IS 531 Document Storage & Retrieval

Chapter 4-Information Retrieval Evaluation

Dr. Bassam Hammo

IRS Evaluation

Let’s see how good is a retrieval system

Evaluation: Precision & Recall

- Recall and precision measure how good a set of retrieved documents is compared with an *ideal* set of relevant documents
  - Recall: What proportion of relevant documents are actually retrieved?
  - Precision: What proportion of retrieved documents are really relevant?

System says…	In reality, the document is…	
	relevant	irrelevant
document is relevant	A	B
document is irrelevant	C	D

$$\text{Precision} = \frac{A}{A+B}$$

Relevant docs retrieved

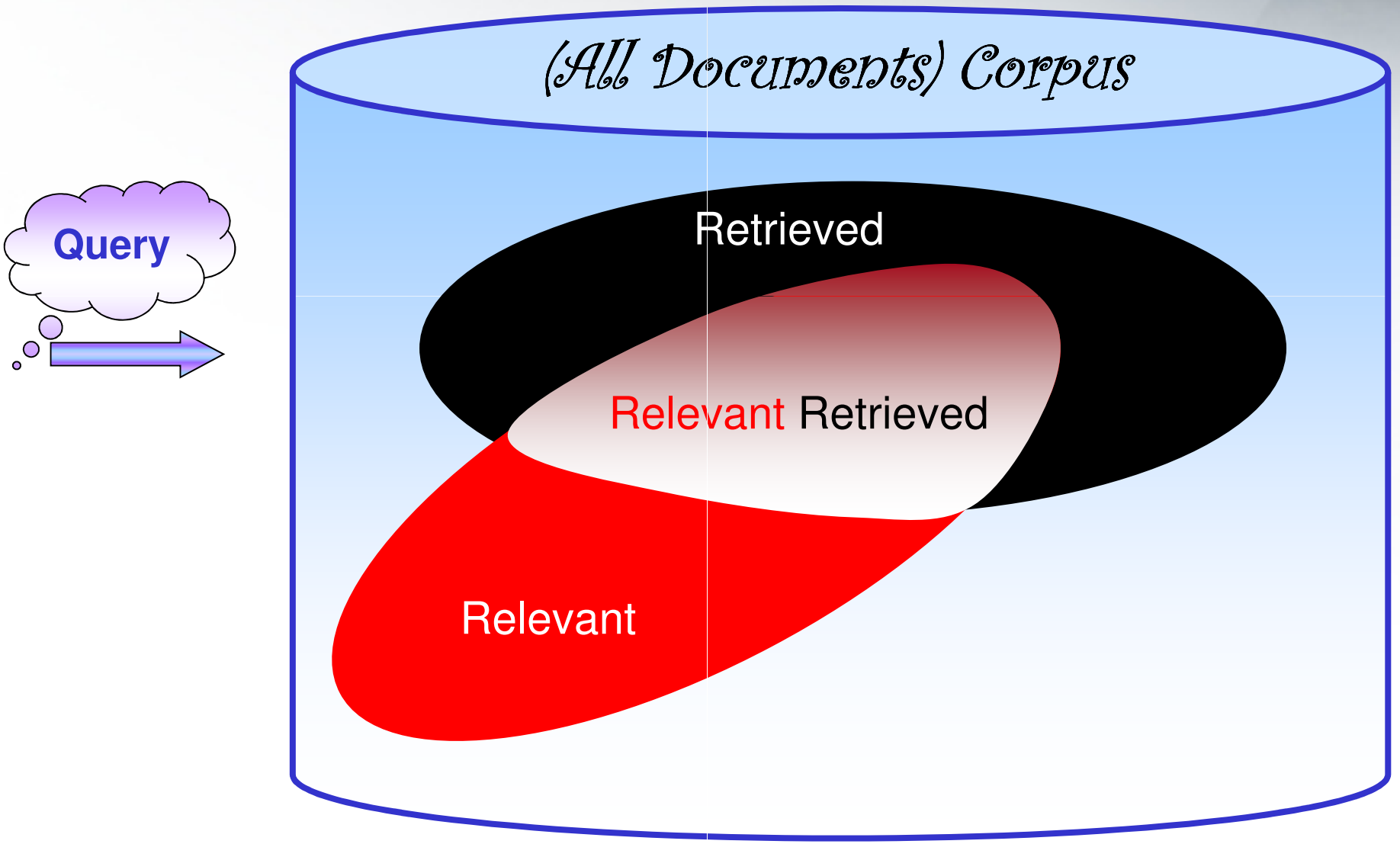
All docs retrieved

$$\text{Recall} = \frac{A}{A+C}$$

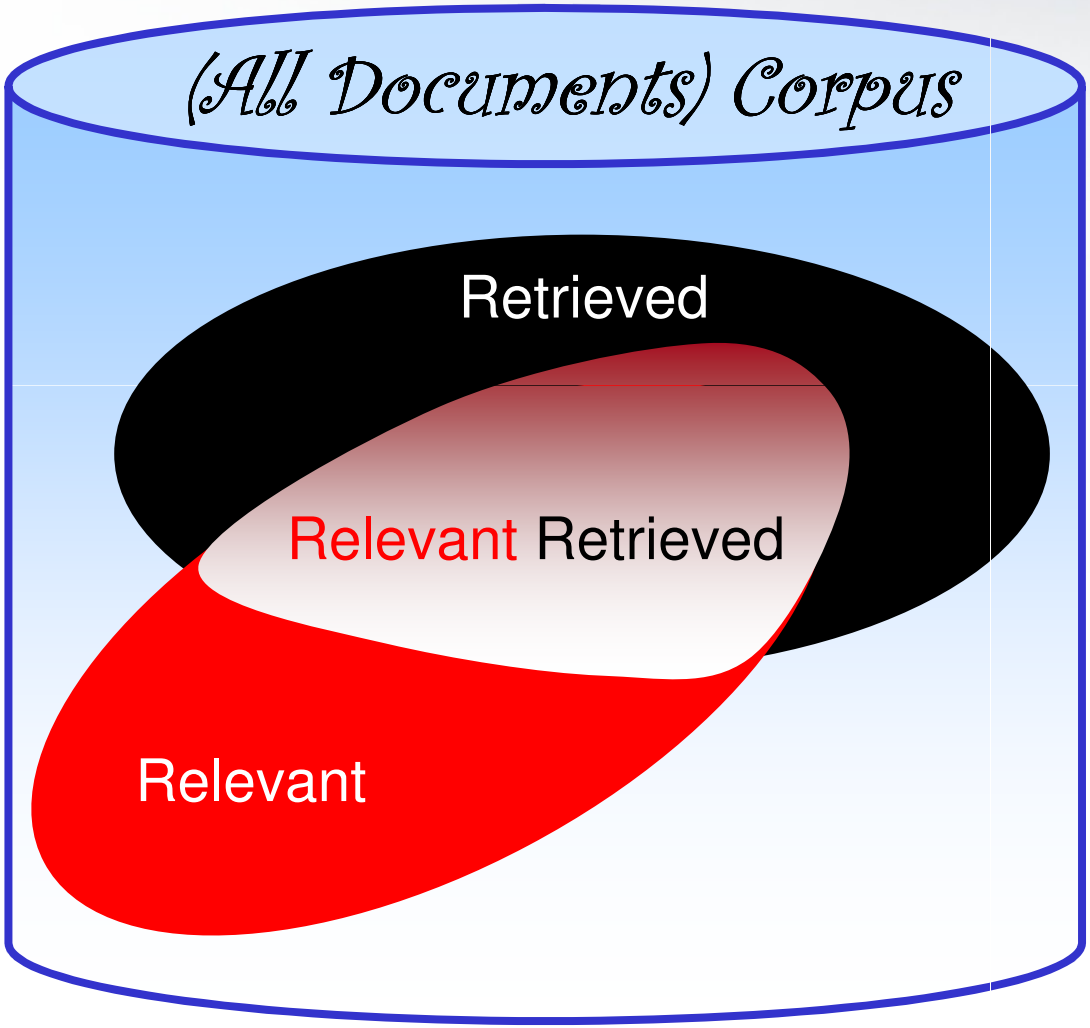
Relevant docs retrieved

All relevant docs (that should have been retrieved)

Relevant vs. Retrieved



Precision vs. Recall



$$\text{Precision} = \frac{\text{Relevant Retrieved}}{\text{Retrieved}}$$

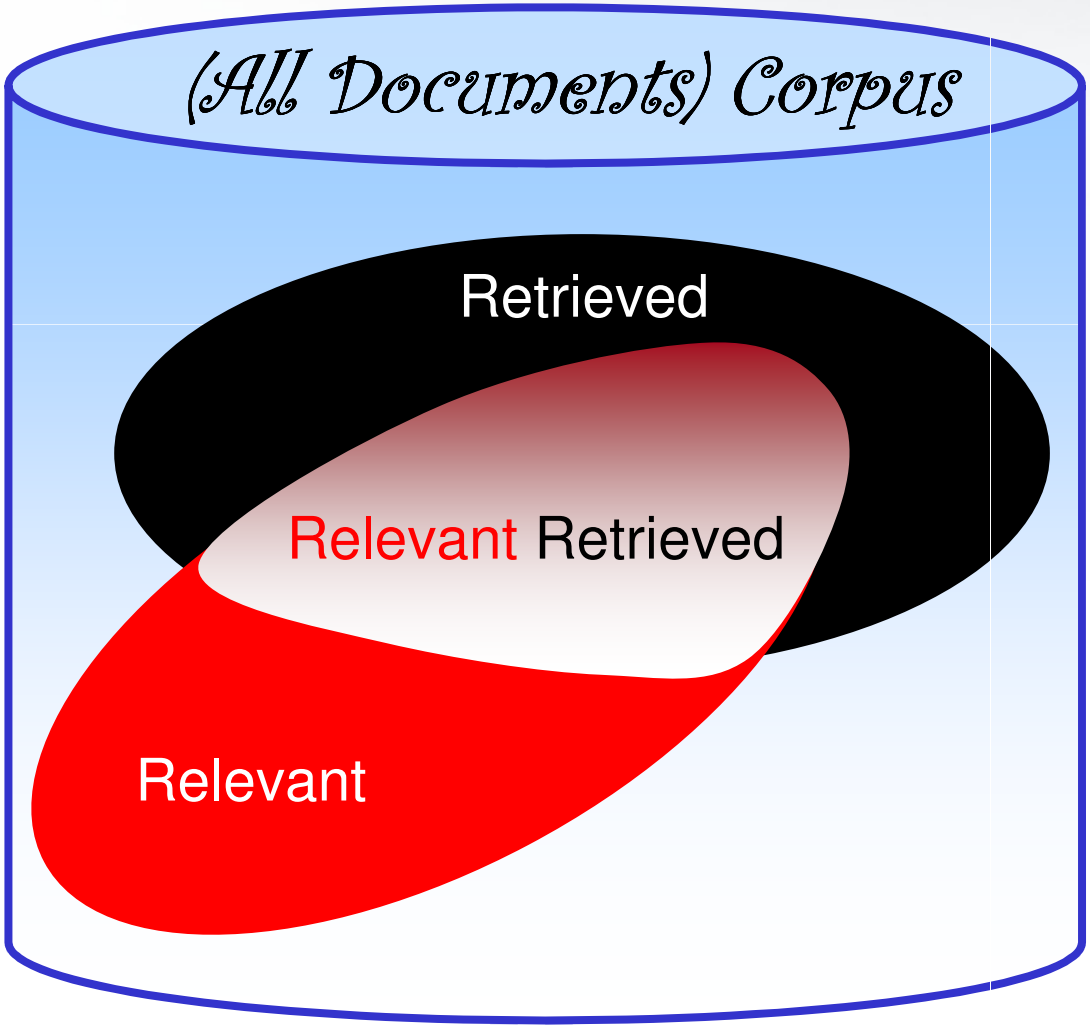
$$\text{Recall} = \frac{\text{Relevant Retrieved}}{\text{Relevant}}$$

Why Precision and Recall?

- Get as much good stuff while at the same time getting as little junk (irrelevant) as possible.

Retrieved vs. Relevant Documents

Very high precision, very low recall

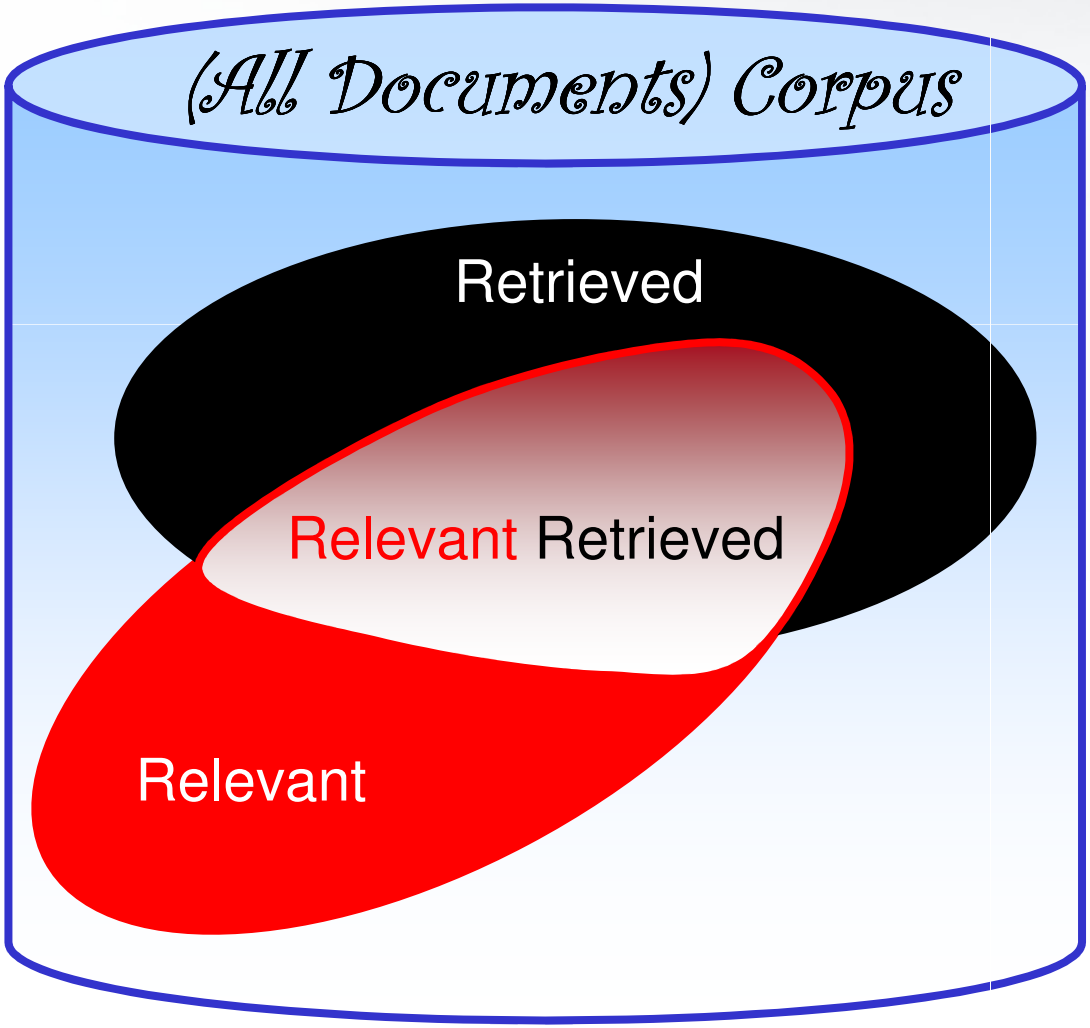


$$\text{Precision} = \frac{\text{Relevant Retrieved}}{\text{Retrieved}}$$

$$\text{Recall} = \frac{\text{Relevant Retrieved}}{\text{Relevant}}$$

Retrieved vs. Relevant Documents

High recall, but low precision

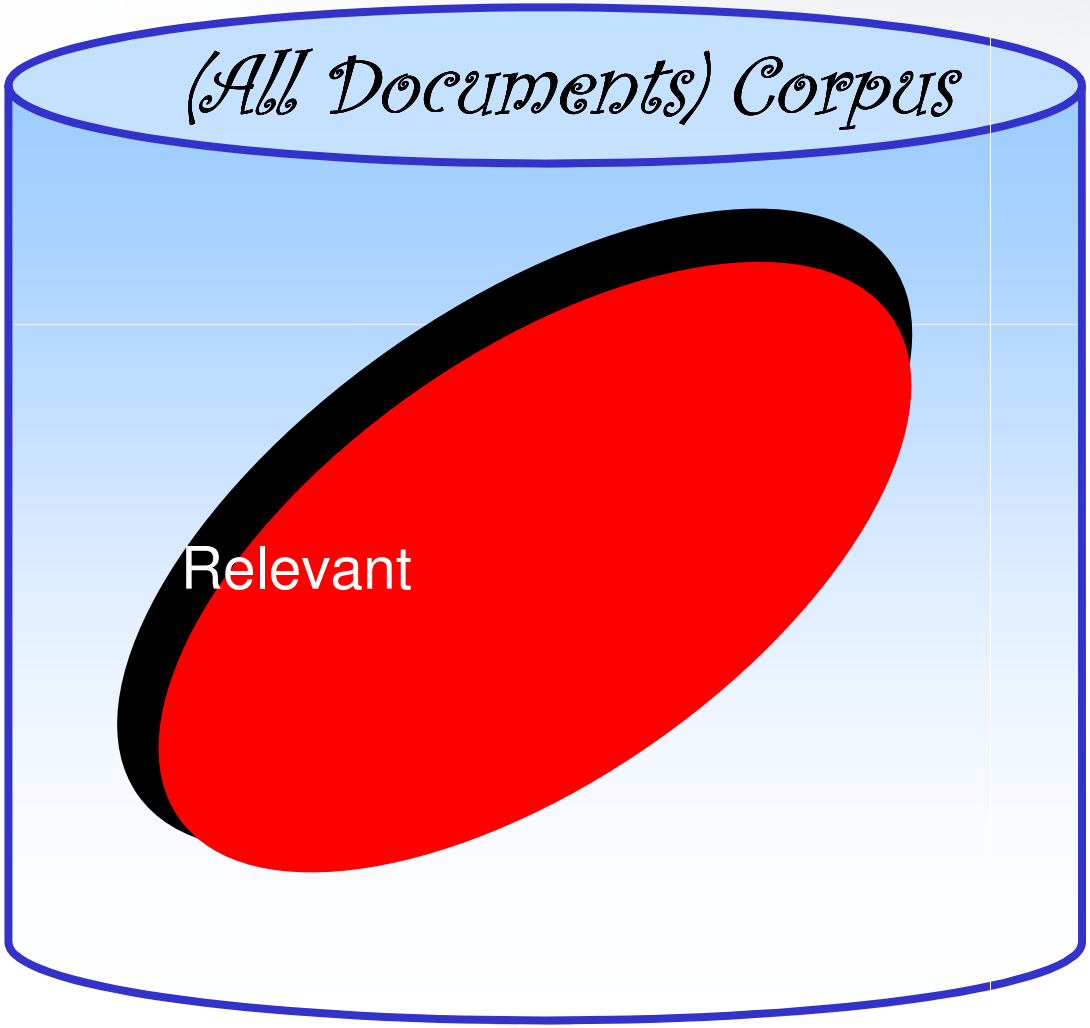


$$\text{Precision} = \frac{\text{Relevant Retrieved}}{\text{Retrieved}}$$

$$\text{Recall} = \frac{\text{Relevant Retrieved}}{\text{Relevant}}$$

Retrieved vs. Relevant Documents

At last! high precision & high recall



$$\text{Precision} = \frac{\text{Relevant Retrieved}}{\text{Retrieved}}$$

$$\text{Recall} = \frac{\text{Relevant Retrieved}}{\text{Relevant}}$$

Evaluation: Example of P&R

- Relevant: d_3 d_5 d_9 d_{25} d_{39} d_{44} d_{56} d_{71} d_{123} d_{389}
- system1: d_{123} d_{84} d_{56}
  - Precision : ??
  - Recall : ??
- system2: d_{123} d_{84} d_{56} d_6 d_8 d_9
  - Precision : ??
  - Recall : ??

Evaluation: Example of P&R

- Relevant: d_3 d_5 d_9 d_{25} d_{39} d_{44} d_{56} d_{71} d_{123} d_{389}

- system1: d_{123}√ d_{84}× d_{56}√
  - Precision: 66% (2/3)
  - Recall: 20% (2/10)

- system2: d_{123}√ d_{84}× d_{56}√ d_6× d_8× d_9√
  - Precision: 50% (3/6)
  - Recall: 30% (3/10)

Evaluation: Problems with P&R

- P&R do not evaluate the ranking
  - $d_{123}\surd$ $d_{84}\times$ ≡ $d_{84}\times$ $d_{123}\surd$
- so other measures are often used:
  - Document cutoff levels
  - P&R curves
  - ...

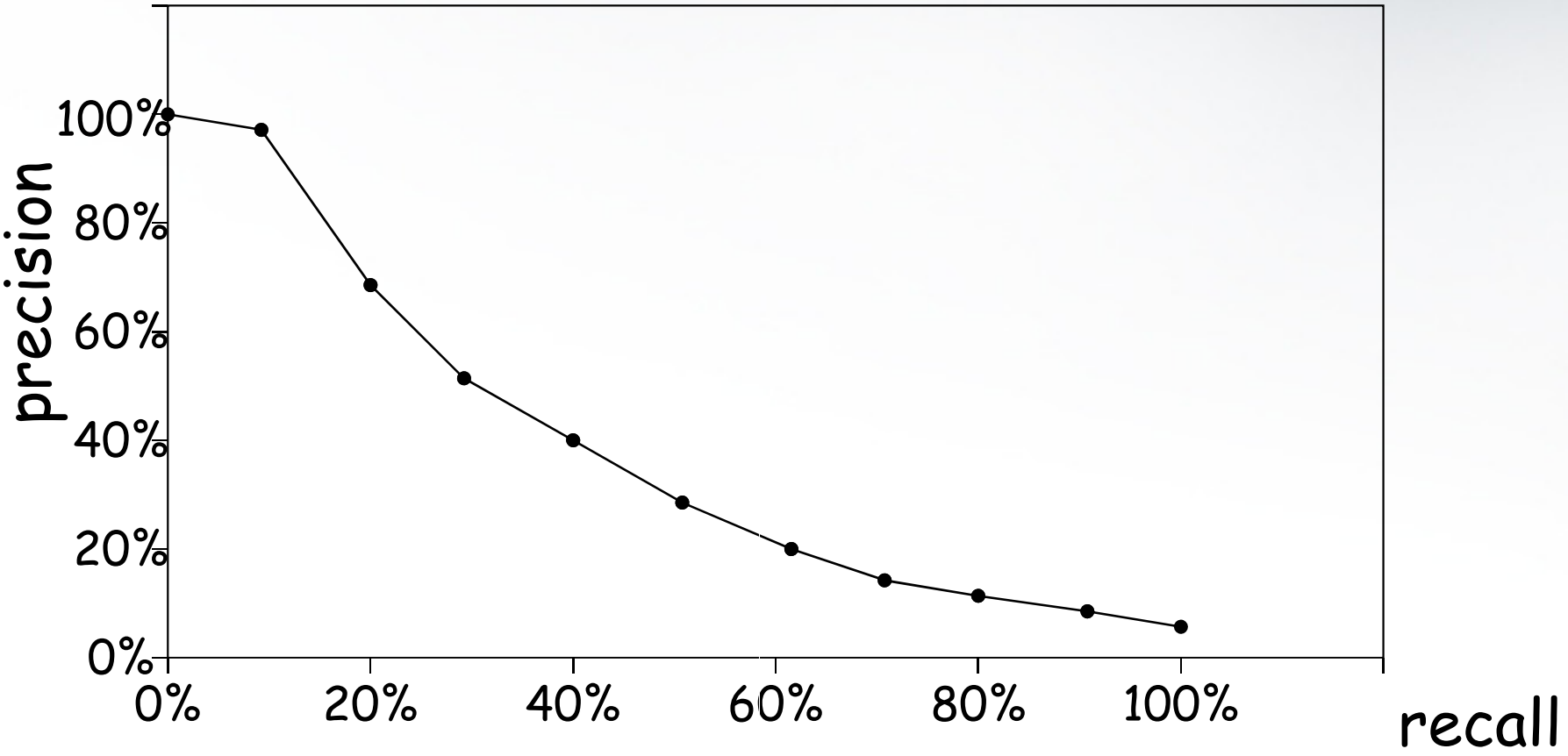
Evaluation: Document cutoff levels

- fix the number of documents retrieved at several levels
  - ex. top 5, top 10, top 20, top 100, top 500…
- measure precision at each of these levels
- Ex:

	system 1	*system 2*	*system 3*
	d1 √	d10 ×	d6 ×
	d2 √	d9 ×	d1 √
	d3 √	d8 ×	d2 √
	d4 √	d7 ×	d10 ×
	d5 √	d6 ×	d9 ×
	d6 ×	d1 √	d3 √
	d7 ×	d2 √	d5 √
	d8 ×	d3 √	d4 √
	d9 ×	d4 √	d7 ×
	d10 ×	d5 √	d8 ×
precision at 5	1.0	0.0	0.4
precision at 10	0.5	0.5	0.5

Evaluation: P&R curve

- measure precision at different levels of recall
  - usually, precision at 11 recall levels (0%, 10%, 20%, …, 100%)



Which system performs better?