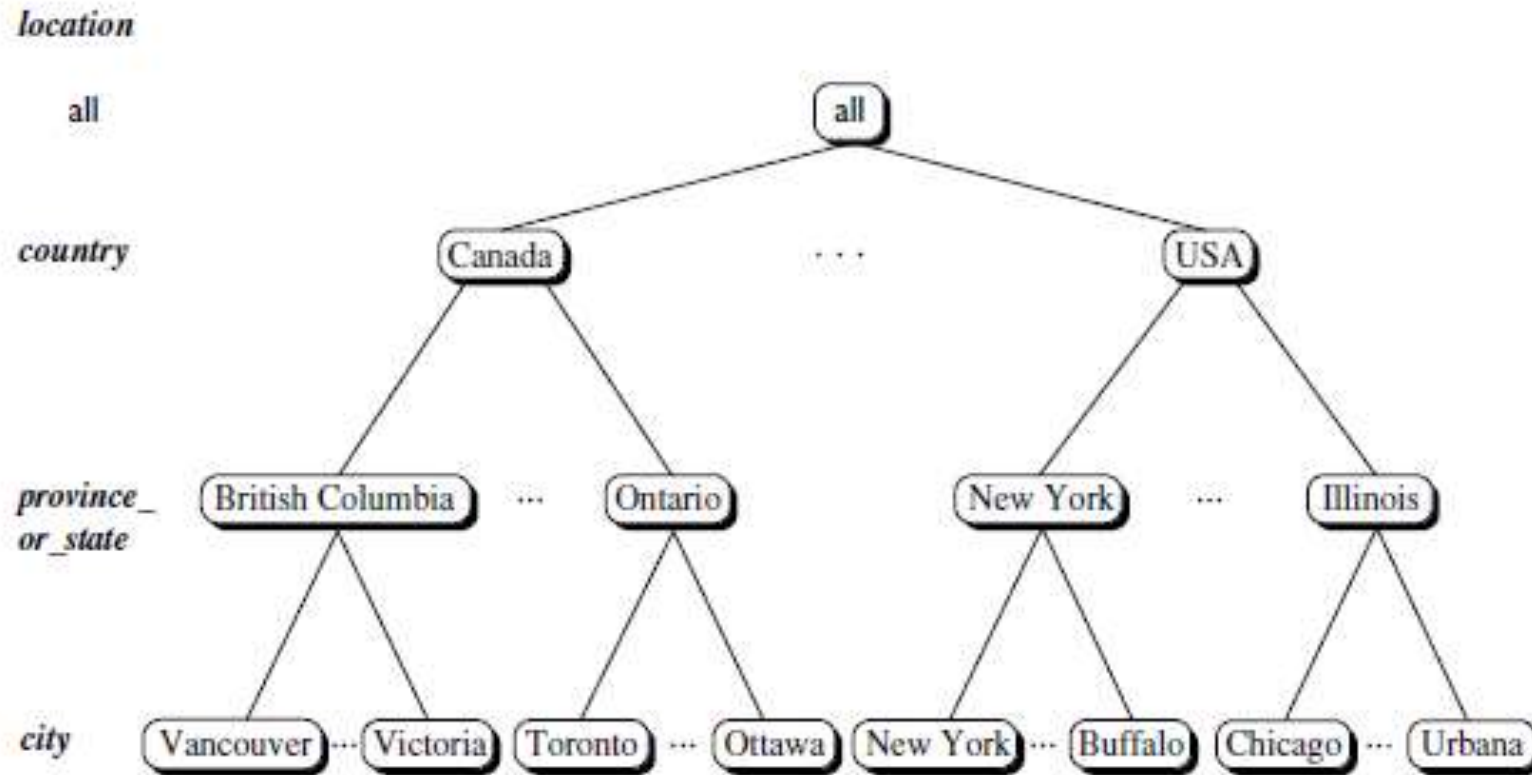
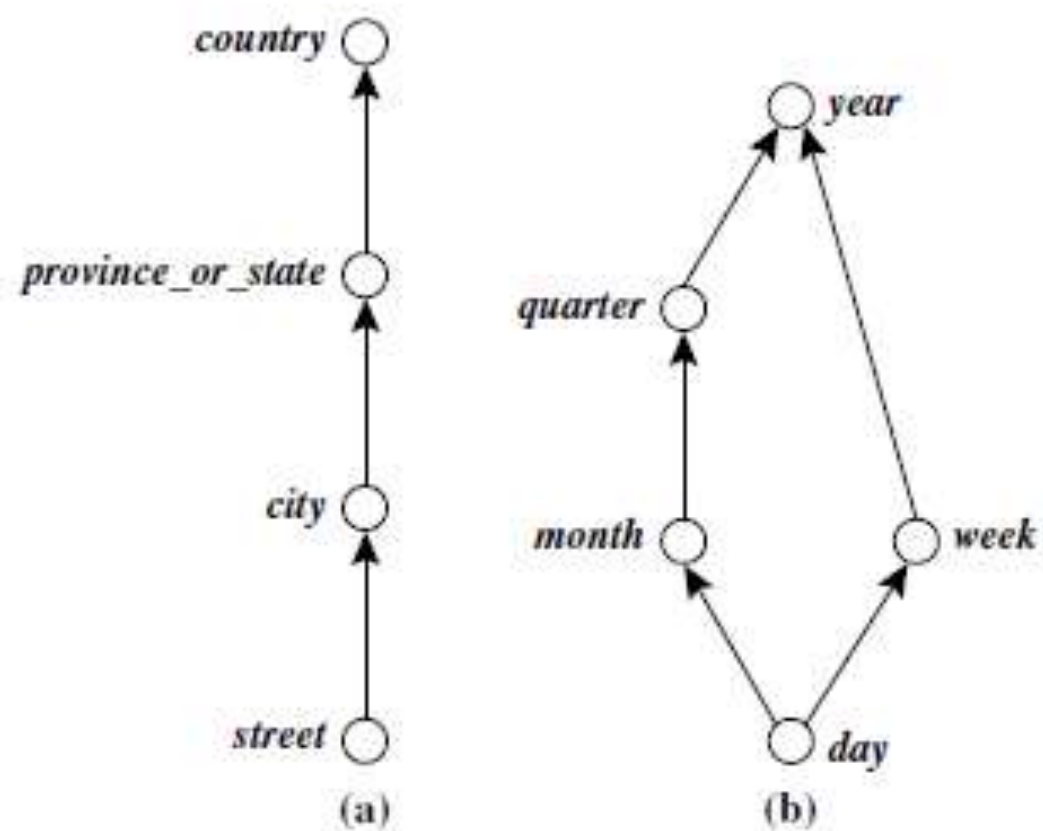


# Concept Hierarchy

- A **concept hierarchy** defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.



**Figure 4.9** A concept hierarchy for *location*. Due to space limitations, not all of the hierarchy nodes are shown, indicated by ellipses between nodes.



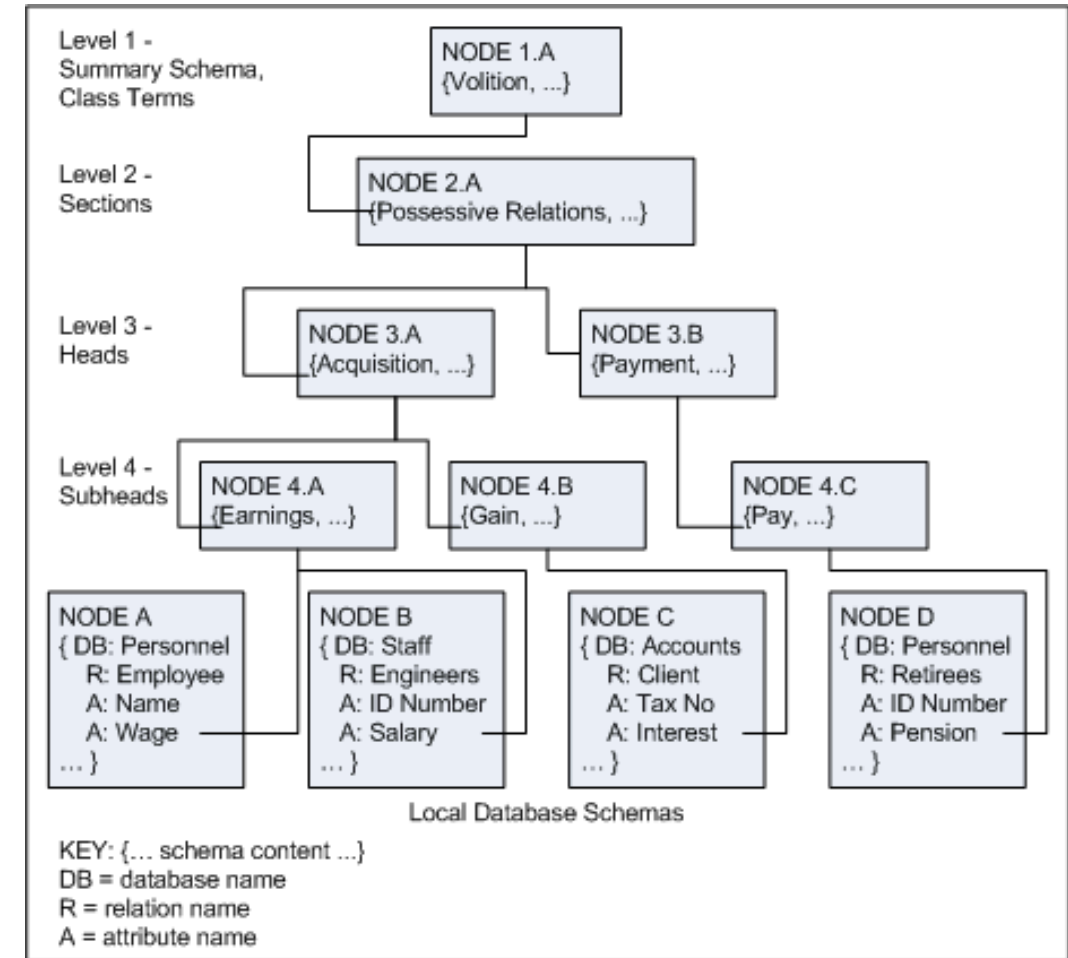
**Figure 4.10** Hierarchical and lattice structures of attributes in warehouse dimensions: (a) a hierarchy for *location* and (b) a lattice for *time*.

- Hierarchical organization -more efficient and effective data analysis.
- Ability to drill down to more specific levels of detail when needed.
- Use - to organize and classify data in a way that makes it more understandable and easier to analyze.
- Main idea behind –
  - *the same data can have different levels of granularity or levels of detail*
  - *By organizing the data in a hierarchical fashion, it is easier to understand and perform analysis.*

# Types of Concept Hierarchies

## Schema Hierarchy

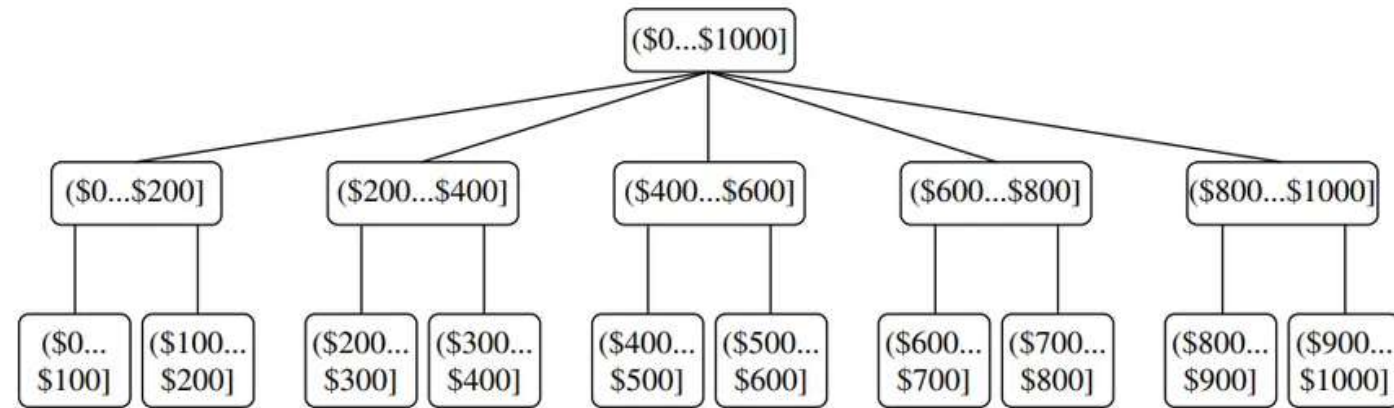
- Used to **organize the schema of a database in a logical and meaningful way**, grouping similar objects together.
- Can be used to **organize different types of data**, such as tables, attributes, and relationships, in a logical and meaningful way.
- **Useful in data warehousing**, where data from multiple sources needs to be integrated into a single database.



# Types of Concept Hierarchies

## Set-Grouping Hierarchy

- Based on [set theory](#)
- Each set in the hierarchy is [defined in terms of its membership in other sets](#).
- Can be used for [data cleaning](#), [data pre-processing](#) and [data integration](#).
- Can be used to
  - identify and remove outliers, noise, or inconsistencies from the data.
  - to integrate data from multiple sources.



---

A concept hierarchy for the attribute *price*, where an interval  $(\$X... \$Y]$  denotes the range from  $\$X$  (exclusive) to  $\$Y$  (inclusive).

# Types of Concept Hierarchies

## Operation-Derived Hierarchy

- Organize data by **applying a series of operations or transformations** to the data.
- The operations are applied in a **top-down fashion**.
- **Each level of the hierarchy** representing a more general or abstract view of the data than the level below it.
- Typically used in data mining tasks such as **clustering** and **dimensionality reduction**.
- The operations applied can be **mathematical or statistical operations** such as **aggregation, normalization**
- Eg: email address: login name< department< university< Country
- **abc@cs.ktu.in**

# Types of Concept Hierarchies

## Rule-based Hierarchy

- Used to organize data by applying a set of rules or conditions to the data.
- Useful in data mining tasks such as classification, decision-making, and data exploration.
- It allows to the assignment of a class label or decision to each data point based on its characteristics
- Identifies patterns and relationships between different attributes of the data.

**Example 3.9** Suppose we have a database **university**, in which a relation **student** is defined by the schema **student**(*name, status, sex, major, age, birthPlace, GPA*). A rule-based concept hierarchy is shown in Figure 3.6 for its graphical expression and Figure 3.7 for its generalization rules. Using DMQL, we can define this hierarchy by statements such as:

```
define hierarchy gpaHier on student as
    level3: "2.0~2.5" < level2: average
    if status = "undergraduate"
```



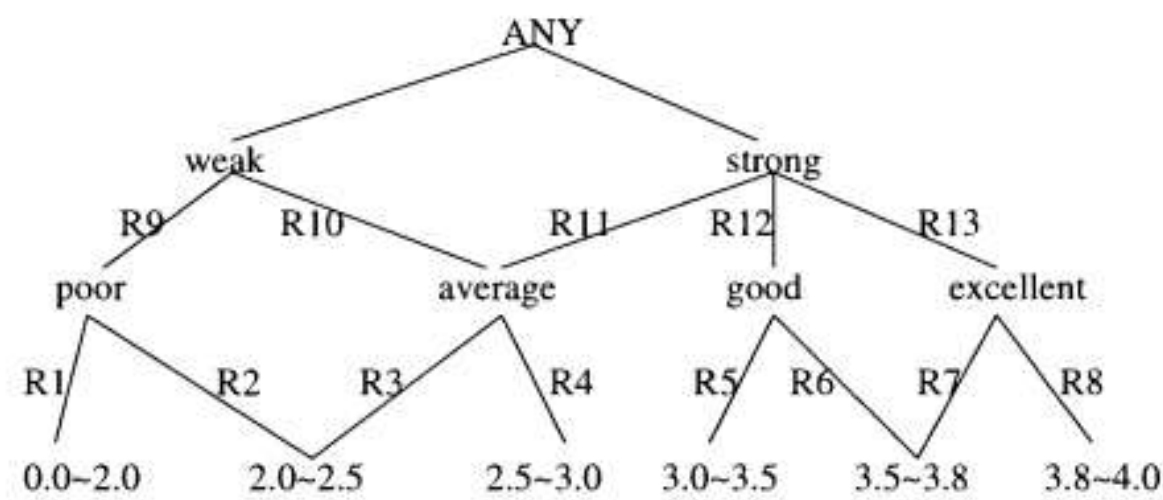


Figure 3.6: A rule-based concept hierarchy *gpaHier* for attribute GPA

- $R_1 : \{0.0 \sim 2.0\} \rightarrow \text{poor};$
- $R_2 : \{2.0 \sim 2.5\} \wedge \{\text{graduate}\} \rightarrow \text{poor};$
- $R_3 : \{2.0 \sim 2.5\} \wedge \{\text{undergraduate}\} \rightarrow \text{average};$
- $R_4 : \{2.5 \sim 3.0\} \rightarrow \text{average};$
- $R_5 : \{3.0 \sim 3.5\} \rightarrow \text{good};$
- $R_6 : \{3.5 \sim 3.8\} \wedge \{\text{graduate}\} \rightarrow \text{good};$
- $R_7 : \{3.5 \sim 3.8\} \wedge \{\text{undergraduate}\} \rightarrow \text{excellent};$
- $R_8 : \{3.8 \sim 4.0\} \rightarrow \text{excellent};$
- $R_9 : \{\text{poor}\} \rightarrow \text{weak};$
- $R_{10} : \{\text{average}\} \wedge \{\text{senior, graduate}\} \rightarrow \text{weak};$
- $R_{11} : \{\text{average}\} \wedge \{\text{freshman, sophomore, junior}\} \rightarrow \text{strong};$
- $R_{12} : \{\text{good}\} \rightarrow \text{strong};$
- $R_{13} : \{\text{excellent}\} \rightarrow \text{strong}.$

# Need of Concept Hierarchy in Data Mining

- There are several reasons why a concept hierarchy is useful in data mining:

- 1. Improved Data Analysis**
- 2. Improved Data Visualization and Exploration**
- 3. Improved Algorithm Performance**
- 4. Data Cleaning and Pre-processing**
- 5. Domain Knowledge**

# Applications of Concept Hierarchy

There are several applications of concept hierarchy in data mining, some examples are:

- **Data Warehousing**
- **Business Intelligence**
- **Online Retail**
- **Healthcare**
- **Natural Language Processing**
- **Fraud Detection**

# OLAP Operations

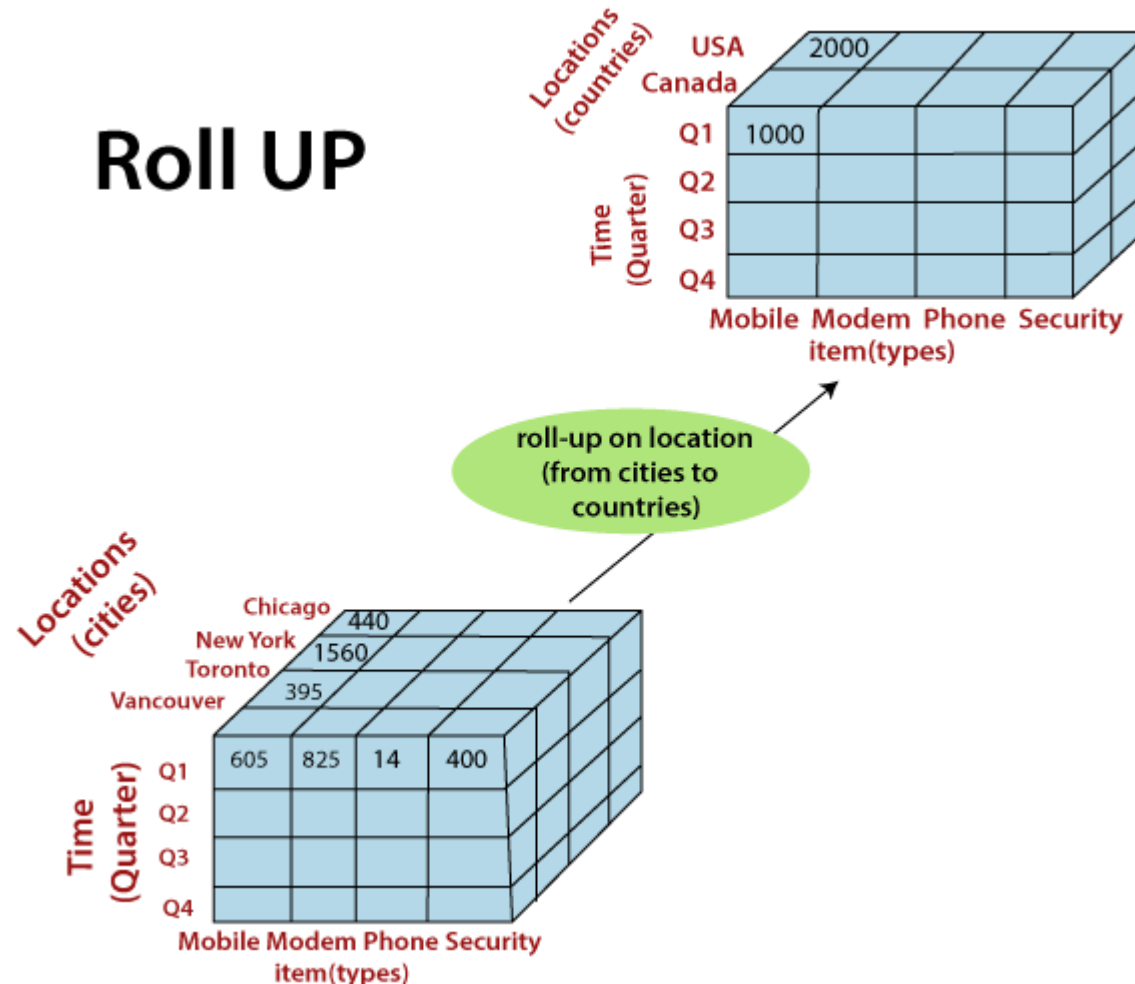
- OLAP ONLINE ANALYTICAL PROCESSING (OLAP) provides a user-friendly environment for Interactive data analysis.
- In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies.

# OLAP operations

- ROLL-UP (aka DRILL UP): summarize data
- ROLL DOWN or DRILL-DOWN : reverse of roll up
- SLICING AND DICING : project and select
- PIVOT (ROTATE): reorient the cube
- Additional
  - Drill across
  - Drill through

# Roll Up/Drill Up/Aggregation

- Performs aggregation on a data cube, either by *climbing up a concept hierarchy* for a dimension or by *dimension reduction*.



# Roll Down/Drill-down

- Drill-down is the reverse of roll-up.
- Drill-down is like **zooming-in** on the data cube.
- It navigates from **less detailed data to more detailed data**.
- Drill-down can be realized by either *stepping down a concept hierarchy* for a dimension or *introducing additional dimensions*.

**Locations (cities)**

**Time (Quarter)**

**Mobile Modem Phone Security item(types)**

Locations (cities)	Mobile	Modem	Phone	Security
Chicago	440			
New York	1560			
Toronto	395			
Vancouver	605	825	14	400
	Q1	Q2	Q3	Q4

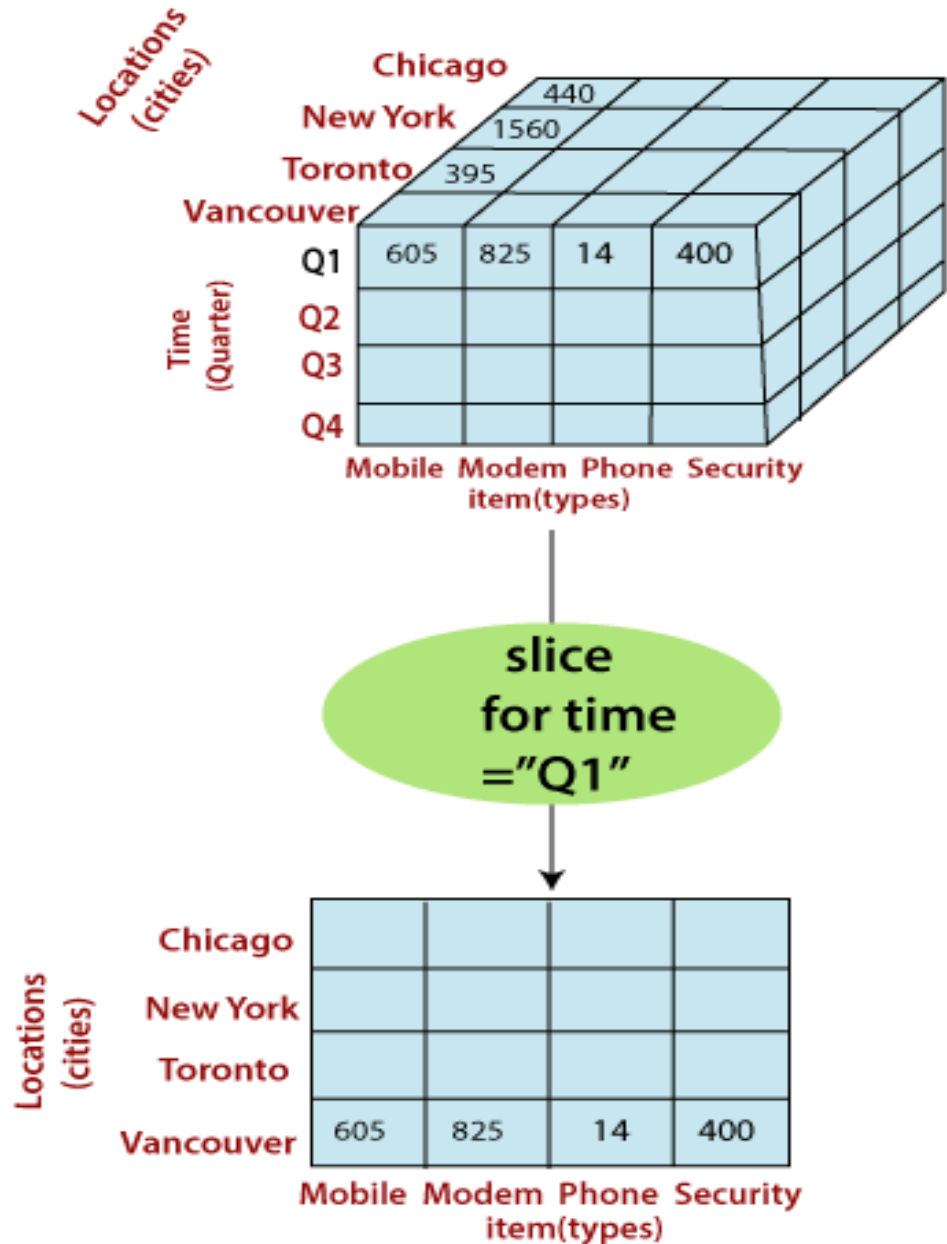
The diagram illustrates a 3D data cube with three dimensions: Time (months), Locations (countries), and Mobile Modem Phone Security item (types). The Time dimension is represented by a vertical axis with labels for each month from Jan to Dec. The Locations dimension is represented by a horizontal axis with labels for Chicago, New York, Toronto, and Vancouver. The Mobile Modem Phone Security item dimension is represented by a depth axis with labels 440, 1560, and 395. The cube is composed of smaller cells, with some cells containing numerical values such as 150, 100, and 150.



# Slice

- A **slice** is a **subset of the cubes** corresponding to a **single value for one or more members of the dimension**.
- Eg: when the customer wants **a selection on one dimension** of a three-dimensional cube resulting in a two-dimensional site.
- Slice operations perform a selection on one dimension of the given cube, thus resulting in a **subcube**.

# Slice



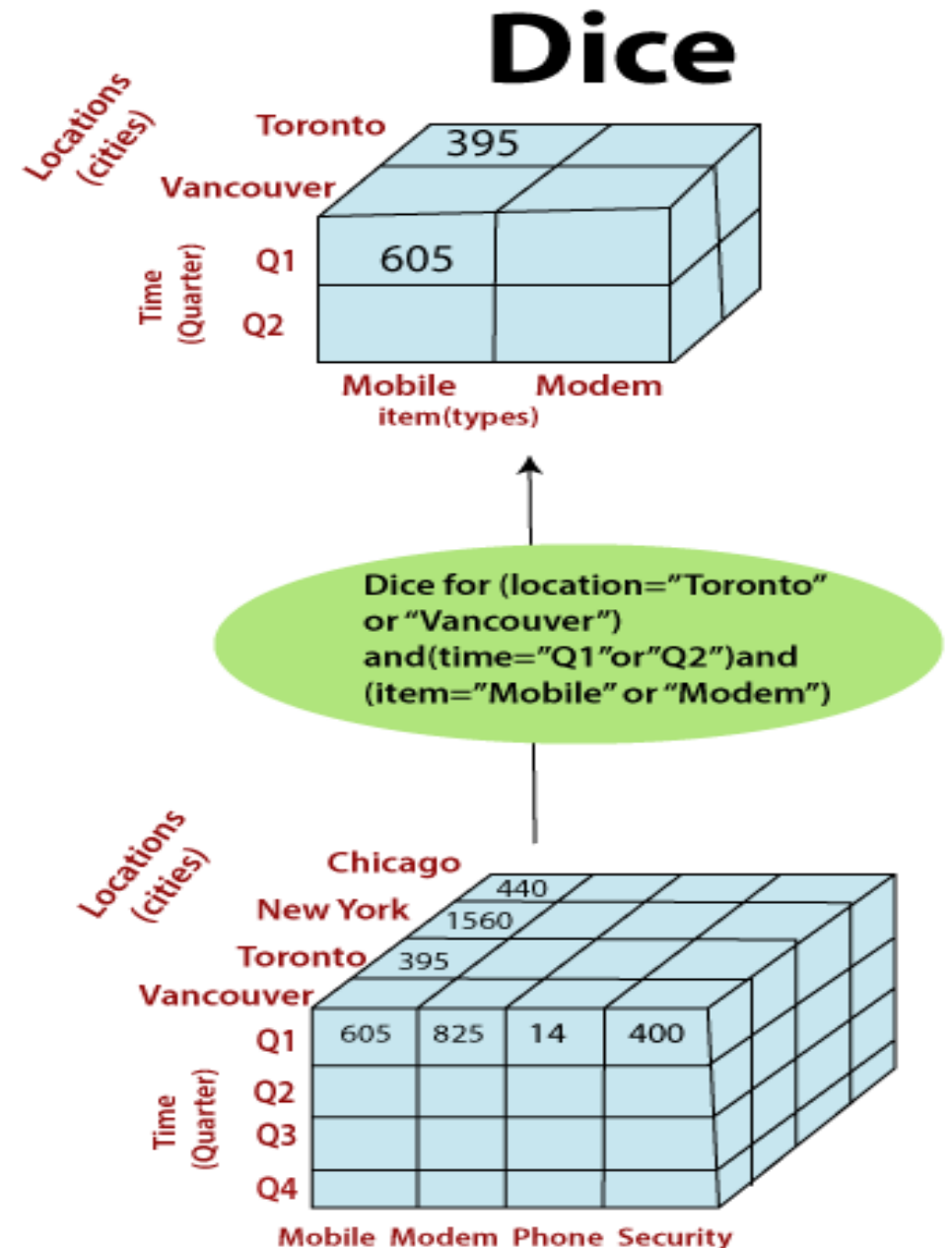
- A slice operation where the sales data are selected from the central cube for the dimension *time* using the criterion *time* = "Q1."

# Dice

- The *dice* operation defines a subcube by performing a selection on two or more dimensions.
- A dice operation on the central cube based on the following selection criteria that involve three dimensions:

(*location* = “Toronto” or “Vancouver”)  
and (*time* = “Q1” or “Q2”)

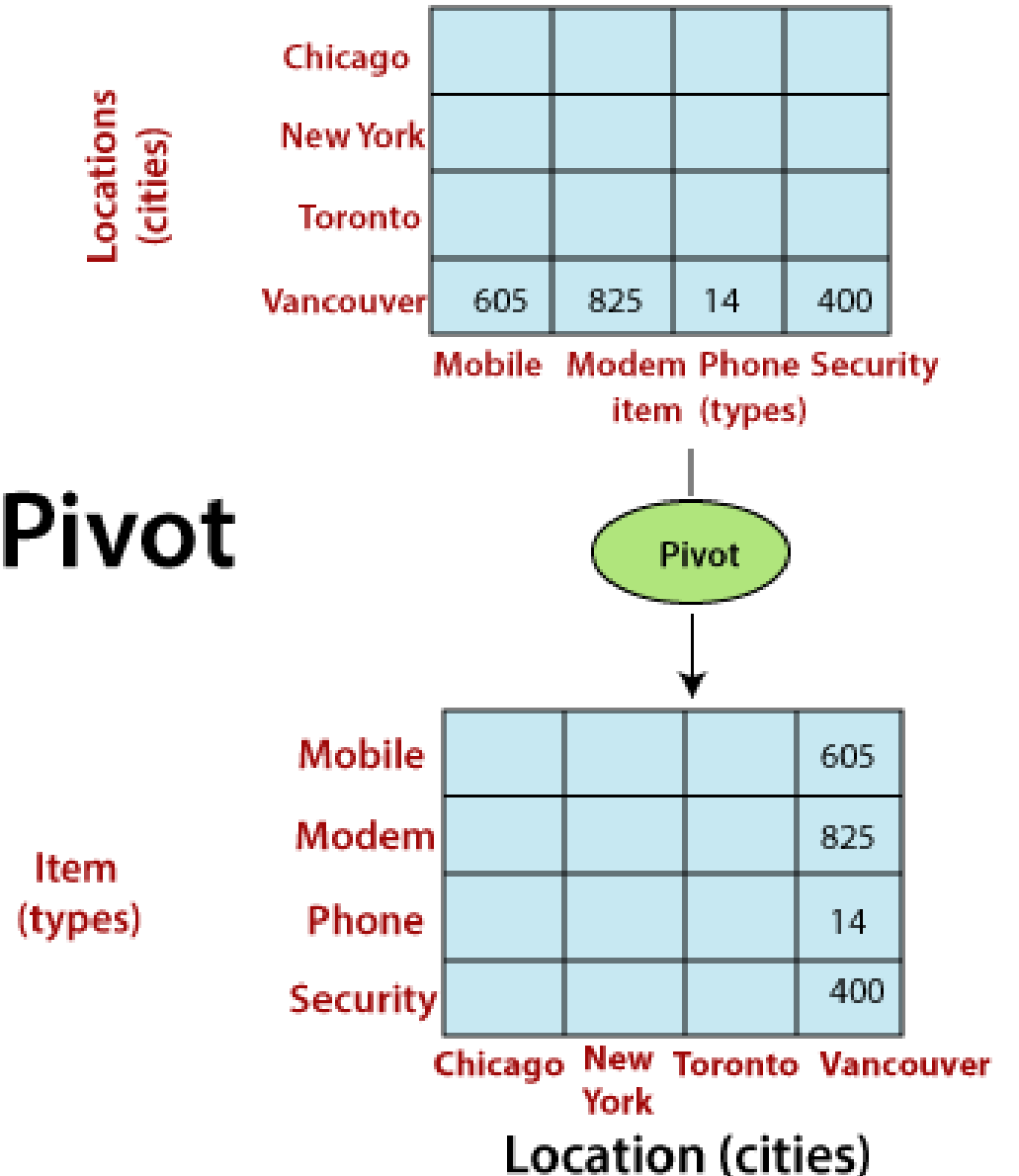
and (*item* = “mobile ” or “modem”).



# Pivot

- The pivot operation is also called a **rotation**.
- Pivot is a **visualization operation**.
- Rotates the data axes in view to **provide an alternative presentation of the data**.
- May **swap the rows and columns** or **move one of the row-dimensions into the column dimensions**.

## Pivot



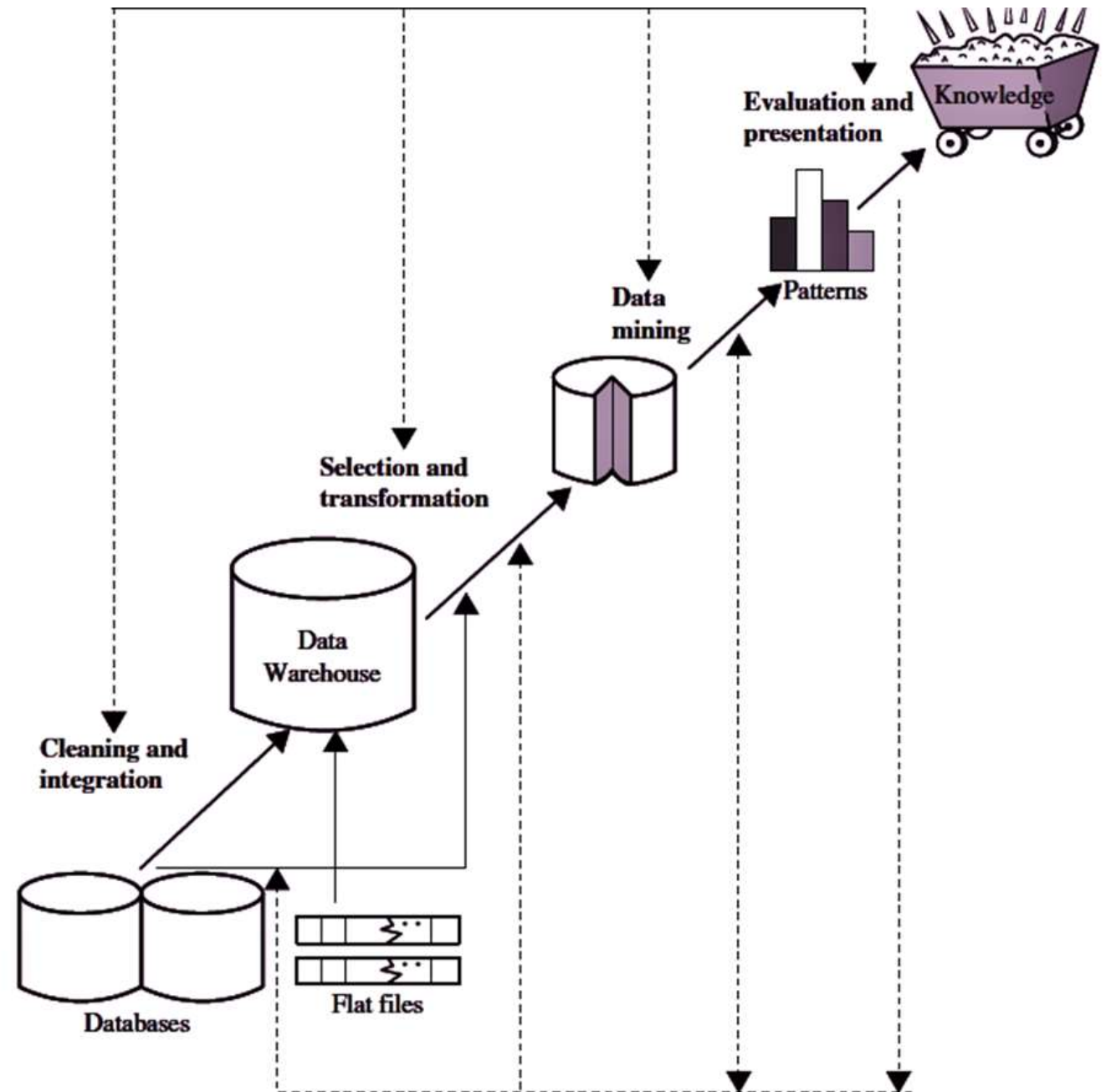
# Other OLAP Operations

- **Drill-across** executes queries involving (i.e., across) **more than one fact table**.
- The **drill-through** operation **uses relational SQL facilities to drill through** the bottom level of a data cube down to its back-end relational tables.

# Introduction to KDD process

KDD- Knowledge Discovery in Datasets

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



# Advantages of KDD

- Improves decision-making
- Increased efficiency
- Better customer service
- Fraud detection
- Predictive modeling



# Disadvantages of KDD

- **Privacy concerns**
- **Complexity**
- **Unintended consequences**
- **Data Quality**
- **High cost**
- **Overfitting**

# Data Mining

# Definition

**Data mining** is the *process* of discovering interesting patterns and knowledge from *large* amounts of data.

- The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

## Key Outcomes of Data Mining

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large datasets and databases

# What is Data Mining?

- The process of extracting knowledge or insights from large amounts of data using various statistical and computational techniques.
- The data can be structured, semi-structured or unstructured.
- Data can be stored in various forms such as databases, data warehouses, and data lakes.
- Primary goal - to discover hidden patterns and relationships in the data that can be used to make informed decisions or predictions.
- How? – By exploring the data using various techniques such as clustering, classification, regression analysis, association rule mining, and anomaly detection.
- Applications- marketing, finance, healthcare, and telecommunications.
- Eg: in marketing, data mining can be used to identify customer segments and target marketing campaigns, while in healthcare, it can be used to identify risk factors for diseases and develop personalized treatment plans.



# Alternative names for Data Mining

1. Knowledge discovery (mining) in databases (KDD)
2. Knowledge extraction
3. Data/pattern analysis
4. Data archaeology
5. Data dredging
6. Information harvesting
7. Business intelligence

# Data Mining on what kinds of data?

- Flat Files
- Relational Databases
- Data Warehouse
- Transactional Database
- Multimedia Database
- Spatial Database
- Time-series database
- WWW

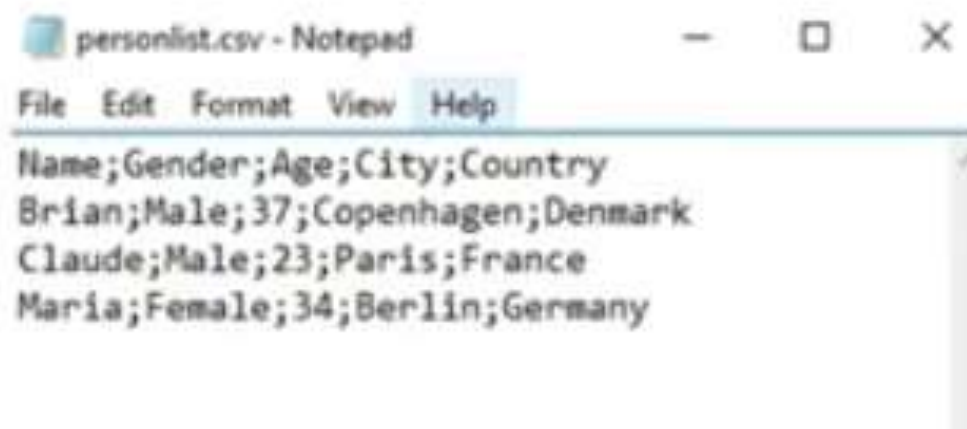
# WHAT KIND OF DATA CAN BE MINED?

In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository.

Here are some examples in detail:

## Flat files

- Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied.
- The data in these files can be transactions, time-series data, scientific measurements, etc.

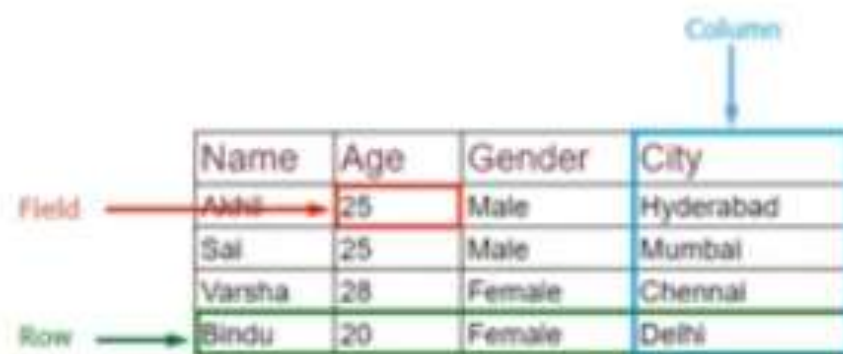


```
personlist.csv - Notepad
File Edit Format View Help
Name;Gender;Age;City;Country
Brian;Male;37;Copenhagen;Denmark
Claude;Male;23;Paris;France
Maria;Female;34;Berlin;Germany
```

# WHAT KIND OF DATA CAN BE MINED? (CONT.)

## Relational Database

- Briefly, a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships.
- Tables have columns and rows, where columns represent attributes and rows represent tuples.
- Data mining algorithms using relational databases can be more versatile than data mining algorithms specifically written for flat files, since they can take advantage of the structure inherent to relational databases.
- While data mining can benefit from SQL for data selection, transformation and consolidation, it goes beyond what SQL could provide, such as predicting, comparing, detecting deviations, etc.



The diagram shows a table with four columns: Name, Age, Gender, and City. The first row contains the values 'Anil', '25', 'Male', and 'Hyderabad'. The second row contains 'Sai', '25', 'Male', and 'Mumbai'. The third row contains 'Varsha', '28', 'Female', and 'Chennai'. The fourth row contains 'Bindu', '20', 'Female', and 'Delhi'. Annotations include a blue arrow labeled 'Column' pointing to the 'City' header, a red arrow labeled 'Field' pointing to the 'Anil' value, and a green arrow labeled 'Row' pointing to the 'Bindu' value.

Name	Age	Gender	City
Anil	25	Male	Hyderabad
Sai	25	Male	Mumbai
Varsha	28	Female	Chennai
Bindu	20	Female	Delhi



# WHAT KIND OF DATA CAN BE MINED? (CONT.)

## Data Warehouse

- A data warehouse is defined as the collection of data integrated from multiple sources that will queries and decision making.
- A data warehouse gives the option to analyze data from different sources under the same roof.
- To facilitate decision-making and multi-dimensional views, data warehouses are usually modeled by a multi-dimensional data structure (cube).



(279) What kind of data can be mined - YouTube - Google Chrome

# WHAT KIND OF DATA CAN BE MINED? (CONT.)

## Transactional Databases

- Transactional databases is a collection of data organized by time stamps, date, etc to represent transaction in databases.
- This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed.
- Highly flexible system where users can modify information without changing any sensitive information.
- Follows ACID property of DBMS.

Tid	Items
1	1 5 9 14 19 23 27 32 37
2	1 5 9 14 19 23 27 32 38
3	2 8 12 14 19 24 27 35 38
4	3 5 9 14 19 23 27 36 38
5	1 8 9 14 19 23 27 32 38
6	2 8 12 14 19 24 27 35 38
7	1 5 9 14 19 23 27 32 37
8	1 5 9 14 19 23 27 32 38
9	2 8 12 14 19 24 27 35 38
10	1 5 9 14 19 23 27 34 37
11	1 5 9 14 19 23 27 32 38
12	1 5 9 14 19 23 27 32 38
13	2 8 12 14 19 24 27 35 38
14	1 5 9 14 19 23 27 34 37
15	1 8 9 14 19 23 27 32 38
16	2 8 12 14 19 24 27 35 38
17	1 5 9 14 19 23 27 32 37
18	1 5 9 14 19 23 27 32 38
19	2 8 12 14 19 24 27 35 38
20	1 5 9 14 19 23 27 34 37

# WHAT KIND OF DATA CAN BE MINED? (CONT.)

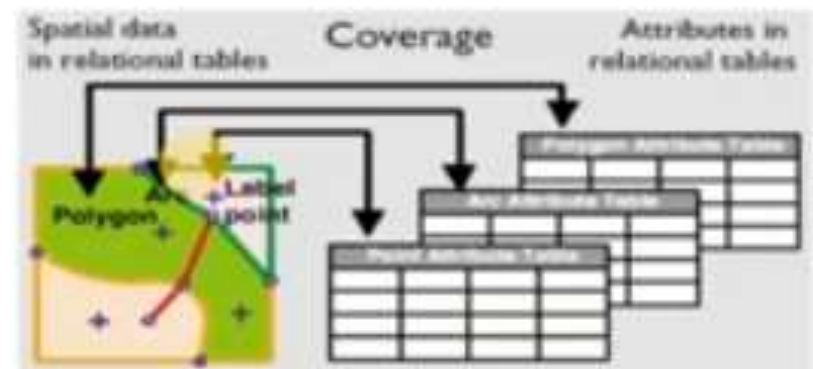
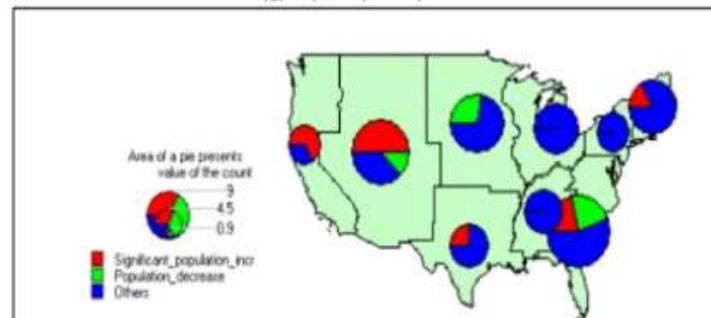
## Multimedia Databases

- Multimedia databases consists audio, video, images and text media.
- They can be stored on Object-Oriented Databases.
- They are used to store complex information in a pre-specified formats.



## Spatial Database

- Store geographical information.
- Stores data in the form of coordinates, topology, lines, polygons, etc.





# WHAT KIND OF DATA CAN BE MINED? (CONT.)

## WWW (World Wide Web)

- WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network.
- It is the most heterogeneous repository as it collects data from multiple resources.
- It is dynamic in nature as Volume of data is continuously increasing and changing.



Parameter	KDD	Data Mining
Definition	KDD refers to a process of identifying valid, novel, potentially useful, and ultimately understandable patterns and relationships in data.	Data Mining refers to a process of extracting useful and valuable information or patterns from large data sets.
Objective	To find useful knowledge from data.	To extract useful information from data.
Techniques Used	Data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge representation and visualization.	Association rules, classification, clustering, regression, decision trees, neural networks, and dimensionality reduction.
Output	Structured information, such as rules and models, that can be used to make decisions or predictions.	Patterns, associations, or insights that can be used to improve decision-making or understanding.
Focus	Focus is on the discovery of useful knowledge, rather than simply finding patterns in data.	Data mining focus is on the discovery of patterns or relationships in data.
Role of domain expertise	Domain expertise is important in KDD, as it helps in defining the goals of the process, choosing appropriate data, and interpreting the results.	Domain expertise is less critical in data mining, as the algorithms are designed to identify patterns without relying on prior knowledge.

# Data Mining Functionalities

- Data mining functionalities specify the kind of patterns to be found in data mining tasks.
- In general, data mining tasks can be classified into two categories: **descriptive** and **predictive**.
- **Descriptive mining** tasks characterize the general properties of the data in the target data set.
- **Predictive mining** tasks perform inference on the current data in order to make predictions.

# **Functionalities of Data Mining**

Class/ Concept Descriptions

Mining Frequent Patterns

Association Analysis

Classification

Cluster Analysis

Evolution & Deviation Analysis

Correlation Analysis

Prediction

Outlier Analysis

# Concept/Class Description

Data can be associated with classes or concepts.

***Class*** : A collection of things sharing a common attribute

Classes of items for sale include *computers and printers*

***Concept***: An abstract or general idea inferred or derived from specific instances

Concepts of customers include *bigSpenders and budgetSpenders*.

Summarized, concise and precise descriptions of individual **classes** and **concepts** are called ***class/concept descriptions***.

These descriptions can be derived using

- (1) *data characterization*, by summarizing the data of the class under study (often called the **target class**) in general terms
- (2) *data discrimination*, by comparison of the target class with one or a set of comparative classes (often called the **contrasting classes**), or
- (3) both data characterization and discrimination.



# Data characterization

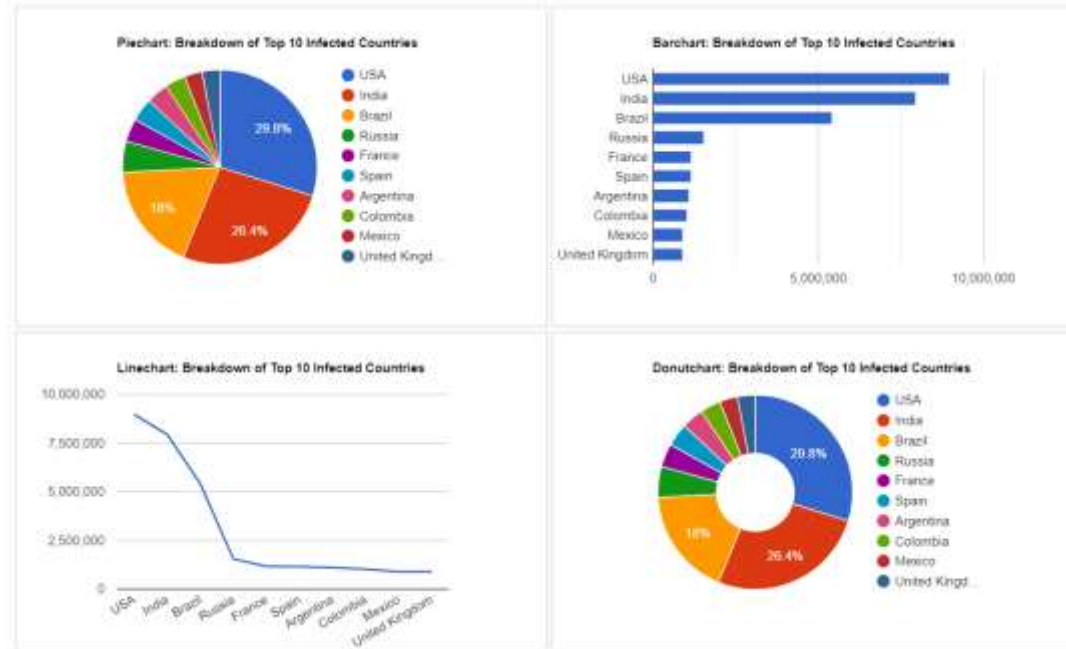
- **Data characterization** is a **summarization of the general characteristics** or features of a target class of data.
- The **data** corresponding to the user-specified class are **typically collected by a query**.

For example, to study the characteristics of **software products with sales that increased by 10% in the previous year**, the data related to such products can be collected by executing an SQL query on the sales database.

- Simple data summaries can be done based on **statistical measures and plots**.
- The **data cube–based OLAP roll-up operation** can be used to perform data summarization along a specified dimension.
- An ***attribute-oriented induction technique*** can be used to perform data generalization and characterization without step-by-step user interaction

# Data characterization

- The output of data characterization can be presented in various forms.
- **Eg: pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables, including crosstabs.**



- The resulting descriptions can also be presented as **generalized relations** or in rule form (called **characteristic rules**).

# Eg: Data characterization

A customer relationship manager at *AllElectronics* may order the following data mining task:

*“Summarize the characteristics of customers who spend more than \$5000 a year at AllElectronics.”*

The result is a general profile of these customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings.

The data mining system should allow the customer relationship manager to drill down on any dimension, such as on *occupation* to view these customers according to their type of employment.

# Data discrimination

- **Data discrimination** is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.
- The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries.
  - *For example, a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period.*
- The methods used for data discrimination are similar to those used for data characterization.

# Data discrimination

- The forms of output presentation are similar to those for characteristic descriptions, although discrimination descriptions should include comparative measures that help to distinguish between the target and contrasting classes.
- Discrimination descriptions expressed in the form of rules are referred to as **discriminant rules**.

# Eg:Data discrimination

*A customer relationship manager at AllElectronics may want to compare two groups of customers—those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g., less than three times a year).*

The resulting description provides a general comparative profile of these customers, such as that

- 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education,*
- whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree.*

Drilling down on a dimension like occupation, or adding a new dimension like income level, may help to find even more discriminative features between the two classes.

# Mining Frequent Patterns

**Frequent patterns** are patterns that occur frequently in data.

*Frequent itemset* - refers to a set of items that often appear together in a transactional data set;

*Eg: milk and bread, which are frequently bought together in grocery stores by many customers.*

*Sequential pattern* A frequently occurring subsequence.

*Eg: customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card*

*Frequent substructure* refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (*frequent*) *structured pattern*.

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

*Frequent itemset mining* is a fundamental form of frequent pattern mining.

# Association analysis.

Suppose that, as a marketing manager at *AllElectronics*, you want to know which items are frequently purchased together (i.e., within the same transaction).

An example of such a rule, mined from the *AllElectronics* transactional database, is:

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

where  $X$  is a variable representing a customer.

A **confidence**, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well.

A 1% **support** means that 1% of all the transactions under analysis show that computer and software are purchased together.



# Association analysis.

- The association rule involves a single attribute or predicate (i.e., *buys*) that repeats.
- Association rules that contain a single predicate are referred to as **single-dimensional association rules**.
- Dropping the predicate notation, the rule can be written simply as

*computer*  $\Rightarrow$  *software* [1%, 50%]

# Example: Multi dimensional Association rules

*AllElectronics* relational database related to purchases, a data mining system may find association rules like

$$\text{age}(X, \text{"20..29"}) \wedge \text{income}(X, \text{"40K..49K"}) \Rightarrow \text{buys}(X, \text{"laptop"})$$

[support = 2%, confidence = 60%].

# Example: Multi dimensional Association rules

Of the *AllElectronics* customers under study

- *2% are 20 to 29 years old with an income of \$40,000 to \$49,000 and have purchased a laptop (computer) at AllElectronics.*
- *There is a 60% probability that a customer in this age and income group will purchase a laptop.*
- An association involving more than one attribute or predicate (i.e., *age*, *income*, and *buys*).
- Each attribute is referred to as a dimension-> referred to as a ***multidimensional association rule***.

# Classification and Regression for Predictive Analysis

## Classification

- **Classification** is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts.
- The model are derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known).
- The model is **used to predict the class label of objects** for which the the class label is unknown.
- How is the derived model presented?
  - *classification rules (i.e., IF-THEN rules)*
  - *Decision trees*
  - *Mathematical formulae*
  - *neural networks*

## Decision tree

- A flowchart-like tree structure
- Each node denotes a test on an attribute value
- Each branch represents an outcome of the test
- Tree leaves represent classes or class distributions.
- Decision trees can easily be converted to classification rules.

## Neural network

- used for classification
- A collection of **neuron-like** processing **units with weighted connections** between the units.

## Other Classification Models:

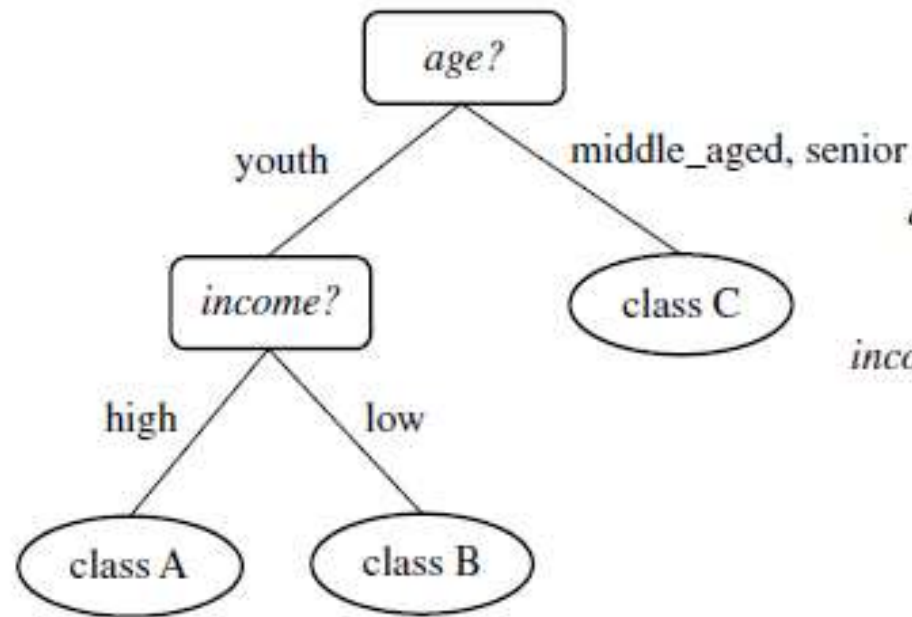
Naïve Bayesian classification

Support Vector Machines

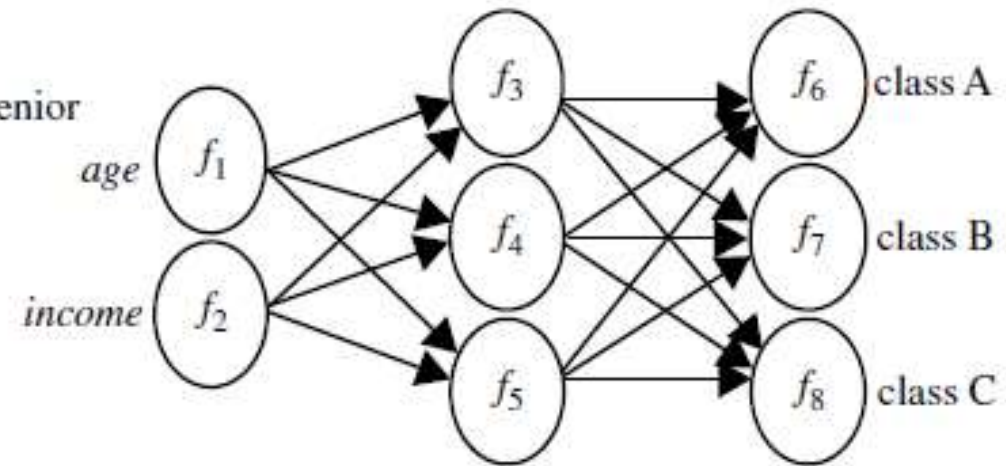
*k*-nearest-neighbor classification.

$age(X, \text{"youth"}) \text{ AND } income(X, \text{"high"}) \longrightarrow class(X, \text{"A"})$   
 $age(X, \text{"youth"}) \text{ AND } income(X, \text{"low"}) \longrightarrow class(X, \text{"B"})$   
 $age(X, \text{"middle\_aged"}) \longrightarrow class(X, \text{"C"})$   
 $age(X, \text{"senior"}) \longrightarrow class(X, \text{"C"})$

(a)



(b)



(c)

**Figure 1.9** A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

# Regression

- **Regression** models **continuous-valued functions**.
- Used to **predict missing or unavailable *numerical data values*** rather than (discrete) class labels.
- *Prediction* -> both numeric prediction and class label prediction.
- ***Regression analysis*** - a statistical methodology that is most often used for numeric prediction.
- Regression also encompasses the **identification of distribution *trends* based on the available data**.

# Classification and Regression for Predictive Analysis

## Relevance analysis

- Classification and regression may need to be **preceded by relevance analysis**.
- Attempts to **identify attributes that are significantly relevant** to the classification and regression process.
- Other attributes, which are **irrelevant**, can then be **excluded** from consideration.



# Eg: Classification

Suppose as a sales manager of *AllElectronics* you want to classify a large set of items in the store, based on three kinds of responses to a sales campaign: *good response*, *mild response* and *no response*.

Derive a model for each of these three classes based on the descriptive features of the items, such as *price*, *brand*, *place made*, *type*, and *category*.

The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.

# Eg: Regression

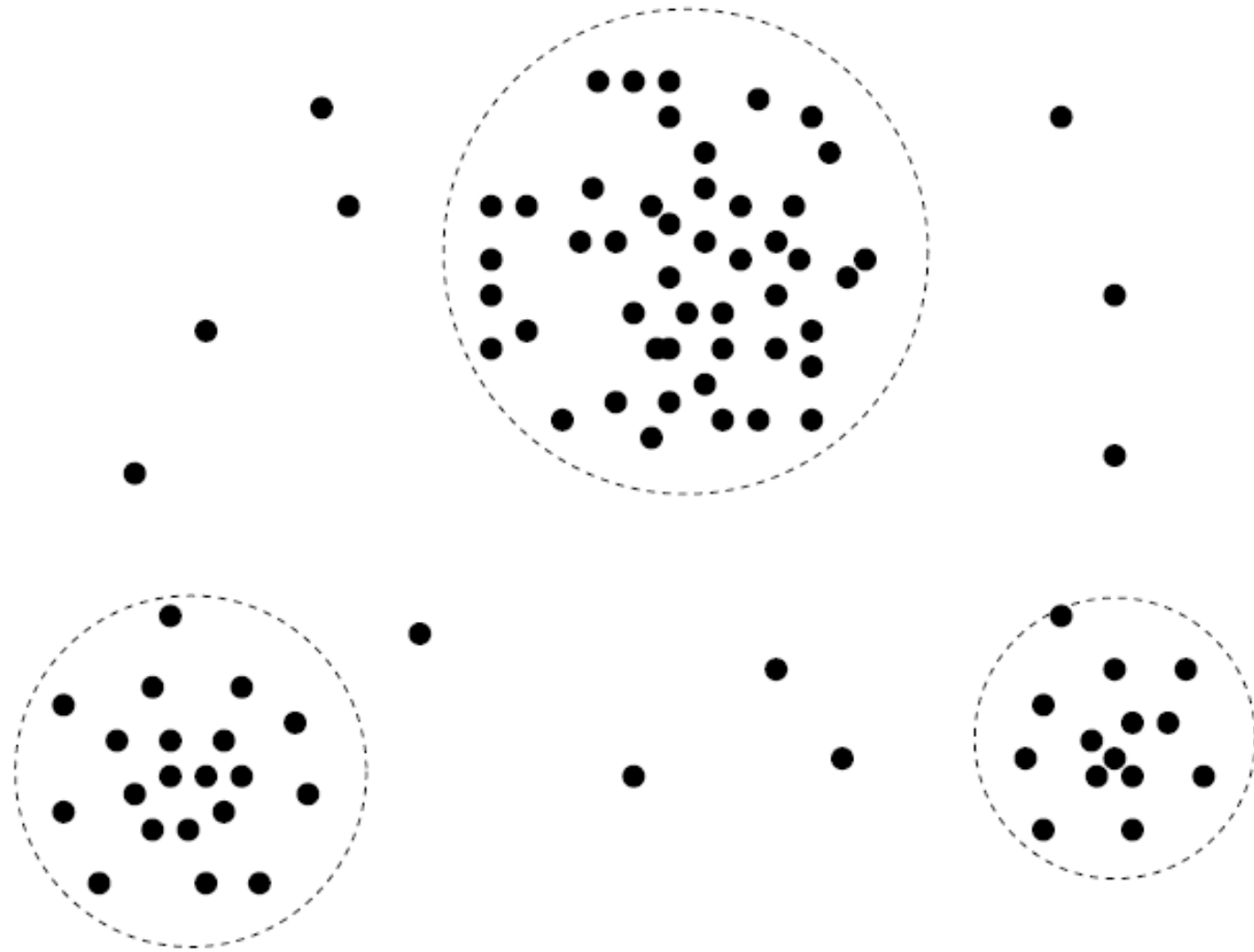
- Predict the amount of revenue that each item will generate during an upcoming sale at *AllElectronics*, based on the previous sales data.
- An example of regression analysis because the regression model constructed will predict a continuous function (or ordered value.)

# Cluster Analysis

- **Clustering** analyzes data objects **without consulting class labels**.
- Clustering **can be used to generate class labels** for a group of data.
- The objects are clustered or grouped based on **the principle of *maximizing the intraclass similarity and minimizing the interclass similarity***.
- Objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters.
- Each cluster so formed can be viewed as **a class of objects**, from which **rules can be derived**.
- Clustering facilitate **taxonomy formation** -> the organization of observations into a hierarchy of classes that group similar events together.

# Example

- Cluster analysis can be performed on *AllElectronics* customer data to identify homogeneous subpopulations of customers.
- These clusters may represent individual target groups for marketing.
- Three clusters of data points are evident.



---

**Figure 1.10** A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

# Outlier Analysis

- A data set may contain **objects that do not comply with the general behavior** or model of the data- ***Outliers***.
- Many data mining methods discard outliers as **noise or exceptions**.
- In some applications the rare events can be more interesting than the more regularly occurring ones.
- The analysis of outlier data is referred to as **outlier analysis** or **anomaly mining**.
- Detected using:
  - **statistical tests** that assume a distribution or probability model for the data
  - **distance measures** where objects that are remote from any other cluster are considered outliers.
  - **Density-based methods** may identify outliers in a local region, although they look normal from a global statistical distribution view.

# Example -Outlier analysis

- Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account.
- Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.

# Are All Patterns Interesting?

- No—only a small fraction of the patterns potentially generated would actually be of interest to a given user.

*“What makes a pattern interesting?”*

*Can a data mining system generate all of the interesting patterns?*

*Can the system generate only the interesting ones?”*



# *“What makes a pattern interesting?”*

A pattern is **interesting** if it is

- (1) *easily understood* by humans
- (2) *valid* on new or test data with some degree of *certainty*
- (3) potentially *useful*
- (4) *novel*.

A pattern is also interesting if it validates a hypothesis that the user *sought to confirm*.

An interesting pattern represents **knowledge**.

# Objective measures of pattern interestingness

- Based on the structure of discovered patterns and the statistics underlying them.
- An objective measure for association rules of the form  $X \Rightarrow Y$  is rule **Support**,
- Represent the percentage of transactions from a transaction database that the given rule satisfies.
- This is taken to be the probability  $P(X \cup Y)$ , where  $X \cup Y$  indicates that a transaction contains both  $X$  and  $Y$ , that is, the union of itemsets  $X$  and  $Y$ .
- **Confidence**, which assesses the degree of certainty of the detected association.
- This is taken to be the conditional probability  $P(Y|X)$ , that is, the probability that a transaction containing  $X$  also contains  $Y$ .
- More formally, support and confidence are defined as
$$\text{support}(X \Rightarrow Y) = P(X \cup Y),$$
$$\text{confidence}(X \Rightarrow Y) = P(Y|X).$$

# Objective measures of pattern interestingness

- **Accuracy** -the percentage of data that are correctly classified by a rule.
- **Coverage** is similar to support- the percentage of data to which a rule applies.
- Although objective measures help identify interesting patterns, they are often insufficient unless combined with subjective measures that reflect a particular user's needs and interests.
- For example, *patterns describing the characteristics of customers who shop frequently at AllElectronics should be interesting to the marketing manager, but may be of little interest to other analysts studying the same database for patterns on employee performance.*
- Many patterns that are interesting by objective standards may represent common sense and, therefore, are actually uninteresting.

# Subjective interestingness measures

- Based on user beliefs in the data.
- These measures find patterns interesting if the patterns are **unexpected** (contradicting a user's belief) or offer strategic information on which the user can act(**actionable patterns**).
- For example, patterns like “a large earthquake often follows a cluster of small quakes” may be highly actionable if users can act on the information to save lives.
- Patterns that are **expected** can be interesting if they confirm a hypothesis that the user wishes to validate or they resemble a user's hunch.
- Eg: During a clinical trial for a new medication, researchers might expect the medication group to show improvement in certain symptoms compared to the placebo group. Observing this expected pattern strengthens the evidence for the medication's effectiveness.

# *Can a data mining system generate all of the interesting patterns?*

- Refers to the **completeness** of a data mining algorithm.
- It is often unrealistic and inefficient for data mining systems to generate all possible patterns.
- Instead, user provided constraints and interestingness measures should be used to focus the search.
- For some mining tasks, such as association, this is often sufficient to ensure the completeness of the algorithm.
- Association rule mining is an example where the use of constraints and interestingness measures can ensure the completeness of mining.

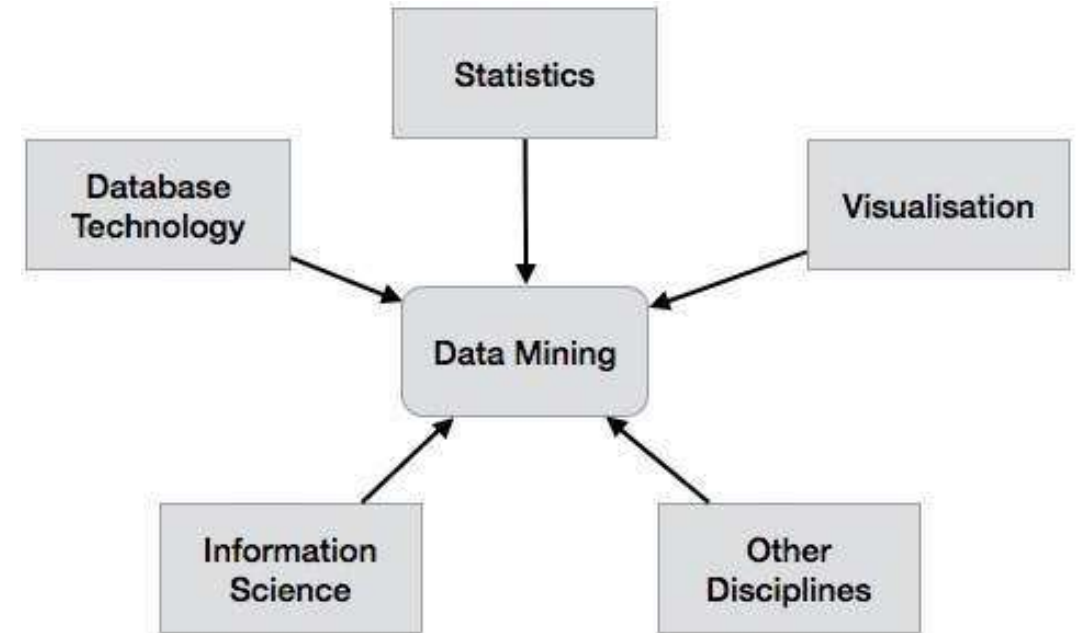
# *Can a data mining system generate only interesting patterns?*

- An optimization problem in data mining.
- It is highly desirable for data mining systems to generate only interesting patterns.
- Users and data mining systems would have to search through the patterns generated to identify the truly interesting ones.
- Progress made but optimization remains a challenging issue in data mining.
- **Measures of pattern interestingness** are essential for the efficient discovery of patterns by target users.
- Such measures can be used after the data mining step to **rank the discovered patterns according to their interestingness**, filtering out the uninteresting ones.
- Can be used to guide and constrain the discovery process, **improving the search efficiency by pruning away subsets of the pattern space** that do not satisfy pre-specified interestingness constraints.

# Data Mining System Classification

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

- Database Technology
- Statistics
- Machine Learning
- Information Science
- Visualization
- Other Disciplines



- Apart from these, a data mining system can also be classified based on the kind of
  - (a) databases mined
  - (b) knowledge mined
  - (c) techniques utilized
  - (d) applications adapted

# Classification Based on the Databases Mined

- We can classify a data mining system according to the kind of databases mined.
- Database system can be classified according to different criteria such as data models, types of data, etc.
- The data mining system can be classified accordingly.
- For example, *if we classify a database according to the data model, then we may have a relational, transactional, object-relational, or data warehouse mining system.*



# Classification Based on the kind of Knowledge Mined

- We can classify a data mining system according to the kind of knowledge mined.
- Data mining system is classified on the basis of functionalities such as –
  - Characterization
  - Discrimination
  - Association and Correlation Analysis
  - Classification
  - Prediction
  - Outlier Analysis
  - Evolution Analysis

# Classification Based on the Techniques Utilized

- We can classify a data mining system according to the kind of techniques used.
- We can describe these techniques according to the degree of user interaction involved or the methods of analysis employed.
- Machine learning, visualization, pattern recognition, neural networks, database-oriented or data-warehouse oriented techniques.
- Classification by User Interaction:
  - **Supervised Learning:** Decision Trees, Support Vector Machines (SVMs)
  - **Unsupervised Learning:** Clustering, Association Rule Learning
- Classification by Analysis Methods:
  - **Statistical Techniques:** Linear Regression, Logistic Regression
  - **Machine Learning Techniques:** Artificial Neural Networks (ANNs), Random Forests

# Classification Based on the Applications Adapted

- We can classify a data mining system according to the applications adapted.
- Eg:
  - Finance
  - Telecommunications
  - DNA
  - Stock Markets
  - E-mail