Unit 1

Introduction to Visualization

Data analysis is the process of cleaning, processing and changing raw data and extracting relevant information from it. Its main purpose is to look for trends, grouping and other relationship between different types of data. **Data visualization** is the process of translating information into visual context such as graph, charts, map etc. to make data easier for human to understand and gain insight from it. The aim of data visualization is to make it easier to identify patterns, trends and outlier in data sets. Visualization represents analyzed data in form of pictorial form that helps to communicate results in simpler form. Data visualization eccentrics includes formats like bar charts, line graph, scatter plot, heat maps etc. Data visualization is a way to represent complex information in easier and more understandable format. Effective data visualization are designed to be clear, accurate and visually appealing such that human brain can perceive and understand meaning in easier way.

Good visualization helps to :

- Provide rapid access to data
- Represent the data in proper format and tell the scenario
- Express complex ideas in simple form
- Shows relationships between abstract concepts

Visualization stages:

it generally includes four basic stages:

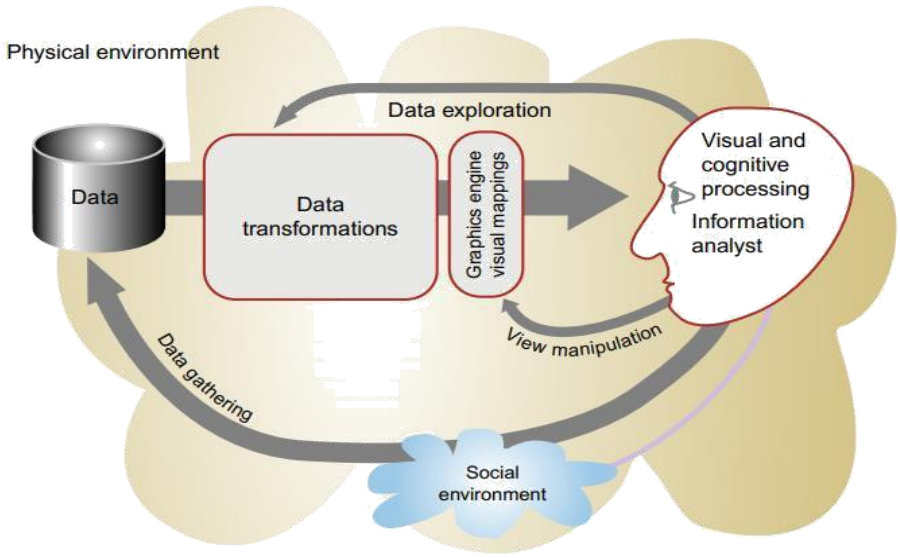
Step 1: collection and storage of data

Step 2: preprocessing stage which transform the data into something that is easier to manipulate.

Step 3: mapping from the selected data to a visual representation which is done through the algorithms that produce an image on the screen.

Step 4: the human perceptual and cognitive system.

Visual perception refers to the way in which human perceive and interpret visual information represented by visual eccentrics like charts, graph etc. it refers to how human's visual system process and interpret the surrounding such as color, shapes, sizes to extract meaningful information and pattern from data. It is the ability to interpret the surrounding environment by processing information that is contained in visible light. It explains how people express surroundings through light that enters our eyes. For effective data visualization, visual perception should take into consideration such that accurate and understandable visualization can be created



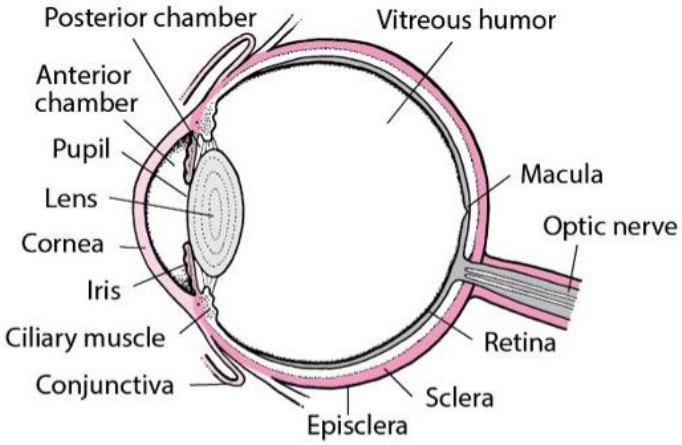
In above figure, longest loop involves data gathering. A data seeker may choose to gather more data to follow up on an interesting lead. Another loop controls the computational preprocessing that takes place prior to visualization which helps to give up the meaning form subjected data. Both the physical environment and social environment are involved in the data gathering lop. The physical environment is a source of data while the social environment determines the way the data is collected and how it is interpreted. The computer is treated as a universal tool for producing interactive graphics i.e. once we figure out the best way to visualize data then algorithm is constructed to create appropriate images

What is visual perception?

- The process of knowing or being aware of information through the eyes.
- The process of acquiring, interpreting, selecting, and organizing sensory information.
- The term refers not only to seeing with the eyes but also to the complex organizational process in which the brain determines how to understand visual data

Visual System

- Light path
  - Cornea, pupil, lens, retina
  - Optic nerve, brain
- Retinal cells
  - Rods and cones
  - Unevenly distributed
- Cones
  - Three "color receptors"
  - Concentrated in fovea
- Rods
  - Low-light receptor
  - Peripheral vision



Img 1: Anatomical structure of an eye, Source

Pictures

- Generally produced by purpose built cameras, or computer applications, or drawings.
- For instance a camera has good optics, focus, white balance exposure controls. It captures a large image at constant high quality resolution (spatial, luminance, hue).

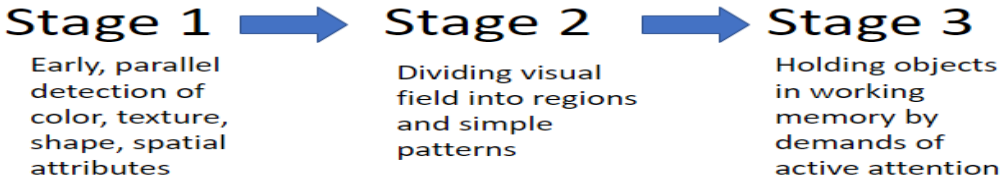
Eyes

- Heuristics developed by evolution using inexpensive biophysical hardware to keep operator alive.
- Human visual system is different in that it has relatively poor optics. It is constantly scanning, adjusting focus, white balance, exposure. It captures detail only in foveal area and very coarsely in the peripheral vision areas. We produce a 3D spatial reconstruction of our world based on our 2D snapshots.

One Simple Model of Perceptual Processing

Three stage process

- Selection
- Organization
- Interpretation



Why it is important?

Visual perception is extremely important because it allows people to interact with the world around them. While blind people and people with limited visual perception certainly participate in the world, they undeniably face accessibility challenges when doing so, because human society is so strongly biased toward visual experiences.

How does human visual system analyze images?

- Some things seem to be done preattentively, without the need for focused attention
- Generally take less than 200-250 msecs (eye movements take 200 msecs)
- Seems to be done in parallel by low-level vision system
- An important contribution vision science makes to data visualization is that a limited set of visual properties can be detected very rapidly and accurately by the low-level visual system

A visual representation of data refers to use of graphical or visualization tool to convey patterns, information, relationship in dataset in easier and simpler form. The main goal of visual representation is to enforce effective communication and analysis of information. The choice of visual representation depends on the type of data, meaning we want to convey and the audience familiarity with the data domain. Some common types of visual representation are:

1. **Charts and graphs:** includes bar chart, line charts, scatter plot, pie chart that are used to display quantitative data to reveal trends and comparison.

2. **Maps**: use to represent geographical data using heatmaps, maps and other spatial visualization. It helps to reveal geographic pattern and their relationship

3. **Tree map:** use nested rectangle or circular segment to represent hierarchical data and its distribution within categories.

4. **Heatmap**: use to represent two-dimensional data by using color to show density of data points and highlight patterns and concentration.

5. **Word clouds:** use to display words from a text dataset with the varying size of text to indicate its frequency. It helps to visualize word frequencies and marks the important topic in text analysis.

6. **Time series visualization:** shows how data change over time and includes charts to visualize chart.

- A powerful tool to help understand complex datasets.

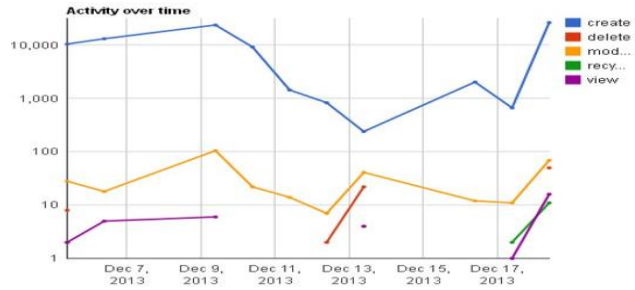
- It enables us to quickly interpret large amounts of data and draw meaningful conclusions.

- Visual representations of data can take various forms, from simple bar charts to complex interactive visualizations.

TYPES OF VISUAL REPRESENTATION OF DATA

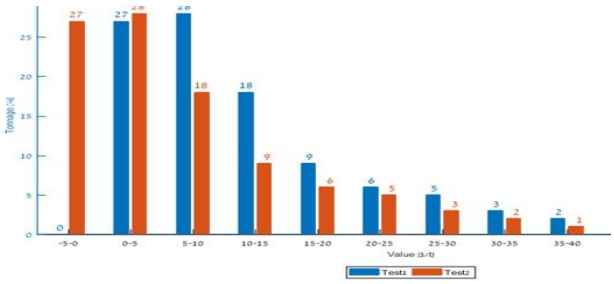
1. Line Charts

These are used to show trends over time. They are ideal for demonstrating how a variable changes over a period.



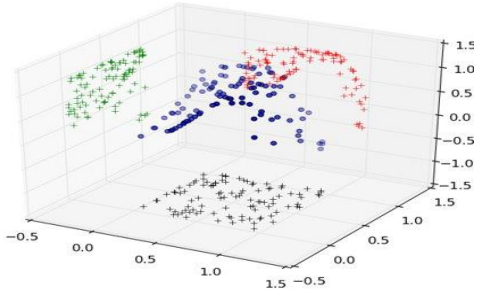
2. Bar Charts

These are used to compare data across different categories. They are simple and effective, making them a popular choice for presenting data.



3. Scatter Plots

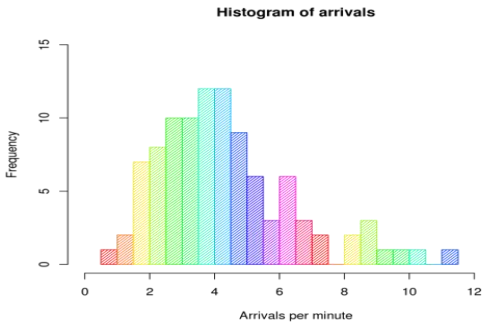
These are used to show the relationship between two variables. They are useful in identifying trends and patterns and help identify outliers.



4. Pie Charts

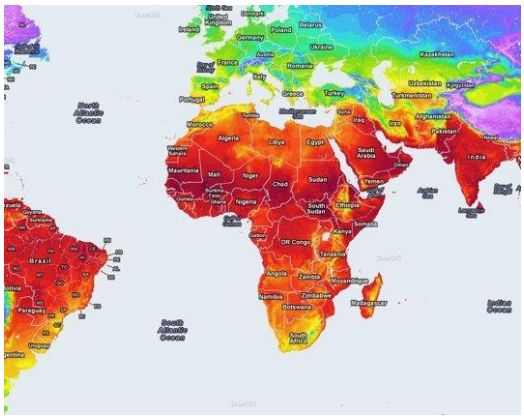
These are used to show the proportion of different categories in a dataset. They are a popular choice for presenting data in a clear and concise manner.

5.Histograms



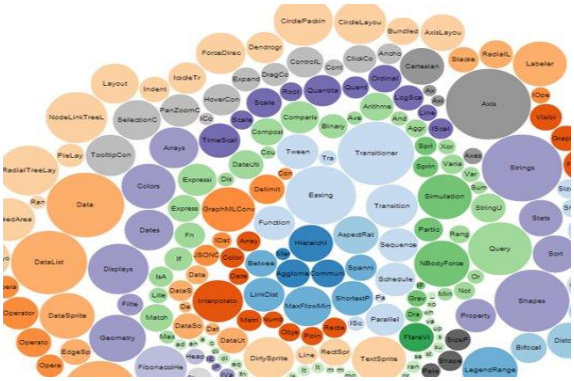
These are used to show the distribution of a dataset. They are helpful in identifying patterns, clusters, and outliers.

6. Heat Maps



These are used to show the density of data in a particular area. They are useful in identifying areas of high and low concentration of data.

7. Bubble Charts



These are like scatter plots but add a third variable, usually represented by the size of the bubble.

Visual representations of data offer several advantages :

- They allow for quick and easy interpretation of large datasets and can reveal patterns and trends that may otherwise be difficult to spot.
- Visualizations can also be used to communicate complex ideas in a more accessible way.
- Visualizations can also be used to highlight key points and draw attention to important aspects of data. This can be especially helpful when presenting data to a wider audience.

Visual representations of data can also have some disadvantages.

- Poorly designed visualizations can be misleading or difficult to interpret.
- Limitations; as they cannot always capture all the nuances of a dataset.
- Visualizations can be used to manipulate data or draw false conclusions, so it is important to be aware of potential biases and use visualizations responsibly.

Visual representations of data are a powerful tool for understanding and communicating complex datasets. They offer several advantages over traditional methods of data analysis and can be used to effectively communicate complex ideas. However, it is important to be aware of the limitations of visualizations and use them responsibly.

Data abstraction is the process of summarizing and simplifying complex data to highlight required patterns, trend and insight. It is a crucial process to make data more manageable and understandable for data analysis and visualization pattern. It is used to reduce level of details from data set. Effective data abstraction helps to enhance the efficiency of data by focusing on most relevant aspect of data. Data abstraction is required in data analysis and visualization for:

i. **Reducing complexity:**

   Complex data sets contain null values, noise and irrelevant details. So data abstraction helps to filter out noise, null values and focus on the important aspect of data. Simplification in data makes easier to identify pattern and relationship between variables.

ii. **Aggregation:**

   Involves combining individual data points into group to calculate summary statistics of each group. For example: daily sales figure, population trends, disease trend etc. aggregation helps in understanding overall trends and patterns in each data points.

iii. **Summarization:**

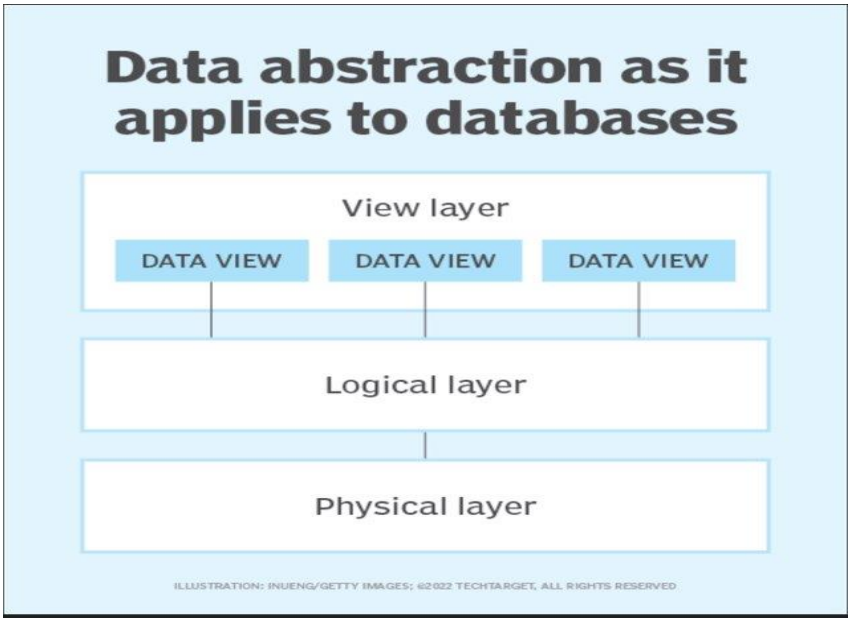
   Involves creating lot of information to capture essence of the data sets. It includes calculation measures like mean, median, mode, range and percentile that can helps to provide quick overview of the data distribution.

iv. **Data transformation:**

   Includes operation like scaling, log transformation and normalization which can improve the distribution and relationships within the data to make it more suitable for data analysis and visualization.

v. **Sampling:**

Sampling is used to select a representative subset of data for analysis. This helps to reduce computational and time requirements.



View Layer

At this level, the user sees the data as it's presented by an application interface. The user might be able to interact with the data or might only be able to view it. In either case, the user typically has access to only certain types of data, has a very limited view of the data in its entirety and has no concept of how or where the data is stored.

Logical Layer

- Provides a conceptual understanding of the data
- Describes the type of data and how that data is related.
- Think as an ERD and how elements are related with each other.

Physical Layer

- Data's physical storage.
- It is concerned with where and how the data is stored.
- How the files are managed?
- And everything that is related to physically storing and maintaining data.

Abstraction makes things simpler, easier to change, easier to implement, easier to document. It can avoid writing tricky code multiple times, simplify design, optimization and indexing, and ultimately increase the maintainability of the solution.

Visual encoding

Visual encoding Visual encoding efers to process of mapping data attributes to visual eccentrics in order to create meaningful information and graphical representation. It involves choosing appropriate visual actor like position, color, size shape and texture to represent different aspect of data. It helps to effectively communicate data insights and patterns to the user. Effective visual encoding ensures that user can quickly and accurately understand the patterns, trend and relationship present in the data. Some common visual properties used in visual encodings are:

i. **Position;** the position of axis in different visual eccentrics like scatter plot, bar graph etc. can represent two different variables.

ii. **Size:** size of graphical elements like circle, bar can be used to represent quantitative values. Larger size typically indicates larger values.

iii. **Color:** color can represent wide range of information and distinguish the information like red color indicates danger, green color indicates safe. Color should be used carefully to ensure accessibility and avoid misinterpretation.

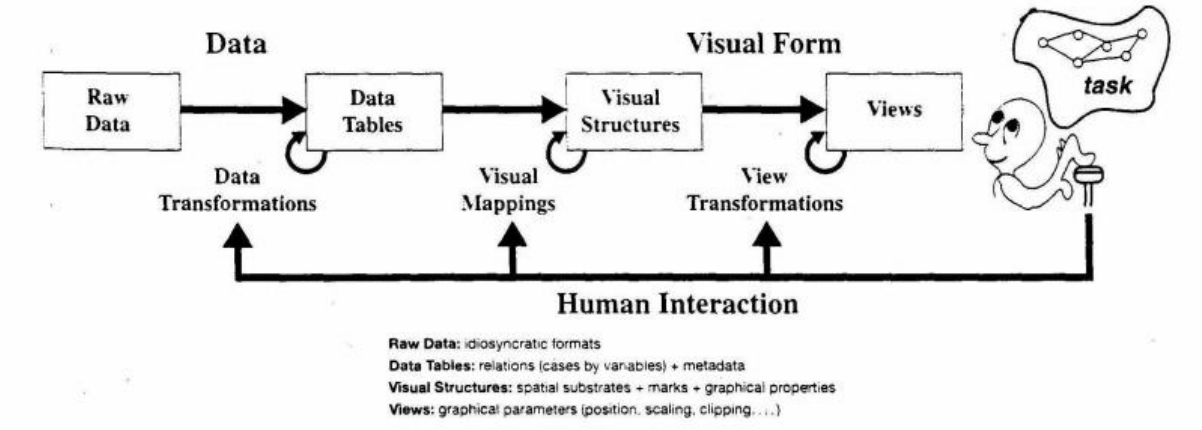
iv. **Shape:** different shape like circle, triangle, square etc. and be used to distinguish categories or represent data points.

v. **Texture:** used to differentiate elements in a visualization when color or shape is limited

vi. **Opacity:** the level of opacity can be used to show density or emphasize specific data points.

vii. **Connection:** lines or links connecting elements that can represent relationships or connection between data points. Mostly used in network visualization.

Mapping Data to Visual Form



Visual encoding is the mapping of information to display elements.

Choosing Visual Encodings

Assume k visual encodings and n data attributes. We would like to pick the “best” encoding among a combinatorial set of possibilities of size (n+1)^k

Principle of Consistency

The properties of the image (visual variables) should match the properties of the data.

Principle of Importance Ordering

Encode the most important information in the most effective way.

Design Criteria

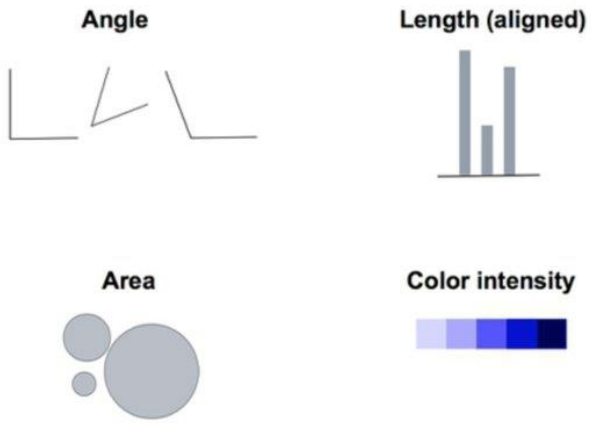
Expressiveness

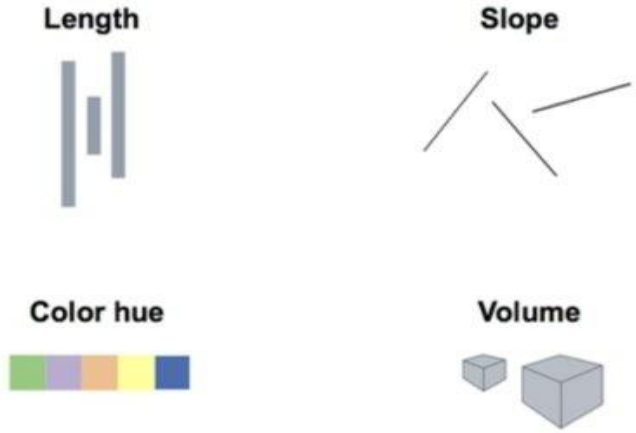
- Express all the facts in the set of data

Effectiveness

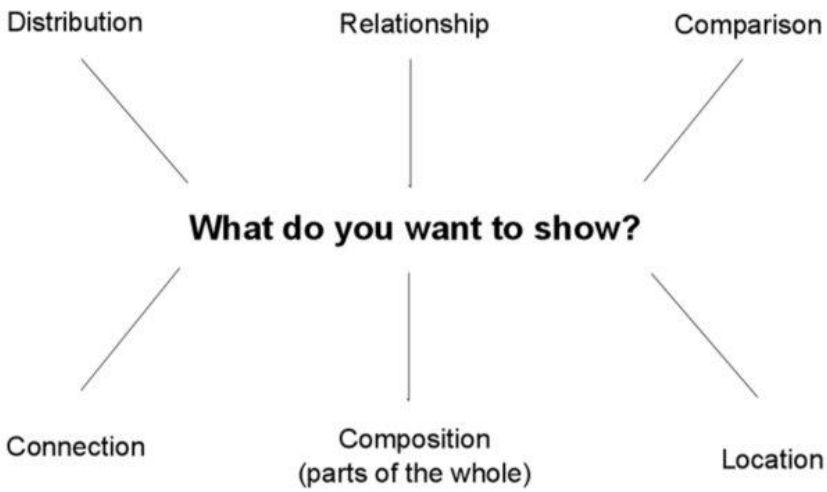
- One Vs Other
- One visualization is more readily perceived than the information in the other visualization.

Encoding data using visual cues

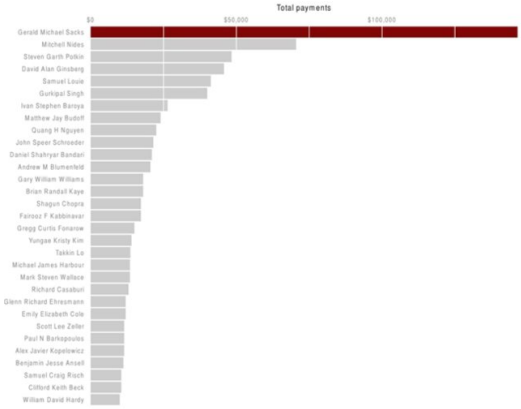




Why Visualize ?

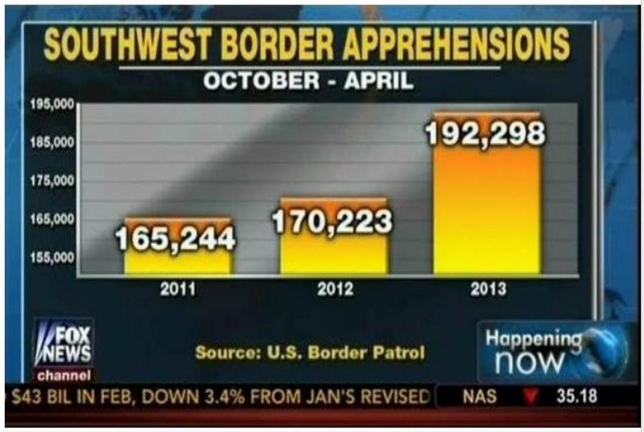


Doctor Vs Payment

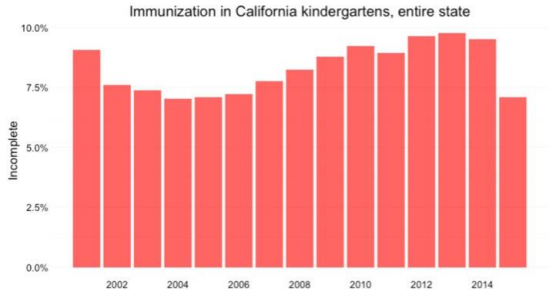


Secondary Color

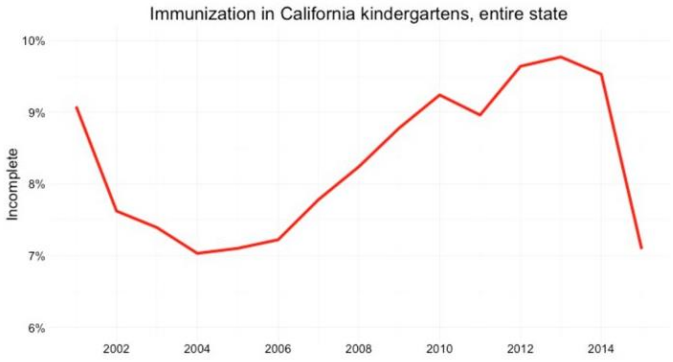
Bar Vs alignment



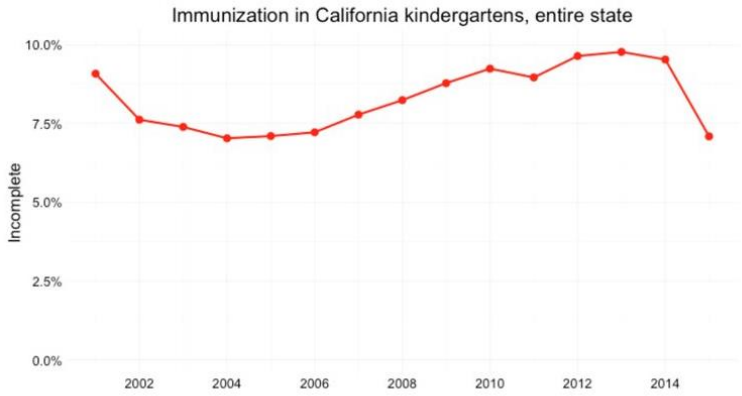
Same Data, Different Visualization



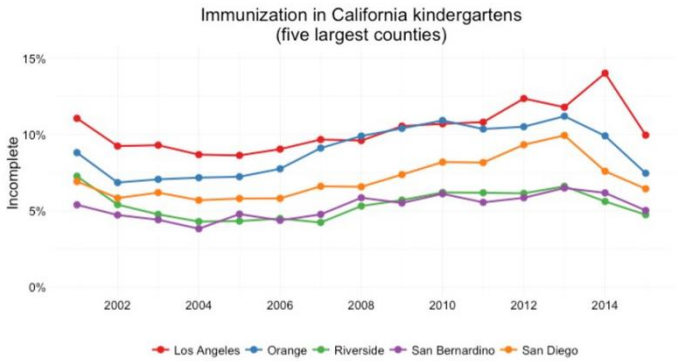
Starting Point=?



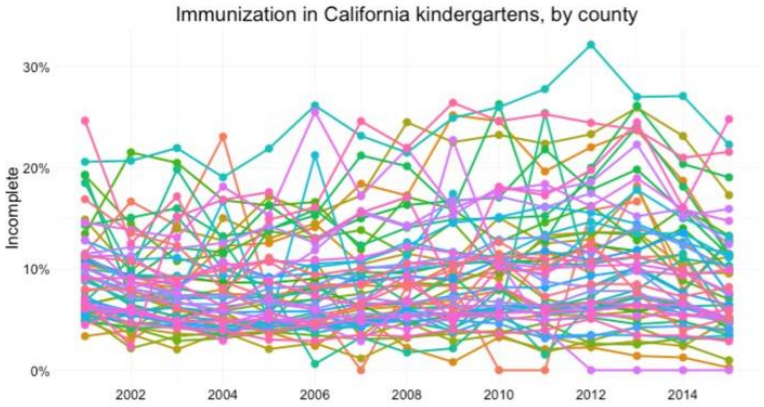
Starting Point =?

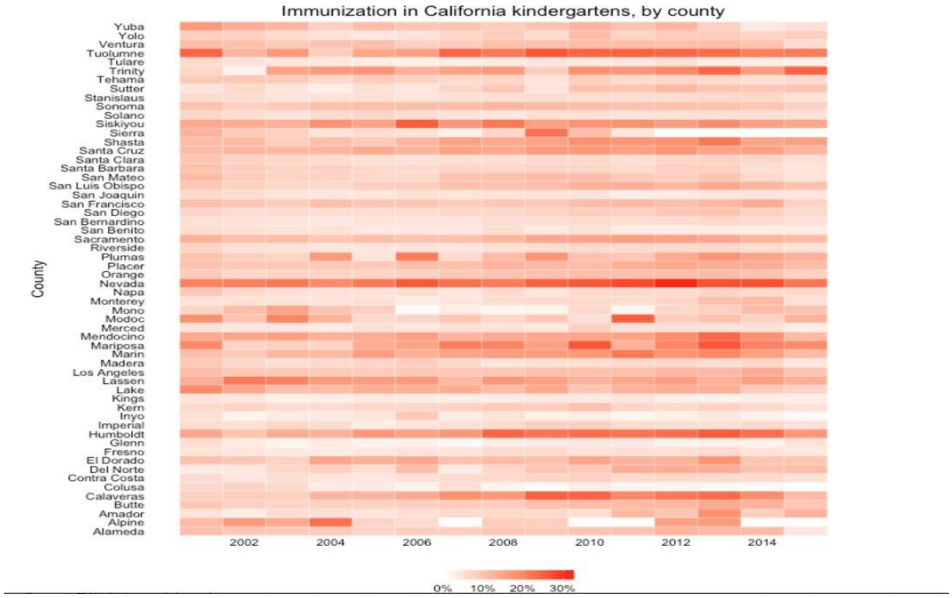


For Handful of countries



For Numerous States/ Countries





Color in visualization: Use of color

Color is a tool to convey information, highlight patterns and engage viewers. In helps to give meaning on datasets. It should be used carefully as it can convey different meaning of data. Color plays the important role on ensuring correct visualization. Color is valuable tools which can enhance the clarity and impact of your data. Color are used to visualize following data:

i. **Categorical data**. Color is used to differentiate groups in dataset. For example, color can be used to represent different product categories, population categories in bar graph, comparison of two or more data in scatter plot, distinguish countries on a map.

ii. **Sequential data:** color can be used to represent ordered data such as values that increase or decrease progressively. For example, different color used in heatmaps to show similarities and dissimilarities between variables, color in line chart to distinguish two or more variable.

iii. **Diverging data**: color can be used to point the data that are above or below a central value. It helps to show distinct midpoint such as comparing positive and negative changes.

iv. **Data density:** color can be used to indicate density of data points in specific area. Lighter or dark shades may represent area of higher or lower concentration.

v. **Time series:** different color can be used in line chart or area chart to represent converging and diverting time fluctuation in dataset that can help user to track trend and fluctuation

vi. **Emphasis and highlighting:** color can be used to draw attention to specific data points or area of interest in data visualization. Highlight data can helps to understand main insights.

vii. **Error:** color can be used to represent error interval, range on data points. Lighter color can indicate higher error and uncertainty whereas dark color can represent more emphasis on data points.

- Choice of color is a major factor in creating effective charts. A good set of colors will highlight the story you want the data to tell, while a poor one will hide or distract from a visualization's purpose.
- In this article, we will describe the types of color palette that are used in data visualization, provide some general tips and best practices when working with color, and highlight a few tools to generate and test color palettes for your own chart creation.

Following points should be taken into consideration while choosing appropriate color in visualization:

- Choose a color palette that is appropriate for the data type and message of the visualization
- Avoid using too many colors as this can lead to confusion and convey different information
- Maintain color consistency across related visualization to make easier for comparison

- COLORS SPEAK A LANGUAGE LOUDER THAN WORDS
- COLORS CREATE A CHEMICAL REACTION IN THE BRAIN THAT PRODUCES AN EMOTIONAL RESPONSE
- THEY TRIGGER THOUGHTS, MEMORIES AND ASSOCIATIONS TO PLACES, PEOPLE, AND EVENTS
- LONG WAVELENGTHS COLOR HAVE FASTER RECOGNITION RESPONSE IN THE BRAIN
- SHORTER WAVELENGTH CAN LOWER BLOOD PRESSURE
- MEANINGS OF COLORS AFFECT HOW VIEWERS PERCEIVE DATA.
- Blue vs Yellow | Calm vs anxiety
- LIGHT RED VS DARK RED - Love, sexuality and Joy | anger, passion and confidence
- Do's

1. Influence overall feel and give depth.

--How you want audience to feel about your data.

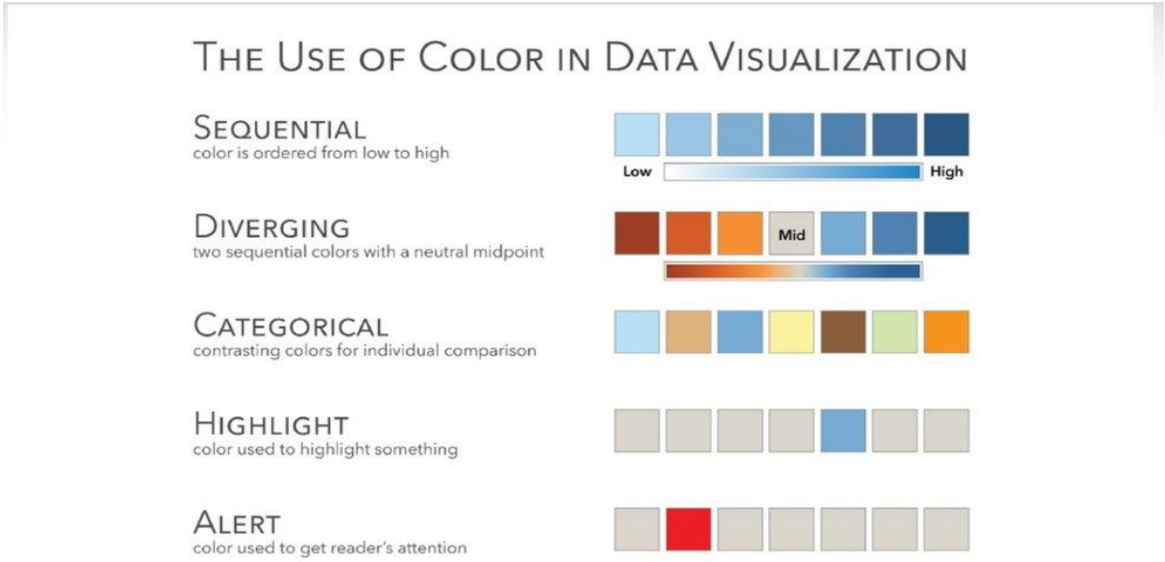
2. Make Important element stand out.

3. Create Contrast.

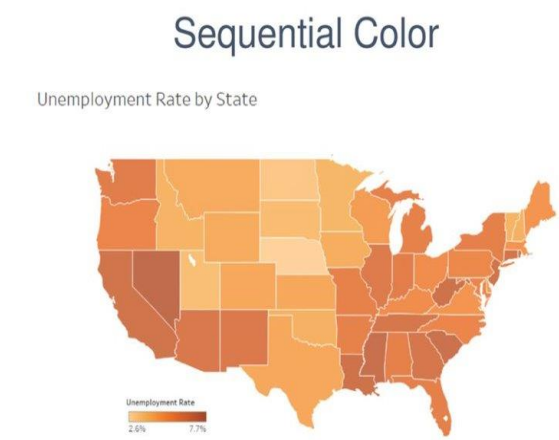
4. Use color consistently

-- Repeated variables across different sets.

5. Use of Gradients



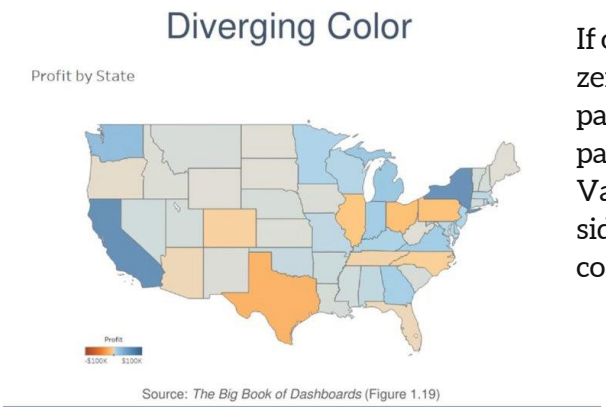
Source: *The Big Book of Dashboards* (Figure 1.16)



Source: *The Big Book of Dashboards* (Figure 1.17)

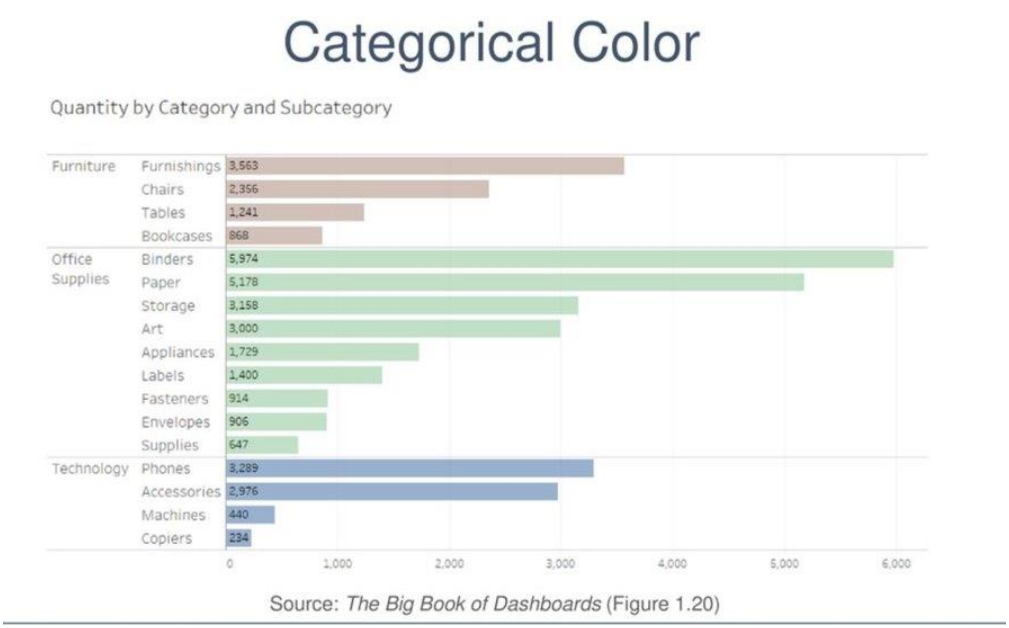
When the variable assigned to be colored is numeric or has inherently ordered values, then it can be depicted with a sequential palette. Colors are assigned to data values in a continuum, usually based on lightness, hue, or both.

The most prominent dimension of color for a sequential palette is its lightness. Typically, lower values are associated with lighter colors, and higher values with darker colors. However, this is because plots tend to be on white or similarly light backgrounds. On a dark background, it's common to have the reverse case,



Source: *The Big Book of Dashboards* (Figure 1.19)

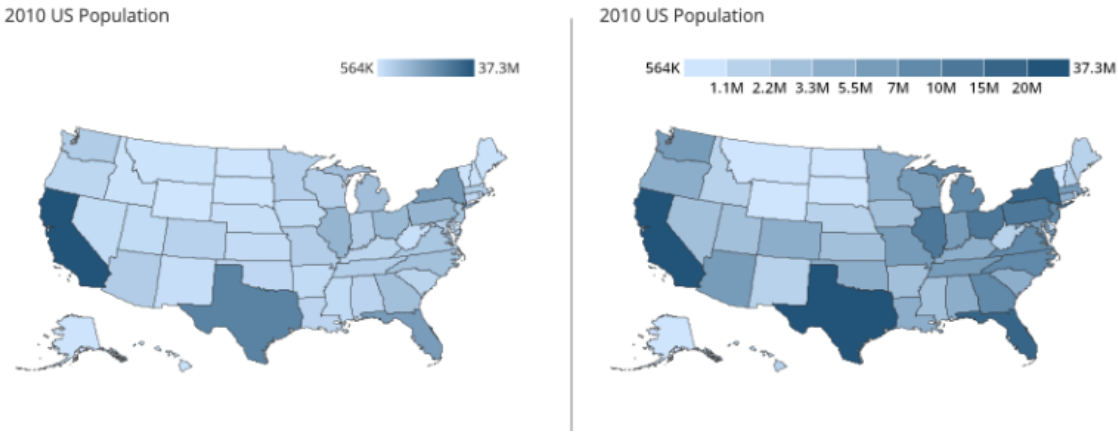
If our numeric variable has a meaningful central value, like zero, then we can apply a diverging palette. A diverging palette is essentially a combination of two sequential palettes with a shared endpoint sitting at the central value. Values larger than the center are assigned to colors on one side of the center, while smaller values get assigned to colors on the opposing side.



Source: *The Big Book of Dashboards* (Figure 1.20)

Discrete VS continuous

- Tools for creating palettes will generally follow the first type, while tools that create visualizations often have the capacity to build a continuous association. While it may seem automatically better to have a continuous function between value and color, there are still merits to the discretized palette.
- Our ability to distinguish differences in color is weaker than that of position or length, so we are already at a disadvantage for associating color to precise values.
- Discretization of values can reduce cognitive load by bringing out the broad patterns in the data.



The Eye with Normal Color Vision

Three types of color sensitive cones

1. Short (S) – respond to short wave lengths
2. Medium (M) - respond to medium wave lengths
   - more sensitive to green colors
3. Long (L) - respond to long wave lengths
   - more sensitive to red colors

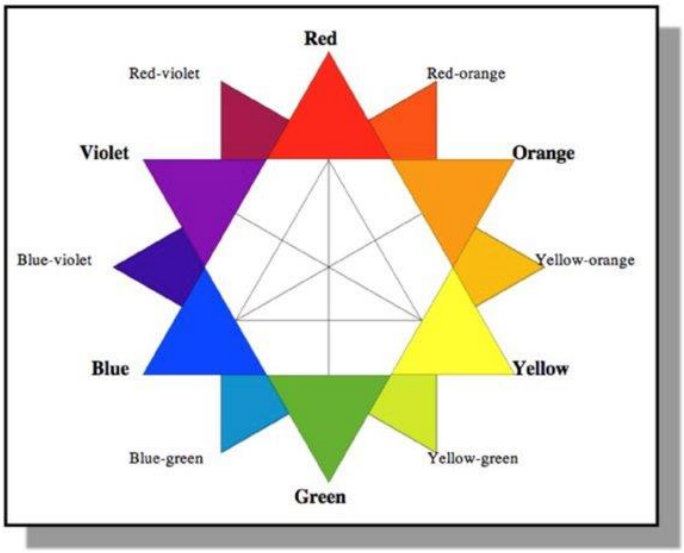
Color Vision Deficiency

2 cones (dichromat)		3 cones (trichromat)	
(L) protanopia	(M) deuteranopia	(L) protanomaly	(M) deuteranomaly
red-blind	green-blind	red-weak	green-weak
1 out of 100 men	1 out of 100 men	1 out of 100 men	5 out of 100 men

no perceptible difference between red, orange, brown, and green

= approximately 8% of men have color vision deficiency

Color



Perceptual issues:

Perceptual issues are those that can lead to misinterpretation of data due to perception of human brain i.e., it refers to challenges and considerations related to how human perceive and interpret visual representation of data. These issues should be addressed as effective data visualization depends on how well it conveys information to the viewer. Addressing perceptual issues in data visualization requires a deep understanding of both the data being presented and the target audience. It is important to continually test and refine visualization to ensure that they effectively convey the information and minimize misinterpretation. Some of the common perceptual issues are:

- Color choice: the human eye is more sensitive to some color that others so using such similar color and dissimilar color together can make difficult for user to understand the meaning of data. Using inappropriate color can lead to confusion of data for eg. Using red color and green color together for danger symbol, choosing a color scale that is not perceptually uniform can also distort the representation of quantitative data.

- Size and position: using size too similar for different situation can make it difficult to see the difference in data values. Distortion in scale such as unequal axis, intervals exaggerated size in chart and graph can misrepresent the data and lead to inaccurate interpretation. Placing the data points in unexpected or cluttered location can make the information difficult to understand.

- Misleading visualization types: choosing wrong types of visual eccentrics like chart and

  graphs for data can result in misinterpretation. For example: using pie chart to show time series data, using heat map to represent one dimensional data etc.

- Labeling and annotation: improper labeling and annotation can misinterpret the meaning of data and leave viewers unsure about the context.

- Data density: too much information can overwhelm the user while too little can lead to miss the main points on decision making. So, how much data is to represent should be clearly analyzed.

- Complexity: human eye can only process limited amount of visual information so using visualization that are too complex can make it difficult to understand the data.

To overcome perceptual issue, following points should be taken into consideration:

- Using limited number of colors
- Using sizes that are clearly differentiated
- Using familiar shapes that are not distorted
- Keeping visualization simple and easy to understand
- Keeping limited amount of data for visualization.

What is data?

- Raw Facts
- It simply exists and has no significance beyond its existence
- Data Preprocessing
- Character/ Set of Characters that are collected, Why?
- They are put into Certain Context. Why?
- May be useable , may be not
- No added interpretation or analysis

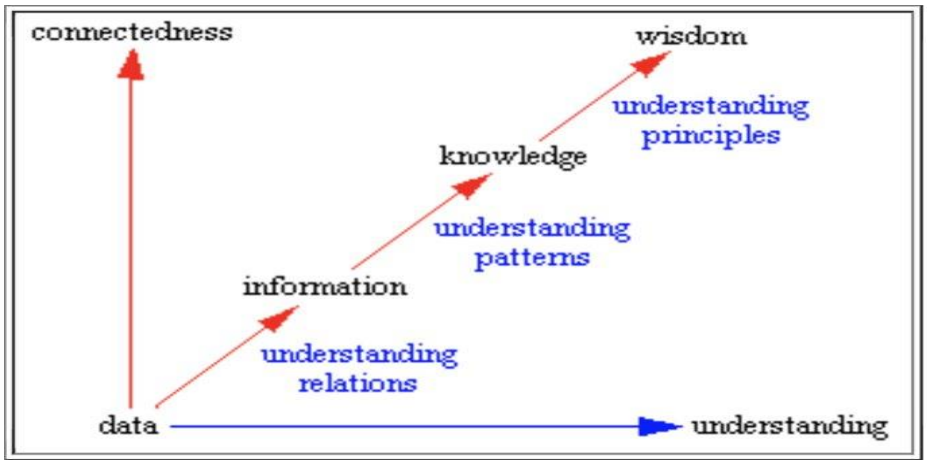
Types of Data

- Single character
- Boolean
- Text

- Number
- Images
- Audio/Video

What can we do with Data?

- Transformation of Data towards Wisdom



Data Flow

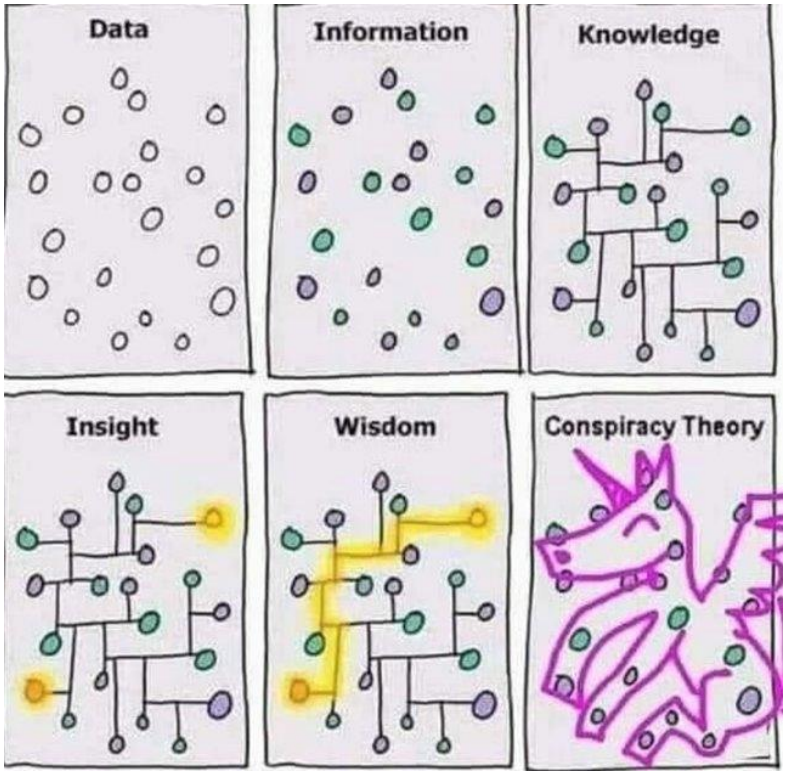
Data - Symbols - Observation

- Price of petrol is Rs.100 per litre

Information

 Descriptive - Who, What, Where, When

 yesterday the price of petrol rose from 80 to 100 per litre in nepal



Knowledge = Information + Information - How, Why - Instruction + Understanding - High Demand less Supply, Rise of petroleum price in International market.

Wisdom - When Knowledge is evaluated - Futuristic and predictive.

Data Analysis

A process of
- Cleaning
- Transforming and
- Modelling

Useful Information

Decision Making

For ?
- Useful Information
- Decision Making

Process and Types

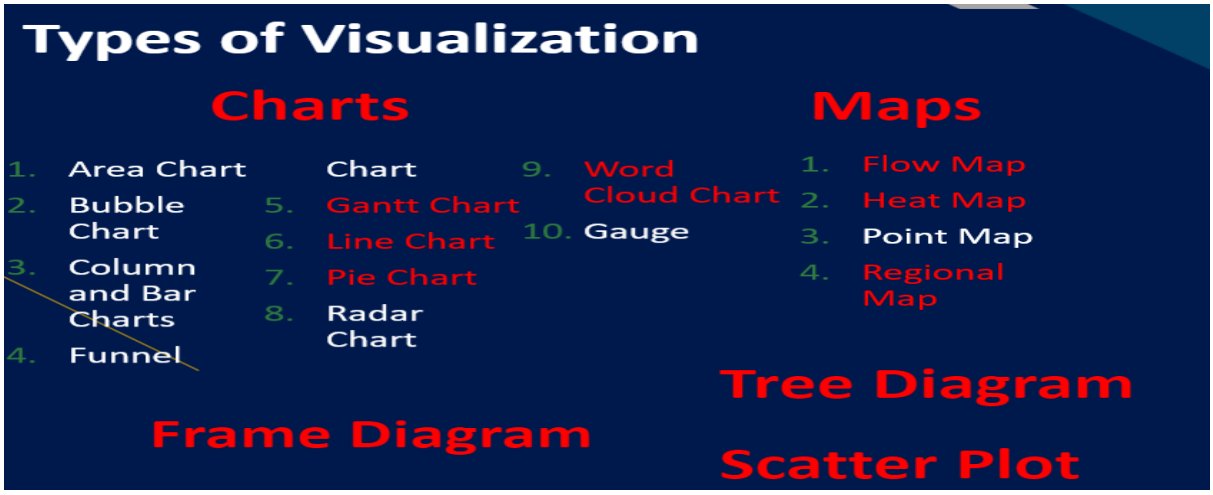
- Data Requirement Gathering
- Data Collection
- Data Cleaning
- Data Analysis and Interpretation
- Diagnostic Analysis
- Predictive Analysis
- Prescriptive Analysis
- Statistical Analysis
- Text Analysis -> Data Mining

Finding Datasets to Practice

- **The Museum of Modern Art (MoMA) Collection (GitHub)**
- **Covid Dataset (GitHub)**
- **IMDB datasets ( )**

Data Visualization

- We are visually oriented creatures.
- Images and displays attract our attention and stay in our memory longer.
- Graphical representation of information and data
- Helps to see and understand trends, outliers, patterns


Types of Visualization
Charts
1. Area Chart
2. Bubble Chart
3. Column and Bar Charts
4. Funnel Chart
5. Gantt Chart
6. Line Chart
7. Pie Chart
8. Radar Chart
9. Word Cloud Chart
10. Gauge
Maps
1. Flow Map
2. Heat Map
3. Point Map
4. Regional Map
Tree Diagram
Frame Diagram
Scatter Plot

Information Overload:

Information overload is the situation where the amount of data presented is too much for the user to understand or process which can happen if there is too much data in a single visualization or when multiple visualization is presented together. Information overload can lead to confusion and decrease ability to understand the data. Information overload can lead to following problems:

- The user may become confused and unable to understand data
- The user may not be able to find information they are searching for.
- The user may become frustrated and give up trying to understand data

Following are the elements that can lead information overload:

- Excess data points: includes excess number of points such as categories, series.
- Complex visual elements: includes complicated charts or graph, multiple lines etc
- Redundant information
- Too many variables
- Lack of hierarchy: if all element are treated equally then it can be challenging for viewer to prioritize and understand the most crucial information.
- Real time data: constant updates and information change can overwhelm the user
- Insufficient context: insufficient explanation can leave viewers in confusion.
- Poor organization of data
- Overuse of color: excessive use of color, styles or formatting options can create visual clutter and make it difficult to focus on the data.

Following factor should be consider to address information overload

- Simplify data: reduce the amount of data and focus on most critical variable and insight
- Prioritize information: highlight the most important data points or trends to guide the viewer attention. Use visual cues, annotation to drew attention
- Use limited interaction: provide clear instruction and options for customization
- Provide context: include clear title, labels, annotation, to provide context and guide. Use caption or summary that can provide more insight to viewer.
- Use hierarchy and consistent design

Quality and Value of information

- The quality and value of information can be described as how it contributes for effective decision making
- The quality of information is high, if it creates managerial impact leading to attention, decision and action

Value of Information

- *Timeliness*
- *Presentation*
- *Accuracy*
- *Completeness*
- *Integrity*

The quality of information can be measured on the four dimensions

- *Utility*
- *Satisfaction*
- *Error*
- *Bias*

1)The utility Dimension

- The utility dimension has four components namely the form, the time, the access and the procession
- If the information is presented in the form of manager requires, then its utility increases
- If it is available(time) when needed, the utility is optimized
- If the information is easily and quickly accessible through the online access system, its utility gets an additional boost
- If the information is processed by the manager who needs it, then its utility is the highest.

2) Satisfaction Dimension

- The degree of satisfaction would determine the quality of the information
- If the organization has a high degree of satisfaction, then one can be safe in saying that information systems are designed properly to meet the information needs of the managers at all the levels

2) Error Dimension

The error creep in on account of various reasons, namely:

- An incorrect data measurement
- An incorrect collection method
- Failure to follow the prescribed data processing procedure
- Loss of data or incomplete data
- Poor application of data validation and control systems
- A deliberate falsification
- The data should be avoided of errors, care should be taken that the information is processed after ensuring the correctness of the data in terms of time and the number of document, and the transactions in the period

4) Bias Dimension

- The procedure of communicating the information should be such that the system is able to detect the degree and the nature of the bias and correct the information accordingly

Information Overload

- It is a term popularized by Alvin Toffler
- Information overload occurs when the amount of input to a system exceeds its processing capacity
- It is simply the inability to digest and apply all of the information provided
- It results in anxiety and stress, delay in decision making, lack of job satisfaction, waste of time and working longer hours

Information Overload

- It refers to difficulty of a person can have understanding an issue and making decisions that can they caused by the presence of too much information
- It is a situation where a manager is given too much information and as a result his is confused and cannot make optimal decisions
- Information overload occurs when the amount of input to a system exceeds its processing capacity

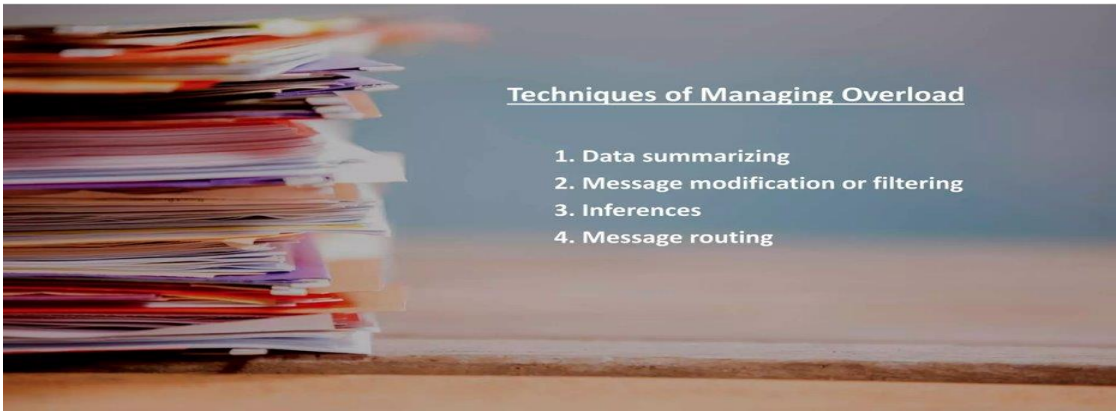
Causes of Information Overload

Causes of Information Overload

- Increasing new information
- Duplication and transmission
- Increase channels of incoming information (e.g. telephone, e-mail, instant messaging)
- Historical information
- Contradictions and inaccuracies
- A low signal-to-noise ratio
- A lack of comparing and processing

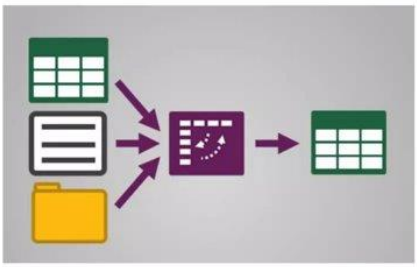
Tips to solve information overload problem

1. Focus and specialize in one thing
2. Take Control
3. Follow only valuable sources
4. Unsubscribe from most of the unwanted subscription
5. Organize the time

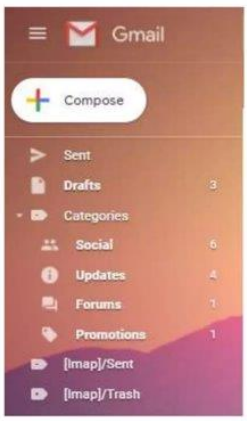


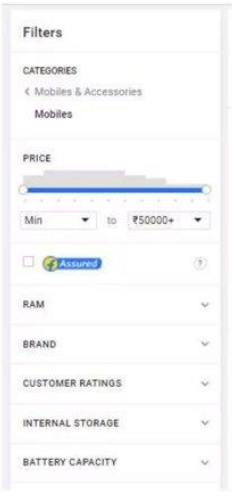
Data Summarizing

- Reduction of data to a meaningful and concise form
- Summarization is a commonly used method of data reduction and it refers to the reduction of data to a meaningful and concise form
- Summarization reduces the amount of data transmission with out changing the essential meanings of the original message



Message Modification or Filtering





3. Inferences

- Process of reaching a conclusion based on facts or evidence
- **"Conclusion based on facts"**

- "This must be either a new pencil or one which has not been used a lot"

"a long pencil with an eraser"

- "This must be either an old pencil or one which has been used for a long period"

"a small pencil without an eraser"