Data Science in the Wild

Lecture 13: Information Visualization

Eran Toch

CORNELL TECH

Agenda

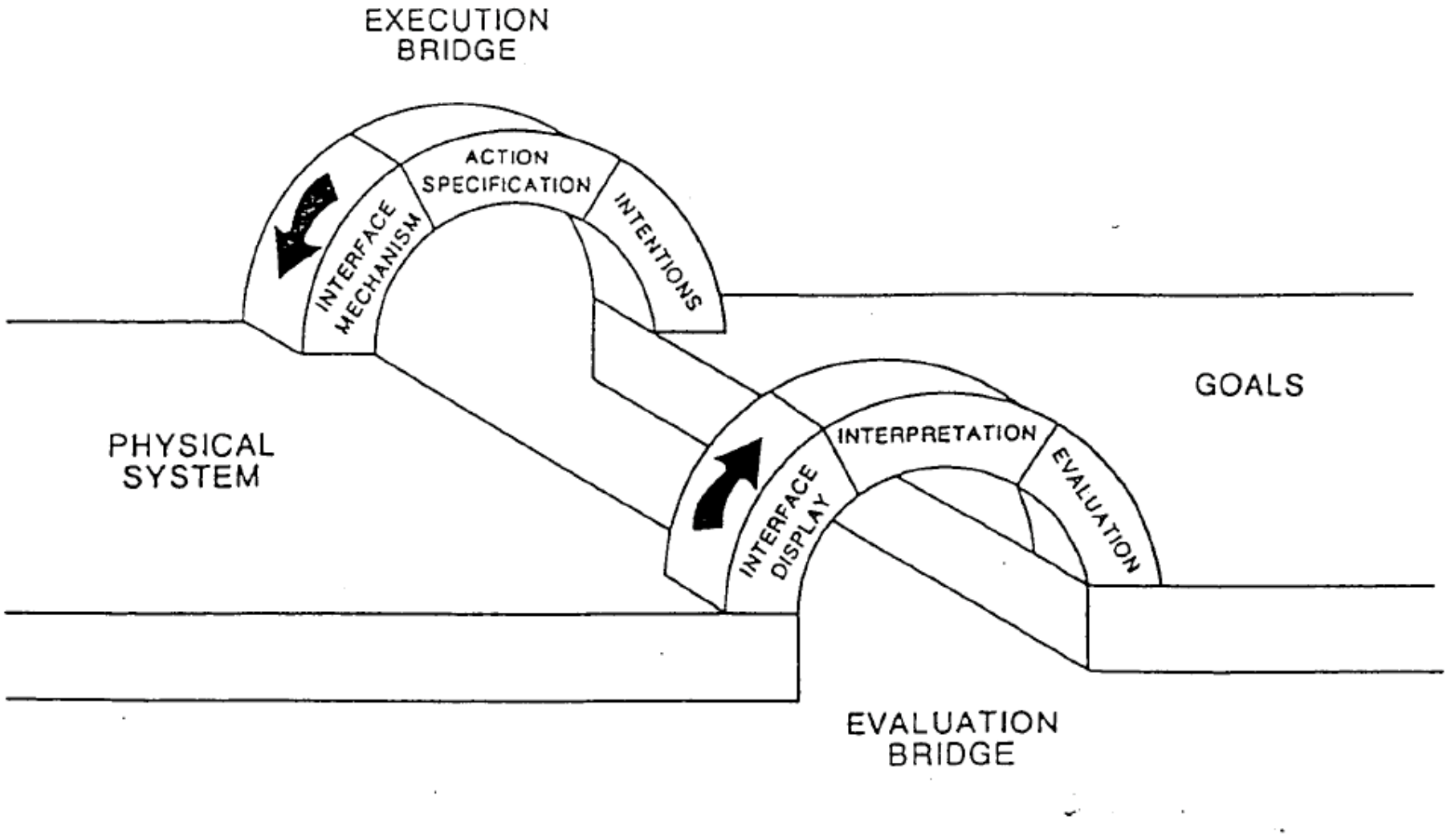
1. What is visualization?
2. Types of visualization
3. Good visualizations

A definition

- “Transformation of the symbolic into the geometric” (McCormick et al., 1987)
- The depiction of information using spatial or graphical representations, to facilitate comparison, pattern recognition, change detection, and other cognitive skills by making use of the visual system

By Marti Hearst

Gulf of Execution and Evaluation



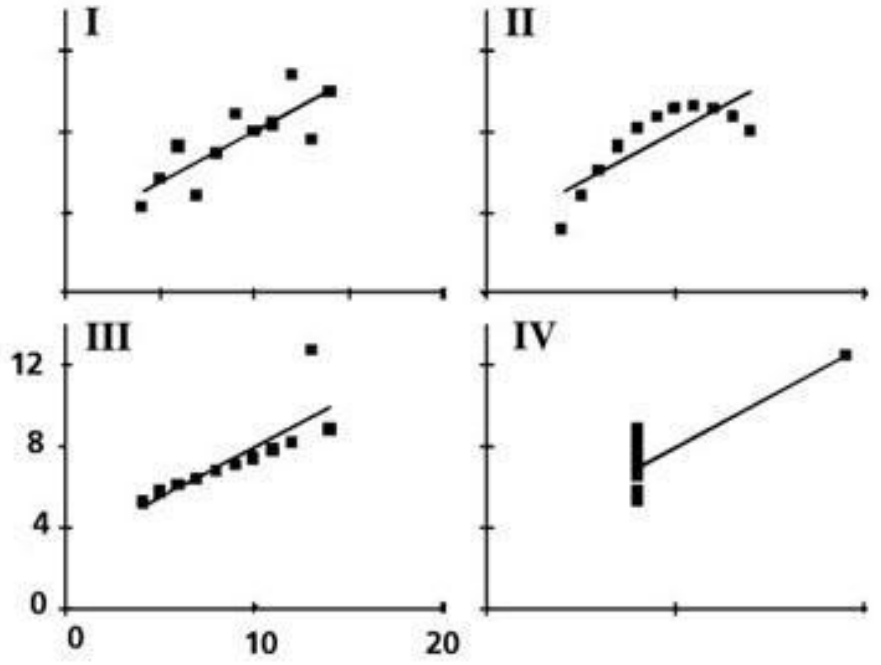
Norman 1986

The limitations of symbolic data

The regression lines for all these data points are exactly the same

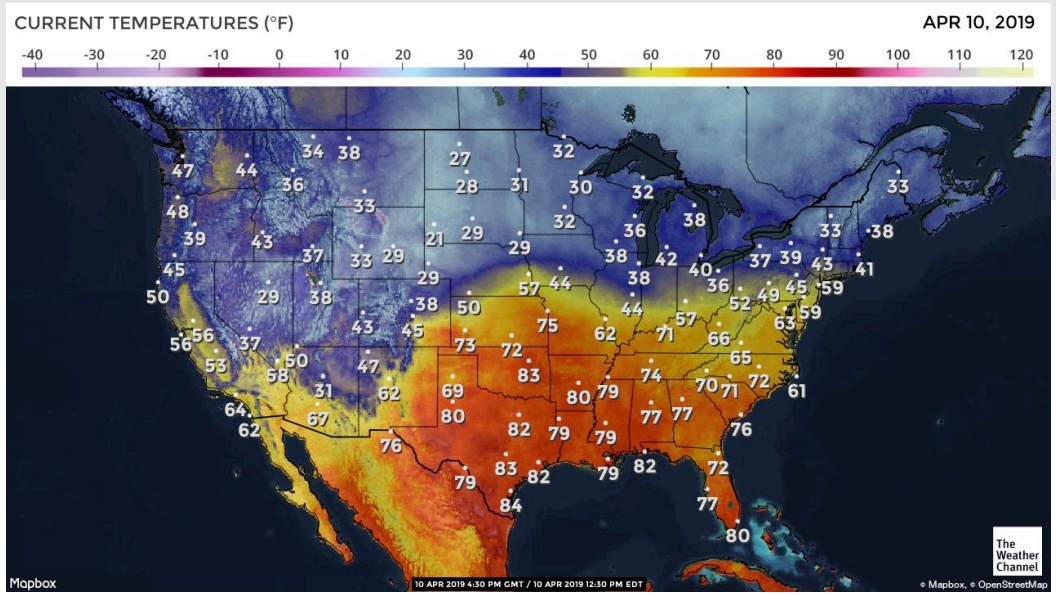
Anscombe's Quartet

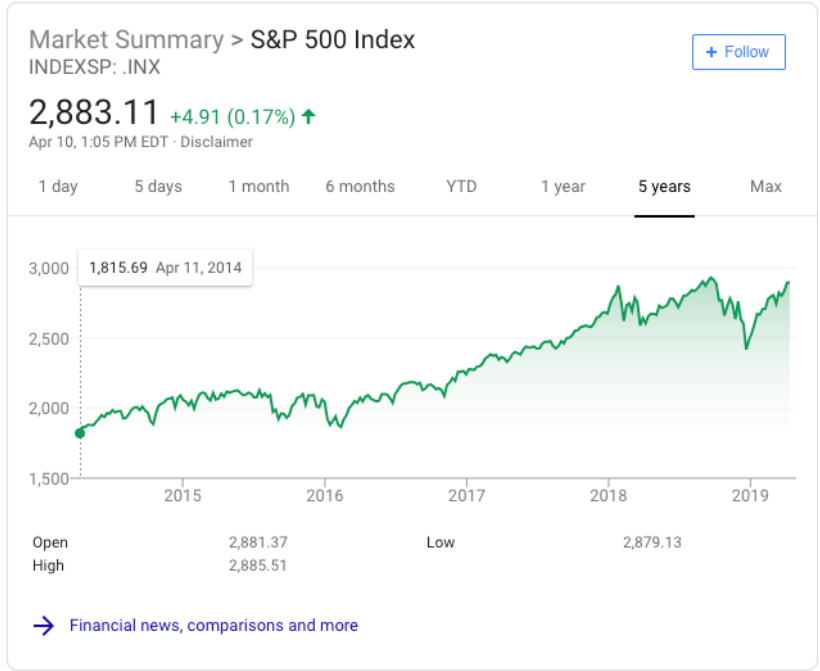
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



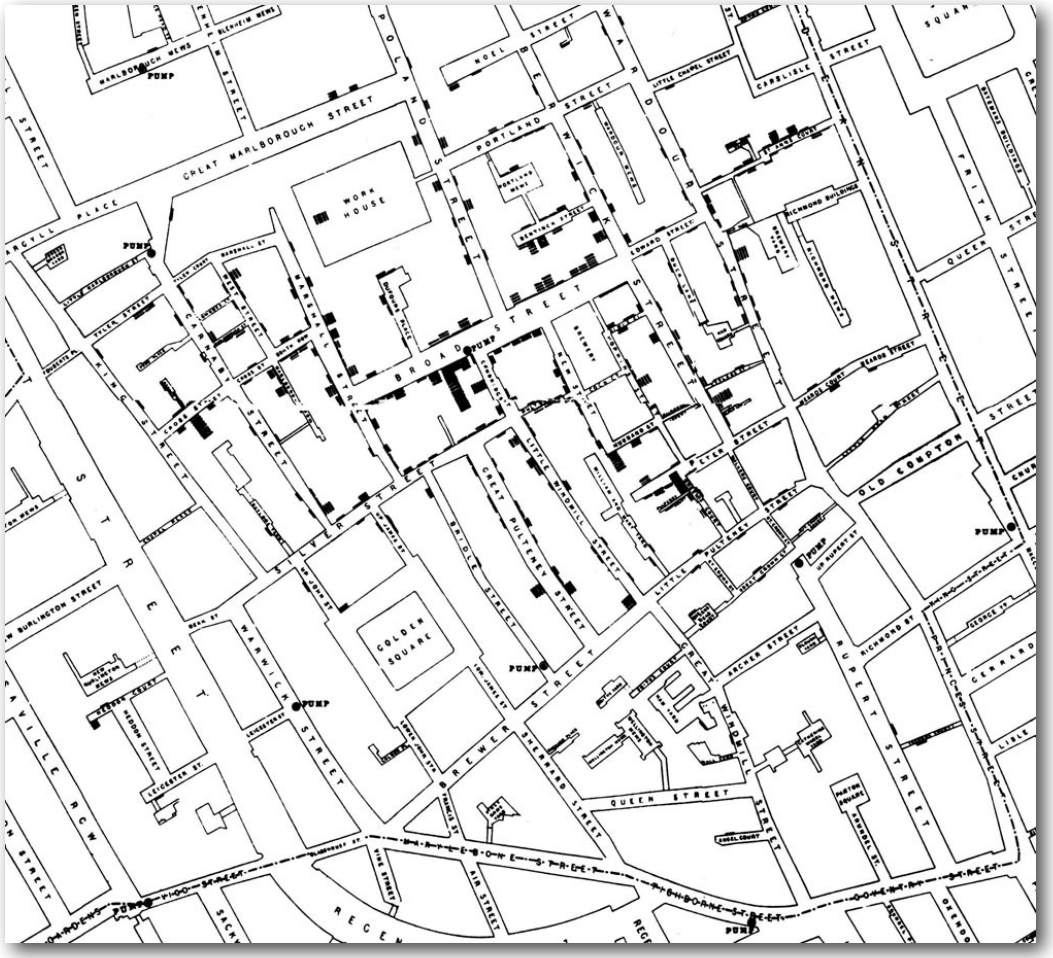
When do we use visualization

- **Explore the data**
  - Make large datasets accessible
  - Support scanning, outlier detection, recognizing and exploring
- **Support analysis**
  - Pattern recognition
  - Comparisons
- **Tell stories about the data**
  - Explain
  - Reason

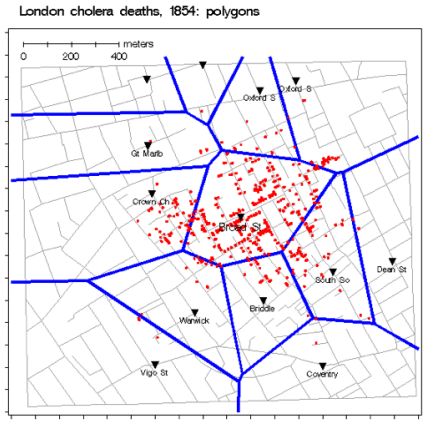




Good Visualizations

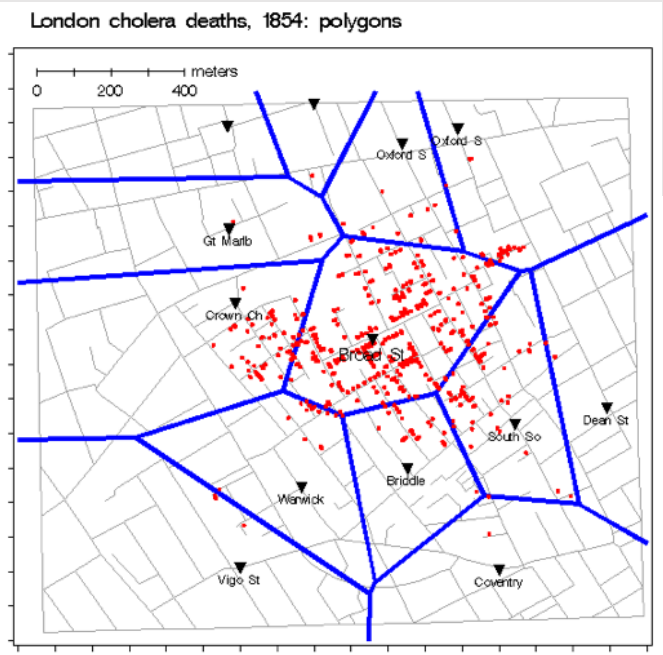


Dr. John Snow's map of deaths from a cholera outbreak in London, 1854, in relation to the locations of public water pumps.



Broadwick Street showing the John Snow memorial and public house.

Theory of Visualization

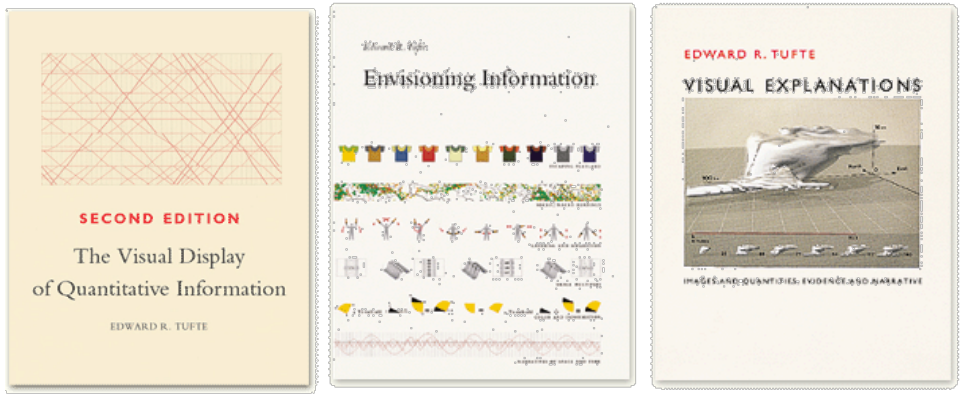


Mapping data questions to proper visualization.

- How to display dimensions of different types
- How to use color? Shapes?
- How to present multiple dimensions?
- What are the cognitive properties of visualization?
- How to differentiate between good and bad visualization?

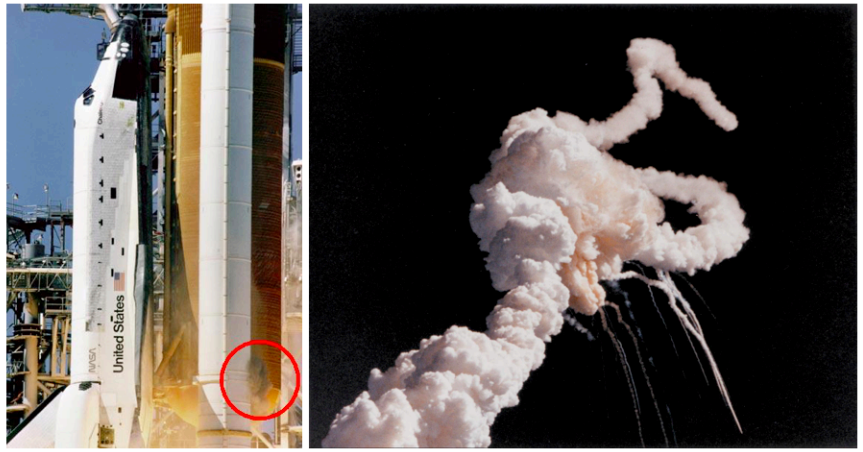
Theory

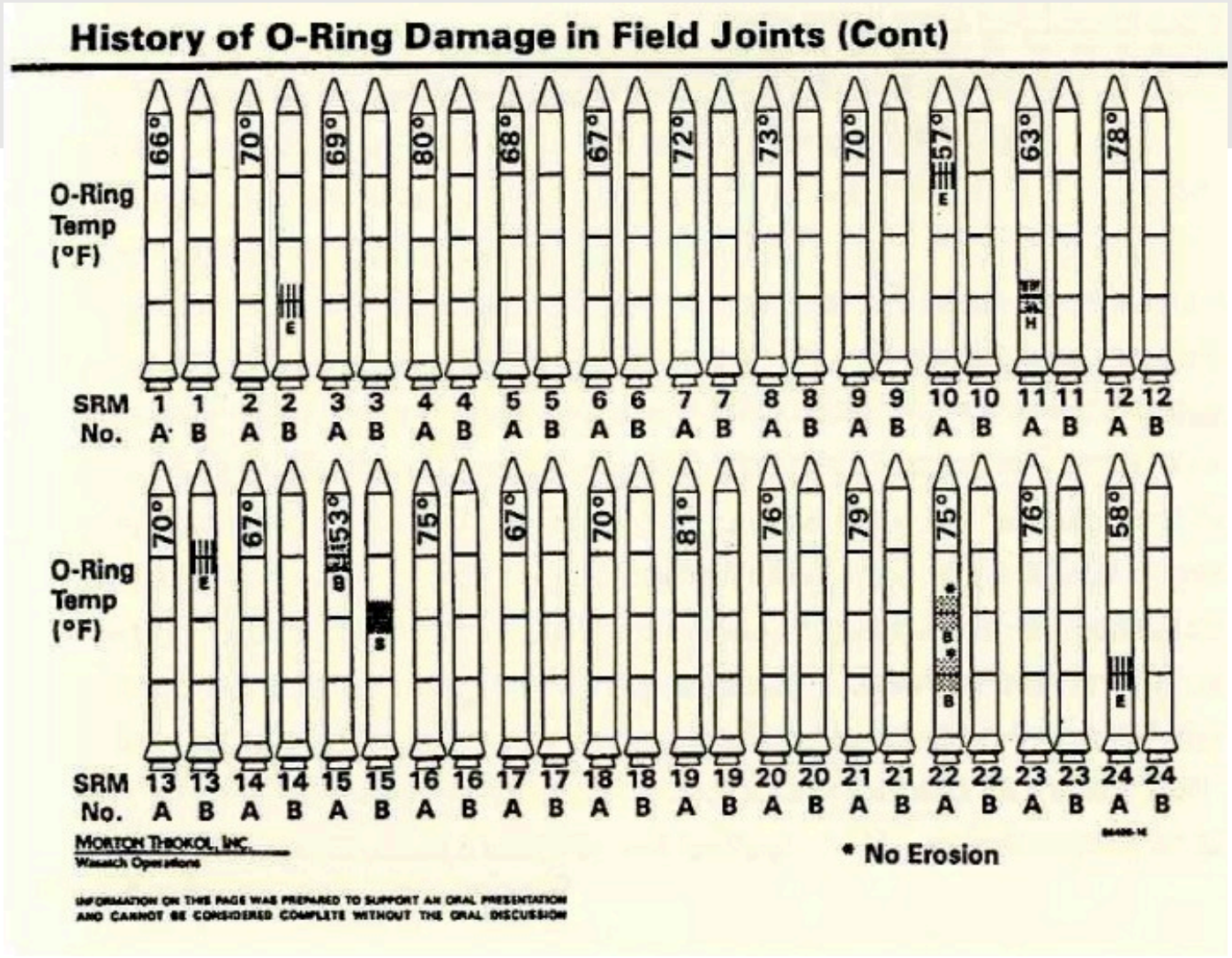
- Kosslyn: Types of Visual Representations
- Lohse et al: How do people perceive common graphic displays
- Bertin, MacKinlay: Perceptual properties and visual features
- Tufte/Wainer: Representing honesty and context



Tufte

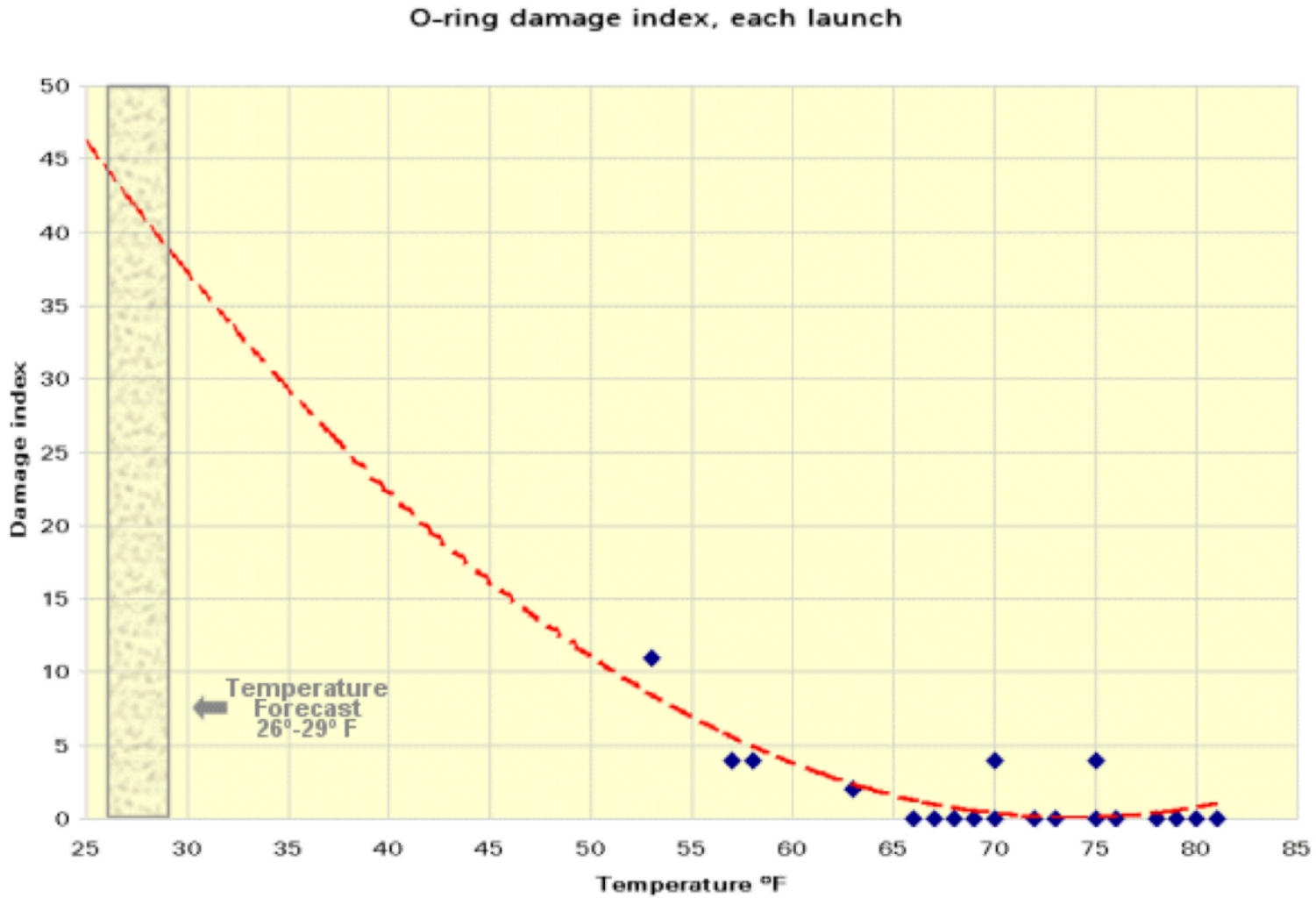
Bad Visualization





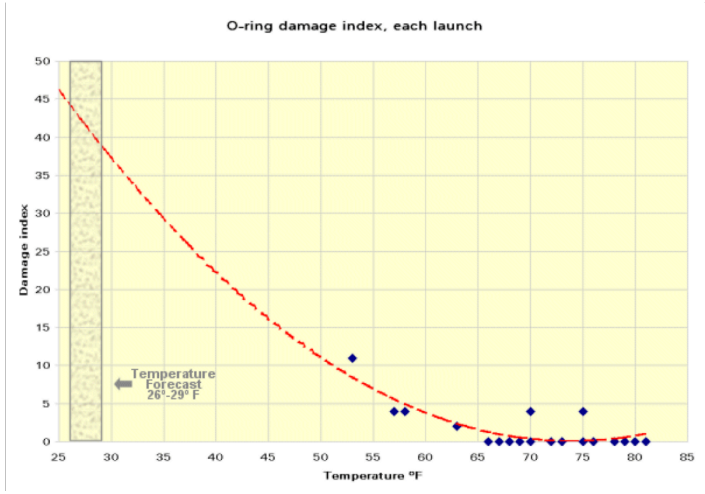
The damage to the O-rings and the temperature is associated with each launch

A (possibly) better visualization



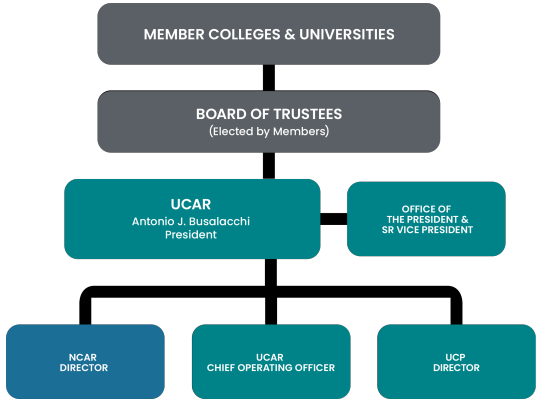
https://medium.com/the-data-experience/four-questions-you-should-ask-before-visualizing-your-data-cd20a302eb65

Visualization Types



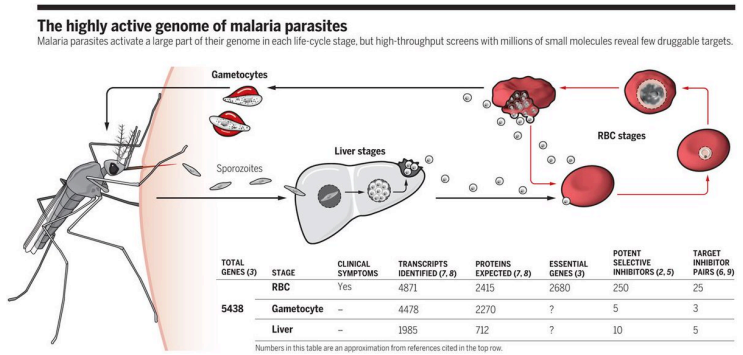
Graph

At least two scales, one of them quanitative



Chart

Discrete relations and structures between entities

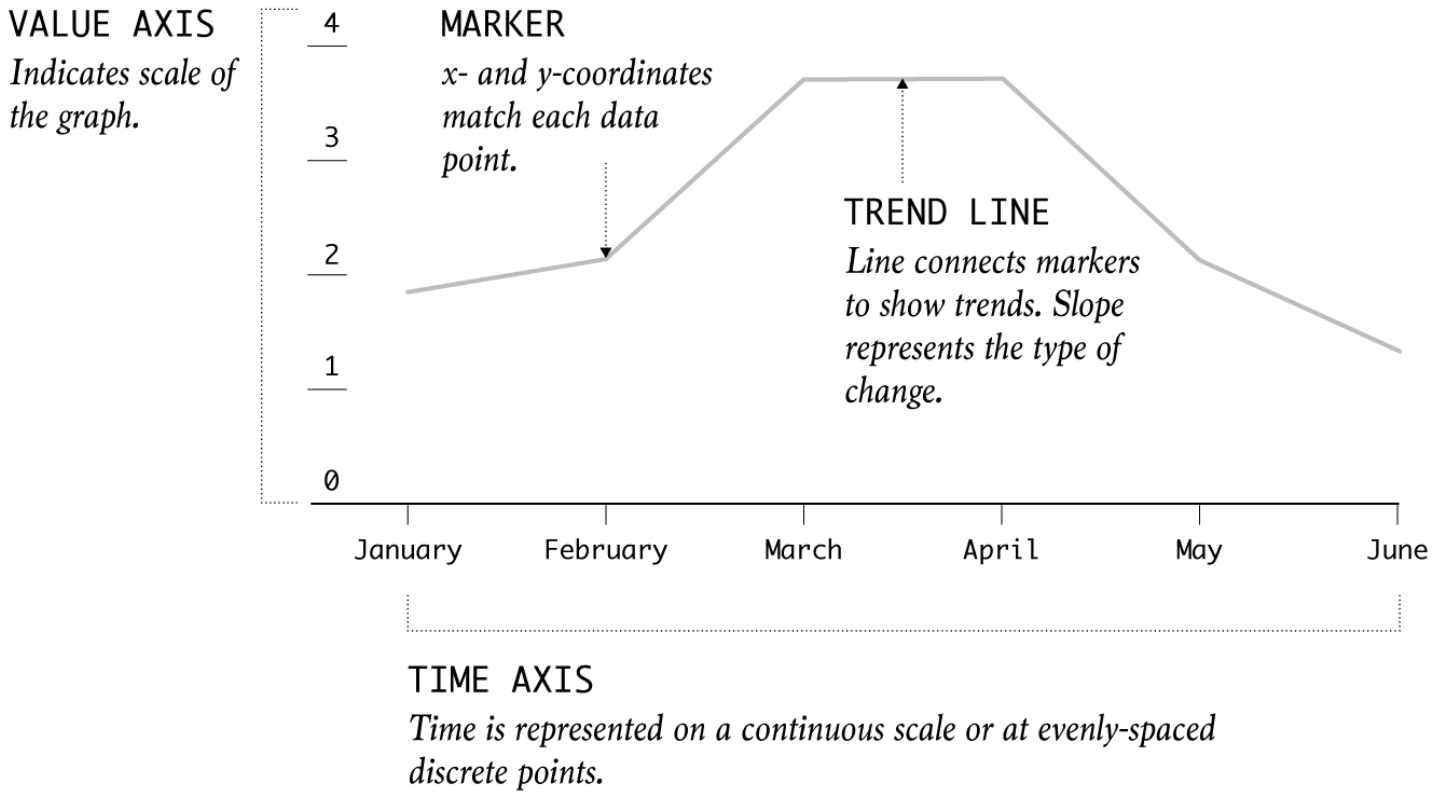


Diagram

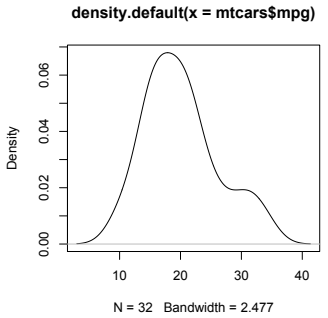
Mixing graphic schematics and symbolic information

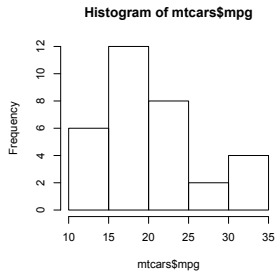
What makes a graph?

- Framework
  - Scale
  - Layers
- Content
  - Graphics: points, lines, areas, bars
  - Marks
- Labels
  - Titles
  - Axes
  - Tic marks

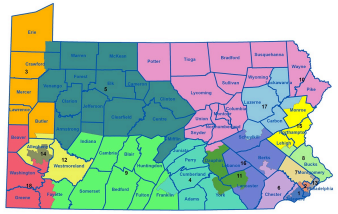


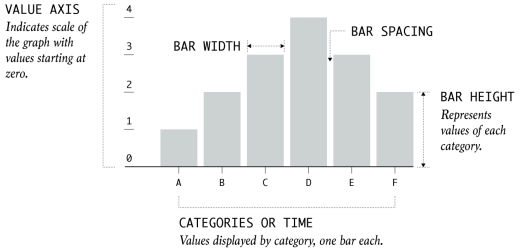
Graph types

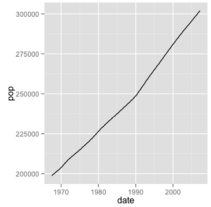




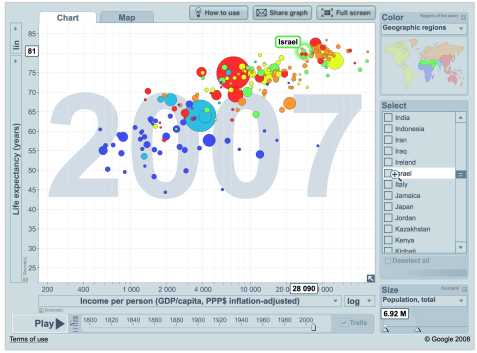
Distribution



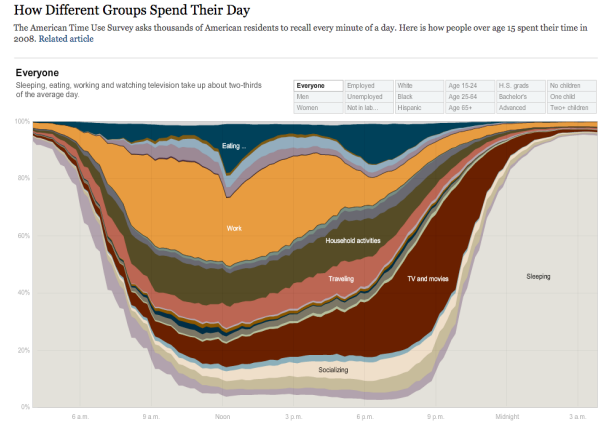


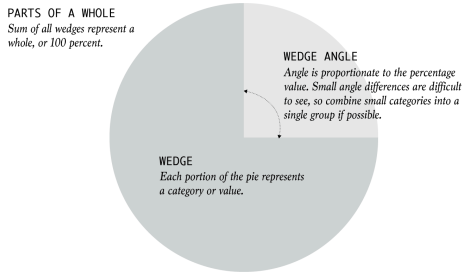


Comparison



Relationship



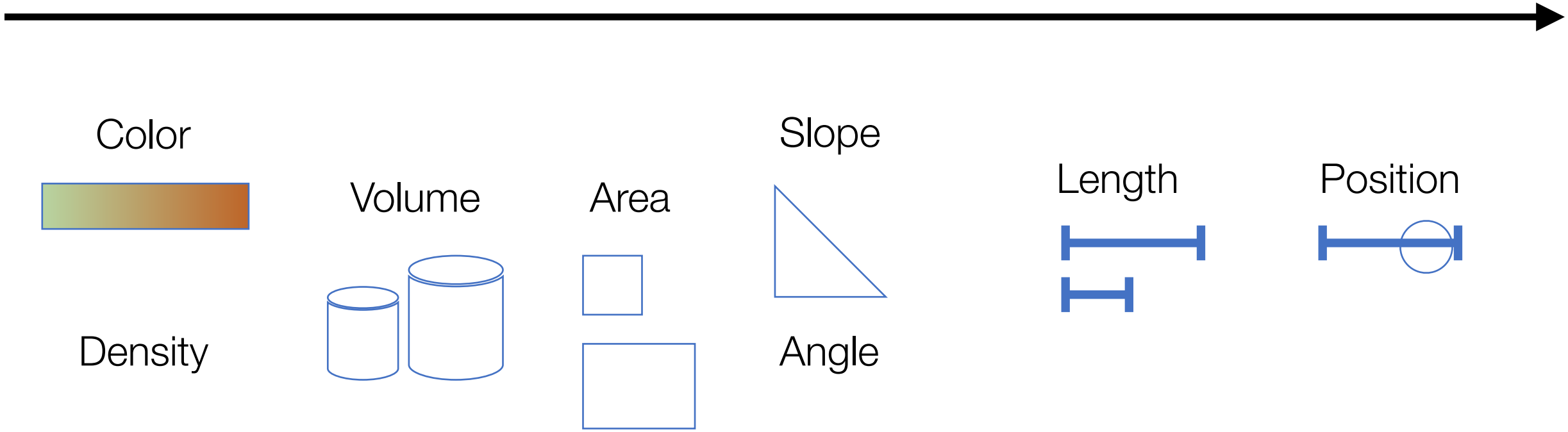


Composition

Accuracy and Visualization

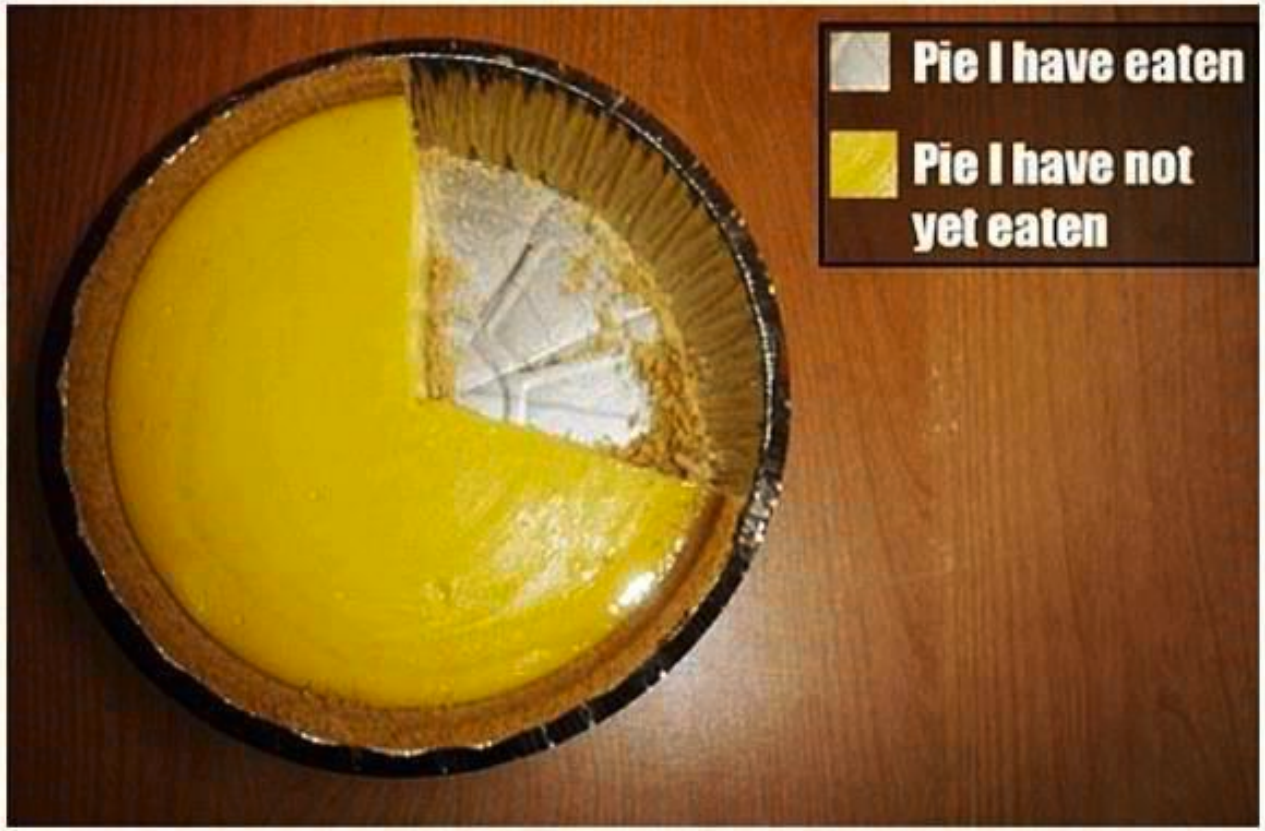
Less Accurate

More Accurate



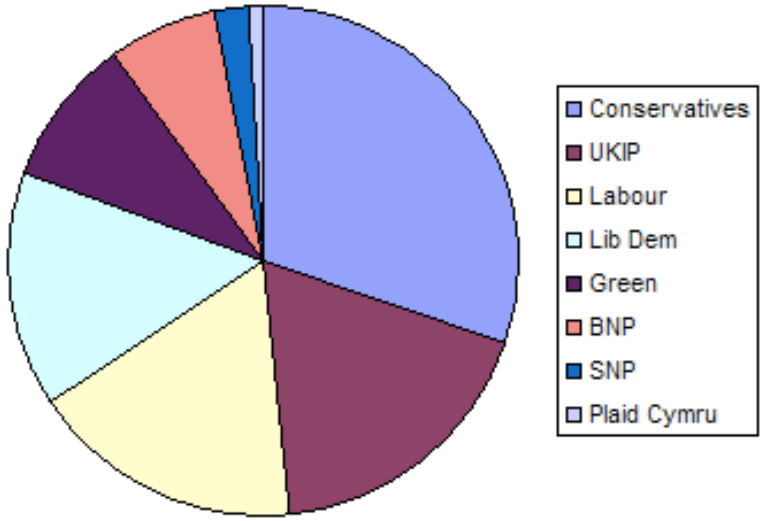
[Mackinlay 88 from Cleveland & McGill

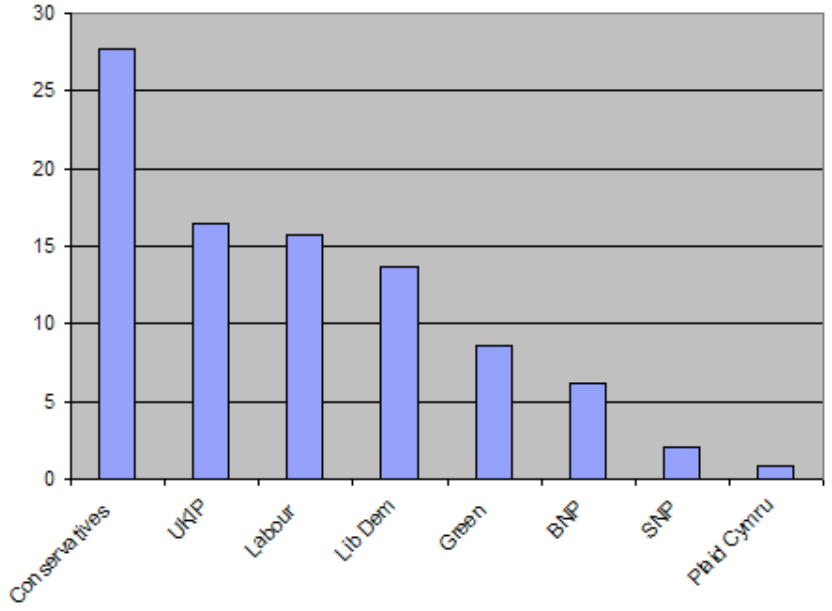
World's Most Accurate Pie Chart



Hey, what do you have against Pie Charts?

- Hard to compare similar values
  - With more than two values
- What is the difference between the second, third and fourth parties?

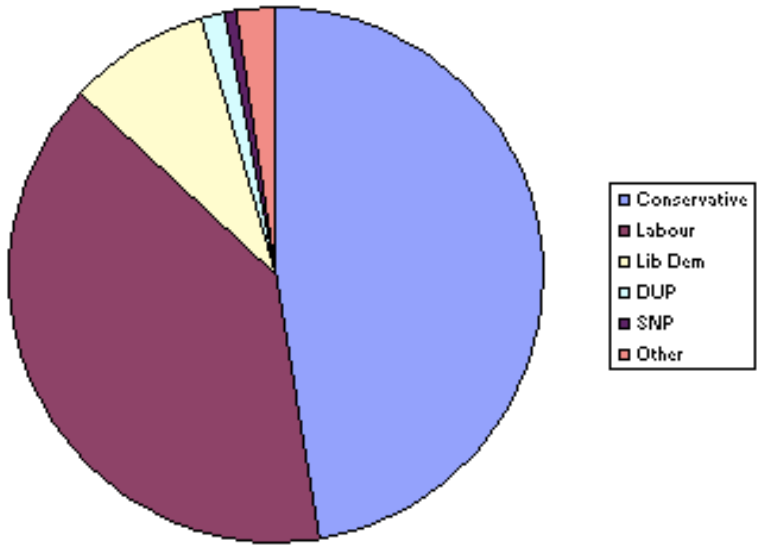


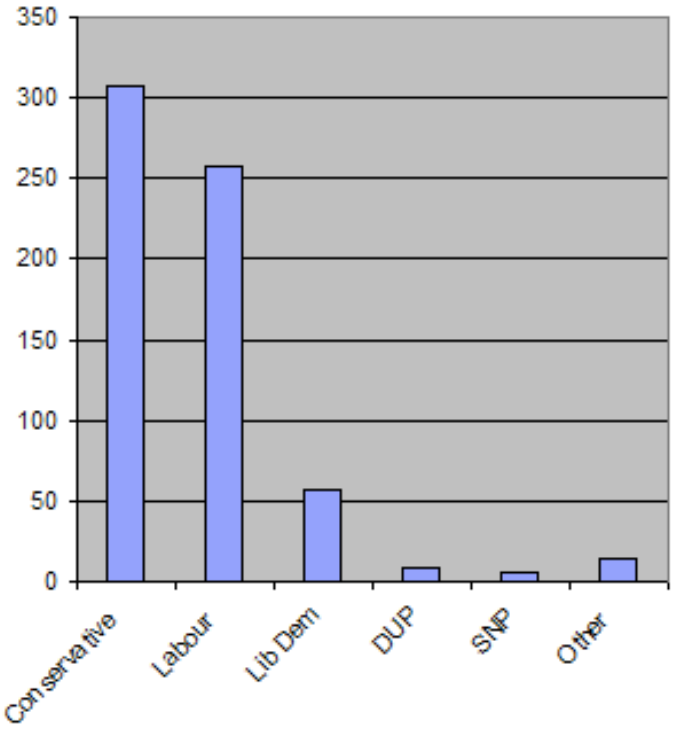


http://www.stevefenton.co.uk/Content/Pie-Charts-Are-Bad/

Pie-Charts are Evil

- Difficult to evaluate combined values.
- Can the Labour and Lib-Dem parties form a coalition?



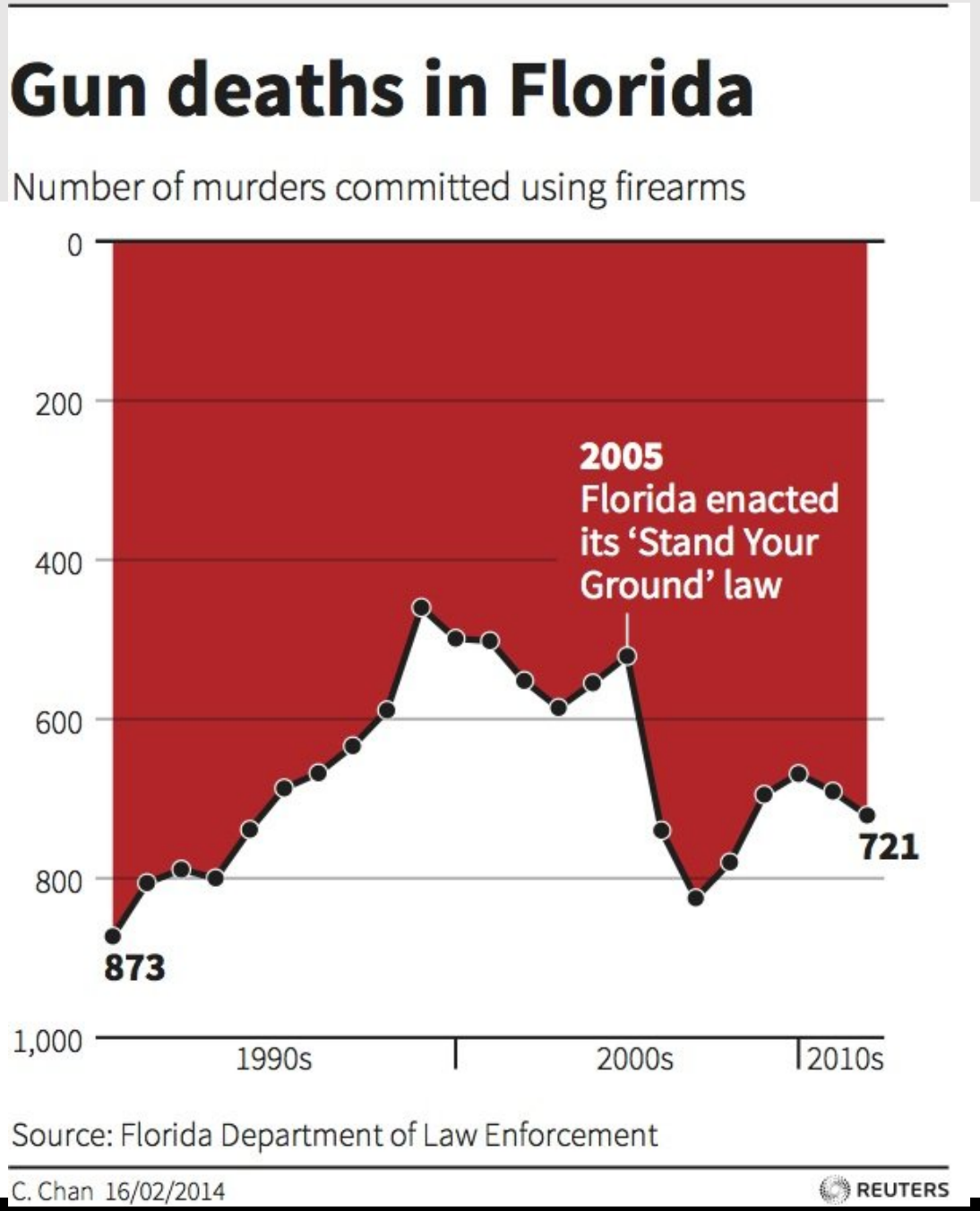


Good Visualization Theory

- Tufte defines several rules for creating good visualizations:
  - Avoid distorting the data
  - Maximize the data-ink ratio
    - Every pixel requires a reason
  - Encourage Eye to Compare Different Pieces of Data
    - Putting data in context: linear averages, examples
  - Closely integrate Statistical and Verbal Descriptions

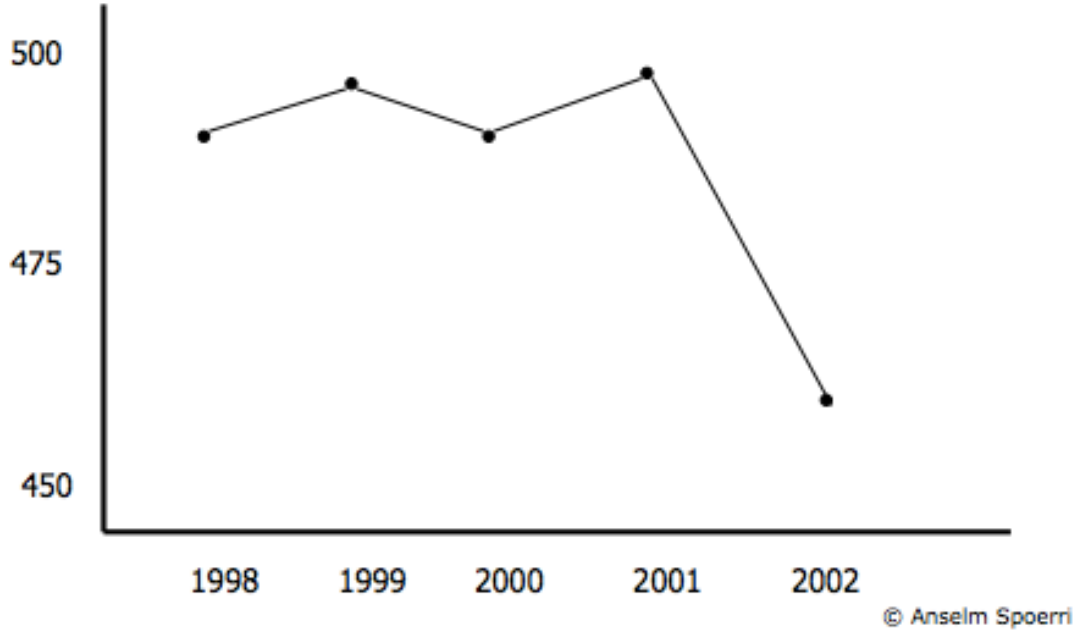
Data Distortion

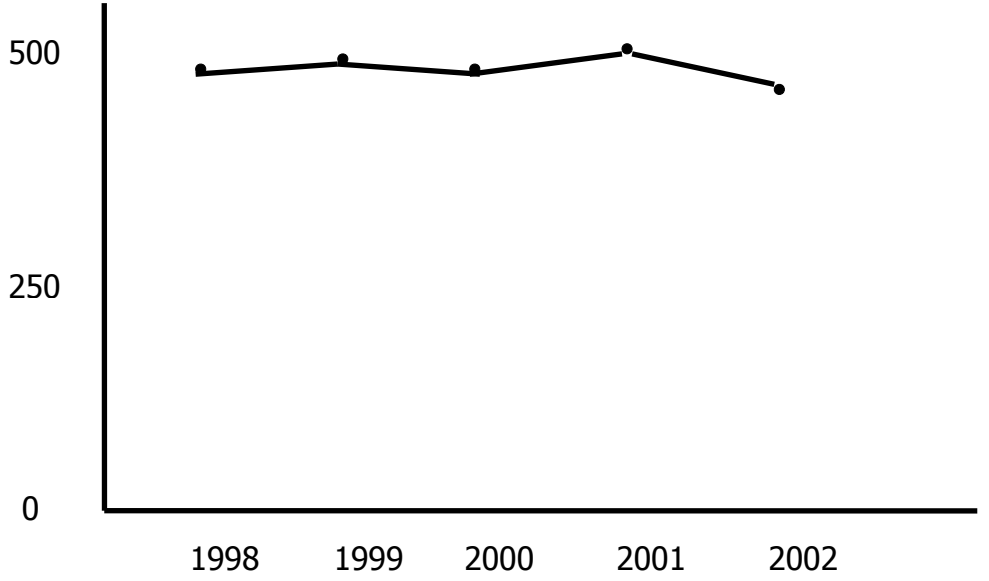
- Graphical integrity:
  - Misleading uses of area
  - Misleading uses of perspective
  - Leaving out important context



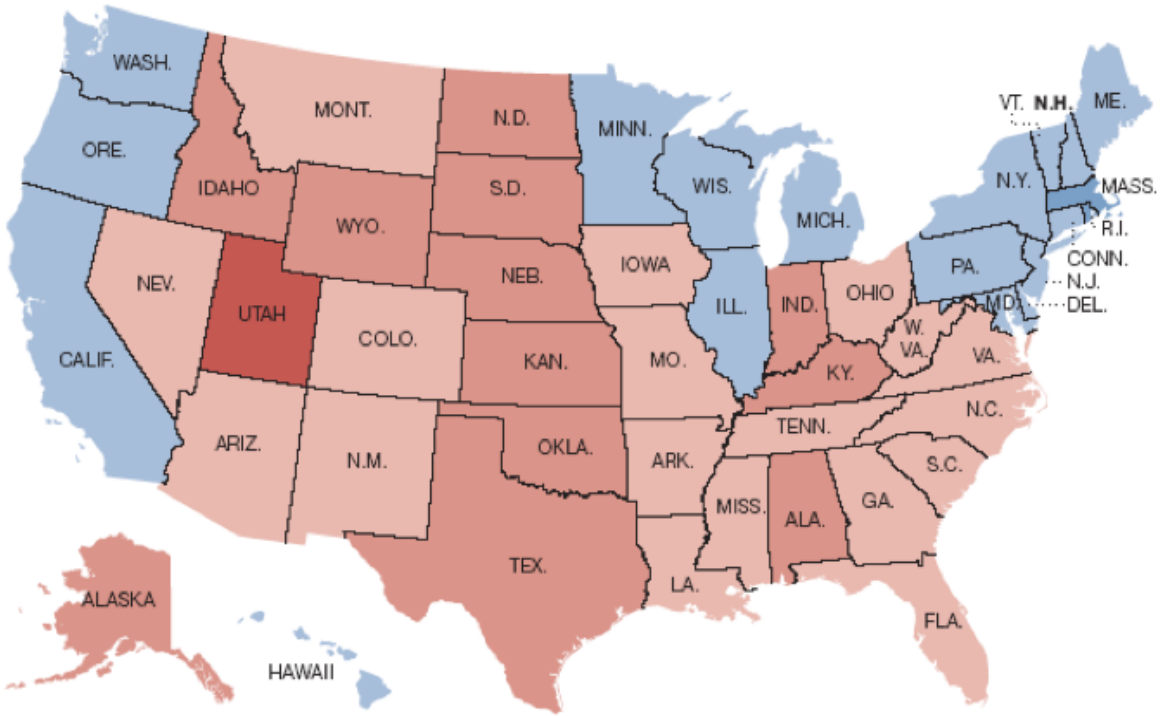
Scale

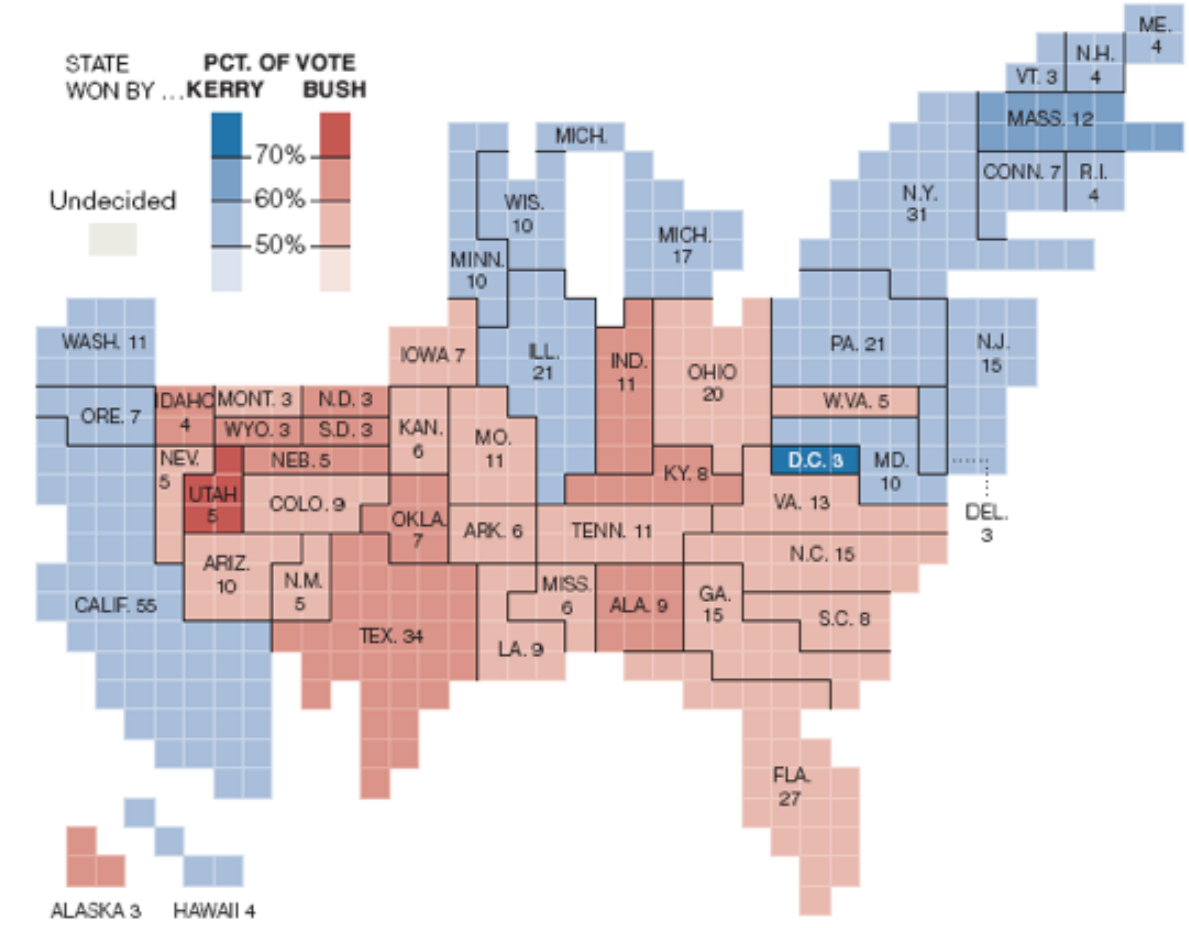
$$\text{Lie factor} = \frac{\text{size of effect in graph}}{\text{Size of effect in data}}$$





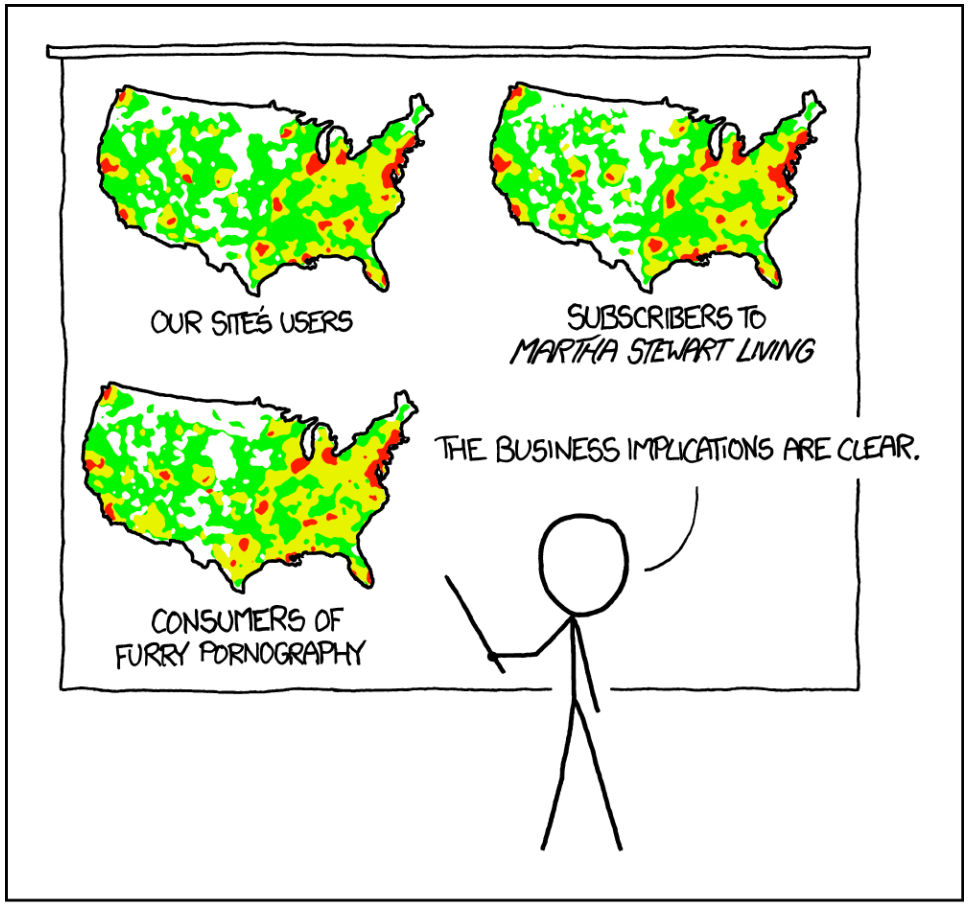
Misleading uses of area





The Problem with Absolute Values

- Sometimes, data should be relative



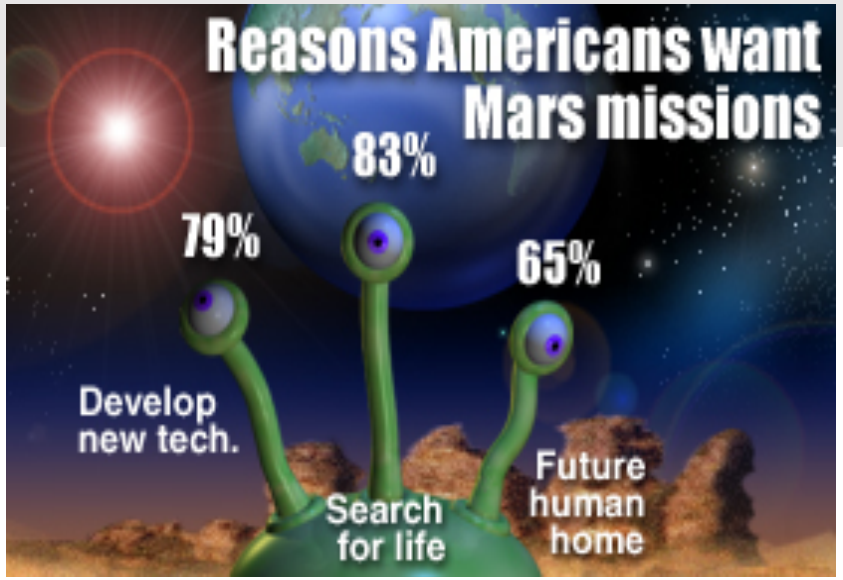
PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

Data/Ink Ratio

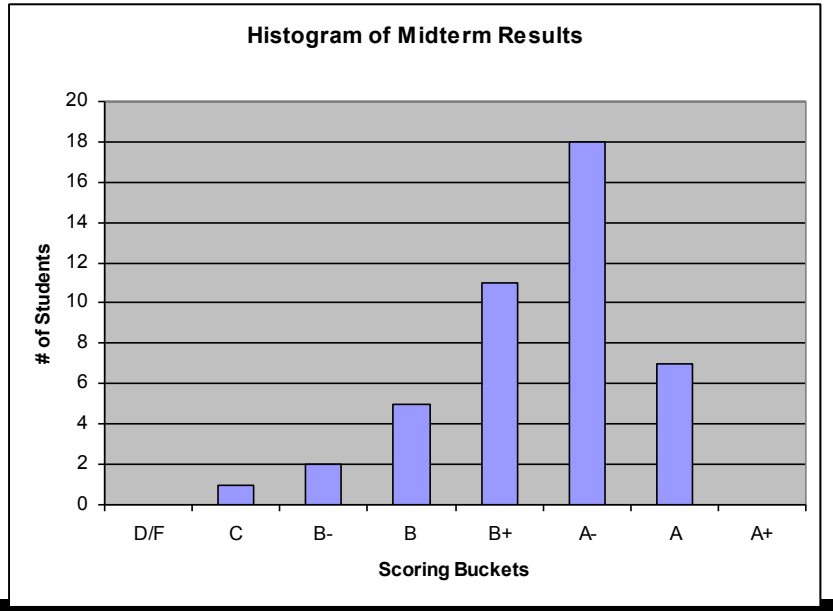
$$\text{Data-ink ratio} = \frac{\text{data-ink}}{\text{Total ink used to print graphic}}$$

= Proportion of a graphic's ink devoted to the non-redundant display of data-information.

< 0.001

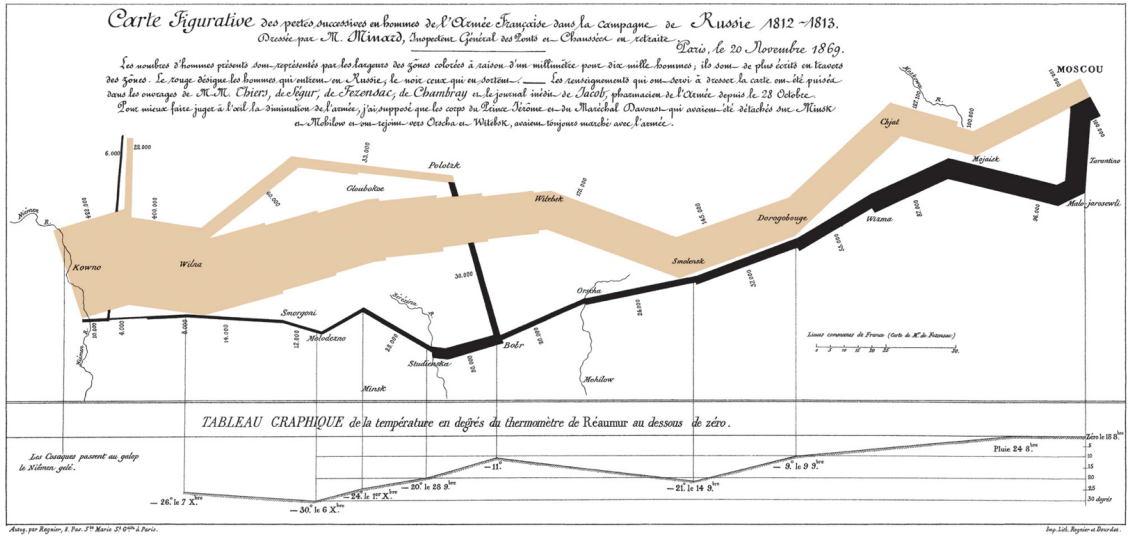


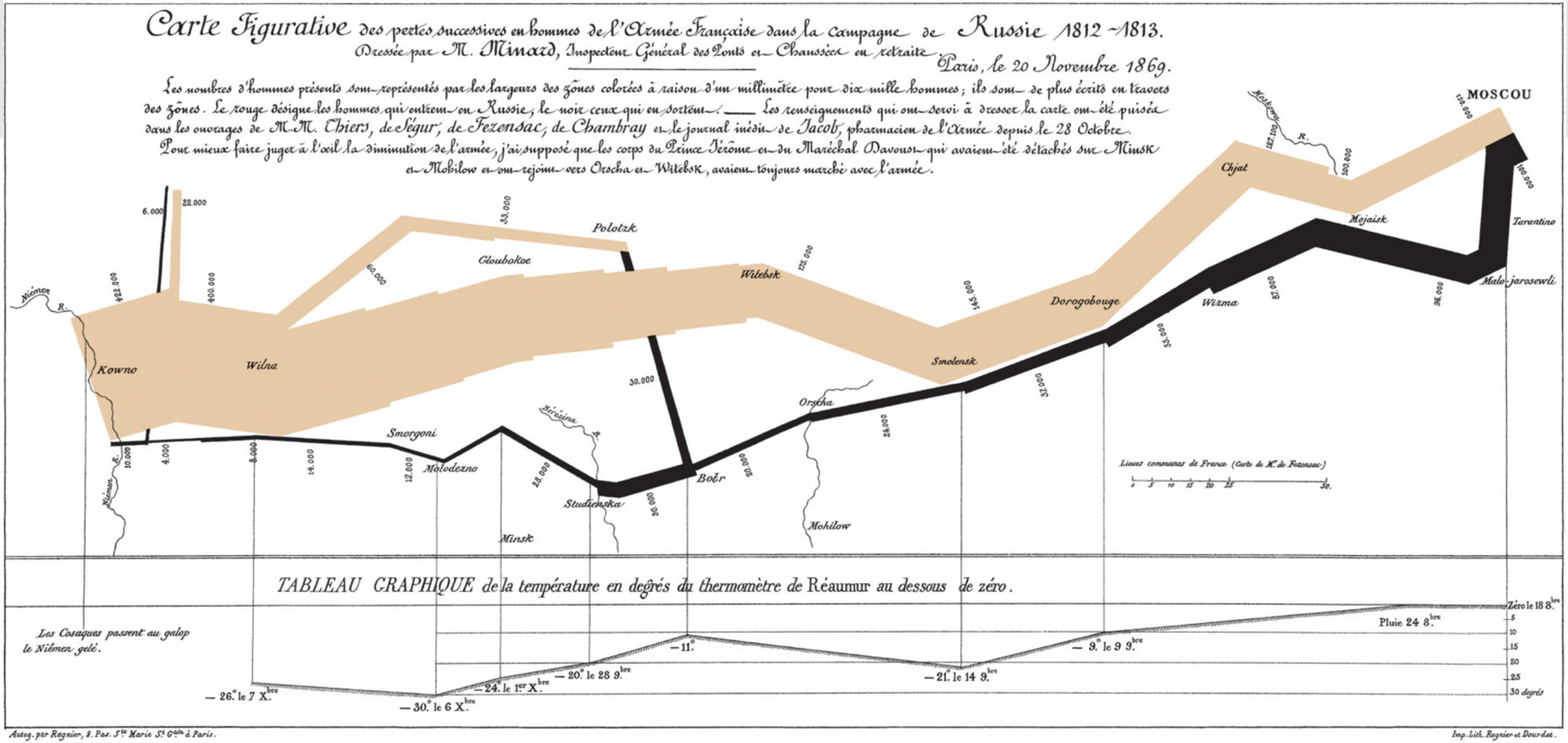
< 0.05 !!!



Helping Comparison

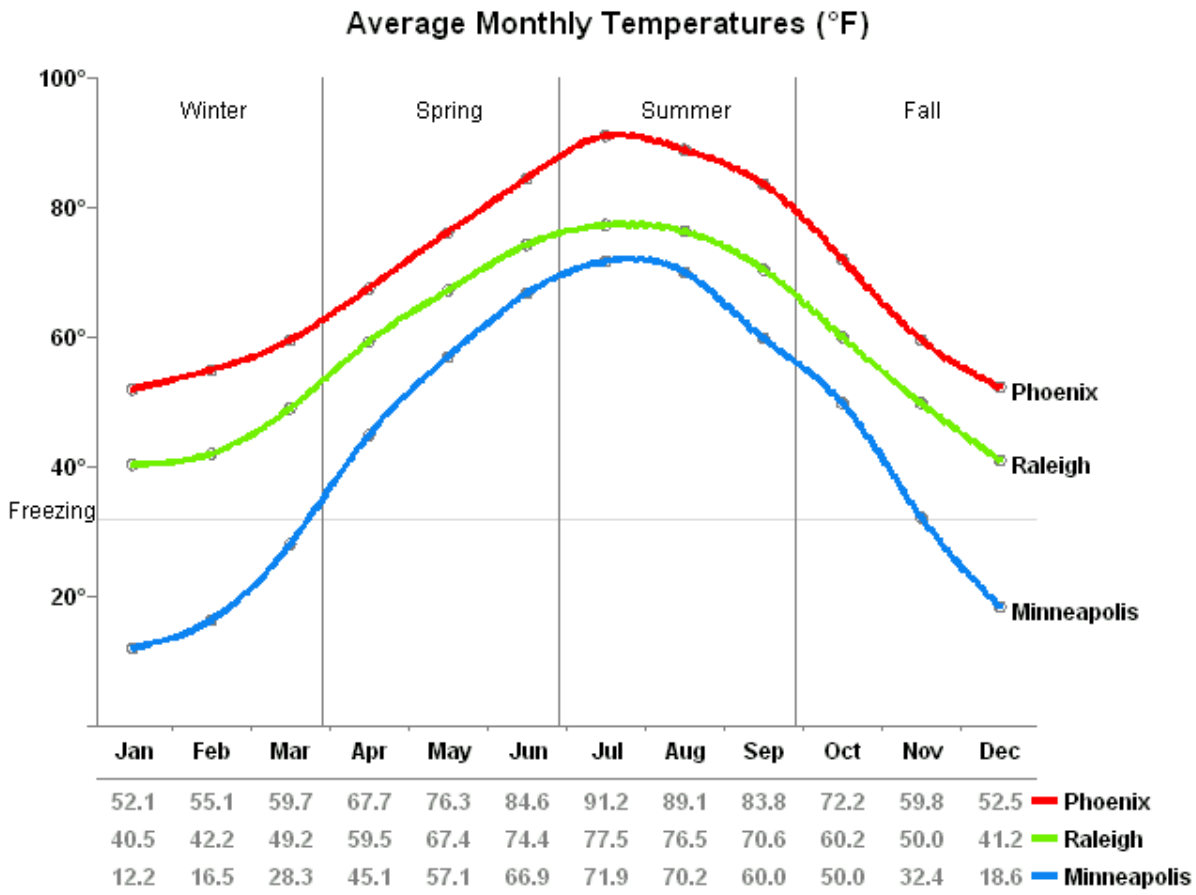
- Make the scale clear
  - Use intuitive scales
  - Make the baseline clear
- Provide context:
  - Examples
  - Generalization and Comparisons
  - Annotations and explanations





The French engineer, Charles Minard (1781-1870), illustrated the disastrous result of Napoleon's failed Russian campaign of 1812. The graph shows the size of the army by the width of the band across the map of the campaign on its outward and return legs, with temperature on the retreat shown on the line graph at the bottom.

Make comparisons explicit



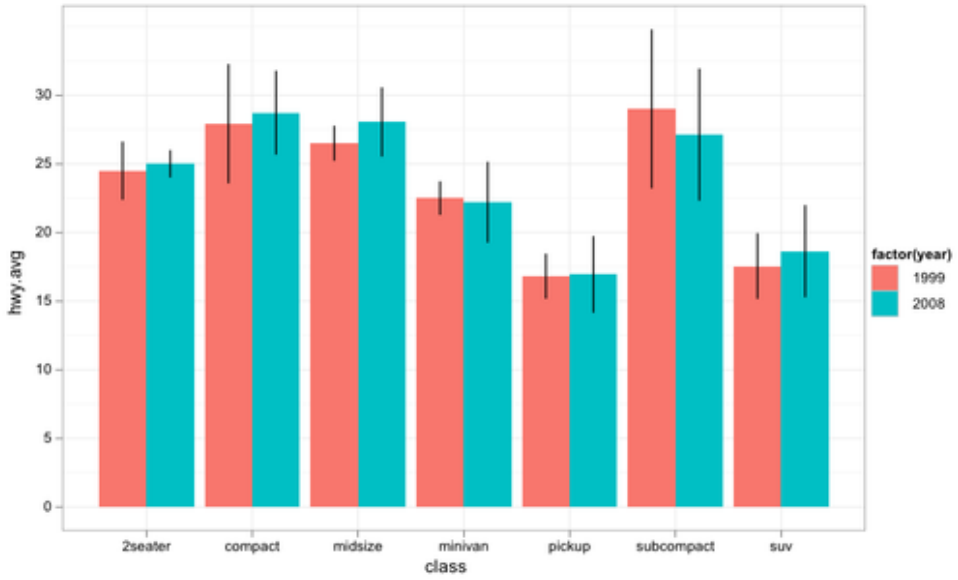
A great visualization can help you understand why the data is a certain

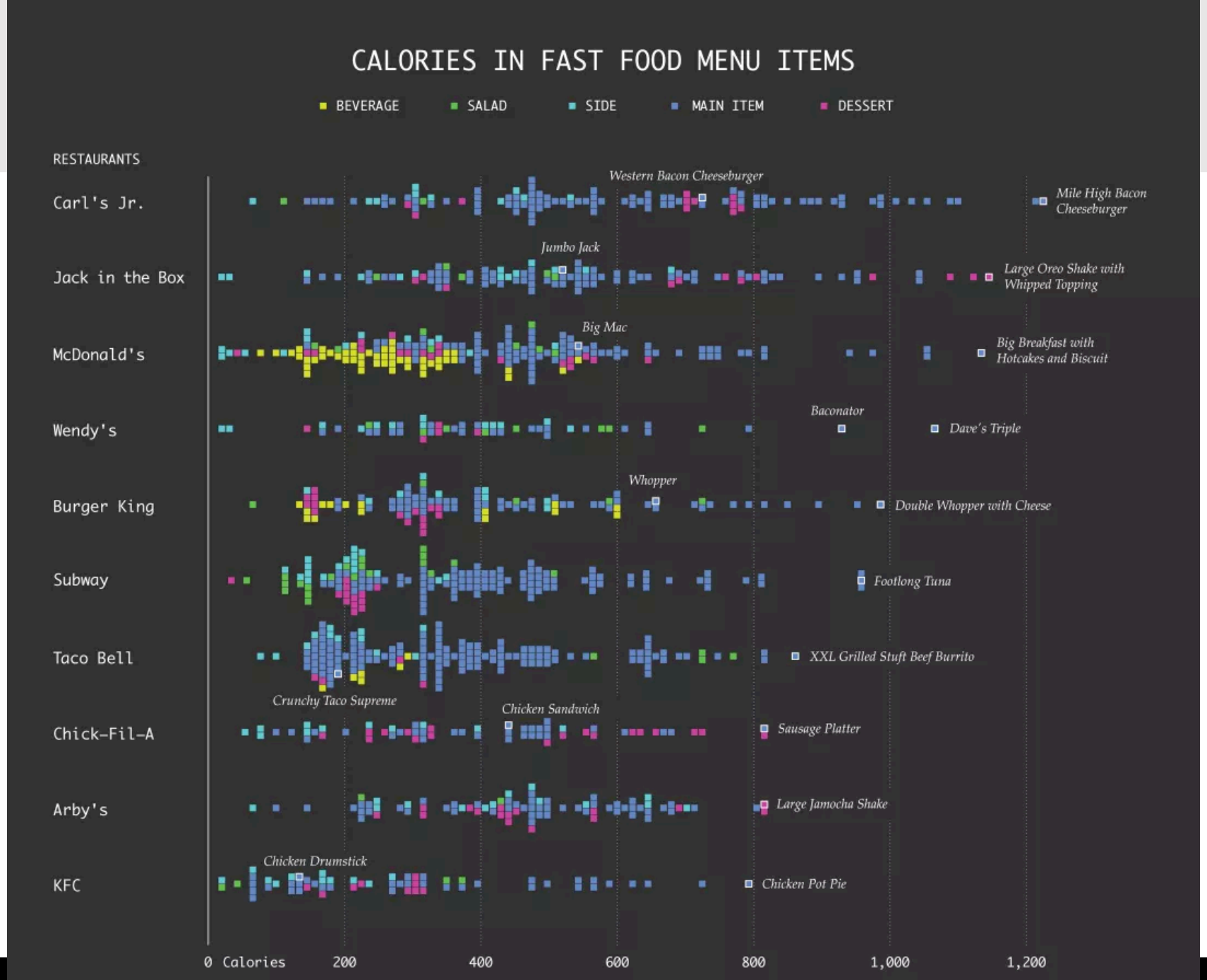
Baseline Example

Show baseline: The Dow Jones Average provides a baseline for comparison.

Providing Context: Distribution

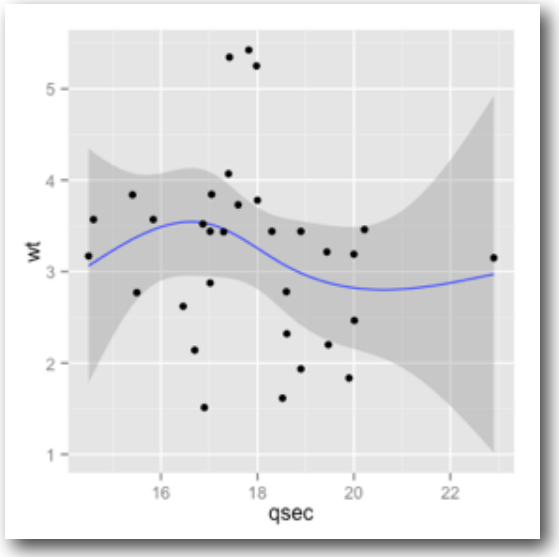
- If bars are used to represent the average, error bars are necessary!
- The error bar visualize some form of variance
- Mostly standard error or standard deviation.



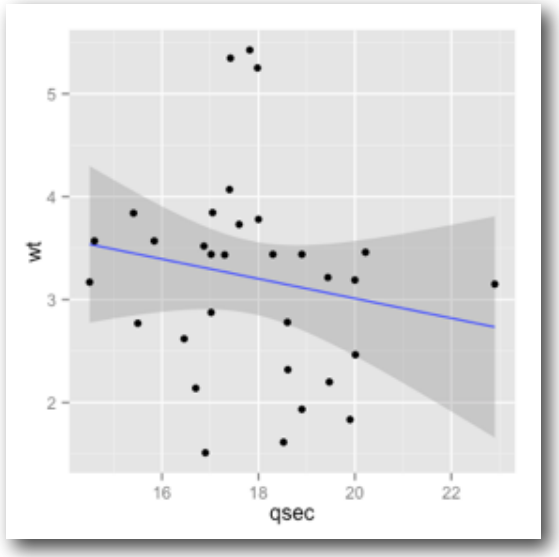

CALORIES IN FAST FOOD MENU ITEMS
BEVERAGE
SALAD
SIDE
MAIN ITEM
DESSERT
RESTAURANTS
Carl's Jr.
Western Bacon Cheeseburger
Mile High Bacon Cheeseburger
Jack in the Box
Jumbo Jack
Large Oreo Shake with Whipped Topping
McDonald's
Big Mac
Big Breakfast with Hotcakes and Biscuit
Wendy's
Baconator
Dave's Triple
Burger King
Whopper
Double Whopper with Cheese
Subway
Footlong Tuna
Taco Bell
XXL Grilled Stuft Beef Burrito
Crunchy Taco Supreme
Chick-Fil-A
Chicken Sandwich
Sausage Platter
Arby's
Large Jamocha Shake
KFC
Chicken Drumstick
Chicken Pot Pie
0 Calories
200
400
600
800
1,000
1,200

Context: Generalization

Adding generalizations to the graphs to provide aggregative data.



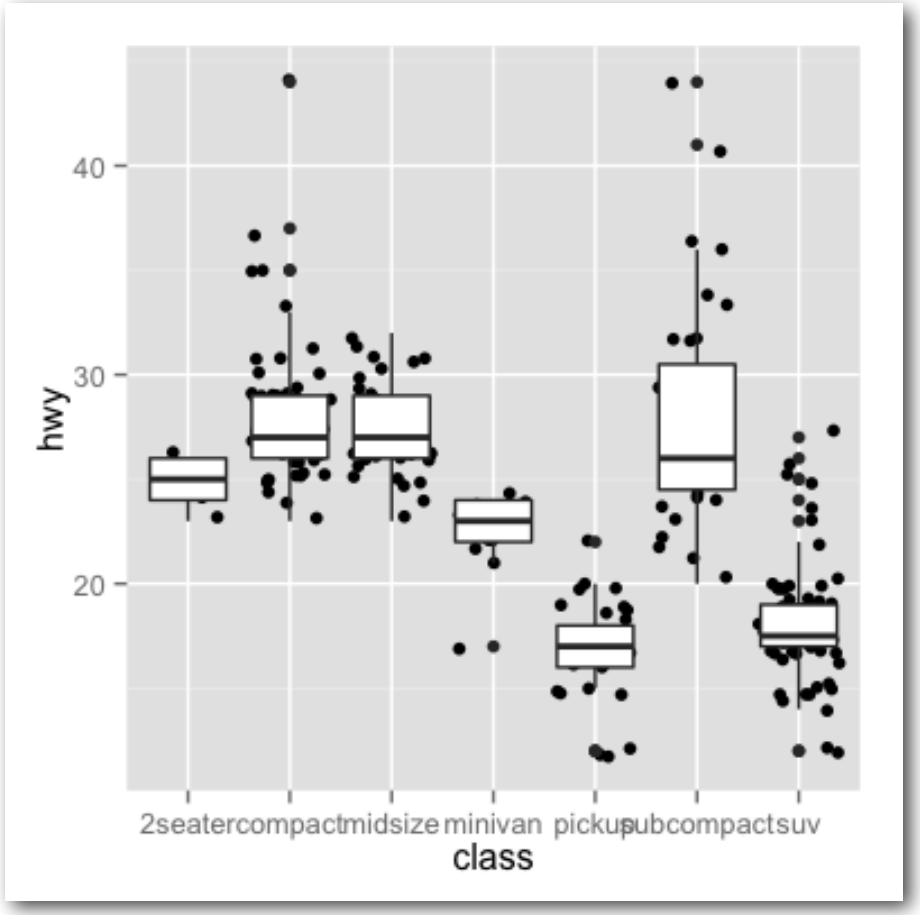
Local Regression Model



Linear Model

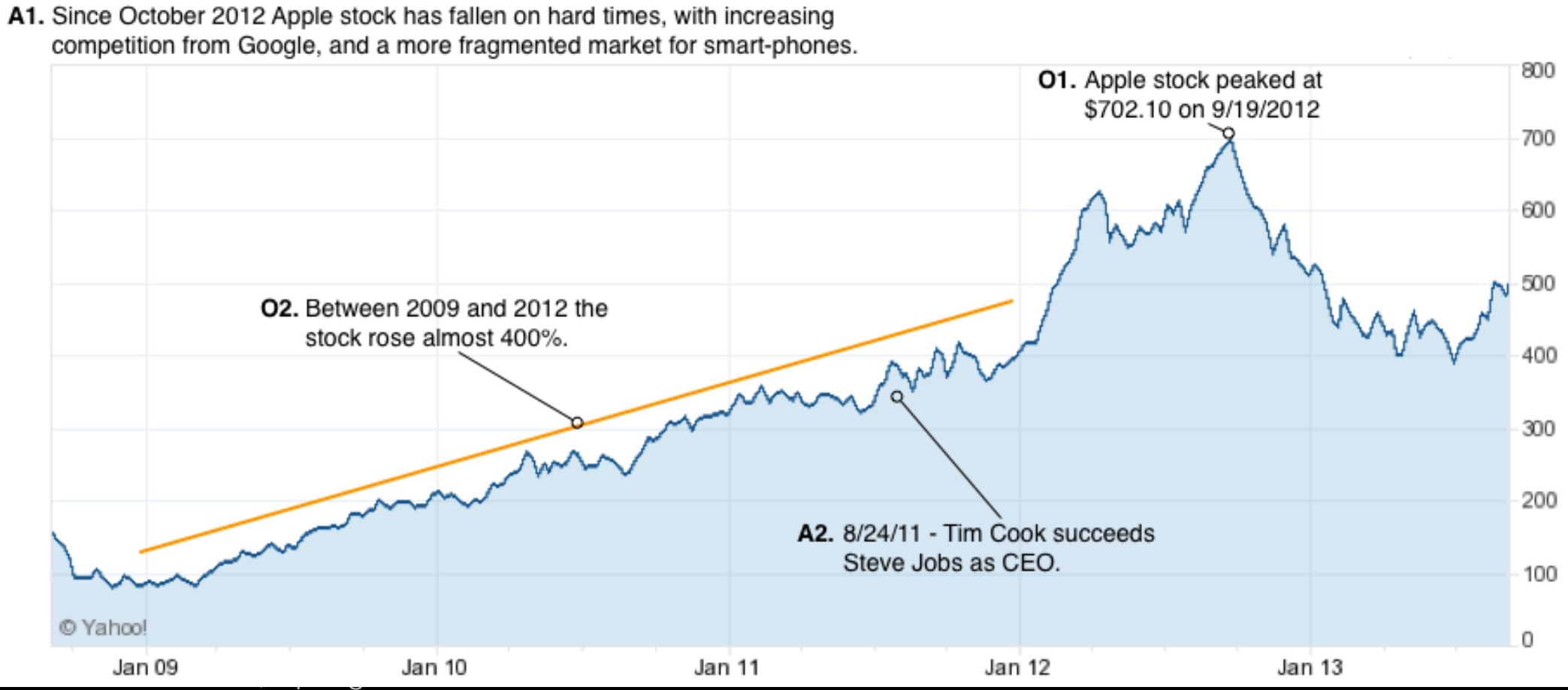
Context: Examples

Use example data to provide context to generalizations.

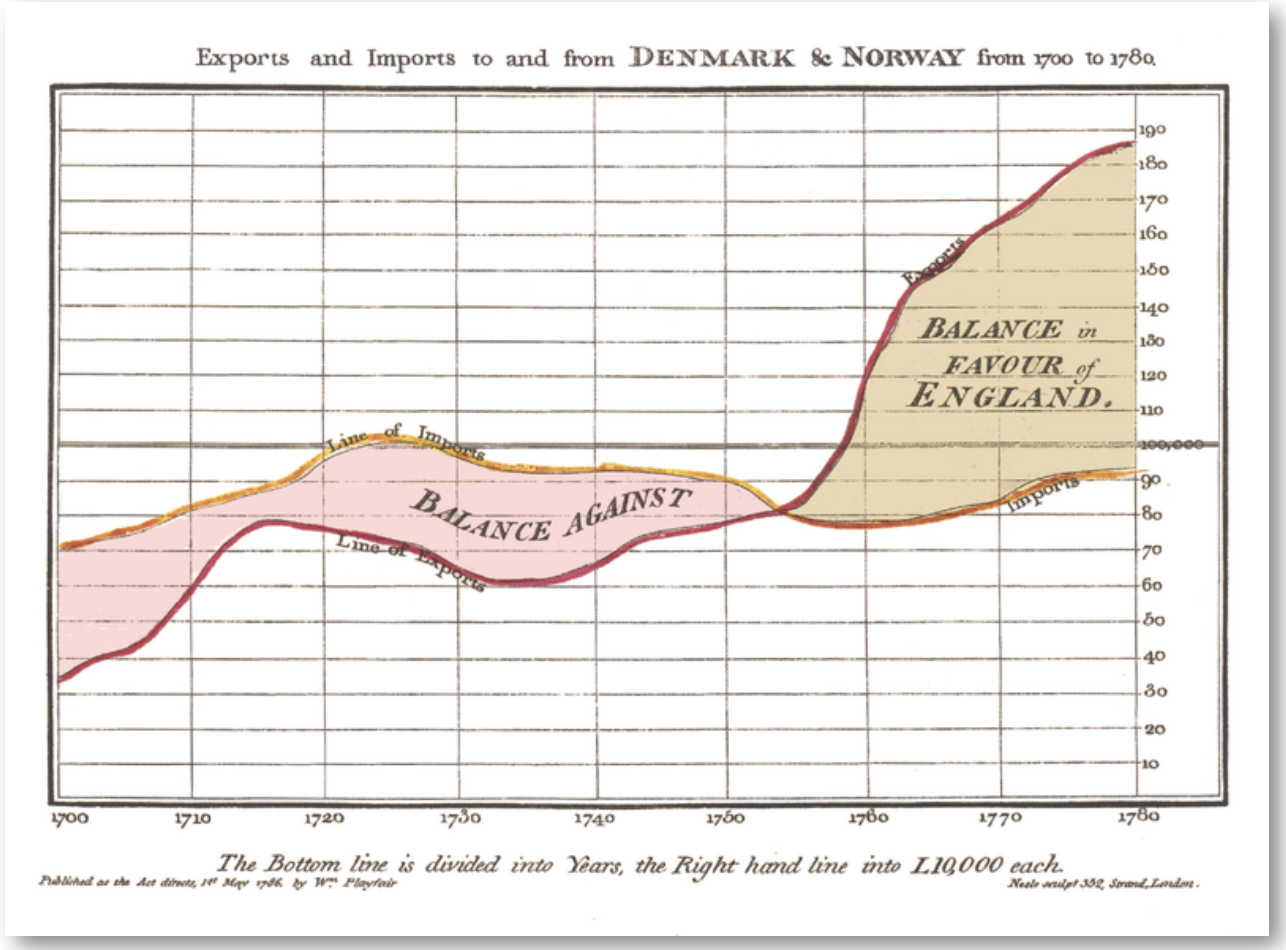


Annotations and explanations

Annotations allow people to understand the data by themselves, and draw conclusions from the data.



Time Series Visualization



Playfair's Denmark/Norway imports/exports, 1786

Summary