Data mining Handwritten Unit-1

Data Mining And Analytics (SRM Institute of Science and Technology)

Data Mining

Unit - 1

Data :- Data is the raw facts and figures upon which we perform some procedure to gather information.

HDD
Tapes → .txt / .excl → Extract → Transform (in single form) → Load
Excell

Complete Process → ETL

Data Mining :- Data mining is the method of extraction of data or previously unknown data patterns from huge sets of data. It is also known as Knowledge Discovery ~~Process~~ in Database (KDD).

* Gregory Piatetsky Shapiro coined the term KDD.

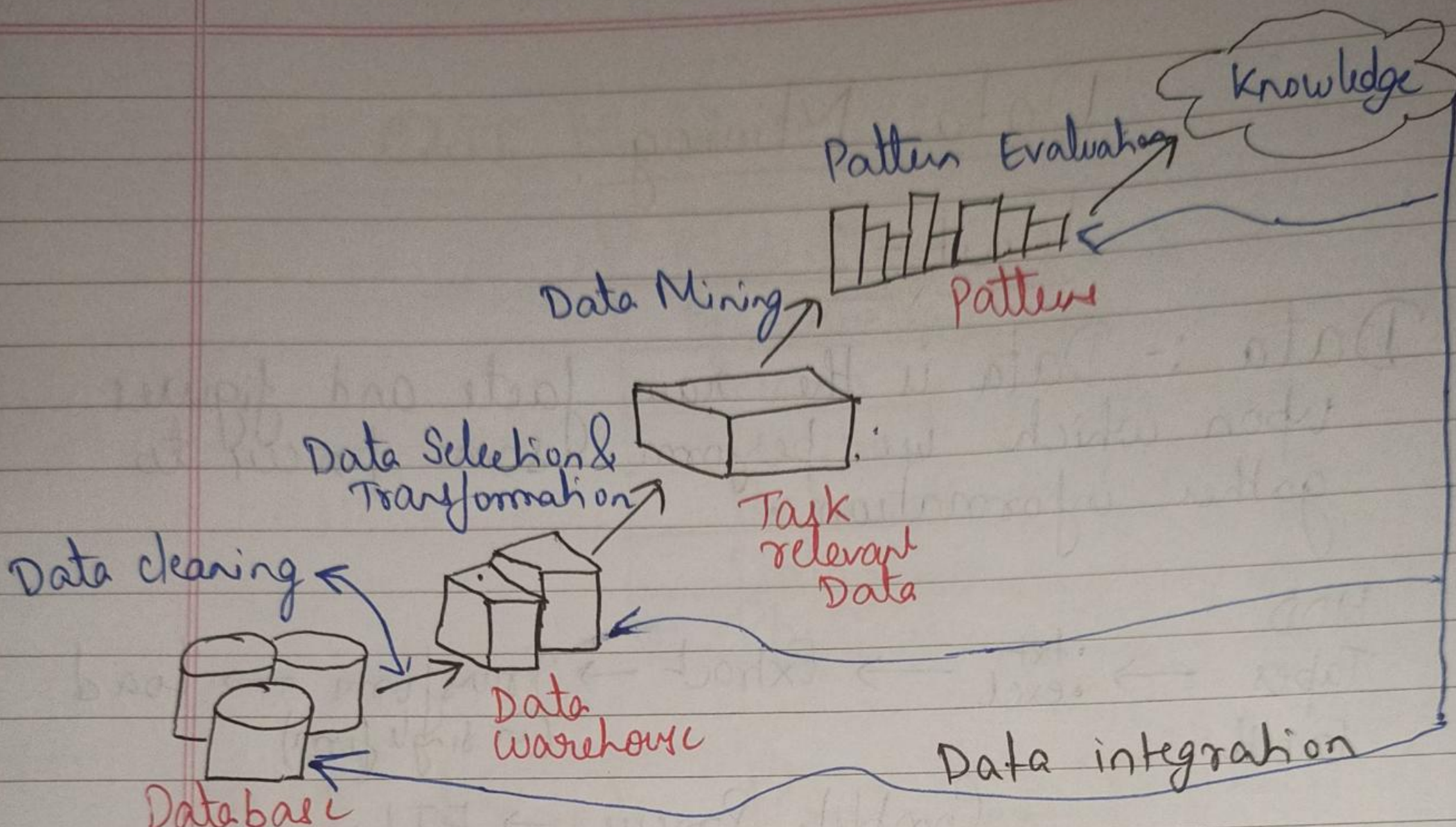
Why Data Mining?

Data-Mining is used in Business to make better managerial decision by:

(a) Automatic summarization of data.
(b) Extracting essence of information stored
(c) Discovering patterns in raw data.

KDD

Steps involved in KDD include :-



(i) Data Cleaning :- Data cleaning is defined as removal of noisy, inconsistent & irrelevant data from collection.

* Cleaning in case of ~~noisy~~ missing data
* Cleaning in noisy data, where noise is random
* cleaning with data discrepancy detection.

(ii) Data integration :- Data integration is defined as heterogenous data from multiple sources combined in a common source.

* Data integration using Data Migration Code.
* Data integration using Data Synchronization tools
* Data integration using ETL process.

Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.

- Data selection using **Neural network**.
- Data selection using **Decision Trees**.
- Data selection using **Naive bayes**.
- Data selection using **Clustering**, **Regression**, etc.

Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure. Data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

Data Transformation is a two step process:

- **Data Mapping**: Assigning elements from source base to destination to capture transformations.
- **Code generation**: Creation of the actual transformation program.

Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful. It is an essential process where intelligent methods are applied in order to extract data pattern.

- Transforms task relevant data into **patterns**.
- Decides purpose of model using **classification** or **characterization**.

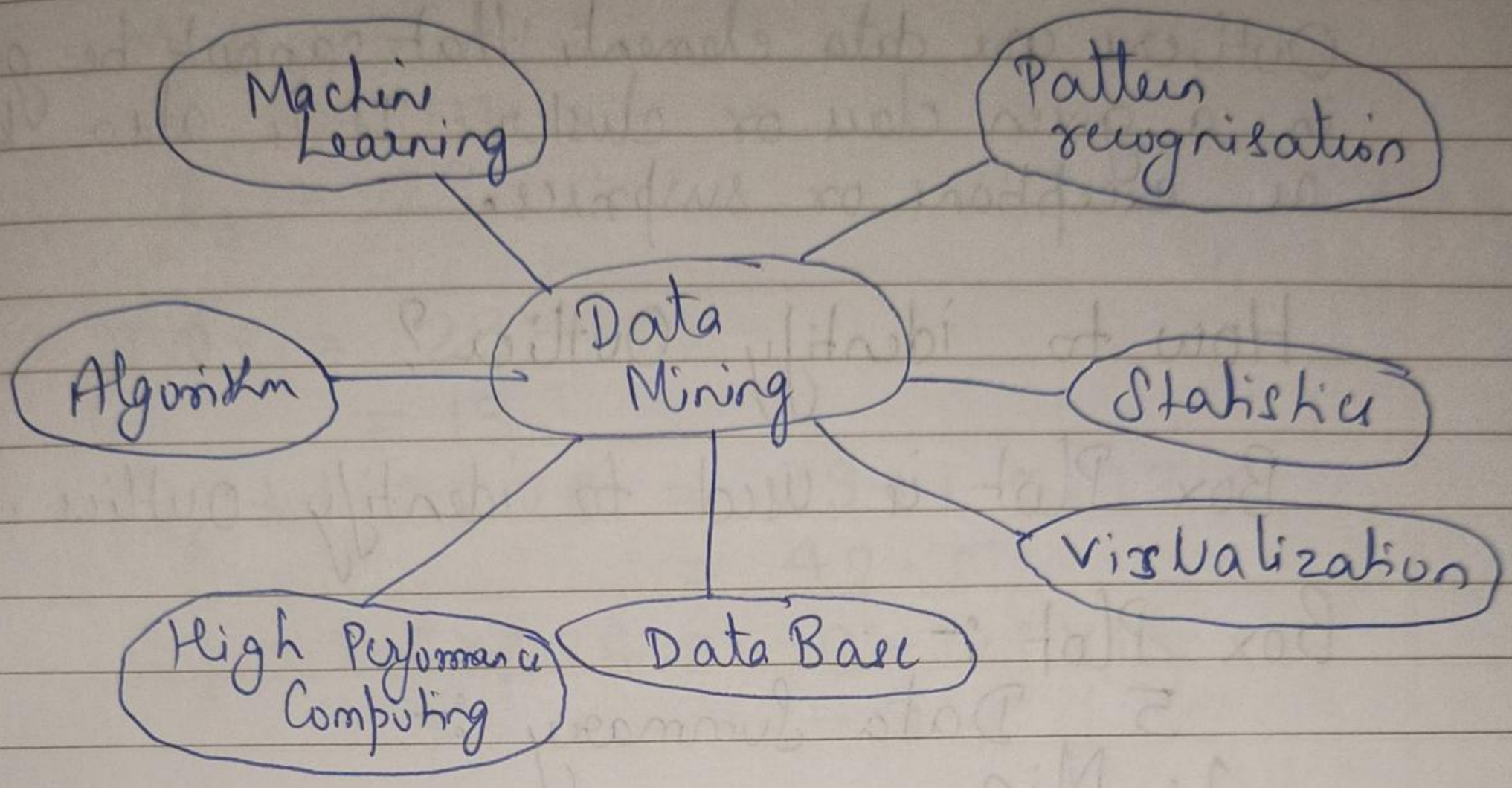
Pattern Evaluation: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.

- Find **interestingness score** of each pattern.
- Uses **summarization** and **Visualization** to make data understandable by user.

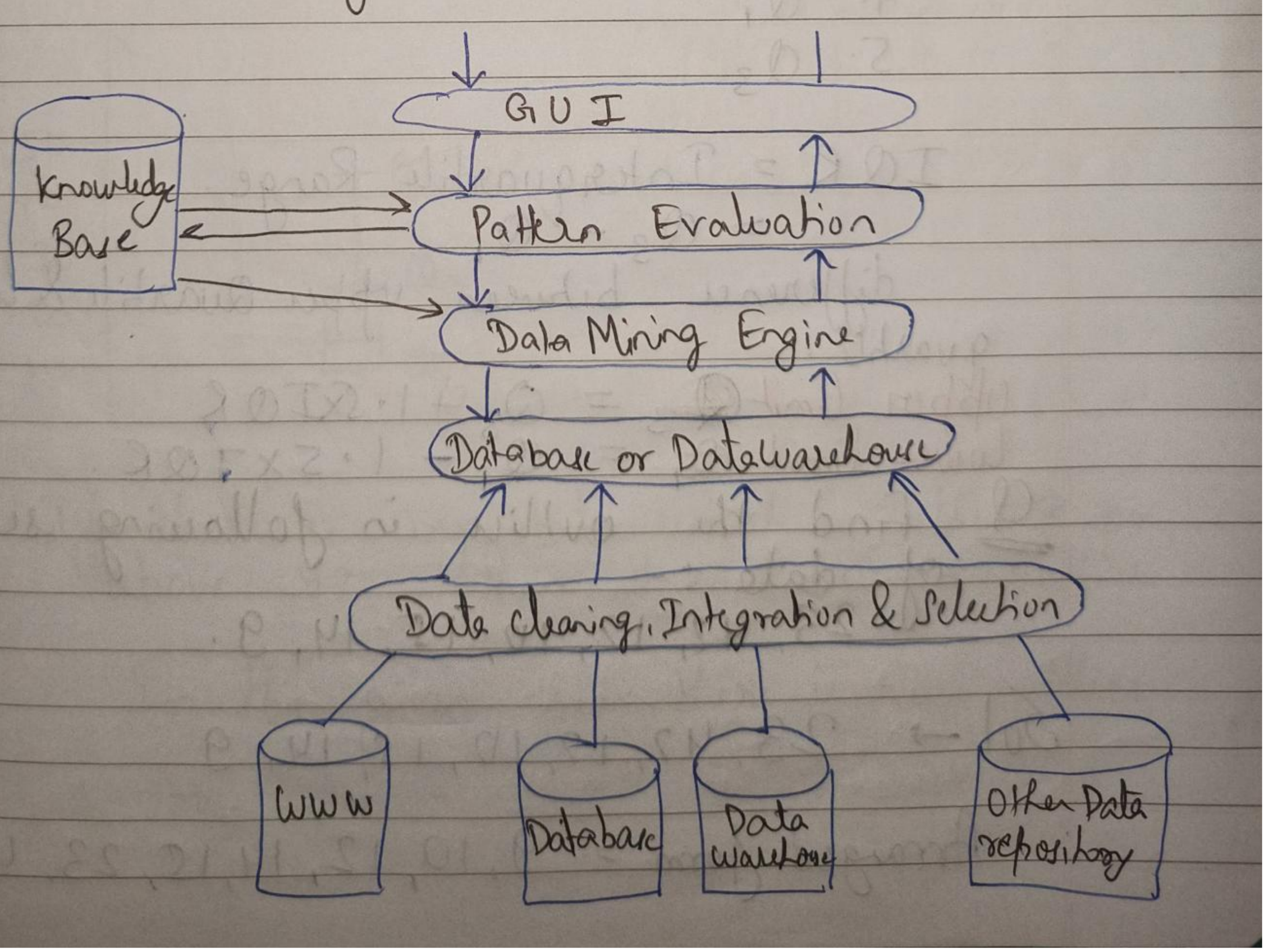
Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

- Generate **reports**.
- Generate **tables**.
- Generate **discriminant rules**, **classification rules**, **characterization rules**, etc.

Tools for Data Mining :-



Data Mining Architecture :-



Outliers :-

Outliers are data elements that cannot be grouped in a given class or cluster. It is also known as exceptions or surprises.

How to identify Outlier?

Box Plot is used to identify outlier.

Box Plot :-

5 Data Summary :-

1. Min
2. Max
3. Median
4. Q_1
5. Q_3

IQR = Interquartile Range.

$= Q_3 - Q_1$

difference between upper Quartile & lower quartile.

Upper limit $= Q_3 + 1.5 \times IQR$

lower limit $Q_1 = Q_1 - 1.5 \times IQR$.

Q find the outlier in following set of data :-

23, 42, 12, 10, 15, 14, 9.

Sol → 23, 42, 12, 10, 15, 14, 9

Arranged form = 9, 10, 12, 14, 15, 23, 42.

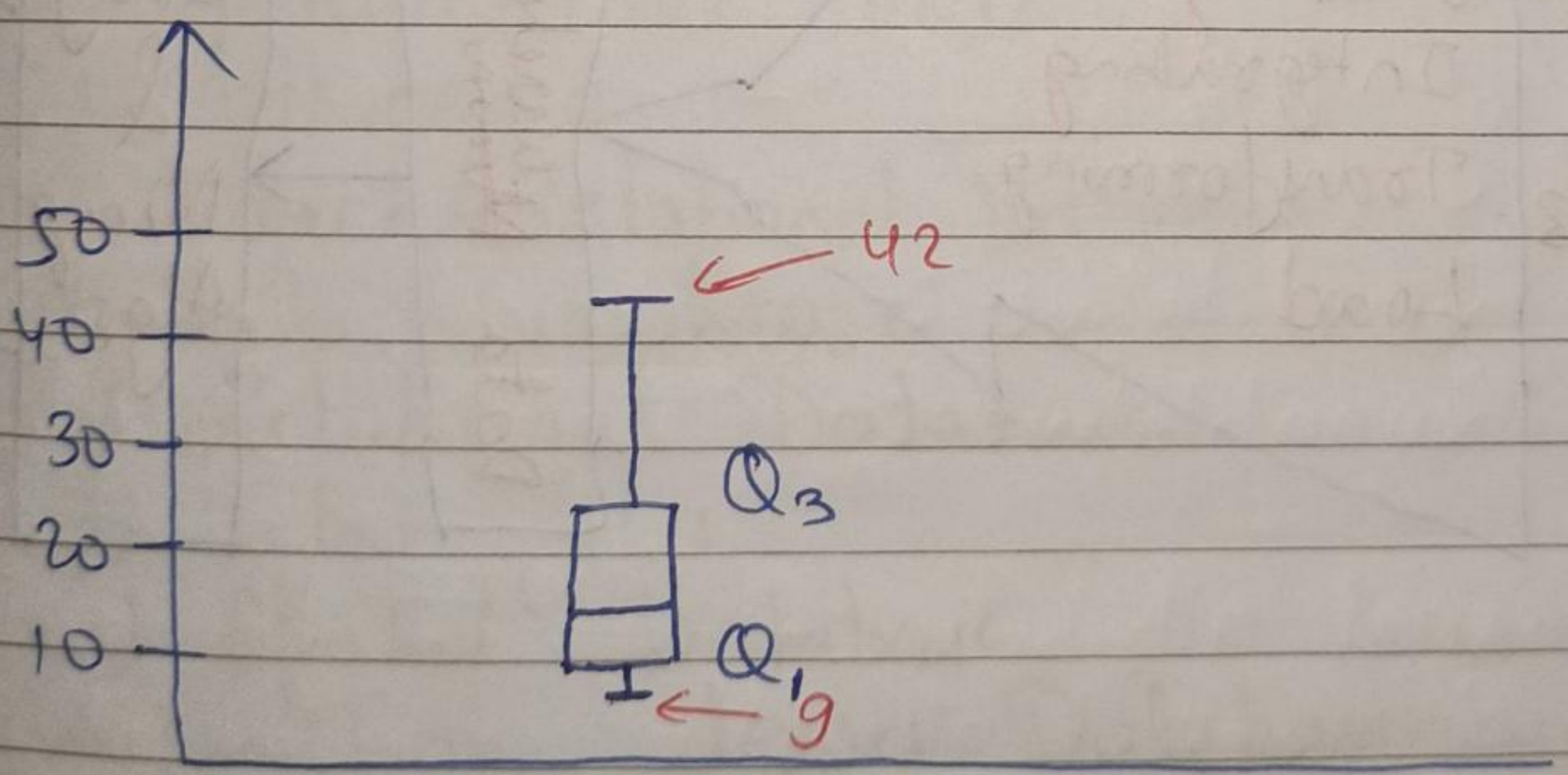
Min = 9
Max = 42
Median = 14.
Q_1 = 10
Q_3 = 23.

IQR = 23 − 10
= 13.

$Q_3 + 1.5 \times IQR = 23 + 1.5 \times 13$
$= 42.5$

$Q_1 - 1.5 \times IQR = 10 - 1.5 \times 13$
$= -9.5$.

min (upper limit, Max) = 42
max (lower limit, Min) = 9.



If there is any data which is below 9 or above 42, then it will be an outlier.

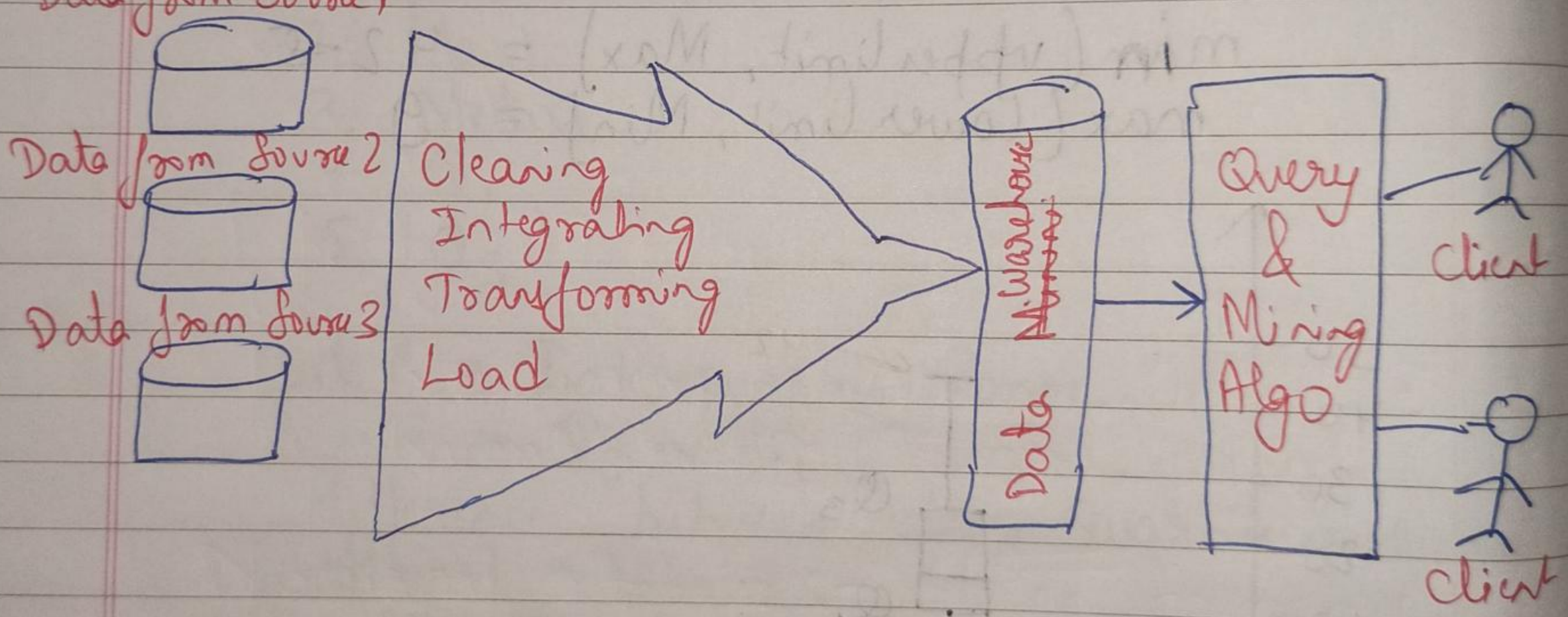
In this given set there is no outliers.

Question for Practice :-

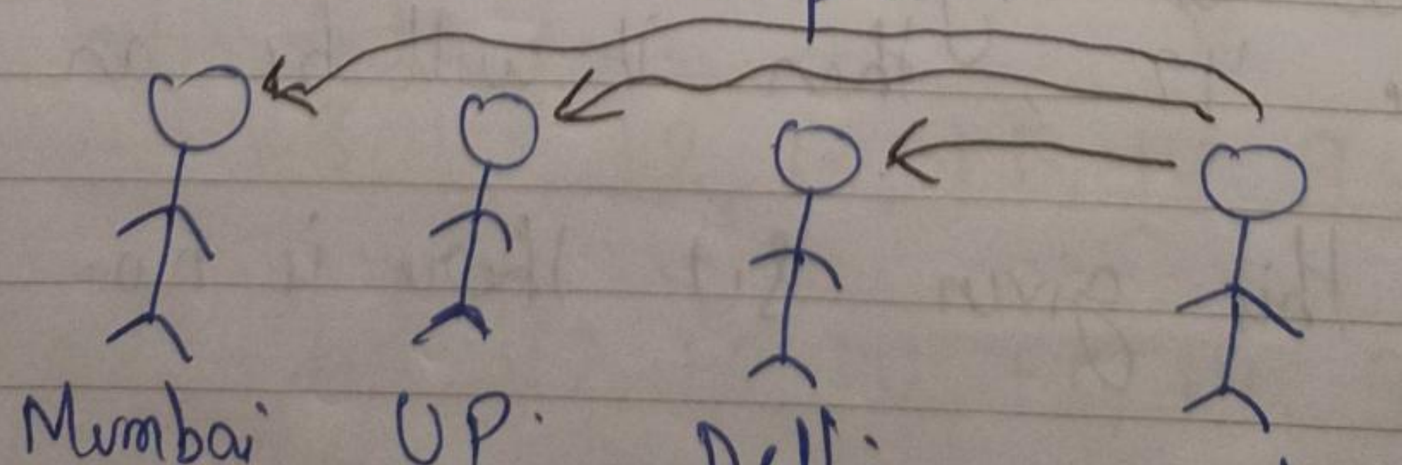
Q. Find the outlier 45, 23, 12, 10, 15, 4, -15.

Sol → Try yourself. Ans → If there is any data between 45 & -15 then there will be outlier. In this there is no outlier.

Data Warehouse :-



To understand why we needed Data warehouse lets see an example :-



If is a Business, many branches can have their own data and these data when asked are processed individually, which is time consuming.
So, data warehouse is created in which all the branches data is stored and can be accessed by head in single processing.

* B-Tree is used in database
* Data cube is used in Data warehouse.

Types of Data :-

(i) Transactional Data
(ii) Multimedia
(iii) Object - Relational Database
(iv) Data - Warehouse
(v) Relational Database
(vi) Flat files
(vii) Spatial Database
(viii) Time Series Database
(ix) WWW.

1. Flat Files: Flat files is defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms. Data stored in flat files have no relationship or path among themselves, like if a relational database is stored on flat file, then there will be no relations between the tables. Flat files are represented by data dictionary. Eg: CSV file. Another example of a flat file is a name-and-address list with the fields Name, Address, and Phone Number. A list of names, addresses, and phone numbers written by hand on a sheet of paper is a flat-file database. Databases created in spreadsheet applications (like Microsoft Excel) are **flat file databases**. An old fashioned example of a flat file or two-dimensional database is the old printed telephone directory. **Application**: Used in DataWarehousing to store data, Used in carrying data to and from server, etc.

2. Relational Databases: A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization. Physical schema in Relational databases is a schema which defines the structure of tables. Logical schema in Relational databases is a schema which defines the relationship among tables. Standard API of relational database is SQL. **Application**: Data Mining, ROLAP model, etc.

3. Data Warehouse: A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision- making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing. There are three types of data warehouse: **Enterprise** data warehouse, **Data Mart** and **Virtual** Warehouse. Two approaches can be used to update data in Data Warehouse: **Query-driven** Approach and **Update-driven** Approach. **Application**: Business decision making, Data mining, etc.

4. Object-Relational Database:

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc.

One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

5. Transactional Databases: A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities. Transactional databases are a collection of data organized by time stamps, date, etc to represent transaction in databases. This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed. Highly flexible system where users can modify information without changing any sensitive information. It follows ACID property of DBMS. **Application**: Banking, Distributed systems, Object databases, etc.

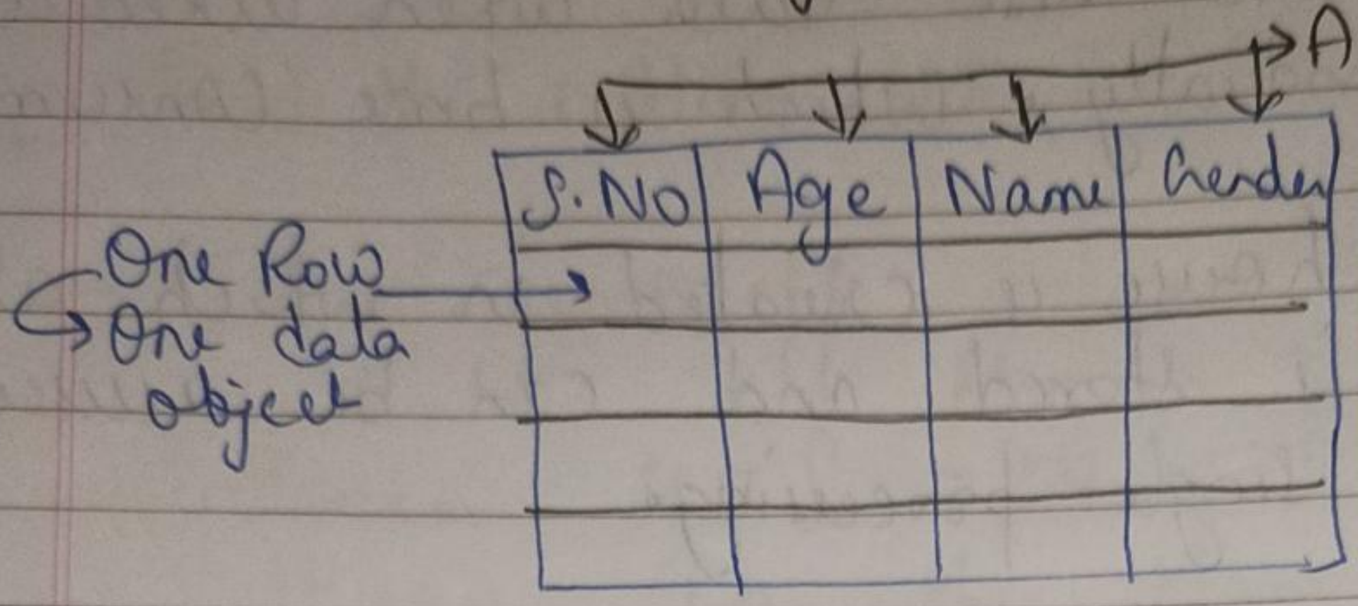
6. Multimedia Databases: Multimedia databases consists audio, video, images and text media. They can be stored on Object-Oriented Databases. They are used to store complex information in pre-specified formats. **Application**: Digital libraries, video-on demand, news-on demand, musical database, etc.

7. Spatial Database: Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning. It stores data in the form of coordinates, topology, lines, polygons, etc. **Application**: Maps, Global positioning, etc.

8. Time-series Databases: Time-series databases contain time related data such as stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time. Handles array of numbers indexed by time, date, etc. It requires real-time analysis. **Application**: eXtremeDB, Graphite, InfluxDB, etc.

9. WWW: WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network. It is the most heterogeneous repository as it collects data from multiple resources. It is dynamic in nature as Volume of data is continuously increasing and changing. **Application**: Online shopping, Job search, Research, studying, etc.

Data Objects & Attributes types :-



Data objects :- vector of attributes may be refferred to as data objects.

Attributes :- Set of attributes used to describe a given object are known as attribute vector or feature vector.

Type of Attributes :-

1. Qualitative :- (Nominal, Ordinal, Binary)
2. Quantitative :- (Numeric, Discrete, Continous)

Nominal attributes :- Related to names. The value of nominal attributes are name of things.
e.g. Colors of hair, eye, etc.

Ordinal attributes :- having some order (related to order).
e.g. Ranks, Grade, etc.

Binary attributes :- True or false

· Assymetric Binary :- True or false having unequal weithage.

Eg. True has more weightage in Medical.

Symmetric Binary :- Both true & false have equal weithage.

Numerical attributes :- it is called quantitative attributes because it is measurable quantity, represented in ~~numb~~ integers or real values.

(i) Interval-Scaled :- has values whose difference is interpretable, but the numerical attribute do not have correct reference point, or we call zero points. Data can be added & subtracted but can't be multiplied or divided.

Eg. Calender dates, temp.

(ii) Ratio-Scaled :- Numeric attribute with fix zero-point. a value is a multiple of another value.

Basic Statistical Desciption of Data

It is used to identify properties of data and highlight which data value should be treated as noise or outlier.

Three area of basic statistical Description :-

(a) Measure of Central Tendency.
(b) Dispersion of data
(c) Graphic display of basic statistical Description

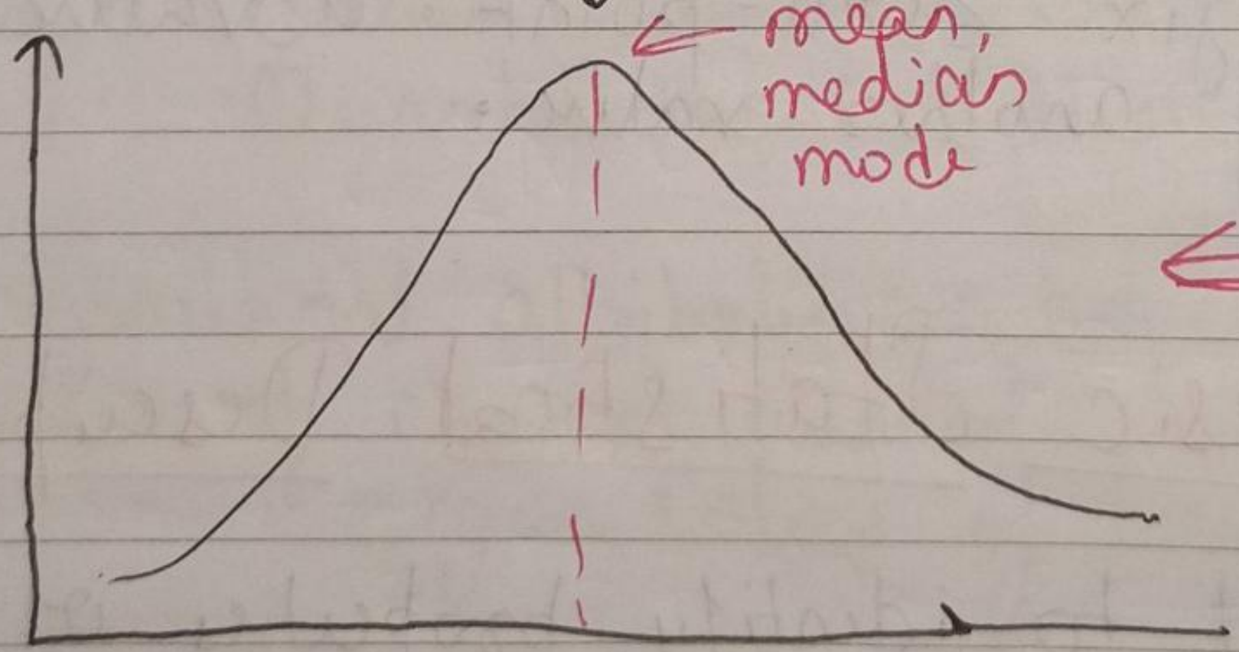
(i) Measure of Central Tendency :-
It measure the location of middle or center of data distribution.

Mean, Median, Mode, Middle Range.

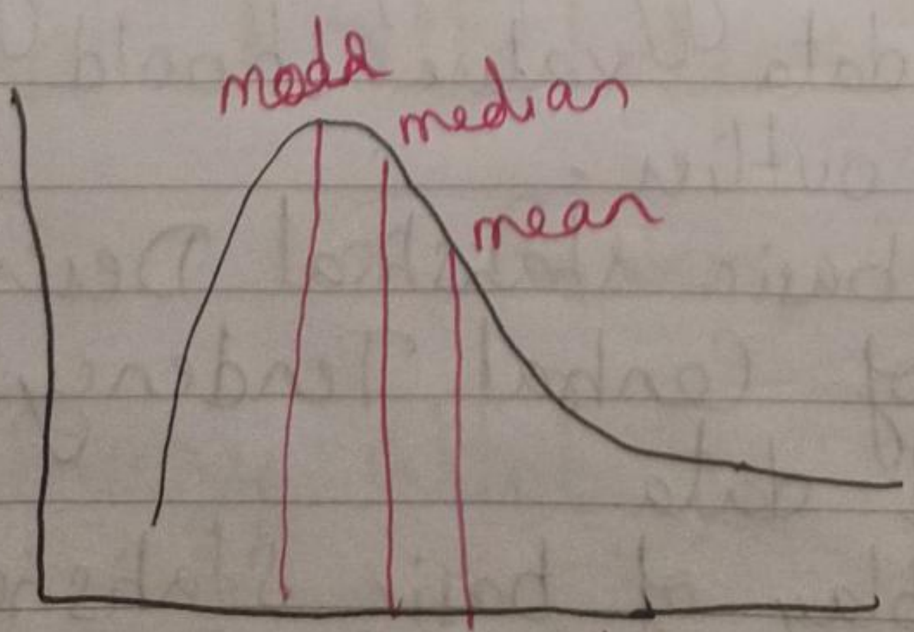
(ii) Dispersion of data :-
Range, Quartiles, Inter-Quartiles Range, Five Number Summary & Boxplot & Variance & S.D.

(iii) Graphic display :-
Bar chart, Pie chart, Line graph. Histogram, Quantile-Quantile plot, Quantile Plots, Scatter plot.

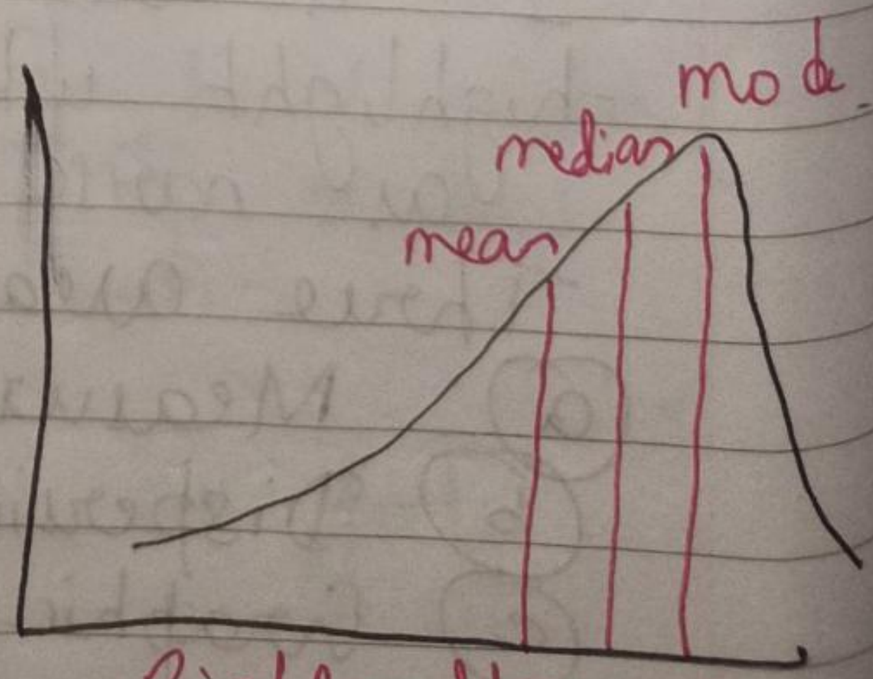
Measure of Central Tendency :-



Ideal situation



Left Skewed



Right-Skewed

Left-skewed :- Mean < Median < Mode

Right-skewed :- Mode < Median < Mean.

Mean → Average of all data.

$$\bar{X} = \frac{x_1 + x_2 + x_3 - - x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Median → First sort in ascending order &
if n is odd, Median = Middle element
if n is even, Median = $\frac{\text{Sum of 2 Middle}}{2}$.

Mode → Highest frqvency.

* Mode can not be determine for Multimodal data. Eg. 20, 30, 40, 50, 60, 70, 80, 10, 20, 30
↳ Mode doesn't exist

Mode can be find from UniModal data.
Eg. 10, 20, 30, 40, 50, 60, 70, 80, 80, 10, 20, 20.
↳ Mode = 20

Standard deviation :- deviation from mean.

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\text{variance}} = \sqrt{\sigma^2} = \sigma.$$

$$SD = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2}$$

Q. Given the data. 30, 36, 47, 50, 52, 52, 60, 63, 70, 70, 110. Compute (a) Mean. (b) Median (c) Mode (d) Variance (e) S.D. Also state whether the data is left skewed or right skewed.

Sol → Data :- 30, 36, 47, 50, 52, 52, 60, 63, 70, 70, 110.

(a) Mean $\Rightarrow \frac{30+36+47+50+52+52+60+63+70+70+110}{11}$

(b) Mean = 58.18

(b) Median = 52

(c) Mode = Doesn't exist

(d) Variance $= \left(\left(\frac{1}{N}\left(\sum_{i=1}^{n}(x_i)^2\right) - \bar{X}^2\right)\right)$

$= \left[\frac{1}{N}\left[41782\right] - 3384.91\right]$

$= \frac{1}{N} \times 38397.09.$

$= 3798.36 - 3384.91$

$= 413.45$ Ans

(e) S.D $= \sqrt{\text{variance}} = \sqrt{413.45}$

$= 20.33$ Ans

(f) Right skewed

* Measuring Data, Similarity & Dissimilarity

S.No	Nominal Data	Ordinal Data	Numeric Data
1	A	Excellent	45
2	B	Fair	22
3	C	Good	64
4	A	Excellent	28
	↑		
	P		

* Similarity → Numerical measure of how alike two data object are. Value is in range (0, 1) and is higher when object are more alike.

* Dissimilarity → Measure of how different two data objects are. It is lower when objects are alike.
* Proximity refers to similarity or dissimilarity.

* Nominal Data :-

No. of nominal attributes, P = 1.

$$d(i,j) = \frac{P - M}{P}$$

M = 0 ← (because it is different)

$$d(1,2) = \frac{1-0}{1} = 1 \text{ (dissimilar)}.$$

$$d(1,3) = \frac{1-0}{1} = 1. \qquad d(1,4) = \frac{1-1}{1} = 0$$

(similar)

$$d(2,3) = \frac{1-0}{1} = 1$$

$$d(2,4) = \frac{1-0}{1} = 1$$

$$d(3,4) = \frac{1-0}{1} = 1$$

On Basis of this we make Dissimilarity Matrix.

	1	2	3	4
1	0			
2	1	0		
3	1	1	0	
4	0	1	1	0

* Ordinal :-

First give rank No to the data.

Excellent → 3
Fair → 1
Good → 2

Now, use Formula $\boxed{\frac{Z - min}{max - 1}}$

1. For 1 :- $\frac{Z - min}{max - 1} = \frac{3-1}{3-1} = 1$

For 2 :- $\frac{1-1}{3-1} = 0$

For 3 :- $\frac{2-1}{3-1} = 0.5$

For 4 :- $\frac{3-1}{3-1} = 1$

So, ordinal data is now converted to numerical.

Excleret	1
Fair	0
Good	0.5
Excellent	1
↑ old	↑ new

* Various Kind of Distances :-

(i) Eucleadian :-

$i = \langle x_{i1}, x_{i2}, x_{i3} \dots x_{in} \rangle$

$j = \langle x_{j1}, x_{j2}, x_{j3} \dots x_{jn} \rangle$

$$\sqrt{(x_{j1}-x_{i1})^2 + (x_{j2}-x_{j2})^2 + \dots (x_{jn}-x_{jn})^2}$$

1	(1, 2)
2	(2, 3)
3	(4, 5)
4	(3, 1)

	1	2	3	4
1	0			
2	$\sqrt{2}$	0		
3	$3\sqrt{2}$	$2\sqrt{2}$	0	
4	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{17}$	0

$$d(1,2) = \sqrt{(2-1)^2 + (3-2)^2}$$

$$= \sqrt{2}$$

$$d(1,3) = \sqrt{(4-1)^2 + (5-2)^2}$$

$$= 3\sqrt{2}$$

$$d(1,4) = \sqrt{(3-1)^2 + (1-2)^2}$$

$$= \sqrt{5}$$

$$d(2,3) = \sqrt{(4-2)^2 + (5-3)^2}$$

$$= 2\sqrt{2}$$

$$d(2,4) = \sqrt{(3-2)^2 + (3+1)^2}$$

$$= \sqrt{5}$$

$$d(3,4) = \sqrt{(3-4)^2 + (1-5)^2}$$

$$= \sqrt{17}$$

(2) Manhaton Distance :-

$$|x_{j1} - x_{j,1}| + |x_{j2} - x_{j2}| + \cdots |x_{jn} - x_{in}|$$

$$\Rightarrow \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ |1|+|1| & 0 & & \\ |3|+|3| & |2|+|2| & 0 & \\ |2|+|-1| & |1|+|-2| & |-1|+|-4| & 0 \end{bmatrix} \end{array}$$

$$\Rightarrow \begin{bmatrix} 0 & & & \\ 2 & 0 & & \\ 6 & 4 & 0 & \\ 3 & 3 & 5 & 0 \end{bmatrix}$$

(3) Supremum Distance :-

$$\max\left((x_{j1} - x_{j1}), (x_{j2} - x_{i2})\right).$$

$$d(1,2) = \max(2-1, 3-2)$$
$$= \max(1,1)$$
$$= 1$$

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ \max(3,3) & \max(2,2) & 0 & \\ \max(2,1) & \max(1,2) & \max(1,4) & 0 \end{bmatrix} \end{array}$$

$$\Rightarrow \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 3 & 2 & 0 & \\ 2 & 2 & 4 & 0 \end{bmatrix}$$

* Minkwoski Generalise Equation :-

$$\text{distance} = \sqrt[k]{|x_{j1}-x_{j1}|^k + |x_{j2}-x_{i2}|^k + \cdots + |x_{jn}-x_{in}|^k}$$

Here, when $k = 2$
$\Rightarrow$ Eucleadean

$k = 1$
$\Rightarrow$ Manhaton.

* Weighted Eucleadean distance :-

$< w_1, w_2, w_3 \; \cdots \; w_n >$

$$\text{distance} = \sqrt[2]{w_i |x_{j1}-x_{i1}|^2 + w_2 |x_{j2}-x_{i2}|^2 + \cdots + w_n |x_{jn}-x_{in}|^2}$$

* Binary Data :-

For Binary data we use Contigency Matrix.

	1	0	Sum
1	q	r	q+r
0	s	t	s+t
Sum	q+s	r+t	q+r+s+t = p

$$d(i,j) = \frac{r+s}{p=(q+r+s+t)}$$

for Symmetric Binary

$$d(i,j) = \frac{r+s}{q+r+s}$$

for asymmetric Binary. In this t4 not worthy.

Name	Gender	Fever	Cough	Test 1	Test 2	Test 3	Test 4
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N

$d(Jack, Jim) =$

Jack \ Jim	1	0
1	1	1
0	1	3

No. of Value where Jack & Jim both is True or 1

No. of values where Jack is 1 & Jim is 0.

where both is 0

where Jack is 0, Jim is 1

$$d(Jack, Jim) = \frac{1+1}{3} = \frac{2}{3} = 0.67$$

By using formula $\frac{r+s}{q+r+s}$ $(1+1)$ $(1+1+1)$

$d(Jack, Marry)$

Jack \ Marry	1	0
1	2	0
0	1	3

$$d(Jack, Marry) = \frac{1+0}{2+1+0} = \frac{1}{3}$$
$$= 0.33.$$

$d(Jim, Marry) =$

Jim \ Marry	1	0
1	1	1
0	2	2

$$d(Jim, marry) = \frac{1+2}{1+1+2}$$
$$= \frac{3}{4} = 0.75$$

* $Similarity(i,j) = 1 - Dis(i,j)$.

* Cosine Similarity :-

Text 1 :- Ram is a good Boy

Text 2 :- Ramesh is a clever Boy.

$\Rightarrow$ Let X be Text 1 , & Y = Text 2.

Ram is a good Boy Ramesh cleves.

X(1, 1, 1, 1, 1, 0, 0). ← frequency of each word in X

Y(0, 1, 1, 0, 1, 1, 1) ← in Y.

$$\boxed{Sim(x,y) = \frac{xy}{\|x\|\,\|y\|}}$$

$$= \frac{1\times0 + 1\times1 + 1\times1 + 1\times0 + 1\times1 + 0\times1 + 0\times1}{\sqrt{1^2+1^2+1^2+1^2+1^2+0^2+0^2}\ \sqrt{0^2+1^2+1^2+0^2+1^2+1^2+1^2}}$$

$$= \frac{3}{\sqrt{5}\sqrt{5}} = \frac{3}{5} = 0.6$$

$$\Rightarrow \|u\| = (x_1, x_2, x_3, x_4, x_5 \ldots x_n) = \sqrt{x_1^2 + x_2^2 + \ldots x_n^2}$$

Q Compute the cosine similarity of

X(5, 0, 3, 0, 2, 0, 0, 2, 0, 0)

Y(3, 0, 2, 0, 1, 1, 0, 1, 0, 1).

Sol → $\frac{xy}{\|x\|\,\|y\|}$ =

$$= \frac{5\times3 + 0\times0 + 3\times2 + 0\times0 + 2\times1 + 0\times1 + 0\times0 + 2\times1 + 0\times0 + 0\times1}{\sqrt{5^2+0^2+3^2+0^2+2^2+0^2+0^2+2^2+0^2+0^2}\ \sqrt{3^2+2^2+1^2+1^2+1^2+1^2}}$$

$$= \frac{15+6+2+2}{\sqrt{42}\ \sqrt{17}} = \frac{25}{\sqrt{42}\ \sqrt{17}}$$

$$= 0.935$$

$$\approx 0.94 \text{ (approx)}$$

* Tanimoto Coefficient

$$= \frac{X \cdot Y}{X \cdot X + Y \cdot Y - X \cdot Y}$$

Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.

Quality of Data :-

(i) Accuracy
(ii) Completeness
(iii) Consistency

(iv) Timeliness
(v) Believability
(vi) Interpretability.

Steps involved in Data preprocessing :-

1) Data cleaning :-
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done.

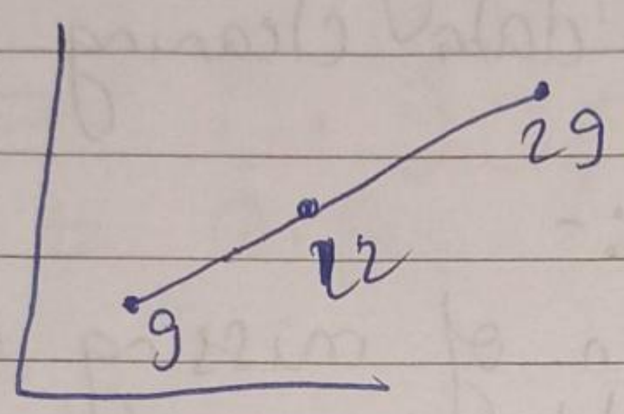
(a) Missing data :-
Possible solution of missing data :-
(i) Delete the tuples
(ii) Filling missing data ~~man~~ manually.
(iii) Central tendency of data.
(iv) Fixed value
(v) Mean/Median of that particular class
(vi) Use regression b

(b) Noisy Data :- (Random error in data) :-
Methods :-
(i) Binning :-
→ Sort the data
→ Creat Bin with equal size.
1. Bin with mean
2. Bin with boundary values.

E.g. 4, 8, 15, 21, 21, 24, 25, 28, 34.

	Mean	Binned
1. B_1 : 4, 8, 15	9	9, 9, 9.
B_2 : 21, 21, 24	22	22, 22, 22
B_3 : 25, 28, 34.	29	29, 29, 29.

2. B_1 (4, 15) → 4, 4, 15.
B_2 (21, 24) → 21, 21, 24.
B_3 (25, 34) → 25, 25, 34.



(ii) Regression analysis :-

(iii) Clustering (Outlier analysis) :-

Data cleaning as a Process :-

→ Data about Data → Meta data
→ Field overload → offering data format
↳ 31/08/2008
2008/03/31

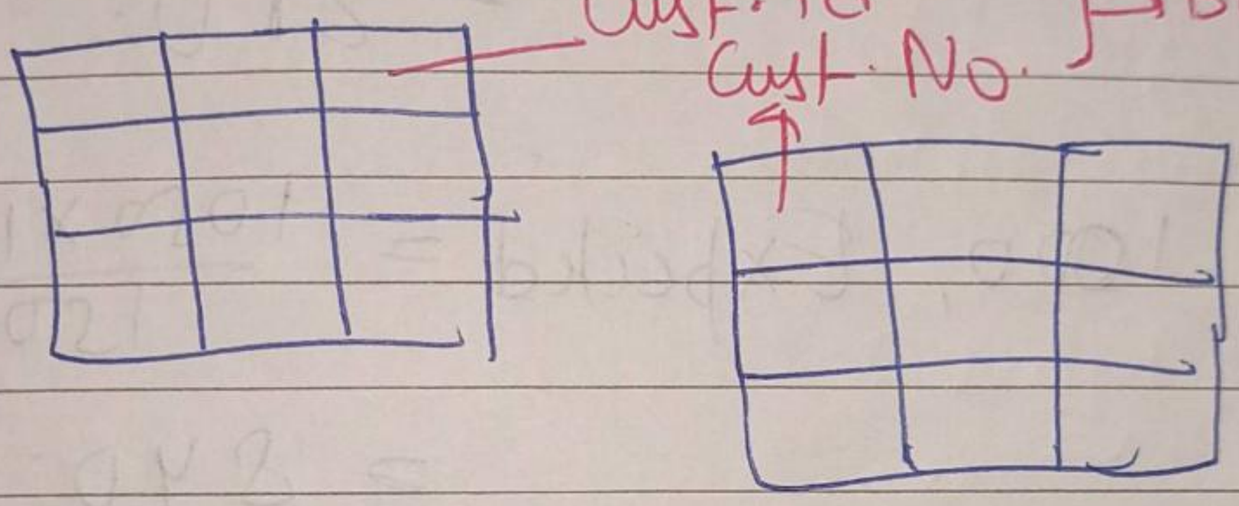
→ Unique data should be specified
→ Database should be not NULL.

Data Integration :-

Data integration merges data from several heterogeneous sources to attain meaningful data.

While integrating the data, we have to deal with several issues as discussed below :-

(a) Unique Entity Identification problem :-



(b) Redudancy :- Redudancy is one of the issues during data integration. Redundant data is an unimportant data that is no needed now. It can also arise due to attributes that could be derived using another attributes in data set.

* The redudancy can be detected using correlation analysis.

Two method for correlation analysis :-

(a) x^2 Test (Chi-Square test) :-

	Fiction	Non fiction	
Male	250	200	450
female	50	1000	1050
	300	1200	1500

$$\text{Expected} = \frac{\text{Count}(a_i) * \text{Count}(b_i)}{\text{Total}}.$$

For 250, $\text{Expected} = \frac{300 \times 450}{1500}$

$= 90$

For 50, $\text{Expected} = \frac{300 \times 1050}{1500}$

$= 210.$

For 1000, $\text{Expected} = \frac{1050 \times 1200}{1500}$

$= 840.$

for 200 $\text{Expected} = \frac{450 \times 1200}{1500} = 360$

$$\sum_{i=1}^{n} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

$= 284.44 + 71.11 + 121.90 + 30.47$

$= 507.92.$

Now, then check the value with x^2 table for $\lambda = (r-1) \times (c-1)$.

(b) Karlpearian Coefficient :-

$$V_{ab} = \frac{\sum_{i=1}^{n} a_i b_i - n\bar{A}\bar{B}}{n\sigma_A \sigma_B} = \frac{\text{Covariance}(A, B)}{\sigma_A \sigma_B}$$

	A	B	$A_i - \bar{A}$	$B_i - \bar{B}$	$(A_i - \bar{A})(B_i - \bar{B})$
t_1	6	20	2	9.2	18.4
t_2	5	10	1	-0.8	-0.8
t_3	4	14	0	3.2	0
t_4	3	5	-1	-5.8	5.8
t_5	2	5	2	-5.8	11.6
	$\bar{A} = 4$	$\bar{B} = 10.8$			35

$$Cov(a,b) = \frac{\sum_{i=1}^{n} (A_i - \bar{A}) \times (B_i - \bar{B})}{n}$$

$$= \frac{35}{n} = \frac{35}{5} = 7.$$

Then find V_{ab} by Covariance formula.

Data Reduction :-

Data Reduction is the preprocessing techniques that helps in obtaining reduced representation of dataset from the available dataset.

Data Reduction Techiques :-

(i) Dimensionality Reduction :-

It is the process of reducing the number of dimension the data is spread across. It mean the attributes or features that the data set carries as the number of dimensions increases the sparsity. This sparsity is critical to clustering, outlier analysis and other algorithms. With reduced dimensionality, it is easy to visualize and manipulate data.

Types :-

(a) PCA :- PCA stands for Principal Component Analysis. It involves identification of a few independent tuples with 'n' attributes that can represent the entire data set. This can be applied to skewed or sparse data.

Basically, in this, if the data is in K dimension, then this method searches for C dimension which can be used to represent the data where C ≤ K. The original data is thus projected on much smaller place.

Steps :-

1) Normalize the data so that it will fall within same range.

2) PCA computes orthogonal vectors that provide a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to each other. These vector are referred to as principle Components.

3) The Principle Components are sorted according to decreasing order of significance or strength.

4) Since data is sorted according to decreasing order of significance, eliminating the weaker component can reduce the size of data.

(b) Attribute Subset Selection :-
In this, attribute irrelevant to data mining or redundant one are not included in a core attribute subset.

Brute force approach is very expensive in which each subset (2^n) of the data having n attributes can be analysied. The best way to do the task is to use the statistical significance tests such that best (or worst) attributes can be recognized. Statistical significance test assumes that attributes are independent of one another. This is a kind of greedy approach in which a significance level is delided and module is tested again & again until p-value of all attributes is less than or equal to selected Significance level. Attribute having higher p-value is discarded.

Methods of Attributes Subset Selection :-

(a) Stepwise Forward Selection :-
Empty set is choosen and selected attribute is added.

(b) Stepwise Backward ~~Selection~~ Elimination :-
All attributes is selected and is Eliminated one-by one.

(c) Combination of Forward & Backward :-
It is combination to select the relevant attribute most efficintly.

(d) Decision Tree Induction :-
It uses decision tree for attribute selection. It consist of flow chart like structure having nodes denoting a test on an attribute.

(ii) Numerosity Reduction :-
This method uses alternat, small form of data representation thus reducing data volume.

Parametric

This method assumes a model into which the data fits. Data model parameters are estimated, and only those parameters are stored, and the rest of the data is discarded. For example, a regression model can be used to achieve parametric reduction if the data fits the Linear Regression model.

Linear Regression models a linear relationship between two attributes of the data set. Let's say we need to fit a linear regression model between two attributes, x and y, where y is the dependent attribute, and x is the independent attribute or predictor attribute. The model can be represented by the equation y=wx + b, Where w and b are regression coefficients. A multiple linear regression model lets us express the attribute y in terms of multiple predictor attributes.

Another method, the Log-Linear model discovers the relationship between two or more discrete attributes. Assume, we have a set of tuples in n-dimensional space; the log-linear model helps to derive the probability of each tuple in this n-dimensional space.

Non-Parametric

A non-parametric numerosity reduction technique does not assume any model. The non-Parametric technique results in a more uniform reduction, irrespective of data size, but it may not achieve a high volume of data reduction like the parametric one. There are at least four types of Non-Parametric data reduction techniques: Histogram, Clustering, Sampling, Data Cube Aggregation, Data Compression.

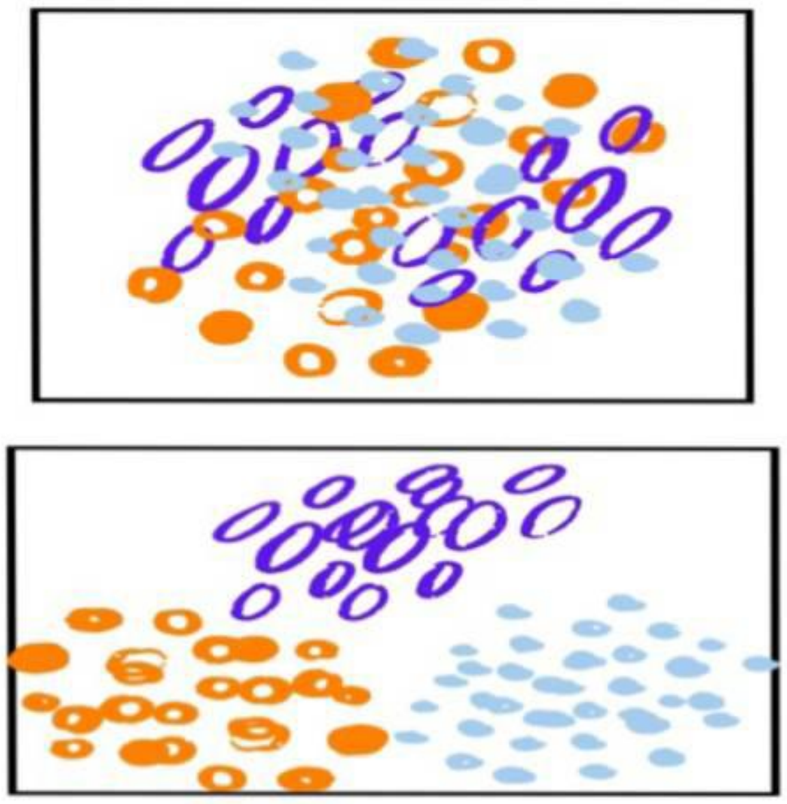
(i) Histogram

A histogram can be used to represent dense, sparse, skewed or uniform data, involving multiple attributes, effectively up to 5 together.

(ii) Clustering

In Clustering, the data set is replaced by the cluster representation, where the data is split between clusters depending on similarities to each other within-cluster and dissimilarities to other clusters. The more the similarity within-cluster, the closer they appear within the cluster.

The quality of the cluster depends on the maximum distance between any two data items in the cluster.



(iii) Sampling

Sampling is capable of reducing large data set into smaller sample data sets, reducing it to a representation of the original data set. There are four types of sampling data reduction methods.

- Simple Random Sample without Replacement of size s
- Simple Random Sample with Replacement of size s
- Cluster Sample
- Stratified Sample

(iv) Data Cube Aggregation

Data Cube Aggregation is a multidimensional aggregation that uses aggregation at various levels of a data cube to represent the original data set, thus achieving data reduction. Data Cube

Aggregation, where the data cube is a much more efficient way of storing data, thus achieving data reduction, besides faster aggregation operations.

(v) Data Compression

It employs modification, encoding or converting the structure of data in a way that consumes less space. Data compression involves building a compact representation of information by removing redundancy and representing data in binary form. Data that can be restored successfully from its compressed form is called Lossless compression while the opposite where it is not possible to restore the original form from the compressed form is Lossy compression.

Data reduction achieves a reduction in volume, making it easy to represent and run data through advanced analytical algorithms. Data reduction also helps in the reduplication of data reducing the load on storage and the algorithms serving data science techniques downstream. It can be achieved in two principal ways. One by reducing the number of data records, or the features and the other by generating summary data and statistics at different levels.

Data Transformation :-

Data transformation in data mining is done for combining unstructured data with structured data to analyze it later.

1. Smoothing :- Proceed to remove noise.
2. Attribute Creation :- New attribute is created from old one.
3. Aggregation
4. Normalization
5. Concept hirarchy.
6. Discreatization.

→ Methods :-

(i) Min-Max Normalization :-

$$v' = \frac{v_i - min}{max - min}(new\ max - new\ min) + new\ min$$

E.g. Data → 10, 20, 200, 500, 30, 40, 60.
Normalize it under (0, 1).

⇒ Min = 10
Max = 500

$$V_1 = 0 + \frac{10-10}{500-10} * (1-0)$$

$$= 0$$

$$V_2 = 0 + \frac{20-10}{500-10}(1) = 0.204.$$

$$V_3 = 0 + \frac{200-10}{500-10} \times 1$$

$$= 0.39$$

$$V_4 = 0 + \frac{560-10}{50-10} \times 1$$

$$= 1$$

$$V_5 = 0 + \frac{30-10}{500-10} \times 1 = 0.408$$

$$V_6 = 0 + \frac{400-10}{500-10} \times 1 = 0.7959$$

$$V_7 = 0 + \frac{50-10}{500-10} \times 1 = 0.10204.$$

(ii) Decimal Scaling Normalization :-

$$V_i' = \frac{V_i}{10^j}$$

j is smallest integer such that $\max(|V_i'|) < 1$

Data	
10	0.001
20	0.002
200	0.02
-10	-0.001
-20	-0.002
-30	-0.003
-1000	-0.1

$$\left|\frac{-1000}{10^j}\right| \leq 1$$

$j = 4$

(iii) Z-Score Normalization :-

$$V_i' = \frac{V_i - \bar{A}}{\sigma_A}$$

E.g

A	$A - \bar{A}$	$(A - \bar{A})^2$	$\frac{V_i - \bar{A}}{\sigma_A}$
1	-3	9	-1.5
2	-2	4	-1
3	-1	1	-0.5
4	0	0	0
5	1	1	0.5
6	2	4	1
7	3	9	1.5
$\bar{A} = 4$		28	

$$\sigma = \sqrt{\frac{28}{7}} \Rightarrow \sqrt{4} = 2$$

Mean = 0 in the values of $\frac{V_i - \bar{A}}{\sigma_A}$. This is the speciality of this method

Q Suppose the mean & S.D of values for the attribute income is 54000 & 16000 respectively. Compute Z-Score normalization for value of 73600.

Sol → $\frac{73600 - 54000}{16000} = 1.225$ Ans

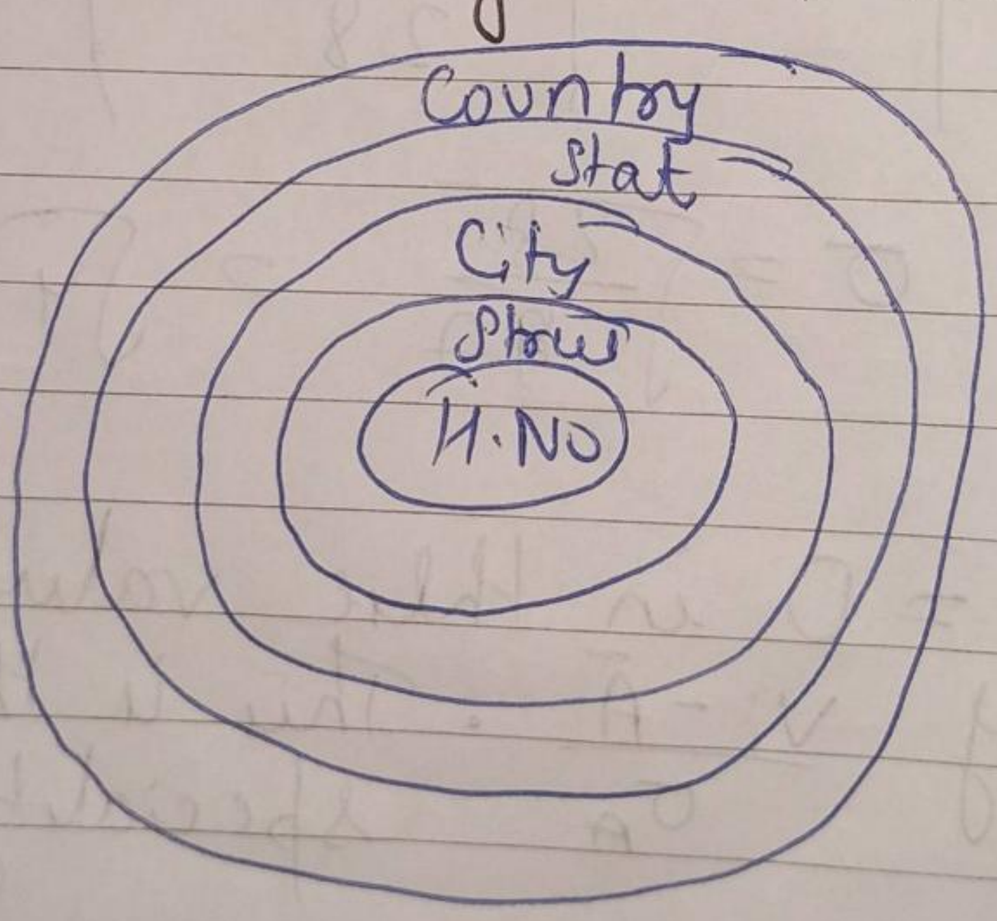
Range → min-max
(-1, 1) → decimal
(Mean & S.D) → Z-Score

→ When to use which one.

Concept hierarchy :-

Only for Nominal Data.

(i) Total ordering or Partial ordering.



(ii) For large database :- Not include all attributes just take certain attributes

{Delhi, Lahore, Dhaka} ⊆ {India, Pakistan, Bangladesh}

(iii) Count unique values of all attributes.

40000 → H.No
10000 → City Stat
1000 → City
65 → Stat
15 → Country

Country
↓
Stat
↓
City
↓
City Stat
↓
H.No.

(iv) Database should define associative relation such that other details comes out automatically.