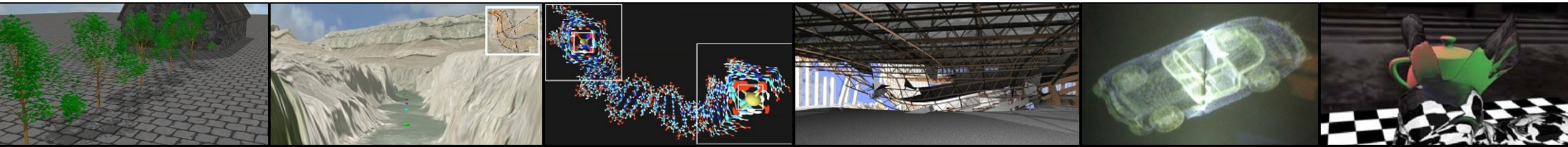
CIS 4930/6930-002
DATA VISUALIZATION

Descriptive Statistics and Visualization

Ghulam Jilani Quadri
University of South Florida

Slide credits D.A. Forsyth



REMINDERS

4/10/2018 – Project 7 Due

4/22/2018 – Paper 3 Review Due

4/15/2018 – Project 7 Peer Review Due

4/24/2018 – Project 8 Due

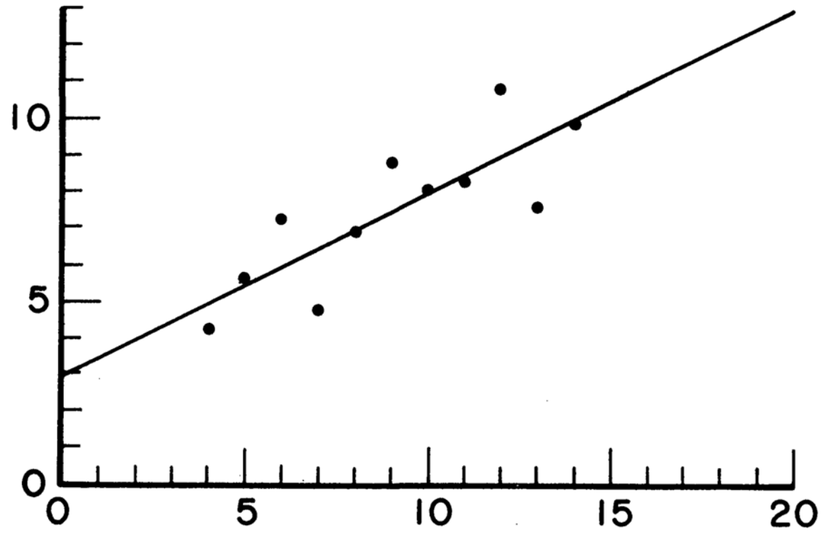
ANSCOMBE'S QUARTET

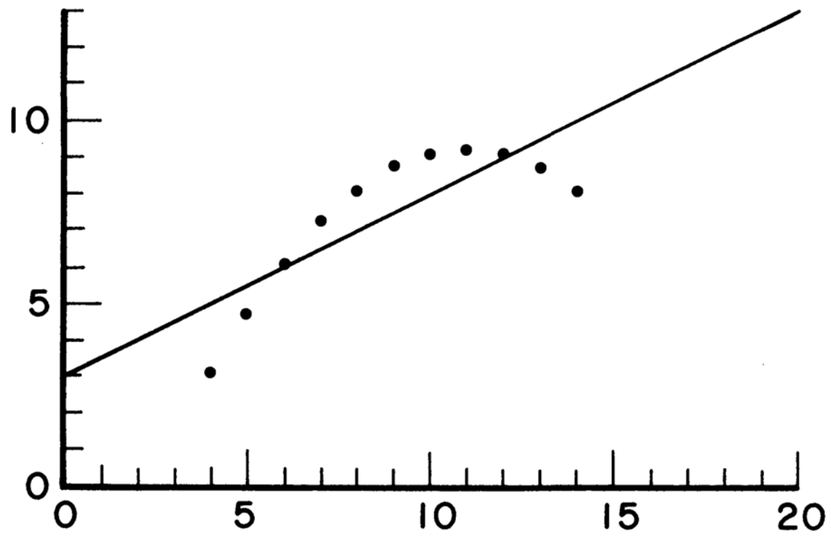
Data set		1-3	1	2	3		4	4
Variable		x	y	y	y		x	y
Obs. no.	1	10.0	8.04	9.14	7.46		8.0	6.58
	2	8.0	6.95	8.14	6.77		8.0	5.76
	3	13.0	7.58	8.74	12.74		8.0	7.71
	4	9.0	8.81	8.77	7.11		8.0	8.84
	5	11.0	8.33	9.26	7.81		8.0	8.47
	6	14.0	9.96	8.10	8.84		8.0	7.04
	7	6.0	7.24	6.13	6.08		8.0	5.25
	8	4.0	4.26	3.10	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

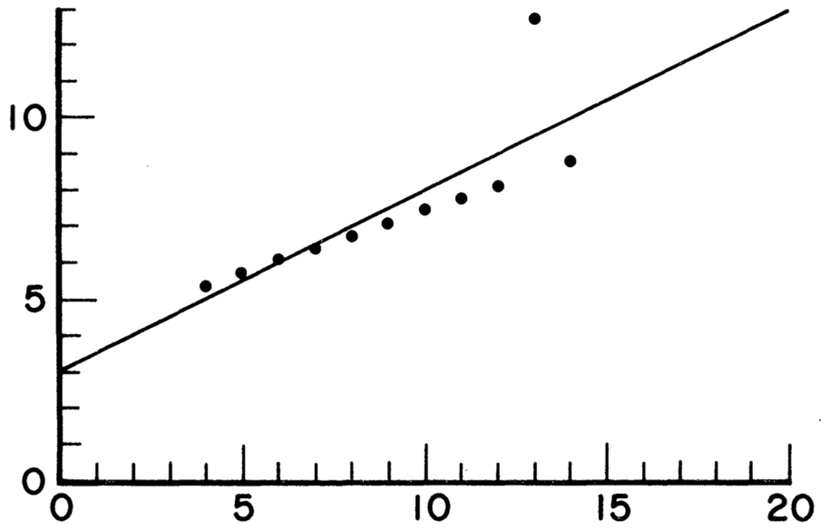
TABLE. Four data sets, each comprising 11 (x, y) pairs.

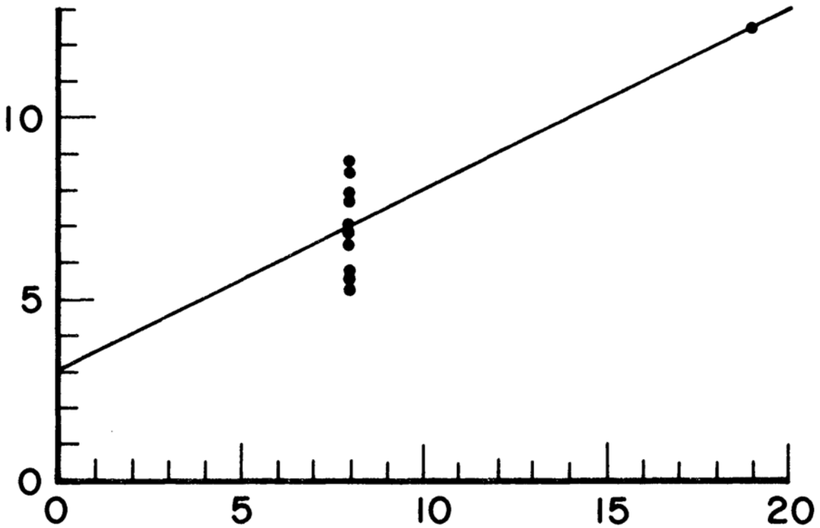
Number of observations (n) = 11
Mean of the x's ($\bar{x}$) = 9.0
Mean of the y's ($\bar{y}$) = 7.5
Regression coefficient (b_1) of y on x = 0.5
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x}$ = 110.0
Regression sum of squares = 27.50 (1 d.f.)
Residual sum of squares of y = 13.75 (9 d.f.)
Estimated standard error of b_1 = 0.118
Multiple R^2 = 0.667

GRAPHS IN STATISTICAL ANALYSIS. F. J. ANSCOMBE. THE AMERICAN STATISTICIAN, VOL. 27, NO. 1. (FEB., 1973), PP. 17-21.
HTTP://WWW.SJSU.EDU/FACULTY/GERSTMAN/STATPRIMER/ANSCOMBE1973.PDF


10
5
0
0
5
10
15
20


10
5
0
0
5
10
15
20


10
5
0
0
5
10
15
20


10
5
0
0
5
10
15
20

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

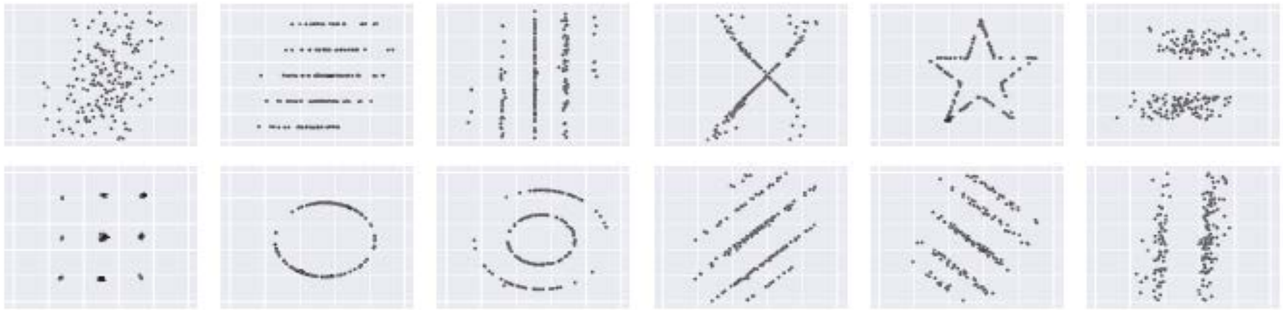
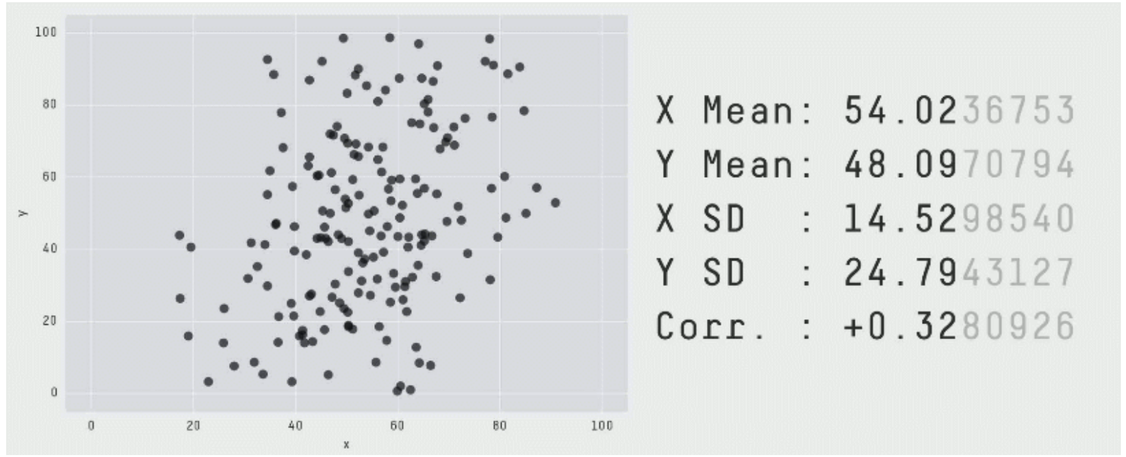


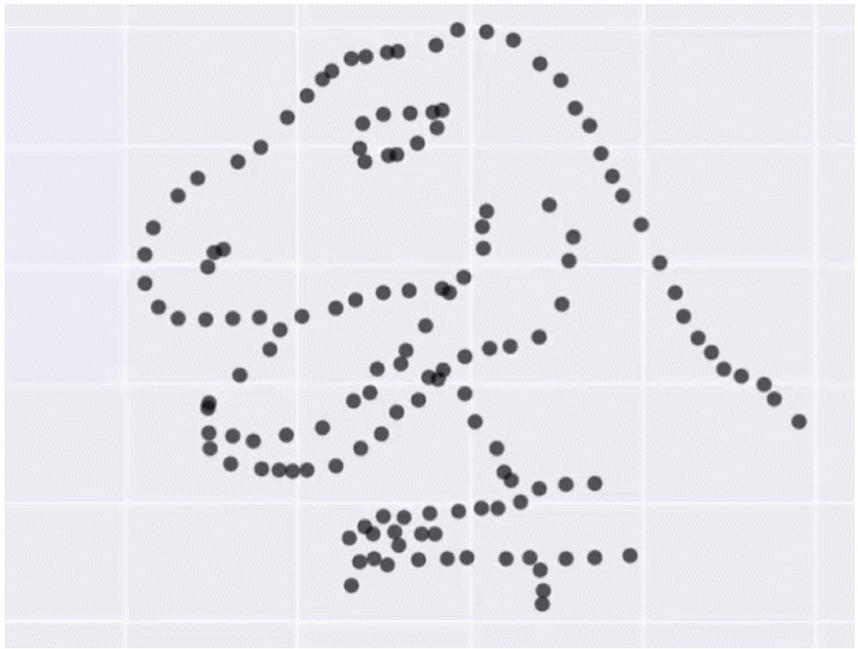
Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. ($\bar{x}$ =54.02, $\bar{y}$ = 48.09, sd_x = 14.52, sd_y = 24.79, Pearson's r = +0.32)

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.

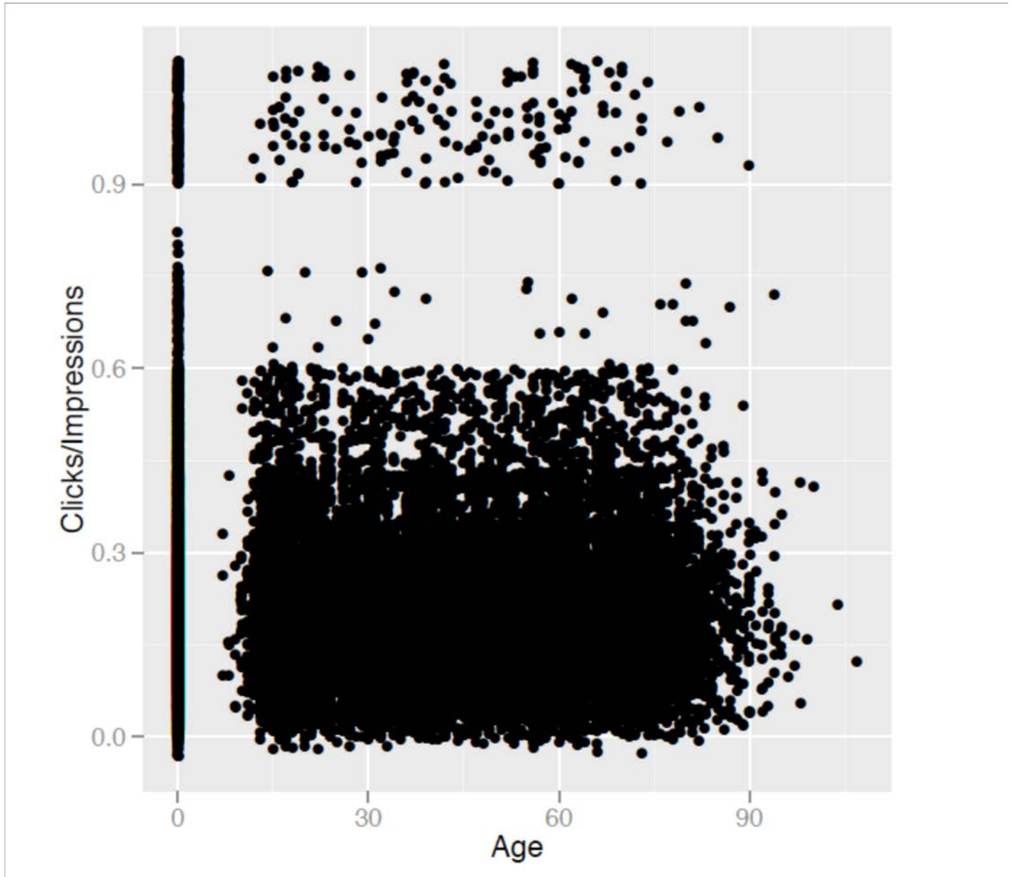
While very popular and effective for illustrating the





PROBLEM #1:

We have too many data points to show

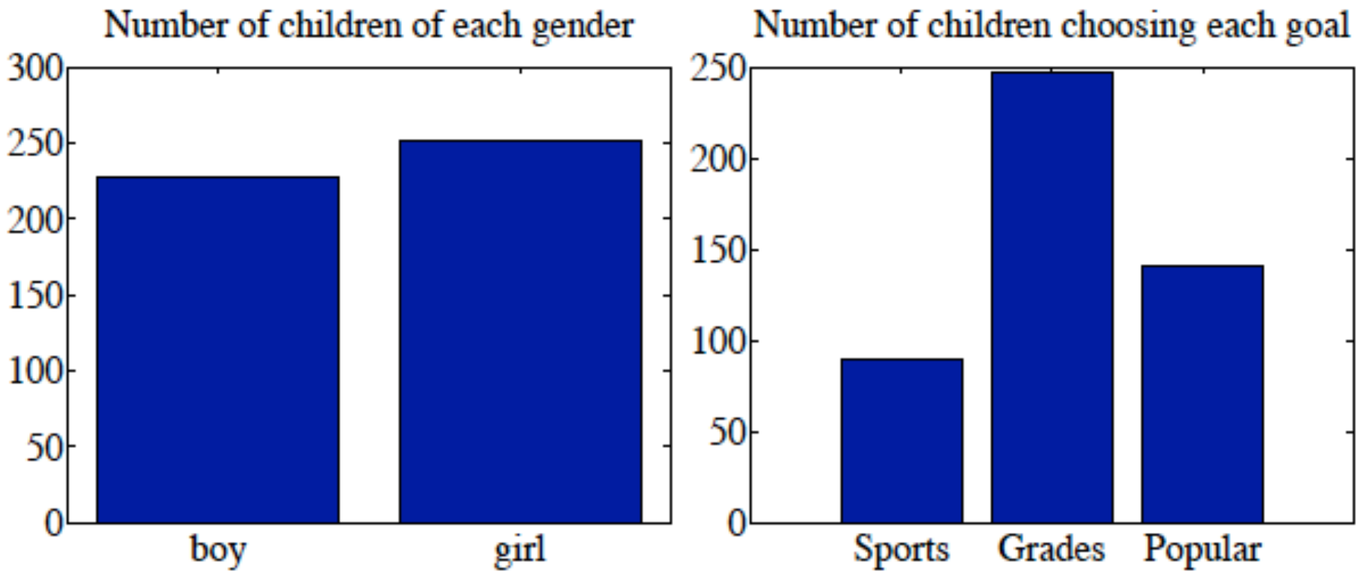


HISTOGRAMS

Bar chart-based visualization that allows evaluating distribution of values.

CATEGORICAL DATA

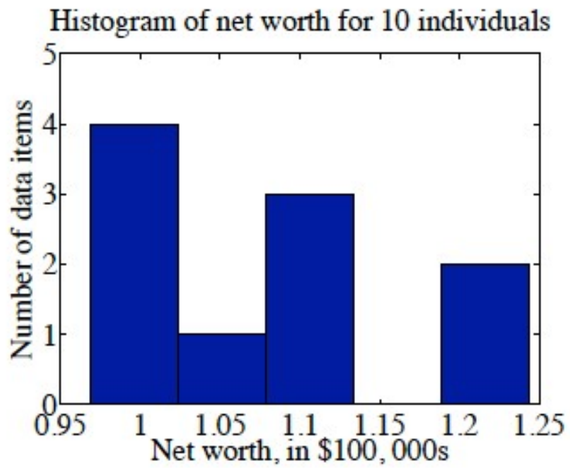
Gender	Goal	Gender	Goal
boy	Sports	girl	Sports
boy	Popular	girl	Grades
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	girl	Grades
girl	Popular	girl	Sports
girl	Grades	girl	Popular
girl	Sports	girl	Grades
girl	Sports	girl	Sports

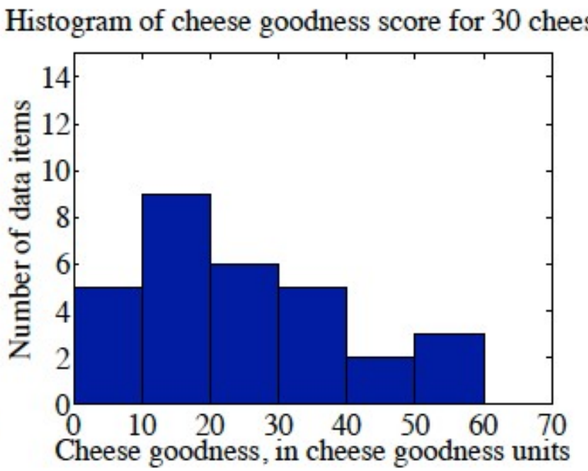


Continuous Data Histograms

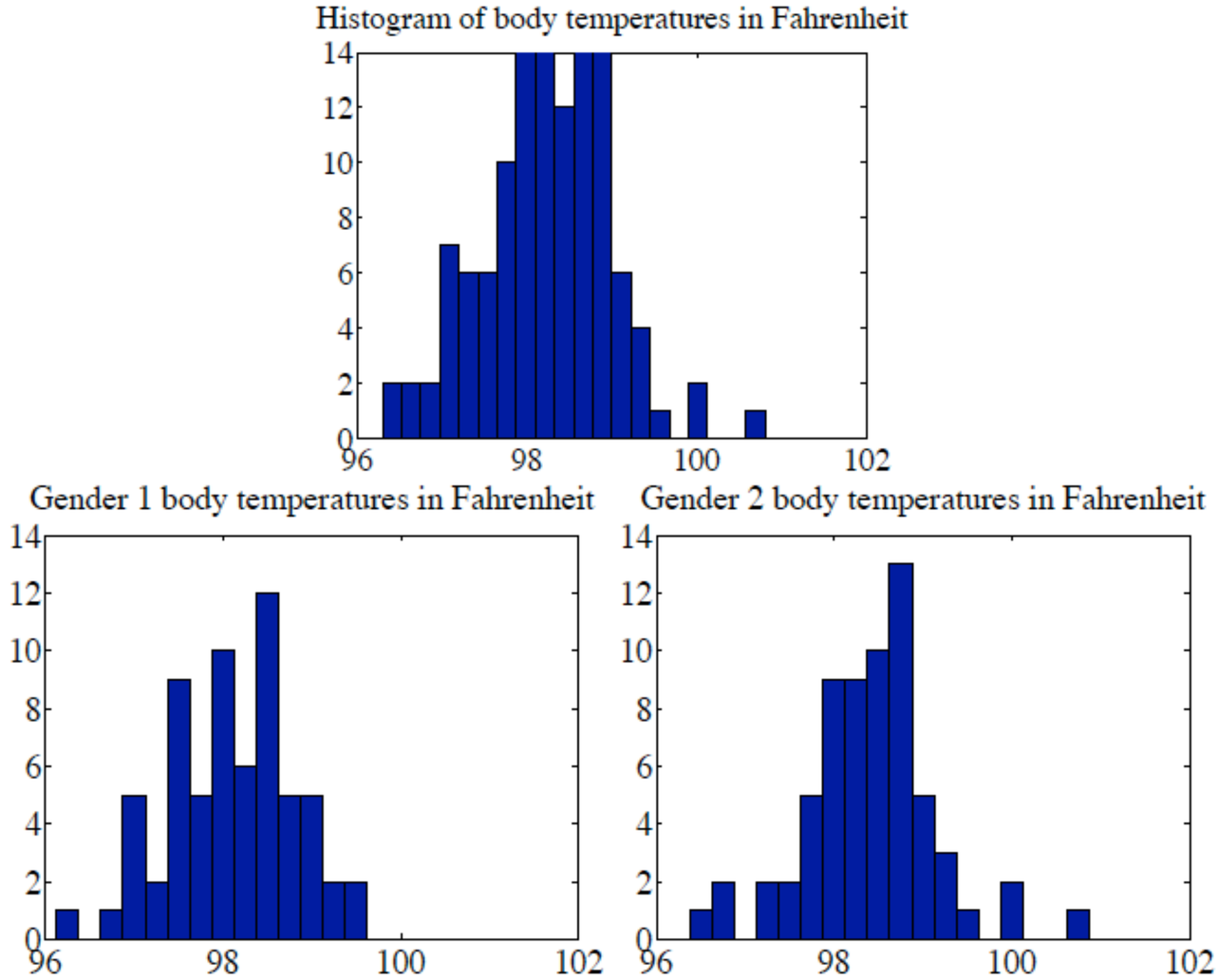
Index	net worth
1	100, 360
2	109, 770
3	96, 860
4	97, 860
5	108, 930
6	124, 330
7	101, 300
8	112, 710
9	106, 740
10	120, 170

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





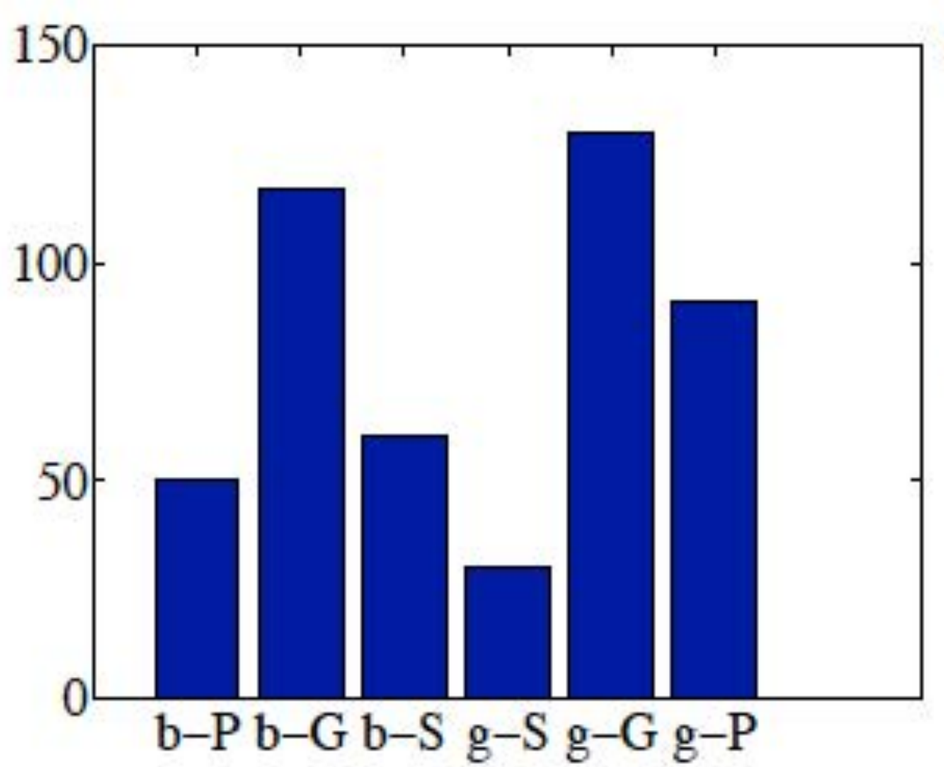
Conditional Histograms



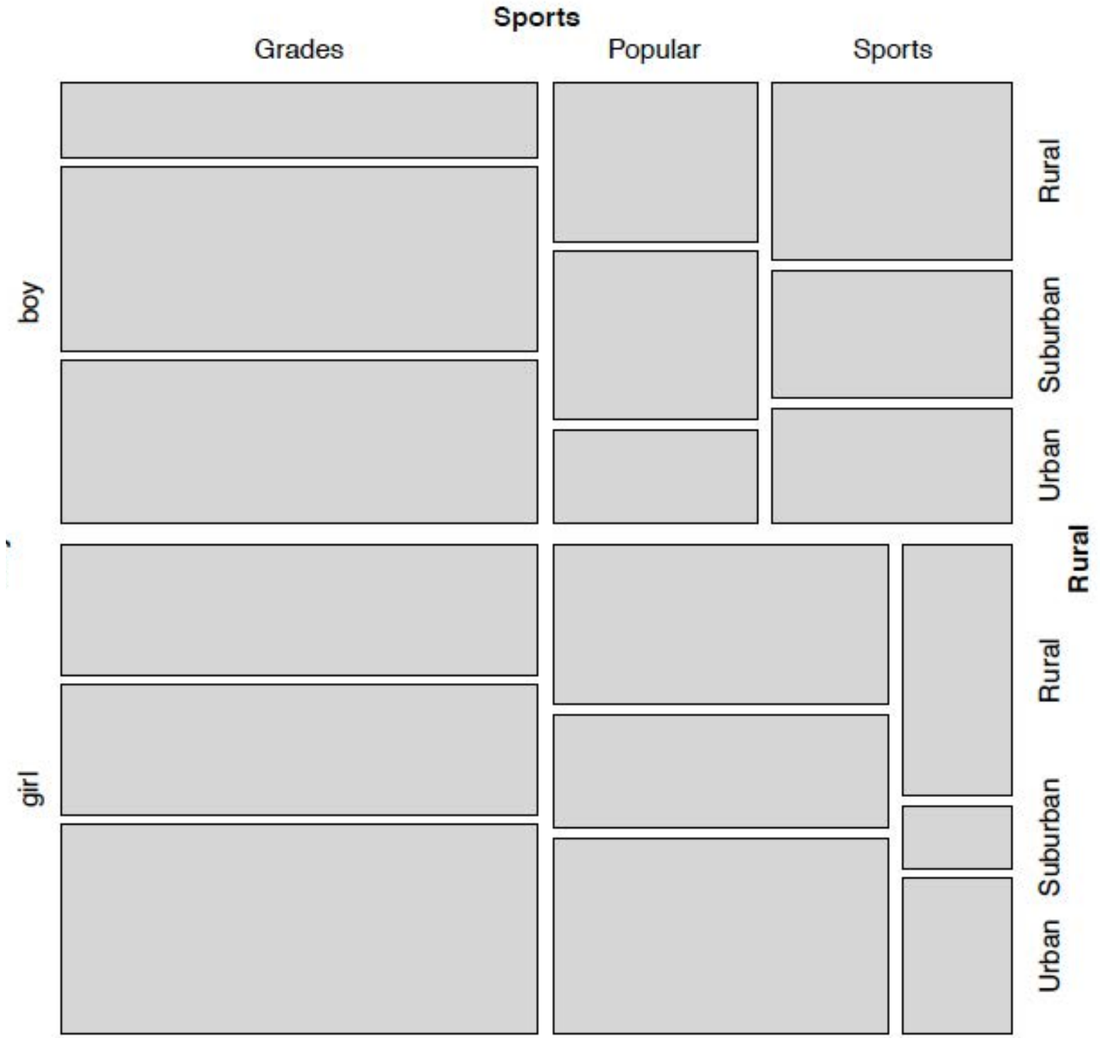
2D Data

CATEGORICAL DATA

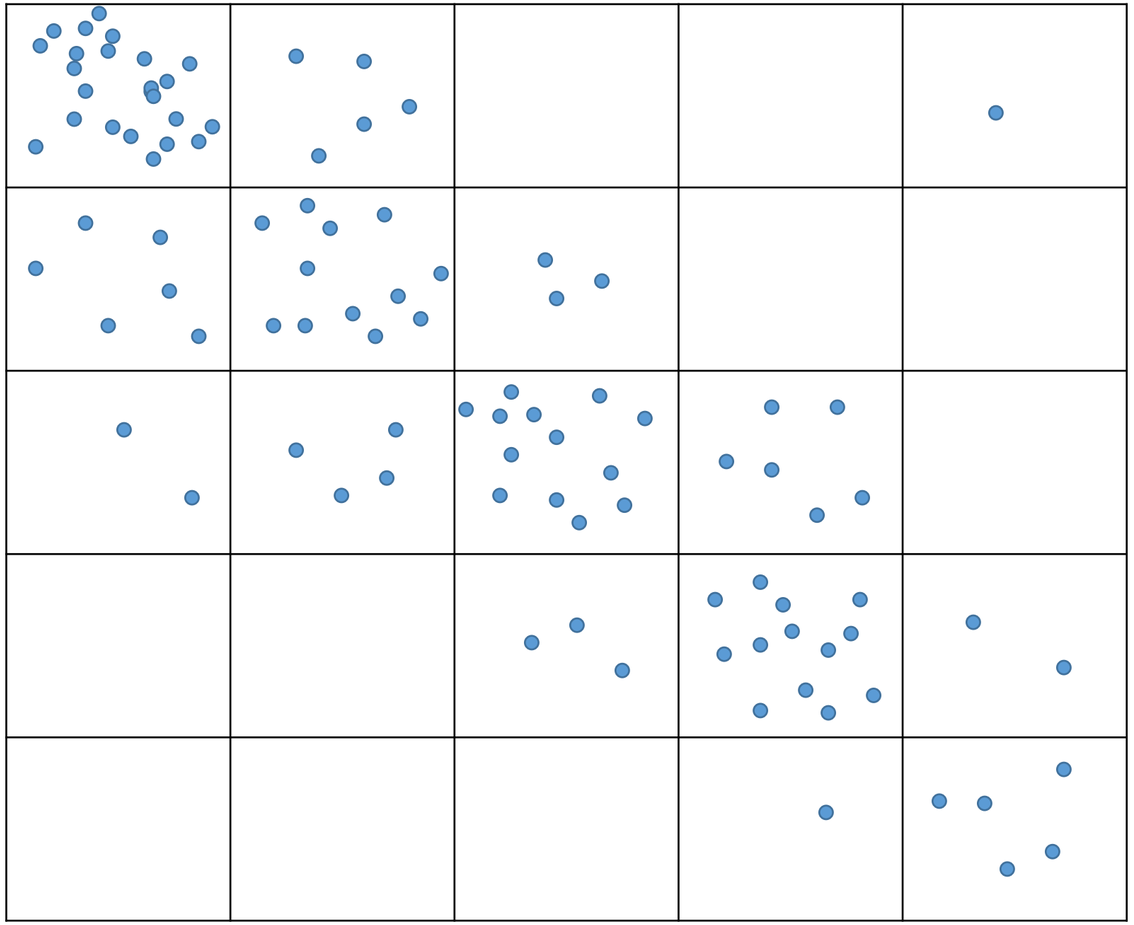
Gender	Goal	Gender	Goal
boy	Sports	girl	Sports
boy	Popular	girl	Grades
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	girl	Grades
girl	Popular	girl	Sports
girl	Grades	girl	Popular
girl	Sports	girl	Grades
girl	Sports	girl	Sports



Mosaic Plots



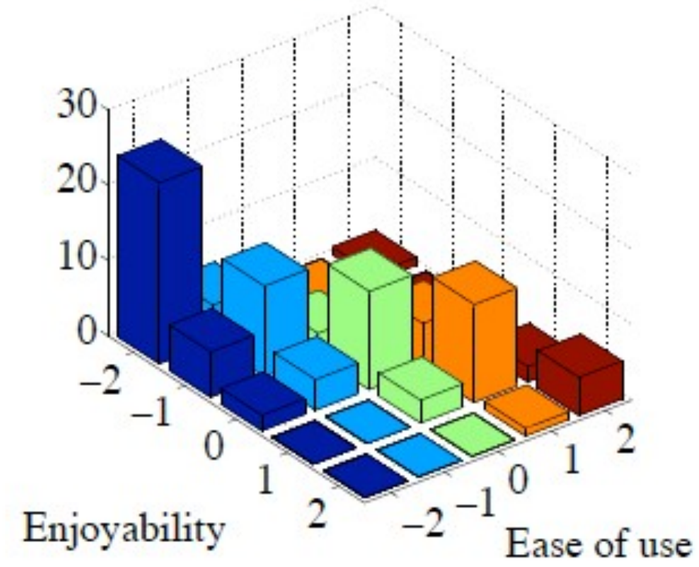
ORDINAL DATA

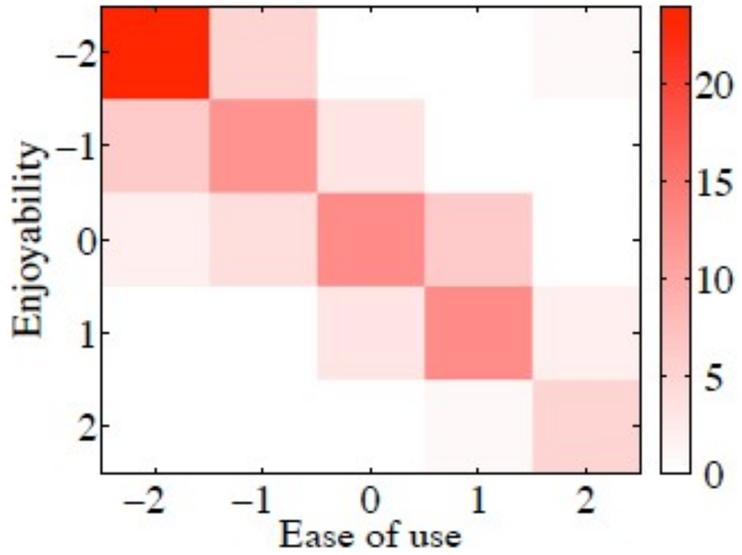


	-2	-1	0	1	2
-2	24	5	0	0	1
-1	6	12	3	0	0
0	2	4	13	6	0
1	0	0	3	13	2
2	0	0	0	1	5

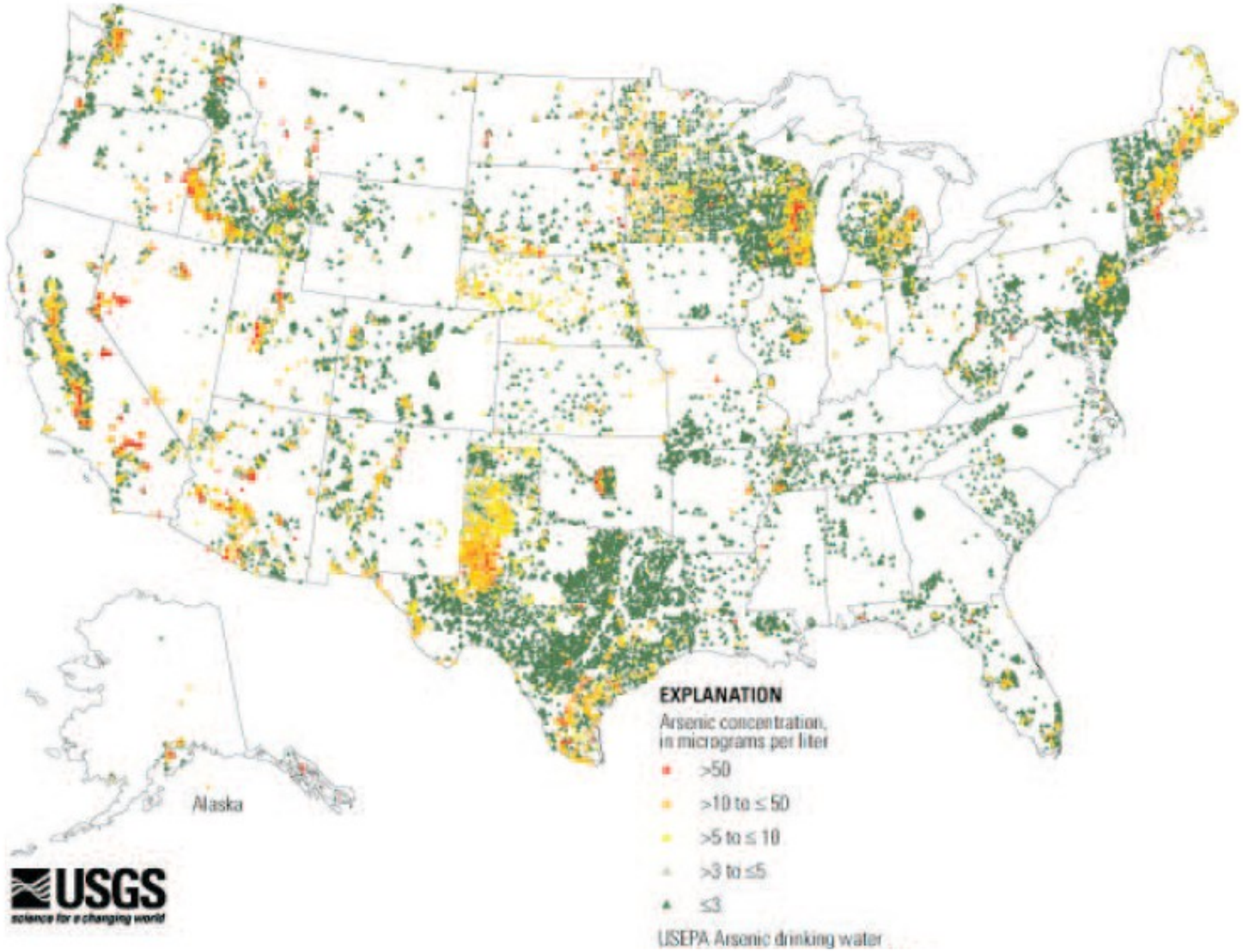
ORDINAL DATA

Counts of user responses for a user interface

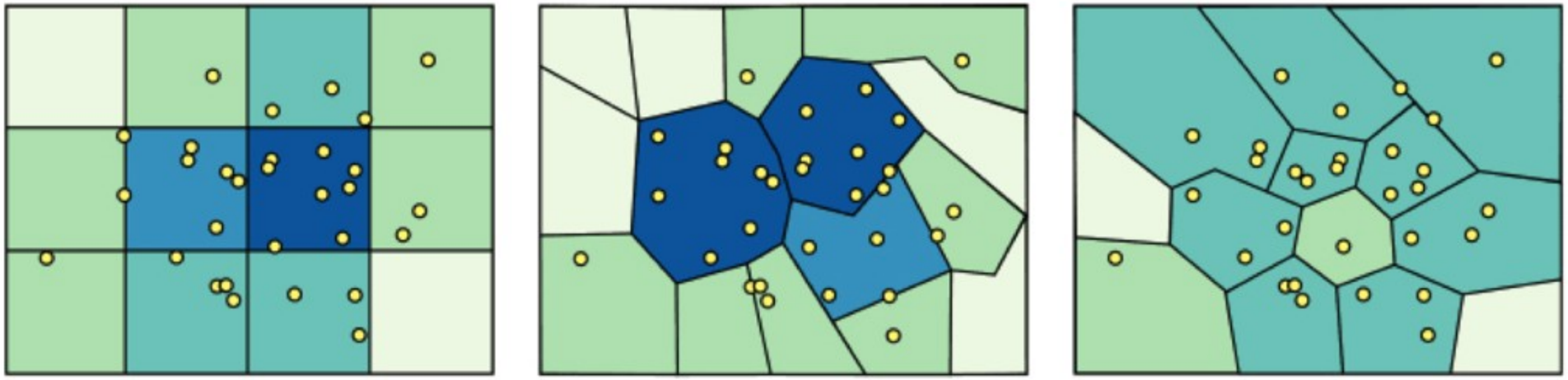




ARSENIC IN WELL WATER



SPATIAL AGGREGATION



MODIFIABLE AREAL UNIT PROBLEM

in cartography, changing the boundaries of the regions used to analyze data can yield dramatically different results

Modeling Data

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:

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

Think of this as a scale

Average distance from mean

STANDARD SCORE (AKA Z SCORE)

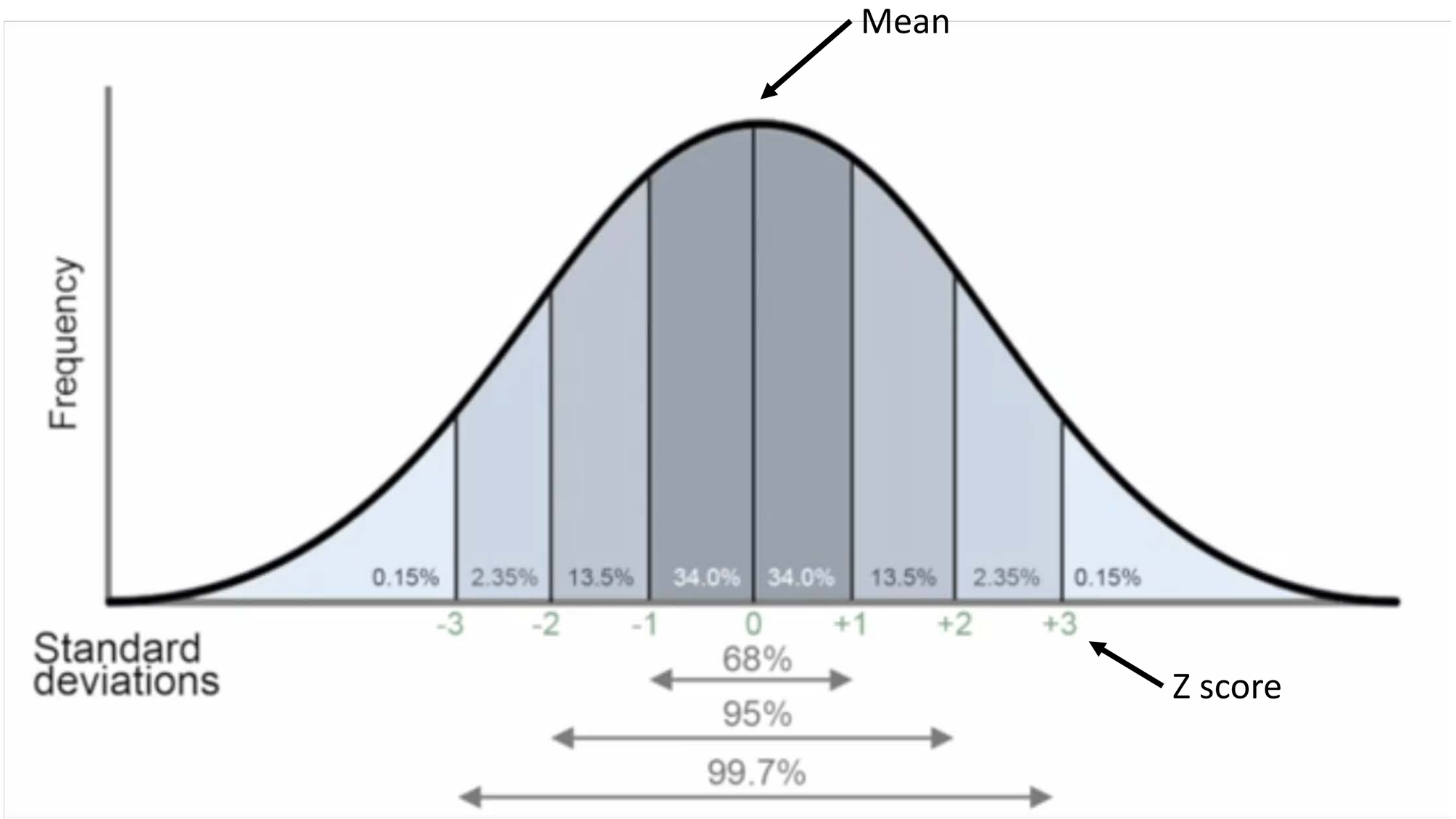
Definition: 3.8 *Standard coordinates*

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

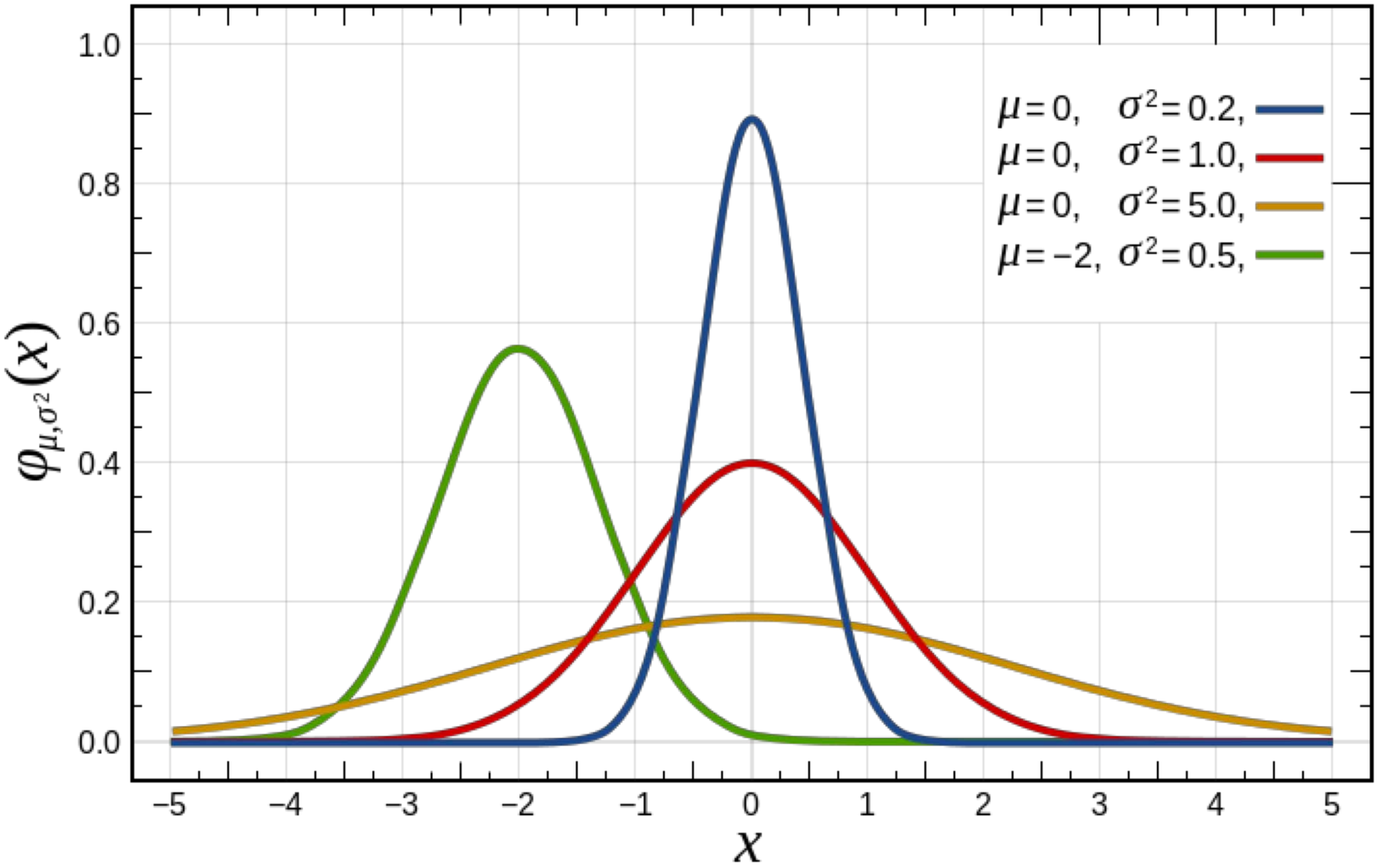
$$\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(x)}.$$

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

Number of standard deviations a point is away from mean

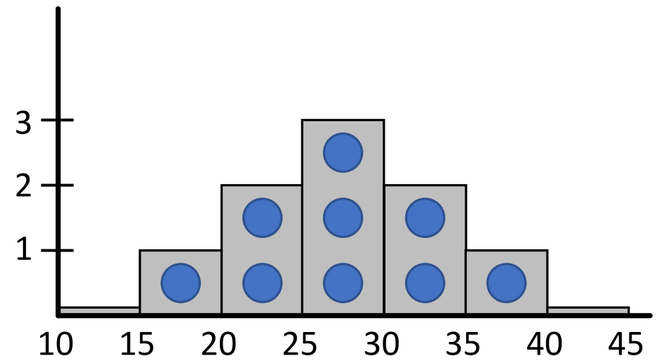

Mean
Frequency
0.15%
2.35%
13.5%
34.0%
34.0%
13.5%
2.35%
0.15%
-3
-2
-1
0
+1
+2
+3
Standard deviations
68%
95%
99.7%
Z score

Normal Distribution



An Example: Histogram

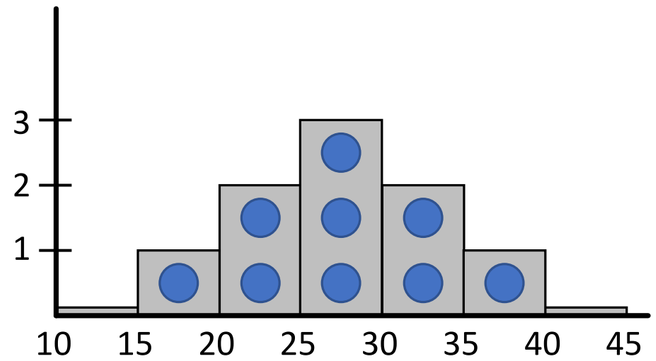
16
27
29
31
26
22
32
36
24

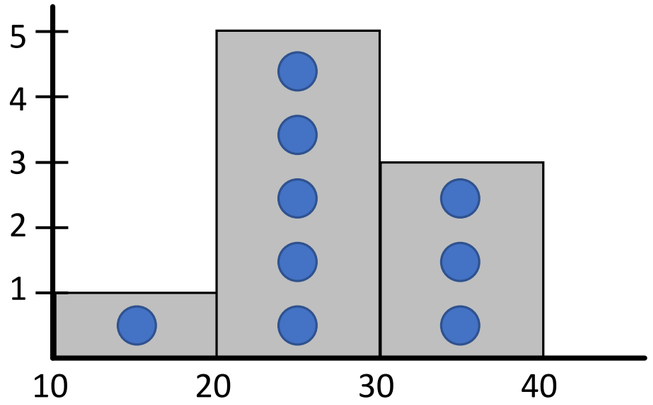


Mean (Average) = 27
Standard Deviation = 6

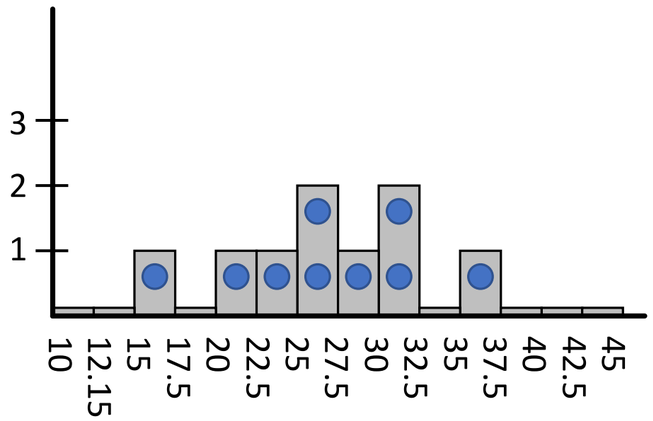
An Example: Histogram Resolution

16
27
29
31
26
22
32
36
24

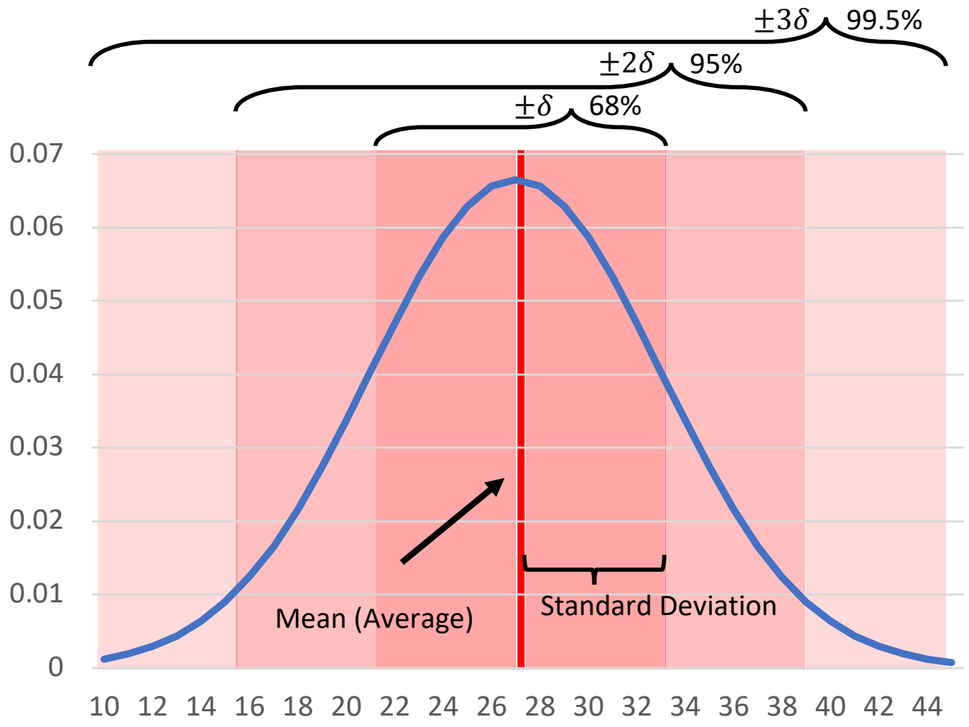


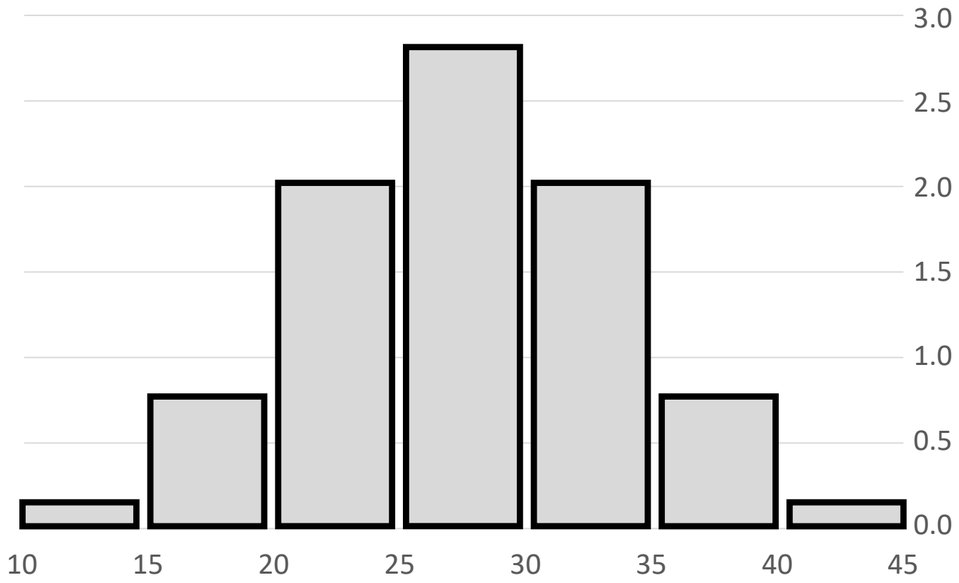


Mean (Average) = 27
Standard Deviation = 6

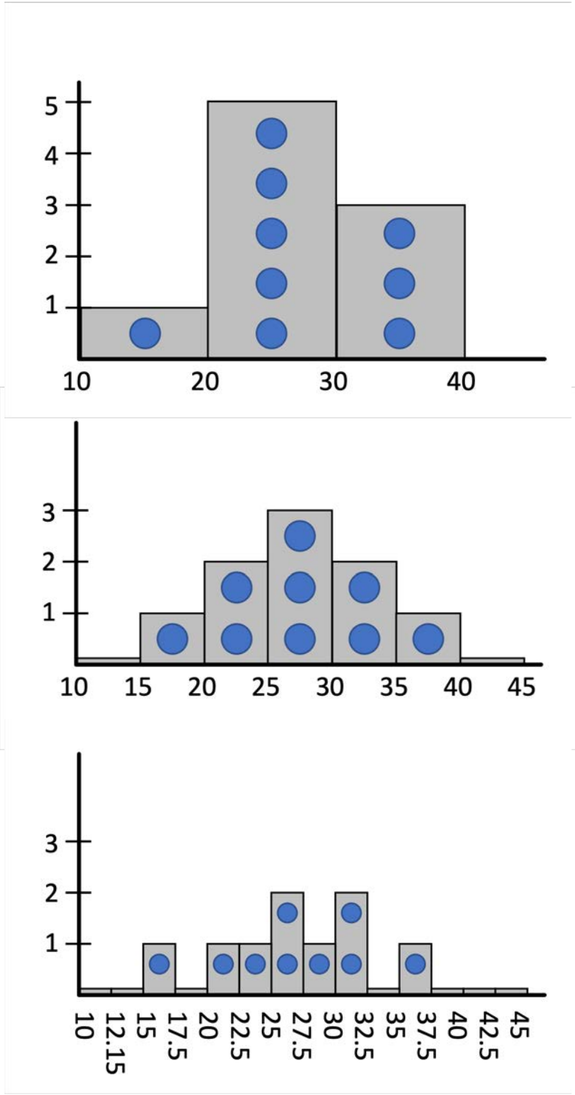


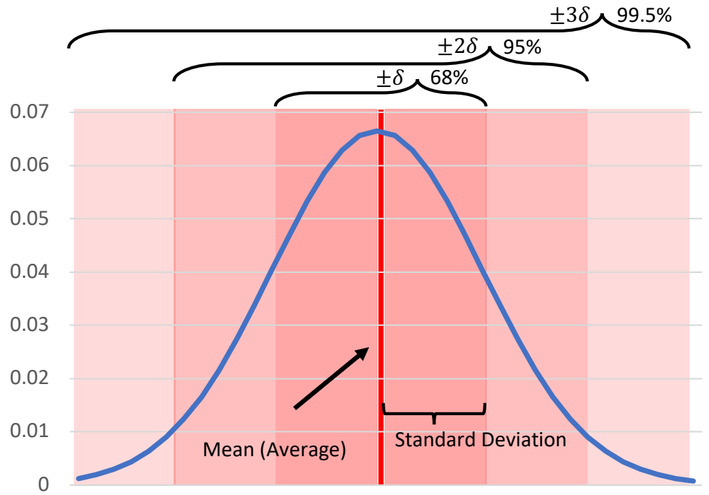
An Example: Statistical Distribution

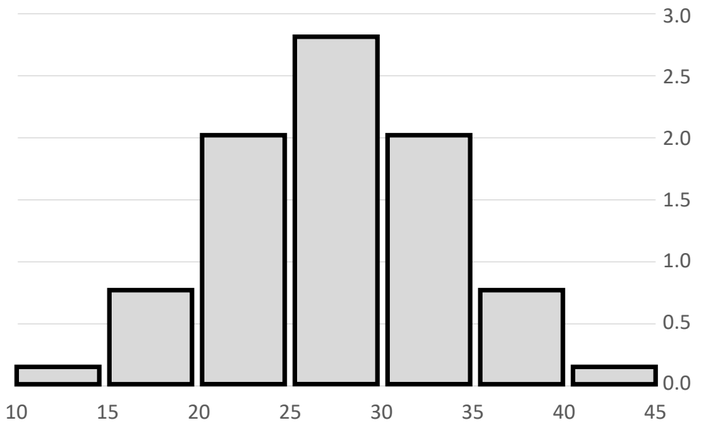




An Example: Comparing Histogram & Distribution

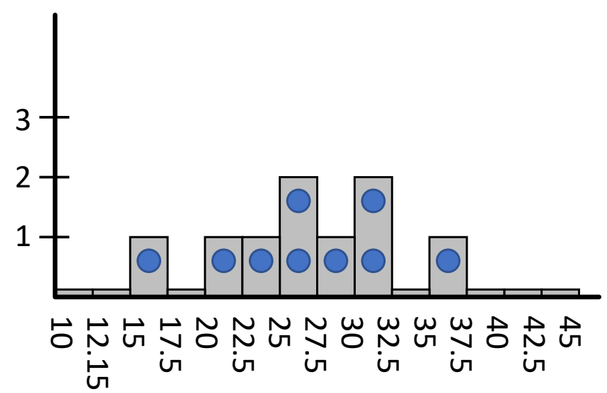


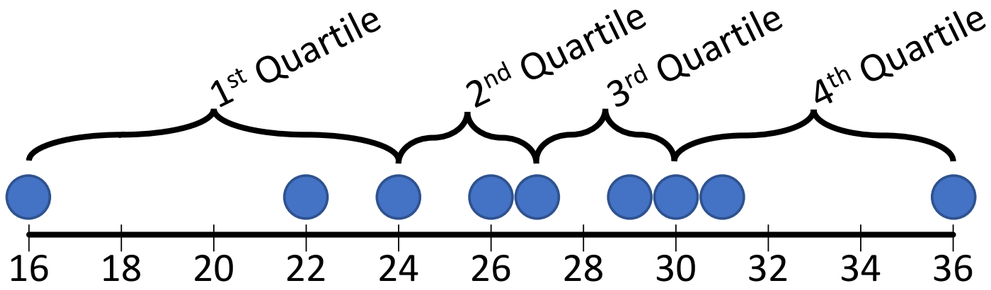




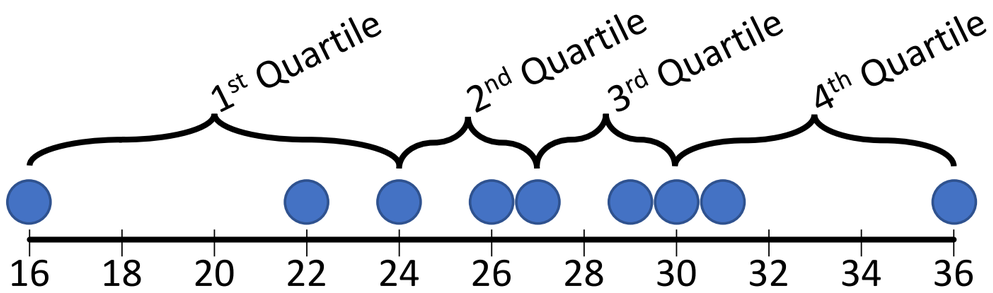
An Example: Comparing Histogram & Distribution

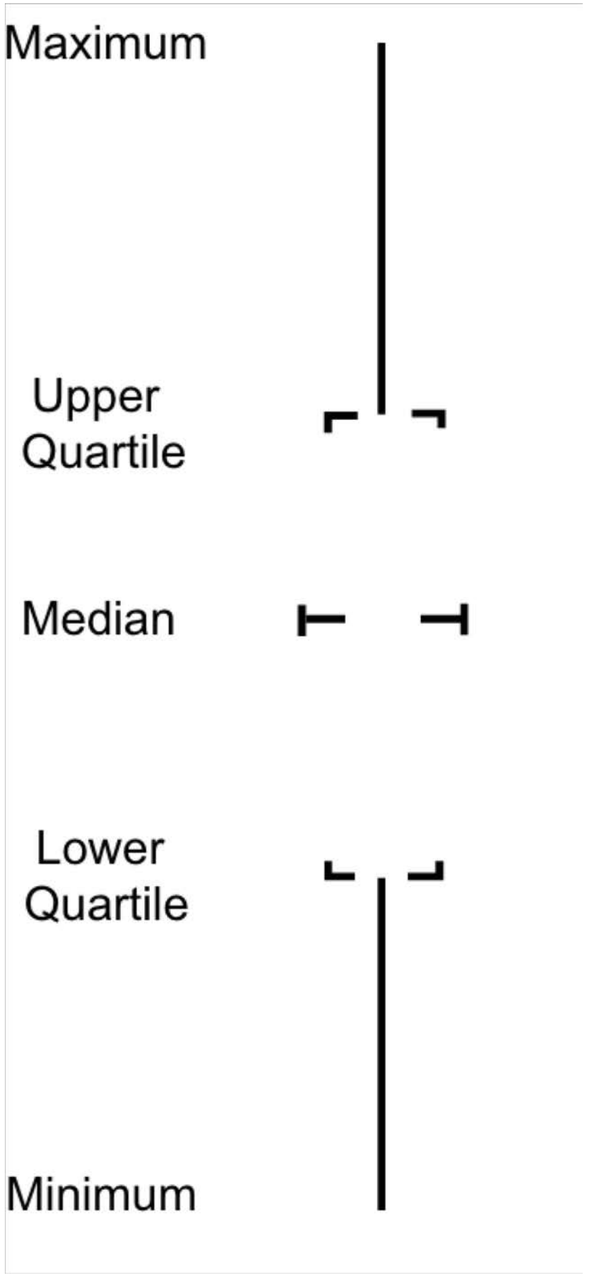
16
27
29
31
26
22
32
36
24





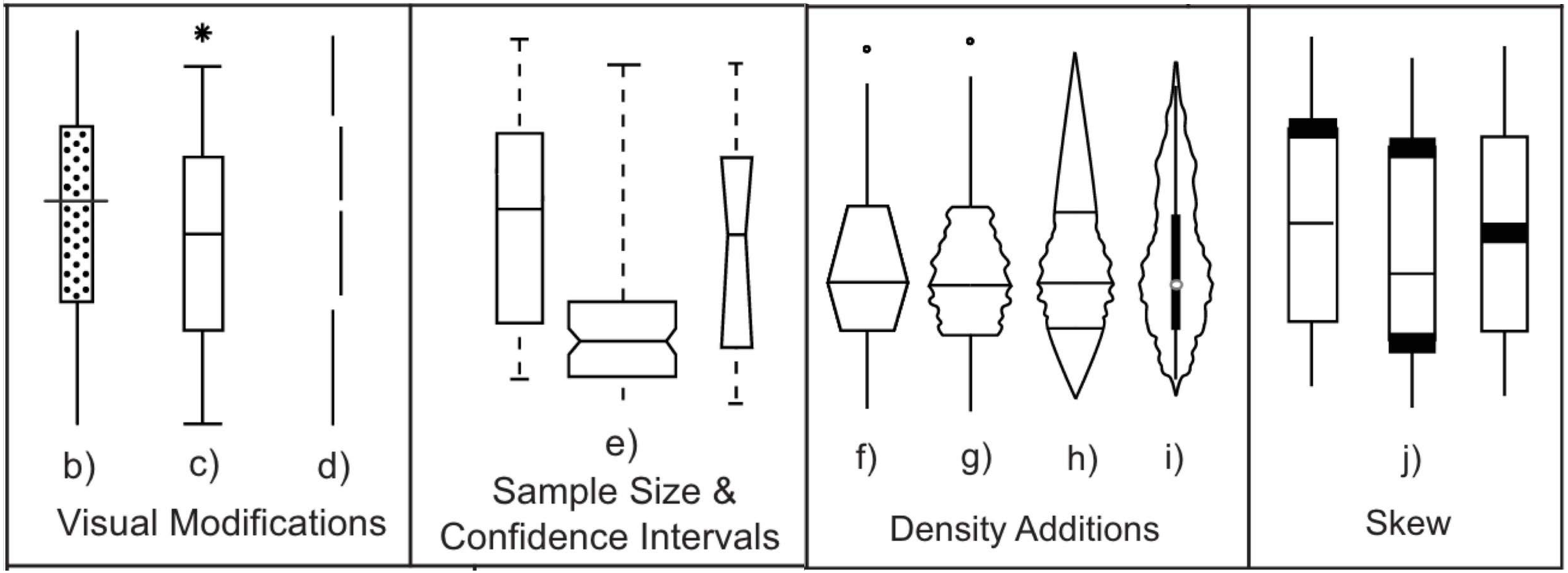
BOXPLOT





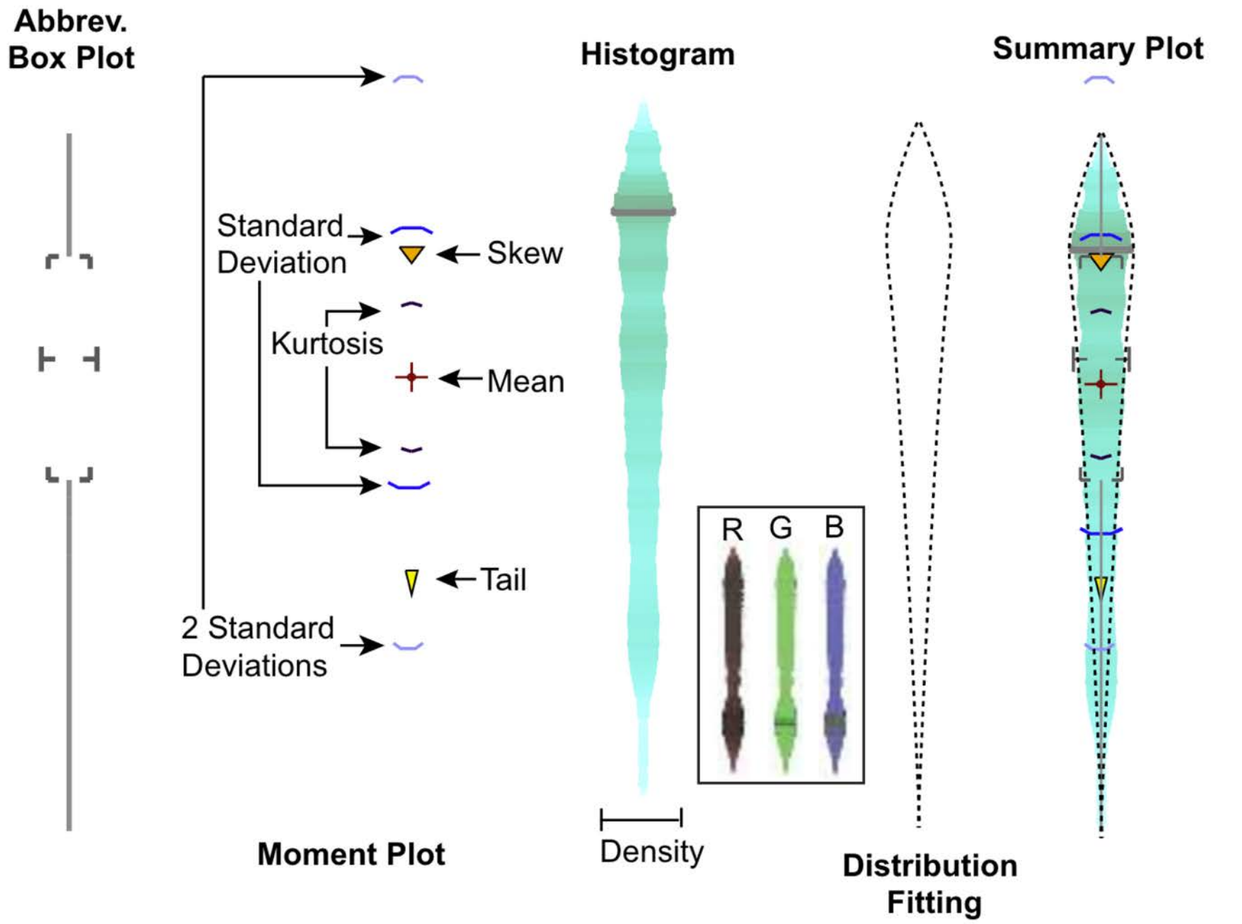
POTTER ET AL., 2010

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.

POTTER ET AL., 2010

PROBLEM #2:

We have too many attributes to show

Pearson Correlation Coefficient

A measure of the linearity between 2 sets

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

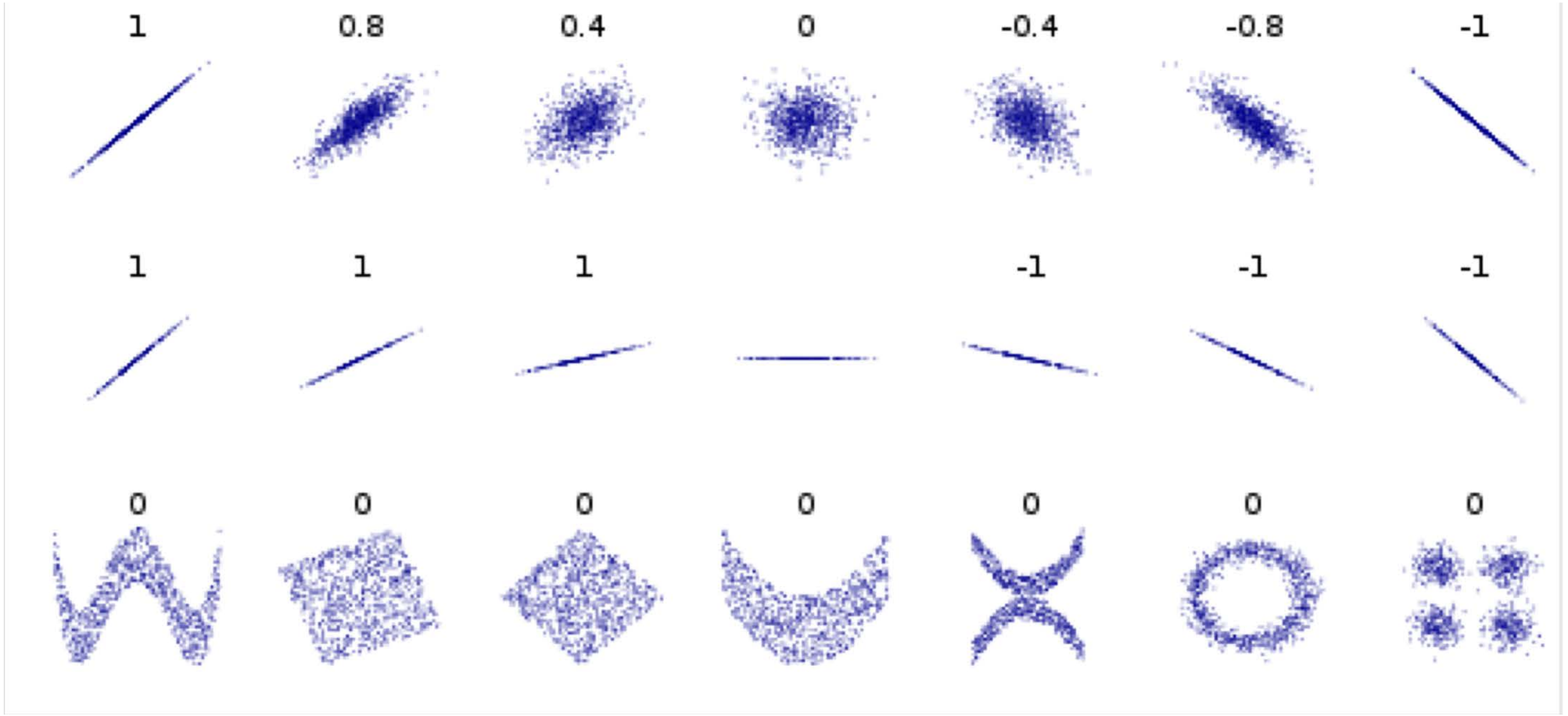
where:

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

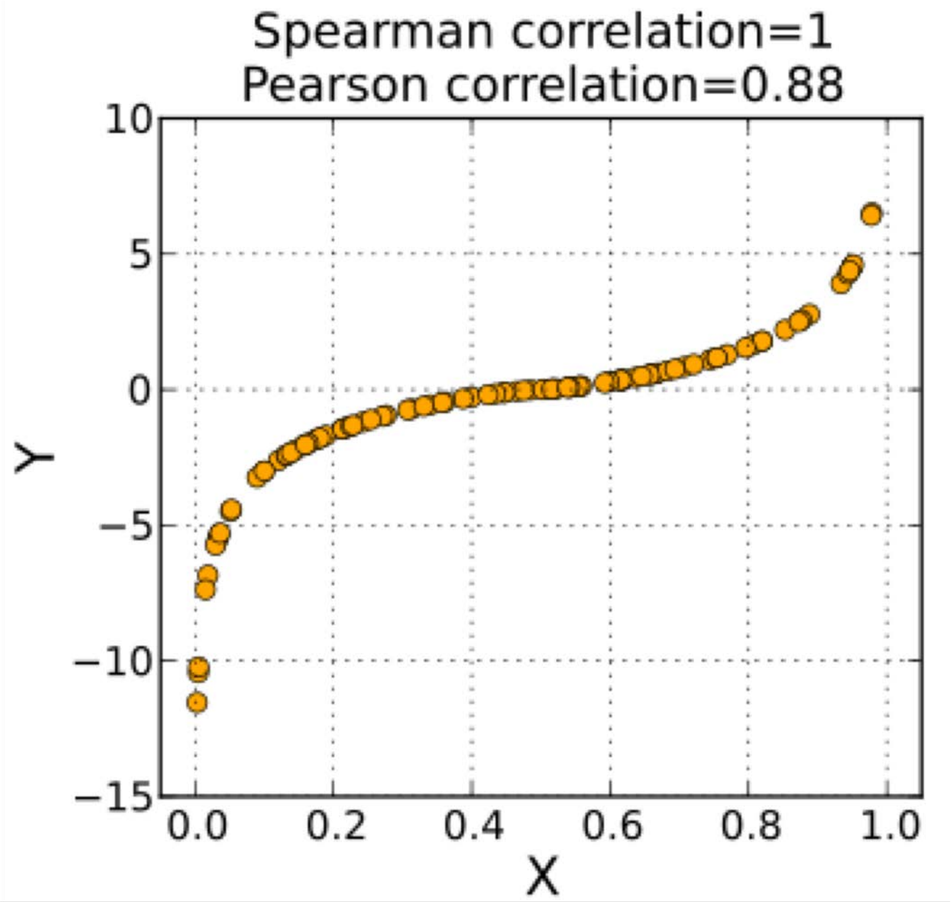
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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 $\bar{y}$


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

Spearman Rank Correlation



Spearman Rank Correlation

sort(X) and sort(Y)

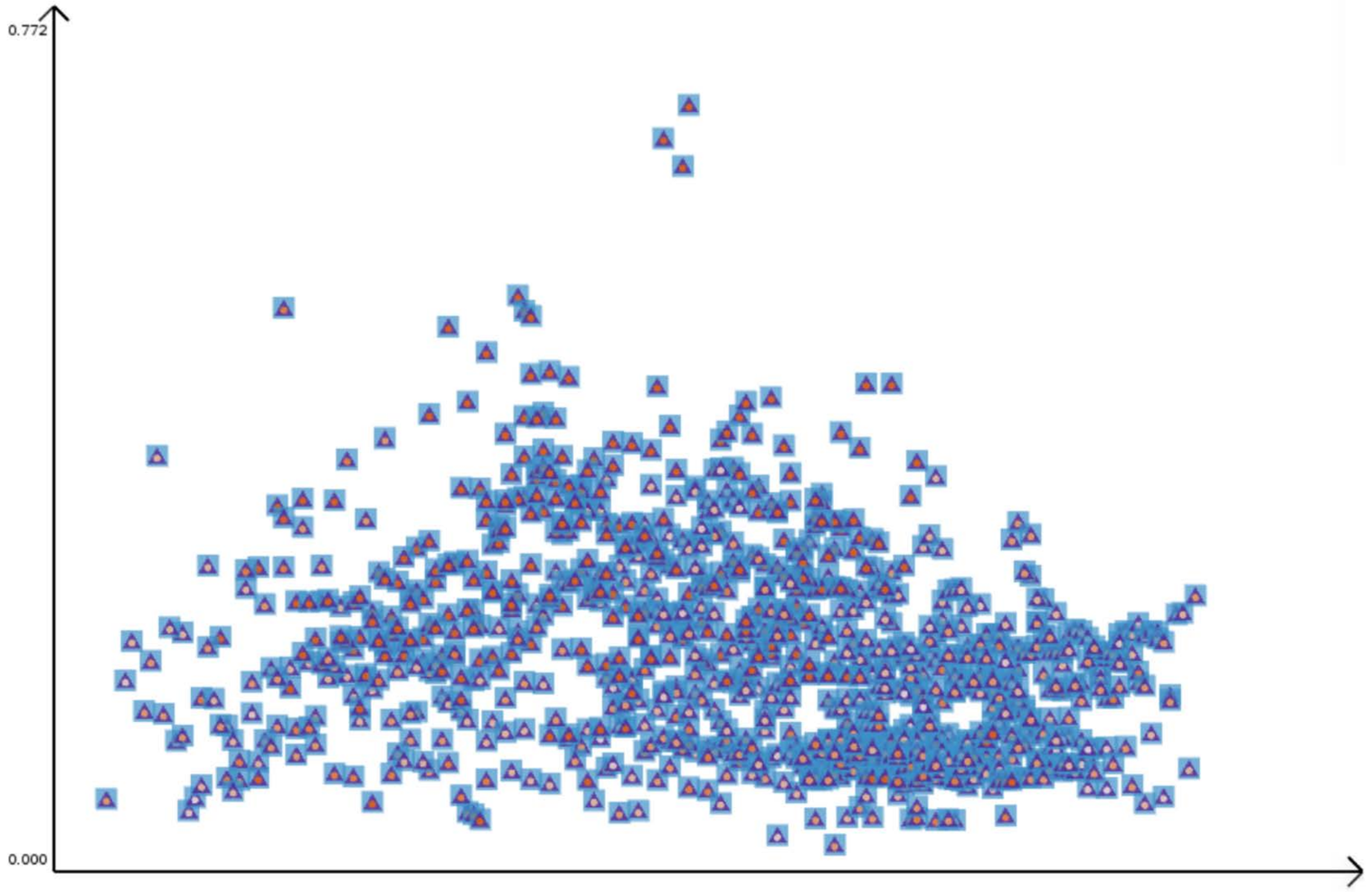
assign X'/Y' rank in sorted list

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

MANY ATTRIBUTES MULTIPLE CORRELATION



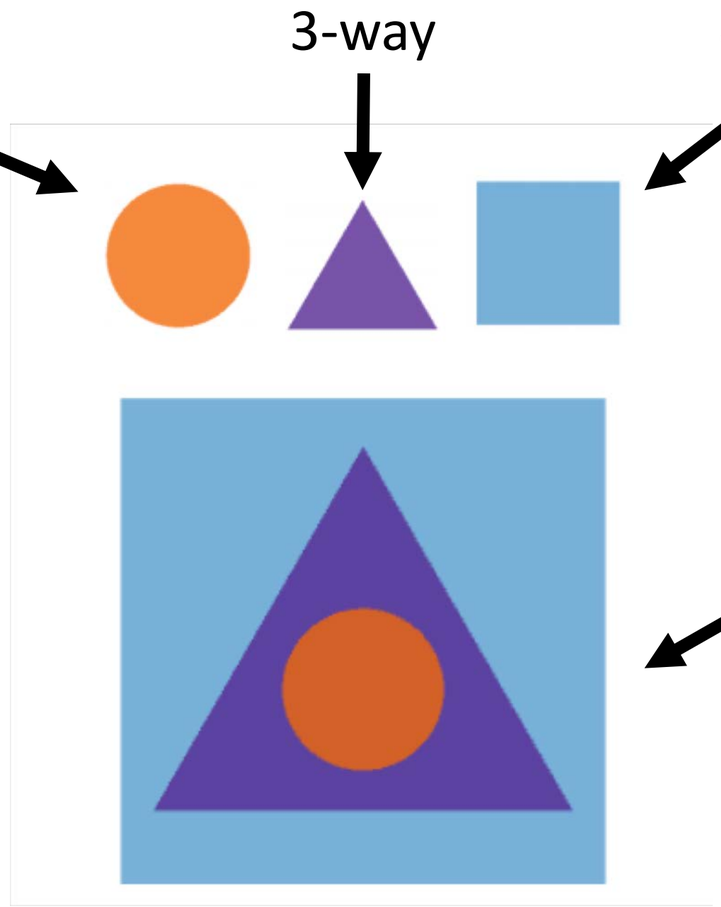
NGUYEN 2016

Multiple Correlation

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

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

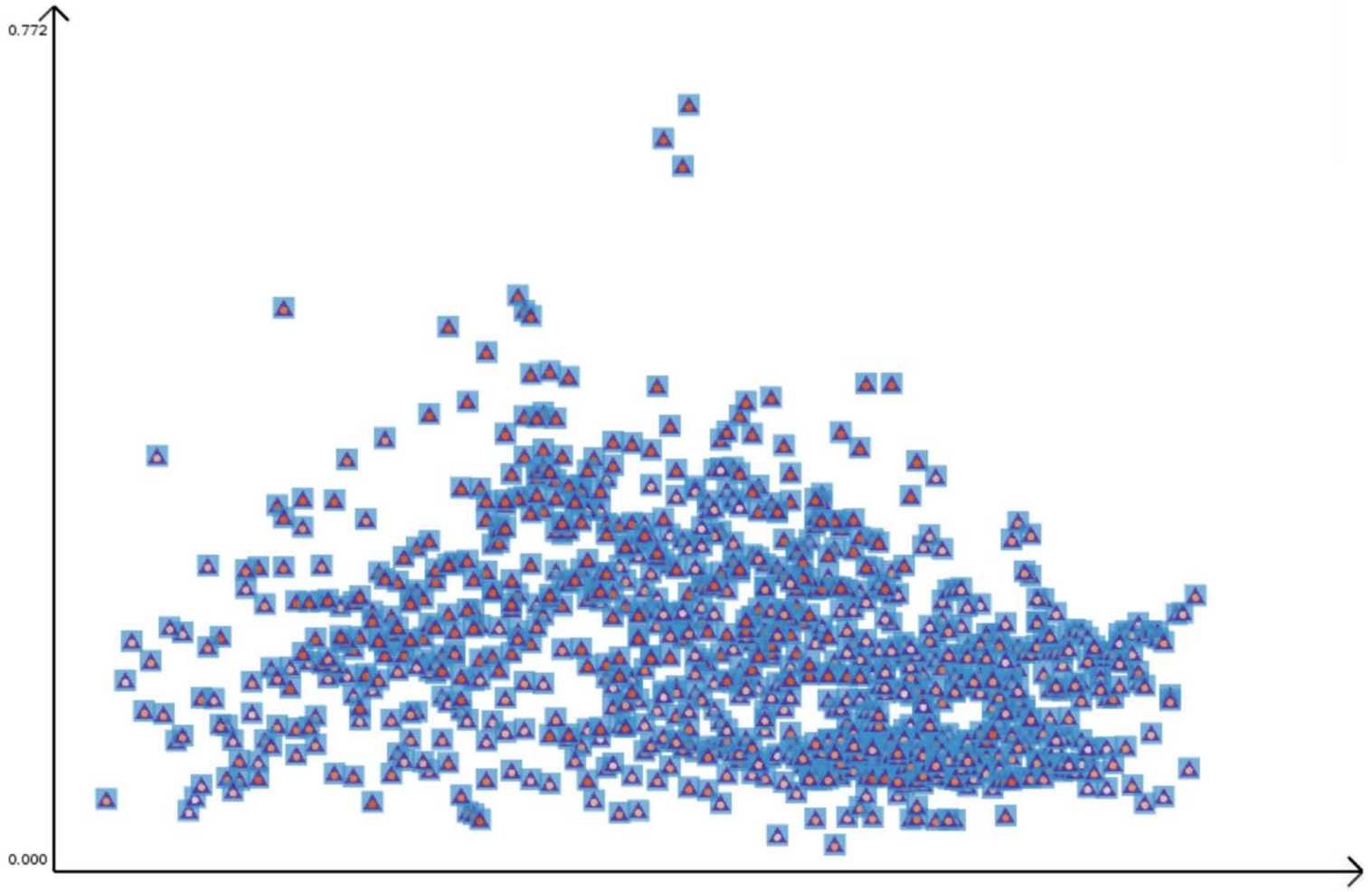
MULTIPLE CORRELATION



Composite Glyph

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

MANY ATTRIBUTES MULTIPLE CORRELATION

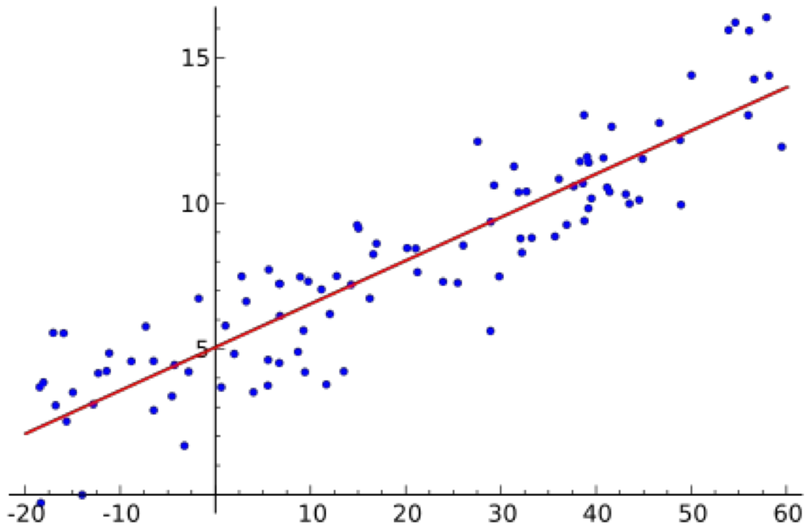


NGUYEN 2016

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. $\Sigma \varepsilon_i^2$)

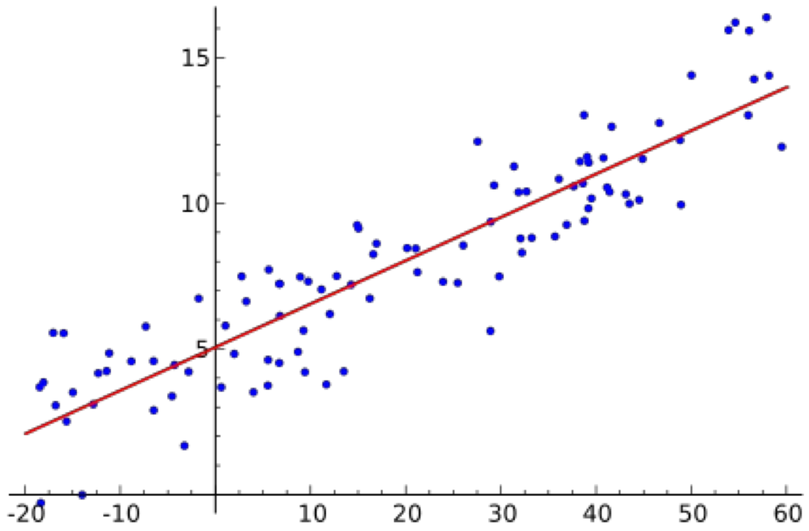


Regression: Fitting a Model to Data

Can be computed directly

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

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



PROBLEM #3

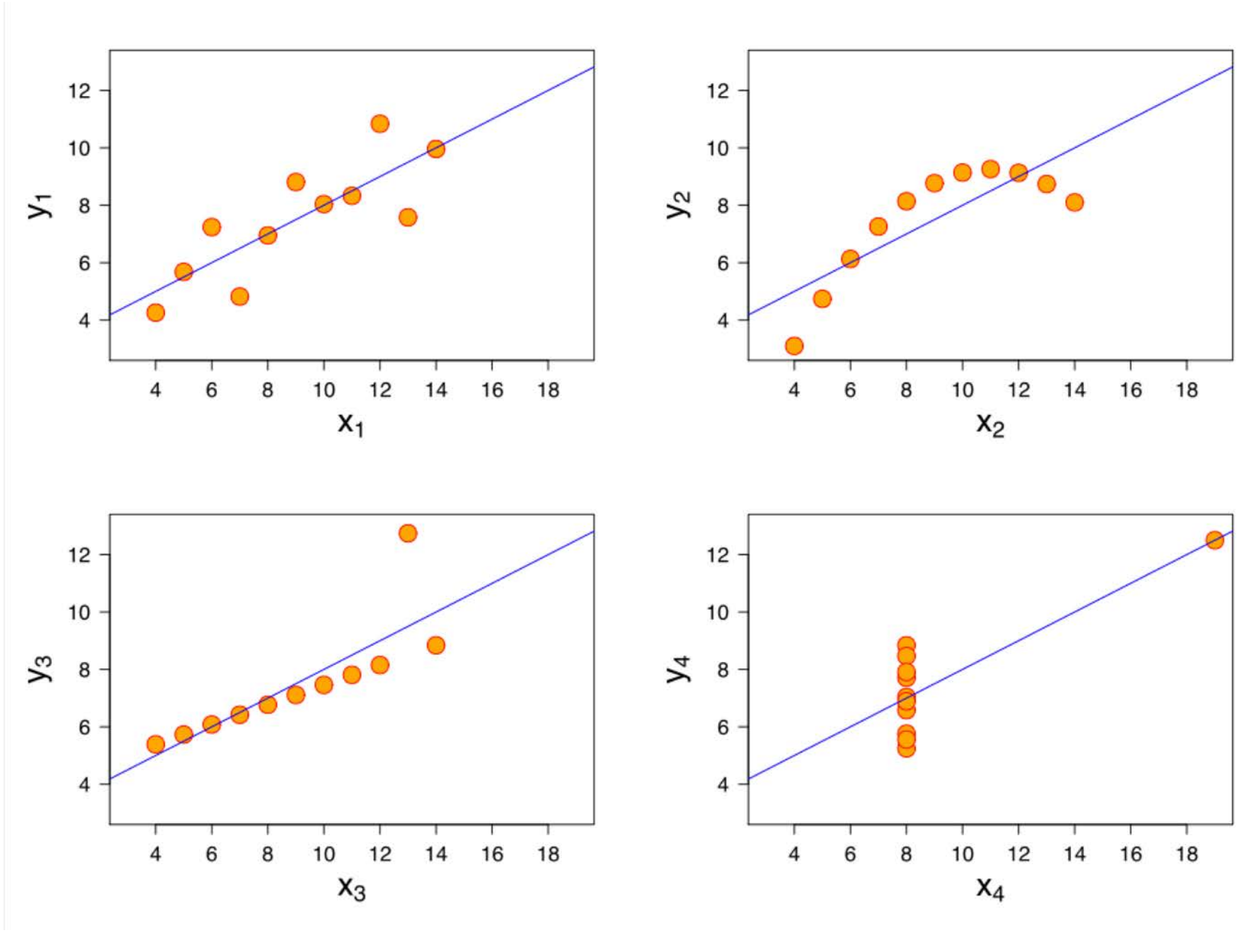
What is lost or misinterpreted…

So, what's the difference between Correlation and Regression and how does it impact our visualization?

KNOW THE SHAPES (INFORMATION) YOUR STATISTIC CAPTURES

Statistical Limitations

Anscombe's quartet

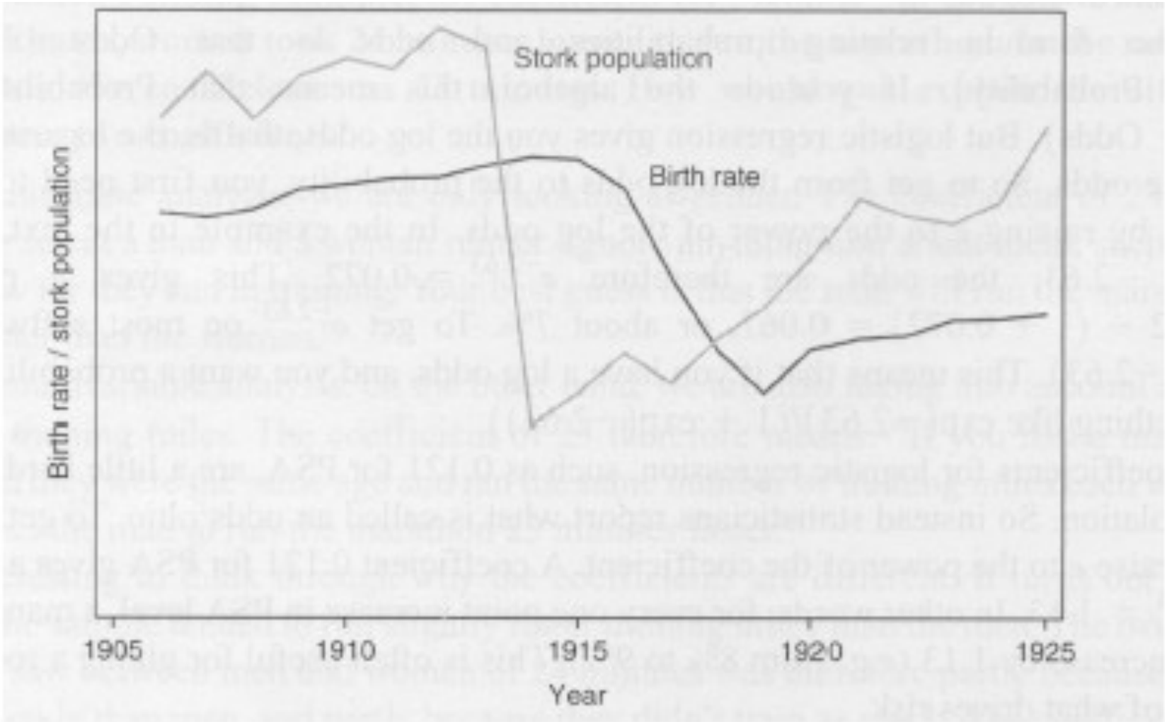


Statistical Limitations

Anscombe's quartet

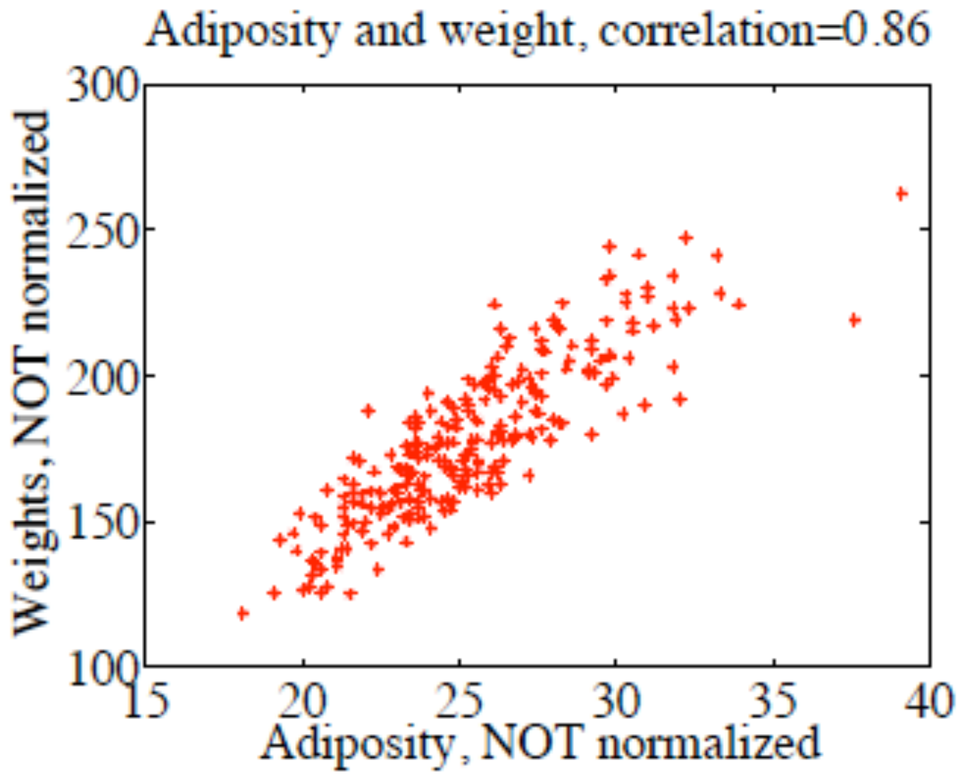
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	plus/minus 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively

CORRELATION != CAUSALITY



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

BUT CAN BE USED TO PREDICT

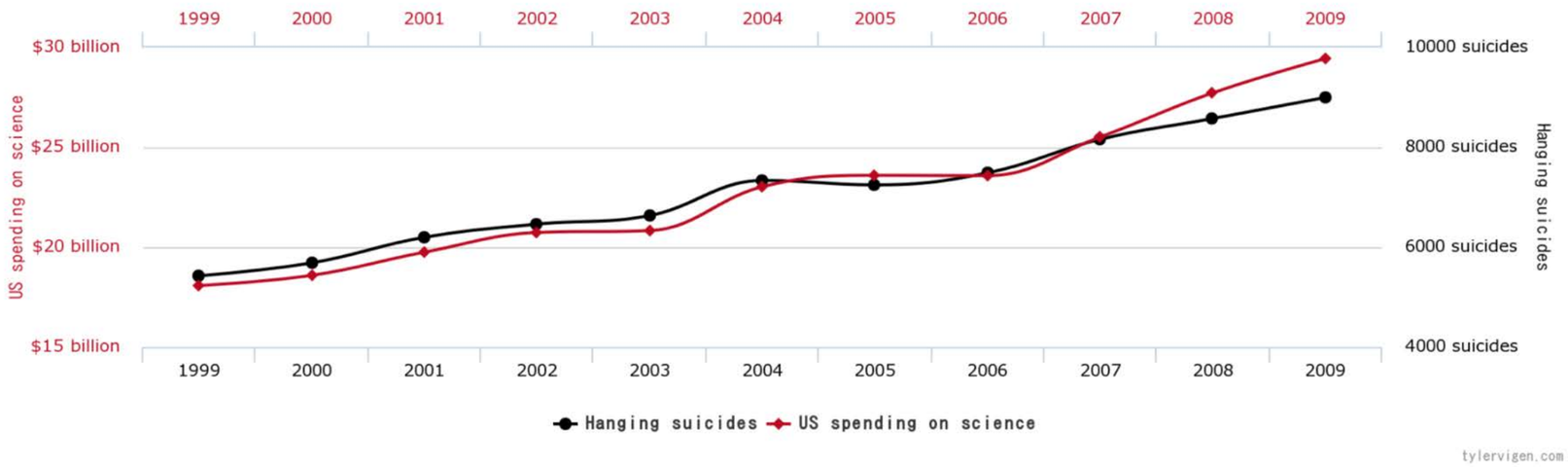


Spurious correlations

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation

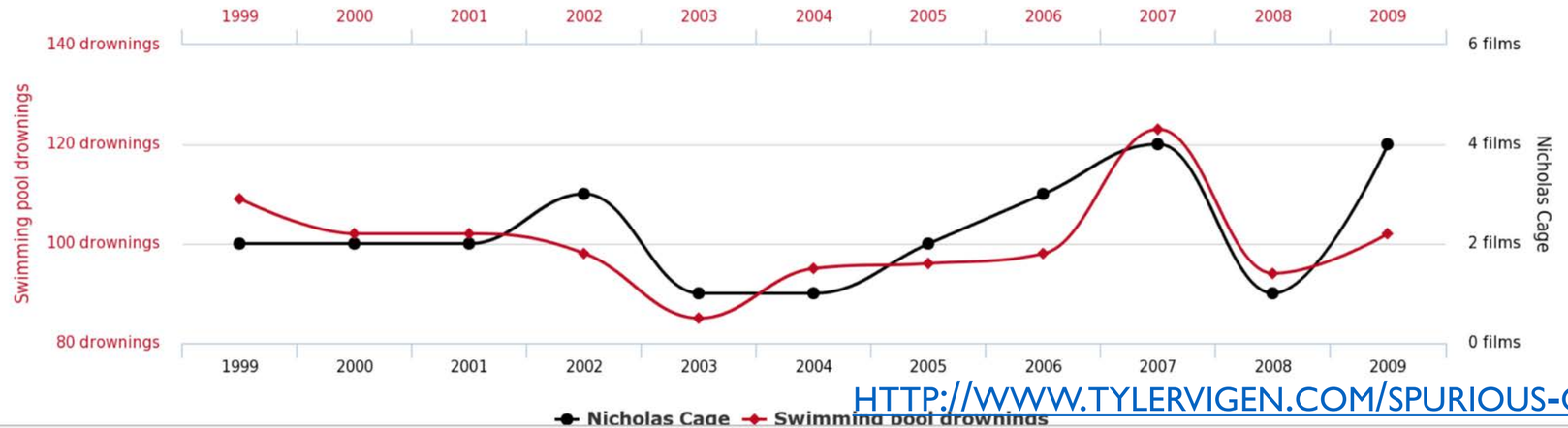


tylervigen.com

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in



HTTP://WWW.TYLERVIGEN.COM/SPURIOUS-CORRELATIONS

Statistical Vis Tools: R

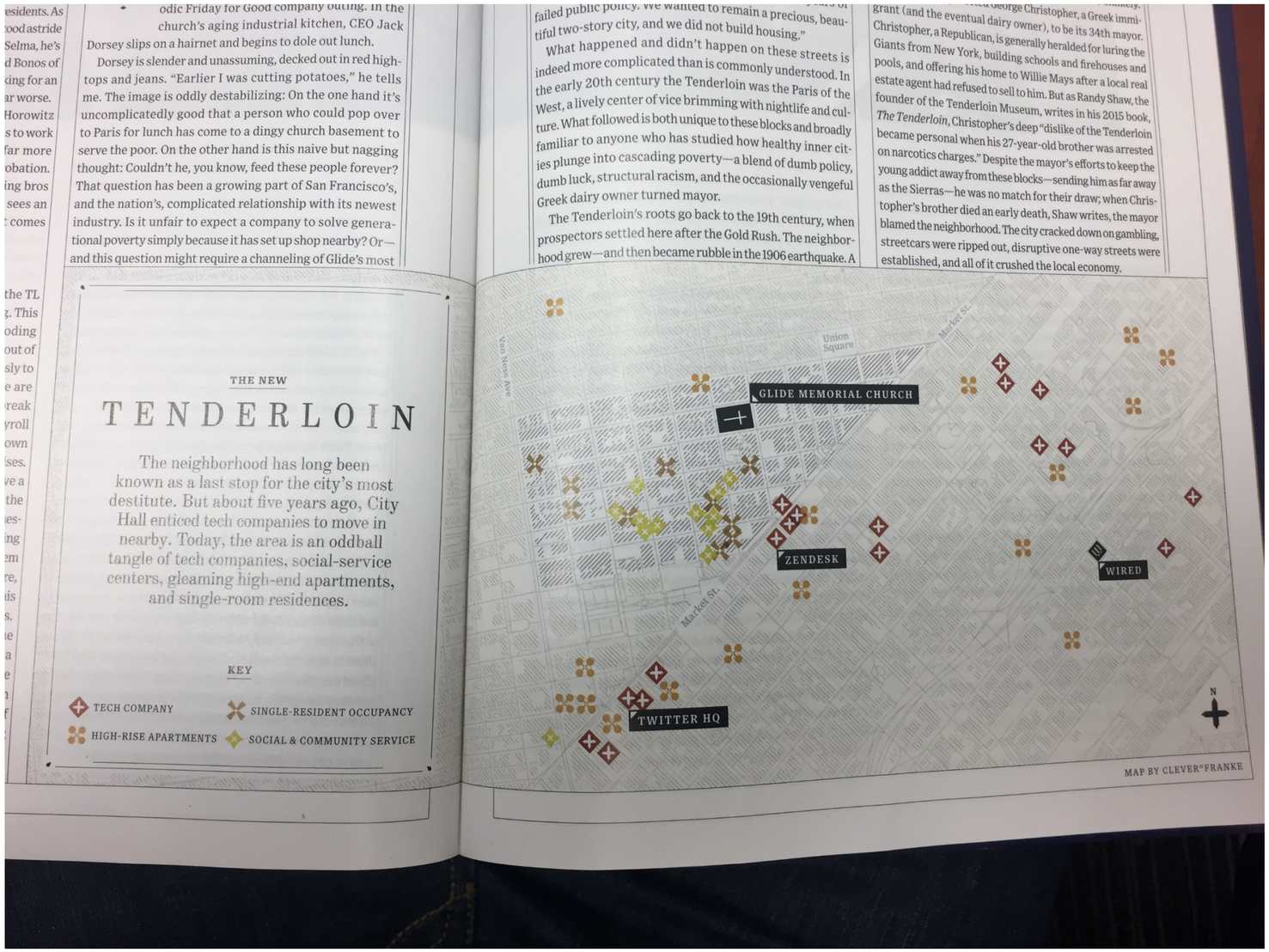
It's free

It's easy to get pictures up and going

Many, many (almost too many) tools already available

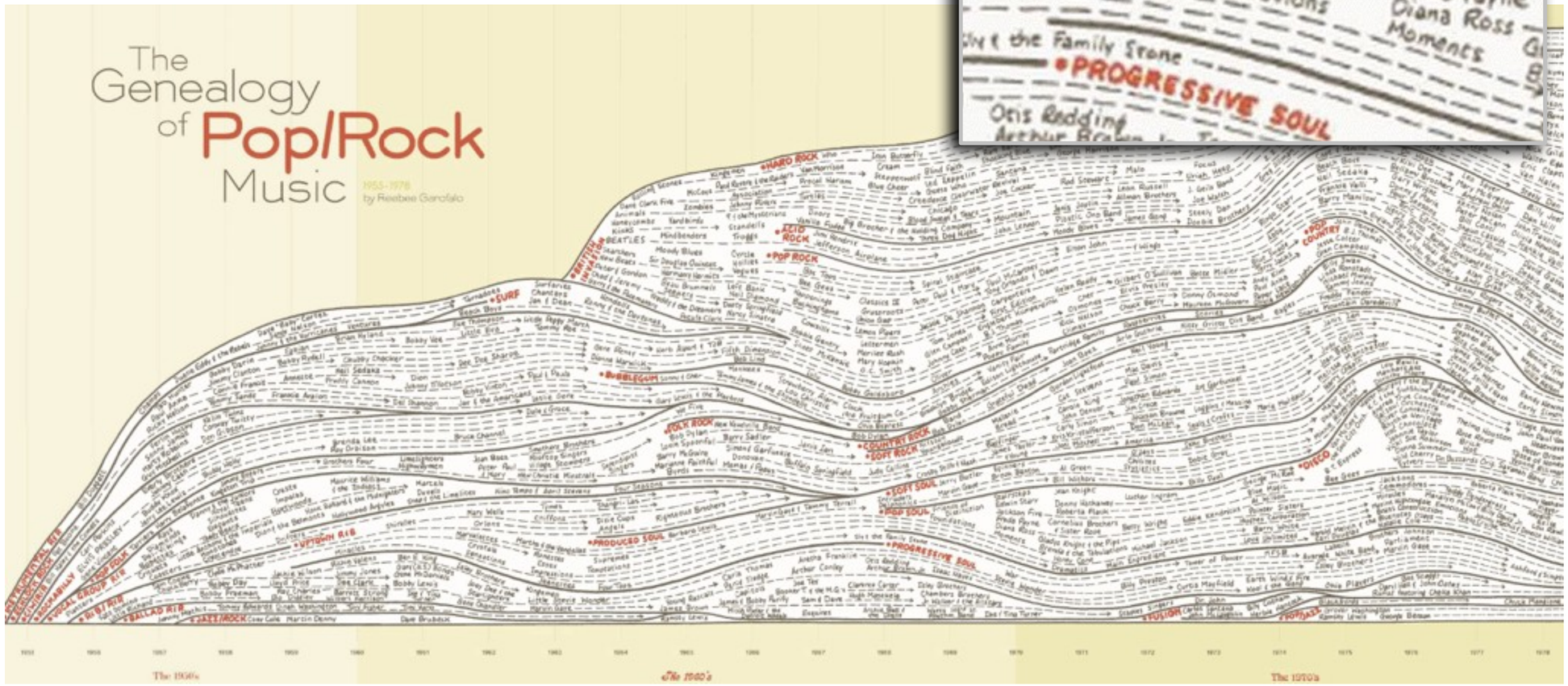
Let's you work with tools *without* implementing them. However, you should UNDERSTAND them.

HTTP://STUDENTS.BROWN.EDU/SEEING-THEORY/INDEX.HTML

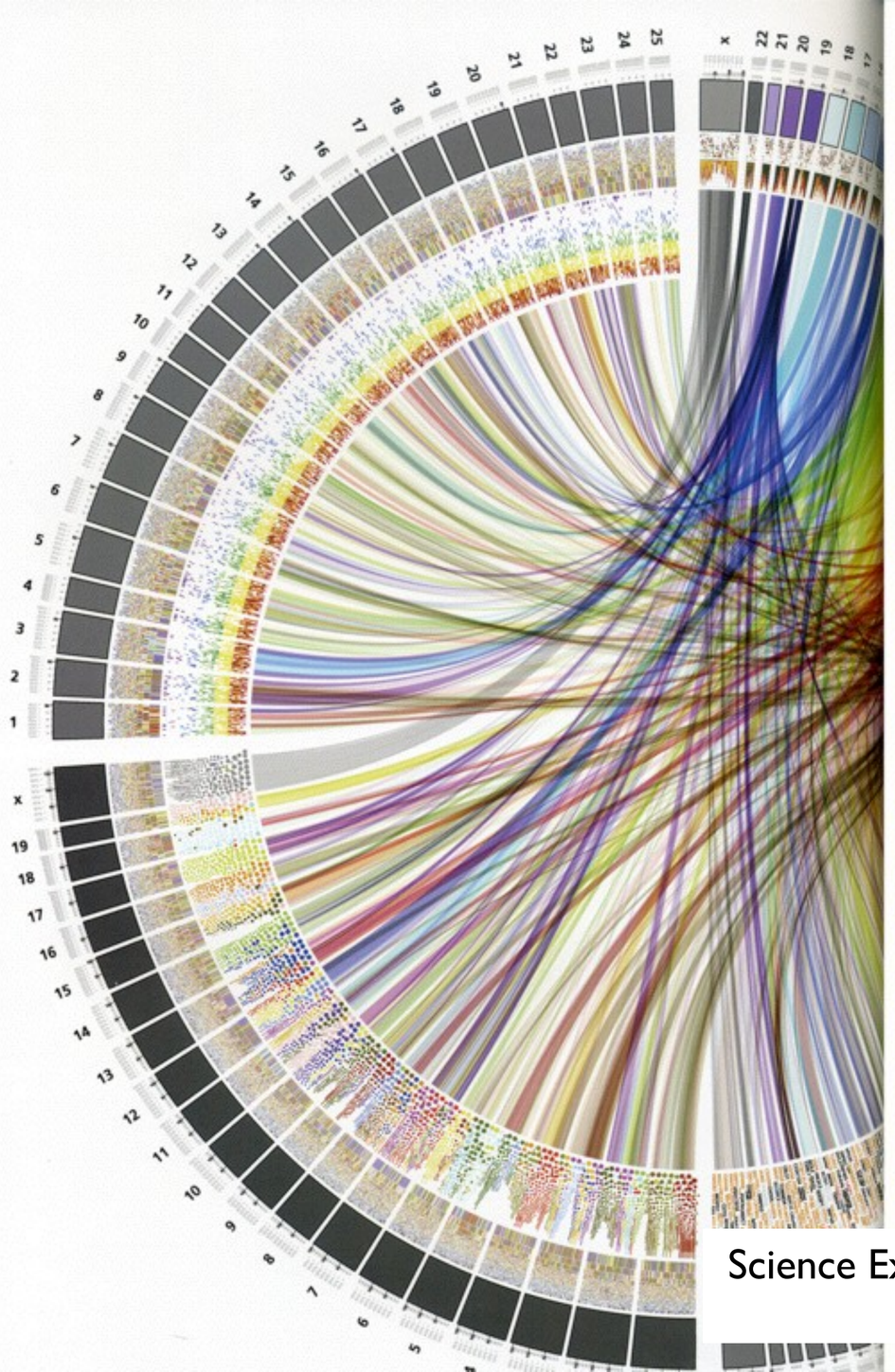


WIRED, JAN 2017

MAXIMIZE AMOUNT OF DATA SHOWN



STEVE CHAPPEL AND REEBE GAROFALO
IN ROCK 'N' ROLL IS HERE TO PAY: THE HISTORY AND POLITICS OF THE MUSIC INDUSTRY, 1977



On the road to a digital society

Computer technology is an ubiquitous element of our world, and fast networks are spanning the globe. This is changing the way we live and work and communicate. A new digital world is emerging, an environment in which creativity and innovation can flourish in many new ways. As a result, science and research have a greater influence on our life in the 21st century than ever before. This is attributable to massive investments in research and development, but also to intensive cooperation and tough competition. The convergence of nano-, bio-, information- and neurotechnologies facilitates completely new applications. Taking its place beside the more traditional factors of land, capital and employment, knowledge is fast becoming the decisive factor for prosperity – and also for the resolution of global problems. In this, the appropriate balance between digital freedom and digital security must be maintained.

Science 2020: Systematically surveying the world *Millions of scientists are getting to the bottom of the secrets of our world, across the whole spectrum of space, time, energy and complexity. Fundamentally new knowledge is emerging from research into inter-disciplinary topics or extreme states of matter. Science long ago escaped the constraints of working only in the realm of our natural living conditions and our perceptions. Considerable investment is flowing into efforts to decode the smallest building blocks of our world and to understand how their interplay produces brand new qualities. The drivers of innovation in research today are data capture via digital sensors; storage, analysis and visualisation via computer and software; and the global exchange of information and knowledge.*

The cost of new knowledge is rising There is now no part of our life that is not the subject of research. At the same time, it is becoming ever more difficult to generate new knowledge. These days, new research methods and technologies enable us to study even the ›farthest frontiers‹ of the world: extremely fast or slow processes, the tiniest building blocks or the largest structures, extreme cold or extreme heat.

Networked knowledge takes on global challenges *Thanks to worldwide information and communication networks, the challenges our civilisation faces in the long term are known to us sooner and more clearly than ever before. We can start developing solutions together at an earlier stage. Research on many topics is global – taking place in close cooperation or in international competition for the fastest and best solutions. National boundaries are becoming irrelevant. Millions of scientists work across countries, continents and time zones in thousands of labs. Their global networking enhances the diversity and efficiency of science and technology. And this, in turn, reinforces globalisation and networking. In a world changing at such a pace, each country must redefine its place.*

The end of distance Mankind faces enormous challenges both locally and globally – the challenge of using resources sustainably and of organising a global economy. Across the globe, complex processes are being recorded in detail, collated in databases and analysed in computer networks. New visualisation techniques make it possible to analyse larger and larger data records and to draw conclusions from the results.

Global networking as the driving force of science In the early days, the Internet linked up scientists, large-scale equipment and information; now it networks computational power and enormous amounts of data through grid and cloud computing. A global Semantic Web is emerging, bringing together data, expertise and knowledge that had previously been distributed among virtual libraries and observatories. The information is being intelligently developed,

must have access to the Internet. This is where new knowledge is freely available.

Science Express: How Science and Technology change our life.
Herausgegeben von der Max-Planck-Gesellschaft

QUESTIONS