Information Retrieval

IR Evaluation

Danilo Montesi
Stefano Giovanni Rizzo

Measuring Relevance

- Methods pioneered by Cyril Cleverdon in the *Cranfield Experiments* in the 1960s

- Three elements:
  1. A benchmark **document collection**
  2. A benchmark suite of **queries**
  3. A **human assessment** of either Relevant or Nonrelevant for each query and each document

Cyril Cleverdon

Assessments

- Note: **user need** is translated into a **query**
- Relevance is assessed relative to the **user need**, *not* the **query**
- E.g., <u>Information need</u>: *My swimming pool bottom is becoming black and needs to be cleaned.*
- <u>Query</u>: ***pool cleaner***
- Assess whether the doc addresses the underlying need, not whether it has these words

Relevance judgments

- Binary (relevant vs. non-relevant) in the simplest case, or more precisely (0, 1, 2, 3 …) in others
- If, for each query, we consider all the set of documents to be judged, the relevance assessment can be huge and expensive
- The depth-**k** pooling solution:
  - Take in consideration the top-**k** (e.g. 100) documents of **N** (e.g. 100) different information retrieval systems
  - Humans must judge a “pool” of no more than **k** x **N** documents (e.g. 10’000), which is far less than the entire document collection (could be millions of documents).

Qualified Test Collections

Collection	*NDocs*	*NQrys*	*Size (MB)*	*Term/Doc*	*Q-D RelAss*
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

Typical TREC

TREC Collections

Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

- The U.S. *National Institute of Standards and Technology* (NIST) has run a large IR test bed evaluation series since 1992. Within this framework, there have been **many tracks** over a range of **different test collections**.

- TREC GOV2 is now the largest Web collection easily available for research purposes, including 25 million pages.

Mechanical Turk

- Present query-document pairs to low-cost labor on online crowd-sourcing platforms
  - Hope that this is cheaper than hiring qualified assessors
- Lots of literature on using crowd-sourcing for such tasks
- Main takeaway – you get some signal, but the variance in the resulting judgments is very high

Effectiveness measures

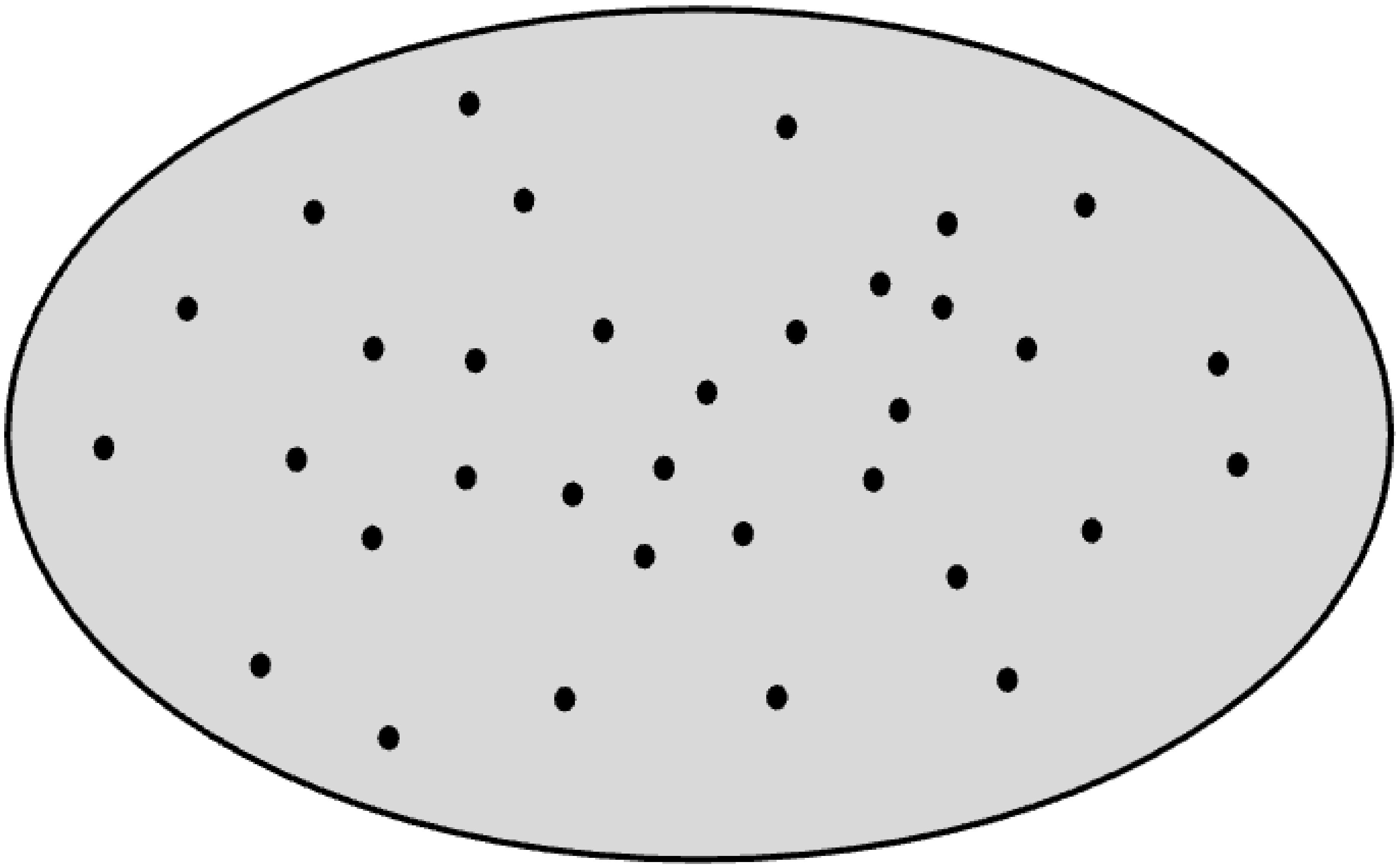
- To assess the *effectiveness* of an IR system (the quality of its search results), there are two parameters about the system's returned results for a query:

  - ***Precision***: What fraction of the returned documents are relevant to the information need?

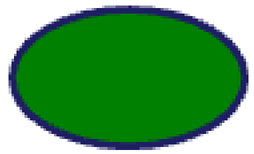
  - ***Recall***: What fraction of the relevant documents in the collection were returned by the system?

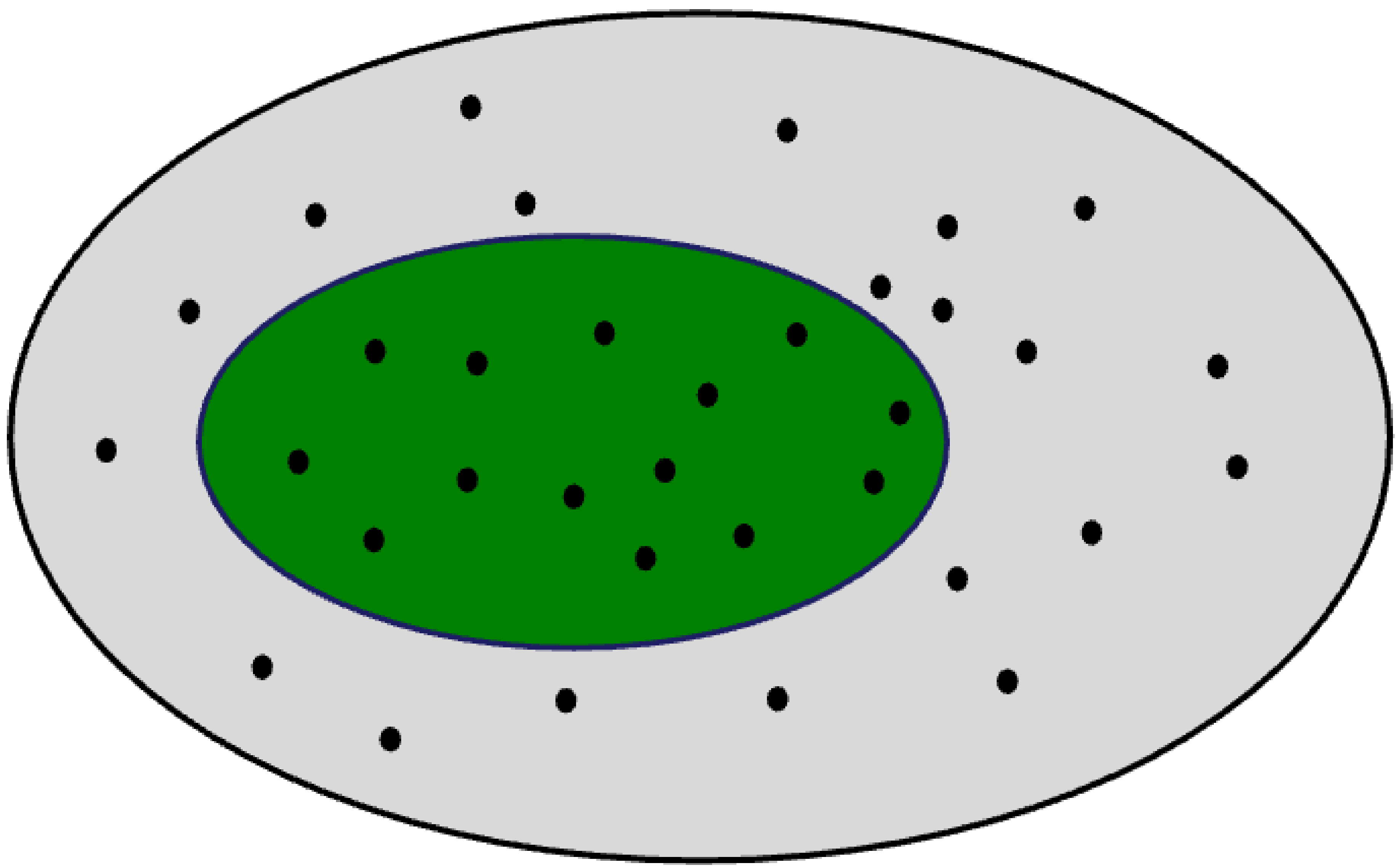
Collection of documents

Each dot • is a document of the collection

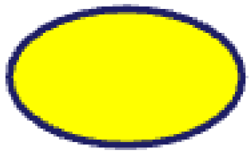


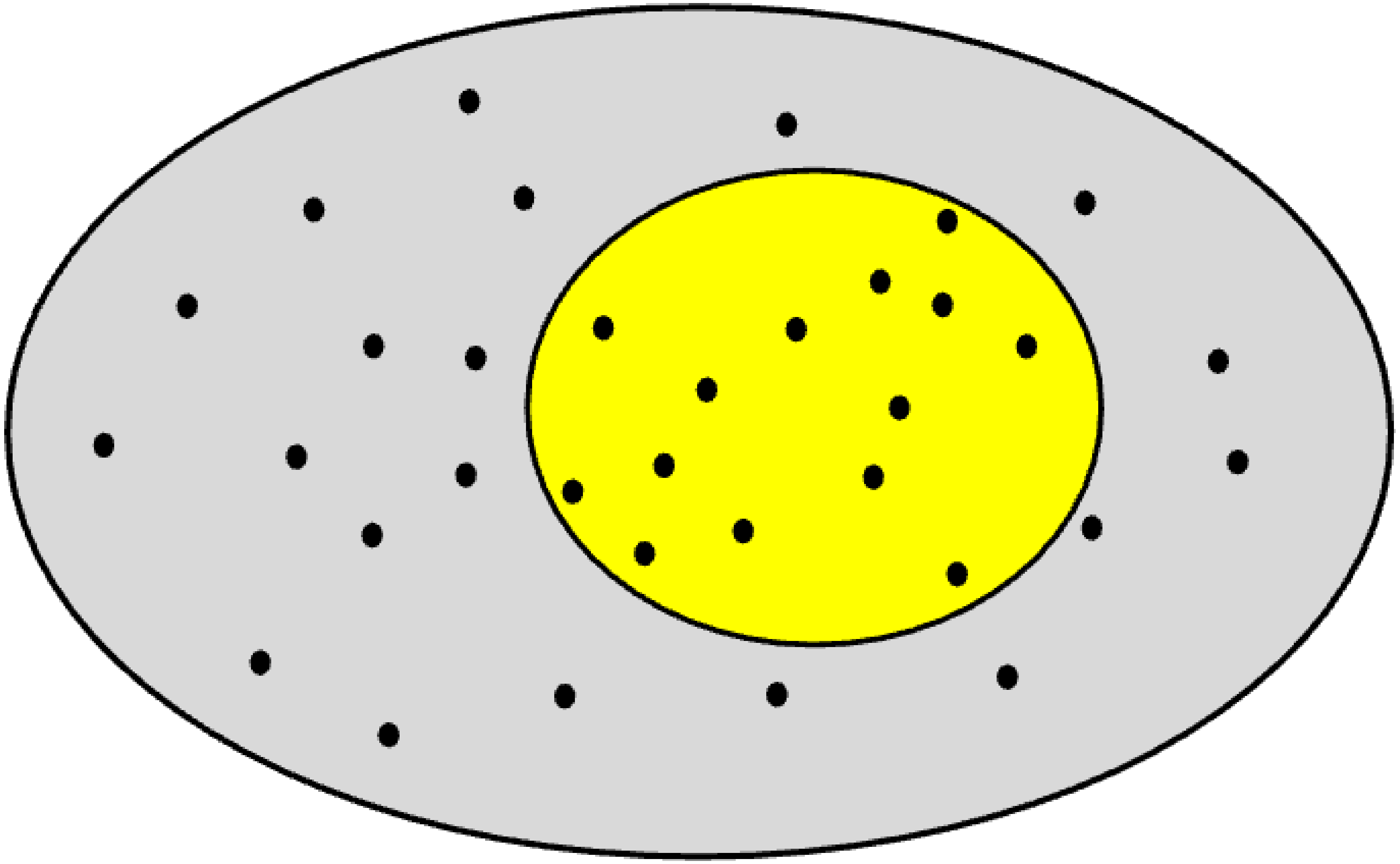
Relevant documents

Given a query, the set in green  is the set of all the documents **really relevant** to the query.



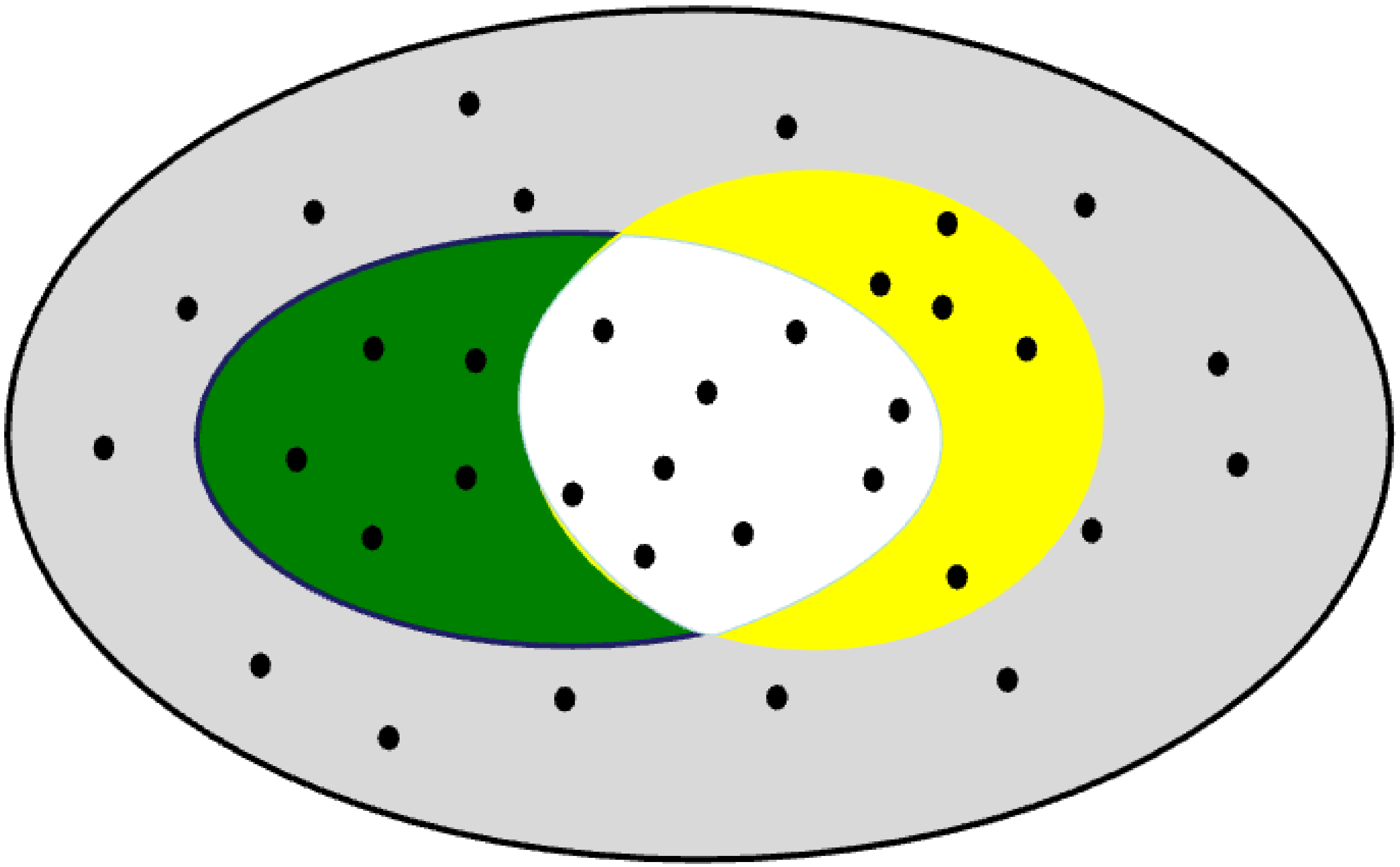
Returned documents

Given the same query, the set in yellow  is the set of all the documents **returned by the system** we want to **evaluate**.



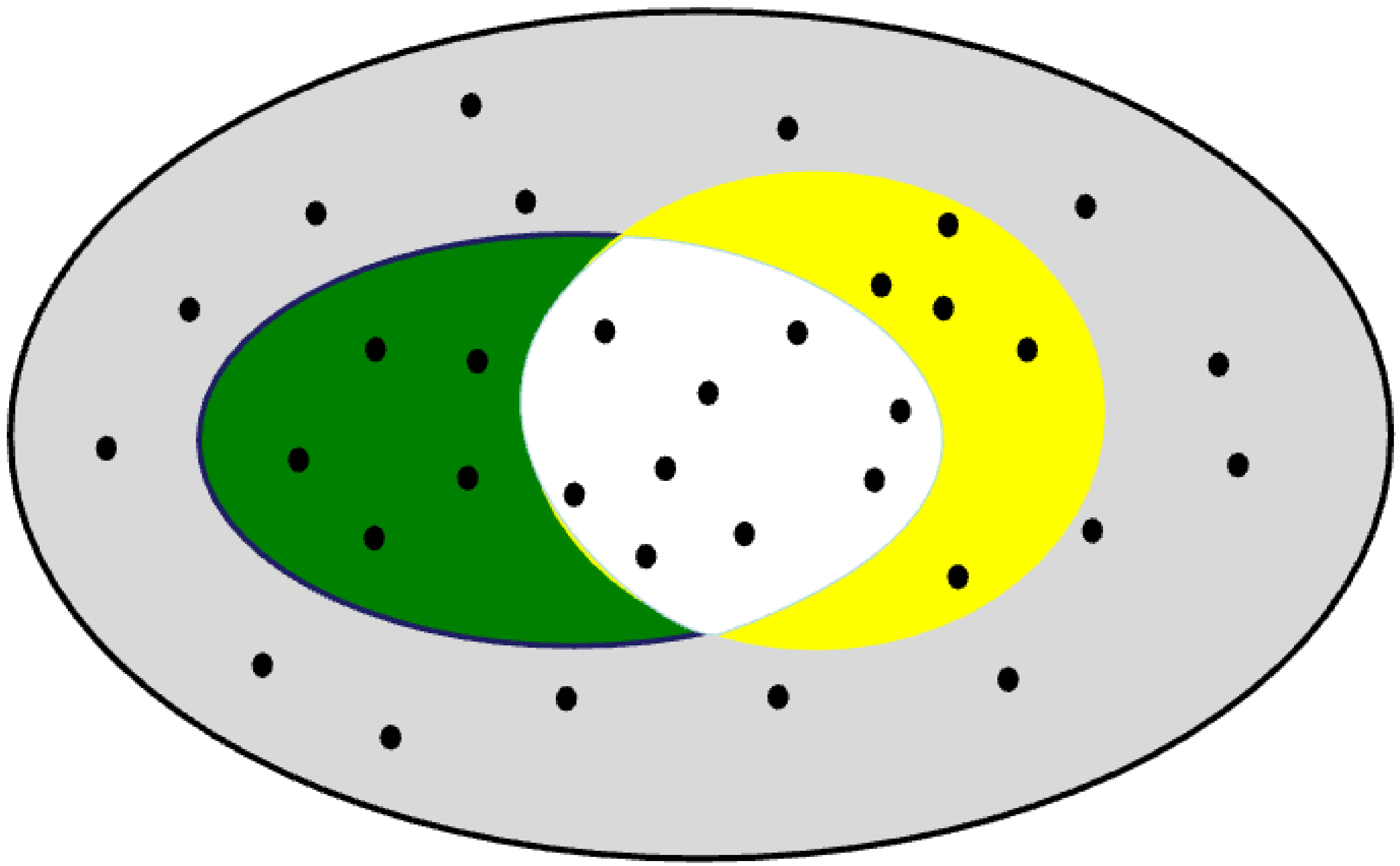
Relevant retrieved documents

Relevant Retrieved Documents = Relevant Documents ∩ Retrieved Documents



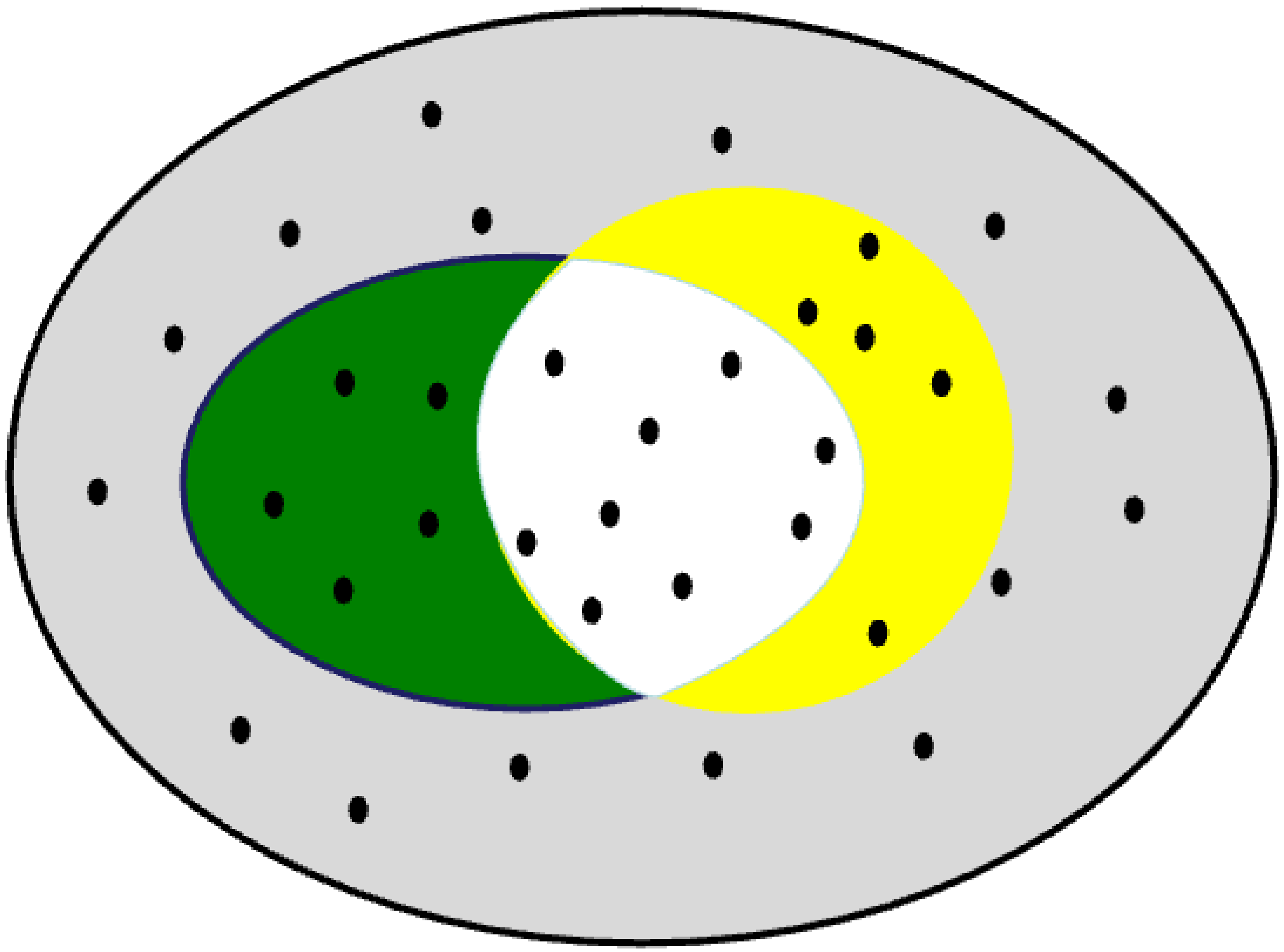
Relevant retrieved documents

Relevant Retrieved Documents = Relevant Documents ∩ Retrieved Documents



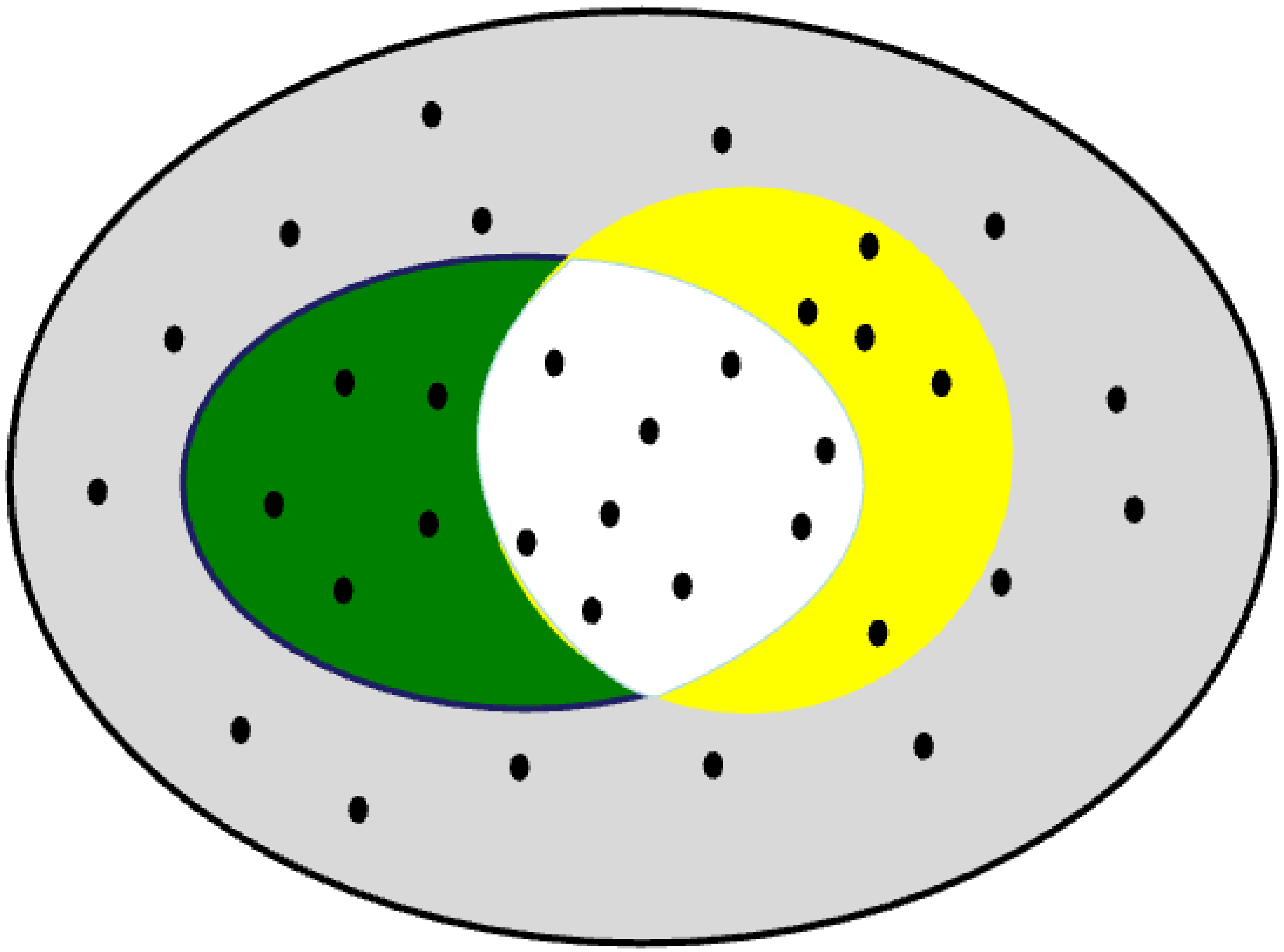
Precision

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



Recall

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



Precision @ K

- Set a rank threshold K
- Compute % relevant in top K
- Ignores documents ranked lower than K

- Ex:
- Precision@1 is 1 rel. /1 ret.
- Precision@2 is 1 rel. /2 ret.
- Precision@3 is 2 rel. /3 ret.

#1 is relevant
#2 is not relevant
#3 is relevant
#4 is not relevant
#5 is relevant

- In similar fashion we have Recall@K

Recall @ K

- Set a rank threshold K
- Compute % relevant in top K
- Ignores documents ranked lower than K

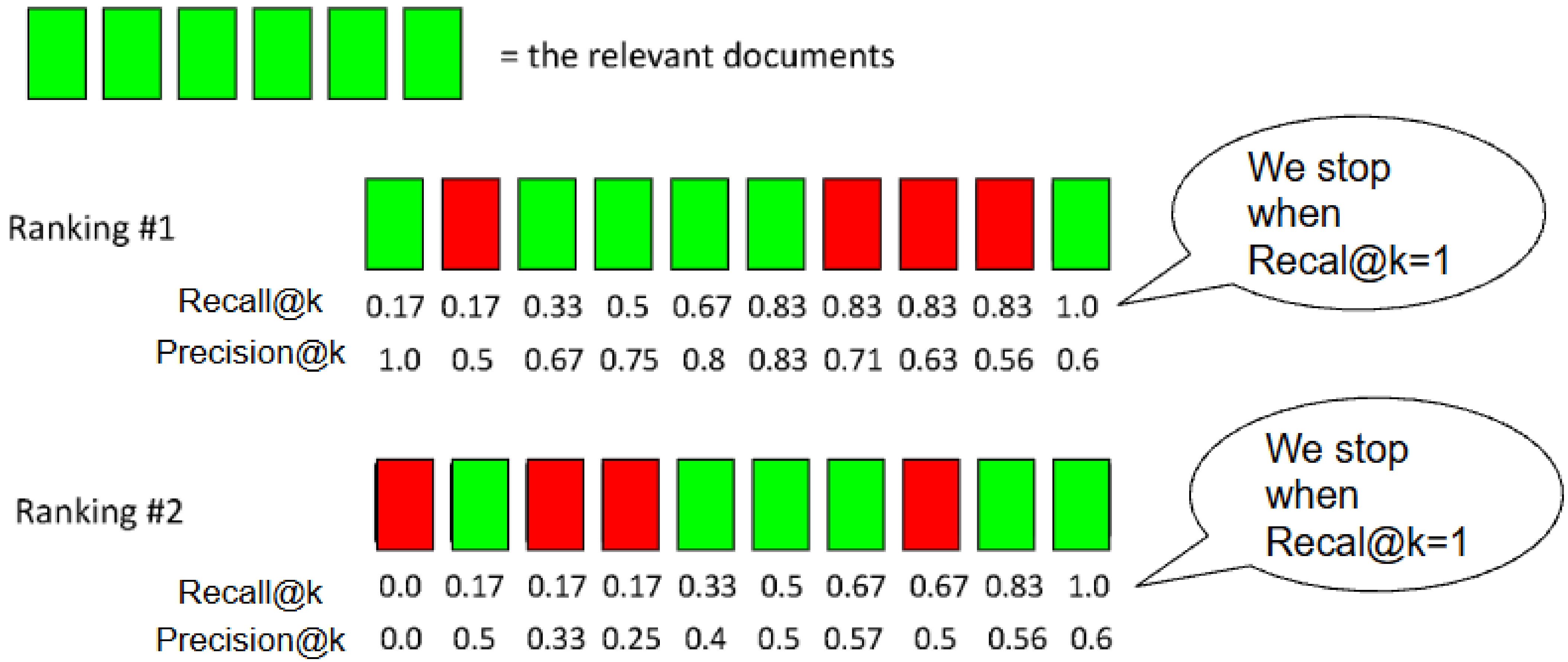
- Ex:
- Recall@1 is 1 rel. /3 rel. tot
- Recall@2 is 1 rel. /3 rel. tot
- Recall@3 is 2 rel. /3 rel. tot

#1 is relevant
#2 is not relevant
#3 is relevant
#4 is not relevant
#5 is relevant

Average Precision (AP)

- Average Precision is an aggregated measure for ranked results.
- It is computed as follows:
  - Instead of setting an arbitrary K we stop only when **all the relevant documents** are retrieved.
  - This coincides with the first K for which Recall@K is equal to 1.
  - We compute the Precisions @K **only for those K where relevant result is retrieved**.
  - The **average** of this Precision measures is the **Average Precision**.

Example: Average Precision (AP)



$$\text{Ranking \#1: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

$$\text{Ranking \#2: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

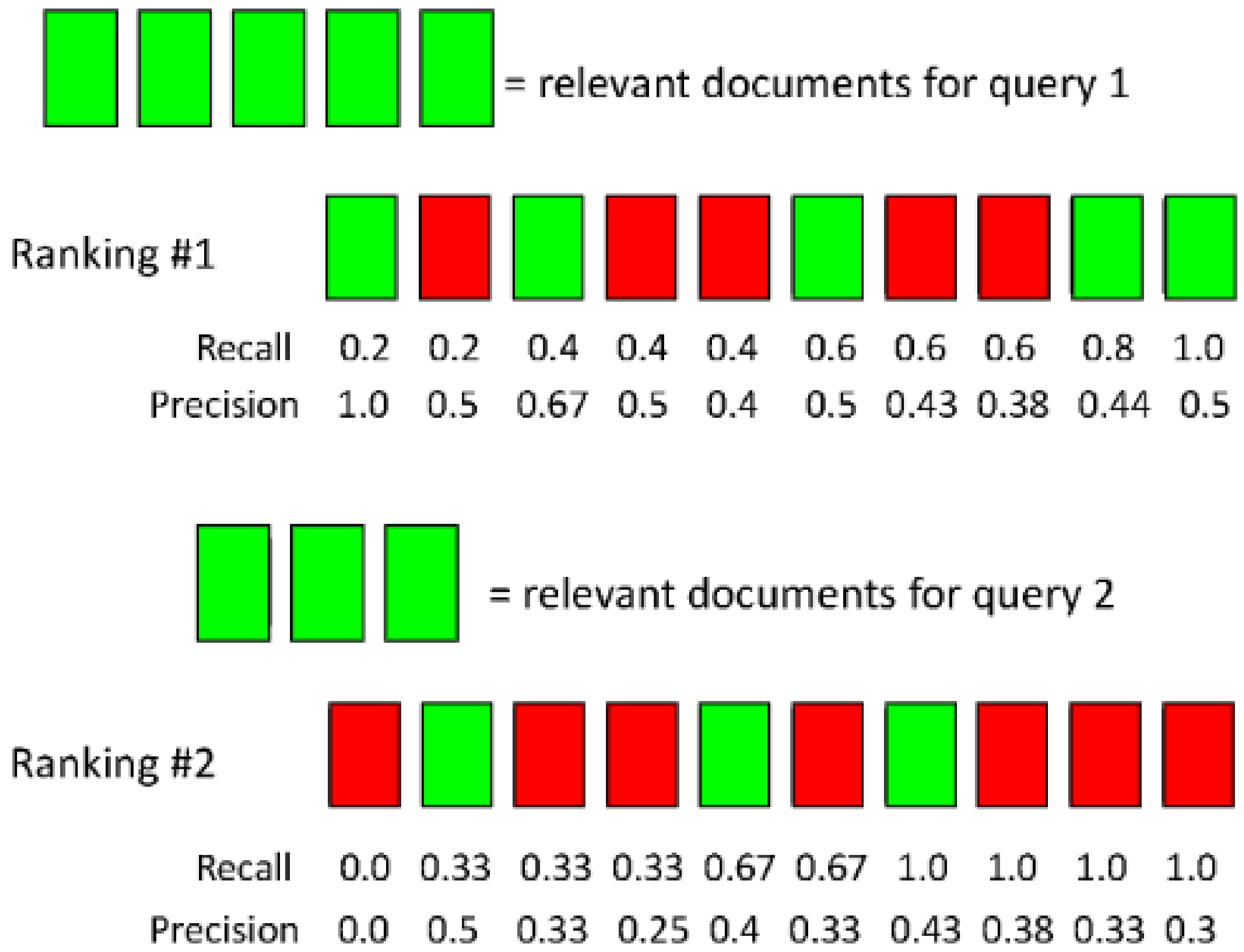
Recall@10 and Precision@10 is equal for the two rankings.
However, AP is able to capture that Ranking #1 is better, as it ranks more relevant documents in higher positions.

Mean Average Precision

- When evaluating a system we usually measure the effectiveness over **more than one query**.
  - Test collections usually span from 50 to 500 queries.

- After computing the Average Precision of each query in the test collection, the **Mean Average Precision (MAP)** is the average of the Average Precision over all the queries.

Mean Average Precision (MAP)



$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$mean\ average\ precision = (0.62 + 0.44)/2 = 0.53$$

MAP - Observations

- If a relevant document never gets retrieved, we assume the precision corresponding to that relevant doc to be zero.
- MAP is macro-averaging: each query counts equally
- Now perhaps most commonly used measure in research papers
- MAP assumes user is interested in finding many relevant documents for each query
- MAP requires many relevance judgments in text collection

Beyond binary relevance

- We assumed a **binary notion of relevance**:
  - either a document is relevant to the query or
  - it is non relevant to the query.
- Some documents can be **less relevant** than others, but still relevant (non binary notion)
  - Specific measure with non-binary assessments: **DCG** (Discounted Cumulative Gain) or **NDCG** (Normalized Discounted Cumulative Gain).
- Binary relevance is still more common and provide a good estimation for IR evaluation

References

[Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze
Introduction to Information Retrieval
Cambridge University Press.
2008]

The book is also online for free:

- HTML edition (2009.04.07)
- PDF of the book for online viewing (with nice hyperlink features, 2009.04.01)
- PDF of the book for printing (2009.04.01)