The Probabilistic Model:
Introduction

COMP3009J: Information Retrieval

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

UCD School of Computer Science
Beijing Dublin International College

The Probabilistic Model

- Along with the Boolean and Vector Space models, the other “classic” model is the **Probabilistic Model**.
  - Also known as the “binary independence retrieval” model.
- Originally proposed by Robertson and Sparck Jones in 1976.
- Pages 79-86 of Modern Information Retrieval (2nd Edition)

Basics

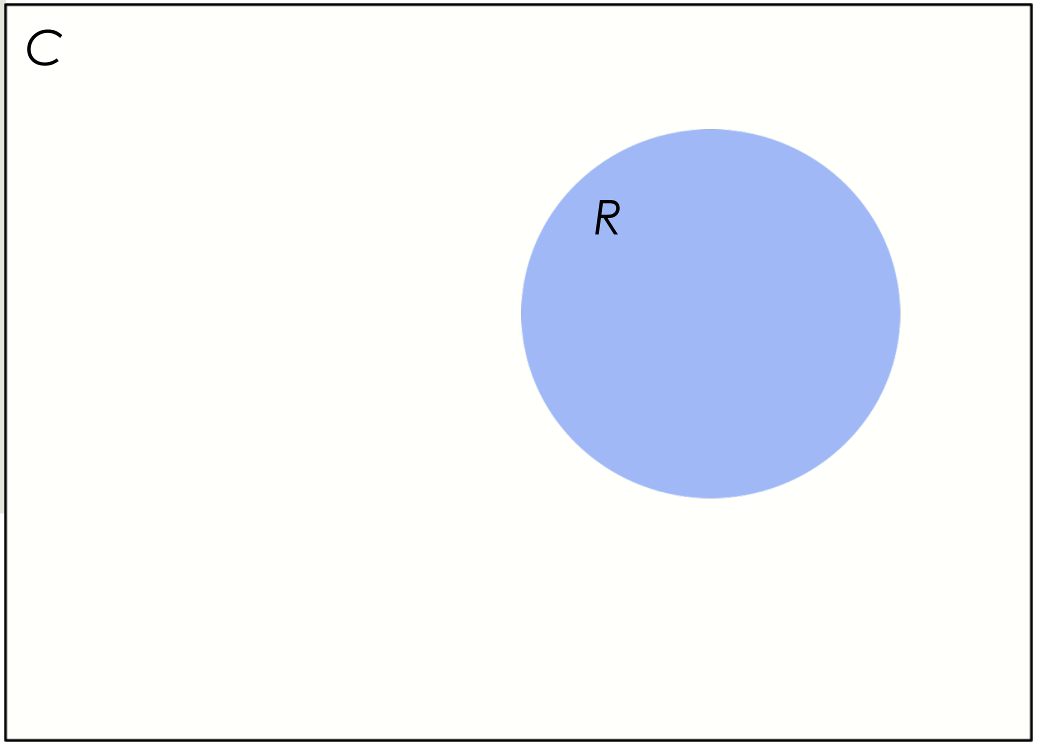
- The premise underlying the probabilistic model is that there exists a set of documents that contains all the documents that are relevant to the user's information need and no others.
- This is referred to as the **ideal answer set**, or *R*.
- If we had a full description of this ideal answer set, we would have no problems retrieving its documents.
- We can think of the query process as the process of trying to specify what the properties of *R* actually are.

R: the ideal answer set

C: the **corpus** (i.e. the set of all documents in the index).

R: the **ideal answer set** (i.e. the set of documents that are relevant to the user's information need).

All documents outside *R* are not relevant. This can also be described as $\bar{R}$



Basics

- We know that there are index **terms** available that describe the documents in some way.
- These terms should be used to describe the properties of the ideal answer set.
- At the beginning of the retrieval process, we don't know what the properties of the ideal answer set are, so we must attempt an **initial guess** so as to return a **reasonable initial set** of documents to the user.
- The user can then interact with the system to help describe what the ideal answer set should be.
- This **user interaction** is a key difference between the Probabilistic Model and the others we have seen so far.

Basics

- The user looks at the documents that have been retrieved and marks which ones are **relevant.**
- **Remember** only the user knows what the information need is: a query is just an attempt to express this.
- The belief behind the Probabilistic Model is that the information need can be better defined by having users expressly state which documents help to satisfy their information need, rather than reducing it to a few keywords.

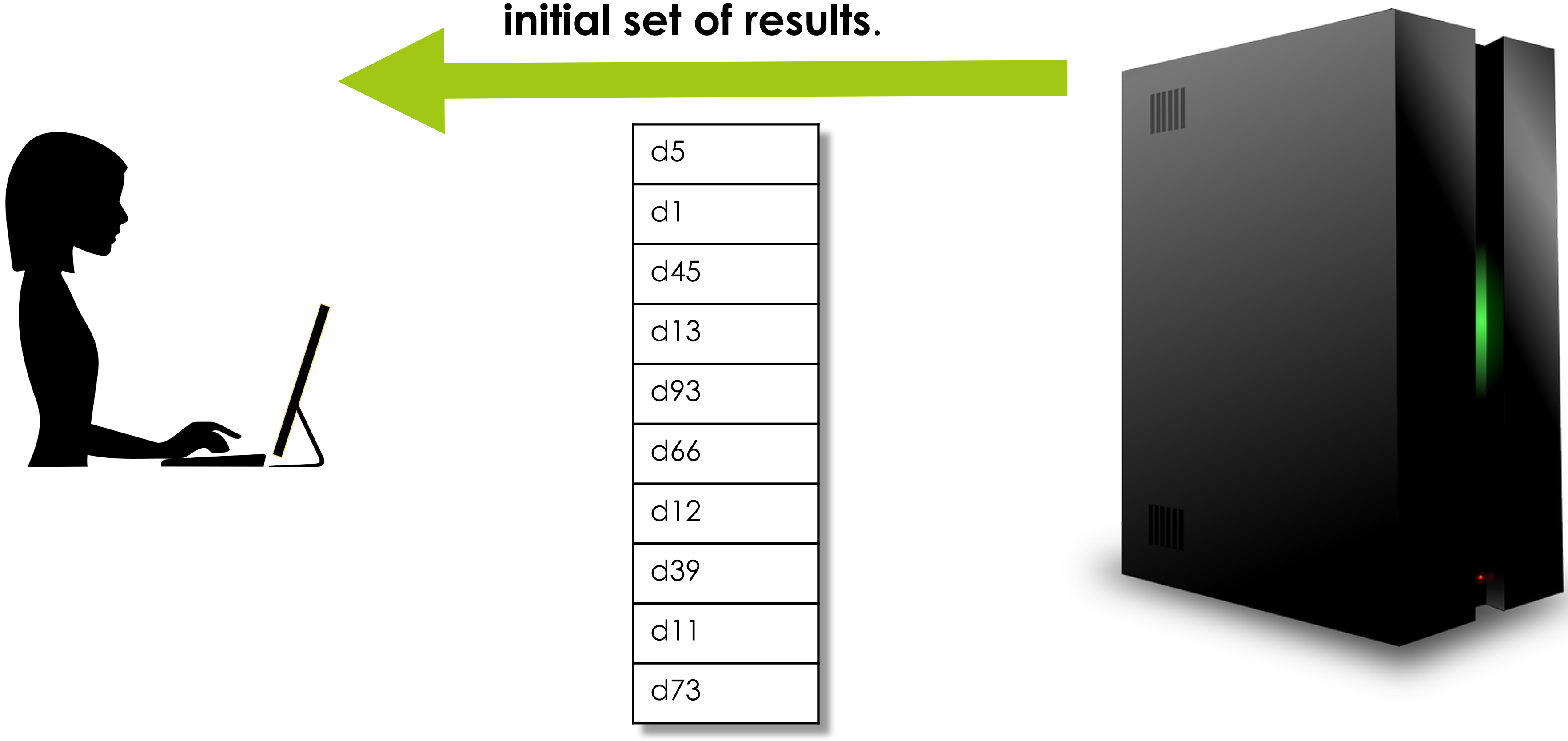
Probabilistic Model Process

1. User has an **information need.**

Probabilistic Model Process

2. User **sends a query** to the IR system.

Probabilistic Model Process



3. System sends back an **initial set of results.**

d5
d1
d45
d13
d93
d66
d12
d39
d11
d73

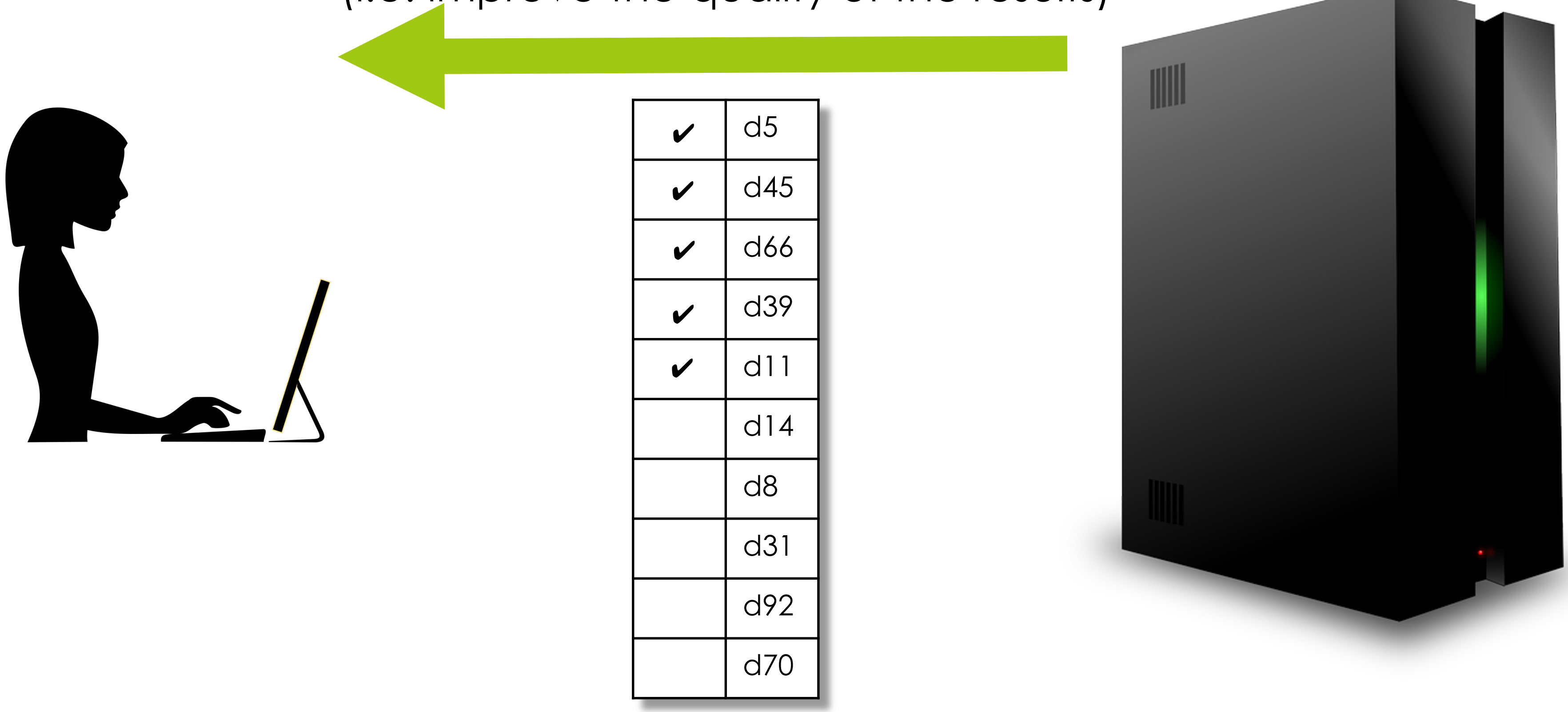
Probabilistic Model Process

4. User **tells** the system which of these results are **relevant**.

✓	d5
	d1
✓	d45
	d13
	d93
✓	d66
	d12
✓	d39
✓	d11
	d73

Probabilistic Model Process

5. System uses this information to **refine** the results
(i.e. improve the quality of the results)



✔	d5
✔	d45
✔	d66
✔	d39
✔	d11
	d14
	d8
	d31
	d92
	d70

Probabilistic Model Process

5. This process may be **repeated** many times.
Eventually, the user is satisfied with the results.

✓	d5
✓	d45
✓	d66
✓	d39
✓	d11
✓	d14
✓	d8
✓	d31
✓	d92
✓	d70

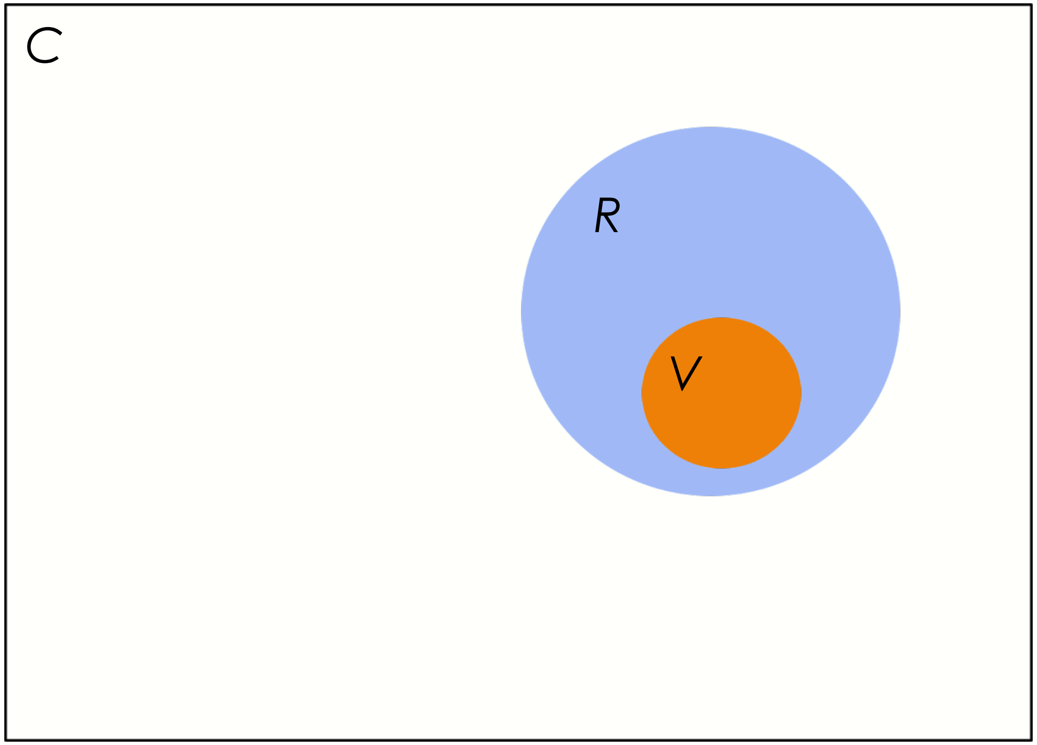
R: the ideal answer set

C: the **corpus** (i.e. the set of all documents in the index).

R: the **ideal answer set** (i.e. the set of documents that are relevant to the user's query).

V: the set of documents that the **user** has **verified** as being **relevant**.

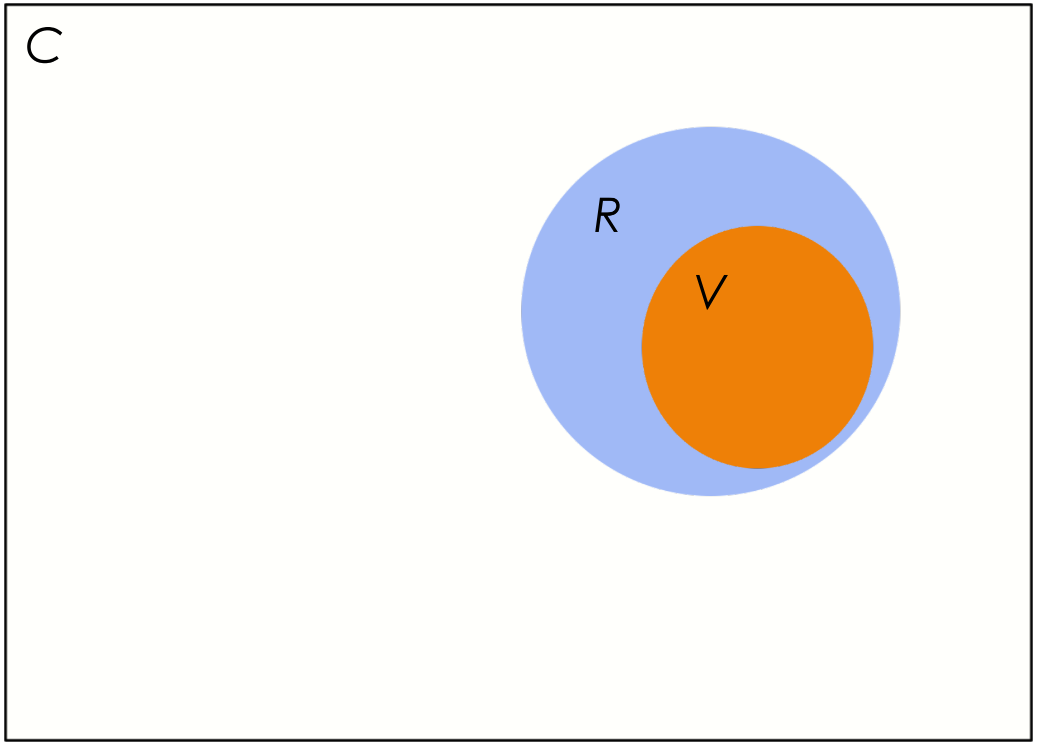
V must be a subset of *R* (user will only mark documents relevant that are actually relevant).



R: the ideal answer set

As more interaction occurs, the user marks more documents as relevant:

V grows to become closer to *R*.

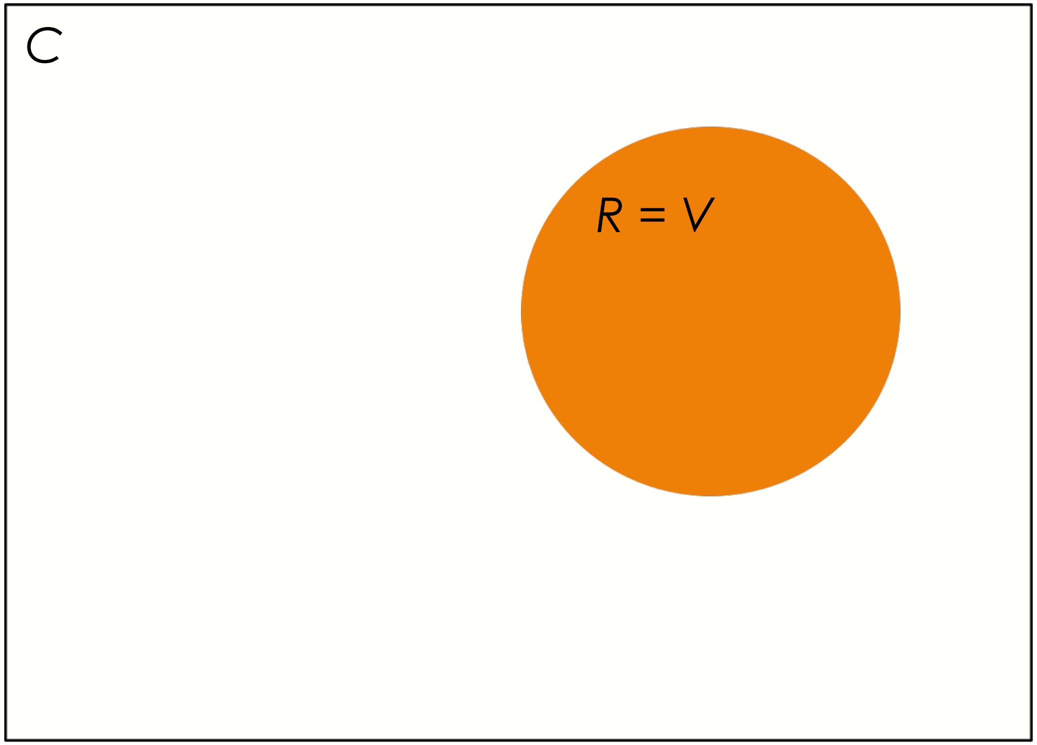


R: the ideal answer set

As more interaction occurs, the user marks more documents as relevant:

V grows to become closer to *R*.

Ideally, *V* and *R* will be equal at the end of the process.



Basics

- The system uses this user feedback to refine the description of the ideal answer set.
- By repeating this process many times, it is expected that the description will evolve and become closer to the real description of the ideal answer set.
- A conscious effort is made to use **probabilities** to create this description.