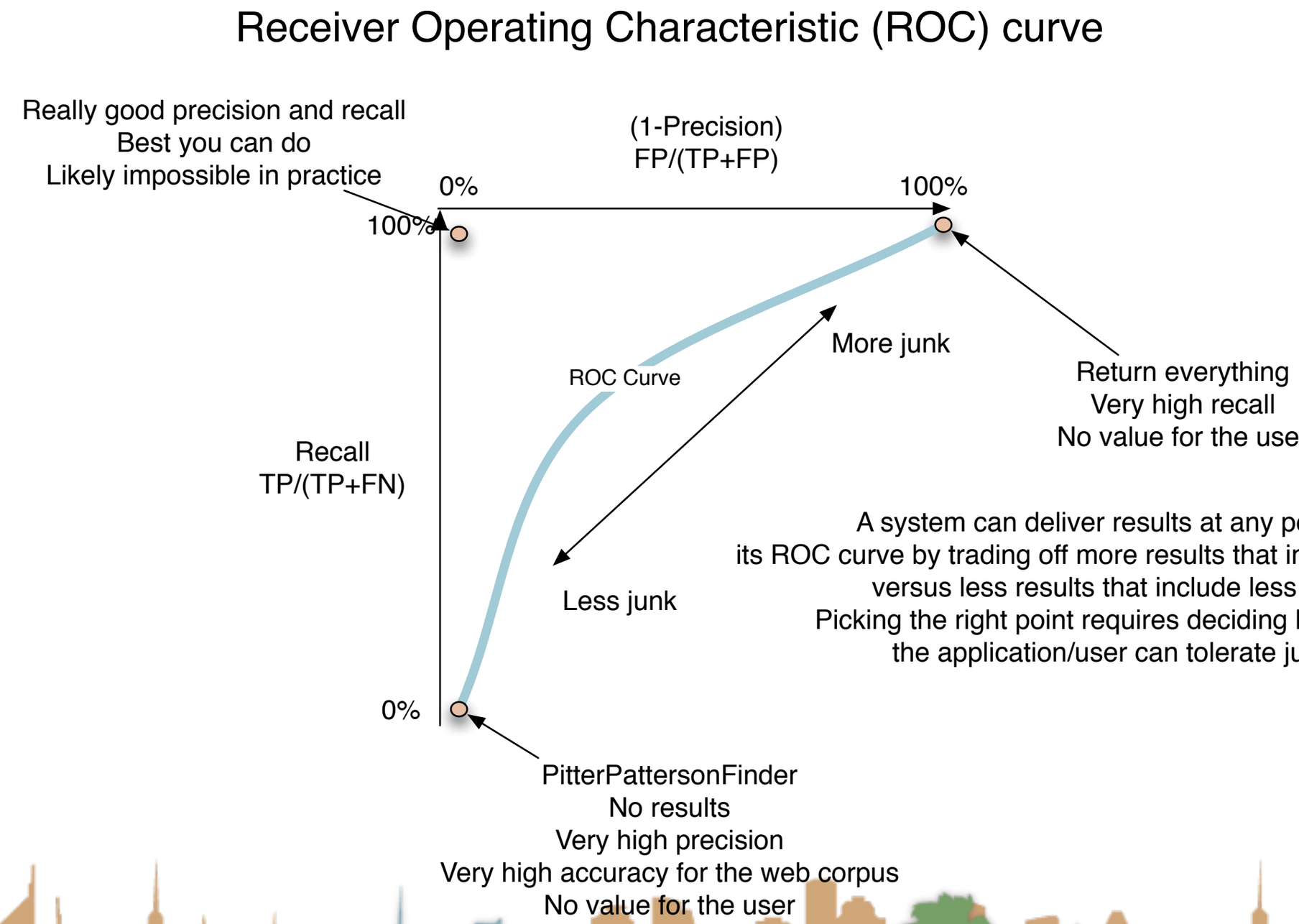
Unranked retrieval - ROC curve



Ranked Retrieval

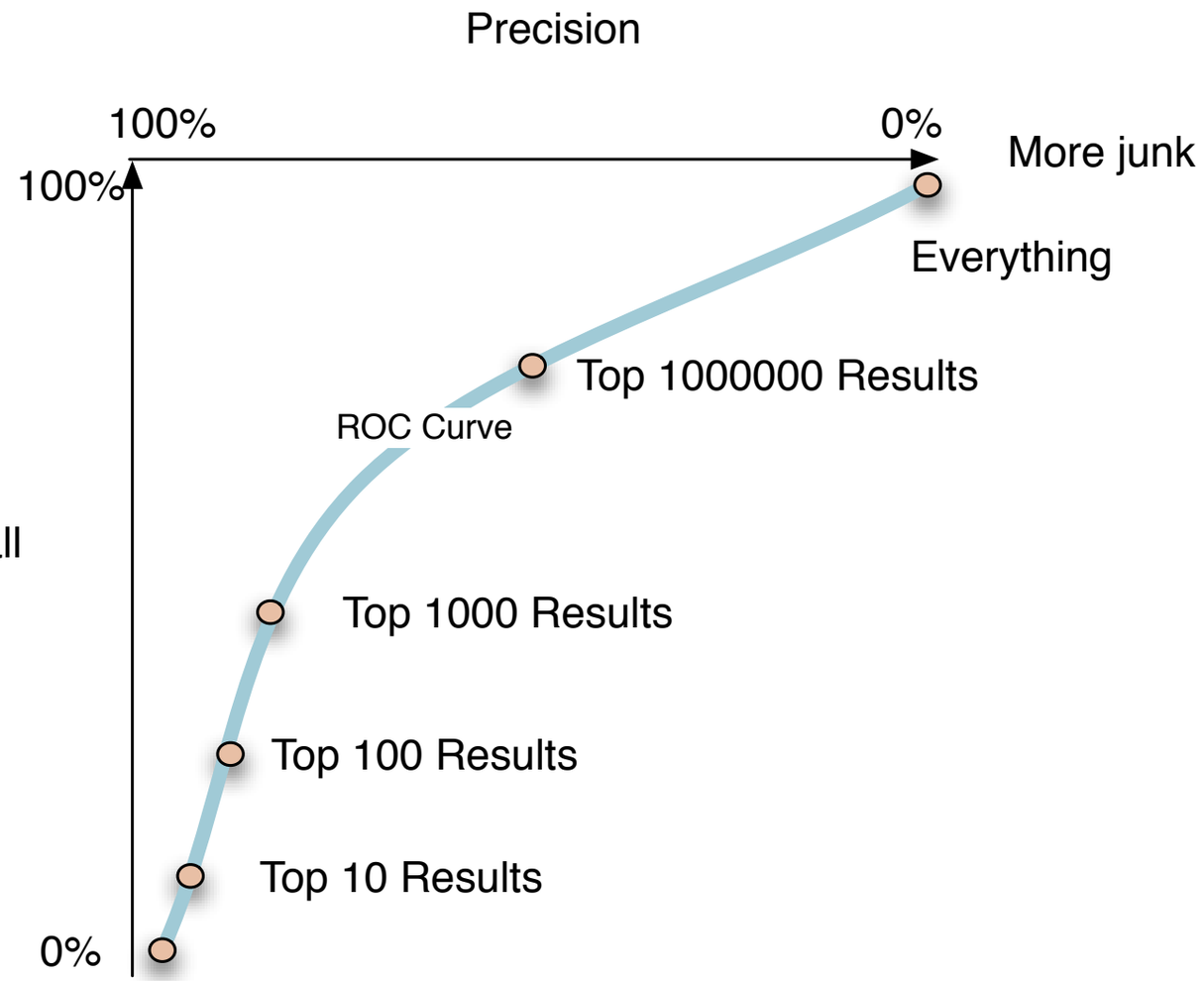
- Precision and Recall are set-based measures
  - They are computed independent of order
  - But, web search return things in lists
  - Lists have order.
  - A better metric of user happiness/relevance is warranted

Ranked Retrieval

- Let's use our existing metrics and extend them to ranked retrieval
  - In one system we can get many samples
  - We can get the top X results:
    - X= 10, 20, 30, 40, etc…
- Each one of those sets has a precision and recall value
- Each of those sets corresponds to a point on the ROC curve.

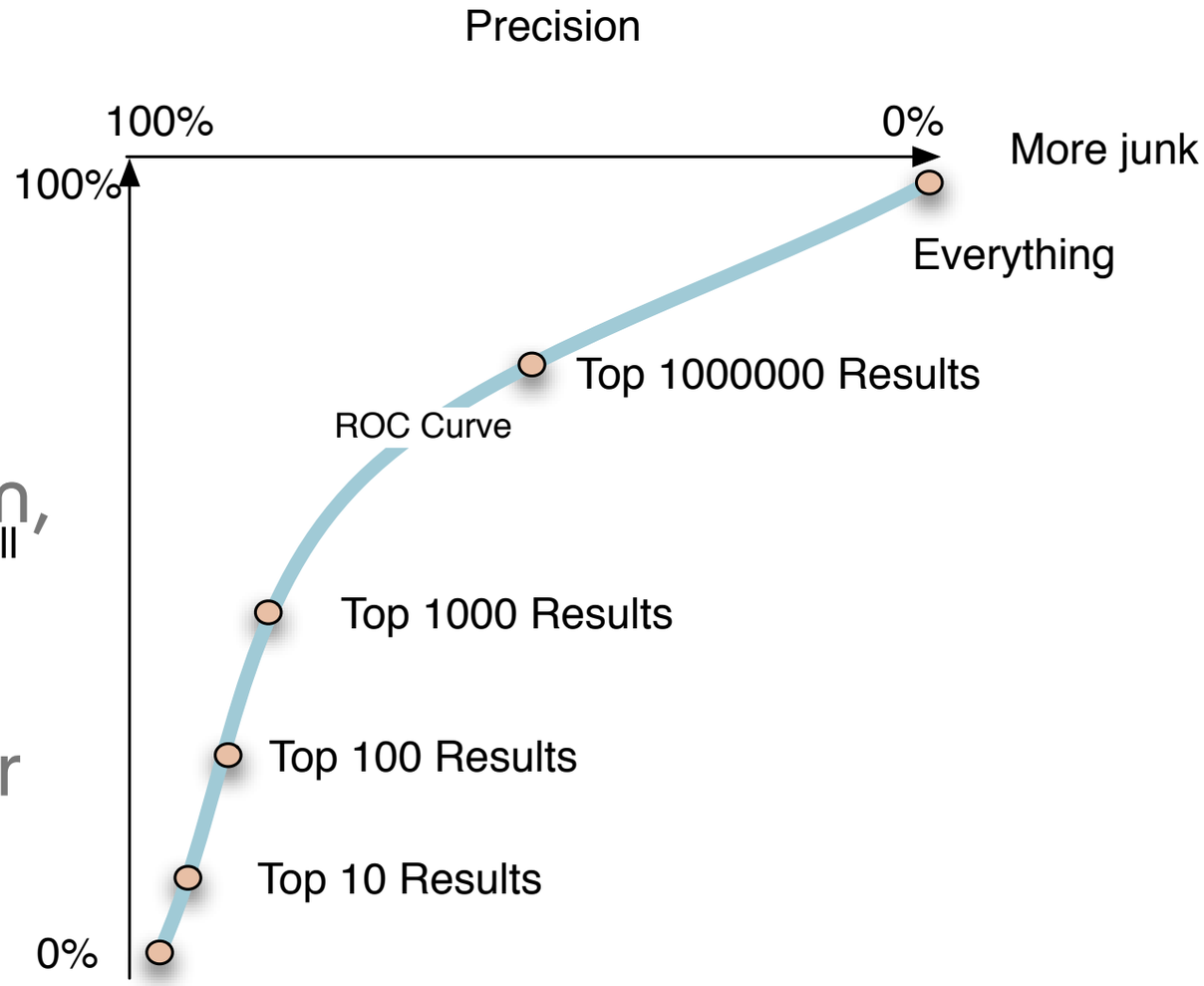
Ranked Retrieval

- Each of those sets corresponds to a point on the ROC curve.



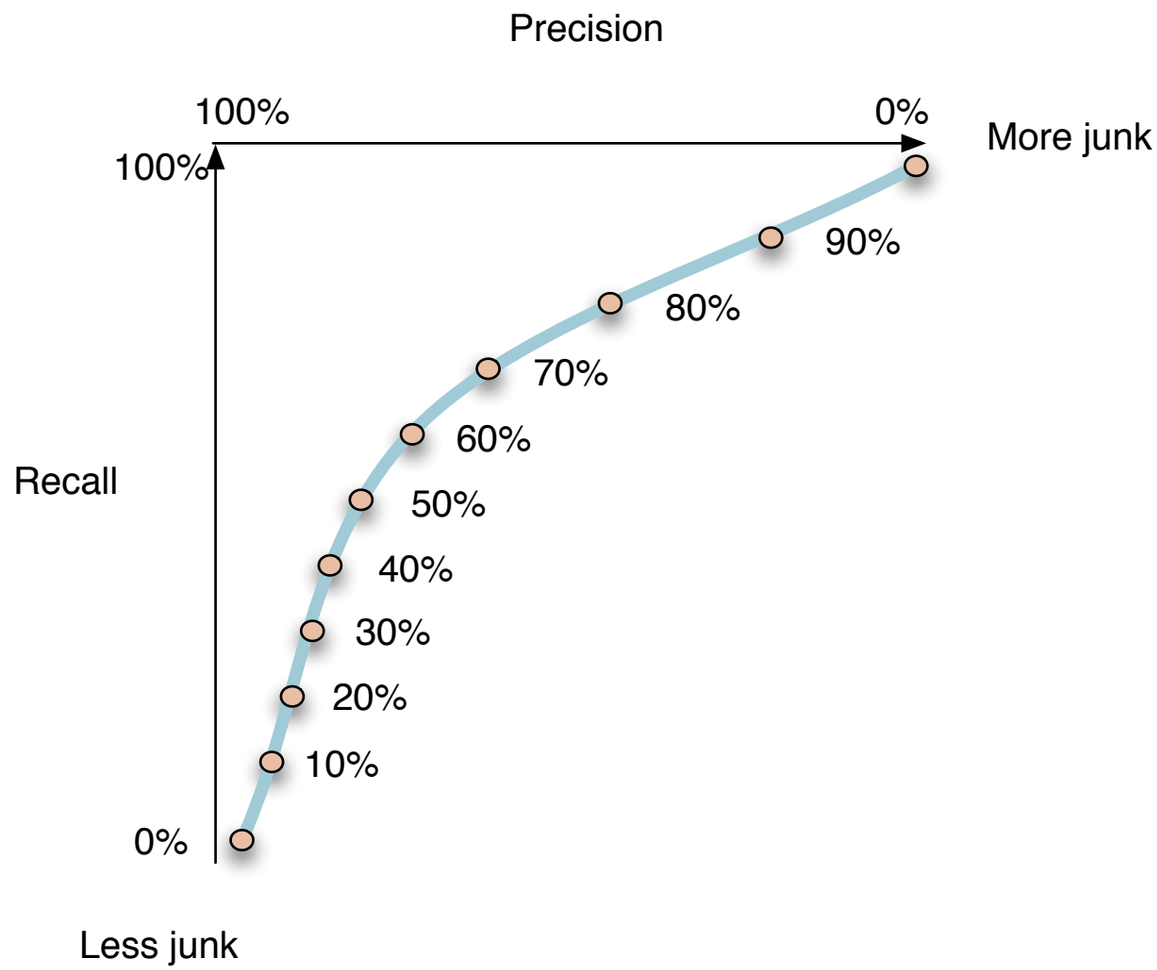
Ranked Retrieval

- One option is to average the precision scores at discrete points on the ROC curve
- But which points?
- We want to evaluate the system, not the corpus
- So it can't be based on number of documents returned



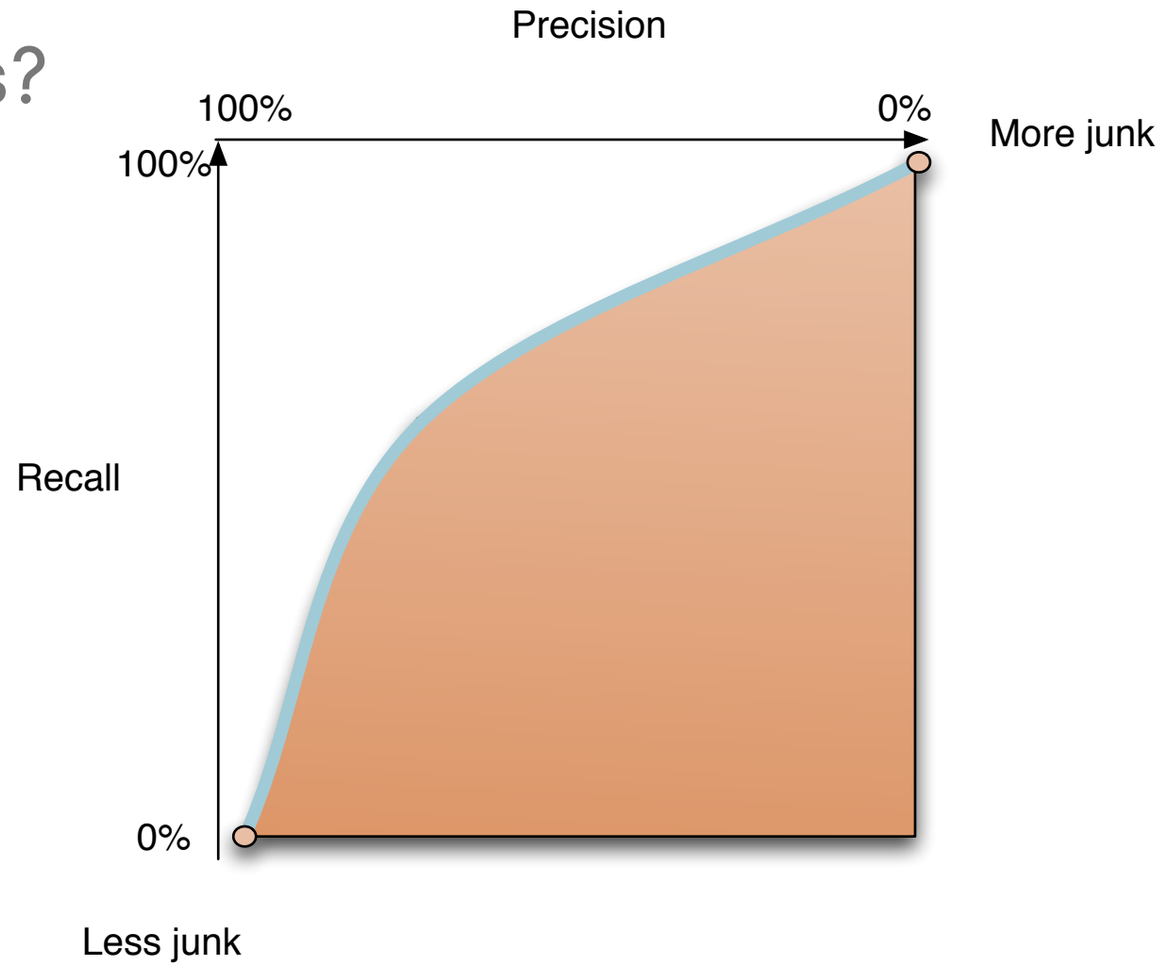
Ranked Retrieval - 11 point precision

- Evaluate based on precision at defined recall points
- Average the precision at 11 points
- This can be compared across corpora
- because it isn't based on corpus size or number of results returned



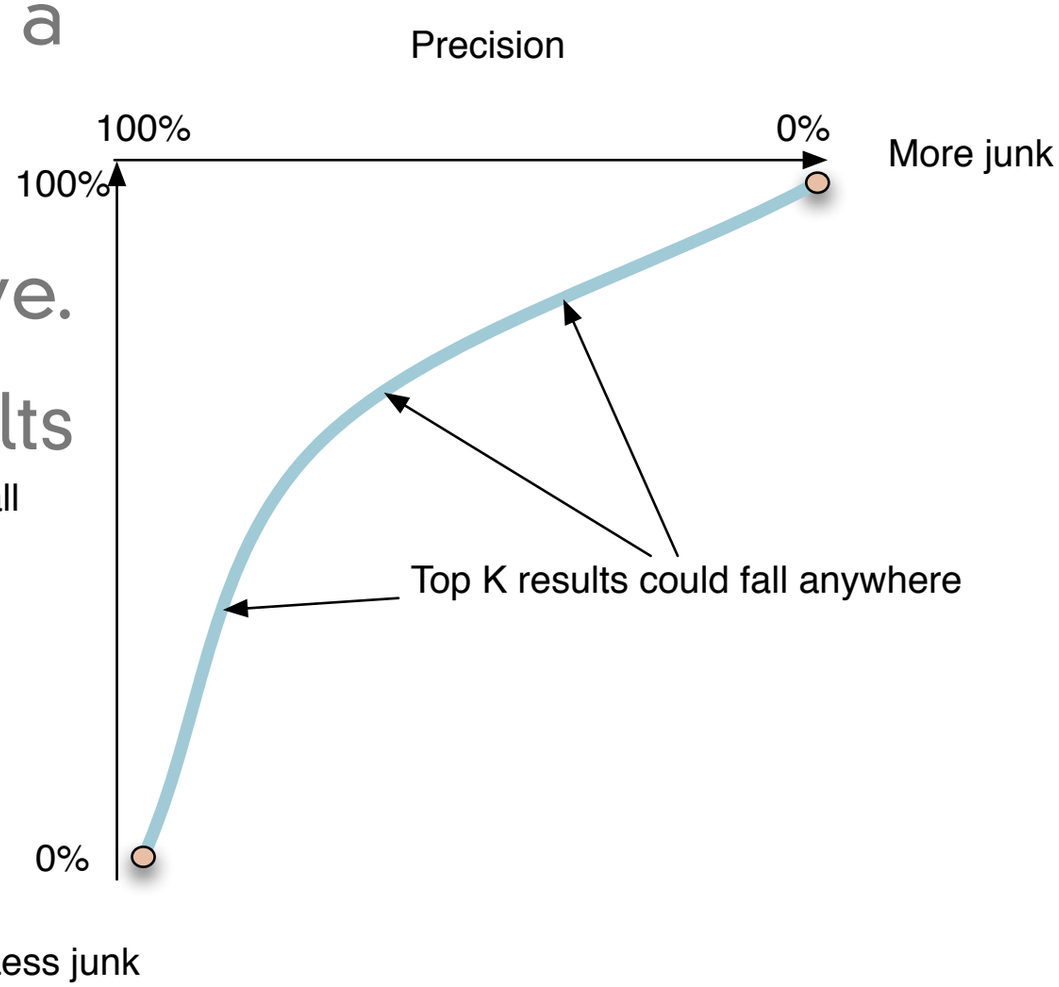
Ranked Retrieval - Mean Average Precision

- Why just 11 points?
- Why not average over all points?
- This is roughly equivalent to measuring the area under the curve.



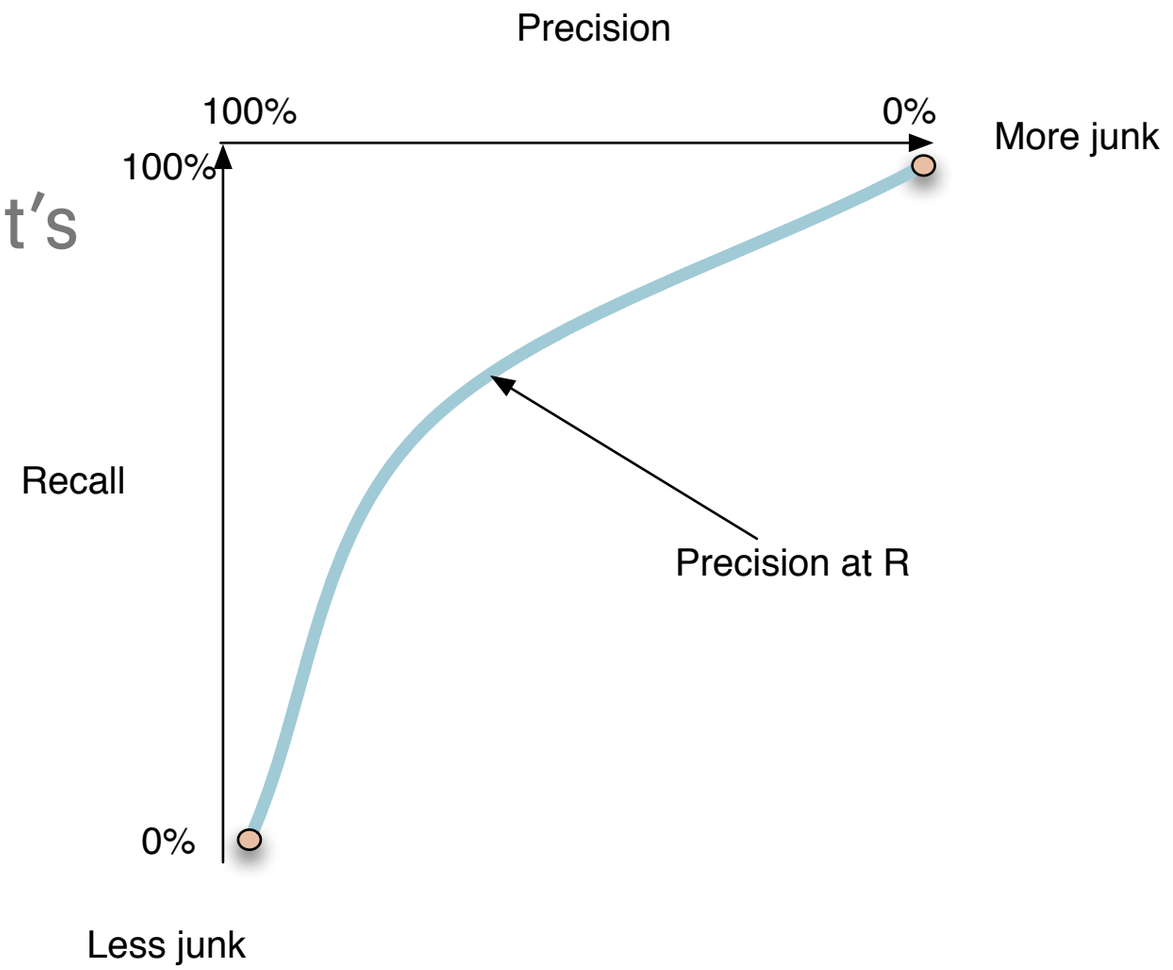
Ranked Retrieval - Precision at k

- Users don't care about results past a page or two
- So area under the curve is too naive.
- Let's evaluate precision with k results instead.
- Highly dependent on number or relevant documents
- If k is 20 and relevant docs is 8
  - best score is 8/(8+12) =0.4



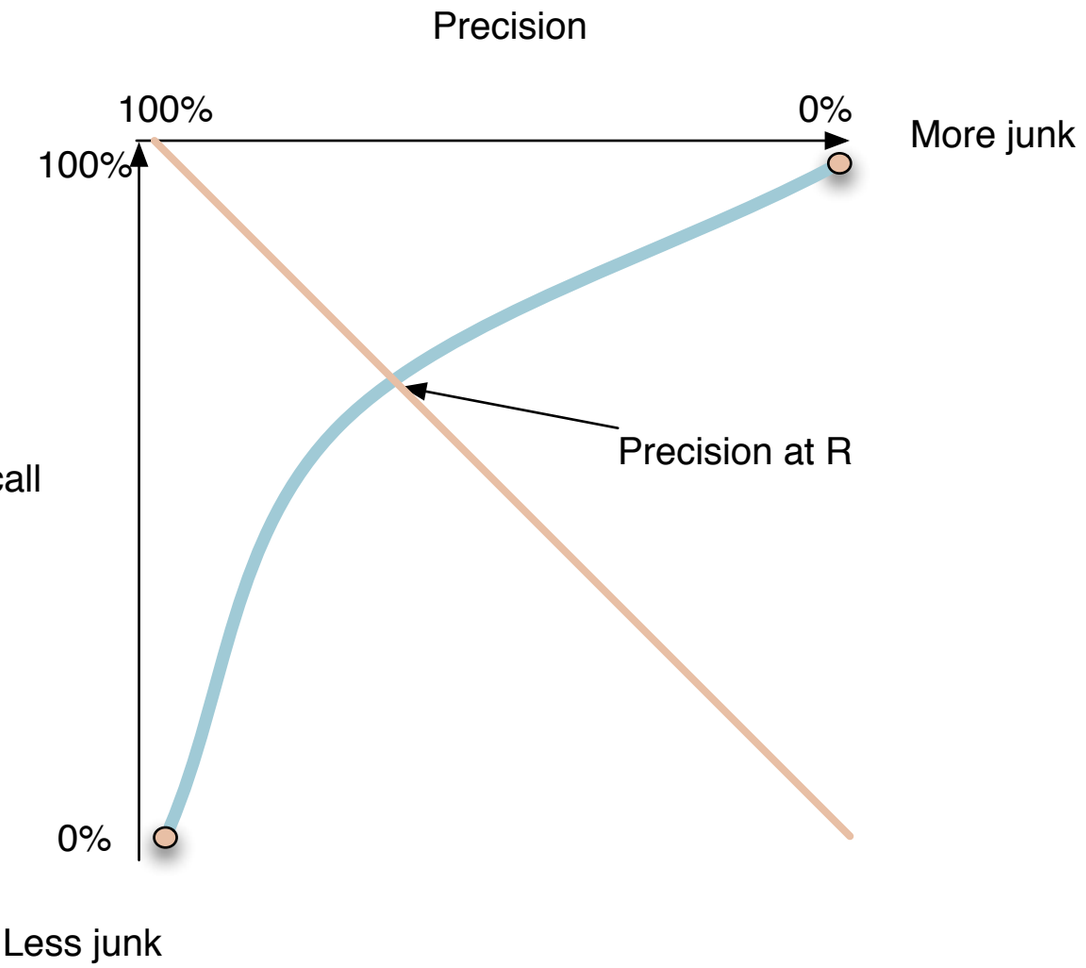
Ranked Retrieval - Precision at R

- We know the number of relevant documents, r, so
- rather than looking at k results let's look at the top r results
- If r is 20
- best score is 20/(20) =1.0
- best score is always 1.0



Ranked Retrieval - Precision at R

- It turns out that Precision at R is the break-even point
  - When Precision and Recall are equal
  - Do we care about this point for any rational reason?



Critiques of relevance

- Is the relevance of one document independent of another?
- Is a gold standard possible?
  - Is a gold standard static?
  - Uniform?
  - Binary?
- Perhaps relevance as a ranking is better.
- Relevance versus marginal relevance
  - what does another document add?