Paul Rosen
paul.rosen@utah.edu
@paulrosenphd
https://cspaul.com







# Visualization for Data Science
# DS-4630 / CS-5630

## Reducing Data by Aggregation

# Learning objectives

- Explain what aggregation shows, and what it hides
- Choose bins, bandwidths, and 2D aggregation for large data
- Summarize groups with uncertainty
- Know when to use and read correlation, clustering, and dimensionality-reduction output
- Recognize when the level of aggregation can change the conclusion

# why reduce?

- Too many data items and/or too many attributes to focus on what is important in the data

## Reducing Items and Attributes

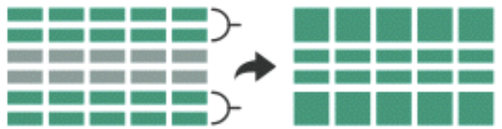
- **Filter**
  - Items



  - Attributes



- **Aggregate**
  - Items

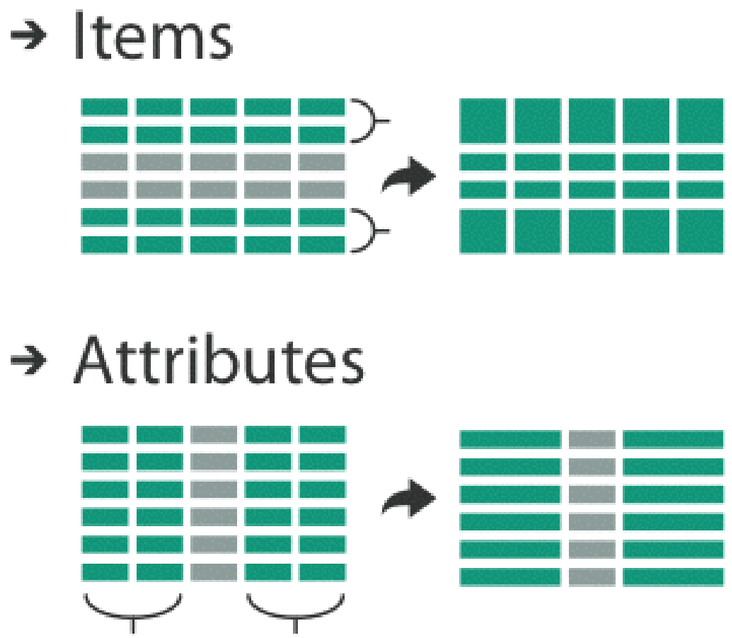


  - Attributes



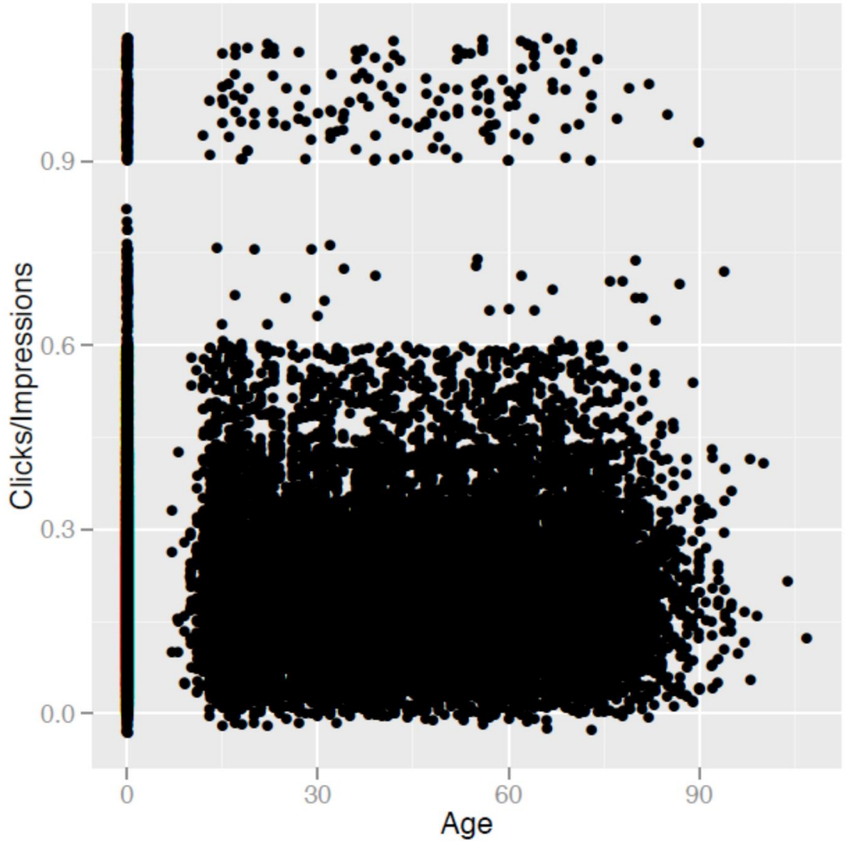
# Aggregate

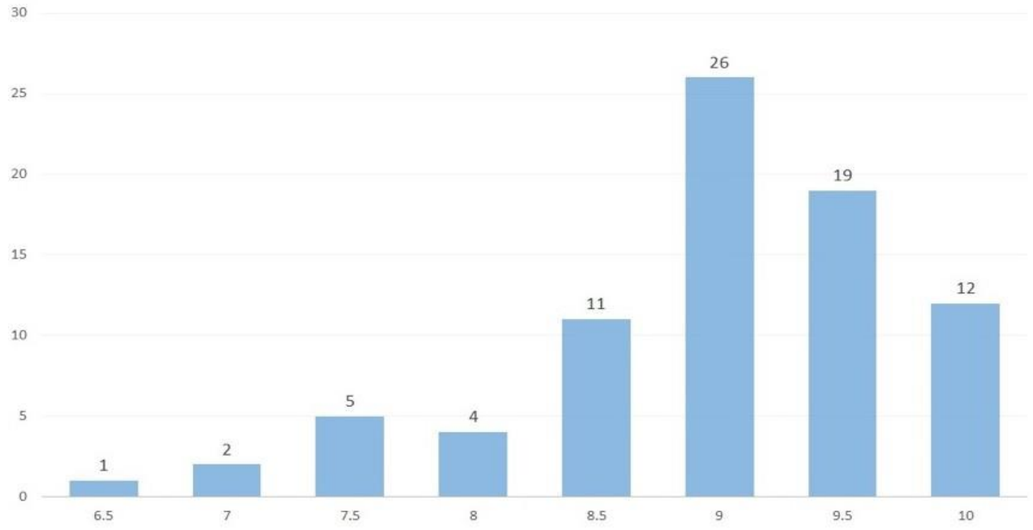
- A group of elements is represented by a new derived element that stands in for the entire group
- Numerous ways to reduce include statistics, topology, machine learning, etc.



# Problem #1: Aggregate Items

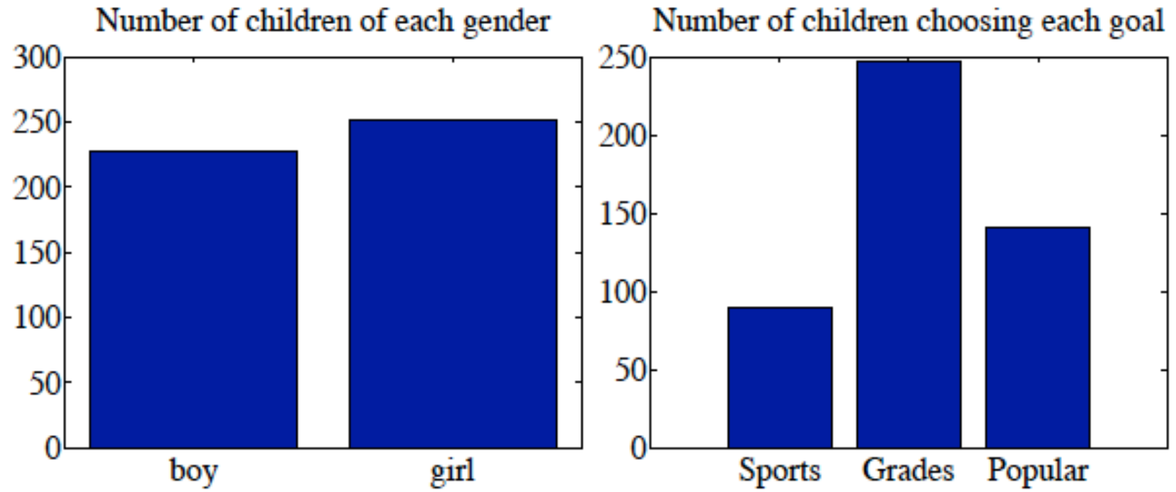
- We have too many data points to show

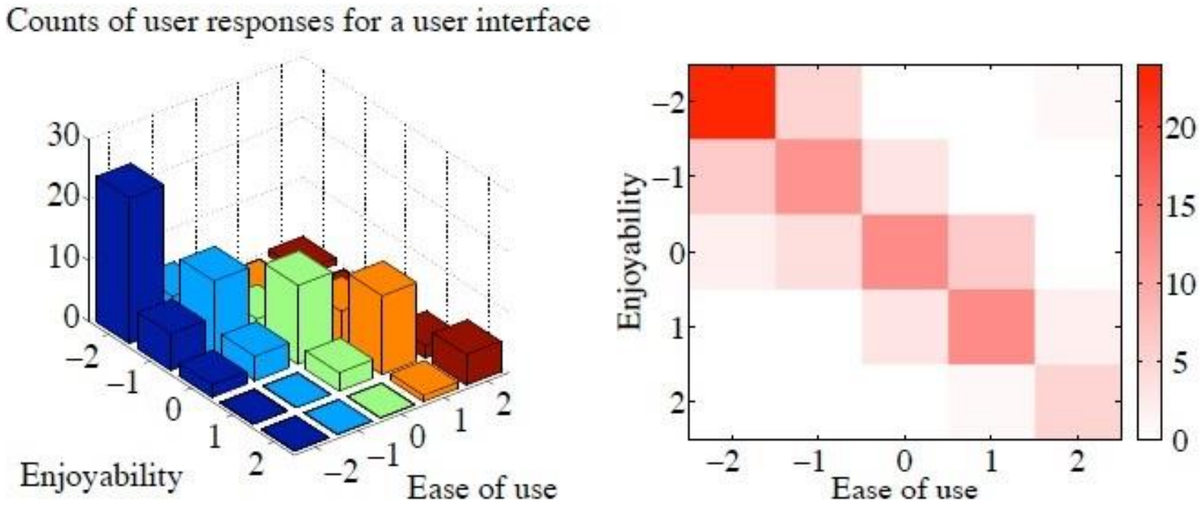


# Histograms

- Generally referring to a bar chart-based visualization that allows evaluating distribution of values.
- Really, histograms capture a distribution of data

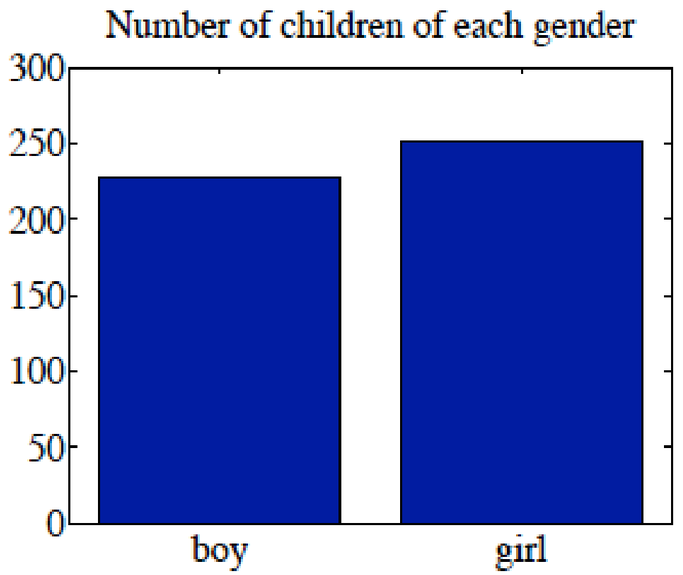


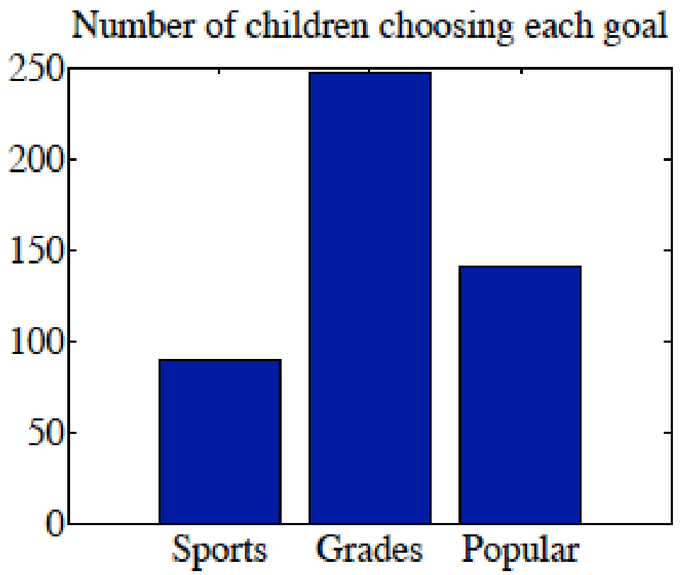



# Categorical data

- Simply count occurrences of each type and visualize

| Gender | Goal | Gender | Goal |
|---|---|---|---|
| boy | Sports | girl | Sports |
| boy | Popular | girl | Grades |
| girl | Popular | boy | Popular |
| girl | Popular | boy | Popular |
| girl | Popular | boy | Popular |
| girl | Popular | girl | Grades |
| girl | Popular | girl | Sports |
| girl | Grades | girl | Popular |
| girl | Sports | girl | Grades |
| girl | Sports | girl | Sports |

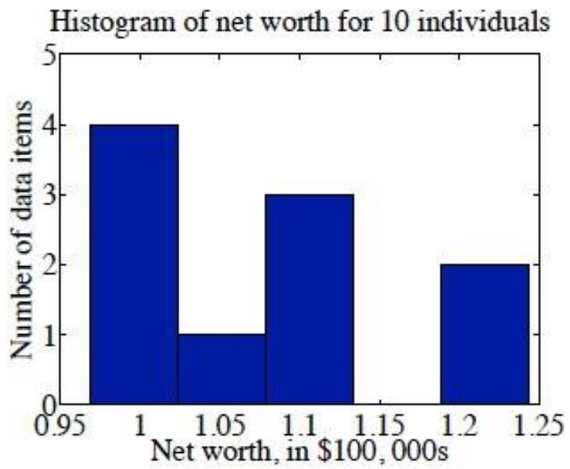


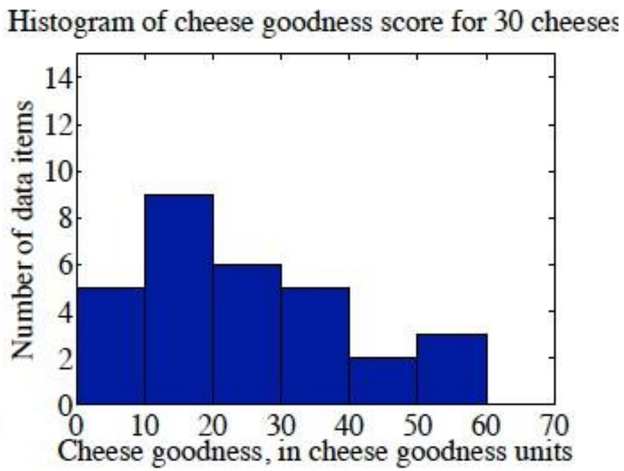



# Continuous Data Histograms

| Index | net worth |
|---|---|
| 1 | 100, 360 |
| 2 | 109, 770 |
| 3 | 96, 860 |
| 4 | 97, 860 |
| 5 | 108, 930 |
| 6 | 124, 330 |
| 7 | 101, 300 |
| 8 | 112, 710 |
| 9 | 106, 740 |
| 10 | 120, 170 |

| Index | Taste score | Index | Taste score |
|---|---|---|---|
| 1 | 12.3 | 11 | 34.9 |
| 2 | 20.9 | 12 | 57.2 |
| 3 | 39 | 13 | 0.7 |
| 4 | 47.9 | 14 | 25.9 |
| 5 | 5.6 | 15 | 54.9 |
| 6 | 25.9 | 16 | 40.9 |
| 7 | 37.3 | 17 | 15.9 |
| 8 | 21.9 | 18 | 6.4 |
| 9 | 18.1 | 19 | 18 |
| 10 | 21 | 20 | 38.9 |



# Calculating a continuous histogram

- Given: $X=\{x_0,\ldots,x_n\}$
- Select: k bins

- $bin_i = \text{floor}( k * (x_i - \min X) / (\max X - \min X) )$

# Calculating a continuous histogram

- X={1,2.5,3,4}
- k = 3



# Calculating a continuous histogram

- X={1,2.5,3,4}
- k = 3
- $bin_i$ = floor( k * ($x_i$ – min X) / (max X – min X) )



# Calculating a continuous histogram

- X={1,2.5,3,4}
- k = 3
- $bin_i$ = floor( 3 * ($x_i$ – 1) / (4 – 1) )



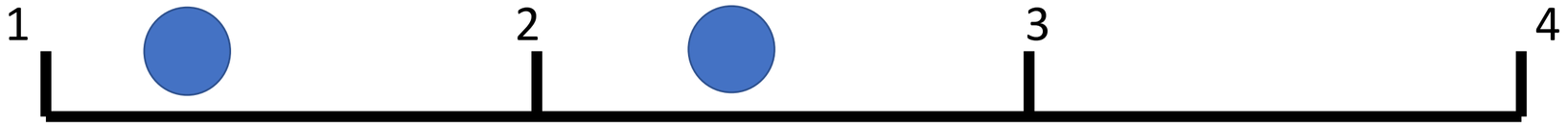
# Calculating a continuous histogram

- X={1,2.5,3,4}
- k = 3
- 1 -> floor( 3 * (1 – 1) / (4 – 1) ) = Bin 0



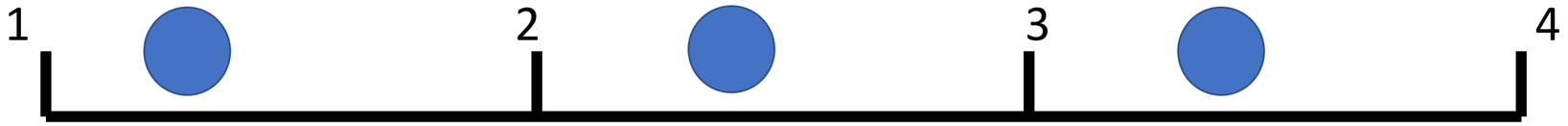
# Calculating a continuous histogram

- X={1,2.5,3,4}
- k = 3
- 2.5 -> floor( 3 * (2.5 – 1) / (4 – 1) ) = Bin 1



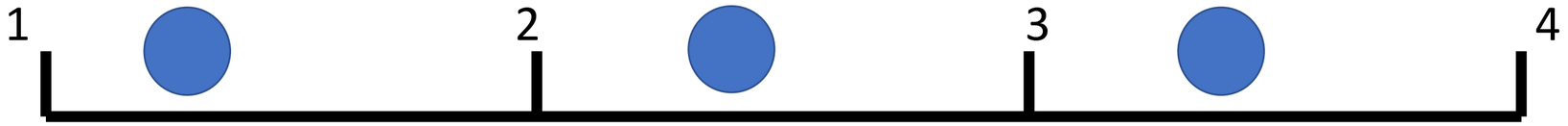
# Calculating a continuous histogram

- X={1,2.5,3,4}
- k = 3
- 3 -> floor( 3 * (3 – 1) / (4 – 1) ) = Bin 2



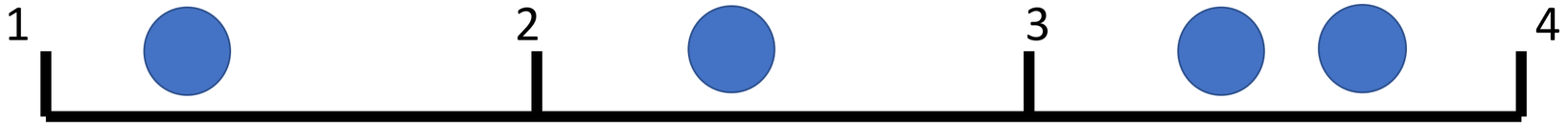
# Calculating a continuous histogram

- X={1,2.5,3,4}
- k = 3
- 4 -> floor( 3 * (4 – 1) / (4 – 1) ) = Bin 3?



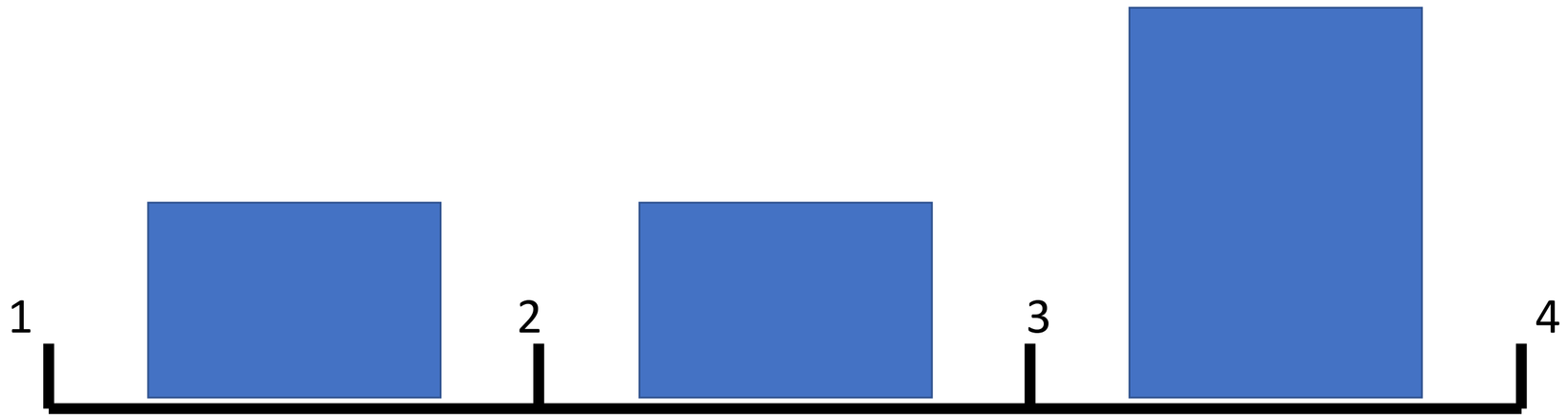
# Calculating a continuous histogram

- X={1,2.5,3,4}
- k = 3
- 4 -> floor( 3 * (4 – 1) / (4 – 1) ) = Bin 2

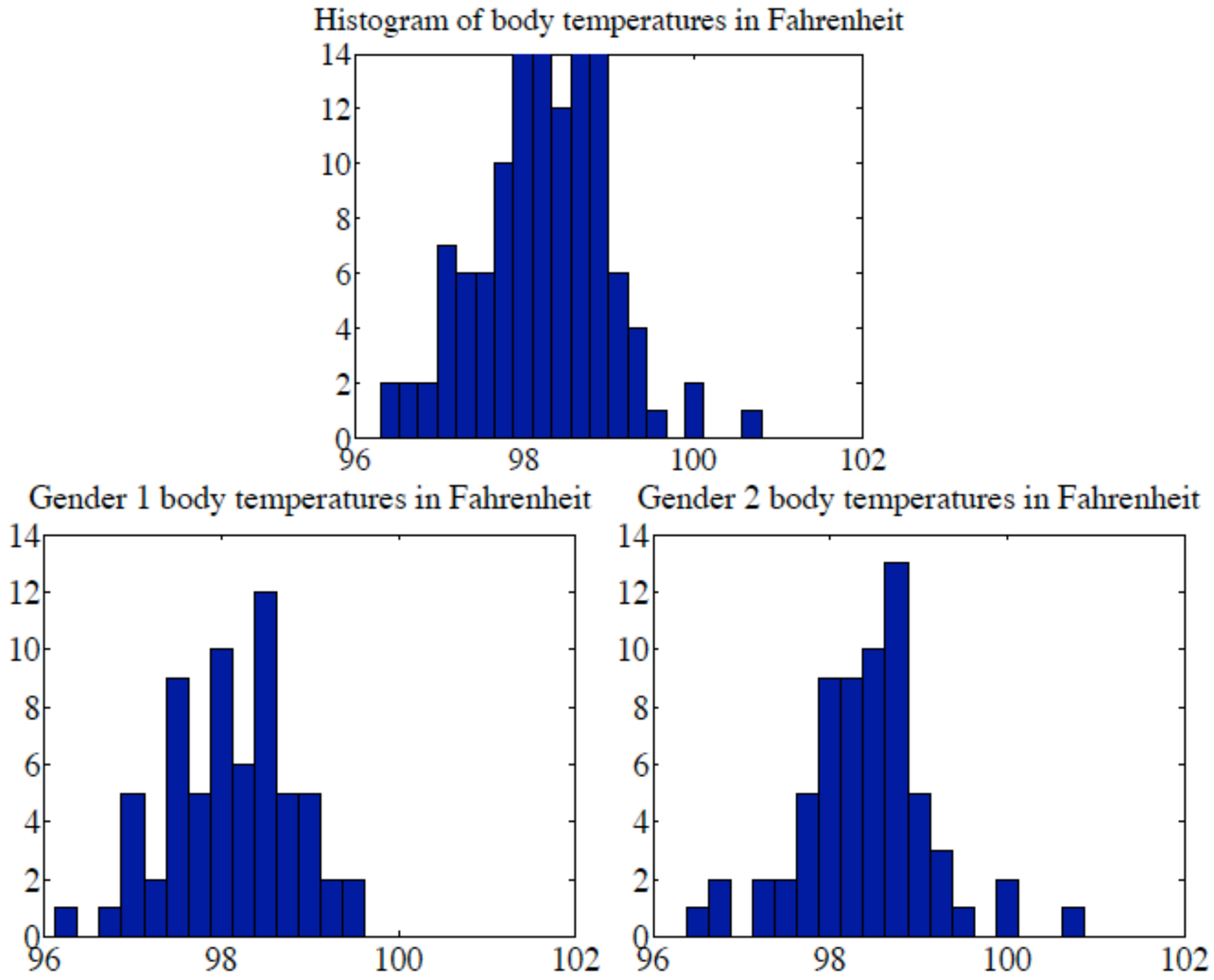


# Calculating a continuous histogram

- X={1,2.5,3,4}
- k = 3

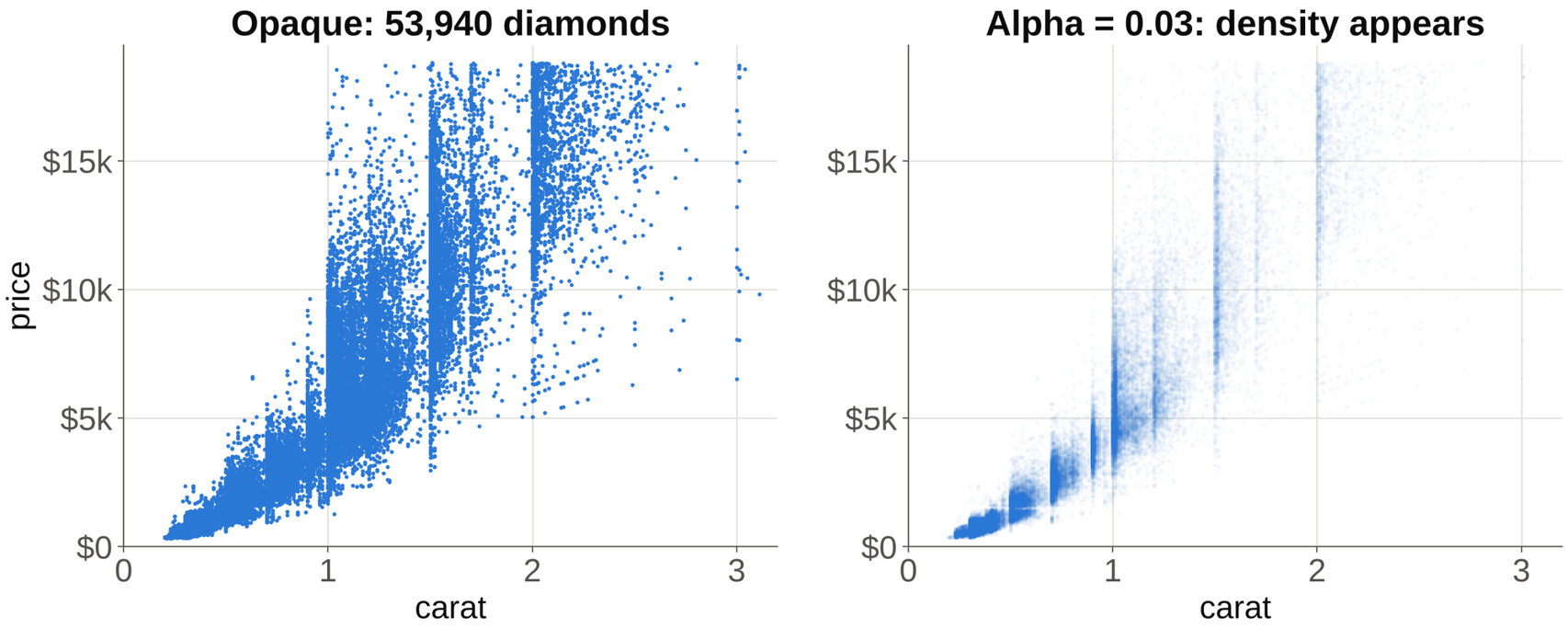


# Conditional Histograms



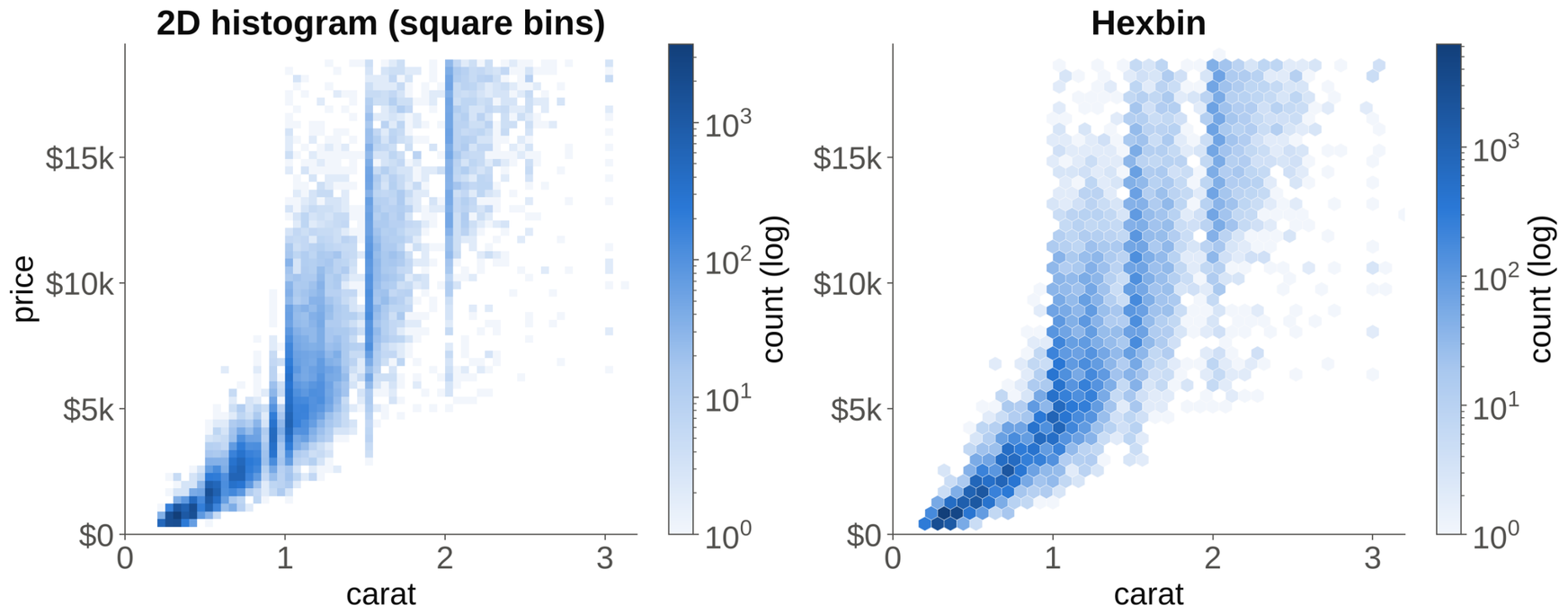
# 2D Data

- Opaque points saturate, so where is most of the data?
- Blending is the cheapest fix, but sparse regions fade away



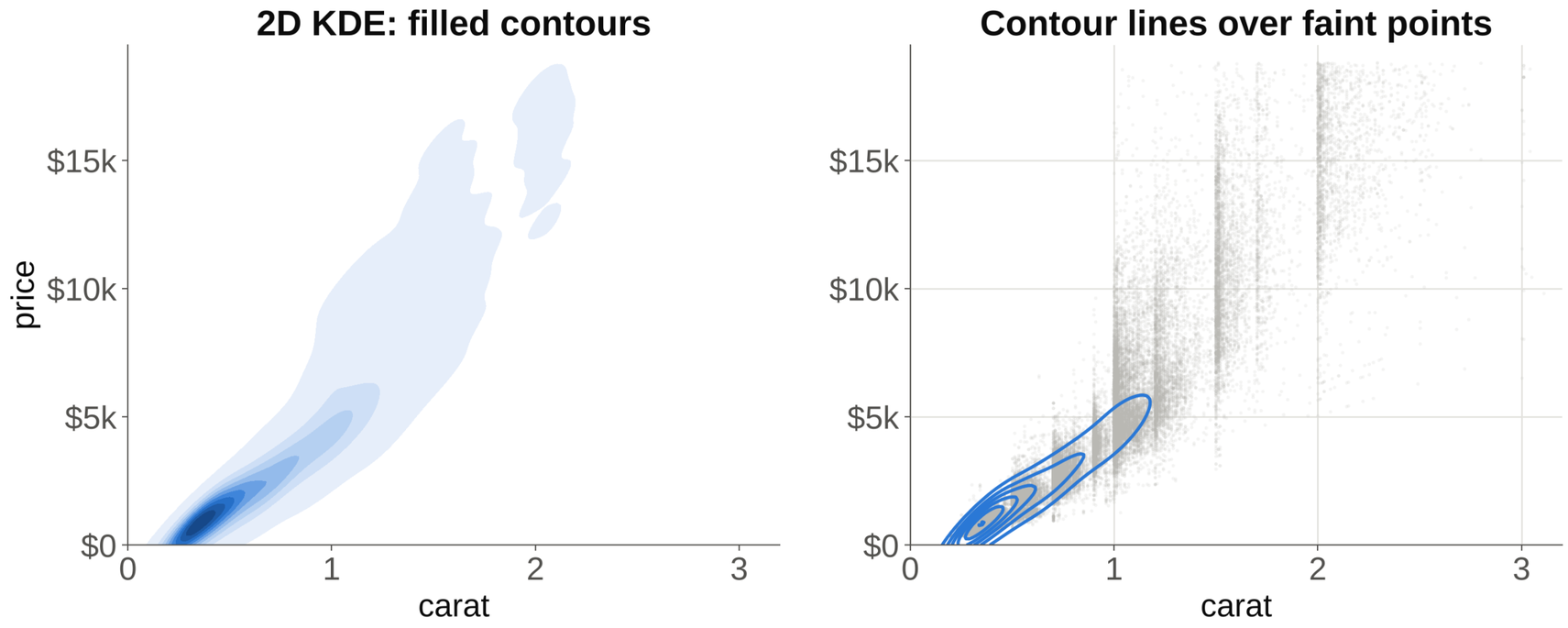
# Aggregate into 2D bins

- Count points per cell; encode the count with a sequential color
- Hexagons: neighbors are equidistant, fewer grid artifacts; use a log scale when counts span 1 to 1,000s

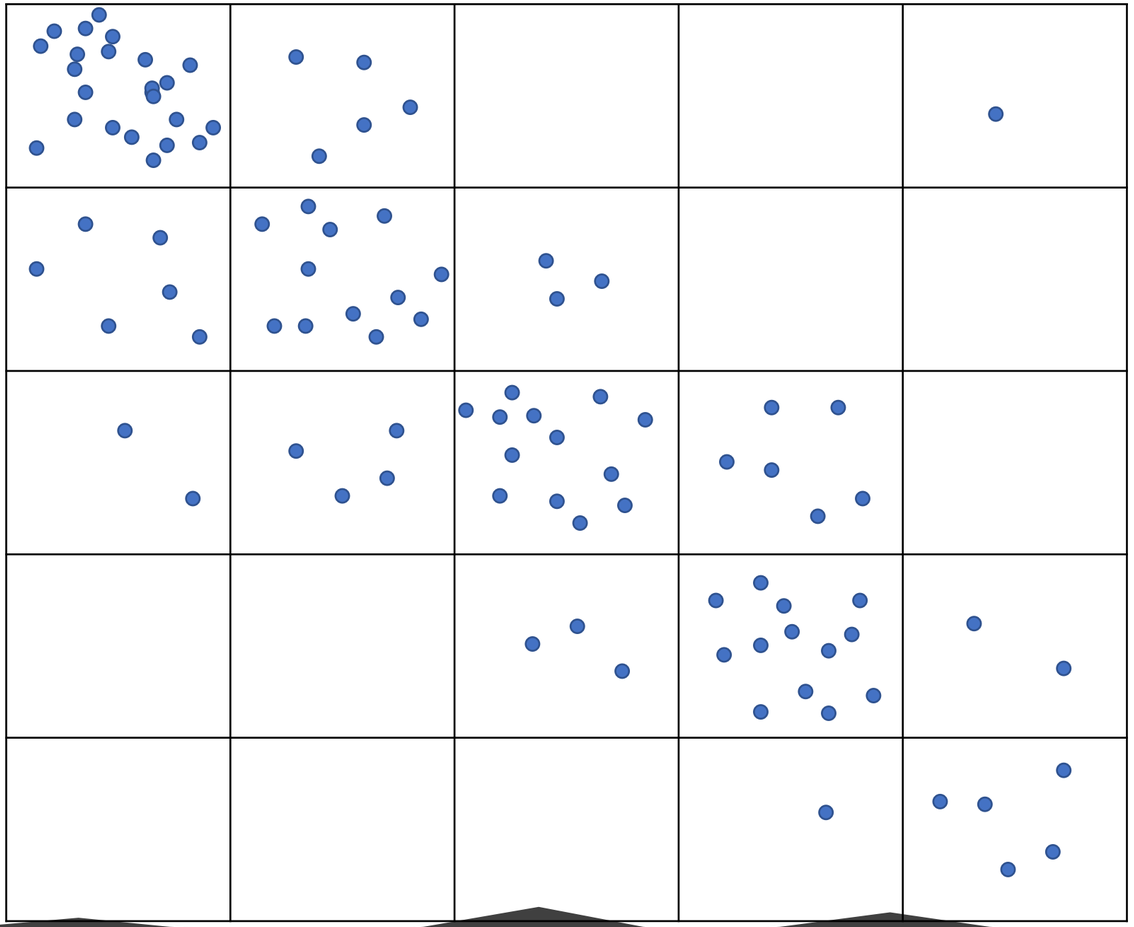


# Smooth instead of bin: 2D density

- A 2D kernel density estimate, drawn as filled regions or contour lines
- Contours over faint points keep the outliers visible



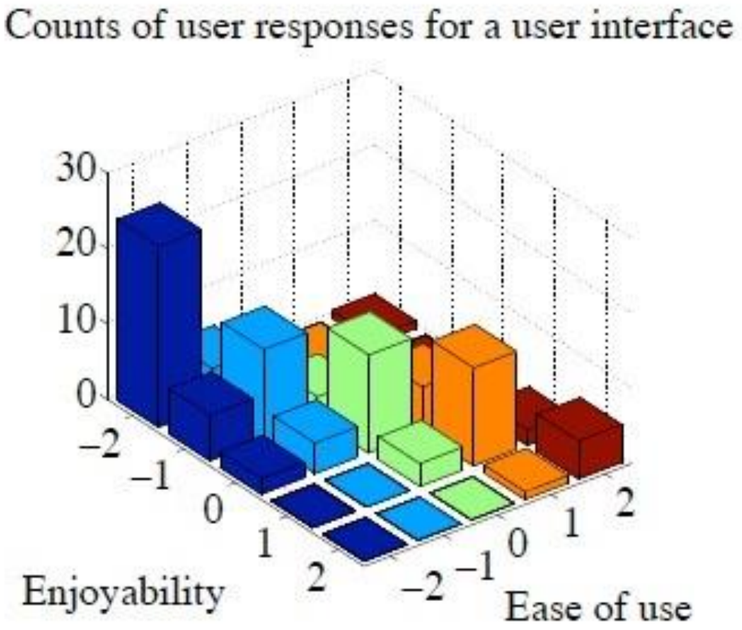
# Ordinal data



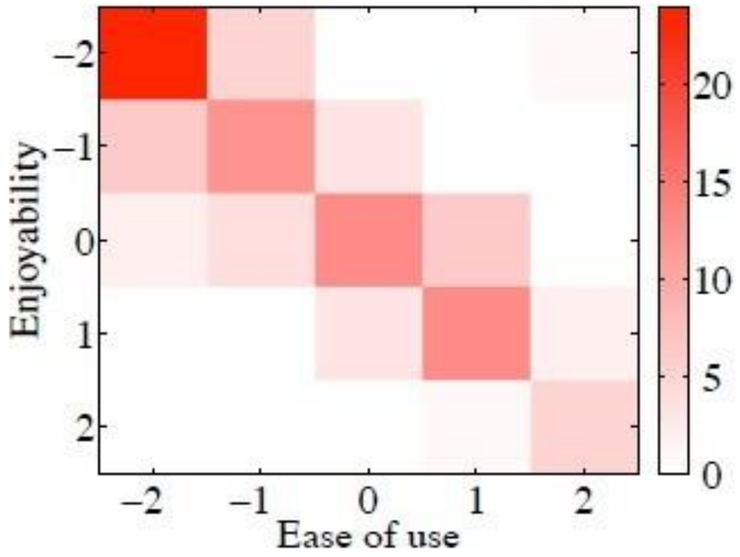
| | -2 | -1 | 0 | 1 | 2 |
|---|---|---|---|---|---|
| -2 | 24 | 5 | 0 | 0 | 1 |
| -1 | 6 | 12 | 3 | 0 | 0 |
| 0 | 2 | 4 | 13 | 6 | 0 |
| 1 | 0 | 0 | 3 | 13 | 2 |
| 2 | 0 | 0 | 0 | 1 | 5 |

# Ordinal data

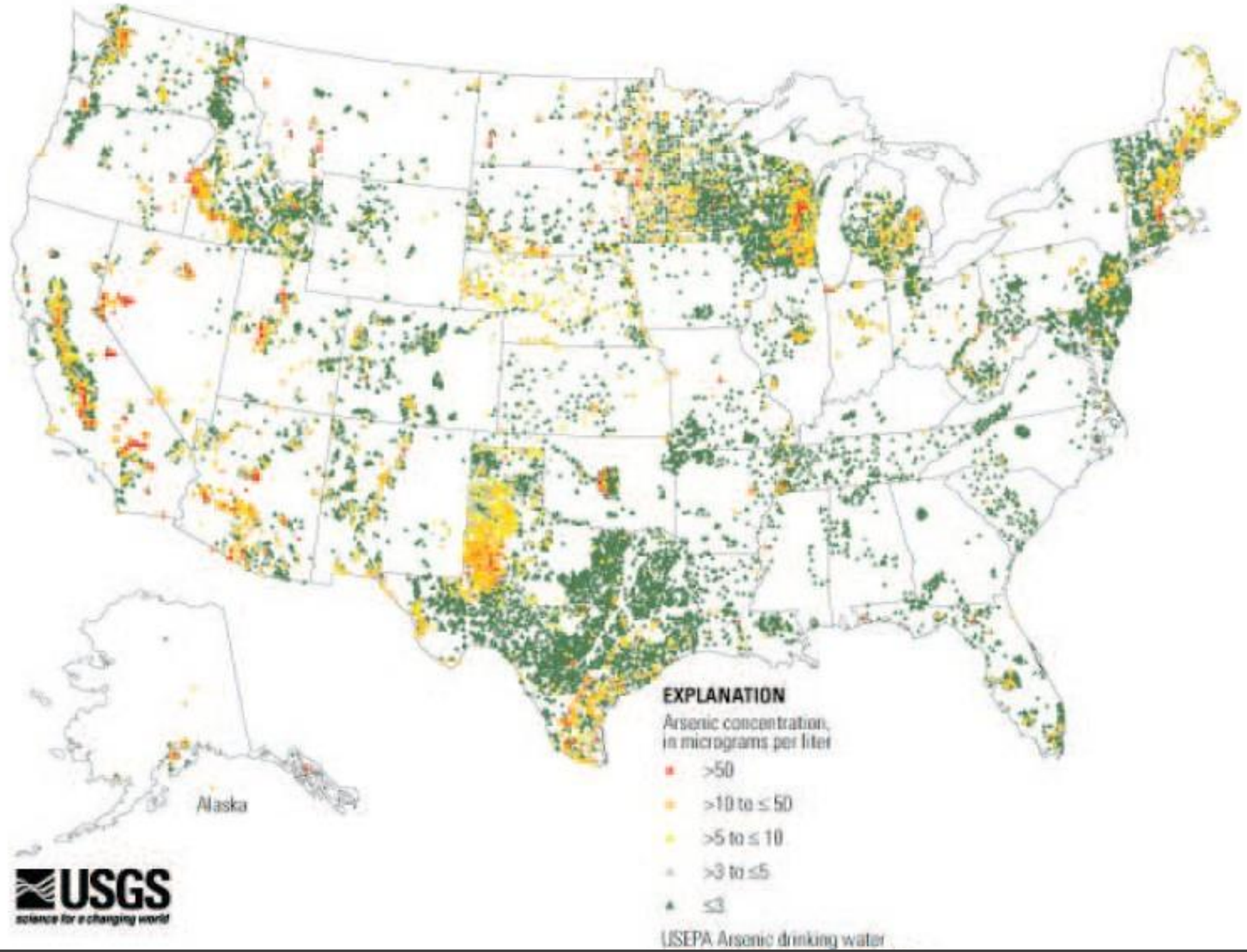
**Don't: 3D bars hide each other and distort heights**



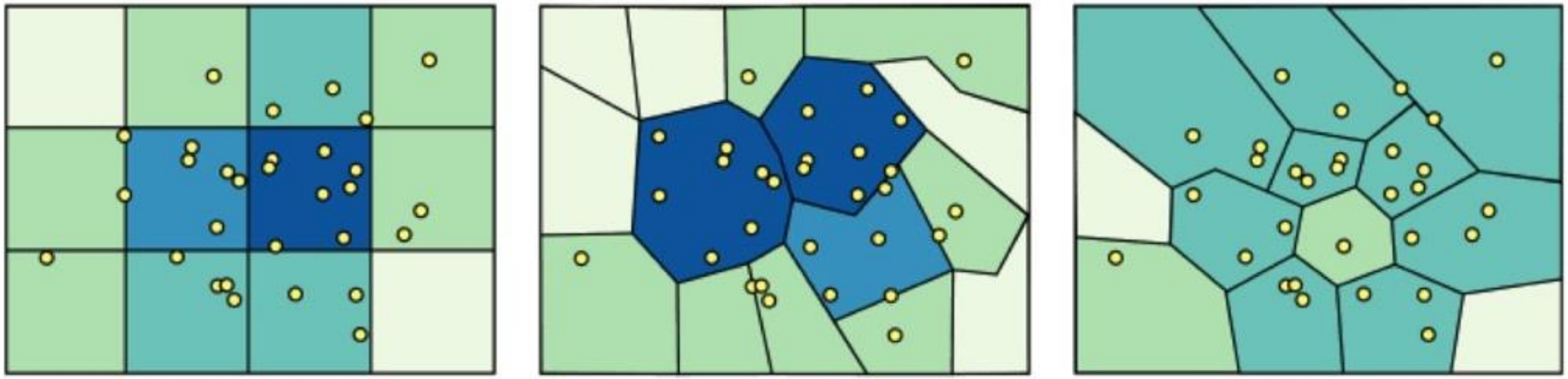

**Do: 2D heatmap, color = count**



# Arsenic in well water

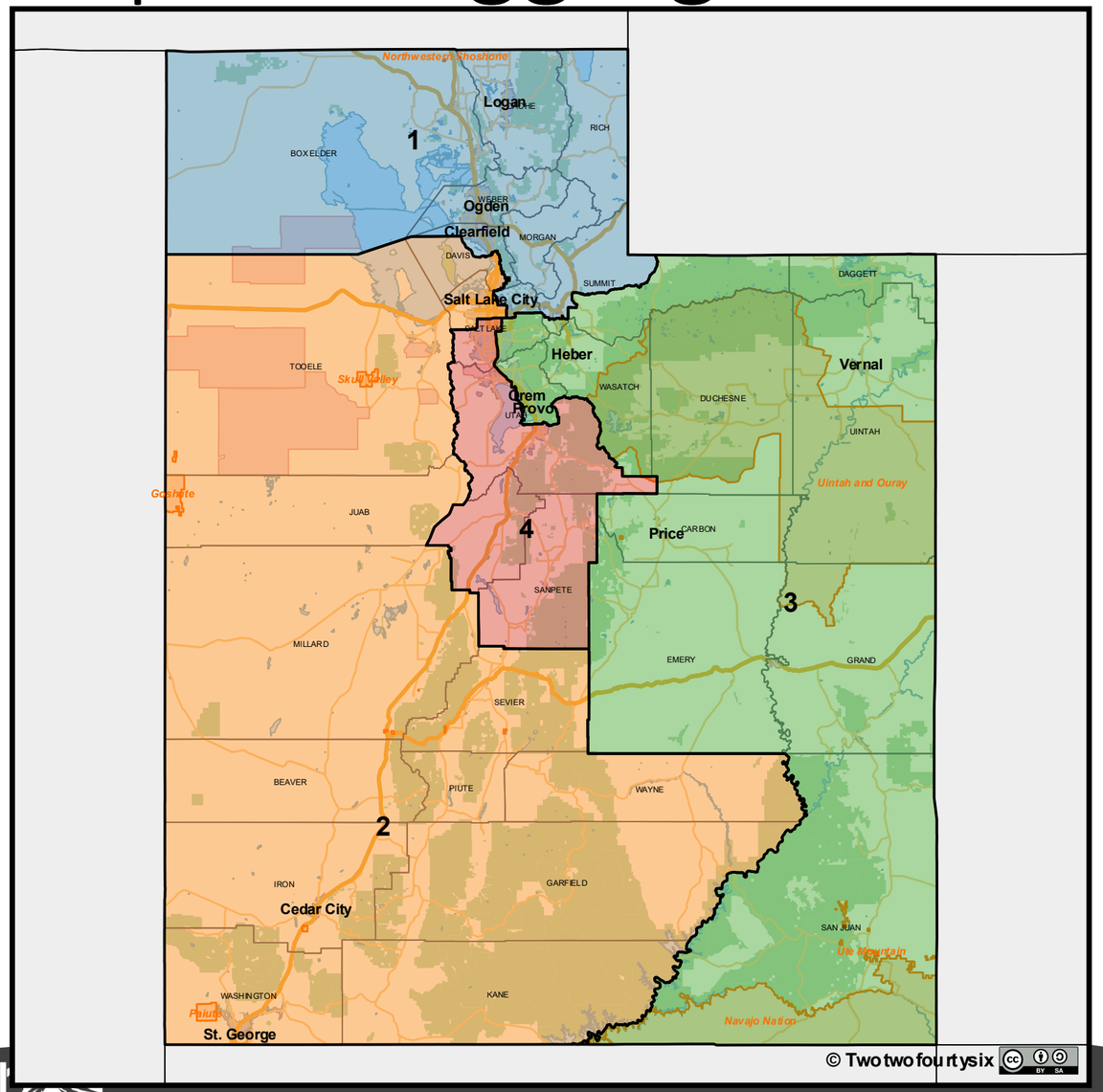


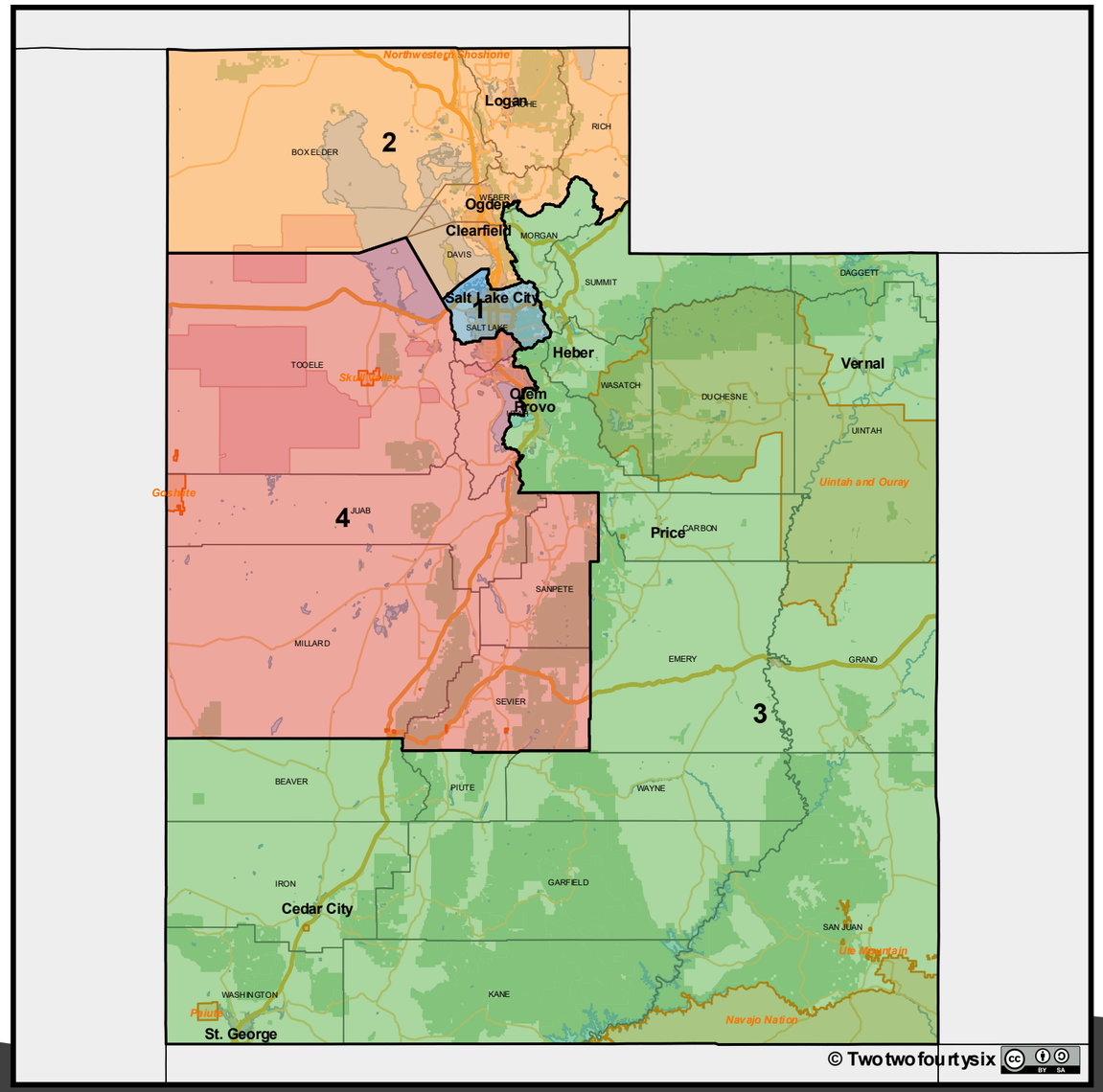
# spatial aggregation



- modifiable areal unit problem
  - in cartography, changing the boundaries of the regions used to analyze data can yield dramatically different results

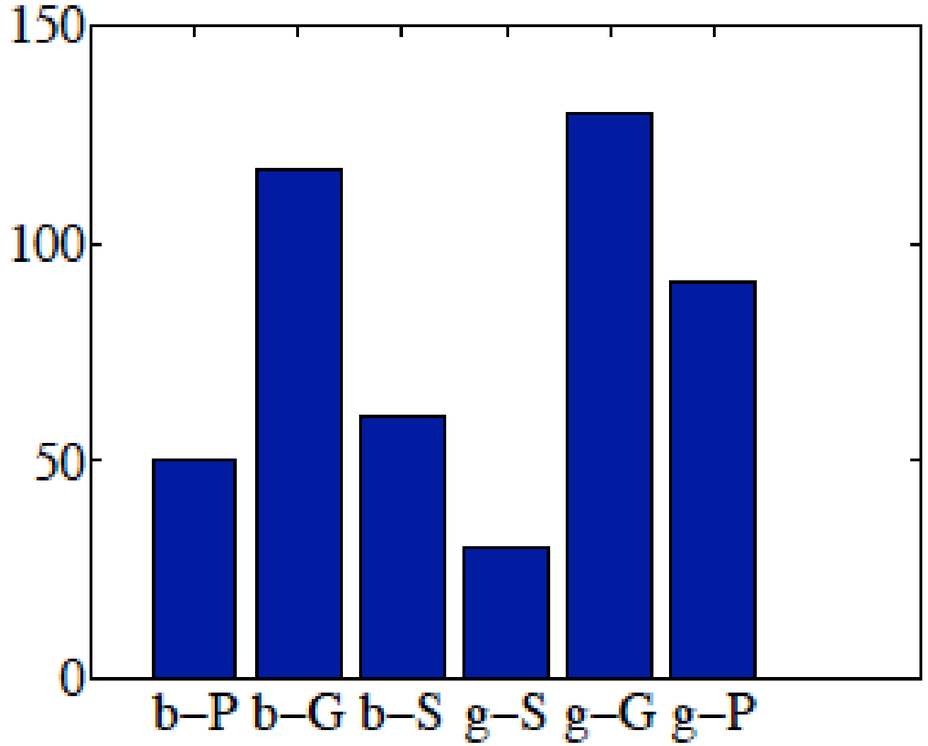
# Spatial Aggregation: Congressional Districts



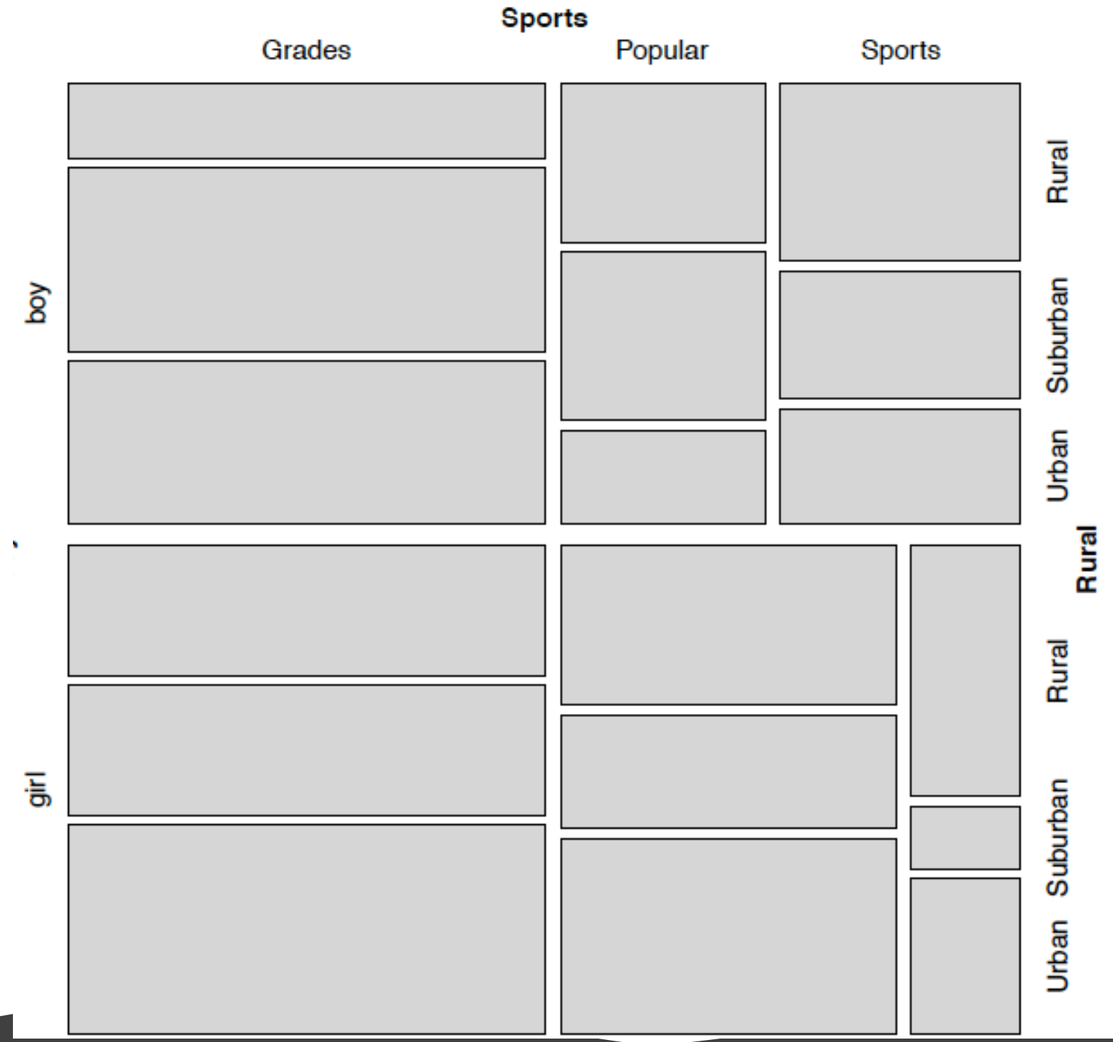



# Categorical data

| Gender | Goal | Gender | Goal |
|---|---|---|---|
| boy | Sports | girl | Sports |
| boy | Popular | girl | Grades |
| girl | Popular | boy | Popular |
| girl | Popular | boy | Popular |
| girl | Popular | boy | Popular |
| girl | Popular | girl | Grades |
| girl | Popular | girl | Sports |
| girl | Grades | girl | Popular |
| girl | Sports | girl | Grades |
| girl | Sports | girl | Sports |

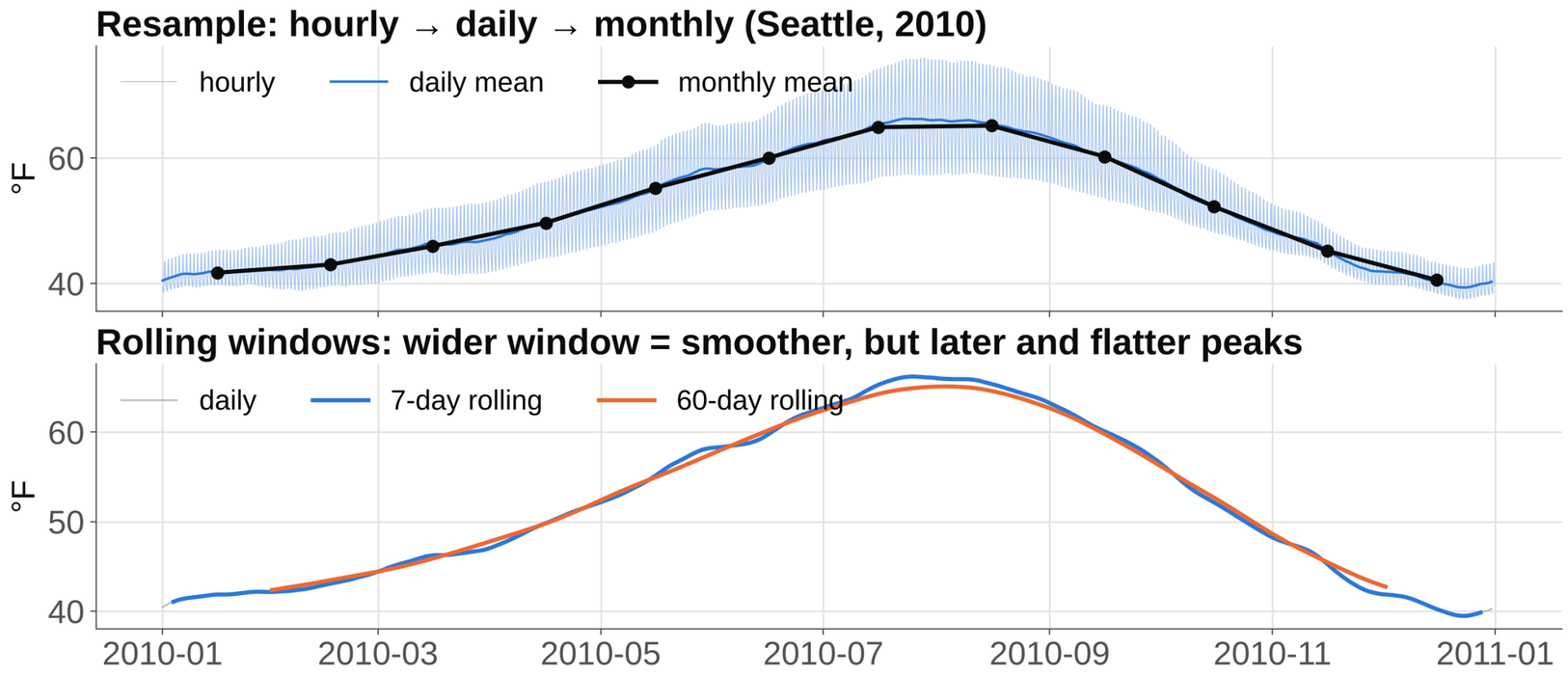


# Mosaic Plots



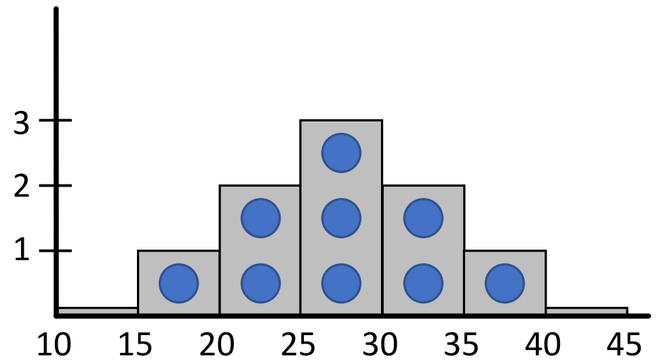
# Temporal aggregation

- Resample to a coarser unit (hour → day → month) with sum, mean, or max
- Rolling windows smooth noise, but wide windows flatten peaks



# Histograms: selecting resolution

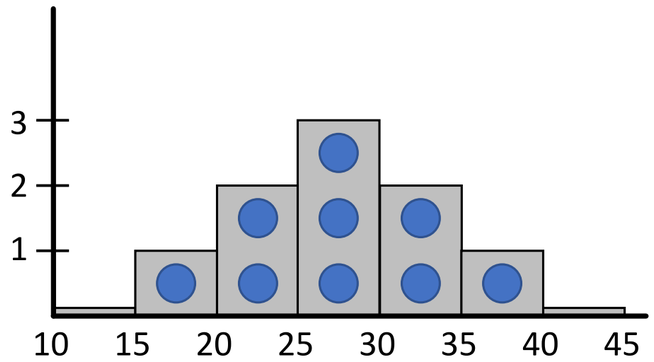
16
27
29
31
26
22
32
36
24

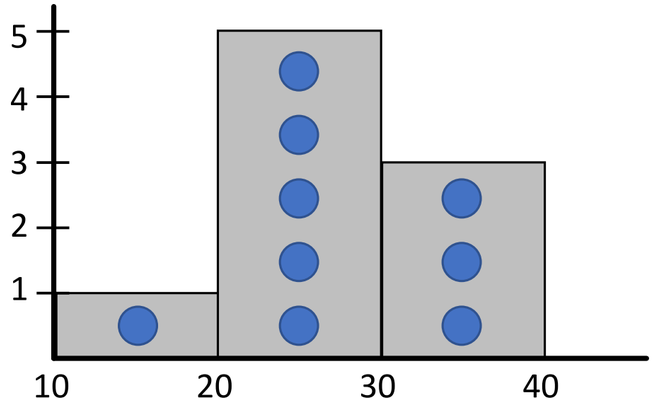


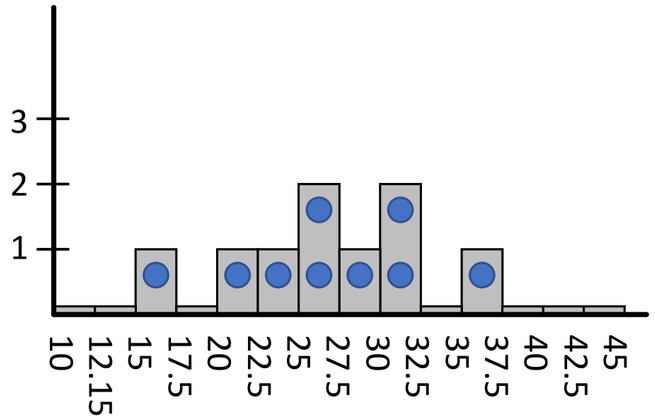

Mean (Average) = 27
Standard Deviation = 6

# Histograms: selecting resolution

16
27
29
31
26
22
32
36
24



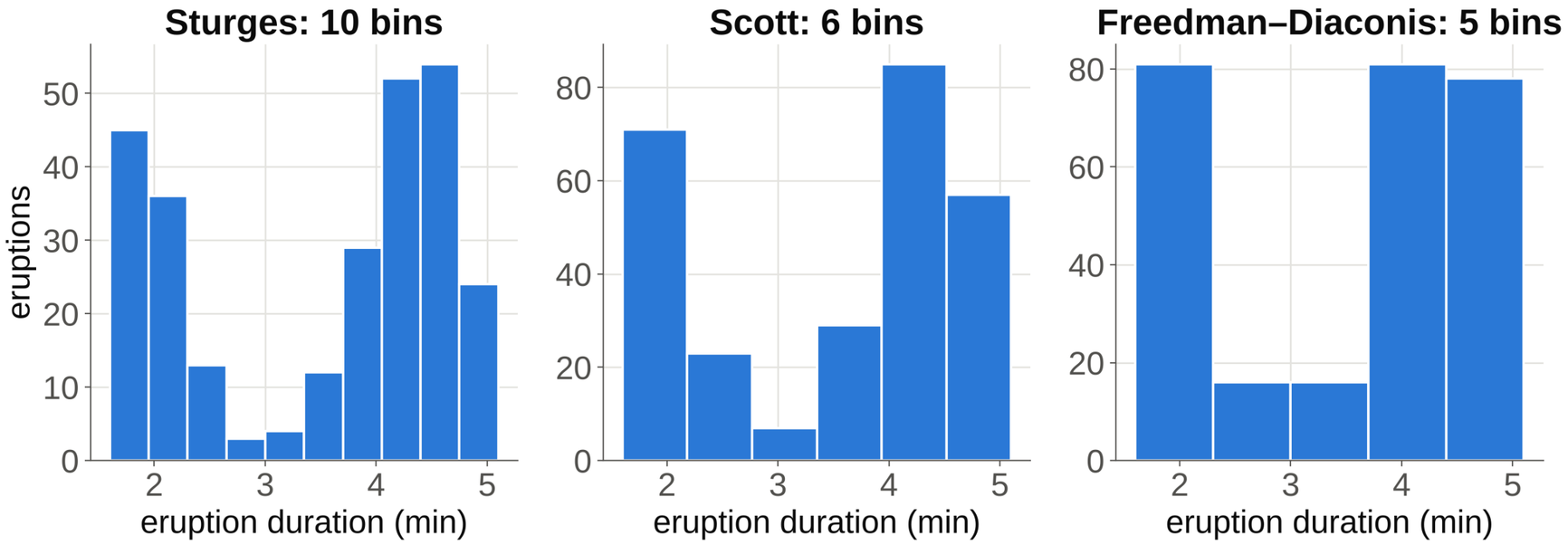




Mean (Average) = 27
Standard Deviation = 6

# Choosing a bin width: rules of thumb

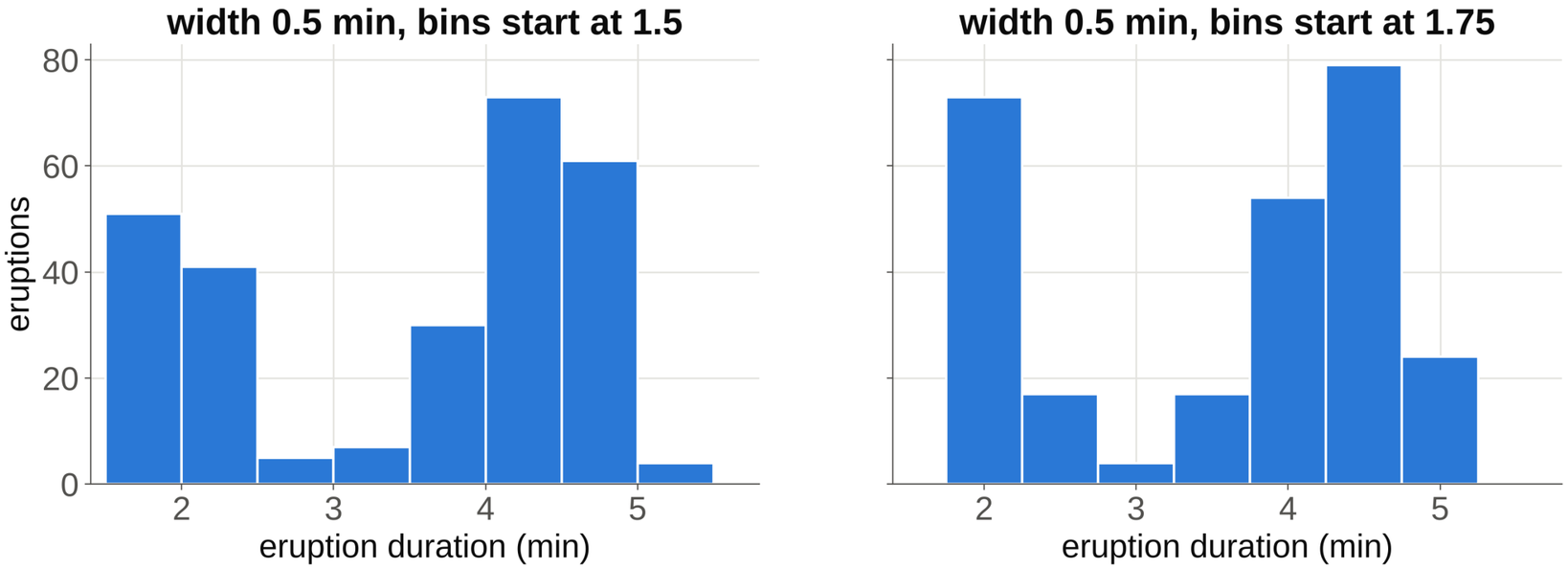
- Sturges: $k = \log_2 n + 1$ bins (assumes roughly normal data)
- Scott: width = 3.49 σ n^(−1/3)
- Freedman–Diaconis: width = 2 IQR n^(−1/3)
- Rules disagree, so treat them as starting points



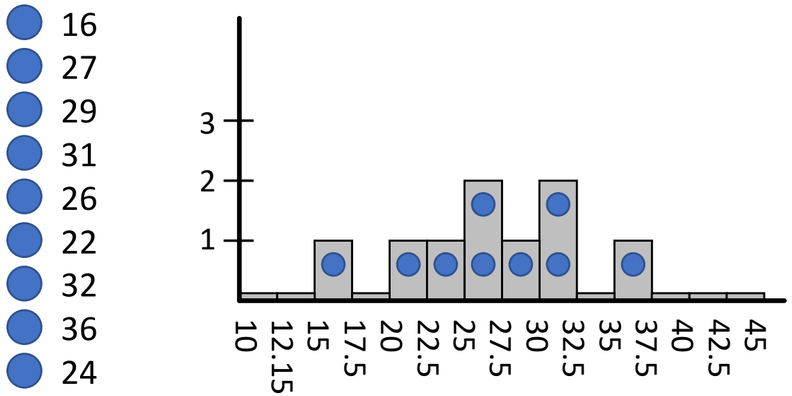

Sturges 1926; Scott 1979; Freedman & Diaconis 1981

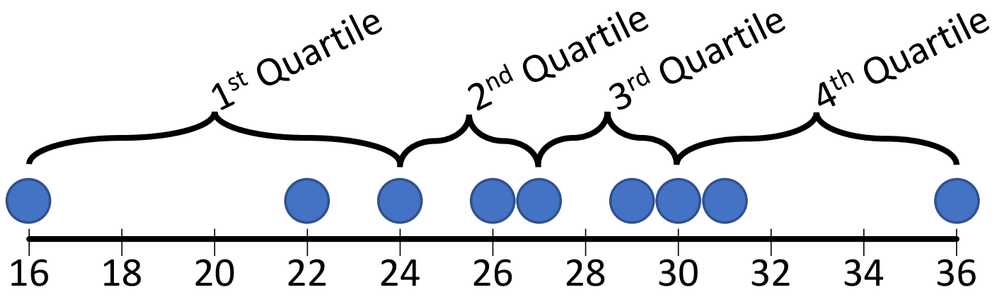
# Bin offset matters too

- Same data, same bin width; only the starting edge moved by 0.25 minutes
- The left peak looks twice as tall: try several offsets, or use a density estimate

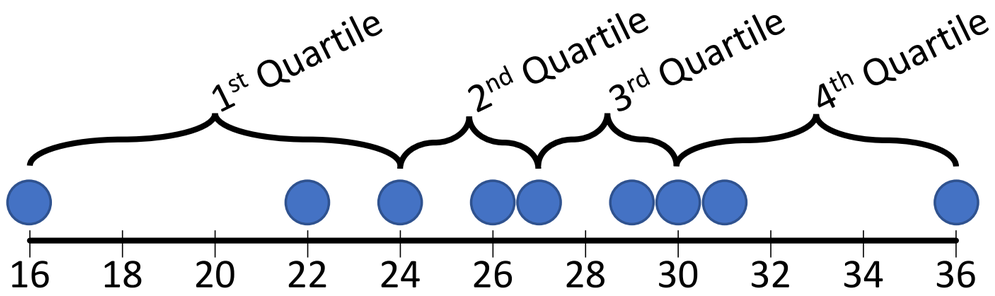


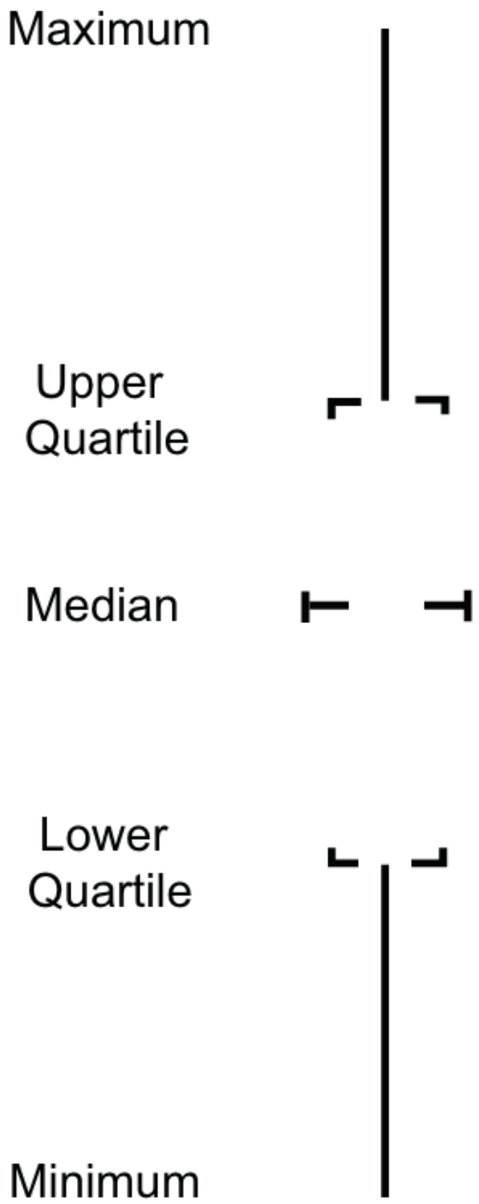
# Using distributions instead of histograms



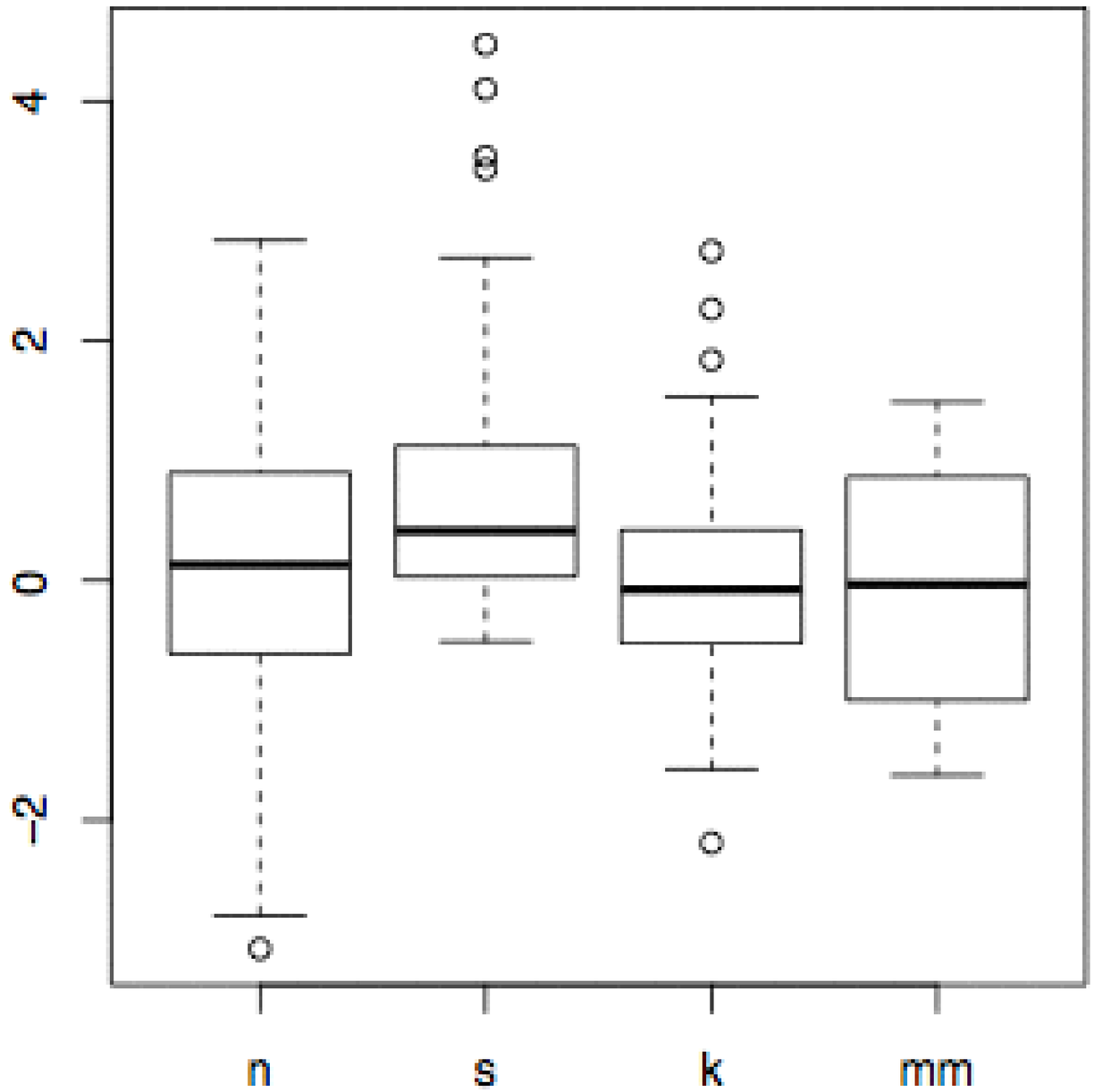
# Boxplot



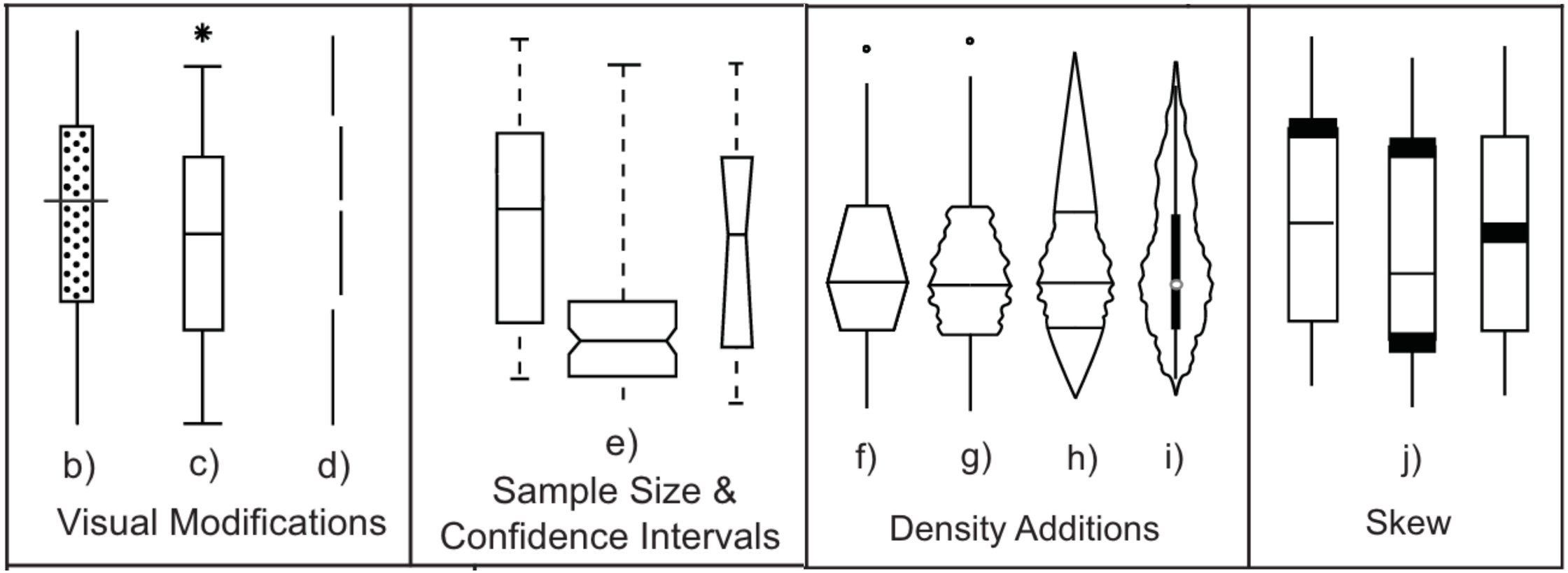




Potter et al., 2010

# Boxplot

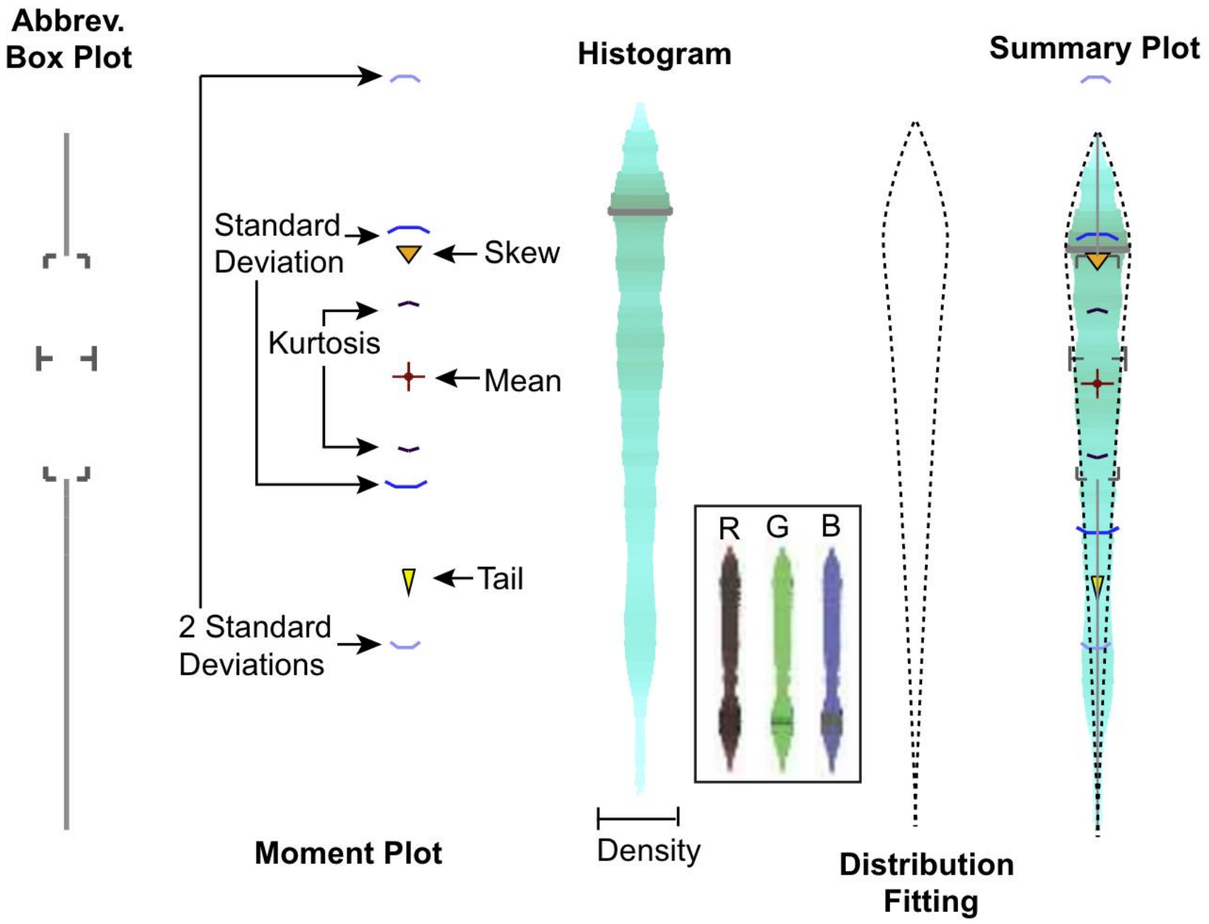


# Boxplots




Potter et al., 2010

# Boxplots




Given a data set $\{x_i\}_{i=1}^N$, we define the following quantities:

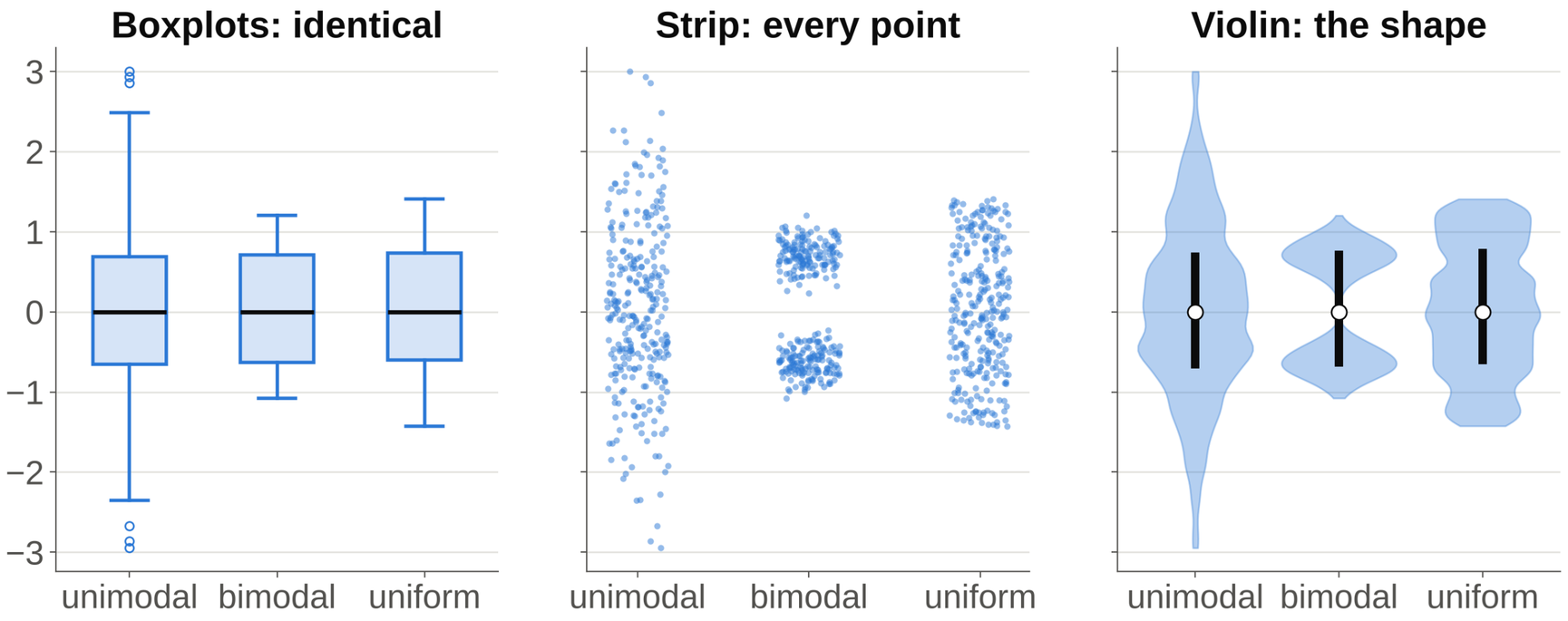
| | |
|---|---|
| $k$th Central Moments: | $\mu_k \simeq \frac{1}{N}\sum_{i=1}^N (x_i - \mu_1)^k$ |
| Mean: | $\mu_1 \simeq \frac{1}{N}\sum_{i=1}^N x_i$ |
| Variance: | $\mu_2 \simeq \frac{1}{N}\sum_{i=1}^N (x_i - \mu_1)^2$ |
| Standard Deviation: | $\sigma = \sqrt{\mu_2}$ |
| Skew: | $\gamma = \frac{\mu_3}{\sigma^3}$ |
| Kurtosis: | $\kappa = \frac{\mu_4}{\sigma^4}$ |
| Excess Kurtosis: | $\kappa_e = \kappa - 3$ |
| Tailing: | $\tau = \frac{\mu_5}{\sigma^5}$ |

where $N$ is the number of data samples.

Potter et al., 2010

# Boxplots can hide the data

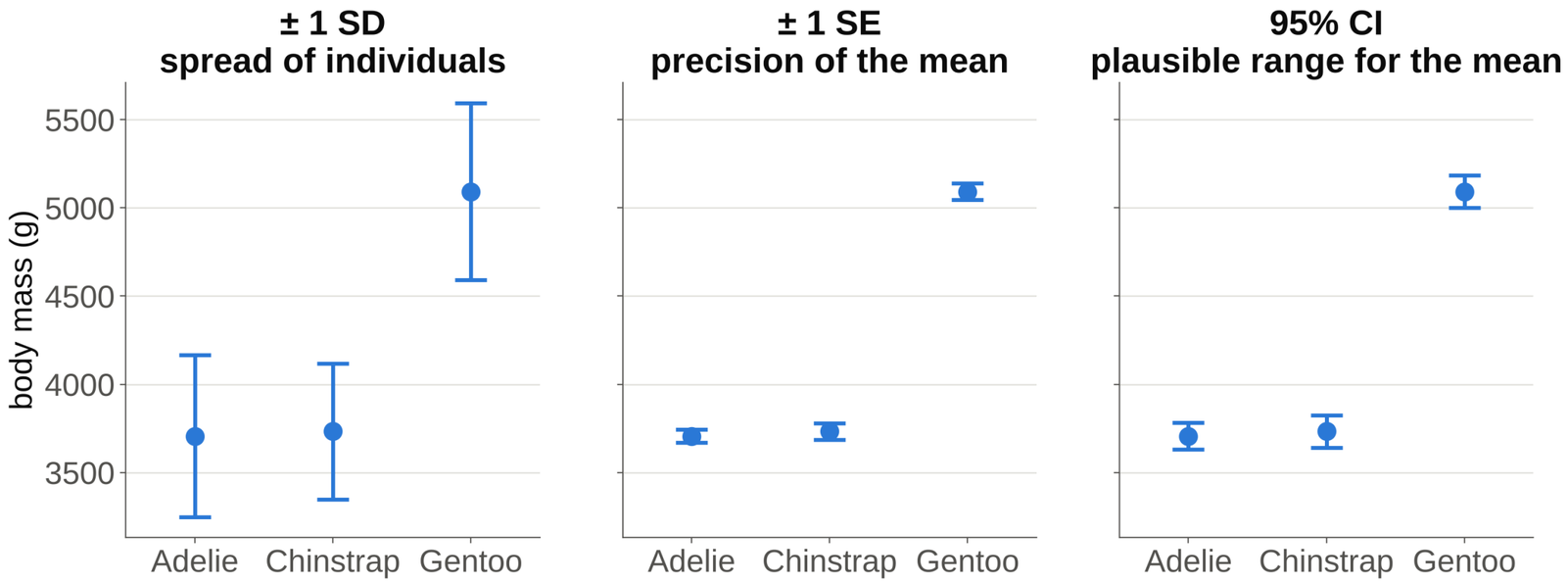
- Three simulated groups with the same median and quartiles
- When n allows, show the points (strip, beeswarm) or the shape (violin, raincloud)




Matejka & Fitzmaurice 2017; Weissgerber et al. 2015

# Error bars: say which one

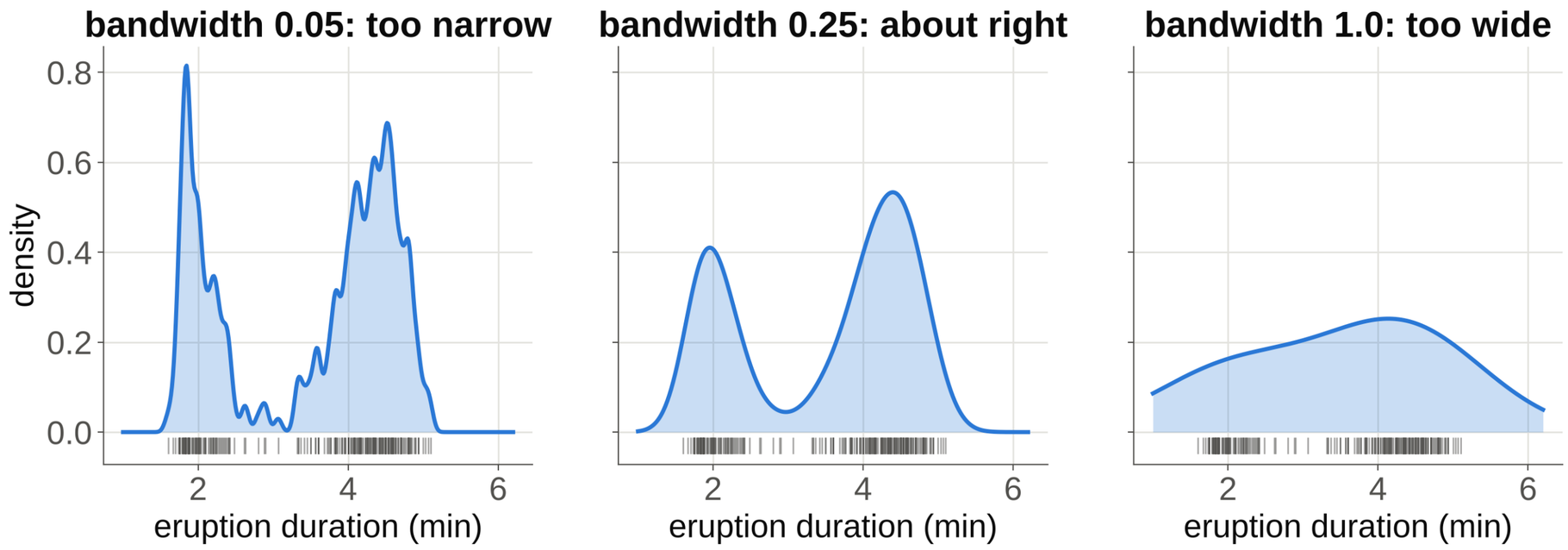
- Standard Deviation (SD) describes individuals
- Standard Error (SE) and Confidence Intervals (CI) describe the mean




Cumming & Finch 2005; Palmer penguins data (Gorman et al. 2014)

# Kernel density estimation (KDE)

- Put a small bump (kernel) on every point and add them up
  - a smooth histogram with no bin edges
- Bandwidth plays the role of bin width: too narrow is noisy, too wide hides the two modes




Old Faithful geyser data (Azzalini & Bowman 1990)

# Problem #2: Aggregate Attributes
# We have too many attributes to show

# attribute aggregation

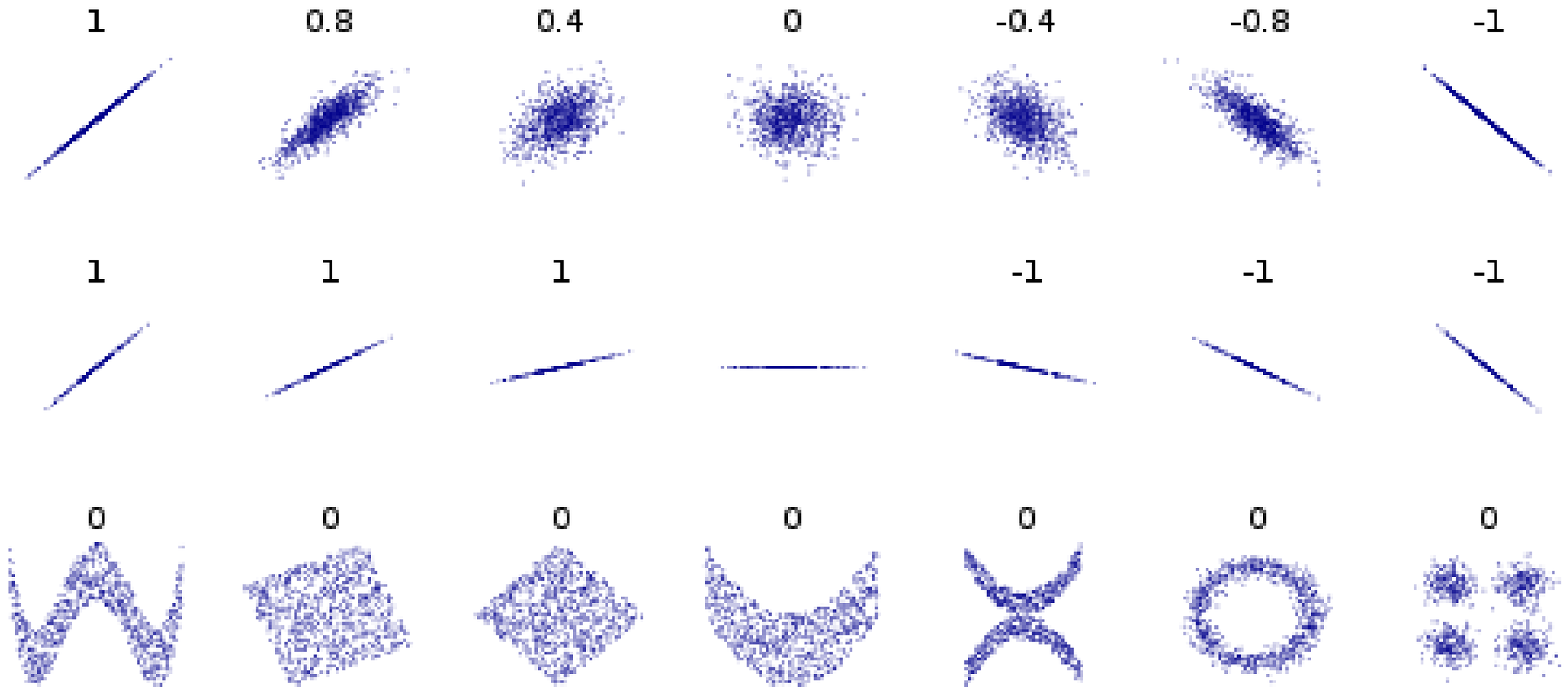
- group attributes and compute a similarity score across the set
- dimensionality reduction to preserve meaningful structure

# Similarity scores

- correlation
  - measure of similarity between 2 or more attributes
  - many variants—pearson, rank, multi-way, etc.
- regression
  - fit a model to the data
  - measure the quality of fit (i.e. $R^2$)
- dimension reduction
  - Use information from all attributes to draw data points in low dimension

# Pearson Correlation Coefficient

- A measure of the linearity between 2 sets

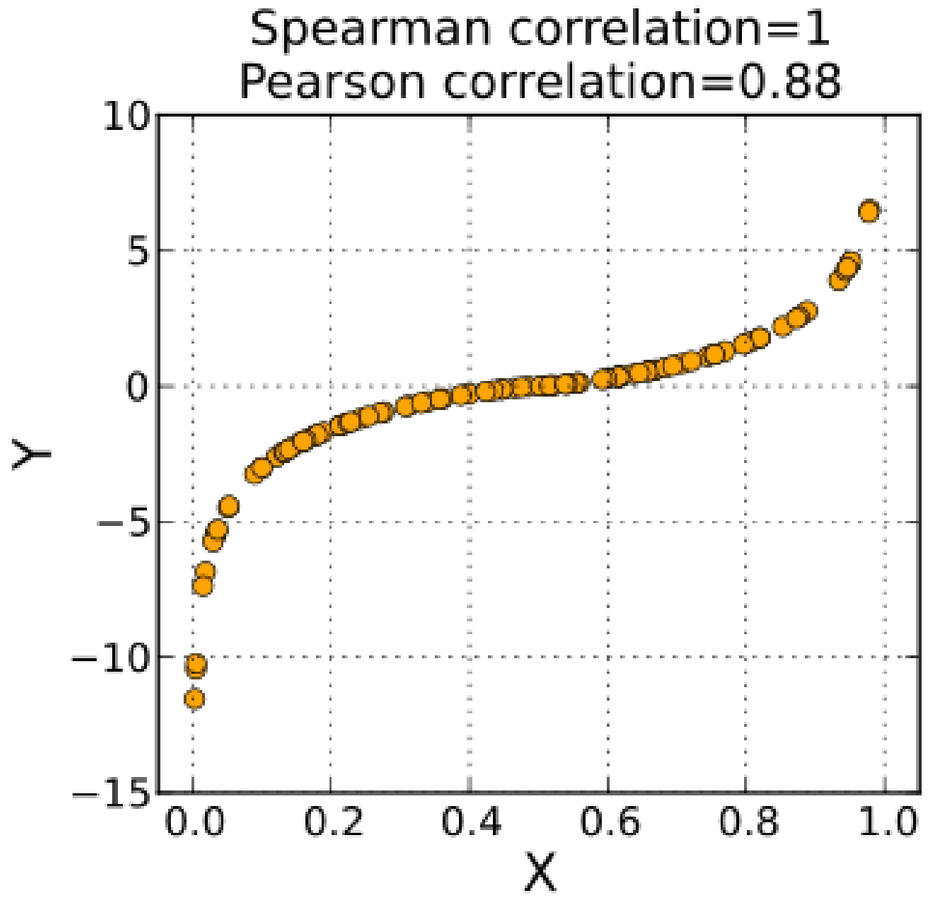


$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- $\operatorname{cov}$ is the covariance
- $\sigma_X$ is the standard deviation of $X$
- $\sigma_Y$ is the standard deviation of $Y$
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

# Spearman Rank Correlation



# Spearman Rank Correlation

- Non-parametric correlation measurement

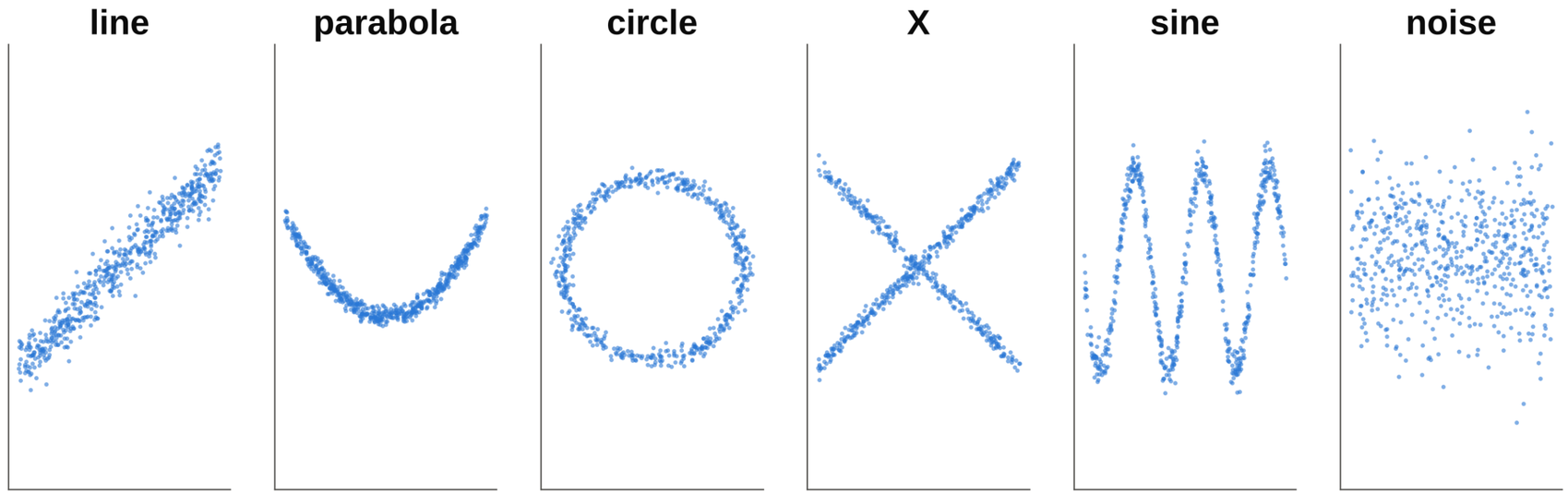
- sort(X) and sort(Y)
- assign X’/Y’ rank in sorted list
- Calculate PCC( X’, Y’ )

# Spearman Rank Correlation

| IQ, (X) | Hours of TV per week, (Y) | rank (X') | rank (Y') |
|---|---|---|---|
| 86 | 0 | 1 | 1 |
| 97 | 20 | 2 | 6 |
| 99 | 28 | 3 | 8 |
| 100 | 27 | 4 | 7 |
| 101 | 50 | 5 | 10 |
| 103 | 29 | 6 | 9 |
| 106 | 7 | 7 | 3 |
| 110 | 17 | 8 | 5 |
| 112 | 6 | 9 | 2 |
| 113 | 12 | 10 | 4 |

# Beyond Pearson and Spearman

- Pearson r and Spearman ρ only detect straight-line or monotonic trends
- Distance correlation (dCor) and mutual information (MI) are 0 only for independent variables, so they catch curves, rings, and crosses



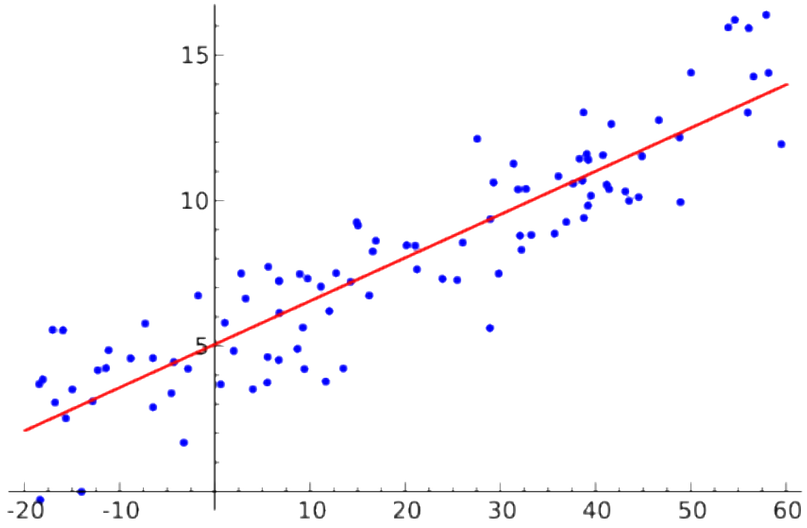

Székely et al. 2007 (dCor); Kraskov et al. 2004 (MI); simulated data

# Regression: Fitting a Model to Data

- Given: $y_i = \alpha + \beta x_i + \varepsilon_i$
- Find $\alpha$ and $\beta$ that minimize $\varepsilon_i$ in the linear least squares sense (i.e. $\Sigma \varepsilon_i^2$)
- Can be computed directly

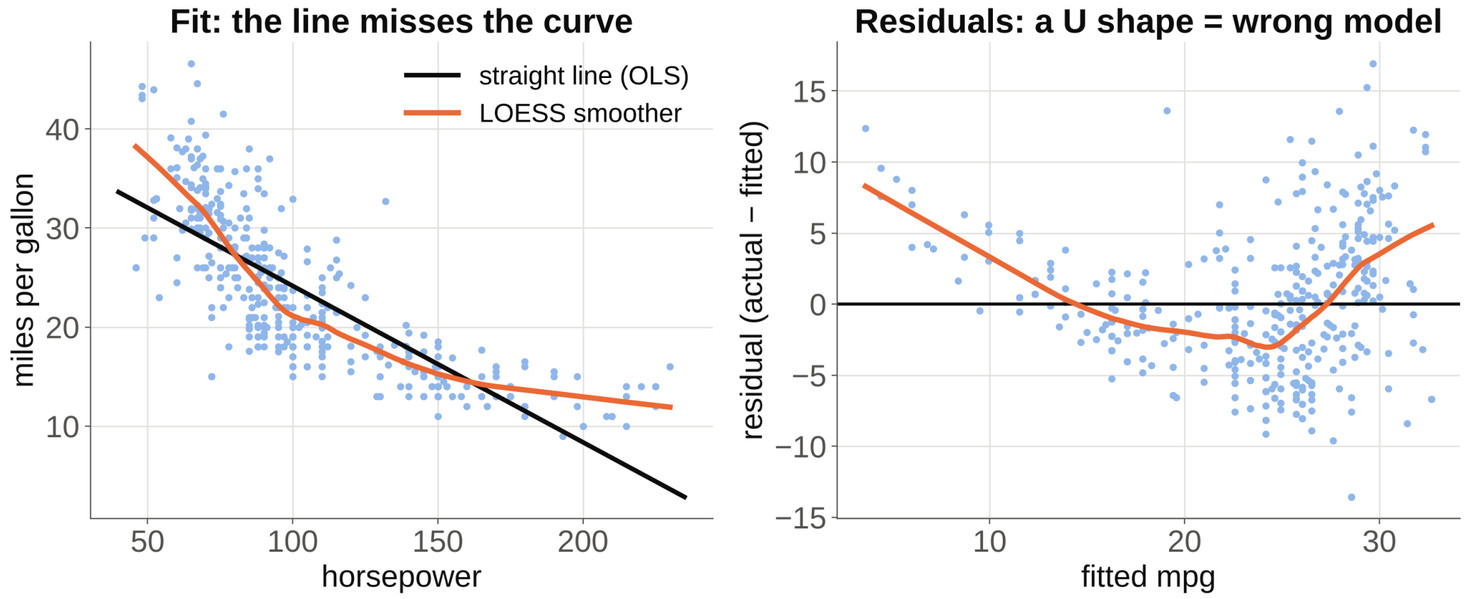
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$



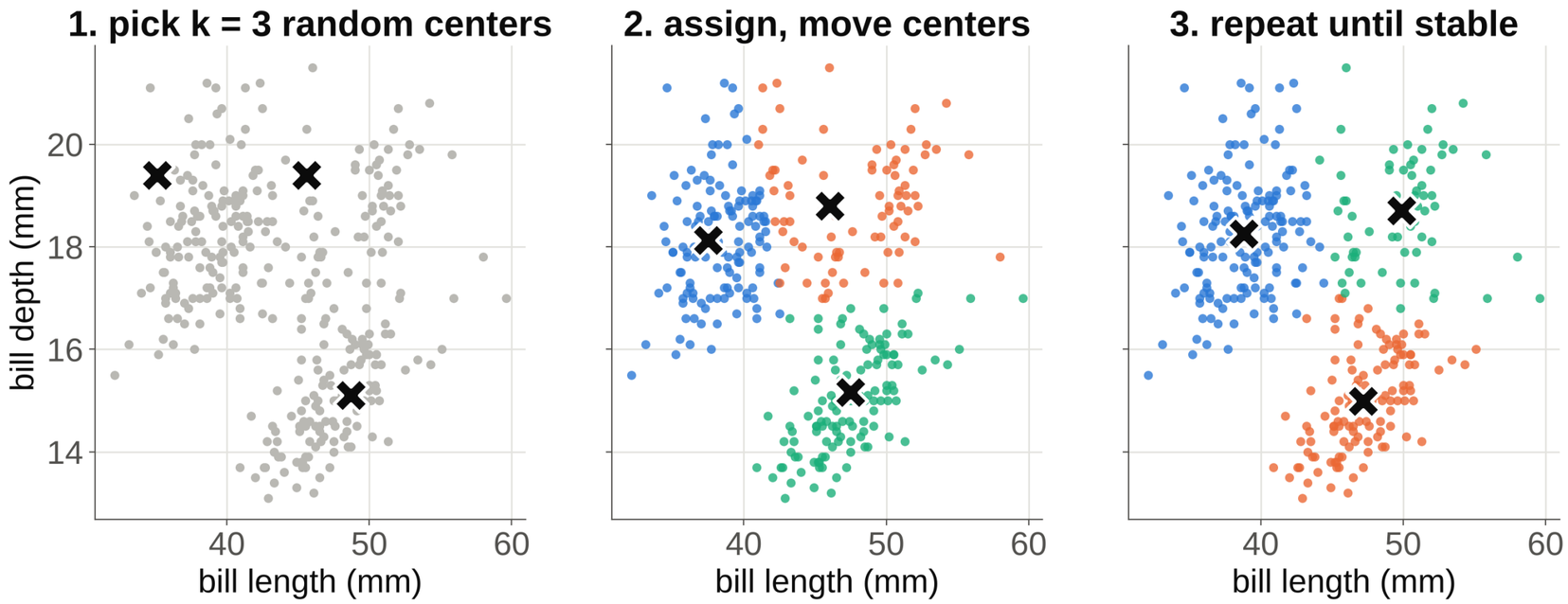
# Check the fit: residuals and smoothers

- Plot residuals (actual − fitted) against fitted values: they should look like noise
- A LOESS smoother follows the data locally and shows the shape a straight line misses



# Clustering: grouping similar items

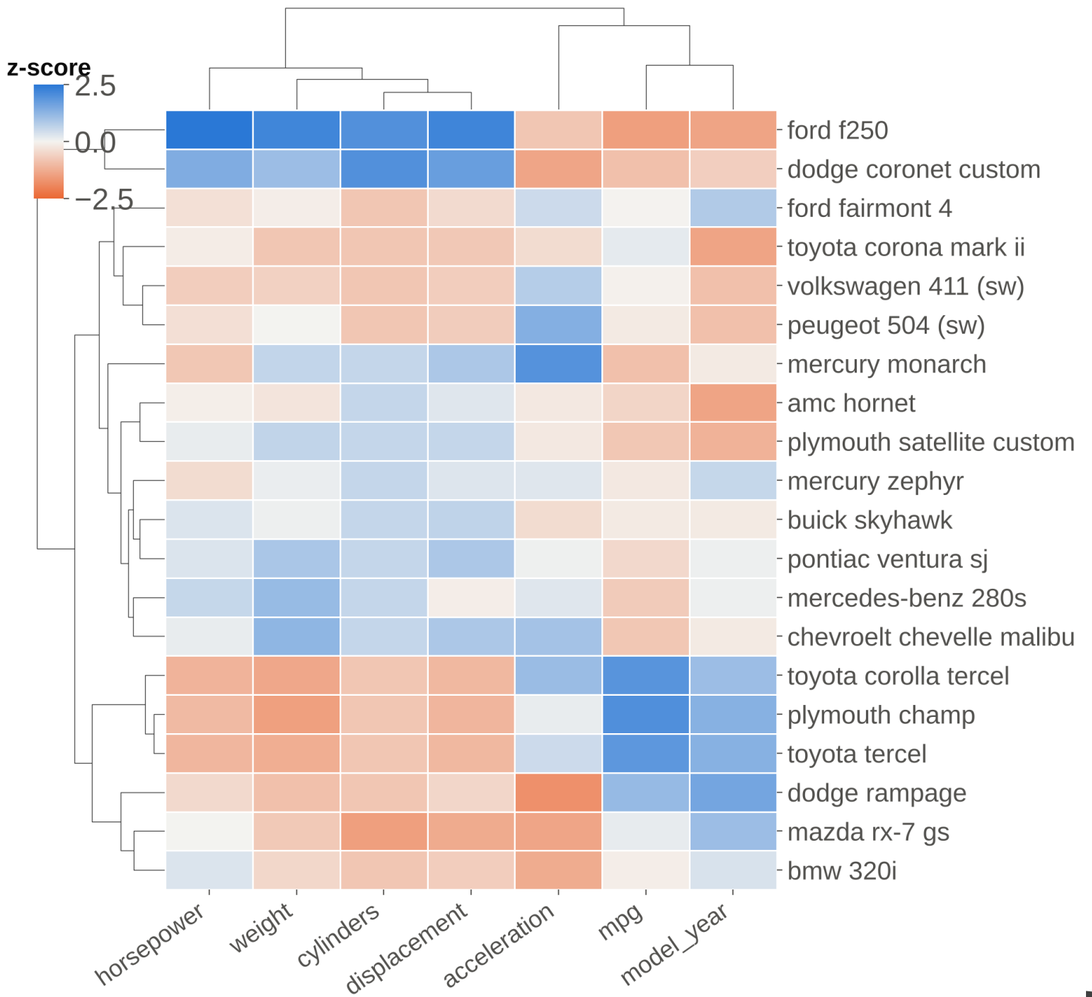
- k-means: pick k centers, assign each point to the nearest, move each center to its points' mean, repeat
- Each cluster becomes one derived item; you choose k, and should standardize attributes first




Lloyd 1982; Palmer penguins data (Gorman et al. 2014)

# Hierarchical clustering

- Repeatedly merge the two most similar items; the dendrogram records every merge
- Reorder rows and columns by the tree so similar items sit together
- 20 cars × 7 standardized attributes: big, heavy cars vs. small, efficient ones
- Cut the tree at any height to get k clusters



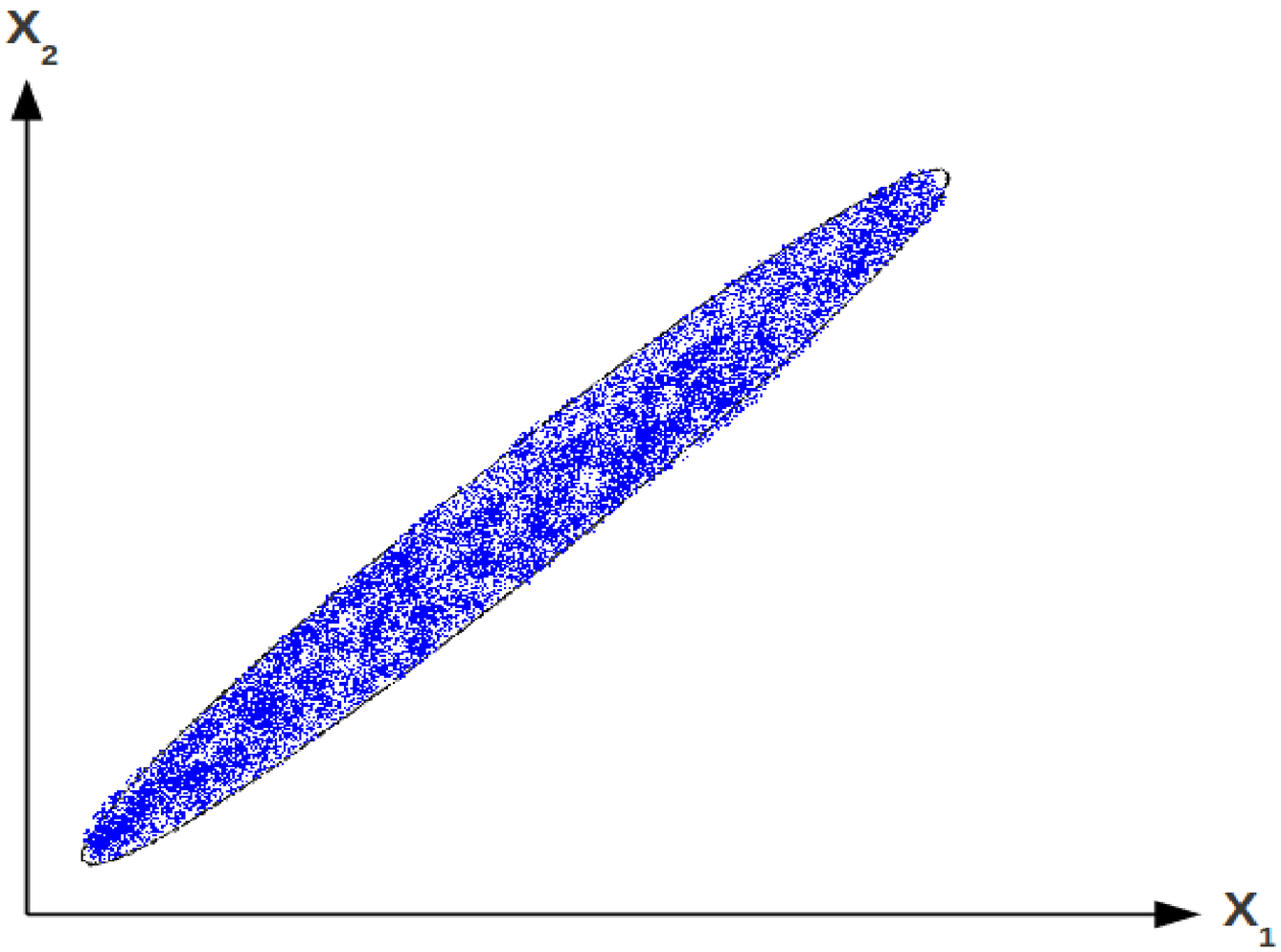
# Dimensionality Reduction

- Takes data with many attributes and represents them in a lower dimension
- Comes in 2 basic flavors, linear and nonlinear, with lots of subvariants

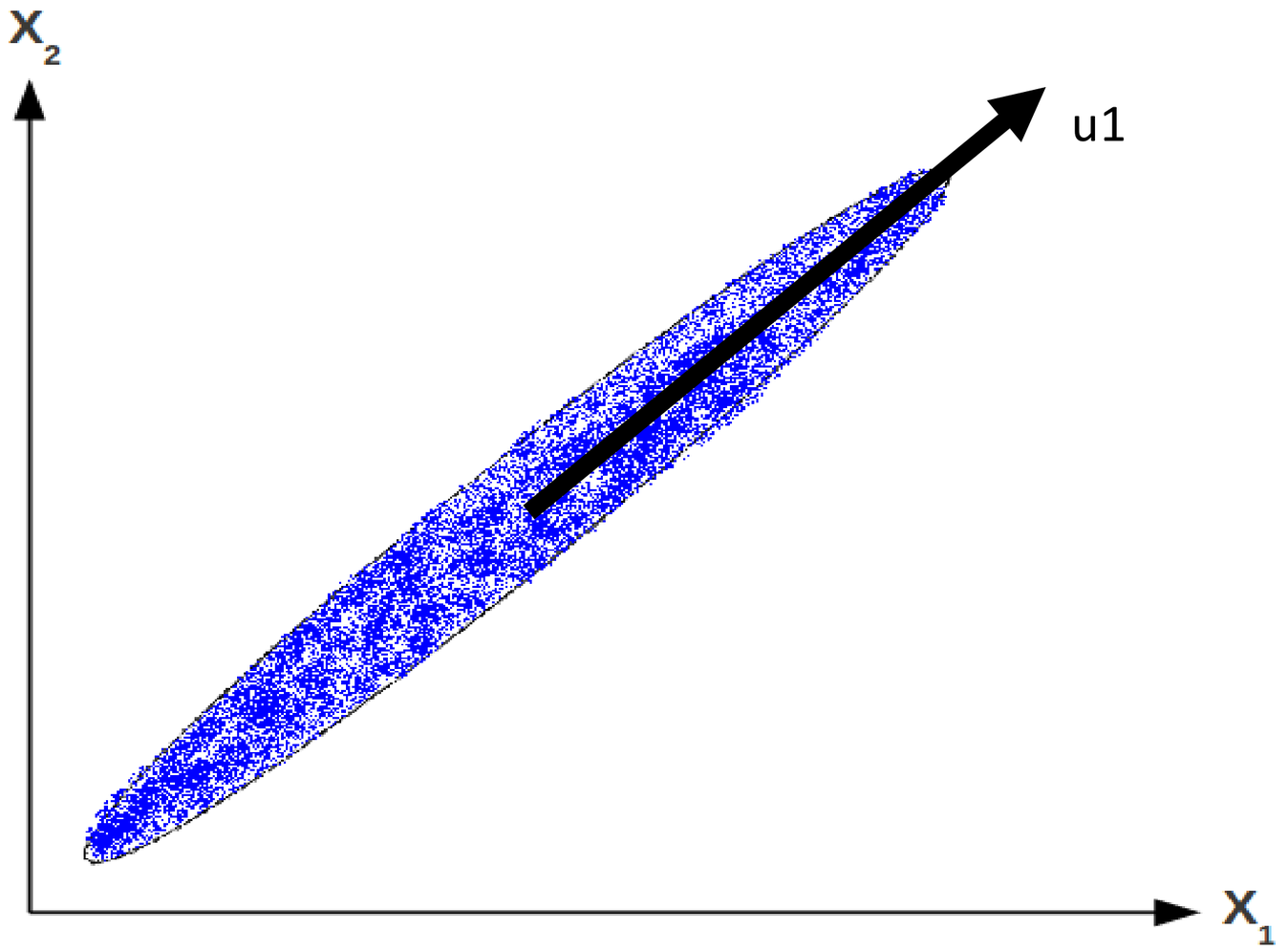
# Principal Component Analysis (PCA)

- Given a set of data points embedded in any number of dimensions
- Find the vector (direction) with the largest variance
- Repeat for additional vectors, finding the next largest variance orthogonal to all previous principal components

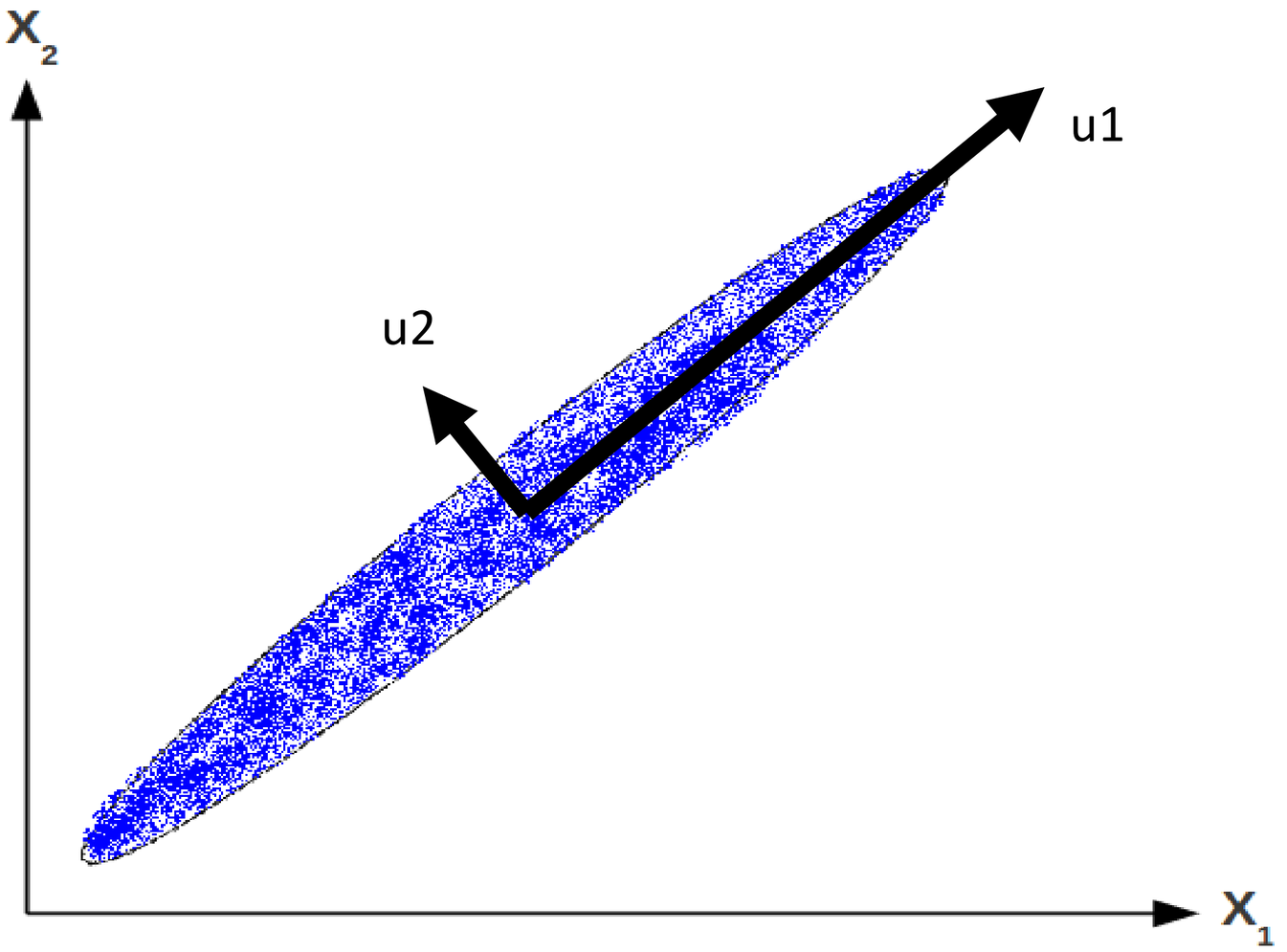
# PCA: an example



# PCA: an example



# PCA: an example

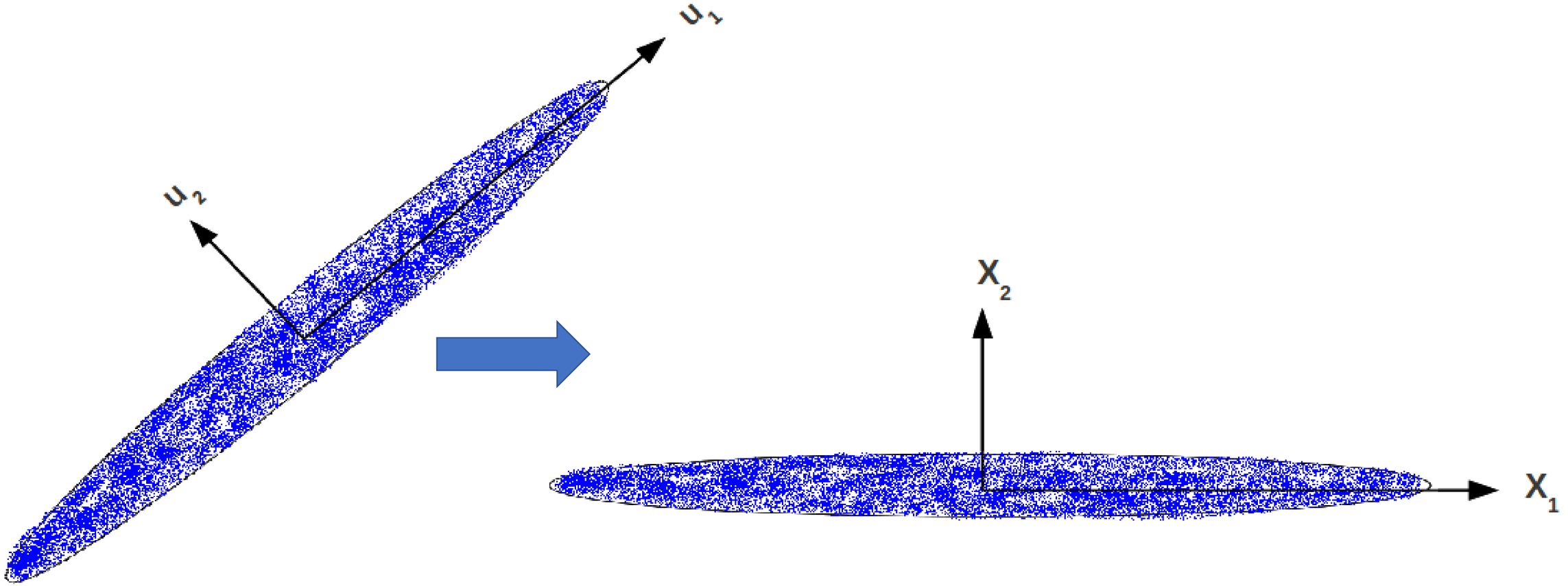


# PCA Provides

- A new basis that maximizes variance, where the length of the vectors indicates variance in that direction

- Common practice is to reproject points by subtracting the mean and adjusting basis

# PCA: an example

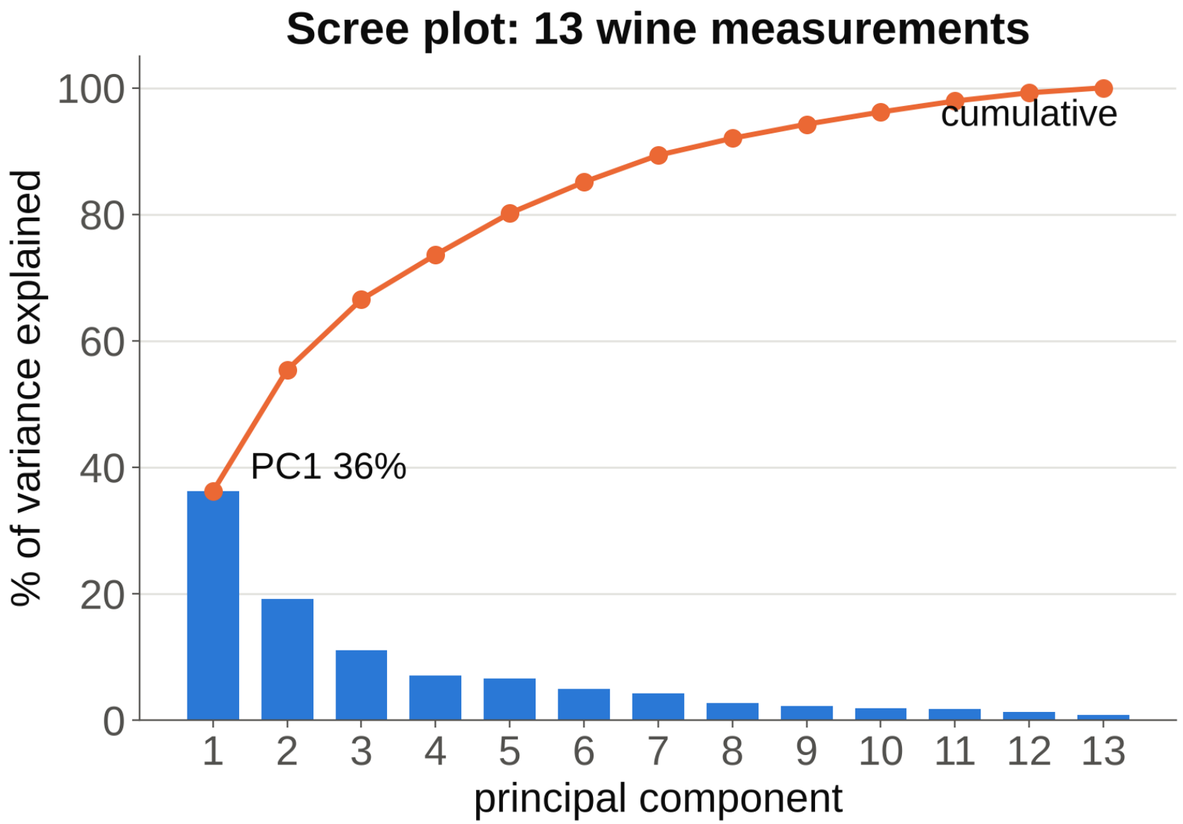


# How’s it calculated?

- Eigenvalue/eigenvector decomposition of covariance matrix
  - (how each attribute varies from each of the other ones)
- Select vectors of the top k eigenvalues

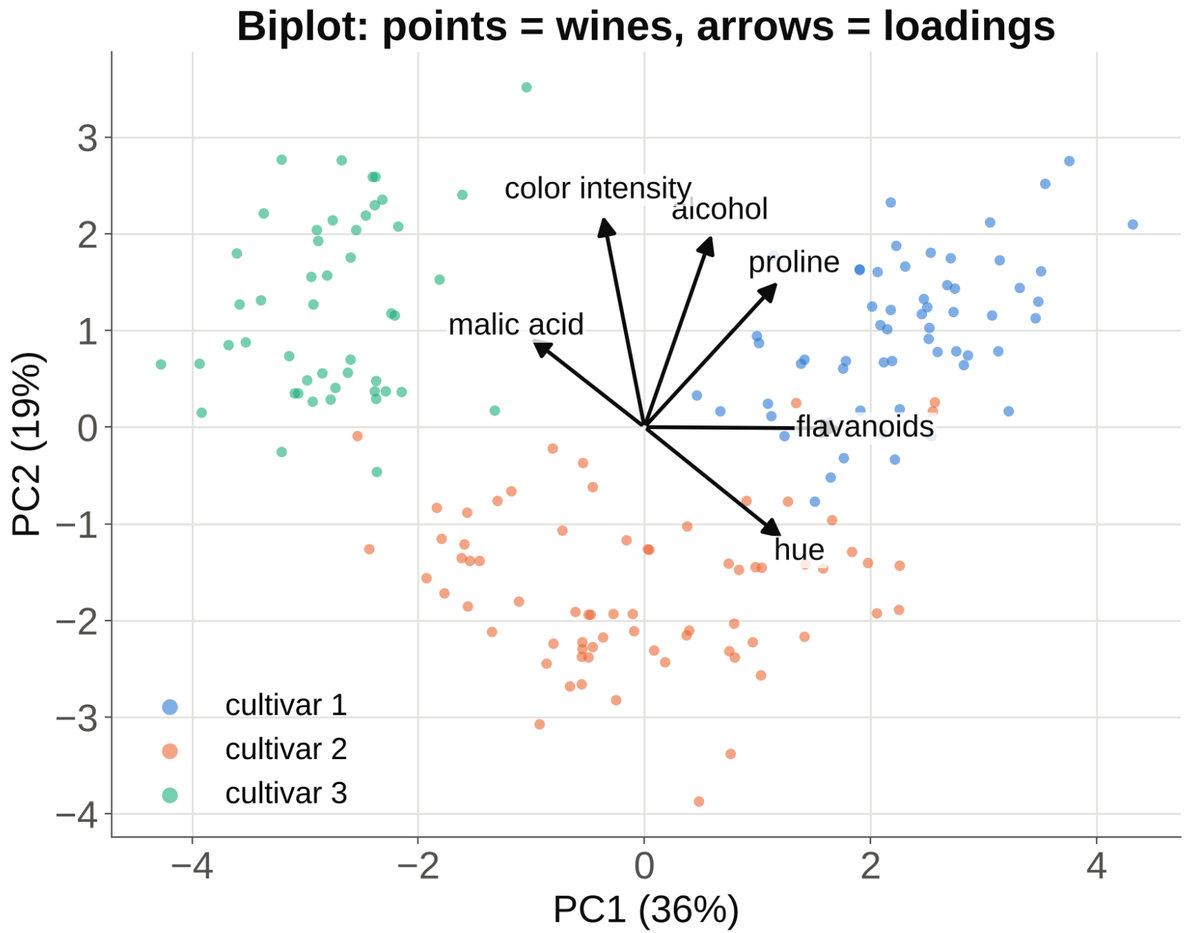
# Reading PCA: how many components?

- Each principal component explains a share of the total variance
- Scree plot: look for the elbow, or keep enough PCs for 80–90% cumulative
- Here 2 PCs keep 55% of the 13-D wine data, so a 2D plot hides almost half
- `explained_variance_ratio_`



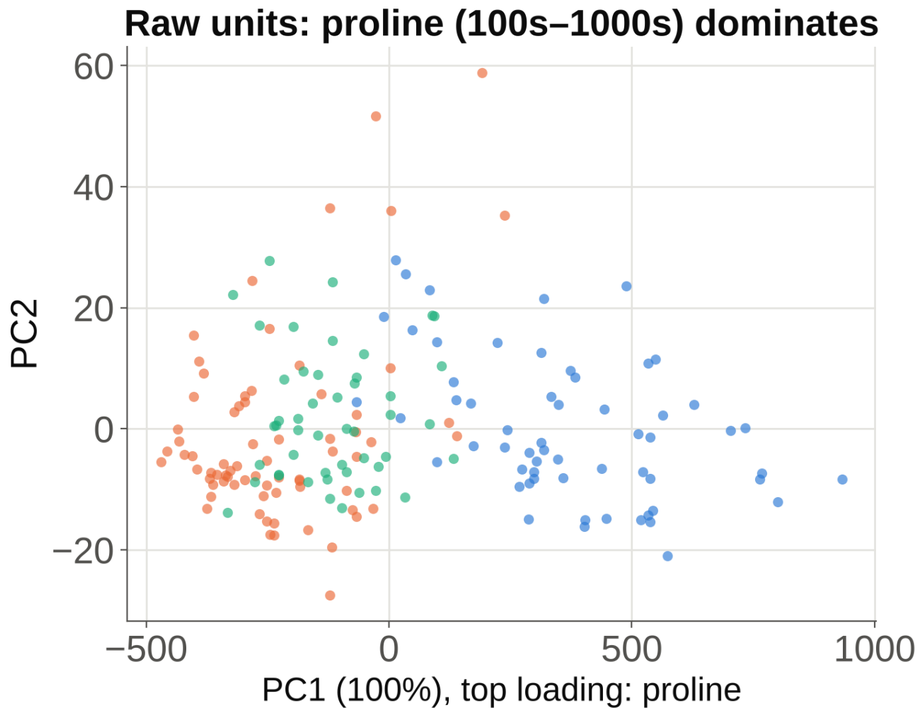
# Reading PCA: loadings and biplots

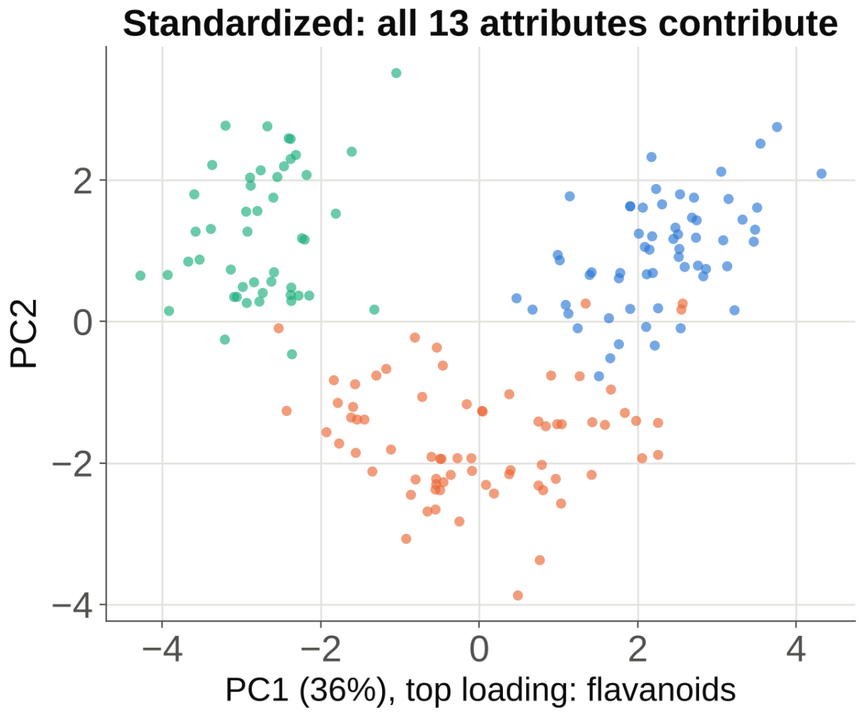
- Loadings: how much each original attribute contributes to a PC
- Biplot: points are items, arrows are attributes
- Arrows pointing the same way are correlated; longer = more influence in this view
- Wines on the right are high in flavanoids and proline



# Standardize before PCA

- PCA maximizes variance, so the attribute with the biggest units wins
- Proline is measured in the hundreds: it alone is PC1. Standardize (z-scores) unless all attributes share units



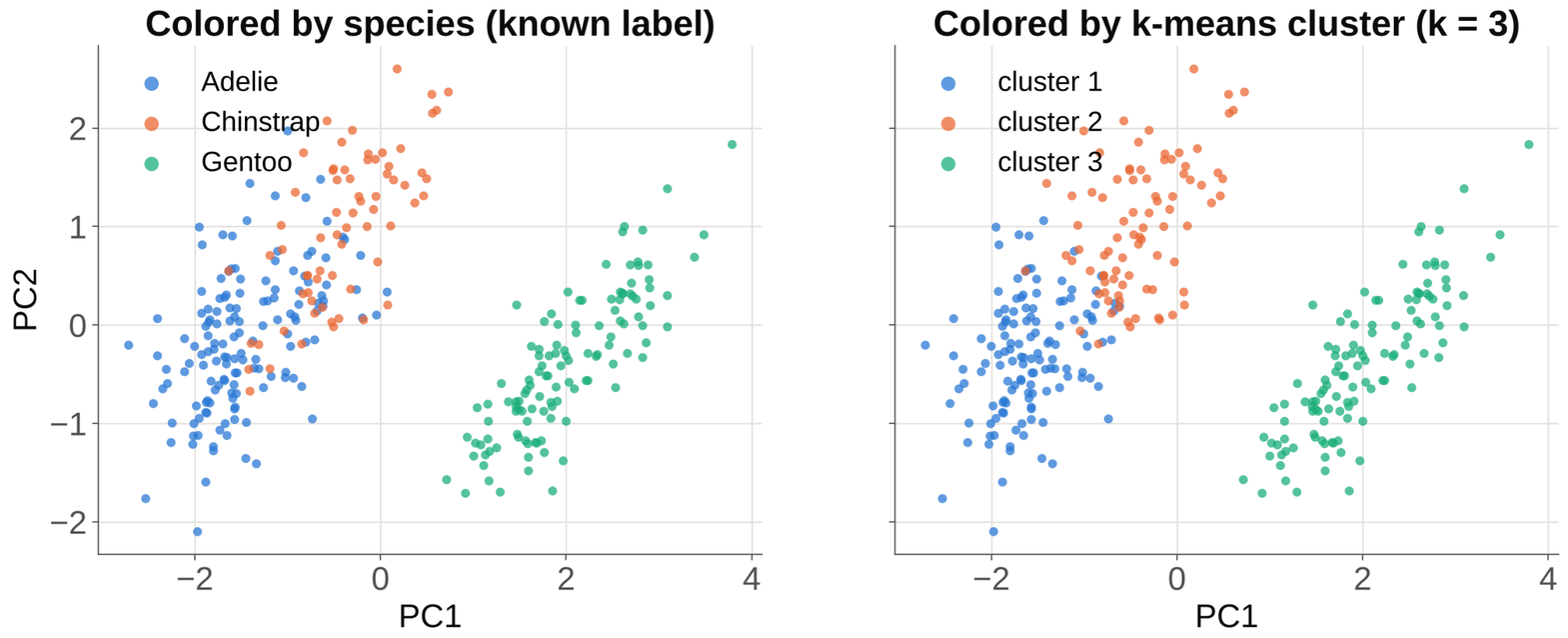



# PCA: what do we gain and lose?

- Gain: one 2D view of all attributes; clusters, outliers, and correlated groups become visible
- Lose: the variance held by the dropped components (check the scree plot)
- Lose: axes are mixtures of attributes, so read the loadings to interpret them
- Lose: PCA is linear; curved structure needs nonlinear methods (coming up)

# Clusters on an embedding

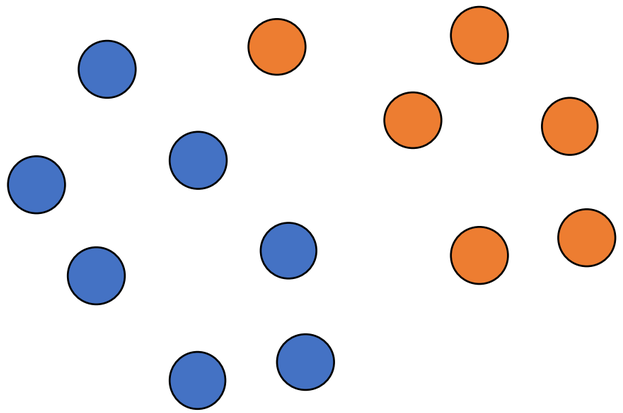
- Color a PCA (or UMAP) projection by cluster to see how well the groups separate
- k-means on 4 penguin measurements matches species for 92% of birds; the misses sit where Adelie and Chinstrap overlap




Palmer penguins data (Gorman et al. 2014)

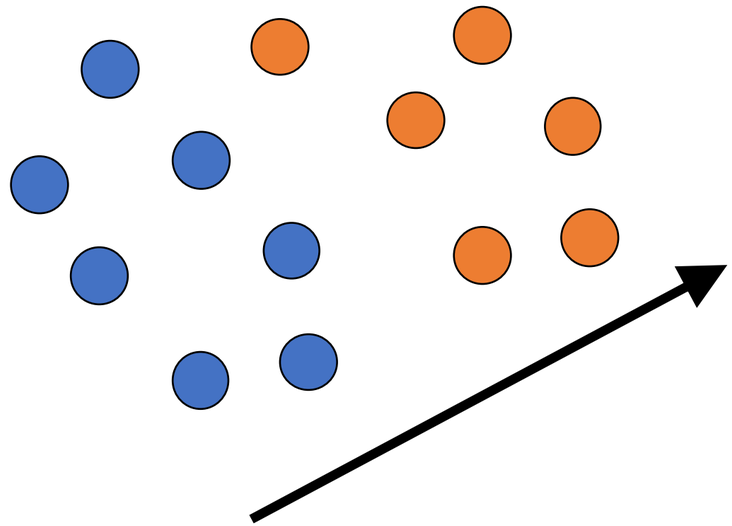
# Linear Discriminant Analysis

- Project points linearly (similarly to PCA)
- Projection axes are calculated to maximize the separability of classes



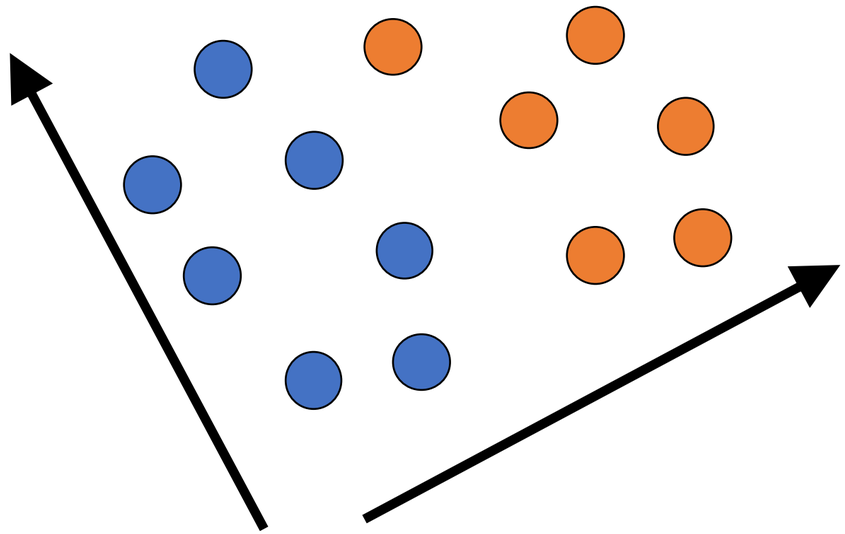
# Linear Discriminant Analysis

- Identify the axis that most separates classes



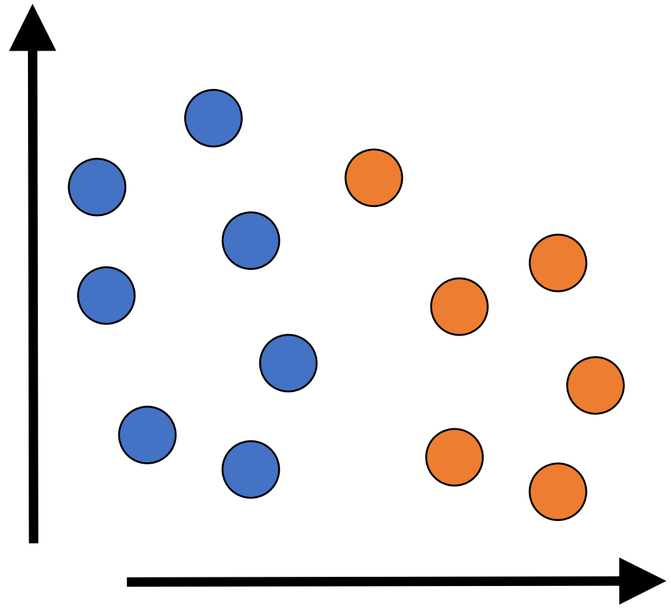
# Linear Discriminant Analysis

- Identify a perpendicular axis that most separates classes



# Linear Discriminant Analysis

- Project (similarly to PCA)



# How is it calculated?

- Eigenvalue/eigenvector decomposition of the composition of two matrices
  - Within-class scatter (variance of data within their class)
  - Between-class scatter (distance between class mean and overall mean)

# Nonlinear Dimensionality Reduction

# Multidimensional Scaling (MDS)

- Given pairwise distances/dissimilarities
- Find a map that preserves distances
- Many “flavors”
- We’ll look at classical MDS

# Basic MDS Process

- Given a dissimilarity (distance) matrix D, where D[i, j] says how far apart items i and j are
- MDS finds points in P dimensions whose pairwise distances match D[i, j] as closely as possible

# Basic MDS Process

- Use eigenvector/eigenvalue decomposition of the dissimilarity matrix
- Although this is a linear operation, the effect to data is nonlinear

# Classical MDS: airline distances

**TABLE 13.2.** *Airline distances (km) between 18 cities. Source: Atlas of the World, Revised 6th Edition, National Geographic Society, 1995, p. 131.*

| | Beijing | Cape Town | Hong Kong | Honolulu | London | Melbourne |
|---|---|---|---|---|---|---|
| Cape Town | 12947 | | | | | |
| Hong Kong | 1972 | 11867 | | | | |
| Honolulu | 8171 | 18562 | 8945 | | | |
| London | 8160 | 9635 | 9646 | 11653 | | |
| Melbourne | 9093 | 10338 | 7392 | 8862 | 16902 | |
| Mexico | 12478 | 13703 | 14155 | 6098 | 8947 | 13557 |
| Montreal | 10490 | 12744 | 12462 | 7915 | 5240 | 16730 |
| Moscow | 5809 | 10101 | 7158 | 11342 | 2506 | 14418 |
| New Delhi | 3788 | 9284 | 3770 | 11930 | 6724 | 10192 |
| New York | 11012 | 12551 | 12984 | 7996 | 5586 | 16671 |
| Paris | 8236 | 9307 | 9650 | 11988 | 341 | 16793 |
| Rio de Janeiro | 17325 | 6075 | 17710 | 13343 | 9254 | 13227 |
| Rome | 8144 | 8417 | 9300 | 12936 | 1434 | 15987 |
| San Francisco | 9524 | 16487 | 11121 | 3857 | 8640 | 12644 |
| Singapore | 4465 | 9671 | 2575 | 10824 | 10860 | 6050 |
| Stockholm | 6725 | 10334 | 8243 | 11059 | 1436 | 15593 |
| Tokyo | 2104 | 14737 | 2893 | 6208 | 9585 | 8159 |

| | Mexico | Montreal | Moscow | New Delhi | New York | Paris |
|---|---|---|---|---|---|---|
| Montreal | 3728 | | | | | |
| Moscow | 10740 | 7077 | | | | |
| New Delhi | 14679 | 11286 | 4349 | | | |
| New York | 3362 | 533 | 7530 | 11779 | | |
| Paris | 9213 | 5522 | 2492 | 6601 | 5851 | |

# Classical MDS: airline distances

- Negative eigenvalues indicate Airline distance is non-Euclidean
- Should only use at most first 3 eigenvectors (negative eigenvalues produce resulting in the imaginary domain)

**TABLE 13.6.** *Eigenvalues of* B *and the eigenvectors corresponding to the first three largest eigenvalues (in red) for the airline distances example.*

| | Eigenvalues | Eigenvectors | | |
|---|---|---|---|---|
| 1 | 471582511 | 0.245 | –0.072 | 0.183 |
| 2 | 316824787 | 0.003 | 0.502 | -0.347 |
| 3 | 253943687 | 0.323 | –0.017 | 0.103 |
| 4 | –98466163 | 0.044 | –0.487 | -0.080 |
| 5 | –74912121 | –0.145 | 0.144 | 0.205 |
| 6 | –47505097 | 0.366 | –0.128 | –0.569 |
| 7 | 31736348 | –0.281 | –0.275 | –0.174 |
| 8 | –7508328 | –0.272 | –0.115 | 0.094 |
| 9 | 4338497 | –0.010 | 0.134 | 0.202 |
| 10 | 1747583 | 0.209 | 0.195 | 0.110 |
| 11 | –1498641 | –0.292 | –0.117 | 0.061 |
| 12 | 145113 | –0.141 | 0.163 | 0.196 |
| 13 | –102966 | –0.364 | 0.172 | –0.473 |
| 14 | 60477 | –0.104 | 0.220 | 0.163 |
| 15 | –6334 | –0.140 | –0.356 | –0.009 |
| 16 | –1362 | 0.375 | 0.139 | –0.054 |
| 17 | 100 | –0.074 | 0.112 | 0.215 |
| 18 | 0 | 0.260 | –0.214 | 0.173 |

# Classical MDS: airline distances

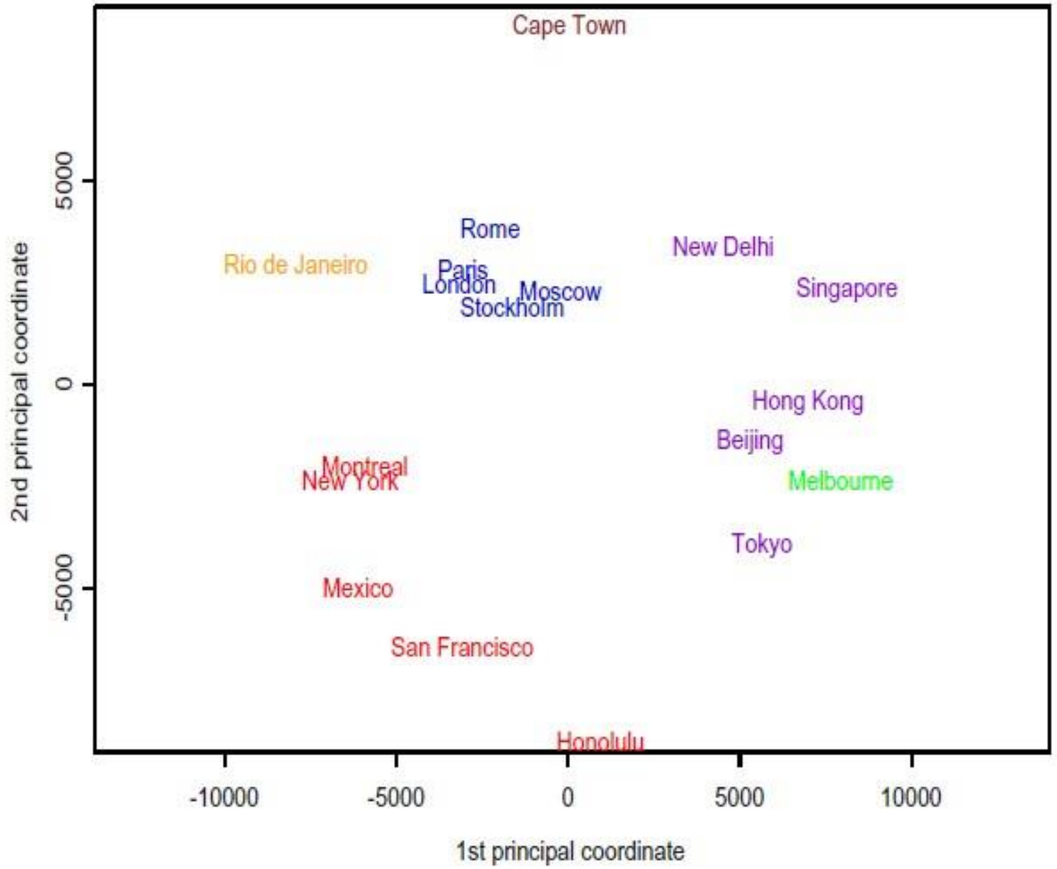



**FIGURE 13.1.** *Two-dimensional map of 18 world cities using the classical scaling algorithm on airline distances between those cities. The colors*

# Classical MDS: airline distances

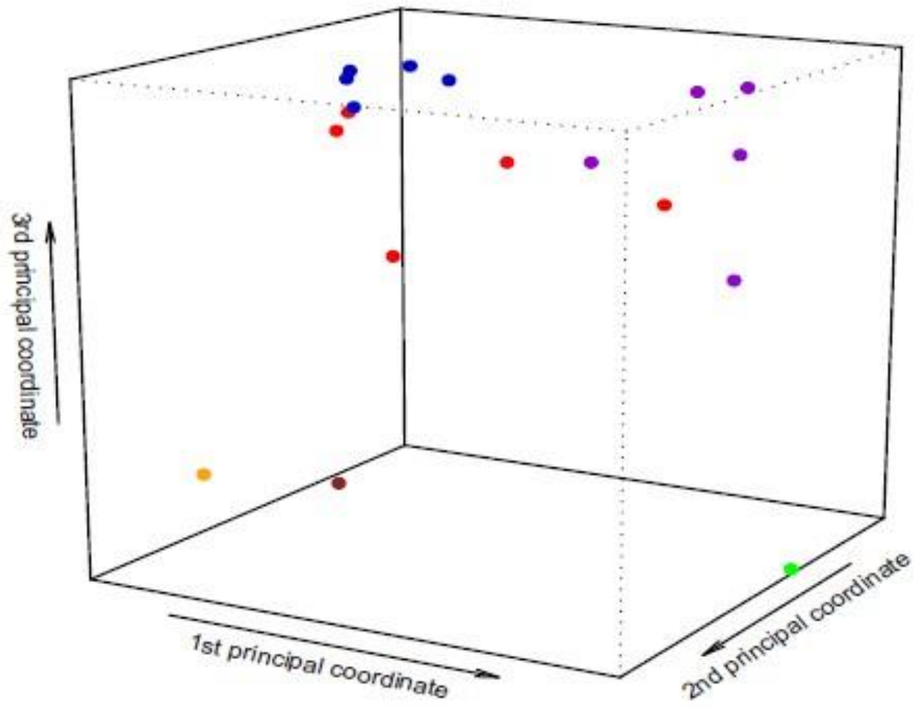



**FIGURE 13.2.** *Three-dimensional map of 18 world cities using the classical scaling algorithm on airline distances between those cities. The colors reflect the different continents: Asia (purple), North America (red), South America (yellow), Europe (blue), Africa (brown), and Australasia (green).*

# Classical MDS: airline distances

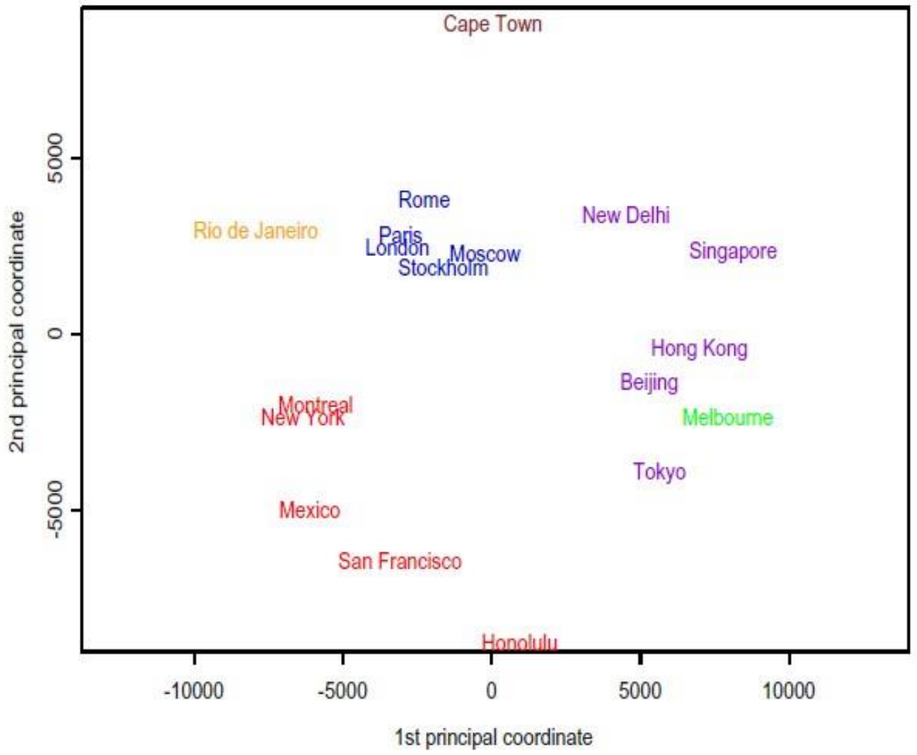



FIGURE 13.1. *Two-dimensional map of 18 world cities using the classical scaling algorithm on airline distances between those cities. The colors*

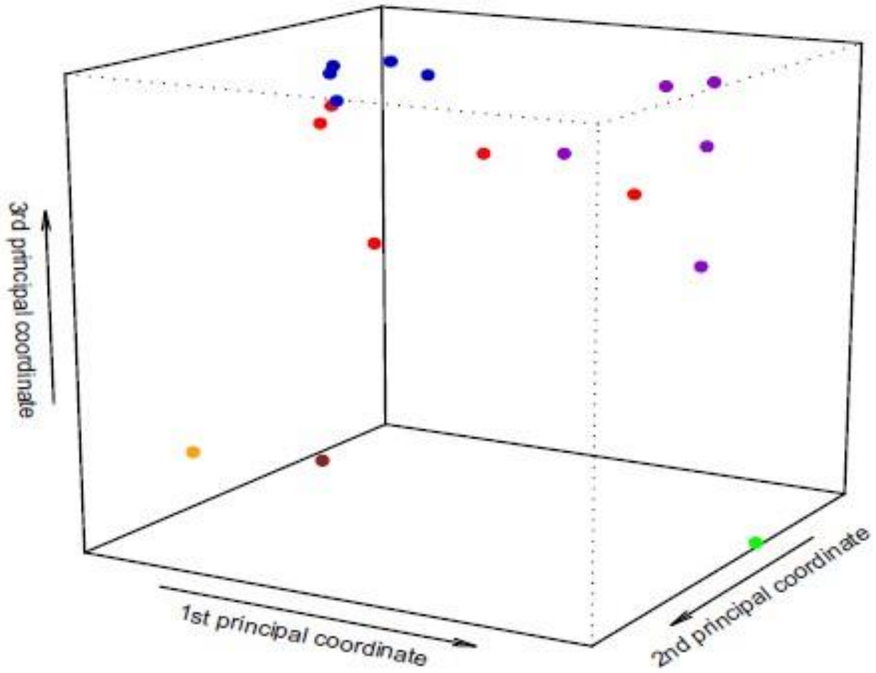
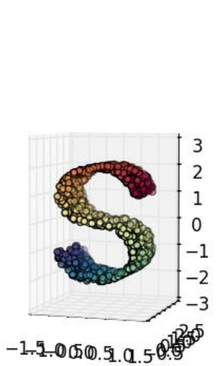
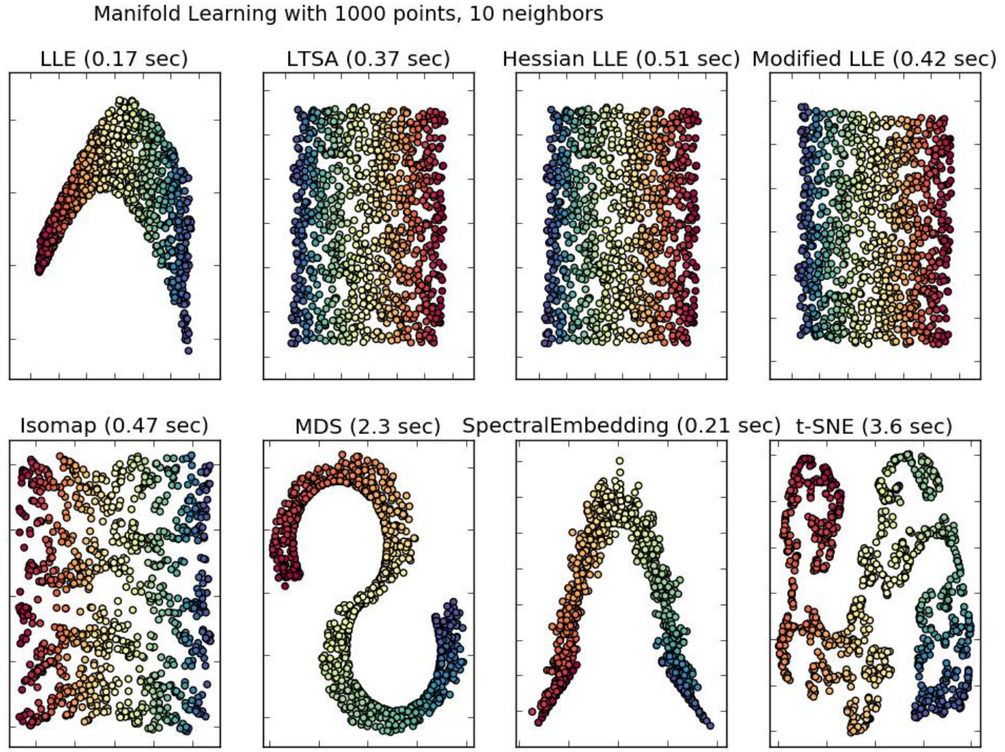



FIGURE 13.2. *Three-dimensional map of 18 world cities using the classical scaling algorithm on airline distances between those cities. The colors reflect the different continents: Asia (purple), North America (red), South America (yellow), Europe (blue), Africa (brown), and Australasia (green).*

# Nonlinear (manifold) methods

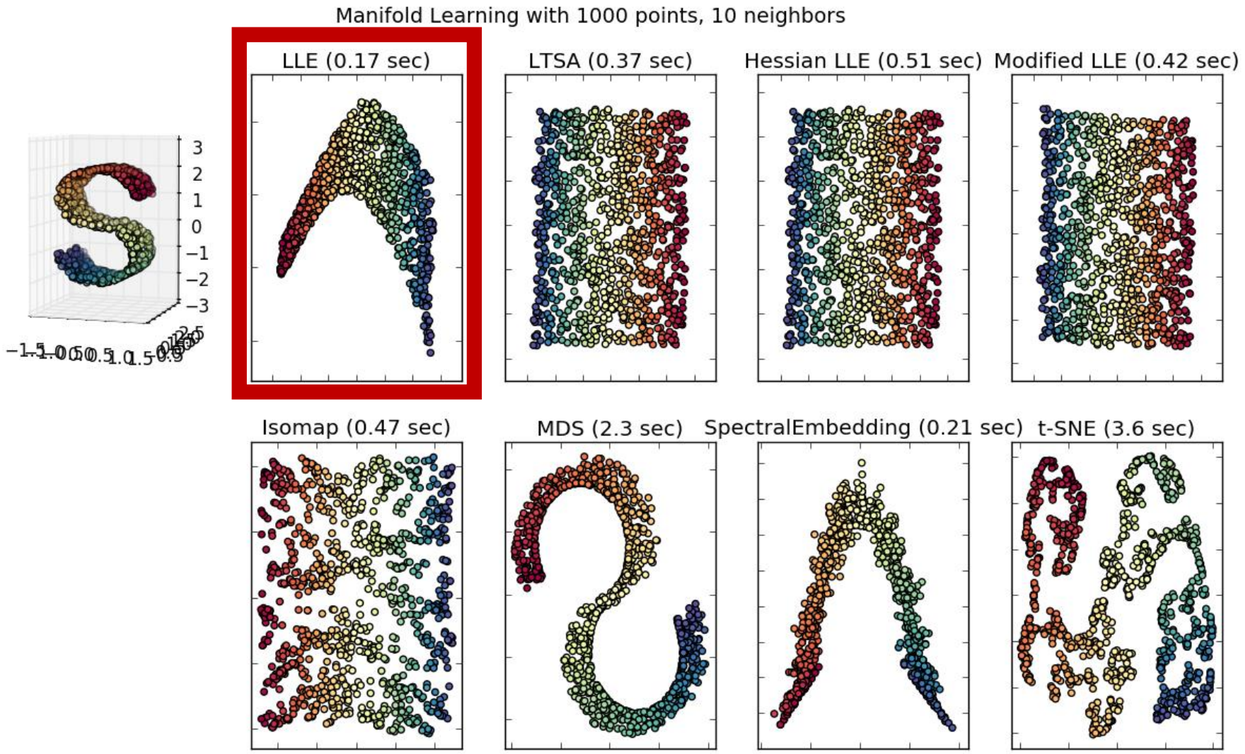
- A lot of options, all preserve different features of the high-dimensional space



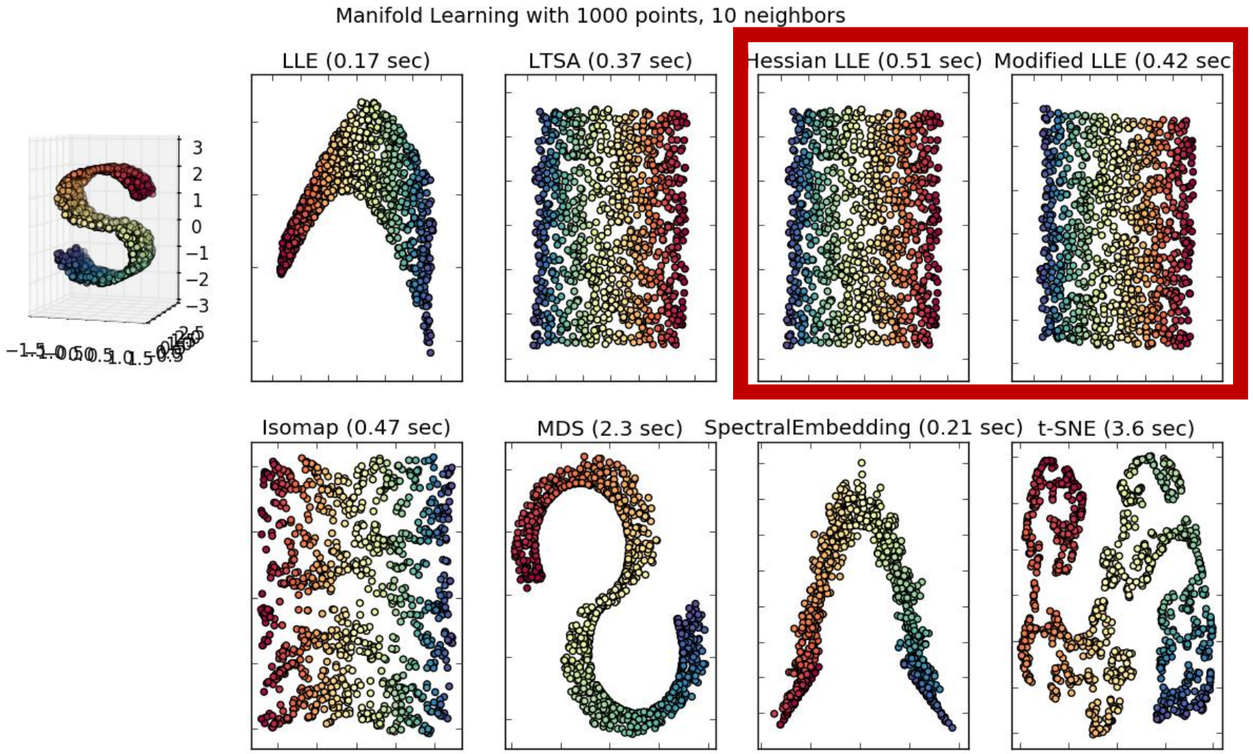
# Local Linear Embedding (LLE)

- Tries to embed such that local neighborhood distances are preserved as well as possible



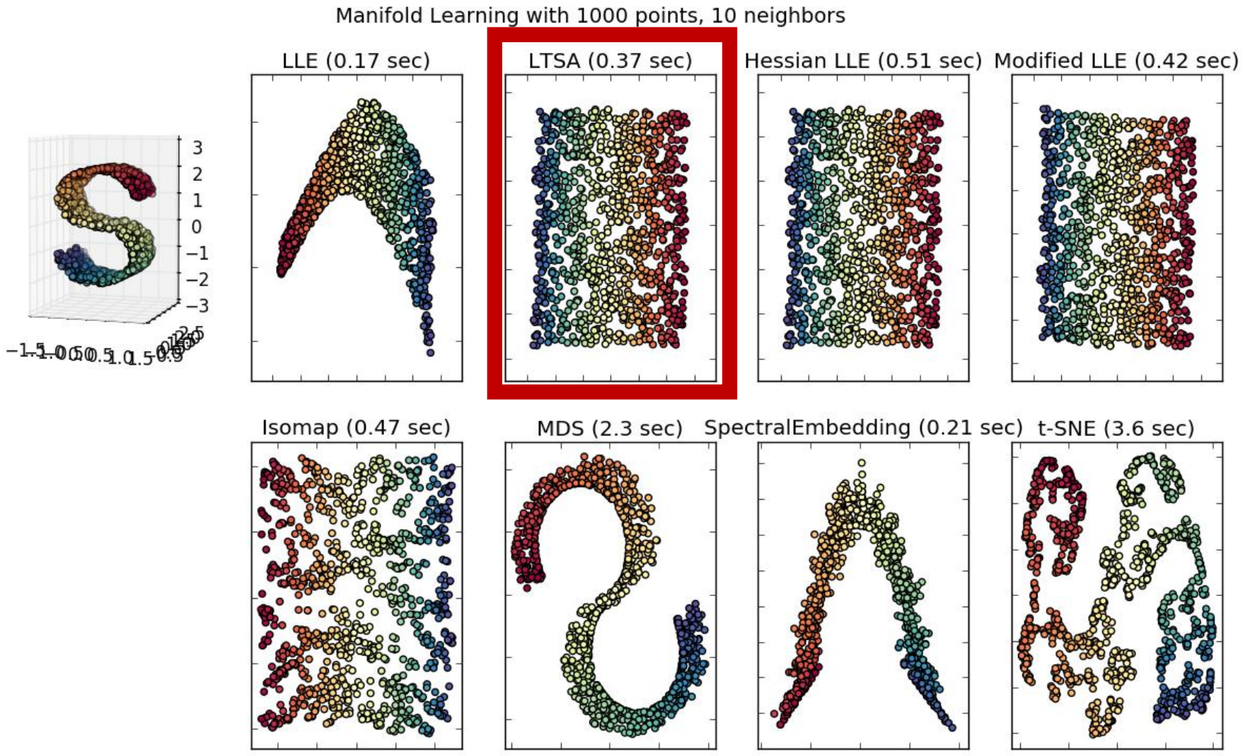
# LLE Extensions

- Fix various issues associated with regularization of the domain in LLE



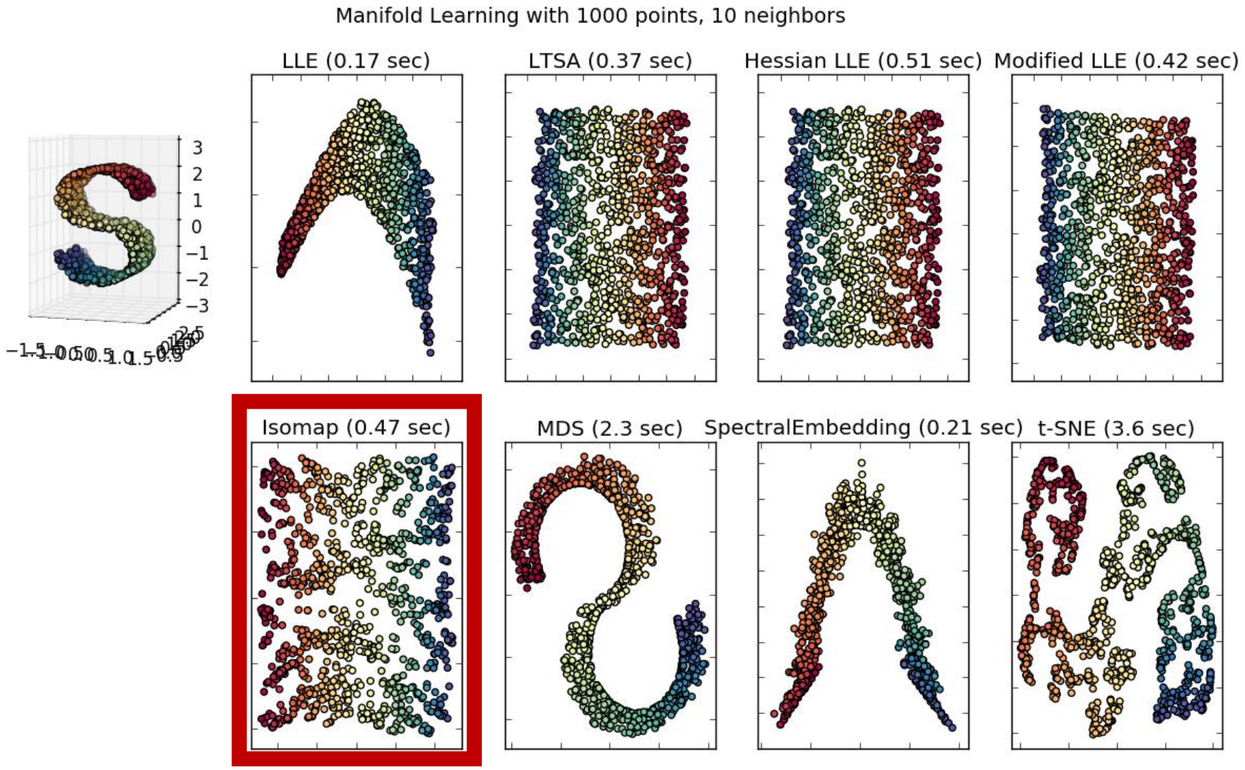
# Local Tangent Space Alignment (LTSA)

- Tries to characterize geometry via local tangent directions



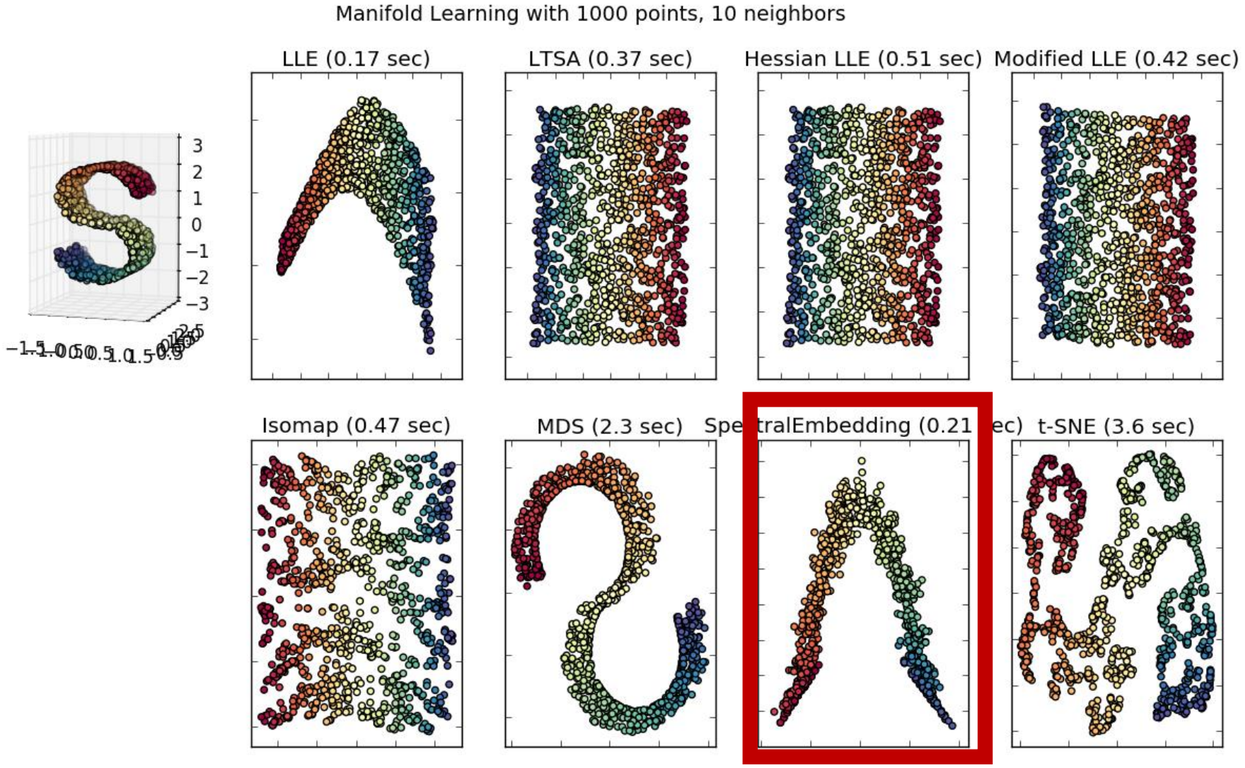
# Isomap

- Extension of MDS. Tries to preserve geodesic distance as well as possible



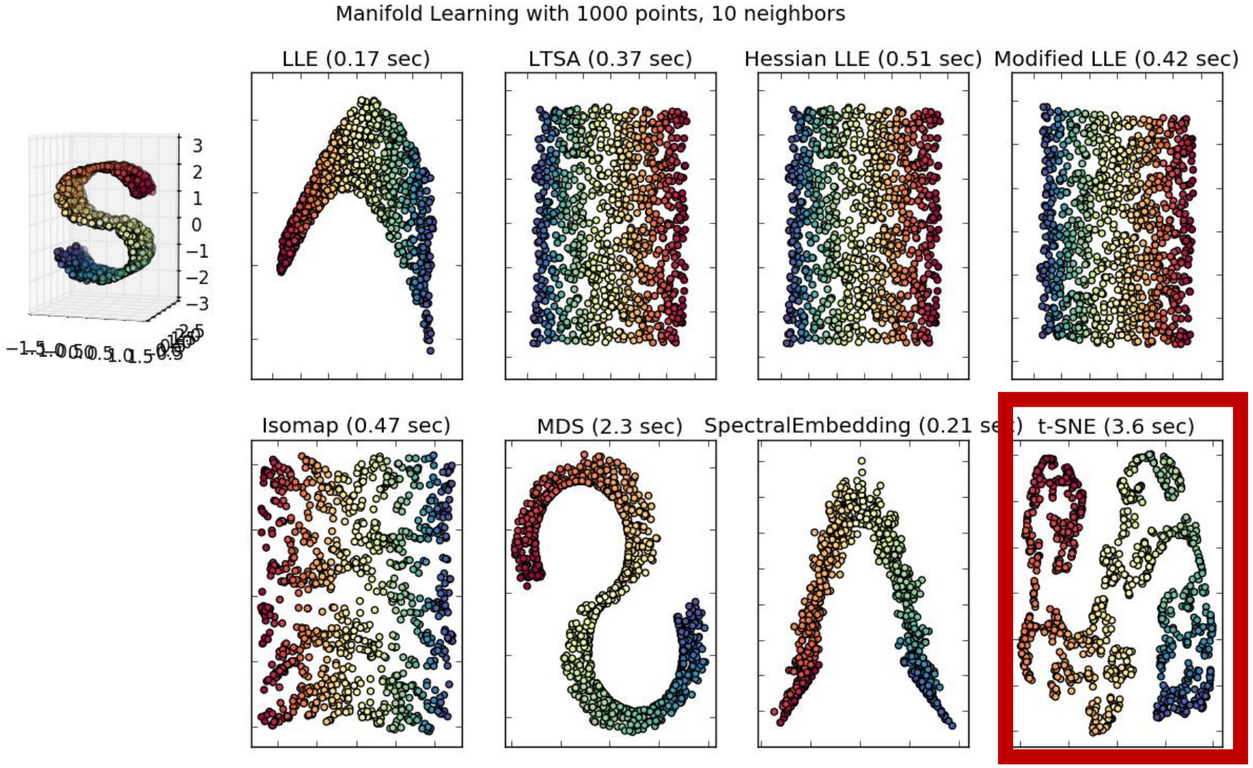
# Spectral Embedding

- Uses a spectral decomposition of the graph Laplacian



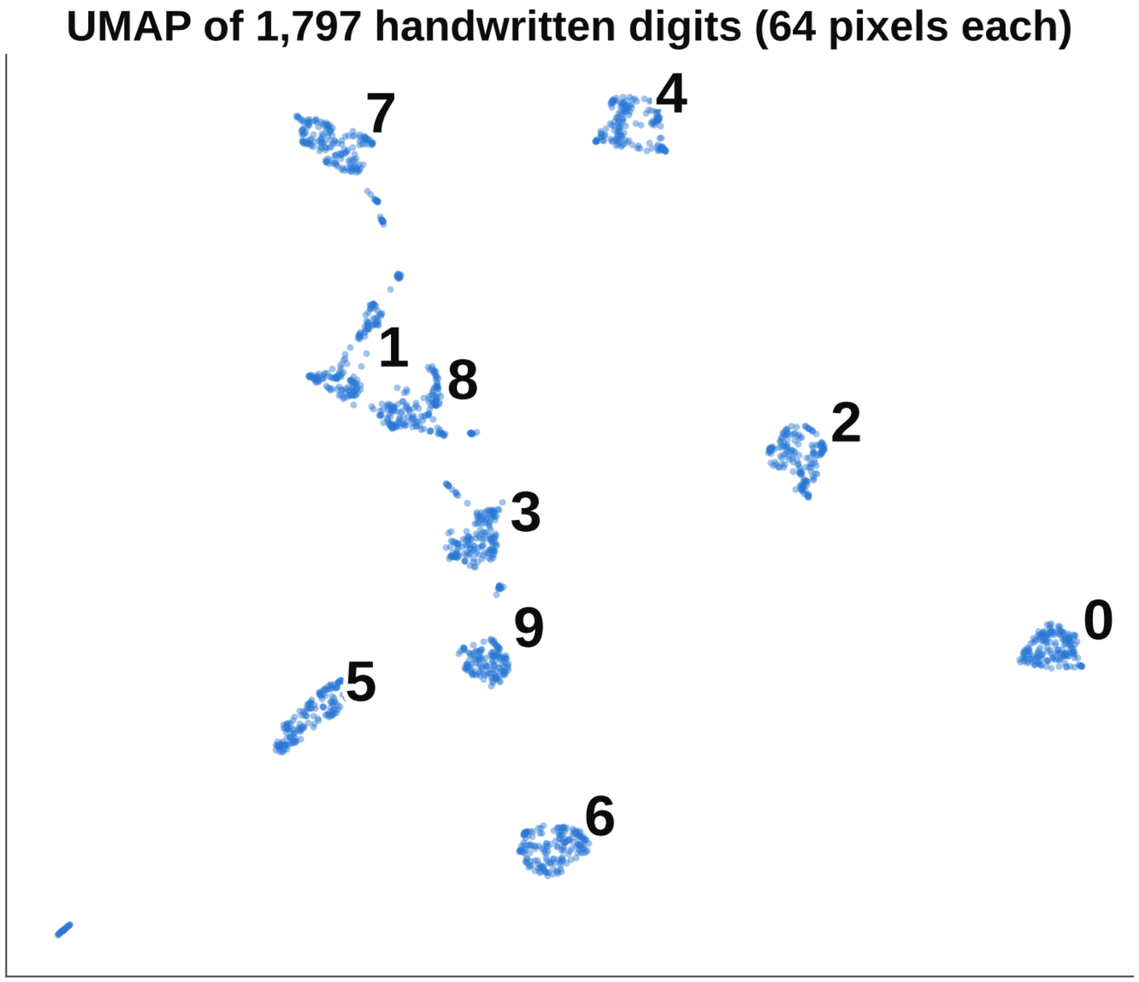
# t-SNE

- t-distributed Stochastic Neighbor Embedding
- Matches each point's neighbor probabilities in high-D and in 2D
- Keeps local neighborhoods; global distances and cluster sizes are not kept



# UMAP

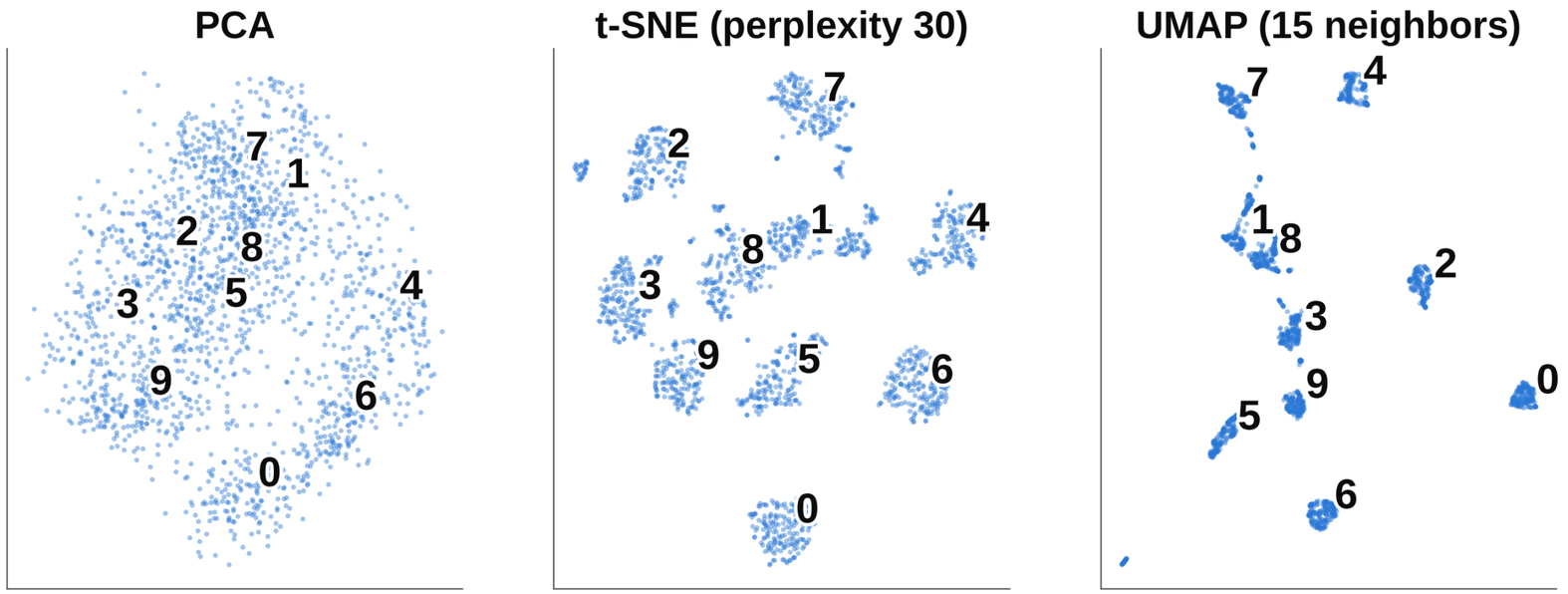
- Uniform Manifold Approximation and Projection
- Builds a nearest-neighbor graph, then lays the graph out in 2D
- Keeps local neighborhoods like t-SNE; faster, and often keeps more of the global layout
- Knobs: n_neighbors (local vs. global), min_dist (packing)




McInnes, Healy & Melville 2018; sklearn digits data (UCI)

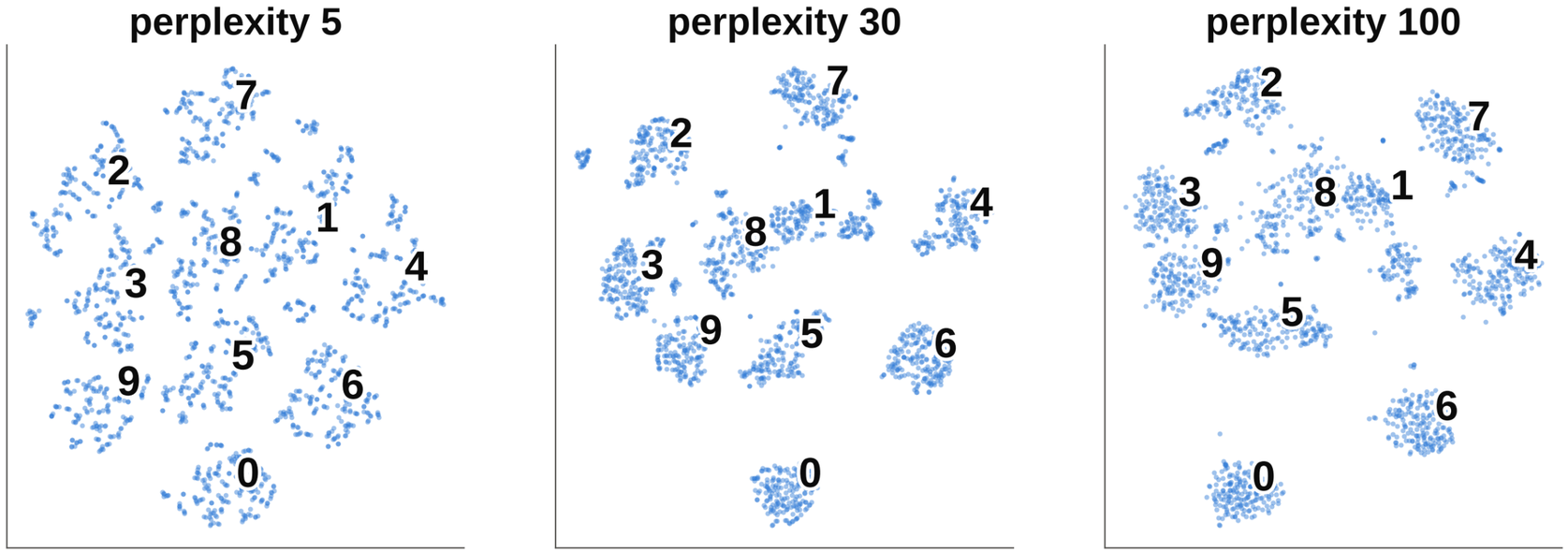
# Same data, three projections

- 1,797 handwritten digits, 64 pixel values each; labels are added afterwards
- PCA overlaps most digits; t-SNE and UMAP separate them without ever seeing the labels



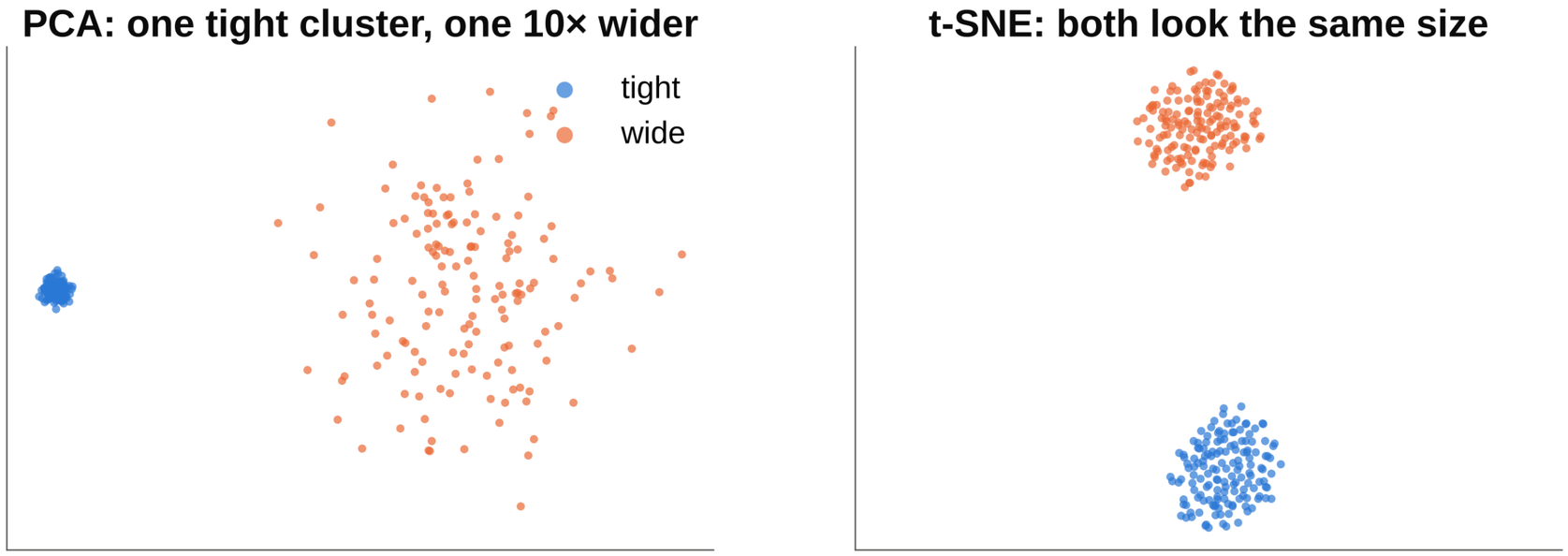
# t-SNE: perplexity changes the picture

- Perplexity is roughly how many neighbors each point pays attention to
- Different settings give different maps: try several and trust what persists



# t-SNE and UMAP: what not to read

- Cluster sizes mean nothing: a tight cluster and a 10× wider one come out the same size
- Gaps between clusters are unreliable; check clusters against the original data and several seeds




Wattenberg, Viégas & Johnson 2016 (Distill); simulated data

# Problem #3
# What is lost or misinterpreted…

In other words, know the shapes (information) your statistic captures

# Anscombe's Quartet

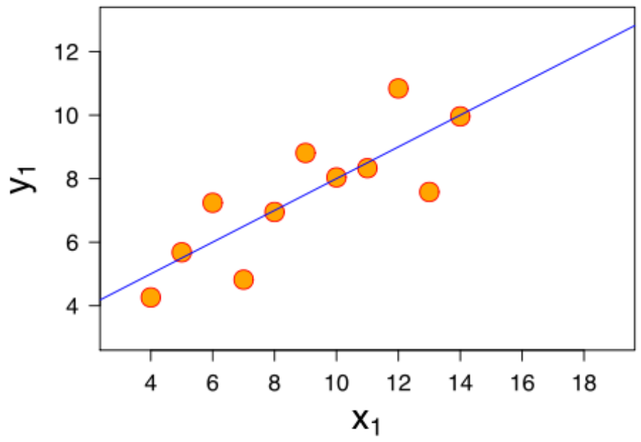
| Data set | | 1-3 | 1 | 2 | 3 | | 4 | 4 |
|---|---|---|---|---|---|---|---|---|
| Variable | | x | y | y | y | | x | y |
| Obs. no. | 1 | 10.0 | 8.04 | 9.14 | 7.46 | | 8.0 | 6.58 |
| | 2 | 8.0 | 6.95 | 8.14 | 6.77 | | 8.0 | 5.76 |
| | 3 | 13.0 | 7.58 | 8.74 | 12.74 | | 8.0 | 7.71 |
| | 4 | 9.0 | 8.81 | 8.77 | 7.11 | | 8.0 | 8.84 |
| | 5 | 11.0 | 8.33 | 9.26 | 7.81 | | 8.0 | 8.47 |
| | 6 | 14.0 | 9.96 | 8.10 | 8.84 | | 8.0 | 7.04 |
| | 7 | 6.0 | 7.24 | 6.13 | 6.08 | | 8.0 | 5.25 |
| | 8 | 4.0 | 4.26 | 3.10 | 5.39 | | 19.0 | 12.50 |
| | 9 | 12.0 | 10.84 | 9.13 | 8.15 | | 8.0 | 5.56 |
| | 10 | 7.0 | 4.82 | 7.26 | 6.42 | | 8.0 | 7.91 |
| | 11 | 5.0 | 5.68 | 4.74 | 5.73 | | 8.0 | 6.89 |

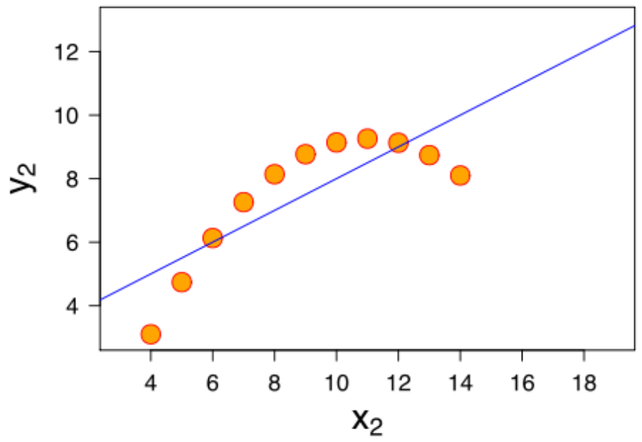
TABLE. Four data sets, each comprising 11 (x, y) pairs.

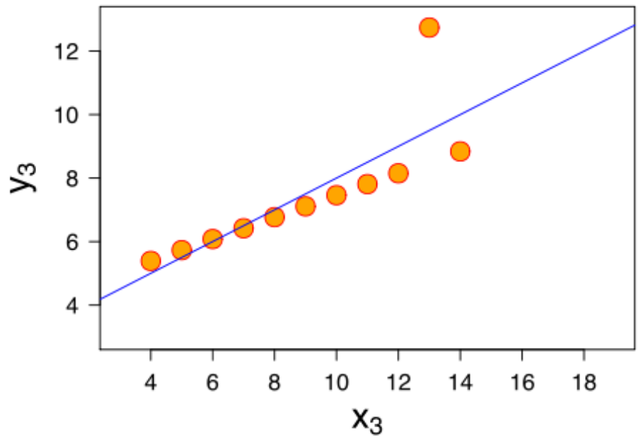
Graphs in Statistical Analysis. F. J. Anscombe. The American, Statistician, Vol. 27, No. 1. (Feb., 1973), pp. 17-21.

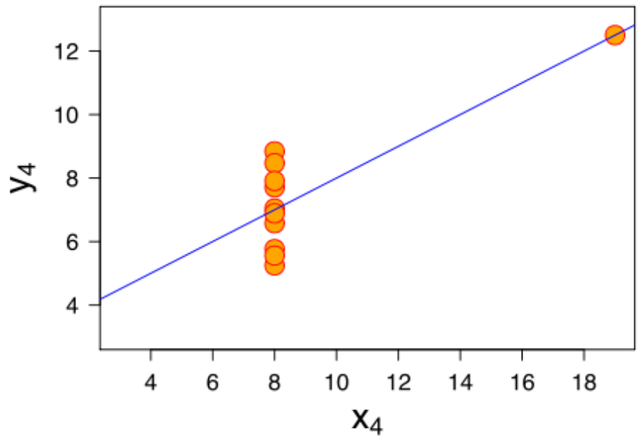
| Property | Value | Accuracy |
|---|---|---|
| Mean of x | 9 | exact |
| Sample variance of x | 11 | exact |
| Mean of y | 7.50 | to 2 decimal places |
| Sample variance of y | 4.125 | plus/minus 0.003 |
| Correlation between x and y | 0.816 | to 3 decimal places |
| Linear regression line | y = 3.00 + 0.500x | to 2 and 3 decimal places, respectively |

# Anscombe's quartet




I (top left): a noisy line; the fit is fine

II (top right): a smooth curve; a line is the wrong model

III (bottom left): a perfect line plus one outlier

IV (bottom right): one far point creates the whole slope

# Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

**Justin Matejka and George Fitzmaurice**
Autodesk Research, Toronto Ontario Canada
{*first.last*}@autodesk.com

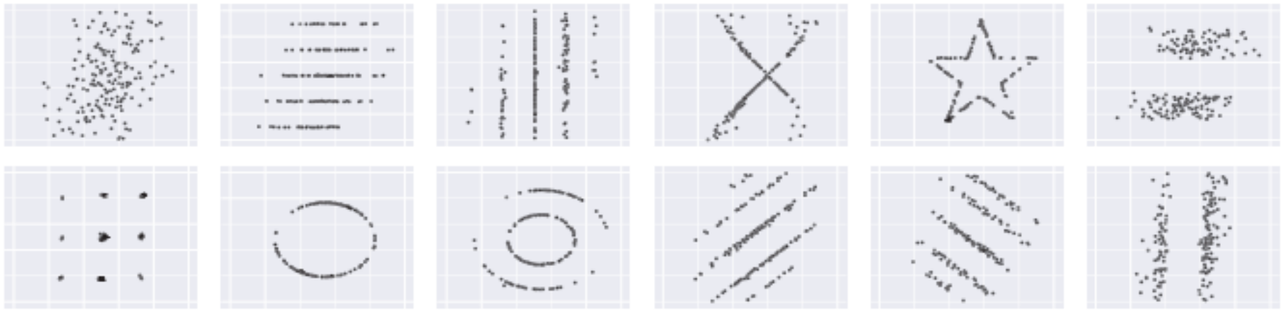


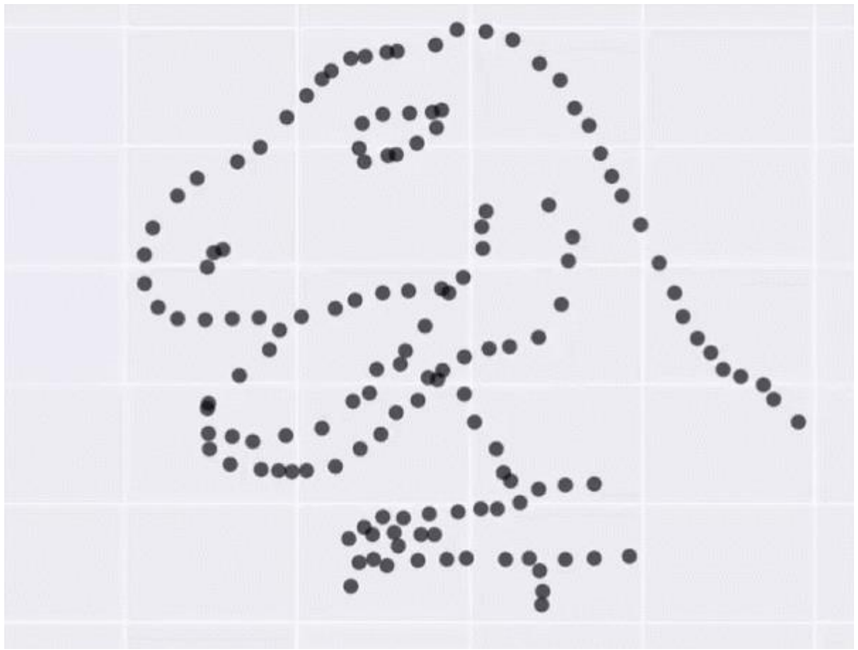
**Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. ($\bar{x}$ =54.02, $\bar{y}$ = 48.09, $sd_x$ = 14.52, $sd_y$ = 24.79, Pearson's r = +0.32)**

**ABSTRACT**

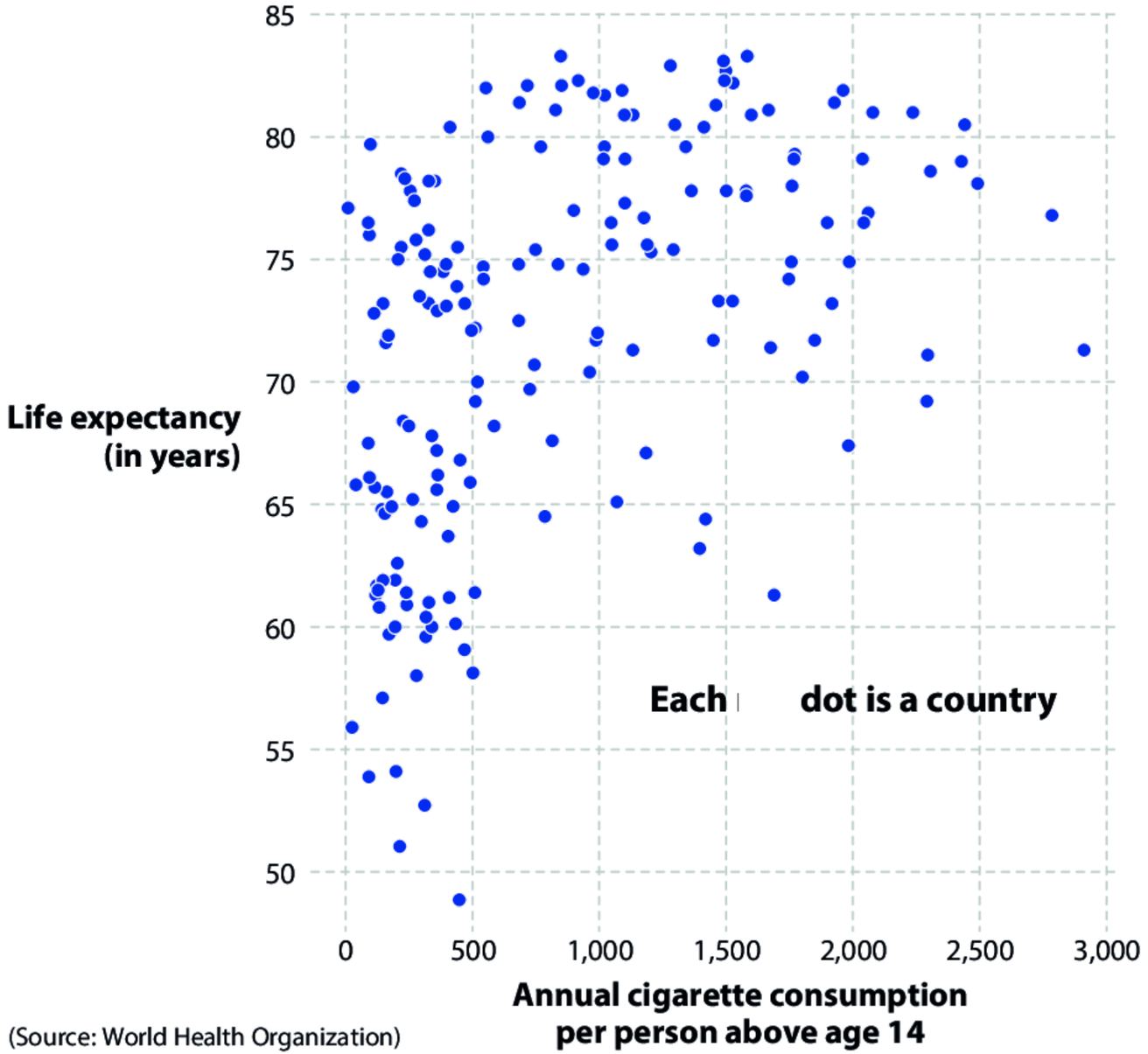
Datasets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This paper presents a novel method for generating such datasets, along with several examples. Our technique varies from previous approaches in that new datasets are iteratively generated from a seed dataset through random perturbations of individual data points, and can be directed towards a desired outcome through a simulated annealing optimization strategy. Our method has the benefit of being agnostic to the particular statistical properties that are to remain constant between the datasets, and allows for same statistical properties, it is that four *clearly different* and *identifiably distinct* datasets are producing the same statistical properties. Dataset I appears to follow a somewhat noisy linear model, while Dataset II is following a parabolic distribution. Dataset III appears to be strongly linear, except for a single outlier, while Dataset IV forms a vertical line with the regression thrown off by a single outlier. In contrast, Figure 2B shows a series of datasets also sharing the same summary statistics as Anscombe's Quartet, however without any obvious underlying structure to the individual datasets, this quartet is not nearly as effective at demonstrating the importance of graphical representations.

While very popular and effective for illustrating the



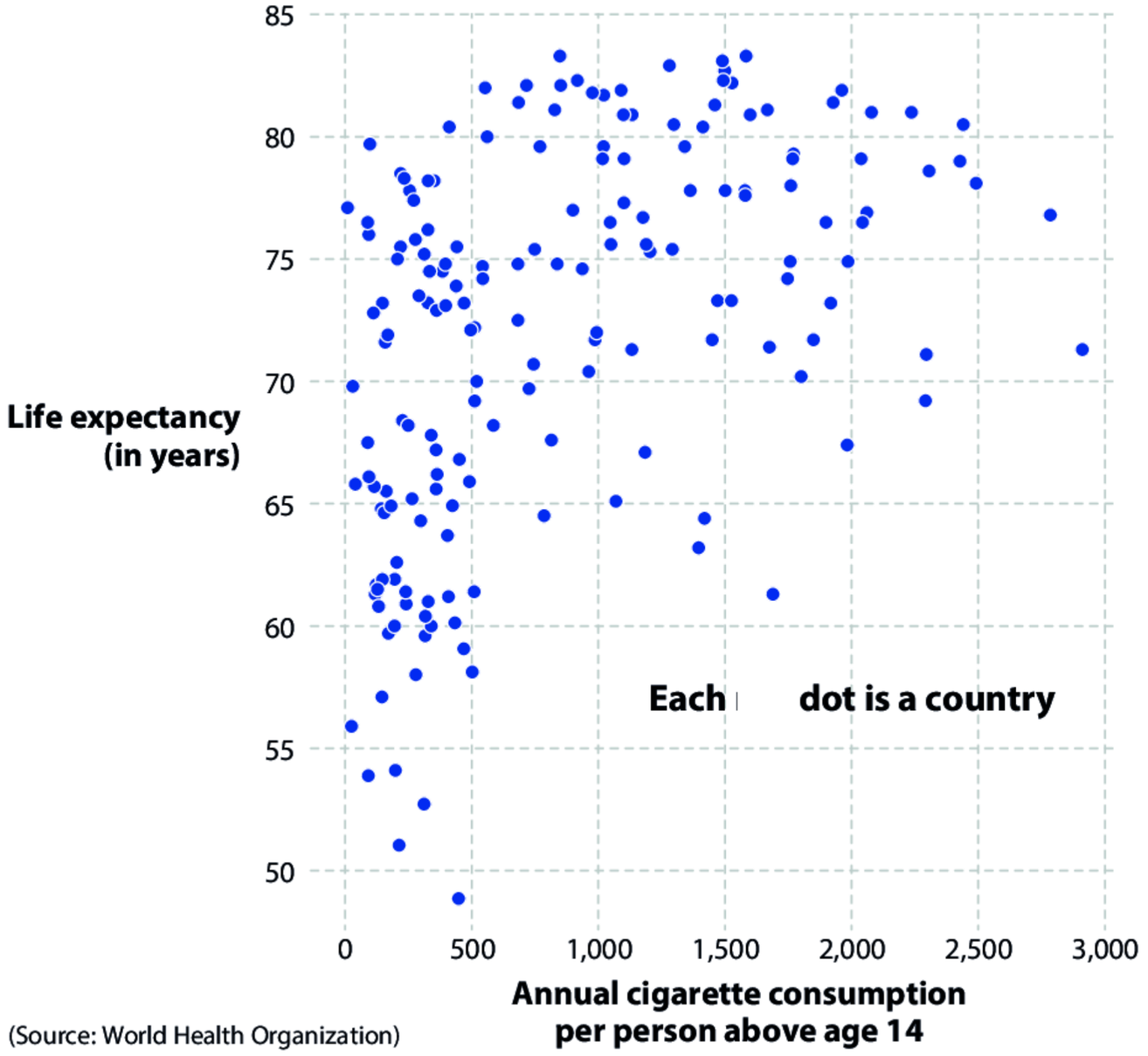



- What does this chart tell you?




(Source: World Health Organization)

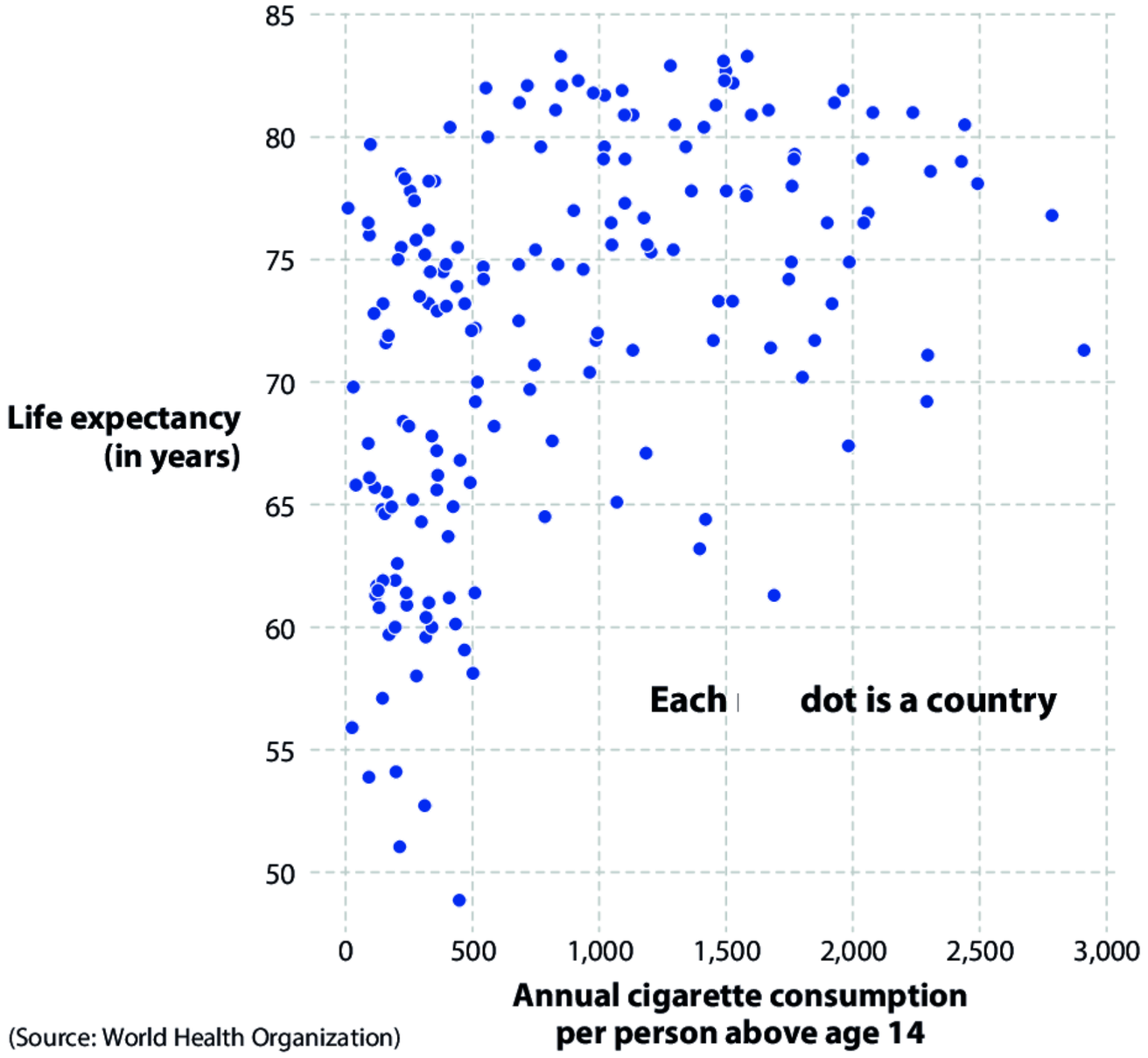
- “The more cigarettes we consume, the longer we live!”
- “There is a positive relationship between cigarette consumption and life expectancy at a country-by-country level!”




(Source: World Health Organization)

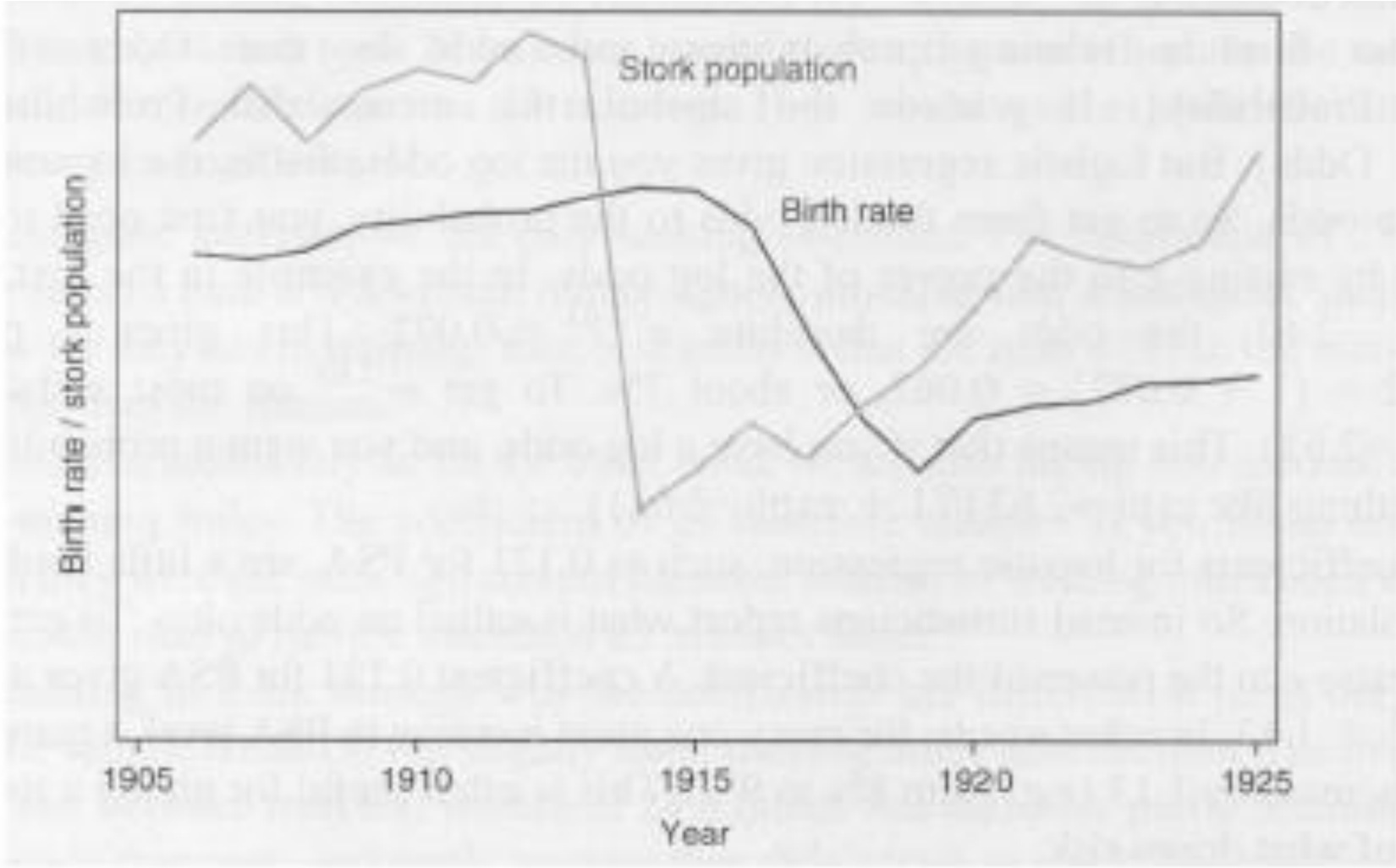
- ~~"The more cigarettes we consume, the longer we live!"~~

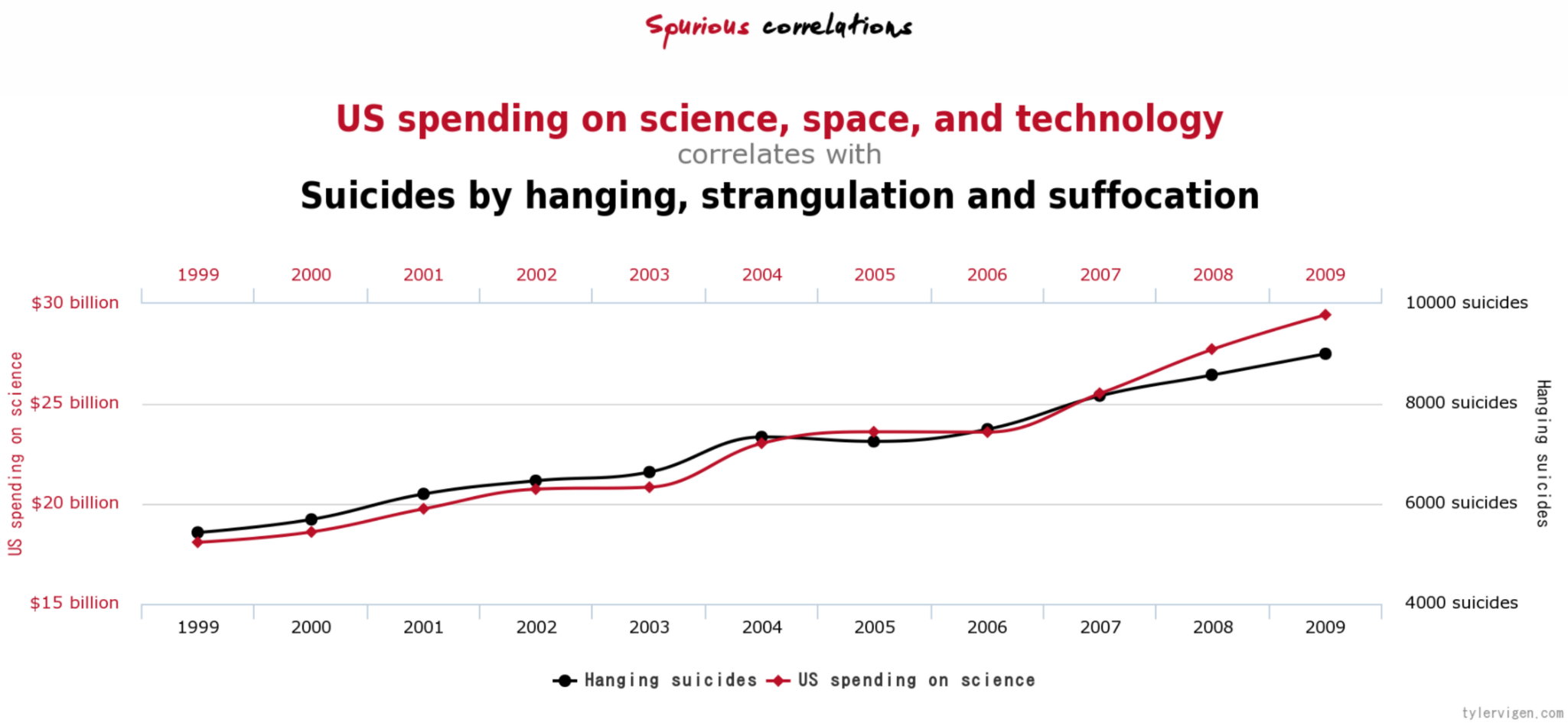
- "There is a positive relationship between cigarette consumption and life expectancy at a country-by-country level!"



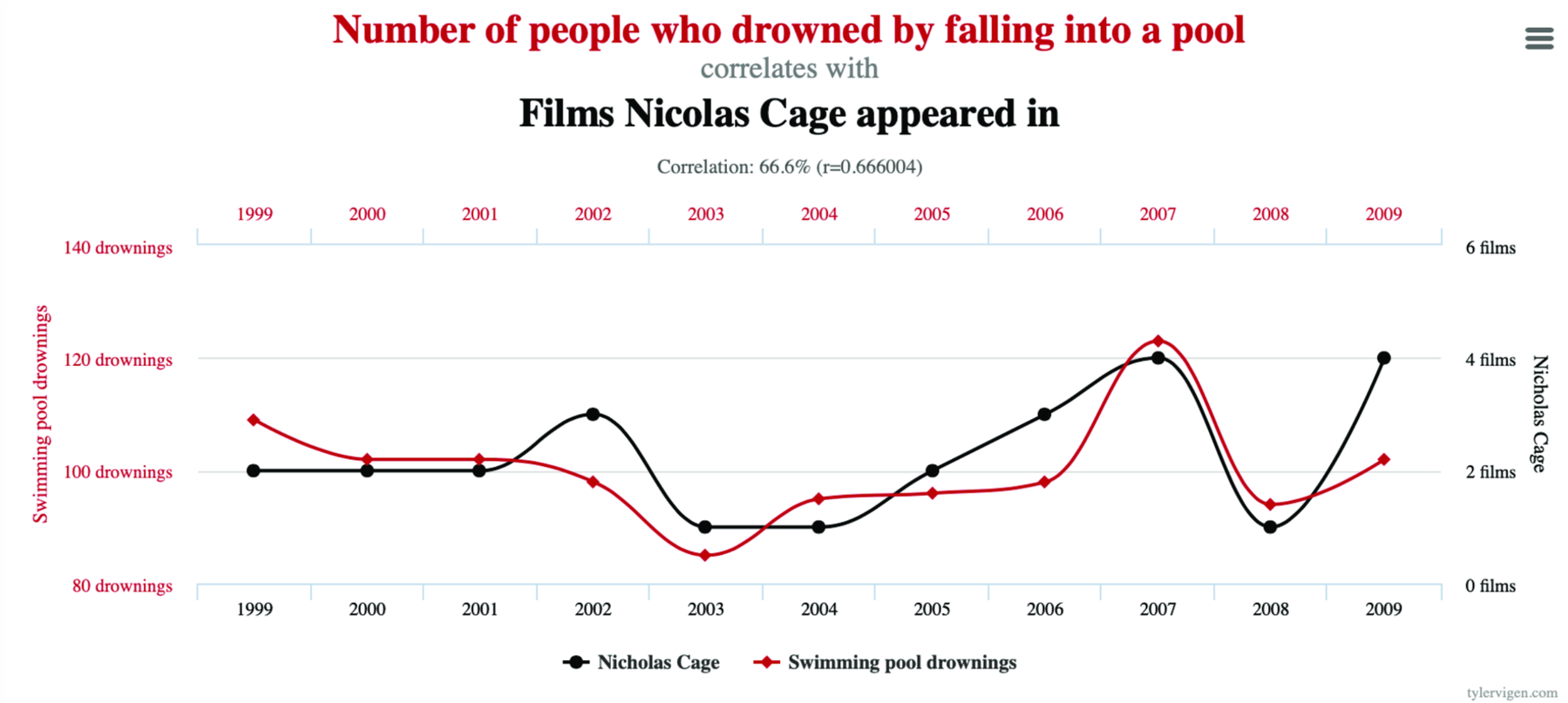

(Source: World Health Organization)

# Correlation != causality

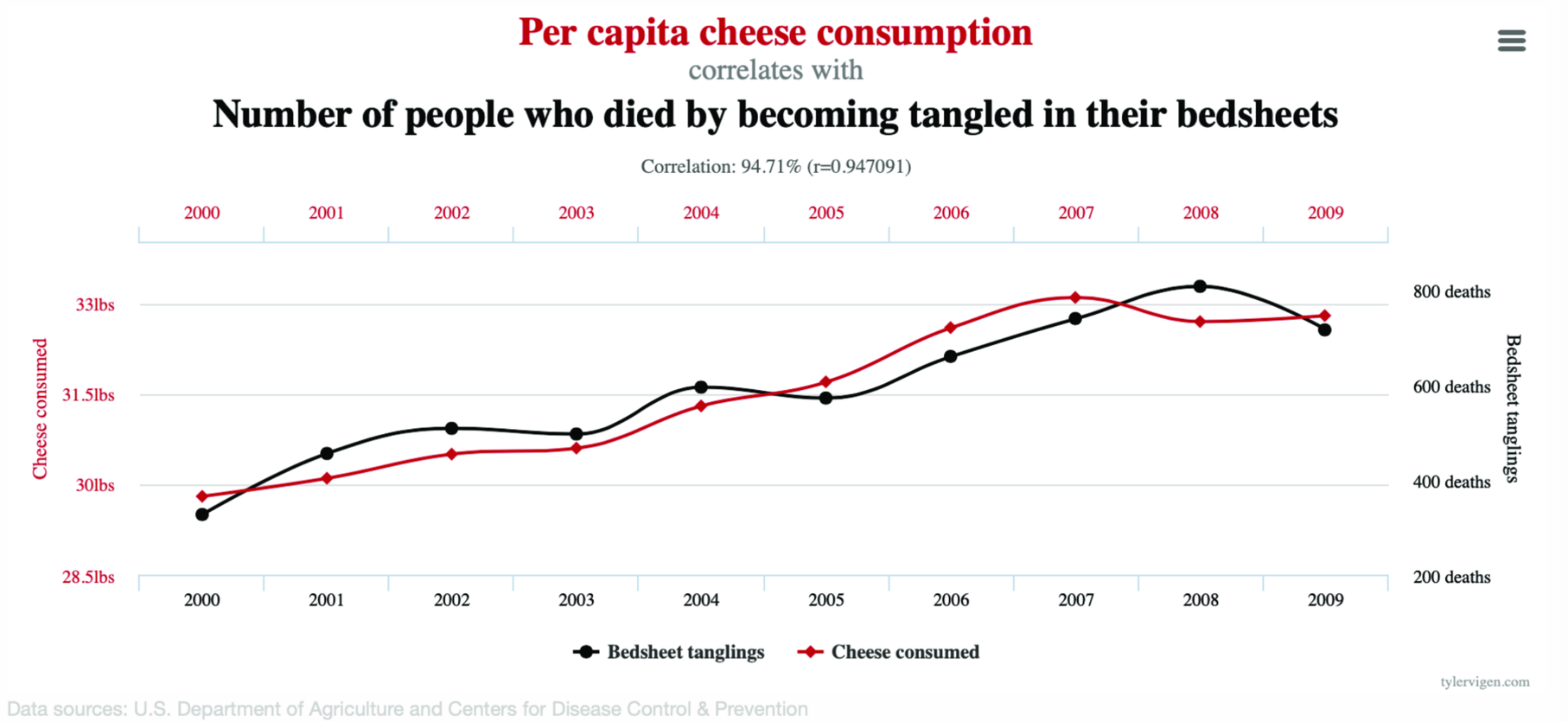





tylervigen.com/spurious-correlations



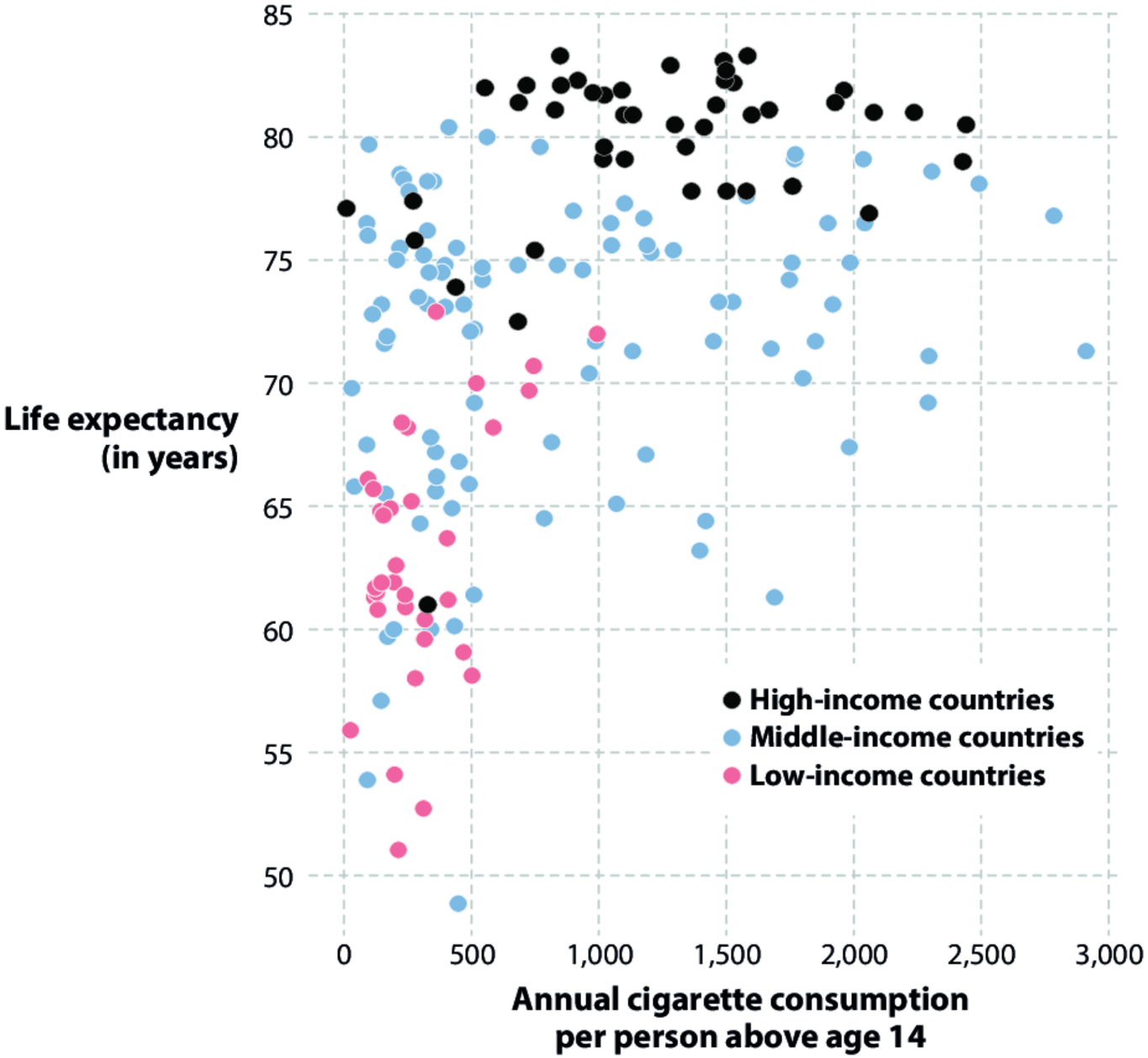

tylervigen.com/spurious-correlations

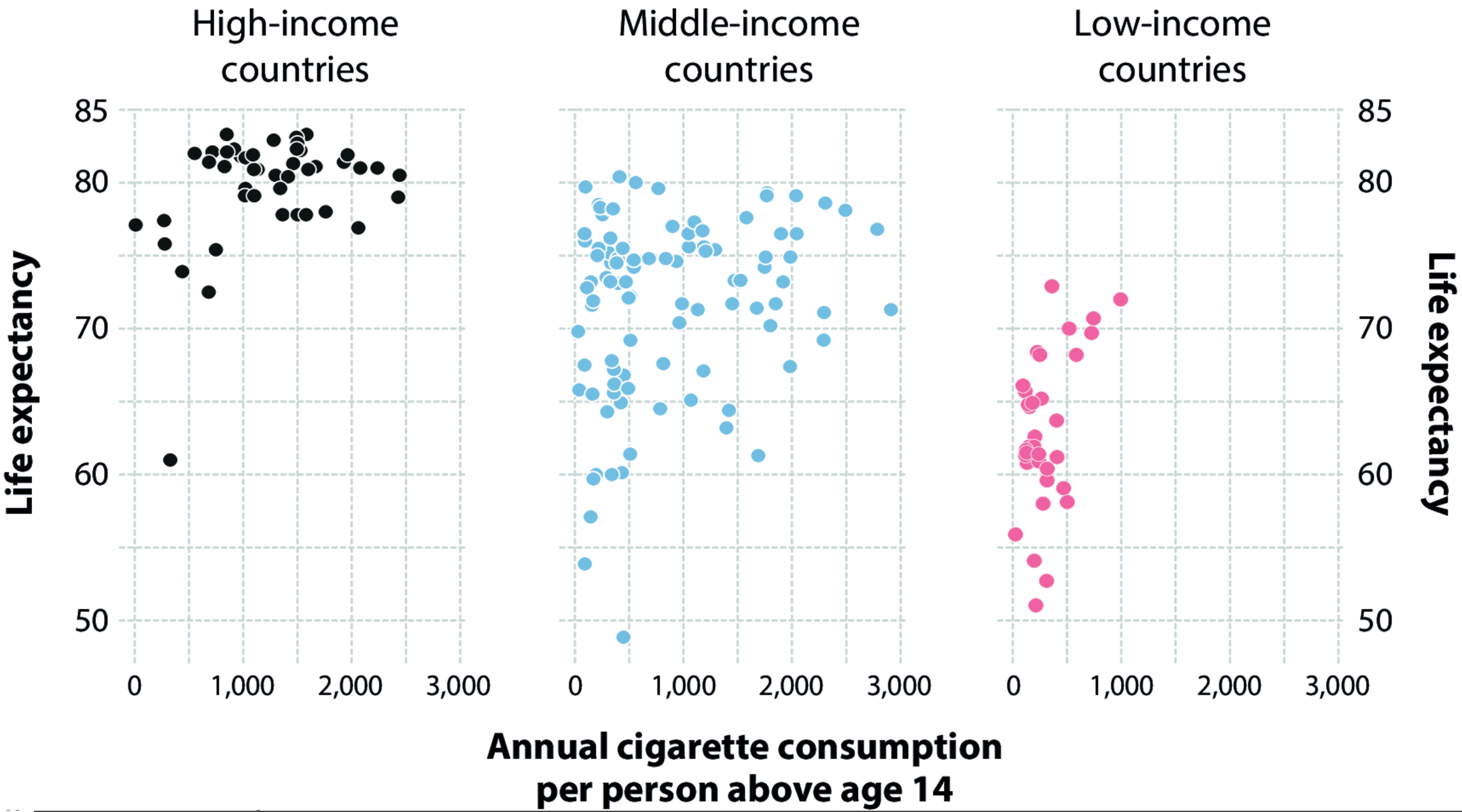


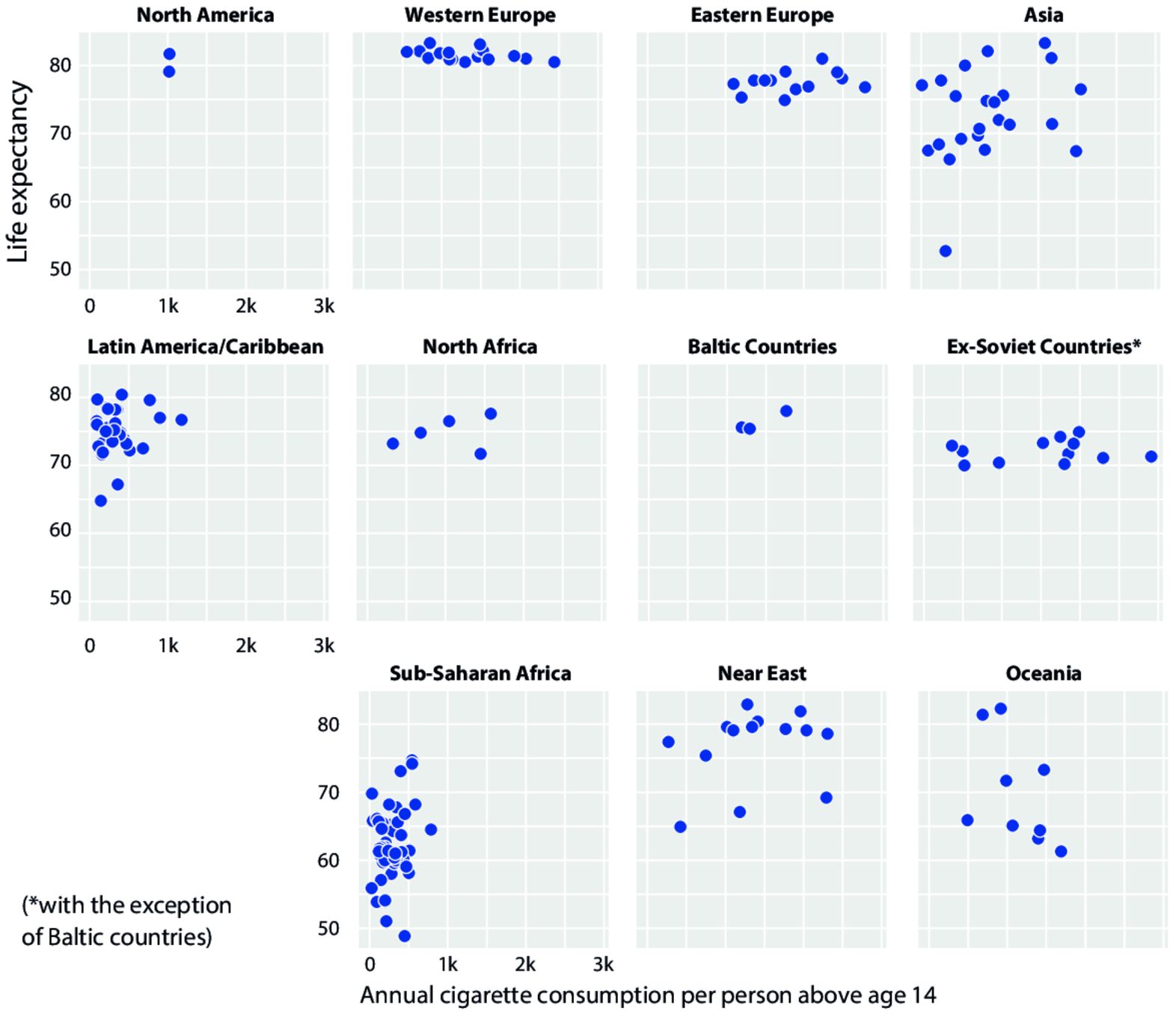

tylervigen.com/spurious-correlations

- ~~"The more cigarettes we consume, the longer we live!"~~

- "There is a positive relationship between cigarette consumption and life expectancy at a country-by-country level!"



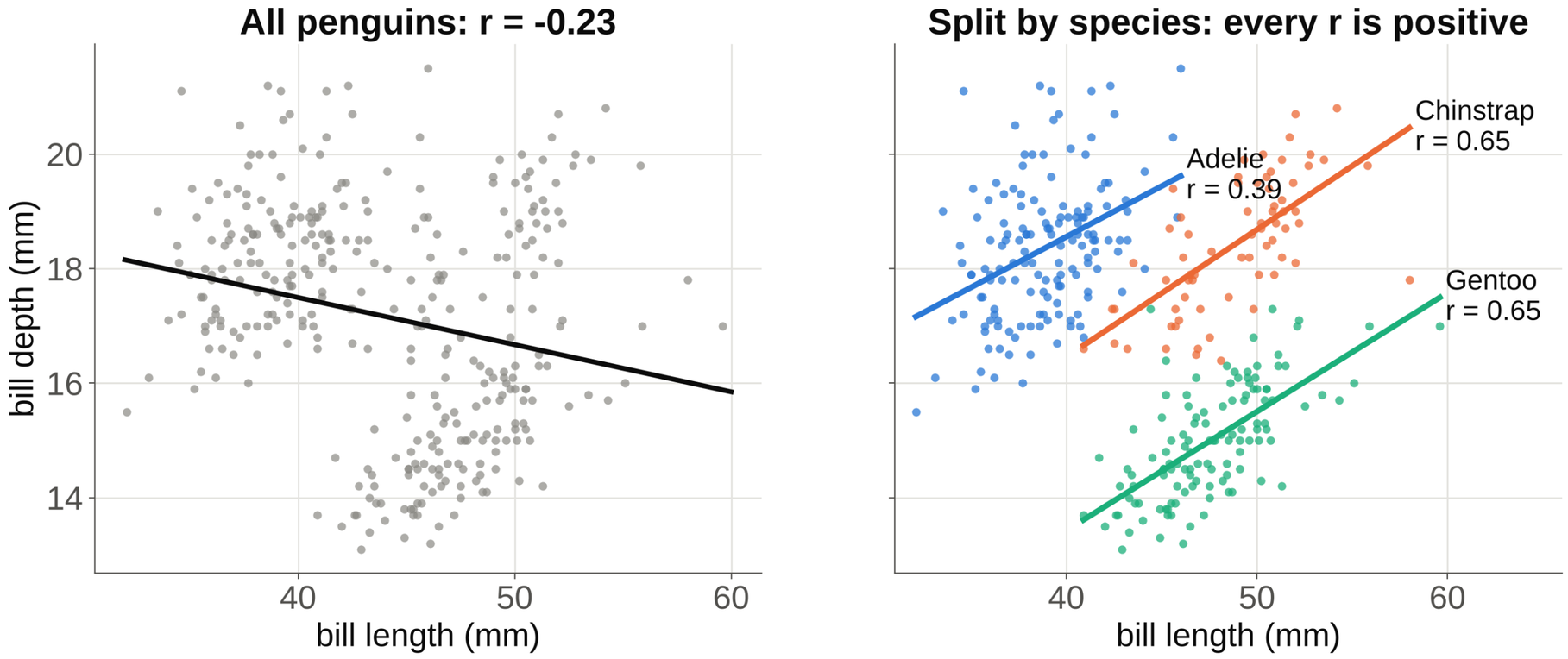

High-income countries
Middle-income countries
Low-income countries
Life expectancy
85
80
70
60
50
0
1,000
2,000
3,000
Annual cigarette consumption per person above age 14




(*with the exception of Baltic countries)

# Simpson's Paradox

- A trend that appears in every group can disappear or reverse when the groups are combined




Palmer penguins data (Gorman et al. 2014)

# Simpson's paradox: Berkeley, 1973

- Overall, men were admitted more often (44% vs. 35%)
- Yet 4 of 6 departments admitted women at a higher rate
- What happened?

| | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| **Total** | 8442 | **44%** | 4321 | 35% |

| Department | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| **A** | ***825*** | 62% | 108 | **82%** |
| **B** | ***560*** | 63% | 25 | **68%** |
| **C** | 325 | **37%** | ***593*** | 34% |
| **D** | 417 | 33% | 375 | **35%** |
| **E** | 191 | **28%** | ***393*** | 24% |
| **F** | 373 | 6% | 341 | **7%** |

Bickel, Hammel & O'Connell 1975

# Simpson's paradox: U.S. wages

- 2000–2013: median wage rose 0.9% overall…
- …but fell in every education group
- Why?

Table 1: Change in Median Wage by Education from 2000 to 201

| Segment | Change in Median Wage (%) |
|---|---|
| Overall | +0.9% |
| No degree | -7.9% |
| HS, no college | -4.7% |
| Some college | -7.6% |
| Bachelor's + | -1.2% |



**Economix**

**Explaining the Science of Everyday Life**

## Can Every Group Be Worse Than Average? Yes.

**BY FLOYD NORRIS** MAY 1, 2013 12:17 PM

Table 2: Number Employed (in millions) by Education: 2000, 2013

| Segment | Employed 2000 | Employed 2013 | Change (%) |
|---|---|---|---|
| Overall | 89.4 | 95.0 | +6.4% |
| No degree | 8.8 | 7.0 | -21.3% |
| HS, no college | 28.0 | 25.0 | -10.6% |
| Some college | 24.7 | 26.0 | +5.4% |
| Bachelor's + | 27.8 | 37.0 | +33.0% |

Floyd Norris, NYT Economix, May 1, 2013

# Aggregation level changes conclusions

- MAUP: redraw the regions and the map tells a different story
- Ecological fallacy: a country-level trend (cigarettes vs. life expectancy) says nothing about individual smokers
- Simpson's paradox: a trend inside every group can reverse when groups are pooled
- Before trusting an aggregate, ask what the units are and whether another grouping changes the answer

# Recap

- Aggregation replaces many items or attributes with fewer derived ones
- Items: counts, histograms, 2D bins, densities, boxplots, clusters; choose bins and bandwidths on purpose
- Show uncertainty, and the data behind the summary when you can
- Attributes: correlation, regression, PCA/LDA, MDS, t-SNE/UMAP; read loadings, don't over-read embeddings
- Always check what was lost: Anscombe, Datasaurus, ecological fallacy, Simpson's