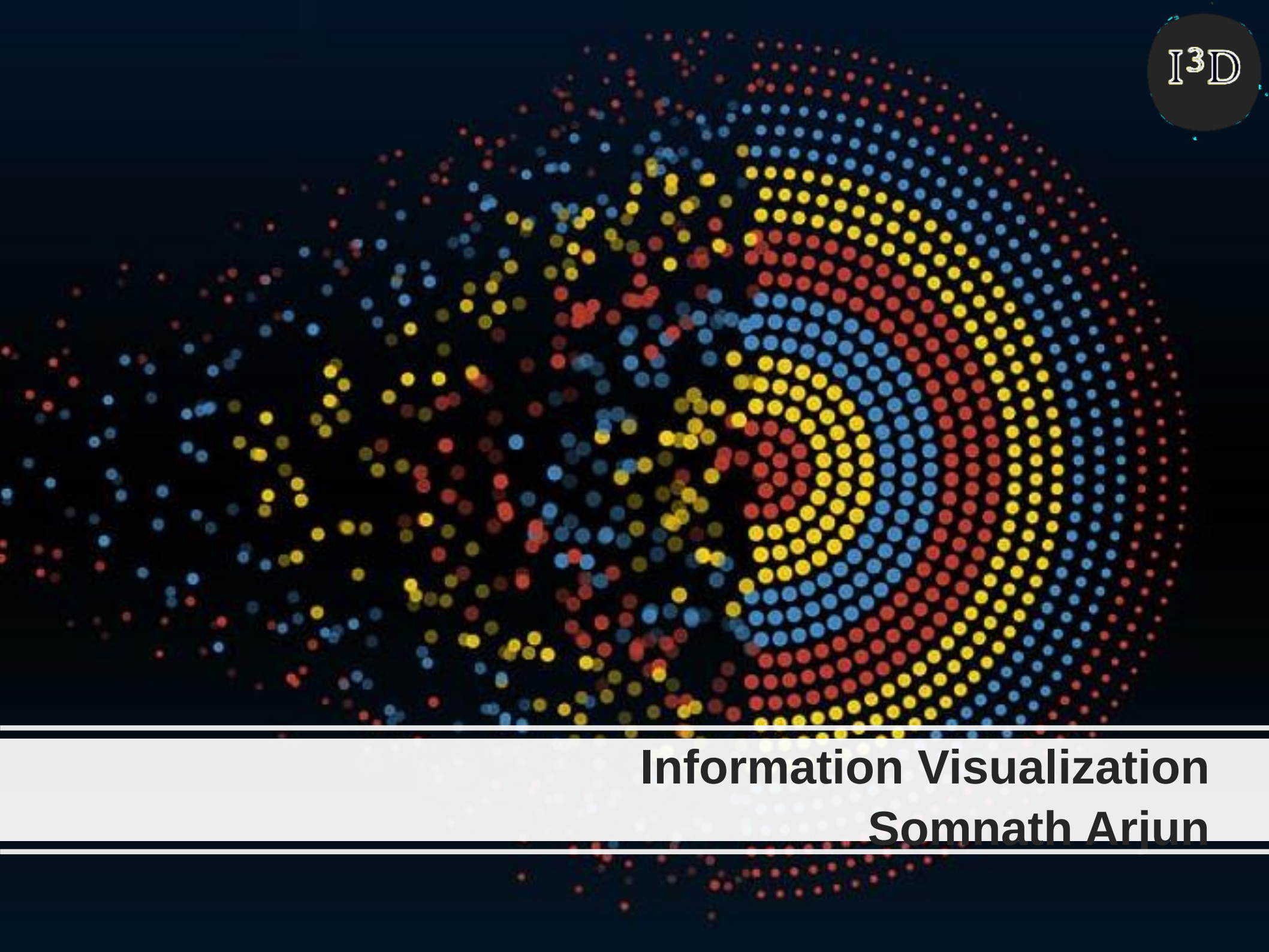


Information Visualization

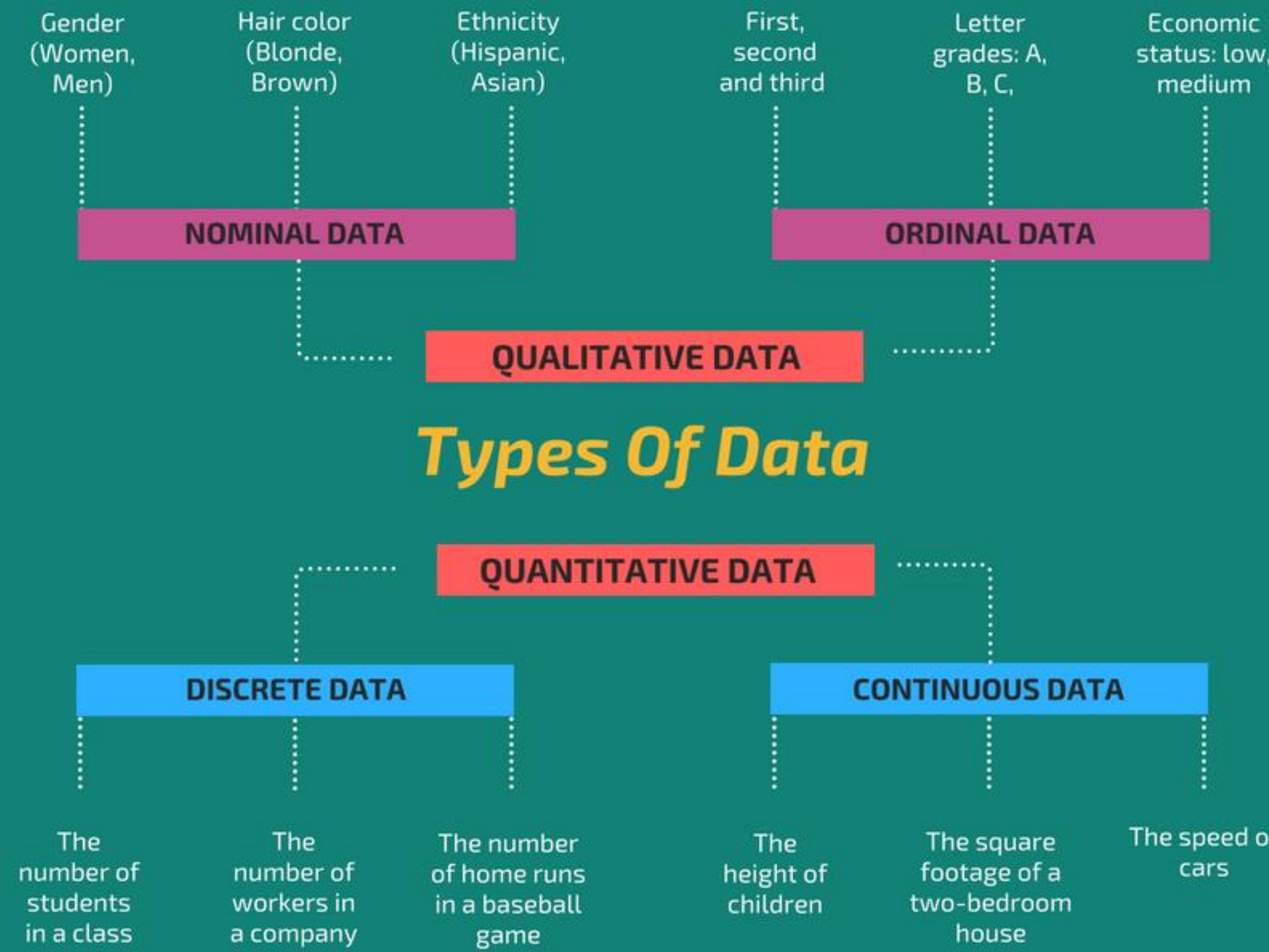
Somnath Arjun

Contents

- What is Information
- What is visualization
- Why Visualization
- One Dimensional Visualization
- Two Dimensional Visualization
- Three Dimensional Visualization
- High Dimensional
  - Pipeline
  - Data Transformation
  - Visual Mapping
  - View Transformation
- Scalar Visualization
- Geo visualization

Information

- Facts or details about someone or something
- Information also called data are collected through observations
- Quantitative and Qualitative data
- Quantitative Data can be expressed as a number or can be quantified. They can be measured by numerical values
- Qualitative Data cannot be expressed as a number and cannot be measured. They consist of words, pictures, symbols but not numbers

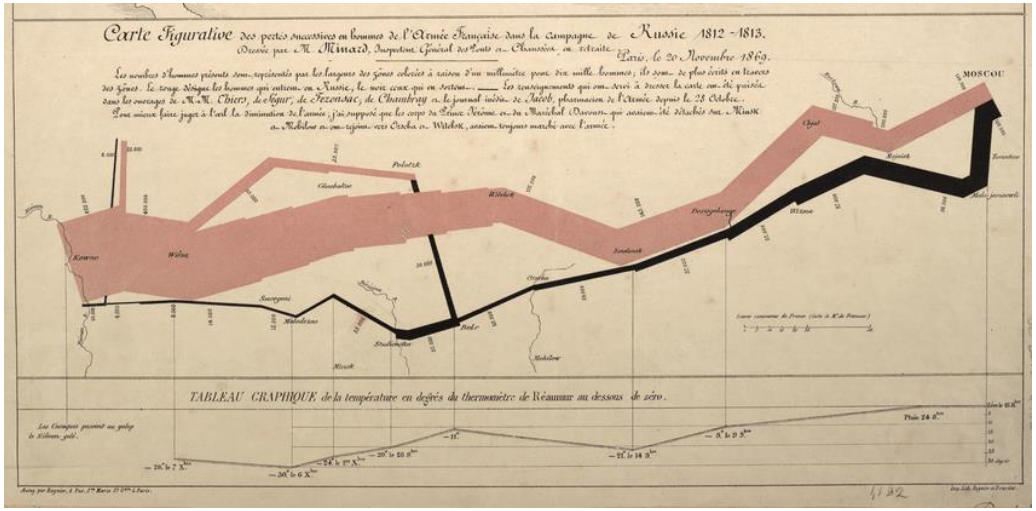

Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed o cars

Why and What

- Huge and complex Datasets : Difficult to understand the information hidden, measure features, finding relation of interest, observe patterns and structure
- Communicate between known features
- Explore data to identify features

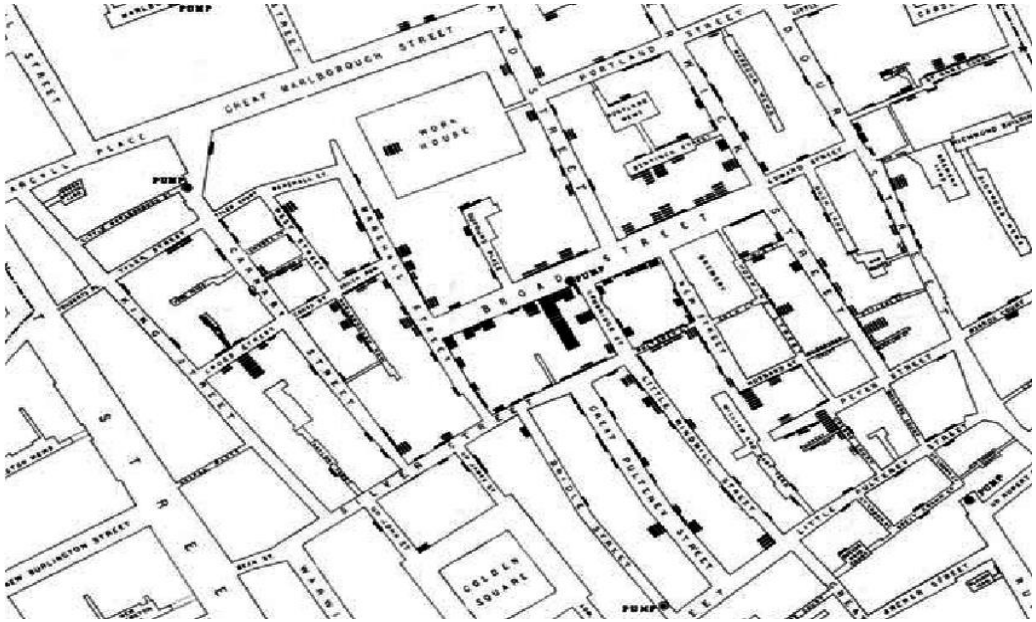
- How to convey information graphically
- How to gain insight by looking at the data
- How to take better advantage of human perceptual system to extract meaning from the data, focus attention, reveal structure and patterns
- How to map data to visual objects

Motivations



Napolean March Map

In 1812, Napoleon marched to Moscow in order to conquer the city. It was a disaster: having started with around 470,000 soldiers, he returned with just 10,000. This chart tells the story of that campaign and has become one of the most famous visualizations of all time. The map details the out-and-back journey of Napoleon's troops. The width of the line represents the total number of soldiers and the color represents the direction (yellow for towards Moscow, black for the return trip). Below the central visualization is also a simple temperature line graph illustrating the rapidly dropping winter cold



Cholera Outbreak Map

It uses small bar graphs on city blocks to mark the number of cholera deaths at each household in a London neighborhood. The concentration and length of these bars show a specific collection of city blocks in an attempt to discover why the trend of deaths is higher than elsewhere. The finding: the households that suffered the most from cholera were all using the same well for drinking water.

Goals of Visualization

Explore/Calculate

- Analyse
- Information Retrieval

Communicate

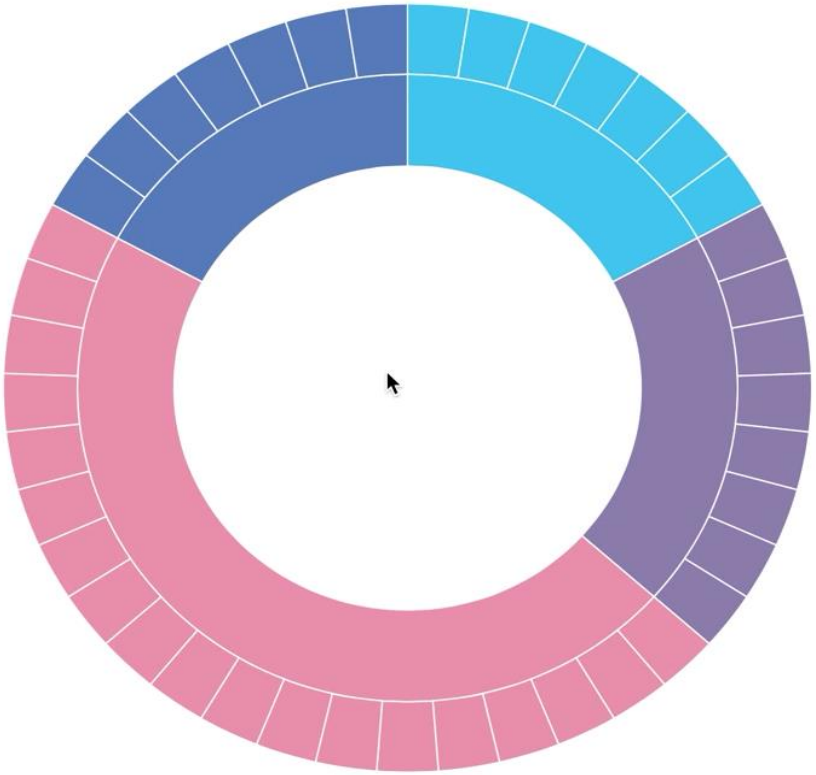
- Explain
- Make decisions
- Reason about Information

Goals of Visualization

- Make large datasets coherent (Present huge amounts of information compactly)
- Present information from various viewpoints
- Present information at several levels of detail
- Examine data relationships

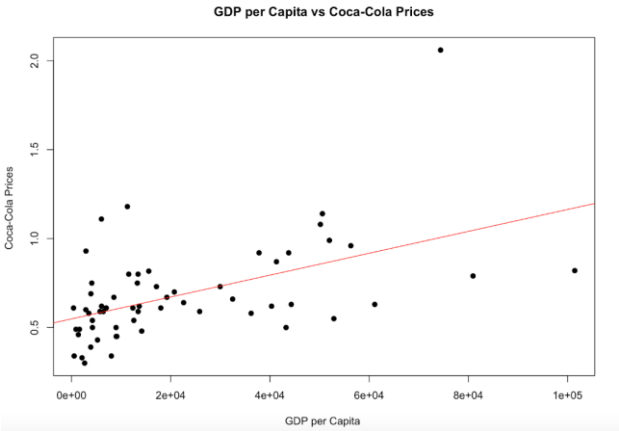
Information at several levels of detail

Sunburst is classic example for getting information at several levels of details



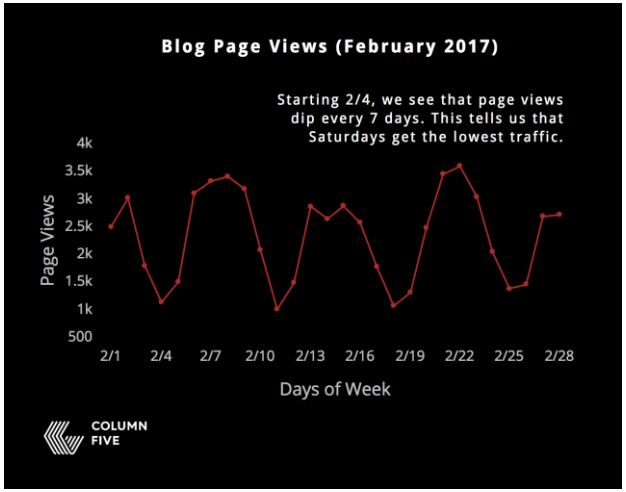
Examine data relationships

- Correlation



Scatterplot with a fitted line that shows the relationship between GDP per Capita and Coca-Cola prices for different countries. The line shows that there is a positive relationship. Through visual inspection we can see the dots don't make a perfect line, so we can say the correlation is only moderately strong. In fact, after calculating Pearson's R, the correlation coefficient is 0.51.

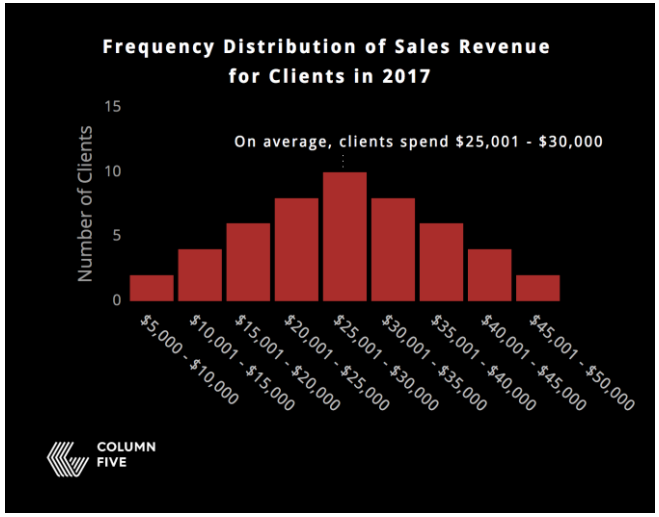
- Trends



Line charts showing how many page views your website gets every day in a month to identify which days of the week generate the most traffic.

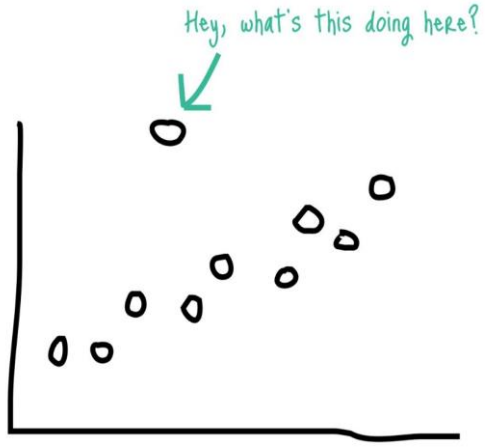
Examine data relationships

- Distribution



Histogram that shows what the average client spends, as well as the range a client might be expected to spend.

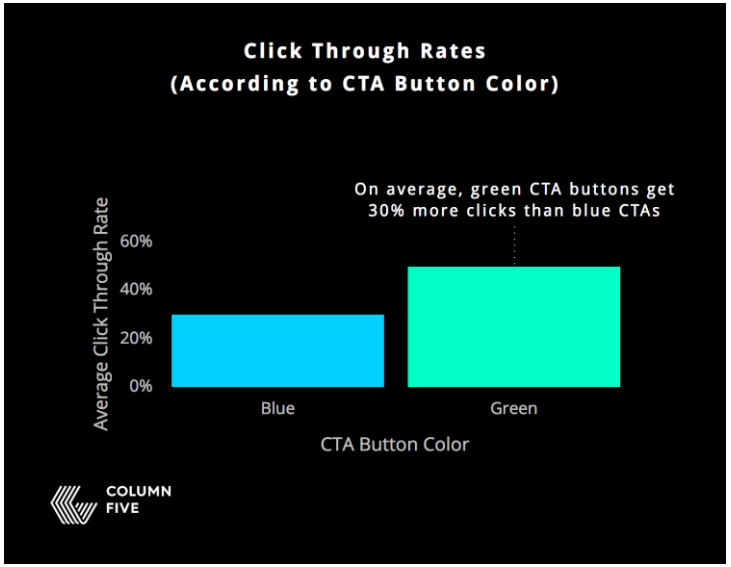
- Outliers



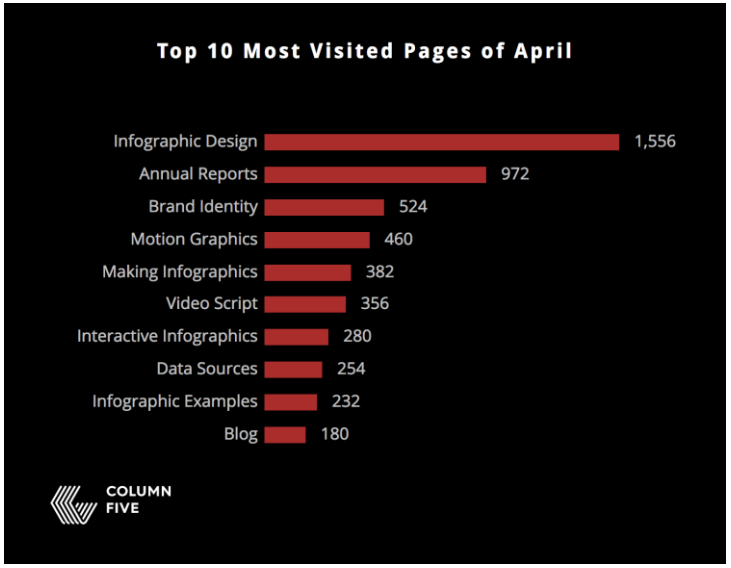
Scatterplot showing data point which is at distant from other similar points

Examine data relationships

- Visual Comparison

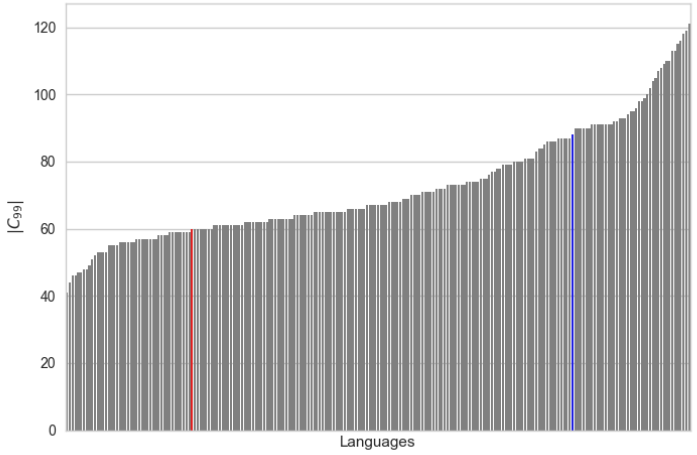


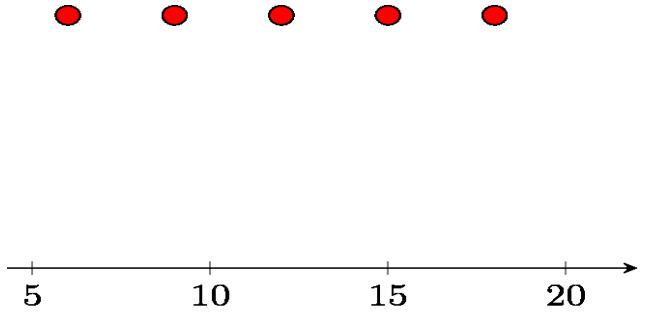
Bar charts showing data comparing click through rates for different colored CTA buttons.



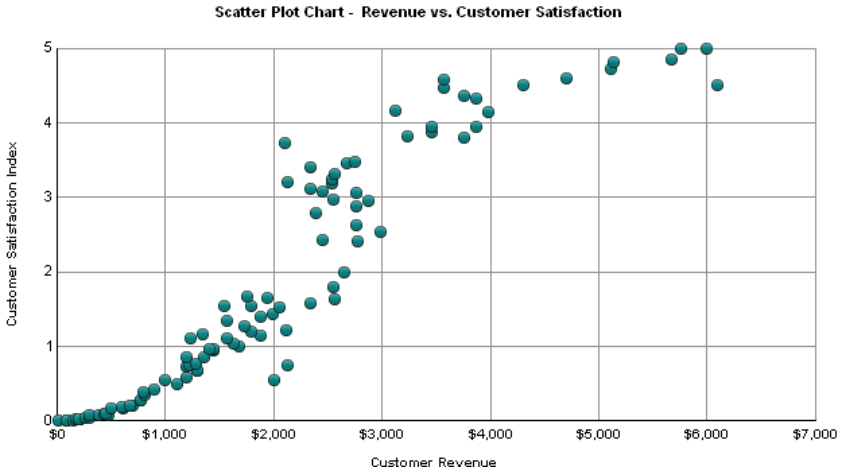
Horizontal bar charts helps us easily compare how much traffic a page is generating.

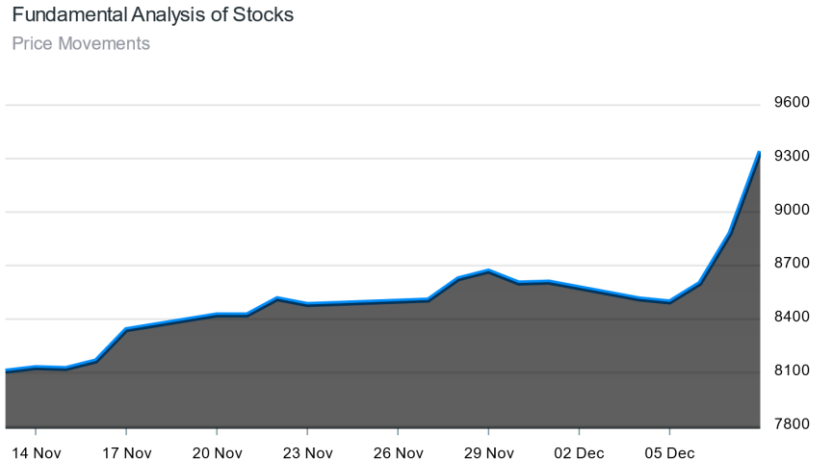
One Dimensional

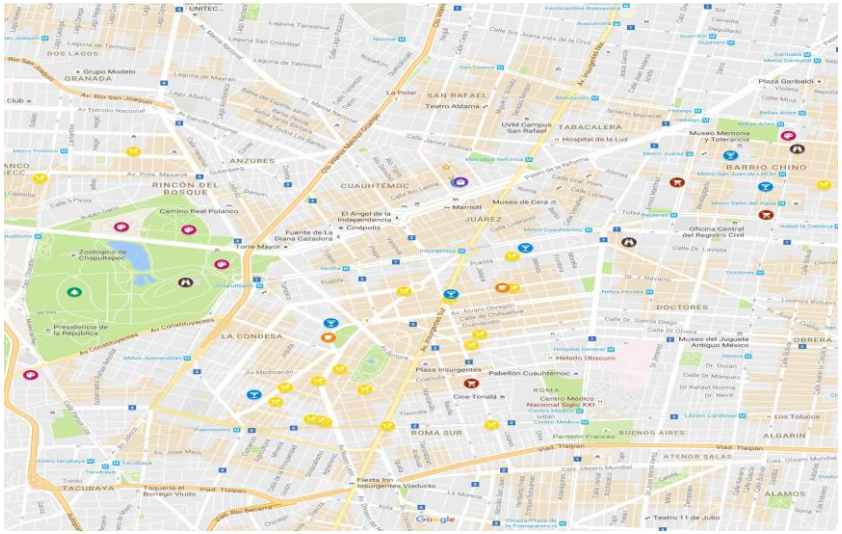


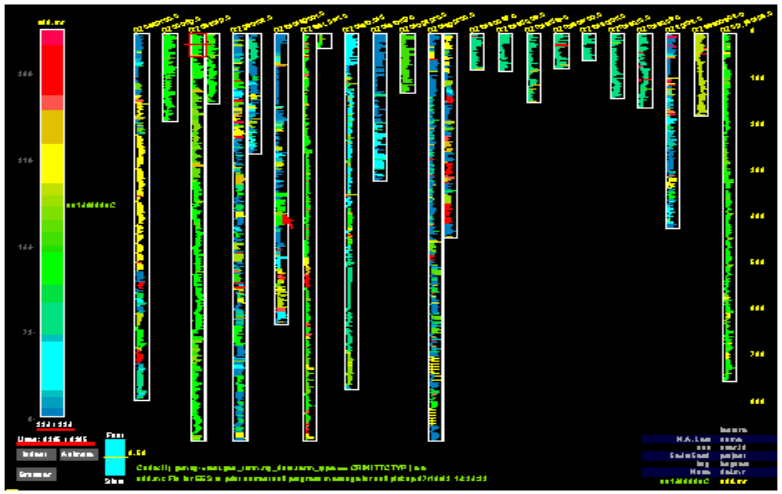


Two Dimensional

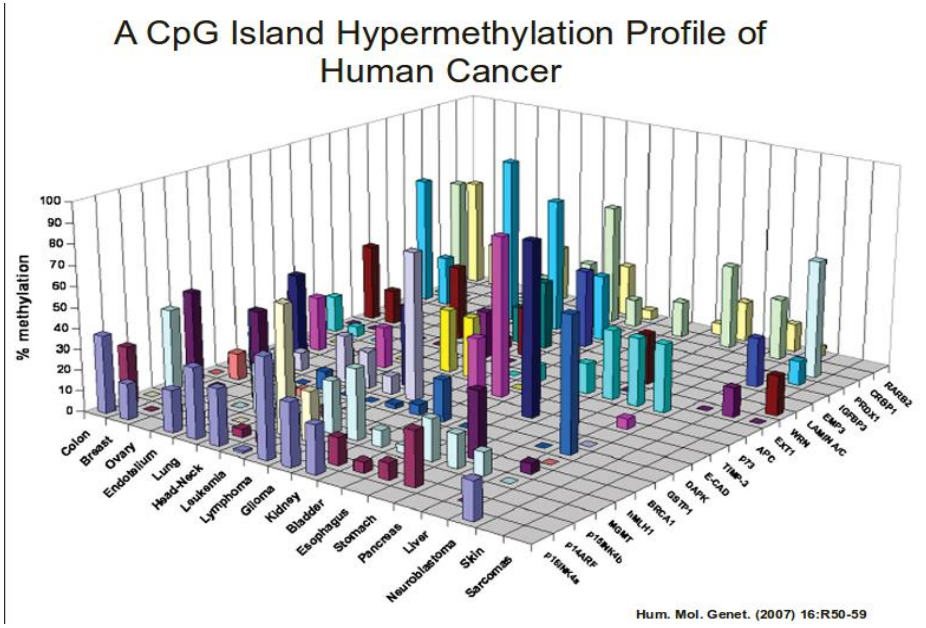


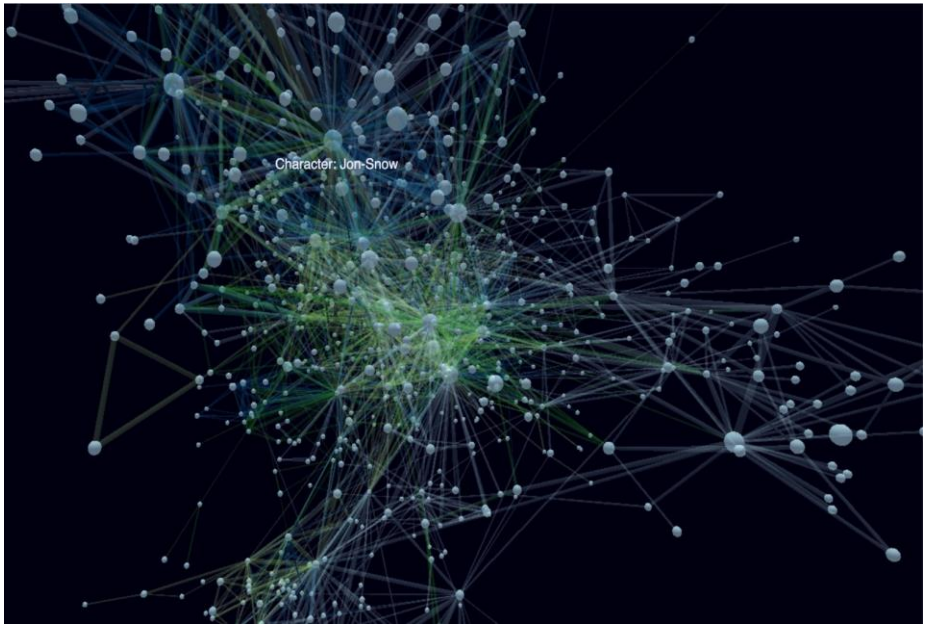


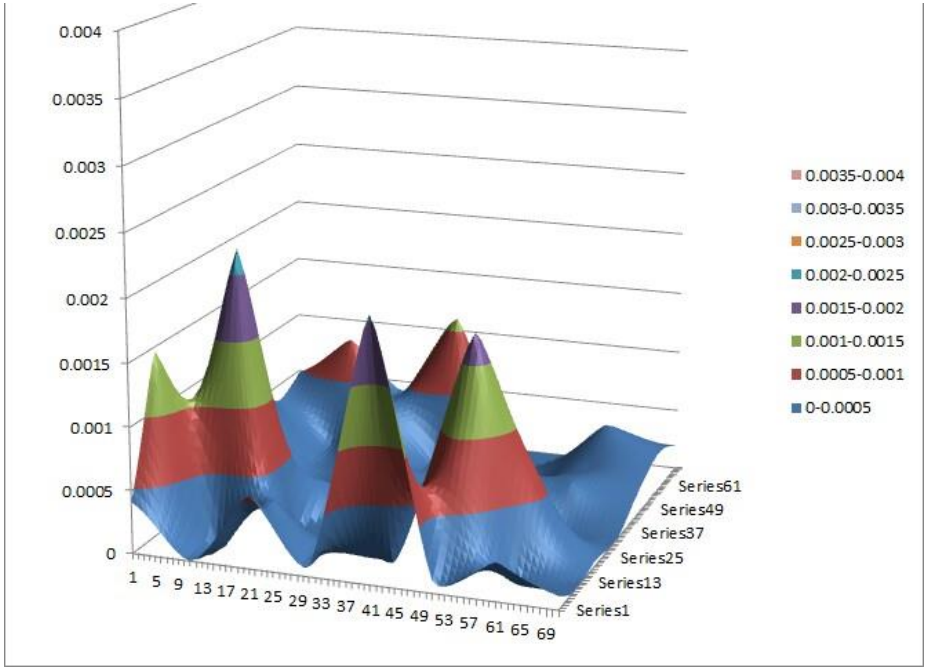


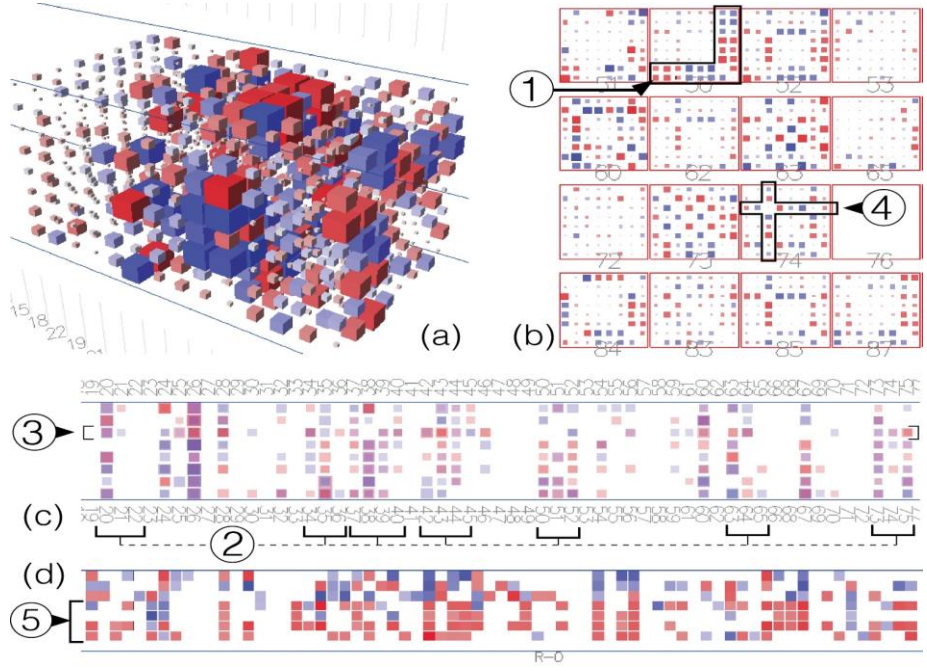


Three Dimensional

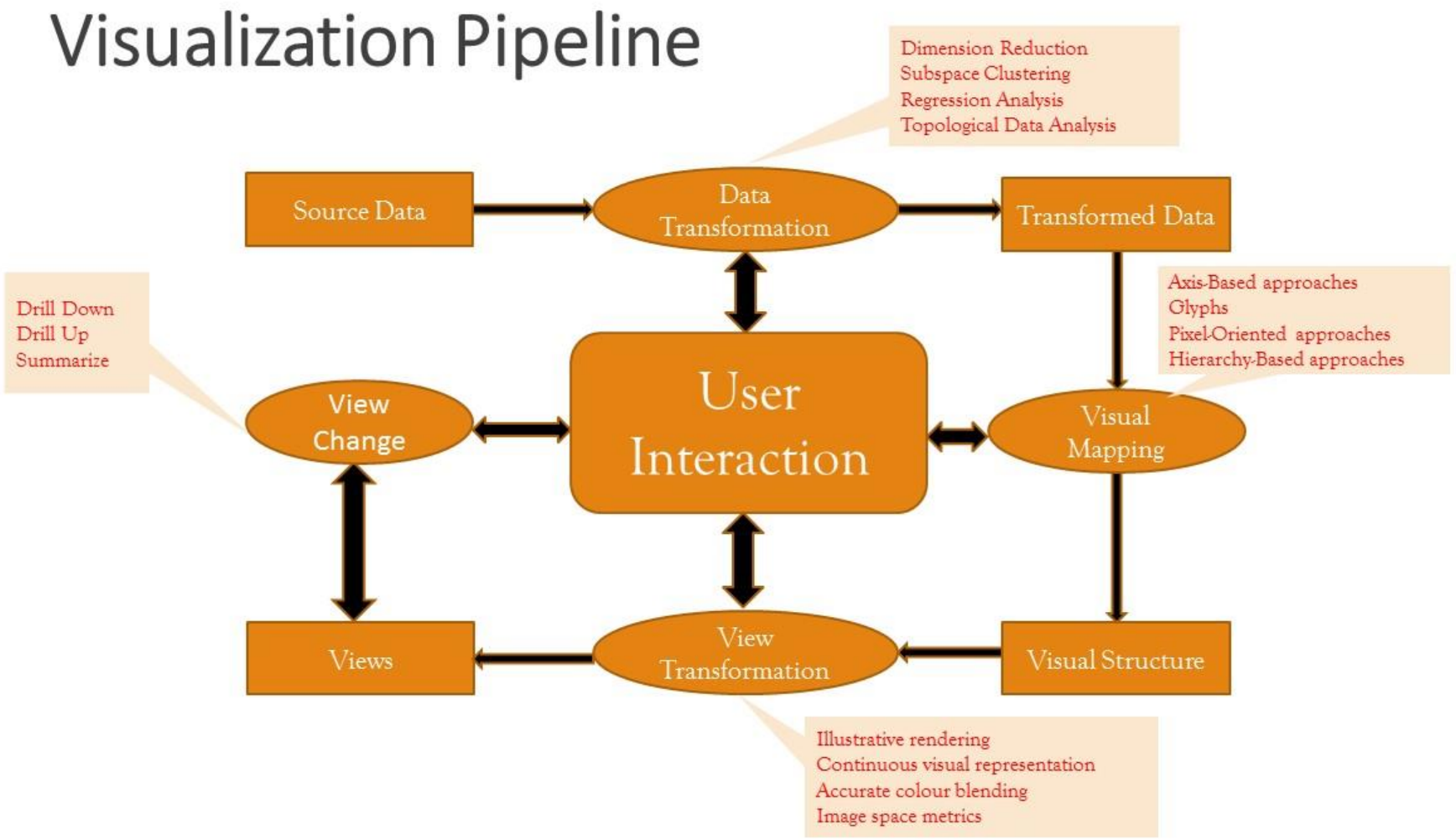








High Dimensional : Pipeline



Data Transformation

Data transformation is the process of converting **data** or information from one format to another, usually from the format of a source system into the required format of a new destination system.

1. Dimension reduction
2. Subspace clustering
3. Topological Data Analysis

Dimension Reduction

It is the process of reducing the number of random variables or features in a data record under consideration, by obtaining a set of principle variables

- PCA
- MDS
- LDA
- Factor Analysis
- t-SNE
- Autoencoder
- Isomap

PCA (Principal Component Analysis)

PCA reduces dimensionality of large datasets by transforming a large set of variables into smaller one that still contain most of the information in the large set

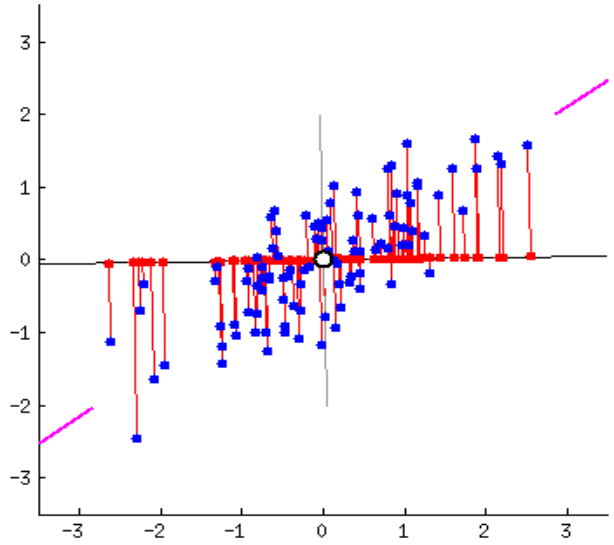
When to use PCA :

- Reduce the number of variables, but not able to identify variables to completely remove from consideration
- Ensure variables are independent of one another
- Feature transformation will help remove irrelevant information

- Standardization
- Covariance matrix computation
  - Why covariance
  - What this matrix tell us about the correlations

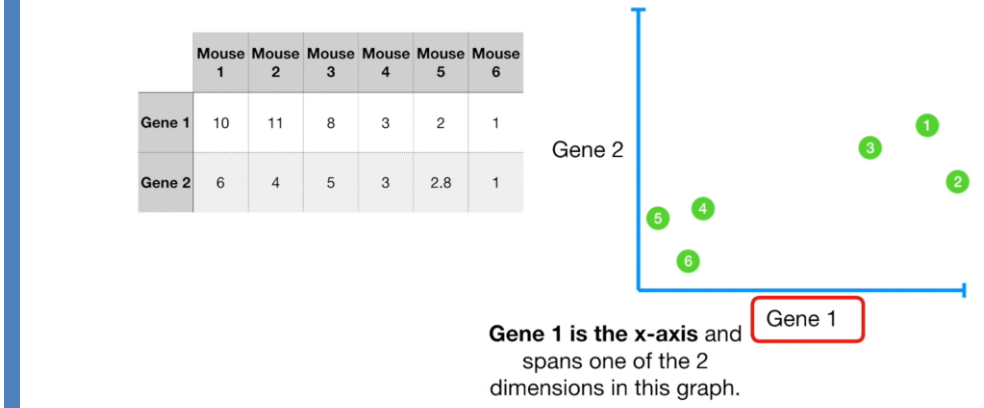
$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

- Compute Eigen vectors and Eigen values of the covariance matrix to identify principal components
  - What are principal components
  - What principal components capture
  - How principal components are constructed



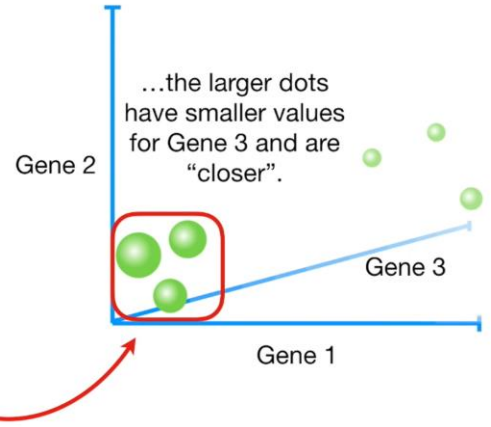
Without Covariance matrix

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



Gene 1 is the x-axis and spans one of the 2 dimensions in this graph.

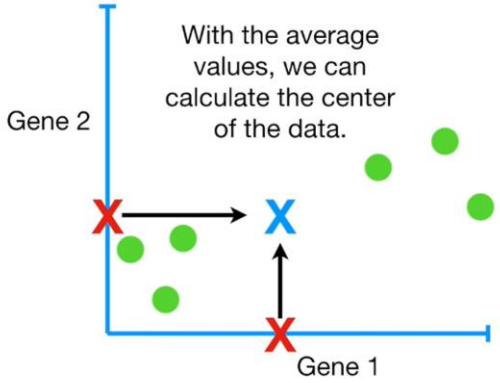
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2

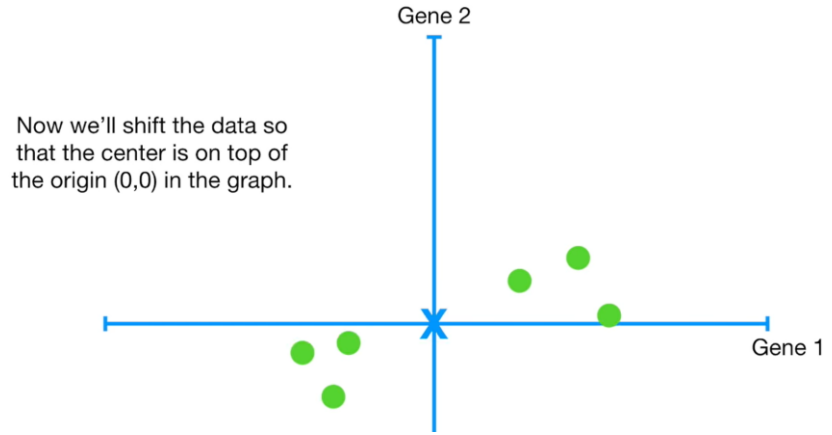


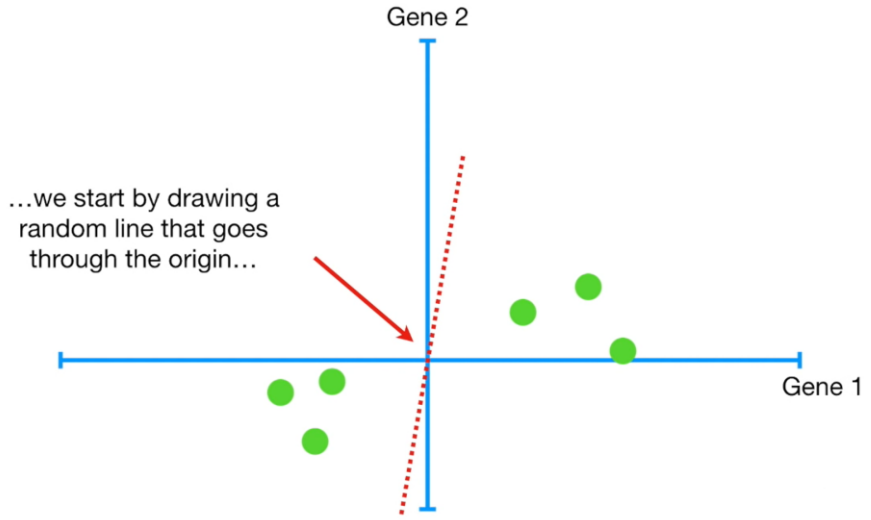
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

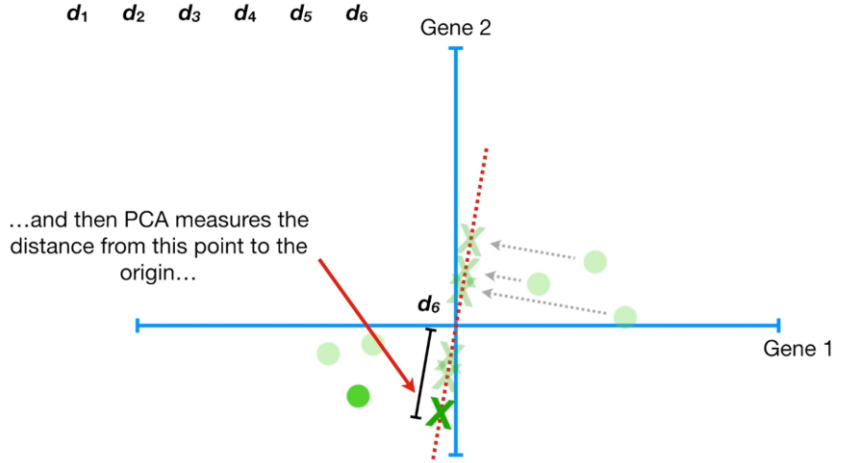
If we measured 4 genes, however, we can no longer plot the data - 4 genes require 4 dimensions.

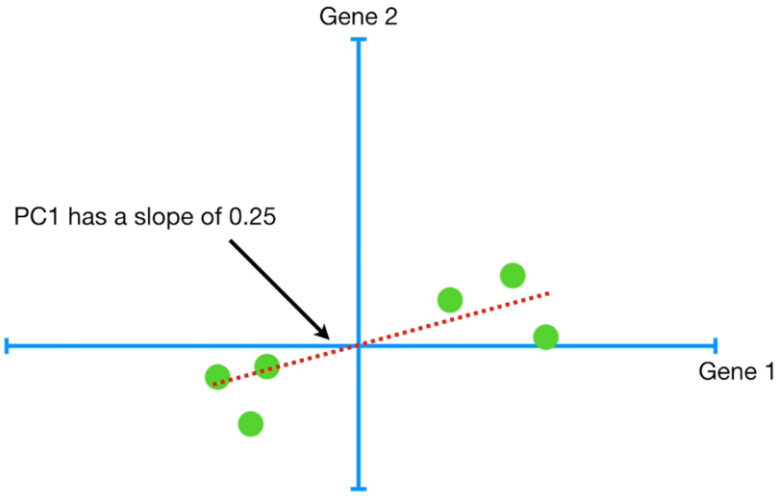
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

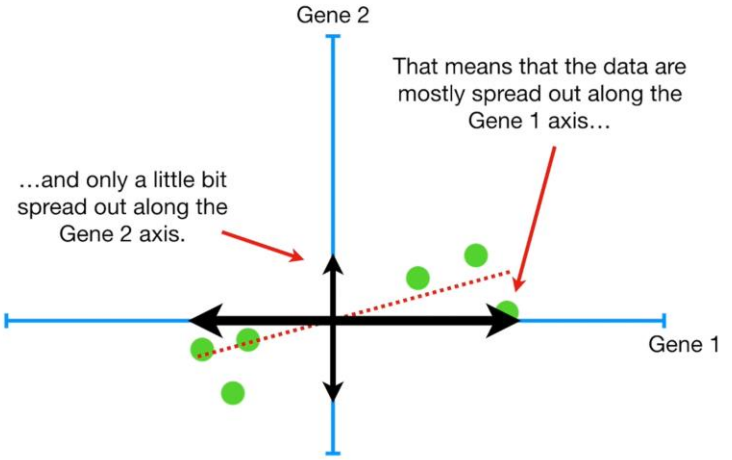






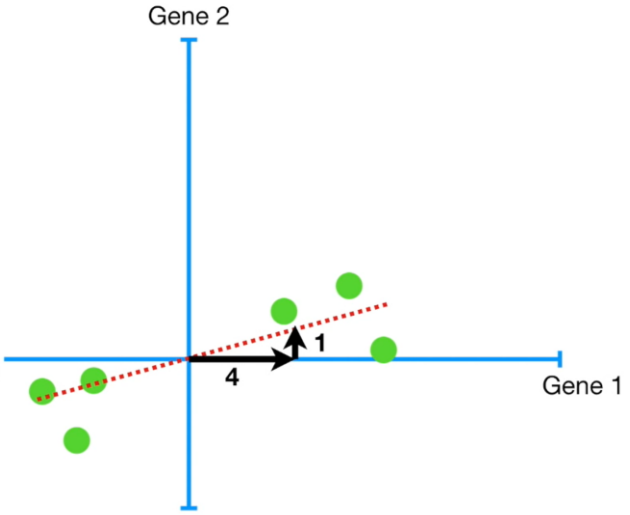






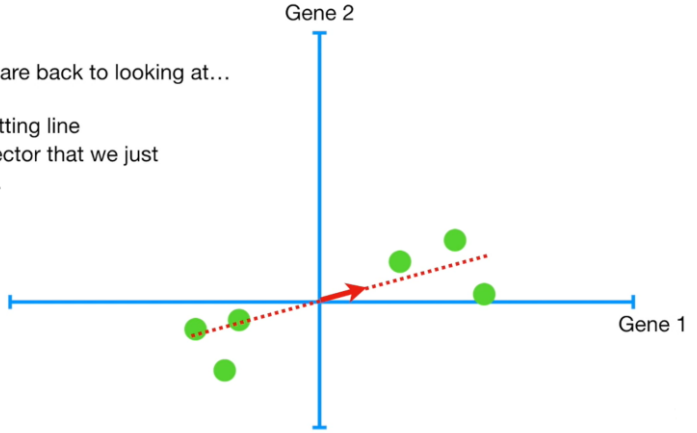
To make PC1
Mix **4** parts Gene 1 with **1** part Gene 2

The ratio of Gene 1 to Gene 2 tells you that Gene 1 is more important when it comes to describing how the data are spread out..



So now we are back to looking at…

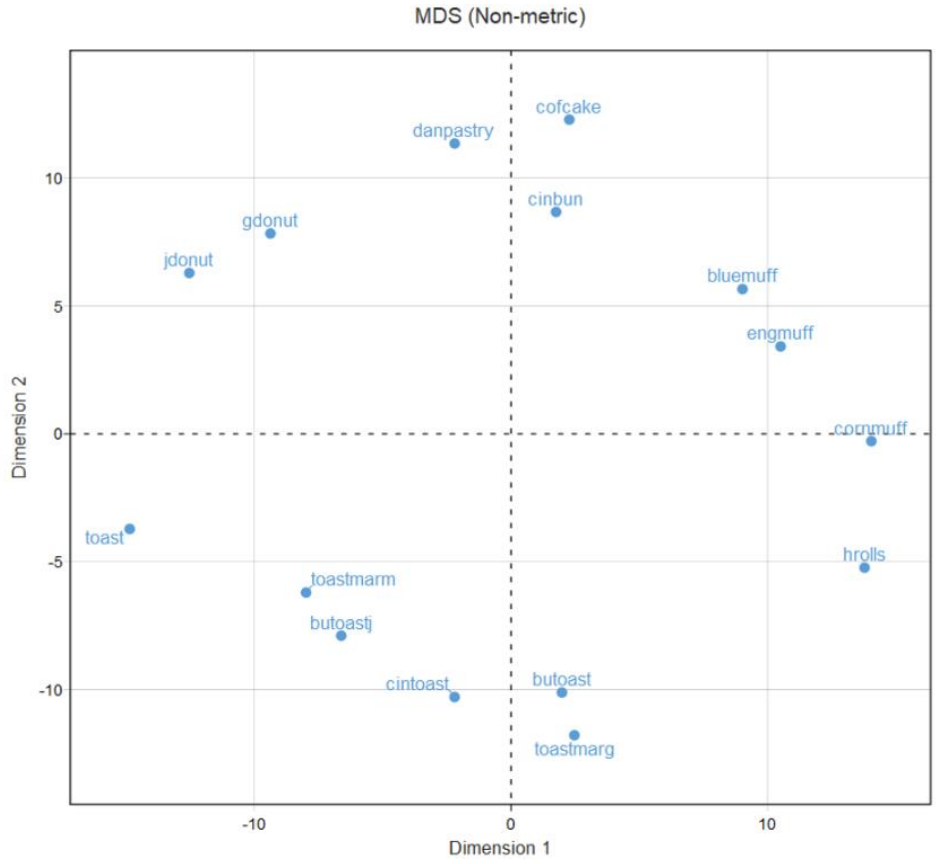
- The data
- The best fitting line
- The unit vector that we just calculated.



MDS (Multidimensional Scaling)

Techniques for visualizing distances between objects, where the distance is known between pair of the objects. The input to MDS is a distance matrix and the output is typically a two-dimensional scatterplot, where each of the objects is represented as a point

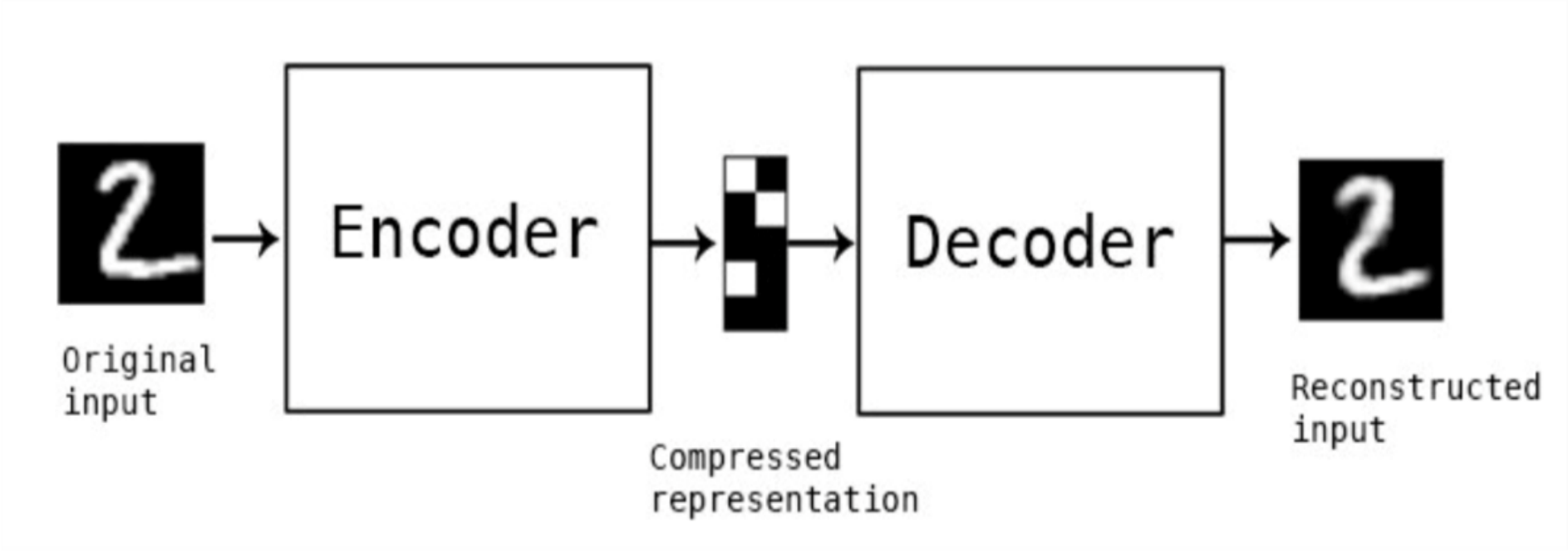
	toast	butoast	engmuff	jdonut	cintoast	bluemuff	hrolls	toastmarm	butoastj	toastmarg	cinbun	danpastry	gdonut	cofcake
butoast	15													
engmuff	25	15												
jdonut	3	24	22											
cintoast	14	3	17	22										
bluemuff	24	17	2	21	19									
hrolls	28	8	4	27	18	8								
toastmarm	7	7	20	11	6	18	23							
butoastj	8	6	21	12	5	19	22	2						
toastmarg	16	2	16	25	4	18	9	8	7					
cinbun	26	17	10	17	12	7	18	20	19	18				
danpastry	21	25	11	5	19	10	22	17	16	26	2			
gdonut	20	18	24	2	23	22	25	11	12	17	4	11		
cofcake	16	22	11	13	21	7	21	21	20	23	6	7	11	
cornmuff	27	11	3	26	16	4	5	25	24	12	12	16	24	16



Autoencoder

Type of artificial neural network used to learn efficient data codings in an unsupervised manner. The aim of an autoencoder is to learn a representation (encoding) for a set of data, for dimensionality reduction by training the network to ignore signal "noise"

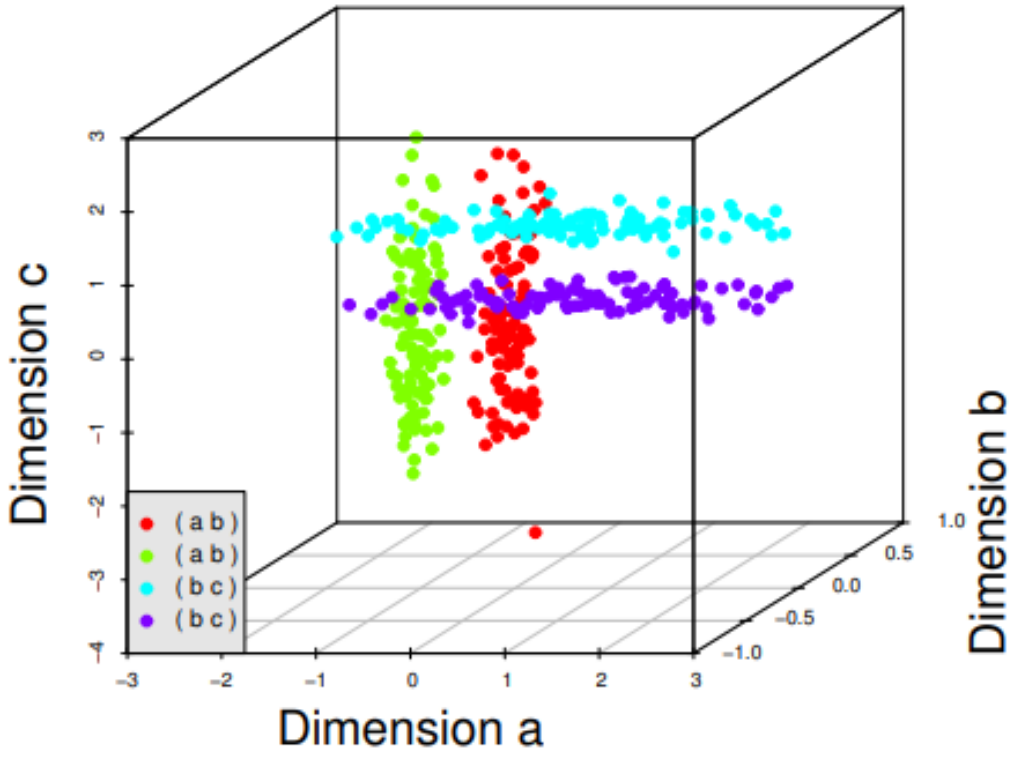
- It efficiently compresses and encode data
- Learns how to reconstruct the data back from the reduced encoded representation
- Final output representation is as close to the original input as possible



Subspace Clustering

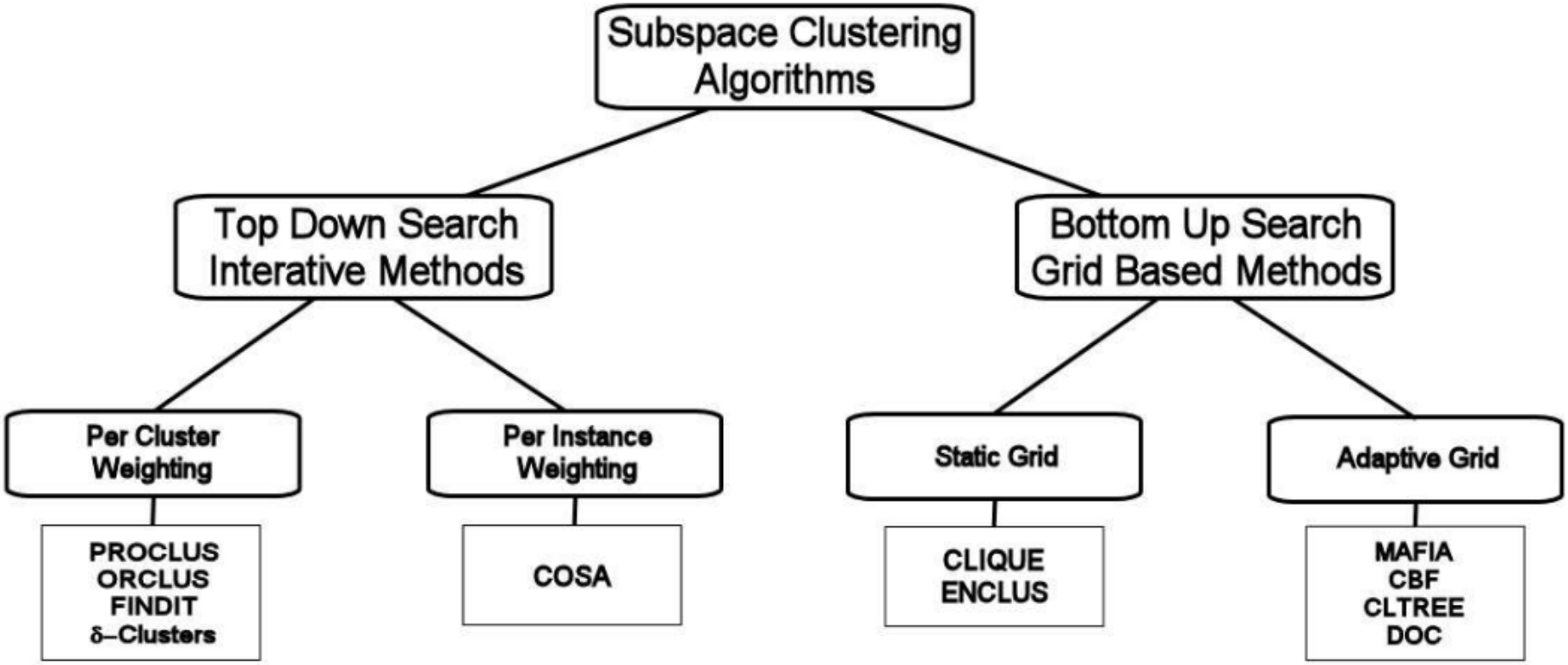
- Subspace clustering is a technique which finds clusters within different subspaces (a selection of one or more dimensions)
- The underlying assumption is that we can find valid clusters which are defined by only a subset of dimensions
- Patient data observing gene expression level can have more than 20000 features
- A cluster of patients suffering from Alzheimer can be found only by looking at the expression data of a subset of 100 genes
- The resulting clusters may be overlapping both in space of features and observations

- A sample dataset with 400 instances in three dimensions
- The dataset is divided into four clusters of 100 instances, each existing in only two of the three dimensions
- First two clusters exist in dimensions a and b
- When k-means is used to cluster this sample data, it does a poor job of finding the clusters.



Sample dataset with four clusters, each in two dimensions with the third dimension being noise

Types of Subspace Clustering



Topological Data Analysis

Topological Data Analysis is an approach to the analysis of datasets using techniques from topology. Extraction of information from datasets that are high-dimensional, incomplete and noisy is generally challenging. TDA provides a general framework to analyse such data in a manner that is insensitive to the particular metric chosen and provides dimensionality reduction and robustness to noise

Reeb Graph

Contour Tree

Visual Mapping

Visual Mapping converts the analysis result from data transformation stage or original dataset into visual structures for rendering in the view transformation stage

- Axis-Based
- Glyphs
- Pixel-Oriented approaches
- Hierarchy-Based approaches

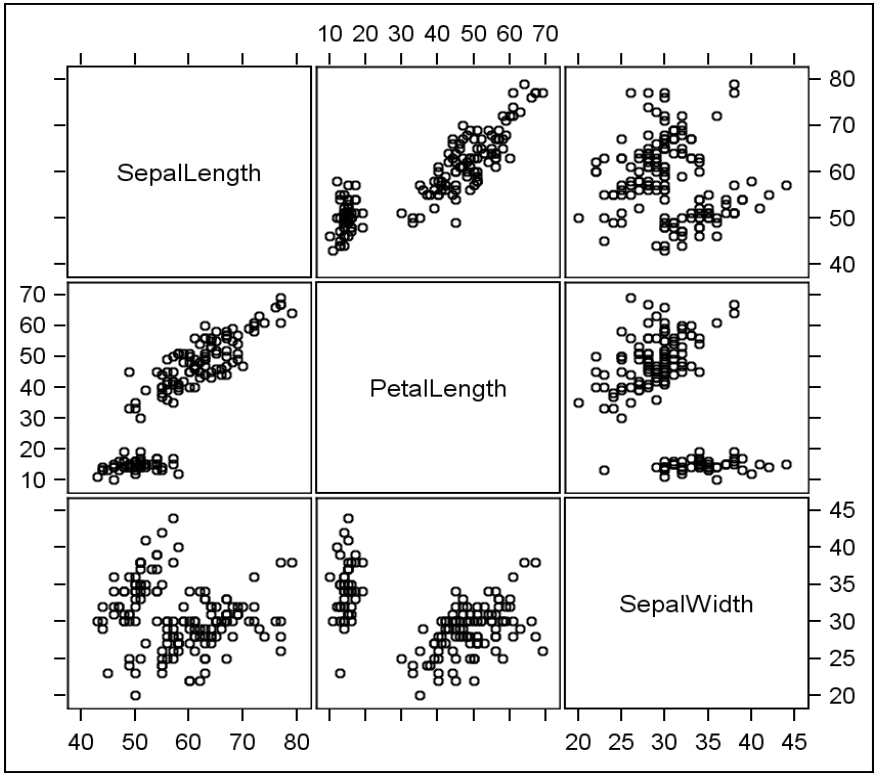
Axis-Based

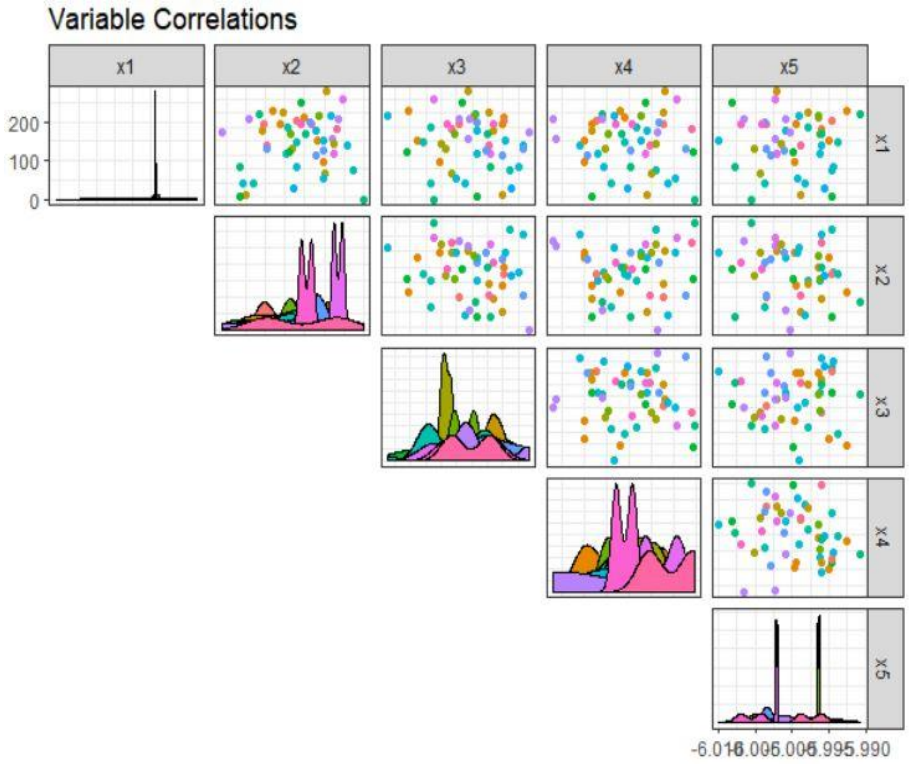
Axis-based methods refer to visual mappings where element relationships are expressed through axes representing the data dimensions

- Scatterplot Matrix
- Parallel Coordinates
- Radial Layout
- Hybrid construction

Scatterplot Matrix

Collection of bivariate scatterplots that allows users to view multiple bivariate relationships simultaneously. Scalability is one of the primary drawback

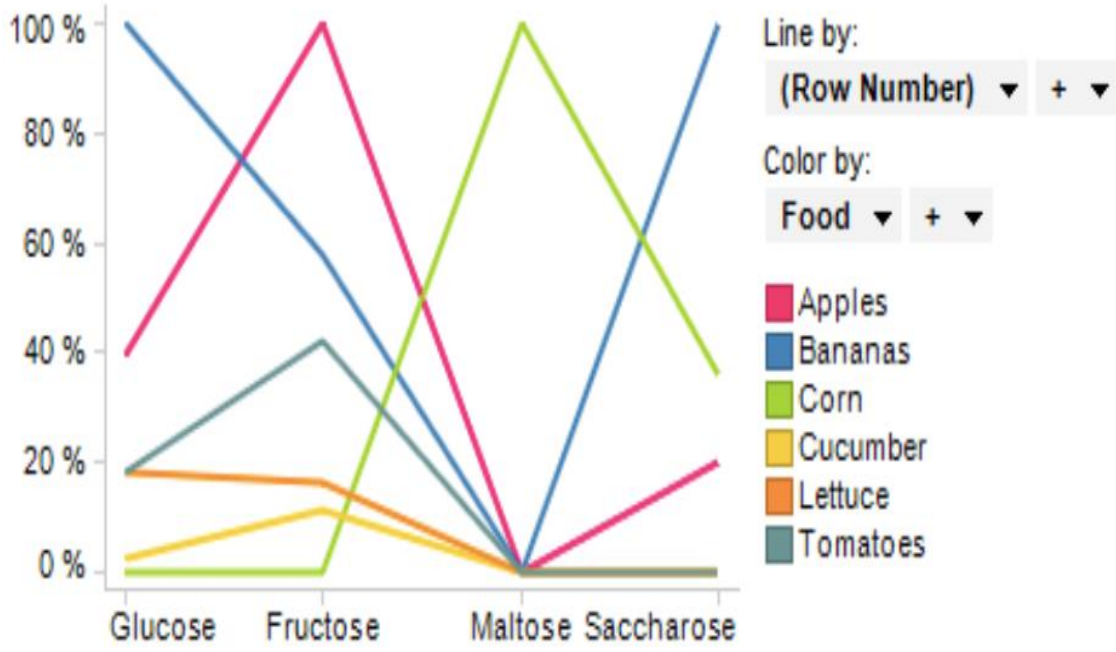




Parallel Coordinates

Maps each row in the data table as a line. Each attribute of a row is represented by a point on the line. In other words each of the dimensions corresponds to a vertical axis

Food	Glucose	Fructose	Maltose	Saccharose
Apples	2.10	4.50	0.00	1.30
Bananas	4.40	2.70	0.00	6.40
Corn	0.60	0.20	0.30	2.30
Cucumber	0.70	0.70	0.00	0.00
Lettuce	1.30	0.90	0.00	0.00
Tomatoes	1.30	2.00	0.00	0.00



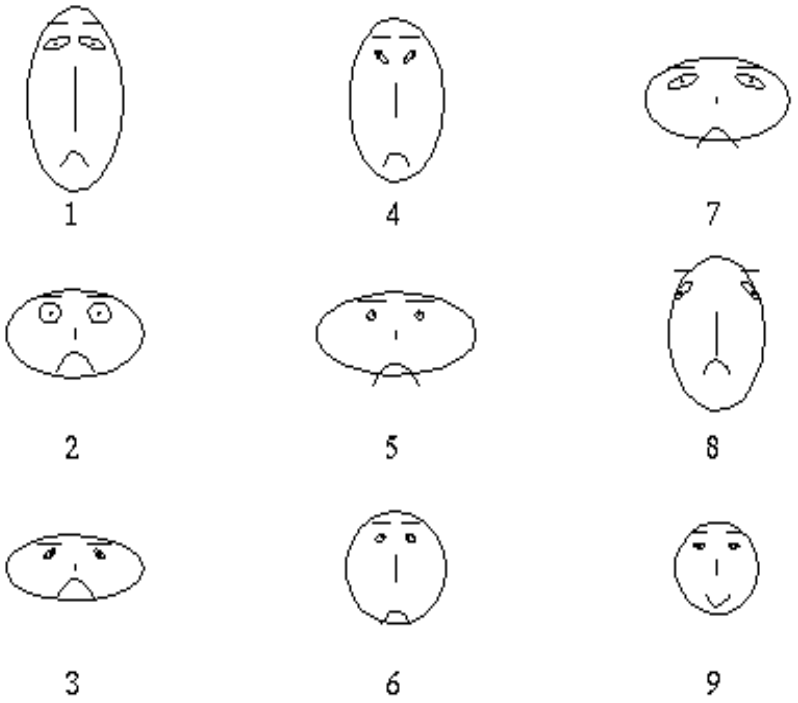
Glyphs

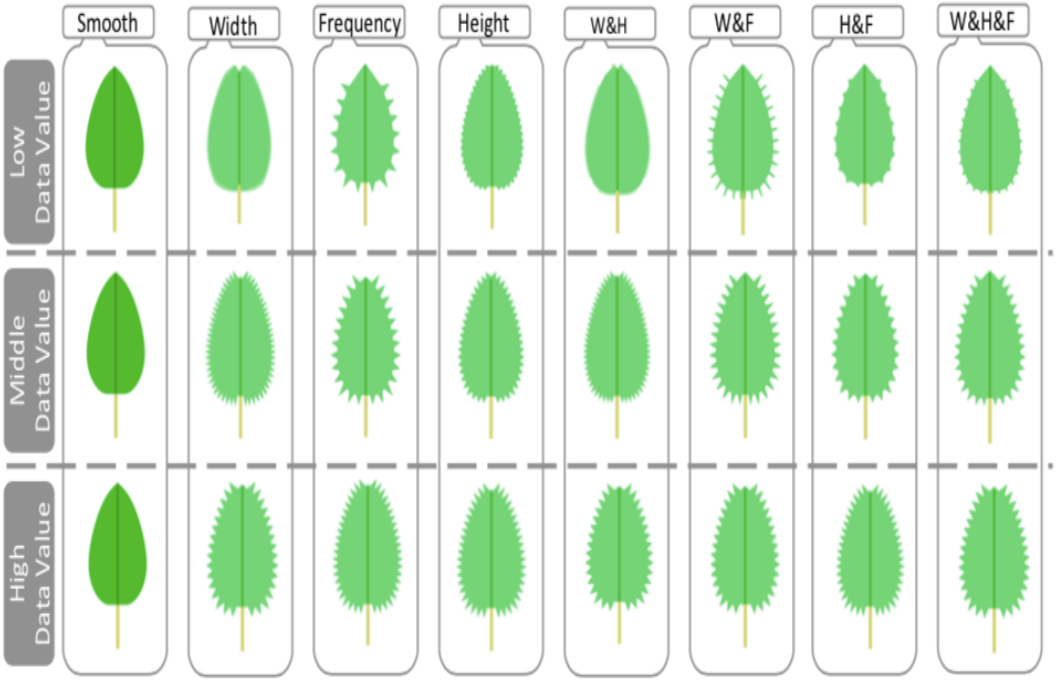
Glyph based visualization is a common form of visual design where a dataset is depicted by a collection of visual objects referred to as glyphs

- Per Element glyphs
- Multi object glyphs

Per Element Glyphs

Each data record is designated for each glyph object and features of the object is represented by the dimension of the data





Pixel Oriented

Encoding data values as individual pixels and creating separate displays or sub-windows for each dimensions

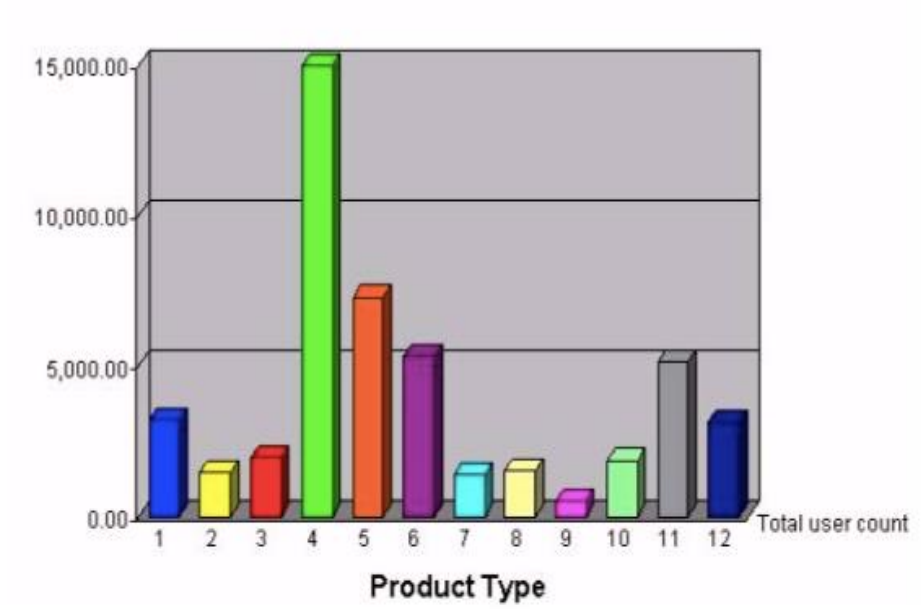
Pixel-Bar charts

Jigsaw Map

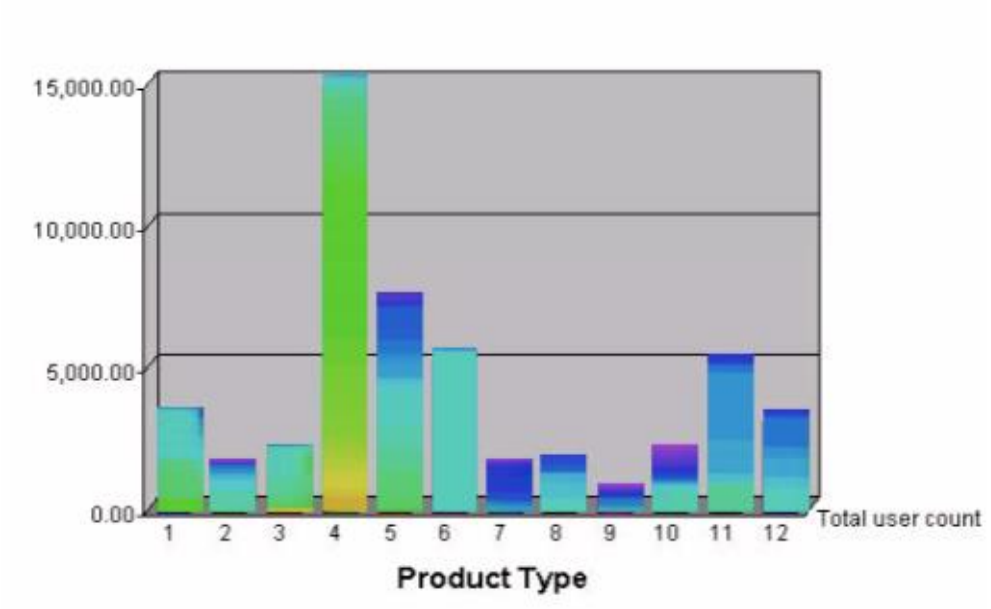
Value and Relation

Pixel Bar Charts

Derived from regular bar charts. It presents the data values directly instead of aggregating them into a few data values. It uses the pixels within the bars to present the detailed information of the data record



a) Equal-Width Bar Chart



a) Equal-Width Pixel Bar Chart

Hierarchy-Based approaches

Hierarchical structures can be used to capture dimensional relationships and to provide summaries for representing high-dimensional datasets

Dimension Hierarchies

Topological Hierarchies

Other Hierarchical structures

View Transformation

View transformation dictate what we ultimately see on the screen.
Methods in these categories focus on screen space and rendering

- Illustrative rendering
- Continuous visual representation
- Image space metrics

Illustrative Rendering

An illustrative visualization is a visualisation technique for measured, simulated, and modelled data that are inspired by traditional illustration. Also depicted as non-photorealistic rendering or stylized rendering, employs abstraction techniques to convey the relevant information and de-emphasize less important details

- Illustrative PCPs
- Illuminated 3D scatterplot
- PCP Transfer Function
- Magic Lens

Continuous Visual Representation

For most high-dimensional visualization techniques, a discrete visual representation is assumed since each element usually corresponds to a single data point. However, due to limitations such as visual clutter and computational cost, many applications prefer a continuous representation

- Continuous Scatterplot
- Continuous parallel coordinates
- Splatterplots

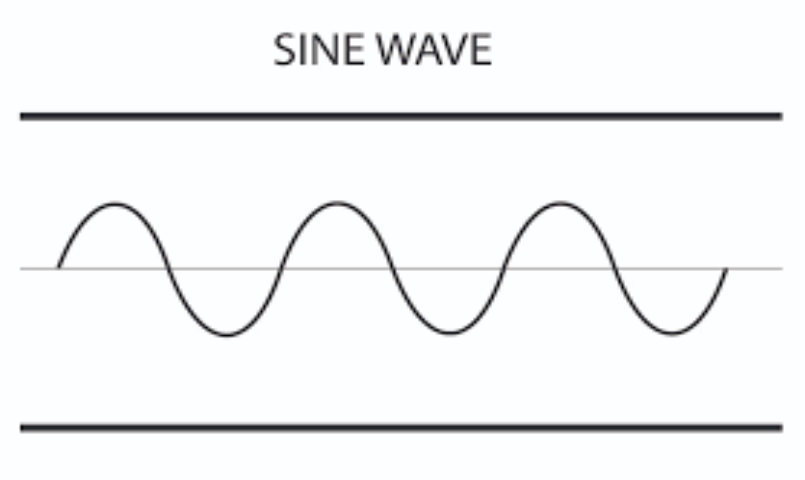
Image Space Metrics

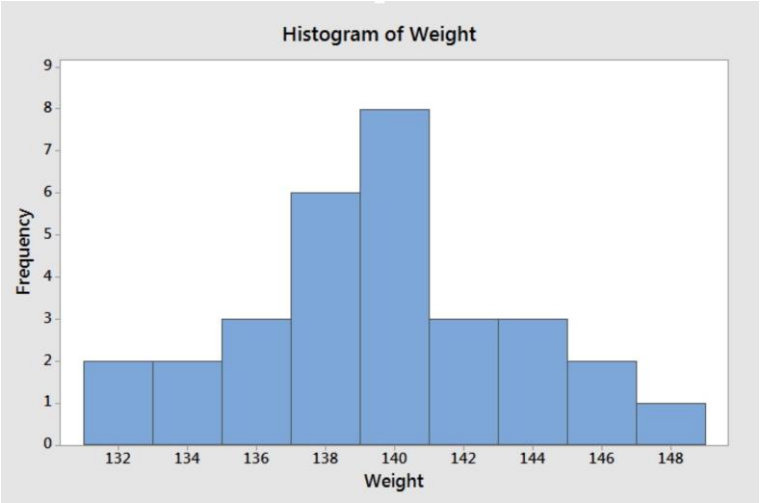
Number of quality measures have been proposed to analyse the visual structure and automatically identify interesting patterns in the techniques, the image space based quality measures is one of the quality measure technique that are applied in the screen space

- Clutter Reduction
- Pargnostics
- Pixnostic

Scalar Visualization

$f : \mathbb{R} \rightarrow \mathbb{R}$





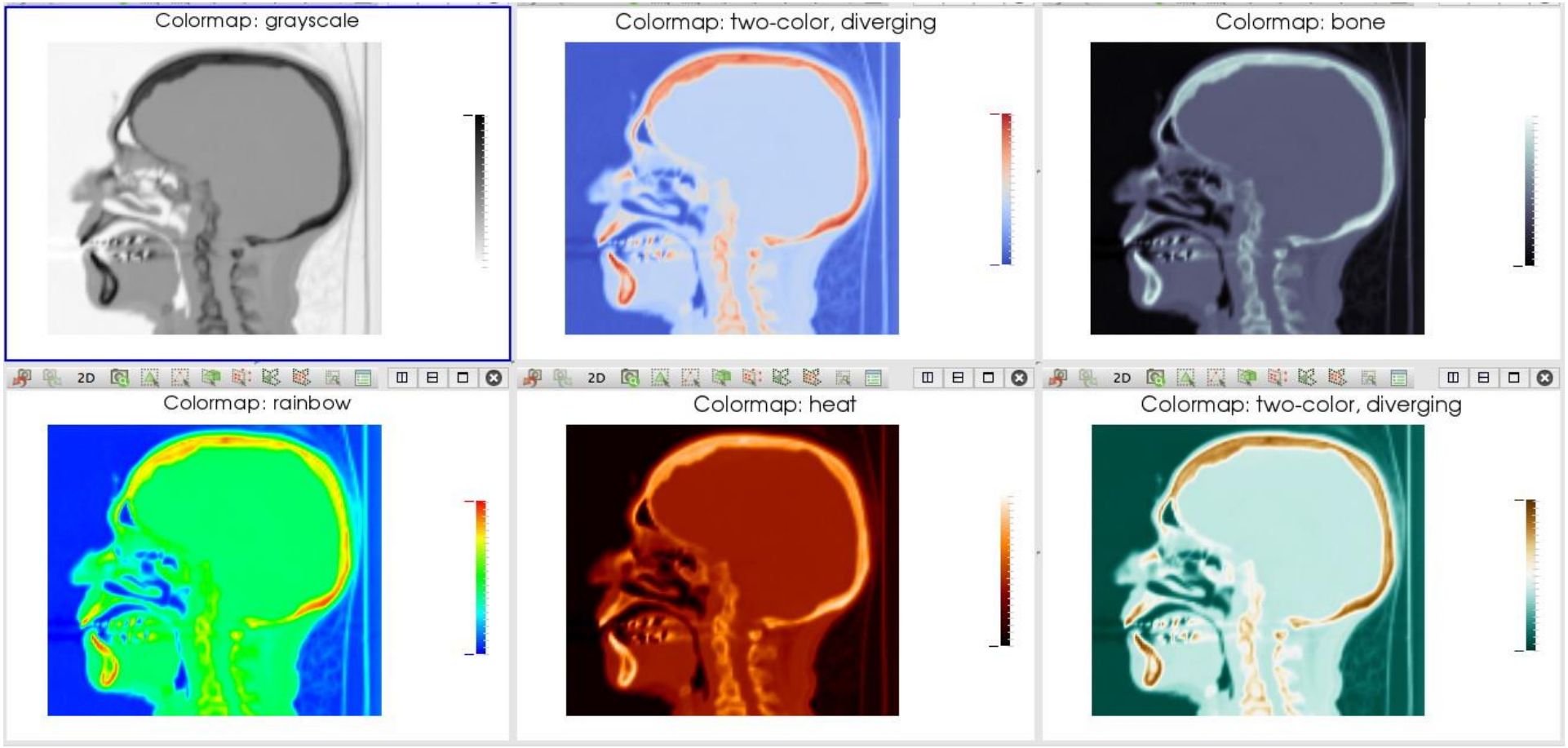
Line graph showing the sine wave function

$f(x) = \sin x$

Scalar Visualization-Two Dimensional

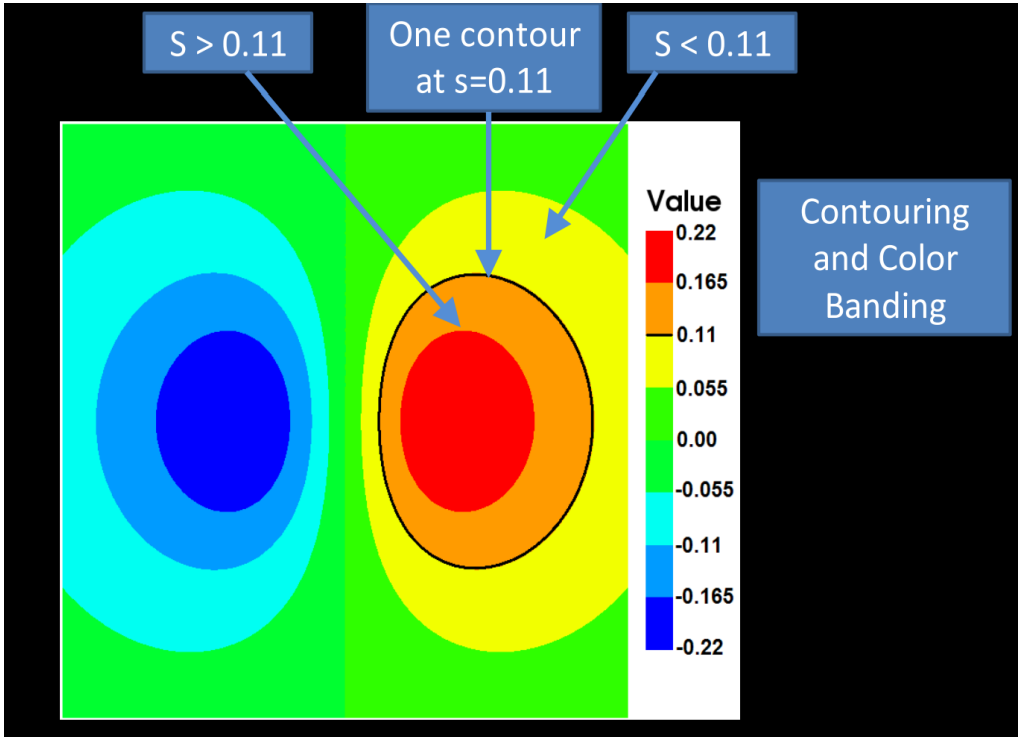
$f : \mathbb{R}^2 \rightarrow \mathbb{R}$

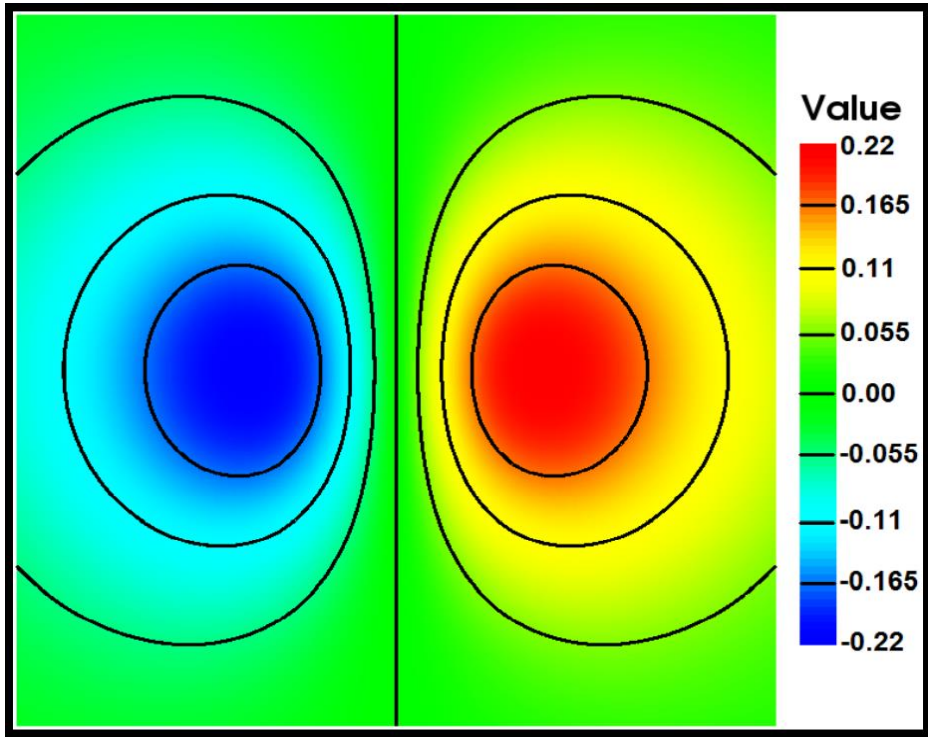
Colour Mapping : Common scalar visualization technique that maps scalar data to colours



Scalar Visualization-Two Dimensional

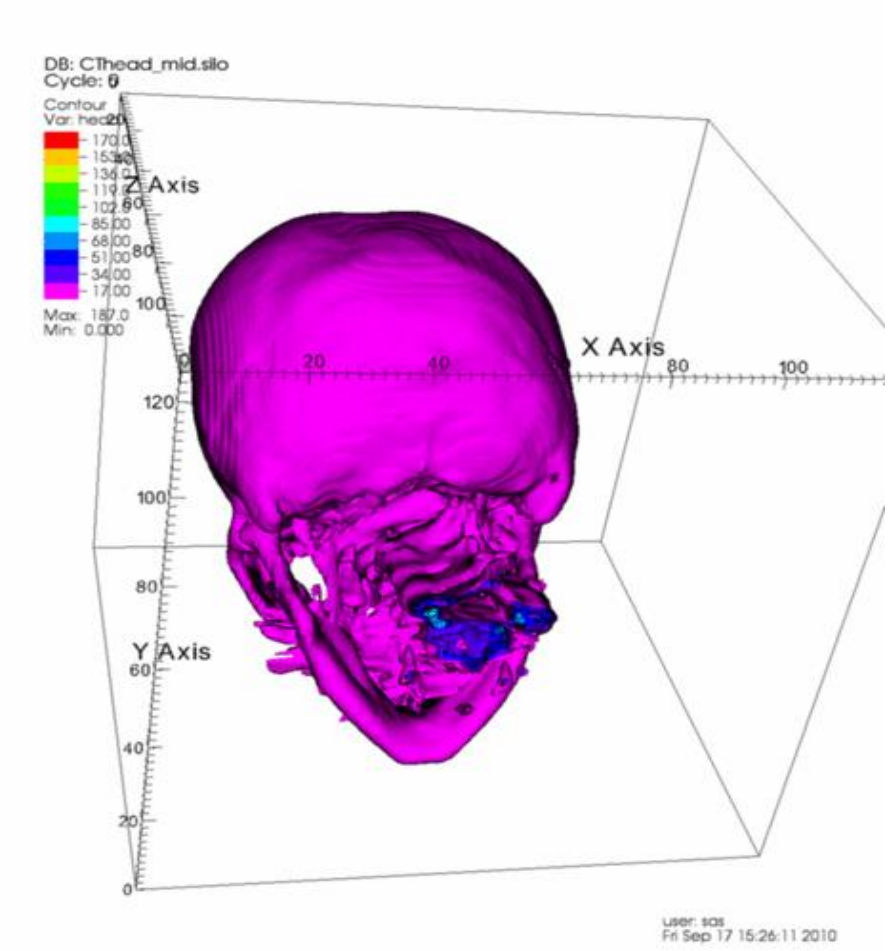
Contouring : A contour line C is defined as all points p in a dataset D that have the same scalar value, or isovalue s(p)=x, $C(x) = \{p \in D | s(p)]x\}$

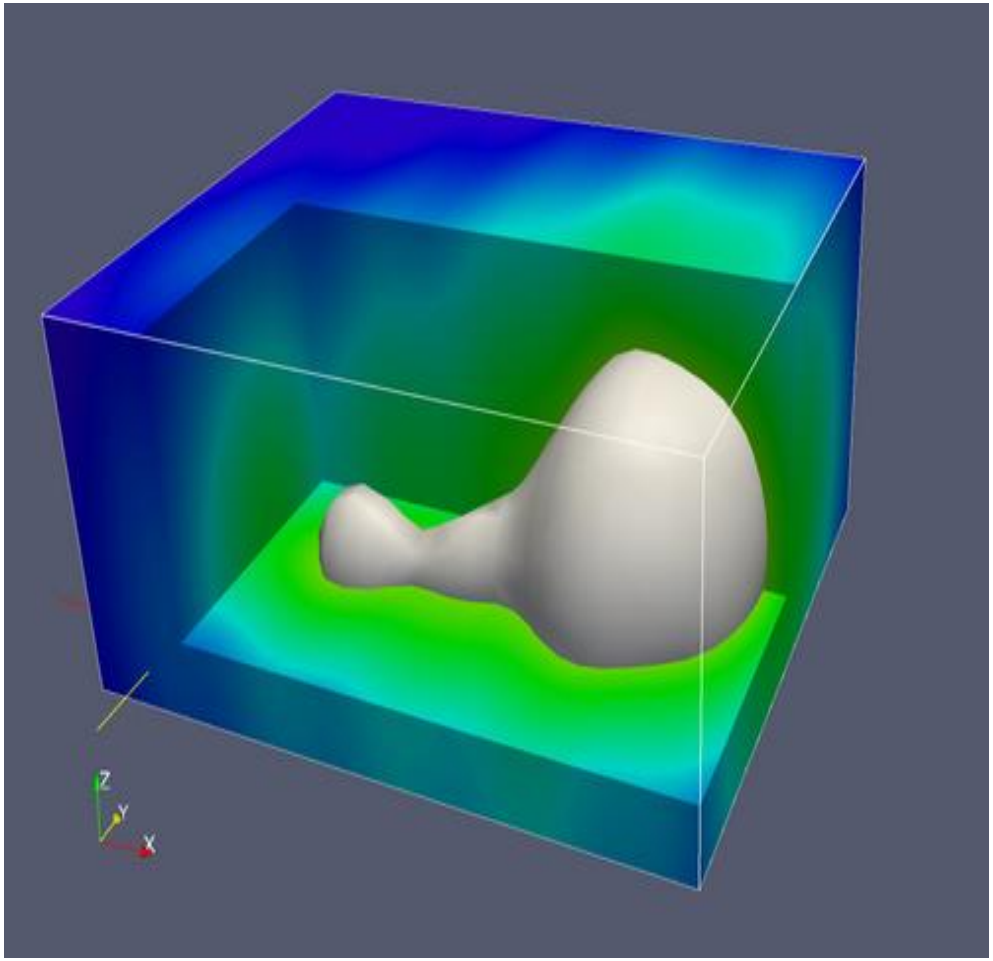




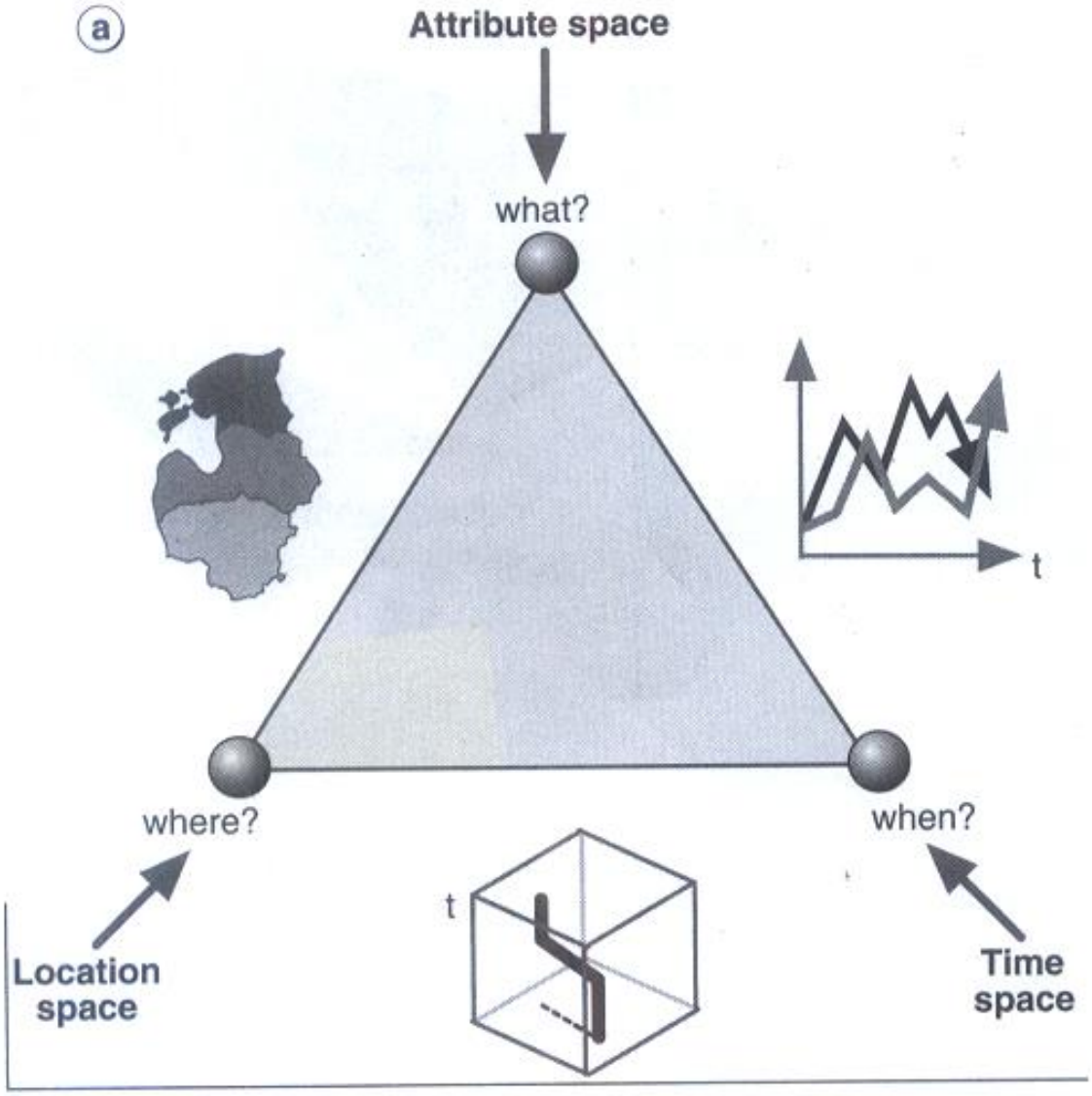
Scalar Visualization-Three Dimensional
$f : \mathbb{R}^3 \rightarrow \mathbb{R}$

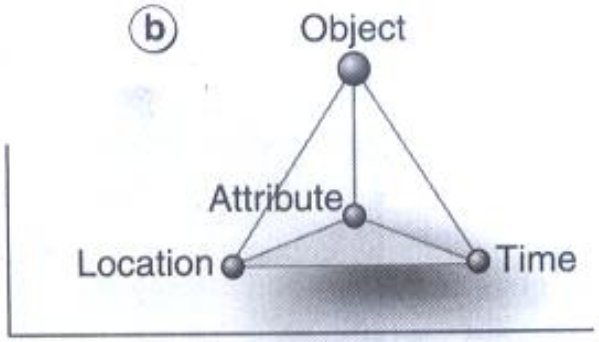
Contouring : For 3D datasets, a contour is a 2D surface, called isosurface

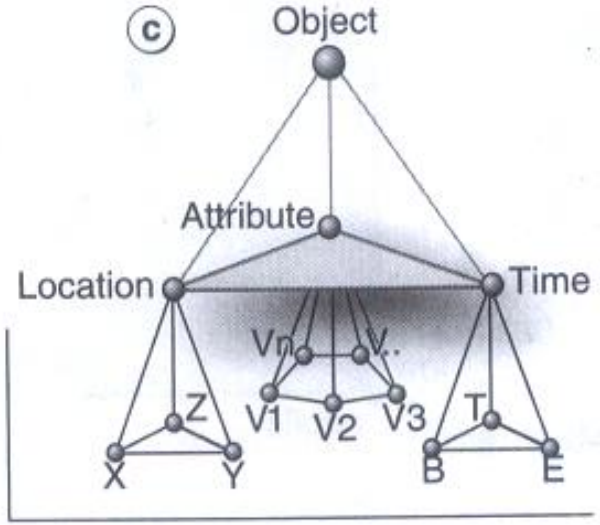




Geo-Visualization







Geo-Visualization - Process

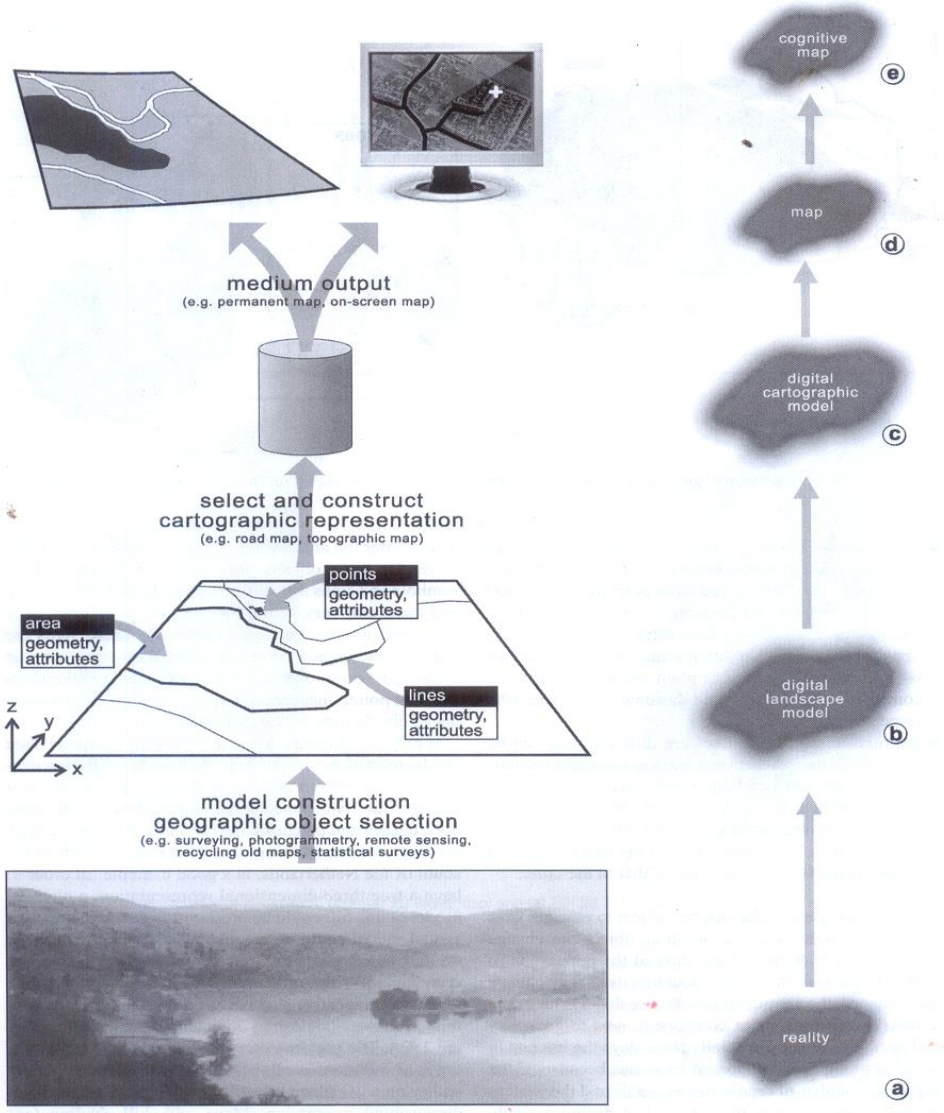
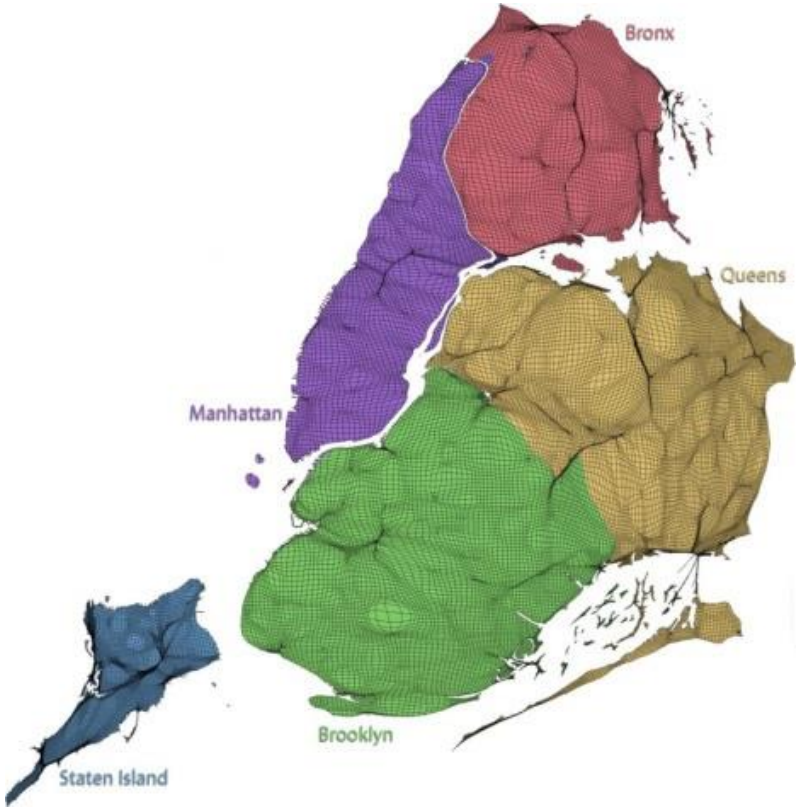
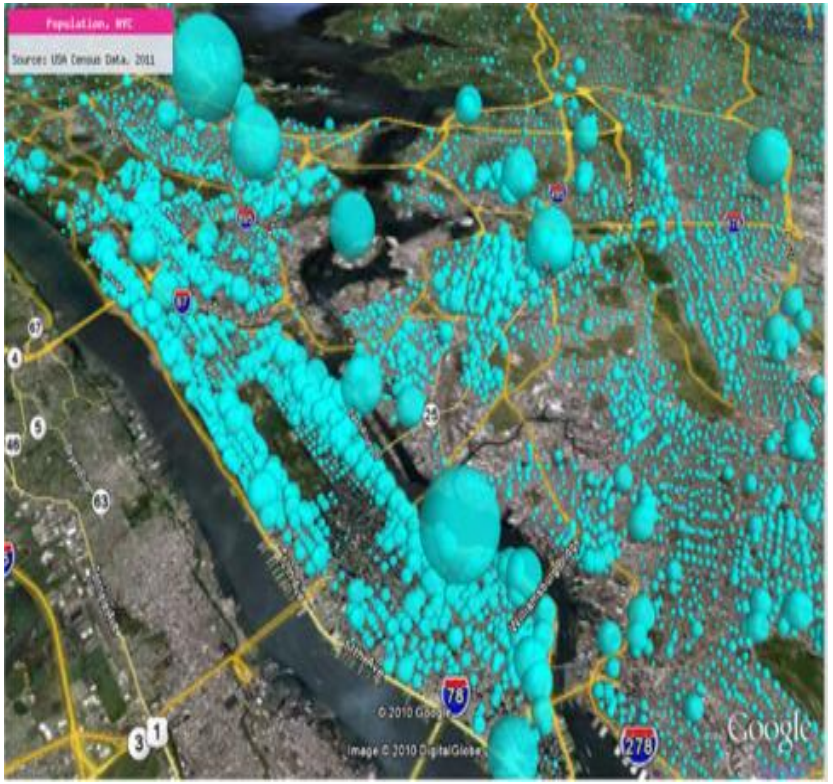


Figure 1.4 The nature of geospatial data: from reality (a), via model construction and selection to a digital landscape model (b), followed by selection and construction of a cartographic representation towards a digital cartographic model (c), presented as a map (d), which results in the user's cognitive map (e)


Bronx
Queens
Manhattan
Brooklyn
Staten Island


Population, NYC
Sources: USA Census Data, 2011
Google

Take Away Points

- Increase in dimension and complexity of the data makes it harder to analyze feature and relationships in the data
- Information visualization is a principal method to understand and observe those datasets
- Two types of data – Quantitative and Qualitative
- One-, Two- and Three-dimensional data can be visualized with various techniques available
- Data with high dimensions must go through a process before it can be visualized