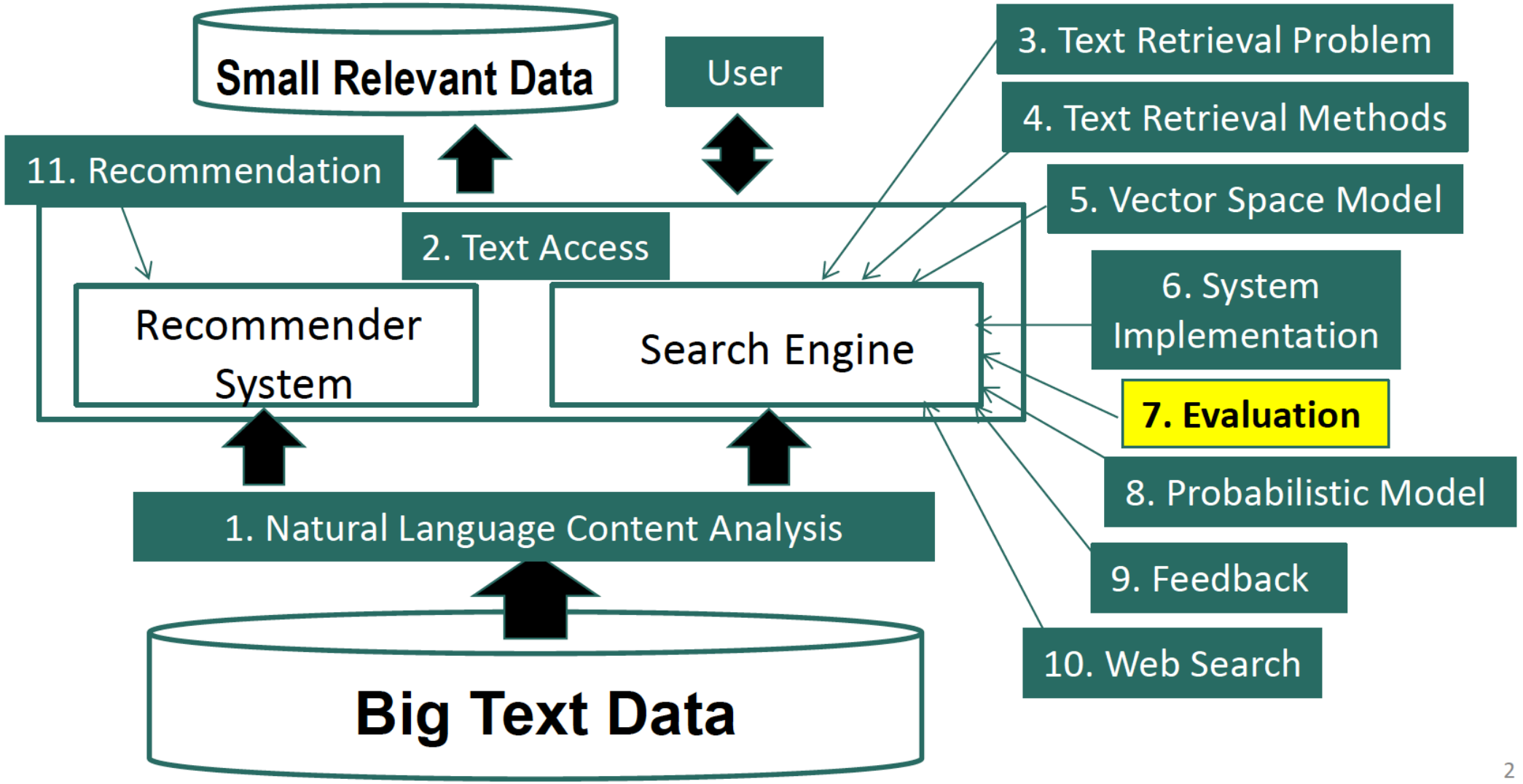
# Information Retrieval

## Evaluation of Text Retrieval Systems:

## Evaluating a Ranked List - Part 1 & 2

Dr. Iqra Safder

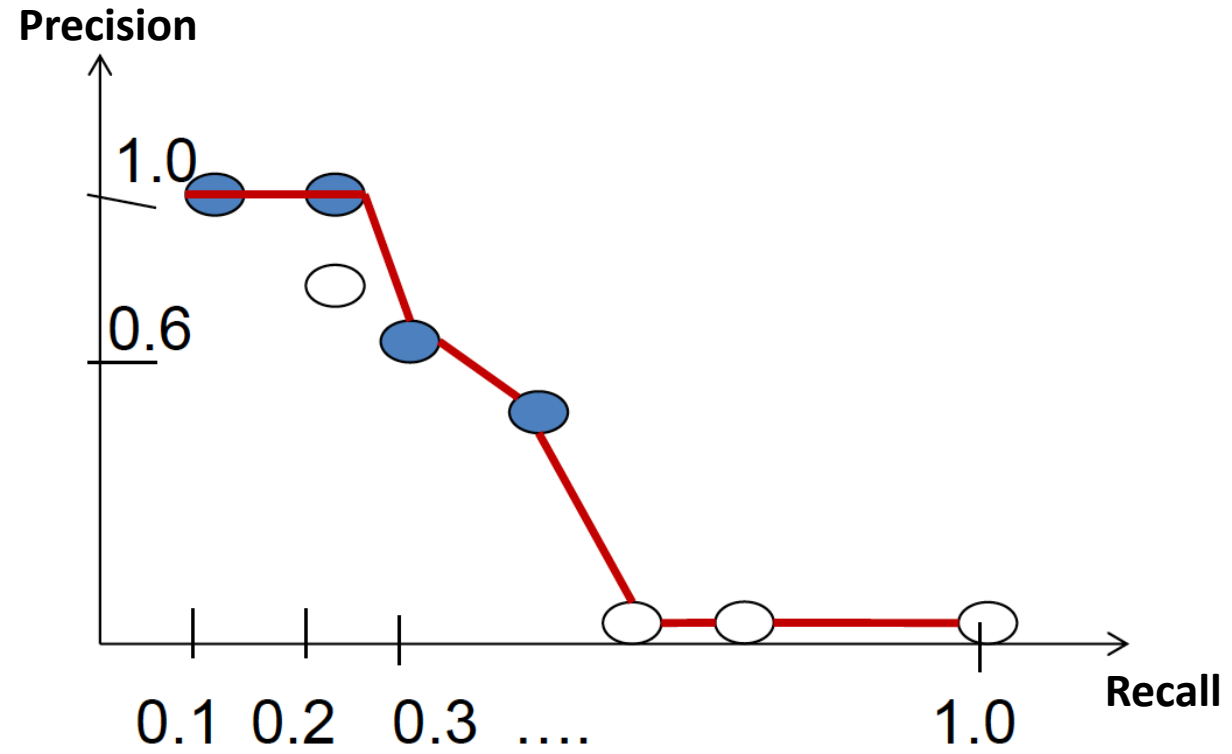
# Evaluation of Text Retrieval Systems



# Evaluating Ranking: Precision-Recall (PR) Curve

Total number of relevant documents in collection: 10

| | Precision | Recall |
|---|---|---|
| D1 + | 1/1 | 1/10 |
| D2 + | 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 | |

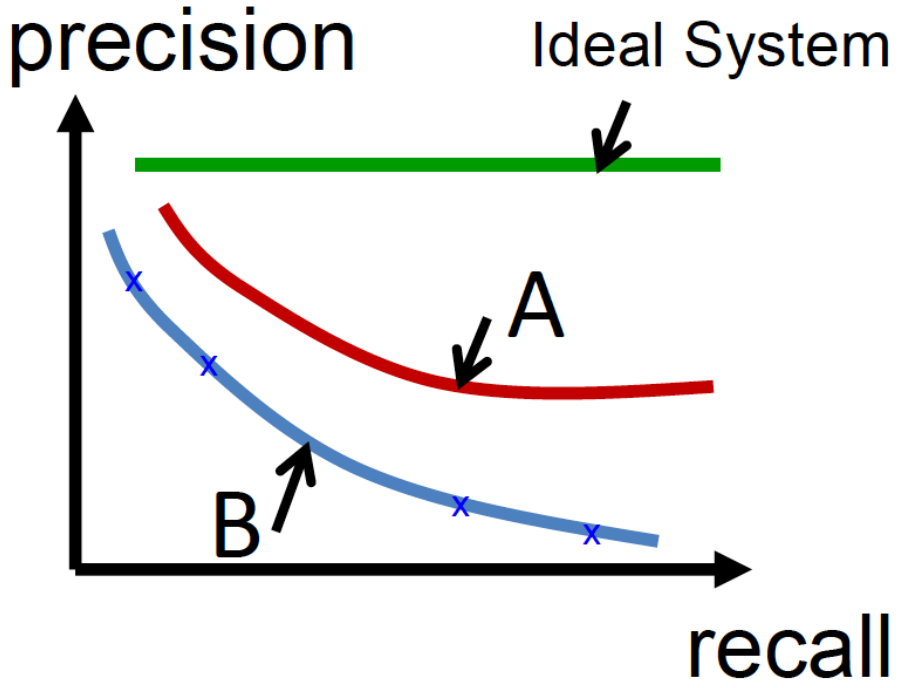


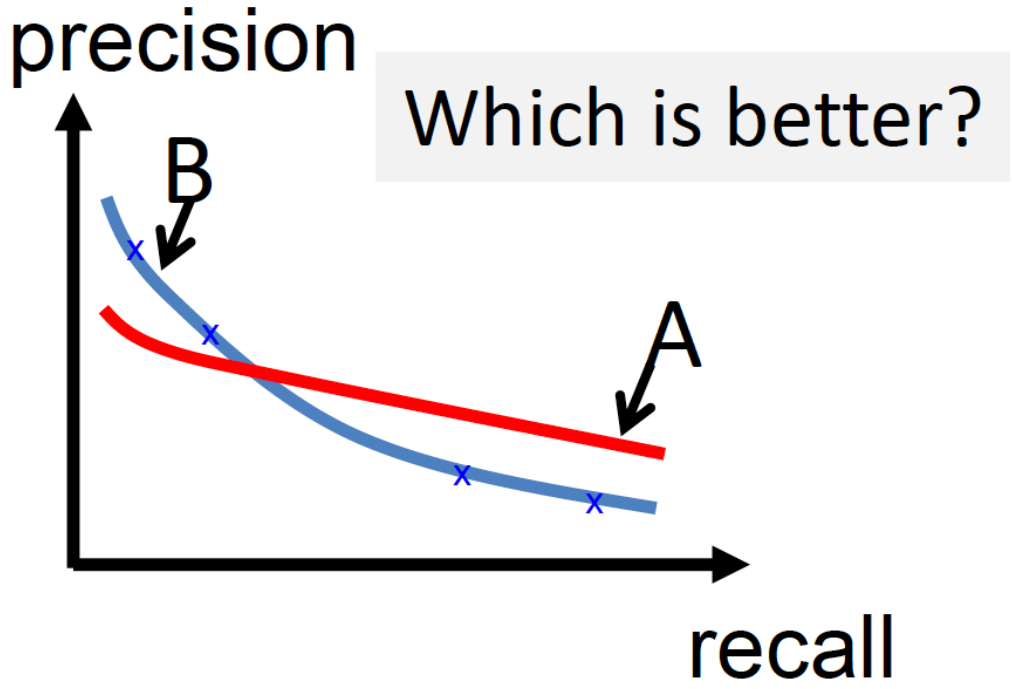

Assume Precision=0?

**Precision is zero at beyond the search results.**

**The key is to avoid any baise, its ok for a measure to deviate a little bit from true number.**

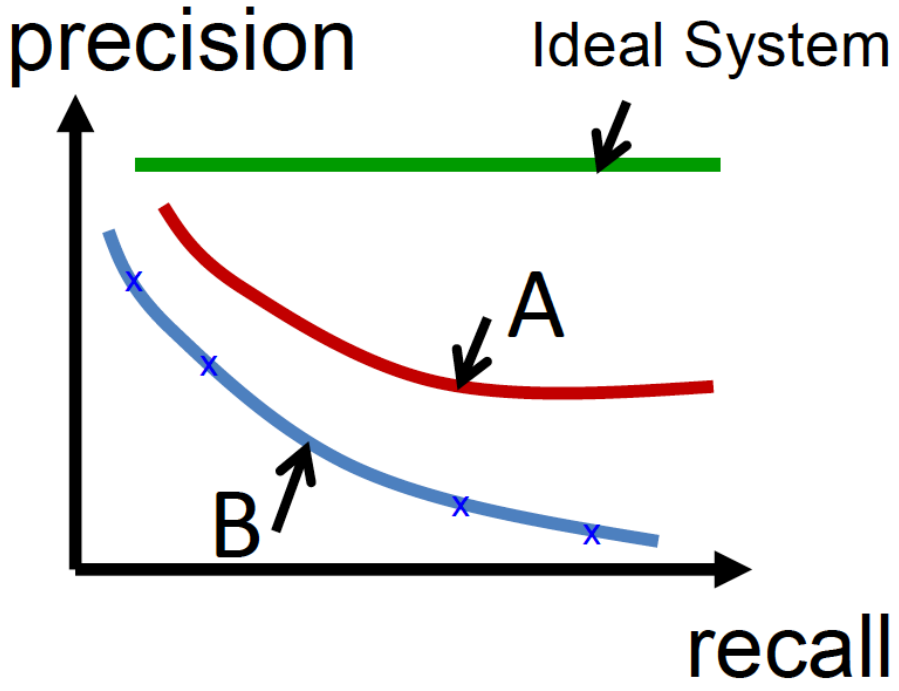
# Comparing PR Curves

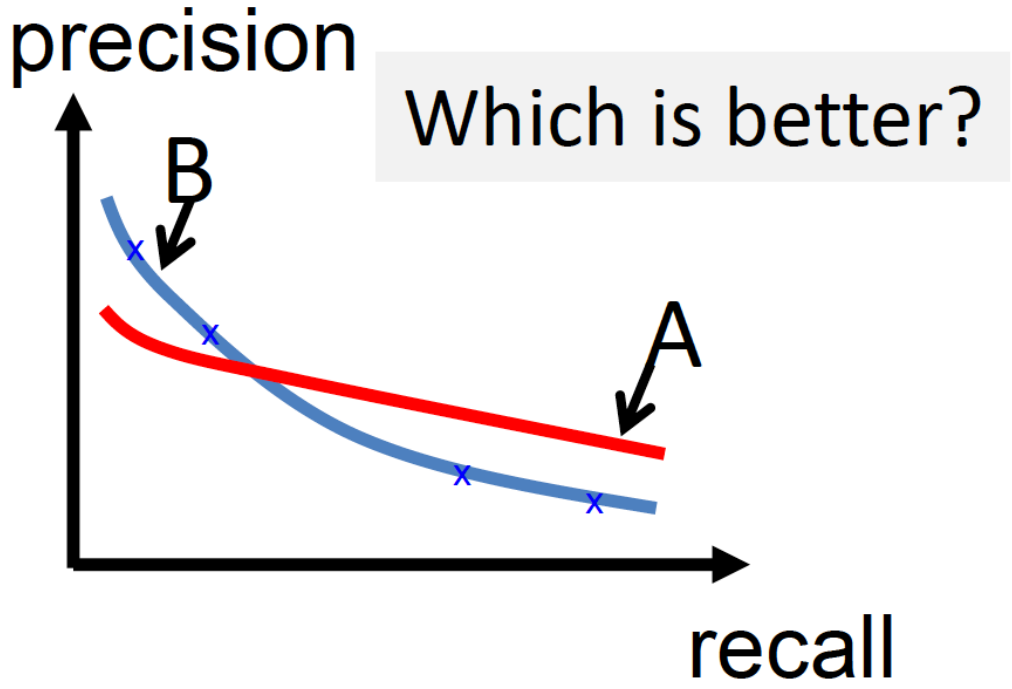



Higher the curve, its better.

# Comparing PR Curves



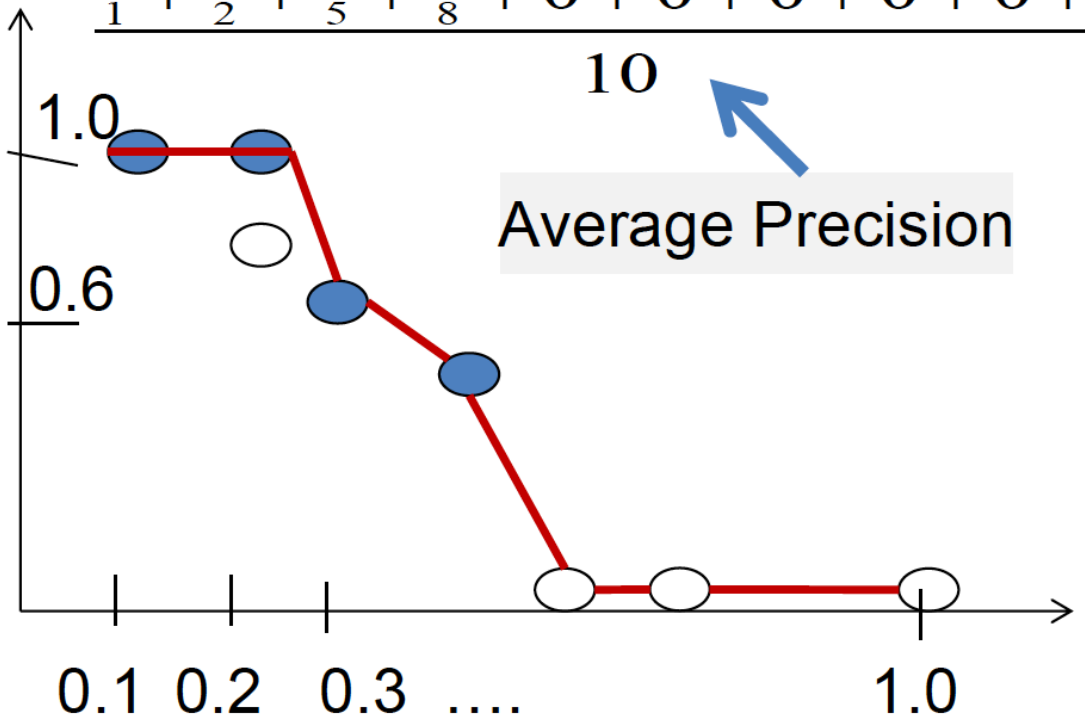

System B is better in lower recall region and System A is better in high recall region.
While Searching for a news user doesnot care about high recall. However while searching papers related to a problem, high recall is the need.
It all depends on users, more prcisely on users context.

# 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 |

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




Area under the curve (AUC)

We look at precison at different recall levels.

# How to Summarize a Ranking

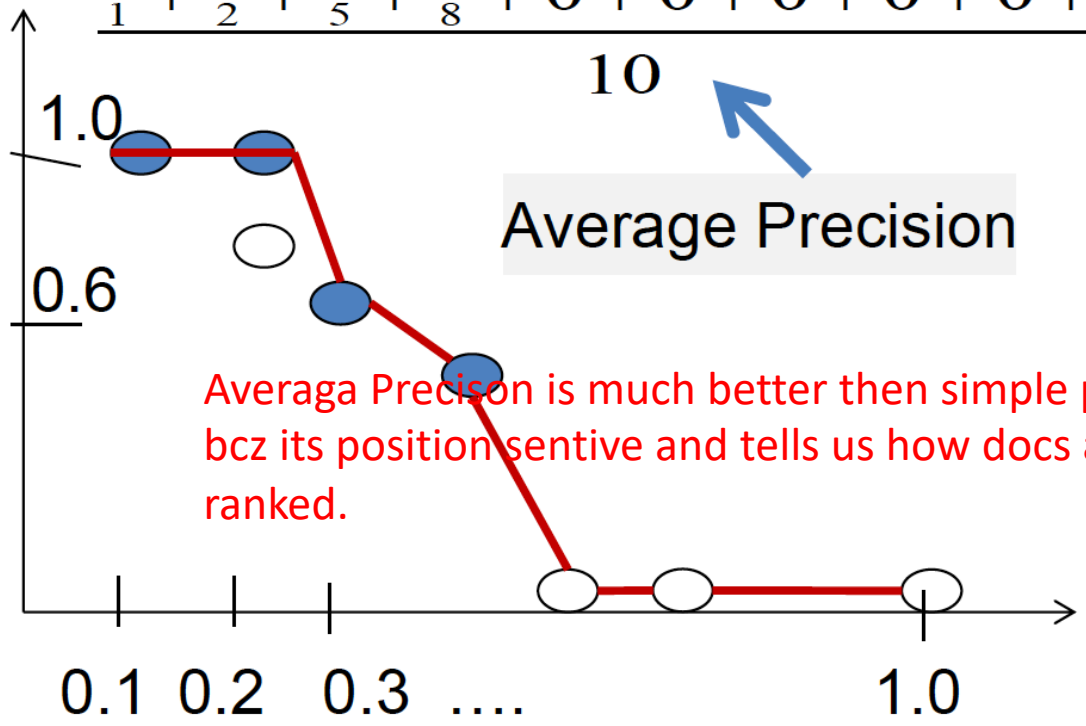
Total number of relevant documents in collection: 10

| | 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 |

This measure is very senstive to the rank of a document. It can tell small differences between two systems

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

Average Precision




Averaga Precison is much better then simple precision, bcz its position sentive and tells us how docs are ranked.

Why we divided with 10 not 4 (retrieved relevant). This way we will be favouring a system that shows few results…so avoid this common mistake always.

# How to Summarize a Ranking

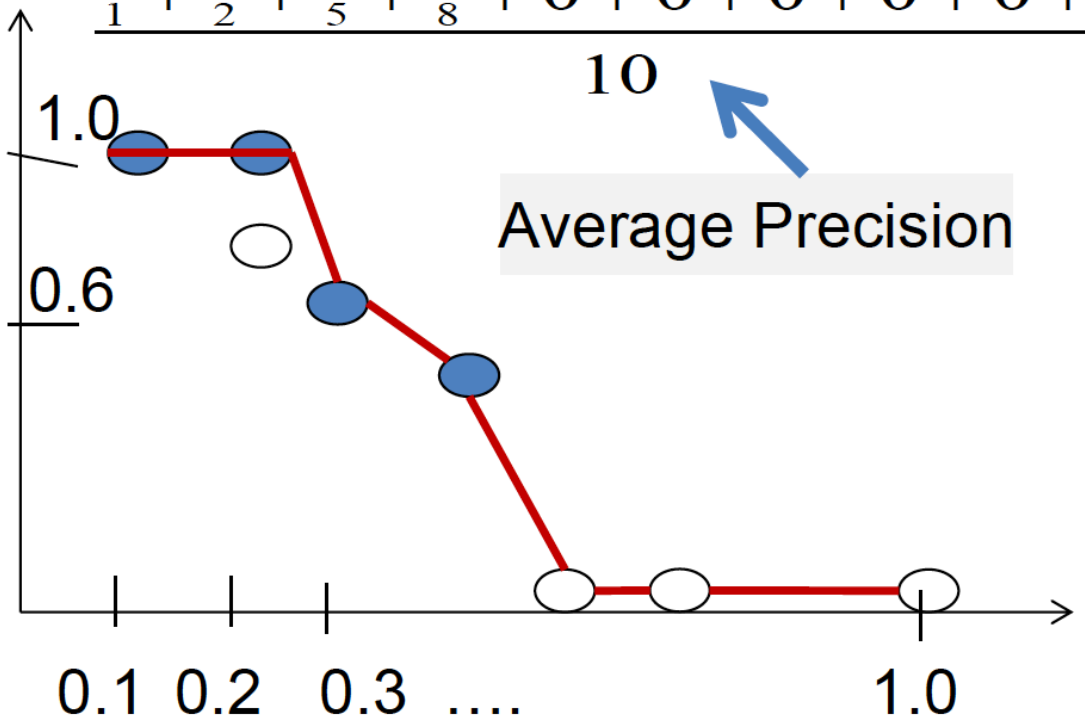
Total number of relevant documents in collection: 10

| | 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 |

This measure is very senstive to the rank of a document. It can tell small differences between two systems

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

Average Precision




Averaga Precison is much better then simple precision, becaucse it is position sentive and tells us where docs are ranked. Whereas simple precision at 10 will not have any effect with the change in positions in the rank list.

# 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
  - gMAP = geometric mean of average precision over a set of queries
  - Which is better: MAP or gMAP?

# 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

What if we have multiple levels of relevance judgments?