Data Visualization for Sensor Data Science

YoungTae Noh

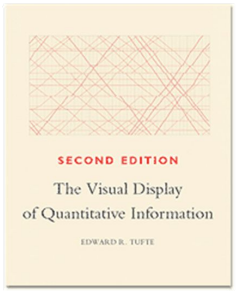
Overview

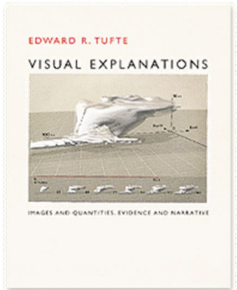
- Part I
  - Graphical Integrity – “functional art”
  - Visualization Design Principles
    - Maximize data-ink ratio
    - Avoid chart junk
    - Increase data density
  - Graphic Design Principles: CRAP
    - Contrast, Repetition, Alignment, Proximity
- Part II
  - Data Visualization Steps & Visual Encoding
  - Visualization Taxonomy & Statistical Graphs – A Tour through the visualization zoo

Part 1

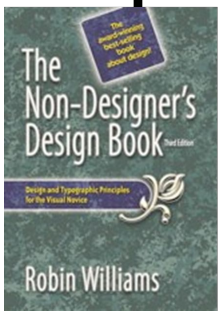
- Graphical Integrity
- Visualization Design Principles
- Graphic Design Principles

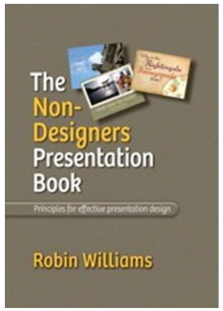
Edward Tufte 





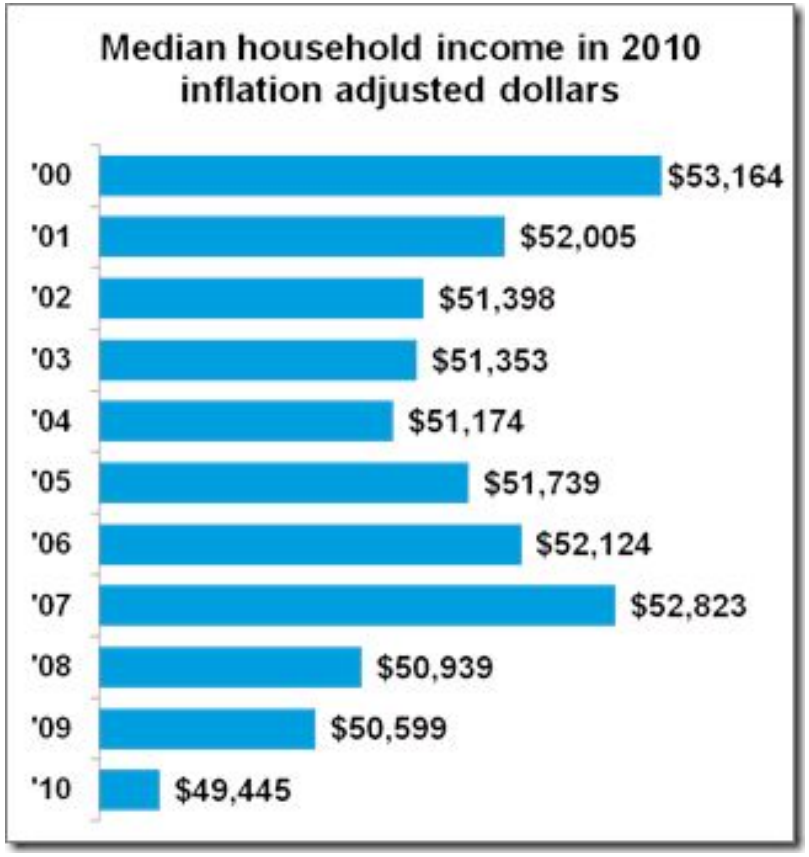
Robin Williams 





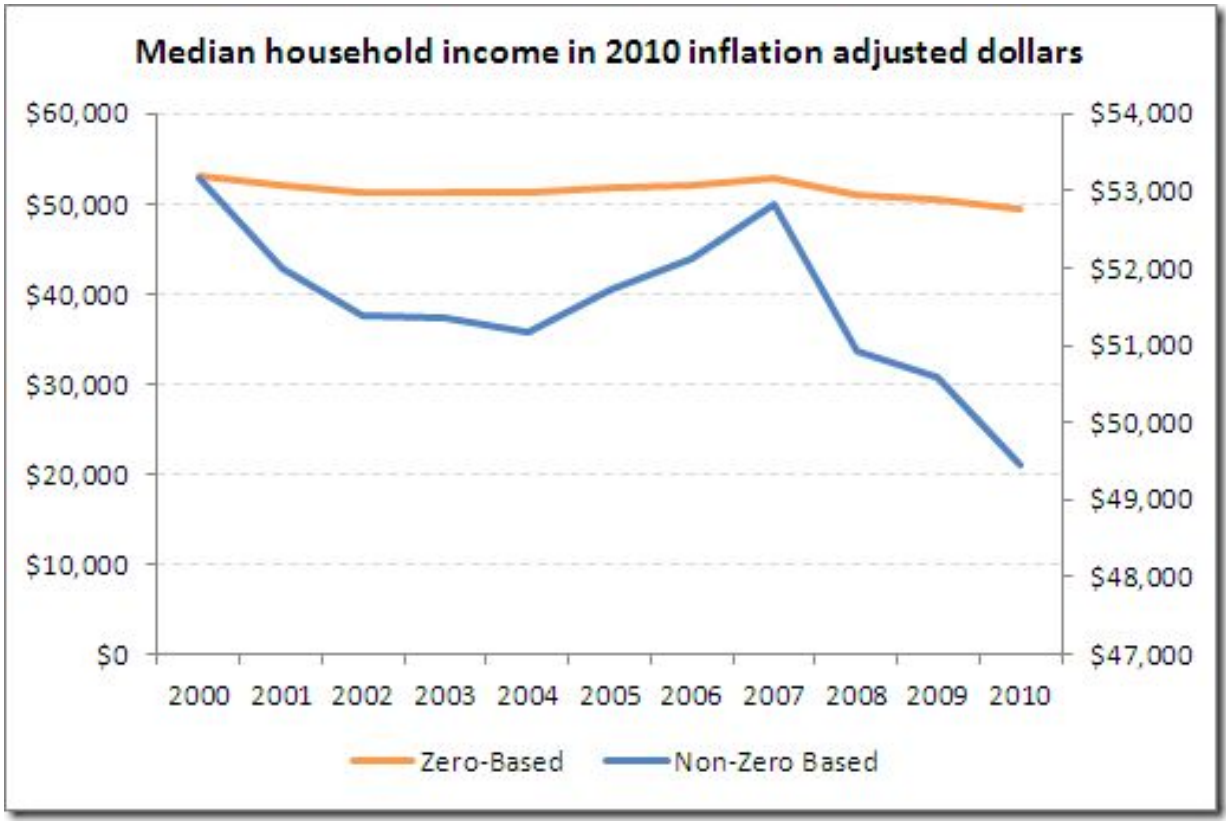
Graphical Integrity

Scale Distortions



Nielson

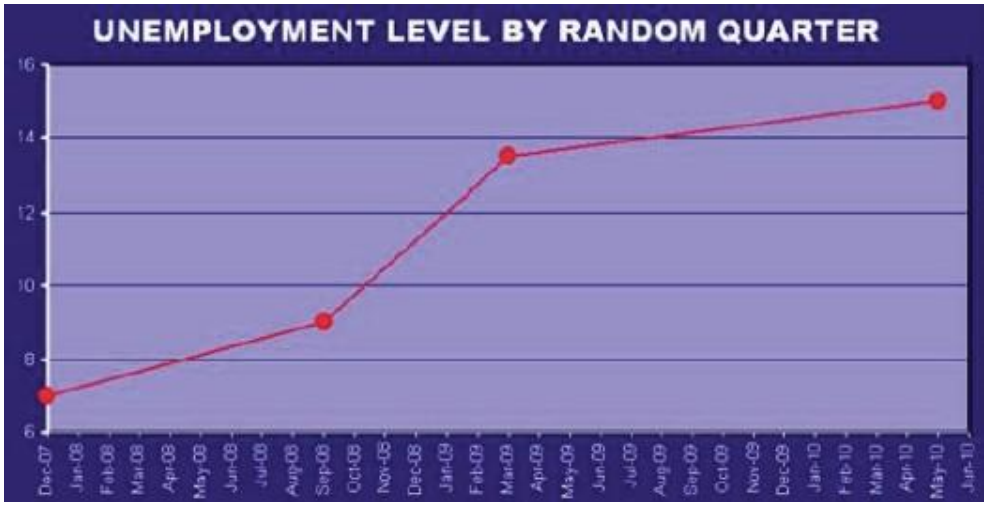
Scale Distortions



A. Kriebel, VizWiz

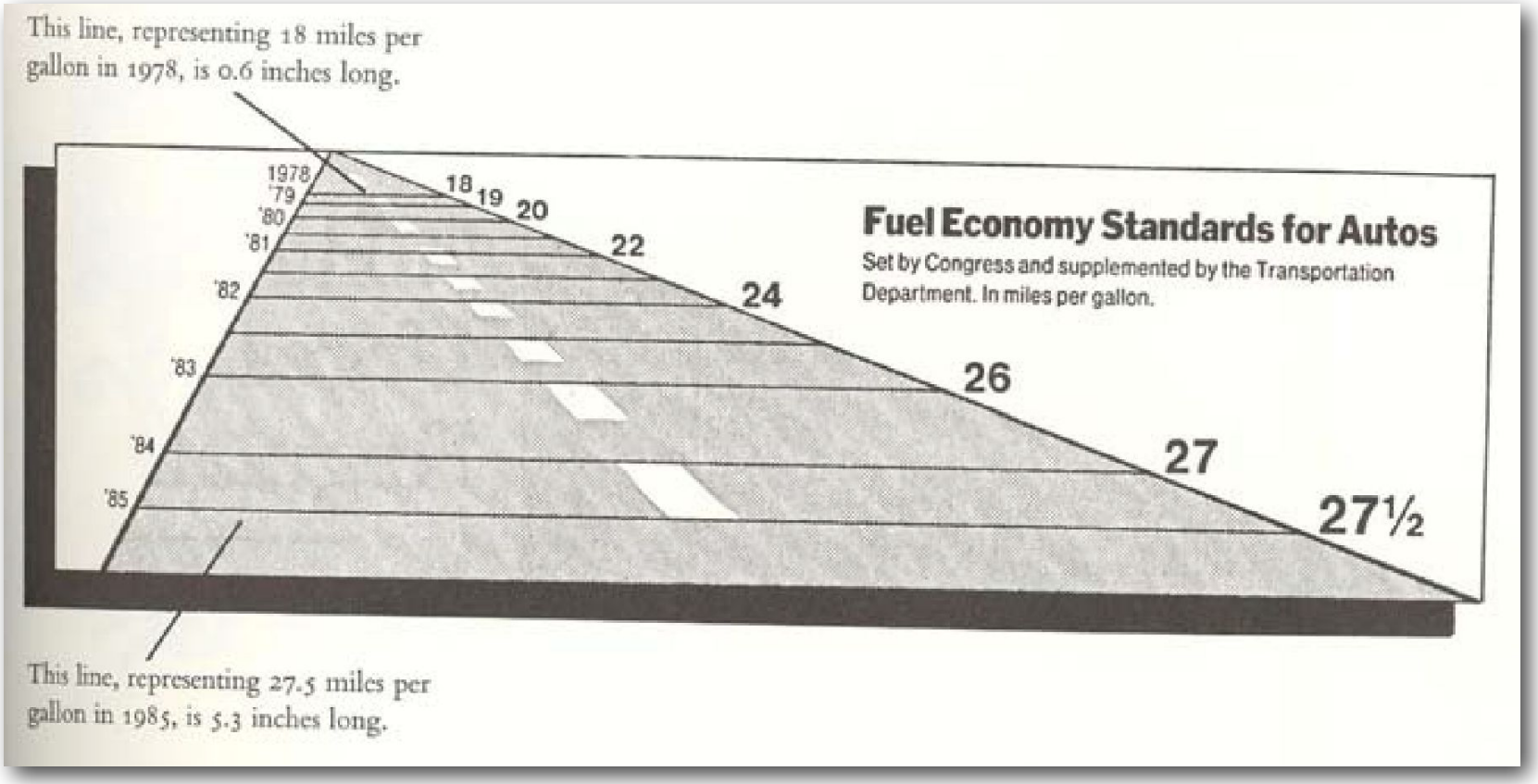
Scale Distortions

Scale Distortions



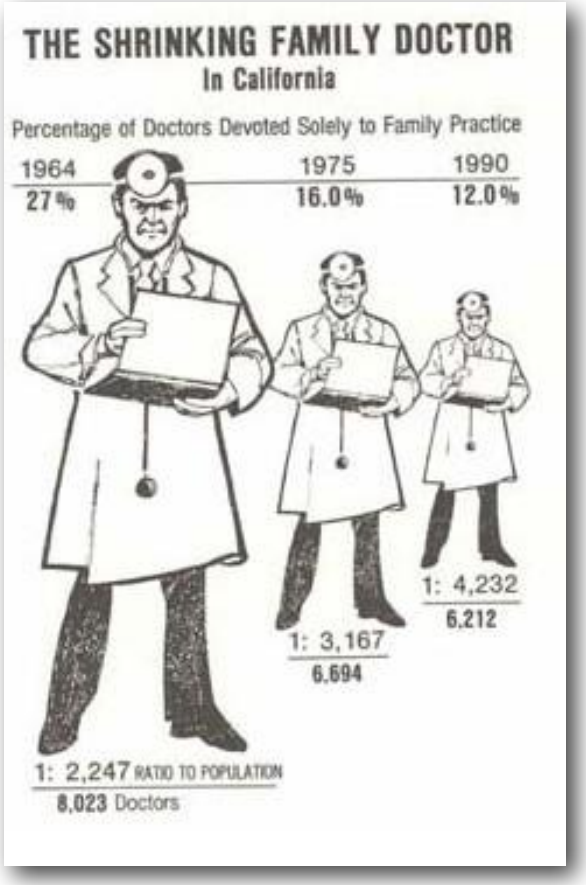
The Lie Factor

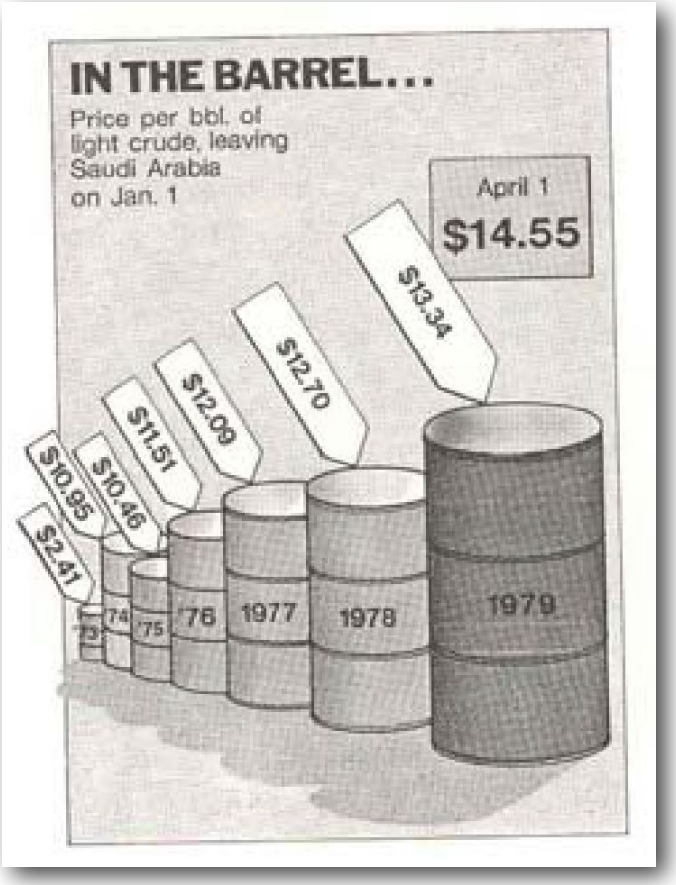
$$\frac{\text{Size of effect shown in graphic}}{\text{Size of effect in data}}$$

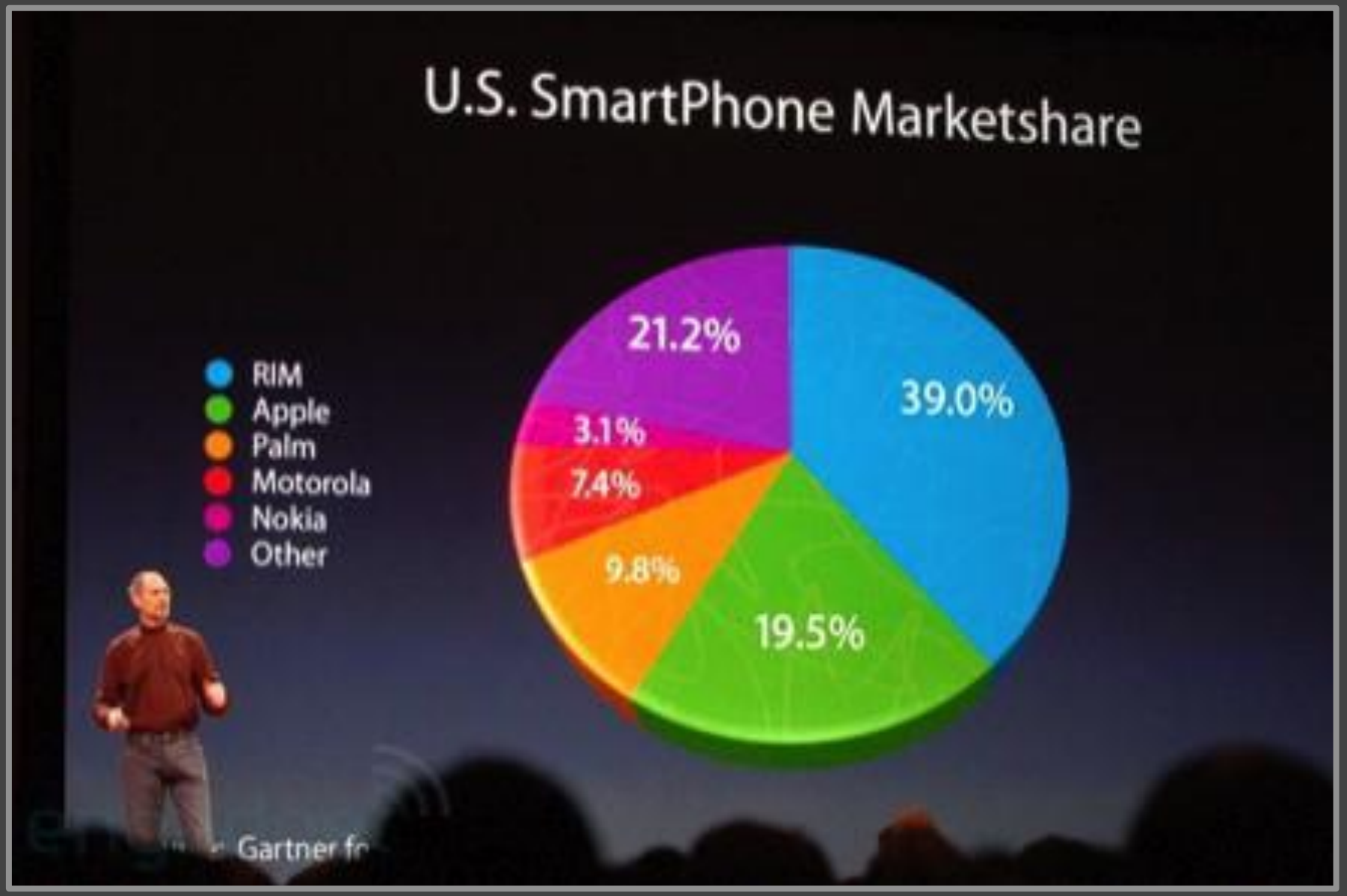


Tufte, VDQI

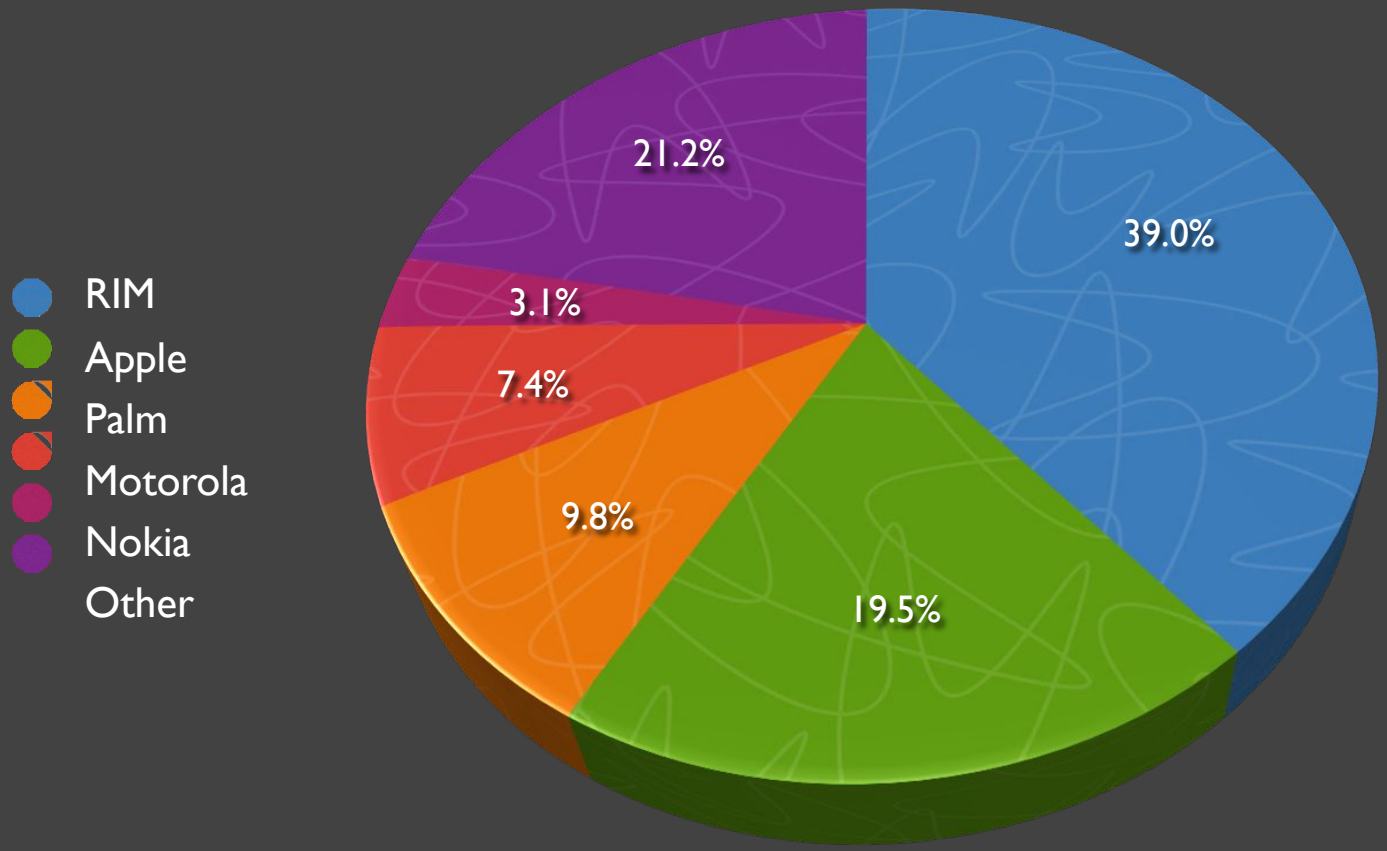
The Lie Factor



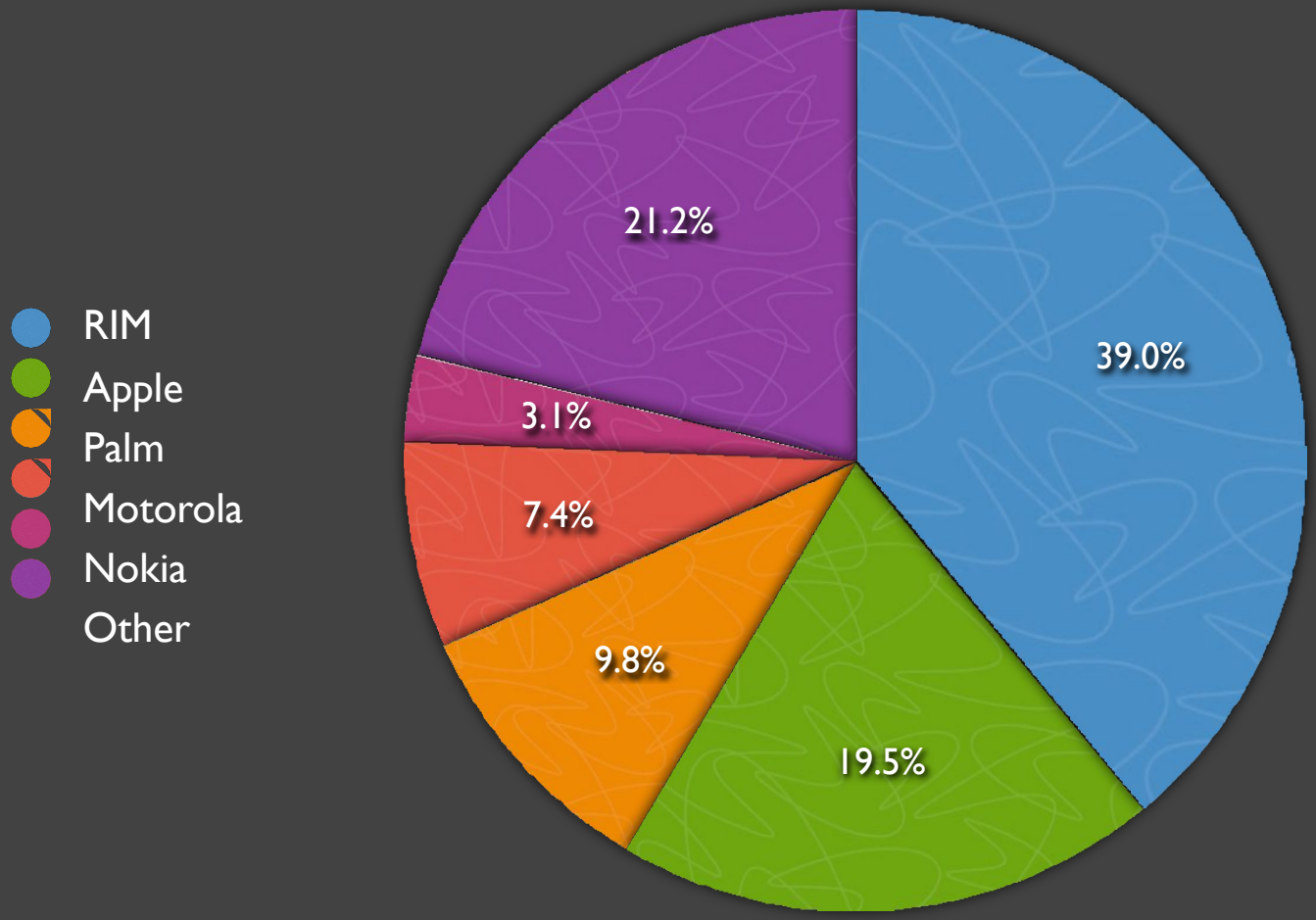



U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
39.0%
19.5%
9.8%
7.4%
3.1%
21.2%
Gartner fr

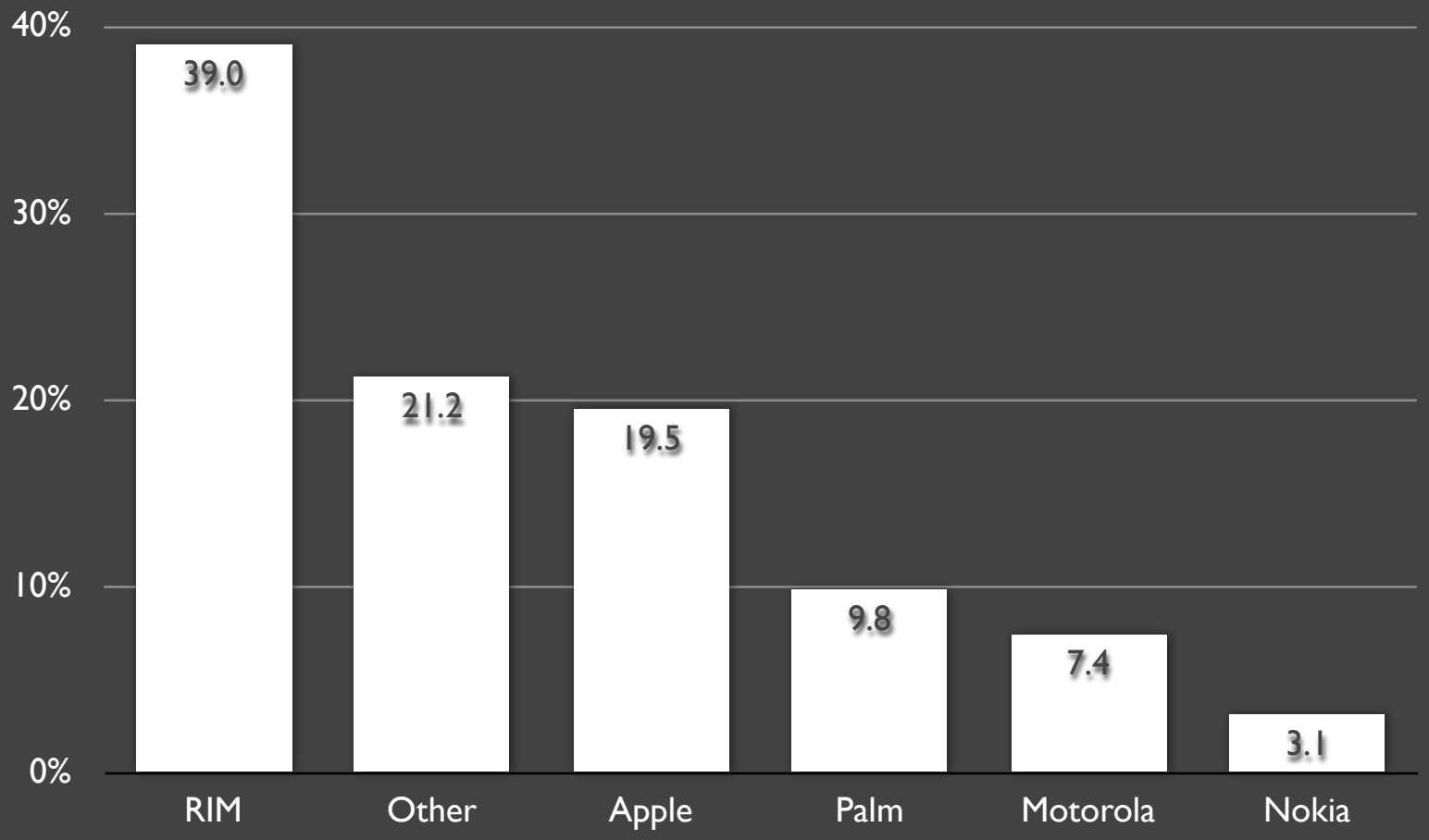
U.S. SmartPhone Marketshare



U.S. SmartPhone Marketshare



U.S. SmartPhone Marketshare



Tufte'sIntegrity Principles

- Clear, detailed, and thorough labeling and appropriate scales
- Size of the graphic effect should be directly proportional to the numerical quantities ("lie factor")
- Show data variation, not design variation

Visualization
Design Principles

Maximize Data-Ink Ratio

, with reason

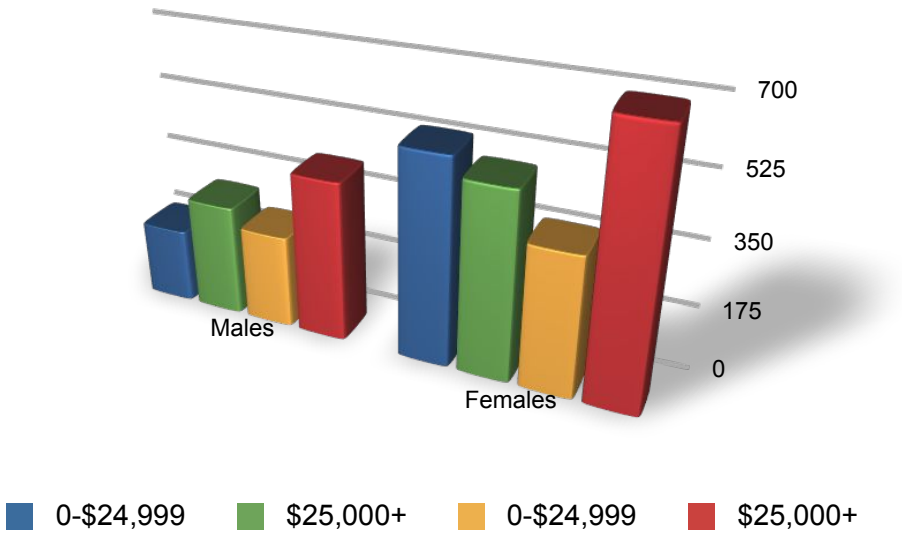
$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$

Above all else show the data!

Maximize Data-Ink Ratio

, with reason

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$

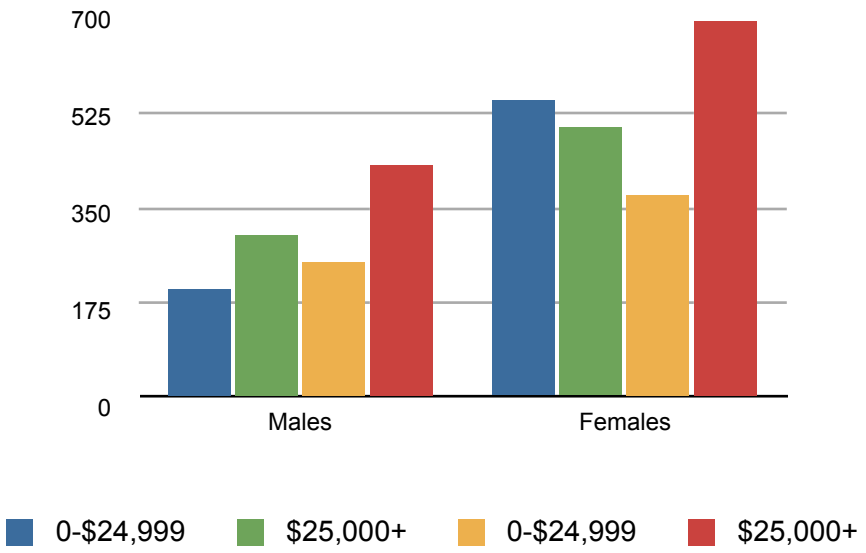


How can we maximize data-ink ratio?

Maximize Data-Ink Ratio

, with reason

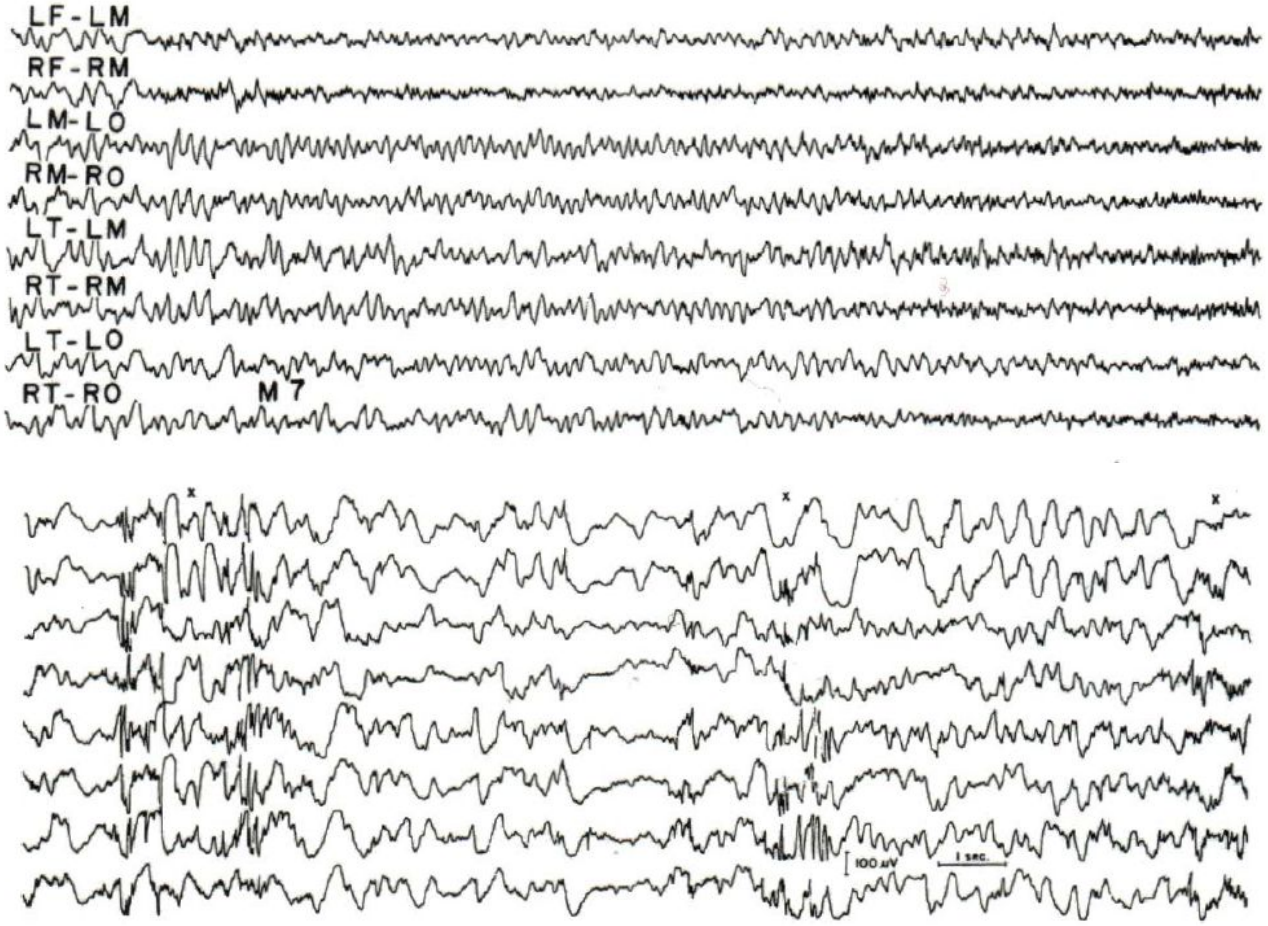
$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$



After erasing non-data ink

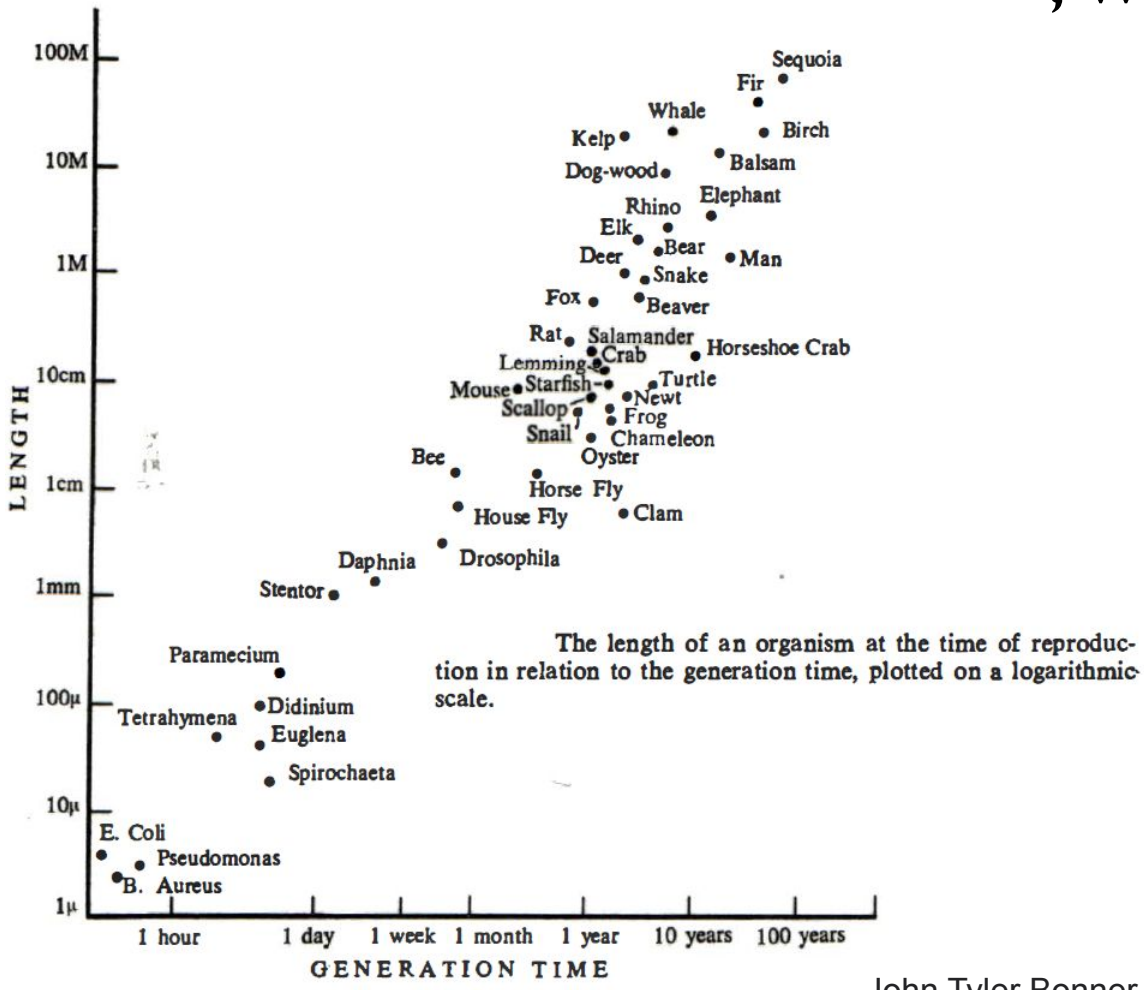
Maximize Data-Ink Ratio

, with reason



Maximize Data-Ink Ratio

, with reason

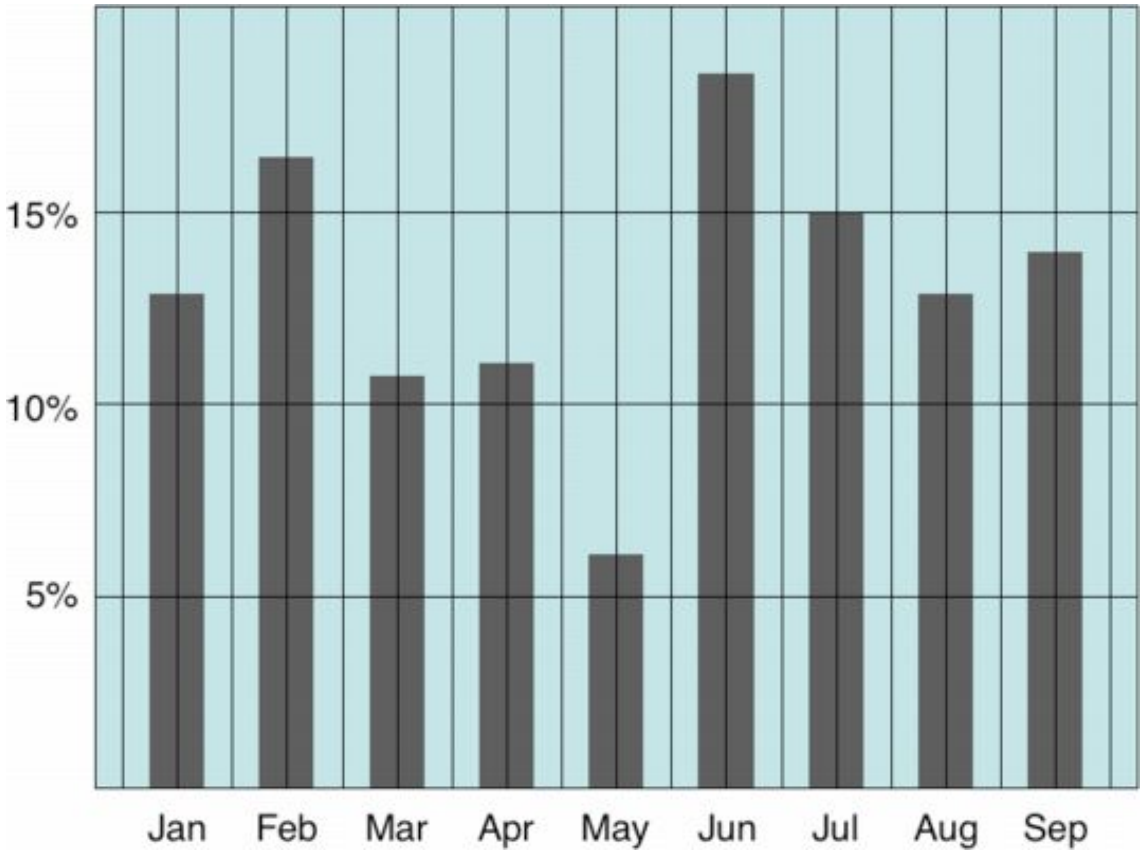


The length of an organism at the time of reproduction in relation to the generation time, plotted on a logarithmic scale.

John Tyler Bonner, Size and Cycle: An Essay on the Structure of Biology (Princeton, 1965), p. 17.

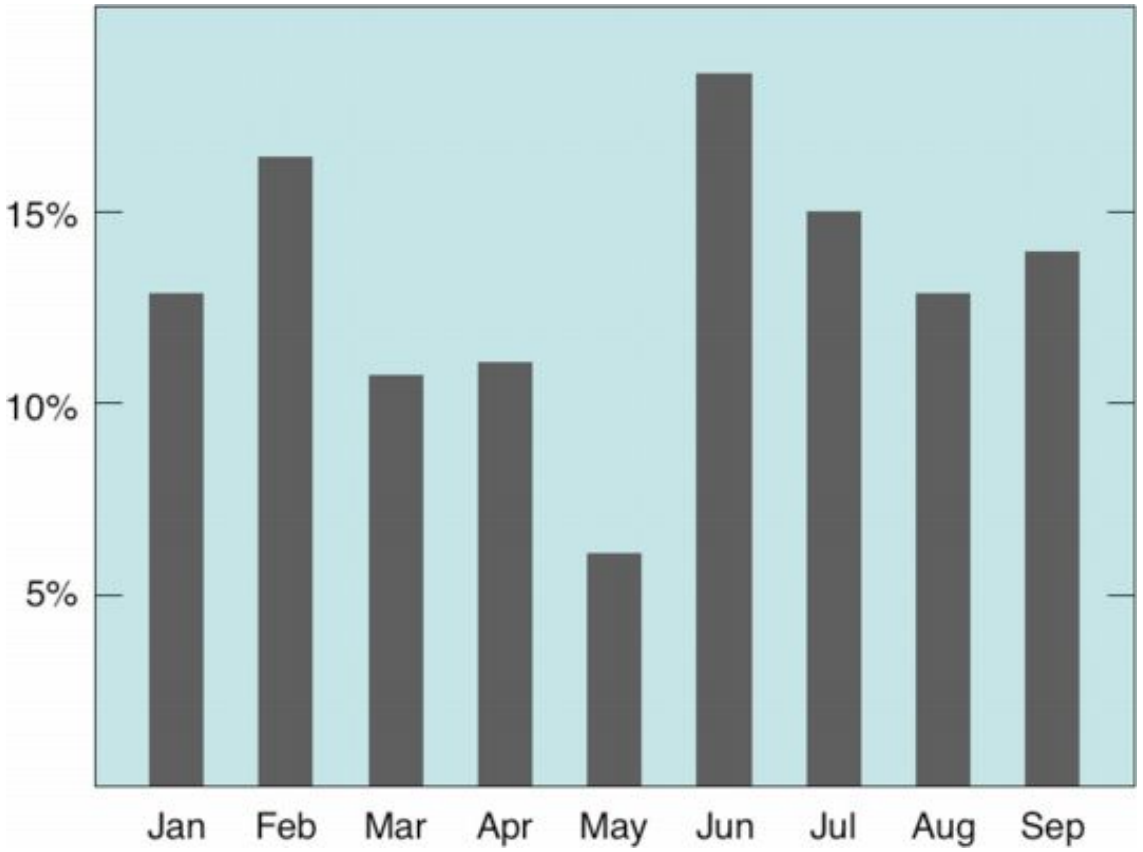
Avoid Chartjunk

Extraneous visual elements that distract from the message



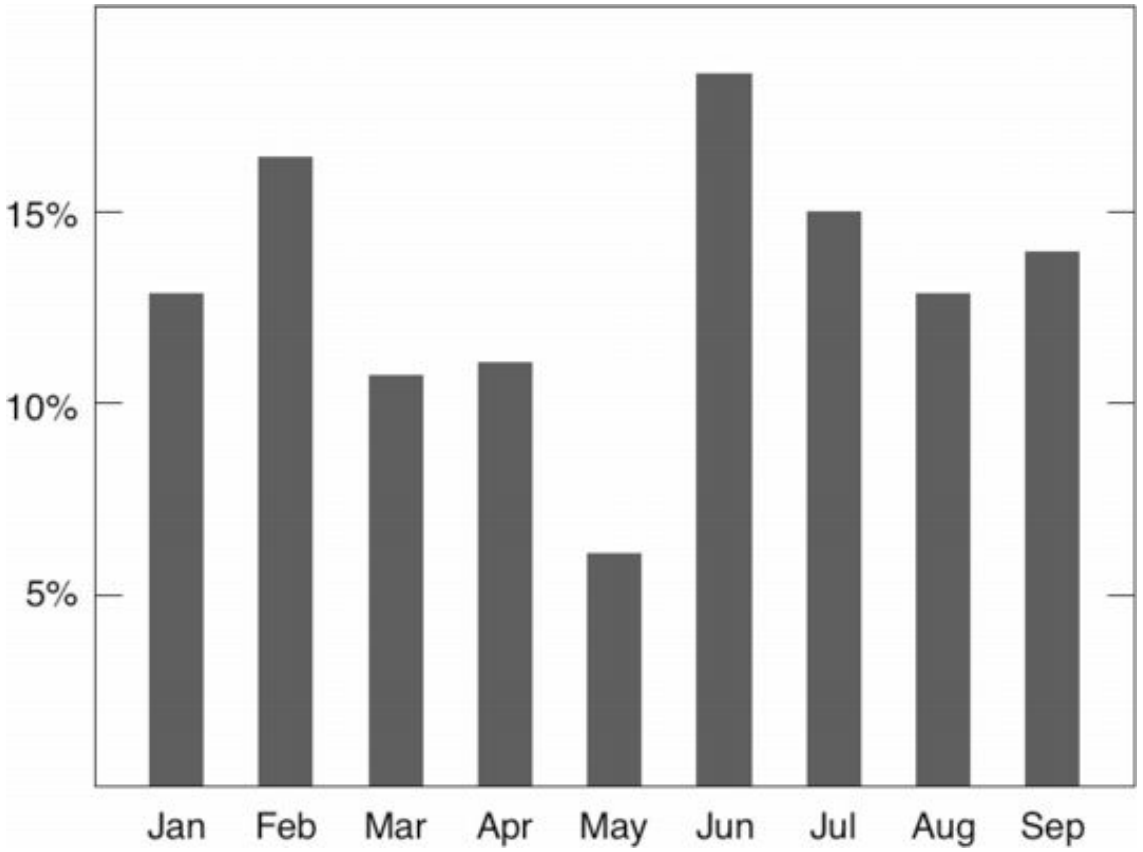
ongoing,Tim Brey

Avoid Chartjunk



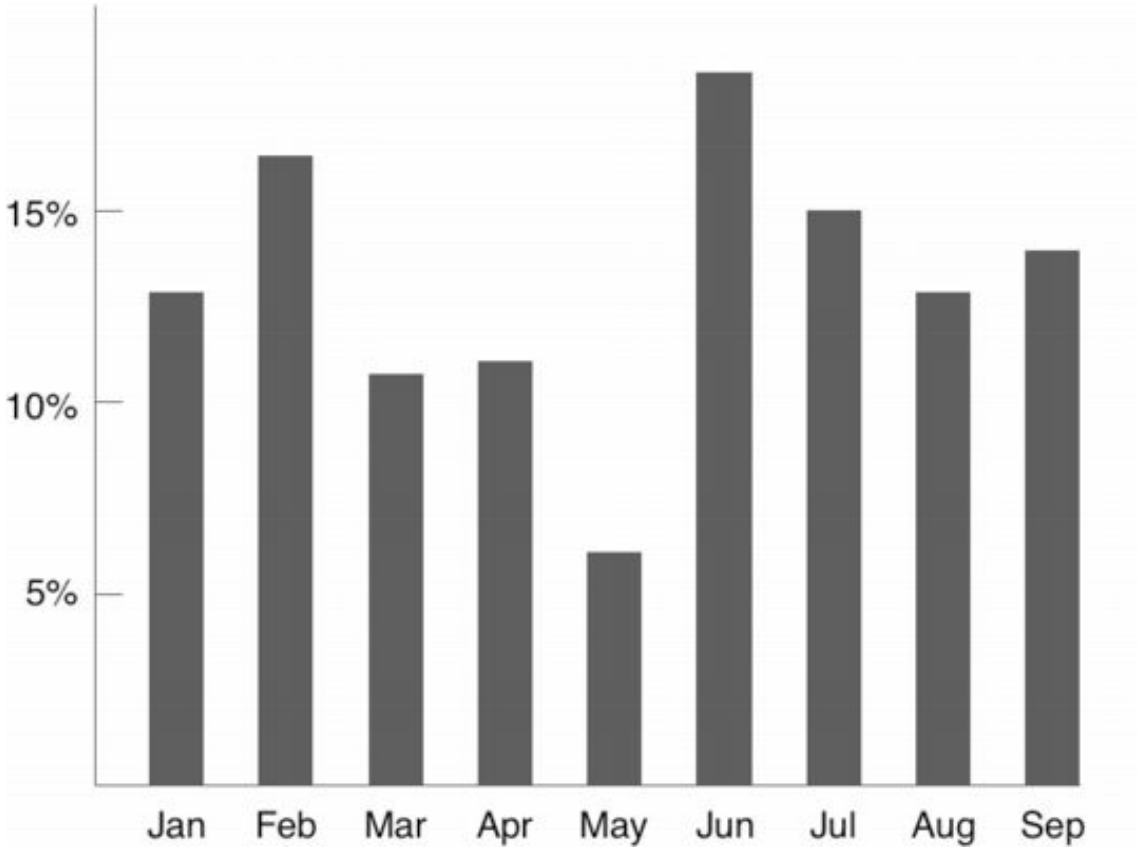
ongoing, Tim Brey

Avoid Chartjunk



ongoing, Tim Brey

Avoid Chartjunk



ongoing, Tim Brey

Avoid Chartjunk

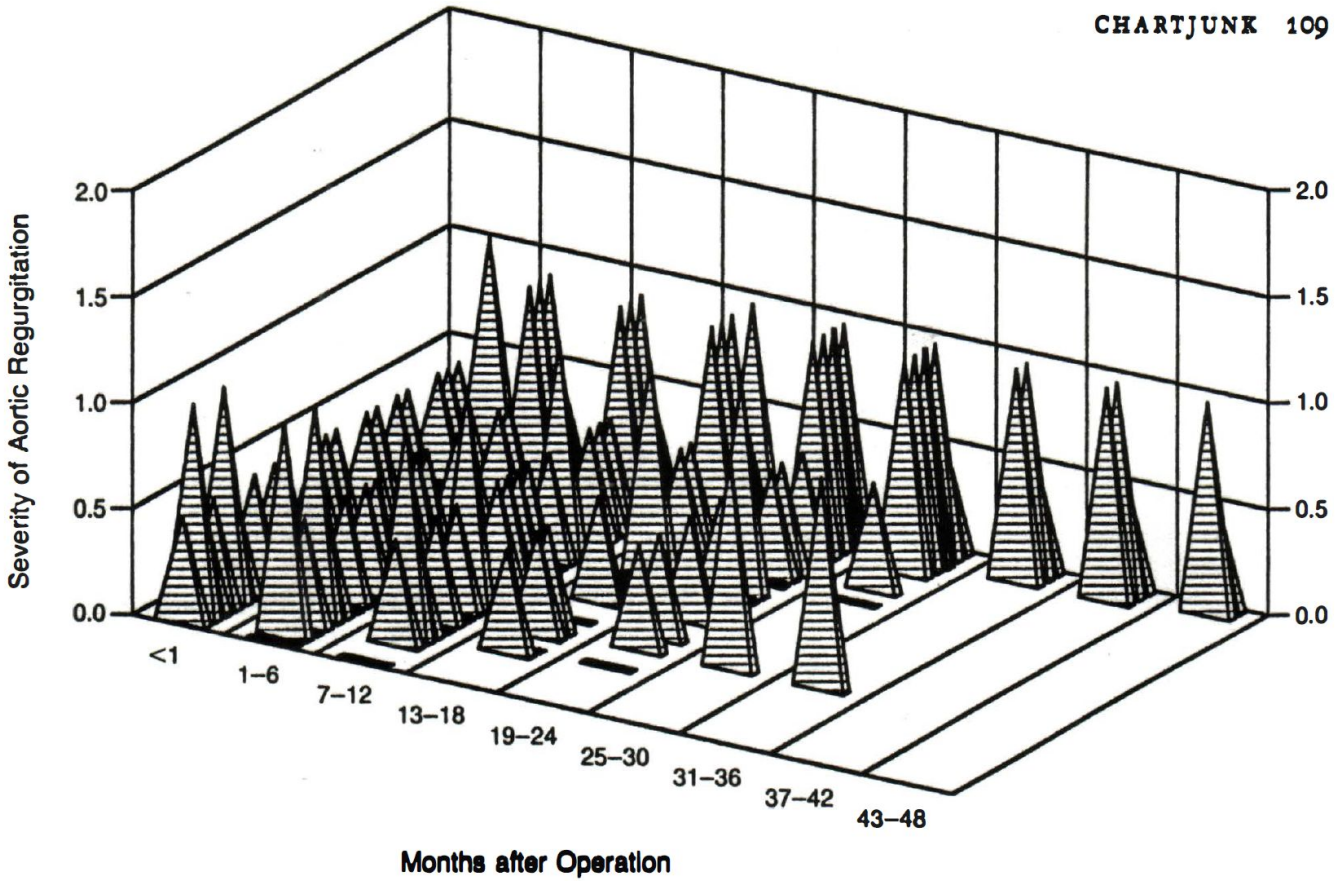
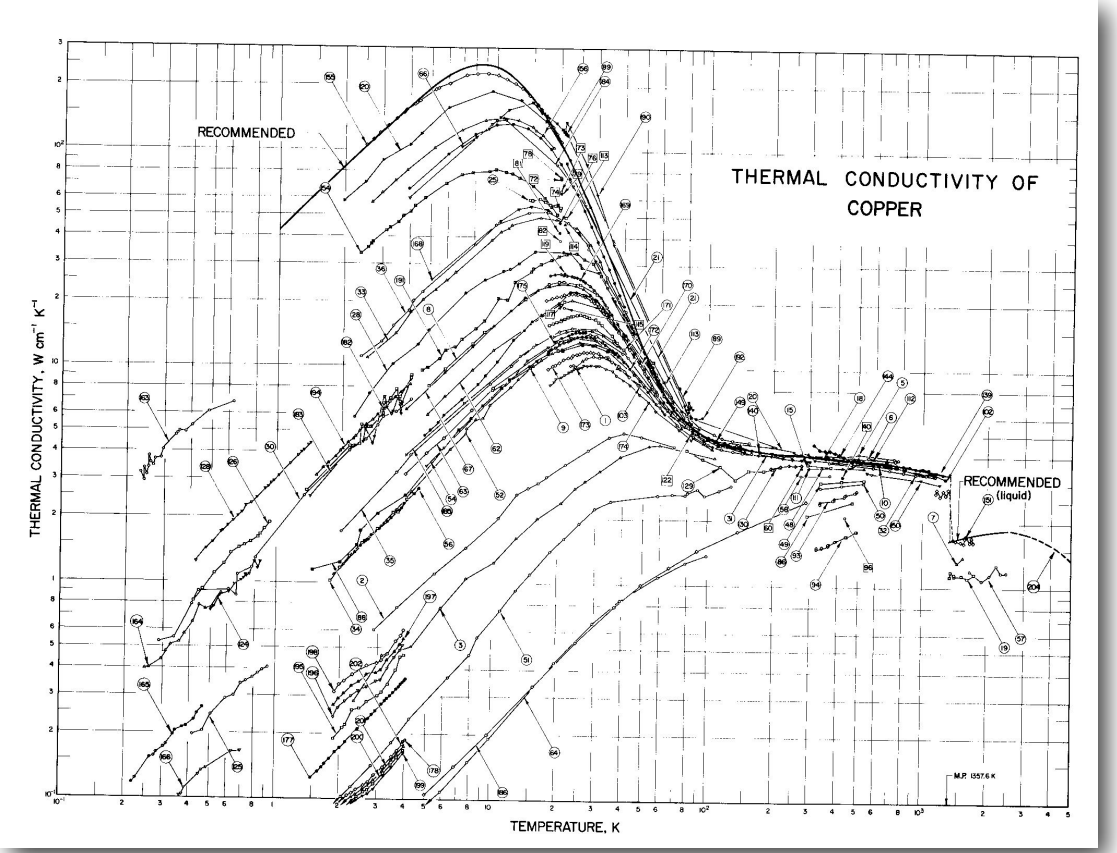


Figure 2. Serial Echocardiographic Assessments of the Severity of Regurgitation in the Pulmonary Autograft in 31 Patients. The numerical grades were assigned according to the severity of regurgitation, as follows: 0, none; 0.5, trivial; 1.0 to 1.5, mild; 2.0, moderate; and 3.0, severe.

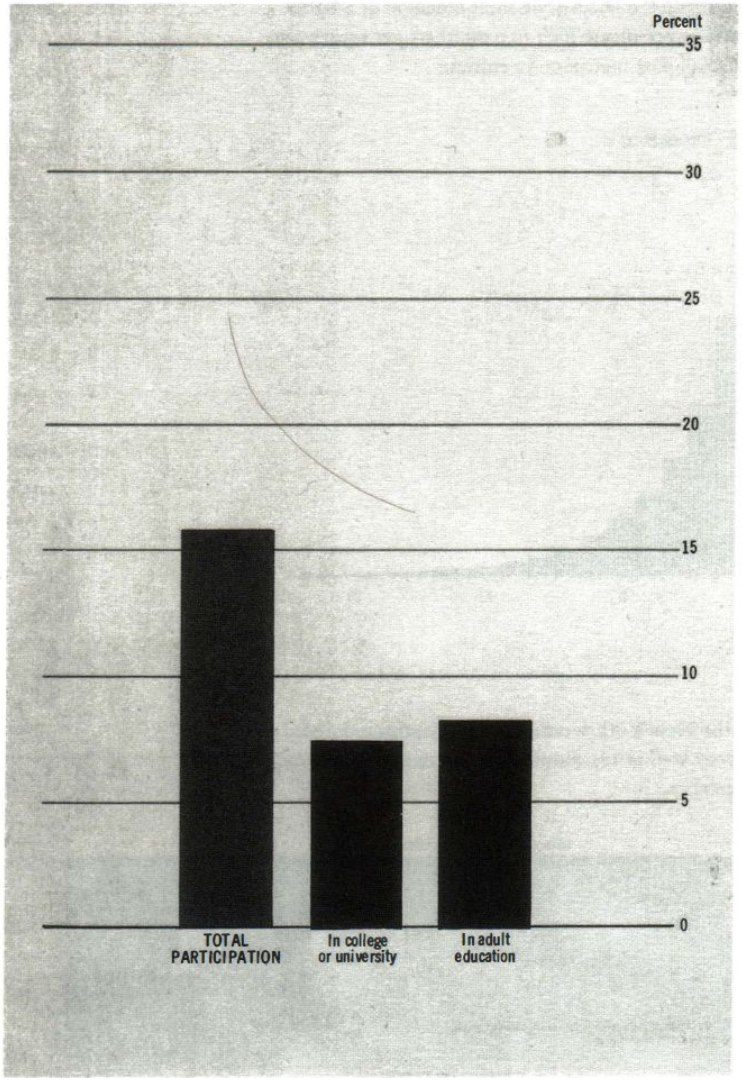
Increase Data Density

$$\text{Data density} = \frac{\text{Number data items}}{\text{Area of data in graphic}}$$



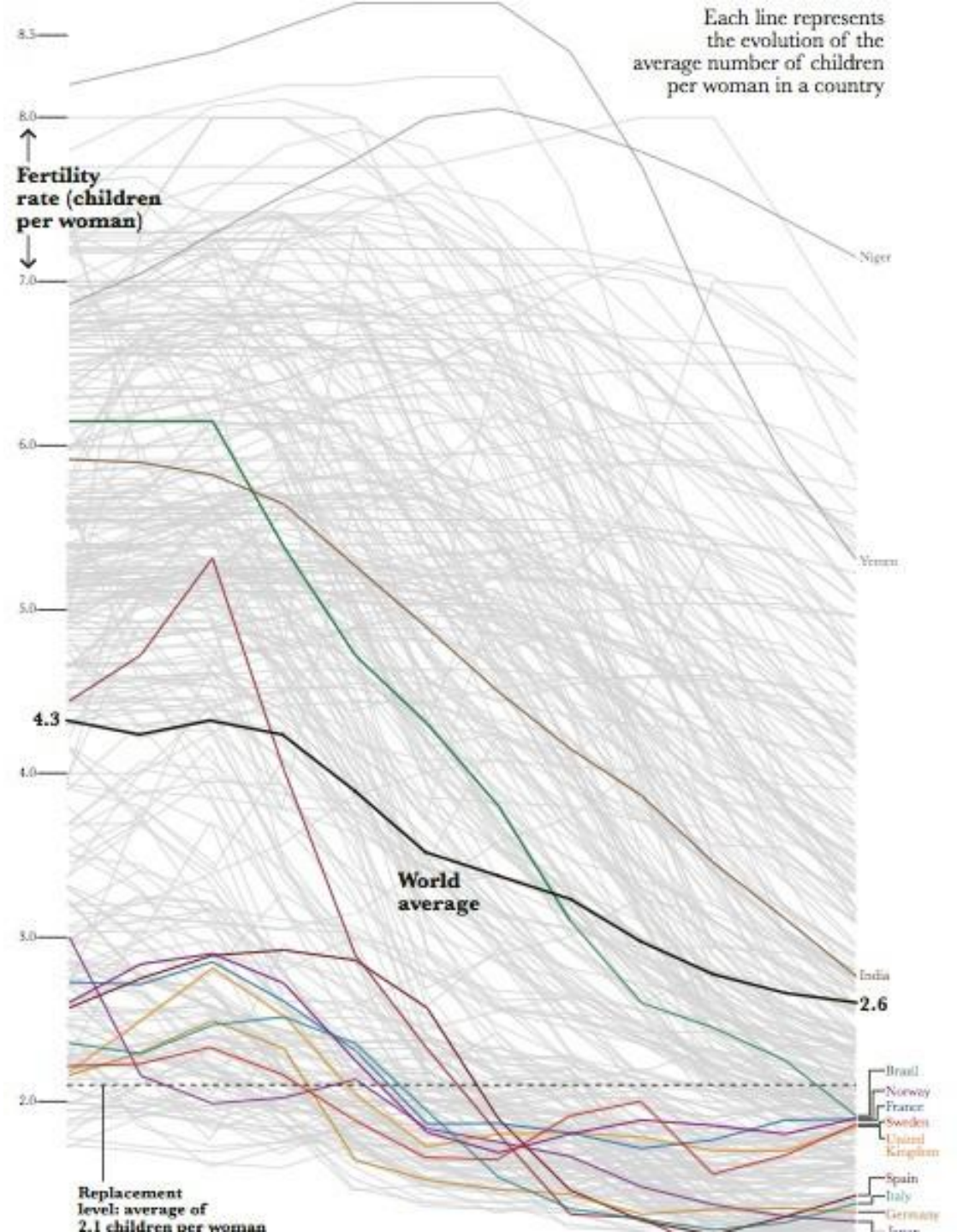
Ho et al., "Thermal Conductivity of the Elements: A Comprehensive Review" J. Phys. Chem. 1974

Increase Data Density



$$\text{ta density} = \frac{\text{Number data items}}{\text{Area of data in graphic}}$$

Executive Office of the President, Office of Management and Budget, Social Indicators, 1973 (Washington, D.C., 1973). p. 86.



A. Cairo,
The Functional

Tufte's Design Principles

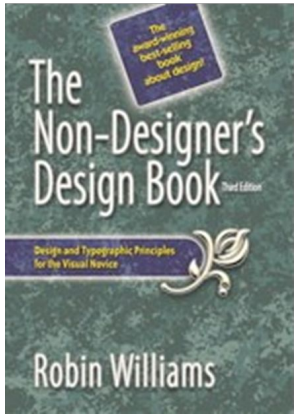
- Maximize data-ink ratio
- Avoid chart junk
- Increase data density
-

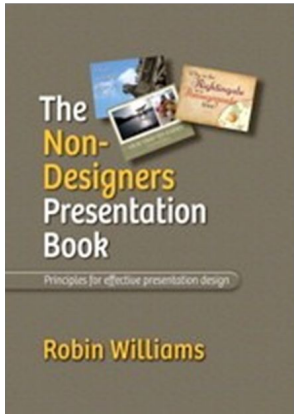
Subjective Dimensions

- **Aesthetics**: Attractive things are perceived as more useful than unattractive ones
- **Style**: Communicates brand, process, who the designer is
- **Playfulness**: Encourages experimentation and exploration
- **Vividness**: Can make a visualization more memorable

Graphic Design Principles

Robin Williams





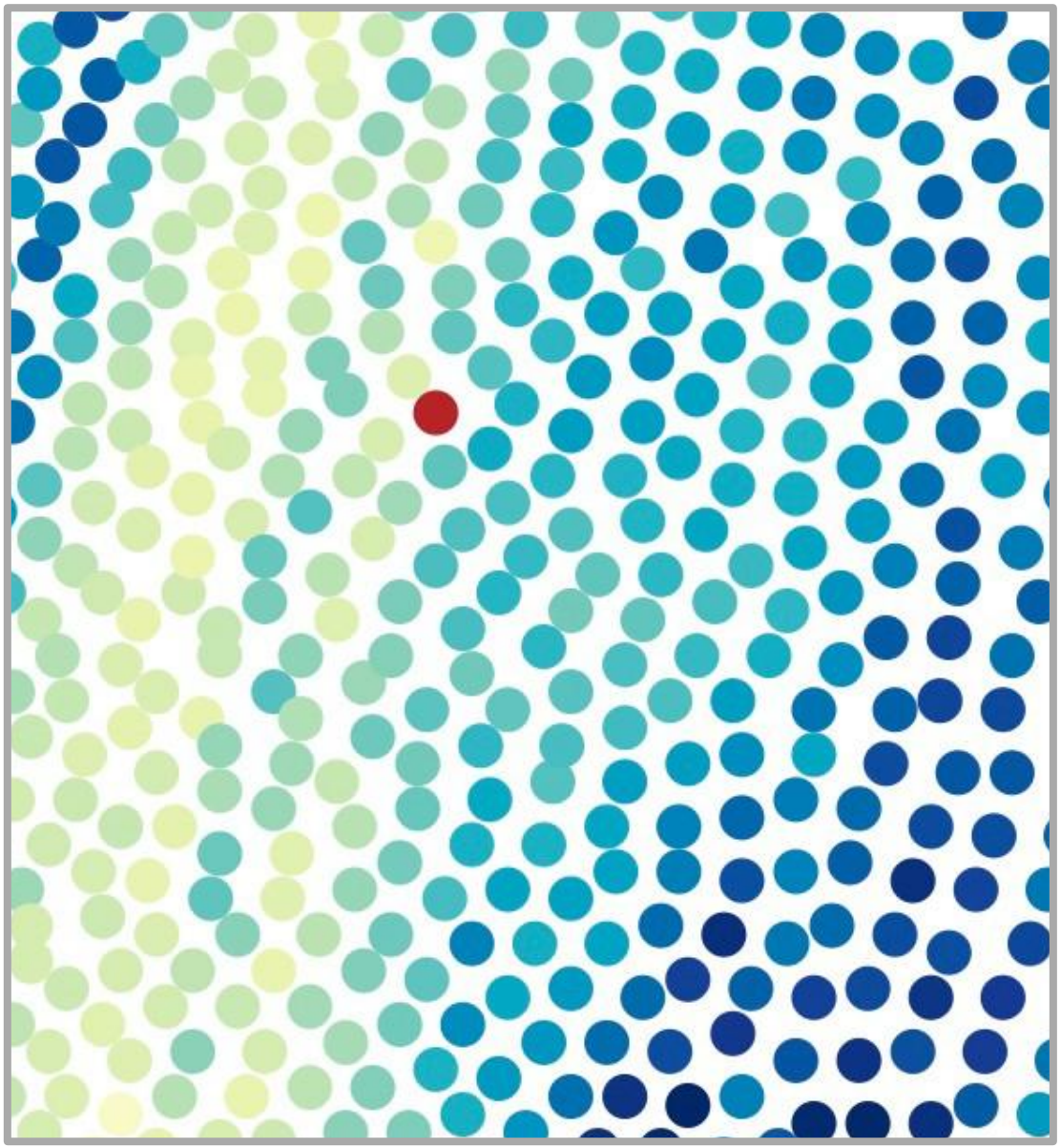
Contrast
Repetition
Alignment
Proximity

Robin Williams

Principle of Contrast

If twoitems are not exactly the same, then make them different. Really different.

Don't be a wimp.



M. Meyer

Unemployment rates by region (in October)

Percentage change compared to previous month

Average: +0.82

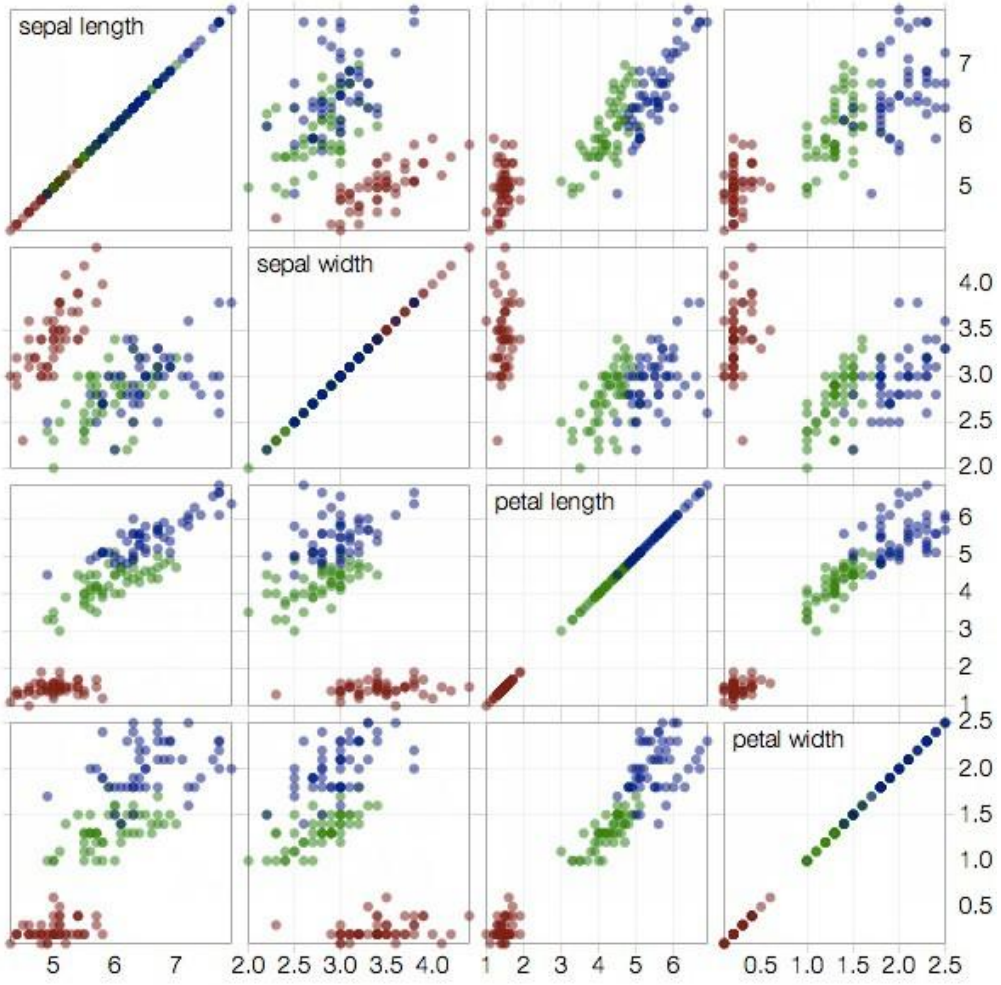
Canarias	+3.42
Aragón	+2.48
Madrid	+2.33
C. Valenciana	+2.08
Melilla	+1.93
Ceuta	+1.81
Murcia	+1.78
C.-La Mancha	+1.78
Cataluña	+1.39
La Rioja	+1.02
País Vasco	+0.84
C. y León	+0.77
Cantabria	+0.54
Navarra	+0.39
Andalucía	-0.30
Galicia	-0.39
Asturias	-0.82
Extremadura	-1.86
Baleares	-4.27

A. Cairo,
The Functional

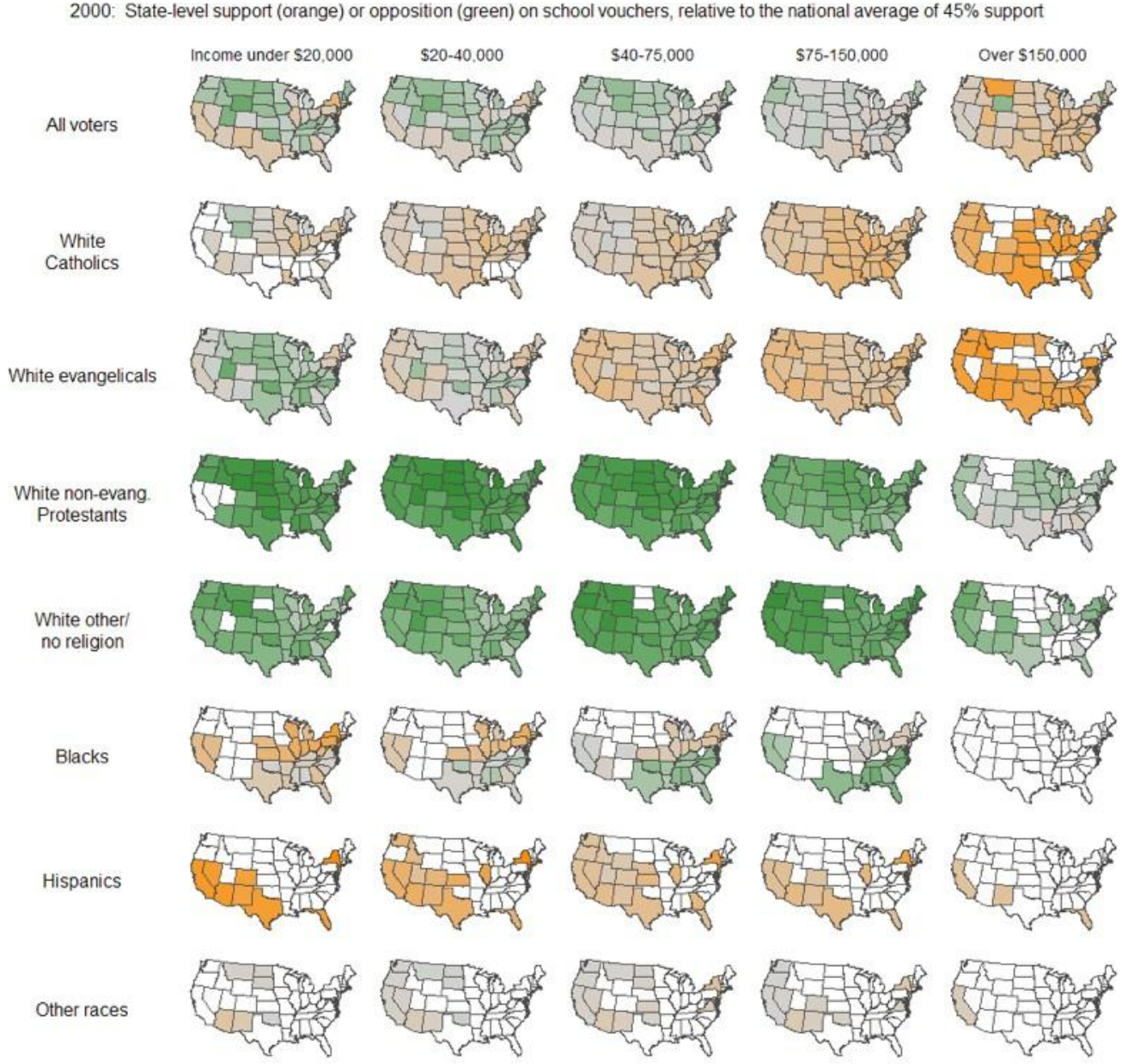
Principle of Repetition

Repeat some aspects of the design throughout the entire piece

Small Multiples



D3, M. Bostock


2000: State-level support (orange) or opposition (green) on school vouchers, relative to the national average of 45% support
Income under $20,000
$20-40,000
$40-75,000
$75-150,000
Over $150,000
All voters
White Catholics
White evangelicals
White non-evang. Protestants
White other/ no religion
Blacks
Hispanics
Other races
Orange and green colors correspond to states where support for vouchers was greater or less than the national average.
The seven ethnic/religious cagetories are mutually exclusive. "Evangelicals" includes Mormons as well as born-again Protestants.
Where a category represents less than 1% of the voters of a state, the state is left blank.

A. Gelmann

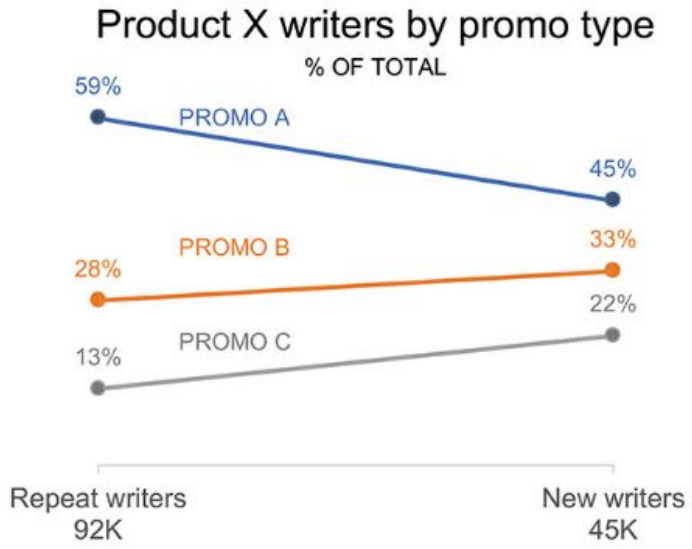
Principle of Alignment

Nothing should be placed on the page arbitrarily

Every item should have a visual connection with something else

There were 45K new writers in the past year.

The distribution across promo types looks different than repeat writers.



Though Promo A makes up the biggest segment overall, they contribute less to new writers than to repeat writers.

Both Promo B and Promo C brought in higher proportion of new writers compared to repeat writers.

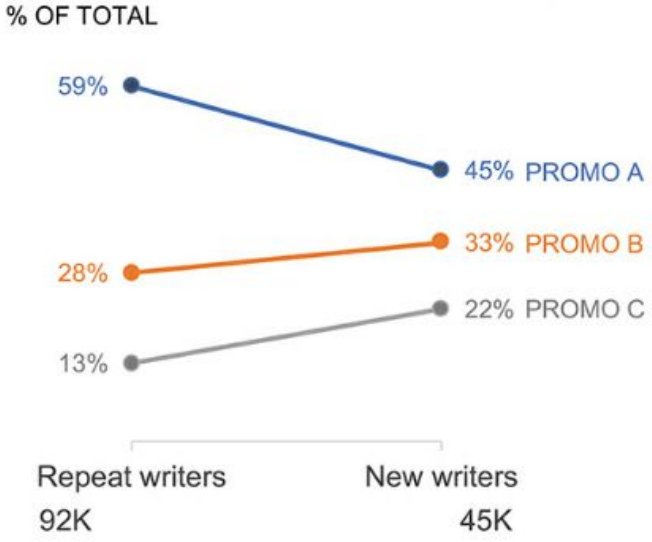
How should we use this data for our future promotion strategy?

What changes would you make when it comes to alignment and white space to improve this visual? Are there other changes you would suggest? Write them down.

There were 45K new writers in the past year.

The distribution across promo types looks different than repeat writers.

Product X writers by promo type



Though Promo A makes up the biggest segment overall, it contributes less to new writers than to repeat writers.

Both Promo B and Promo C brought in higher proportions of new writers compared to repeat writers.

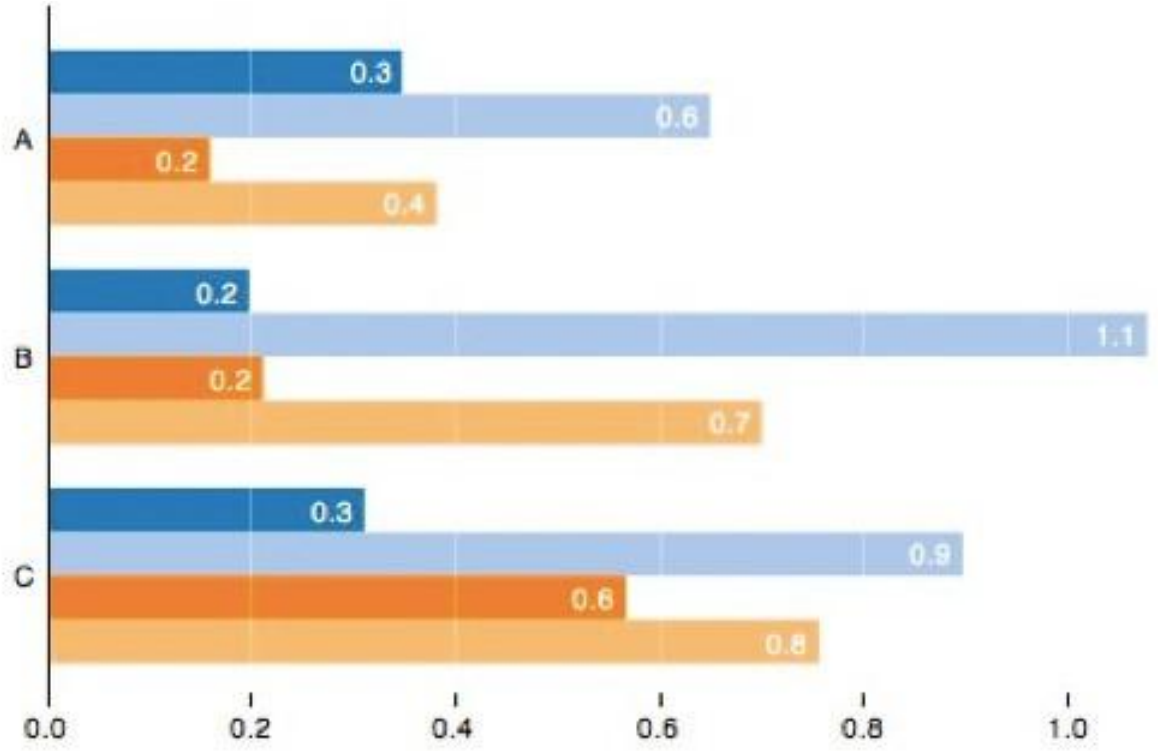
How should we use this data for our future promotion strategy?

Principle of Proximity

Group related items together . . .

as physical closeness implies a relationship

Proximity

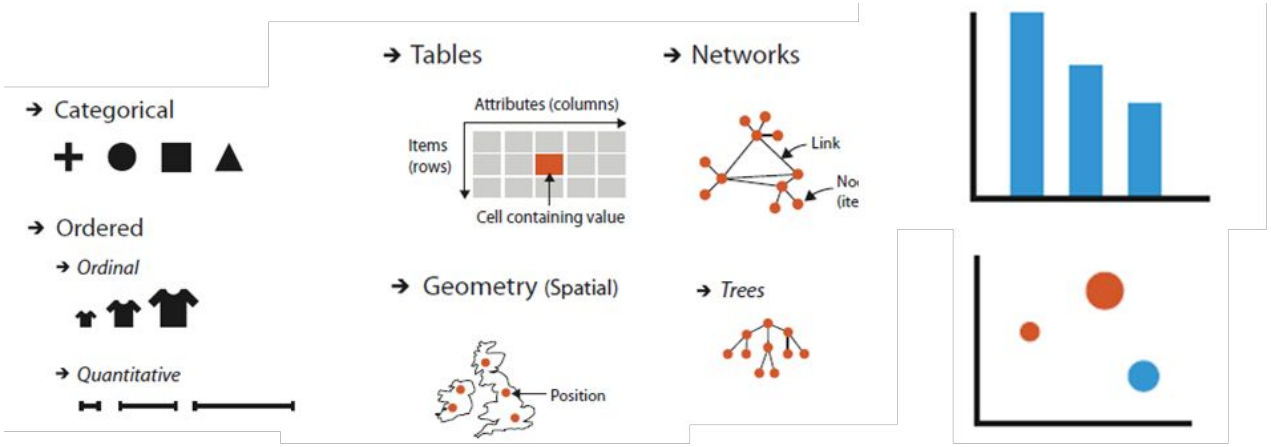


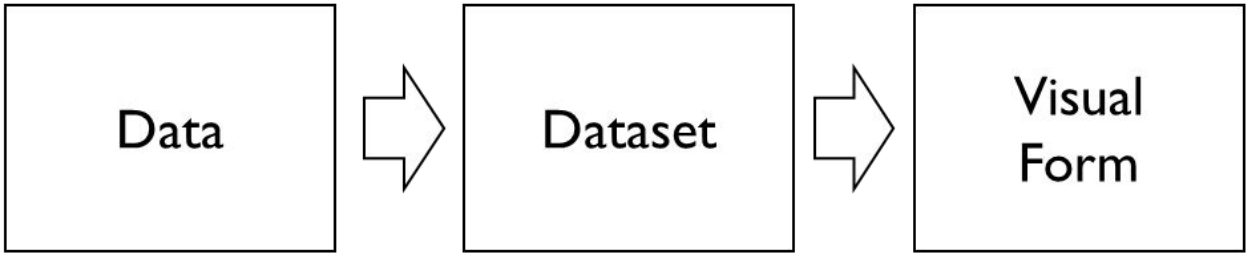
Protovis, M. Bostock

Overview

- Part I
  - Graphical Integrity – “functional art”
  - Visualization Design Principles
    - Maximize data-ink ratio
    - Avoid chart junk
    - Increase data density
  - Graphic Design Principles: CRAP
    - Contrast, Repetition, Alignment, Proximity
- Part II
  - Data Visualization Steps & Visual Encoding
  - Visualization Taxonomy & Statistical Graphs – A Tour through the visualization zoo

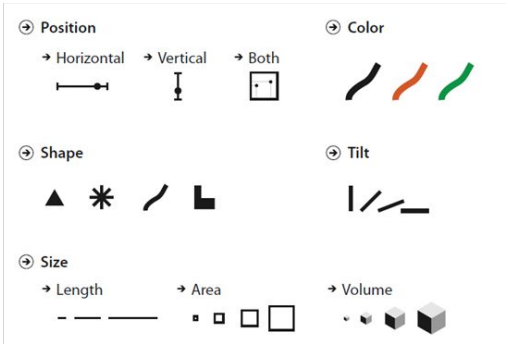
Data Visualization Steps





Data Transformation

Visual Mapping (Encoding)



Data Types

- **N**ominal (categorical)

  Are = or ≠ to other values

  Apples, Oranges, Bananas,...

- **O**rdinal (ordered)

  Obey a < relationship

  Small, medium, large

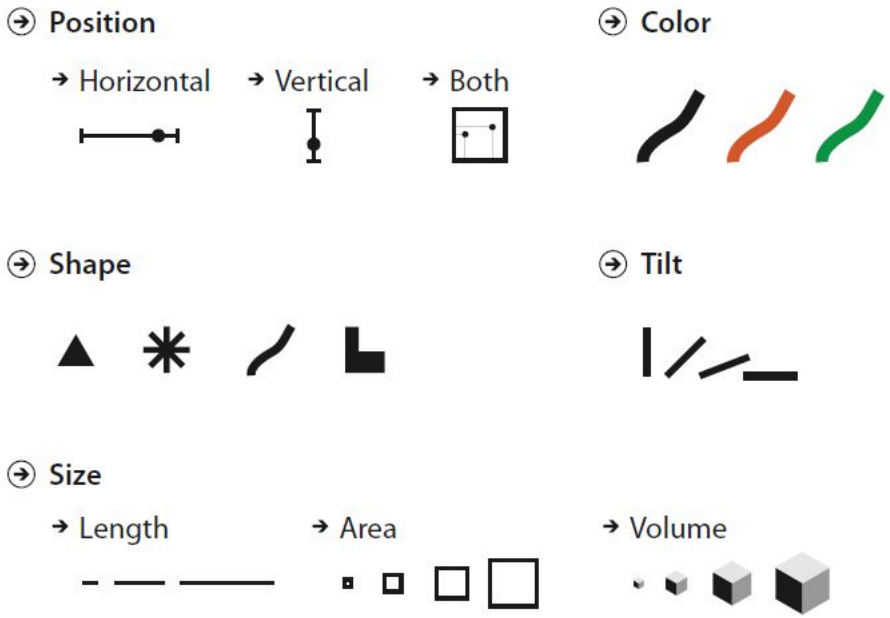
- **Q**uantitative

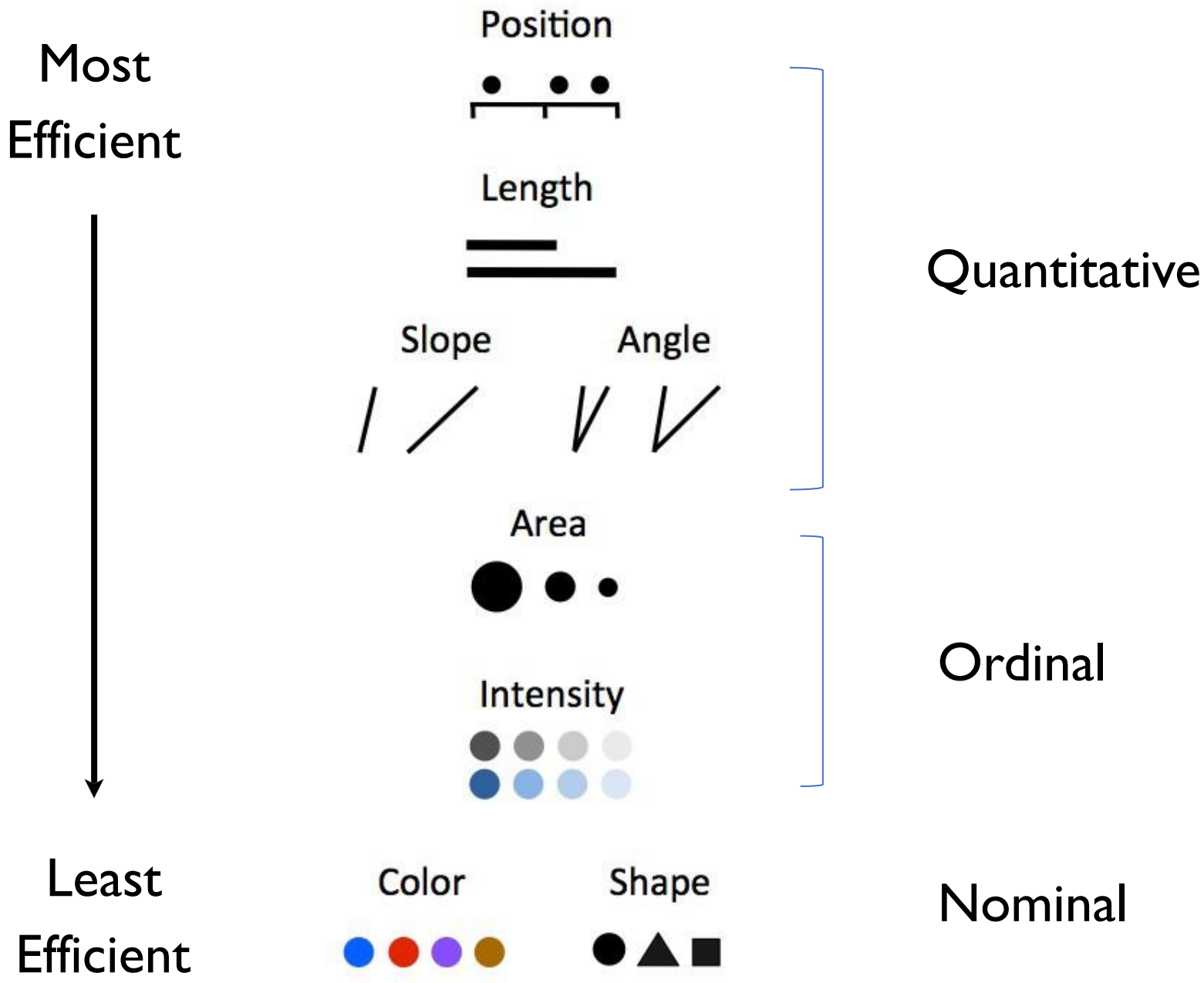
  Can do arithmetic on them

Visual Mapping (Encoding)

Marks: geometric primitives

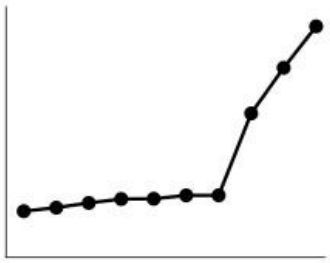
Visual variables (channels) control the appearance of marks

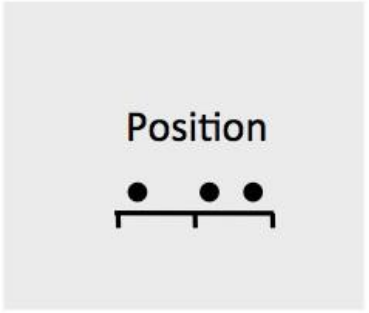


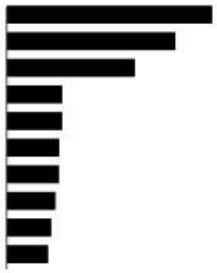


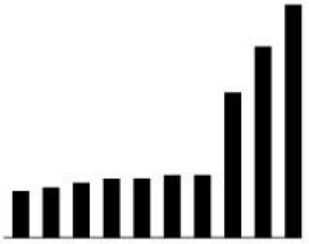
C. Mulbrandon
VisualizingEconomics.com

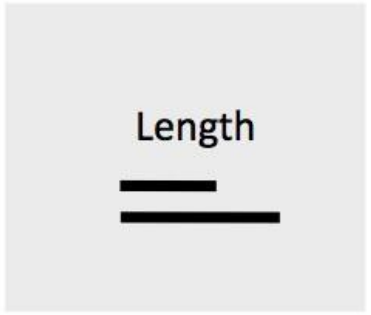
Most Efficient





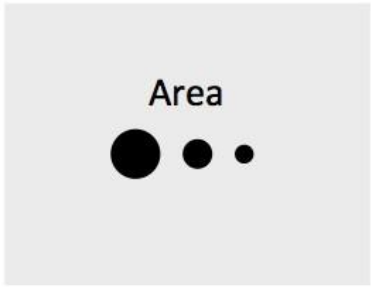






C. Mulbrandon
VisualizingEconomics.com

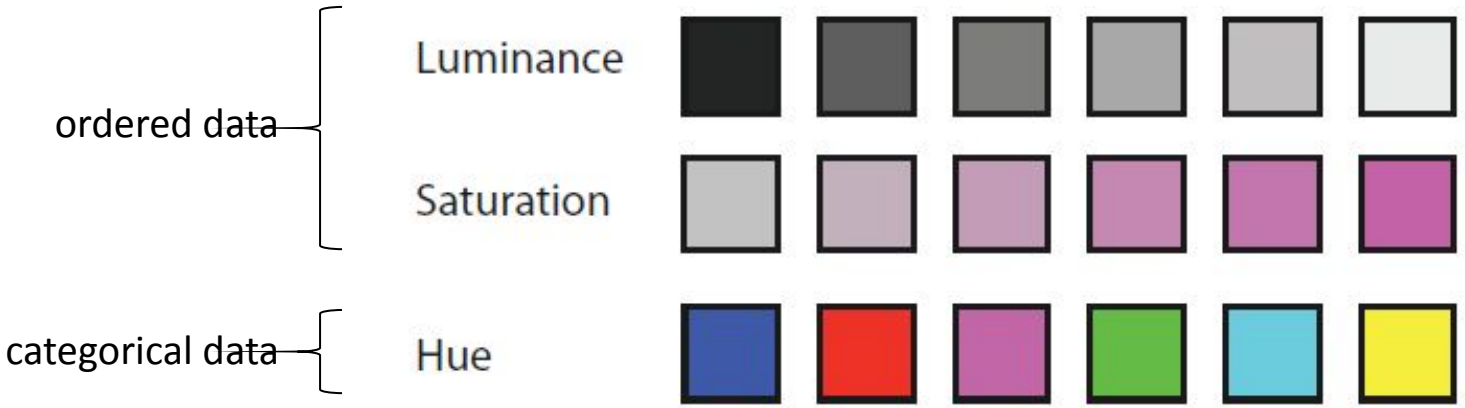
Pie & Donut Charts



C. Mulbrandon
VisualizingEconomics.com

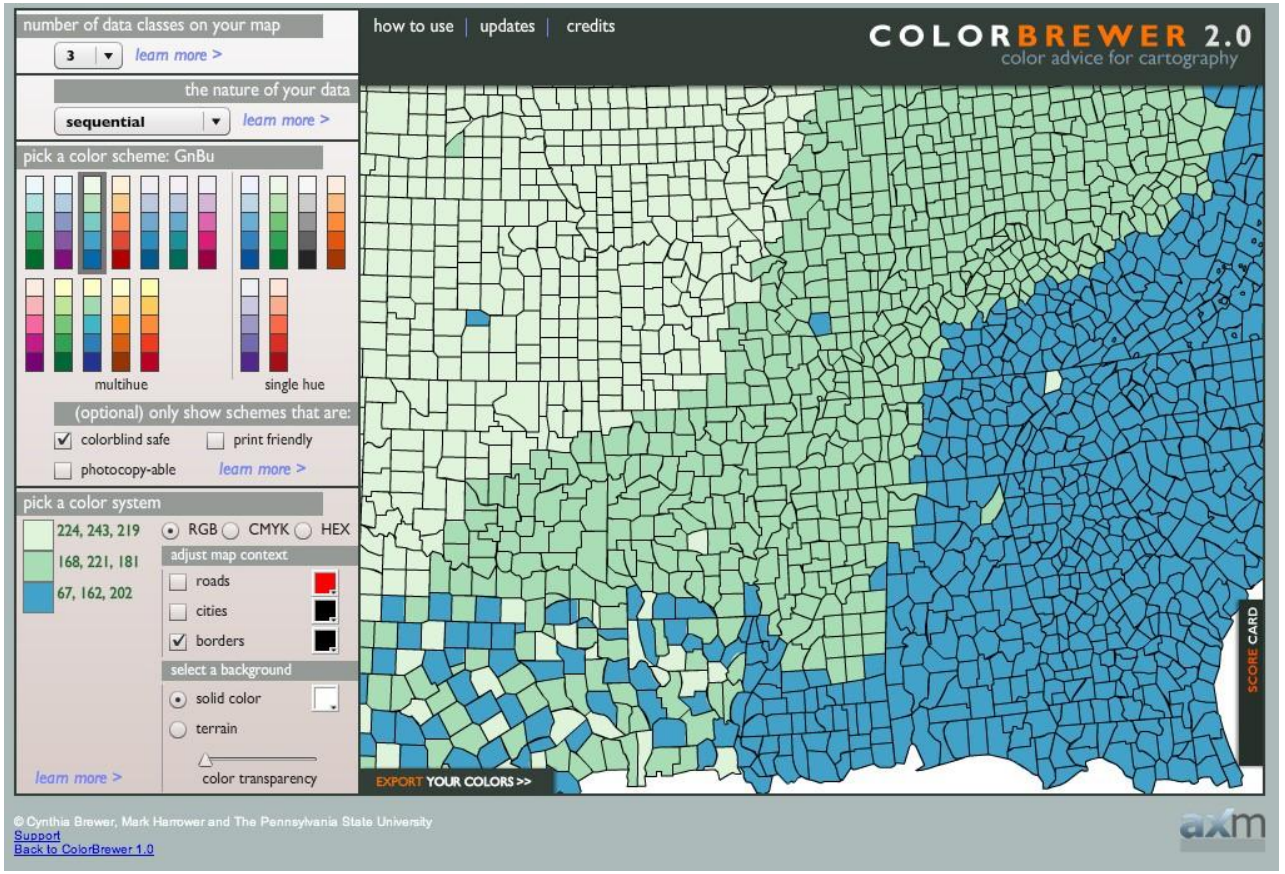
Color

- luminance/saturation for ordered data types
  - But limited accuracies (<5 steps for luminance, <3 steps for saturation)
- hue: effective for categorical data and showing groupings



Sanity check: make sure that it also looks good in gray scale

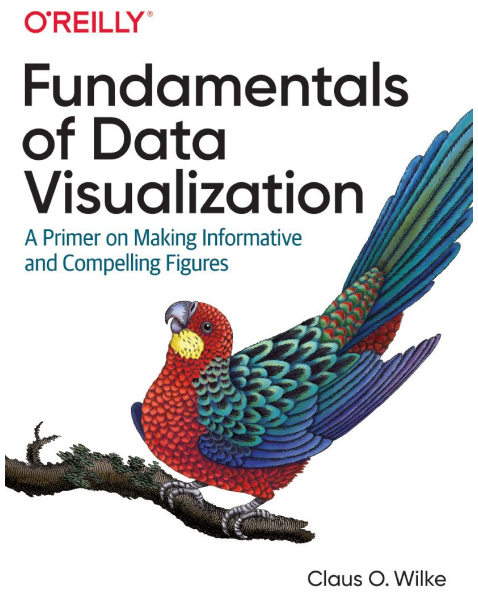
Color - ColorBrewer



http://www.colorbrewer2.org

Data Visualization

Visualization Taxonomy & Statistical Graphs



Axis & Legends & Captions

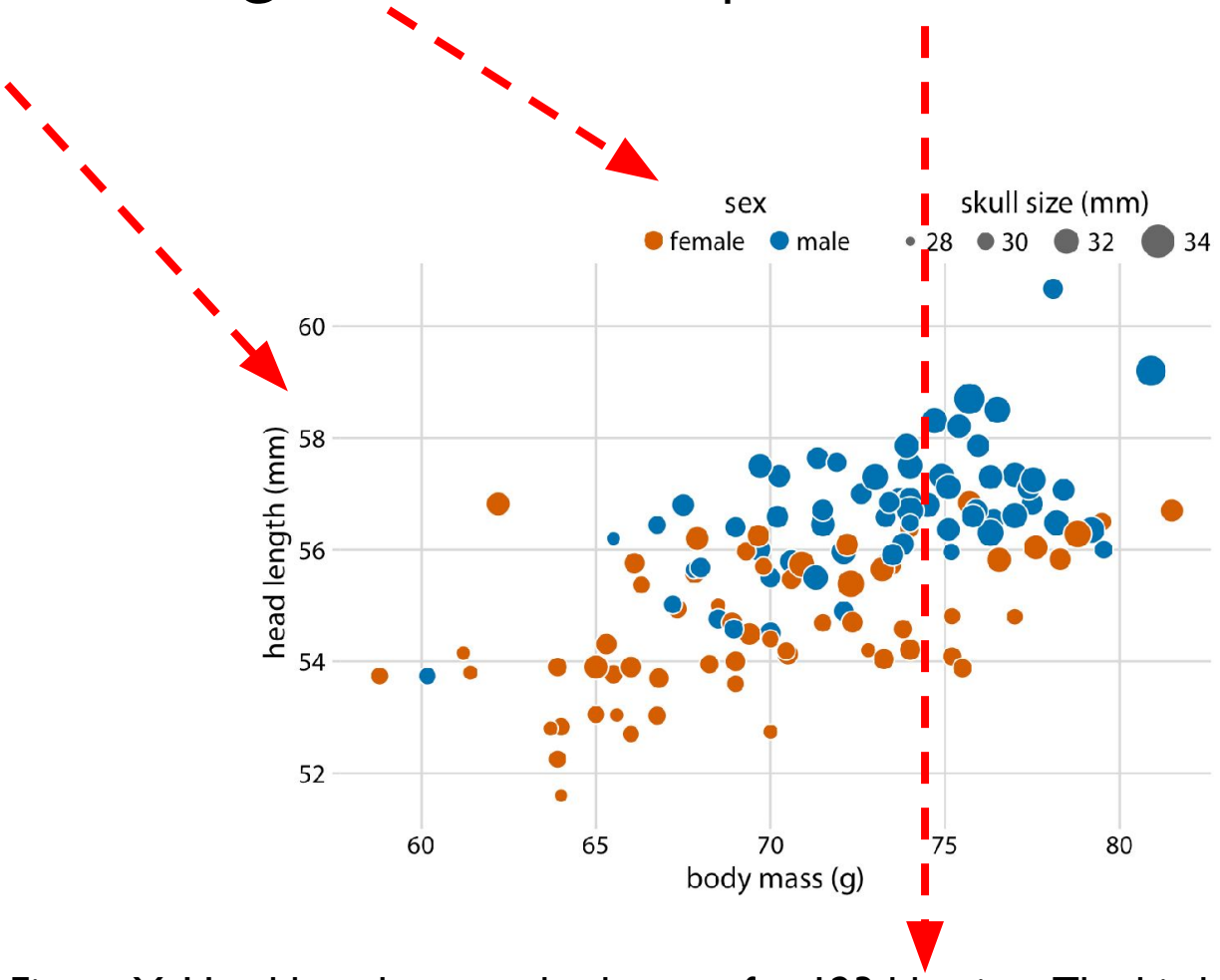


Figure X: Head length versus body mass for 123 blue jays. The birds' sex is indicated by color, and the birds' skull size by symbol size. Head length measurements include the length of the bill while skull size measurements do not.

Tables

a

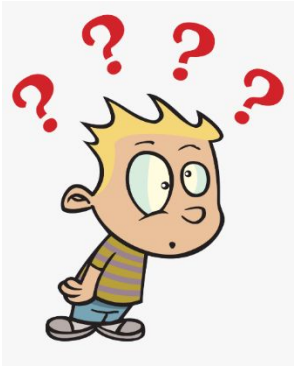
ugly

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

b

ugly

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746



Tables

a

ugly

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

b

ugly

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

c

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

d

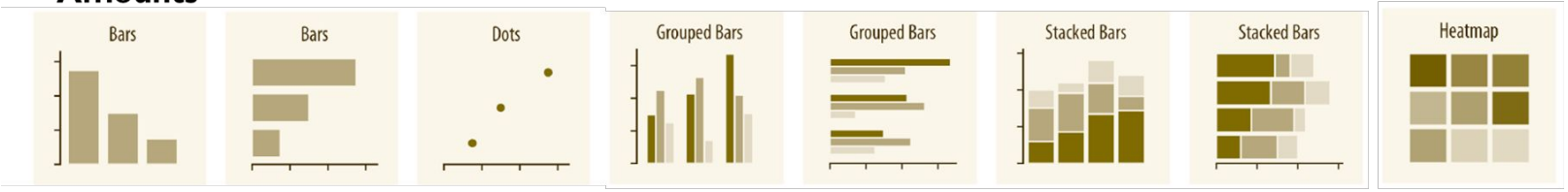
Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

1. Do not use vertical lines.
2. Do not use horizontal lines between data rows. (Horizontal lines as a separator between the title row and the first data row or as a frame for the entire table are fine.)
3. Text columns should be left aligned.
4. Number columns should be right aligned and should use the same number of decimal digits throughout.
5. Columns containing single characters should be centered.
6. The header fields should be aligned with their data; i.e., the heading for a text column will be left aligned and the heading for a number column will be right aligned.

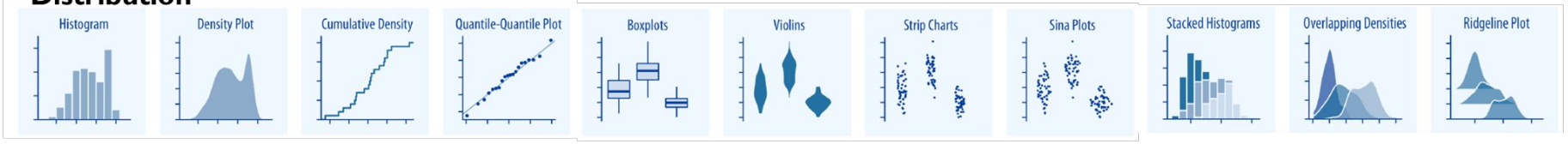
A Tour Through the Visualization Zoo

- Time-Series Data
  - Index Charts
  - Stacked Graphs
  - Small Multiples
  - Horizon Graphs
- Statistical Distributions
  - Histograms & Box Plots
  - Density Plots
  - Stem-and-Leaf Plots
  - Q-Q Plots
  - SPLOM (Scatter Plot Matrix)
  - Parallel Coordinates
- Maps
  - Choropleth Maps
  - Cartograms
  - Flow Maps
  - Graduated Symbol Maps
- Hierarchies
  - Node-link diagrams
  - Adjacency Diagrams
  - Enclosure Diagrams
- Networks
  - Force-directed Layouts
  - Arc Diagrams
  - Matrix Views (Adjacency Diagrams)

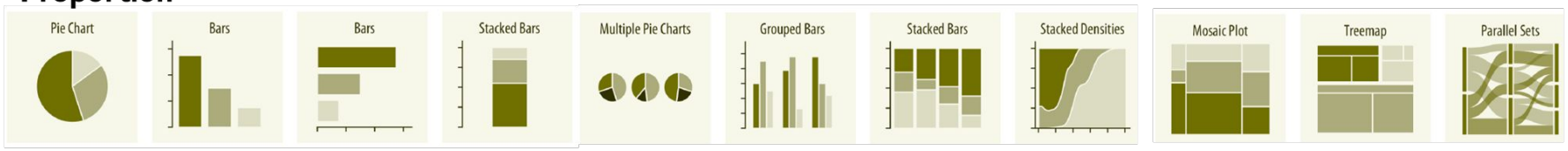
Amounts



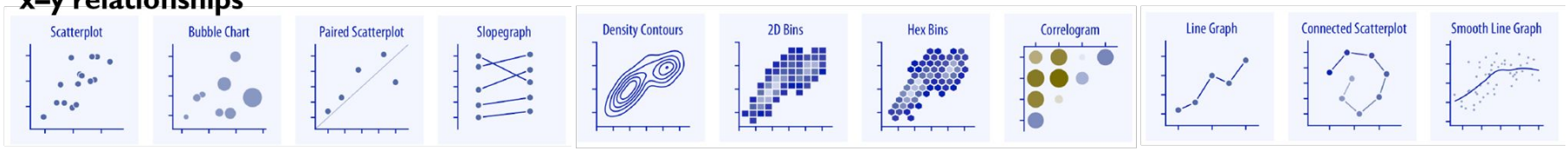
Distribution



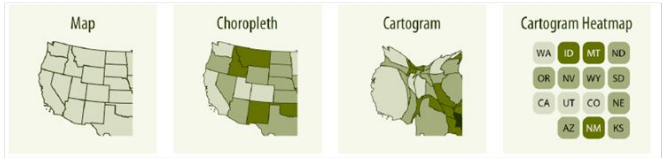
Proportion



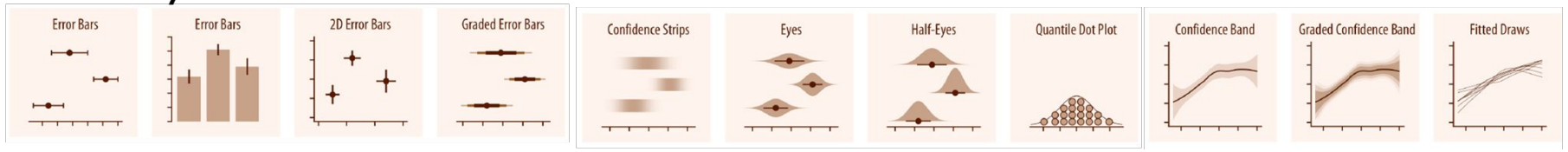
x–y relationships



Geospatial Data

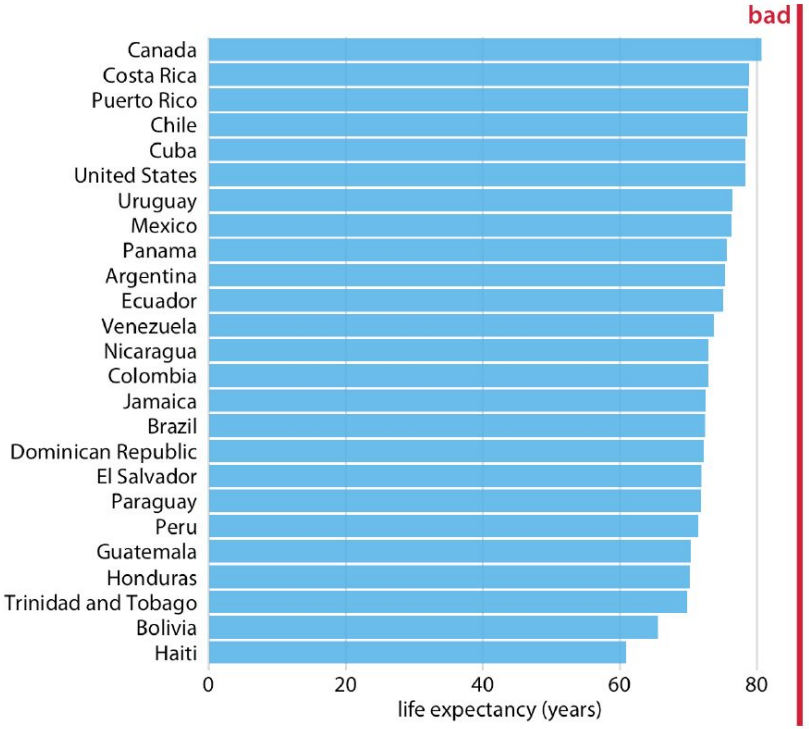


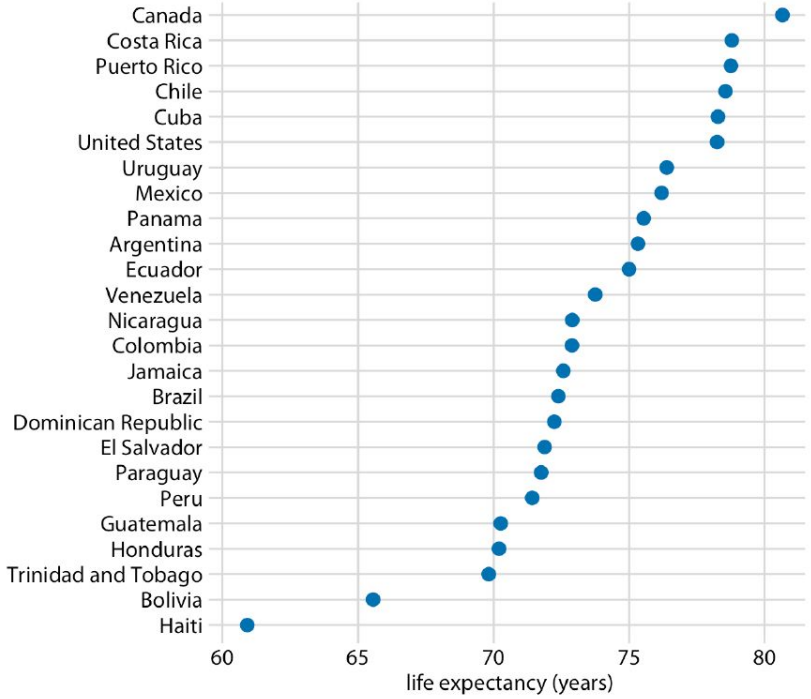
Uncertainty



Amounts

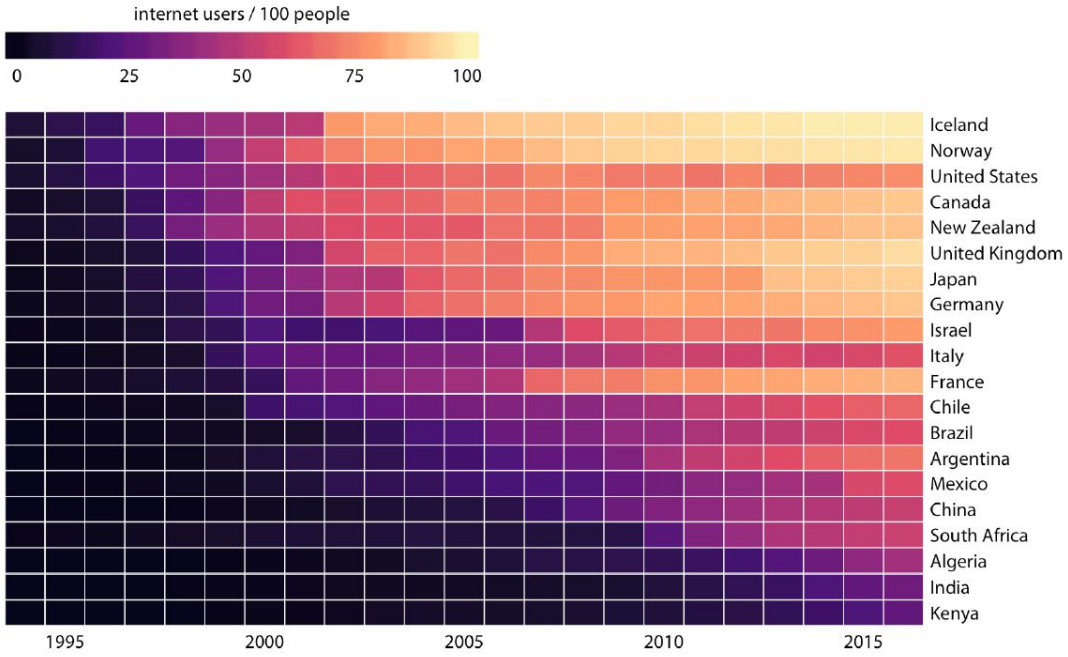
- Dot Plots and Heatmaps





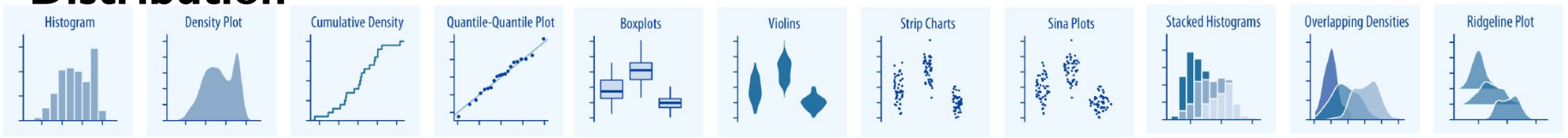
Amounts

- Dot Plots and Heatmaps

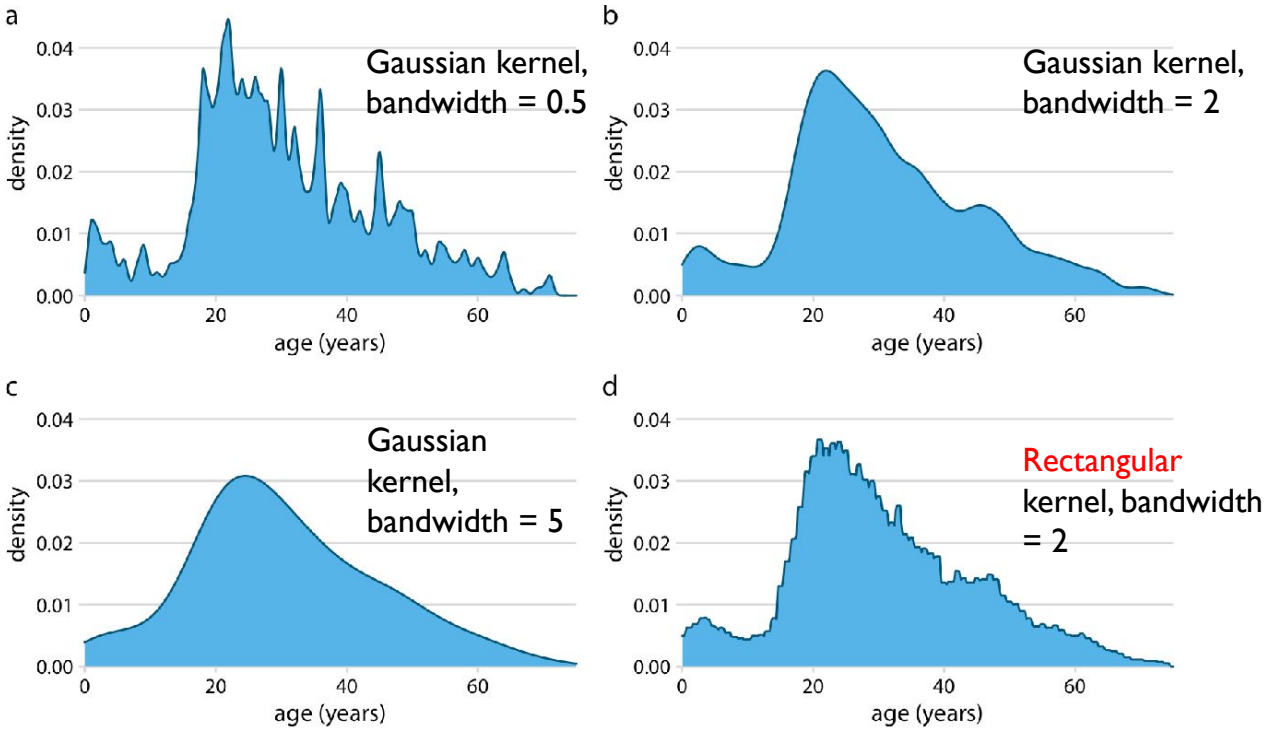


Internet adoption over time, for select countries. Countries were ordered by the year in which their internet usage first exceeded 20%.

Distribution

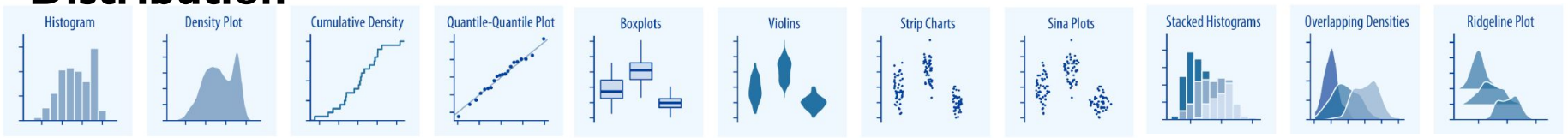


- Visualizing a Single Distribution

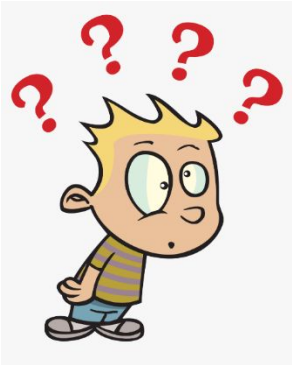


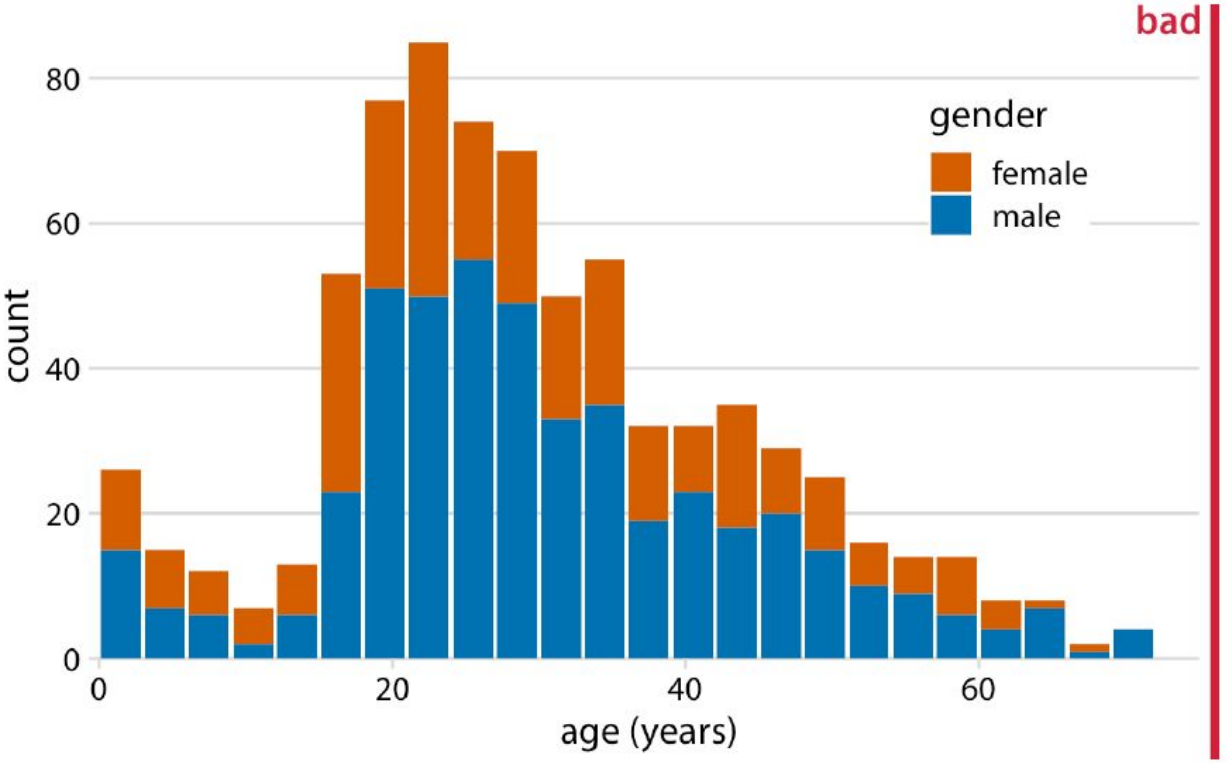
Kernel density estimates depend on the chosen kernel and bandwidth. Here, the same age distribution of Titanic passengers is shown for four different combinations of these parameters: Kernel function + bandwidth (smoothing factor)

Distribution



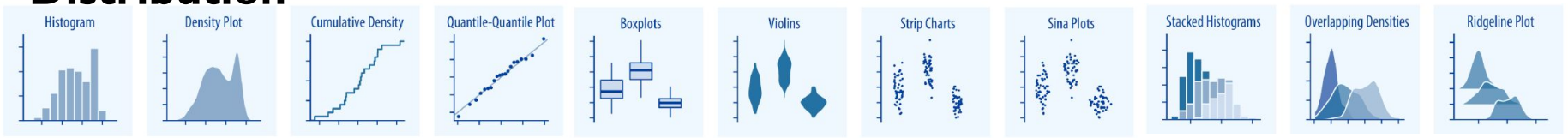
- Visualizing Multiple Distributions at the Same Time



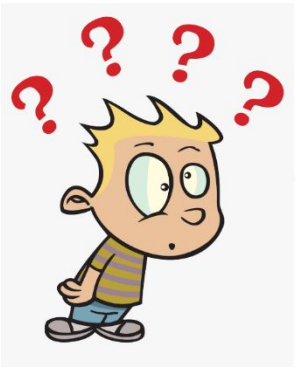


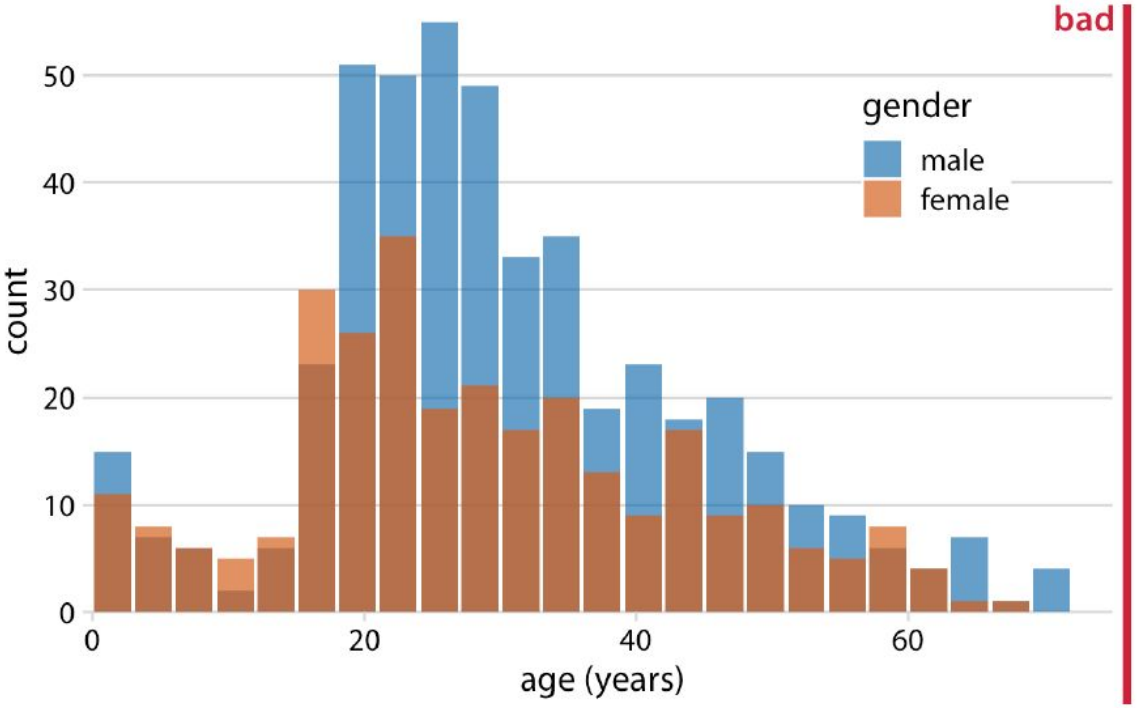
Histogram of the ages of Titanic passengers stratified by gender, shows as overlapping histograms (it's not stacked histograms)

Distribution



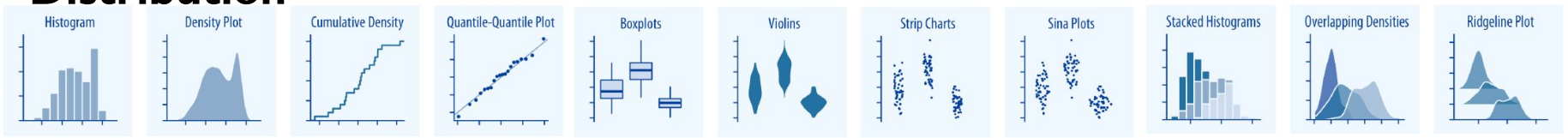
- Visualizing Multiple Distributions at the Same Time



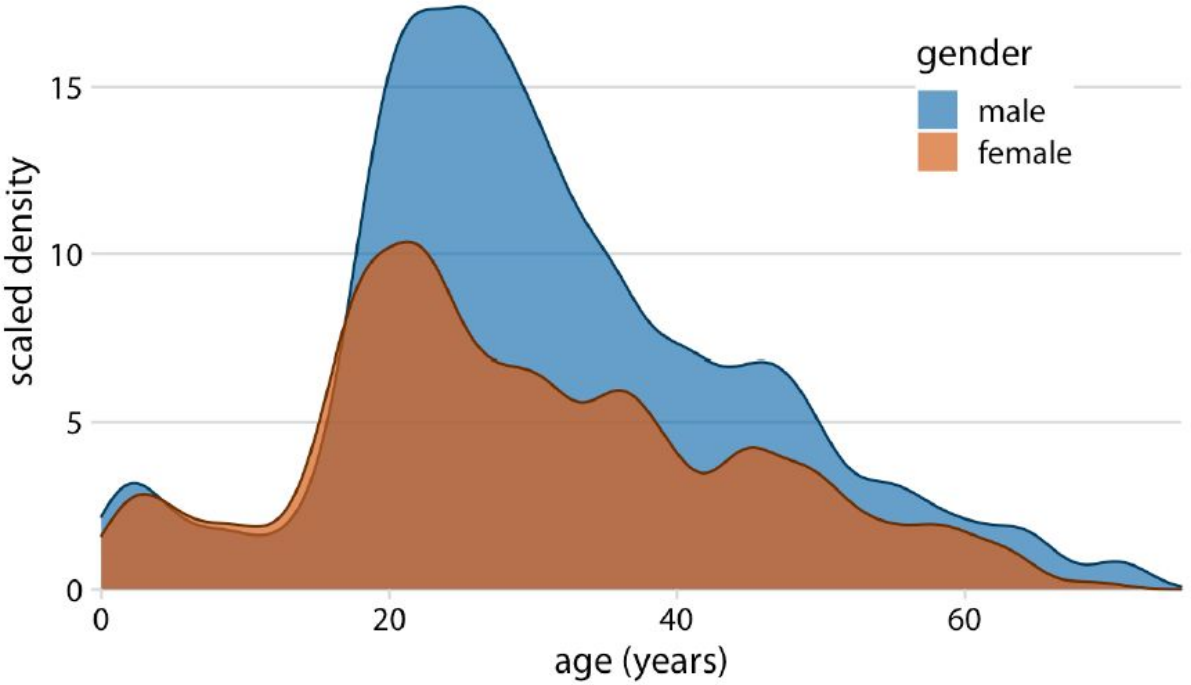


Age distributions of male and female Titanic passengers, shown as two overlapping histograms

Distribution

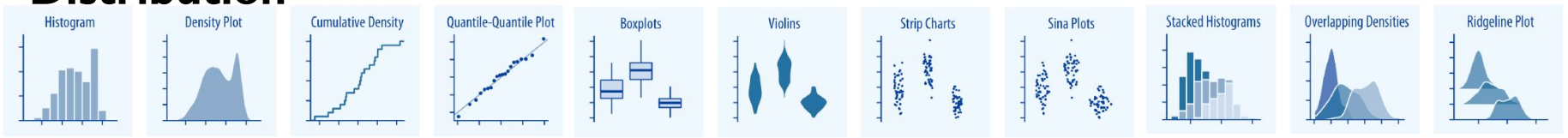


- Visualizing Multiple Distributions at the Same Time

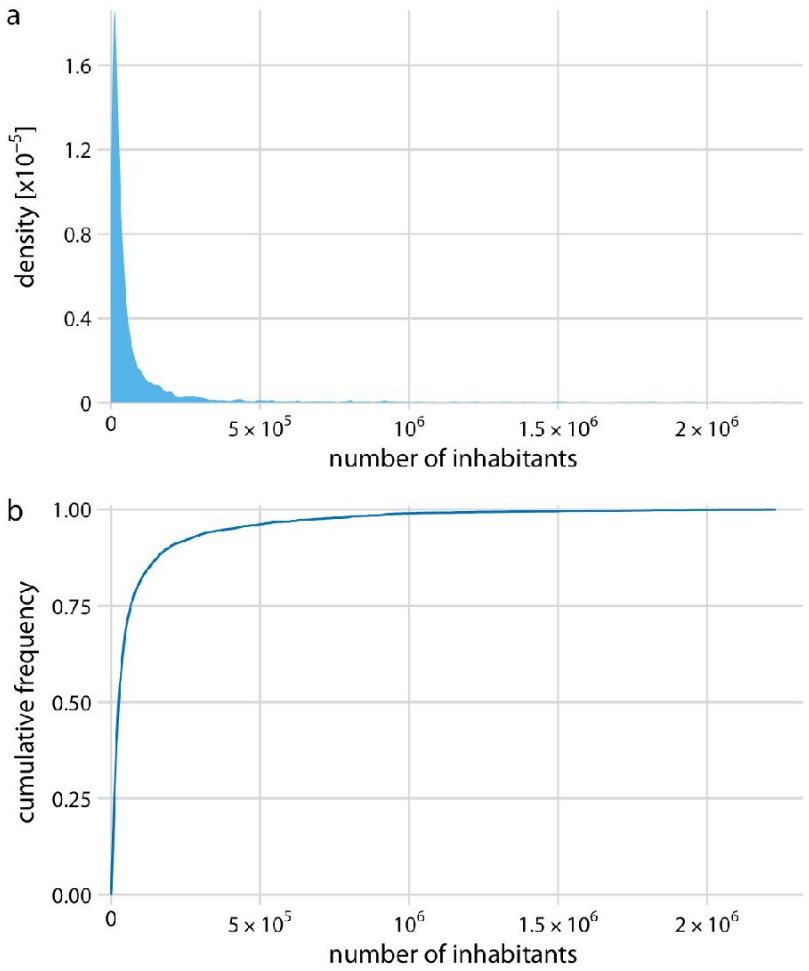


Density estimates of the ages of male and female Titanic passengers. To highlight that there were more male than female passengers, the density curves were scaled such that the area under each curve corresponds to the total number of male and female passengers with known age (468 and 288, respectively).

Distribution

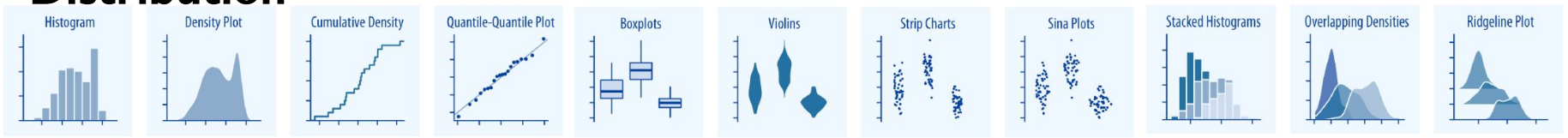


- Highly Skewed Distributions

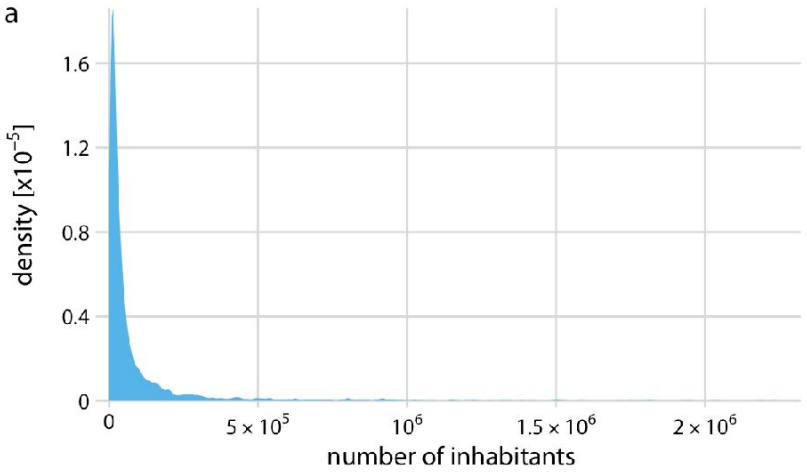


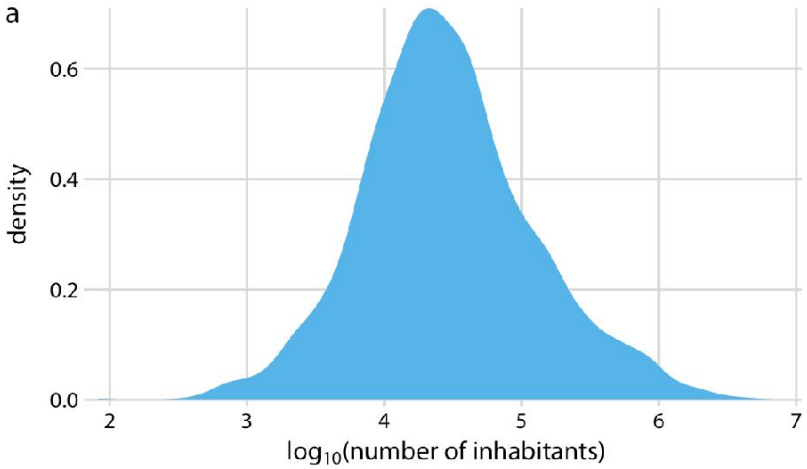
Distribution of the number

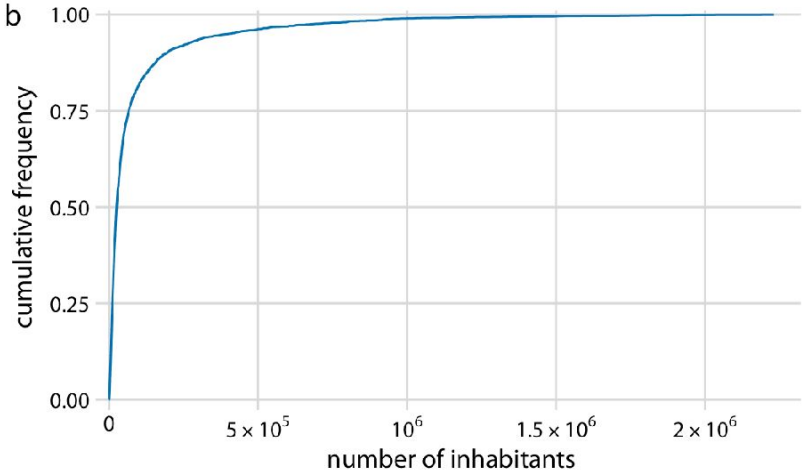
Distribution

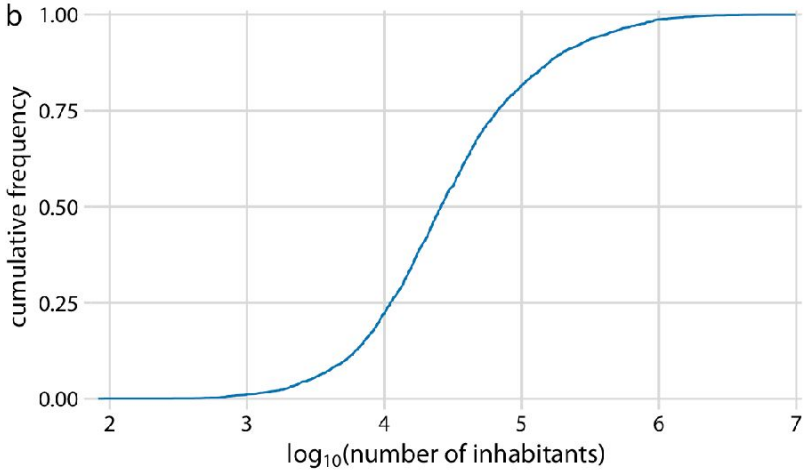


- Highly Skewed Distributions



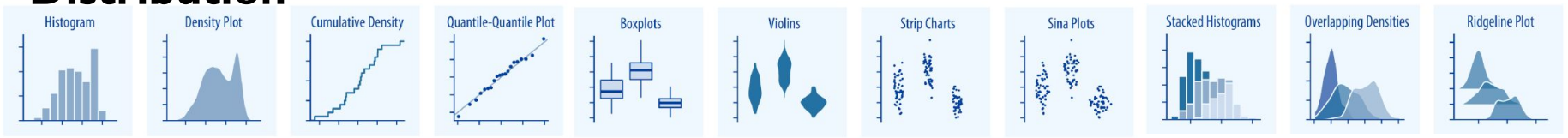




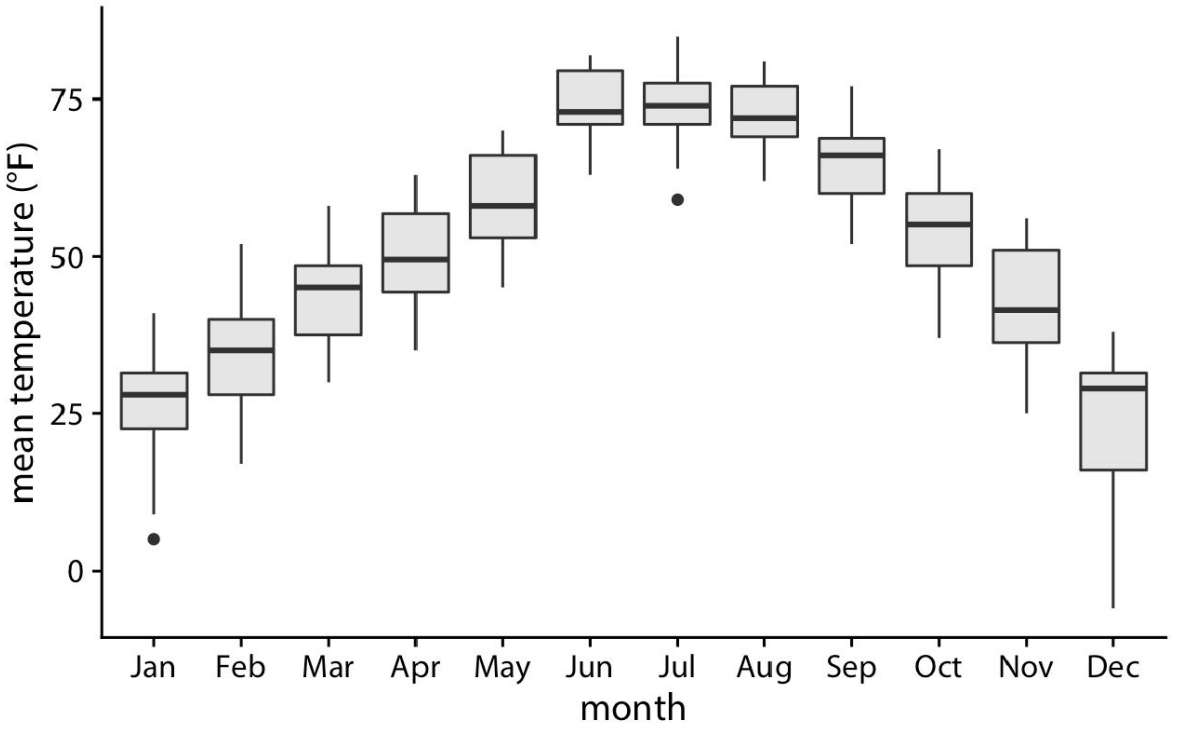


Distribution of the number

Distribution

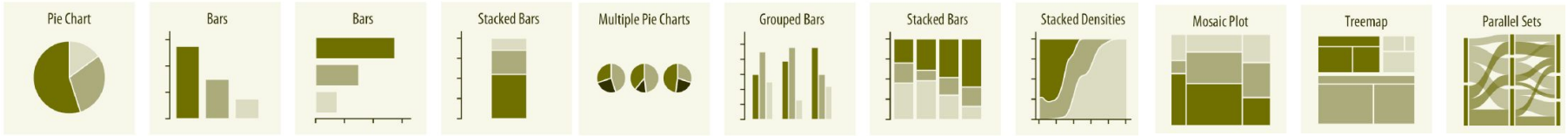


- Visualizing Many Distributions Along the Vertical Axis

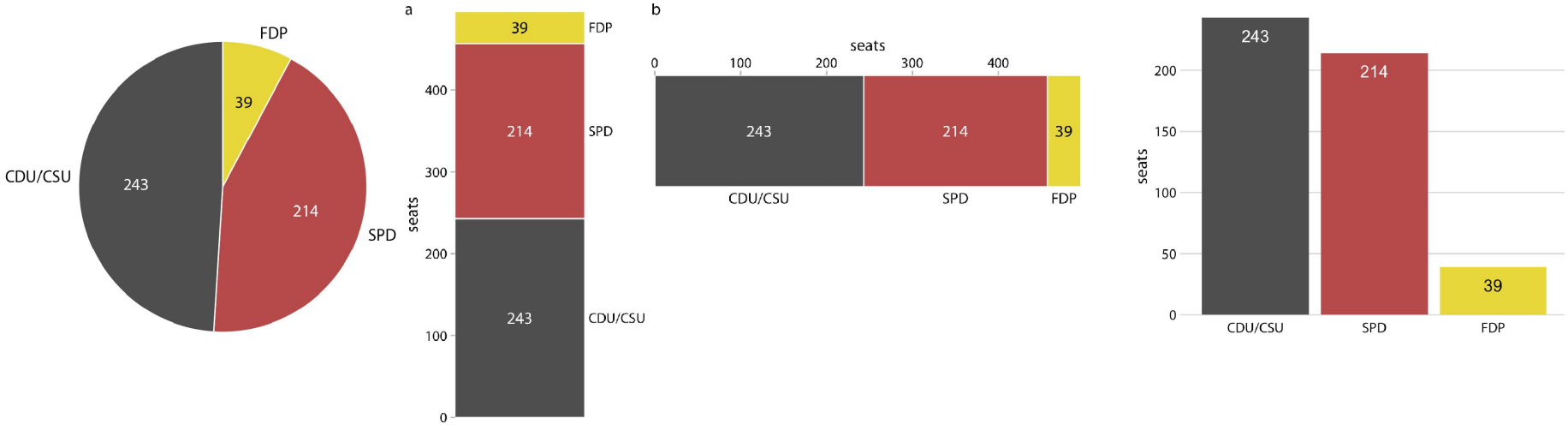


Mean daily temperatures in Lincoln, NE, visualized as boxplots

Proportion



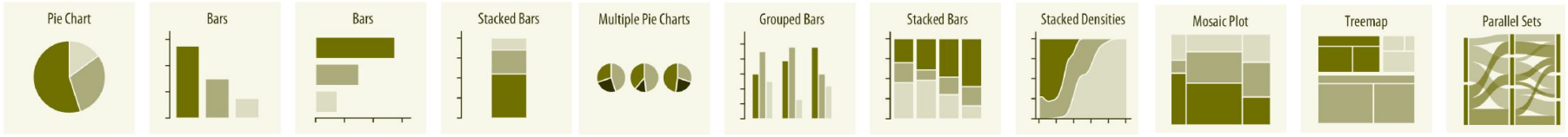
- A Case for Pie Charts



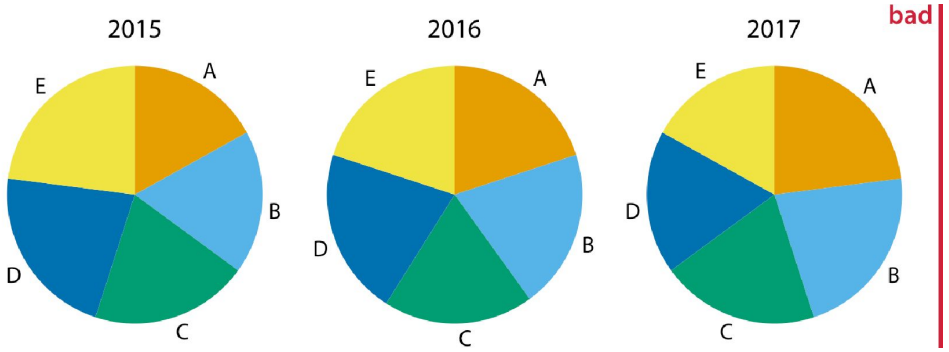
Pie chart Stacked bars Side-by-side bars

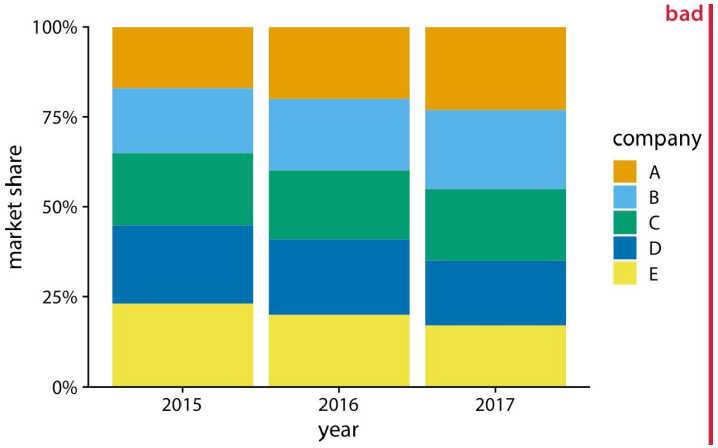
Party composition of the eighth German Bundestag, 1976–1980, visualized as a pie chart. Only the pie chart highlights that the ruling coalition of SPD and FDP had a small majority over the opposition CDU/CSU. Data source: Wikipedia

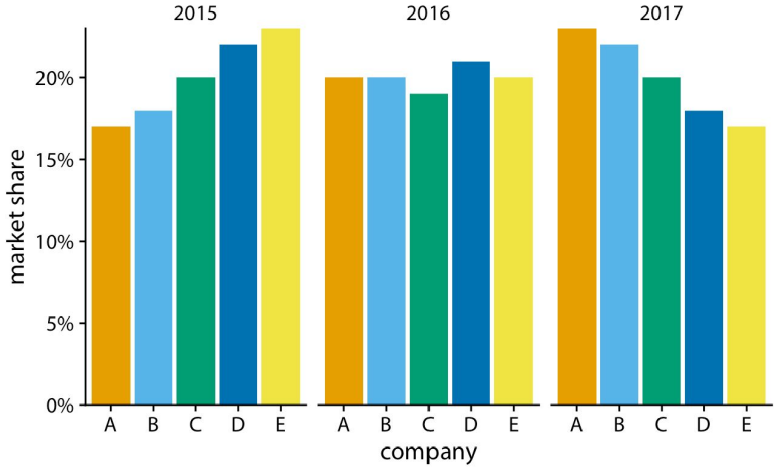
Proportion



- A Case for Side-by-Side Bars

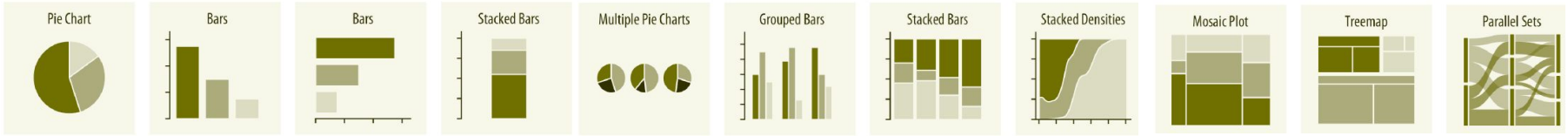




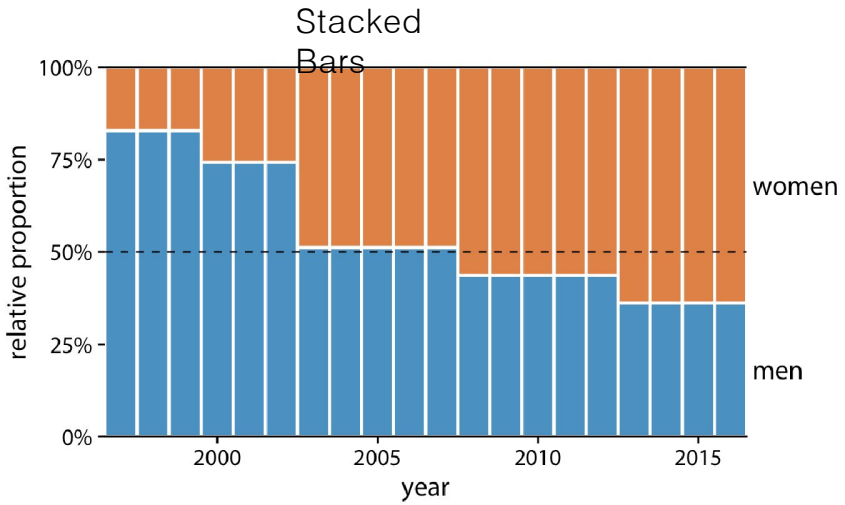


Market share of five hypothetical companies, A–E, for the years 2015–2017

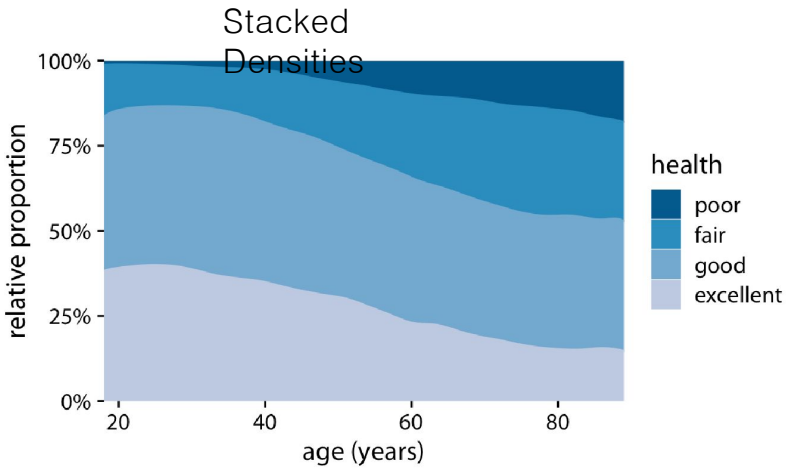
Proportion



- A Case for Stacked Bars and Stacked Densities

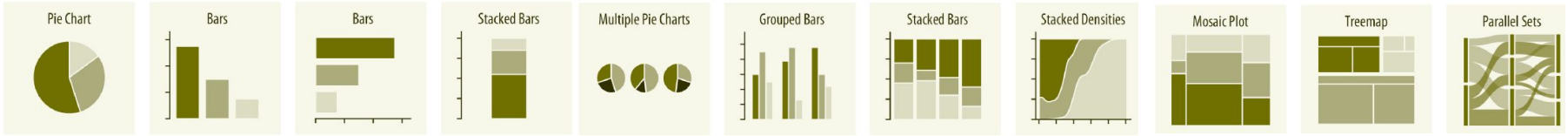


Change in the gender composition of the Rwandan parliament over time, 1997 to 2016.



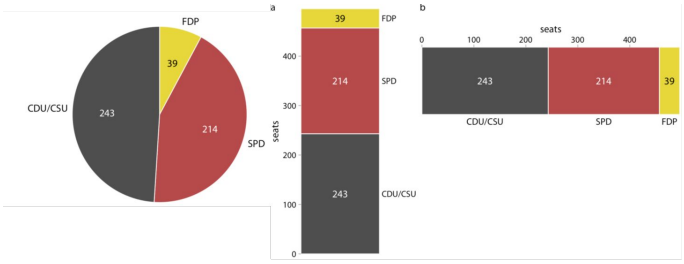
Health status by age

Proportion



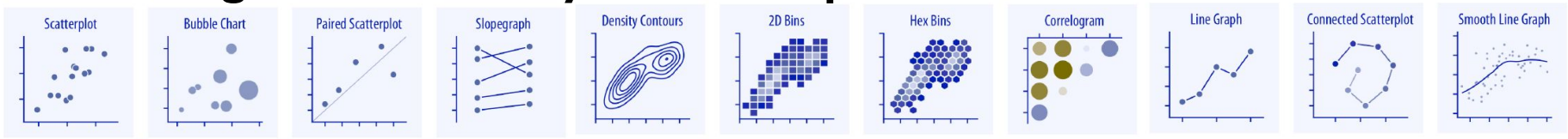
- Pros and cons of common approaches to visualizing proportions: pie charts, stacked bars, and side-by-side bars

	Pie chart	Stacked bars	Side-by-side bars
Clearly visualizes the data as proportions of a whole	✓	✓	✗
Allows easy visual comparison of the relative proportions	✗	✗	✓
Visually emphasizes simple fractions, such as 1/2, 1/3, 1/4	✓	✗	✗
Looks visually appealing even for very small datasets	✓	✗	✓
Works well when the whole is broken into many pieces	✗	✗	✓
Works well for the visualization of many sets of proportions or time series of proportions	✗	✓	✗

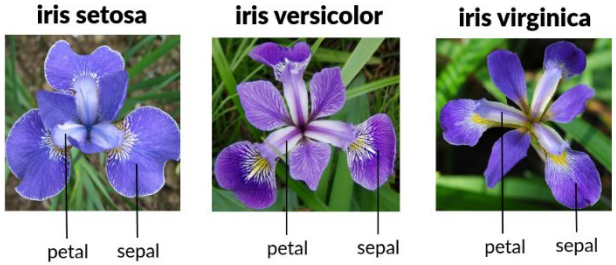


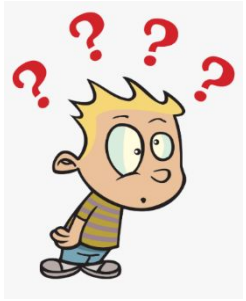
Pie chart Stacked bars Side-by-side bars

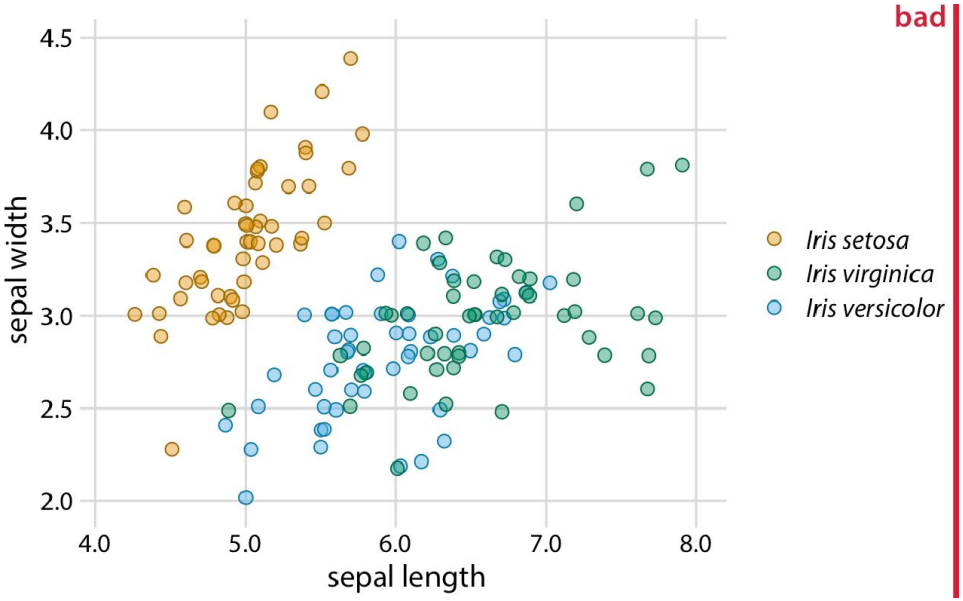
Visualizing association: x–y relationships



- Visualizing Associations: Scatterplots

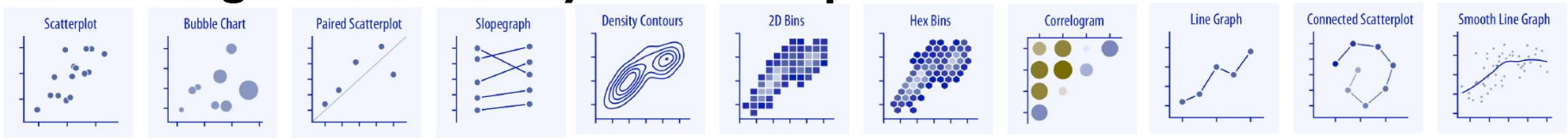




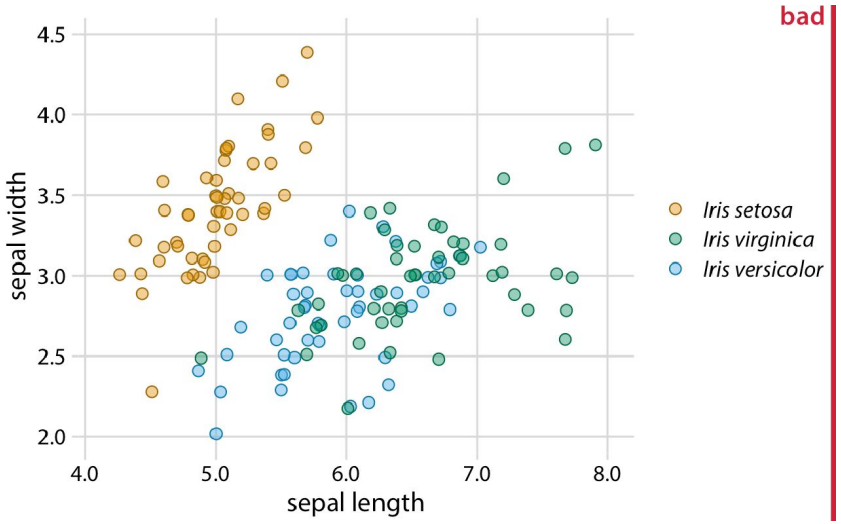


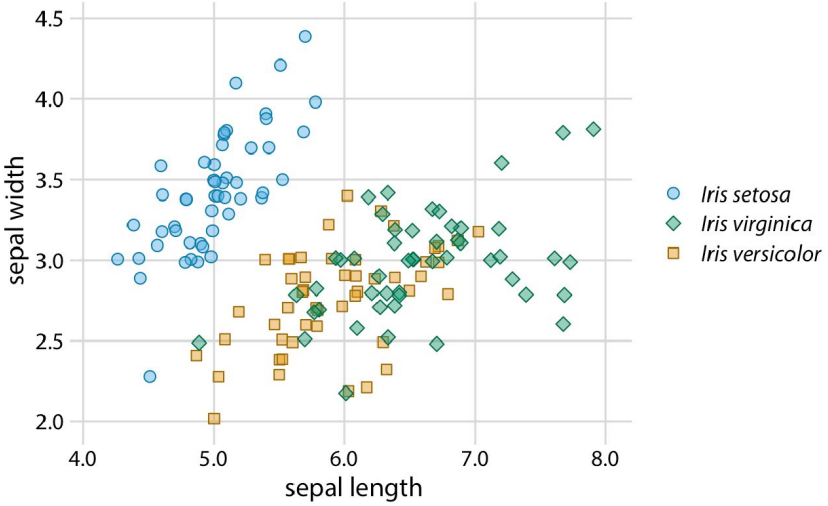
Sepal width versus sepal length for three different Iris species (Iris setosa, Iris virginica, and Iris versicolor). Each point represents the measurements for one plant sample. A small amount of jitter has been applied to all point positions to prevent overplotting

Visualizing association: x–y relationships



- Visualizing Associations: Scatterplots

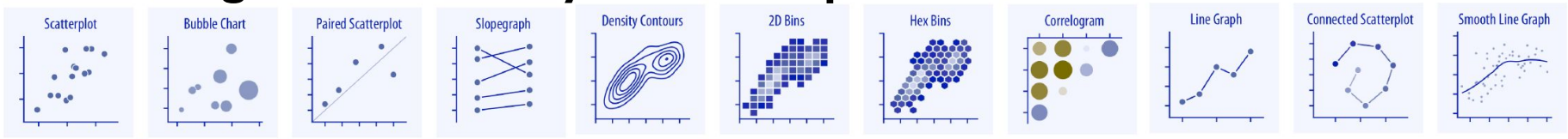




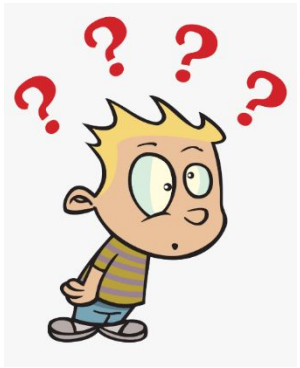
Better visual encoding:
different color + point shape (mark)

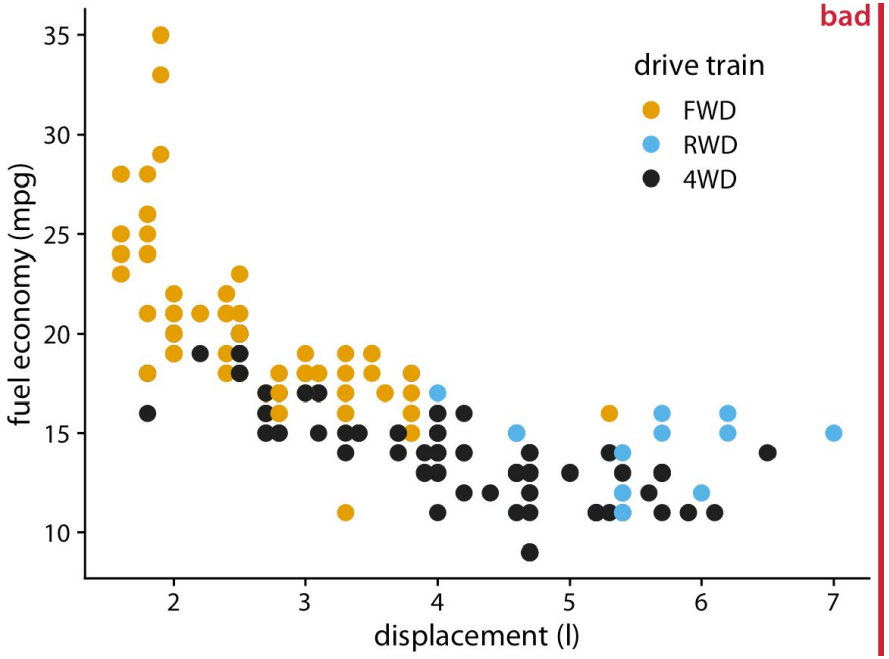
Sepal width versus sepal length for three different Iris species (Iris setosa, Iris virginica, and Iris versicolor). Each point represents the measurements for one plant sample. We have swapped the colors for Iris setosa and Iris versicolor and we have given each Iris species its own point shape

Visualizing association: x–y relationships



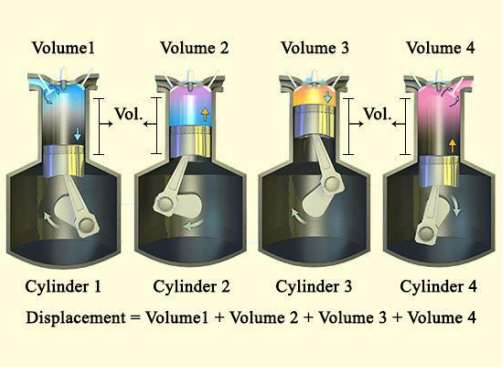
- Visualizing Associations: Scatterplots





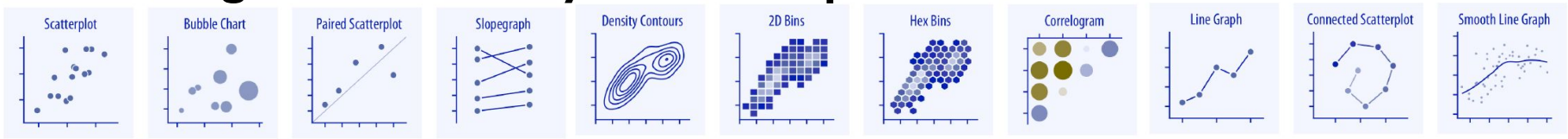
Many points are plotted on top of others and obscure them (popular cars have similar engine displacements)

City fuel economy versus engine displacement, for popular cars released between 1999 and 2008. Each point represents one car. The point color encodes the drive train: front-wheel drive (FWD), rear-wheel drive (RWD), or four-wheel drive (4WD).

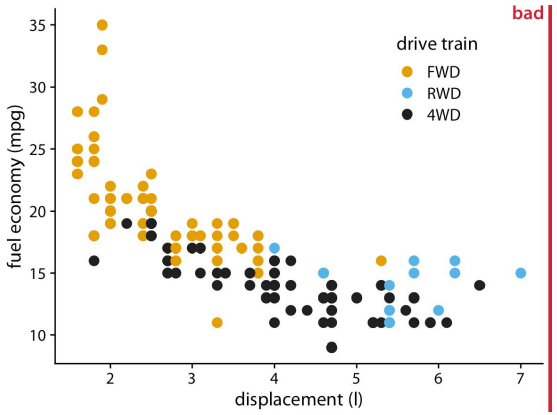


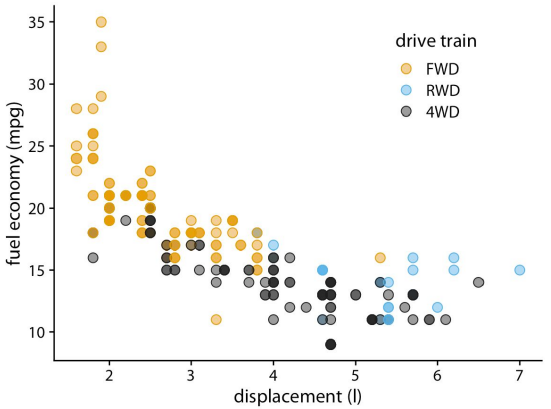
engine displacement

Visualizing association: x–y relationships

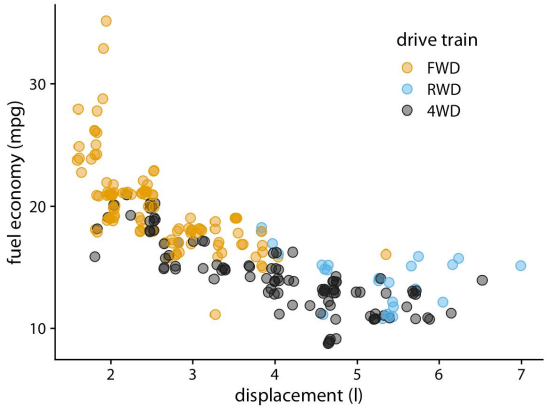


- Visualizing Associations: Scatterplots

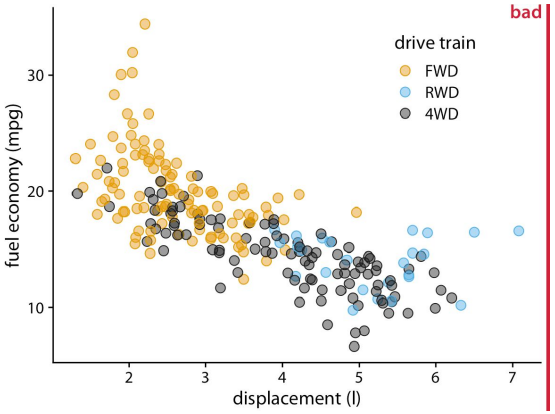




partially transparent

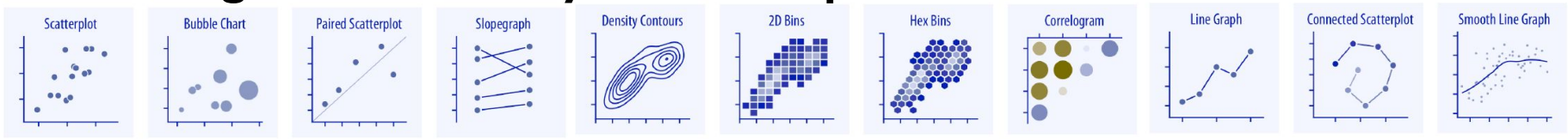


adding a small amount of jitter

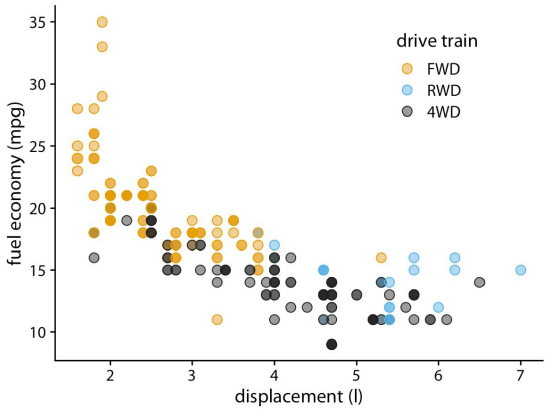


adding too much jitter

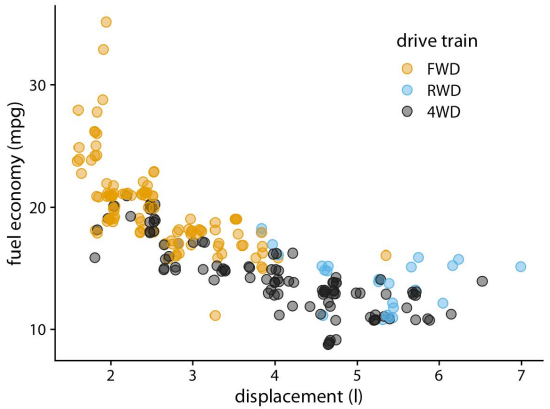
Visualizing association: x–y relationships



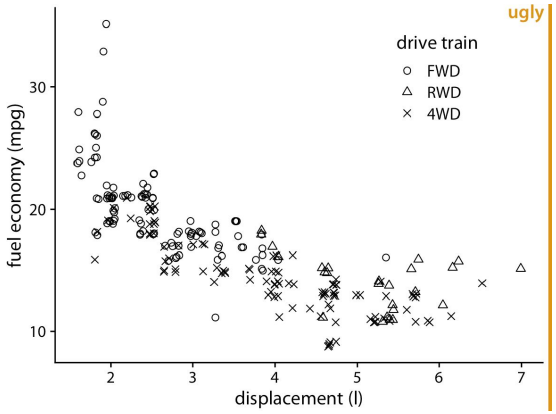
- Visualizing Associations: Scatterplots



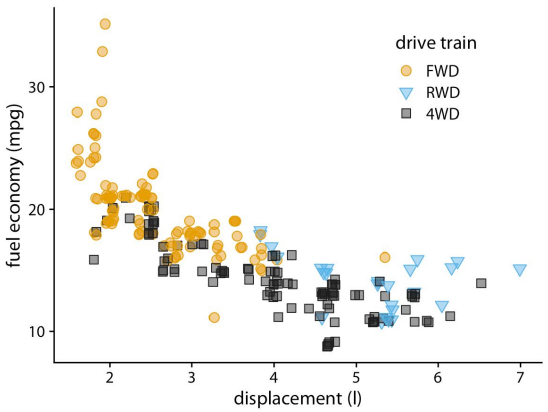
partially transparent



adding a small amount of jitter

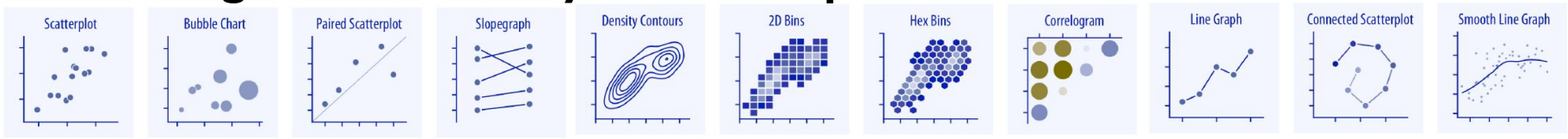


difficult to read



reproducible in grayscale

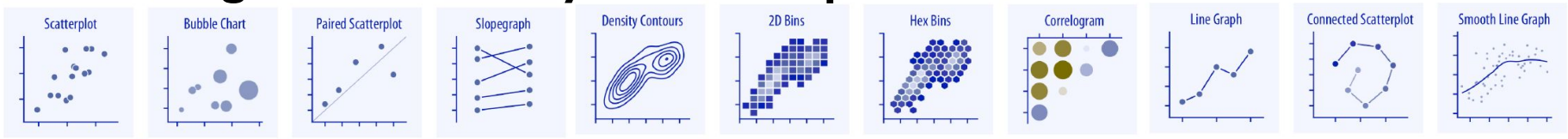
Visualizing association: x–y relationships



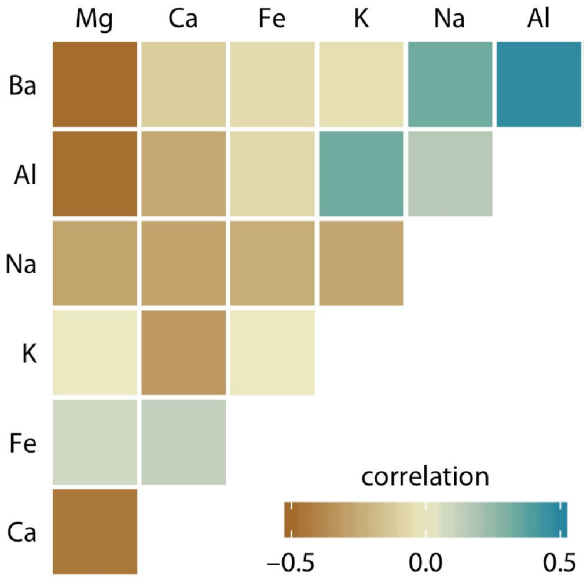
- Visualizing Associations: Scatterplots

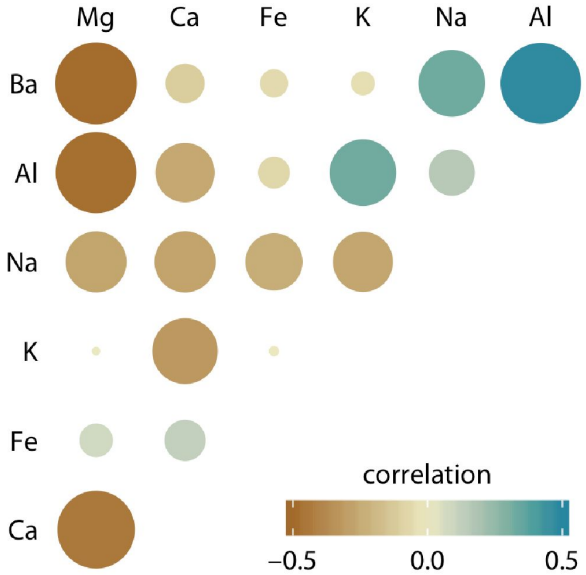
All-against-all scatterplot matrix of head length, body mass, and skull size, for 123 blue jays. This figure shows the exact same data. Because we are better at judging position than symbol size, correlations between skull size and the other two variables are easier to perceive in the pairwise scatterplots than in the previous plot

Visualizing association: x–y relationships



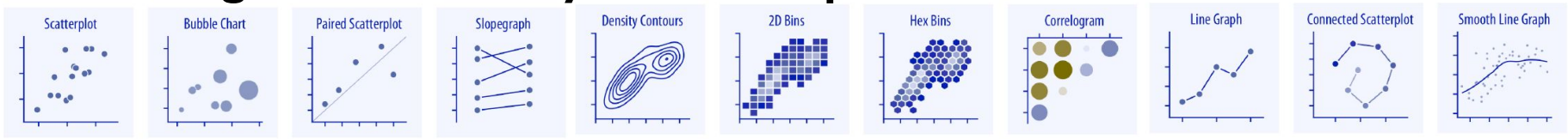
- Visualizing Associations: Scatterplots



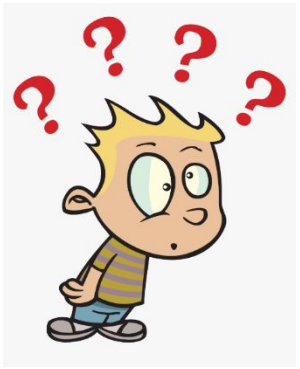


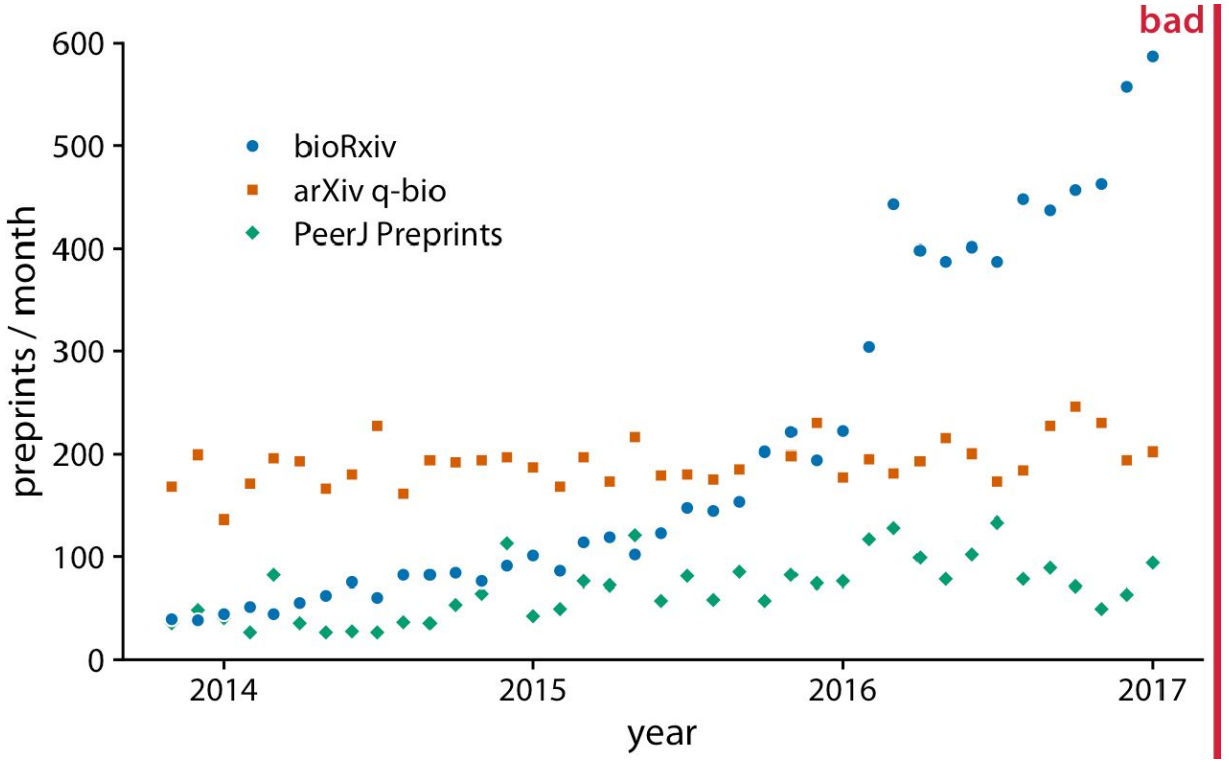
Correlations in mineral content for 214 samples of glass fragments obtained during forensic work. The dataset contains seven variables measuring the amounts of magnesium (Mg), calcium (Ca), iron (Fe), potassium (K), sodium (Na), aluminum (Al), and barium (Ba) found in each glass fragment. The colored tiles represent the correlations between pairs of these variables. The magnitude of each correlation is also encoded in the size of the colored circles. This choice visually deemphasizes cases with correlations near zero.

Visualizing association: x–y relationships



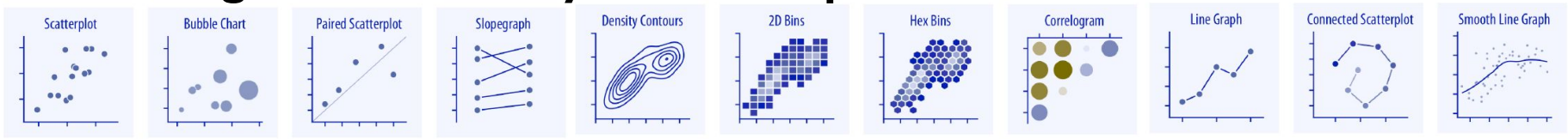
- Multiple time series



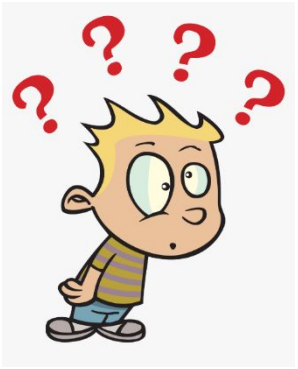


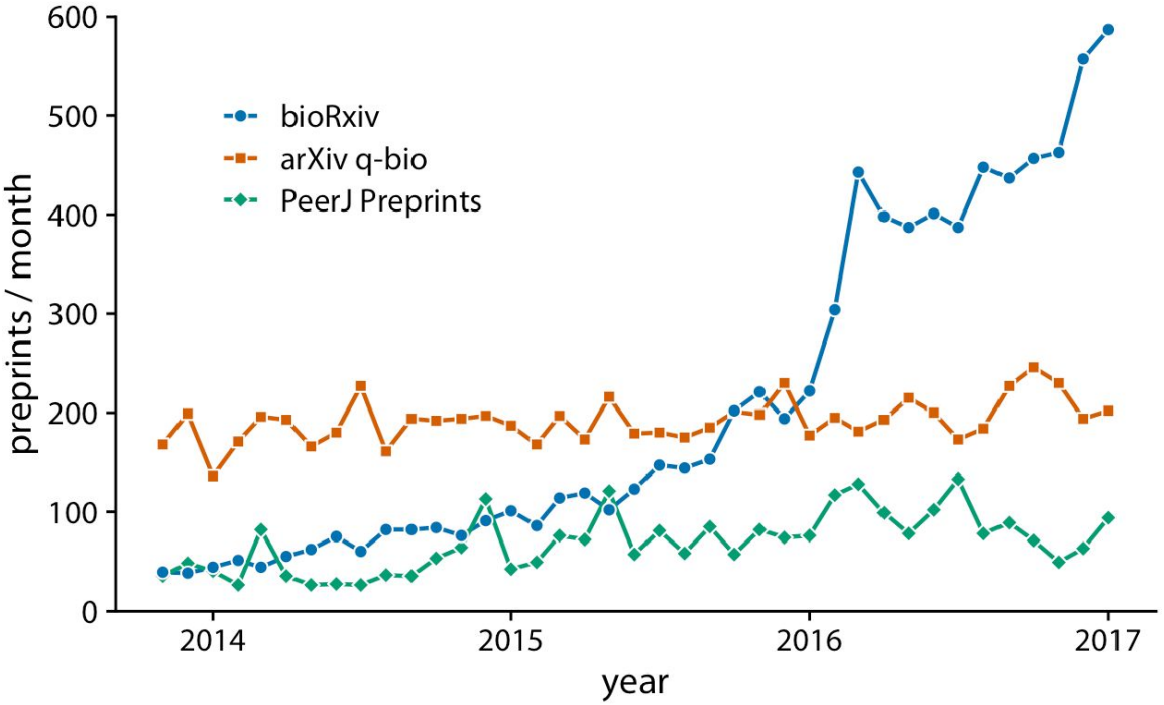
Monthly submissions to three preprint servers covering biomedical research: bioRxiv, the q-bio section of arXiv, and PeerJ Preprints

Visualizing association: x–y relationships



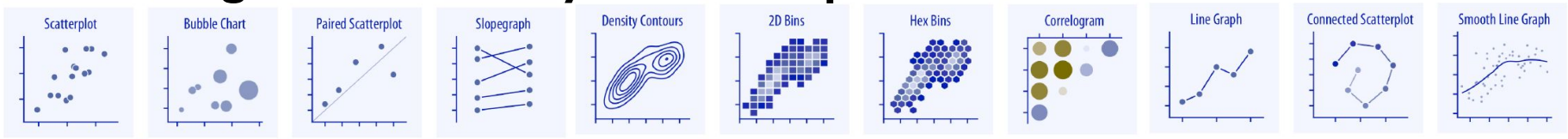
- Multiple time series



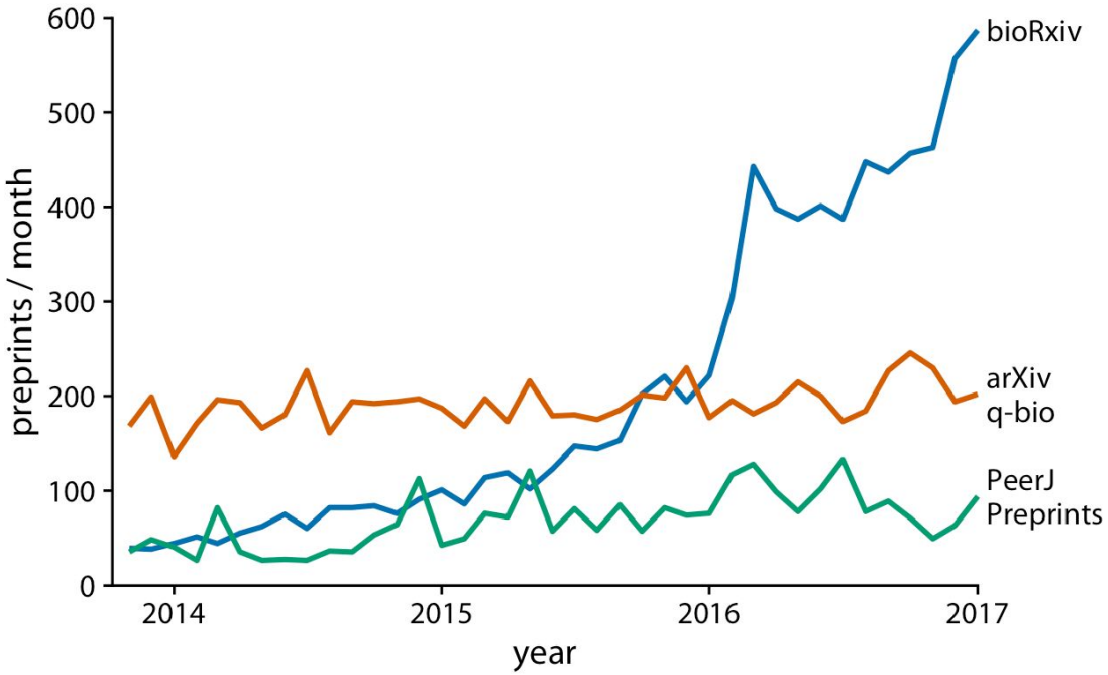


Monthly submissions to three preprint servers covering biomedical research.
By connecting the dots with lines, we help the viewer follow each individual time course.

Visualizing association: x–y relationships

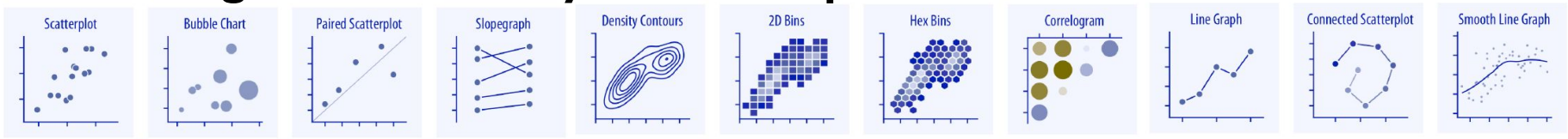


- Multiple time series

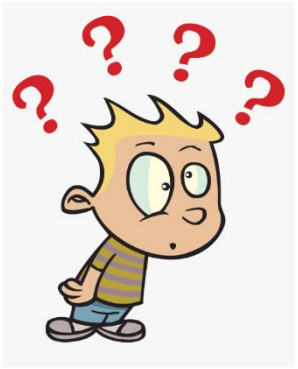


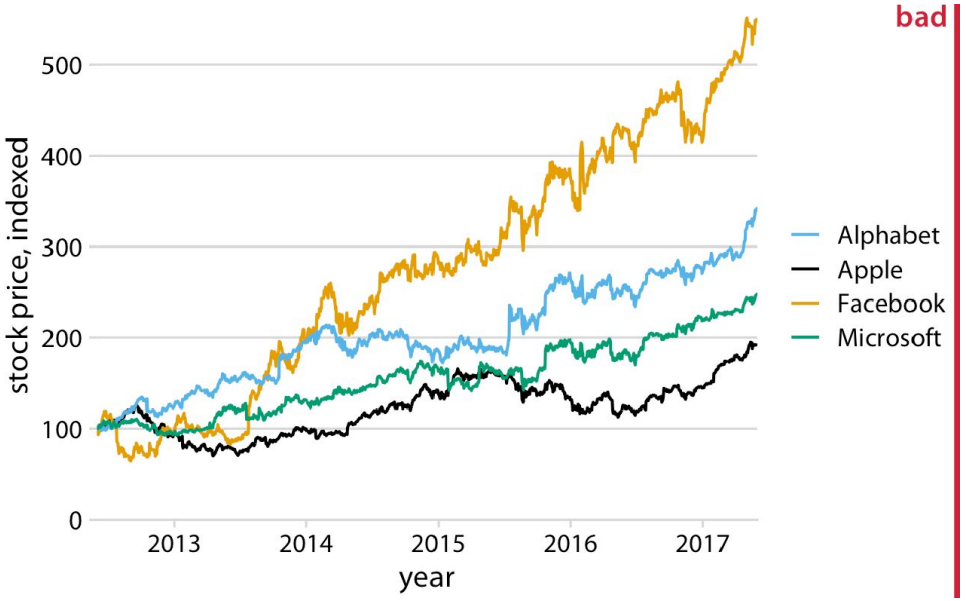
Monthly submissions to three preprint servers covering biomedical research. Directly labeling the lines instead of providing a legend reduces the cognitive load required to read the figure, and eliminating the legend removes the need for points of different shapes. This enables us to streamline the plot further by eliminating the dots.

Visualizing association: x–y relationships



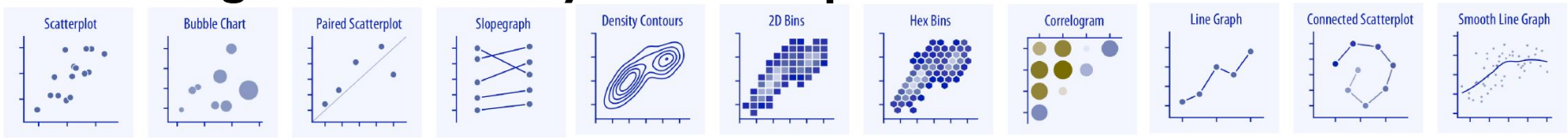
- Multiple time series



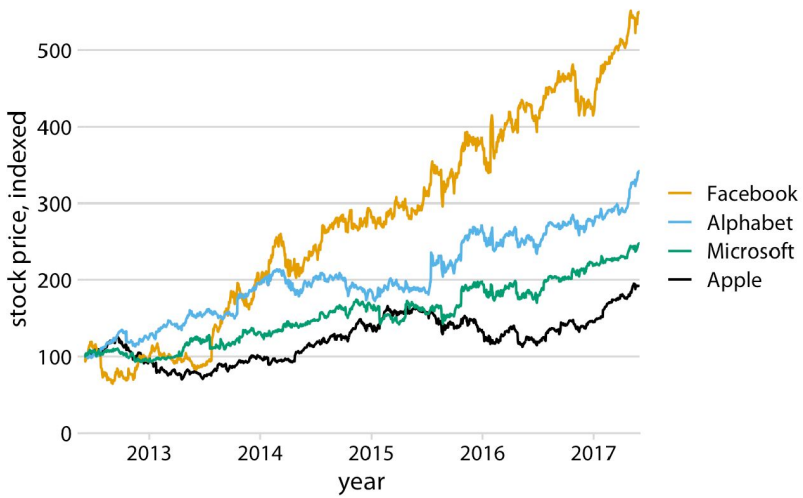


Stock price over time for four major tech companies. The stock price for each company has been normalized to equal 100 in June 2012

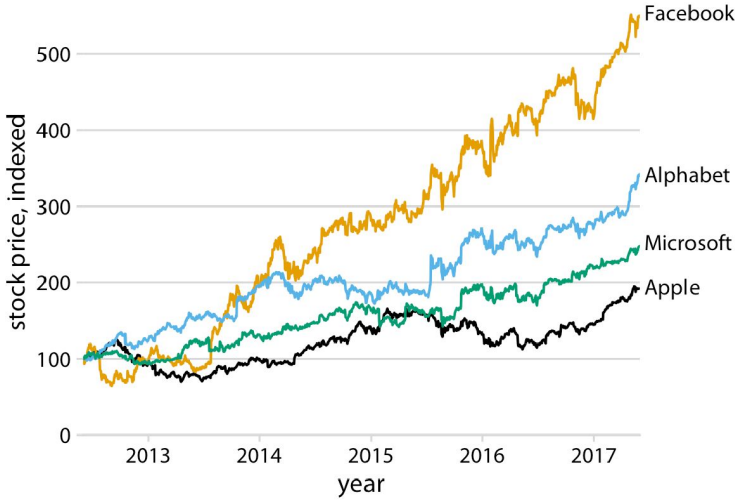
Visualizing association: x–y relationships



- Multiple time series
  - Match ordering, and if possible, design your figures so they don't need a separate legend



Order matching



No legend – direct labeling

Stock price over time for four major tech companies. The stock price for each company has been normalized to equal 100 in June 2012

Summary

- Part I
  - Graphical Integrity – “functional art”
  - Visualization Design Principles
    - Maximize data-ink ratio
    - Avoid chart junk
    - Increase data density
  - Graphic Design Principles: CRAP
    - Contrast, Repetition, Alignment, Proximity
- Part II
  - Data Visualization Steps & Visual Encoding
  - Visualization Taxonomy & Statistical Graphs – A Tour through the visualization zoo