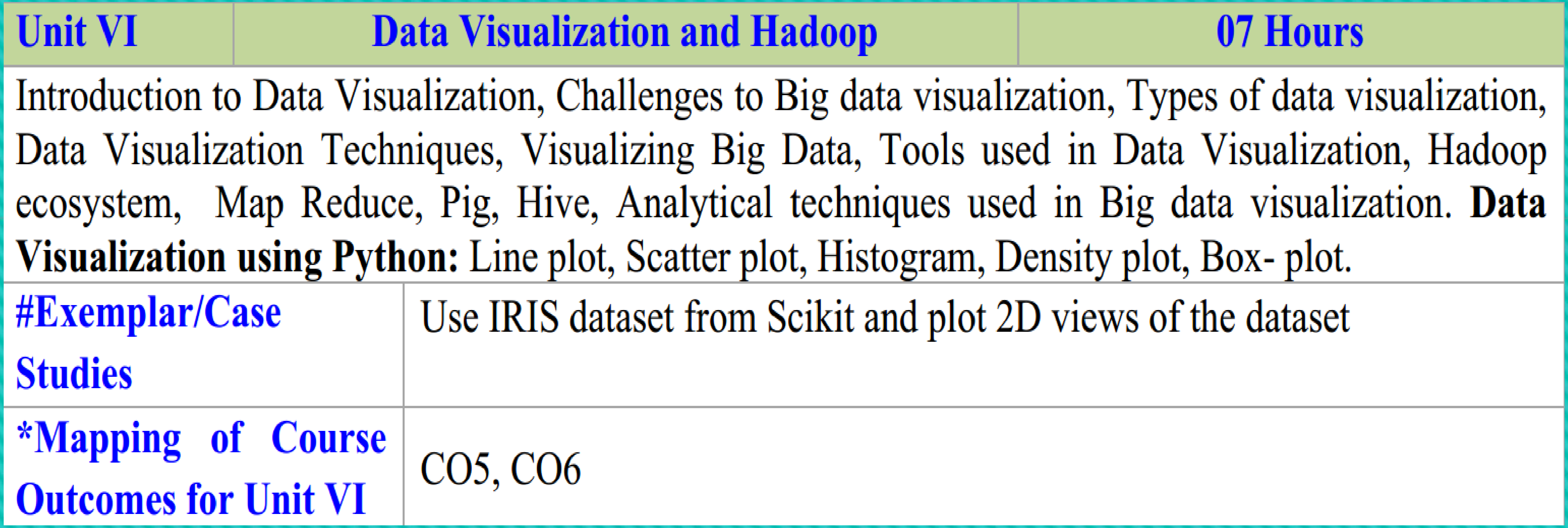
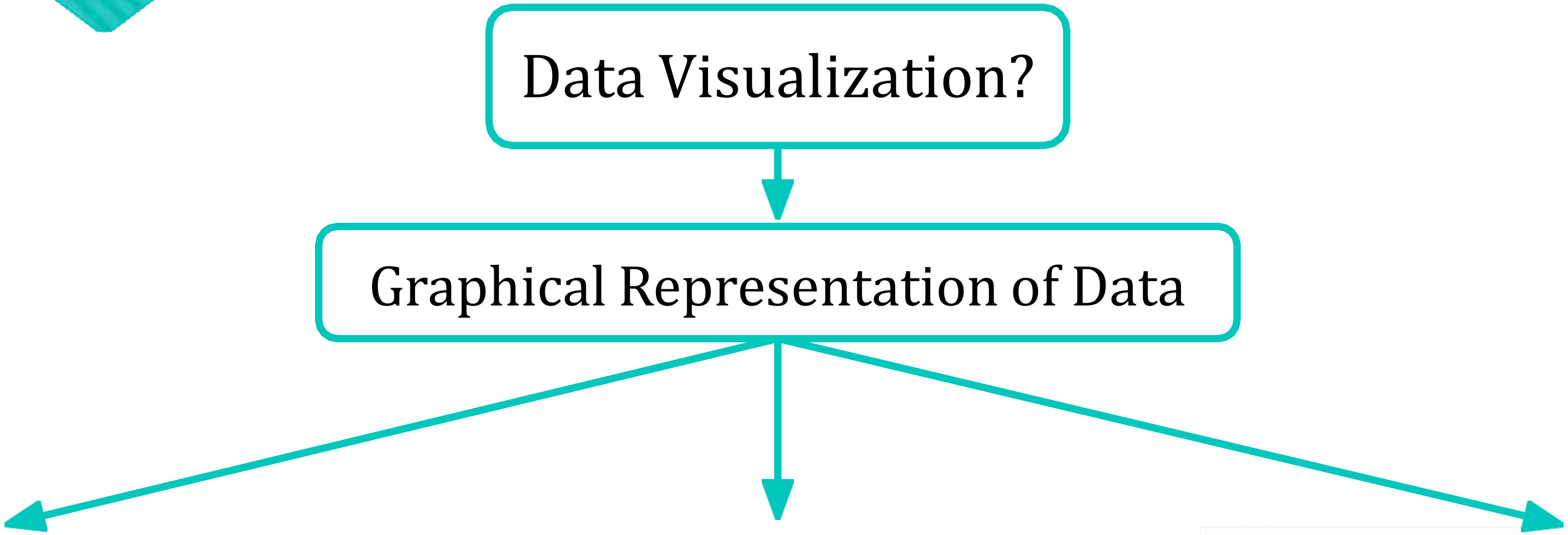
Data Analytics
Unit- VI
Big Data Visualization



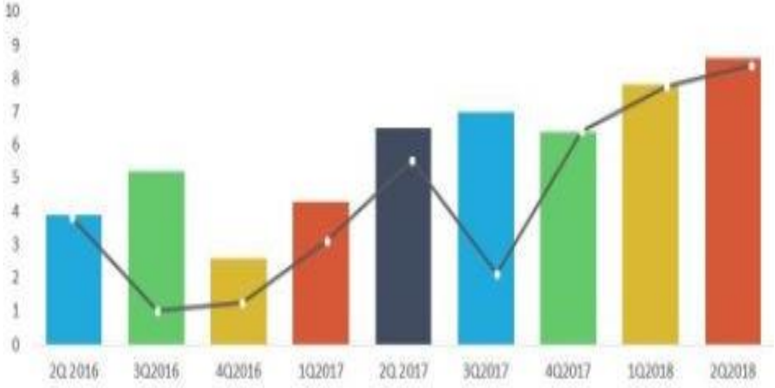
Unit VI	Data Visualization and Hadoop	07 Hours
Introduction to Data Visualization, Challenges to Big data visualization, Types of data visualization, Data Visualization Techniques, Visualizing Big Data, Tools used in Data Visualization, Hadoop ecosystem, Map Reduce, Pig, Hive, Analytical techniques used in Big data visualization. **Data Visualization using Python:** Line plot, Scatter plot, Histogram, Density plot, Box- plot.		
#Exemplar/Case Studies	Use IRIS dataset from Scikit and plot 2D views of the dataset	
***Mapping of Course Outcomes for Unit VI**	CO5, CO6	

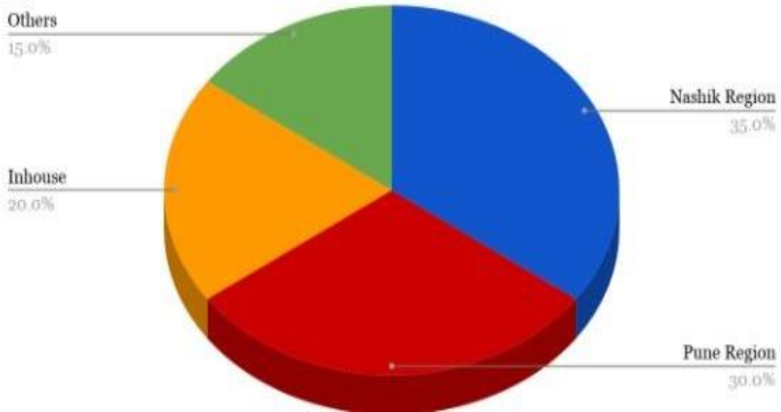
Introduction to Data visualization

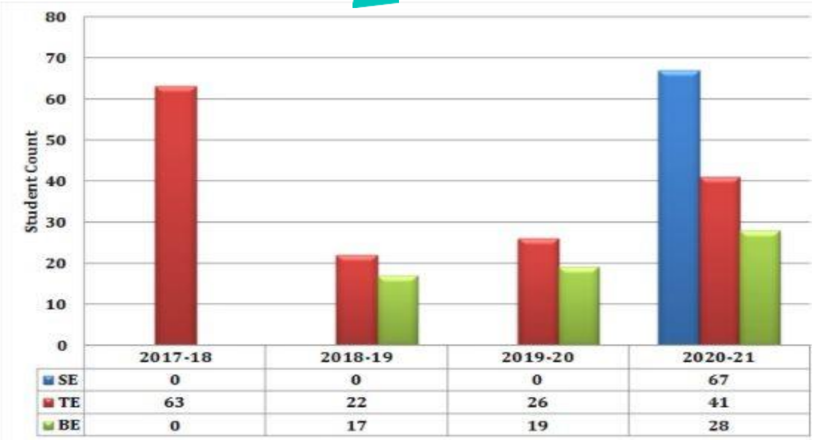


Data Visualization?

Graphical Representation of Data







	2017-18	2018-19	2019-20	2020-21
🟦 SE	0	0	0	67
🟥 TE	63	22	26	41
🟩 BE	0	17	19	28

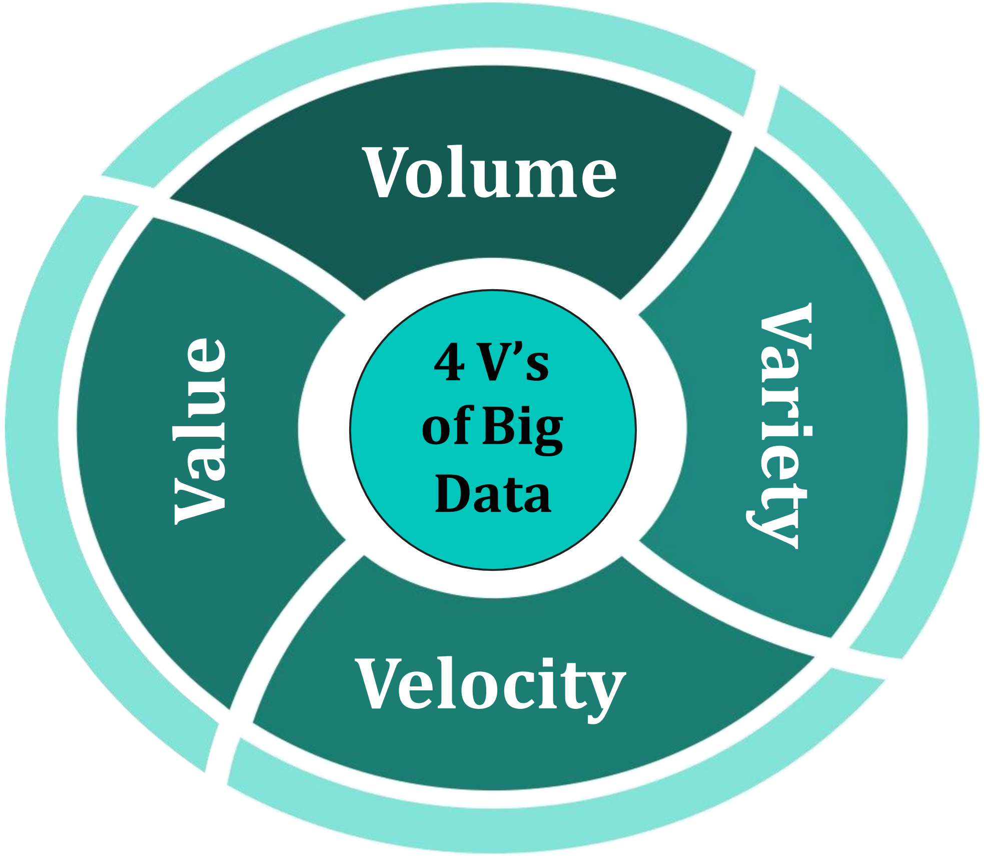
Introduction to Data visualization

- Data visualization is a graphical representation of any data or information.
- Visual elements such as charts, graphs, and maps are the few data visualization tools that provide the viewers with an easy and accessible way of understanding the represented information.
- Data visualization enables you or decision-makers of any enterprise or industry to look into analytical reports and understand concepts that might otherwise be difficult to grasp.

Reasons/Objectives for using data visualization

- To explore sources
- To find patterns and Relationship among the data.
- To illustrate or hide the data.
- to predict sales volumes
- to identify areas that need attention or improvement
- to understand what factors influence customers' behavior
- to know which products to place where
- to discover how to increase revenues or reduce expenses
- spreadsheets are hard to visualize
- patterns and trends can be spotted quickly and easily
- Saves time and energy

Big data visualization Challenge?



Challenges to Big data visualization

The visualization-based methods take the challenges presented by the “four Vs” of big data and turn them into following opportunities . [1].

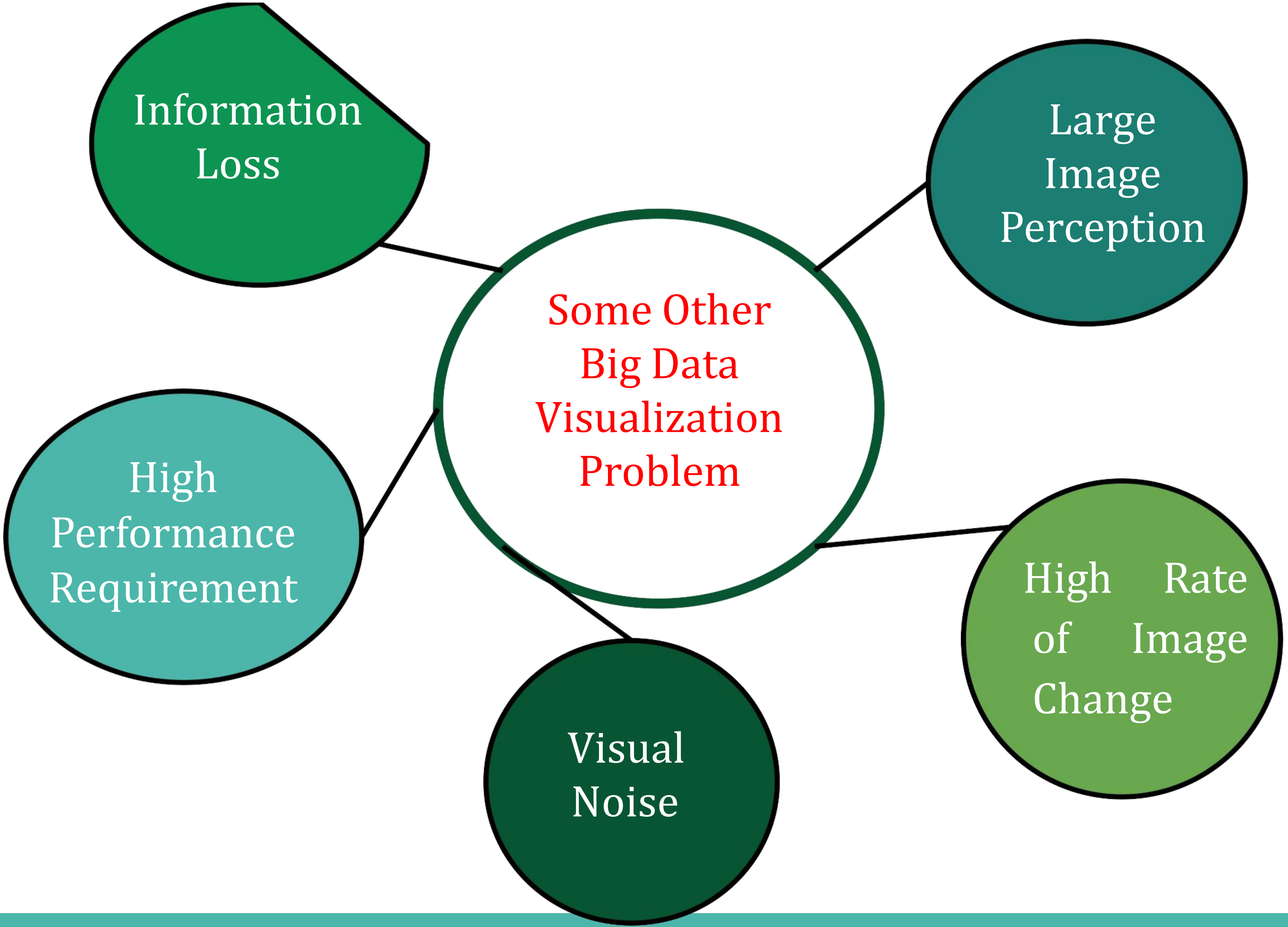
- Volume: The methods are developed to work with an immense number of datasets and enable to derive meaning from large volumes of data.
- Variety: The methods are developed to combine as many data sources as needed.
- Velocity: With the methods, businesses can replace batch processing with real-time stream processing.
- Value: The methods not only enable users to create attractive infographics and heatmaps, but also create business value by gaining insights from big data.

Challenges to Big data visualization

- Visualization of big data with diversity and heterogeneity (structured, semi-structured, and unstructured) is a big problem.
- Speed is the desired factor for the big data analysis.
- Designing a new visualization tool with efficient indexing is not easy in big data.
- Cloud computing and advanced graphical user interface can be merged with the big data for the better management of big data scalability
- Visualization systems must contend with unstructured data forms such as graphs, tables, text, trees, and other metadata.
- Big data often has unstructured formats.
- Due to bandwidth limitations and power requirements, visualization should move closer to the data to extract meaningful information efficiently.
- Visualization software should be run in an in situ manner. Because of the big data size, the need for massive parallelization is a challenge in visualization. [1]

Challenges to Big data visualization

Some Other Problems For Big data visualization


Information Loss
Large Image Perception
Some Other Big Data Visualization Problem
High Performance Requirement
High Rate of Image Change
Visual Noise

Challenges to Big data visualization

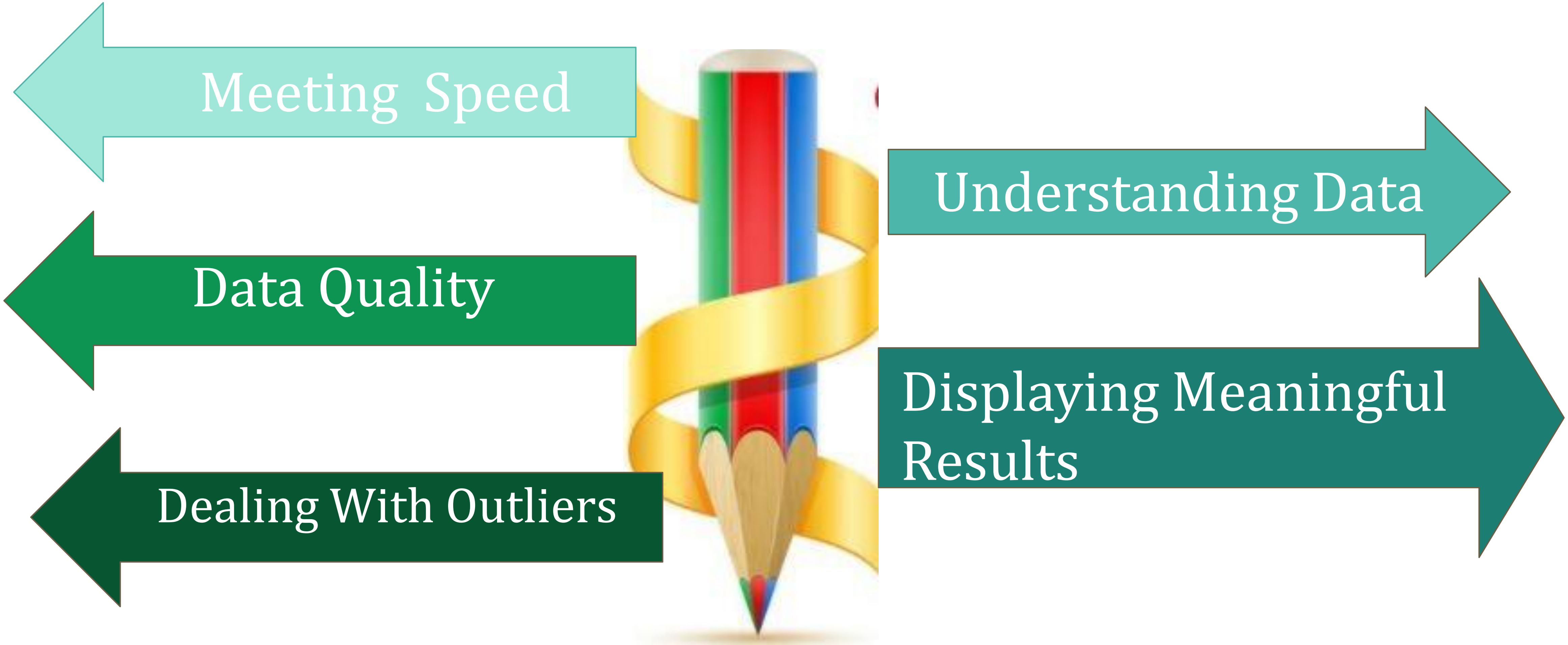
There are also following problems for big data visualization: [1]:

- **Visual noise:** Most of the objects in dataset are too relative to each other. Users cannot divide them as separate objects on the screen.
- **Information loss:** Reduction of visible data sets can be used, but leads to information loss.
- **Large image perception:** Data visualization methods are not only limited by aspect ratio and resolution of device, but also by physical perception limits.
- **High rate of image change:** Users observe data and cannot react to the number of data change or its intensity on display.
- **High performance requirements:** It can be hardly noticed in static visualization because of lower visualization speed requirements--high performance requirement.

Challenges to Big data visualization

- In Big Data applications, it is difficult to conduct data visualization because of the large size and high dimension of big data.
- Most of current Big Data visualization tools have poor performances in scalability, functionalities, and response time.
- Uncertainty can result in a great challenge to effective uncertainty-aware visualization and arise during a visual analytics process [1]

Solution



Meeting Speed

Data Quality

Dealing With Outliers

Understanding Data

Displaying Meaningful Results

Solution

Potential solutions to some challenges or problems about visualization and big data were presented

1.**Meeting the need for speed**: One possible solution is hardware. Increased memory and powerful parallel processing can be used. Another method is putting data in-memory but using a grid computing approach, where many machines are used.

2. **Understanding the data**: One solution is to have the proper domain expertise in place.

3.**Addressing data quality:** It is necessary to ensure the data is clean through the process of data governance or information management.

4.**Displaying meaningful results**: One way is to cluster data into a higher-level view where smaller groups of data are visible and the data can be effectively visualized.

5.**Dealing with outliers:** Possible solutions are to remove the outliers from the data or create a separate chart for the outliers.

Conventional data visualization tools

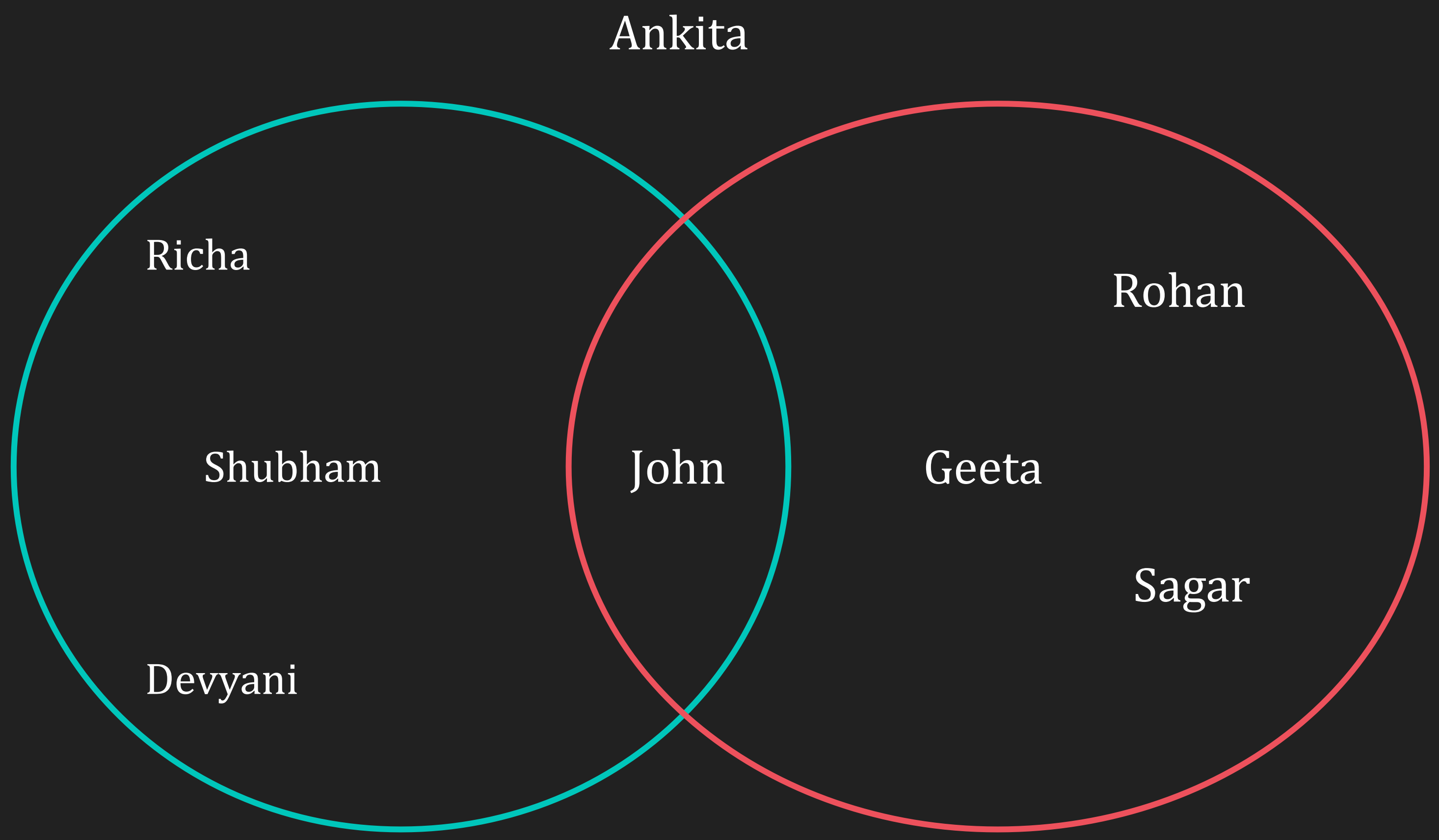
- Many conventional data visualization methods are often used.
- They are:
  - Table, histogram, scatter plot, line chart, bar chart, pie chart, area chart, flow chart, bubble chart,
  - multiple data series or combination of charts, timeline,
  - Venn diagram, data flow diagram, and entity relationship diagram, etc.
  - The additional methods are: parallel coordinates, treemap, cone tree, and semantic network, etc

Conventional data visualization methods

1. Table, histogram, scatter plot, line chart, bar chart, pie chart, area chart, flow chart, bubble chart
2. multiple data series or combination of charts, timeline
3. Venn diagram, data flow diagram, and entity relationship diagram
4. Parallel coordinates, treemap, cone tree, and semantic network

Conventional data visualization tools

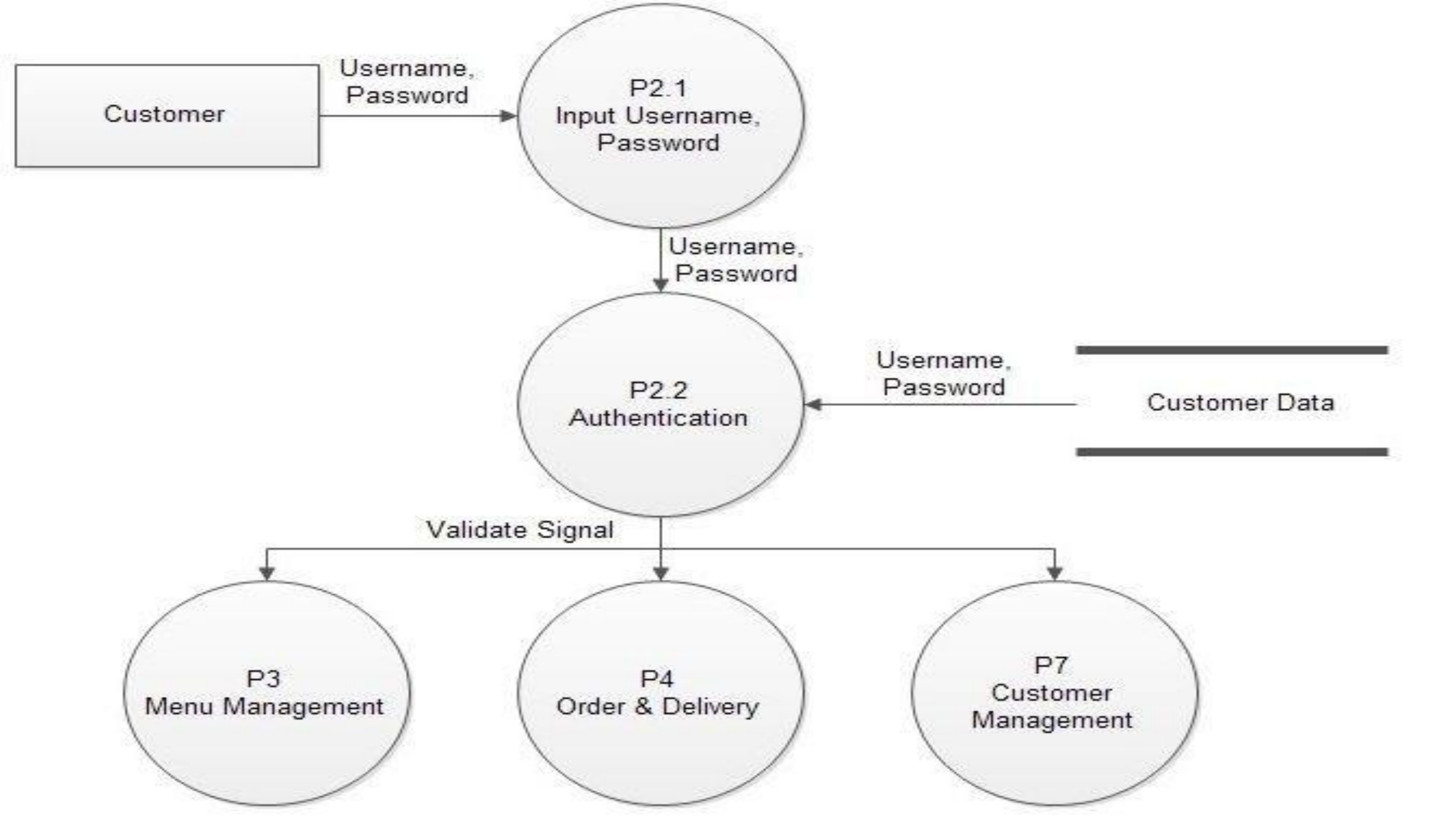
Venn Diagram



Conventional data visualization tools

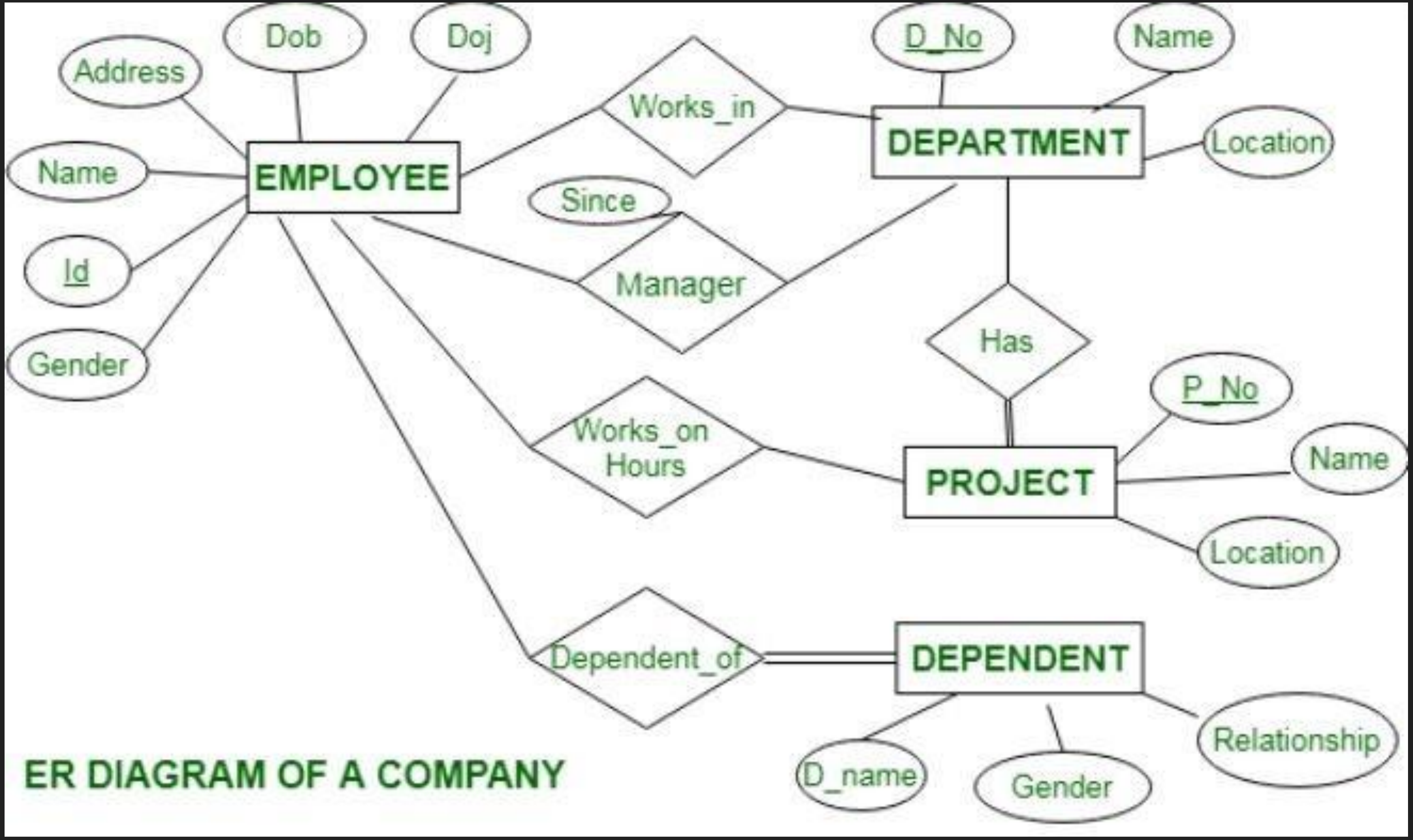
DFD- Data Flow Diagram

https://sites.google.com/site/bookstoreintelligent/analysis/context-data-flow-diagram



Conventional data visualization tools

ER- Entity Relationship Diagram

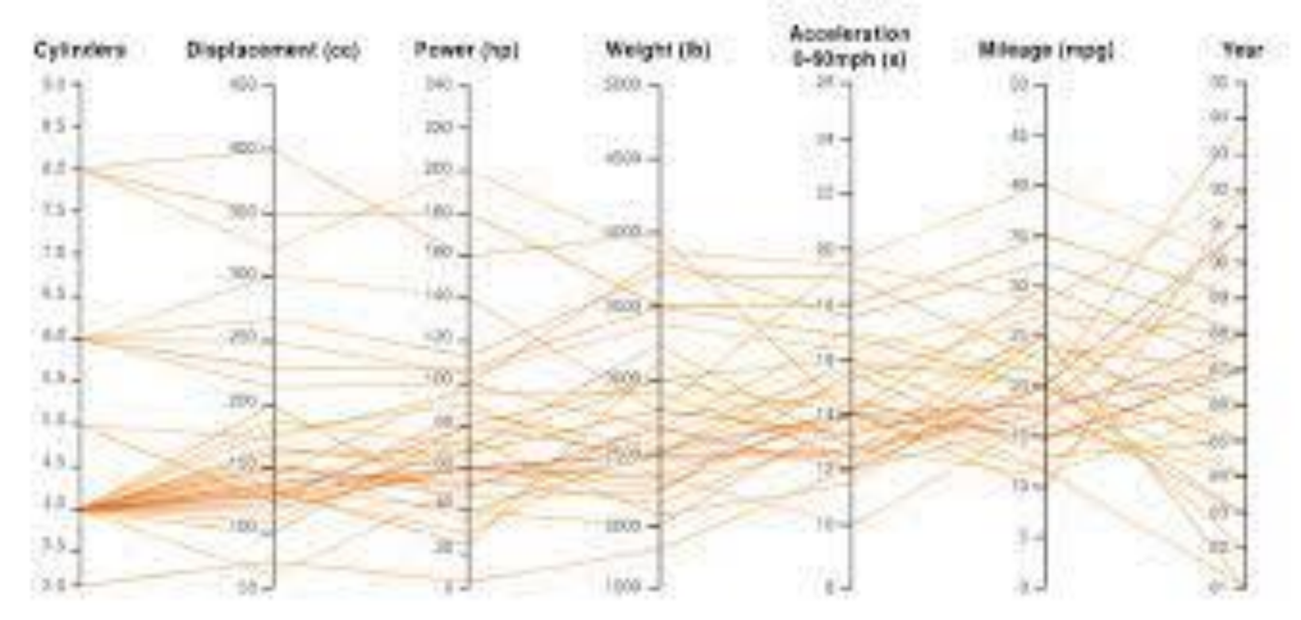


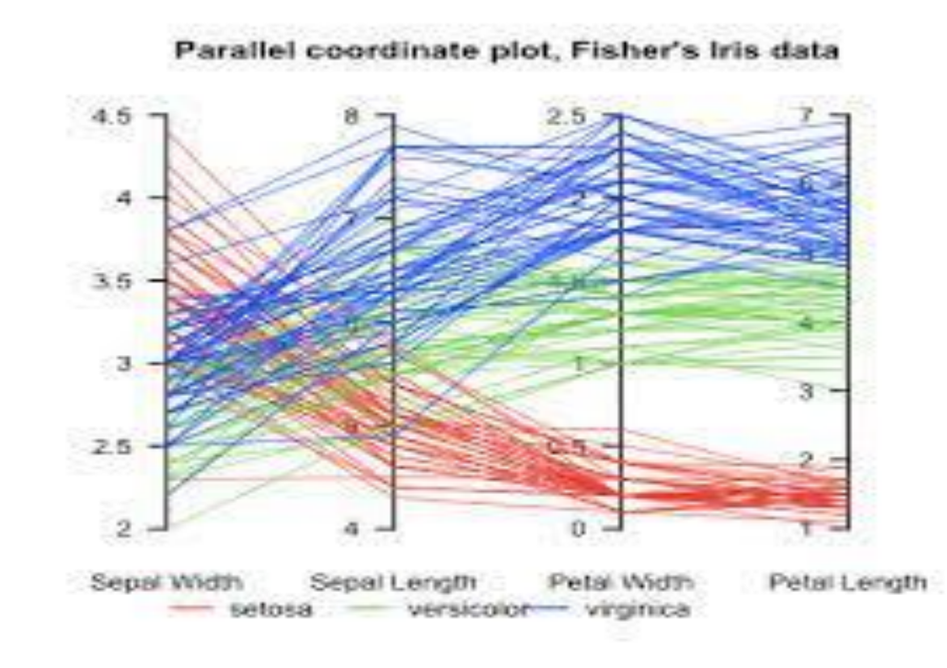
ER DIAGRAM OF A COMPANY

https://www.geeksforgeeks.org/er-diagram-of-a-company/

Conventional data visualization tools

- **Parallel coordinates** is used to plot individual data elements across many dimensions.
- Parallel coordinate is very useful when to display multidimensional data.

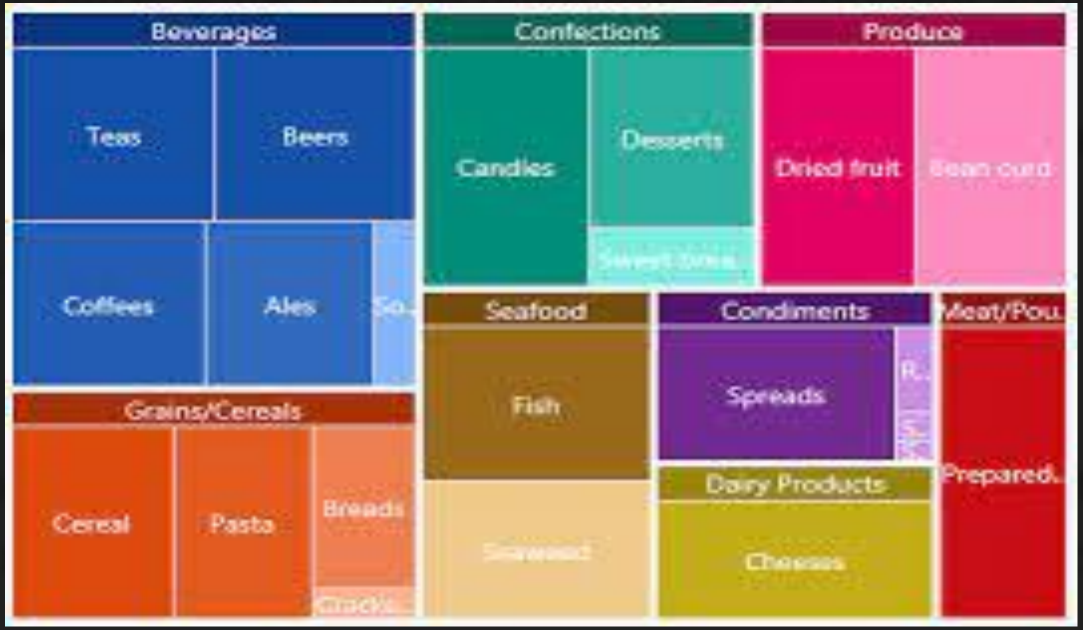




- A semantic network is a graphical representation of logical
- It generates directed graph, the combination of nodes or vertices, edges or arcs, and label over each edge [1].

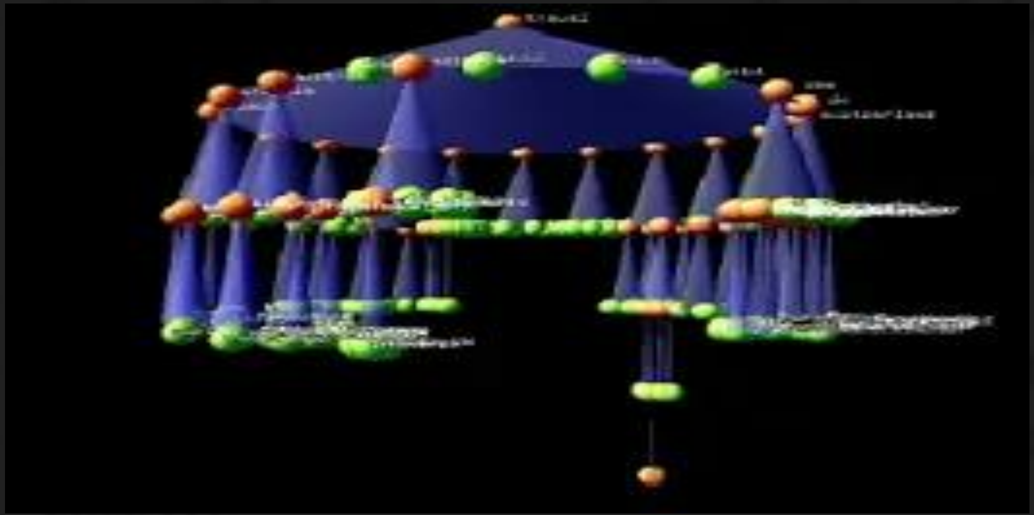
Conventional data visualization tools

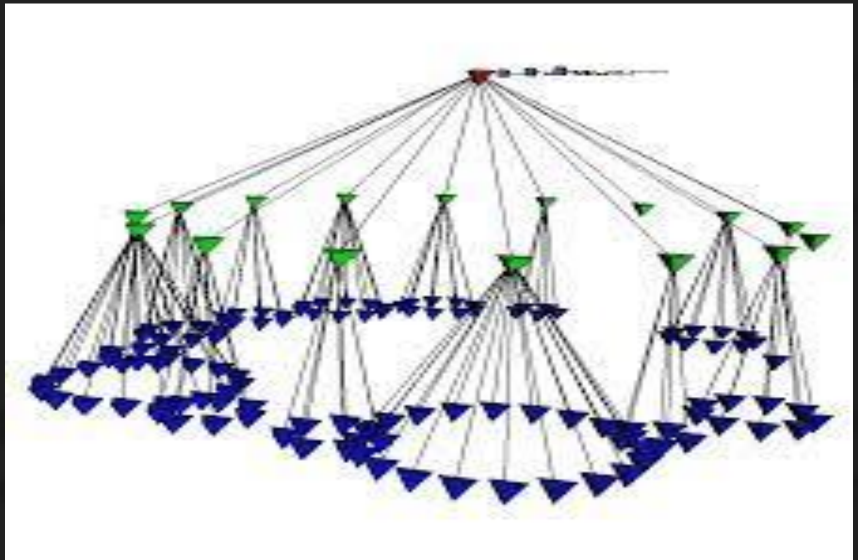
- **Treemap i**s an effective method for visualizing hierarchies.
- The size of each sub-rectangle represents one measure, while color is often used to represent another measure of data.
- A treemap of a collection of choices for streaming music and video tracks in a social network community.



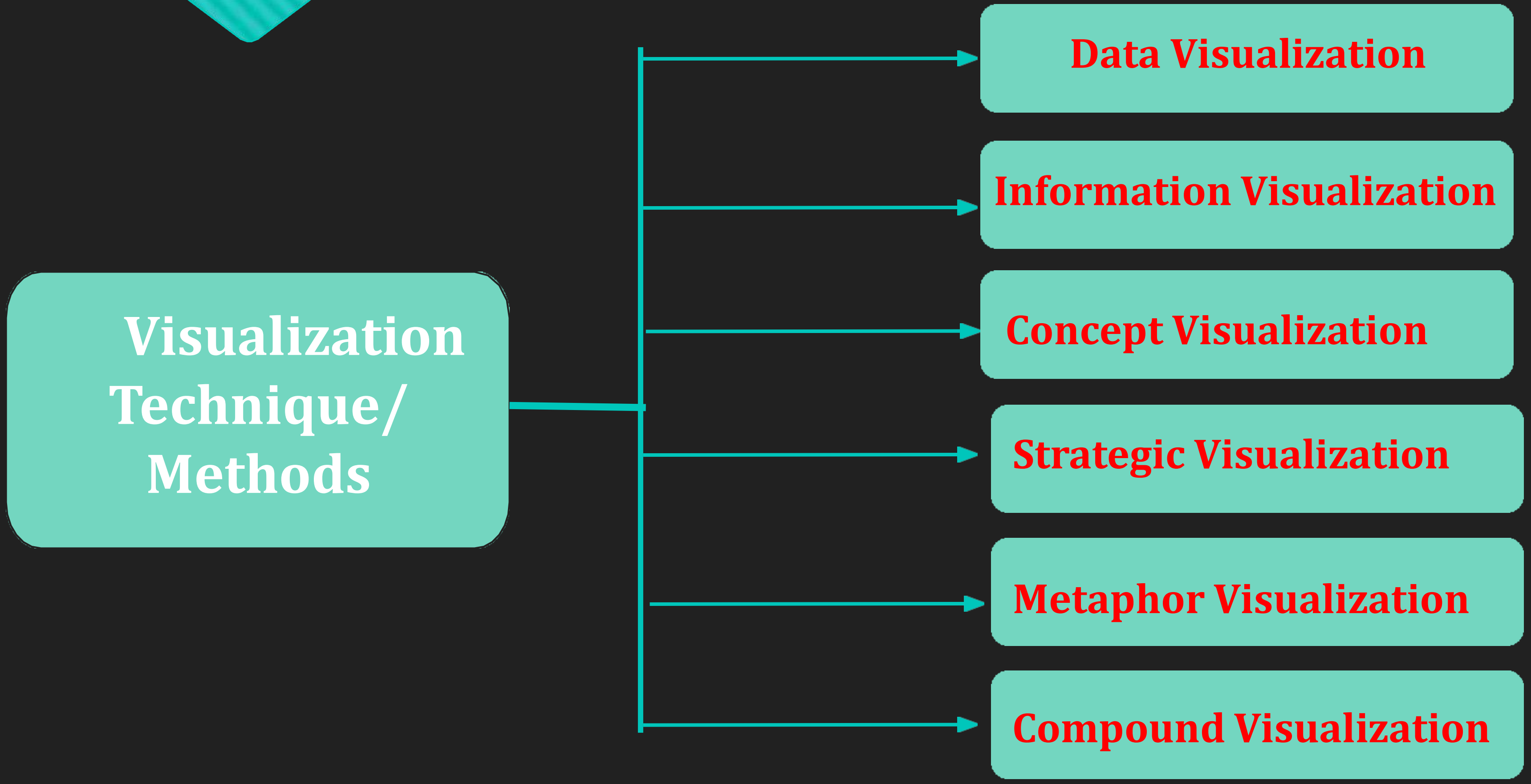
Conventional data visualization tools

- **Cone tree** is another method displaying hierarchical data such as organizational body in three dimensions.
- The branches grow in the form of cone.
- A semantic network is a graphical representation of logical relationship between different concepts.
- It generates directed graph, the combination of nodes or vertices, edges or arcs, and label over each edge.





Techniques for visual data representations



Visualization Technique/ Methods

Data Visualization

Information Visualization

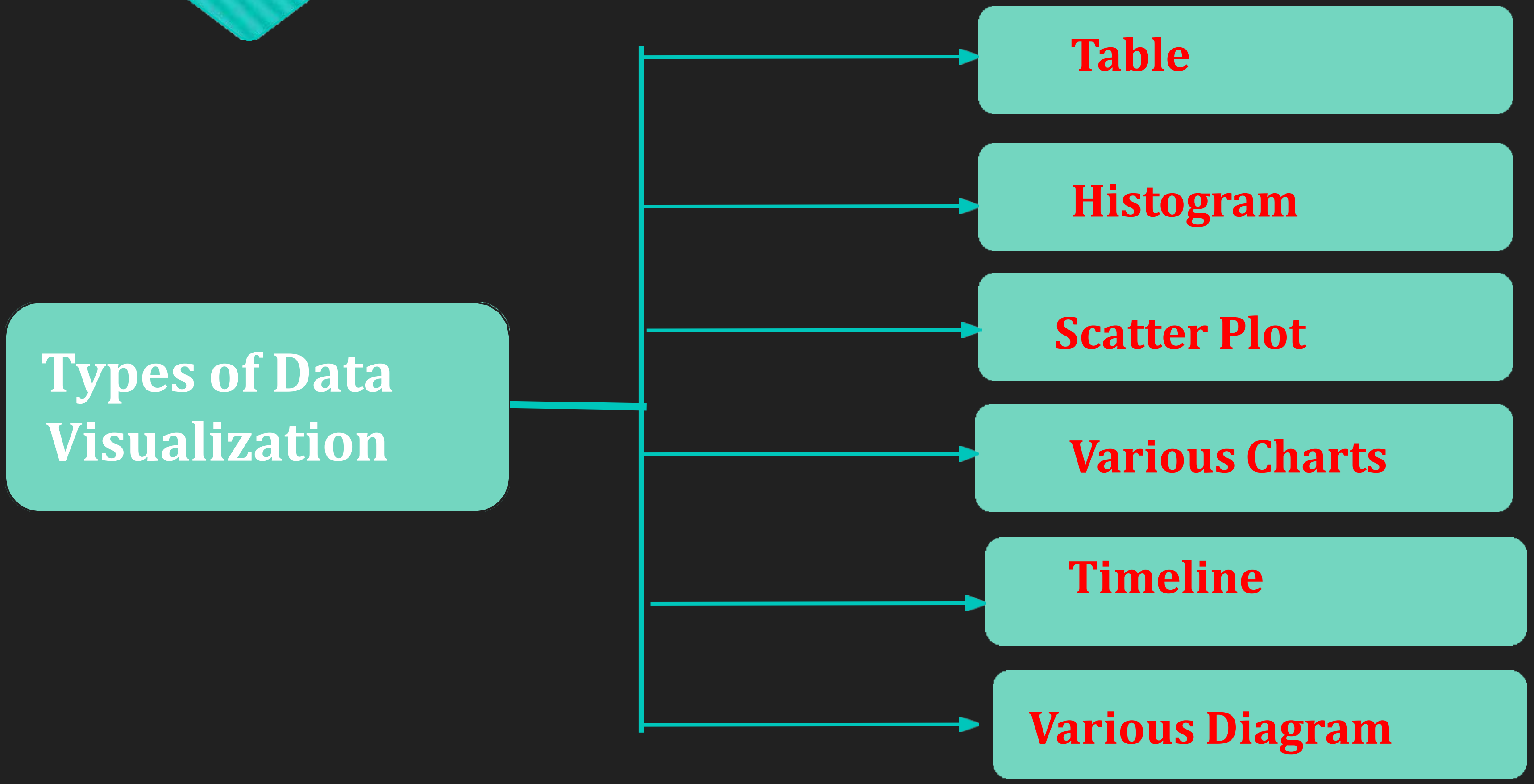
Concept Visualization

Strategic Visualization

Metaphor Visualization

Compound Visualization

Types of data visualization



Types of Data Visualization

Table

Histogram

Scatter Plot

Various Charts

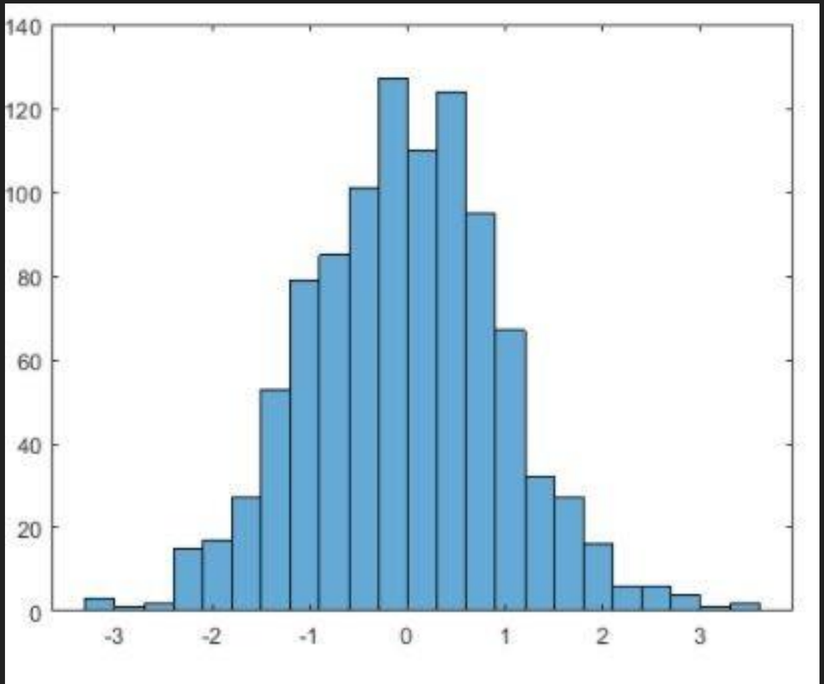
Timeline

Various Diagram

Types of data visualization- Table

Eid	Ename	Post	Salary
101	Kunal	MAnager	95000
102	Ishita	Project Leader	85000
103	Diksha	Tester	35000

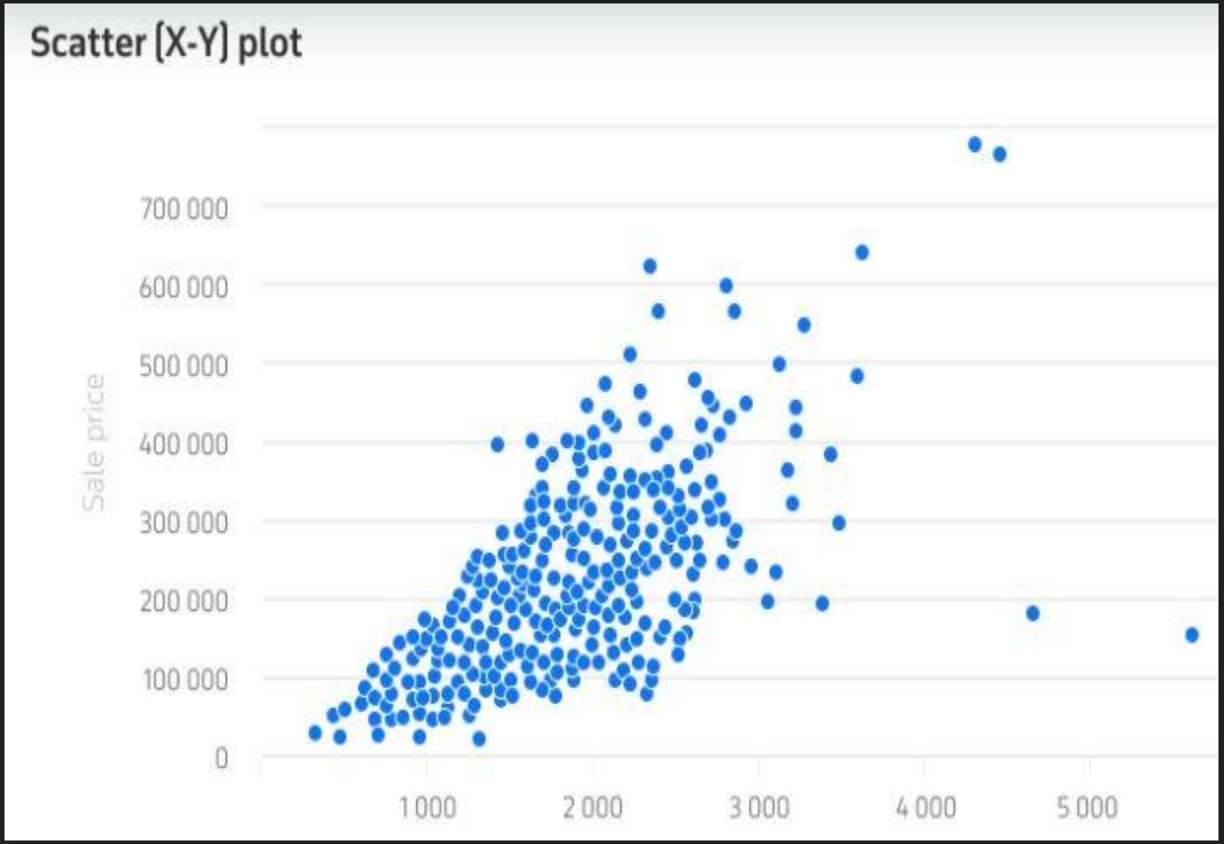
Types of data visualization- Histogram



- An approximate representation of the distribution of numerical data. Divide the entire range of values into a series of intervals and then count how many values fall into each interval this is called binning.

- For example, determining frequency of annual stock market percentage returns within particular ranges (bins) such as 0-10%, 11-20%, etc. The height of the bar represents the number of observations (years) with a return % in the range represented by the respective bin.

Types of data visualization- Scatter Plot

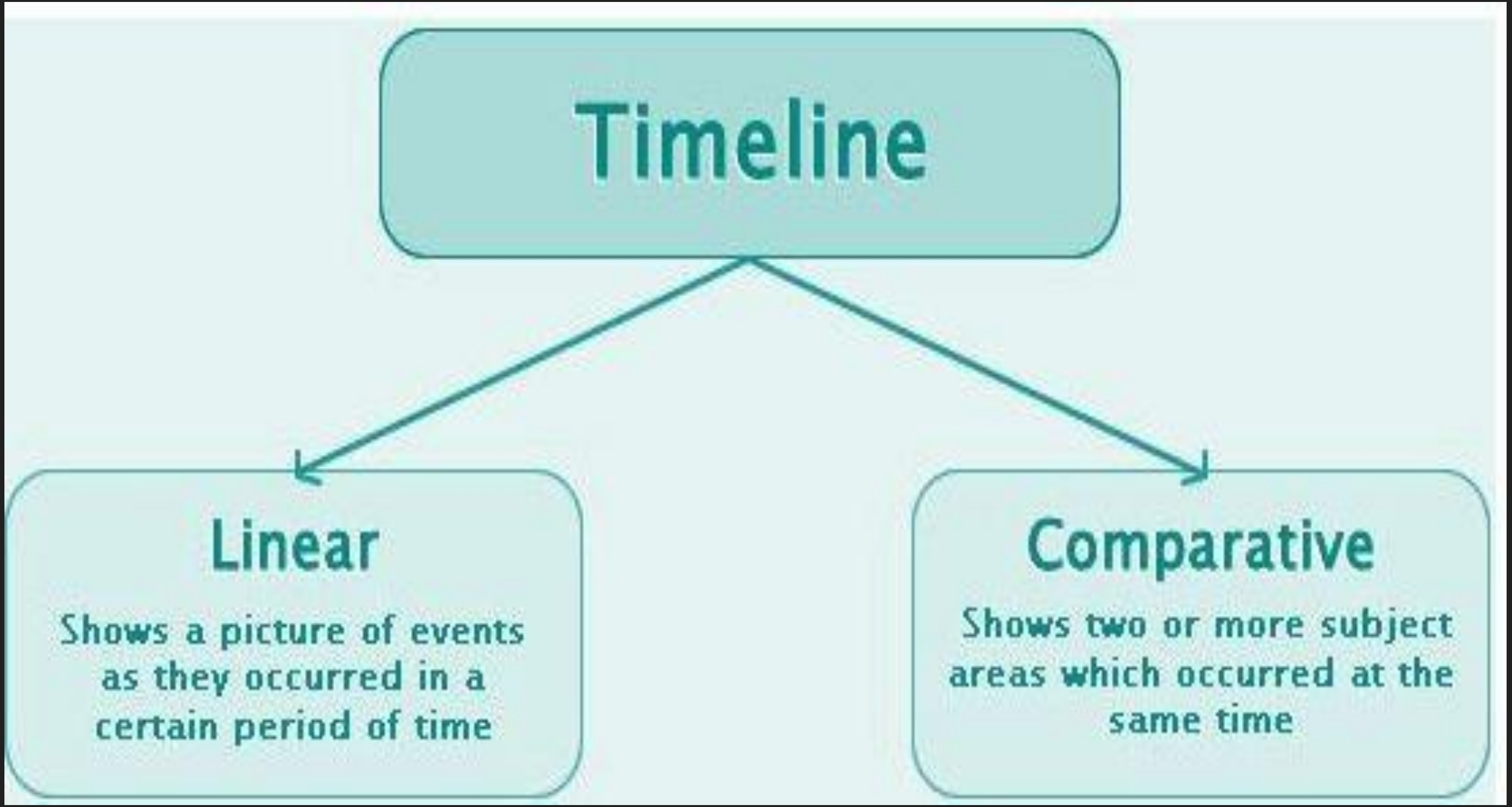


When you have multiple data points and need to examine the correlation between X and Y variables. Consequently, variables should depend on each other or influence each other in some way. For example, supply is usually related to demand.

Types of data visualization- Timeline

- A **timeline chart** is an effective way to visualize a process using chronological order.

- Since details are displayed graphically, important points in time can be easy seen and understood.



Timeline

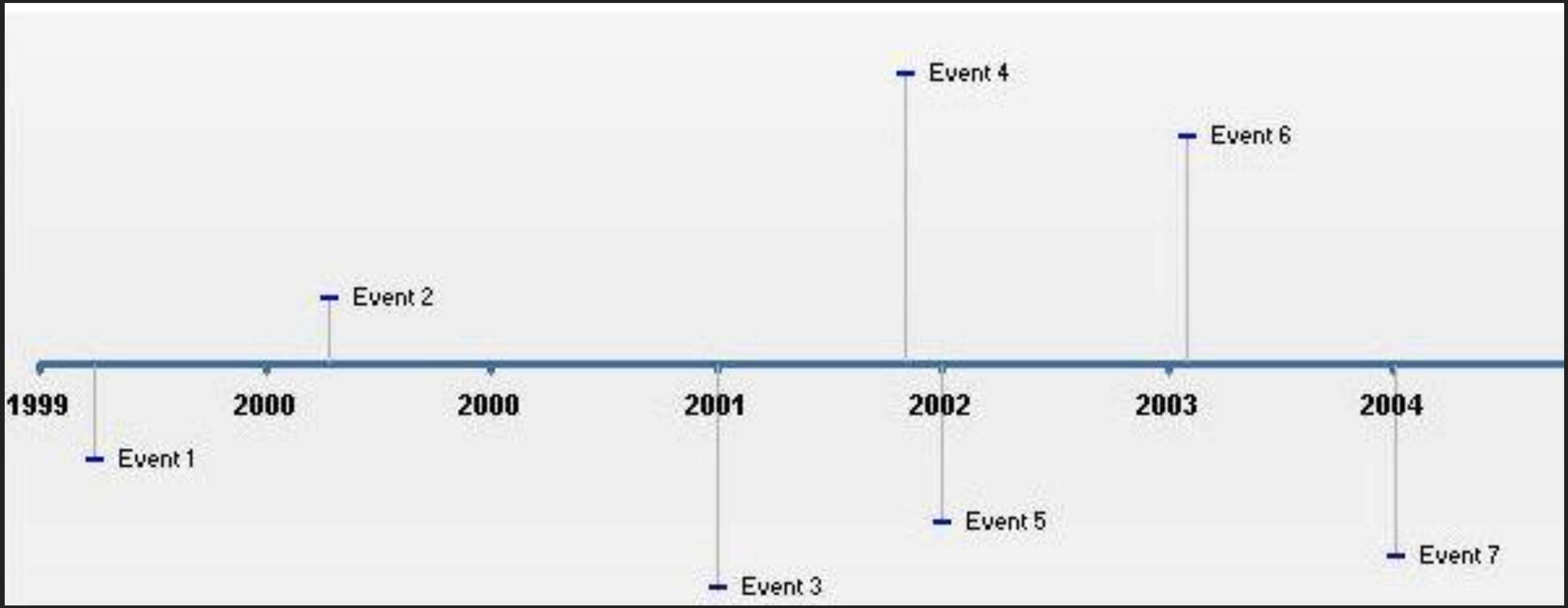
Linear

Shows a picture of events as they occurred in a certain period of time

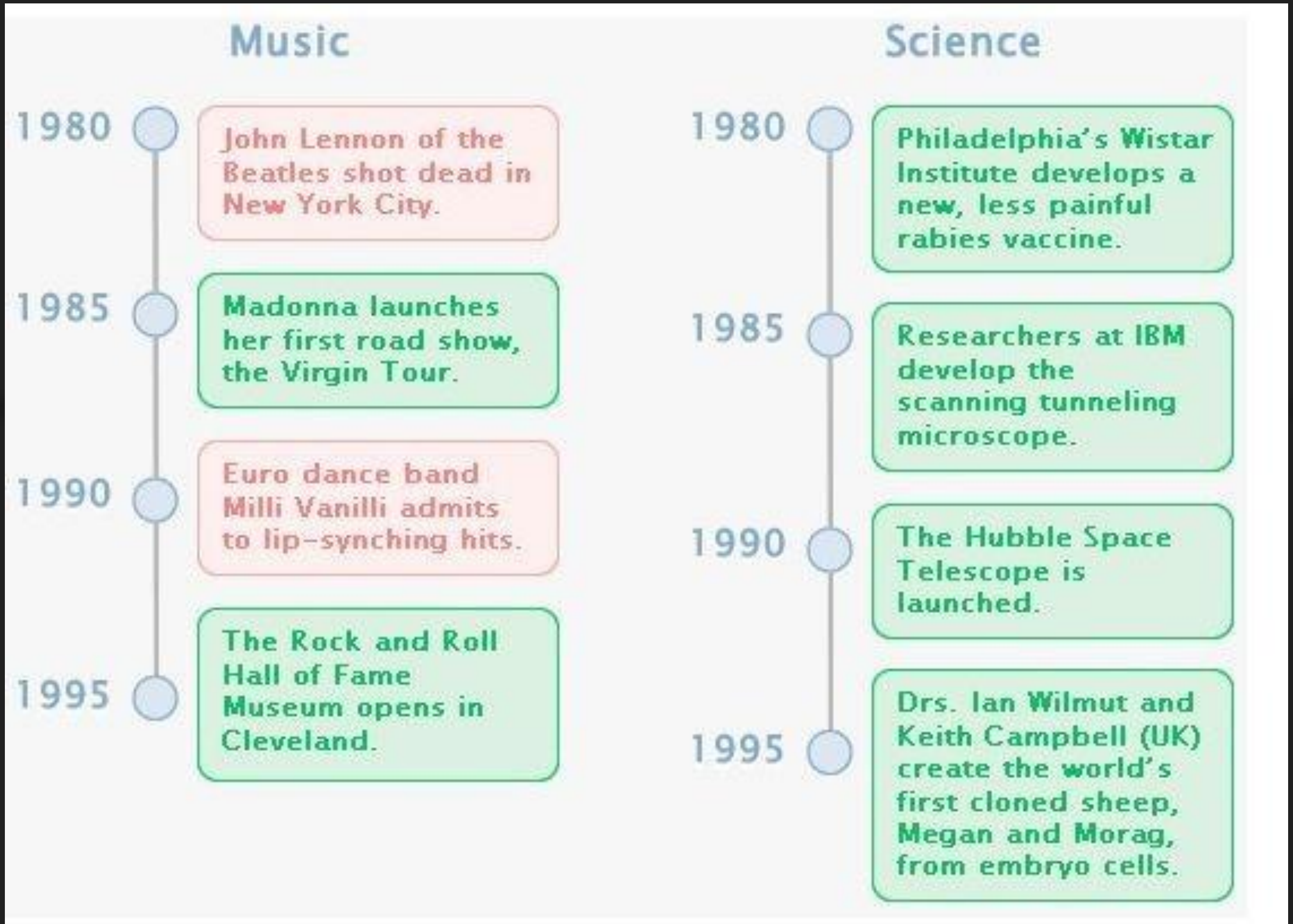
Comparative

Shows two or more subject areas which occurred at the same time

Types of data visualization- Timeline



Types of data visualization- Timeline



Music

1980 John Lennon of the Beatles shot dead in New York City.

1985 Madonna launches her first road show, the Virgin Tour.

1990 Euro dance band Milli Vanilli admits to lip-synching hits.

1995 The Rock and Roll Hall of Fame Museum opens in Cleveland.

Science

1980 Philadelphia's Wistar Institute develops a new, less painful rabies vaccine.

1985 Researchers at IBM develop the scanning tunneling microscope.

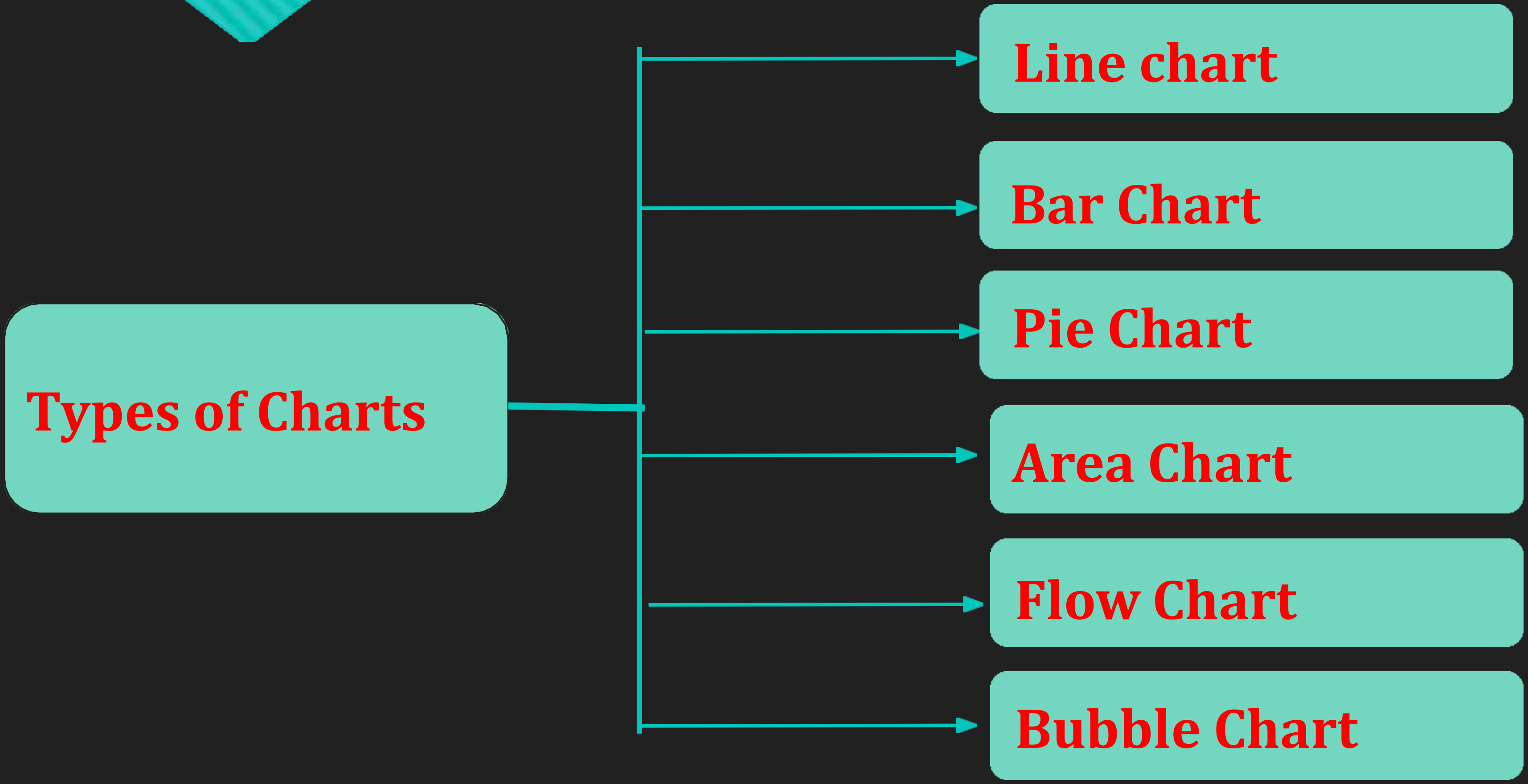
1990 The Hubble Space Telescope is launched.

1995 Drs. Ian Wilmut and Keith Campbell (UK) create the world's first cloned sheep, Megan and Morag, from embryo cells.

https://www.upsid cs-designing-a-tim eline/

Types of data visualization-

Various Charts



Types of Charts

Line chart

Bar Chart

Pie Chart

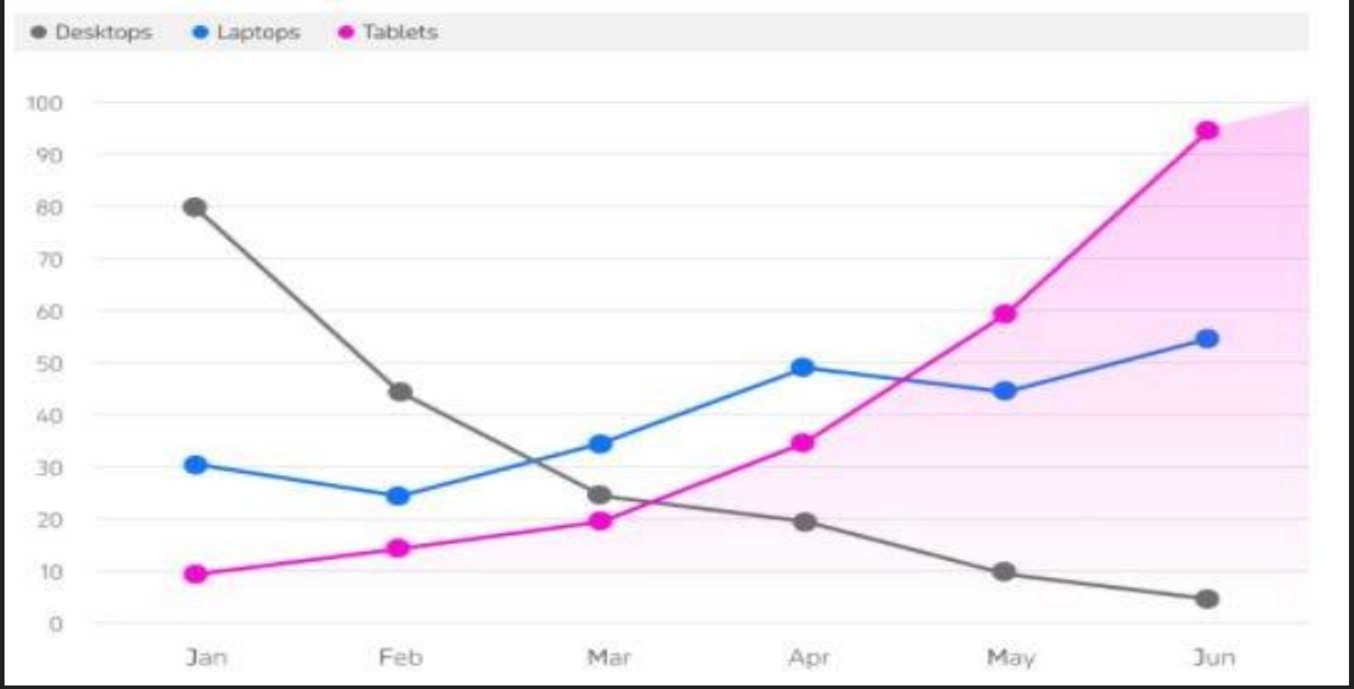
Area Chart

Flow Chart

Bubble Chart

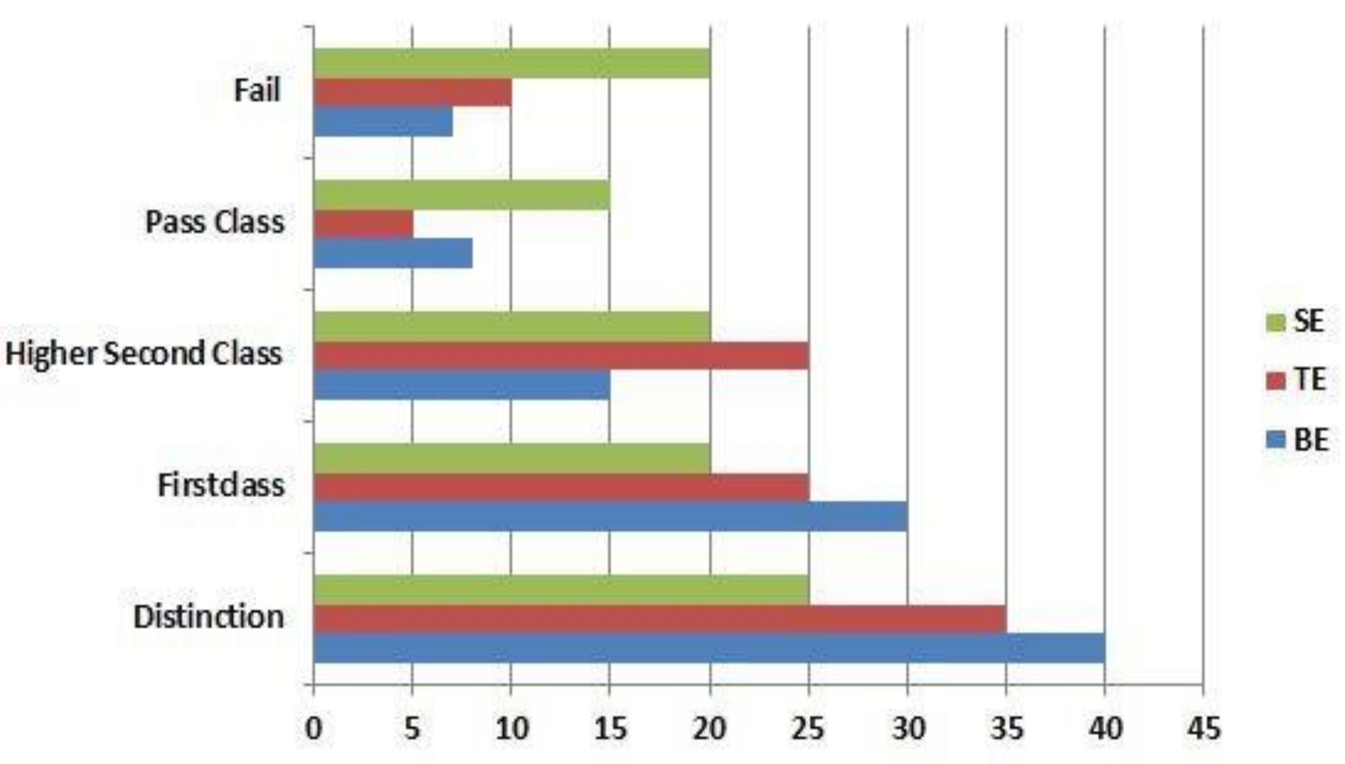
Types of data visualization-

Line Chart



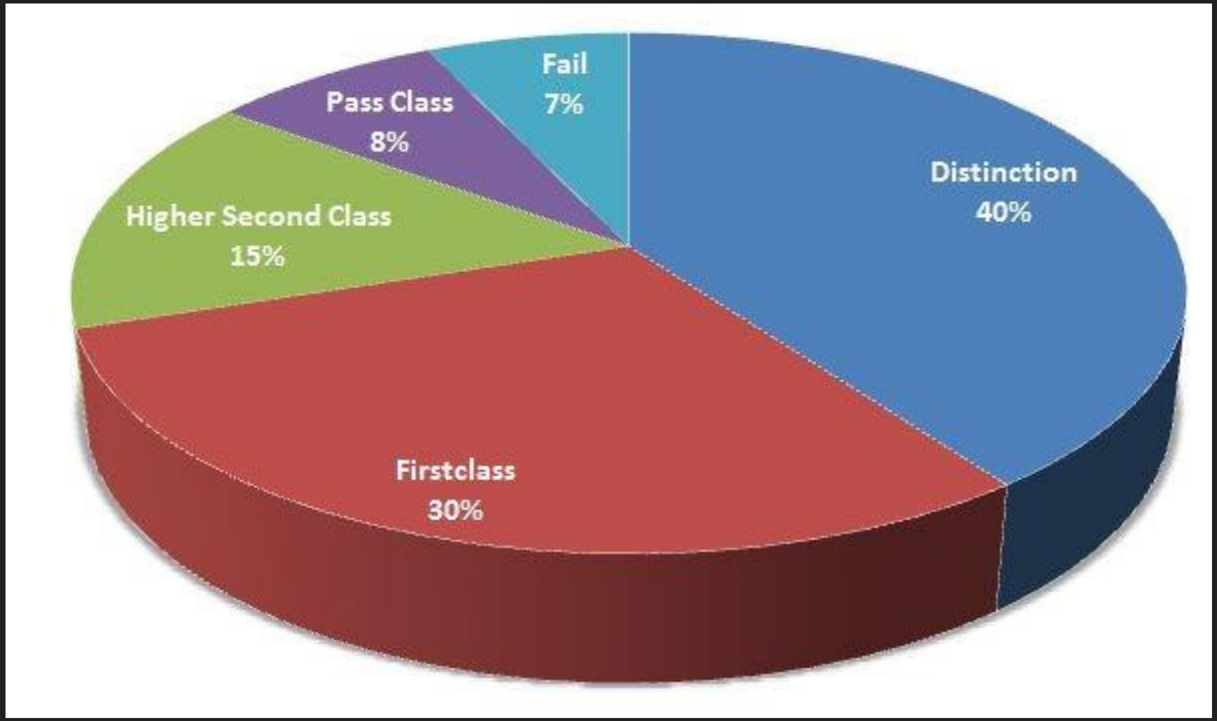
Types of data visualization-

Bar Chart



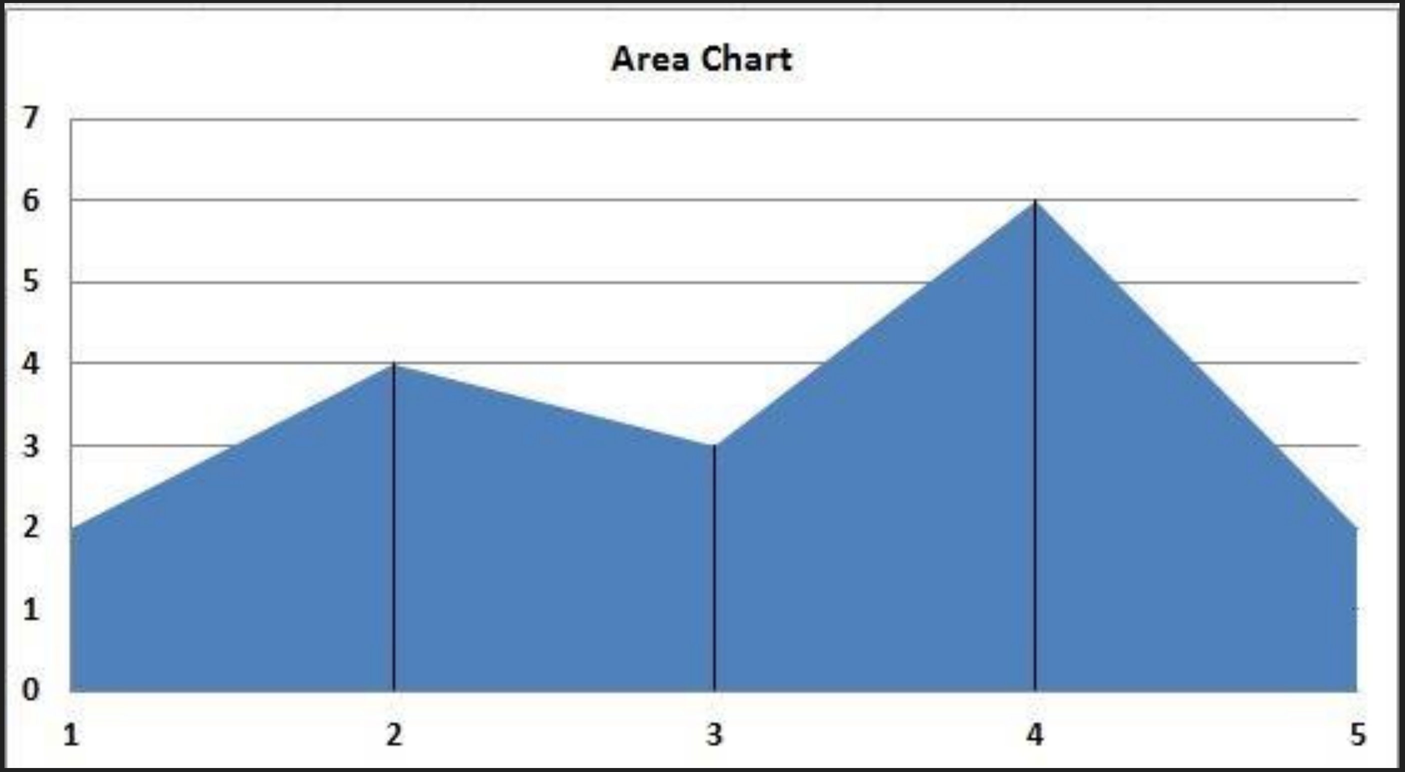
Types of data visualization-

Pie Chart

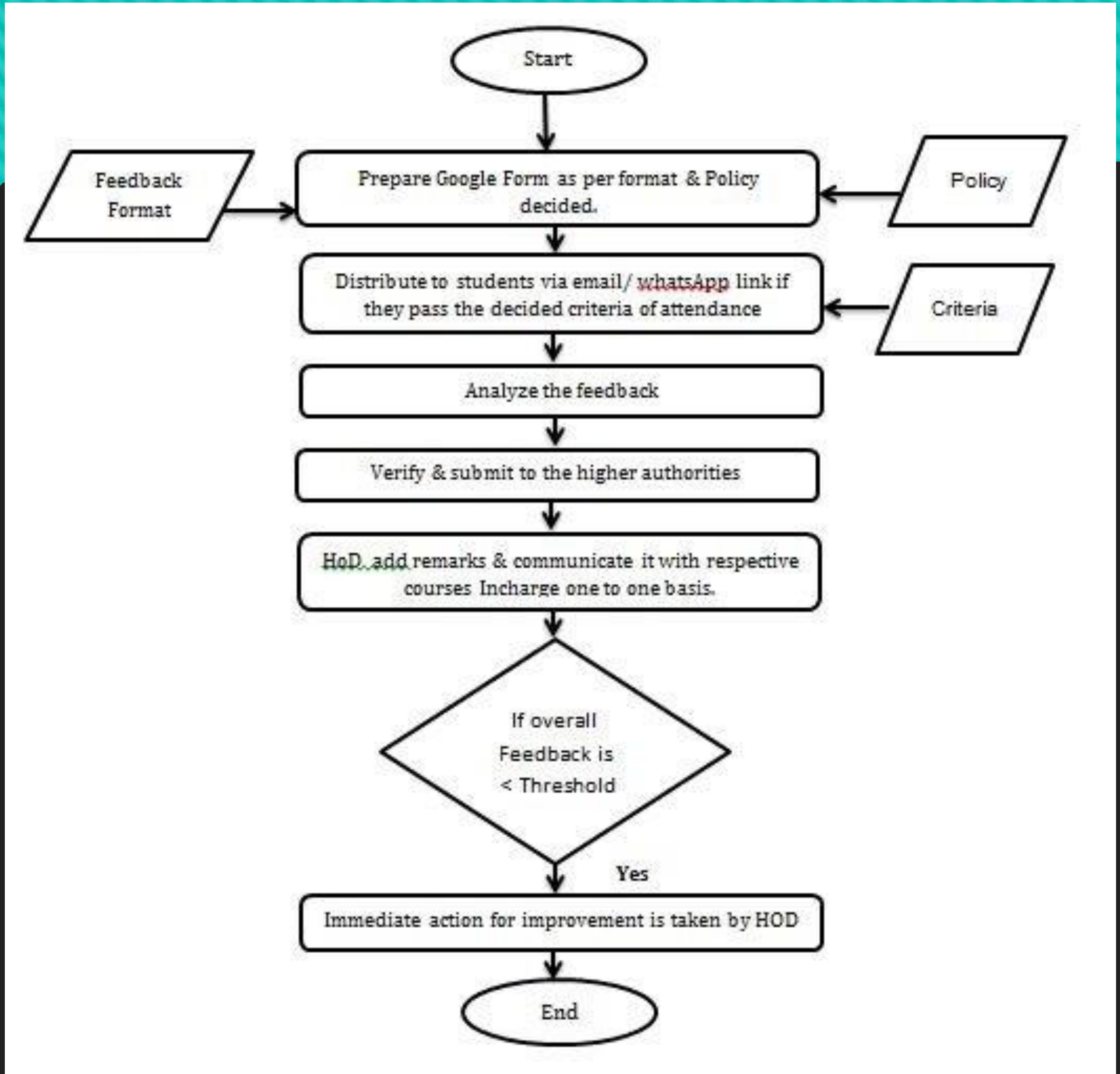


Types of data visualization-

Area Chart



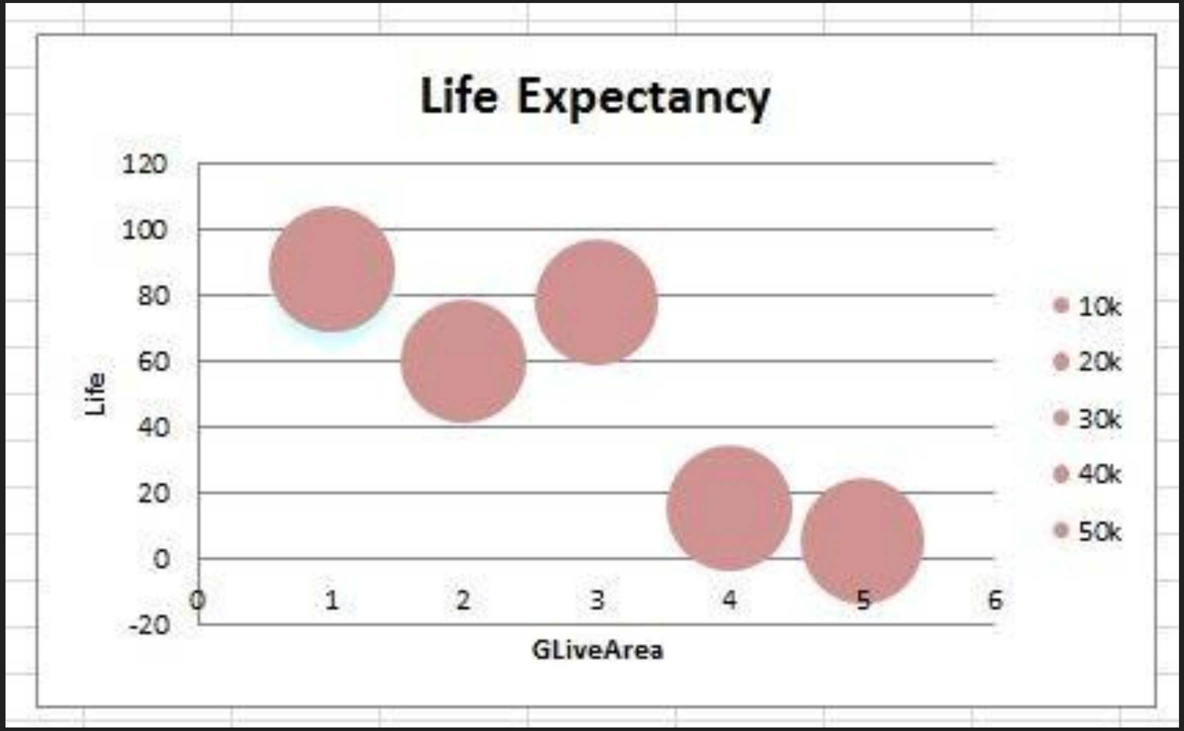
Types of data visualization-Flow Chart



Start
Feedback Format
Prepare Google Form as per format & Policy decided.
Policy
Distribute to students via email/ whatsApp link if they pass the decided criteria of attendance
Criteria
Analyze the feedback
Verify & submit to the higher authorities
HoD add remarks & communicate it with respective courses Incharge one to one basis.
If overall Feedback is < Threshold
Yes
Immediate action for improvement is taken by HOD
End

Types of data visualization-

Bubble Chart



Visualizing Big Data

- Big Data visualization involves the presentation of data of almost any type in a graphical format that makes it easy to understand and interpret.
- But it goes far beyond typical corporate graphs, histograms and pie charts to more complex representations like *heat maps* and *fever charts*, enabling decision makers to explore data sets to identify correlations or unexpected patterns.
- A defining feature of Big Data visualization is scale.
- Today's enterprises collect and store vast amounts of data that would take years for a human to read, let alone understand.
- But researchers have determined that the human retina can transmit data to the brain at a rate of about 10 megabits per second.
- Big Data visualization relies on powerful computer systems to ingest raw corporate data and process it to generate graphical representations that allow humans to take in and understand vast amounts of data in seconds.

Tools used in data visualization

1. **Google charts-** A powerful, easy to use and an interactive data visualization tool for browsers and mobile devices. It has a rich gallery of charts and allows you to customize as per your needs. Rendering of charts is based on HTML5/SVG technology.

Pros:

Simple to learn and user-friendly.

Fast & accurate, Highly interactive.

Completely free, Interactive dashboard.

Cross-platform portability with any additional plugins. Supports iPhone, iPad & Android.

Can read from multiple data sources – Excel, SQL databases, CSV, Google Spreadsheets, etc.

Open-Source/Licensed: Open Source Chart Library.

Cons:

A network connection is mandatory while using this tool.

Lacks demos on advanced features.

Working with API for complex presentations is sometimes difficult to learn.

Lacks sophisticated statistical processing.

Tools used in data visualization

2. Tableau : A business intelligence tool that aids people in visualizing and understanding their data. It is widely used in the field of Business intelligence. It allows you to design interactive graphs and charts in the shape of dashboards and worksheets to obtain business visions.

Open-Source/Licensed: Licensed. It has a free trial available.

Pros:

- Outstanding visualization capabilities.
- Easy to use,Mobile friendly.
- Connectivity to multiple data sources.
- Healthy community and forum.
- Powerful computation, Quick insights.

Cons:

- Very costly and has inflexible pricing.
- No option for scheduling and auto-refresh of reports.
- Restrictive visual imports.
- Static parameters that need to be updated manually each time when the data gets modified are present.
- Column table formatting is difficult.

Tools used in data visualization

3. Sisense: Provides instant insights for anyone, anywhere in your organization. It allows you to create visual dashboards and reports to state any piece of data, uncover underlying trends & patterns and make data-driven decisions.

Open-Source/Licensed: Licensed. It has a free trial available.

Pros:

It has a very friendly user interface.
Great analysis performance on huge datasets.
Excellent support
Easy upgrades
Integrates very well with different data sources.
This product is very flexible and allows for easy customization.

Cons:

Difficult to maintain and develop analytic cubes.
It does not have any inbuilt data type to support the time format.
Limited type of visualizations.
If a cube rebuild is required then the cube becomes inaccessible during that period.

Tools used in data visualization

4.Datawrapper is increasingly becoming a popular choice, particularly among media organizations which frequently use it to create charts and present statistics. It has a simple, clear interface that makes it very easy to upload csv data and create straightforward charts, and also maps, that can quickly be embedded into reports.

5.Chartio- This is a cloud-based analytics platform that offers interactive dashboards, beautiful charts, and data exploration features.

It does not require any SQL knowledge.

6.IBM Watson Analytics:It provides automatic data visualization which aids in figuring out patterns, trends and complex relationships in business data.

Tools used in data visualization

Open Source Data Visualization tools

- Candela
- Charted
- Chart JS
- D3.js
- Datawrapper
- Dygraphs
- Leaflet
- RAW Graphs

Data visualization with Tableau

- Tableau is a Business Intelligence tool for visually analyzing the data.

- Data visualization is the process of describing information through visual rendering.

Tableau-PRODUCTS OF TABLEAU

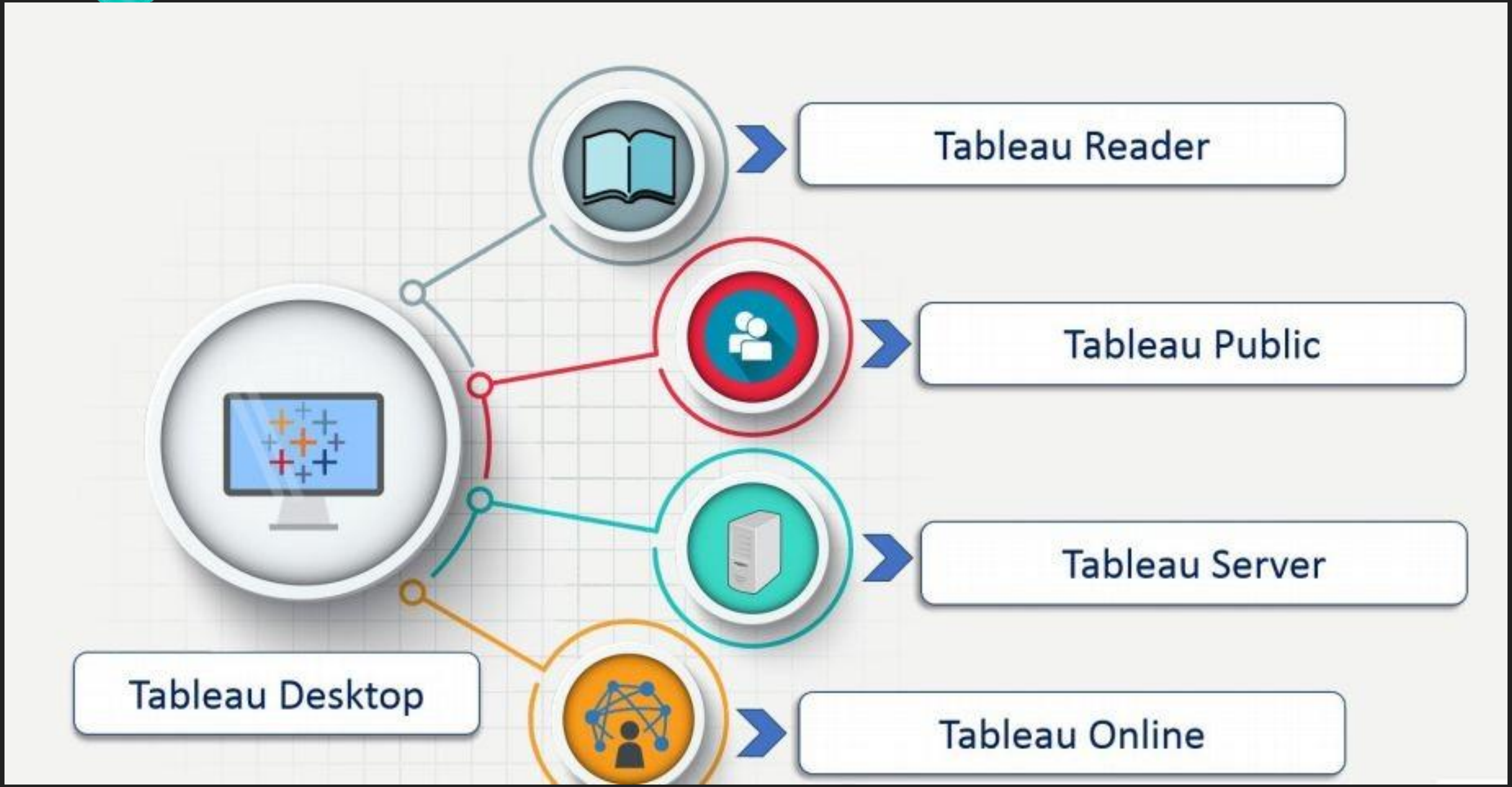
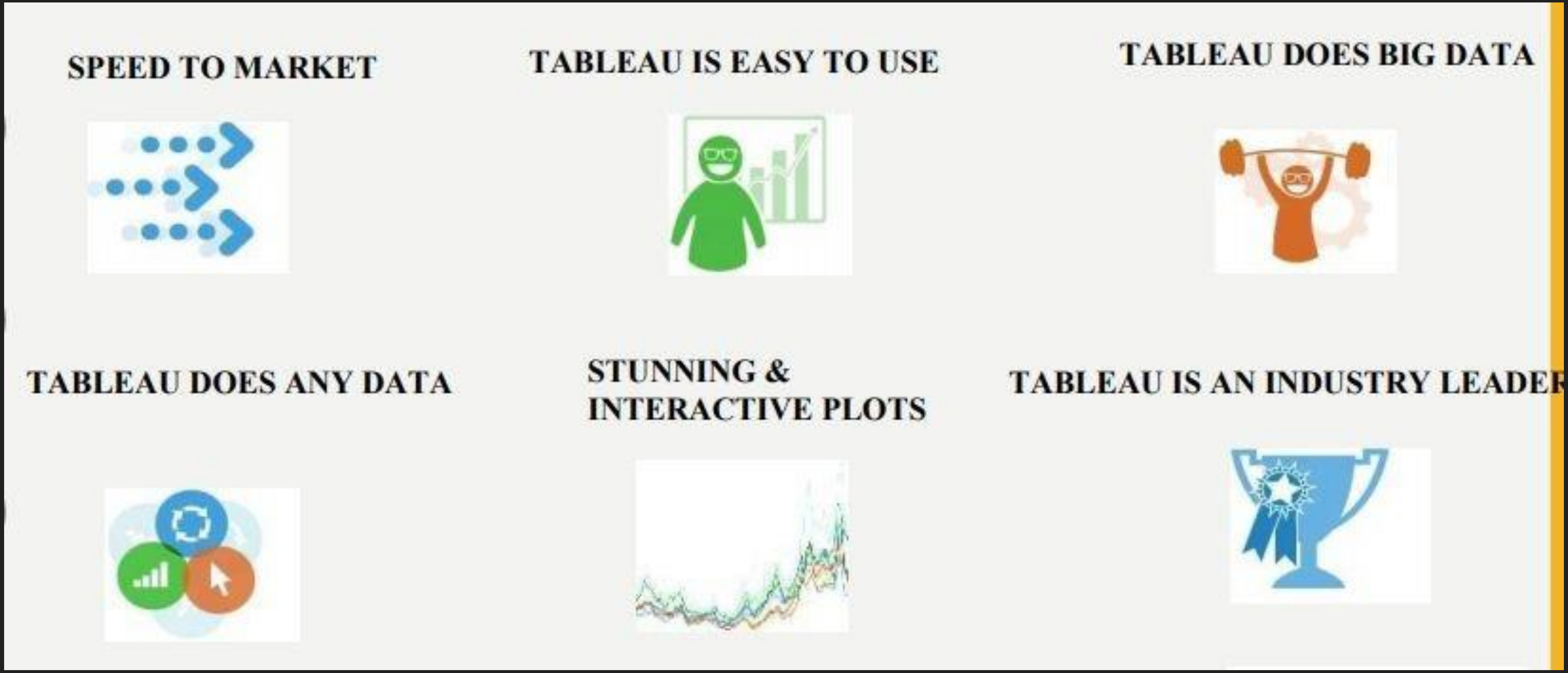


Tableau-FEATURES



SPEED TO MARKET

TABLEAU IS EASY TO USE

TABLEAU DOES BIG DATA

TABLEAU DOES ANY DATA

STUNNING & INTERACTIVE PLOTS

TABLEAU IS AN INDUSTRY LEADER

Tableau-Other FEATURES

1.Rapidly analyze data

2.Blend Diverse Data Sets

3.User-friendly Dashboards

4.Numeric Calculation

5.Creating your own

Tableau-THE EASY SEVEN STEPS

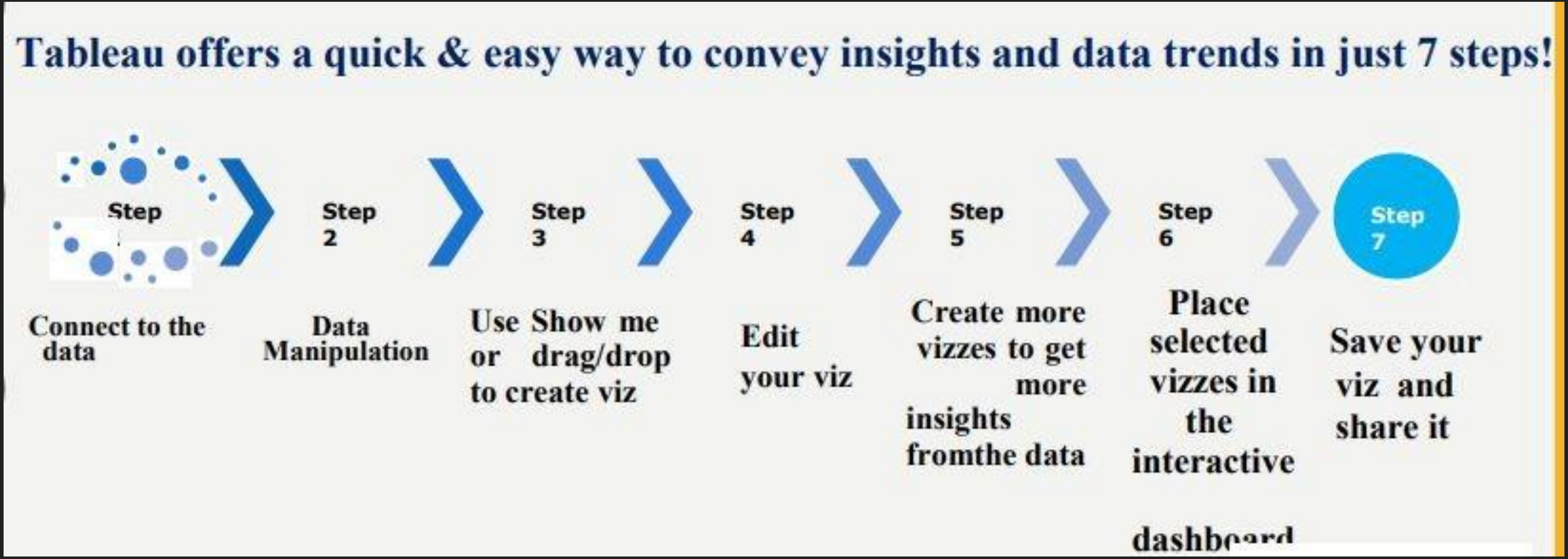


Tableau offers a quick & easy way to convey insights and data trends in just 7 steps!

Step – Connect to the data

Step 2 – Data Manipulation

Step 3 – Use Show me or drag/drop to create viz

Step 4 – Edit your viz

Step 5 – Create more vizzes to get more insights fromthe data

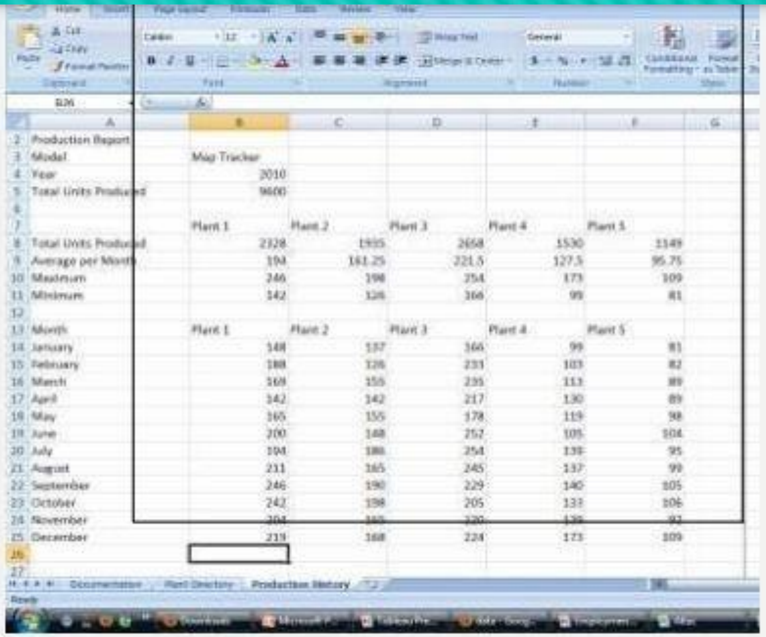
Step 6 – Place selected vizzes in the interactive dashboard

Step 7 – Save your viz and share it

Tableau-CONNECTING TO DATA

ORACLE
PostgreSQL
& etc
VERTICA
TERADATA
NETEZZA

Tableau-Converting Data into Visual Formats



	A	B	C	D	E	F
2	Production Report					
3	Model	Map Tracker				
4	Year	2010				
5	Total Units Produced	9600				
6						
7		Plant 1	Plant 2	Plant 3	Plant 4	Plant 5
8	Total Units Produced	2328	1935	2658	1530	1149
9	Average per Month	194	161.25	221.5	127.5	95.75
10	Maximum	246	198	254	173	109
11	Minimum	142	126	166	99	81
12						
13	Month	Plant 1	Plant 2	Plant 3	Plant 4	Plant 5
14	January	148	137	166	99	81
15	February	188	126	233	103	82
16	March	169	155	235	113	89
17	April	142	142	217	130	89
18	May	165	155	178	119	98
19	June	200	148	252	105	104
20	July	194	186	254	139	95
21	August	211	165	245	137	99
22	September	246	190	229	140	105
23	October	242	198	205	133	106
24	November	204	169	220	139	92
25	December	219	168	224	173	109
26						
27						

Converting data into visual format

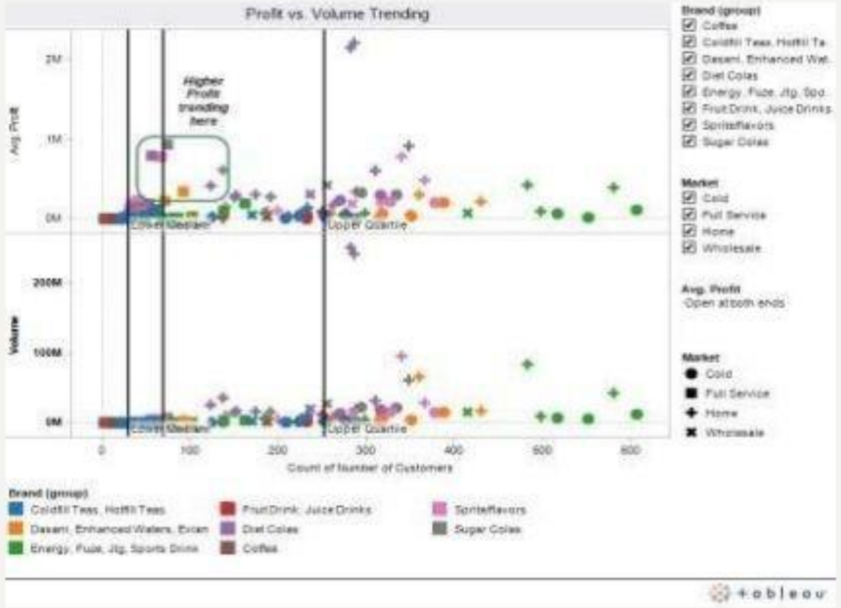


tableau SOFTWARE

Tableau-Advantages

- Data visualization
- Drag-drop functionalities
- Ease of implementation – Easy to learn
- Can handle large amount of data – With no compromise on performance
- Multi-device support – Responsive customization

Tableau-Disadvantages

- No option of scheduling
- No custom visual imports
- Poor custom formatting –Limits to 16 columns table –Time consuming manual formatting
- Manual parameter updating
- Poor handling of screen resolution
- Expensive!