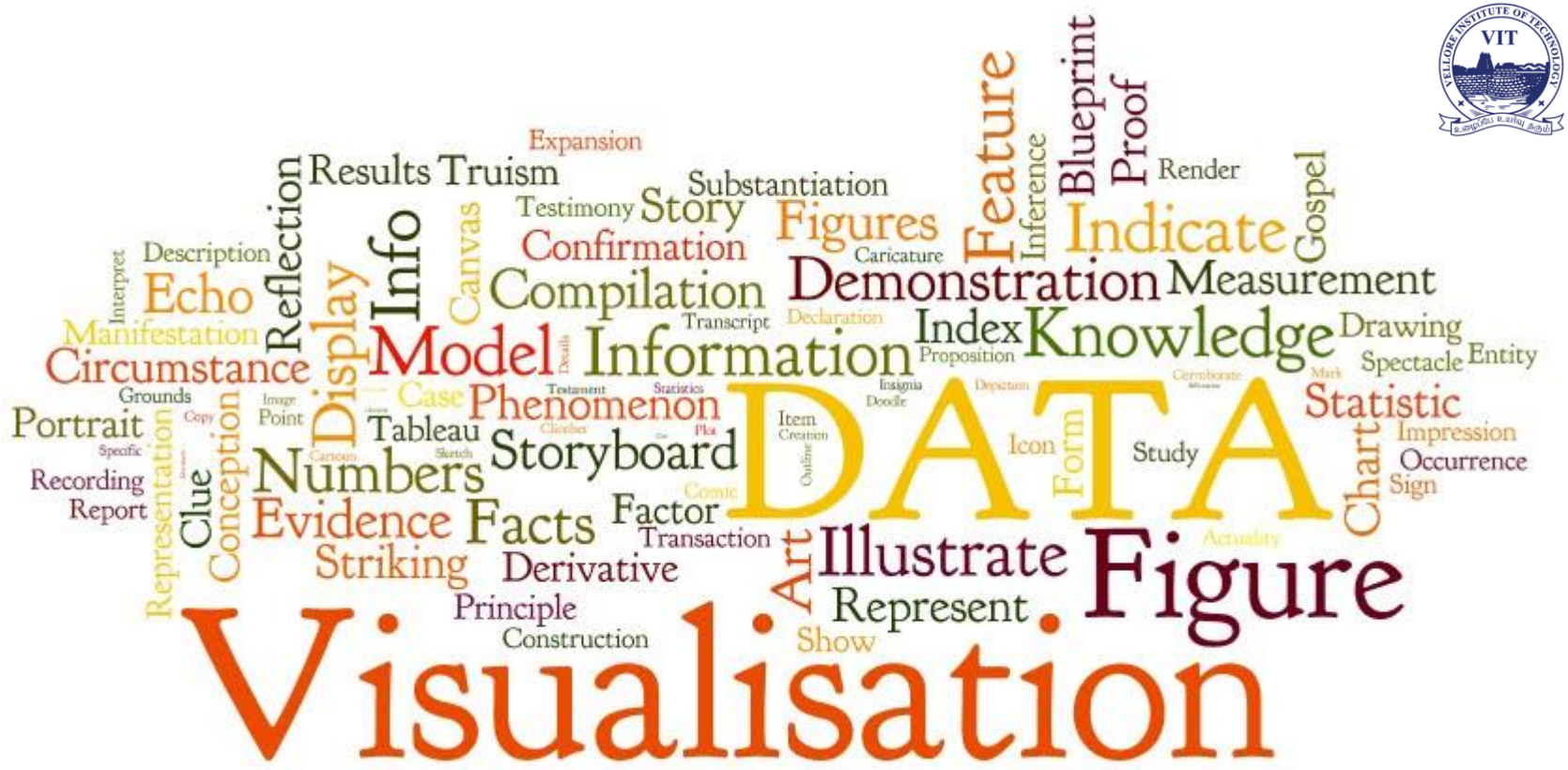


Introduction

Prof. Pattabiraman.V

SCSE, VIT, Chennai

Outline

- What is Data Visualization?
- Why do we need Data Visualization?
- Goals of Data Visualization
- Characteristics of effective graphical displays
- Different Types of Data
- Visual Variables

?

01:54

What is Data Visualization?

- Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context.
- Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

Why Data Visualization?

- Did you know that 25% of your brain power is connected to visual stimulus, and 70% of our sensory receptors are in our eyes?
- No wonder we "get the picture" faster when presenting information visually

Why Data Visualization?

- A picture is worth 1000 words.
- A picture can also be worth 1000 data points.
  - In 1973, the statistician Francis Anscombe demonstrated the importance of graphing data.
  - The Anscombe's Quartet shows how four sets of data with identical simple summary statistics can vary considerably when graphed.

Why Data Visualization?

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Anscombe's Quartet dataset

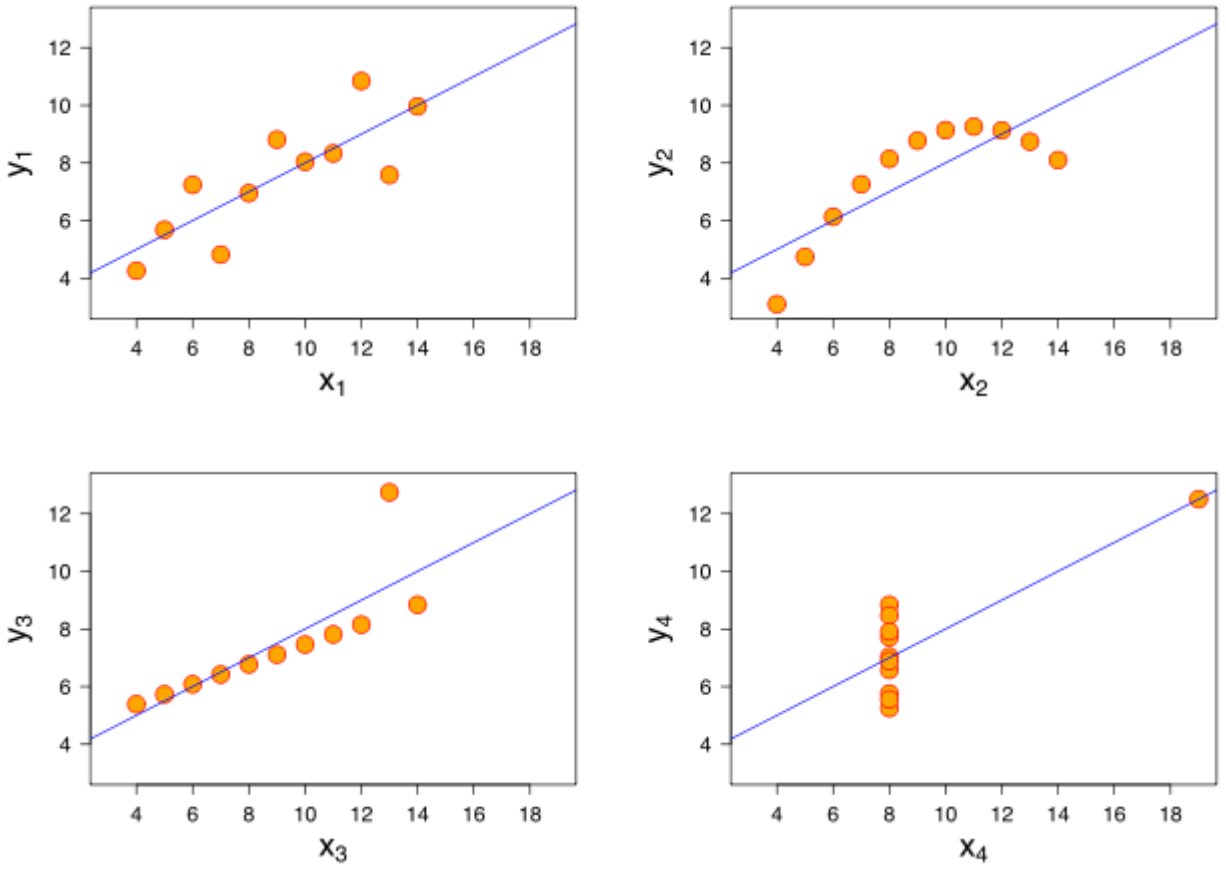
Why Data Visualization?

- Simple Summary Statistics of Anscombe's Quartet Data Table

Property	Value
Mean of x of each data set	9 (exact)
Variance of x in each data set	11 (exact)
Mean of y in each data set	7.50 (to 2 decimal places)
Variance of y in each data set	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each data set	0.816 (to 3 decimal places)
Linear regression line for each data set	y = 3.00 + 0.500x (to 2 and 3 decimal places, respectively)

Why Data Visualization?

- Graph of Anscombe's Quartet Data Table



Goals of Data Visualization

- A primary goal of data visualization is to communicate information clearly and efficiently via statistical graphics, plots and information graphics.
- Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message.

Goals of Data Visualization

- Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable.
- Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic follows the task.
- Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables

Characteristics of effective graphical displays

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else
- avoid distorting what the data has to say
- present many numbers in a small space

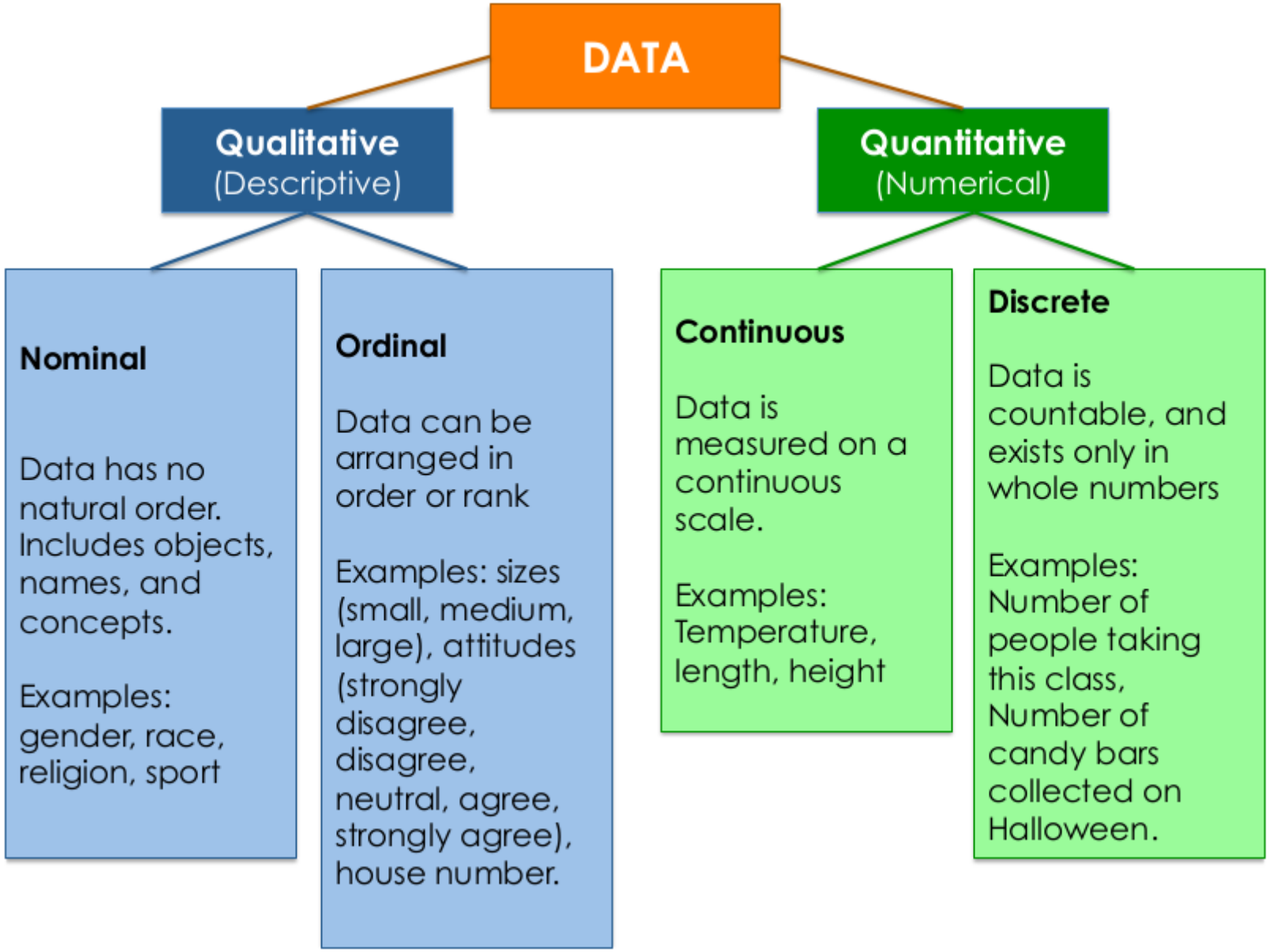
Characteristics of effective graphical displays

- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal descriptions of a data set

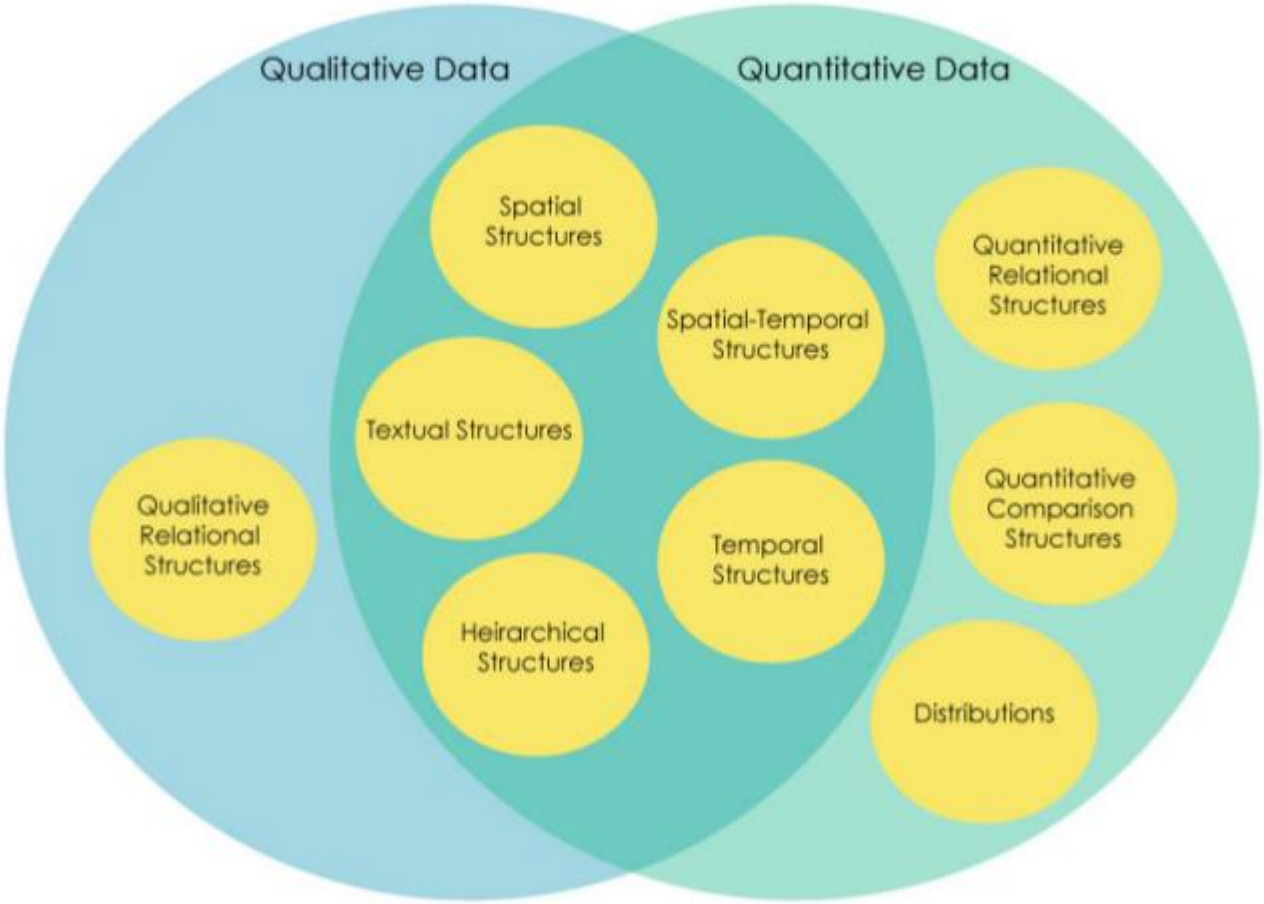
Characteristics of effective graphical displays

- The greatest value of a picture is when it forces us to notice what we never expected to see. - John Tukey

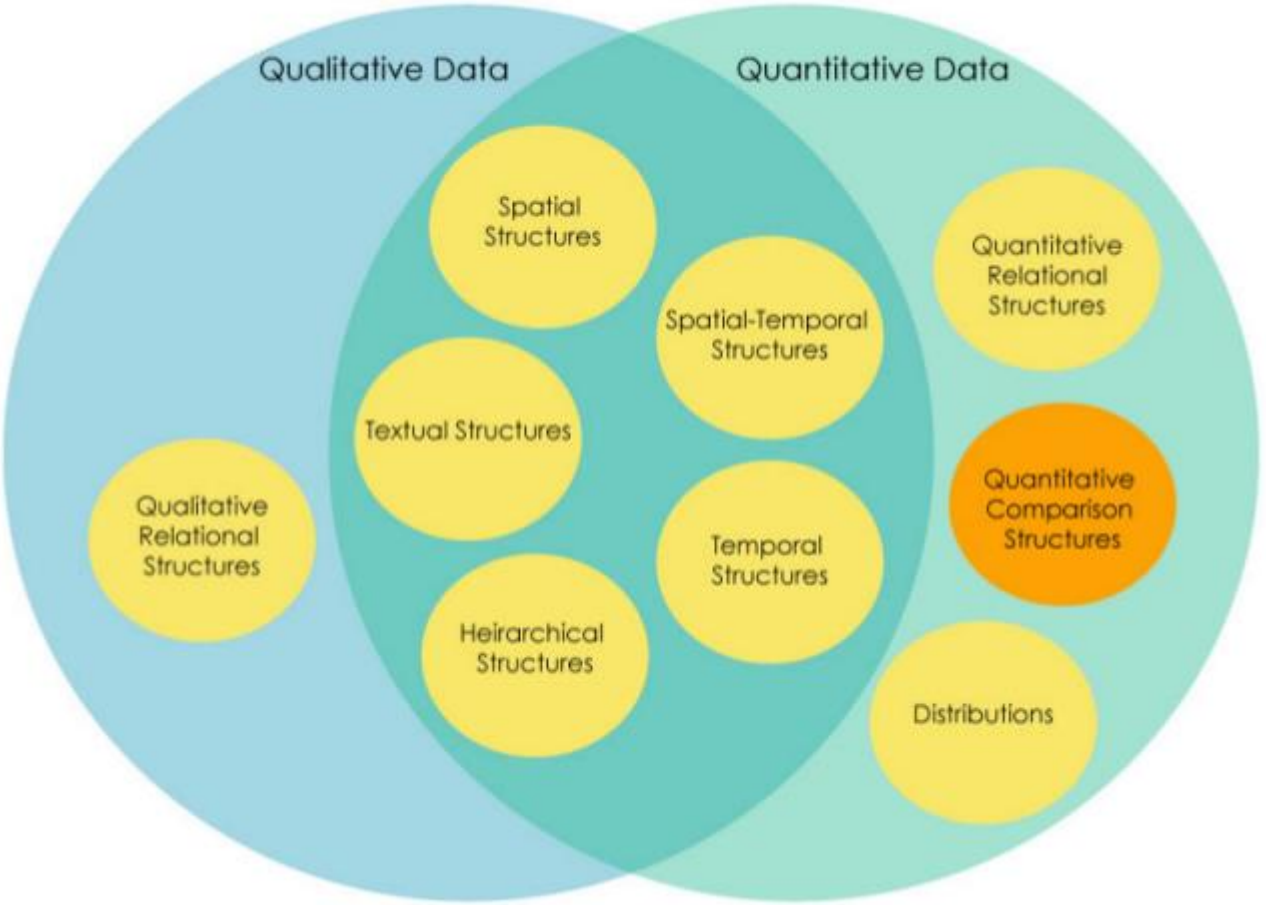
Different Types of Data



Different Types of Data

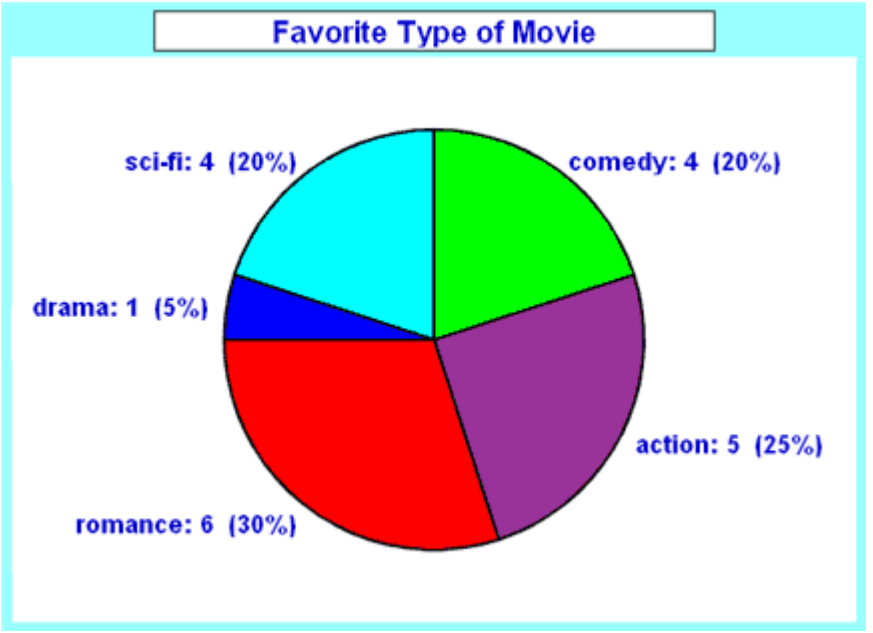


Different Types of Data

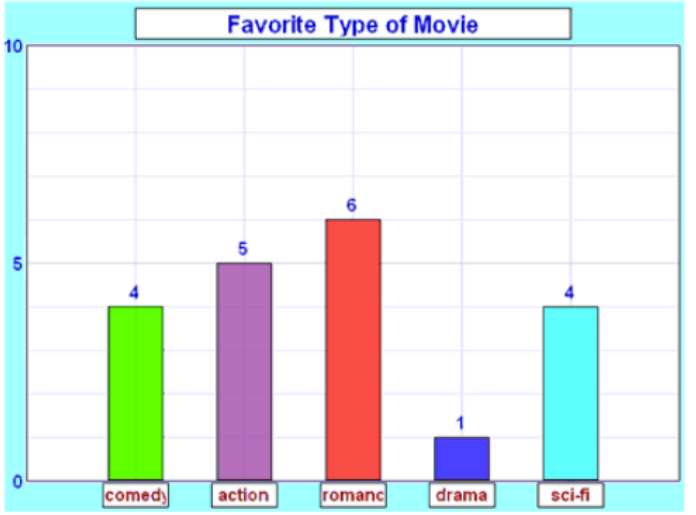


Quantitative Comparison

- Use sparingly
- No more than six components.
- Not useful when values of each component are similar



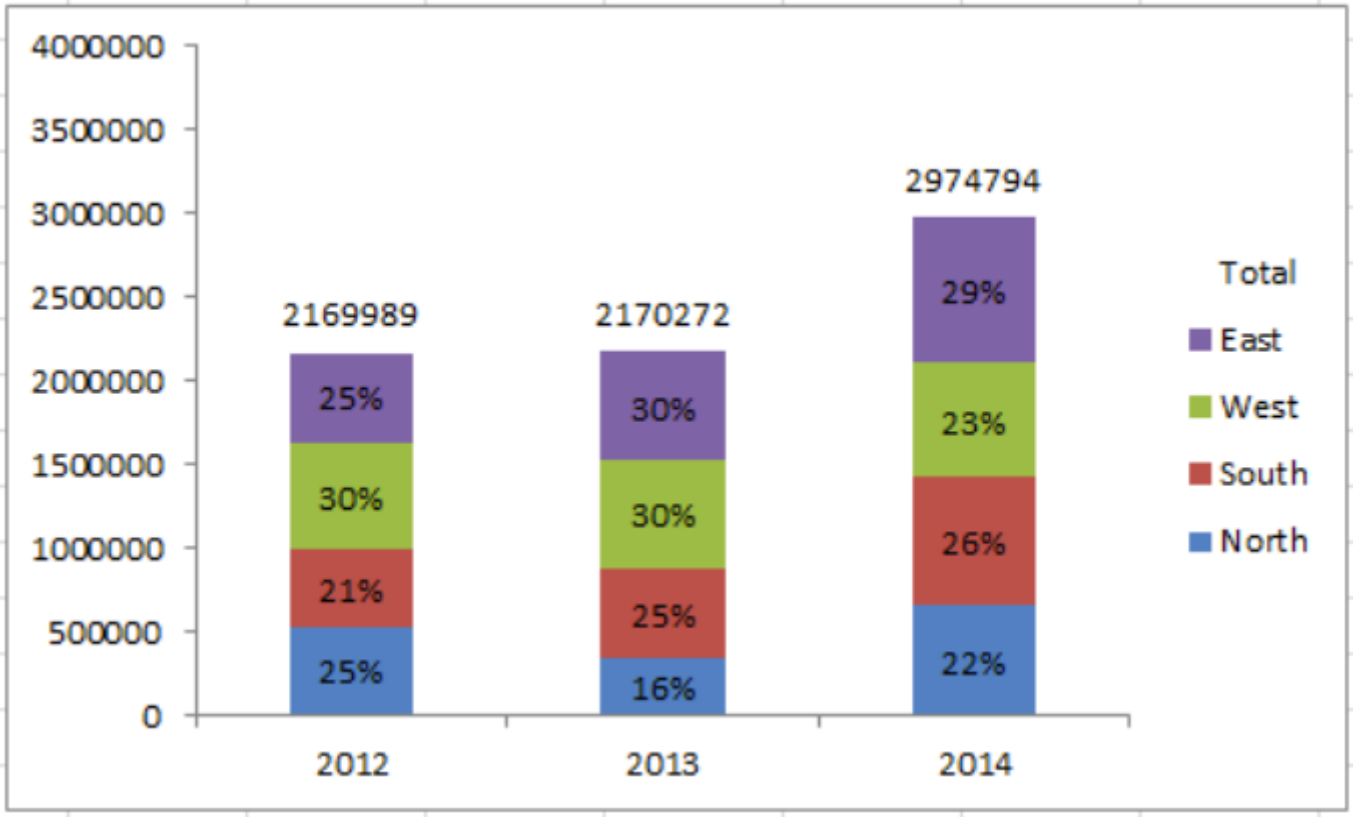
Quantitative Comparison



- Bar graph
- Best for comparing categories.
- Best Practices
  - Make bars and columns wider than the space between them.
  - Do not allow grid lines to pass through columns or bars.
  - Use a single font type on a graph.

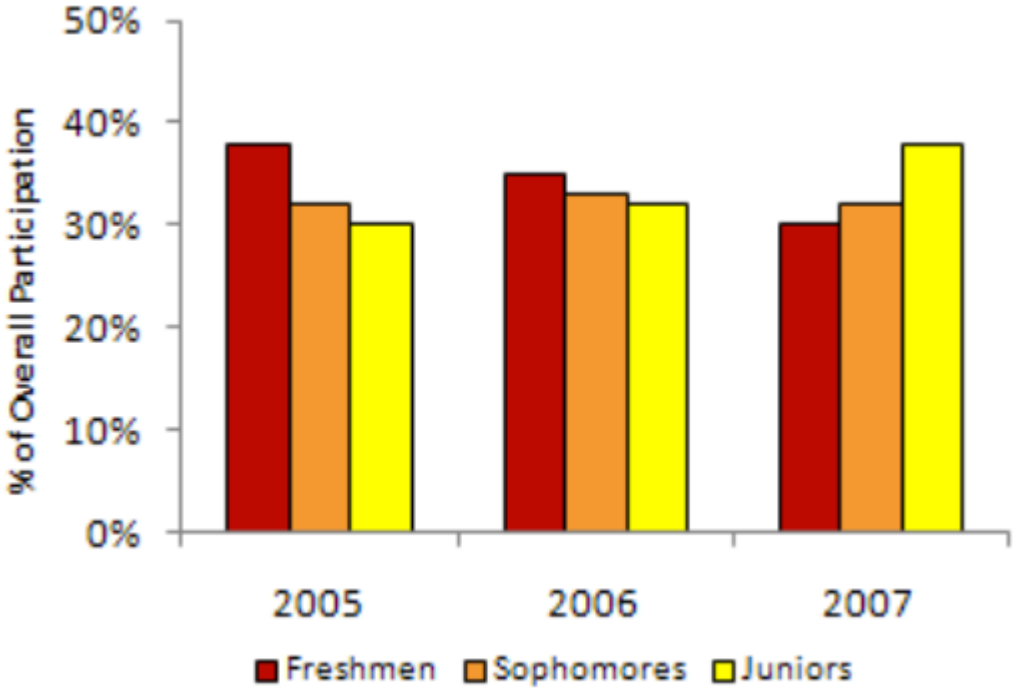
Quantitative Comparison

- Stacked bar graph



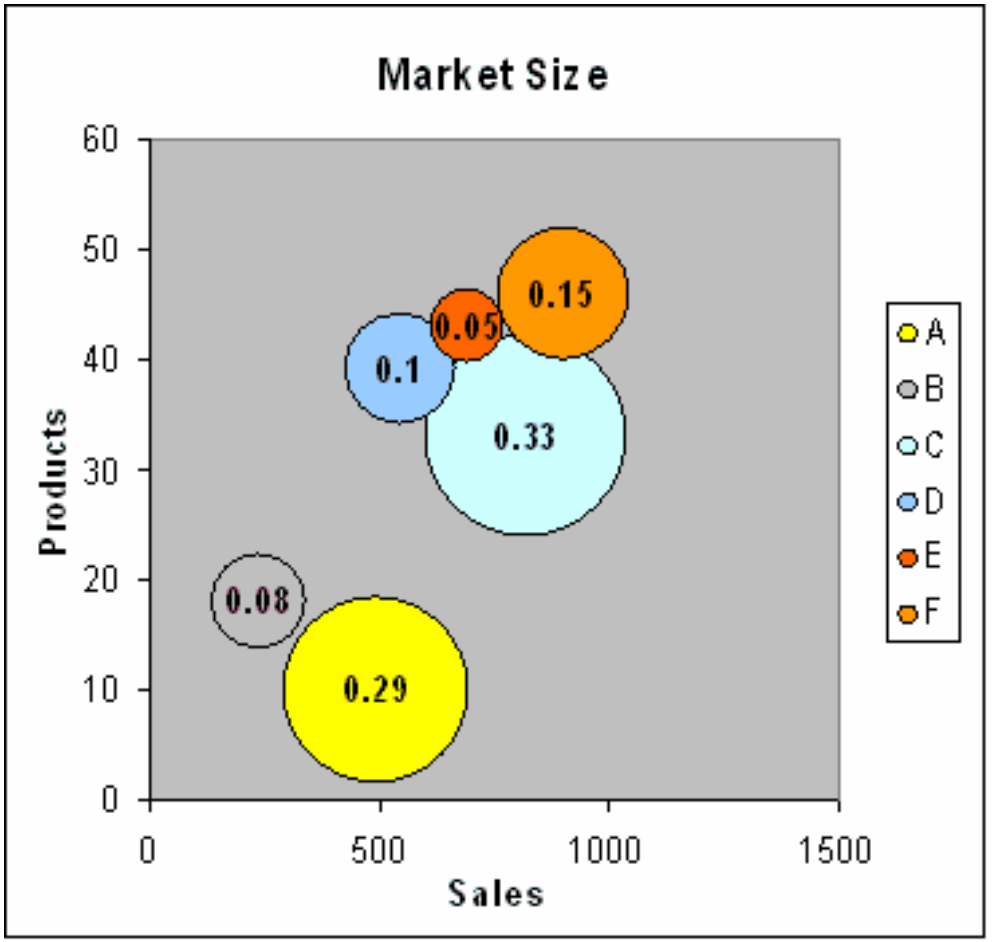
Quantitative Comparison

- Group Bar Plot or Clustered bar graph



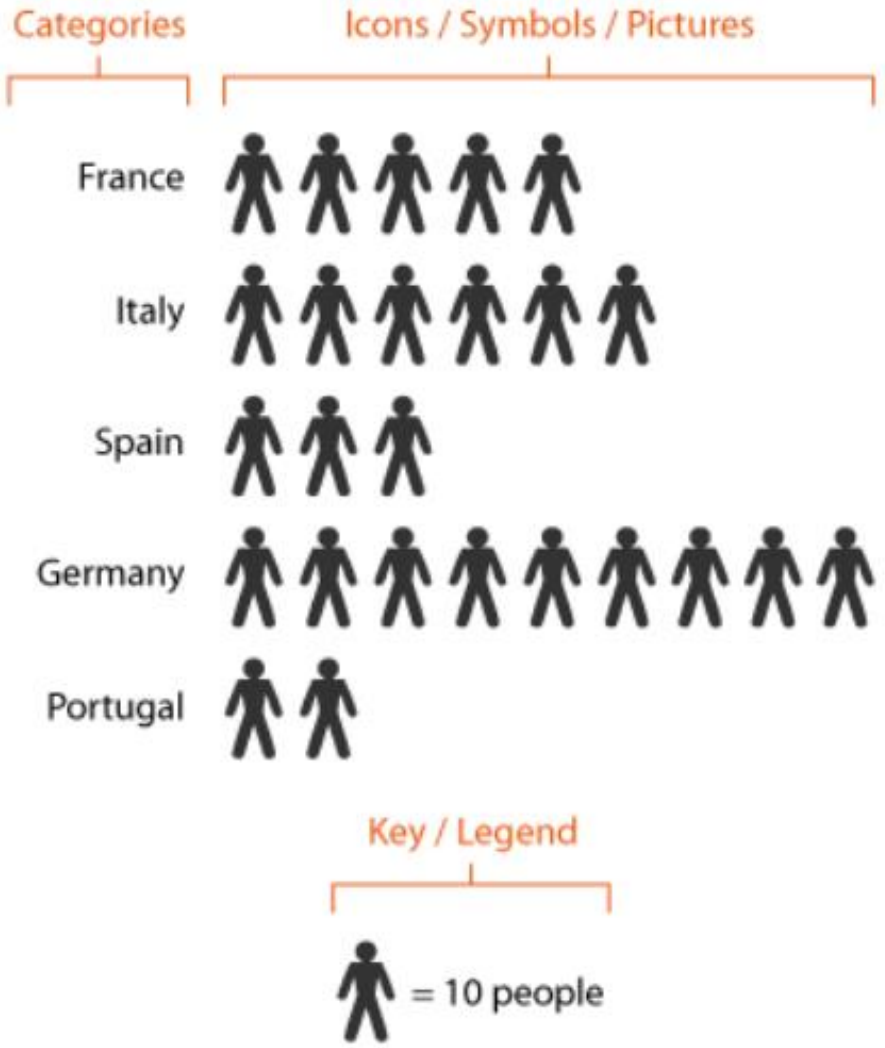
Quantitative Comparison

- Bubble Charts

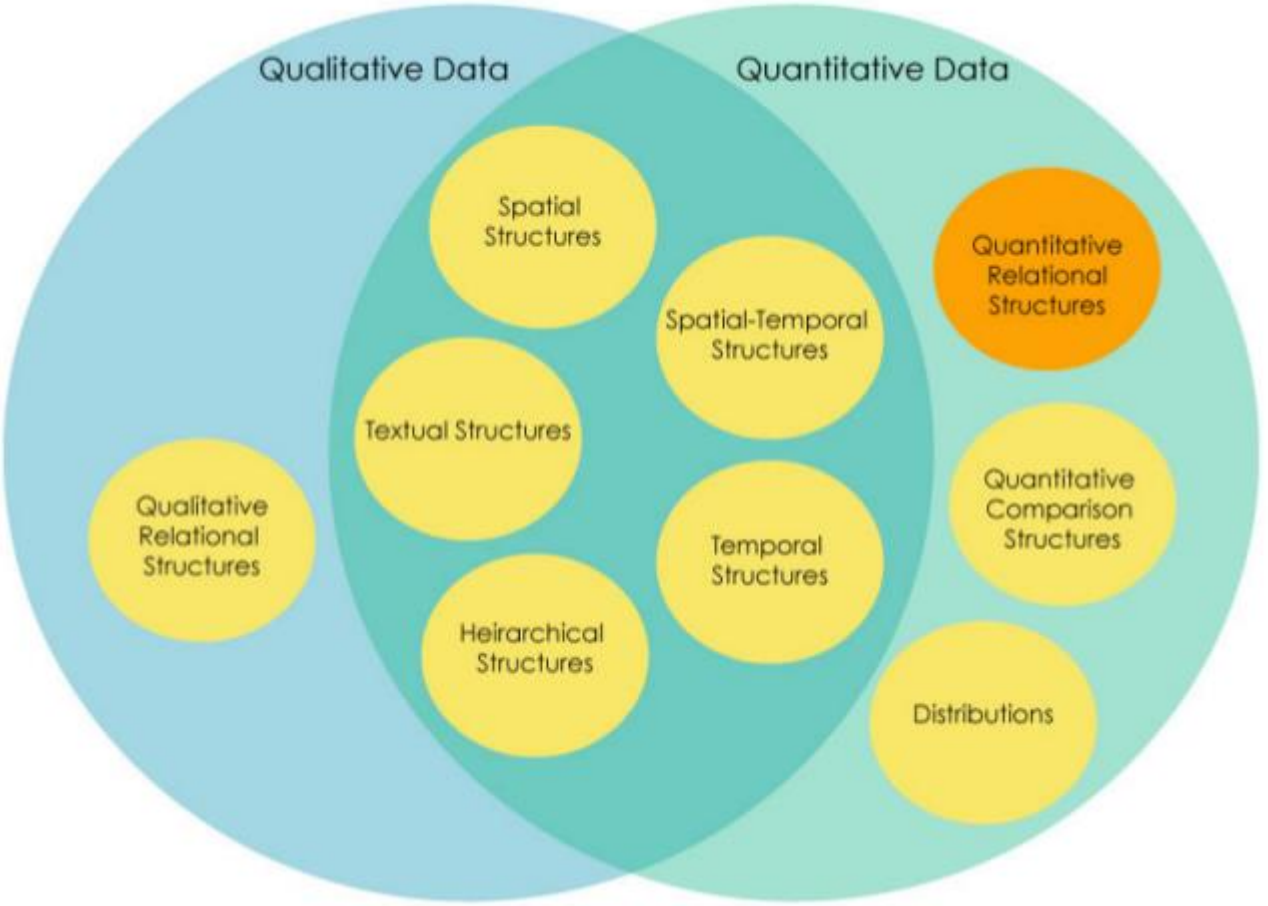


Quantitative Comparison

- Pictogram Chart
  - For discrete data

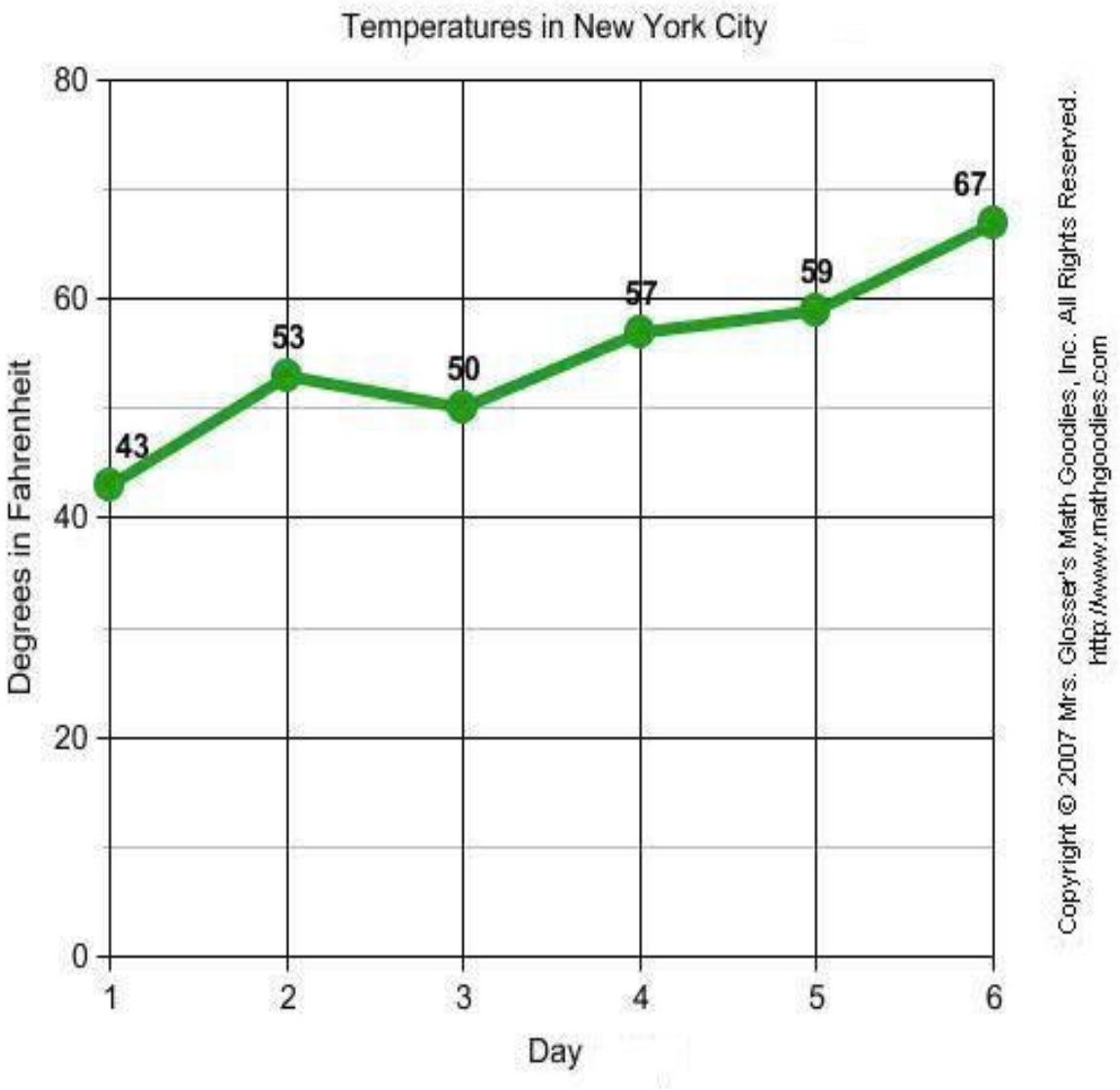


Different Types of Data



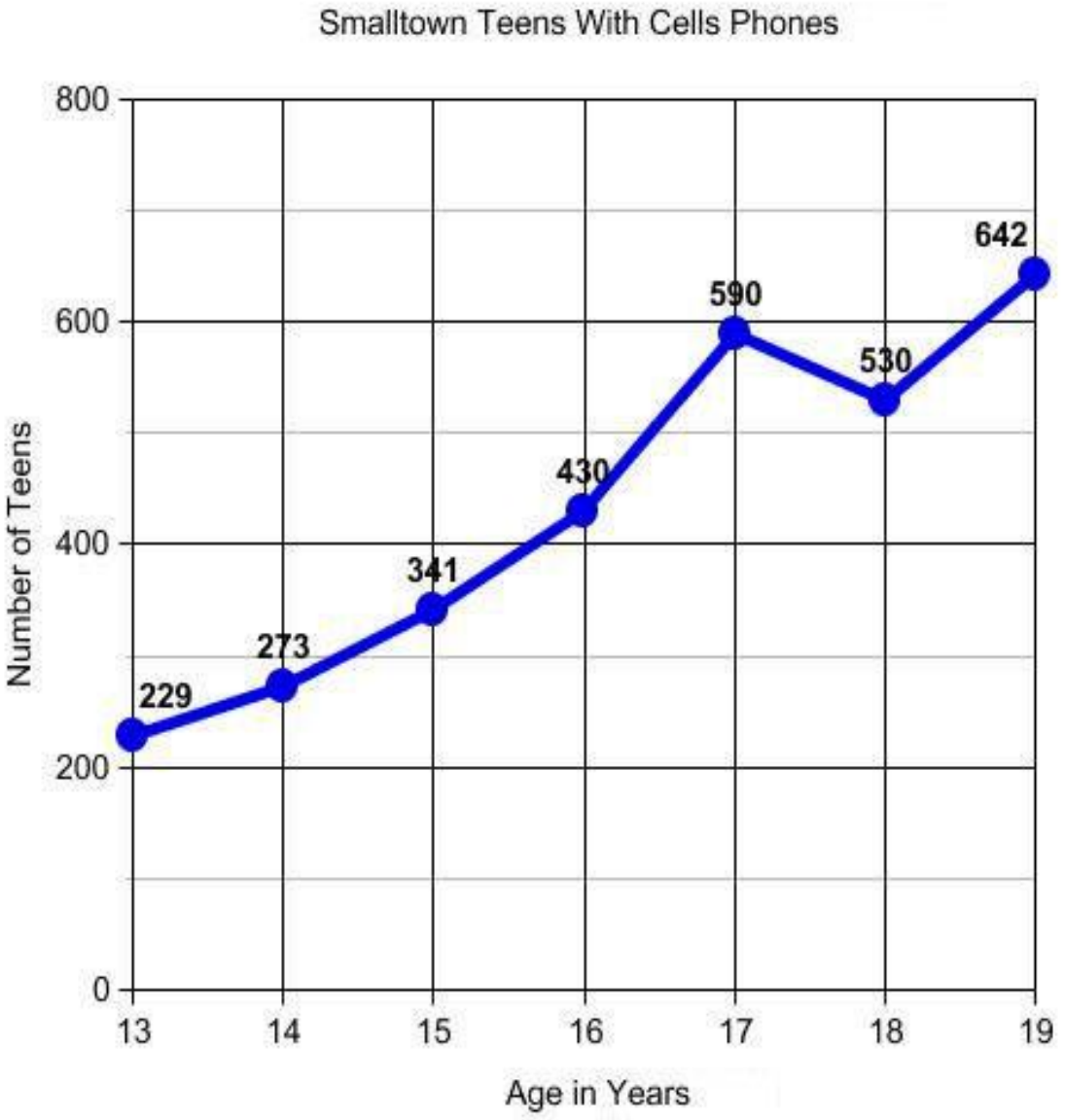
Quantitative Relational

- Line Charts
  - For identifying
  - trends.



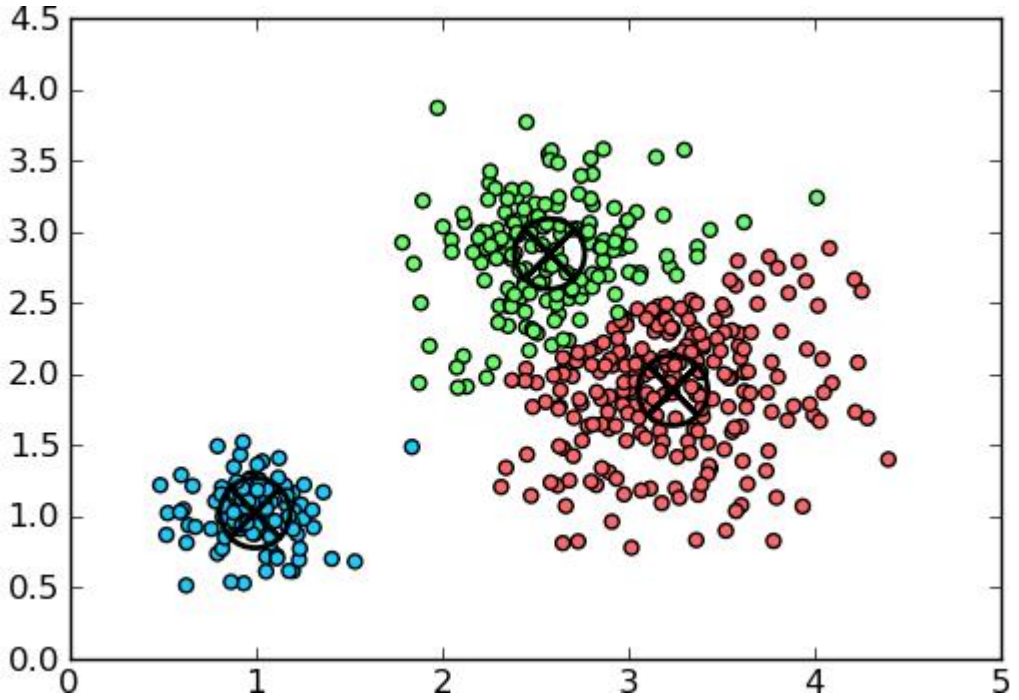
Quantitative Relational

- Line Charts
  - For identifying
  - trends



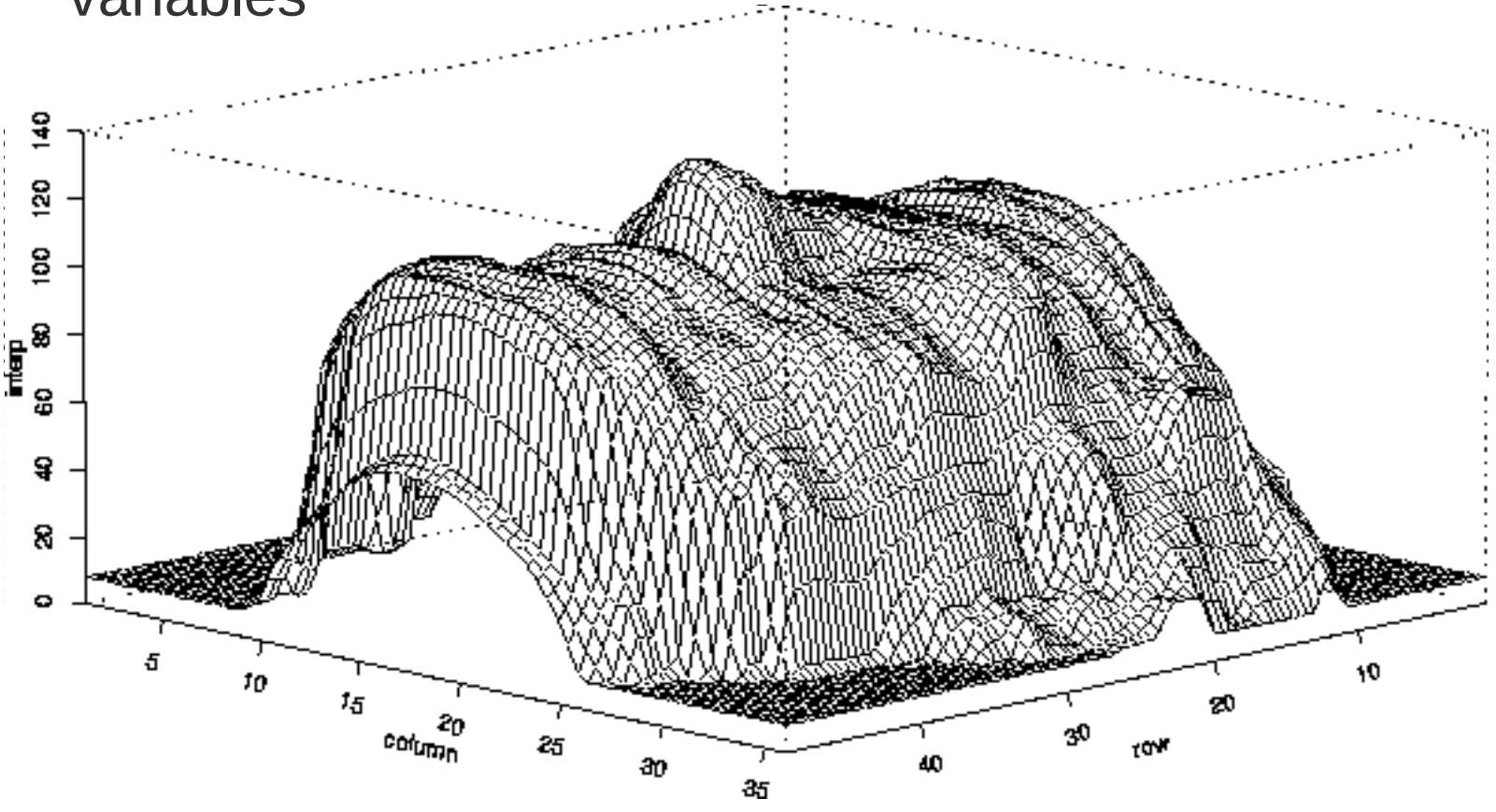
Quantitative Relational

Scatter Plots- For testing and identifying relationships, and statistical correlations



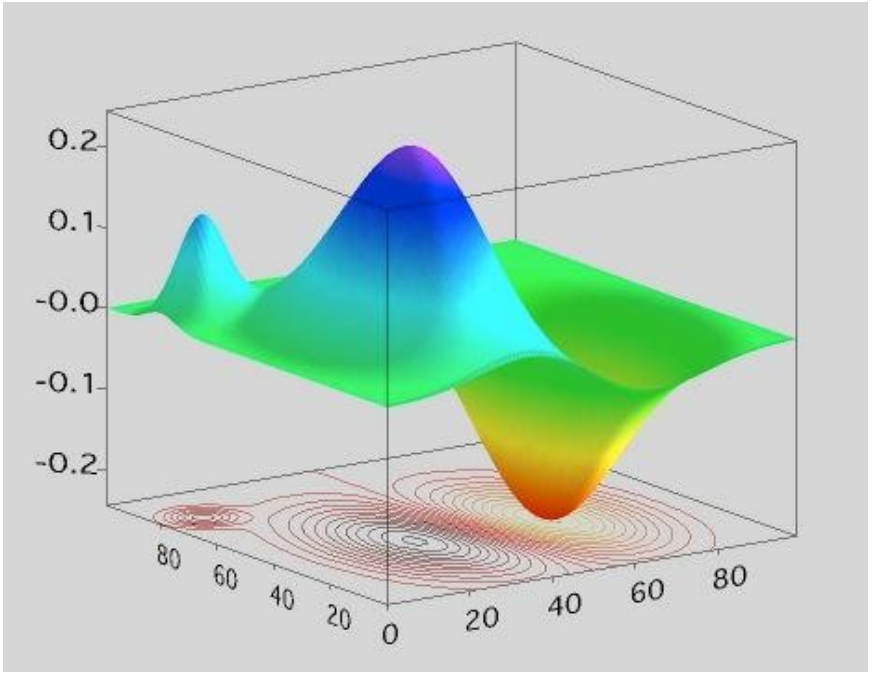
Quantitative Relational

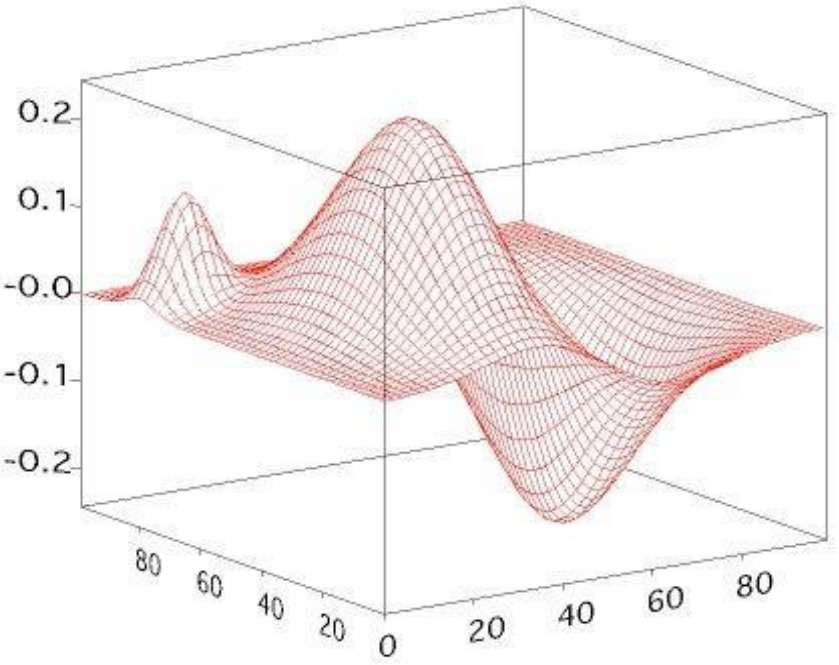
- Surface plots
  - Topography, Density Functions that have two dependent variables



Quantitative Relational

- Surface plots
  - Topography, Density Functions that have two dependent variables



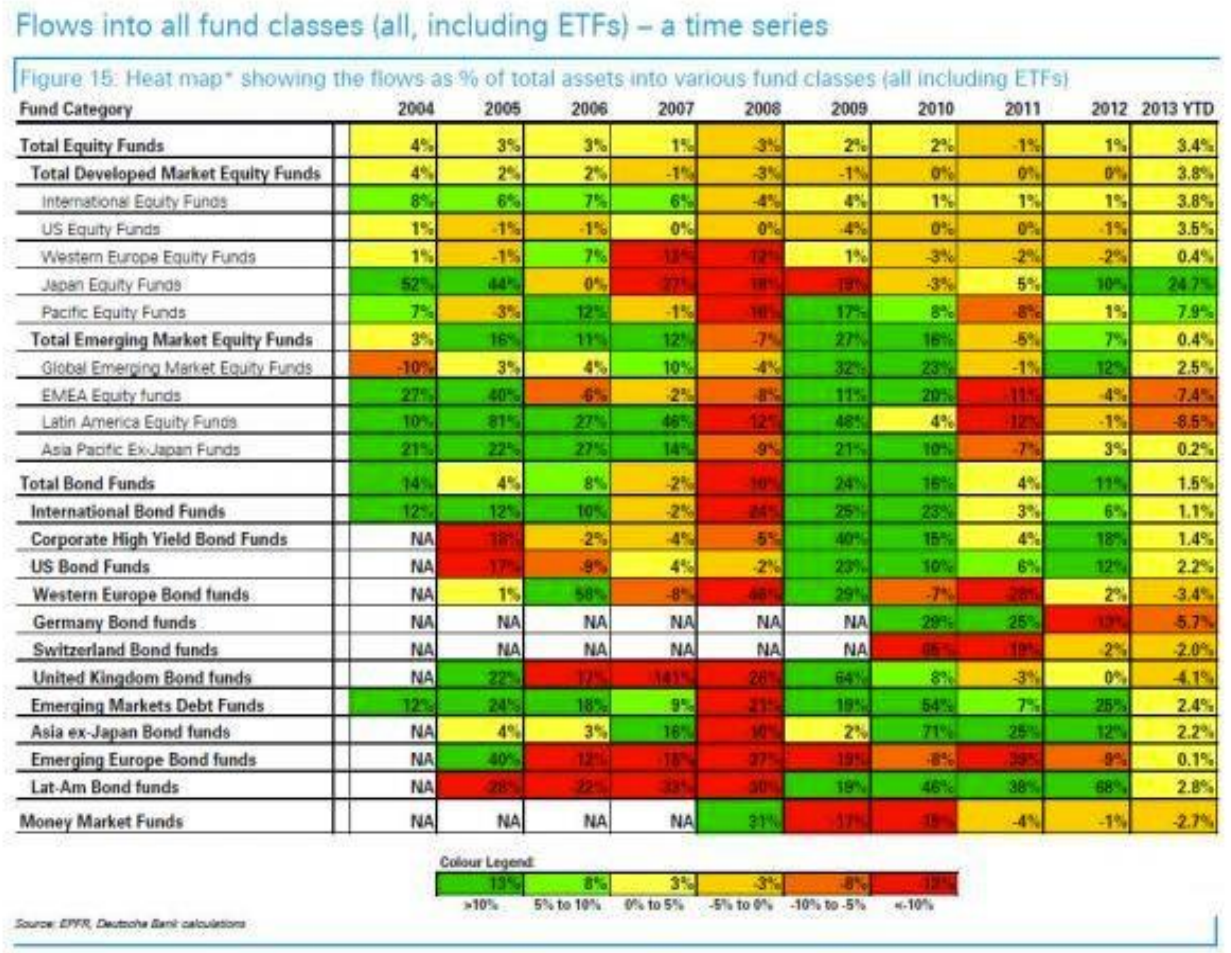


Quantitative Relational

- Heat Map

Quantitative Relational

- Co-occurrence matrix/ Heat map



Flows into all fund classes (all, including ETFs) – a time series

Figure 15: Heat map* showing the flows as % of total assets into various fund classes (all including ETFs)

Fund Category	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013 YTD
Total Equity Funds	4%	3%	3%	1%	-3%	2%	2%	-1%	1%	3.4%
Total Developed Market Equity Funds	4%	2%	2%	-1%	-3%	-1%	0%	0%	0%	3.8%
International Equity Funds	8%	6%	7%	6%	-4%	4%	1%	1%	1%	3.8%
US Equity Funds	1%	-1%	-1%	0%	0%	-4%	0%	0%	-1%	3.5%
Western Europe Equity Funds	1%	-1%	7%	13%	12%	1%	-3%	-2%	-2%	0.4%
Japan Equity Funds	52%	44%	0%	27%	18%	19%	-3%	5%	10%	24.7%
Pacific Equity Funds	7%	-3%	12%	-1%	16%	17%	8%	-8%	1%	7.9%
Total Emerging Market Equity Funds	3%	16%	11%	12%	-7%	27%	16%	-5%	7%	0.4%
Global Emerging Market Equity Funds	-10%	3%	4%	10%	-4%	32%	23%	-1%	12%	2.5%
EMEA Equity funds	27%	40%	-6%	-2%	-8%	11%	20%	-11%	-4%	-7.4%
Latin America Equity Funds	10%	81%	27%	48%	12%	48%	4%	-12%	-1%	-8.5%
Asia Pacific Ex-Japan Funds	21%	22%	27%	14%	-9%	21%	10%	-7%	3%	0.2%
Total Bond Funds	14%	4%	8%	-2%	10%	24%	16%	4%	11%	1.5%
International Bond Funds	12%	12%	10%	-2%	6%	25%	23%	3%	6%	1.1%
Corporate High Yield Bond Funds	NA	18%	-2%	-4%	-5%	40%	15%	4%	18%	1.4%
US Bond Funds	NA	17%	-9%	4%	-2%	23%	10%	6%	12%	2.2%
Western Europe Bond funds	NA	1%	58%	-8%	46%	29%	-7%	-28%	2%	-3.4%
Germany Bond funds	NA	NA	NA	NA	NA	NA	29%	25%	13%	-5.7%
Switzerland Bond funds	NA	NA	NA	NA	NA	NA	6%	19%	-2%	-2.0%
United Kingdom Bond funds	NA	22%	17%	141%	26%	64%	8%	-3%	0%	-4.1%
Emerging Markets Debt Funds	12%	24%	18%	9%	21%	19%	54%	7%	25%	2.4%
Asia ex-Japan Bond funds	NA	4%	3%	16%	10%	2%	71%	25%	12%	2.2%
Emerging Europe Bond funds	NA	40%	12%	18%	37%	19%	-8%	38%	-8%	0.1%
Lat-Am Bond funds	NA	28%	22%	33%	30%	19%	46%	38%	68%	2.8%
Money Market Funds	NA	NA	NA	NA	31%	17%	19%	-4%	-1%	-2.7%

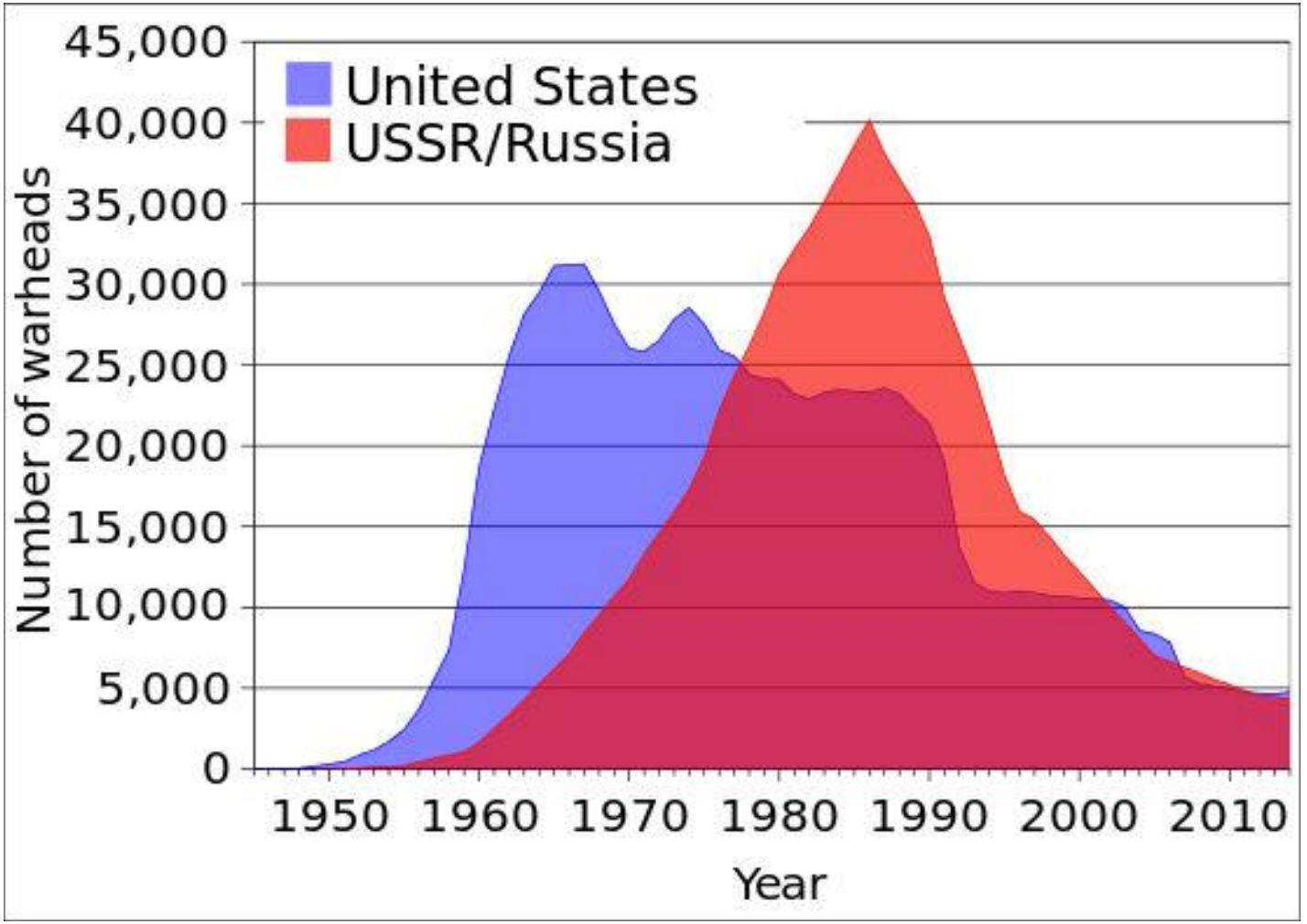
Colour Legend:

>10%	5% to 10%	0% to 5%	-5% to 0%	-10% to -5%	<-10%

Source: EPFR, Deutsche Bank calculations

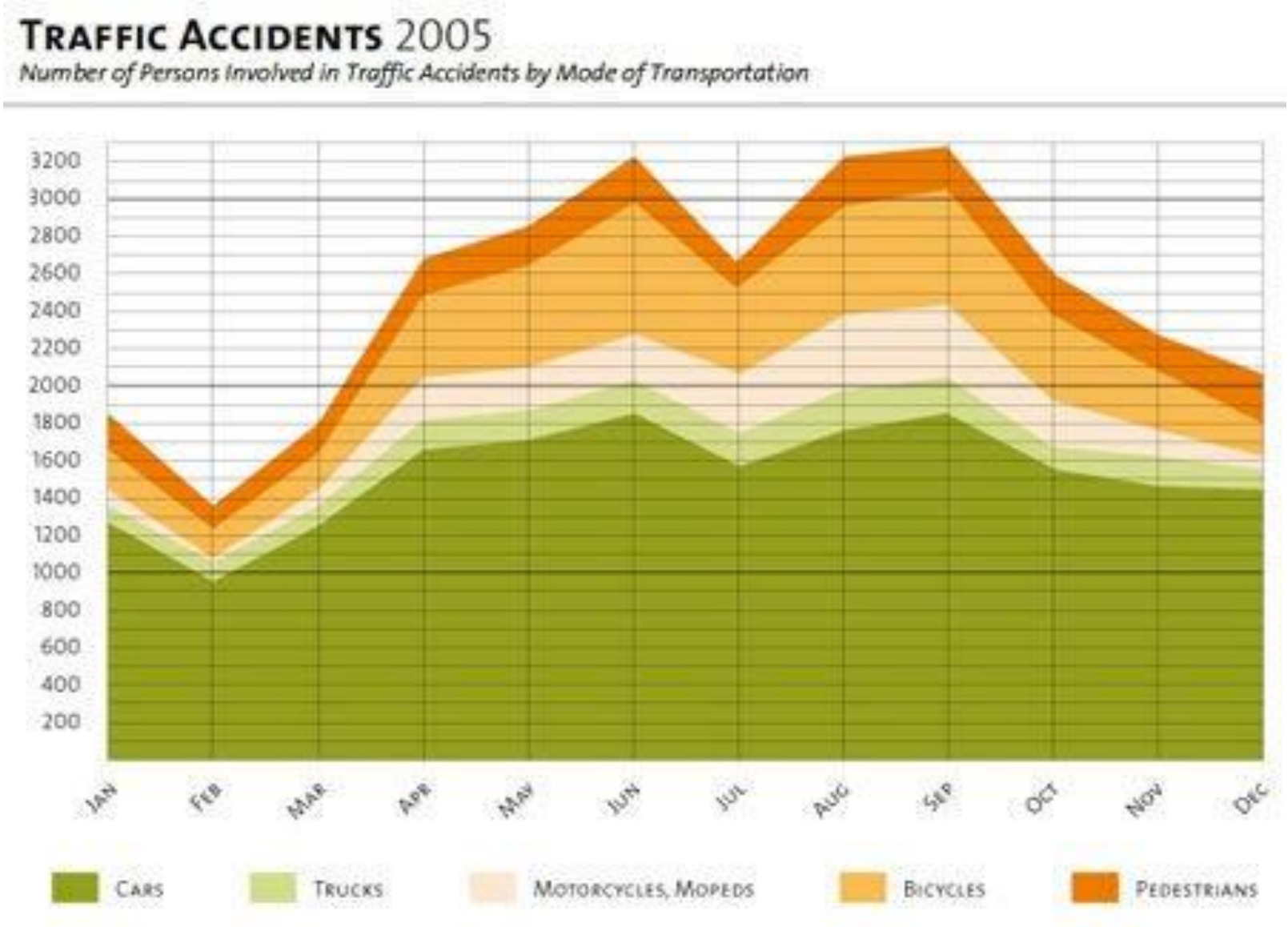
Quantitative Relational and Comparison

- Area Graph

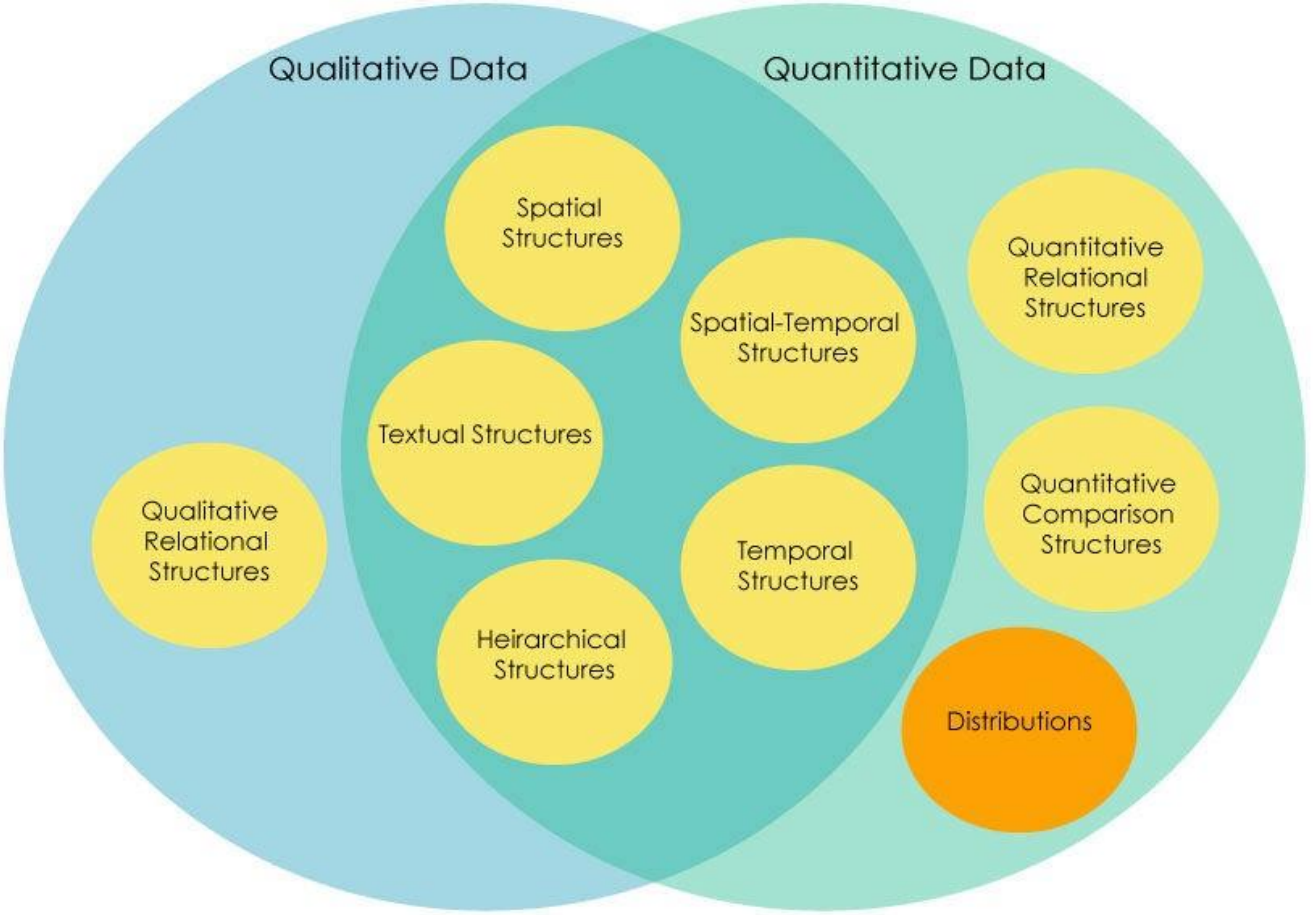


Quantitative Relational and Comparison

Stacked Area Graph

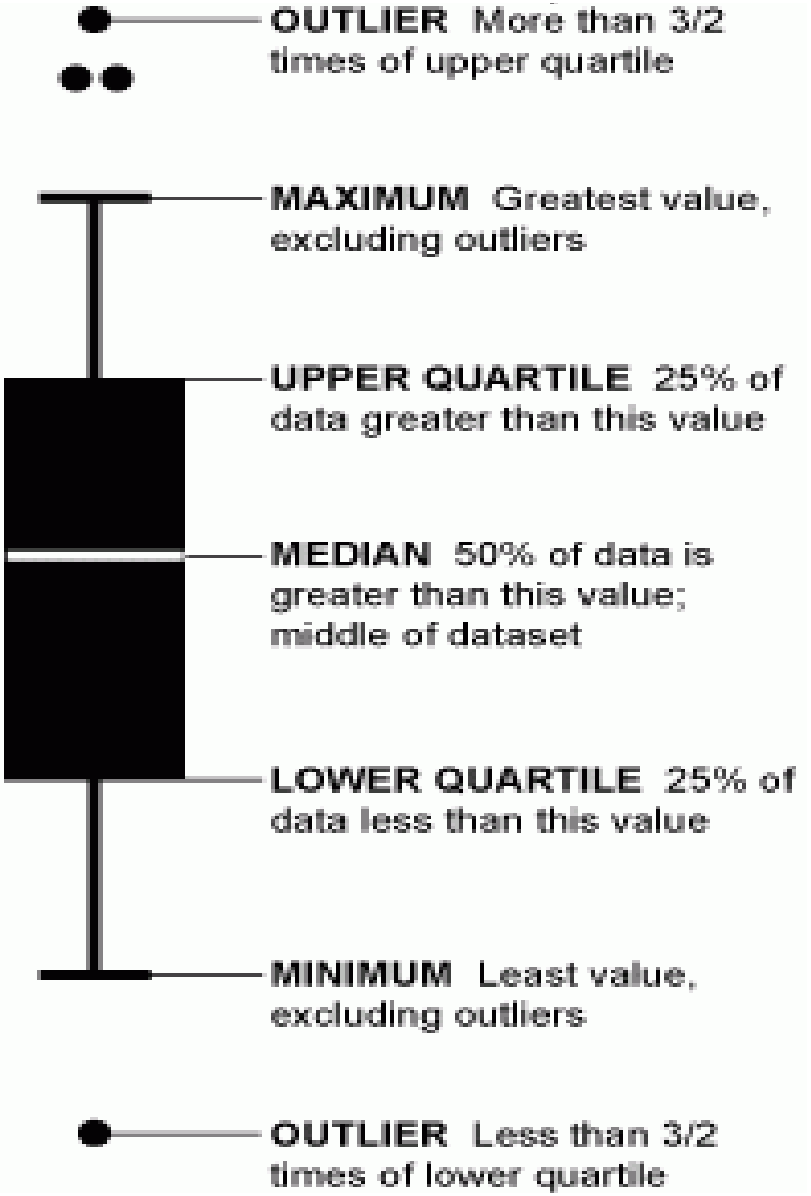


Different Types of Data

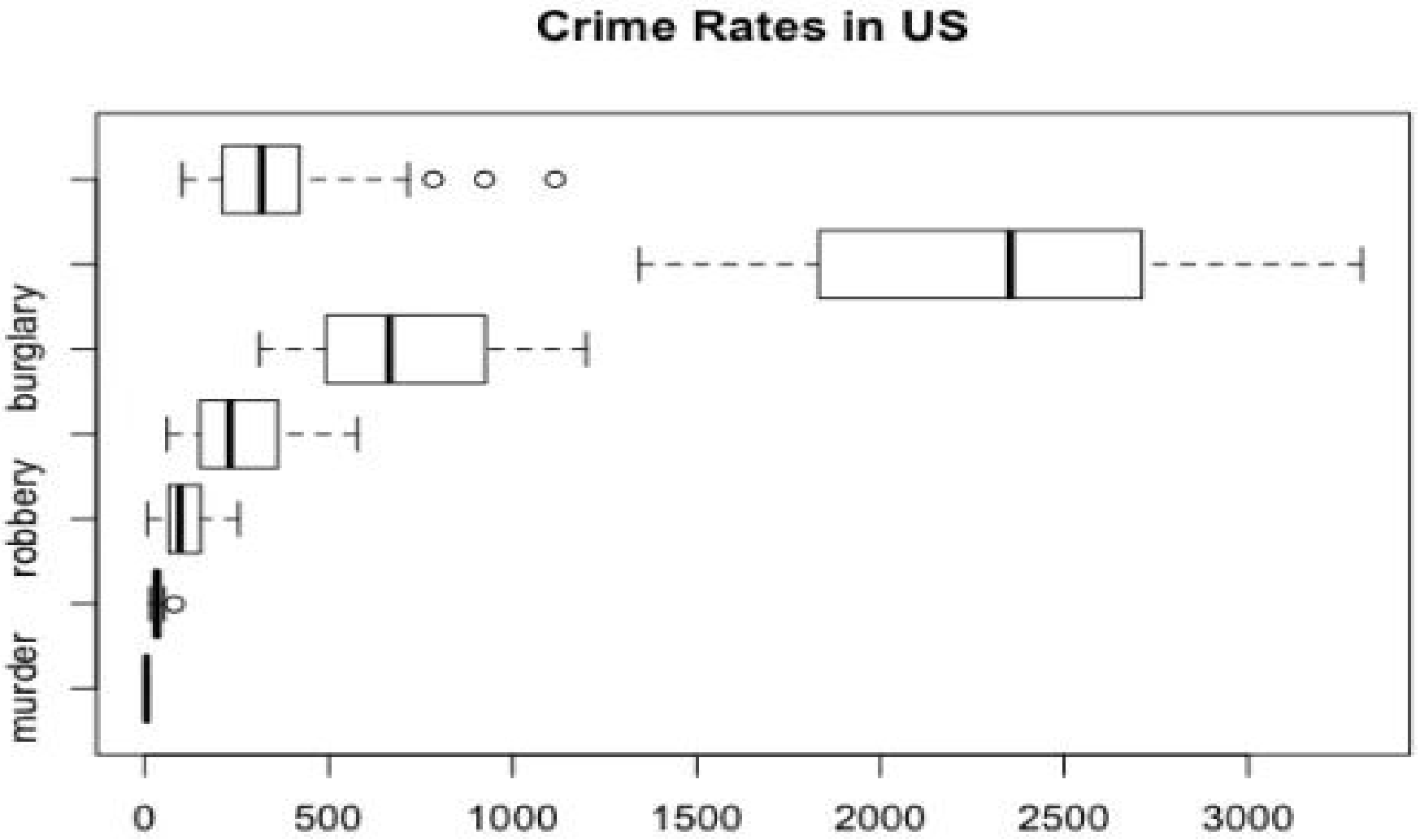


Distributions

Box and Whisker Plot

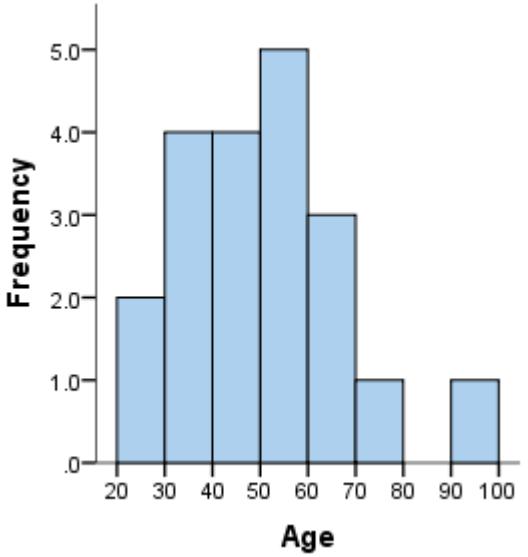


Distributions

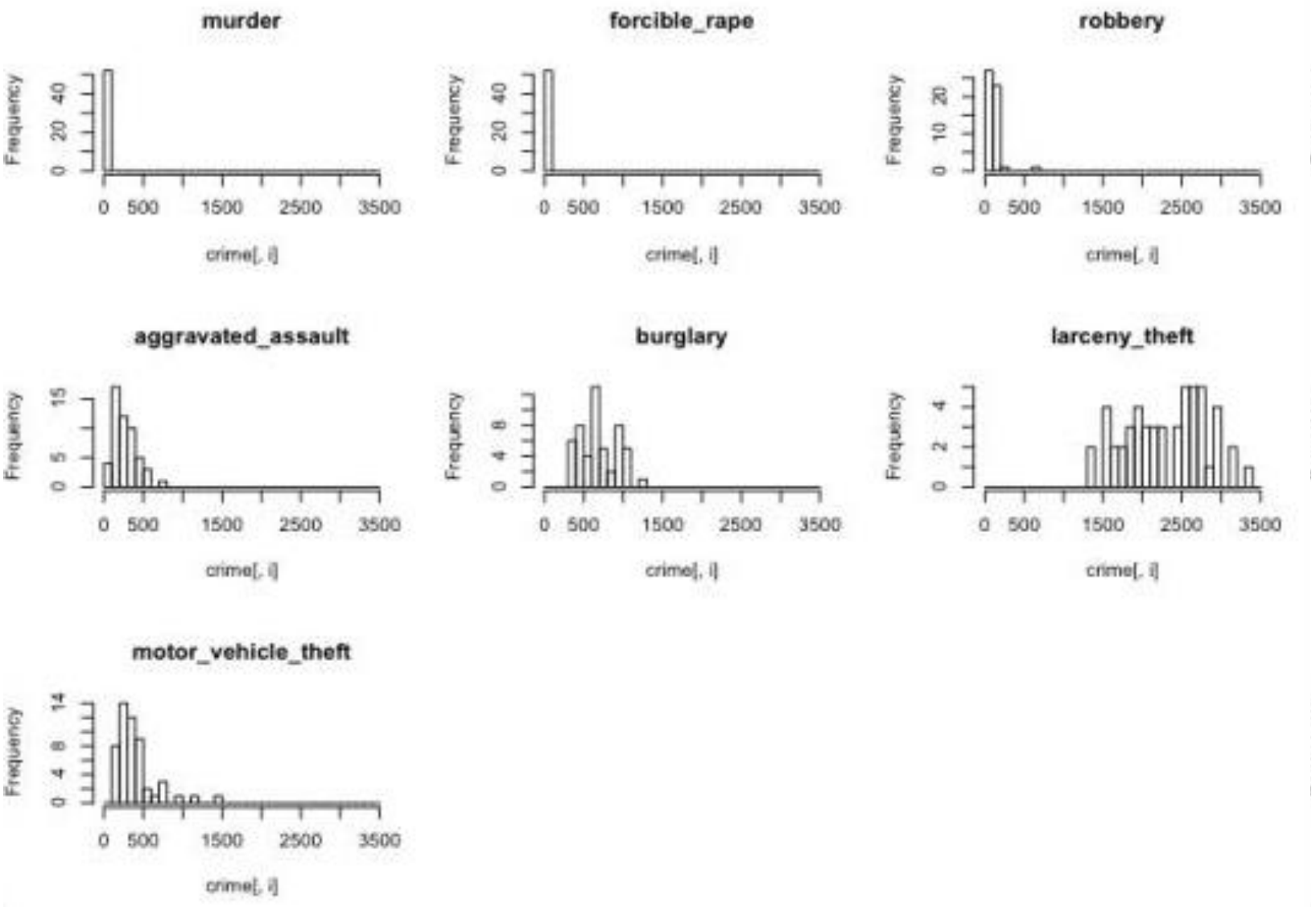


Distributions

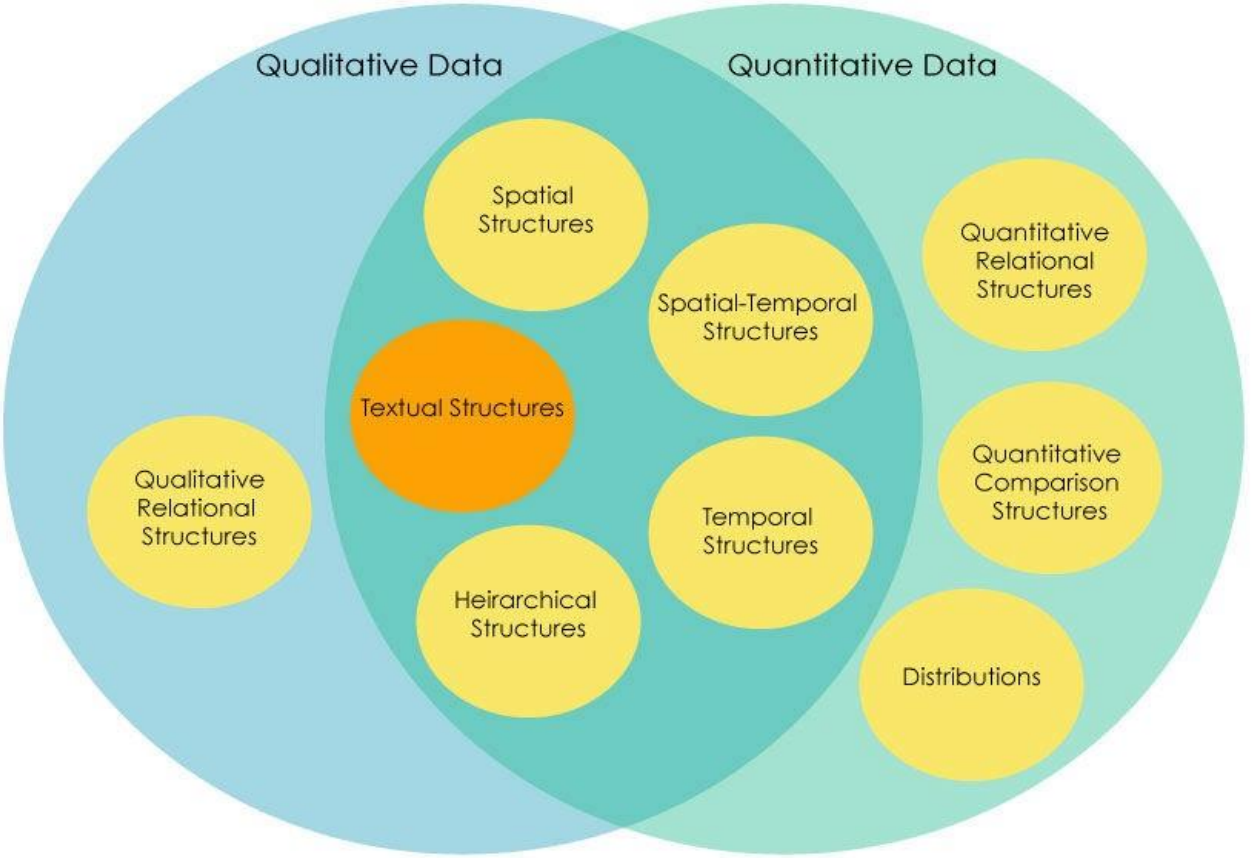
- Histograms
  - A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data.



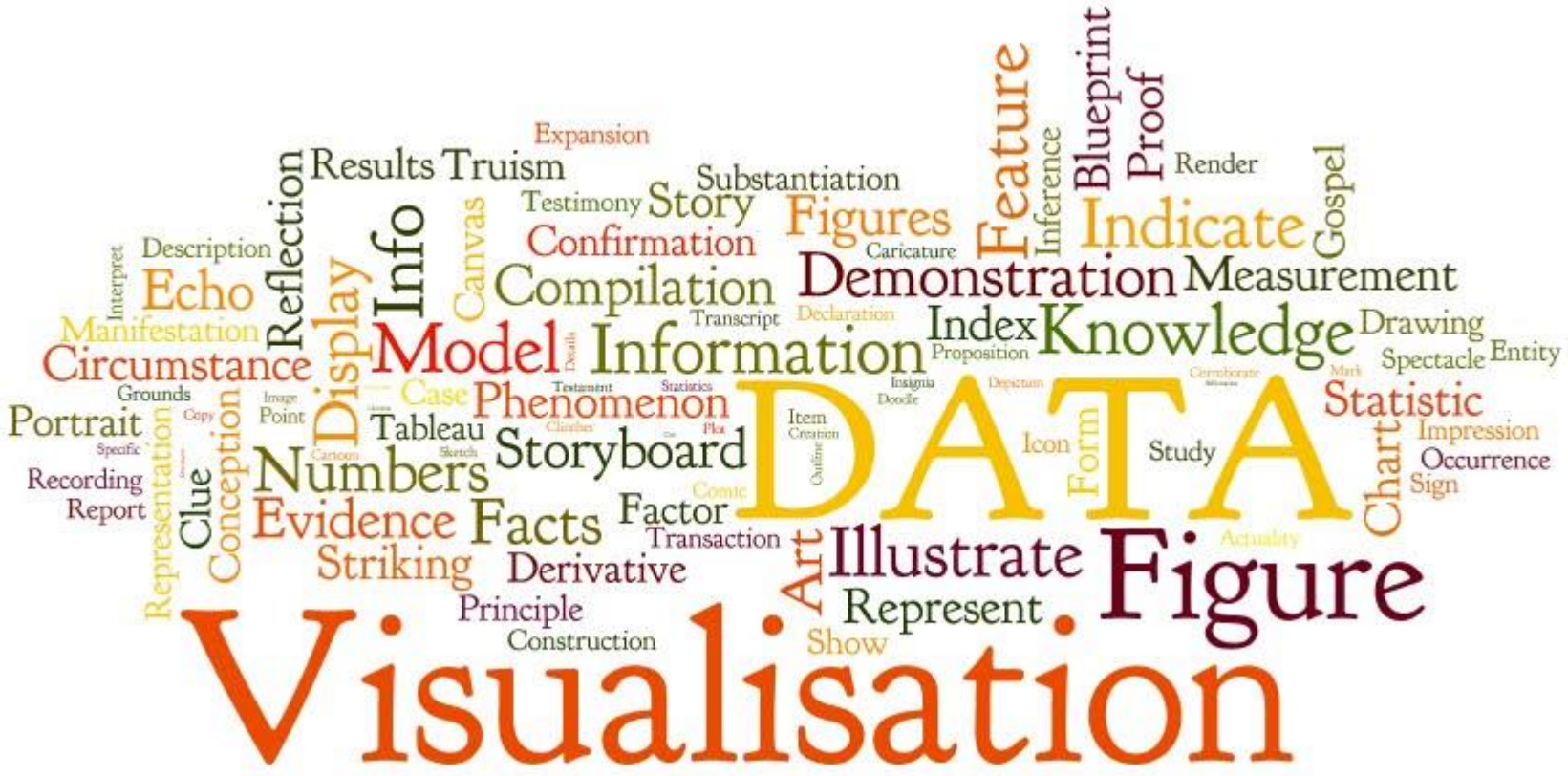
Distributions



Different Types of Data



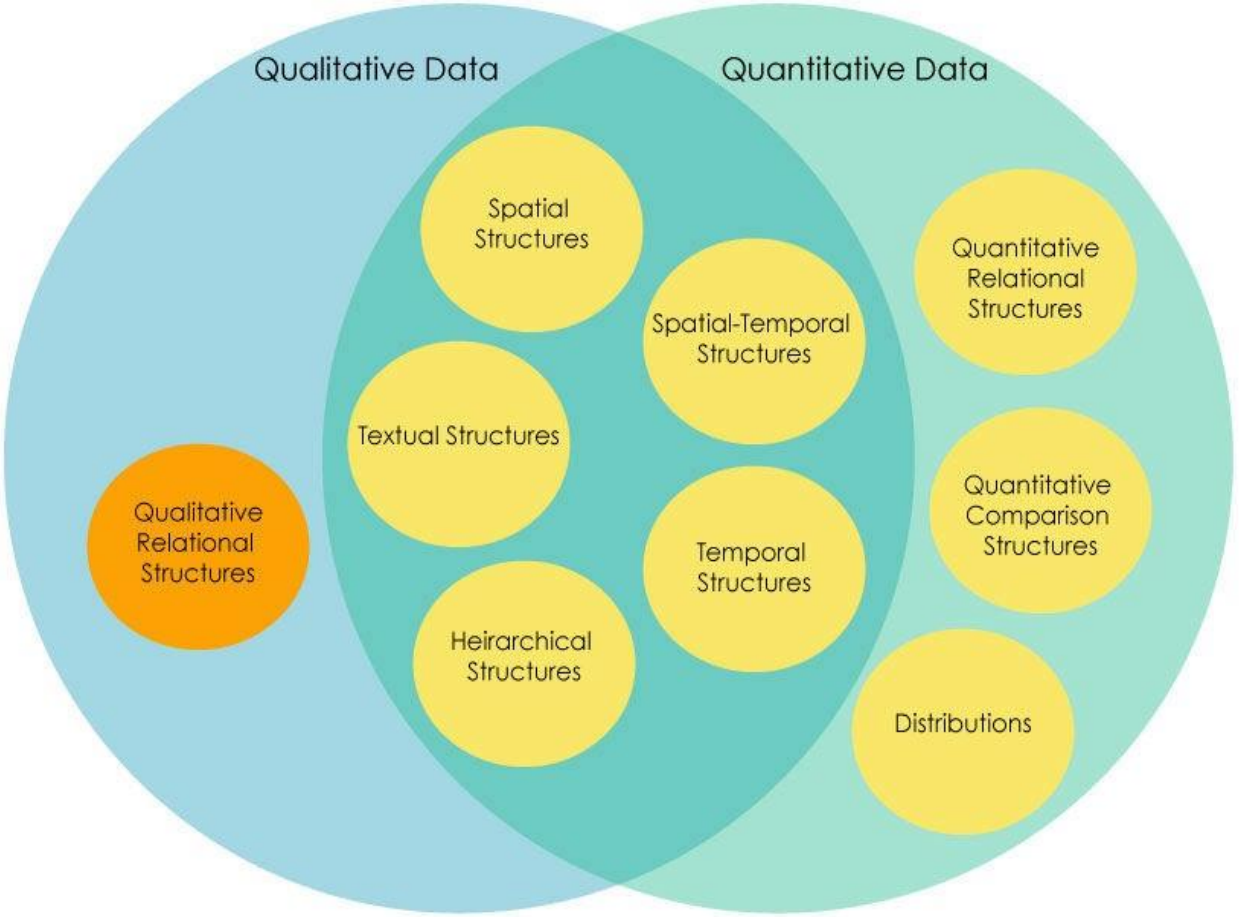
Qualitative Data: Textual Structures



Qualitative Data: Textual Structures

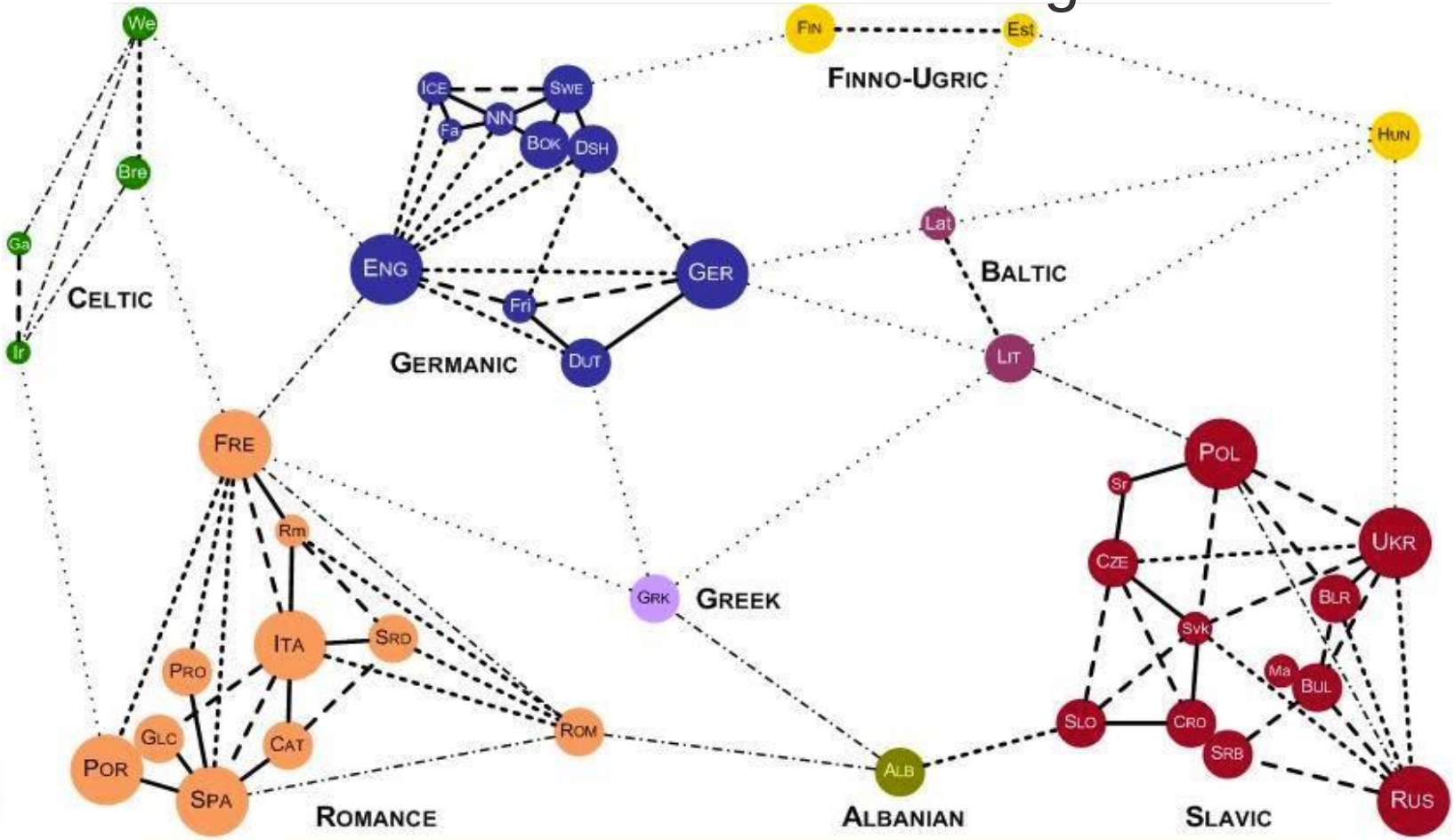
- Word Tree

Different Types of Data



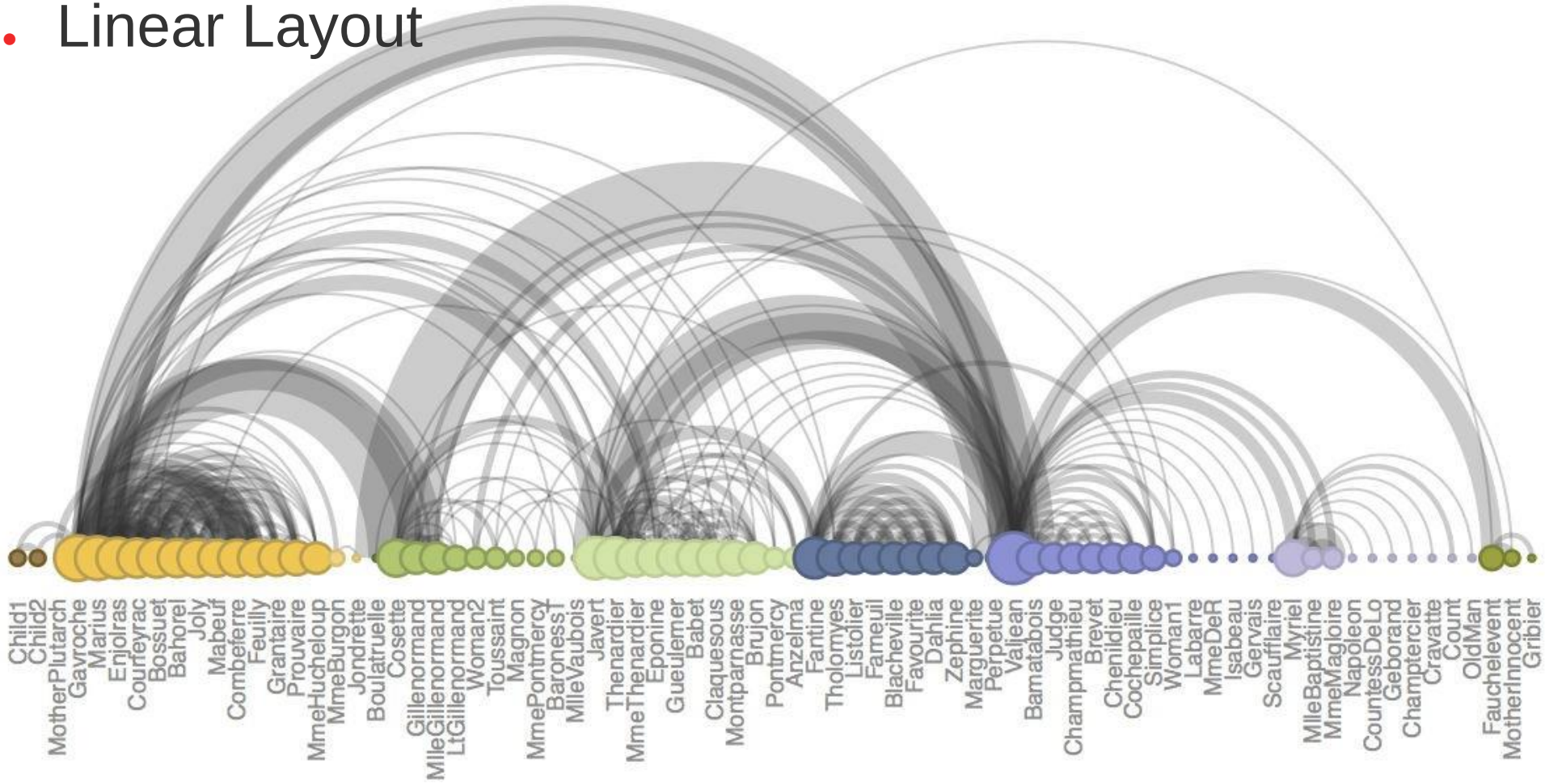
Qualitative Relational Structures

- Source Directed Node Link Diagram

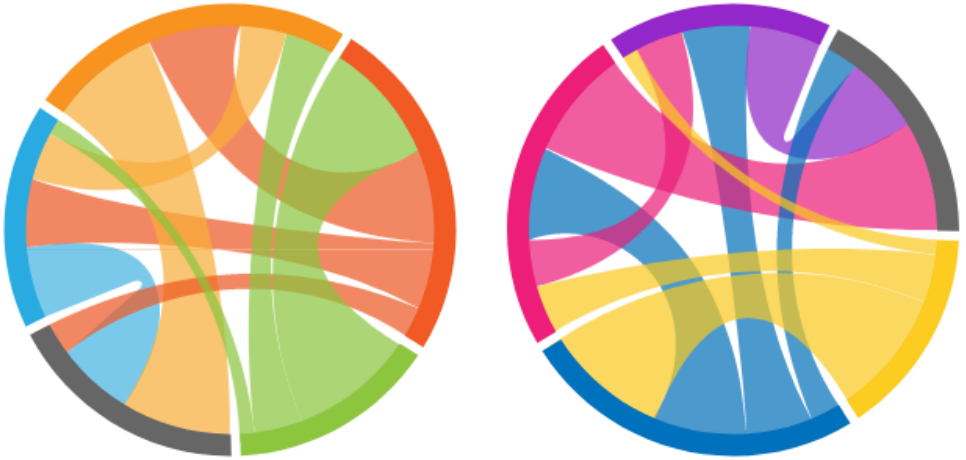


Qualitative Relational Structures

- Linear Layout

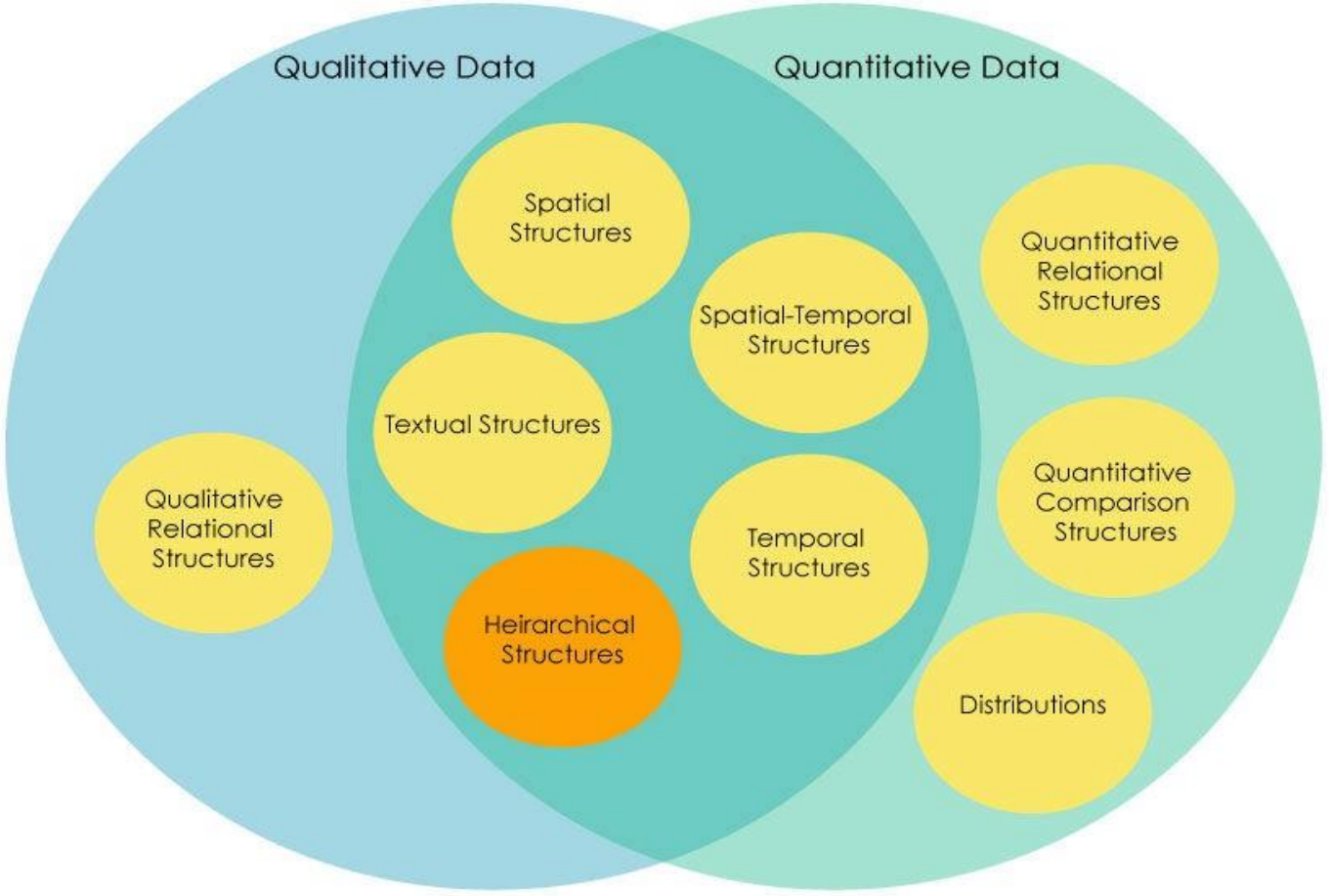


Qualitative Relational Structures



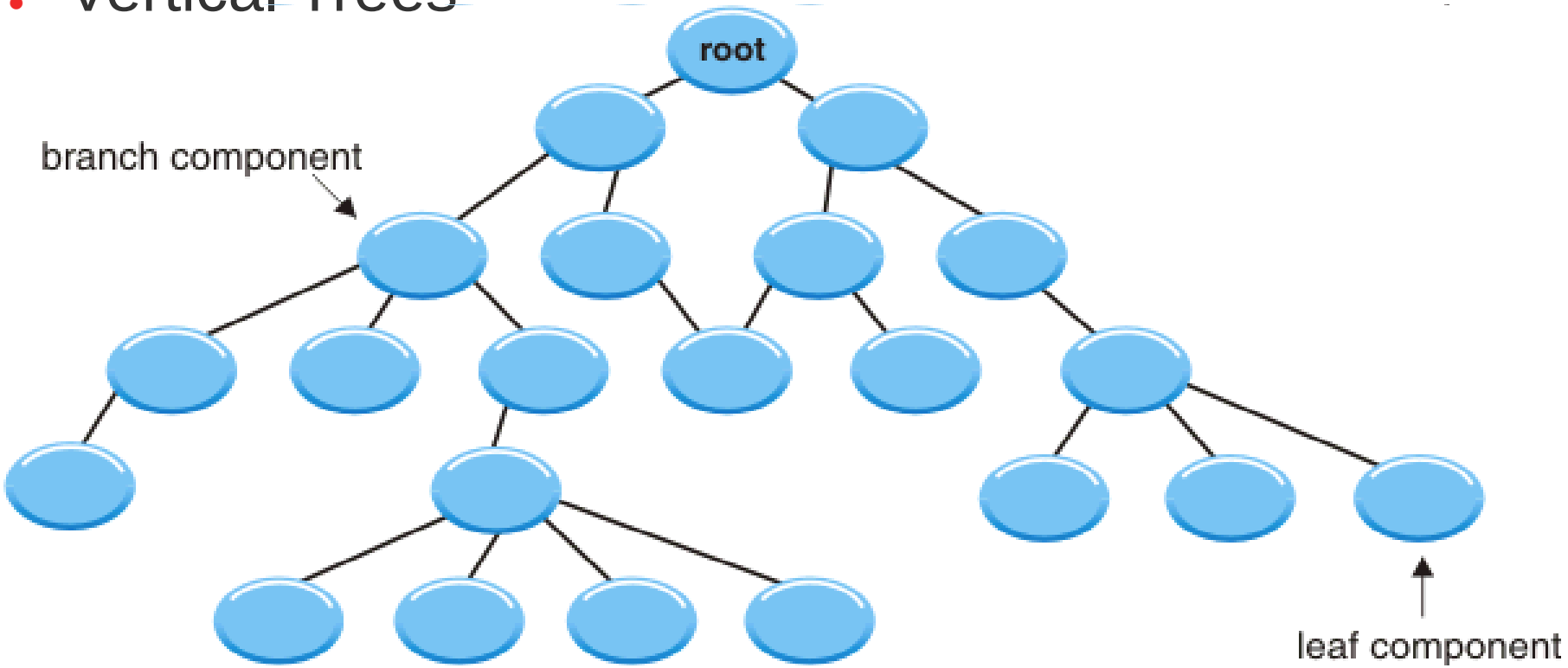
- Chord Diagram
  - visualises the inter-relationships between entities. The connections between entities are used to display that they share something in common.
  - The size of the arc is proportional to the importance of the flow.

Different Types of Data



Hierarchical Structures

- Vertical Trees

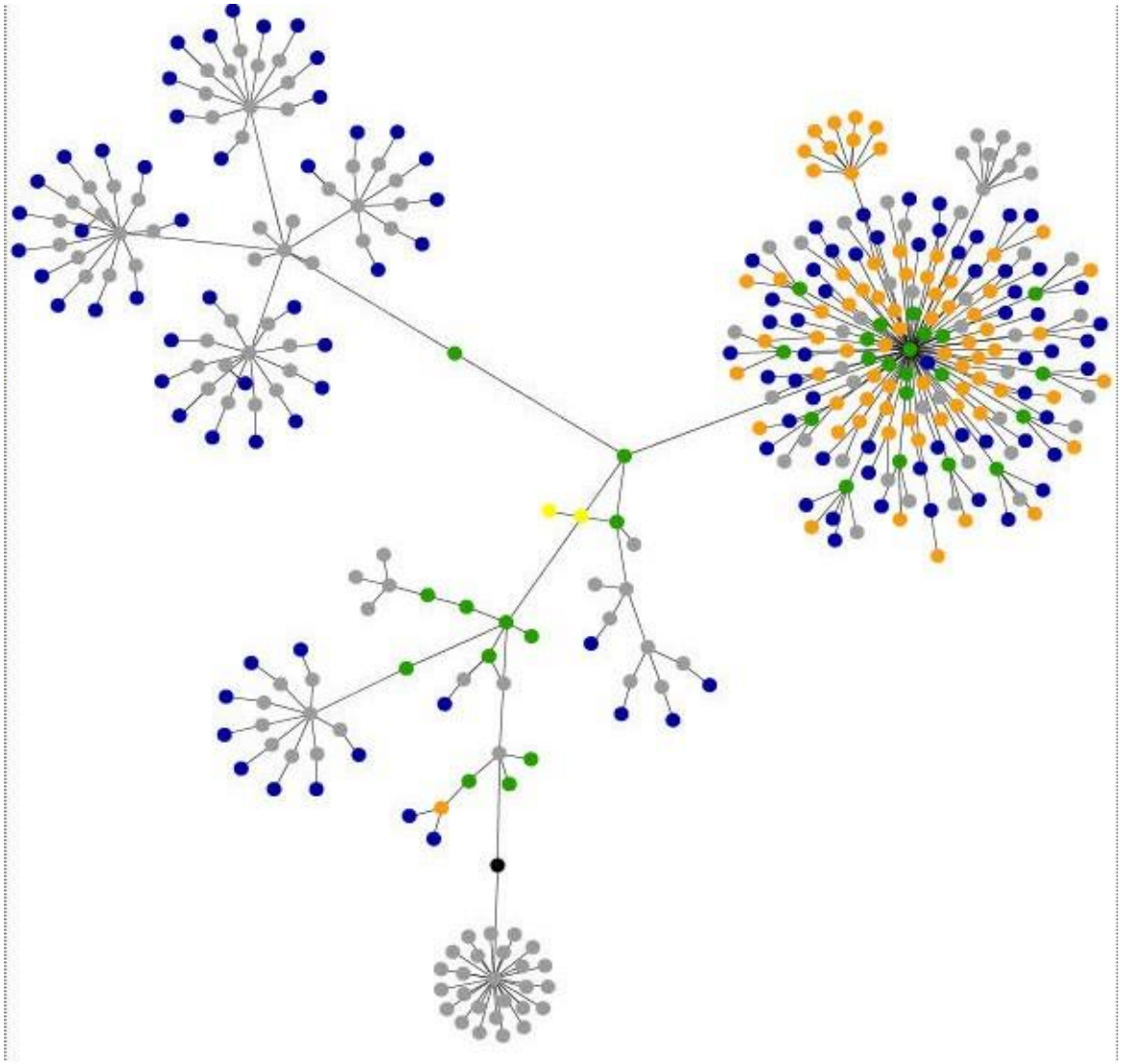


Hierarchical Structures

- Horizontal Trees

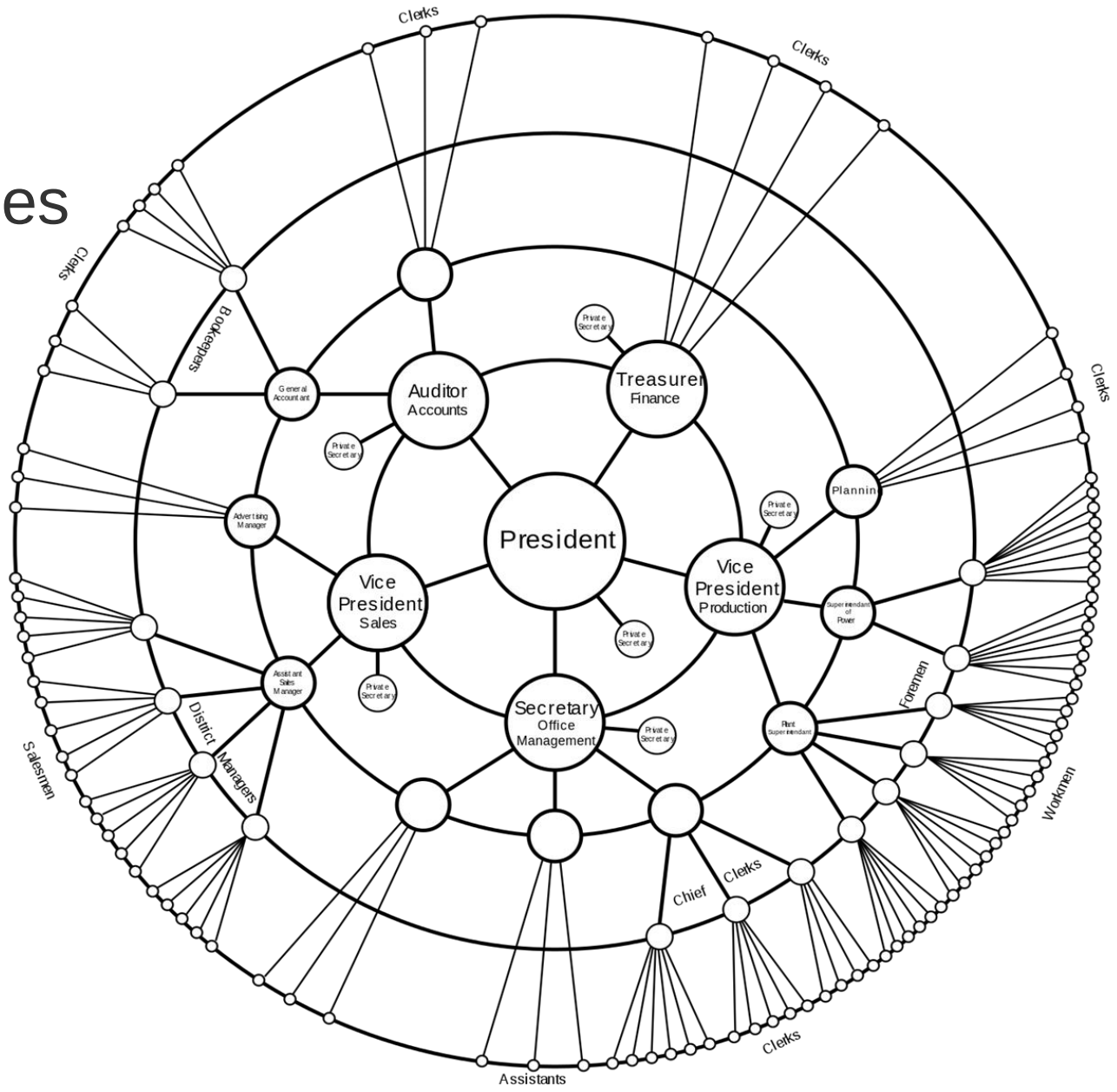
Hierarchical Structures

- Multi-Directional
- Trees
- Websites as Graphs



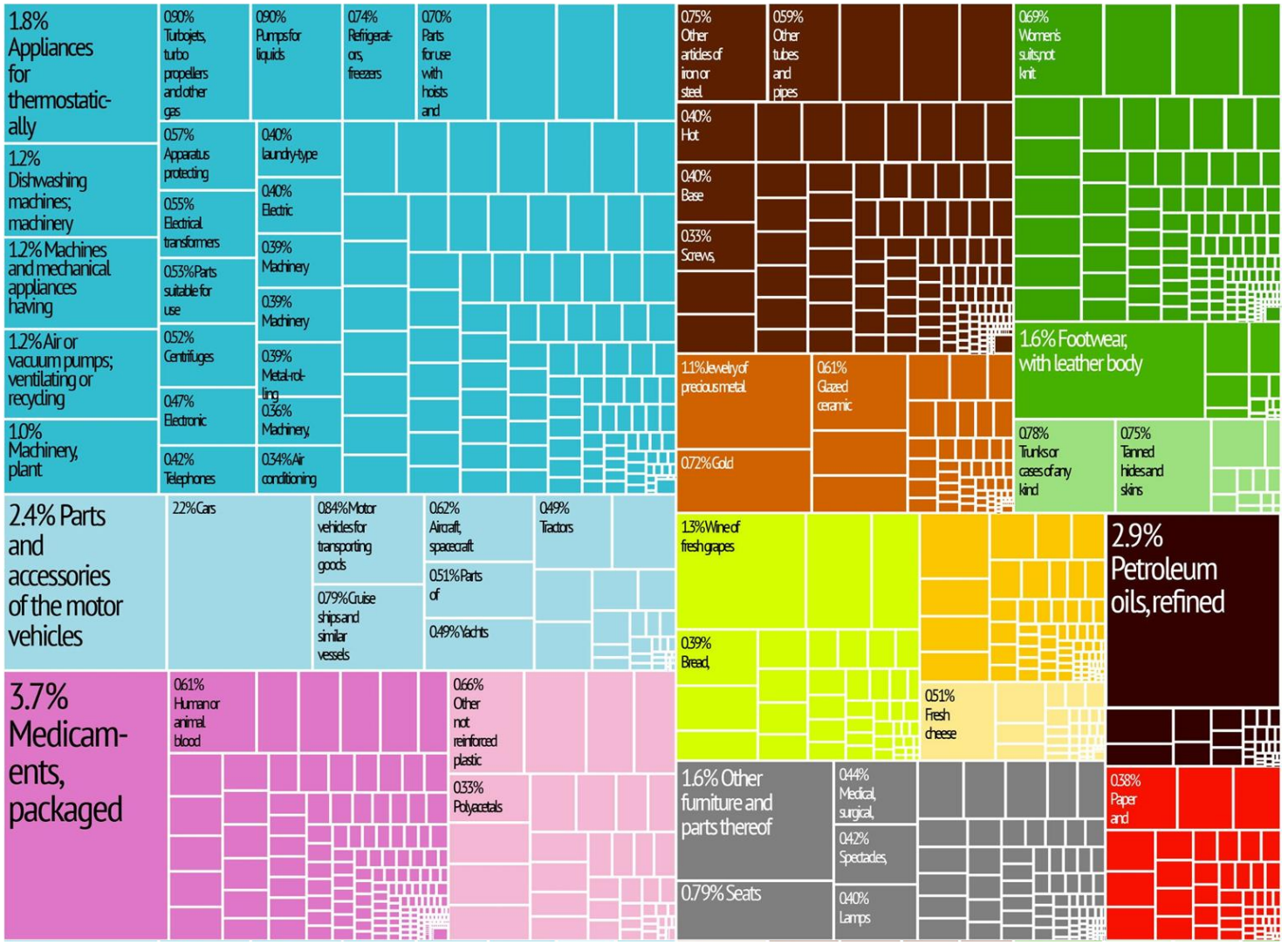
Hierarchical Structures

- Radial Trees



Hierarchical Structures

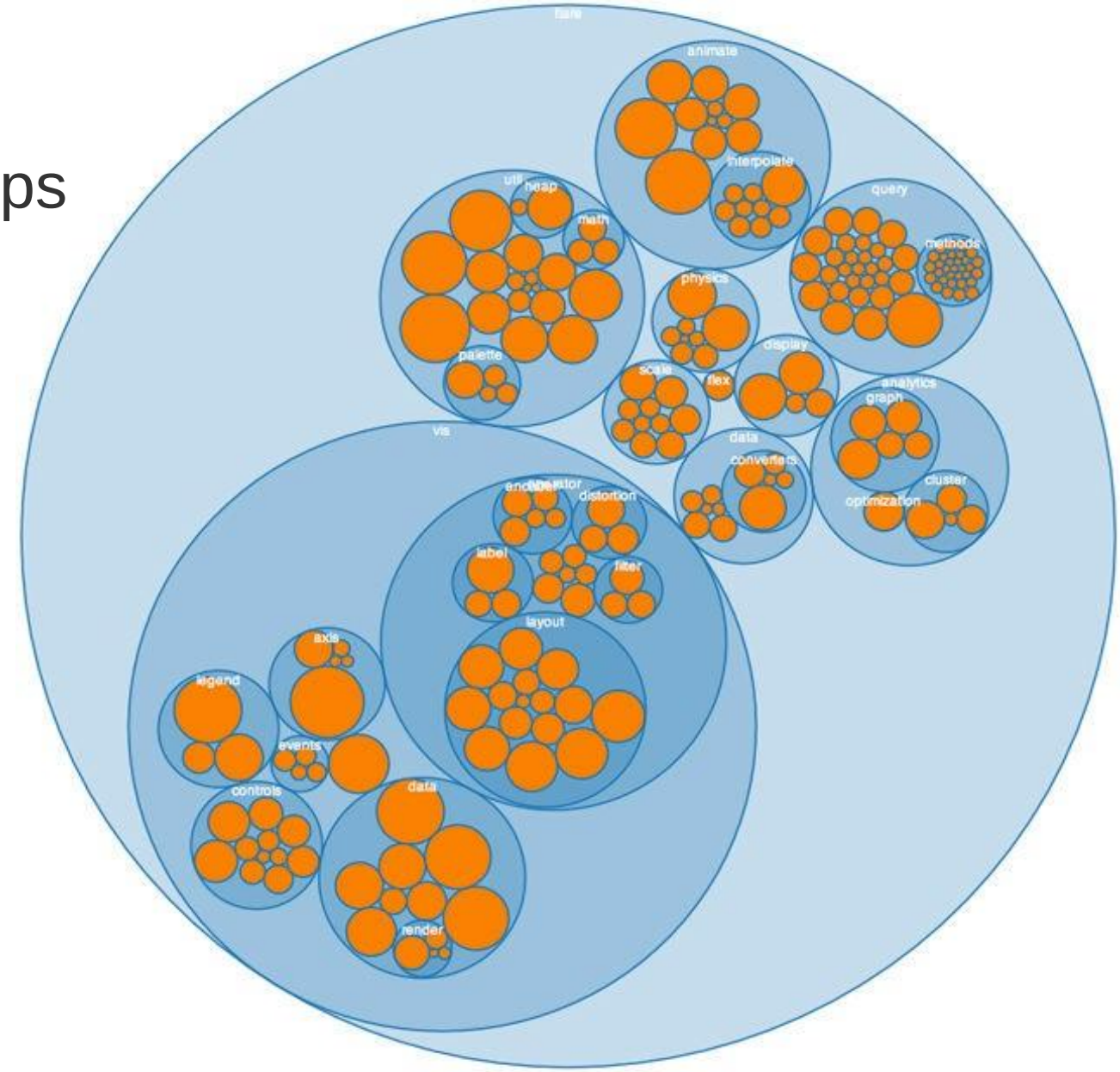
- Rectangular Tree Maps



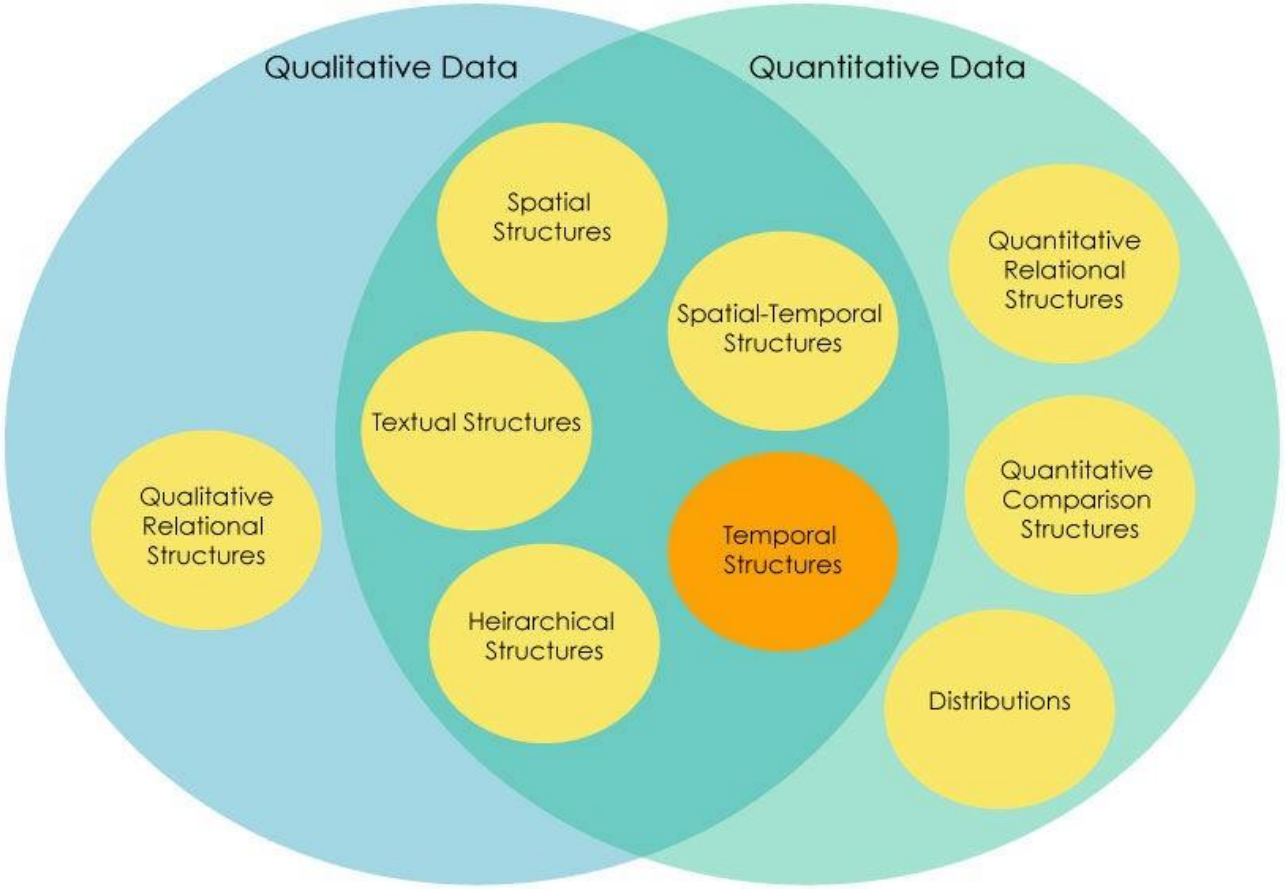
Hierarchical Structures

Hierarchical Structures

- Circular Tree Maps

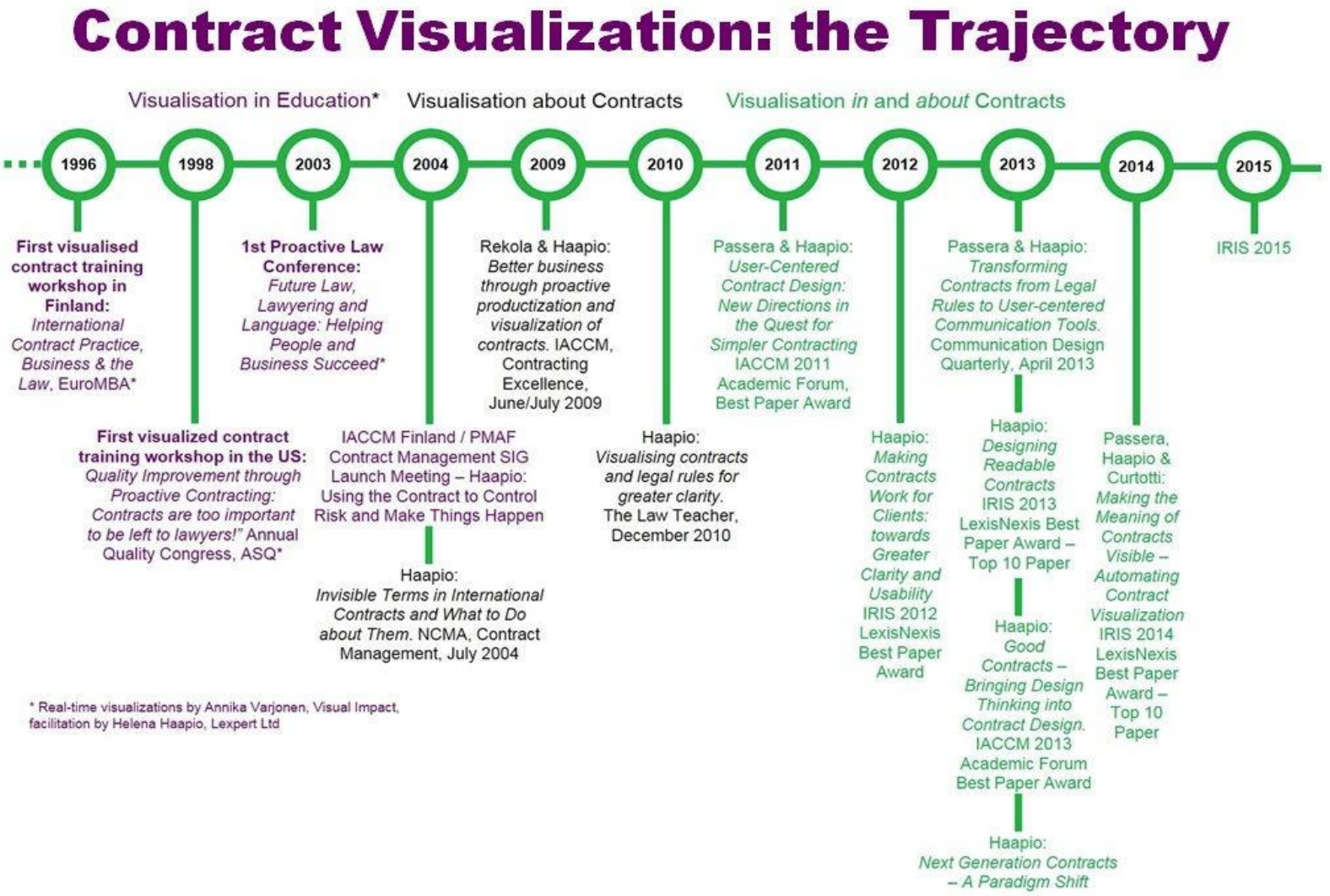


Different Types of Data

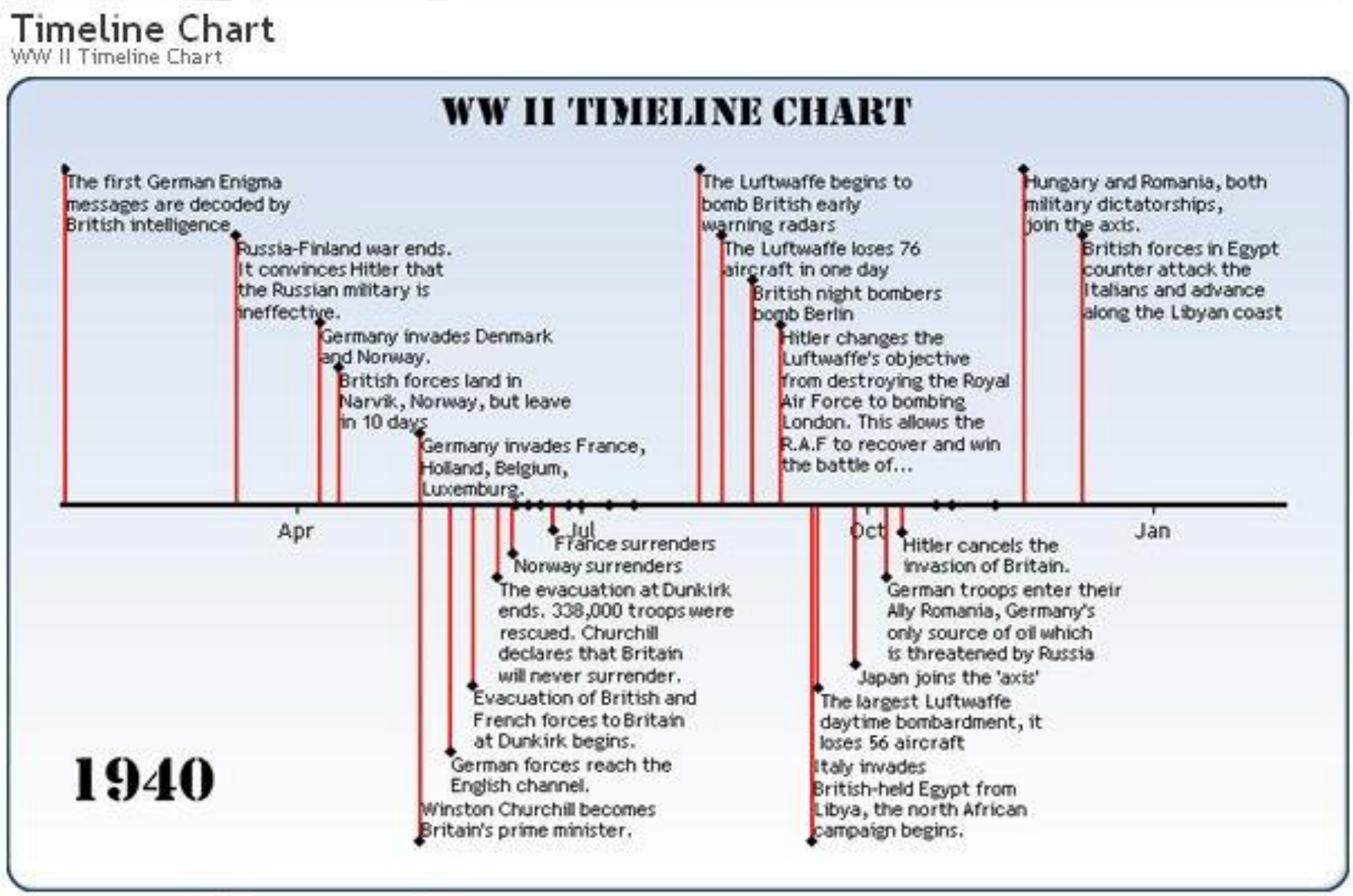


Temporal Structures

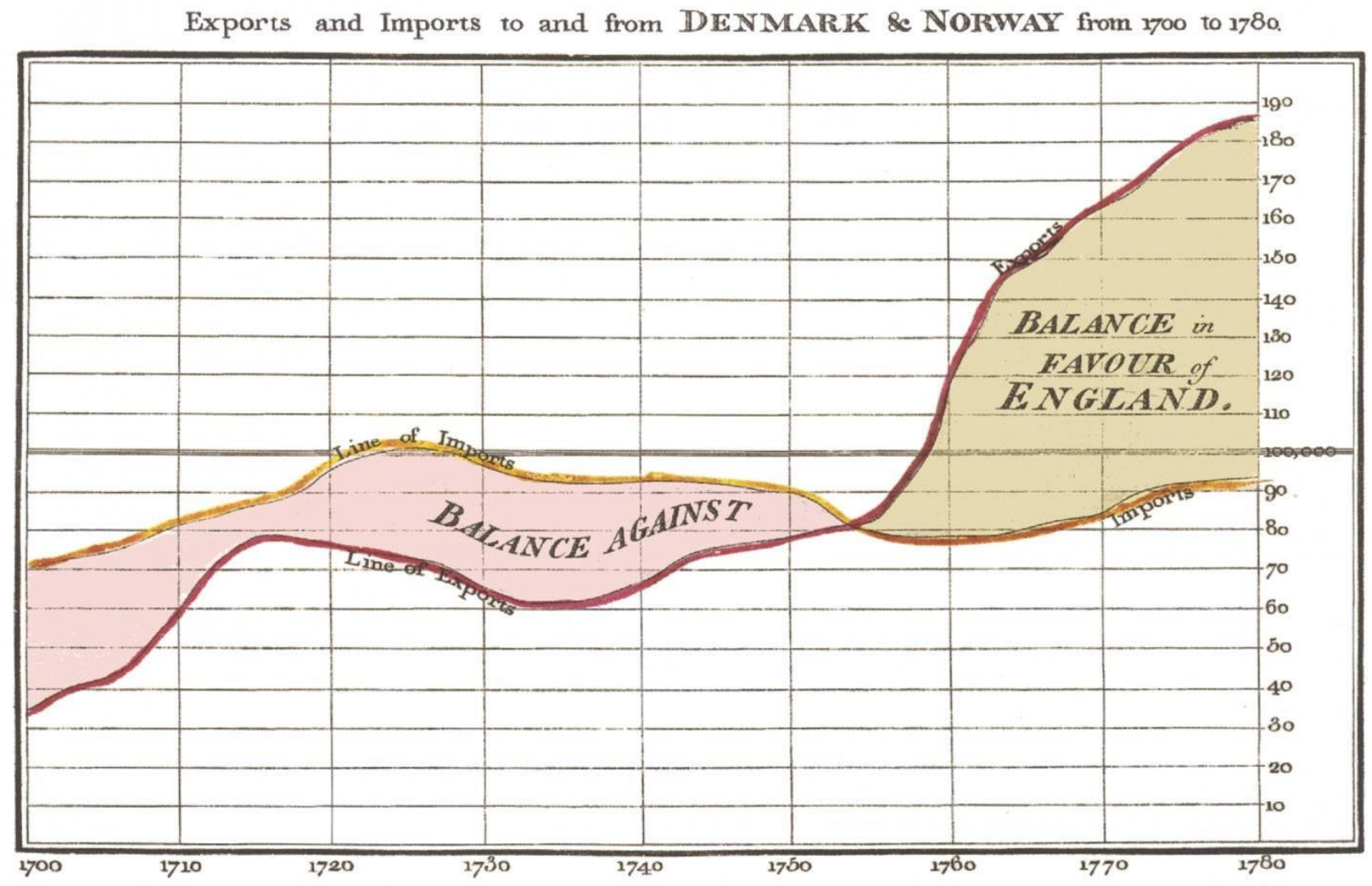
- Time Lines



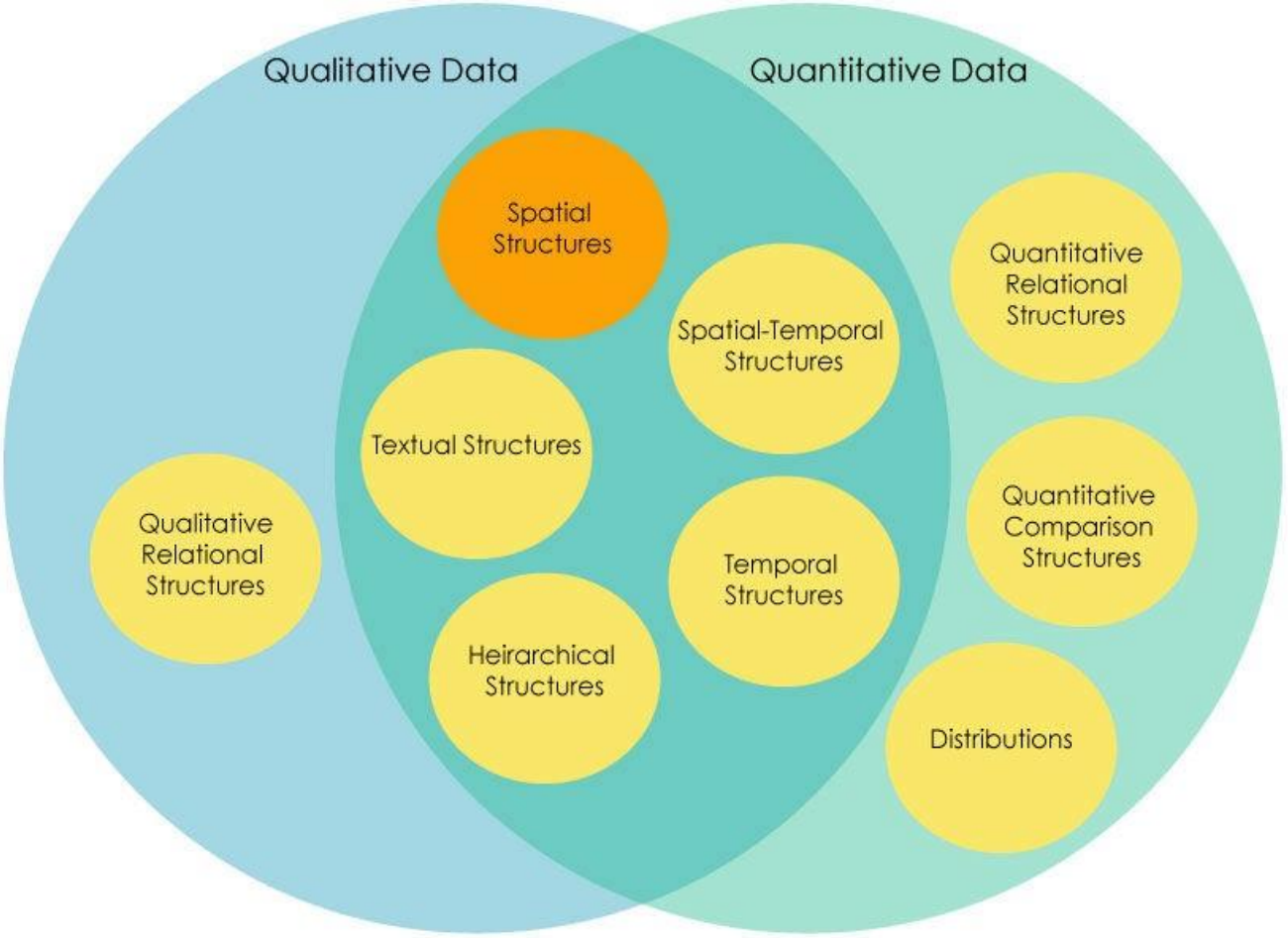
Temporal Structures



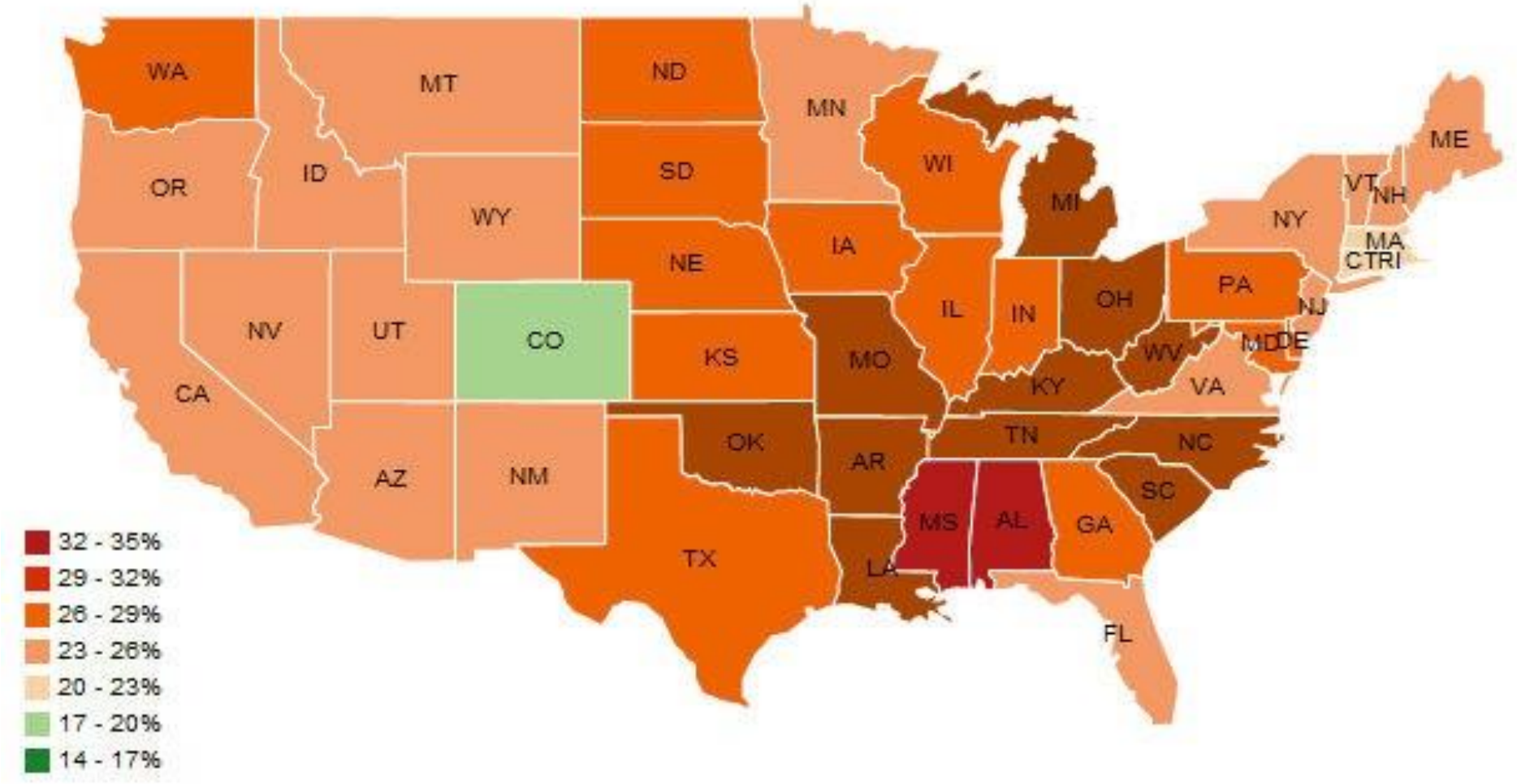
Temporal Structures



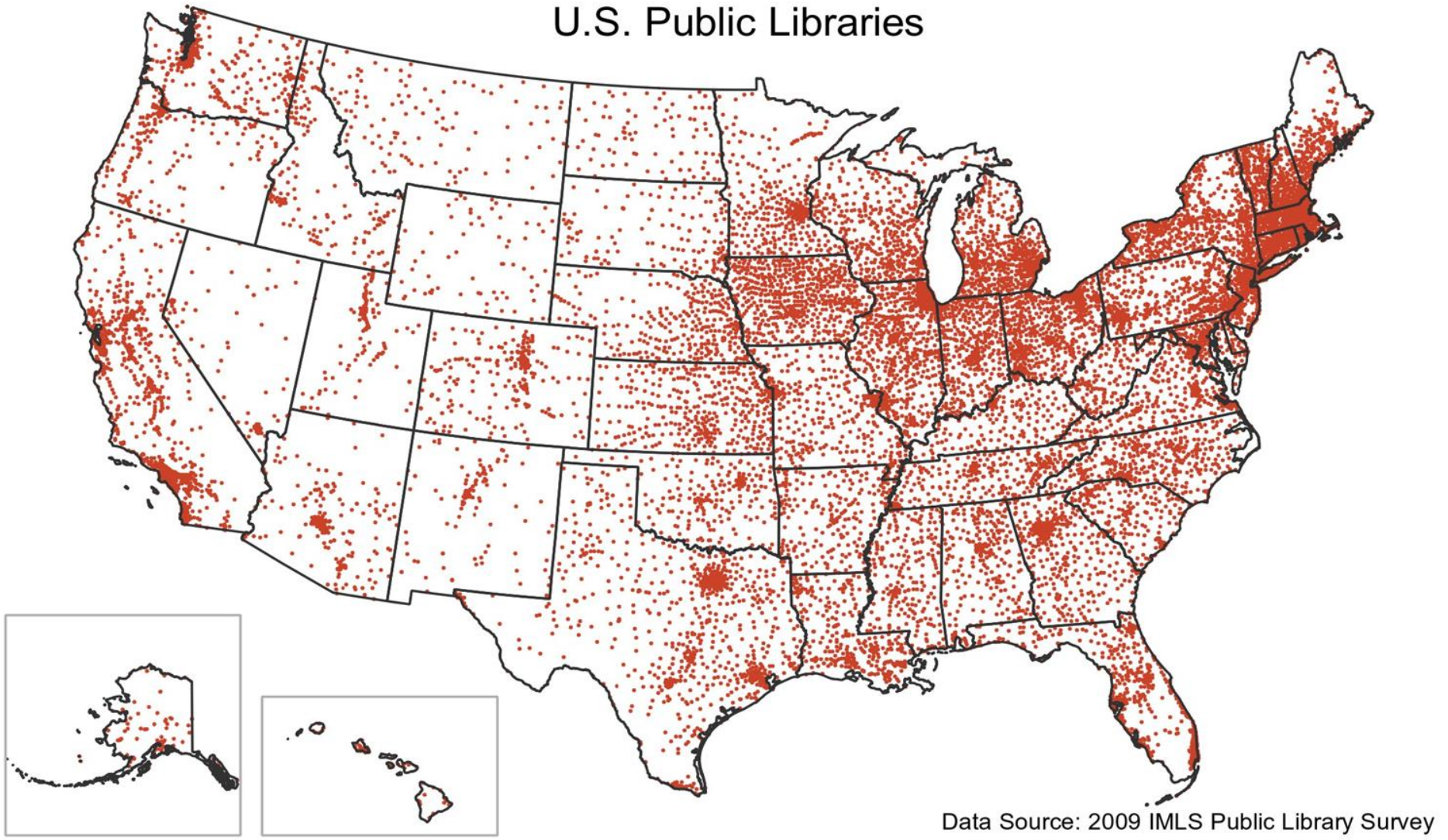
Different Types of Datas



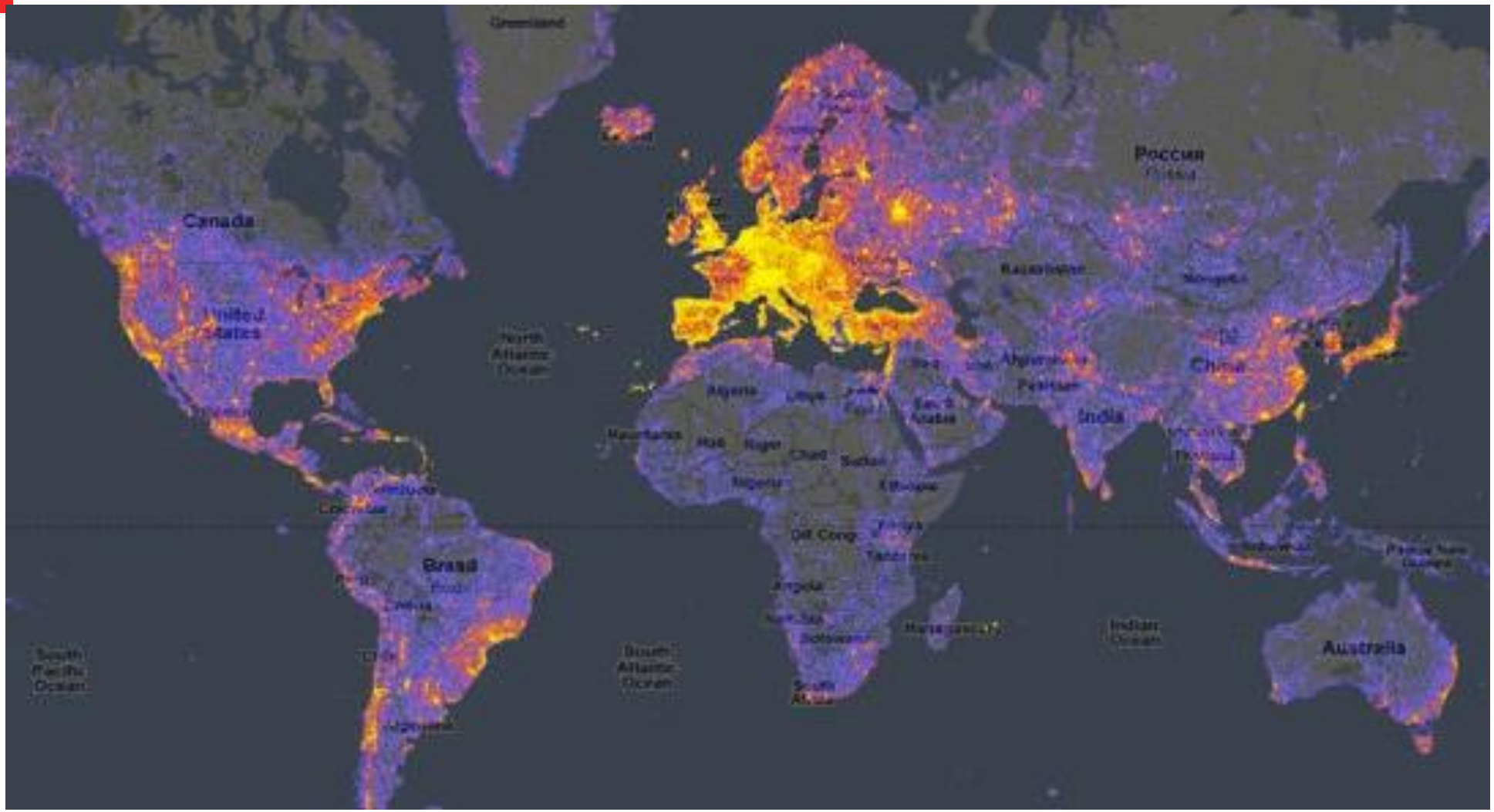
Spatial Structures: Maps



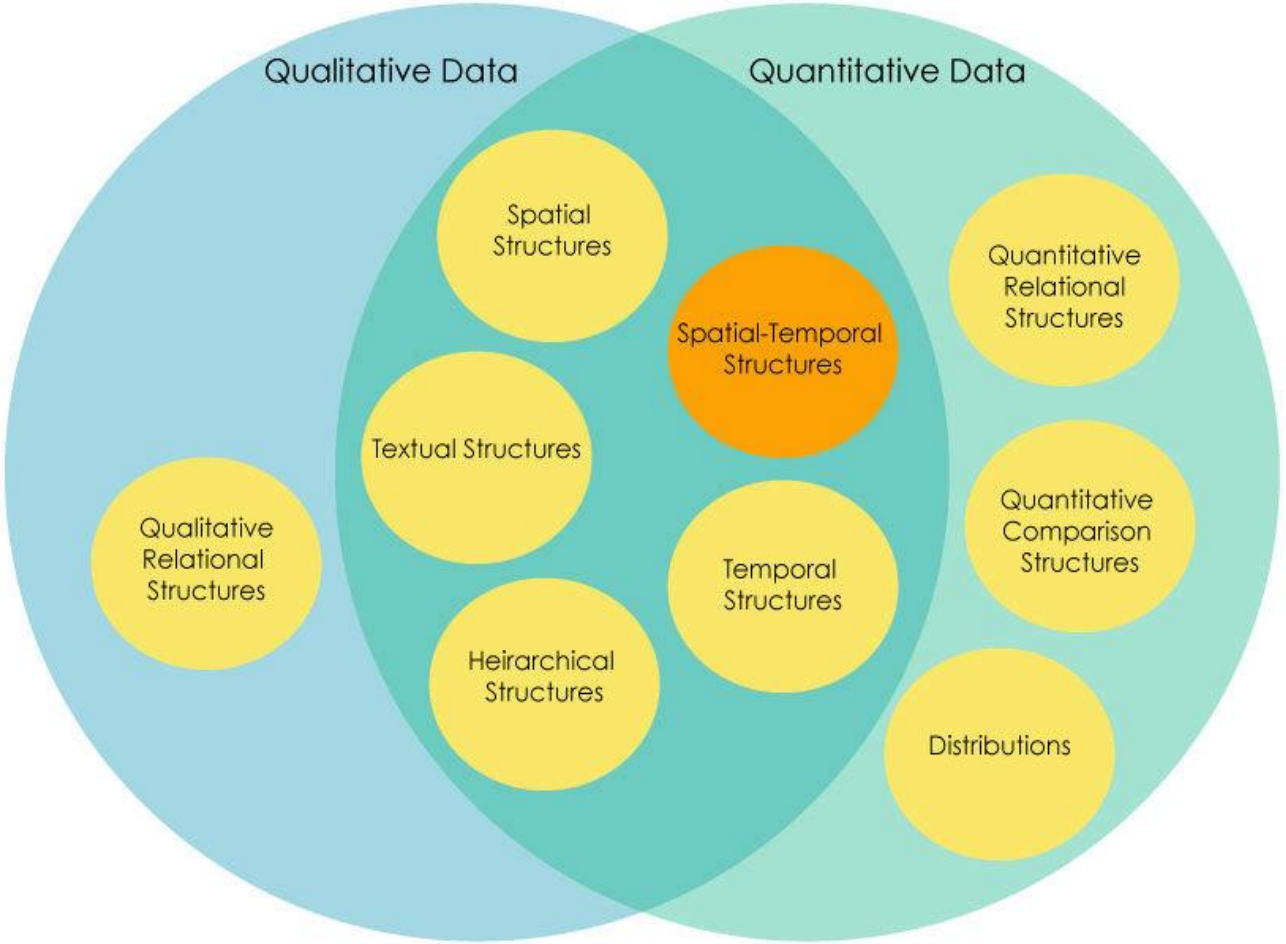
Spatial Structures: Maps



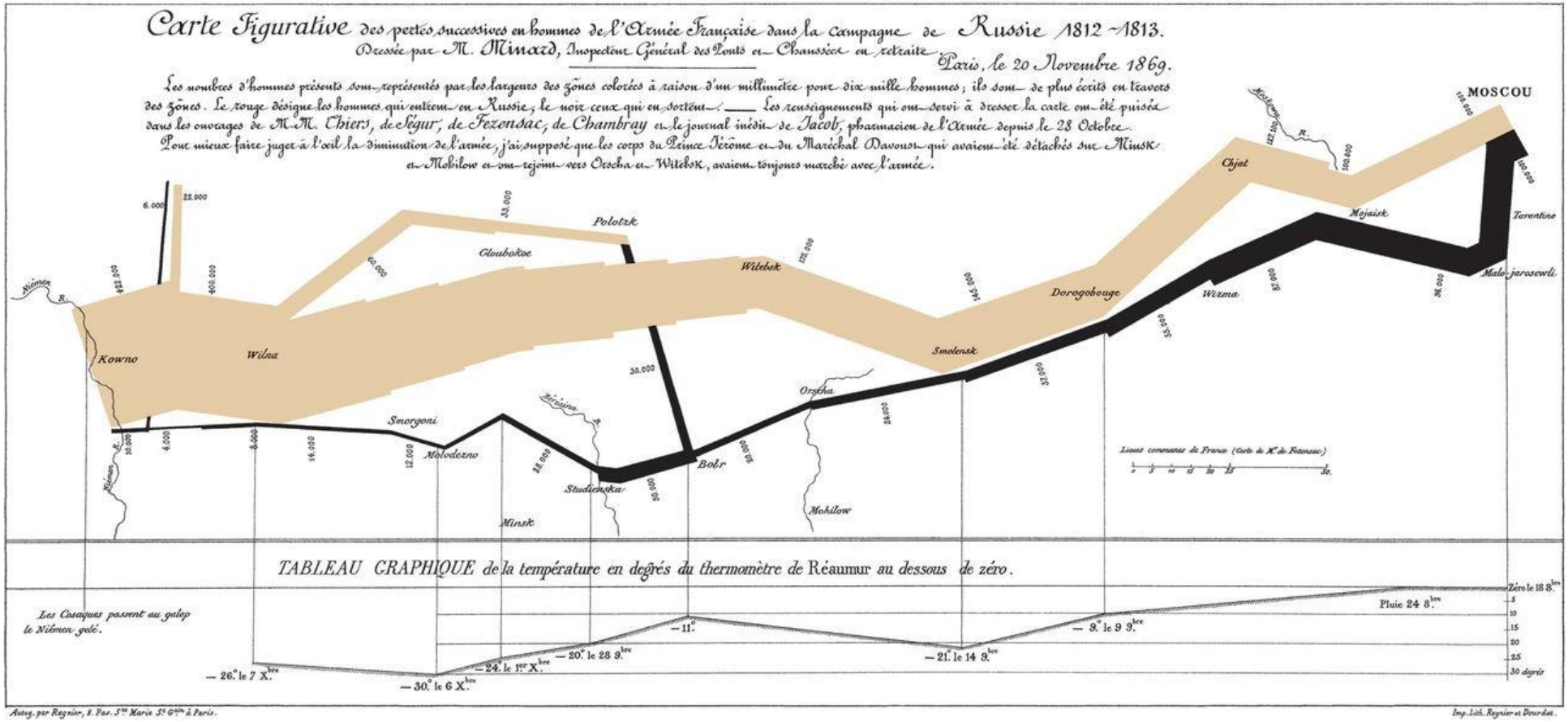
Spatial Structures: HeatMap



Different Types of Data

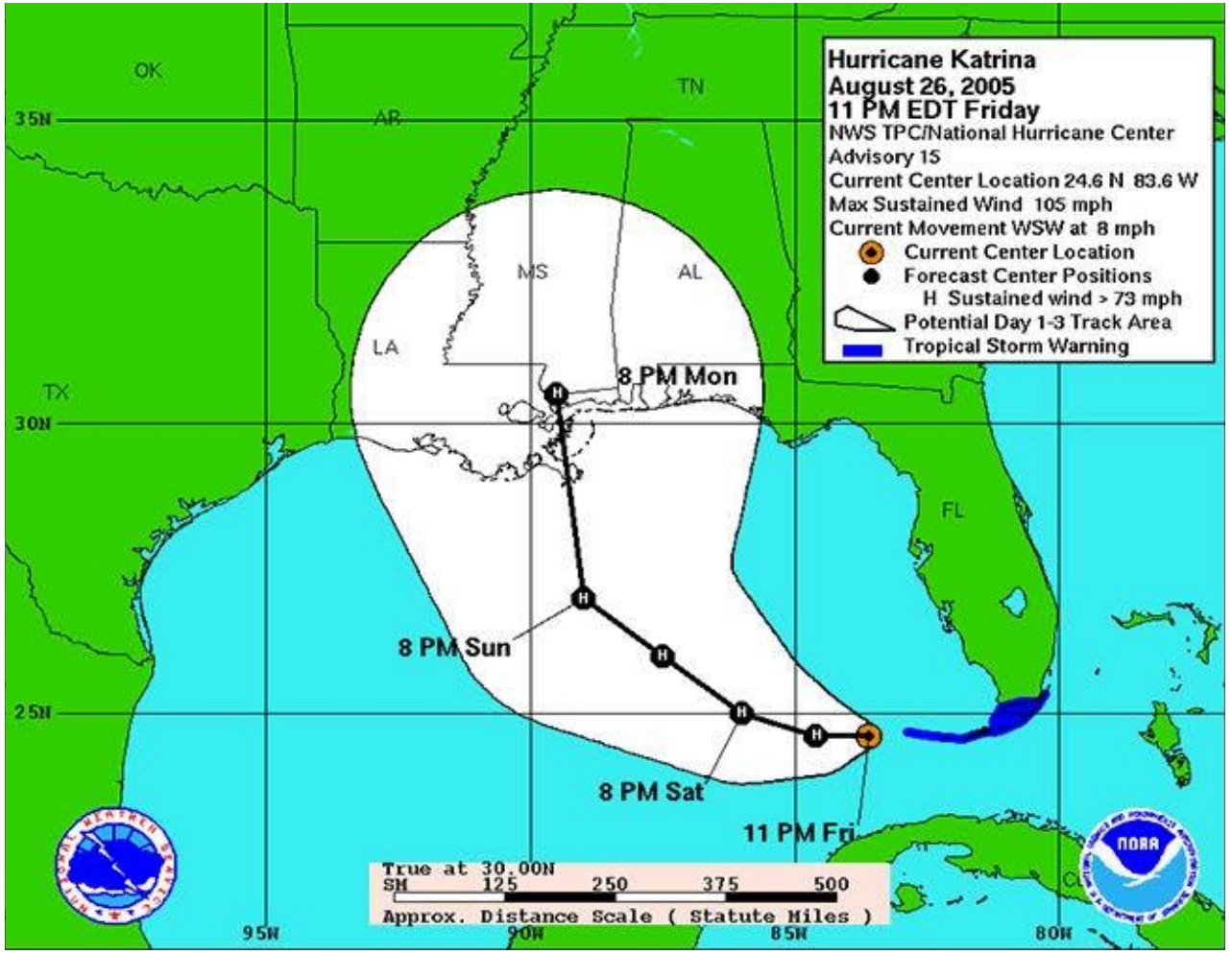


Spatial-Temporal Structures



Charles Minard's map of Napoleon's disastrous Russian campaign of 1812. The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates

Spatial-Temporal Structures



Class Activity

- Find a visualization online.
- Answer the following questions:
  - In one or two sentences, what story does it tell? Identify the data.
  - What type of data is it?
  - How many dimensions are being visually mapped? Identify the visual variables used.
  - Identify the type of visualization, or methods used.
  - If it is interactive, describe the interaction, and the data revealed.

Chart Suggestions—A Thought-Starter

What would you like to show?

Comparison

- Among Items
  - Two Variables per Item
    - Variable Width Column Chart
  - One Variable per Item
    - Many Categories
      - Table or Table with Embedded Charts
    - Few Categories
      - Many Items
        - Bar Chart
      - Few Items
        - Column Chart
- Over Time
  - Many Periods
    - Cyclical Data
      - Circular Area Chart
    - Non-Cyclical Data
      - Line Chart
  - Few Periods
    - Single or Few Categories
      - Column Chart
    - Many Categories
      - Line Chart

Relationship

- Two Variables
  - Scatter Chart
- Three Variables
  - Bubble Chart

Distribution

- Single Variable
  - Few Data Points
    - Column Histogram
  - Many Data Points
    - Line Histogram
- Two Variables
  - Scatter Chart
- Three Variables
  - 3D Area Chart

Composition

- Changing Over Time
  - Few Periods
    - Only Relative Differences Matter
      - Stacked 100% Column Chart
    - Relative and Absolute Differences Matter
      - Stacked Column Chart
  - Many Periods
    - Only Relative Differences Matter
      - Stacked 100% Area Chart
    - Relative and Absolute Differences Matter
      - Stacked Area Chart
- Static
  - Simple Share of Total
    - Pie Chart
  - Accumulation or Subtraction to Total
    - Waterfall Chart
  - Components of Components
    - Stacked 100% Column Chart with Subcomponents

Thank You