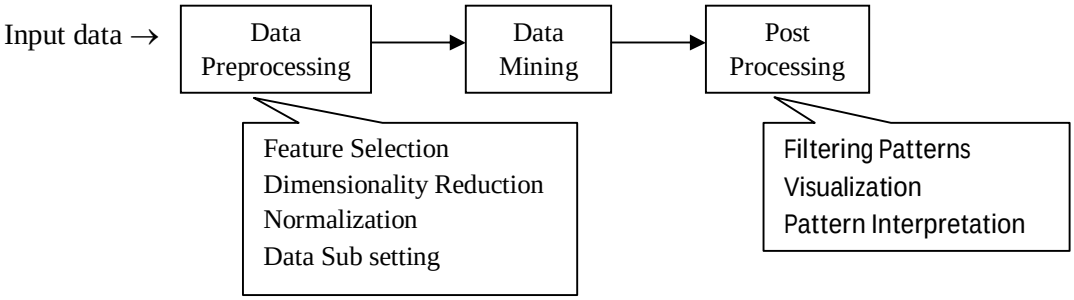
What is is Data Mining?

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown.

Data Mining and Knowledge Discovery:



Data mining is an integral part of knowledge discovery in databases (KDD), which is the overall process of converting row data into useful information, as shown in above Figure. This process consists of a series of transformation steps, from data preprocessing to post processing of data mining results.

The input data can be can be stored in variety of formats (flat files, spreadsheets or relational tables) and may reside in a centralized data repository or to be distributed across multiple sites. The purpose of preprocessing is to transform the raw input data into an appropriate format for subsequent analysis.

The steps involved in data preprocessing include fusing data from multiple sources, cleaning data to remove noise and duplicate observations, and selecting records and features that are relevant to the data-mining task at hand. Because of the many ways, data can be collected and stored, data preprocessing step in the overall knowledge discovery process.

"Closing the loop" is the phrase often used to refer to the process of integrating data mining results into decision support systems. For example, in business applications, the insights offered by data mining results can be integrated with campaign management tools so that effective marketing promotions can be conducted and tested. Such integration requires a post-processing step that ensures that only valid and useful results are incorporated into the decision support system. An example of post processing is visualization, which allows analysis to explore the data and the data mining results from a variety of viewpoints. Statistical measures or hypothesis testing methods can also be applied during post processing to eliminate spurious data mining results.

MOTIVATING CHALLENGES:

As mentioned earlier, traditional data, analysis techniques have often encountered practical difficulties in meeting the challenges posed by new data sets. The following are some of the specific challenges that motivated the development of data mining.

1) Scalability (Meaning from **Wikipedia**)

"Scalability is the ability of a system, network, or process to handle a growing amount of work in capable manner or its ability to be enlarged to accommodate that growth".

Scalability: Because of advances in data generation and collection, data sets with sizes of gigabytes, terabytes, or even pet bytes are becoming common. If data mining algorithms are to be scalable. Many data mining algorithms employ special search strategies to handle exponential search problems. Scalability may also require the implementation of novel data structures to access individual records in an efficient manner.

For instance, out-of-core algorithms may be necessary when processing data sets that cannot fit into main memory. Scalability can also be improved by using sampling or developing parallel and distributed algorithms.

2) High Dimensionality: It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago. In bioinformatics, progress in microarray technology has produced gene expression data involving thousands of features. Data sets with temporal or spatial components also tend to have high dimensionality.

3) Heterogeneous and complex data: Traditional data analysis often deals with data sets containing attributes of the same type, either continuous or categorical. As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes. There are some non-traditional types of data as web pages containing semi-structured text and hyperlinks, DNA data and climate data i.e., temperature, pressure, moisture etc,

4) Data Ownership and Distribution: Sometimes the data needed for analysis may not store in one location. The data is distributed among resources belonging to multiple entities. This requires the development of distributed data mining techniques which includes some challenges

a) How to reduce the amount of communication needed to perform the distributed computation.

b) How to consolidate the data mining results

c) How to address data security issues.

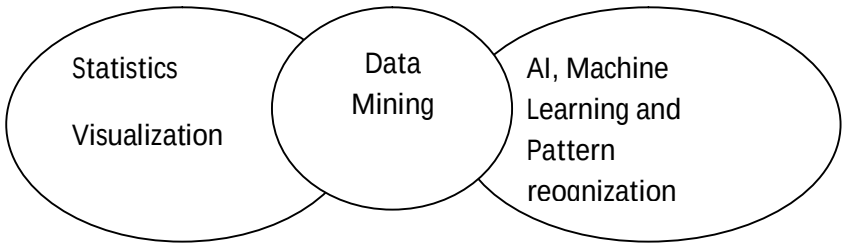
5) Non-traditional Analysis: The traditional statistical approach is based on hypothesize and test paradigm. A hypothesis is proposed, an experiment is designed to gather the data and then the data is analyzed with respect to the hypothesis. But this process is extremely labor intensive. Current data analysis tasks require the generation and evaluation of thousands of hypothesis and the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation.

The Origins of Data Mining:

⇨ Data mining is an interdisciplinary field, the confluence of a set of disciplines including database technology, machine learning, visualization, statistics and information science.

⇨ There are some other disciplines may applied such as neural network, fuzzy or rough set theory, knowledge representation include logic programming and high-performance computing.

⇨ Data mining draws upon ideas such as sampling, estimation and hypothesis testing from statistics.



Data base Technology, Parallel & Distributed Computing

⇨ Depending on the kinds of data to be mined or on the given data mining application, data mining system also integrate techniques from optical data analysis, information retrieval, image analysis, signal processing, computer graphics, web technology, economics, business, bioinformatics or psychology.

⇨ Data mining has also adopted ideas from other areas including optimization, evolutionary computing, information theory, signal processing and information retrieval.

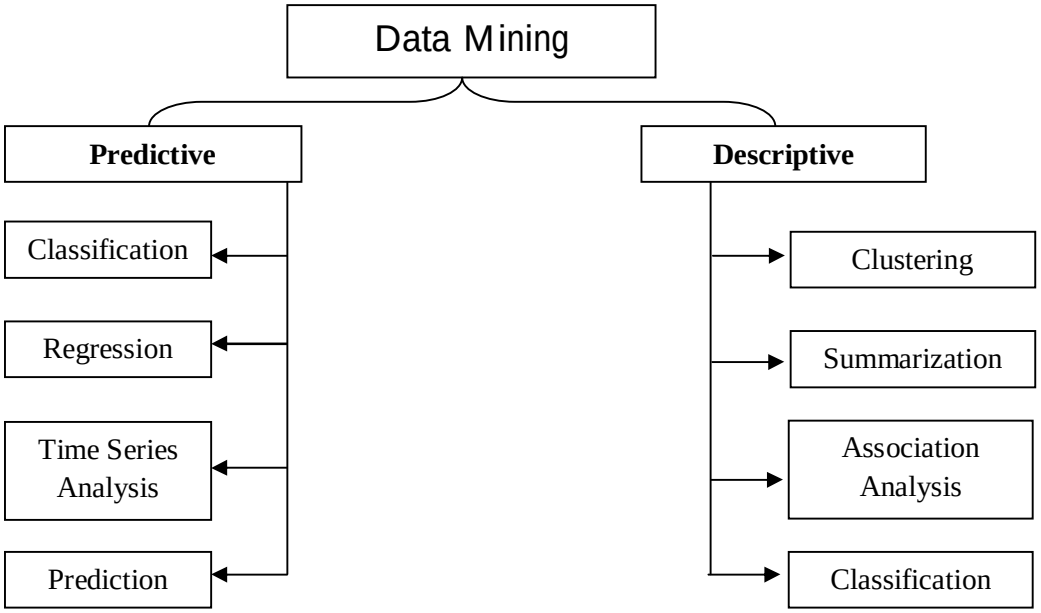
Data Mining Tasks:

⇨ Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.

⇨ Data mining tasks are divided into two major categories

i) Predictive tasks: These tasks used to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is known as target or dependent variable and the attribute used for making the prediction is called explanatory or independent variable.

ii) Descriptive tasks: These tasks characterize the general properties of the data in the database. These tasks are exploratory in nature and frequently require preprocessing techniques to validate and explain the results.



i) Predictive Tasks:

a) Classification: Classification I used for discrete target variables and maps data into predefined groups or classes. All the classes are determined before examining the data.

Eg: Determining whether to make a bank loan and identifying credit risks.

b) Regression: Regression is used for continuous target variables. It is used to map a data item to a real-valued prediction variable. Regression involves the learning of the function that does this mapping.

c) Time-series Analysis: with time series analysis the value of an attribute is examined as it varies over time. The values usually are obtained as evenly speed time points i.e., daily, weekly, hourly.

Eg: Stock market data analysis.

d) Prediction: Prediction can be viewed as a type of classification. The difference is that prediction is predicting a future state rather than a current state. Prediction applications include flooding, speech recognition, machine learning and pattern recognition.

ii) Descriptive Tasks:

a) Clustering: Clustering is similar to classification, except that the groups are not pre-defined, but rather defined by the data alone. Clustering also called as segmentation. The most similar data are grouped into clusters. Clustering is used to identify outliers or anomalies.

b) Summarization: Summarization maps data into subsets with associated simple descriptions. It is also called as characterization. Data characterization is summarizing the data of the class i.e., target class

c) Association analysis: It is used to discover associated or uncovering relationships among the data. The discovered patterns are represented in the form of implication rules or association rules. The goal of association analysis is to extract the most interesting patterns in the data.
Eg: marker basket analysis

d) Sequence Discovery: Sequence analysis or discovery is used to determine sequential patterns in data. These patterns are based on a time sequence of actions. These patterns are similar to associations but the relationship is based on time.
Eg: the person who purchases CD players may be found to purchase CDs with in a particular time.

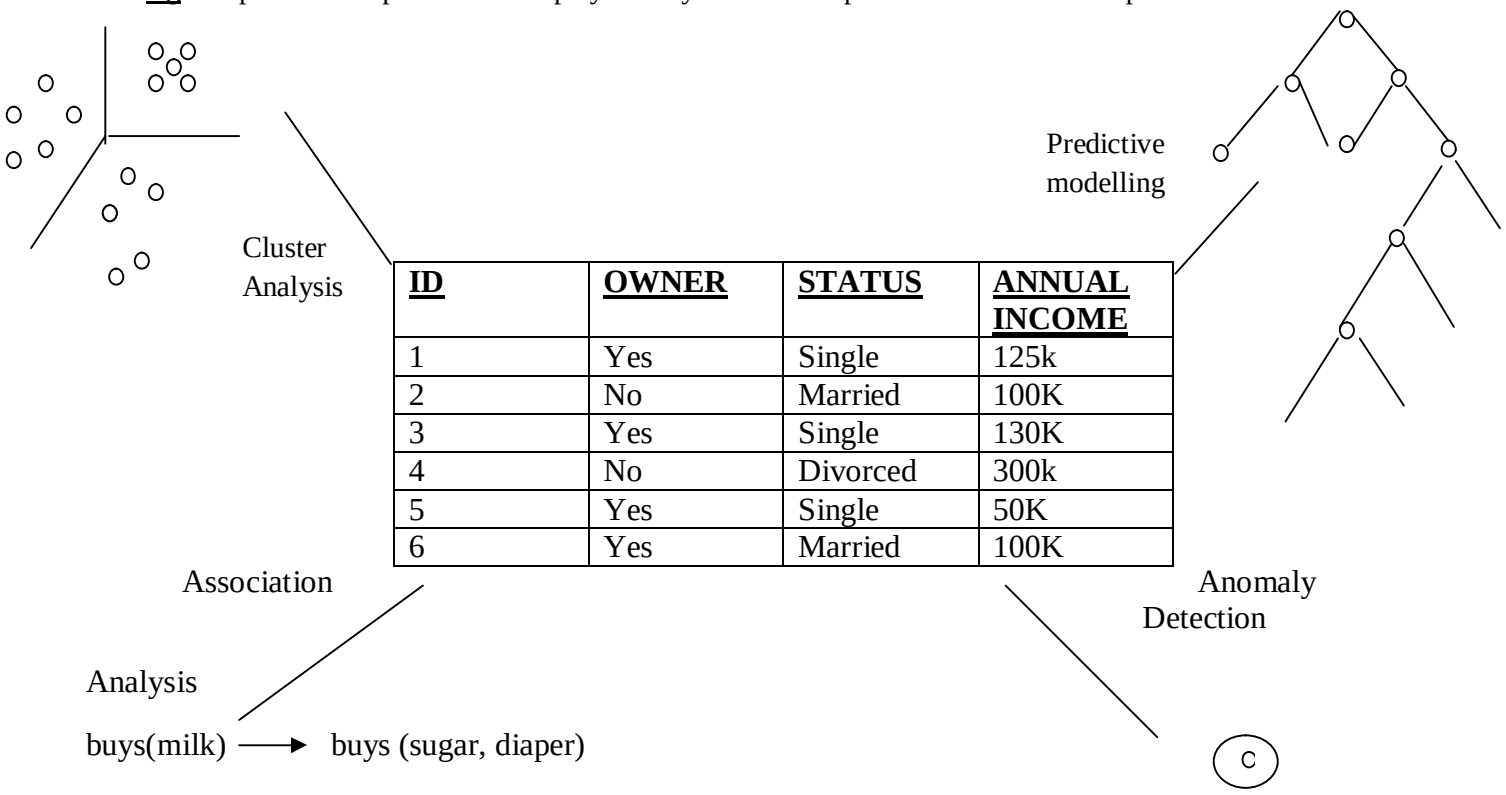


fig: four of the core data mining tasks

Types of Data:

⇨ A data set can often be viewed as a collection of data objects.

⇨ A data object can also call as record, point, vector, pattern, event, case, sample, observation or entity.

⇨ The data objects are described by a number of attributes that capture the basic characteristics of an object.

⇨ An attribute is also called as variable, characteristic, field, feature or dimension.

Eg: student information: sometimes a data set is a file in which the objects are records/rows in the file and each field/column corresponds to an attribute.

Std ID	Year	Grade Point Average (GPA)
1034262	Senior	3.24
1052638	Freshman	3.62
1082246	Sophomore	3.51

Table: A sample data set containing std information

Attributes and Measurements:

Attribute: An attribute is a property or characteristic of an object that may vary, either from one object to another or from one time to another.

Eg: Eye colour varies from person to person, while the temperature of an object varies over time.

Measurement: A measurement scale is a rule/function that associates a numerical or symbolic value with an attribute of an object. The purpose of measurement is the application of a measurement scale to associate a value with a particular attribute of a specific object.

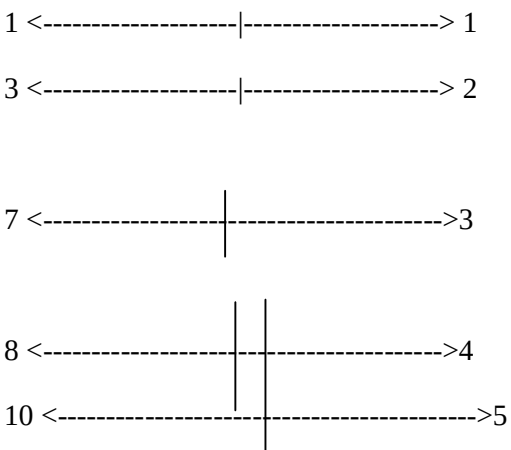
The type of an Attribute: The values used to represent an attribute may have properties that are not properties of the attribute itself and vice versa. This is illustrated with two examples.

Eg: **1)** Employee Age and IO Number. Two attributes that might be associated with an employee are ID and age. Both of these attributes can be represented as integers. We can take an average of age but not the average employee ID.

For the age of attribute, the properties of the integers used to represent age are very much the properties of the attribute. Even so, the correspondence is not complete since ages have a maximum while integers do not.

2) Length of dine segments: Consider some line segments, which mapped to numbers in two different ways.

Each successive line segment, going from the top to the bottom is formed by appending the topmost line segment to itself. Thus the second line segment from the top is formed by appending the topmost line segment to itself twice and so on.



A mapping of lengths to numbers that captures only the order properties of length

A mapping of length to numbers that captures both the order and additivity properties of length

⇨ All the line segments are multiplies of the first. This fact is captured by the measurements on the right-hand side of the figure, but not by those on the left-hand side.

⇨ The measurements scale on the left-hand side captures only the ordering of the length attribute while the scale on the right-hand side captures both the ordering and additivity properties. Thus an attribute can be measured in a way that does not capture all the properties of the attribute.

The Different Types of Attributes:

A specific way to specify the type of an attribute is to identify the properties of numbers that corresponds to underlying properties of the attribute.

Eg: An attribute length has many of the properties of numbers.

The following properties of numbers are used to describe attributes

1. Distinctness = and ≠
2. Order <, ≤, > and ≥
3. Addition + and –
4. Multiplication * and /

⇨ By these properties we can define four types of attributes: nominal, ordinal, interval and ratio.

⇨ Nominal and Ordinal attributes are referred to as categorical or qualitative attributes.
Eg: Employee ID

⇨ Interval and Ratio attributes are referred to as quantitative or numeric attributes.
Eg: Integer valued and continuous.

(i) **Nominal Attribute:** The value of a nominal attribute are just different names i.e., they provide information to distinguish one object from another (=, ≠)
Eg: zip codes, Emp.ID, eye color, gender.

(ii) **Ordinal Attribute**: The values of an ordinal attribute provide enough information to order objects. (<, >)
Eg: hardness of minerals, {good, better, best}, grades, street numbers.

Operations: median, percentiles, rank correlation, run tests, sign tests.

(iii) **Interval Attribute:** For interval attributes, the differences between values are meaningful i.e., a unit of measurement exists.
Eg: calendar dates, temperature in Celsius or Fahrenheit.

Operations: Mean, standard deviation, Pearson's correlation, t and F tests.

(iv) **Ratio attributes**: For ratio variables, both differences and ratios are meaningful.
Eg: temperature in Kelvin, counts, age, mass, length, electric current.

Operations: geometric mean, harmonic mean, percent variation.

⇨ The types of attributes can also be described in terms of transformations that do not change the meaning of attribute.

Attribute Type	Transformation	Comment
Nominal	Any one-to-one mapping Eg: a permutation of values	If all emp ID nums are reassigned, it will not make any difference.
Ordinal	An order-preserving change of values i.e. new-value=f(old-value) Where f=monotonic function	An attribute encompassing the notion of good, better, best can be represented equally well by values {1, 2, 3} or by {0.5, 1, 10}
Interval	new-value=a*old-val+b a, b are constants	The Fahrenheit and Celsius temp scales differ in the location of their zero value and size of a degree.
Ratio	new-value=a*old-value	Length can be measured in meters or feet

⇨ The statistical operations that make sense for a particular type of attribute must yield the same results when the attribute is transformed using a transformation that preserves the attribute's meaning.

Describing Attributes by the Numbers of Values:

An independent way of distinguishing between attributes is by the number of values they can take.

a) **Discrete attributes:** A discrete attribute has a finite or infinite set of values. Such attributes are categorical, such as zip codes or ID numbers or numeric such as counts. Binary attributes are a special case of discrete attributes and having only two values i.e. true/false, yes/no, male/female, 0/1

b) **Continuous Attributes**: A continuous attribute is one whose values are real numbers. These attributes are temperature height or weight. Continuous attributes are represented as floating-point variables.

⇨ Nominal and Ordinal attributes are binary or discrete while interval and ratio attributes are continuous.

⇨ Countable attributes are discrete and are ratio attributes.

c) **Asymmetric Attributes**: Binary attributes where only non-zero values are important are called asymmetric binary attributes. These types of attributes are important for association analysis.

Types of Data Sets:

The data sets are grouped into three groups: Record data, graph-based data, ordered data.

General characterization of Data Sets:

Three characteristics apply to data sets and have a significant impact on the data mining techniques used.

(i) **Dimensionality:** The dimensionality of a data set is the number of attributes that the objects in the data in the data set possess. Data with a small number of dimensions tends to be qualitatively different than high dimensional data. The difficulties associated with analysing high-dimensional data are referred to as curse of dimensionality.

(ii) **Sparsity**: For some data sets, with asymmetric features. Most attributes of an object have values of 0, in many cases less than 1% of the entries are non-zero. Sparsity allows only non-zero values need to be stored and manipulated. This saves computation time and storage.

(iii) **Resolution:** The properties of the data are different at different resolutions. For e.g., the surface of the earth seems very uneven at a resolution of a few meters, but it is relatively smooth at a resolution

of tens of kilometers. The patterns in the data depend on the level of resolution. If the resolution is too fine, a pattern may not be visible or may be buried in the noise.

a) Record Data:

⇨ In this, the data set is a collection of records/data objects each of which consists of a fixed set of data fields/attributes.

⇨ Every record has the same set of attributes.

⇨ Record data is usually stored either in flat files or in relational databases. There are different types of record data as

Tid	Refund	Marital Status	Taxable Income	Defaulted Browser
1	Yes	Single	125k	No
2	No	Married	100k	No
3	No	Single	70k	No
4	Yes	Married	120k	No
5	No	Divorced	95k	yes
6	No	Married	60k	No
7	yes	Married	220k	No

⇨ **Transaction/market Basket Data**: Transaction data is a special type of record data, where each record involves a set of items. In a grocery store, the set of products purchased by a customer data is called market basket data. Transaction data is a collection of set of items.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Milk, Sugar
4	Soda, Diaper, Milk
5	Beer, Soda, Diaper, Milk

⇨ **The Data matrix**: If the data objects have the same fixed set of numeric attributes, then the data objects can be specified as points in a multidimensional space, where each dimension represents a distinct attribute describing the object. A set of such data objects can be interpreted as an m by n matrix where there are m rows, one for each object, and m columns one for each attribute. This matrix is called a data matrix or a pattern matrix. A data matrix is a variation of record data, but because it consists of numeric attributes, standard matrix operation can be applied to transform and manipulate the data.

Production of X Load	Production of Y Load	Distance	Load
10.23	5.27	15.22	1.2
12.65	6.25	16.22	22
13.54	7.23	17.34	23
14.27	8.43	18.45	25

	Team	coach	play	ball	score	game	coin
Document1	3	0	5	0	2	6	0
Document2	0	7	0	2	1	0	0
Document3	0	1	0	0	1	2	2

Fig: Document-term matrix

⇨ **The Sparse Data Matrix**: A sparse data matrix is a special case of a data matrix in which the attributes are of the same type and are asymmetric i.e., only non-zero values are important. Transaction data is an example of a sparse data matrix that has only 0-1 entries. Another common example is document data.

If the order of the terms in a document is ignored, then a document can be represented as a terms vector where each term is a component/attribute of a vector and the value of each component is the number of times the corresponding term occurs in a document. This representation of a collection of documents is often called a document-term matrix. In the above figure the documents are the rows of this matrix, while the terms are the columns.

b) Graph-Based Data:

⇨ A graph can be a convenient and powerful representation for data. Consider two specific cases:

1. The graph captures relationships among data objects.
2. The data objects themselves are represented as graphs.

⇨ **Data with Relationships among Objects**: The relationships among objects frequently convey important information. In such cases the data represented as a graph. The data objects are mapped to nodes, while the relationships among objects are captured by the links between objects and link properties such as direction and weight.

Eg: web pages on the WWW, which contain both text and links to other pages.

⇨ **Data with objects that Are Graphs**: If objects have structure i.e. the objects contain sub objects that have relationships then such objects are frequently represented as graphs.

Eg: The structure of chemical compounds can be represented by a graph, where the nodes are atoms and the links between nodes are chemical bonds. The below **figure** shows **Linked web-pages**

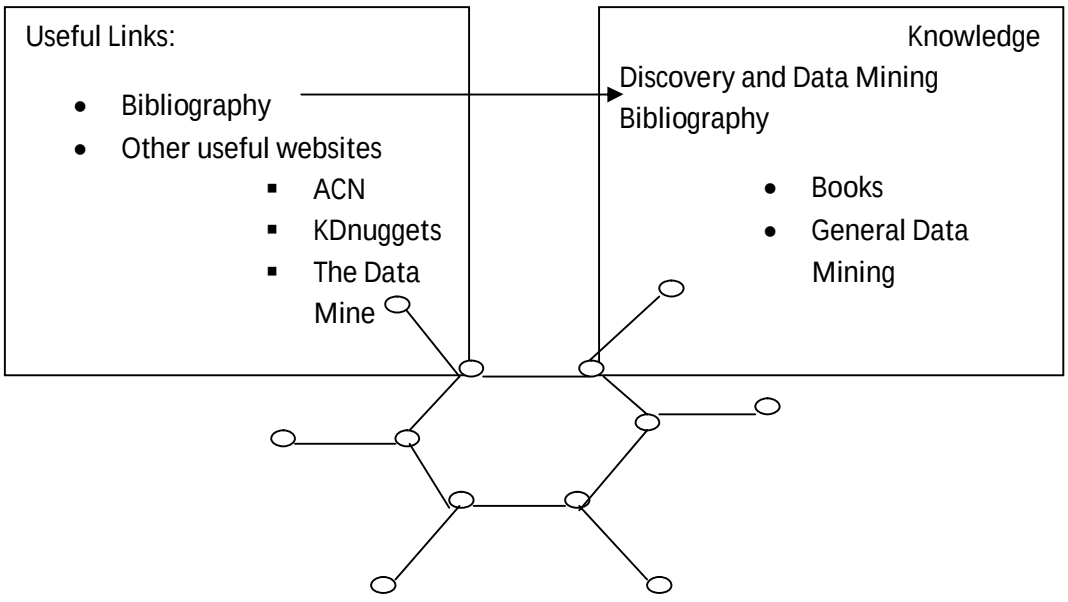


Fig: Benzene Molecule

⇨ Chemical compound benzene contains atoms of (black) carbon and hydrogen (gray). A graph representation makes it possible to determine which substructures occur frequently in a set of compounds and to ascertain whether the presence of any other substructures is associated with the

presence or absence of certain chemical properties. Substructure mining, which is a branch that analyzes such data.

c) <u>Ordered Data</u>:

⇨ The data in which the attributes have relationships that involve order in time or space.

⇨ There are different types of ordered data.

⇨ Sequential Data: It is also referred to as temporal data, can be thought of as an extension of record data, where each record has a time associated with it.

E.g.: people who buy DVD players tend to buy DVDs in the period immediately following the purchase.

Consider a sequential transaction data with five different times t_1, t_2, t_3, t_4 *and* t_5 three customers c_1, c_2 *and* c_3 and five items A, B, C, D and E. At time t_3, customer c_2 purchased items A and D. In the bottom table, each row corresponds to a particular customer. Each row contains information on each transaction involving the customer, where a transaction is considered to be a set of items and the time they purchased.

E.g. customer c_3 bought items A and C at time t_2

Time	customer	Items purchased
t_1	c_1	A,B
t_2	c_3	A,C
t_3	c_1	C,D
t_4	c_2	A,D
t_5	c_1	A,E

Customer	Time & items purchase
c_1	$\{t_1: A, B\}\{t_3: A, E\}$
c_2	$\{t_4: A, D\}$
c_3	$\{t_2: A, C\}$

Fig: sequential transaction data

⇨ Sequence Data: sequence data consists of a data set that is a sequence of individual entities, such as a sequence of words or letters. It is quite similar to sequential data; except that there are no time stamps, instead there are positions in an ordered sequence.

E.g.: The generic information of plants and animals i.e., genes.

⇨ Time series Data: Time series data is a special type of sequential data in which each record is a time series i.e. a series of measurements taken over time.

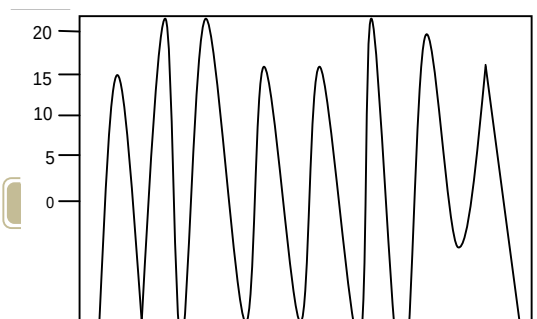
E.g.: Financial data set might contain objects that are time series of the daily pieces of various stocks.

⇨ It is also called as temporal data which consider temporal auto correlation i.e. if two measurements are close in time, then the values of those measurements are very similar.

GGTTCCGCTTAGCCCCGCC

GCCTACCTACTTTAAGCCCGC

GAGAAGACCTCCTAAGAAGC



CCAACCGAGTCCGACCAGGT

TGGGCTGCCTGCTCGACCACG

Fig: Geometric sequence data

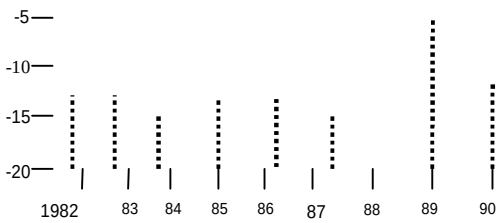


Fig: Temperature ....es Series

⇨ **Spatial Data**: some objects have spatial attributes such as positions or areas etc.

E.g. weather data-precipitation, temperature; pressure is collected for a variety of geographical locations.

⇨ An important aspect of spatial data is spatial auto-correlation i.e. objects that are physically close tend to be similar in other.

⇨ **Handling Non-Record Data**: Most data mining algorithms are designed for record data or its variations, such as transaction data and data matrices. Record-oriented techniques can be applied to non-record data by extracting features from data objects and using these features to create a record corresponding to each object.

E.g. consider the chemical structure data i.e. Benzene molecule given a set of common substructures, each compound can be represented as a record with binary attributes that indicate whether a compound data set, where the transactions are the components and the items are the substructures.

Data Quality:

⇨ To perform effective data mining we need to concentrate on data quality issues at source.

⇨ Data mining focuses on

1. Detection and correlation of data quality problems(Data cleaning)
2. Use of algorithms that can tolerate poor data quality.

Measurement and Data collection Issues:

- Databases are not static in nature due to human error, limitations of measuring devices or flaws in the data collection process. Due to these problems values or even entire data object may be missing.
- There may be some objects duplicated i.e. multiple data objects that all correspond to a single real object.
  E.g. A person lived at two different addresses.
- There may be some inconsistencies such as a person has a height of 2 meters but weights only 2 kilograms.
- The definition of measurement and data collection errors considers a variety of problems that involve measurement error: noise, artifacts, bias, precision and accuracy.
- So we need to concentrate data quality issues that may involve both measurement and data collection problems: outliers, missing and inconsistent values and duplicate values.

a) Measurement and Data collection Errors:

- A common measurement error is that the value recorded differs from the true value to some extent. For continuous attributes the numerical difference of the measured and true value is called the error.

- The data collection errors such as omitting data objects or attribute values or inappropriately including a data object.

b) Noise and Artifacts:

- Noise is the random component of a measurement error.
- It is a variance of the measured value.

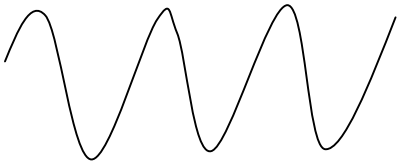


Fig: Noise in Times Series Context

- Noise is removed using data smoothing techniques and by devising robust algorithms that produce acceptable results even when noise is present.
- Artifacts are referred to as deterministic distortions of data. Such as a streak in the same place on a set of photographs.

c) Precision, Bias and Accuracy:

- In statistics the quality of the measurement process and the resulting data are measured by precision and bias.
- **Precision:** The closeness of repeated measurements to one another precision is measured by the standard deviation of a set of values.
- **Bias:** A systematic variation of measurements from the quantity being measured. Bias is measured by taking the difference between the mean of the set of values and the known value of the quantity.
  E.g. suppose the mass is weighted five times as {1.015, 0.990, 1.013, 1.001, 0.986}. The mean of these values is 1.001. The bias is 0.001. The precision is measured by the standard deviation is 0.013.
- **Accuracy**: The closeness of measurements to the true value of the quality being measured. Accuracy depends on precision and bias but there is no specific formula of accuracy. Accuracy is specified by significant digits.
  E.g.: If the length of an object is measured with a meter stick whose smallest markings are millimetres then we should only record the length of data to the nearest millimetre. The precision of such a measurement would be ±0.5mm.

d) Outliers: Outliers are either (1) data objects with characteristics that are different from most of the other data objects in a data set or (2) values of an attribute that are unusual with respect to the typical values for that attribute. In fraud objects or events from a large number of normal ones.

e) Missing values: Many attributes may not have the values recorded such as customer income. So we need to fill the missing values for the attribute by specific methods.

i) Eliminate Data Objects or Attributes: A simple and effective strategy is to eliminate objects with missing values. This is done when there is no information provided and is not suitable for efficient mining results.

ii) Estimate missing values: Sometimes missing data can be reliably estimated. Consider a time series data that changes in a time period but has widely scattered missing values. In such cases the missing values can be estimated by using the remaining values. Using a global constant to fill in the missing values. If the attribute is continuous use the attribute mean to fill in the missing values.

iii) Ignore the missing value during Analysis: Many data mining approaches can be modified to ignore missing values.

E.g.: suppose that objects are being clustered and the similarity between pairs of data objects needs to be calculated. If one or both objects of a pair have missing values for some attribute then the similarity can be calculated by using only the attributes that do not have missing values.

f) Inconsistent Values:

- Data may contain inconsistent values. Consider an address where both a zip code and city are listed, but the specified zip code area is not contained in that city. It may be entered this information transposed two digits or a digit was misread when the information was scanned from a handwritten from.
- It is important to detect and correct such problems.
- Some types of inconsistencies are easy to detect.
  E.g.: person's height should not be negative.
- In some cases it is necessary to consult an external source of information.
  E.g.: When an insurance company processes claims for reimburse against a database of its customer.

g) Duplicate Data:

- A data set may be including data objects that are duplicates of one another. Many people receive duplicate mailings because they appear in a database multiple times under sight different names.
- To detect and eliminate such duplicates two issues be addressed:
  1) If there are two objects that actually represent a single object then the values of corresponding attributes may differ and these inconsistent values must be resolved.
  2) Avoiding accidently combining data objects that are similar but not duplicates such as two distinct people with identical names.
- The deduplication is used to refer to the process of dealing with these issues.

Issues Related to Applications:

- Data quality issues can also be considered from an application viewpoint as expressed by the statement.
  "Data is of high quality if it is suitable for its intended use".
- Consider a few of the general issues:
  (i)Timeliness: Some data starts to age as soon as it has been collected. If the data is out of date, then so are the models and patterns that are based on it.
  (ii)Relevance: The available data must contain the information necessary for the application. Consider the test of building a model that predicts the accident rate for drivers. If the information about the age and gender of the driver is omitted, then it is having limited accuracy unless this information is indirectly available through other attributes.
  (iii)Knowledge about the Data: Data sets are accomplished by documentation that describes different aspects of the data. If the related these attributes are likely to provide highly redundant information. If the documentation is poor then analysis of precision of the data, the types of features (normal, ordinal, interval, ratio), the scale of measurement (e.g. meters or feet for length) and the origin of the data.