# Data types:

* Nominal Data: Non Ordered Categorical variables
  + Only deals in Binary operators
* Ordinal Data: Categorical variables which have an order
  + Binary operators and can also use > or <
* Quantitative (Ratio): Defined zero point.
  + Includes variables like mass, height, age
  + Operators: !=, = ,>,<, -, %
* Quantitative (Interval): Lacks absolute zero point
  + Operators: !=, = ,>,<, -, % etc
  + Includes variables like Year, temperature

# Definitions

* Data-ink ratio = (data – ink)/total ink
* Expressive: Encodes all and only facts in data
* Effective: If information is perceived more readily/easily
* Principle of Importance Ordering: Encode more important info more effectively
* Hypothesis: Educated guess about some aspect of world
* Map: Scaled down visualisation of the physical world
* Discriminability: Choose channel which can accommodate number of distinct categories wish to encode
* Separability: Channels ability to interfere with each other
* Redundancy: Avoid encoding redundant information or information already encoded
* Stimuli: Features of environment which senses are capable of detecting
* Sensation: Physical response of a sense organ to stimuli
* Perception: Psychological process of actively selecting and organising stimulus info detected by sensory organs so to create awareness
* Visual Attention: Various mechanisms which help determine which regions of an image are selected for detailed analysis
* Fixations: Detailed info from a small region is visible
* Saccades: Brief period which eyes flick to a new location
* Change Blindness: Major changes to a visual representation going unnoticed
* Inattentional Blindness: When we fail to notice something in our field of vision as we are occupied with something else
* Pre-attentive properties: Set of visual features which are detected by low level, fast-acting processes within period of single fixation
* Colour blindness: Decreased ability to perceive colour differences under normal lighting conditions
  + Types include Protanopia, Deuteranopia and Tritanopia

# Visualisation Methods:

* Table:
  + Column= Attribute
  + Row= Item/observation
  + Can use aggregation (min, max, count, average, mean)
  + Can use data values from other columns to create a new one
  + Can group data
* Bar Chart:
  + Each bar is different item
  + Uniform width as data isn’t encoded in width
  + Height represents data value
  + One quantitative data attribute for bar length
  + Horizontal ones can be used if length of text is long
  + Can have multiple attributes for y axis (Can either repeat graph for different regions or use colours)
  + Stacked bar charts are an option
  + Data is encoded in x, y, colour
* Dot Plots:
  + X: Values
  + Y: Categories
  + Dots are observation
  + Height/length is dependent on graph
  + Mark used is dot and data can be encoded on x ,y , colour and size
* Line Chart:
  + Show trends/cycles over given time
  + Y is usually value and x is time/must be continuous
  + Show points for emphasis
  + Can show multiple lines for comparisons
  + Can use slope graphs when only two times of interest exist
  + Mark is line and data is encoded in x, y, colour
* Scatter graphs:
  + Point represents one observation
  + X and Y both quantitative
  + Used to show correlation between two quantitative variable
  + Can identify groups and outliers
  + Can plot every pair in a scatter plot matrix to compare all variables
  + Mark used is point and data is encoded in x , y , colour, shape
  + Can show multivariate data
* Pie Chart:
  + Shows percentages of data
  + Data is encoded in colour and size
* Histograms:
  + In depth view of numeric variable
  + Construct guide:
    - Divide data into bins
    - Count occurrence of each bin
    - Normalize counts
    - Pot on a bar graph of normalized counts
  + Width=range, height = frequency/probability
  + Shapes can reveal skewness/ modes
  + Data is encoded in x, y, length, width
* Dot maps
  + x encodes longitude while y encodes latitude
* Line Maps
  + mark line
  + usually have encodings for size and colour
* Choropleth maps (area mark):
  + Used to show regional variation.
  + Issue where data isn’t necessarily clear
* Cartogram (Area mark):
  + map based visualisations which distort geography of map to display data attributes
* Tile grid maps:
  + Constant area to represent different regions.
  + Colour represents data attribute
* We can select features to represent multivariate data.
* Heatmap
  + Used to show multivariate data
  + Each row = Row in data. Each column = Attribute in data
  + Each cell is a single value
  + Usually normalize values by point-min(x)/ Range of data
  + Data attributes must share common features or scales
* Radar charts
  + Display for multivariate data
  + Distance from centre indicates quantitative attribute.
  + Bad for nominal/Categorical data display
* Parallel Coordinates
  + Helps represent multivariate data
  + Each line in this plot represents row in dataset
  + Each access scaled to min and max of each data attribute
  + Messy
  + Patterns are order dependant
  + Better if interactive
* Chernoff Faces uses
  + a humanlike face for a glyph
  + Face is a whole unit or row of data
  + Parts of face are adjusted based on data they represent
  + Wide range of disadvantages such as subjectivity and complexity. They are also inefficient and prone to being misinterpreted. Lack of precision
* Mercator Projection:
  + Pros: Good for navigation, preserves shapes and directions
  + Cons: Its use of distortion exaggerates sizes and distances
* Gall-Peters Projection
  + Pros: Preserves areas
  + Cons: Doesn’t preserve shape nor distances
* PCA (Principle Component Analysis:
  + Center data to mean
  + Build covariance matrix
  + Find eigenvalues and eigen vectors
  + Plot data along with eigenvectors
* Possible Encoding Channels:
  + Position (horizontal, vertical or both)
  + Colour
  + Shape
  + Tilt
  + Size like length area and volumes
* GeoJSON: open standard format for representing simple geographical features using JSON
* TopoJSON is an extension of GeoJSON which exploits redundancy in geometry
* Useful encodings for spatial dimensions:
  + Size
  + Shape
  + Luminance
  + Hue

# Good visualization practices:

* + Left align text columns
  + Use consistent rounding
  + Right alight numerical columns
  + Show data values if precision is needed
  + If visualization doesn’t add anything, no need to add
  + Avoid angled/ vertical text
  + Remove background and chart background colour
  + No legend depending on if data is labelled properly
  + Monochrome depending on graph type (most time keep monochrome), use colour deliberately. Limit colours as hard to tell apart. Avoid rainbow and use other scales. Turn everything else grey if wishing to highlight something
  + No need for grid lines
  + Sometimes one can remove vertical axis but add data values if doing so
  + Chart border is pointless
  + Don’t use 3D graphs, multiple graphs are better
  + For pie graphs, adding labels helps a lot especially if slices are similar in size
  + Bar chart axes must be at zero
  + Don’t obscure data with support
  + Consistent fonts

# Gestalt

* Gestalt principles: Principles of how we organize disparate elements of sensory experience into a sensible whole
* Confounding Gestalt principles: Principles which lessens the effect of any individual Gestalt principle
* List of Gestalt principles:
  + Proximity Principle: elements tend to be perceived as aggregated into  
    groups if they are near each other
  + Similarity Principle: elements tend to be integrated into groups if they  
    are like each other
  + Connectedness Principle: elements tend to be grouped together if they  
    are connected by other elements
  + Enclosure principle: elements tend to be grouped together if they  
    are enclosed by other elements

# Colour Models

* CIE Chromaticity Diagram:
  + - X Coordinate is a combination of red and green wavelengths while the Y coordinates is a combination of green and blue wavelengths.
    - Colour gamut is a subset of colours that can be represented by mixing colours at its corners
    - Spectral locus: The curved edge. Each point represents a pure hue/unsaturated color of specific wavelength
    - Chromaticity = Hue + Saturation
    - Entire colour range cant be displayed based on any 3 colours
    - 2D
    - Only perceivable colours
* RGB (Red Green Blue)
  + Based on Cartesian coordinate system
  + X = Red, Y = Green, Z = Blue
  + Corners are RGB, CYM, BW values with black being at origin and white being (1,1,1)
* HSL (Hue Saturation Lightness)
  + 3 Measures
    - Hue: Colour attribute which describes pure colour
    - Saturation: Gives a measure of how much a pure colour is diluted with white light
    - Lightness
* The Munsell system
  + Perceptually uniform version of HSL space
  + 5 Principal Hues: Red, Yellow, Green, Blue and Purple with 5 intermediate Hues between principle hues
  + Value: Measures perceptual lightness (0 being black and white being 10)
  + Chroma: Purity of colour

# Visualisation Experiment

* Creating a visualization experiment:
  + Proposed hypothesis: Make a guess about aspect of world. This is the what we aim to find. Must be specific and testable
  + Measured variables: What we measure.
    - Independent variables: Variables which are manipulated to create different experiment conditions
    - Dependant: Variables which are measured to see how they are affected by independent variables
    - Confounding variables: Variables which interfere with measurement
  + Experimental methods/conditions: Levels and values of independent variables used in an experiment
    - Between groups: Subject assigned one experimental condition
      * Avoids interference effects but needs large number of subjects
    - Within Groups: subject performs under all different condition
      * Less people needed and there is variability across all conditions but order of conditions is important to avoid learning effects
      * Latin square: Way to order conditions in within group experiments to control learning effects. In a balanced one, each condition precedes and follows each other condition a equal number of times
  + Selected Subjects: Should represent population hypothesis refers to
    - Larger than 10 always
  + Data Collection
    - Gathered either manually or automatically
    - Measurements either objective or subjective
  + Data analysis: Use of statistical techniques
    - Between two results:
      * Unpaired t-test: Unpaired Groups
      * Paired test: Paired groups
      * Mann-Whitney. No Normality
    - More than two results:
      * ANOVA
      * Bonferroni test
      * Kruskal Wallis
  + Experiment management
    - Pilot study: Smaller version of experiment to test design
    - All participants have consistent experience
* Evaluation: Gathering data about usability of a design or product by specific users for a particular activity in specified environment
  + Used to:
    - Compare effectiveness and usability of approaches
    - Explore aspects of visual perception
    - Receive feedback to better understand users to refine visualisations
  + 2 types:
    - Controlled Experiment: Performed in a fully controllable environment allowing us to be specific about what we test
    - Field Study: Performed in a natural setting allowing us to see how people interact with visualisations

# Network Data

* Network Data: Row is relationship between people/items
* 3 types of networks:
  + Directed
  + Undirected
  + Weighted
* Tasks that are done on networks:
  + Finding paths
  + Identifying clusters/groups/communities
* We can use the following to visualize network data:
  + Node-Link Diagrams
    - Node colours, size and edge width can be encoded
  + Adjacency Matrix
    - Marks tend to be point with two nominal data on x and y
    - Edges can have weights/attributes
* In most network diagrams:
  + Minimize line crossings
  + Maintain pleasing aspect ratio
  + Minimize total area
  + Minimize total length of edges
  + Minimise the number of bends in edges
  + Minimise number of distinct angles used
  + Strive for symmetry
* Spring force model
  + We replace edges with springs (attractive forces)
  + Replace vertices with charged particles to repel each other
* Radial tree Algorithm:
  + Place nodes on concentric circles by level
  + Partition circle into sectors where the width of each sector reflects number of leaves
  + Draw subtrees into their sectors
  + Make sure order is preserved even if planarity is not guaranteed
* Treemaps can be used
  + They use enclosure to indicate hierarchy with area indicating some data attribute