Programming with Data Bootcamp: Lecture 4

http://dsg.csail.mit.edu/6.S079/

Slides courtesy of Emanuel Zgraggen / Sam Madden / Tim Kraska (6.S079)

Key ideas:
Data Visualization
Lying w/Statistics and Visualizations

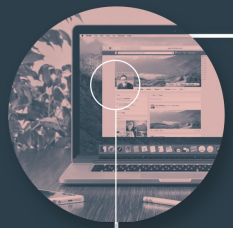
A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter



4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

463EB

of data will be created every day by 2025

IDC

95m

photos and videos are shared on Instagram

Instagram Business

294bn

billion emails are sent

Radicati Group

320bn emails to be sent each day by 2021

306bn emails to be sent each day by 2020

3.9bn

people use emails

65bn

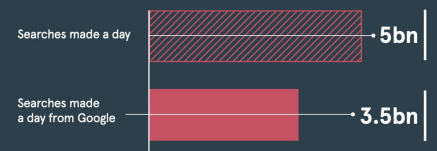
messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

4TB

of data produced by a connected car

Intel



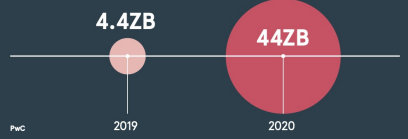
Smart insights

28PB

to be generated from wearable devices by 2020

Statista

ACCUMULATED DIGITAL UNIVERSE OF DATA



PwC

RACONTEUR

https://www.raconteur.net/infographics/a-day-in-data/

How do we make sense of all this data?

How do we use data in decision-making processes?

How do we avoid being overwhelmed?

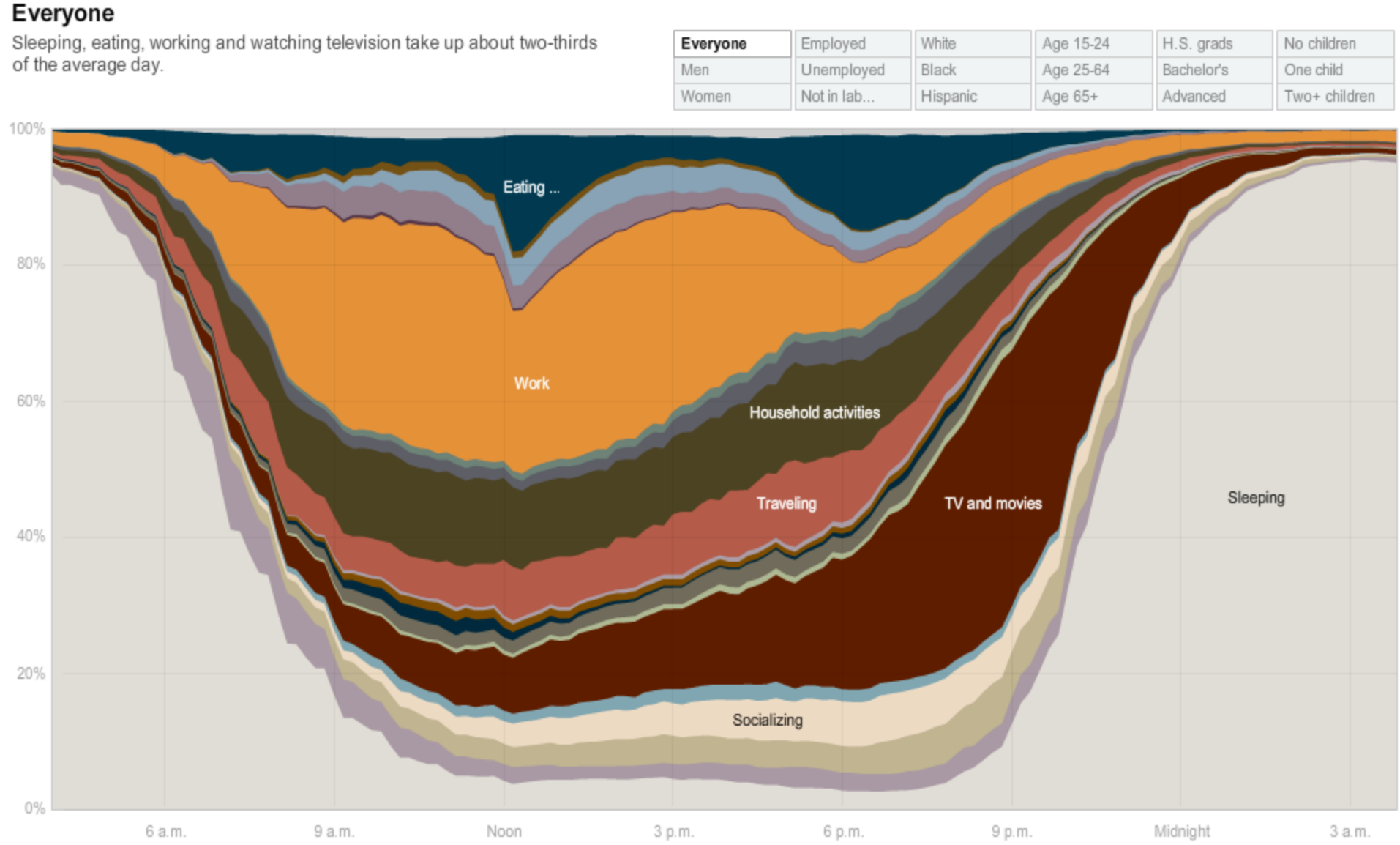
Challenge

Transform the data into understanding and insight thus making it useful to people

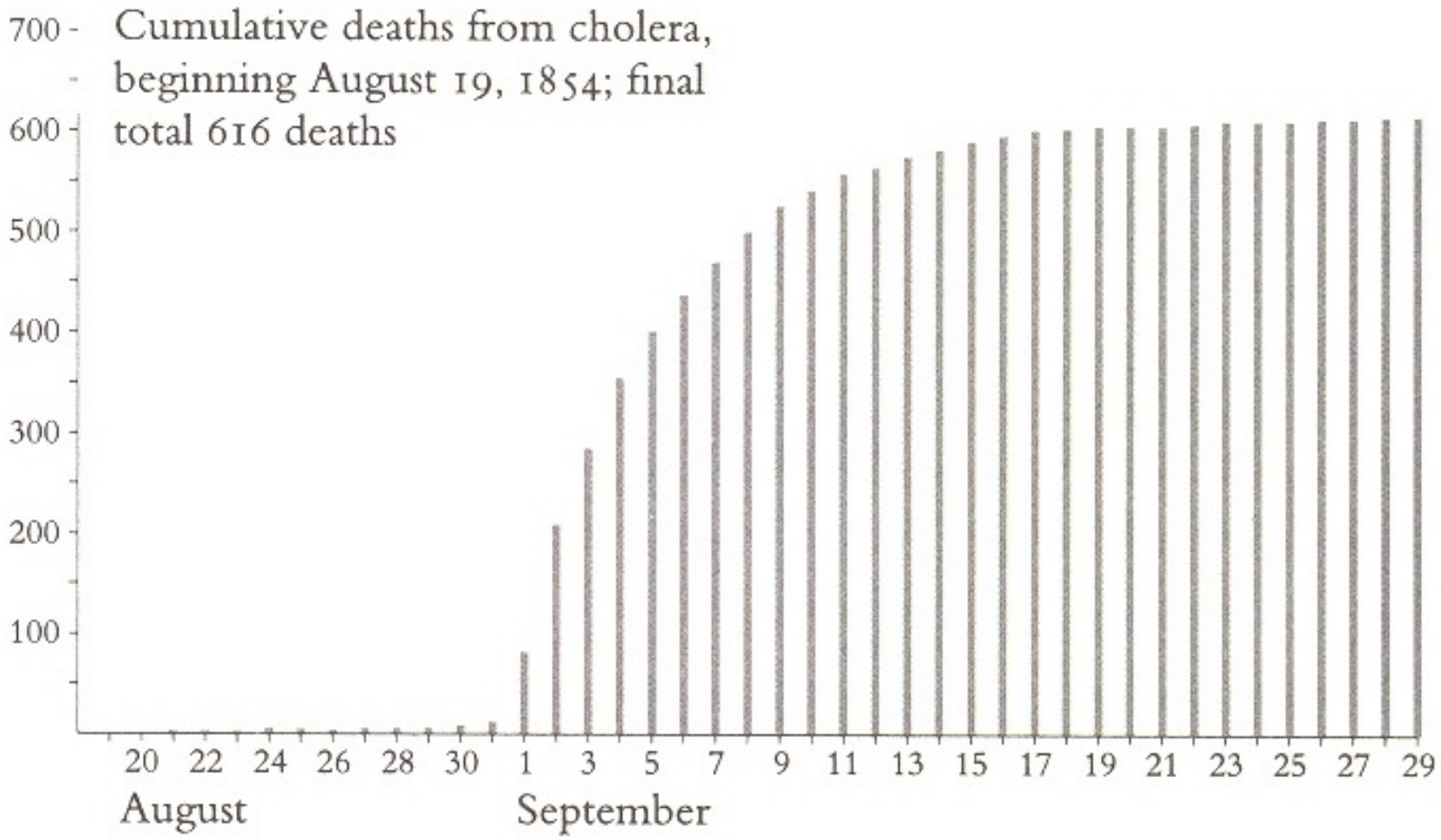
Why create visualizations?

- Answer questions (or discover them)
- Make decisions
- See data in context
- Expand memory
- Support graphical calculation
- Find patterns
- Present argument or tell a story
- Teach

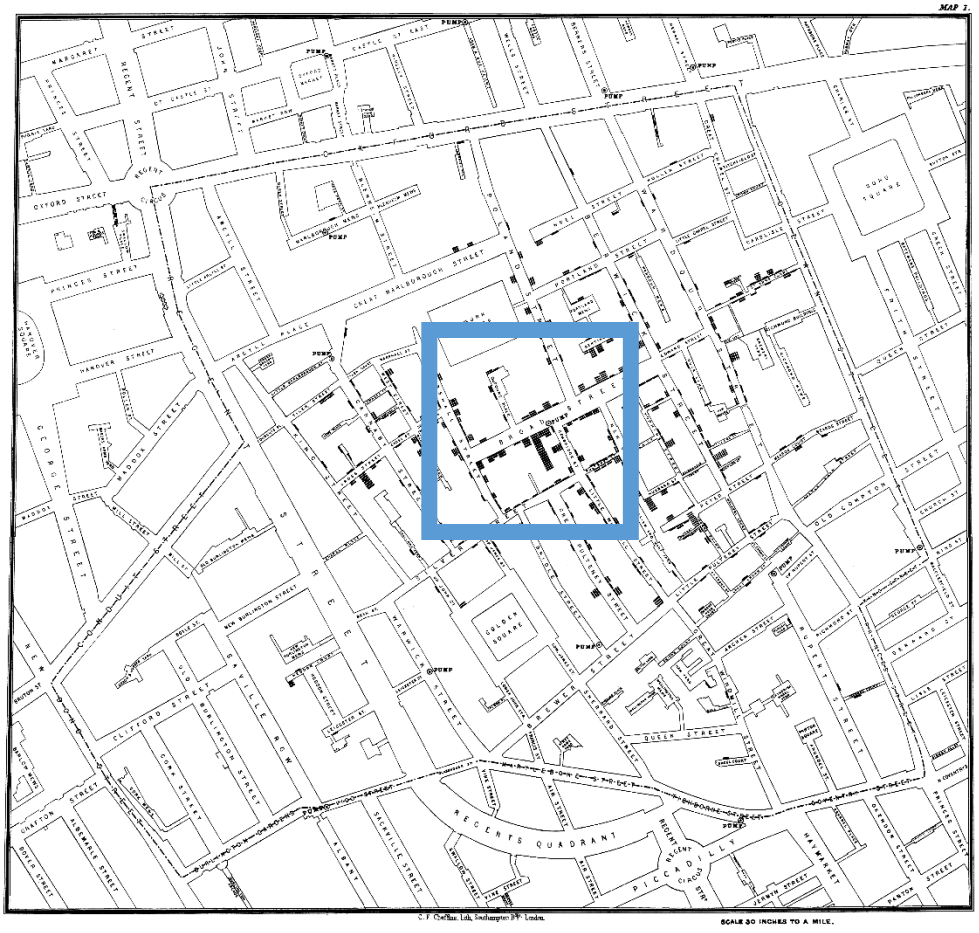
Answer a question



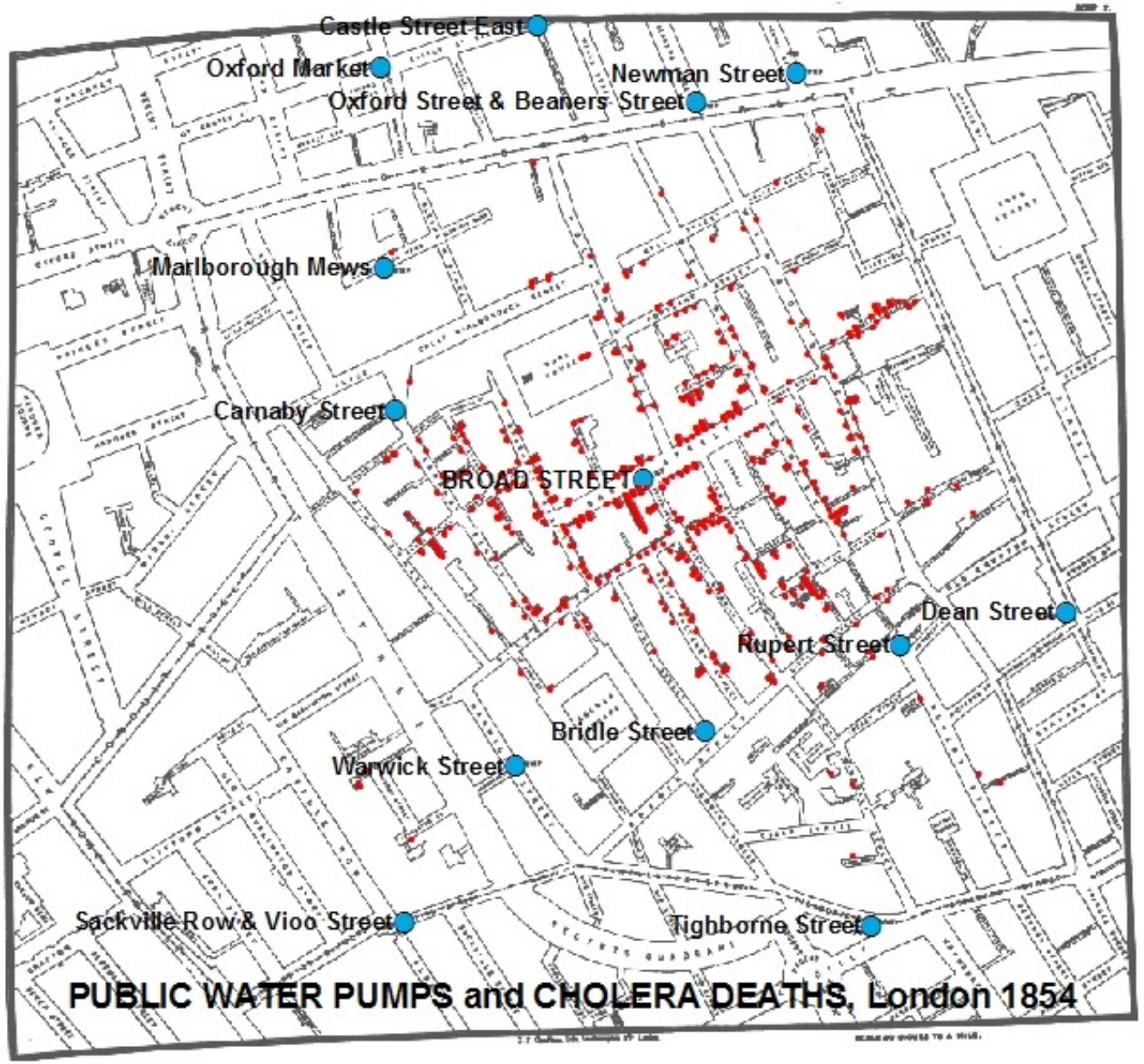
[New York Times]


Cumulative deaths from cholera, beginning August 19, 1854; final total 616 deaths
700
600
500
400
300
200
100
20 22 24 26 28 30 1 3 5 7 9 11 13 15 17 19 21 23 25 27 29
August
September

See data in context

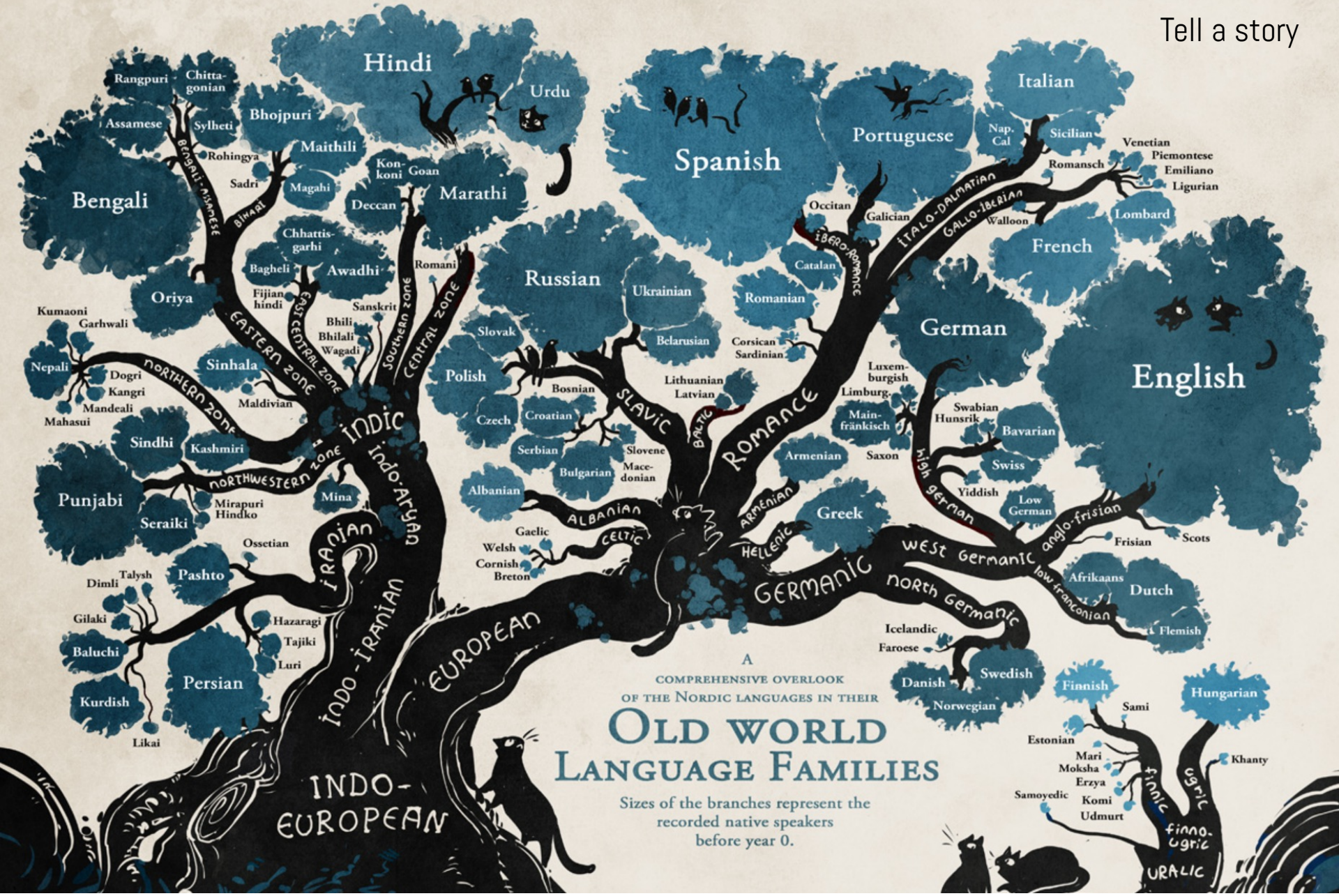


See data in context



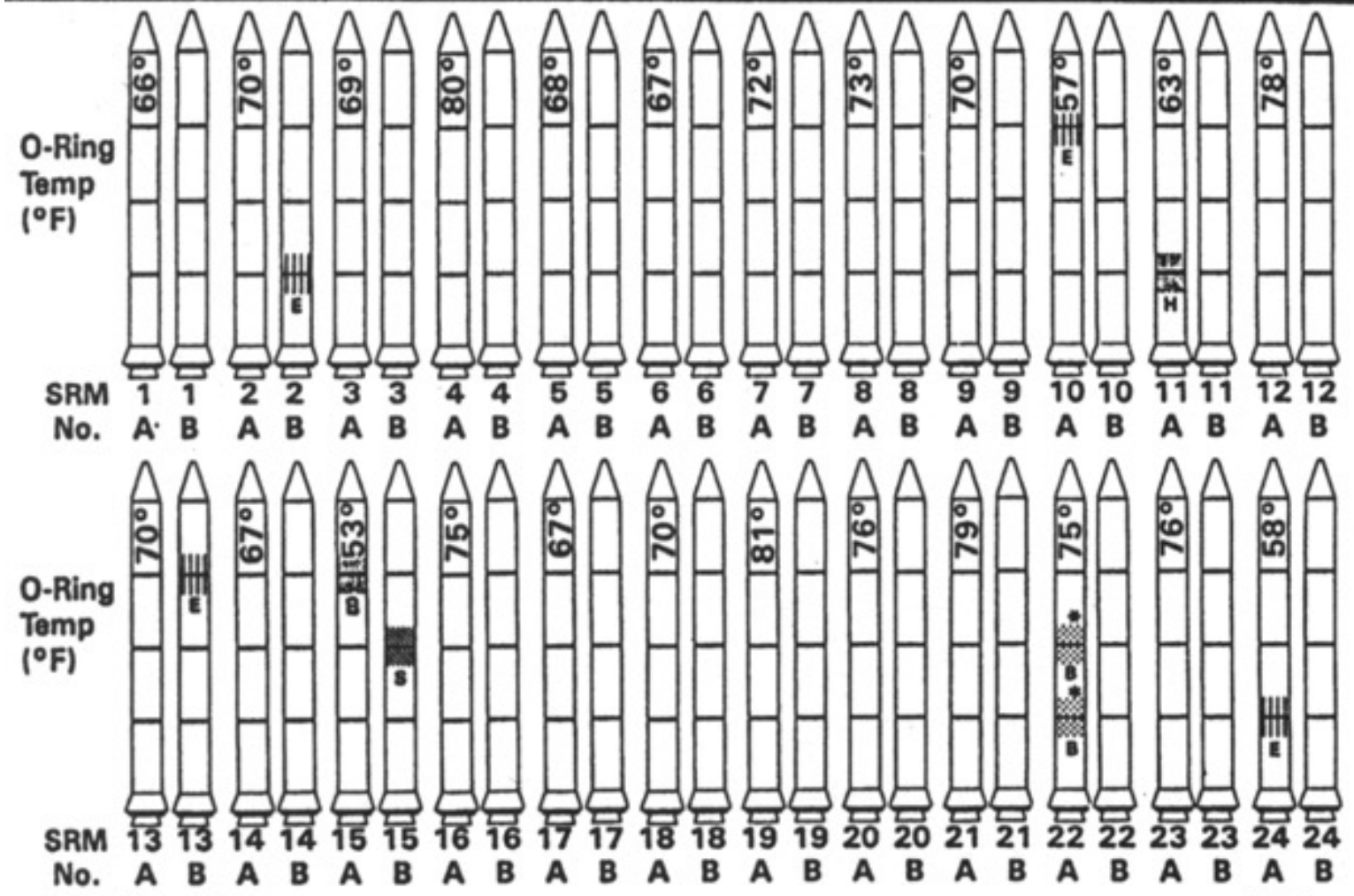
[http://www.udel.edu/johnmack/frec682/cholera/]

Tell a story



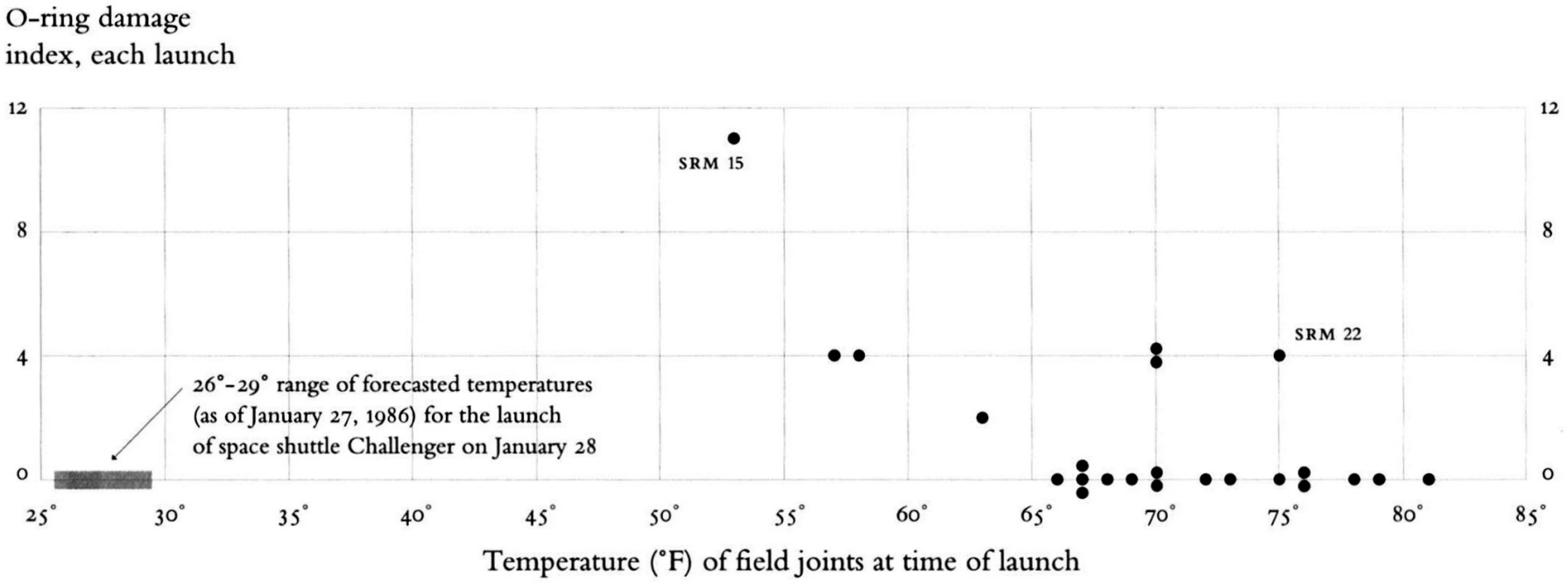
[Minna Sundberg]

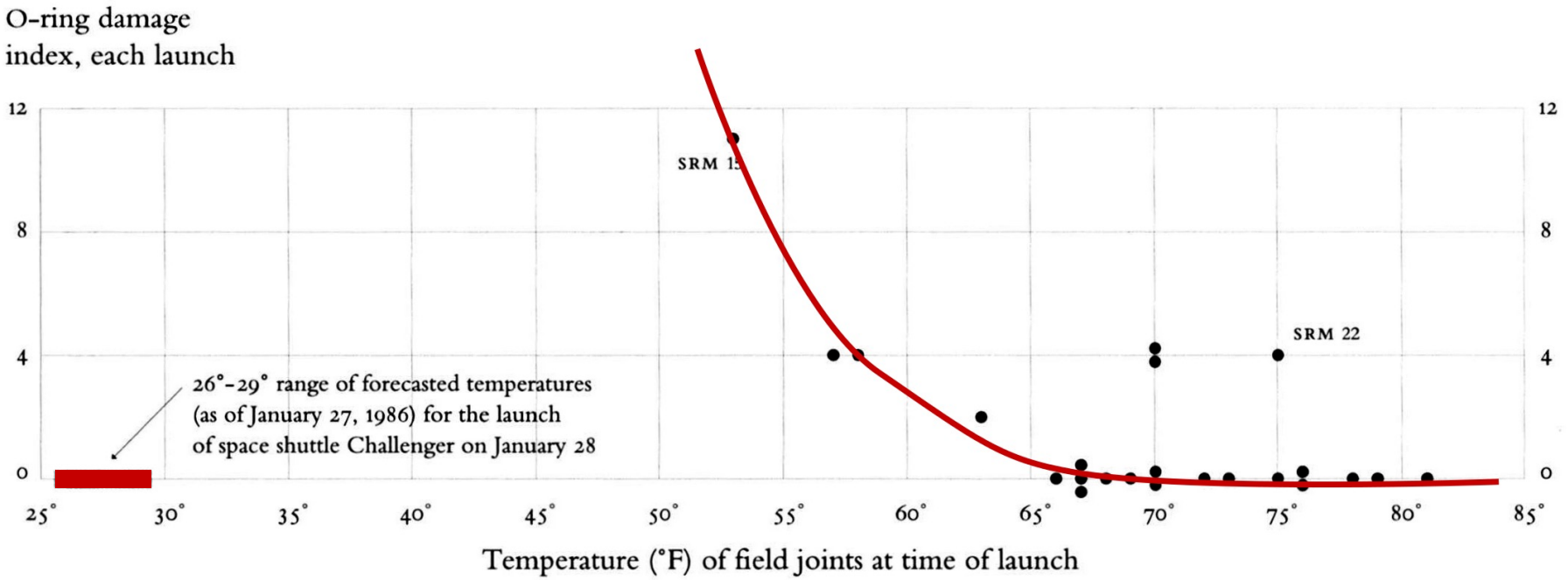
History of O-Ring Damage in Field Joints (Cont)



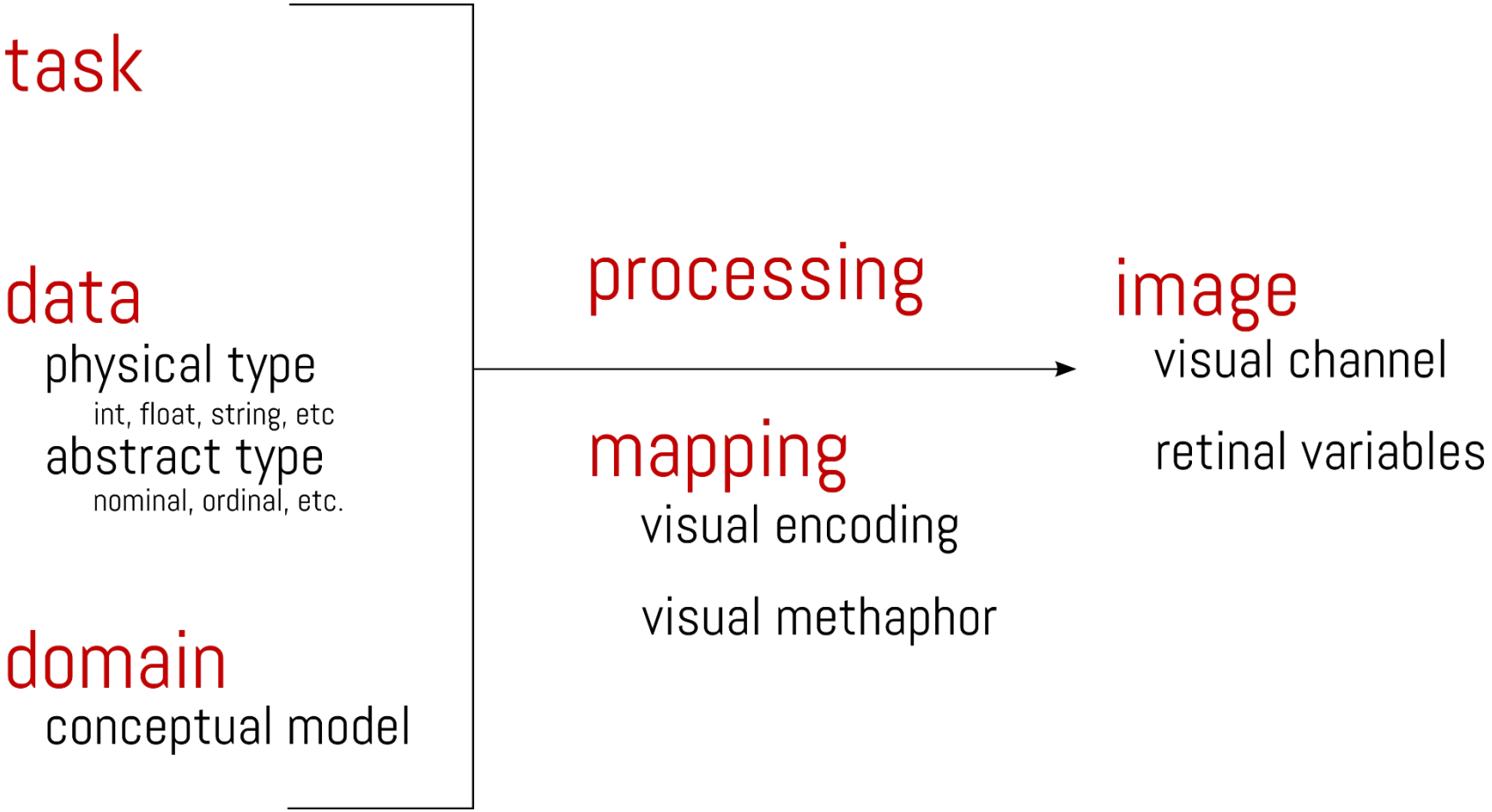
MORTON THIOKOL, INC.
Wasatch Operations

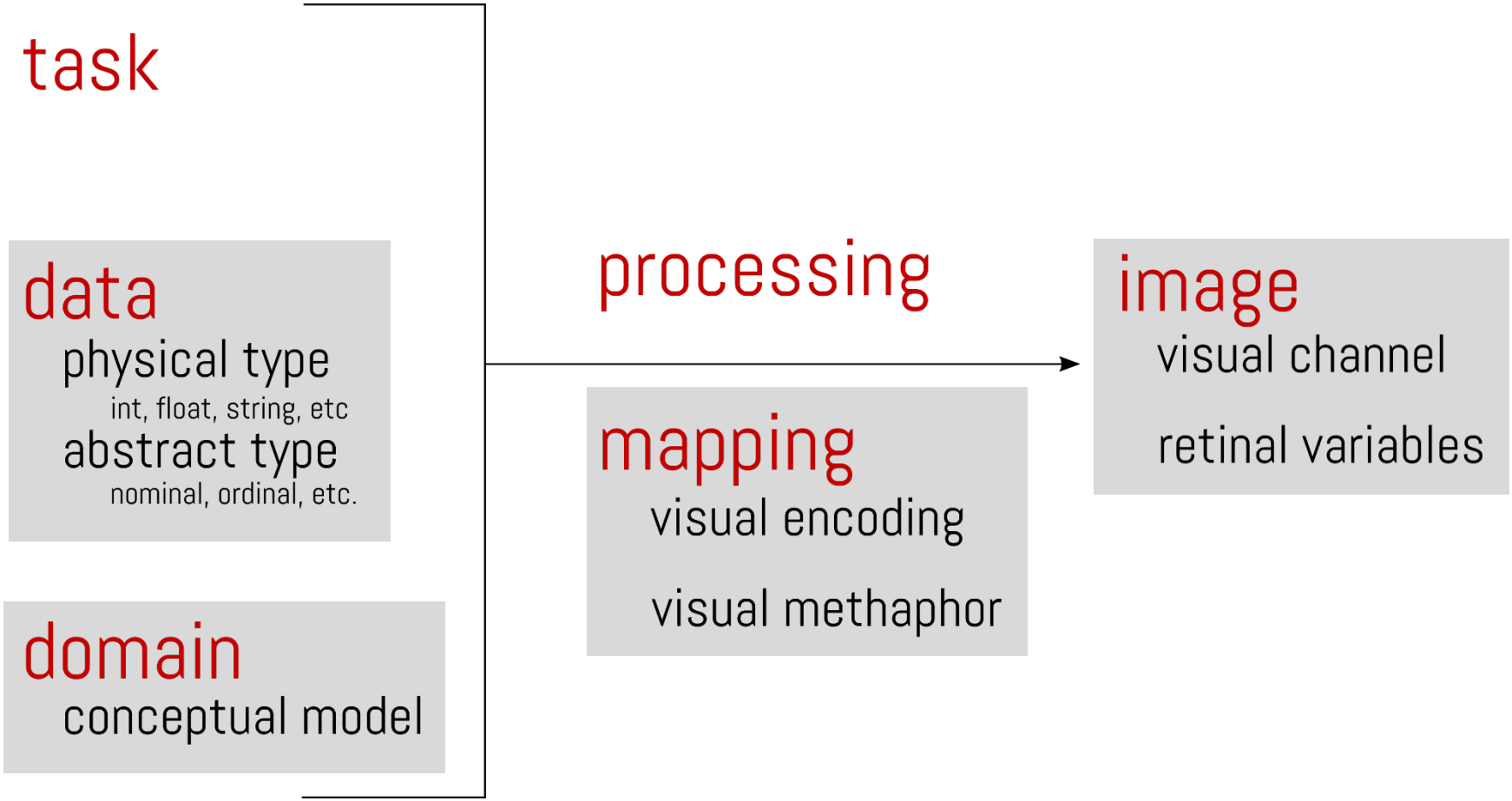
INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION


O-ring damage
index, each launch
12
8
4
0
SRM 15
SRM 22
26°–29° range of forecasted temperatures
(as of January 27, 1986) for the launch
of space shuttle Challenger on January 28
25°
30°
35°
40°
45°
50°
55°
60°
65°
70°
75°
80°
85°
Temperature (°F) of field joints at time of launch

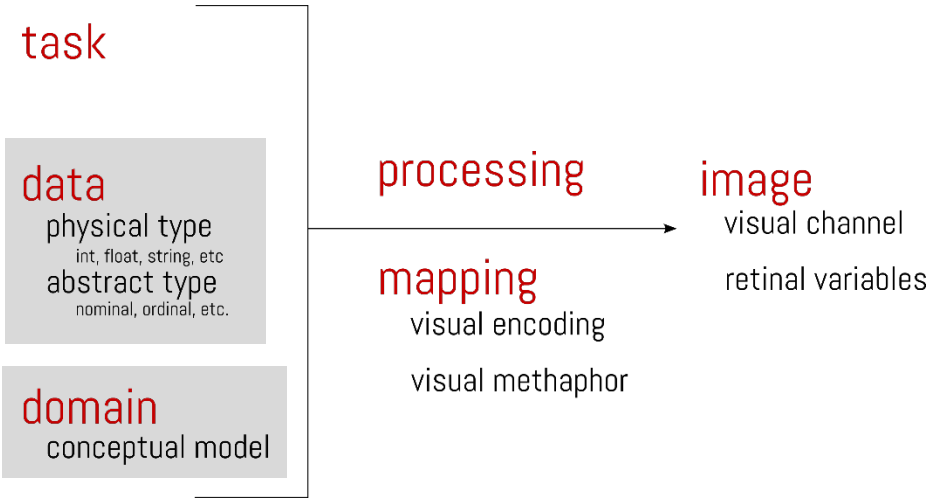

O-ring damage
index, each launch
12
8
4
0
SRM 15
SRM 22
26°–29° range of forecasted temperatures
(as of January 27, 1986) for the launch
of space shuttle Challenger on January 28
25°
30°
35°
40°
45°
50°
55°
60°
65°
70°
75°
80°
85°
Temperature (°F) of field joints at time of launch

How do we create visualizations?


task
data
physical type
int, float, string, etc
abstract type
nominal, ordinal, etc.
domain
conceptual model
processing
mapping
visual encoding
visual methaphor
image
visual channel
retinal variables


task
data
physical type
int, float, string, etc
abstract type
nominal, ordinal, etc.
domain
conceptual model
processing
mapping
visual encoding
visual methaphor
image
visual channel
retinal variables

Data & Domain


task
data
physical type
int, float, string, etc
abstract type
nominal, ordinal, etc.
domain
conceptual model
processing
mapping
visual encoding
visual methaphor
image
visual channel
retinal variables

Data Model

- How the data is organized
- How are data elements related

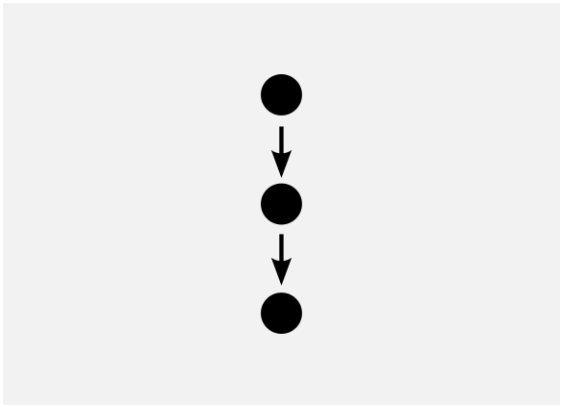
Conceptual Model

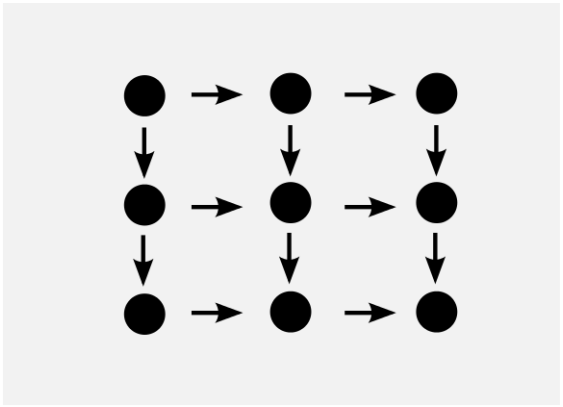
- Mental constructions
- Include semantics and support reasoning

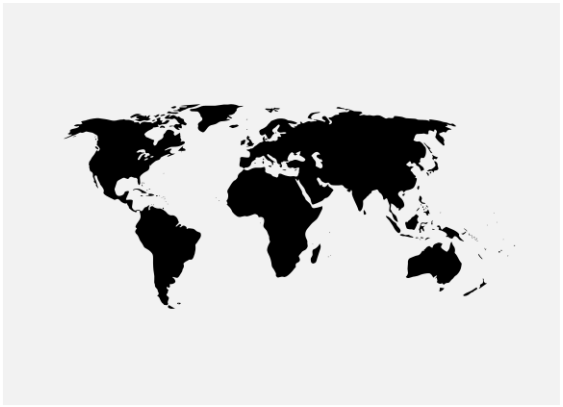
Data vs. Conceptual

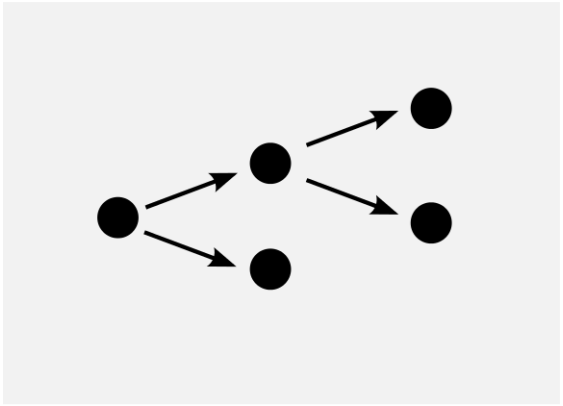
- 1D list of floats vs. Temperature
- 3D list of floats vs. Space

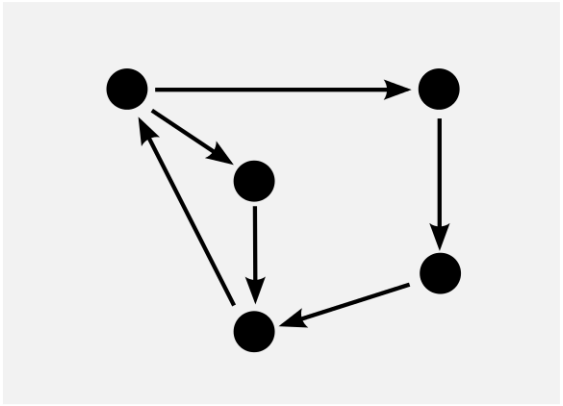
Data Model Taxonomy

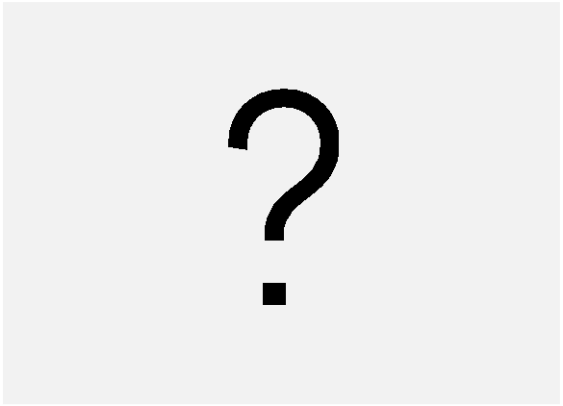












Variables

- Physical types
  - Characterized by storage format
  - Characterized by machine operations
  - Example:
    - bool, short, int32, float, double, string, …
- Level of measurement
  - Describes the relationship among values
    - Nominal
    - Ordinal
    - Quantitative

Nominal, Ordinal and Quantitative

- N – Nominal (labels):
  - Fruits: Apples, Oranges, …
- O – Ordinal
  - Quality of meat: Grade A, AA, AAA
- Q – Interval (Location of zero is arbitrary)
  - Dates: Mar. 14, 1933
  - Lat: 26.1, Long: -110.0
  - Only differences (i.e. Intervals) can be compared
- Q – Ratio (zero fixed)
  - Physical measurements: Length, Mass
  - Counts and amounts

[S. S. Stevens, On the theory of scales of measurements, 1946]

Nominal, Ordinal and Quantitative

- N - Nominal (labels):
  - Operations: =, ≠
- O - Ordinal
  - Operations: =, ≠, <, >, ≤, ≥
- Q - Interval (Location of zero arbitrary)
  - Operations: =, ≠, <, >, ≤, ≥, -
  - Can measure distances or spans
- Q - Ratio (zero fixed)
  - Operations: =, ≠, <, >, ≤, ≥, -, ÷
  - Can measure ratios or proportions

[S. S. Stevens, On the theory of scales of measurements, 1946]

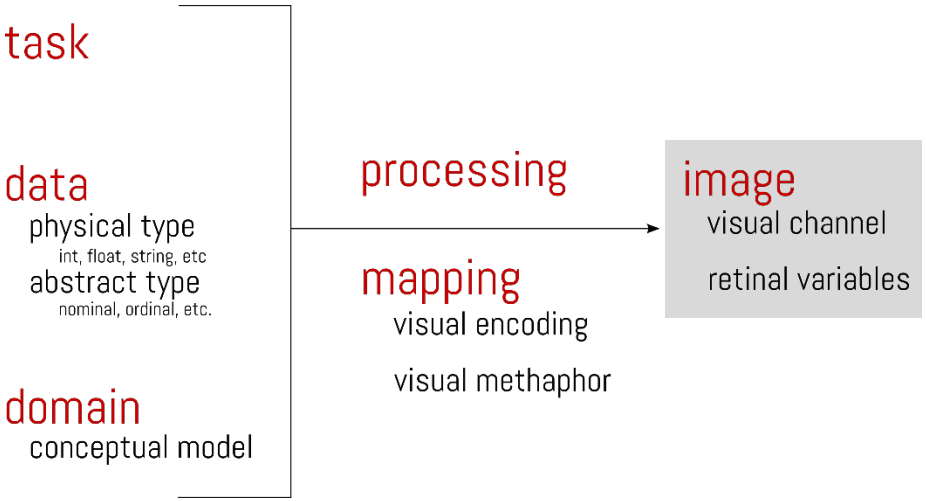
Example

- Data Model
  - 32.50, 54.0, 17.30, …
  - 1D, floats
- Conceptual Model
  - Temperature
- N,O,Q Type
  - Burned vs. Not burned (N)
  - Hot, warm, cold (O)
  - Continuous range of values (Q)

Width	Length	Species
1.2	5.1	setosa
1.4	4.9	versicolor
0.8	4.7	virginica
1.7	4.6	setosa
4.9	4.3	virginica

Q, Q, N

Image


task
data
physical type
int, float, string, etc
abstract type
nominal, ordinal, etc.
domain
conceptual model
processing
mapping
visual encoding
visual methaphor
image
visual channel
retinal variables

Pre-attentive

unconscious, parallel, fast

Attentive

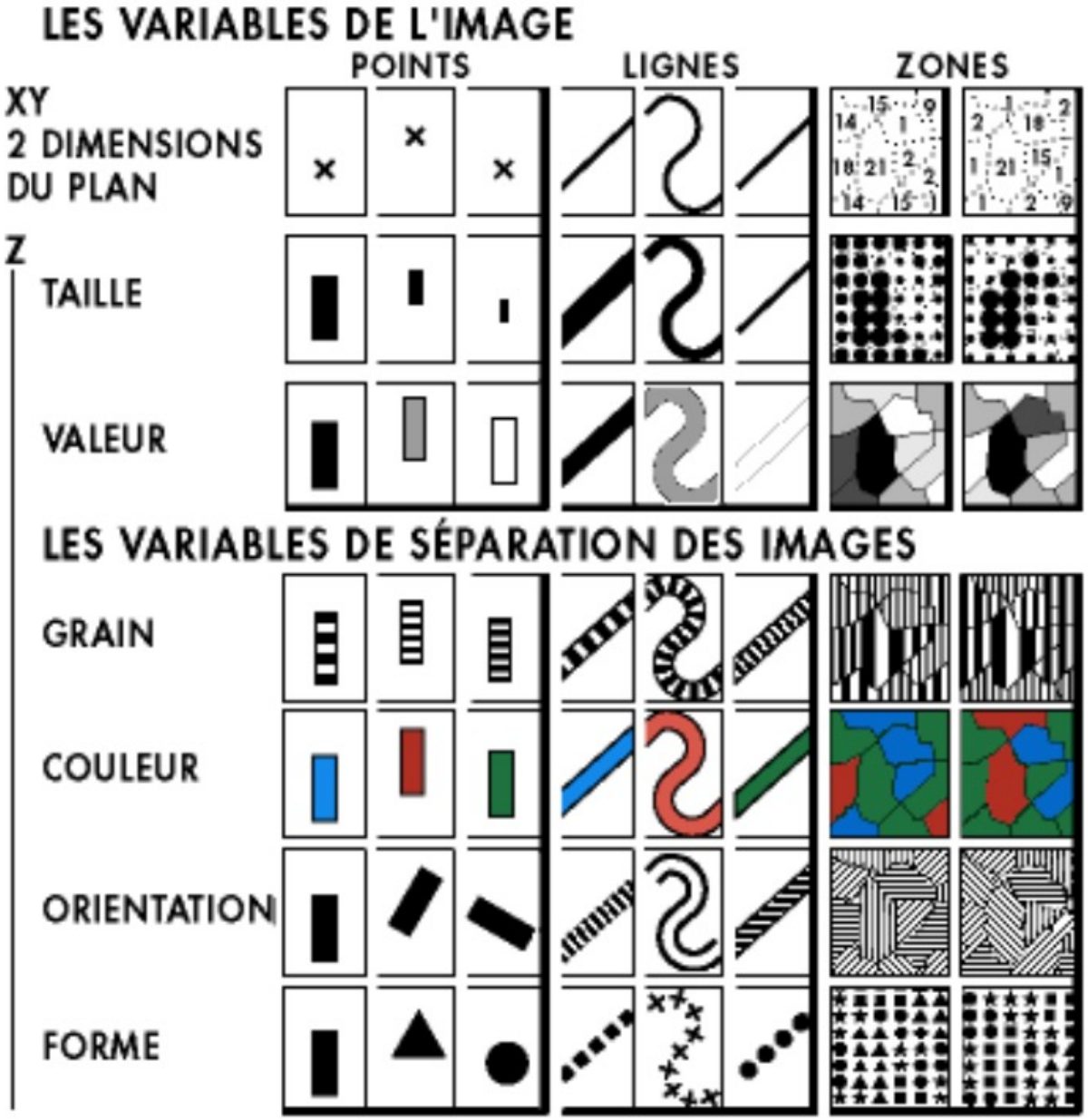
conscious, serial, slow

How many 3's?

128176875613897654698450698560498282676298098584582245098564589450984509809435859091030209905959595772564675050678904567884578980982167765487636490856091294968 6

How many 3's?

1281768756138976546984506985604982826762
9809858458224509856458945098450980943585
9091030209905959595772564675050678904567
8845789809821677654876364908560912949686



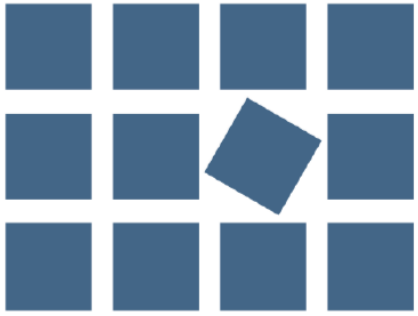
[Bertin, Simonolgy of Graphics, 1983]

Color hue

Color brightness

Position

Orientation

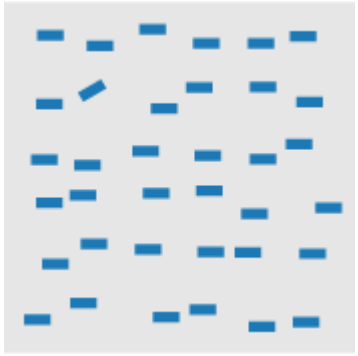


Color saturation

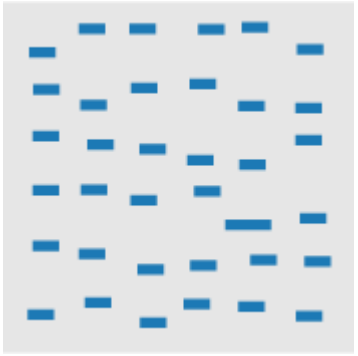
Size

Texture

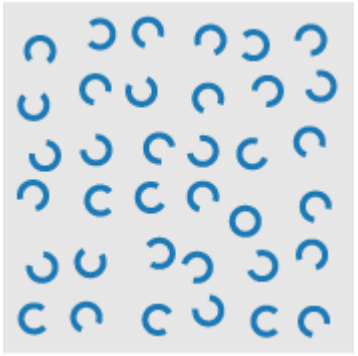
Shape



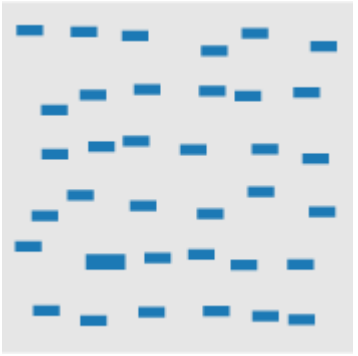
Line orientation



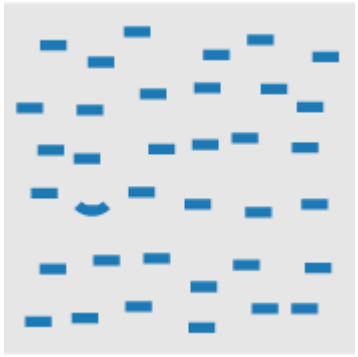
Length, width



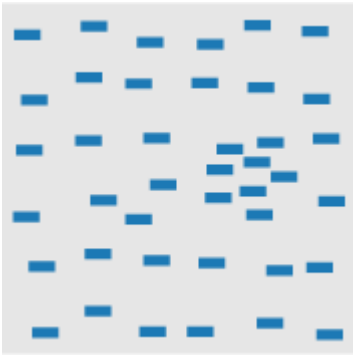
closure



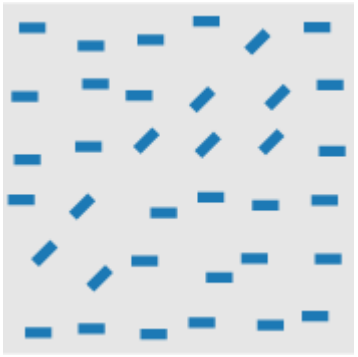
size



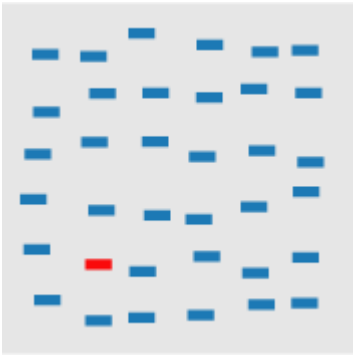
curvature



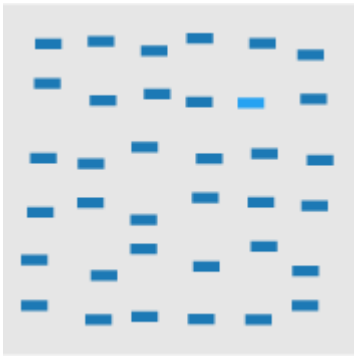
density



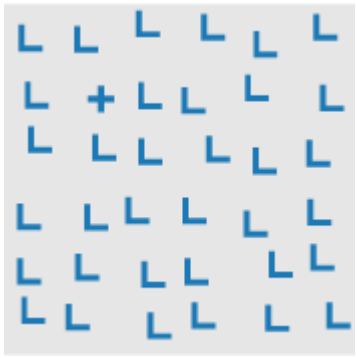
number, estimation



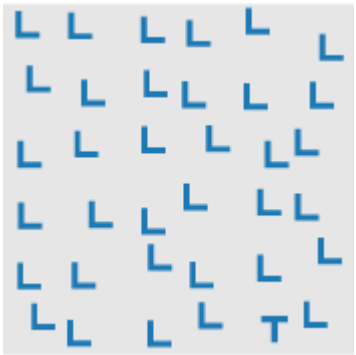
hue



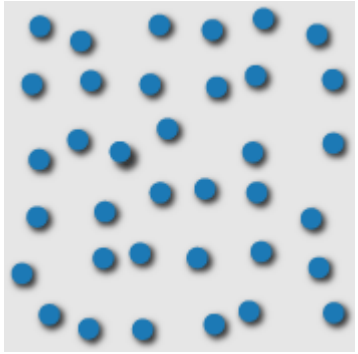
intensity



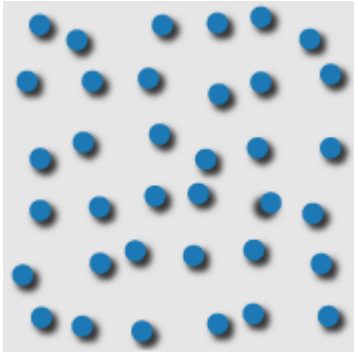
intersection



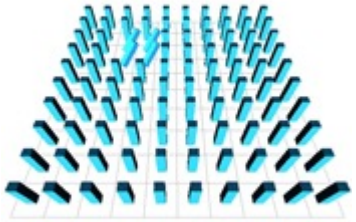
terminators



3D depth cues,
stereoscopic depth

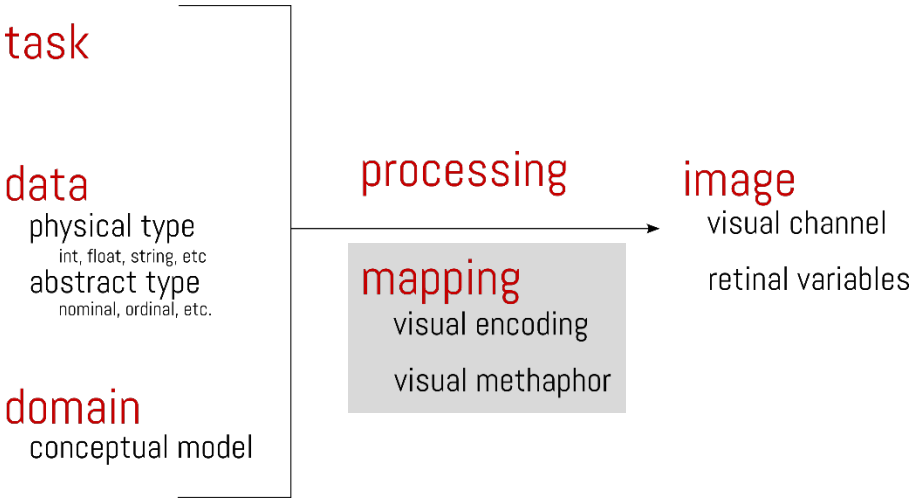


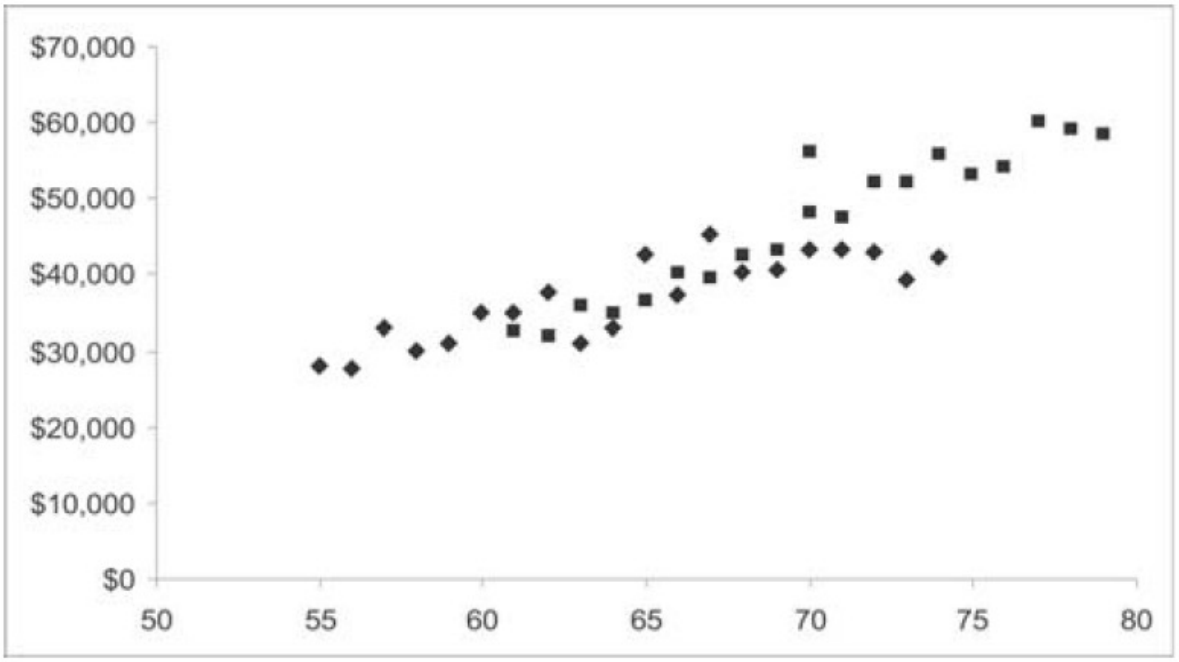
Lighting direction

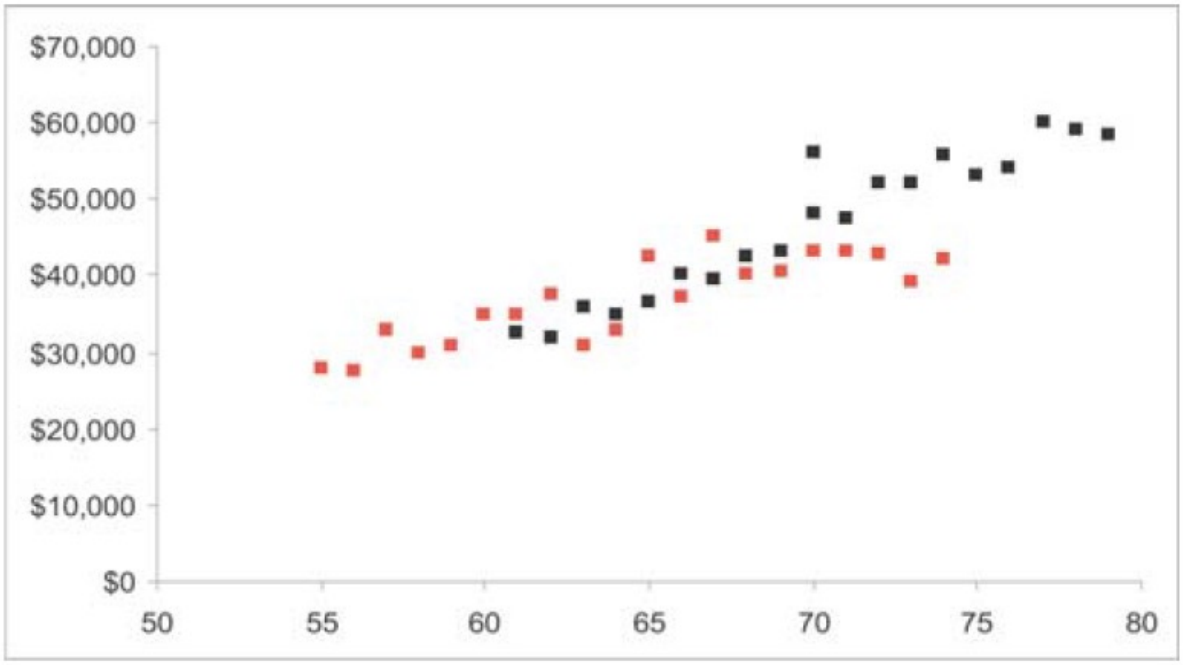


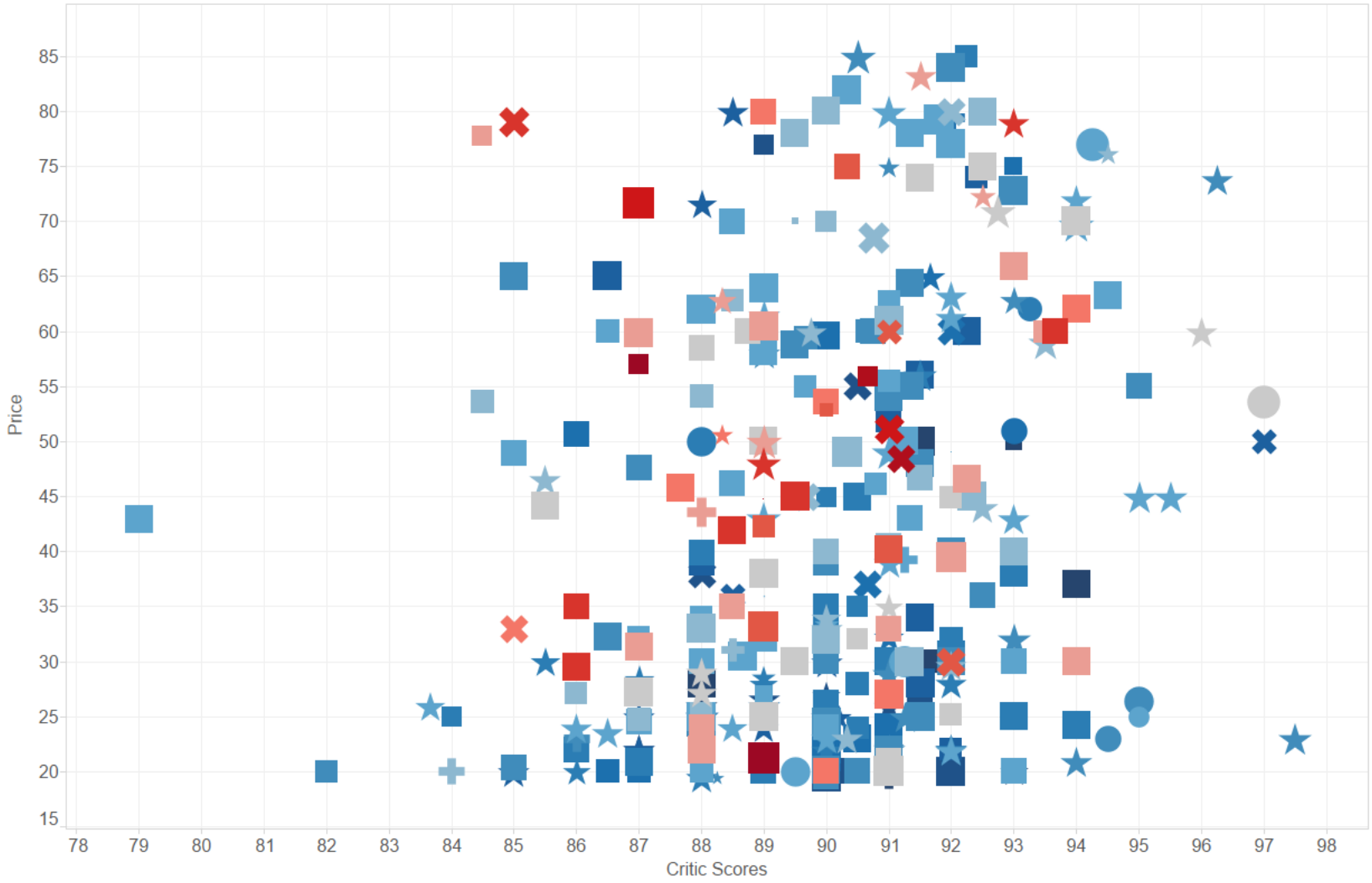
3D orientation

Mapping


task
data
physical type
int, float, string, etc
abstract type
nominal, ordinal, etc.
domain
conceptual model
processing
mapping
visual encoding
visual methaphor
image
visual channel
retinal variables

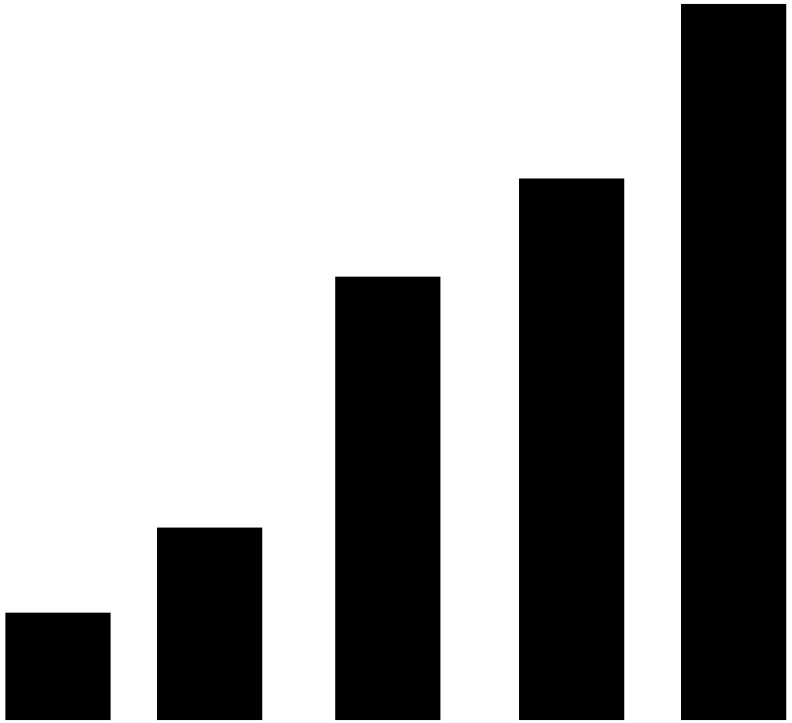

$70,000
$60,000
$50,000
$40,000
$30,000
$20,000
$10,000
$0
50
55
60
65
70
75
80


$70,000
$60,000
$50,000
$40,000
$30,000
$20,000
$10,000
$0
50
55
60
65
70
75
80

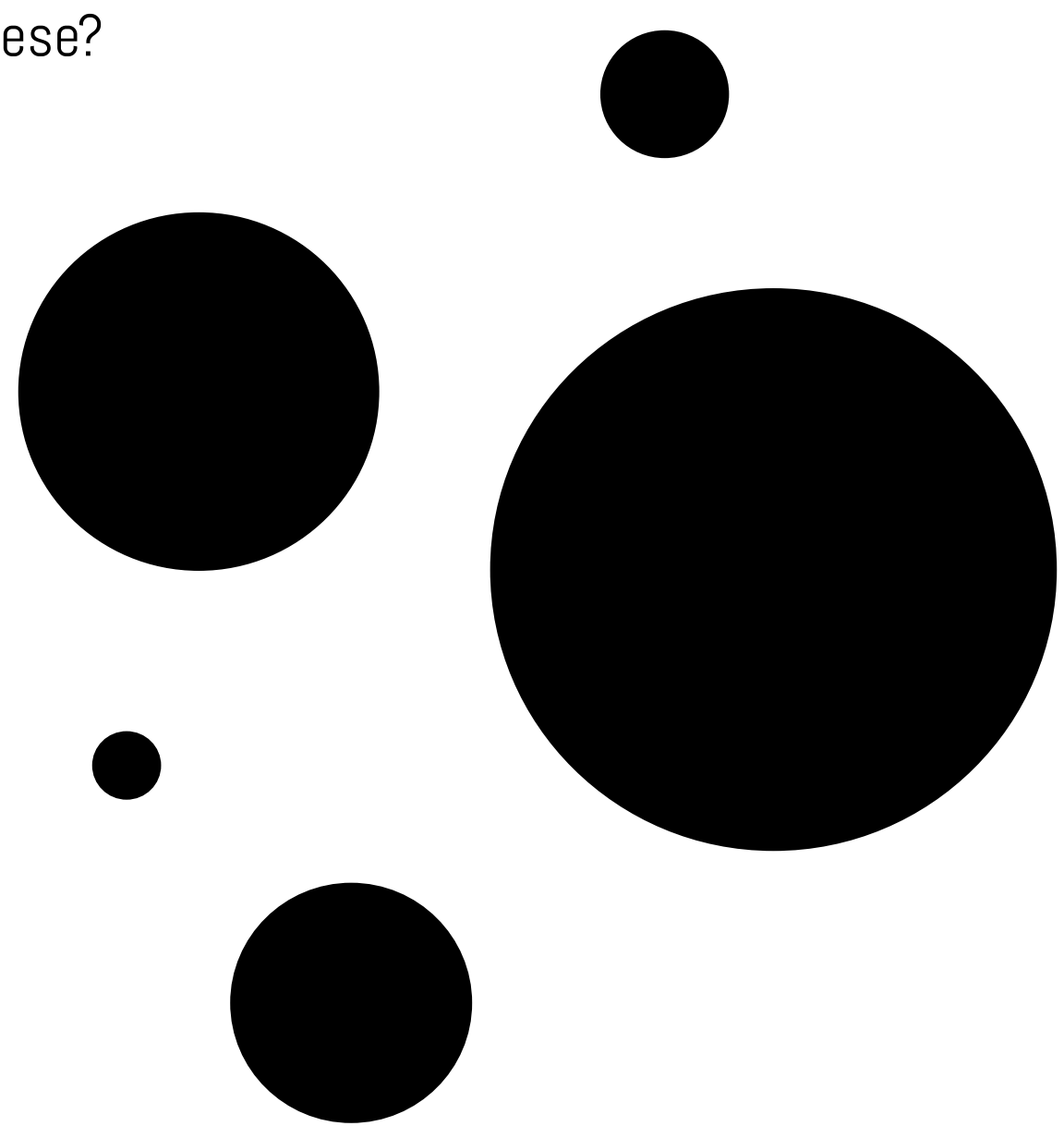


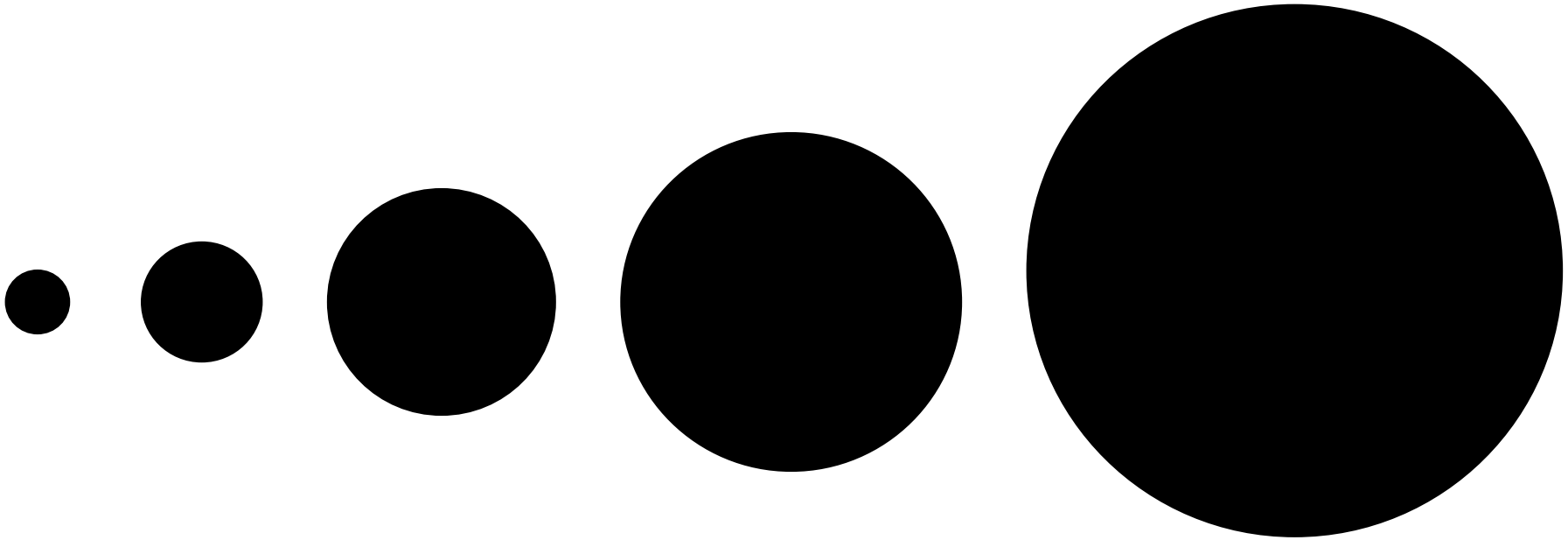
X: Critic Scores, **Y:** Price, **Size:** User Rating, **Color:** Vintage, **Shape:** Type

Can you order these?
(low -> high)



Can you order these?
(low -> high)

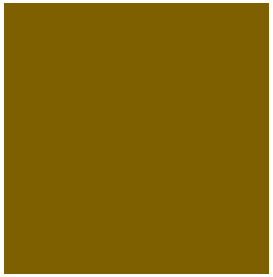


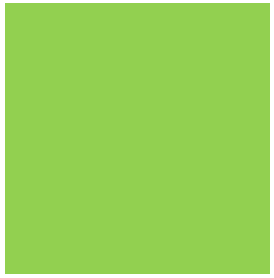


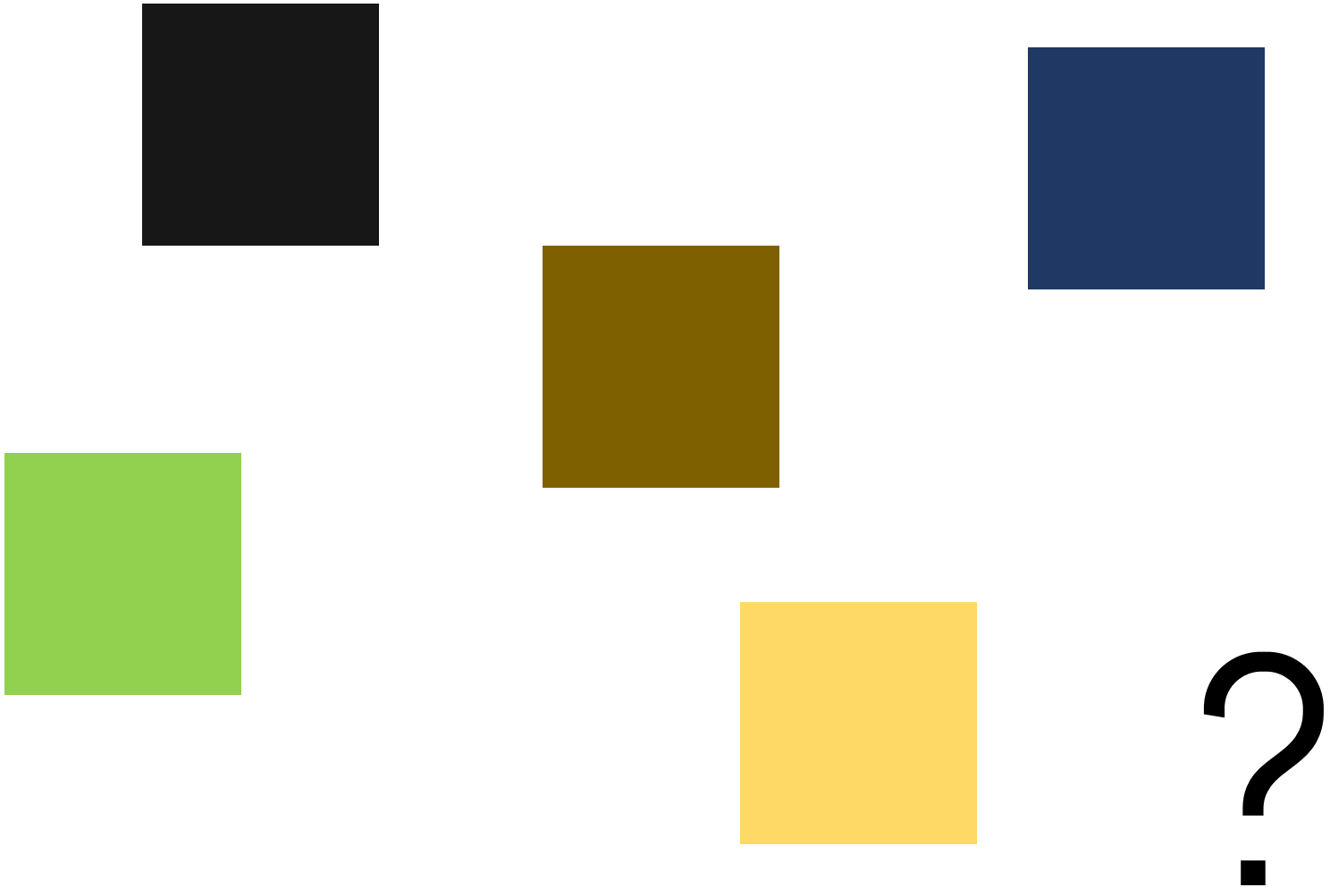
Can you order these?
(low -> high)

Can you order these?

(low -> high)







Nominal, Ordinal and Quantitative

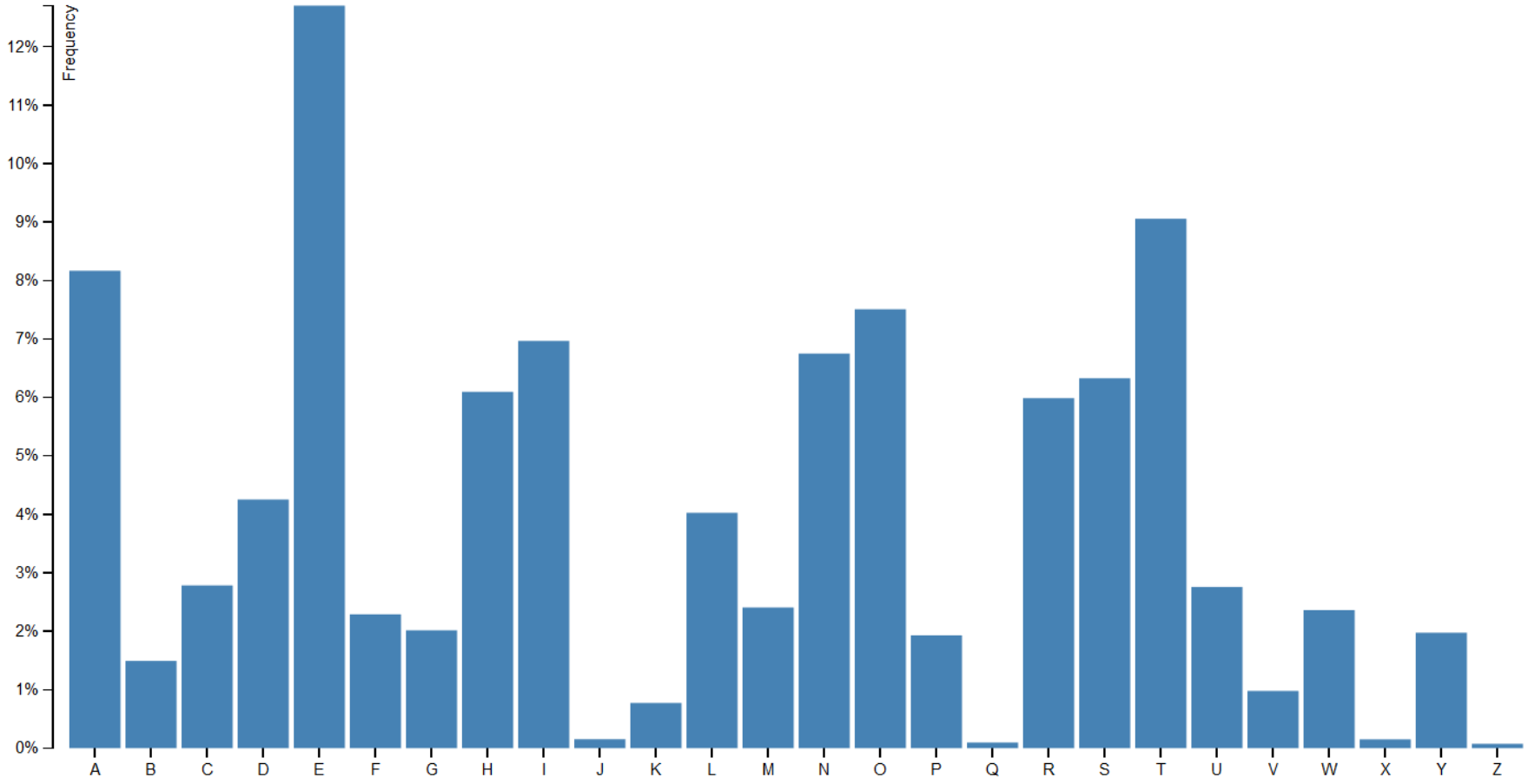
Position	N	O	Q
Size	N	O	Q
Value	N	O	Q
Texture	N	O	
Color	N		
Orientation	N		
Shape	N		

Nominal
Ordered
Quantitative

Some Principles & Guidelines

Bar chart

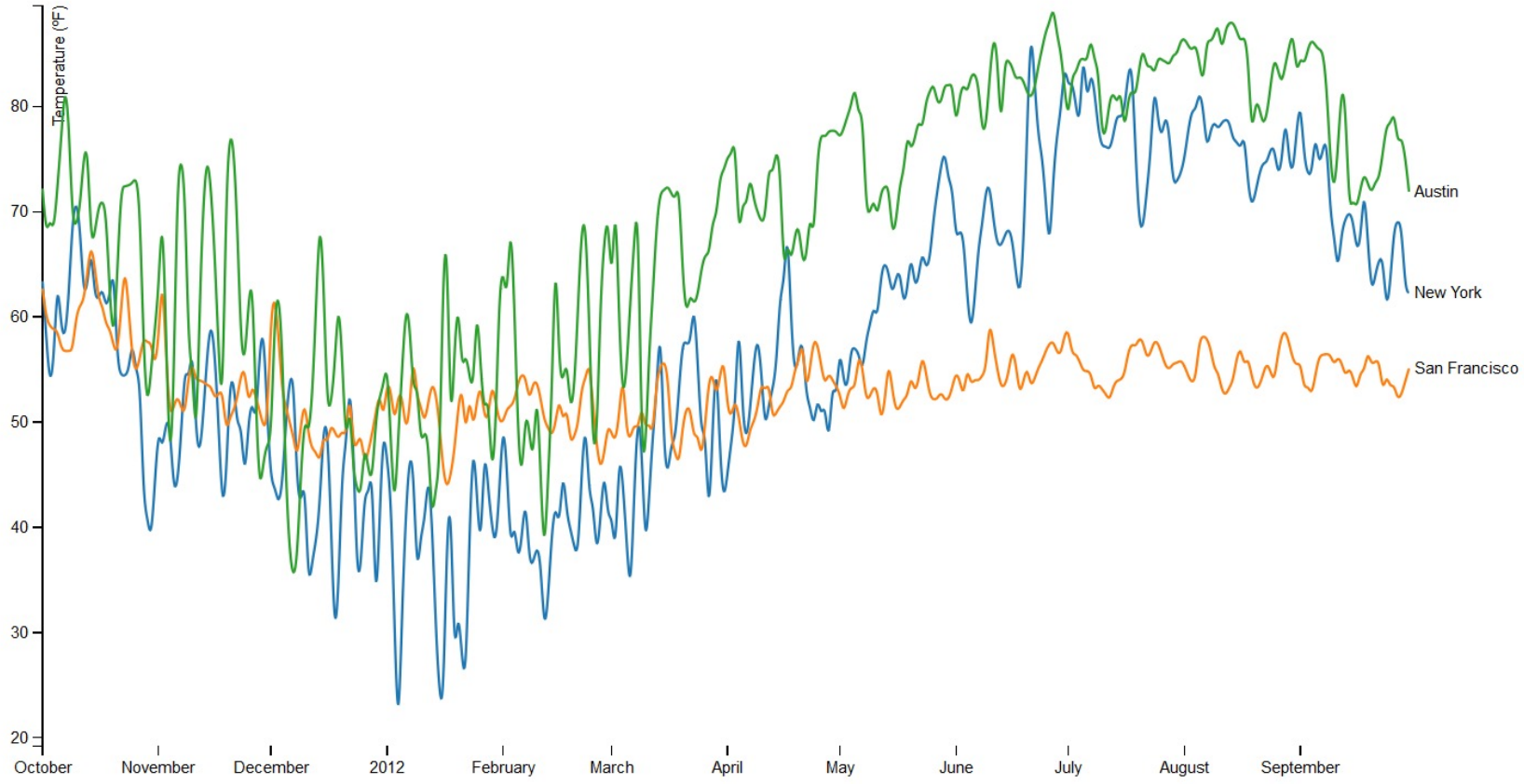
Display different quantities of single-variable data



https://github.com/mbostock/d3/wiki/Gallery

Line chart

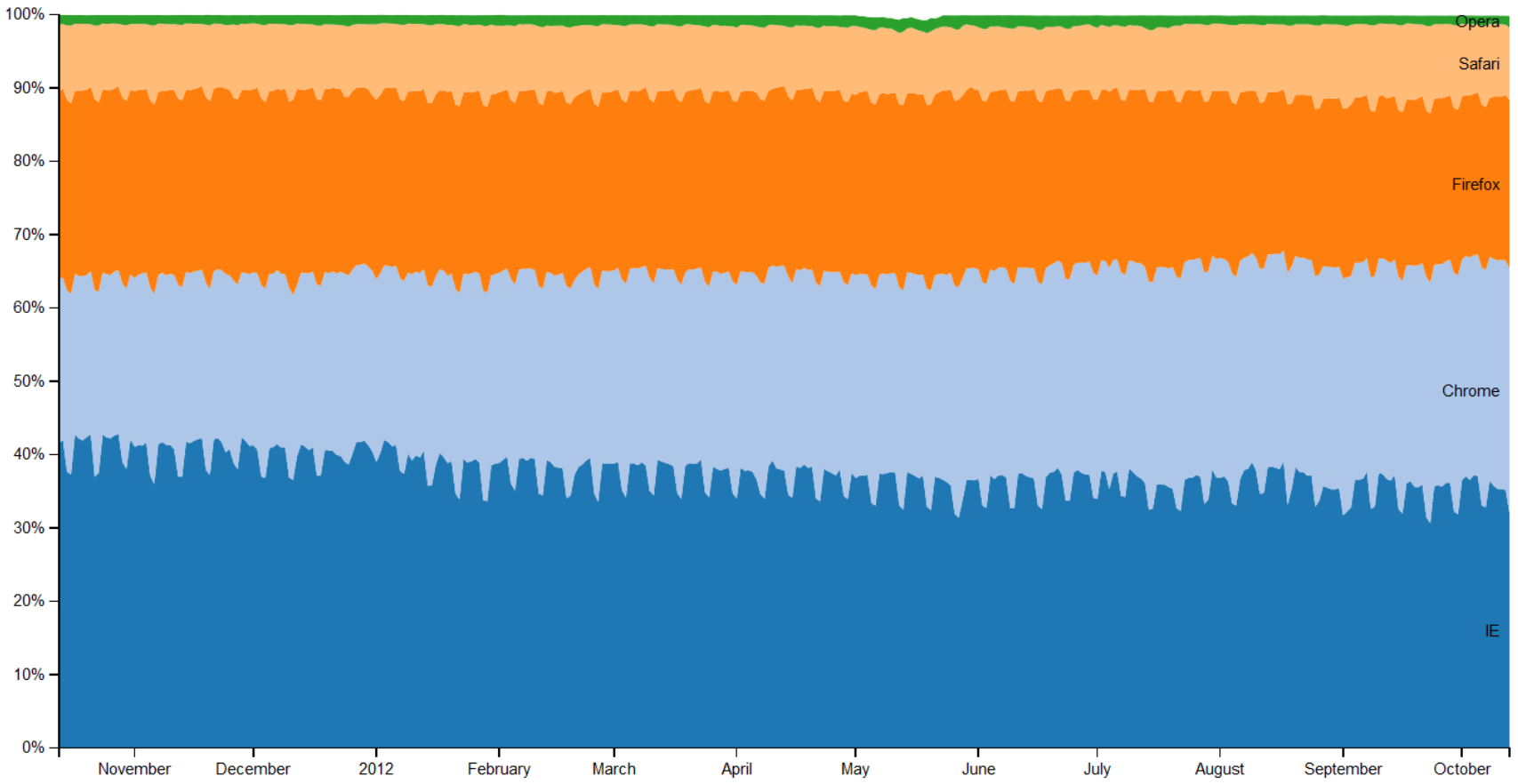
Display how a variable develops over time



https://github.com/mbostock/d3/wiki/Gallery

Stacked area chart

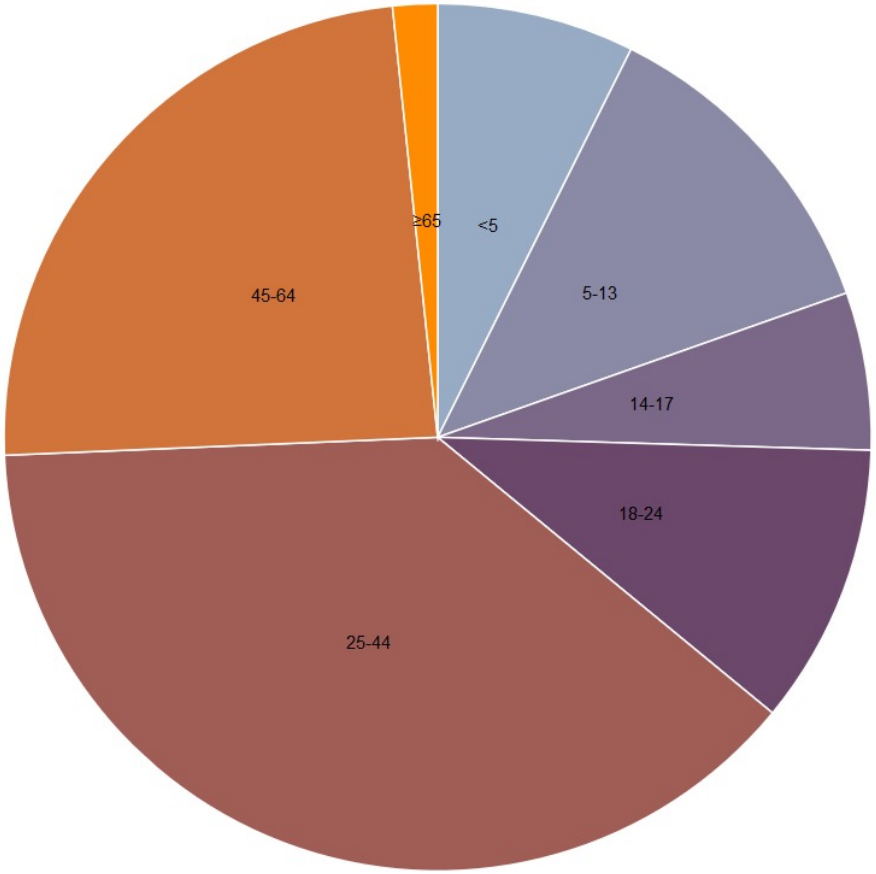
Display total of a variable over time



https://github.com/mbostock/d3/wiki/Gallery

Pie chart

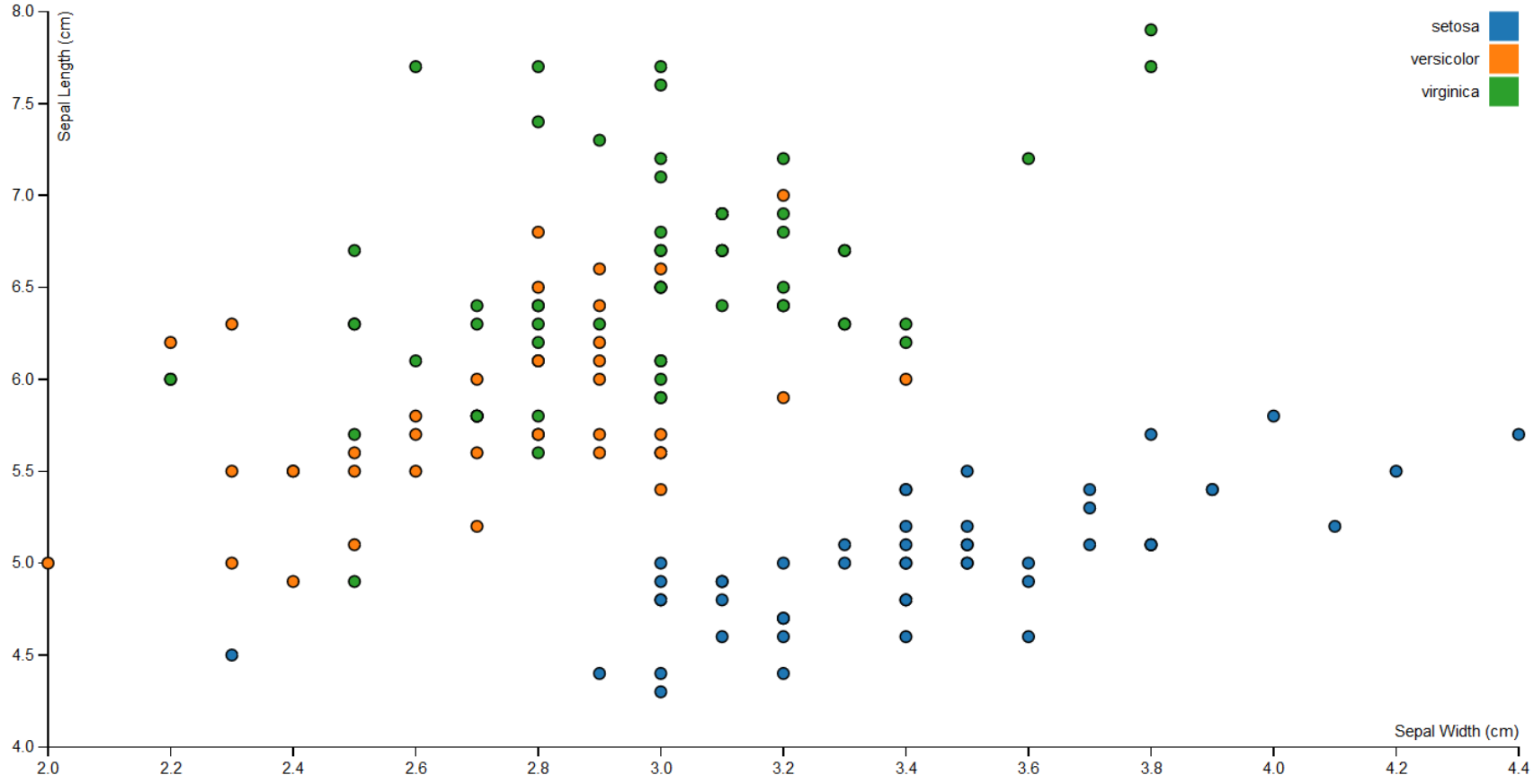
Display distribution of a variable



https://github.com/mbostock/d3/wiki/Gallery

Scatterplot

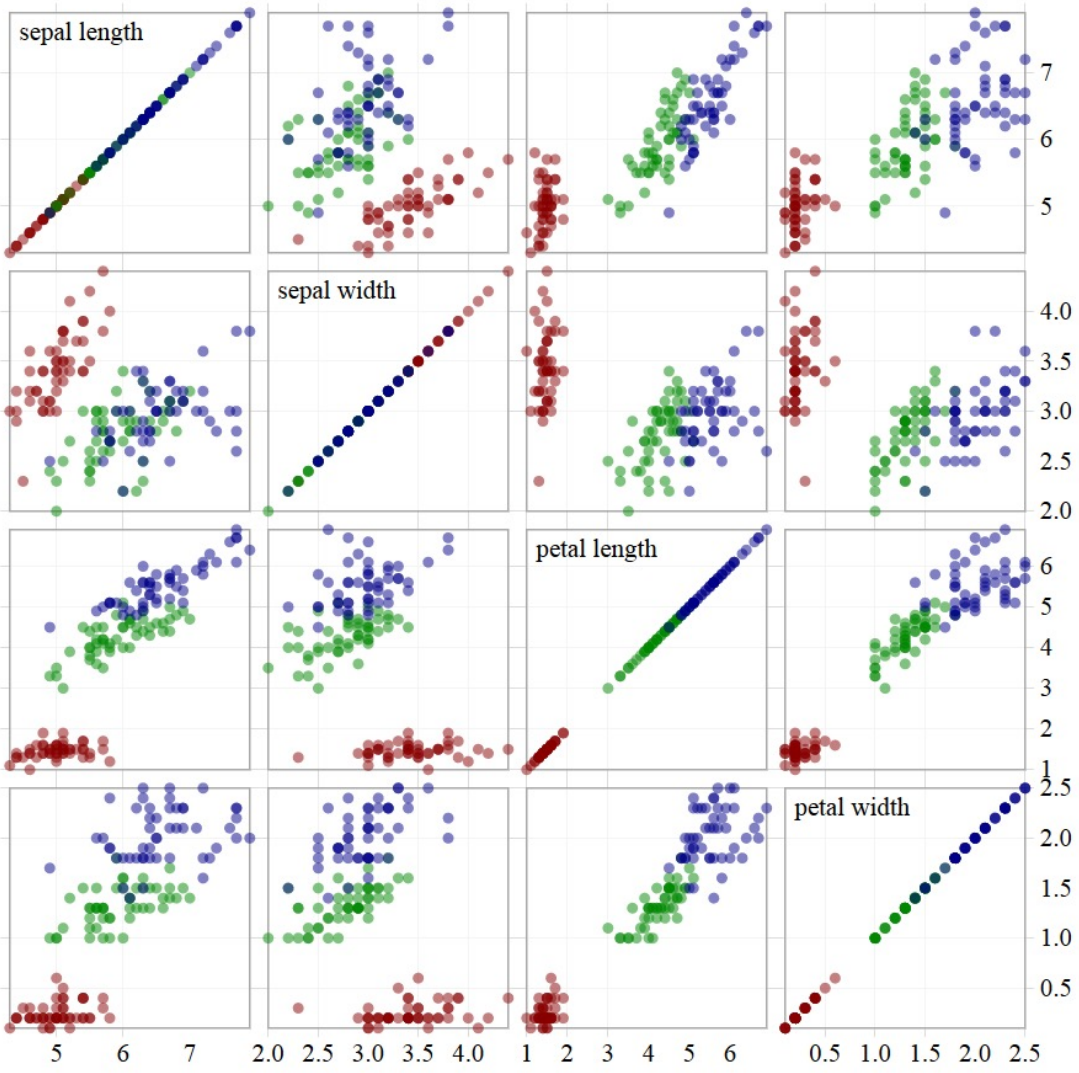
Display relationship between two variables



https://github.com/mbostock/d3/wiki/Gallery

Scatterplot matrix

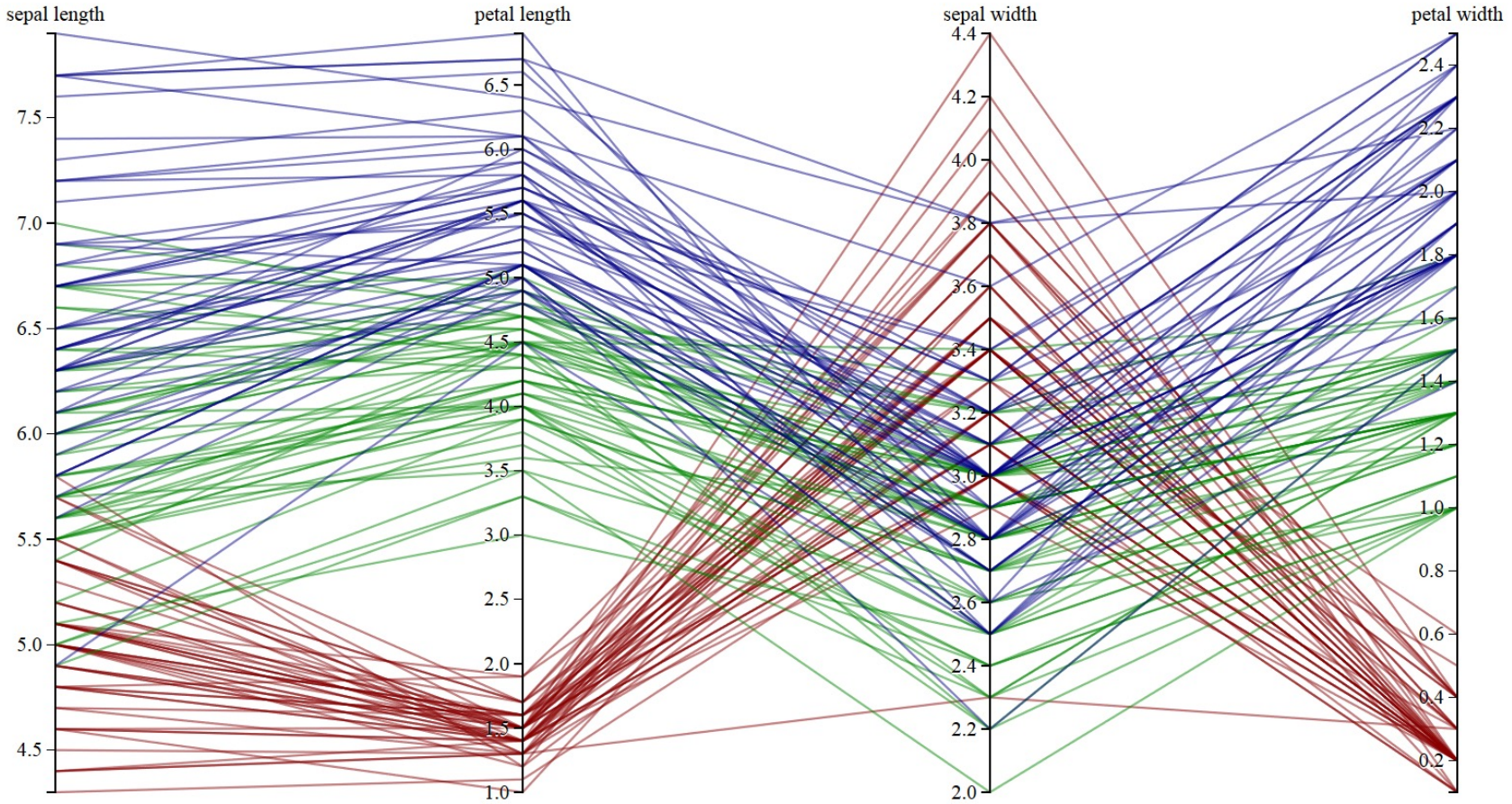
Display relationship between multiple variables



https://github.com/mbostock/d3/wiki/Gallery

Parallel coordinate plot

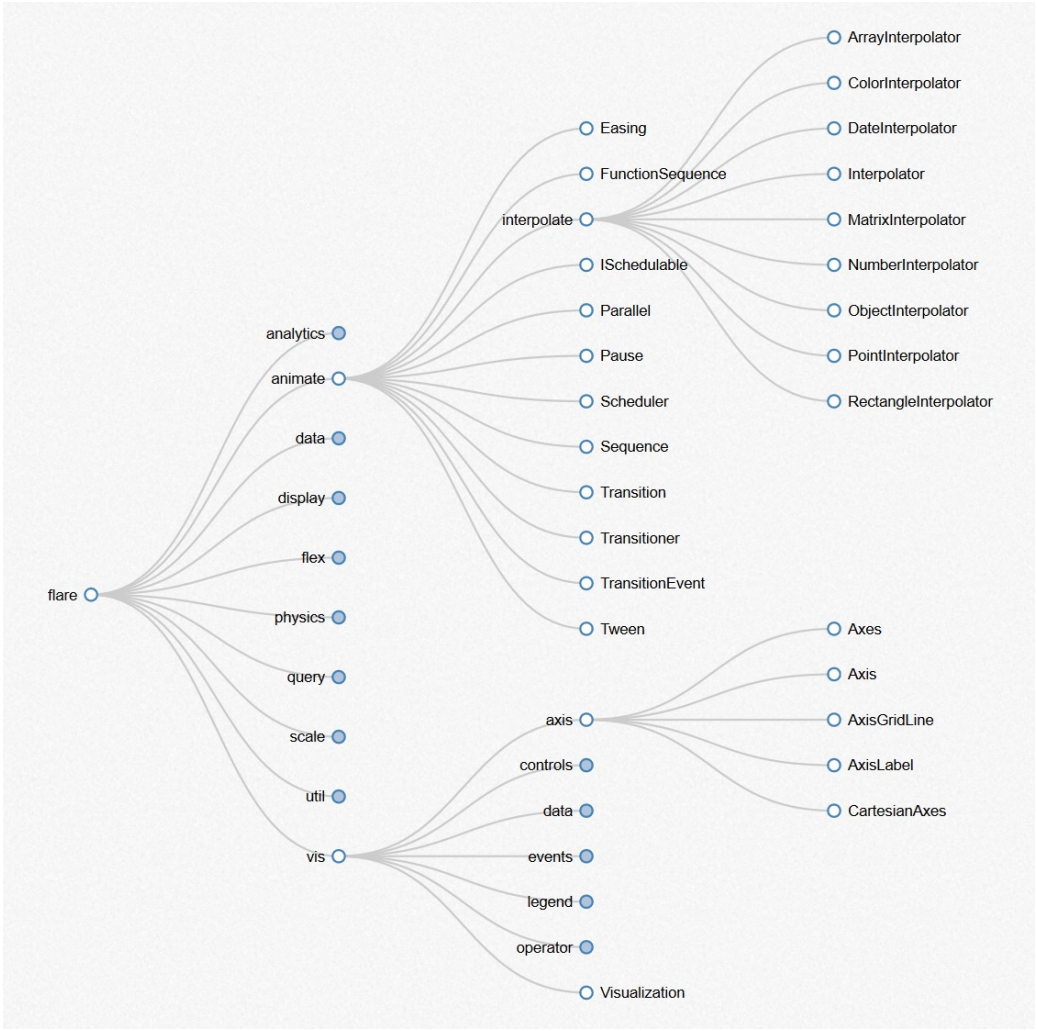
Display relationship between multiple variables



https://github.com/mbostock/d3/wiki/Gallery

Tree

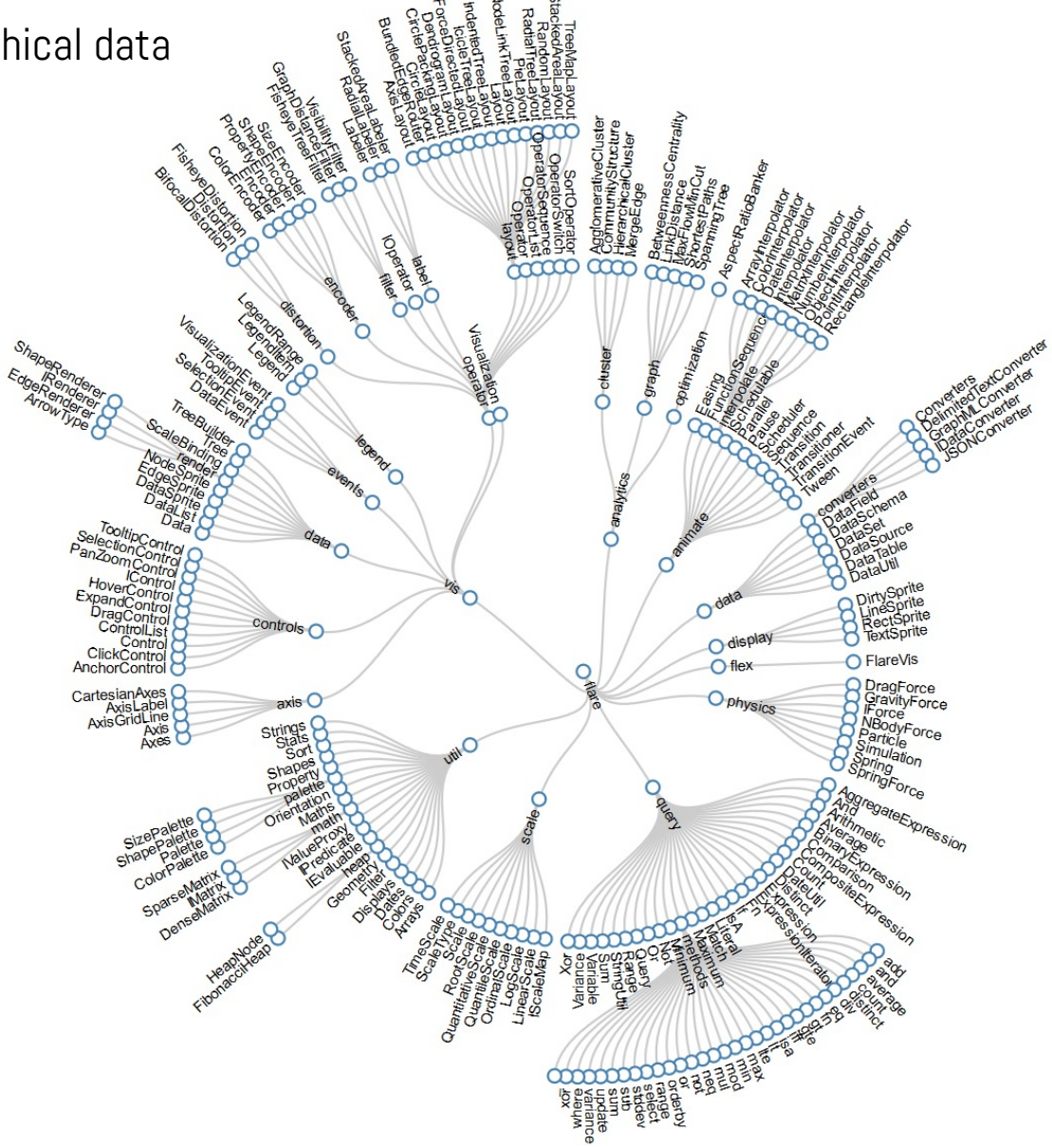
Display hierarchical data



https://github.com/mbostock/d3/wiki/Gallery

Radial tree

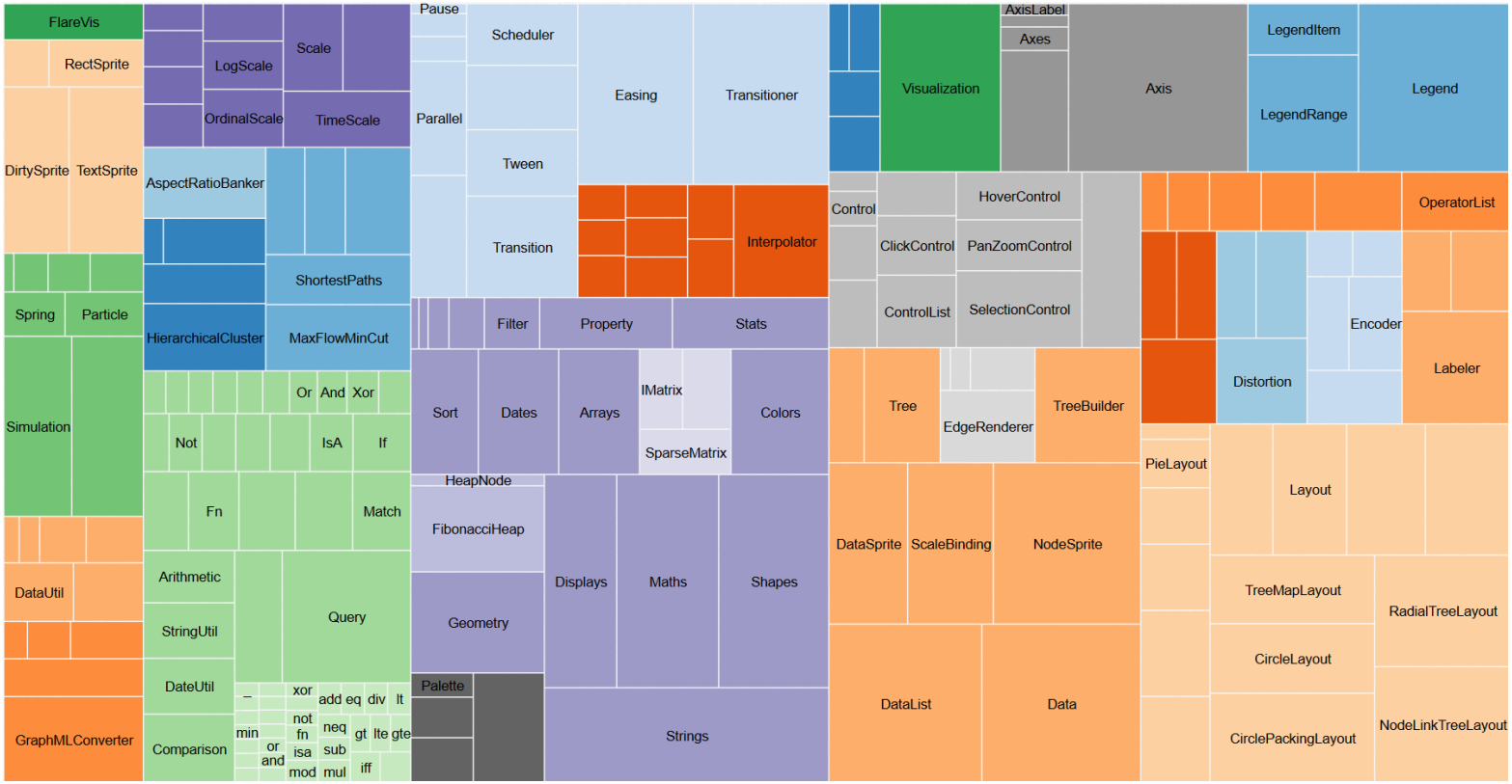
Display hierarchical data



https://github.com/mbostock/d3/wiki/Gallery

Tree map

Display hierarchical data



https://github.com/mbostock/d3/wiki/Gallery

Force-directed graph

Display graphs, networks, relationships

https://github.com/mbostock/d3/wiki/Gallery

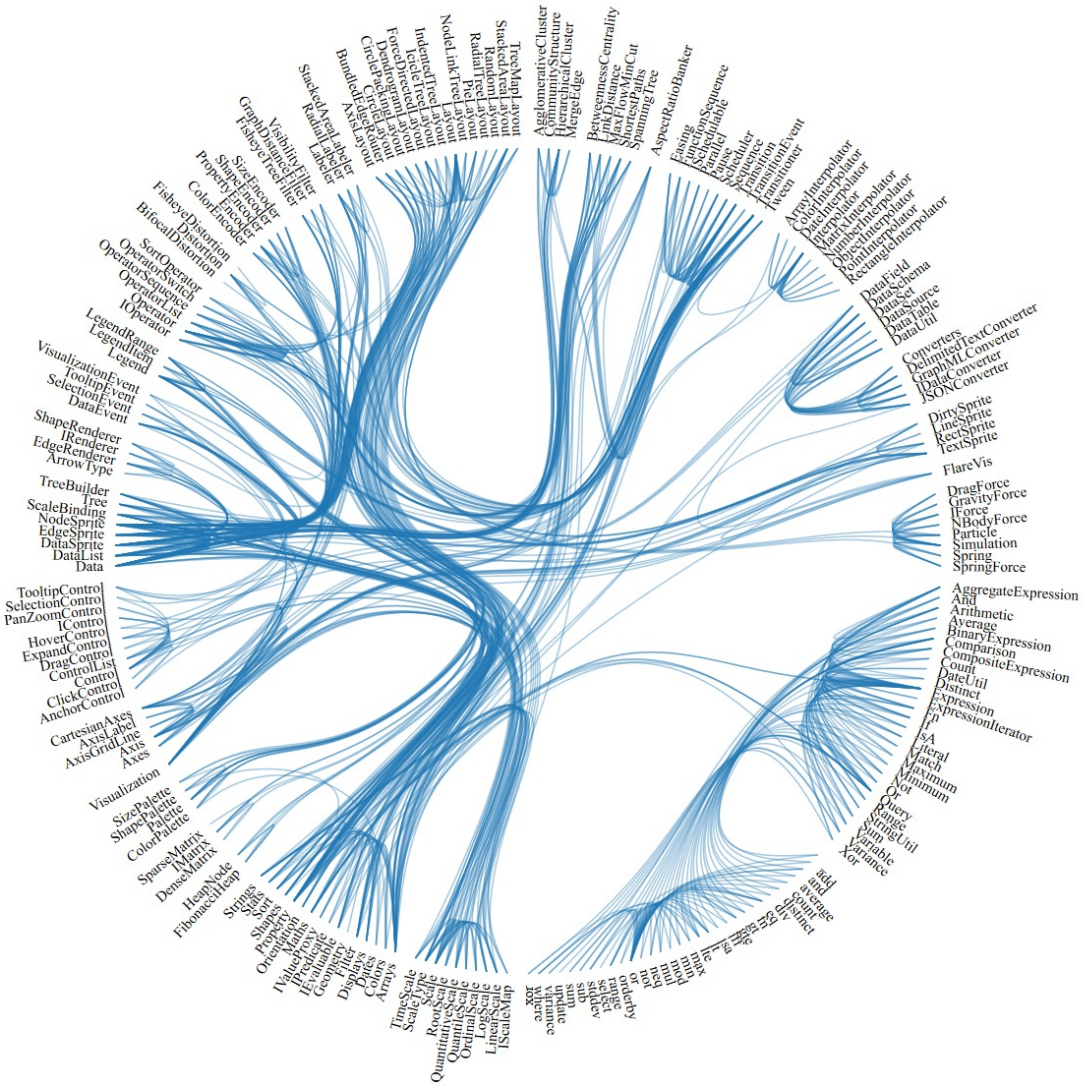
Radial graph

Display graphs, networks, relationships

https://github.com/mbostock/d3/wiki/Gallery

Radial graph

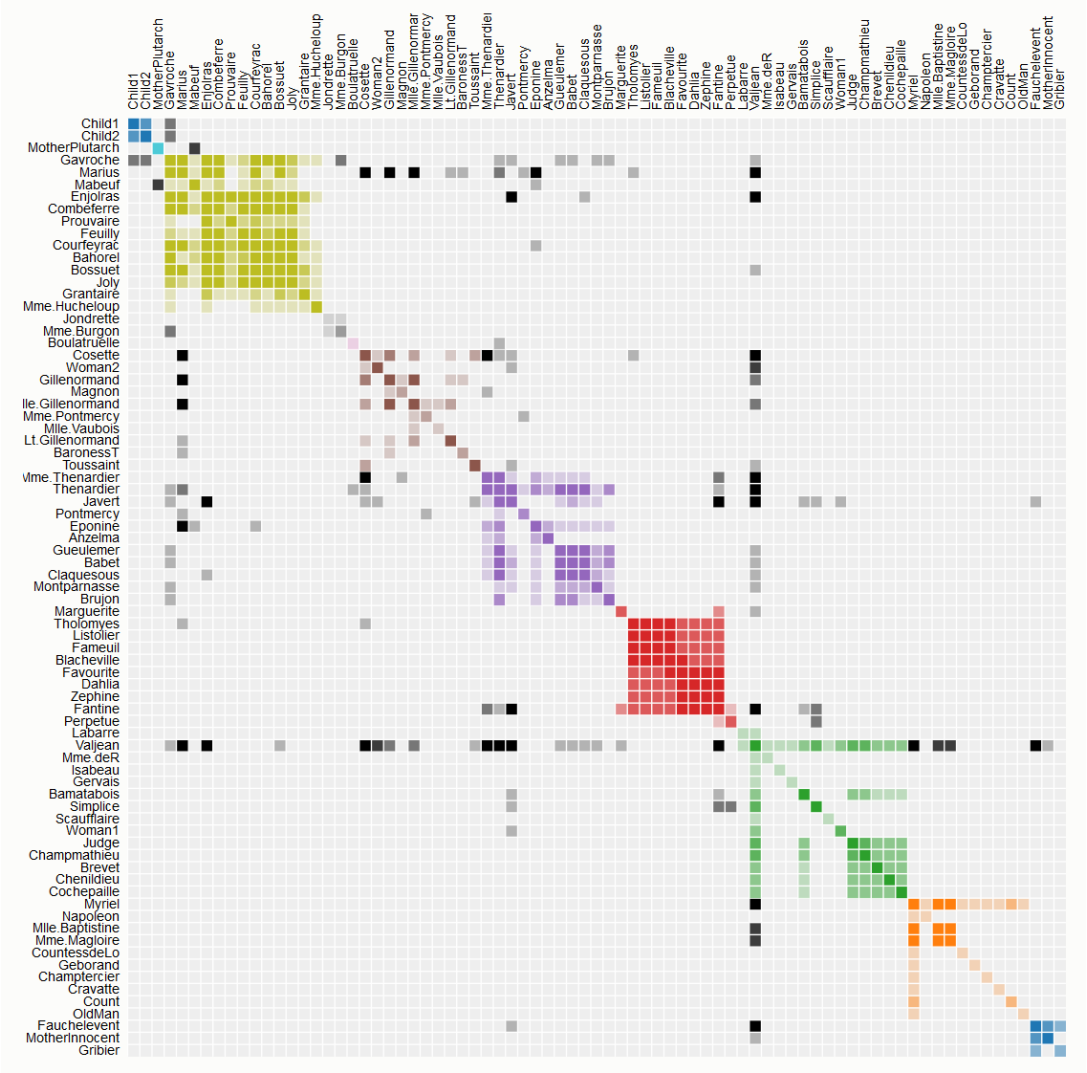
Display graphs, networks, relationships



https://github.com/mbostock/d3/wiki/Gallery

Co-occurrence matrix

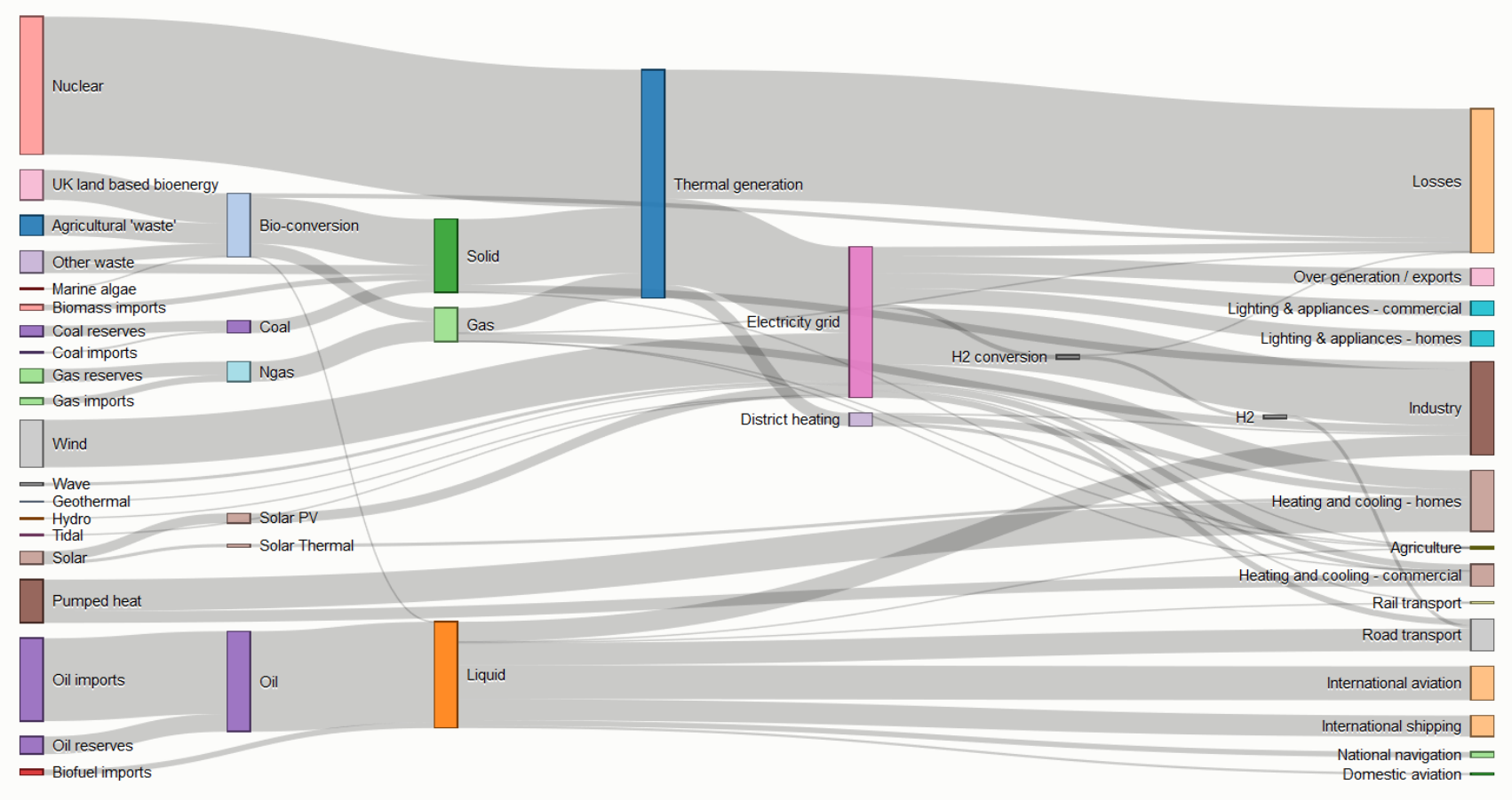
Display graphs, networks, relationships



https://github.com/mbostock/d3/wiki/Gallery

Sankey diagram

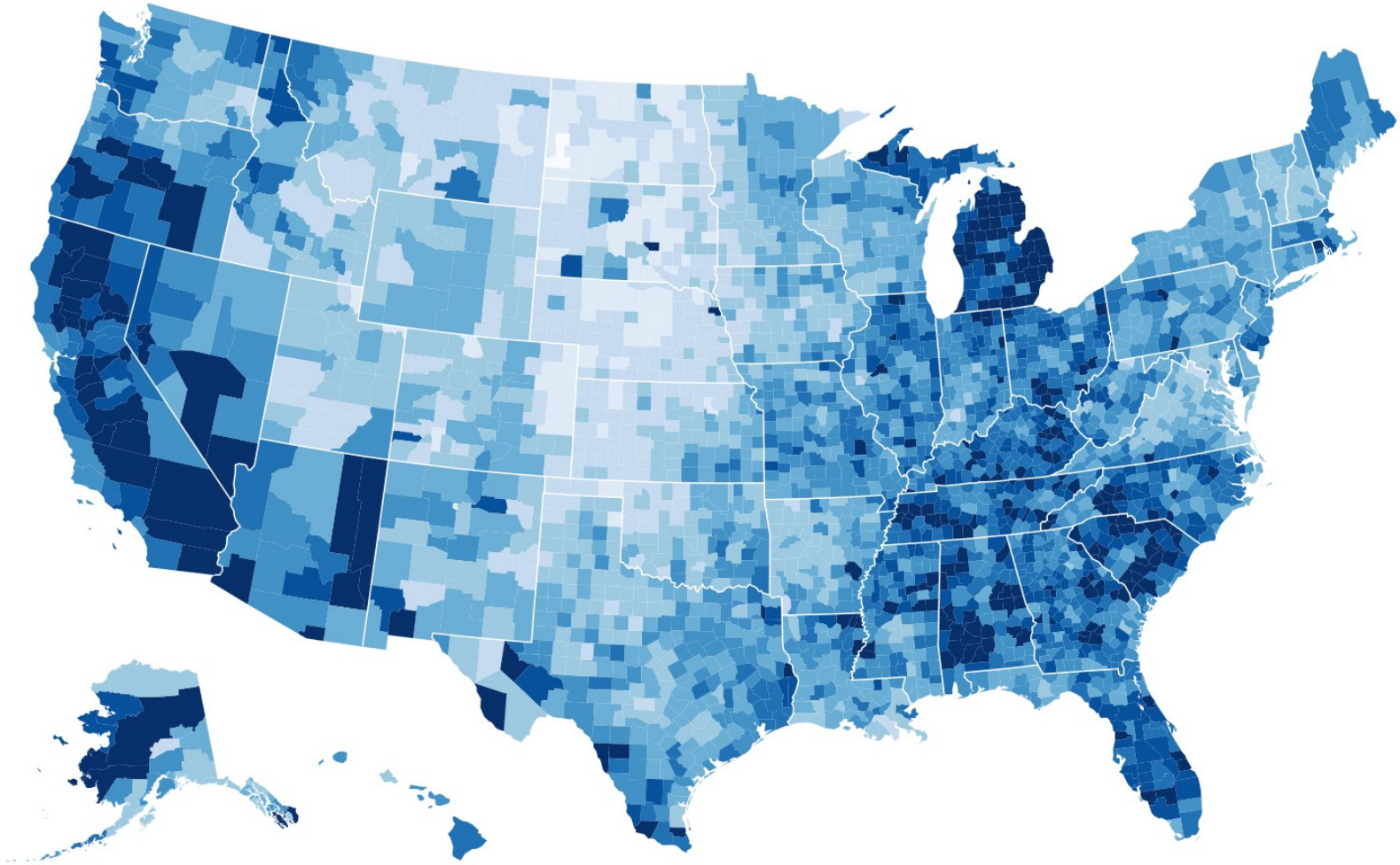
Display flow amongst entities



https://github.com/mbostock/d3/wiki/Gallery

Maps

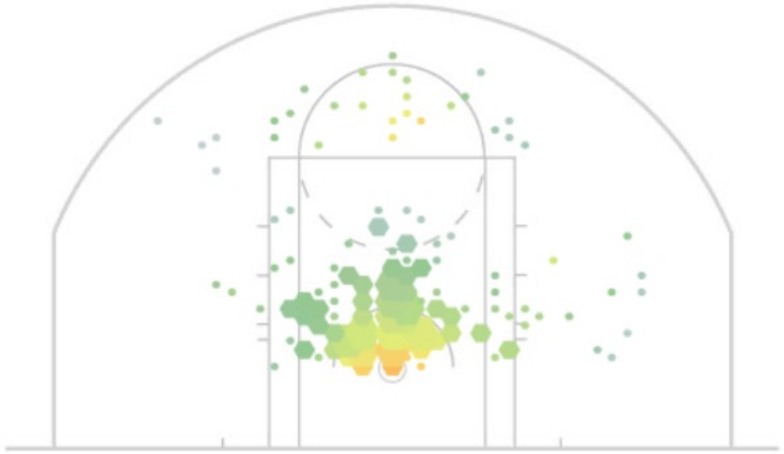
Display a spatial variable



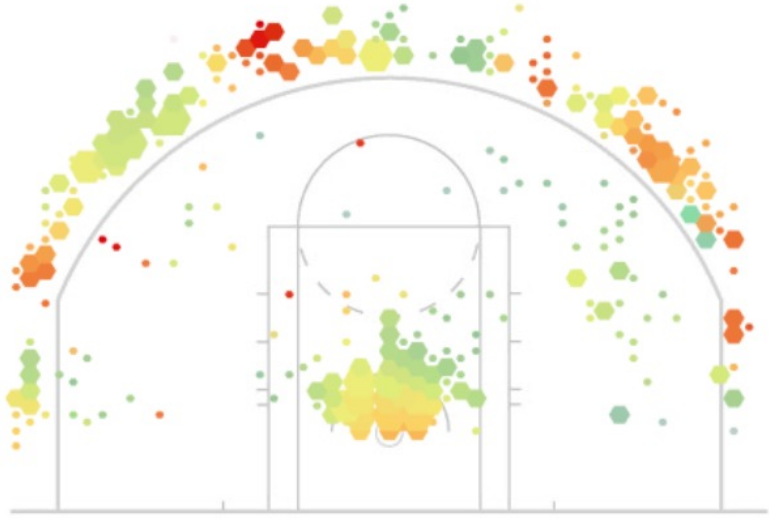
https://github.com/mbostock/d3/wiki/Gallery

Maps

Display a spatial variable



Kendrick Perkins



James Harden

[New York Times]

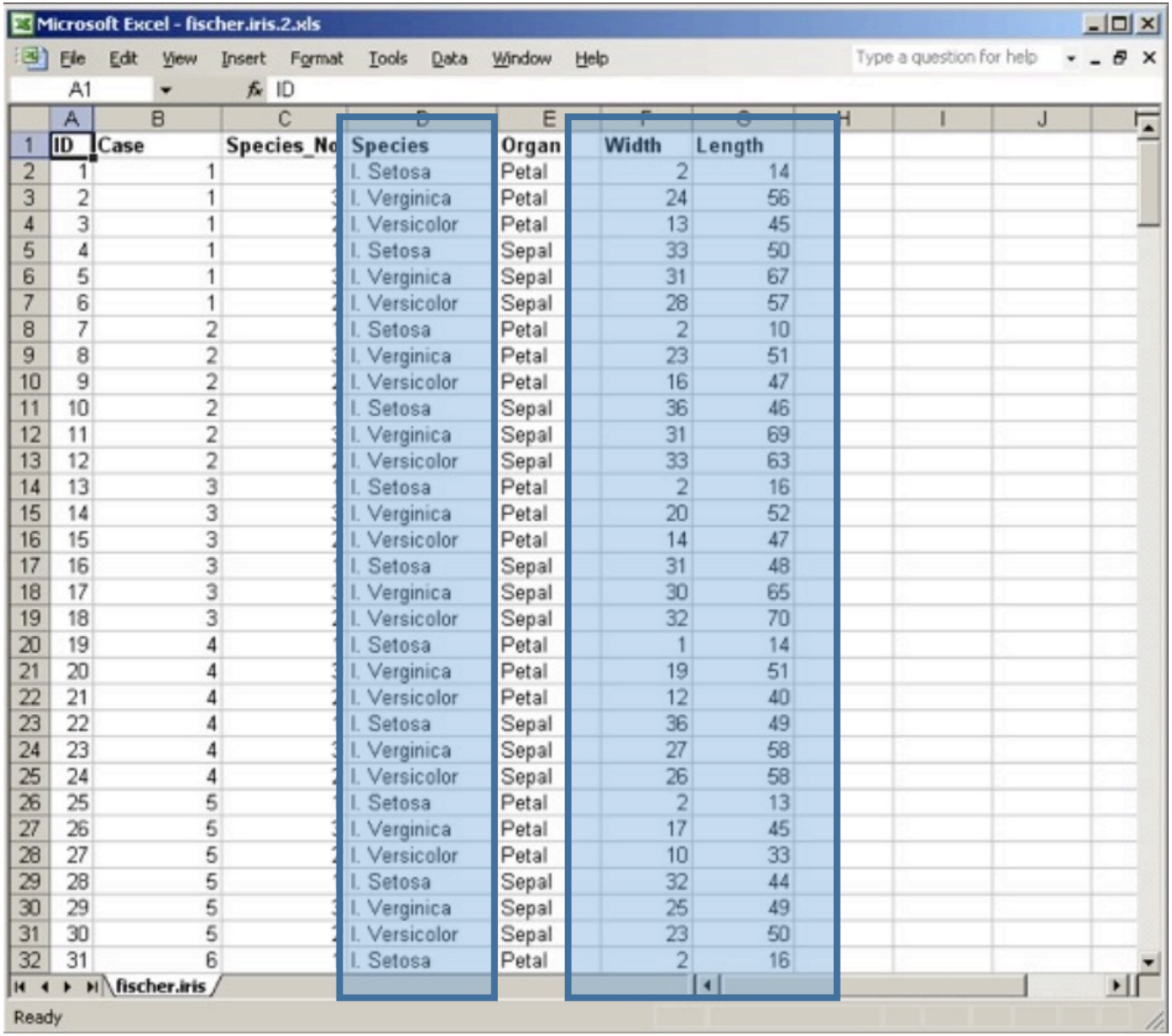
Expressiveness

A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness

A visualization is more effective than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

[Mackinlay, APT, 1986]

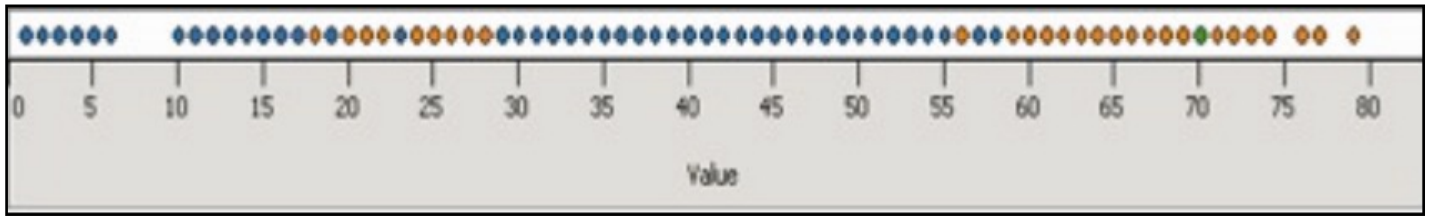


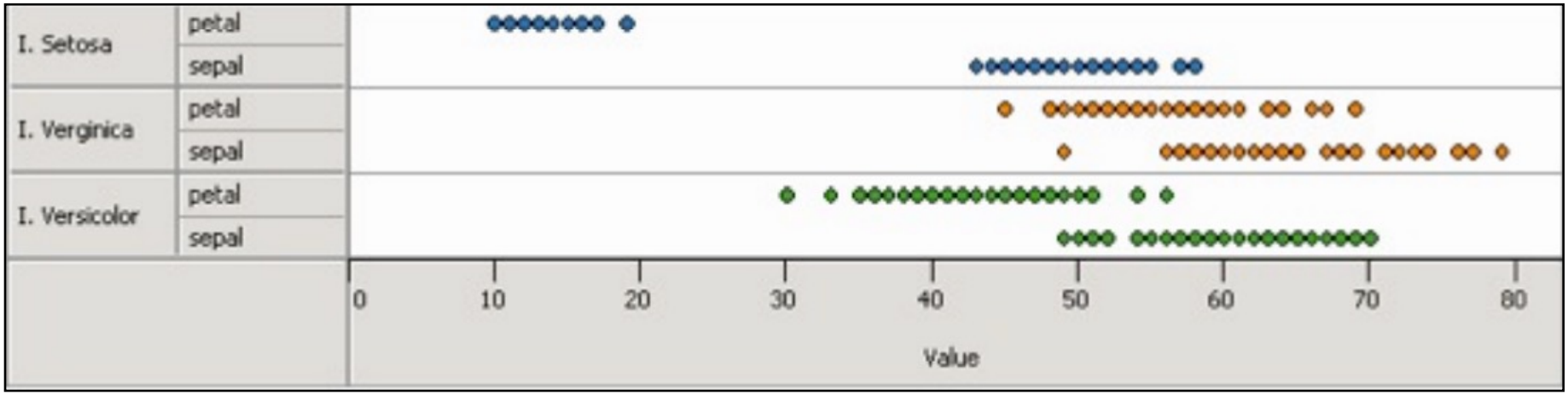
ID	Case	Species_No	Species	Organ	Width	Length
1	1	1	I. Setosa	Petal	2	14
2	1	3	I. Verginica	Petal	24	56
3	1	2	I. Versicolor	Petal	13	45
4	1	1	I. Setosa	Sepal	33	50
5	1	3	I. Verginica	Sepal	31	67
6	1	2	I. Versicolor	Sepal	28	57
7	2	1	I. Setosa	Petal	2	10
8	2	3	I. Verginica	Petal	23	51
9	2	2	I. Versicolor	Petal	16	47
10	2	1	I. Setosa	Sepal	36	46
11	2	3	I. Verginica	Sepal	31	69
12	2	2	I. Versicolor	Sepal	33	63
13	3	1	I. Setosa	Petal	2	16
14	3	3	I. Verginica	Petal	20	52
15	3	2	I. Versicolor	Petal	14	47
16	3	1	I. Setosa	Sepal	31	48
17	3	3	I. Verginica	Sepal	30	65
18	3	2	I. Versicolor	Sepal	32	70
19	4	1	I. Setosa	Petal	1	14
20	4	3	I. Verginica	Petal	19	51
21	4	2	I. Versicolor	Petal	12	40
22	4	1	I. Setosa	Sepal	36	49
23	4	3	I. Verginica	Sepal	27	58
24	4	2	I. Versicolor	Sepal	26	58
25	5	1	I. Setosa	Petal	2	13
26	5	3	I. Verginica	Petal	17	45
27	5	2	I. Versicolor	Petal	10	33
28	5	1	I. Setosa	Sepal	32	44
29	5	3	I. Verginica	Sepal	25	49
30	5	2	I. Versicolor	Sepal	23	50
31	6	1	I. Setosa	Petal	2	16

[Fisher, 1936]

Example 1: Cannot express the facts

A one-to-many (1->N) relation cannot be expressed in a single horizontal dot plot because multiple tuples are mapped to the same position.

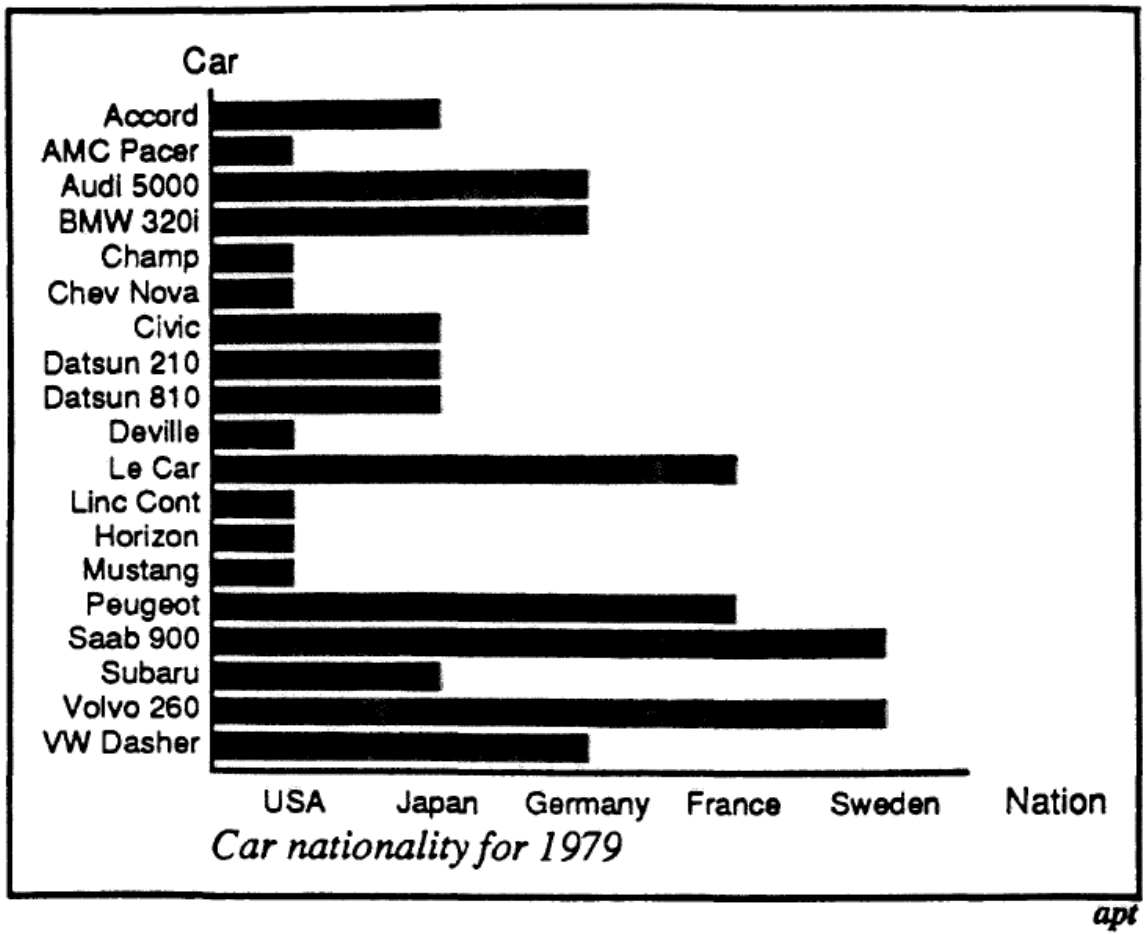


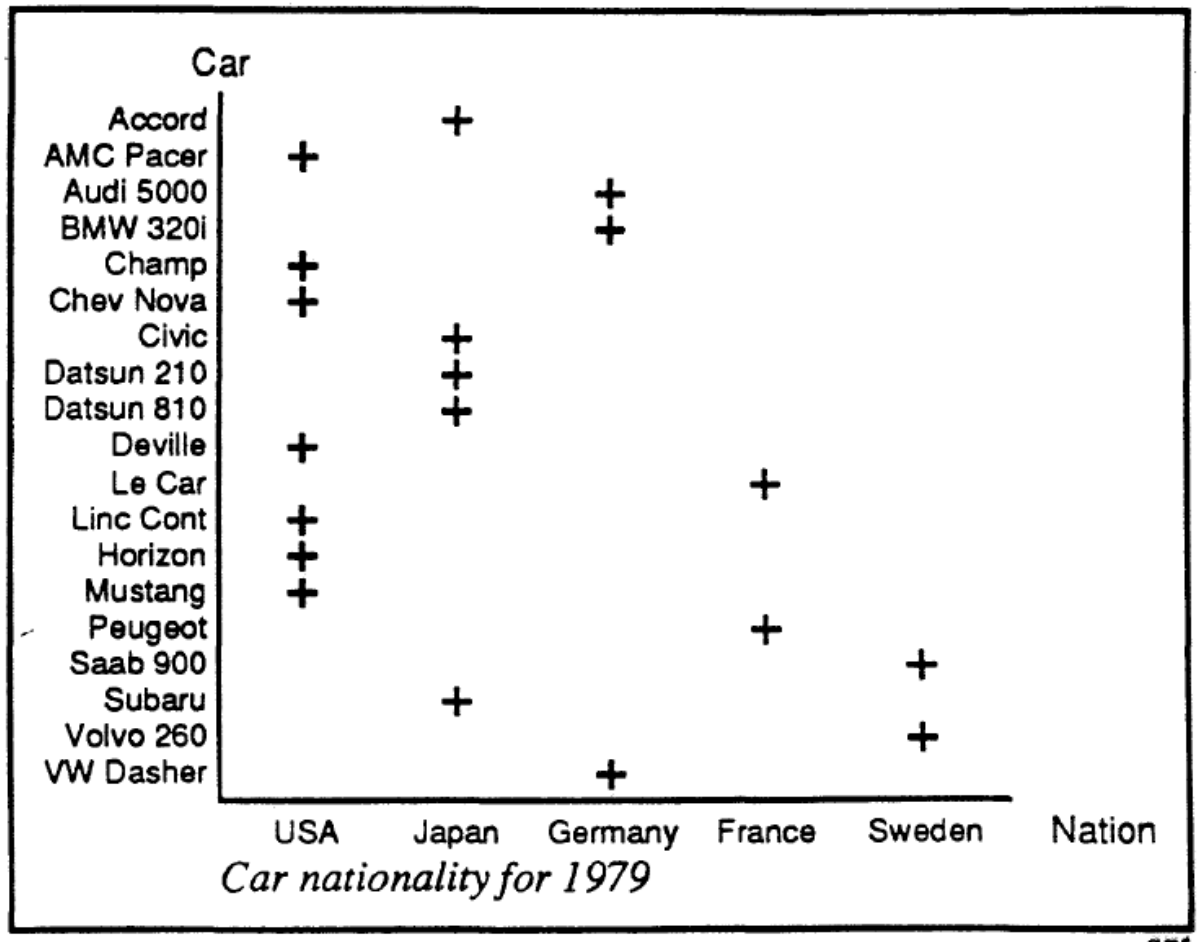


Example 2: Express facts not in the data

A length is interpreted as a quantitative value;

Length of bar says something untrue about Nominal data

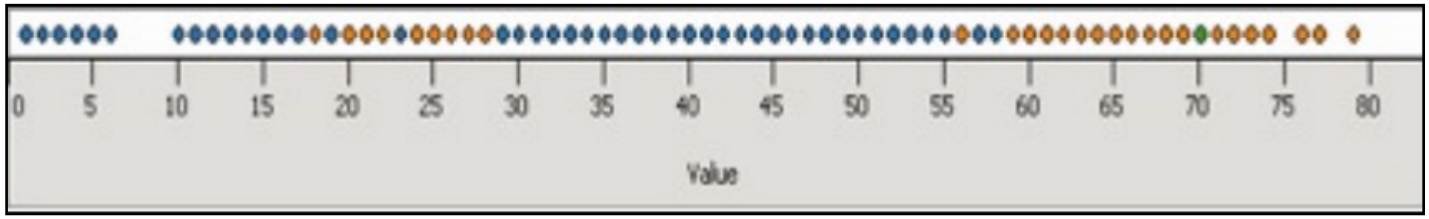


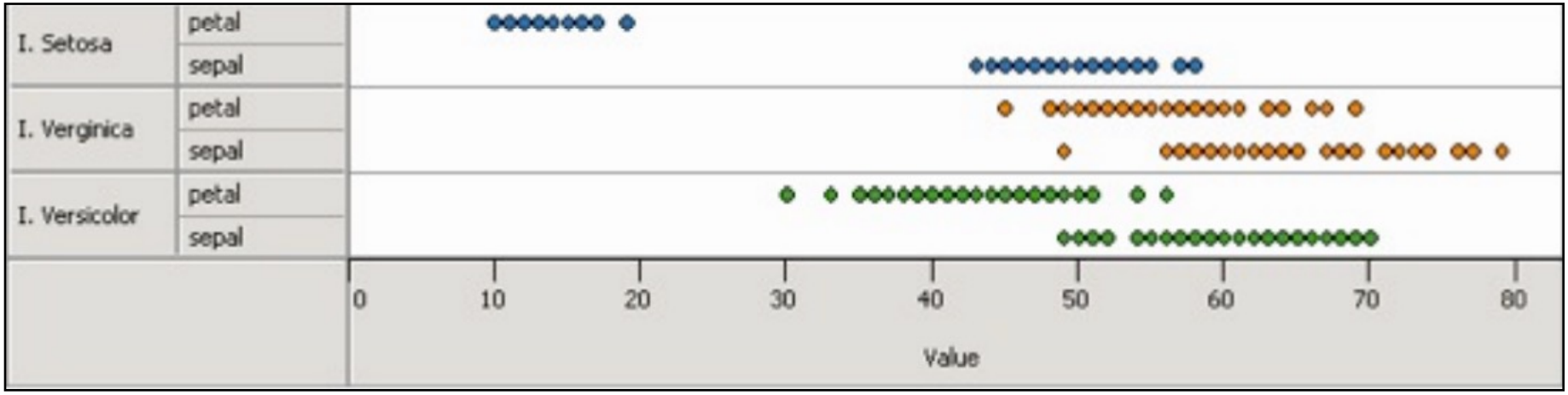


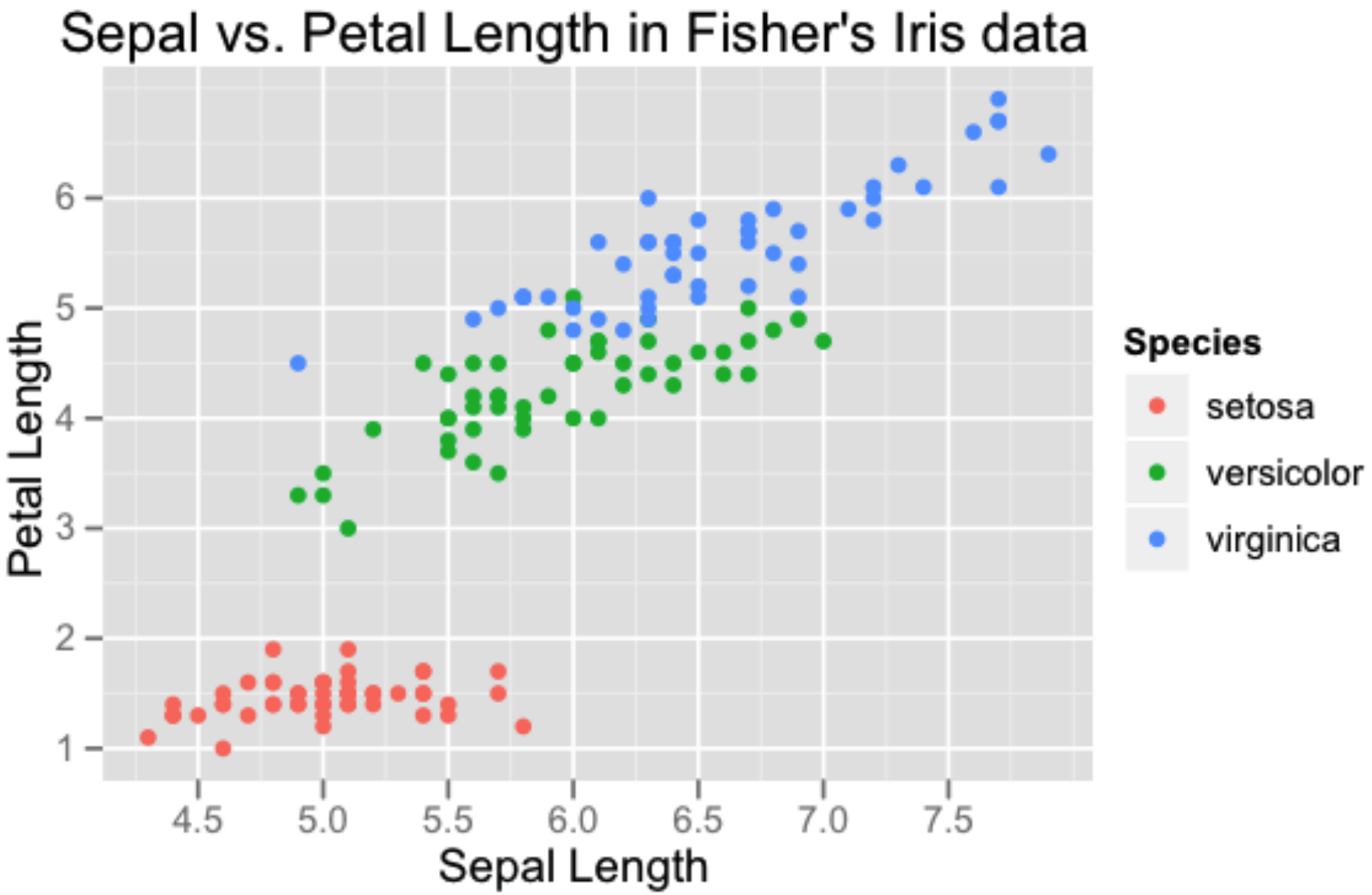
Car nationality for 1979

apt

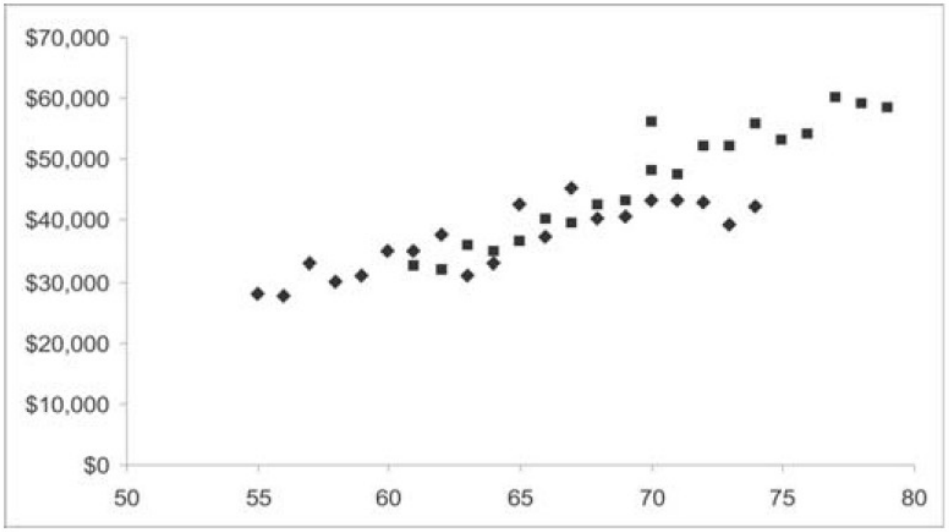
Example 4: Effectiveness

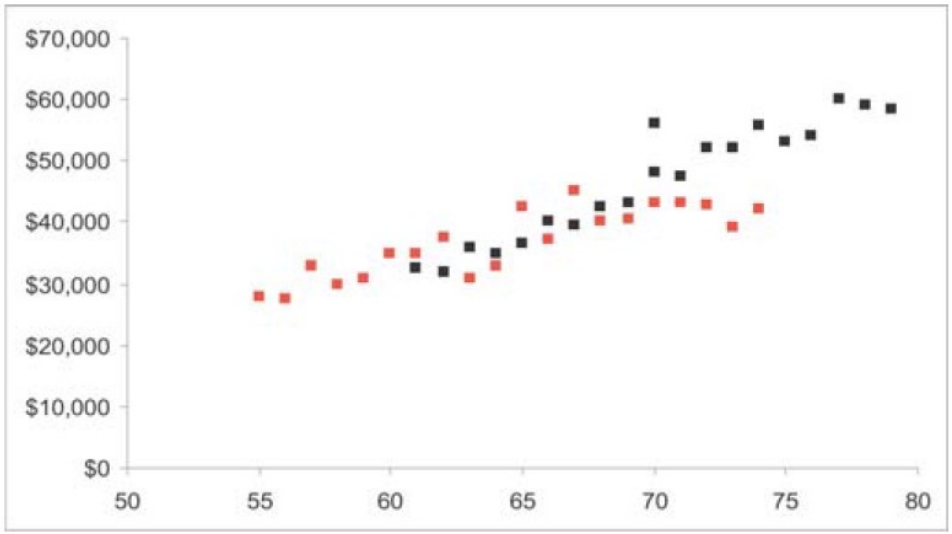


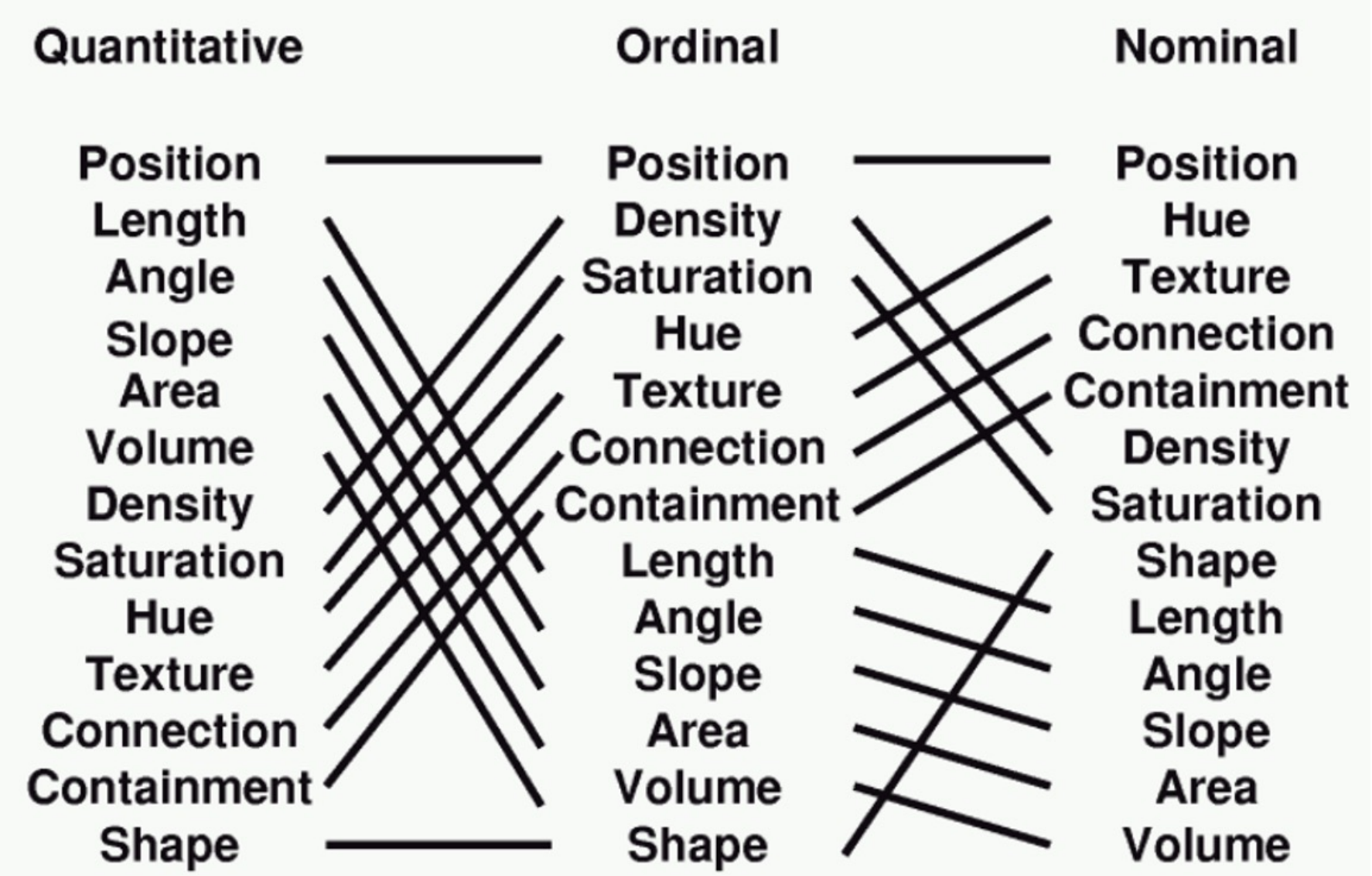



Sepal vs. Petal Length in Fisher's Iris data
Petal Length
Sepal Length
1
2
3
4
5
6
4.5
5.0
5.5
6.0
6.5
7.0
7.5
Species
setosa
versicolor
virginica

Example 3: Effectiveness





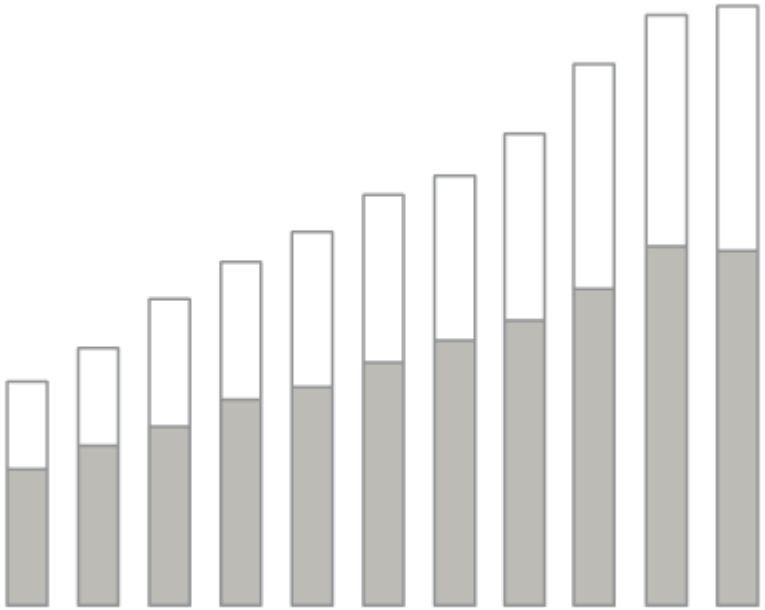

Quantitative
Ordinal
Nominal
Position
Length
Angle
Slope
Area
Volume
Density
Saturation
Hue
Texture
Connection
Containment
Shape
Position
Density
Saturation
Hue
Texture
Connection
Containment
Length
Angle
Slope
Area
Volume
Shape
Position
Hue
Texture
Connection
Containment
Density
Saturation
Shape
Length
Angle
Slope
Area
Volume

[Mackinlay]

Tufte's Rules

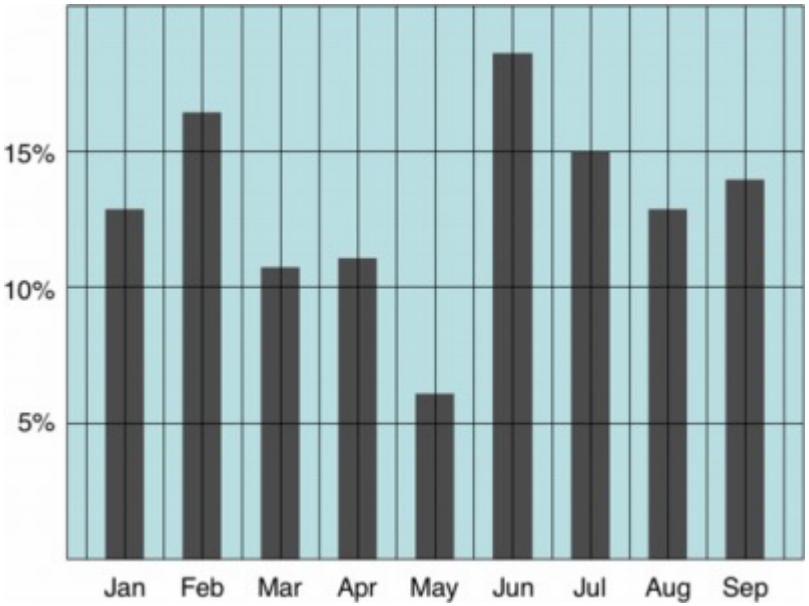
[http://www.sealthreinhold.com/tuftes-rules/rule_three.php]

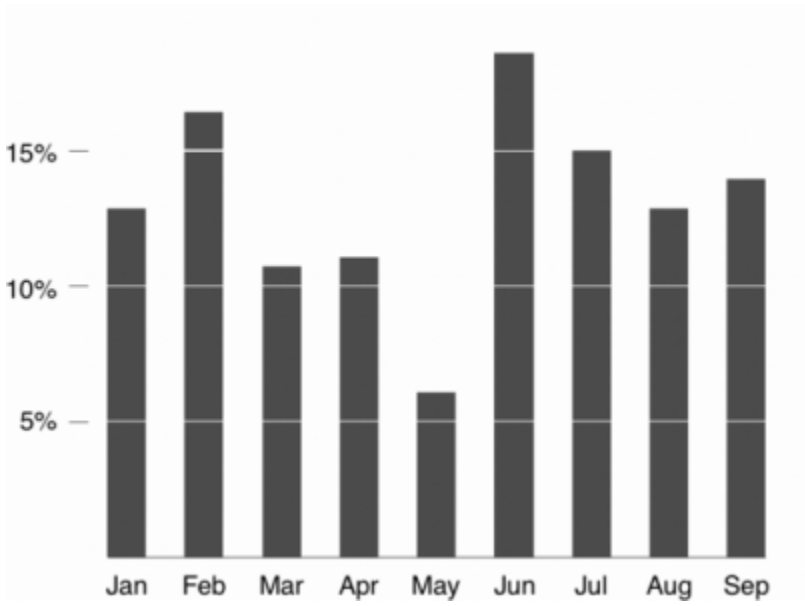
Avoid Chartjunk



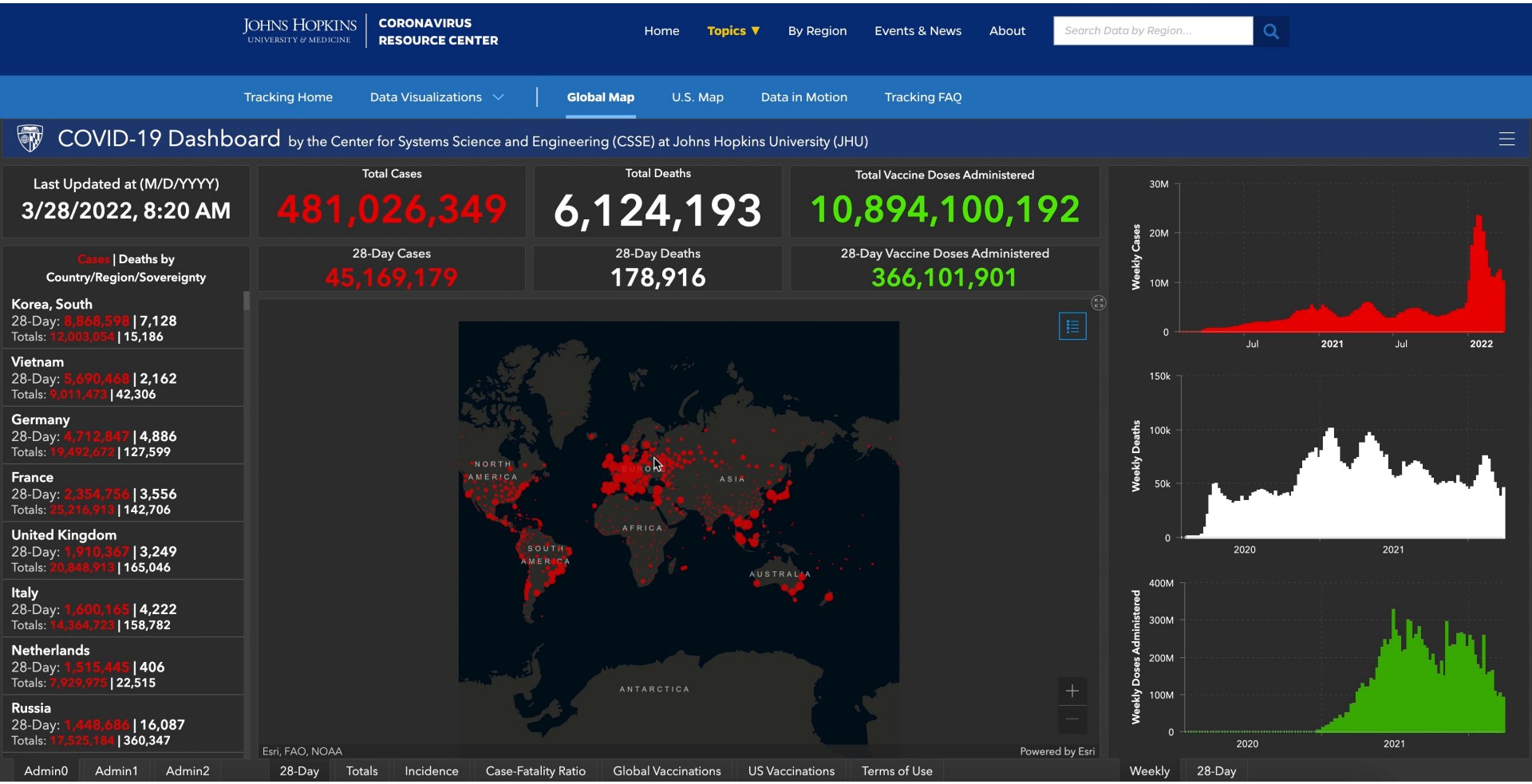
Maximize Data-Ink Ratio

$$\text{Data-Ink Ratio} = \frac{\text{Data-Ink}}{\text{Total Ink used}}$$

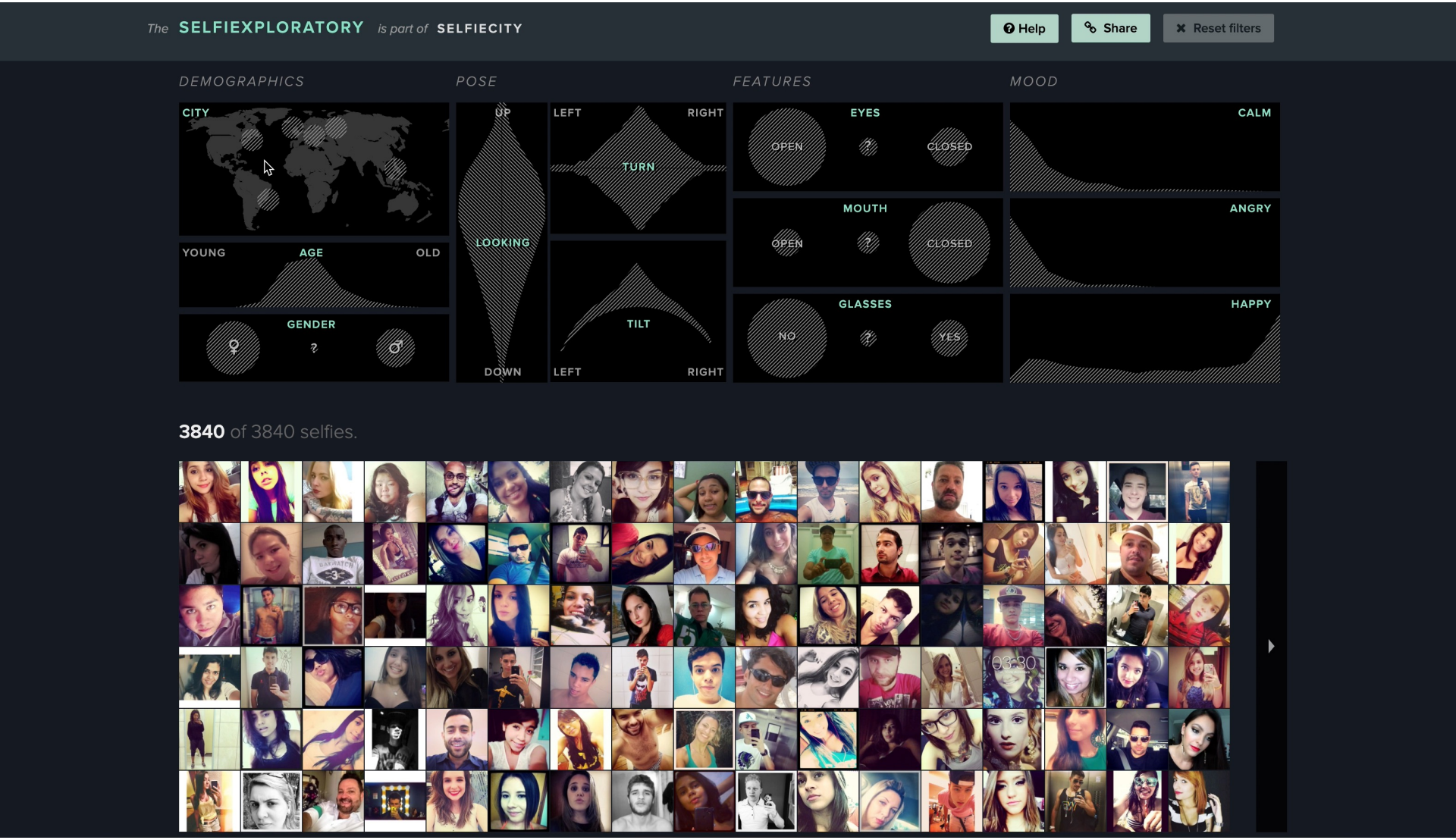

15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep


15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

Interaction



https://coronavirus.jhu.edu/map.html



http://selfiecity.net/selfiexploratory/

Animation

Why Use Animation?

- Visual variable to encode data
- Direct attention
- Understand system dynamics
- Understand state transition (maintain context)
- Increase engagement

Expectation

The expectation of a random variable is a number that attempts to capture the center of that random variable's distribution. It can be interpreted as the long-run average of many independent samples from the given distribution. More precisely, it is defined as the probability-weighted sum of all possible values in the random variable's support,

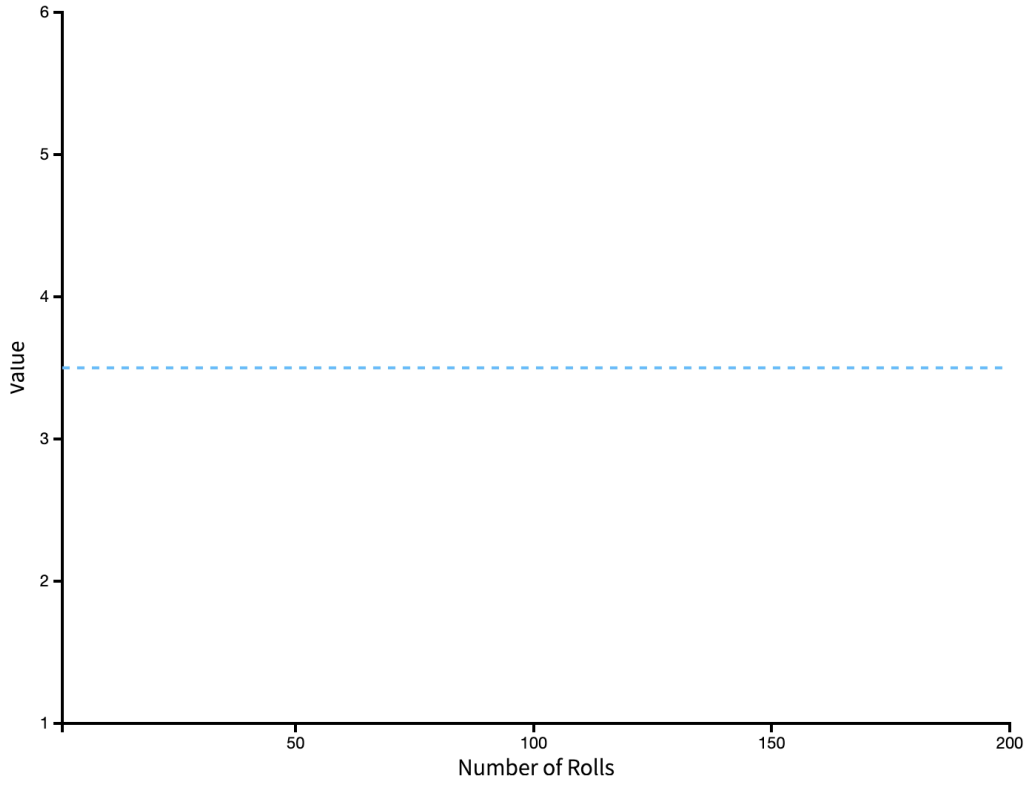
$$\mathrm{E}[X] = \sum_{x \in \mathcal{X}} xP(x)$$

Consider the probabilistic experiment of rolling a fair die and watch as the running sample mean converges to the expectation of 3.5.

Roll the Die

Roll 100 times

Change the distribution of the different faces of the die (thus making the die biased or "unfair") by adjusting the blue bars below and observe how this changes the expectation.

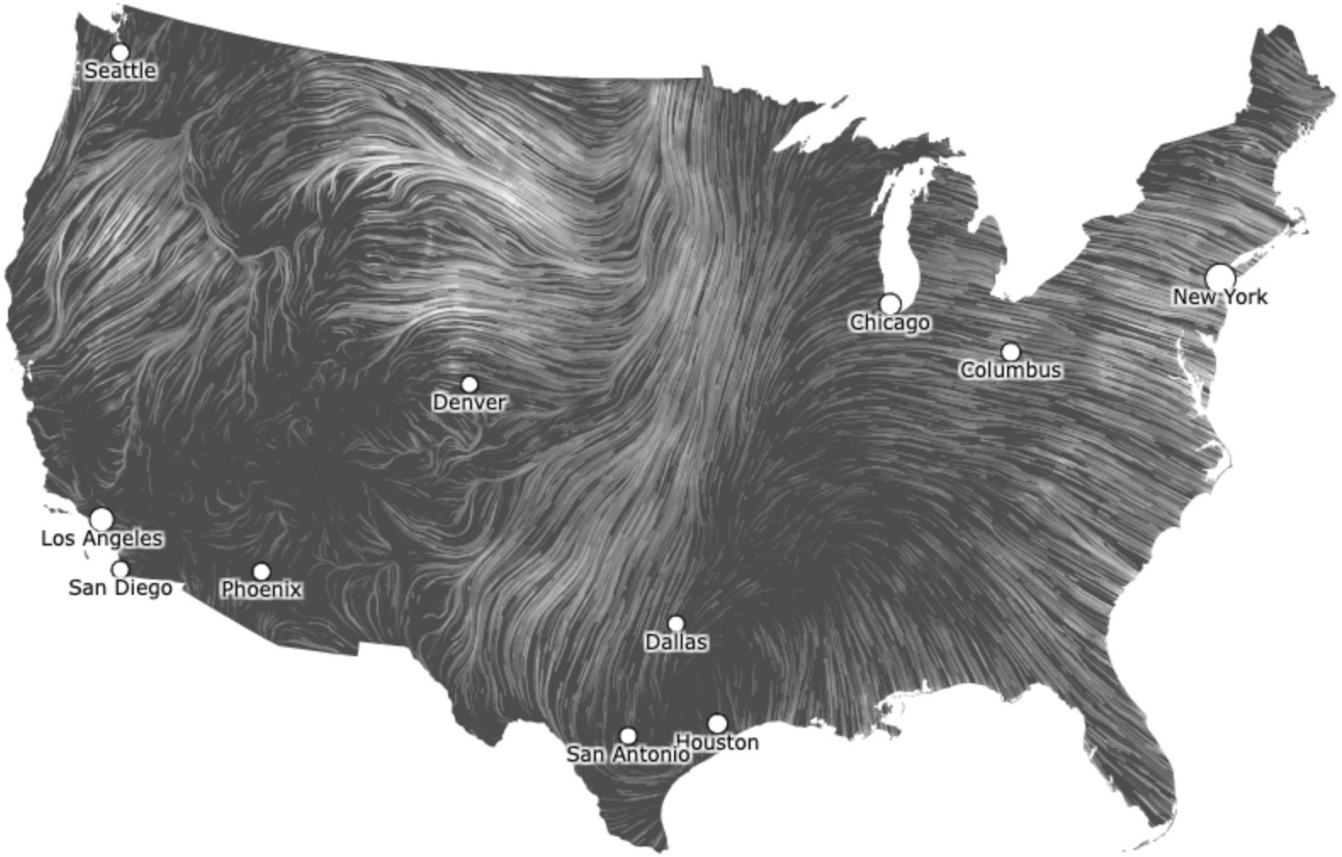


[students.brown.edu/seeing-theory/, Daniel Kunin]

wind map

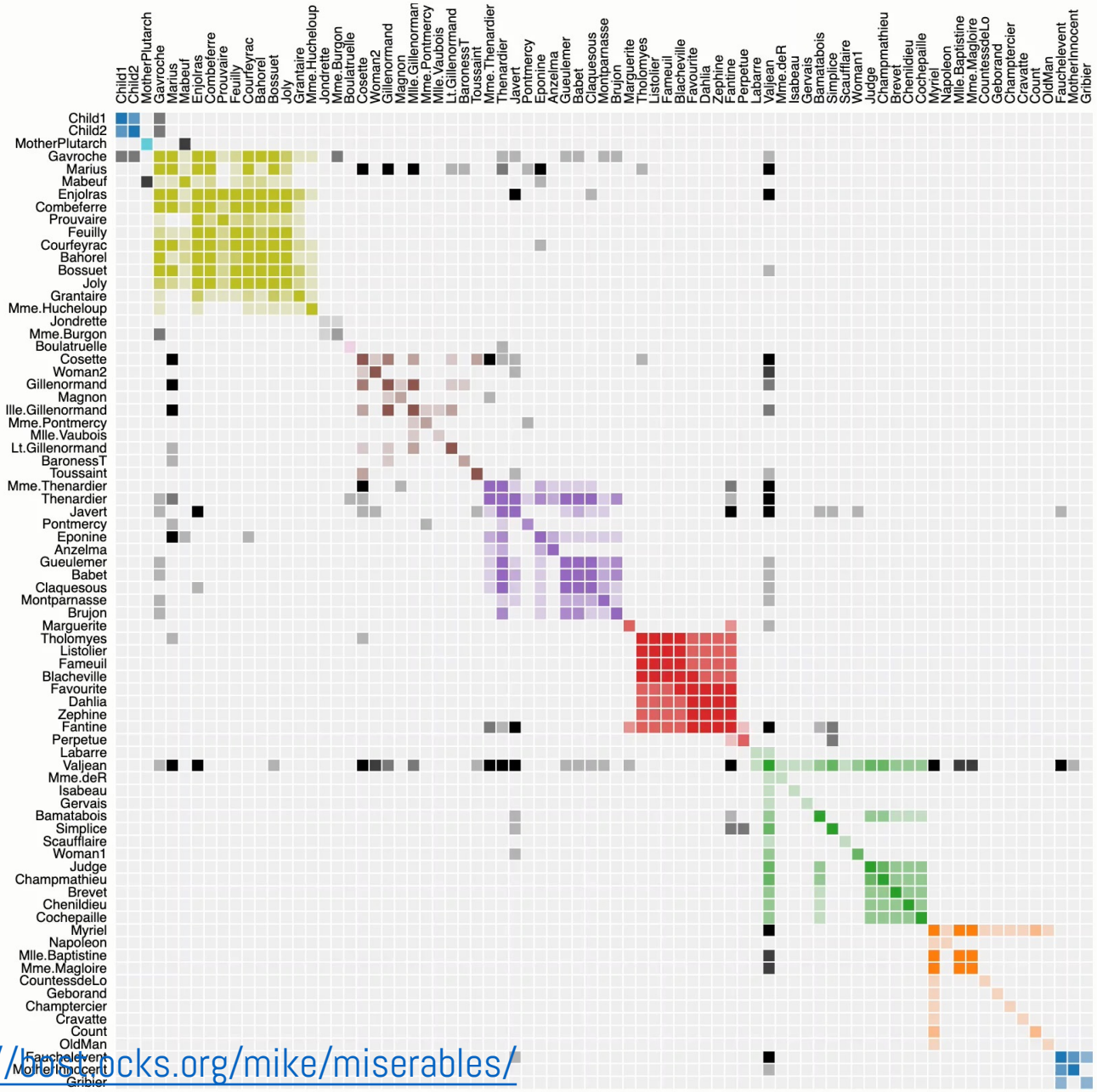
March 12, 2022
5:12 pm EST
(time of forecast download)

top speed: **50.5 mph**
average: **14.4 mph**



http://hint.fm/wind/

Les Misérables Co-occurrence



Order: by Cluster

This matrix diagram visualizes character co-occurrences in Victor Hugo's *Les Misérables*.

Each colored cell represents two characters that appeared in the same chapter; darker cells indicate characters that co-occurred more frequently.

Use the drop-down menu to reorder the matrix and explore the data.

Built with d3.js.

http://bost.ocks.org/mike/miserables/

Courses

- Arvind Satyanarayan
  - http://vis.csail.mit.edu/classes/6.894/
- Maneesh Agrawala
  - http://vis.berkeley.edu/courses/cs294-10-fa14/wiki/index.php/Main_Page
- Jeff Heer
  - http://courses.cs.washington.edu/courses/cse512/14wi/
- John Stasko
  - http://www.cc.gatech.edu/~stasko/7450/

Tools

- Vega ecosystem
  - https://vega.github.io/vega/
  - https://vega.github.io/vega-lite/
  - https://altair-viz.github.io/
- Tableau
  - https://public.tableau.com/en-us/s/
- Einblick
  - https://www.einblick.ai/
  - Fell free to ping me (ez@einblick.ai) with questions, feedback, etc.

Blogs & Websites

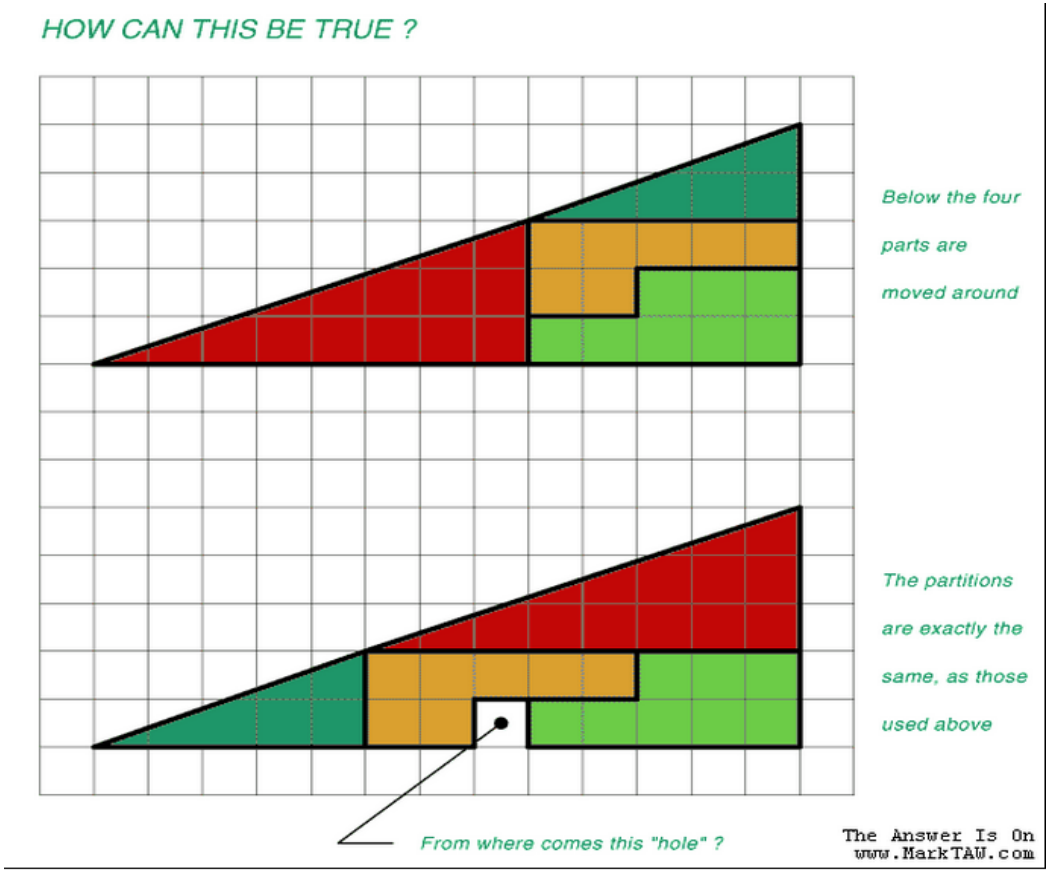
- https://www.reddit.com/r/dataisbeautiful/
- http://fivethirtyeight.com/
- http://flowingdata.com/
- http://www.informationisbeautiful.net/
- http://infosthetics.com/
- http://junkcharts.typepad.com/
- http://datavisualization.ch/
- http://eagereyes.org/
- http://blog.okcupid.com/
- https://twitter.com/nytgraphics

Articles & Others

- Tufte: The Visual Display of Quantitative Information
  - http://www.edwardtufte.com/tufte/
- http://www.csc.ncsu.edu/faculty/healey/PP/index.html
- http://www.sealthreinhold.com/tuftes-rules/rule_three.php
- Ted Talk: Beauty of Data Visualization
  - https://www.youtube.com/watch?v=pLqjQ55tz-U
- http://piksels.com/wp-content/uploads/2009/01/visualizingdata.pdf
- http://homes.cs.washington.edu/~jheer/files/zoo/
- http://www.targetprocess.com/articles/visual-encoding.html
- http://en.wikipedia.org/wiki/Misleading_graph

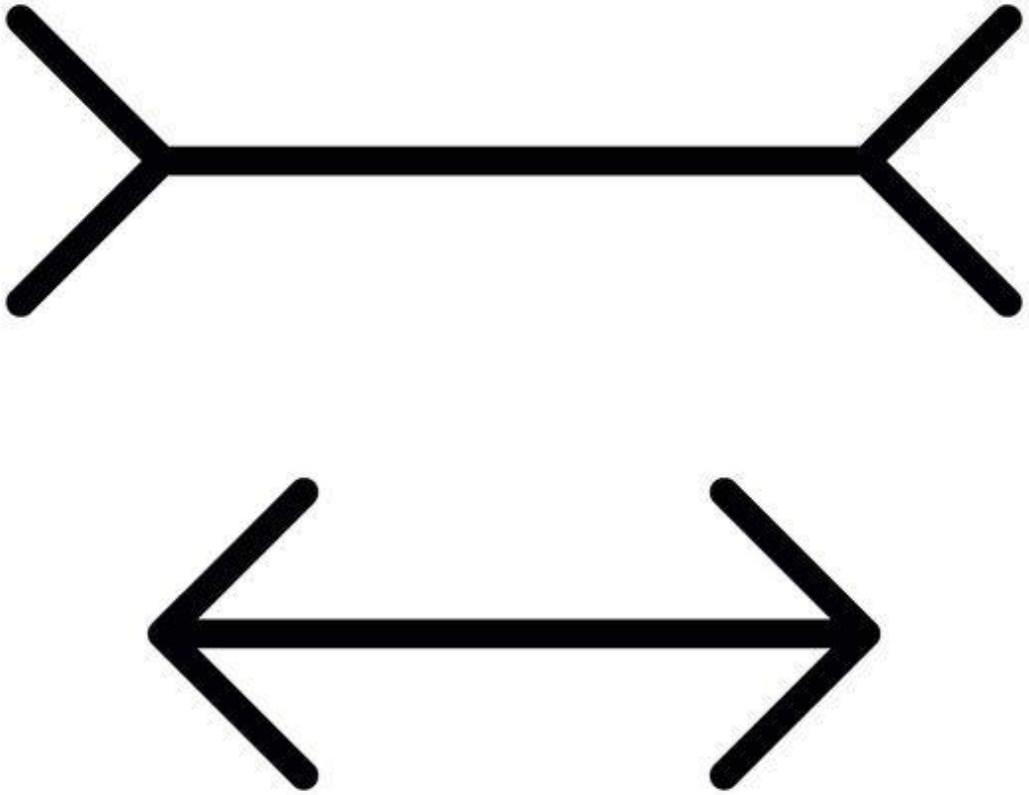
TIM KRASKA

LYING WITH STATISTICS AND VISUALIZATIONS



http://www.marktaw.com/blog/TheTriangleProblem.html

SURVEYS



LYING WITH NUMBERS

"The average market salary for MIT graduates with 0-2 years of experience is $151,000 per year"

In how many ways can this be misleading?

https://www.paysa.com/blog/harvard-vs-mit-which-tech-grads-make-the-most/

SAMPLING BIAS

- **More successful graduates are more likely to respond to surveys**
  - They feel good about their earnings
  - Surveys are only sent to big companies
- **How big is their sample size?**
  - Not disclosed

- **Tendency to exaggerate**
  - Brag about your salary
  - School spirit, want your alma mater to rank highly
- **Tendency to minimize**
  - No one likes tax
- **Do they cancel out each other?**
  - No one knows!

LYING WITH NUMBERS

"The average market salary for MIT graduates with 0-2 years of experience is $151,000 per year"

In how many ways can this be misleading?

https://www.paysa.com/blog/harvard-vs-mit-which-tech-grads-make-the-most/

THE TERM "AVERAGE"

- **Imagine a school with 5 alumni**
  - Bill Doors: $1 million/year
  - Mark Bergkerzuck: $120k/year
  - Larry Sheet: $100k/year
  - Sergey Bin: $80k/year
  - Steve Baller: $80k/year
- **Average can be mean, median, or mode; They can be totally different**
- **Mean: Evenly distributes the total among individuals**
  - Can be unrepresentative when measurements are highly skewed
  - In our example: $276k/year
- **Median: Value dividing distribution into two equal parts (50th percentile)**
  - In our example: $100k/year
- **Mode: Most frequently observed outcome (rarely reported with numeric data)**
  - In our example: $80k/year

Slide by Larry Winner, "Overview of How to Lie with Statistics"

FINAL SENTENCE

"The average market salary for MIT graduates with 0-2 years of experience is $151,000 per year"

PayScale's methodology did not include alums with advanced degrees and only used data from graduates with bachelor's degrees. It also excluded self-employed and contract employees.

Because the salaries of graduates from elite schools vary extensively, the study has a relatively wide margin of error, the report stated.

CORRELATION VS CAUSATION

What conclusions can you make from this data?

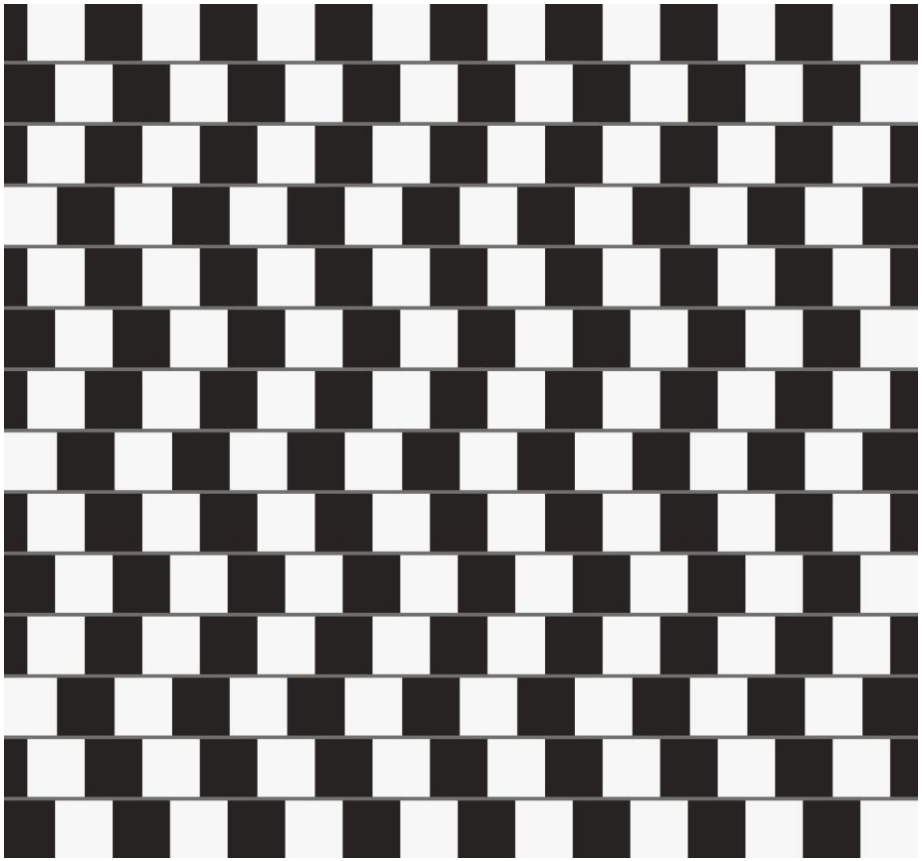
Does going to MIT make you rich?

LYING WITH SURVEYS

Three questions you should ask after you read any paper:

1. **Is there any bias in the sample set?**
    a. Look for unconscious bias
    b. Look for conscious bias
2. **What statistics are they actually talking about?**
3. **What conclusions can we make from their findings?**

LYING WITH VISUALIZATIONS

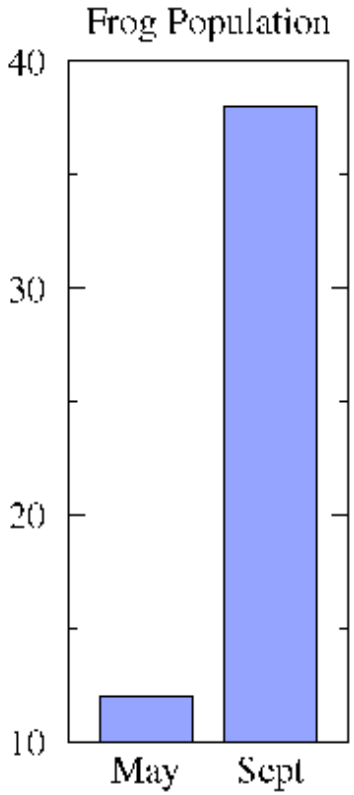


LYING WITH VISUALIZATION

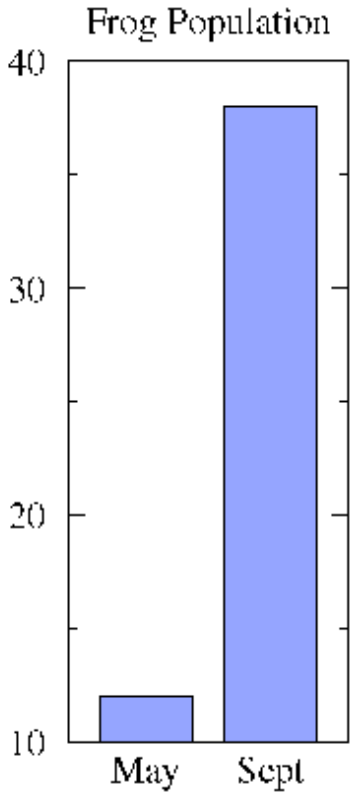
~~*seeing is believing*~~

"don't believe everything you see."

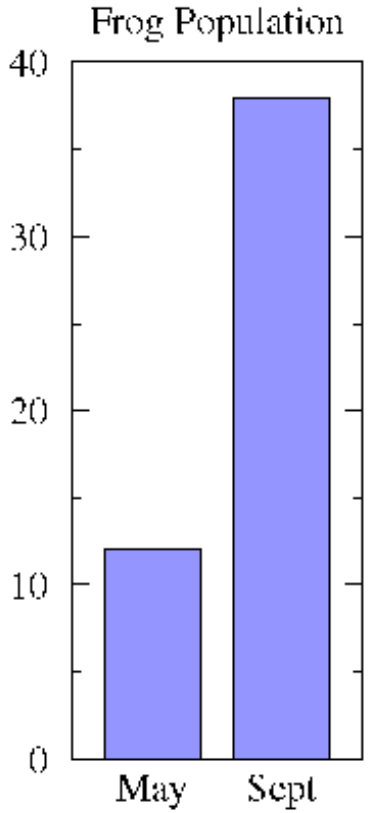
LYING WITH BAR CHARTS



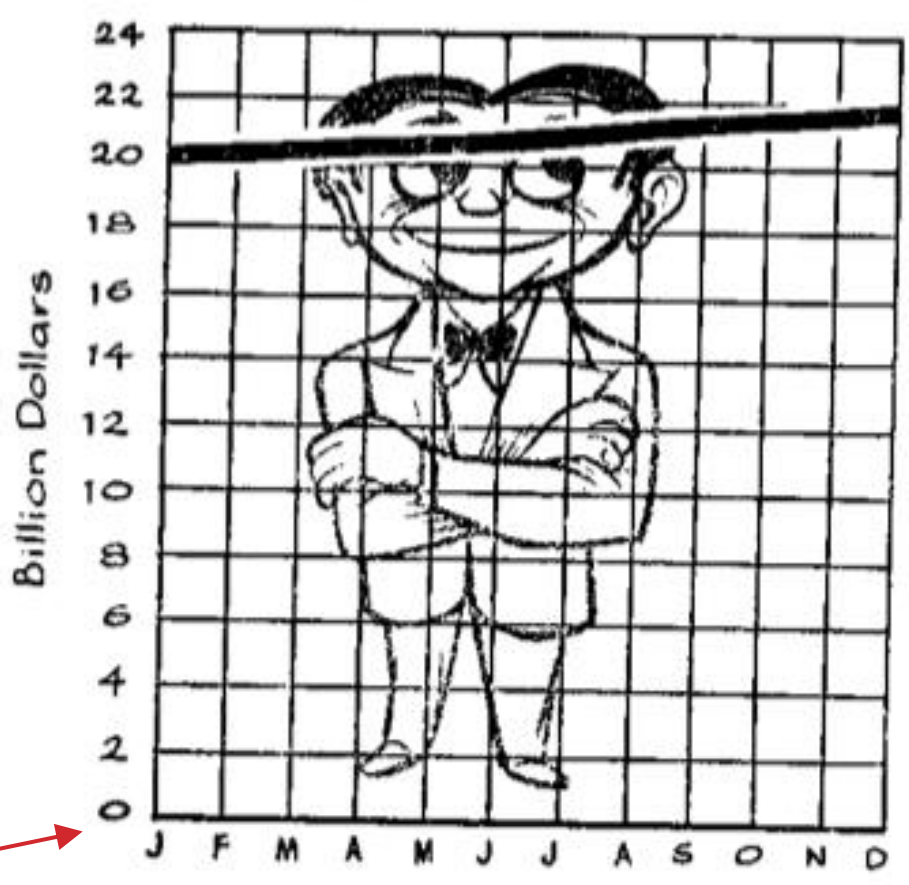
LYING WITH BAR CHARTS



VS

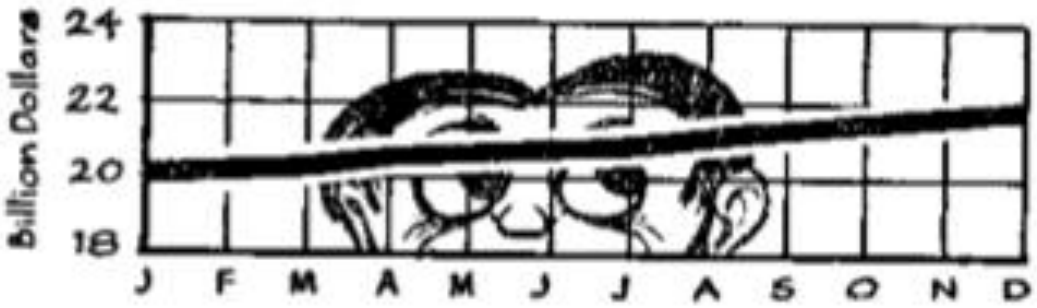


LYING WITH LINE CHART

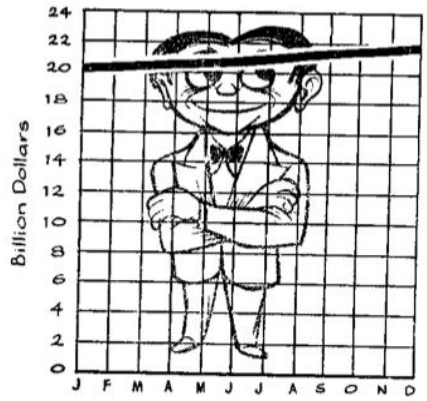


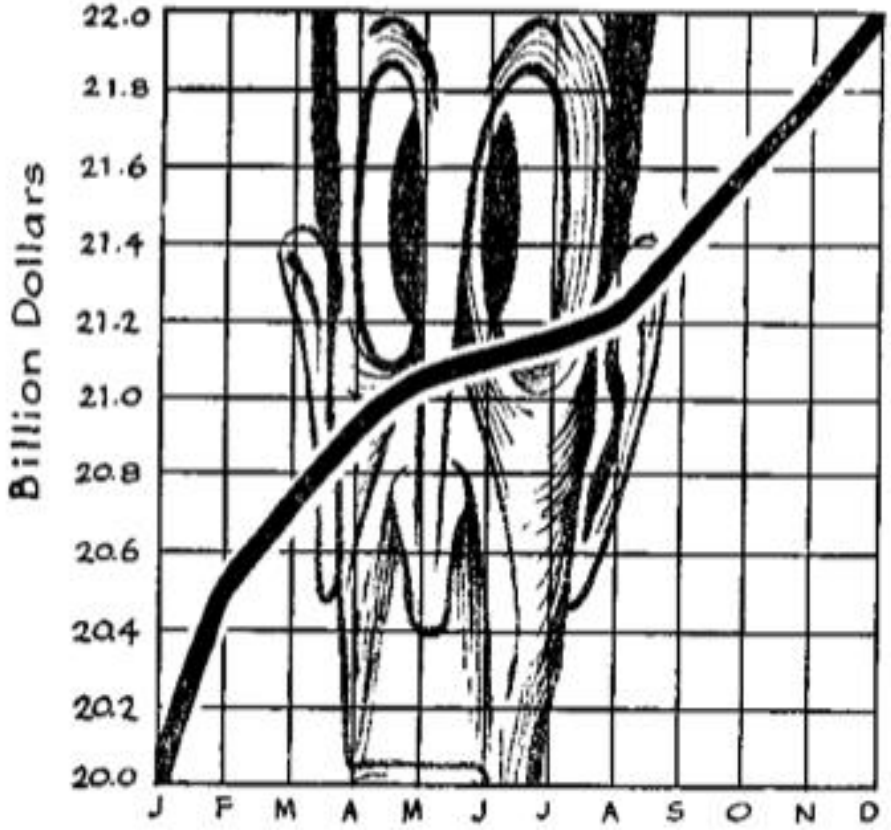
Zero line at the bottom

CHOP OFF THE BOTTOM



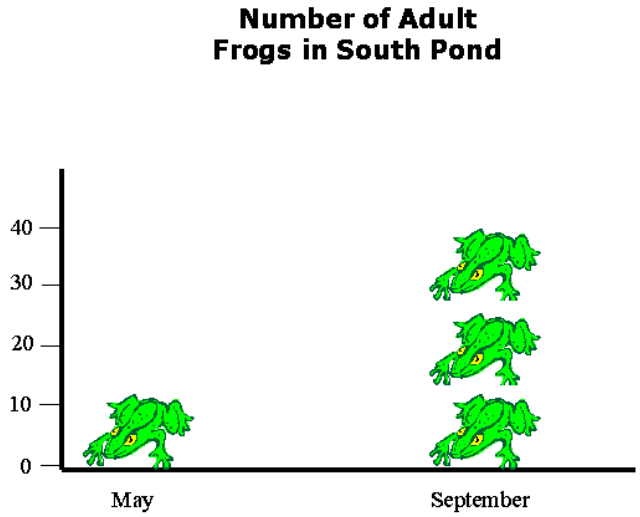
CHANGE THE PORTION OF Y-AXIS





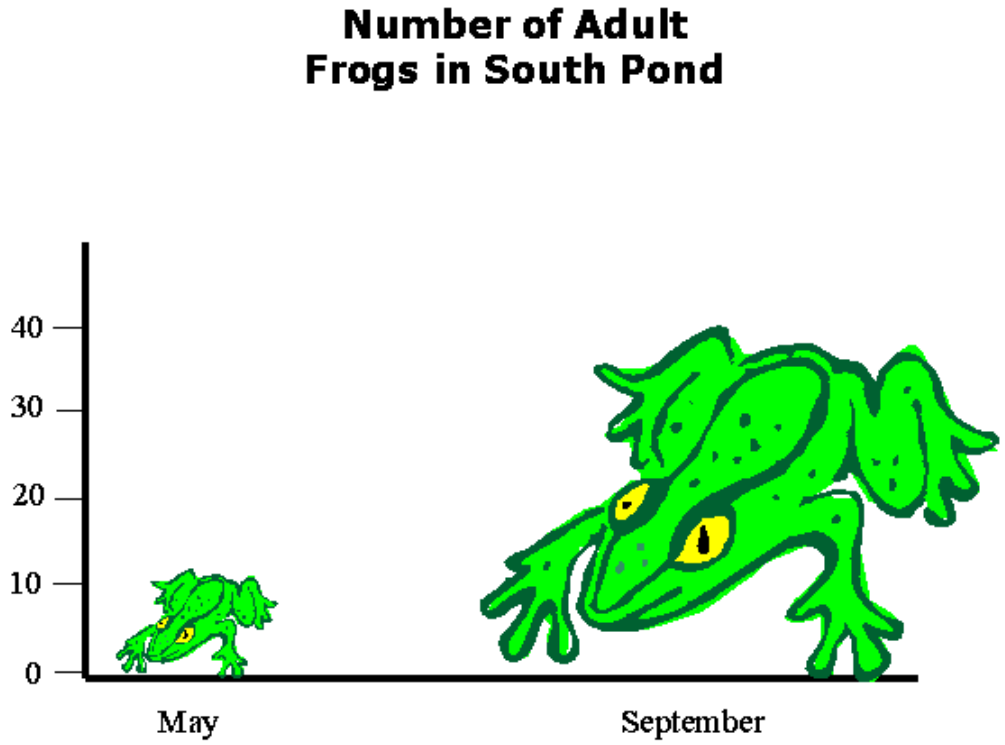
LYING WITH DIAGRAM

- **Say that in a pond, there were**
  - 13 Adult frogs in May
  - 39 Adult frogs in September
- **Represented in a "stacked-frog" plot**



LYING WITH DIAGRAM

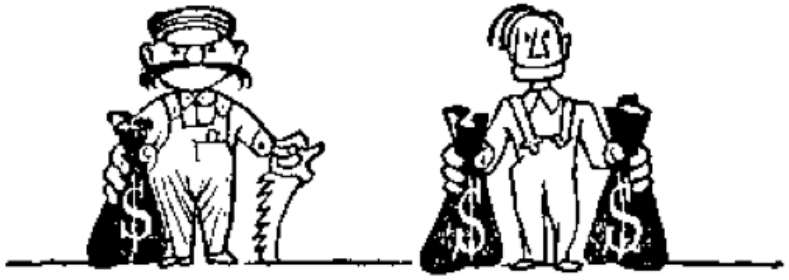
or we can represent in this way…



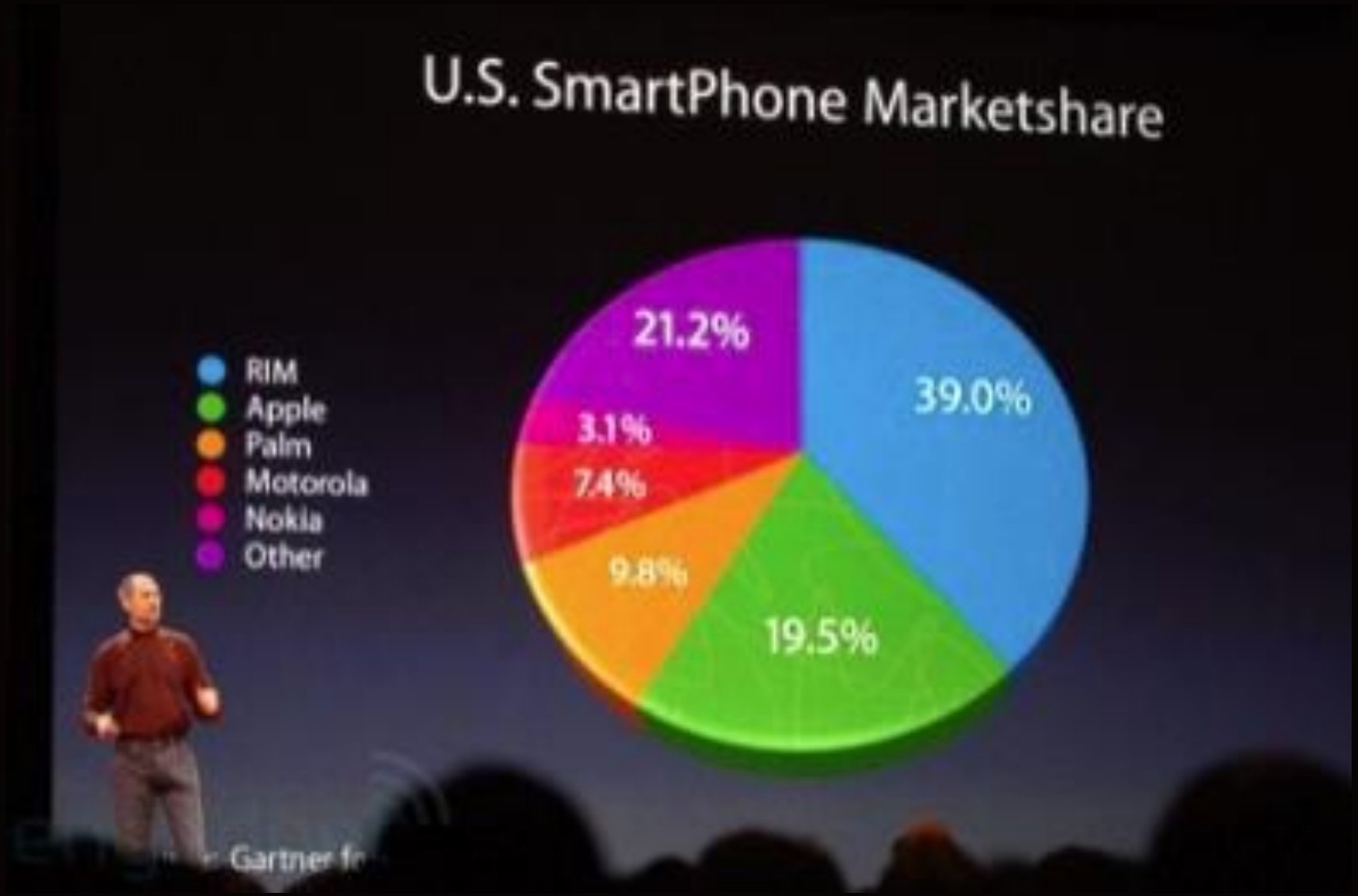
http://www.physics.csbsju.edu/stats/display.html

SOME MORE EXAMPLE

People at “A” get twice pay than people at “B”

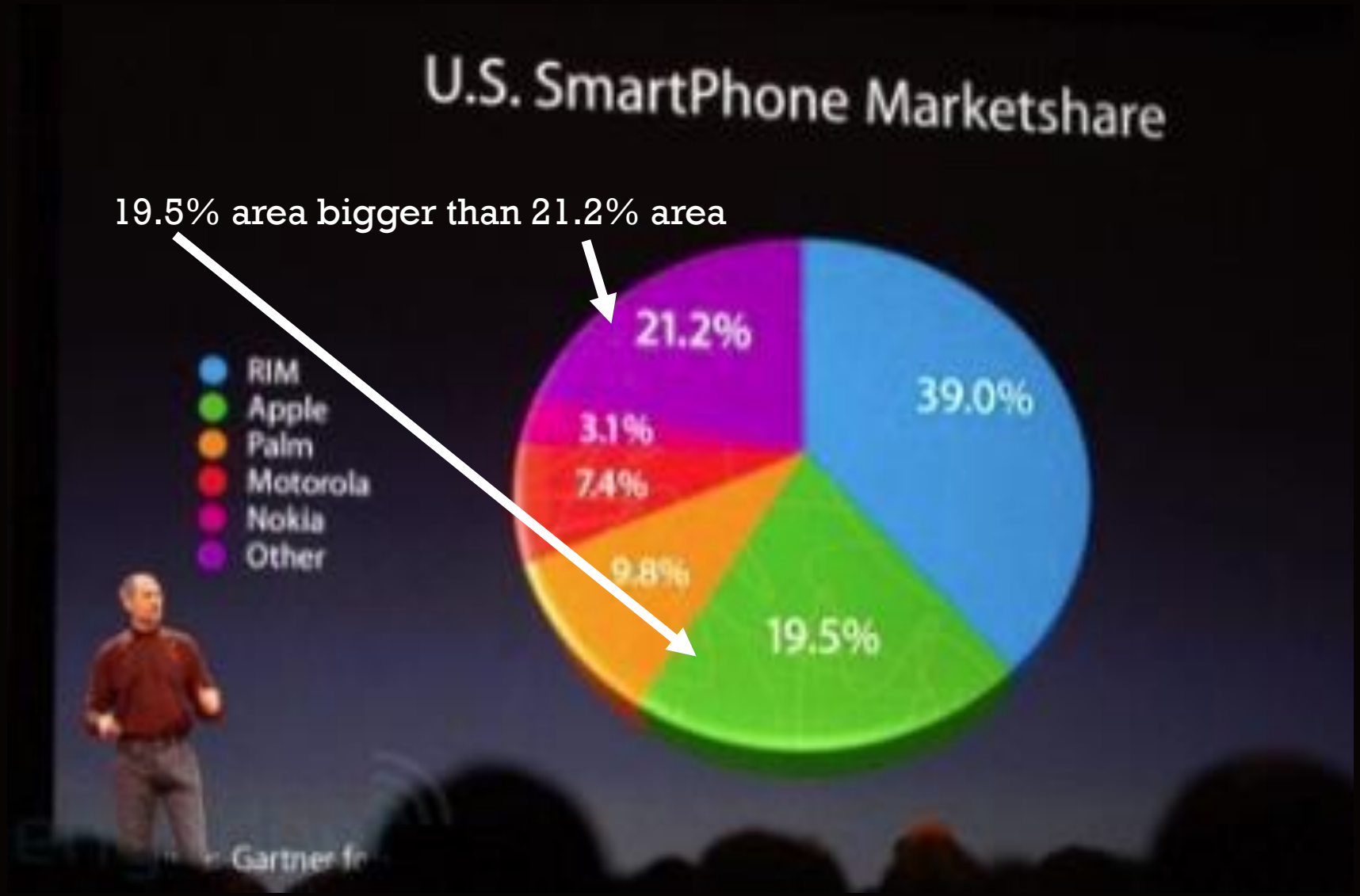


APPLE WWDC 2008



http://www.theguardian.com/technology/blog/2008/jan/21/liesdamnliesandstevejobs

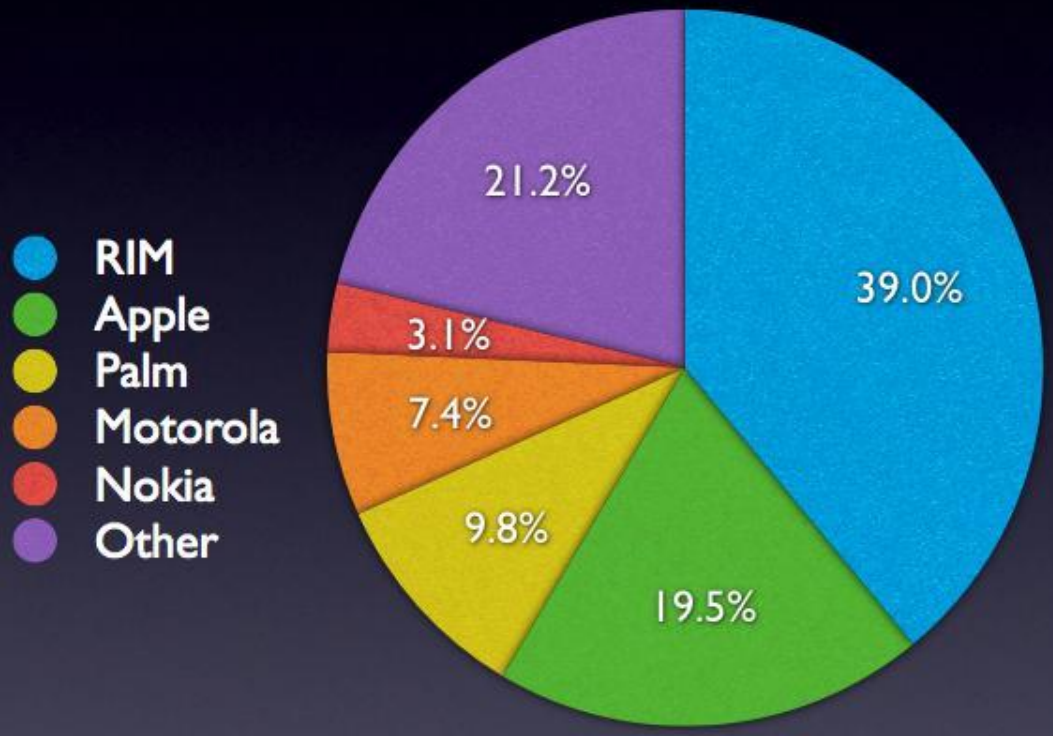
APPLE WWDC 2008



http://www.theguardian.com/technology/blog/2008/jan/21/liesdamnliesandstevejobs

TODAY'S EXAMPLE (APPLE WWDC 2008)

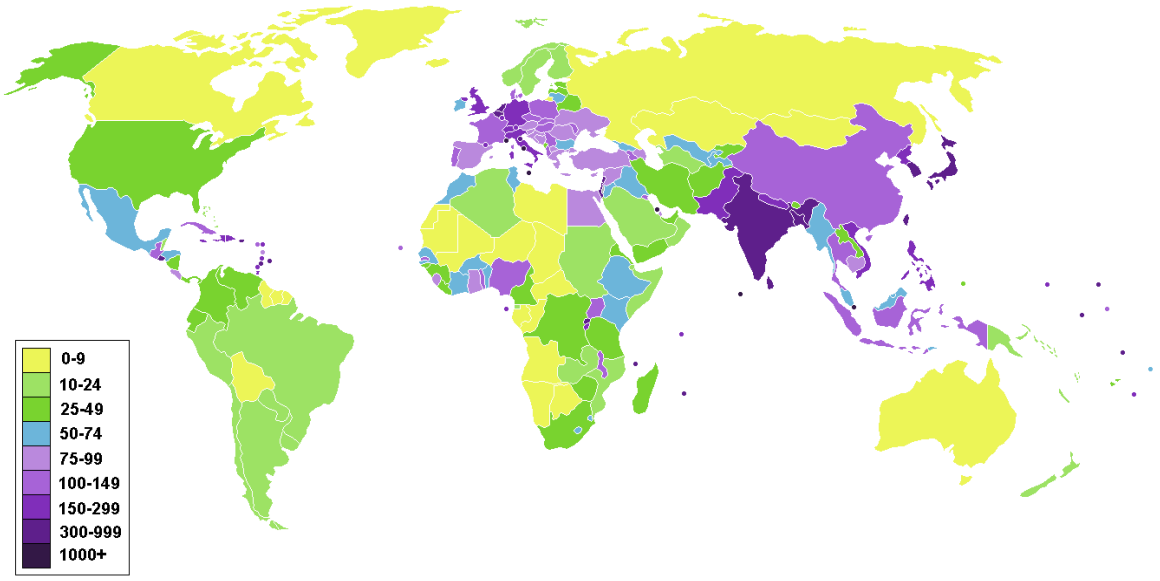
U.S. SmartPhone Marketshare



LYING WITH MAPS

Choropleth map

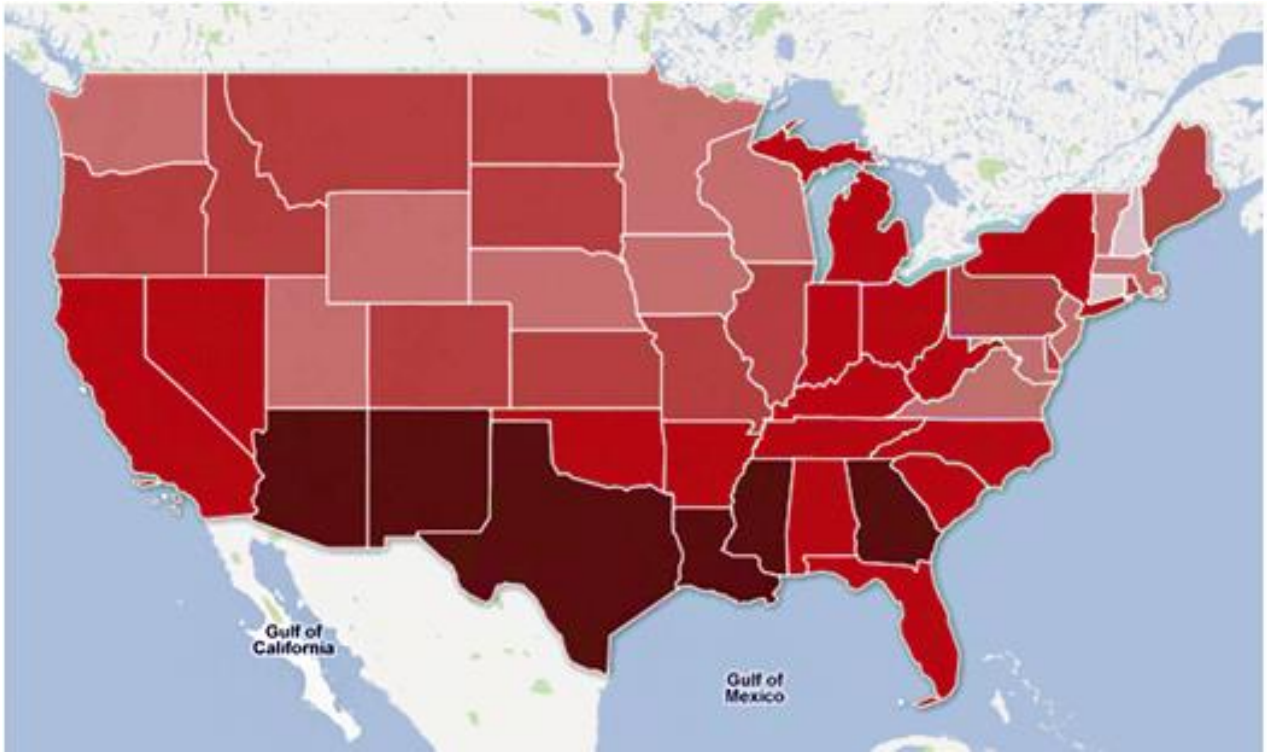
- **provides an easy way to visualize measurement varies across geographic area**



http://upload.wikimedia.org/wikipedia/commons/1/17/World_population_density_map.png

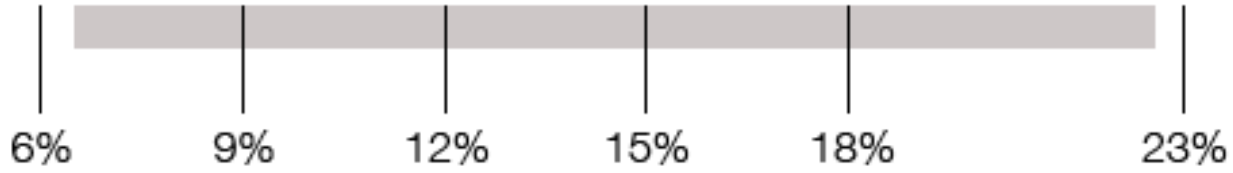
LYING WITH CHOROPLETH MAP

US poverty map from Guardian data blog

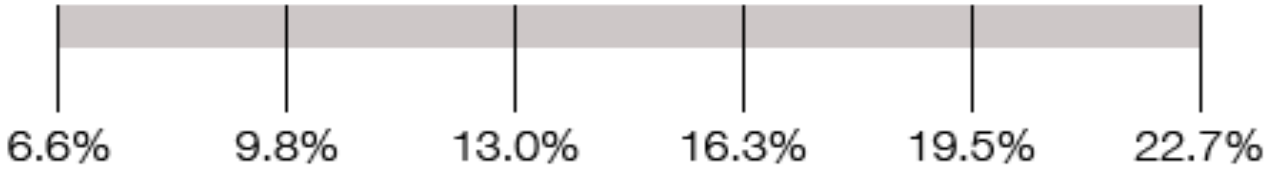


LYING WITH CHOROPLETH MAP

- **Poverty data range from 6.6% to 22.7%**
  - Unequally distributed

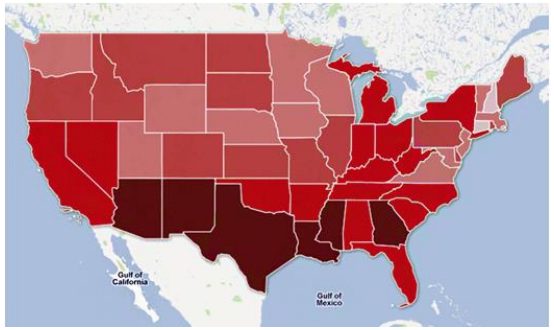


If we are measuring inequality, perhaps we should at least use equally distributed classes



http://vis4.net/blog/posts/choropleth-maps/

CHOICE OF COLOR

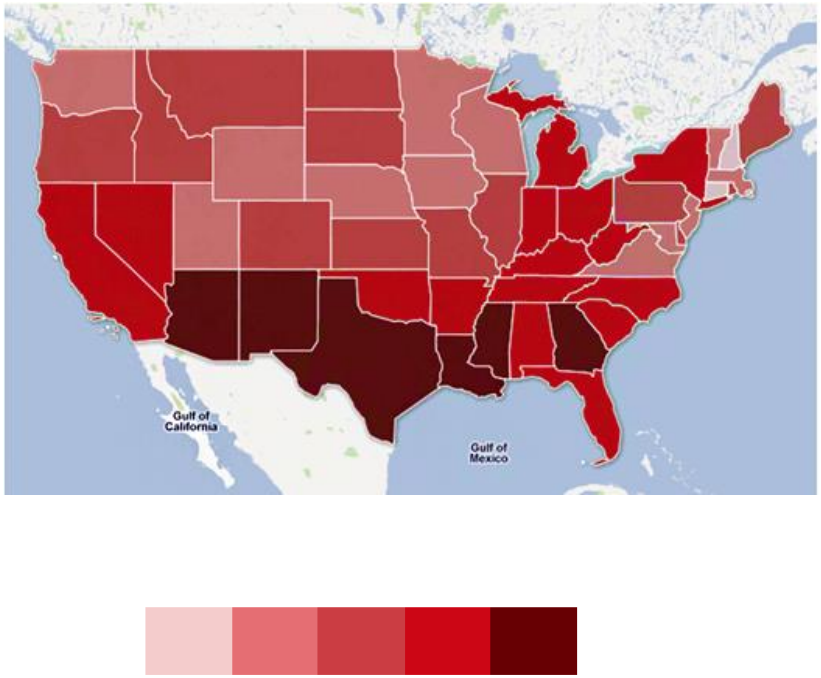


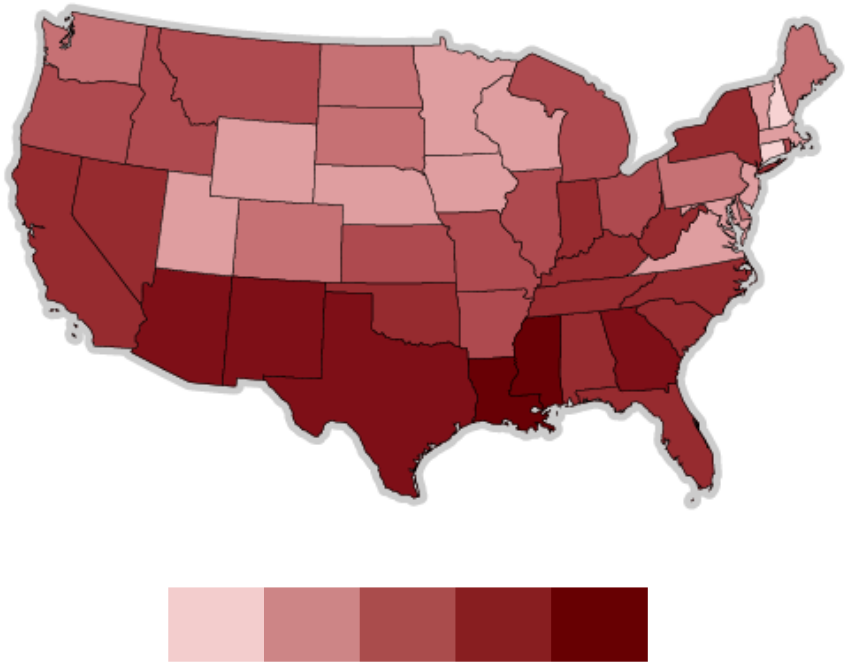
vs

http://vis4.net/blog/posts/choropleth-maps/

LYING WITH CHOROPLETH MAP

With equally distributed classes and equidistant colors from a HSV gradient



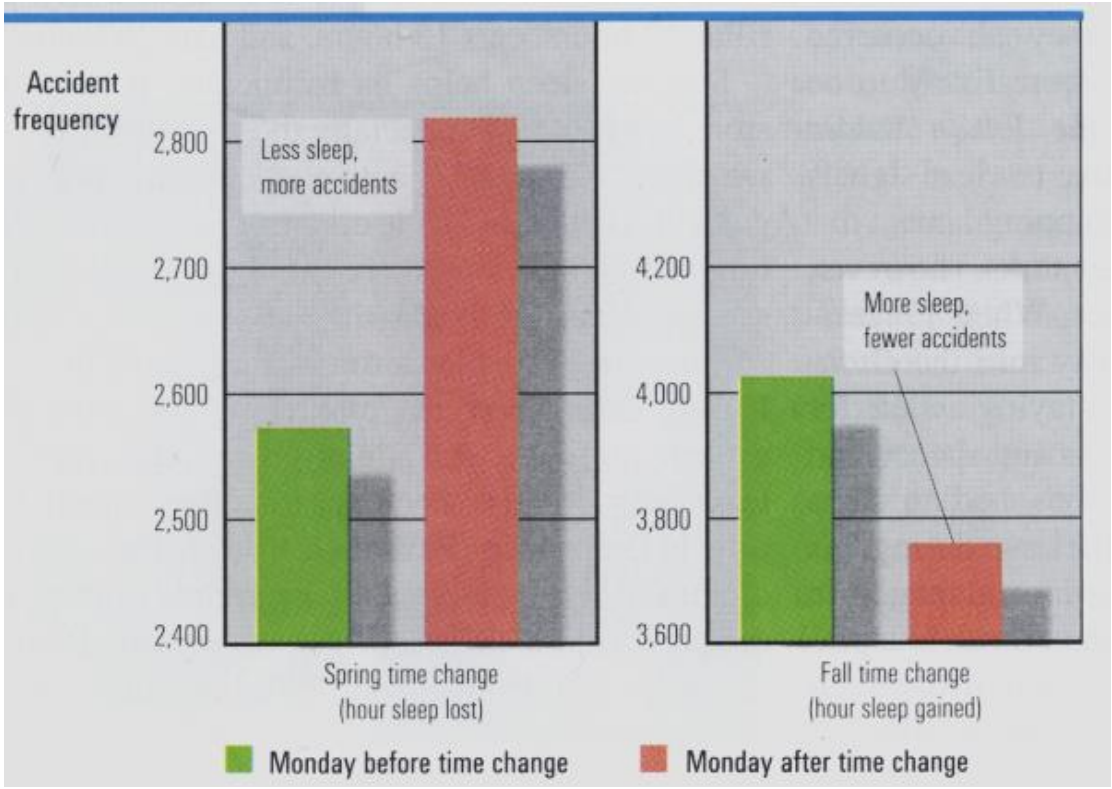


http://vis4.net/blog/posts/choropleth-maps/

LYING WITH CHOROPLETH MAP

- **When look at any choropleth map, be aware of**
  - How they categorize the classes
  - How they choose the colors
- **Choropleth map classification based on**
  - Equal-intervals
  - quantile classing; each class has equal number quantity
  - Iterative algorithm to find “natural breaks”

CLASS QUESTION



The above graphic was copied from a book about sleep research. The bars try to summarize the number of traffic accidents in Canada before and after daylight-savings time adjustments for the years 1991 and 1992 (combined). The goal of the graph is to suggest a correlation between lost sleep and traffic accidents.

Find 3 problems with this visualization

LYING THROUGH AGGREGATIONS

	Smoker?	
	Yes	No
Dead	107	132
Alive	174	175
Total	281	307
% Dying	38.1%	43.0%

(data adapted from Appleton et al. 1996, Am. Stat.)

MOTIVATING EXAMPLE: SMOKING & SURVIVAL

20-year follow-up study, Wickham in UK (Tunbridge et al. 1977)

1972-1974, one-in-six survey of the electoral roll, largely concerned with thyroid disease and heart disease

For simplicity, consider women aged 45 to 75 at the start of the study

- Smoking status: current smoker (Y/N)
- 20-year survival info: determined for all women in the study

	Smoker?	
	Yes	No
Dead	107	132
Alive	174	175
Total	281	307
% Dying	38.1%	43.0%

Protective effect of smoking?

(data adapted from Appleton et al. 1996, Am. Stat.)

SMOKING & SURVIVAL (CON'T)

	Age Group					
	45-54		55-64		65-74	
	Smoker?		Smoker?		Smoker?	
	Yes	No	Yes	No	Yes	No
Dead	27	12	51	40	29	101
Alive	103	66	64	81	7	28
Total	130	78	105	121	36	129
% Dying	20.8%	15.4%	48.6%	33.1%	80.6%	78.3%

Consider 10-year ranges: 45-54,55-64,65-75

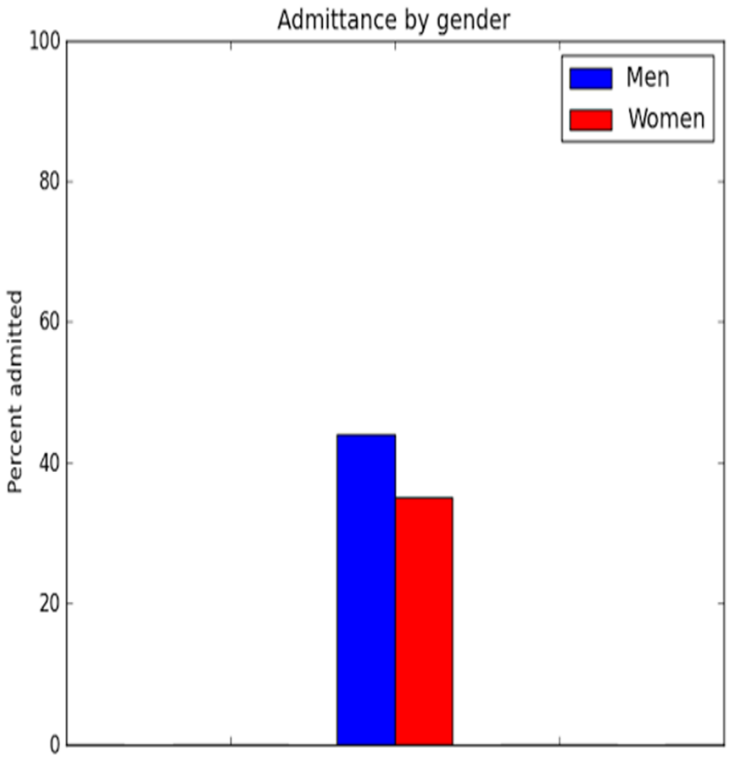
Non-smoking group does better in each case!

GENDER BIAS, OR NOT?

1973, UC Berkeley was afraid to be sued for discrimination against women in graduate school admissions

Percent acceptance: Male vs Female,

44% vs. 35%



GENDER BIAS, OR NOT? (CONT'D)

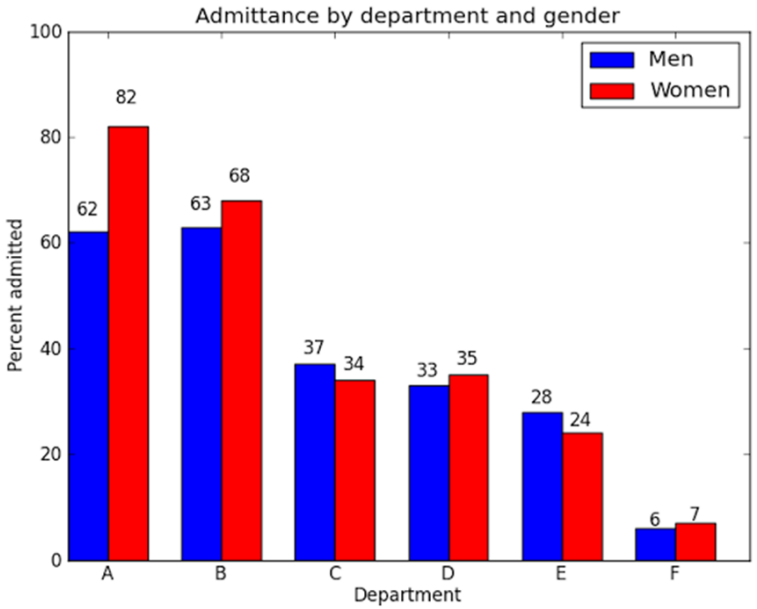
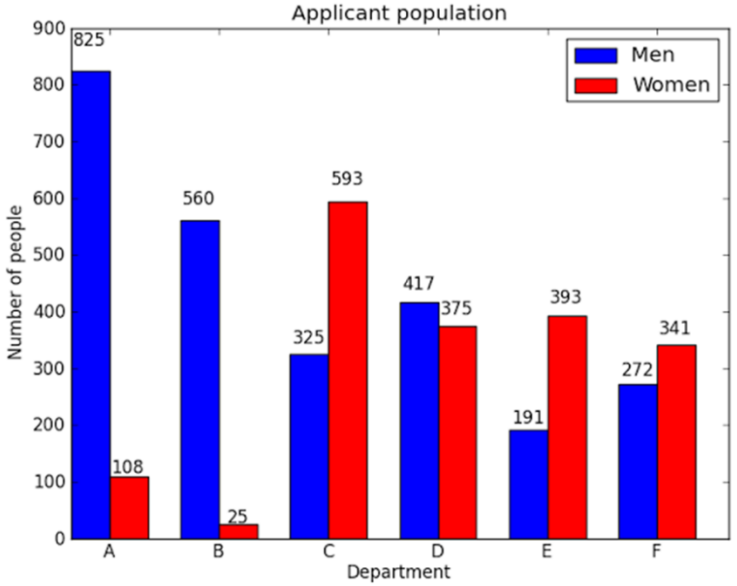


Table 2. Admissions data by sex of applicant for two hypothetical departments. For total, $\chi^2 = 5.71$, d.f. $= 1$, $P = 0.19$ (one-tailed).

Applicants	Outcome				Difference	
	Observed		Expected			
	Admit	Deny	Admit	Deny	Admit	Deny
	Department of machismatics					
Men	200	200	200	200	0	0
Women	100	100	100	100	0	0
	Department of social warfare					
Men	50	100	50	100	0	0
Women	150	300	150	300	0	0
	Totals					
Men	250	300	229.2	320.8	20.8	− 20.8
Women	250	400	270.8	379.2	− 20.8	20.8



P. J. Bickel, E. A. Hammel, J. W. O'Connell. (1975). Sex Bias in Graduate Admissions: Data from Berkeley. Science 187, (4175). pp. 398-404

SIMPSON PARADOX

If we have

$$\frac{a}{b} < \frac{A}{B} \text{ and } \frac{c}{d} < \frac{C}{D},$$

is it also true that

$$\frac{a+c}{b+d} < \frac{A+C}{B+D}?$$

Not necessarily! Note that

$$\frac{1}{3} < \frac{3}{8} \text{ and } \frac{5}{8} < \frac{2}{3}$$

but

$$\frac{1+5}{3+8} > \frac{3+2}{8+3}$$

Be aware of the dangers of ignoring a covariate that is correlated to an outcome variable and an explanatory one.

Simpson, E.H. (1951). "The interpretation of Interaction in Contingency Tables", *Journal of the Royal Statistical Society*, B, 13, 238-241.

MORE RECENT EXAMPLE

TechCrunch

Google found it paid men less than women for the same job

by Megan Rose Dickey https://techcrunch.com/2019/03/04/google-found-it-paid-men-less-than-women-for-the-same-job/

Wired

Are men at Google Paid less than women? Not Really [L SEP]

by Natasha Tiku https://www.wired.com/story/men-google-paid-less-than-women-not-really/

IMPRESSIVE FIGURES

MAKING NUMBERS LOOK BIGGER OR SMALLER

MAKING NUMBERS LOOK BIGGER

"runs up to 10x faster"

(https://www.digitalengineering247.com/article/altair-optistructruns-up-to-10x-faster-on-nvidia-gpus)

"lasts up to 5x longer"

(https://ca.crest.com/en-ca/products/crest-complete-whitening-plus-scope-outlast-toothpaste)

"cleans up to 10x better"

(https://www.youtube.com/watch?v=Yx9iCKKzYR4)

UP-TOS

"runs up to 10x faster"

(https://www.digitalengineering247.com/article/altair-optistructruns-up-to-10x-faster-on-nvidia-gpus)

"lasts up to 5x longer"

(https://ca.crest.com/en-ca/products/crest-complete-whitening-plus-scope-outlast-toothpaste)

"cleans up to 10x better"

(https://www.youtube.com/watch?v=Yx9iCKKzYR4)

"Schism consistently outperforms simple partitioning schemes, …, reducing the cost of distributed transactions **up to 30%**"

Carlo Curino, Yang Zhang, Evan P. C. Jones, Samuel Madden: *Schism: a Workload-Driven Approach to Database Replication and Partitioning*. PVLDB 3(1): 48-57 (2010)

What is the problem with up-tos?

Example (from Colton)

Sex and race distribution of 158 cases of *abdominal aortic aneurysms (AAA)* at metropolitan hospitals in a Southern city

Sex & Race	#AAA
White Males	93
AA Males	30
White Females	22
AA Females	13

Author's conclusion: Incidence of AAA is almost 3 times more frequent in Whites than African-Americans.

Clicker: Do you see a potential problem?

Example (from Colton)

Sex and race distribution of 158 cases of *abdominal aortic aneurysms (AAA)* at metropolitan hospitals in a Southern city

Sex & Race	#AAA
White Males	93
AA Males	30
White Females	22
AA Females	13

Author's conclusion: Incidence of AAA is almost 3 times more frequent in Whites than African-Americans.

Do you see a potential problem?

This fallacy is known as a lack of denominators

EXCEPTION FALLACY

HTTPS://CLICKER.MIT.EDU/6.S079/

4 out of 6 members of the math team representing Canada at the 2018 International Math Olympiad were from Ontario.

Clicker: Does Ontario have the best K-12 math curriculum in Canada?

a) Yes

b) No

c) Impossible to say

d) Scooby-doo

EXCEPTION FALLACY

4 out of 6 members of the math team representing Canada at the 2018 International Math Olympiad were from Ontario.

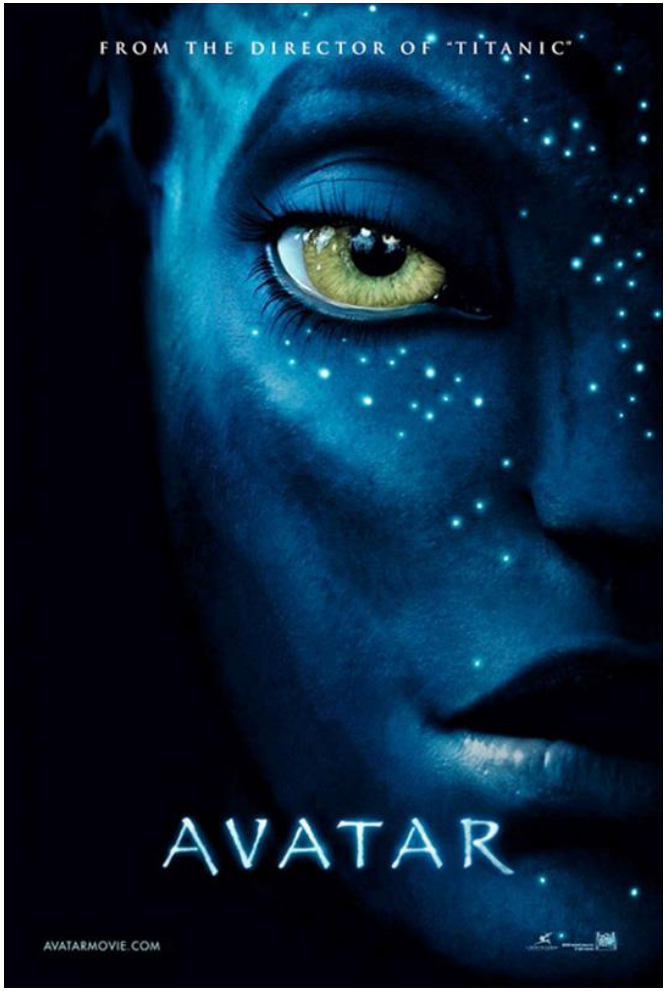
Does Ontario have the best K-12 math curriculum in Canada?

Note that Ontario has 40% of the nation's population.

DO YOU SEE A PROBLEM WITH THIS LIST?

Example: The Top 10 All Time Grossing Films (in Millions – US)

1) **Avatar (2009): $760**
2) **Titanic(1997): $658**
3) **Marvel's the Avengers (2012): $588**
4) **The Dark Knight (2008): $533**
5) **Star Wars I: The Phantom Menace (1999) $474**
6) **Star Wars IV: A New Hope (1977): $460**
7) **The Dark Knight Rises (2012) $449**
8) **Shrek 2 (2011): $441**
9) **E.T. The Extra-Terrestrial (1982): $435**
10) **The Hunger Games: Catching Fire (2013): $424**

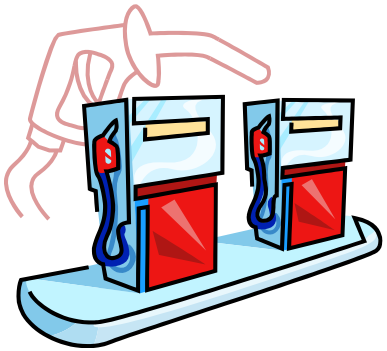


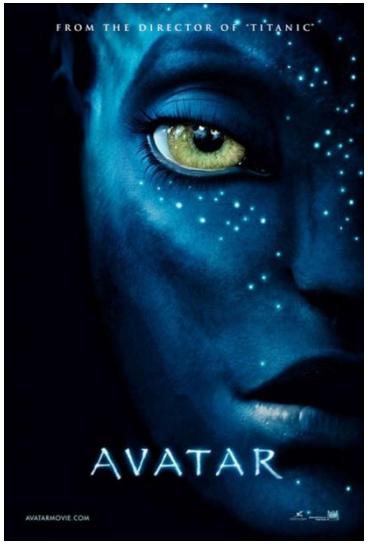
REAL VS. NOMINAL VARIABLES

Nominal Variables are in terms of a current dollars. For example, you're starting salary after college might be $50,000 per year.

Real variables are in terms of some fixed commodity. Real variables measure purchasing power. If a gallon of gas costs $2.00, then we can calculate your "real" income.

$$\textbf{Real Income} = \frac{\textbf{Nominal Income}}{\textbf{Price}} = \frac{\textbf{\$50,000}}{\textbf{\$2.00}} = \textbf{25,000}$$





In 2009, a gallon of gas cost $3.50

$$\text{Real Gross} = \frac{\text{Nominal Gross}}{\text{Price}} = \frac{\$749\text{M}}{\$3.50} = 214\text{M}$$

(Gallons of Gas)

In 1977, a gallon of gas cost $.62

$$\text{Real Income} = \frac{\text{Nominal Gross}}{\text{Price}} = \frac{\$460\text{M}}{\$.62} = 742\text{M}$$

(Gallons of Gas)

The Top 10 All Time Grossing Films– **Inflation Adjusted** (Millions of 2000 Dollars)

1) Gone With the Wind (1939): $1,689
2) Star Wars Episode IV(1977): $960
3) The Sound of Music(1965): $768
4) ET: The Extraterrestrial(1982): $764
5) The Ten Commandments (1956): $706
6) Titanic (1997): $691
7) Jaws (1975): $690
8) Dr. Zhivago (1965): $669
9) The Exorcist (1973): $596
10) Snow White (1937): $587

Notes: Avatar falls to #14 ($516), a movie ticket in 1939 was $0.23

SELF-DRIVING CARS

[Tesla] said Autopilot-enabled cars had covered 130 million miles without a fatality, compared to a national average of one fatality every 94 million miles. **Musk says it would be "morally reprehensible" to delay its rollout.**

Tesla's Cars Have Driven 140M Miles on Autopilot. Here's How *--- Wired, 17 Aug 2016*

Clicker: what is the problem with this statement?

SELF-DRIVING CARS

[Tesla] said Autopilot-enabled cars had covered 130 million miles without a fatality, compared to a national average of one fatality every 94 million miles. **Musk says it would be "morally reprehensible" to delay its rollout.**

Tesla's Cars Have Driven 140M Miles on Autopilot. Here's How *--- Wired, 17 Aug 2016*

A RAND Corporation report concluded that fatalities and injuries are so rare that it would require an automated car to drive as many as hundreds of billions of miles before its performance could be fairly compared with statistics from the much larger population of human drivers.

https://www.technologyreview.com/s/601849/teslas-dubious-claims-about-autopilots-safety-record/

SERVICE UP-TIME

A fictitious school bus status update website claims 99.9% uptime.

Is this good?

SERVICE UP-TIME

A fictitious school bus status update website claims 99.9% uptime.

Is this good?

What if in the morning hours (5am - 8am) of a big snowstorm day, the website is down due to too much traffic.

With three such snowstorms a year, the website is down 9 hours out of a total of 365 x 24 hours per year.

9/8760 is roughly 0.1%. But the website is down when you most need it to be up!

P-VALUE

A New Study shows: A Glass Of Red Wine Is The Equivalent To An Hour At The Gym [Fox News 02/15 and others]

http://www.huffingtonpost.co.uk/2016/01/08/a-glass-of-red-wine-is-the-equivalent-to-an-hour-at-the-gym-says-new-study_n_7317240.html

A new study shows: Secret to winning a Nobel prize? Eat More Chocolate [Time 10/12]

Scientists find the secret of longer life for men

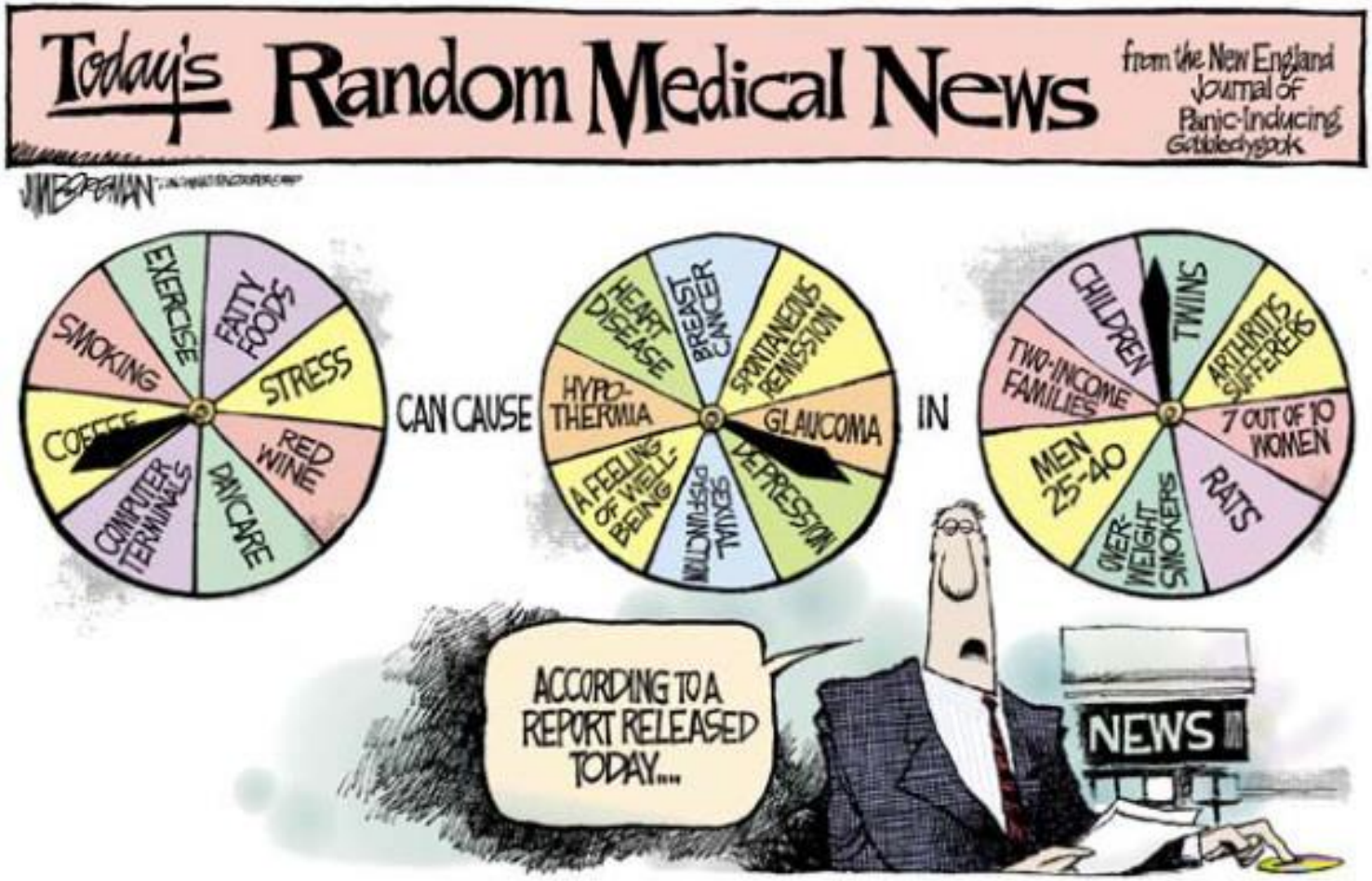
[Daily Mail UK, 09/12]

http://www.dailymail.co.uk/sciencetech/article-2207981/Scientists-secret-living-life-men-bad-news-Castration-key.html

Scientists find the secret of longer life for men (The bad news: castration is the key) [Daily Mail UK, 09/12]

http://www.dailymail.co.uk/sciencetech/article-2207981/Scientists-secret-living-life-men-bad-news-Castration-key.html

Data Dredging



STATISTICAL TEST

Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes

Beer (25):

27 20 21 26 27 31 24 21 20 19
23 24 28 19 24 29 18 20 17 31
20 25 28 21 27

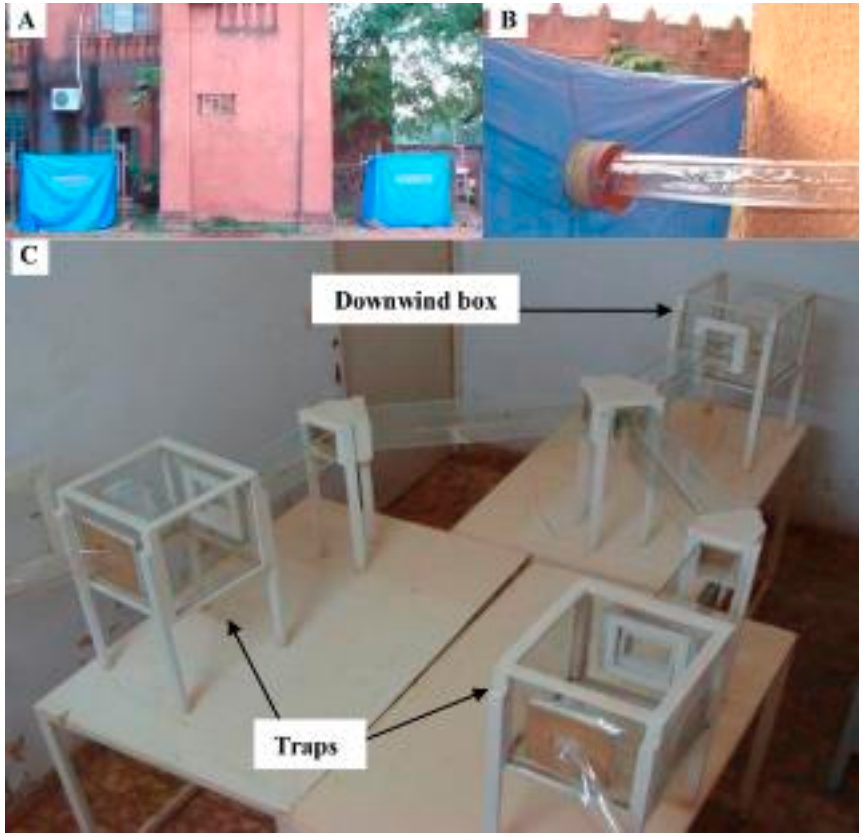
Mean: 23.6

Water (18):

21 22 15 12 21 16 19 15 22 24
19 23 13 22 20 24 18 20

Mean: 19.2

Is a difference of 4.4 significant?



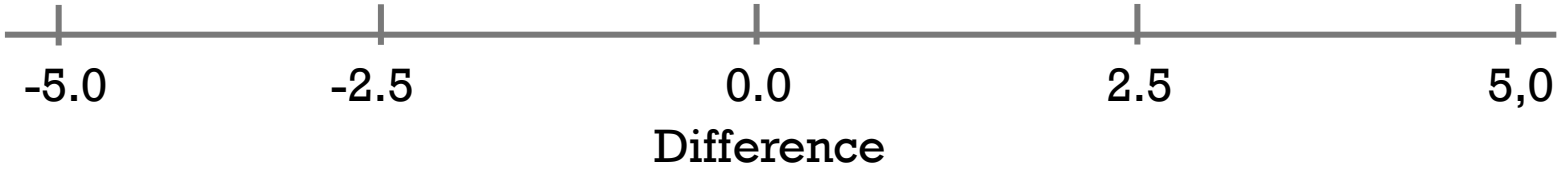
Beer Consumption #Human Attractiveness to Malaria Mosquitoes
Thierry Lefèvre , Louis-Clément Gouagna, Kounbobr Roch Dabiré, Eric Elguero, Didier Fontenille, François Renaud, Carlo Costantini, Frédéric Thomas

PERMUTATION TEST

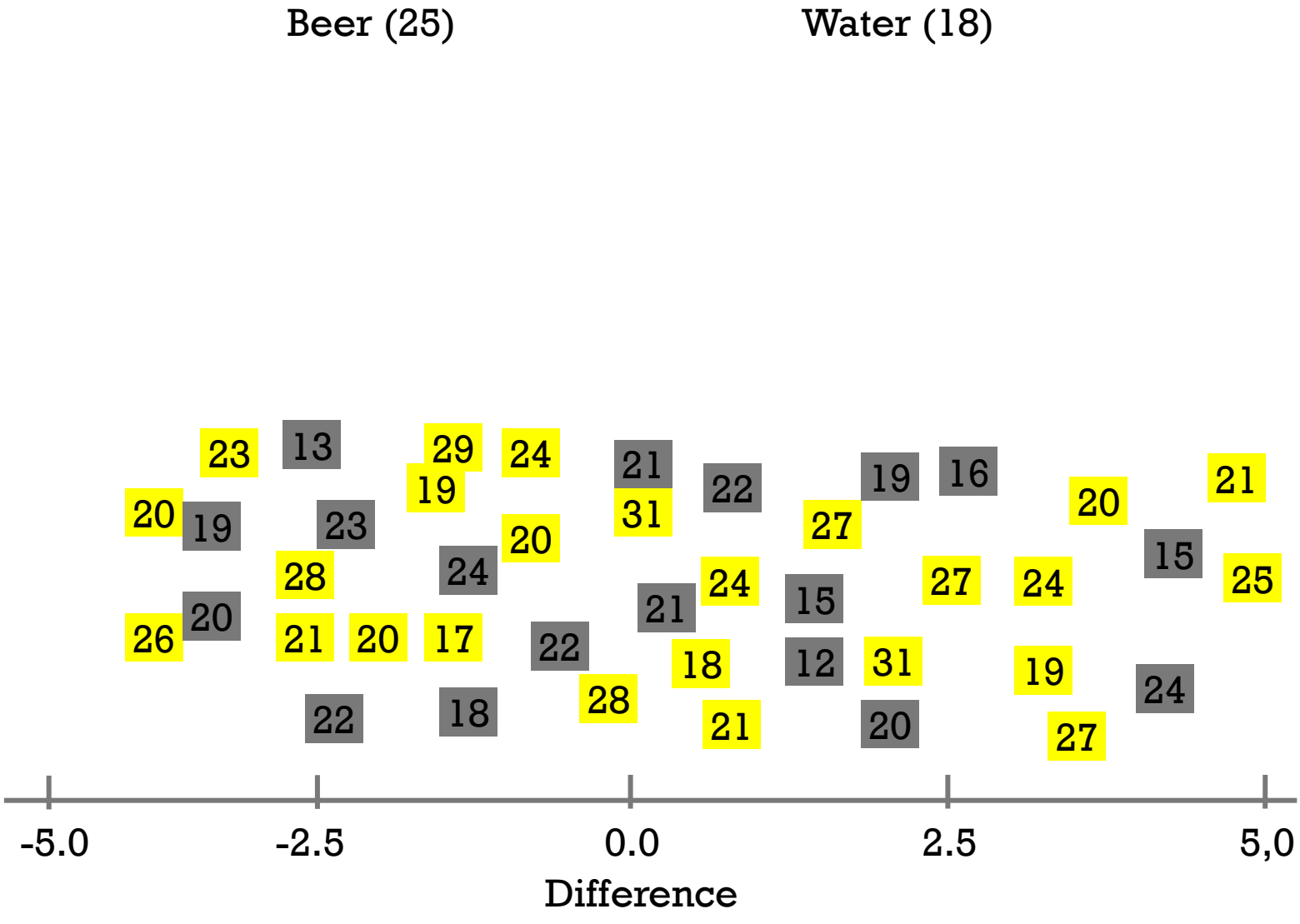
Beer (25)				
27	23	20	31	29
20	24	25	24	18
21	28	28	21	20
26	19	21	20	17
27	24	27	19	31

Water (18)			
21	19	16	24
22	23	19	18
15	13	15	20
12	22	22	
21	20	24	

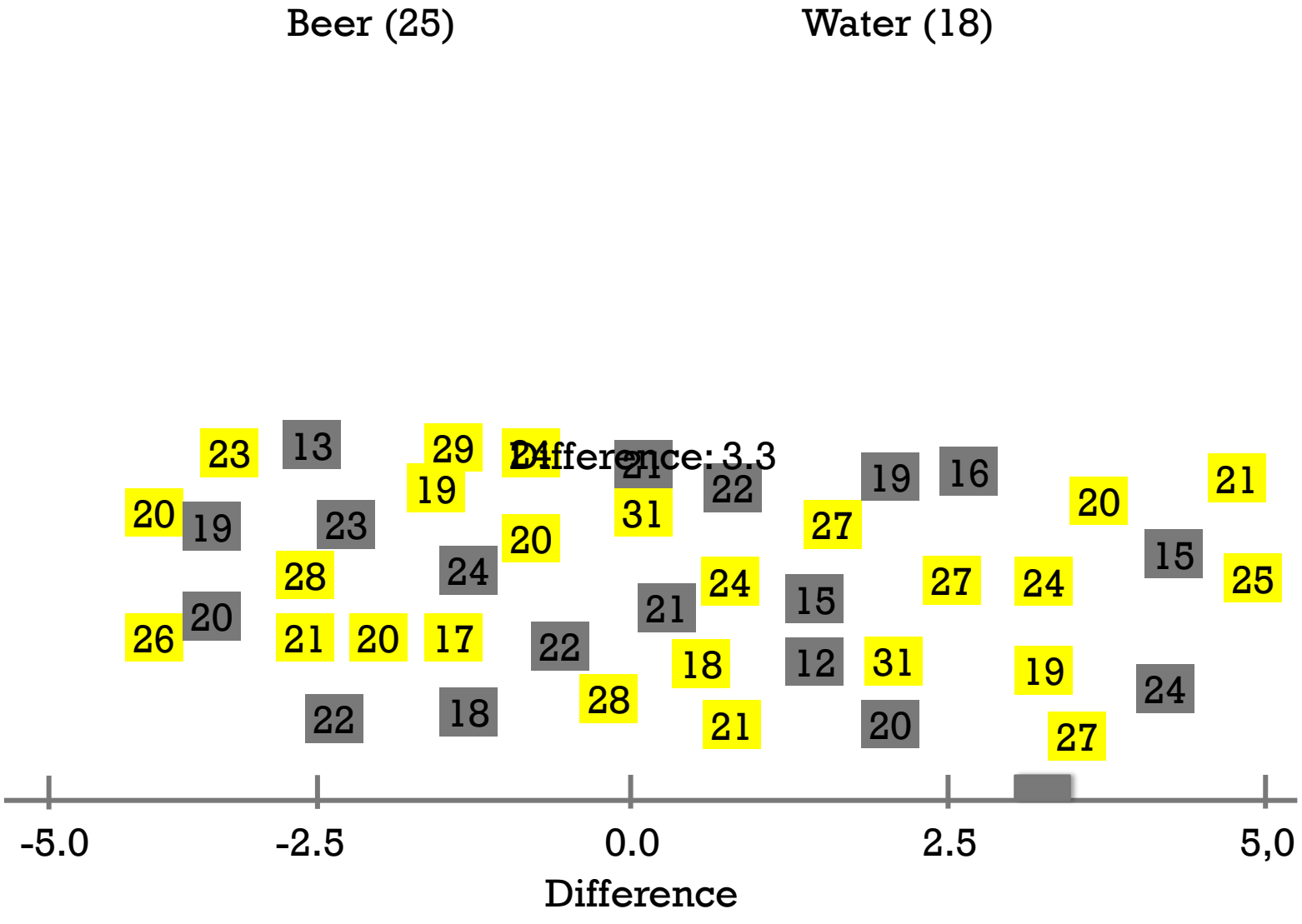
Difference: 4.4



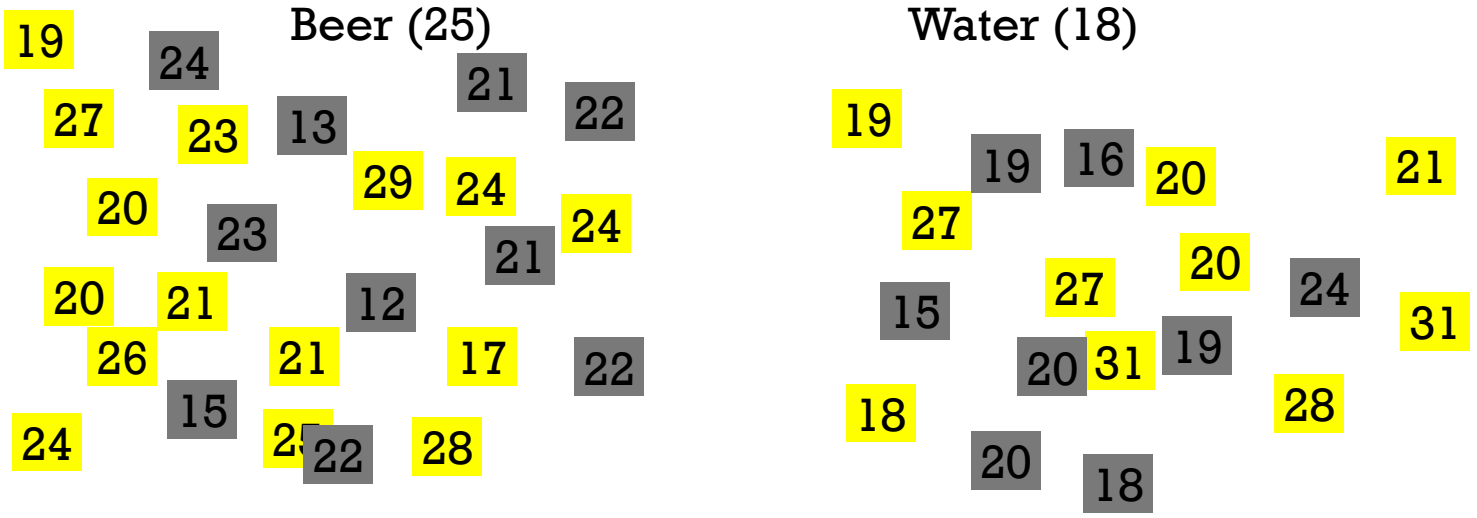
PERMUTATION TEST



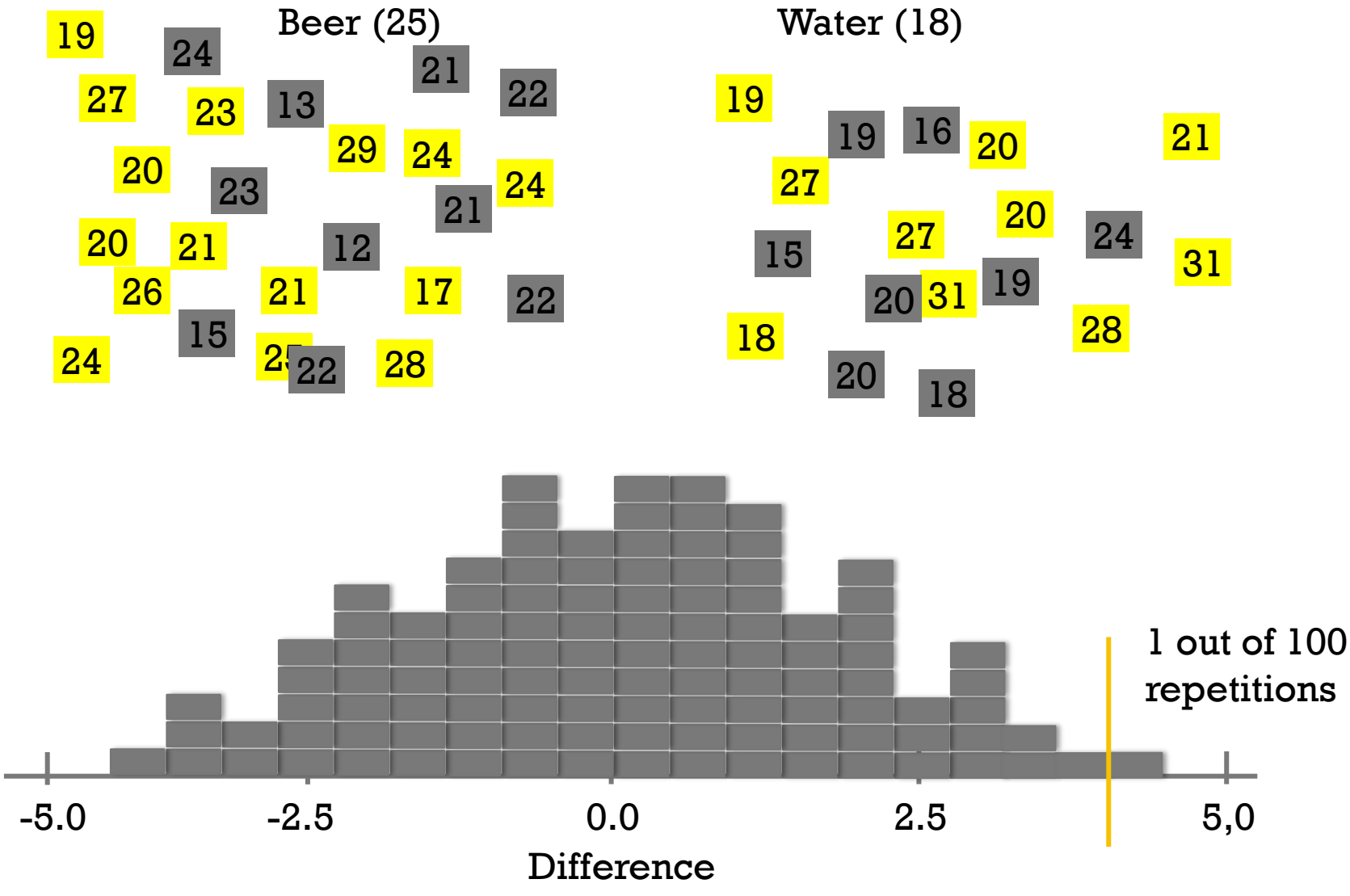
PERMUTATION TEST



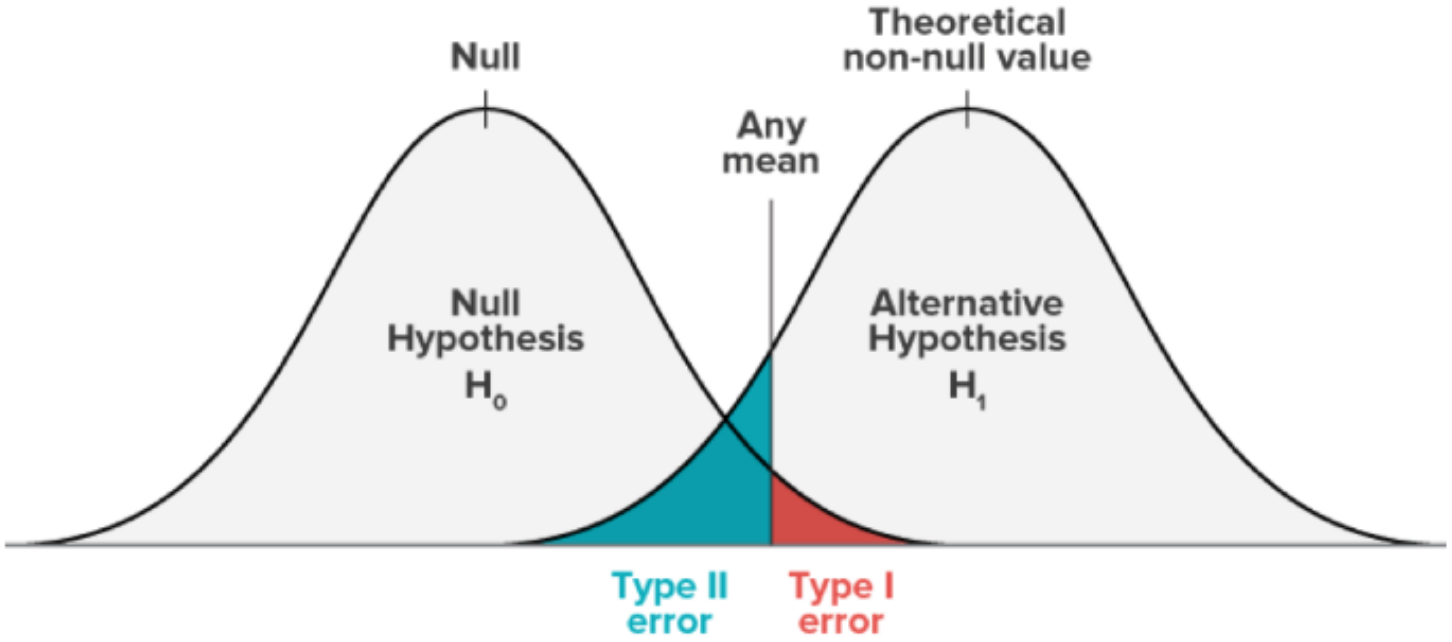
PERMUTATION TEST



PERMUTATION TEST



TYPE I VS TYPE II ERROR



The p value (Type I error) is the probability to obtain an effect equal to or more extreme than the one observed presuming the null hypothesis of no effect is true

P-VALUE HAS PROBLEMS!

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what are the implications for authors? The following are anticipated questions and their corresponding answers.

Question 1. *Will manuscripts with p-values be desk rejected automatically?*

Answer to Question 1. No. If manuscripts pass the

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The problems are well documented (Chihara, 1994; Fisher, 1973; Glymour, 1980; Popper, 1983; Suppes, 1994; Trafimow, 2003, 2005, 2006). However, there have been Bayesian proposals that at least somewhat circumvent

FICTIOUS EXAMPLE: BRAIN CANCER

Hypothesis: Brain cancer causes a headache

Data shows $p<0.01$ (considered very significant)

FICTIOUS EXAMPLE: BRAIN CANCER

Hypothesis: Brain cancer causes a headache

Data shows $p<0.01$ (considered very significant)

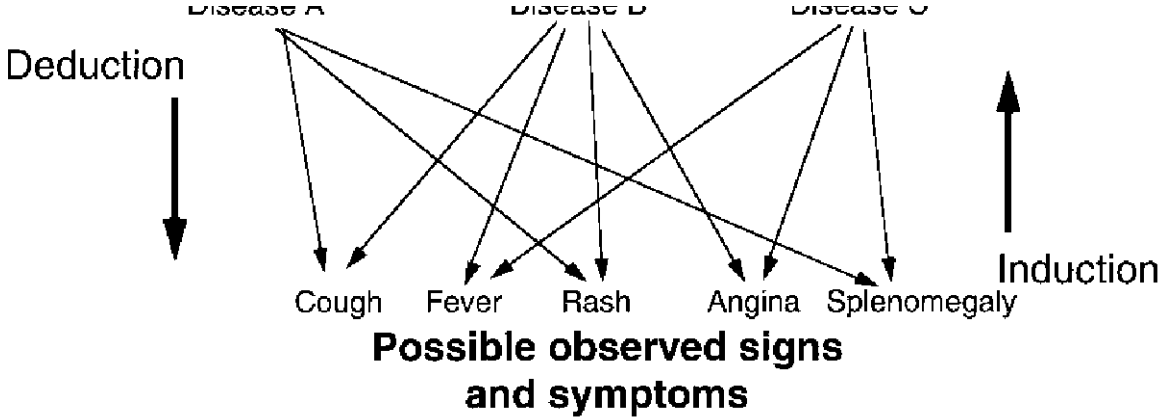
If you have a headache,
how likely is it that you have brain cancer?

FICTIOUS EXAMPLE: BRAIN CANCER

Hypothesis: Brain cancer causes a headache

Data shows $p<0.01$ (considered very significant)

If you have a headache,
how likely is it that you have brain cancer?



Statistical Inference

CLICKER: WHAT IS THE INTERPRETATION OF $P < 0.05$
HTTPS://CLICKER.MIT.EDU/6.S079/

A) The chances are greater than 1 in 20 that a difference would be found if the study were repeated.

B) The probability is less than 1 in 20 that a difference this large could occur by chance alone.

C) The probability is greater than 1 in 20 that a difference this large could occur by chance alone.

D) The chance is 95% that the study is correct

E) None of the above

MISCONCEPTION 1

"In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

What is wrong with this?

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.

MISCONCEPTION 1

"In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

This is an understand- able but categorically wrong interpretation because the *P* value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false. This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true. "

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.

MISCONCEPTION #1

"If P=.05, the null hypothesis has only a 5% chance of being true"

Let us suppose we flip a penny four times and observe four heads, two-sided P =.125. This does not mean that the probability of the coin being fair is only 12.5%.

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

MISCONCEPTION #2

A non significant difference (eg, P .05) means there is no difference between groups.

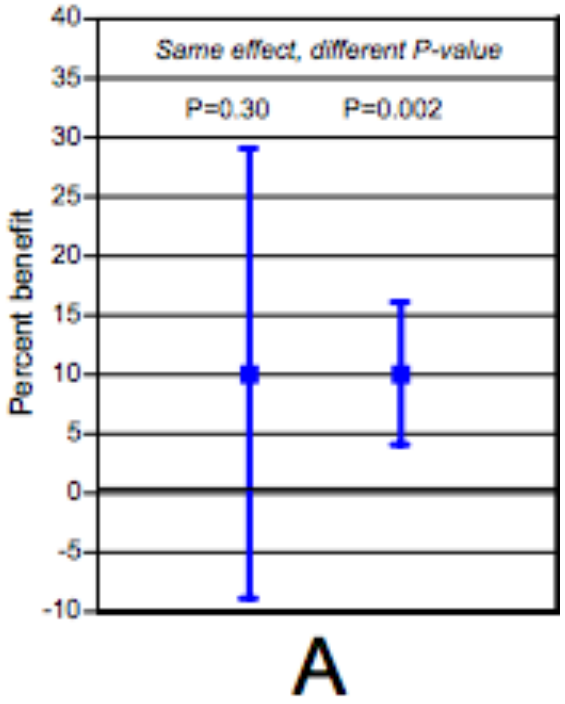
- A non significant difference only means the null effect is statistically consistent with the observation
- It does not make the null effect most likely
- In fact, the observed effect best explains the effect regardless the significance.

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

MISCONCEPTION #3

A statistically significant finding is (clinical) important

The P value carries no information about the magnitude of an effect, which is captured by the effect estimate and confidence interval.



MISCONCEPTION #4

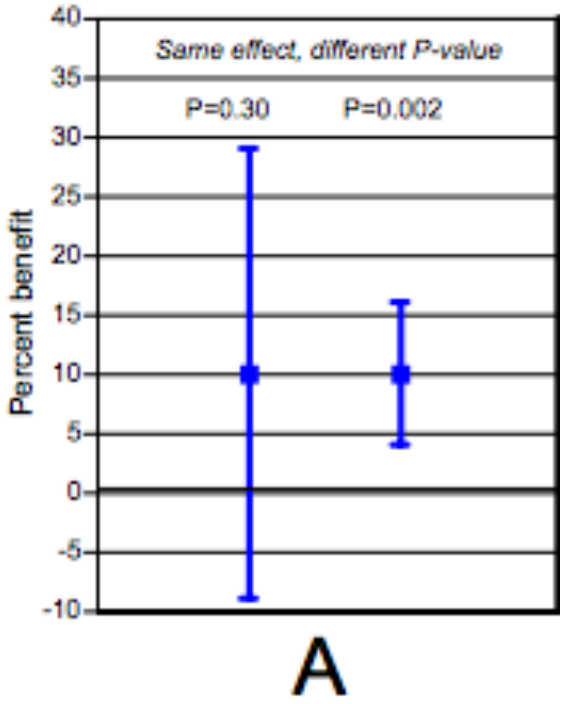
"Studies with P values on opposite sides of .05 are conflicting"

H_0:Drug T has no effect

H_1:Drug T has a positive effect

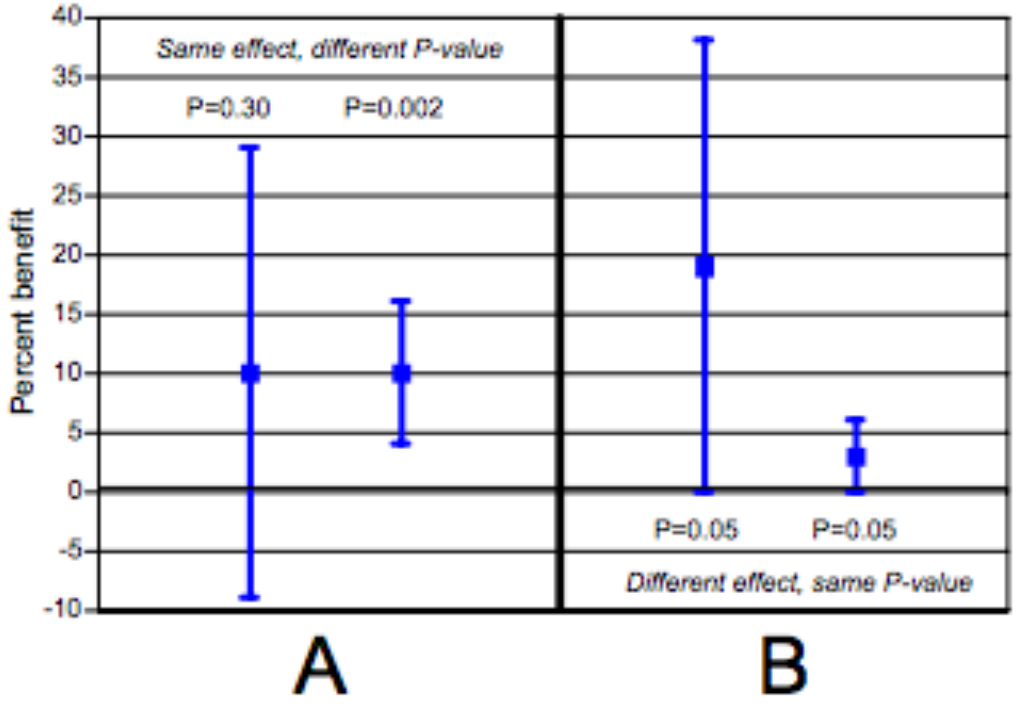
Study I: P=0.3

Study II: P=0.002



MISCONCEPTION #5

Studies with the same P value provide the same evidence against the null hypothesis



P-VALUE HAS PROBLEMS!

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what are the implications for authors? The following are anticipated questions and their corresponding answers.

Question 1. *Will manuscripts with p-values be desk rejected automatically?*

Answer to Question 1. No. If manuscripts pass the

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The problems are well documented (Chihara, 1994; Fisher, 1973; Glymour, 1980; Popper, 1983; Suppes, 1994; Trafimow, 2003, 2005, 2006). However, there have been Bayesian proposals that at least somewhat circumvent

P-HACKING (ALSO DATA DREDGING, DATA FISHING, DATA SNOOPING, DATA BUTCHERY)

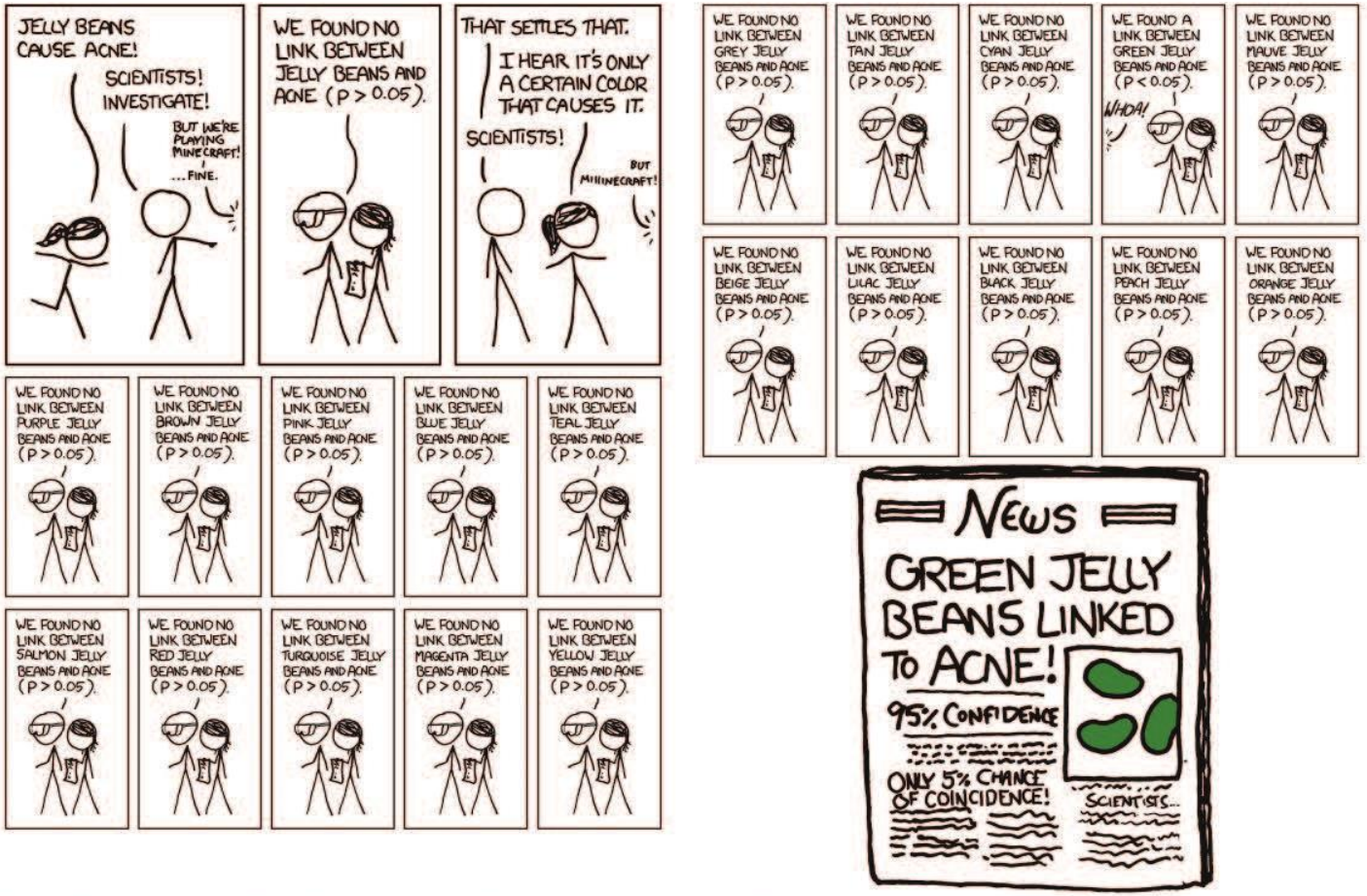


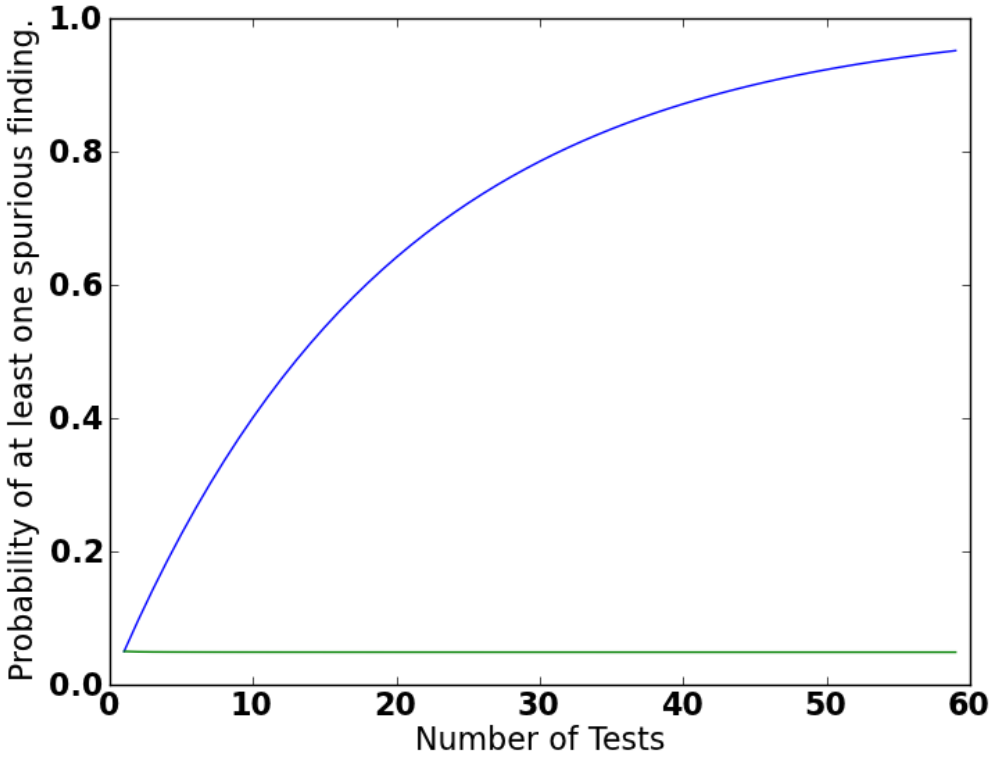
Figure 1. There is no overall effect of jelly beans on acne. Bummer. How about subgroups? Often subgroups are explored without alerting the reader to the number of questions at issue. Courtesy xkcd, http://xkcd.com/882/

P(detecting an effect when there is none) = $\alpha = 0.05$

P(detecting an effect when it exists) = $1 - \alpha$

P(detecting an effect when it exists on every experiment) = $(1 - \alpha)^k$

P(detecting an effect when there is none on <u>at least one</u> experiment) = $1 - (1 - \alpha)^k$

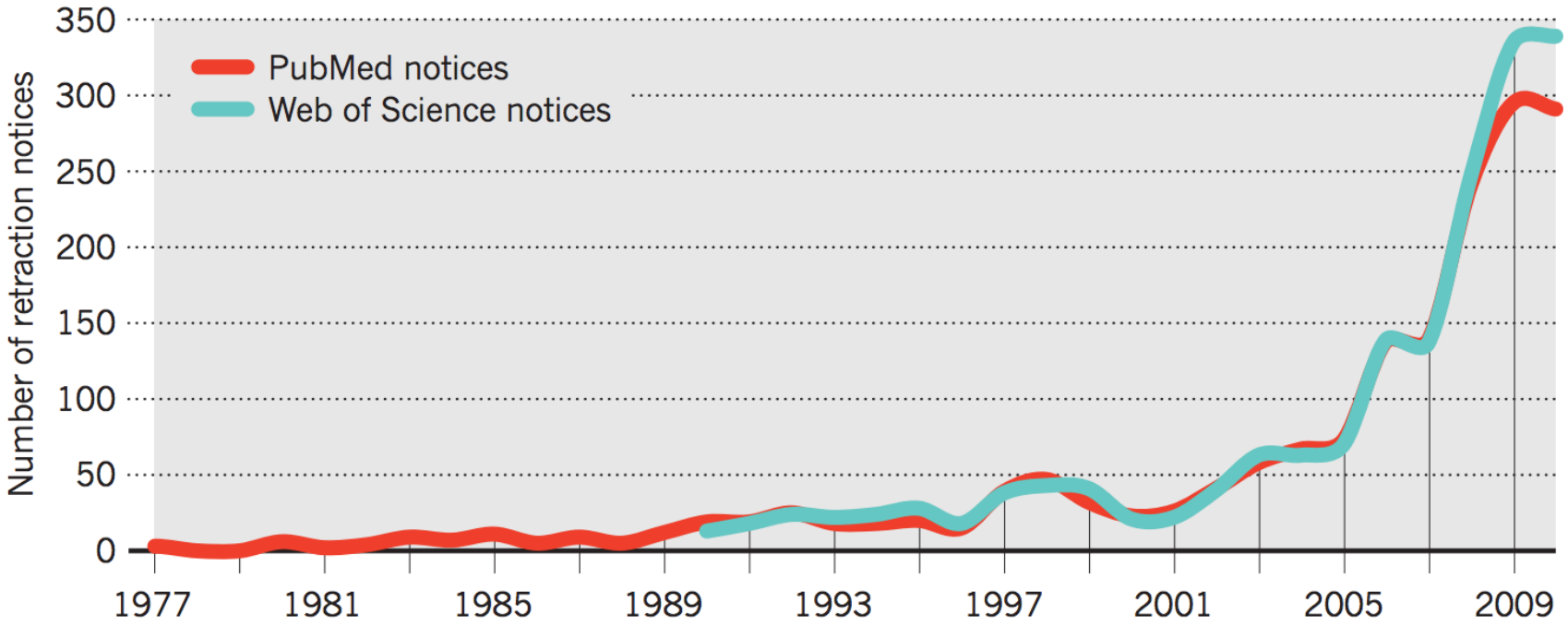


$\alpha = 0.05$

"Familywise Error Rate"

MISTAKES AND FRAUD

2001 – 2011:
- 10X increase in retractions
- only 1.44X increase in papers

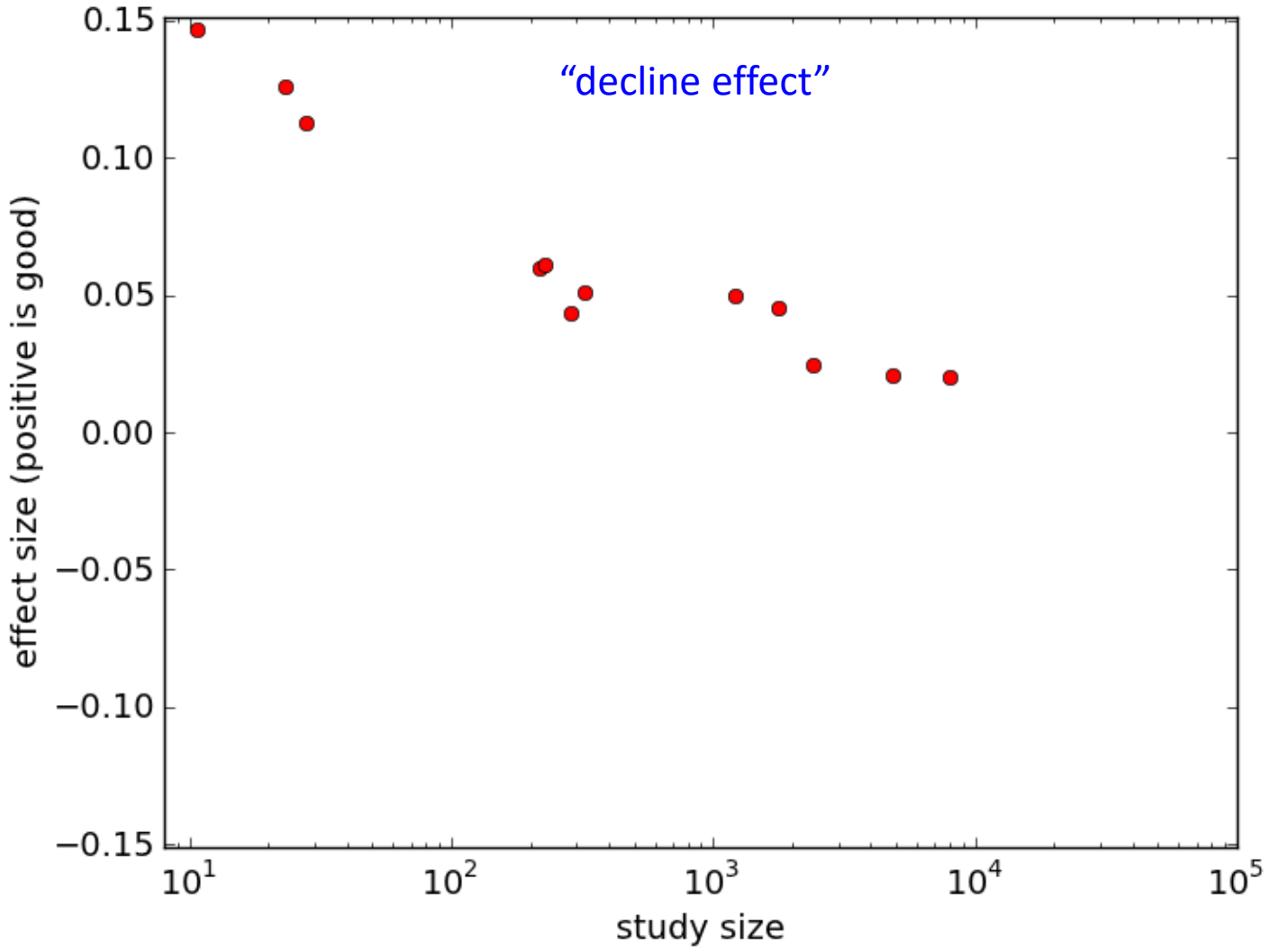


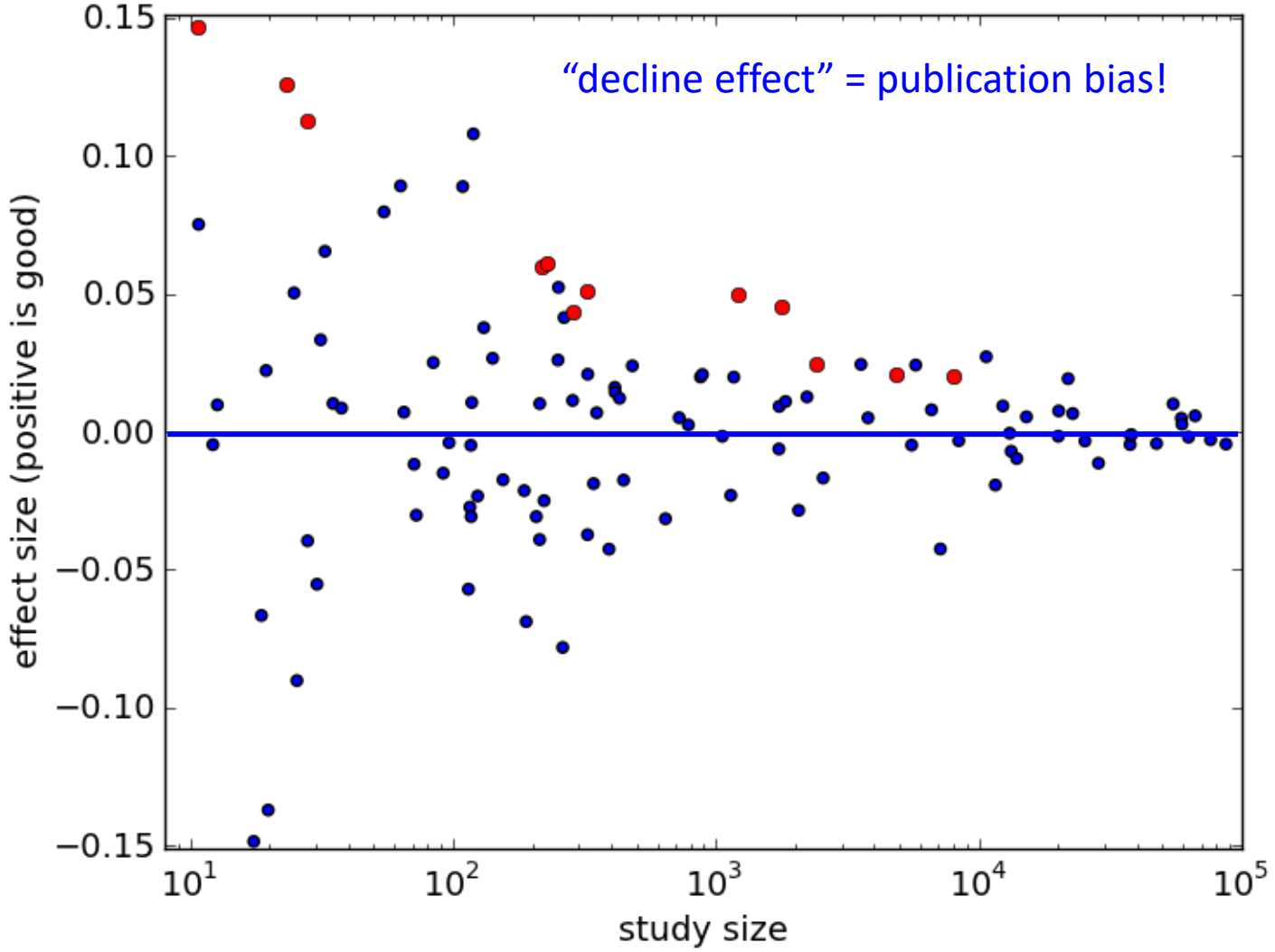
Richard Van Noorden, 2011, Nature 478
The Rise of the Retractions

http://www.nature.com/news/2011/111005/pdf/478026a.pdf

85

PUBLICATION BIAS




"decline effect" = publication bias!
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
effect size (positive is good)
10^1
10^2
10^3
10^4
10^5
study size

FAMILY-WISE ERROR RATE CORRECTIONS

Bonferroni Correction

- Just divide by the number of hypotheses

$$\alpha_c = \frac{\alpha}{k}$$

Šidák Correction

- Asserts independence

$$\alpha = 1 - (1 - \alpha_c)^k$$

$$\alpha_c = 1 - (1 - \alpha)^{\frac{1}{k}}$$

MANY ANALYSTS, ONE DATA SET

MANY ANALYSTS, ONE DATA SET

Variations in Analytic Choices Affect Results

Abstract:

"**Twenty-nine teams involving 61 analysts used the same data** set to address the same research question: whether soccer referees are more likely to give red cards to dark-skin-toned players than to light- skin-toned players. Analytic approaches varied widely across the teams, and the estimated effect sizes ranged from 0.89 to 2.93 (*Mdn* = 1.31) in odds-ratio units. **Twenty teams (69%) found a statistically significant positive effect, and 9 teams (31%) did not observe a significant relationship**. Overall, the 29 different analyses used 21 unique combinations of covariates. Neither analysts' prior beliefs about the effect of interest nor their level of expertise readily explained the variation in the outcomes of the analyses. …. Crowdsourcing data analysis, a strategy in which numerous research teams are recruited to simultaneously investigate the same research question, makes transparent how defensible, yet subjective, analytic choices influence research results."

CLOSING THOUGHTS

"It is easy to lie with statistics, but it is easier to lie without them."

attributed to Frederick Mosteller (1916-2006)

REFERENCES

- How to lie with Statistics - Darrell Huff
- How to lie with Maps - Mark Monmonier
- http://www.sciencebasedmedicine.org/psychology-journal-bans-significance-testing/
- Nuzzo R: Scientific method: statistical errors. Nature. 2014 Feb 13;506(7487)
- Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.
- Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999;130:1005-13.