Information Retrieval

Evaluation of IR

Dr. Seán Russell

School of Computer Science,
University College Dublin

Week 6

Table of Contents

Section Contents

- What does it mean to evaluate something?
- In general, we are asking "how well does the system work?"
- There are many ways that this could be measured:
  - How quickly does the user receive a response?
  - How much memory is used by the system?
  - How much CPU resources are used?
  - Does the user enjoy using the system?
  - How effective is the system in satisfying the user's information need?

- The first three questions relate to the processing done by the system
  - The algorithmic complexity of our algorithm (Big-O) will be reflected in CPU use
  - Other elements of the design (such as caches) will also be reflected in memory use
  - Both of these will combine with the actual system running the code to determine the time taken to get a response
- These are generally important questions, but they relate more to how the system is implemented than it's model

- The fourth question relates to the experience of the user
- User experience (UX) is all about how they perceive the application
  - Is it easy to use?
- Again, an important question but more about the interface than the core of the program

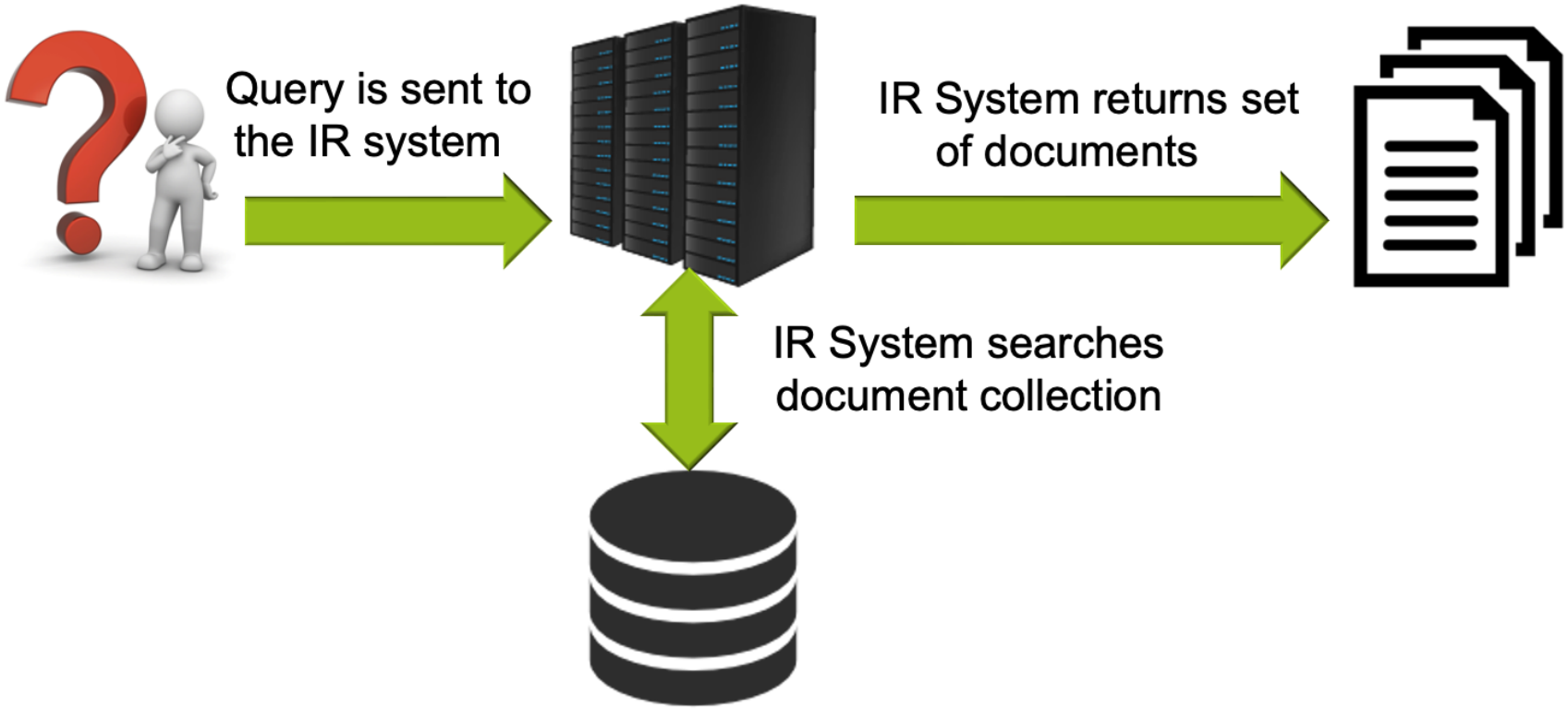
- The final question is about how well the system worked.
- If the system worked correctly, the returned results will contain the information we needed
- This is the most interesting question!
- If we want to compare IR systems, we need to evaluate how well the retrieval algorithms work!

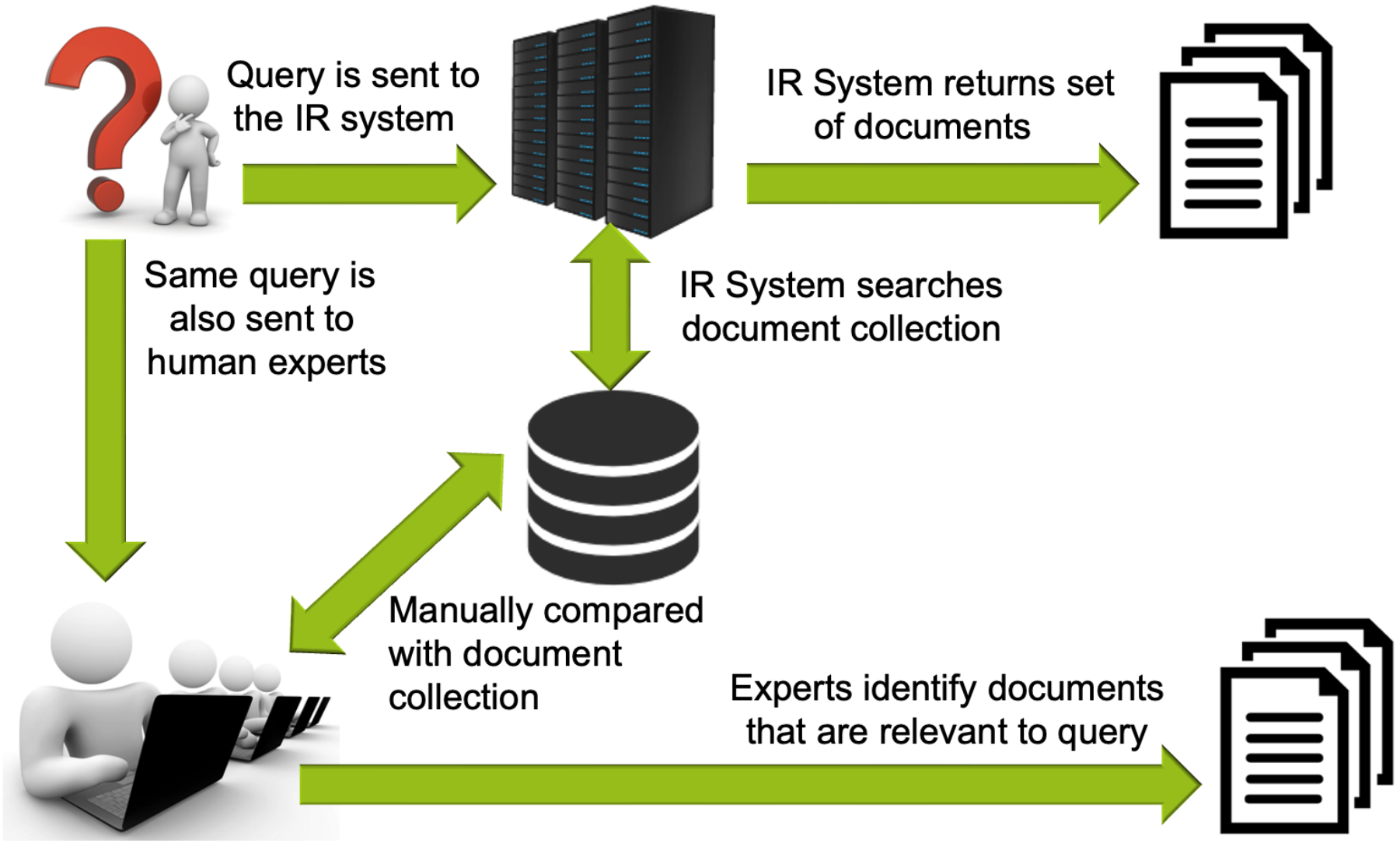
- Evaluation of the effectiveness of an IR system (particularly in the research area) is a vital topic
  - We call this **retrieval evaluation**
- There are many different techniques used in IR and there needs to be accepted ways to quantify their performance
- Many metrics exist to do this

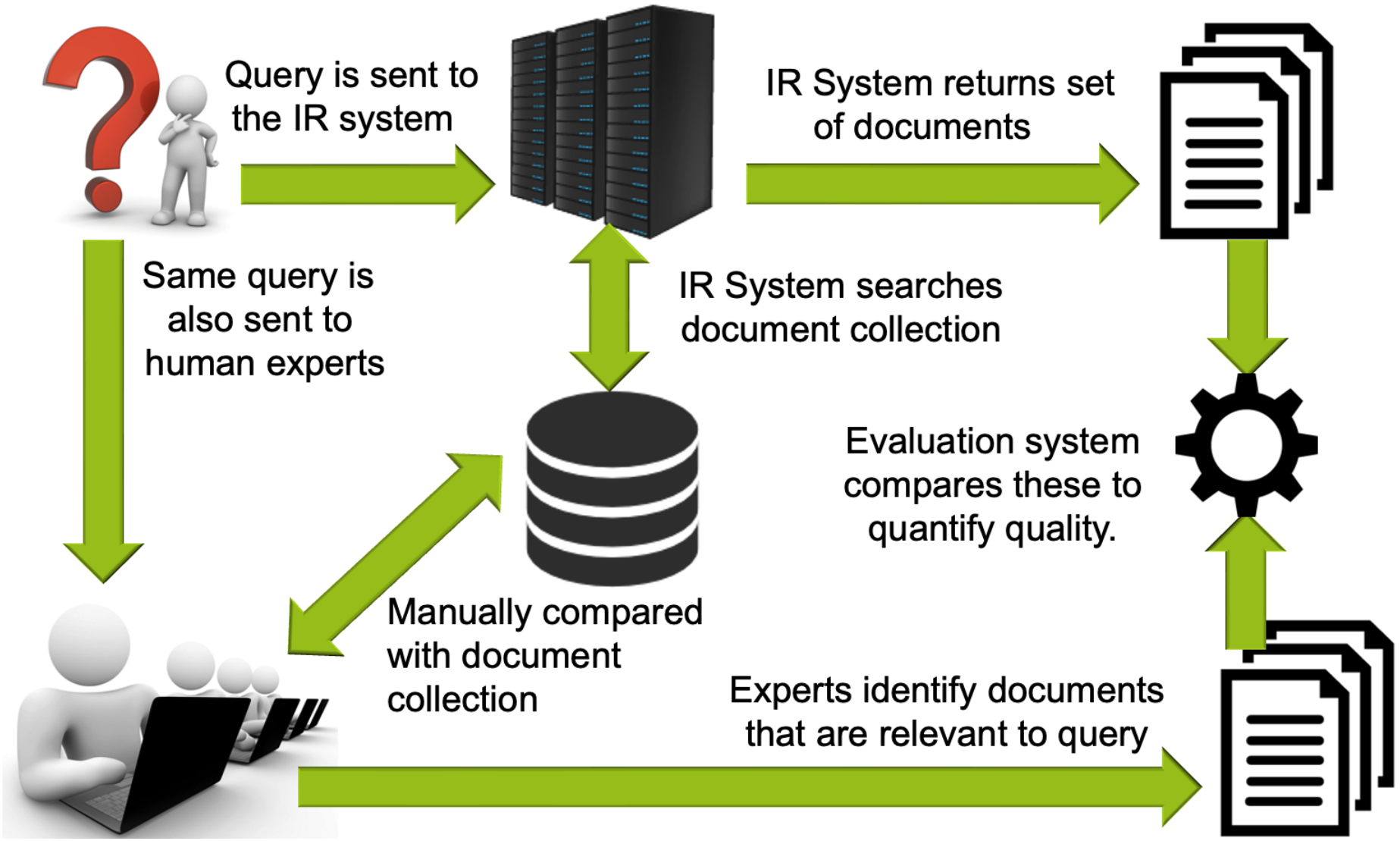
- Below are the commonly used metrics when evaluating IR systems
  - Precision/Recall
  - Precision @ n/R-precision
  - Mean Average Precision (MAP)
  - bPref
  - NDCG
- But first, we need to know how to do the evaluation!

Section Contents

- When we query an IR system, it returns a set of documents
- But how effective was it?
- We need a way of evaluating the response numerically
  - This will allow us to compare the results of different IR systems to the same queries


Query is sent to
the IR system
IR System returns set
of documents
IR System searches
document collection

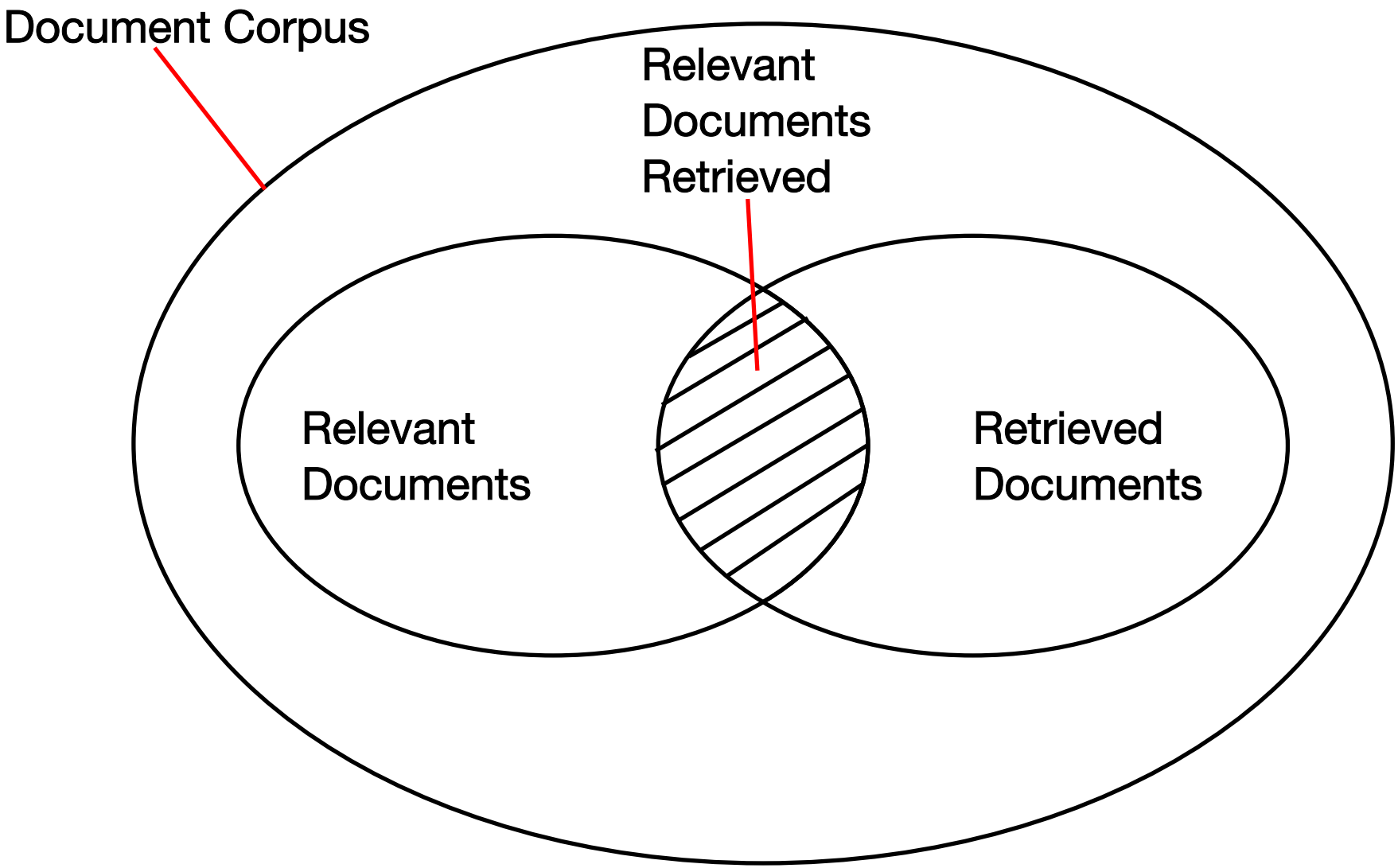

Query is sent to the IR system
IR System returns set of documents
Same query is also sent to human experts
IR System searches document collection
Manually compared with document collection
Experts identify documents that are relevant to query


Query is sent to the IR system
IR System returns set of documents
Same query is also sent to human experts
IR System searches document collection
Evaluation system compares these to quantify quality.
Manually compared with document collection
Experts identify documents that are relevant to query

- This method of evaluation requires that we have experts identify the documents that are **relevant** to a particular query
- A **relevant document** is one that (at least partially) satisfies a user's information need
- Unfortunately, an information need is a very subjective thing
- This is the method used in the Cranfield experiments

- Evaluations carried out in this way has led to the creation of **standard test collections**
- These collections consist of
  - A **standard corpus** used for all queries
  - A set of **standard queries**
  - A set of **relevance judgements**
- I.e. For each of the standard **queries**, the human judges/experts have created a list of all of the documents in the **corpus** that are **relevant**

- When planning our evaluation, we can talk about 3 sets of documents
  - C: The corpus is the set of all available documents
  - Rel: The relevant set is the documents that have been judged as relevant by experts for that query
  - Ret: The retreived set is the documents that the IR system has returned in response to the same query


Document Corpus
Relevant
Documents
Retrieved
Relevant
Documents
Retrieved
Documents

- In reality, the answer set is not really a set!
- Generally it is in the form of a ranked list
  - The boolean model is the exception to this
- When we evaluate the effectiveness of an IR technique, we need to evaluate the quality of this ranked list

- Consider a query q, on a document collection
- $\text{Rel} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$
- Ret:

1. d_{123}
2. d_{84}
3. d_{56}
4. d_6
5. d_8
6. d_9
7. d_{511}
8. d_{129}
9. d_{187}
10. d_{25}
11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_3