Data Mining & Data Warehouse

Associate Professor Dr. Raed Ibraheem Hamed

University of Human Development,
College of Science and Technology
Department of Information Technology

2016 – 2017

Introduction

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data Mining and Business Intelligence
- Data, Information, Knowledge and Decision Making Action Cycle
- The Primary Tasks of Data Mining
- Basic Data Mining Tasks

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of multiple data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but the strongest need for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

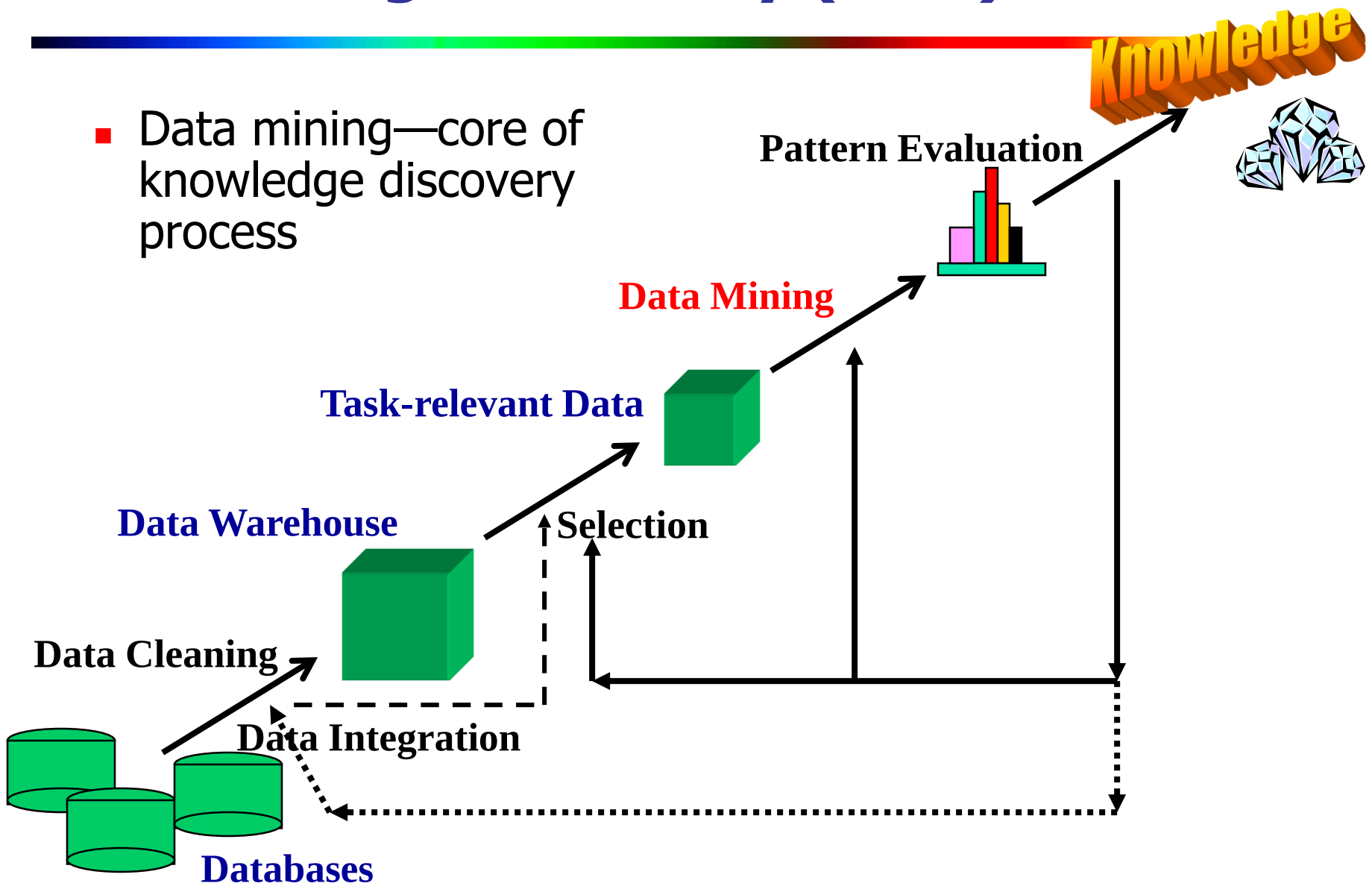
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (**non-trivial**, **implicit**, **previously unknown** and **potentially useful**) patterns or knowledge from huge amount of data.

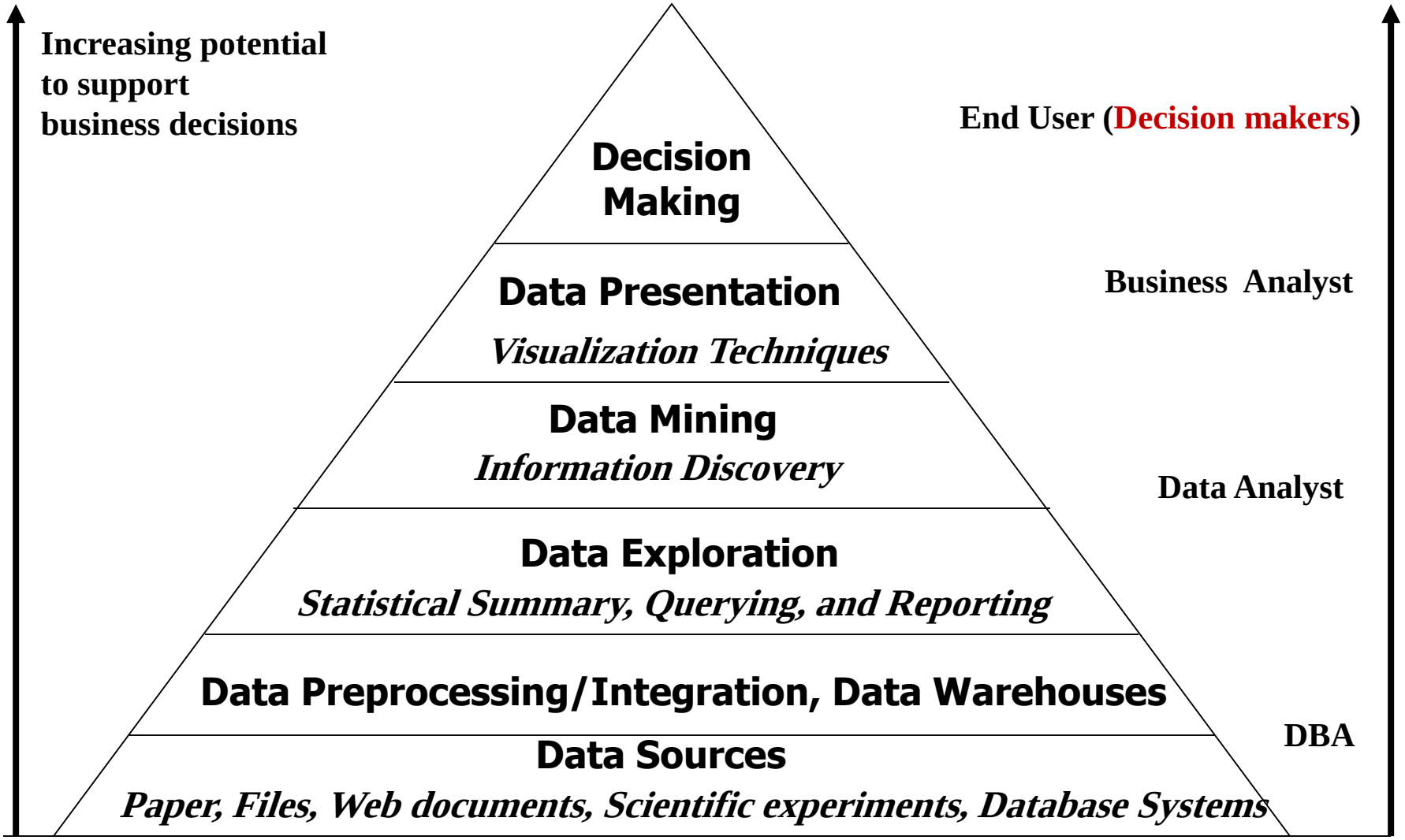
- **Alternative names**
  - Knowledge discovery in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, information gathering, business intelligence, etc.

Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process



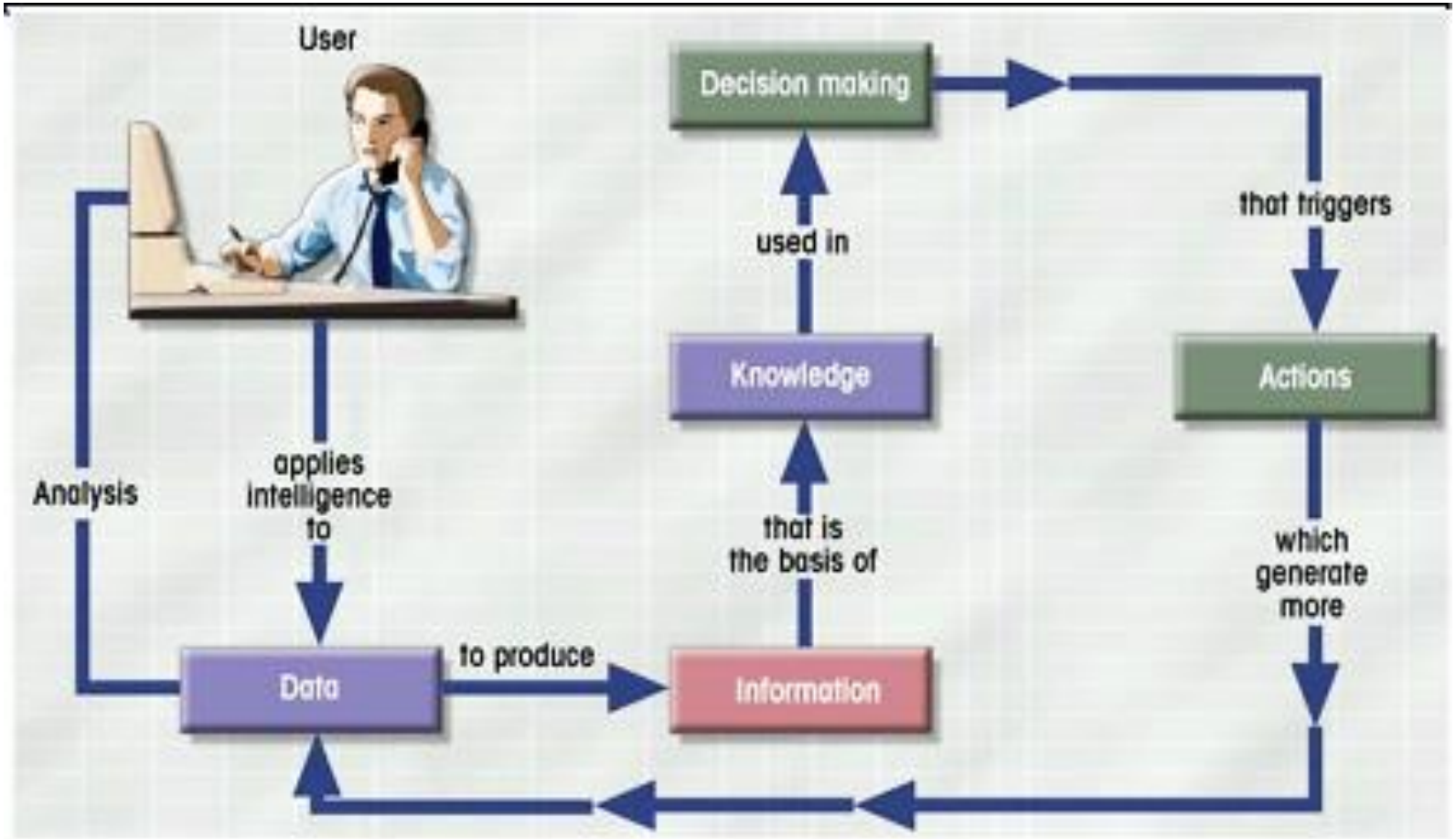
Data Mining and Business Intelligence



Data Mining and Business Intelligence

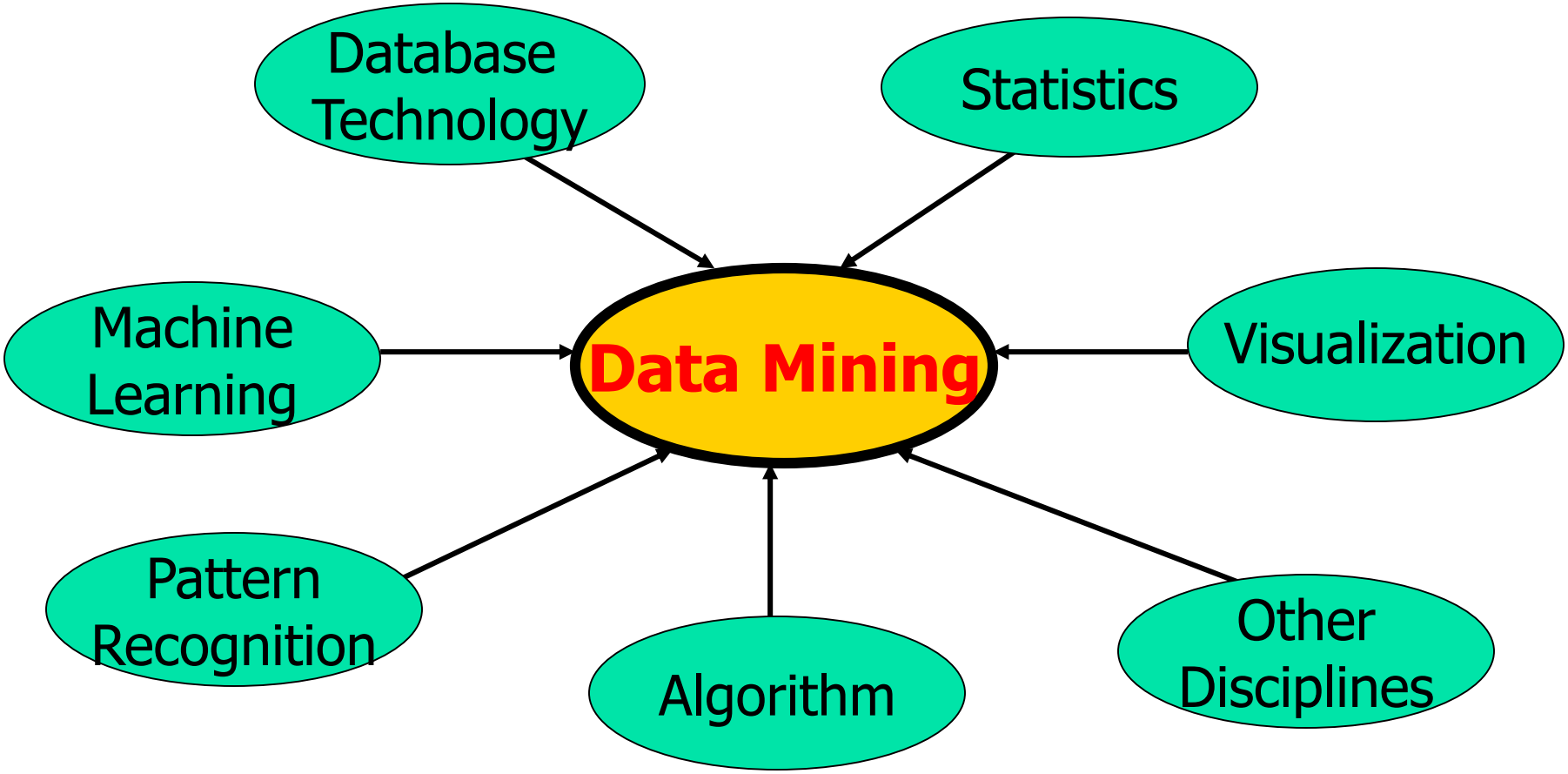
- ✓ **DBA**: Gather the data from different sources like paper based data, transactional data, other form of data. DBA put different form of data in to same format and build a **Data warehouse**.
- ✓ **Data Analyst**: Once Data is ready, job role for data analyst is to understand the requirement from business users (decision makers)
- ✓ **Business Analyst**: work on different tools to create a sophisticated reports and graphs which gives the end user flexibility to view the key processes within the organization.
- ✓ **End user**: Are also known as a **Decision makers** who view these ports and take the decisions about the business.

Data, Information, Knowledge and Decision Making Action Cycle



Data Mining: Confluence of Multiple Disciplines

A data mining system can be classified according to the following criteria



Why Not Traditional Data Analysis?

- Huge amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and inheritance databases
  - Multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Relational, data warehouse, transactional, stream, active, time-series, text, multi-media, heterogeneous, WWW
- **Knowledge to be mined**
  - Characterization, association, classification, clustering, deviation, outlier analysis, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining: Classification Schemes

Different views lead to different classifications:

1. Data view: Kinds of data to be mined
2. Knowledge view: Kinds of knowledge to be discovered
3. Method view: Kinds of techniques utilized
4. Application view: Kinds of applications adapted

Data Mining: On What Kinds of Data?

- **Advanced data sets and advanced applications**
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Relational databases
  - Heterogeneous databases and historical databases
  - Multimedia database
  - Text databases
  - The World-Wide Web

The Primary Tasks of Data Mining

The two "high-level" primary goals of data mining, in practice, are *prediction* and *description*.

1. **Prediction:** involves using some variables or fields in the database to predict unknown or future values of other variables of interest.

2. **Description:** focuses on finding human-interpretable patterns describing the data.

Predictive Model

- Involves using some variables or fields in the database to predict unknown or future values of other variables of interest.
- Or based on the use of other historical data.
- **Example :-**
  1. Credit card fraud
  2. Breast cancer early warning
  3. Terrorist act

Descriptive Model

- Focuses on finding human-interpretable patterns describing the data.
- Serves as a way to explore the properties of data examined, not to predict new properties.
- Always required a domain expert.

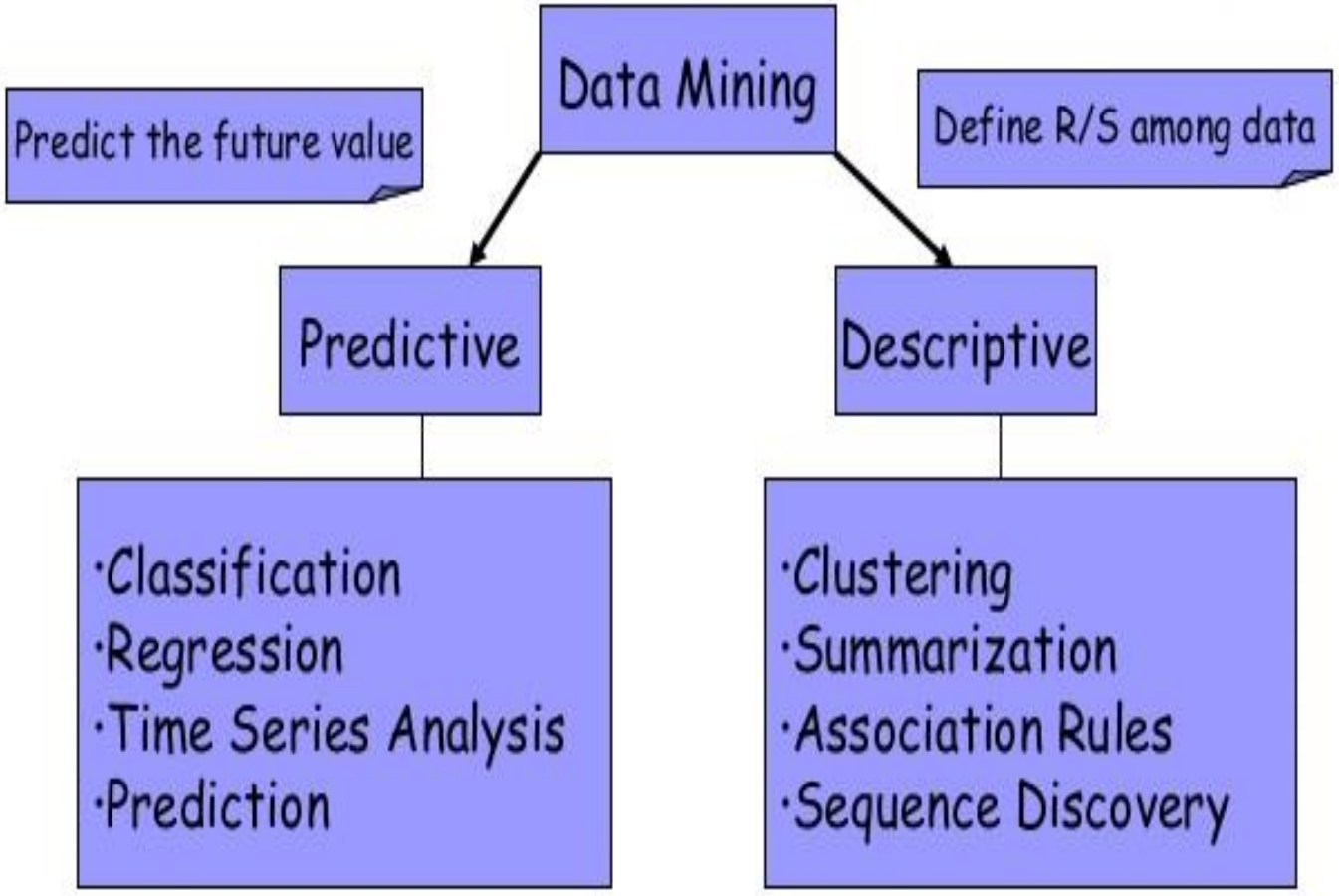
- **Example:**
  1. Segmenting marking area
  2. Profiling student performances
  3. Profiling GooglePlay/ AppleApps customer

Basic Data Mining Tasks

The goals of prediction and description are achieved by using the following primary data mining tasks:

1) Classification
2) Pattern Regression
3) Time serious analysis
4) Prediction
5) Clustering
6) Association rules
7) Summarization
8) Sequence discovery

Basic Data Mining Tasks



Applications of Data Mining

- E-commerce
- Marketing and retail
- Finance
- Telecoms
- Drug design
- Process control
- Space and earth sensing
- Bioinformatics
- Etc.

Requirements for a Data Mining System

- Data mining systems should be:

1) Computationally sound
2) Statistically sound
3) Ergonomically sound

UHD
UNIVERSITY OF HUMAN DEVELOPMENT