Data is a collection of raw facts and figures.

① What Motivated Data Mining? why it is important.

* The main reason that dataMining has, extracted the information. why extracted the information.

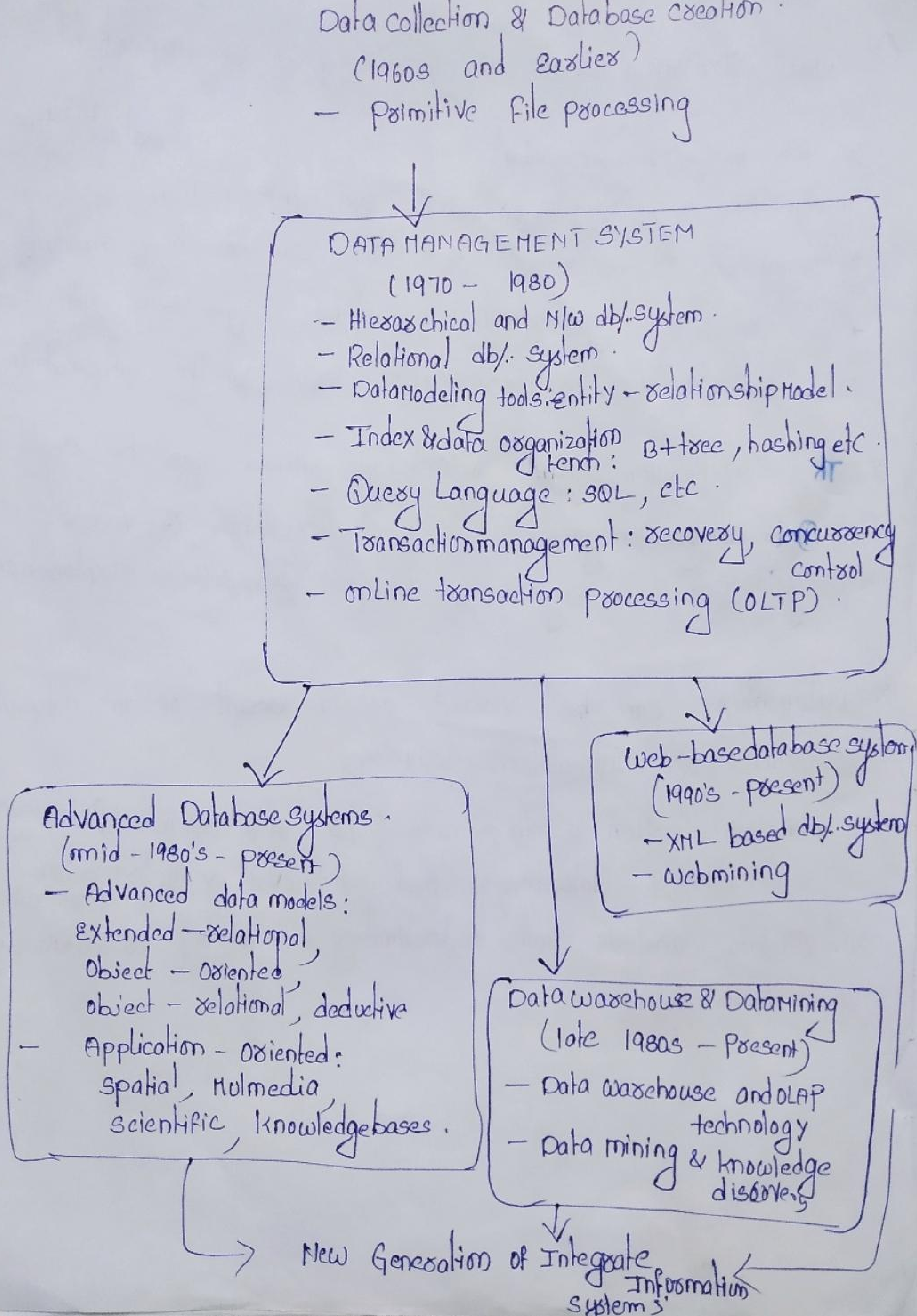
* Today (or) Recent Years the data can be huge amount.

* So, In that huge amount of data, by using the DM/. we can extract the information (or) gain the information.

* This information and knowledge gained can be used for ① business Managgement ② productioncontrol ③ market analysis. ④ To Engineering design ⑤ science exploration.

* Datamining can be viewed as a result of the natural evolution of information technology

*① datacollection & db creation. (1960' and earlier), ②datamanagement (including datastorage and retrieval & db transaction Processing). and ③ data analysis and understanding (involving dawarhousing and datamining)



Data collection & Database creation.
(1960s and Earlier)
- Primitive file processing

DATA MANAGEMENT SYSTEM
(1970 - 1980)
- Hierarchical and N/w db/. system.
- Relational db/. system.
- DataModeling tools: entity - relationship Model.
- Index & data organization tech: B+tree, hashing etc.
- Query Language: SQL, etc.
- Transaction management: recovery, concurrency control
- online transaction processing (OLTP).

Advanced Database systems.
(mid - 1980's - present)
- Advanced data models: extended - relational, Object - oriented, object - relational, deductive
- Application - oriented: spatial, Mulmedia, scientific, knowledgebases.

Web-basedatabase system
(1990's - present)
- XML based db/. system
- webmining

Datawarehouse & DataMining
(late 1980s - Present)
- Data warehouse and OLAP technology
- Data mining & knowledge disoovery

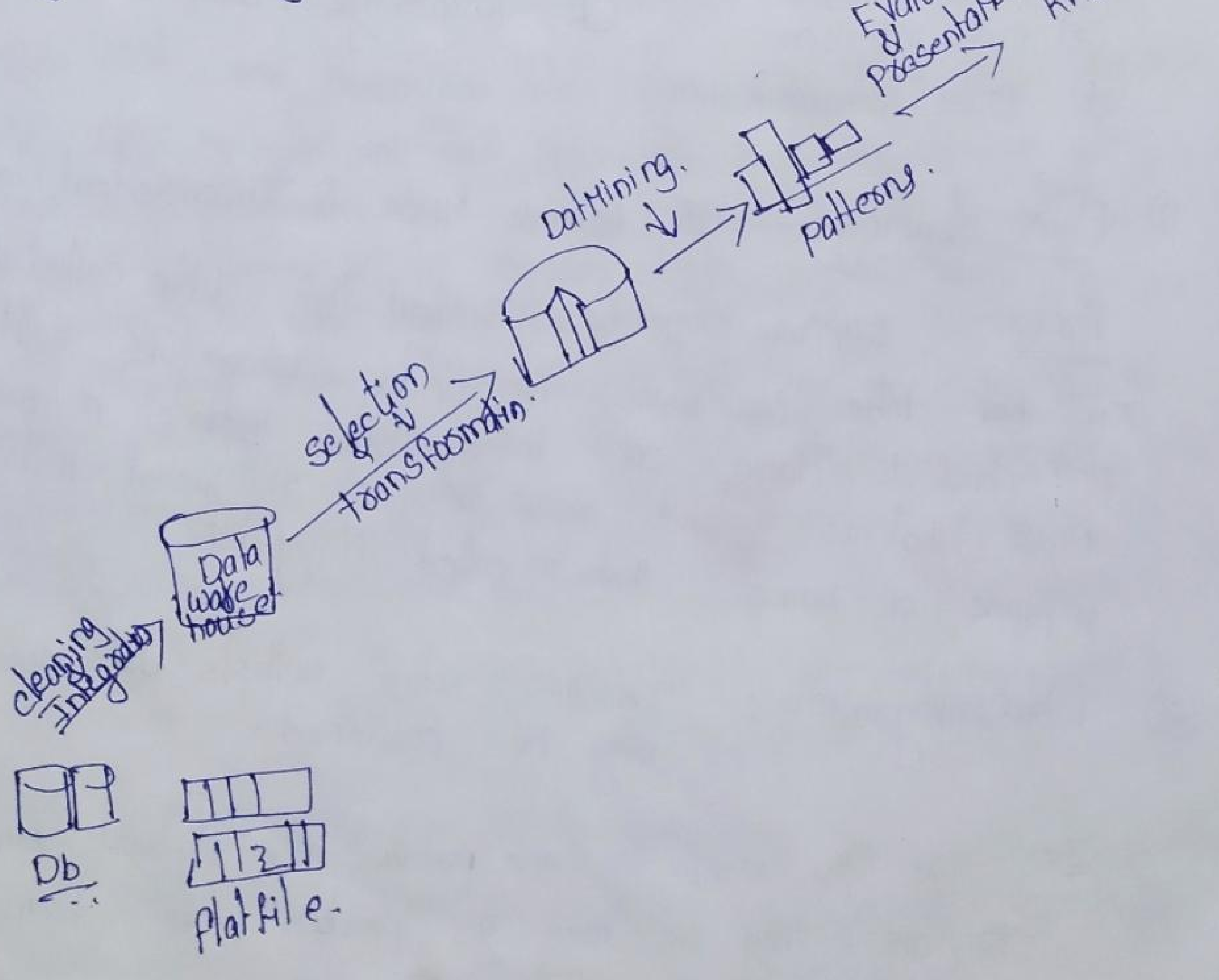
New Generation of Integrate Information systems.

* what is Data Mining ?

Simply stated, datamining refers to extracting (or) "mining" knowledge from large amount of data.

* There are many other terms carrying a similar (or) slightly different meaning to datamining such as knowledge mining from db, knowledge extraction, data/ Pattern analysis, data archaeology, data dredging.

* Data mining another Popularly used term, is knowledge Discovery in Db/. (or) KDD.



KDD:- The obrevation of KDD is knowledge Discovery db/..

* The Main objective of the KDD process is to extract information from data inthe context of large db/..
* ~~I~~ i.e it is like retrive the information from huge amount of data.

* If you want to get the knowledge to entire the db/. (i.e) huge amount of data, we need to follow the 7 steps.

1. Datat cleaning.
2. Data Integration.
3. Data Selection.
4. Data Transformation.
5. DataMining.
6. Pattern evaluation.
7. knowledge presentation.

① Data cleaning:- To Remove Noise & Inconsistent data.

for ex:- Suppose I have student db/. table called studend.. In that table we have two fields 1st one is Hallticket Num. and one is Name. If insert name inplace of Hallticket Num that is called Noise data. If insert Hallticket inplace of Name that is called Noisy data.

② DataIntegration:- In this step Multiple datasources may be combined.

Ex:- If you want to one product, In flipkart, Amazon, Hopscoth, they we may get product details from db/..

If You want purchase that Product, obviously we need to enter the your card details, bank related details. so, bank related data is available in bank db. so, here we are using Multiple db comined means data must be integrated same time this called just like data Integration.

③ Data selection :- Data relevant to the analysis task are retrived from the db/.

Ex:- If You want to know the student percentage in the student table so, need n't to select all the attributes for calculate the percentage, just select Hallticket No field, grade fied and percentage field. By using these three (or) two field i.e Hallticket No and percentage field, so, we will get the details of Percentage about student instead of all attributes which is available in the student table we are going to select two fields this is just like selection.

Another example of selection, if You want know the employee salary in employee table, so obviously we need to select employee id & employee salary field. Instead of the all attribute we are going to select only two fields

so these are way to identify the required field in db table. so, this is called data selection.

④ Data transformation :- After completing the above 3 steps i.e Datacleaning, Data Integration, Data selection, so Data Transformation is required so, In this step, Data is transformed (or) consolidated in forms appropriate for mining by Performing summary (or) aggregation operation.

⑤ Data Mining :– Here Intelligent methods are applied in order to Extract data Patterns.

Ex:– suppose if want to know the highet Paid salary Employee in the organization obviously applied an Intelligent Methd.

⑥ Pattern Evaluation :– In this step, data patterns are evaluated so, based on the User requirement, Enduser Pattern evaluation are useful.

⑦ knowledge presentation :– The last step in the KDD is knowledge is represented i.e Here visualization & knowledge representation techniques are used to Present the mined knowledge to the user.

we want to get the knowledge in the db (or) from the data abviously we have to followed these steps.

First data is available in diff db/.. from that db we are going to get the data, Next we have to clean the data after cleaning the data again we are going to select the data and transform the data after selecting the data. so, obviously the data is transformed to datamining Datamining just like cleaned data so, here we get only aggregate related data so, there will be get only numeric data after that datamining data is send to pattern evaluation so, here based on the used requirement the datamining is converted in to graphical representation i.e we will get chart, barchart, flowchart, piechart a so.on ..

* Based on the pattern evaluation the end user get the knowledge.

* we get the knowledge the end user take decision.

Data Mining Task :-

* Data mining task can be classified in to two categories.

① Descriptive Task ② predictive. Task.

① Descriptive Task :-

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

* predictive mining tasks makes the prediction based on current data.

Descriptive Mining task :-

* It focus on the summarization and conversion of the data in to meaningful information for reporting and monitoring.

* It Permits to examine the data in a detailed way so, that it would be able to answer easily about "what has happened?" & what is happening?

for ex:- A customer is s/w employee. He may purchase computer, labtop, (or) s/w. And depends upon those Particular those characterstic we can able to describe the Particular customer is called descriptive.

* clustering, summarization, association analysis are the technique categorized under descriptive mining"

(i) Association Analysis:-

* association analysis is the ~~disa~~ discovery of association rules showing attribute-value condition that occur frequently together in a given set of data.

* Association analysis is widely used for market basket (or) transaction data analysis.

i.e association analysis is useful for discovering Intersting relation ship hidden in ~~da~~ large datasets.

for ex:- Given the All Electronics relational db/. a data mining system may find association rules like

$$age(x, \text{"}20 \ldots 29\text{"}) \wedge income(x, \text{"}20k \ldots 29k\text{"}) \Rightarrow buys(x, \text{"CDplayer"}).$$

[support = 2%, confidence = 60%] .

where x is a Variable representing a customer. The rule indicates all electronics customers under study 2% (support) are 20 to 29 Years of age with an income of 20k to 29K and have purchase a CD player at All Electronics .

~~clustering~~ cluster analysis :—

* cluster analysis is a statistical classification technique. In which a set of objects (or) points with similar characteristic are grouped together in cluster

* It encompasses a number of different algorithm and methods that are all used for grouping object of similar kinds into respective categories.

* Descriptive mining describe unsupervised .

Predictive Mining Task :-

Material

Data Mining Tasks:

Data mining tasks are generally divided into two major categories:

1) Predictive tasks

2) Descriptive tasks

1. Predictive tasks:

The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as the explanatory or independent variables.

Classification and Regression:

Predictive modeling refers to the task of building a model for the target variable as a function of the explanatory variables.

Classification is used for discrete target variables, and regression, which is used for continuous target variables. For example, predicting whether a Web user will make a purchase at an online bookstore is a classification task because the target variable is **binary-valued.** On the other hand, forecasting the future price of a stock is a regression task because price is a **continuous-valued** attribute. The goal of both tasks is to learn a model that minimizes the error between the predicted and true values of the target variable.

Anomaly detection:

Anomaly detection is the task of identifying observations whose characteristics are significantly different from the rest of the data. Such observations are known as anomalies or outliers. Applications of anomaly detection include the detection of fraud, network intrusions.

2. Descriptive tasks:

Here, the objective is to derive patterns (correlations, trends, clusters, trajectories) that summarize the underlying relationships in data. Descriptive data mining tasks are often exploratory (investigate) in nature and frequently require postprocessing techniques to validate and explain the results.

Association analysis:

Association analysis is used to discover patterns that describe strongly associated features in the data. The discovered patterns are typically represented in the form of implication rules or feature subsets. Useful applications of association analysis include finding groups of genes that have related functionality, identifying Web pages that are accessed together.

Cluster analysis:

Cluster analysis seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other than observations that belong to other clusters. Clustering has been used to group sets of related customers, find areas of the ocean that have a significant impact on the Earth's climate.

Attribute types in datamining:—

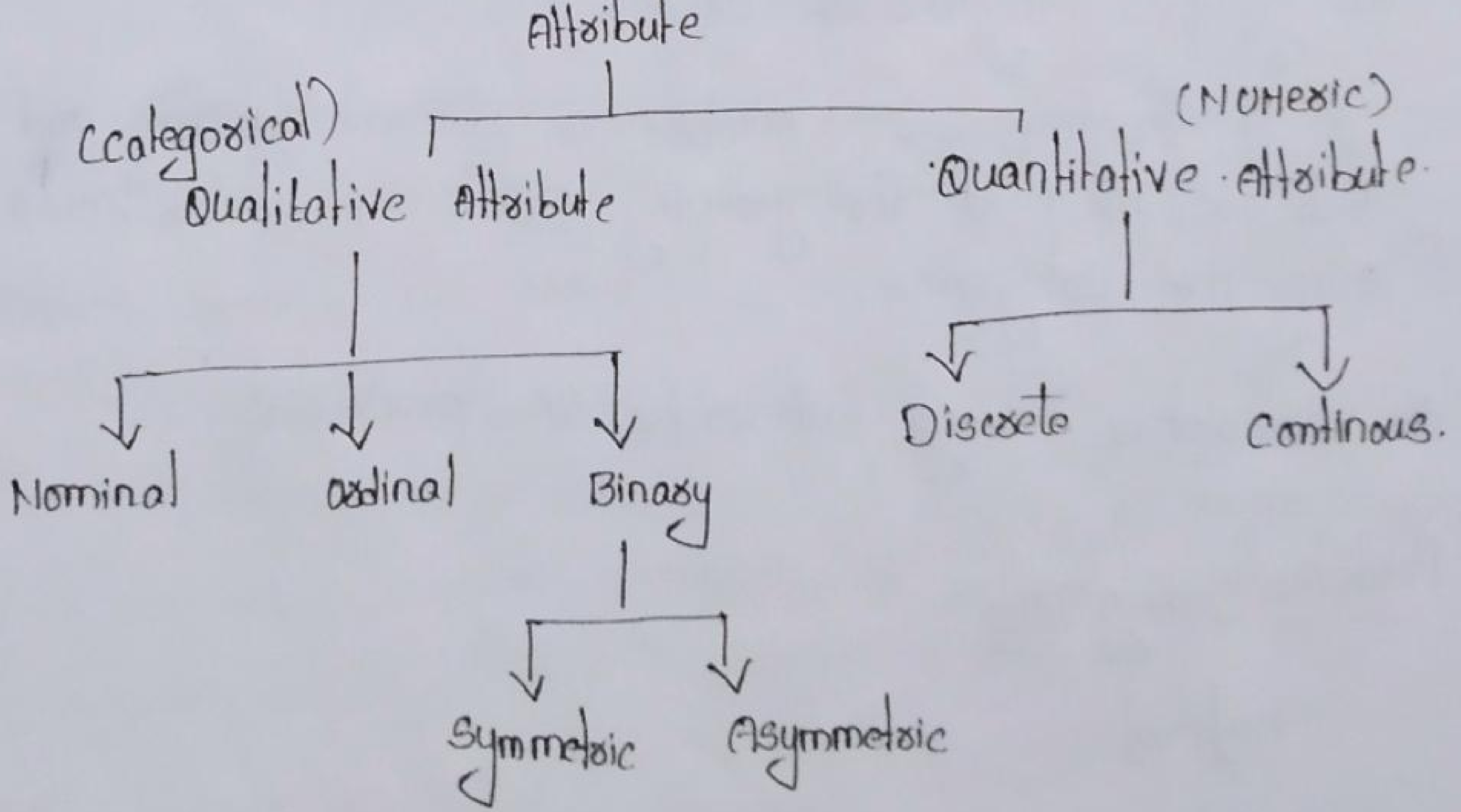
The attribute can be defined as a field for storing the data that represent the characteristic of a data object. The attribute is the Property of the object.

* The attribute represents different features of the object.

ex:- hair color is the attribute of the lady, similarly rollno & marks are attribute of the student.

Types of Attributes:— There are two types of the attributes

① Qualitative Attribute ② Quantitative Attribute.



Example of attribute:-

In this example, Roll No, Name, & Result are attributes of the Object named as a student.

Roll No	Name	Result.
1.	ABC	Pass.
2.	XYZ	Fail.

1. Qualitative Attribute:- (categorical)

Nominal attribute:- Nominal attribute is in alphabetical form and not in an Integer. Nominal attributes are Qualitative Attribute.

Ex:- zipcodes, employee ID numbers, eye color, gender

ordinal attributes:- The values of an ordinal attribute provide enough information to order objects i.e All values have a meaning full order.

Ex:- Hardness of minerals {good, better, best}
Grades - A means highest marks, B means marks are less tha A, C means marks are less than grades A & B.

Quantitative Attribute:- (Numeric)

1. A Numeric attribute is Quantitative because, it is a measurable Quantity, represented in Integer (or) real values.

* Numerical attribute are two types. (i) Interval & (ii) Ratio.

(Here from the name only you can say so, there is some Interval where ever this interval is there the length of that inerval is same)

*(i) Interval-scaled :- An interval-scaled attribute has values, whose differences are interpretable but the numemerical attributes do not have the correct reference Point, (or) we can call zero points. Data can be added and subtracted at an interval but ~~we~~ can not be Multiplied (or) divided.

ex :- ~~ca~~ calender dates, temperature in celsius or) F

(ii) Ratioscaled :- attribute is a numeric attribute with a fix zero-point. if a ~~meause~~ measurement is ratio-scaled, we can say of a value as being a mutiple(or ratio of another value.

ex :- temperature in kelvin, length, age, etc.

Discrete attribute :- A discrete attribute has a finite (or) countably infinite set of values. It can be Numerical and can also be in categorical form.

* profession, : Teacher, Business Man.

zipcode : 301701, 110040.

continous attribute:- A continous attribute is one whose Values are real numbers.

Ex:- Temperature, height, weight.

* continous attribute are typically represented as float-Point variables.

Ex:- Height : 5.4, 6.2 ... etc.
weight : 50.33 .. etc

Binary Attribute:- Binary attributes have two values/states. The binary attribute is of two types

1. symmetric Binary Attribute.
2. Asymmetric Binary Attribute.

symmetric Binary Attribute:- In this type Both values are Equally important.

Ex:- if we have open admission to our university, then it doesn't matter, whether you are a male (or) a female.

Ex:- Gender male female.

· Assy Asymmetric Binary Attribute:- In this type Both values are not equally important.

Types of Data sets:-

There are three general characteristics of Data sets.
① Dimensionality ② sparsity ③ Resolution.

① Dimensionality :- The dimensionality of a dataset is the number of attributes that the objects in the dataset have.

In a Particular dataset if there are high number of attribute (also called high dimensionality) then it can become difficult to analyse such a~~s~~ dataset.

② sparsity :- For some dataset, such as those with asymmetric features, most attribute of an object have values of 0; in many cases fewerthan 1% of the entries are non-zero. such In Particular Practical terms, sparsity is an advantage becoz, usually only the non zero values need to be stored and manipulated.

③ Resolution :- 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 noise. if the resolution is too coarse, the Pattren may disappear.

Finally coming on the types Datasets.

* Thre There are 3 types of datasets.

1. Record Data ② Graph-based Data ③ ordered Data.

1. Record Data :-

* Majority of Datamining work assumes that data is a collection of records.(dat objects).
* The most basic form of record data has no explicit relationship among records (or) data fields, and ever record has the same set of attributes.
* Record data is usually stored either in flat files (or) in relational db/.

Different types of record data are describe.

(i) Transaction (or) Market Basket Data.
(ii) Data matrix
(iii) Document - term matrix.

(i) Transaction data :-

* It is special type of record data, In which each record contains a set of items.

ex:- shopping in a supermarket (or) a grocery store. For any Particular customer, a record will contain a set of items Purchased by the customer in that respective visit to the supermarket or the grocery store.

* This type of data is called market-Basket Data.

TID	Items
1	Bread, Milk, JaM.
2	Bread, Biscuits.
3	Biscuits, Jam, Milk,
4	cookies, Bread, Milk, Jam.
5	Jam, Milk, Bread.

Here the above figure shows a sample transaction. dataset.

Each row represents the purchases of a particular customer at a particular Time.

Data Matrix:- Data object with only numeric attribute can be represented by an m by n matrix, where there are m rows, one for each object and n columns, one for each attribute.

Projection Of xLoad.	Projection. of Yload.	Distance	Load	Thickness.
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22.	2.2.	1.1

It consists entirely numeric attributes, so, this is entirely continous, (or) entirely interval, (or) ratio variables. Then we can think of it as a mathematical matrix rather than just a table. so, we would have an m by n matrix i.e There are mrows, one for each data object and columns, one for Each attribute.

Document - term matrix:-

* It is just like data matrix. Every term, every entry every data attribute has a numeric value. but

The document data matrix is also called A sparse data matrix. A spare datamatrix is a special case of data matrix in which the attributes are of the same type and are asymmetric. i.e only non zero values are important.

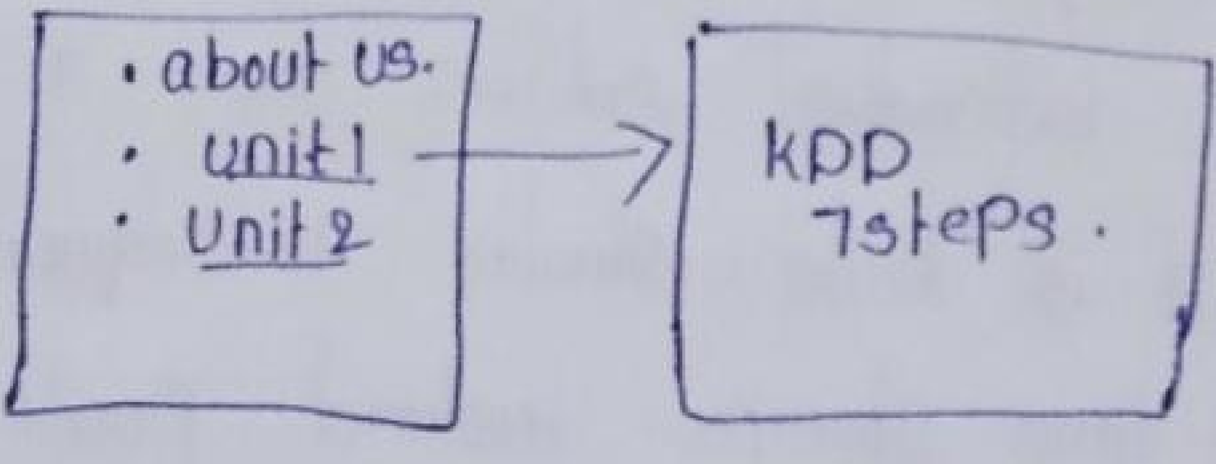
② Graph - Based Data :-

In this type we consider two specific cases.
(i) The graph captures relationships among data objects
(ii) The data objects themselves are represented as graphs.
This TYPE can be divided in to two types.

⑴ Data with Relationships among objects :-

The data objects are mapped to nodes of the graph, while the relationships among objects are captured by the links b/w objects and link properties, such as direction & weight.

consider webpages on the world wide web, which contain both text and links to other pages



② Data with Objects That are Graphs:-

If objects have structure, that is, the objects contain sub objects that have relationships, then such objects are frequently represented as graphs.

Ex:- The structure of chemical compounds can be represented by a graph, where the nodes are atoms and link b/w nodes are chemical bonds.

③ Ordered Data:-

For some types of data, the attributes have relationships that involve order in time (or) space.

In this type there are 4 types of Data.

1. Sequential Data:-

Time	customer	item pur
t1	C1	A,B
t2	C3	A,C
t2	C1	C,D

* Also referred to as temporal data, can be thought of as an extension of record data, where each record has time associated with it.

consider a retail transaction dataset that also stores the time at which the transaction took place.

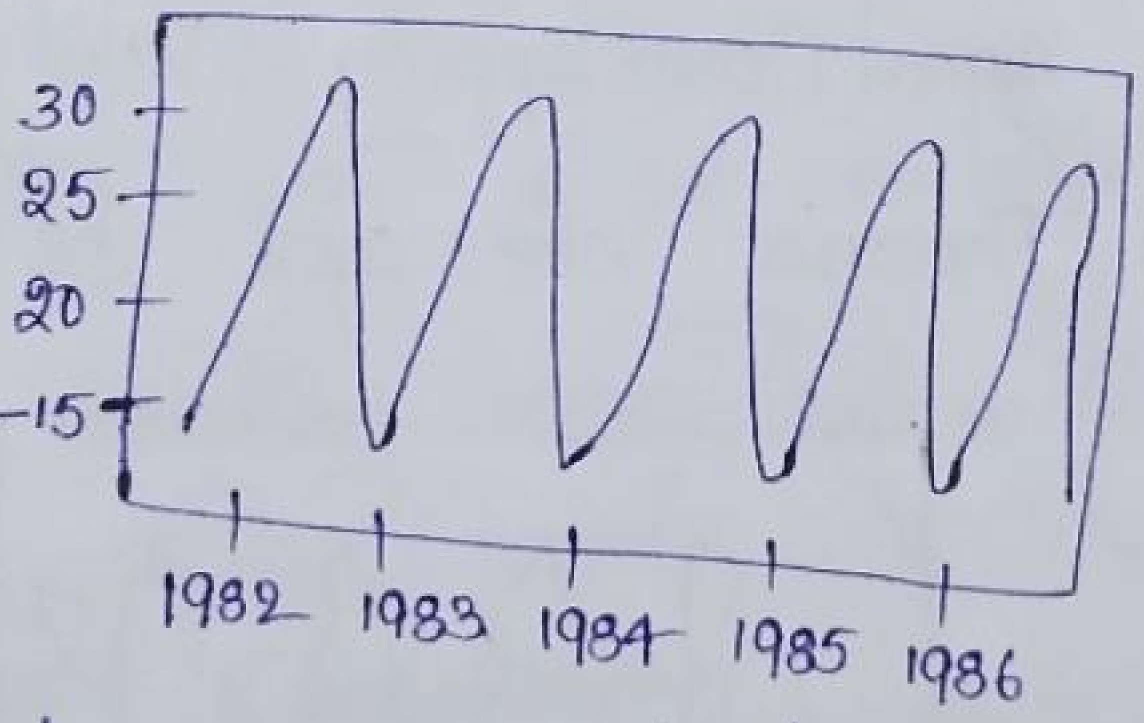
2. Sequence Data:- Sequence data consists of a dataset that is a sequence of individual entities, such as a sequence of word (or) letters. It is Quite similar to sequential data, except that there are no time stamps. instead position in an ordered sequence.

The below figure shows a section of the human genetic code expressed using the four nucleotides from which all DNA is constructed A, T, G, and C

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCC GCCC GCGCCGTC
CA GAAGGGCCCGCCCTGGCGGCG.

(3) 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 measurement taken over time. for ex,

A financial dataset might contain objects that are time series of the daily prices of various stocks.



④ Spatial Data:- Some objects have spatial attribute, such as positions or areas, as well as other types of attributes. An example of spatial data is weather data (temperature, pressure) that is collected for a variety of geographical locations.

Data Quality :— There are three fundamental Question around data Quality.

① what problem should we worry about?
② How can we detect problem with the data?
③ what can we do about these problems?

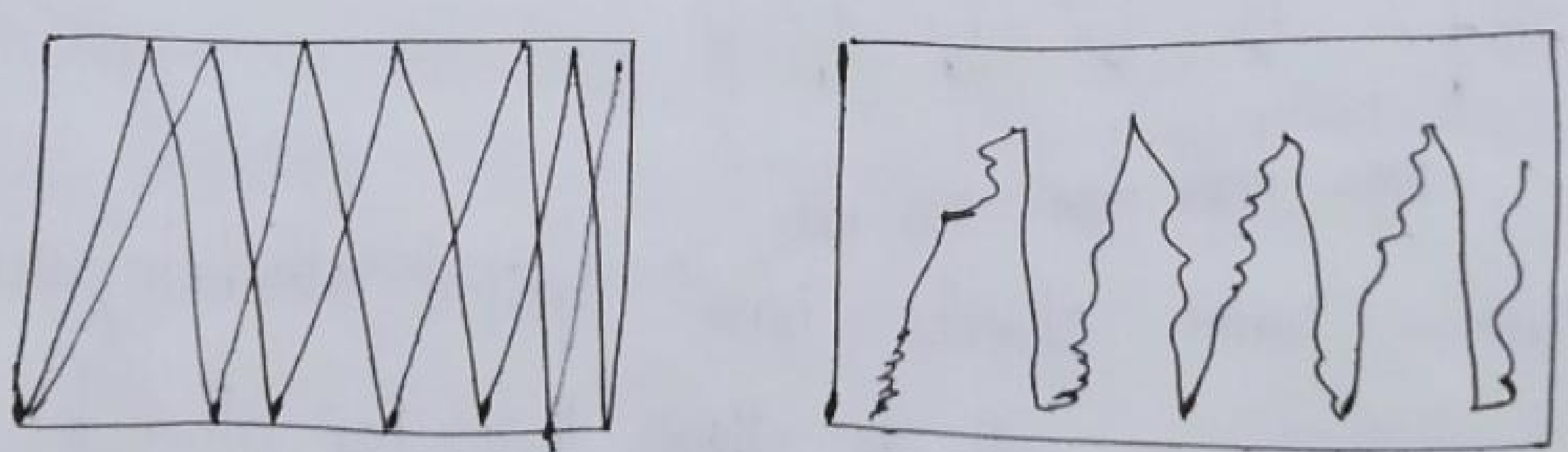
Examples of data Quality Problems: There 3 very common kinds kinds of data Quality Problems.

- Noise and outliers.
- Missing values.
- Duplicate values.

* Noise

Noise :- An invalid signal overlapping valid data

Ex:- distortion of a Person's voice over the phone.



Outliers:- outlier are either.

1. Data object that, in some sense, have characteristics that are different from most of the other data objects in the data set, (or)

2. Values of an attribute that are unusual with respect to the most common (typical) values for the attribute.

it is important to distinguish b/w noise & outliers. outlier is desirable & noise is desirable. outliers can be legitimate data obj/.. (or) values.

Missinvalues:-

if there are Missing values present in the dataset.

Reason for missing values.

* sometimes, missing values are because information is not collected.

eg:- People decline to give their age & weight.

Attributes may not be applicable to all cases.

eg:- annual income is not applicable to children.

How to Handling missing values :-

1. Eliminate Data objects.
2. Estimate Missing values.
3. Ignore the Missing values During Analysis.
4. Replace with all Possible values

Duplicate Data:-

Dataset may include data objects tha are ~~objects~~ duplicates (or) almost duplicates, of one another,. To detect and eliminate such duplicates, two main issues must be addressed.

* Fisst, 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.

* second, care need to be taken to avoid accidentally combining data objects that are similar, but not duplicates, such as two distinct people with identical names.

Measures to check Data Quality

1. precision ② Bias ③ Accuracy

* The Quality of measurement process and the resulting data are measured by precision & Bias

Precision:- The closness of repeated measurements (of the same Quantity) to one another. It is often measured by the standard deviation of a set of Values.

Data Quality:

In this section the data mining focus on measurement and data collection issues(data cleaning), although some application-related issues are also discussed.

1. Measurement and Data Collection Issues:

It is unrealistic to expect that data will be perfect. There may be problems due to human error, limitations of measuring devices, or flaws in the data collection process. Values or even entire data objects may be missing. In other cases, there may be spurious or duplicate objects.

Measurement and Data Collection Errors:

The term measurement error refers to any problem resulting from the measurement process. A common problem 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 term data collection error refers to errors such as omitting data objects or attribute values, or inappropriately including a data object.

Noise and Artifacts:

Noise is the random component of a measurement error. It may involve the distortion of a value or the addition of spurious objects. The deterministic distortions of the data are often referred to as artifacts.

Precision, Bias, and Accuracy:

In statistics and experimental science, the quality of the measurement process and the resulting data are measured by precision and bias.

Precision- The closeness of repeated measurements (of the same quantity) to one another.

Bias- A systematic variation of measurements from the quantity being measured.

Precision is often measured by the standard deviation of a set of values, while bias is measured by taking the difference between the mean of the set of values and the known value of the quantity being measured. Bias can only be determined for objects whose measured quantity is known by means external to the current situation. Suppose that we have a standard laboratory weight with a mass of 1g and want to assess the precision and bias of our new Laboratory scale. We weigh the mass five times, and obtain the following five values: {1.015, 0.9901, 0.0131, 0.001, and 0.986}. The mean of these values is 1.001, and hence, the bias is 0.001. The precision, as measured by the standard deviation, is 0.013.

Accuracy- The closeness of measurements to the true value of the quantity being measured.

Outliers:

Outliers are either (1) data objects that, in some sense, have characteristics that are different from most of the other data objects in the data set. Furthermore, it is important to distinguish between the notions of noise and outliers. Outliers can be legitimate data objects or values. Thus, unlike noise, outliers may sometimes be of interest. In fraud and network intrusion detection.

Missing Values:

In some cases, the information was not collected; e.g., some people decline to give their age or weight. In other cases, some attributes are not applicable to all objects.

Eliminate Data Objects or Attributes:

A simple and effective strategy is to eliminate objects with missing values. However, even a partially specified data object contains some information, and if many objects have missing values, then a reliable analysis can be difficult or impossible.

Estimate Missing Values:

Sometimes missing data can be reliably estimated. For example, consider a time series that changes in a reasonably smooth fashion, but has a few, widely scattered missing values. In such cases, the missing values can be estimated (interpolated) by using the remaining values.

If the attribute is continuous, then the average attribute value of the nearest neighbours is used; if the attribute is categorical, then the most commonly occurring attribute value can be taken.

Ignore the Missing Value during Analysis:

Many data mining approaches can be modified to ignore missing values. For example, 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 attributes, then the similarity can be calculated by using only the attributes that do not have missing values.

Inconsistent Values:

Data can contain inconsistent values. Consider an address field, where both a zip code and city are listed, but the specified zip code area is not contained in that city. It may be that the individual entering this information transposed two digits, or perhaps a digit was misread when the information was scanned from a handwritten form. Regardless of the cause of the inconsistent values, it is important to detect and, if possible, correct such problems.

Duplicate Data:

A data set may include data objects that are duplicates, or almost duplicates, of one another. Many people receive duplicate mailings because they appear in a database multiple times under slightly different names. To detect and eliminate such duplicates, two main issues must be addressed.

First, 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.

Second, care needs to be taken to avoid accidentally combining data objects that are similar, but not duplicates, such as two distinct people with identical names. The term **deduplication** is often used to refer to the process of dealing with these issues.

2. 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."

Timeliness:

Some data starts to age as soon as it has been collected. In particular, if the data provides a snapshot of some ongoing phenomenon or process, such as the purchasing behavior of customers or Web browsing patterns, then this snapshot represents reality for only a limited time. If the data is out of date, then so are the models and patterns that are based on it.

Relevance:

The available data must contain the information necessary for the application. Consider the task of building a model that predicts the accident rate for drivers. If information about the age and gender of the driver is omitted, then it is likely that the model will have limited accuracy unless this information is indirectly available through other attributes.

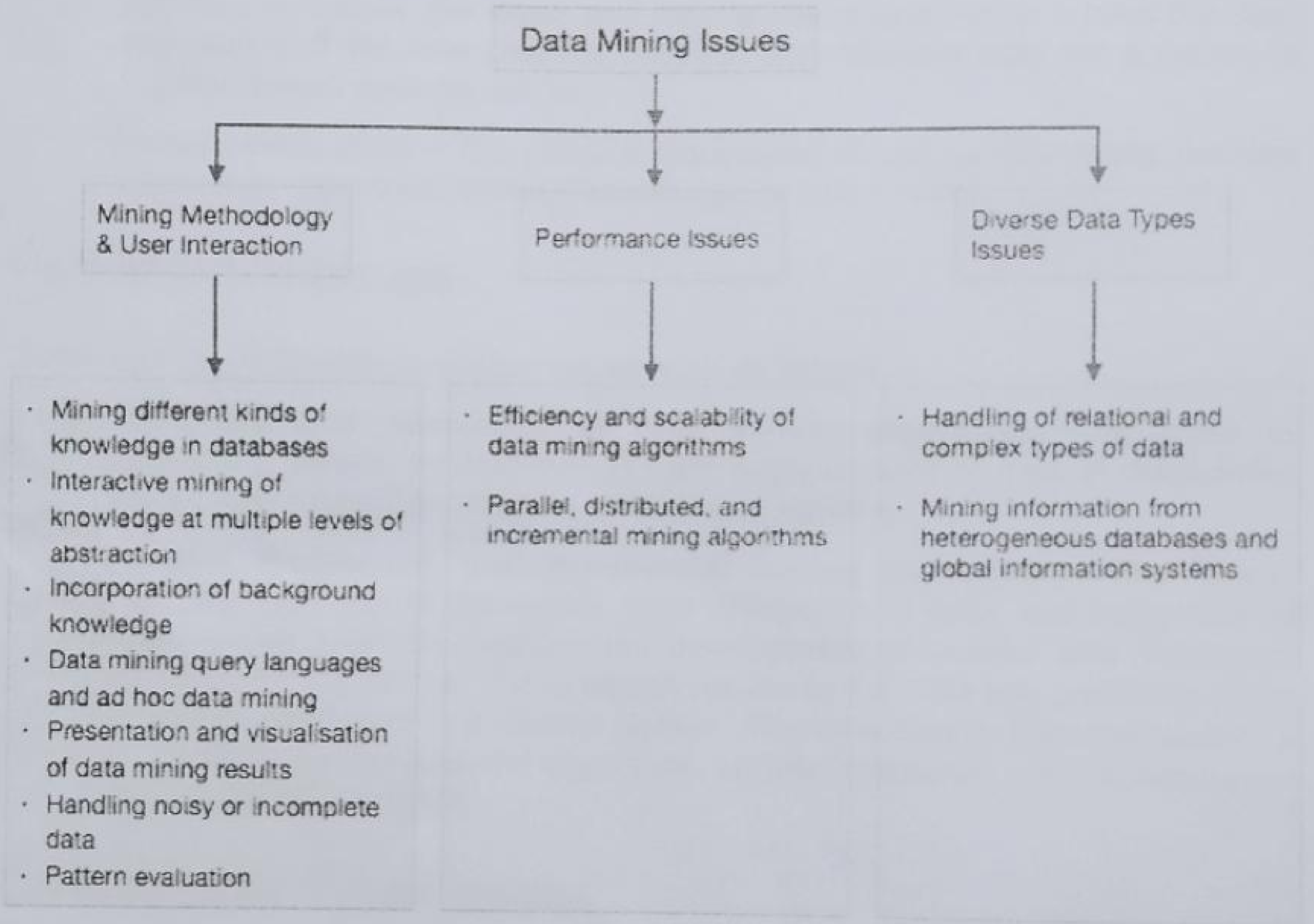
Knowledge about the Data:

Ideally, data sets are accompanied by documentation that describes different aspects of the data; the quality of this documentation can either aid or hinder the subsequent analysis. if the documentation identifies several attributes as being strongly related, these attributes are likely to provide highly redundant information, and we may decide to keep just one.

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding −

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.


Data Mining Issues
Mining Methodology & User Interaction
Performance Issues
Diverse Data Types Issues
Mining different kinds of knowledge in databases
Interactive mining of knowledge at multiple levels of abstraction
Incorporation of background knowledge
Data mining query languages and ad hoc data mining
Presentation and visualisation of data mining results
Handling noisy or incomplete data
Pattern evaluation
Efficiency and scalability of data mining algorithms
Parallel, distributed, and incremental mining algorithms
Handling of relational and complex types of data
Mining information from heterogeneous databases and global information systems

Mining Methodology and User Interaction Issues

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used.

Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.

- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.