Introduction to Information Retrieval-Chapter 1

Dr. Subrat Kumar Nayak

Associate Professor

Dept. of CSE, ITER, SOADU

What is Information Retrieval?

Information retrieval(IR) is finding material(usually documents) of an unstructured nature(usually text) that satisfies an information need from within large collections(usually stored on computers).

- As defined in this way, information retrieval used to be an activity that only a **few people** engaged in: **reference librarians, paralegals, and similar professional searchers.**

- Now the world has changed, and **hundreds of millions of people engage in information retrieval every day** when they use a web search engine or search their email. Information retrieval is fast becoming the dominant form of information access, **overtaking traditional database style searching**

- These days we frequently think first of **web search**, but there are many other cases:
  - E-mail search
  - Searching your laptop
  - Corporate knowledge bases
  - Legal information retrieval

IR vs. databases: Structured vs unstructured data

- Structured data tends to refer to information in “tables”

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

- Typically allows numerical range and exact match

(for text) queries, e.g.,

Salary < 60000 AND Manager = Smith.

Unstructured Data

- Typically refers to free text
- Allows
  - Keyword queries including operators
  - More sophisticated “concept” queries e.g.,
    - find all web pages dealing with *drug abuse*
- Classic model for searching text documents

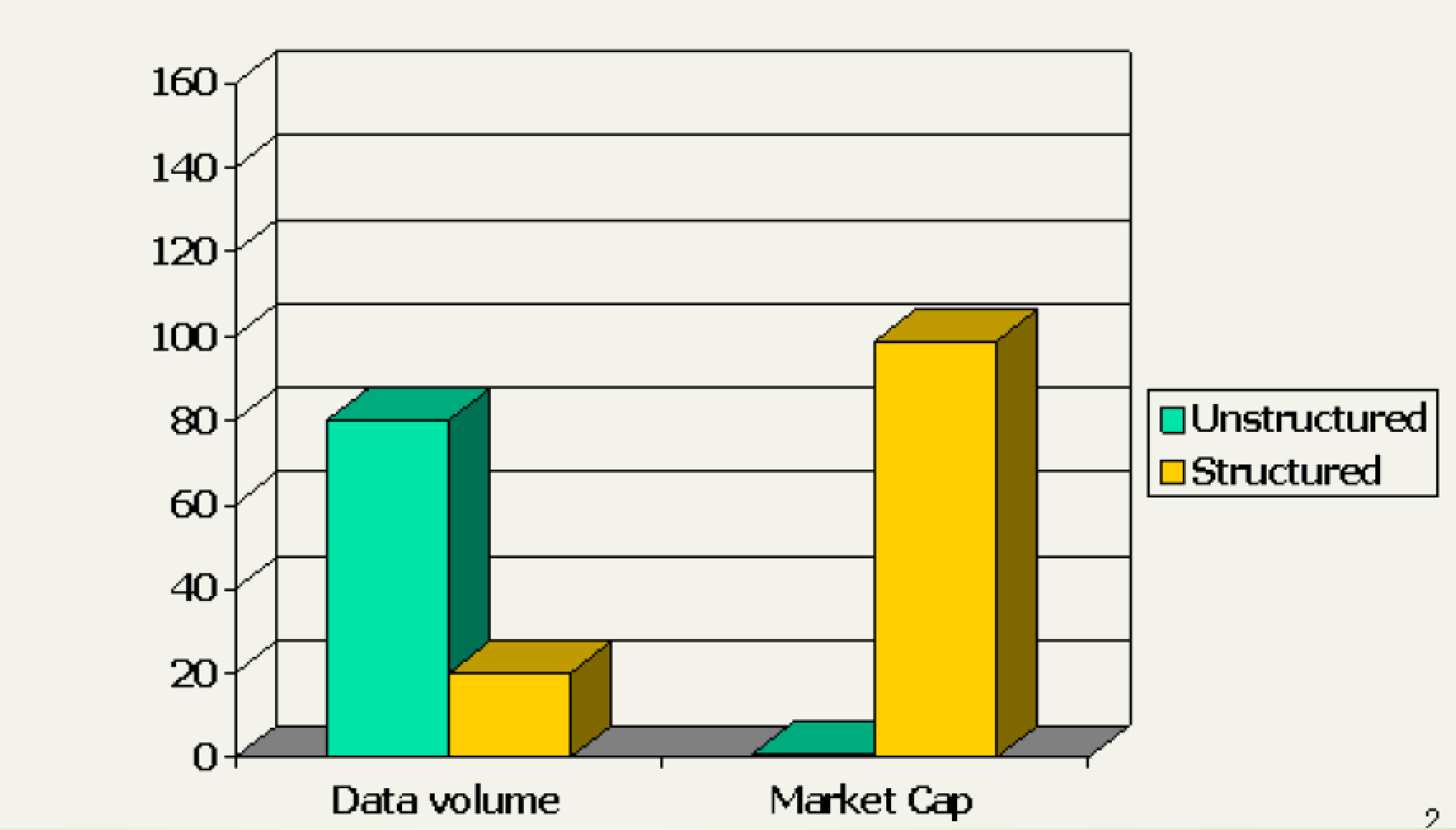
Semi-structured Data

- In reality, almost no data are truly “unstructured”.

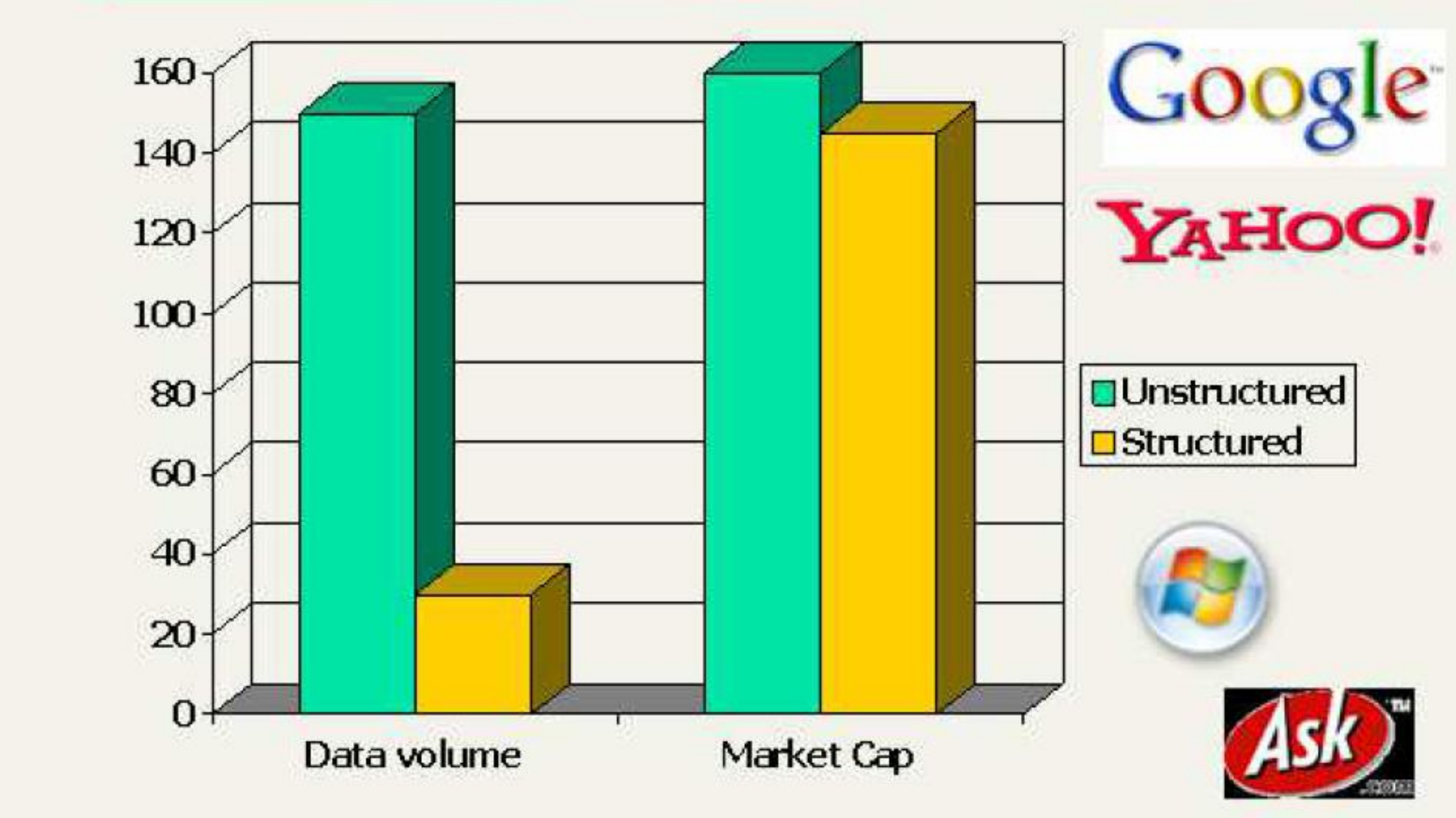
- Facilitates “semi-structured” search such as
  - ***Title*** contains <u>data</u> AND ***Bullets*** contain <u>search</u>

- So, IR is also used to facilitate “semi-structured” search such as finding a document where the **title contains Java** and **the body contains threading.**

Unstructured (text) vs structured (database) Data in 1996



Unstructured (text) vs Structured (database) Data Today



What is Document Collection?

- Document Collection: text units we have built an IR system over.
- Usually documents
- But could be
  - Memos
  - book chapters
  - paragraphs
  - scenes of a movie
  - Turns in a conversation...

Some Terminologies…

- **Information Need**

An ***information need*** is the topic about which the user desires to know more.

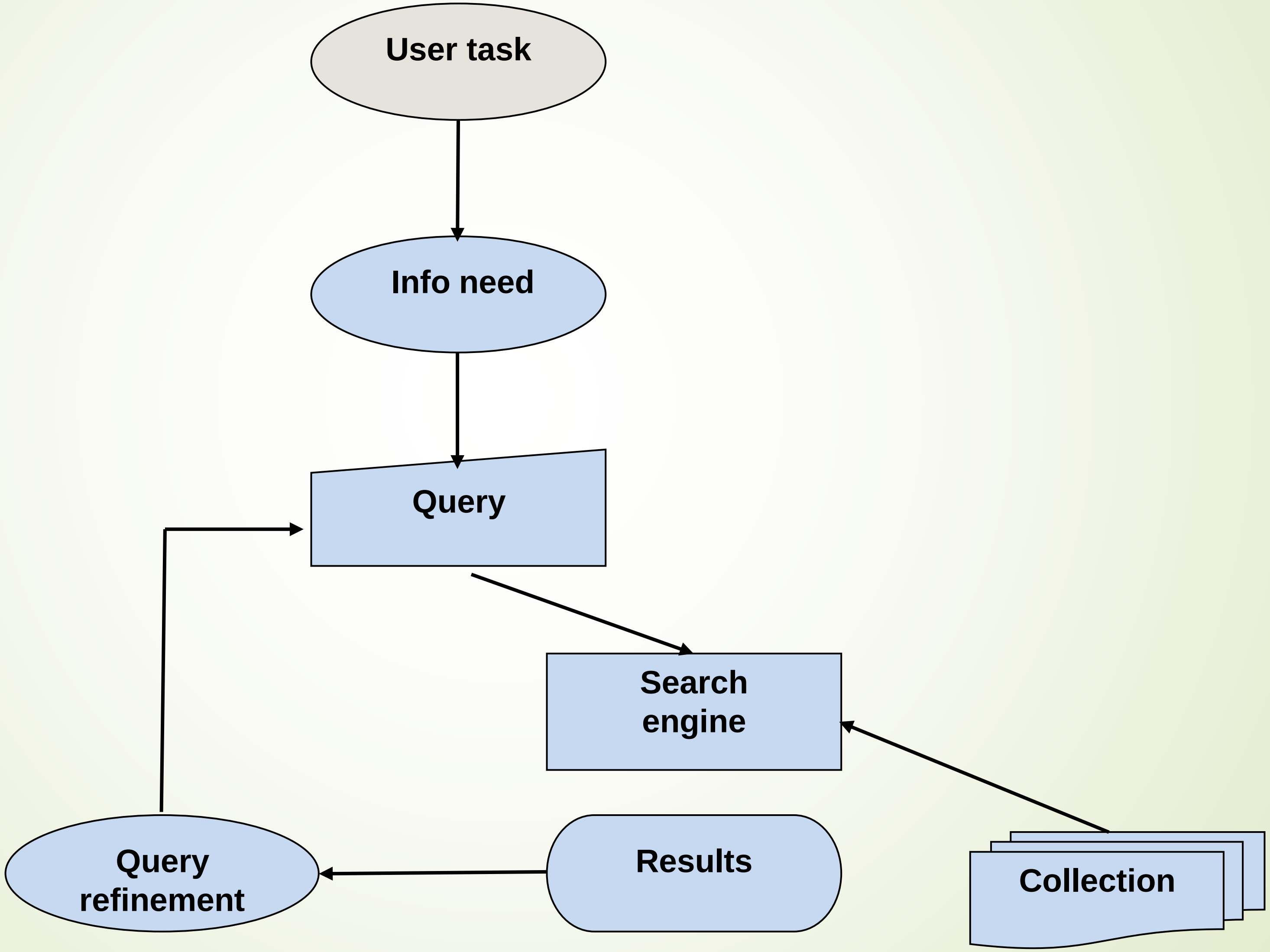
- **Query**

A query is what the user conveys to the computer in an attempt to communicate the information need.

- **Relevant document**

A document is ***relevant*** if it is one that the user perceives as containing information of value with respect to their personal information need.

The classic search model…



How well has the system performed?

- The ***effectiveness*** of an IR system (i.e., the quality of its search results) is determined by two key statistics about the system's returned results for a query:
  - ***Precision*:** What fraction of the returned results are relevant to the information need?

  i.e., Fraction of retrieved docs that are relevant to the user's information need.

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

  i.e., Fraction of relevant docs in collection that are retrieved

- What is the best balance between the two?
  - Easy to get perfect recall: just retrieve everything
  - Easy to get good precision: retrieve only the most relevant

Basic Assumption of Information Retrieval?

Collection: A set of documents

- Assume it is a **static collection** for the moment

Goal: Retrieve documents with information that is relevant to the user's information need and helps the user complete a task.

IR vs DBMS

- One of the most distinguish features of **DBMSs** is that they offer an advance Data Modelling Facility(DMF) including Data Definition Language and Data Manipulation Language for modelling and manipulating data.
  - **IRSs** do not offer an advance DMF. Usually data modelling in IRs is restricted to classification of objects.
- A major strength of the Data Definition Language of **DBMSs** is the capability to define the data integrity constraints.
  - In **IRSs** such validation mechanisms are less developed.
- **DBMSs** provide precise semantics.
  - **IRSs** most of the time provides imprecise semantics.
- **DBMSs** has structured data format.
  - Whereas, **IRSs** is characterized by unstructured data format.
- Query specification is complete in **DBMS**.
  - In **IRS** query specification is incomplete.
- Query language is artificial in **DBMS**.
  - In **IRS** query language is near to natural language.

IR vs Data Mining

- **Information Retrieval** is the process of organising data(usually textual data) and building algorithms so people can write queries to retrieve the data they want. Think of Google.
- **Data Mining** is the process of discovering patterns in data.
- Let’s say you have a shop, data about your customers, their previous purchases, and want to predict which one of them will purchase the brand new product you are about to release next month. This process can be carried via Machine Learning, Statistics, or just via simple Database Queries. In that sense, Information Retrieval can also be considered as a subset of Machine Learning.

Final remarks on Information Retrieval?

- IR is related to many areas:
  - ❑ NLP, AI, database, machine learning, user modeling…
  - ❑ library, Web, multi media search,…
- Relatively **week theories**
- Very strong tradition of experiments
- Many remaining (and exciting) problems
- **Difficult area**: Intuitive methods do not necessarily improve effectiveness in practice