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



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

FILTERING, AGGREGATION, & STATS

#### **Reducing Items and Attributes**

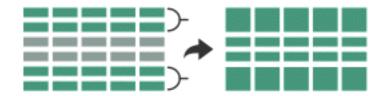
- → Filter
  - → Items



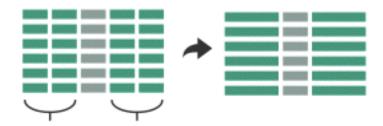
→ Attributes



- **→** Aggregate
  - → Items



→ Attributes





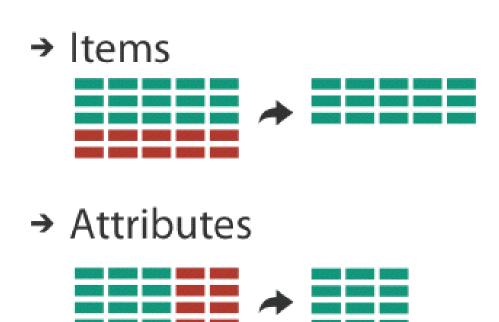
# why reduce?

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



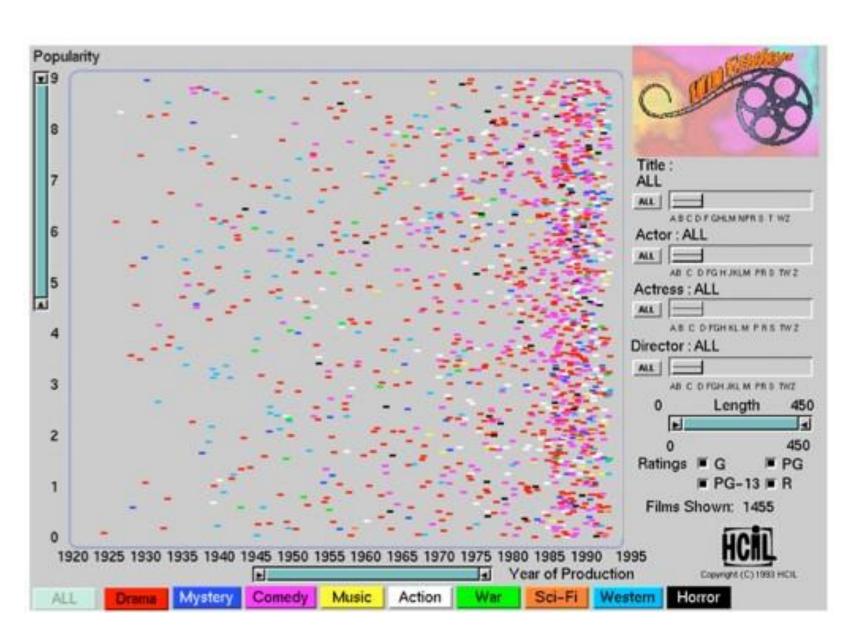
#### filter

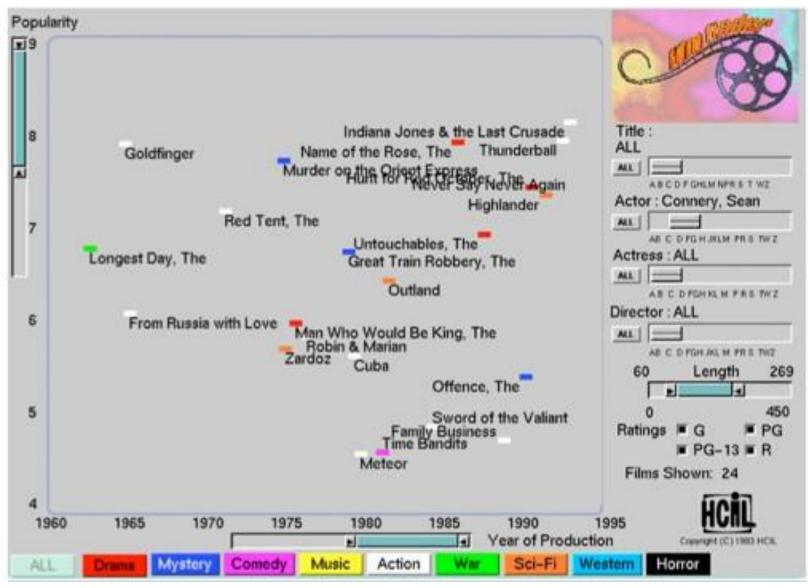
- elements are eliminated to support dynamic queries
  - coupling between encoding and interaction so that user can immediately see the results of an action





#### ITEM FILTERING



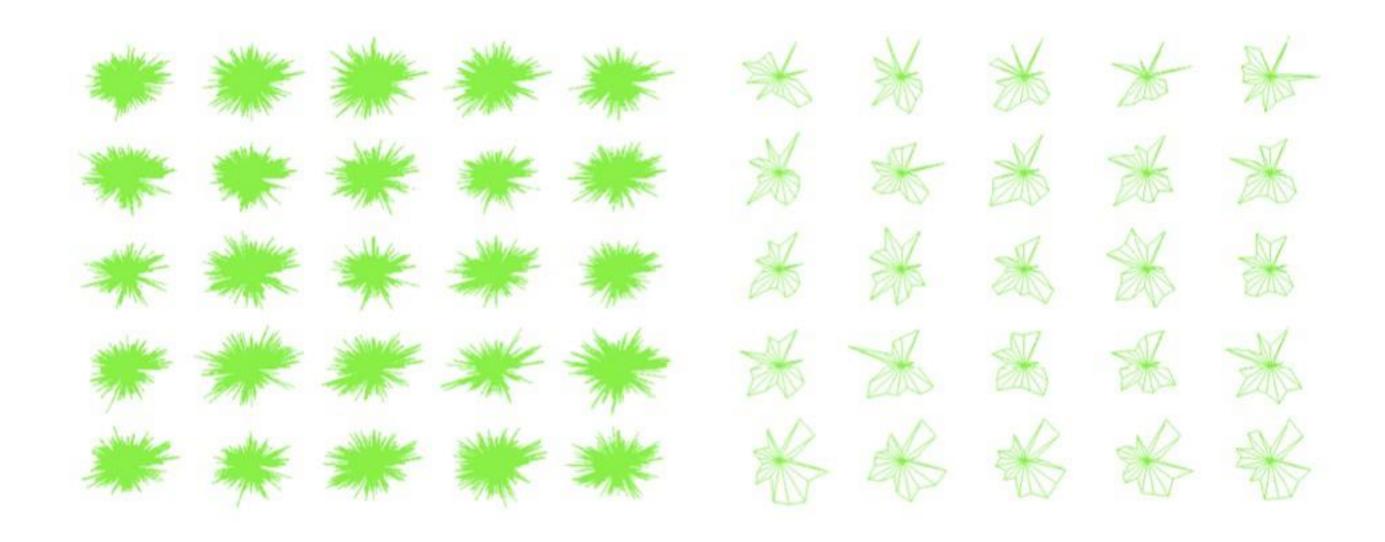








#### ATTRIBUTE FILTERING







#### 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

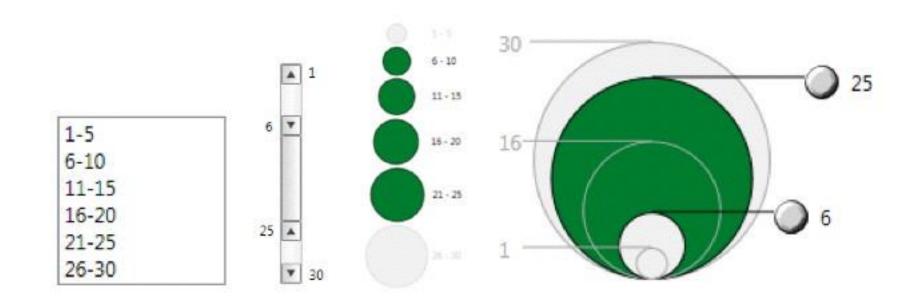






#### Controlling Filtering: interactive legends

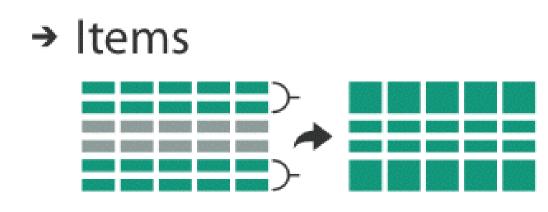
- controls combining the visual representation of static legends with interaction mechanisms of widgets
- define and control visual display together



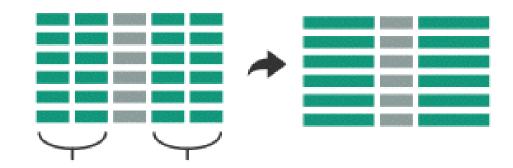


#### aggregate

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



→ Attributes





#### Numerous ways to reduce...

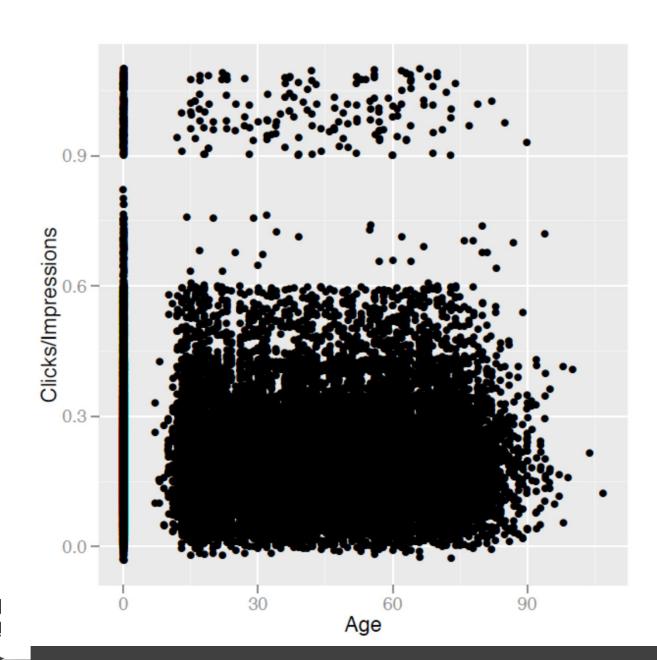
• statistics, topology, machine learning, etc.

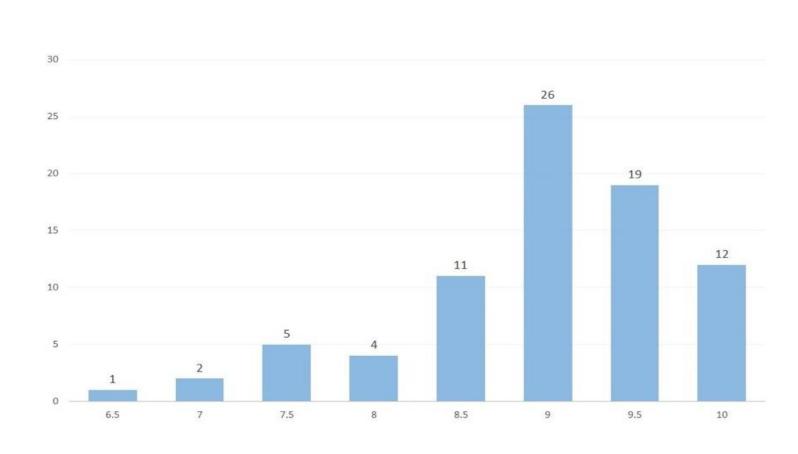




#### Problem #1: Aggregate Items

We have too many data points to show

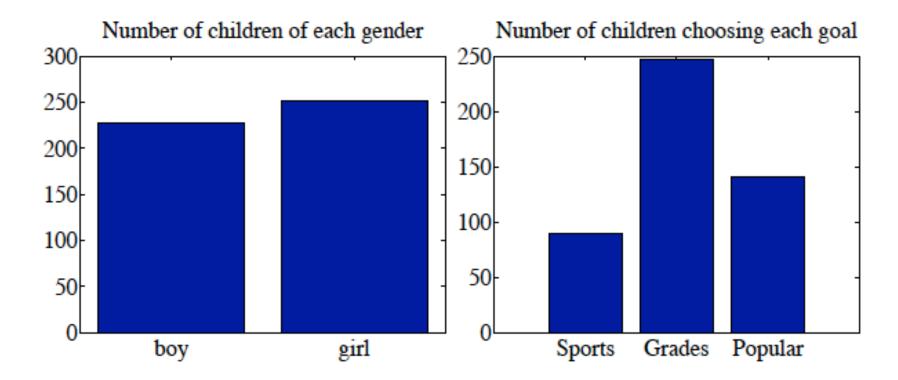




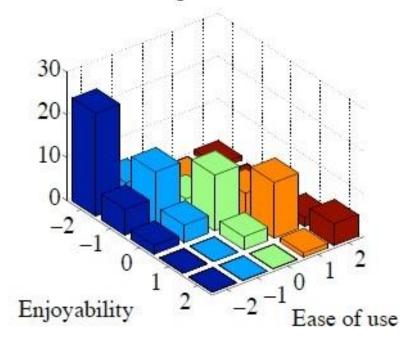


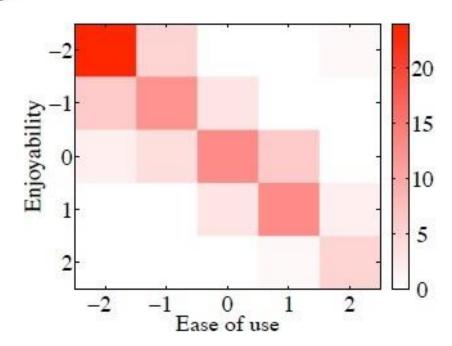
#### Histograms

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



Counts of user responses for a user interface



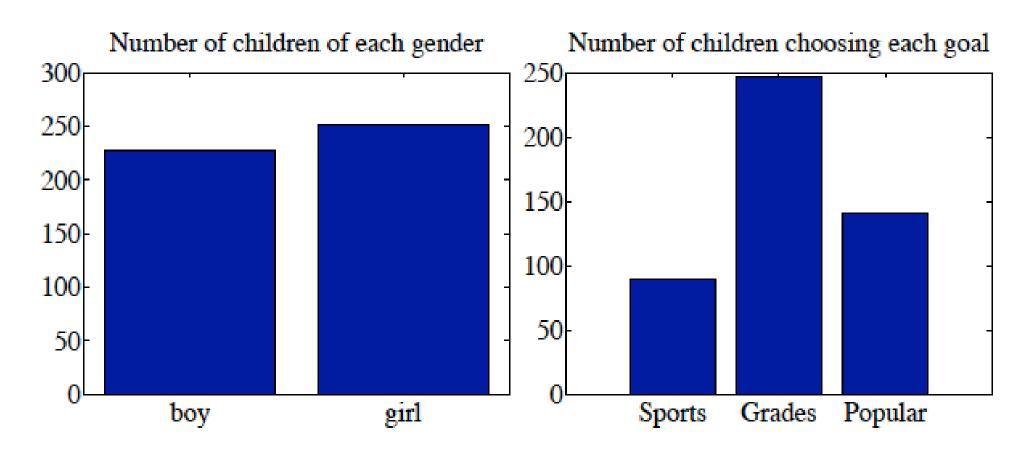




#### Categorical data

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

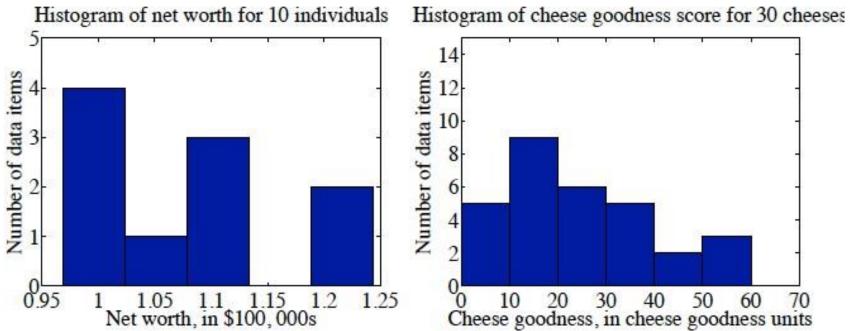




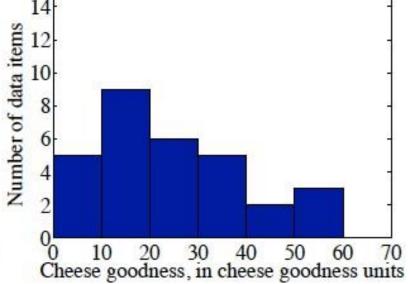
#### Continuous Data Histograms

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









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

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



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



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





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





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





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





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





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





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



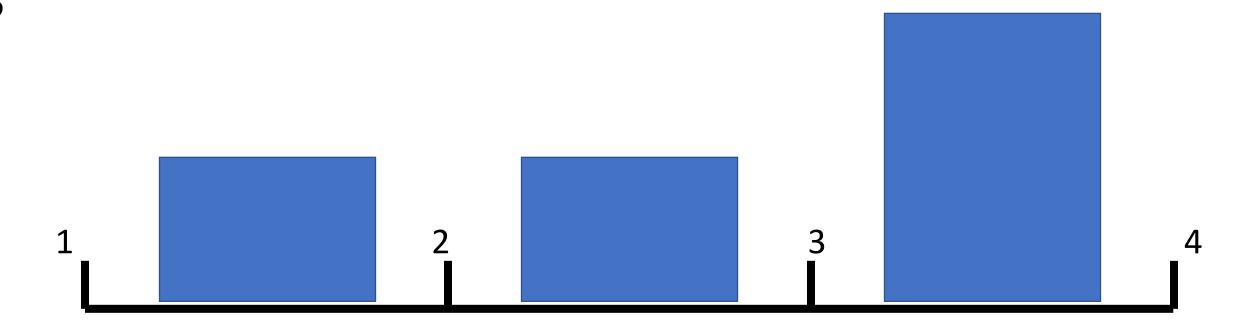


- $X=\{1,2.5,3,4\}$
- k = 3
- 4 -> floor(3\*(4-1)/(4-1)) = Bin 2



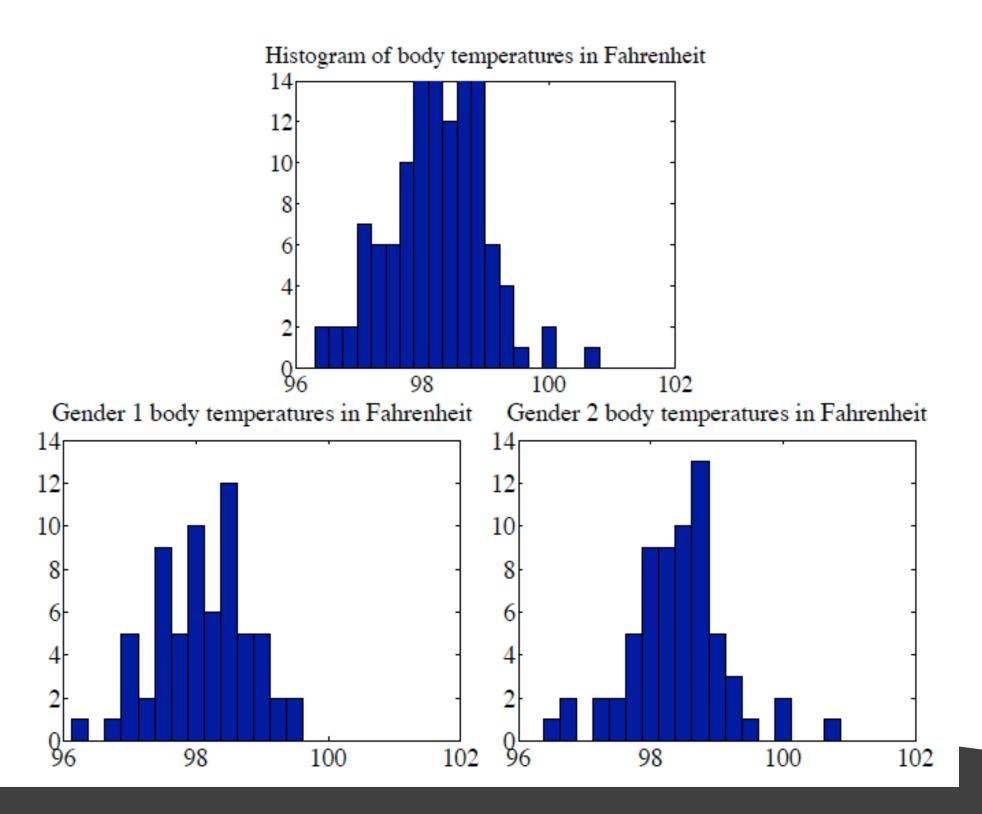


- X={1,2.5,3,4}
- k = 3





## Conditional Histograms





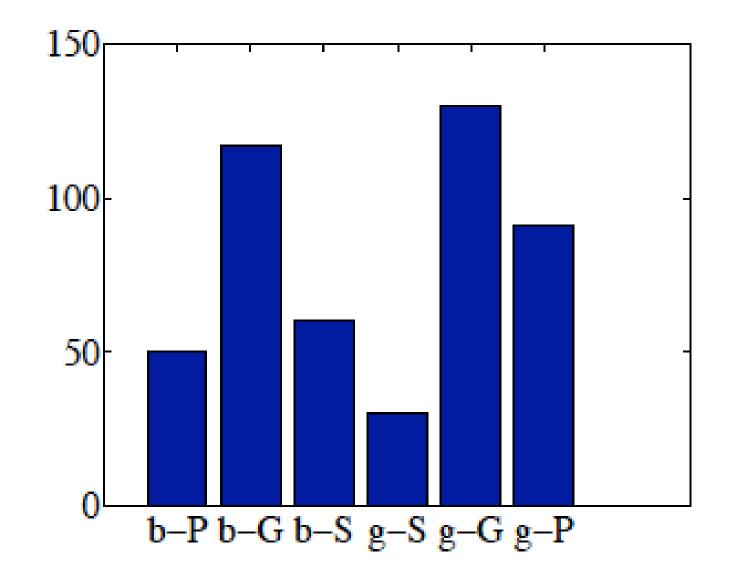
# 2D Histograms





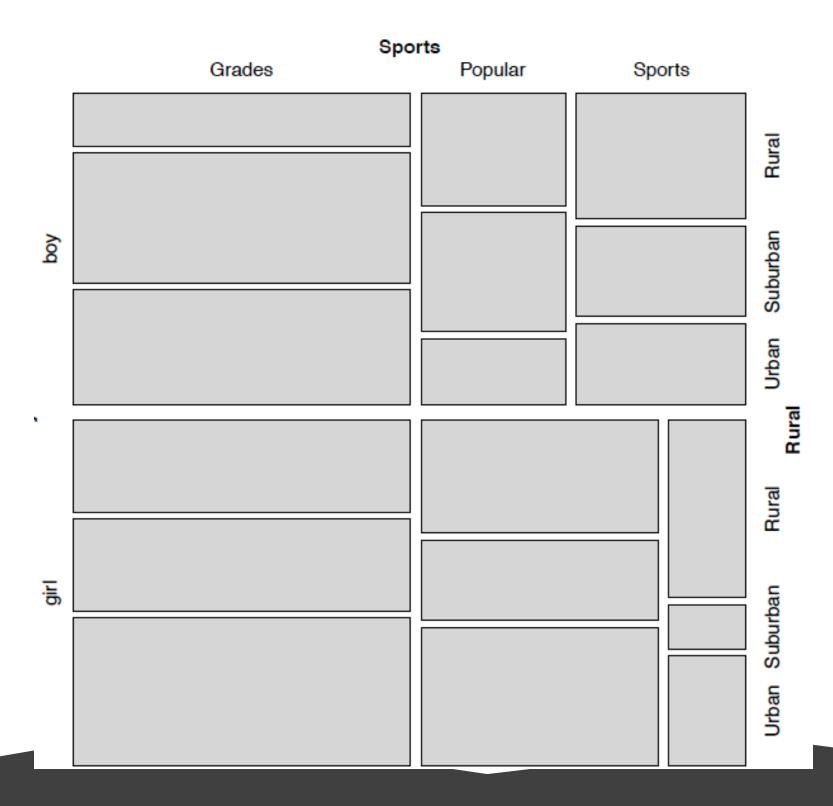
# Categorical data

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



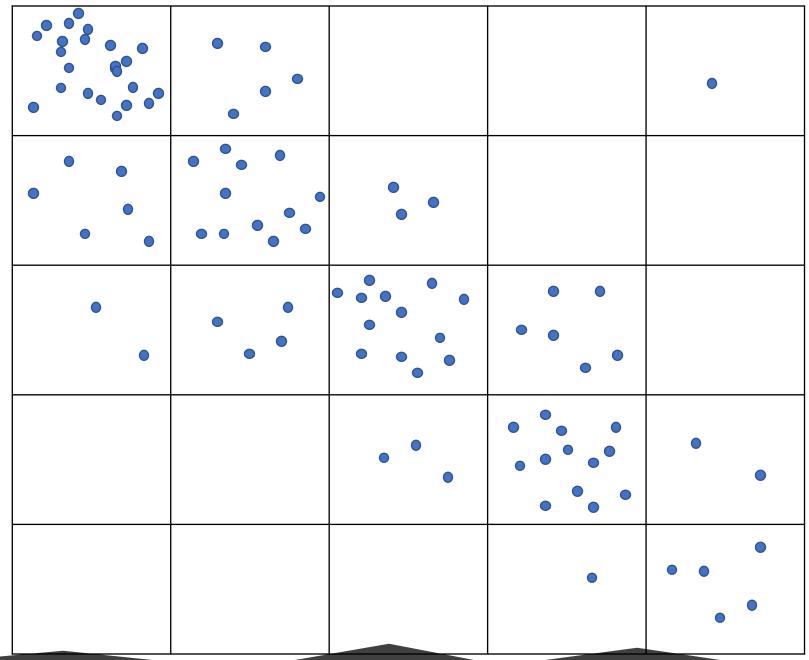


#### Mosaic Plots





#### Ordinal data

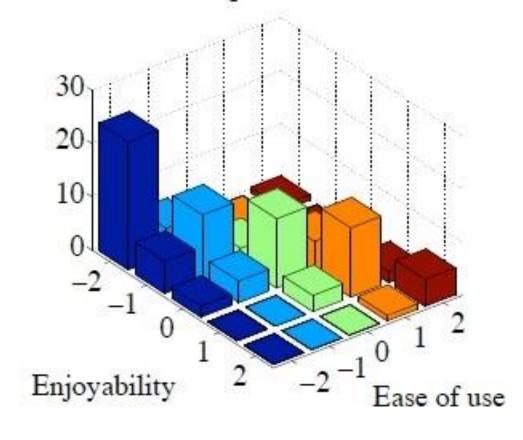


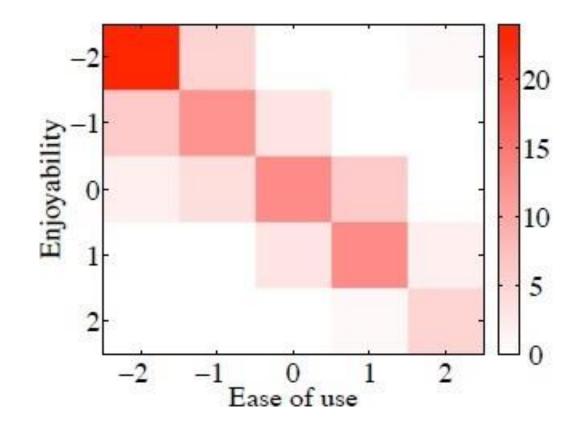
	-2	-1	0	1	2
-2	24	5	0	0	1
-1	6	12	3	0	0
0	2	4	13	6	0
1	0	0	3	13	2
2	0	0	0	1	5



#### Ordinal data

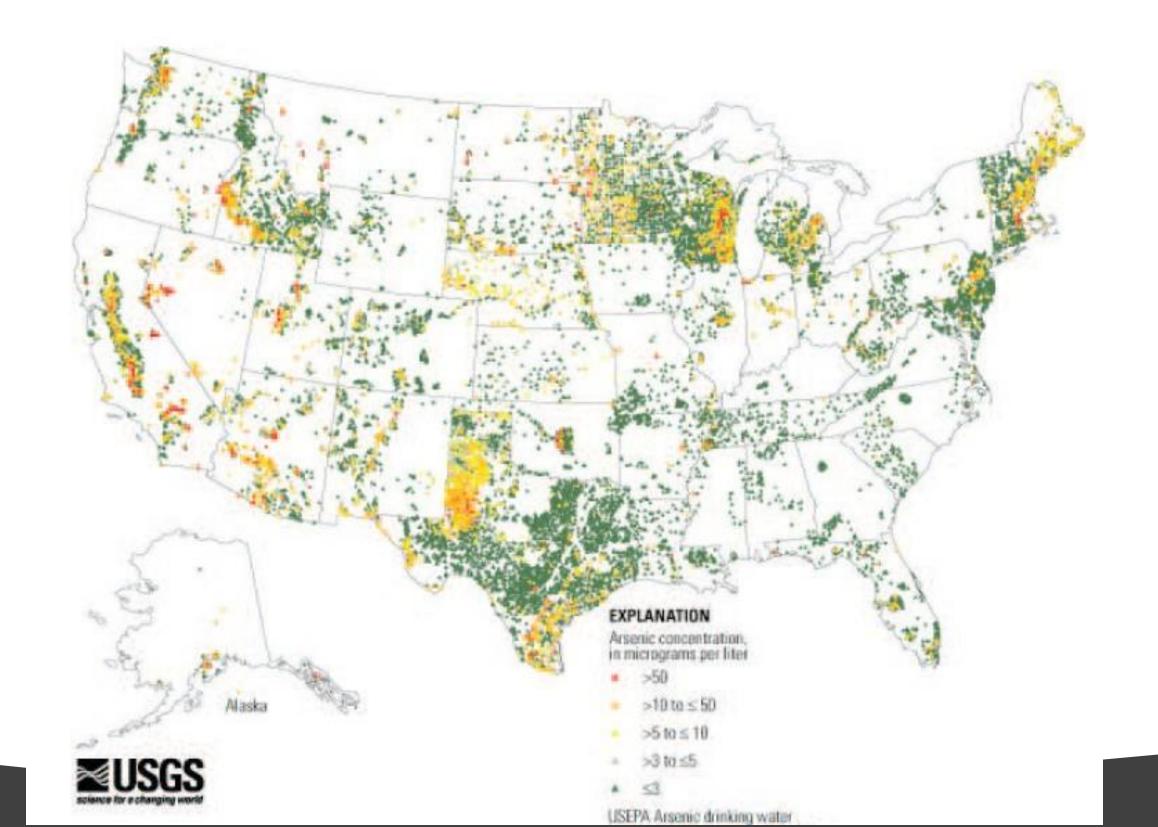
#### Counts of user responses for a user interface





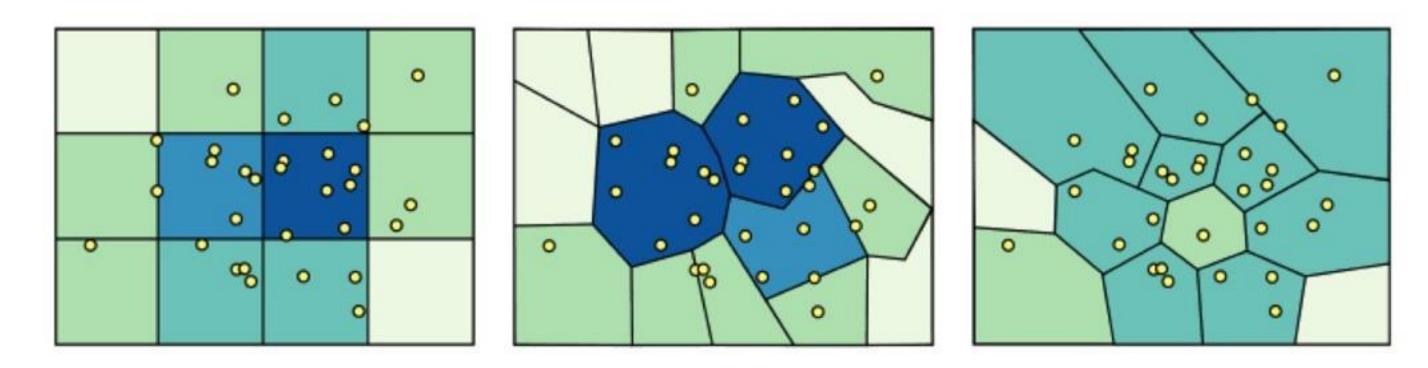


#### Arsenic in well water





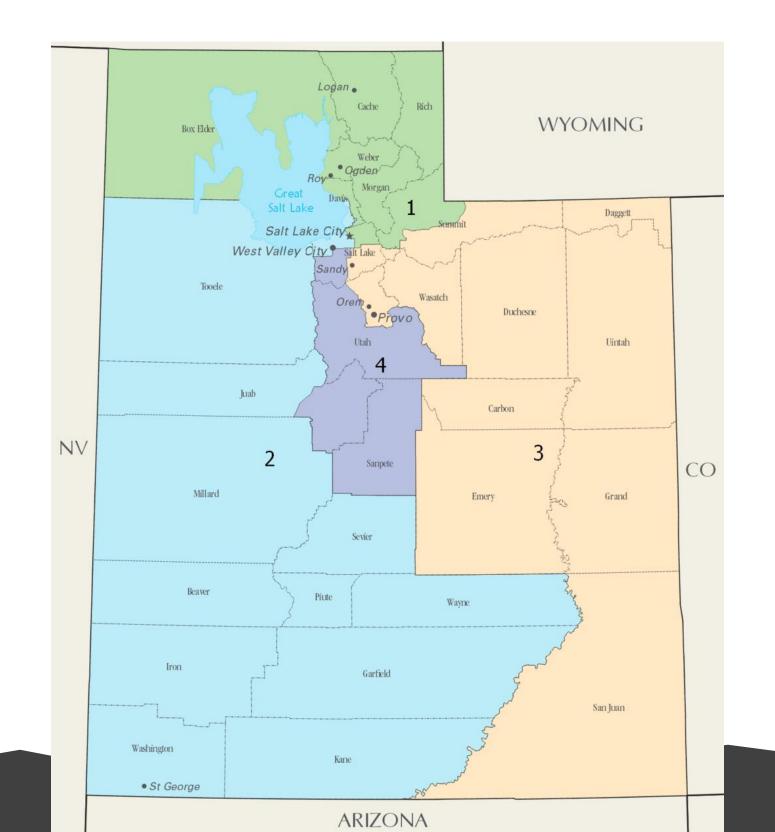
#### spatial aggregation



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

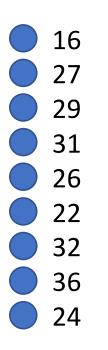


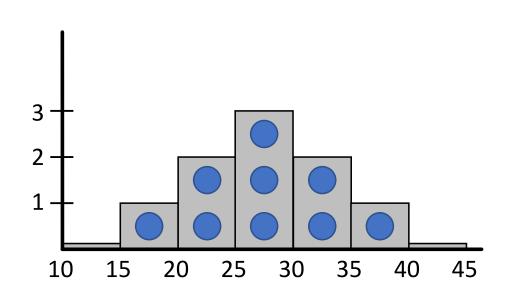
#### spatial aggregation: Congressional Districts





## Histogram Challenges: Selecting Resolution



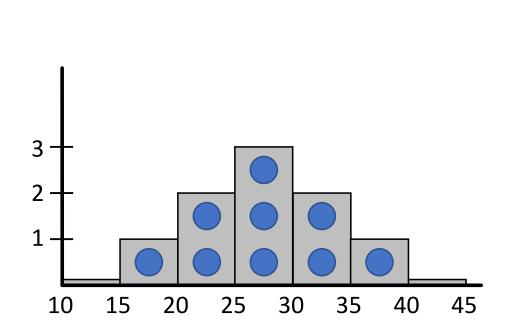


Mean (Average) = 27 Standard Deviation = 6

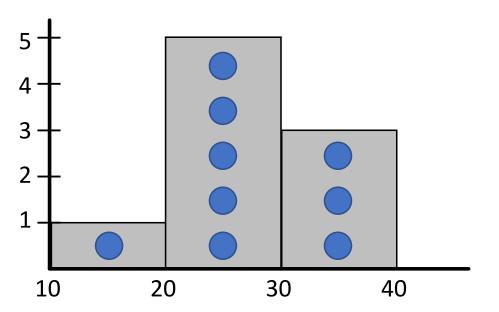


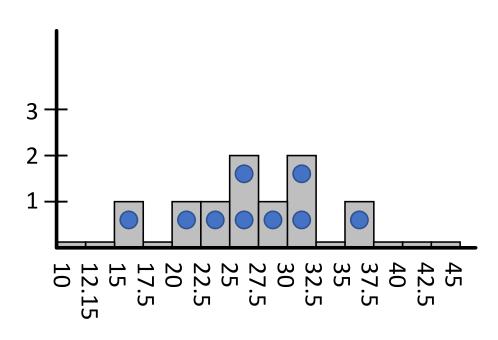
## Histogram Challenges: Selecting Resolution





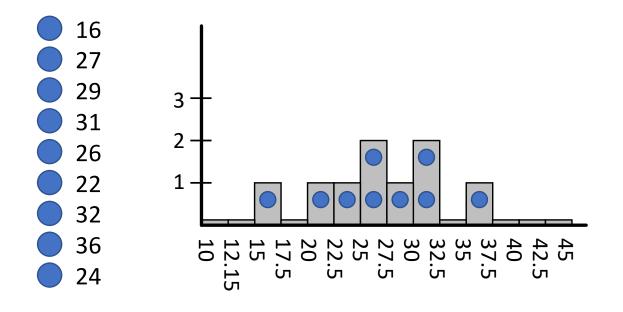


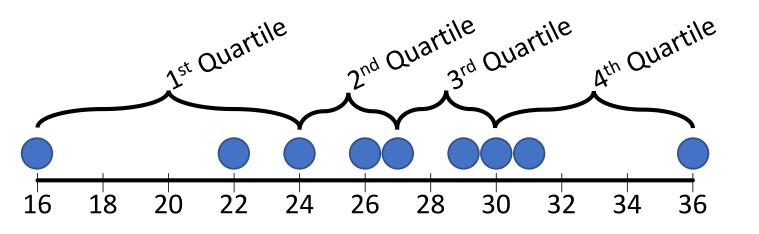






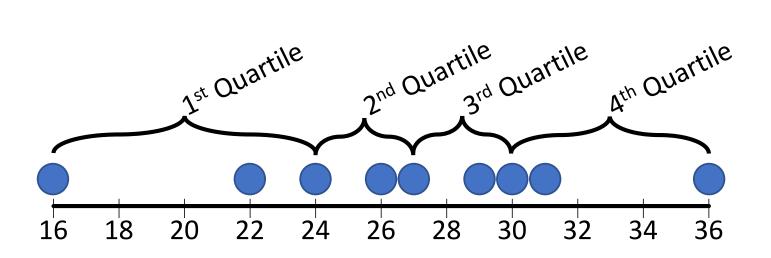
# An Example: Comparing Histogram & Distribution

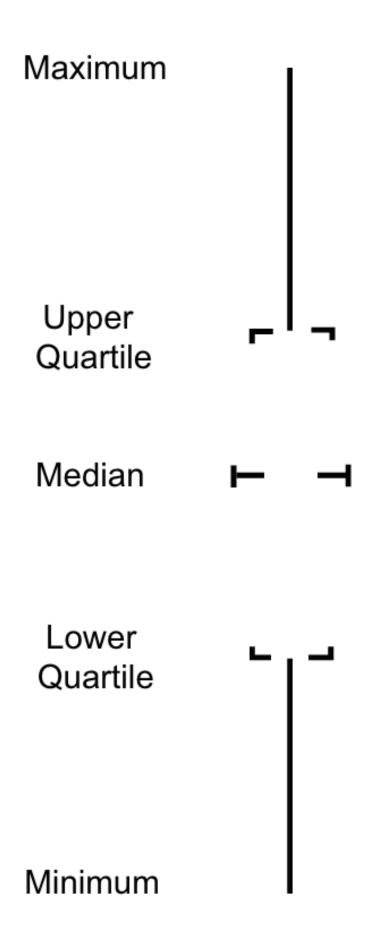






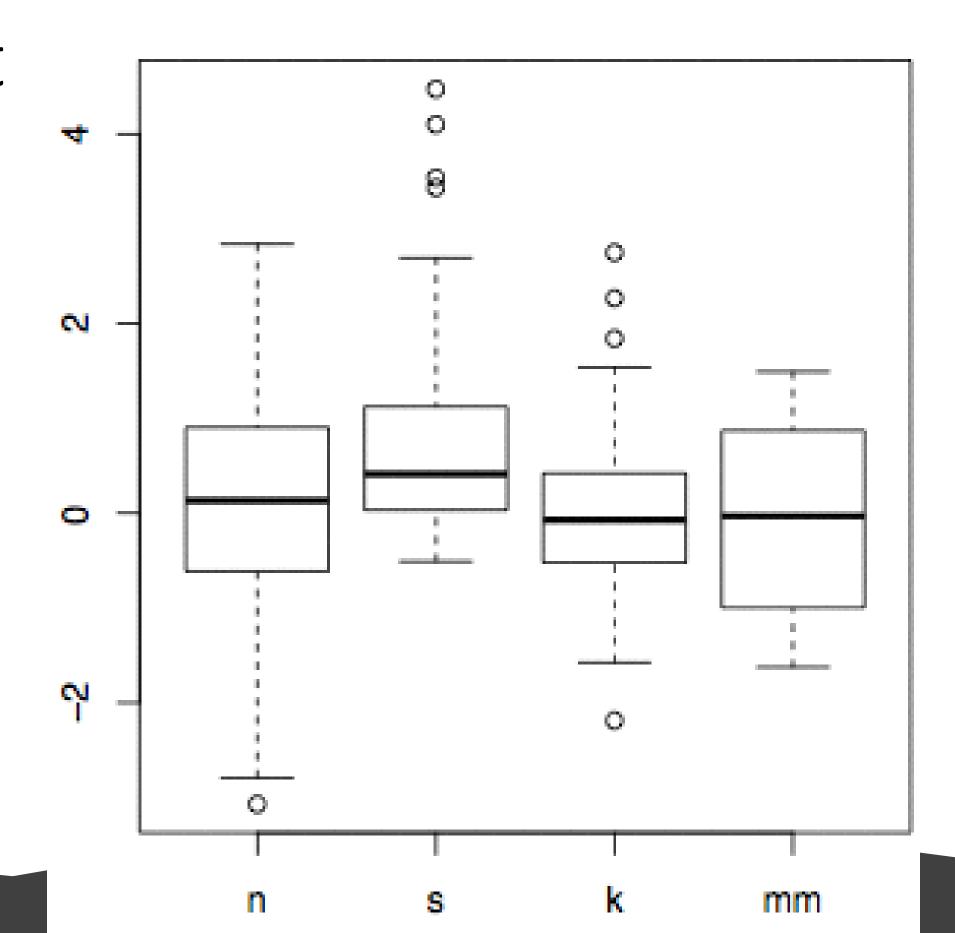
## Boxplot





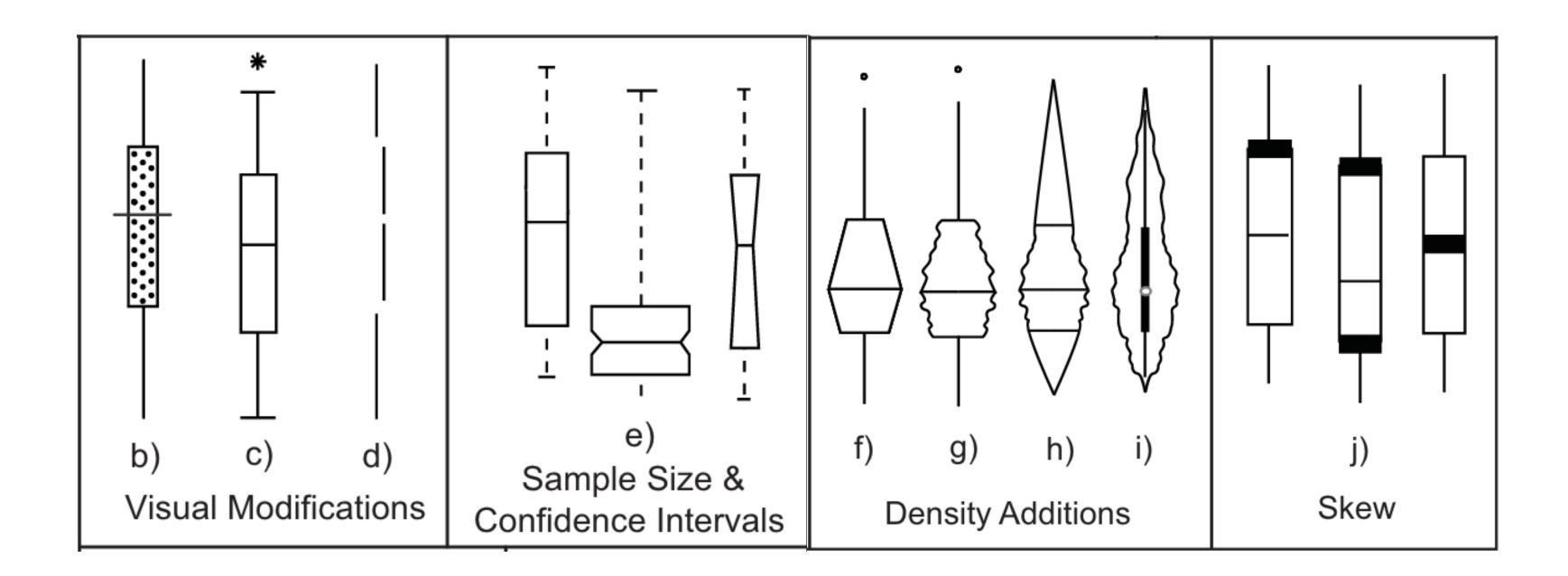


## Boxplot



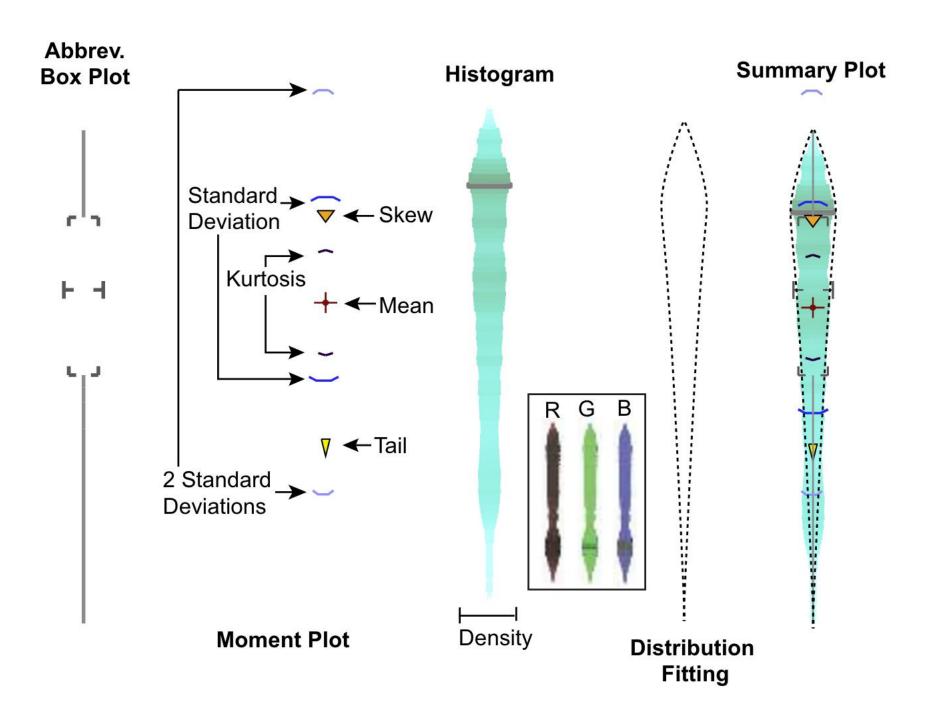


## Boxplots





## 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



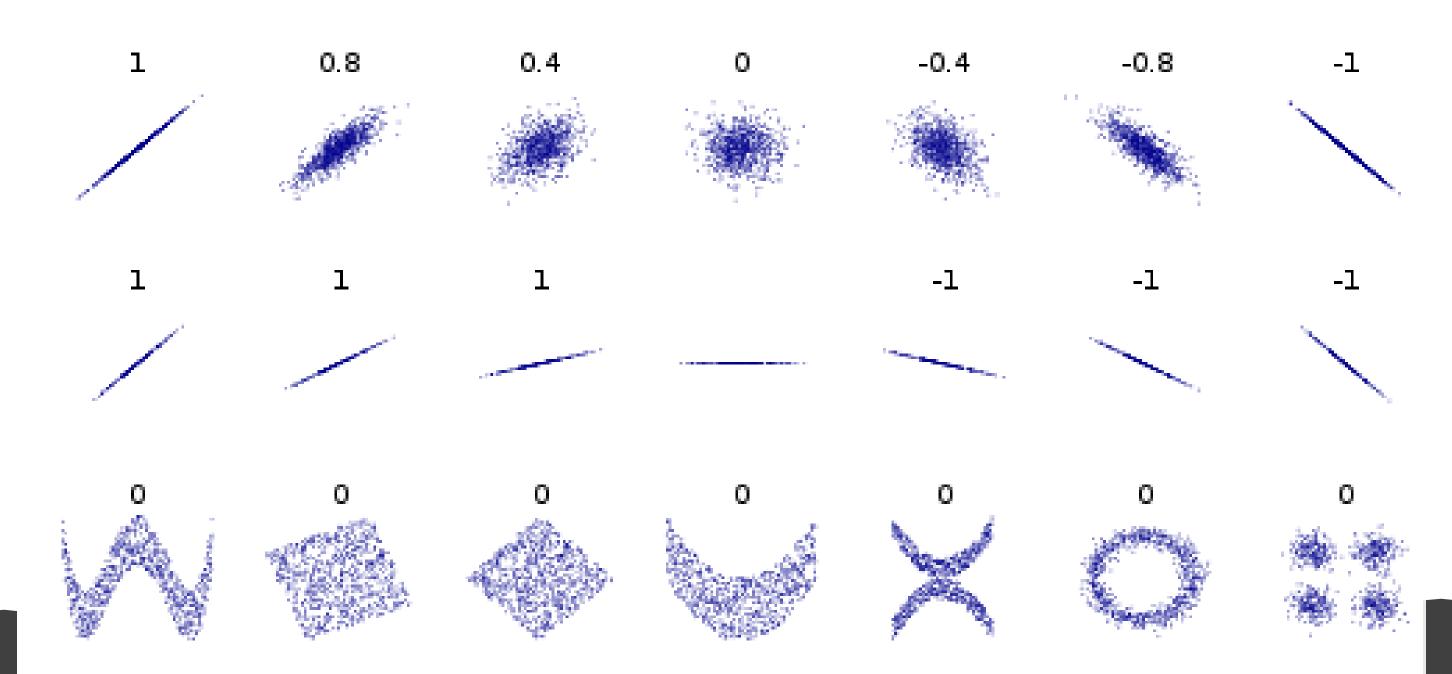
## Similarity scores

- 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<sup>2</sup>)



## Pearson Correlation Coefficient

• A measure of the linearity between 2 sets





$$ho_{X,Y} = rac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

### where:

- cov is the covariance
- ullet  $\sigma_X$  is the standard deviation of X
- ullet  $\sigma_Y$  is the standard deviation of Y



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

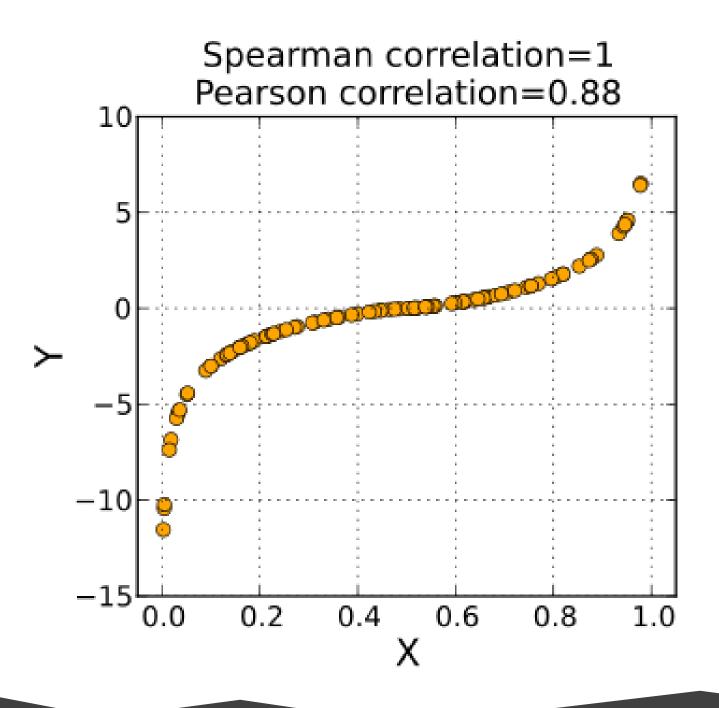
where:

ullet  $n, x_i, y_i$  are defined as above

$$ullet ar x = rac{1}{n} \sum_{i=1}^n x_i$$
 (the sample mean); and analogously for  $ar y$ 



## 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

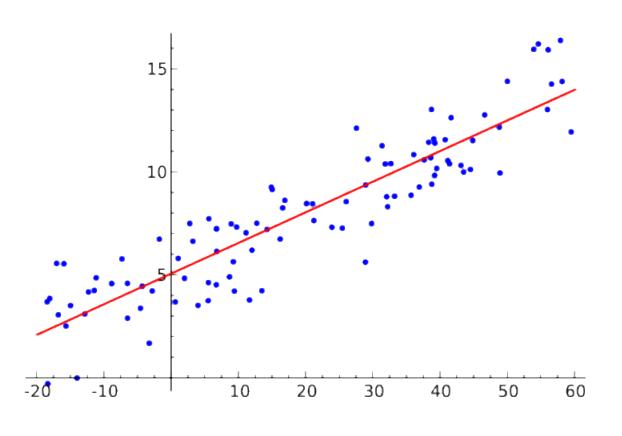
<u>IQ</u> , (X)	Hours of TV per week, (Y)	rank (X')	rank (Y')
86	0	1	1
97	20	2	6
99	28	3	8
100	27	4	7
101	50	5	10
103	29	6	9
106	7	7	3
110	17	8	5
112	6	9	2
113	12	10	4



## 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.  $\Sigma \varepsilon_i^2$ )



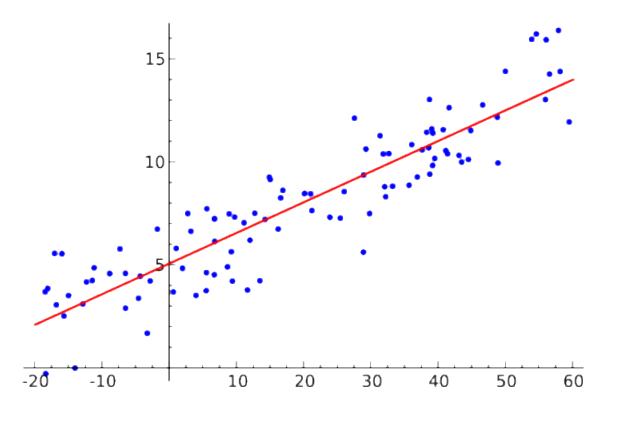


## Regression: Fitting a Model to Data

Can be computed directly

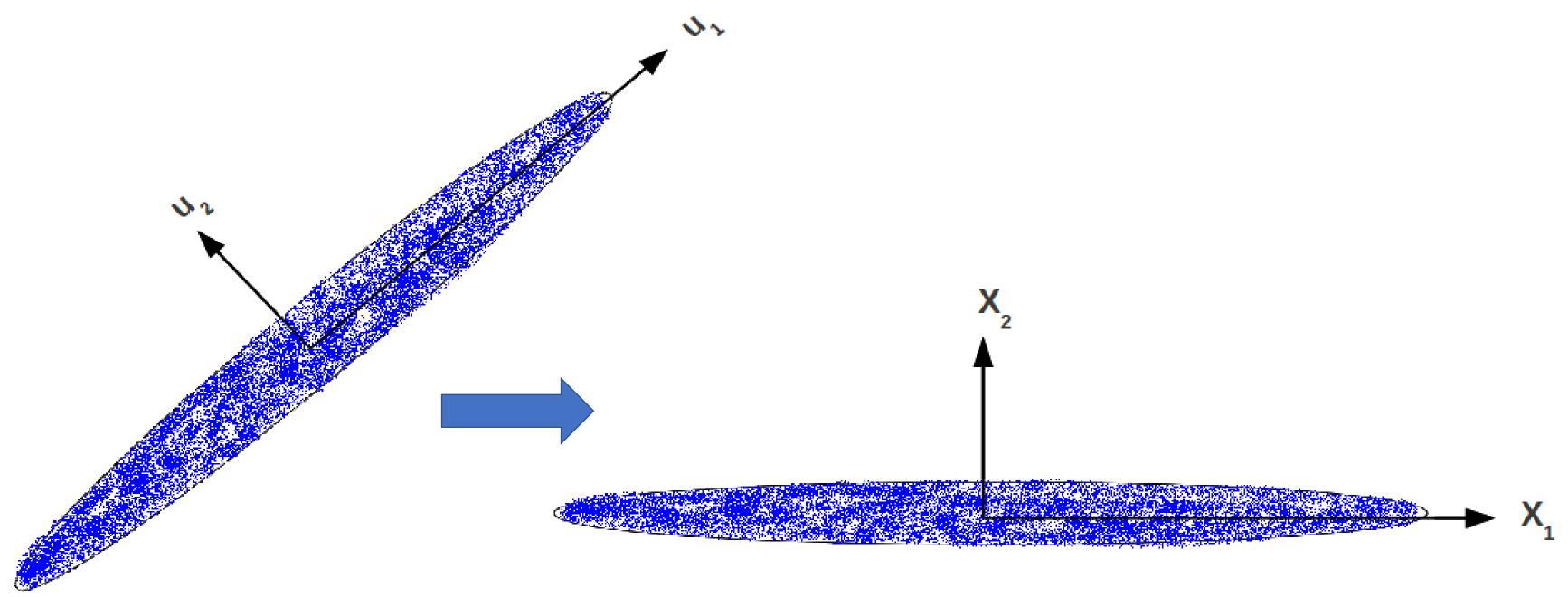
$$\hat{eta} = rac{\sum_{i=1}^n (x_i - ar{x})(y_i - ar{y})}{\sum_{i=1}^n (x_i - ar{x})^2}$$

$$\hat{lpha}=ar{y}-\hat{eta}\,ar{x}$$



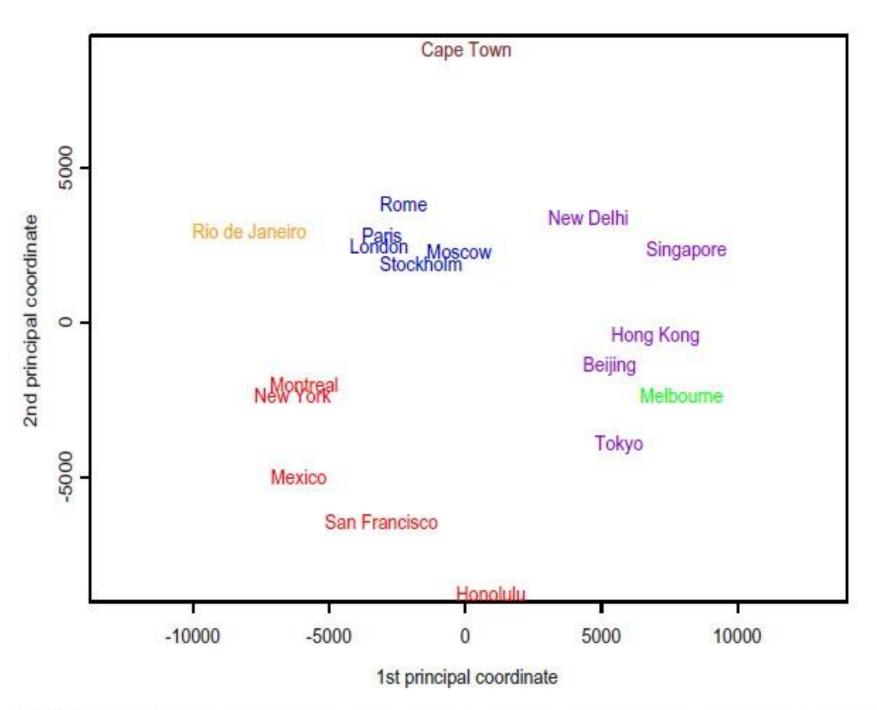


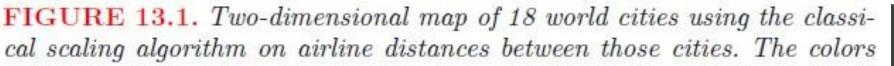
# Linear Dimensionality reduction: Principal Component Analysis (PCA)





# Nonlinear Dimensionality Reduction: Multidimensional Scaling (MDS)







## 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	У	У	У		X	У
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

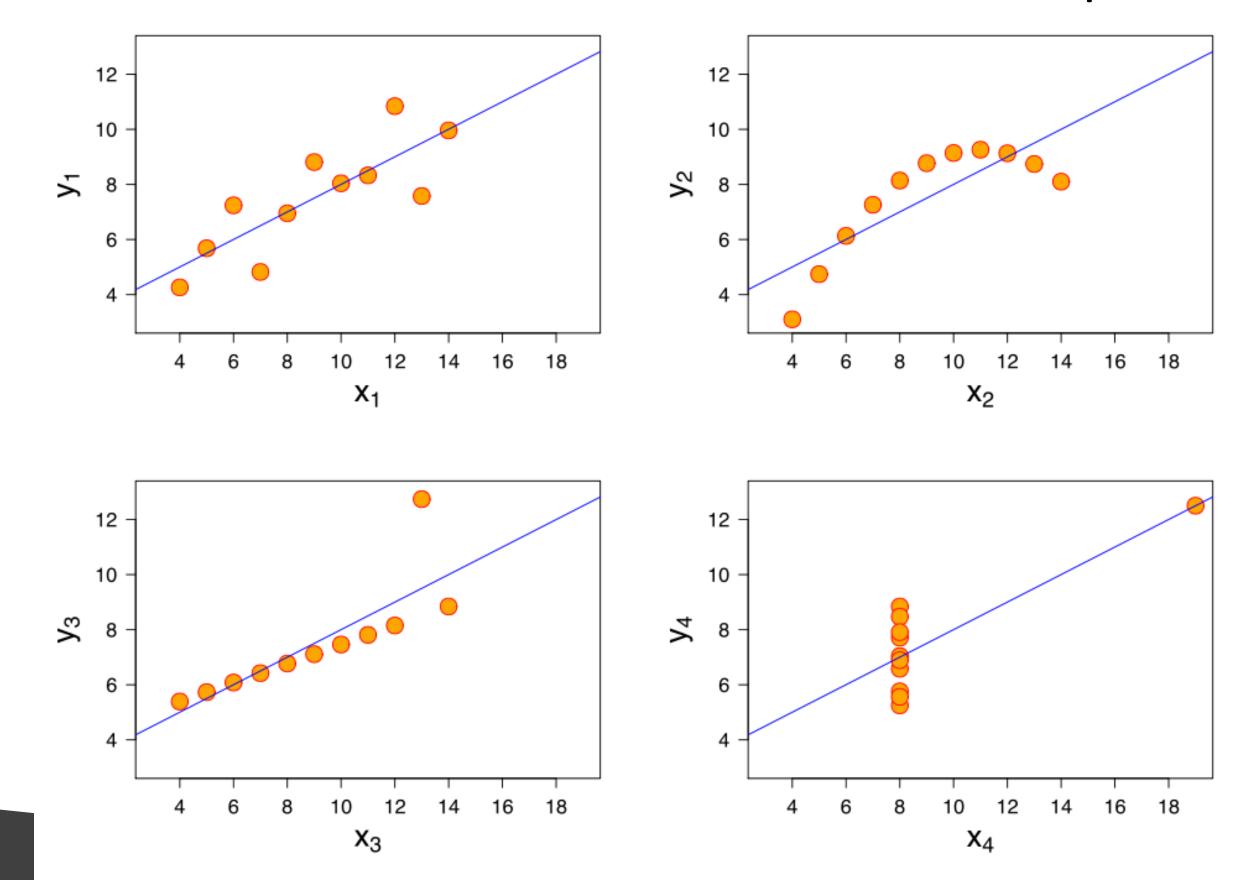
TABLE. Four data sets, each comprising 11 (x, y) pairs.



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

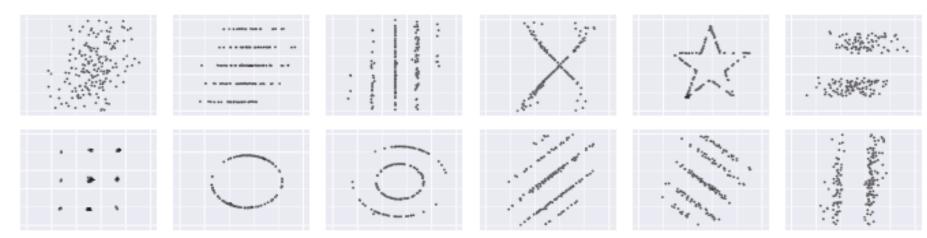


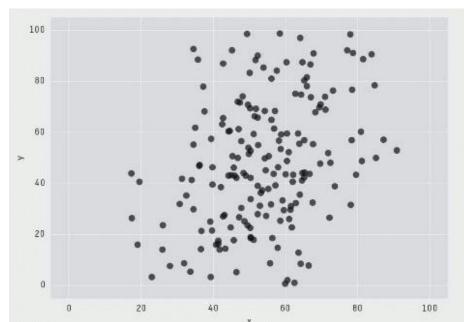
Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. ( $\bar{x} = 54.02$ ,  $\bar{y} = 48.09$ ,  $sd_x = 14.52$ ,  $sd_y = 24.79$ , Pearson's r = +0.32)

#### 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.

While year popular and affective for illustrating the



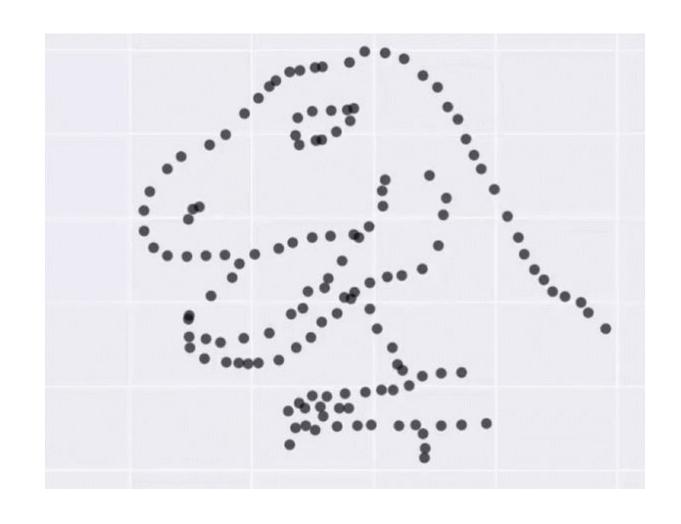
X Mean: 54.0236753

Y Mean: 48.0970794

(SD: 14.5298540

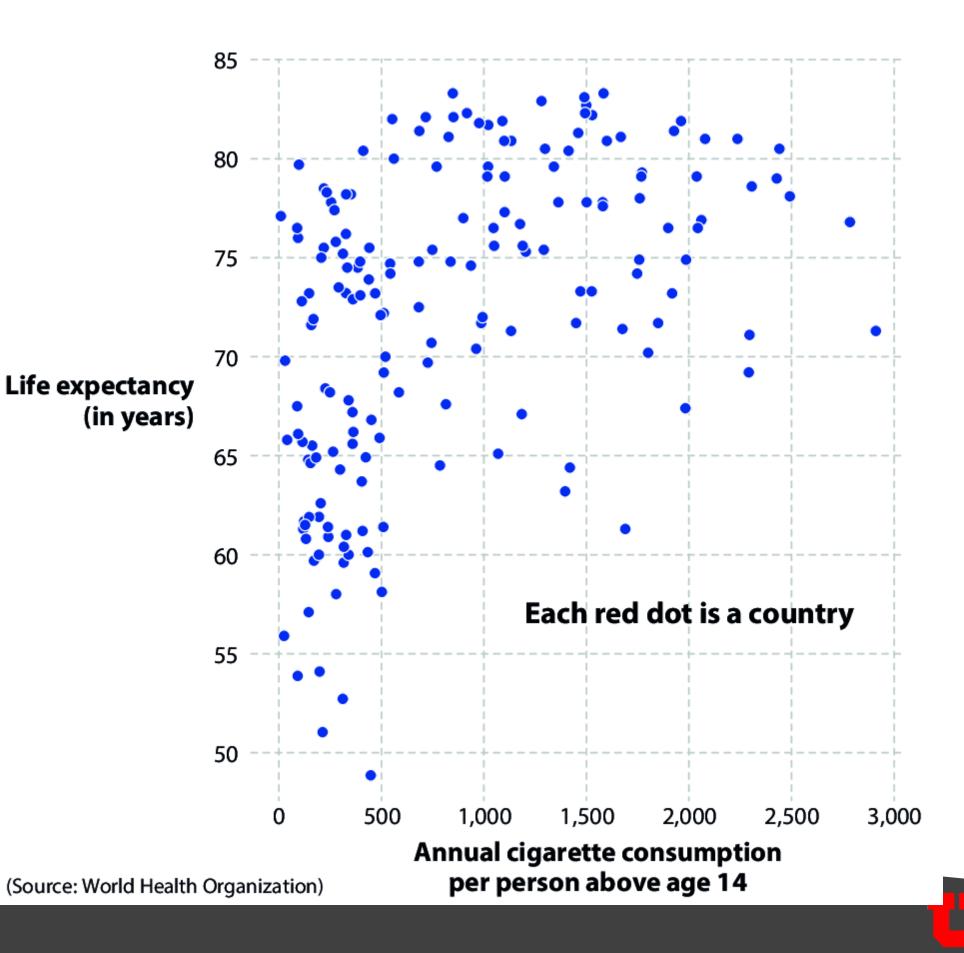
Y SD : 24.7943127

Corr. : +0.3280926



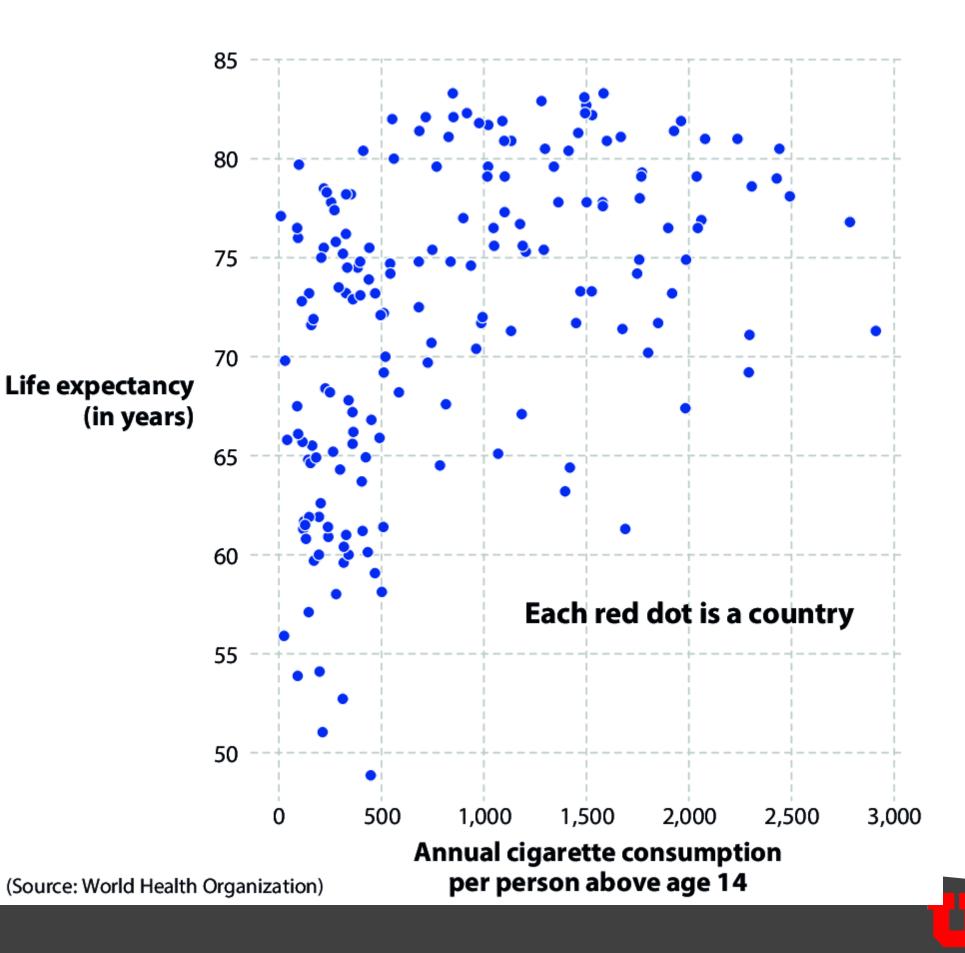


- "The more cigarettes we consume, the longer we live!"
- "There is a positive relationship between cigarette consumption and life expectancy at a countryby-country level!"



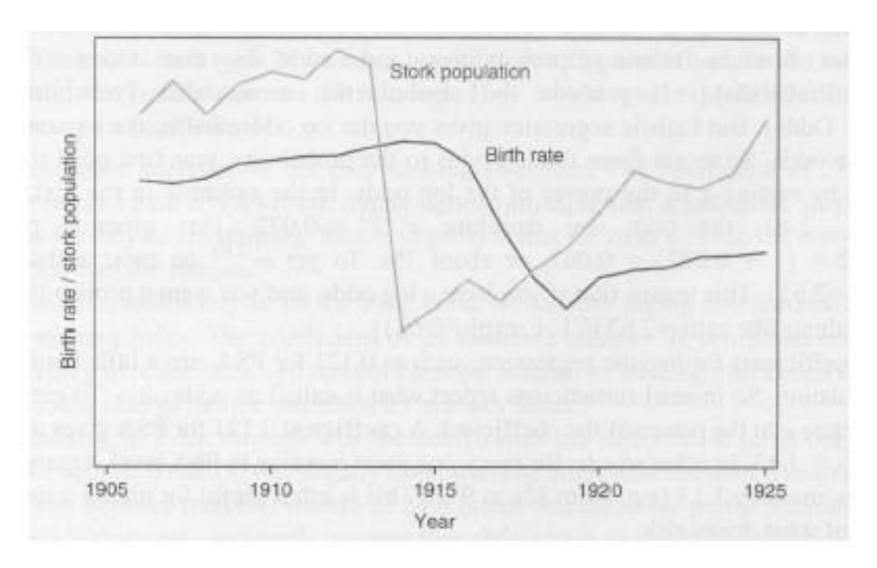


- "The more cigarettes we consume, the longer we live!"
- "There is a positive relationship between cigarette consumption and life expectancy at a countryby-country level!"





## Correlation != causality



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



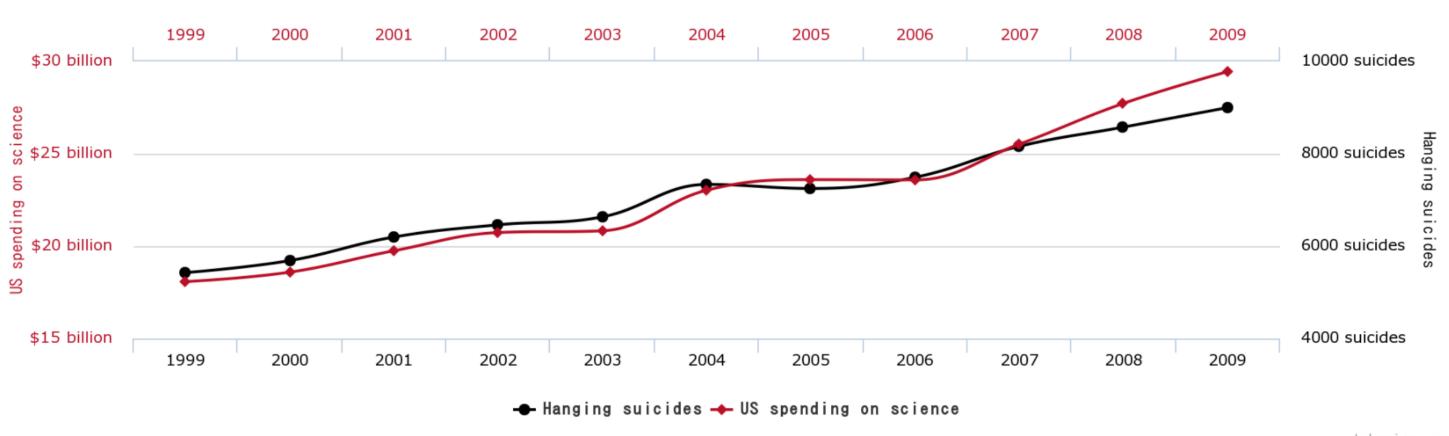


### Spurious correlations

### **US** spending on science, space, and technology

correlates with

### Suicides by hanging, strangulation and suffocation



tylervigen.com

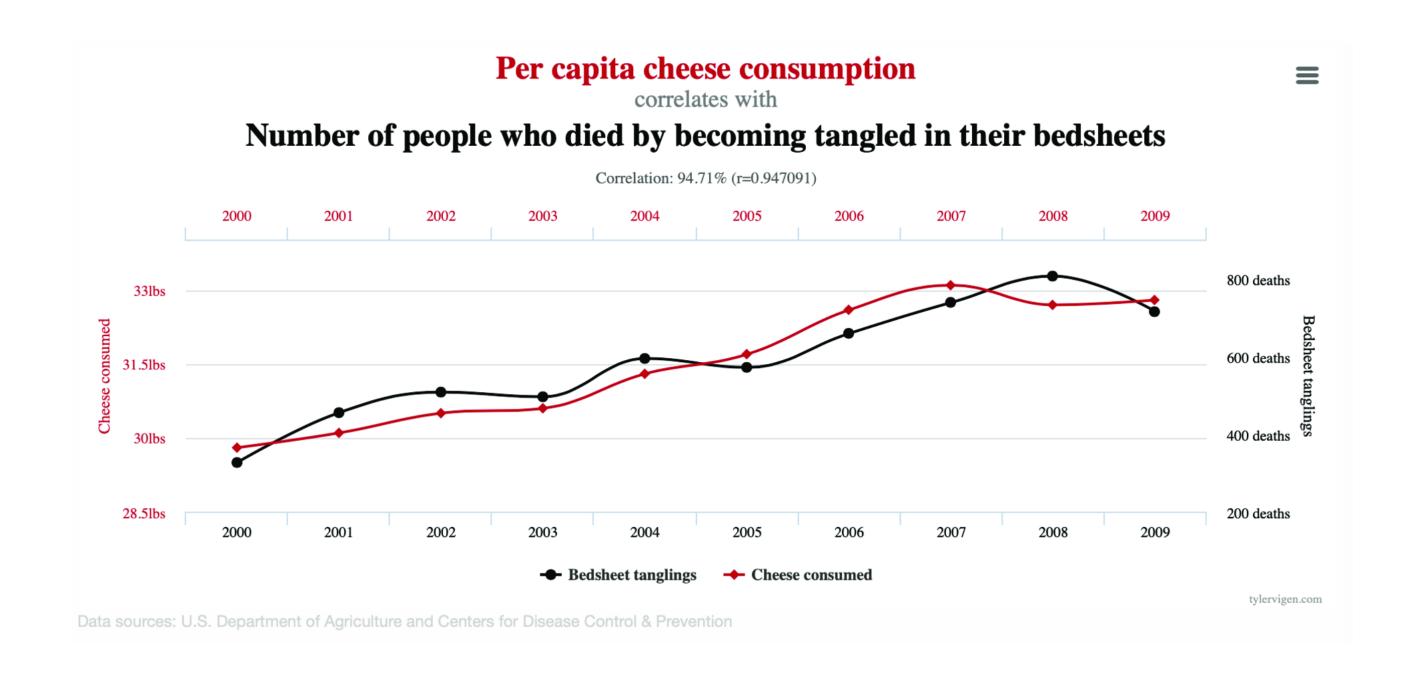
http://www.tylervigen.com/spurious-correlations



#### Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in Correlation: 66.6% (r=0.666004) 1999 2001 2002 2006 2007 2008 6 films Swimming pool drownings 120 drownings 2 films 100 drownings 80 drownings 0 films 2004 2003 2005 1999 2001 2002 2007 2008 2009 **→** Swimming pool drownings Nicholas Cage tylervigen.com Data sources: Centers for Disease Control & Prevention and Internet Movie Database





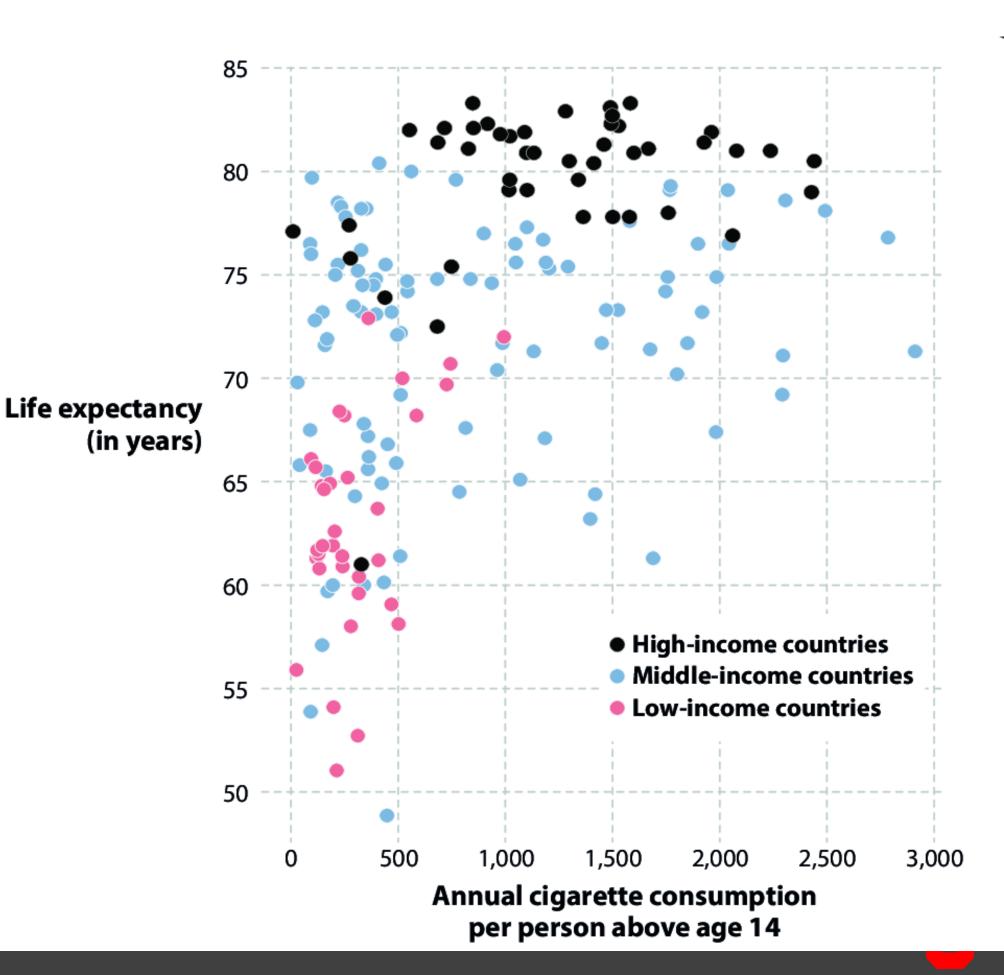




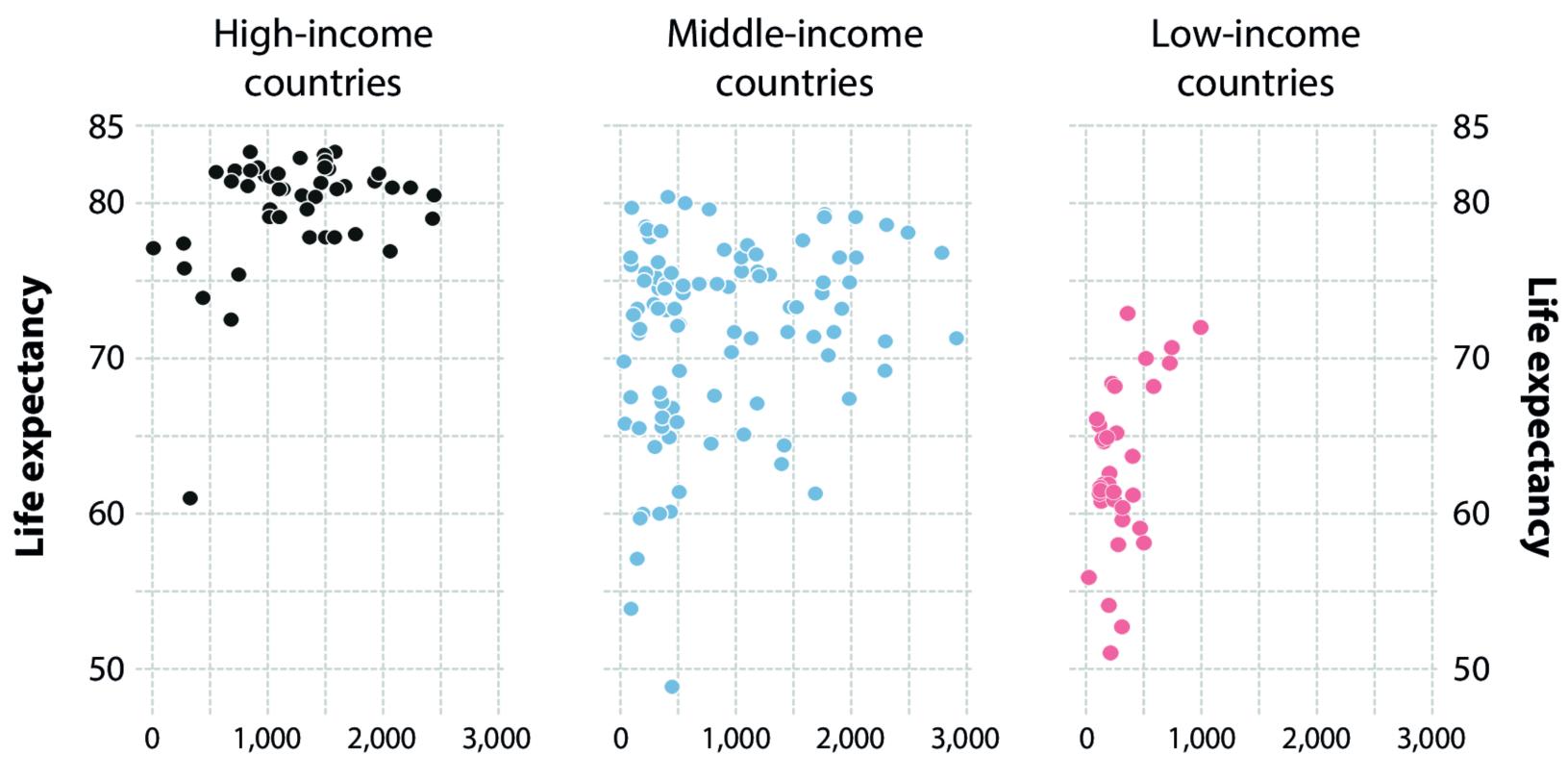


\* "The more cigarettes we consume, the longer we live!"

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

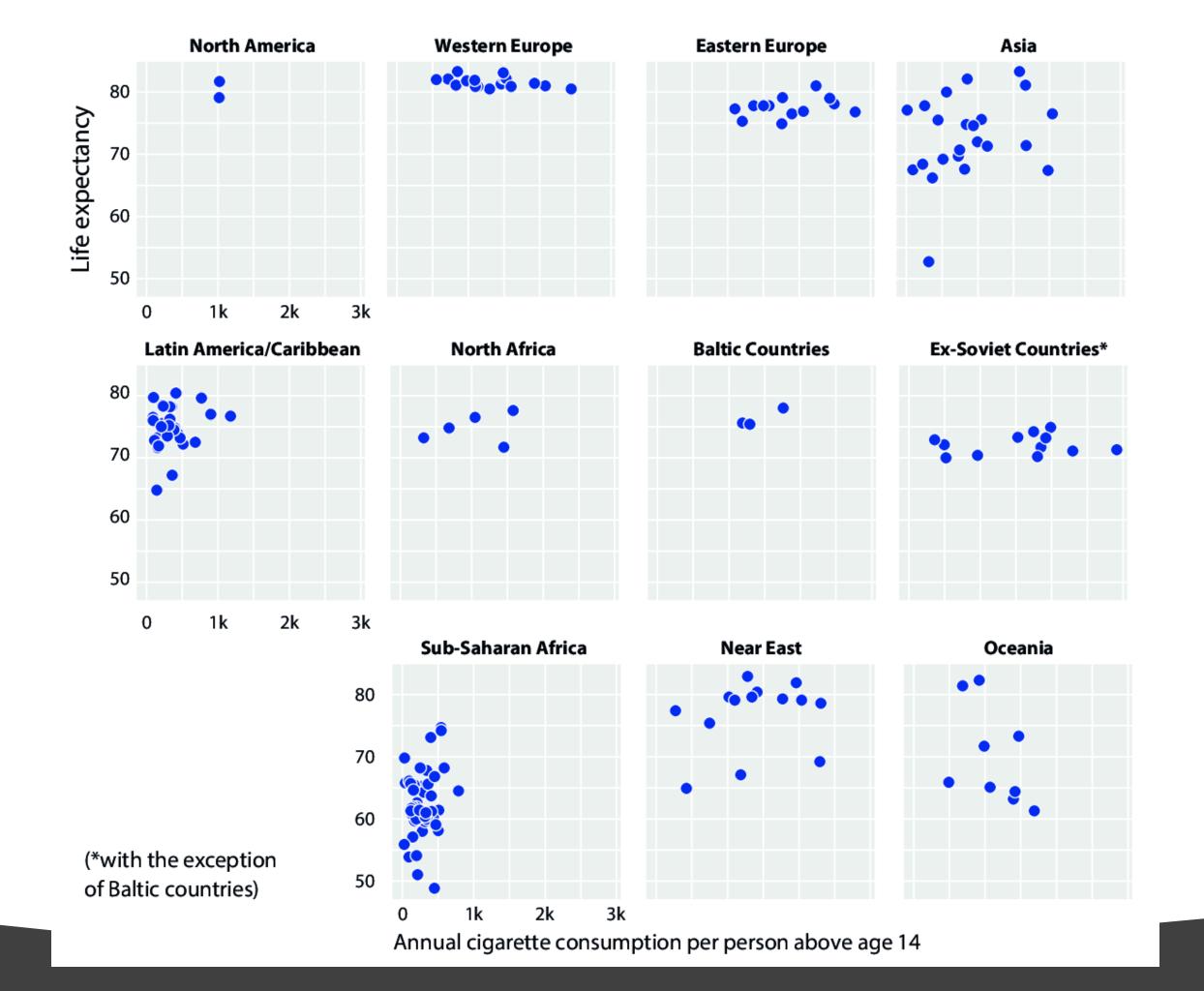






Annual cigarette consumption per person above age 14







## Simpson's Paradox

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

	Ме	n	Women	
	Applicants Admitted		Applicants	Admitted
Total	8442	44%	4321	35%

Damantmant	Ме	n	Women		
Department	Applicants	Admitted	Applicants	Admitted	
Α	825	62%	108	82%	
В	560	63%	25	68%	
С	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 2013

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%		



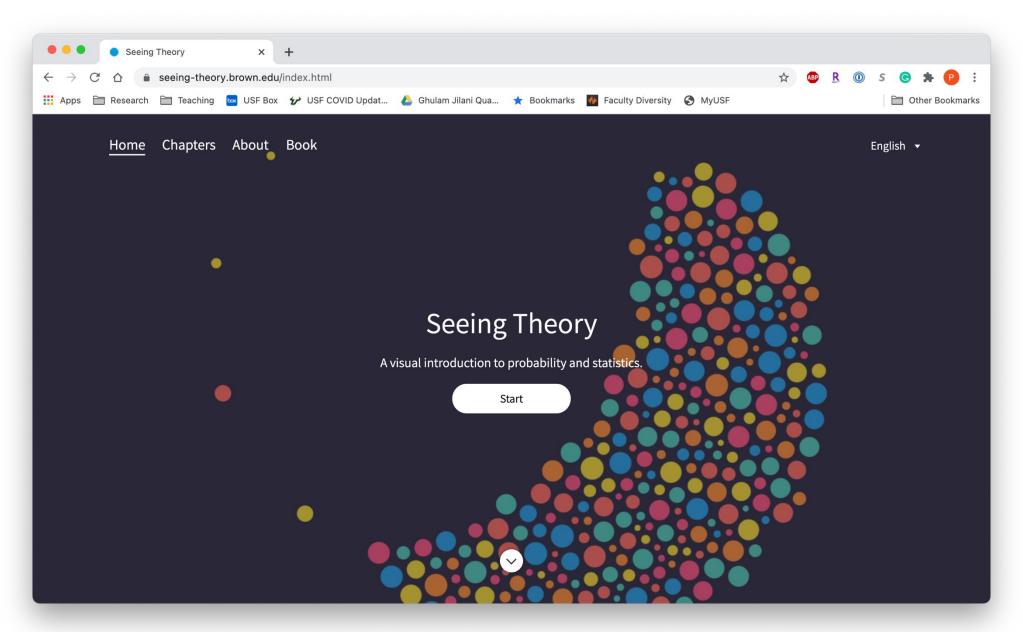
#### Can Every Group Be Worse Than Average? Yes.

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





