**Unit One**

**Objective Of A Data Warehouse**

* Collect Data-Scrub, Integrate & Make It Accessible
* Provide Information – For Our Businesses
* Start Managing Knowledge
* So Our Business Partners Will Gain Wisdom!

**Data Warehouse Definition**

* A Data Warehouse is a Structured Repository of Historic Data.
* It is Developed in an Evolutionary Process by integrating data from Non-integrated legacy systems.
* A Data Warehouse is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
* Data Warehouses are constructed via a process of DATA CLEANING, DATA INTEGRATION, DATA TRANSFORMATION, DATA LOADING, and PERIODIC DATA REFRESHING.

**Definition of Data Mining**

* Data Mining refers to extracting or “mining” Knowledge from large amounts of data.

Scenario:

Remember that the mining of gold from the rocks or sand is referred to as gold mining rather than rock or sand mining.

* The data mining should have been more appropriate named “ knowledge mining from data,” which unfortunately somewhat long, “Knowledge Mining”.
* Data Mining as an analytic process designed to explore data in search for consistent patterns and/or systematic relationship among them.
* The ultimate goal of data mining is prediction.
* The predictive data mining is the most common type of data mining and one that has most direct business application.

**For example:**

For example, a credit card company may want to engage in predictive data mining, to derive a (trained) model or set of models (e.g., neural networks, [meta-learner](http://www.obgyn.cam.ac.uk/cam-only/statsbook/stdatmin.html)) that can quickly identify transactions which have a high probability of being fraudulent.

**Data Mining Tasks**

* In data mining tasks can be classified into two categories:
* Descriptive
* Predictive

Descriptive mining tasks characterize the general properties of the data in the database.

Predictive mining tasks perform inference on the current data in order to make predictions.

**Data Mining Functionalities And Patterns**

The functionalities of data mining , and the kinds of patterns are:

* + Concepts/ Class Description
  + Mining Frequently Patterns, Association, and Correlations
  + Classification and Prediction
  + Clustering Analysis

Evolution Analysis

**Concept/Class Description: Characterization And Discrimination**

Data can be associated with classes or concepts

For Example:

In the All Electronic store, classes of items for sale include computers and printers, and concepts of customers include bigspenders .

Such descriptions of a class or concept are called Class/Concept descriptions.

Data characterization, by summarizing the data of the class under study (often called the **target classes**)

**For Example**:

A data mining system should be able to produce a description summarizing the characteristics of customers who spend more than $1,000 a year at All Electronics. The result could be a profile of the customers, such as they are 40-50 years old, employed, and have excellent credit ratings.

Data discrimination, by comparison of target class with one or a set of comparative classes (often called the **contrasting classes**)

**For Example:**

A data mining system should be able to compare two groups of All Electronics customers, such as those who shops for computer products regularly versus those who rarely shop for such products. The resulting description provides a general comparative profile of the customers, such as 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.

**Mining Frequent Patterns Association And Correlations**

* Frequently patterns, as the name suggests, are patterns that occur frequently in data.
* There are many kinds of frequently patterns, including itemsets, subsequences, and substructures.
* A frequently itemset typically refers to a set of items that frequently appear together in a transactional data sets, such as milk and bread.

Mining frequently patterns leads to the discovery of interesting associations and correlations with data

**Example: Association Analysis**

Suppose, as a marketing managers of ALLElectronics, you would like to determine which items are frequently purchased together within the same transactions.

An example of such a rule, mined from the ALLElectronics transactional database, is

**Buys(X,”computer”)🡺buys(X,”software” [support =1% , confidence=50%]**

**Classification and Prediction**

* Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.
* For the purpose of being able to use the model to predict the class of objects whose class label is unknown.
* The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks.

Example:

Suppose, as sales manager of ALLElectronics, you would like to classify a large set of items in the store, based on three kinds of responses to a sales campaign: good, mild and no response. You would like to drive 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 others, presenting an organized picture of the data set.**

**Cluster Analysis**

* The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.
* The clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another.
* But are very dissimilar to objects in other clusters.

**Example:**

Cluster analysis can be performed on ALLElectronics customer data in order to identify homogenous subpopulations of customers. These clusters may represent individual target groups for marketing.

**Outlier Analysis**

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are called **outliers.**

Most data mining methods discard outliers as noise or exceptions.

In some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones.

**Example:**

Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account.

**Evolution Analysis**

* Data evolution analysis describes and models regularities or trends for objects whose behavior change over time.
* Although this may include characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of time related data.
* Distinct features of such as analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Example:

Suppose that you have the major stock market (time-series) data of the last several years available from the New York Stock Exchange and you would like to invest in shares of high-tech industrial companies.

A data mining study of stock exchange data my identify stock evolution regularities for overall stocks and for the stocks of particular companies. Such regularities may help predict future trends in stock market prices, contributing to your decision making regarding stock investments.

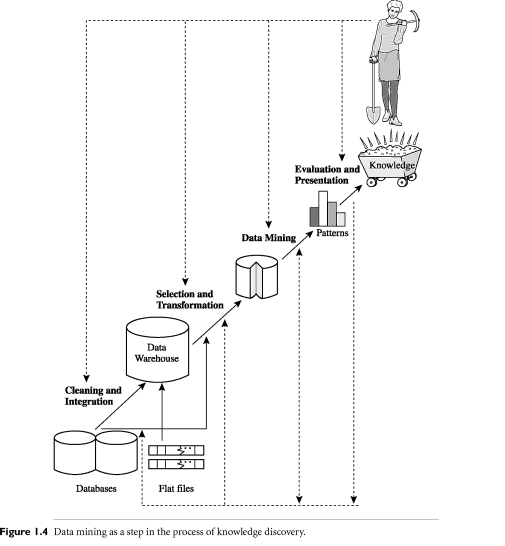
**Data Mining Vs KDD**

* **Knowledge Discovery in Databases (KDD):**

The process of finding useful information and patterns in data.

* **Data Mining:**

Use of algorithms to extract the information and patterns derived by the KDD process.



**The Stages of KDD (Knowledge Discovery in Database)**

The KDD process as follows:

Data Cleaning

Data Integration

Data Selection

Data Transformation

Data Mining

Pattern Evaluation

Knowledge Presentation

**Cleaning:**

To remove noise and inconsistent data

**Data Integration:**

Where multiple data sources may be combined

**Data Selection:**

Where data relevant to the analysis task are retrieved from the database.

**Data Transformation:**

Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operation, for instance.

**Data Mining:**

An essential process where intelligent methods are applied in order to extract data patterns.

(We agree that data mining is a step in the knowledge discovery process)

**Pattern Evaluation:**

To identify the truly interesting patterns representing knowledge based on some interestingness measures

**Knowledge Presentation:**

Where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

**The Major Components of Typical Data Mining**

* Database, Data Warehouse, World Wide Web, or other information repository
* Database or Data Warehouse Server
* Knowledge Based
* Data Mining Engine
* Pattern Evaluation Module

User Interface

**The Issues In Data Mining**

* Security and Social Issues
* User Interface Issues
* Mining Methodology Issues
* Performance Issues

Data Source Issues

**Application of Data Warehouse and Data Mining**

Some of the applications of data warehousing include:

* Agriculture
* Biological data analysis
* Call record analysis
* Churn Prediction for Telecom subscribers, Credit Card users etc.
* Decision support
* Financial forecasting
* Insurance fraud analysis
* Logistics and Inventory management
* Trend analysis

**Review of basic concepts of data warehousing and data mining**

* The Explosive Growth of Data: from terabytes to petabytes
* Data accumulate and double every 9 months
* High-dimensionality of data
* High complexity of data
* New and sophisticated applications
* There is a big gap from stored data to knowledge; and the transition won’t occur automatically.
* Manual data analysis is not new but a bottleneck
* Fast developing Computer Science and Engineering generates new demands

**What is Data Mining?**

Art/Science of extracting non-trivial, implicit, previously unknown, valuable, and potentially Useful information from a large database

**Data mining is**

* A hot buzzword for a class of techniques that find patterns in data
* A user-centric, interactive process which leverages analysis technologies and computing power
* A group of techniques that find relationships that have not previously been discovered
* Not reliant on an existing database
* A relatively easy task that requires knowledge of the business problem/subject matter expertise

**Data mining is not**

1. Brute-force crunching of bulk data
2. “Blind” application of algorithms
3. Going to find relationships where none exist
4. Presenting data in different ways
5. A difficult to understand technology requiring an advanced degree in computer science

**Data mining is not**

* A cybernetic magic that will turn your data into gold. It’s the process and result of knowledge production, knowledge discovery and knowledge management.
* Once the patterns are found Data Mining process is finished.
* Queries to the database are not DM.

**What is Data Warehouse?**

* According to W. H. Inmon, a **data warehouse** is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions.
* “A data warehouse is a copy of transaction data specifically structured for querying and reporting” – Ralph Kimball
* **Data Warehousing** is the process of building a data warehouse for an organization.
* Data Warehousing is a process of transforming data into information and making it available to users in a timely enough manner to make a difference

**Subject Oriented**

* Focus is on Subject Areas rather than Applications
* Organized around major subjects, such as customer, product, sales.
* Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

**Integrated**

* Constructed by integrating multiple, heterogeneous data sources
* Integration tasks handles naming conventions, physical attributes of data
* Must be made consistent.

**Time Variant**

* Only accurate and valid at some point in time or over some time interval.
* The time horizon for the data warehouse is significantly longer than that of operational systems.

Operational database provides current value data.

Data warehouse data provide information from a historical perspective (e.g., past 5-10 years)

**Non Volatile**

* Data Warehouse is relatively Static in nature.
* Not updated in real-time but data in the data warehouse is loaded and refreshed from operational systems, it is not updated by end users.

**Data warehousing helps business managers to :**

* + Extract data from various source systems on different platforms
  + Transform huge data volumes into meaningful information
  + Analyze integrated data across multiple business dimensions
  + Provide access of the analyzed information to the business users anytime anywhere

**OLTP vs. Data Warehouse**

* Online Transaction Processing (OLTP) systems are tuned for known transactions and workloads while workload is not known a priori in a data warehouse
* OLTP applications normally automate clerical data processing tasks of an organization, like data entry and enquiry, transaction handling, etc. (access, read, update)
* Special data organization, access methods and implementation methods are needed to support data warehouse queries (typically multidimensional queries)
  + e.g*., average amount spent on phone calls between 9AM-5PM in Kathmandu during the month of March, 2012*
* **OLTP Data Warehouse**
* Application Oriented Subject Oriented
* Used to run business Used to analyze business
* Detailed data Summarized and refined
* Current up to date Snapshot data
* Isolated Data Integrated Data
* Repetitive access Ad-hoc access
* Clerical User Knowledge User (Manager)
* **OLTP Data Warehouse**
* Performance Sensitive Performance relaxed
* Few Records accessed at a time (tens) Large volumes accessed at a time(millions)
* Read/Update Access Mostly Read (Batch Update)
* No data redundancy Redundancy present
* Database Size 100MB -100 GB Database Size 100 GB - few terabytes
* **OLTP Data Warehouse**
* Transaction throughput is the performance metric Query throughput is the performance metric
* Thousands of users Hundreds of users
* Managed in entirety Managed by subsets

**Why Data Mining?**

Because it can improve customer service, better target marketing campaigns, identify high-risk clients, and improve production processes. In short, because it can help you or your company make or save money.

**Data mining has been used to:**

* Identify unexpected shopping patterns in supermarkets.
* Optimize website profitability by making appropriate offers to each visitor.
* Predict customer response rates in marketing campaigns.
* Defining new customer groups for marketing purposes.
* Predict customer defections: which customers are likely to switch to an alternative supplier in the near future.
* Distinguish between profitable and unprofitable customers.
* Identify suspicious (unusual) behavior, as part of a fraud detection process.
* **Data analysis and decision support**
  + Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation

Risk analysis and management

* + - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  + Fraud detection and detection of unusual patterns (outliers)
* **Other Applications**
  + Text mining (news group, email, documents) and Web mining
  + Stream data mining
  + Bioinformatics and bio-data analysis

**Market Analysis and Management**

* Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
* Target marketing
  + Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
  + Determine customer purchasing patterns over time
* Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
* Customer profiling—What types of customers buy what products (clustering or classification)
* Customer requirement analysis
  + Identify the best products for different groups of customers
  + Predict what factors will attract new customers
* Provision of summary information
  + Multidimensional summary reports
  + Statistical summary information (data central tendency and variation)

**Corporate Analysis & Risk Management**

* Finance planning and asset evaluation
  + cash flow analysis and prediction
  + contingent claim analysis to evaluate assets
  + cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)
* Resource planning
  + summarize and compare the resources and spending
* Competition
  + monitor competitors and market directions
  + group customers into classes and a class-based pricing procedure
  + set pricing strategy in a highly competitive market

**Fraud Detection & Mining Unusual Patterns**

* Approaches: Clustering & model construction for frauds, outlier analysis
* Applications: Health care, retail, credit card service, telecomm.
  + Auto insurance: ring of collisions
  + Money laundering: suspicious monetary transactions
  + Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  + Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  + Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  + Anti-terrorism

**Knowledge Discovery in Databases Process**

* Data selection
* Cleaning
* Enrichment
* Coding
* Data Mining
* Reporting

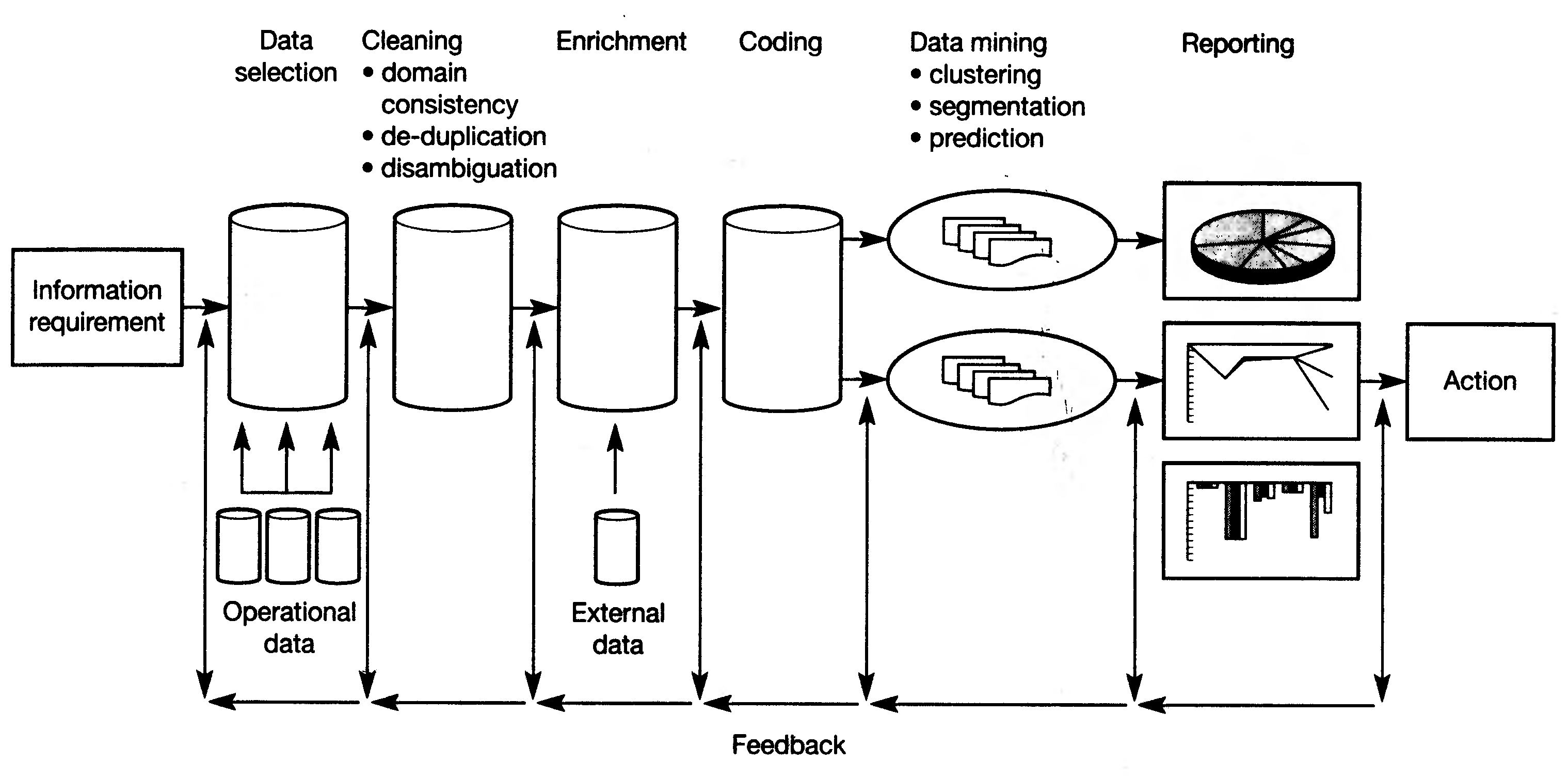


Figure: Knowledge Discovery in Databases (KDD) Process

**Data Selection**

Once you have formulated your informational requirements, the nest logical step is to collect and select the data you need. Setting up a KDD activity is also a long term investment. A data environment will need to download from operational data on a regular basis, therefore investing in a data warehouse is an important aspect of the whole process.

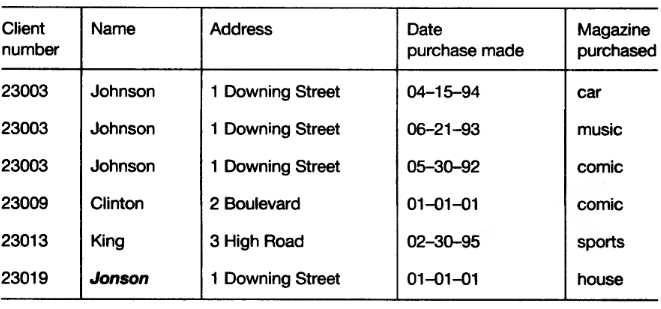


Figure: Original Data

**Cleaning**

Almost all databases in large organizations are polluted and when we start to look at the data from a data mining perspective, ideas concerning consistency of data change. Therefore, before we start the data mining process, we have to clean up the data as much as possible, and this can be done automatically in many cases.

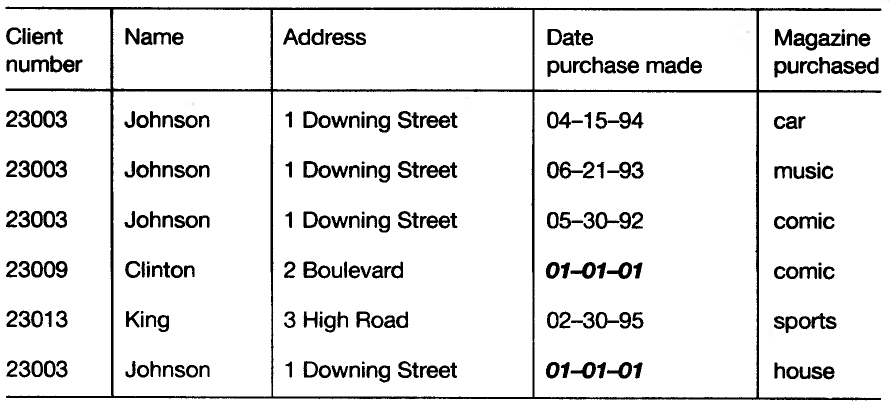


Figure: De-duplication

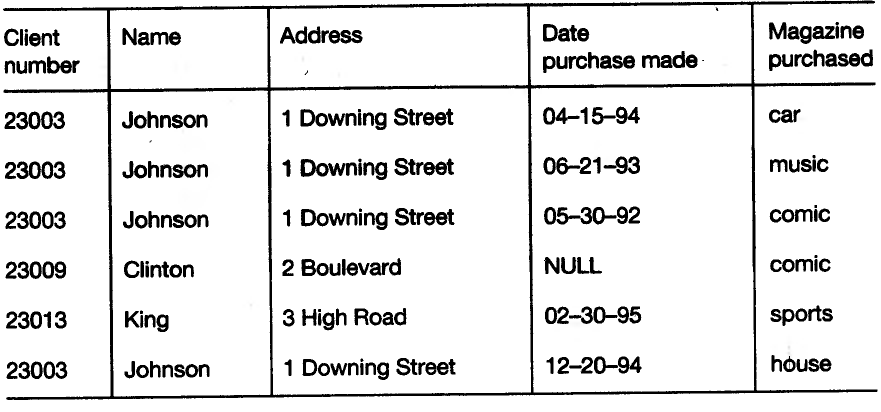


Figure: Domain Consistency

**Enrichment**

Matching the information from bought-in databases with your own databases can be difficult. A well-known problem is the reconstruction of family relationships in databases. In a relational environment, we can simply join this information with our original data.

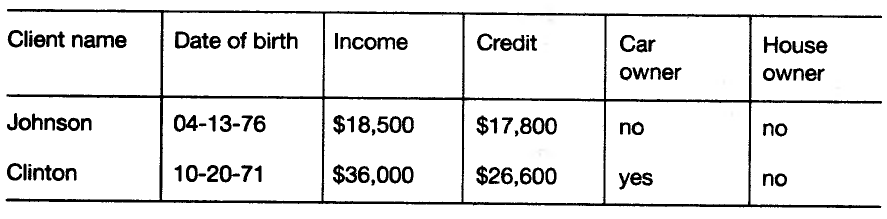


Figure: Enrichment

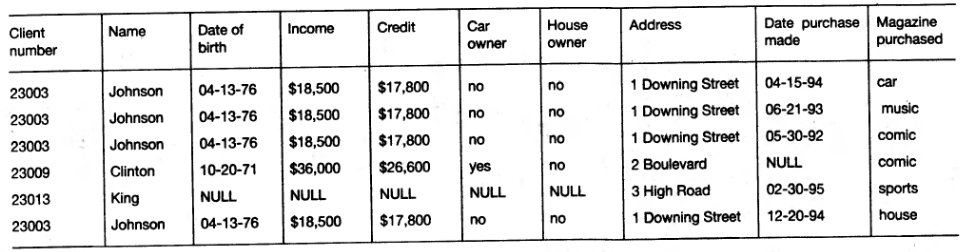


Figure: Enriched Table

**Coding**

We can apply following coding technique:

1. Address to regions
2. Birthdate to age
3. Divide income by 1000
4. Divide credit by 1000
5. Convert cars yes/no to 1/0
6. Convert purchased date to months numbers

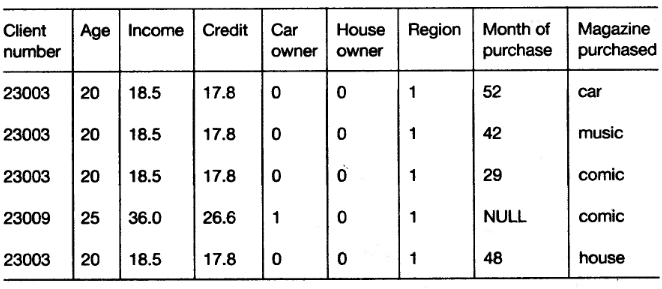


Figure: After Coding Stage

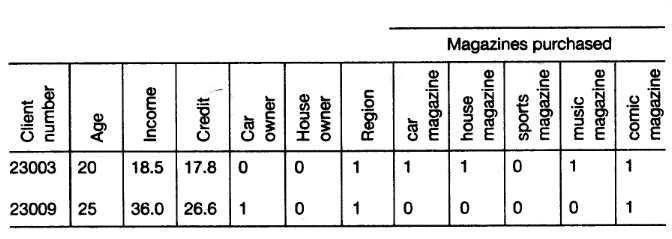


Figure: Final Table

**Data Mining**

It is a discovery stage in KDD process.

* Data mining refers to extracting or “mining” knowledge from large amounts of data.
* Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Database, or KDD.
* Alternatively, others view data mining as simply an essential step in the process of knowledge discovery.

**Some Alternative names to data mining are:**

* + Knowledge discovery (mining) in databases (KDD)
  + Knowledge extraction
  + Data/pattern analysis
  + Data archeology
  + Data Dredging
  + Information Harvesting
  + Business intelligence, etc.

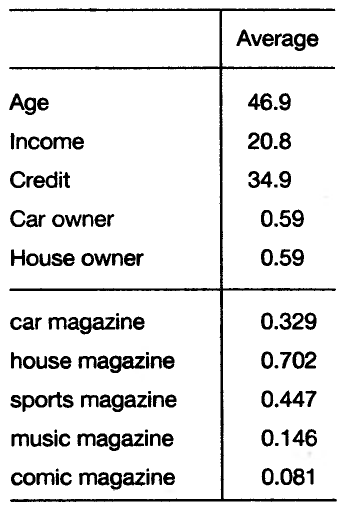


Figure: Averages

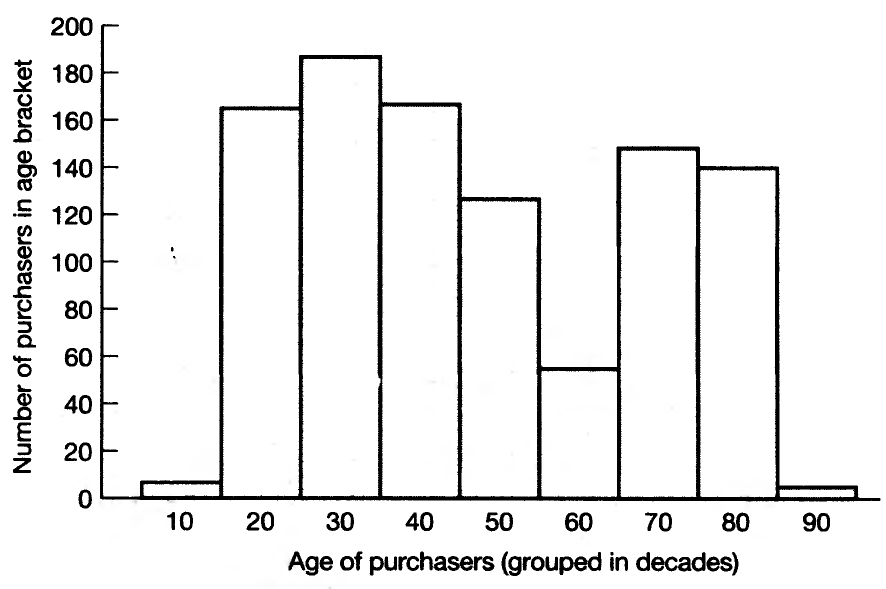


Figure: Age distribution of readers

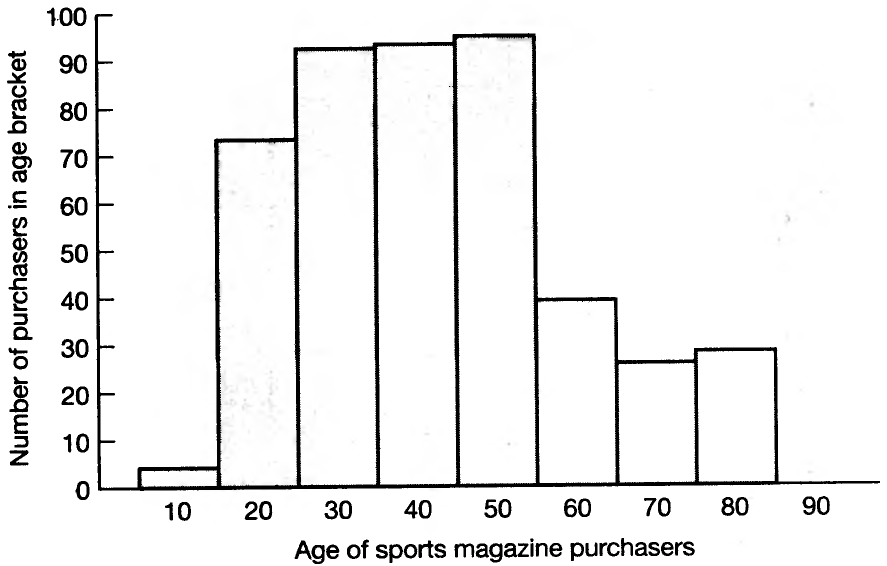


Figure: Age distribution of readers of sports magazines

**Reporting**

* It uses two functions:
  1. Analysis of the results
  2. Application of results
* Visualization and knowledge representation techniques are used to present the mined knowledge to the user.

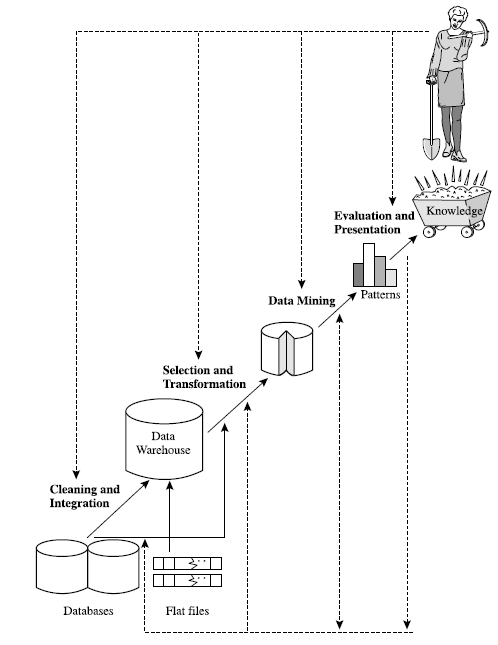
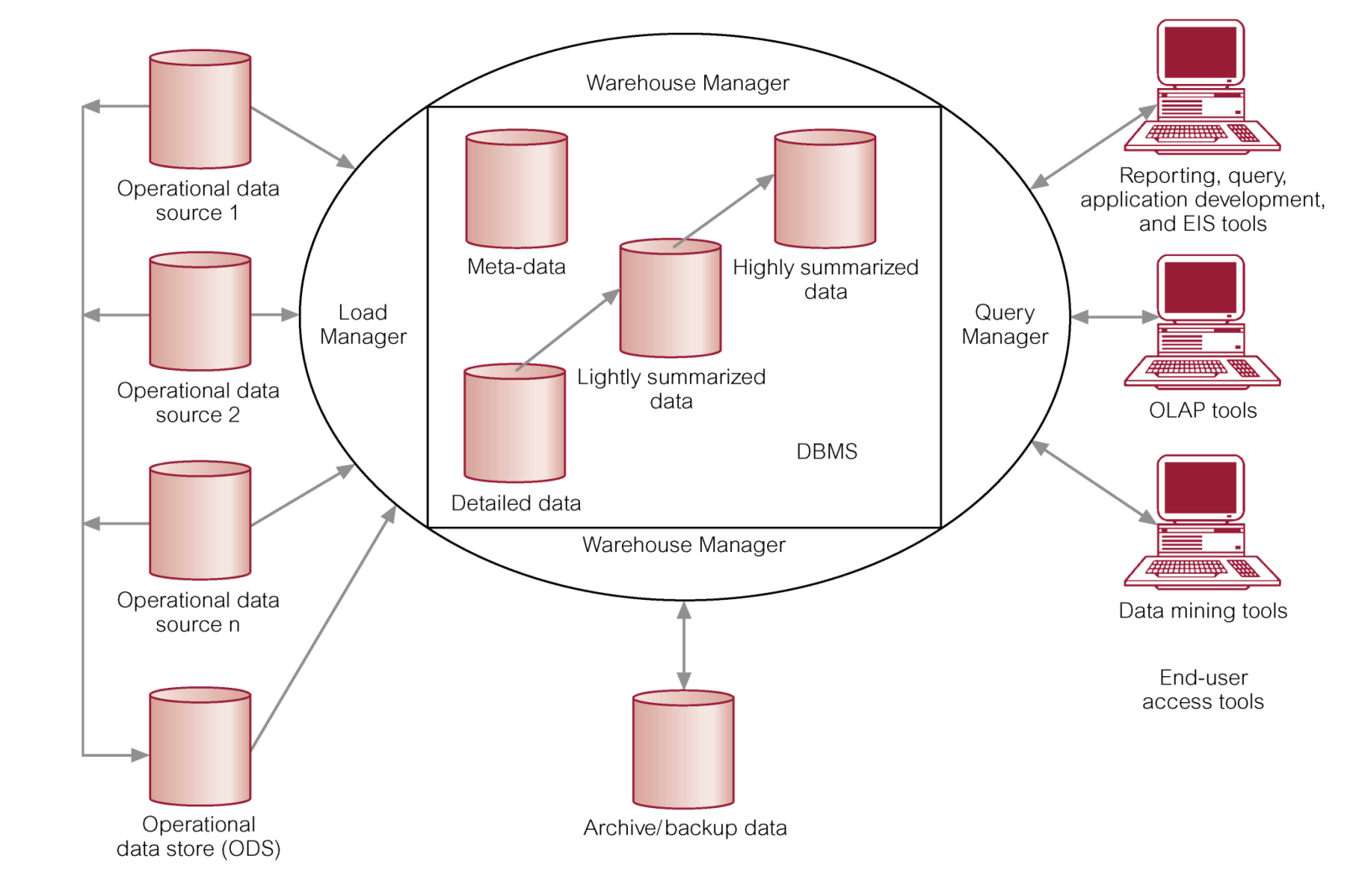


Figure: Data mining as a step in the process of knowledge discovery.

**Data Mining: Confluence of Multiple Discipline**

**Data Warehouse Architecture**



**Operational Data Sources:** Itmay include:

* Network databases.
* Departmental file systems and RDBMSs.
* Private workstations and servers.
* External systems (Internet, commercially available databases).

**Operational Data Store (ODS):** It is a repository of current and integrated operational data used for analysis.

* Often structured and supplied with data in same way as DW.
* May act simply as staging area for data to be moved into the warehouse.

Provides users with the ease of use of a relational database while remaining distant from decision support functions of the DW.

**Warehouse Manager** (Data Manager):

* Operations performed include:
  + Analysis of data to ensure consistency.
  + Transformation/merging of source data from temp storage into DW
  + Creation of indexes.
  + Backing-up and archiving data.

**Query Manager** (Manages User Queries):

* Operations include:
  + directing queries to the appropriate tables and
  + scheduling the execution of queries.
* In some cases, the query manager also generates query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.

**Meta Data:** This area of the DW stores all the meta-data (data about data) definitions used by all the processes in the warehouse.

* Used for a variety of purposes:
  + Extraction and loading processes
  + Warehouse management process
  + Query management process
* End-user access tools use meta-data to understand how to build a query.
* Most vendor tools for copy management and end-user data access use their own versions of meta-data.

**Lightly and Highly Summarized Data:** It stores all the pre-defined lightly and highly aggregated data generated by the warehouse manager.

* The purpose of summary info is to speed up the performance of queries.
* Removes the requirement to continually perform summary operations (such as sort or group by) in answering user queries.

**Archive/Backup Data:** It stores detailed and summarized data for the purposes of archiving and backup.

* May be necessary to backup online summary data if this data is kept beyond the retention period for detailed data.
* The data is transferred to storage archives such as magnetic tape or optical disk.

**End-User Access Tools:**

* The principal purpose of data warehousing is to provide information to business users for strategic decision-making.
* Users interact with the warehouse using end-user access tools.
* There are three main groups of access tools:
  1. Data reporting, query tools
  2. Online analytical processing (OLAP) tools *(Discussed later)*
  3. Data mining tools *(Discussed later)*

**Benefits of Data Warehousing**

* Queries do not impact Operational systems
* Provides quick response to queries for reporting
* Enables Subject Area Orientation
* Integrates data from multiple, diverse sources
* Enables multiple interpretations of same data by different users or groups
* Provides thorough analysis of data over a period of time
* Accuracy of Operational systems can be checked
* Provides analysis capabilities to decision makers
* Increase customer profitability
* Cost effective decision making
* Manage customer and business partner relationships
* Manage risk, assets and liabilities
* Integrate inventory, operations and manufacturing
* Reduction in time to locate, access, and analyze information (Link multiple locations and geographies)
* Identify developing trends and reduce time to market
* Strategic advantage over competitors
* Potential high returns on investment
* Competitive advantage
* Increased productivity of corporate decision-makers
* Provide reliable, High performance access
* Consistent view of Data: Same query, same data. All users should be warned if data load has not come in.
* Quality of data is a driver for business re-engineering.

**Applications of Data Mining**

* Data mining is an interdisciplinary field with wide and diverse applications
  + There exist nontrivial gaps between data mining principles and domain-specific applications
* Some application domains
  + Financial data analysis
  + Retail industry
  + Telecommunication industry

Biological data analysis

**Data Mining for Financial Data Analysis**

* Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality
* Design and construction of data warehouses for multidimensional data analysis and data mining
  + View the debt and revenue changes by month, by region, by sector, and by other factors
  + Access statistical information such as max, min, total, average, trend, etc.
* Loan payment prediction/consumer credit policy analysis
  + feature selection and attribute relevance ranking
  + Loan payment performance
  + Consumer credit rating
* Classification and clustering of customers for targeted marketing
  + multidimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group
* Detection of money laundering and other financial crimes
  + integration of from multiple DBs (e.g., bank transactions, federal/state crime history DBs)
  + Tools: data visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)

**Data Mining for Retail Industry**

* Retail industry: huge amounts of data on sales, customer shopping history, etc.
* Applications of retail data mining
  + Identify customer buying behaviors
  + Discover customer shopping patterns and trends
  + Improve the quality of customer service
  + Achieve better customer retention and satisfaction
  + Enhance goods consumption ratios
  + Design more effective goods transportation and distribution policies
* Example 1. Design and construction of data warehouses based on the benefits of data mining
  + Multidimensional analysis of sales, customers, products, time, and region
* Example 2. Analysis of the effectiveness of sales campaigns
* Example 3. Customer retention: Analysis of customer loyalty
  + Use customer loyalty card information to register sequences of purchases of particular customers
  + Use sequential pattern mining to investigate changes in customer consumption or loyalty
  + Suggest adjustments on the pricing and variety of goods
* Example 4. Purchase recommendation and cross-reference of items

**Data Mining for Telecommunication Industry**

* A rapidly expanding and highly competitive industry and a great demand for data mining
  + Understand the business involved
  + Identify telecommunication patterns
  + Catch fraudulent activities
  + Make better use of resources
  + Improve the quality of service
* Multidimensional analysis of telecommunication data
  + Intrinsically multidimensional: calling-time, duration, location of caller, location of callee, type of call, etc.
* Fraudulent pattern analysis and the identification of unusual patterns
  + Identify potentially fraudulent users and their typical usage patterns
  + Detect attempts to gain fraudulent entry to customer accounts
  + Discover unusual patterns which may need special attention
* Multidimensional association and sequential pattern analysis
  + Find usage patterns for a set of communication services by customer group, by month, etc.
  + Promote the sales of specific services
  + Improve the availability of particular services in a region
* Use of visualization tools in telecommunication data analysis

**Biomedical Data Analysis**

* DNA sequences: 4 basic building blocks (nucleotides): adenine (A), cytosine (C), guanine (G), and thymine (T).
* Gene: a sequence of hundreds of individual nucleotides arranged in a particular order
* Humans have around 30,000 genes
* Tremendous number of ways that the nucleotides can be ordered and sequenced to form distinct genes
* Semantic integration of heterogeneous, distributed genome databases
  + Current: highly distributed, uncontrolled generation and use of a wide variety of DNA data
  + Data cleaning and data integration methods developed in data mining will help
* Similarity search and comparison among DNA sequences
  + Compare the frequently occurring patterns of each class (e.g., diseased and healthy)
  + Identify gene sequence patterns that play roles in various diseases
* Association analysis: identification of co-occurring gene sequences
  + Most diseases are not triggered by a single gene but by a combination of genes acting together
  + Association analysis may help determine the kinds of genes that are likely to co-occur together in target samples
* Path analysis: linking genes to different disease development stages
  + Different genes may become active at different stages of the disease
  + Develop pharmaceutical interventions that target the different stages separately
* Visualization tools and genetic data analysis

**Problems in Data Warehousing**

* Underestimation of resources for data loading
* Hidden problems with source systems
* Required data not captured
* Increased end-user demands
* Data homogenization
* High demand for resources
* Data ownership
* High maintenance
* Long duration projects

Complexity of integration

**Major Challenges in Data Warehousing**

* Data mining requires single, separate, clean, integrated, and self-consistent source of data.
  + A DW is well equipped for providing data for mining.
* Data quality and consistency is essential to ensure the accuracy of the predictive models.
  + DWs are populated with clean, consistent data
* Advantageous to mine data from multiple sources to discover as many interrelationships as possible.
  + DWs contain data from a number of sources.
* Selecting relevant subsets of records and fields for data mining
  + requires query capabilities of the DW.
* Results of a data mining study are useful if can further investigate the uncovered patterns.
  + DWs provide capability to go back to the data source.
* The largest challenge a data miner may face is the sheer volume of data in the data warehouse.
* It is quite important, then, that summary data also be available to get the analysis started.
* A major problem is that this sheer volume may mask the important relationships the data miner is interested in.
* The ability to overcome the volume and be able to interpret the data is quite important.

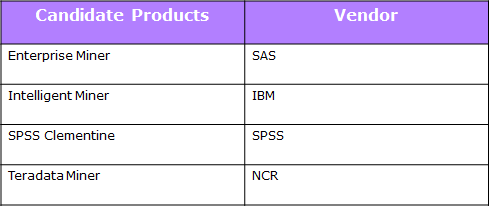
**Major Challenges in Data Mining**

* Efficiency and scalability of data mining algorithms
* Parallel, distributed, stream, and incremental mining methods
* Handling high-dimensionality
* Handling noise, uncertainty, and incompleteness of data
* Incorporation of constraints, expert knowledge, and background knowledge in data mining
* Pattern evaluation and knowledge integration
* Mining diverse and heterogeneous kinds of data: e.g., bioinformatics, Web, software/system engineering, information networks
* Application-oriented and domain-specific data mining
* Invisible data mining (embedded in other functional modules)
* Protection of security, integrity, and privacy in data mining

**Warehouse Products**

* Computer Associates -- CA-Ingres
* Hewlett-Packard -- Allbase/SQL
* Informix -- Informix, Informix XPS
* Microsoft -- SQL Server
* Oracle -- Oracle7, Oracle Parallel Server
* Red Brick -- Red Brick Warehouse
* SAS Institute -- SAS
* Software AG -- ADABAS
* Sybase -- SQL Server, IQ, MPP

**Data Mining Products**



**Unit Two**

**DBMS Vs Data Warehouse**

DBMS is the whole system used for managing digital databases, which allows storage of database content, creation/maintenance of data, search and other functionalities.

Whereas a data warehouse is a place that store data for archival, analysis and security purposes. A data warehouse is made up of a single computer or several computers connected together to form a computer system.

DBMS, sometimes just called a database manager, is a collection of computer program that is dedicated for the management (i.e organization, storage and retrieval) of all databases that are installed in the system (i.e hard drive or network).

Data warehouses play a major role in Decision Support Systems (DSS). DSS is a technique used by organizations to develop and identify facts, trends or relationships that would help them to make better decisions to achieve their organizational goals.

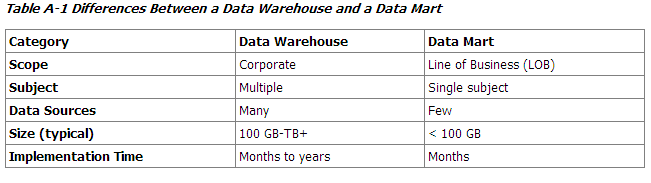
The key difference between DBMS and data warehouse is the fact that a data warehouse can be treated as a type of a database or a special kind of database, which provides special facilities for analysis, and reporting while, DBMS is the overall system which manages a certain database.

Data warehouses mainly store data for the purpose of reporting and analysis that would help an organization in the process making decisions, while a DBMS is a computer application that is used to organize, store and retrieve data. A data warehouse needs to use a DBMS to make data organization and retrieval more efficient.

**Data Mart**

**Definition**

* A data mart is a simple form of data warehouse that is focused on a single subject ( or functional area) such as Sales, Finance, or Marketing.
* Data marts are often built and controlled by a single department within an organization.
* Given their single-subject focus, data marts usually draw data from only a few sources.
* The sources could be internal operational systems, a central data warehouse, or external data.



**Types Of Data Marts**

There are two basic types of data marts:

Dependent

Independent

**Dependent Data Mart**:

The dependent data marts draw data from a central data warehouse that has already been created.

**Independent Data Mart**:

In contrast, are standalone systems built by drawing data directly from operational or external sources of data, or both.

**Metadata**

* Database that describes various aspects of data in the warehouse
* Administrative Metadata: Source database and contents, Transformations required, History of Migrated data
* End User Metadata:
  + - * + Definition of warehouse data
        + Descriptions of it
        + Consolidation Hierarchy

**Uses of Metadata**

* Map source system data to data warehouse tables
* Generate data extract, transform, and load procedures for import jobs
* Help users discover what data are in the data warehouse
* Help users structure queries to access data they need

**A Multidimensional Data Model**

Data warehouse and OLAP tools are based on a multidimensional data model. This model views data in the form of **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.

**Example:**

AllElectronics may create a sales data warehouse in order to keep records of the store’s sales with respect to the dimensions **time, item, branch, and locations.** These dimensions allow the store to keep track of things like monthly sales of items and branches and locations at which the items where sold.

**Different n-Dimension Data Model**

* From Tables and Spreadsheets to Data Cubes
* Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Databases
* Examples for Defining Star, Snowflake, and Fact Constellation Schemas
* OLAP operations in the Multidimensional Data Model

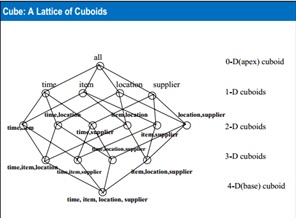
**From Tables and Spreadsheets to Data Cubes**

* A data warehouse is based on a multidimensional data model which views data in the form of a data cube
* A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions

» Dimension tables, such as item (item\_name, brand, type), or time(day, week, month, quarter, year)

» Fact table contains measures (such as dollars\_sold) and keys to each of the related dimension tables

In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.



**Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Databases**

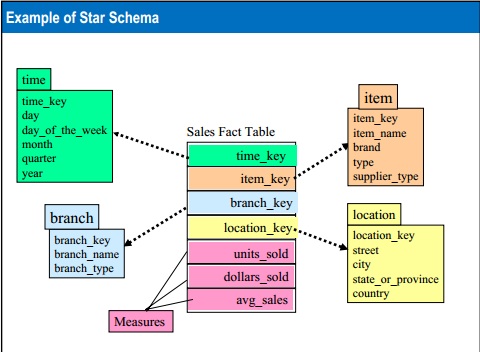
* The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them.
* Such a data model is appropriate for **on-line transaction processing**.
* A **data warehouse**, however, requires a concise, subject-oriented schema that facilitates **on-line data analysis**.
* The most popular data model for a data warehouse is a **multidimensional model**.
* Such a model can exist in the form of a **star schema**, a **snowflake schema**, or a **fact constellation schema**.

**Stars Schema**

The most common modeling paradigm is the star schema, in which the data warehouse contains

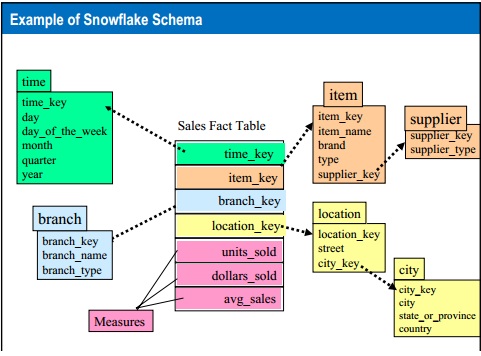
* + A large central table (Fact Table) containing the bulk of the data, with no redundancy
  + A set of smaller attendant tables(Dimension Tables), one for each dimension.

The Schema graph resembles a starburst, with the dimension tables displays displayed in a radial pattern around the central fact table.



**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.
* The resulting schema graph forms a shape similar to a snowflake.
* The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies.
* Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.



**Fact Constellation**

Sophisticated application 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**.

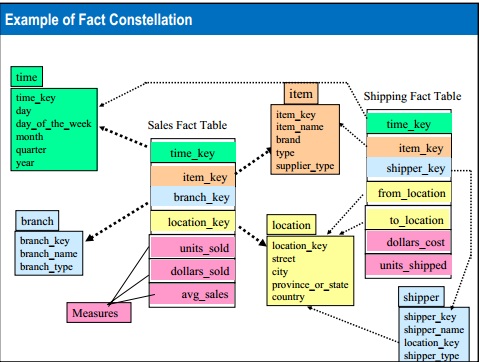


Fig: Fact Constellation Schema of a data warehouse for sales

**Data Warehouse Vs Data Mart**

In data warehousing, there is a distinction between a data warehouse and a data mart.

A data warehouse collects information about subjects that span the entire organizations, such as customers, items, sales, assets, and personnel, and thus its scope is **enterprise-wide**.

For data warehouse, the **fact constellation schema** is commonly used, since it can model multiple, interrelated subjects.

A data mart, on the other hand, is a department subset of the data warehouse that focuses on selected subjects, and thus its scope is **department-wide**. For data mart, the **star or snowflake schema** are commonly used, since both are geared toward modeling single subjects, although the star schema is more popular and efficient.

**Examples for Defining Star, Snowflake, and Fact Constellation Schemas**

* In particular, we examine how to define data warehouses and data marts in our SQL-Based mining query language, **DMQL**.
* Data warehouses and data marts can be defined using two language primitives, one for **cube definition** and one for **dimension definition**.

The cube definition statement has the following syntax:

**define cube** (cube\_name) **[**<dimension\_list>**]:** <measure\_list>

The dimension definition statement has the following syntax:

**define dimension** <dimension\_name> **as** **(**<attribut\_or\_dimension\_list>**)**

**Examples:**

**Star Schema:**

**define cube** sales\_star [time, item, branch, location]: dollar\_sold=sum(sales\_in\_dollars), units\_sold=count(\*)

define dimension time as (time\_key, day, day\_of\_weeks, month, quarter, year )

define dimension item as (item\_key, item\_name, brand, type, supplier\_type)

define dimension branch as (branch\_key, branch\_name, branch\_type)

Define dimension location as (location\_key, street, city, province\_or\_state, country)

**OLAP Operations in the Multidimensional Data Model**

A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand.

Example: (OLAP Operations)

Let’s look at some typical OLAP operations for multidimensional data. Each of the operations described below is illustrated in the fig.

At the center of the fig. is a data cube for ALLElectronics sales. The cube contains the dimensions **location, time, and item.**

*Where location is aggregated with respect to city values, time is aggregated with respect to quarters and item is aggregated with respect to item types.*

The data examined are for the cities Chicago, New York, Toronto and Vancouver.

**OLAP Operations: Rollup, Drill-down, and Slice and Dice**

**Rollup:**

The roll-up operation ( also called the drill-up operation by some vendors) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.

Rather than grouping the data by city, the resulting cube groups the data by country.

When roll-up is performed by dimension reduction, one or more dimensions are removed from the given cube.

**Drill-Down:**

The Drill-down is the reverse of roll-up. It navigates from less details data to more detailed data.

Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.

A drill-down adds more detail to the given data, it can also be performed by adding new dimensions to a cube.

**Slice and Dice:**

The slice operation performs a selection on one dimension of the given cube, resulting in a subcube.

The dice operation defines a subcube by performing a selection on two or more dimensions.

**Pivot (rotate):**

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

**Unit Three**

**Data Warehouse Architecture**

In this section, we discuss issues regarding data warehouse architecture:

* + - To design and construct a data warehouse
    - A three tier data warehouse architecture

**The Design of a Data Warehouse: A Business Analysis Framework**

“What can business analysts gain from having a data warehouse?

First, having a data warehouse may **provide a competitive** advantage by presenting relevant information from which to measure performance and make critical adjustments in order to help win over competitors.

Second, a data warehouse can **enhance business productivity** because it is able to quickly and efficiently gather information that accurately describes the organization.

Third, a data warehouse **facilitates customer relationship management** because it provides a consistent view of customers and item across all lines of business, all departments and all markets.

Finally, a data warehouse may **bring about cost reduction** by tracking trends, patterns, and exceptions over long periods in a consistent and reliable manner.

Four Different views regarding the design of a data warehouse must be considered:

The top-down view

The data source view

The data warehouse view

The business query view

**The top-down**

It view allows the selection of the relevant information necessary for the data warehouse. This information matches the current and future business needs.

**The data source view**

It exposes the information being captured, stored, and managed by operational systems. This information may be documented at various levels of details and accuracy.

From individual data source tables to integrated data source tables.

Data sources are often modeled by traditional data modeling techniques, such as ER model or CASE tools.

**The data warehouse view**

It includes fact tables and dimension tables. It represents the information that is stored inside the data warehouse, including pre-calculated totals and counts, as well as information regarding the source, date, and time of origin, added to provide historical context.

**The business query view**

It is the perspective of data in the warehouse from the viewpoint of the end user.

Building and using a data warehouse is a complex task because it requires business skills, technology skills and program management skills.

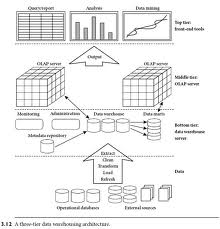
**The Warehouse Design Process**

In general, the warehouse design process consists of the following steps:

* + 1. Choose a business process to model
    2. Choose the grain of the business process
    3. Choose the dimensions that will apply to each fact table record

Choose the measures that will populate each fact table record

**A Three-Tier Data Warehouse Architecture**



1. The **bottom tier** is a warehouse database server that is almost always a relational database system.
2. The **middle tier** is an OLAP server that is typically implemented using either (i) a relational OLAP (ROLAP) model or (ii) a multidimensional OLAP (MOLAP) model.
3. The **top tier** is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools.

**Distributed Data Warehouse**

* (DDW) Data shared across multiple data repositories, for the purpose of OLAP. Each data warehouse may belong to one or many organizations. The sharing implies a common format or definition of data elements (e.g. using XML).
* Distributed data warehousing encompasses a complete enterprise DW but has smaller data stores that are built separately and joined physically over a network, providing users with access to relevant reports without impacting on performance.
* A distributed DW, the nucleus of all enterprise data, sends relevant data to individual data marts from which users can access information for order management, customer billing, sales analysis, and other reporting and analytic functions.

**Data Warehouse Manager**

* Collects data inputs from a variety of sources, including legacy operational systems, third-party data suppliers, and informal sources.
* Assures the quality of these data inputs by correcting spelling, removing mistakes, eliminating null data, and combining multiple sources
* Releases the data from the data staging area to the individual data marts on a regular schedule.
* Measures the costs and benefits.
* Estimates the cost and benefits

**Virtual Warehouse**

* The data warehouse is a great idea, but it is complex to build and requires investment. Why not use a cheap and fast approach by eliminating the transformation steps of repositories for metadata and another database.
* This approach is termed the 'virtual data warehouse'. To accomplish this there is need to define 4 kinds of information:
  + - A data dictionary containing the definitions of the various databases.
    - A description of the relationship among the data elements.
    - The description of the way user will interface with the system.
    - The algorithms and business rules that define what to do and how to do it.

**Disadvantages of VDW**

* + - Since compete with production data transactions, performance can be degraded.
    - There is no metadata, no summary data or no individual DSS (Decision Support System) integration or history. All queries must be repeated, causing additional burden on the system.

There is no refreshing process, causing the queries to be very complex.

**Data Warehouse Back-End Tools and Utilities**

Data warehouse systems use back-end tools and utilities to populate and refresh their data. These tools and utilities include the following functions:

* + - * + Data extraction
        + Data Cleaning
        + Data Transformation
        + Load

Refresh

**Types of OLAP Servers**

OLAP servers present business users with multidimensional data from data warehouses or data marts, without concerns regarding how or where the data are stored.

However, the physical architecture and implementation of OLAP servers must consider data storage issues.

Implementations of a warehouse server for OLAP processing include the following:

* + - * Relational OLAP (ROLAP)
      * Multidimensional OLAP (MOLAP)
      * Hybrid OLAP (HOLAP)

**Relational OLAP (ROLAP) Server**

* + ROLAP servers include optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services.
  + These are the intermediate servers that stand in between a relational back-end server and client front-end tools.
  + They use a relational or extended-relational DBMS to store and manage warehouse, and OLAP middleware to support missing pieces.

ROLAP technology tends to have greater scalability than MOLAP technology.

**Multidimensional OLAP (MOLAP) Server**

* These servers supports multidimensional views of data through array-based multidimensional storage engines
* They map multidimensional views directly to data cube array structures.
* The advantages of using a data cube is that it allows fast indexing to pre-computed summarized data.
* In multidimensional data stores, the storage utilization may be low if the data set is sparse.
* Many MOLAP servers adopt a two-level storage representation to handle dense and sparse data sets.

**Hybrid OLAP (HOLAP) Servers**

* The hybrid OLAP approach combines ROLAP and MOLAP technology.
* Benefiting from the greater scalability of ROLAP and the faster computation of MOLAP.
* The Microsoft SQL Server 2000 supports a hybrid OLAP server.

**Unit Four**

**Computation of Data Cubes**

* Data warehouses contain huge volumes of data.
* OLAP servers demand that decision support queries to answered in the order of seconds.

It is crucial for data warehouse systems to support highly efficient cube computation techniques, access methods and query processing techniques.

**Efficient Computation of Data Cubes**

* At the core of multidimensional data analysis is the efficient computation of aggregations across many sets of dimensions.
* In SQL’s terms, these aggregations are referred to as group-by’s.
* Each group-by can be represented by a cuboid.
* Where the set of group-by’s forms a lattice of cuboids defining a data cube.

**The compute cube Operator and the Curse of Dimensionality**

* One approach to cube computation extends SQL so as to include a compute cube operator.
* The compute cube operator computes aggregates over all subsets of the dimensions specified in the operation.
* This can be require excessive storage space, especially for large numbers of dimensions.
* A cube computation operator was first proposed and studied by Gray et al.

**Example:**

**A data cube** is a lattice of cuboids.

Suppose that you would like to create a data cube for ALLElectronics sales that contains the following: city, item, year, and sales\_in\_dollars.

You would like to be able to analyze the data, with queries such as following:

“Compute the sum of sales, grouping by city and item.”

“Compute the sum of sales, grouping by city.”

“Compute the sum of sales, grouping by item.”

Taking the three attributes, city, item and year.

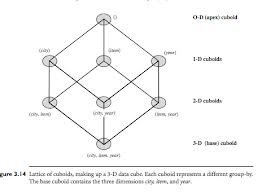
The dimensions for the data cube, and sales\_in\_dollars as the measure, the total number of cuboids, or group-by’s that can be computed for this data cube is 2 power 3 =8.

The possible group-by’s are the following:

{(city, item, year), (city, item), (city, year), (item, year), (city), (item), (year), () },

Where () means that the group-by is empty (i.e. the dimensions are not grouped)

Therefore, for cube with n dimensions, there are a total of 2 power n cuboids, including the base cuboid.



**Fig:** Lattice of cuboids, making up a 3-D data cube. Each cuboid represents a different group-by. The base cuboid contains the three dimensions city, item, and year.

**Curse of Dimensionality**

* OLAP may need to access different cuboids for different queries.
* Therefore, it may seen like a good idea to compute all or at least some of the cuboids in a data cube in advance.
* Pre-computation leads to fast response time and avoids some redundant computation.

A major challenge related to this pre-computation, however, is that the required storage space may explode if all the cuboids in a data cube are pre-computed, especially when the cube has many dimension.

* The storage requirements are even more excessive when many of the dimensions have associated concept hierarchies, each with multiple levels.
* This problem is referred to as the **curse of dimensionality**.
* If there are many cuboids, and these cuboids are large in size, a more reasonable option is **partial materialization**, that is, to materialize only some of the possible cuboids that can be generated.

**Partial Materialization**

There are three choices for data cube materialization given a base cuboid:

**No Materialization**

Do not pre-compute any of the “nonbase” cuboid. This leads to computing expensive multidimensional aggregates on the fly, which can be extremely slow.

**Full Materialization**

Pre-compute all of the cuboids. The resulting lattice of computed cuboids is referred to as the full cube. This choice typically requires huge amounts of memory space in order to store all of the pre-computed cuboids.

**Partial Materialization**

Selectively compute a proper subset of the whole set of possible cuboids. It represents an interesting trade-off between storage space and response time.

**The partial materialization of cuboids or sub-cubes should consider three factors:**

* + - Identify the subset of cuboids or sub-cubes to materialize
    - Exploit the materialized cuboids or sub-cubes during query processing
    - Efficiently update the materialized cuboid or sub-cubes during load and refresh.

**Indexing OLAP DATA**

* To facilitate efficient data accessing, most data warehouse systems support index structures and materialized views (using cuboids).
* The **bitmap indexing** method is popular in OLAP products because it allows quick searching in data cubes.
* The bitmap index is an alternative representation of the record\_ID (RID) list.
* In the bitmap index for a given attribute, there is a distinct bit vector, **Bv**, for each **value v** in the domain of the attribute.
* If the domain of a given attribute consists of **n values**, then **n bits** are needed for each entry in the bitmap index (i.e. there are n bit vectors).
* If the attribute has the **value v** for a given row in the data table, then the bit representing that value is set to **1** in the corresponding row of the bitmap index.
* All other bits for that row are set to 0.
* Bitmap indexing is advantageous compared to hash and tree indices.
* It is especially useful for low-cardinality domains because comparison, join, and aggregation operations are then reduced to bit arithmetic, which substantially reduces the processing time.
* Bitmap indexing leads to significant reductions in space and I/O since a string of characters can be represented by a single bit.
* The **join indexing** method gained popularity from its use in relational database query processing.
* Join indexing registers the joinable rows of two relations from a relational database.
* Join indexing is especially useful for maintaining the relationship between a foreign key and its matching primary keys, from the joinable relation.
* The star schema model of data warehouses makes join indexing attractive for crosstable search, because the linkage between a fact table and its corresponding dimension tables comprises the foreign key of the fact table and the primary key of the dimension table.

**Efficient Processing of OLAP Queries**

The purpose of materializing cuboids and constructing OLAP index structures is to speed up query processing in data cubes. Given materialized views, query processing should proceed as follows:

1. Determine which operations should be performed on the available cuboids.

Determine to which materialized cuboid(s) the relevant operations should be applied.

1. **Determine which operations should be performed on the available cuboids.**
   * This involves transforming any selection, projection, roll-up (group-by), and drill-down operations specified in the query into corresponding SQL and/or OLAP operations.
   * For example, slicing and dicing a data cube may correspond to selection and/or projection operations on a materialized cuboid

**. Determine to which materialized cuboid(s) the relevant operations should be applied.**

* + This involves identifying all of the materialized cuboids that may potentially be used to answer the query, pruning the above set using knowledge of “dominance” relationships among the cuboids.
  + Estimating the costs of using the remaining materialized cuboids, and selecting the cuboid with the least cost.
* The storage model of a MOLAP server is an n-dimensional array, the front-end multidimensional queries are mapped directly to server storage structures, which provides direct addressing capabilities.
* The storage structure used by dense and sparse arrays may differ, making it advantageous to adopt a two-level approach to MOLAP query processing.
* The two-dimensional dense arrays can be indexed by B-trees.

**Tuning and Testing of Data Warehouse**

ETL or Data warehouse **testing** is categorized into four different engagements:

* **New Data Warehouse Testing** – New DW is built and verified from scratch. Data input is taken from customer requirements and different data sources and new data warehouse is build and verified with the help of ETL tools.
* **Migration Testing** – In this type of project customer will have an existing DW and ETL performing the job but they are looking to bag new tool in order to improve efficiency.

**Change Request** – In this type of project new data is added from different sources to an existing DW. Also, there might be a condition where customer needs to change their existing business rule or they might integrate the new rule.

**Report Testing** – Report are the end result of any Data Warehouse and the basic propose for which DW is build. Report must be tested by validating layout, data in the report and calculation.

**Tuning**

* There is little that can be done to **tune** any business rules enforced by constraints. If the rules are enforced by using SQL or by triggers code, that can needs to be tuned to maximal efficiency.
* The load can be also improved by using parallelism.
* The data warehouse will contain two types of query. There will be fixed queries that are clearly defined and well understood, such as regular reports, common aggregations, etc.
* Often the correct tuning choice for such eventualities will be to allow an infrequently used index or aggregation to exist to catch just those sorts of query.
* To create those sorts of indexes or aggregations, you must have an understanding that such queries are likely to be run.
* Before you can tune the data warehouse you must have some objective measures of performance to work with. Measures such as
  + - * + Average query response times
        + Scan rates
        + I/O throughput rates
        + Time used per query
        + Memory usage per process
* These measure should be specified in the service level agreement (SLA).

**Unit Five**

**What is KDD?**

* Computational theories and tools to assist humans in extracting useful information (i.e. knowledge) from digital data
* Development of methods and techniques for making sense of data
* Maps low-level data into other forms that are:
  + - more compact (i.e. short reports)
    - more abstract (i.e. model of process generating the data)
    - more useful (i.e. a predictive model of future cases)
  + Core of KDD process employees “data-mining”

**Why KDD?**

* The size of datasets are growing extremely large – billions of records – hundreds of thousand of fields
* Analysis of data must be automated
* Computers enable us to generate amounts of data for humans to digest, thus we should use computers to discover meaningful patterns and structures from the data

**Current KDD Applications**

A few examples from many fields

* Science – SKYCAT : used to aid astronomers by classifying faint sky objects
* Marketing – AMEX : used customer group identification and forecasting. Claims 10%-15% increase in card usage
* Fraud Detection – HNC Falcon, Nestor Prism: credit card fraud detection : US Treasury money-laundering detection system

Telecommunications – TASA (Telecommunication Alarm-Sequence Analyzer) : locates patterns of frequently occurring alarm episodes and represents the patterns as rules

**KDD vs Data Mining**

**KDD:**

* The nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [Fayyad, et al]
* KDD is the overall process of discovering useful knowledge from data.

**Data mining:**

* An application of specific algorithms for extracting patterns from data.
* Data mining is a step in the KDD process.

**Primary Research and Application Challenges for KDD**

* Larger Databases
* High Dimensionality
* Assessment of Statistical Significance
* Understandability of Patterns
* Integration with Other Systems

**Data Mining Techniques**

* + Association Rules
  + Sequence Mining
  + CBR or Similarity Search

* + Deviation Detection
  + Classification and Regression
  + Clustering

**Data Mining Techniques**

**Association Rules:**

Detects sets of attributes that frequently co-occur, and rules among them, e.g. 90% of the people who buy cookies, also buy milk (60% of all grocery shoppers buy both)

**Sequence Mining (Categorical):**

Discover sequences of events that commonly occur together.

**CBR or Similarity Search:**

Given a database of objects, and a “query” object, find the object(s) that are within a user-defined distance of the queried object, or find all pairs within some distance of each other.

**Deviation Detection:**

Find the record(s) that (are) the most different from the other records, i.e. find all outliers. These may be thrown away as noise or may be the “interesting” ones.

**Classification and regression:**

Assign a new data record to one of several predefined categories or classes. Regression deals with predicting real-valued fields. Also called supervised learning

**Clustering:**

Partition the dataset into subsets or groups such that elements of a group share a common set of properties, with high within group similarity and small inter-group similarity. Also called unsupervised learning

**Types of Data Mining Tasks**

**Descriptive Mining Task**

Characterize the general properties of the data in the database

**Predictive Mining Task**

Perform inference on the current data in order to make predictions

**Data Mining Application Examples**

The areas where data mining has been applied recently include:

* + Science
    - astronomy,
    - bioinformatics,
    - drug discovery, ...
  + Web:
    - search engines, bots, ...
  + Government
    - anti-terrorism efforts (we will discuss controversy over privacy later)
    - law enforcement,
    - profiling tax cheaters
* Business
  + - advertising,
    - Customer modeling and CRM (Customer Relationship management)
    - e-Commerce,
    - fraud detection
    - health care, ...
    - investments,
    - manufacturing,
    - sports/entertainment,
    - telecom (telephone and communications),
    - targeted marketing,

**Data Mining Tools**

Most data mining tools can be classified into one of three categories:

* + - * Traditional Data Mining Tools
      * Dashboards

Text-mining Tools

**Traditional Data Mining Tools**

* Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques.
* Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database.
* The majority are available in both Windows and UNIX versions, although some specialize in one operating system only.
* In addition, while some may concentrate on one database type, most will be able to handle any data using [online analytical processing](http://en.wikipedia.org/wiki/OLAP) or a similar technology.

**Dashboards**

* Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen — often in the form of a chart or table — enabling the user to see how the business is performing.
* Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year).
* This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

**Text-mining Tools**

* The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text — from Microsoft Word and Acrobat PDF documents to simple text files, for example.
* These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications.
* Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database).

**Unit Six**

**Data Mining Query Language**

* DMQL

A Data Mining Query language for Relational Databases

* Create and manipulate data mining models through a SQL-based interface
* Approaches differ on what kinds of models should be created, and what operations we should be able to perform
* Commands specify the following:
  + The set of data relevant to the data mining task (the training set)
  + The kinds of knowledge to be discovered

Generalized relation

Characteristic rules

Discriminant rules

Classification rules

Association rules

* Background knowledge
  + - Concept hierarchies based on attribute relationship, etc.
* Various thresholds
  + - Minimum support, confidence, etc.

**DMQL**

**The primitives specify the following**

* The set of task-relevant data to be mined
* The kind of knowledge to be mined
* The background to be mined
* The background knowledge to be used in the discovery process
* The interestingness measures and thresholds for pattern

evolution

* The expected representation for visualizing the discovered patterns
* Based on these primitives, we design a query language for data mining called DMSQL (data mining query language),
* DMQL allows the ad hoc mining of several kinds of knowledge from relational databases and data warehouses at multiple levels of abstraction.
* The language adopts an SQL-like syntax, so that it can easily be integrated with the relational query language SQL.

**Data Specifications**

* The first step in defining a data-mining task is the specification of the task-relevant data, that is, the data on which mining to be perform.
* This involves specifying the database and tables or data warehouse containing the relevant data, conditions for selecting the relevant data, the relevant attributes or dimensions for exploration, and instructions regarding the order of grouping of the data retired.

**DMQL** provides clauses for the specification of such information, as follows:

**Use database (database\_name) or use the data warehouse (data\_warehouse\_name):**

The use clause directs the mining task to the database or data warehouse specified.

**From (relation (s)/cube (s)) [where (condition)]:**

The from and where clauses respectively specify the database tables or data cubes involved, and the conditions defining the data to be retrieved

**Hierarchy Specification**

* Concept hierarchies allow the mining of knowledge at multiple levels of abstraction.
* In order to accommodate the different viewpoints of users with regard to the data, there may be more than one concept hierarchy per attribute or dimension.
* For instance, some users may prefer to organize branch locations by provinces and states, while others may prefers to organize them according to languages used.
* In such cases, a user can indicate which concept hierarchy is to be used with statement use hierarchy (hierarchy\_name for {attribute\_or\_dimension) Otherwise, a default hierarchy per attribute or dimension is used.

**Interestingness Measure Specification**

* The user can help control the number of uninteresting patterns returned by the data mining system by specify measures of pattern interestingness and their corresponding thresholds
* Interestingness measures and thresholds can be specified be the user with the statement with

{(interest\_measure\_name)} threshold (threshold\_value)

**Pattern Presentation and Visualization Specification**

**How can users specify the forms of presentation and visualization to be used in displaying the discovered patterns?**

Our data mining query language needs syntax that allows user to specify the display of discovered patterns in one or more forms, including rules, table cross tabs, pie or bar charts, decision trees, cubes, etc.

We define the DMQL statement for this purpose

Display a

(Result\_form)

Where the (Result\_form) could be any of the knowledge presenation or visualization forms listed above.

* Interactive mining should allow the discovered patterns to be viewed at different concept level or from different angles.
* This can be accomplished with roll-up and drill-down operations.

**A Data Mining GUI Functional Components**

* Data collection and data mining query composition
* Presentations of discovered patterns
* Hierarchy specification and manipulation
* Manipulation of data mining primitives

**Note:**

* We presented DMQL syntax for specifying data mining queries in terms of the five data mining primitives.
* For a given query, these primitives define the task-relevant data, the kind of knowledge to be mined, the concept hierarchies and interestingness measures to be used, and the presentation forms for pattern visualization.

**The Data Mining Standards**

1. The overall process by which data mining models are produced, used, and deployed
2. A standard representation for data mining and statistical models
3. A standard representation for cleaning, transforming, and aggregating attributes to provide the inputs for data mining models
4. A standard representation for specifying the settings required to build models and to use the outputs of models in other systems.
5. Interfaces and Application Programming Interfaces (APIs) to other languages and systems.
6. Standards for viewing, analyzing, and mining remote and distributed data

**Unit Seven**

**Background**

* **Frequent patterns** are patterns (such as itemsets, subsequences, or substructures) that appear in a data set frequently.
* For Example:
* A set of items, such as milk and bread, that appear frequently together in a transaction data set is a **frequently itemset**.
* **A subsequence**, such as buying first a PC, the a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern.
  + **A substructure** can refer to different structural forms, such as subgraphs, subtree, or sublattices, which may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern.
  + Finding such frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data.
  + Moreover, it helps in data classification, clustering, and other data mining tasks as well.
  + Thus, frequent pattern mining has become an important data mining task and a focused theme in data mining research.

**Why Association Mining is necessary?**

* In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.
* It is intended to identify strong rules discovered in database using different measures of interestingness.

**Data Mining Concepts**

Associations and Item-sets

An association is a rule of the form if X then Y,

It is denoted as X🡪Y

Example:

If Nepal wins in cricket, sales of sweets go up.

For any rule if X🡪Y🡪Y🡪X, then X and Y are called an “interesting item-set”

Example:

People buying school uniforms in June also buy school bags

(People buying school bags in June also buy school uniforms)

**Support and Confidence:**

The support for a rule R is the ratio of the number of occurrences of R, given all occurrences of all rules.

The confidence of a rule X🡪Y, is the ratio of the number of occurrences of Y given X, among all other occurrences given X

**The Apriori Algorithm**

Given minimum required support s as interestingness criterion:

1. Search for all individual elements (1 element item set) that have a minimum support of s
2. Repeat

2.1 From the result of the previous search for i-element item-sets, search for all i+1 element item-sets that have a minimum support of s

2.2 This becomes the sets of all frequent (i+1) element item-sets that are interesting

1. Until item-set size reaches maximum

**Type of Data**

* Tabular (Ex: Transaction Data)

-Relational

- Multi-dimensional

* Spatial (Ex: Remote Sensing Data)
* Temporal (Ex: Log Information)

-Streaming (Ex multimedia, network traffic)

* Tree (Ex: XML Data)
* Graphs (Ex: WWW, BioMolecular Data)
* Sequence (Ex: DNA, activity Logs)
* Text, Multimedia

**Unit eight**

**Contents**

**Classification** maps each data elements to one of a set of pre-determined classes based on the difference among data elements belonging to different classes.

**Clustering** groups data elements into different groups based on the similarity between elements with a single group.

**Definition**

* Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends.
* Such analysis can help provide us with a better understanding of the data at large.
* Whereas classification predicts categorical (discrete, unordered) labels, prediction model continuous valued functions.
* Many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition, and statistics.

**Issue Regarding Classification and Prediction**

* Preparing the Data for Classification and Prediction
* Comparing Classification and Prediction Methods

**Preparing the Data for Classification and Prediction**

The following preprocessing steps may be applied to the data to help improve the accuracy, efficiency, and scalability of the classification or prediction process:

* **Data Cleaning**

This refers to the preprocessing of data in order to remove or reduce noise and the treatment of missing values.

* **Relevance analysis**

Many of the attributes in the data may be redundant. Correlated analysis can be used to identify whether any two given attributes are statistically related.

* **Data Transformation and reduction**

The data may be transformed by normalization, particularly when neural networks or methods involving distance measurements are used in the learning steps.

**Comparing Classification and Prediction Methods**

Classification and prediction methods can be compared and evaluated according to the following criteria:

* **Accuracy**

This refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data (i.e., tuples without class label information)

* **Speed**

This refers to the computational costs involved in generating and using the given classifier or predictor.

* **Robustness**

This refers to ability of the classifier or predictor to make correct predictions given noisy data or data with missing values

* **Scalability**

This refers to the ability to construct the classifier or predictor efficiently given large amounts of data

* **Interpretability**

This refers to the level of understanding and insight that is provided by the classifier or predictor.

**Classification Techniques**

**Hunt’s Method for decision tree identification:**

Given N element types and m decision classes:

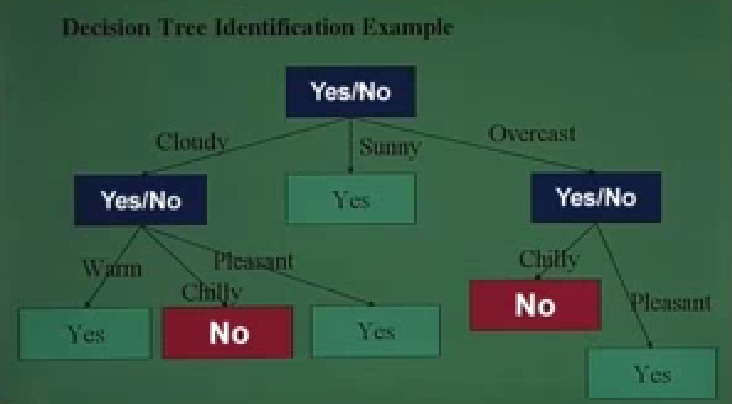
1. For i🡨 1 to N do
2. Add element I to the i-1 element item-sets from the previous iteration

ii. Identify the set of decision classes for each item-set

iii. If an item-set has only one decision class, then that item-set is done, remove that item-set from subsequent iterations.

.2 done

**Classification Techniques**



**Decision Tree Identification Example:**

* Top down technique for decision tree identification
* Decision tree created is sensitive to the order in which items are considered
* If an N-item-set does not result in a clear decision.

Classification classes have to be modeled by rough sets.

**Clustering Techniques**

* Clustering partitions the data set into clusters or equivalence classes.
* Similarity among members of a class more than similarity among members across classes.

Similarity measures: Euclidian distance or other application specific measures.

**Clustering Techniques**

Nearest Neighbour Clustering Algorithm:

Given n elements X1, X2, …. Xn, and threshold t, .

1. j 🡨 1, k 🡨 1, cluster = { }
2. Repeat
   * 1. Find the nearest neighbour of xj
     2. Let the nearest neighbour be in cluster m
     3. If distance to nearest neighbour >t, then create a new cluster and k 🡨k+1; else assign xj to cluster m
     4. j 🡨j+1
3. until j>n

**Iterative partitional Clustering**

Given n element x1, x2, . . . Xn, and k clusters, each with a center.

1. Assign each element to its closest cluster center
2. After all assignments have been made, compute the cluster centroids for each of the cluster
3. Repeat the above two steps with the new centroids until the algorithm converges

**Regression**

Numeric prediction is the task of predicting continuous (or ordered) values for given input

For example:

We may wish to predict the salary of college graduates with 10 years of work experience, or the potential sales of a new product given its price.

The mostly used approach for numeric prediction is regression

A statistical methodology that was developed by Sir Frances Galton (1822-1911), a mathematician who was also a cousin of Charles Darwin

In many texts use the terms “regression” and “numeric prediction” synonymously

Regression analysis can be used to model the relationship between one or more independent or predictor variables and a dependent or response variable (which is continuous value)

In the context of data mining, the predictor variables are the attributes of interest describing the tuple

The response variable is what we want to predict

**Types of Regression**

The types of Regression are as:

* + - Linear Regression
    - NonLinear Regression

**Linear Regression**

Straight-line regression analysis involves a response variable, y, and a single predictor variable, x.

It is the simplest form of regression, and models y as a linear function of x.

That is,

y=b+wx

Where the variance of y is assumed to be constant, and b and w are regression coefficients specifying the Y-intercept and slope of the line, respectively.

The regression coefficient, w and b, can also be thought of as weight, so that we can equivalent write,

y=w0+w1x.

The regression coefficient can be estimated using this method with the following equations:

[Refer to write board:]

Example Too:

**Multiple Linear Regression**

The multiple linear regression is an extension of straight-line regression so as to involve more than one predictor variable.

An example of a multiple linear regression model based on two predictor attributes or variables, A1 and A2, is

y=w0+w1x1+w2x2,

Where x1 and x2 are the values of attributes A1 and A2, respectively, in X.

Multiple regression problems are instead commonly solved with the use of statistical software packages, such as SPSS(**Statistical Package for the Social Sciences)**, etc..

**NonLinear Regression**

The straight-line linear regression case where dependent response variable, y, is modeled as a linear function of a single independent predictor variable, x.

If we can get more accurate model using a nonlinear model, such as a parabola or some other higher-order polynomial?

Polynomial regression is often of interest when there is just one predictor variable.

Consider a cubic polynomial relationship given by

y=w0+w1x+w2xsq2+w3xcu3

**NonLinear Regression**

In statistics, **nonlinear regression** is a form of [regression analysis](http://en.wikipedia.org/wiki/Regression_analysis) in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations.

Contents

Clustering

**Definition**

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression.

 First the set is partitioned into groups based on data similarity (e.g., using clustering), and then labels are assigned to the relatively small number of groups.

It is also called **unsupervised learning**. Unlike classification, clustering and unsupervised learning do not rely on predefined classes and class-labeled training examples. For this reason, clustering is a form of learning by observation, rather than learning by examples.

**Definition**

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.

**Advantages**

**Advantages** of such a clustering-based process:

* + - * Adaptable to changes
      * Helps single out useful features that distinguish different groups.

**Applications** of Clustering

* + - Market research
    - Pattern recognition
    - Data analysis
    - Image processing
    - Biology
    - Geography
    - Automobile insurance
    - Outlier detection

**K-Mean Algorithm**

**Input:**

k: the number of clusters,

D: a data set containing n objects.

**Output:** A set of k clusters.

**Method:**

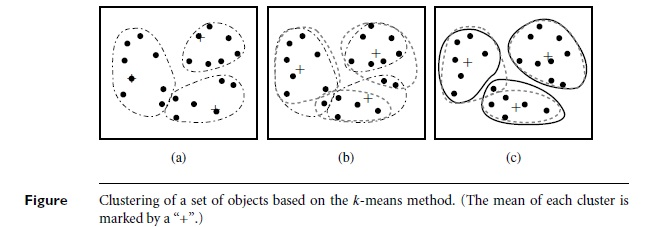
1. arbitrarily choose k objects from D as the initial cluster centers;

(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, i.e., calculate the mean value of the objects for each cluster;

(5) until no change;



**K-Medoids Algorithm**

**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 in *D*as the initial representative objects or seeds;

(2) repeat

(3) assign each remaining object to the cluster with the nearest representative object;

(4) randomly select a nonrepresentative object, ***o***random;

(5) compute the total cost, *S*, of swapping representative object, ***oj***, with ***o***random;

(6) if *S*< 0 then swap ***oj***with ***o***random to form the new set of *k*representative objects;

(7) until no change;

**Bayesian Classification**

**Bayesian Classification** is based on Baye’s theorem.

Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers.

Bayesian classifiers have also exhibited high accuracy and speed when applied to large database.

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence.

Bayesian belief networks are graphical models, which unlike naïve Bayesian classifiers, allow the representation of dependencies among subsets of attributes.

**Bayes’ Theorem**

Bayes’

**Unit Nine**

**Contents**

Spatial Data Mining

Multimedia Data Mining

Text Data Mining

Web Data Mining

**Spatial Data Mining**

**A spatial database** stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, and VLSI(very large-scale integration) chip layout data.

They carry topological and/or distance information, usually organized by sophisticated, multidimensional spatial indexing structures that are accessed by spatial data, geometric computation, etc.

**Definition**

**Spatial data mining** refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases.

It is expected to have wide applications in geographic information systems, geo-marketing, remote sensing, image database exploration, medical imaging, navigation, traffic control, environmental studies, and many other area where spatial data are used.

Statistical spatial data analysis has been a popular approach to analyzing spatial data and exploring geographic information.

**Spatial Data warehouse**

There are several challenging issues regarding the construction and utilization of spatial data warehouses.

The first challenge is the integration of spatial data from heterogeneous sources and systems. Spatial data are usually stored in different industry firms and government agencies using various data formats.

The second challenge is the realization of fast and flexible on-line analytical processing in spatial data warehouses

**The types of dimensions in a spatial data cube:**

**A nonspatial dimesion** contains only nonspatial data. Nonspatial dimensions temperature and precipitation as “hot” for temperature and “wet” for precipitation.

**A spatial-to-nonspatial** dimension is a dimension whose primitive-level data are spatial but who generalization, starting at a certain high level, becomes nonspatial.

For example:

The spatial dimension city relays geographic data for the U.S. map. Suppose that the dimension’s spatial representation of, say, Seattle is generalized to the string “pacific\_northwest”. Although “pacific\_northwest” is a spatial, its representation is not spatial

**A spatial-to-spatial dimension** is a dimension whose primitive level and all of its high level generalization data are spatial.

For example:

The dimension equi\_temperature\_region contains spatial data, as do all of its generalizations, such as with regions covering 0-5\_degrees(Celsius), 5-10\_degrees, and so on.

**Mining Raster Databases**

**Spatial database systems** usually handle vector data that consist of points, lines, polygons(regions), and their compositions, such as networks or partitions.

Typical examples of such data include maps, design graphs, and 3-D representations of the arrangement of the chains of protein molecules.

However, a huge amount of space-related data are in digital raster (image) forms, such as satellite images, remote sensing data, and computer tomography.

It is important to explore data mining in raster or image databases. Methods for mining raster and image data are examined in the following section regarding the mining of multimedia data.

**Multimedia Data Mining**

A **multimedia database system** stores and manages a large collection of multimedia data, such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text markups, and linkages.

**Multimedia database systems** are increasingly common owing to the popular use of audio-video equipment, digital cameras, CD-ROMs, and the internet.

Typical multimedia database systems include NASA’s (National Aeronautics and Space Administration) EOS (Earth Observation System), various kinds of image and audio-video databases, and internet databases.

**Similarity Search in Multimedia Data**

“When searching for similarities in multimedia data, we can search on either the data description or the data content?”

For similarity searching in multimedia data, we consider two main families of multimedia indexing and retrieval systems:

1. **description-based retrieval systems**, which build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creations.

**content-based retrieval systems**, which support retrieval based on the image content, such as color histogram, texture, pattern, image topology, and the shape of the objects and their layouts and locations within the image.

**Description-based retrieval** is labor-intensive if performed manually. If automated, the results are typically of poor quality.

*Recently development of Web-based image clustering and classification methods has improved the quality of description-based Web image retrieval, because image surrounded text information as well as Web linkage information can be used to extract proper description and group images describing a similar theme together.*

**Content-based retrieval** uses visual features to index images and promotes object retrieval based on feature similarity, which is highly desirable in many applications

In a **content-based image retrieval system**, there are often two kinds of queries:

i. image-sample-based queries

ii. Image feature specification queries

**Content-based retrieval** has wide applications, including medical diagnosis, weather prediction, TV production, Web Search engines for images, and e-commerce.

**Image-sample-based queries** find all of the images that are similar to the given image sample.

This search compares the feature vector( or signature) extracted from the sample with the feature vectors of images that have already been extracted and indexed in the image database.

Based on this comparison, images that are closed to the sample image are returned.

**Image feature specification queries** specify or sketch image features like color, texture, or shape, which are translated into a feature vector to be matched with the feature vectors of the images in the database.

*Some systems, such as QBIC (Query By Image Content), support both sample-based and image feature specification queries.*

*There are also systems that support both content-based and description-based retrieval.*

Several approaches have been proposed and studied for **similarity-based retrieval in image databases**, based on image signature:

* + 1. Color histogram-based signature
    2. Multifeature composed signature
    3. Wavelet-based signature
    4. Wavelet-based signature with region-based granularity

**Audio and Video Data Mining**

Besides still images, an incommensurable amount of **audiovisual information** is *becoming* available in digital form, in digital archives, on the World Wide Web, in broadcast data streams, and in personal and professional database.

There are great demands for effective **content-based retrieval** and data mining methods for audio and video data.

*Typical examples include searching for and multimedia editing of particular video clips in a TV studio, detecting suspicious persons or scenes in surveillance videos, finding a particular melody or tune in your MP3 audio album.*

To facilitate the recording, search, and analysis of audio and video information from multimedia data, industry and standardization committees have made great strides towards **developing a set of standards for multimedia information description and compression**.

For example,

MPEG-k (developed by MPEG: Moving Picture Experts Group) and JPEG are typical video compression schemes.

The most recently released MPEG-7, formally named “Multimedia Content Description Interface” is a standard for describing the multimedia content data.

**Text Mining**

Most previous studies of data mining have focused on structured data, such as relational, transactional, and data warehouse data.

However, in reality, a substantial portion of the available information is stored in **text database** ( or document databases) as news articles, research papers, books, digital libraries, e-mail message, and Web pages.

**Text databases** are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mail, and WWW.

Data stored in the most **text databases are semi-structured data** in that they are neither completely unstructured nor completely structured.

**Basic Measures For Text Retrieval**

“Suppose that text retrieval system has just retrieved a number of documents for me based on my input in the form of a query. How can we assess how accurate or correct the system was?”

Let the set of documents relevant to a query be denoted as **{Relevant}**, and the set of documents retrieved be denoted as **{ Retrieved}**. The set of documents that are both relevant and retrieved is denoted as { Relevant} Intersection {Retrieved}.

There are two basic measures for assessing the quality of text retrieval:

**Precision**: This is the percentage of retrieved documents that are in fact relevant to the query (i.e. “correct” responses ). It is formally defined as (see the notation in book page nos. 616)

**Recall**: The is the percentage of documents that are relevant to the query and were, in fact, retrieved. It is formally defined as (see the notation in book page nos. 616)

**Text Mining**

An information retrieval system often needs to trade off recall for precision or vice versa. One commonly used trade-off is the **F-score**, which is defined as harmonic mean of recall and precision. It is formally defined as (see the notation in book page nos. 616)

**Precision, recall, and F-score are the basic measures of a retrieved set of documents**.

**Text Retrieval Methods**

“What methods are there for information retrieval?”

Broadly speaking, retrieval methods fall into two categories: They generally either view the retrieval problem as a **document selection problem** or as **a document ranking problem**.

In **document selection method**, the query is regarded as specifying constraints for selecting relevant documents.

The retrieval system would take such a **Boolean query** and return documents that satisfy the Boolean expression.

*Because of the difficulty in prescribing a user’s information need exactly with a Boolean query, the Boolean retrieval method generally only works well when the user knows a lot about the document collection and can formulate a good query in this way.*

**Document Ranking methods** use the query to rank all documents in the order of relevance.

For ordinary users and exploratory queries, these methods are more appropriate than document selection methods.

Most modern information retrieval systems present a ranked list of documents in response to user’s keyword query.

There are many different ranking methods based on a large spectrum of mathematical foundation, including algebra, logic, probability, and statistics.

**Text Indexing Techniques**

There are several popular text retrieval indexing techniques, including **inverted indices** and **signature files**.

An **inverted index** is an index structure that maintains two hash indexed or B+ -tree indexed tables: **document\_table** and **term\_table**, where

**Document\_table** consists of a set of document records, each containing two fields: doc\_id and posting\_list, where posting\_list is a lists of terms ( or pointers to terms) that occur in the document, sorted according to some relevance measure.

**Term\_table** consists of a set of term records, each containing two fields: term\_id and posting\_list, where posting\_list specifies a list of document identifiers in which the term appears.

A **signature file** is a file that stores a signature record for each document in the database.

Each signature has a fixed size of b bits representing terms.

A simple encoding scheme goes as follows, each bit of a document signature is initialized to 0.

A bit is set to 1 if the term it represents appears in the document.

**Mining the World Wide Web**

The **World Wide Web** serves as huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other services.

It also contains a **rich and dynamic collection of hyperlink information**, and access and usage information, providing rich sources for data mining.

Web Mining includes mining Web linkage structures, Web contents, and Web access patterns.

This involves **mining the Web page layout structure**, mining the Web’s link structures to identify authoritative Web pages, mining multimedia data on the Web, automatic classification of Web documents, and Web usage mining.

Based on the following observation, the Web also poses great challenges for effective resource and knowledge discovery:

* + The Web seems to be too huge for effective data warehousing and data mining.
  + The complexity of Web pages is far greater than that of any traditional text document collection
  + The Web is a highly dynamic information source
  + The Web serves a broad diversity of user communities

Only a small portion of the information on the Web is truly relevant or useful

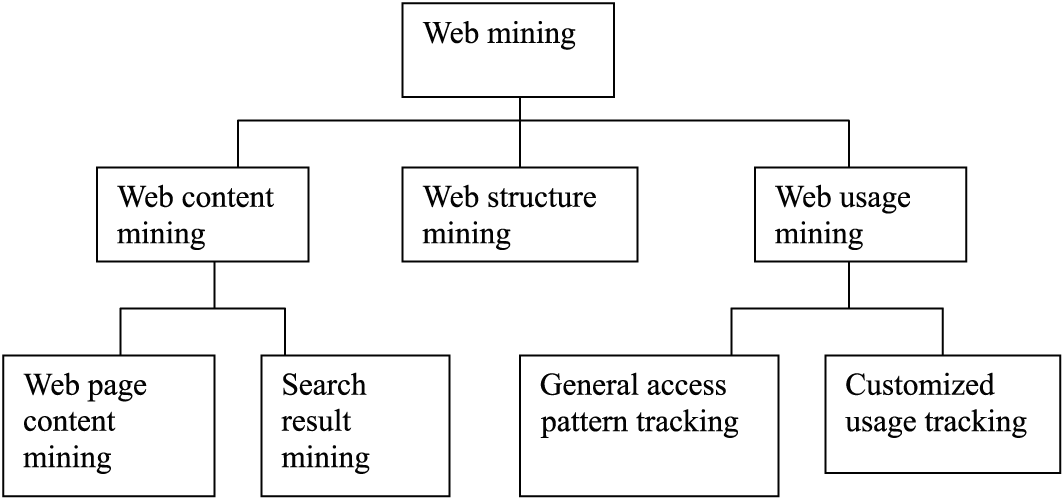
**Challenges for Effective Resource and KD**

Web mining is mining of data related to the World Wide Web. This may be the data actually present in Web pages or data related to Web activity. Web data can be:

* + Content of actual Web pages
  + Intra page structure includes the HTML or XML node for the page
  + Inter page structure is the actual linkage structure between Web pages
  + Usage data that describe how Web pages are accessed by visitors

User profile include demographic and registration information obtained about users

**Taxonomy of Web Mining**



**Web content mining** examines the content of Web pages as well as results of Web searching.  The content includes text as well as graphics data. Web content mining is further divided into Web page content mining and search results mining.

**Web page content mining** is traditional searching of Web pages via content, while Search results mining is a further search of pages found from a previous search.

**With Web structure mining**, information is obtained from the actual organization of pages on the Web.

**Web usage mining** looks at the logs of Web access. General access pattern tracking is a type of usage mining that looks at a history of Web pages visited. This usage may be general or may be targeted to specific usage or users. Usage mining also involves mining of these sequential patterns.

**Uses of Web Mining**:

* + - * Personalization for a user can be achieved by keeping track of previously accessed pages.
      * Web usage patterns can be used to gather business intelligence to improve sales and advertisement.
      * Collection of information can be done in new ways.
      * Testing of relevance of content and web site architecture can be done.

There are many **index-based Web search engines**.

These search the Web, **index Web pages**, and build and store huge keyword-based indices that help locate sets of Web pages containing certain keywords.

With such search engines, an experienced user may be able to quickly locate documents by providing a set of tightly constrained key words and phrases.

This indicates that a simple **keyword-based Web search engine** is not sufficient for Web resource Directory

However, Web mining can be used to substantially enhance the power of a Web search engine.

Since Web mining may identify authoritative Web pages, classify Web documents, and resolve many ambiguities and subtleties raised in keyword-based Web search.

In general, Web mining tasks can be classified into three categories:

* + - * Web Content Mining
      * Web Structure Mining
      * Web Usage Mining