Presented by

Dr. S.P. Siddique Ibrahim /Assistant Professor

Outline

- Introduction
- What is data mining?

Data Measures

Bit
Byte
KB
MB
GB
TB
PB
EB
ZB
YB

Units of storage in computers

YouTube Video Quality	Data Used per Hour	Data Used per Day
480p	562.5MB	13.5GB
720p	1.86GB	44.64GB
1080p	3.04GB	72.96GB
4k	15.98GB	383.52GB

Spotify

Default Spotify settings use **2MB+** per 3-minute song.

That's **40MB** every hour. Or **960MB** per day.

Netflix

Each standard definition Netflix stream uses **1GB** of data per hour (**24GB** per day).

High definition Netflix streams can use as much as **3GB** of data each hour (**72GB** per day).

And ultra HD uses **7GB** per hour (**168GB** per day.)

Sources: Sandvine, Domo, TechJury, iNews

Data Mining

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Valid: The patterns hold in general.

Novel: We did not know the pattern beforehand.

Useful: We can devise actions from the patterns.

Understandable: We can interpret and comprehend the patterns.

- Data Mining is:
- (1) The efficient discovery of previously unknown, valid, potentially useful, understandable patterns in large datasets.

- (2) The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner

- Data Mining is:
- (3) Data mining is automated (or) convenient extraction of patterns representing knowledge implicitly stored in large databases, data warehouses and other massive information repositories.

Real definition of data mining

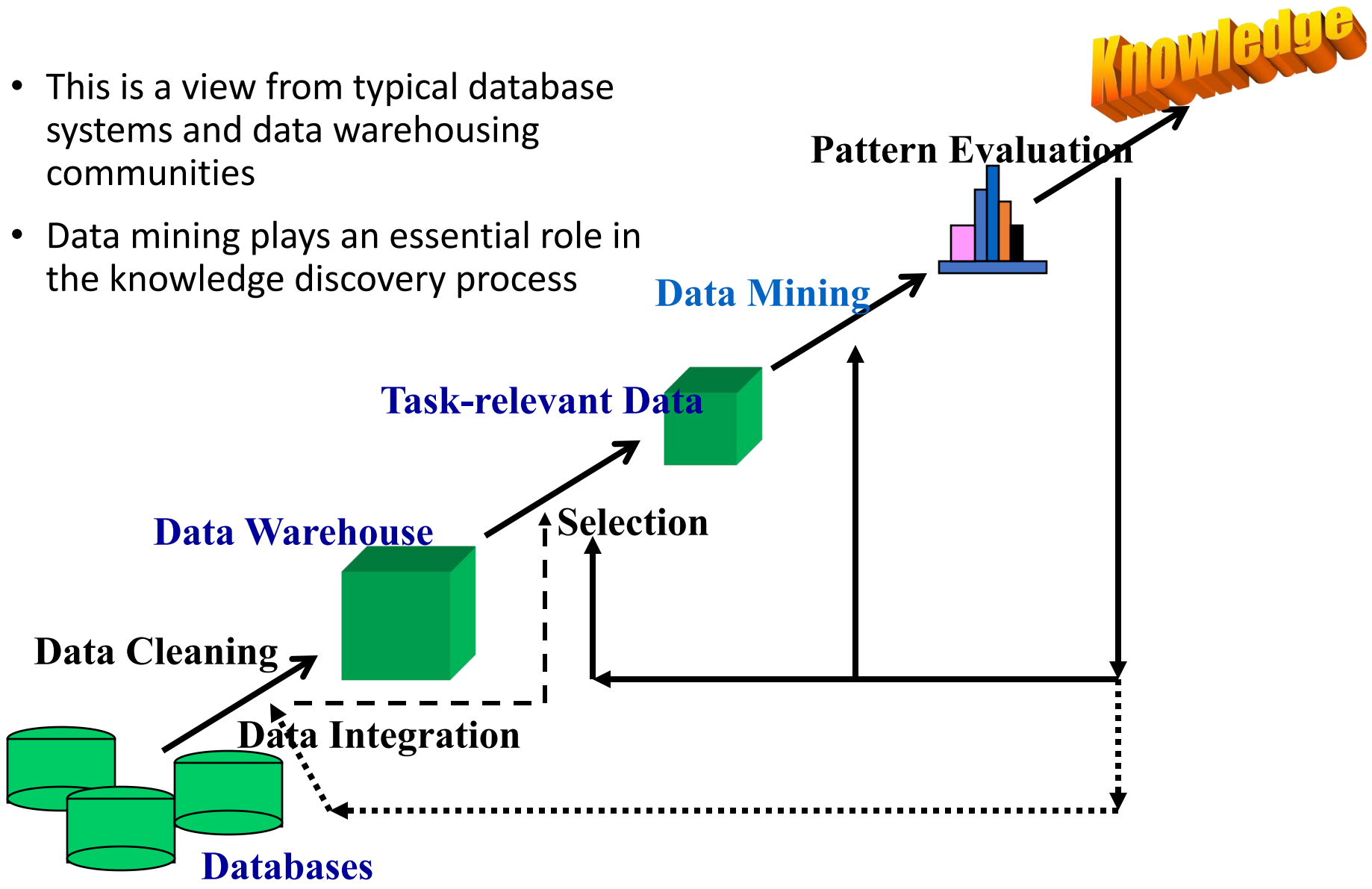
- Name is misnomer
- Similar to Gold mining
- Knowledge mining from data “Knowledge mining”
- Knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaelogy, and data dredging

What is data mining?

- Essential step in the process of knoledge discovery in databases.

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



- Data Mining is a multidisciplinary field, drawing work from areas including
- database technology,
- AI,
- ML,
- Neural networks,
- statistics,
- pattern recognition,
- knowledge based systems,
- knowledge acquisition,
- information retrieval,
- high-performance computing, and
- data visualization.

Why Data Mining?

- The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and immnent need for turning such data into useful information and knowledge.

- Field such as business management, production control, and market analysis to engineering design and science exploration.

- It can be viewed as natural evolution of information technology.

- An evolutionary path has been witnessed in the database industry in the development of the following functionalities

- Data collection
- database creation
- data management(including data storage,retrieval and data transaction processing)
- data analysis
- understanding data warehouse and
- data mining

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s—2000s:
  - Data mining and data warehousing, multimedia databases, and Web databases

Data Mining Vs Database

- **Data Mining:**
- It can be defined as finding hidden information in a database. Also, it refers to extracting or mining knowledge from large amount of data.
- **Database:**
- Database is known as set of data's or the data's are stored in a structured manner using the Query.
  eg: database can be assume like a shelf. inside the shelf there are many folders for assigning each items. the particular item will go towards the particular shelf. this is the basic concept

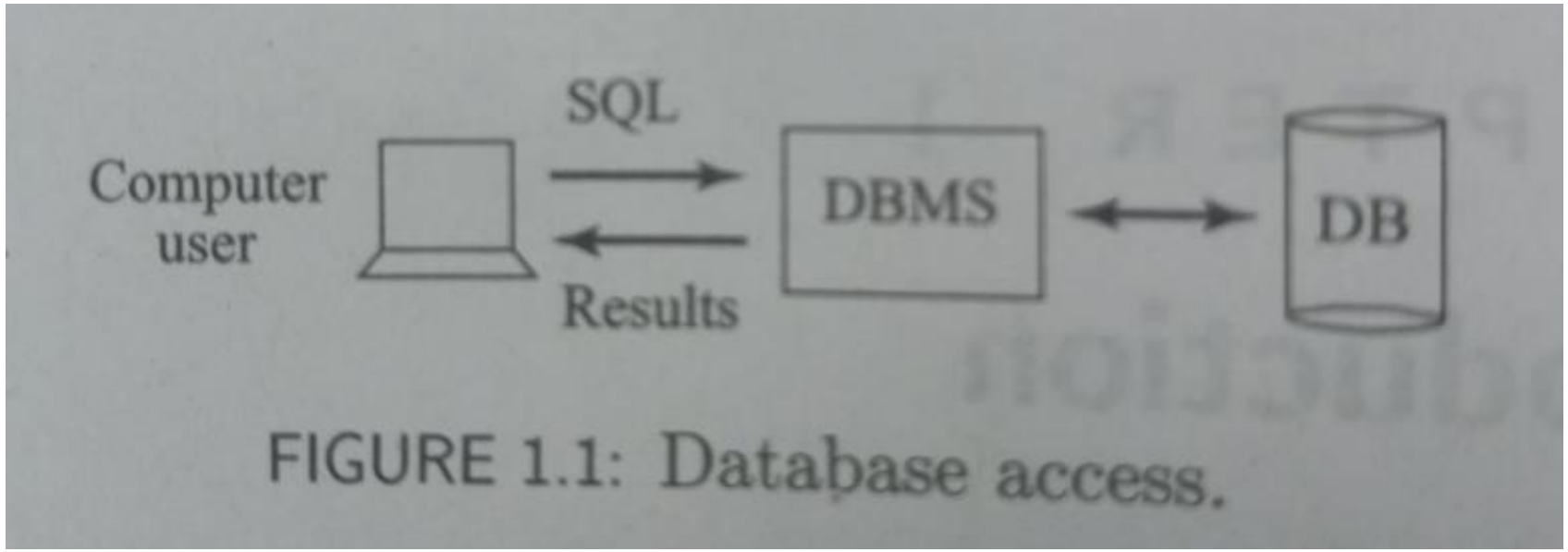


FIGURE 1.1: Database access.

Introduction

- Data is growing at a phenomenal rate
- Users expect more sophisticated information
- How?
- Simple query not enough!!!!

UNCOVER HIDDEN INFORMATION

DATA MINING

Data Mining Definition

- Finding hidden information in a database
- Fit data to a model
- Similar terms
  - Exploratory data analysis
  - Data driven discovery
  - Deductive learning

Example

- Credit Card companies must determine whether to authorize credit card purchases.
- Suppose that based on past historical information about purchases, each purchase is placed into one of four classes

- Authorize
- Ask for further identification before authorization
- Do not authorize
- Do not authorize but contact police

Data Mining Task

- First, the historical data must be examined to determine how the data fit into the four catogories.

- Secondly, the problem is to apply this model to each new purchase.

Data Mining Task contd.,

- Eventually, the second part indeed may be stated as a simple database query, the first part cannot be.....

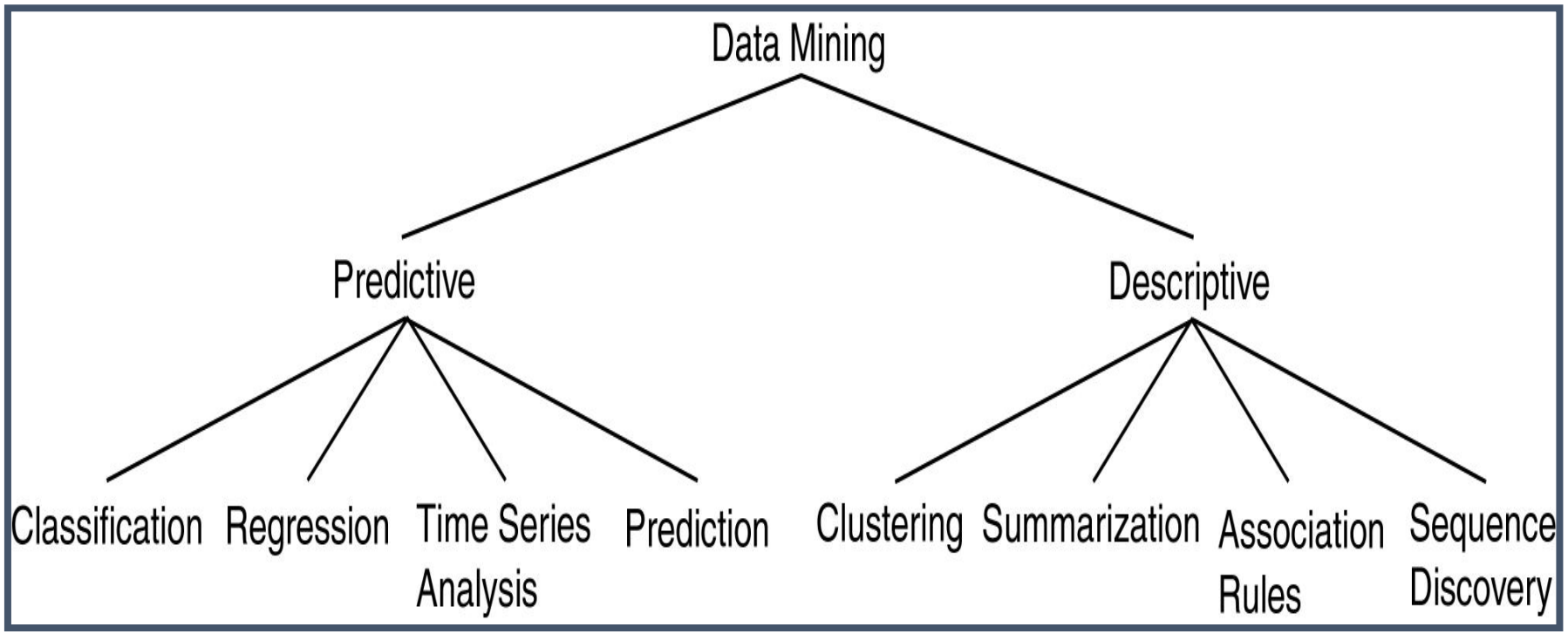
Query Examples

- **Database**
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
- **Data Mining**
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (association rules)

Data Mining Algorithm

- Objective: Fit Data to a Model
  - Descriptive
  - Predictive
- Preference – Technique to choose the best model
- Search – Technique to search the data
  - “Query”

Data Mining Models and Tasks



Summary

- Why data mining?
- Who contributed for large data generation?
- Difference between data mining and database