DATA MINING AND DATA WAREHOUSING

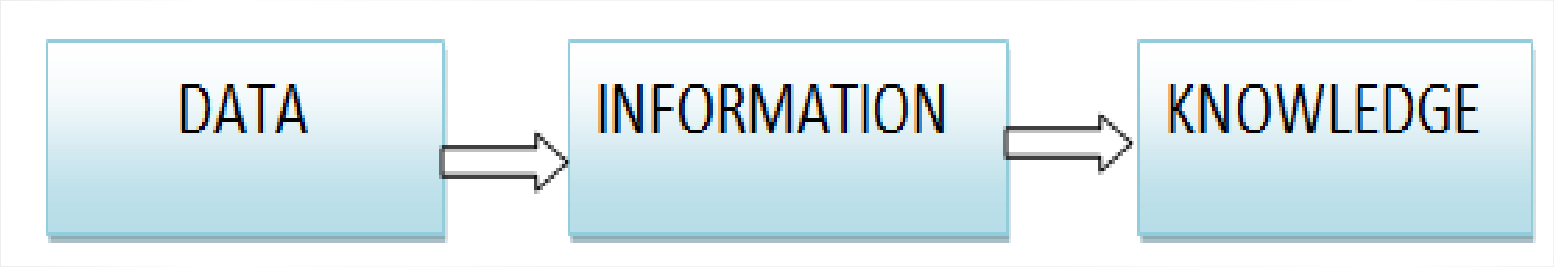
ARYA S V
Lecturer in Computer Science
School of Distance Education
University of Kerala

OVERVIEW

- Introduction
- Data Mining
- Data pre-processing
- Data Warehousing
- Data Cube
- OLAP
- Market Basket Analysis
- Association Rule
- Apriori Algorithm
- Classification vs Prediction
- Decision Tree
- Bayesian Classifier
- Lazy Classifier
- K-Nearest Neighbor method
- Rule based Classification
- Cluster Analysis
- Partition Methods
- K-means and K-medoids
- Outlier Detection

Introduction to Data Mining

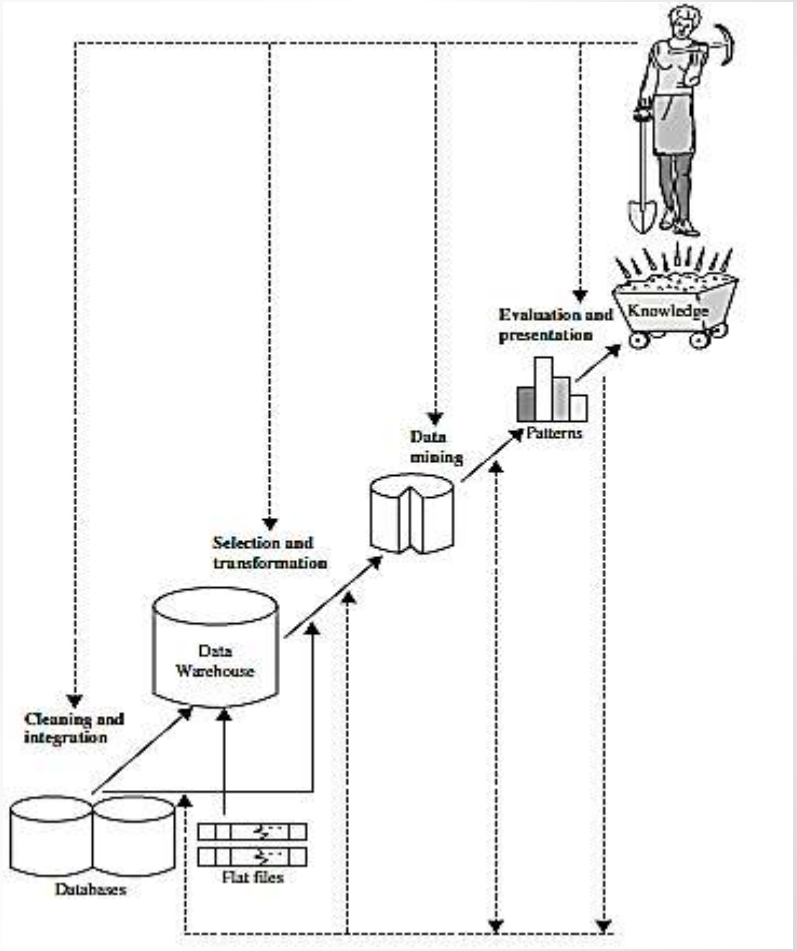
- **Data** is raw fact or disconnected fact.
- **Information** is the Processed data.
- **Knowledge** is derived from information by applying rules to it.



- **Data mining** is the process of extracting hidden, valid, and potentially useful patterns in huge data sets.
- Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data.

Steps in Data Mining

1. **Data cleaning** (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)



Types of Data for Mining

1. **Flat files** (The data for transactions, time-series data, scientific measurements, etc can be represented in these files.)
2. **database data** (Relational databases are one of the most commonly available and richest information repositories)
3. **data warehouse data** (A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.)
4. **transactional data.** (A transaction typically includes a unique transaction identity number (*trans ID*) and a list of the **items** making up the transaction, such as the items purchased in the transaction.)

Application Domains

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

Two highly successful and popular application examples of data mining:

- ✓ Business intelligence
- ✓ Search engines.

Data mining tasks can be classified into two categories:

- ✓ Descriptive mining tasks
- ✓ Predictive mining tasks.

Data Pre-Processing

Data Processing is a task of converting data from a given form to a much more usable and desired form.

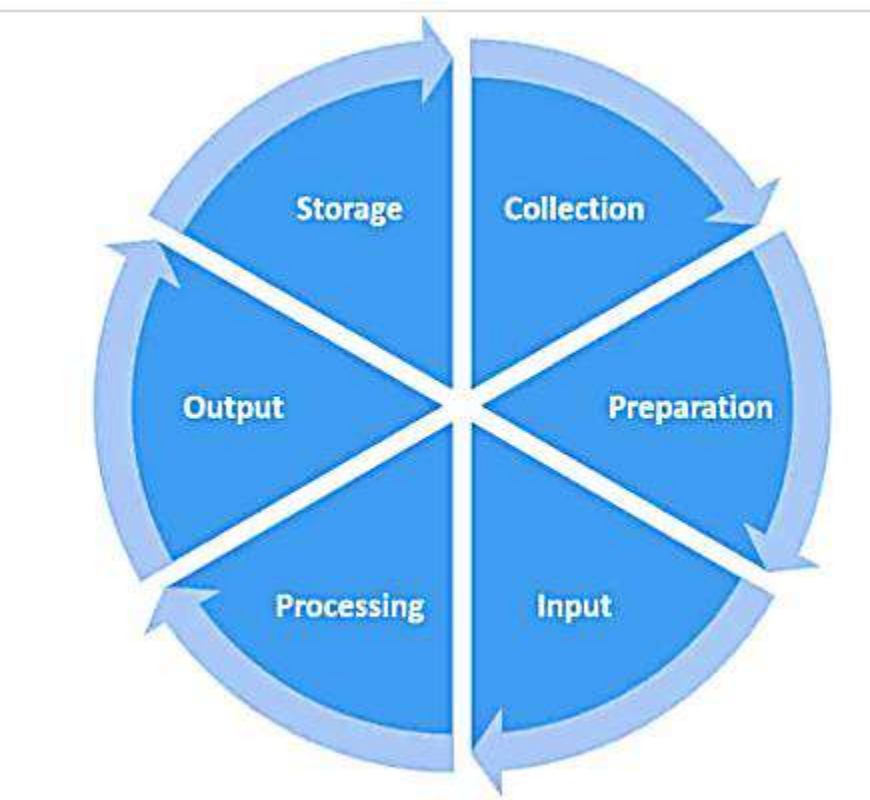
Why Preprocess the Data?

Data have quality if they satisfy the requirements of the intended use.

There are many factors comprising **data quality**:

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Interpretability

Stages of Data Processing Cycle



- ***Collection***
- ***Preparation***
- ***Input***
- ***Processing***
- ***Output***
- ***Storage***

Major tasks in Data Preprocessing

- **Data cleaning**

  ✓ Clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

- **Data integration**.

  ✓ Integrating multiple databases, data cubes, or files.

- **Data reduction Data reduction**

  ✓ Reduces the size of data and makes it suitable and feasible for analysis.

- **Data transformation**.

  ✓ Converting data from one format or structure into another format or structure.

Data Warehouse

- A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.
- Warehouse are the very large databases.

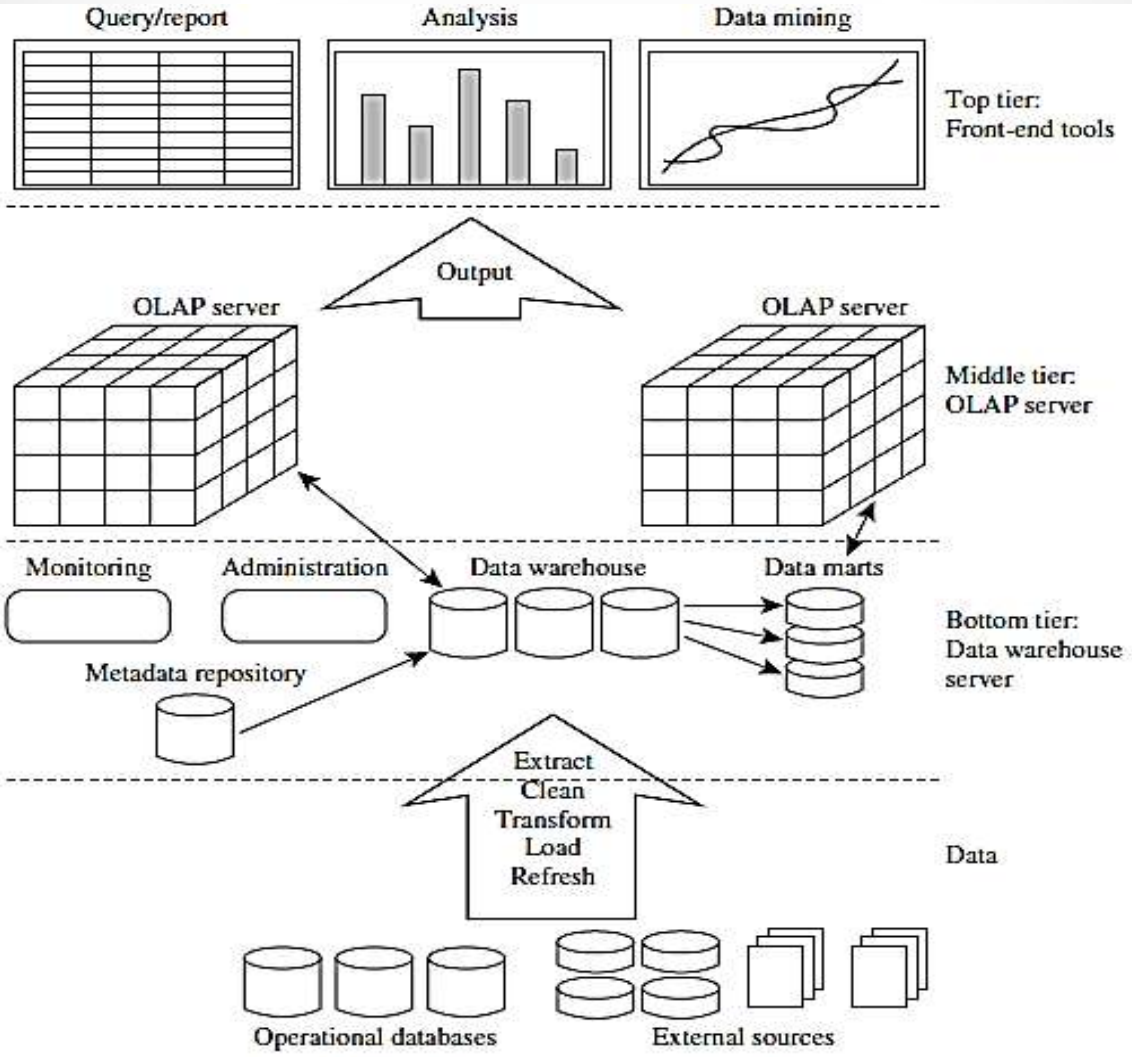
What is Data Warehousing?

- The process of transforming data into information and making it available to users in a timely enough manner to make a difference is known as data warehousing.
- "A data warehouse is a **subject-oriented**, **integrated**, **time-variant**, and **nonvolatile** collection of data in support of management's decision making process"
- Data warehouses provide **online analytical processing (OLAP)** tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data generalization and data mining.

Differences Between Database Systems And Data Warehouses

	OLAP	OLTP
Users and system orientation	An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.	OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals
Data contents	An OLAP system manages large amounts of historic data.	An OLTP system manages current data.
Database design	typically adopts either a star or a snowflake model and a subject -oriented database design.	An OLTP system usually adopts an entity-relationship (ER) data model and an application -oriented database design.
View	An OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization	An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historic data or data in different organizations
Access patterns	Accesses to OLAP systems are mostly read-only operations	The access patterns of an OLTP system consist mainly of short, atomic transactions.

Data Warehouse Architecture

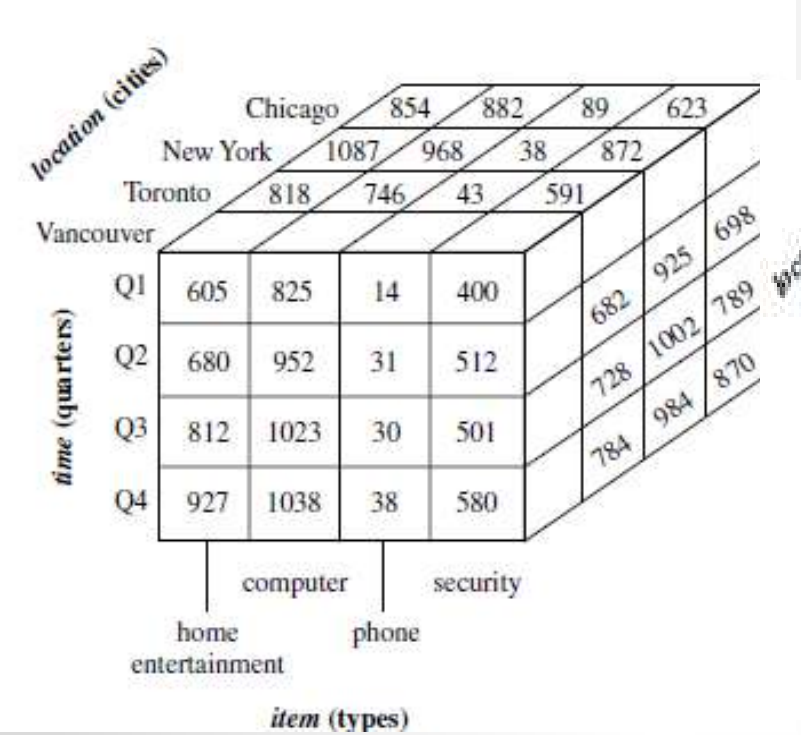


A three-tier data warehousing architecture.

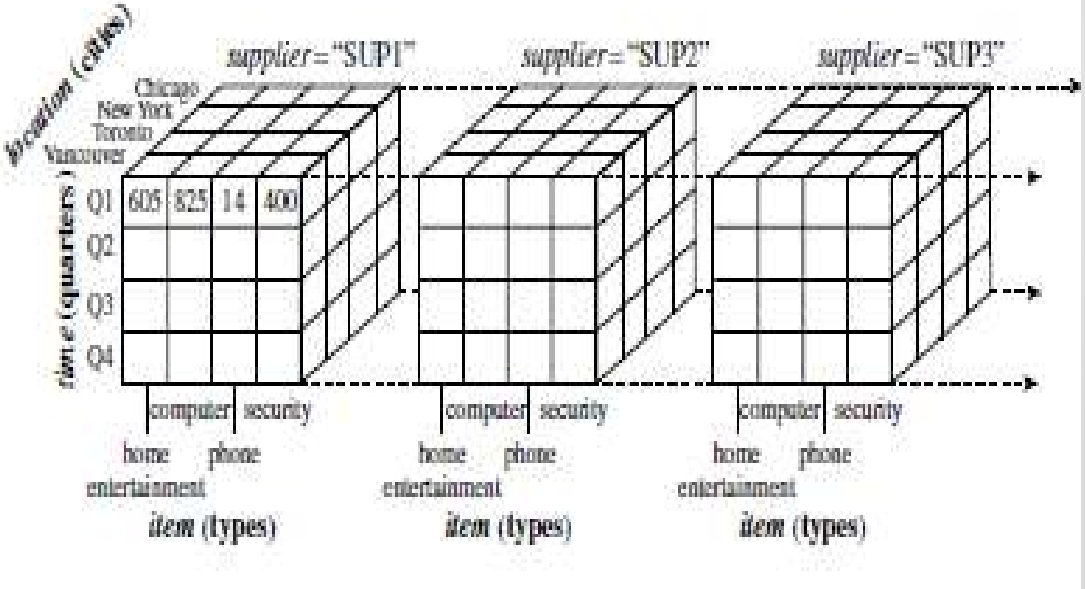
Data Cube

- Data warehouses and OLAP tools are based on a **multidimensional data model**. This model views data in the form of a data cube.
- A **data cube** allows data to be modeled and viewed in multiple dimensions.
- It is defined by dimensions and facts.
- In general terms, **dimensions** are the perspectives or entities with respect to which an organization wants to keep records. Each dimension may have a table associated with it, called a **dimension table.**
- **Facts** are numeric measures. The **fact table** contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

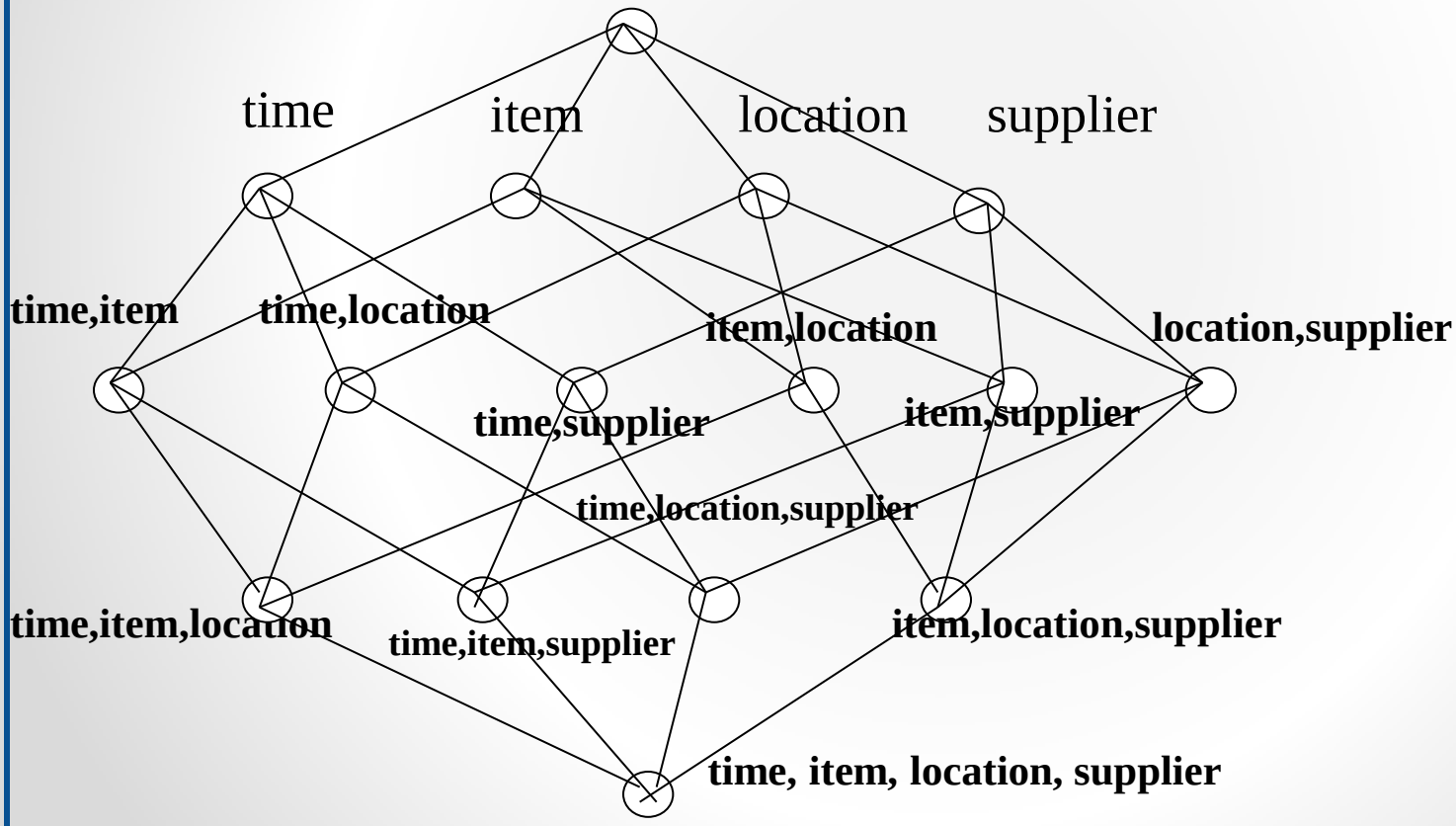
A 3-D data cube representation of the data according to time, item, and location. Here the measure displayed is dollars_sold (in thousands)



A 4-D data cube representation of sales data, according to time, item, location, and supplier. The measure displayed is dollars sold (in thousands).(only some of the cube values are shown.)



Lattice of Cuboids, making up a 4-D Data Cube For Time, Item, Location, And Supplier.



0-D(apex) cuboid

1-D cuboids

2-D cuboids

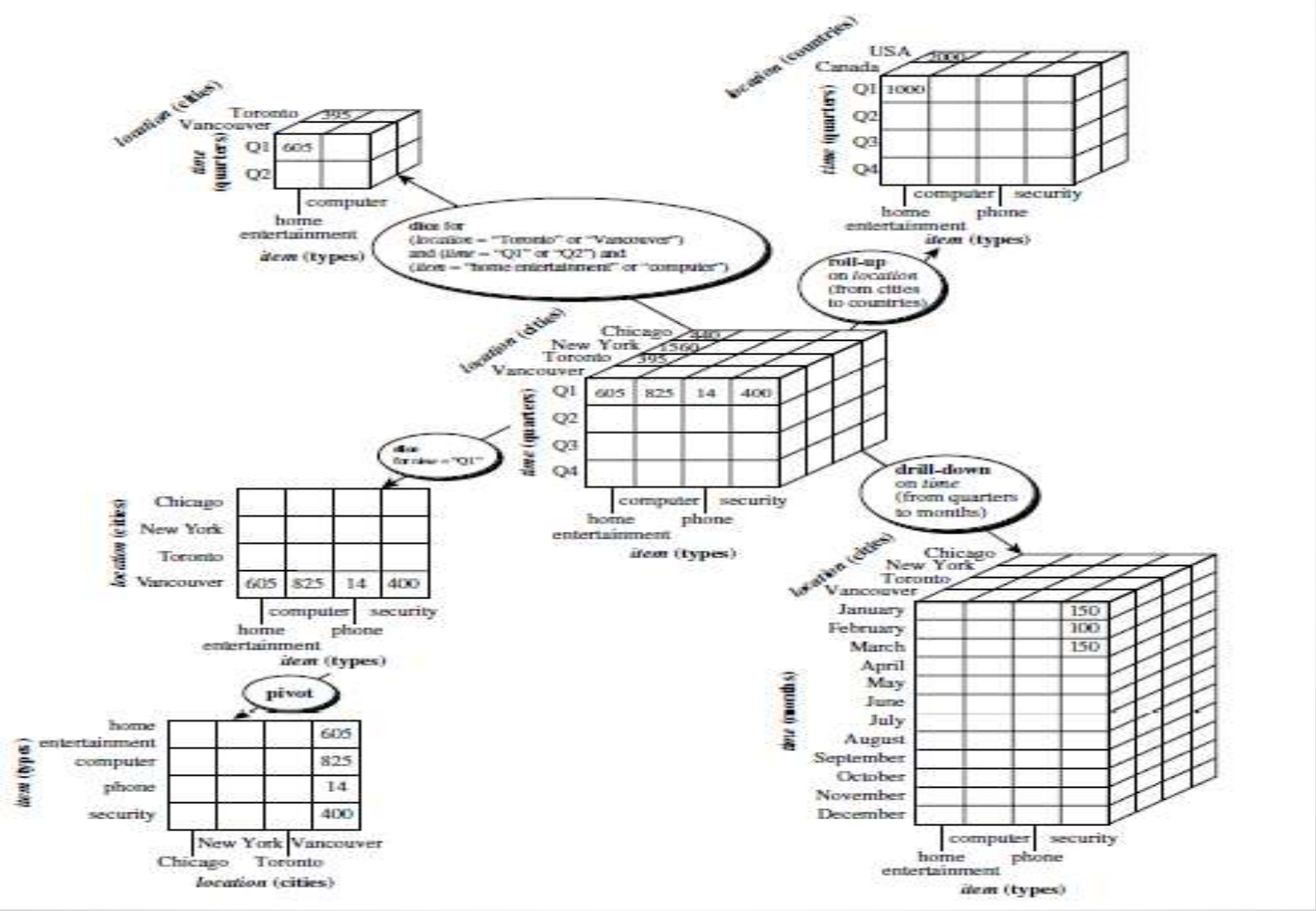
3-D cuboids

4-D(base) cuboid

OLAP
(Online Analytical Processing)

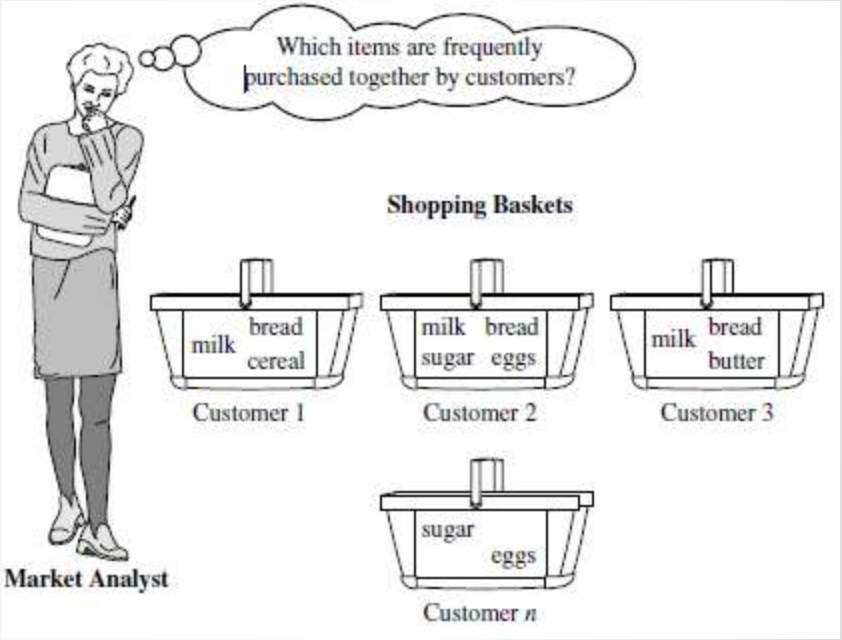
- OLAP provides a user-friendly environment for interactive data analysis.
- An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.
- An OLAP system manages large amounts of historic data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity.
- OLAP systems can organize and present data in various formats in order to accommodate the diverse needs of different users.

Examples of Typical OLAP Operations on Multidimensional Data.



Market Basket Analysis

- **Market basket analysis** analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets"
- The discovery of these associations can help retailers develop marketing strategies by analyzing which items are frequently purchased together by customers.



Association Rule

- Association rule mining finds interesting associations and relationships among large sets of data items.
- This rule shows how frequently a item set occurs in a transaction. A typical example is Market Based Analysis.
- Market Based Analysis is one of the key techniques used by large relations to show associations between items.
- It allows retailers to identify relationships between the items that people buy together frequently.

The basic definitions:

- **Support Count() –** Frequency of occurrence of a itemset.
- **Frequent Item set –** An item set whose support is greater than or equal to minsup threshold.
- **Association Rule –** An implication expression of the form X -> Y, where X and Y are any 2 item sets.

Apriori Algorithm

- It uses prior knowledge of frequent item set properties.
- We apply an iterative approach or level-wise search where k-frequent item sets are used to find k+1 item sets.
- To improve the efficiency of level-wise generation of frequent item sets, an important property is used called Apriori property which helps by reducing the search space.
- **Apriori Property**

  All non-empty subset of frequent item set must be frequent. The key concept of Apriori algorithm is its anti-monotonicity of support measure.

Apriori Algorithm Steps

1. Scan the transaction data base to get the support 'S' each 1-itemset, compare 'S' with min_sup, and get a support of 1-itemsets,
2. Use join to generate a set of candidate k-item set. Use apriori property to prune the unfrequented k-item sets from this set.
3. Scan the transaction database to get the support 'S' of each candidate k-item set in the given set, compare 'S' with min_sup, and get a set of frequent k-item set
4. If the candidate set is NULL, for each frequent item set 1, generate all nonempty subsets of 1.
5. For every nonempty subsets of 1, output the rule "s=>(1-s)" if confidence C of the rule "s=>(1-s)" min_conf
6. If the candidate set is not NULL, go to step 2.

Classification vs Prediction

- Classification is the process of finding a model that describes and distinguishes data classes and concepts.
- Classification is the problem of identifying on the basis of a training set of data containing observations and whose categories membership is known.
- It is a two-step process,
  - ✓Learning step (where a classification model is constructed)
  - ✓Classification step (where the model is used to predict class labels for given data).
- A medical researcher wants to analyze breast cancer data to predict which one of three specific treatments a patient should receive.

Classification vs Prediction(cont..)

- Here data analysis task is **classification**,

where a model or **classifier** is constructed to predict class (categorical) labels, such as "safe" or "risky" for the loan application data;

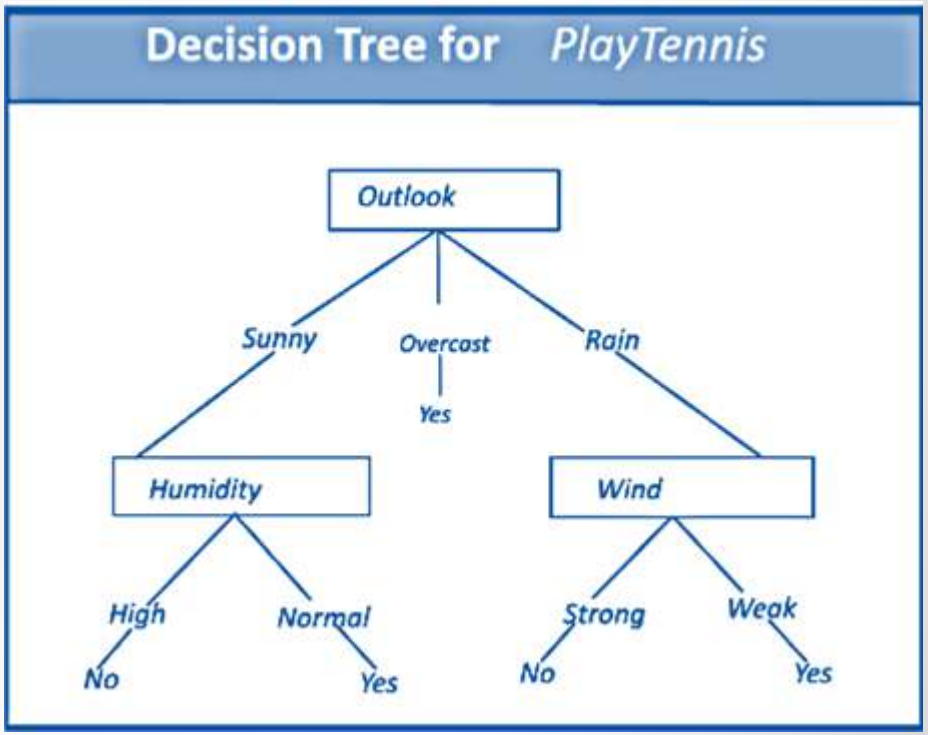
"yes" or "no" for the marketing data;

"treatment A," "treatment B," or "treatment C" for the medical data.

- Suppose that the marketing manager wants to predict how much a given customer will spend during a sale at a shop.
- This data analysis task is an example of **numeric prediction**, where the model constructed predicts a continuous-valued function, or ordered value, as opposed to a class label. This model is a **predictor**.

Decision Trees for Classification

- Decision tree is the most powerful and popular tool for classification and prediction.
- A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.
- The topmost node in a tree is the **root** node.



Naive Bayesian Classifier

- Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle,

  i.e. every pair of features being classified is independent of each other.

- Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where A and B are events.

Lazy Learners

- Lazy Learners are the most intuitive type of learners and are used in many practical scenarios.
- The process of modeling the training data until it is needed to classify the testing data. Techniques that employ this strategy are known as Lazy Learners.
- A lazy learner simply stores the training data and only when it sees a test tuple starts generalization to classify the tuple based on its similarity to the stored training tuples.
- Lazy learners do less work while training data is given and more work when classification of a test tuple is given.
- Lazy learners can be computationally very expensive while doing classification or predictions which do not require any model building.

k–Nearest Neighbor Method

- Mainly used to finds intense application in pattern recognition, data mining and intrusion detection.
- It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data

Algorithm

Let m be the number of training data samples. Let p be an unknown point.

1. Store the training samples in an array of data points arr[]. This means each element of this array represents a tuple (x, y).
2. for i=0 to m: Calculate Euclidean distance d(arr[i], p).
3. Make set S of K smallest distances obtained. Each of these distances corresponds to an already classified data point.
4. Return the majority label among S.

Rule-Based Classification

- Rule-based classifier makes use of a set of IF-THEN rules for classification. We can express a rule in the following from –

  Let us consider a rule R1,

  R1: IF age = youth AND student = yes THEN buy_computer = yes

Points to remember

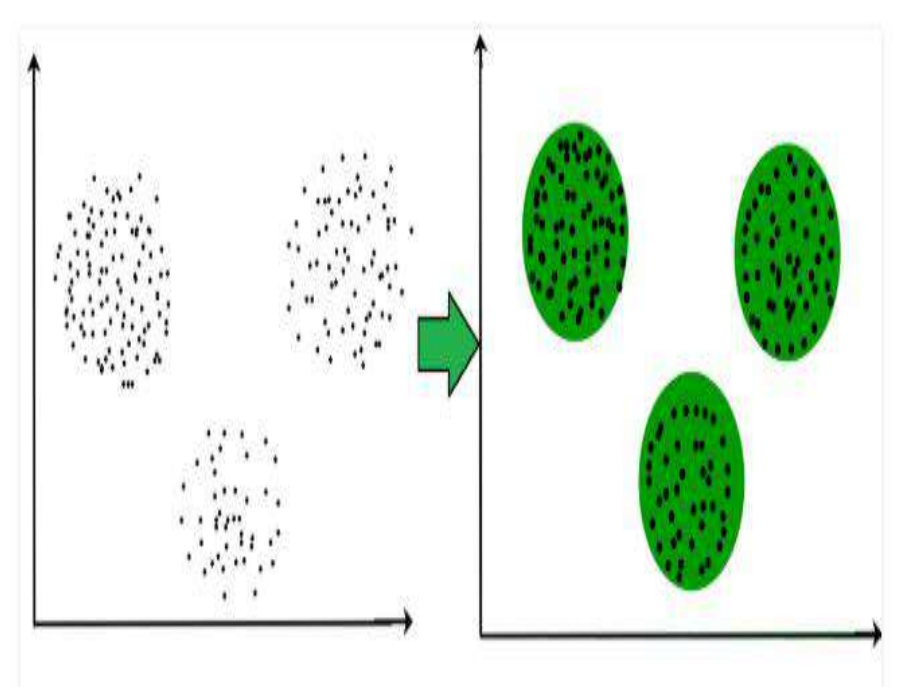
- The IF part of the rule is called **rule antecedent** or **precondition**.
- The THEN part of the rule is called **rule consequent**.
- The antecedent part the condition consist of one or more attribute tests and these tests are logically ANDed.
- The consequent part consists of class prediction.

Cluster Analysis

- **Clustering** is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters.
- **Cluster analysis** or simply clustering is the process of partitioning a set of data objects (or observations) into subsets.
- Each subset is a **cluster**, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as **clustering**.
- Clustering is also called **data segmentation** in some applications because clustering partitions large data sets into groups according to their similarity

- Clustering can also be used for **outlier detection**, where outliers (values that are "far away" from any cluster) may be more interesting than common cases.
- Clustering is known as **unsupervised learning** because the class label information is not present.

The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



Requirements for Cluster Analysis

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Requirements for domain knowledge to determine input parameters
- Ability to deal with noisy data
- Incremental clustering and insensitivity to input order
- Capability of clustering high-dimensionality data
- Constraint-based clustering
- Interpretability and usability

Partitioning Methods

- The simplest and most fundamental version of cluster analysis is partitioning, which organizes the objects of a set into several exclusive groups or clusters.
- The general criterion of a good partitioning is that objects in the same cluster are "close" or related to each other, whereas objects in different clusters are "far apart" or very different.
- General Characteristics of Partitioning methods
  - ✓ Find mutually exclusive clusters of spherical shape
  - ✓ Distance-based
  - ✓ May use mean or medoid (etc.) to represent cluster center
  - ✓ Effective for small- to medium-size data sets

K-Means: A Centroid-Based Technique

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

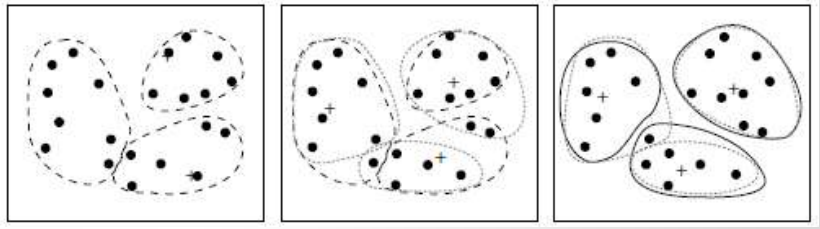
(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;

(2) **repeat**

(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;

(5) **until** no change;



k-Medoids

- It is also called as Partitioning Around Medoid algorithm.
- A medoid can be defined as the point in the cluster, whose dissimilarities with all the other points in the cluster is minimum.
- The dissimilarity of the medoid(Ci) and object(Pi) is calculated by using $E = |Pi - Ci|$

Algorithm:

1. Initialize: select k random points out of the n data points as the medoids.
2. Associate each data point to the closest medoid by using any common distance metric methods.
3. While the cost decreases:

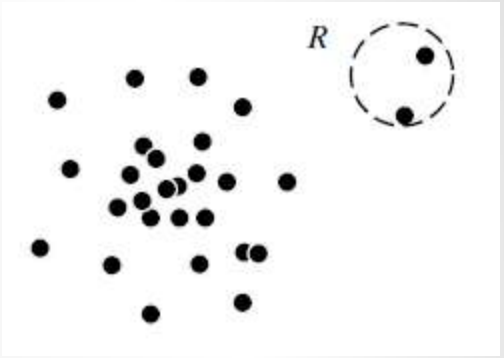
For each medoid m, for each data o point which is not a medoid:

1. Swap m and o, associate each data point to the closest medoid, recompute the cost.
2. If the total cost is more than that in the previous step, undo the swap.

Outlier Detection in Clustering

- An **outlier** is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism.
- Outliers are referred as "abnormal" data.
- Outliers are different from noisy data. Noise is a random error or variance in a measured variable.
- Outliers are interesting because they are suspected of not being generated by the same mechanisms as the rest of the data.
- Outlier detection is also related to novelty detection in evolving data sets.

The objects in region R are outliers



Types of Outliers

✓ Global Outliers
✓ Contextual Outliers
✓ Collective Outliers

Outlier detection techniques.

Supervised, Semi-Supervised, and Unsupervised Methods

- Supervised methods model data normality and abnormality.
- In some application scenarios, objects labeled as "normal" or "outlier" are not available. Thus, an unsupervised learning method has to be used.
- In some cases where only a small set of the normal and/or outlier objects are labeled, but most of the data are unlabeled.

Statistical Methods, Proximity-Based Methods and Clustering-Based Methods

- Statistical methods (also known as model-based methods) make assumptions of data normality.
- The effectiveness of proximity-based methods relies heavily on the proximity (or distance) measure used.
- Clustering-based methods assume that the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters.

Thank You