



# **Data Mining: Concepts and Techniques**

## **— Introduction —**

# Chapter 1. Introduction

---

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality

# 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 abundant data
    - Business: Web, e-commerce, transactions, stocks, ...
    - Science: Remote sensing, bioinformatics, scientific simulation, ...
    - Society and everyone: news, digital cameras, YouTube
- Applications: Market analysis, fraud detection, customer retention etc.

# Evolution of Database Technology

- Before 1960, **data collection & DB creation**
  - Primitive file processing
- 1970-1980s, **database management system**
  - Hierarchical, relational, n/w, sql, query processing & optimization, OLTP
- mid 1980s, **advanced data base systems**
  - Extended relational, object relational, object oriented
  - Advanced applications:multimedia, active stream & sensor etc.
- late 1980s, **advanced data analysis**
  - Data warehouse, OLAP, data mining
- 1990s, **web based databases**
  - XML based DBS, information integration, information retrieval
- 2000s
  - Stream data management and mining, Data mining and its applications, Web technology (XML, data integration) and global information systems



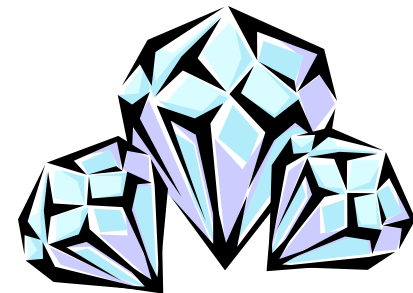
*data rich but information  
poor!*

*Requires Powerful tool*

# 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
  - Data mining: a misnomer? (gold mining, not sand/rock mining)
  - Knowledge mining from data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

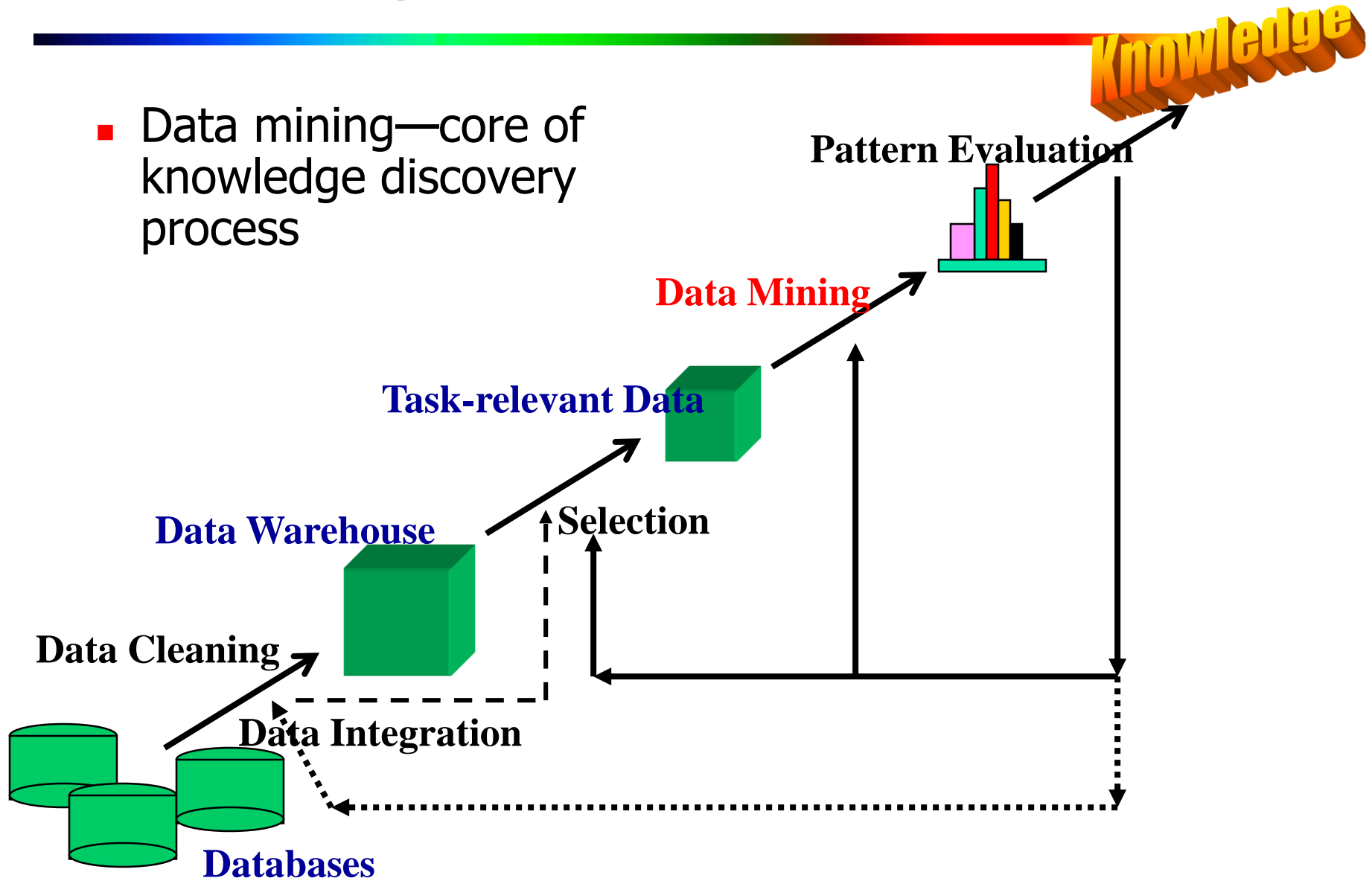


# What Is Data Mining?



# Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process





# Knowledge Discovery (KDD) Process

---

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)<sup>1</sup>
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations)<sup>2</sup>
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

# Why Not Traditional Data Analysis?

---

- Tremendous 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 , graphs, social networks and multi-linked data  
Heterogeneous databases , scientific simulations
- New and sophisticated applications

# Data Mining: On What Kinds of Data?

---

## 1. Database-oriented data sets

### ■ Relational database - ***AllElectronics***

- *Customer* (*cust ID*, customer name, address, age, occupation, annual income, credit information, category)
- Similarly, each of the relations *item*, *employee*, and *branch* consists of a set of attributes describing their properties.
- Tables can also be used to represent the relationships between or among multiple relation tables. For this example, these include purchases (customer purchases items, creating a sales transaction that is handled by an employee), items sold (lists the items sold in a given transaction), and works at (employee works at a branch of AllElectronics).

# Queries



- Q1: Show list of all items that were sold in the last quarter
- Q2: Show the total sales of the last month, grouped by branch
- Q3: How many sales transactions occurred in the month of December
- Q4: Which sales person had the highest amount of sales?
- data mining searches for trends & patterns - DM can analyze customer data to predict the credit risk of new customers based on their income, age, and previous credit information

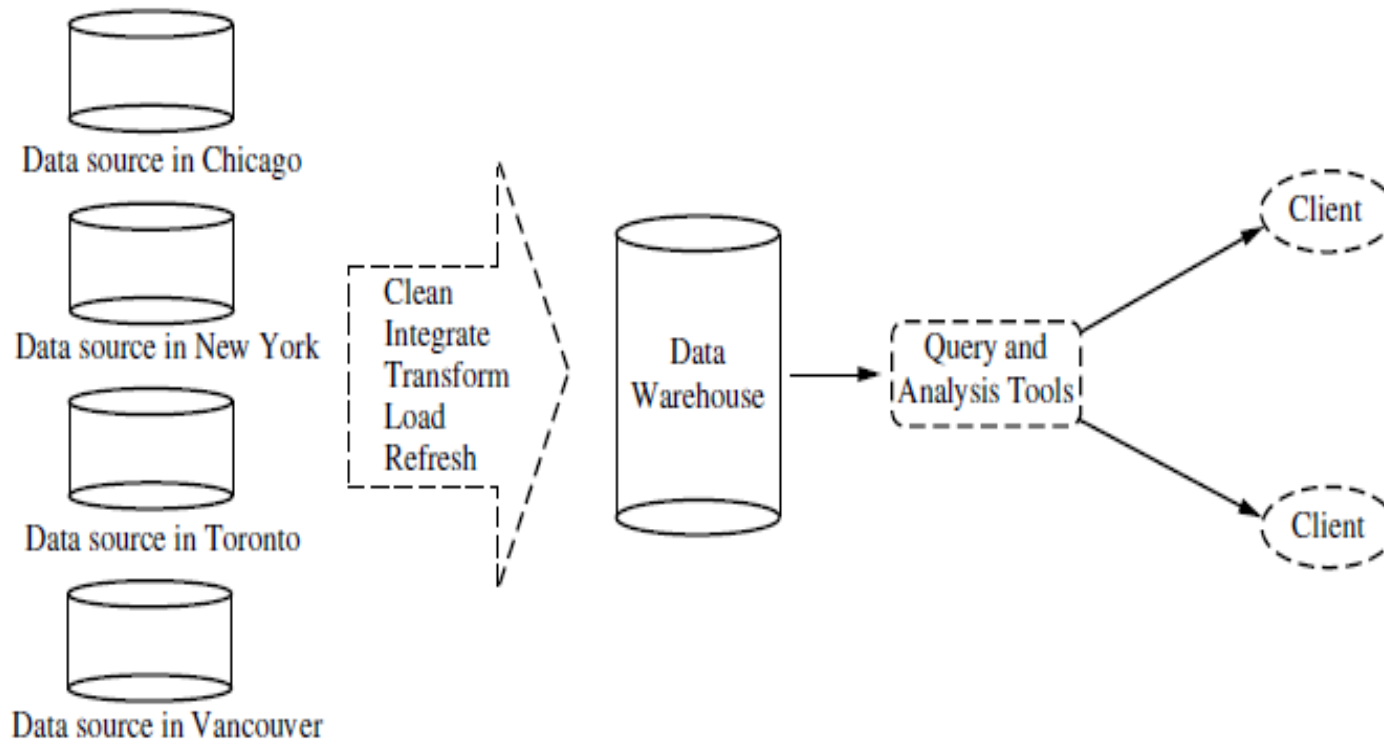
# Data Mining: On What Kinds of Data?



## 2. data warehouse

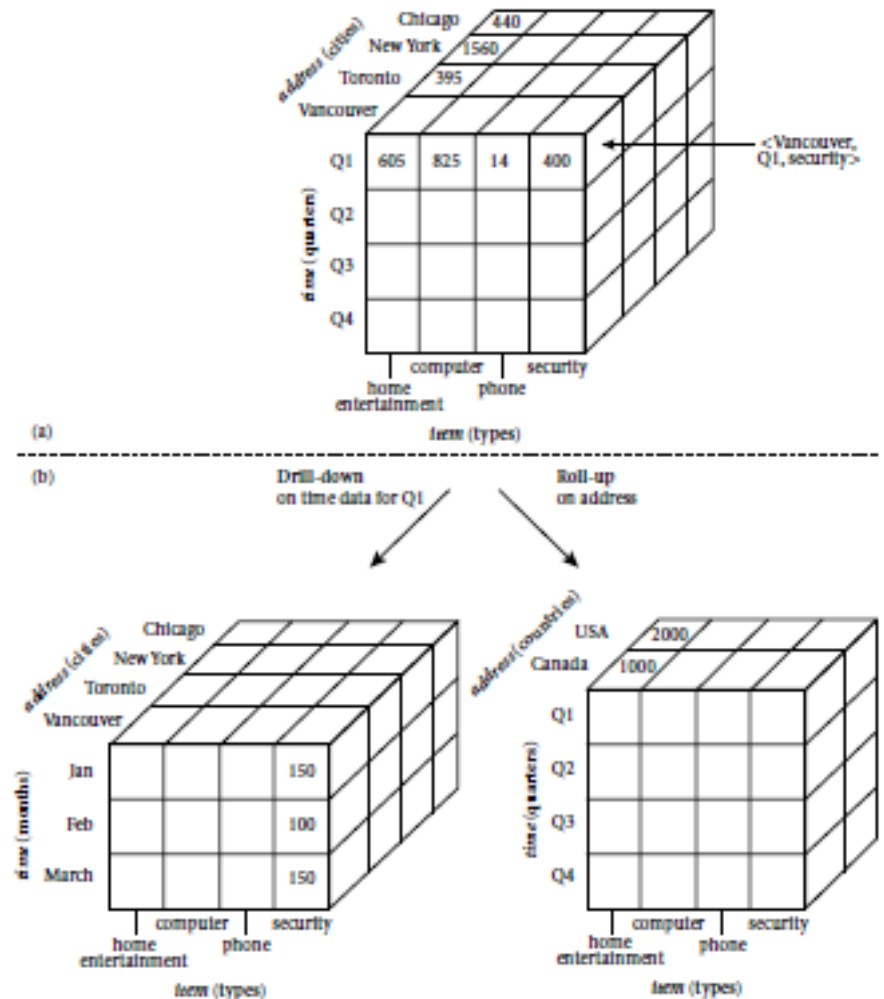
- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site
- A data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process
  
- Q: analysis of the company's sales per item type per branch for the third quarter
- difficult task - since the relevant data are spread out over several databases, physically located at numerous sites

# Data Mining: On What Kinds of Data?



A data warehouse collects information about subjects that span an *entire organization*, and thus its scope is *enterprise-wide*.

A data mart, is a department subset of a data warehouse. It focuses on selected subjects, and thus its scope is *department-wide*.



# Data Mining: On What Kinds of Data?

## 3. transactional data bases

- Q: Show all the items purchased by Sandy Smith” or “How many transactions include item number I3?”
  - may require a scan of the entire transactional database.
- Q: Which items sold well together?”
  - *market basket data analysis*
  - *Identify frequent item sets: computer & printer*

<i>trans_ID</i>	<i>list of item_IDs</i>
T100	I1, I3, I8, I16
T200	I2, I8
...	...



# Data Mining: On What Kinds of Data?



## 4. Advanced data sets and advanced applications

- Data streams and sensor data
- Time-series data (values obtained over repeated measurement of time : temperature), temporal data (time related attribute; eg.stock exchange), sequence data (incl. bio-sequences)
- Spatial data and spatiotemporal data (spatial obj change with time)
- Text databases (unstructured / semi structured / structured)
- Multimedia database (image, audio, video)
- Heterogeneous databases and legacy databases
- Structure data, graphs, social networks and multi-linked data
- Object-relational databases
- The World-Wide Web

# Data Mining Functionalities-What kinds of patterns may be mined?

---

- 2 categories
  - Descriptive
    - Descriptive mining tasks characterize the general properties of the data in the database
  - Predictive
    - Predictive mining tasks perform inference on the current data in order to make predictions.

# Data Mining Functionalities

## 1. Concept/class description: Characterization and discrimination

- Data can be associated with classes or concepts
- *AllElectronics* store –
  - classes of items for sale - *computers* and *printers*
  - concepts of customers - *bigSpenders* and *budgetSpenders*
- *Descriptors derived from*
  - a. *Data characterization* - summarization of the general characteristics or features of a target class of data.

Eg. summarize the characteristics of customers who spend more than \$1,000 a year at *AllElectronics*.

result - a general profile of the customers, such as they are 40–50 years old, employed, and have excellent credit ratings
  - *methods*
    - *Statistical measures & plots , OLAP roll-up operation, Attribute oriented induction technique*

# Data Mining Functionalities



*b. Data discrimination* - comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

Eg. compare two groups of *AllElectronics* customers, such as those who shop for computer products regularly (more than two times a month) versus those who rarely shop for such products (i.e., less than three times a year).

Result:

80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education

60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree

*c. Both characterization and discrimination*

# Data Mining Functionalities

---

## 2. Mining frequent patterns, association, correlation

Frequent patterns - patterns that occur frequently in data

3 types of frequent patterns:

a. *frequent itemset* - a set of items that frequently appear together in a transactional data set, such as bread and butter

b. frequent subsequence - the pattern that customers tend to purchase. first a PC, followed by a digital camera, and then a memory card, is a (*frequent*) *sequential pattern*

c. *frequent substructure* - different structural forms, such as graphs, trees, or lattices

- Mining frequent patterns leads to the discovery of interesting associations and correlations within data.


# Data Mining Functionalities

Association and correlation analysis:

- determine items that are frequently purchased together within the same transactions.

$buys(X, \text{"computer"}) \Rightarrow buys(X, \text{"software"})$  [ $support = 1\%$ ,  $confidence = 50\%$ ]  
 $computer \Rightarrow software$  [ $1\%$ ,  $50\%$ ]

- $age(cust, \text{"20:::29"}) \wedge income(cust, \text{20K:::29K}) \Rightarrow buys(cust, \text{"CD player"})$  [ $support = 2\%$ ,  $confidence = 60\%$ ]
- association rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold



$$\text{support} = \frac{(X \cup Y).count}{n}$$

$$\text{confidence} = \frac{(X \cup Y).count}{X.count}$$

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

Association rules:

$A \rightarrow D$  (60%, 100%)

$D \rightarrow A$  (60%, 75%)

# Data Mining Functionalities

---

## 3. Classification and Prediction

- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts
- Classification maps data into predefined groups or classes.
- supervised learning - the classes are determined before examining the data.
- purpose - use the model to predict the class of objects whose class label is unknown.
- Methods - *classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks*

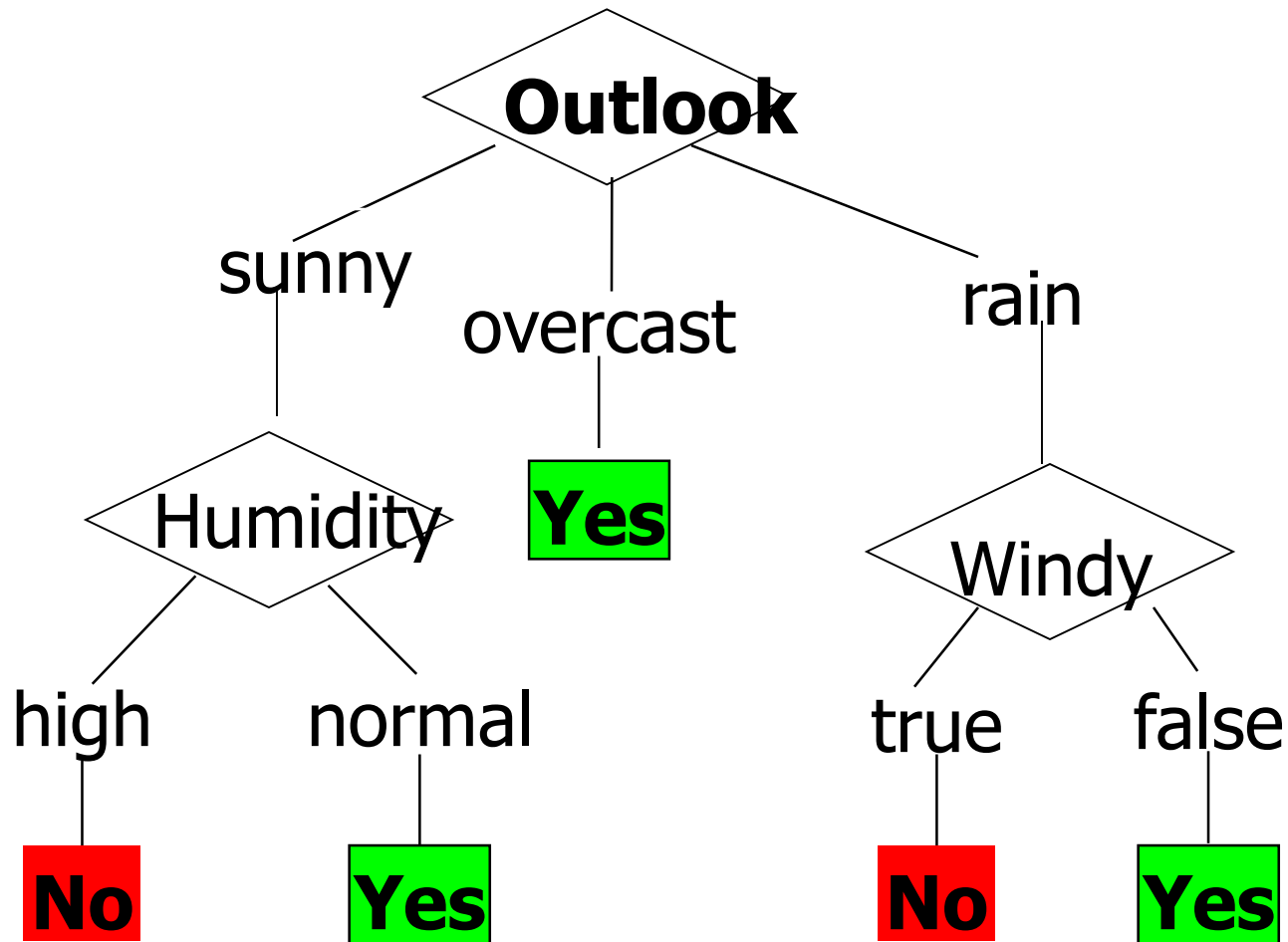


# Weather Data: Play or not Play?

---

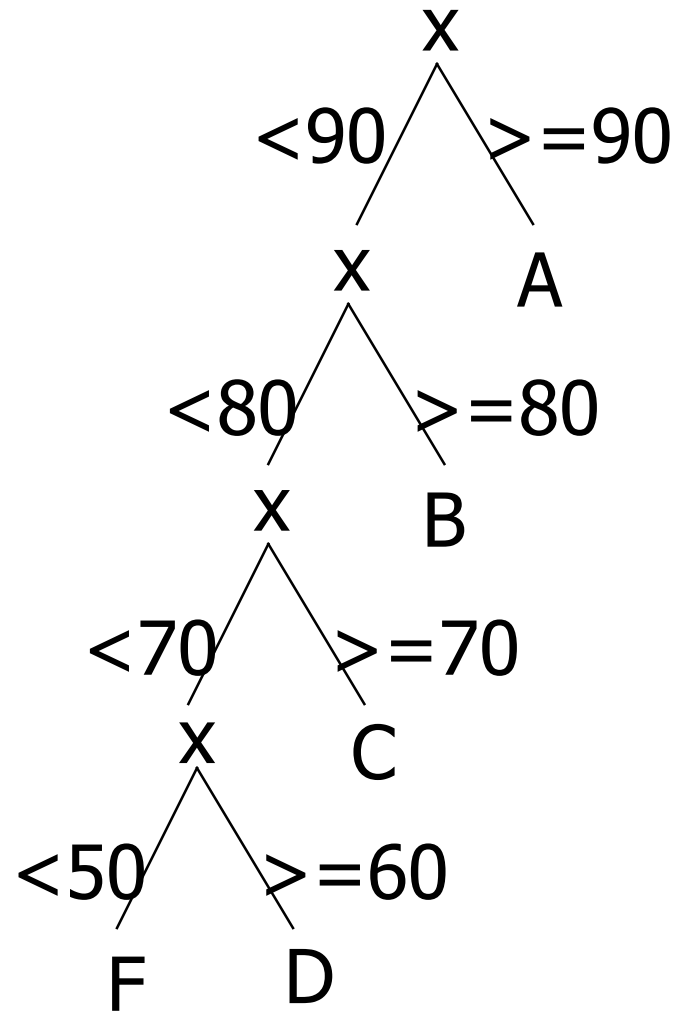
Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

# Example Tree for “Play?”



# Classification Ex: Grading

- If  $x \geq 90$  then grade = A.
- If  $80 \leq x < 90$  then grade = B.
- If  $70 \leq x < 80$  then grade = C.
- If  $60 \leq x < 70$  then grade = D.
- If  $x < 60$  then grade = F.



# Data Mining Functionalities

---

- Prediction - models continuous-valued functions. used to predict missing or unavailable *numerical data values* rather than class labels.
- Method: regression analysis
- Eg. Predict the amount of revenue that each item will generate during upcoming sale at AllElectronics

# Data Mining Functionalities



## 4. Cluster analysis

- Class label is unknown: Group data to form new classes – unsupervised learning
- Maximizing intra-class similarity & minimizing interclass similarity
- A *cluster* is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters



# Data Mining Functionalities



## 5. Outlier analysis

- Outlier: Data object that does not comply with the general behavior of the data
- Noise or exception? Useful in fraud detection, rare events analysis
- fraudulent usage of credit cards
  - detect purchases of extremely large amounts for a given account number
  - compare to regular charges incurred by the same account


# Are All the “Discovered” Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting!
- **Interestingness measures**
  - A pattern is **interesting** if it is easily understood by humans, valid on new\_or test data with some degree of **certainty**, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on **statistics and structures of patterns**, e.g., support and confidence for association rules
$$\text{support}(X \Rightarrow Y) = P(X \cup Y).$$
$$\text{confidence}(X \Rightarrow Y) = P(Y|X).$$
  - Subjective: based on **user's belief** in the data, e.g., unexpectedness, novelty, actionability, etc.
  - Manager – frequent customer, analyst – employee performance pattern



# Find All and Only Interesting Patterns?

- Find all the interesting patterns: **Completeness**
  - Can a data mining system find all the interesting patterns? Do we need to find all of the interesting patterns? - inefficient
  - User provided constraints & interestingness measure
  - Association , classification , clustering
- Search for only interesting patterns: An **optimization** problem – desirable but, challenging issue
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First generate all the patterns and then filter out the uninteresting ones
    - Generate only the interesting patterns—mining query optimization

- 
- How is a *data warehouse* different from a database? How are they similar?
  - Differences: A data warehouse is a repository of information collected from multiple sources, over a history of time, stored under a unified schema, and used for data analysis and decision support; whereas a database, is a collection of interrelated data that represents the current status of the stored data. There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another. A database system supports ad-hoc query and on-line transaction processing.
  - Similarities: Both are repositories of information, storing huge amounts of persistent data.