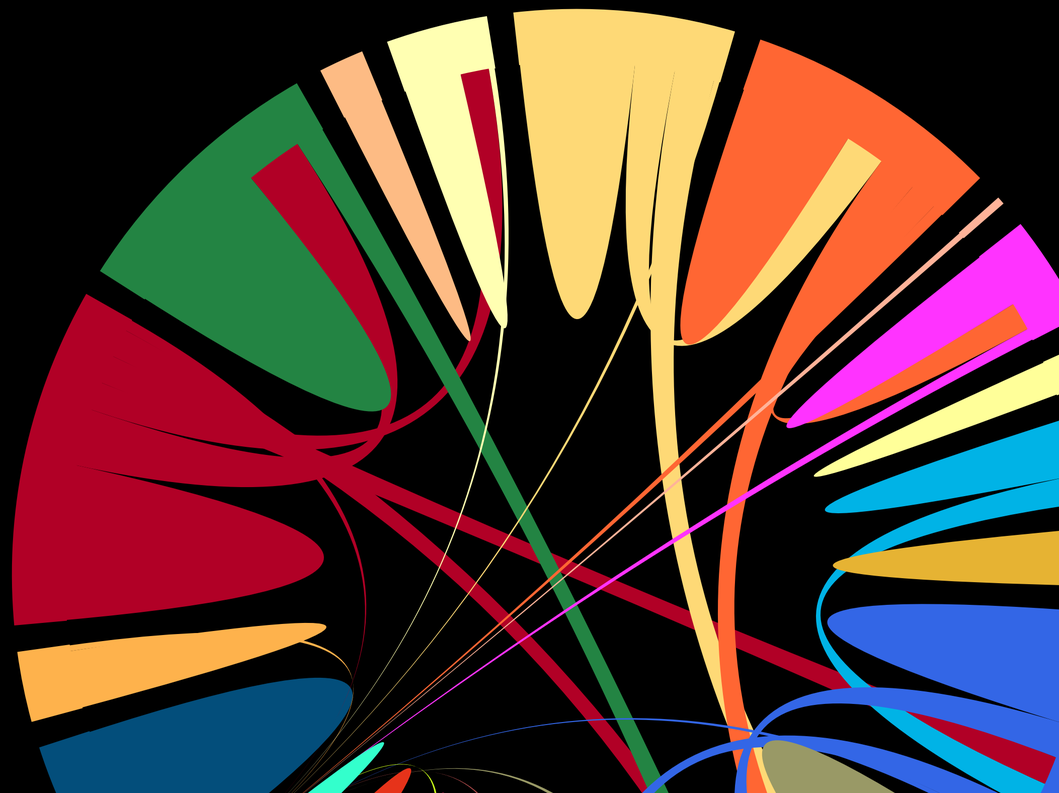
Information Visualization INFO250

Chapter 2:
- Ontologies of visualization
- Unwanted biases



Luís Cruz - l.cruz@tudelft.nl

Overview of the chapter

- In the following two weeks, we will talk about **classification of visualization**. In this class specifically, we will talk about the classification by (1) data type and (2) complexity and how that is related to the dataset that you are working on.
- We will also discuss some principles and pitfalls in the designing of some common visualization types.
- This topic is important for two reasons:
  - It is a basic element to take into consideration during the design of visualization.
  - It is important because you need to learn to **speak the language** before becoming a professional.

Ontology

What is an ontology?

- a set of concepts and **categories** in a subject area or **domain** that shows their **properties and the relations** between them.
- Another word: **taxonomy**

Ontology of visualizations based on Data Type

By Shneiderman, 1996

- Classification of visualizations according to **data type (not the visualization itself⚠️)**:
  - 1D/Linear
  - 2D/Planar
  - 3D/Volumetric
  - Temporal
  - Multi-dimensional
  - Tree
  - Network
- https://www.cs.umd.edu/~ben/papers/Shneiderman1996eyes.pdf

The Eyes Have It:
A Task by Data Type Taxonomy for Information Visualizations

Ben Shneiderman
Department of Computer Science,
Human-Computer Interaction Laboratory, and Institute for Systems Research
University of Maryland
College Park, Maryland 20742 USA
ben@cs.umd.edu

Abstract

A useful starting point for designing advanced graphical user interfaces is the Visual Information-Seeking Mantra: overview first, zoom and filter, then details on demand. But this is only a starting point in trying to understand the rich and varied set of information visualizations that have been proposed in recent years. This paper offers a task by data type taxonomy with seven data types (one-, two-, three-dimensional data, temporal and multi-dimensional data, and tree and network data) and seven tasks (overview, zoom, filter, details-on-demand, relate, history, and extracts).

> Everything points to the conclusion that the phrase 'the language of art' is more than a loose metaphor, that even to describe the visible world in images we need a developed system of schemata.
>
> E. H. Gombrich *Art and Illusion,* 1959 (p. 76)

1. Introduction

Information exploration should be a joyous experience, but many commentators talk of information overload and anxiety (Wurman, 1989). However, there is promising evidence that the next generation of digital libraries for structured databases, textual documents, and multimedia will enable convenient exploration of growing information spaces by a wider range of users. Visual language researchers and user-interface designers are inventing powerful information visualization methods, while offering smoother integration of technology with task.

The terminology swirl in this domain is especially colorful. The older terms of information retrieval (often applied to bibliographic and textual document systems) and database management (often applied to more structured relational database systems with orderly attributes and sort keys), are being pushed aside by newer notions of information gathering, seeking, or visualization and data mining, warehousing, or filtering. While distinctions are subtle, the common goals reach from finding a narrow set of items in a large collection that satisfy a well-understood information need (known-item search) to developing an understanding of unexpected patterns within the collection (browse) (Marchionini, 1995).

Exploring information collections becomes increasingly difficult as the volume grows. A page of information is easy to explore, but when the information becomes the size of a book, or library, or even larger, it may be difficult to locate known items or to browse to gain an overview.

Designers are just discovering how to use the rapid and high resolution color displays to present large amounts of information in orderly and user-controlled ways. Perceptual psychologists, statisticians, and graphic designers (Bertin, 1983; Cleveland, 1993; Tufte, 1983, 1990) offer valuable guidance about presenting static information, but the opportunity for dynamic displays takes user interface designers well beyond current wisdom.

2. Visual Information Seeking Mantra

The success of direct-manipulation interfaces is indicative of the power of using computers in a more visual or graphic manner. A picture is often cited to be worth a thousand words and, for some (but not all) tasks, it is clear that a visual presentation—such as a map or photograph—is dramatically easier to use than is a textual description or a spoken report. As computer speed and display resolution increase, information visualization and graphical interfaces are likely to have an expanding role. If a map of the United States is displayed, then it should be possible to point rapidly at one of 1000 cities to get tourist information. Of course, a foreigner who knows a city's name (for example, New Orleans), but not its location, may do better with a scrolling alphabetical list.

0-8186-7469-5/96 $05.00 © 1996 IEEE

1-dimensional

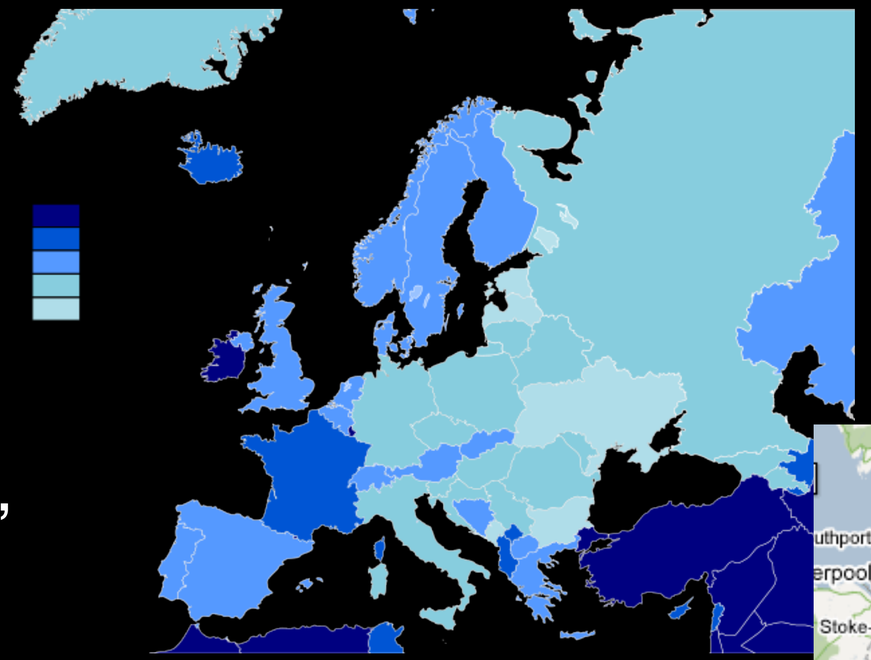
- Source code, text, alphabetical lists

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">

<html xmlns="http://www.w3.org/1999/xhtml">
    <head>
        <meta http-equiv="Content-Type" content=
        "text/html; charset=us-ascii" />
        <script type="text/javascript">
            function reDo() {top.location.reload();}
            if (navigator.appName == 'Netscape') {top.onresize = reDo;}
            dom=document.getElementById;
        </script>
    </head>
    <body>
    </body>
</html>
```

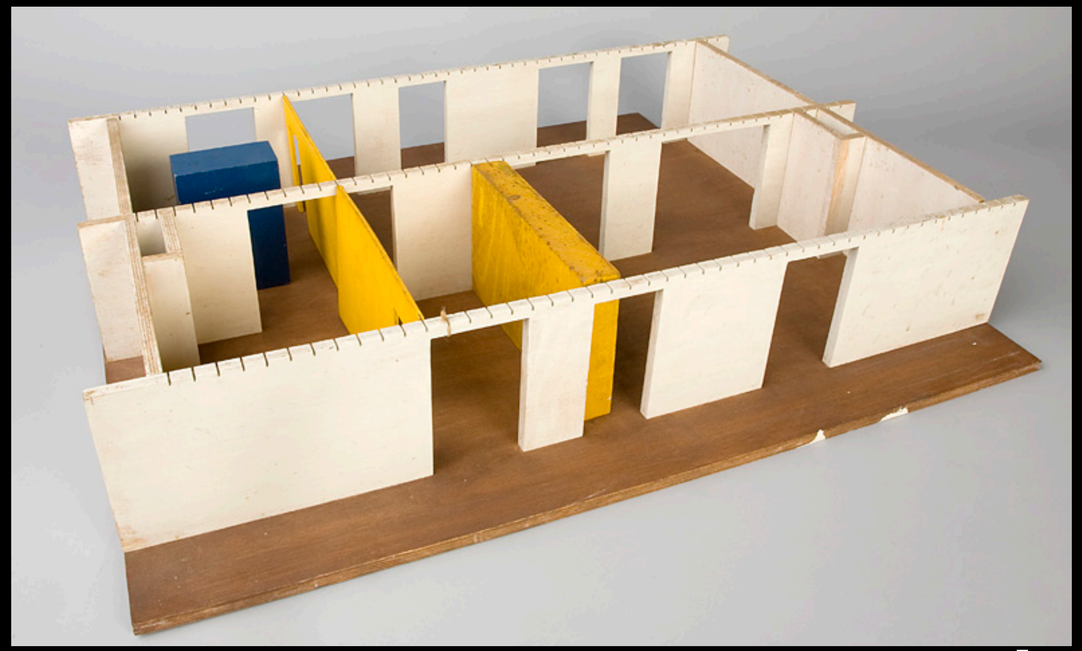
2-dimensional

- Planar or map data.
  - E.g: geographic maps, floorplans, or newspaper layouts
- Each data point is mapped into the 2d plan, covering a dot or area in that plan (not necessarily squared)



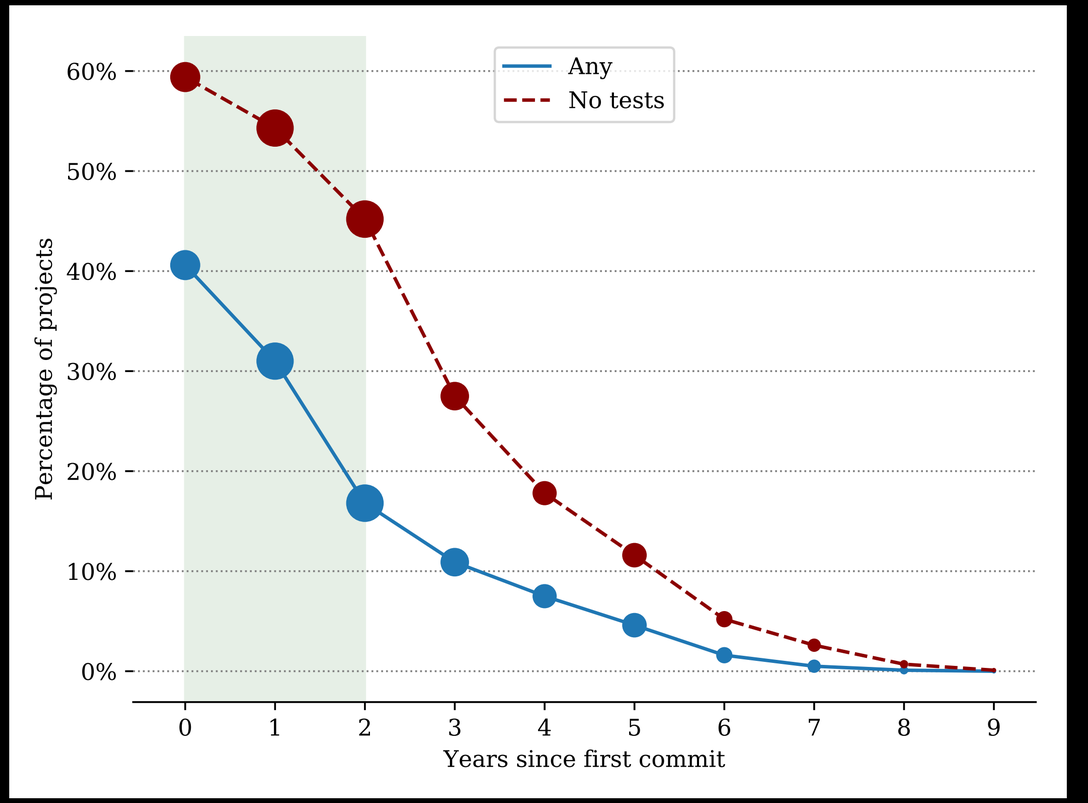
3-dimensional

- Object modeling



Temporal

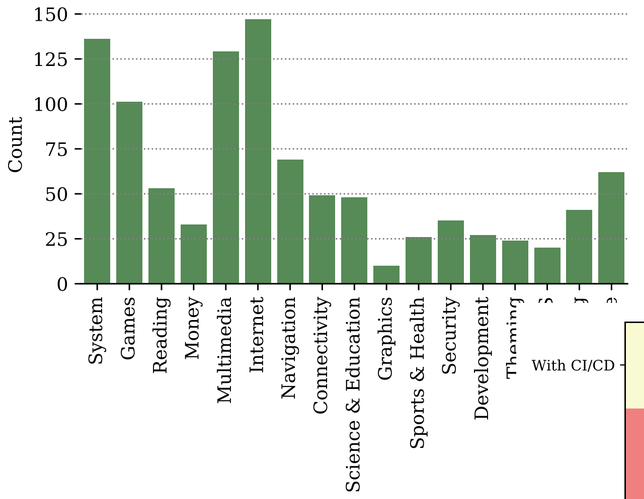
- Temporal data variable can be regarded as a normal type of data point.
- Some common graph types for temporal data include: Gantt Chart, line chart, area chart.
- Temporal data is typically represented with timestamps along the x-axis.

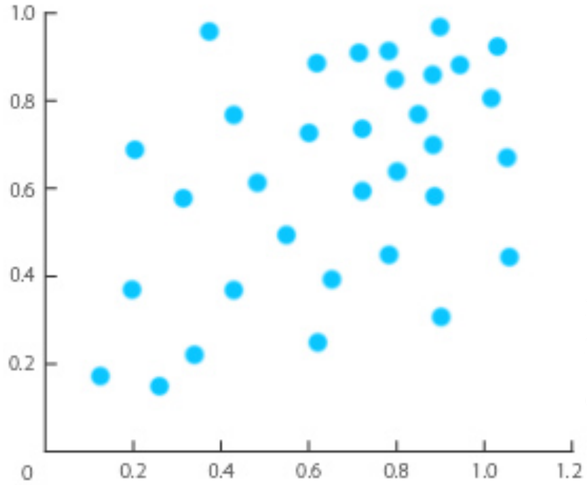


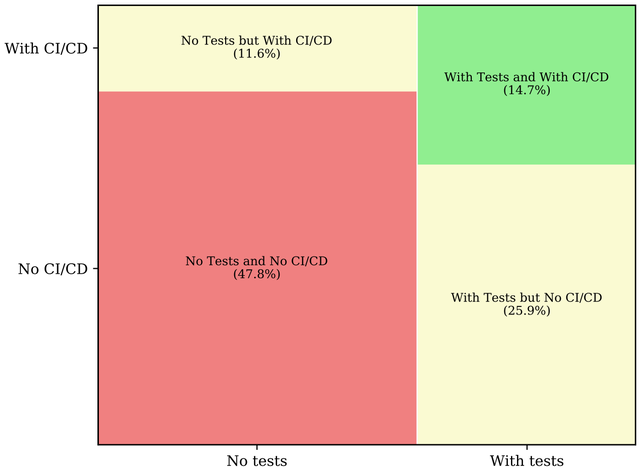
		2011				2012				2013				2014			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Task 1	Bootstrap																
	Thesis Proposal																
Task 2	Non-Interactive Simulator																
	Interactive Simulator																
Task 3	Deductive Fault Localization																
Task 4	Inductive Fault Localization																
Task 5	Self-healing Framework																
Task 6	Thesis																
	Publications																
	Milestones						1			3		4	2	5	6		

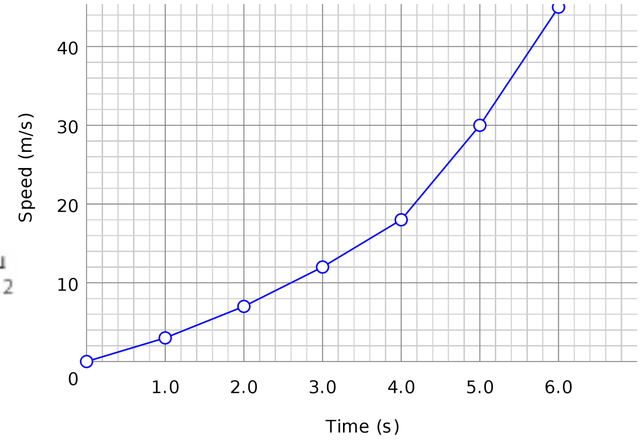
Multi-dimensional

- most relational and statistical databases are processed as multi-dimensional data
- The data is typically represented using 2D visualizations. (1D or 3D could also be used)



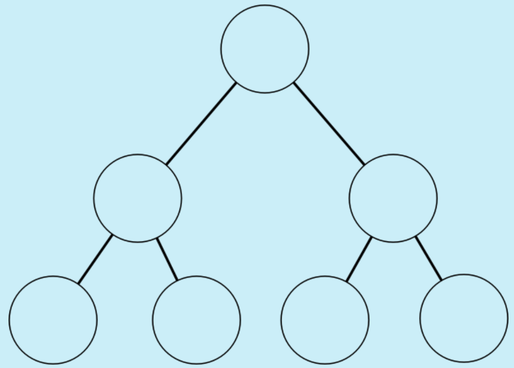


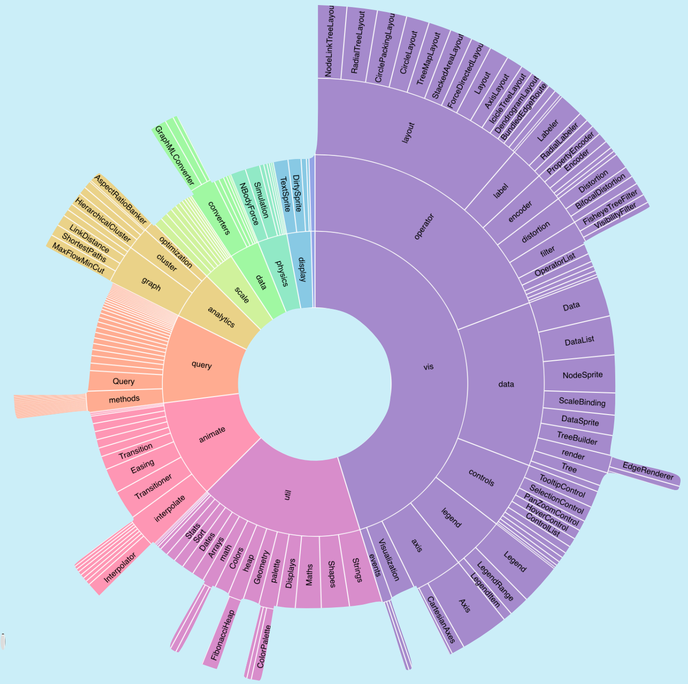




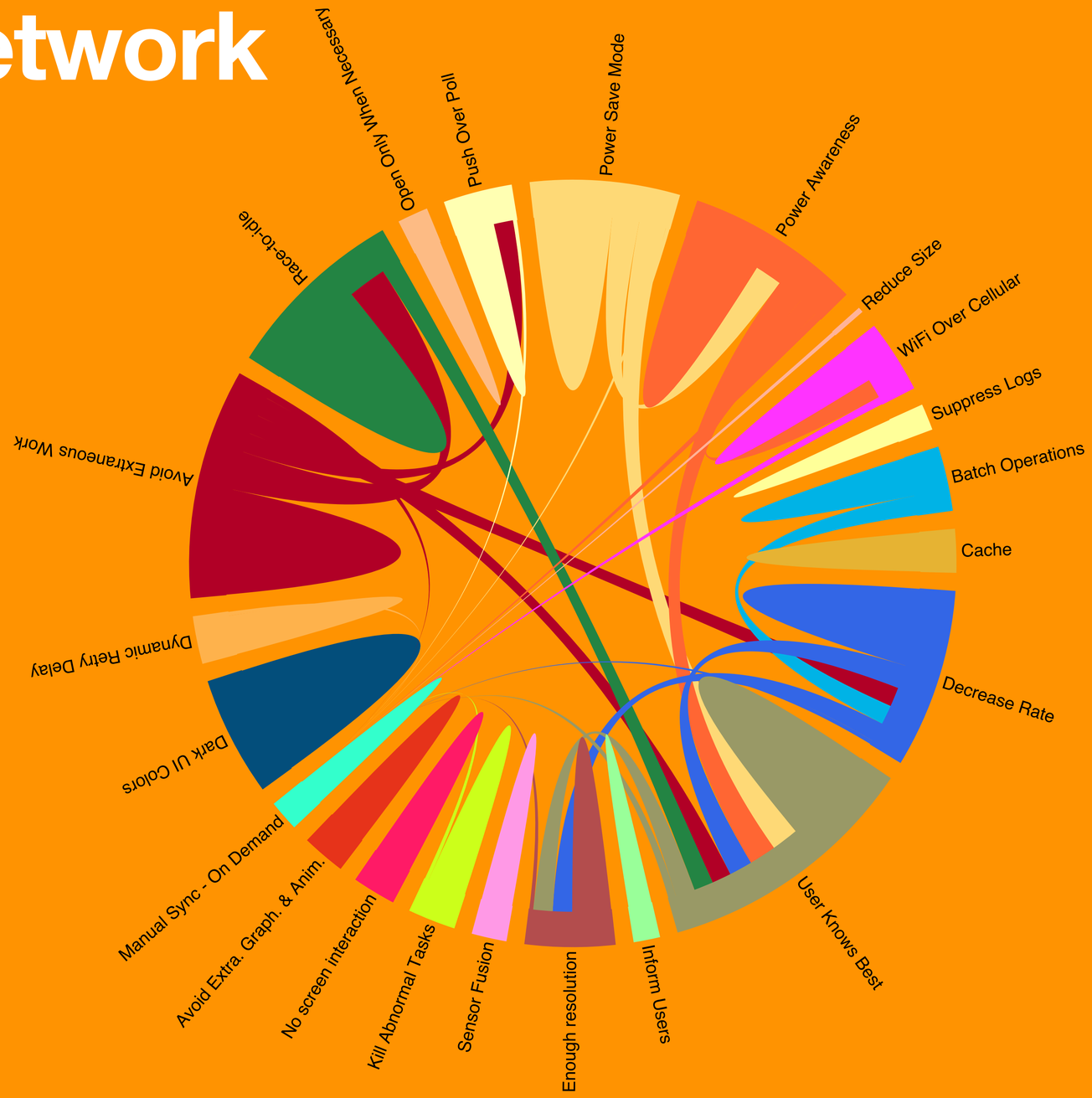
Tree

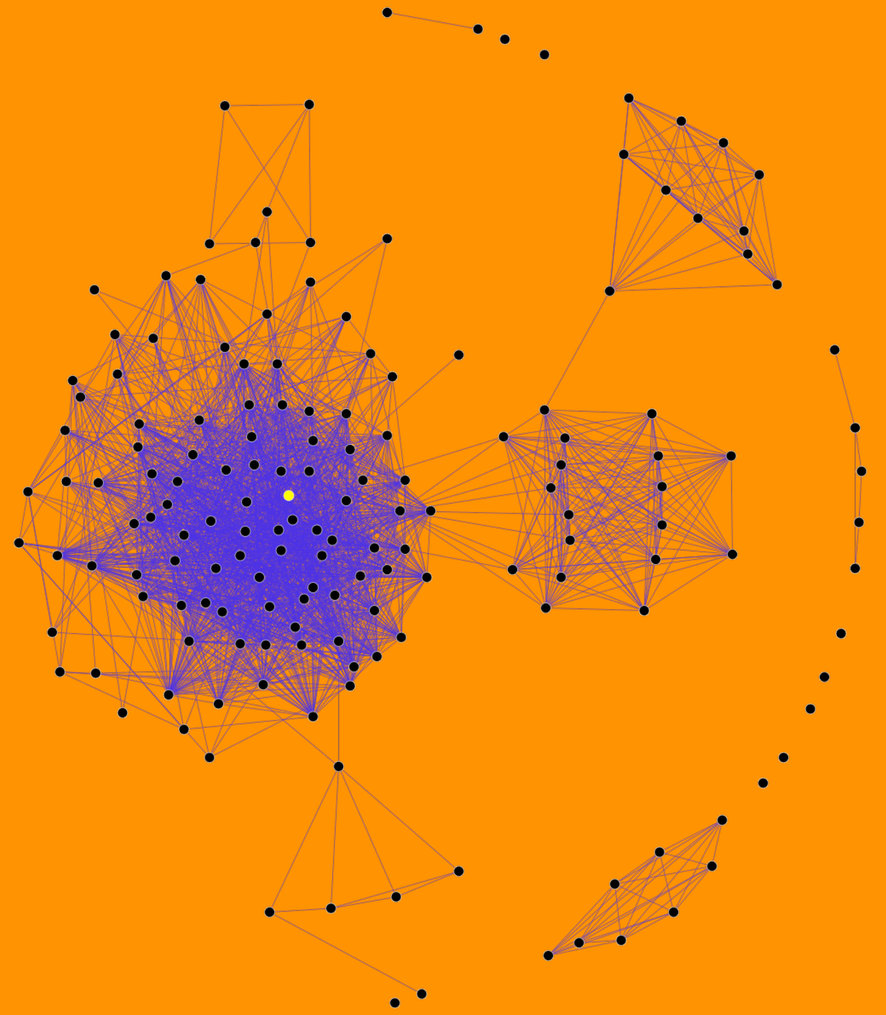
- Collections of items with each item having a link to one parent item (except the root)
- Represents hierarchical relationships between items
- Examples: file system, decision-making





Network


Open Only When Necessary
Push Over Poll
Power Save Mode
Power Awareness
Reduce Size
WiFi Over Cellular
Suppress Logs
Batch Operations
Cache
Decrease Rate
User Knows Best
Inform Users
Enough resolution
Sensor Fusion
Kill Abnormal Tasks
No screen interaction
Avoid Extra. Graph. & Anim.
Manual Sync - On Demand
Dark UI Colors
Dynamic Retry Delay
Avoid Extraneous Work
Race-to-idle

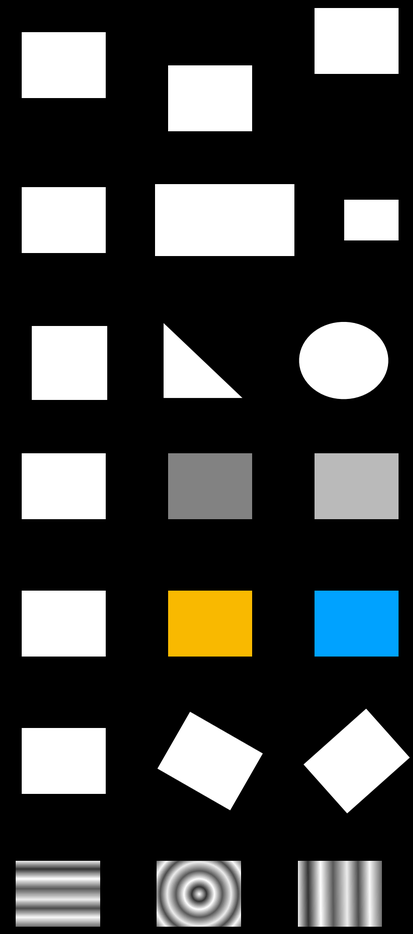


More ontologies

Bertin’s Original Retinal Variables

More ontologies

- **Position**. Changes in the x, y location
- **Size**. Change in length, area or repetition
- **Shape**. Infinite number of shapes
- **Value**. Changes from light to dark
- **Colour**. Changes in hue at a given value
- **Orientation**. Changes in alignment
- **Texture**. Variation in ‘grain’



Data types
More ontologies

- Numerical Data:
  - **Discrete**
  - **Continuous**
    - **Interval**
    - **Ratio**
- Categorical Data:
  - **Nominal**
  - **Ordinal**

Graph types
More ontologies

- Each graph type fits a particular visualization problems.
- A good visualization selects the best graph type according to the information it conveys.
- Examples of graph types: pie chart, scatter plot, line chart, bar plot, chord diagram, and so on.


MY FAVORITE BARS
(In Percentage of Awesomeness)
McGee's
MacLaren's
28%
30%
12%
27%
White Horse Tavern
3%
This is a pie chart
describing my favorite bars.


MY FAVORITE PIES
(In Percentage of Tastiness)
And this is a bar graph
describing my favorite pies.

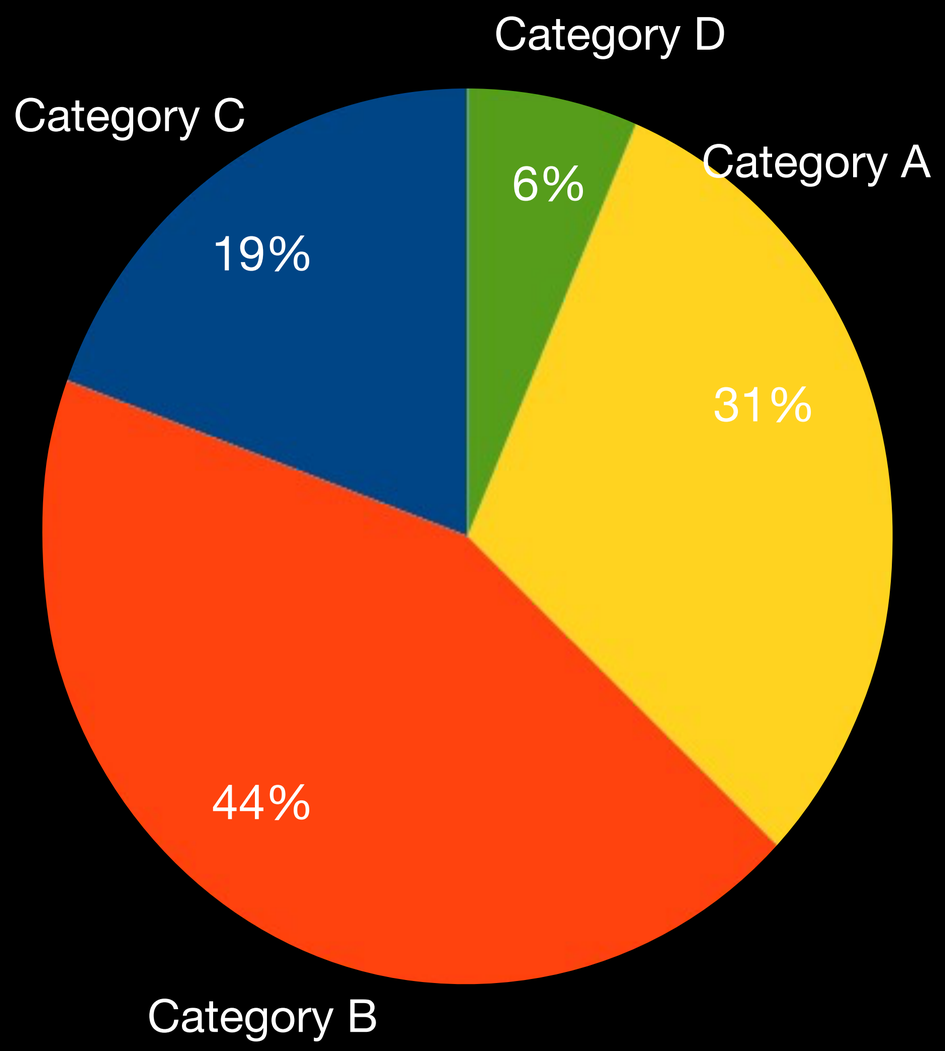
Analytical Framework to analyze graphs
(Homework)

- What **visual patterns** are used in the graph?
- What **data types** are represented in the visual patterns?
- What are some general **relationships between data points** (story) one can draw?
- What are some **pitfalls** of the graph?
- **How to address** the pitfalls?

Pie Chart

Graph types

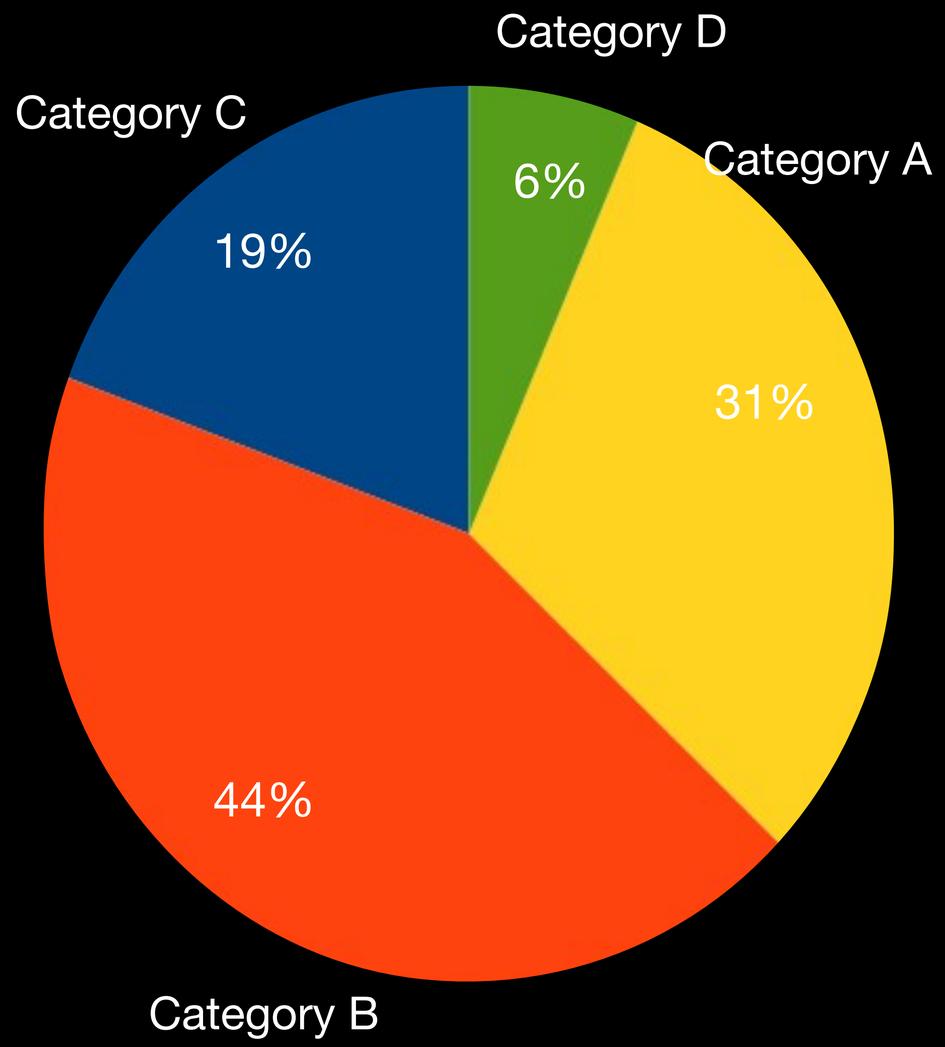
- Pie chart displays how the size of each category (by the size of the slice of pie) adds up to the whole population. It visualizes portions (or percentages) that sum up to 1 (or 100%).
- Visual pattern: **size (angle)** of the circle for the portion, and **color** for the category
- Data: Nominal with Ratio
- Annotation is important for pie chart!
- Ratio is normally displayed, if not combined with raw numbers.



Pie Chart (II)

Graph types

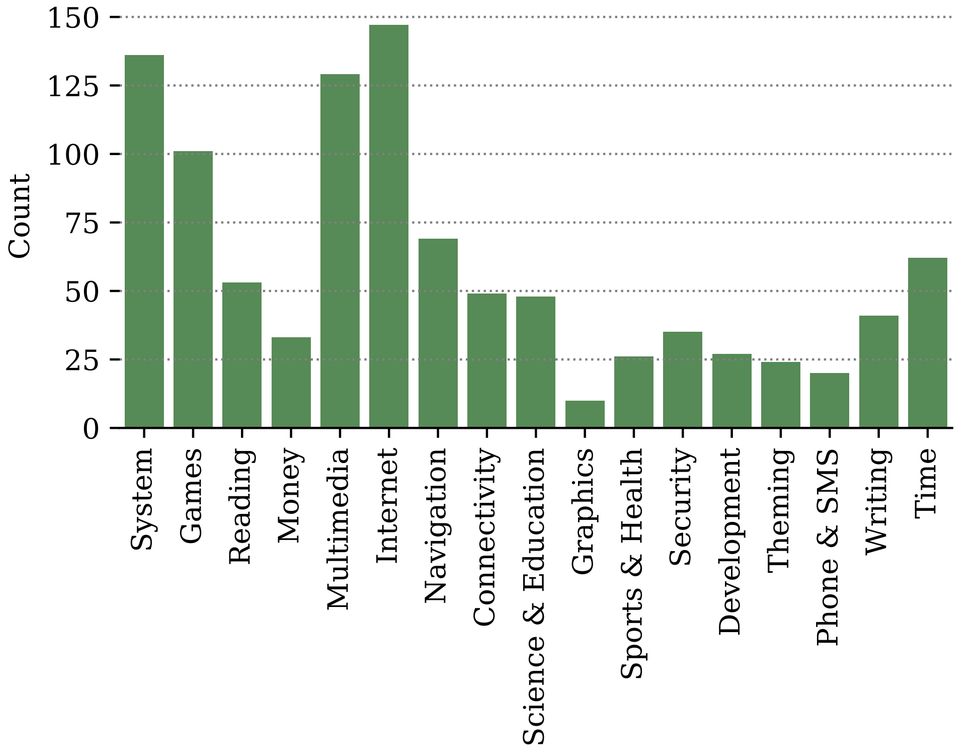
- Pie chart (or donut chart) is based on three assumptions:
  - Categories should be mutual-exclusive;
  - All categories of the population are ideally included;
  - Ratios of all categories should be added up to 100%.
- If any assumption is violated, the data is better visualized in bar chart.



Bar Chart

Graph types

- Bar chart shows the size of an observation:
  - **Size** (height) of the bar, **position** of the bar
  - Data: Categorical and Discrete
- Think about how bar chart should be used differently than pie chart.



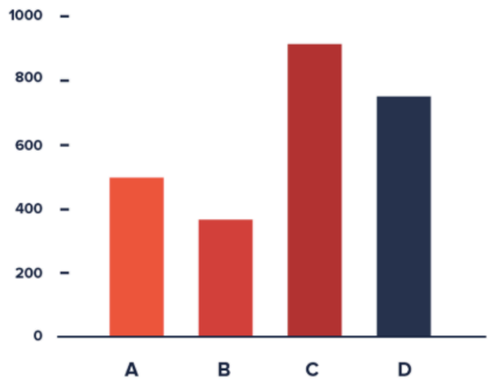
Bar Chart

Graph types

- Always start in the baseline. **Do not cut axes**:
  - Tufte principle: proportional representation of numbers!
  - This applies to nearly all graph types (1) with linear scale and (2) the proportion of numeric values bears meanings.
- Avoid using 3-D representations when not necessary.

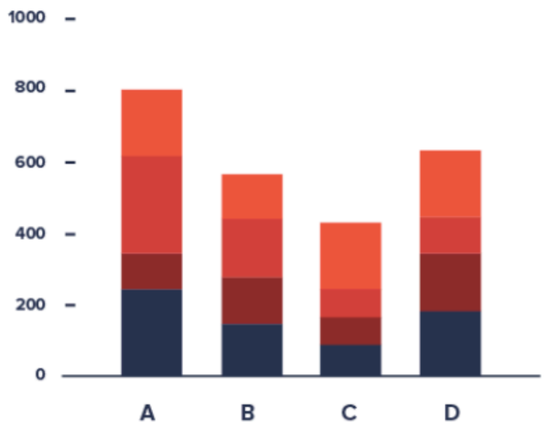
Grouped, Stacked and Opposite bar charts

Bar Chart (vertical)

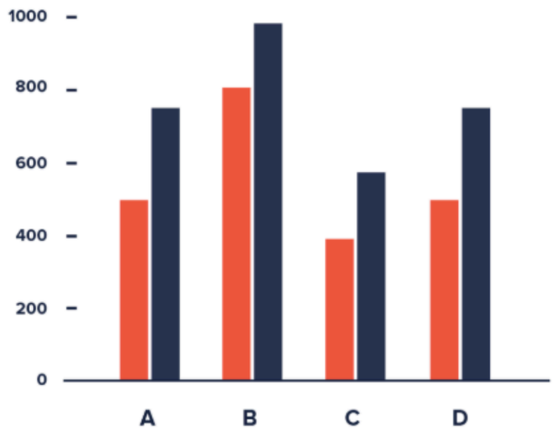


What different stories can the latter three graphs tell?

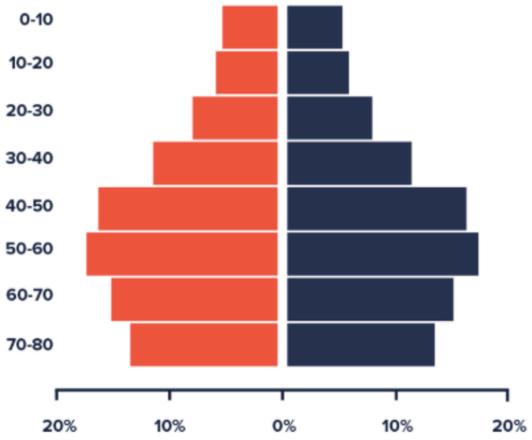
Stacked Bar Chart



Grouped Bar Chart

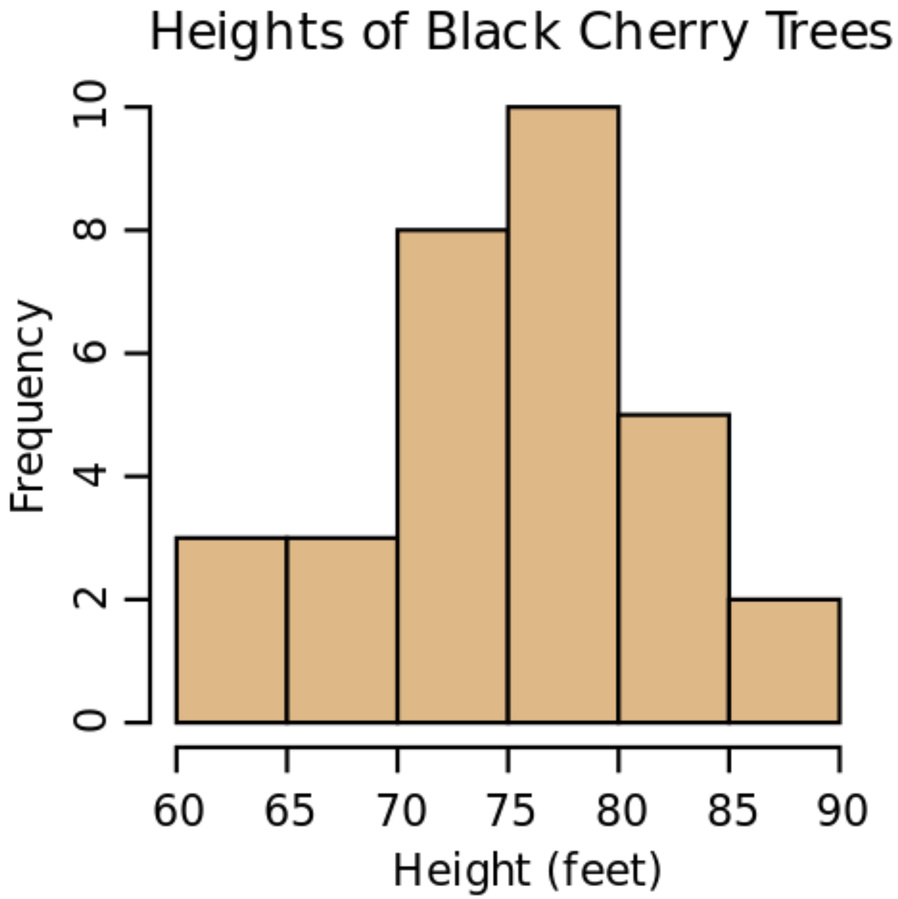


Population Pyramid



Histogram
Chart types

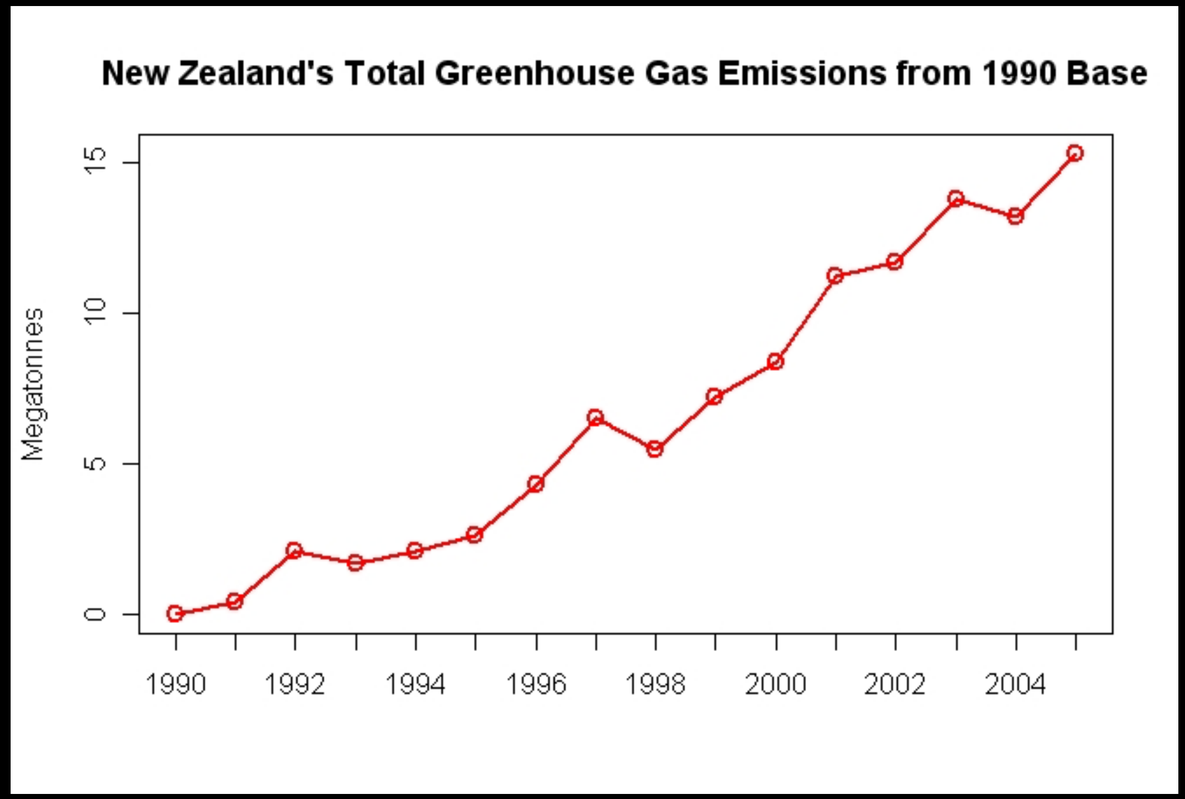
- Special case of bar chart
- Represent the distribution of numerical data
- **Data type:** It maps continuous values into a series of bins — i.e., **intervals**.
- Each bin is represented with a bar. The height of the bar is proportional to the number of data points in that bin.
- Main pitfall: the selection of bins can leave out important information, or make the visualization too complex.



Line Chart

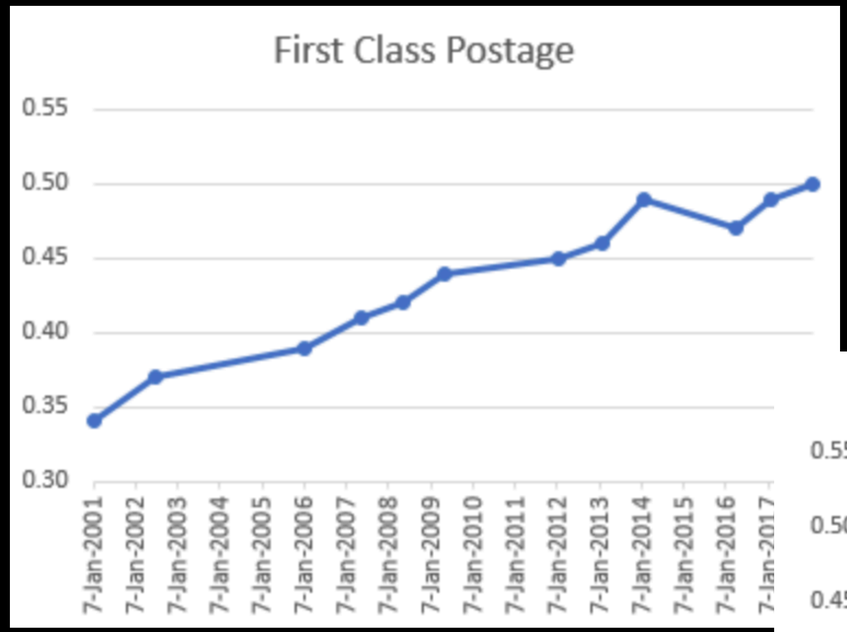
Graph types

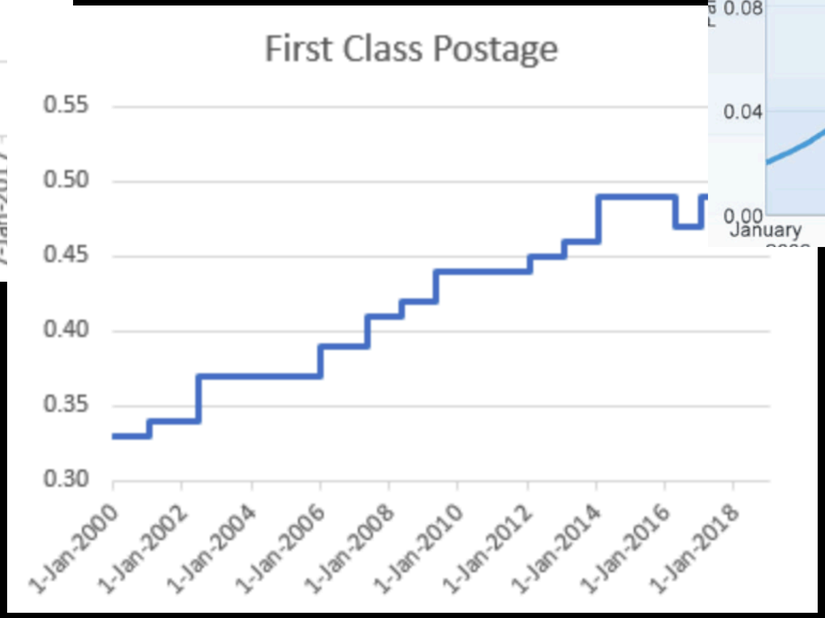
- Line chart uses the line (connected dots) to draw the pattern of an observation or several observations over timeline or some other categories.
  - Visual pattern: **position** in x- and y-axes…
  - Data:
    x – datetime, numerical (or other ordinal data types)
    y – numerical
- For two-dimensional visualization timeline data is conventionally plotted on the x-axis.

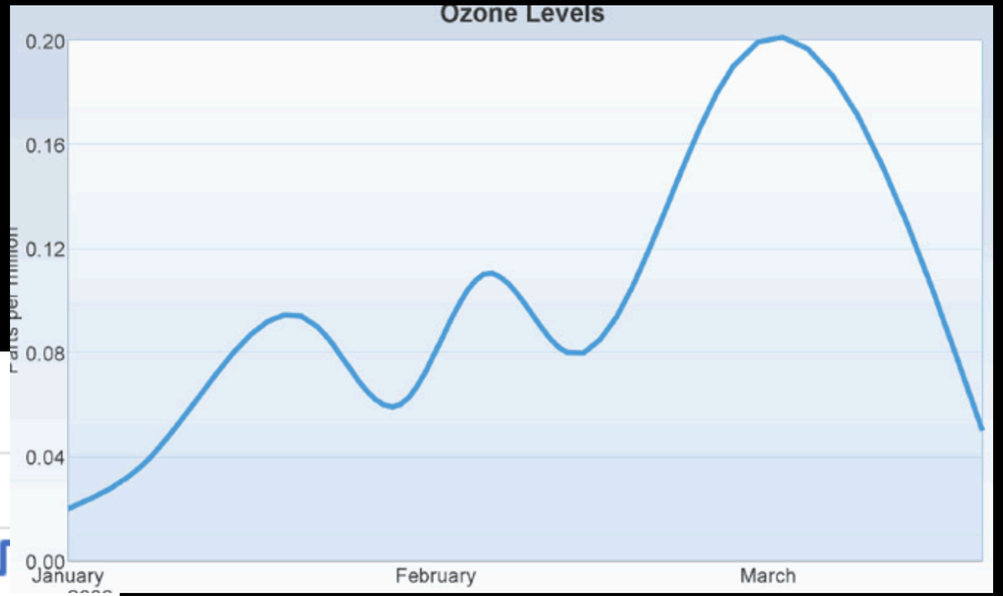


Stepped, curved, and dotted line charts

Chart types



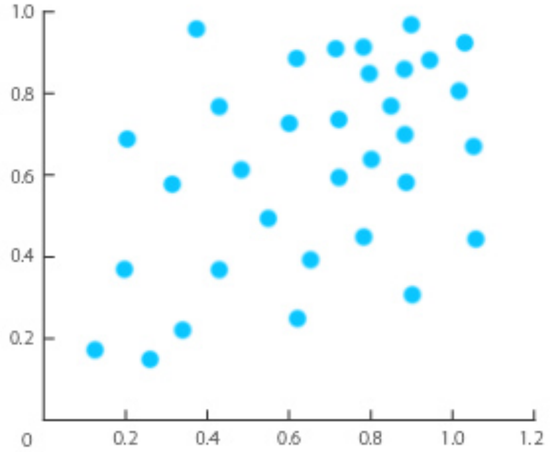


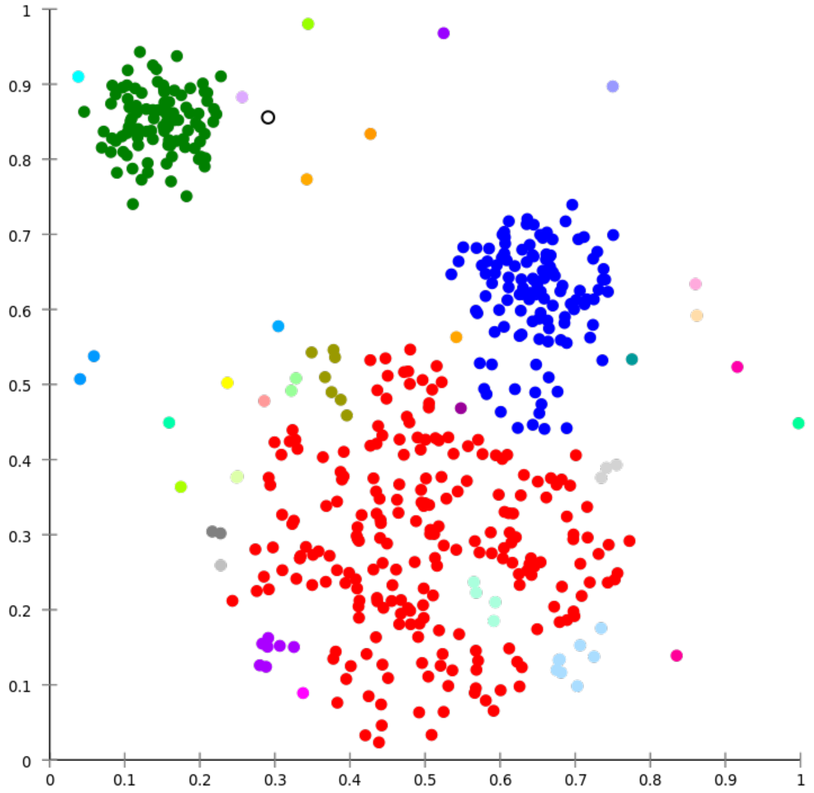


Scatter plot

Graph types

- Scatterplot plots two aspects of objects in a two-dimensional space.
- Visual patterns: **position** in x- and y-axes…
- Data: In most cases, both axes should be mapped to numerical data.*
- Scatterplot is frequently used in correlation and clustering analyses.

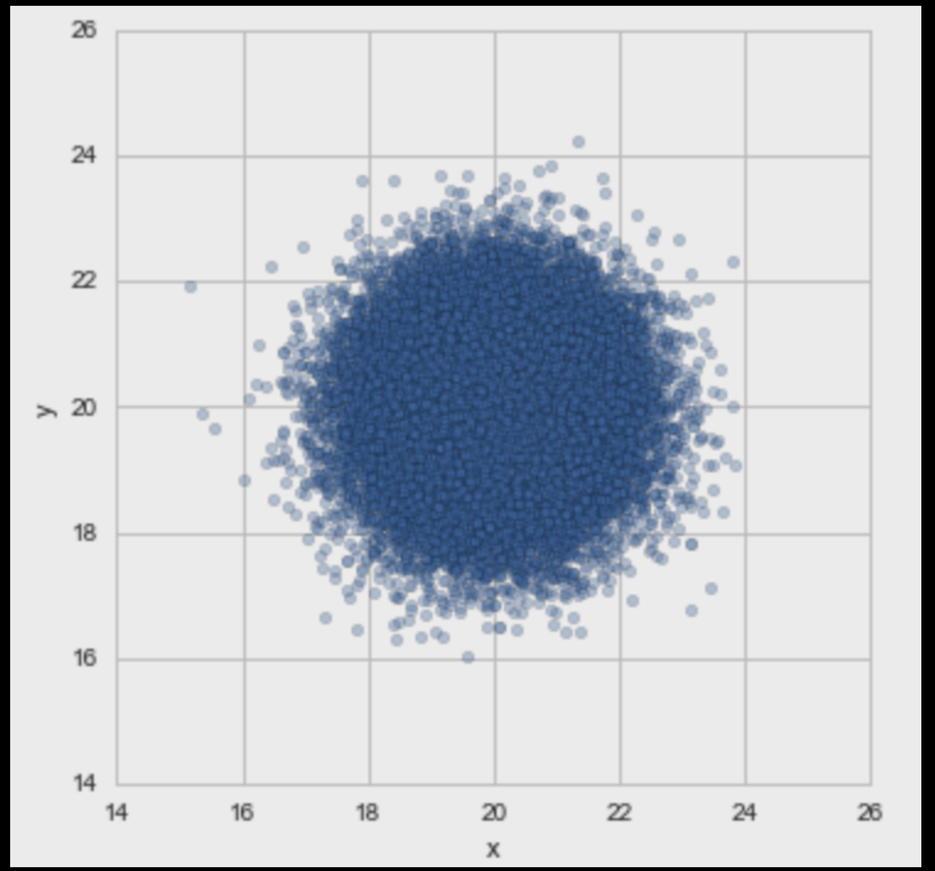




Scatter plot

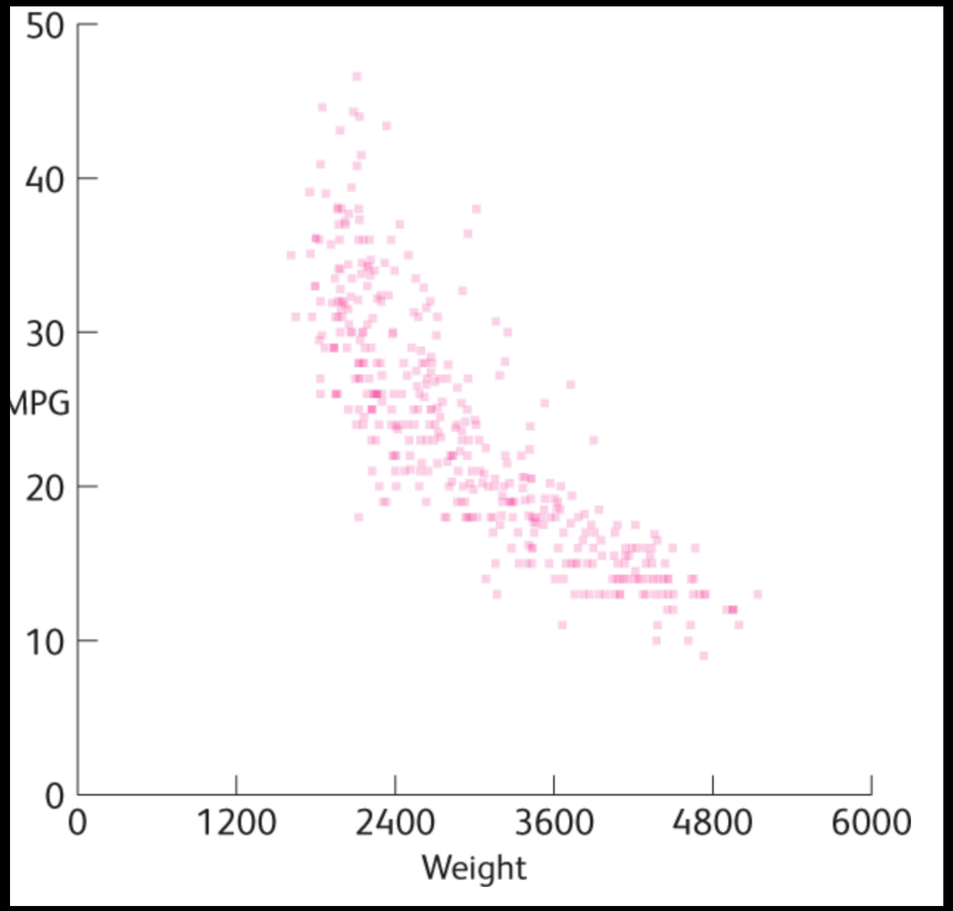
Chart types

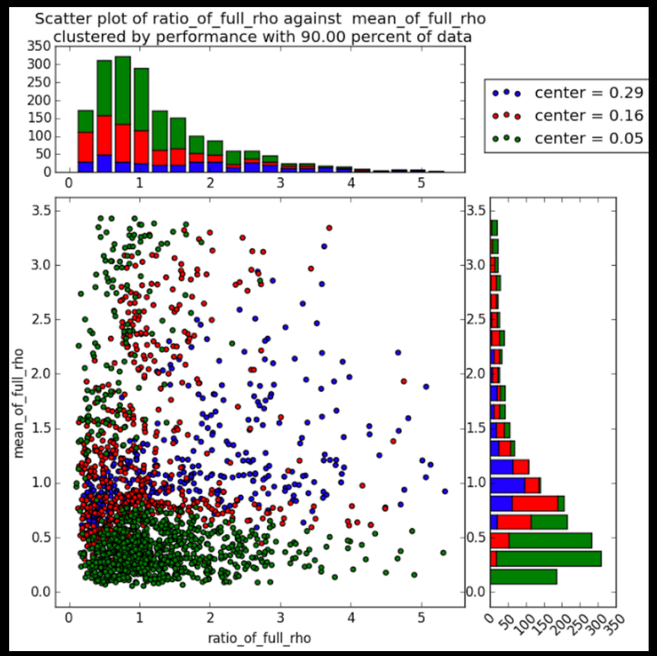
- As compared to other two-dimensional graphs, scatterplot can represent much more information, especially more observations.
- However, this often causes the problem of information overload, which makes it difficult to understand the data.
- Some solutions to information overload in scatterplot:
  - Use transparency in the points.
  - Use extra graphs to offer the summary of values.
  - Use **small multiples**.
  - Scatterplot may not be the correct answer to all data!
  - Such as data with too many entities and data variables with very uneven distributions.

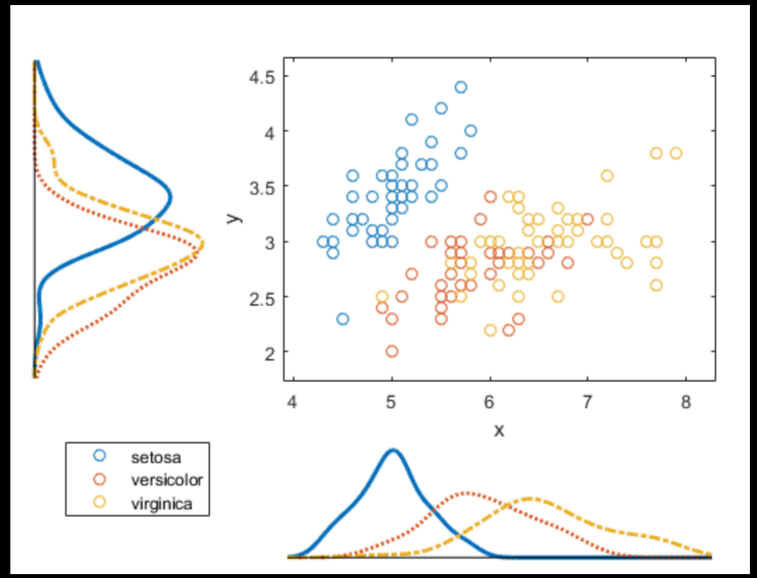


Scatter plot: solutions

Chart types



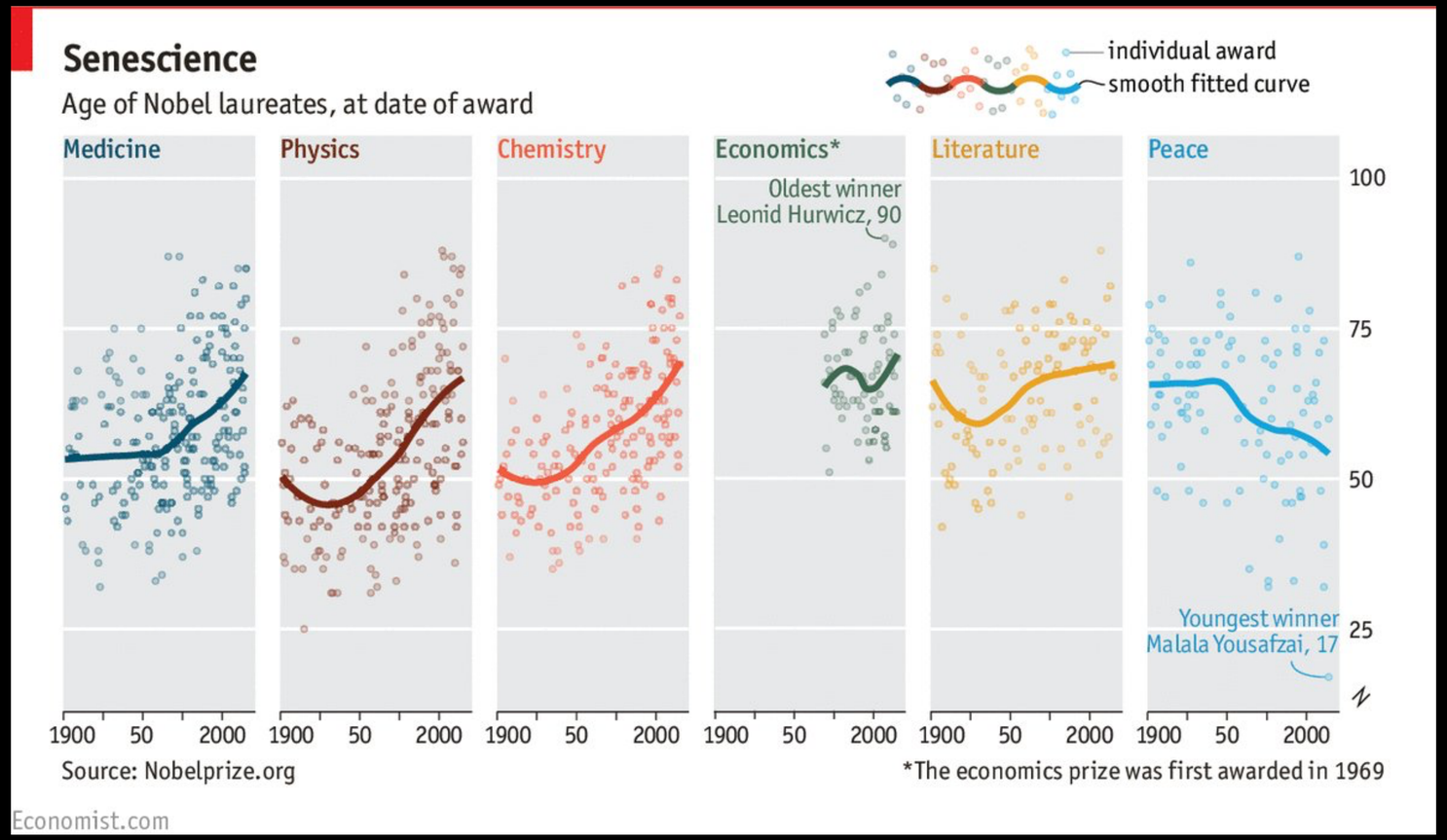




Small Multiples

Graph types

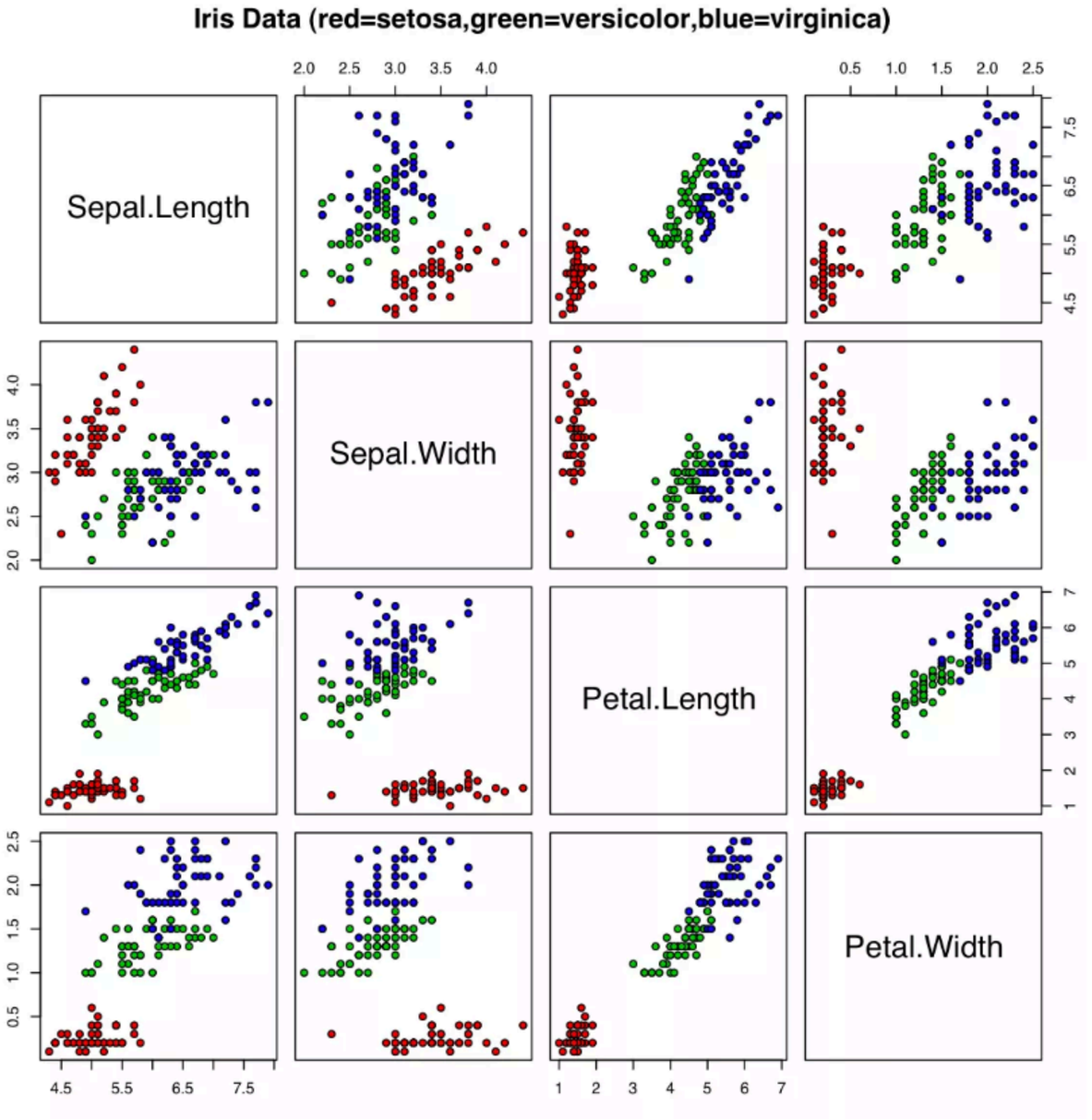
- “A small multiple (sometimes called trellis chart, lattice chart, grid chart, or panel chart) is **a series of similar graphs** or charts using the same scale and axes, allowing them to be **easily compared**.” https://en.wikipedia.org/wiki/Small_multiple



Small Multiples

Graph types

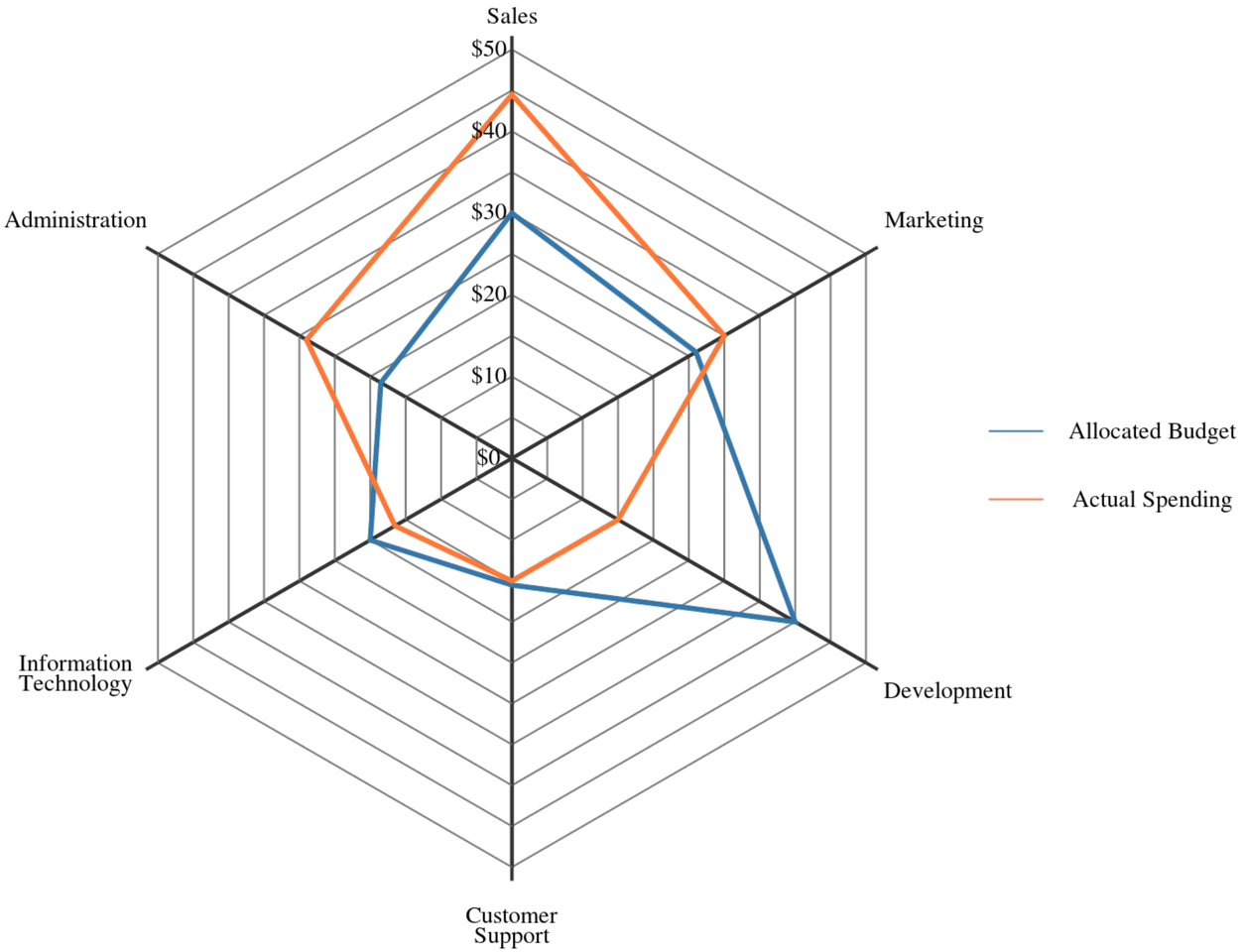
- Small multiples, as a meta-type of graph, can be applied to different visualization types, such as scatterplot, line chart, and bar chart.
- Besides making inter-group comparisons, another use of small multiples is to give a comprehensive summary of the dataset.



Radar Chart

Graph types

- Useful to show data with more than 3 dimensions (e.g., relational databases)
- Two questions can be answered by radar chart:
  - How individuals are compared in all parameters in the graph?
  - How is one's overall performance?
- What are the cons of this graph?



Assignments

- Assignment 5
- https://luiscruz.github.io/course_infovis/assignments/assignment5
- Tutorial 2
- https://luiscruz.github.io/course_infovis/Tutorials/tutorial2

Assignment 5
YAML

- Considerable improvement on YAML compliance.
  - Common mistakes:
    - Wrong file name. “Assignment5” is not the requested filename.
    - Student ID should not have “less than” < and “more than” > symbols.
    - Multiple-line answers must have quotation marks.

Assignment 5
Visual Variables

- **Visual variable** is different than (data) **variable.**
- Refer to **Bertin’s Original Retinal Variables**.
  - Shape, size, colour, position, value, orientation, texture

Assignment 5

Data types

- **Data type** is different from (data) **variable**
- Check **Data Type Ontology**.

Repetition

Data types

More ontologies

- Numerical Data:
  - **Discrete {1,2,3,4,5,…}**
  - **Continuous {-2, -1, 0, 1.3, 1.4, 1.1113, 2, e, π, 5.612314542}**
  - **Interval {[1, 3], [3, 5], [5, 6]** (special type of continuous)
  - **Ratio** (special type of continuous)
- Categorical Data:
  - **Nominal (Blue, Yellow, Green)**
  - **Ordinal (Elementary, High-school, Undergraduate, Graduate)**

Repetition

Data types
More ontologies

- Numerical Data:
  - **Discrete {1 2**
  - **Con**

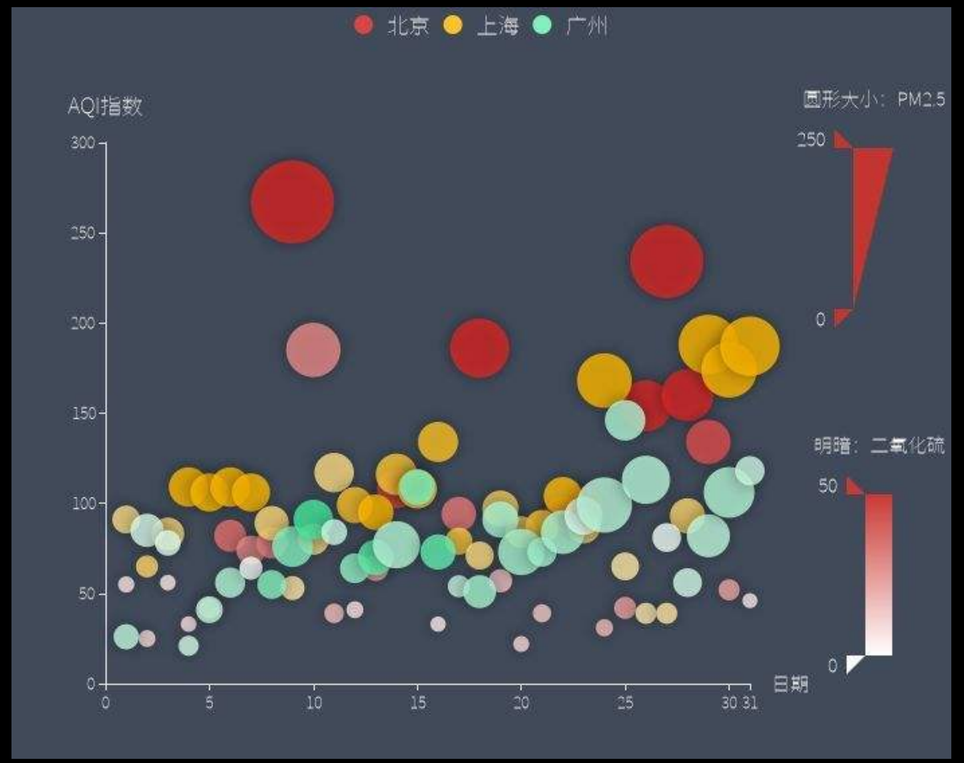
Too basic

ol, Undergraduate, Graduate)

More charts

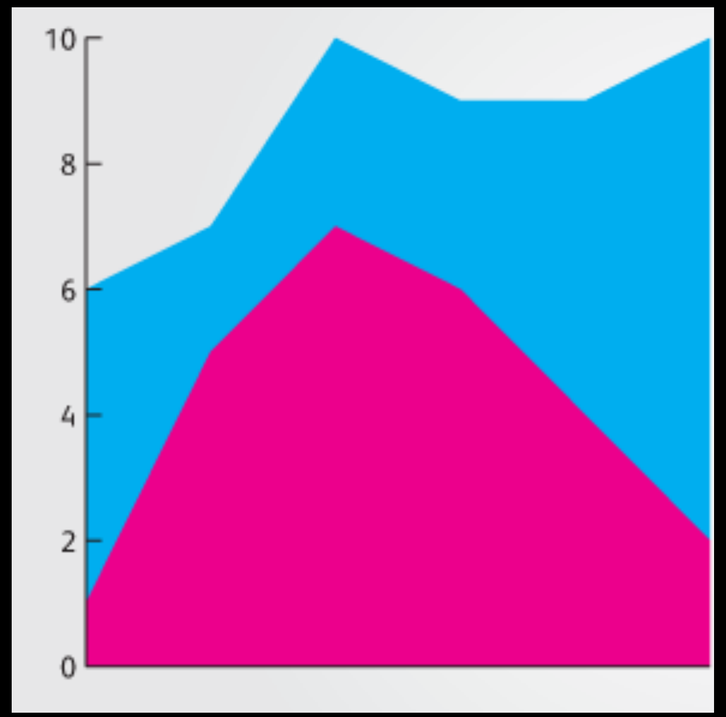
Bubble Chart

- Special type of scatter plot
- Visual variables: **position** and **size.** (Colour is optional)
- Data type: **continuous**
- Relationship: **correlation between** 3 variables: x, y, and another represented by the size of the bubble
- Pitfalls: bubbles can **hide each other** when they are close to each other. This is emphasised with big circles.
- Solution to pitfalls: reduce **opacity** of circles.



Area Chart

- Based on line chart.
- Fill the area under the curve.
- Visual variables: **position** and **size** and **colour** (or shade, or texture).
- Data type: **numerical** in y-axis and **numerical** or **ordinal** in x-axis.
- Relationship: part-to-whole
- Pitfall: The irregular shape do not give a clear notion of the differences between the areas of the two groups.
- Solution: add labels with the value of the areas; or add a label with the part-to-whole ratio.



Area Chart

- What is the difference with a line plot??

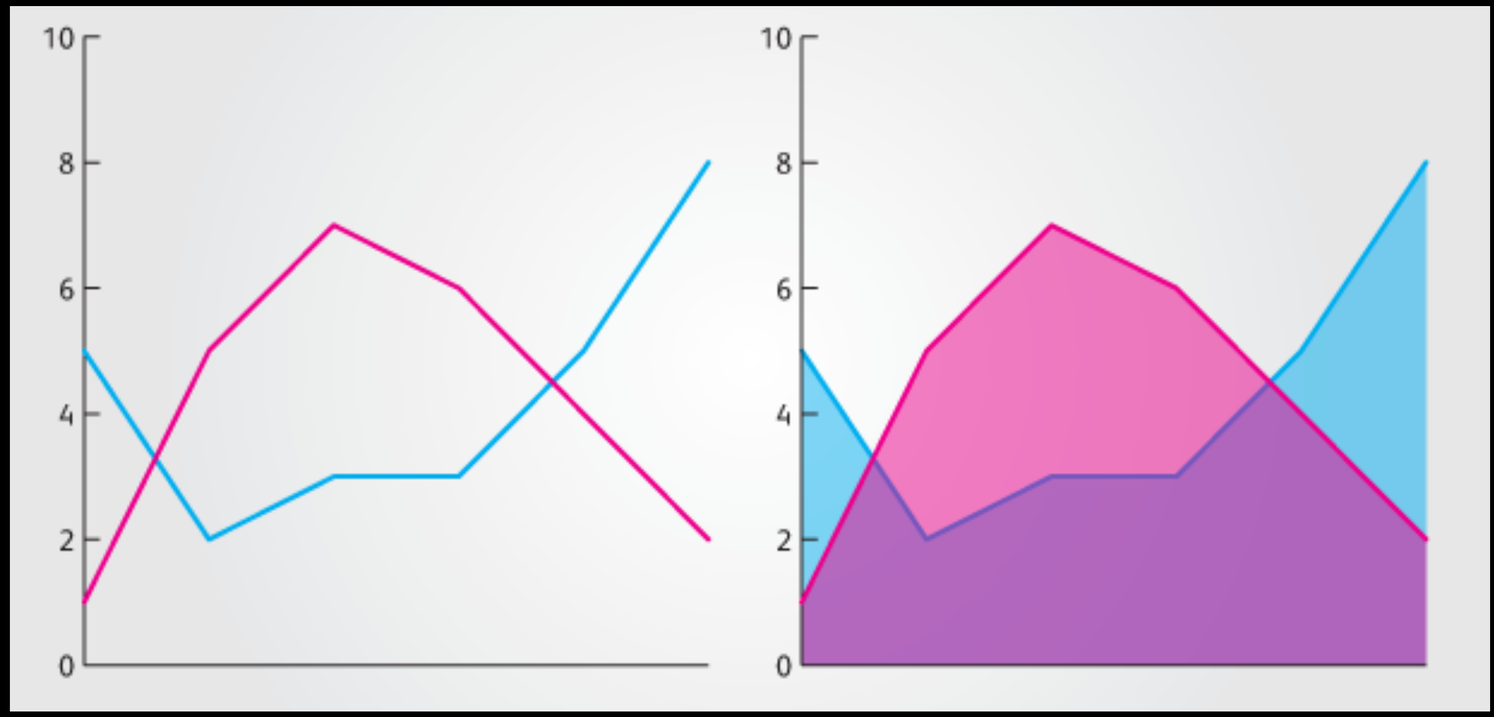


Image from https://en.rockcontent.com/blog/line-vs-area-charts/

Area Chart

- What is the difference with a line plot??

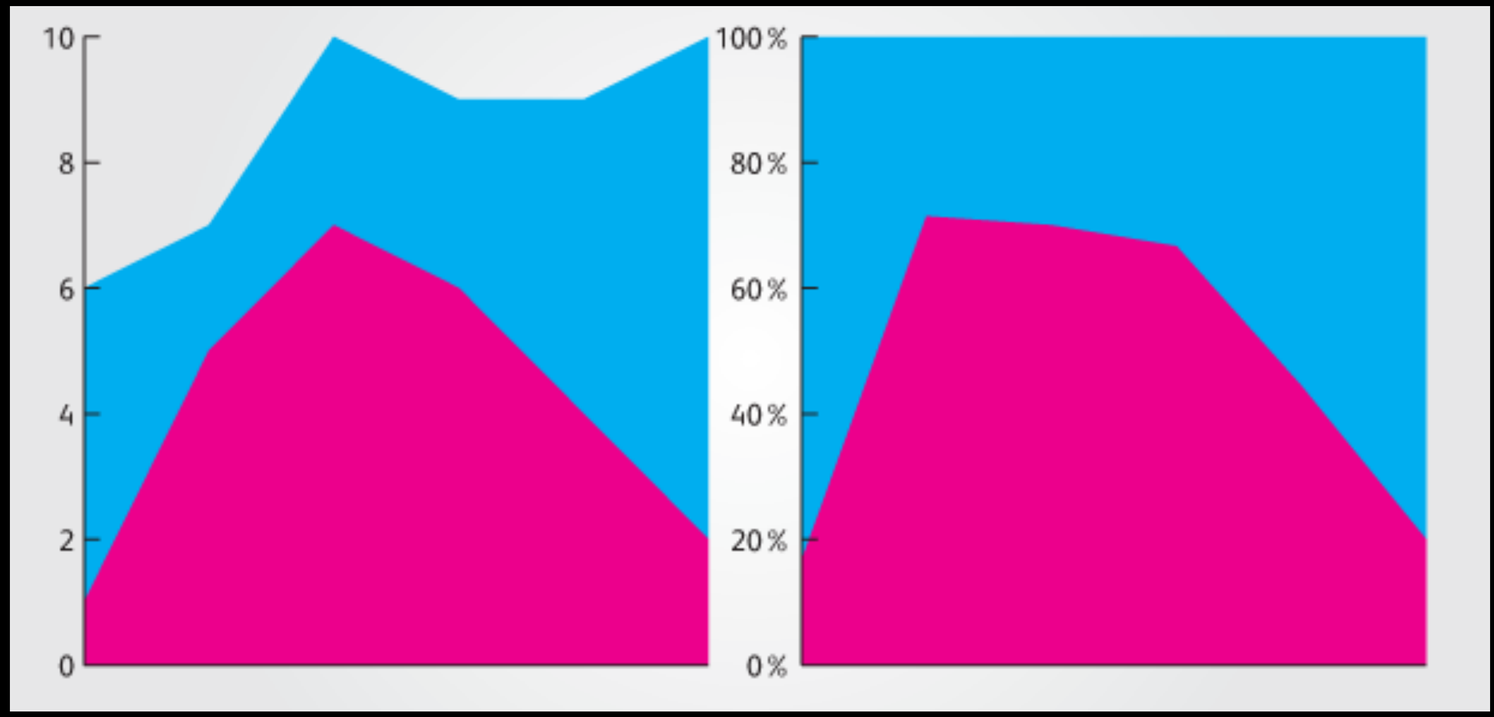
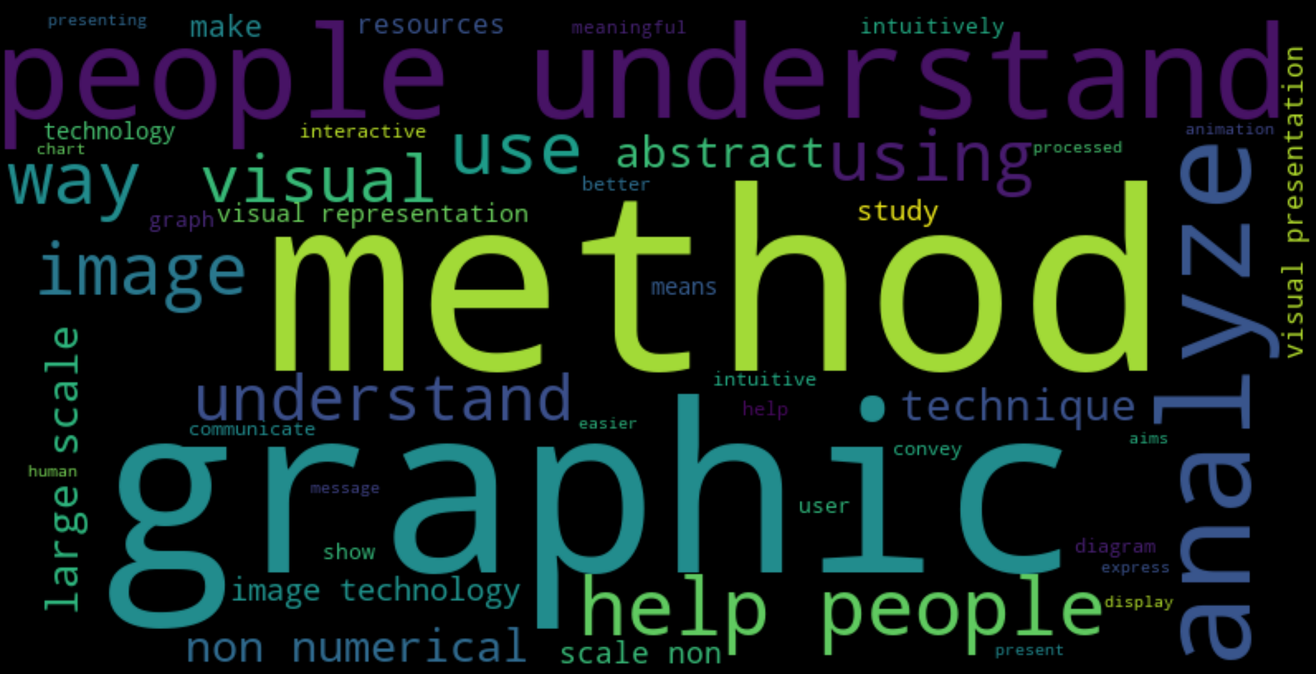


Image from https://en.rockcontent.com/blog/line-vs-area-charts/

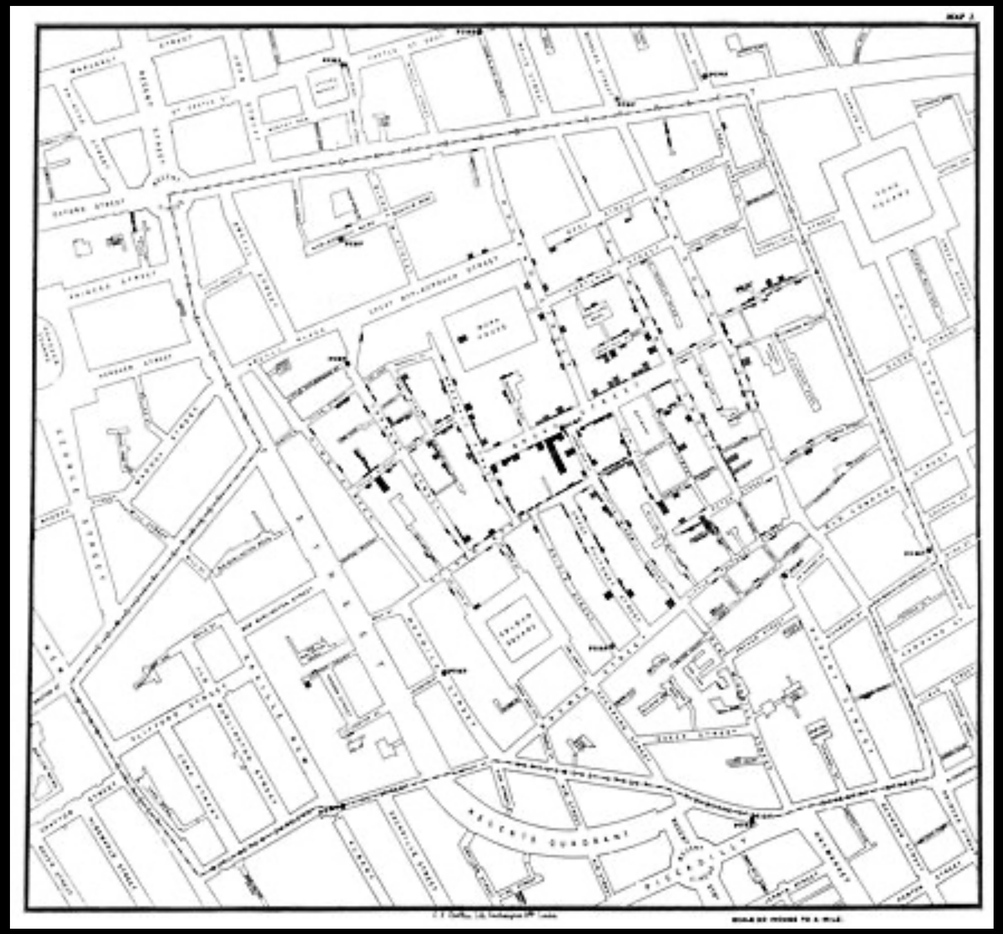
Word cloud

- A.k.a **wordle**, **tag cloud**.
- Visual variables: only **size**. Color and position do not represent any change in data.
- Data types: **text**



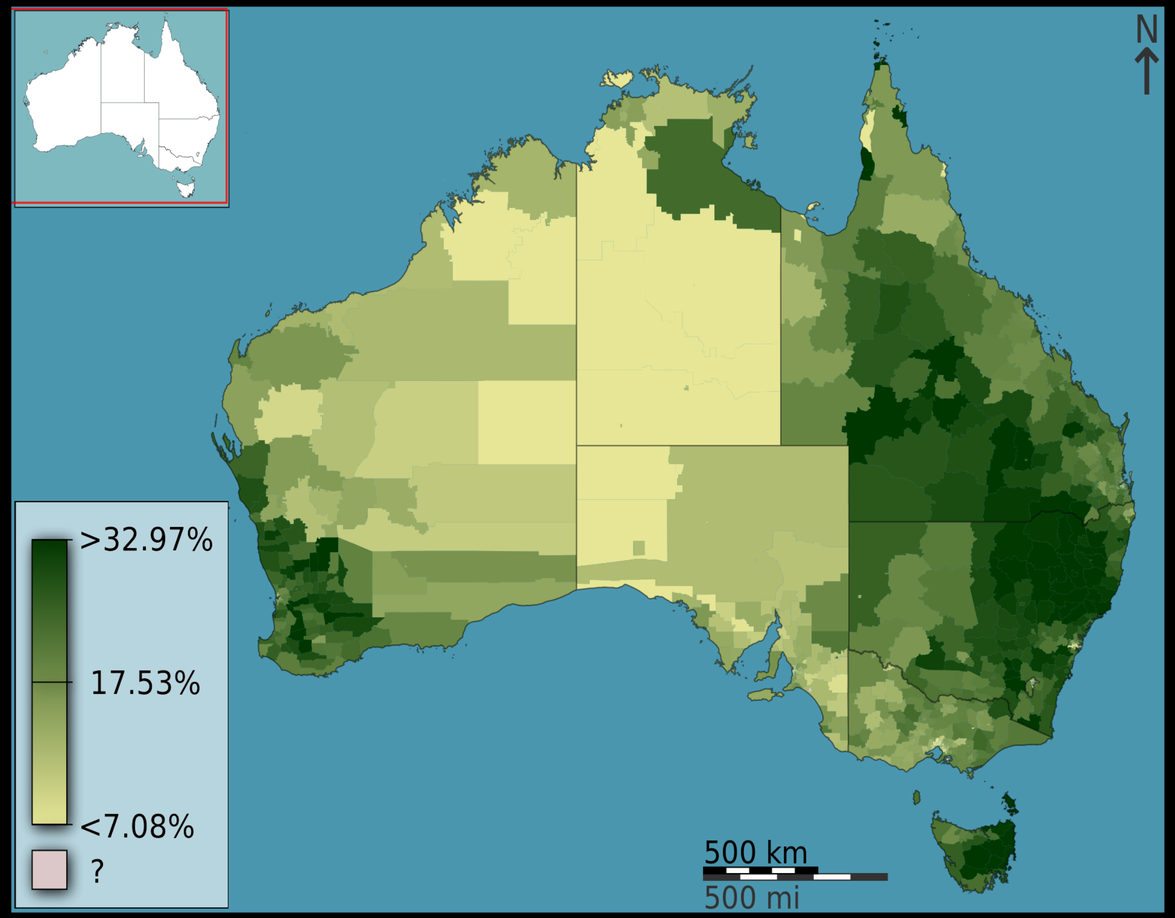
Thematic map

- Visual variables: **position**
- Data type: geospatial
- Relationship: geographical.
- Pitfall: Some areas in the map maybe overwhelmed with too mach data points while others may be to sparse.
- Solution: Provide interactive tools to zoom in and zoom out.
- Special types of thematic maps: **Choropleth map**, **Bubble map**



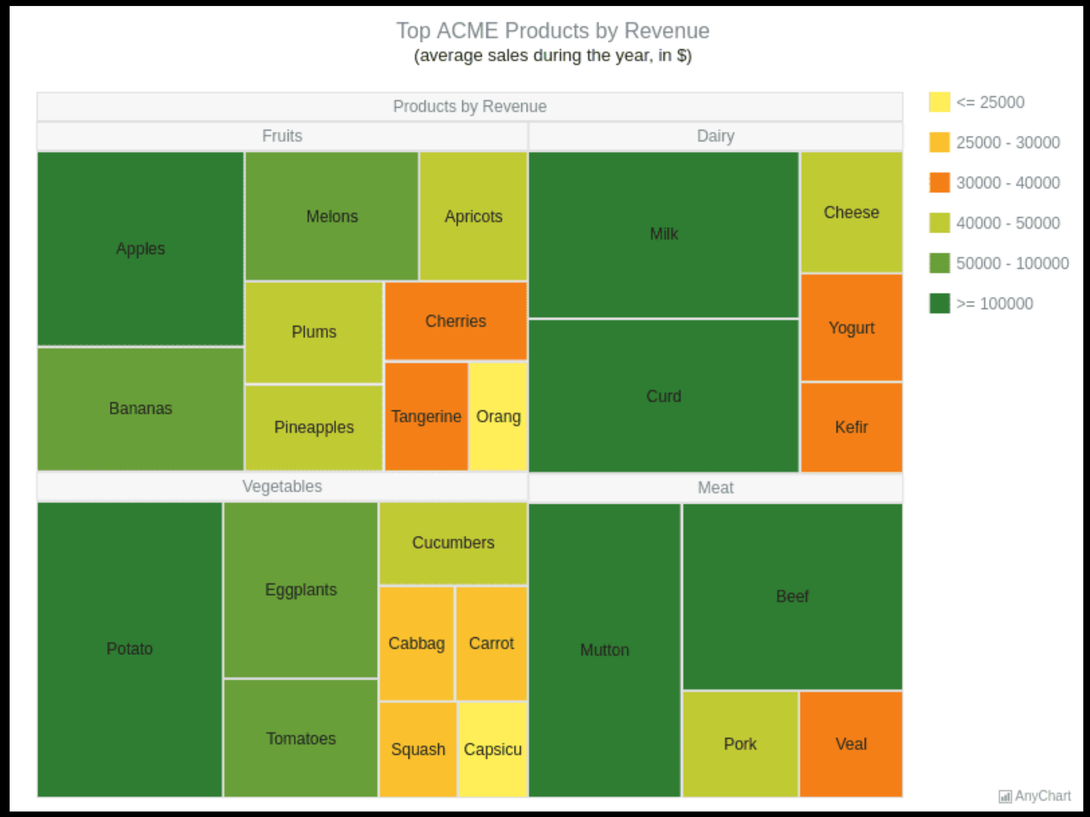
Choropleth map

- Visual variables: **position, color.**
- Data type: geospatial, numerical
- Relationship: geographical.
- Pitfall: Some areas may look bigger because of the size of the geographical region but that may not be related with the data being presented. E.g., size of population is different than size of the country.
- Solution: Use **bubble map**.



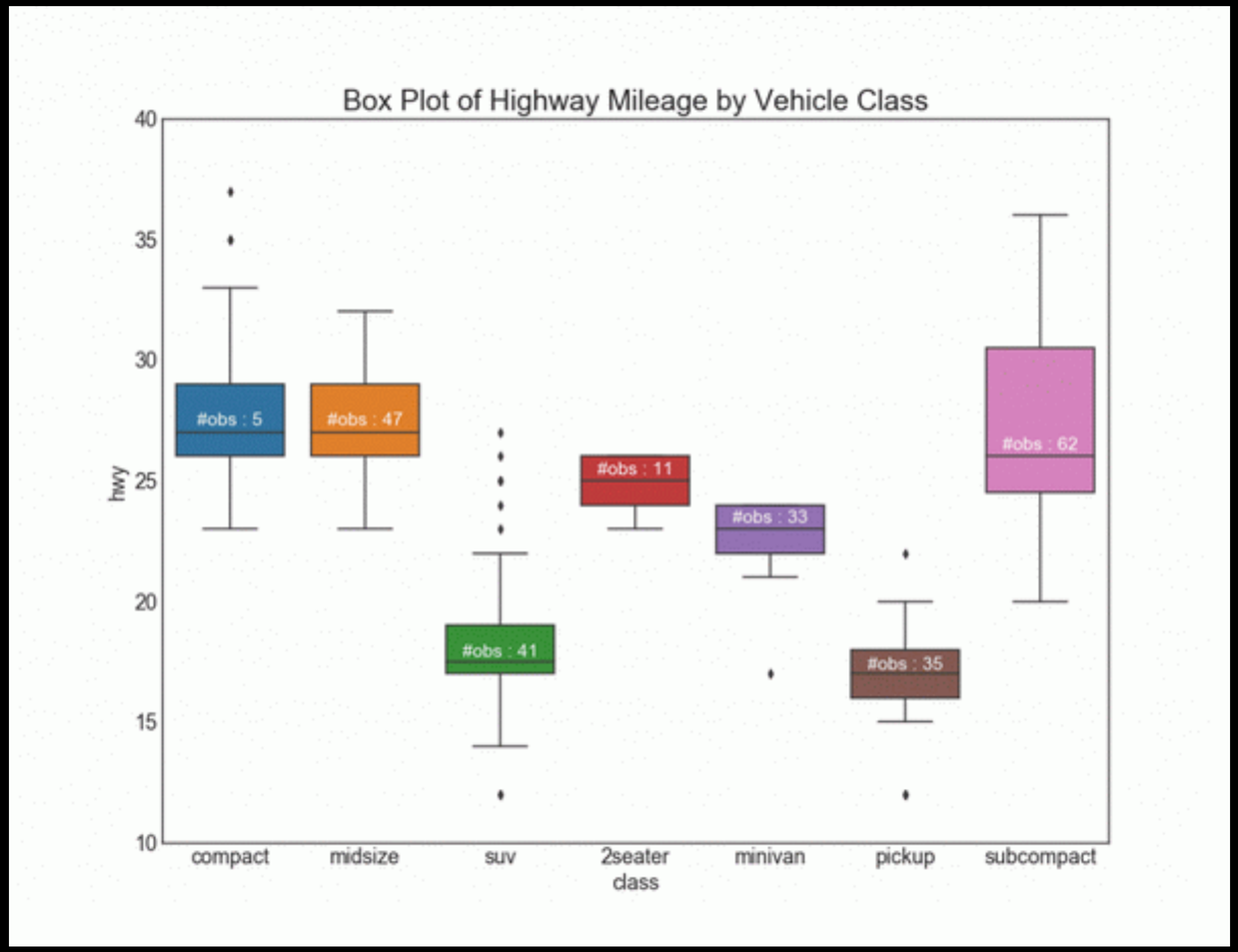
Treemap

- Visual variable: Size. (**Color** is optional)
- Data type: **Nominal**, *Abstract Data Type - tree*
- Relationship: hierarchical relationship, proportion between different groups in the same hierarchical level.
- Pitfalls: it is not straightforward to compare areas.
- Solution: use a **bar plot** to show the size of the groups and a **tree graph** to show hierarchy.



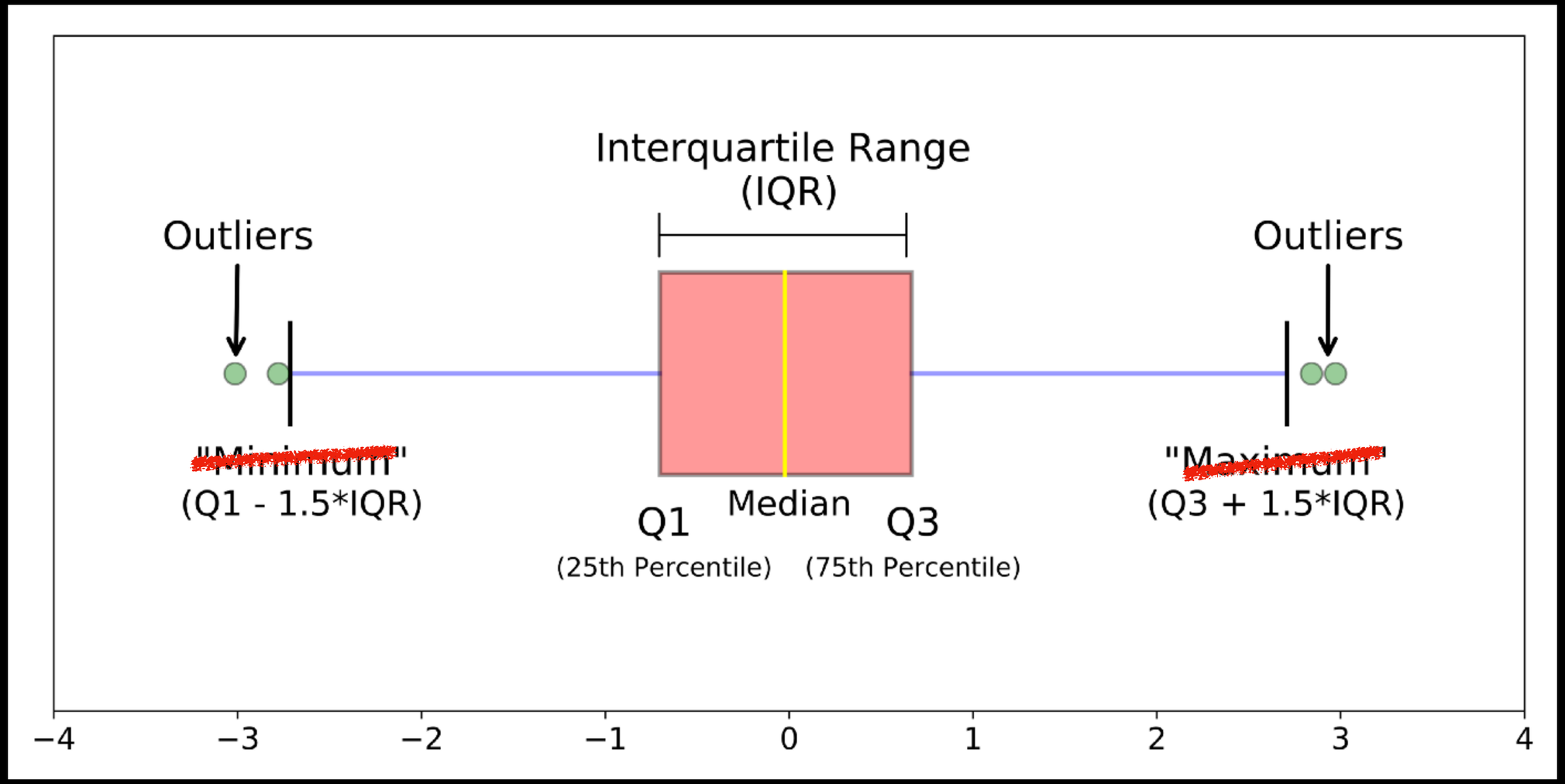
Box Plot

- A.k.a **box and whisker plot**
- Visual variable: **size** and **position**.
- Data Type: **continuous** (y) and **categorical** (x)
- Relationship: difference between **distributions**
- Pitfalls: may hide important information if the distribution is not Normal.
- Solution: Use **violin** plots when the shape of the distribution is unknown.
- Variations: Notched Box Plots; variable-width Box



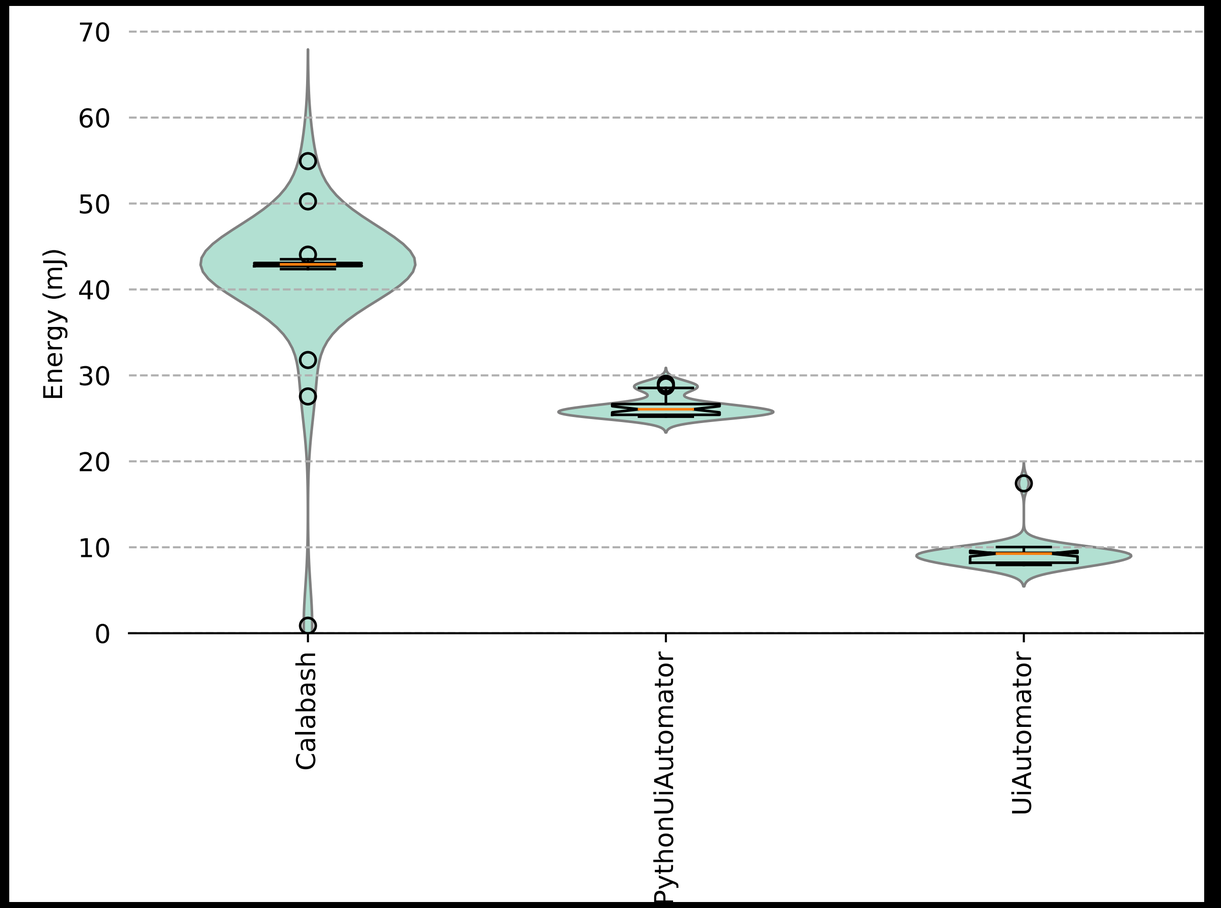
Box Plot

In detail



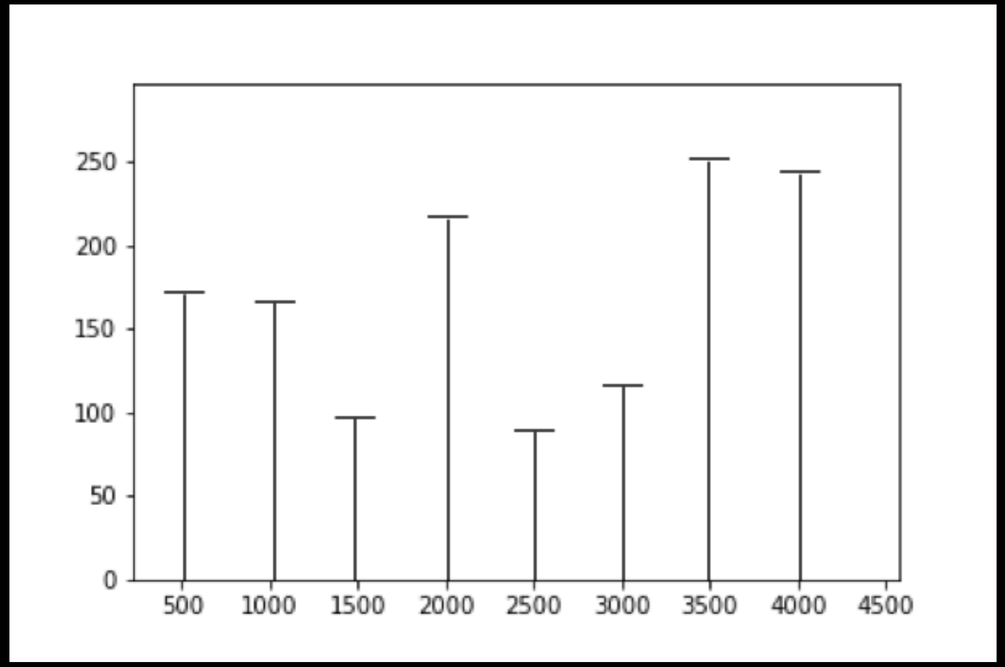
Violin Plot

- Visual variable: **size, position, shape**.
- Data Type: **continuous** (y) and **categorical** (x)
- Relationship: difference between the shape of **distributions**
- Pitfalls: do not show descriptive statistics to compare distributions. Harder to compare.
- Solution: Combine **box plots with violin plots**.



Assignment 6

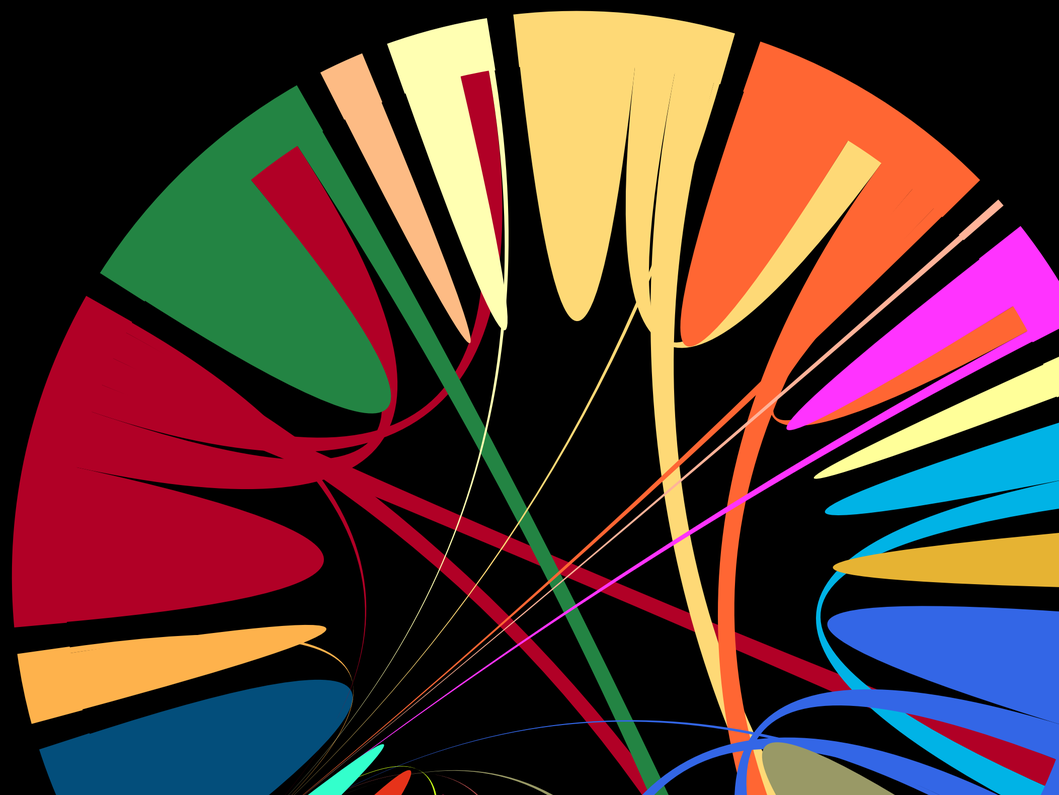
- https://luiscruz.github.io/course_infovis/assignments/assignment6
- “Skinnybar” plots



Information Visualization INFO250

Class 7:
- Ontologies based on User Task
- Unwanted biases
- Assignment 7

Luís Cruz - l.cruz@tudelft.nl



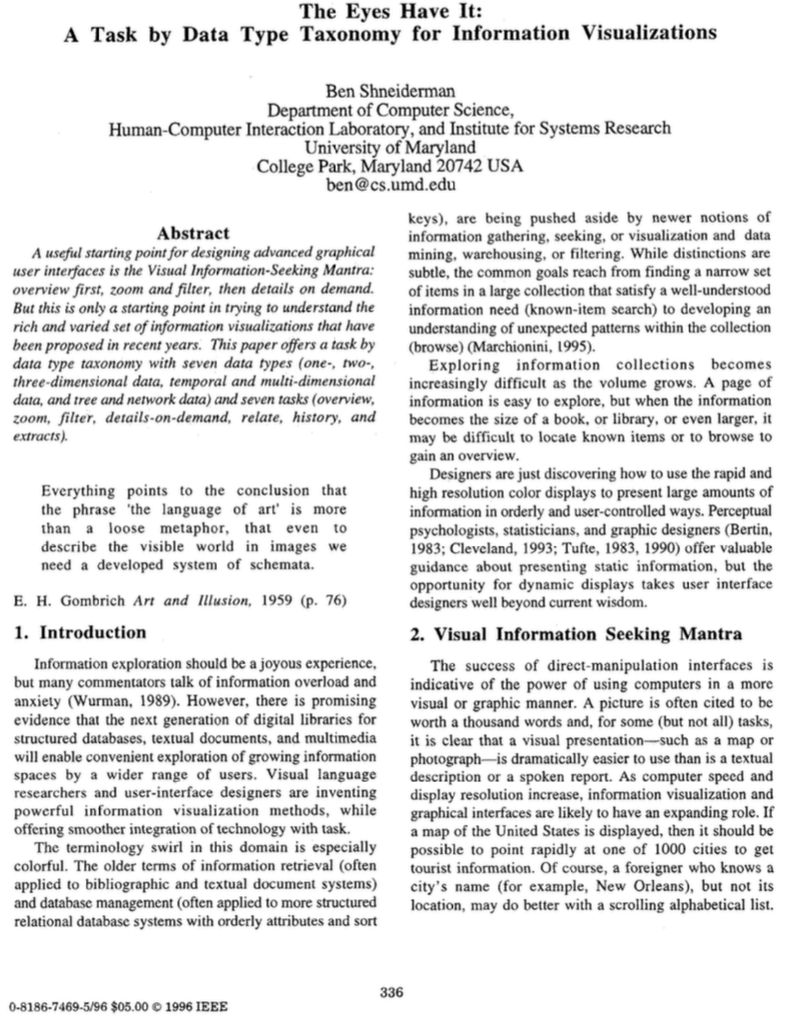
Previous Week

- Ontology of visualizations based on data type
- Ontologies of data types and retinal/visual variables
- Different types of visualizations
- Assignment 6 - Skinny plots

Ontology of visualizations based on User Task

- Classification of visualizations according to **user task**:
  - **Overview.** Gain an overview of the entire collection.
  - **Zoom.** Zoom in on items of interest.
  - **Filter.** Filter out uninteresting items.
  - **Details-on-demand.** Select an item or group and get details when needed.
  - **Relate.** View relationship among items.
  - **History.** Keep a history of actions to support undo, replay, and progressive refinement.
  - **Extract.** Allow extraction of sub-collections and of the query parameters.

By Shneiderman (1996)



The Eyes Have It:
A Task by Data Type Taxonomy for Information Visualizations

Ben Shneiderman
Department of Computer Science,
Human-Computer Interaction Laboratory, and Institute for Systems Research
University of Maryland
College Park, Maryland 20742 USA
ben@cs.umd.edu

Abstract

A useful starting point for designing advanced graphical user interfaces is the Visual Information-Seeking Mantra: overview first, zoom and filter, then details on demand. But this is only a starting point in trying to understand the rich and varied set of information visualizations that have been proposed in recent years. This paper offers a task by data type taxonomy with seven data types (one-, two-, three-dimensional data, temporal and multi-dimensional data, and tree and network data) and seven tasks (overview, zoom, filter, details-on-demand, relate, history, and extracts).

Everything points to the conclusion that the phrase 'the language of art' is more than a loose metaphor, that even to describe the visible world in images we need a developed system of schemata.

E. H. Gombrich *Art and Illusion*, 1959 (p. 76)

1. Introduction

Information exploration should be a joyous experience, but many commentators talk of information overload and anxiety (Wurman, 1989). However, there is promising evidence that the next generation of digital libraries for structured databases, textual documents, and multimedia will enable convenient exploration of growing information spaces by a wider range of users. Visual language researchers and user-interface designers are inventing powerful information visualization methods, while offering smoother integration of technology with task.

The terminology swirl in this domain is especially colorful. The older terms of information retrieval (often applied to bibliographic and textual document systems) and database management (often applied to more structured relational database systems with orderly attributes and sort keys), are being pushed aside by newer notions of information gathering, seeking, or visualization and data mining, warehousing, or filtering. While distinctions are subtle, the common goals reach from finding a narrow set of items in a large collection that satisfy a well-understood information need (known-item search) to developing an understanding of unexpected patterns within the collection (browse) (Marchionini, 1995).

Exploring information collections becomes increasingly difficult as the volume grows. A page of information is easy to explore, but when the information becomes the size of a book, or library, or even larger, it may be difficult to locate known items or to browse to gain an overview.

Designers are just discovering how to use the rapid and high resolution color displays to present large amounts of information in orderly and user-controlled ways. Perceptual psychologists, statisticians, and graphic designers (Bertin, 1983; Cleveland, 1993; Tufte, 1983, 1990) offer valuable guidance about presenting static information, but the opportunity for dynamic displays takes user interface designers well beyond current wisdom.

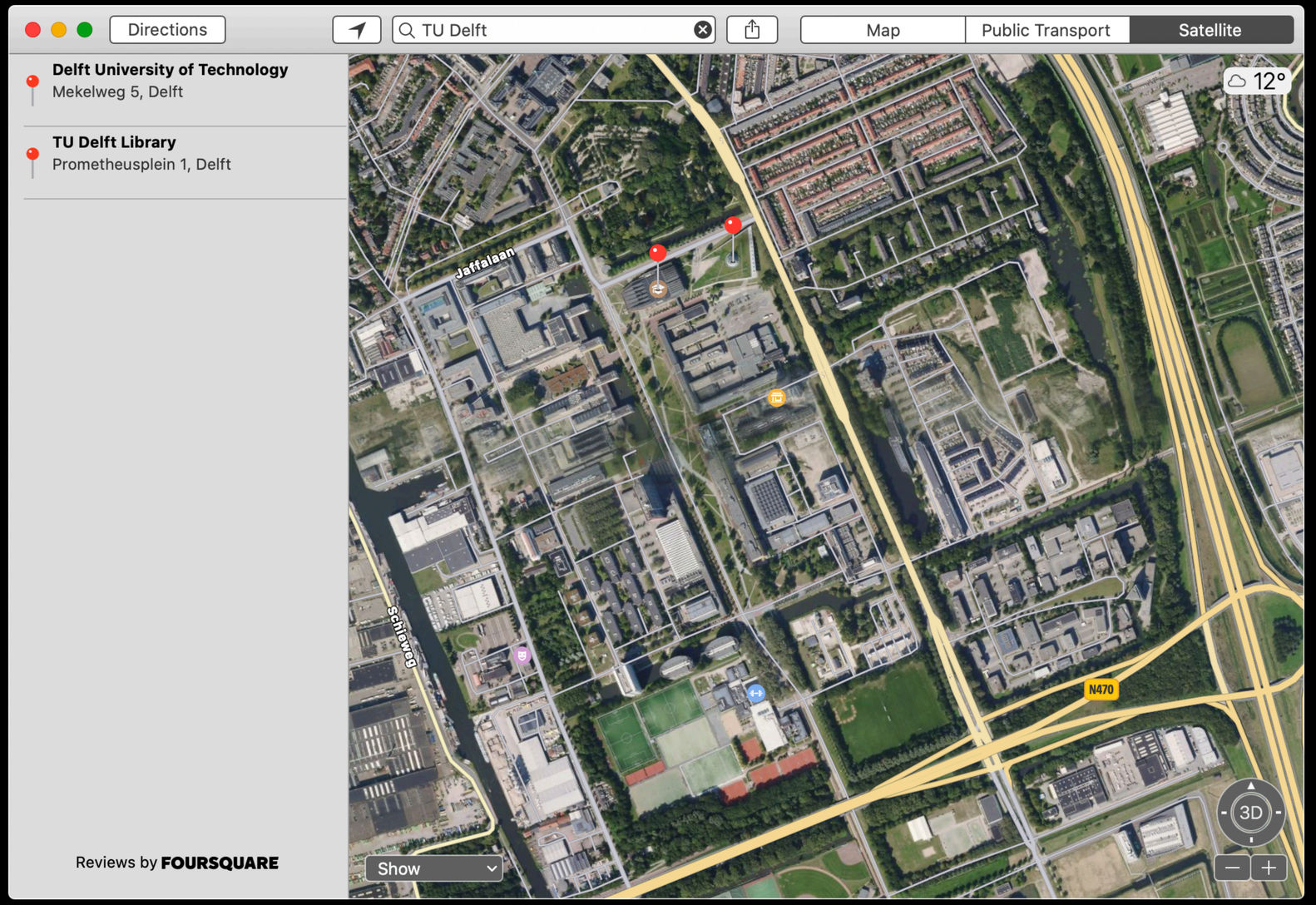
2. Visual Information Seeking Mantra

The success of direct-manipulation interfaces is indicative of the power of using computers in a more visual or graphic manner. A picture is often cited to be worth a thousand words and, for some (but not all) tasks, it is clear that a visual presentation—such as a map or photograph—is dramatically easier to use than is a textual description or a spoken report. As computer speed and display resolution increase, information visualization and graphical interfaces are likely to have an expanding role. If a map of the United States is displayed, then it should be possible to point rapidly at one of 1000 cities to get tourist information. Of course, a foreigner who knows a city's name (for example, New Orleans), but not its location, may do better with a scrolling alphabetical list.

0-8186-7469-5/96 $05.00 © 1996 IEEE

https://www.cs.umd.edu/~ben/papers/Shneiderman1996eyes.pdf

Example I: Apple Maps

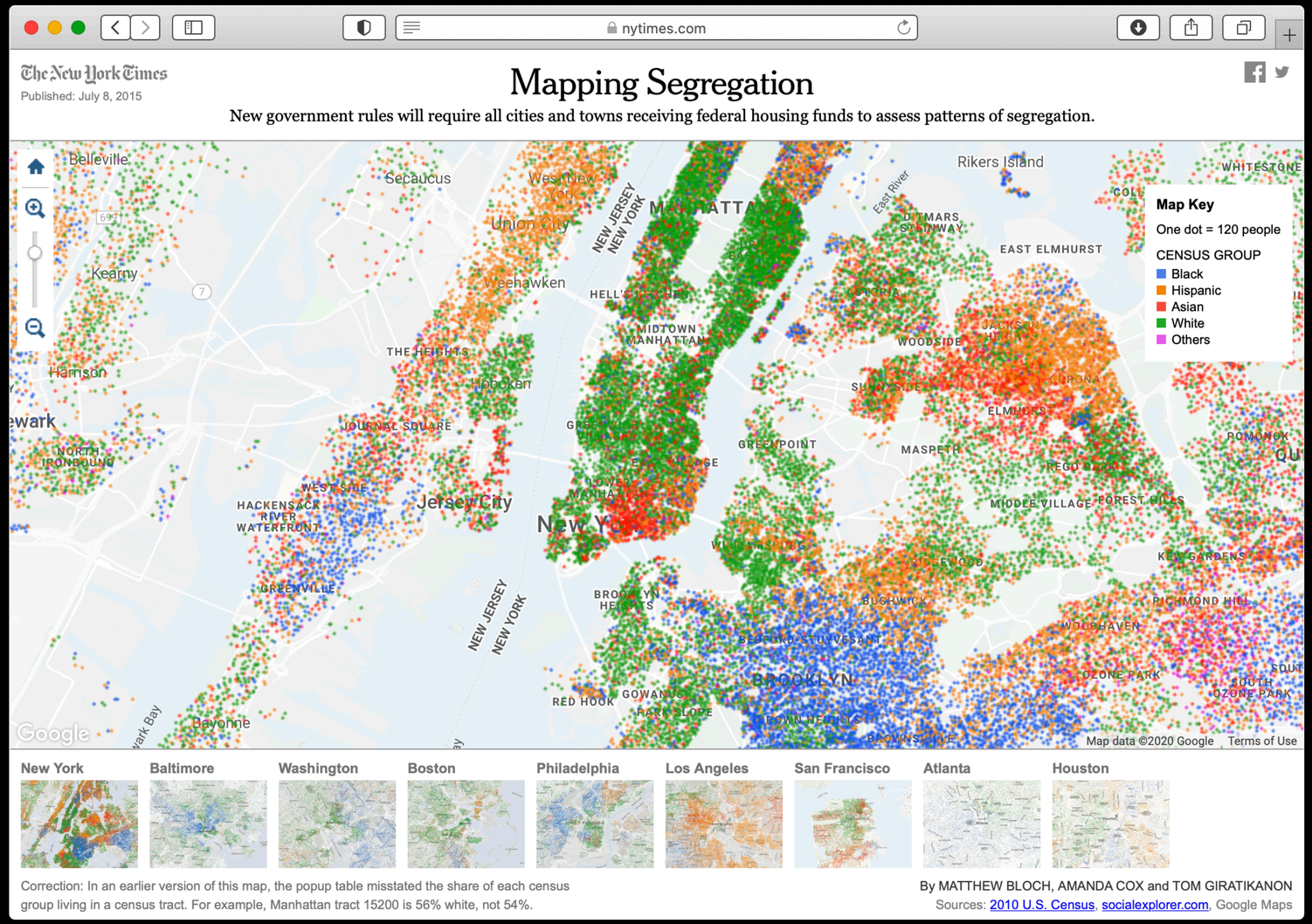


Example I: Apple Maps

- We can apply this system to any information system with a visual interface.
- Apple Maps is designed for a broad spectrum of user needs from a very large user base.

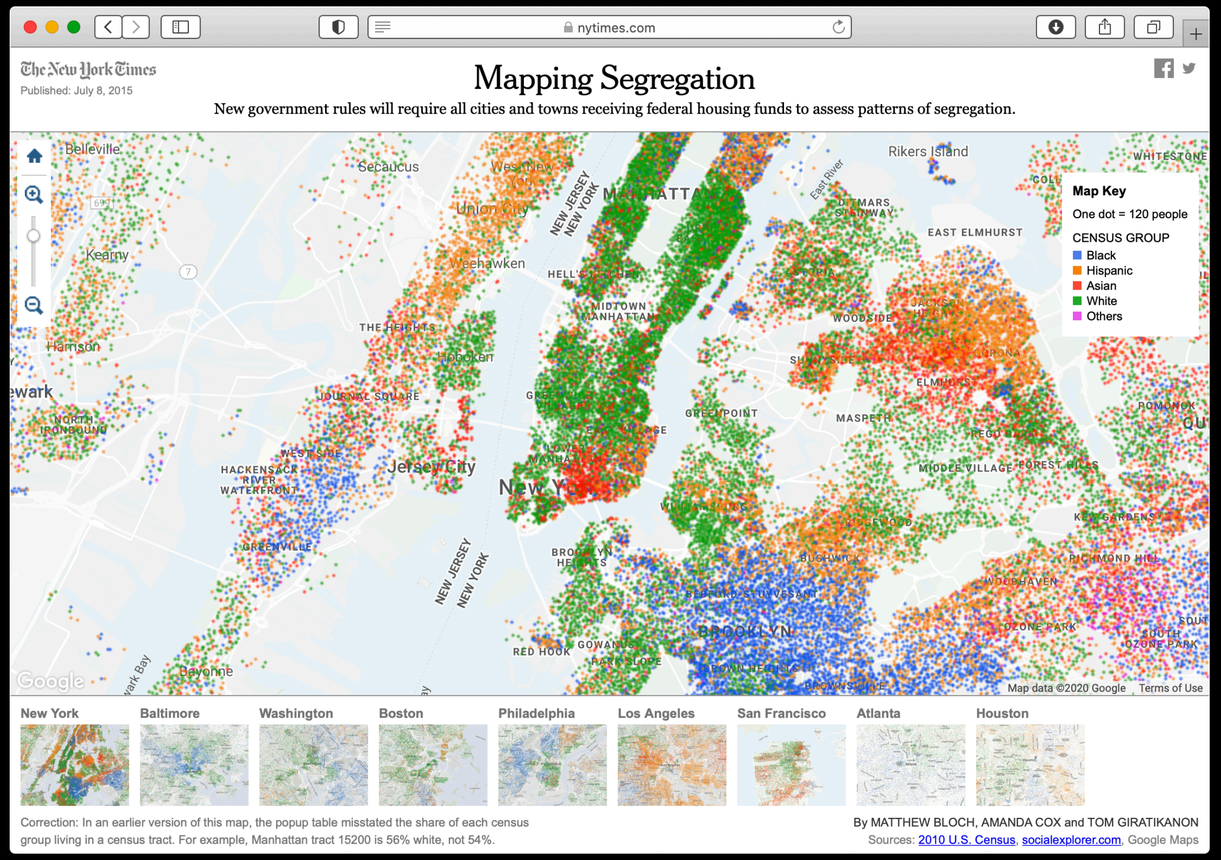
- **Overview**
- **Zoom**
- **Filter**
- **Details-on-demand**
- **Relate**
- **History**
- **Extract**

Example II: NYT Mapping Segregation



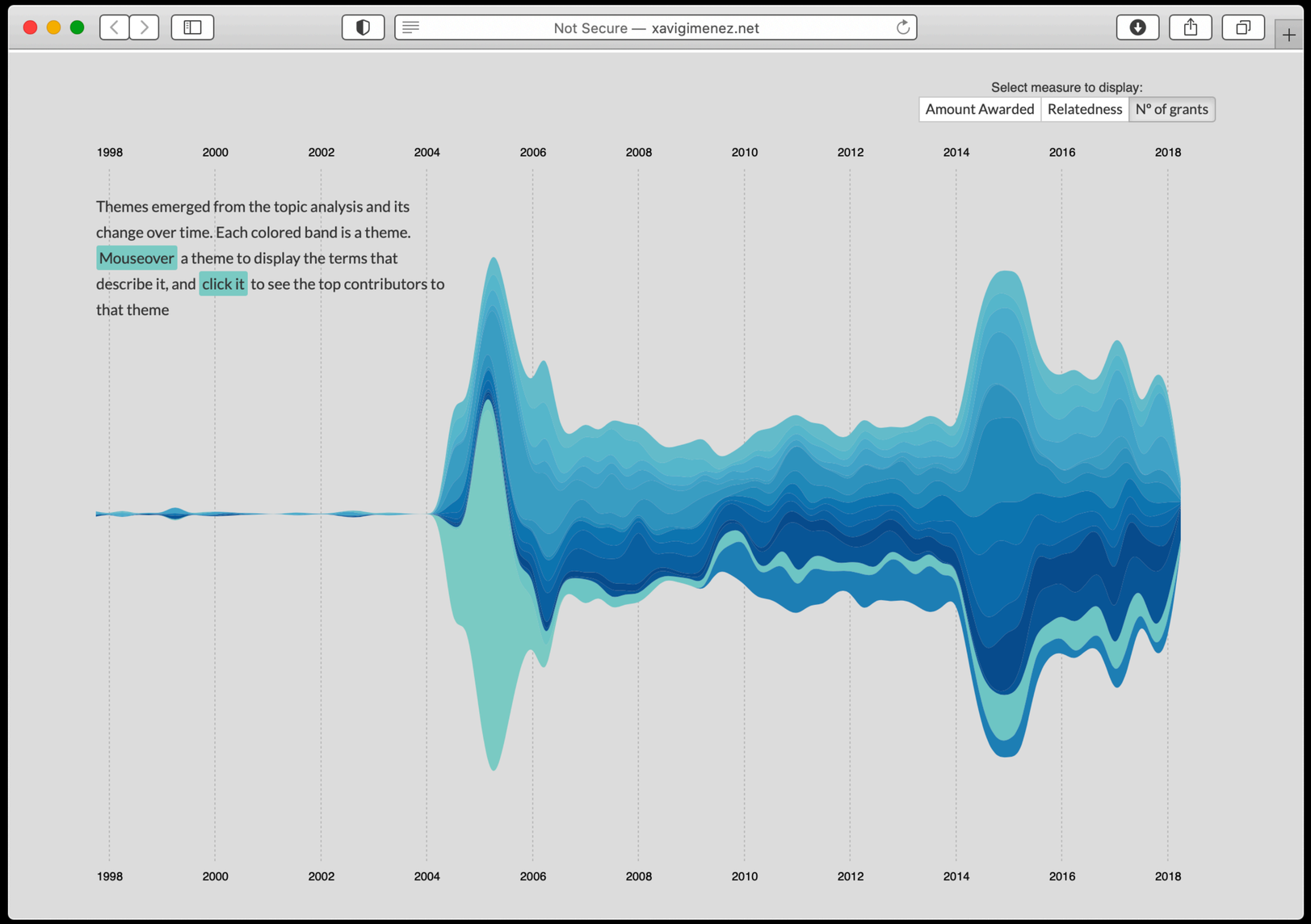
- https://www.nytimes.com/interactive/2015/07/08/us/census-race-map.html

Example III: Who has funded what theme



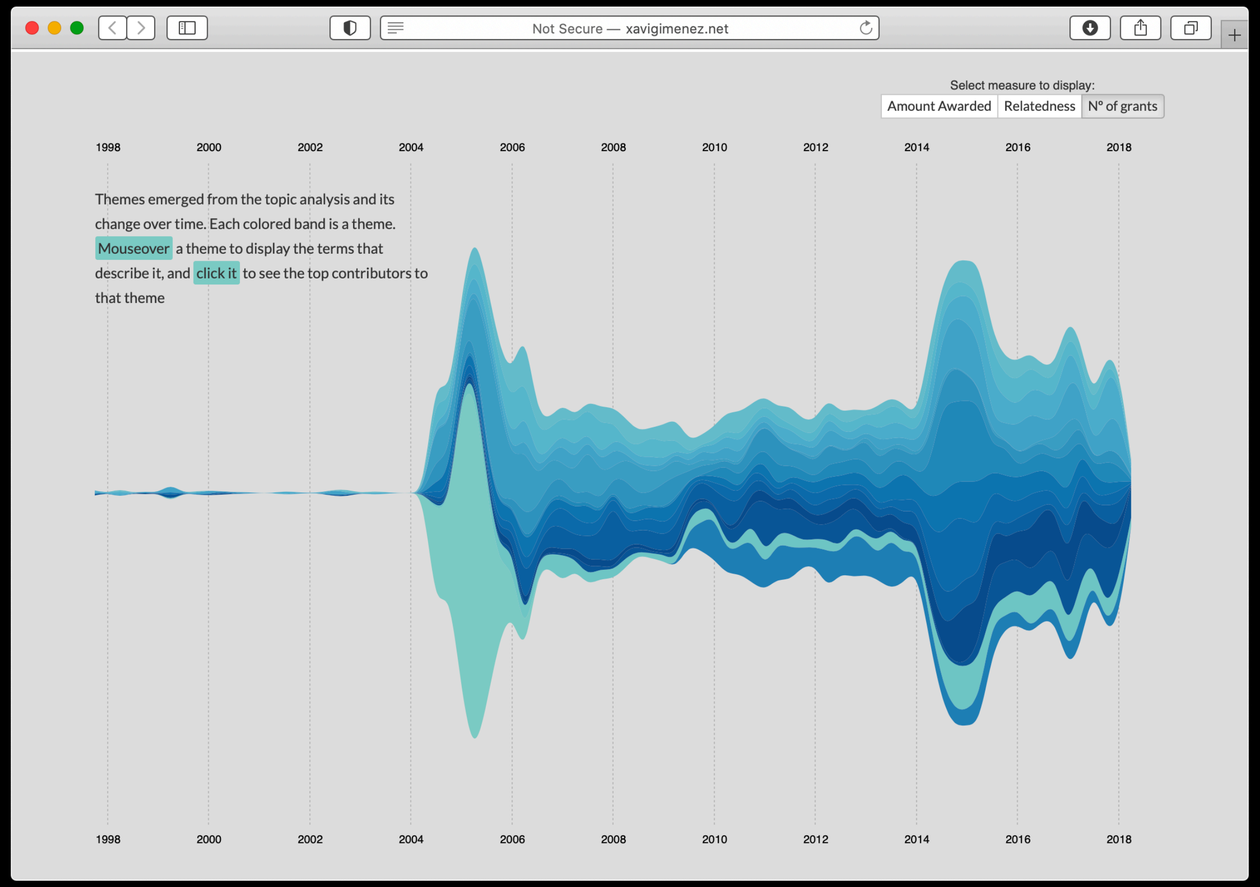
- Overview
- Zoom
- ~~Filter~~
- Details-on-demand
- Relate
- ~~History~~
- ~~Extract~~

Example III: Who has funded what theme



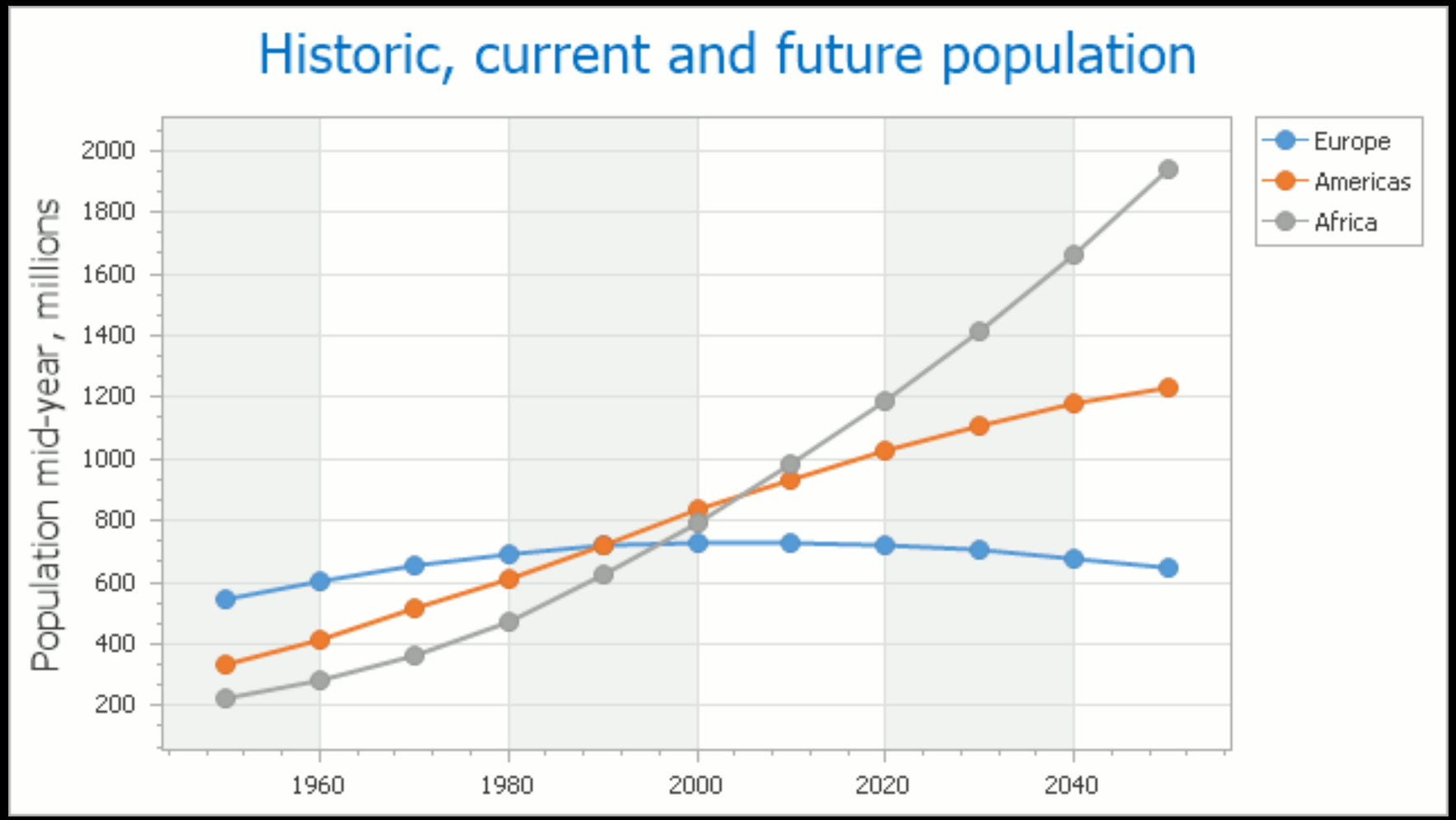
http://www.xavigimenez.net/360giving-grantmaking-themes/

Example III: Who has funded what theme

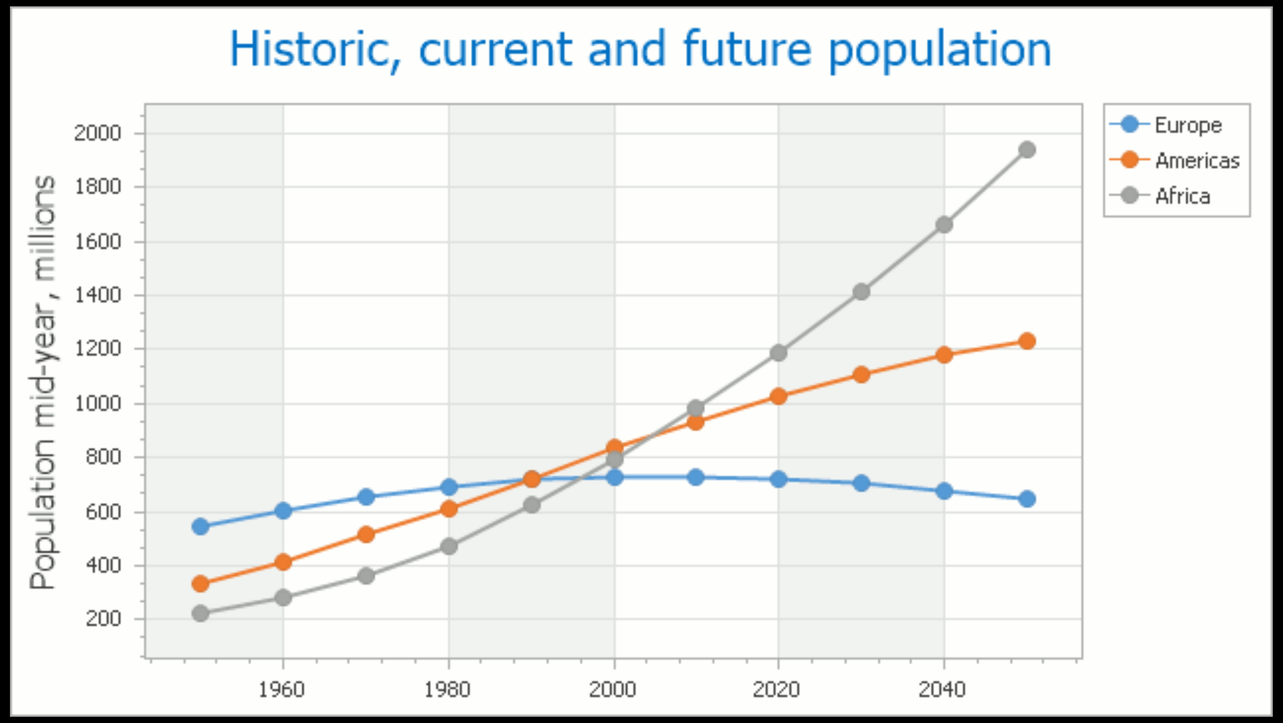


- Overview
- Zoom
- Filter
- Details-on-demand
- Relate
- ~~History~~
- ~~Extract~~

Example IV: A regular (static) line chart



Example IV: A regular (static) line chart



- **Overview**
- ~~Zoom~~
- ~~Filter~~
- ~~Details-on-demand~~
- **Relate**
- ~~History~~
- ~~Extract~~

Data and biases

- Trash in trash out
- Provenance of data
- Time sensitivity
- P hacking
- Cherry picking
- Data-driven, not question-driven
- Trust the system
- Correlation issues (e.g., spurious correlation)
- Lack of context to understand the data

Trash in Trash out

Data and biases

- Bad data leads to bad visualizations
- Is your data clean?
- Common errors include **data duplication**, **missed data**, **NA values not marked**, and so on.

Provenance of Data

Data and biases

- The data used to create a visualisation is more credible if data is made public
- Without having access to the data, visualizations cannot be verified nor replicated.

Time sensitivity

Data and biases

- Particular insights only make sense in a particular time frame
- If a graph is time-sensitive, it means that its insights can only be used in a particular point in time.
- It is important that the visualizations communicates the validity of the visualisation across time.

P-hacking

Data and biases

- The term was coined in 2014 by Regina Nuzzo in Nature News.
- P-value is a common metric to check statistical significance when conducting hypothesis testing.
  - In sum, to find evidence, p-value needs to be below a significance level, typically 0.05.
  - P-hacking is manipulating data in a way that produces a desired p-value

STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had discovered that extremists quite literally see the world in black and white.

The results were "plain as day", recalls Motyl, a psychology PhD student at the University of Virginia in Charlottesville. Data from a study of nearly 2,000 people seemed to show that political moderates saw shades of grey more accurately than did either left-wing or right-wing extremists. "The hypothesis was sexy," he says, "and the data provided clear support." The *P* value, a common index for the strength of evidence, was 0.01 — usually interpreted as 'very significant'. Publication in a high-impact journal seemed within Motyl's grasp.

But then reality intervened. Sensitive to controversies over reproducibility, Motyl and his adviser, Brian Nosek, decided to replicate the study. With extra data, the *P* value came out as 0.59 — not even close to the conventional level of significance, 0.05. The effect had disappeared, and with it, Motyl's dreams of youthful fame[1].

It turned out that the problem was not in the data or in Motyl's analyses. It lay in the surprisingly slippery nature of the *P* value, which is neither as reliable nor as objective as most scientists assume. "*P* values are not doing their job, because they can't," says Stephen Ziliak, an economist at Roosevelt University in Chicago, Illinois, and a frequent critic of the way statistics are used.

For many scientists, this is especially worrying in light of the reproducibility concerns. In 2005, epidemiologist John Ioannidis of Stanford University in California suggested that most published findings are false[2]; since then, a string of high-profile replication problems has forced scientists to rethink how they evaluate results.

At the same time, statisticians are looking for better ways of thinking about data, to help scientists to avoid missing important information or acting on false alarms. "Change your statistical philosophy and all of a sudden different things become important," says Steven Goodman, a physician and statistician at Stanford. "Then 'laws' handed down from God are no longer handed down from God. They're actually handed down to us by ourselves, through the methodology we adopt."

DALE EDWIN MURRAY

OUT OF CONTEXT

P values have always had critics. In their almost nine decades of existence, they have been likened to mosquitoes (annoying and impossible to swat away), the emperor's new clothes (fraught with obvious problems that everyone ignores) and the tool of a "sterile intellectual rake" who ravishes science but leaves it with no progeny[3]. One researcher suggested rechristening the methodology "statistical hypothesis inference testing"[3], presumably for the acronym it would yield.

The irony is that when UK statistician Ronald Fisher introduced the *P* value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the

© 2014 Macmillan Publishers Limited. All rights reserved

Cherry Picking

Data and biases

- Fallacy of **selecting evidence** that supports an argument while ignoring evidence that contradicts it.

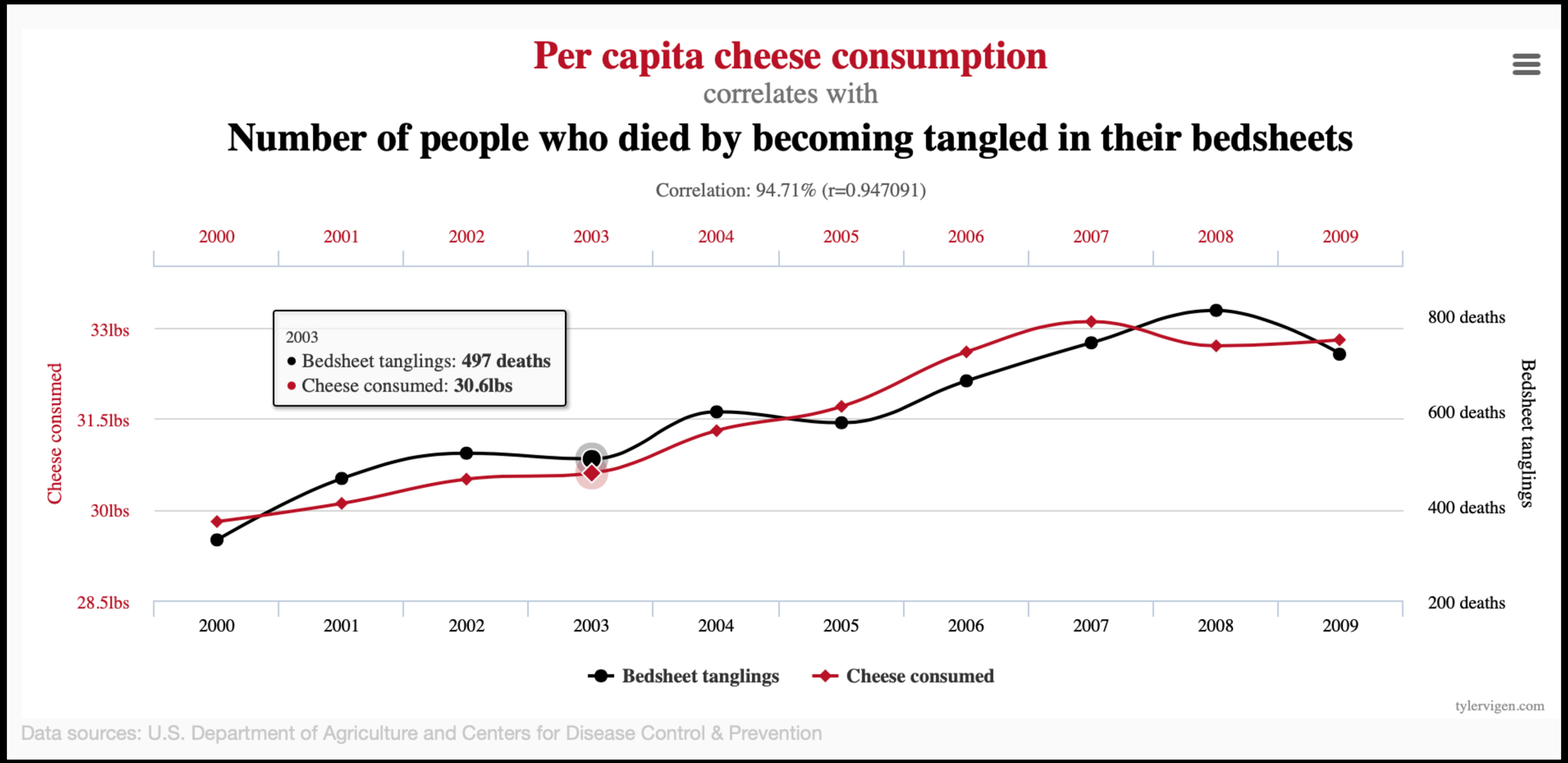
Trust the system

Data and biases

- Errors may occur anywhere in the data pipeline
- Replication is important to make sure there are no systematic errors.

Spurious correlation

Data and biases



https://www.tylervigen.com/spurious-correlations

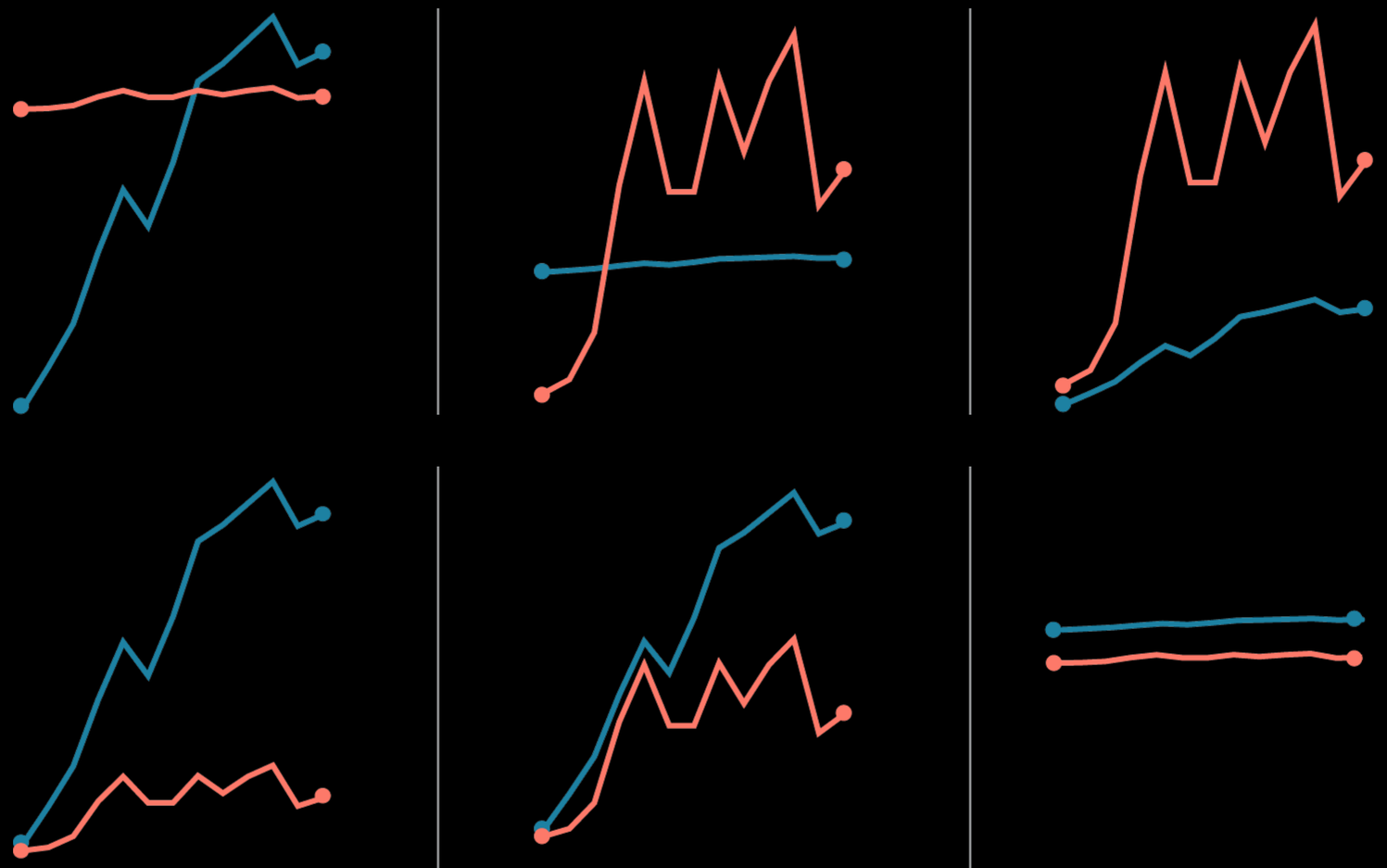
Lack of context to understand the data

Data and biases

- Insights from data cannot be provided without context.
- Without the right context, visualizations can be misleading.
- **Purpose**, **Readership**, **Media**, **Visual guides**.
-

Spurious correlation: How to avoid?

- To avoid spurious correlation, **we should NOT**:
  - Plot unrelated things together;
  - Using two y-axes in any graph;
  - Cut numerical axes in any graph;
  - Use Pareto chart (https://datavizproject.com/data-type/pareto-chart/)



https://blog.datawrapper.de/dualaxis/

Spurious correlation and Big data

- Spurious correlation (some events are **correlated but not causally related**) is a major potential problem for big data analysis, because when a dataset is large enough, everything could be correlated with everything.
- https://link.springer.com/article/10.1007/s10699-016-9489-4

Assignment 7

- https://luiscruz.github.io/course_infovis/assignments/assignment7