

Evaluation in IR

Introduction to Information Retrieval

INF 141/ CS 121

Donald J. Patterson

Content adapted from Hinrich Schütze

<http://www.informationretrieval.org>



Information need

- Remember the user has an **information need**
 - not a query
- Relevance is assessed in relation to the information need, not the query
 - e.g., I am looking for information on whether drinking red wine is more effective than eating chocolate at reducing risk of heart attacks
 - Query: red wine heart attack effective chocolate risk
 - Does the document address the **need**, not the query



Relevance benchmarks

- TREC - National Institute of Standards and Testing (NIST)
has run a large IR test bed for many years
- Reuters and other benchmark document collections
- Retrieval tasks which are specified
 - sometimes as queries
- Human experts mark, for each query and for each document
 - Relevant or Irrelevant



Unranked retrieval

- Precision:
 - Fraction of retrieved documents that are relevant
- Recall:
 - Fraction of relevant documents that are retrieved



Unranked retrieval

- Precision:
 - Fraction of retrieved documents that are relevant
- Recall:
 - Fraction of relevant documents that are retrieved

	<i>Relevant</i>	<i>Not Relevant</i>
<i>Retrieved</i>	<i>TP</i>	<i>FP</i>
<i>Not Retrieved</i>	<i>FN</i>	<i>TN</i>



Unranked retrieval

- Precision:
 - Fraction of retrieved documents that are relevant
- Recall:
 - Fraction of relevant documents that are retrieved

	<i>Relevant</i>	<i>Not Relevant</i>
<i>Retrieved</i>	<i>TP</i>	<i>FP</i>
<i>Not Retrieved</i>	<i>FN</i>	<i>TN</i>

$$? \text{ Precision} = \frac{TP}{TP + FP}$$

$$? \text{ Recall} = \frac{TP}{TP + FN}$$



Unranked retrieval - Accuracy

- The difficulty with measuring “accuracy”
- In one sense accuracy is how many judgments you make correctly

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

	<i>Relevant</i>	<i>Not Relevant</i>
<i>Retrieved</i>	<i>TP</i>	<i>FP</i>
<i>Not Retrieved</i>	<i>FN</i>	<i>TN</i>



Exercise

- Documents A - F, Query q

<i>Document</i>	<i>Relevant(q)</i>	<i>Not Relevant(q)</i>
<i>A</i>	✓	
<i>B</i>		✓
<i>C</i>		✓
<i>D</i>	✓	
<i>E</i>		✓
<i>F</i>	✓	

- If my system returns A,C,D,E to query q
- How many TP, TN, FP, FN do I have?

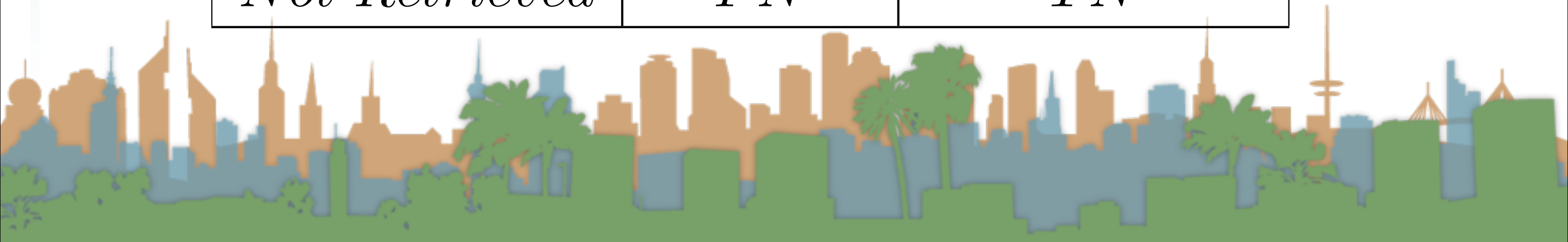


Exercise

Retrieved : A C D E

<i>Document</i>	<i>Relevant(q)</i>	<i>Not Relevant(q)</i>
<i>A</i>	✓	
<i>B</i>		✓
<i>C</i>		✓
<i>D</i>	✓	
<i>E</i>		✓
<i>F</i>	✓	

	<i>Relevant</i>	<i>Not Relevant</i>
<i>Retrieved</i>	<i>TP</i>	<i>FP</i>
<i>Not Retrieved</i>	<i>FN</i>	<i>TN</i>



Exercise

Retrieved : A C D E

<i>Document</i>	<i>Relevant(q)</i>	<i>Not Relevant(q)</i>
<i>A</i>	✓	
<i>B</i>		✓
<i>C</i>		✓
<i>D</i>	✓	
<i>E</i>		✓
<i>F</i>	✓	

	<i>Relevant</i>	<i>Not Relevant</i>
<i>Retrieved</i>	<i>TP</i>	<i>FP</i>
<i>Not Retrieved</i>	<i>FN</i>	<i>TN</i>



Exercise

- What is our precision?

$$Precision = \frac{TP}{TP + FP}$$

TP	2
FP	2
FN	1
TN	1

- What is our recall?

$$Recall = \frac{TP}{TP + FN}$$

- What is our accuracy?

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$



Exercise

- If my system returns A,C,D,E to query q....

<i>Document</i>	<i>Relevant(q)</i>	<i>Not Relevant(q)</i>
<i>A</i>	✓	
<i>B</i>		✓
<i>C</i>		✓
<i>D</i>	✓	
<i>E</i>		✓
<i>F</i>	✓	

Precision

$\frac{1}{2}$

Recall

$\frac{2}{3}$

Accuracy

$\frac{1}{2}$

- What do I want Precision to be?

	<i>Relevant</i>	<i>Not Relevant</i>
<i>Retrieved</i>	<i>TP</i>	<i>FP</i>
<i>Not Retrieved</i>	<i>FN</i>	<i>TN</i>

$$Precision = \frac{TP}{TP + FP}$$



Exercise

- If my system returns A,C,D,E to query q....

<i>Document</i>	<i>Relevant(q)</i>	<i>Not Relevant(q)</i>
<i>A</i>	✓	
<i>B</i>		✓
<i>C</i>		✓
<i>D</i>	✓	
<i>E</i>		✓
<i>F</i>	✓	

Precision

$\frac{1}{2}$

Recall

$\frac{2}{3}$

Accuracy

$\frac{1}{2}$

- What do I want Recall to be?

	<i>Relevant</i>	<i>Not Relevant</i>
<i>Retrieved</i>	<i>TP</i>	<i>FP</i>
<i>Not Retrieved</i>	<i>FN</i>	<i>TN</i>

$$\text{Recall} = \frac{TP}{TP + FN}$$



Exercise

- If my system returns A,C,D,E to query q....

<i>Document</i>	<i>Relevant(q)</i>	<i>Not Relevant(q)</i>
<i>A</i>	✓	
<i>B</i>		✓
<i>C</i>		✓
<i>D</i>	✓	
<i>E</i>		✓
<i>F</i>	✓	

Precision

$\frac{1}{2}$

Recall

$\frac{2}{3}$

Accuracy

$\frac{1}{2}$

- What do I want Accuracy to be?

	<i>Relevant</i>	<i>Not Relevant</i>
<i>Retrieved</i>	<i>TP</i>	<i>FP</i>
<i>Not Retrieved</i>	<i>FN</i>	<i>TN</i>

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$



Unranked retrieval - Accuracy

- Welcome to my search engine
- I guarantee a 99.9999% accuracy.
- Bring on the venture capital

Beta

PITTERPATTERSONFINDER

Search for:



Unranked retrieval - Accuracy

- Most people **want to find something** and can tolerate some junk

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

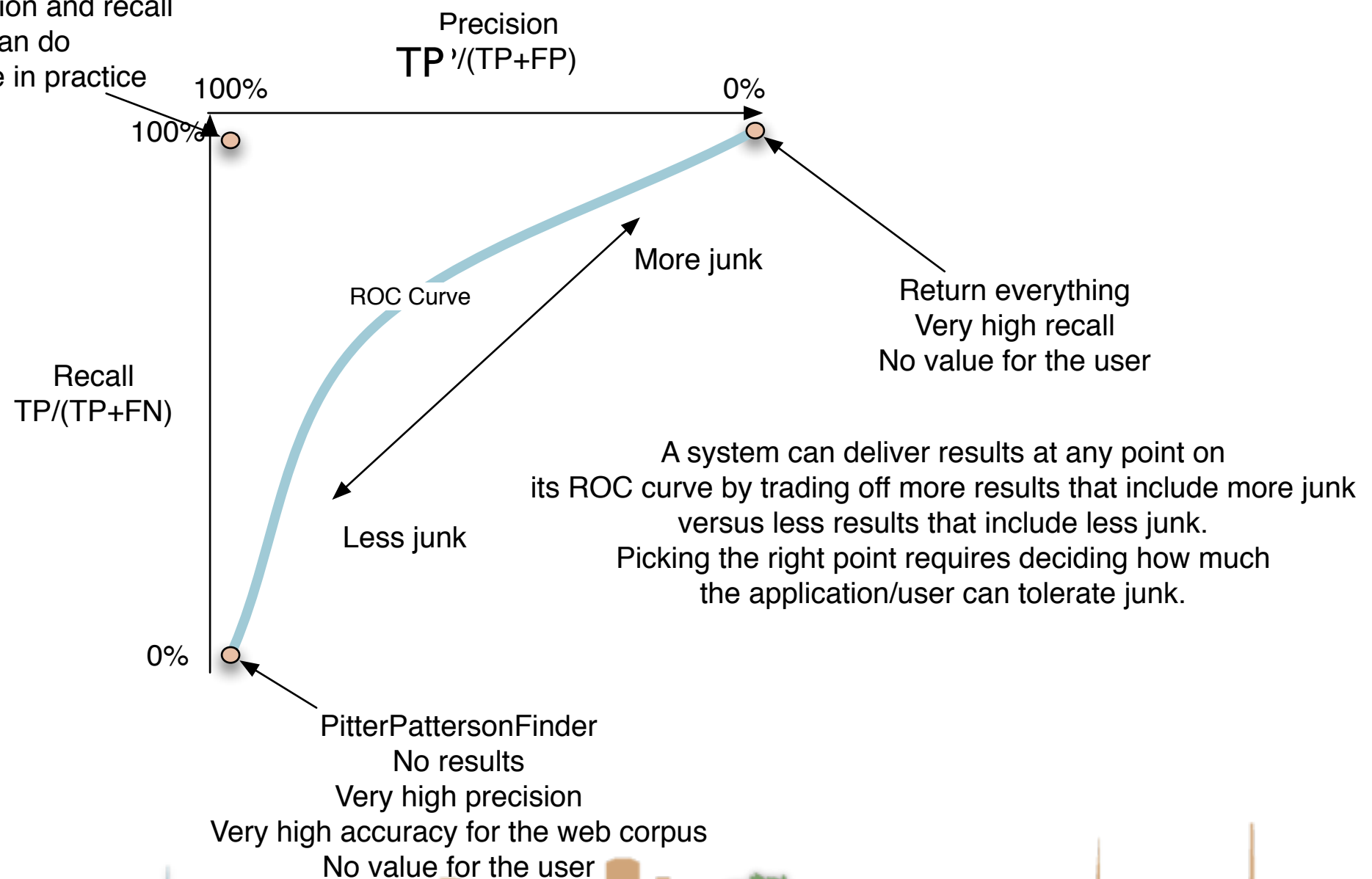
$$Accuracy = \frac{0 + \uparrow}{0 + 0 + \epsilon + \uparrow}$$



Unranked retrieval - ROC curve

Receiver Operating Characteristic (ROC) curve

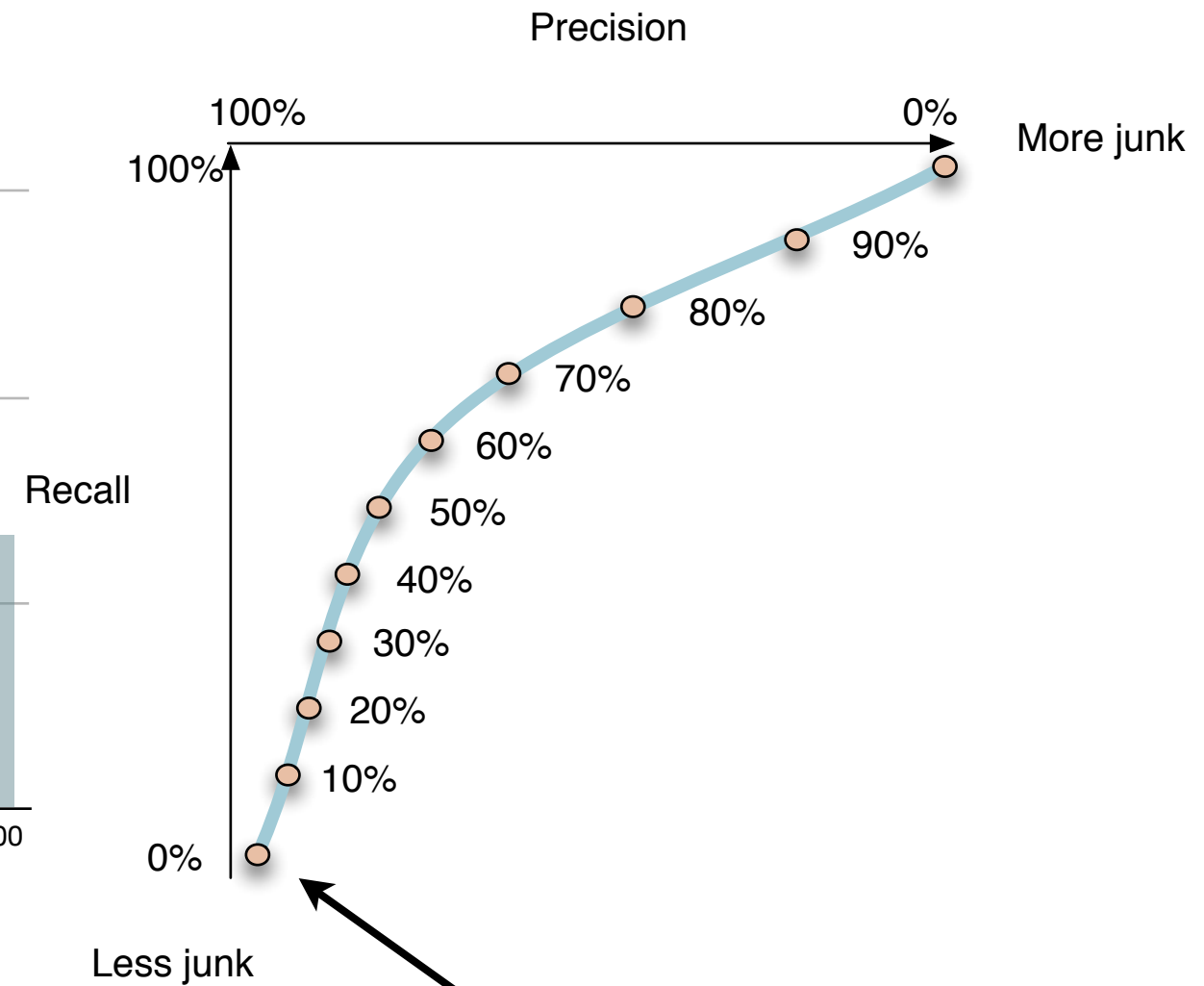
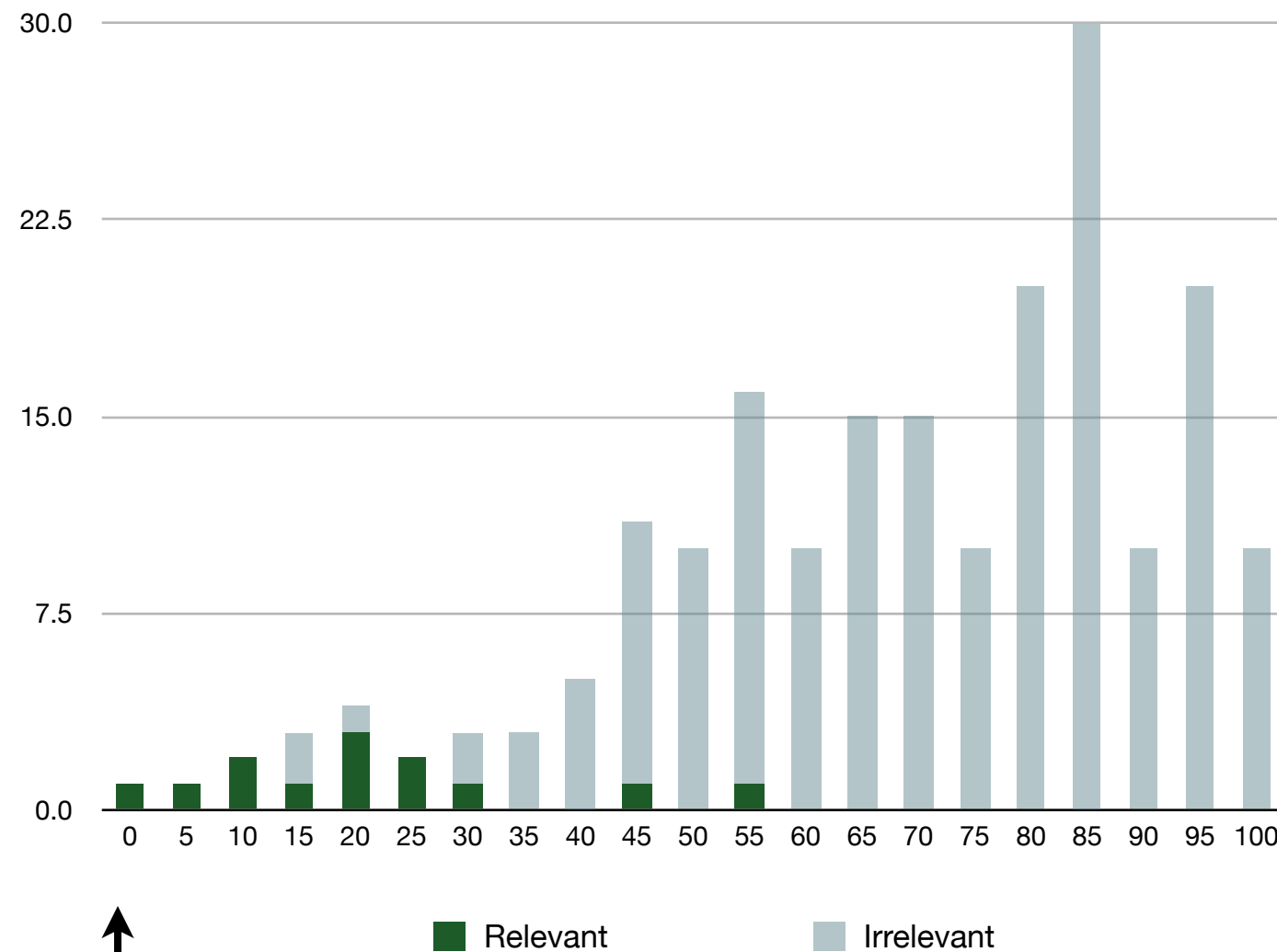
Really good precision and recall
Best you can do
Likely impossible in practice



Unranked retrieval - ROC curve

Receiver Operating Characteristic (ROC) curve

Example Histogram of Documents versus relevance score



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



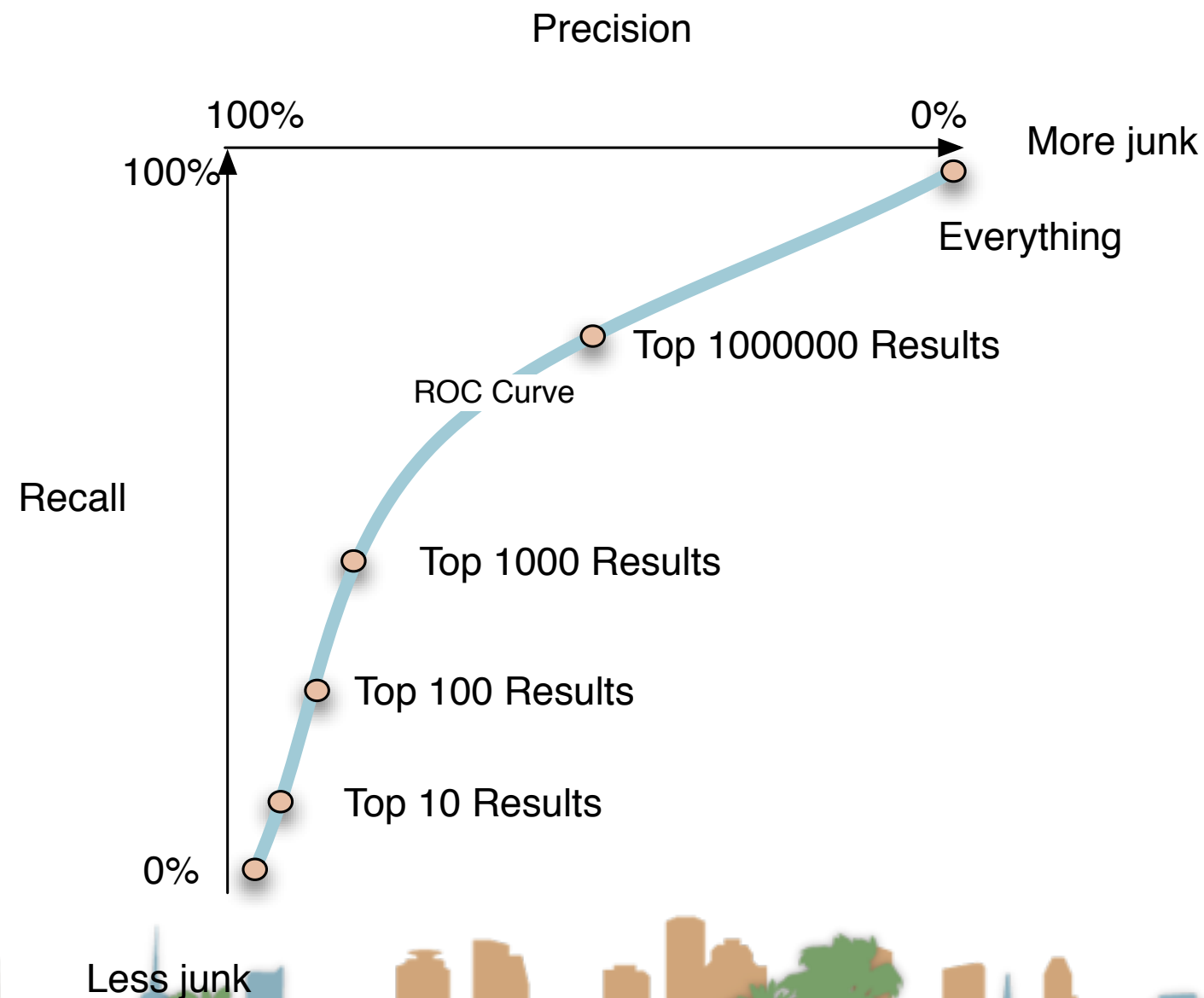
Ranked Retrieval

- 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.



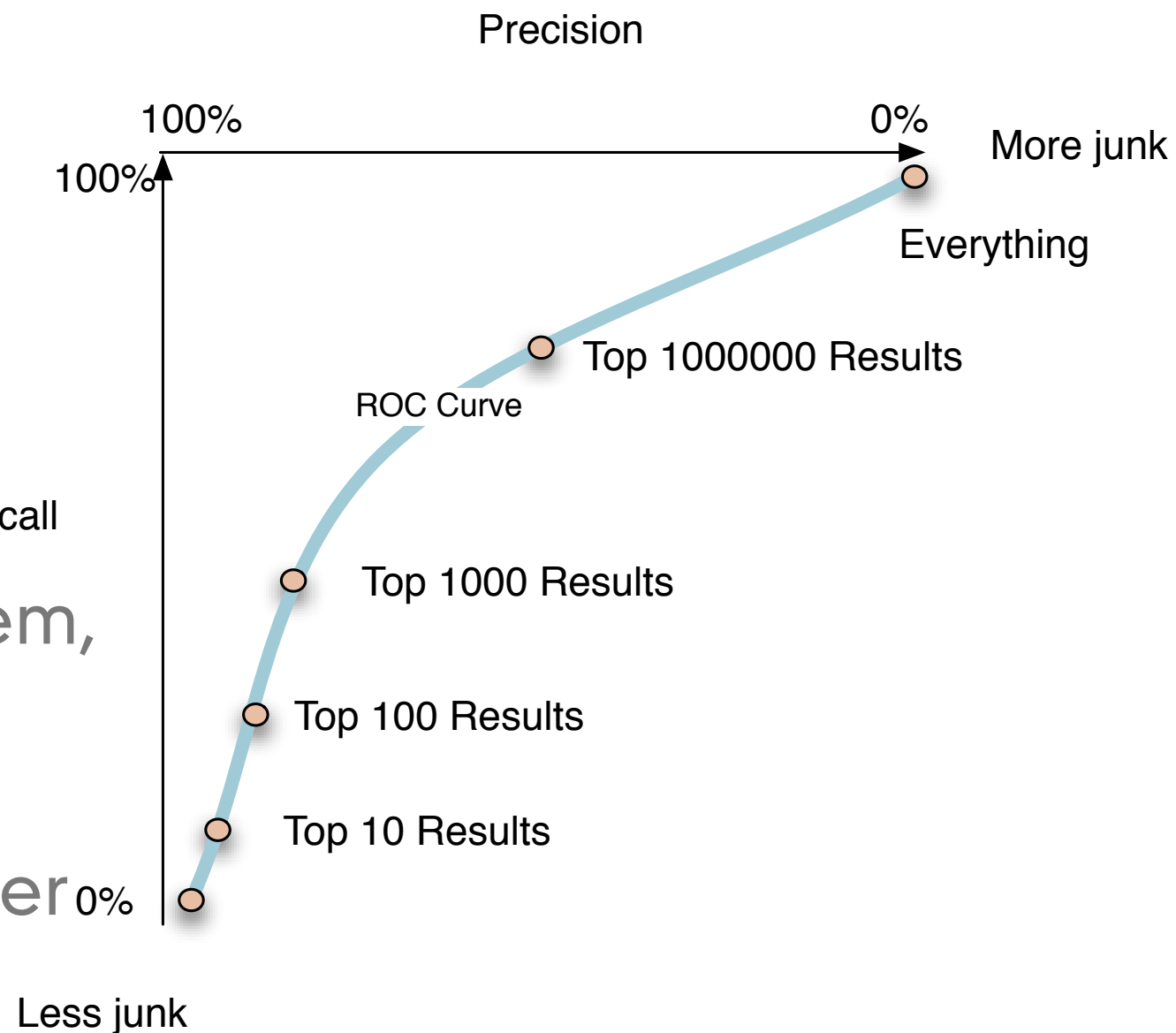
Ranked Retrieval

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



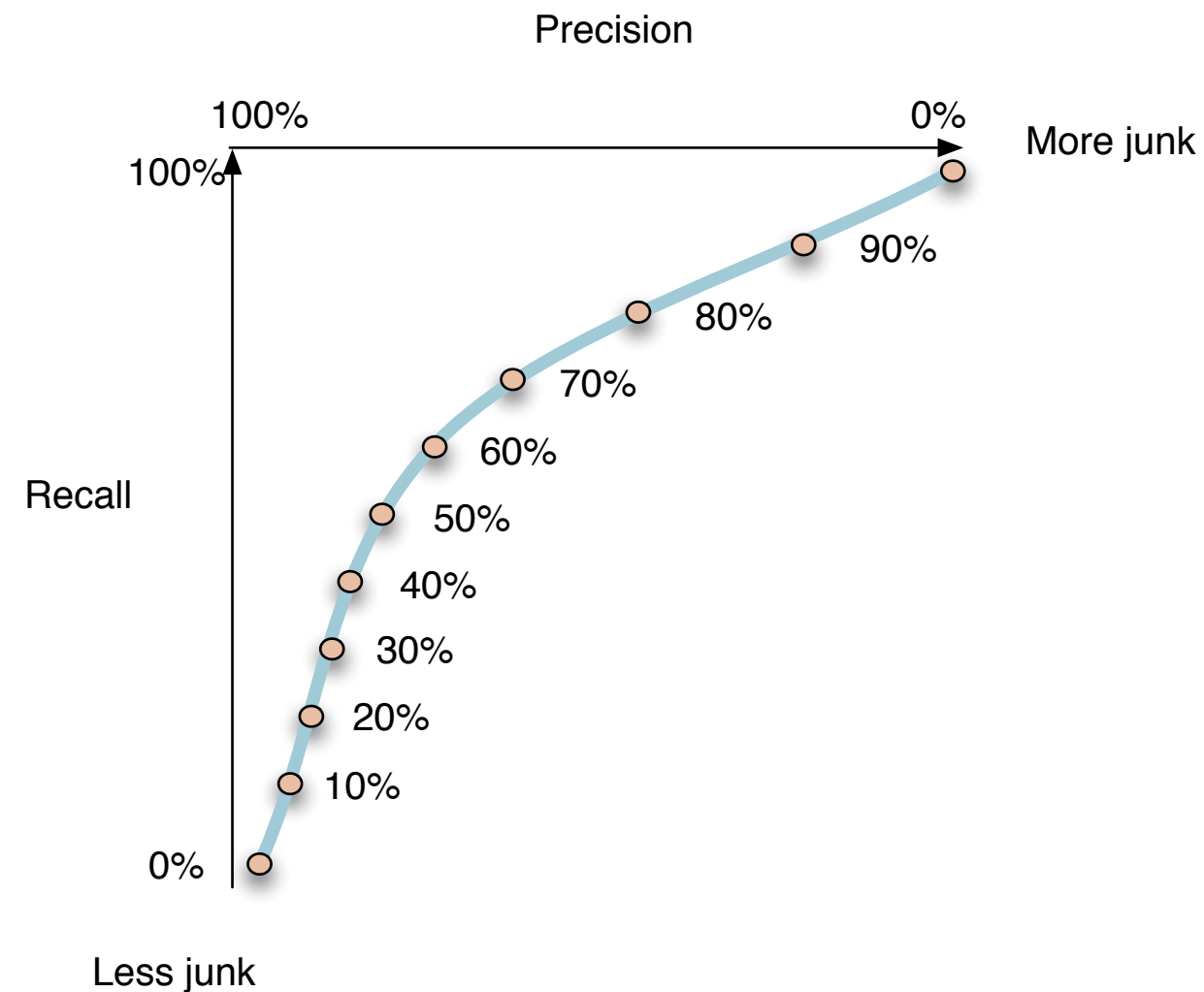
Ranked Retrieval

- 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



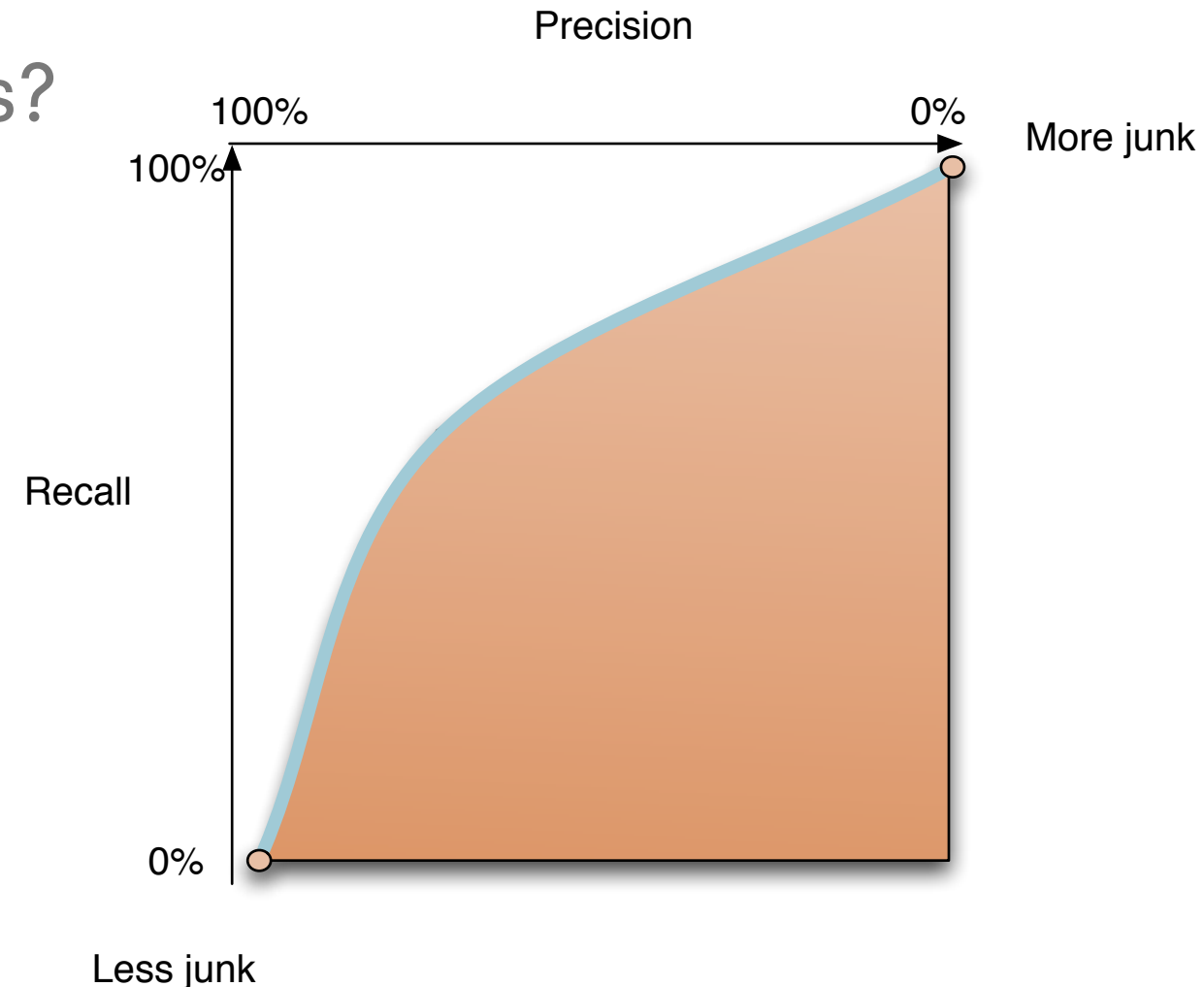
Ranked Retrieval - 11 point precision

- 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



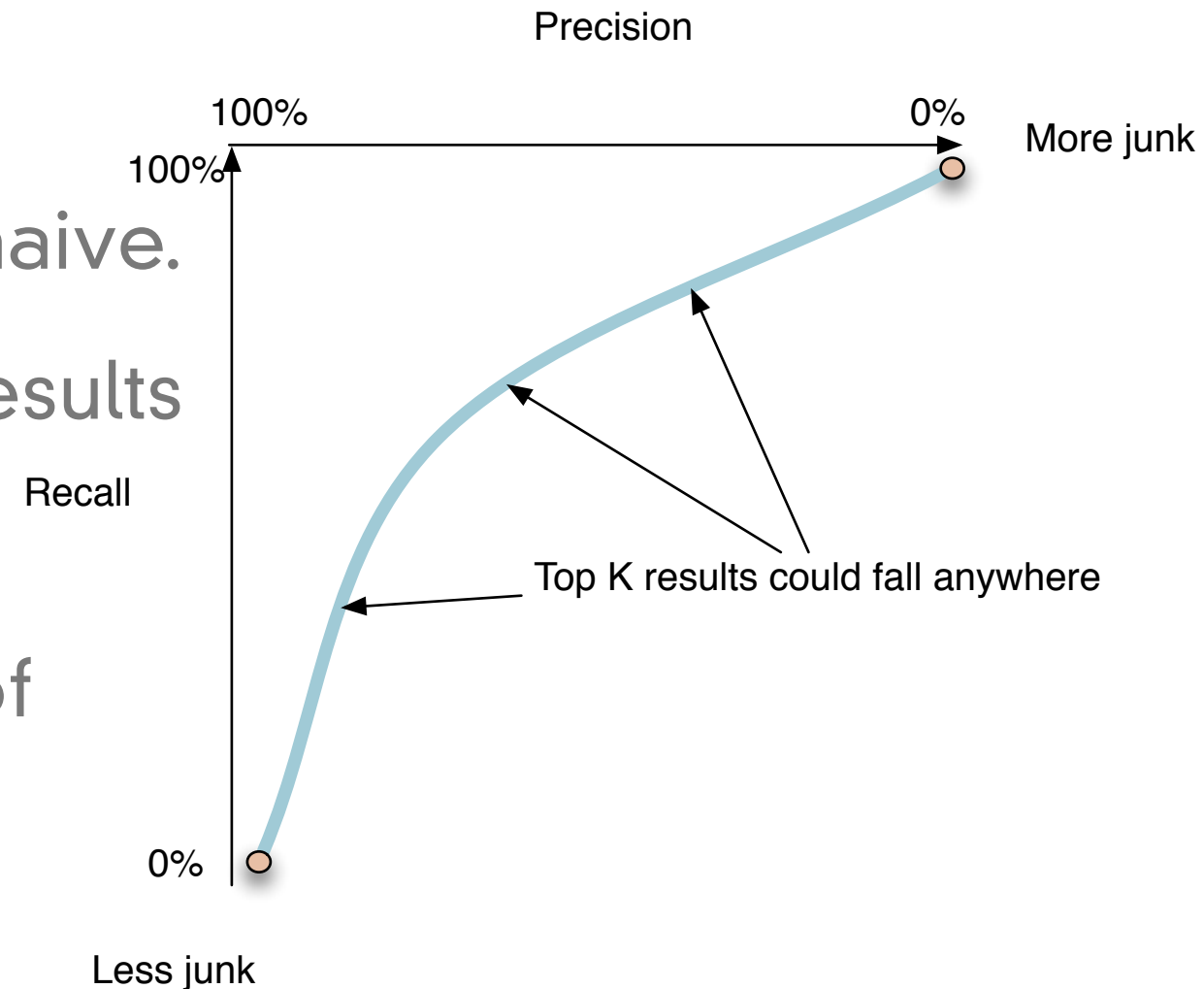
Ranked Retrieval - Mean Average Precision

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



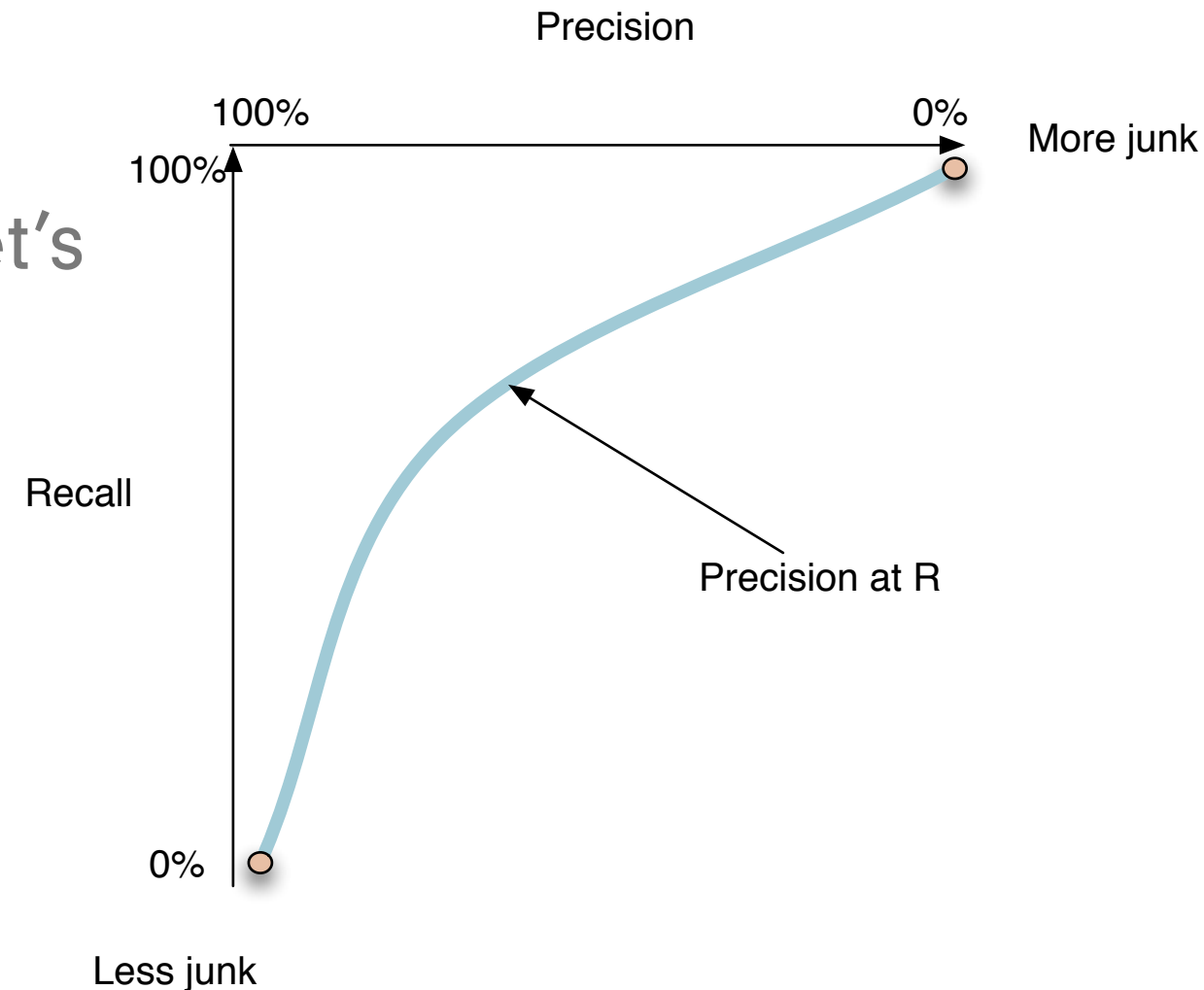
Ranked Retrieval - Precision at k

- 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 of 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?

