Paul Rosen

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

www.sci.utah.edu

THE UNIVERSITY OF UTAH

Visualization for Data Science
DS-4630 / CS-5630 / CS-6630

FILTERING, AGGREGATION, & STATS

Reducing Items and Attributes

Filter

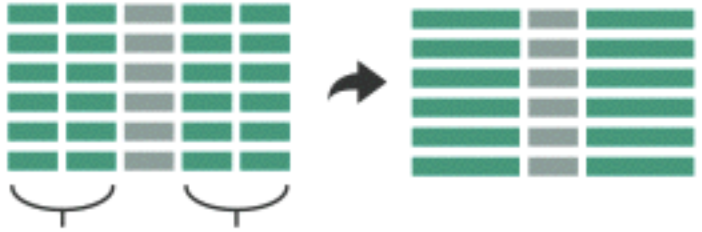
- Items

- Attributes

Aggregate

- Items

- Attributes



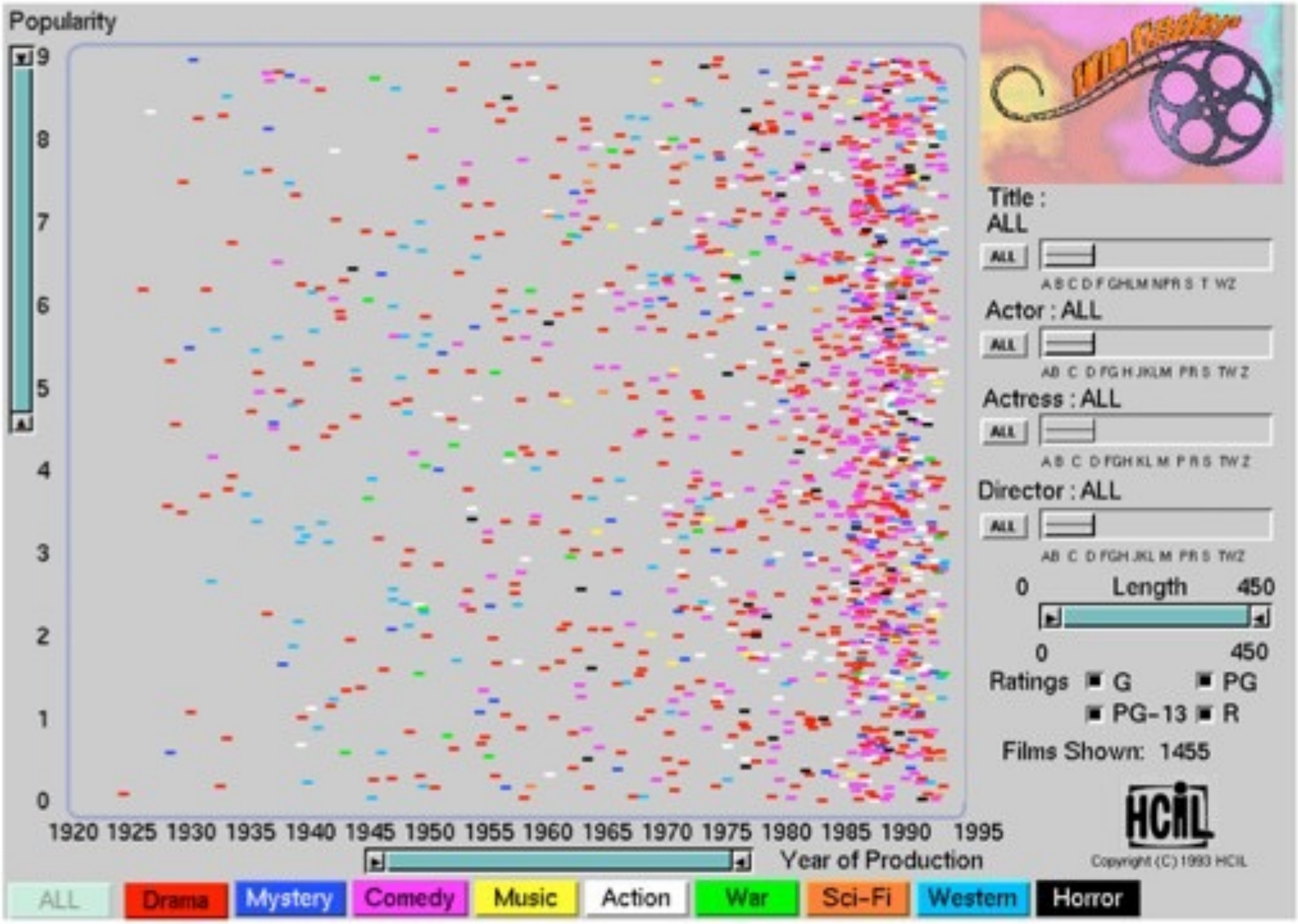
why reduce?

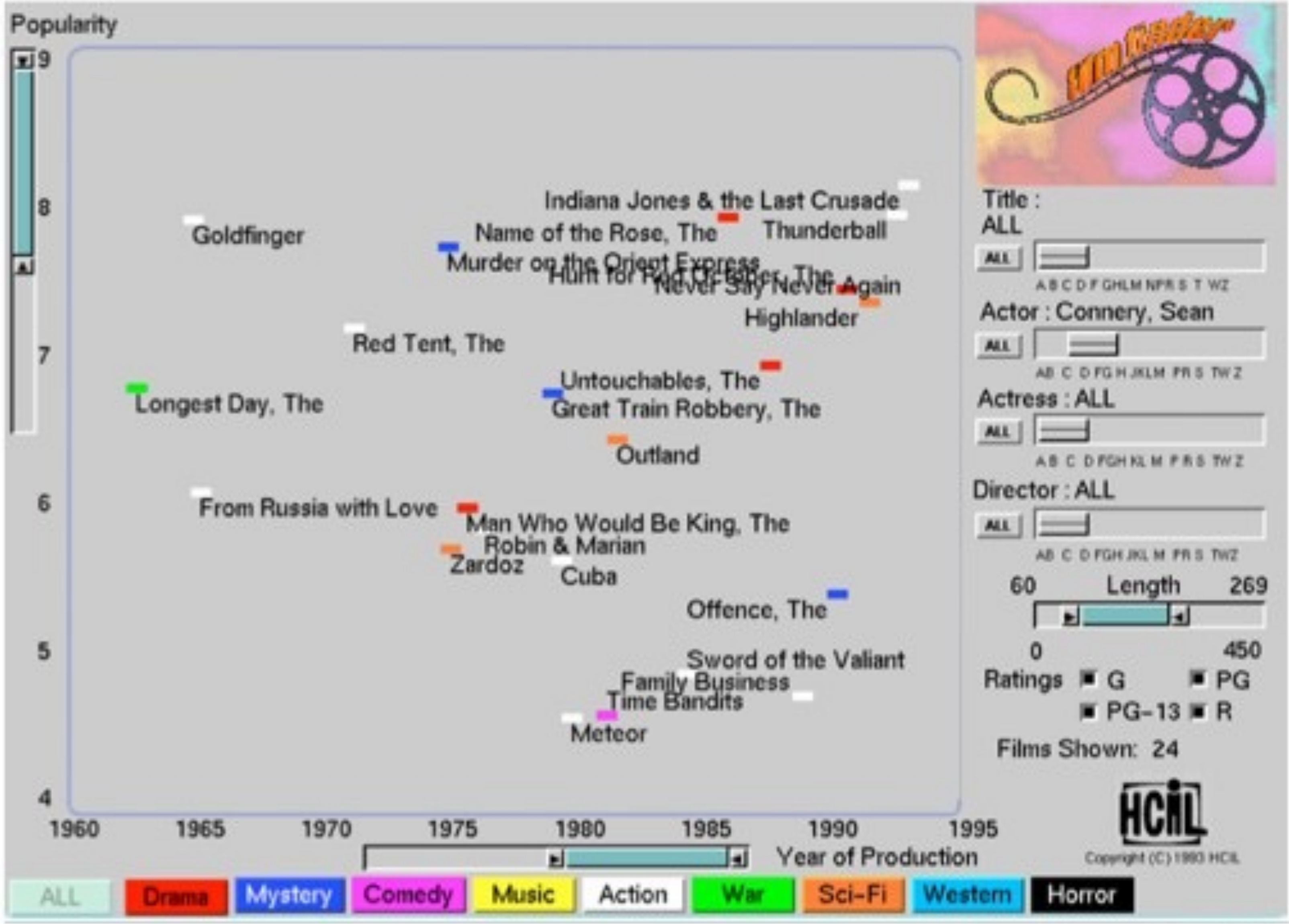
- Too many data items and/or too many attributes to focus on what is important in the data

filter

- elements are eliminated to support dynamic queries
  - coupling between encoding and interaction so that user can immediately see the results of an action

ITEM FILTERING





Ahlberg 1994

UPDATED June 25, 2012

RECOMMEND TWITTER LINKEDIN SIGN IN TO E-MAIL SHARE

New York Health Department Restaurant Ratings Map

The New York City Department of Health and Mental Hygiene performs unannounced sanitary inspections of every restaurant at least once per year. Violation points result in a letter grade, which can be explored in the map below, along with violation descriptions. The information on this map will be updated every two weeks. For menus and reviews by New York Times critics, visit our restaurants guide. Related Article »

FIND A RESTAURANT | FIND A LOCATION

Name of restaurant

FILTER

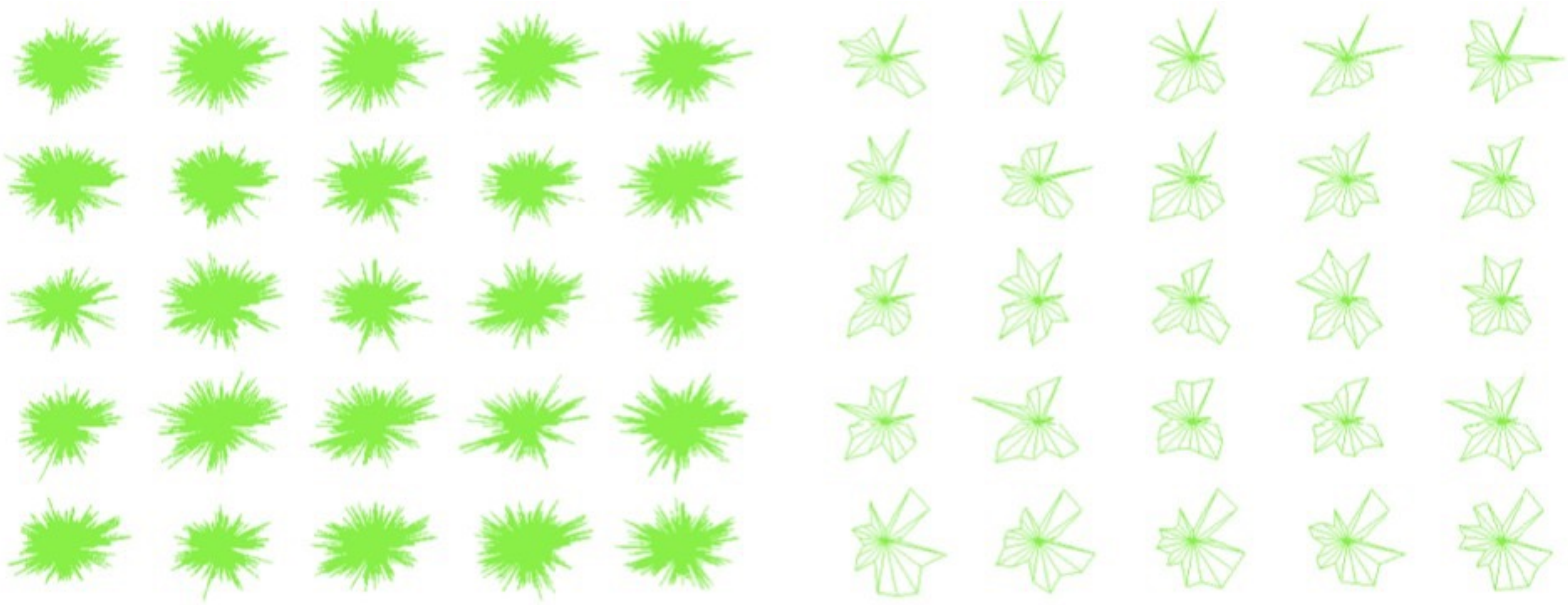
All grades | All violations | All cuisines

Restaurant locations are derived from the New York City Department of Health and Mental Hygiene database. Due to the limitations of the Health Department's database, some restaurants could not be placed.

By JEREMY WHITE | Send Feedback

Source: New York City Department of Health and Mental Hygiene

ATTRIBUTE FILTERING



• Wang 2003

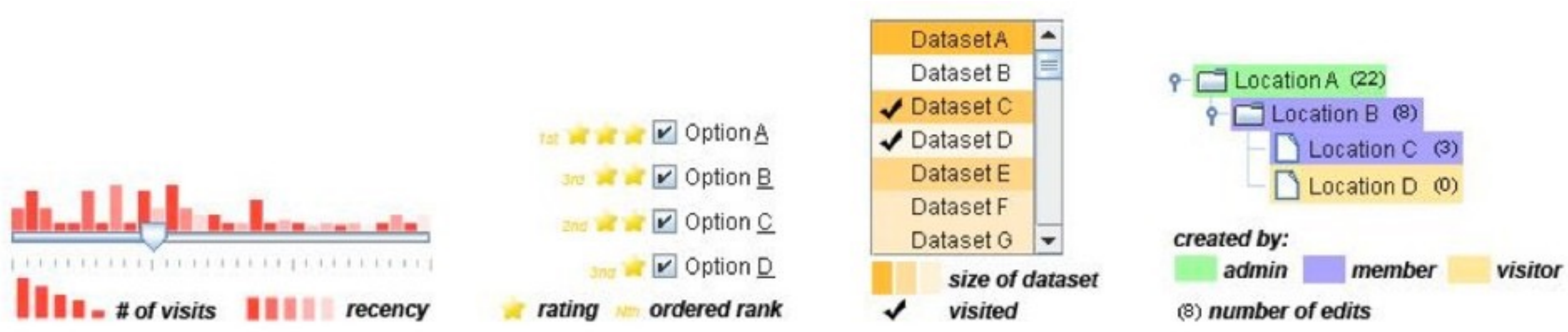
Controlling filtering

- Driven by 2 approaches
  - Widget-based filtering

  - Visualization-based filtering

Controlling Filtering: scented widgets

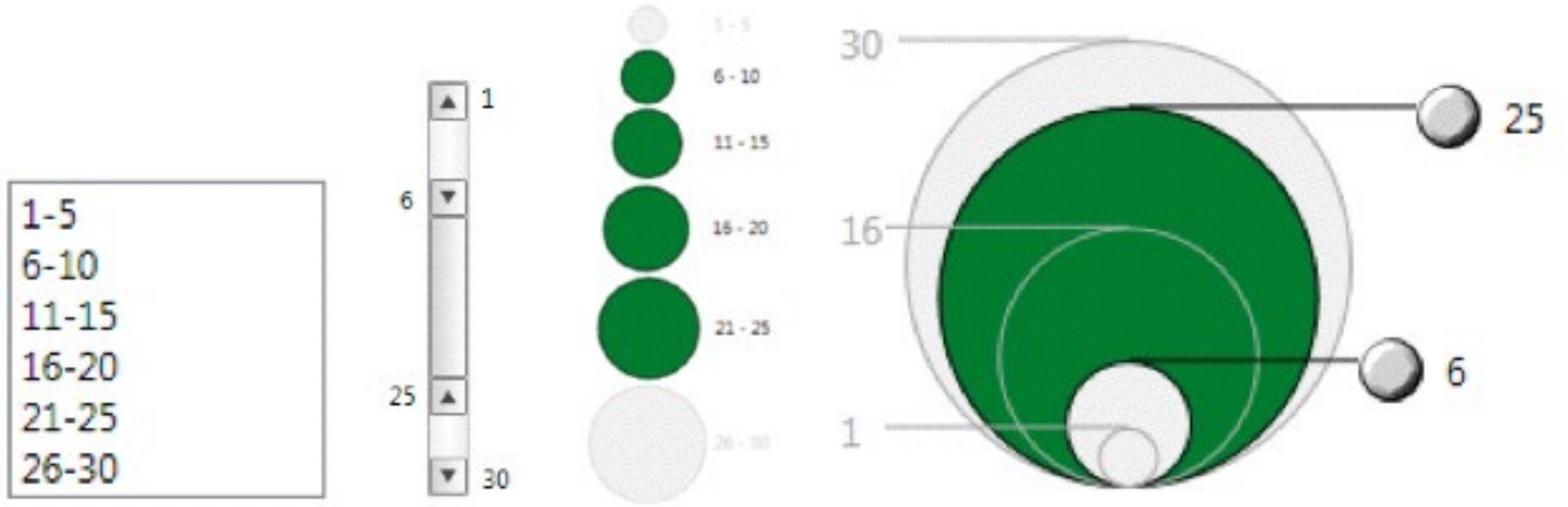
- information scent: user gets sense of data
- GOAL: lower the cost of information forging through better cues



- Willett 2007

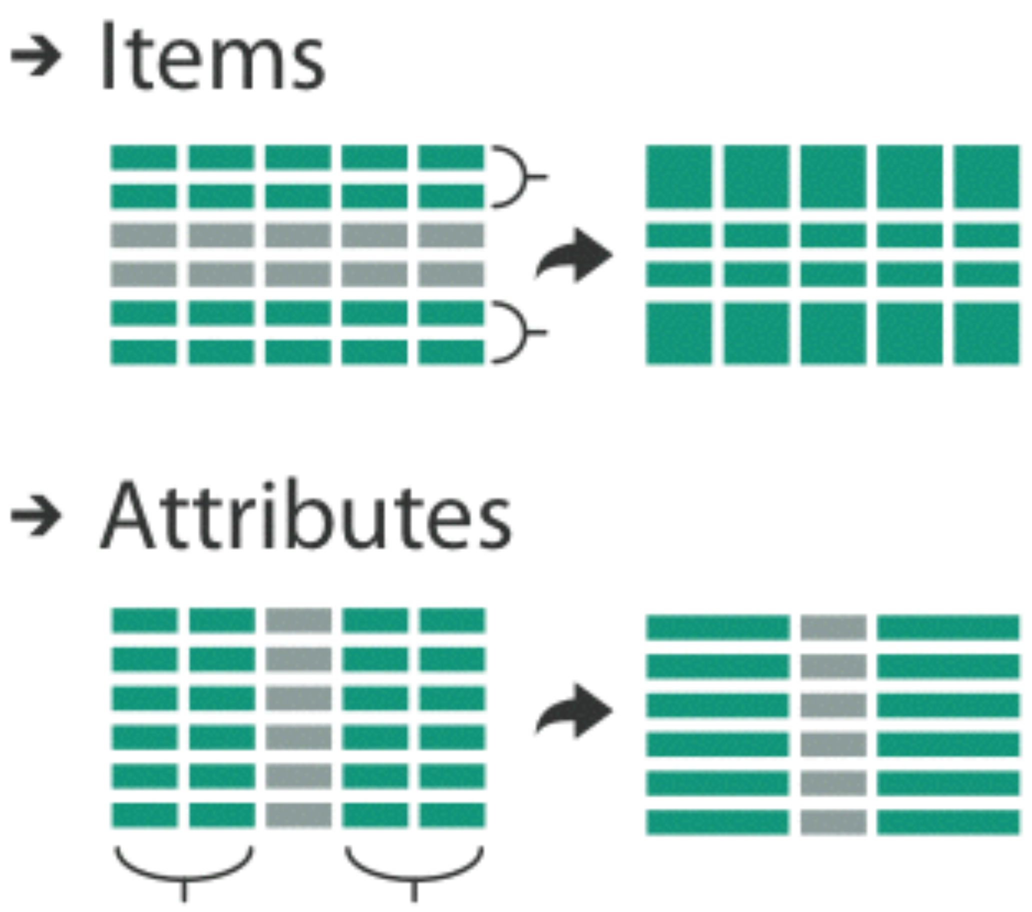
Controlling Filtering: interactive legends

- controls combining the visual representation of static legends with interaction mechanisms of widgets
- **define and control visual display together**



aggregate

- a group of elements is represented by a new derived element that stands in for the entire group

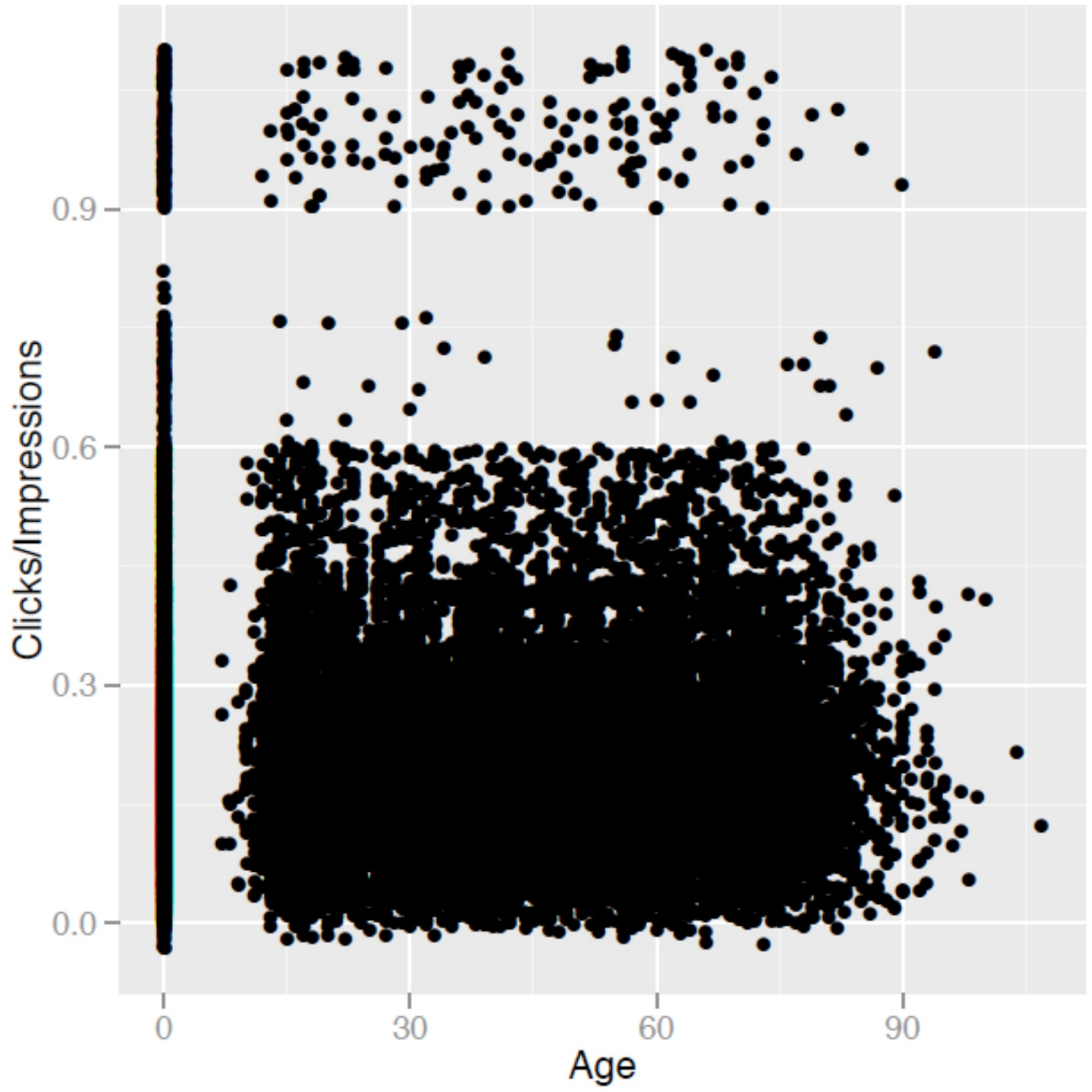


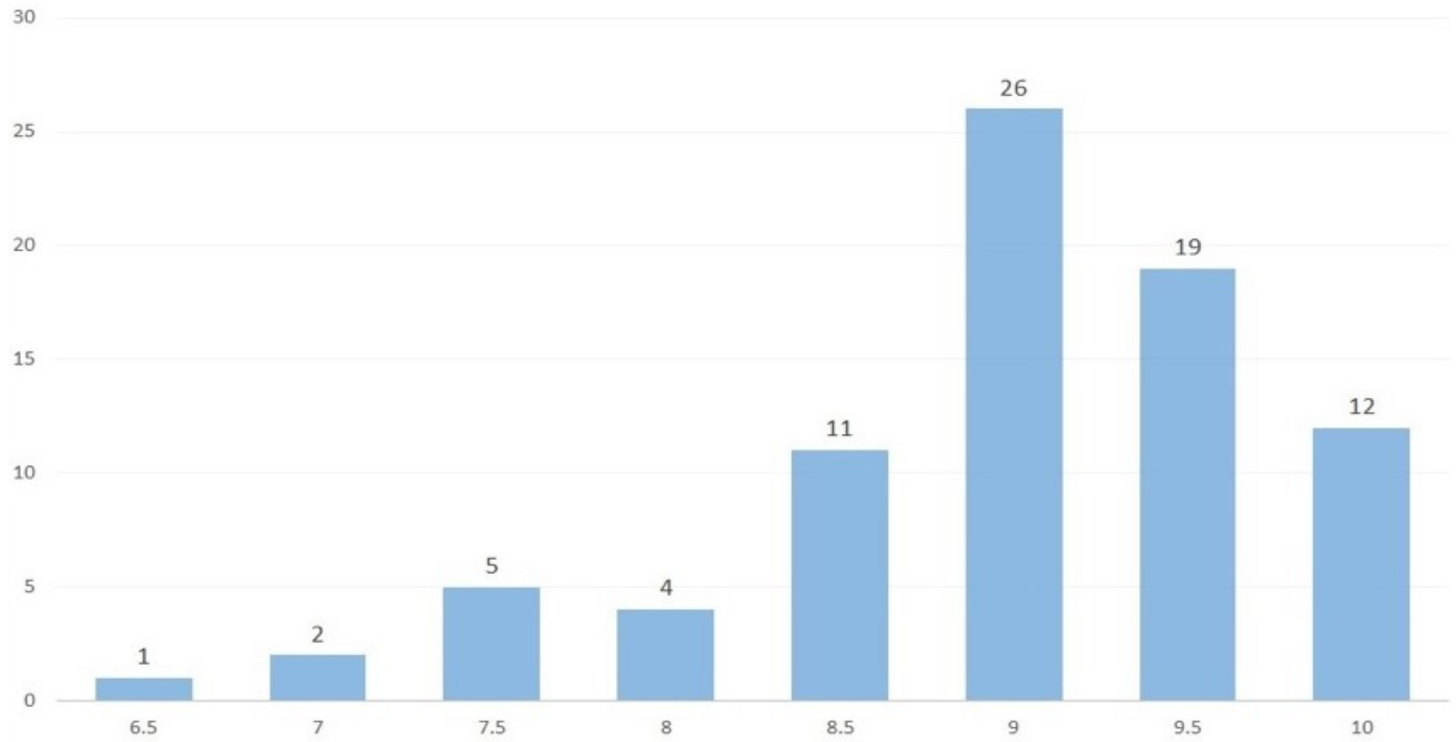
Numerous ways to reduce…

- statistics, topology, machine learning, etc.

Problem #1: Aggregate Items

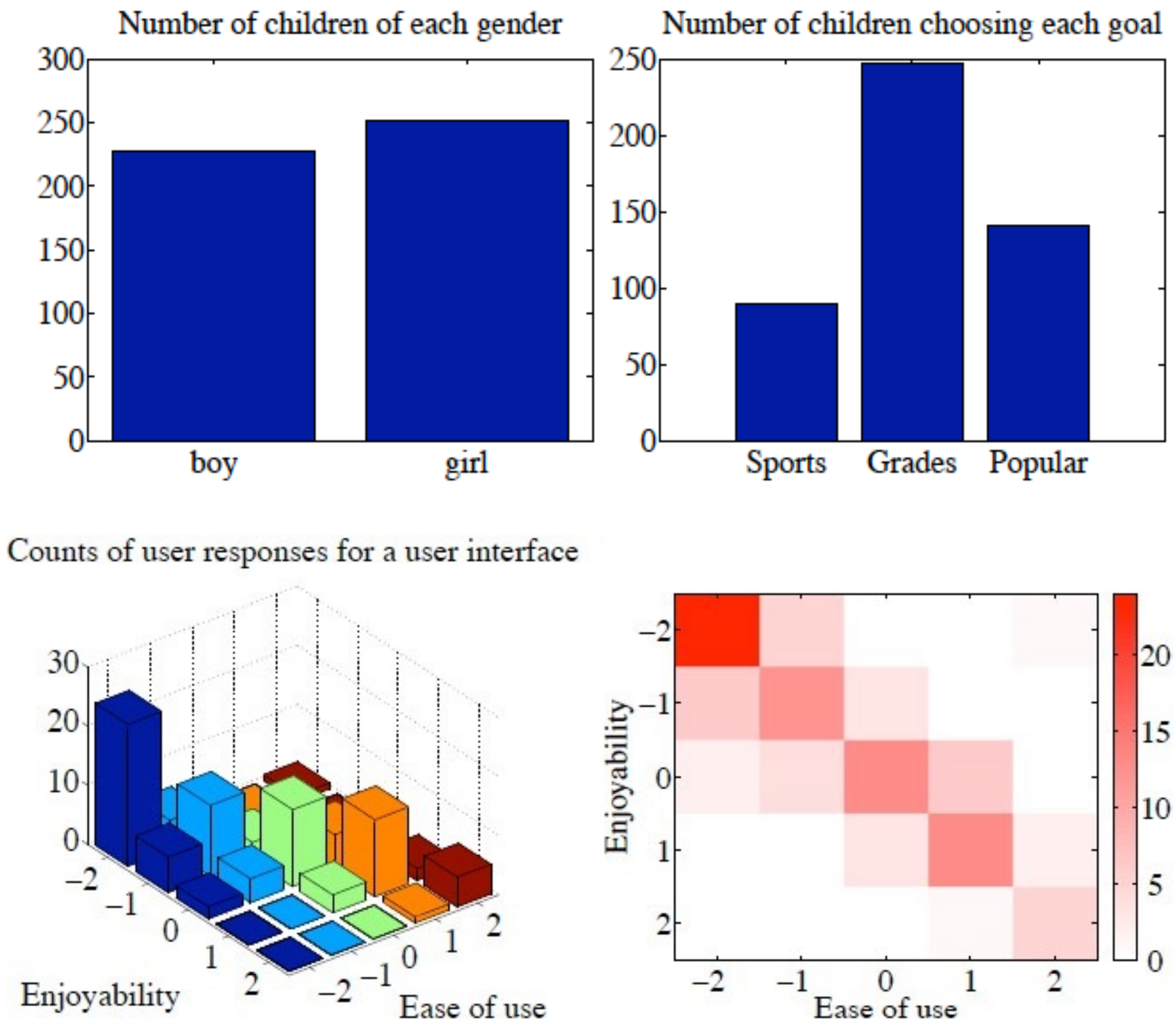
- We have too many data points to show





Histograms

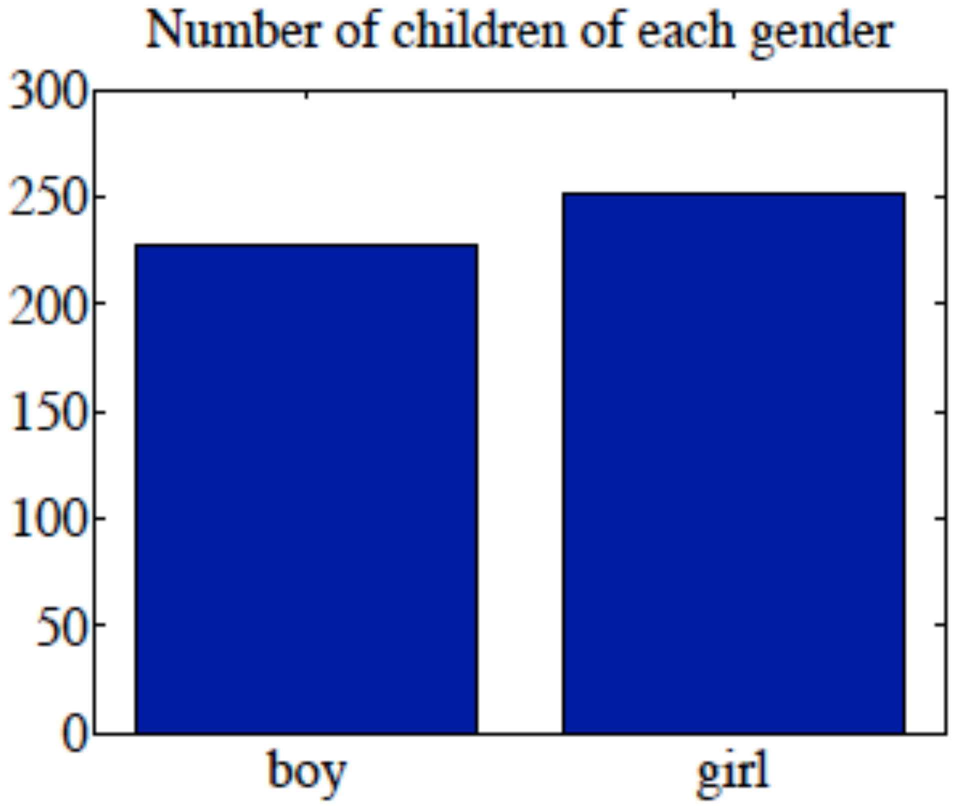
- Generally referring to a bar chart-based visualization that allows evaluating distribution of values.
- Really, histograms capture a distribution of data

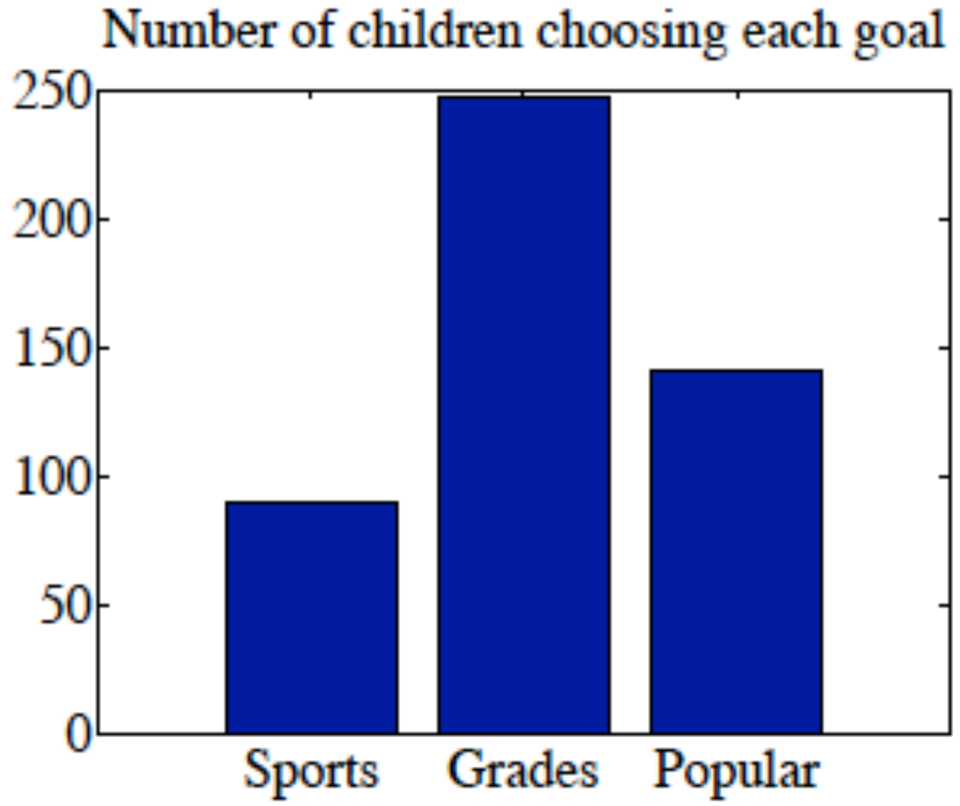


Categorical data

- 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

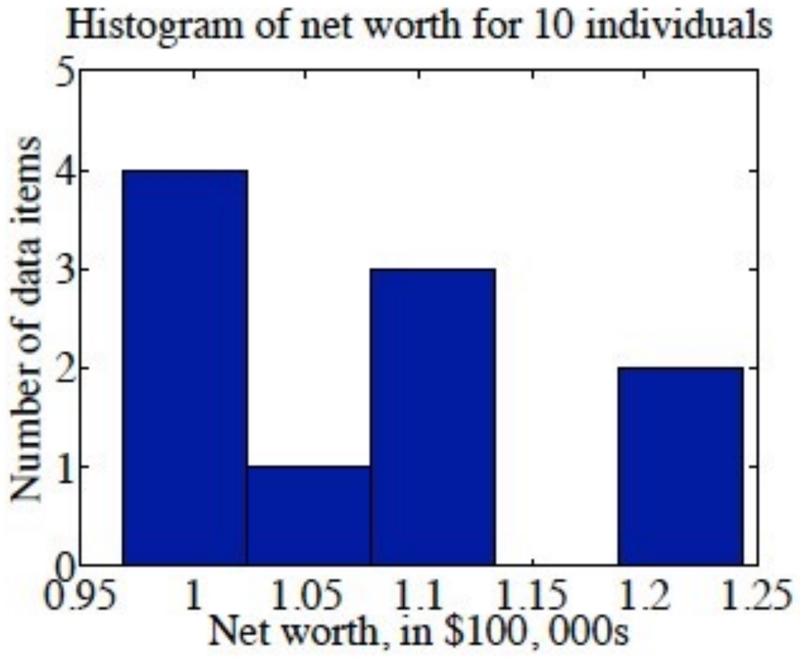


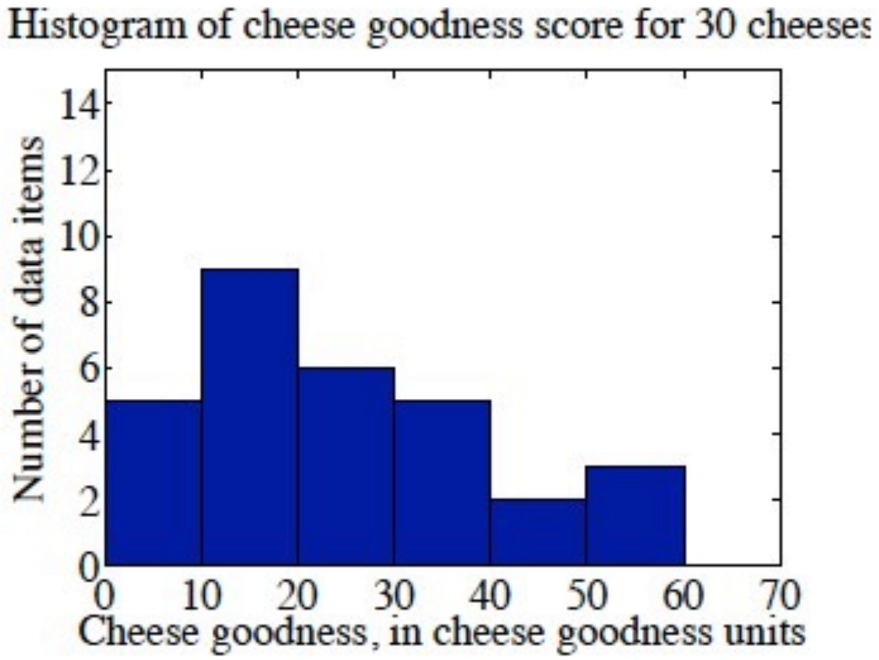


Continuous Data Histograms

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

- $bin_i = k * (x_i - \min X) / (\max X - \min X)$

Calculating a continuous histogram

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

Calculating a continuous histogram

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

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

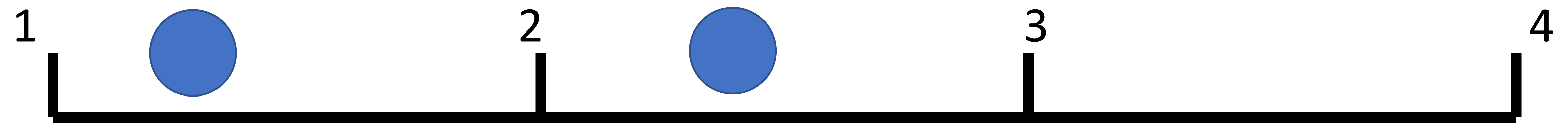
- X={1,2.5,3,4}
- k = 3
- bin_i = floor(3 * (x_i – 1) / (4 – 1))

Calculating a continuous histogram

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

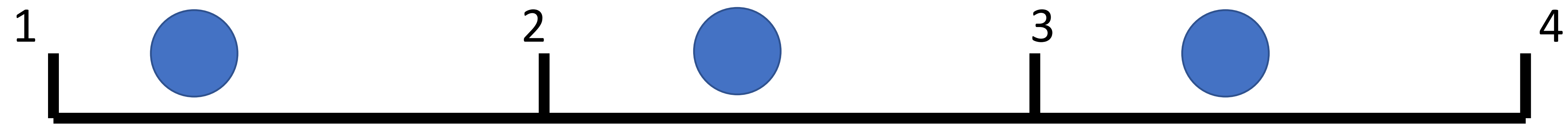
Calculating a continuous histogram

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



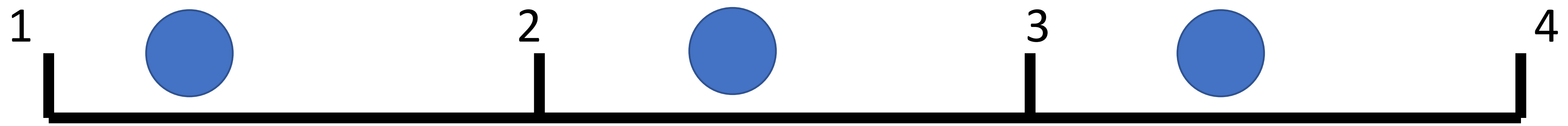
Calculating a continuous histogram

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



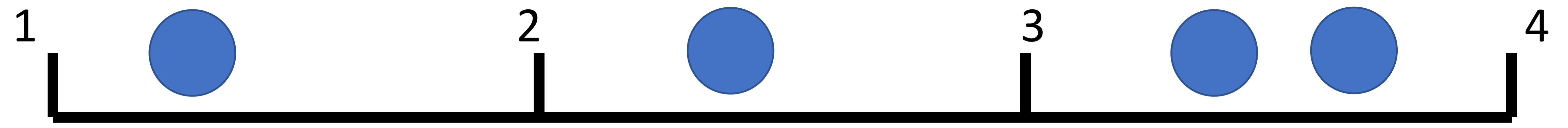
Calculating a continuous histogram

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



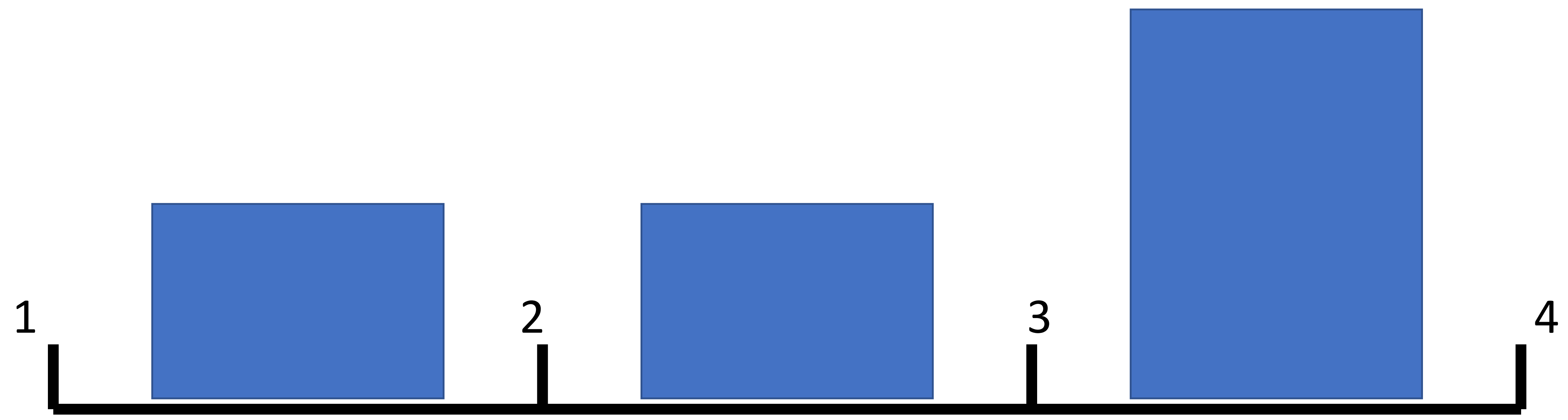
Calculating a continuous histogram

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

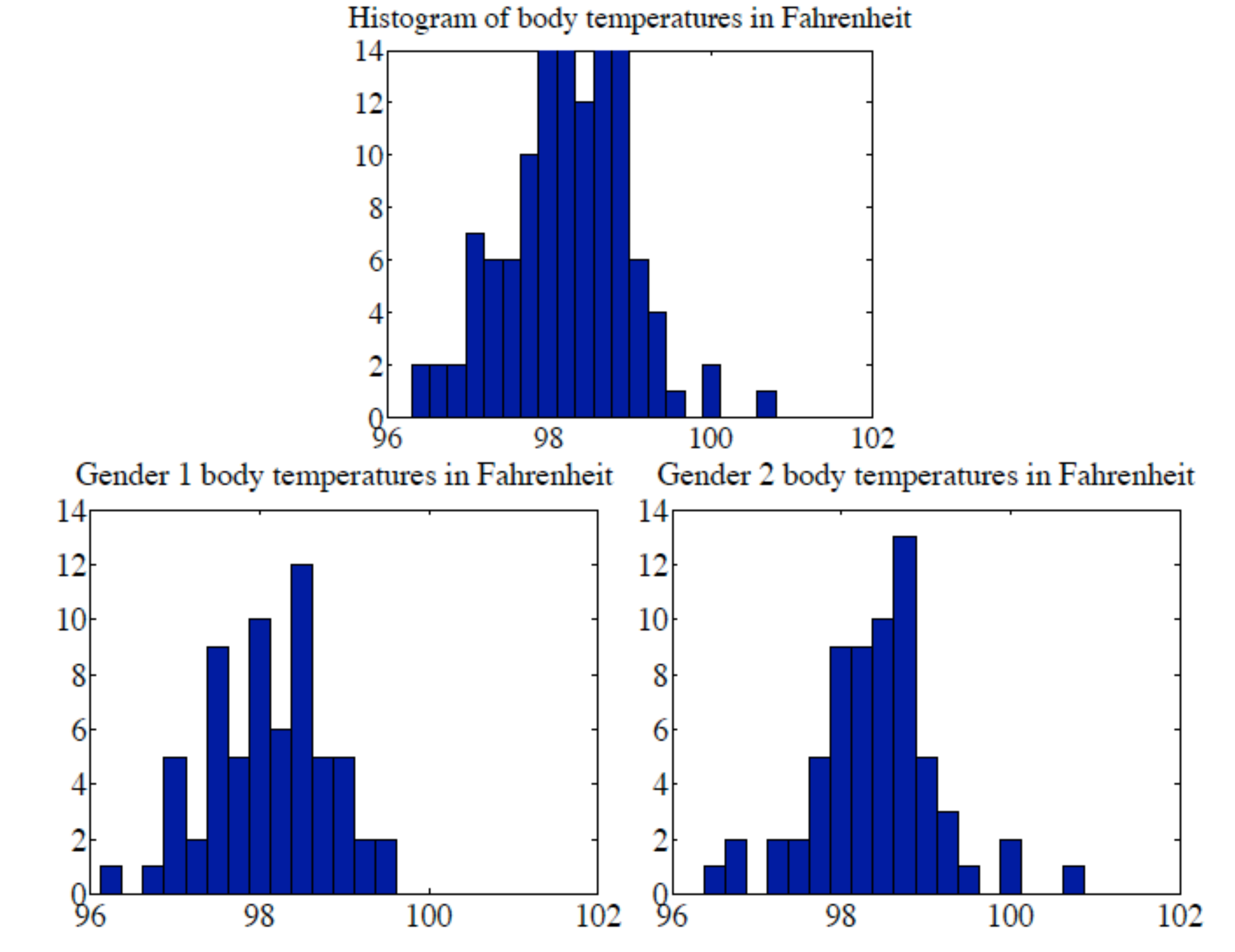


Calculating a continuous histogram

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



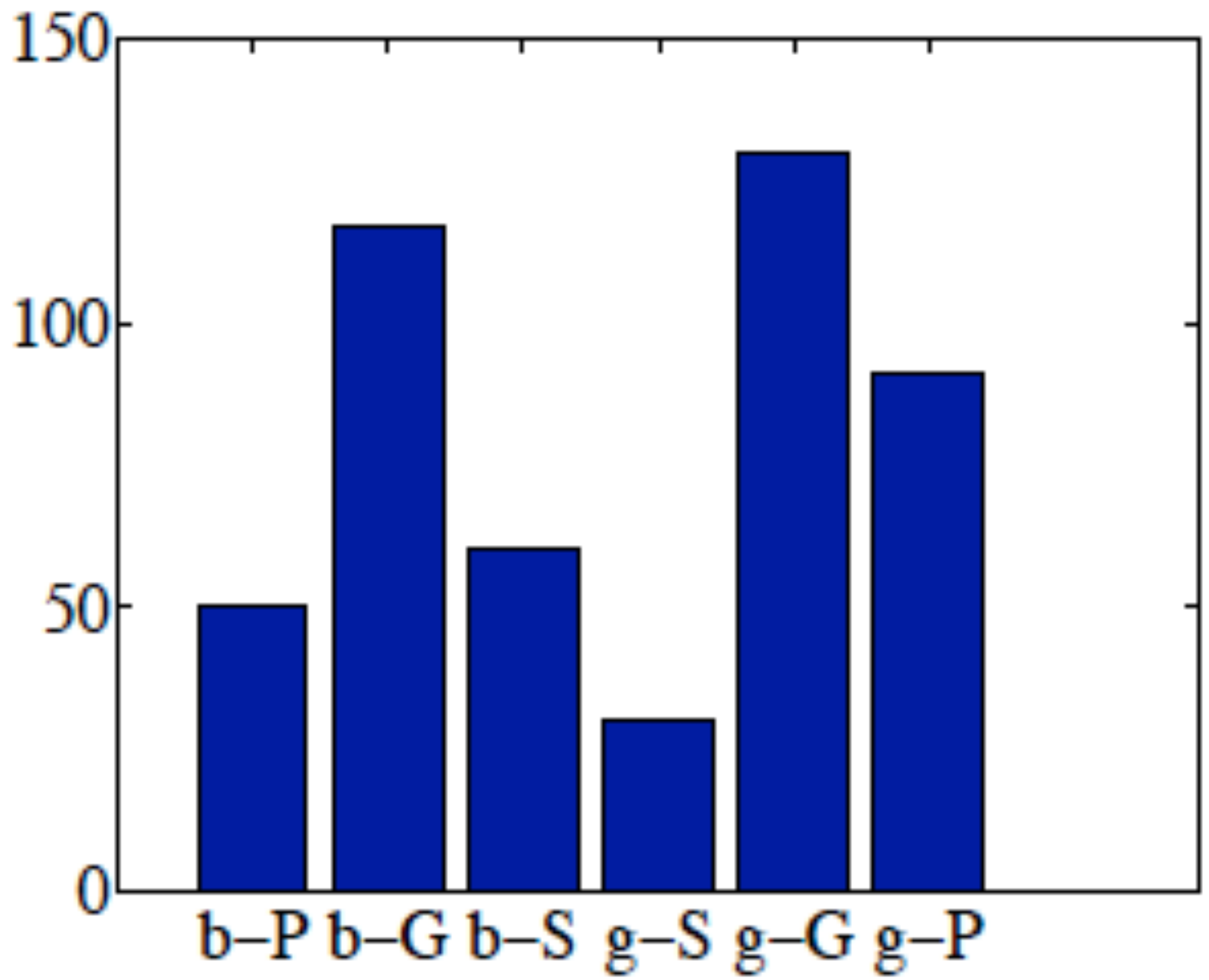
Conditional Histograms



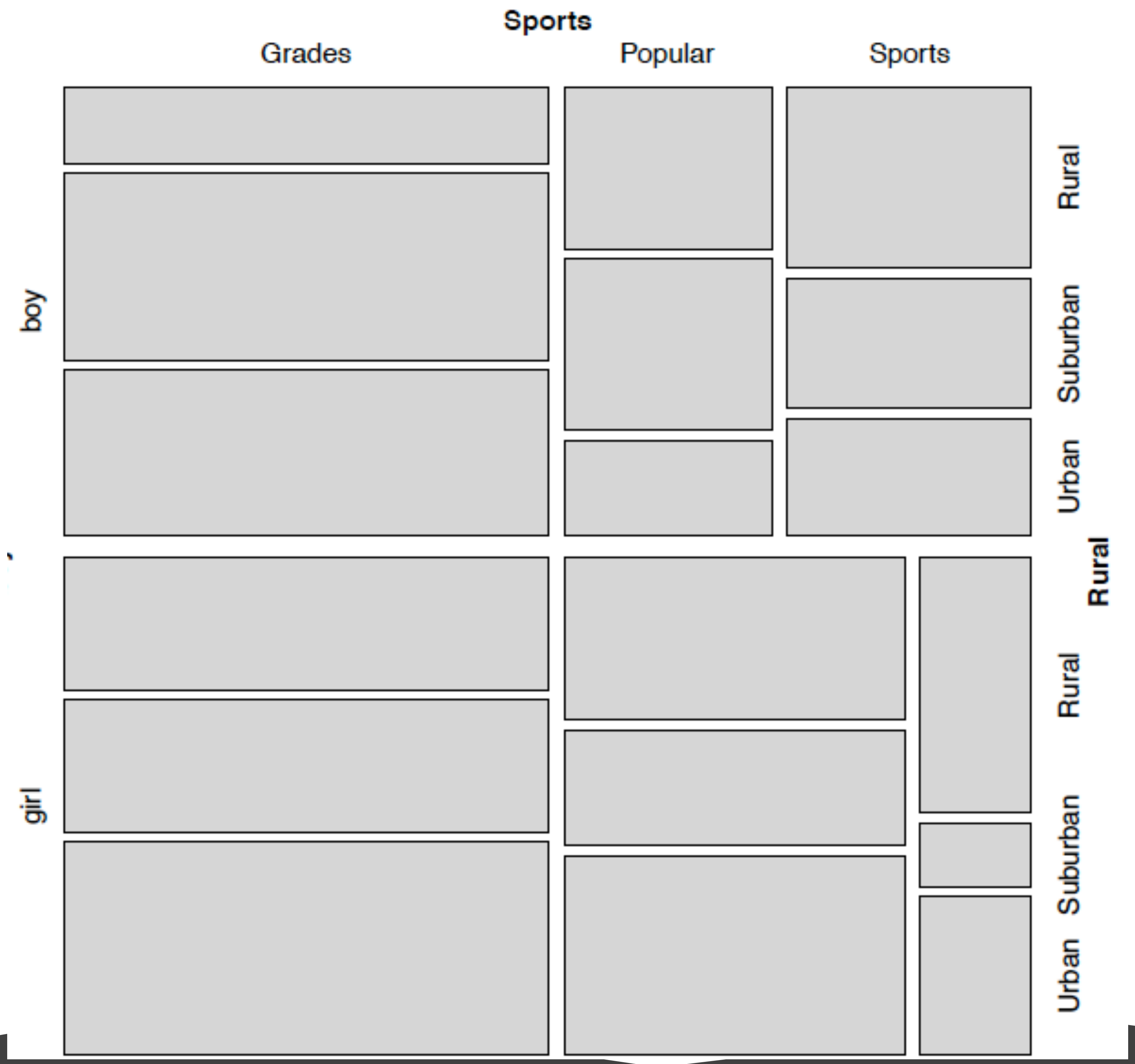
2D Histograms

Categorical data

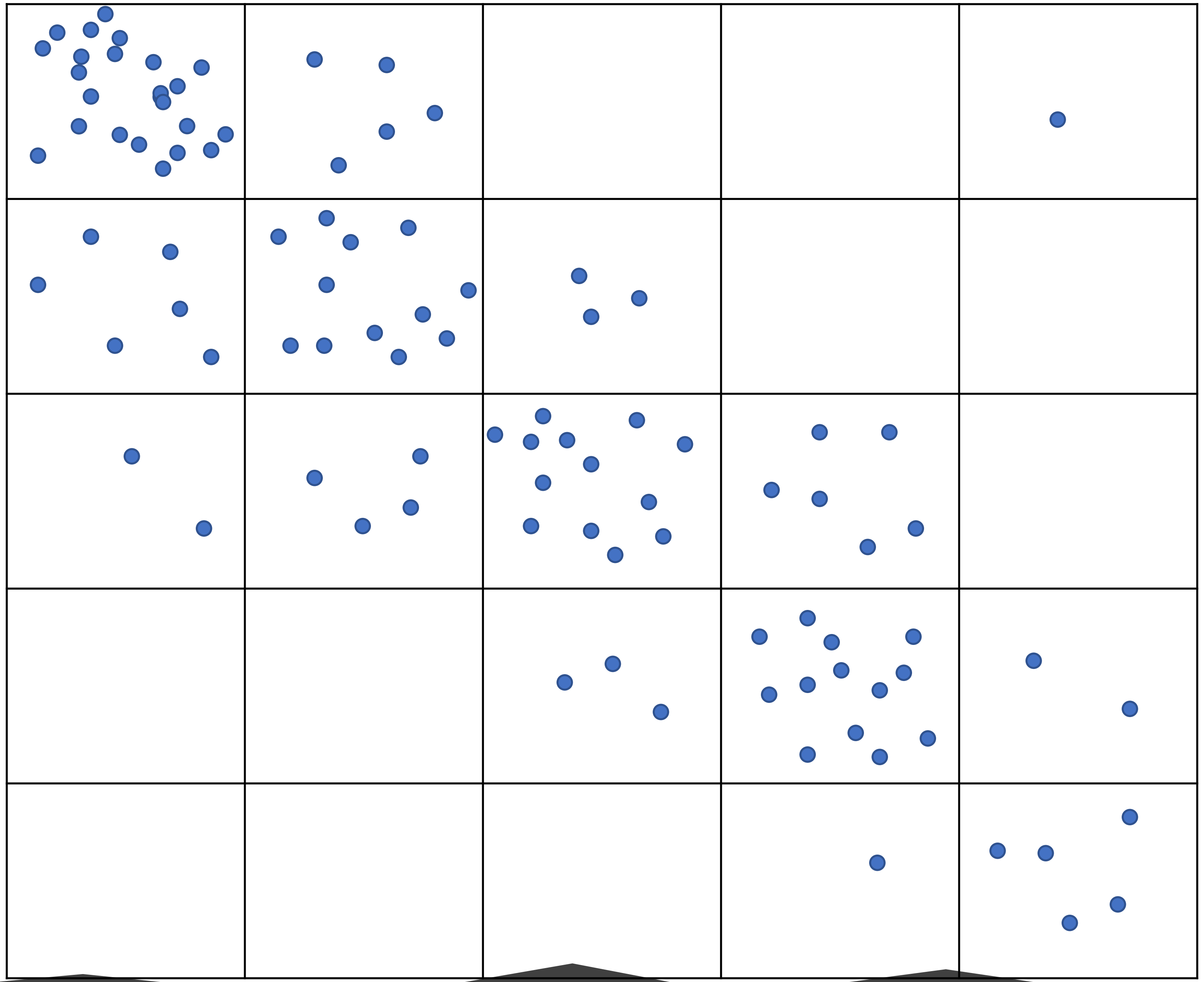
Gender	Goal	Gender	Goal
boy	Sports	girl	Sports
boy	Popular	girl	Grades
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	girl	Grades
girl	Popular	girl	Sports
girl	Grades	girl	Popular
girl	Sports	girl	Grades
girl	Sports	girl	Sports



Mosaic Plots

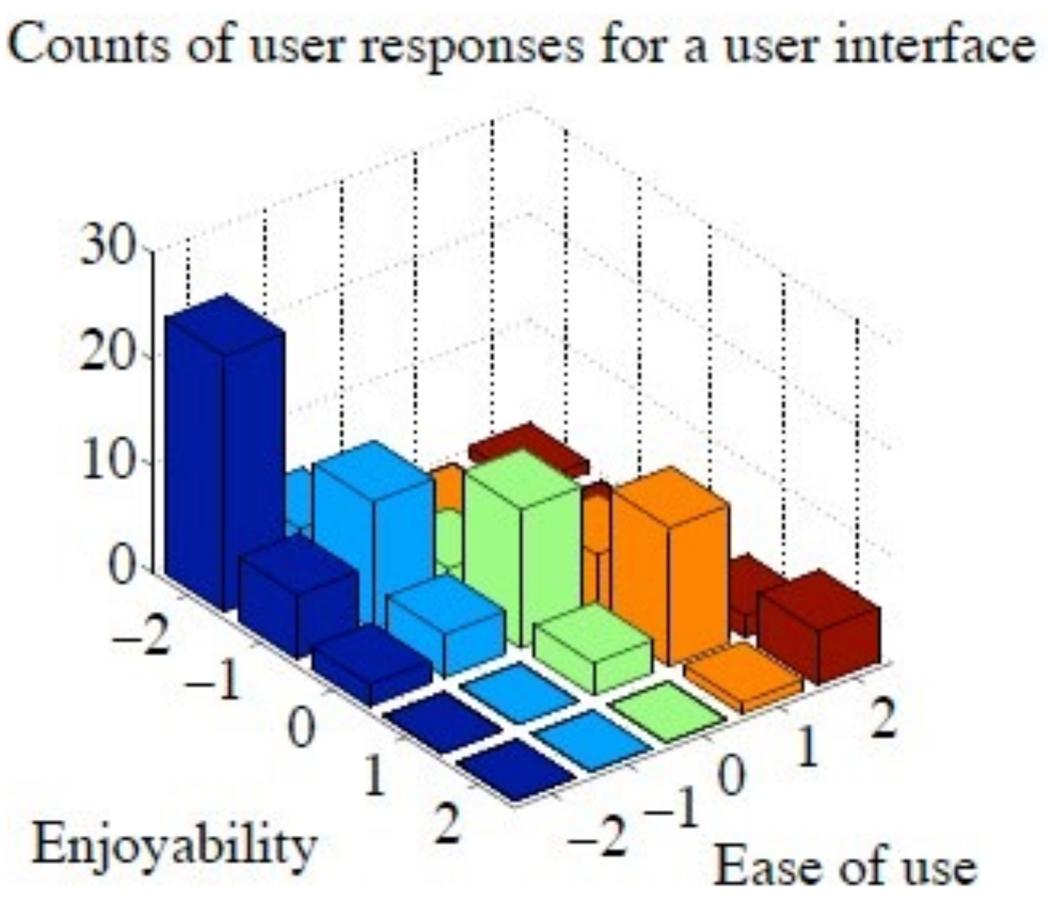


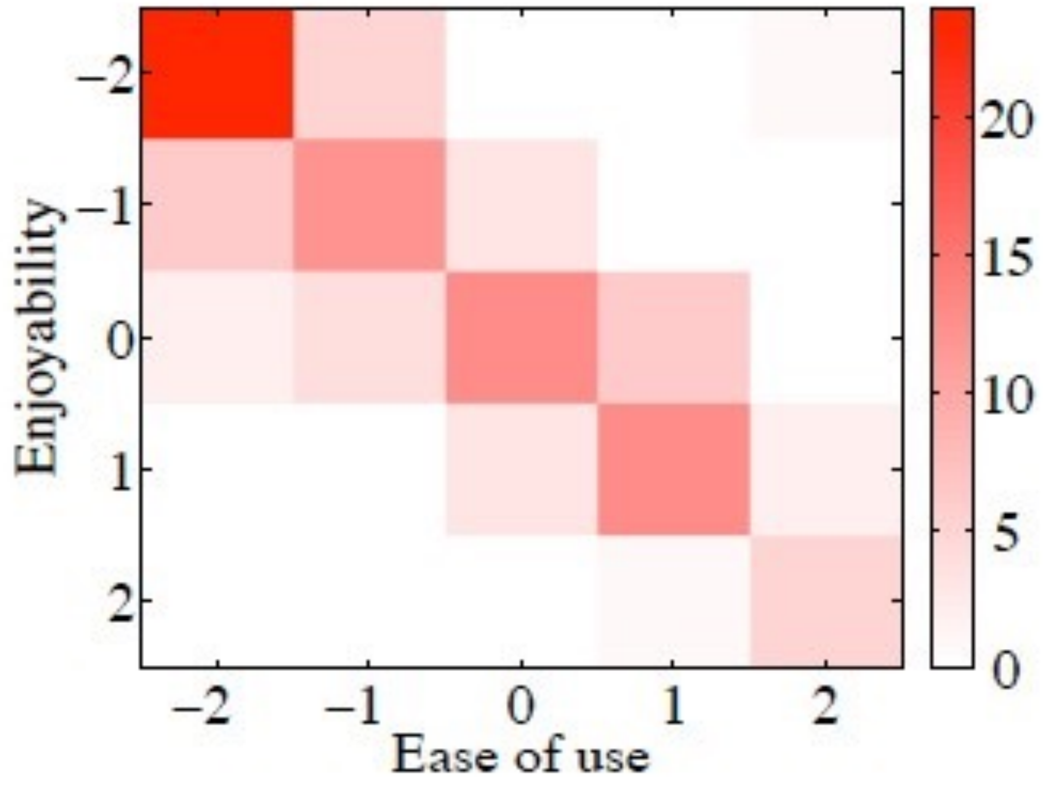
Ordinal data



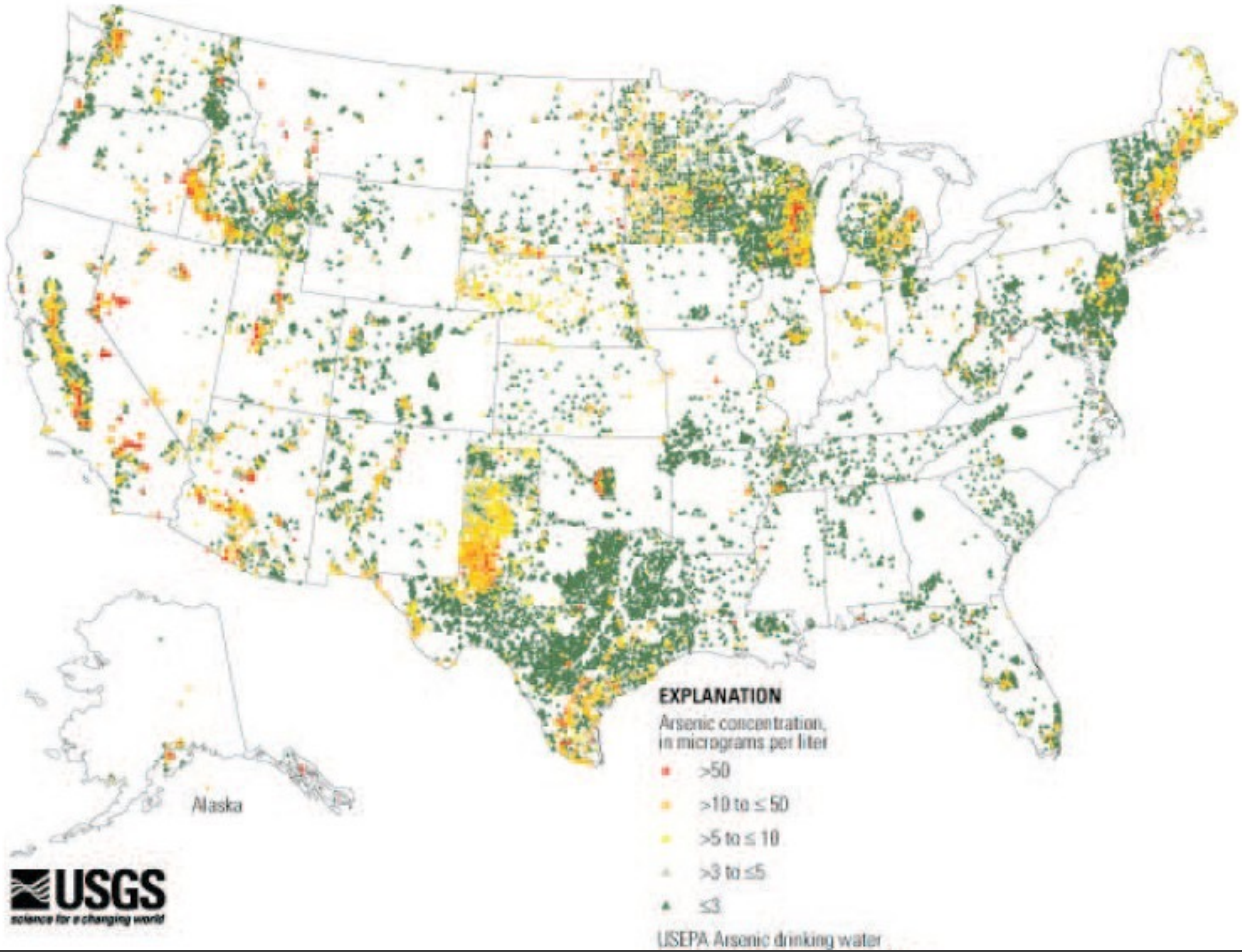
	-2	-1	0	1	2
-2	24	5	0	0	1
-1	6	12	3	0	0
0	2	4	13	6	0
1	0	0	3	13	2
2	0	0	0	1	5

Ordinal data

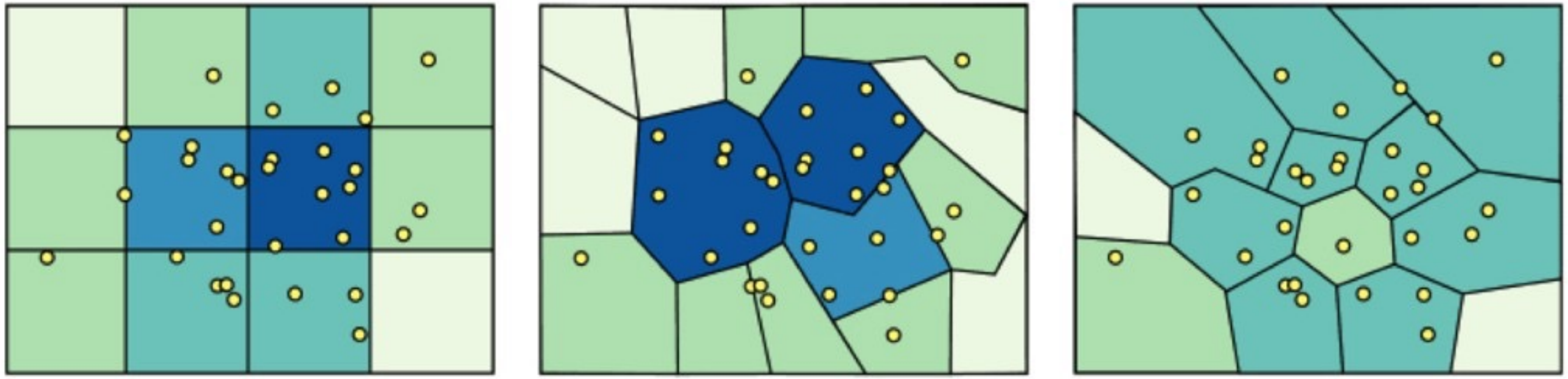




Arsenic in well water

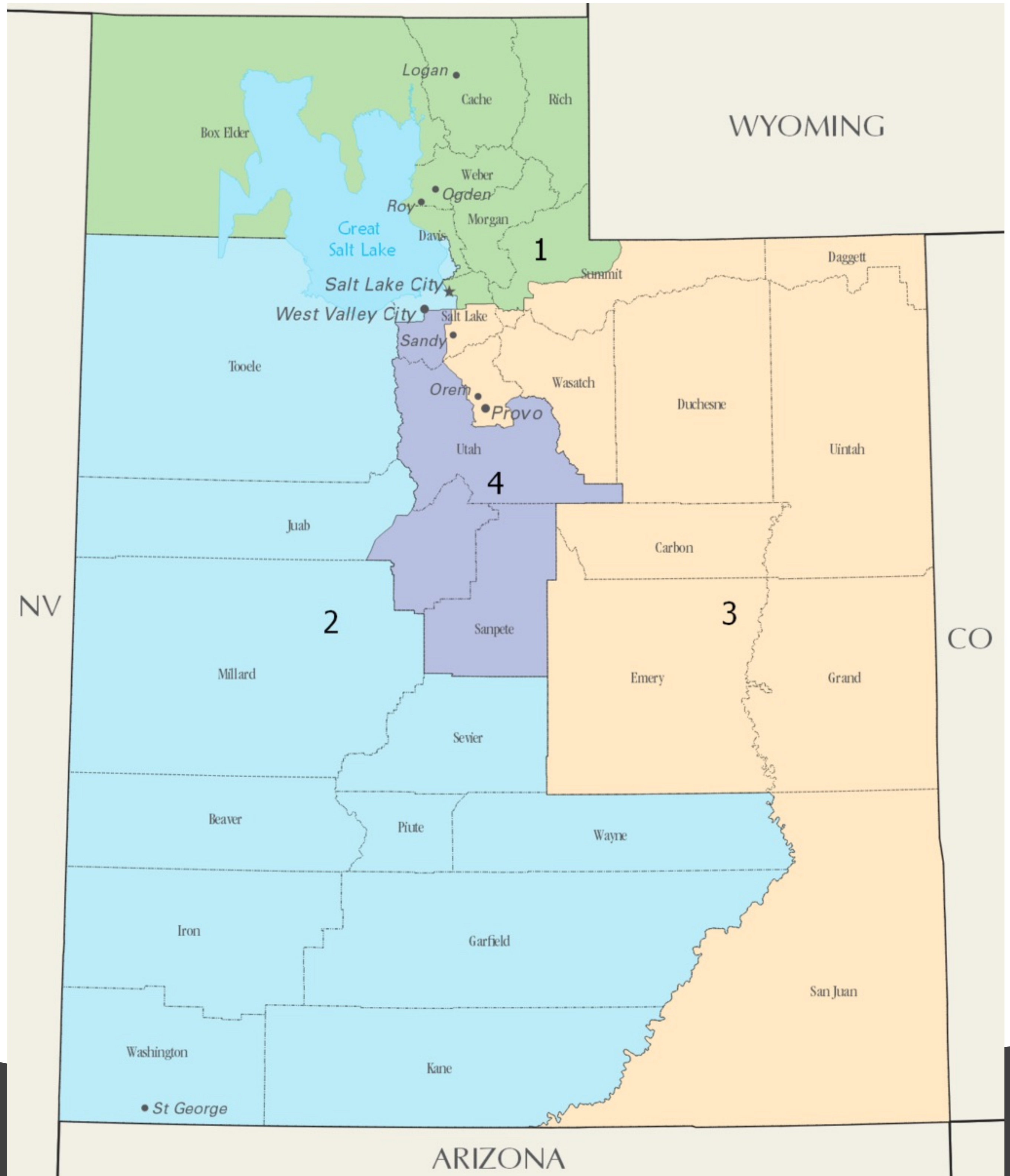


spatial aggregation



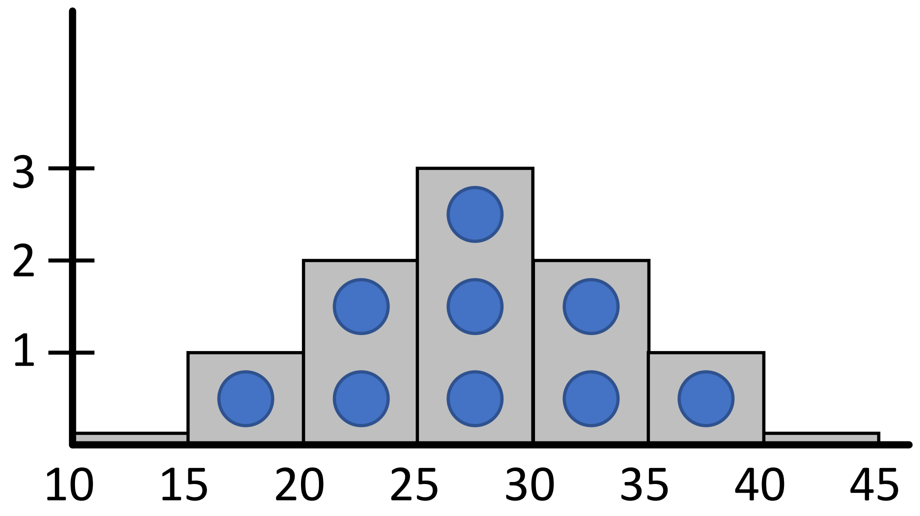
- modifiable areal unit problem
  - in cartography, changing the boundaries of the regions used to analyze data can yield dramatically different results

spatial aggregation: Congressional Districts



Histogram Challenges: Selecting Resolution

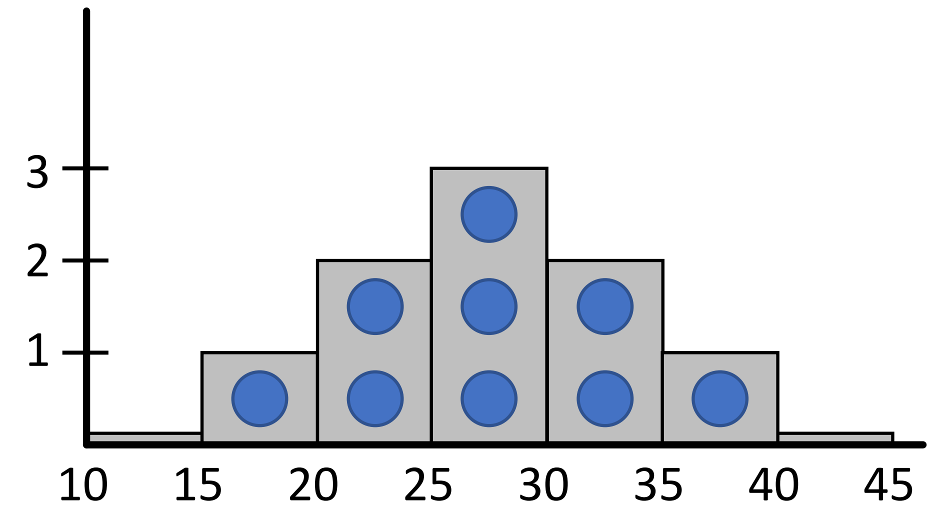
- 16
- 27
- 29
- 31
- 26
- 22
- 32
- 36
- 24

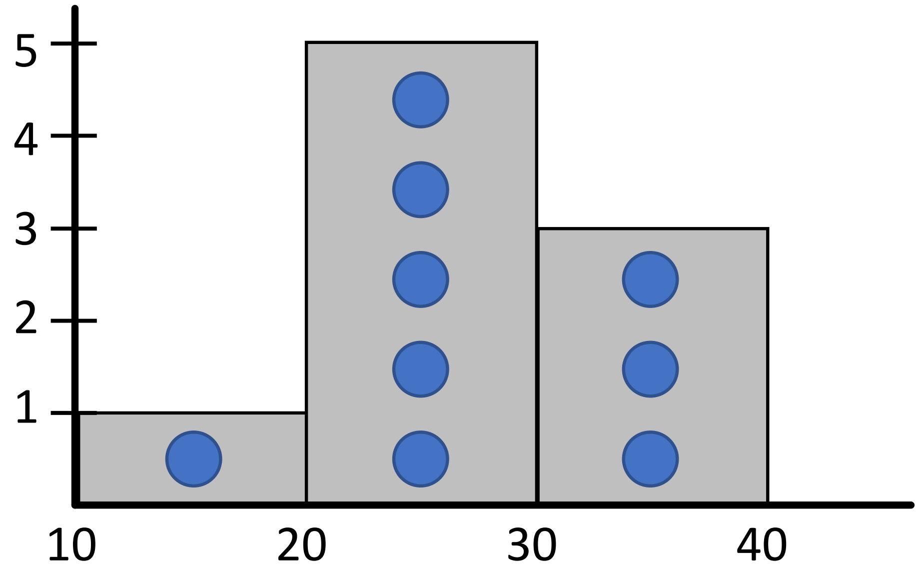


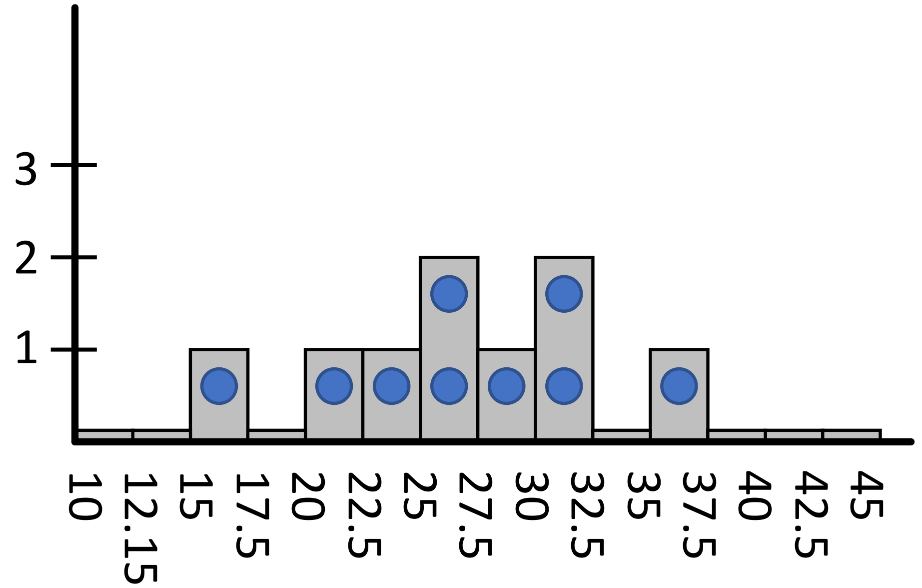
Mean (Average) = 27
Standard Deviation = 6

Histogram Challenges: Selecting Resolution

- 16
- 27
- 29
- 31
- 26
- 22
- 32
- 36
- 24

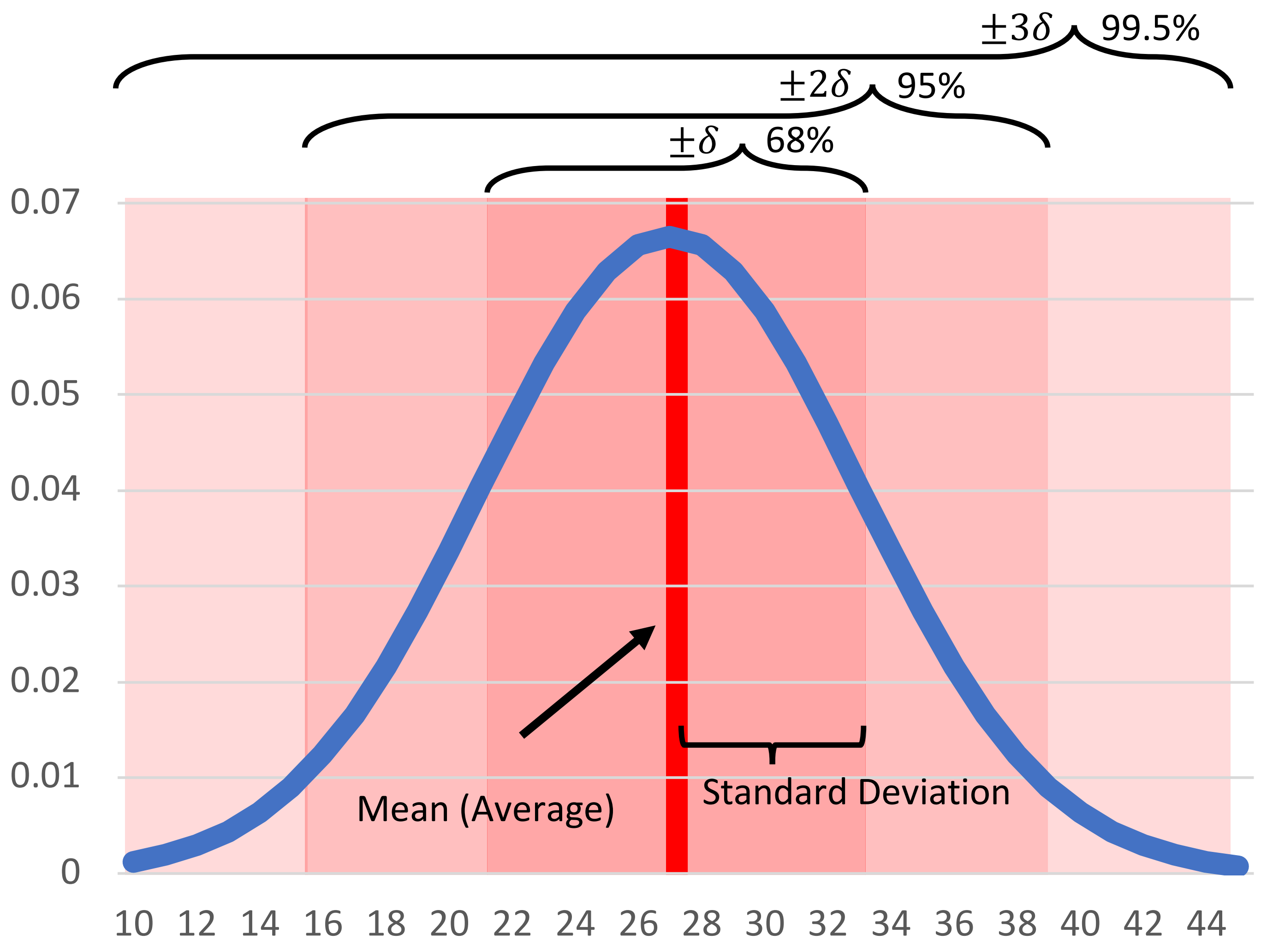


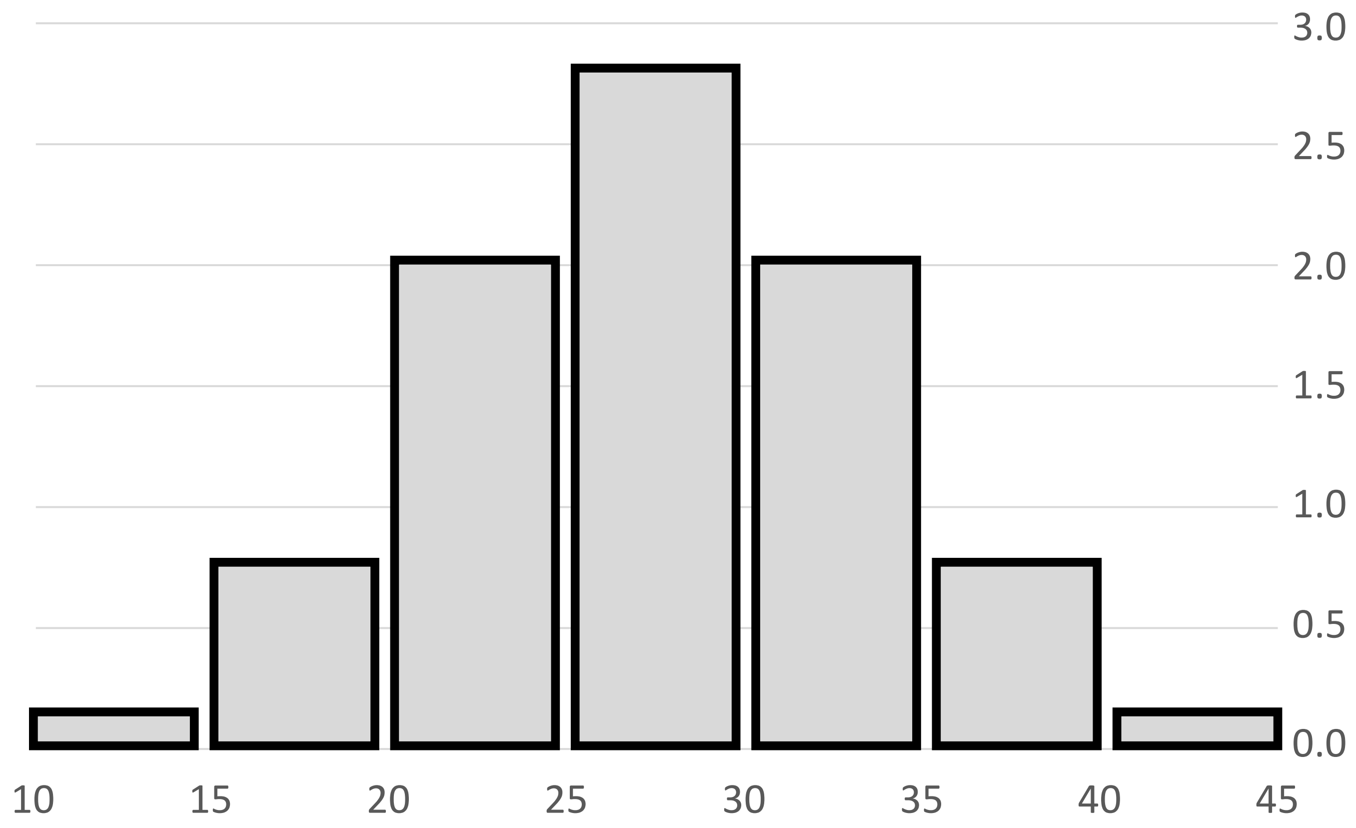




Mean (Average) = 27
Standard Deviation = 6

Statistical Modeling





Summary Statistics – mean

Definition: 3.1 *Mean*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. Their mean is

$$\text{mean}(\{x\}) = \frac{1}{N} \sum_{i=1}^{i=N} x_i.$$

- The average
- The best estimate of the value of a new data point in the absence of any other information about it

Summary statistics - Standard deviation

Definition: 3.2 *Standard deviation*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. The standard deviation of this dataset is is:

$$\mathsf{std}(x_i) = \sqrt{\frac{1}{N}\sum_{i-1}^{i=N}(x_i - \mathsf{mean}(\{x\}))^2} = \sqrt{\mathsf{mean}(\{(x_i - \mathsf{mean}(\{x\}))^2\})}.$$

- Think of this as a scale
- Average distance from mean

Standard Score (aka z score)

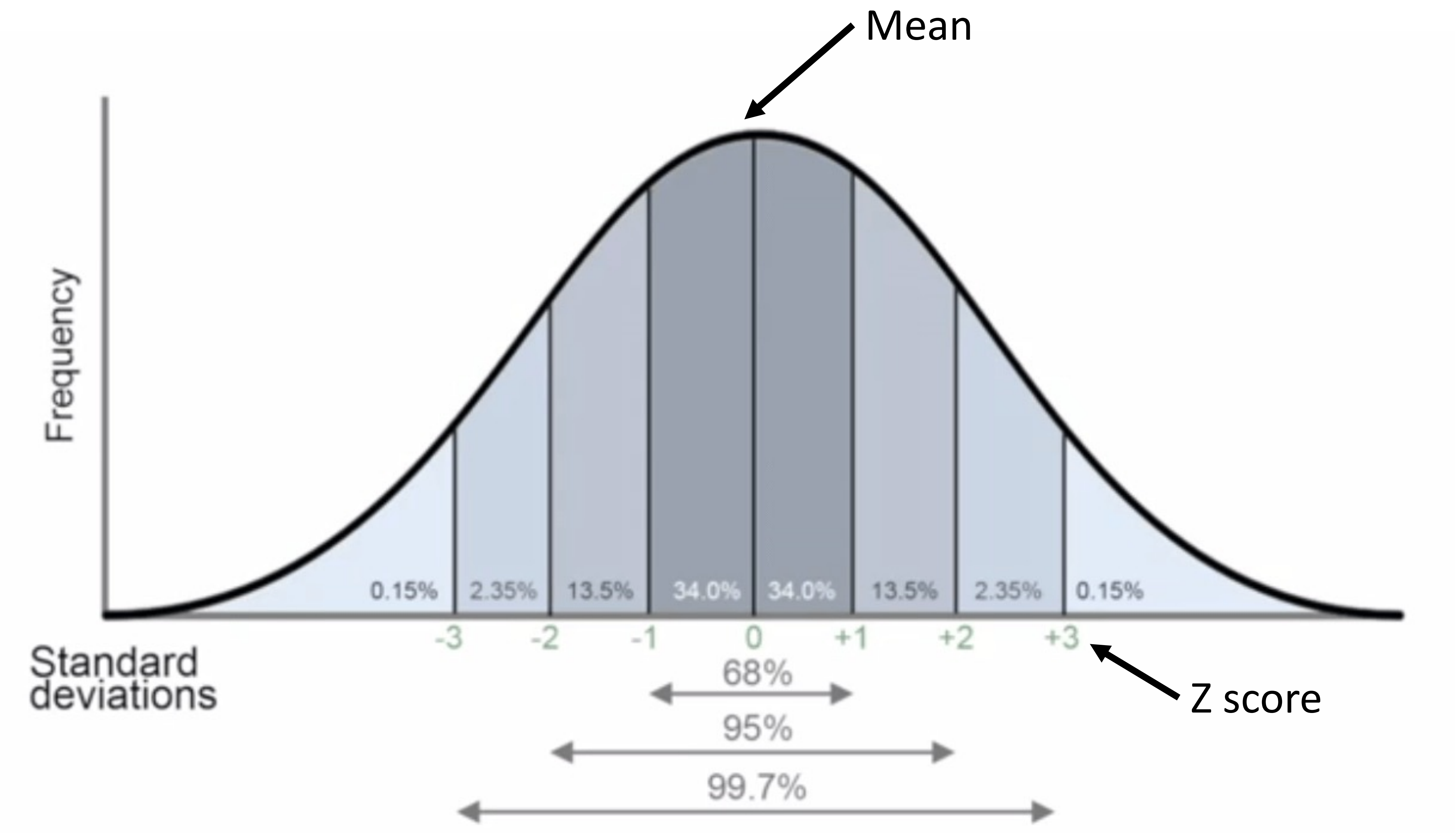
Definition: 3.8 *Standard coordinates*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

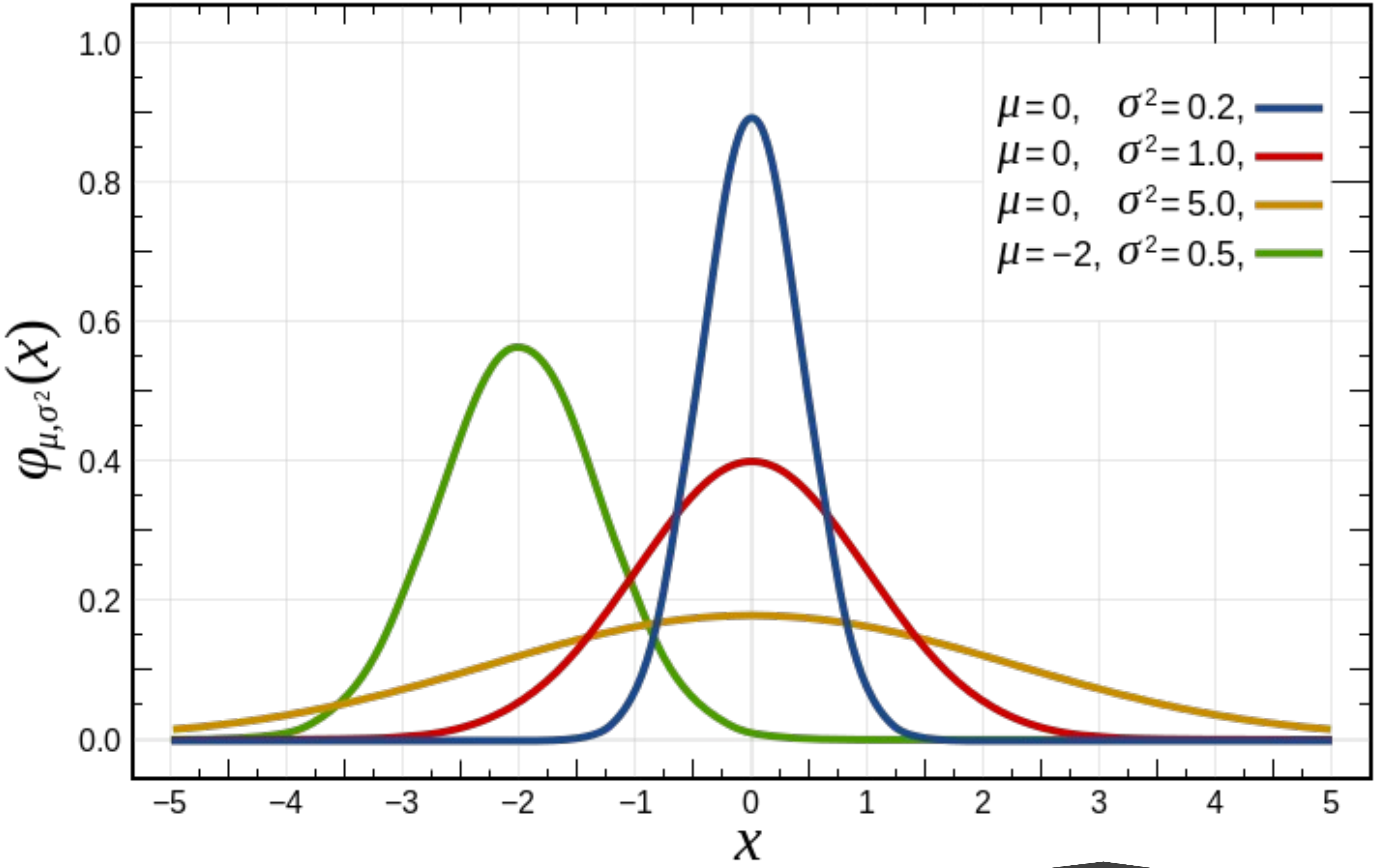
$$\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(x)}.$$

We write $\{\hat{x}\}$ for a dataset that happens to be in standard coordinates.

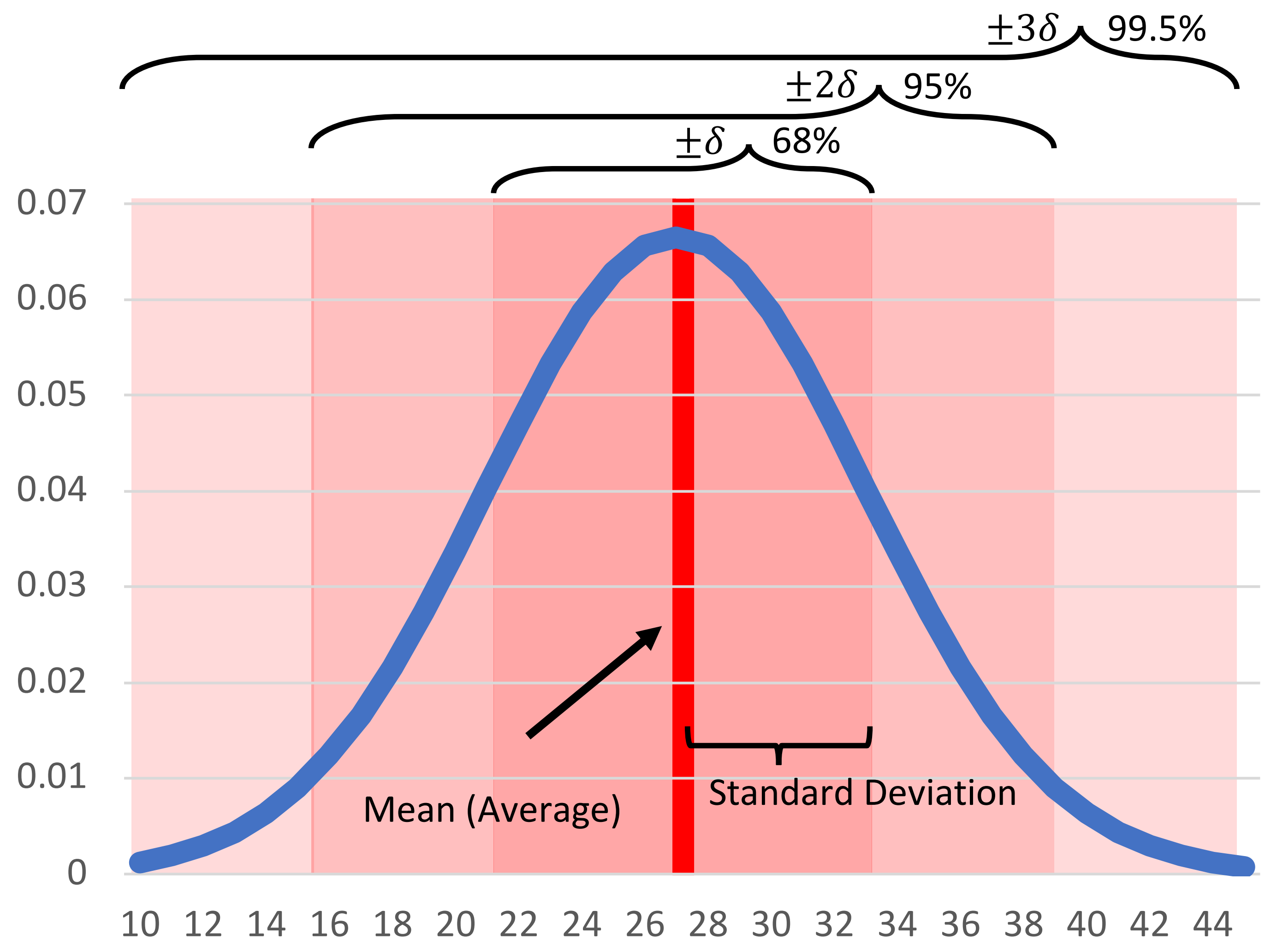
- Number of standard deviations a point is away from mean

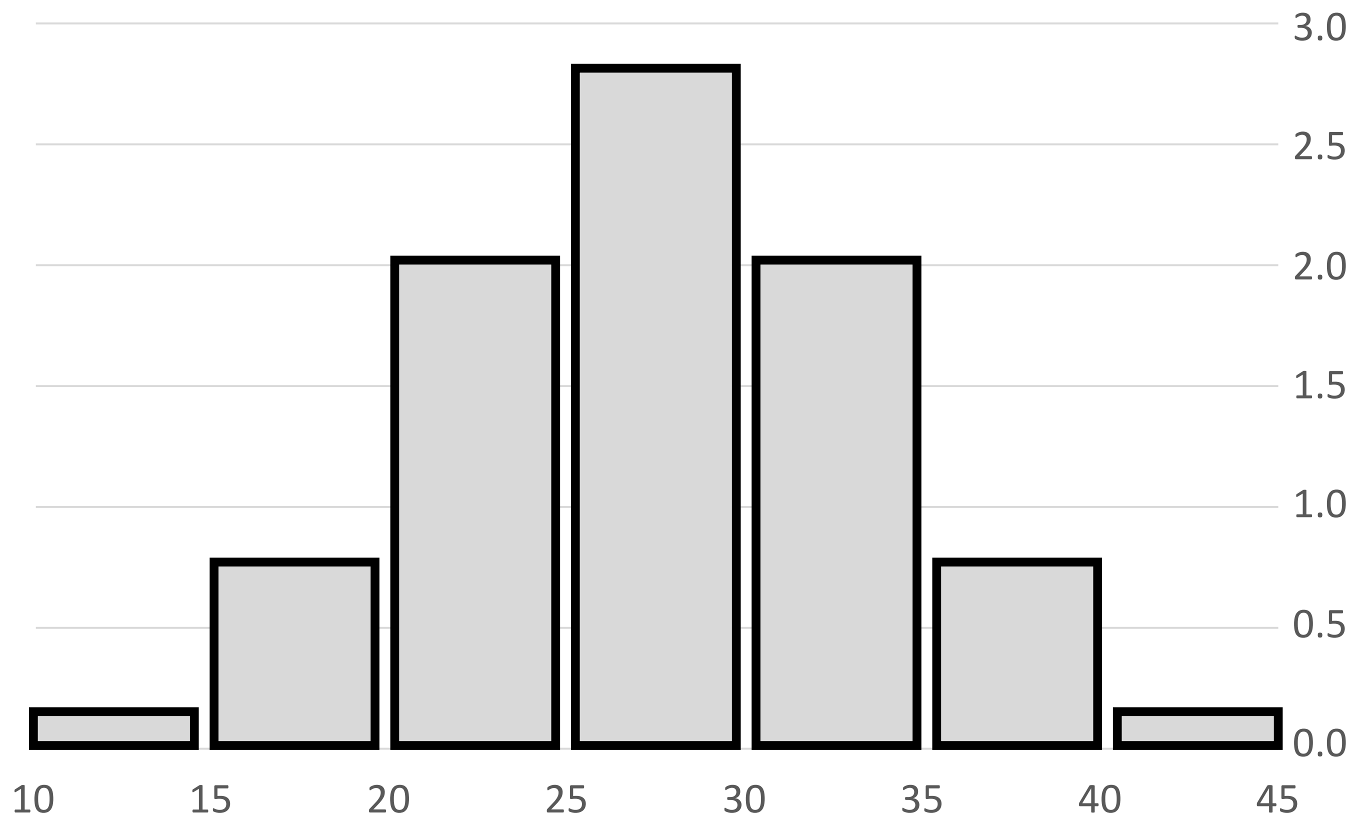

Mean
Frequency
0.15%
2.35%
13.5%
34.0%
34.0%
13.5%
2.35%
0.15%
-3
-2
-1
0
+1
+2
+3
Standard deviations
68%
95%
99.7%
Z score

Normal Distribution

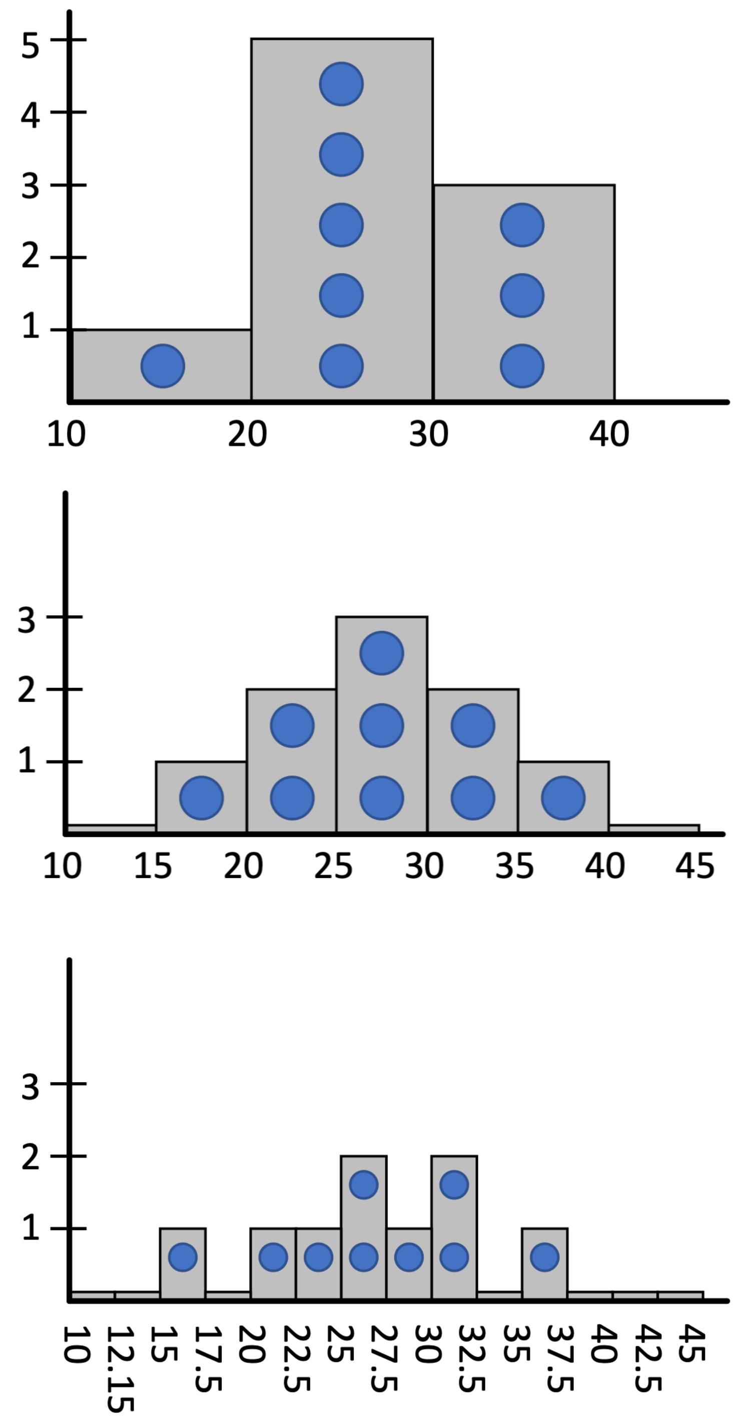


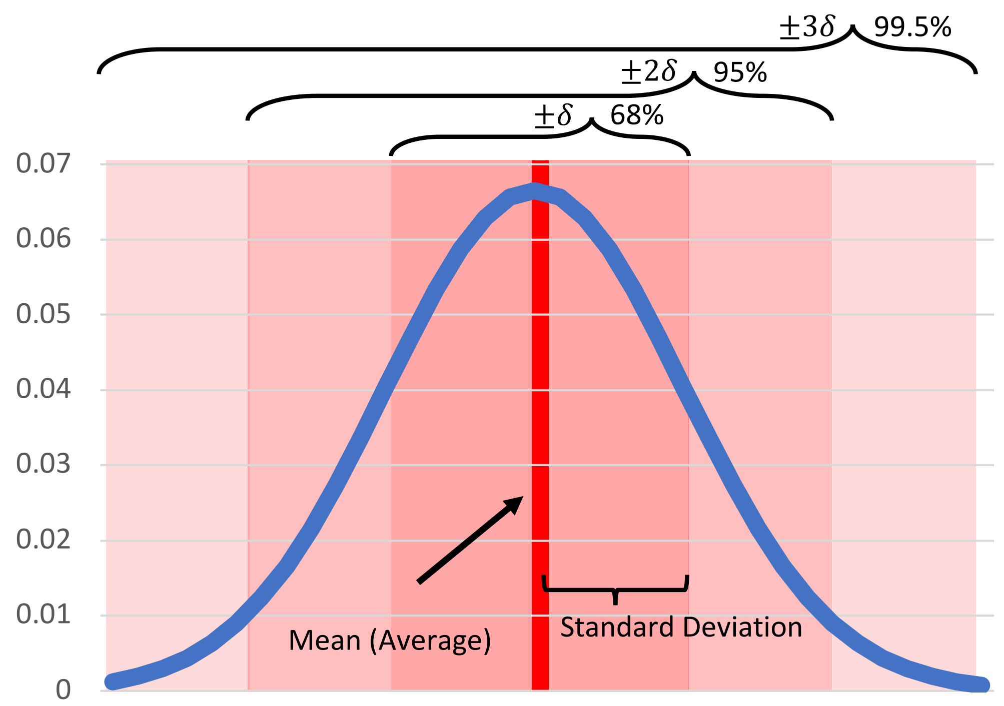
An Example: Statistical Distribution

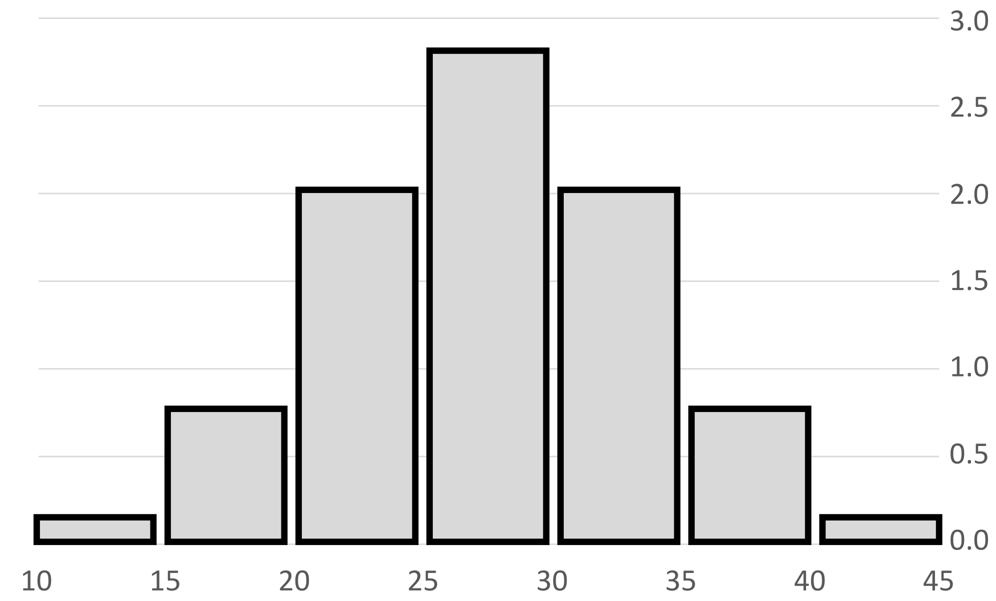




An Example: Comparing Histogram & Distribution

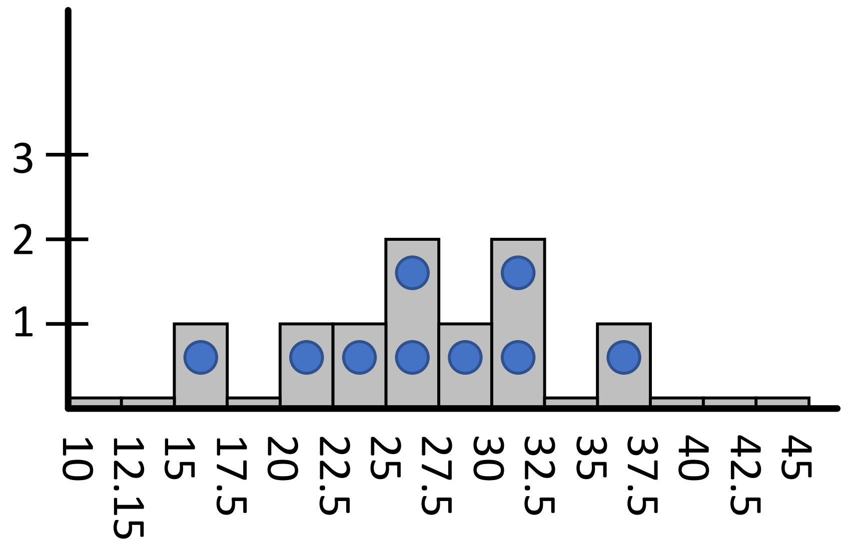


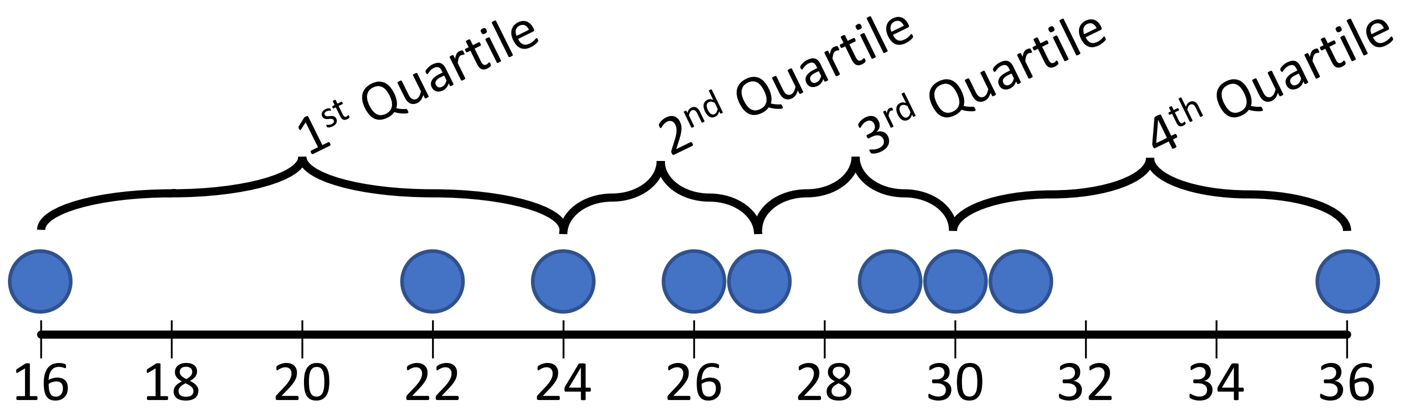




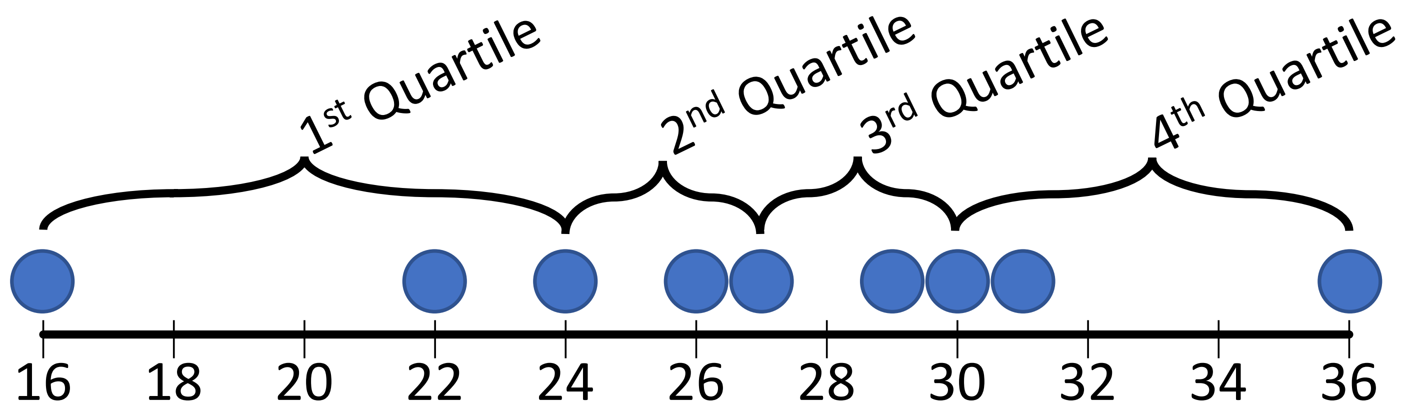
An Example: Comparing Histogram & Distribution

- 16
- 27
- 29
- 31
- 26
- 22
- 32
- 36
- 24

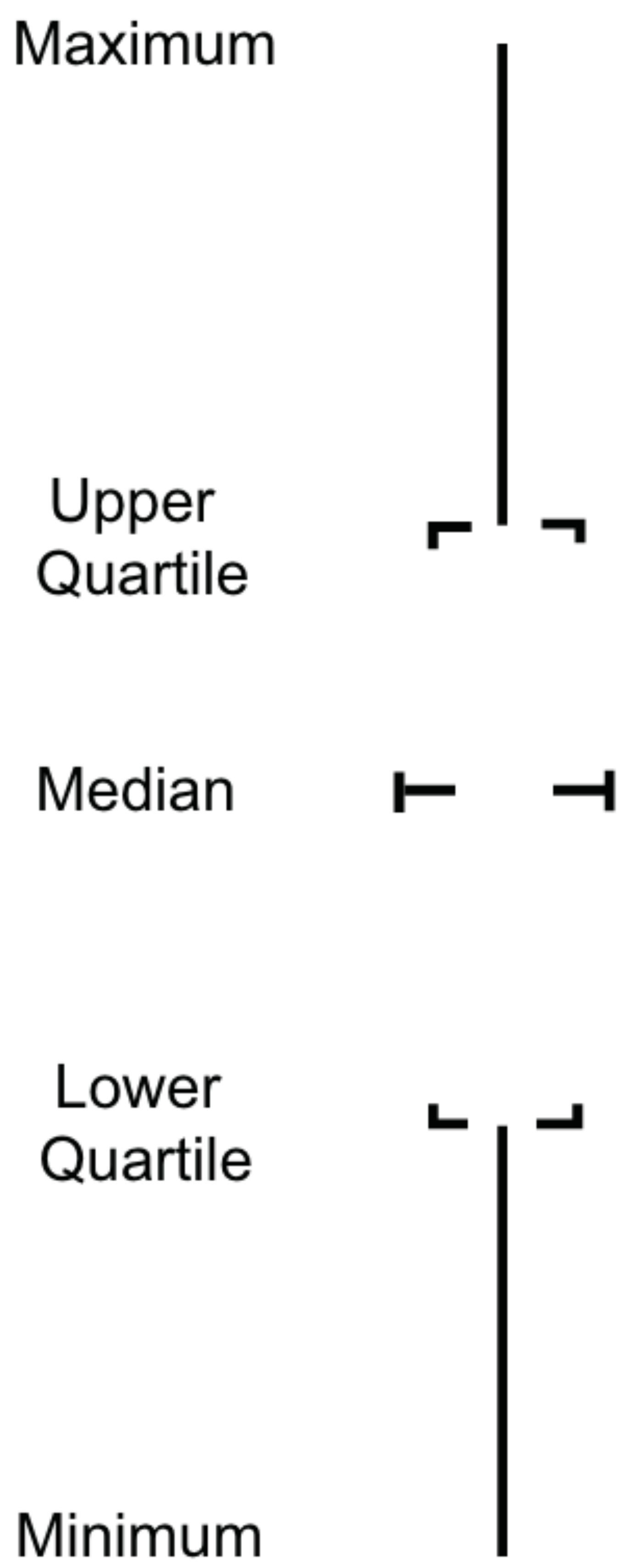




Boxplot

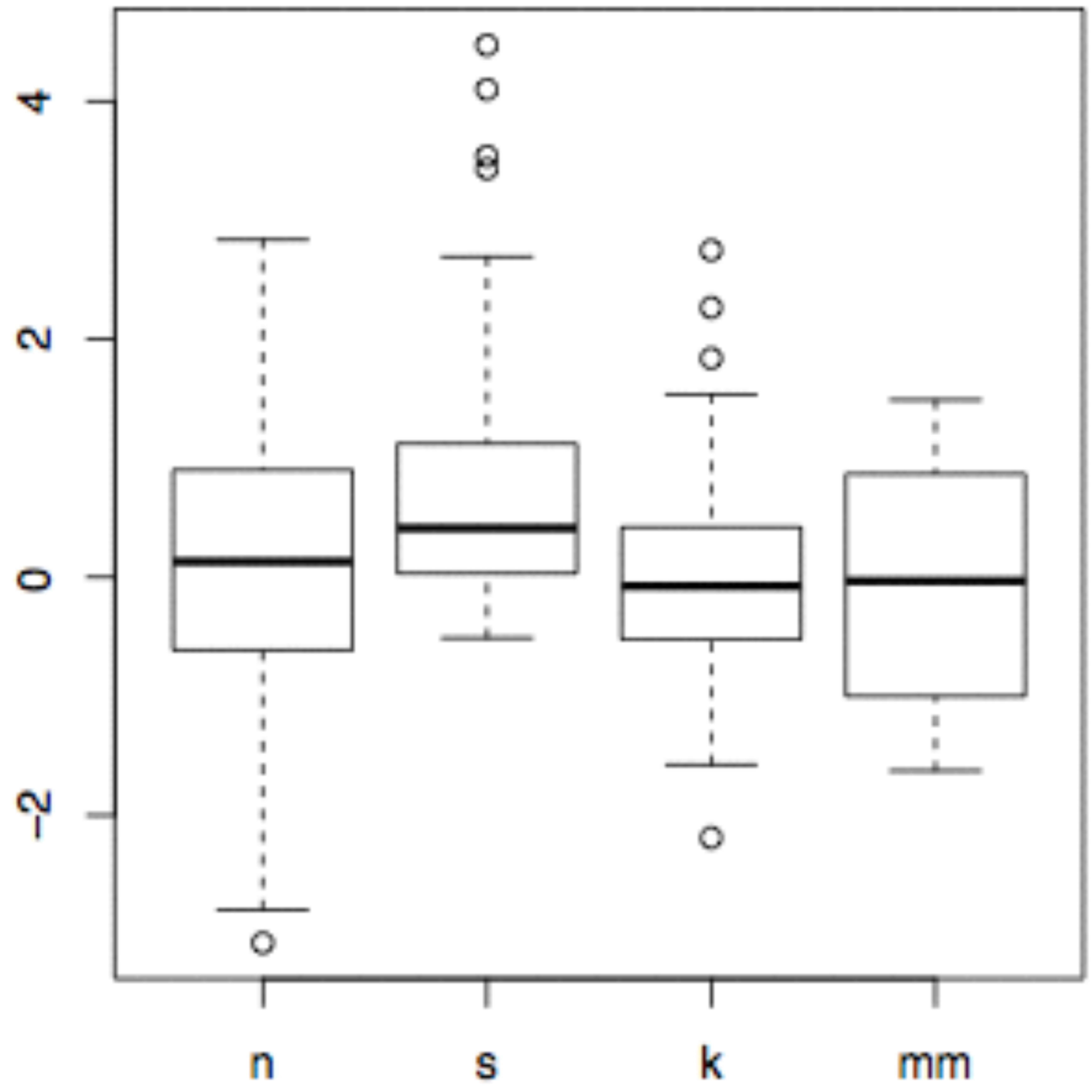


1st Quartile
2nd Quartile
3rd Quartile
4th Quartile
16
18
20
22
24
26
28
30
32
34
36

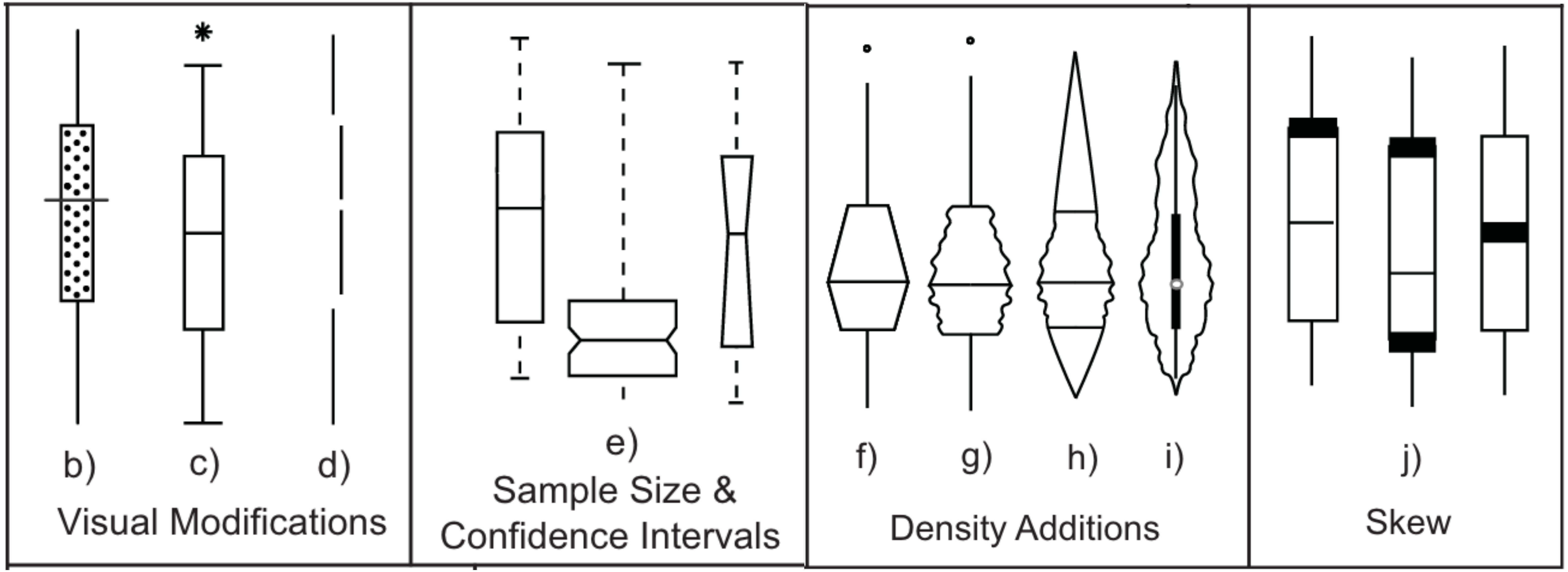


Maximum
Upper Quartile
Median
Lower Quartile
Minimum

Boxplot

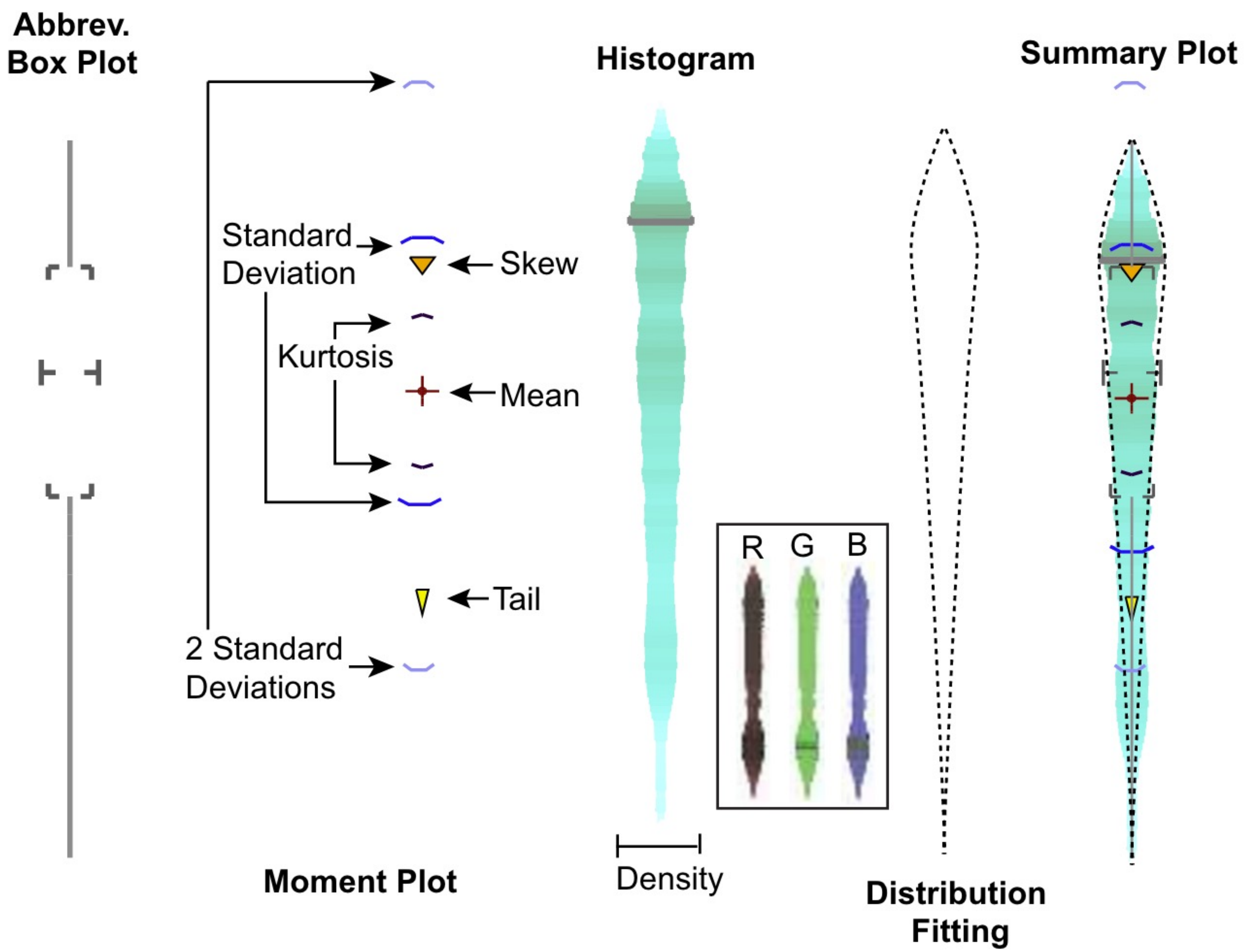


Boxplots



Potter et al., 2010

Boxplots



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

- kth 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.

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

attribute aggregation

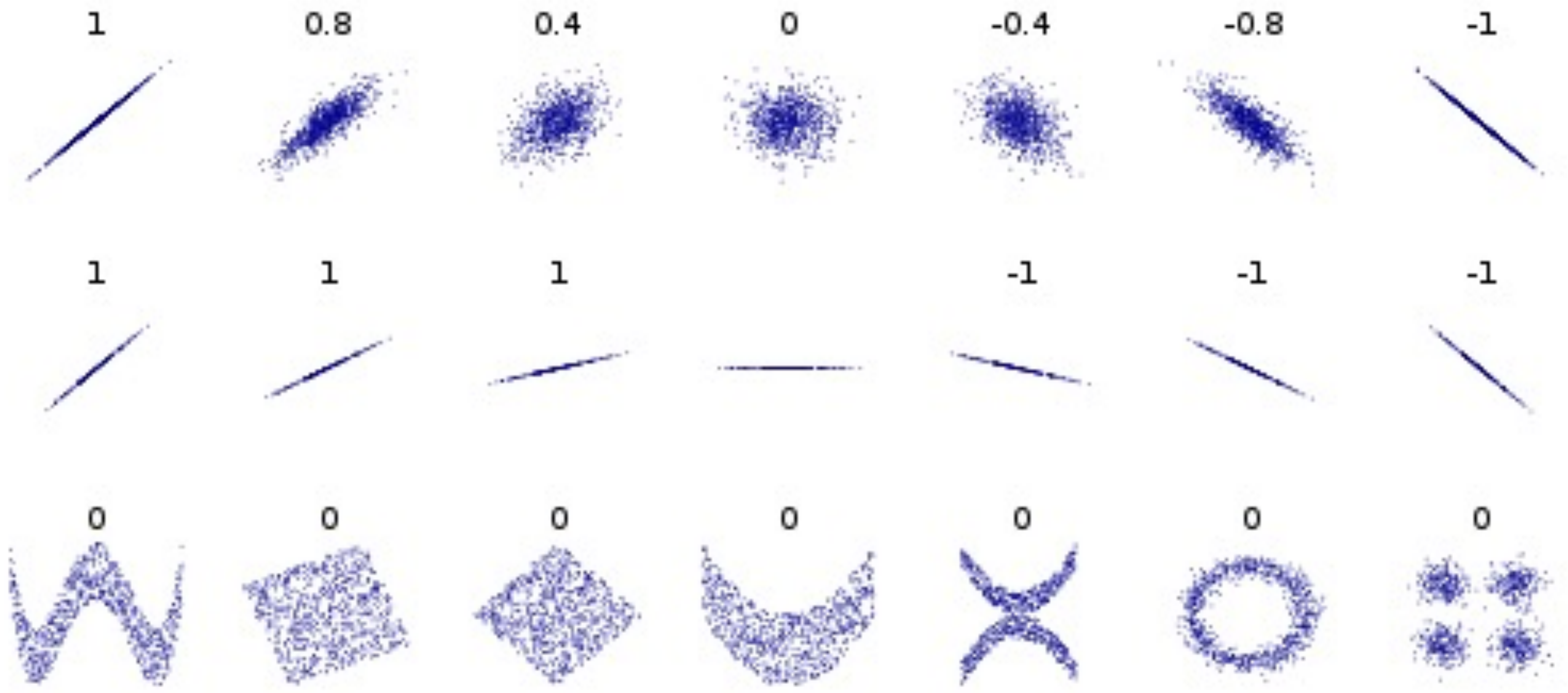
- group attributes and compute a similarity score across the set
- dimensionality reduction to preserve meaningful structure

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)

Pearson Correlation Coefficient

- A measure of the linearity between 2 sets


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

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

where:

- n, x_i, y_i are defined as above
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

- Given: $X=\{x_0,\ldots,x_n\}$, $Y=\{y_0,\ldots,y_n\}$
- Calculate mean(X), mean(Y), stdev(X), stdev(Y)
  - mean(X) = $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
  - stdev(X) = $$\sigma_X \quad \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$$

$$\boxed{r = \frac{1}{n}\,\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_X \sigma_Y}}$$

- X={1,2.5,3,4.5}
- Y={2,2.5,3.5,4}

- mean(X) = 2.75, mean(Y) = 3

- stdev(X)= sqrt($(1-2.75)^2 + (2.5-2.75)^2 + (3-2.75)^2 + (4.5-2.75)^2$ / 4) = 1.25
- stdev(Y)= sqrt($(2-3)^2 + (2.5-3)^2 + (3.5-3)^2 + (4-3)^2$ / 4) = 0.79

- X={1,2.5,3,4.5}
- Y={2,2.5,3.5,4}

- mean(X) = 2.75, mean(Y) = 3
- stdev(X)= 1.25, stdev(Y)= 0.79

- Covariance(X,Y)

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

= 1/4 * (1-2.75)(2-3) + (2.5-2.75)(2.5-3) +
(3-2.75)(3.5-3) + (4.5-2.75)(4-3)

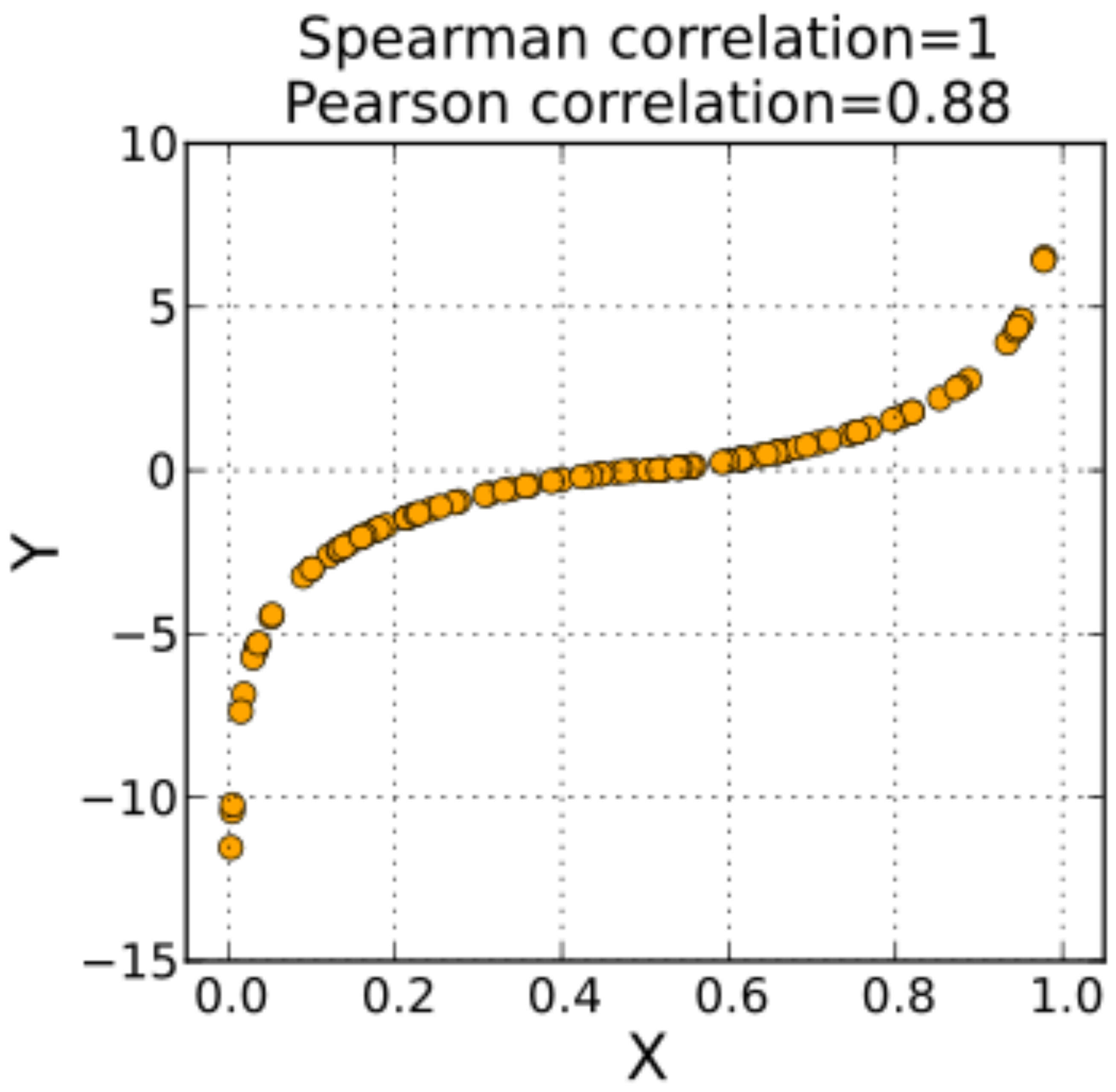
= 3.75 / 4 = 0.94

- $X=\{1,2.5,3,4.5\}$
- $Y=\{2,2.5,3.5,4\}$

- mean(X) = 2.75, mean(Y) = 3
- stdev(X)= 1.25, stdev(Y)= 0.79
- Cov(X,Y)= 0.94

$$\boxed{r = 0.94 / (1.25 * 0.79) = 0.95}$$

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

- X={1, 2.5, 3, 4.5}
- Y={2, 3.5, 2.5, 4}

- X’ = rank(X)
- Y’ = rank(Y)

- SRC = PCC(X’, Y’)

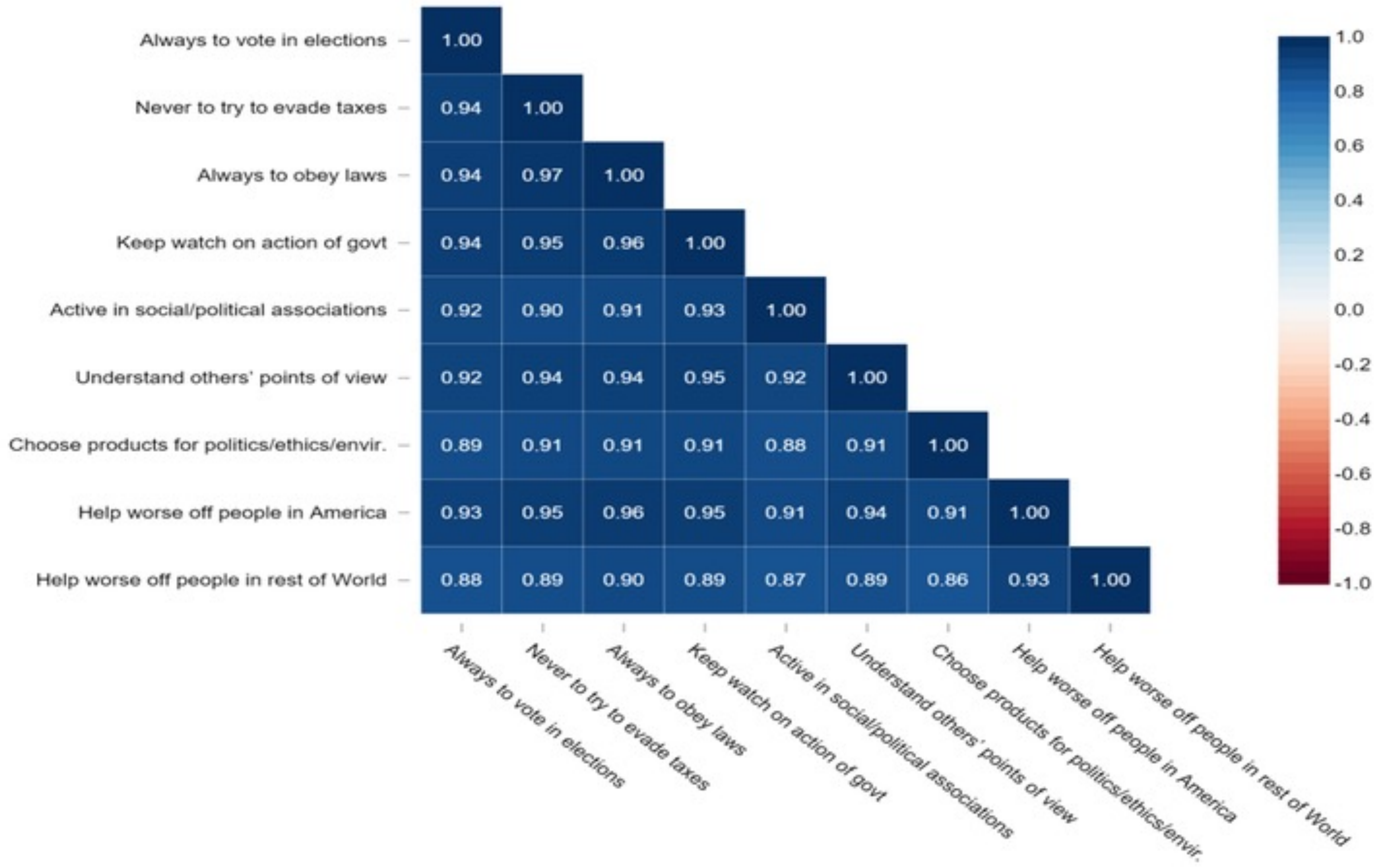
- X={1, 2.5, 3, 4.5}
- X Sorted {1, 2.5, 3, 4.5}

- X’ = rank(X)
- X’ = { rank(1), rank(2.5), rank(3), rank(4.5) }
- X’ = { 1, 2, 3, 4 }

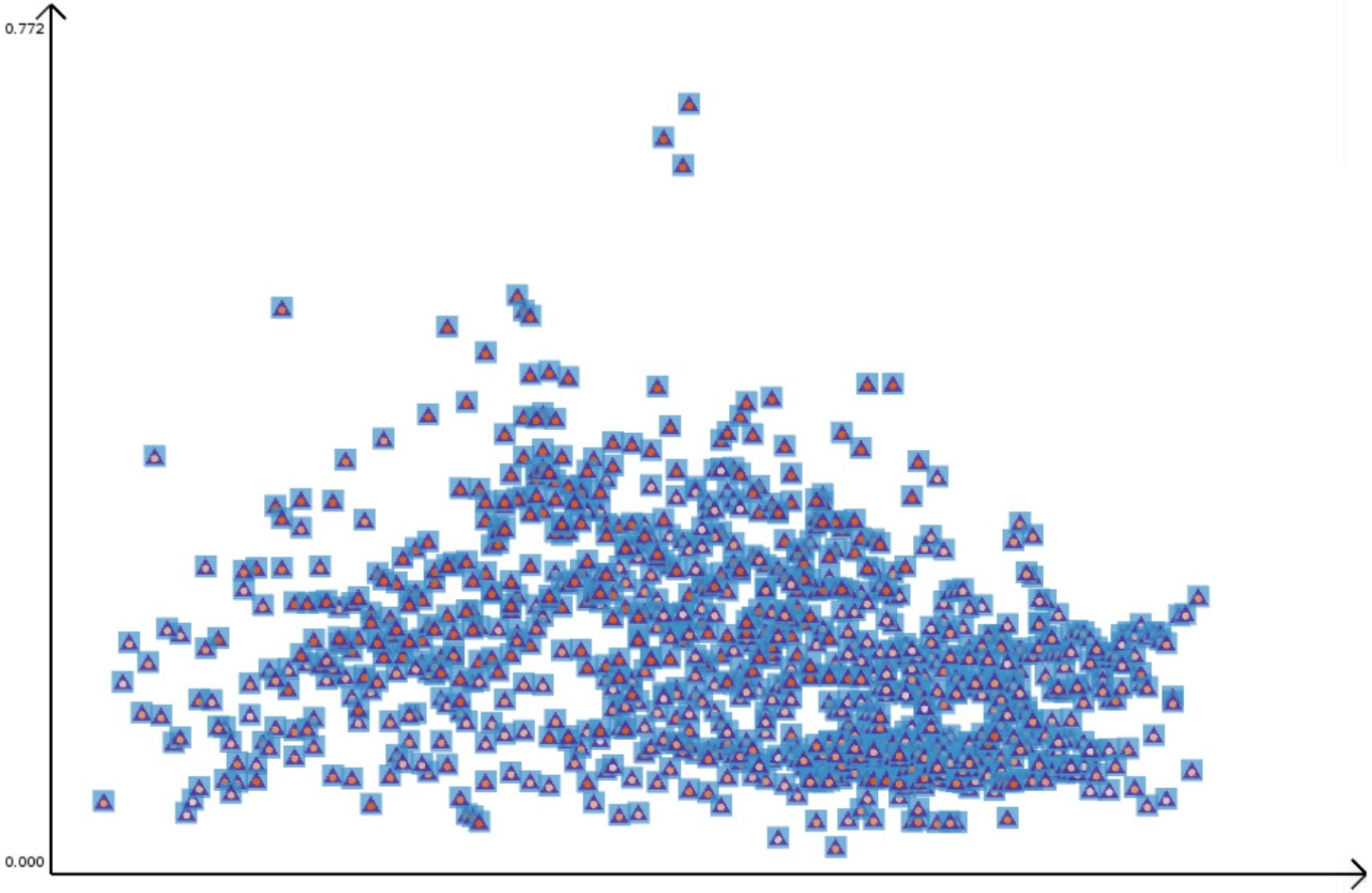
- Y={2, 3.5, 2.5, 4}
- Y Sorted {2, 2.5, 3.5, 4}

- Y’ = rank(Y)
- Y’ = { rank(2), rank(3.5), rank(2.5), rank(4) }
- Y’ = { 1, 3, 2, 4 }

Multiple Attributes – Correlation Matrix



Many Attributes Multiple Correlation

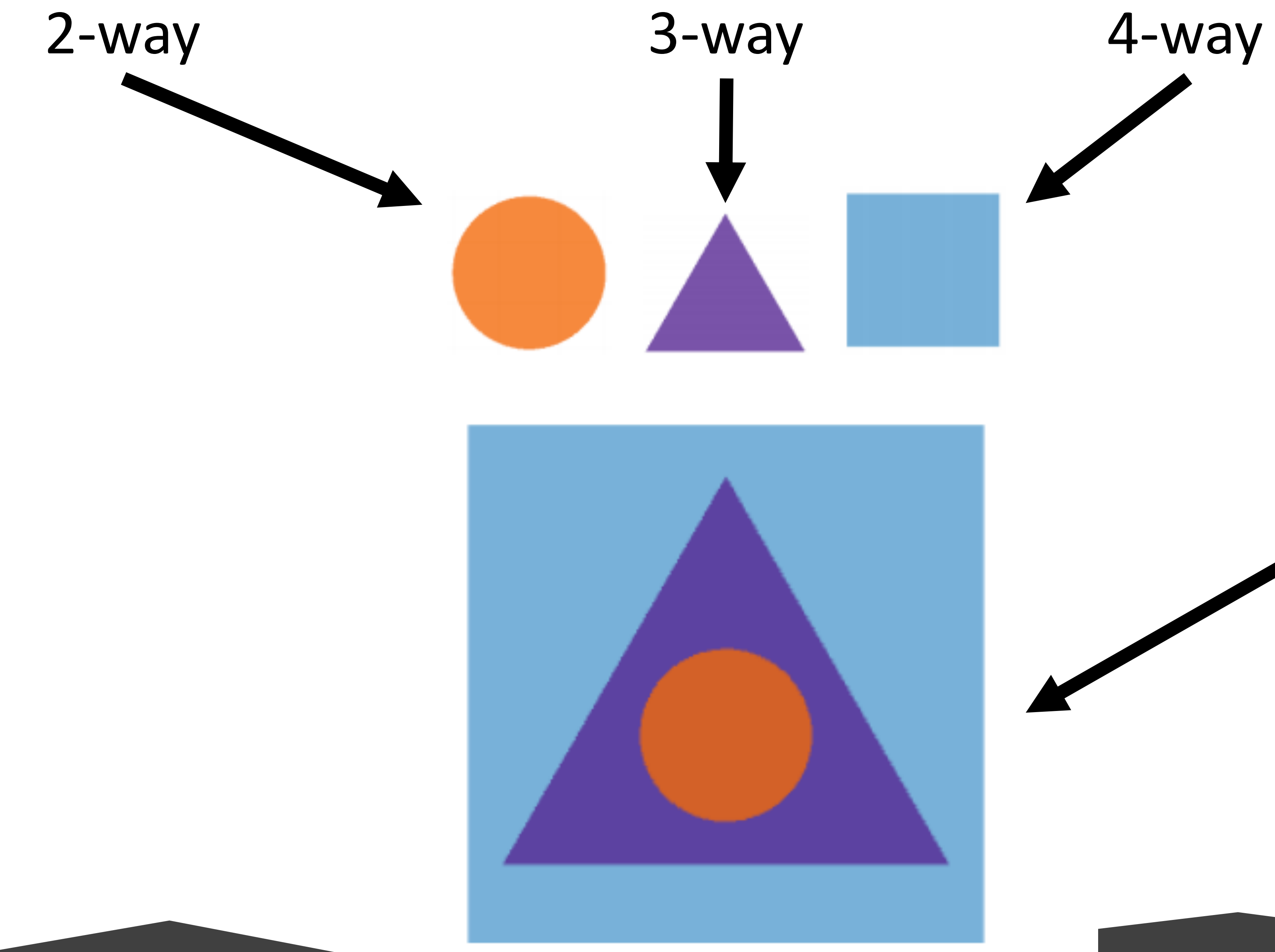


Multiple Correlation

$$R^2 = \mathbf{c}^\top R_{xx}^{-1}\,\mathbf{c},$$

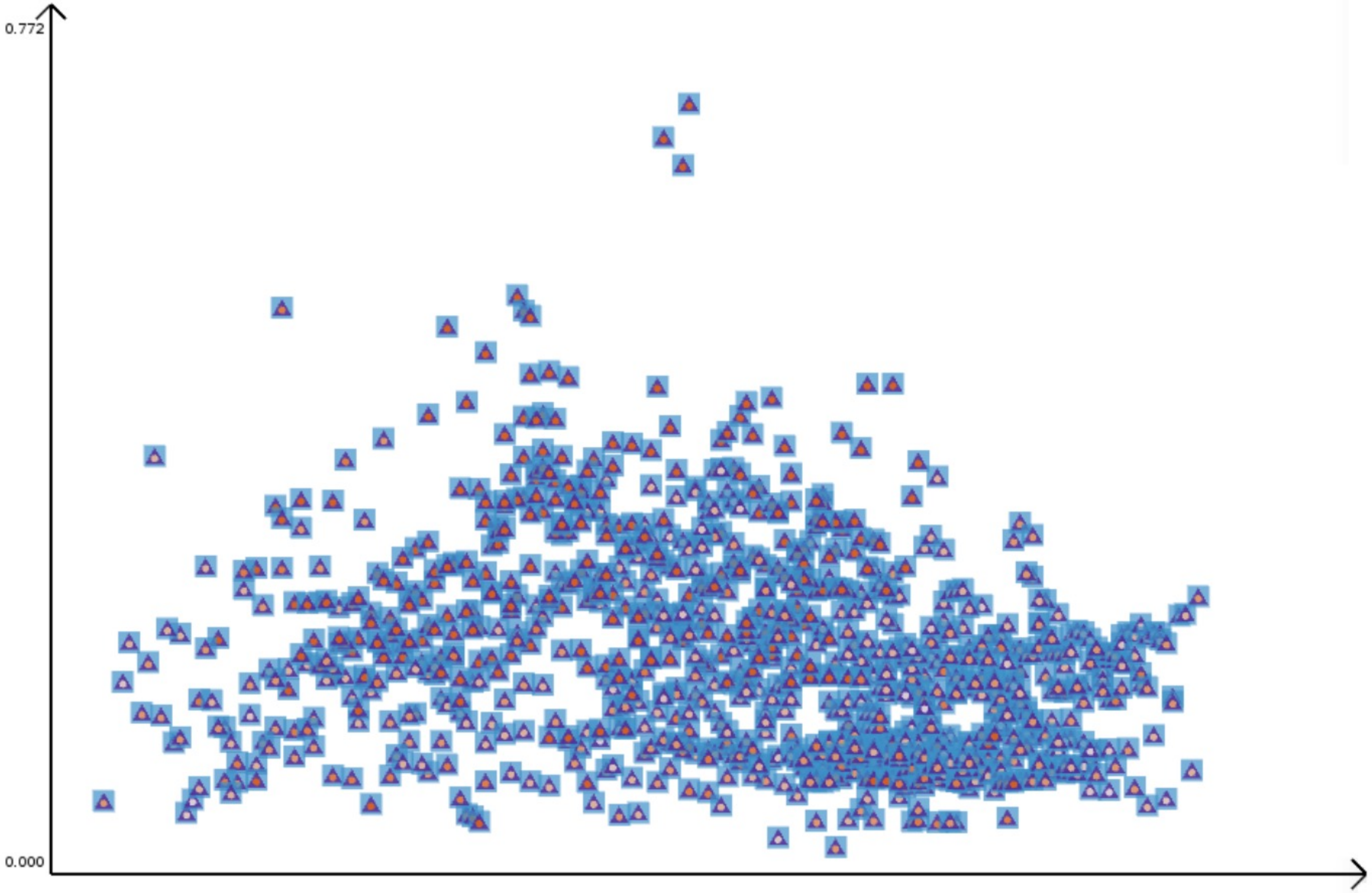
$$R_{xx} = \begin{pmatrix} r_{x_1x_1} & r_{x_1x_2} & \cdots & r_{x_1x_N} \\ r_{x_2x_1} & \ddots & & \vdots \\ \vdots & & \ddots & \\ r_{x_Nx_1} & \cdots & & r_{x_Nx_N} \end{pmatrix}.$$

Multiple Correlation



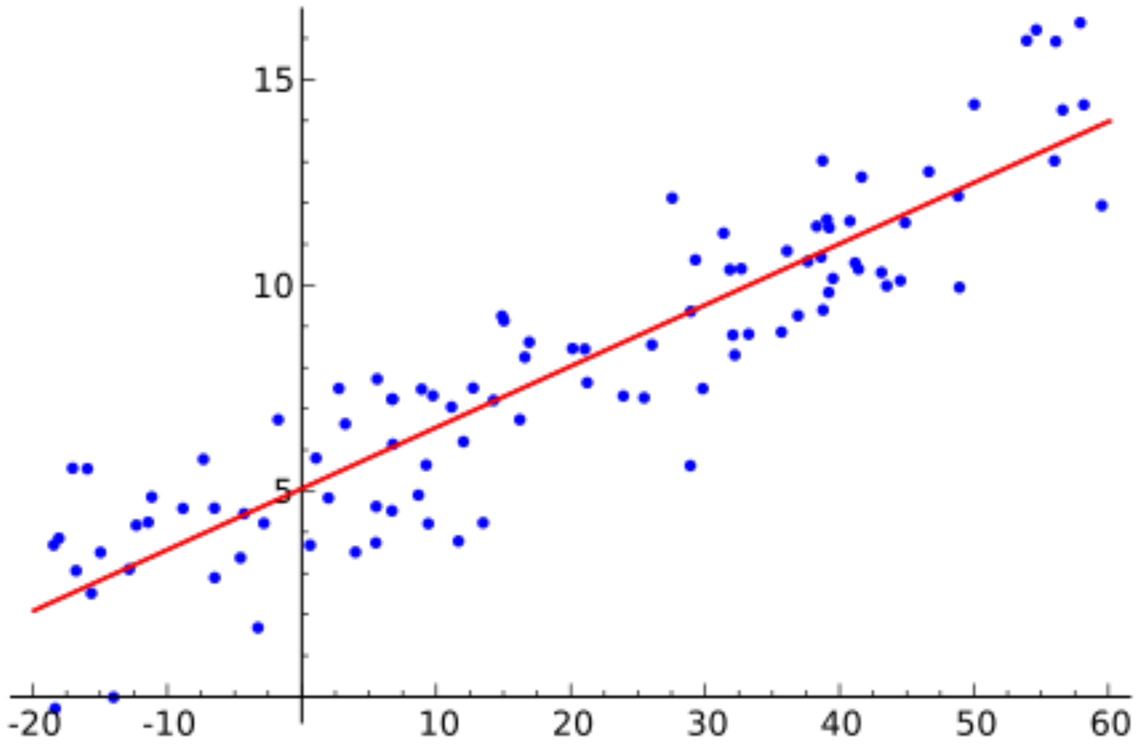
Composite Glyph
- (6) 2-way
- (12) 3-way
- (4) 4-way

Many Attributes Multiple Correlation



Regression: Fitting a Model to Data

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

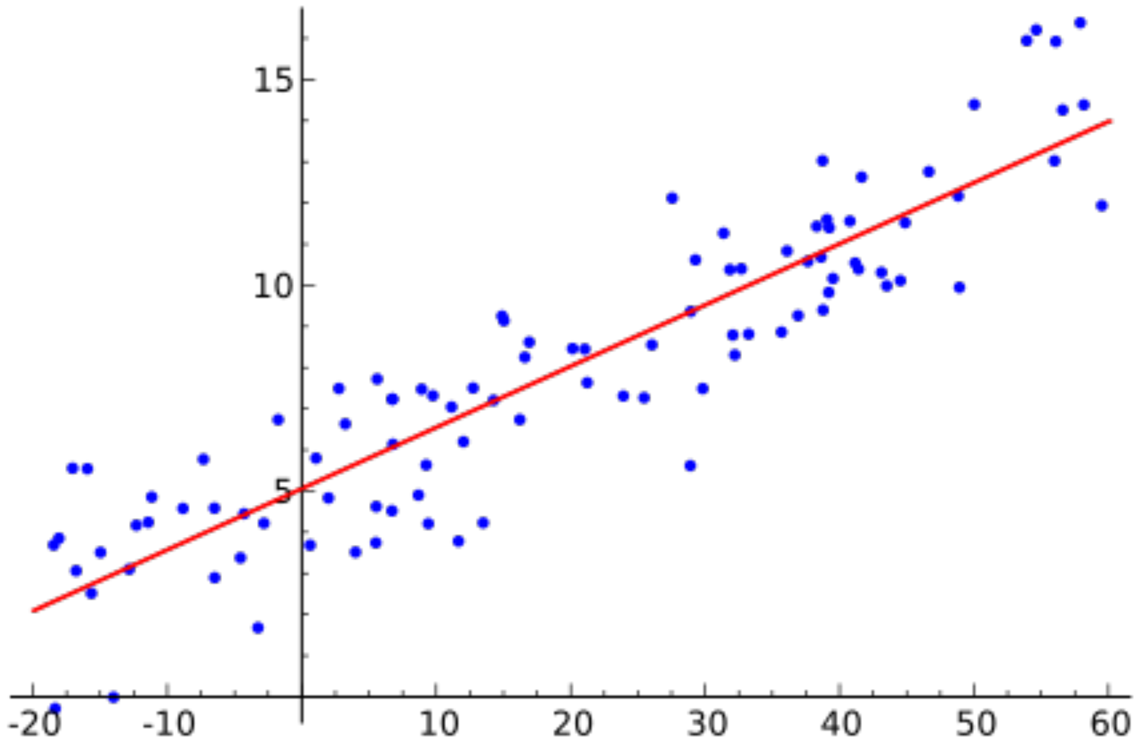


Regression: Fitting a Model to Data

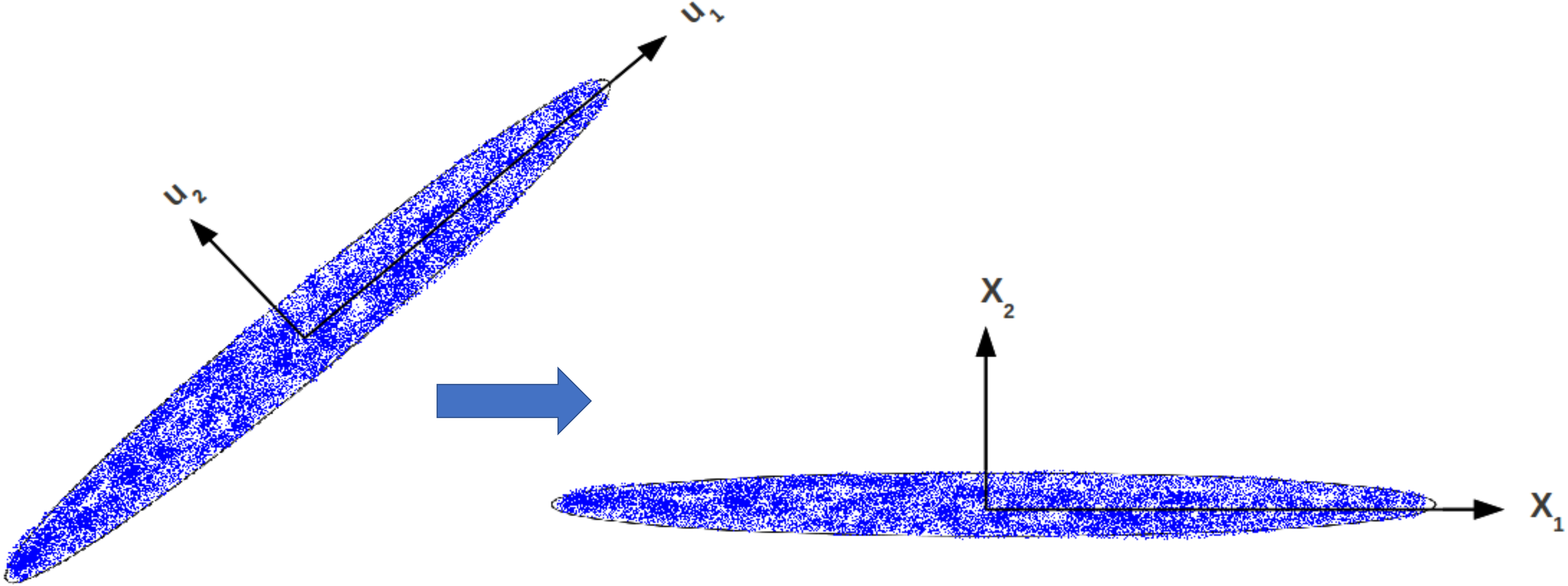
- Can be computed directly

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

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



Linear Dimensionality reduction: Principal Component Analysis (PCA)



Nonlinear Dimensionality Reduction: Multidimensional Scaling (MDS)

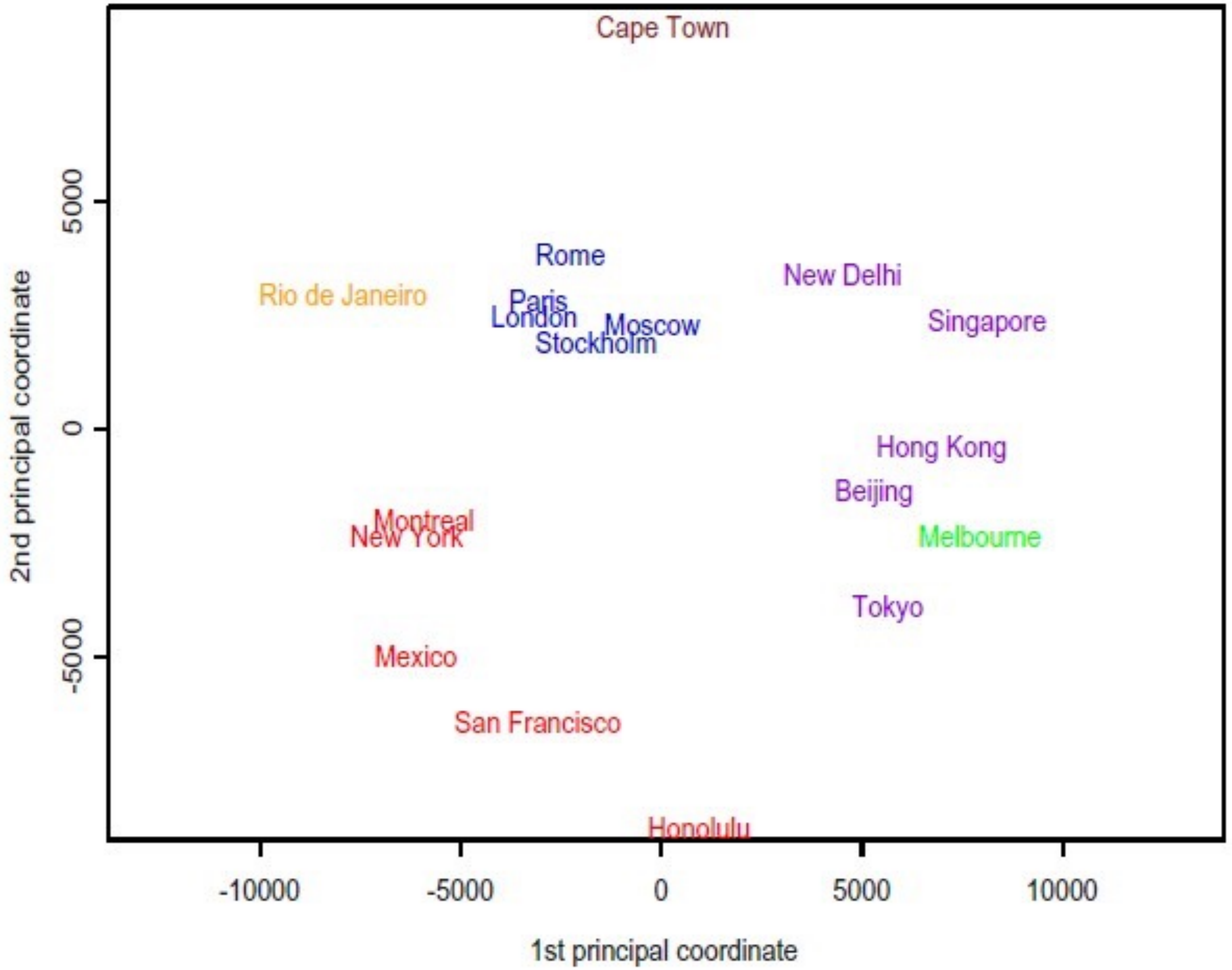


FIGURE 13.1. *Two-dimensional map of 18 world cities using the classical scaling algorithm on airline distances between those cities. The colors*

Problem #3
What is lost or misinterpreted…

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

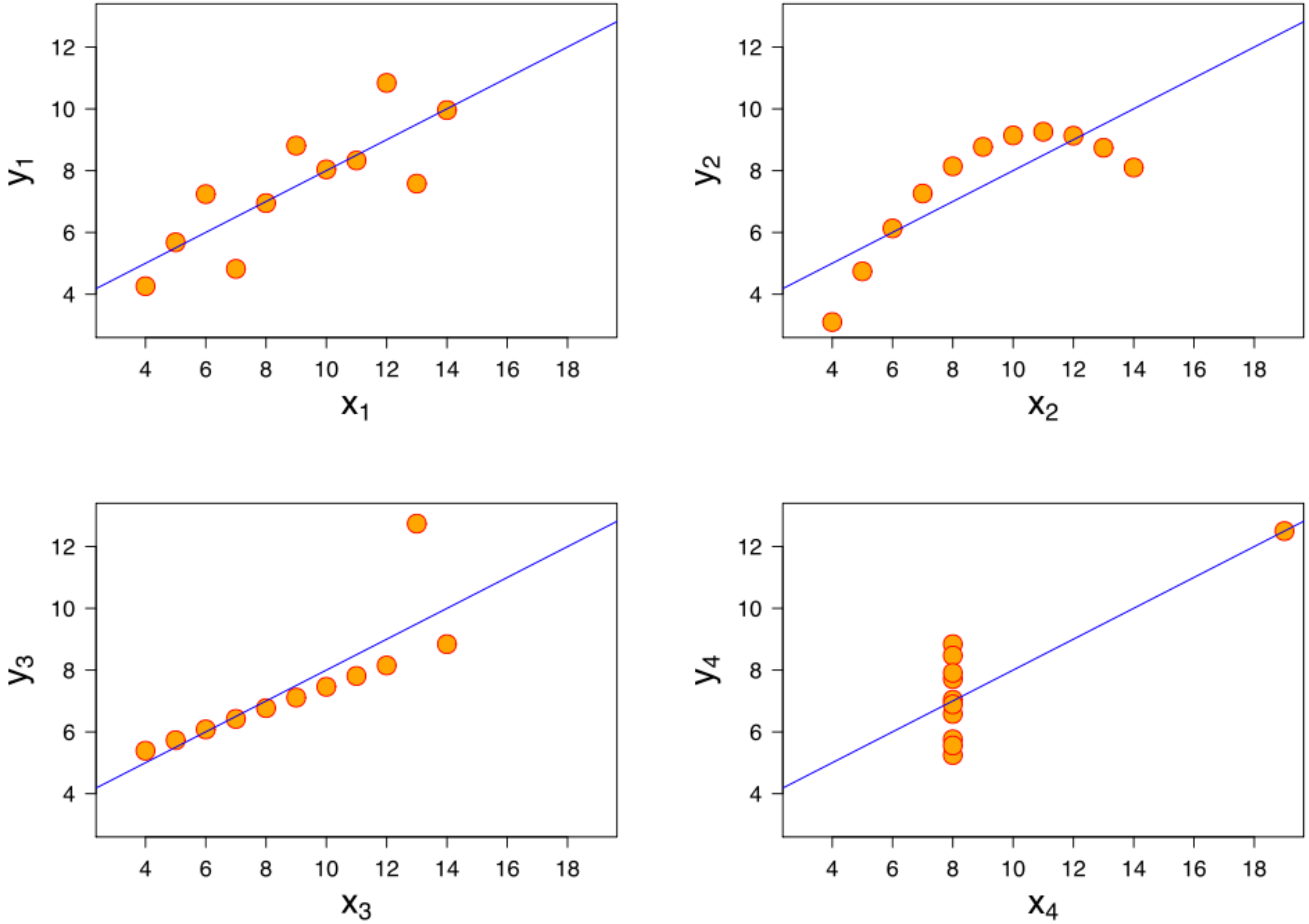
Anscombe's Quartet

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.

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

Statistical Limitations: Anscombe's quartet



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

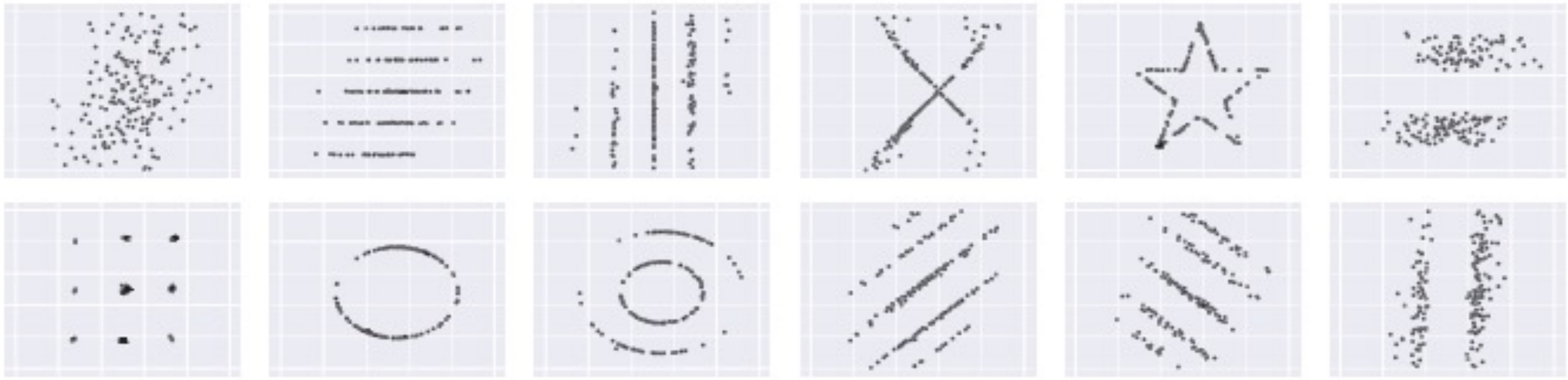
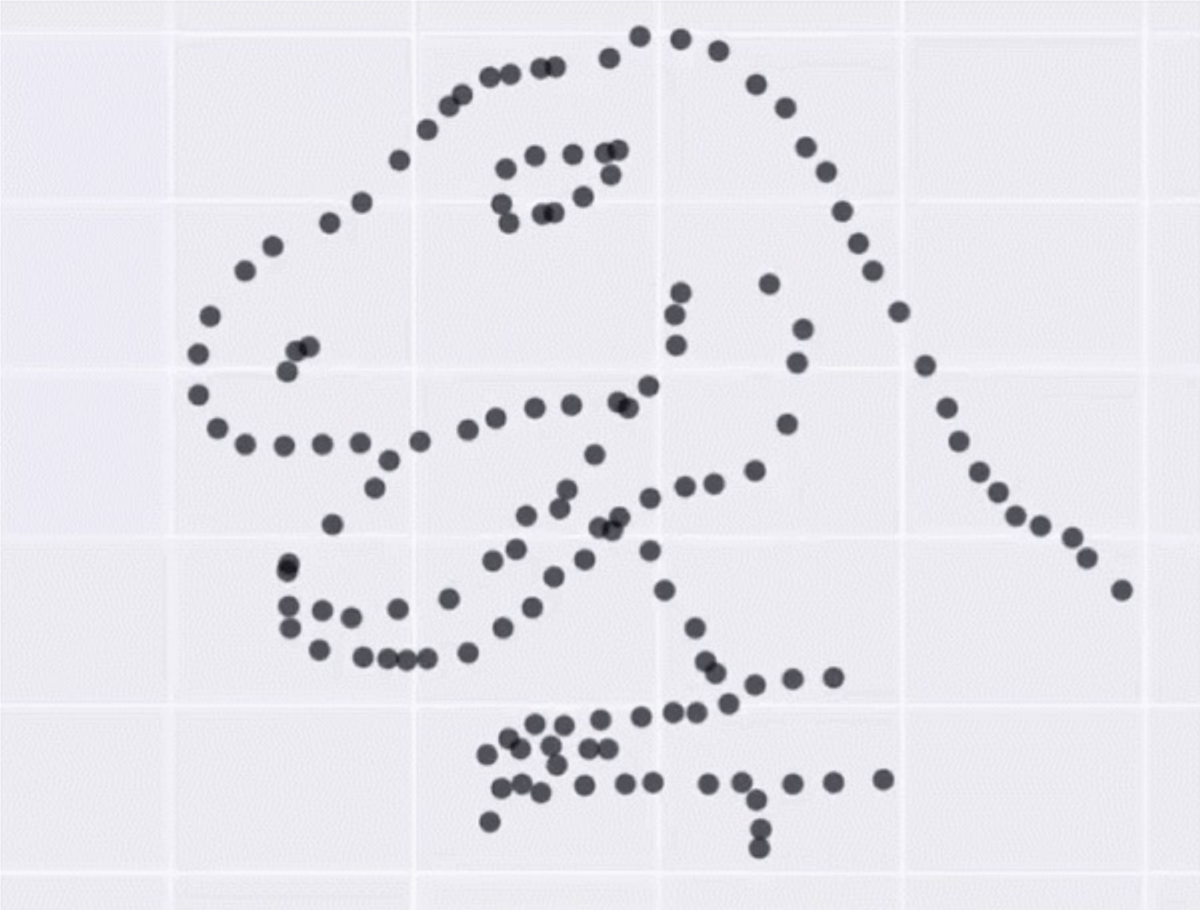


Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. (*$\bar{x}$ = 54.02, $\bar{y}$ = 48.09, sd_x = 14.52, sd_y = 24.79, Pearson's r = +0.32*)

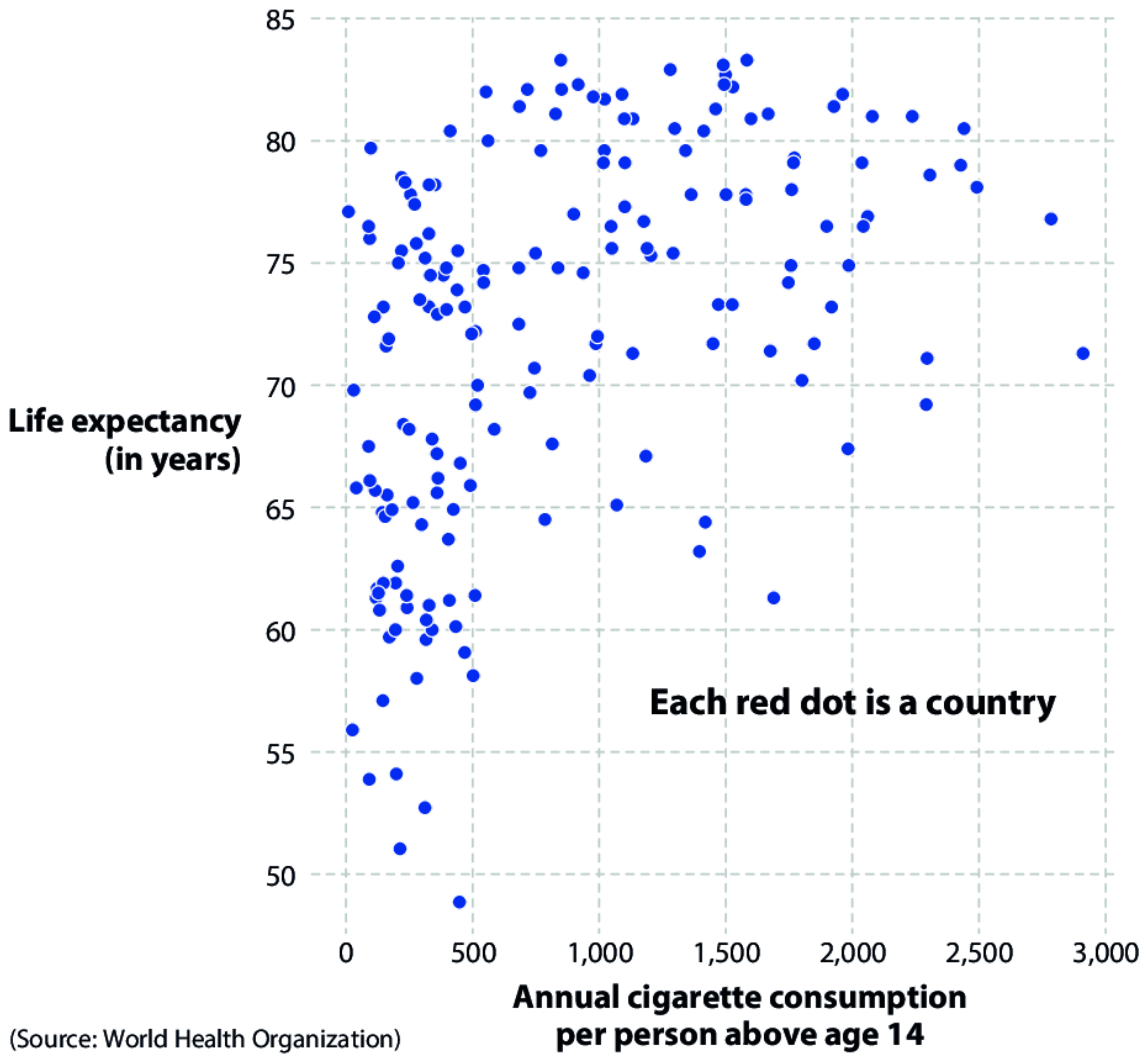
ABSTRACT
Datasets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This paper presents a novel method for generating such datasets, along with several examples. 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.



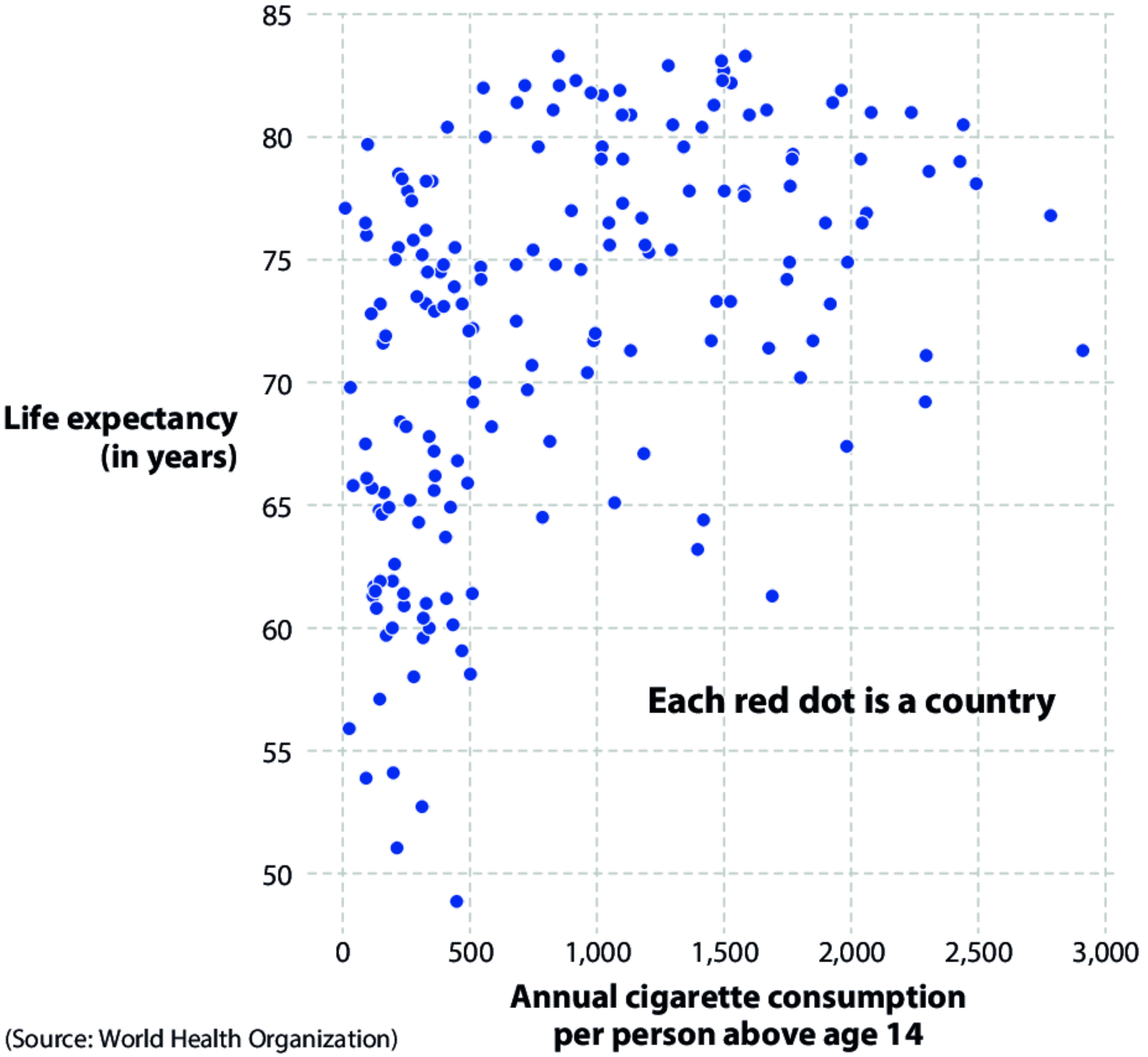
- “The more cigarettes we consume, the longer we live!”

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



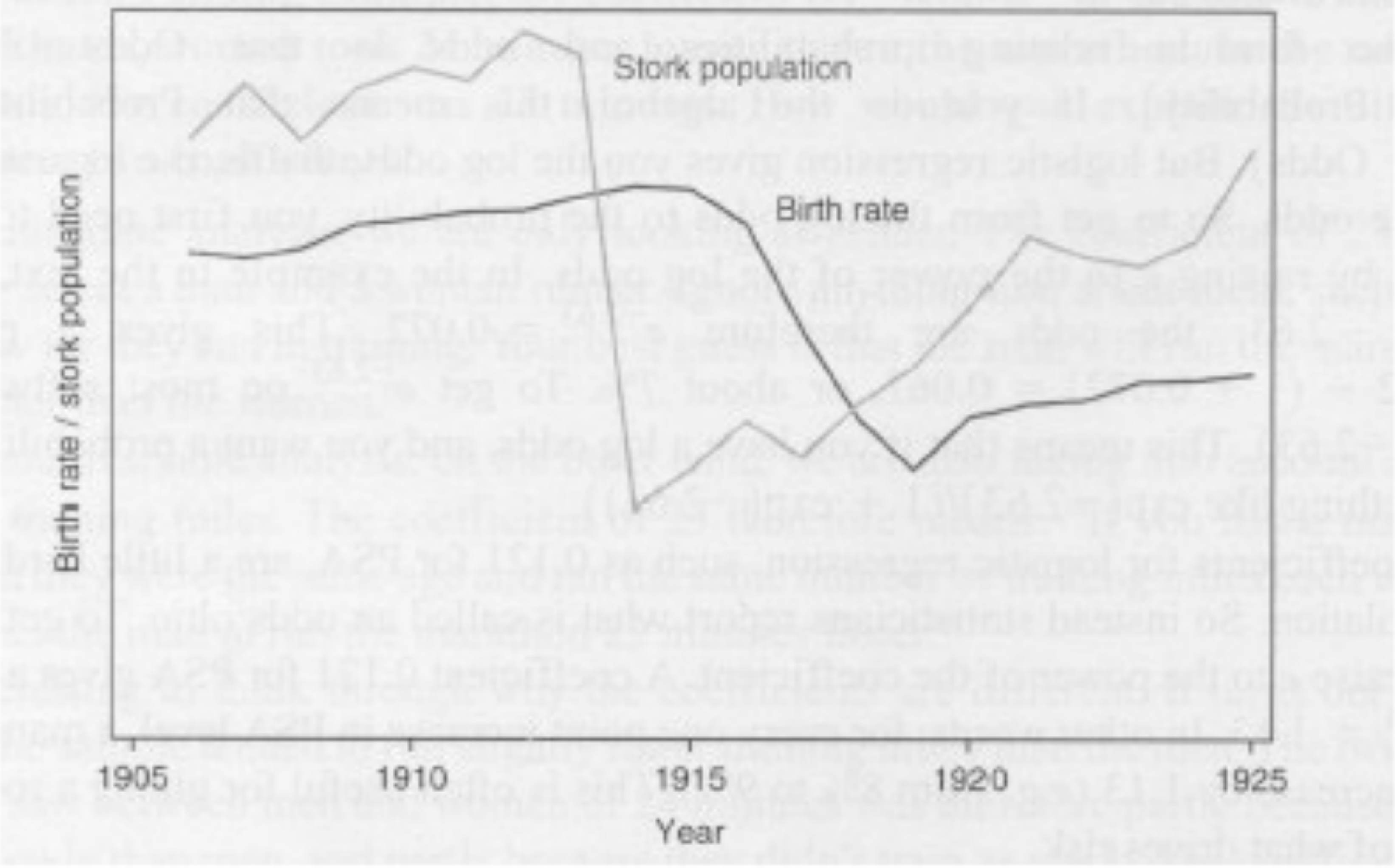
- ~~“The more cigarettes we consume, the longer we live!”~~

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

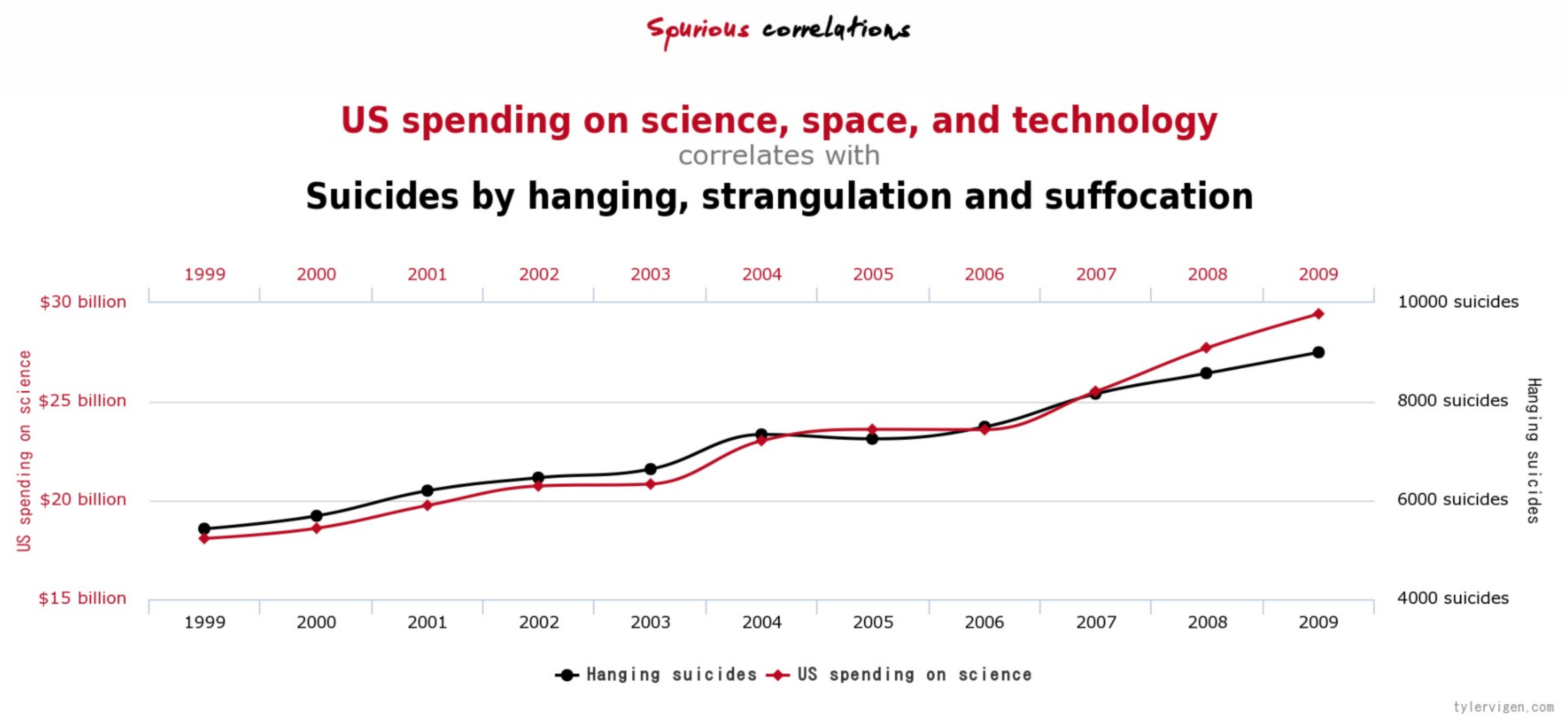


(Source: World Health Organization)

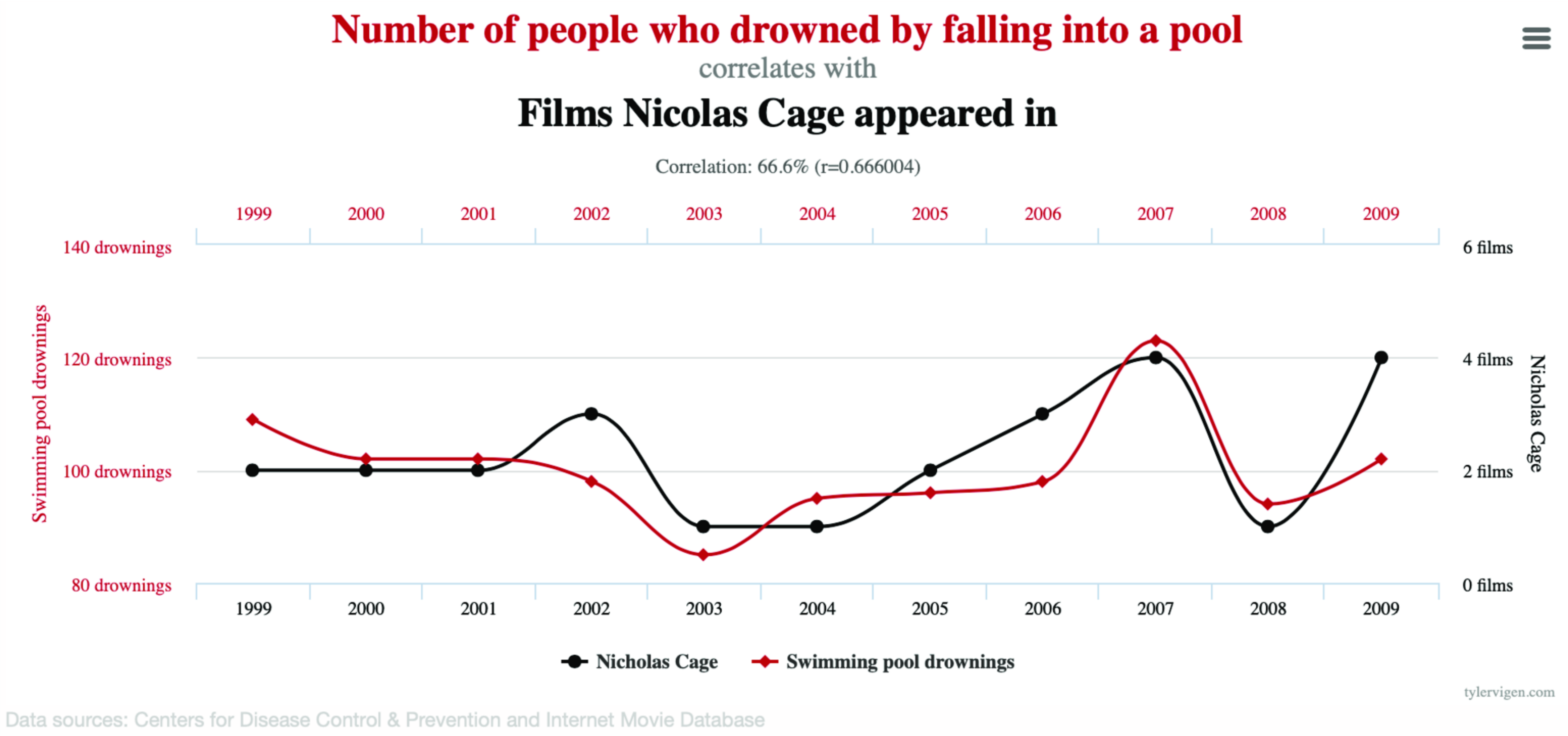
Correlation != causality



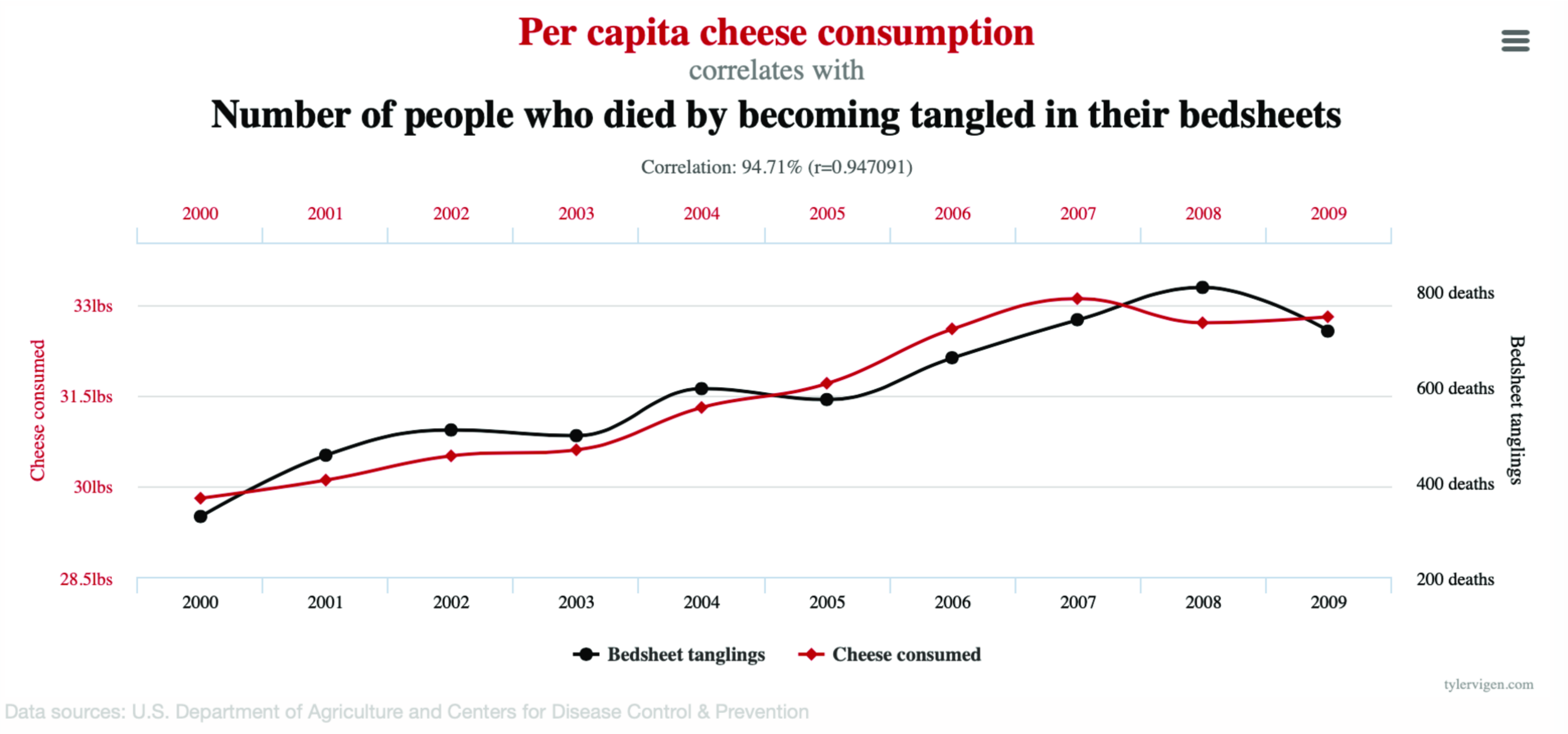
and foot size is positively correlated with reading ability, etc.



http://www.tylervigen.com/spurious-correlations



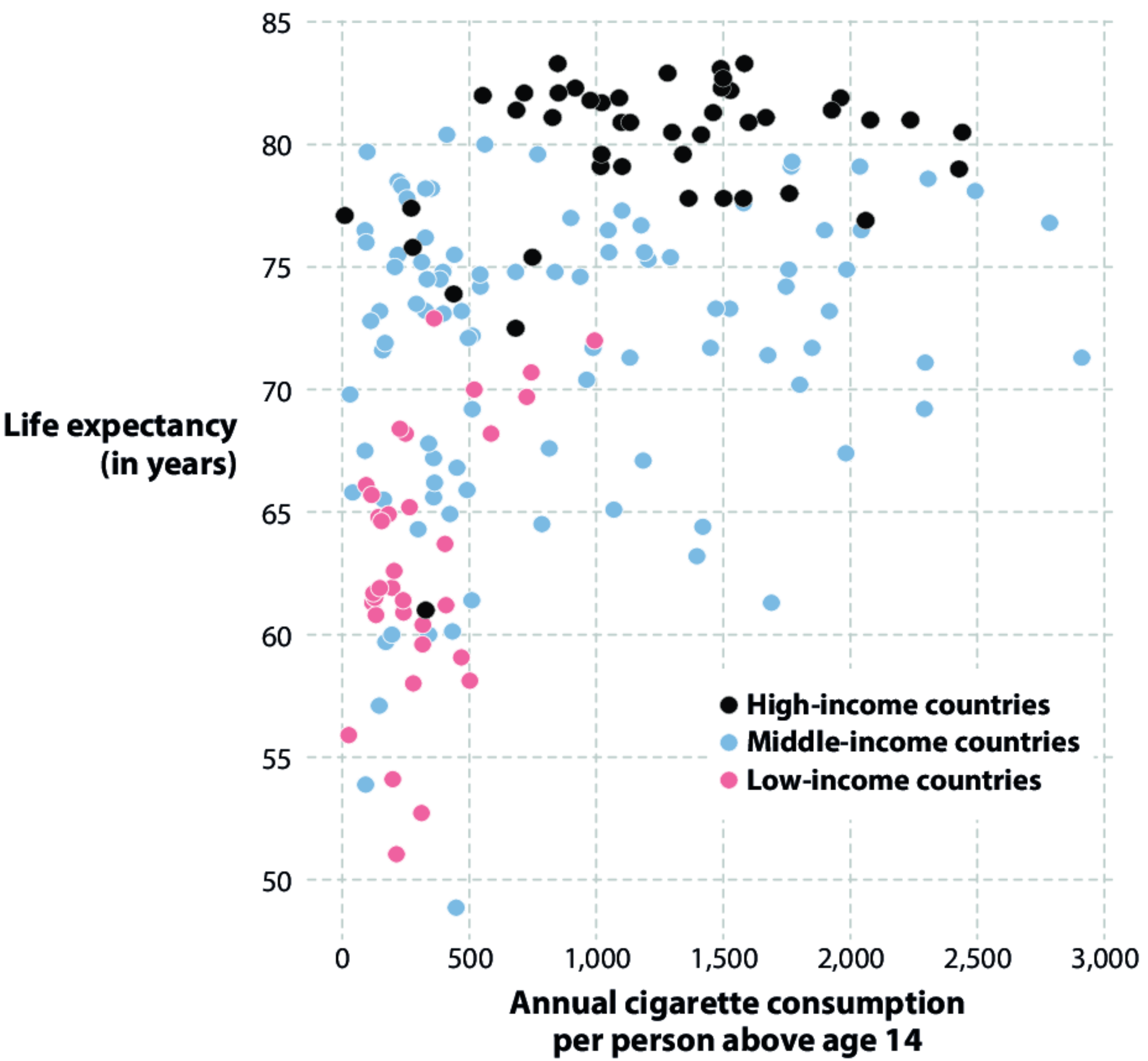
http://www.tylervigen.com/spurious-correlations

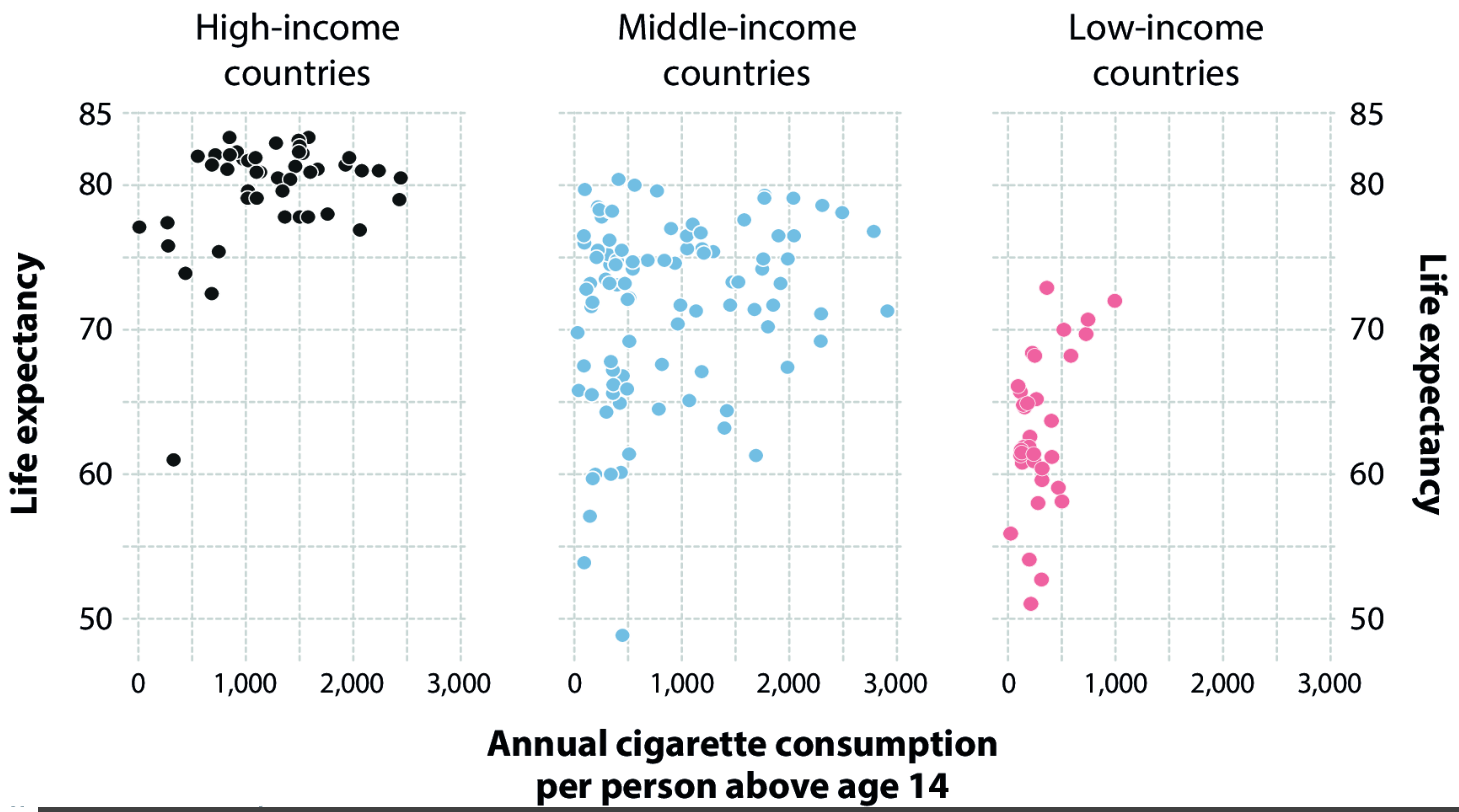


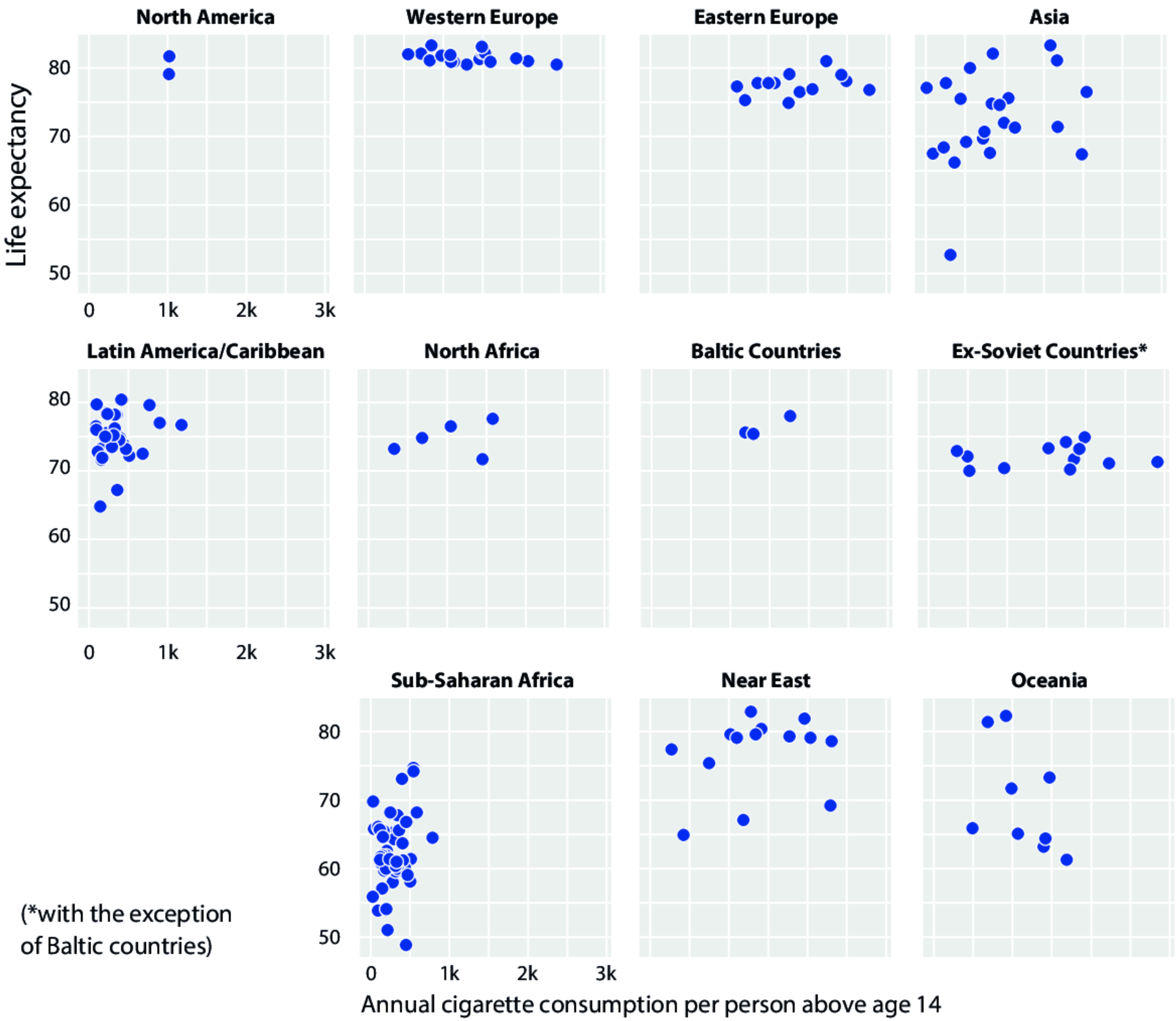
http://www.tylervigen.com/spurious-correlations

- ~~“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
85
80
70
60
50
0
1,000
2,000
3,000
Life expectancy
Annual cigarette consumption per person above age 14



(*with the exception of Baltic countries)

Simpson's Paradox

- trend that appears in several different groups of data but disappears or reverses when these groups are combined

	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%**

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%

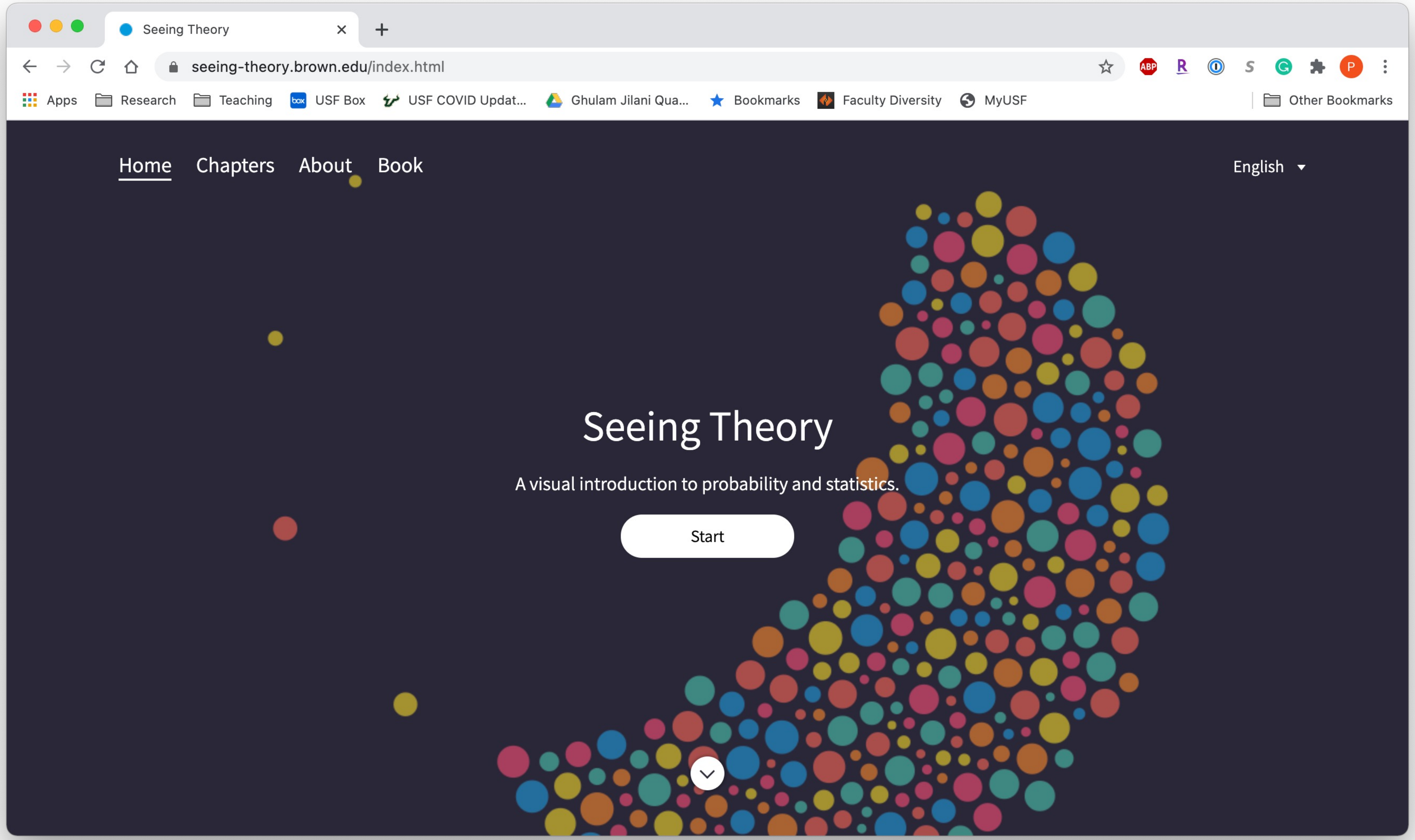
Economix
Explaining the Science of Everyday Life

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%



http://students.brown.edu/seeing-theory/index.html