IR evaluation

Ack: Victor Lavrenko

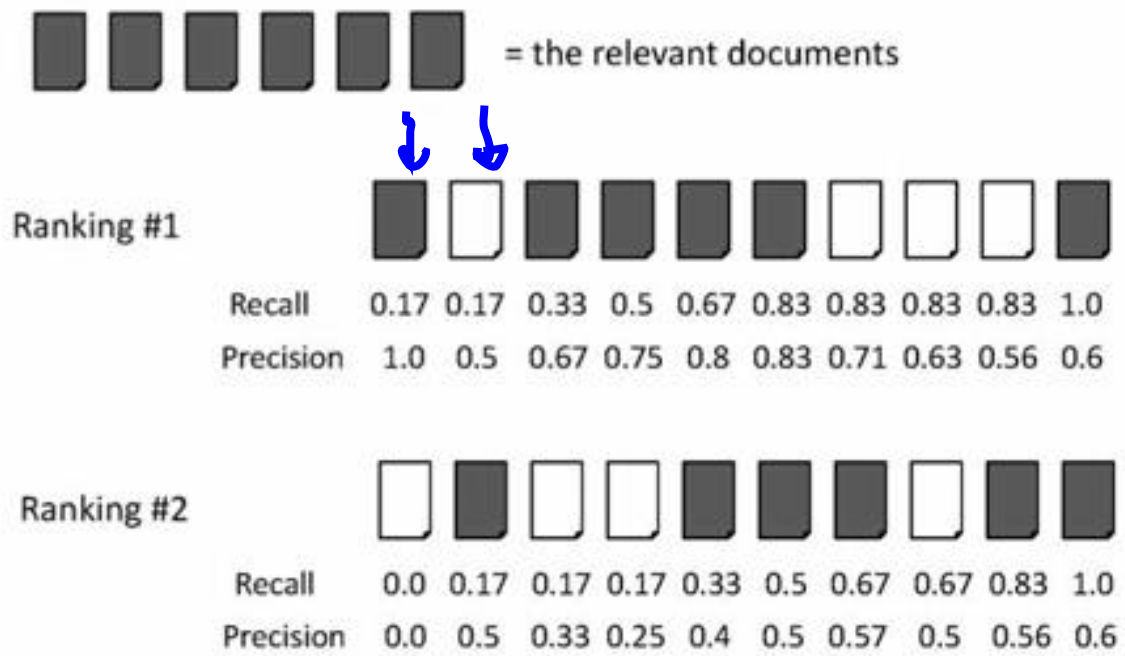
Why not accuracy?

Comparing recall / precision

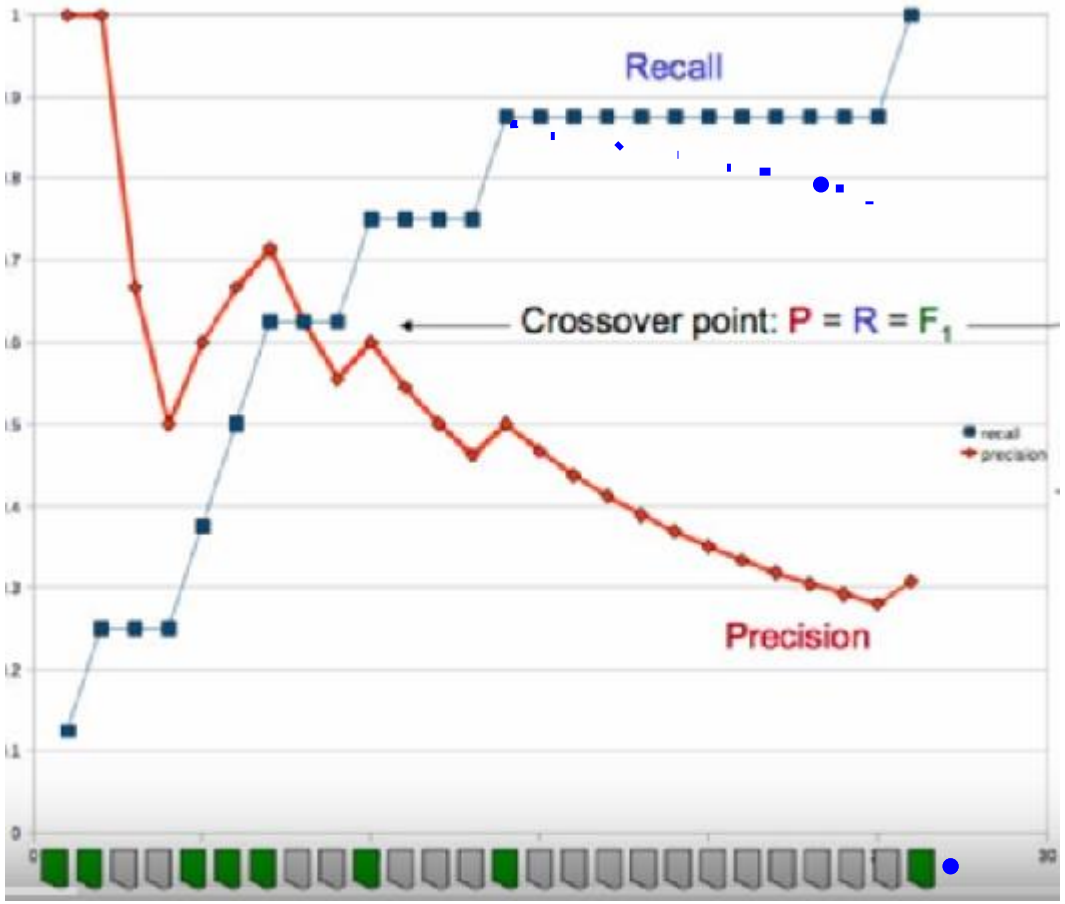
- Which of the following is a better system?
  - system A: recall = 50%, precision = 57%, F_1=53%
  - system B: recall = 100%, precision = 40%, F_1=57%
- Could be the same exact system
  - using different threshold settings
  - R/P, F_1 comparisons often meaningless
  - more informative to compare ranking against ranking

Recall / Precision and ranking

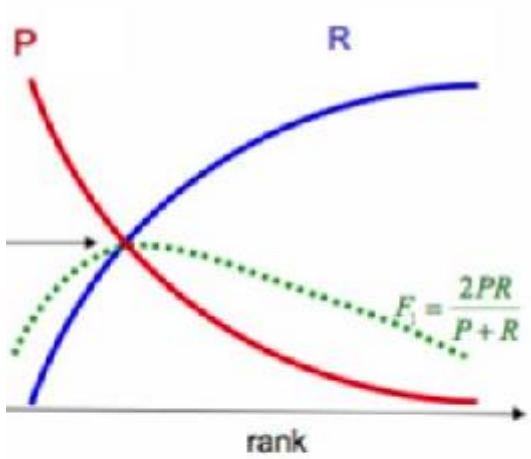
- Search engine produces a ranking, not a set
  - can compute recall, precision at every rank

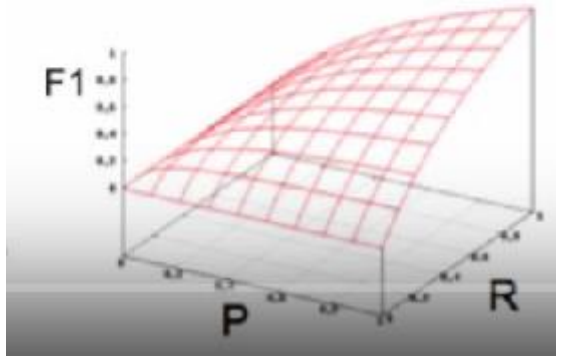


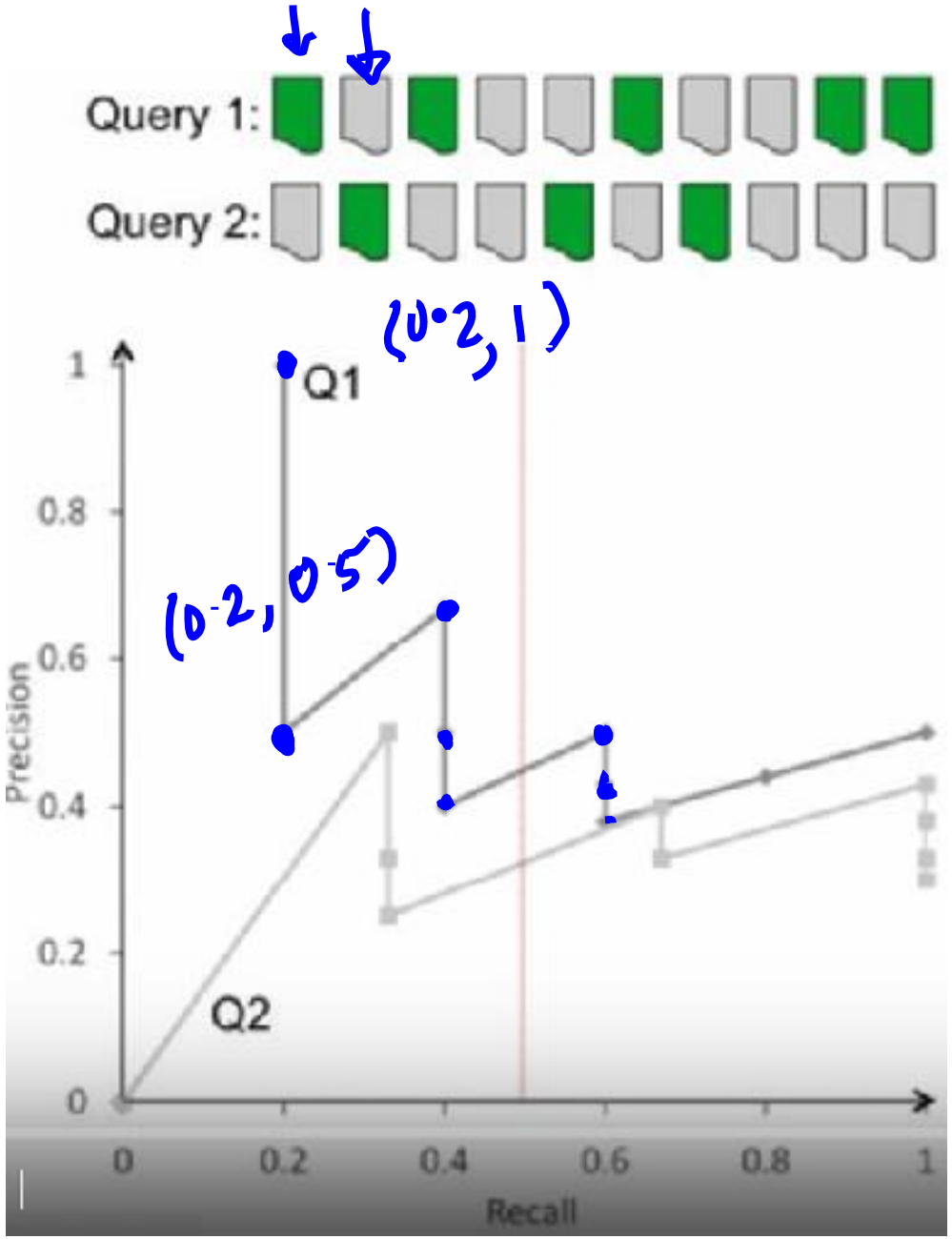
F = P = R

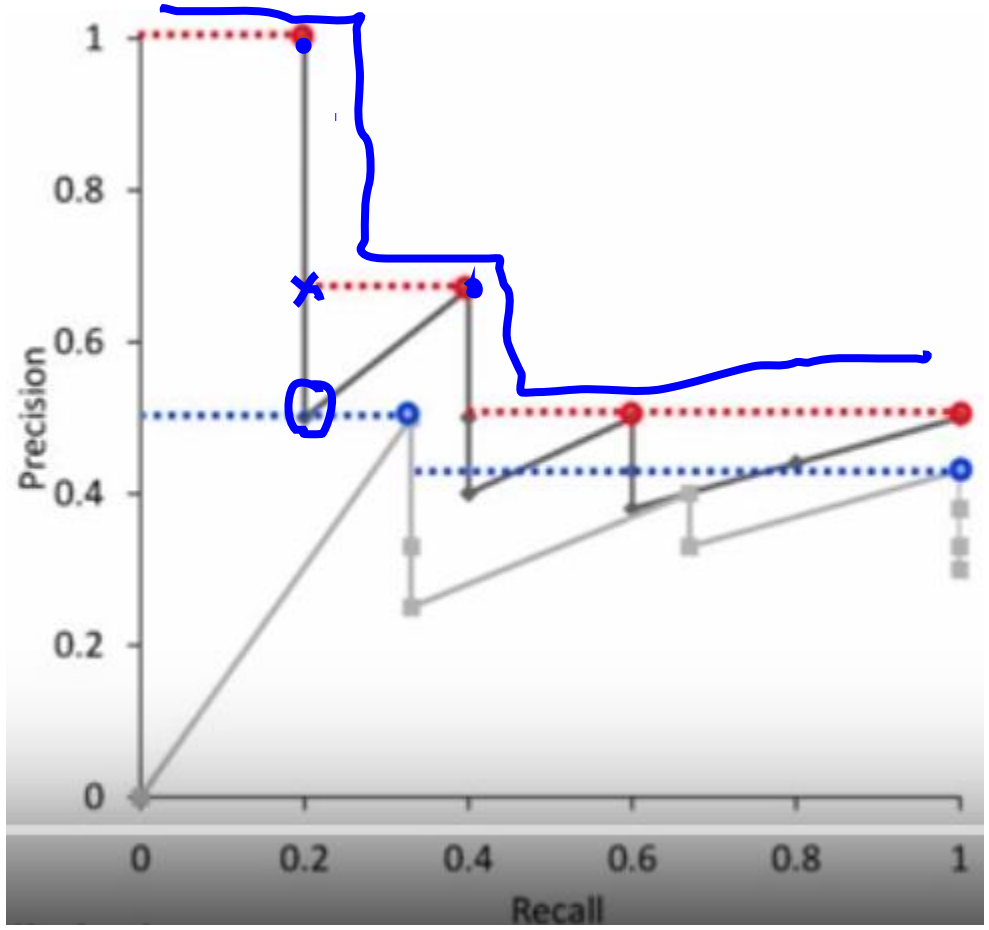

Recall
Crossover point: P = R = F1
Precision
recall
precision

→ Ranking

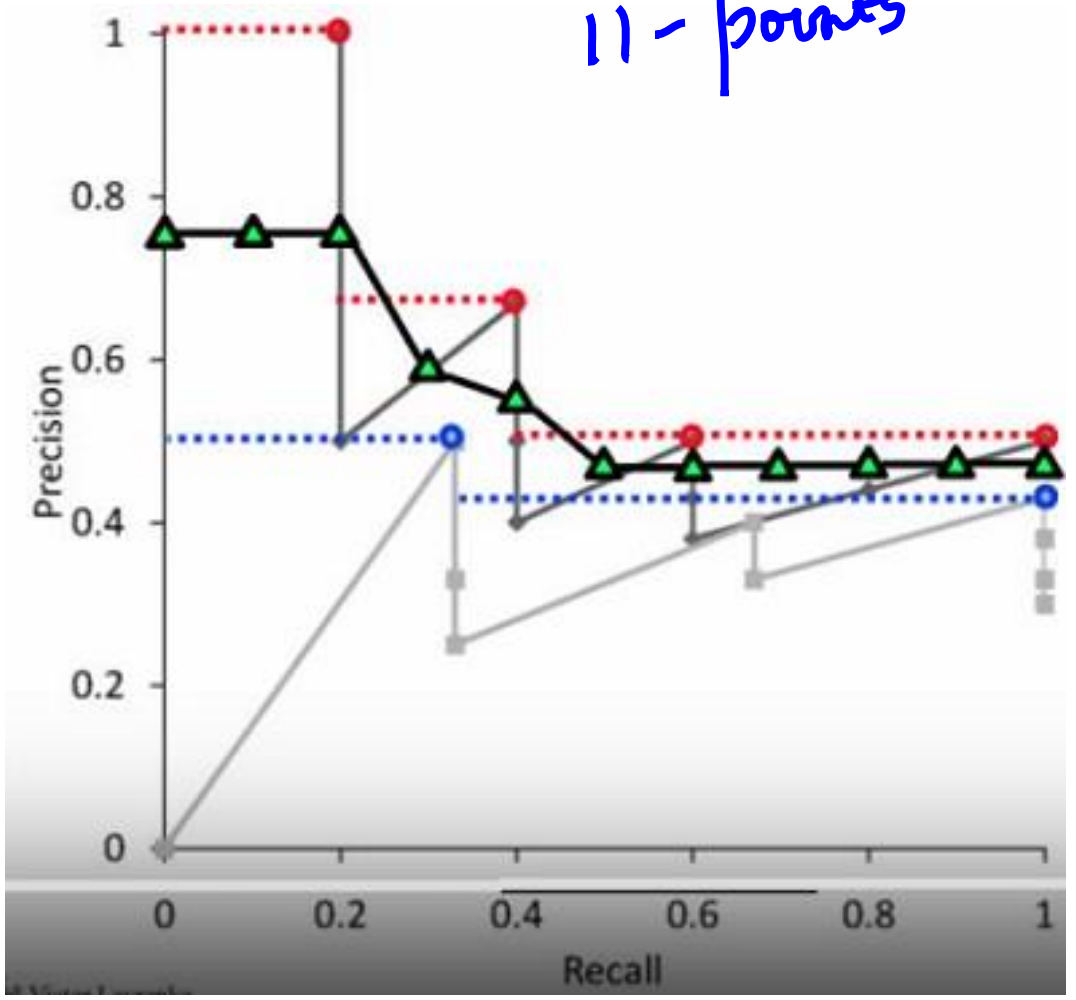

P
R
$F_1 = \frac{2PR}{P+R}$
rank

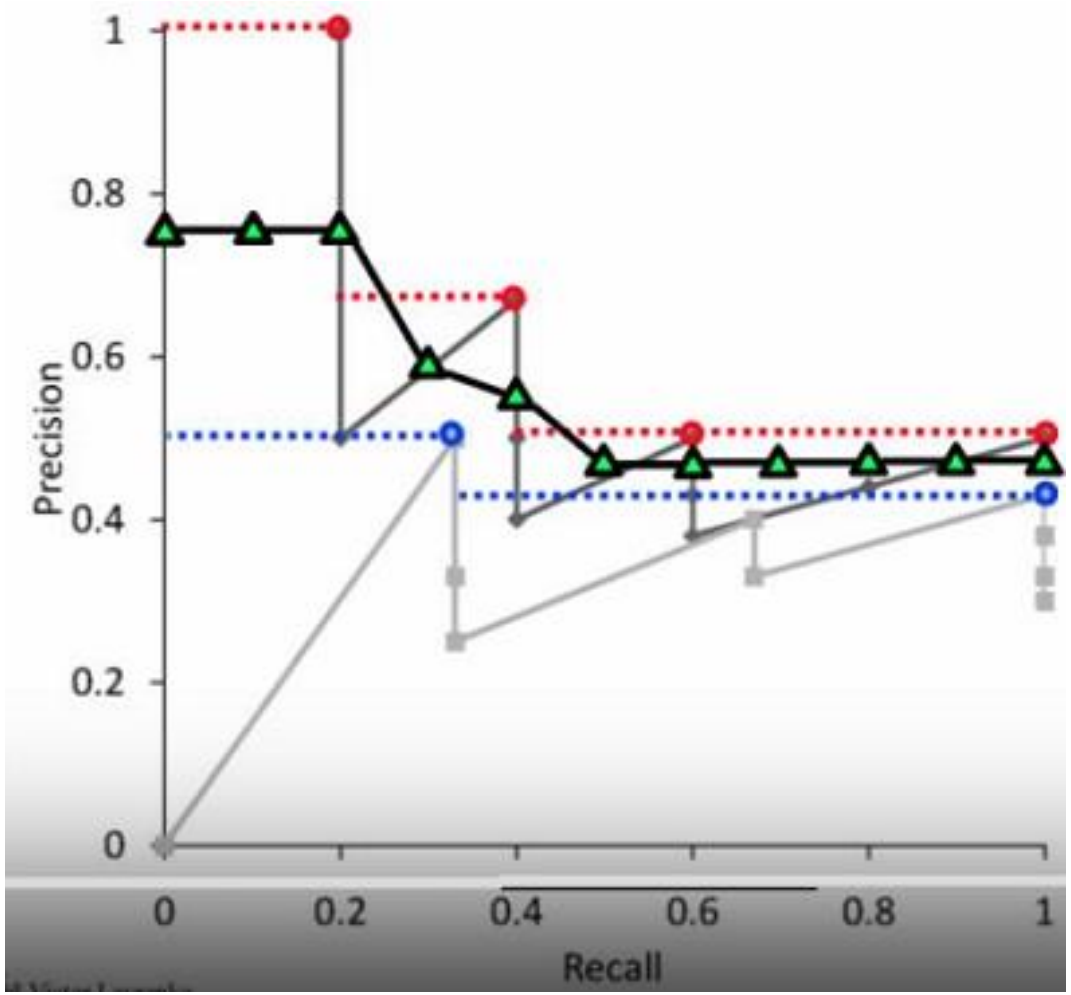

F1
P
R

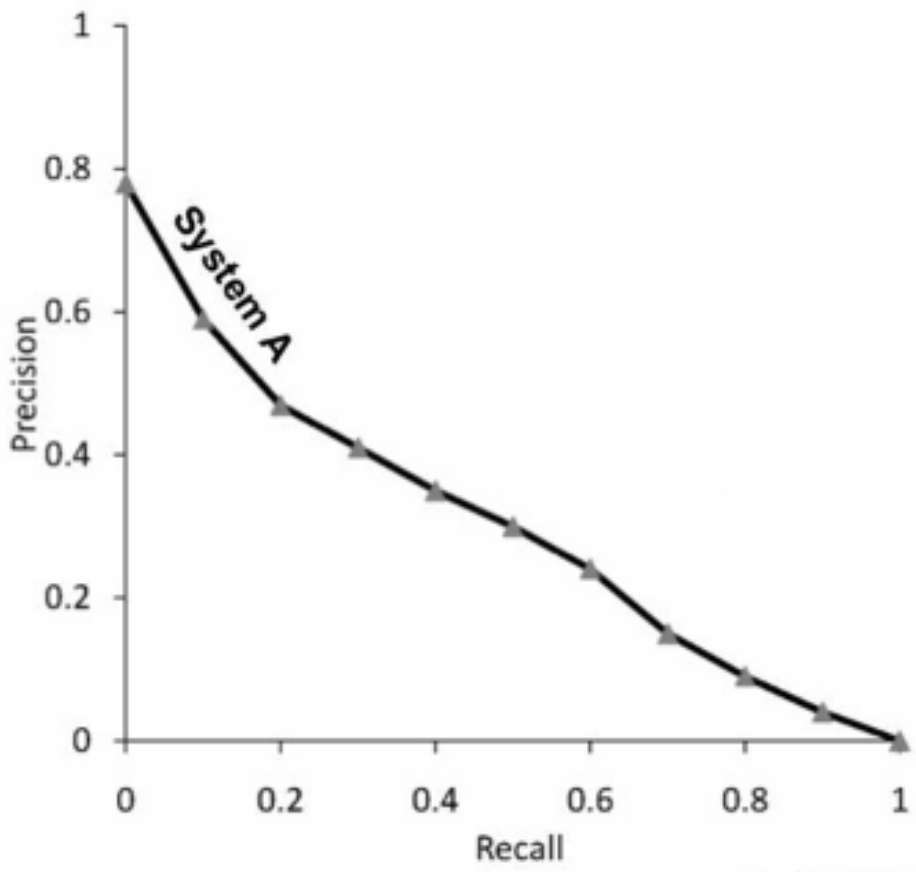

Query 1:
Query 2:
(0.2, 1)
Q1
(0.2, 0.5)
Precision
1
0.8
0.6
0.4
0.2
0
Q2
0
0.2
0.4
0.6
0.8
1
Recall

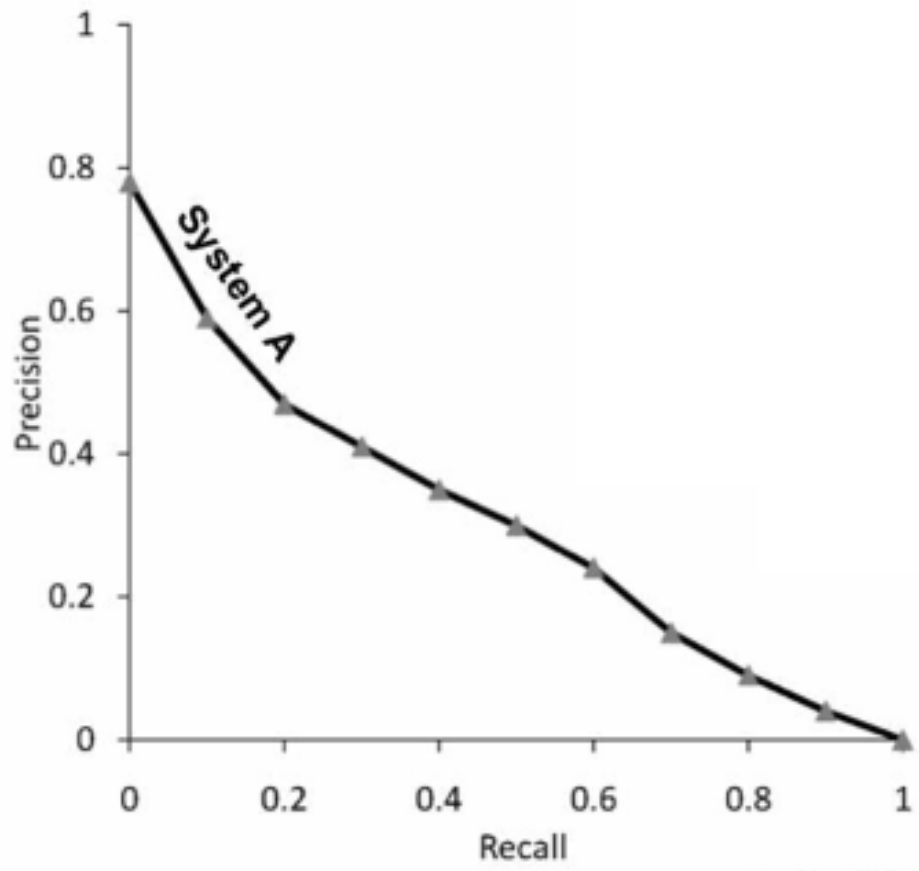

Precision
Recall
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

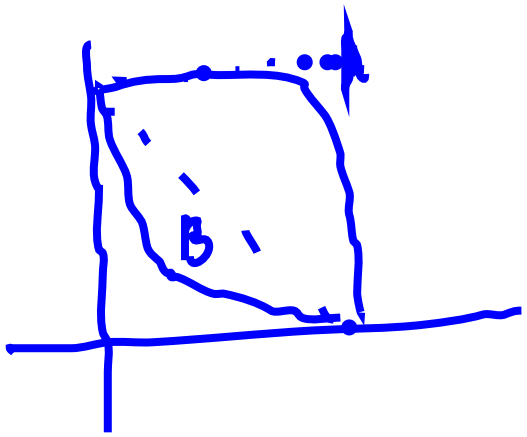
11 - points

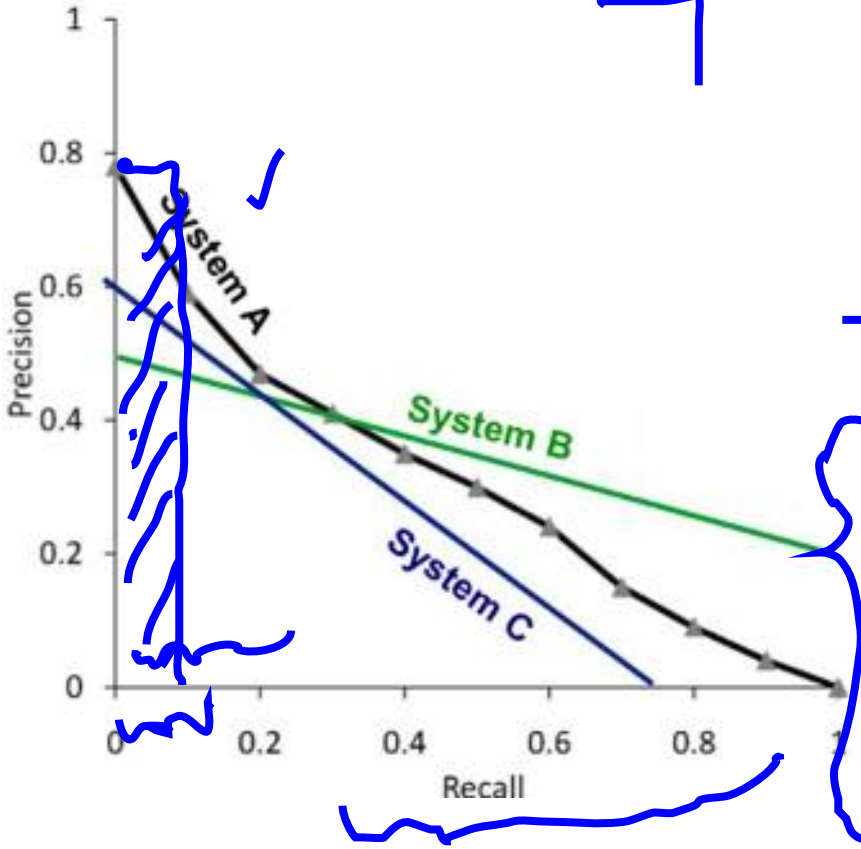

Precision
Recall
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1


1
0.8
0.6
0.4
0.2
0
Precision
0
0.2
0.4
0.6
0.8
1
Recall


1
0.8
0.6
0.4
0.2
0
Precision
System A
0
0.2
0.4
0.6
0.8
1
Recall







→ Mean Reciprocal Rank

→ Mean Avg. Precision

→ nDCG