SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Data visualization

PROF. D. HERREMANS

50.038 Computational Data Science

Guest speaker

Dr Brian Ang is currently an Assistant Director with the Digital Services Lab of Infocomm Media Development Authority (IMDA). He leads a team of AI scientists, engineers and project managers in the area of speech-related technologies. Prior to joining IMDA, he was a Senior Member of Technical Staff with the Cognition and Fusion Lab of DSO National Laboratories. In DSO, he was the Principal Investigator for projects relating to computational cognitive approaches for intent inference, prediction and activity recognition. He has published several papers in the area of machine learning, deep learning and computational cognitive models. He obtained his B.Eng (Electrical Engineering) and Ph.D (Computational Intelligence Techniques for Data Analysis), both from the National University of Singapore.

Guest speaker

Sharing on the National Speech Corpus and Natural &Transcription Technologies

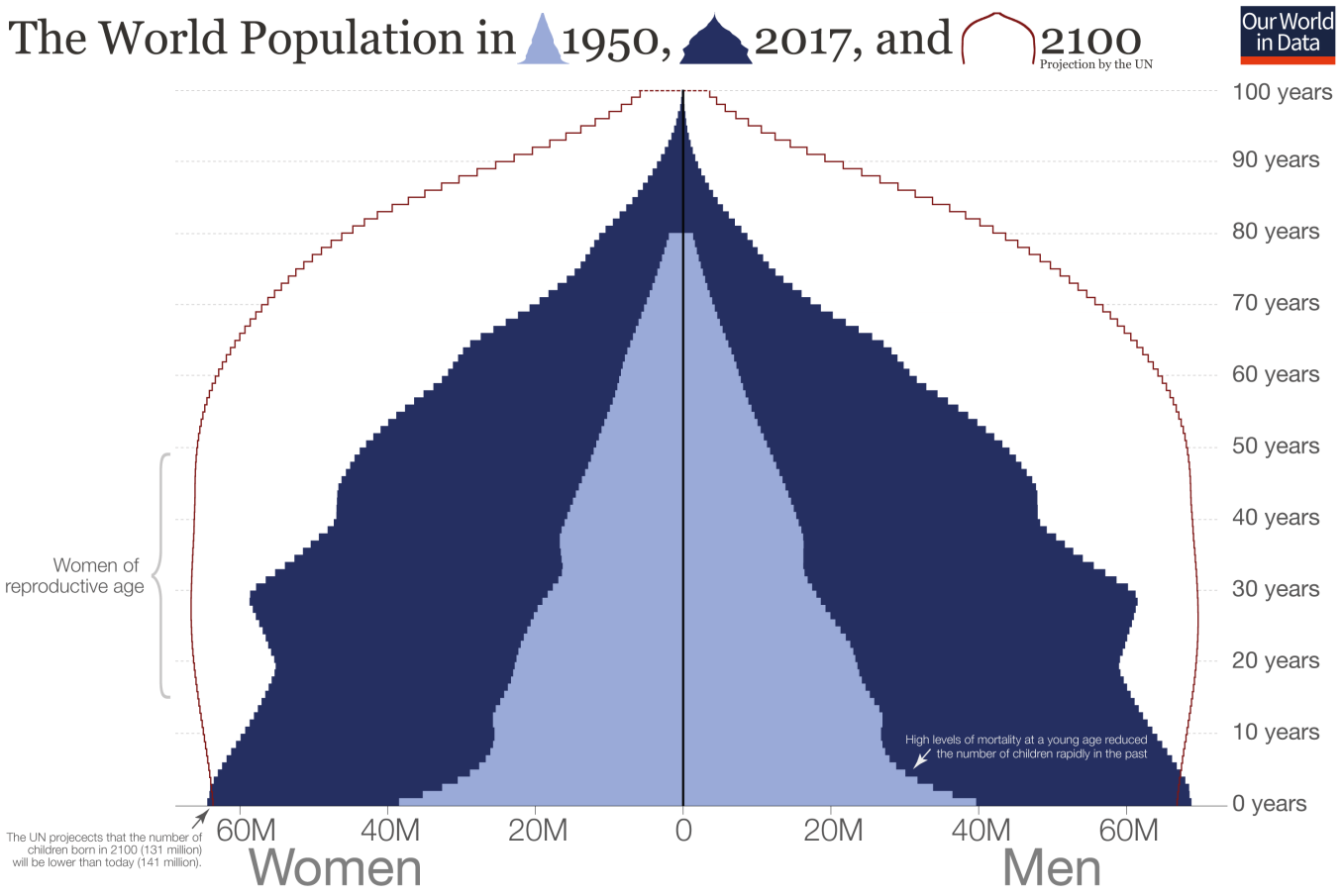
In this session, the speaker will introduce Infocomm Media Development Authority (IMDA) Digital Services Lab (DSL) and share a few speech-related projects that DSL is currently working on. The projects include the National Speech Corpus (NSC) and the Natural Speech & Transcription Technologies (NSTT). Examples of speech-related applications will also be presented. The NSC is a corpus of audio and transcription files, and can be used to develop an Automatic Speech Recognition (ASR) engine for local context. The NSTT consists of modules such as a Speech Activity Detection Engine to separate speech and non-speech portions of an audio file, ASR Engine for speech to text transcription, and Speech Synthesis Markup Language Module for text to speech synthesis.

SEMMA Methodology (SAS)

- **S**ample from datasets
- **E**xplore datasets, e.g. visualization
- **M**odify data, e.g. create/transform features
- **M**odel -> use algorithms to fit model
- **A**ssess: compare models, test datasets, evaluate reliability/usefulness

A picture says more…

This goes for graphs as well…



Data source: United Nations – World Population Prospects 2015. Data in 1-year-brackets is only available up to the age of 100 years in 2017 and 2100 and only up to 80 years in 1950.
The interactive data visualization is available at OurWorldinData.org. There you find the raw data and more visualizations on this topic. Licensed under CC-BY-SA by the author Max Roser.

Why visualize?

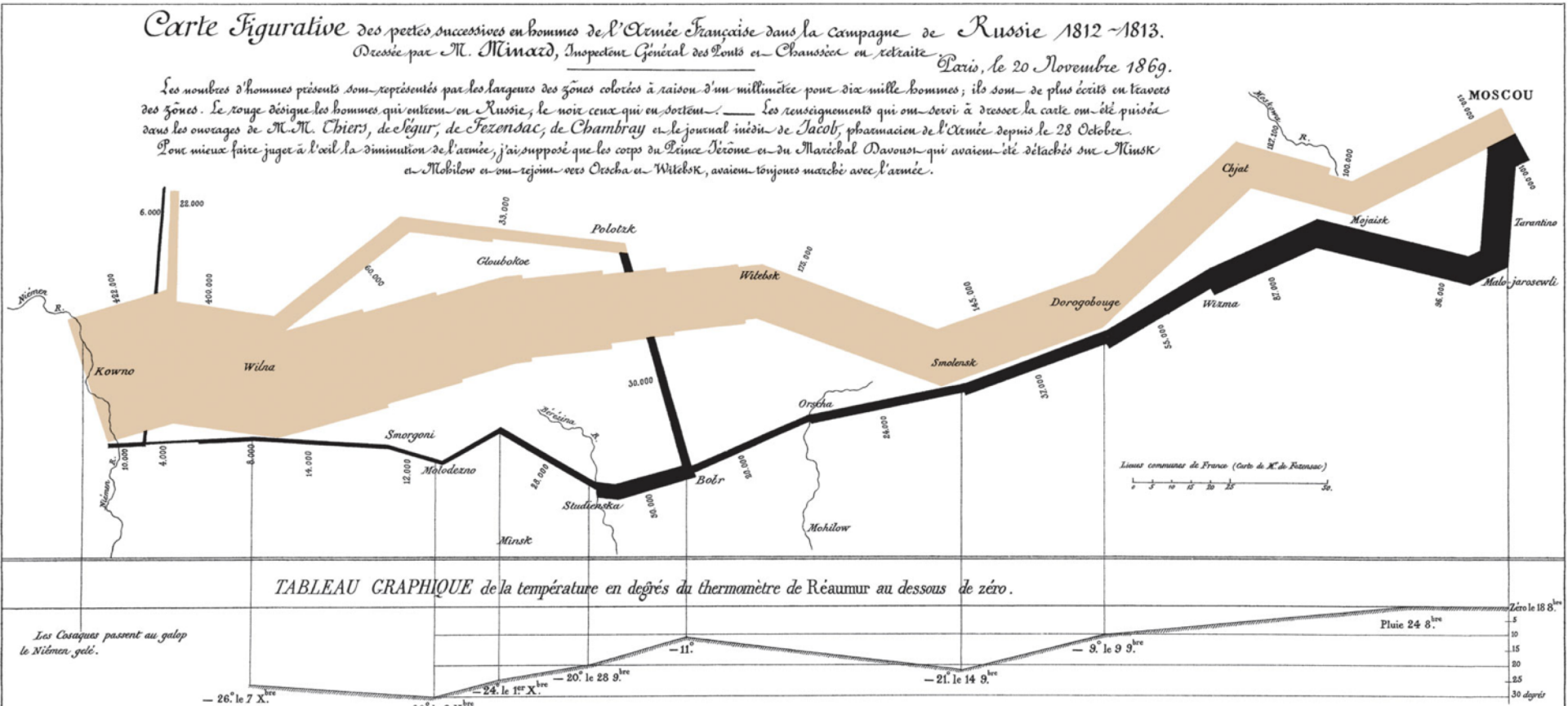
- 2 Objectives:
  - Data analysis, to:
    - understand the data
    - derive information from them

    (involves comprehensiveness)
  - Communication:
    - of information

    (involves simplification)

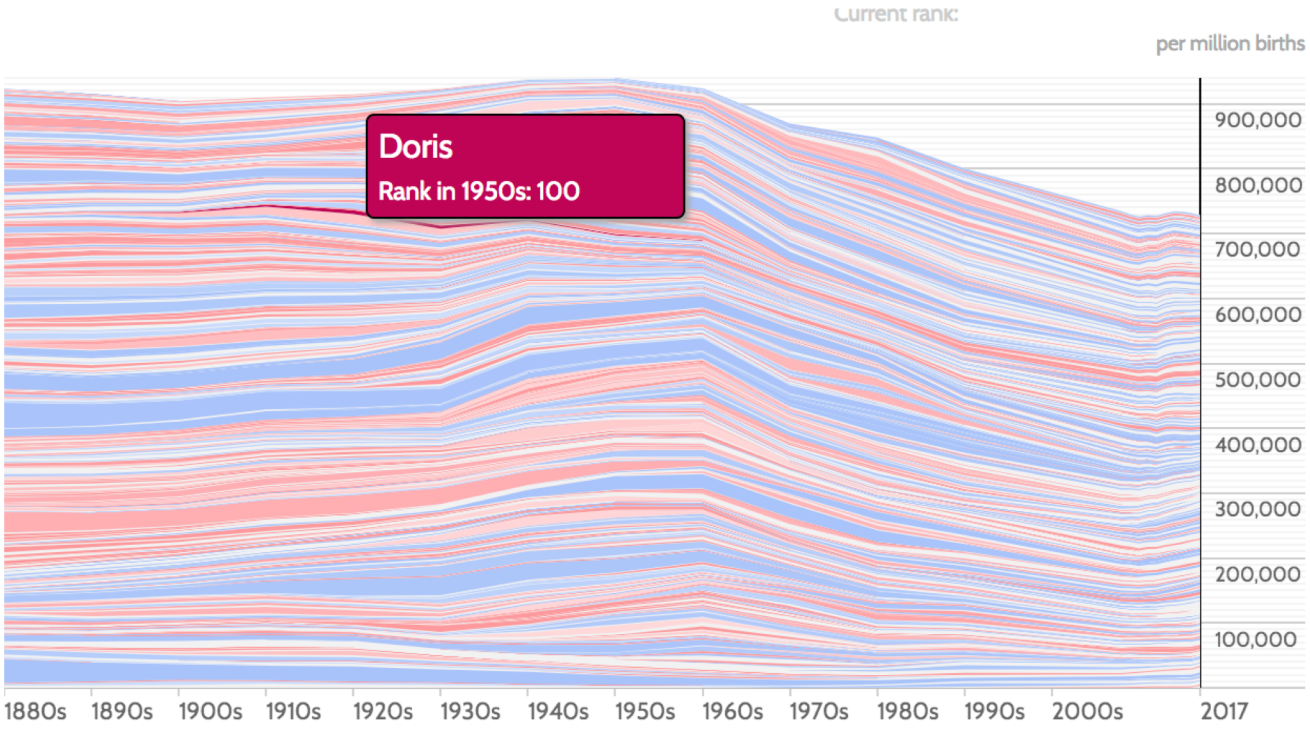
A classic: Charles Joseph Minard 1869 Napoleon's March

- According to Tufte: "It may well be the best statistical graphic ever drawn. 5 variables: Army Size, location, dates, direction, temperature during retreat



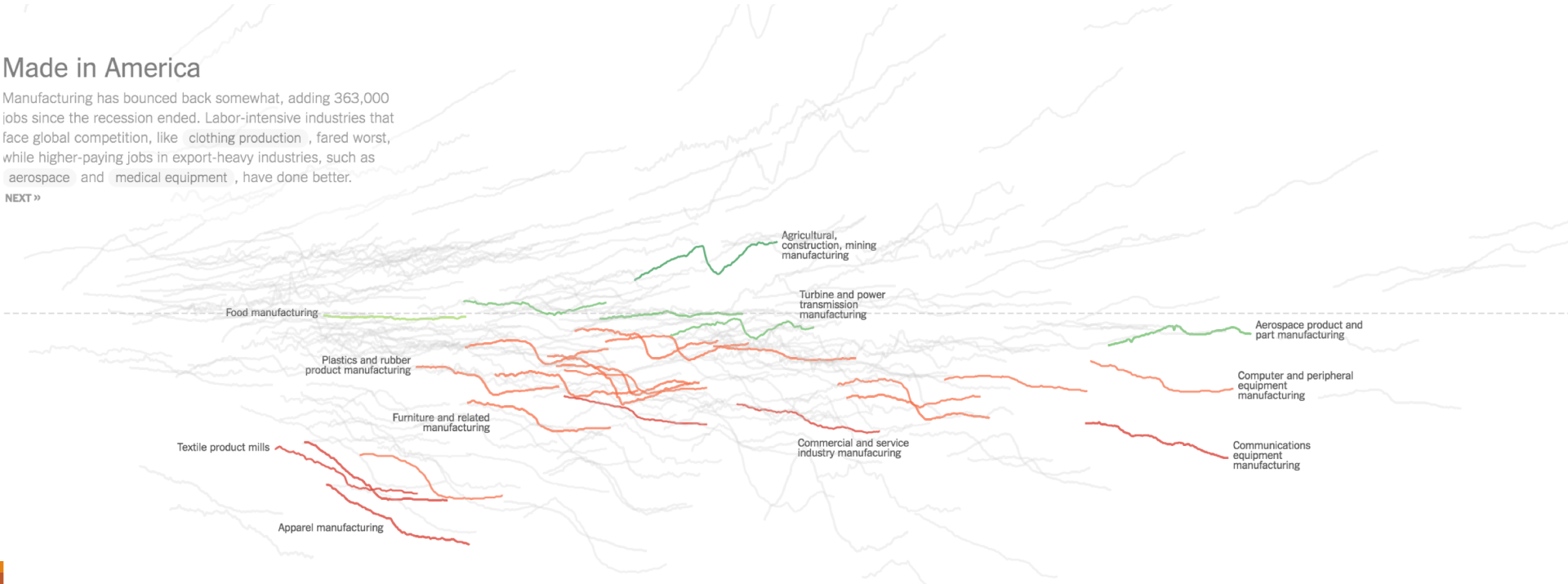
Interactivity: Baby Names Voyager

- http://www.babynamewizard.com/voyager#prefix=&sw=both&exact=false. (Wattenberg et al. 2005)



How the Recession Reshaped the Economy, in 255 Charts

- NY Times Interactive Visualizations (recession/recovery 2014) http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html



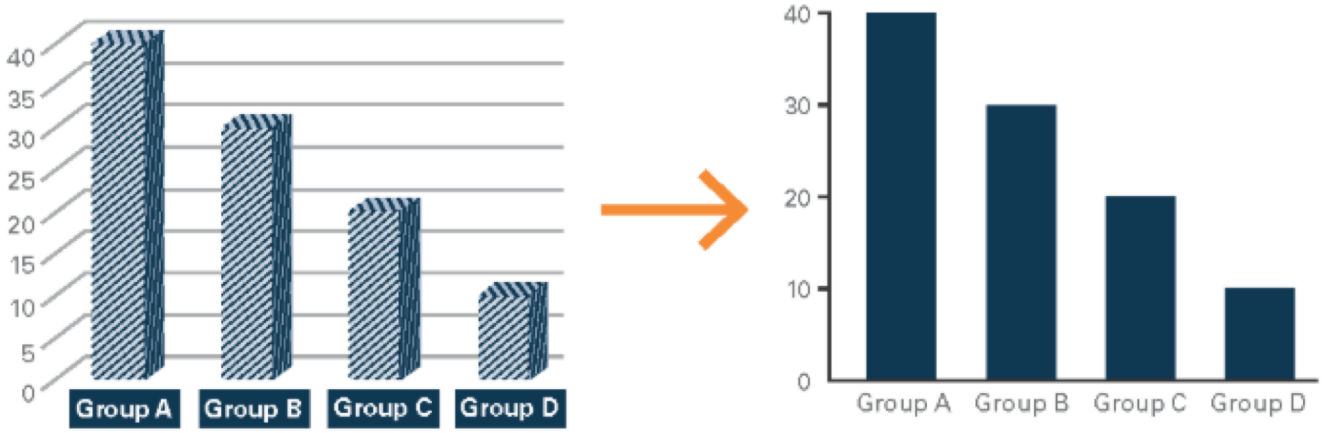
Best practices

- Start with a question, what information are you trying to communicate? What is the goal of the visualization?
- What data do you have available?
- What level of detail does it go down to?
- How can you use other data to supplement your data?

- Core principles (next slides)

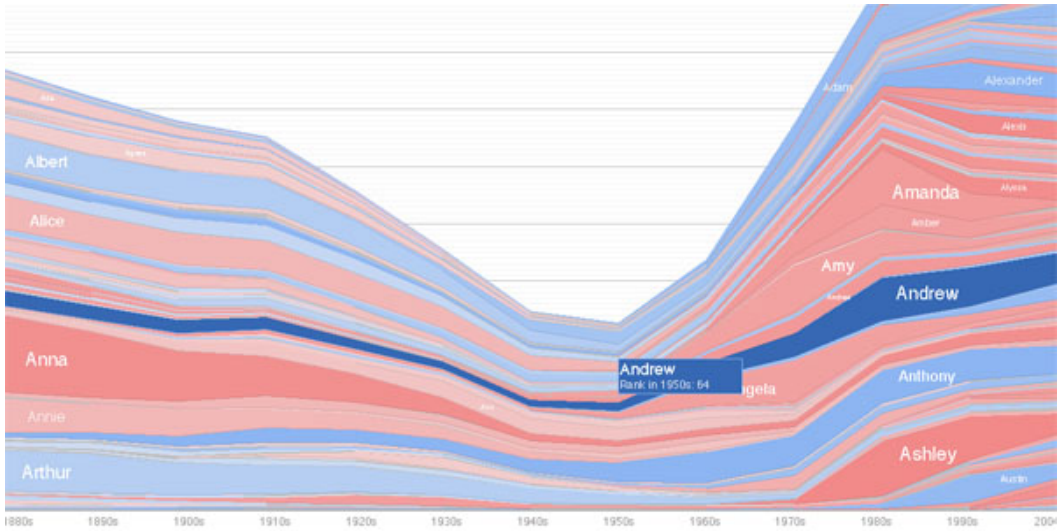
Reduce clutter

- Simple is better:
  - Gratuitous features (does it review more information?)
  - Reduce chartjunk/tablejunk



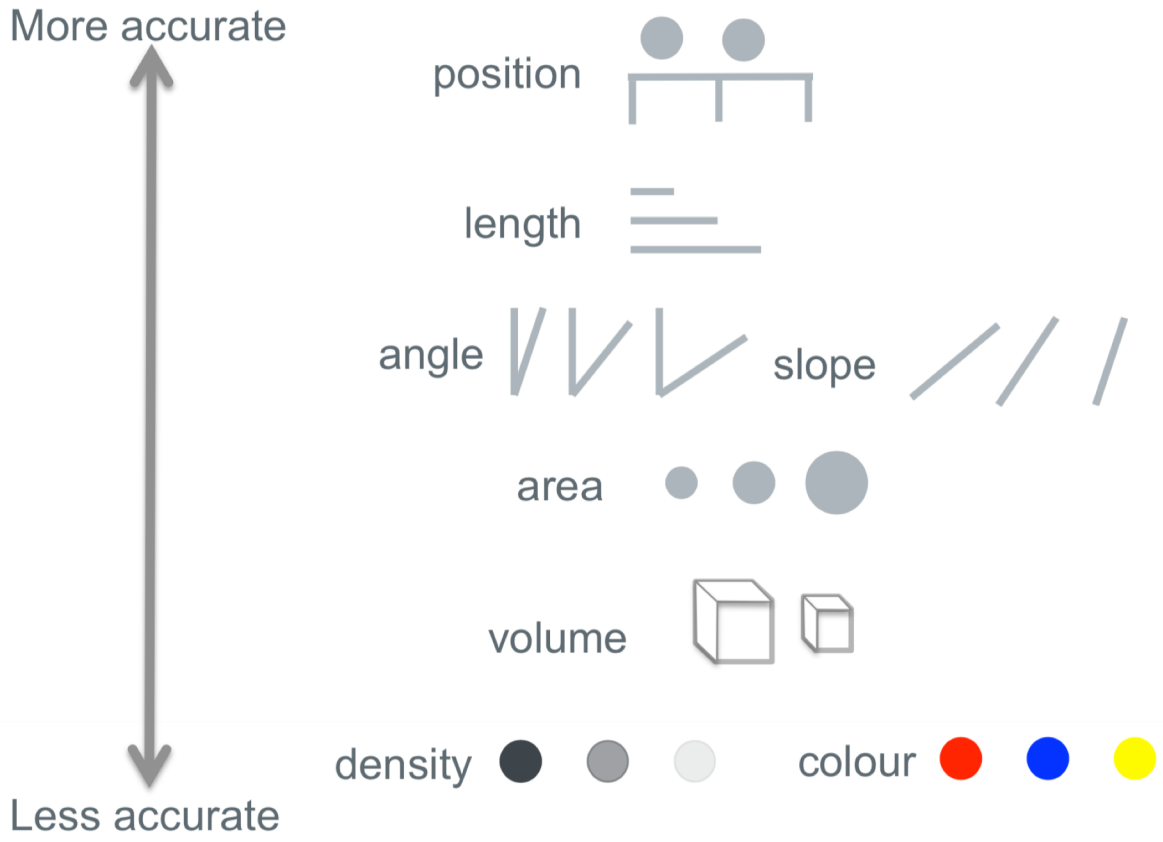
Interactive chart design

- With interactive charts you can keep things very simple by **hiding** and **dynamically revealing** important structure.
- On an interactive chart, you reveal the information most useful for **navigating** the chart.



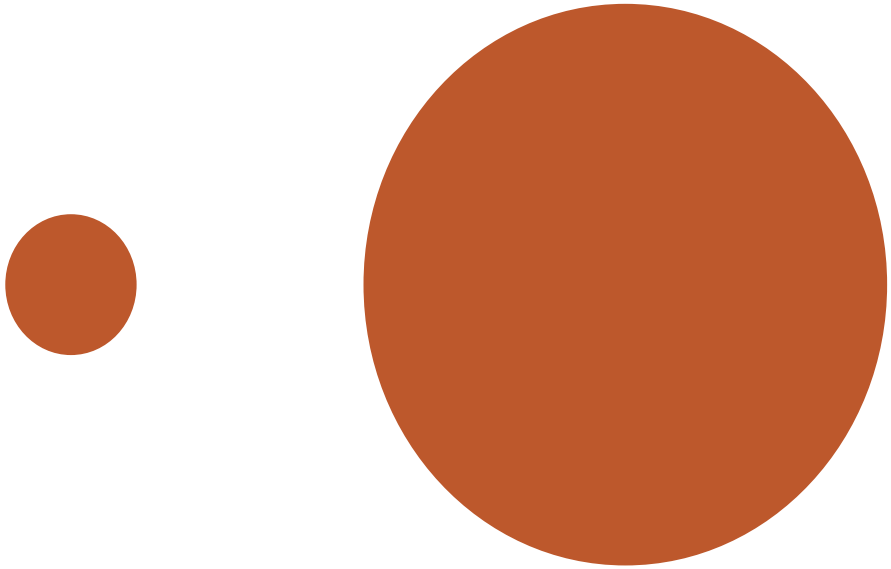
Encoding schemes

Accuracy of Quantitative Perceptual Tasks

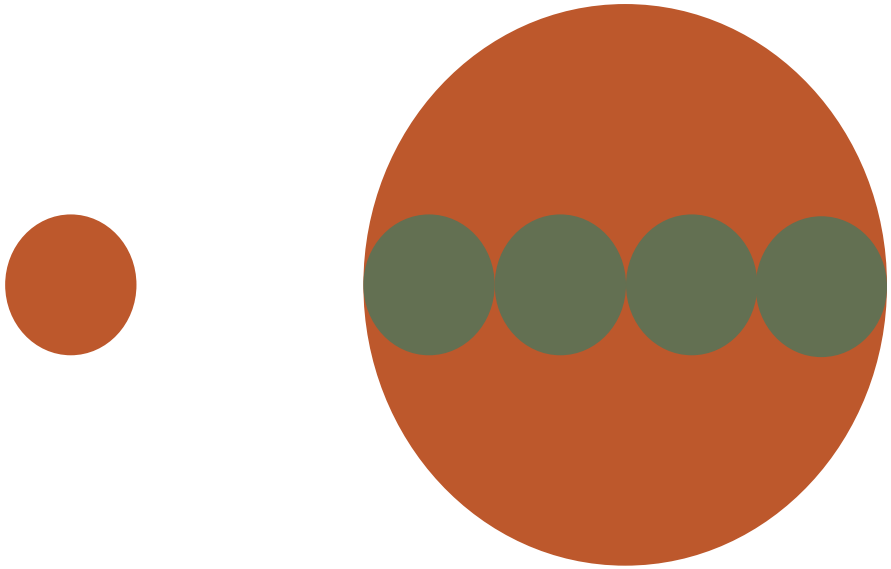


Cleveland, W.S. & McGill, R. Science 229, 828-833 (1985)

Compare area of circles



Compare area of circles



Which is brighter?

Which is brighter?

(128, 128, 128)

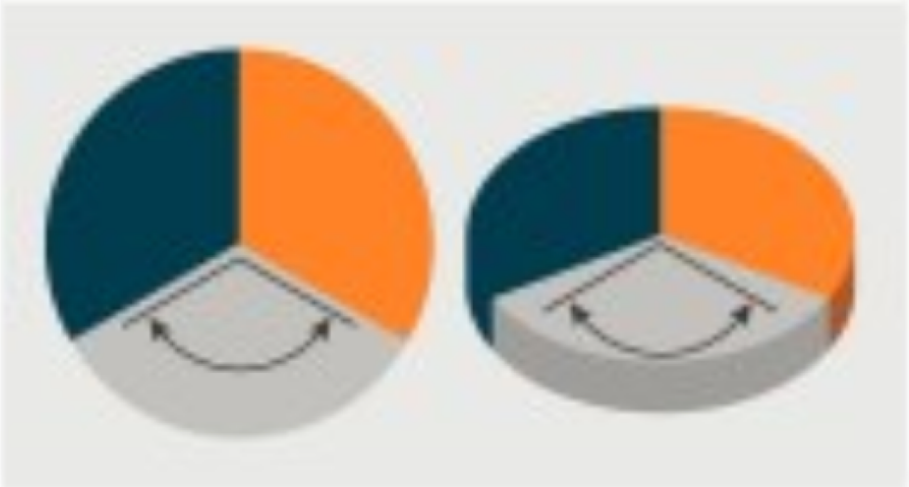
(144, 144, 144)

Just noticeable Difference

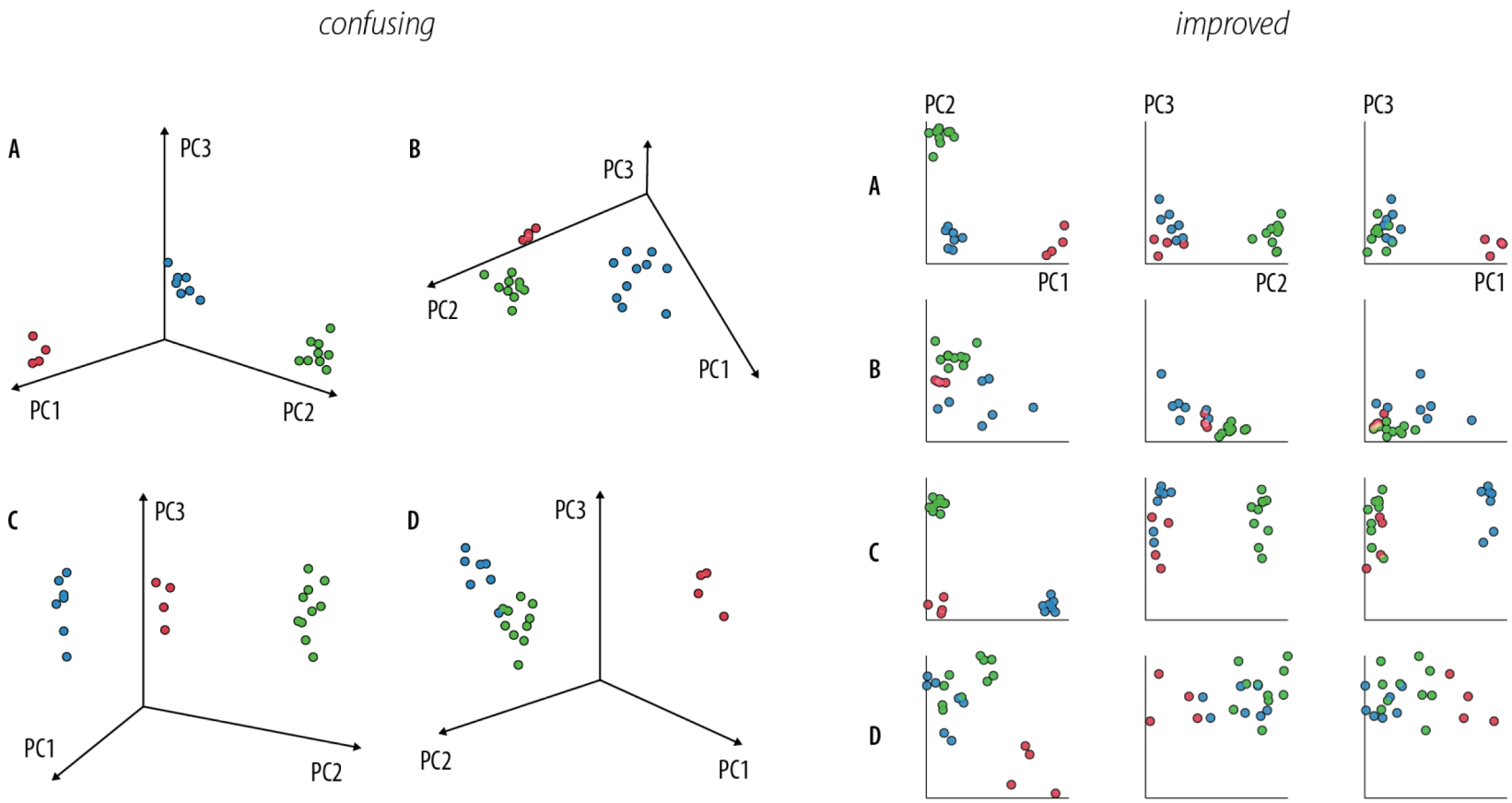
- Ratios more important than magnitude

- Most continuous variations in stimuli are perceived in discrete steps

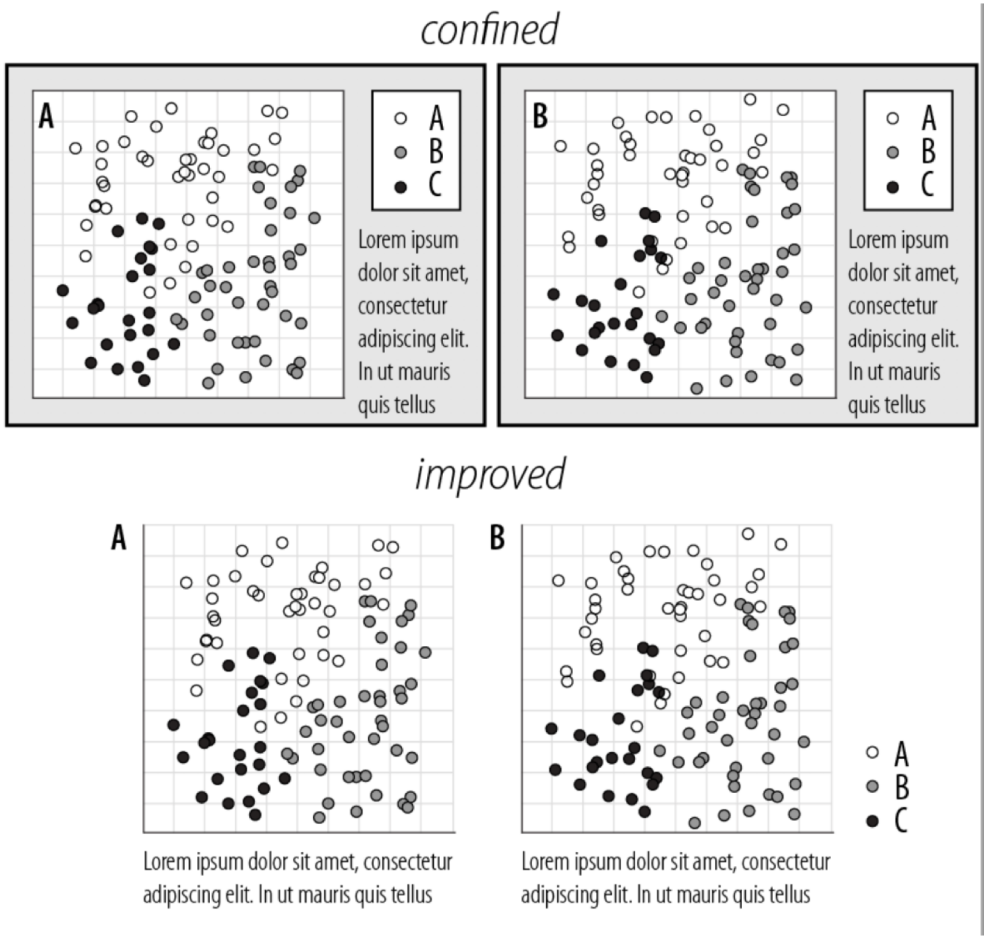
Careful with 3D and pie charts



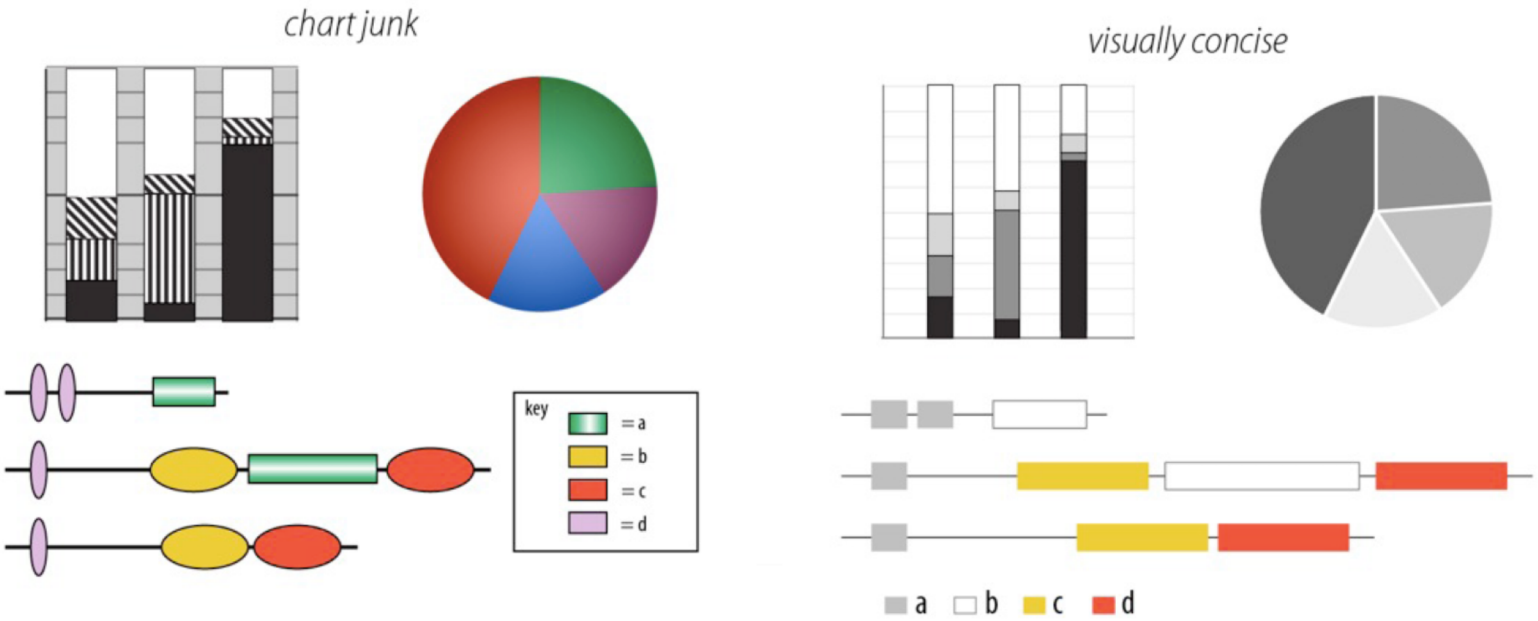
Avoid 3D


confusing
improved
A
PC3
PC1
PC2
B
PC3
PC2
PC1
C
PC3
PC1
PC2
D
PC3
PC2
PC1
A
PC2
PC1
PC3
PC2
PC3
PC1
B
C
D

Increase data to ink ratio



Avoid chart junk



Use color! – Brewer Palettes

- ! Brewer palettes (colorbrewer2.org) provide a range of palettes based on HSV model which make life easier for us….

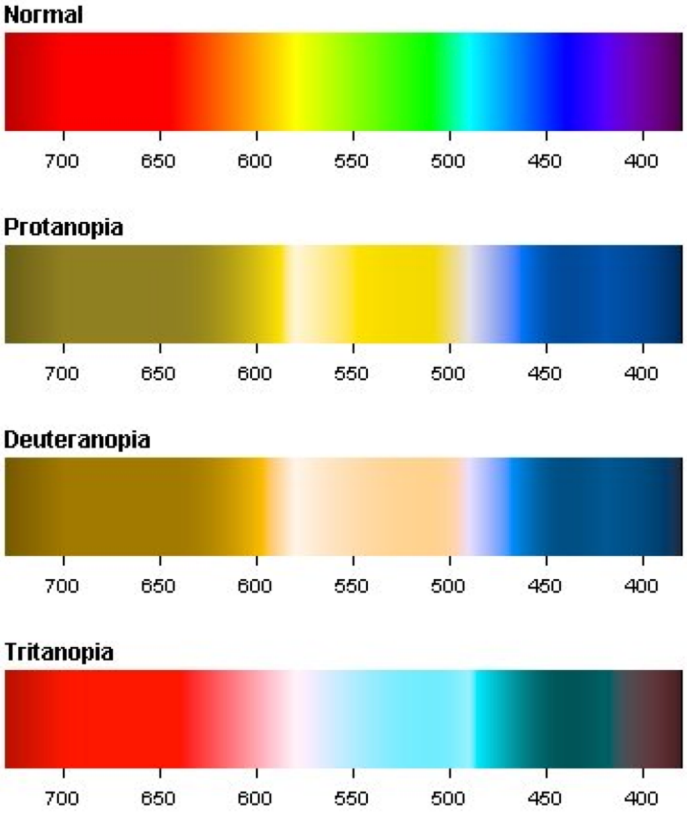
Avoid the use of hue to encode quantitative variables

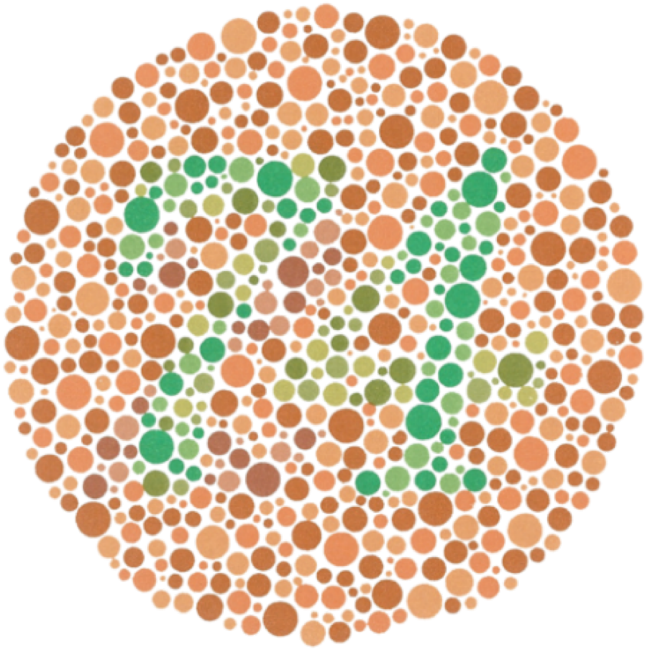
Quantitative encoding e.g. heat maps

Two-sided quantitative encodings

Color

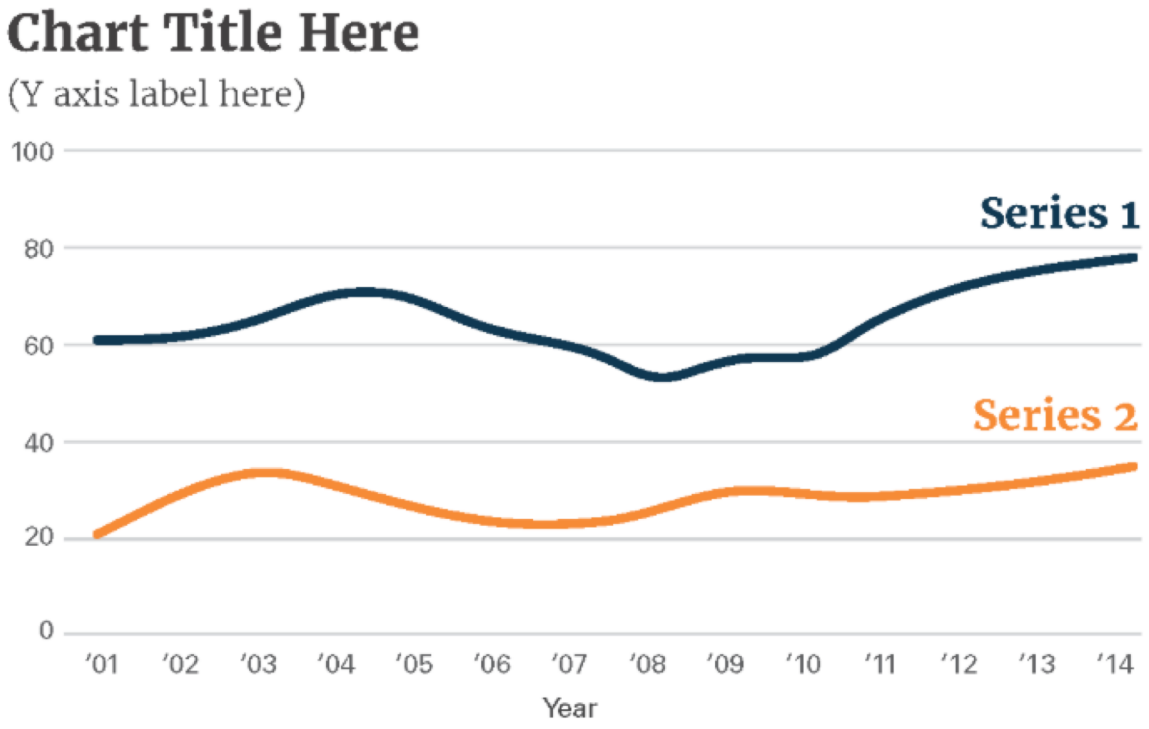
- 10-20% of population are red/green color blind!





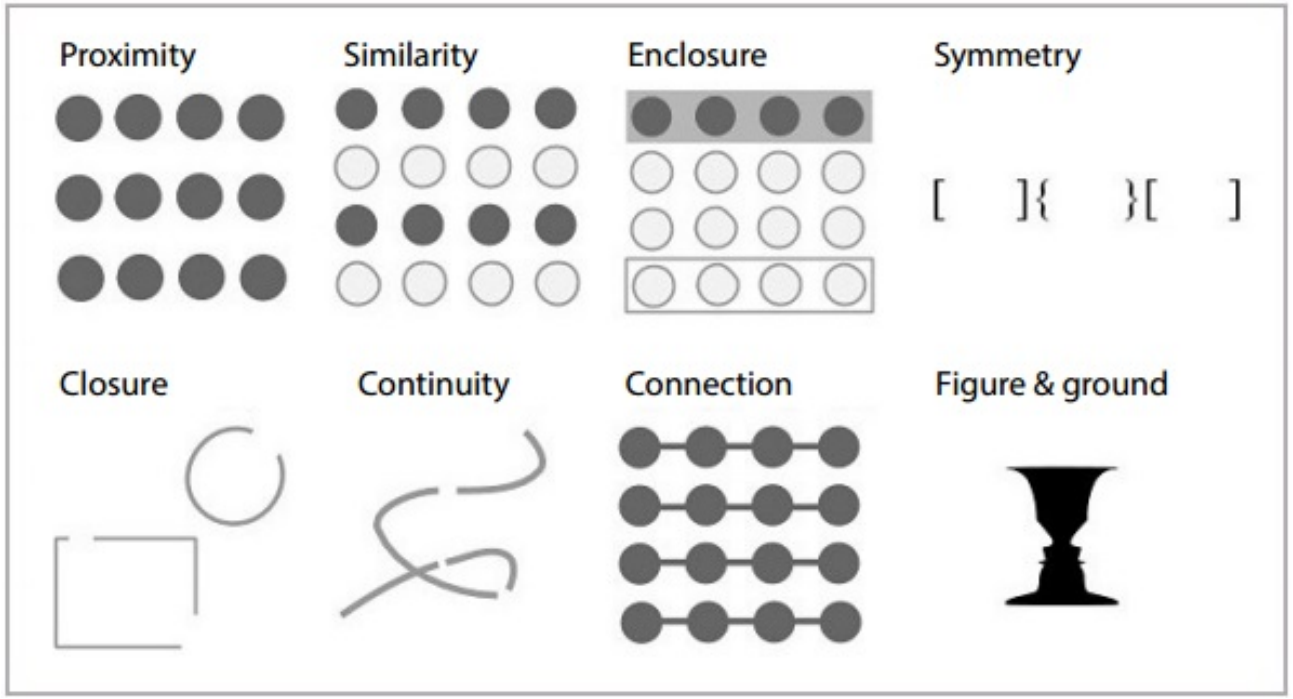
Integrate text and graphics

- Avoid slideshow effect: narration of images in text.
- Better: visuals complement text, but can be read standalone as well



Gestalt psychology principles (1912)

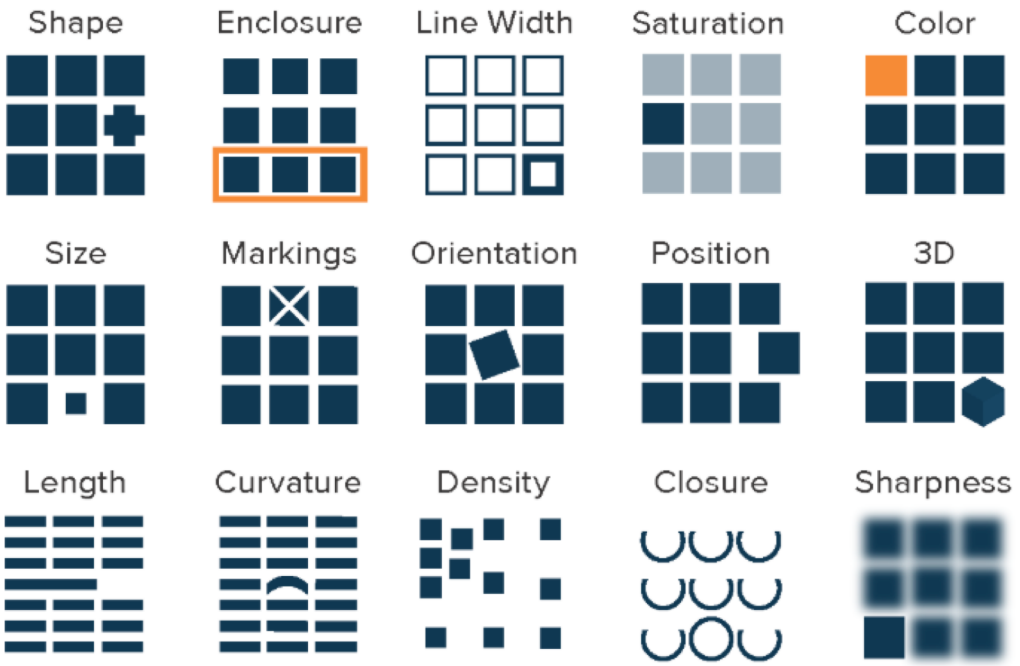
- Because our eyes detect a limited set of visual characteristics (e.g. shape/contrast), so we unconsciously group them.



Source http://blog.fusioncharts.com/2014/03/how-to-use-the-gestalt-principles-for-visual-storytelling-podv/

Preattentive processing

- Preattentive processing → What is perceived without dedicated attention.
- To draw attention, clearly break the groups

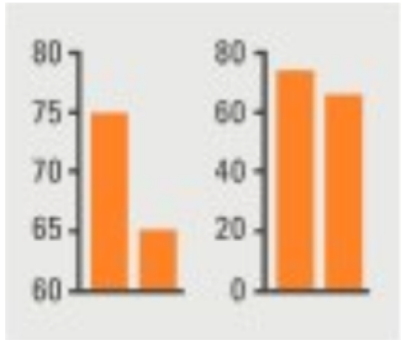


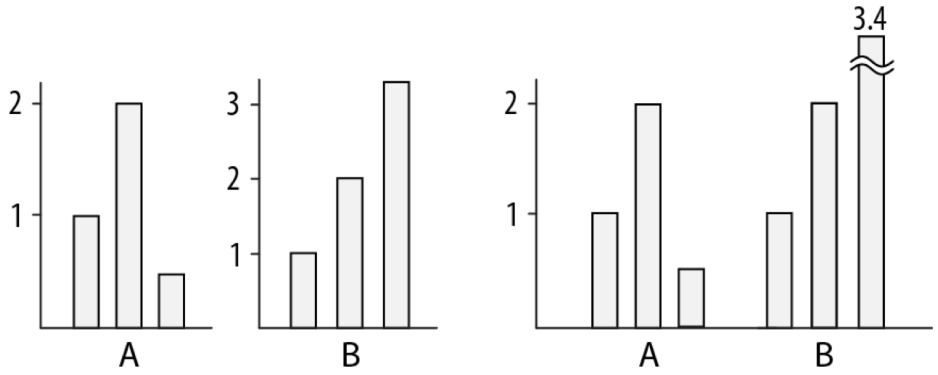
Who is your audience?

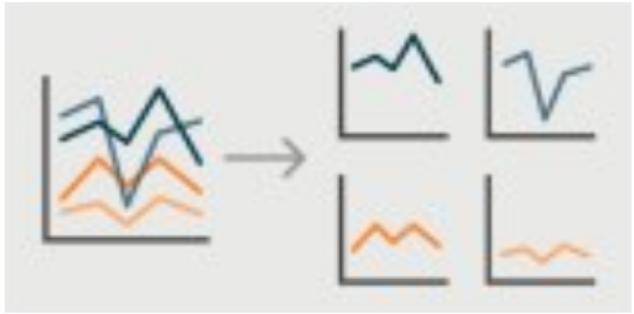
- How detailed do they want to see the data?
- Do they have a technical background?

Other pointers

- Break up complicated graphs
- Labels should be readable
- Include annotation to facilitate annotation/where to look

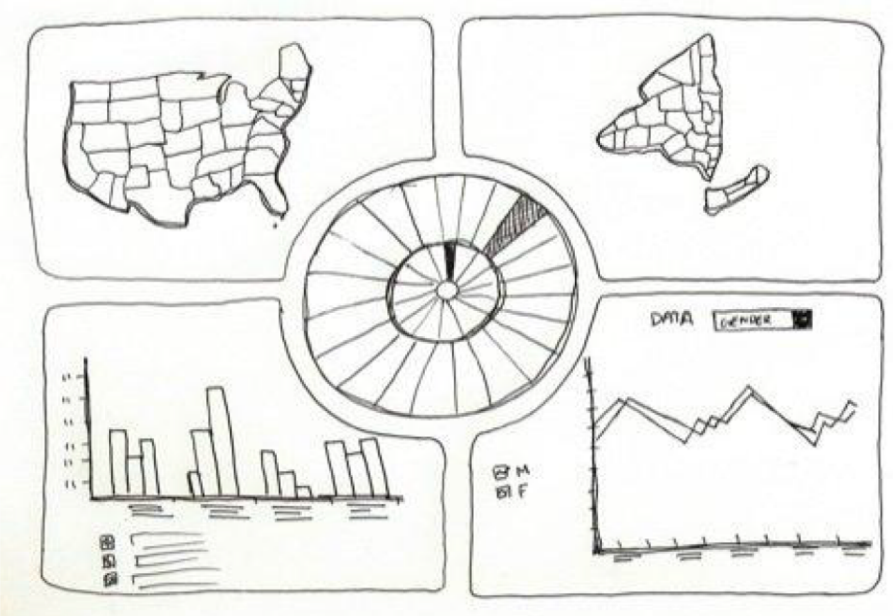






Make a sketch (pencil & paper)

- How will the visualization(s) be viewed? (desktop, mobile, print)

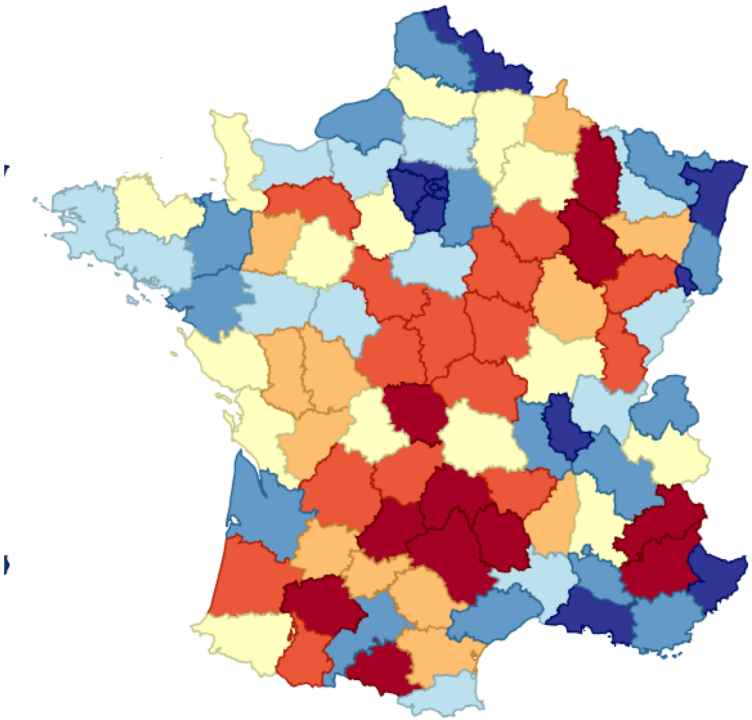


Common graphs

2D Graphs

Choropleth

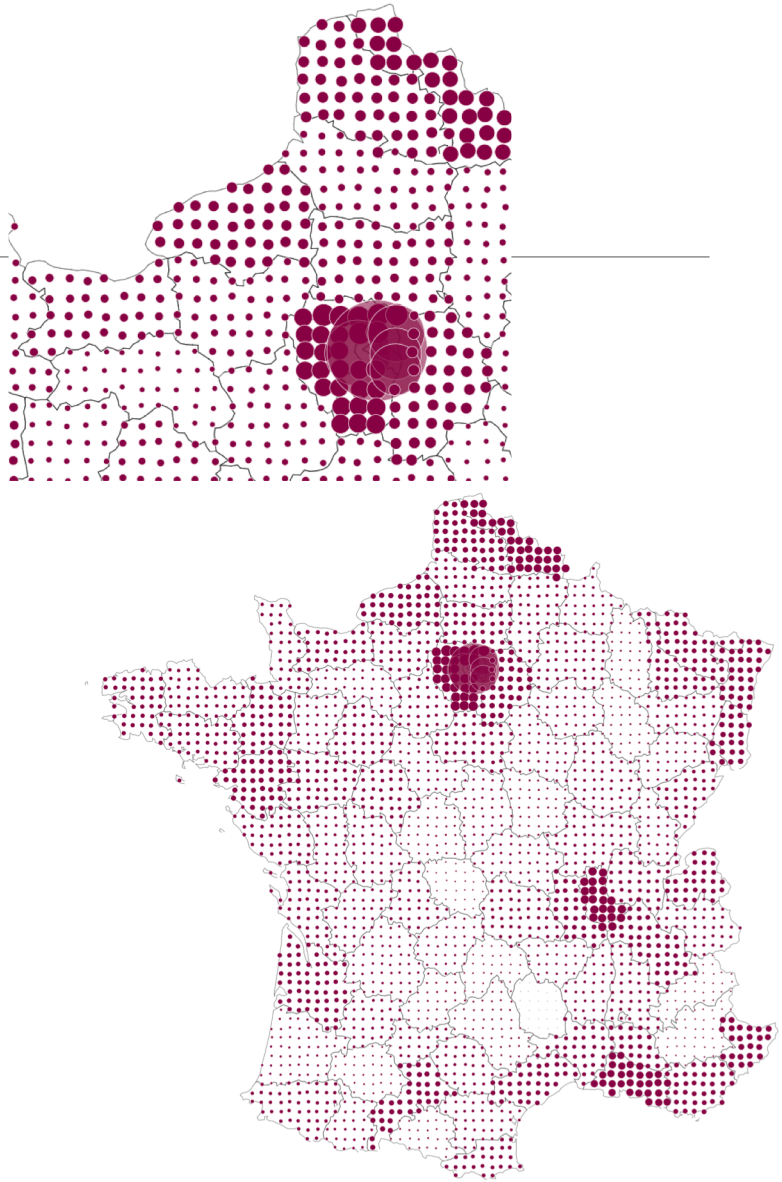
- Displays divided geographical areas or regions that are colored, shaded or patterned in relation to a data variable
- Example: population data for each of the dept. of France.
- Downsides:
  - You don't know real values (hoover?)
  - Common error: encoding raw data values (e.g. population), instead of normalized (e.g. population per square km).



Map source: GADM,

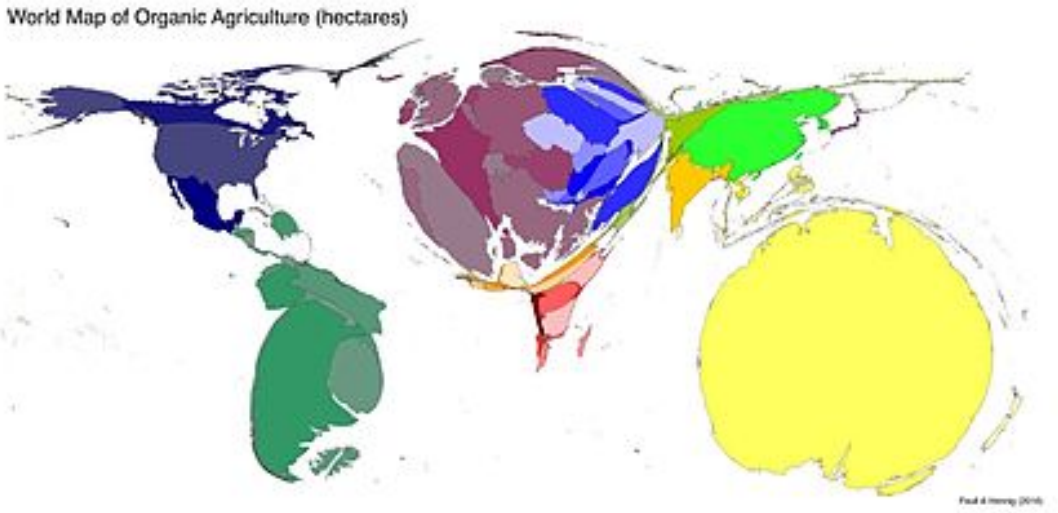
Dot Grid Maps

- Advantage: shows quantity or density of a distribution
- Example: population distribution across departments in France
- We can compare the population density by looking at individual circles while still getting an impression of the total population for each department.



Cartograms

- Distorts the geometry or space of a map to convey the information of an alternative variable (e.g. population or travel time).
- The two main types:
  - Area cartograms (ex: organic farming per country)
  - Distance cartograms



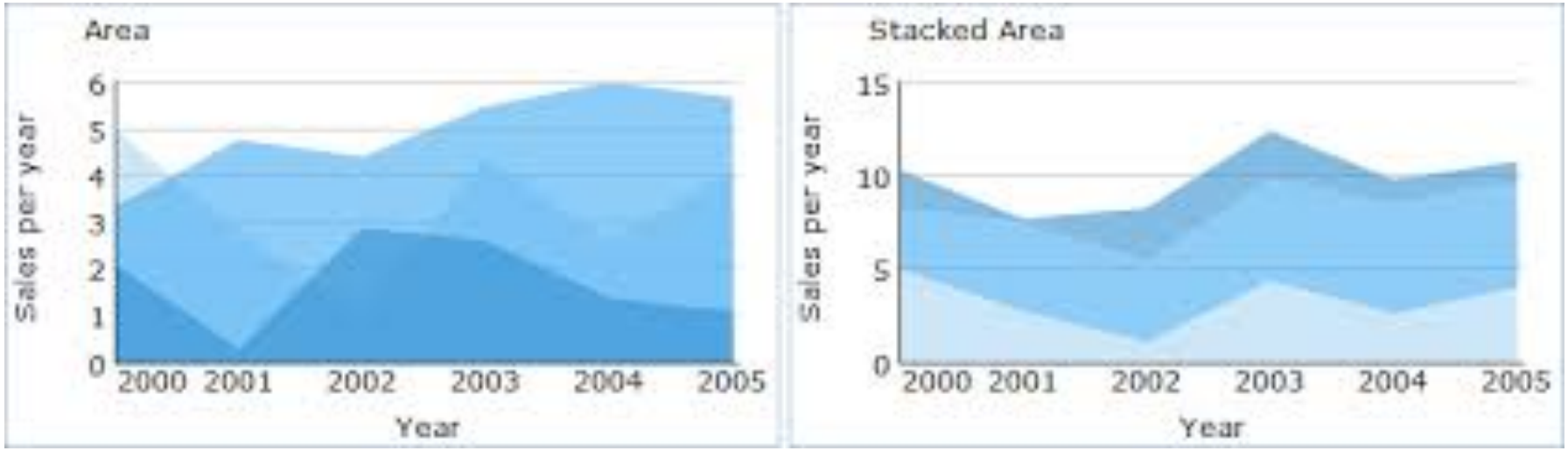
Temporal

Time Series

- Can be connected scatterplot

Area chart

- Similar to a line chart, it allows you to represent more than one value on a line while the transparent shading and colors allow you to see the other data lines.
  → Allows for easier comparison between data, as all values are visible.

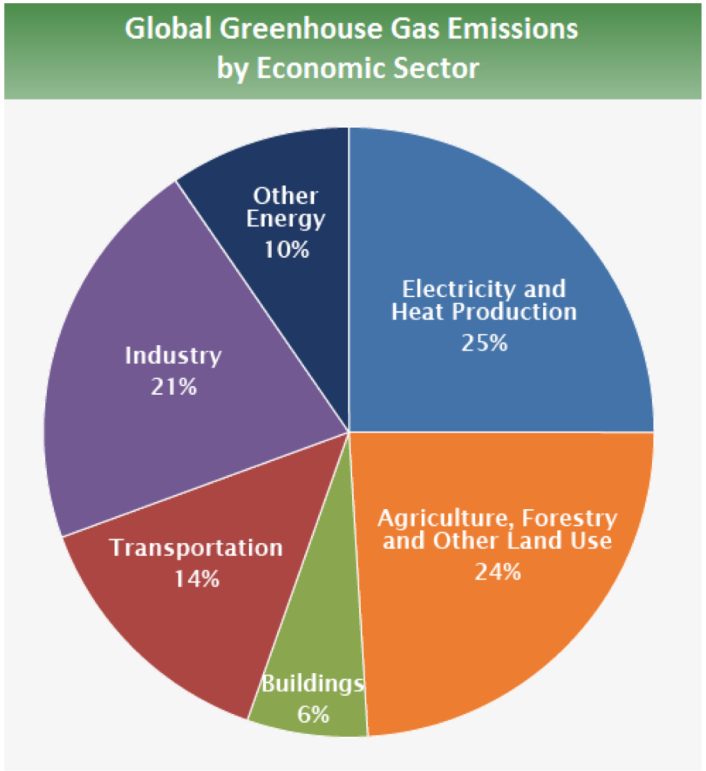


Arc diagrams

Multidimensional

Pie Chart

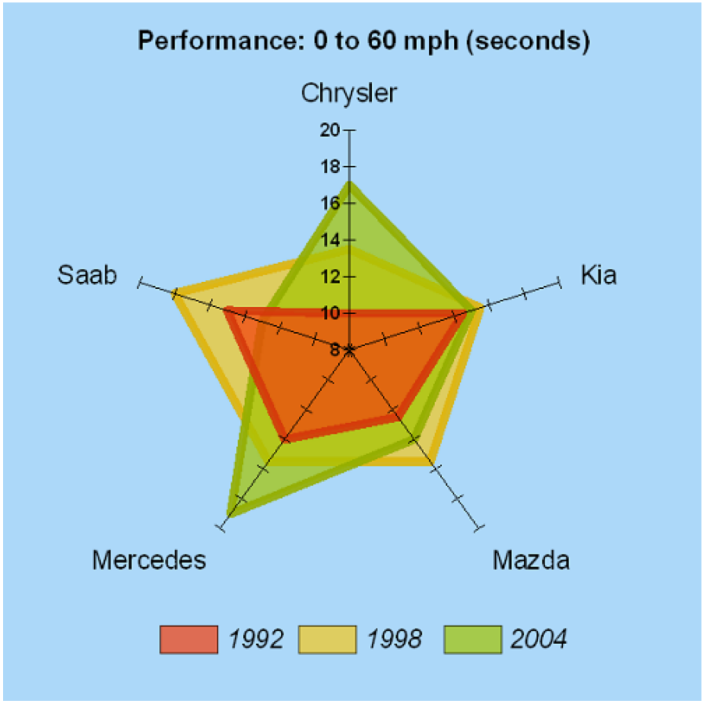
- Divided into sectors to illustrate numerical proportion; the arc length and angle of each sector is proportional to the quantity it represents.



Source: EPA

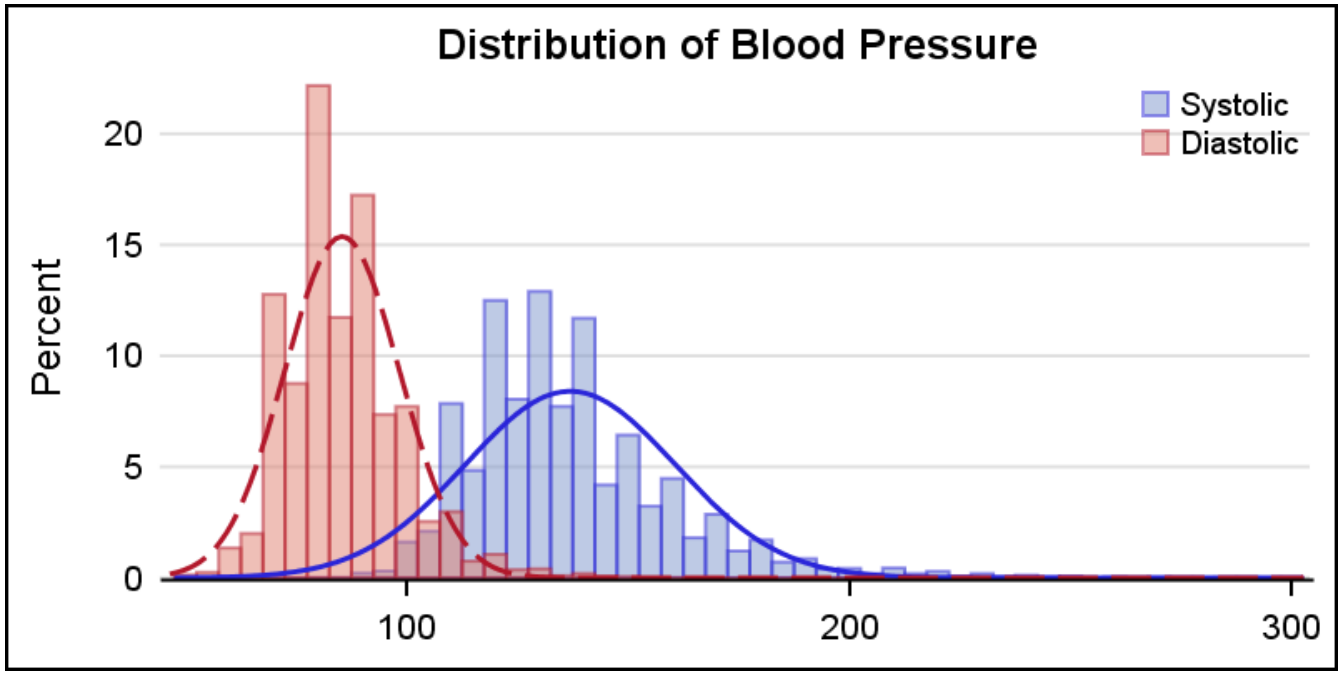
Radar/spider chart

- Used to plot values along a series of separate axis, starting at the center and working outwards to represent relative value.
- Usually, it's used when more than five values need representations and visualisation within on chart.
- Often used for sports statistics, budgets.



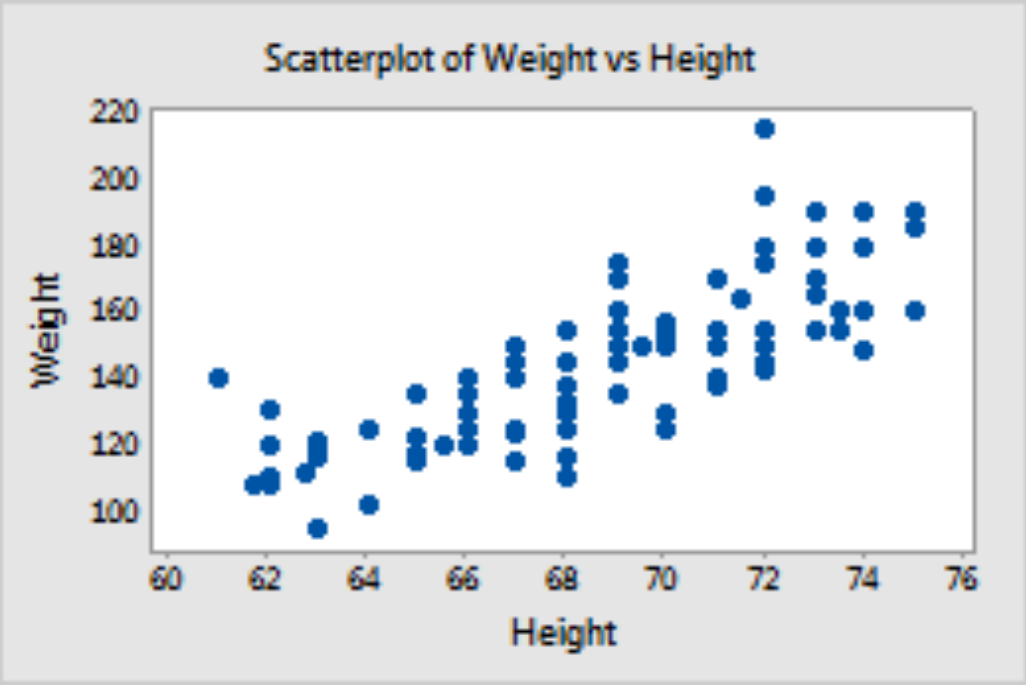
Histogram

- Display the distribution of data
- Uses rectangles with heights proportional to the count and widths equal to the “bin size” or range of small intervals.



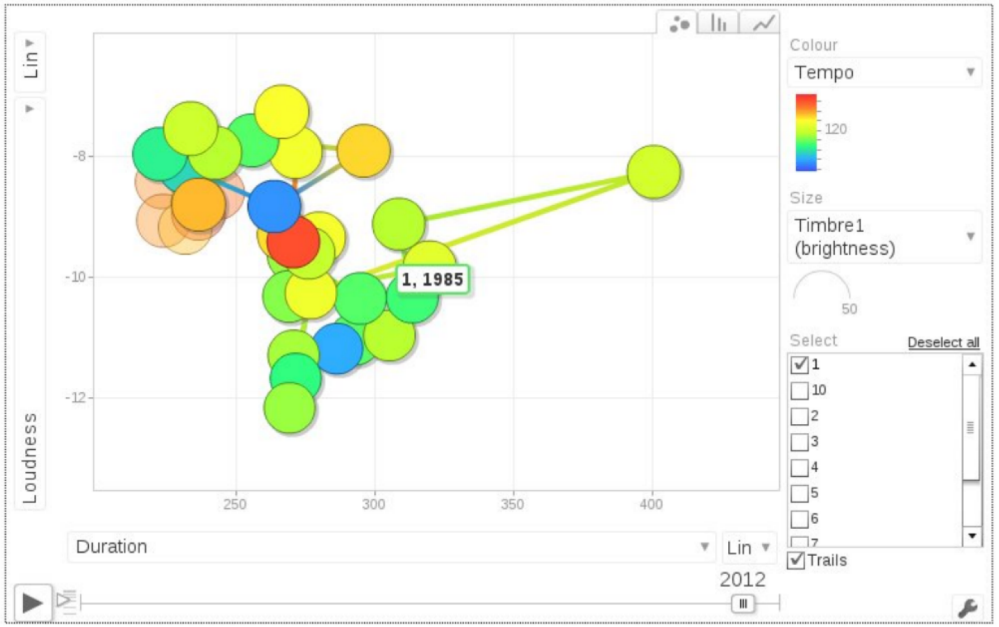
Scatter plot

- Displays values for two variables for a set of data as a collection of points.



Bubble chart

- Display data by location on an axis and add a third variable in size of the bubble, similar to a scatter chart
- Other dimensions: color, animation (motion chart)

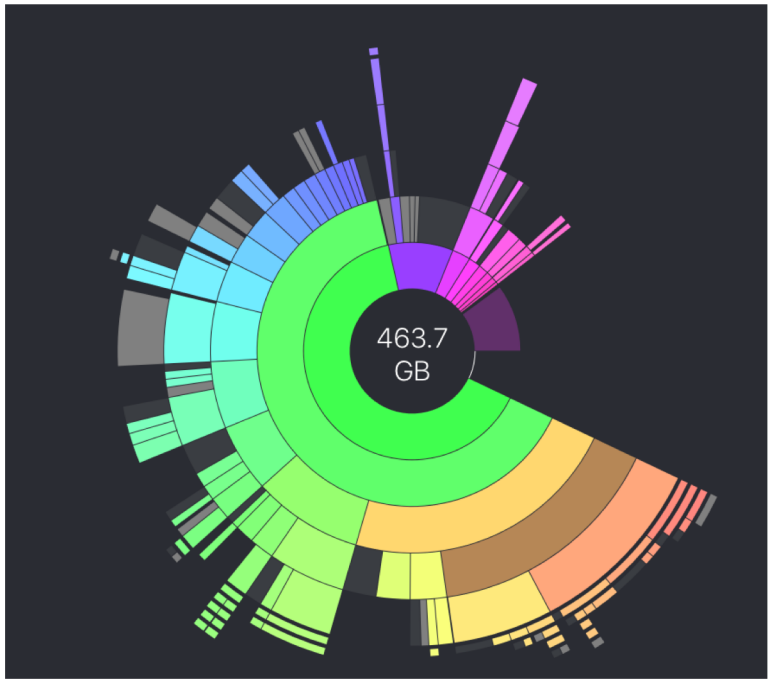


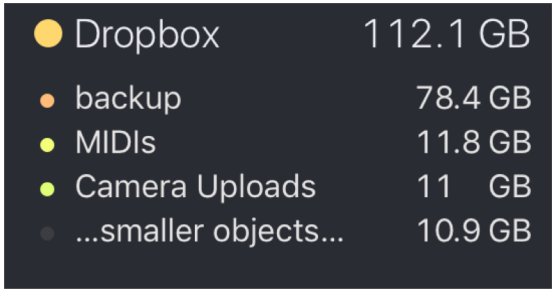
dorienherremans.com/dance

Hierarchical

Ring diagram

- Multilevel pie chart that visualises hierarchical data with concentric circles.

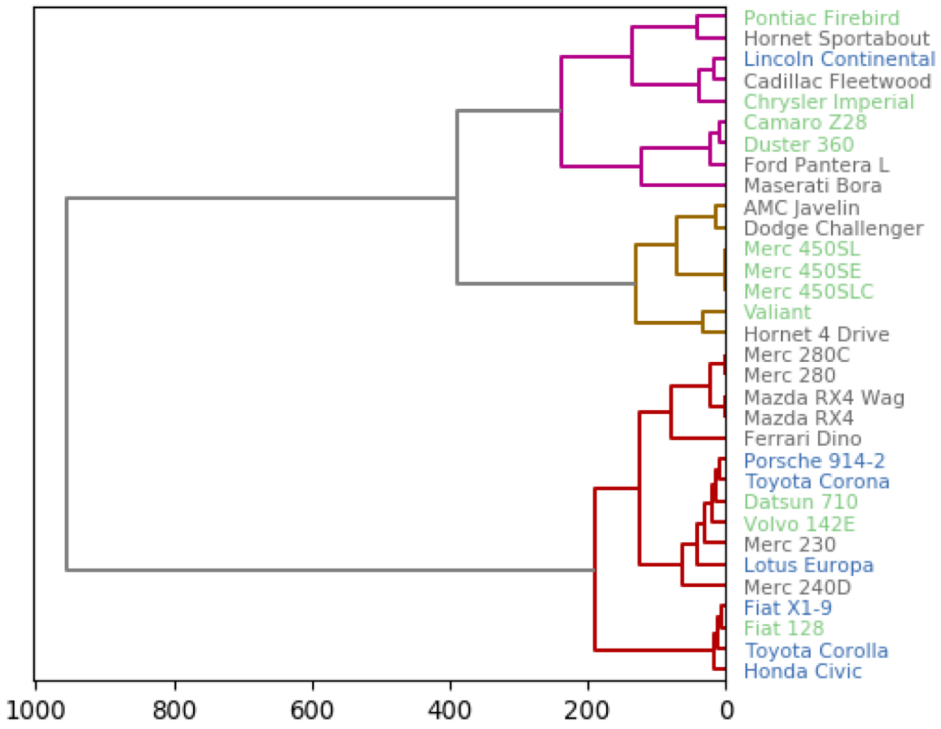




Source: Daisydisk

Tree diagram

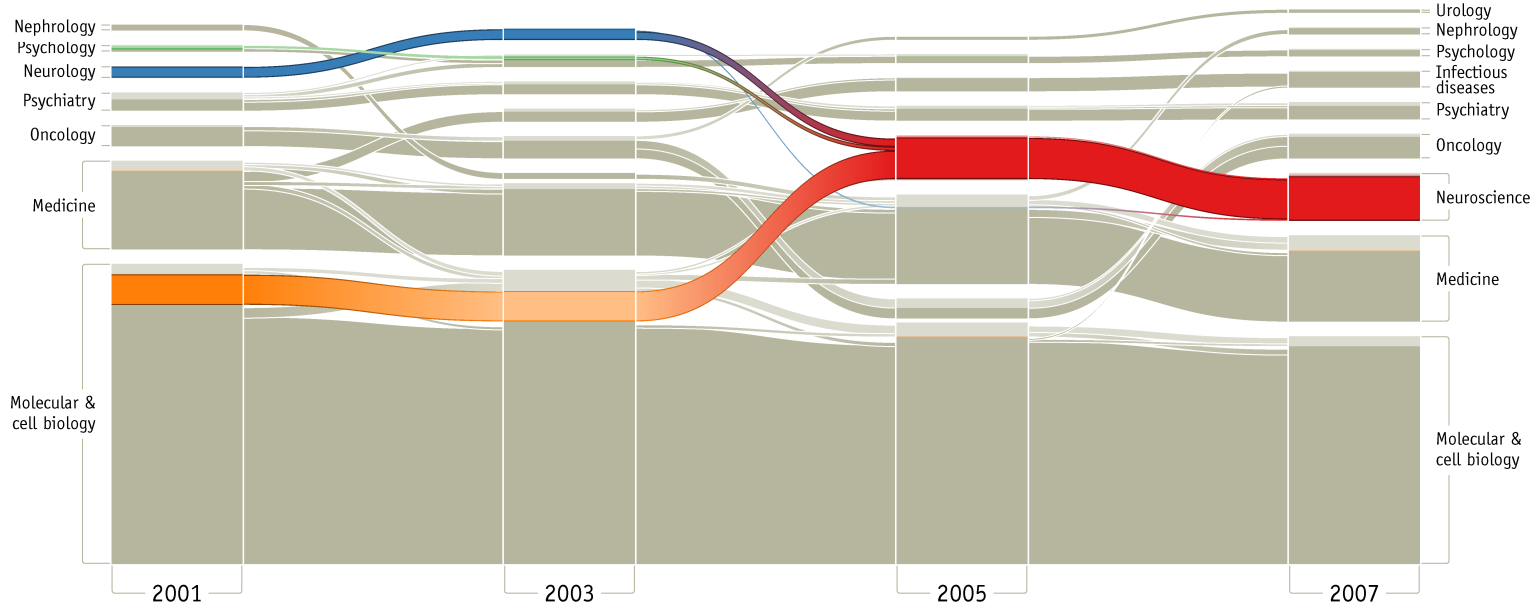
- Represents the hierarchical nature of a structure in graph form. It can be visually represented from top to bottom or left to right.
- Also called dendrogram when representing clusters



Network

Alluvial diagram

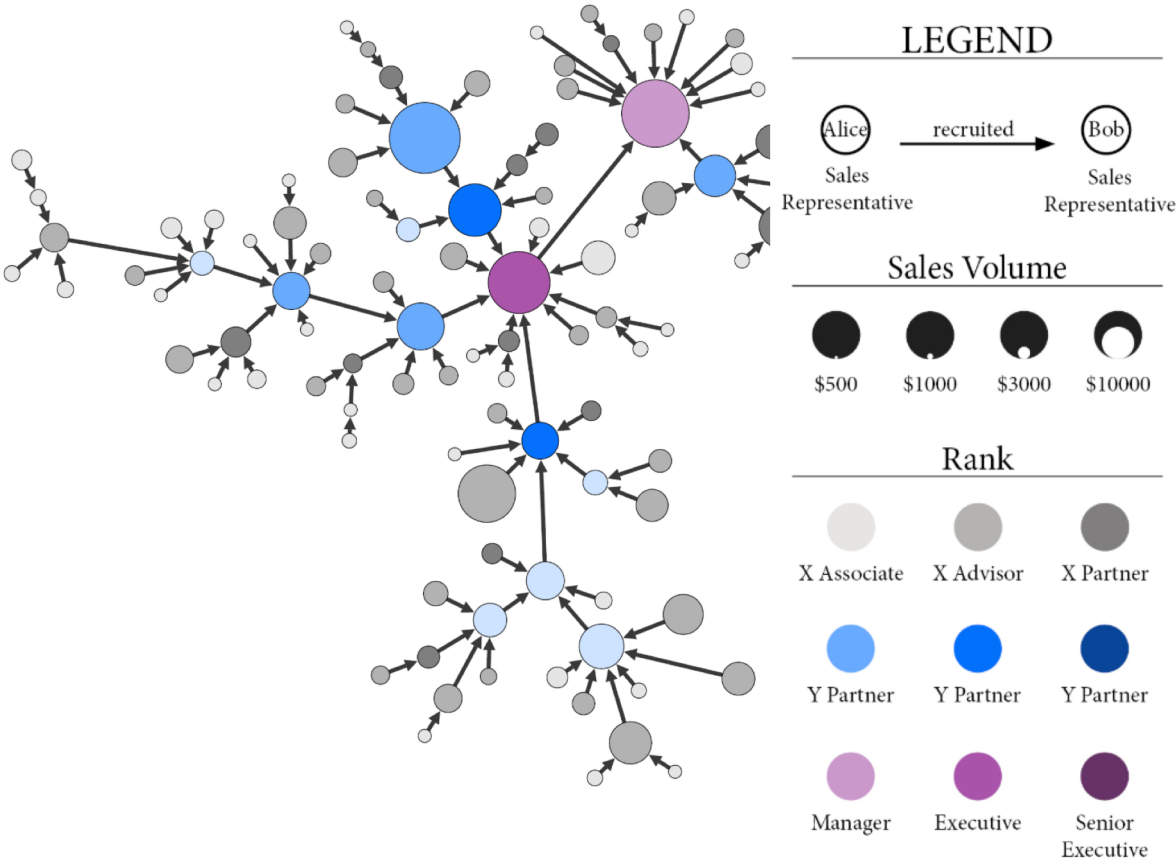
- An alluvial diagram is a type of flow diagram that represents changes in network structure over time.



Rosvall, M., & Bergstrom, C. T. (2010). Mapping change in large networks. *PLoS ONE, 5*(1), e8694.CC BY 2.5

Node-link diagram

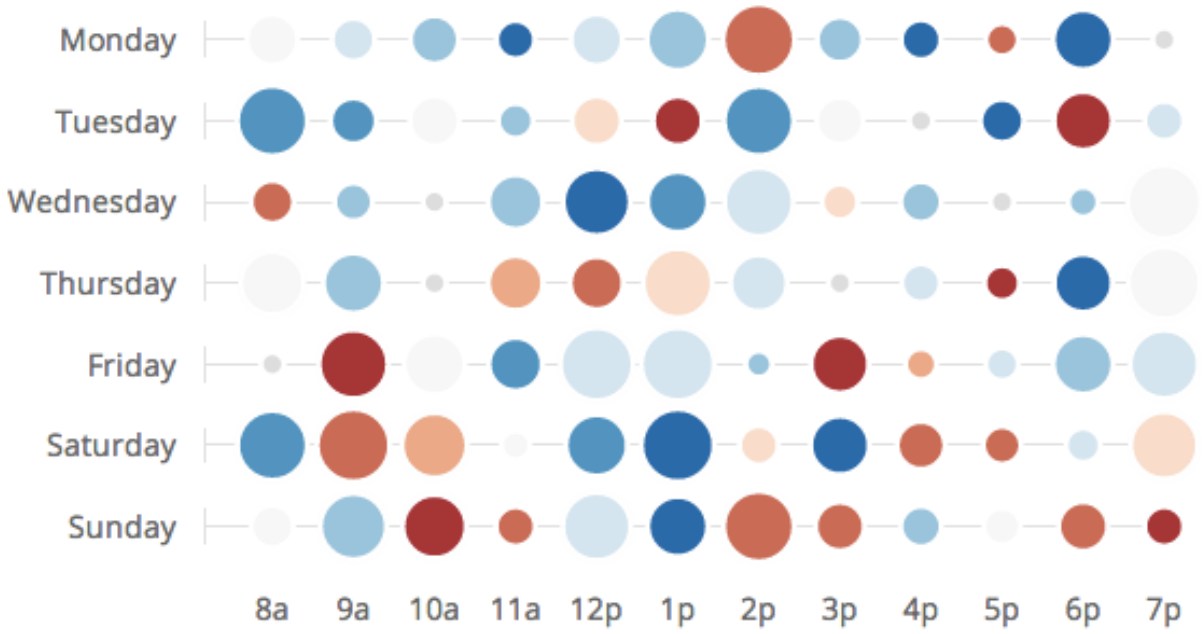
- Represents nodes as dots and links as line segments to show how a data set is connected.
- In 3D: hypergraph
- Example: network of company sales



Source: https://linkurio.us/blog/graph-viz-101-visual-language-node-link-diagrams/

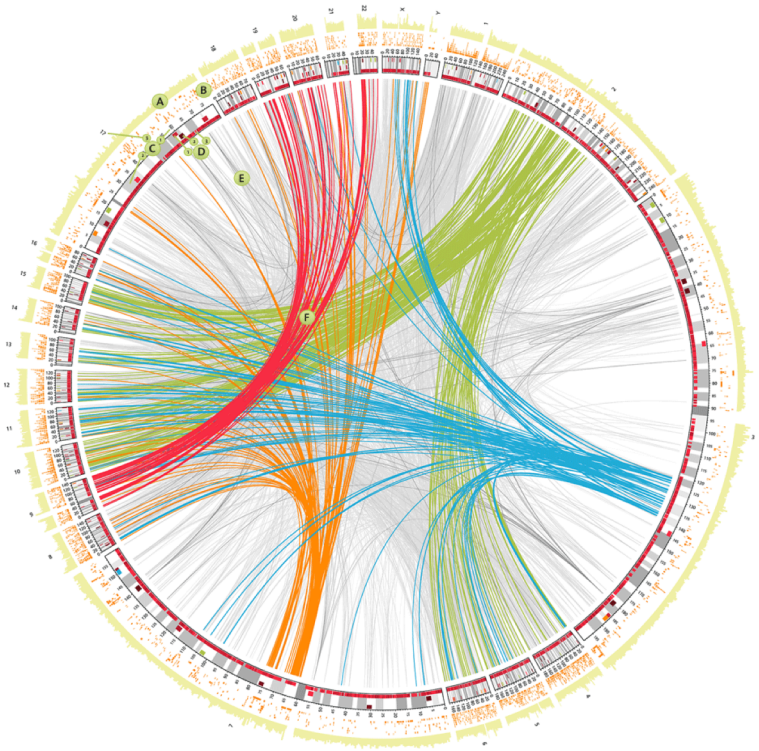
Matrix

- A matrix chart or diagram shows the relationship between two, three, or four groups of information and gives information about said relationship.



Circos – human genome

- Location of genes implicated in disease
- Regions of self-similarity
  - Structural variation
  - Within populations
- Uses:
  - links, heat maps, tiles, histograms
  - Use of colour, good continuity, length, transparency, ..

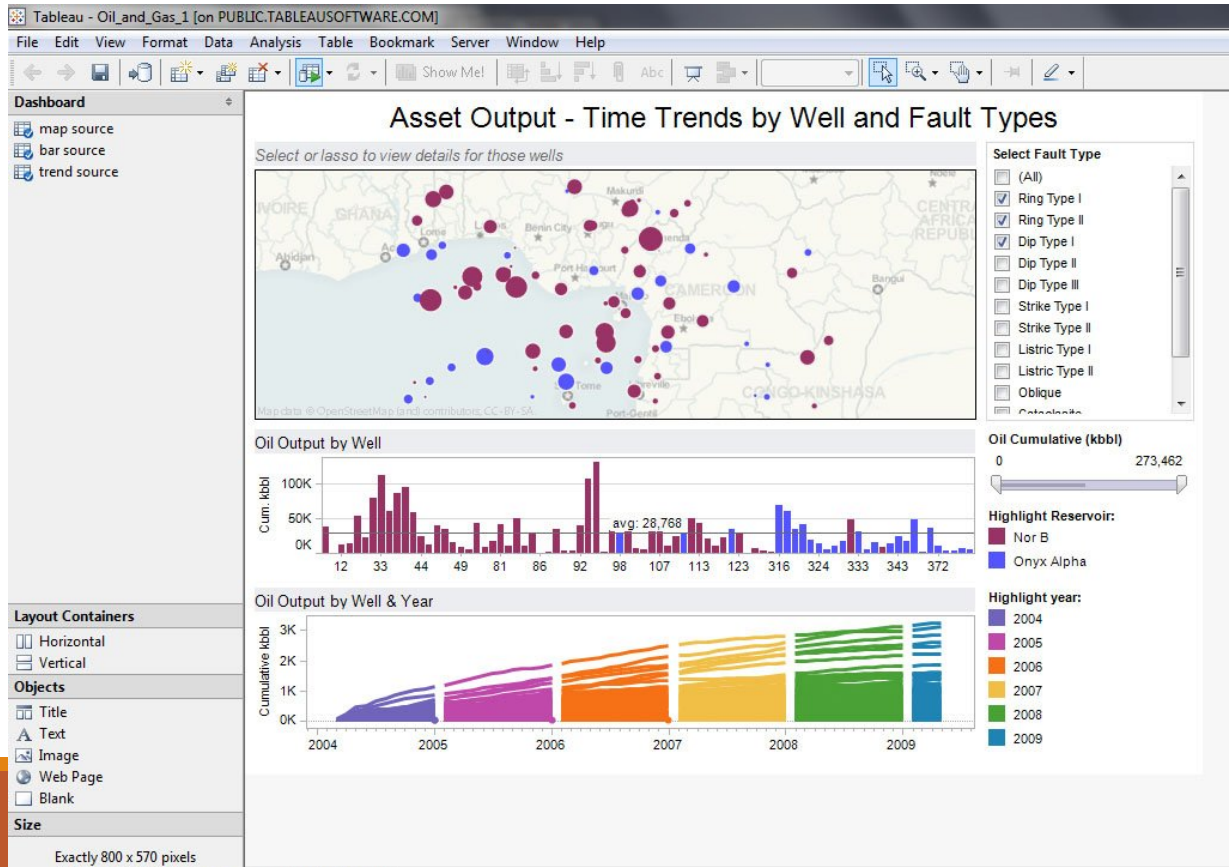


Tools

FOR MAKING GRAPHS

Tableau

- Focus on Business Intelligence
- Origin: to commercialize research which had been conducted at Stanford University



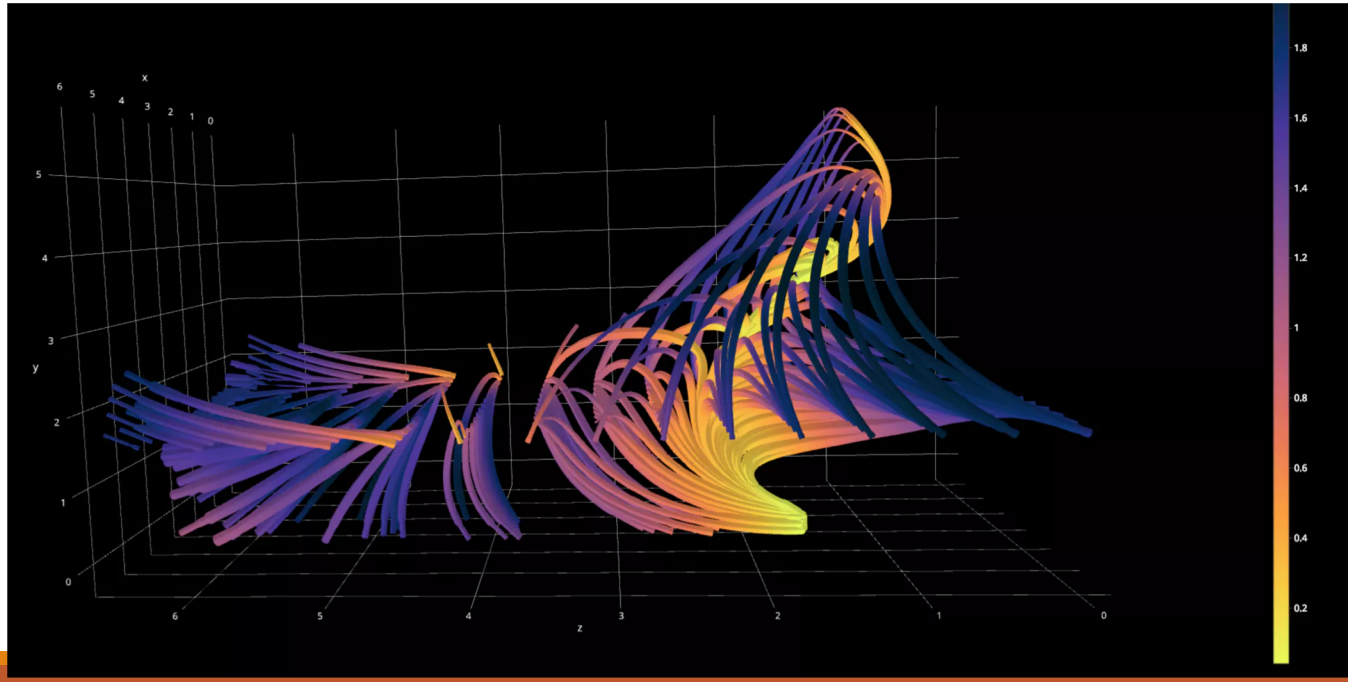
Python libraries - matplotlib

- Matplotlib: Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms

  ➔ Lab exercises

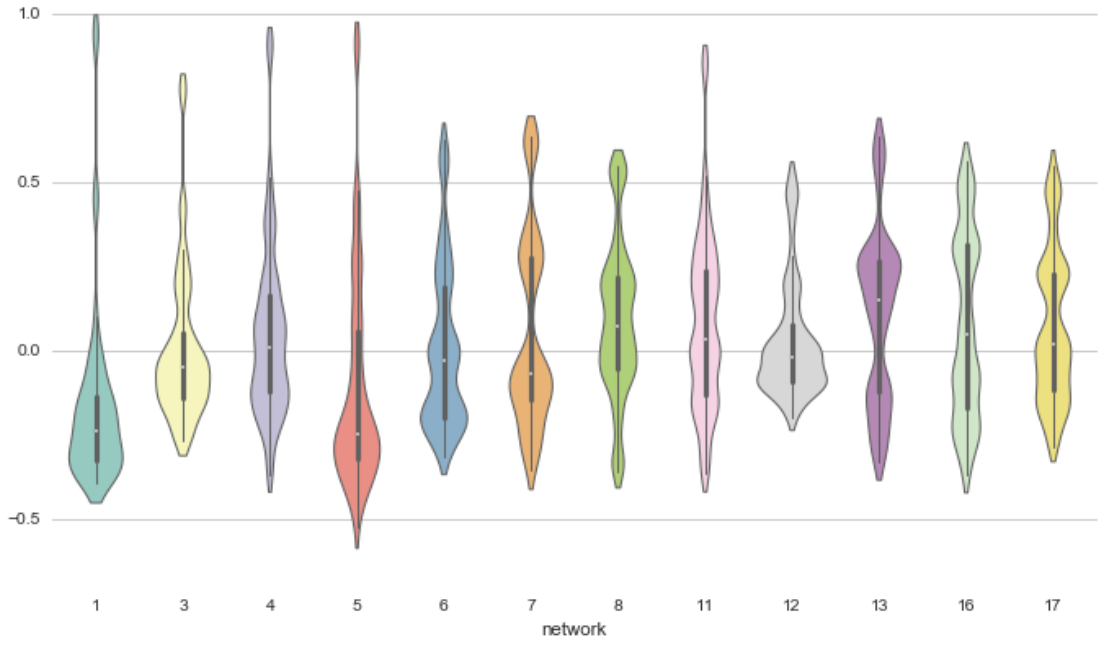
Python libraries - plotly

- A Python framework for building analytics web apps.
- Also offer maps.
- Interactive.



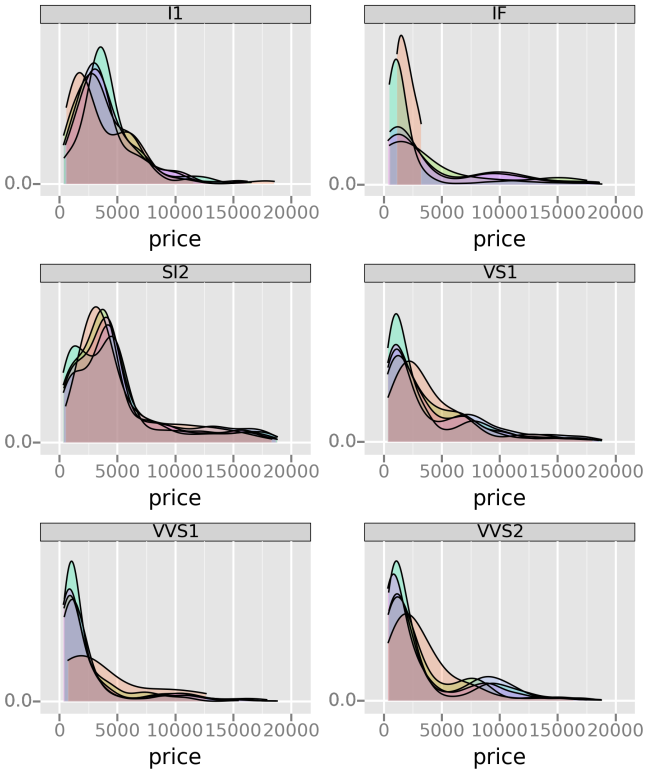
Python lib - seaborn

- ”Seaborn is a Python visualization library for **statistical plotting**. It comes equipped with preset styles and color palettes so you can create complex, aesthetically pleasing charts with a few lines of code.
- Seaborn is built on top of Python’s core visualization library matplotlib.



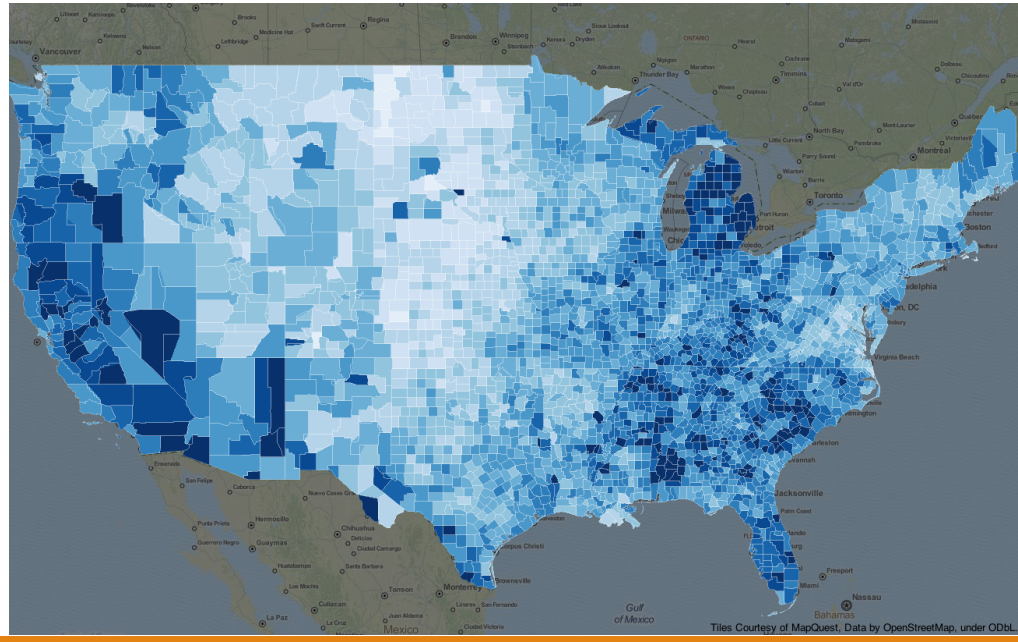
Python lib - ggplot

- Based on **ggplot2**, an **R** plotting system, and concepts from The Grammar of Graphics.
- ggplot operates differently than matplotlib: it lets you layer components to create a complete plot. For instance, you can start with axes, then add points, then a line, a trendline, etc.



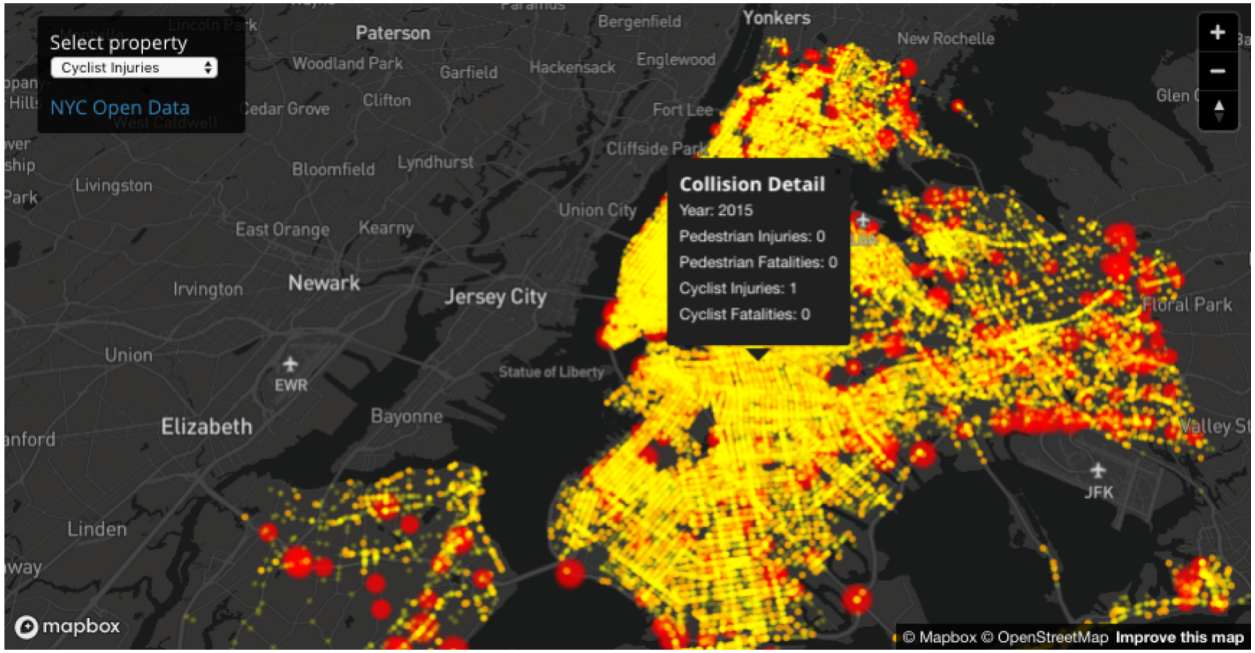
Python - geoplotlib

- geoplotlib is a toolbox for creating **maps** and plotting geographical data.
- You must have Pyglet (an object-oriented programming interface) installed to use geoplotlib.



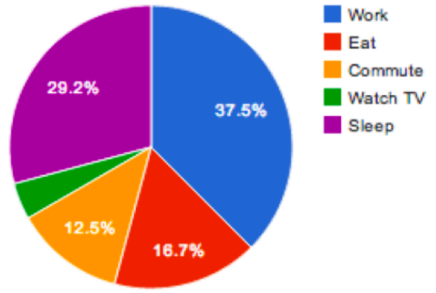
Mapbox

- Online maps for websites and applications (Foursquare, Lonely Planet, Facebook, the Financial Times, The Weather Channel and Snapchat,…)
- API & SDK's

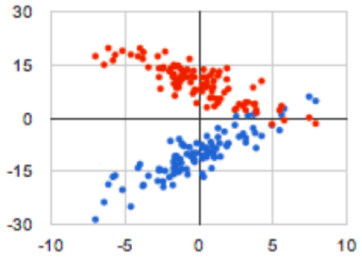


Google Chart Tools

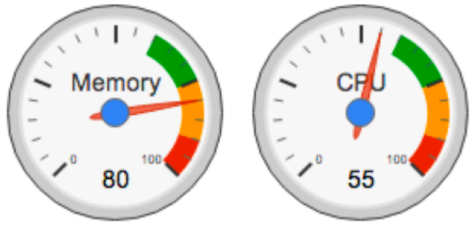
Pie Chart



Scatter Chart



Gauge

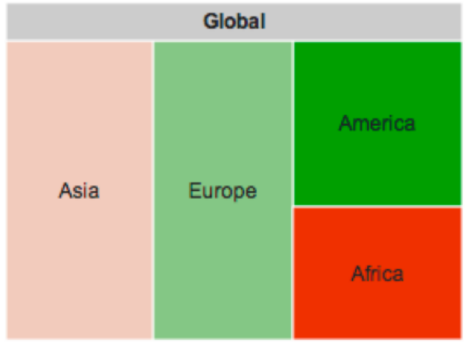


Geo Chart

Table

	Name	Salary	Full Time
1	Mike	$10,000	✓
2	Jim	$8,000	✗
3	Alice	$12,500	✓
4	Bob	$7,000	✓

Treemap



Upcoming lab

- Python 3 & Matplotlib

- Guest lecture