Unit-6 : Framework & Visualizations

Big Data (CS-3032)

Kalinga Institute of Industrial Technology
Deemed to be University
Bhubaneswar-751024

School of Computer Engineering

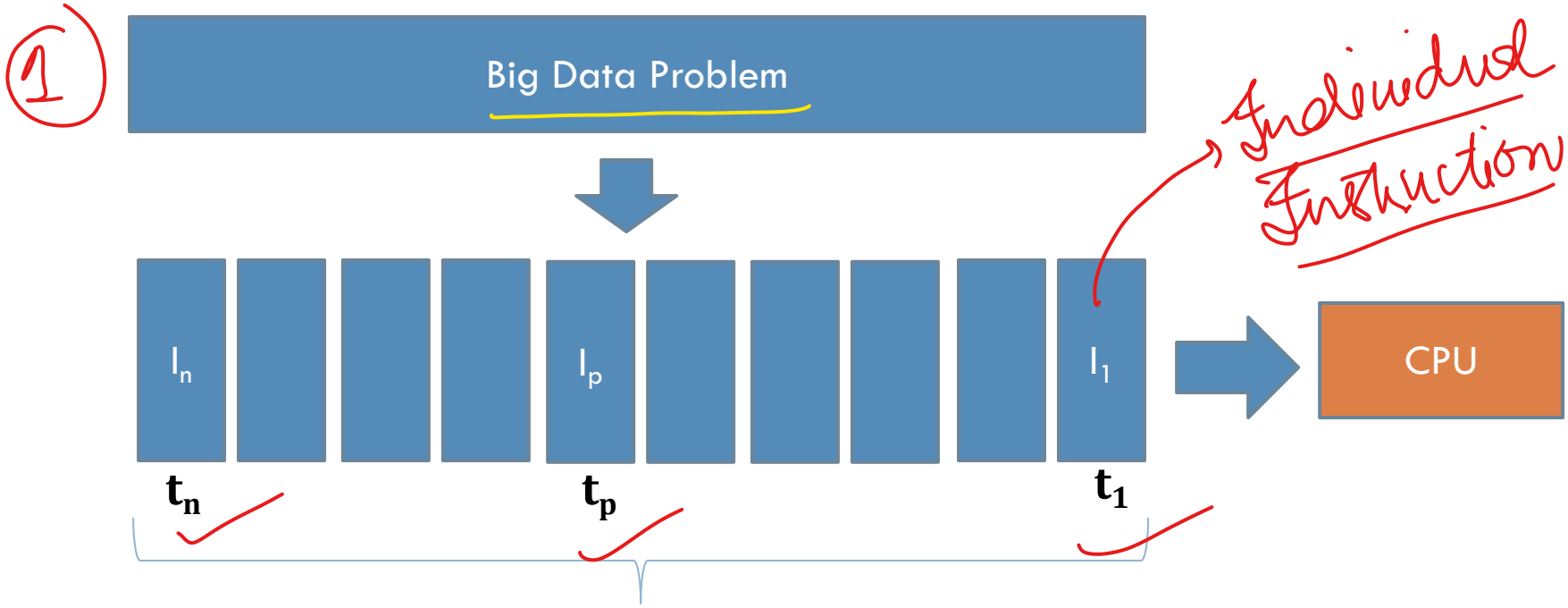
3 Credit

Lecture Note

Course Contents

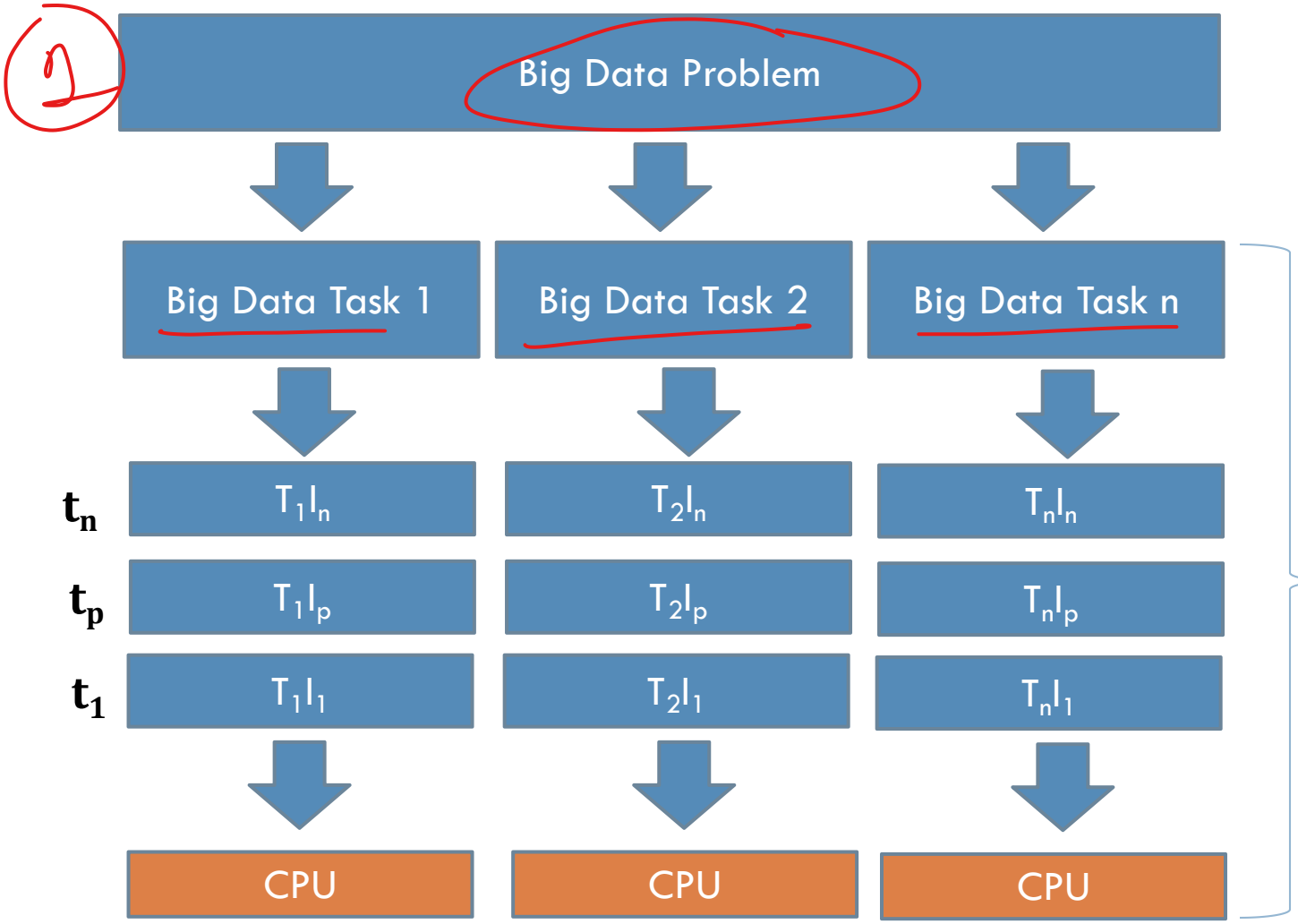
Sr #	Major and Detailed Coverage Area	Hrs
5	**Frameworks And Visualization**	6
	Distributed and Parallel Computing for Big Data, Big Data Visualizations – Visual data analysis techniques, interaction techniques, applications	

Sequential Computing in Big Data



Big Data Instructions processed by a CPU at different time slot

Parallel Computing in Big Data



Big Data instructions processed by multiple CPU at different time slot in a single machine

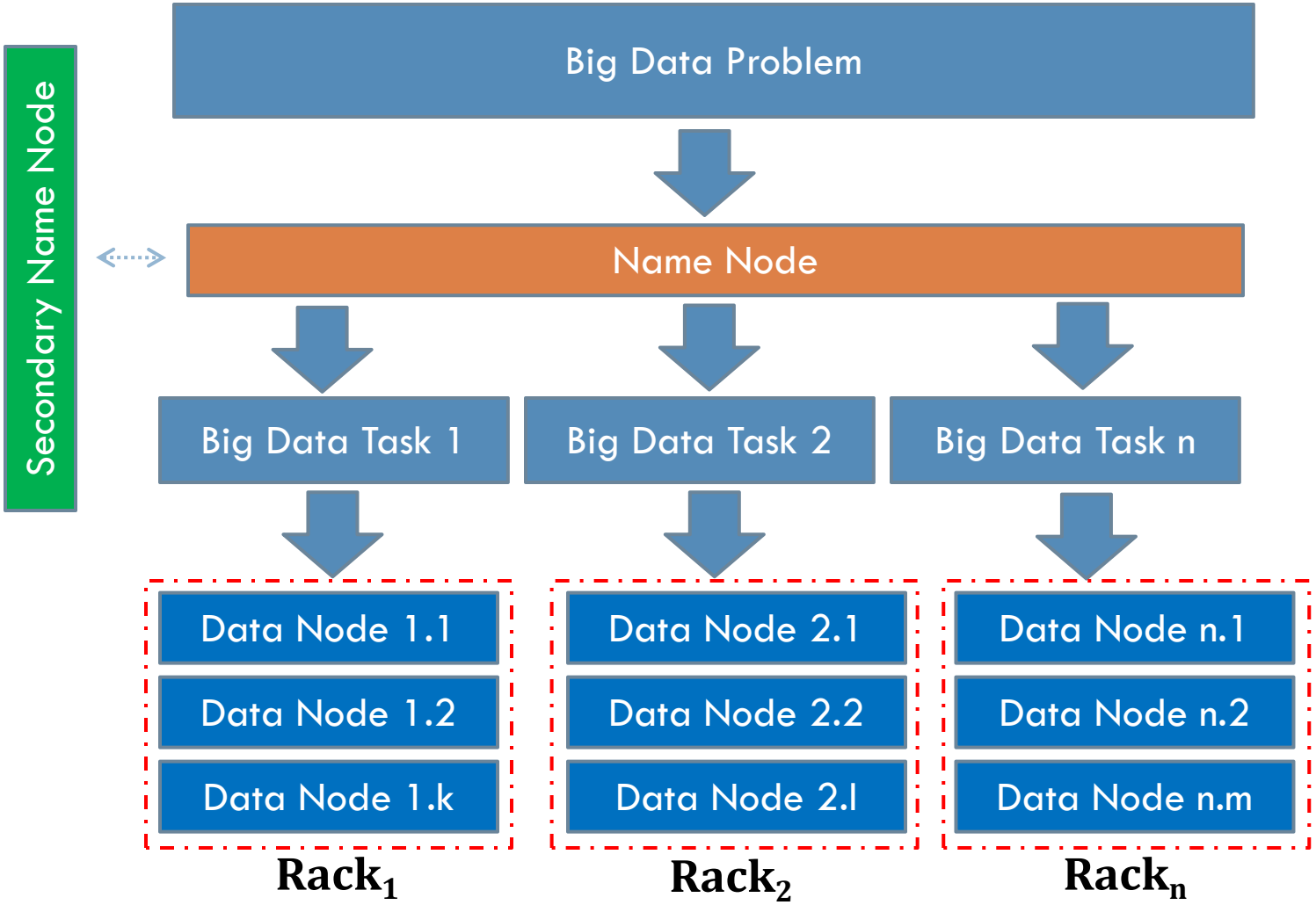
Difference between Sequential Computing and Parallel Computing

Sequential Computing	Parallel Computing
All the instructions are executed in a sequence, one at a time.	All the instructions are executed parallely.
It has a single processor.	It is having multiple processors.
It has low performance and the workload of the processor is high due to the single processor.	It has high performance and the workload of the processor is low because multiple processors are working simultaneously.
Bit-by-bit format is used for data transfer.	Data transfers are in bytes.
It requires more time to complete the whole process.	It requires less time to complete the whole process.
Cost is low	Cost is high

Class Work

- What kind of tasks can be split up into separate threads while using parallel processing?
- What are the advantages and disadvantages of using parallel computation?
- What is the role of memory in parallel processing?
- Give some examples of real-world applications that use parallel processing?
- What is the maximum number of threads that can exist at any given moment during the execution of a program?
- How parallel processing improves computational efficiency?
- How would you design a parallel algorithm for a given problem?

Distributed Computing in Big Data



Difference between Parallel Computing and Distributed Computing

Parallel Computing	Distributed Computing
Single computer is required	Uses multiple computers
Multiple processors perform multiple operations	Multiple computers perform multiple operations
Improves the system performance	Improves system scalability, fault tolerance and resource sharing capabilities
Processors communicate with each other through bus	Computer communicate with each other through message passing.

Distributed and Parallel Computing in Big Data

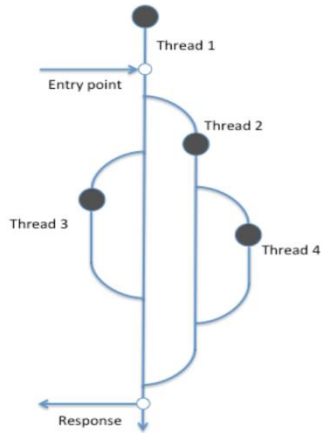
- Multiple computing resources are connected in a network and computing tasks are distributed across the resources. The sharing of tasks increases the speed as well as the efficiency of the system. Because of such reason, the distributed computing is more efficient than traditional methods of computing. It is primarily suitable to process huge amounts of data in a limited time.
- Another way to improve the processing capability of the system is to add additional computational resources to it. This will help in dividing complex computations into subtasks, which can be handled individually by processing units that are running in parallel. Such system are called parallel systems in which multiple parallel computing resources are involved to carry out computation simultaneously.
- Organization use both parallel and distributed computing techniques to process Big Data due to the important constraint for business, so called "Time".

Distributed and Parallel Computing in Big Data cont'd

- Sometimes, computing resources develop technical challenge, and fail to respond. Such situations can be handled by virtualization, where some processing and analytical tasks are delegated to other resources.
- Another problem that often hampers data storage and processing activities is latency. It is defined as the aggregate delay in the system because of delays in the completion of individual tasks. Such a delay automatically leads to the slowdown in system performance as whole and is often termed as system delay. It affects data management and communication within and across various business units thereby, affecting the productivity and profitability of the organizations. Implementing distributed and parallel computing methodologies helps in handling both latency and data-related problem.

Class Work

- The reasons to why a system should be built distributed, not just parallel are: **Scalability, Reliability, Data sharing, Resources sharing, Heterogeneity and modularity, Geographic construction, and Economic.** Explain each of the terms in details.
- Design considerations for distributed systems are: **No global clock, Geographical distribution, No shared memory, Independence and heterogeneity, Fail-over mechanism, and Security concerns.** Explain each of the terms.
- Explain the following image.



Visualization

Visualization is a pictorial representation technique. Anything which is represented in pictorial or graphical form, with the help of diagrams, charts, pictures, flowcharts etc. is known as visualization. Data visualization is a pictorial or visual representation of data with the help of visual ads such as graphs, bar, histograms, tables, pie charts, mind maps etc.

Ways of Representing Visual Data

The data is first analyzed and then the result is visualized. There are 2 ways to visualize a data, namely, Infographics and data visualization.

Infographics – It is the visual representation of information or data rapidly.

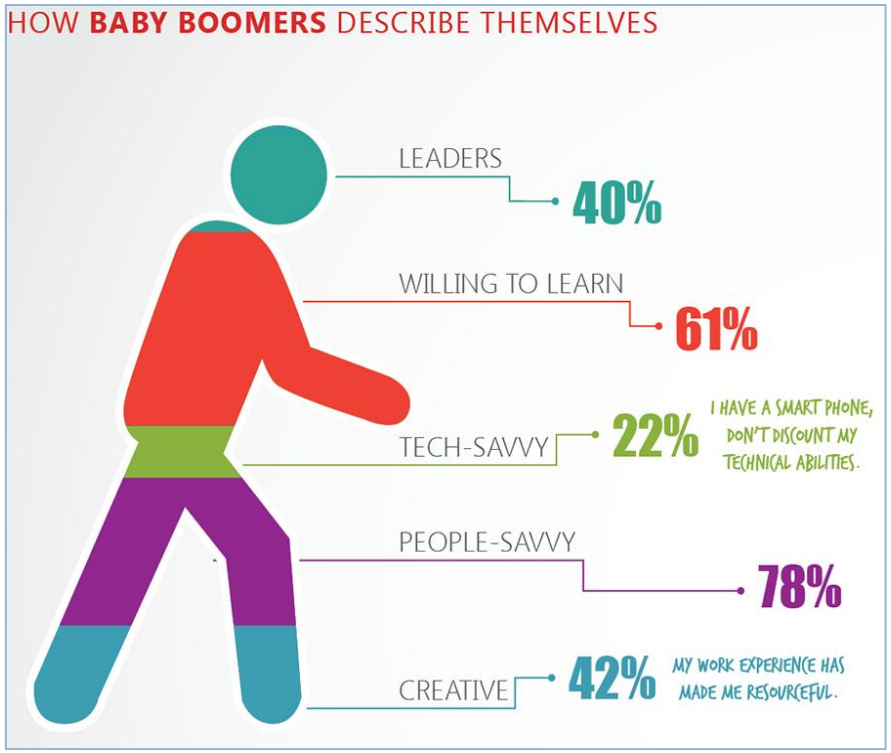
Data Visualization – It is the study of representing data or information in a visual form.

Difference: Infographics tell a premeditated story to guide the audience to conclusions (subjective). **Data visualizations** let the audience draw their own conclusions (objective).

An infographic can contain data visualizations but not the other way around.

Infographics vs. Data Visualization

Infographics



Data Visualization

Infographics vs. Data Visualization cont'd

Infographics are:

- Best for telling a premeditated story and offer subjectivity
- Best for guiding the audience to conclusions and point out relationships
- Created manually for one specific dataset
- It is used for Marketing content, Resumes, Blog posts, and Case studies etc.

Data visualizations are:

- Best for allowing the audience to draw their own conclusions, and offer objectivity
- Ideal for understanding data at a glance
- Automatically generated for arbitrary datasets
- It is used for Dashboards, Scorecards, Newsletters, Reports, and Editorials etc.

Class Work

- Design an Infographics narrating the advantages of online education.

Data Visualization Purpose

- Data presented in the form of graphics can be **analyzed better** than the data presented in words.
- Patterns, trends, outliers and correlations that **might go undetected in text-based data** can be **exposed and recognized easier** with data visualization software.
- Data scientists can use data visualizations to make their information **more actionable.** Illustrations, graphs, charts and spreadsheets can turn dull reports into something illuminating, where it's **easier to gather insight** and **actionable results**.
- Data Visualization help to **transmit a huge amount of information** to the human brain **at a glance**.
- Data Visualization **point out key or interesting breakthrough in a large dataset.**

Techniques Used for Big Data Visualization

Big data can be presented in various forms, which include simple line diagrams, bar graphs tables, metrics etc. Techniques used for a visual representation of the big data are as follows:

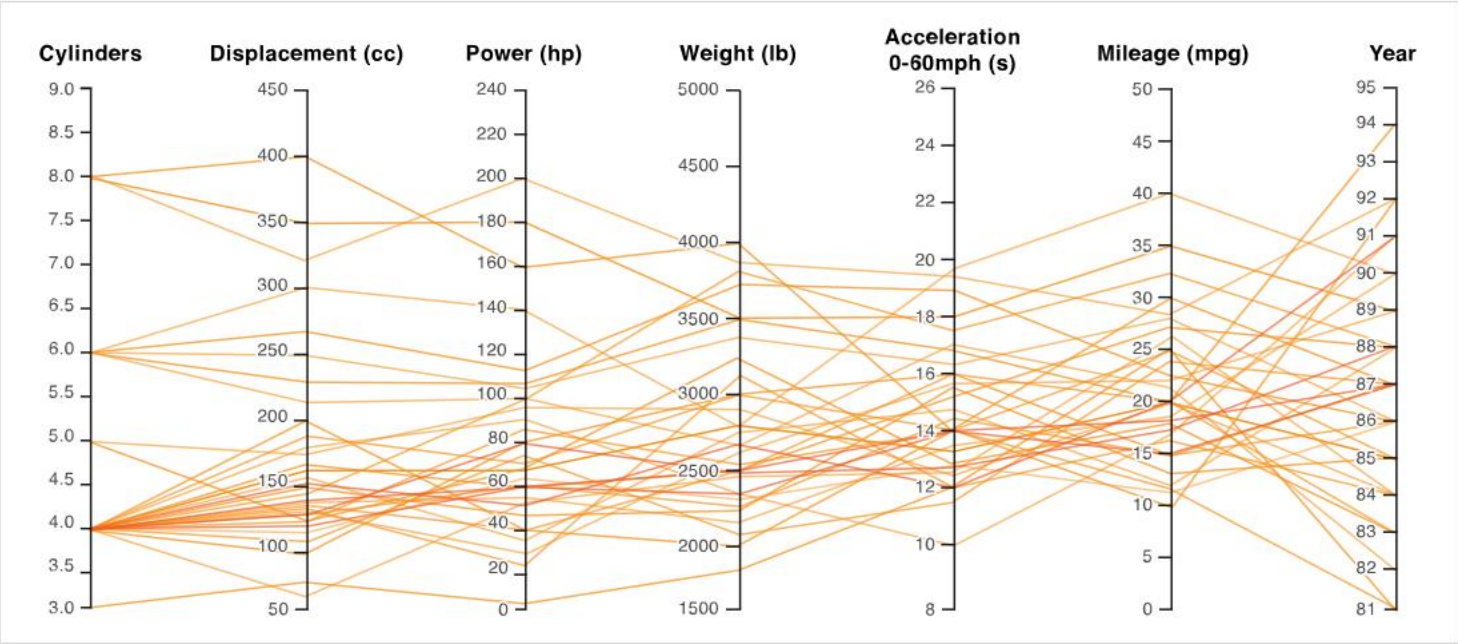
- Map
- Parallel Coordinate Plot
- Venn Diagram
- Timeline
- Euler Diagram
- Hyperbolic Trees
- Cluster Diagram
- Ordinogram
- Isoline
- Isosurface
- Streamline
- Direct Volume Rendering (DVR)

Map

It is generally used to represent the location of different areas of a country and is generally drawn on a plain surface. Google maps is generally widely used for big data visualization. Now–a-days it is widely used for finding the location in different domains of country.

Parallel Coordinate Plot

It is a visualization technique of representing multidimensional data. This type of visualisation is used for plotting multivariate, numerical data. Parallel Coordinates Plots are ideal for comparing many variables together and seeing the relationships between them.



Class Work

Consider the characteristics of each item to respective variables. Draw a parallel coordinates plot for the following multivariate numerical data.

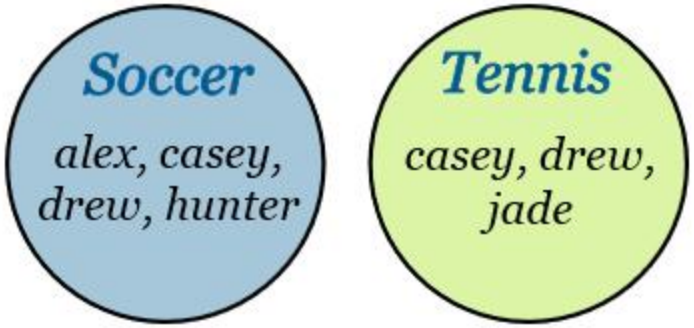
	Variable X	Variable Y	Variable Z
Item 1	50	100	2.0
Item 2	20	115	1.5
Item 3	70	120	3.0
Item 4	90	200	4.0

The highest and lowest values of each variable are as follows:

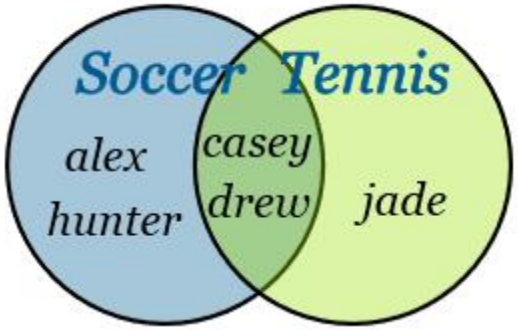
	Highest	Lowest
Variable X	100	25
Variable Y	250	100
Variable Z	5.0	1.0

Venn Diagram

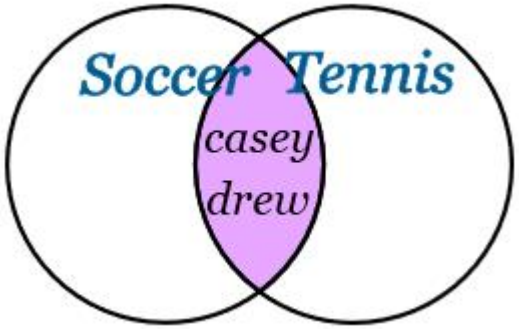
It is used to represent the logical relations between finite collection of sets.



List your friends that play Soccer OR Tennis



List your friends that play Soccer AND Tennis



- Draw the Venn Diagram to show people that play Soccer but NOT Tennis
- Draw the Venn Diagram to show people that play Soccer or play Tennis, but not the both.

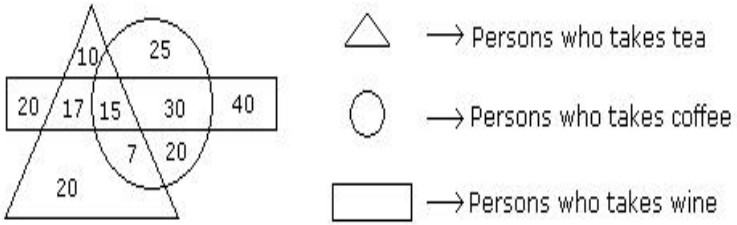
Class Work

Let A, B and C represent people who like apples, bananas, and carrots respectively. The number of people in A = 10 (A1 to A10), B = 12 (B1 to B12) and C = 16 (C1 to C16). Three people are such that they enjoy apples, bananas as well as carrots. Two of them like apples and bananas. Let three people like apples and carrots. Also, four people are such that they like bananas and carrots. Draw the Venn diagram illustrating:

- How many people like apples only?
- How many people like only one of the three?
- How many people like all three?

Class Work

Consider the following diagram.

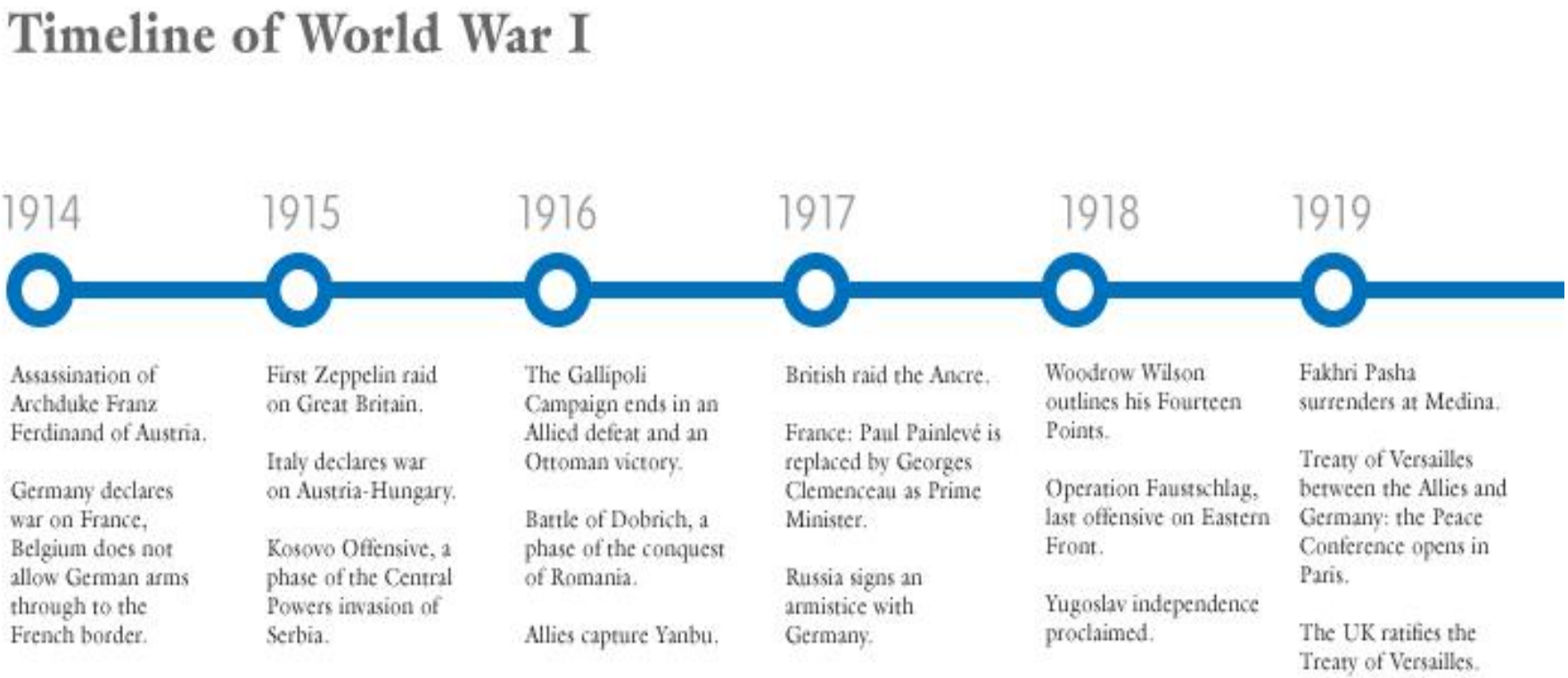


Draw the Venn diagram illustrating:

- How many people like only tea and wine?
- How many people like tea only?
- How many people like only tea and coffee?
- How many people like tea, coffee and wine?

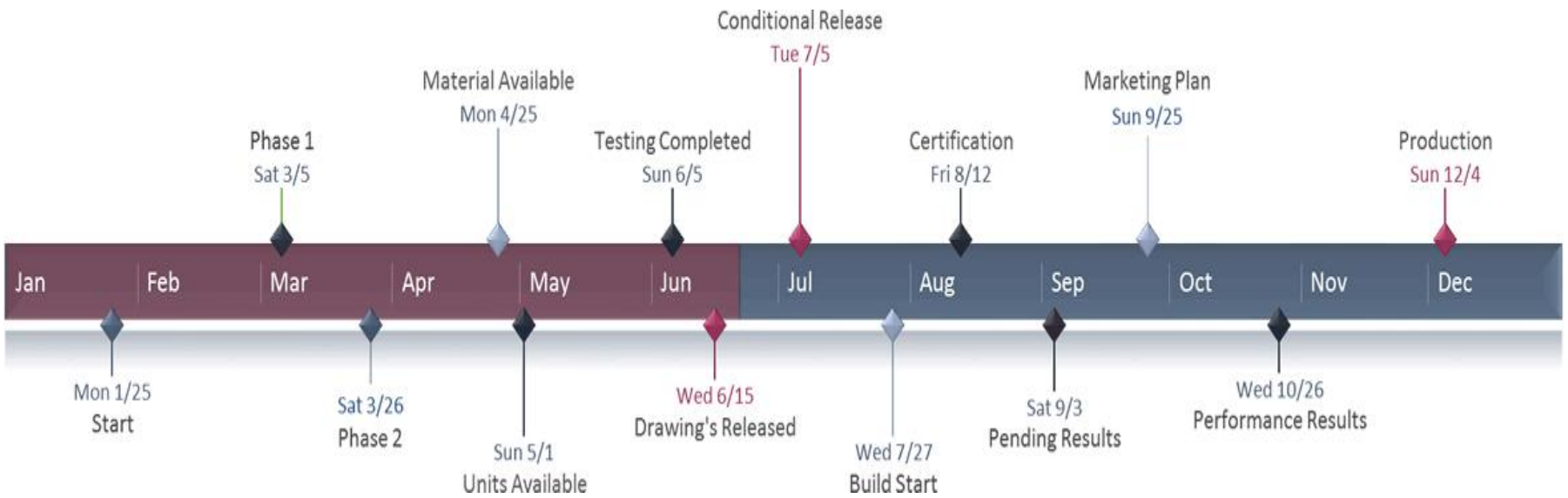
Timeline

It is used to represent a chronological display of events.



Source: datavizcatalogue.com

Timeline cont'd



Source: officetimeline.com

Class Work

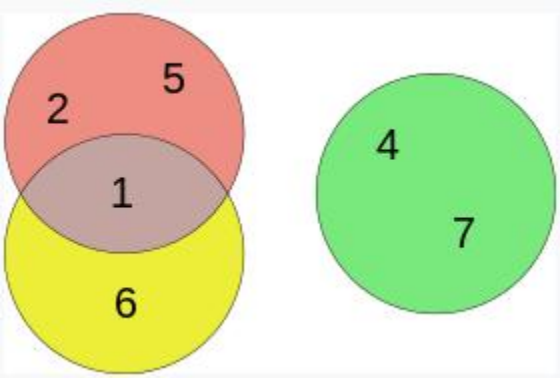
Draw a chronological sequence of the following dataset

Year	Event
1941	First electronic computer
1947	First commercial stored program counter
1972	PROLOG language revealed
1950	Birth of AI
1956	Dartmouth conference
1957	Logic Theorist development
1968	Micro-world program SHRDLU created
1958	Lisp language developed
1963	Start of DoD's advance research project
1970	First Expert System
1991	AI system beats human chess master

Euler Diagram

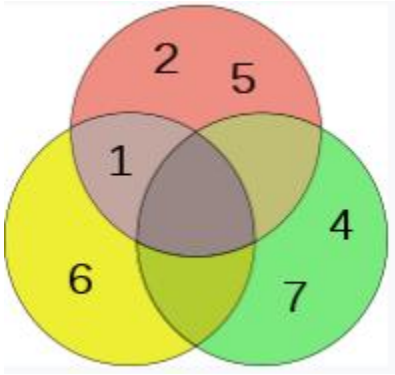
It is a representation of the relationships between sets.

Example: Let's take 3 sets namely A = {1, 2, 5}, B = {1, 6} and C={4, 7}. The Euler diagram of the sets looks lie:



Source: wikipedia

Draw the Equivalent Venn Diagram



Class Exercise

Draw the Euler diagram of the sets, X = {1, 2, 5, 8}, Y = {1, 6, 9} and Z={4, 7, 8 , 9}. Then draw the equivalent Venn Diagram.

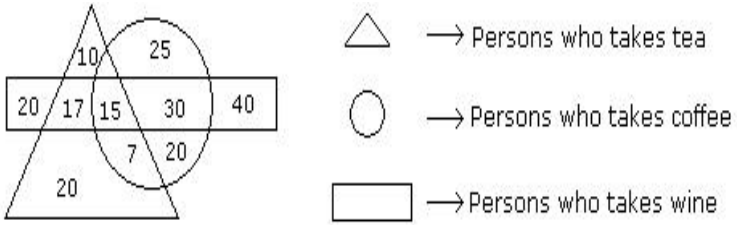
Class Work

Let A, B and C represent people who like apples, bananas, and carrots respectively. The number of people in A = 10 (A1 to A10), B = 12 (B1 to B12) and C = 16 (C1 to C16). Three people are such that they enjoy apples, bananas as well as carrots. Two of them like apples and bananas. Let three people like apples and carrots. Also, four people are such that they like bananas and carrots. Draw the Euler diagram illustrating:

- How many people like apples only?
- How many people like only one of the three?
- How many people like all three?

Class Work

Consider the following diagram.

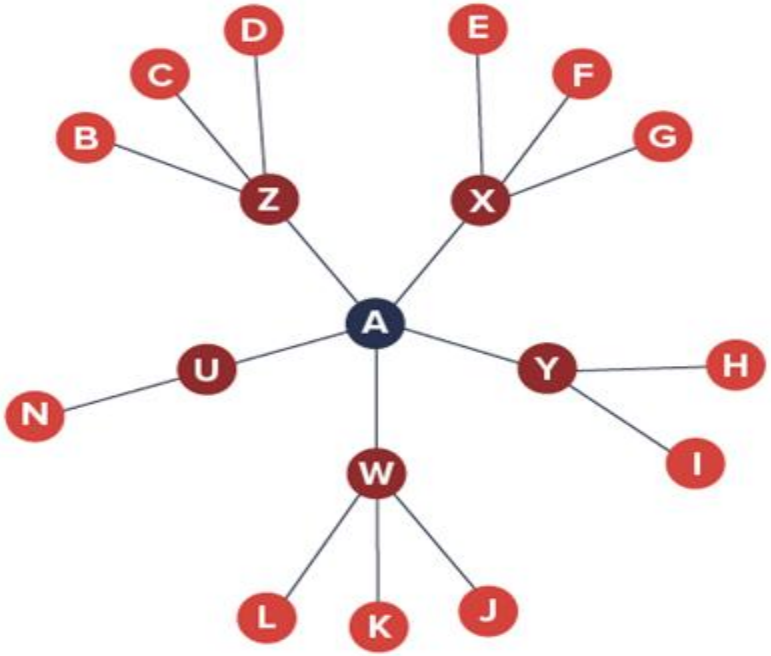


Draw the Euler diagram illustrating:

- How many people like only tea and wine?
- How many people like tea only?
- How many people like only tea and coffee?
- How many people like tea, coffee and wine?

Hyperbolic Trees

A hyperbolic tree (often shortened as hypertree) is an information visualization and graph drawing method inspired by hyperbolic geometry. These are special types of graphs composed of vertices and edges (connecting lines).

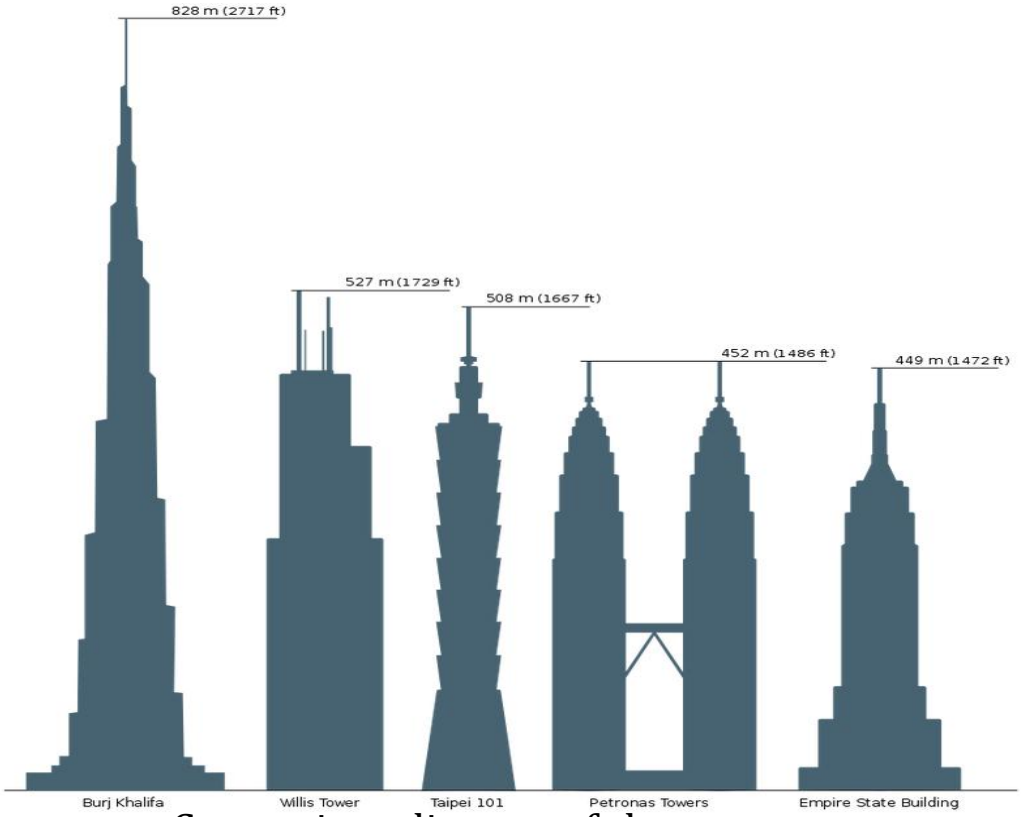


Class Work

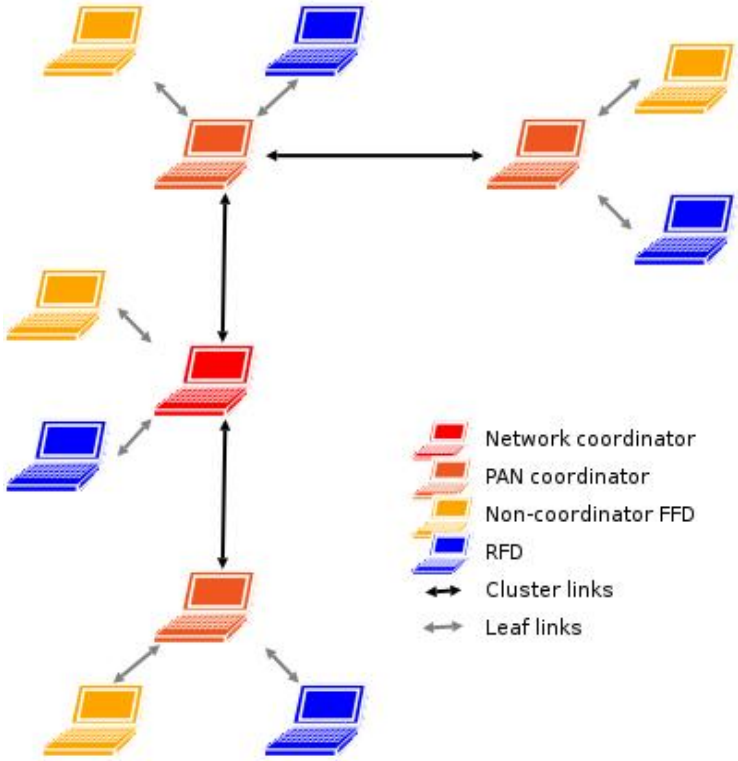
Your parent has asked you to reduce monthly expenses. So you adopted 2 techniques called as “Buy Less” and “Pay Less”. So, you have started in buying less food and cloths. While buying less food, you have started cooking; reduce eating outside meals, buy nonperishable items in bulk. While paying, you have started sharing the cost, bought items that are on sale, and bought generic items. Draw the hyperbolic tree.

Cluster Diagram

A cluster diagram or clustering diagram is a general type of diagram, which represents some kind of cluster. A cluster in general is a group or bunch of several discrete items that are close to each other.



Comparison diagram of sky scraper



Computer network diagram

Class Work

- Draw a cluster diagram for big data environment by including name node, secondary name node, standby name node and data nodes arranged in 10 racks. Each rack has different number of data nodes.
- Draw a cluster diagram for city-planning by identifying groups of houses according to their house type, value, and geographical location.
- Draw a cluster diagram for an insurance firm by identifying groups of motor insurance policyholders with a high average claim cost.

Ordinogram

It is generally used to perform the analysis operation of various sets of **multivariate objects** which are generally used in different domain. *Simple two-dimensional graph is an example of ordinogram.*

Univariate data – This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.

Height (cm)	164	167	170	170.4	176.5	180	179.2	165	175

Multivariate data – This type of data involves two or more than two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the variables.

Temperature in Celsius	Ice Cream Sales
20	2000
35	5000

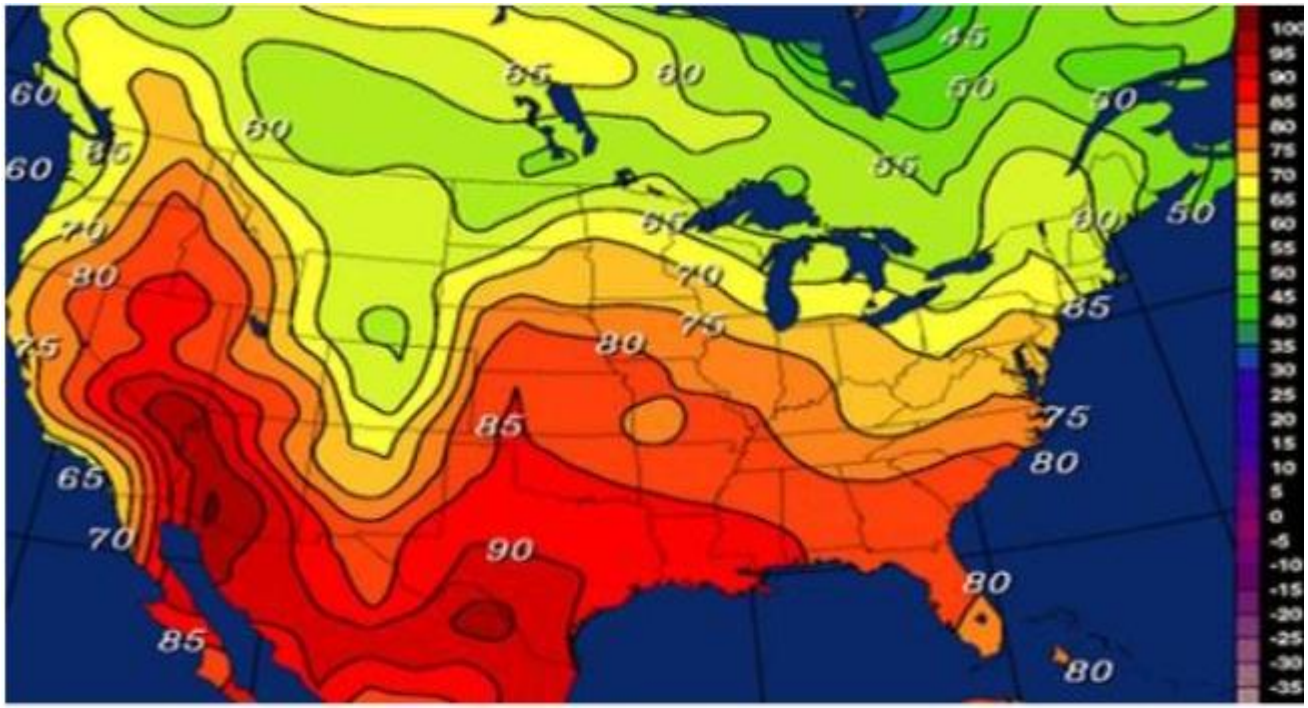
Class Work

For the following dataset, draw the ordinogram.

Ice Cream Sales vs. Temperature												
Temp	14.2	16.4	11.9	15.2	18.5	22.1	19.4	25.1	23.4	18.1	22.6	17.2
Sales	215	325	185	332	406	522	412	614	544	421	445	408

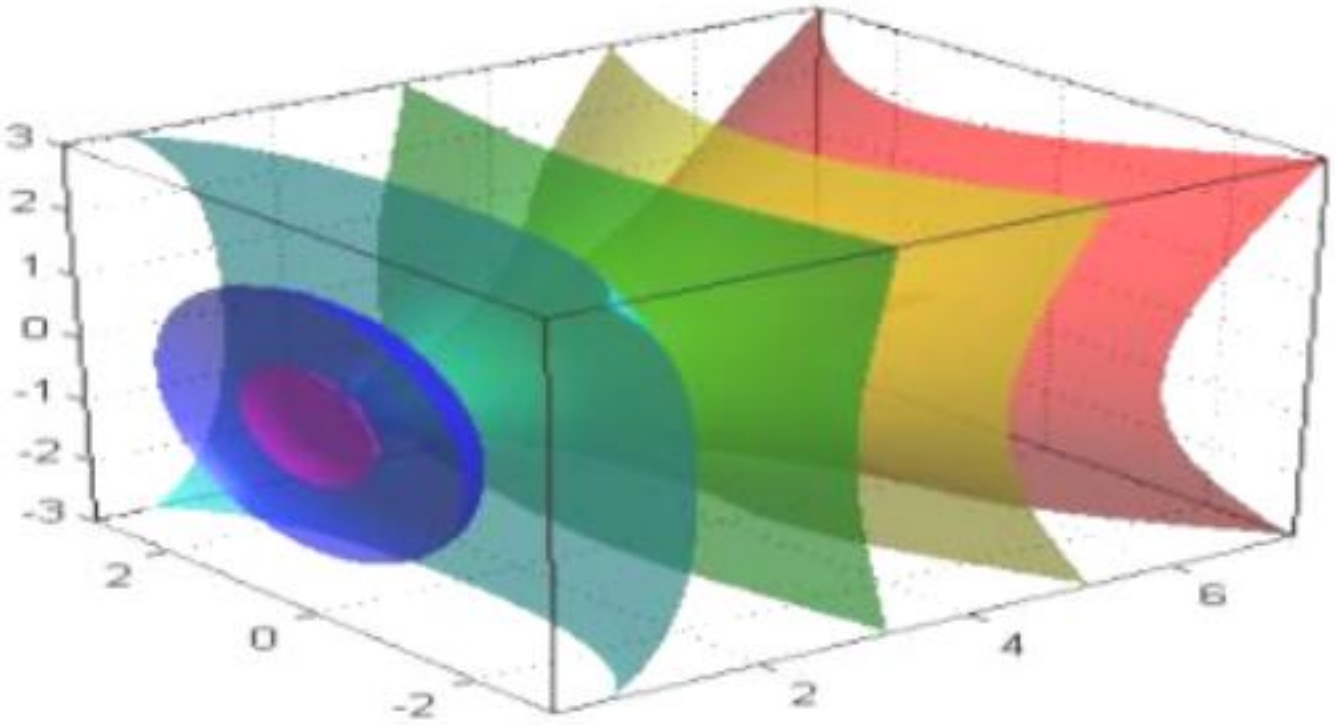
Isoline

It show a range of quantity i.e., it is basically a 2D data representation of a curved line that generally transfers constantly on the surface of the graph, and the plotting of line generally drawn on the basis of data arrangement. Radar maps, temperature maps and rainfall maps are all Isoline.



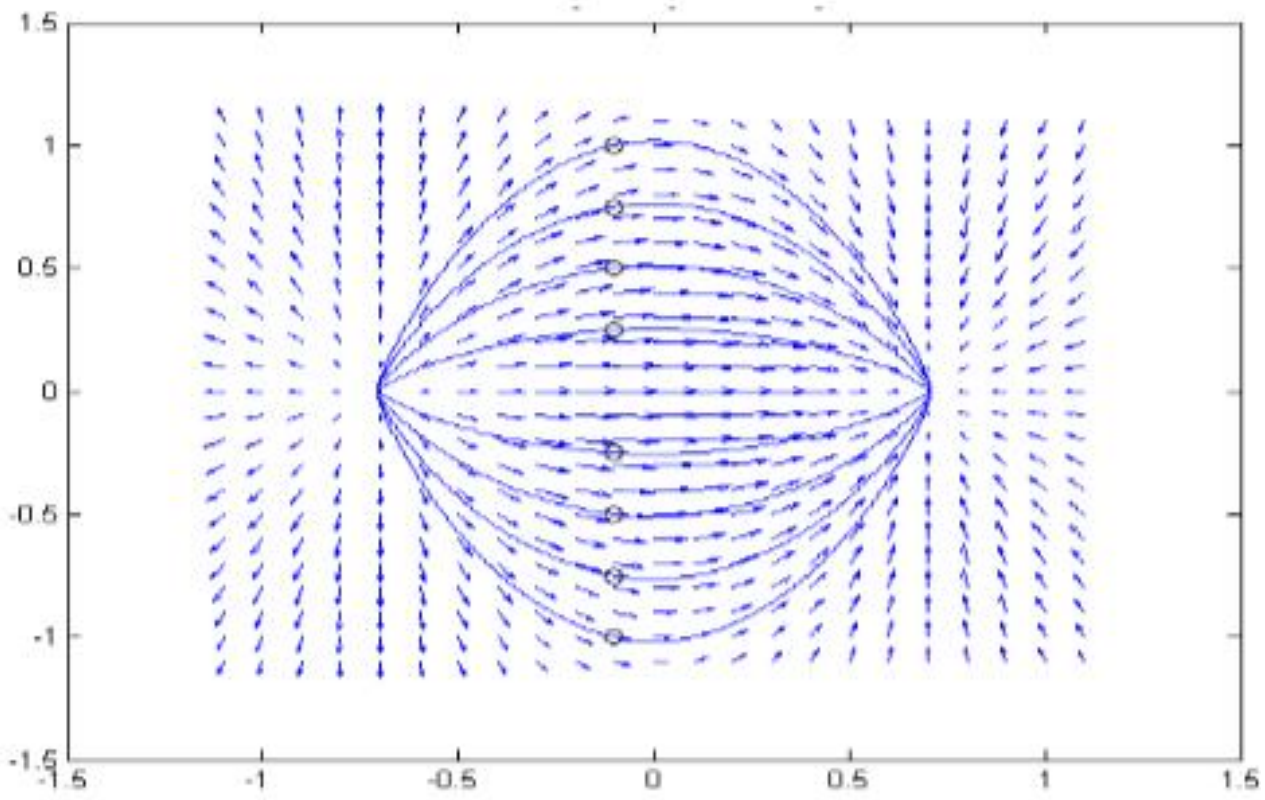
Isosurface

It is a 3D representation of an Isoline. Isosurfaces are designed to present points that are bound by a constant value in a volume of space i.e. in a domain that covers 3D space.



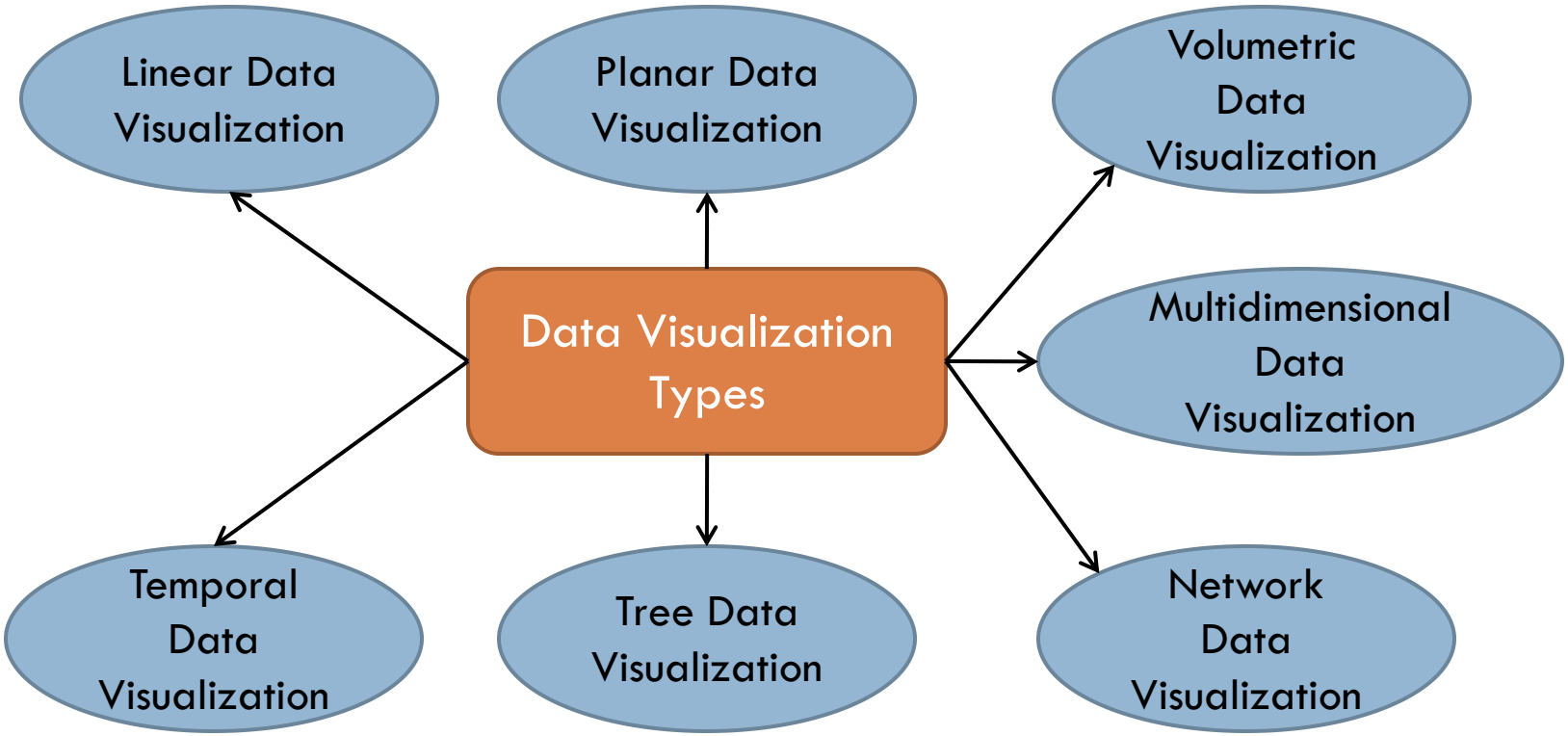
Streamline

It is a field that is generated from the description of velocity vector field of the data flow.



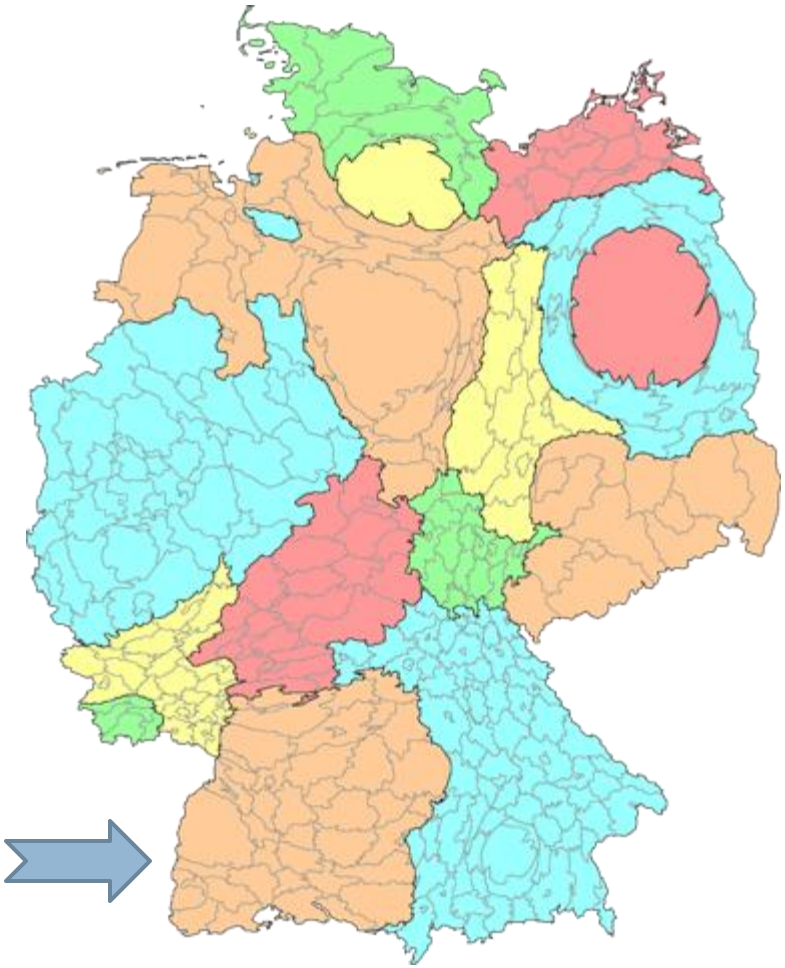
Types of Data Visualization

Data visualization can be done in different ways such as:



Types of Data Visualization cont'd

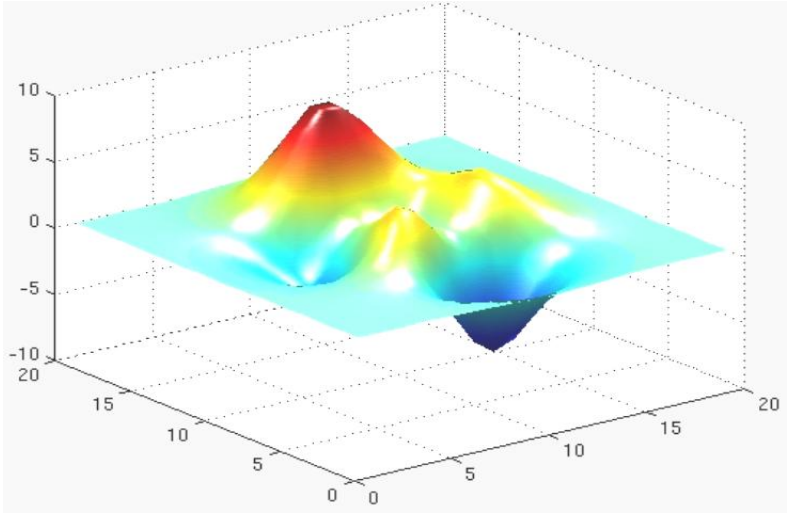
- **Linear Data Visualization:** Data always represented in list format. Basically it's not consider it as a visualization technique rather is a data organization technique. No tool is used to visualize the data. It is also called as 1D data visualization.
- **Planar Data Visualization:** Data generally take in the form of images, diagrams or charts over a plane surface. The best example of this type of data visualization is Cartogram and dot distribution. A cartogram is a map in which some thematic mapping variable – such as travel time, or population is substituted for land area or distance. Some tools used to build planar data visualization are GeoCommons, Polymaps, Google Maps, Tabeleau Public etc.



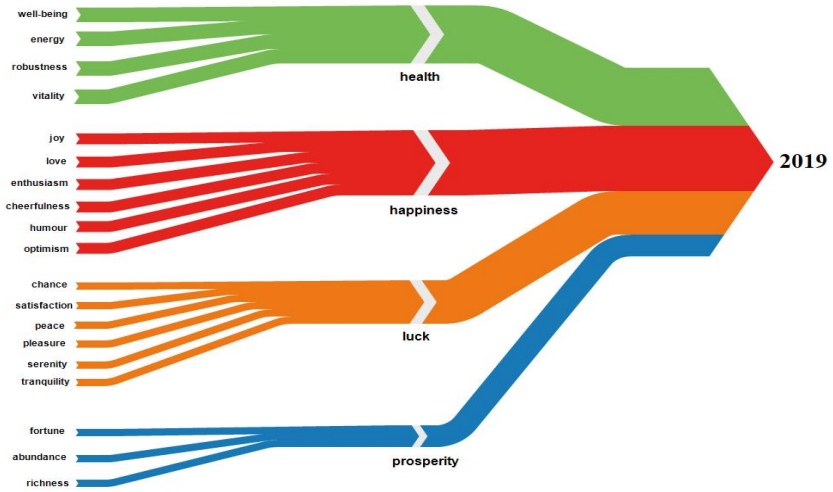
Germany-population-cartogram

Types of Data Visualization cont'd

- **Volumetric Data Visualization:** the presentation of data generally involves exactly with three dimensions to present simulations, surface and volume rendering etc. and commonly used scientific studies. Basic tools used for it are AC3D, AutoQ3D, TrueSpace etc.

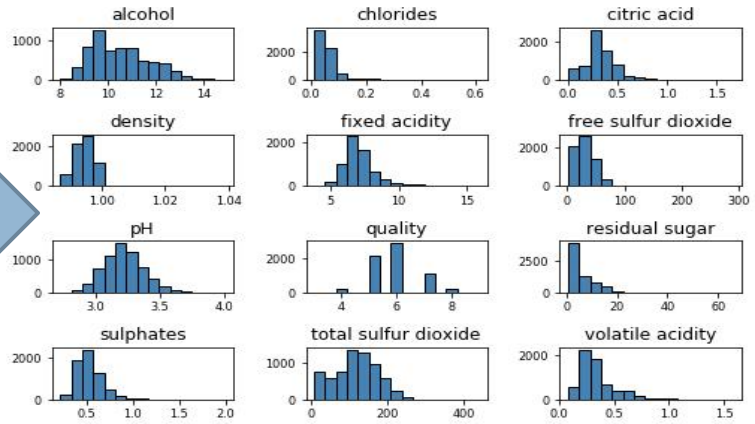


- **Temporal Data Visualization:** Sometimes, visualizations are time dependent so to visualize the dependence of analyses of time, the temporal data visualization is used which include Gantt chart, Time series and Sanky diagram etc.

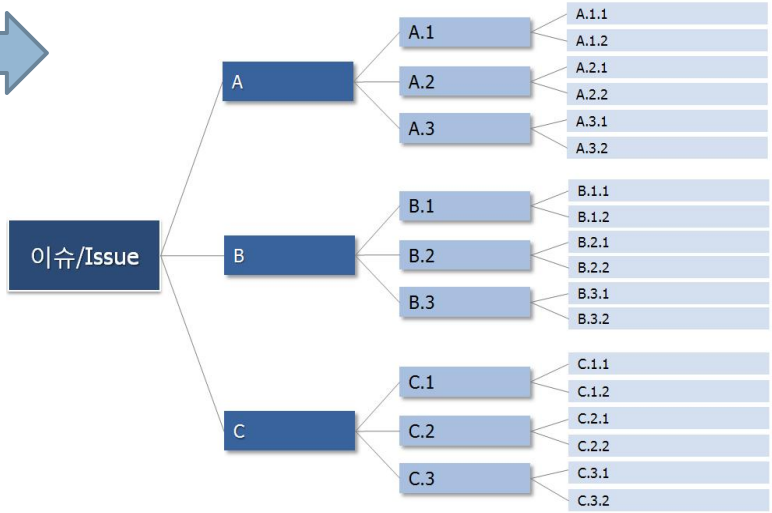


Types of Data Visualization cont'd

- **Multidimensional Data visualization:** Numerous dimension are generally used to represent the data. Generally pie charts, histograms, bar charts etc are generally used. Many Eyes, Google Charts, Tableau Public, etc. are some tools used to create such visualization.

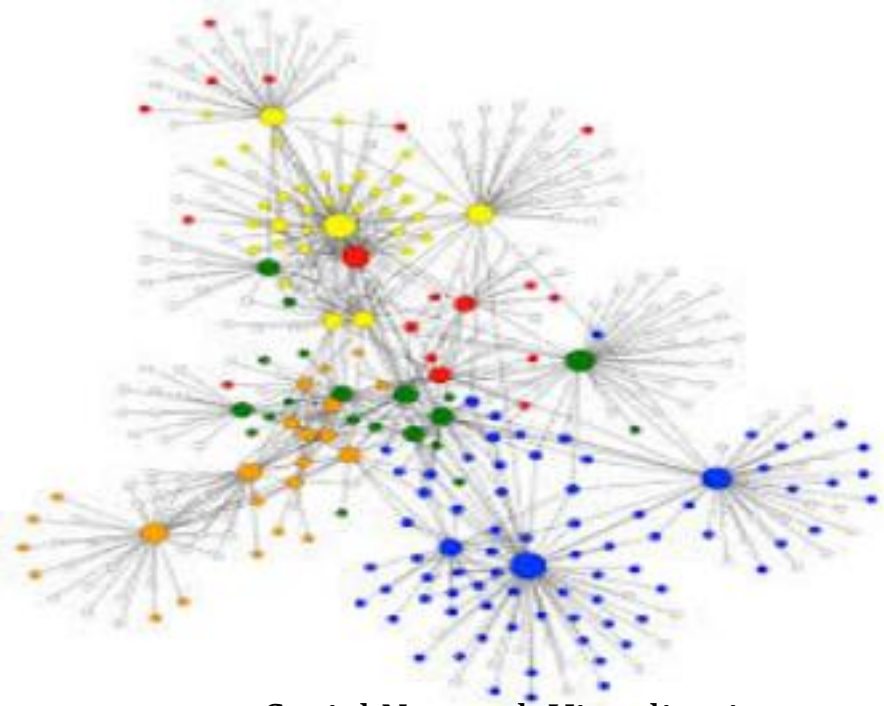


- **Tree/Hierarchical Data visualization:** Sometimes, data relationships need to be shown in the form of hierarchies and to represent it, tree or hierarchical data visualization. Examples include hyperbolic tree, wedge-stack graph, etc. Google Charts, d3, etc. are some tools used to create such visualization.



Types of Data Visualization cont'd

- **Network Data Visualization:** This approach is generally used to represent the relations that are too complex in the form of hierarchies. Some of the basic tools used for network data visualization are hive plot, Pajek, Gephi, NodeXL, Google Fusion Tables, Many Eyes, d3/Protovis etc.



Social Network Visualization

Visualization Interaction Techniques

The following interaction techniques are used in information visualization to overcome various limitations such as maximum amount of information is limited by the resolution.

- **Zooming:** It is one of the basic interaction techniques of information visualizations. It allows the user to specify the scale of magnification and increasing or decreasing the magnification of an image by that scale. This allows the user focus on a specific area and information outside of this area is generally discarded.
- **Filtering**: It is one of the basic interaction techniques often used in information visualization used to limit the amount of displayed information through filter criteria.
- **Details on demand**: This technique allows interactively selecting parts of data to be visualized more detailed and additional information on a point-by-point basis.
- **Overview-plus-Detail:** Two graphical presentation, wherein one shows a rough overview of the complete information space and neglects details, and the other one shows a small portion of the information space and visualizes details. Both are either shown sequentially or in parallel.

Application of Data Visualization

There are 3 ways to use data visualization in a company.

1. **Internal Communication:** Any key data that influences decision-making is prime for data visualization. This is specifically true for the information delivered to higher-ups such as boss or other key stakeholders. Examples are presentation, reports or financial statements.
2. **Reporting**: With data visualization, results reporting to clients or customers is more impactful.
3. **Marketing Content**: Public-facing content for thought leadership or promotion is more credible with data. Content such as blogs, whitepapers, infographics etc. can be beneficial.

THANK YOU!