HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Visualization for table, multi-dimensional data

Outline

- Previous lesson
  - Coordinate systems and axes
  - Color scales
  - Visualizing amounts
  - Visualizing distributions
  - Visualizing many distributions at once
- Today lesson
  - Visualizing proportions
  - Visualizing nested proportions
  - Visualizing associations
  - Visualizing trends
  - Visualizing uncertainty

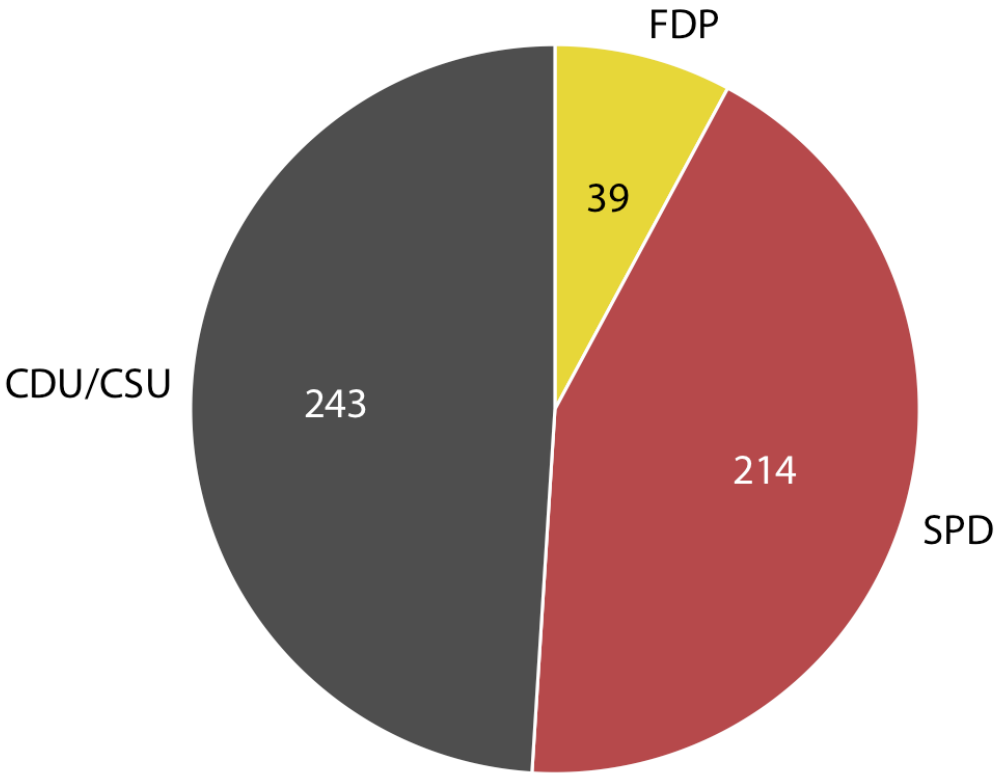
Visualizing proportions

Scenarios

- Show how some group, entity, or amount breaks down into individual pieces that each represent a proportion of the whole.
- Examples:
  - The proportions of men and women in a group of people.
  - The percentages of people voting for different political parties in an election.
  - The market shares of companies.

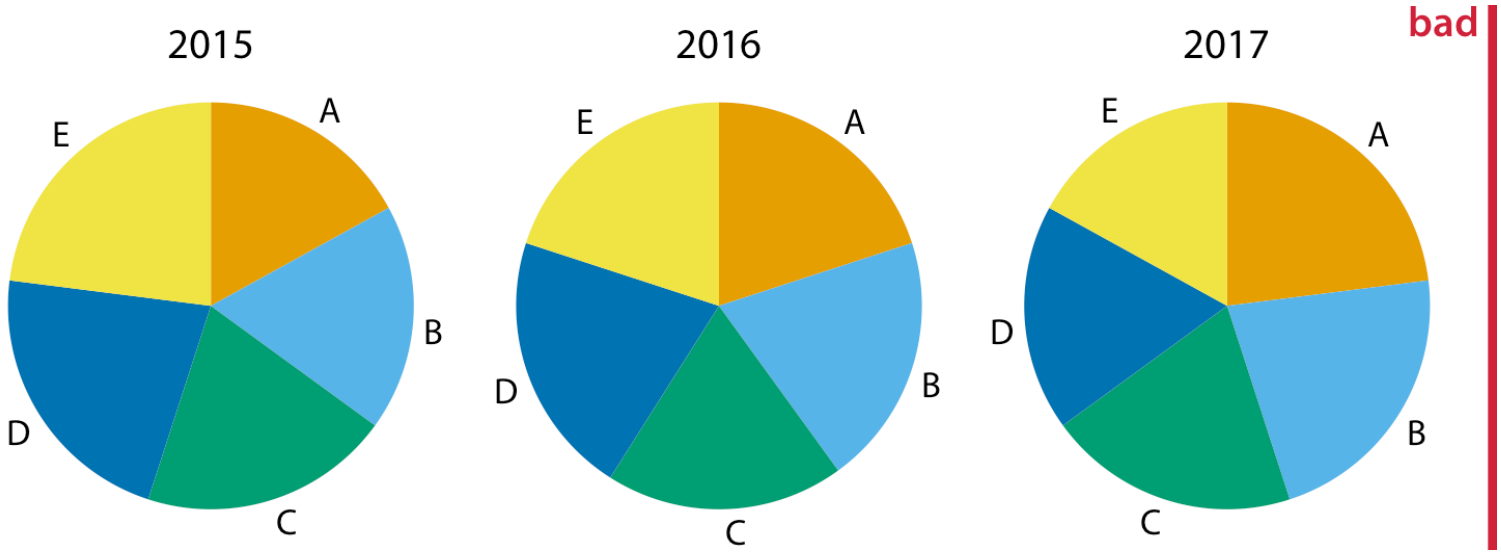
Example: Party composition of the eighth German Bundestag, 1976–1980

- Visualized as a pie chart
  - Breaks a circle into slices such that the area of each slice is proportional to the fraction of the total it represents.



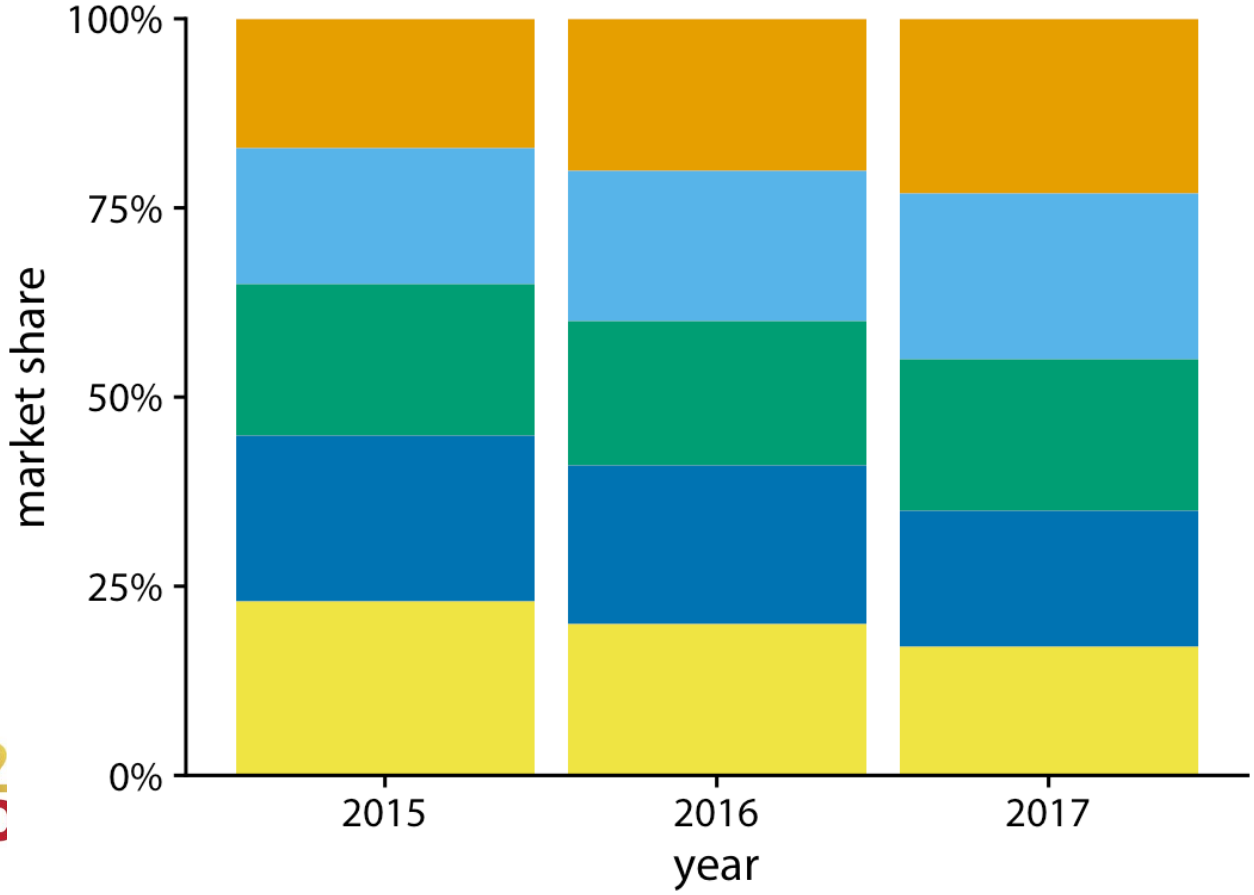
A case where pie charts fail

- Market share of five hypothetical companies, A–E, for the years 2015–2017
  - A comparison of relative market share within years is nearly impossible.
  - Changes in market share across years are difficult to see.



Stacked bars

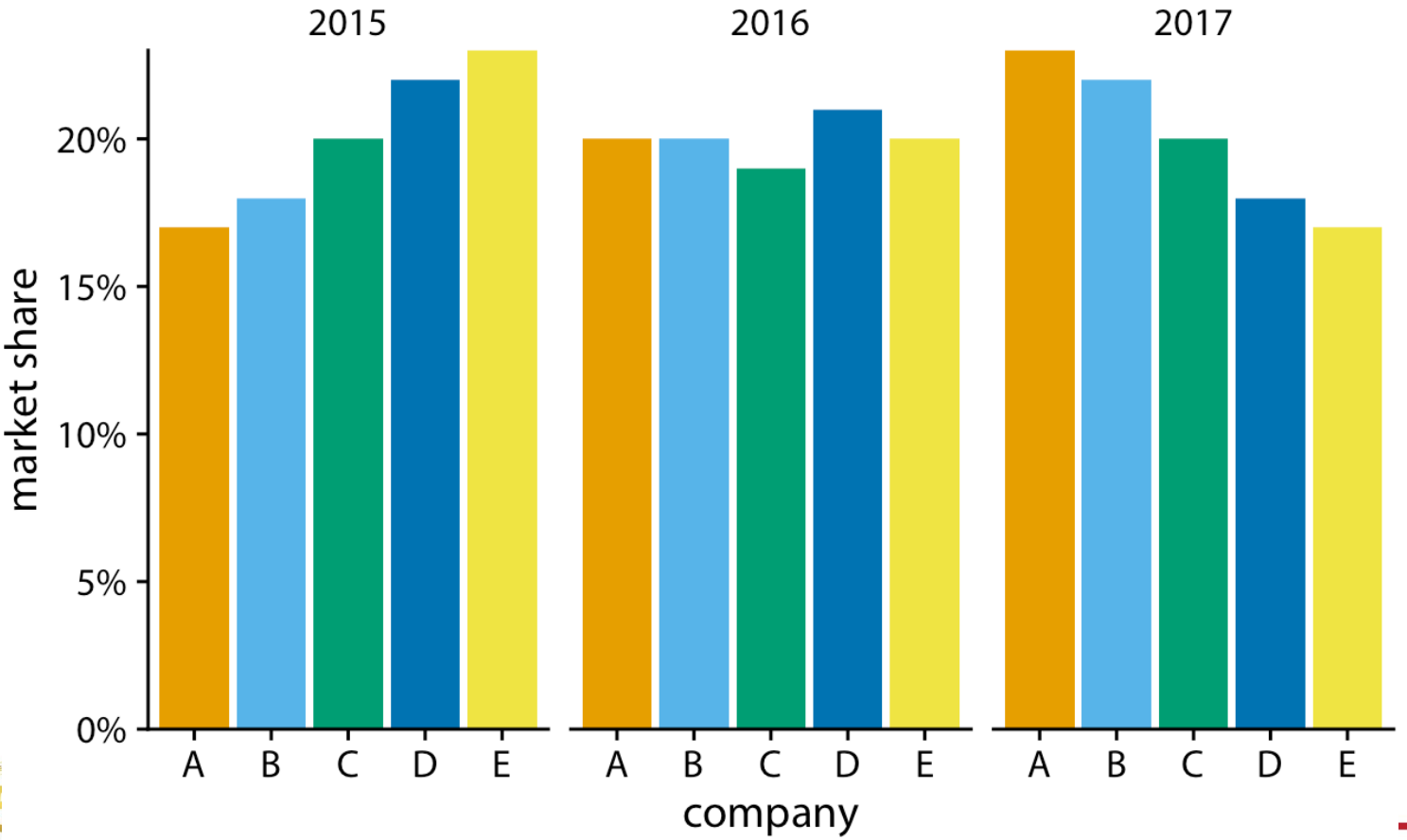
- The trends of a growing market share for company A and a shrinking market share for company E are clearly visible.
- The relative market shares of the five companies within each year are still hard to compare



bad

Side-by-side bars

- Market share of five hypothetical companies for the years 2015–2017, visualized as side-by-side bars.

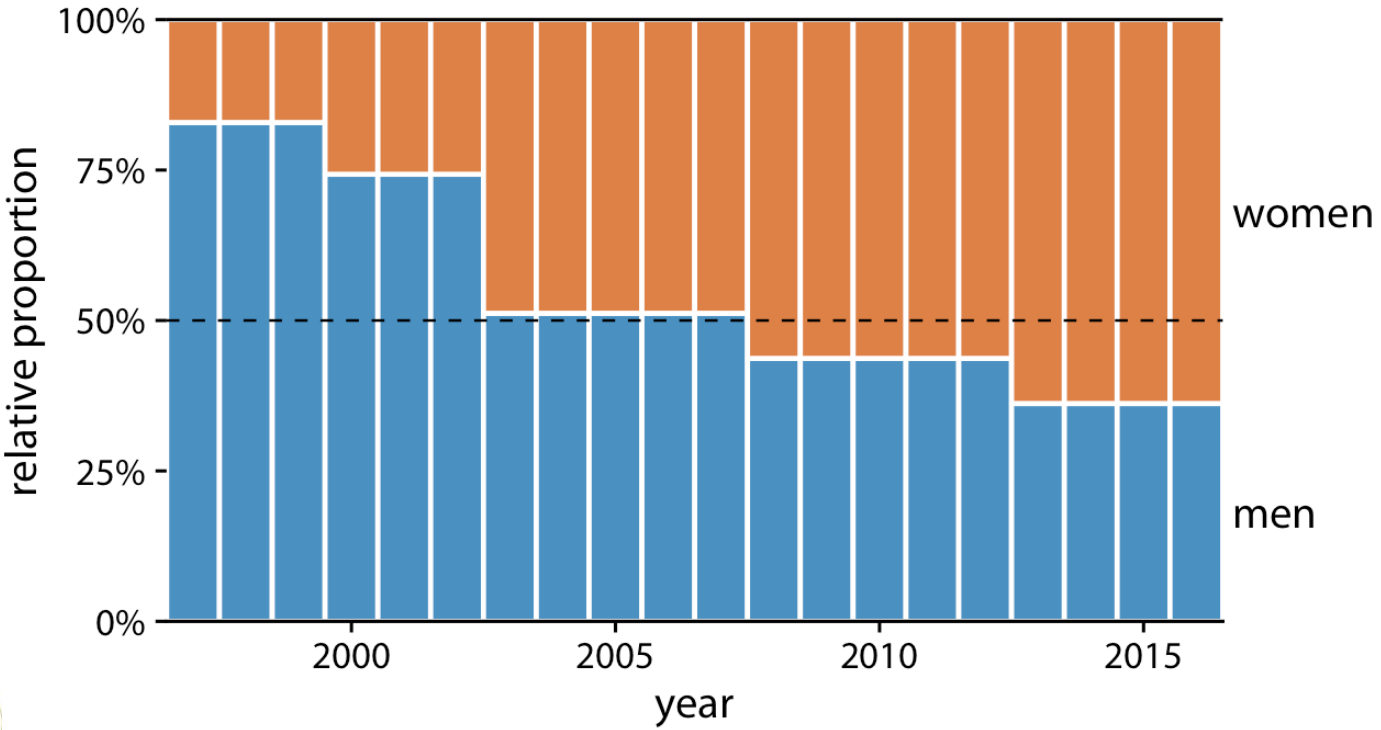


Pros and cons of common approaches to visualizing proportions

	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	✗	✓	✗

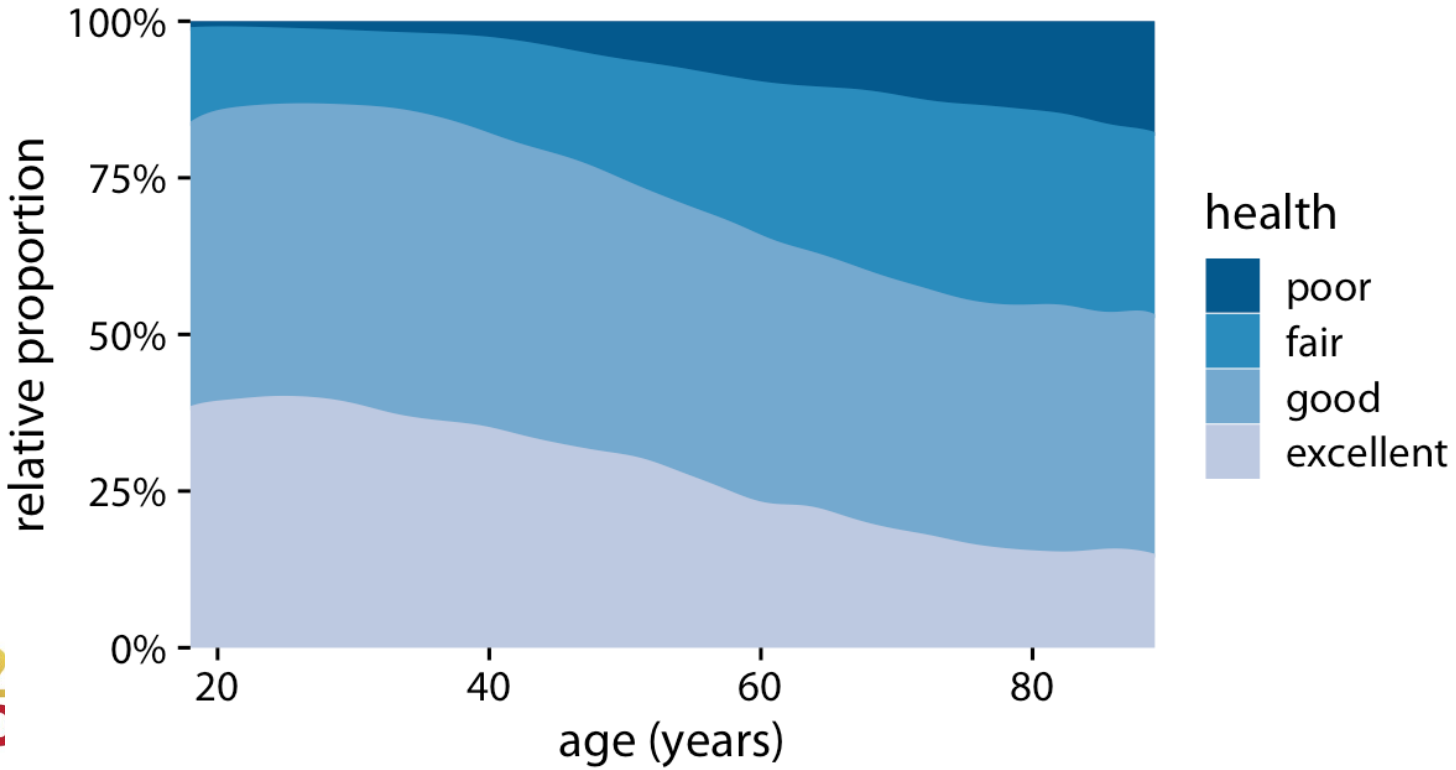
A case for Stacked Bars

- The problem of shifting internal bars disappears if there are only two bars in each stack.
- Example: Change in the gender composition of the Rwandan parliament over time, 1997 to 2016.



A case for Stacked Densities

- Stacked densities can be thought of many infinitely small stacked bars arranged side-by-side
- Visualize how proportions change in response to a continuous variable
- Example: Health status by age

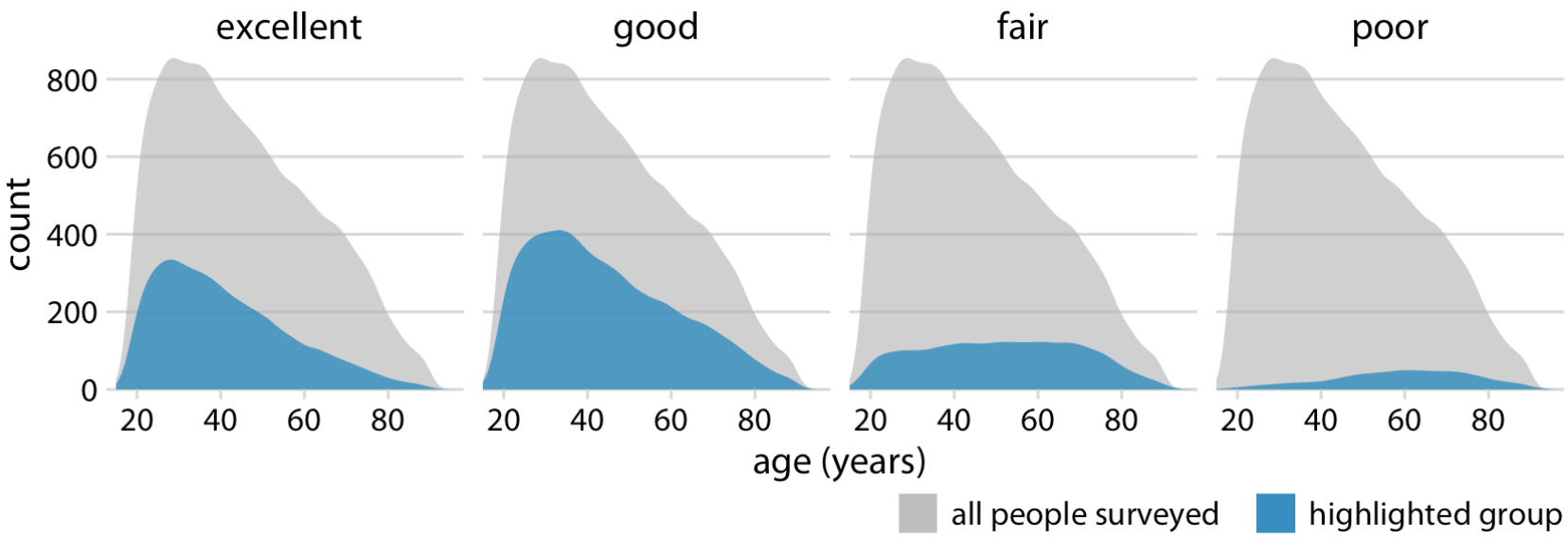


Visualizing proportions separately as parts of the total

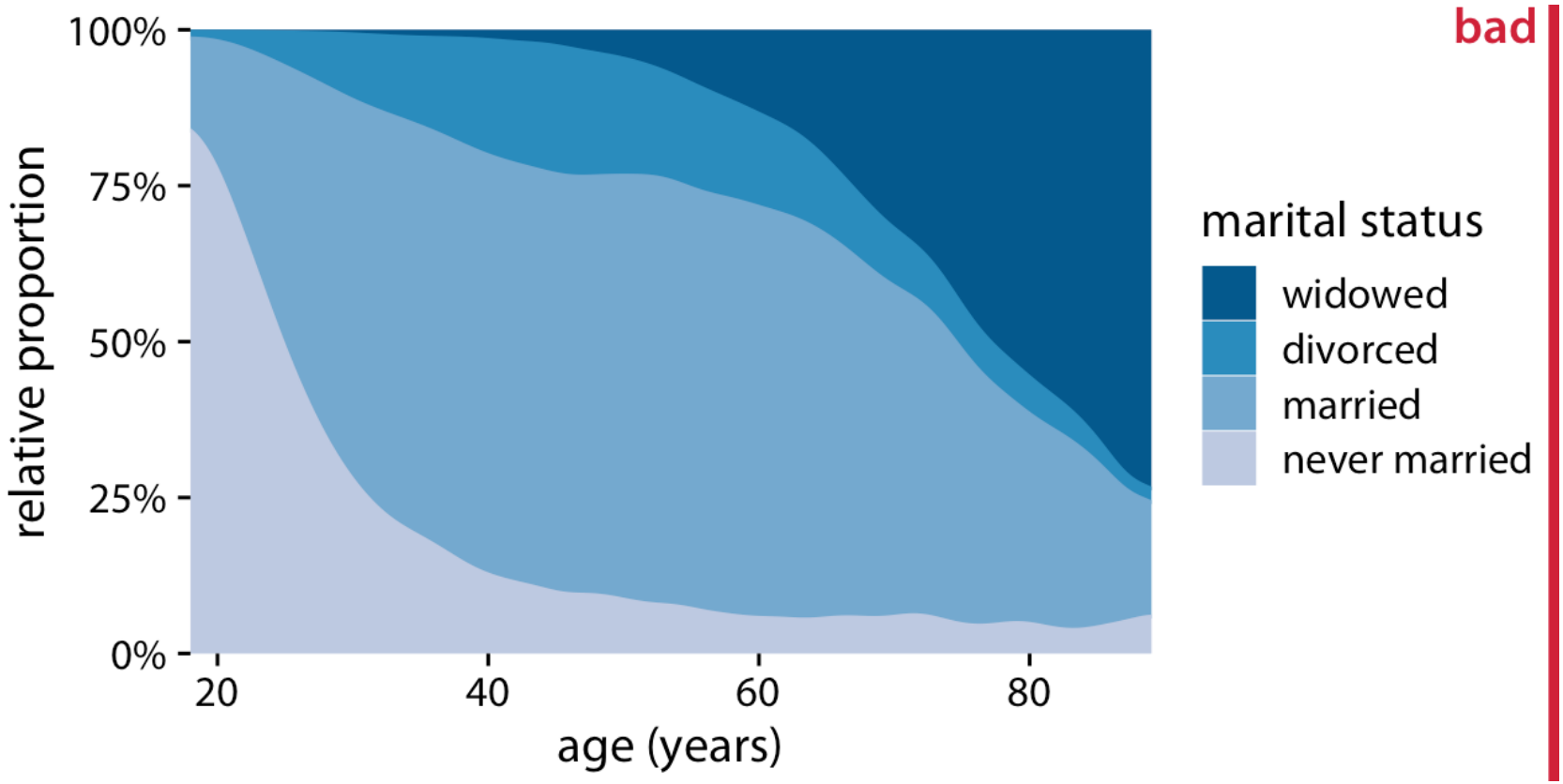
- Side-by-side bars have the problem that they don't visualize the size of the individual parts relative to the whole
- Stacked bars have the problem that the different bars cannot be compared easily because they have different baselines

Example: Health status by age, shown as proportion of the total number of people

- The colored areas show the density estimates of the ages of people with the respective health status
- The gray areas show the overall age distribution

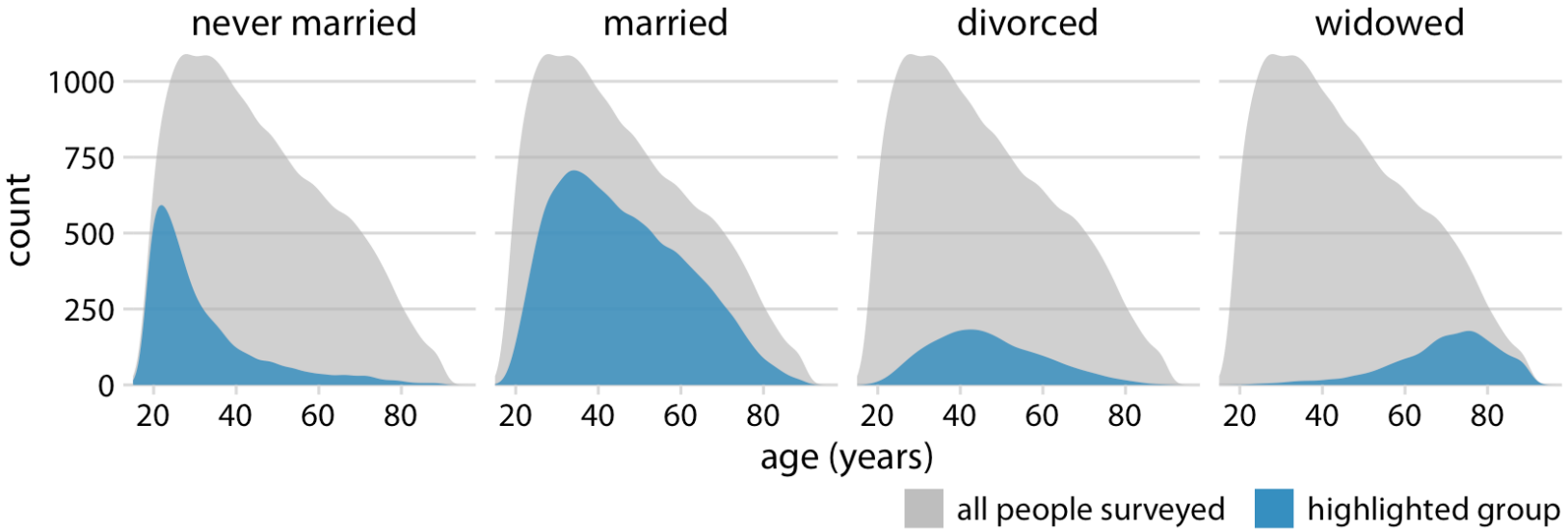


Example: Marital status by age



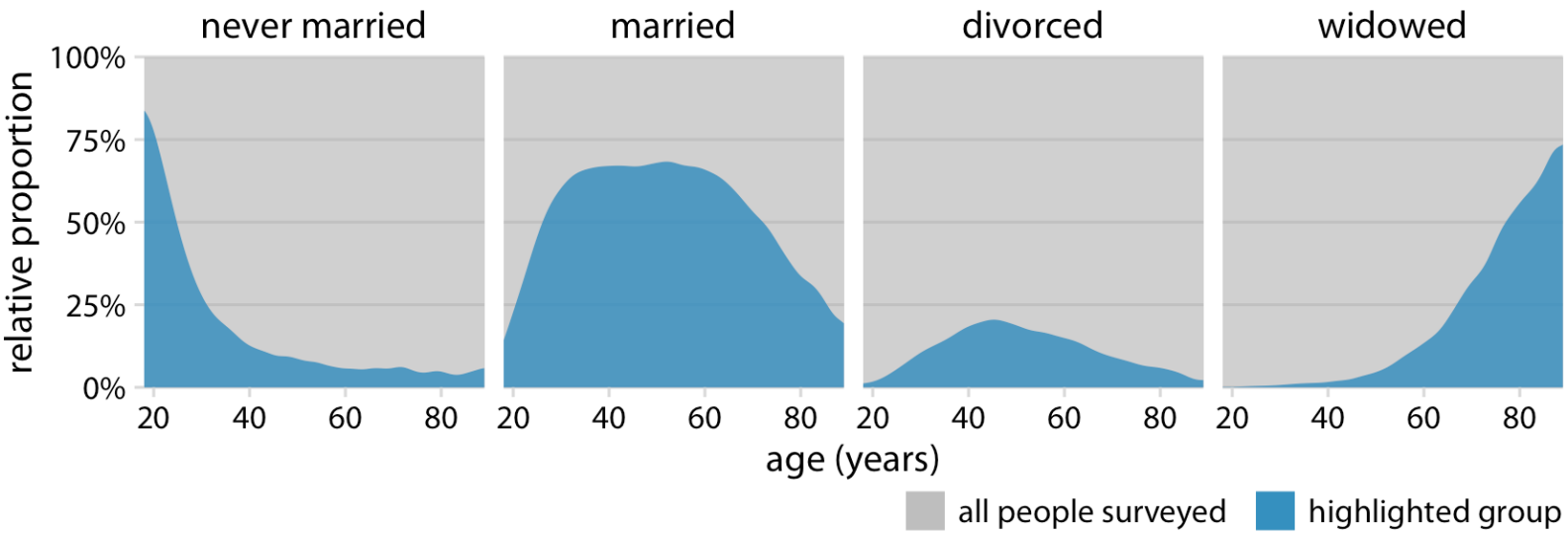
Example: Marital status by age, shown as proportion of the total number of people

- The colored areas show the density estimates of the ages of people with the respective marital status.
- The gray areas show the overall age distribution.
- Still, this representation doesn't make it easy to determine relative proportions at any given point in time.



Example: Marital status by age, shown as proportion of the total number of people

- Show relative proportions instead of absolute counts along the y axis



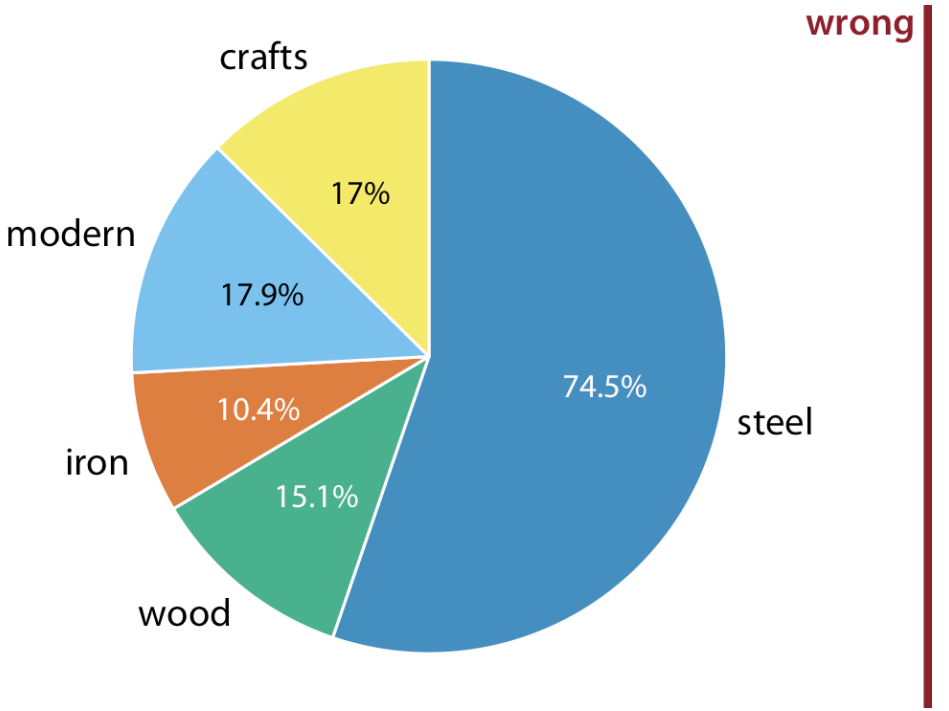
Visualizing nested proportions

Scenarios

- Break down a dataset by multiple categorical variables at once.
- Example
  - Visualize both the fraction of bridges made from steel, iron, or wood and the fraction that are crafts or modern.

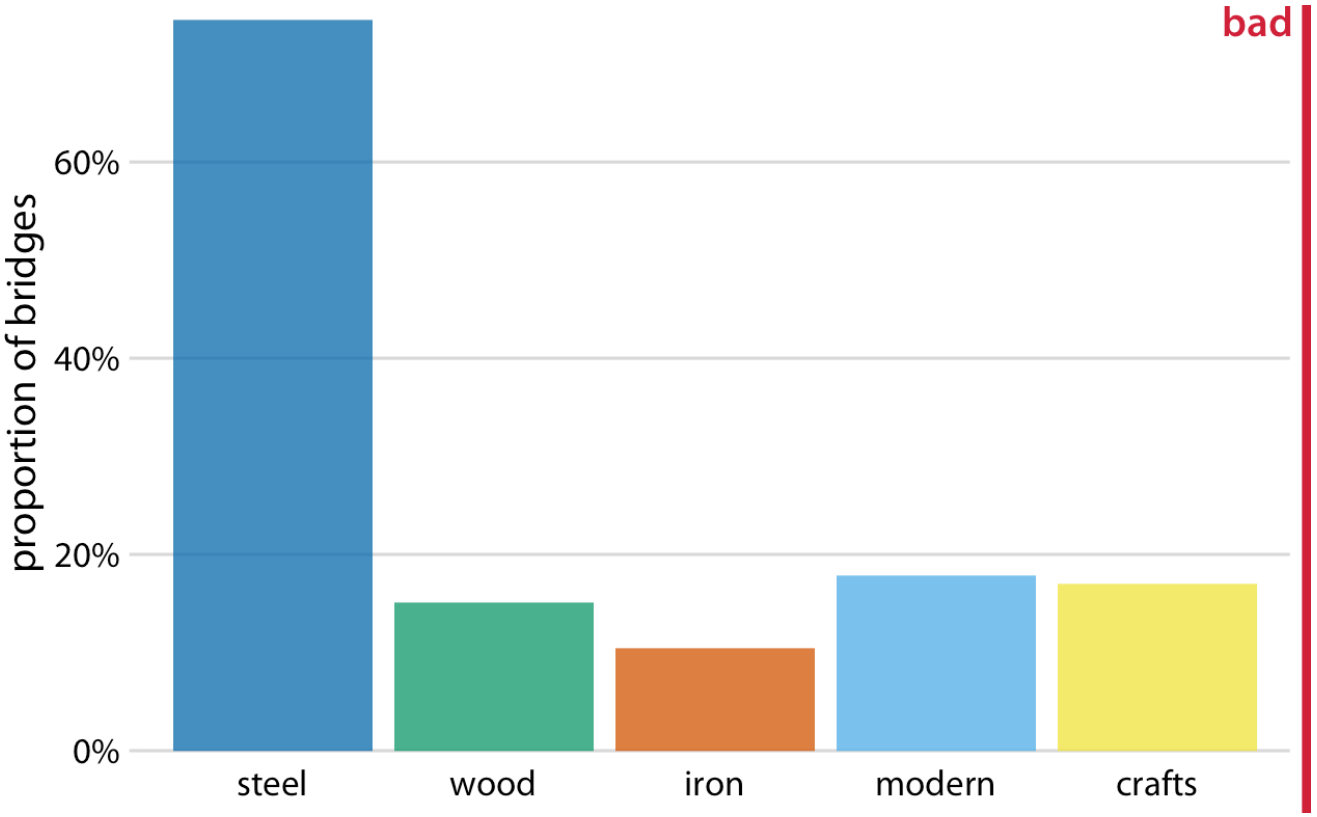
Example: Breakdown of bridges in Pittsburgh

- By construction material (steel, wood, iron) and by date of construction (crafts, before 1870, and modern, after 1940).
- The percentages add up to more than 100%.



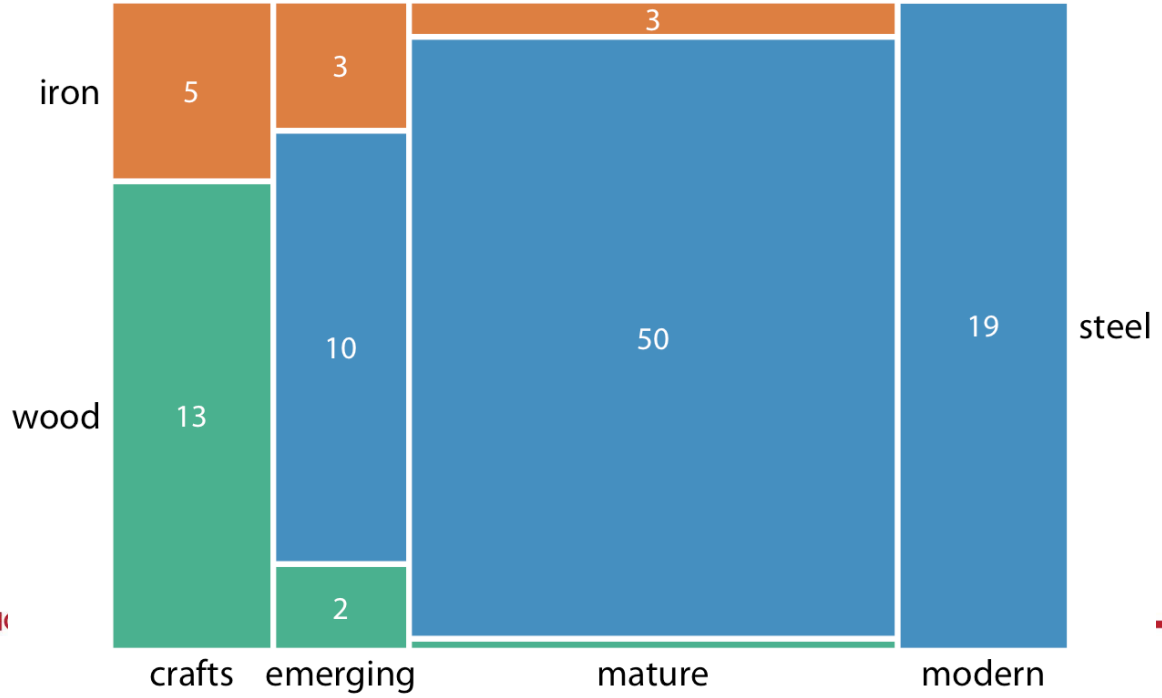
Example: Breakdown of bridges in Pittsburgh as a bar plot

- Does not clearly indicate the overlap among different groups



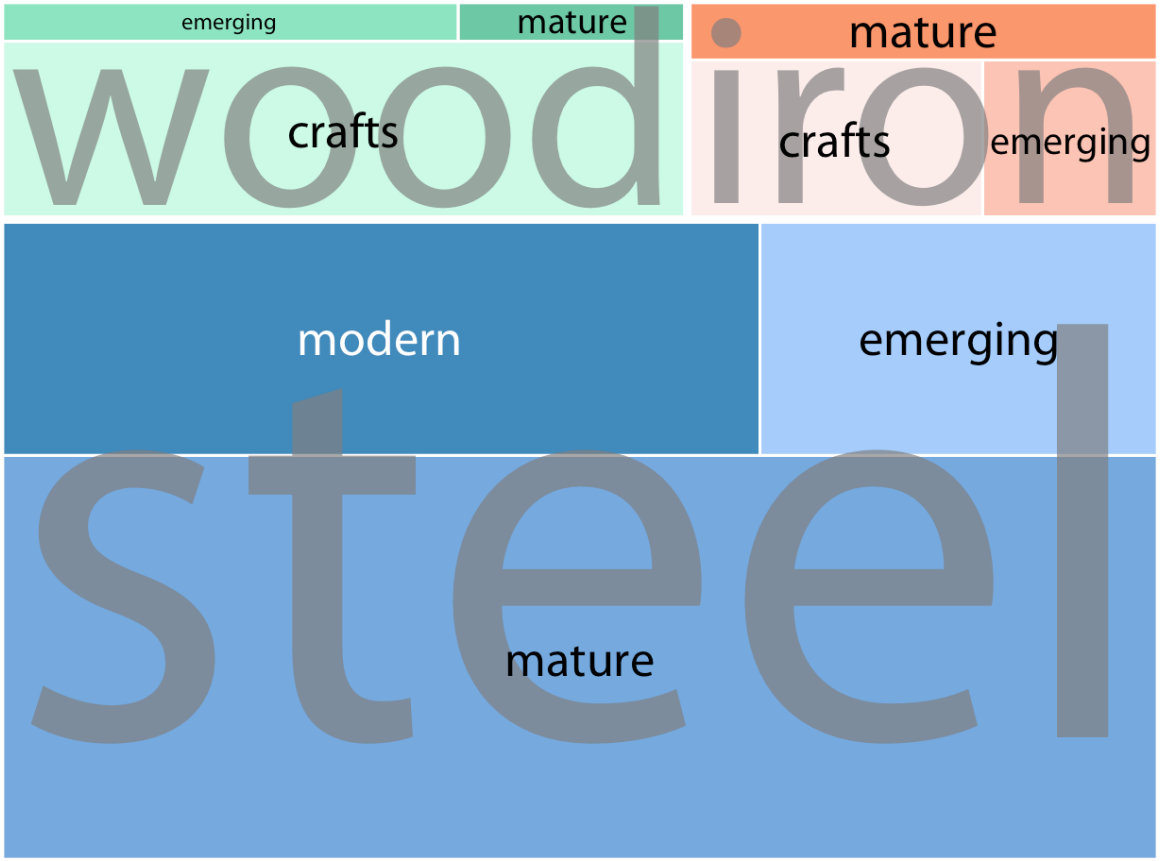
Mosaic plots

- The widths of each rectangle are proportional to the number of bridges constructed in that era.
- The heights are proportional to the number of bridges constructed from that material.
- Numbers represent the counts of bridges within each category.
- A critical condition for a mosaic plot: every categorical variable shown must cover all the observations in the dataset.

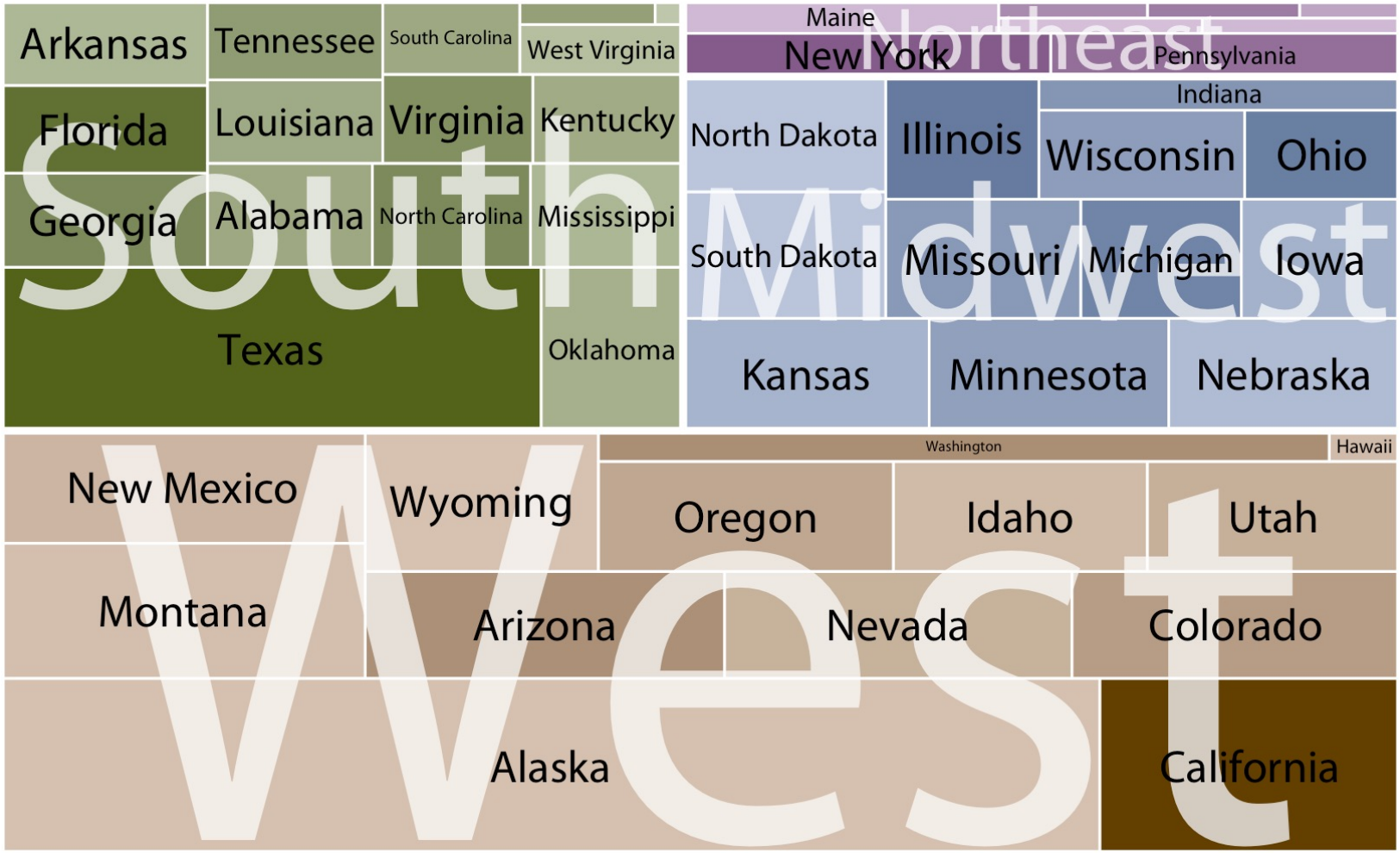


Tree map

- Take an enclosing rectangle and subdivide it into smaller rectangles whose areas represent the proportions.
  - Recursively nest rectangles inside each other.



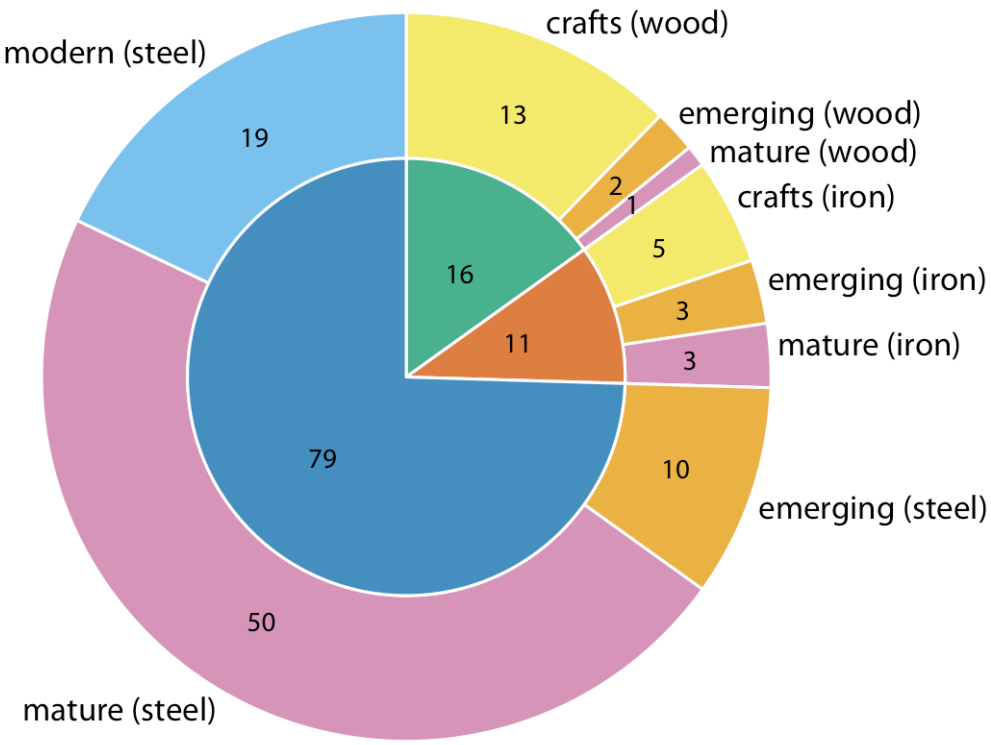
Example: States in the US visualized as a treemap



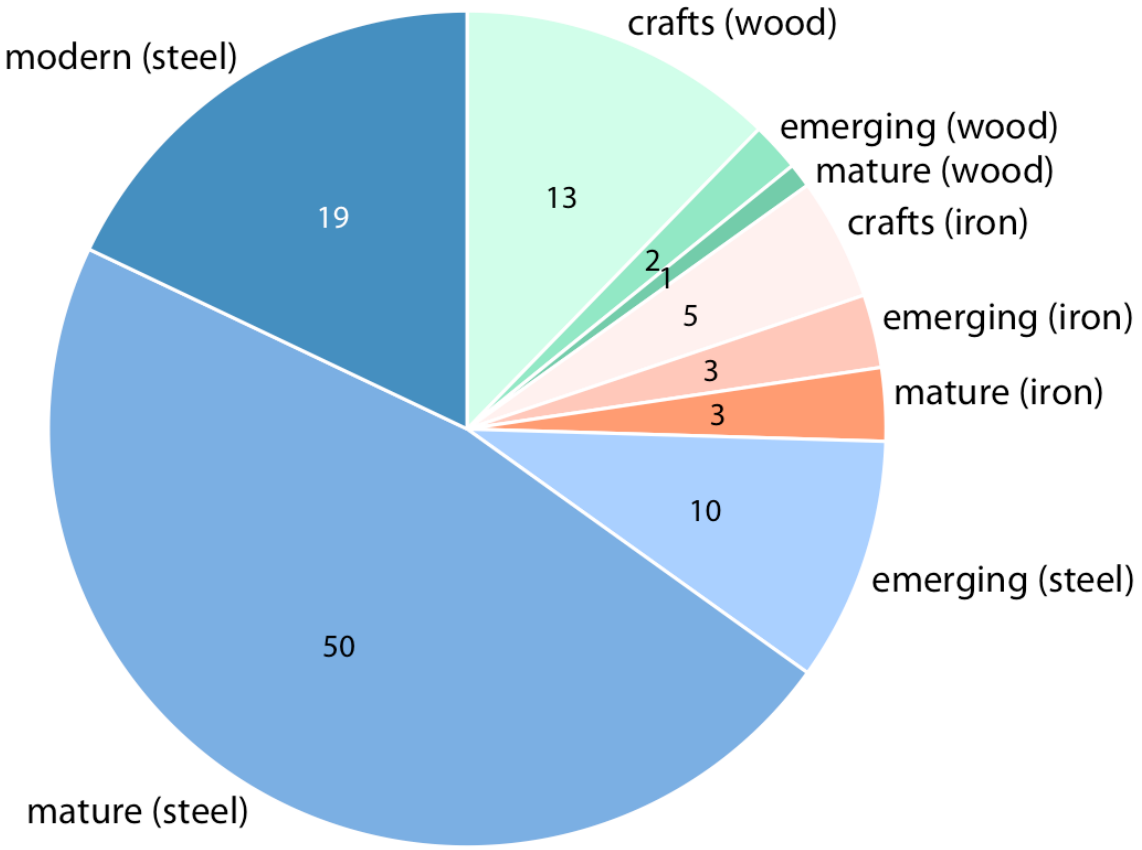
Nested pies

- Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron; inner circle) and by era of construction (crafts, emerging, mature, modern; outer circle).
- Numbers represent the counts of bridges within each category

ugly

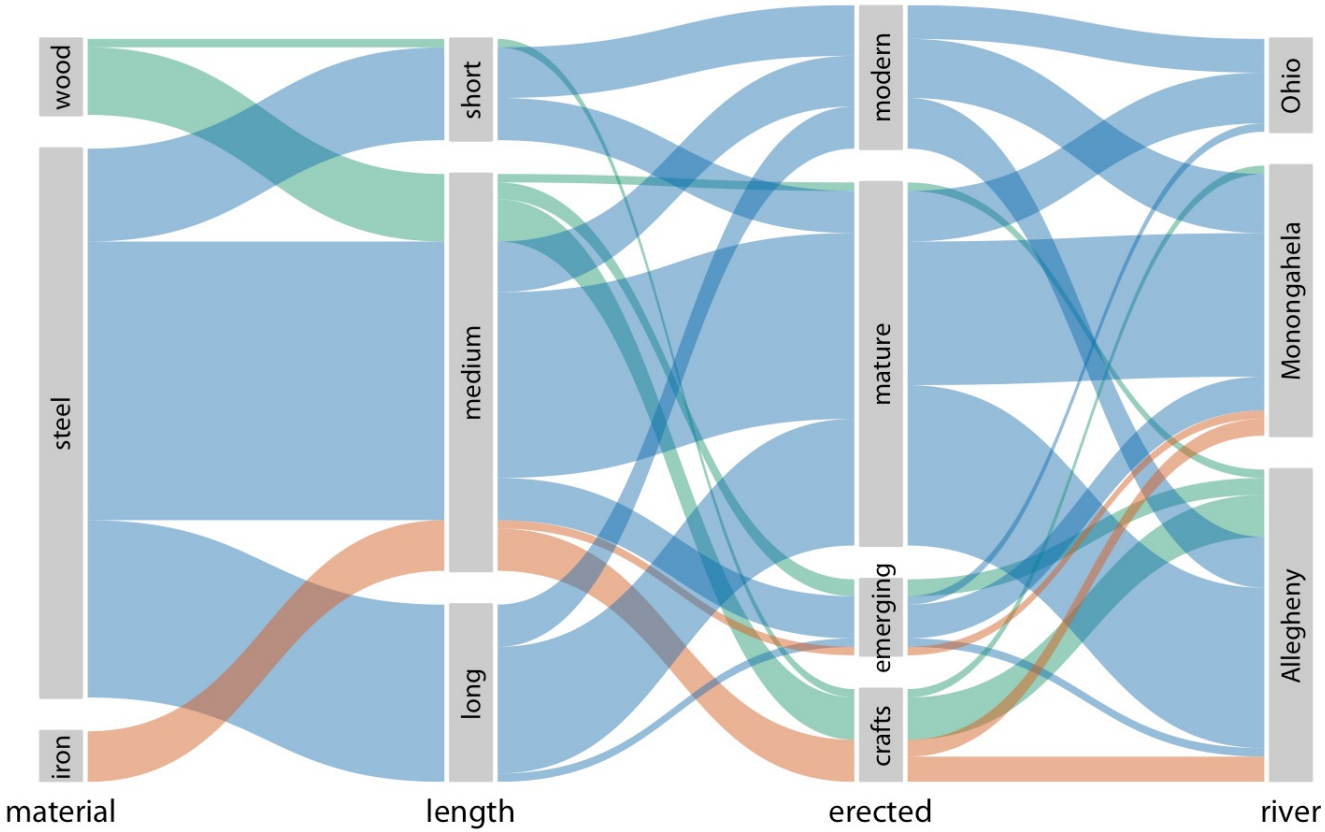


Example: Breakdown of bridges in Pittsburgh by construction material and by era of construction



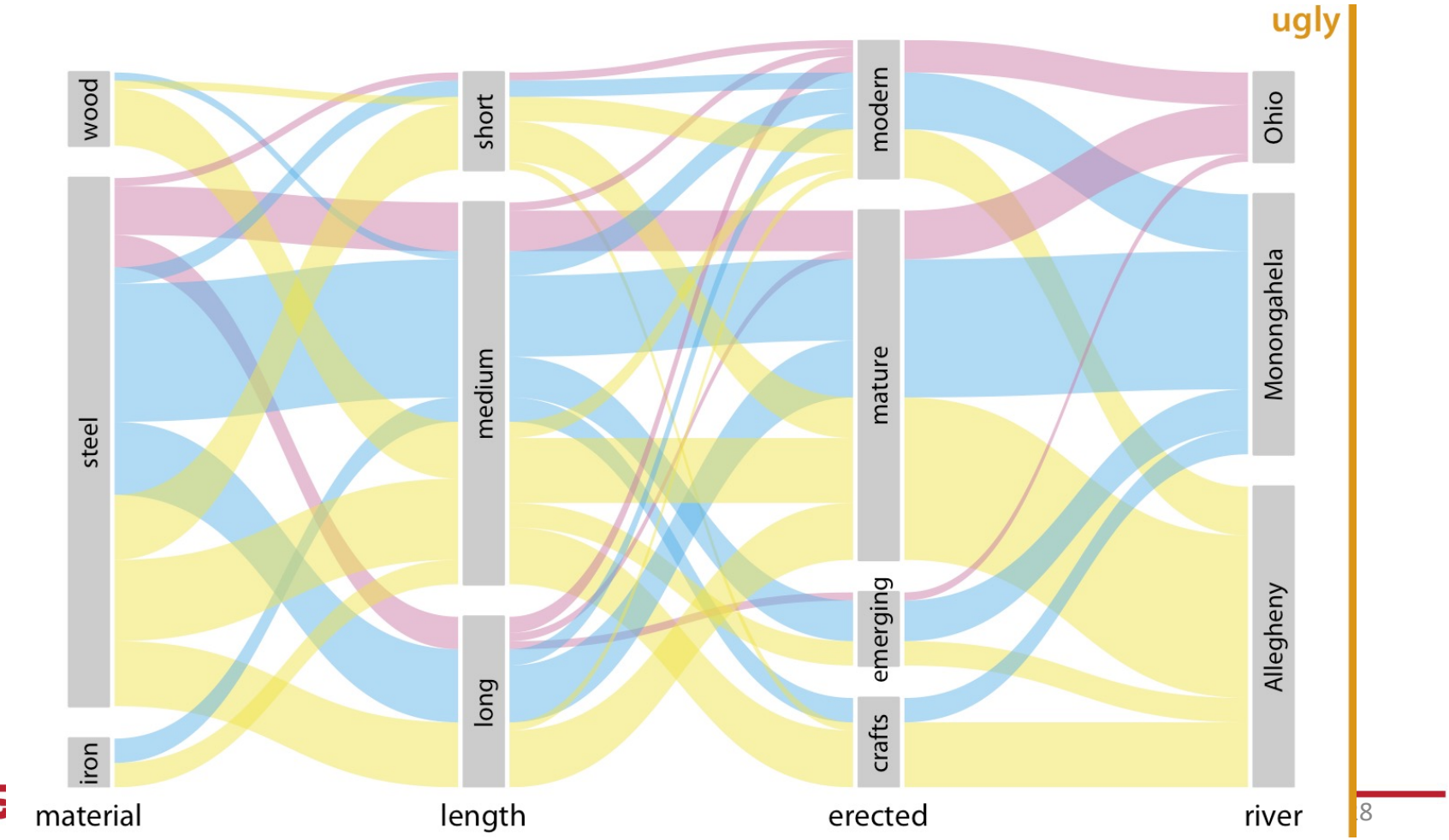
Parallel sets

- Show how the total dataset breaks down by each individual categorical variable.
- Then draw shaded bands that show how the subgroups relate to each other.



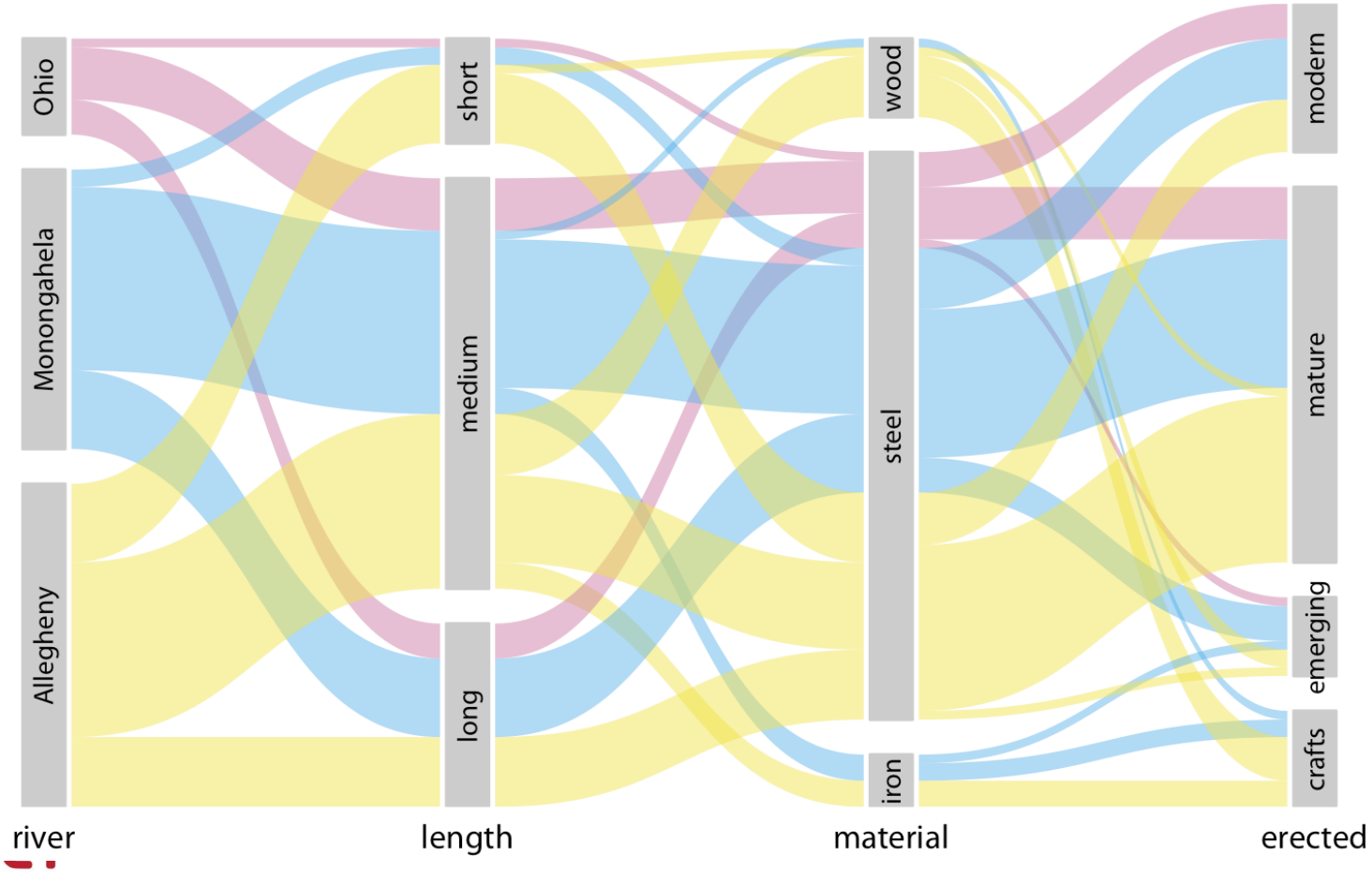
Example: Breakdown of bridges in Pittsburgh by construction material, length, era of construction, and the river they span

- The coloring of the bands highlights the river spanned by the different bridges.



Example: Breakdown of bridges in Pittsburgh by construction material, length, era of construction, and the river they span

- The modified order results in a figure that is easier to read and less busy.



Visualizing associations

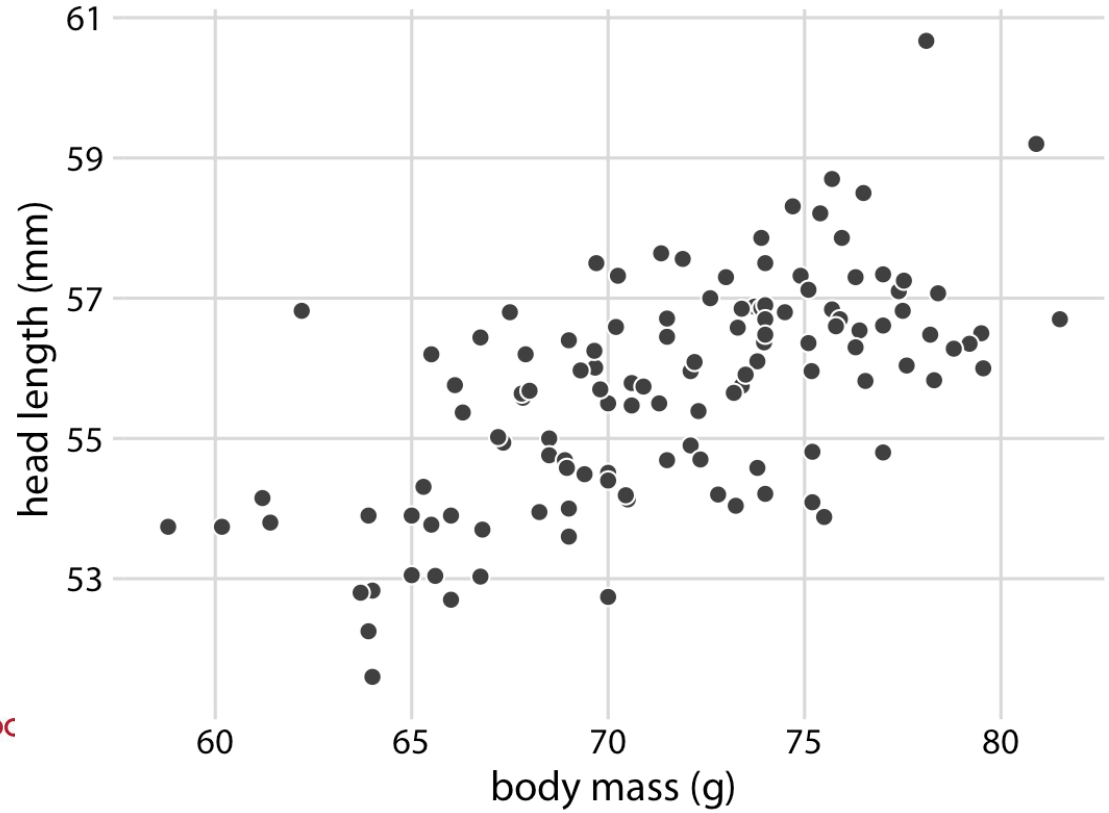
Among two or more Quantitative Variables

Scenarios

- Many datasets contain two or more quantitative variables.
- We may be interested in how these variables relate to each other.
- Example:
  - A dataset of quantitative measurements of different animals, such as the animals' height, weight, length, and daily energy demands.

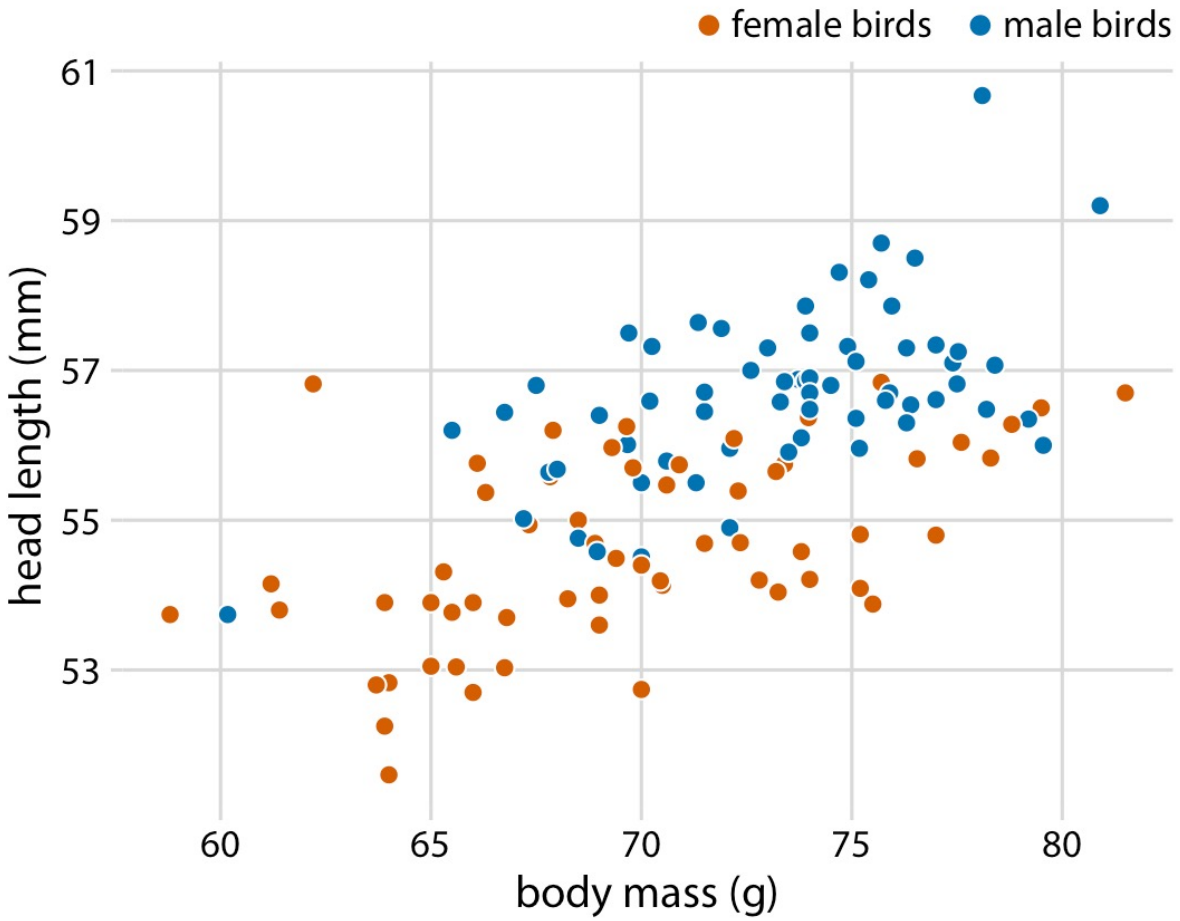
Scatterplots

- Head length (measured from the tip of the bill to the back of the head, in mm) versus body mass (in grams), for 123 blue jays.
- Each dot corresponds to one bird.
- There is a moderate tendency for heavier birds to have longer heads.



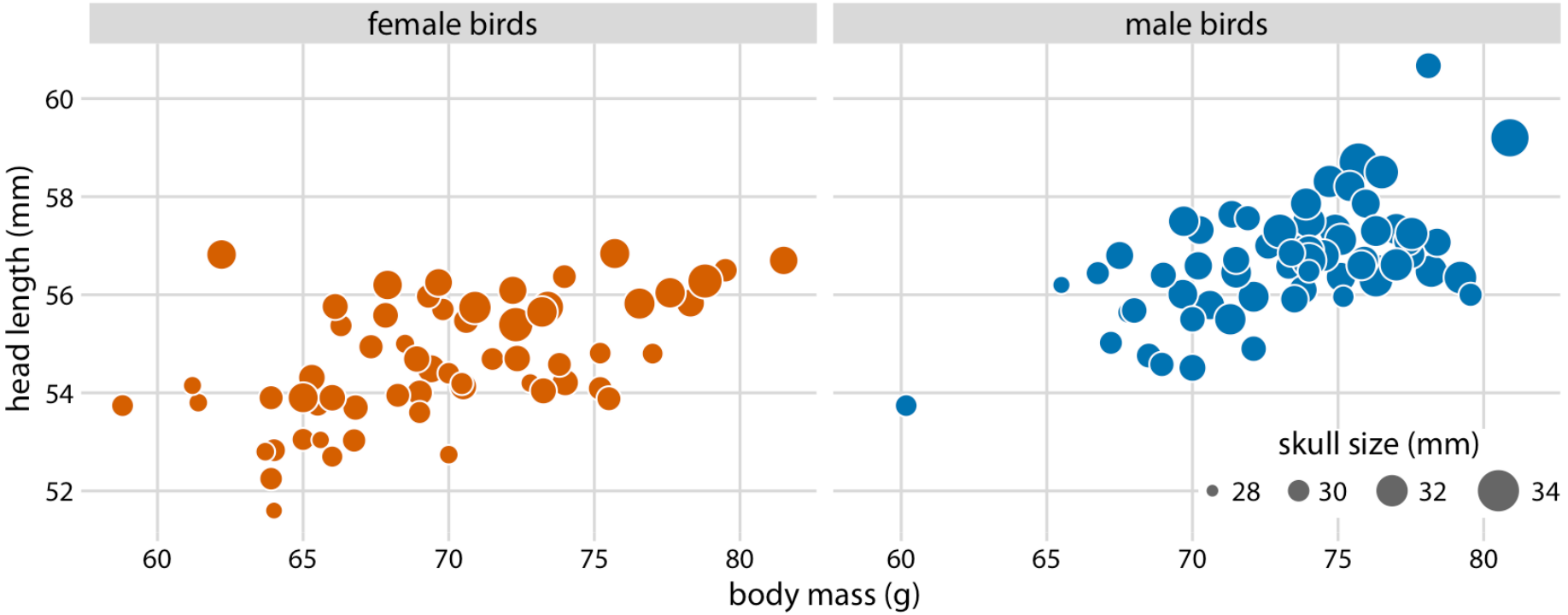
Example: Head length versus body mass

- The birds' sex is indicated by color.

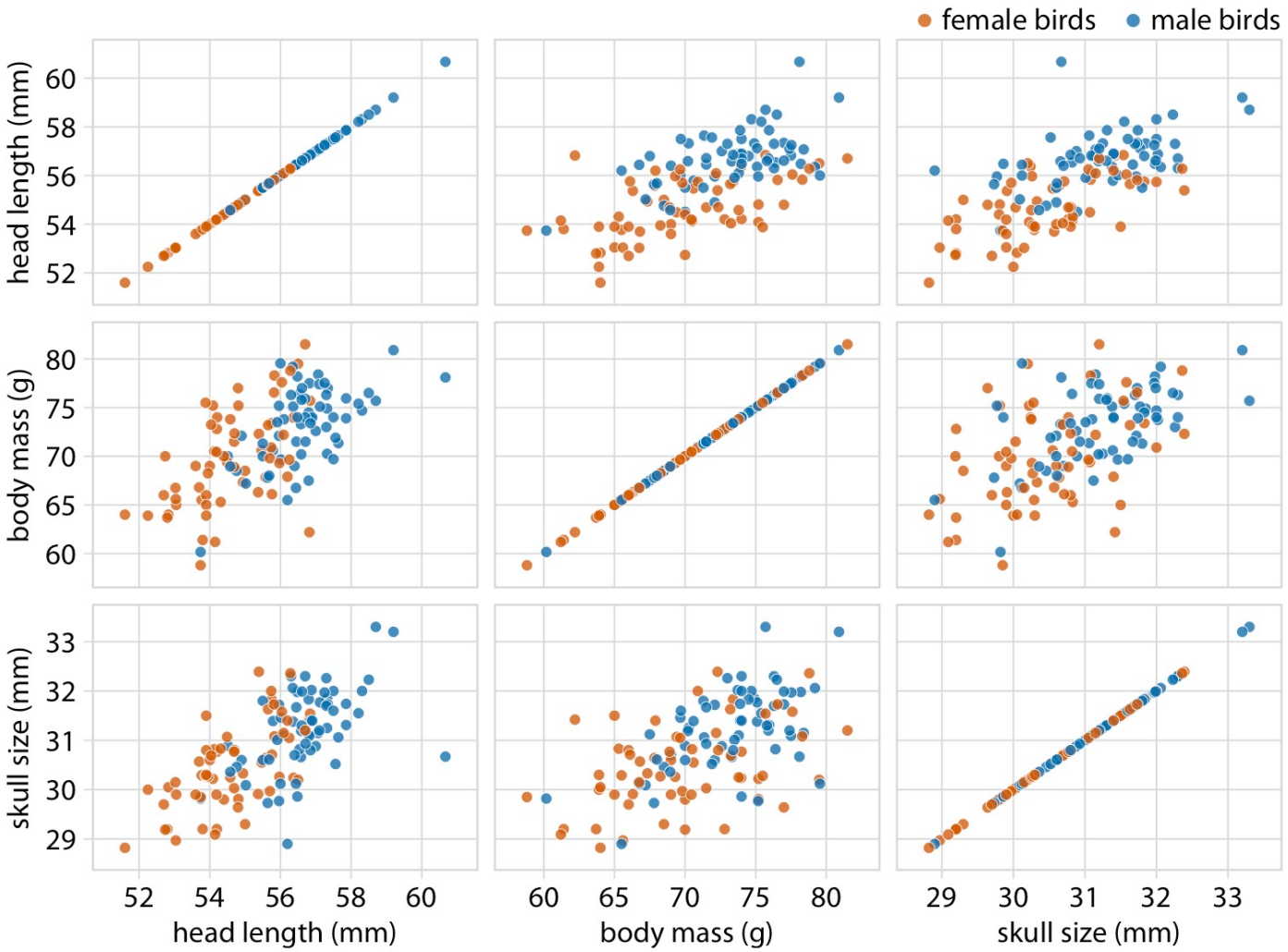


Bubble chart

- The birds' sex is indicated by color.
- The birds' skull size by symbol size.

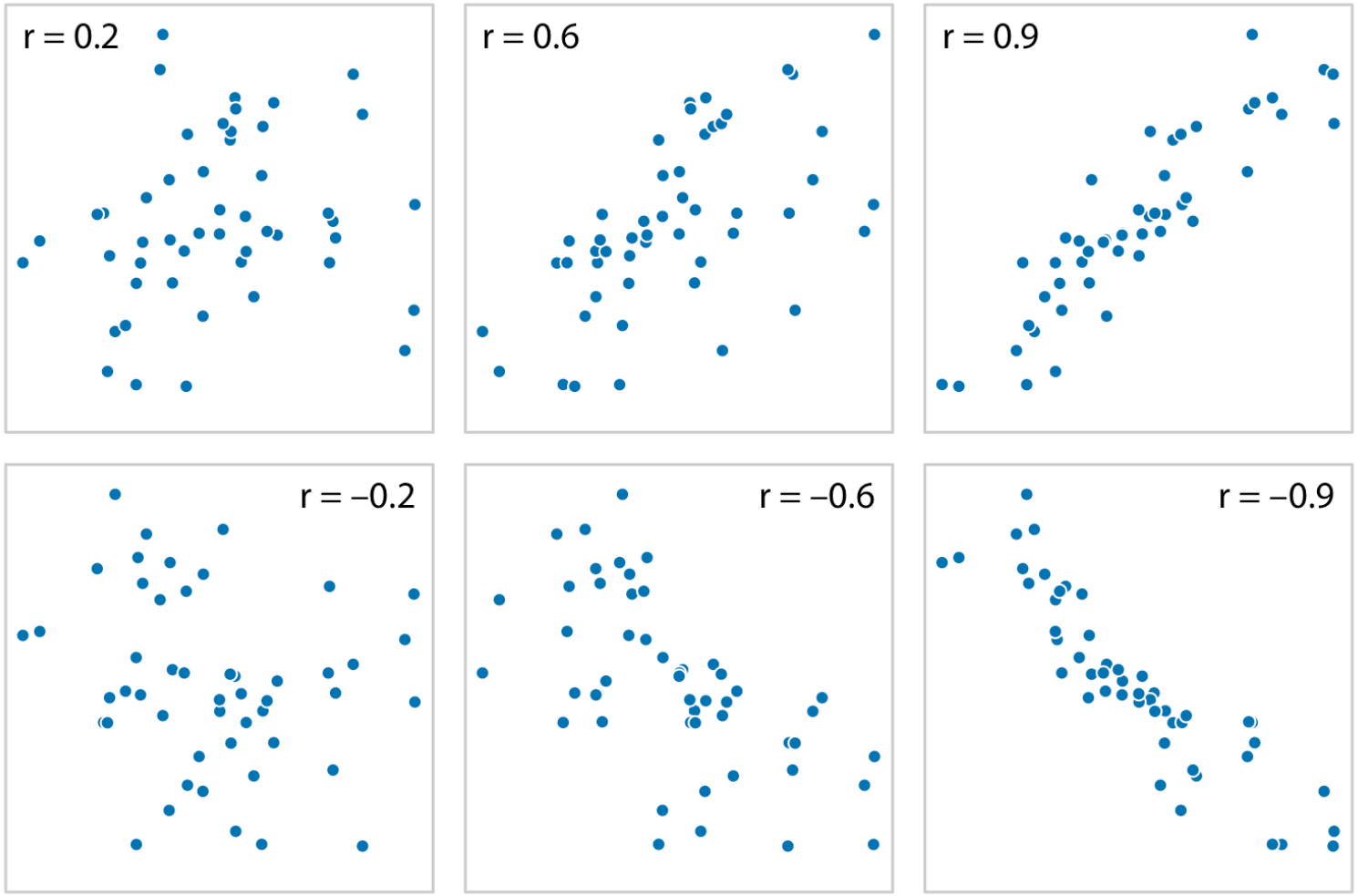


All-against-all scatterplot matrix



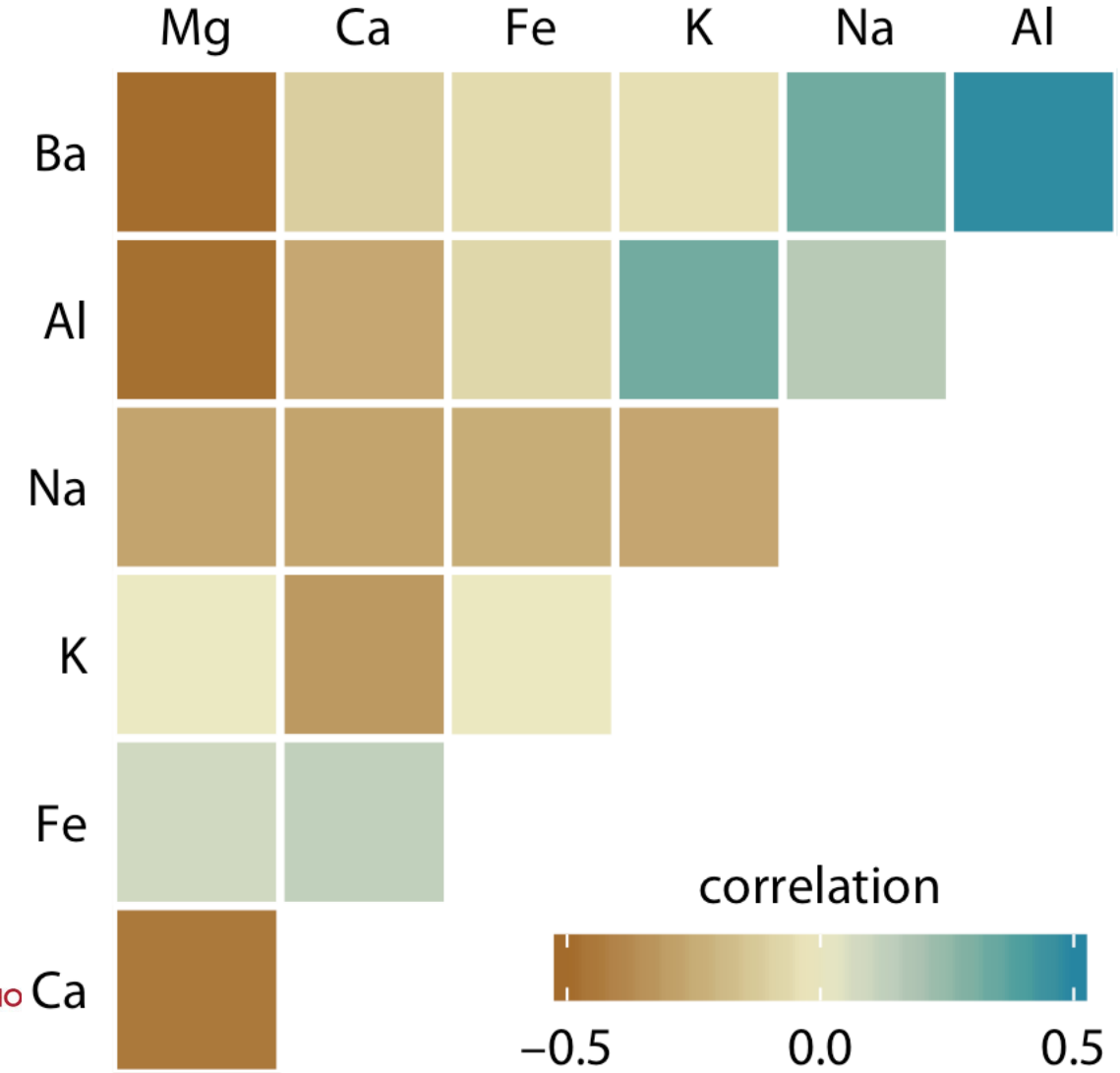
Correlation coefficient

- Examples of correlations of different magnitude and direction, with associated.
-



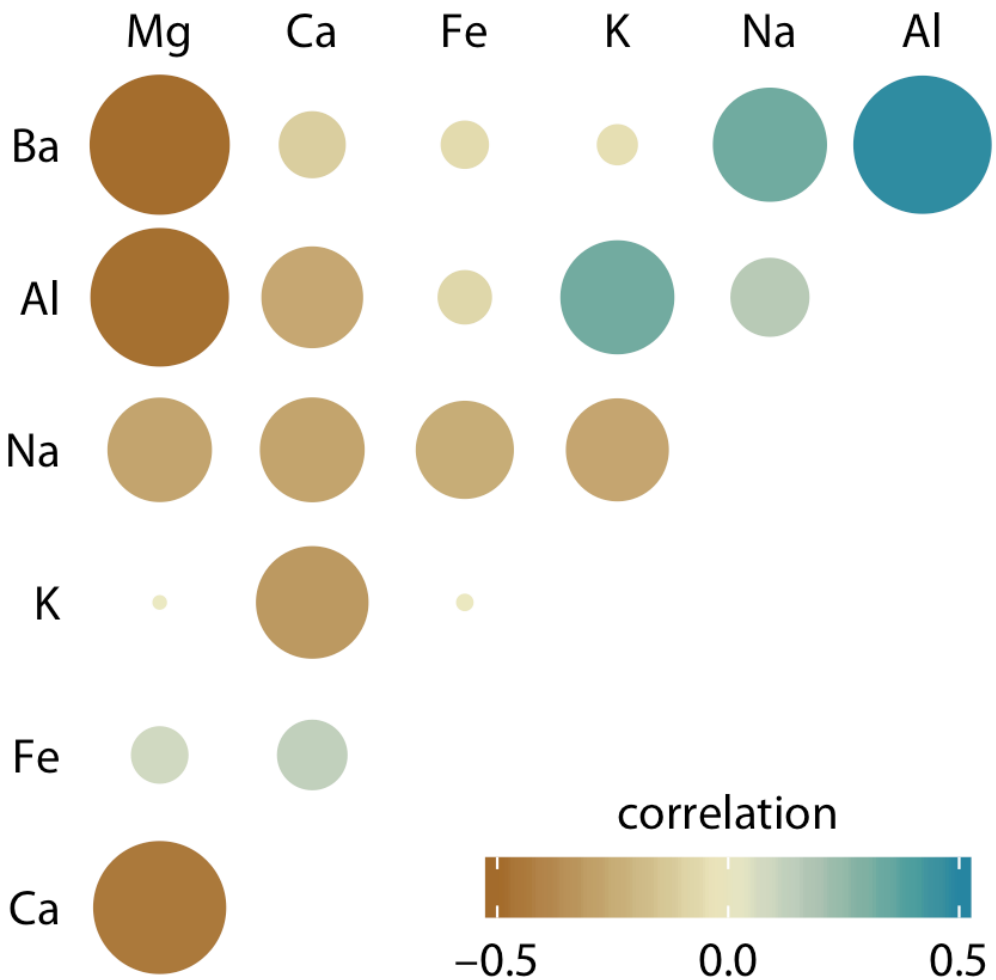
Correlograms

- Visualizations of correlation coefficients



Example: Correlations in mineral content for forensic glass samples

- The magnitude of each correlation is also encoded in the size of the colored circles.

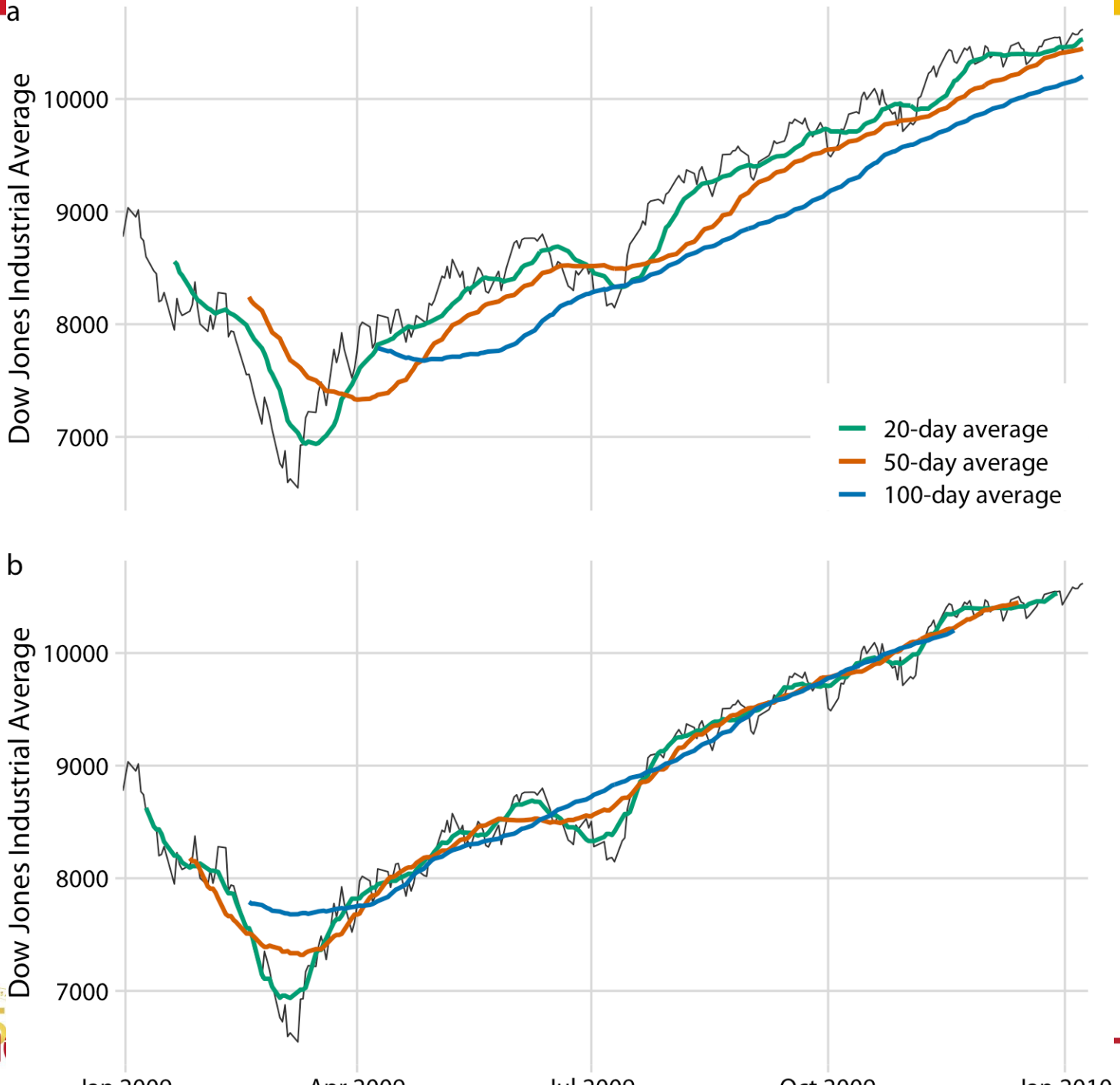


Visualizing trends

- Two fundamental approaches to determining a trend
  - Smoothing the data by some method, such as a moving average
  - Or fitting a curve with a defined functional form and then draw the fitted curve.

Smoothing

- Daily closing values of the Dow Jones Industrial Average for the year 2009, shown together with their 20-day, 50-day, and 100-day moving averages.
  - (a) The moving averages are plotted at the ends of the moving time windows.
  - (b) The moving averages are plotted in the centers of the moving time windows.

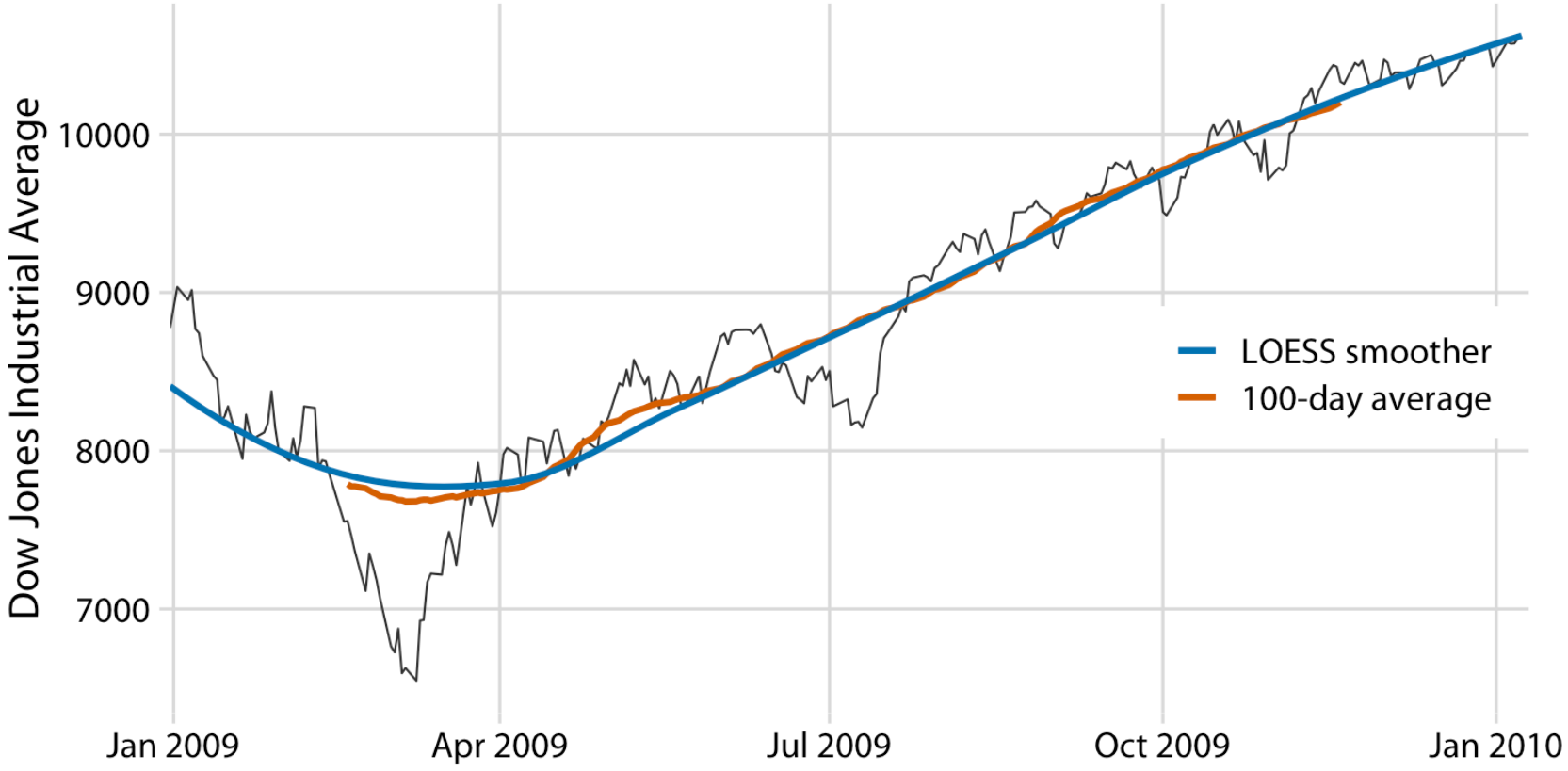

a
Dow Jones Industrial Average
10000
9000
8000
7000
20-day average
50-day average
100-day average
b
Dow Jones Industrial Average
10000
9000
8000
7000
Jan 2009
Apr 2009
Jul 2009
Oct 2009
Jan 2010

Moving average limitations

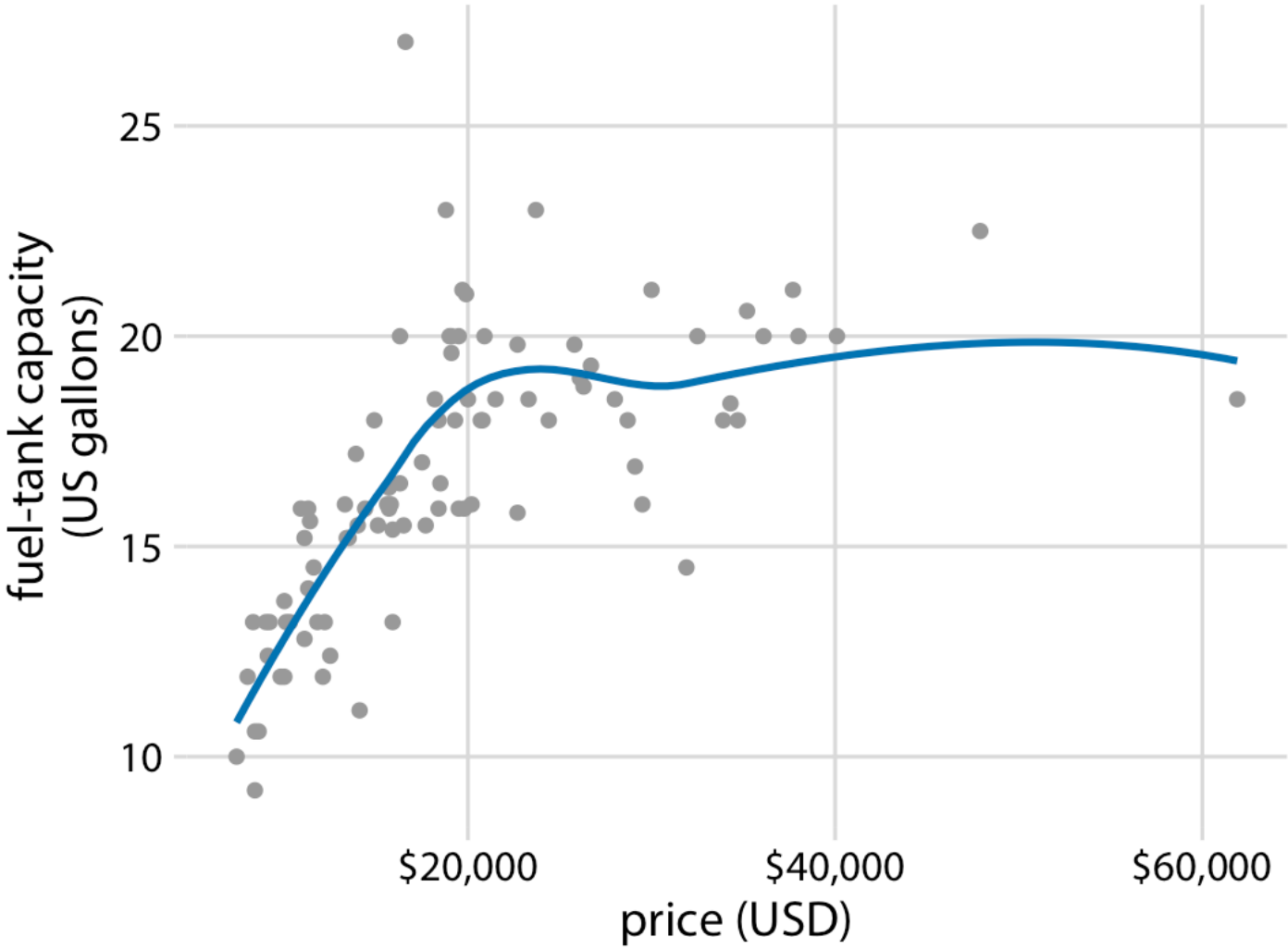
- Results in a smoothed curve that is shorter than the original curve
  - Parts are missing at either the beginning or the end or both.
- Even with a large averaging window, a moving average is not necessarily that smooth. It may exhibit small bumps and wiggles even though larger-scale smoothing.

Locally estimated scatterplot smoothing (LOESS)

- Fits low-degree polynomials to subsets of the data.



Example: Fuel-tank capacity versus price of 93 cars released for the 1993 model year

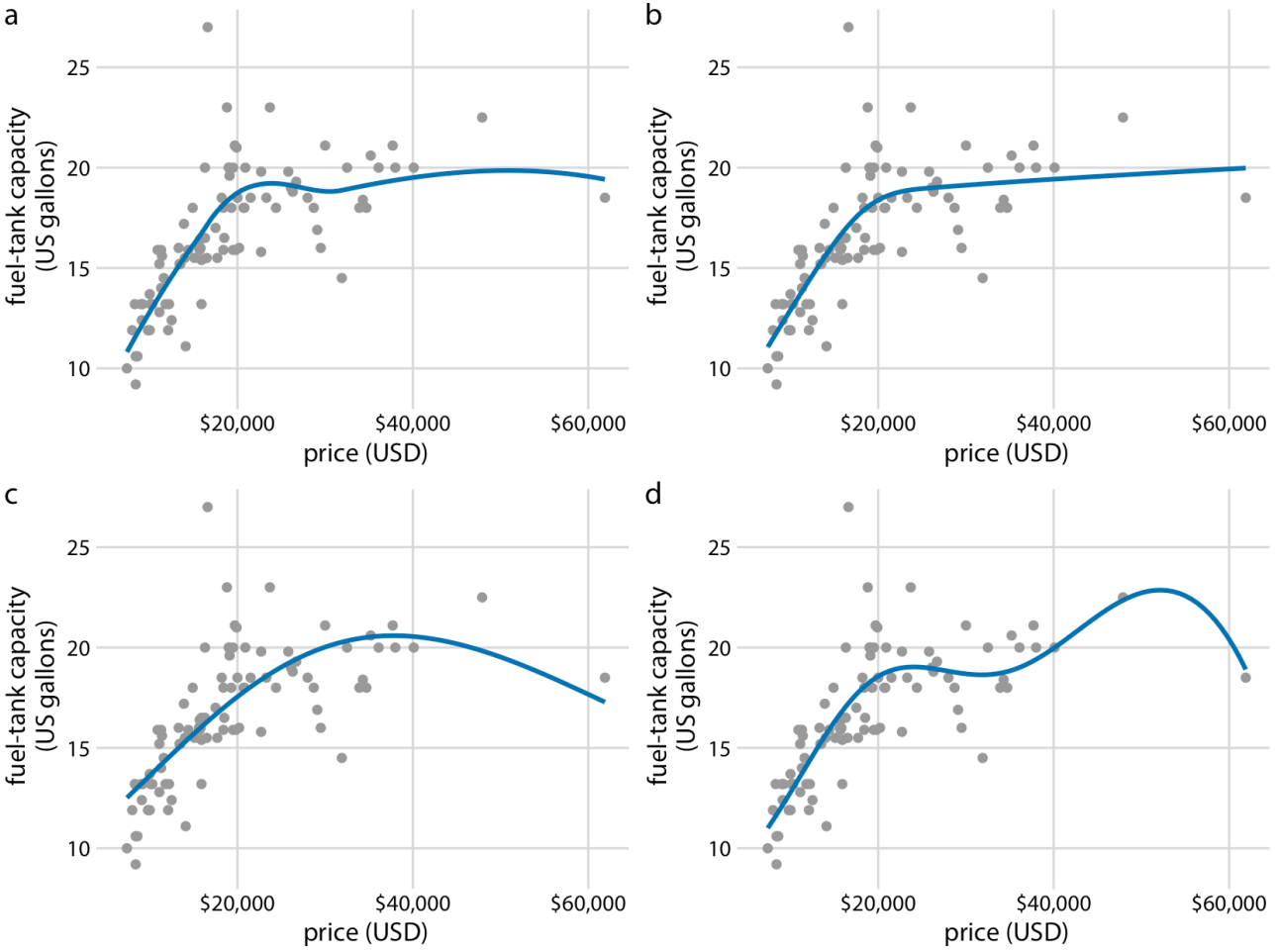


Spline models

- LOESS requires fitting of many separate regression models
- Spline models
  - A faster alternative to LOESS
  - A spline is a piecewise polynomial function
  - Fit a spline with k segments, we need to specify k + 1 knots.

Example: Different smoothing models

- Display widely different behaviors, near the boundaries of the data.
  - (a) LOESS smoother, as in Figure 14-4.
  - (b) Cubic regression splines with 5 knots.
  - (c) Thin-plate regression spline with 3 knots.
  - (d) Gaussian process spline with 6 knots.



Visualizing uncertainty

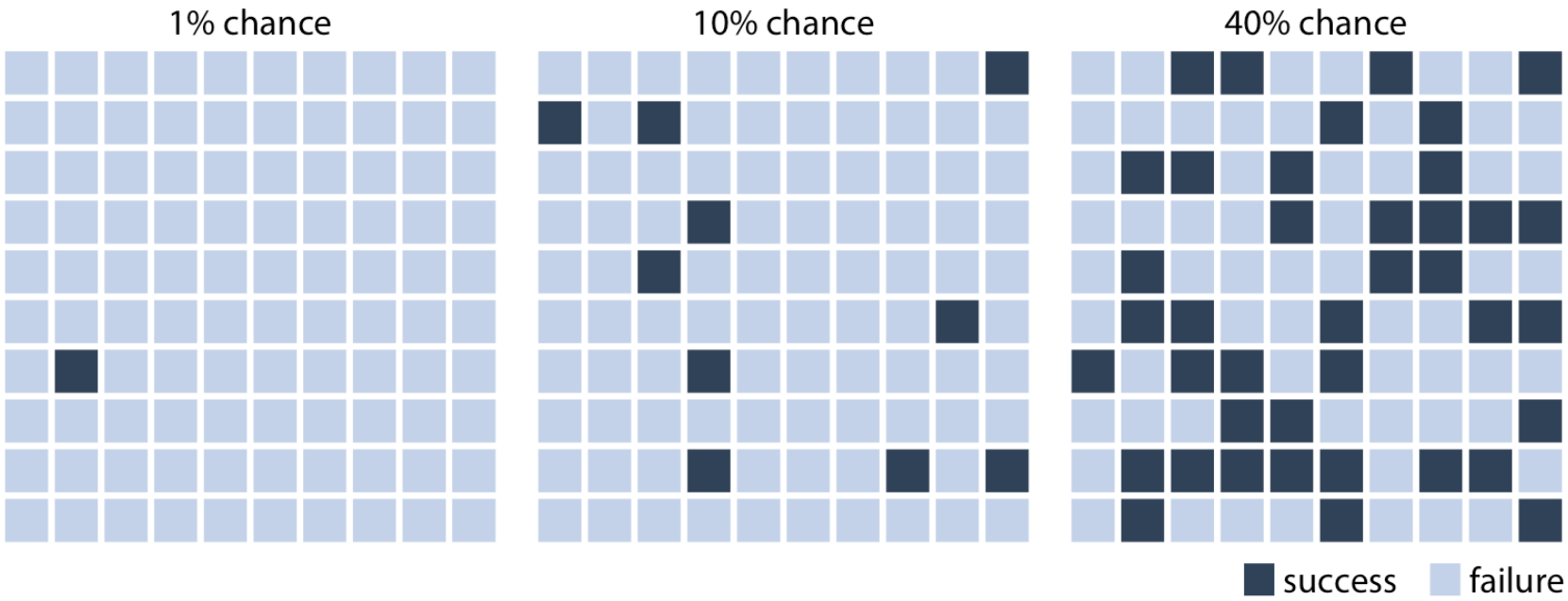
Scenarios

- One of the most challenging aspects of data visualization is the visualization of uncertainty.
- Nearly every dataset has some uncertainty.
  - Whether and how we choose to represent this uncertainty can make a major difference in how accurately our audience perceives the meaning of the data.
- Common approaches
  - Error bars
  - Confidence bands

Examples of uncertainty

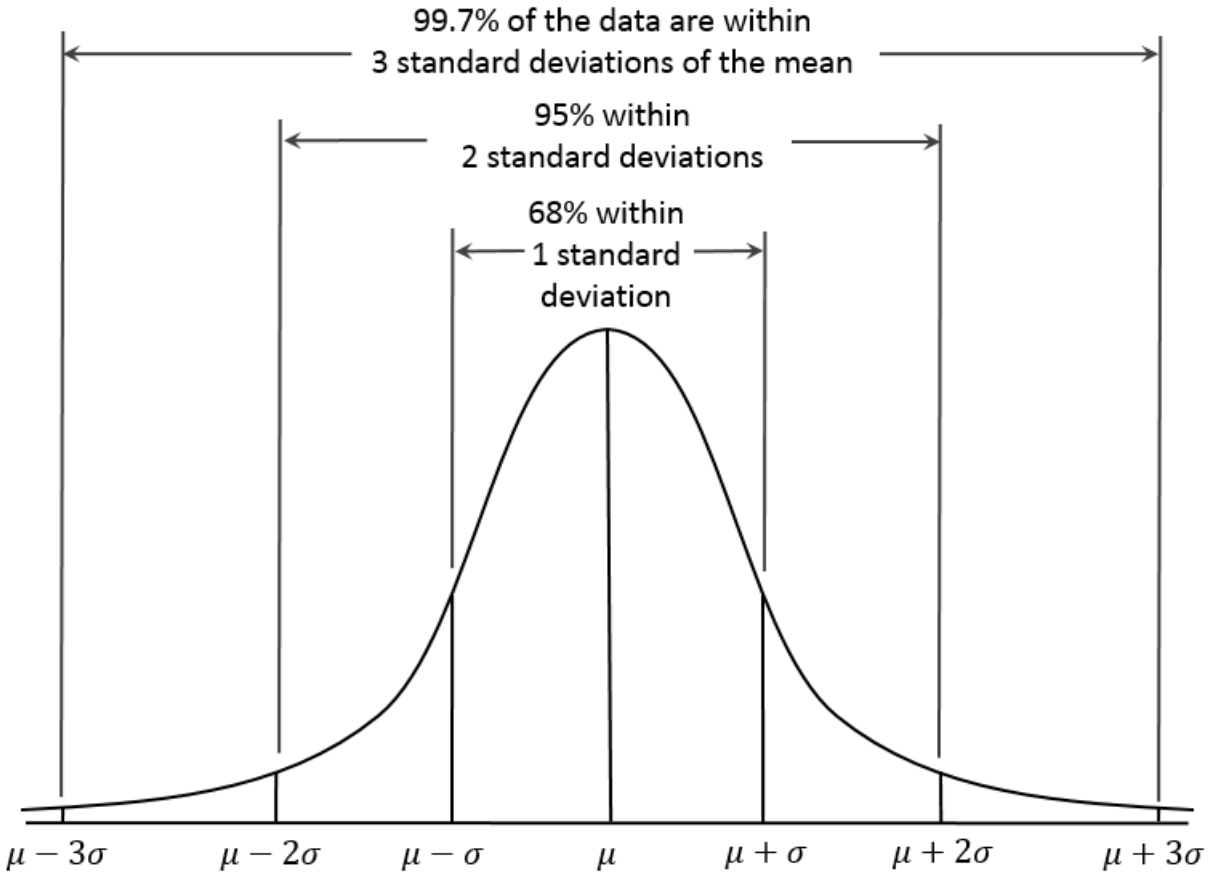
- In the context of future events, the eventual outcome is uncertain
- An event in the past can also be uncertain
  - A red car parked across the street at 8 a.m. but not at 4 p.m., then we can conclude the car left at some point during the 8-hour window, but we don't know exactly when.
- Mathematically, we deal with uncertainty by employing the concept of probability.

Visualizing probability as frequency

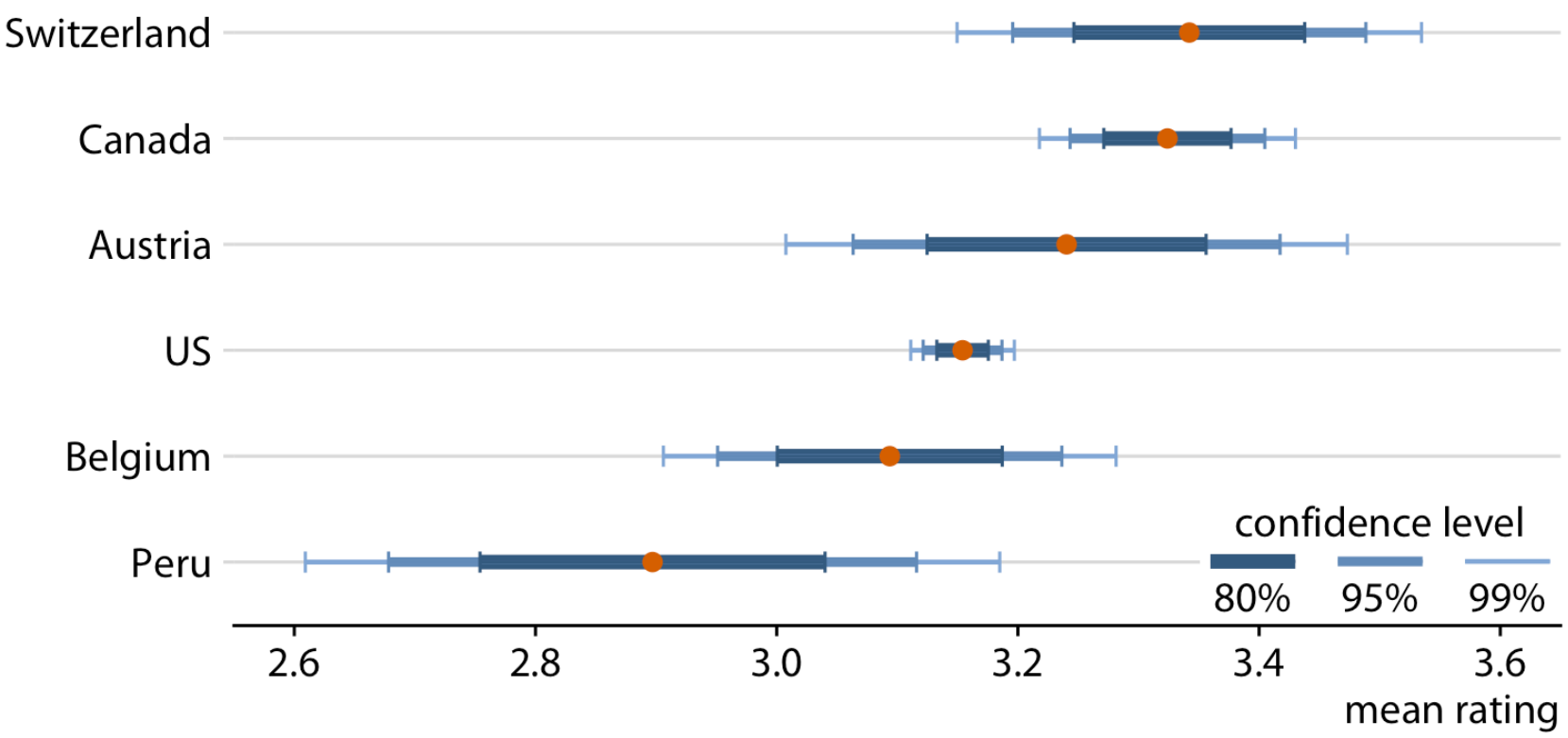


Probability distribution

- Confidence levels are expressed as a percentage and indicate how frequently that percentage of the target population would give an answer that lies within the confidence interval.

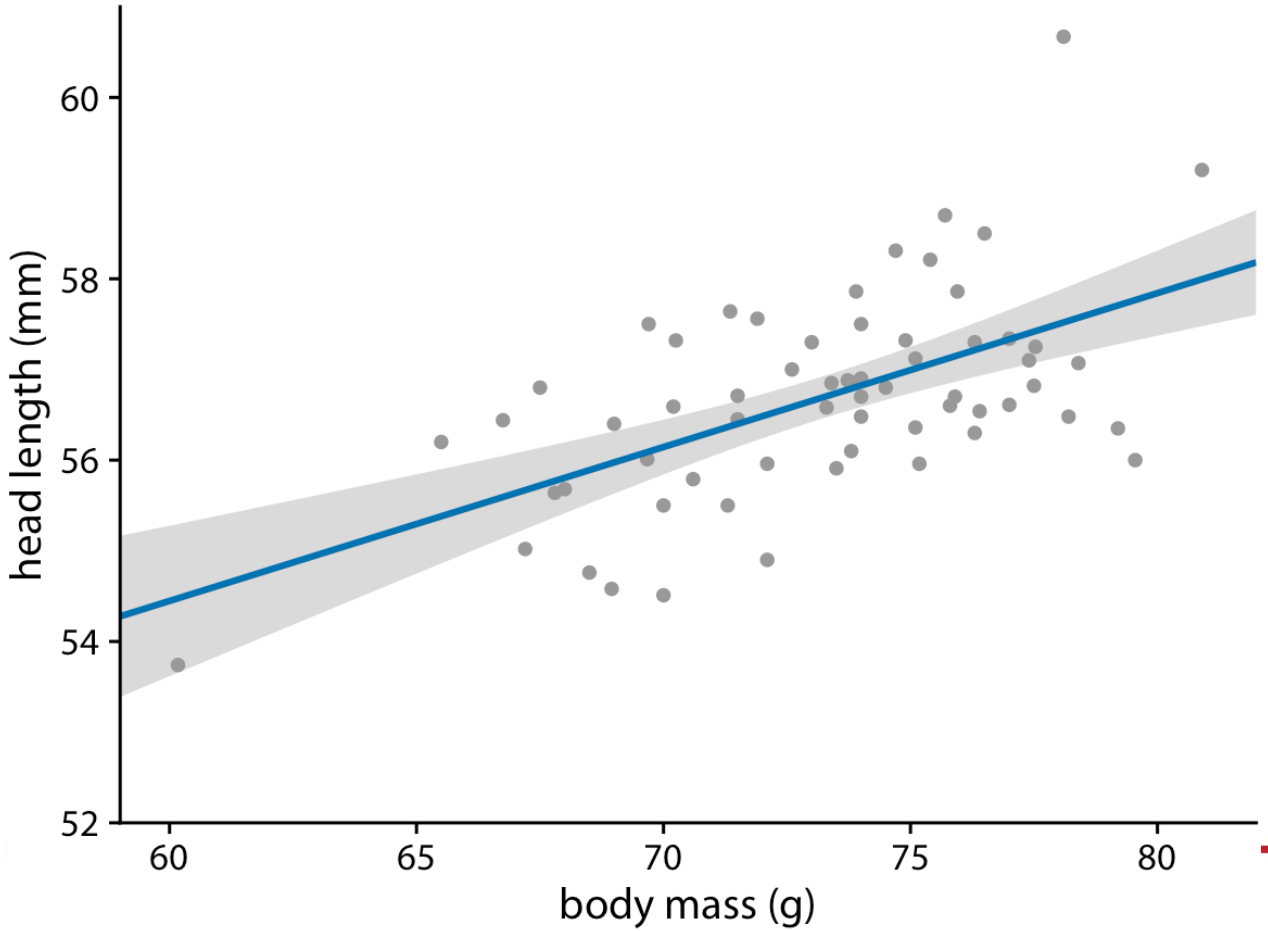


Example: Mean chocolate flavor ratings and associated confidence intervals



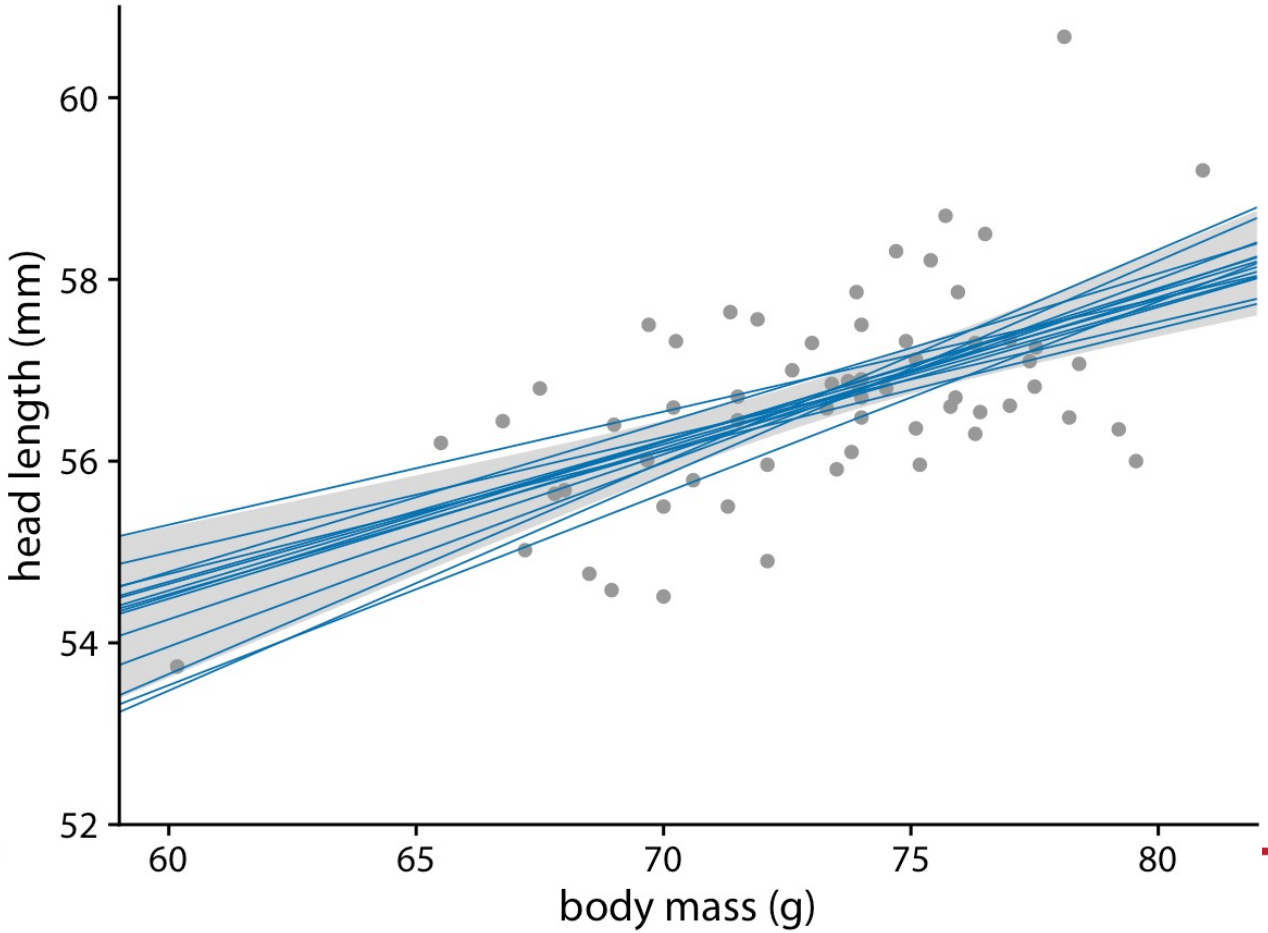
Visualizing the uncertainty of curve fits

- Show the uncertainty in a trend line with a confidence band
  - Range of different fit lines that would be compatible with the data
- Example: Head length versus body mass for male blue jays



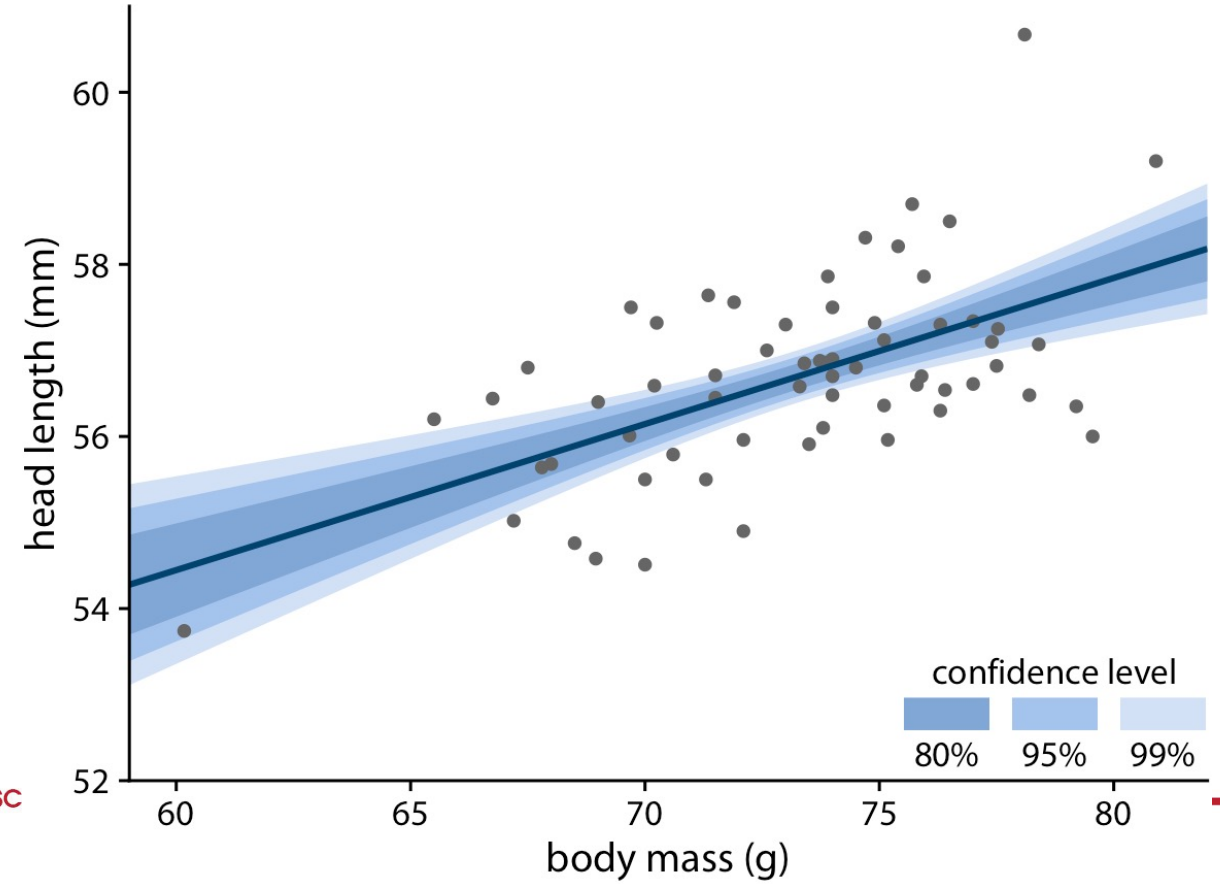
Example: Head length versus body mass for male blue jays

- The straight blue lines now represent equally likely alternative fits randomly drawn from the posterior distribution



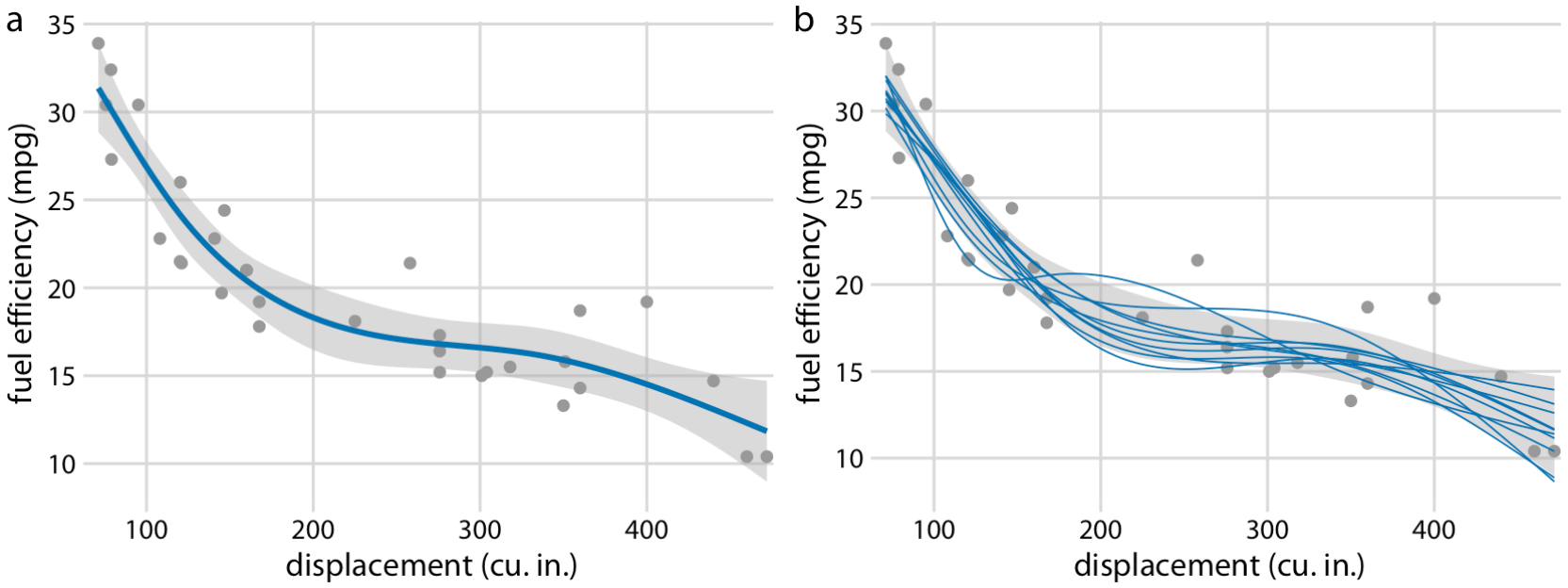
Graded confidence band

- Shows several confidence levels at once
  - Enhances the sense of uncertainty in the reader.
  - Forces the reader to confront the possibility that the data might support different alternative trend lines.



Confidence bands for nonlinear curve fits

- The confidence band represents a family of curves that are all quite a bit wigglier than the overall best fit shown in part (a).



VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attention!!!