Evaluation: Introduction & The Cranfield Paradigm

COMP3009J: Information Retrieval

Dr. David Lillis (david.lillis@ucd.ie)

UCD School of Computer Science
Beijing Dublin International College

Further reading:

- Modern Information Retrieval (2nd ed.): Chapter 4
- Modern Information Retrieval (1st ed.): Chapter 3
- An Introduction to Information Retrieval: Chapter 8

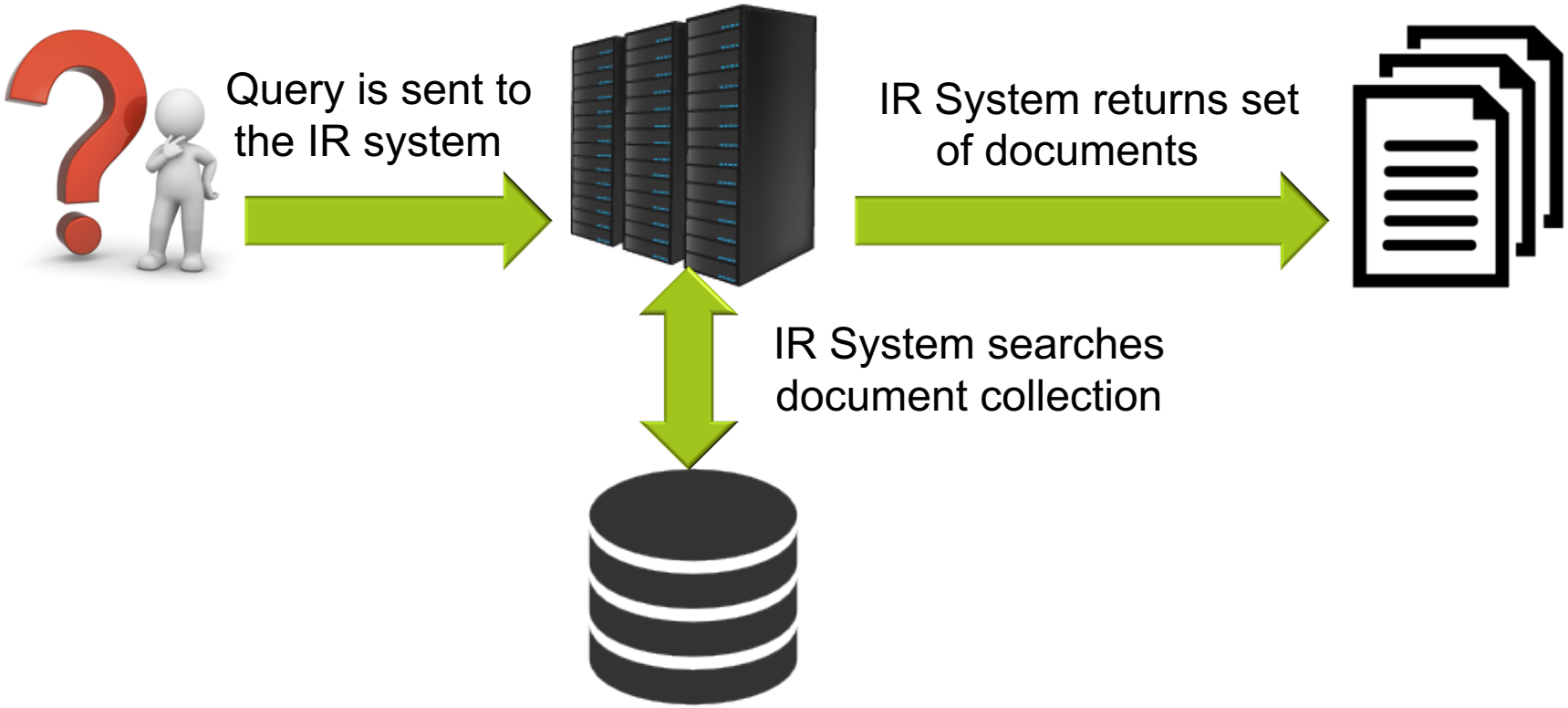
Introduction

- Evaluation is concerned with the question "How well does the system work?"
- There are many measureable quantities for this:
  1. **Processing:** How quickly does the user receive a response? How well are resources utilised?
  2. **User Experience**: Does the user enjoy using the system?
  3. **Search:** How effective is the system in satisfying the user's information need?
- Question 3 is of most interest in this lecture. This evaluate the actual retrieval algorithms that we are using.

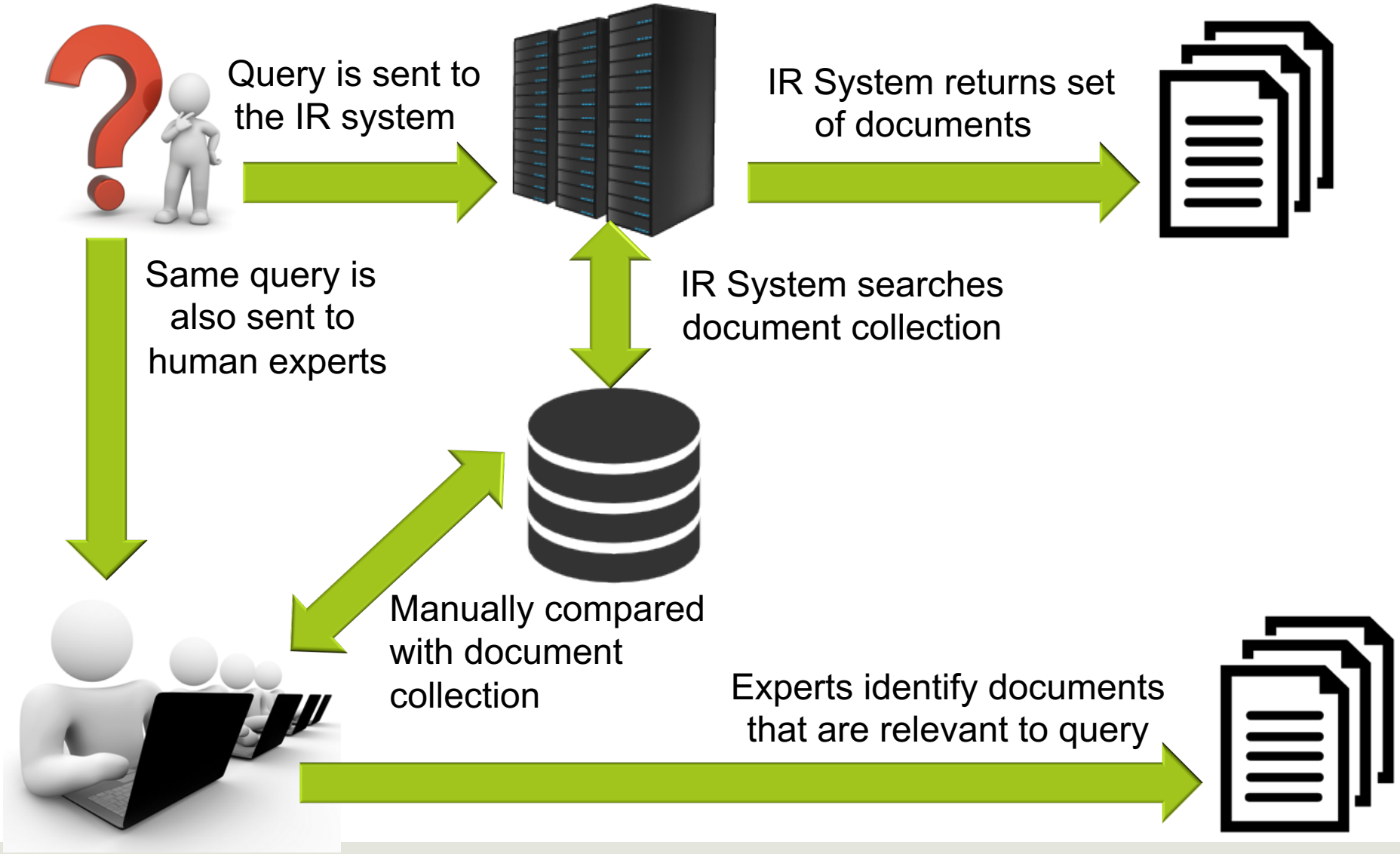
Introduction

- Evaluation of the effectiveness of an IR system (particularly in the research area) is a vital topic.
- 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.
- We will look at the following commonly used metrics:
  - Precision/Recall
  - Precision @ n/R-precision
  - Mean Average Precision (MAP)
  - bPref
  - NDCG

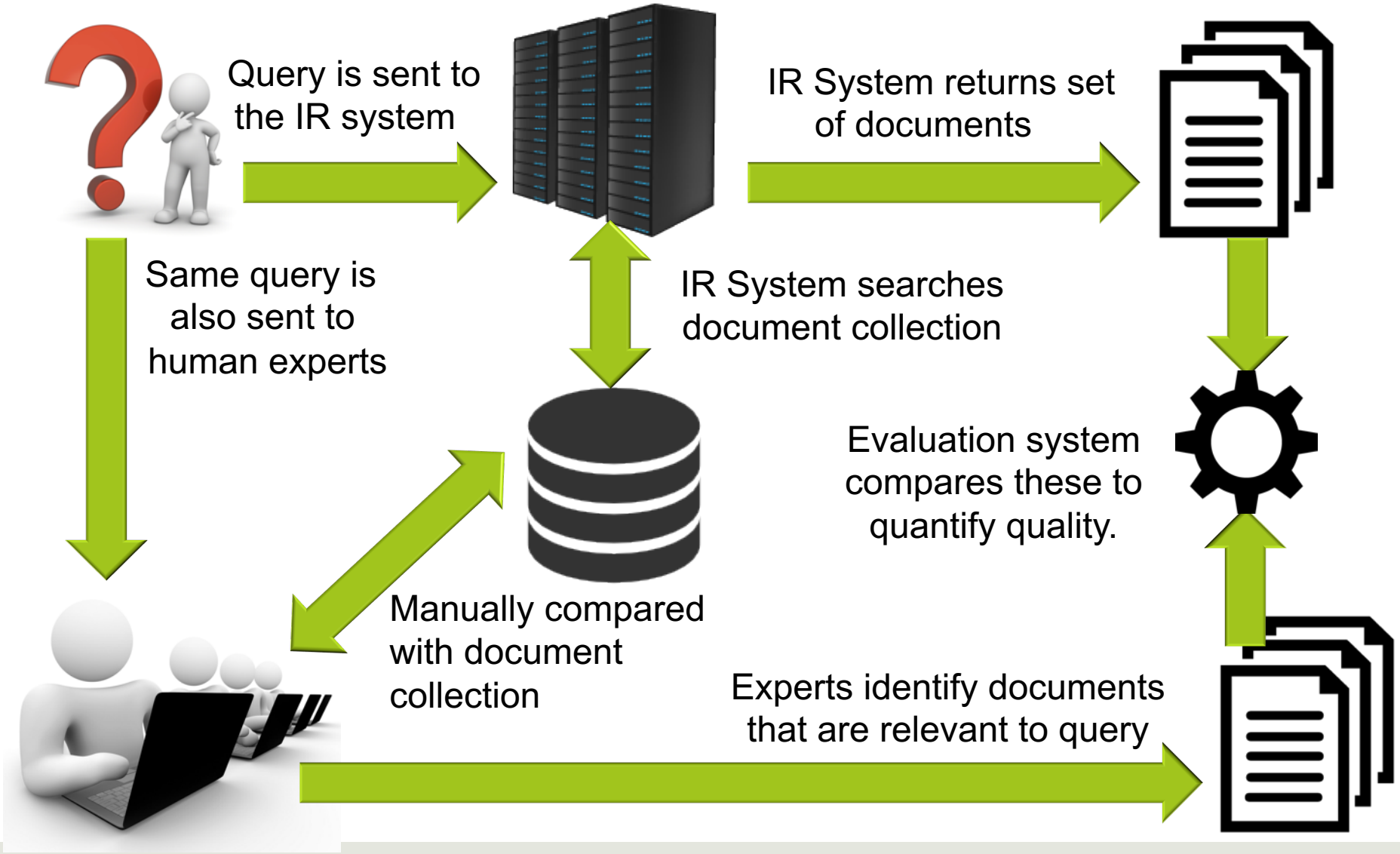
Cranfield Paradigm



Cranfield Paradigm



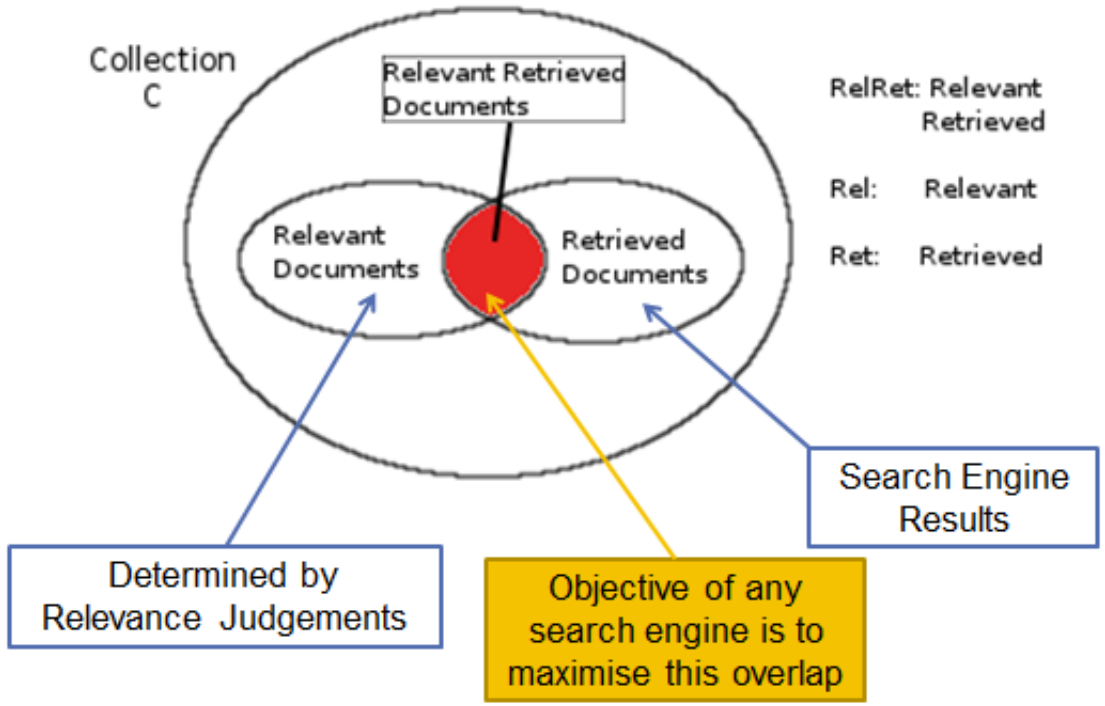
Cranfield Paradigm



Introduction: The Cranfield Paradigm

- A **relevant document** is one that (at least partially) satisfies a user's information need.
- Unfortunately, an information need is a very subjective thing.
- As an alternative, we use experts to judge whether each document is relevant to each query (the Cranfield experiments used aeronautical engineers).
- We can talk about 3 **sets** of documents.
  - The **collection** is the set of all available documents.
  - The **answer set** is the set of documents that an IR system has returned in response to a query.
  - The **relevant set** is the set of documents that have been judged by the experts to be relevant for that query.

Relevance


Collection
C
Relevant Retrieved
Documents
Relevant
Documents
Retrieved
Documents
RelRet: Relevant
Retrieved
Rel: Relevant
Ret: Retrieved
Search Engine
Results
Determined by
Relevance Judgements
Objective of any
search engine is to
maximise this overlap

Introduction

- In reality, the answer set is normally not really a set.
- Generally it is in the form of a **ranked list**.
  - The Boolean Model is the exception to this.
- The purpose of evaluating the effectiveness of an IR technique is to evaluate the quality of this ranked list.

Example (from Modern Information Retrieval)

- Consider a query *q*, on a document collection C where |C|=800
- *Rel* = {d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}}
- The ranked list of retrieved documents, *Ret* is given by:

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