MC 5403- ADVANCED DATABASES AND DATA MINING

UNIT II: DATAMINING & DATA PREPROCESSING

UNIT - II : Data Mining & Data Preprocessing

Syllabus:

UNIT II DATAMINING & DATA PREPROCESSING

Introduction to KDD process – Knowledge Discovery from Databases - Need for Data Preprocessing – Data Cleaning – Data Integration and Transformation – Data Reduction – Data Discretization and Concept Hierarchy Generation.

Table of Contents

Total Pages: 58

UNIT - II : Data Mining & Data Preprocessing

2.1 INTRODUCTION TO DATA MINING

2.1.1 Introduction to Data Mining

What Is Data Mining?

- **An iterative and interactive process of discovering " novel", " valid", "useful", "comprehensive and understandable patterns" and " model" s in MASSIVE data sources (databases).**
  - **Novel:** something we are not aware of
  - **Valid:** generalize to the future
  - **Useful:** some reaction is possible
  - **Understandable:** leading to insight
  - **Iterative:** many steps and many passes
  - **Interactive:** human is a part of the system.

- ***DM is an extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.***
- **DM is the finding interesting structure in data**
  - ***Structure:*** refers to statistical patterns, predictive models, hidden relationships.

Other Meaning information about DM

- 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
- ***A class of database application*** that analyze data in a database using tools which look for trends or anomalies.
- Data mining was invented by IBM.

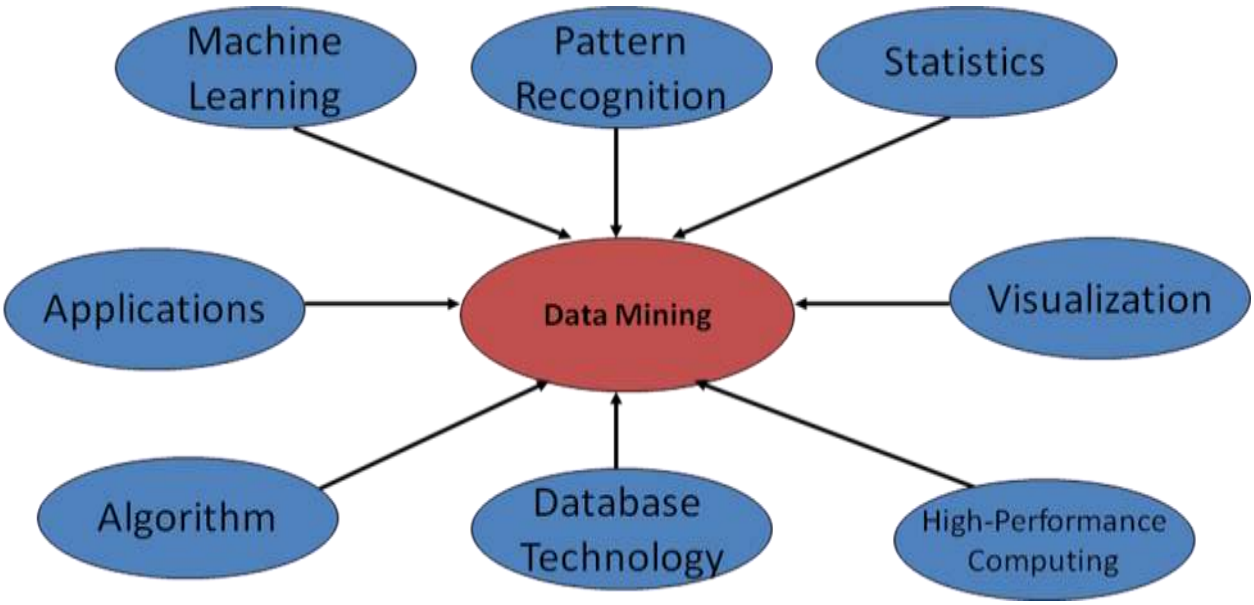
Data mining is multidisciplinary topic

- Data mining refers to **extracting or "mining" knowledge from large amounts of data.**

UNIT - II : Data Mining & Data Preprocessing

- Most data-mining problems and corresponding solutions ***have roots in classical data analysis.***
- **Data mining has its origins** in various disciplines, of which the two most important **are** ***statistics and machine learning.***

Data Mining: Confluence of Multiple Disciplines



Why Confluence (union) of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

UNIT - II : Data Mining & Data Preprocessing

2.1.2 Alternate meaning (names) of DATA MINING(DM)

- Data mining: a **misnomer?**
- **Alternative names**
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- **Watch out:** Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Data Mining Motivation

- Changes in the Business Environment
  - Customers becoming more demanding
  - Markets are saturated
- Databases today are huge:
  - More than 1,000,000 entities/records/rows
  - From 10 to 10,000 fields/attributes/variables
  - Gigabytes and terabytes
- Databases a growing at an unprecedented rate
- Decisions must be made rapidly
- Decisions must be made with maximum knowledge

Motivation: "Necessity is the Mother of Invention"

- **Data explosion problem**
  - Automated data collection tools and mature database technology lead to ***tremendous amounts of data stored in databases***, data warehouses and other information repositories
  - ***Very little data will ever be looked at by a human***.
  - ***We are drowning in data, but starving for knowledge!***
  - ***"The greatest problem of today is:***
    - how to teach people to ignore the irrelevant,?
    - how to refuse to know things, before they are suffocated.?
  - For too many facts are as bad as none at all.
- **We are drowning in data, but starving for knowledge!**

UNIT - II : Data Mining & Data Preprocessing

- **Solution: Data warehousing and data mining**
  - Data warehousing and on-line analytical processing
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

2.1.1 Need of DM

Why Data Mining?

- **The Explosive Growth of Data:** from terabytes to petabytes
  - **Data collection and data availability**
    - Automated data collection tools, database systems, Web, computerized society
  - **Major sources of abundant data**
    - **Business:** Web, e-commerce, transactions, stocks, …
    - **Science:** Remote sensing, bioinformatics, scientific simulation, …
    - **Society and everyone**: news, digital cameras, YouTube

- **We are drowning in data, but starving for knowledge!**
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets
  - Human analysis skills are inadequate:
    - Volume and dimensionality of the data
    - High data growth rate
  - Availability of:
    - Data
    - Storage
    - Computational power
    - Off-the-shelf software
    - Expertise

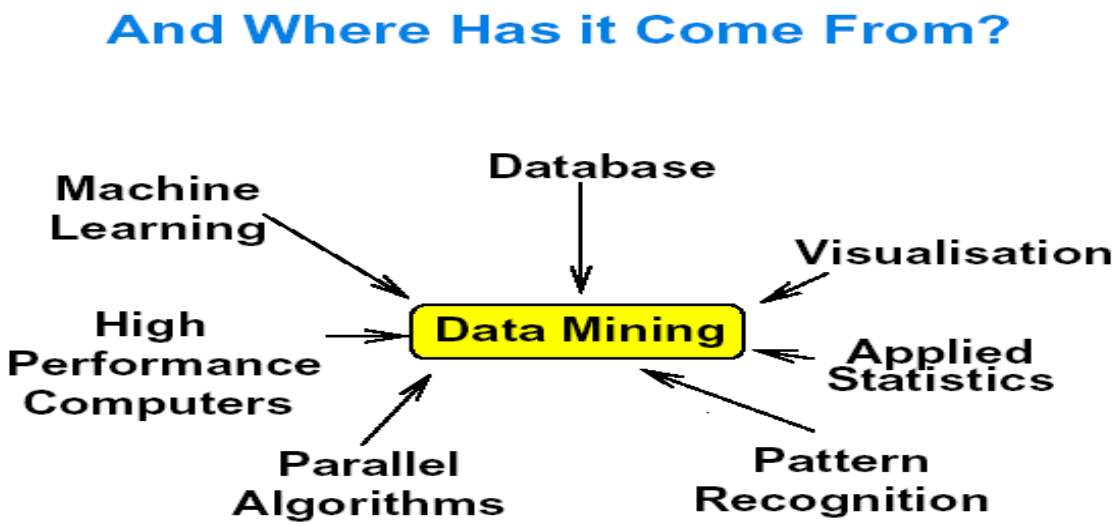
- **The Explosive Growth of Data:** from terabytes to petabytes
  - ***Data collection and data availability***
    - Automated data collection tools, database systems, Web, computerized society
  - ***Major sources of abundant data***
    - ***Business***: Web, e-commerce, transactions, stocks, …

- ***Science:*** Remote sensing, bioinformatics, scientific simulation, …
- ***Society and everyone***: news, digital cameras, YouTube

The need of DM:

Data Mining Objectives

- With data mining, it is possible to better:
  - manage product warranties,
  - predict purchases of retail stock,
  - unearth fraud,
  - determine credit risk, and
  - define new products and services.



Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions
  - Computers have become cheaper and more powerful
- Competitive Pressure is Strong

- Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

Examples: What is (not) Data Mining?

What is not Data Mining?

- Look up phone number in phone directory
- Query a Web search engine for information about "Amazon"

What is Data Mining?

- Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Data mining is <u>not</u>

- Brute-force crunching of bulk data
- "Blind" application of algorithms

- Going to find relationships where none exist
- Presenting data in different ways
- A database intensive task
- A difficult to understand technology requiring an advanced degree in computer science

2.1.2 Goals of Data Mining

What are the goals of Data Mining?

There are four major goals of DM.

- **Prediction**: To foresee the possible future situation on the basis of previous events.
  - *Given sales recordings from previous years:*
    - *can we predict what amount of goods we need to have in stock for the forthcoming season?*
- **Description**: What is the reason that some events occur?
  - *What are the reasons for the cars of one producer to sell better that equal products of other producers?*

- **Verification**: We think that some relationship between entities occur.
  - *Can we check if (and how) the thread of cancer is related to environmental conditions?*
- **Exception detection**: There may be situations (records) in our databases that correspond to something unusual.
  - I*s it possible to identify credit card transactions that are infact frauds?*

Data Mining: Classification Schemes

- Decisions in data mining
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted

Decisions in Data Mining

- **Databases to be mined**

  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

Data Mining Tasks

The Major DM tasks are:

1) ***Description Tasks***:- What has happened?
2) ***Diagnostic Tasks*** :- Why it happened?
3) ***Prediction Tasks***:- What will happen?
4) ***Prescriptive Tasks***:- What to do about it?

- Note:1) & 2) are current scenario. 3) & 4) are future expectation.

Common data mining tasks are:

  - Classification [Predictive]
  - Clustering [Descriptive]
  - Association Rule Discovery [Descriptive]
  - Sequential Pattern Discovery [Descriptive]
  - Regression [Predictive]
  - Deviation Detection [Predictive]

- Prediction Tasks
  - Use some variables to predict unknown or future values of other variables
- Description Tasks
  - Find human-interpretable patterns that describe the data.

UNIT - II : Data Mining & Data Preprocessing

2.1.3 Data Mining Activities

The Data-mining activities are of one of two categories:

- 1) ***Predictive data mining***, which *produces the model of the system described by the* given data set, or
- 2) ***Descriptive data mining***, which *produces new, nontrivial information based on* the available data set.

- **Examples of tasks addressed by Data Mining**
  - Predictive Modeling (classification, regression)
  - Segmentation (Data Clustering)
  - Summarization
  - Visualization

- On the predictive end of the spectrum:
  - the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks.

- On the other, descriptive, end of the spectrum:
  - the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets.

- The goals of prediction and description are achieved by using data-mining techniques.

2.1.4 DM (Data Mining) Vs DW (Data Warehousing)

DM	DW
I. Provides the Enterprise with intelligence	**I. Provides the Enterprise with a memory.**

II.	Data Mining tools can analyze massive databases to deliver answers to questions such as, " Which customers are most likely to respond to my next promotional mailing, and why?"	II.	DM is acting as a place holder to massive data grinded from DM activities.
III.	Data Mining does not require that a Data Warehouse be built. Often, data can be downloaded from the operational files to flat files that contain the data ready for the data mining analysis.	III.	DM is a physical element may support DM (if required)
IV.	DM is equipped with effective techniques for DM to be consistent.	IV.	A data warehouse is well equipped for providing data for mining.
V.	Major challenge to exploit data mining is identifying suitable data to mine.	V.	Major challenge to DW is hold suitable data to mine.
VI.	Data mining requires single, separate, clean, integrated, and self-consistent source of data.	VI. Data warehouse holds single, separated, cleaned, integrated, and self-consistent data.	
VII.	DM itself has patterns, rules for providing Query capabilities.	VII. Selecting relevant subsets of records and fields for data mining requires query capabilities of the data warehouse.	
VIII.	Results of a data mining study are useful if there is some way to further investigate the uncovered patterns.	VIII. Data warehouses provide capability to go back to the data source.	

2.1.5 Major phases of Data Mining

UNIT - II : Data Mining & Data Preprocessing

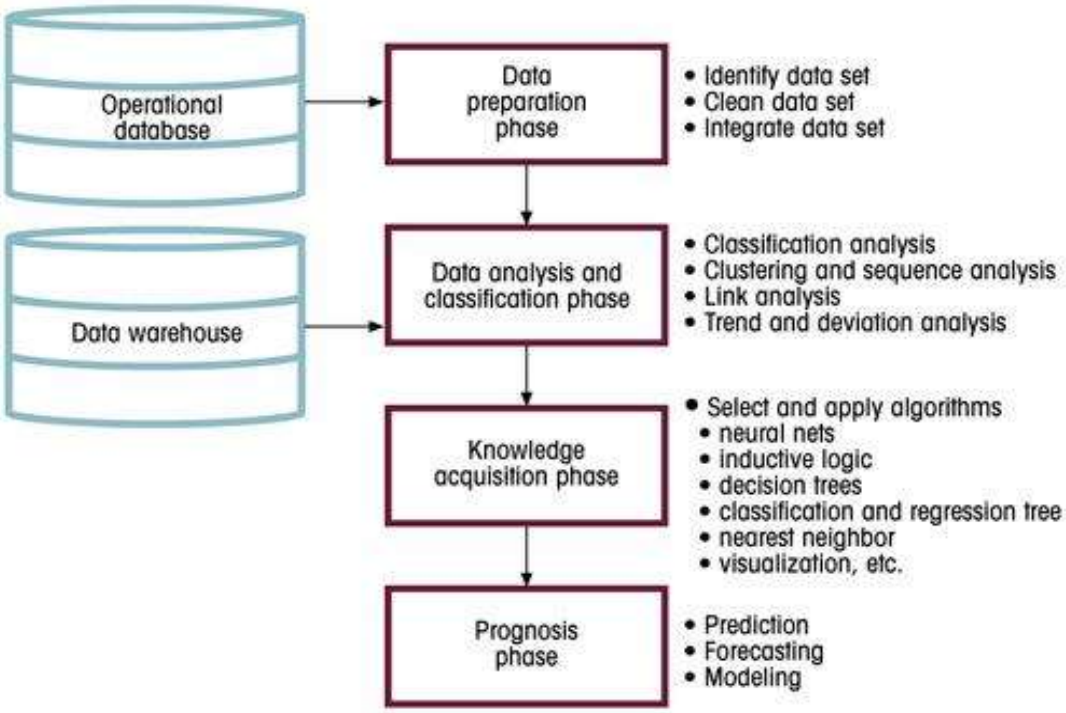


FIGURE 13.23 DATA-MINING PHASES

Four Phases of Data Mining

1. **Data Preparation**
   - Identify and cleanse data sets.
   - Data Warehouse is usually used for data mining operations.
2. **Data Analysis and Classification**
   - Identify common data characteristics or patterns using
     - Data groupings, classifications, clusters, or sequences.
     - Data dependencies, links, or relationships.
     - Data patterns, trends, and deviations.

3. **Knowledge Acquisition**
   - Select the appropriate modeling or knowledge acquisition algorithms.
   - Examples: neural networks, decision trees, rules induction, genetic algorithms, classification and regression tree, memory-based reasoning, or nearest neighbor and data visualization).
4. **Prognosis**

- Predict future behavior and forecast business outcomes using the data mining findings.

Data Mining Yielding

- Data mining yields five basic type of information:
  - Association - occurrences are linked to a single event.
  - Sequences - events are linked over time.
  - Classification - patterns are recognized that describe the characteristics of a group, such as customers who cancel credit cards
  - Clustering - discovers undiscovered groupings ``*Buyers of expensive sport cars are typically young urban professionals whereas luxury sedans are bought by elderly wealthy persons."*
  - Forecasting - estimates future value such as inventory turnover

List out the Data Mining Application areas

- **Science :‰** astronomy, bioinformatics, drug discovery, … „
- **Business**: ‰ advertising, CRM (Customer Relationship management), investments, manufacturing, sports/entertainment, telecom, eCommerce, targeted marketing, health care, … „
- **Web:** ‰ search engines, bots, … „
- **Government ‰**: law enforcement, profiling tax cheaters, anti-terror(?)

2.1.6 Influences of DM in real time scenarios

What are influences of DM in real time scenarios?

Scenario 1:

Customer Attrition: Case Study „

- **Situation:** Attrition rate at for mobile phone customers is around 25-30% a year!
- **Task:** " Given customer information for the past N months, predict who is likely to attrite next month." Also, estimate customer value and what is the cost effective offer to be made to this customer.
- **Customer Attrition Results**
- Verizon Wireless built a customer data warehouse
- Identified potential attriters

- Developed multiple, regional models
- Targeted customers with high propensity to accept the offer
- Reduced attrition rate from over 2%/month to under 1.5%/month (huge impact, with >30 M subscribers) (Reported in 2003)

Scenario 2:

Assessing Credit Risk: Case Study

- **Situation:** Person applies for a loan
- **Task:** Should a bank approve the loan?
- **Note:** People who have the best credit don't need the loans, and people with worst credit are not likely to repay. Bank's best customers are in the middle
-
- **Credit Risk - Results**
- Banks develop credit models using variety of machine learning methods.
- Mortgage and credit card proliferation are the results of being able to successfully predict if a person is likely to default on a loan
- Widely deployed in many countries

Scenario 3:

Successful e-commerce – Case Study „

- A person buys a book (product) at Amazon.com.
- **Task:** Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought: ‰
- Customers who bought "**Advances in Knowledge Discovery and Data Mining**", also bought "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations** "
- Recommendation program is quite successful

Scenario 4:

Unsuccessful e-commerce case study

- **Security and Fraud Detection - Case Study „**
- Credit Card Fraud Detection „
- Detection of Money laundering ‰ FAIS (US Treasury)
- Securities Fraud ‰ NASDAQ KDD system
- Phone fraud ‰ AT&T, Bell Atlantic, British Telecom/MCI
- Bio-terrorism detection at Salt Lake Olympics 2002

UNIT - II : Data Mining & Data Preprocessing

2.2 INTRODUCTION TO KDD PROCESS

2.2.1 Introduction

Data Understanding: Needs

- **I can't find the data I need**
  - data is scattered over the network
  - many versions, subtle differences
- **I can't get the data I need**
  - need an expert to get the data.
- **I can't understand the data I found**
  - available data poorly documented
- **I can't use the data I found**
  - results are unexpected
  - data needs to be transformed from one form to other

Data Understanding: Quantity

- **Number of instances** (records, objects)
  - *Rule of thumb: 5,000 or more desired*
  - if less, results are less reliable; use special methods (boosting, …)
- **Number of attributes** (fields)
  - *Rule of thumb: for each attribute, 10 or more instances*
  - If more fields, use feature reduction and selection
- **Number of targets**
  - *Rule of thumb: >100 for each class*
  - if very unbalanced, use stratified sampling

Data Understanding: Relevance

- What data is available for the task?
- Is this data relevant?
- Is additional relevant data available?
- How much historical data is available?
- Who is the data expert ?

UNIT - II : Data Mining & Data Preprocessing

2.2.2 KDD (Knowledge Discovery Databases)

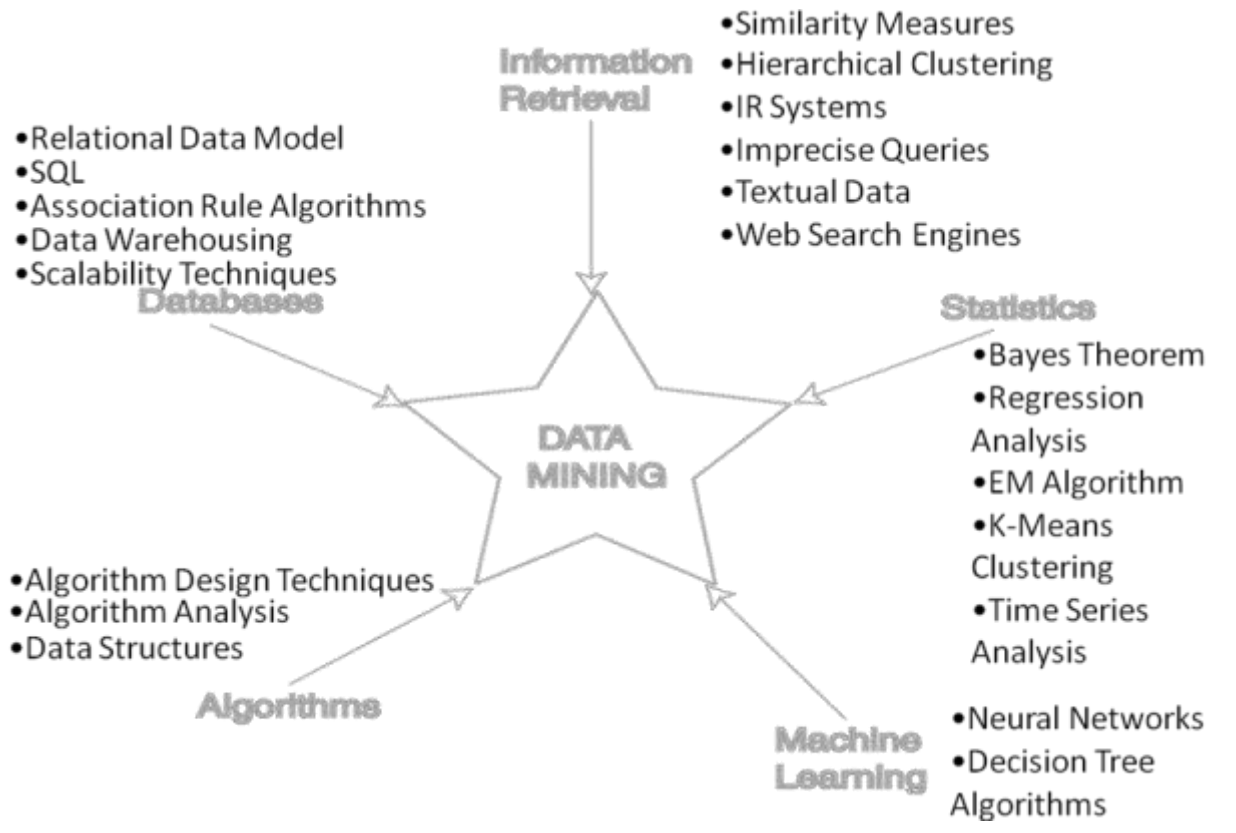
What is KDD?

- **The basic task of KDD is to extract knowledge (or information) from lower level data (databases).**
- **"Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."**
  by Fayyad, Piatetsky-Shapiro and Smyth

Data Mining vs. KDD

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

Data Mining Development



UNIT - II : Data Mining & Data Preprocessing

DM to KDD

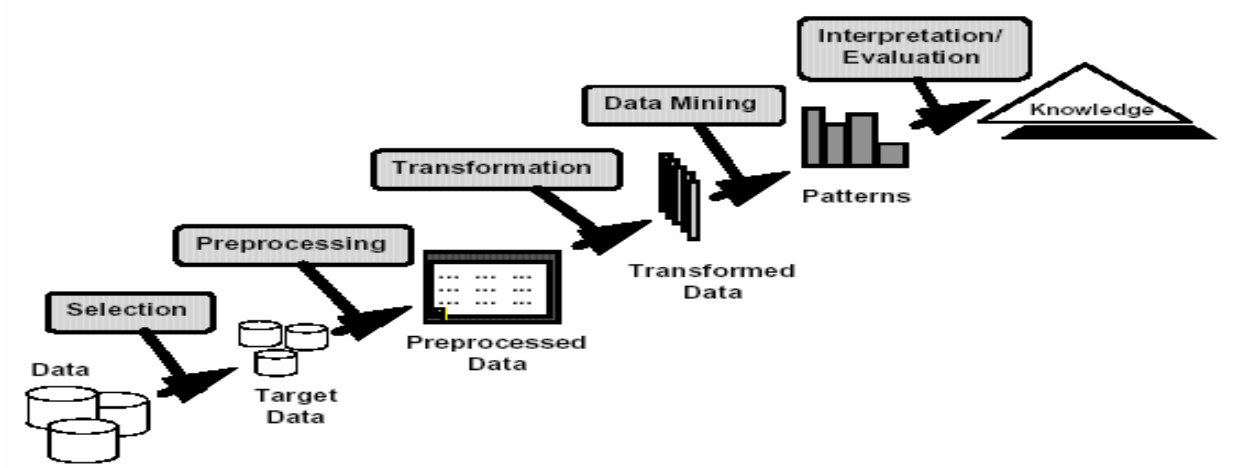
- **Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to:**
- **the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases.**
- **While data mining and KDD are frequently treated as synonyms.**
  - **Data mining is actually part of the knowledge discovery process.**

- **Knowledge Discovery (KDD) Process**
  - **This is a view from typical database systems and data warehousing communities**
  - **Data mining plays an essential role in the knowledge discovery process**

Knowledge Discovery (KDD) Processes

- **KDD key processes**
  1. **Learning the application domain.**
  2. **Relevant prior knowledge and goals of application.**
  3. **Creating a target data set: data selection.**
  4. **Data cleaning and preprocessing: (may take 60% of effort!)**
  5. **Data reduction and transformation.**
  6. **Data mining .**
  7. **Summarization, classification, regression, association, clustering.**
  8. **Pattern evaluation and knowledge presentation.**
  9. **Use of discovered knowledge.**

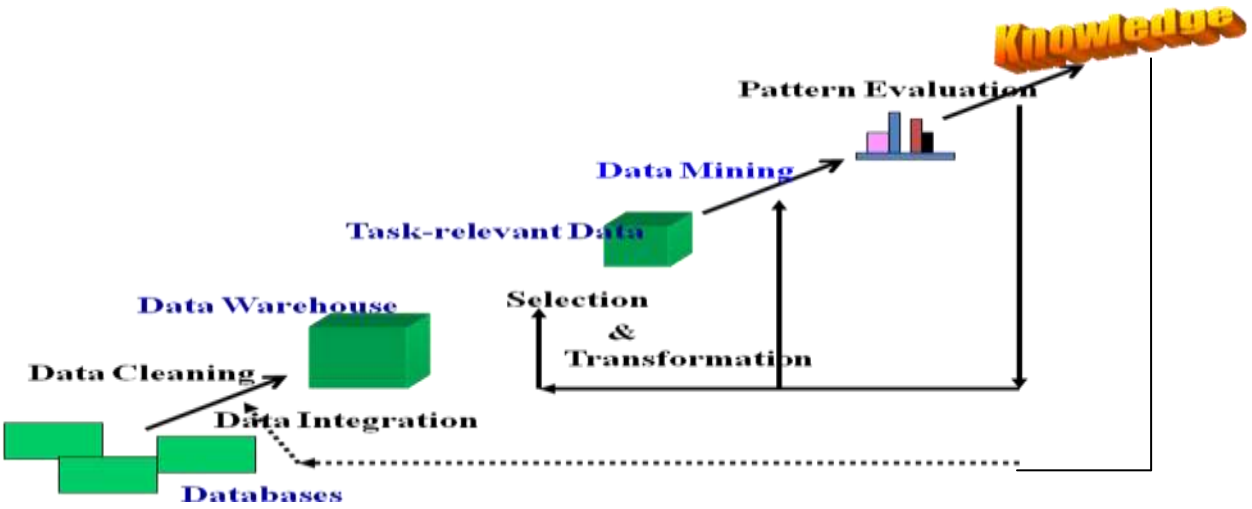
UNIT - II : Data Mining & Data Preprocessing



Knowledge Discovery (KDD) Process of selected DM

Knowledge Discovery (KDD) Process – Enhanced

- **This is a view from typical database systems and data warehousing communities**
- **Data mining plays an essential role in the knowledge discovery process**



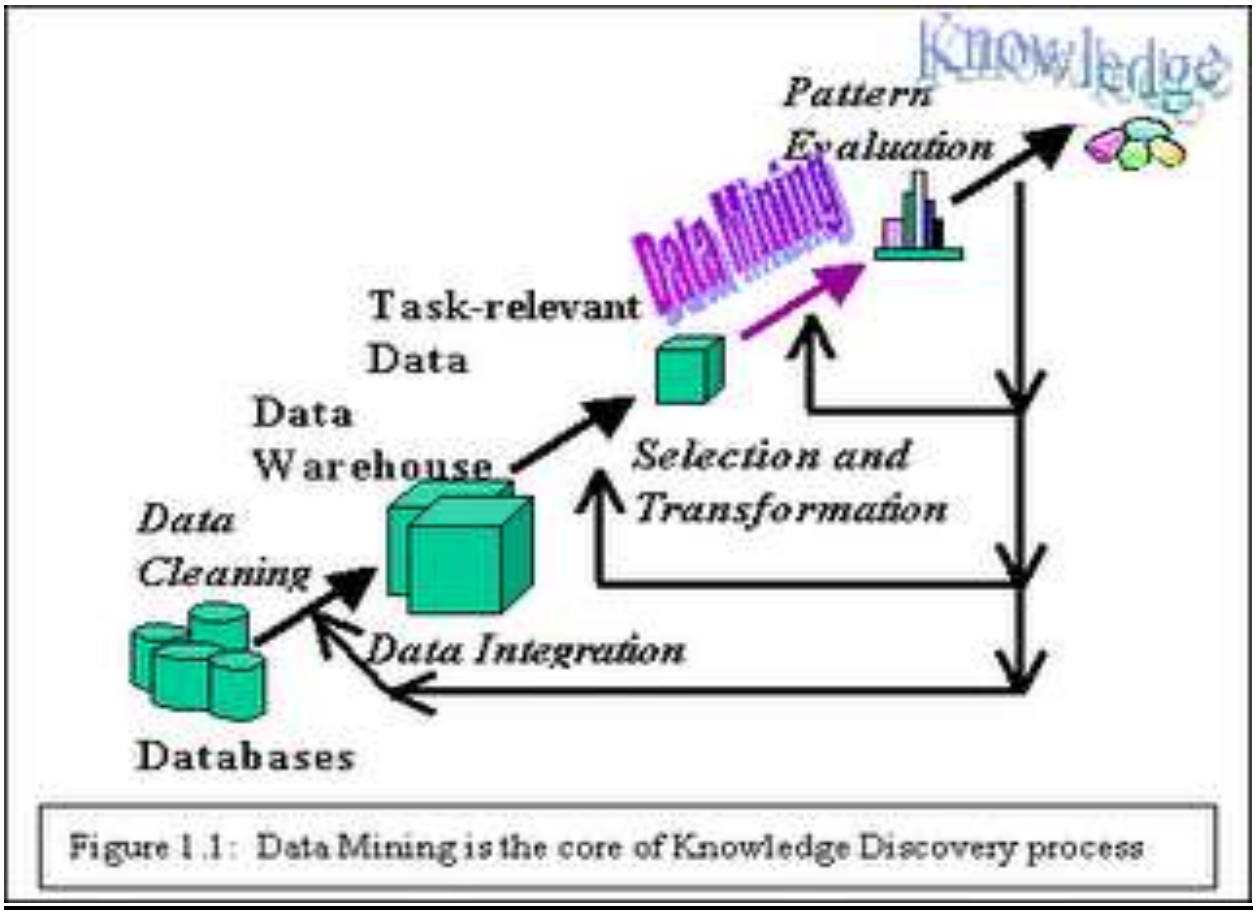


Figure 1.1: Data Mining is the core of Knowledge Discovery process

- The Knowledge Discovery in Databases process *comprises of a few steps leading from raw data collections to some form of new knowledge.*

Knowledge discovery as a process is consists of an iterative sequence of the following steps:

1. **Data cleaning**
2. **Data integration**
3. **Data selection**
4. **Data transformation**
5. **Data mining**
6. **Pattern evaluation**
7. **Knowledge presentation**

KDD Process: Several Key Steps

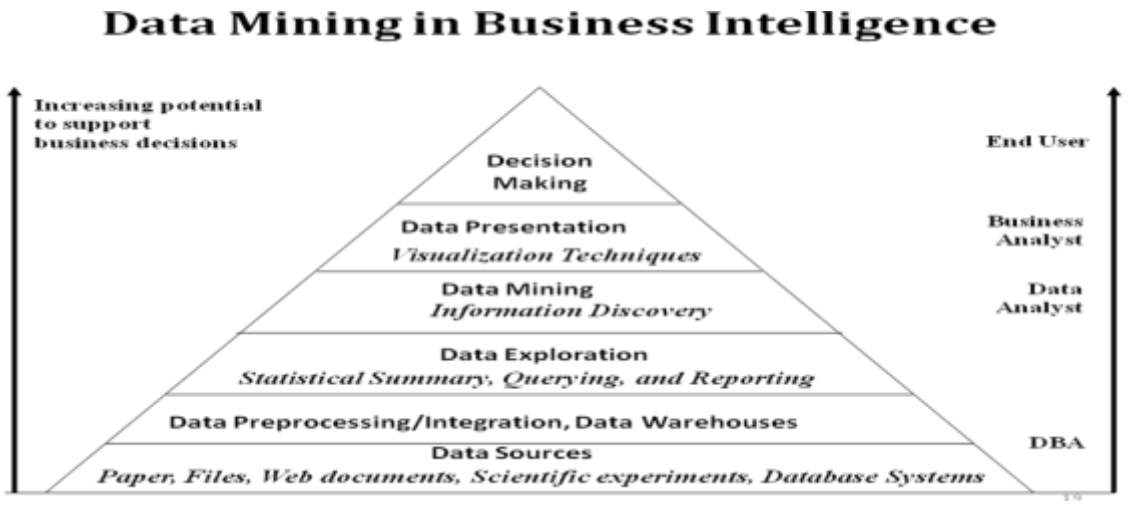
1. Data cleaning (to remove noise and inconsistent data)

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

- The data mining step may interact with the user or a knowledge base.

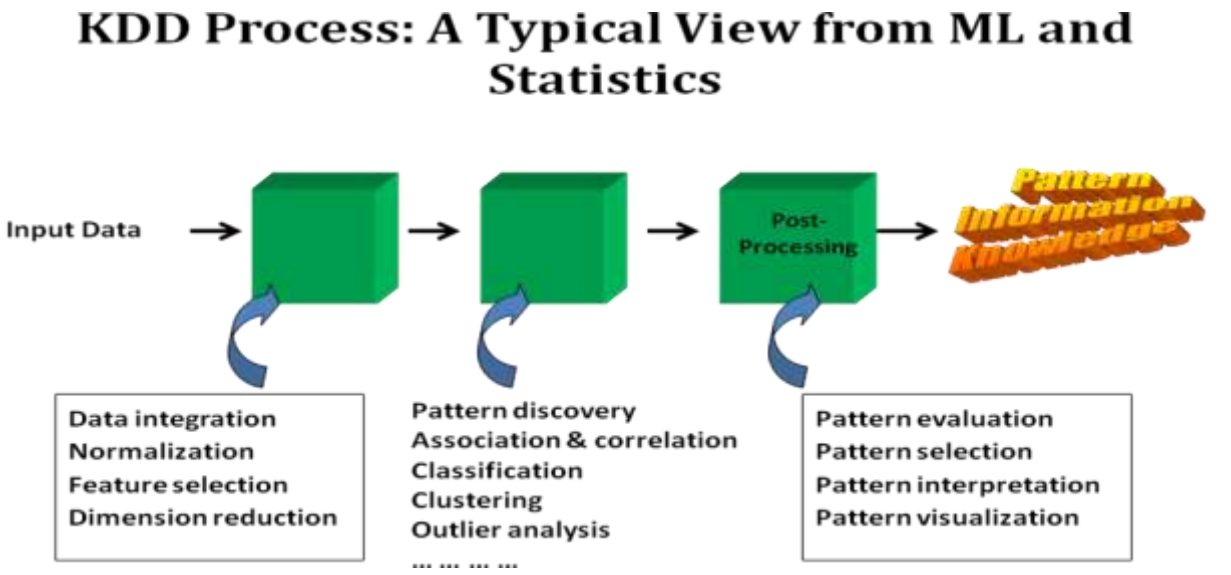
KDD Process : example: A Web Mining Framework

- **Web mining usually involves**
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base



Example: Mining vs. Data Exploration

- **Business intelligence view**
  - Warehouse, data cube, reporting but not much mining
- **Business objects vs. data mining tools**
- **Supply chain example:** tools
- **Data presentation**
- **Exploration**



- **This is a view from typical machine learning and statistics communities**

Multi-Dimensional View of Data Mining

- ***Data to be mined***
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- ***Knowledge to be mined (or: Data mining functions)***
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels

UNIT - II : Data Mining & Data Preprocessing

- ***Techniques utilized***
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- ***Applications adapted***
  - **Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.**

Data Mining: On What Kinds of Data?

- **Database-oriented data sets and applications**
  - Relational database, data warehouse, transactional database
- **Advanced data sets and advanced applications**
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Evaluation of Knowledge

- **Are all mined knowledge interesting?**
  - One can mine tremendous amount of "patterns" and knowledge
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient(brief), …

- **Evaluation of mined knowledge → directly mine only interesting knowledge?**
  - Descriptive (expressive) vs. predictive (projecting)
  - Coverage
  - Typicality vs. novelty(innovation)
  - Accuracy
  - Timeliness
  - **…**

UNIT - II : Data Mining & Data Preprocessing

2.3 KNOWLEDGE DISCOVERY FROM DATABASES

2.3.1 Introduction

- Knowledge Discovery in Databases (KDD) is **an automatic, exploratory analysis and modeling of large data repositories.**
- **KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets.**

- Data Mining (DM) is the **core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns.**
- The model is **used for understanding phenomena from the data, analysis and prediction.**

- Knowledge Discovery in Databases is **the process of searching for hidden knowledge in the massive amounts of data that we are technically capable of generating and storing.**

- Data, in its **raw form, is simply a collection of elements, from which little knowledge can be gleaned.**
- With the **development of data discovery techniques the value of the data is significantly improved.**

2.3.2 Evaluation of Knowledge Discovery from Databases (KDD)

- KDD has evolved from interaction and cooperation among such different fields as
  - machine learning,
  - pattern recognition,
  - database,
  - statistics,
  - artificial Intelligence,
  - knowledge representation, and
  - knowledge acquisition for intelligent systems.
- The main idea in KDD is to discover a high level knowledge (abstract knowledge) from lower levels of relatively raw data, or to discover a higher level of interpretation and abstraction than those previously known.

UNIT - II : Data Mining & Data Preprocessing

- The KDD process is interactive and iterative.
- One has to make several decisions in the process of KDD.

- Data mining, the pattern extraction phase of KDD, can take on many forms, the choice dependent on the desired results.
- KDD is a multi-step process that facilitates the conversion of data to useful information.

Evaluation of Knowledge Discovery from Databases (KDD)

- **The goal is:**
  - to distinguish from unprocessed data,
    - something that may not be obvious
    - but is valuable or enlightening in its discovery.
- Extraction of knowledge from raw data is accomplished **by applying Data Mining methods.**
- KDD has a much **broader scope**, of which data mining is one step in a multidimensional process.
- DM & KDD have been ***attracting a significant amount of research, industry, and media attention of late.***
- the KDD field is ***concerned with the development of methods and techniques for making sense of data.***

- The basic problem addressed by the KDD process is:
  - **one of mapping low-level data**
    - (which are typically too voluminous to understand and digest easily)
  - **into other forms that might be more compact**
    - (for example, a short report),
  - **more abstract**
    - (for example, a descriptive approximation or model of the process that generated the data),
  - **or more useful**
    - (for example, a predictive model for estimating the value of future cases).

2.3.3 Why Do We Need KDD?

- The traditional method of turning data into knowledge relies on manual analysis and interpretation.
  - For example, in the health-care industry, it is common for specialists to periodically analyze current trends and changes in health-care data, say, on a quarterly basis.
- The specialists then provide a report detailing the analysis to the sponsoring health-care organization; this report becomes the basis for future decision making and planning for health-care management.
- In a totally different type of application, planetary geologists sift through remotely sensed images of planets and asteroids, carefully locating and cataloging such geologic objects of interest as impact craters.

- KDD includes **multidisciplinary activities.**
- This encompasses ***data storage and access, scaling algorithms to massive data sets and interpreting results.***
- The ***data cleansing and data access process included in data warehousing facilitate the KDD*** process.
- **Artificial intelligence also supports KDD** by discovering empirical laws from experimentation and observations.
- Data mining is a step in the KDD process :
  - that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data.

- Be it science, marketing, finance, health care, retail, or any other field, the classical approach to data analysis relies fundamentally on one or more analysts becoming intimately familiar with the data and serving as an interface between the data and the users and products.
- There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data.
- The basic problem addressed by the KDD process is one of mapping low-level data into other forms that might be more compact, more abstract, or more useful.
- Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data.

2.3.4 DM as KDD

- It is important to note that **KDD is not accomplished without human interaction.**
- Two main types of Data Mining:
  - **verification-oriented** (the system verifies the user's hypothesis) and
  - **discovery-oriented** (the system finds new rules and patterns autonomously).
- ***Discovery methods are those that automatically identify patterns in the data.***
- The ***discovery method branch consists of prediction methods versus description methods.***
- Descriptive methods are oriented to data interpretation, which focuses on understanding
  - (by visualization for example) the way the underlying data relates to its parts.

- Prediction-oriented methods aim to build a behavioral model, which obtains new and unseen samples and is able to predict values of one or more variables related to the sample.
- It also develops patterns which form the discovered knowledge in a way which is understandable and easy to operate upon.
- Some prediction-oriented methods can also help provide understanding of the data.

- Most of the discovery-oriented Data Mining techniques (quantitative in particular) are ***based on inductive learning, where a model is constructed, explicitly or implicitly, by generalizing from a sufficient number of training examples.***
- The underlying assumption of the ***inductive approach is that the trained model is applicable to future unseen examples.***

- Verification methods, on the other hand***, deal with the evaluation of a hypothesis proposed by an external source*** (like an expert etc.).
- These methods include the most common methods of traditional statistics,
  - like goodness of fit test, tests of hypotheses (e.g., t-test of means), and analysis of variance (ANOVA).

UNIT - II : Data Mining & Data Preprocessing

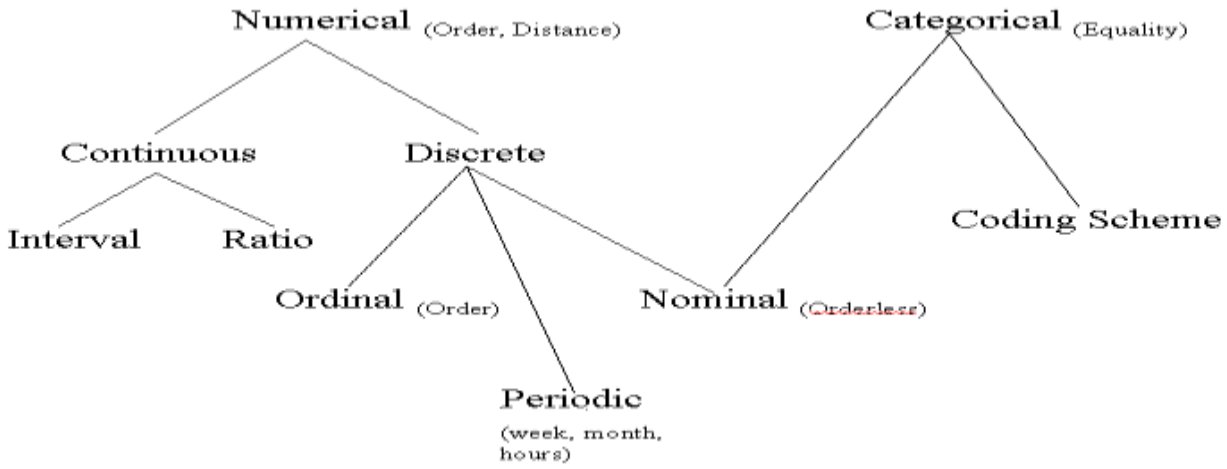
2.4 NEED FOR DATA PREPROCESSING

2.4.1 Introduction to Data Preprocessing

- **What is Data Preprocessing?**
  - **Data preprocessing describes**
    - *any type of processing performed on raw data*
    - *to prepare it for another processing procedure.*
  - **Data preprocessing transforms**
    - *the data into a format* that will be
    - more easily and effectively processed for the purpose of the user

What does *Data Preprocessing* mean?

A1	A2	…	An	C



2.4.2 Need of *Data Preprocessing*

- **Real-world data is often:**
  - *incomplete*,
  - *inconsistent*, and/or
  - lacking in certain behaviors or trends, and
  - is likely to *contain many errors.*
- *Data preprocessing is a proven method* of resolving such issues.
- Data preprocessing *prepares raw data for further processing.*

UNIT - II : Data Mining & Data Preprocessing

Data Types and Forms is the deciding factor

- **Attribute-value data**
- **Data types**
  - numeric, categorical (see the hierarchy for its relationship)
  - static, dynamic (temporal)
- **Other kinds of data**
  - distributed data
  - text, Web, meta data
  - images, audio/video

Data Quality: Need of *Data Preprocessing* ?

- ***Measures for data quality: (*A multidimensional view)**
  - **Accuracy:** correct or wrong, accurate or not
  - **Completeness:** not recorded, unavailable, …
  - **Consistency:** some modified but some not, dangling, …
  - **Timeliness:** timely update?
  - **Believability:** how trustable the data are correct?
  - **Interpretability:** how easily the data can be understood?

2.4.3 Importance of Data Preprocessing

Why preprocess the data? (or) Why Is Data Preprocessing Important?

- **No quality data, no quality mining results!**
  - *Quality decisions must be based on quality data*
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse *needs consistent integration of quality data*
- **Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse**
- **Data in the real world is dirty**
  - **incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - **e.g., occupation=“ ”**
  - **noisy:** containing errors or outliers

- **e.g., Salary="-10"**

– **inconsistent:** containing discrepancies in codes or names
  - **e.g., Age="42" Birthday="03/07/1997"**
  - **e.g., Was rating "1,2,3", now rating "A, B, C"**
  - **e.g., discrepancy between duplicate records**

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - **Database data (extended-relational, object-oriented, heterogeneous, legacy)**
  - **data warehouse**
  - **transactional data**
  - **stream, spatiotemporal**
  - **time-series,**
  - **sequence**
  - **text and web**
  - **multi-media**
  - **graphs &**
  - **social and information networks**

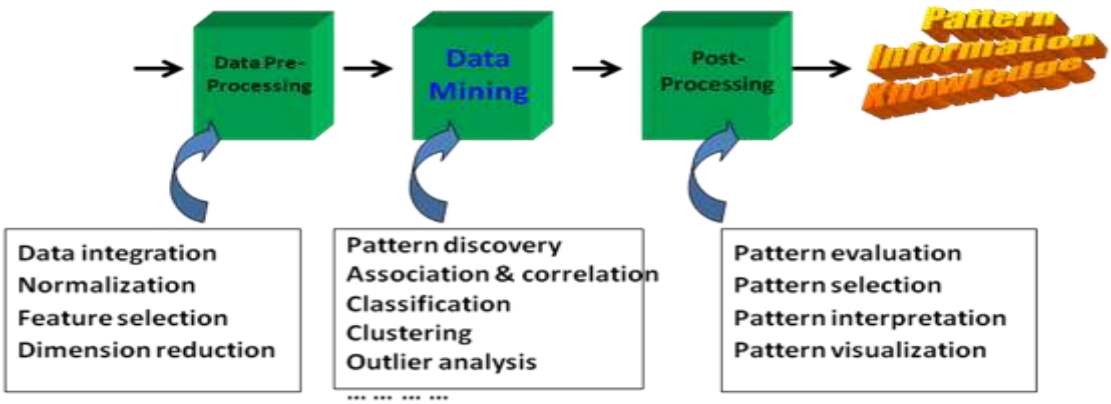
- **Knowledge to be mined (or Data mining functions)**
  - **Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.**
  - **Descriptive vs. predictive data mining**
  - **Multiple/integrated functions and mining at multiple levels**

Multi-Dimensional View of Data Mining

- **Techniques utilized**
  - **Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.**
- **Applications adapted**
  - **Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.**

Major Tasks in Data Preprocessing

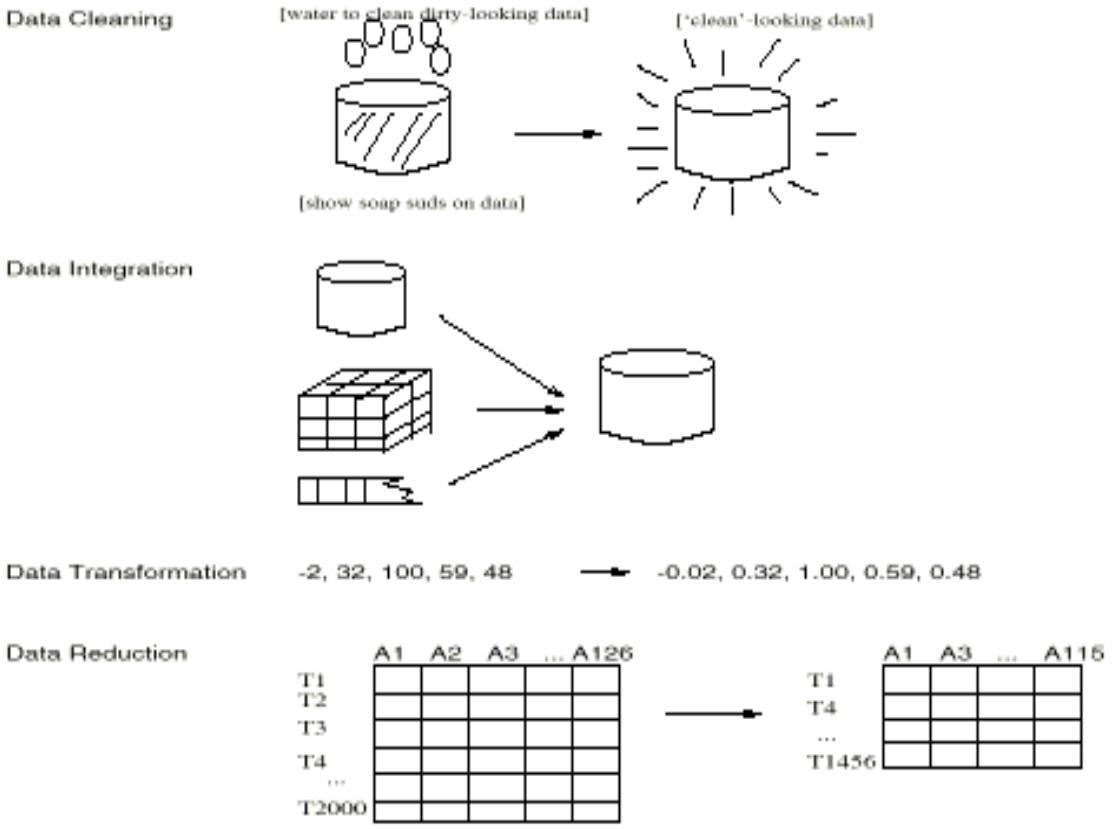
- **Data cleaning**
  - **Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies**
- **Data integration**
  - **Integration of multiple databases, data cubes, or files**
- **Data transformation**
  - **Normalization and aggregation**
- **Data reduction**
  - **Obtains reduced representation in volume but produces the same or similar analytical results**
- **Data discretization**
  - **Part of data reduction but with particular importance, especially for numerical data**



- This is a view from typical machine learning and statistics communities

Forms of data preprocessing

UNIT - II : Data Mining & Data Preprocessing



Different tools and methods used for preprocessing

- ***includes:***
  - ***sampling*, which selects a representative subset from a large population of data;**
  - ***transformation*, which manipulates raw data to produce a single input;**
  - ***denoising*, which removes noise from data;**
  - ***normalization,* which organizes data for more efficient access; and**
  - **feature extraction, which pulls out specified data that is significant in some particular context.**

UNIT - II : Data Mining & Data Preprocessing

2.5 DATA CLEANING

2.5.1 Introduction

What is DC (Data Cleaning) ? Why is it Important?

- **"Data cleaning is one of the three biggest problems in data warehousing"—Ralph Kimball**
- **"Data cleaning is the number one problem in data warehousing"—DCI survey**

- **Data cleaning tasks**
  - **Fill in missing values**
  - **Identify outliers and smooth out noisy data**
  - **Correct inconsistent data**
  - **Resolve redundancy caused by data integration**

Data cleaning : What is it?

2.5.2 DC activities

- **Fill in missing values,**
- **smooth noisy data,**
- **identify or remove outliers, and**
- **resolve inconsistencies**

Data Cleaning: Acquisition

- **Data can be in DBMS**
  - **ODBC, JDBC protocols**
- **Data in a flat file**

  - **Fixed-column format**
- **Delimited format: tab, comma ",", other**
  - **E.g. C4.5 and Weka "arff"use comma-delimited data**
- **Attention: Convert field delimiters inside strings**
  - **Verify the number of fields before and after**

Data Cleaning: Example

- **Original data (fixed column format)**
- **000000000130.06.19971979--10-3080145722 #000310 111000301.01.000100000000004 0000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000. 000000000000000.000000000000000.0000000...… 000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.00 0000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.0000 00000000000.000000000000000.000000000000000.000000000000000.00 0000000000300.00 0000000000300.00**
- **Clean data**
  - **0000000000300.000000000001,199706,1979.833,8014,5722 , ,#000310 …. ,111,03,000101,0,04,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0300,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0300,0300.00**

Data Cleaning: Metadata

- **Field types:**
  - **binary, nominal (categorical),ordinal, numeric, …**
  - **For nominal fields: tables translating codes to full descriptions**
- **Field role:**
  - **input : inputs for modeling**
  - **target : output**
  - **id/auxiliary : keep, but not use for modeling**
  - **ignore : don't use for modeling**
  - **weight : instance weight**

  - …
- **Field descriptions**

2.5.3 Missing Data

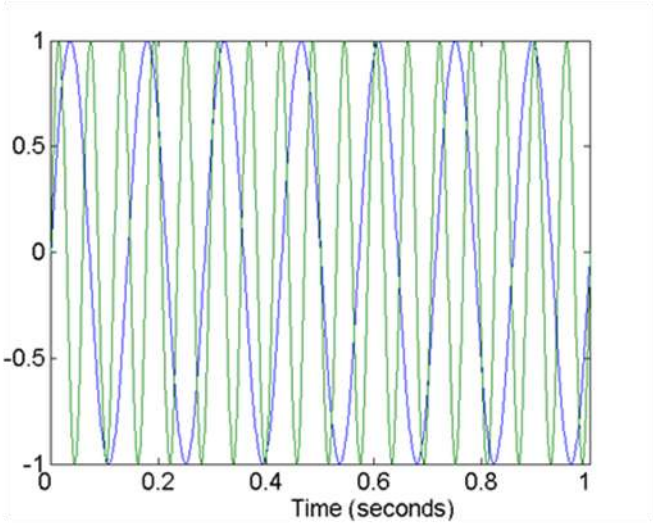
- **Data is not always available**
  - **E.g., many tuples have no recorded value for several attributes, such as customer income in sales data**
- ***Missing data may be due to***
  - **equipment malfunction**
  - **inconsistent with other recorded data and thus deleted**
  - **data not entered due to misunderstanding**
  - **certain data may not be considered important at the time of entry**
  - **not register history or changes of the data**
- **Missing data may need to be inferred (indirect or conditional).**

How to Handle Missing Data?

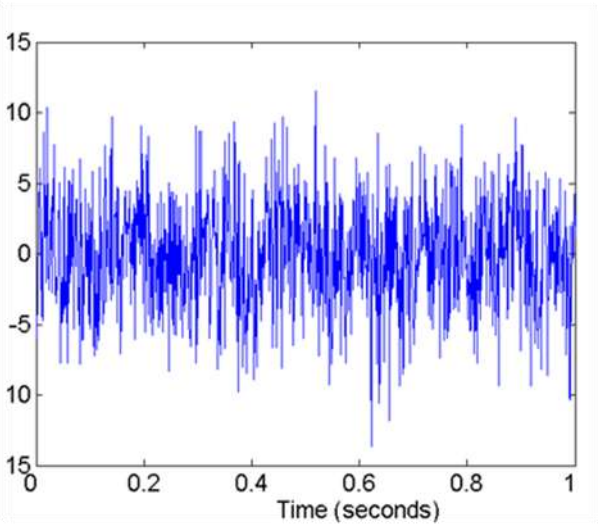
- ***Ignore the tuple:*** **usually done when class label is missing**
- **Fill in the missing value *manually***
- **Use a *global constant to fill in the missing value*: ex. "unknown"**
- ***Use the attribute mean to fill in the missing value***
- ***Use the attribute mean for all samples belonging to the same class to fill in the missing value***
- ***Use the most probable value to fill in the missing value:*** **inference-based such as Bayesian formula or decision tree**

2.5.4 Noisy Data

- **Noise refers to modification of original values**
  - **Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen**



Two Sine Waves



Two Sine Waves + Noise

- ***Noise:*** **random error or variance in a measured variable**
- ***Incorrect attribute values may due to***
  - **faulty data collection instruments**
  - **data entry problems**
  - **data transmission problems**
  - **technology limitation**
  - **inconsistency in naming convention**
- ***Other data problems which requires data cleaning***
  - ***d*uplicate records**
  - **incomplete data**
  - **inconsistent data**

2.5.5 Methods for DC

2.5.5.1 Introduction

How to Handle Noisy Data?

- **Binning method:**
  - **first sort data and partition into (equal-frequency) bins**
  - **then one can smooth by bin means, smooth by bin median, smooth by bin boundaries**

UNIT - II : Data Mining & Data Preprocessing

- **Clustering**
  - **detect and remove outliers**
- **Regression**
  - **smooth by fitting the data to a regression functions – linear regression**

Simple Discretization Methods: Binning

- ***Equal-width (distance) partitioning:***
  - **It divides the range into *N* intervals of equal size: uniform grid**
  - **if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: *W* = (*B*-*A*)/*N*.**
  - **The most straightforward**
  - **But outliers may dominate presentation**
  - **Skewed data is not handled well.**
- ***Equal-depth (frequency) partitioning:***
  - **It divides the range into *N* intervals, each containing approximately same number of samples**
  - **Good data scaling**
  - **Managing categorical attributes can be tricky.**

Binning Methods for Data Smoothing

- **Sorted data for price (in dollars):**

  4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Partition into (equi-depth) bins: (size = 4)

- **Bin 1: 4, 8, 9, 15**
- **Bin 2: 21, 21, 24, 25**
- **Bin 3: 26, 28, 29, 34**

*** *Smoothing* by bin means:**

- **Bin 1: 9, 9, 9, 9 (3 + 8 + 9 + 15) /4 = 9**
- **Bin 2: 23, 23, 23, 23 (21+ 21+ 24 + 25)/4 = 23**
- **Bin 3: 29, 29, 29, 29 (26 + 28 + 29 + 34)/4 = 29**

*** *Smoothing* by bin boundaries: (select min or max value)**

- **Bin 1: 4, 4, 4, 15**
- **Bin 2: 21, 21, 25, 25**
- **Bin 3: 26, 26, 26, 34**

- **Sorted data for price (in dollars):**

4, 8, 9, 12, 15, 21, 21, 21, 24, 25, 26, 26, 28, 29, 34

Partition into (equi-depth) bins: (size = 5)

*** *Smoothing* by bin median:**

- **Bin 1: 9, 9, 9, 9, 9**
- **Bin 2: 21, 21, 21, 21, 21**
- **Bin 3: 28, 28, 28, 28, 28**

2.5.5.2 Regression Analysis

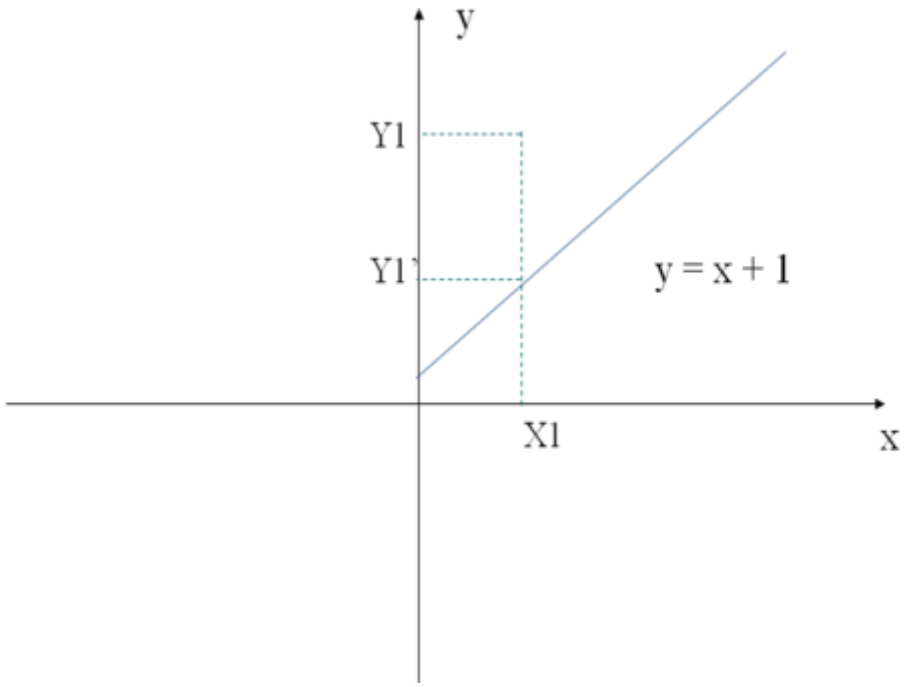
What is Regression?

- **Regression is the *measure of the average relationship between two or more variables in terms of the original units of data.***
- **Data can be filtered / smoothed by fitting the data to a function such as with a regression.**

- **What is regression analysis?**
- **Regression Analysis is a technique used for modeling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables.**
  - **Dependent variable is a single variable *being explained /predicted by the regression model (response variable)***
  - **Independent variable is the *explanatory variable(s) used to predict the dependent variable (predictor variable)***

Regression Analysis categories

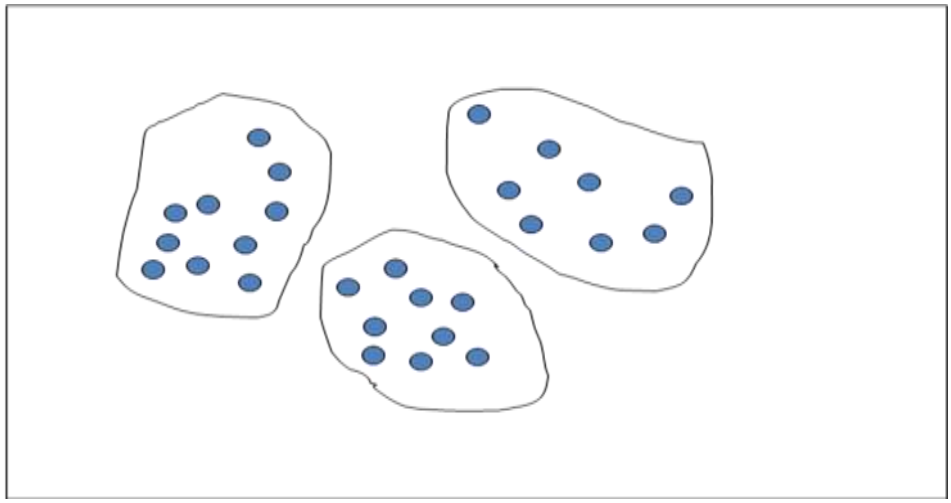
- **A) Linear regression: It involves finding the best line to fit two attributes, so that one can be used to predict the other.**
- **B) Multiple Linear Regression: It involves more number of attributes and the data are fit to a multidimemnsional surface.**



2.5.5.3 Cluster Analysis

- **The data element or objects that are entirely different from others or are inconsistent in comparison to other data elements are refer to as Outliers.**
- ***Outliers may be detected by Clustering.***
- **Goal: To determine the intrinsic grouping in a set of unlabeled data.**

Cluster Analysis



2.5.6 Data Cleaning as a Process

- **Data discrepancy detection**
  - **Use metadata (e.g., domain, range, dependency, distribution)**
  - **Check field overloading**
  - **Check uniqueness rule, consecutive rule and null rule**
  - **Use commercial tools**
    - ***Data scrubbing:*** **use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections**
    - ***Data auditing:*** **by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)**

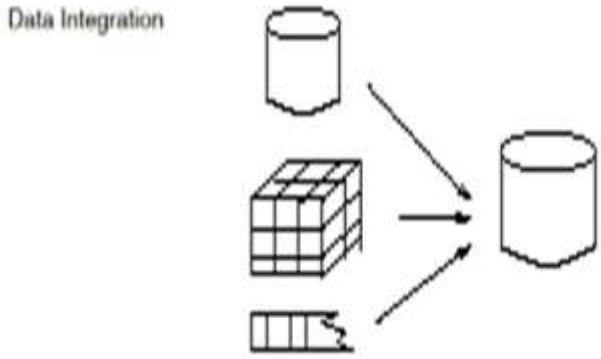
Thus the data cleaning process is a continuous data mining activity that makes the information ready to boost the ETL process.

****

2.6. DATA INTEGRATION AND TRANSFORMATION

2.6.1 Introduction

- **Data integration:**
  - Combines data from multiple sources into a coherent store
- **Schema integration**: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- **Entity identification problem:**
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton



- **Detecting and resolving data value conflicts**
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

2.6.2 Handling Redundancy in Data Integration

- ***Redundant data occur often when integration of multiple databases***
- ***Sources of Redundancy:***
  - ***Object identification*: The same attribute or object may have different names in different databases**
  - ***Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue**

- **Redundant attributes** may be able to be **detected by *correlation analysis***
- ***Careful integration*** of the data from multiple sources may help:
  - ***reduce/avoid redundancies and inconsistencies*** and
  - ***improve mining speed and quality.***

2.6.3 Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values
  - s.t. each old value can be identified with one of the new values

Data Transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Tranformation methods:

- ***Smoothing:*** remove noise from data
- ***Aggregation:*** summarization, data cube construction
- ***Generalization:*** concept hierarchy climbing

- ***Normalization:*** Scaled to fall within a small, specified range
  - min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

  - normalization by decimal scaling

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that Max(| |)<1}$$

- **Attribute/feature construction**
  - New attributes constructed from the given ones

The mining process – data transformation is achieved through selective mining techniques.

****

UNIT - II : Data Mining & Data Preprocessing

2.7 DATA REDUCTION

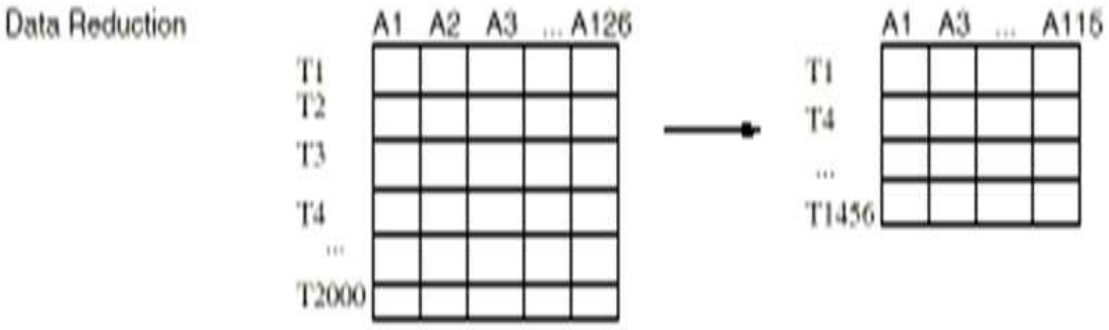
2.7.1 Introduction

- **Definition: Data reduction**
  - **DR is *a pre-processing technique which helps in reduced representation of the data set that is much smaller in volume* but yet produce the same (or almost the same) analytical results.**
- **Data reduction: Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results.**

Why Data Reduction?

- **Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set.**
- **A database/data warehouse may store terabytes of data.**
- ***Advantages:***
  - **Even after DR, integrity of original data is still maintained.**

- **Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set.**



- **Reducing the number of attributes**
  - ***Data cube aggregation:* applying roll-up, slice or dice operations.**
  - ***Removing irrelevant attributes:* attribute selection (filtering and wrapper methods), searching the attribute space (see Lecture 5: Attribute-oriented analysis).**
  - ***Principle component analysis (PCA) (numeric attributes only):* searching for a lower dimensional space that can best represent the data..**

- ***Reducing the number of attribute values***
  - **Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).**
  - **Clustering: grouping values in clusters.**
  - **Aggregation or generalization**
- ***Reducing the number of tuples***
  - **Sampling**

2.7.2 Data Reduction Strategies

- **Data reduction strategies**
  - **Data cube aggregation:**
  - **Dimensionality reduction — e.g., remove unimportant attributes**
  - **Data Compression**
  - **Numerosity reduction — e.g., fit data into models**
  - **Discretization and concept hierarchy generation**

Data Cube Aggregation

- **The lowest level of a data cube**
  - **the aggregated data for an individual entity of interest**
  - **e.g., a customer in a phone calling data warehouse.**
- **Multiple levels of aggregation in data cubes**
  - **Further reduce the size of data to deal with**
- **Reference appropriate levels**
  - **Use the smallest representation which is enough to solve the task**
- **Queries regarding aggregated information should be answered using data cube, when possible.**

Dimensionality Reduction

- ***Curse of dimensionality***
  - **When dimensionality increases, data becomes increasingly sparse**
  - **Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful**
  - **The possible combinations of subspaces will grow exponentially**

UNIT - II : Data Mining & Data Preprocessing

Dimensionality Reduction steps

- **Avoid the curse of dimensionality**
- **Help eliminate irrelevant features and reduce noise**
- **Reduce time and space required in data mining**
- **Allow easier visualization**

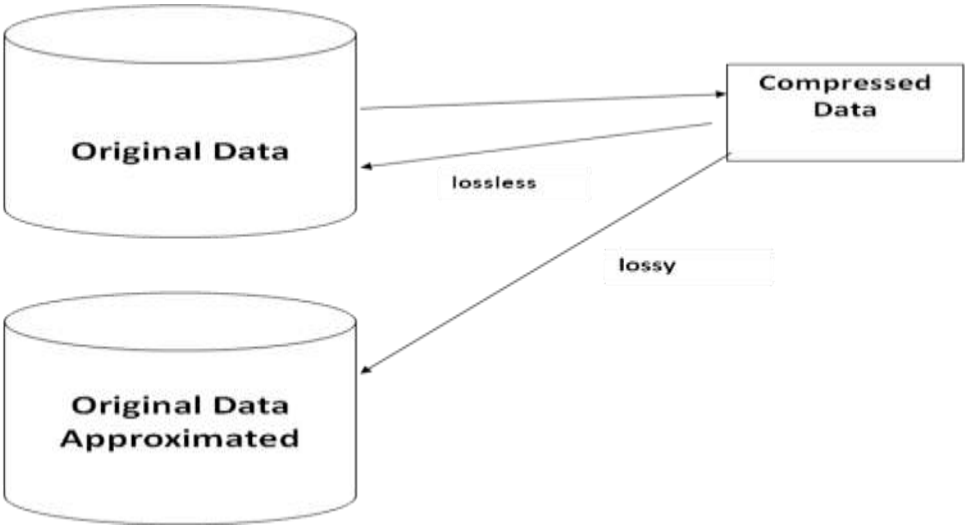
Feature selection (attribute subset selection):

- **Select a minimum set of features such that**
  - **the probability distribution of different classes given the values for those features is**
    - **as close as possible to the original distribution given the values of all features**
- **reduce # of patterns in the patterns, easier to understand**

- **Heuristic methods**
  - **step-wise forward selection**
  - **step-wise backward elimination**
  - **combining forward selection and backward elimination**
  - **decision-tree induction**

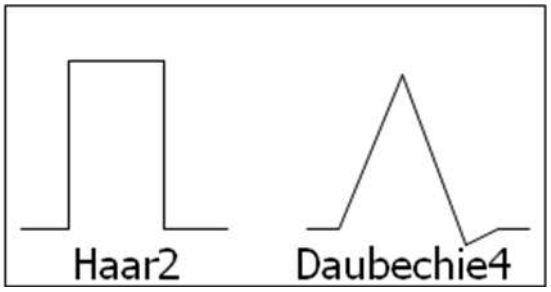
Data Compression

- **String compression**
  - **There are extensive theories and well-tuned algorithms**
  - **Typically lossless**
  - **But only limited manipulation is possible without expansion**
- **Audio/video compression**
  - **Typically lossycompression, with progressive refinement**
  - **Sometimes small fragments of signal can be reconstructed withoutreconstructing the whole**
- **Time sequence is not audio**
  - **Typically short and vary slowly with time**

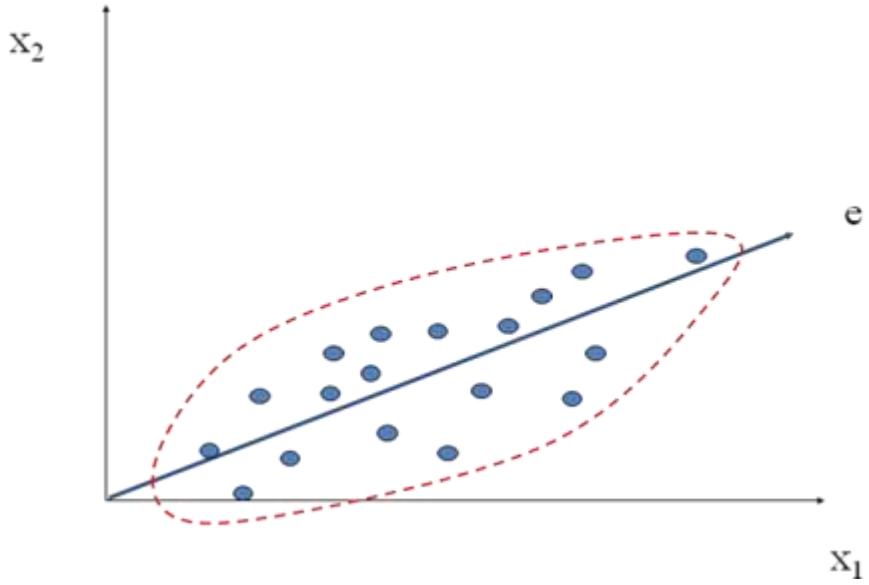


2.6.3.1 Wavelet Transforms

- **Discrete wavelet transform (DWT): linear signal processing**
- **Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients**
  - **Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space**



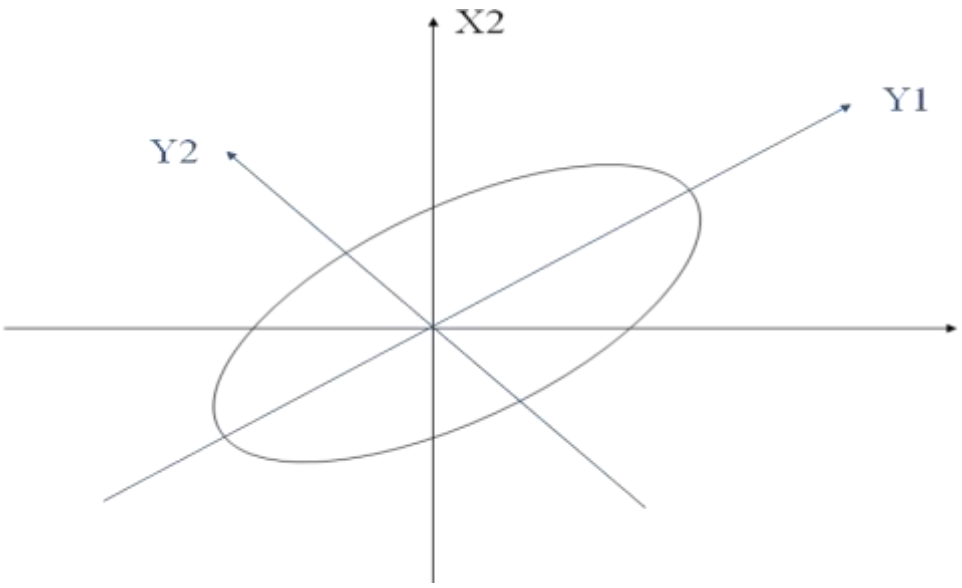
- **Method:**
  - **Length, L, must be an integer power of 2 (padding with 0s, when necessary)**
  - **Each transform has 2 functions: smoothing, difference**
  - **Applies to pairs of data, resulting in two set of data of length L/2**
  - **Applies two functions recursively, until reaches the desired length**



2.6.3.2 Principal Component Analysis (PCA)

- **Given *N* data vectors from *k*-dimensions,**
- **find *c* <= *k* orthogonal vectors that can be best used to represent data**
  - **The original data set is reduced to one consisting of N data vectors on c principal components (reduced dimensions)**
- **Each data vector is a linear combination of the *c* principal component vectors**
- **Works for numeric data only**
- **Used when the number of dimensions is large**

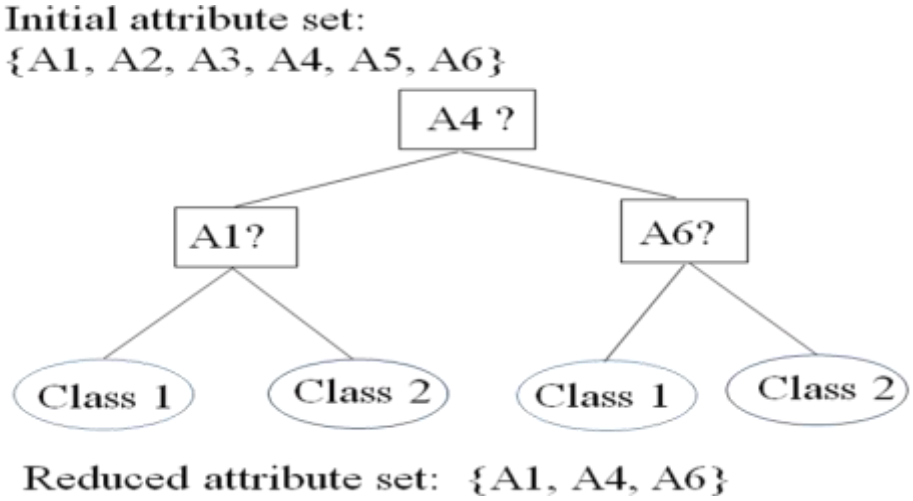
Principal Component Analysis



Heuristic Selection Methods

- **There are 2^d possible sub-features of *d* features**
- **Several heuristic selection methods:**
  - **Stepwise forward selection**
  - **Stepwise backward elimination**
  - **Combination of forward selection and backward elimination**
  - **Decision tree induction**

Example of Decision Tree Induction



UNIT - II : Data Mining & Data Preprocessing

Numerosity Reduction (Multiplicity Reduction)

- **Reduce data volume by**
  - **choosing alternative,**
  - **smaller forms of data representation**

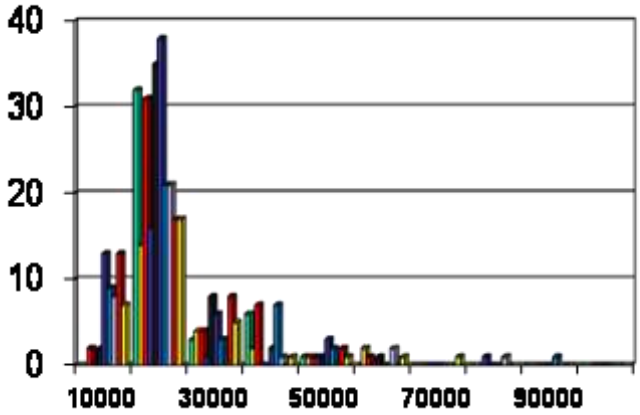
- **Parametric methods**
  - **Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)**
  - **Log-linear models: obtain value at a point in m-D space as the product on appropriate marginal subspaces**
- **Non-parametric methods**
  - **Do not assume models**
  - **Major families: histograms, clustering, sampling**

Data Reduction Method
Regression and Log-Linear Models

- ***Linear regression*: Data are modeled to fit a straight line**
  - **Often uses the least-square method to fit the line**
- **Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector**
- **Log-linear model: approximates discrete multidimensional probability distributions**

Histograms

- **Divide data into buckets and store average (sum) for each bucket**
- **Partitioning rules:**
  - **Equal-width: equal bucket range**
  - **Equal-frequency (or equal-depth)**
  - **V-optimal: with the least *histogram variance* (weighted sum of the original values that each bucket represents)**
  - **MaxDiff: set bucket boundary between each pair for pairs have the β–1 largest differences**



Clustering

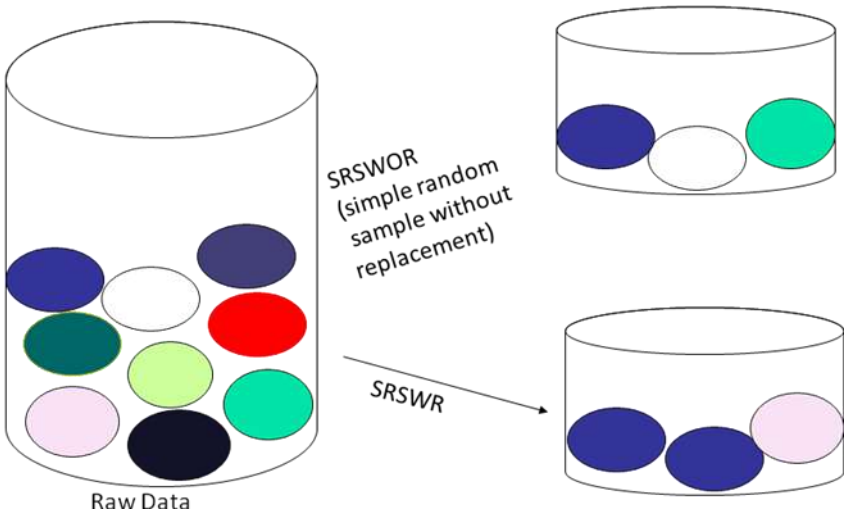
- **Partition data set into clusters based on similarity, and**
- **store cluster representation (e.g., centroid and diameter) only**
  - **can be very effective if data is clustered but not if data is "smeared"**
  - **Can have hierarchical clustering and be stored in multi-dimensional index tree structures.**

Sampling

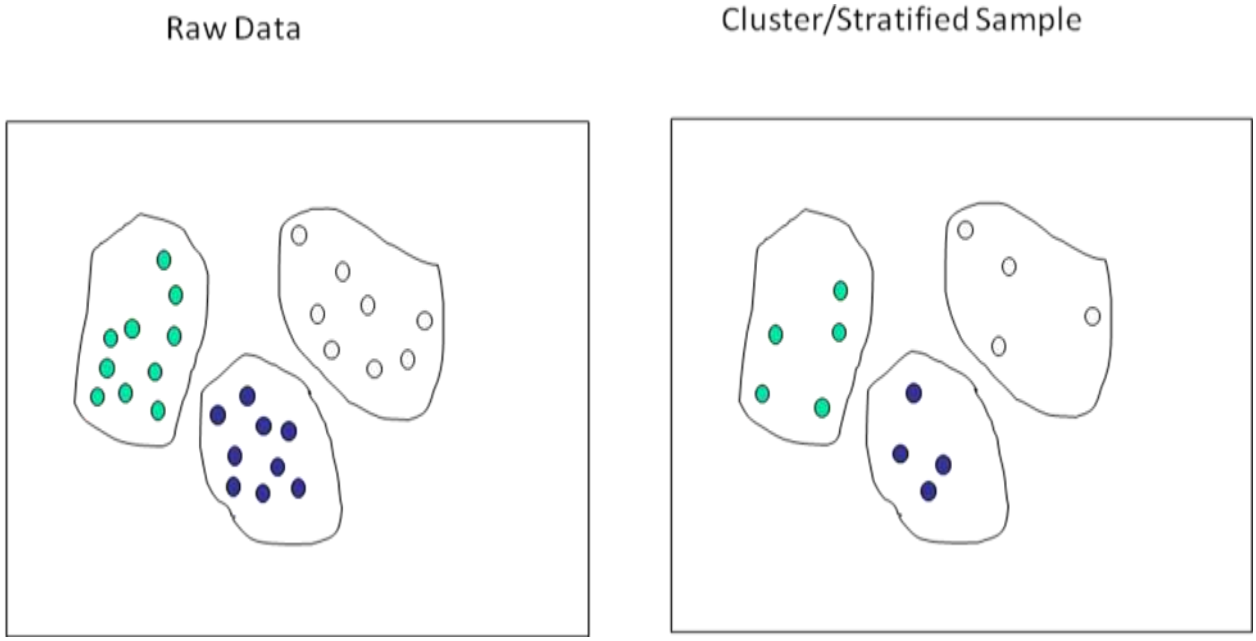
- **Sampling: obtaining a small sample *s* to represent the whole data set *N***
- **Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data**
- **Choose a representative subset of the data**
  - **Simple random sampling may have very poor performance in the presence of skew**

- **Develop adaptive sampling methods**
  - **Stratified sampling:**
    - **Approximate the percentage of each class (or subpopulation of interest) in the overall database**
    - **Used in conjunction with skewed data**
- **Note: Sampling may not reduce database I/Os (page at a time)**

Sampling: with or without Replacement



Sampling: Cluster or Stratified Sampling



Thus the data reduction is achieved through the adoptive techiques avaiable for the dataware housing architect.

****

UNIT - II : Data Mining & Data Preprocessing

2.8 DATA DISCRETIZATION AND CONCEPT HIERARCHY GENERATION

2.8.1 Introduction

Discretization (dividing by groups)

- **What is Discretization?**
  - **Dividing the range of values by possible groups.**
  - In general, ***attributes are to be ordered based on the use.***

- **How to divide?**
  - **Attribute Types!!!**
  - **Three types of attributes:**
  - ***Nominal*** — values from an unordered set
  - ***Ordinal*** — values from an ordered set
  - ***Continuous*** — real numbers
- **Purpose & Use of Discretization:**
- **Purpose: divide the range of a continuous attribute into intervals**
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

Data Discretization

- Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals.
- Thus ,

  DD = range of the attribute / intervals
- Interval labels can then be used to replace actual data values.
- **Objective of D-D:**
  - Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data.
- This leads to a concise, easy-to-use, knowledge-level representation of mining results.

2.8.2 Data Discretization Types

- Discretization techniques can be categorized based on whether it uses class information, as and also based on Split or merge options:

- **Supervised discretization**
  - the discretization process uses class information
- **Unsupervised discretization**
  - the discretization process does not use class information

- **Top-down**
  - If the process starts by first finding one or a few points (called split points or cut points) to split the entire attribute range, and then
  - repeats this recursively on the resulting intervals
- **Bottom-up**
  - starts by considering all of the continuous values as potential split points,
  - removes some by merging neighborhood values to form intervals,
  - and then recursively applies this process to the resulting intervals.

2.8.3 Discretization and Concept hierarchy

- **Discretization**
  - reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.
- **Concept hierarchies**
  - ***reduce the data by collecting and replacing low level concepts*** (such as numeric values for the attribute age) by ***higher level concepts*** (such as young, middle-aged, or senior).

- **Typical methods:**
  - Binning
- **Top-down split, unsupervised,**
  - Clustering analysis (covered above)
  - Either top-down split or bottom-up merge, unsupervised
  - Interval merging by c2 Analysis
- **unsupervised, bottom-up merge**
- All the methods can be applied recursively

Generalization

- ***Generalization is the generation of concept hierarchies***

for categorical data

- Categorical attributes have a finite (but possibly large) number of distinct values, with no ordering among the values.
- **Data Transformation**
- Examples include
- – geographic location,
- – job category, and
- – itemtype.

generation for numeric data

- Binning
- Histogram analysis
- Clustering analysis
- Entropy-based discretization
- Discretization by intuitive partitioning

Entropy-Based Discretization

- Given a set of samples S, if S is partitioned into two intervals S1 and S2 using boundary T, the entropy after partitioning is

$$E(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization.
- The process is recursively applied to partitions obtained until some stopping criterion is met, e.g.,

$$Ent(S) - E(T,S) > \delta$$

- Experiments show that it may reduce data size and improve classification accuracy.

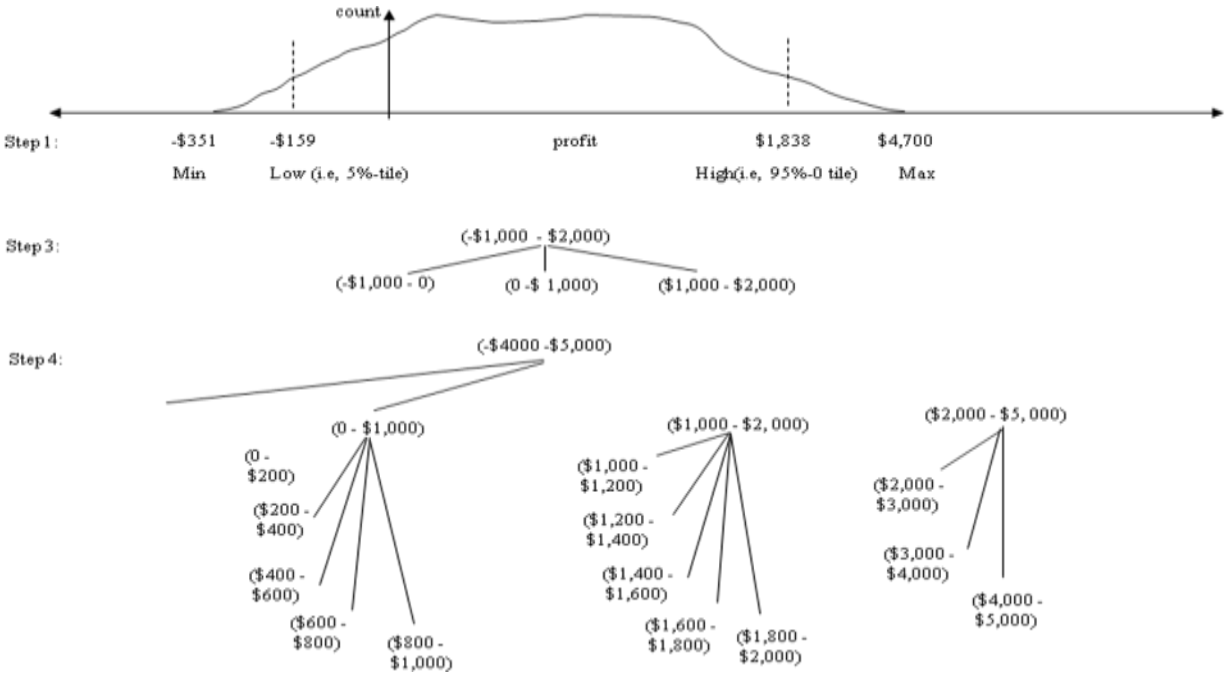
Discretization by intuitive partitioning

- 3-4-5 rule can be used to segment numeric data into relatively uniform, "natural" intervals.

* If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equal-width intervals

* If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
* If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals

Example of 3-4-5 rule

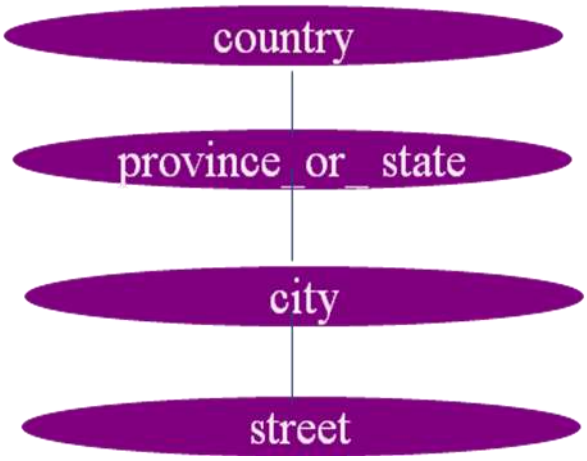


Concept hierarchy generation for categorical data

- Specification of a **partial ordering of attributes explicitly at the schema level by users or experts**
- Specification of a portion of a **hierarchy by explicit data grouping.**
- Specification of a **set of attributes, but not of their partial ordering.**
- Specification of **only a partial set of attributes.**

Specification of a set of attributes

- Concept hierarchy can be automatically generated based on the number of distinct values per attribute in the given attribute set.
- The attribute with the most distinct values is placed at the lowest level of the hierarchy.



Discretization by Classification & Correlation Analysis

- **Classification (e.g., decision tree analysis)**
  - **Supervised:** Given class labels, e.g., cancerous vs. benign
  - **Using *entropy*** to determine split point (discretization point)
  - **Top-down**, recursive split
- **Correlation analysis (e.g., Chi-merge: χ^2-based discretization)**
  - **Supervised:** use class information
  - **Bottom-up merge**: find the best neighboring intervals (those having similar distributions of classes, i.e., low χ^2 values) to merge
  - Merge performed recursively, until a predefined stopping condition

Concept Hierarchy Generation

- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is **usually associated with each dimension in a data warehouse**
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity
- **Concept hierarchy formation:** Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth*, *adult*, or *senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed **for both numeric and nominal data.** For numeric data, use discretization methods shown.

Concept Hierarchy Generation for Nominal Data

- **Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts**
  - *street* < *city* < *state* < *country*
- **Specification of a hierarchy for a set of values by explicit data grouping**
  - {Urbana, Champaign, Chicago} < Illinois
- **Specification of only a partial set of attributes**
  - E.g., only *street* < *city*, not others
- **Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values**
  - E.g., for a set of attributes: {*street, city, state, country*}

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):
  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N} \qquad \bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Median: A holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

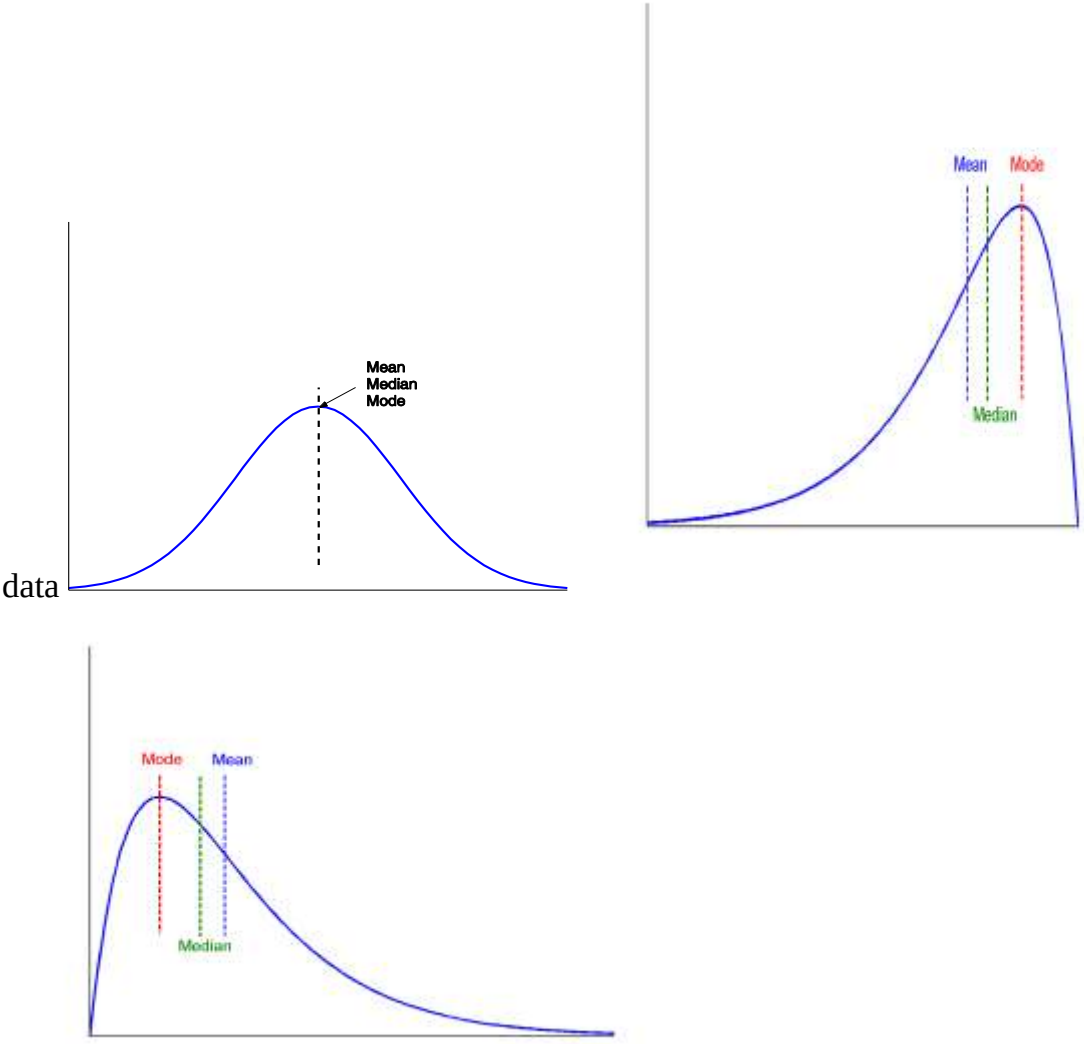
$$median = L_1 + (\frac{n/2 - (\sum f)l}{f_{median}})c$$

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

$$mean - mode = 3 \times (mean - median)$$

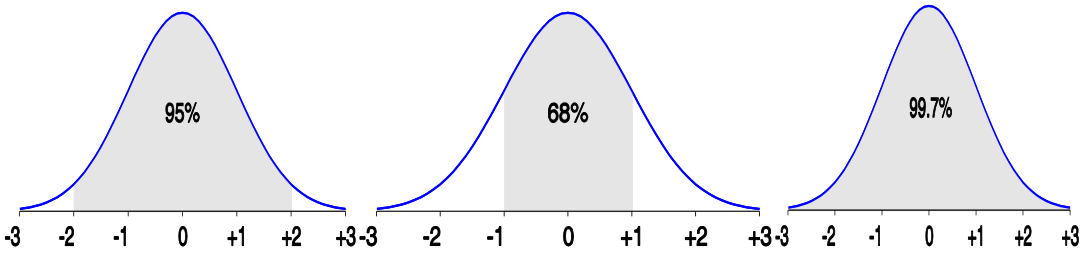
Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed



Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it



Thus the Data discritization and Concept hierarchy are the additional activities that support the Data mining preprocessing effectively.

****