CHAPTER 5

UNIT V

Big Data Visualization

Syllabus Topics

Introduction to Data visualization, Challenges to Big data visualization, Conventional data visualization tools, Techniques for visual data representations, Types of data visualization, Visualizing Big Data, Tools used in data visualization, Analytical techniques used in Big data visualization

Syllabus Topic : Introduction to Data Visualization

5.1 Introduction to Data Visualization

Q. 5.1.1 What is data visualization ? (Refer section 5.1) **(2 Marks)**

DEFINITION

Data visualization *has two terms; data means information and visualization means pictorial representation or graphical representation ; so the data visualization term is defined as the pictorial representation of some information so that the user can analyze the data quickly.*

- Nowadays, many organizations consider the **Data visualization** as a modern equivalent of visual communication.
- The real fact is that, processing the graphically represented data is easier than the data represented using spread sheets or similar tools; and that's why, data visualization is used to convey the information to the users via interactive visual representation of abstract data so that the internal structure of data and their relationship is known by users quickly.
- Data finding with Visualization-based method provides the ability to merge different data from different sources to make various customized analytical views.
- According to the Friedman (2008), main goal of data visualization is to communicate information clearly and effectively through graphical means.
- A primary objective of data visualization is to communicate information in clear and efficient manner to the users through graphical representation such as plots and statistical and / or information graphics.

☞ Objectives of data visualization

- Other objectives behind the use of data visualization are :
  - To enlighten the data or see the data in context.
  - To solve and give solution of a specific problem.
  - For understanding the data more clearly, explore that data so that it will help to make proper decisions.
  - To illustrate or hide the data.
  - To find patterns or relationship among the data.
  - To make comparison between some statistical data or predict outcomes.
- In the data visualization if data is in the numerical form, then it is represented using bars, lines, or dots.

- Effective data visualization helps to make complex data more accessible, understandable and usable. In data visualization the Tables are used to keep a specific measurement, whereas the various types of charts are used to demonstrate various patterns or relationships among variables in the visualized data.
- Data visualization is the technique which is used to communicate data or information by representing the data and information using visual graphic objects like points, lines or bars. One of the steps in data analysis or data science is data visualization.

Guiding points for visualization

- Following points are used as guidance for visualization :
  - The metadata i.e. data about data can be more informative.
  - Participation of Interactive Visualization tools and user involvement is more essential.
  - Interactive visualization tools are more efficient to discover some information instead of using static data tools.

Syllabus Topic : Challenges to Big Data Visualization

5.2 Challenges to Big Data Visualization

Q. 5.2.1 What are the challenges and their possible solutions in Big data visualization ? (Refer section 5.2) **(4 Marks)**

- Before learning the challenges in big data visualization let's know about the 5Vs of big data which causes big data challenges.

5 V's of Big data

- The 5Vs of big data are volume, variety, velocity, value, and veracity. These are the characteristics of big data. These 5Vs defines the dimensionality of big data.

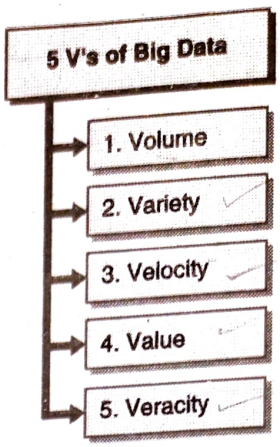

Fig. 5.2.1 : 5 V's of Data

➜ 1. Volume

- Refers to the measurable amount of data (specifically machine-generated data).
- This characteristic of Big data defines size of the data set which creates problem in its storage and analysis while using the traditional database technology.

➜ 2. Variety

- Refers to the various types and forms of data used in an organization.
- These characteristics can categorize the data into Structured, semi-structured, or unstructured data format. The financial data is of structured type data and unstructured type data includes social media chatting, images, videos, voice recordings, etc.
- Handling huge variety of data / information becomes more difficult.

➜ 3. Velocity

- Refers to the processing speed of that data or speed of generating and distributing new data over the organization.
- This characteristic leads to implement real-time processing for the streaming data analysis on social media, different types of transactions or trading systems, etc.

➜ 4. Value

Refers to generating the status of an organization / business by analyzing the big data.

5. Veracity

Refers to the complexity of data due to which quality and accuracy of data is decreased. Veracity in data causes uncertainty which causes the omitting or skipping the valuable data.

- As these dimensions increases, all the aspects of big data become challenging; it also hamper effectively to the data visualization.
- Visual analytics has two challenges as :
  - Scalability
  - Dynamics
- The visualization-based methods accept the challenges presented by the "5Vs" of big data and turn them into following opportunities :

1. **Volume :** There are some methods which are developed to work with huge number of datasets and obtain meaningful information from huge amount of data.
2. **Variety :** There are some methods which are developed to join as many data sources as required.
3. **Velocity :** There are some methods which are developed to replace batch processing with real-time stream processing.
4. **Value :** There are some methods which are not only developed to enable users for producing attractive info-graphics and heat-maps, but also produces business value by analyzing the big data.
5. **Veracity :** There are some methods which are developed to avoid decision making depending on analysis of uncertain and rough big data, by assigning a veracity grade or veracity score for particular datasets.

- Diversity and heterogeneity in big data creates a big problem while visualizing that data.
- Analysis Speed is the most preferred factor in the big data analysis.
- For handling the big data scalability the Cloud computing and advanced GUI are combined with the big data.
- Usually the big data is in unstructured format. And to visualize the unstructured data, tables, texts, trees, graphs, and other metadata is used.
- The visualization of big data is done as possible as closer to the data so as to efficiently extract meaningful information.
- As the size of big data is vast, providing huge parallelization is a challenge in big data visualization.
- The required parallelization can be achieved through using parallel visualization algorithm; it is also challenging to decompose the problem into several independent tasks which can execute concurrently.
- For discovery process in big data, the effective data visualization is needed. As the discovery process introduces few challenges like high complexity and high dimensionality due to the amount of data.
- Effective data visualization handles these challenges by using different dimensionality reduction methods. Sometimes these methods can be applicable or can't be applicable.
- As the dimensions are effectively visualized, the probability of identifying possible patterns, correlations, or outliers is higher.
- Other challenges of big data visualization are: Perceptive and Interactive Scalability.
- As the size of big data is vast, visualizing every data point leads to over-plotting; or reducing the data via sampling or filtering which leads to elimination of some valuable structures, objects, or outliers.
- The problems like, High latency and disruption in interaction are resulted by querying large data sets.
- It is difficult to design new big data visualization tool which results efficiency.
- Due to the large size and dimensions of big data the visualization becomes more challenging.
- There are several Big Data visualization tools which are currently in use which may results poor scalability, functionalities, and response time.

Some other problems occur in big data visualization

- Instead of above problems the big data visualization also has some other problems as given below :

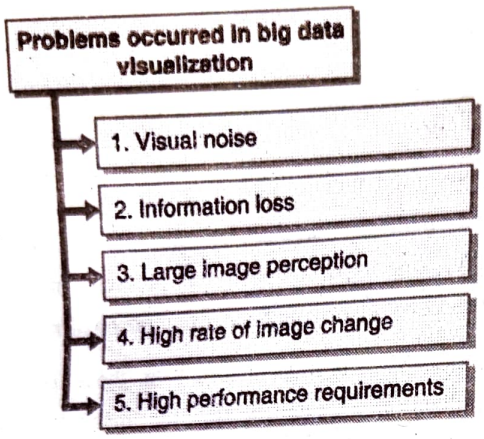

Fig. 5.2.2 : Problems occurred in big data visualization

➜ 1. Visual noise

A data set can have the objects which are too relative to each other. So users unable to separate them on the screen while visualizing.

➜ 2. Information loss

It is possible to perform reduction on the visible data sets; but there may be information loss while doing so.

➜ 3. Large image perception

The aspect ratio, device resolution, and physical perception limits the Data visualization methods.

➜ 4. High rate of image change

Users can only observe data and can't respond to the amount of data change or image intensity on display.

➜ 5. High performance requirements

It can be hardly noticed in static visualization because of lower visualization speed requirements.

Solutions for big data visualization problems

- Possible solutions for the big data visualization challenges or problems we have discussed yet are as follows :

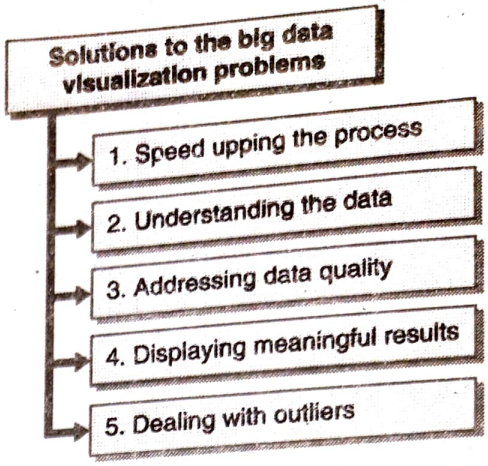

Fig. 5.2.3 : Solutions to the big data visualization problems

➜ 1. Speed upping the process

- By using faster hardware we can speed up the analysis process.
- By increasing the memory and using powerful parallel processing we can increase the speed.
- By using grid approach where number of machines are used but storing the data in-memory.

➜ 2. Understanding the data

Taking help of proper domain expertise to understand it correctly.

➜ 3. Addressing data quality

Assuring the data quality by using the data governance or information management process.

➜ 4. Displaying meaningful results

Effective visualization of data is done by clustering the data into a higher-level view, where the smaller groups of data are visible.

➜ 5. Dealing with outliers

Removing the outliers from the data or create a separate chart for those outliers are the possible solutions to deal with them.

Syllabus Topic : Conventional Data Visualization Tools

5.3 Conventional Data Visualization Tools

Q. 5.3.1 Explain interactive visualization steps. (Refer section 5.3) (2 Marks)

- Here the conventional means the methods and ideas used by organization for visualizing the data.
- Even though the new technologies are invented, developed and, deployed to provide newer facilities of data visualization process, it is necessary to understand first the basics of data visualization and how to choose proper visualization method which is most effective among several data visualization methods.
- In previous section we have learned about the basics of data visualization now let's know the points based on which selection of most effective data visualization method is done.

Selection points on which interactive visualization takes place

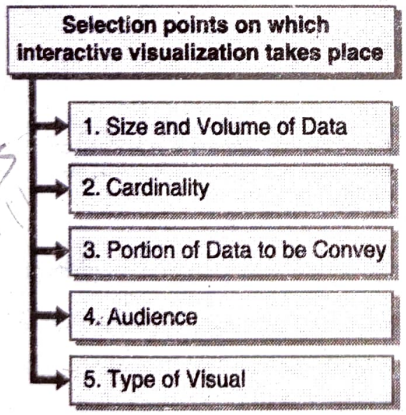

Fig. 5.3.1 : Selection points on which interactive visualization takes place

1. Size and volume of data

To make a perfect choice, the size and volume of data should be visualized.

2. Cardinality

The cardinality and the context of data should also be visualized.

3. Portion of Data to be Convey

Visualizing the point or portion of data which user wants to convey.

4. Audience

To whom the user wants to convey the portion of data / information.

5. Type of visual

- Which type of visual should convey that point to the audience correctly and very quickly.
- To make the appropriate choice is depending on the trial and error basis.
- Visualizations can be static as well as interactive.

Interactive visualization approaches

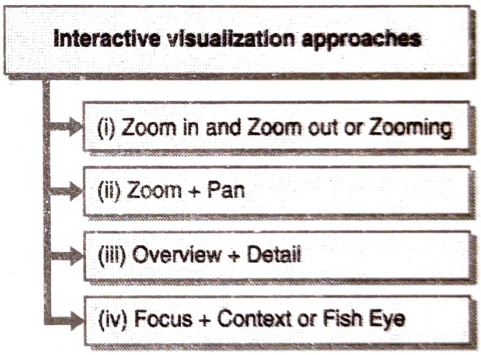

Fig. 5.3.2 : Interactive visualization approaches

(i) Zoom in and Zoom out or Zooming

- To display the contents more clearly and in detail, the user interfaces brings the content closer.
- In computing, ZUI (Zooming User Interface or Zoomable User Interface) allows users to change the scale of the interfaces area according to the user's choice. When the user wants to see the contents in maximized size for detail view then zoom in technique is used, and when the user wants to make the contents size smaller then zoom out technique is used.
- Zoom out technique is useful for mobile device because the screen size of mobile device is limited as compare to the area required for displaying the content.

➜ **(ii) Zoom + Pan**

- Initially the Zoomable visual interface started with the overview and then allows users to use zoom in technique on some part of the interface to view corresponding information in detail.
- The Zoomed visual interface can be pan over the remaining interface area without zooming out the interface.
- For returning to the overview the zoom in , zoom out techniques are applied on the interface alternatively, and to focus on the another portion of information in detail again zoom in technique is used on the interface area including the information to be focused, and again pan zoom in state and pan in zoom out state.
- It has two advantages :
  - Helps to efficiently use the screen
  - Provides unlimited scalability.
- It has two disadvantages
  - While zoom in it loses the overview
  - Slower navigation.

➜ **(iii) Overview + Detail**

- Multiple views (overview, and detail view) are used simultaneously by this technique.
- Such technique is used in maps.
- This interactive technique even though keeps overview of the interface to avoid confusion in the detail view, but suffers with the visual discontinuity among overview and detail view.
- This technique has advantages as:
  - The overview is stable and presents multiple overviews (i.e. chained views or scalable views).
  - While visual is separate between views as back and forth views.
  - The views compete for screen space.

➜ **(iv) Focus + Context or Fish Eye**

- This technique expands the area to be focused directly within the overall view or overview context.
- The focused area represents the details about that part of information space.
- Also the user can navigates easily across the overview just by sliding the focus across the overview to know information about other parts in detail.
- To keep context and the space for focused region in the overview while expanding that on interface, the rest part of overview is partially pushed back by warping the overview. Therefore, this technique is also called as fisheye view.
- This technique has the main advantage as :

  While focusing some region the technique keeps its context, which means that it keeps view of non-focus area.
- But the distortion and unstable overview are the main disadvantages of this technique.

☞ **Steps used to perform interactive visualization**

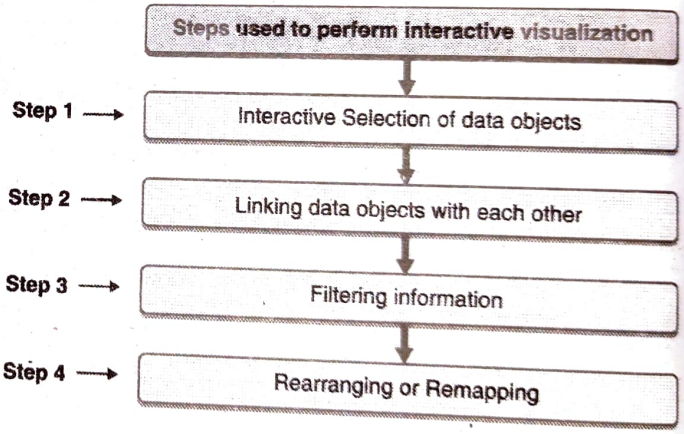

Fig. 5.3.3

Step 1 : Interactive Selection of data objects

According to the user data entities or subset or part of whole data or whole data sets are selected for visualization.

Step 2 : Linking data objects with each other

Linking is used for connecting multiple views to relate the information presented by them.

Step 3 : Filtering information

By filtering the information only valuable data is focused and unrelated or valueless data is removed to decrease the amount of information.

Step 4 : Rearranging or Remapping

- In the visual mapping, the spatial layout is essential factor so while creating different insights the rearrangement of the spatial layout of the information is very effective.
- Due to the Web-based linking technologies, as the data changes the visualization also changes. This will reduce efforts of keeping visualizations up to date.
- Web-based tools are able to do live linking whereas most of the visualization tools used by scientists can't do live linking.

Syllabus Topic : Techniques for Visual Data Representations

5.4 Techniques for Visual Data Representations

Q. 5.4.1 What are the visual data representation techniques ? (Refer section 5.4) **(4 Marks)**

According to different authors the categorization of Visualization techniques or methods is different. As mentioned in periodic table of visualization there are six main categories as given below :

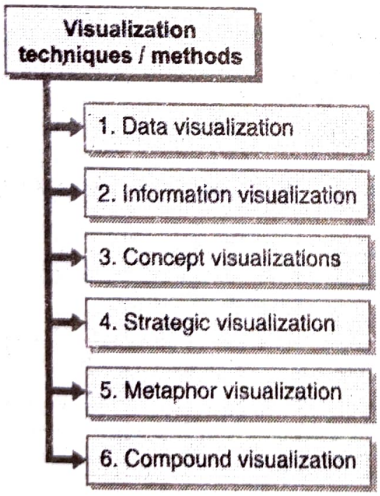

Fig. 5.4.1 : Visualization techniques/methods

➜ **1. Data visualization**

This type of visualization helps to represents quantitative data with or without axes visually in schematic or diagrammatic forms for example, Table, Line chart, Pie chart, Histogram, and Scatter plot etc.

➜ **2. Information visualization**

Information visualization facilitates interactivity in the data to increase cognition or perception ability of it. In the visualization the data is transformed into a changeable image (for example: Data map, Tree map, Clustering, Semantic network, Time line, and Venn/ Euler diagram, etc) which are used by users for interaction during the data manipulation.

➜ **3. Concept visualization**

This visualization is a method which is used to explain the ideas, plans, concepts in detail and analyzing them easily with the help of Mindmap, Layer chart, Concentric circle, Decision tree, Pert chart etc.

➜ **4. Strategic visualization**

The strategic visualization is used to represent the organization's strategies of development, formulation, communication, implementation, and some time its analysis visually with the help of Organizational chart, Strategy map, Failure tree, and Portfolio diagram etc. It is a systematic approach used by organizations.

➜ **5. Metaphor visualization**

The next category of visualization is Metaphor visualization which Organizes and structures information graphically. This visualization helps to express insight of information using metaphor characteristics like Metro map, Story template, Funnel, and Tree etc.

➜ **6. Compound visualization**

- Compound visualization allows merging the different graphic representation formats in one single schema or frame. The examples of compound visualization are as : Cartoon, Rich picture, Knowledge map, and Learning map etc.
- In this chapter we are focusing on two visualization categories Data Visualization and Information Visualization.
- These methods are essential to analyse and represent the massive data in meaningful form, so that it will easily understand and interpretable.

Data visualization techniques

There are some **data visualization techniques** like :

1. Table
2. Pie Chart
3. Bar Chart
4. Histogram
5. Line Chart
6. Area Chart
7. Scatter Plot
8. Bubble Chart
9. Multiple Data Series

Information visualization techniques

There are some **Information visualization techniques** like : .

1. Parallel Coordinates
2. Tree Map
3. Entity Relationship Diagram
4. Cone Tree
5. Time Line
6. Flow Chart
7. Data Flow Diagram
8. Venn Diagram
9. Semantic Network

Syllabus Topic : Types of Data Visualization

5.5 Types of Data Visualization

Q. 5.5.1 Explain types of data visualization. (Refer section 5.5) **(2 Marks)**

Most used data visualization methods / types are:

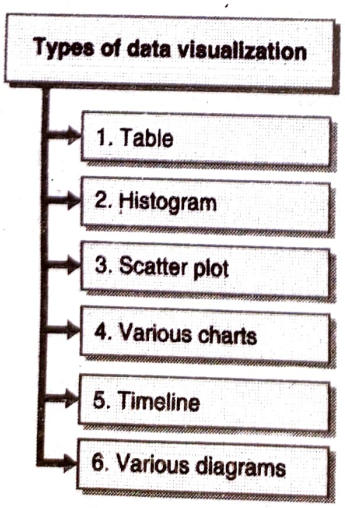

Fig. 5.5.1 : Types of data visualization

1. Table

- The most frequently used data representation technique is Table because it is simple, easy to understand, and easy to interpret.
- Collection of rows and columns is together referred as a table which represents the data into structured format.
- The rows are also known as tuples, records, vectors, etc. whereas the columns are known as fields, parameters, properties, attributes, etc.
- In tabular forms rows represent variables, data, measurement, etc, and the columns represent records with the set of values or vice versa.
- Tabular form offers simultaneous measurement or correlation of two values / variables where, one variable resides in columns and other in rows.
- The smallest unit of table is **cell.** To indicate particular cell in the table for storing and retrieving data to and from, the cell combination of the row and column is used. To indicate the cell at 4^{th} row and 2^{nd} column [4, 2] notation is used. Where first number indicates which row you want to point out and the second number indicates which column you want to point out. And the intersection of the specified row and column indicates certain cell in the table.

Example

- See the following example of employee's data represented in the tabular format.

Emp_id	Emp_name	Emp_designation	Emp_salary	Emp_Mobno
E1001	Kunal	Manager	40000	9868789954
E1002	Ishita	Project Leader	35000	9899443300
E2001	Shravi	Sr. Developer	30000	7833551231
E2002	Shrey	Sr. Tester	25000	9012345678
E2003	Diksha	Jr. Developer	20000	9912345678
E2004	Pratiksha	Jr. Tester	20000	8897654321

- P. E. Hoffman brings a new term for table as "Table Visualization". Table visualization leads to represent the data into tabular form.
- Here, Row is referred as "Dimension", that signifies independent variable of a record or tuple whereas the column is referred as "Variates" and signifies the dependent variable.

- Tables are useful for **comparison, analysis of the relationship,** and **composition** while table contains few variables and data points.
- Tables are the source for constructing all the types of charts. But it is not good idea to generate a chart, if the data can be understandable easily from the table.

Use tables when

- We want to simply represent our data in its raw format, or to represent a large number of text or string values.
- We want to do both at a time, display grand totals as well as compare or look up the individual values.
- We want to present a huge number of precise dimension values and measure values.
- When our audience wants to know the primary data.
- When we have the discrete audience across the world and each wants to see their own section of the table.
- When we want the values involve multiple units of measure.
- When we want that the data has to communicate quantitative information, but not trends.

When to not use tables

- There are many reasons to use a table as well as there are also many instances to use different data visualization types instead of tables because the table is difficult to read as compare to other data representation techniques.
- Also the table takes longer time to digest any information.
- So, for identifying the patterns and relationship, data visualization is needed, at that time using table for visualizing is not good idea, using graph is better choice.

2. Histogram

- A vertical bar chart is used to draw a histogram which represents the distribution of a set of data over a continuous interval or certain time period and relationships of a single variable over a set of classes.
- While representing the tabulated data into histogram, the tabulated frequency at every interval / bin / instance is represented by every bar in a histogram. And the total area of the histogram is equal to the number of data.
- The one of the most commonly used graphical presentation of data is Histogram.
- Histogram is used to graphically represent the huge amount of data / measurements / dimensions contained by table.
- A Histogram organizes and displays the table data in user-friendly format.
- That means the Histogram constructed to visualize the table data will make that data easy to understand by representing the number of majority values into a measurement scale, and representing number of variation presented in the data.

Use histograms when

- We want to summarize large data sets in the form of graphics : It is difficult to understand the set of table data, so by summarizing it on a tally sheet and organizing it into a Histogram we can make it easier to understand.
- We want to compare process results based on specification limits : For effective comparison use of histogram is good idea.
- While comparing the processes if the process specification limits are added to our Histogram then it will help to make quick decisions on whether the current process was capable to generate "good" products or not.
- The specification limits may be the basic features of product generated by the process like, length, weight, density, quantity of materials to be delivered, etc.
- We want to convey data / information graphically : Histogram helps to team members by giving an estimate of where to focus / concentrate, estimate of limits, and estimate of presented any gaps or unusual values in the data set.

- So it will helps to the team members to catch up the most frequently occurred values / variables.
- Histograms also help by giving a rough view of the probability distribution.
- Also depending on the information rendered by histogram the decision making is done.
- See the Fig. 5.5.2 of a Histogram.

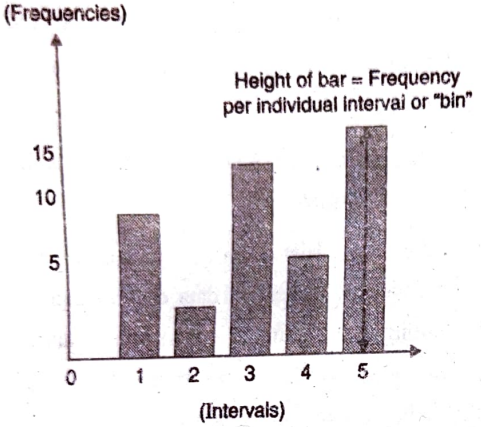

Fig. 5.5.2

➜ 3. Scatter plot

- Scatter plots are also known as X-Y Plots, Scatter Charts, Scatter Graphs, Point Graphs, or Scattergrams.
- Primary use of Scatter plot is to represent correlation and distribution analysis or clustering trends and also it will help to determine the anomalies or outliers in the data set.
- It is helpful to represent the relationship among 2 different variables where one may be or may not be correlates to another.
- Scatter plots uses to place a set of points through the Cartesian Coordinates for displaying values from two variables M and N as shown in Fig. 5.5.3.
- In this Fig. 5.5.3 values of variable **M** are represented on Y-axis and values of variable **N** are represented on X-axis.
- Scatter plots help to detect existence or absence of a relationship or correlation between these two variables (M and N).
- Scatter plots help to identify the **outliers** (means the points plotted far outside the general cluster of points) the Fig. 5.5.3 also shows **outliers.**

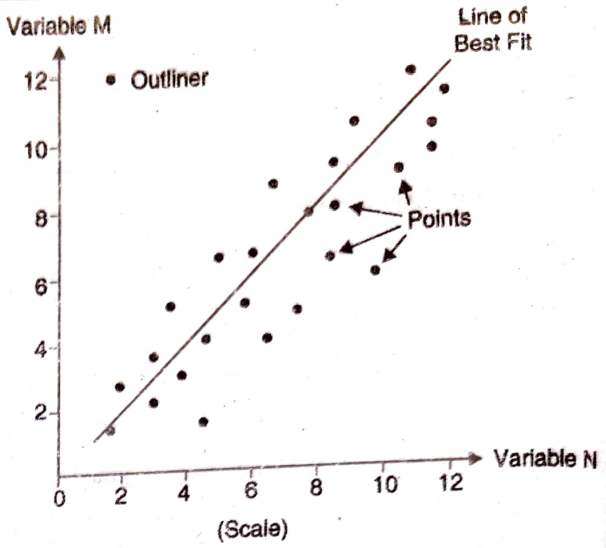

Fig. 5.5.3

- For analysis drawing Lines or curves as close to all the points as possible and showing how all the points were condensed into a single line would look. This is typically known as the **Line of Best Fit** or a **Trend Line** and can be used to make estimates via interpolation. The Fig. 5.5.3 also shows **Line of Best Fit.**

☞ Correlation

Correlation is categorized based on the patterns displayed on Scatter plots. Types of correlation are as follows:

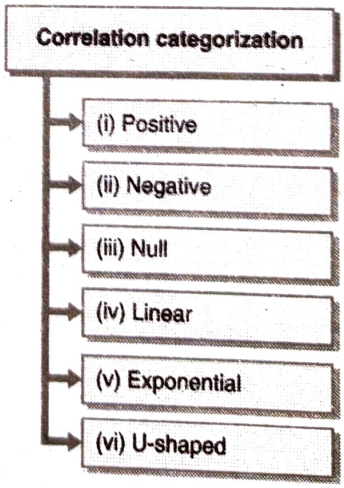

Fig. 5.5.4 : Correlation categorization

➜ **(i) Positive :** As value at X-axis increases the value at Y-axis also increases. We can say that in this correlation values are directly proportional to each other.

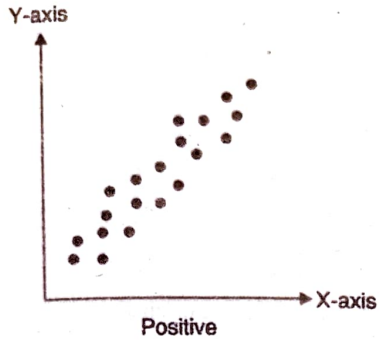

Fig. 5.5.5

➜ **(ii) Negative :** As value at X-axis increases, the value at Y-axis decreases. We can say that in this correlation values are inversely proportional to each other.

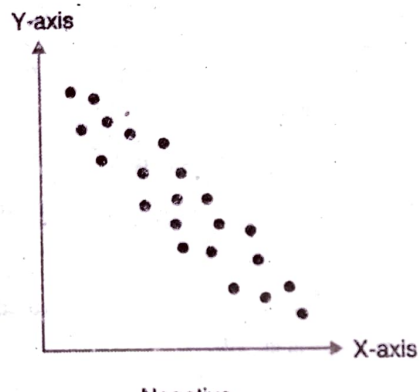

Fig. 5.5.6

➜ **(iii) Null :** No correlation. If a set of points generated by taking any combination of values of two variables and the variables doesn't relate at any combination of values then this pattern falls into null correlation.

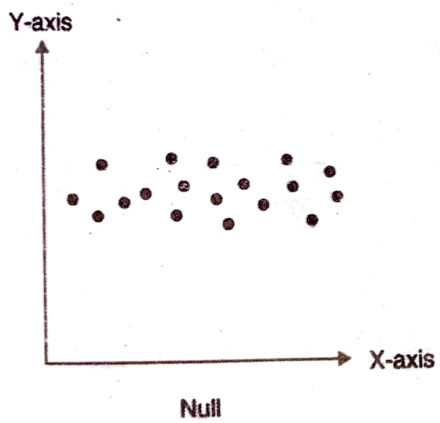

Fig. 5.5.7

➜ **(iv) Linear :** The linear correlation of two variables results following pattern as shown in Fig. 5.5.8.

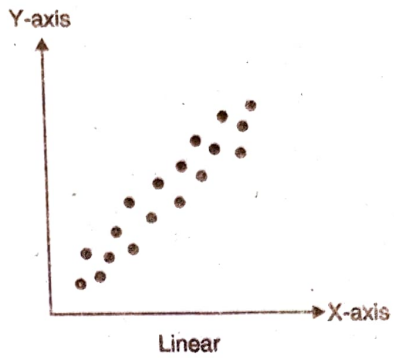

Fig. 5.5.8

➜ **(v) Exponential :** The exponential correlation of two variables results following pattern as shown in Fig. 5.5.9.

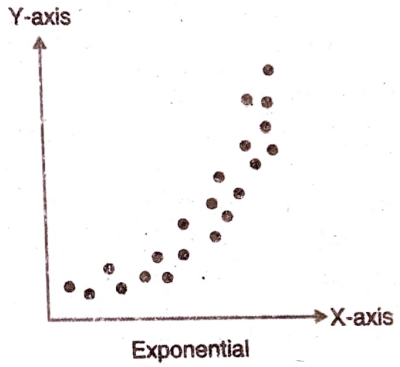

Fig. 5.5.9

➜ **(vi) U-shaped :** The correlation of two variables which results following pattern as shown in Fig. 5.5.10 falls into U-shaped.

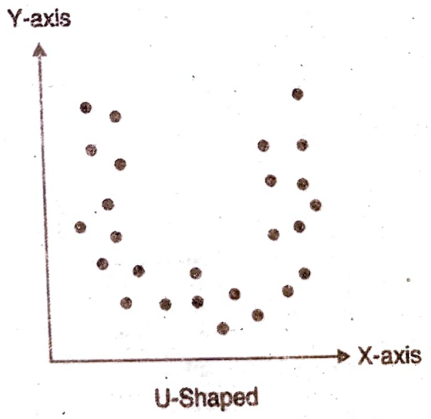

Fig. 5.5.10

– The strength of the correlation measured as strong, weak, or none. The strength is measured by how close the points are plotted to each other on the X-Y plane. For better understanding see the Fig. 5.5.11.

Fig. 5.5.11

- Scatter Plots are used when : When we want to detect the correlation (positive, negative, or null) in a large data set.
- When to not use Scatter plots : If we have only few pieces of information then, using a scatter plot will results into empty and pointless graph.

  If there is no correlation shown by scatter points then the chart also becomes useless.

4. Various charts

- There are various types of charts which are used to visualize the data. Let's talk about when to use charts for data visualization :
  - When we want to convey a message which is in the form of the data.
  - When we want to illustrate a correlation among several values.
  - For example, using charts for demonstrating the change in sells of a product (like extreme growth in sell) is best option as compare to tables because it will easy to grasp the contents shown by charts as compare to tables.

Charts

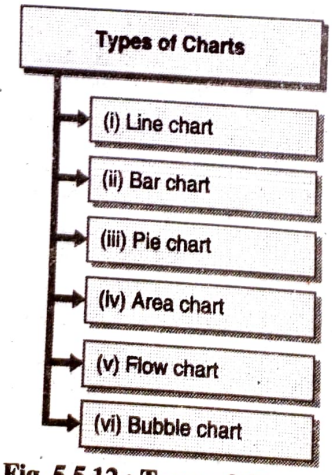

Fig. 5.5.12 : Types of Charts

Let's know one by one.

(i) Line chart

- The extension of Scatter plots is Line graph. One of the most commonly used chart type is Line chart. It is the most basic chart type which is formed by connecting a series of data points together with a line.
- Line chart is used for analyzing the financial status by connecting past recorded financial data together in a sequence with a line.
- Line charts are used to represent trends and analyze how data has changed over time.
- Line charts are used when we have a continuous data set. These are best suited for trend-based visualizations of data to display quantitative values over a continuous interval or period of time.
- Line chart helps to concentrate on the flow of the values changing over time or the continuity of the values.
- A line chart is also a good alternative to column charts when the chart is small.
- The line charts are usually used in stock market to demonstrate how the stock values are developed for a particular company over time on the stock market.
- Line chart connects set of data points which are indicated by icons or symbols with a line, or we can also simply draw a line without icons. See the Fig. 5.5.13 shows an example of line chart using individual symbols to illustrate different data points :

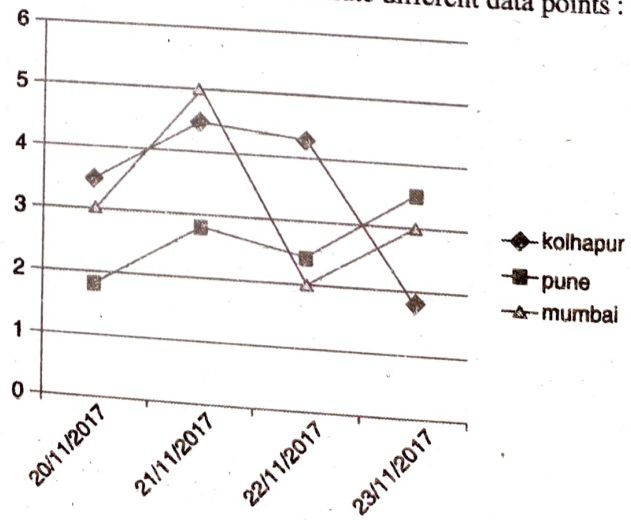

Fig. 5.5.13

- In the Fig. 5.5.13 the chart indicates the average temperature from date 20-11-2017 to 23-11-2017 in 3 cities Kolhapur, Pune, and Mumbai. The horizontal line indicates the temperature in Celsius degree and the vertical line indicates the dates.
- Depending on the data points to be plotted on the chart, the line chart has several forms like, Step line chart, Reverse step line chart, Vertical segment line chart, Horizontal segment line chart, Curve line chart.

(ii) Bar chart

- It is also known as Bar Graph or Column Graph. Usually the Bar chart is used to represent discrete data unlike line chart. Even though the graphical representation of bar chart and histogram is somewhat similar, we can distinguish bar charts from histograms as bar charts can't be constructed to represent continuous data over time intervals.
- The classic bar graph / charts uses either horizontal bars or vertical bars to illustrate discrete comparisons among groups. Where, one axis (it may be X-axis or Y-axis) of the chart represents the particular group being compared, and the other axis represents a discrete value scale.
- The bar chart may contain single (where each bar indicates different category), grouped (known as multi bars where bars are clustered in groups), or stacked bars (where bars are divided into subparts to show cumulative effect).
- See the following example of single bar chart :

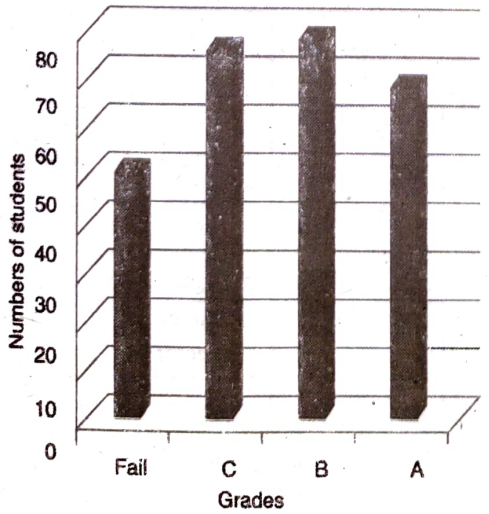

Fig. 5.5.14

(iii) Pie chart

DEFINITION

*The type of chart where a circle is divided to represent different categories of a data set is known as **pie chart** or circle graph.*

- A Pie chart shows the information in the data set in a simple and easy to understand manner.
- In this type of chart a circle is divided into proportional segments (referred as Sectors) each indicating distinct category for representing the proportions and percentages among the categories. The full circle represents the total sum of all the data, which is equal to 100%.
- Each sector shows the relative size of each value. The bigger the slice, the more of that particular data was gathered.
- Mostly for the comparison purpose a pie chart is being used. Various applications of pie charts can be found in presentations, business, offices, school, and at home.
- Data visualization using Pie chart becomes effective in a situation where we want to compare single segment of pie chart with the remaining segments of the pie chart.
- For ease of understanding and interpreting, different colors can be used to represent different segments in a pie chart.
- Data visualizing using Pie chart will be difficult if :
  - We want to compare different pie charts and different segments of different pie charts with each other.
  - We want to visualize several values or segments. For example if a data set has more than 10 values / segments then to visualize all in a circle graph becomes more complex because as the amount of values to represent increases, the size of each segment decreases. So it becomes difficult to analyze data at a glance. The comparison among groups of Pie Charts also becomes difficult as is harder to differentiate the size of items via area.

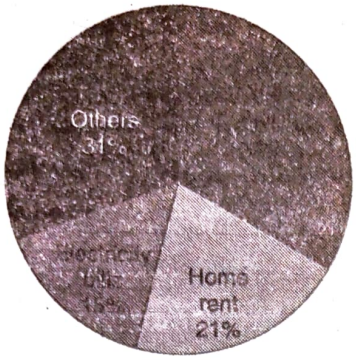

Fig. 5.5.15

(iv) Area chart

- Area chart is also known as area graph. The area chart is used to graphically represent the development of quantitative values / data over the intervals.
- Instead of individual values, area and line chart both are used to represent a time-series relationship, and/or continuity over a dataset, or trends.
- The differentiation between these 2 charts is that the area chart fills the area below the line with a specific color or texture.
- The steps to form an Area Graph / chart are as follows :
  - First plot the data points on a Cartesian coordinate grid
  - Then connect all the data points with a line
  - Finally fill the area below the completed line with a specific color or texture.
- Area chart / graphs can be grouped or stacked.
- The Grouped Area Graphs start from the same zero axes as shown in Fig. 5.5.16 :

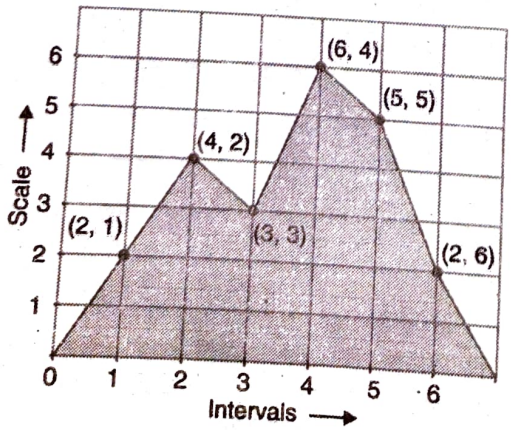

Fig. 5.5.16

- Whereas the Stacked Area Graphs starts each data series from the point left by the previous data series as shown in Fig. 5.5.17 :

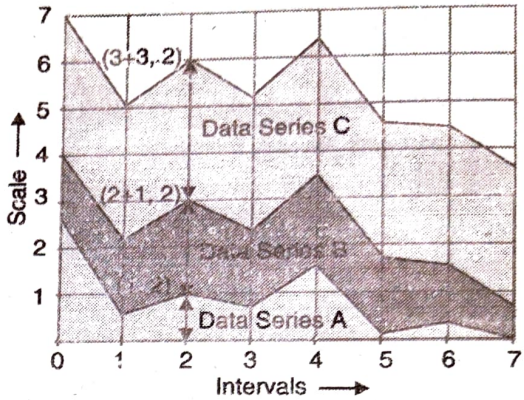

Fig. 5.5.17

(v) Flow chart

- The flow chart also called as *Flow Diagram, Flow Process Chart, Process Chart, Process Map, Process Model, and Work Flow Diagram.* A flow chart is defined as graphical or symbolic representation of sequenced steps of a process.
- Flow chart uses standardize special symbols to represent specific steps in the process. The standard symbols used in flow chart are as follows :

Symbols	Name
	Start/End terminal
	Process
	Input / Output
	Flow line
	Decision
	On-page connector
	Off-page connector

Fig. 5.5.18

- The flow chart is used to represent the flow of a process by using a series of connected symbols. Drawing a Flow chart of a process makes the working of process more understandable and also it helps to plan and develop a new process or improve existing process easily as the control flow is illustrated by flow chart. The Fig. 5.5.19 is an example of flow chart.

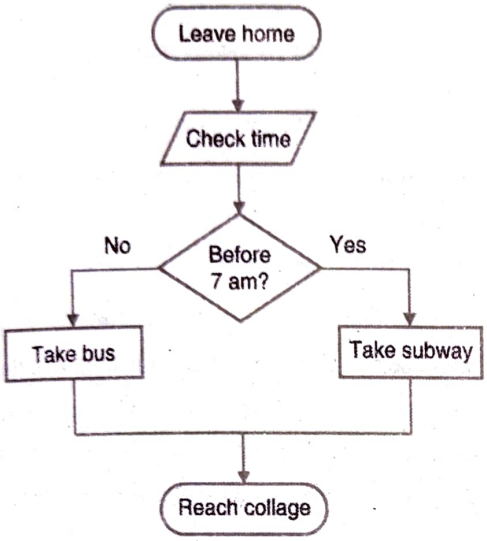

Fig. 5.5.19

(vi) Bubble chart

- The variation of Scatter Plot is bubble chart. In Scatter plot Chart dots are used to represent the data points whereas in Bubble Chart bubbles are used to represent the dots.
- As we discussed earlier the scatter plots compare two values, in addition to this the bubble chart adds one dimension namely bubble size as the third variable.
- In bubble charts the bubbles are plotted by three different dimension values to represent the data points, where one value represents its position along x-axis, second value represents its position along y-axis, and third value represent the size of the bubble in the chart.
- Usually in bubble chart a bubble is distinguished from other bubbles concerning its size and position. And if the bubble chart contains bubbles of same size then, use labels or colors to distinguish them from each other. Fig. 5.5.20 shows the example of bubble chart.

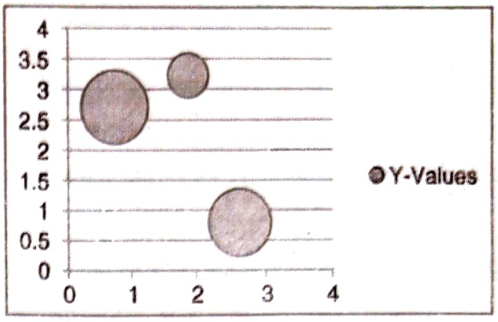

Fig. 5.5.20

- The Bubble Charts are applicable when we want to compare and represent the relationships among classified circles. To do so positioning and proportions are used. At the end we can say that the Bubble Charts are used to analyze patterns / correlations.

Multiple data series or combination of charts

- To represent multiple data series we can use combination of more than one type of charts.
- That's why here different data visualization techniques are enclosed into one chart.
- Fig. 5.5.21 shows the Combination of two chart types : bar chart and line chart to represent the trend of data series :

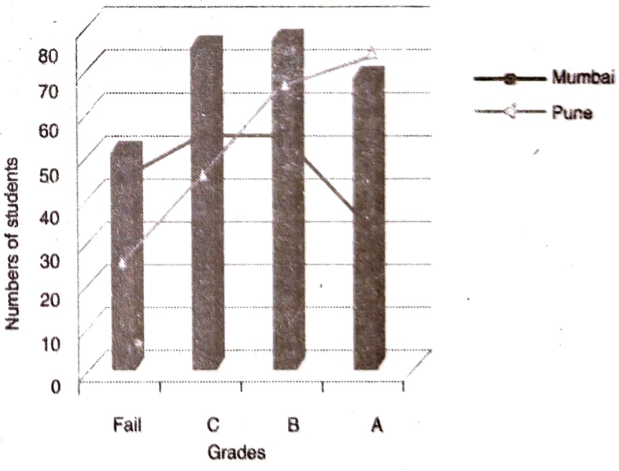

Fig. 5.5.21

- Similarly different graphs can be combined to create one graph. For example one can consider any combination from Area Series, Bar Series, Bubble Series, Candlestick Series, Line Series, and Plot Series etc.

5. Timeline

- A Timeline is a pictorial representation of a series of events in chronological sequence along with drawing straight line, that's why sometimes the Timeline is referred as chronology.
- Because of use of line while representing the chronological sequence of events in timeline, the users can easily understand the relationship among those different events.
- Most of the time the Timeline is a graphical representation of past events which are the part of history.

Categories of timeline

- Timeline categorized as linear timeline and comparative timeline.

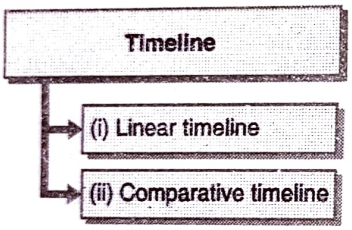

Fig. 5.5.22 : Timeline

(i) Linear timeline : It is a sequence of events that happened in particular period of time, we can represent them in horizontal or vertical manner.

(ii) Comparative timeline : It is two sets of ordered events which are occurred at a place or a set of ordered events which are occurred at two different places which can or can't be compared but can be assumed for some sort of knowledge.

Fig. 5.5.23 shows the example of timeline.

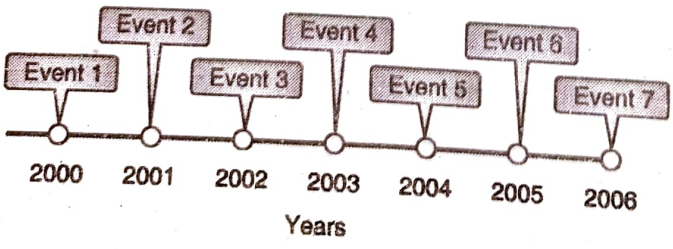

Fig. 5.5.23

6. Various diagrams

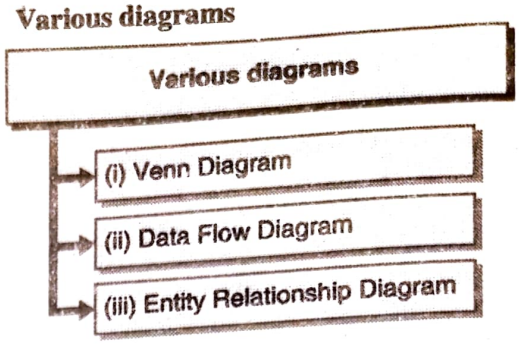

Fig. 5.5.24 : Various diagrams

(i) Venn diagrams

- In 1880 John Venn introduces Venn diagrams which are used to represent the relationship among two or more sets, so they are also known as set diagrams.
- It demonstrates the relationship among items of different sets. That is this diagram is used to determine intersection of sets (common characteristics) and determine complement of sets (dissimilar characteristics). Fig. 5.5.25 shows an example of Venn diagram.
- The Venn diagrams are used in engineering and scientific presentations, computer science and its applications, and theoretical mathematics, probability, logic, statistics, etc.

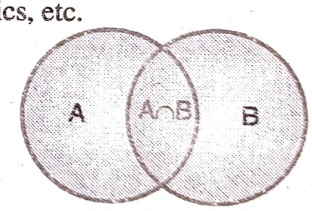

Fig. 5.5.25

- Here first circle represents one set called A and another circle represents other set called B. and the A∩B indicates the common objects in both set A and set B.

(ii) Data flow diagram

- DFD (Data Flow Diagram) is essential to discover the technique of data processing in a system.
- DFD is used to graphically represent the data or information transformation; means it is used to represent the data processing, data storing, and flow of data through the processes.
- DFDs are used to visualize the data processing. DFDs can be of different levels like DFD level0, level1, etc.

- Where, DFD of level 0 constructs overview of the entire system, and DFD of level 1 provide detail view of each process in the system i.e. it explains each process of the system in detail, and DFD of level 2 represents the system in more detail. Various notations are used while constructing the DFD, like, for constructing process, tacking input and generating output, using file or database and to show control flow etc.
- Fig. 5.5.26 shows the simple DFD Diagram.

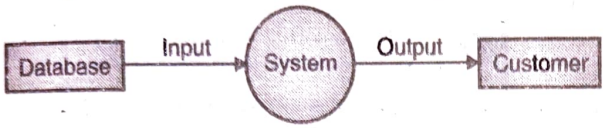

Fig. 5.5.26

(iii) ER diagram

- Entity relationship diagram is used in software engineering. It is an abstract and conceptual data representation technique.
- Basically the ERD is used for database modeling where it can helps to form conceptual schema, semantic schema, and also mostly it is used to build relational database schema for the system.

To construct the ERD several standard symbols are used which signifies various components of the ERD. ER diagram components are: Entity (an object or concept about which information needs to be stored), Relationship (how two or more than two entities share information), and Attributes (unique characteristic of an entity). The following example represents the ERD for hospital :

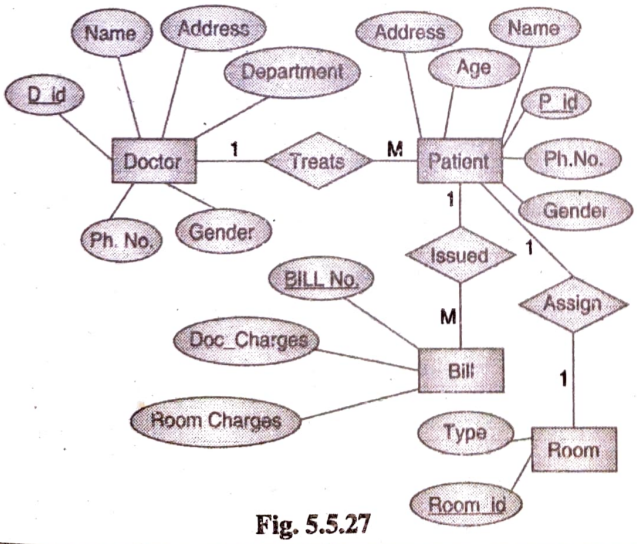

Fig. 5.5.27

Data visualization methods

Following are some other data visualization methods, which are in use even if they are less popular than the above methods:

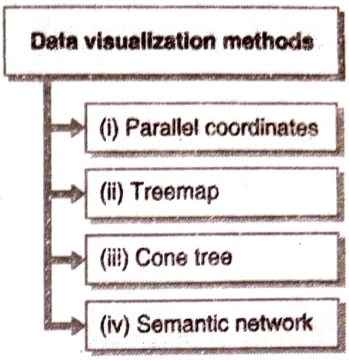

Fig. 5.5.28 : Data visualization methods

(i) Parallel coordinates

To plot individual data elements over several dimensions the parallel coordinates are used. It is very useful while displaying multidimensional data.

(ii) Treemap

For visualizing the hierarchical structured data, Treemap is used. For measurement purpose the size of sub-rectangle and the colors are used where the size of each sub-rectangle represents one measure, and another measure of data is represented by particular color.

(iii) Cone tree Cone tree is also used to represent the hierarchical structured data, for example, organizational body in three dimensions where the branches grow in the form of cone.

(iv) Semantic network

To represent logical relationship among several concepts semantic network is used. It generates directed graph i.e. the collection of nodes or vertices, labeled edges or arcs.

Syllabus Topic : Visualizing Big Data

5.6 Visualizing Big Data

- Some refer the Data visualization as a branch of descriptive statistics, and some refer it as a grounded theory development tool. The amount of data generated

by organization is increased year by year through Internet activities. This data is referred as "**big data**".

- The main problem in data visualization is that the collected data is useful only if valuable outcomes are generated from it by processing, analyzing and communicating this data. While doing so, ethical and analytical challenges are introduced. The data scientists help to deal with these challenges.
- Big data needs high volume, high velocity, and / or high variety datasets for processing the data optimization, discovery and decision making.
- The Big Data Visualization referred as the "front end" of big data.
- Data Visualization approaches are used to represent data in different sensitive objects like, tables, diagrams, images, and other objects.
- But Big Data visualization is not as easier as traditional visualization of small datasets.
- So to handle large-scale data visualization, many data scientists use some techniques to shrink the size of data before the actual data representation. And also it is necessary to choose correct data representation tool when visualizing big data.
- Challenges of Big Data introduces at the time of data capture, data storage, data analysis, data searching, data sharing, and data visualization.

Syllabus Topic : Tools used in Data Visualization

5.7 Tools used in Data Visualization

Q. 5.7.1 Explain tools used in data visualization. (Refer section 5.7) **(4 Marks)**

- For visualizing the data sets in the form of 2D and 3D various tools are used. Some of them provide animation facility through one or more dimensions.
- The simple, traditional tools such as line, bar, column, pie, etc are outdated and nowadays, advanced tools are introduced and used by most of the businesses.
- The use of visualization tools has advantage of quick deployment.
- Based on the nature of data sets of the business and the underlying structure of the business the visualization tools are chosen among several options.

Parts of visualization tools

The visualization tools are divided into 2 parts :

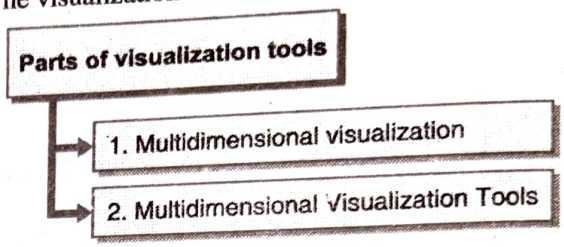

Fig. 5.7.1 : Parts of visualization tools

1. Multidimensional visualization

- There are two categories of Multidimensional Visualizations.
- Where the first type examines the category proportions or category counts.

Examples

- Following Examples of visualizations indicates the first type i.e. category proportions or counts :
  - Pie chart
  - Wordles
  - Bar chart
  - Histogram
  - Rank plot
  - Tree map
- And the second type examines the relationships among the variables.

Examples

- Following Examples of visualizations indicates the second type i.e. relationships among variables:
  - Scatter plot
  - Line chart
  - Step chart
  - Area chart

- Heat map
- Matrices
- Parallel coordinates/sets
- Radar/spider chart
- Box and whisper plots
- Mosaic display
- Waterfall chart
- Pixel bar chart
- Tabular comparison of charts

2. Multidimensional Visualization Tools

Google Charts

- This tool displays live data on our website.
- Google charts contain Introduction, Quick Start, & Chart Gallery for the ideas.

Many Eyes

- Many Eyes is an experiment done by IBM Research and the IBM Cognos s/w group.
- Many Eyes is developed by using JAVA and Flash. It is a public website means all the data and their visualizations are available to all the users and not for specific users only.
- It is a public website which allows users to upload the data and for such data the application generates interactive visualization.
- This Web application also allows the user to share their visualizations and also supports the user to discuss through various approaches from same data.
- Many Eyes is open-source. This kind of tool allows viewing others' visualizations, and also it supports to upload our own data and create our own visualizations.
- It allows different views like :
  - The representation of Relations between points using scatter plot, matrix charts and network diagrams.
  - The representation of Comparison between values using bar, histograms and bubble charts.
  - The representation of Trends which are changing over time using line, bar and category bar graphs.
  - The representation of Parts of a whole using pie chart, treemap, and treemap for comparisons.
  - The representation of Text analyzer using word tree, tag cloud, phrase net, word cloud.
  - The representation of Geographical graphics using charts on maps.

Tableau Public

- The most popular tool is Tableau Public developed by the US Company Tableau Software. It is a free tool and according to their website it "brings data to life".
- This kind of tool allows viewing others' visualizations or creates our own.
- We will discuss this tool in detail in next section.

Weave

- Weave stands for Web-based Analysis and Visualization Environment.
- The Weave tool is designed for visualizing any available data.
- Weave can handle different data types because it has large array of options for working with the various data types.

Wordle

- Wordle tool takes the text as input from user and generates "word clouds".
- The clouds generated by Wordle provide greater importance to words which are most frequently occurred in the source text.
- The clouds can be tweak according to the user's choice like with different fonts, layouts, and color schemes.

Specialized hierarchical and landscape visualization

- Hierarchical or Tree Visualizations represents the data in the form of collection of items linked to one parent item excluding the root of tree or hierarchy.

- Items and the links connecting parent and child have numerous attributes.
- Following Examples of visualizations indicates this kind of visualization:
  - Dendrogram,
  - Phylogenetic tree,
  - Radial tree,
  - Hyperbolic tree,
  - Tree map,
  - Cone tree,
  - Tadial hierarchy,
  - Fecision tree/flow chart.

Hierarchical Visualization Tools

1. Network Workbench

- This tool is generally used for analyzing, modeling and visualizing large-scale networks.
- It handles the data of biomedical, social science and physics research.
- This tool designs, evaluates, and operates a unique distributed, shared resources environment for large-scale network.

2. Protovis

- Protovis offers a graphical approach to visualization, which creates custom views of data with bars and dots.
- This tool also defines simple marks like bars, dots, etc through the dynamic properties which encodes the set of data.
- Protovis is declared and designed such that it can be learned easily with the help of examples.

5.8 Open-Source Data Visualization Tools

Q. 5.8.1 What are the open-source data visualization tools? (Refer section 5.8) **(4 Marks)**

- As we know the data visualization is the method of accepting data in the tabular or spatial form as input and transmitting it in a human-friendly and visual form.
- For visualization, tools are required. There are several open source tools which are useful during the creation of informative and helpful graphs.

Data visualization tools

- Following data visualization tools are open source :

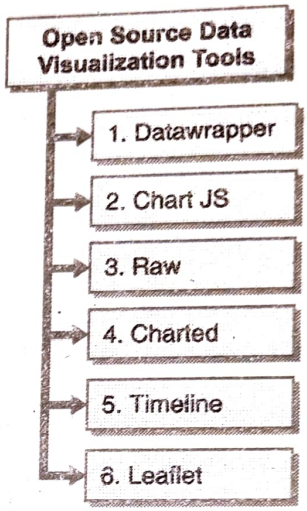

Fig. 5.8.1 : Open Source data visualization tools

D3, Dygraphs we will discuss in next section in detail.

1. Datawrapper

- Datawrapper is fully open-source data visualization which was produced in Europe by the journalism organizations. Datawrapper tool is designed to create data visualization for news institutes.
- Datawrapper assures that a graph can be created in just 4 steps with the help of web based Graphics User Interface (GUI).
- The steps are as follows :
  - To create a graph, click on the "New Chart" link on the top menu bar.
  - You can then paste your data in the text area;
  - Then, the tool analyzes it and shows you the preview.
  - If everything is fine, you can publish it.

2. Chart JS

- Chart JS is open-source, a clean charting library.
- Chart JS is a good selection for users who want control over the look and feel of their charts.

- So to visualize the data, before creating a chart, we required to include the library in our frontend code.
- After that we can add charts and assign values to them with the help of the API from the library.
- If the user doesn't want to code, this is not suitable choice for visualization.

➜ **3. Raw**

- Raw is also open-source, web based tool, which is built on D3.js library.
- It assures that if we paste our data then, graph can be created by following few simple steps with the help of web based GUI.
- As it is built on D3.js library it has all the features of D3. It is simple and ready to use tool of visualization for non-programmable users.

➜ **4. Charted**

- Charted tool is also open-source, which is invented by the product science team at Medium.
- Charted is one of most minimal charting tool which is available online.
- To visualize the data we need to just paste a link of a Google spreadsheet or a .csv file as input to the charted tool.
- Then the tool creates a chart with the data.
- To assure whether the chart is up-to-date or not the Charted fetches the data at some interval usually after every thirty minutes.
- Users can code and host their own version of charted tool. It is freely available online.

➜ **5. Timeline**

- When we want to display the set of events in sequential manner then, we use time line to represent those events in sequence. For this purpose an open-source tool is suitable namely Timeline.
- Before publishing the data using Timeline tool; the first thing the user needs to do is the formatting the data into Google Spreadsheet. And then use timeline generator to publish the data.
- Also it is possible to embed the timeline in the webpage by using the provided embed code.

➜ **6. Leaflet**

- Leaflet is lightweight, mobile friendly JavaScript library. It is used to create interactive maps.
- While designing the Leaflet following things are taken into consideration: performance, simplicity, and usability.
- It takes advantage of HTML5 and CSS3 on advanced browsers, along with older ones.
- Leaflet is a tool which works across all the major desktop and mobile platforms.
- The Leaflet has well-documented, easy to use, and beautiful API and readable source code, so only using plug-ins available in the market we can extend it.

5.9 Data Visualization with Tableau

Q. 5.9.1 Explain Data visualization with Tableau. (Refer section 5.9) (4 Marks)

- In the data analytics and visualization the most popular tool is Tableau. Tableau is a business intelligence (BI) software tool. It supports interactive visualization of data.
- It provides faster visualization because it has an in-memory data engine which helps to speed up the visualization.
- The Hadoop infrastructure can be embed by Tableau.
- For structuring the queries and cache information for in-memory analytics, Tableau uses Hive.
- Caching reduces the latency of a Hadoop cluster, so it is able to offer interactivity among the users of tool and Big Data applications.
- It is a free tool used for data visualization with the help of graphics which merge a graphical interface (which is fast, appealing and efficient) with traditional elements of BI tools (like, the organizational model of variables by dimensions and measures, or connection with other information management systems like, databases, and spreadsheets).

Features

Some of the features of this tool are as given below:

1. **Quick and easy data acquisition**

- It can handle any size of databases and spreadsheets.
- The data inputted to this tool can be in the Microsoft Excel, Access, or plain text formats.
- This tool can work with different graphics like fever, bars, stacked bars, pie, maps with polygons, lines or points, etc.

2. **Publication of interactive graphics**

This tool merges variety of data sources in a single view.

3. **Data are public**

The data Visualized by this tool is public, so to accept the data as raw material it is possible to download that data from the visualization.

4. **Tableau has three main products to process large-scale datasets** as given below :

- Tableau Desktop
- Tableau Sever
- Tableau Public

More information about one of them is given as below:

5. **Tableau Public**

- For desktop clients, the Tableau offers a free version known as Tableau Public so it is refered as Desktop Application which uses Windows and JavaScript technologies.
- Even though the Tableau Public is free version of Tableau, it provides most of the similar powerful visualization capabilities as the paid versions of Tableau provides.
- The users have the ability to analyze data from sources like Excel sheets, or from geographical visualizations like, Gantt charts, Treemaps, etc.
- As Tableau Public provides some features like robust data preparation and visualization, we will require the Professional and/or Server edition so that we can :
  - Share interactive visualizations over our organization.
  - Save files on our computer.
  - Connect with advanced data sources (like, Hadoop, Oracle databases, Microsoft SQL Server and SQL Server Analysis Services etc.).
- Tableau Public allows us to connect with files like, Excel or .CSV spreadsheets, etc.
- Although web data connectors are available to connect with databases which are published in web formats like HTML, most of the businesses require the Professional or Server editions for connecting with their databases.
- Still the Tableau Public is good for visual analytics, if we emphasis on flat files like Excel workbooks.
- The Tableau Public's main feature is its user interface.

Note : Tableau Public is perfect for users who need visual interaction with data for the analysis purpose without accessing complex data sources.

5.10 Introduction to : Pentaho, Flare, Jasper Reports, Dygraphs, Datameer Analytics Solution and Cloudera, Platfora, NodeBox, Gephi, Google Chart API, Flot, D3, Visual.ly

Q. 5.10.1 Give brief introduction of following :
(i) Gephi, (ii) D3.js,
(iii) Pentaho (iv) Dygraph.js
(Refer section 5.10) **(8 Marks)**

- The Traditional data visualization tools are not sufficient to handle big data as the size of data to be handle is huge and complex.
- For visualizing the data some software are developed which contains the functions of visualization and interaction for data visualization.

Some of these software are discussed below :

1. Pentaho

This software supports the spectrum of BI functions like analysis, dashboard, enterprise-class reporting, data mining, and data integration as discussed below:

Analysis

The Mondrian OLAP Server and JPivot library offers the analysis engine for navigating and exploring. It does the multidimensional analysis.

Reporting

It allows the designing, creating, distributing the reports in different known forms (like, HTML, PDF, and so on) from different sources.

Pentaho reports are created based on the JFreeReport library, but the reports created with external reporting libraries like, Jasper Reports or BIRT can be integrated in Pentaho.

Dashboard

It is used for monitoring and analyzing the Key Performance Indicator (KPIs). The latest version of BI Suit includes a set of tools which allows users to create attractive dashboards, with graphs, reports, analysis views, and other Pentaho contents within few efforts.

Data Mining

For predictive analysis and understanding the business, Data Mining is performed on the business data using some algorithms.

Data Integration

It integrates the scattered information from different sources like, databases, files, applications, etc. in order to provide that final integrated information to the user.

- **The versions of Pentaho Server are :**
  - Open Source
  - Professional Standard
  - Professional Premium and Enterprise.
- **There are three layers :**
  - Presentation layer (has reporting, analysis, dashboard, and process management),
  - Business intelligence platform (has security administration, business logic and repository),
  - Data and application Integration (has ETL, Metadata, EII. This can be built on 3rd party application like CRM.)
- Pentaho presents on all three layers with the respective products : **Data layer, server layer, and client layer.**

2. Flare

- It is an Action Script library for creating data visualization which runs in Adobe Flash Player.
- Including the basic charts and graphs to complex interactive graphics the Flare tool supports data management, visual encoding, animation, and interaction techniques.
- A modular design of Flare features leads the developers to create customized visualization techniques without reinventing the creation process.
- Flare is open-source software, meaning it can be freely deployed and modified.
- Flare is used by many popular organizations such as, BBC News, Slate magazine, Wired Italia, etc.

3. Jasper Reports

- It is an open-source java reporting tool.
- It has a novel software layer for generating reports from the big data storages.
- JasperSoft supplies different commercial software product around the JasperReports product.
- The main product of JasperSoft is JasperReport Server which is a Java EE web application. It offers advanced report server capabilities like report scheduling and permissions.
- The reports generated by JasperReports are defined in XML file format also known as JRXML that can be hand-coded, and designed, using a tool.

- JRXML files are saved with the .jrxml extensions and its compiled version is a .jasper file.
- The compilation can be done using the iReport on the fly or using the JasperCompileManager class at runtime. iReport is a visual designer for JasperReports.

4. Dygraphs

- Dygraphs is a fast, flexible, open source JavaScript charting library that helps to search and understand dense data sets.
- Dygraphs produces the interactive zoomable charts of time series.
- It enables the developer to produce interactive charts with X and Y axis to present powerful diagrams.
- As the data being parsed is more, proportionally the functionality of graph is higher.
- The Dygraphs are built for visualizations which contain a huge number of views.
- Dygraphs.js allows analyzing the individual parts of a data set.
- Dygraphs.js library is compatible across all major web browsers, and capable to respond to touch sensitivity.

Features of Dygraphs

1. Huge data set Handling: Dygraphs plots millions of points without using external server or Flash.
2. Supporting error bands around the data series.
3. Interactive Pan or zoom: It provides interactive out of the box zoom or pan
4. Displaying values on mouse-over: by default the mouse-over is on.
5. Highly customizable: In Dygraph wide set of options are available for customization. Using these options and custom call backs we can raise digraphs to do as we want.

Dygraphs can works on all the recent browsers we can squeeze to zoom on the devices like mobile or tablet.

Compatible with Google Visualization API.

5. Datameer Analytics Solution and Cloudera

- The Datameer and Cloudera together provide a complete big data analytics solution.
- The Cloudera's Enterprise Data Hub (EDH) permits us to store our data entirely in Hadoop where the storage is centralized and cost-effective.
- Datameer is end-to-end big data analytics application that runs locally on Cloudera's EDH, which allows us to move quickly from raw data to insights.
- Not just the technical staff but also the business analyst can get the power of big data analyst due to the spreadsheet interface provided by Datameer.
- Datameer, Inc is a big data analytics and visualization company situated in San Francisco in California. It offers self service and schema-free Big Data Analytics application for Hadoop.
- Datameer focuses on the analysis of large amount of data for business users of Apache Hadoop.
- The Datameer, Inc Company's product is Datameer Analytics solution (DAS), which is a business integration platform for Hadoop.
- The DAS contains data source integration, and an analytics engine with a spreadsheet interface which is specially designed for business user with over two hundred analytic functions and visualization including reports, charts and dashboards.
- Major Hadoop distributions like, Apache, Cloudera, EMC Greenplum HD, IBM Big Insights, MapR, Yahoo!, and Amazon uses the DAS.

6. Platfora

- Platfora is number one Big Data Discovery platform built natively on Hadoop and Spark.
- Platfora, Inc. founded in 2011, and situated in San Mateo, California. It offers a big data analytics platform referred as Platfora, which allows organizations to analyze their data like, business transactions, customer interactions with organization, and machine data.

- The company's Platfora combines in-memory acceleration, BI analytics and visualization in order to make more efficient big data analytics and make simpler data discovery.
- The business users and data scientists can visually interact with petabyte-scale data in seconds because of Platfora which allows them to analyze organization's data like, business transactions, customer interactions with organization, and machine data to discover new opportunities and handle risks.
- Platfora provides a way to transform the business unlock insights, make decisions, and generate enhanced outcomes with the help of its industry-defining Customer Analytics, and Internet.
- The Big Data Discovery scales with high-performance interactivity by scaling the in-memory architecture horizontally across infinite nodes.

7. NodeBox

- NodeBox is node-based software, which is used for creating two dimensional graphics and visualizations. It is easy-to-use, fast and efficient application developed for generative design.
- NodeBox can import many data formats like Excel spreadsheet and also it is possible to write our own data importers and exporters. Because of these unique approaches the NodeBox is ideal for rapid data visualization.
- Parameters of NodeBox can be animated. The NodeBox is derived from a programming language Python.

8. Gephi

- Gephi is one of the open-source tools written in java and OpenGL which has JavaScript-based visualization platform.
- It is used to manipulate very large and complex data sets.
- The Gephi is designed to use by scientists and data scientists rather than used by business analysts.
- It is a data explorer.
- Gephi is a graph-based visualization tool which not only separates large data sets and generate attractive visualizations, but also provides the ability to clean and sort the data.
- **It is a platform for :**
  - The interactive visualization
  - Exploration of networks
  - Dynamic and hierarchical graphs.
- **It is used to :**
  - Represent the relationship between data and its evolution, grouping sets, etc.
  - Represent the hierarchies
  - Exporting and importing tables, among other functions.
- **It is a network analysis tool. The network analysis includes :**
  - Biological network analysis
  - Social network analysis
  - Link analysis (It is a type of network analysis where relations among object classes are clarified by treating the objects as nodes of a network.)
- Gephi is best for only graph visualization so it is not made for other types of visualizations.
- Since Gephi has limited data sources, it can't be used as a visual analytics platform for all the purpose.

Note : Even though the Gephi do not have some features like data preparation and integration and supports limited common data sources; it is a great tool used by data scientists in experimental network analysis.

9. Google Chart API

- The Google product, Google Chart Tool allows user to make simple visualizations using an online tool, and make the process of visualization design public.
- Google Chart API has some features like; users can easily enter or start using the tool and less programming knowledge is needed for creating attractive static charts and interactive visualizations.

- With the help of a wide range of libraries and APIs, developers create their own visualizations.
- Because the deprecation of static visualization segment of the toolset, currently the Google Chart Tools site only offers dynamic visualization tools.
- Using the Google Chart Tools we can get right visualization as per our requirements as long as we didn't need customization and happy with the Google look.
- This tool creates graphic images as PNG. It is free to use but with some limitations. Initially, its limit was 50,000 requests per URL, but now the limit is extended to 250,000 requests per URL.
- Using the external server as cache of generated images we can avoid this limitation.
- Without coding of graphical elements the visualization is done and this visualization will be hosted by Google for free, which allows easy sharing across the web.
- To visualize the data many formats are supported like: bubble charts, line plots, treemaps, and geographic maps.

10. Flot

- Flot is a jQuery library for line graphs and bar charts so, the person who is familiar with jQuery can easily handle the calls back, styling and behavior of the graphics.
- Flot can works on all the browsers which support canvas. It provides direct canvas access to draw custom shapes.
- Float allows us to access to lot of call back functions in order to run our own code and style the results when some events occur like hover, click, mouse-out, etc.
- Flot provide more flexibility than other charting tools. It supports lines, plots, field areas in any combinations. And also supports different visualization options for data points, stacked charts, panning and zooming, and plugins which give various capabilities, for specific functionality.
- Flot is one of the oldest charting libraries around and focuses on simple usage and interactive features. But on the other hand it also means that you have full control over the presentation, animation, and user interaction.
- Most modern and backward compatible till IE6 browsers supports Flot. The Flot's plugin repository permits to use additional plot types which are contributed by community. It also plots textual data and their categories.

11. D3.js

- D3 stands for Data-Driven Documents.
- It is a small, free JavaScript library used to operate data based / HTML documents using open web standards.
- D3 creates quick web-based visualization with standards, like, HTML (Hypertext Mark-up Language), SVG (Scalable Vector Graphics), and CSS (Cascading Style Sheets)
- It also handles interactivity and integrates staged animation and smooth transitions into the webpage.
- Without depending on the proprietary software, the browsers may render visualizations which are complex.
- D3 is open-source so it can be used and adapted by other users. The different forms of D3 are possible as vast as the geometry itself (bubbles, Chord diagrams, node links, etc).

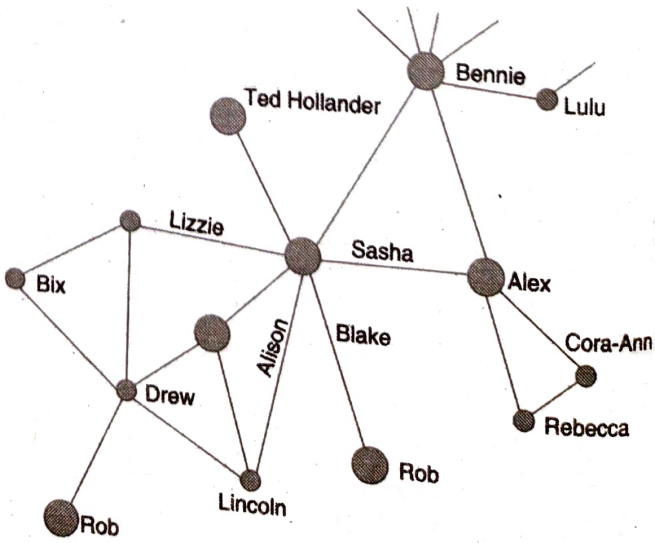

Fig. 5.10.1 : A Character map using D3

- D3 provides the facility to bind the data with DOM (Document Objects Model - for Representation) and apply the transformations on it.
- E.g. set of numbers creating an HTML table, and creating an interactive SVG graphic with transitions and interactions using the data.
- D3.js can generate visualizations like, choropleth, hub plot, motion chart, and fisheye distortion.

12. Visual.ly

- Visual.ly is a community platform for data visualization and infographics, but probably the Visual.ly is most popular for infographic rather than data visualization.
- Visual.ly acts as both the showcase for infographics and marketplace and community for publishers, designers, and researchers.
- Even though it is a marketplace for visualization, the site enables the users to find images by description, tags, and different category sources like education to business or politics.
- Also users can pick a template, connect it to social media like Facebook or Twitter and get back some nice cartoon graphics as a result so that they can publish the graphics as profile or share it through social networks.
- The Twitter Visualizer Tool is the first tool of Visual.ly. This tool accepts Twitter handles of two people from user and creates infographics to compare the personalities in terms of their hobbies, followers count, occupation, etc.

Syllabus Topic : Analytical Techniques used in Big Data Visualization

5.11 Analytical Techniques used in Big Data Visualization

Q. 5.11.1 What are the advanced analytical methods in big data visualization ? (Refer section 5.11) **(8 Marks)**

- In big data visualization some analytical techniques are used to visualize large amount of data so that the data should be perceivable.

☞ Analytical methods

- To do so following analytical methods are introduced :

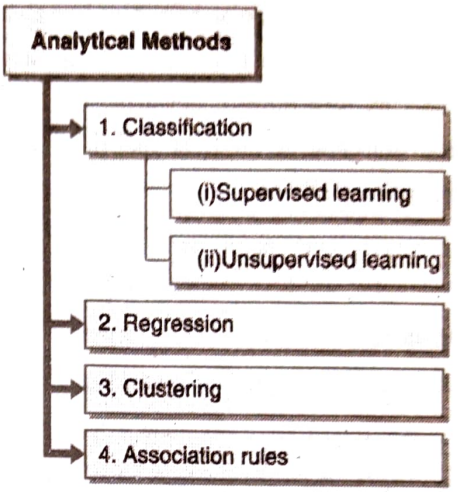

Fig. 5.11.1 : Analytical methods

➜ 1. Classification

- The classification predicts the class of the object belongs to. In this, a classifier accepts a set of pre-classified examples and learns these examples to assign unseen examples. In simpler words, the classifier assigns class labels to new observations.
- There are several classification methods but all are fall into either of following two classification models:

➜ **(i) Supervised learning :** Most classification methods are supervised, where set of pre-labelled data is learned by classifier to classify the unlabeled observations.

➜ **(ii) Unsupervised learning :** In this, hidden structures are discovered from unlabeled data.

One of the most popular classification methods is decision tree.

☞ Decision tree

- It also known as prediction tree. Here the tree structure is used to indicate the decision sequences and consequences.

- Here X = {x1, x1, ••• xn) is input and the goal is to predict a response or output variable Y.

  Where x1,x1, ... xn are the input variables.

- In order to achieve the goal, a decision tree needs to form with test points and branches.

  Where at every test point particular input variable is tested and a decision is made to choose a particular branch, and then traverse down the tree through the chosen branches to reach at the final point which indicates the prediction.

- A decision tree is a structure of test points (called nodes) and branches that indicates the decisions being made.
- In this structure a node which doesn't have further branches is referred as a leaf node.
- These leaf nodes return class labels or probability scores.
- It is possible to convert the decision tree into a set of decision rules.

➜ 2. Regression

- The main difference between classification and regression is that, Classification predicts something will happen, whereas the regression predicts how much of it will happen.
- Regression analysis is performed to demonstrate that the outcome variable is influenced by a set of variables.
- The outcome variable depends on the other variables, so it is known as a dependent variable.
- And the variables on which the outcome variable is depend are called as the input variables or the independent variables.
- To answer the following types of questions Regression analysis can use :
  - How much a person's expected income is?
  - What is the probability that an applicant will fail to clear a loan?
- To answer the first question **linear regression** method is useful.
- And to answer the second question **logistic regression** is useful.

☞ Linear Regression

- This technique is used to form the relationship among many input variables and a continuous outcome variable.
- While forming the relationship among an input variable and the outcome variable assumes that the relation among them is linear.
- Application of Linear regression is in business, government, and other sectors. Some common practical applications are as follows:
  - A simple linear regression analysis can be used for pricing in Real estates.
  - A linear regression analysis can be used in businesses and government sectors to predict demand for goods and services.
  - A linear regression model can be used in medical section to analyze the effect of a proposed radiation treatment on reducing tumour sizes.

☞ Logistic Regression

- The logistic regression analysis is performed with a variety of situations in both the public and the private sectors.

☞ Sectors in which logistic regression model is used

- In Following sectors the logistic regression model is mostly used :

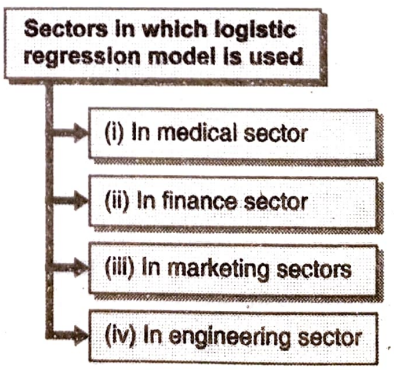

Fig. 5.11.2 : Sectors in which logistic regression model is used

➜ **(i) In medical sector**

To predict the patient's response for specific treatment the logistic regression analysis is used.

➜ **(ii) In finance sector**

With the help of a loan applicant's credit history and the details on the loan, determining the probability that an applicant will clear the loan or not.

➜ **(iii) In marketing sectors**

The logistic regression is used to determine the carrier switching probability of a wireless customer depending on some factors like his / her age, the number of family members in the plan, how much data in the plan is remaining (in months), and social network contacts.

➜ **(iv) In engineering sector**

The logistic regression is used to determine the malfunction or failure probability of mechanical part depending on the operating conditions and various diagnostic measurements.

➜ **3. Clustering**

- In general, clustering is an unsupervised technique which is used for grouping similar objects.
- As the Clustering techniques are unsupervised, the data scientist does not need advance-determination of labels to be applied to the clusters.
- To exploratory analyze the data Clustering method is used.
- The structure of the data describes the significant levels of objects and determines better way to group the objects having same characteristics.
- In this method Predictions are not made; instead this method finds similarities between the objects, and groups similar objects into clusters.
- Clustering techniques are applied in marketing, economics sectors, and different science branches.
- Clustering leads to classification because after cluster identification, each cluster is labelled with specific name according to its characteristics.
- A popular clustering method is k-means.

☞ **K-means**

- k-means is an analytical technique.
- Suppose a collection of objects having n measurable attributes is given,
- Then, k-means identifies k (a chosen value of k) clusters of objects based on the objects' proximity to the middle of the k groups.
- The middle is determined by calculating the arithmetic average (mean) of each cluster's n-dimensional vector of attributes.
- Some specific applications of k-means are image processing, medical, and customer segmentation.

➜ **4. Association Rules**

- Another unsupervised learning method is association rules.
- In this technique predictions are not made, instead it is a descriptive technique which discovers remarkable relationships among the items that are hidden in a large dataset.
- These discovered relationships are denoted as rules or frequent item-sets.
- The database transaction mining is done through Association rules.
- Following questions are answered by the association rules:
  - Which products are dependent on each other means the customer tends to buy together?
  - What products tend to buy by the customers?
  - After purchasing this product by the customers what are the other products do they tend to view or buy?
- For example, see the Fig. 5.11.3 to understand how the assignment rule answers the above questions.

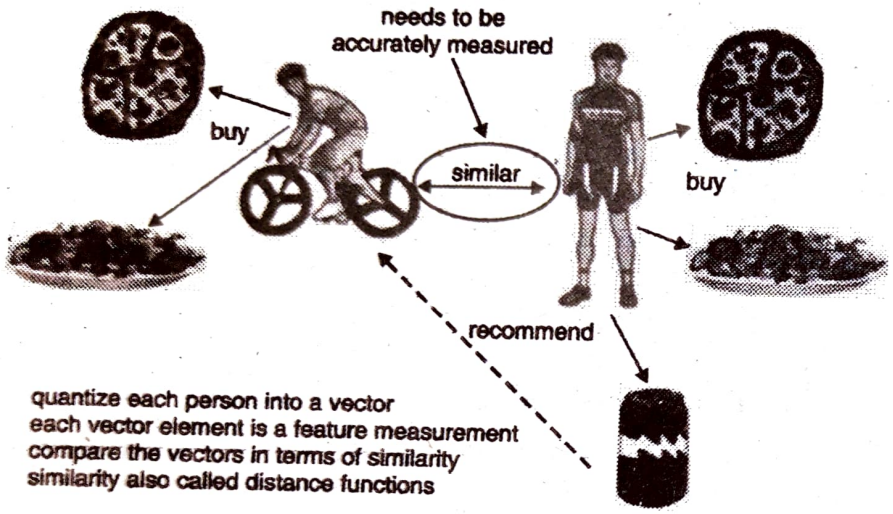

Fig. 5.11.3

Applications of Association Rules

The market basket analysis involves the process of implementing the association rules mining which are used by different companies for following purposes :

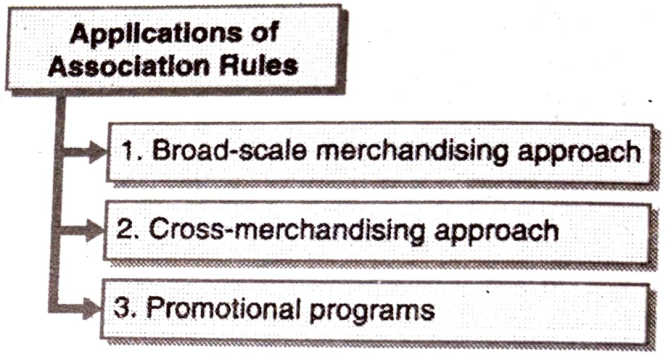

Fig. 5.11.4 : Applications of association rules

➜ **1. Broad-scale merchandising approach**

To specify which products should be included in or excluded from the inventory of every month.

➜ **2. Cross-merchandising approach**

To relate products and high-margin or high-ticket items and compare them to better merchandising.

➜ **3. Promotional programs**

With the help of a loyalty card program managing incentives gain by buying multiple products can be calculated.

❑❑❑