RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

CST 466 Data Mining

Module 1: Introduction to Data Mining and Data Warehousing

Textbooks

1. Dunham M H, “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, 2003.

2. Arun K Pujari, “Data Mining Techniques”, Universities Press Private Limited, 2008.

3. Jaiwei Han and Micheline Kamber, “Data Mining Concepts and Techniques”, Elsevier, 2006

Reference Books

1. M Sudeep Elayidom, “Data Mining and Warehousing”, 1st Edition, 2015, Cengage Learning India Pvt. Ltd.

2. MehmedKantardzic, “Data Mining Concepts, Methods and Algorithms”, John Wiley and Sons, USA, 2003.

3. Pang-Ning Tan and Michael Steinbach, “Introduction to Data Mining”, Addison Wesley, 2006.

Syllabus

- Data warehouse-Differences between Operational Database Systems and Data Warehouses
- Multidimensional data model- Warehouse schema
- OLAP Operations
- Data Warehouse Architecture
- Data Warehousing to Data Mining, Data Mining Concepts and Applications,
- Knowledge Discovery in Database Vs Data mining
- Architecture of typical data mining system
- Data Mining Functionalities, Data Mining Issues.

Data Mining

- **Definition:**
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns
- Finding hidden information in a database
- Fit data to a model

Data Mining System Components

A typical data mining system may have the following components:

Data base/warehouse or other information repositories :- Containing a set of operational/historic data, or other related information.

Database/warehouse server:- Which extracts the data relevant to the data mining requests.

Knowledge base:- Containing domain knowledge in the form of metadata.

Data Mining (DM) engine:- A set of functional modules for tasks, such as classification, prediction, regression, cluster analysis, etc.

Pattern Evaluation module:- Works in tandem with the DM modules, by employing interestingness measures to help focus the search towards the required patterns.

GUI:- Facilitates interactive data mining and knowledge discovery.

Data Warehouse

- **Database** is a collection of related **data** that represents some elements of the real world.
- **Database** is used to perform online transaction and query processing.
- **Data warehouse** is an information system that stores historical and commutative **data** from single or multiple sources.
- **Database** is designed to **record data** whereas the **Data warehouse** is designed to **analyze data and make decisions**.
- Data warehouse is designed to **analyze**, **report**, **integrate** transaction data from different sources.

Data Warehouse

- **Data warehouse** refers to a data repository that is maintained separately from an organization’s operational databases.

- Data warehouse systems allow for integration of a variety of application systems.

- They support information processing by providing a solid platform of consolidated historic data for analysis.

Data Warehouse

A data warehouse is a **subject oriented**, **integrated**, **time variant** and **non-volatile** collection of data in support of the **management decision making** process.

Subject-oriented: A data warehouse is organized around major subjects such as customer, supplier, product, and sales.

Integrated: A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and online transaction records.

Time-variant: Data are stored to provide information from a historic perspective (e.g., the past 5–10 years).

Non-volatile: A data warehouse does not require transaction processing, recovery and concurrency control mechanisms. It usually requires only two operations in data accessing: *initial loading of data* and *access of data*.

Data Warehousing

Data warehousing is the process of constructing and using data warehouses.

The construction of a data warehouse requires: **data cleaning, data integration, and data consolidation.**

Organizations use this information to support **business decision-making activities** like:

- increasing customer focus, which includes the analysis of customer buying patterns
- repositioning products and managing product portfolios by comparing the performance of sales by quarter, year, geographic regions etc.
- analyzing operations and looking for sources of profit
- managing customer relationships, making environmental corrections, and managing the cost of corporate assets

Differences between Operational Database Systems and Data Warehouses

- The major task of online operational database systems is to perform **online transaction**
- **and query processing.**

- These systems are **called online transaction processing (OLTP) systems**.

- They cover most of the day-to-day operations of an organization such as:
    - **Purchasing**
    - **Inventory**
    - **Manufacturing**
    - **Banking**
    - **Payroll**
    - **Registration**
    - **Accounting**

Differences between Operational Database Systems and Data Warehouses

- Data warehouse systems serve users or knowledge workers in the role of data analysis and decision making.

- Such systems can organize and present data in various formats in order to accommodate the diverse needs of different users.

- These systems are known as **online analytical processing (OLAP) systems.**

Differences between Operational Database Systems and Data Warehouses

- Distinguishing features of OLTP and OLAP are summarized as follows:

- **Users and system orientation:** An OLTP system is **customer-oriented** and is used for transaction and query processing by **clerks, clients, and information technology professionals.** An OLAP system is **market-oriented** and is used for data analysis by **knowledge workers, including managers, executives, and analysts**.

- **Data contents:** An OLTP system **manages current data** that, typically, are too detailed to be easily used for decision making. 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. These features make the data easier to use for informed decision making.

Differences between Operational Database Systems and Data Warehouses

- **Database design**: An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design. An OLAP system typically adopts either a star or a snowflake model and a subject-oriented database design.

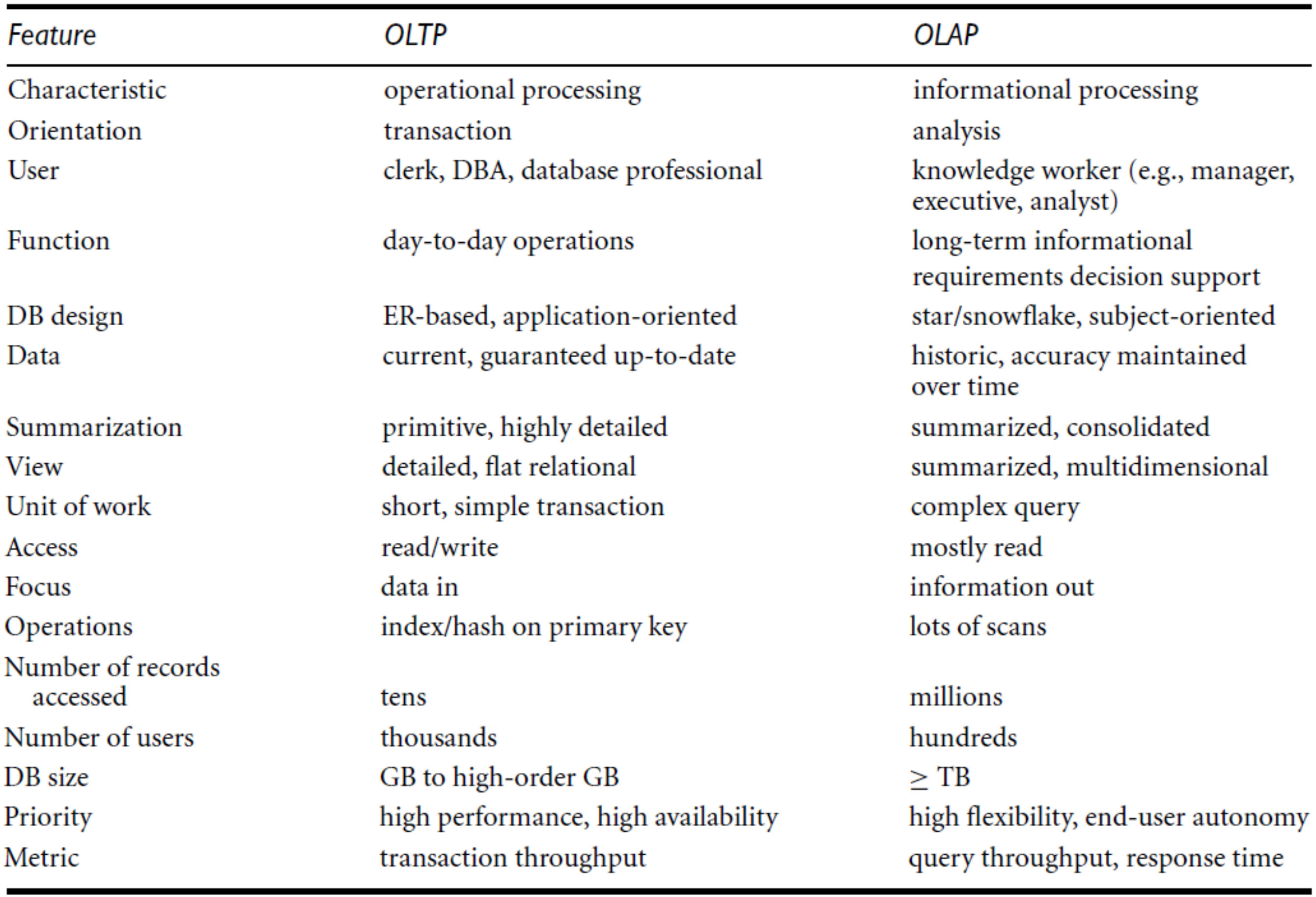
- **View:** An OLTP system focuses mainly on the **current data** within an enterprise or department, without referring to historic data or data in different organizations. In contrast, an OLAP system often spans **multiple versions of a database schema**, due to the evolutionary process of an organization. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.

Differences between Operational Database Systems and Data Warehouses

- **Access patterns:** The access patterns of an OLTP system consist mainly of **short, atomic transactions.** Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read-only operations (because most data warehouses store historic rather than up-to-date information), although many could be complex queries.

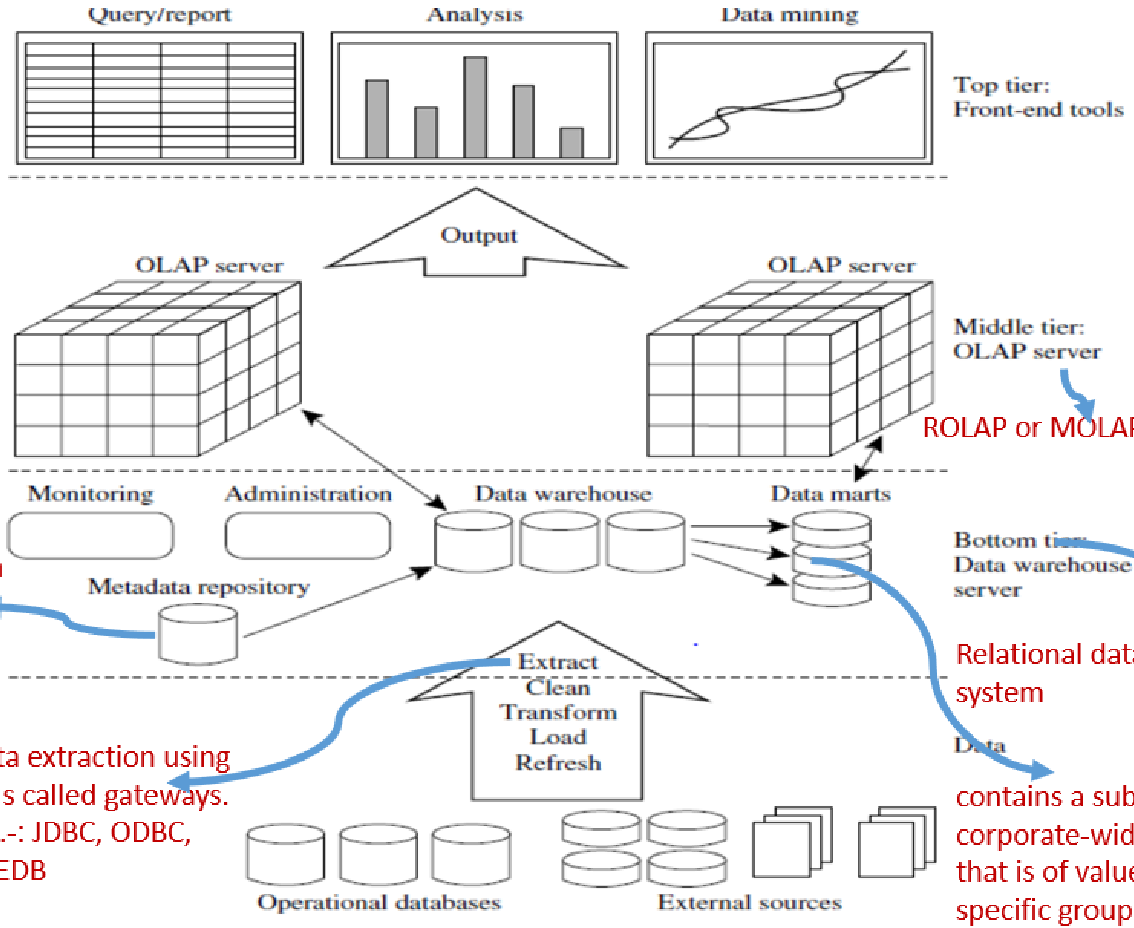
- Other features that distinguish between OLTP and OLAP systems include
- **Database size**
- **Frequency of operations**
- **Performance metrics.**

Differences between Operational Database Systems and Data Warehouses



Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements decision support
DB design	ER-based, application-oriented	star/snowflake, subject-oriented
Data	current, guaranteed up-to-date	historic, accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	GB to high-order GB	≥ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

3 Tier Data Warehousing Architecture



ROLAP or MOLAP

Stores information about the data warehouse and its contents.

Relational data base system

Data extraction using APIs called gateways. E.g.-: JDBC, ODBC, OLEDB

contains a subset of corporate-wide data that is of value to a specific group of users

Data Warehousing: A Multitiered Architecture

- The bottom tier is a **warehouse database server** that is almost always a relational database system.

- Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (e.g., customer profile information provided by external consultants).

- These tools and utilities perform **data extraction, cleaning, and transformation** (e.g., to merge similar data from different sources into a unified format), as well as **load and refresh functions** to update the data warehouse. The data are extracted using application program interfaces known as **gateways.**

- This tier also contains a metadata repository, which stores information about the data warehouse and its contents.

Data Warehousing: A Multitiered Architecture

- The middle tier is an OLAP server that is typically implemented using either:
- **a relational OLAP (ROLAP) model** (i.e., an extended relational DBMS that maps operations on multidimensional data to standard relational operations)
- or a **multidimensional OLAP (MOLAP) model** (i.e., a special-purpose server that directly implements multidimensional data and operations).

- The top tier is a front-end client layer, which contains **query and reporting tools, analysis tools, and/or data mining tools** (e.g., trend analysis, prediction, and so on).

Multidimensional Data Model

- 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.
- **Dimensions** are the perspectives or entities with respect to which an organization wants to keep records.
- Eg: - Data Warehouse for sales of company X have dimensions: **time, item, branch, and location.**

Multidimensional Data Model

- Each dimension may have a table associated with it, called a **dimension table**, which further describes the dimension.

- Eg:- Item is described by the attributes: **item name, brand, and type.**

- A multidimensional data model is typically organized around a central theme (or subject), such as sales.

- This theme is represented by a **fact table.**

Multidimensional Data Model

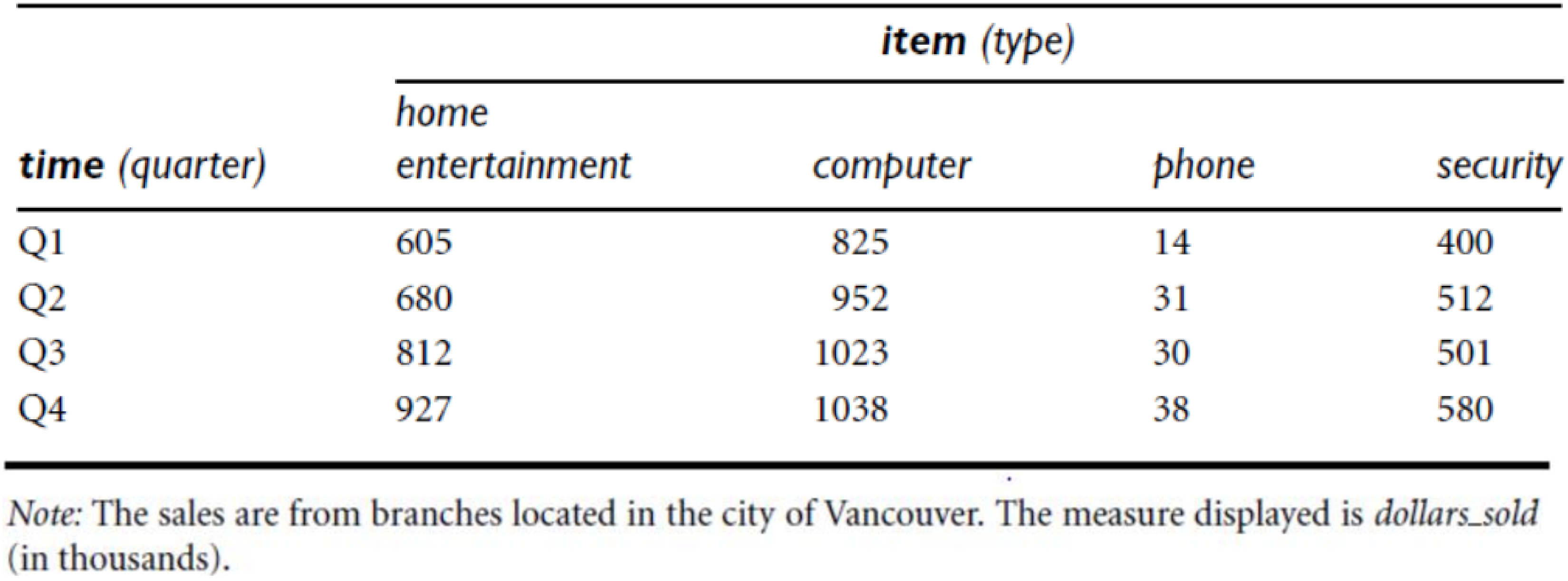
- **Facts are numeric measures.**

- Eg:- For a sales data warehouse, facts include **dollars sold (sales amount in dollars), units sold (number of units sold), and amount budgeted.**

- The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

- In the context of data warehousing, the data cube is **n-dimensional.**

Activity: Identify the dimensions and facts...



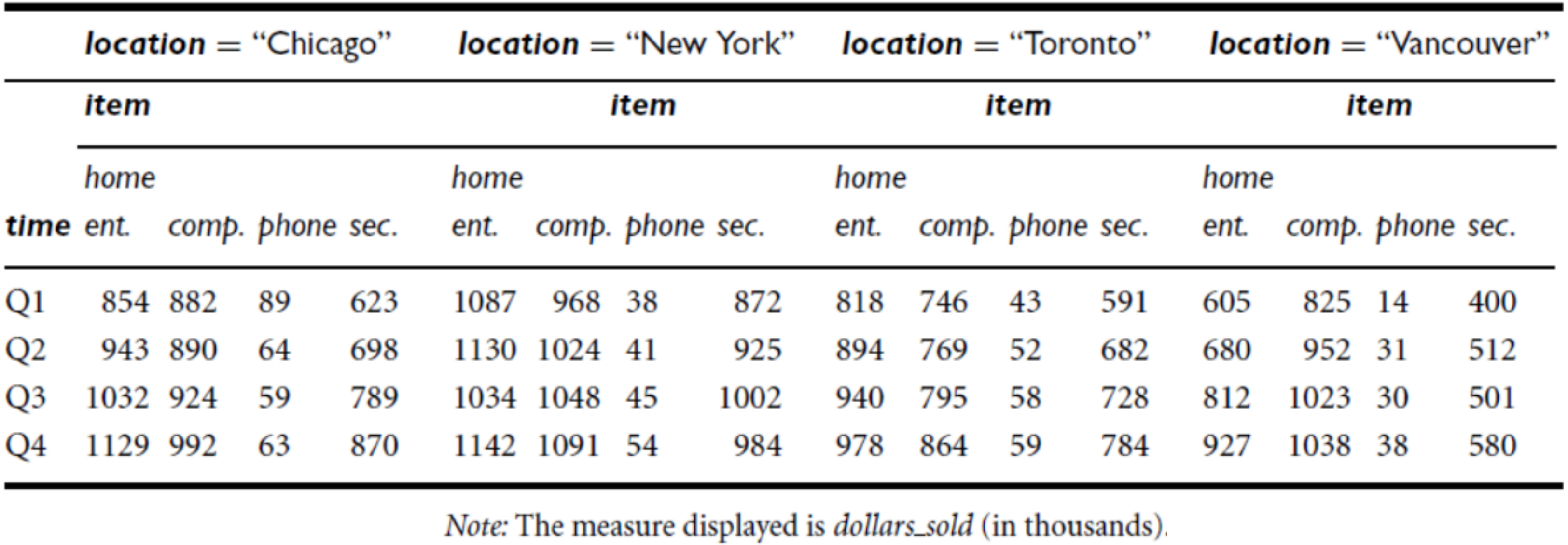
	item *(type)*			
time *(quarter)*	*home entertainment*	*computer*	*phone*	*security*
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

Note: The sales are from branches located in the city of Vancouver. The measure displayed is *dollars_sold* (in thousands).

Dimensions: **item, time**

Fact: **dollars_sold**

Activity: Identify the dimensions and facts...



	location = "Chicago"				***location*** = "New York"				***location*** = "Toronto"				***location*** = "Vancouver"			
	item				***item***				***item***				***item***			
	home				*home*				*home*				*home*			
time	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

Note: The measure displayed is *dollars_sold* (in thousands).

Dimensions: **item, time, location**

Fact: **dollars_sold**

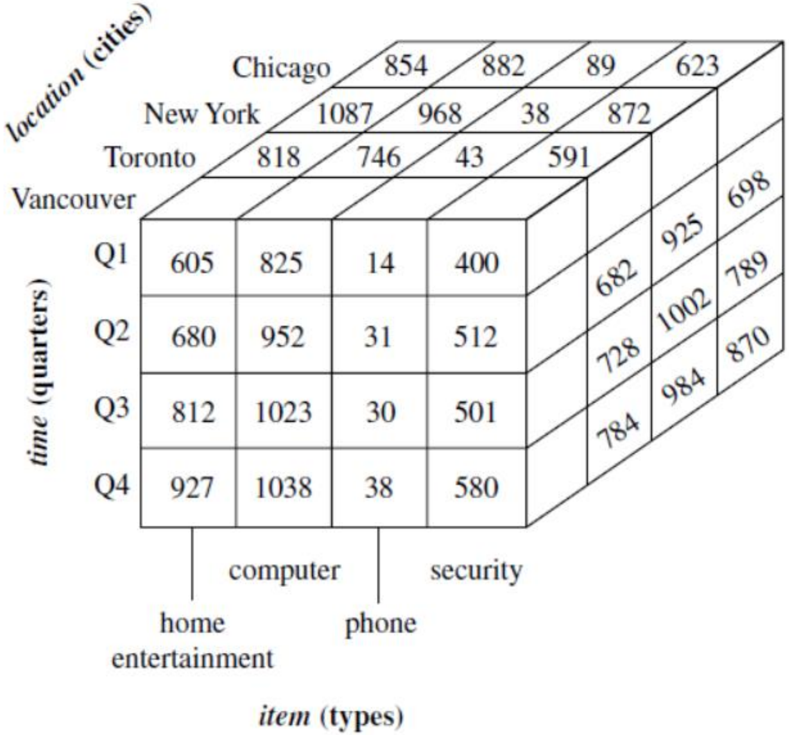
Data Cube : 3D

- In the multidimensional model, data are organized into **multiple dimensions**, and **each dimension** contains **multiple levels of abstraction** defined by **concept hierarchies**.

- The cube contains the dimensions *location, time*, and *item.*

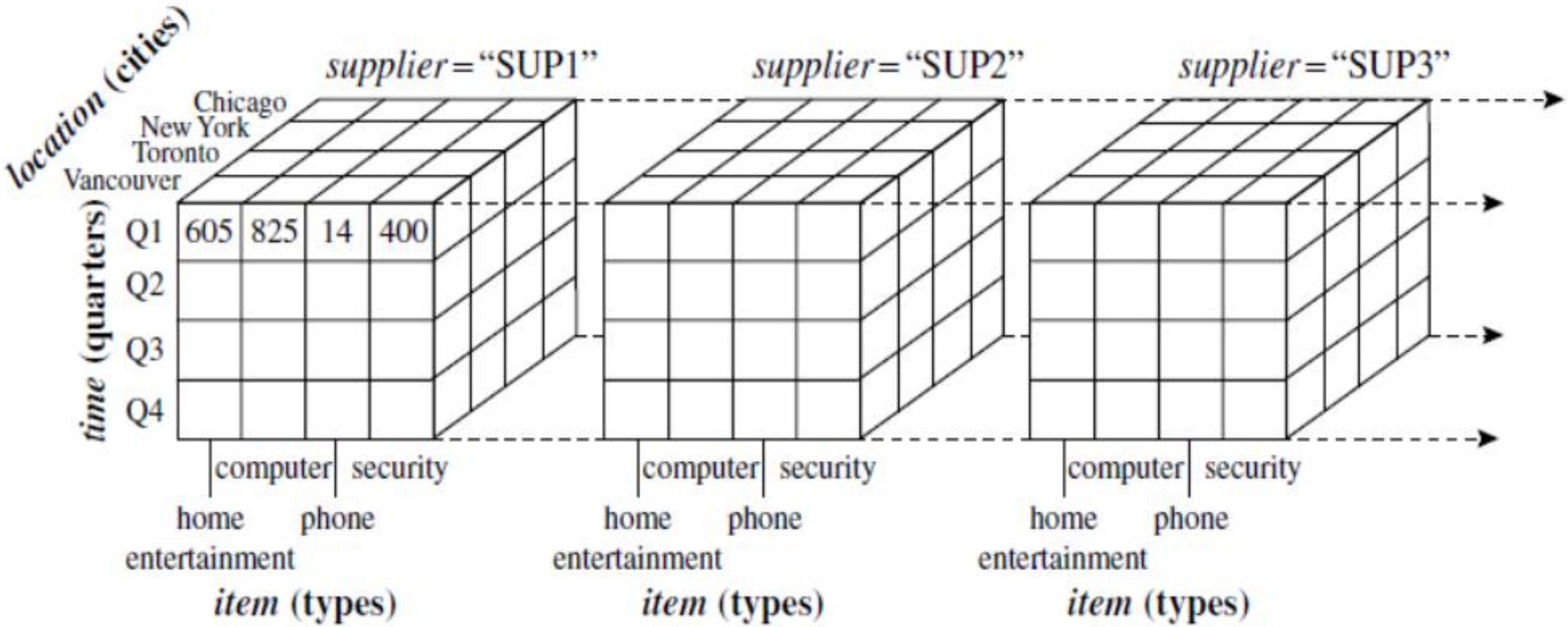
- *location* is aggregated with respect to city values.

- Each city can be mapped to the higher level concept - district (province) , which in turn can be mapped to country.



Vancouver	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

Data Cube : 4D



location (cities)
Chicago
New York
Toronto
Vancouver
supplier = "SUP1"
supplier = "SUP2"
supplier = "SUP3"
time (quarters)
Q1
Q2
Q3
Q4
605
825
14
400
home entertainment
computer
phone
security
item (types)

Warehouse Schema

- The popular data model for a data warehouse is multidimensional model, which can exist in the form of a **star schema**, a **snowflake schema**, or a **fact constellation schema**.

- **Star schema:** The most common modeling paradigm is the star schema, in which the data warehouse contains:
  1. a large central table (**fact table**) containing the bulk of the data, **with no redundancy.**
  2. a set of smaller attendant tables (**dimension tables**), one for each dimension.

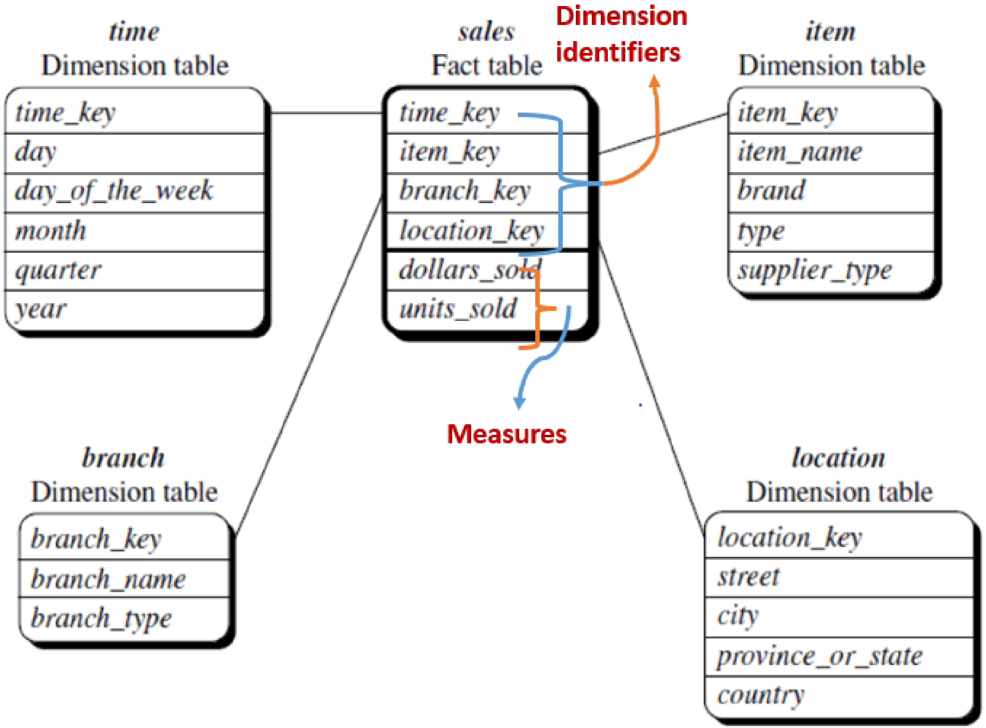
- The schema graph resembles a starburst, with the dimension table displayed in a radial pattern around the central fact table.

Star Schema

- In the star schema, each dimension is represented by only one table, and each table contains a set of attributes.
- *Location* dimension table contains the attribute set (*location_key, street, city, province or state, country*).

- This introduces redundancy.
- Eg:-

(...., Kakkanad, Cochin, Kerala, India),
(...., Edappally, Cochin, Kerala, India),
....



Snowflake Schema

- The **snowflake schema** is a variant of the star schema model, where **some dimension tables are *normalized***, thereby further splitting the data into **additional tables**.

- Such a table is easy to maintain and saves storage space.

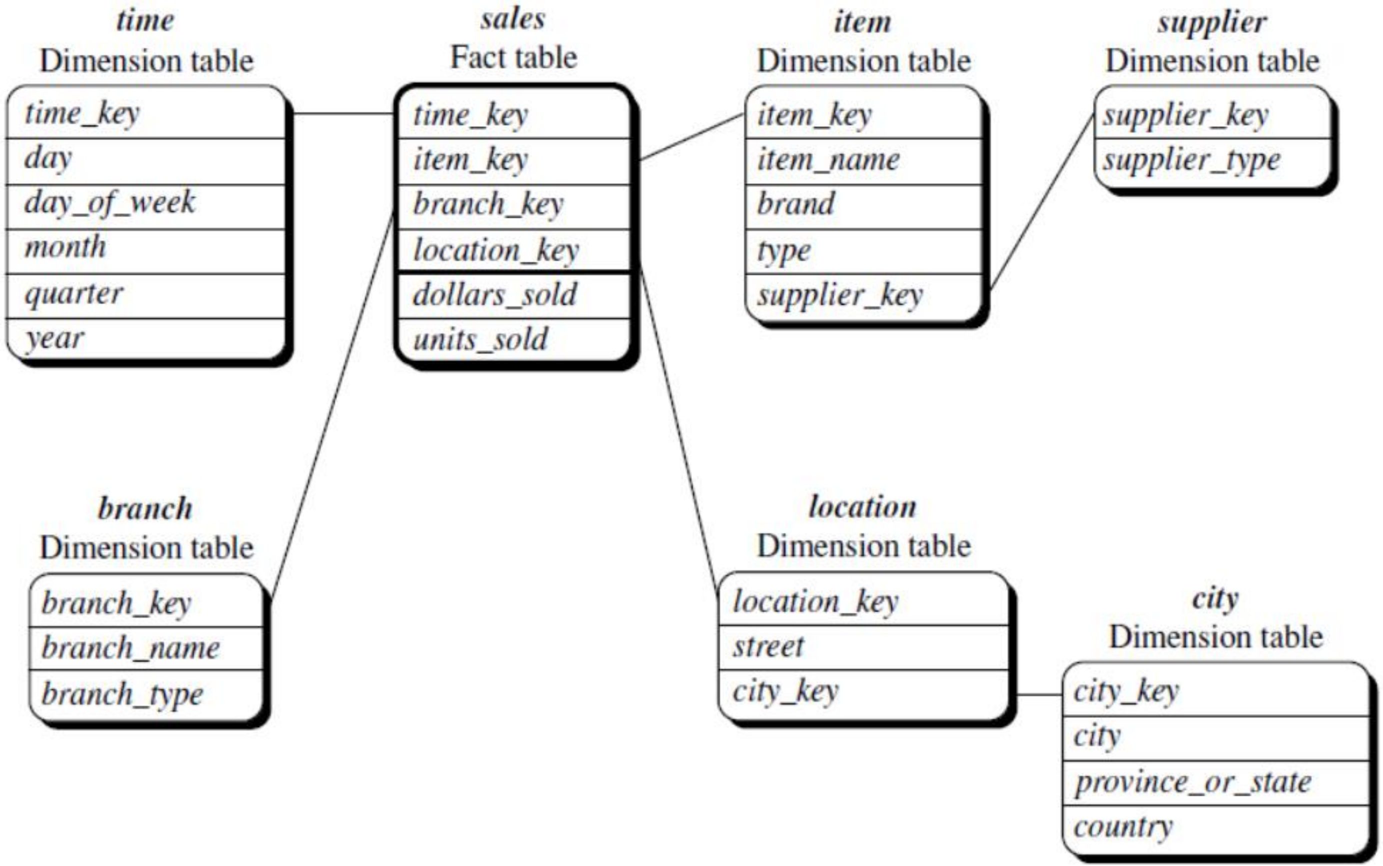
- However, this space savings is negligible in comparison to the typical magnitude of the fact table.

- Furthermore, the snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query.

- Consequently, the system performance may be adversely impacted.

- Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.

Snowflake Schema



time
Dimension table

time_key
day
day_of_week
month
quarter
year

sales
Fact table

time_key
item_key
branch_key
location_key
dollars_sold
units_sold

item
Dimension table

item_key
item_name
brand
type
supplier_key

supplier
Dimension table

supplier_key
supplier_type

branch
Dimension table

branch_key
branch_name
branch_type

location
Dimension table

location_key
street
city_key

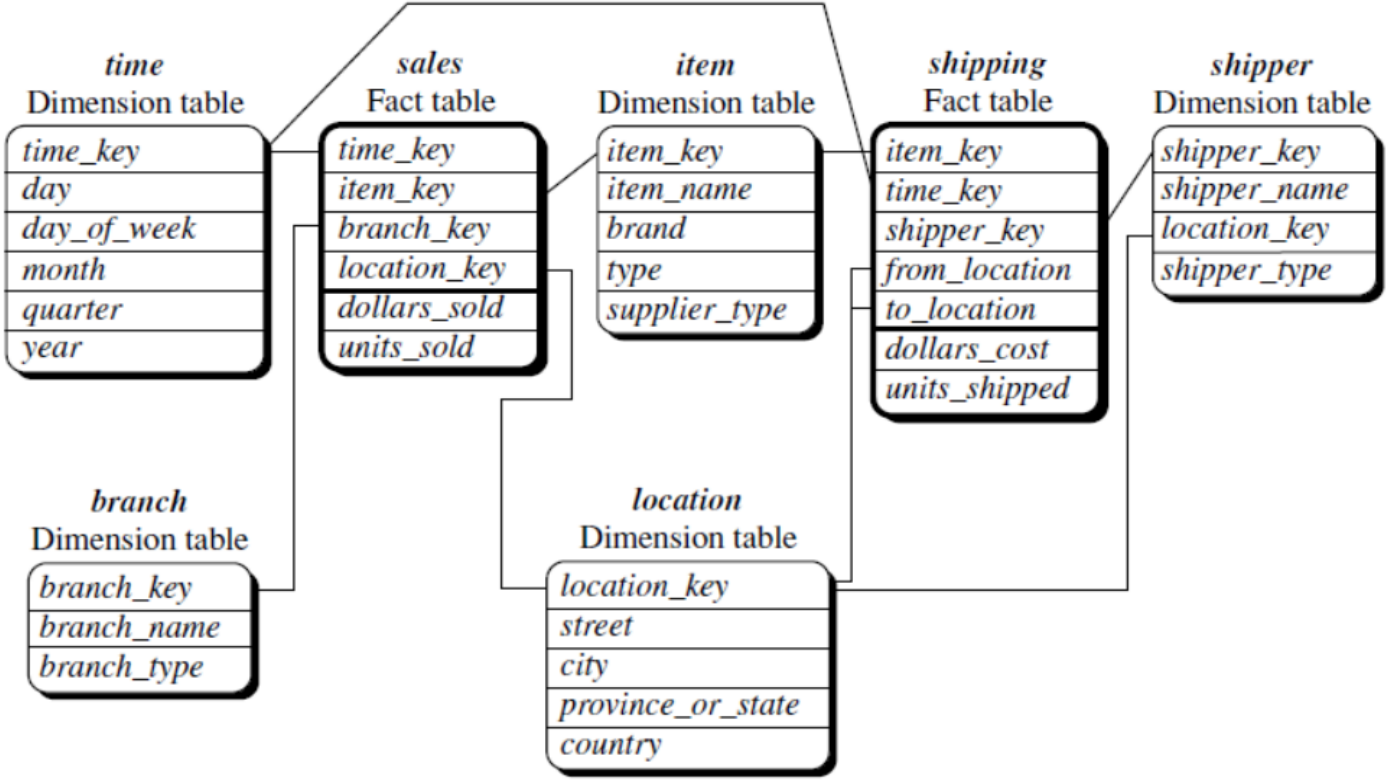
city
Dimension table

city_key
city
province_or_state
country

Fact Constellation or Galaxy Schema

- **Fact constellation:** Sophisticated applications may require multiple fact tables to *share* dimension tables.
- This kind of schema can be viewed as a collection of stars, and hence is called a **galaxy schema** or a **fact constellation**.
- The example shown has 2 fact tables, sales and shipping.
- A fact constellation schema **allows dimension tables to be shared between fact tables**.
- Eg:- the dimensions tables for *time*, *item*, and *location* are shared between the *sales* and *shipping* fact tables.

Fact Constellation or Galaxy Schema



time
Dimension table
- *time_key*
- *day*
- *day_of_week*
- *month*
- *quarter*
- *year*

sales
Fact table
- *time_key*
- *item_key*
- *branch_key*
- *location_key*
- *dollars_sold*
- *units_sold*

item
Dimension table
- *item_key*
- *item_name*
- *brand*
- *type*
- *supplier_type*

shipping
Fact table
- *item_key*
- *time_key*
- *shipper_key*
- *from_location*
- *to_location*
- *dollars_cost*
- *units_shipped*

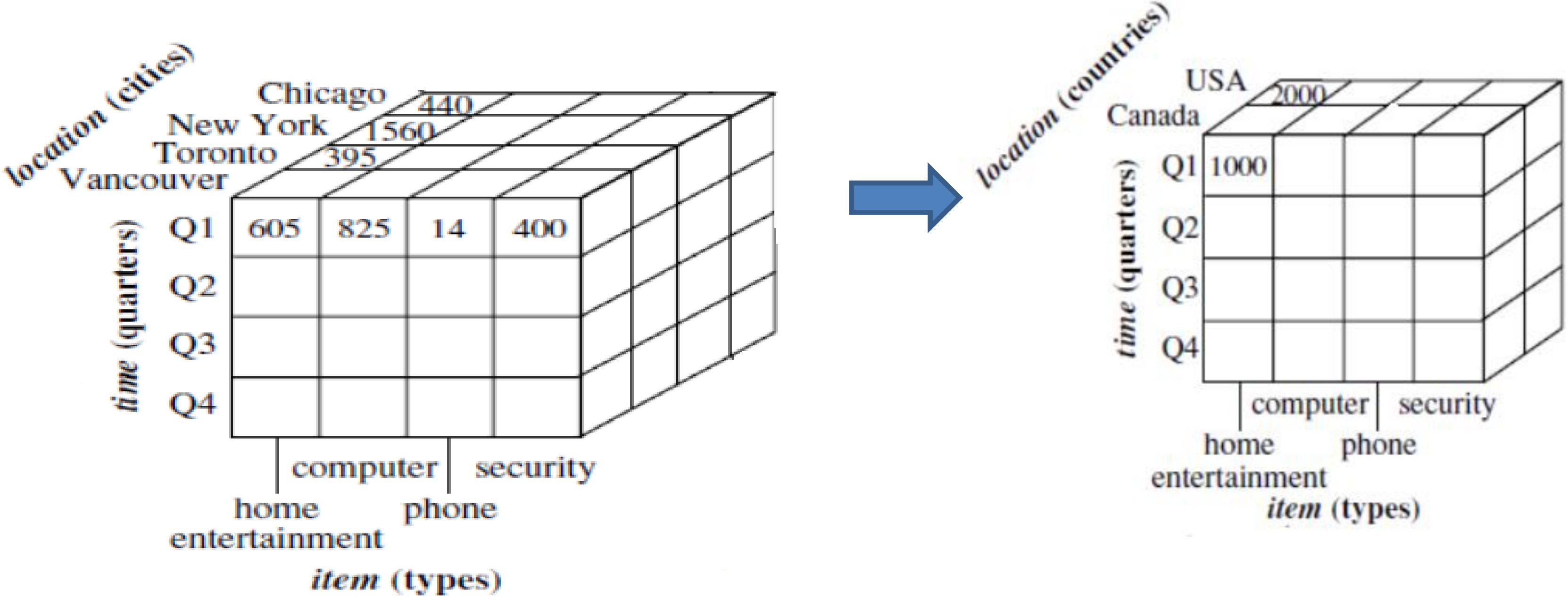
shipper
Dimension table
- *shipper_key*
- *shipper_name*
- *location_key*
- *shipper_type*

branch
Dimension table
- *branch_key*
- *branch_name*
- *branch_type*

location
Dimension table
- *location_key*
- *street*
- *city*
- *province_or_state*
- *country*

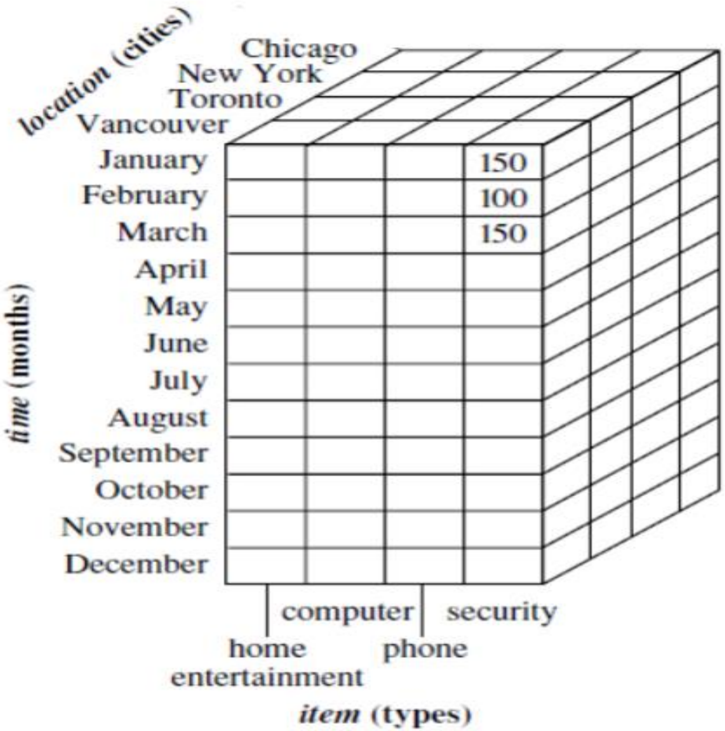
OLAP Operations

Roll up/Drill up :- performs aggregation on a data cube, either by ***climbing up a concept hierarchy* for a dimension** or by ***dimension reduction*** *(eg:- 4D cube to 3D by removing the dimension, supplier).*



OLAP Operations

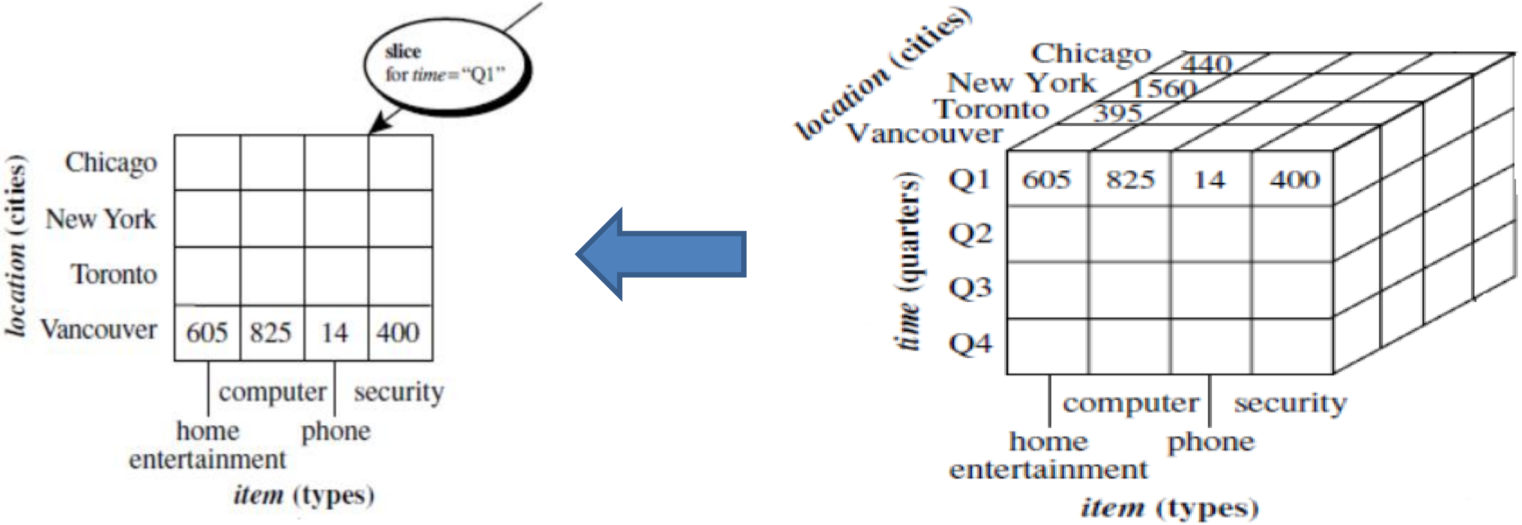
Drill-down:- is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either ***stepping down a concept hierarchy*** **for a dimension** or ***introducing additional dimensions (e.g.:- 3D to 4D cube)***.



OLAP Operations

Slice: The *slice* operation performs a selection on one dimension of the given cube, resulting in a sub-cube.

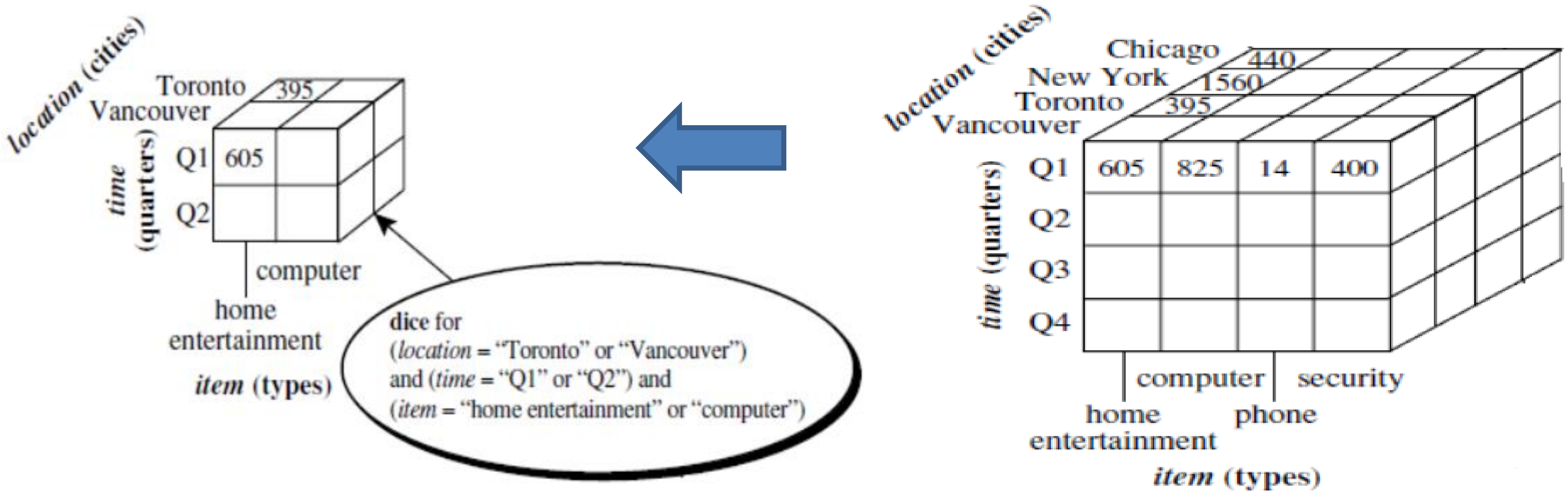
E.g.:- The sales data are selected from the central cube for the dimension *time* using the criterion *time* = "Q1."



OLAP Operations

Dice: The *dice* operation defines a sub-cube by performing a selection on two or more dimensions.

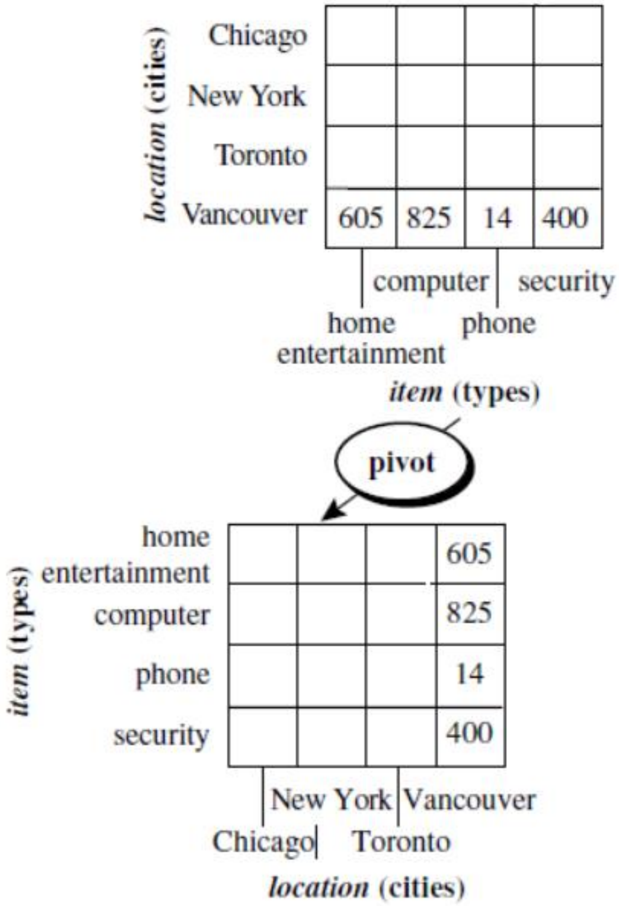
E.g.:- (*location* = “Toronto” or “Vancouver”) and (*time* = “Q1” or “Q2”) and (item = “home entertainment” or “computer”)



dice for (*location* = “Toronto” or “Vancouver”) and (*time* = “Q1” or “Q2”) and (*item* = “home entertainment” or “computer”)

OLAP Operations

Pivot:- *Pivot* (also called ***rotate***) is a visualization operation that rotates the data axes in view to provide an alternative data presentation.



location (cities)	home entertainment	computer	phone	security
Chicago				
New York				
Toronto				
Vancouver	605	825	14	400

item (types)	Chicago	New York	Toronto	Vancouver
home entertainment				605
computer				825
phone				14
security				400

Data Warehousing to Data Mining

- The data mining field has conducted substantial research regarding mining on various data types, including:
  - relational data
  - data from data warehouses
  - transaction data
  - time-series data
  - spatial data
  - text data, and
  - flat files.

- **Multidimensional data mining** (also known as *exploratory multidimensional data mining*, **online analytical mining**, or **OLAM**) integrates OLAP with data mining to uncover knowledge in multidimensional databases.

Data Warehousing to Data Mining

- Multidimensional data mining is particularly important for the following reasons:

- **High quality of data in data warehouses:** Most data mining tools need to work on integrated, consistent, and cleaned data, which requires costly data cleaning, data integration, and data transformation as preprocessing steps. A data warehouse constructed by such preprocessing serves as a valuable source of high-quality data for OLAP as well as for data mining.

- **Available information processing infrastructure surrounding data warehouses:** Comprehensive information processing and data analysis infrastructures have been or will be systematically constructed surrounding data warehouses. It is better to use such available infrastructures is rather than constructing everything from scratch.

Data Warehousing to Data Mining

- **OLAP-based exploration of multidimensional data:** Effective data mining needs exploratory data analysis. A user will often want to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge/ results in different forms. Multidimensional data mining provides facilities for mining on different subsets of data and at varying levels of abstraction—by drilling, pivoting, filtering, dicing, and slicing on a data cube and/or intermediate data mining results. This, together with data/knowledge visualization tools, greatly enhances the power and flexibility of data mining.

- **Online selection of data mining functions:** Users may not always know the specific kinds of knowledge they want to mine. By integrating OLAP with various data mining functions, multidimensional data mining provides users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

Data Warehousing to Data Mining

- **Summary:**
- Data warehousing provides users with large amounts of clean, organized, and summarized data, which greatly facilitates data mining.
- For example, rather than storing the details of each sales transaction, a data warehouse may store a summary of the transactions per item type for each branch or, summarized to a higher level, for each country.
- The capability of OLAP to provide multiple and dynamic views of summarized data in a data warehouse sets a solid foundation for successful data mining.

Data Warehousing to Data Mining

- **Summary:**
- Data mining should be a human-centered process.
- Rather than asking a data mining system to generate patterns and knowledge automatically, a user will often need to interact with the system to perform exploratory data analysis.
- OLAP sets a good example for interactive data analysis and provides the necessary preparations for exploratory data mining.
- Instead of mining associations at a primitive (i.e., low) data level among transactions, users should be allowed to specify roll-up operations along any dimension.

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?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD),
  - knowledge extraction, data/pattern analysis,
  - data archeology,
  - data dredging,
  - information harvesting, business intelligence, etc.

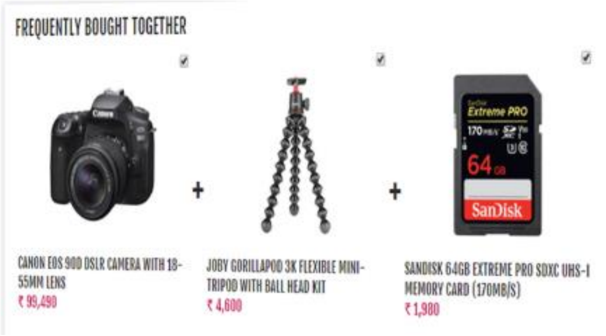
What is Data Mining and what is not?

- **What is not Data Mining?**
  - Looking up phone number in a phone directory
  - Query a web-search engine for information about something

- **What is Data Mining?**
  - Using specific search terms as indicators of some activity, e.g., flu trends. (Those who search for flu symptoms are likely to have flu).
  - Analysing voluminous data for classification, e.g., (loan sanction approval based on the customers' previous payment history.
  - Prediction of the behaviour of a new customer with similar financial and socio-economic background, based on the training set).

Applications of Data Mining: Sales and Marketing

- Enables businesses to understand hidden patterns inside the historical purchasing transaction data.
- Data mining is used for market based analysis:
  - What product combinations were purchased together.
  - Current market trends.
  - Effectiveness of a recent promotional activity.
  - Competitors’ sales strategy analysis


FREQUENTLY BOUGHT TOGETHER
CANON EOS 90D DSLR CAMERA WITH 18-55MM LENS
₹ 99,490
+
JOBY GORILLAPOD 3K FLEXIBLE MINI-TRIPOD WITH BALL HEAD KIT
₹ 4,600
+
SANDISK 64GB EXTREME PRO SDXC UHS-I MEMORY CARD (170MB/S)
₹ 1,980

Applications of Data Mining: Sales and Marketing

- **Academics**
  - Effectively predict the academic performances and placement chances of each student.
  - Predicting students future learning behavior.
  - Learning patterns of students can be identified to develop different teaching methods.

- **Health Sector**
  - Comparing and contrasting symptoms, causes and courses of treatment to find the most effective course of action for a certain illness or condition.
  - Identifying inappropriate referrals or prescriptions, insurance fraud and fraudulent medical claims.

What kind of data can be mined?

- **Database Data:** Consists of a collection of interrelated data. Include DBMS which is a set of software programs to manage and access those data.

- **Data Warehouse Data:** A repository of information collected from multiple sources, stored under a unified scheme and usually reside in a single site.

- **Transactional Data:** Each record in a transactional data captures a transaction.

- **Time related or Sequential Data:** Historical records, stock exchange data, biological sequence data

- **Data Streams:** Video surveillance, sensor data

- **Spatial Data:** Maps

Data Mining vs Knowledge Discovery in Databases

- The terms knowledge discovery in databases (KDD) and data mining are often used interchangeably.

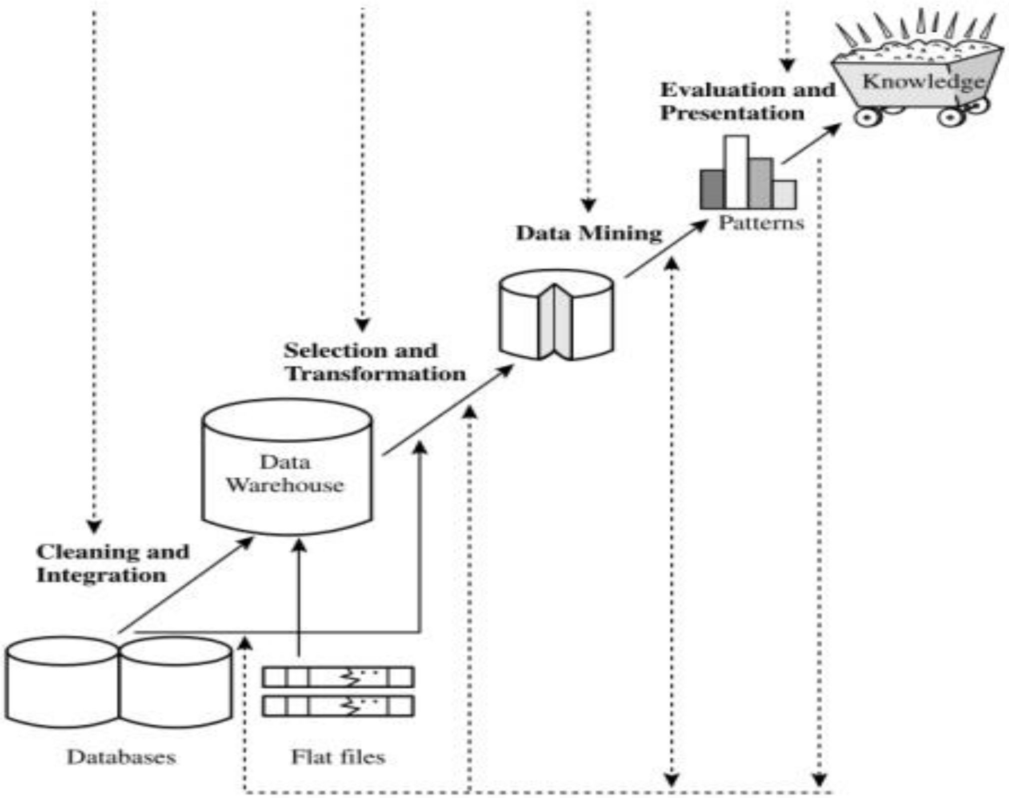
- **Knowledge discovery in databases (KDD)** is the process of finding useful information and patterns in data.

- **Data mining** is the use of algorithms to extract the information and patterns derived by the KDD process.

- KDD is a process that involves many different steps.

- The input to this process is the data, and the output is the useful information desired by the users.

Data Mining as a Step in Knowledge Discovery from Data (KDD)



Data Mining as a Step in Knowledge Discovery from Data (KDD)

- **Data cleaning:**
    - Remove noise and inconsistent data.
    - “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies

- **Data integration:**
    - Multiple data sources may be combined.
    - Involves integrating multiple databases, data cubes, or files

- **Data selection:**
    - Data relevant to the analysis tasks are retrieved from the database

Data Mining as a Step in Knowledge Discovery from Data (KDD)

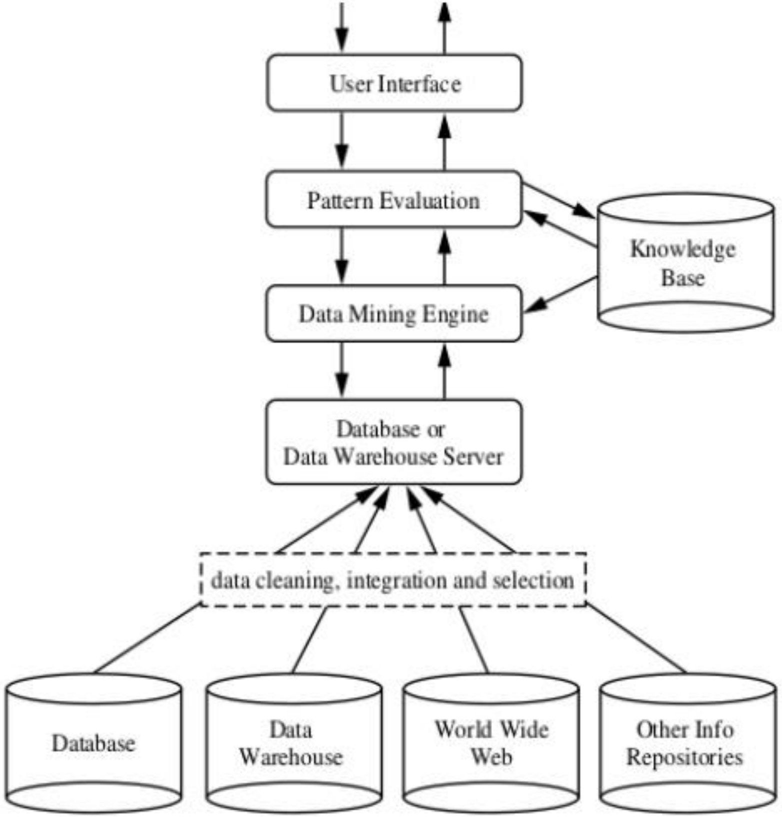
- **Data transformation:**
  - Data are transformed (smoothing, normalization, attribute construction) and consolidated (summarization, aggregation, generalization) into forms appropriate for mining.

- **Data mining:**
  - Essential process where intelligent methods are applied to extract data patterns.

- **Pattern evaluation:**
  - To identify the truly interesting patterns representing knowledge based on interestingness measures.

- **Knowledge presentation:**
  - Visualization and knowledge representation techniques are used to present mined knowledge to the users.

Architecture of a typical data mining system



Architecture of a typical data mining system

- **Database, data warehouse, World Wide Web, or other information repository**
  - Represent different kinds of information repositories where the data resides.
  - Data cleaning and data integration techniques may be performed on the data.

- **Database or data warehouse server**
  - The database or data warehouse server is responsible for fetching the relevant data, based on the user’s data mining request.

- **Knowledge base**
  - This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.
  - Knowledge such as user beliefs, which can be used to assess a pattern’s interestingness based on its unexpectedness, may also be included.

Architecture of a typical data mining system

- **Data mining engine**
  - Consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

- **Pattern evaluation module**
  - This component typically employs interestingness measures
  - Interacts with the data mining modules so as to focus the search toward interesting patterns
  - This module may be integrated with the mining module, depending on the implementation of the data mining method used

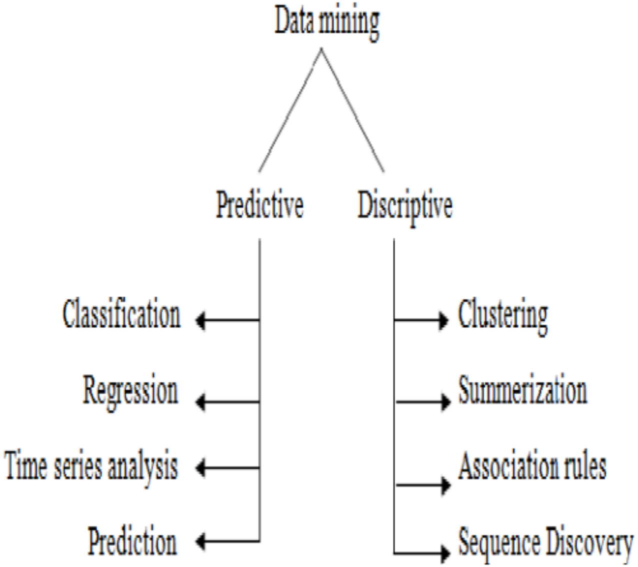
Architecture of a typical data mining system

- **User interface**
  - Module communicates between users and the data mining system
  - Allow the user to interact with the system
  - Users can specify a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.
  - Allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

Data Mining Functionalities

- **Data mining functionalities** are used to specify the kinds of patterns to be found in data mining tasks.

- In general, such tasks can be classified into two categories: **descriptive and predictive.**



Data Mining Functionalities

- **Predictive model:** Makes a prediction about values of data using known results found from different data.

- May be made based on the use of other historical data.

- Classification, Regression, Time Series analysis are predictive in nature

- Eg:- credit card purchase denial due to similarity of the current purchase with earlier purchases subsequently found to be made with stolen cards.

Data Mining Functionalities

- **Descriptive model:** Identifies patterns and relationships in data.
- Serves as a way to explore the properties of the data examined, not to predict new properties.
- Eg:- Clustering, summarization, association rules.

Data Mining Functionalities

- **Data characterization** is a summarization of the general characteristics or features of a target class of data

- Eg: Summarize the characteristics of customers who spend more than $5000 a year at a store.

- 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 Mining Functionalities

- **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.

Data Mining Functionalities

- **Mining Frequent Patterns: Frequent itemset** typically refers to a set of items that often appear together in a transactional data set

- For example: milk and bread, which are frequently bought together in grocery stores by many customers.

- A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card, is a **(frequent) sequential pattern.**

- Mining frequent patterns leads to the discovery of interesting associations and correlations within data **(Association analysis).**

Data Mining Functionalities

- **Classification and Regression for Predictive Analysis:**

- 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).

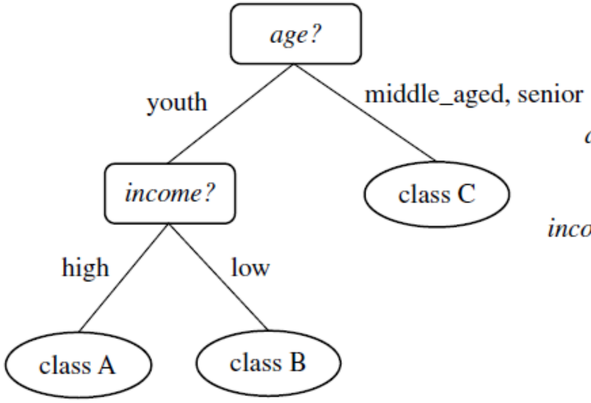
- The model is used to predict the class label of objects for which the class label is unknown.

- The derived model may be represented in various forms, such as classification rules (i.e., IF-THEN rules), decision trees, mathematical formulae, or neural networks.

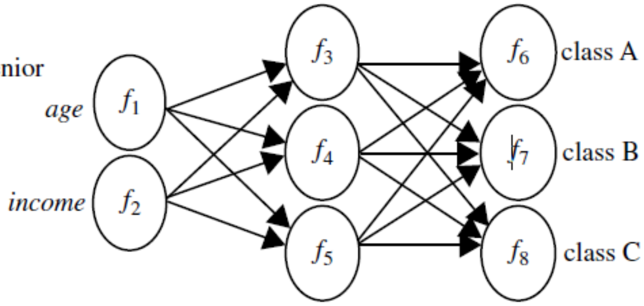
Data Mining Functionalities

age(X, “youth”) AND income(X, “high”) → *class(X, “A”)*
age(X, “youth”) AND income(X, “low”) → *class(X, “B”)*
age(X, “middle_aged”) → *class(X, “C”)*
age(X, “senior”) → *class(X, “C”)*

(a)



(b)



(c)

Data Mining Functionalities

- **Classification and Regression for Predictive Analysis:**
- Classification predicts categorical (discrete, unordered) labels, regression models continuous-valued functions.
- That is, regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels.
- The term prediction refers to both numeric prediction and class label prediction.
- Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well.
- Eg: predict the amount of revenue that each item will generate during an upcoming sale at Company X, based on the previous sales data

Data Mining Functionalities

- **Cluster Analysis:** Unlike classification and regression, which analyze class-labeled (training) data sets, clustering analyzes data objects without consulting class labels.

- In many cases, class labeled data may simply not exist at the beginning.

- 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.

- Eg: Cluster analysis can be performed on Company X customer data to identify homogeneous subpopulations of customers.

Data Mining Functionalities

- **Outlier Analysis:** A data set may contain objects that do not comply with the general behavior or model of the data known as outliers.
- Many data mining methods discard outliers as noise or exceptions.
- However, in some applications (e.g., fraud detection) 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.
- Eg: 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.

Data Mining Issues/Challenges

1. Mining Methodology:

- **Mining various and new kinds of knowledge:** Different data mining tasks to be applied in in diverse applications with ever-growing data.
- **Mining knowledge in multidimensional space:** Exploratory multidimensional data mining.
- **Mining is an inter-disciplinary effort:** To mine data from natural language text, fuse Data Mining + Information retrieval + Natural Language processing.

Data Mining Issues

- **Boosting the power of discovery in a networked environment:** Most data objects reside in a linked environment, such as, inter-linked web documents, interconnected databases through foreign keys, etc. These semantic ties can be used to retrieve information.

- **Handling uncertainty, noise and incompleteness of data:** Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns.

- **Pattern evaluation and pattern or constraint-guided mining:** By using subjective interestingness measures or user-specified constraints to guide the discovery process, we may generate more interesting patterns and reduce the search space.

Data Mining Issues

2. User Interaction:

- **Interactive mining:** Flexible GUIs to facilitate the dynamic exploration of the “cube space” (OLAP operations) by the user, while mining.

- **Incorporation of background knowledge:** Background knowledge related to the mining domain facilitates better querying and information retrieval.

- **Ad-hoc data mining and more sophisticated data mining query languages:** High-level data mining query languages should facilitate specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered patterns.

Data Mining Issues

- **Presentation and visualization of data mining results:** The system must adopt expressive knowledge representations, user-friendly interfaces, and visualization techniques.

3. Efficiency and Scalability:

- **Efficiency and scalability of data mining algorithms:** The running time of a data mining algorithm must be predictable, short, and acceptable by applications.

- **Parallel, distributed and incremental mining algorithms:** Such algorithms first partition the data into “pieces.” Each piece is processed, in parallel, by searching for patterns.

Data Mining Issues

4. Diversity of database types:

- **Handling complex types of data:** Domain or application-dedicated data mining systems are being constructed for in-depth mining of specific kinds of data.

- **Mining dynamic, networked and global data repositories:** Data mining algorithms deal with gigantic, distributed, and heterogeneous global information systems and networks.