Data Visualization

INF552 (2023-2024)

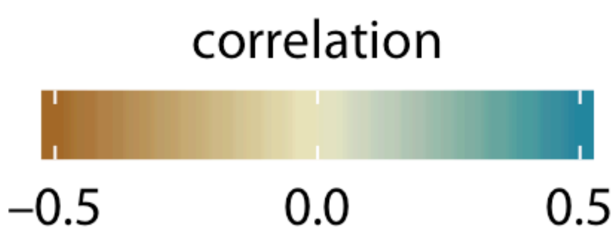
Session 04
Multi-variate Data Visualisation (Part II)

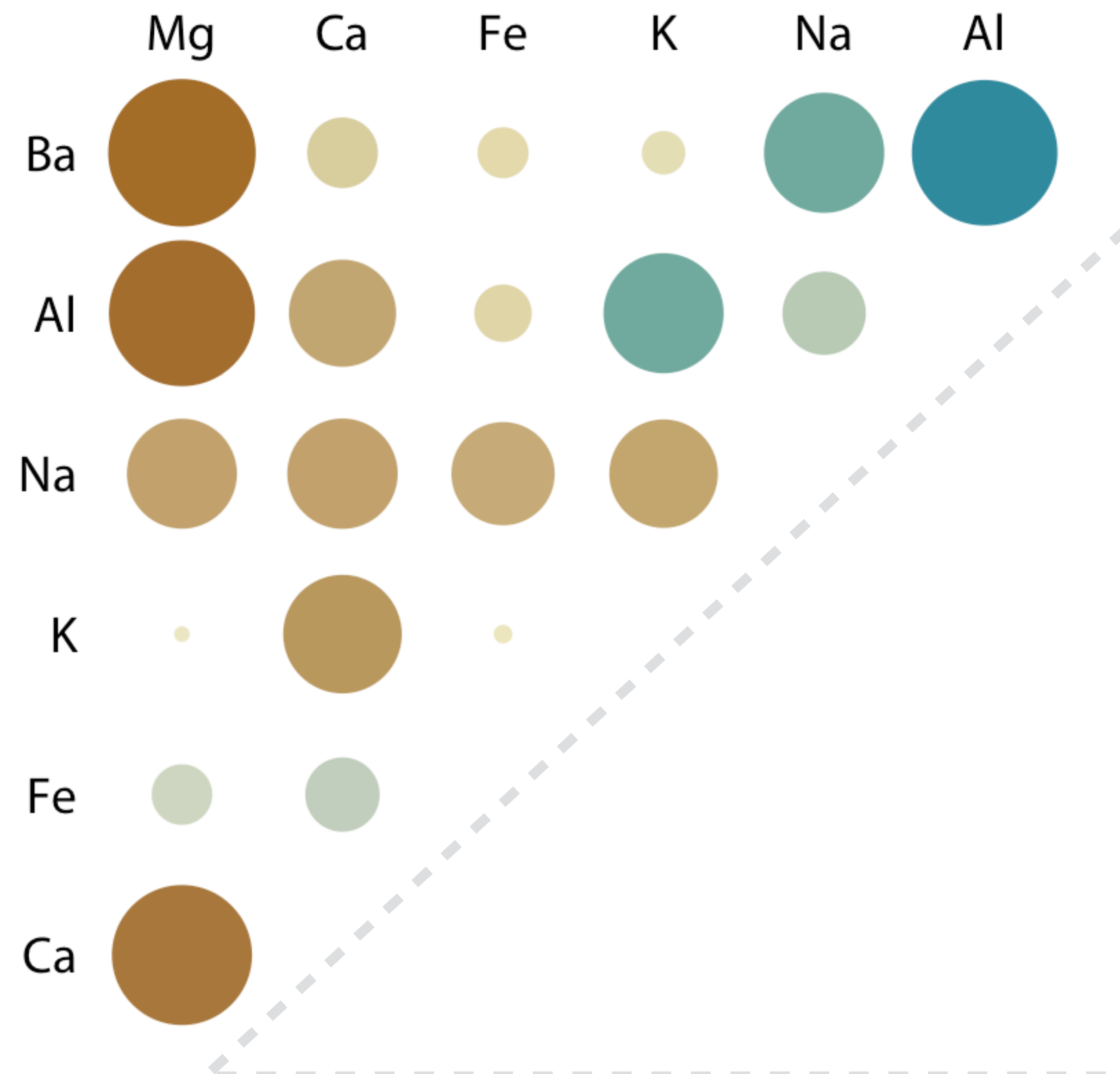
emmanuel.pietriga@inria.fr

Correlograms

When interested in identifying correlations, if there are too many variables, it can be more efficient to compute those coefficients and visualize them directly:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$





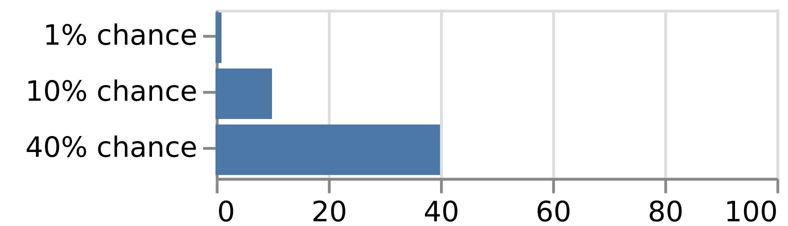
Symmetric formula in x_i and y_i

This is a relatively abstract representation though…

[Example source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

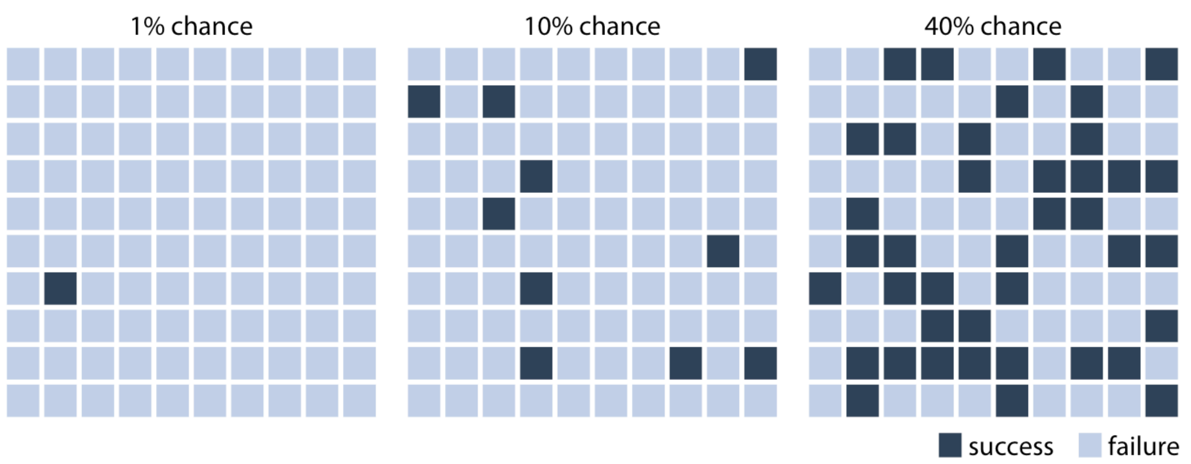
Discrete outcome visualization / Frequency framing

Showing the probability as a single number:



vs.

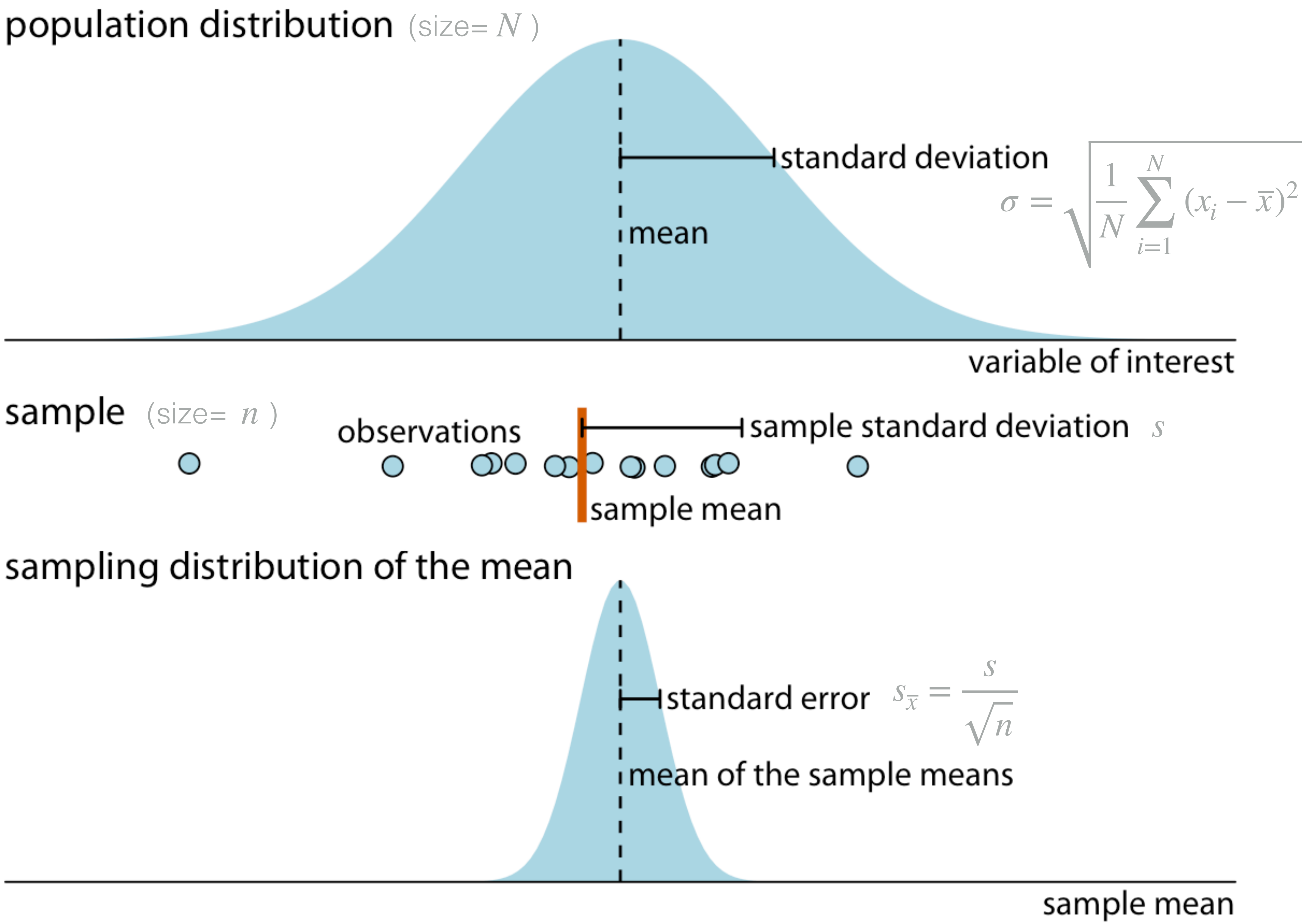
Showing discrete outcomes:



Key idea: emphasize the frequency aspect and the unpredictability of a random trial

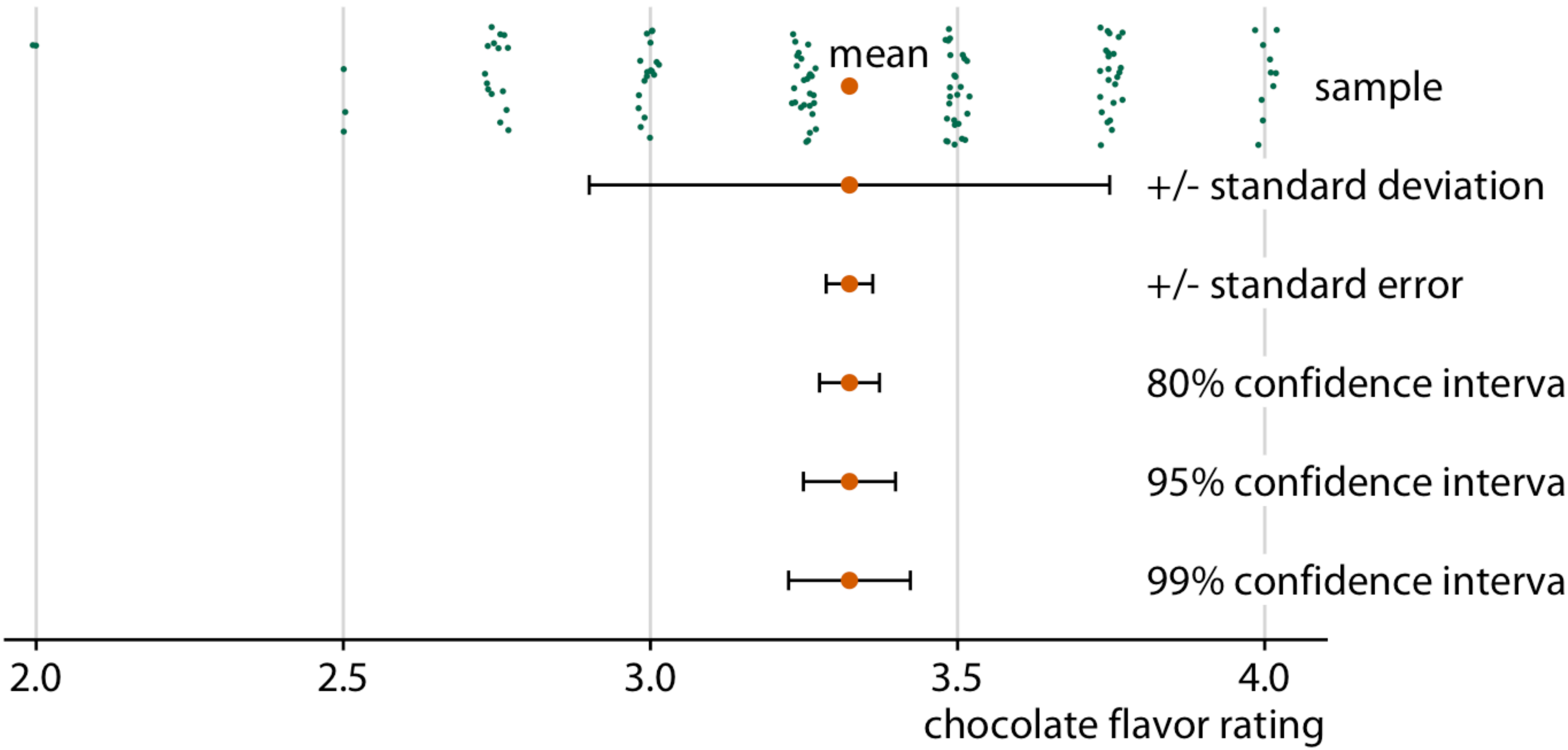
[Image source (right): Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

Reminder - Descriptive Statistics



[Image source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

Error Bars



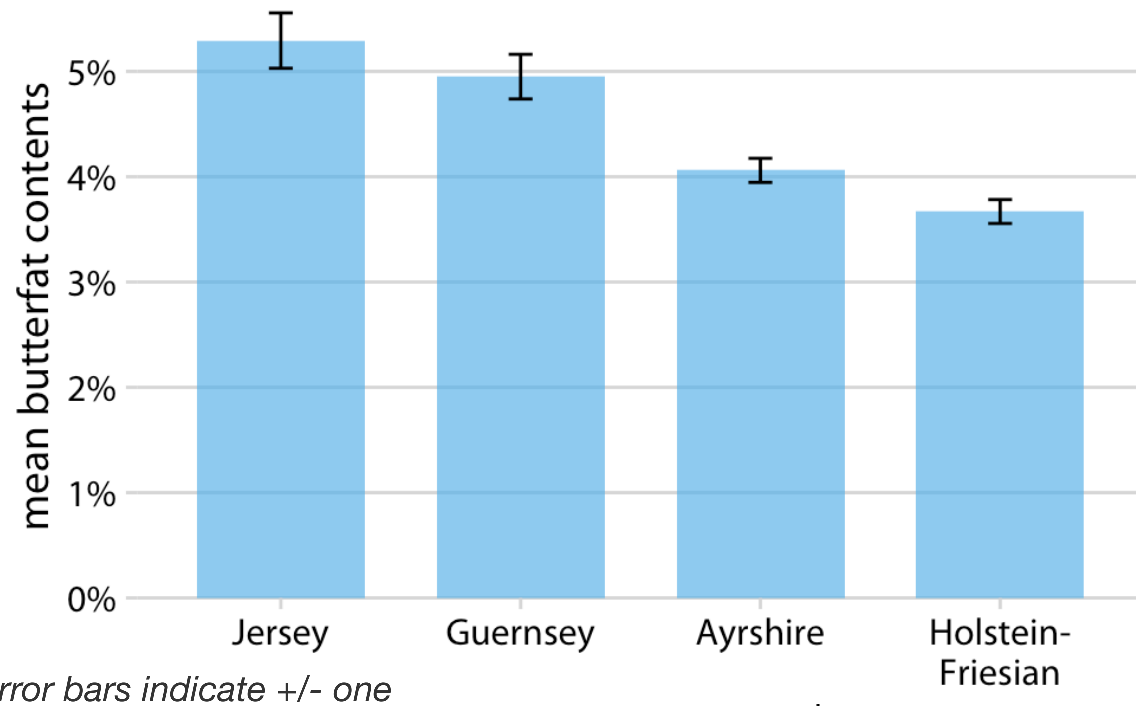
Always indicate what the error bar represents (quantity, confidence level)

[Image source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

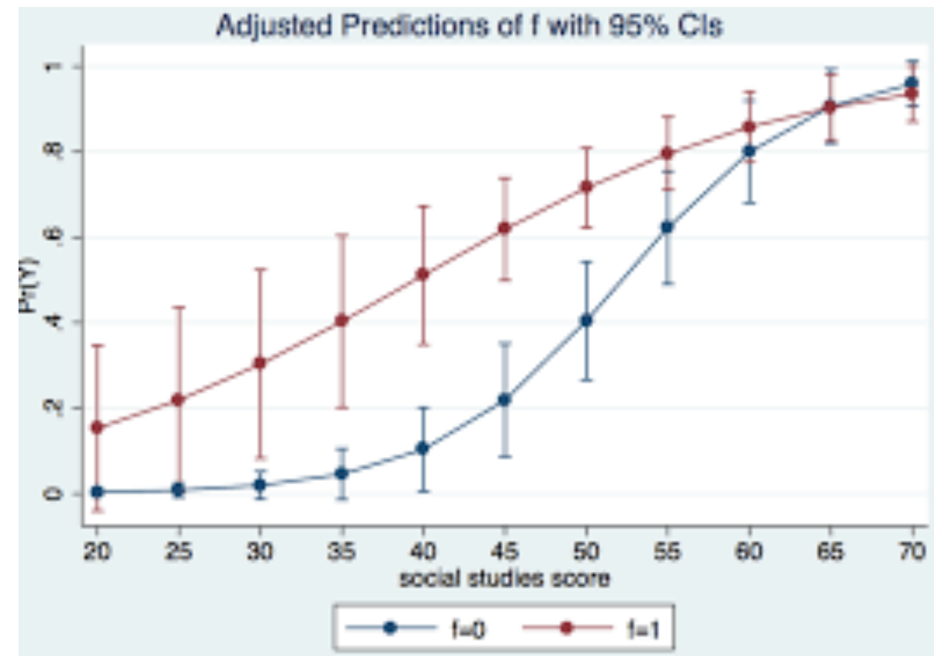
Advantages of error bars

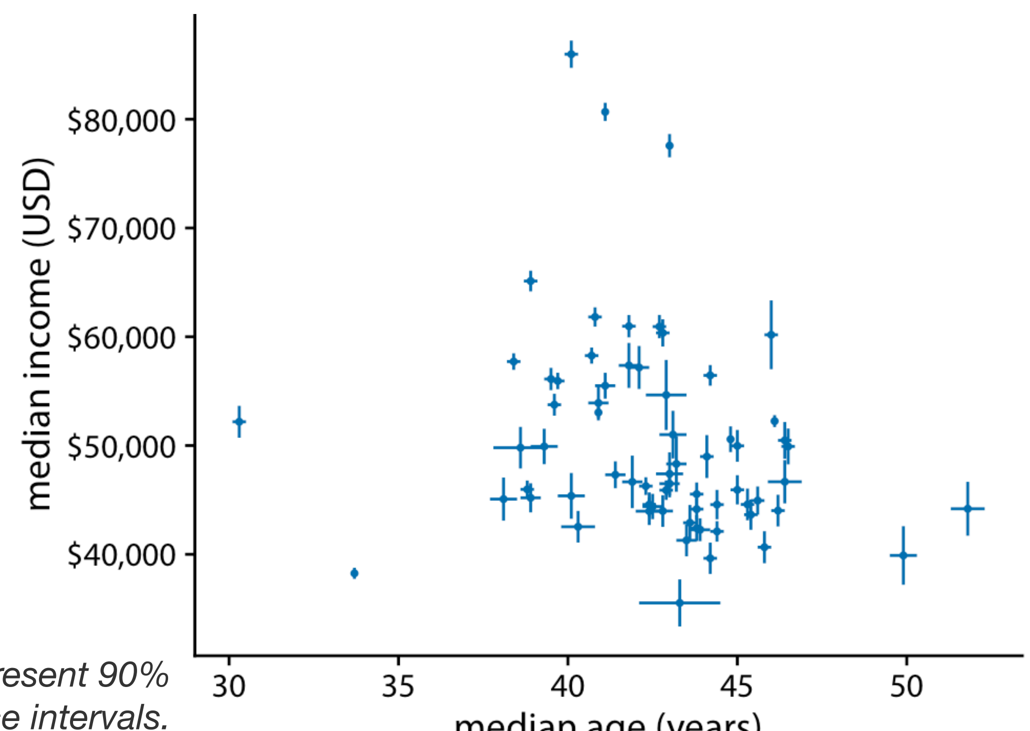
Small visual footprint (usually cause little visual interference)

Apply to other types of plots than bar charts:



Error bars indicate +/- one standard error of the mean.





Error bars represent 90% confidence intervals.

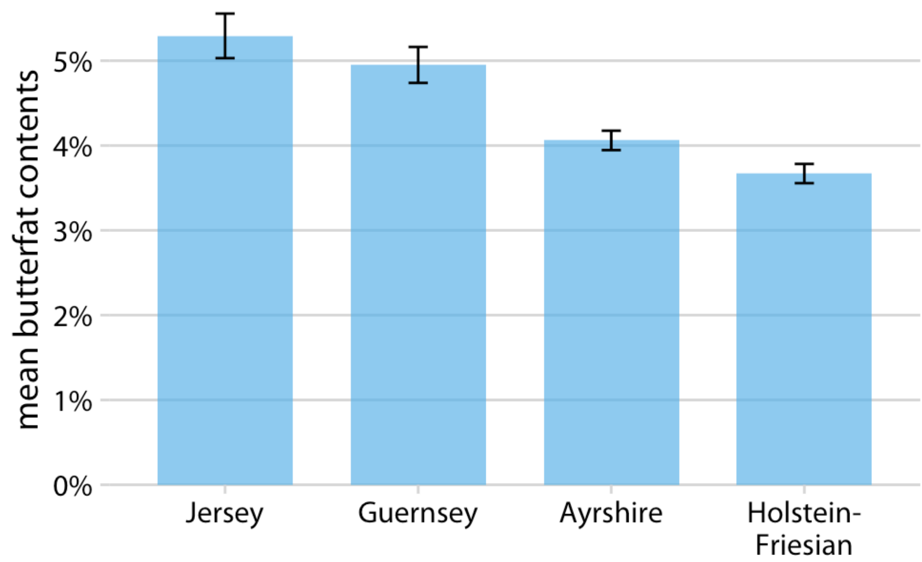
[Image source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

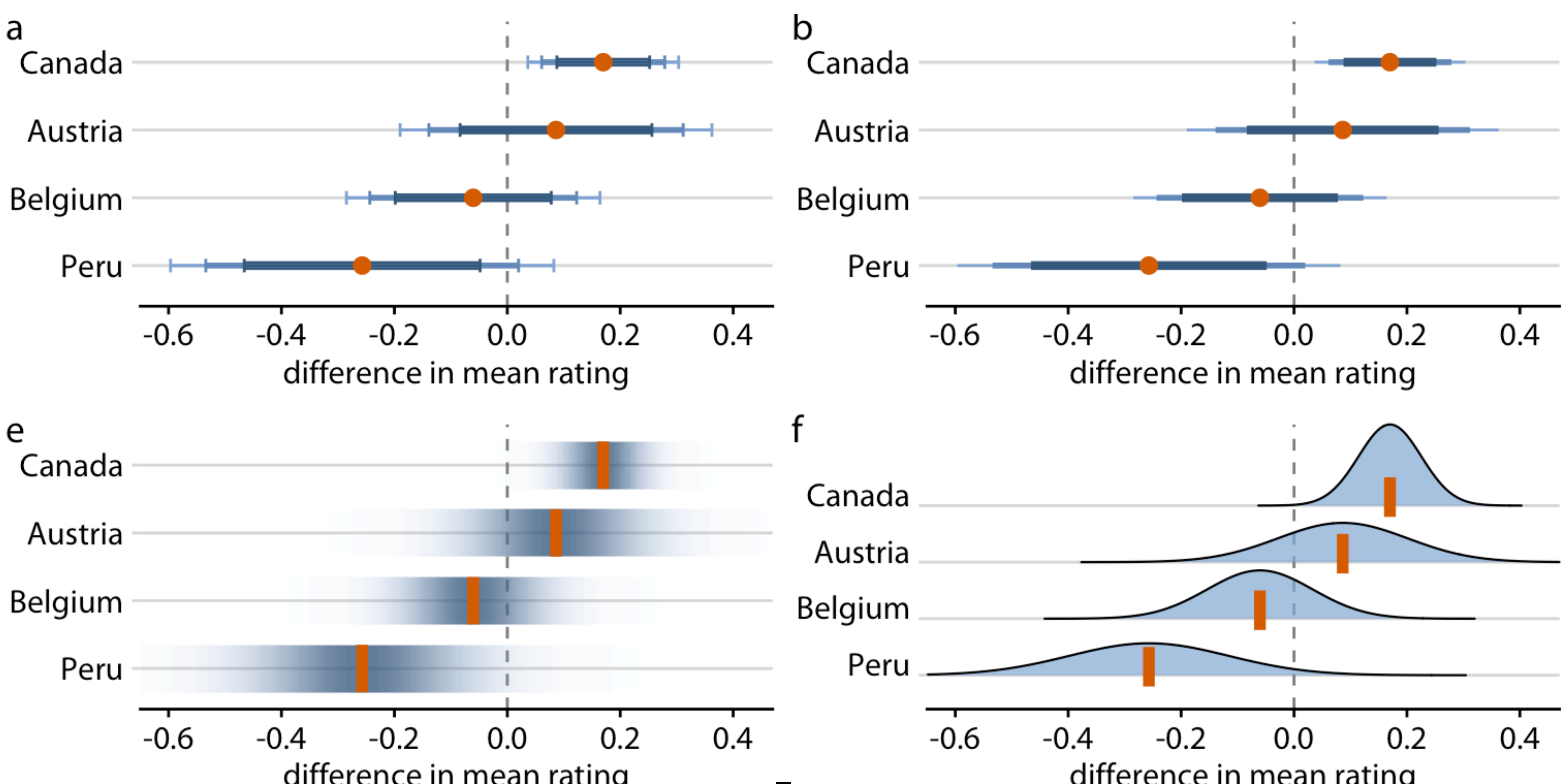
Issues with error bars: deterministic construal errors

With simple error bars, there is a risk that people perceive them deterministically:

- as representing min/max values;
- as precisely delineating the range of parameter estimates.

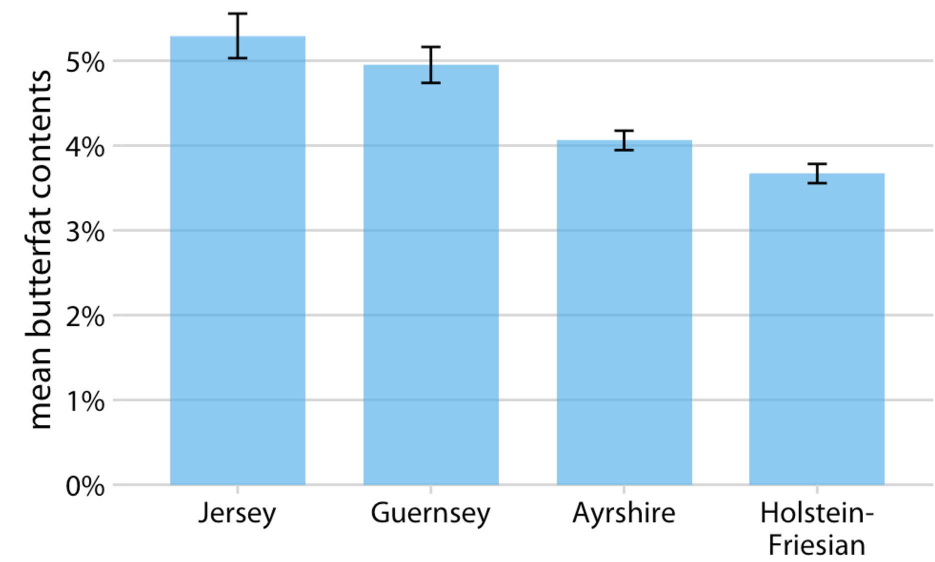
Visually convey the notion of uncertainty with, *e.g.*, graded error bars or confidence strips.





[Image source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

Issues with error bars: can be misleading



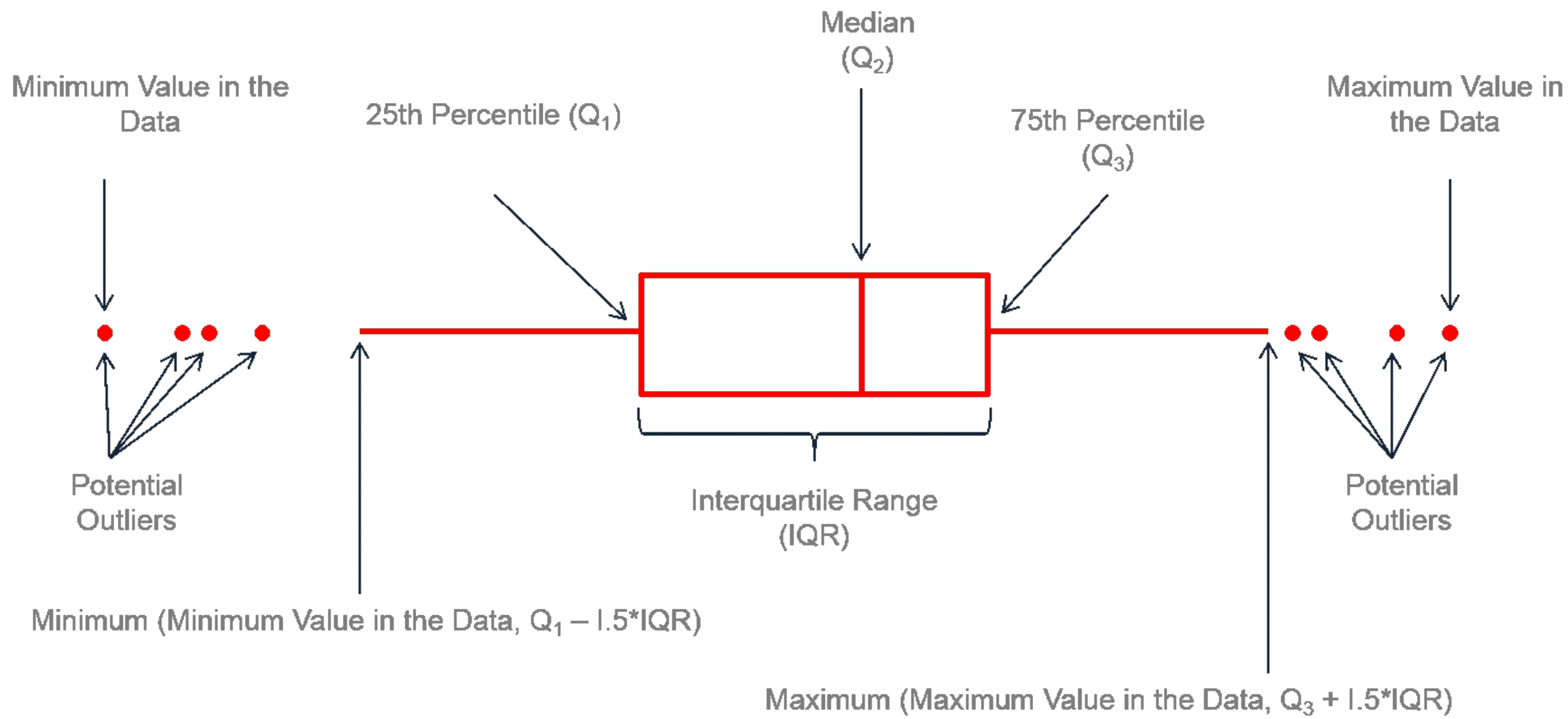
Symmetric error bars are misleading if the data are skewed.

Absence of overlap of two error bars does not necessarily entail that the difference is significant at a given confidence level: *plot confidence intervals of the differences themselves*.

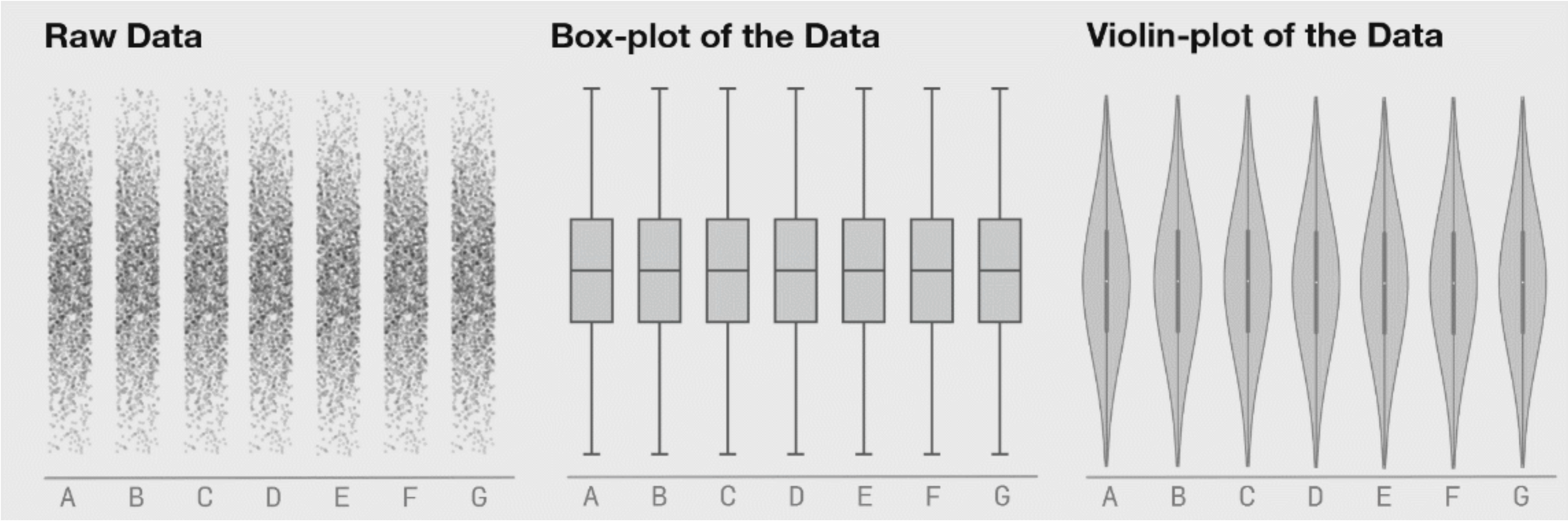
Interpretation of error bars representing confidence intervals (taking 95% as an example confidence level):

- ~~the bar delineates the interval that has 95% chance of containing the true value~~
- percentage of confidence intervals that would include the true value if an infinite number of random samples (same size) were pulled from the data and each time a 95% confidence interval was constructed.

Providing further information about the distribution: Box Plots



Provides more information, but still limited (missing sample size, distribution, …)

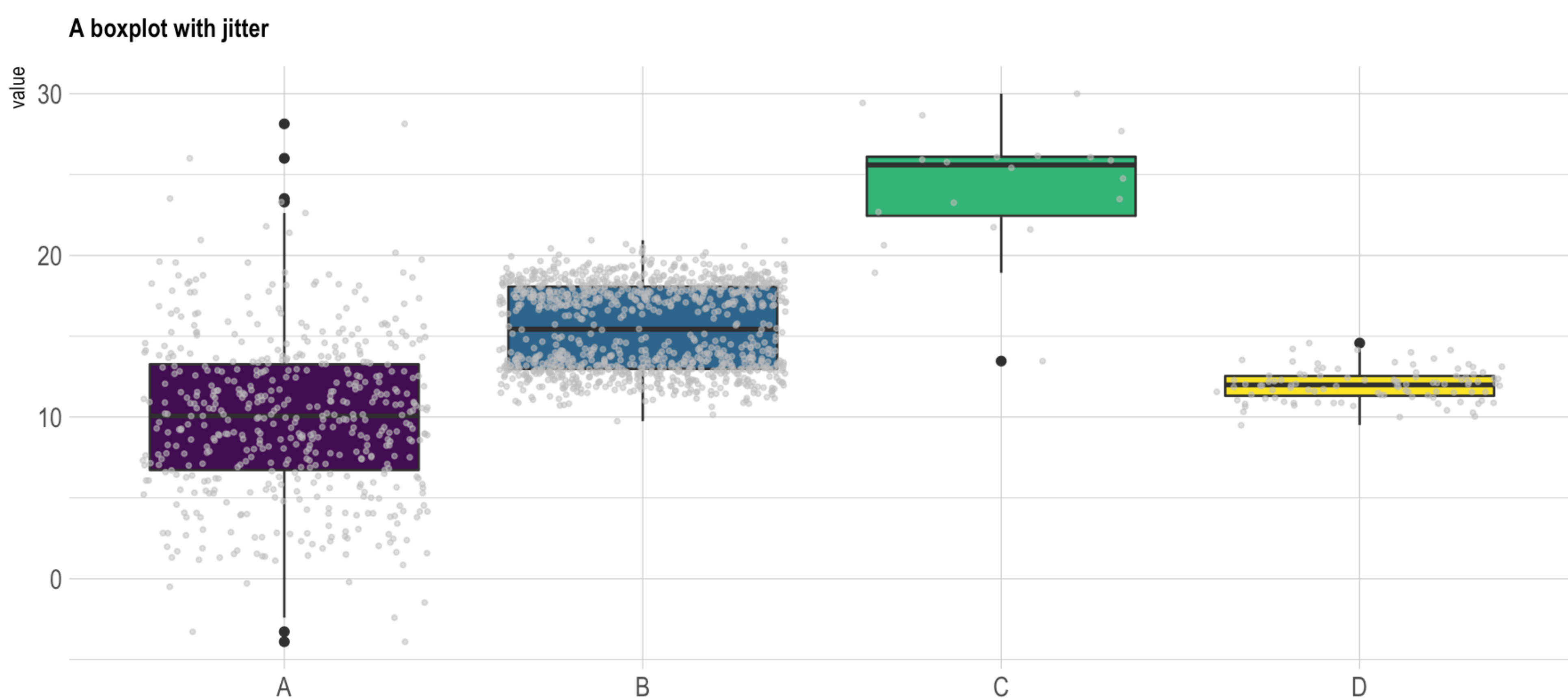


[Matejka et al. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). ACM, 1290-1294.]

Better: show the data *and* summary statistics

If not dealing with too many data points, showing them can be valuable.

An example using jitter:

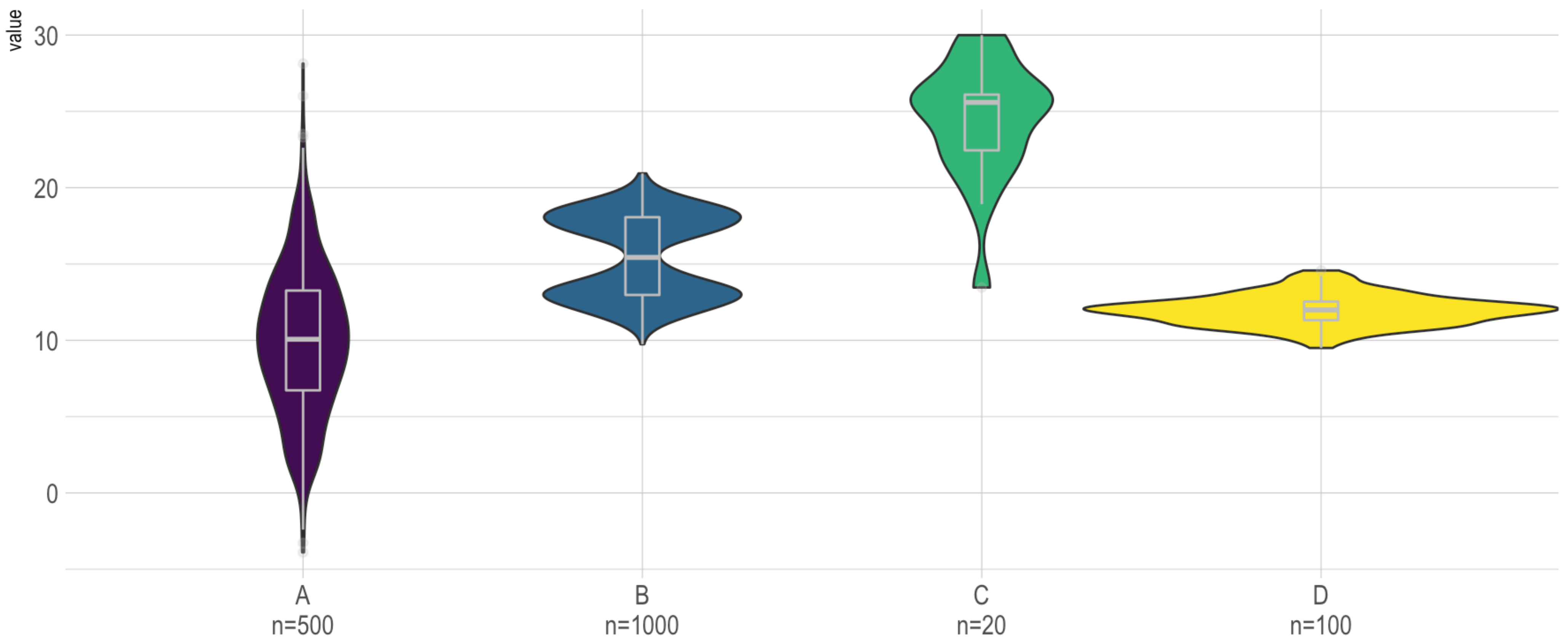


Shows that C's sample size is much smaller, that B seems to have a bimodal distribution.

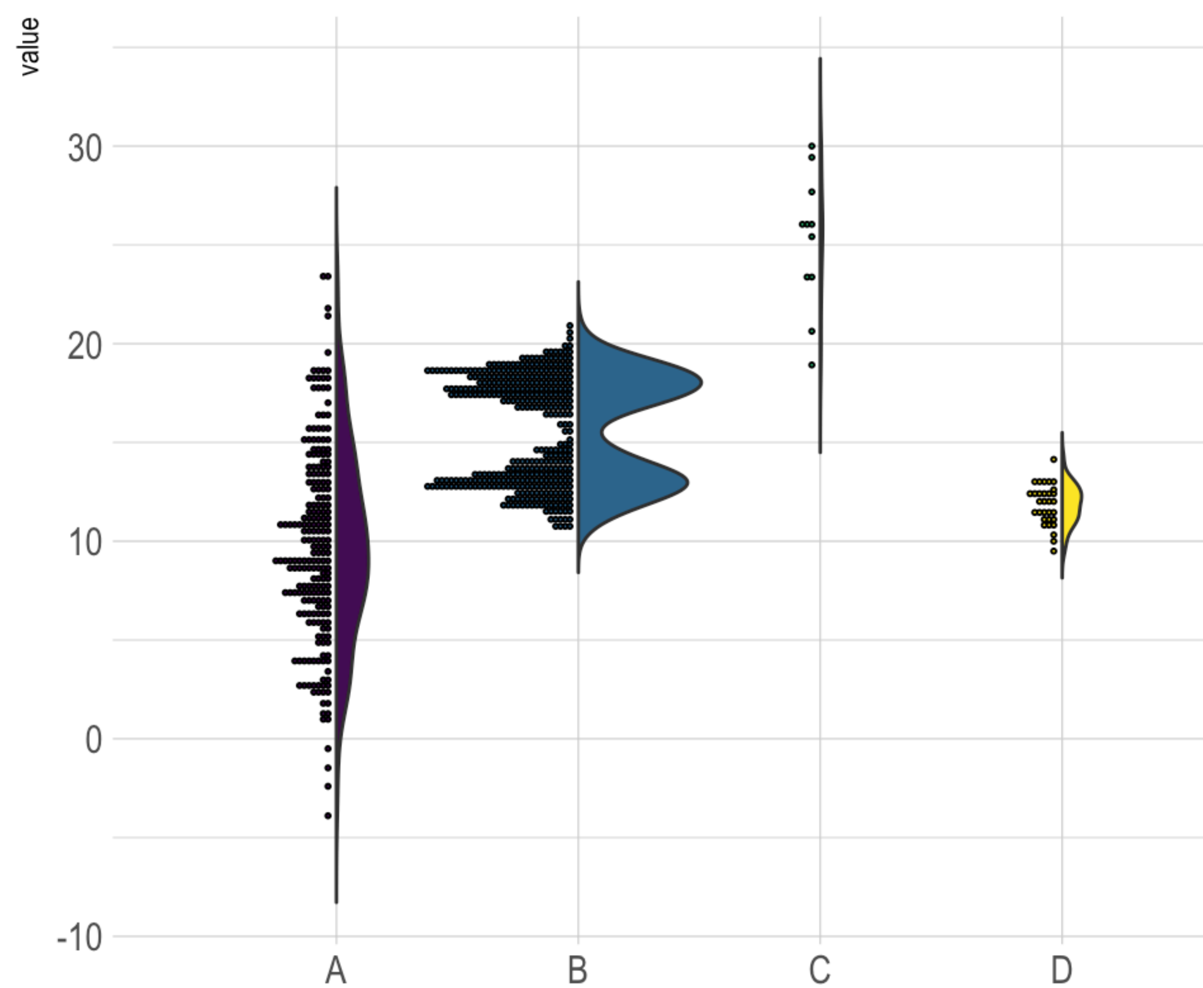
or show *distribution* and summary statistics

If dealing with many data points, showing them no longer works.

Show their distribution instead, using violin plots:

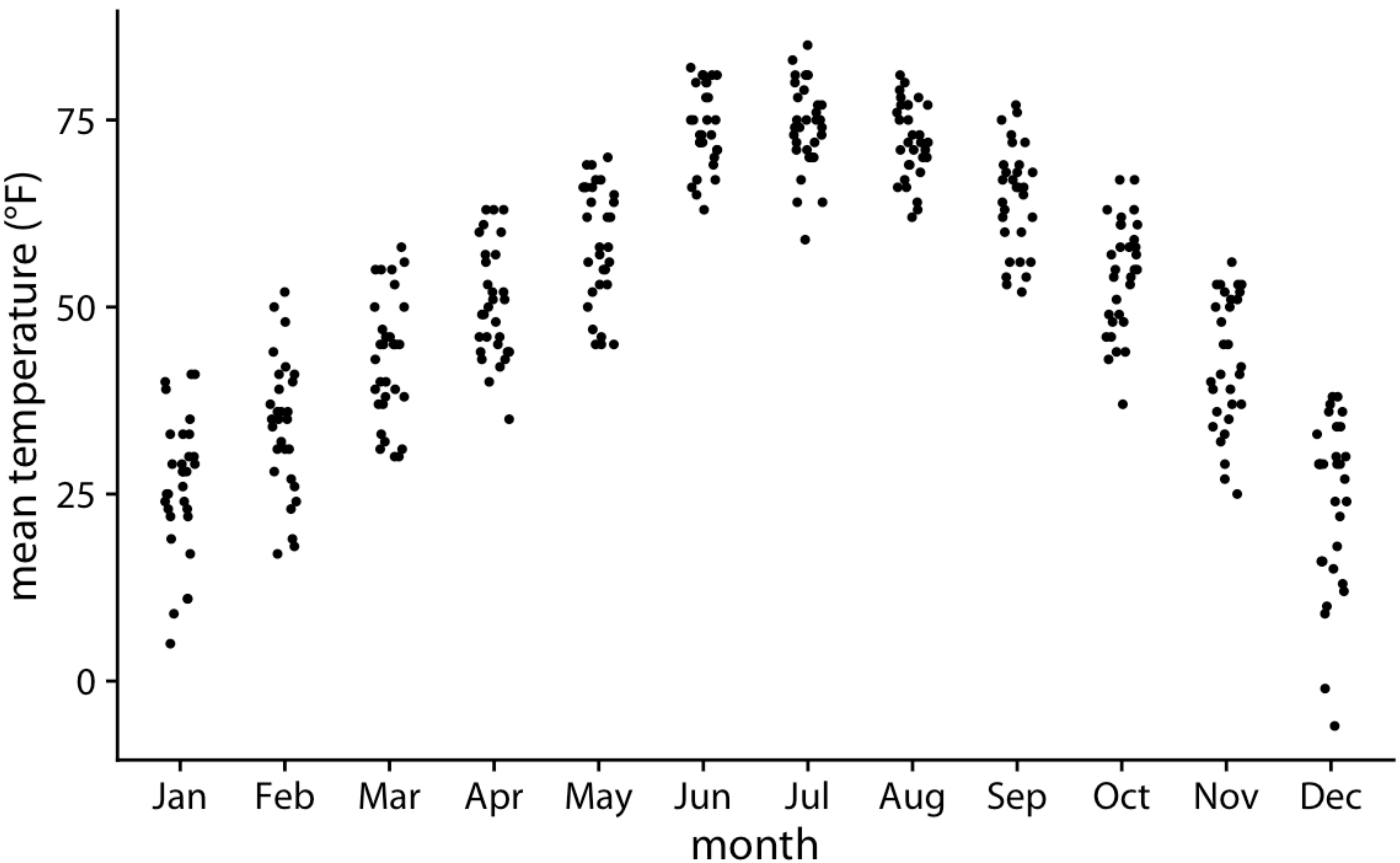


or give a sense of the raw data.



Sparse datasets

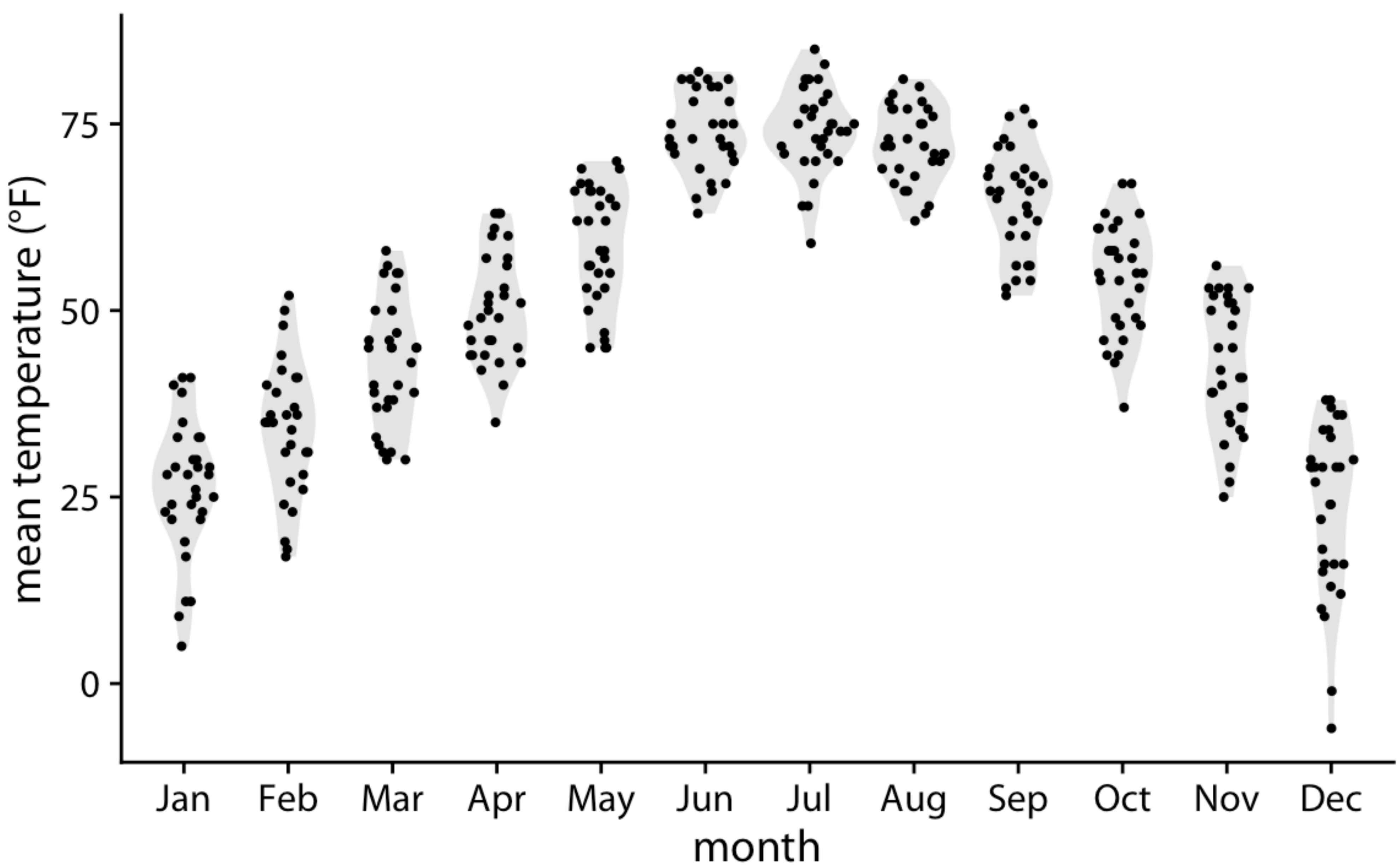
Plot the raw data points…



[Image source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

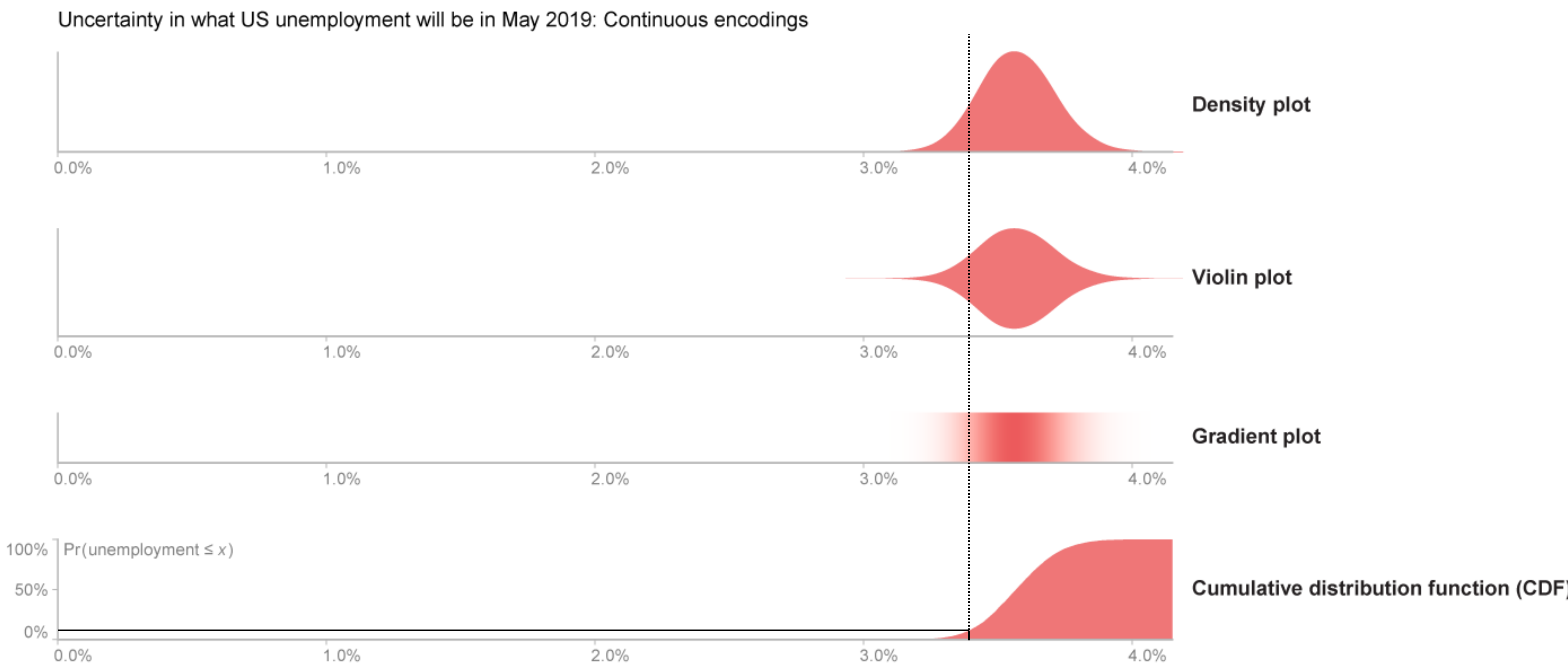
Sparse datasets

Plot the raw data points… and violins.



[Image source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

Visualizing uncertainty, one-sided



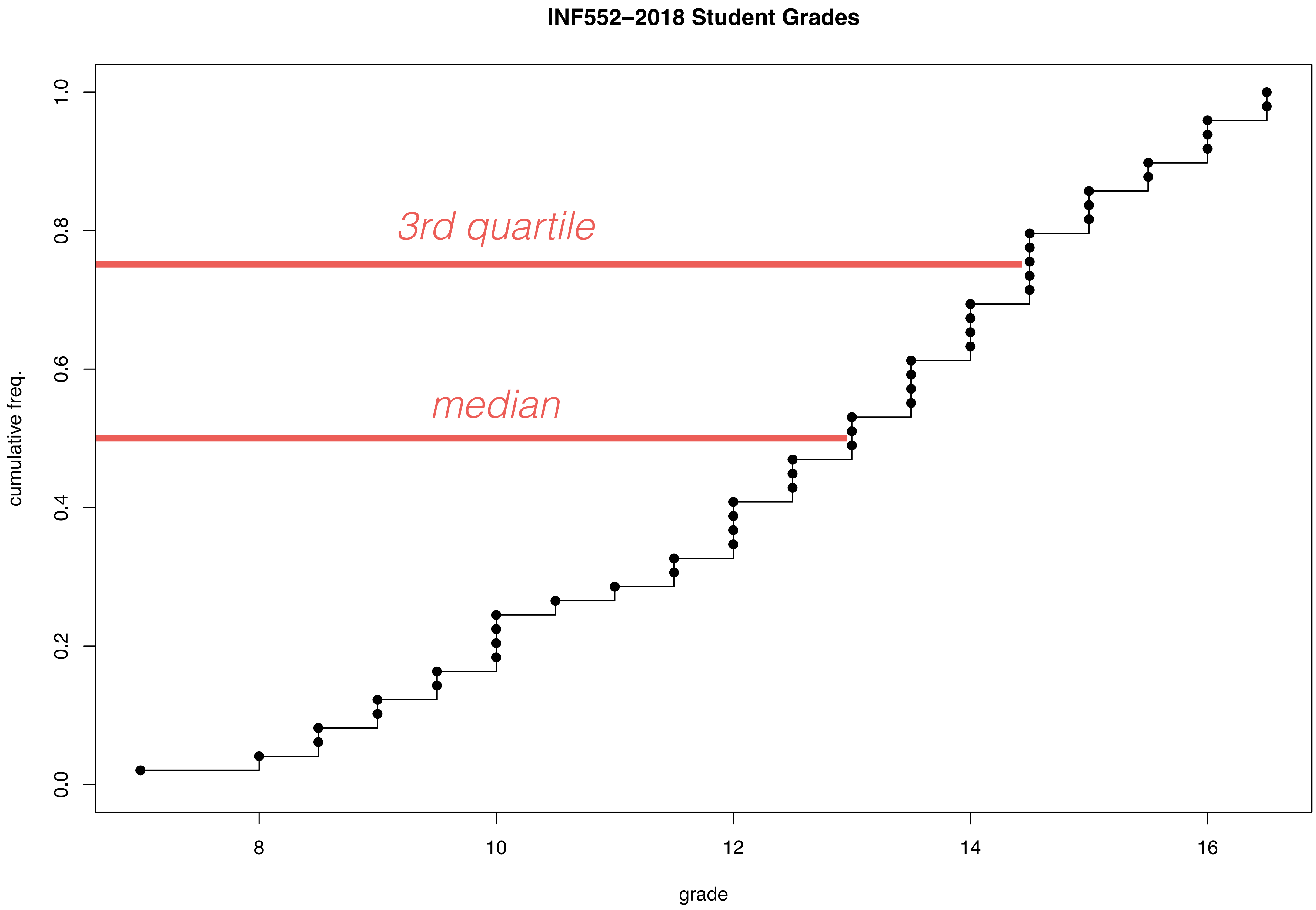
CDF effectively supports answering questions such as:

what is the probability that **`var`** *will be less than* **`x`**

Position judgment *vs.* area/color judgment with the other three encodings.

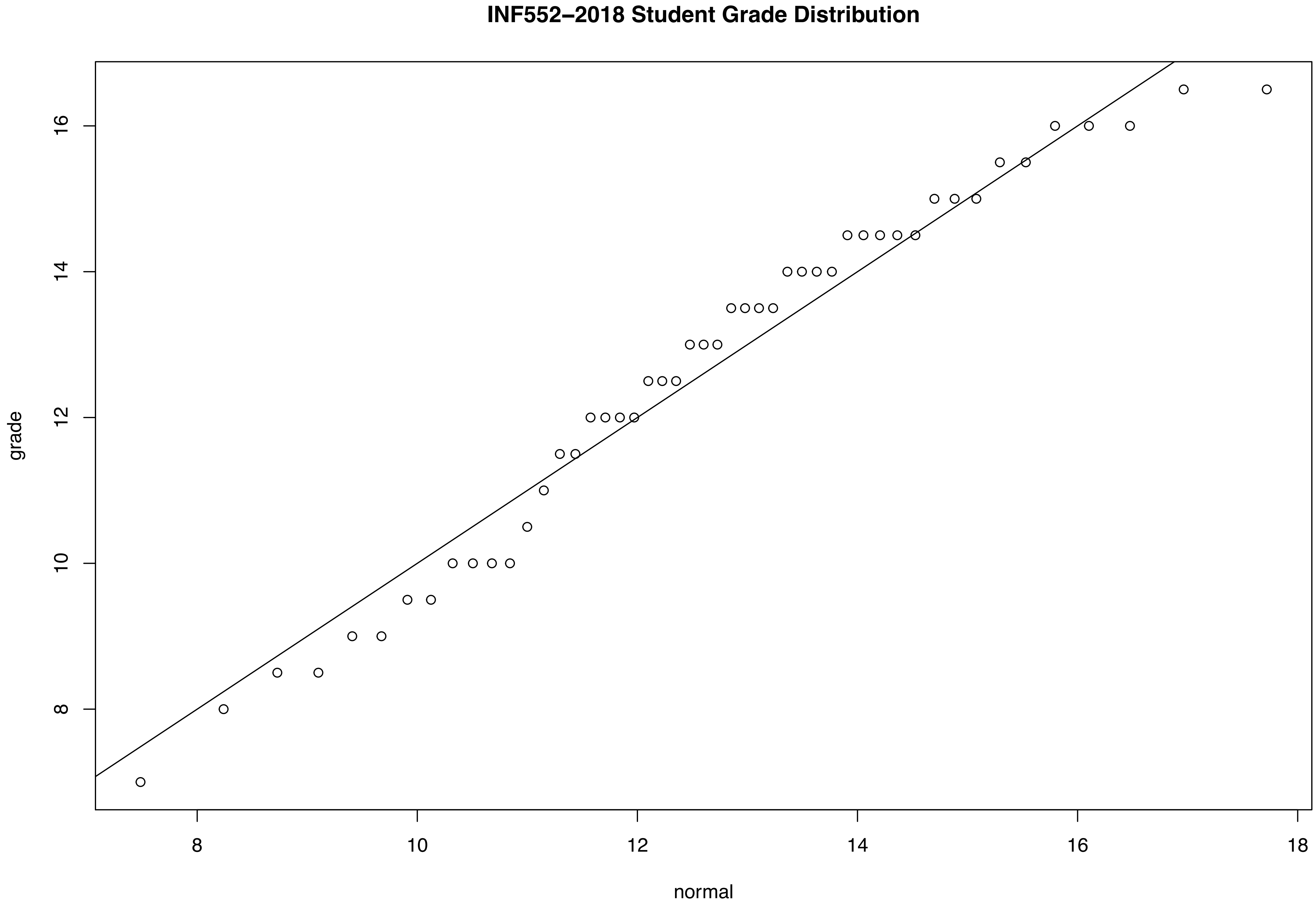
Mean, mode, *etc*. are more difficult to find though.

Empirical Cumulative Distribution Function (ECDF)



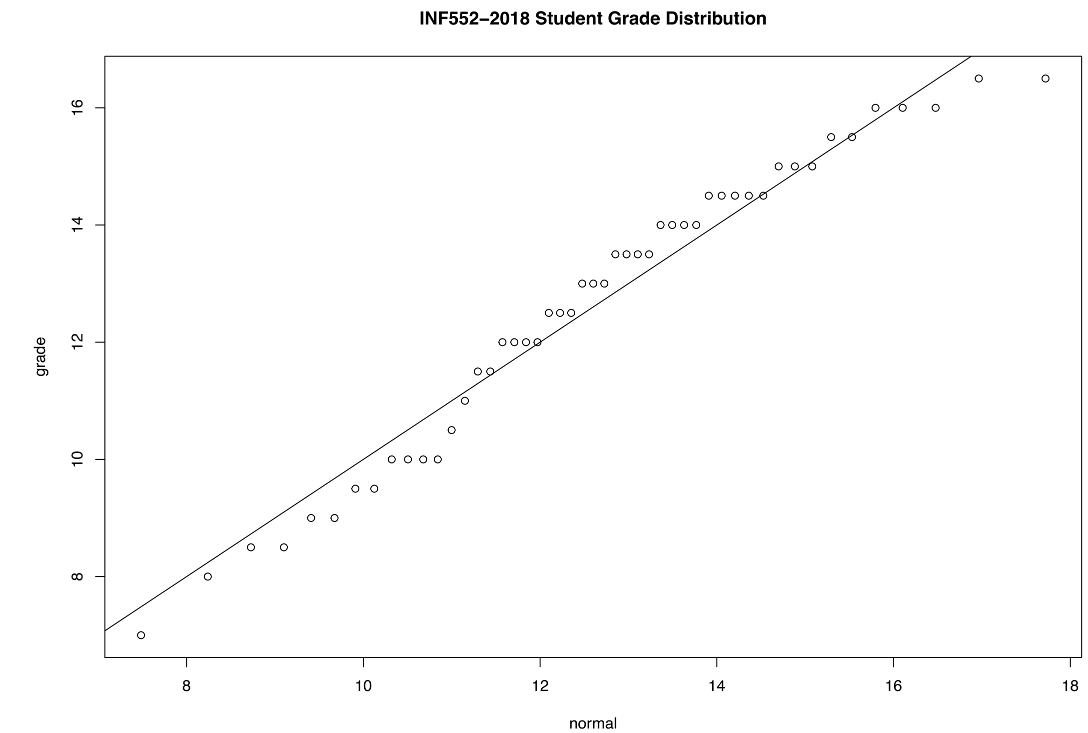
Easy to find the median, quantiles, *etc*.

Quantile-quantile plot (Q-Q plot)



Visual representation of the similarity between two distributions:
- two distinct datasets;
- theoretical *vs.* observed.

Quantile-quantile plot (Q-Q plot)



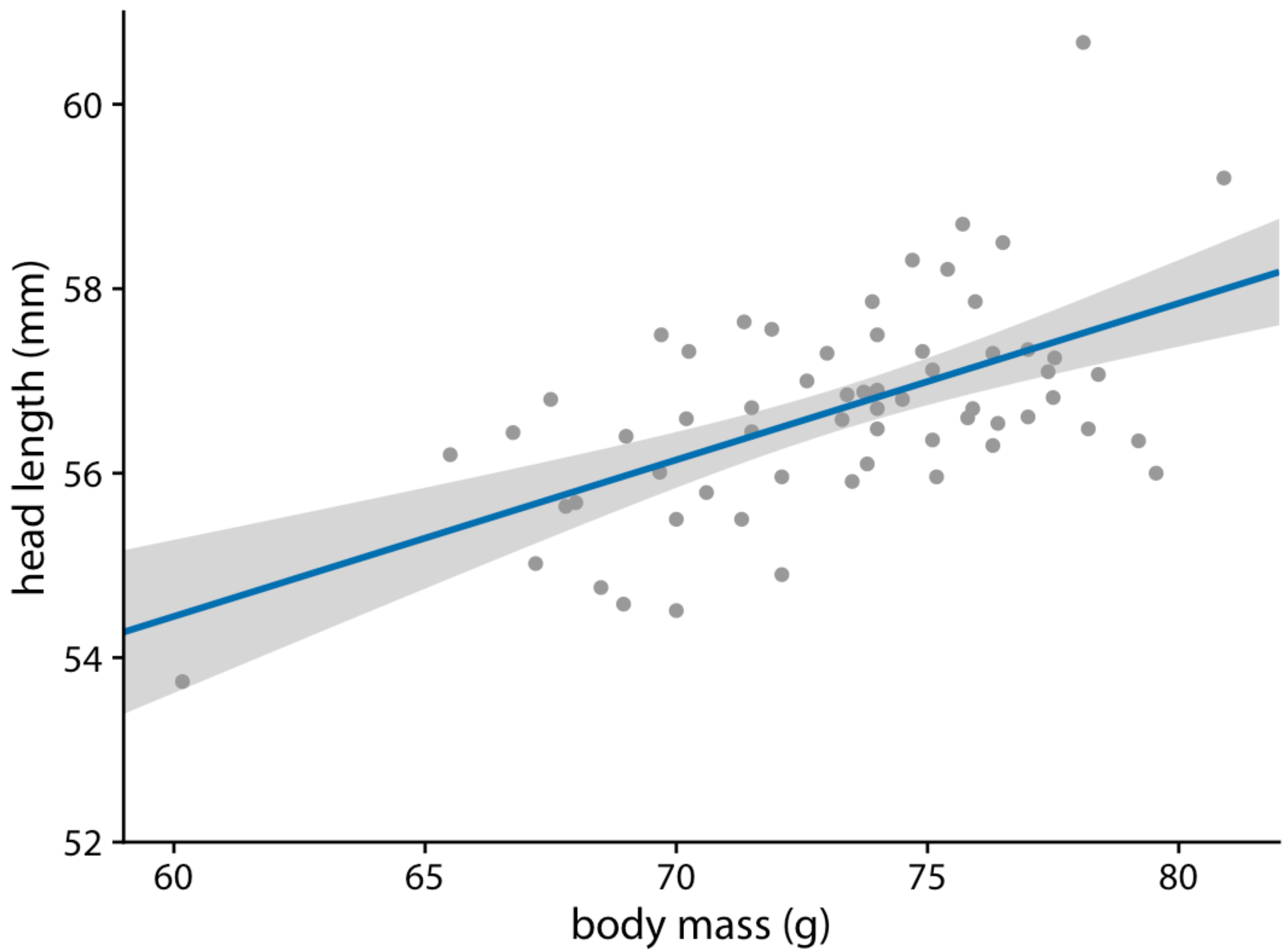
= sample sizes:

- plot the sorted data;

≠ sample sizes, *with card(d1) < card(d2)*:

- quantiles correspond to the sorted data values from *d1*;
- quantiles for *d2* are interpolated.

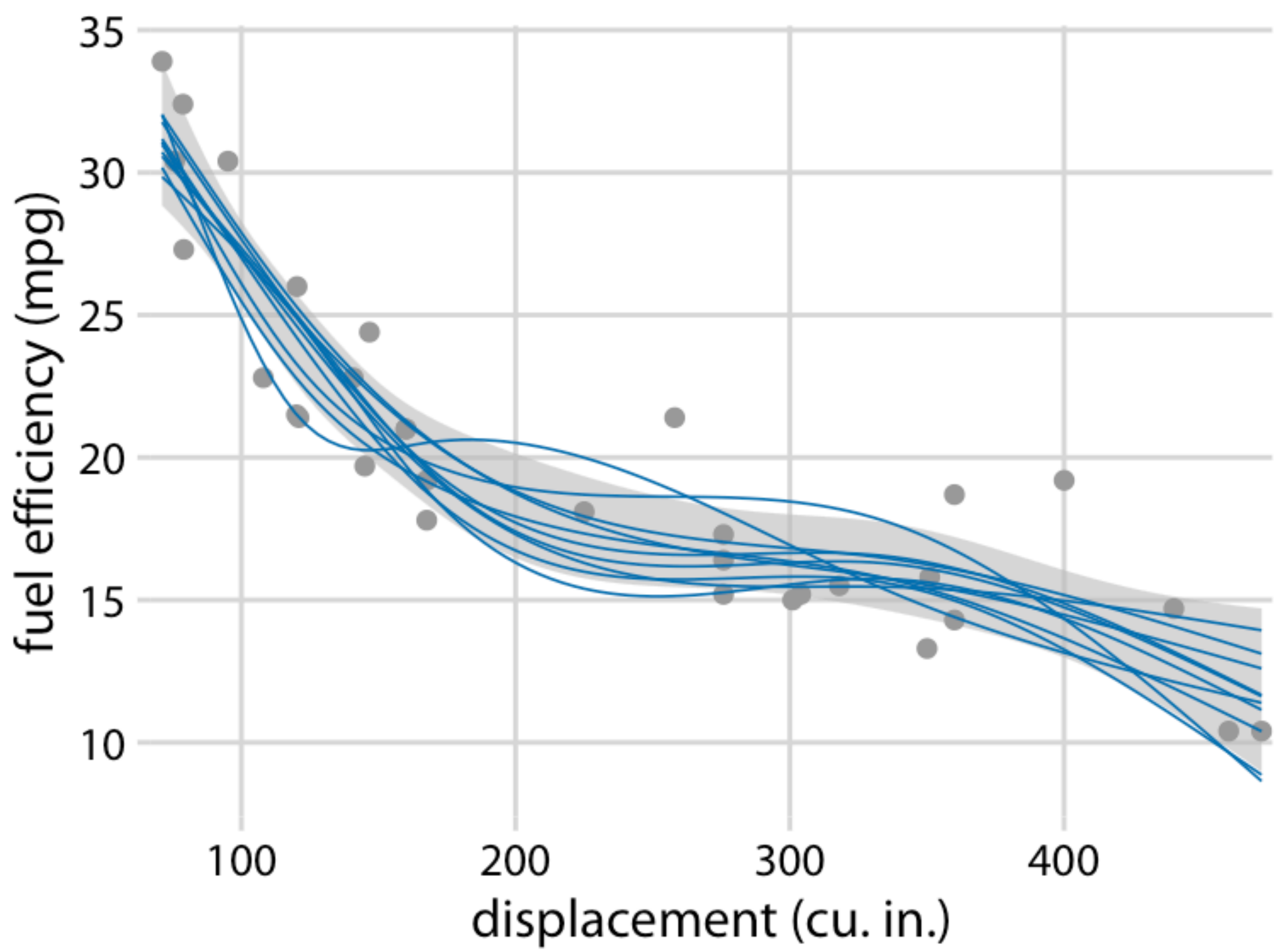
Uncertainty of curve fits: confidence bands (linear fits)



Uncertainty about both slope and intercept of the regression line.

[Image source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

Uncertainty of curve fits: confidence bands (non-linear fits)



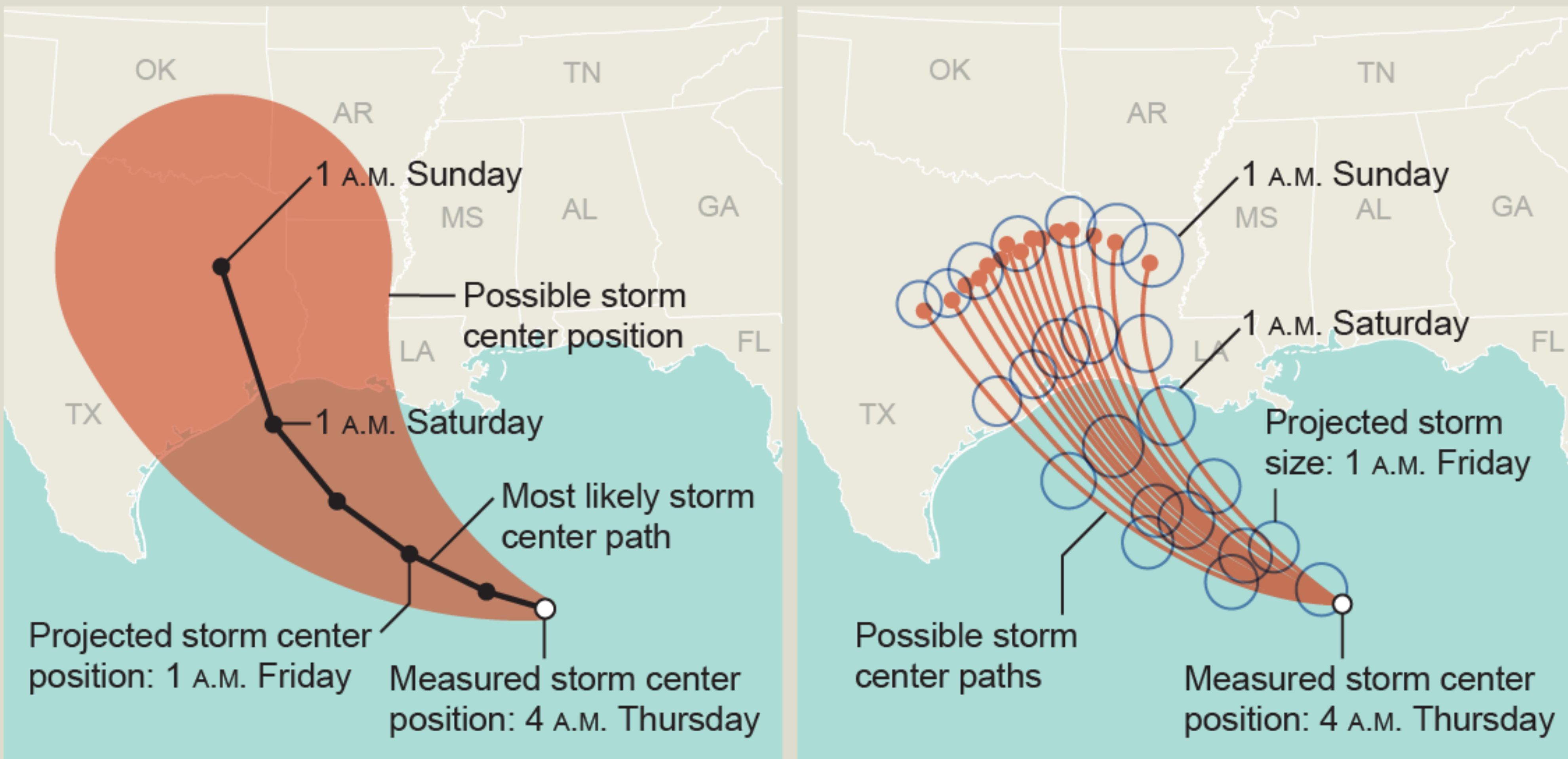
Uncertainty not only about position of the curve, but its actual shape (wiggliness).

Issue of deterministic construal error again.

[Image source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

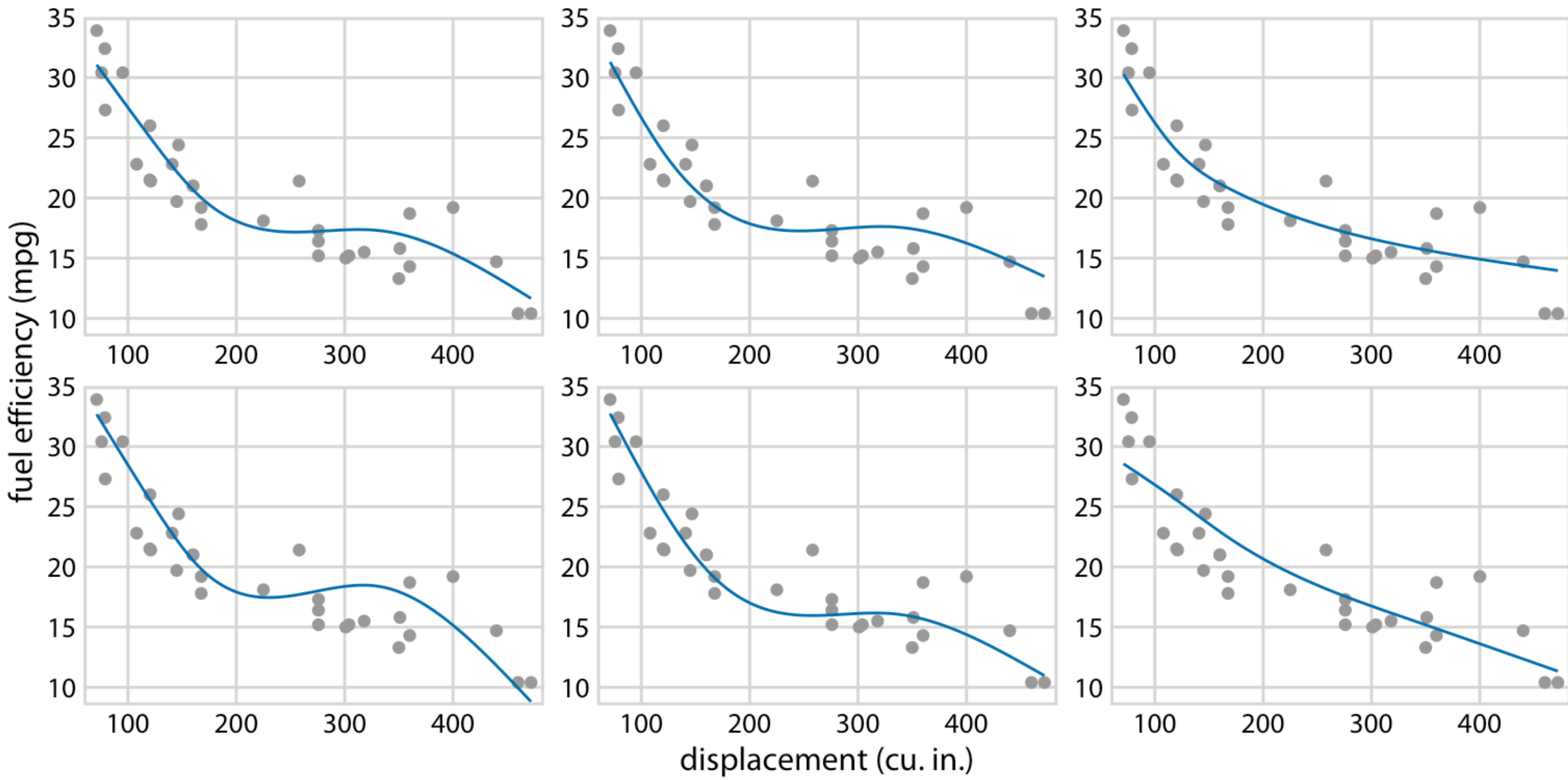
Deterministic construal errors

"Cone of uncertainty" (*left*) shows where a hurricane may head, according to a group of forecasts. An alternative is to show the specific path predicted by each forecast (*right*). Both approaches have pros and cons in helping people judge the risk they may face, but the one on the right makes it clearer that the path is difficult to predict.



[Credit: Tiffany Farrant-Gonzalez; Sources: National Hurricane Center (cone of uncertainty); "Visualizing Uncertain Tropical Cyclone Predictions Using Representative Samples from Ensembles of Forecast Tracks," by Le Liu et al., in IEEE Transactions on Visualization and Computer Graphics, Vol. 25; August 20, 2018 (multiple storm paths)]

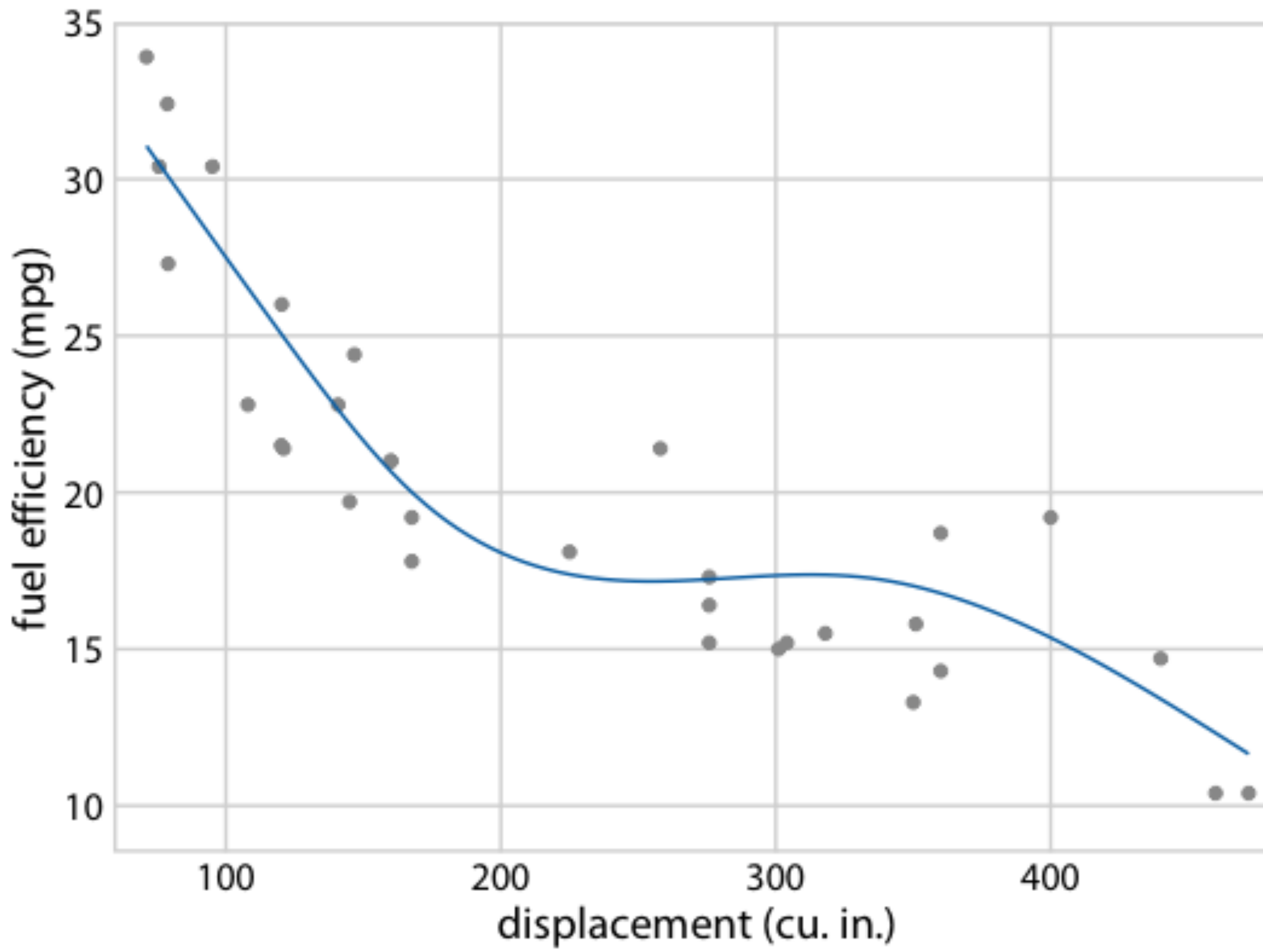
Show different possible outcomes



Plot multiple (hypothetical) outcomes, either as small multiples…

[Image source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

Show different possible outcomes: Hypothetical Outcome Plots



Plot multiple (hypothetical) outcomes, either as small multiples… or animating between them.

- Very effective at conveying uncertainty.
- Hard switch between them, or smooth (quick) fade in/out, but no smooth shape morphing.
- Show outcomes that are representative of the true distribution of possible outcomes.

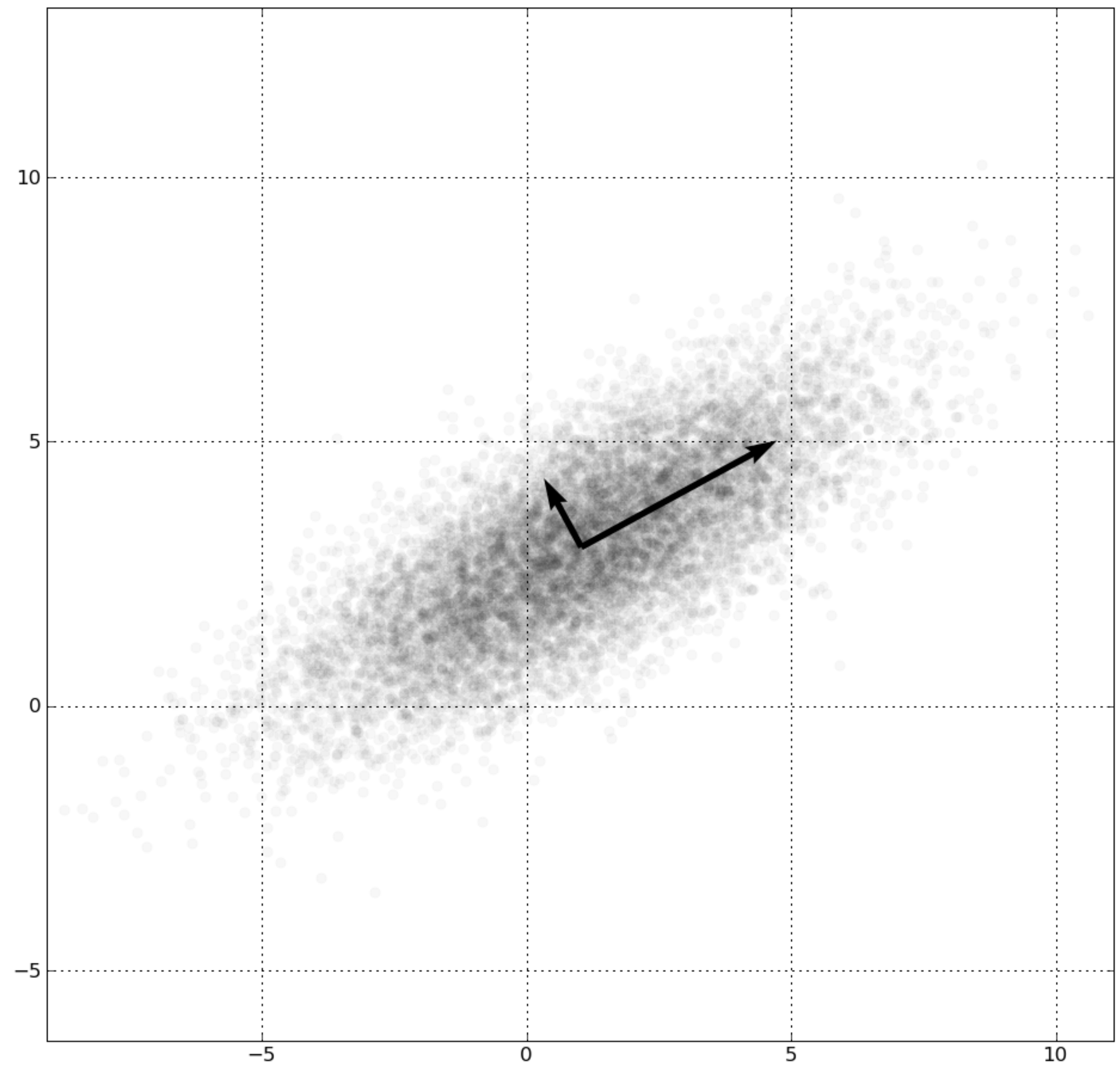
[Image source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

Dimensionality Reduction

Given a set of data points that lie in a high-dimensional space, find a representation of those points in a lower-dimensional space (2D space, 3D volume).

[Image source: R. Cutura et al. VisCoDeR
A tool for Visually Comparing Dimensionality Reduction Algorithms, 2018]

Dimensionality Reduction: PCA



Principal Component Analysis

PCs are **linear** combinations of original variables

PCs are uncorrelated
PCs are ordered by max. variation they yield (decreasingly)

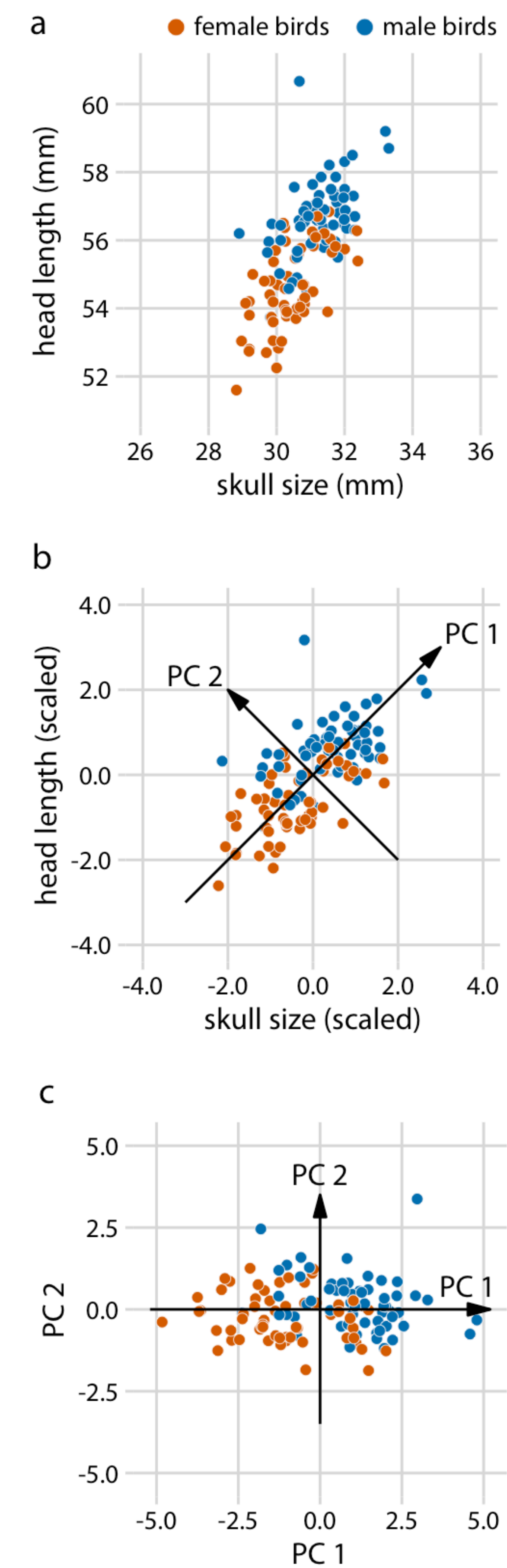
http://en.wikipedia.org/wiki/File:GaussianScatterPCA.png

Dimensionality Reduction: PCA

Standardize variables to zero mean and unit variance.

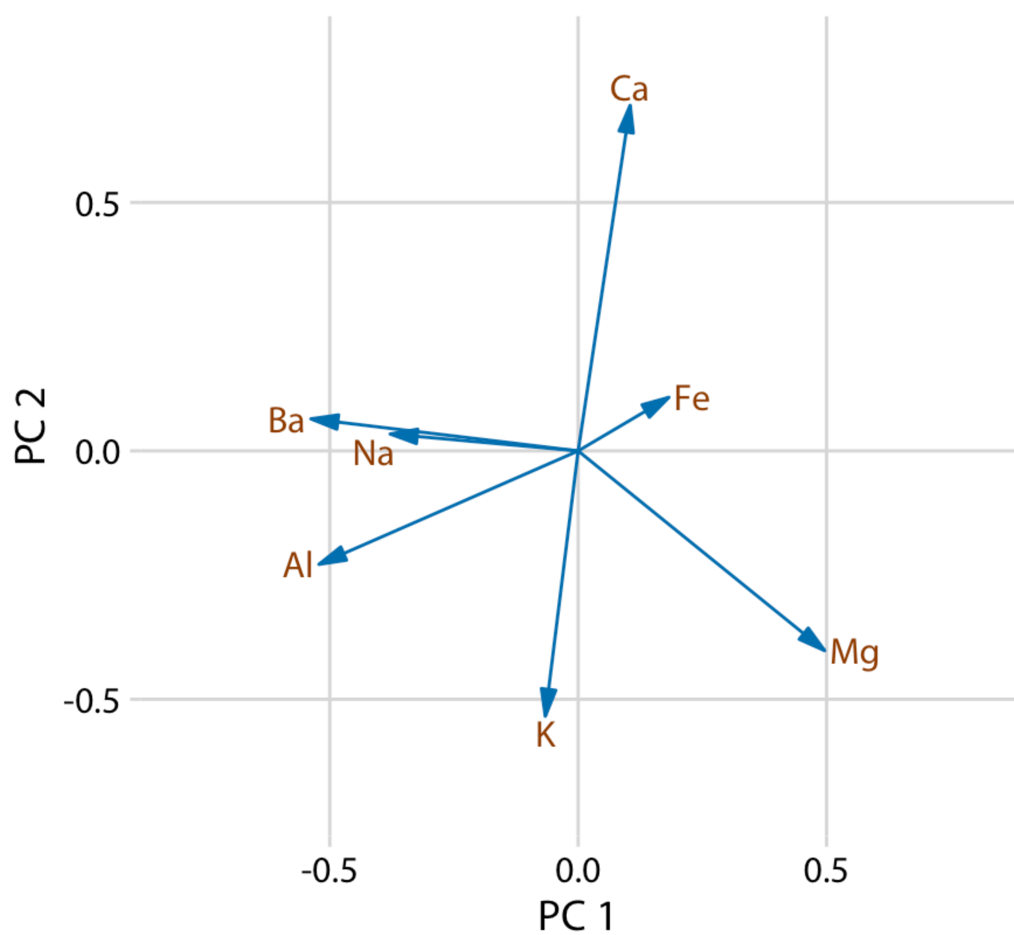
Define PCs along directions of maximum variation.

Project the data in the new coordinates system.



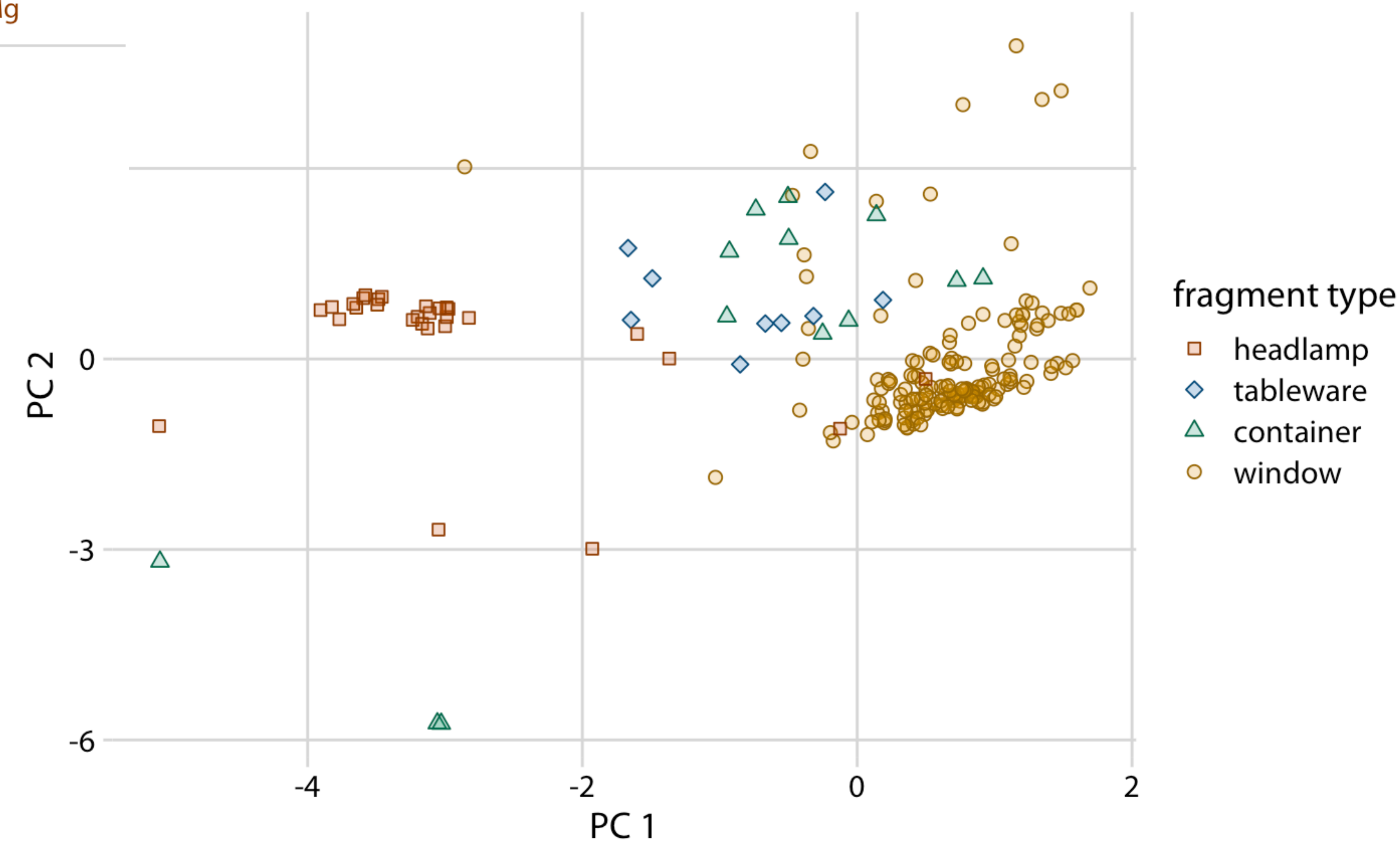
[Example source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

Dimensionality Reduction: PCA



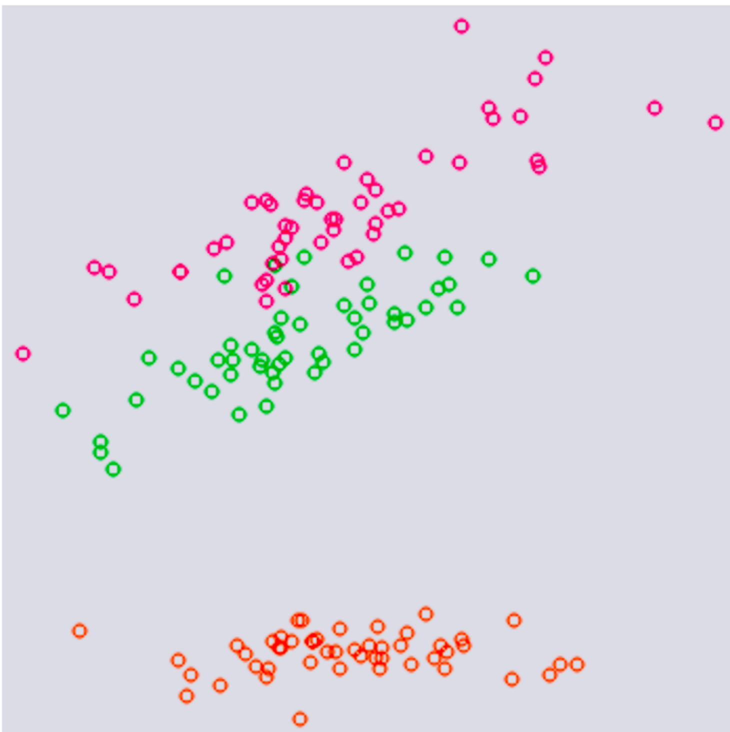
Composition of the first two PCs

Dataset projection along those two PCs



[Example source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

Dimensionality Reduction: MDS



PCA

MDS

Multi-Dimensional Scaling (**non-linear**).

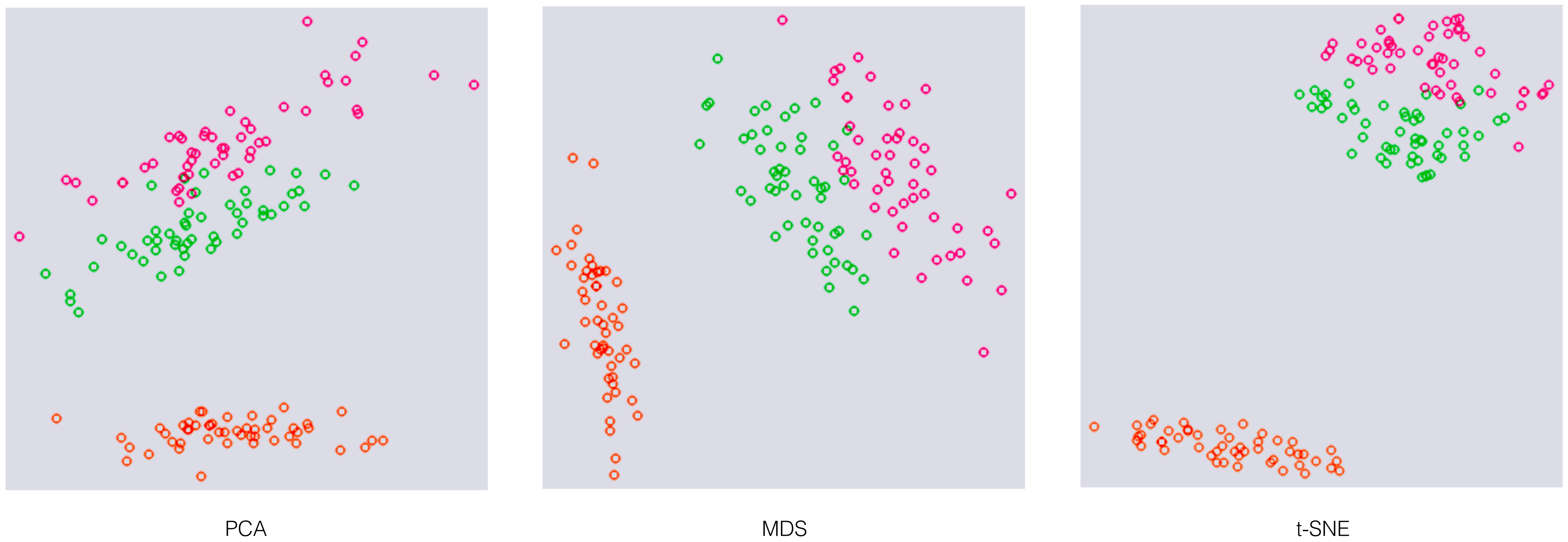
PCA considers the data points themselves, while MDS considers the pairwise distances between data points: it tries to preserve distances from the *n*D space in the 2D plane projection.

Preserves global structure (larger distances) at the expense of local structure (smaller distances).

Analogy with force-directed graph layout algorithms.

ISOMAP only considers the k-nearest-neighbors of each data point.

Dimensionality Reduction: t-SNE



t-distribution Stochastic Neighborhood Embedding (**non-linear**).

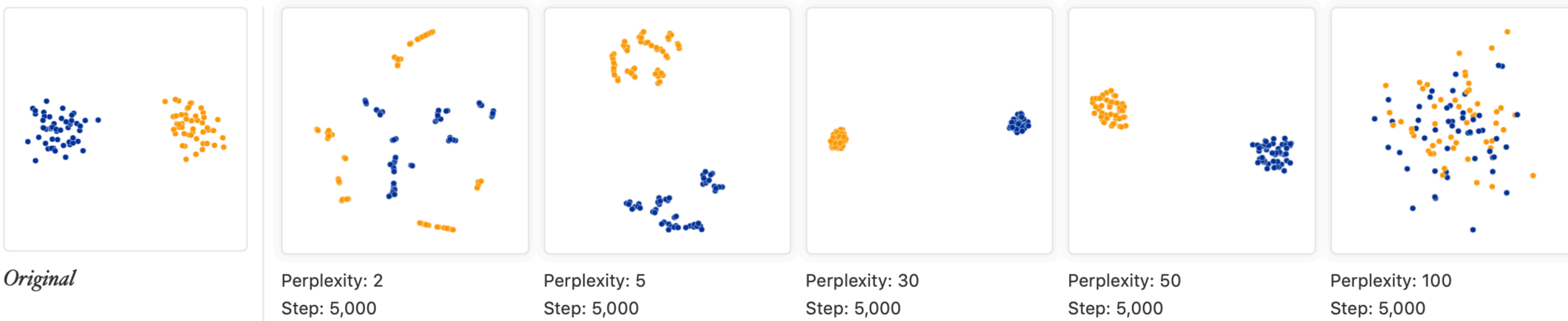
Analogy with force-directed graph layout algorithms (again).

Better preservation of both global *and* local structure.

Can work well with a large number of dimensions (1000+).

Different runs can yield different outputs.

Dimensionality Reduction: t-SNE



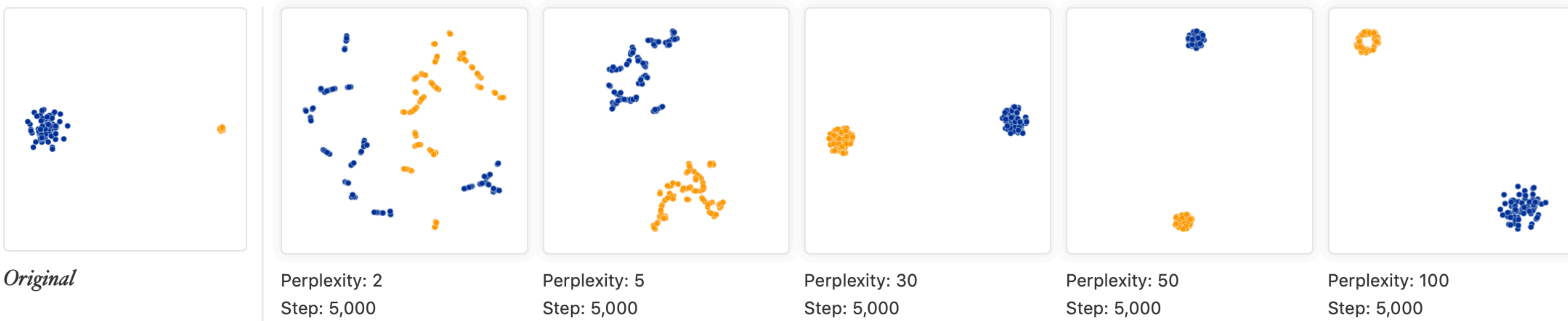
perplexity must be < number of points (here 50)

iterate until a stable configuration is reached

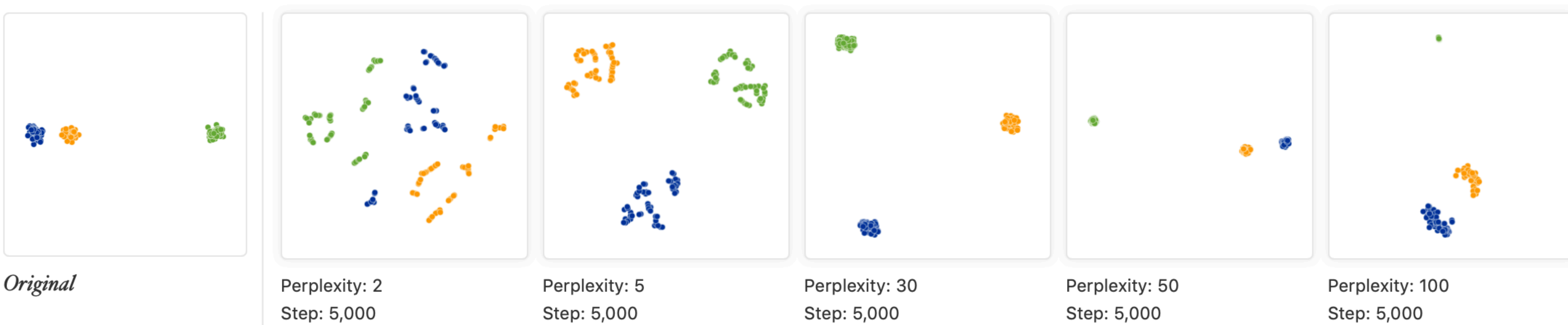
[All examples from Wattenberg et al., How to Use t-SNE Effectively, Distill, 2016. 10.23915/distill.00002]

Dimensionality Reduction: t-SNE

t-SNE adapts the notion of distance to regional density

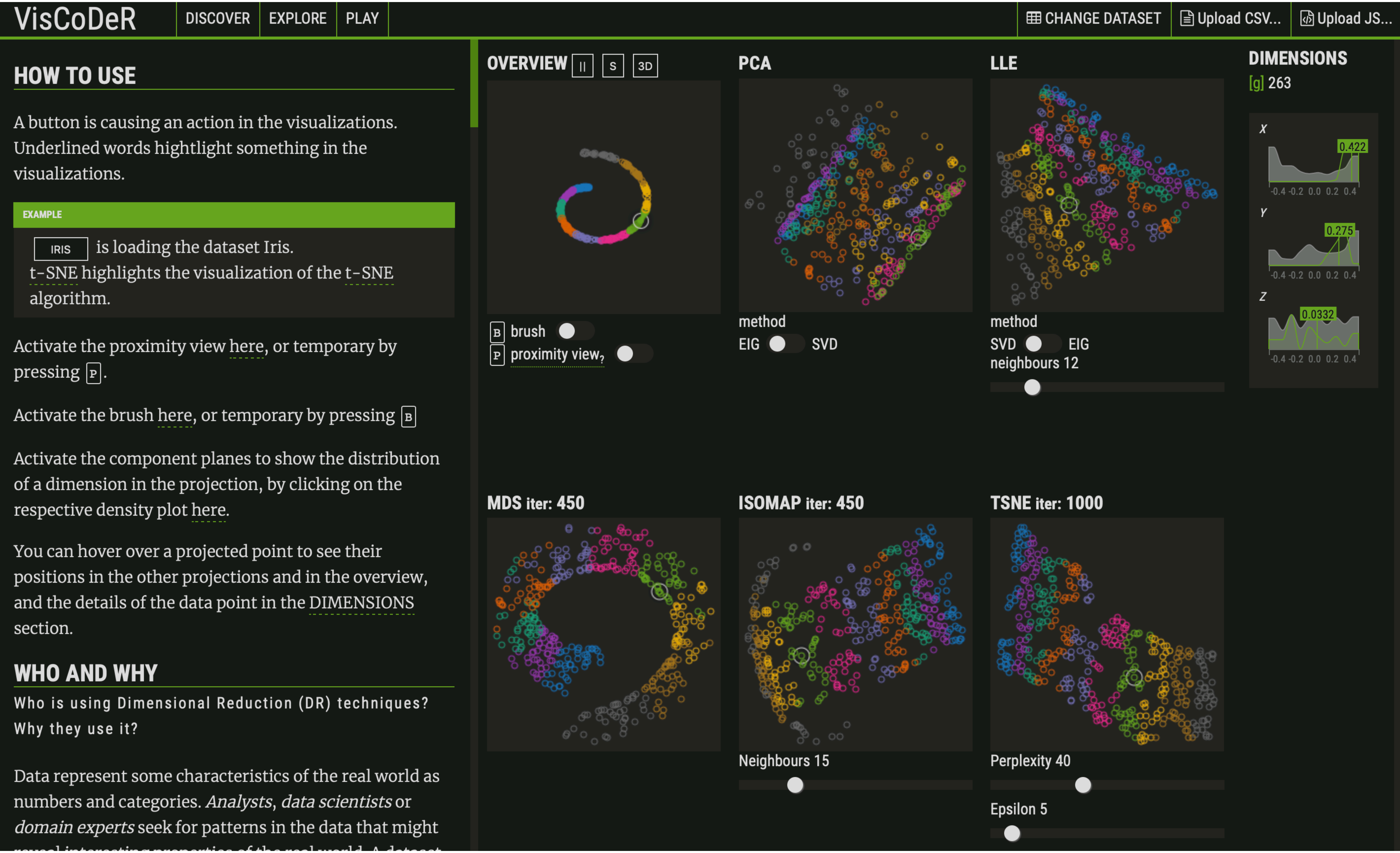


$\Longrightarrow$ cannot compare relative size of clusters



distance is not reliable

[All examples from Wattenberg et al., How to Use t-SNE Effectively, Distill, 2016, 10.23915/distill.00002] https://www.renecutura.eu/viscoder/



viscoder.mp4