CS 109: Data Science

Exploratory Data Analysis & Effective Visualizations

Hanspeter Pfister
pfister@seas.harvard.edu

Joe Blitzstein
blitzstein@stat.harvard.edu

Verena Kaynig
vkaynig@seas.harvard.edu

This Week

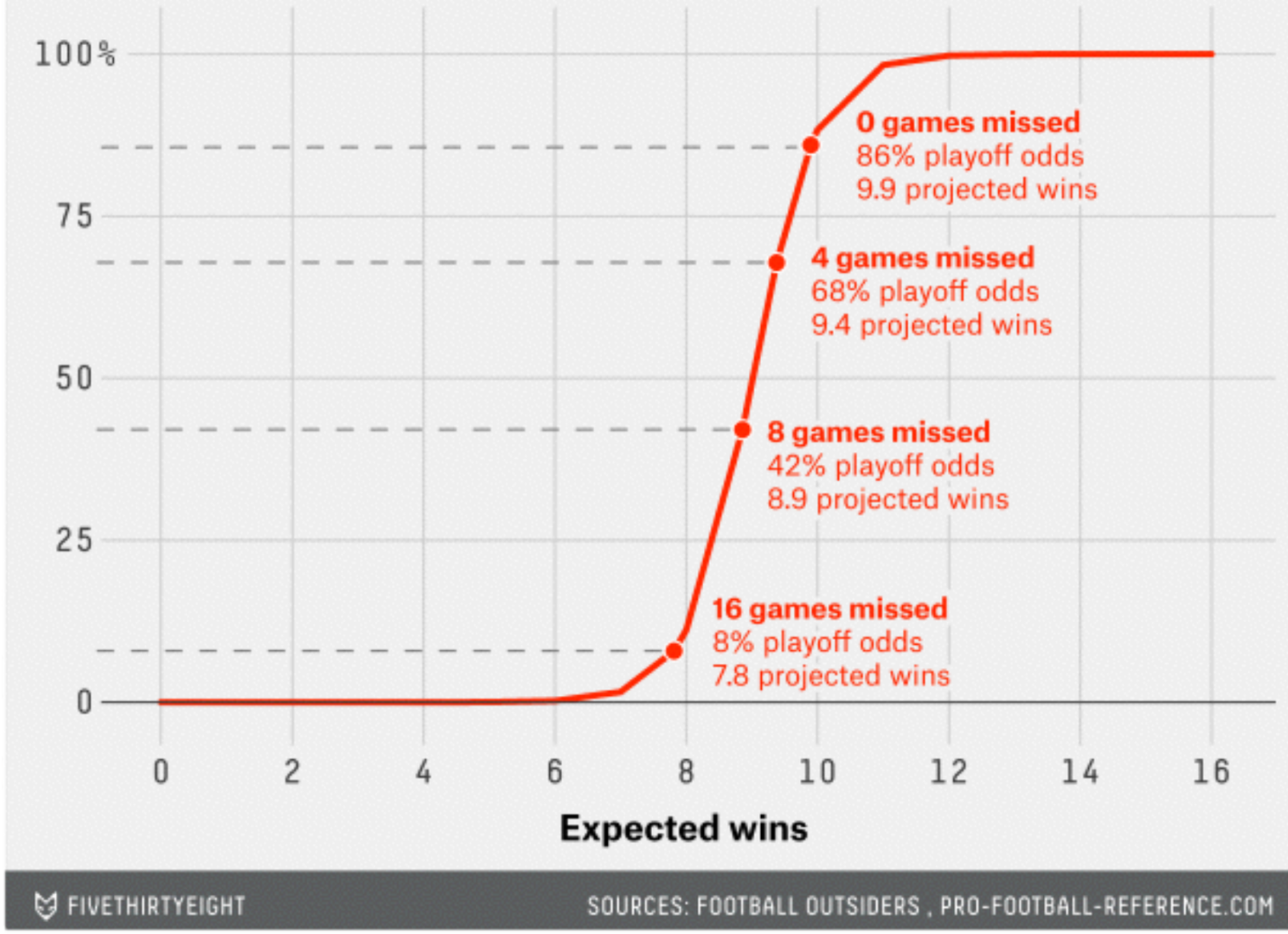
- HW0 - due today (not graded)
- HW1 - out today, due Th 9/24

  Check syllabus for grading / late day / collaboration policies

- Sectioning - keep an eye on Piazza for information on how to indicate preferences

Brady's Value Comes In The Playoff Sweet Spot

Probability of Patriots making playoffs depending on Brady's missed games, modeled using DYAR and seasonal data from 1990-2014



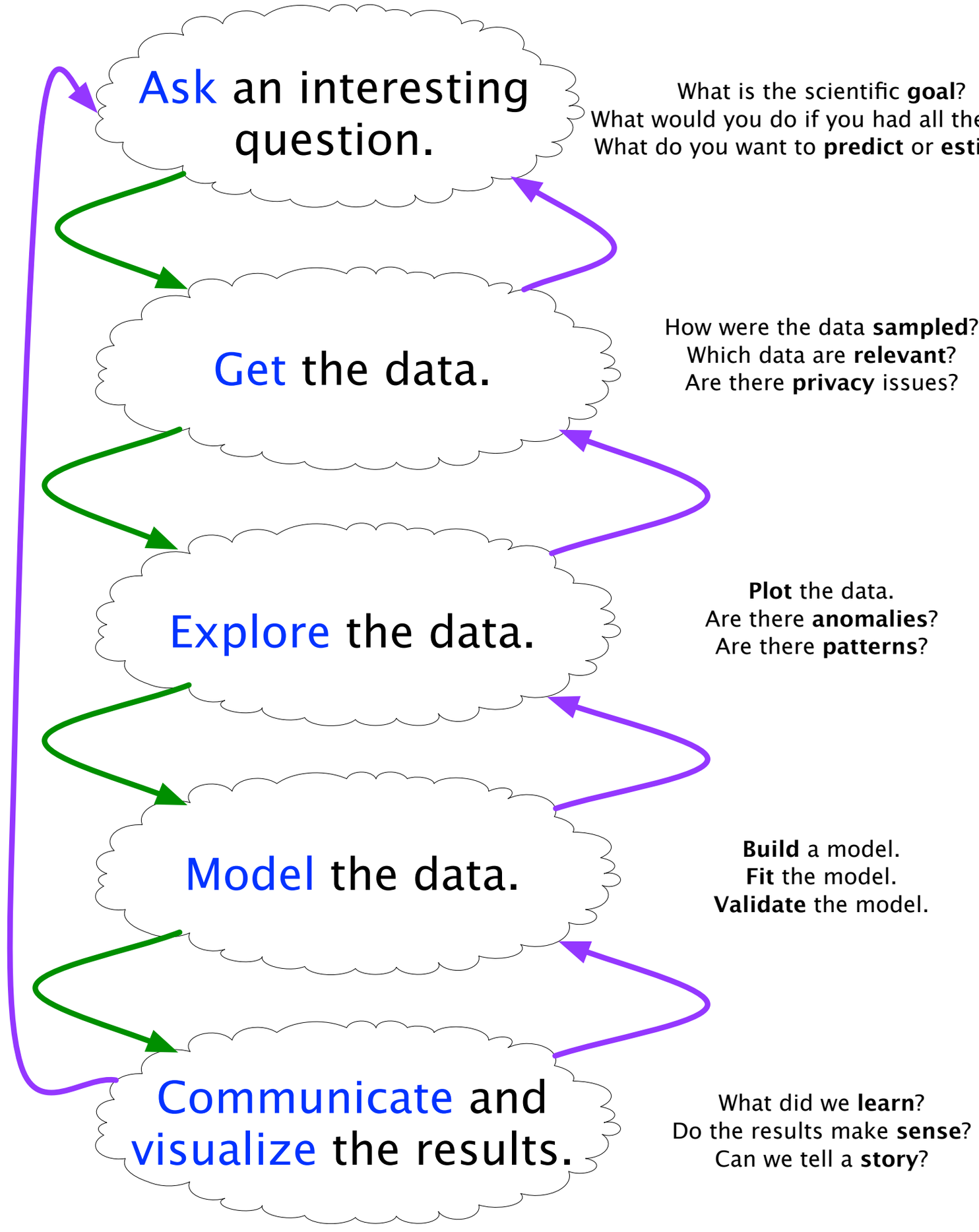
FIVETHIRTYEIGHT SOURCES: FOOTBALL OUTSIDERS , PRO-FOOTBALL-REFERENCE.COM

2015 NF
Brady, T
The AFC

By HARRY ENTEN

In preparation for the 2015 NFL season, FiveThirtyEight is running a series of

sement Primary


Ask an interesting question.
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?
Get the data.
How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?
Explore the data.
Plot the data.
Are there **anomalies**?
Are there **patterns**?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

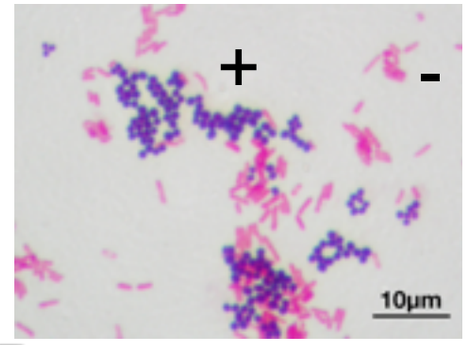
Data Exploration

Not always sure what we are looking for (until we find it)

Example: Antibiotics

Will Burtin, 1951

Data



Genus, *Species*

Min. Inhibitory Concentration [ml/g]

Table 1: Burtin's data.

Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	850	1.2	1	negative
Mycobacterium *tuberculosis*	800	5	2	negative
Proteus *vulgaris*	3	0.1	0.1	negative
Pseudomonas *aeruginosa*	850	2	0.4	negative
Salmonella (Eberthella) *typhosa*	1	0.4	0.008	negative
Salmonella *schottmuelleri*	10	0.8	0.09	negative
Staphylococcus *albus*	0.007	0.1	0.001	positive
Staphylococcus *aureus*	0.03	0.03	0.001	positive
Streptococcus *fecalis*	1	1	0.1	positive
Streptococcus *hemolyticus*	0.001	14	10	positive
Streptococcus *viridans*	0.005	10	40	positive

What Questions?

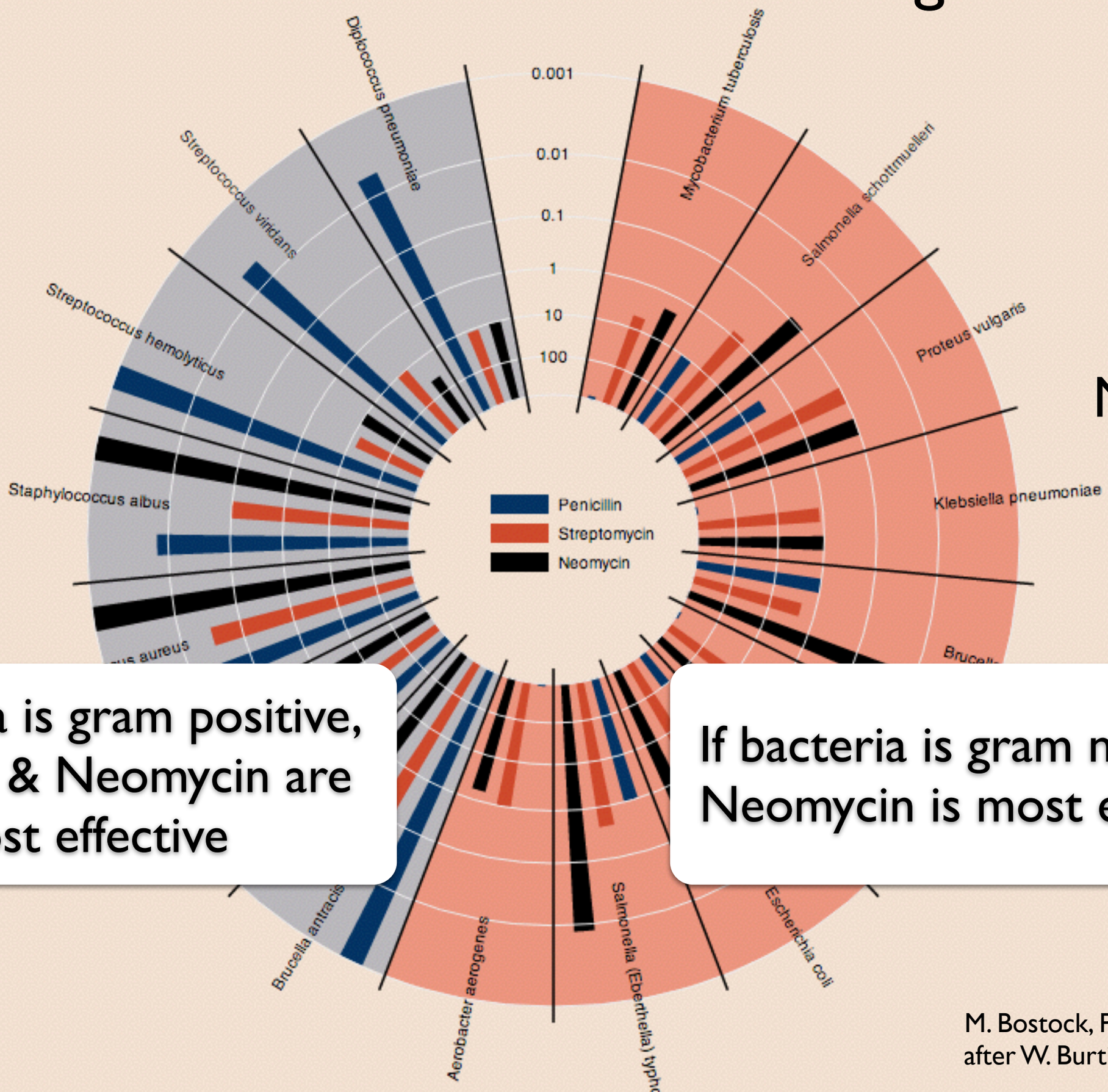
Table 1: Burtin's data.

Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	850	1.2	1	negative
Mycobacterium *tuberculosis*	800	5	2	negative
Proteus *vulgaris*	3	0.1	0.1	negative
Pseudomonas *aeruginosa*	850	2	0.4	negative
Salmonella (Eberthella) *typhosa*	1	0.4	0.008	negative
Salmonella *schottmuelleri*	10	0.8	0.09	negative
Staphylococcus *albus*	0.007	0.1	0.001	positive
Staphylococcus *aureus*	0.03	0.03	0.001	positive
Streptococcus *fecalis*	1	1	0.1	positive
Streptococcus *hemolyticus*	0.001	14	10	positive
Streptococcus *viridans*	0.005	10	40	positive

How effective are the drugs?

Gram Positive

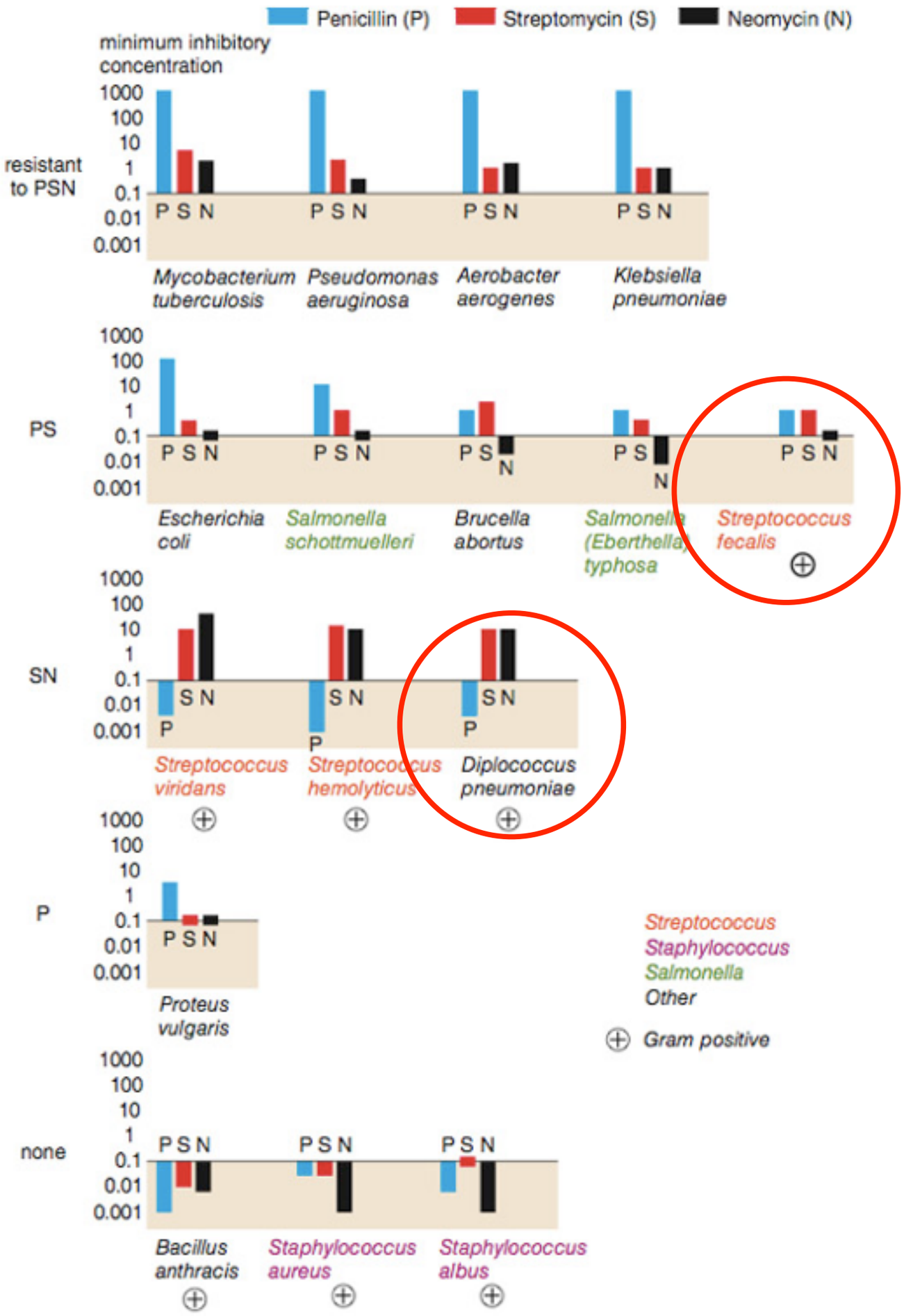
Gram Negative



If bacteria is gram positive, Penicillin & Neomycin are most effective

If bacteria is gram negative, Neomycin is most effective

M. Bostock, Protovis after W. Burtin, 1951



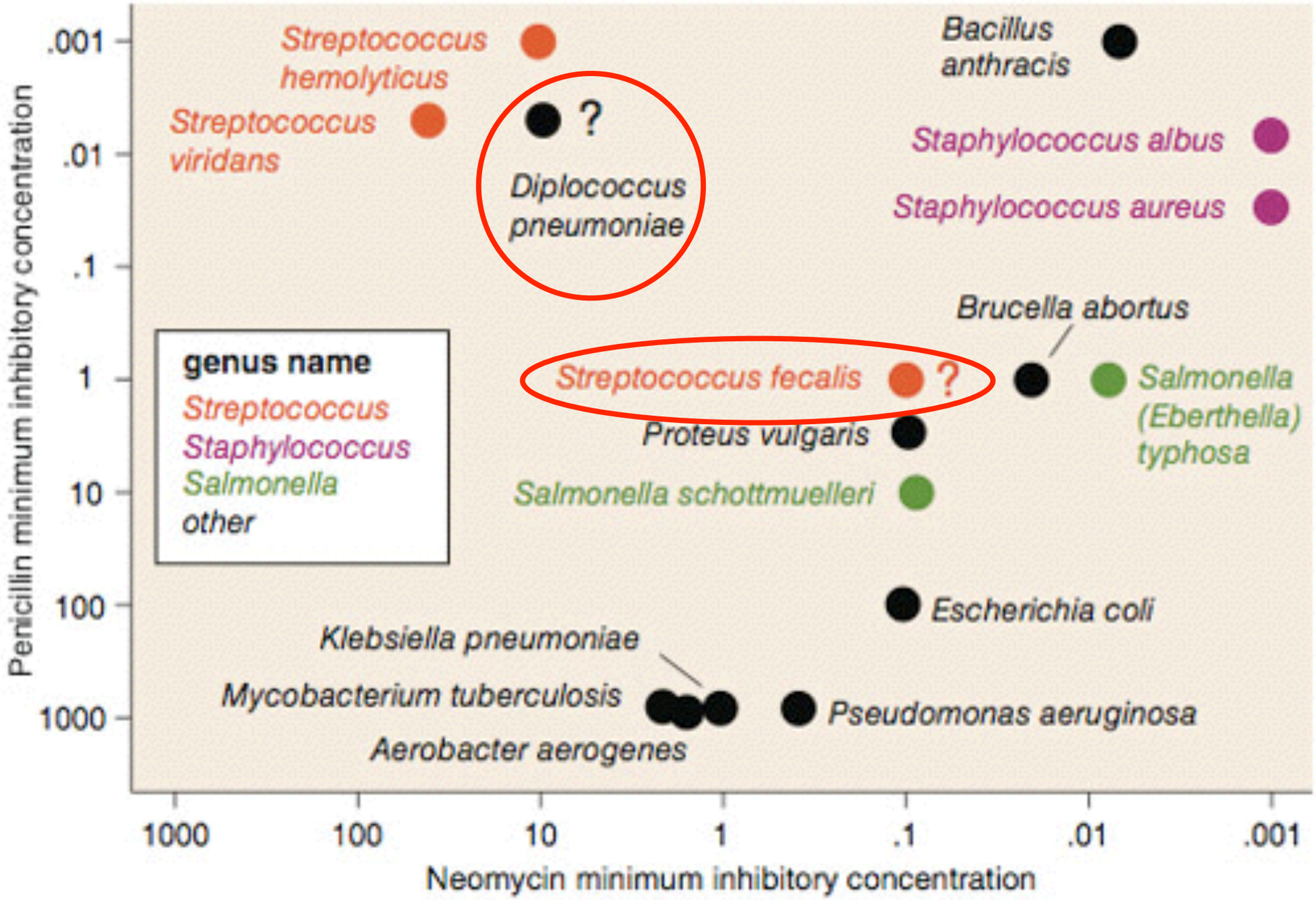
How do the bacteria compare?

Not a streptococcus!
(realized ~30 years later)

Really a streptococcus!
(realized ~20 years later)

Wainer & Lysen, "That's funny..."
American Scientist, 2009
Adapted from Brian Schmotzer

How do the bacteria compare?



Wainer & Lysen, "That's funny..."
American Scientist, 2009

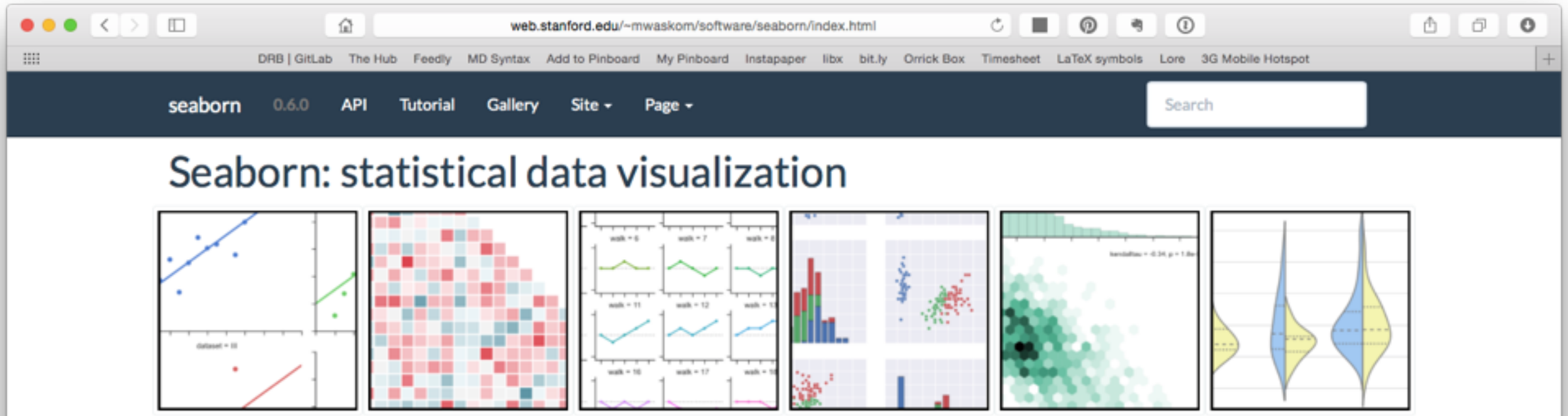
Exploratory Data Analysis

"The greatest value of a picture is when it forces us to notice what we never expected to see."

John Tukey

Visualization

To convey information through graphical representations of data



Visualization Goals

Communicate (Explanatory)

Present data and ideas

Explain and inform

Provide evidence and support

Influence and persuade

Analyze (Exploratory)

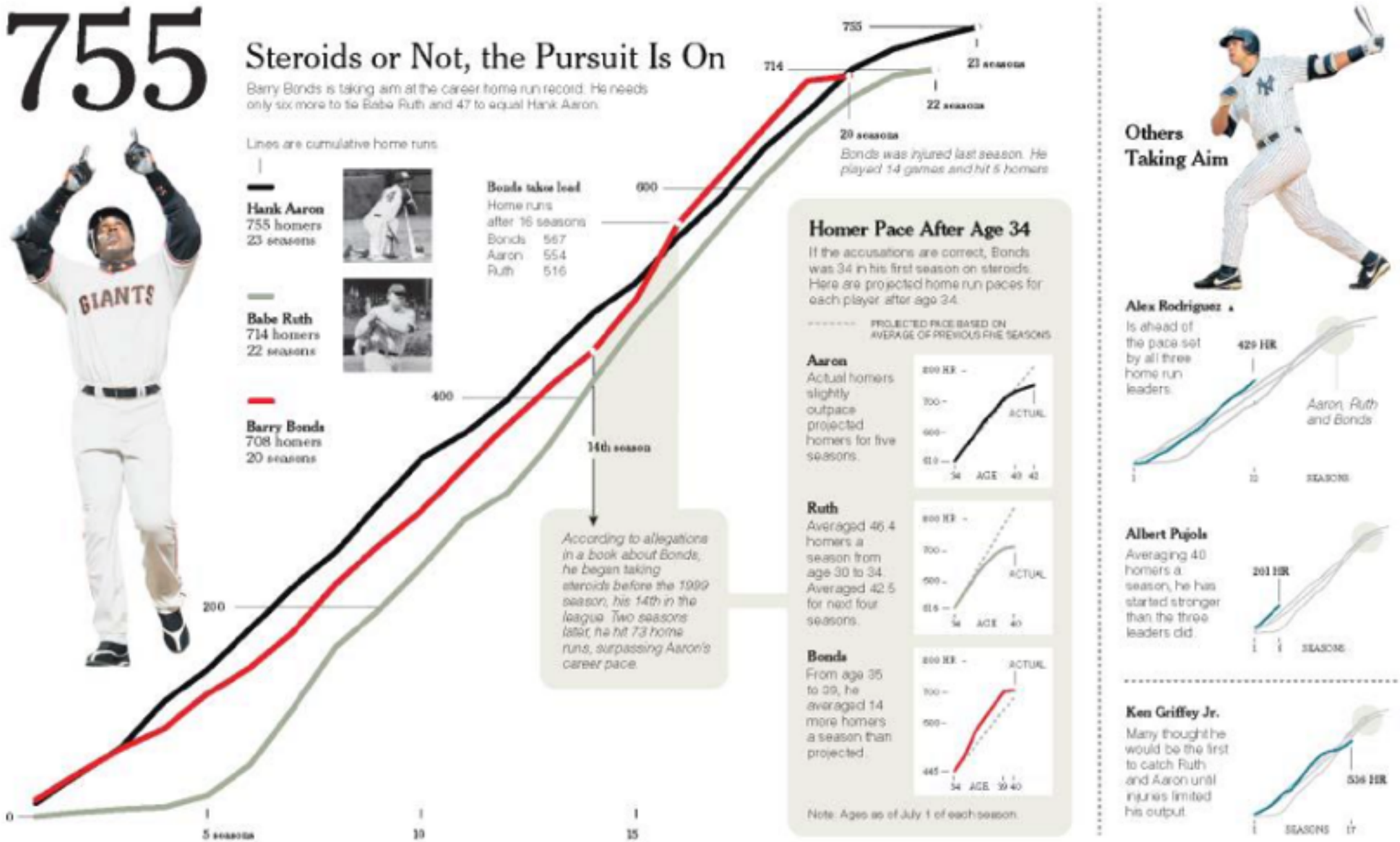
Explore the data

Assess a situation

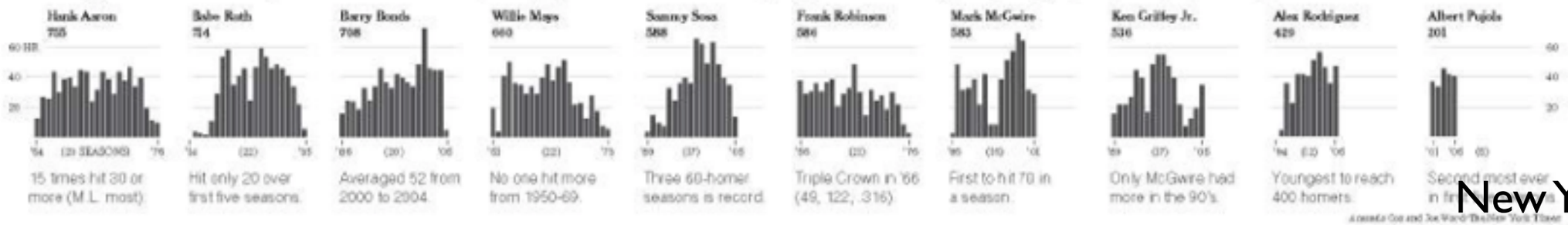
Determine how to proceed

Decide what to do

Communicate



Differing Paths to the Top of the Charts The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).



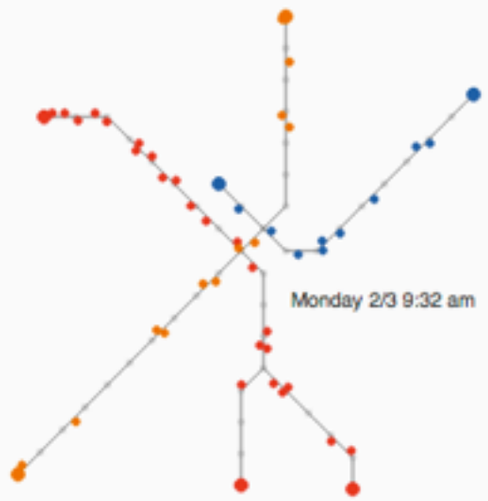
New York Times

Explore

Visualizing MBTA Data

An interactive exploration of Boston's subway system

Mike Barry and Brian Card - June 10, 2014



Boston's Massachusetts Bay Transit Authority (MBTA) operates the 4th busiest subway system in the U.S. after New York, Washington, and Chicago. If you live in or around the city you have probably ridden on it. The MBTA recently began publishing substantial amount of subway data through its public APIs. They provide the full schedule in General Transit Feed Specification (GTFS) format which powers Google's transit directions. They also publish realtime train locations for the Red, Orange, and Blue lines (but not Green or Silver lines). The following visualizations use data captured from these feeds for the entire month of February, 2014. Also, working with the MBTA, we were able to acquire per-minute entry and exit counts at each station measured at the turnstiles used for payment.

We attempt to present this information to help people in Boston better understand the trains, how people use the trains, and how the people and trains interact with each other.

The Trains

In a typical weekday, trains make approximately 1150 trips on the red, orange, and blue lines starting at 5AM and continuing through 1AM the next morning. On Saturdays trains make 870 trips and on Sundays they make 760.

To better understand how the trains operate on a typical day, below are all trips that trains took on the red, orange, and blue lines on Monday February 3 2014. Each vertical line represents a station, and time extends from top to bottom. Steeper lines indicate slower trains. This visualization was first used by Étienne-Jules Marey to visualize train schedules and is typically called a "Marey Diagram."

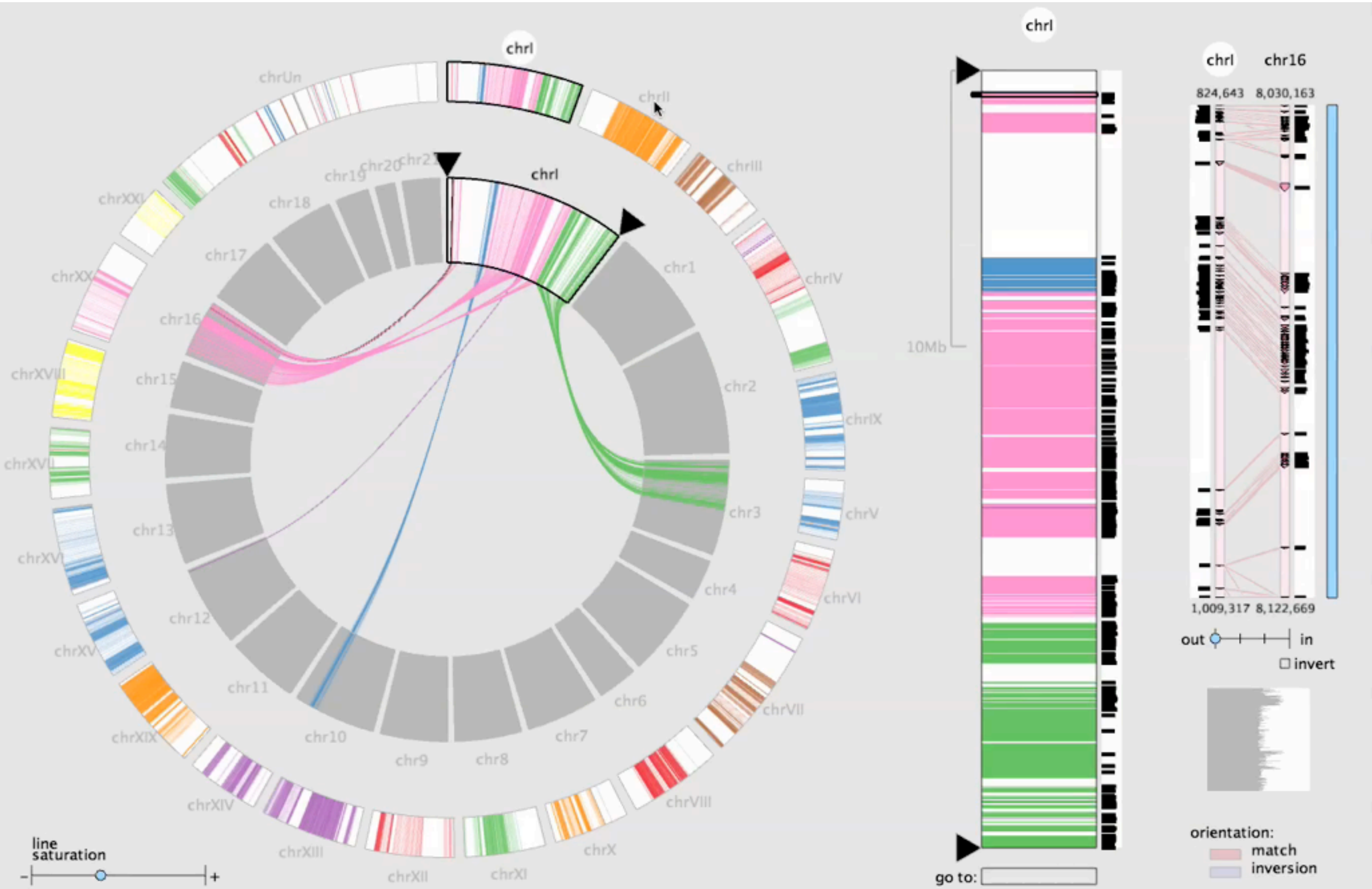
Average Number of Trips per Day

	Weekdays	Saturdays	Sundays
Red	450	350	300
Orange	320	260	220
Blue	380	260	240
Total	**1150**	**870**	**760**

Subway Trips on Monday February 3, 2014

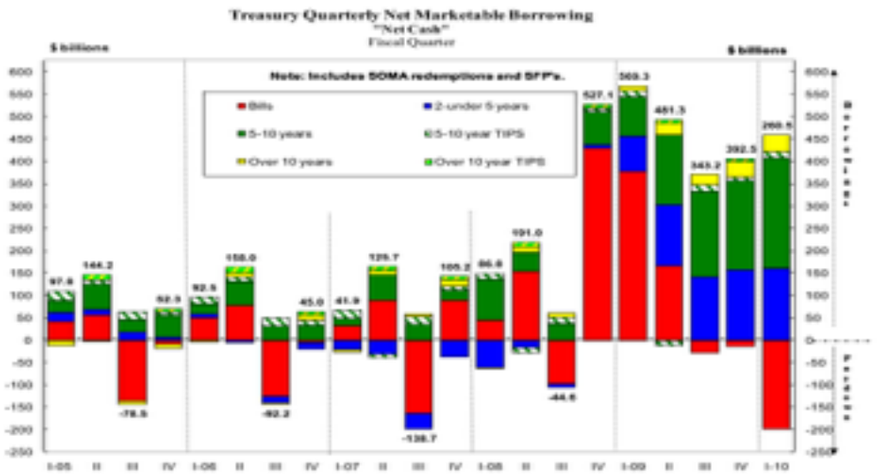
MizBee

[Meyer et al. 2009]



Effective Visualizations

Not Effective....



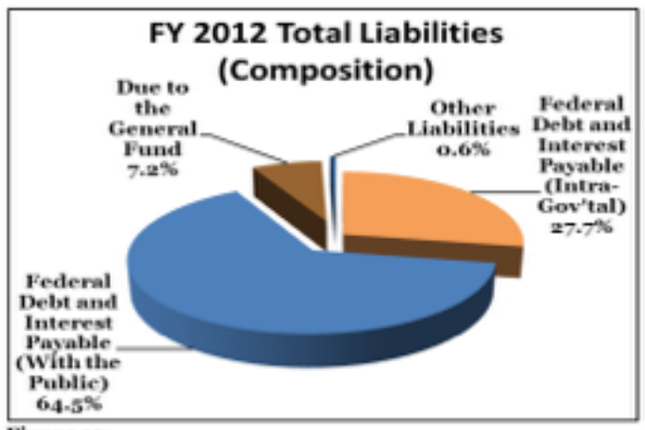
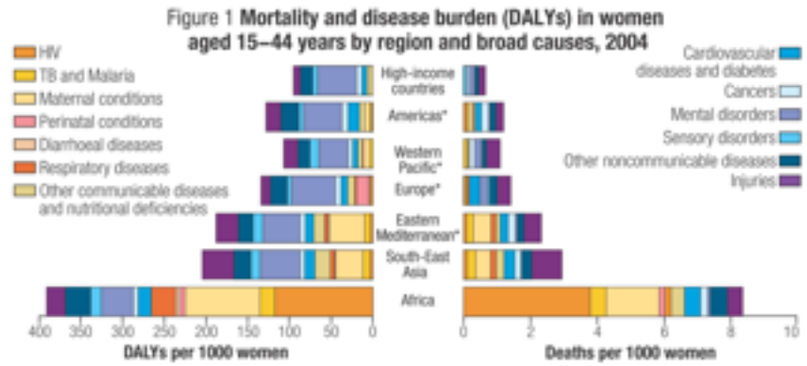
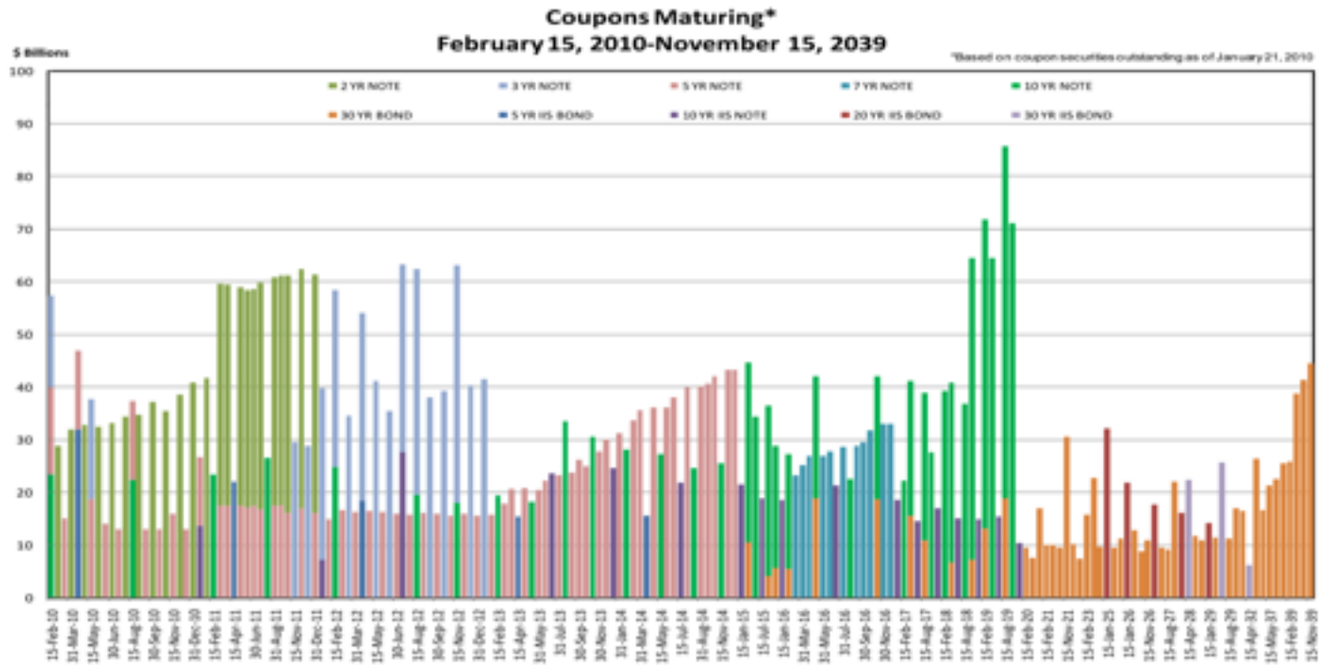


Figure 10





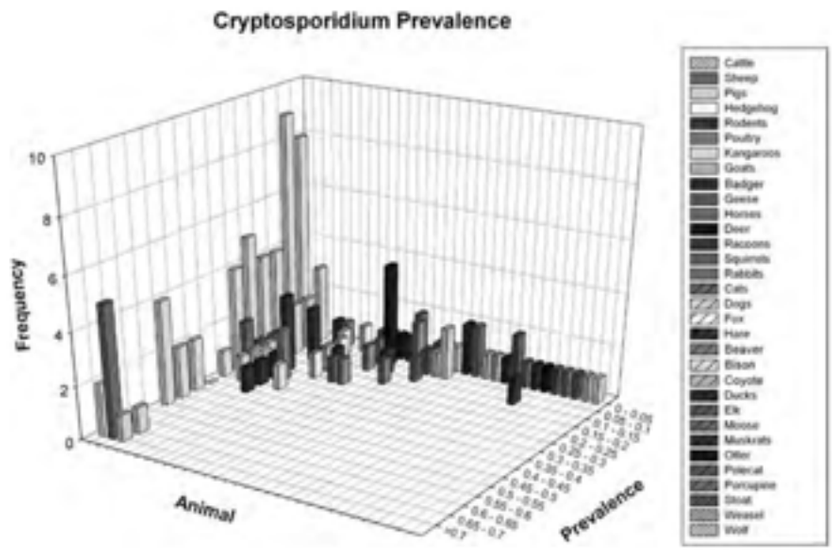


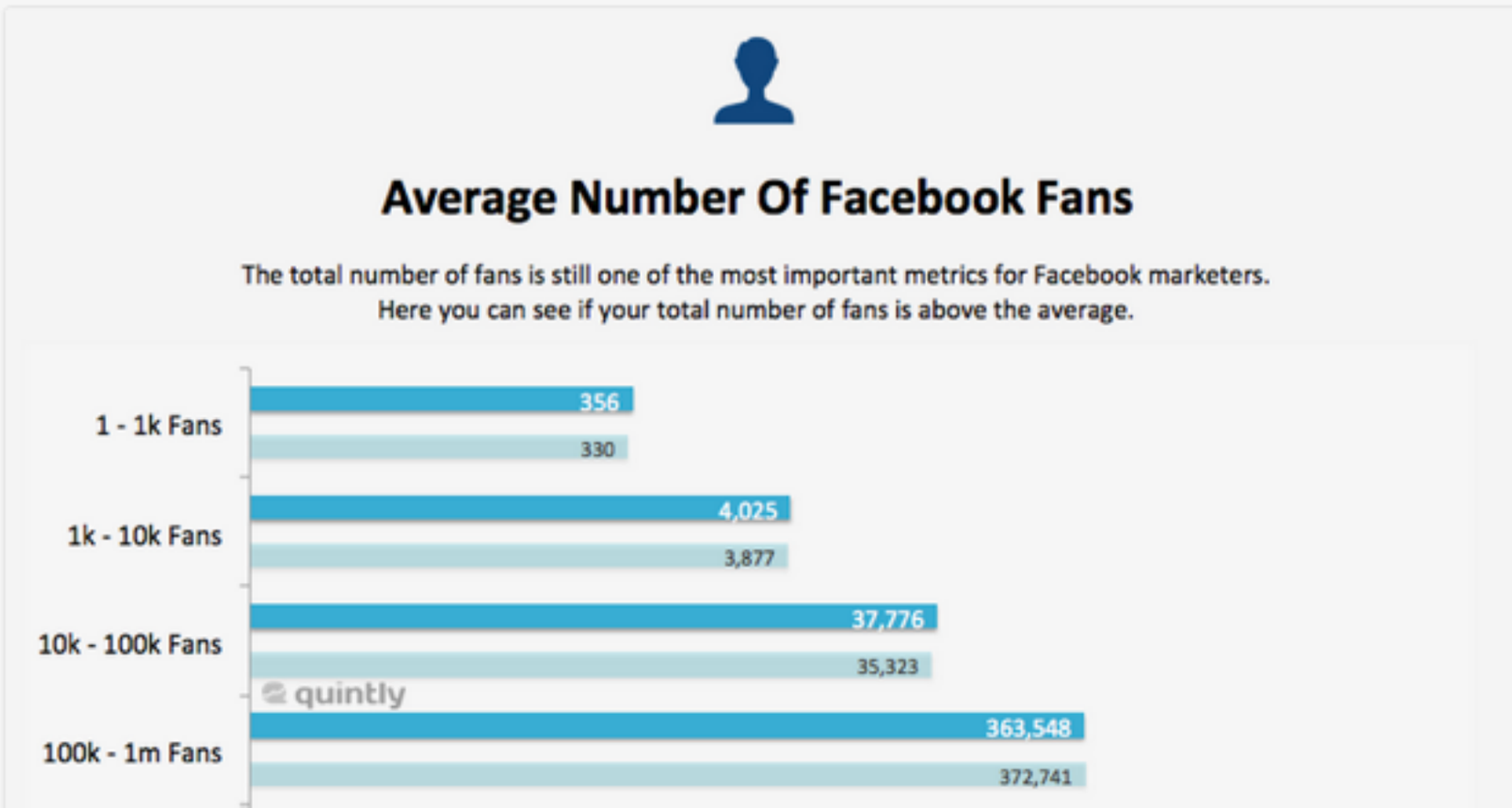
Figure 5.2 Mean prevalence rates of *Cryptosporidium* oocysts by animal species.

Sources: US Treasury and WHO reports

http://viz.wtf

WTF Visualizations

Visualizations that make no sense.
For a discussion of what is wrong with a particular visualization, tweet at us @WTFViz.
Check out our friends Thumbs Up Viz and accidental aRt, or submit.

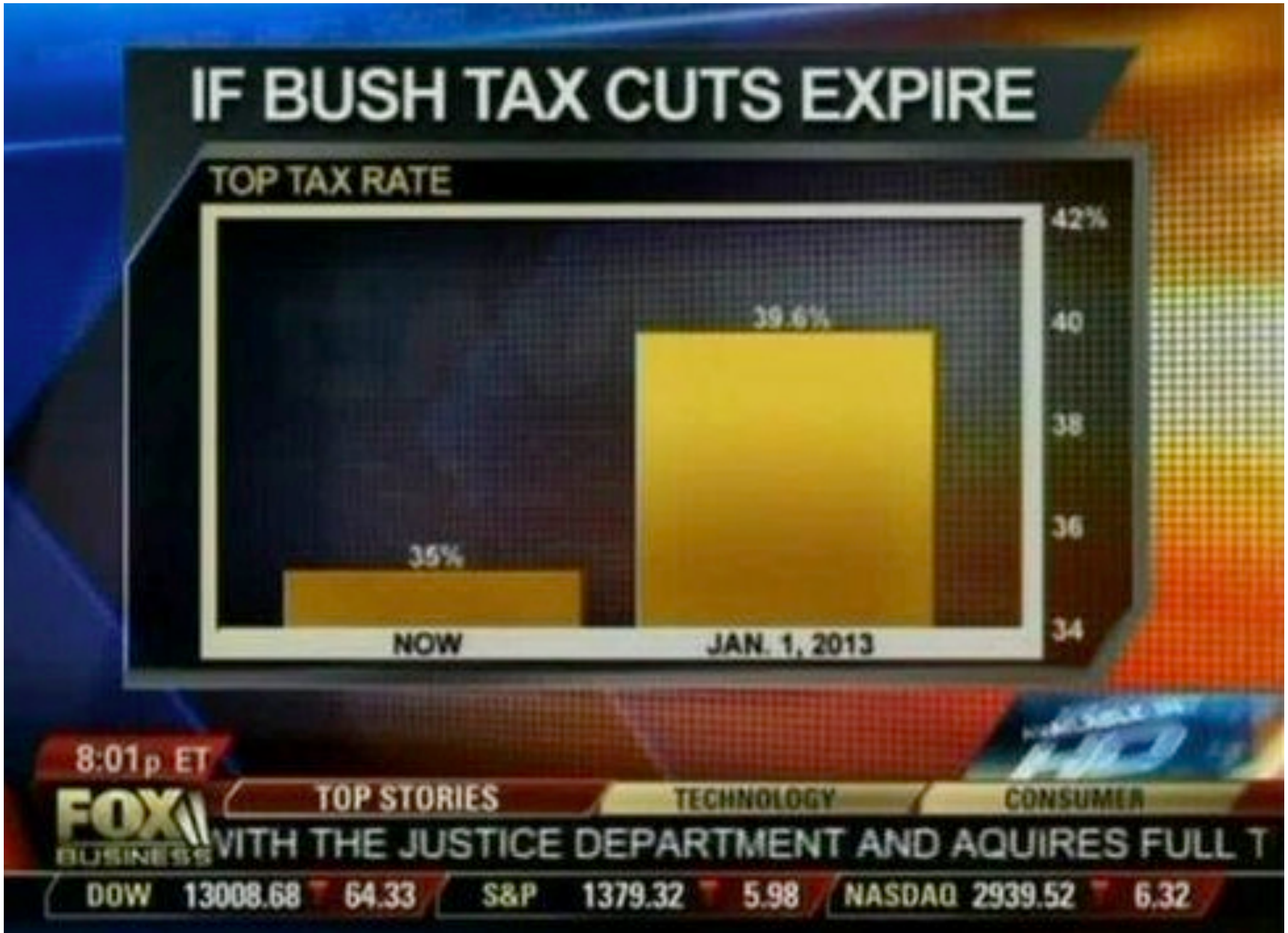


Effective Visualizations

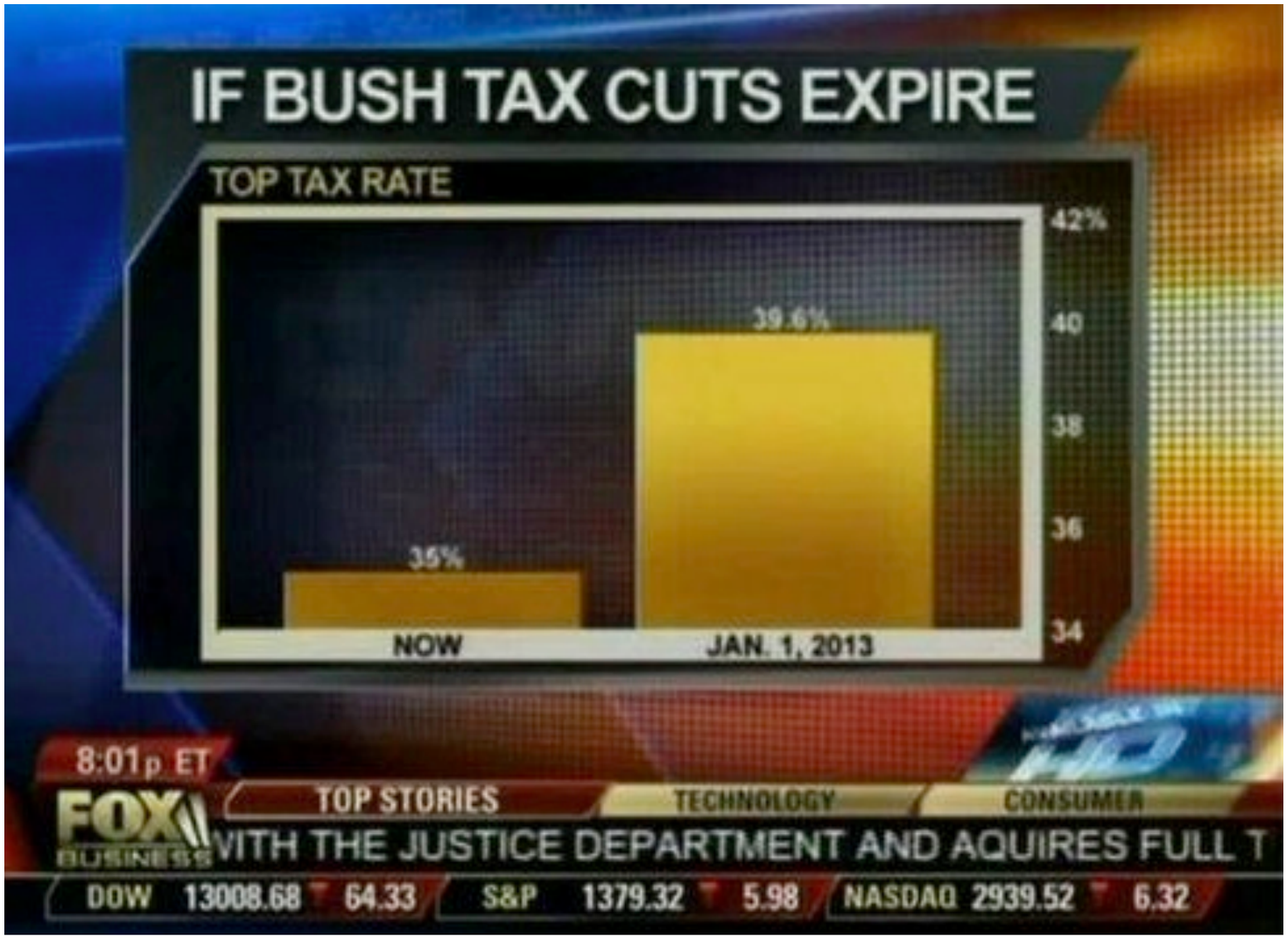
1. Have graphical integrity
2. Keep it simple
3. Use the right display
4. Use color strategically
5. Tell a story with data

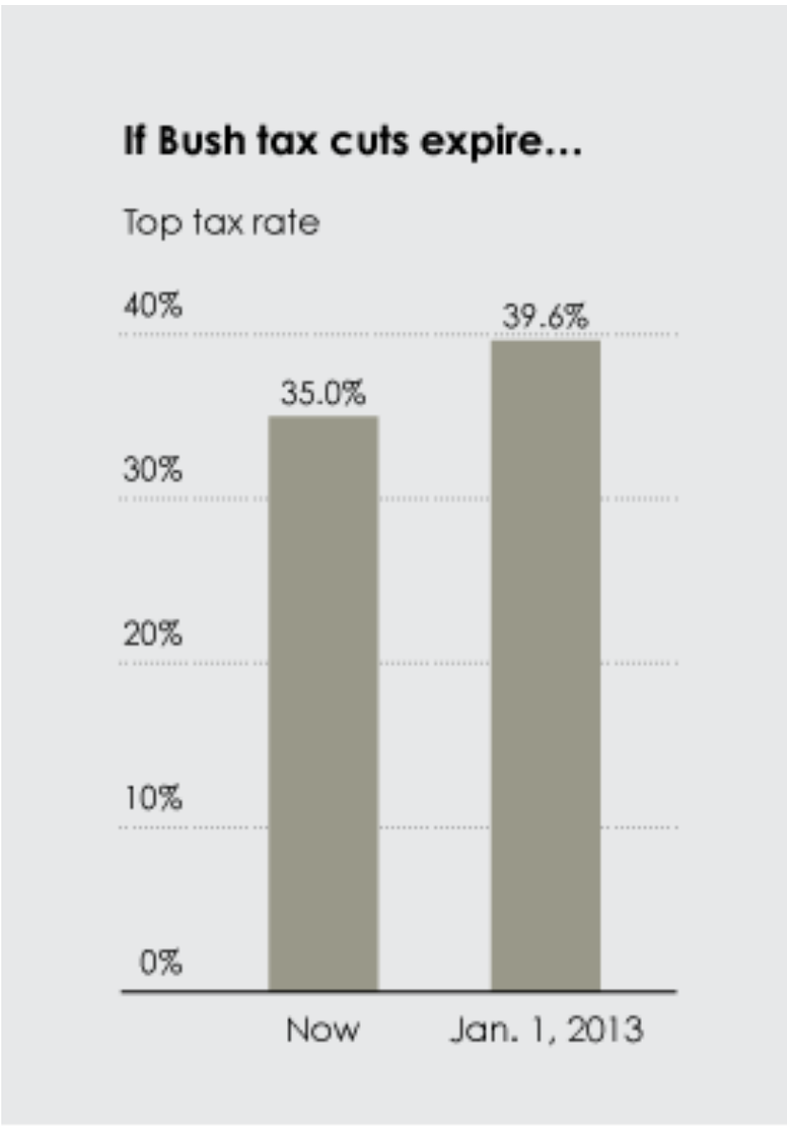
Graphical Integrity

Graphical Integrity



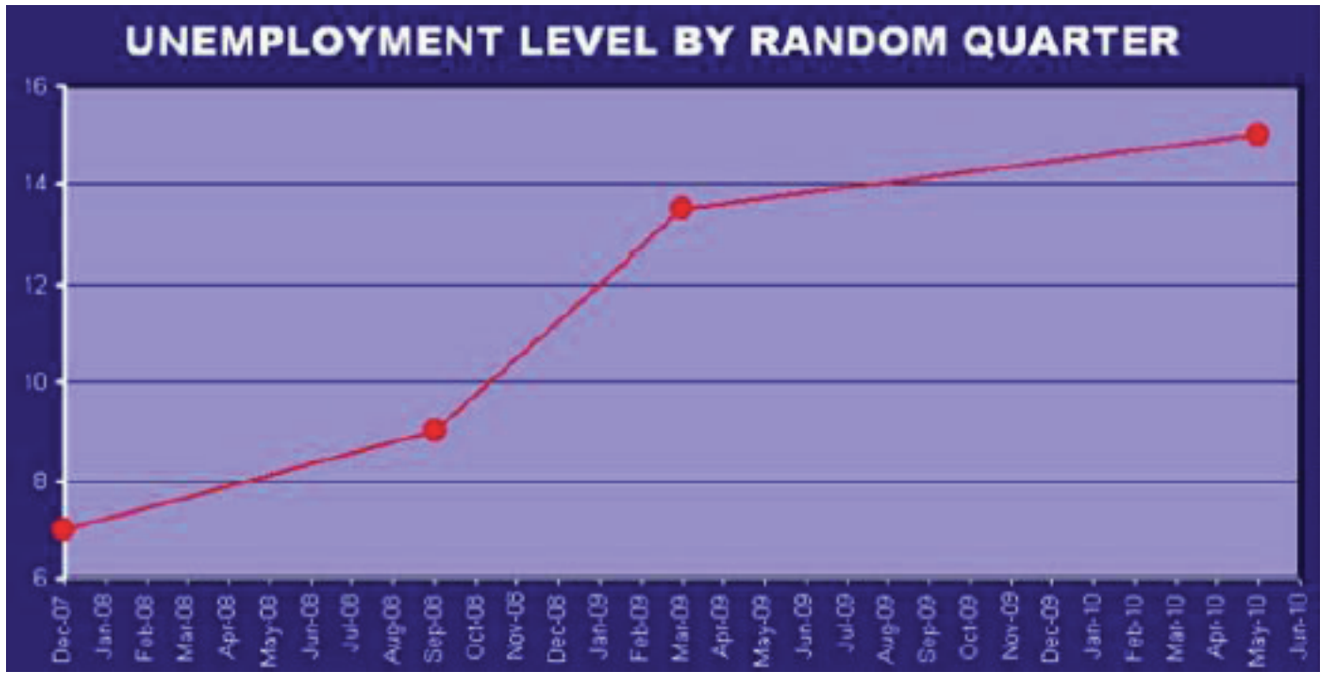
Scale Distortions



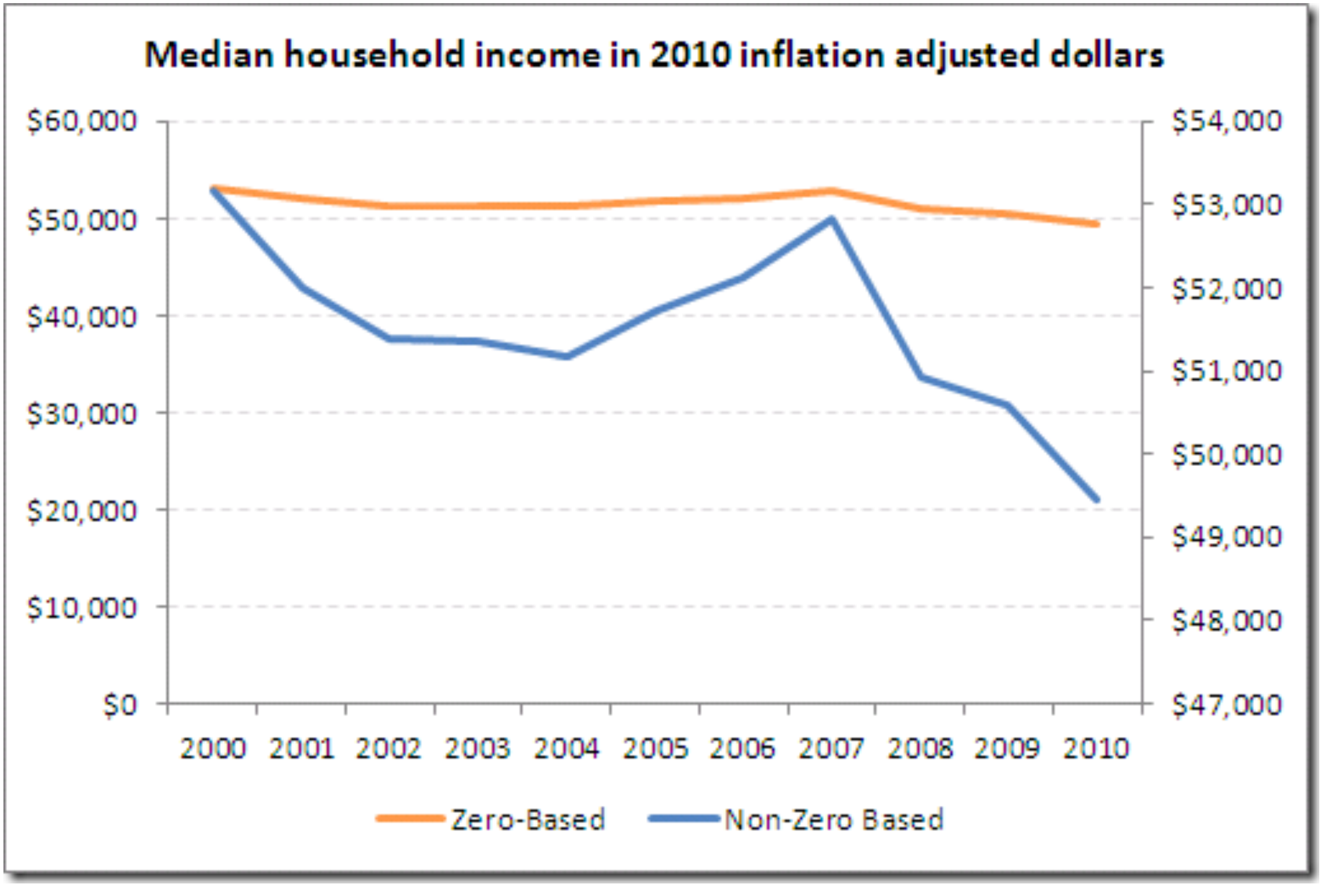



JOB LOSS BY QUARTER
15 MIL
13.5 MIL
9 MIL
7 MIL
DEC '07
SEPT '08
MARCH '09
JUNE '10
FOX NEWS .COM
SOURCE: BLS
AMERICA'S NEWSROOM

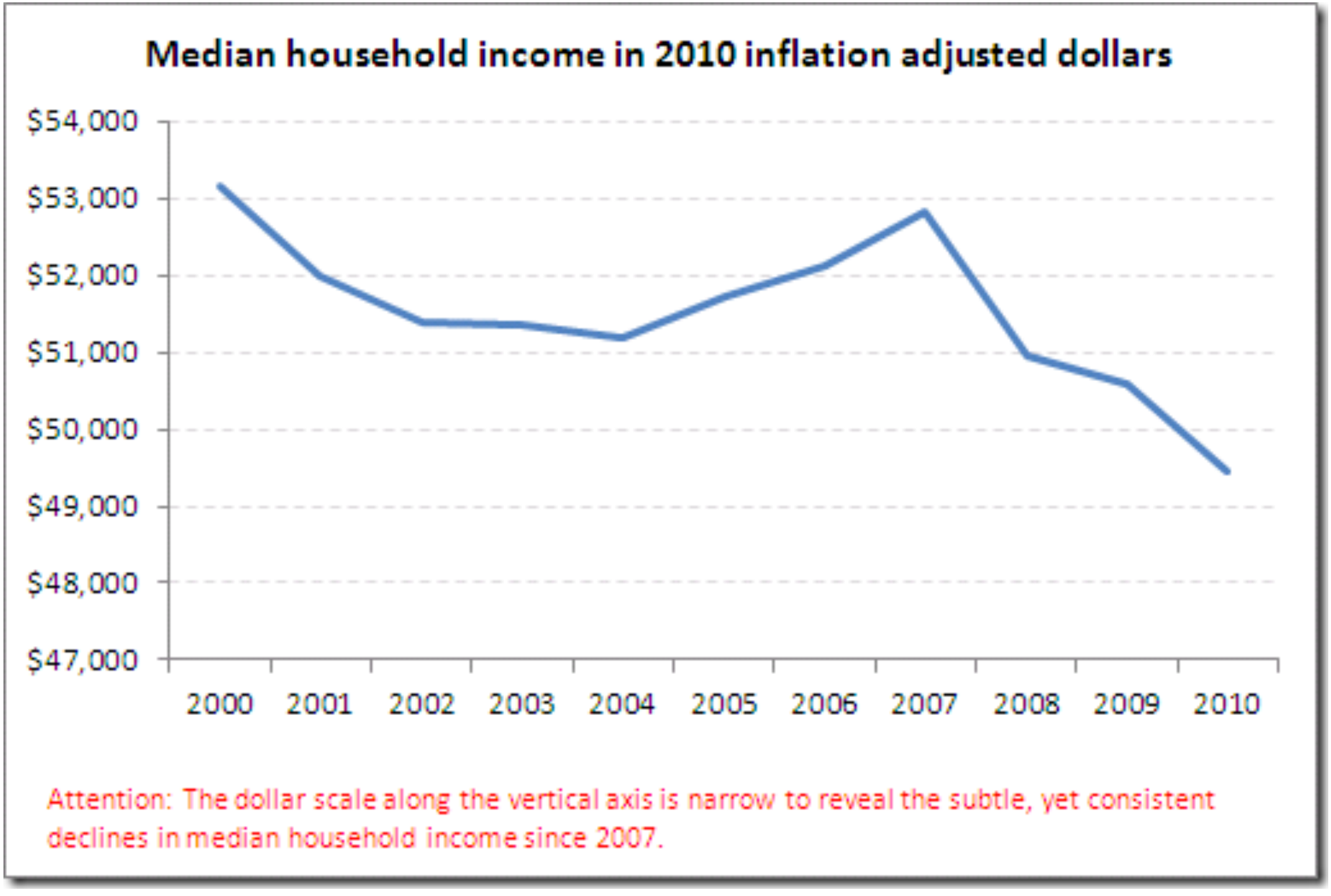
Scale Distortions



Scale Distortions



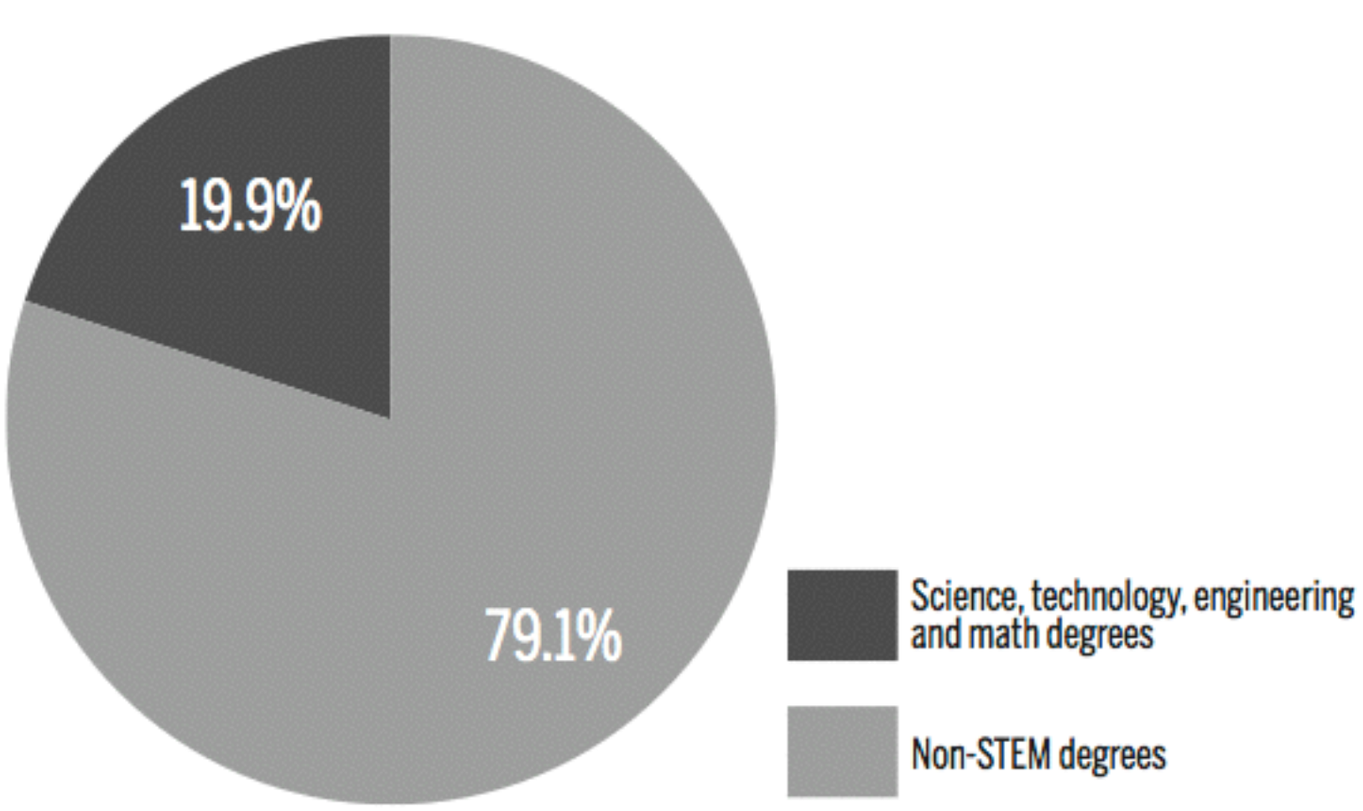
A. Kriebel,VizWiz


Median household income in 2010 inflation adjusted dollars
$54,000
$53,000
$52,000
$51,000
$50,000
$49,000
$48,000
$47,000
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Attention: The dollar scale along the vertical axis is narrow to reveal the subtle, yet consistent declines in median household income since 2007.

Yale Daily News

THE OLDEST COLLEGE DAILY
ESTABLISHED 1878

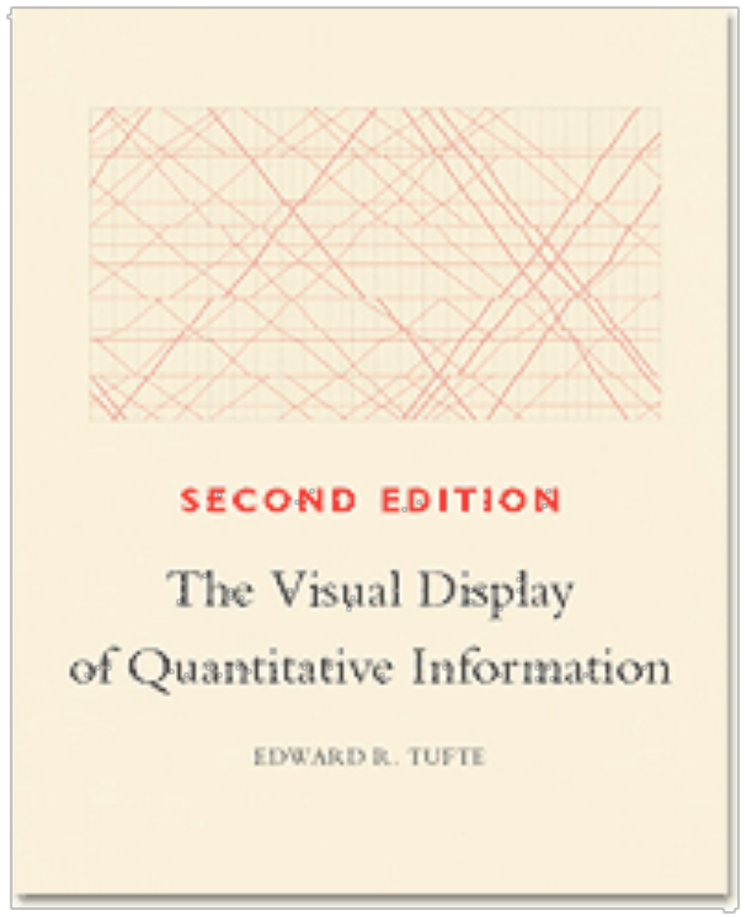
CHART YALE GRADUATES' MAJORS, CLASS OF 2011

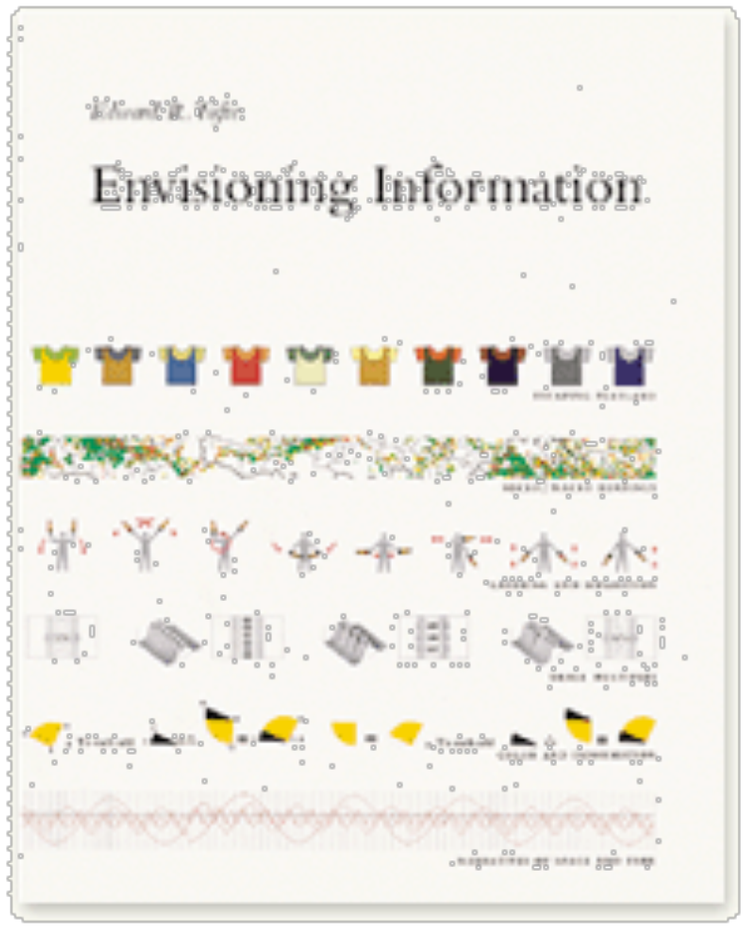


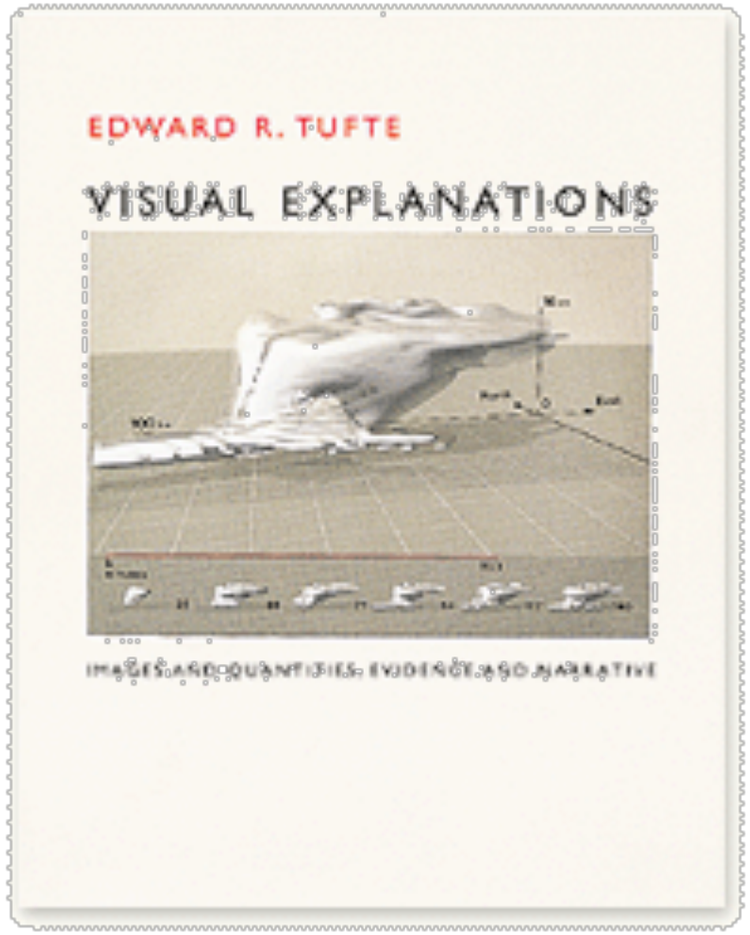
September 13, 2011

Keep It Simple

Edward Tufte

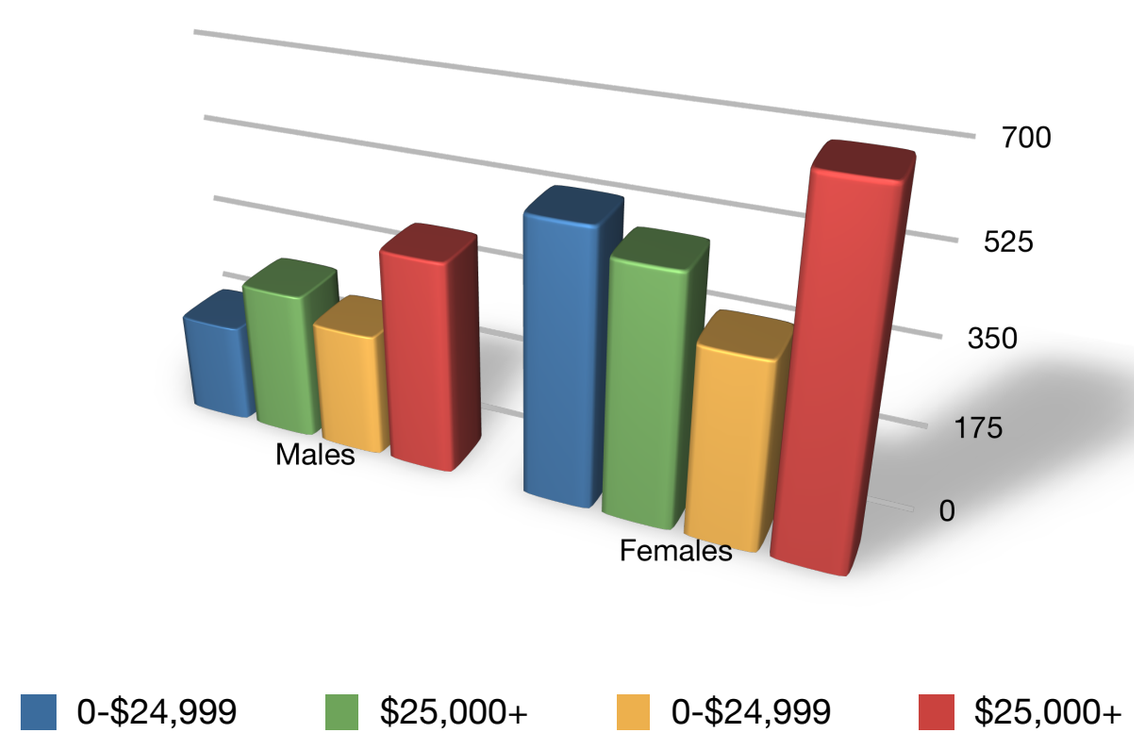






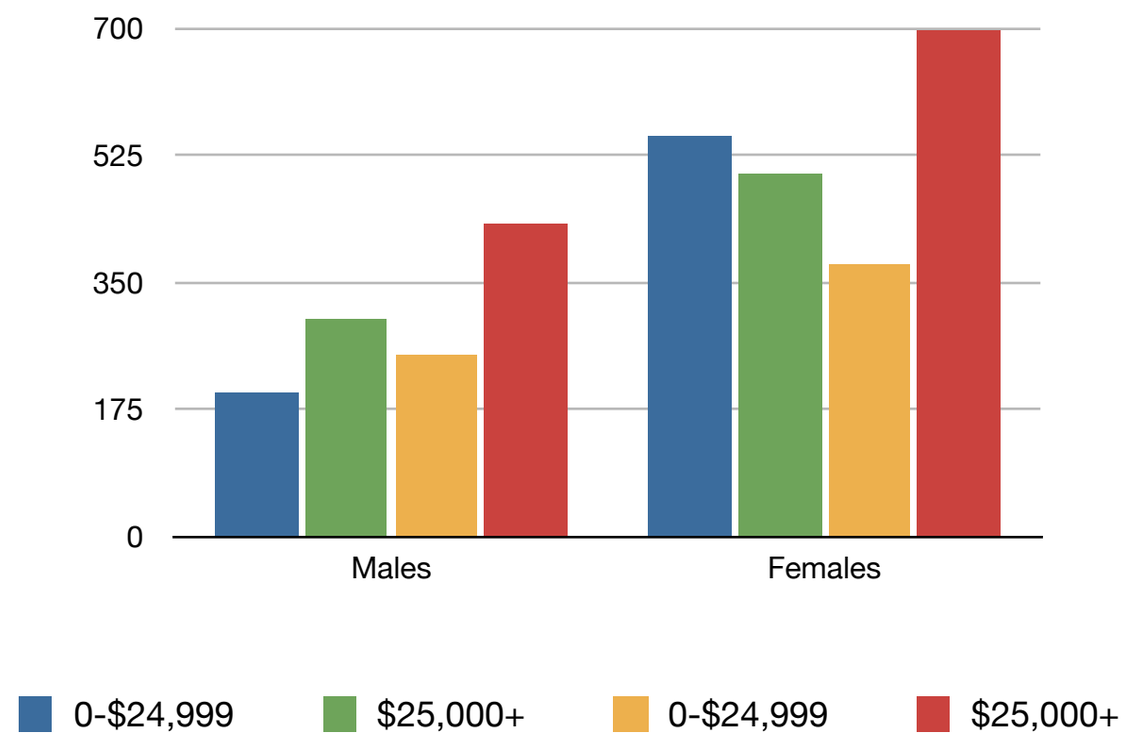
Maximize Data-Ink Ratio

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

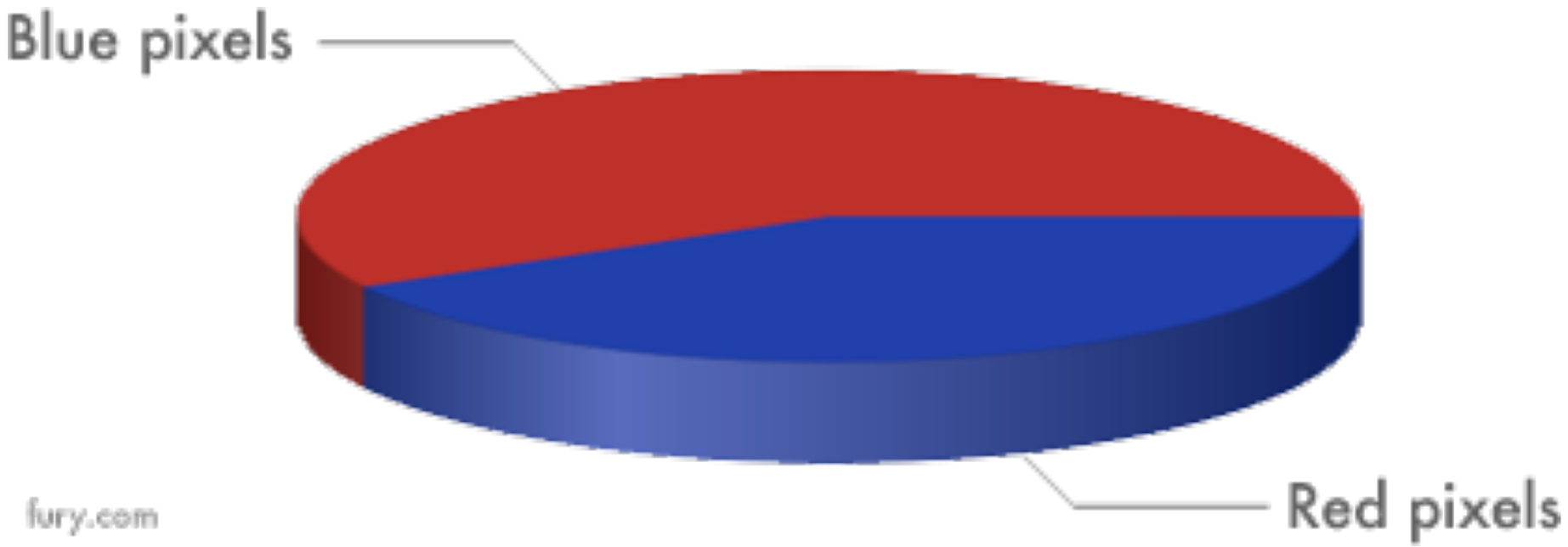


Maximize Data-Ink Ratio

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



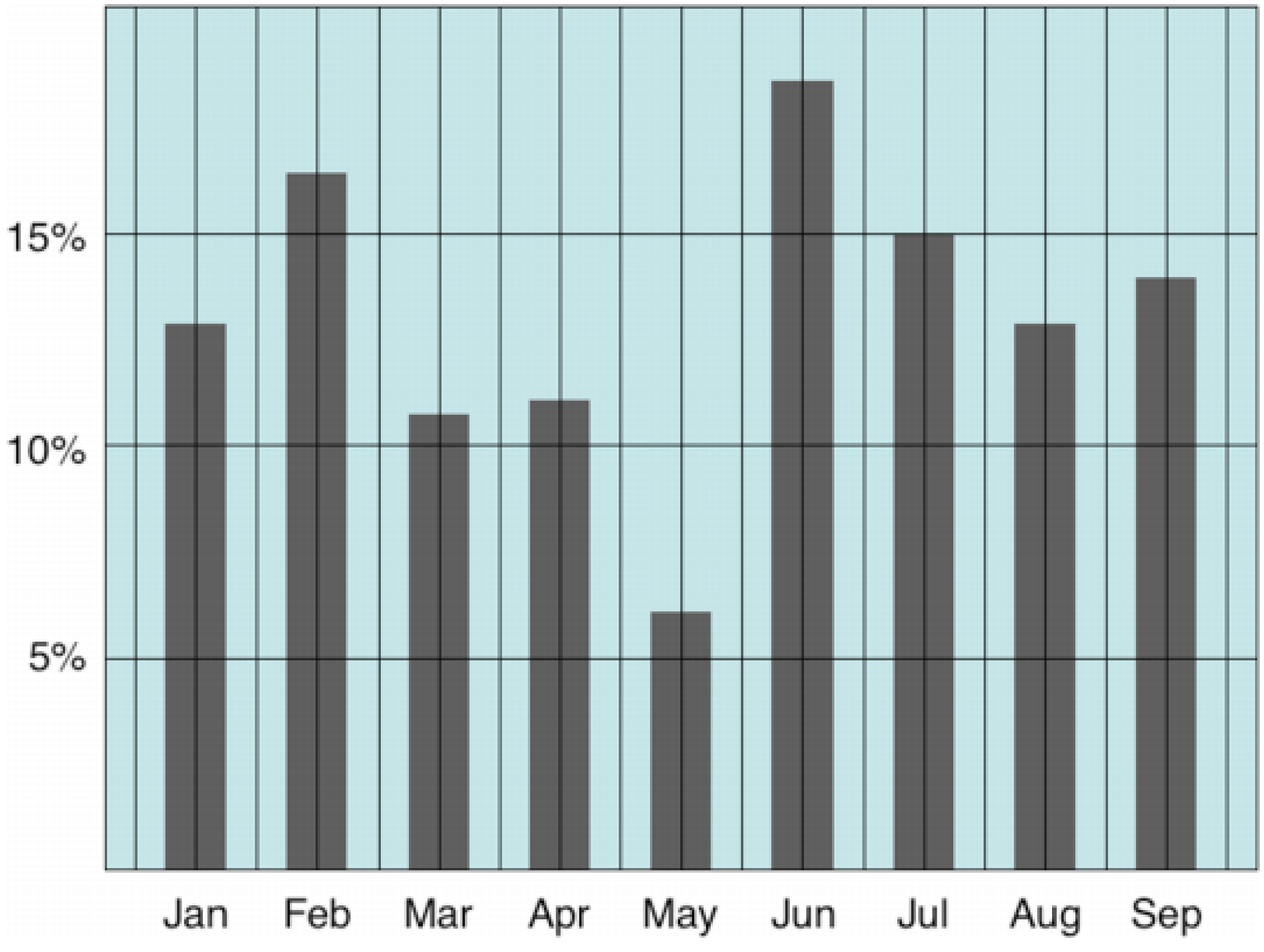
Why 3D pie charts are bad



Kevin Fox

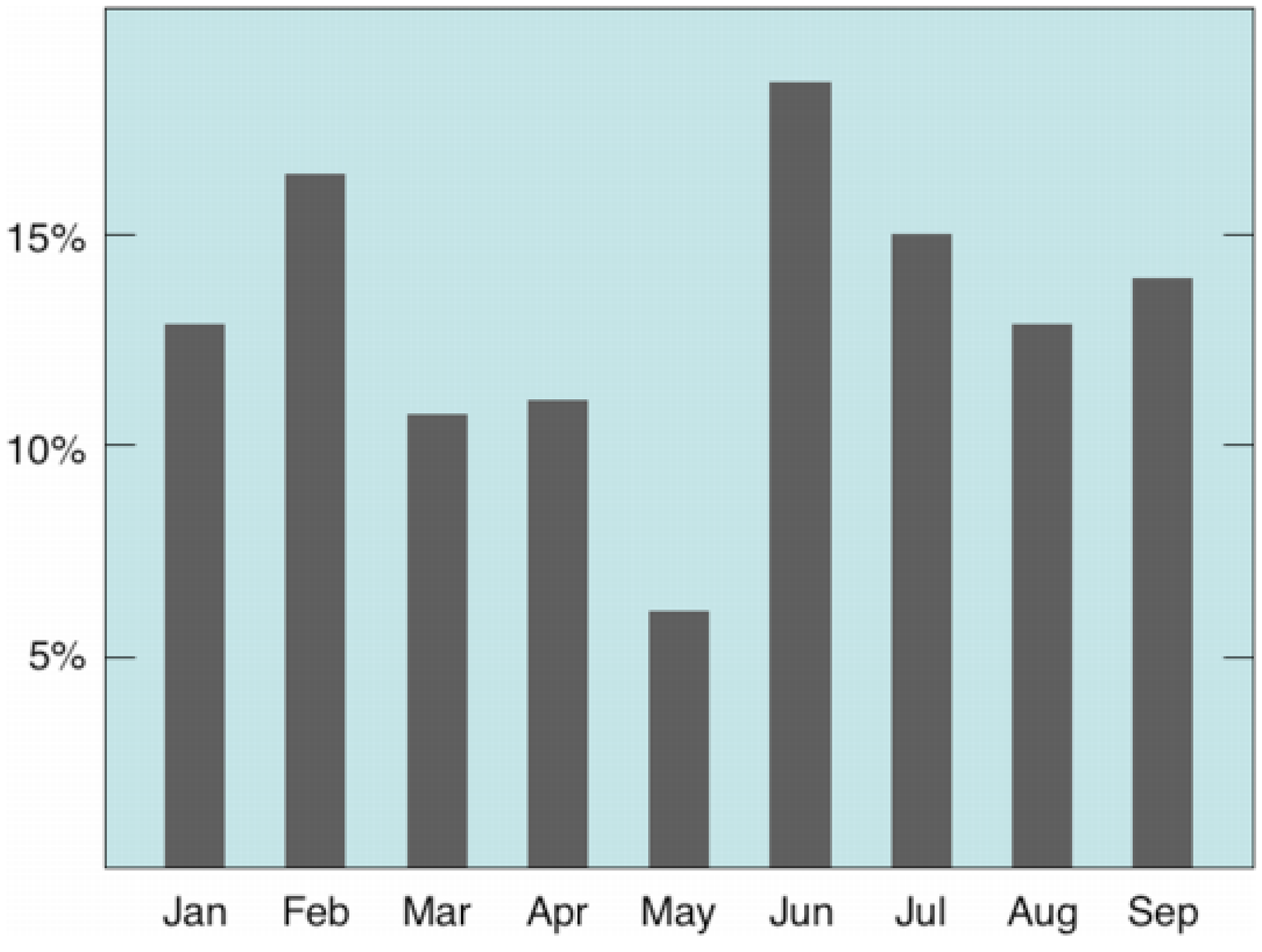
Avoid Chartjunk

Extraneous visual elements that distract from the message



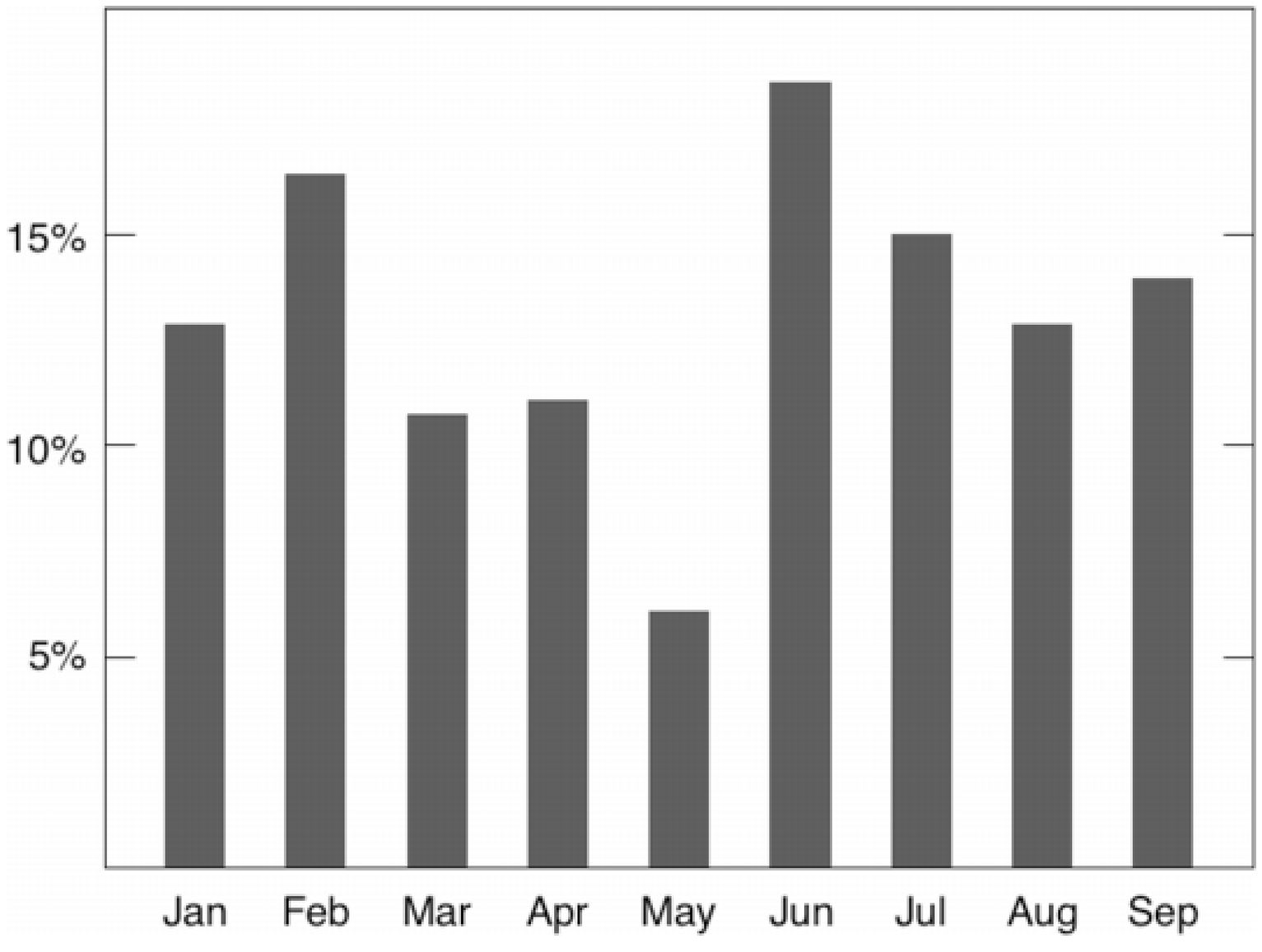
ongoing, Tim Brey

Avoid Chartjunk



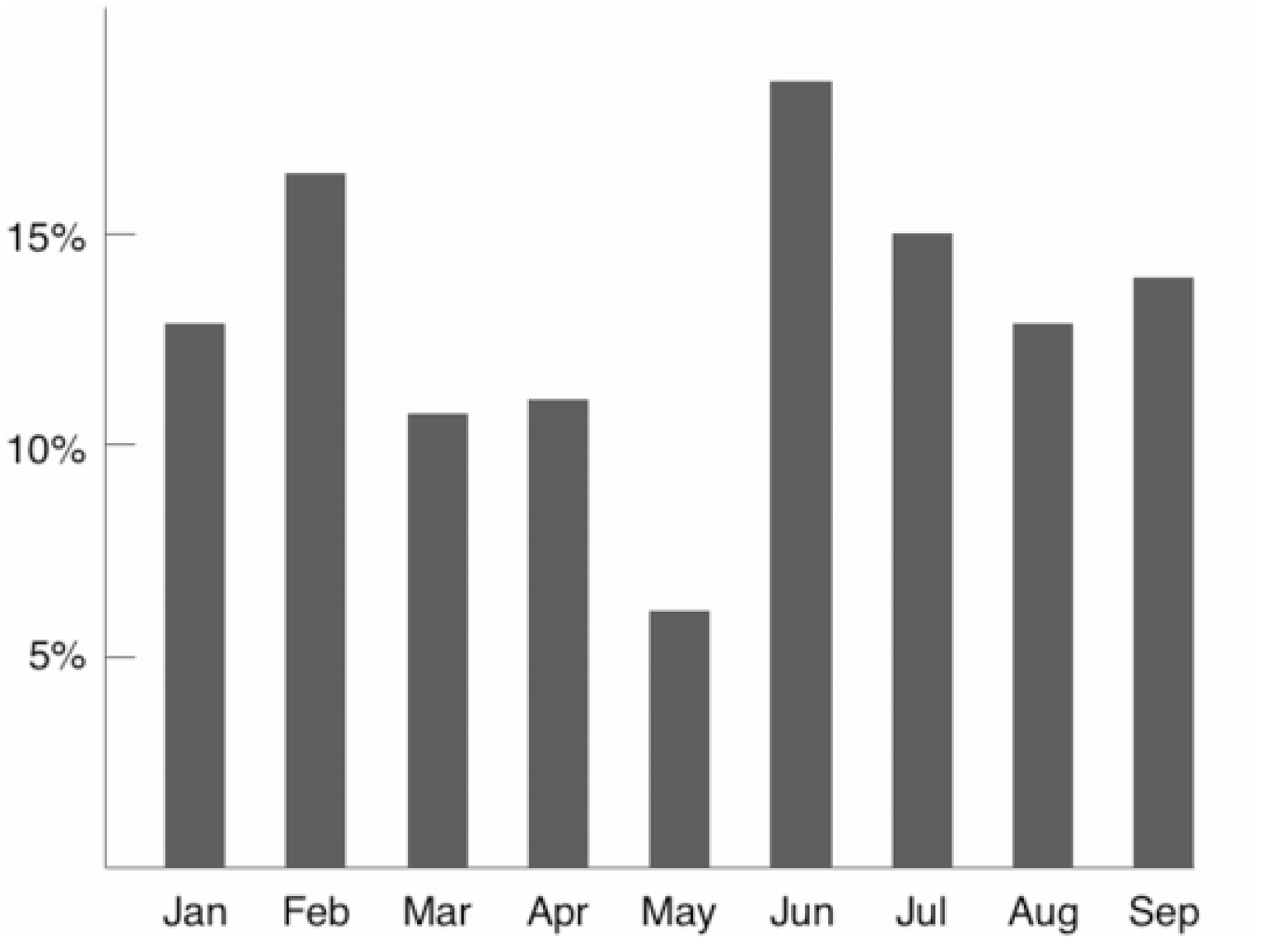
ongoing, Tim Brey

Avoid Chartjunk



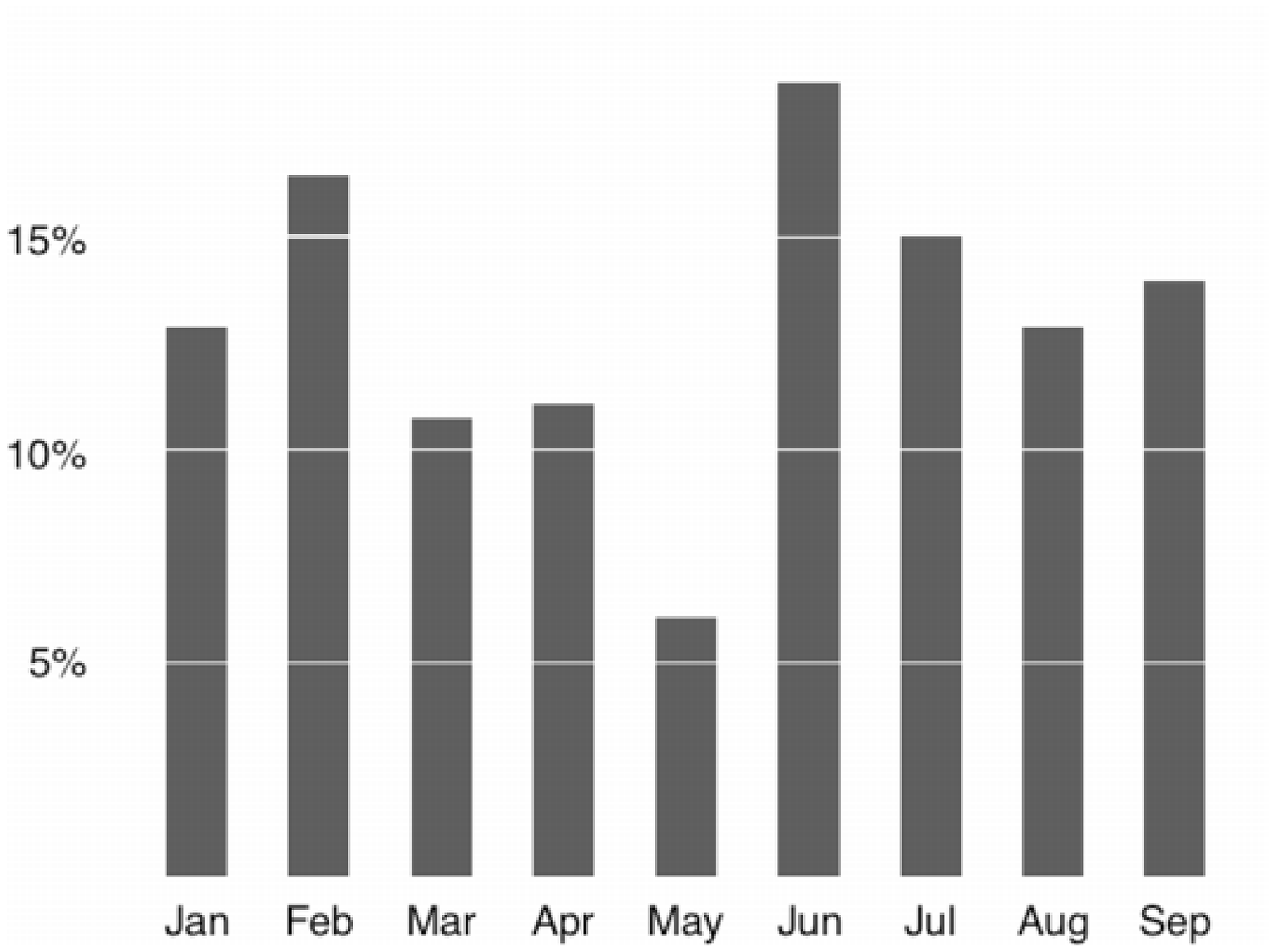
ongoing, Tim Brey

Avoid Chartjunk



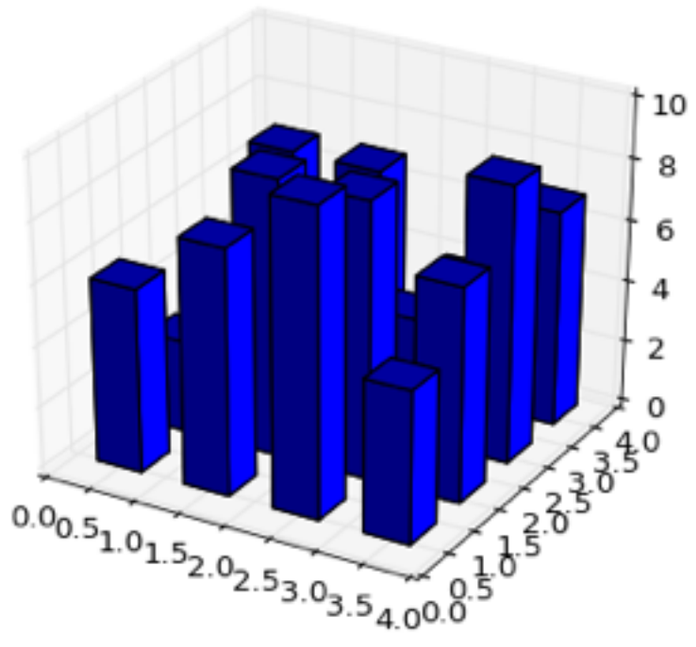
ongoing, Tim Brey

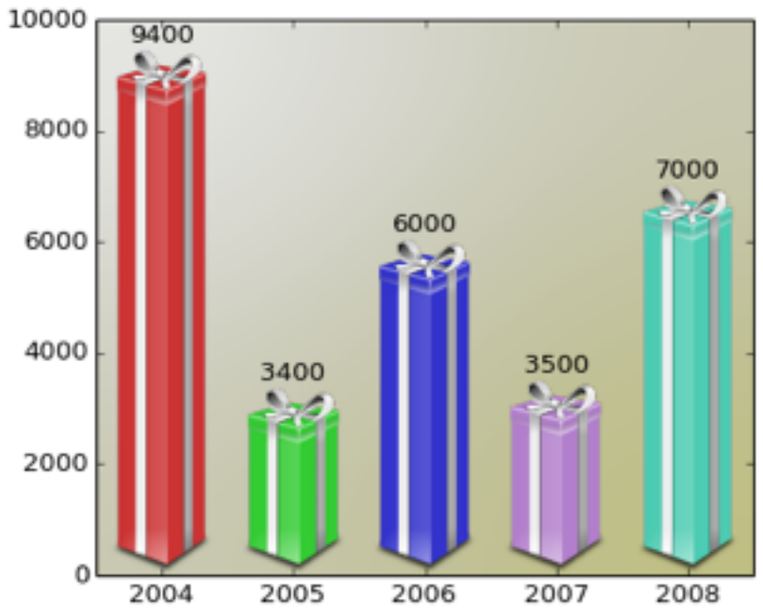
Avoid Chartjunk



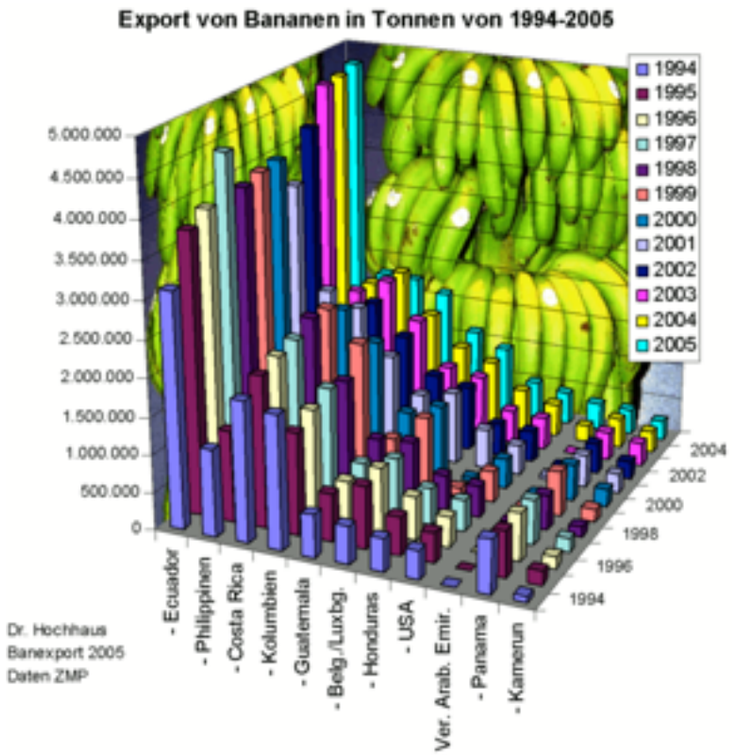
ongoing, Tim Brey

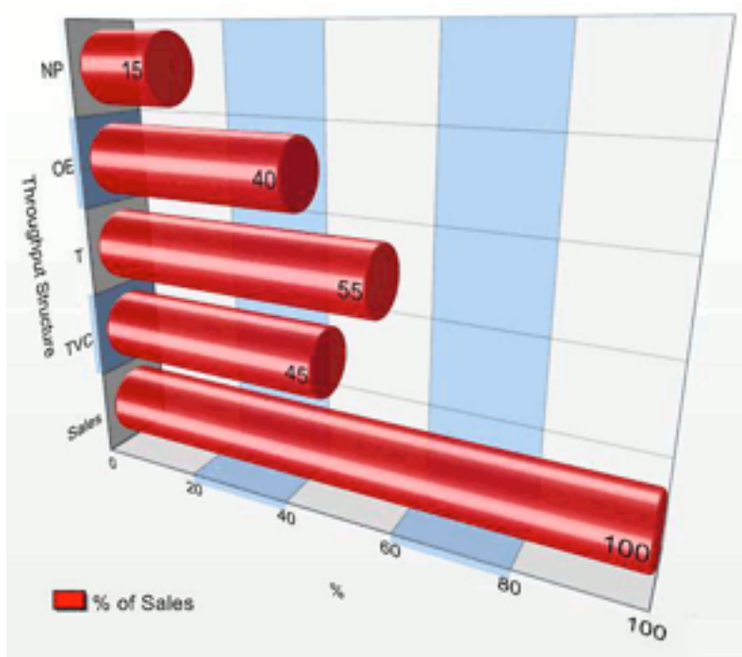
Don’t!





matplotlib gallery

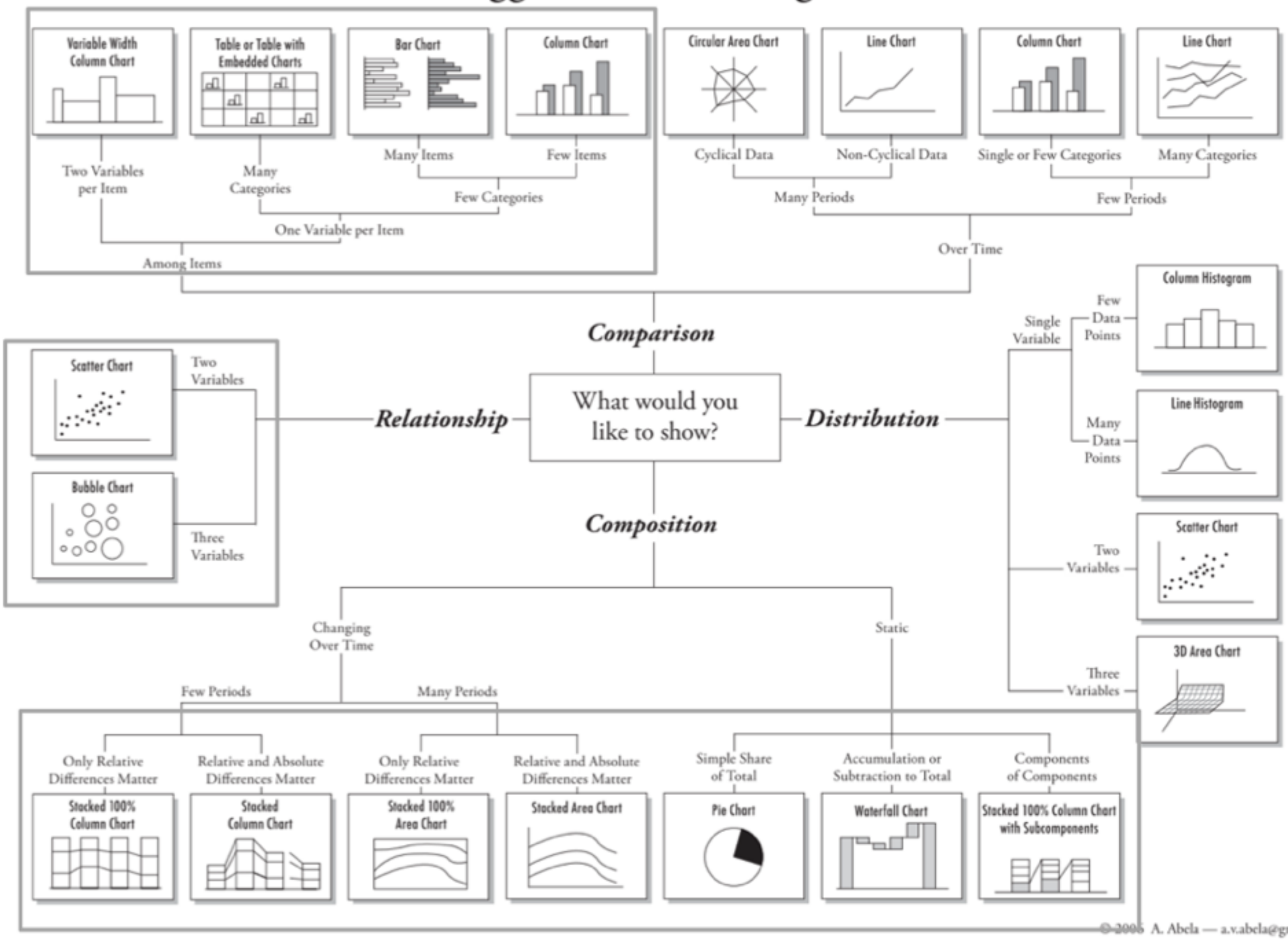




Excel Charts Blog

Use The Right Display

Chart Suggestions—A Thought-Starter



http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

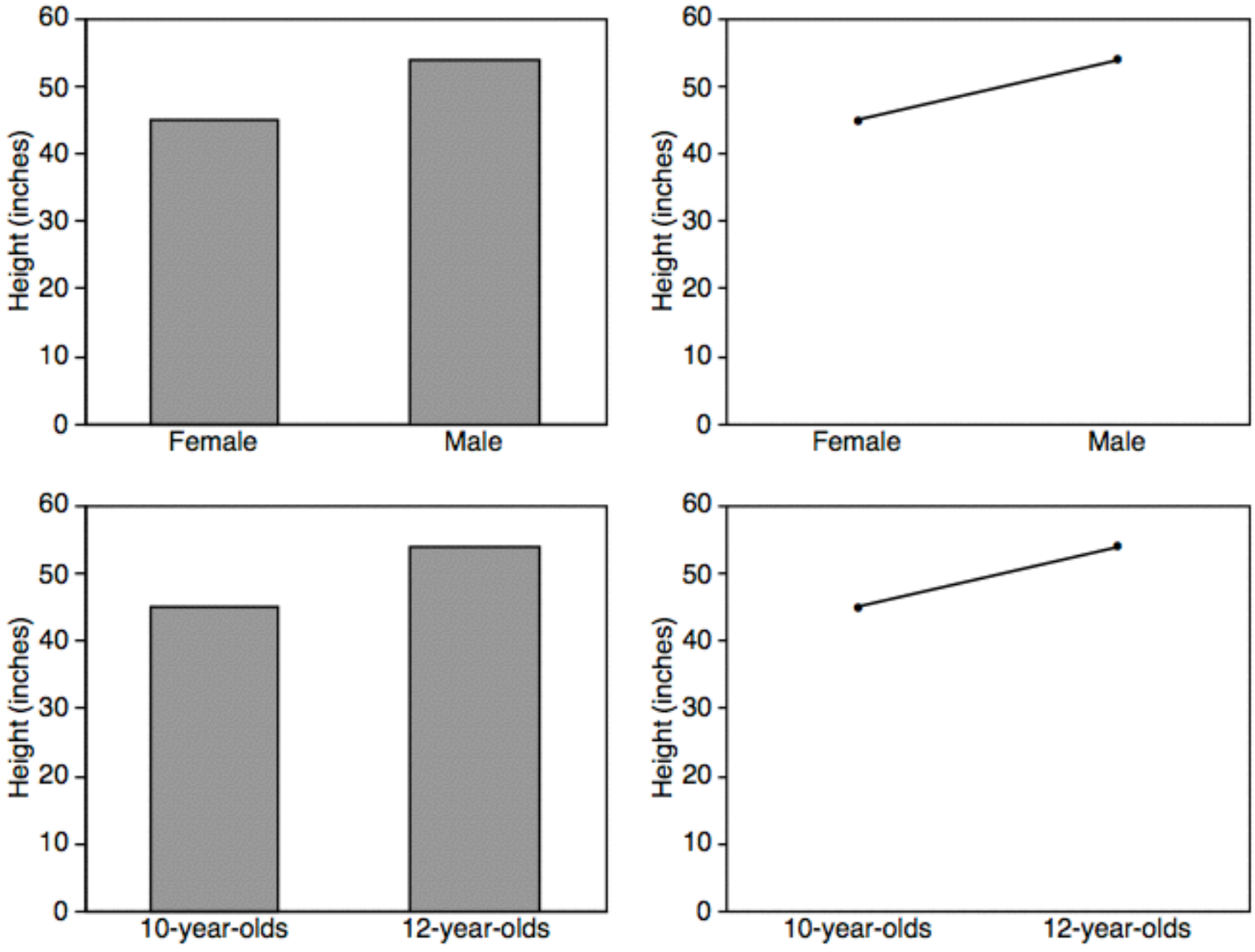
Comparisons

Bar Chart

How Much Does Beer Consumption Vary by Country?

Bottles per person per week

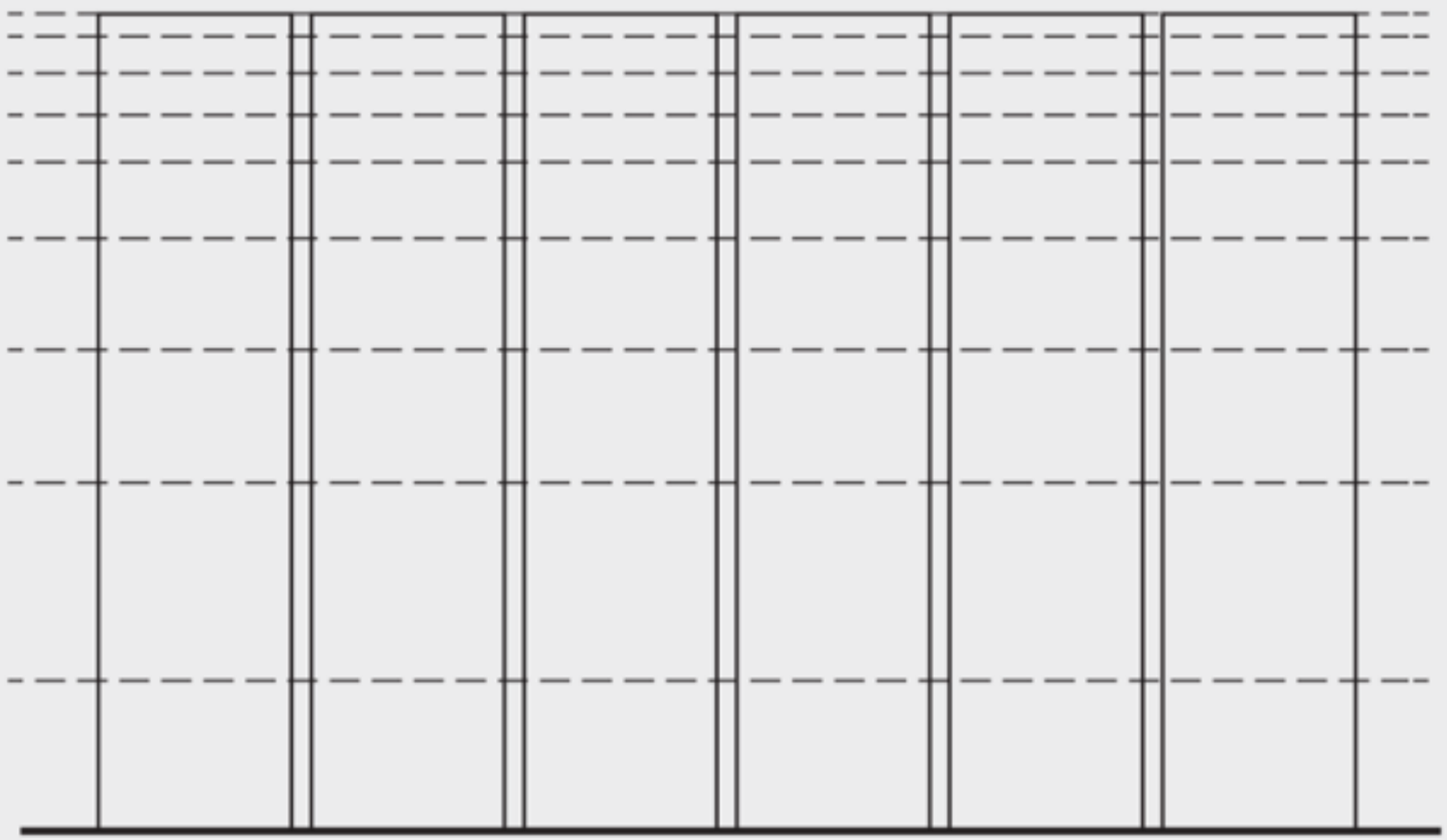
Bars vs. Lines



Zacks 1999

Bar Chart Baselines Start at Zero

BY NATHAN YAU / POSTED TO GUIDES / TAGS: BAR CHART, RULES



There are visualization rules and there are visualization

Nathan Yau

Trends

Apple Inc. (AAPL) - NasdaqGS

Add to Portfolio | Like 6k

601.10 ↑15.53(2.65%) 4:00PM EDT | After Hours: **604.60** ↑3.50 (0.58%) 7:15PM EDT - Nasdaq Real Time Price

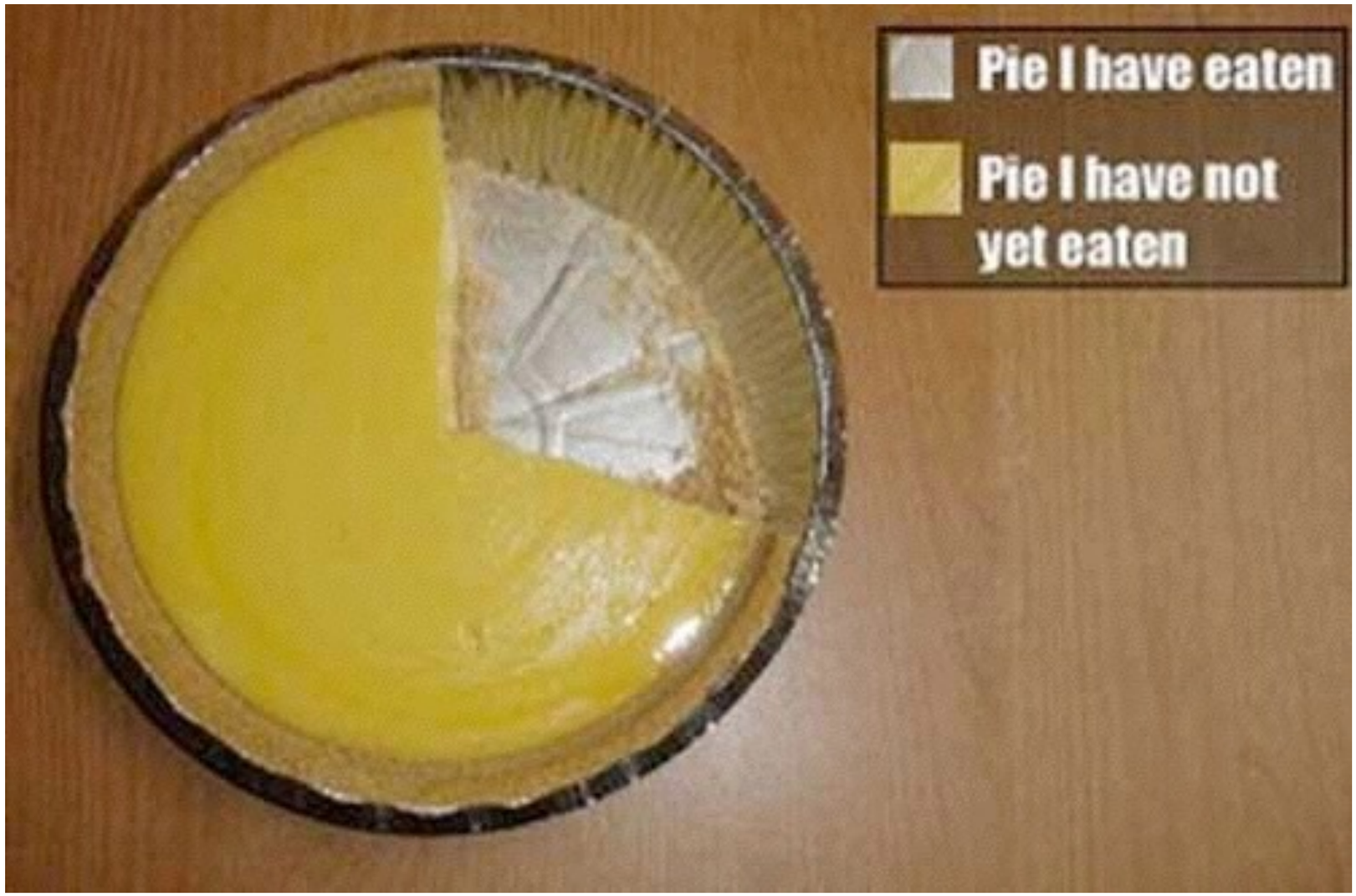
Enter name(s) or symbol(s) | GET CHART | COMPARE | EVENTS | TECHNICAL INDICATORS | CHART SETTINGS | RESET

Basic Chart | Full Screen | Print | Share | Send Feedback

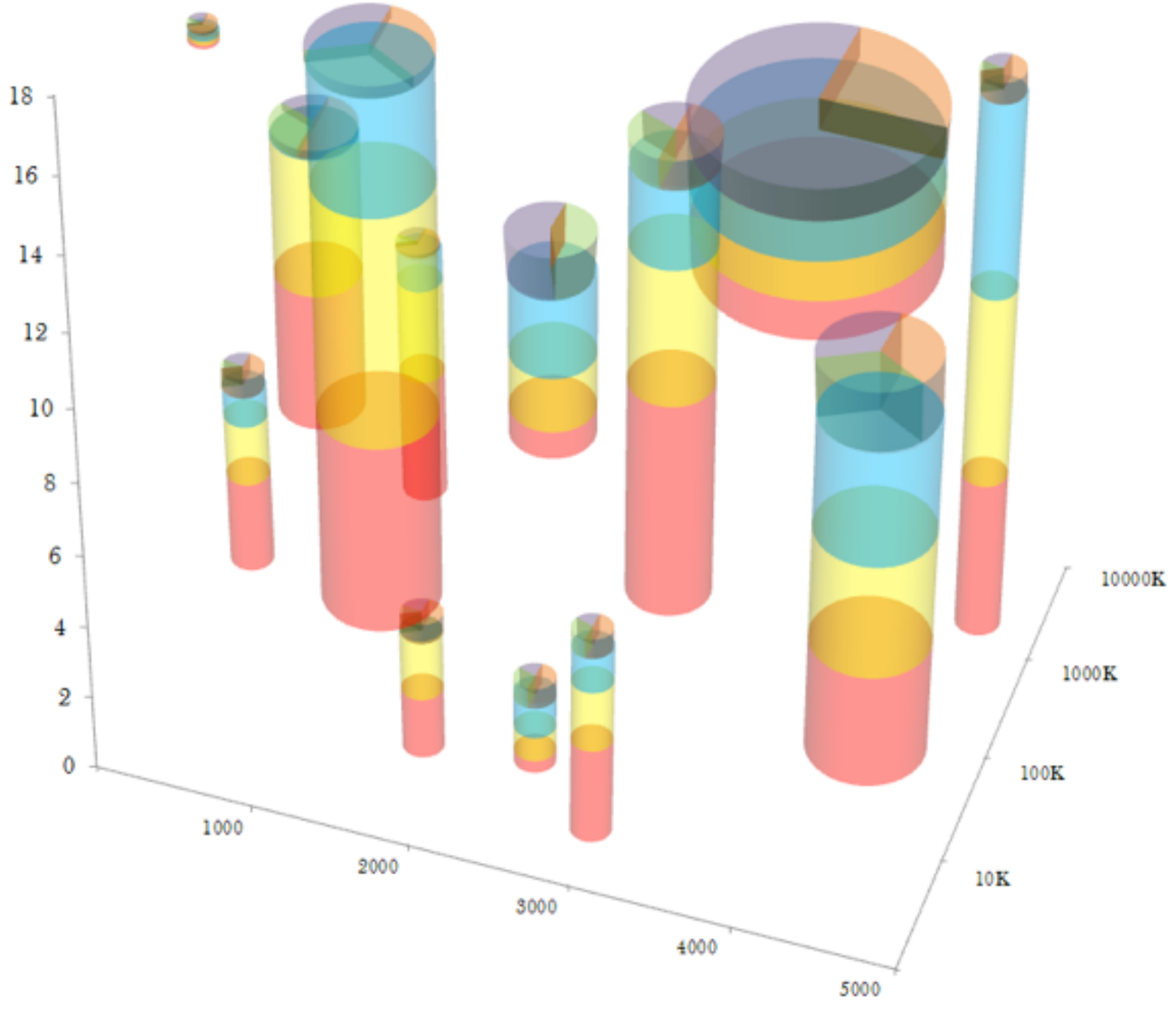
Yahoo! Finance

Proportions

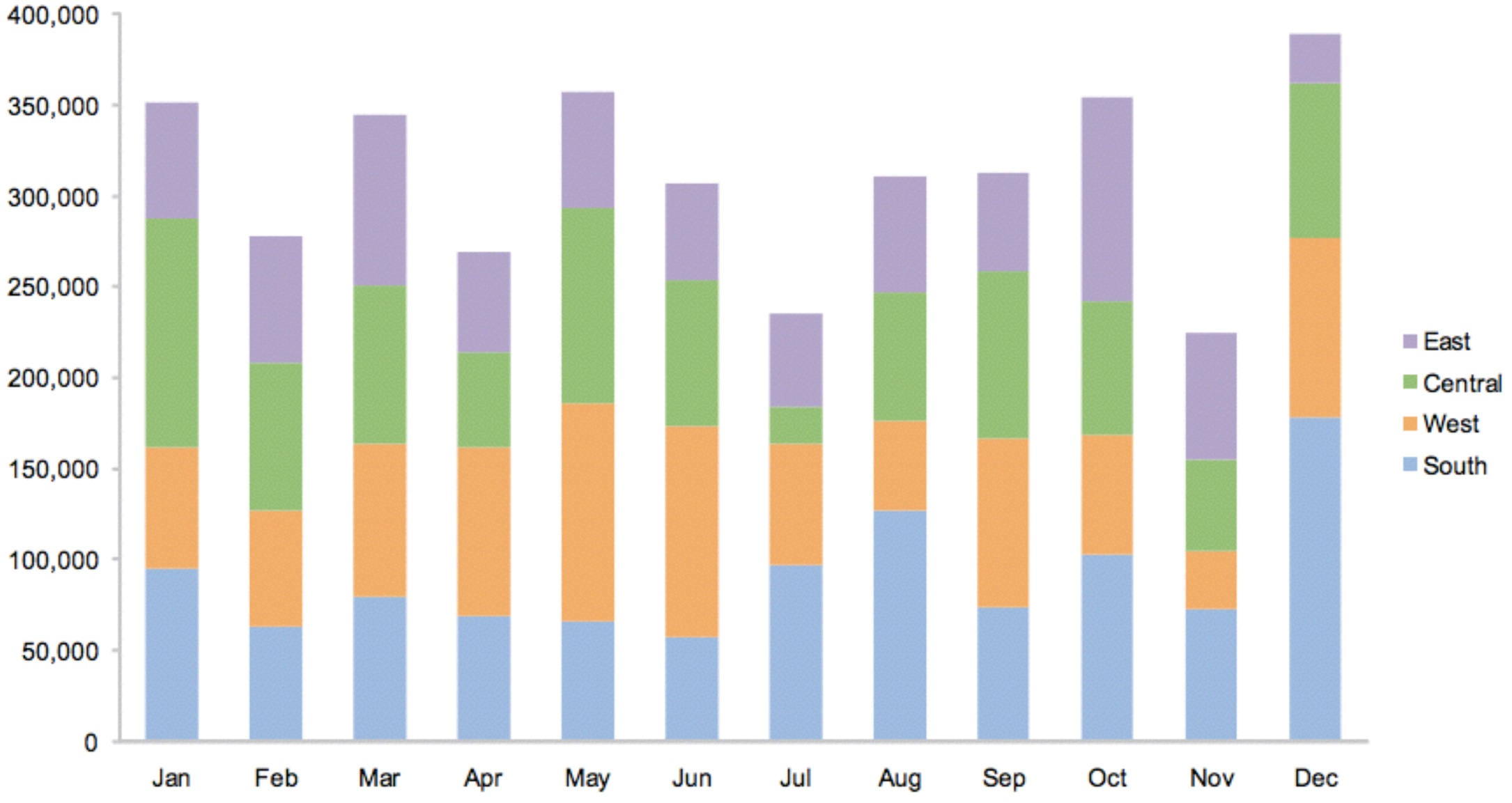
Pie Charts



eagerpies.com

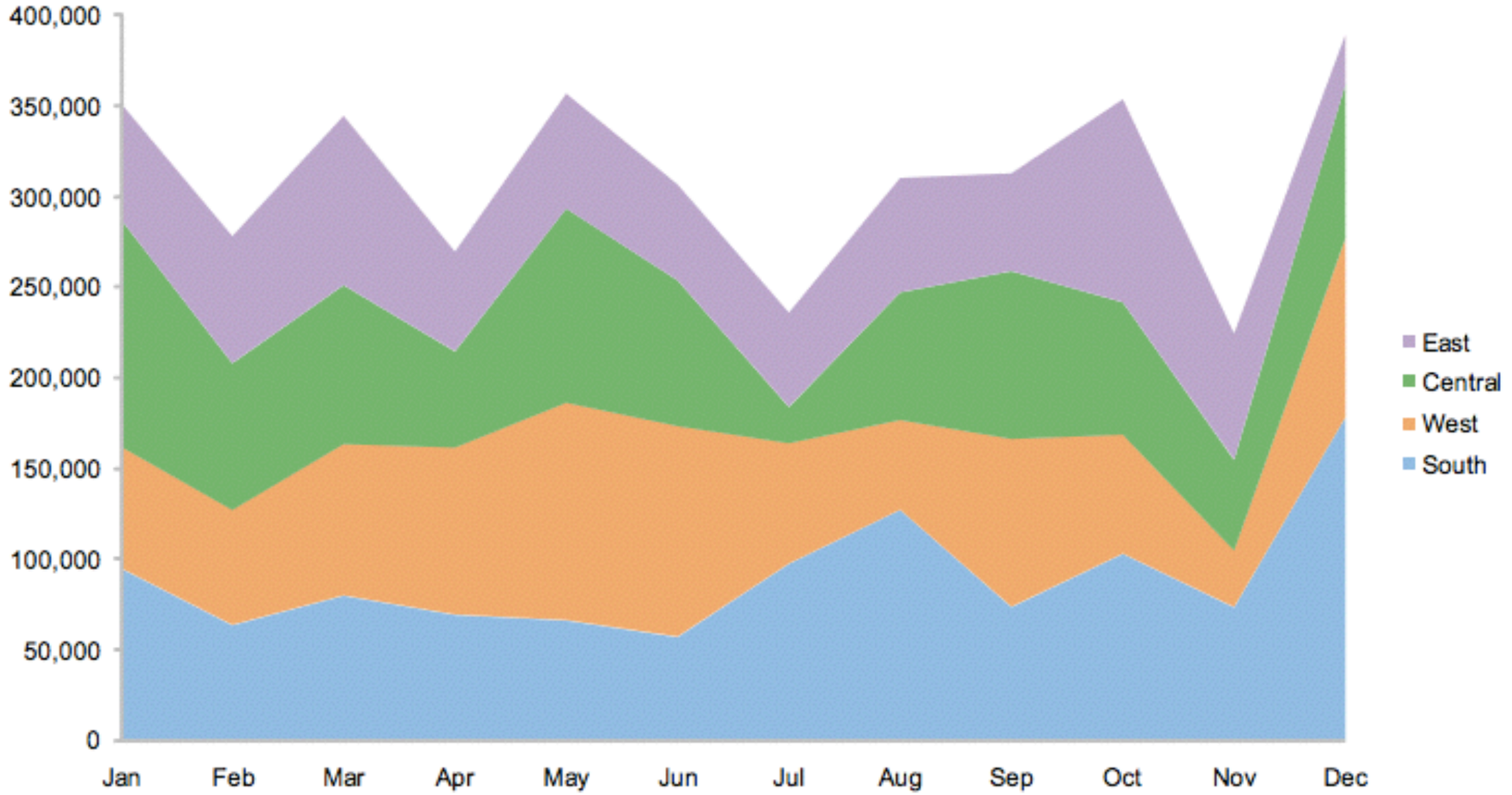


Stacked Bar Chart



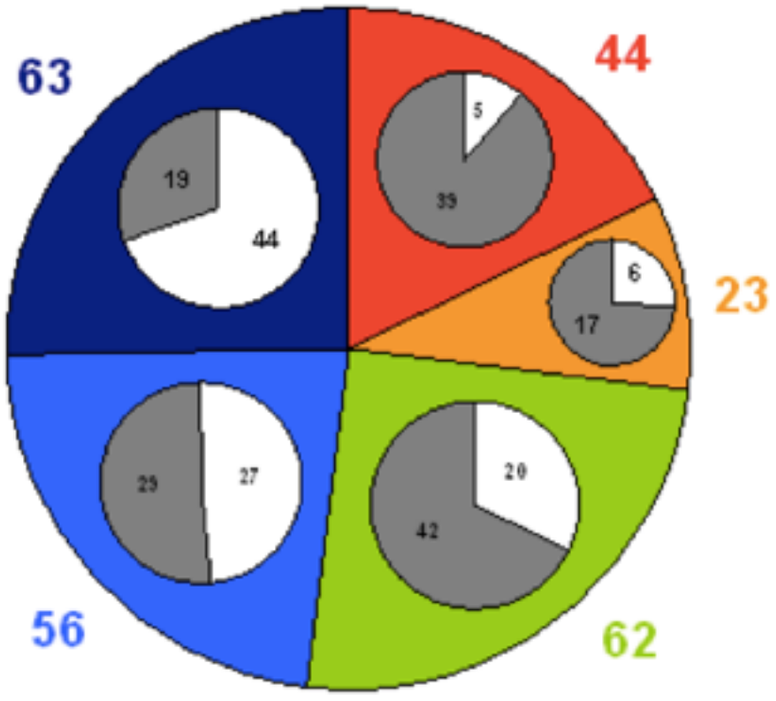
S. Few

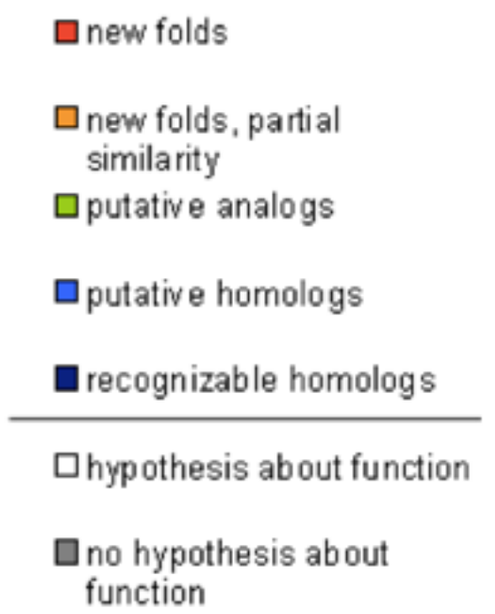
Stacked Area Chart

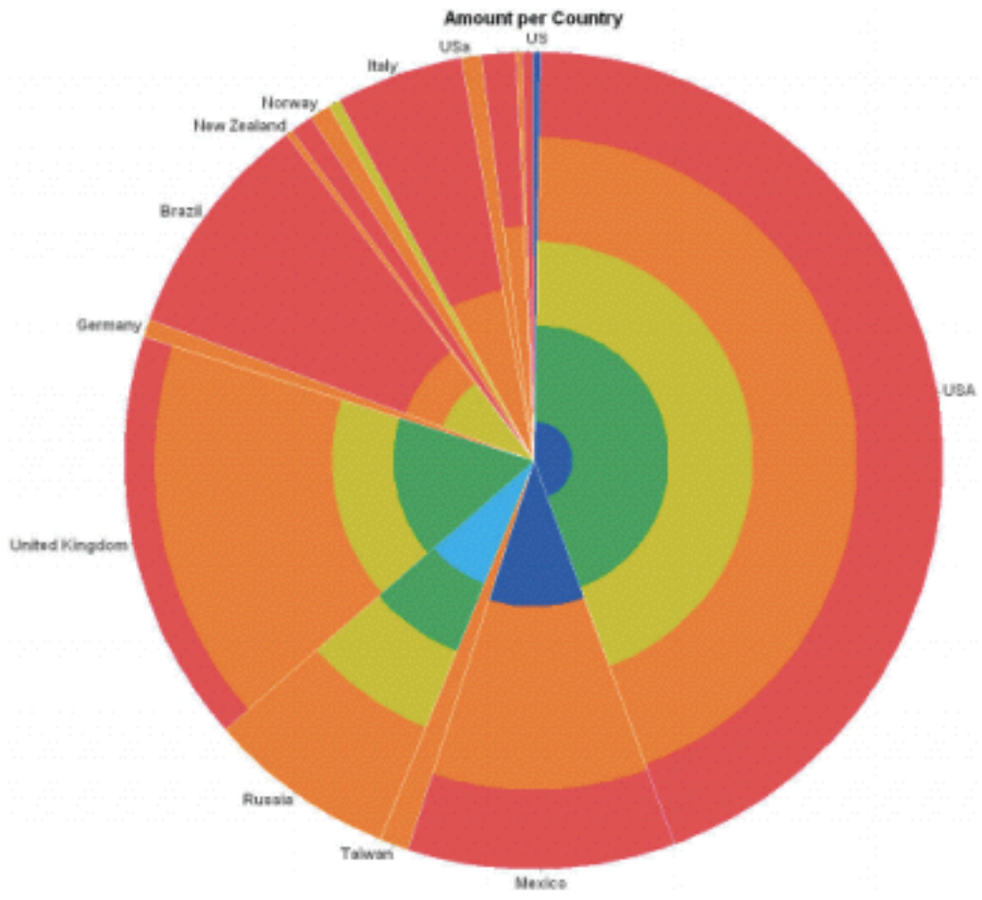


S. Few

Don’t!

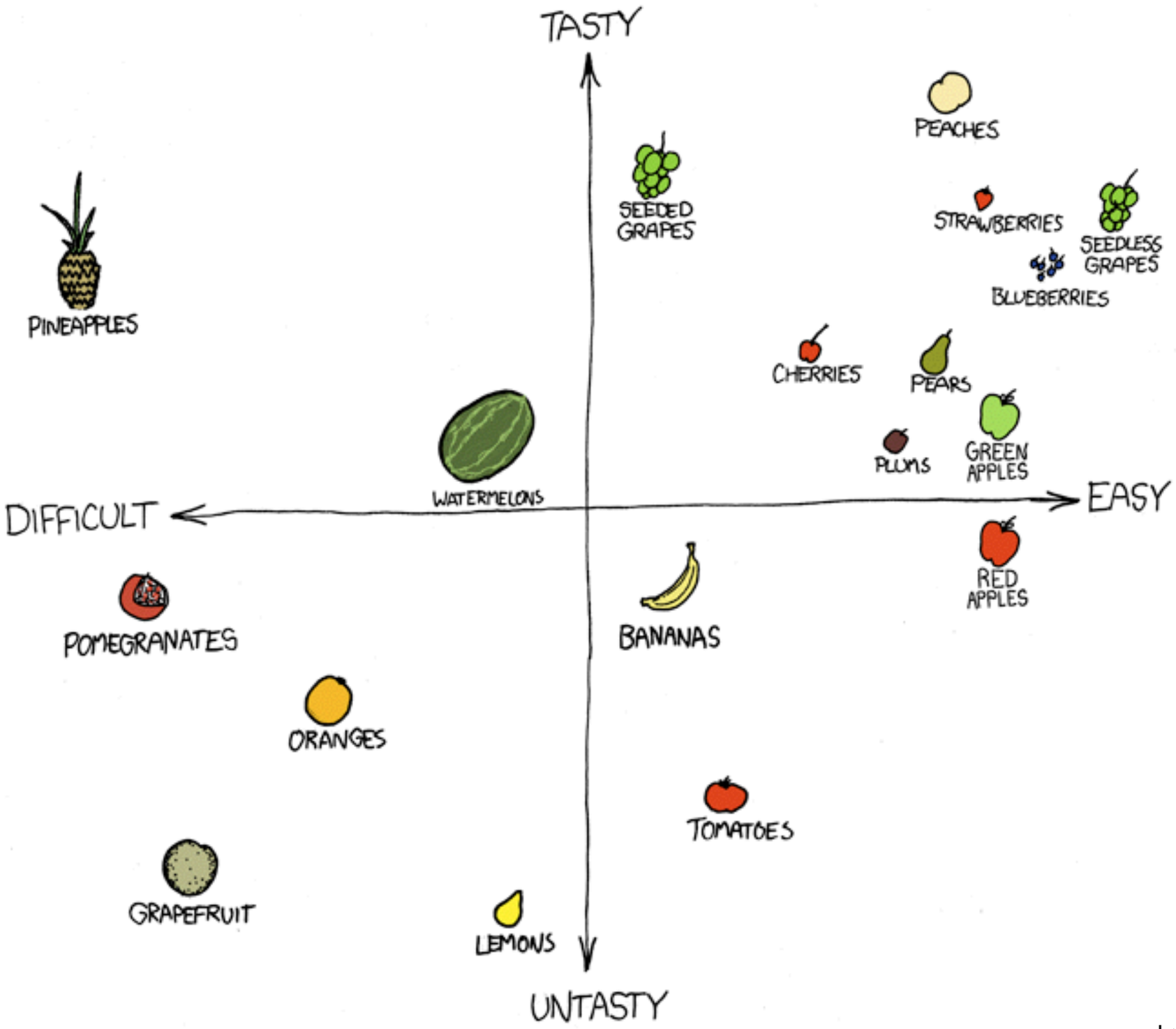

63
44
23
56
62
19
44
5
39
6
17
29
27
20
42


new folds
new folds, partial similarity
putative analogs
putative homologs
recognizable homologs
hypothesis about function
no hypothesis about function


Amount per Country
US
USa
Italy
Norway
New Zealand
Brazil
Germany
USA
United Kingdom
Russia
Taiwan
Mexico

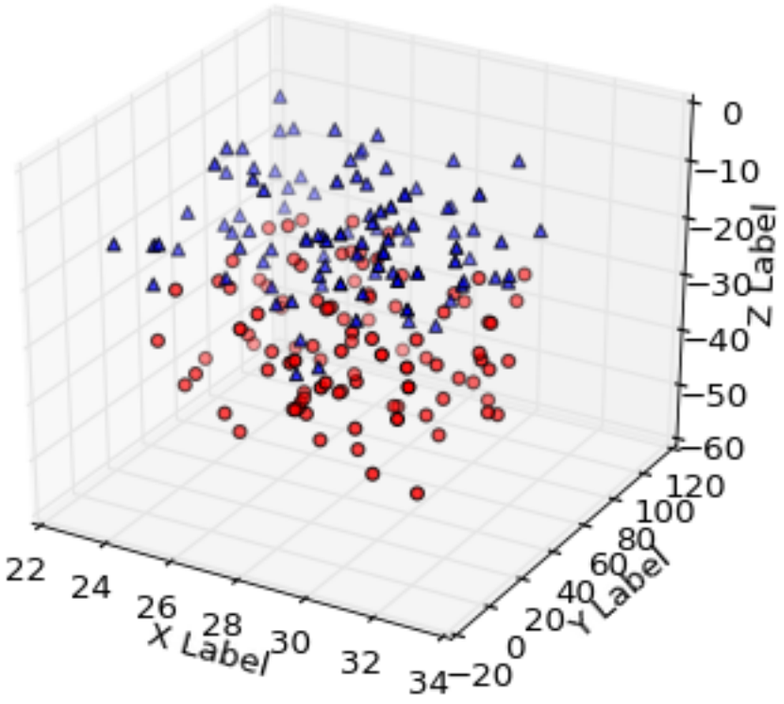
Correlations

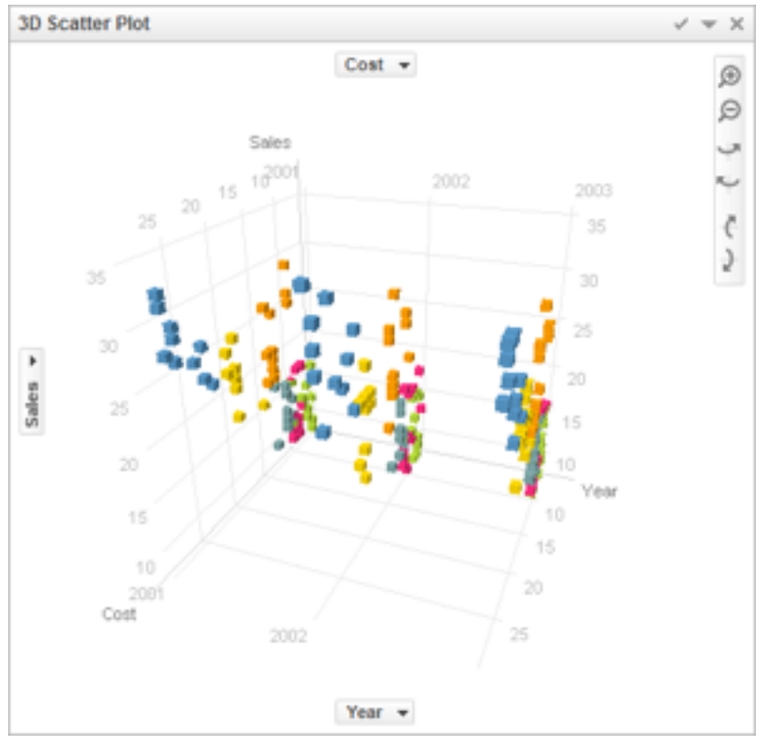
Scatterplots

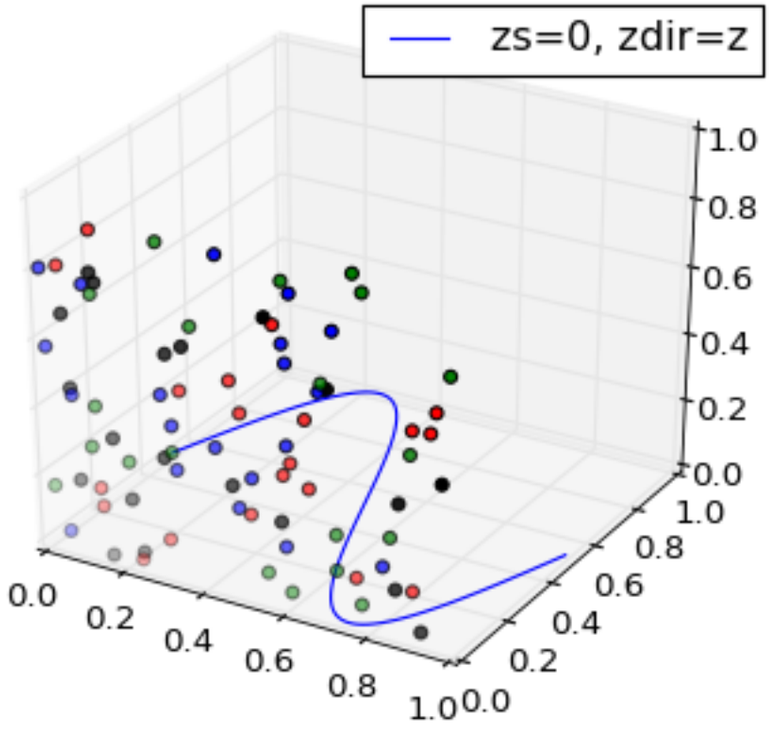


http://xkcd.com/388/

Don’t!

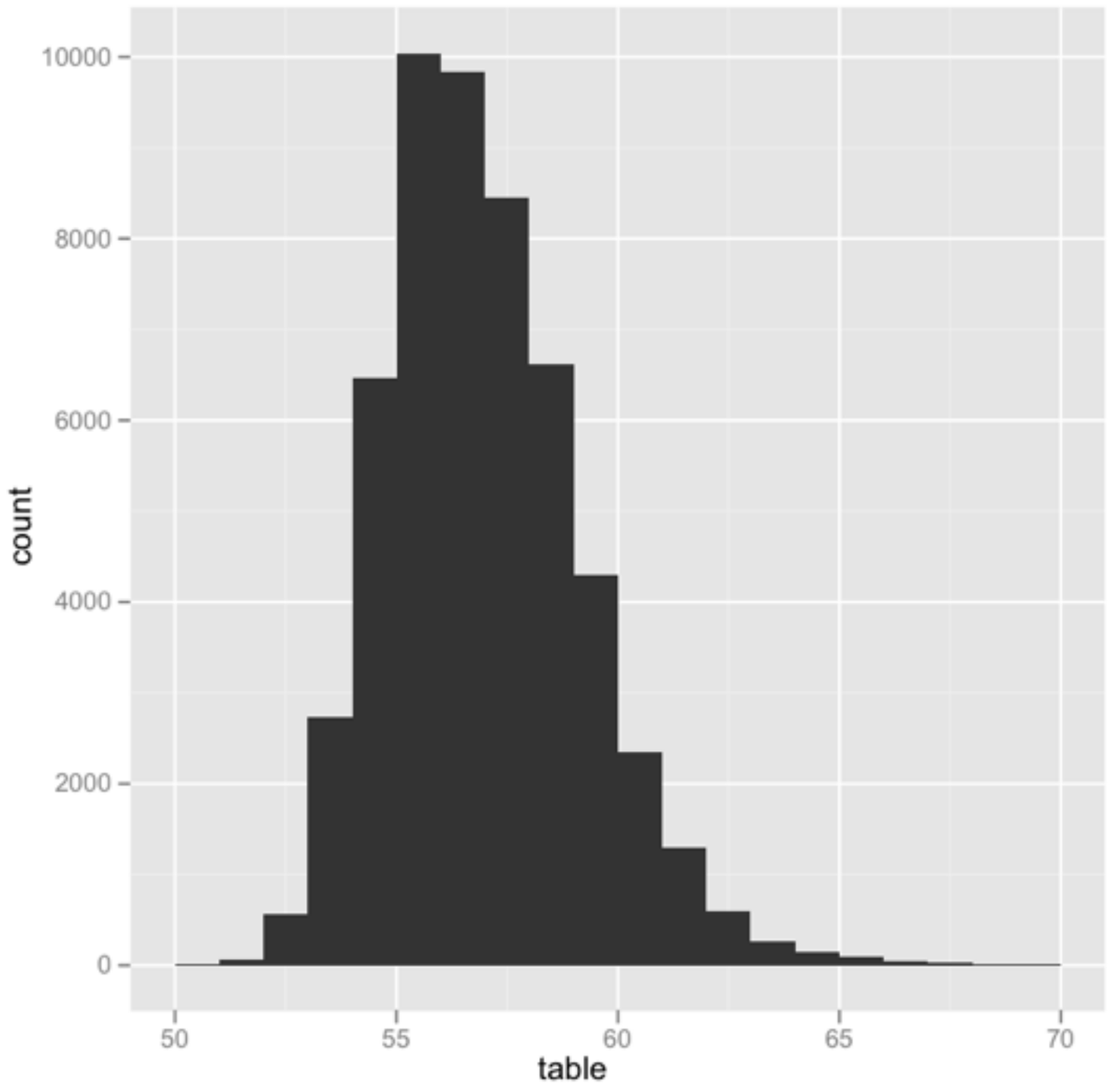

0
−10
−20
−30
−40
−50
−60
Z Label
120
100
80
60
40
20
0
Y Label
22
24
26
28
30
32
34
−20
X Label


3D Scatter Plot
Cost
Sales
2001
2002
2003
Year
Sales
Cost
Year


zs=0, zdir=z
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

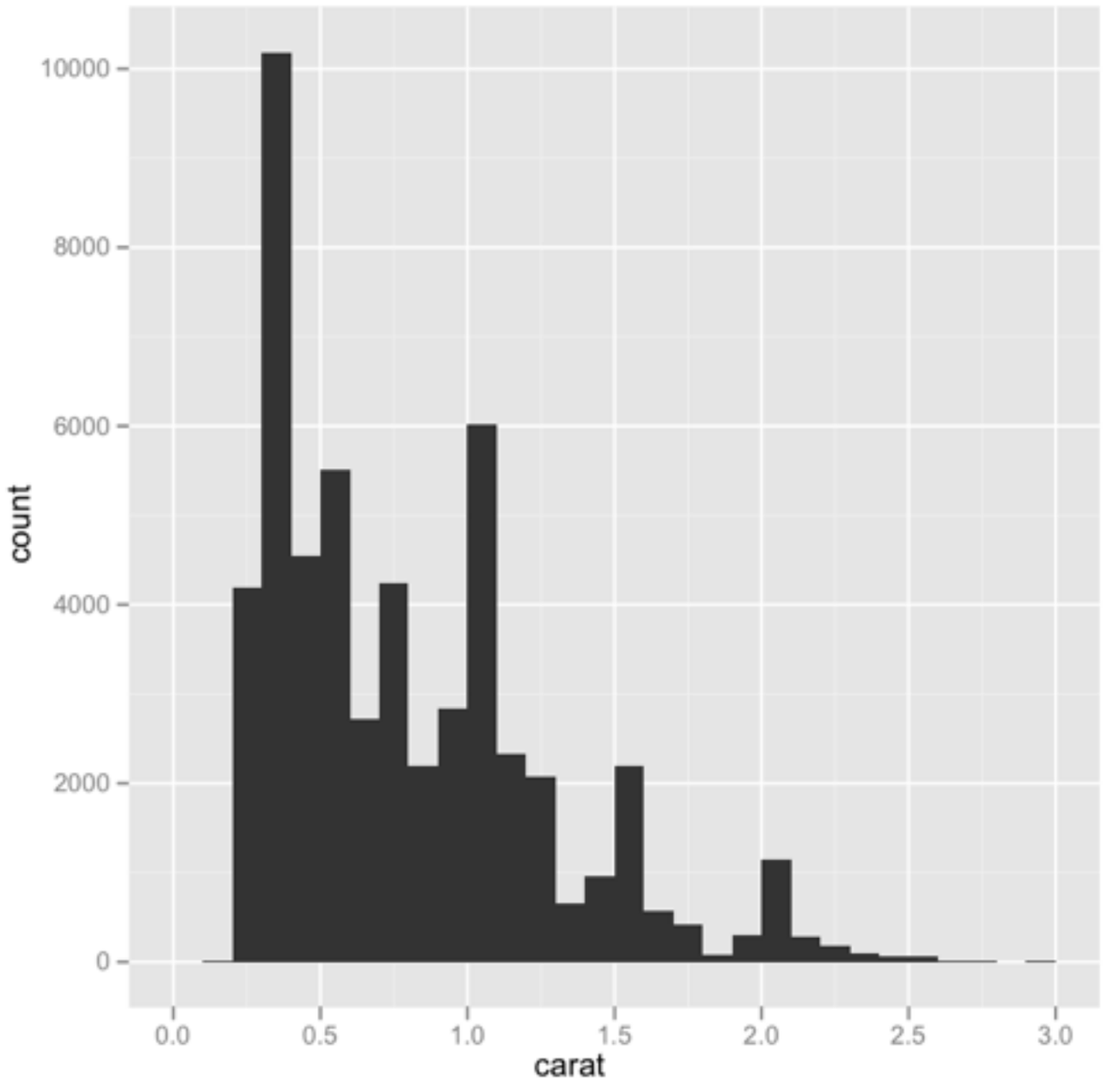
Distributions

Histogram

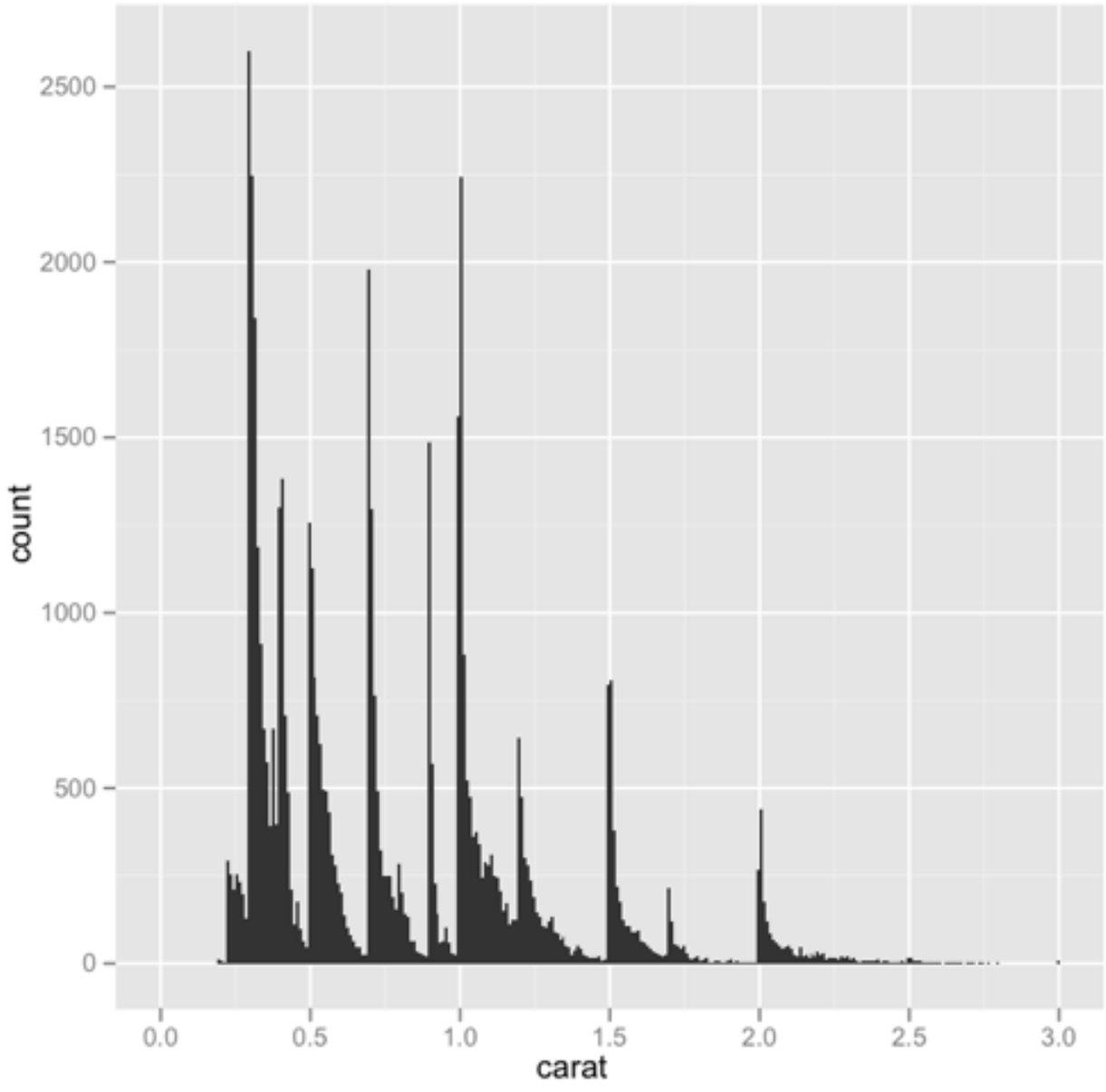


ggplot2

Bin Width

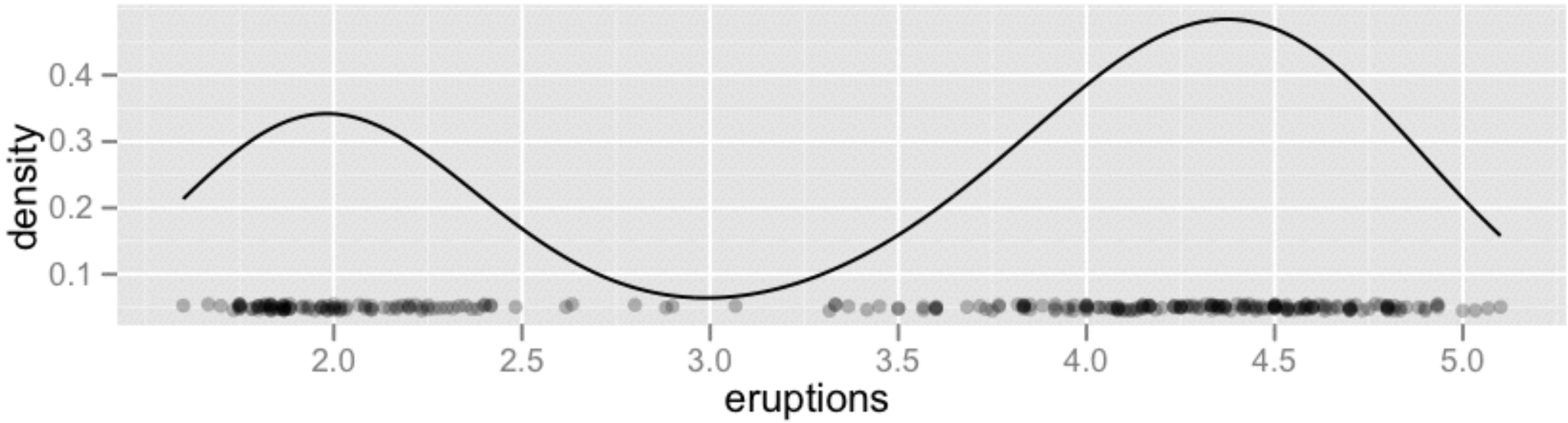


binwidth = 0.1

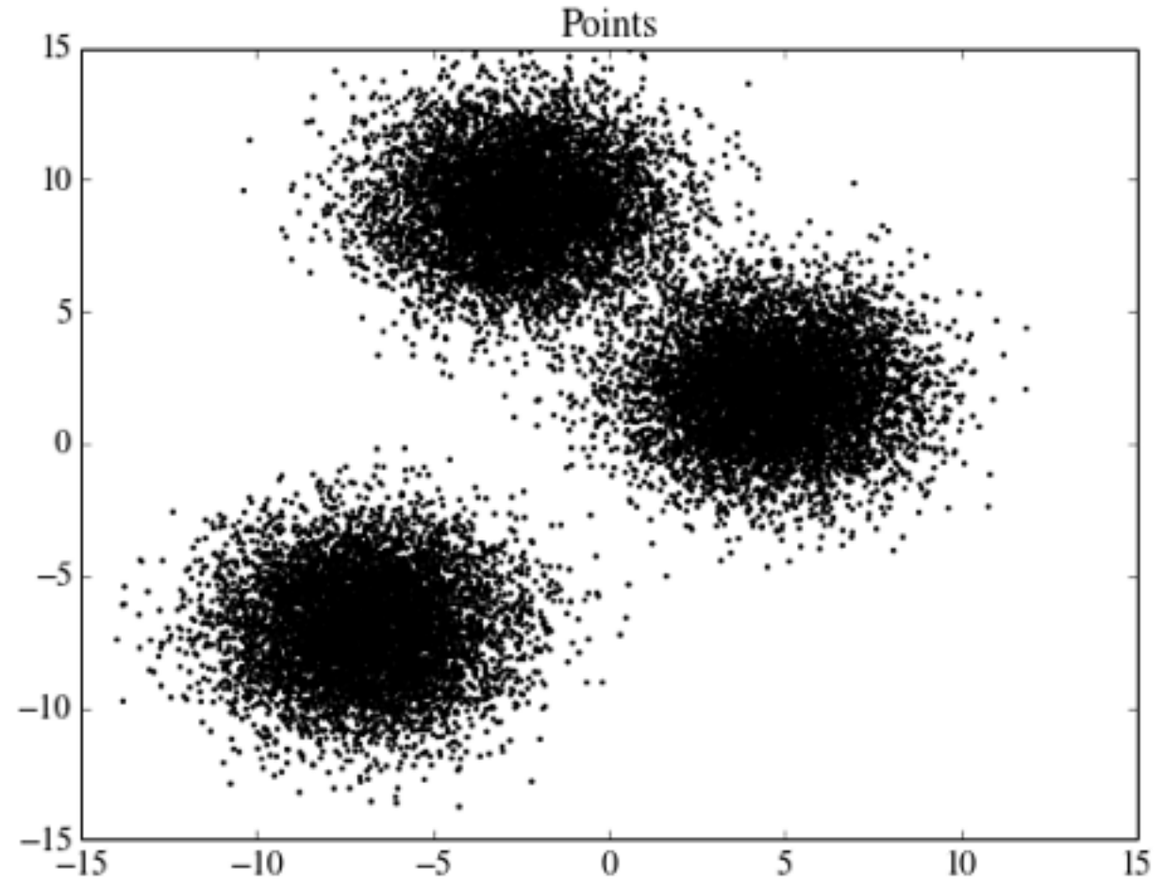


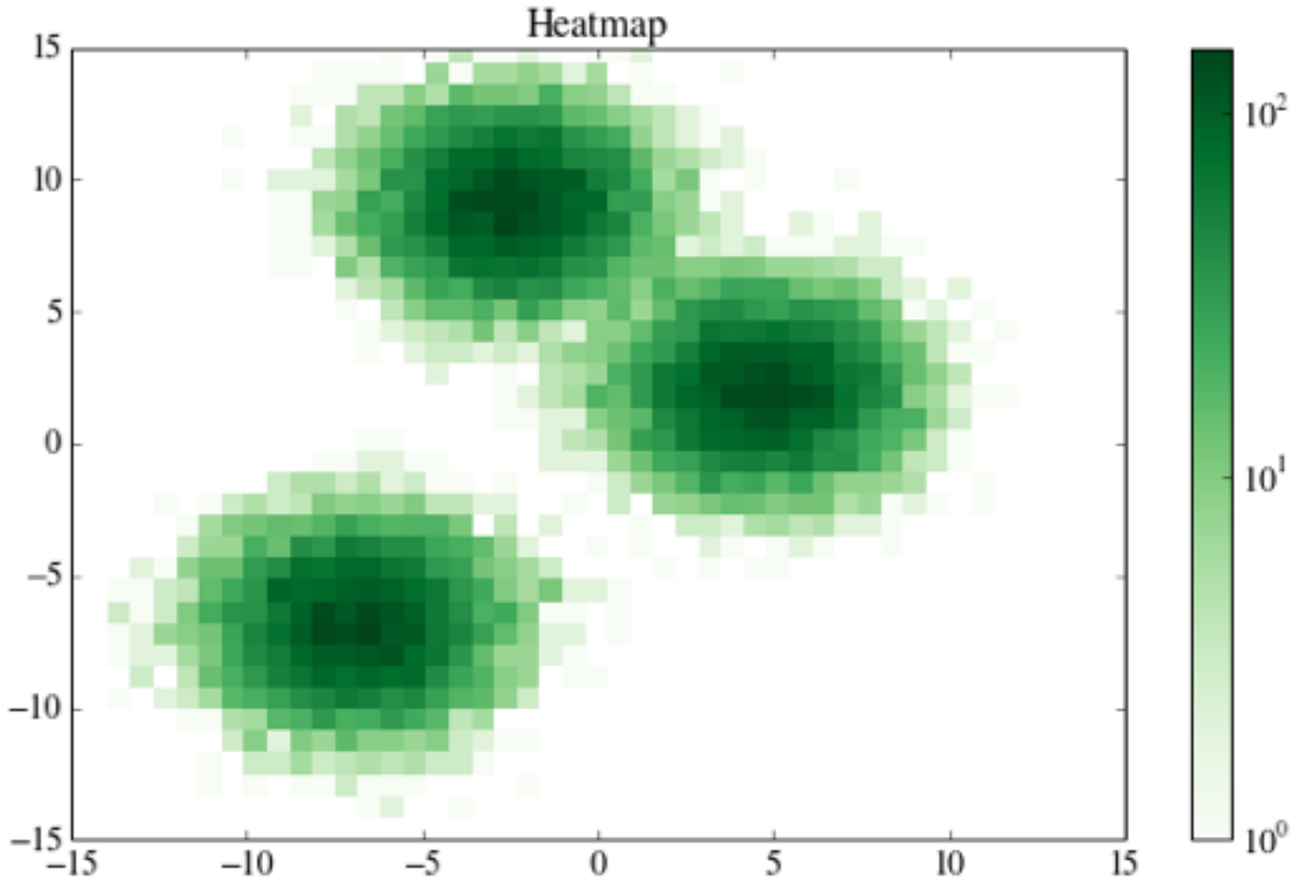
binwidth = 0.01

Density Plots

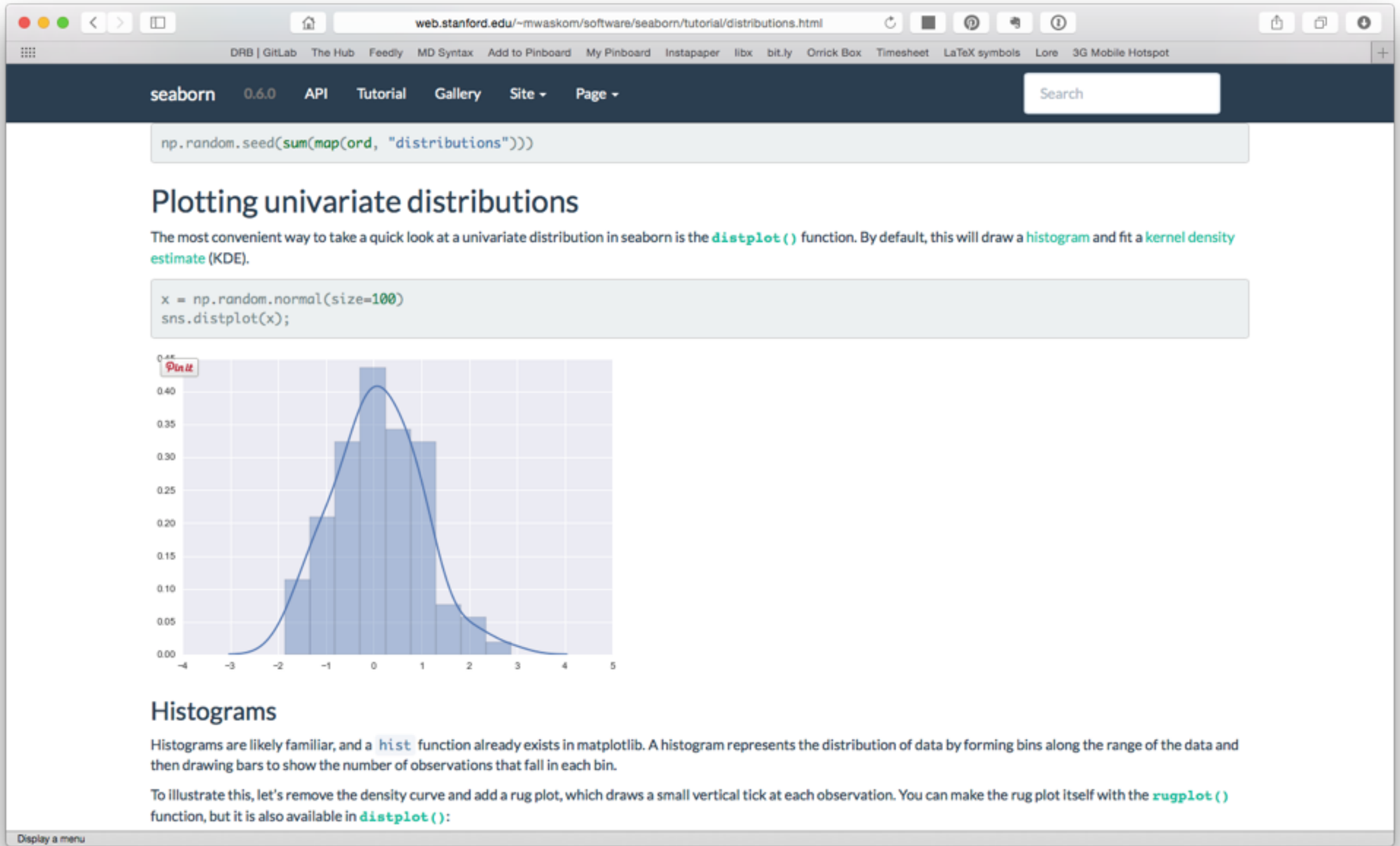


2D Density Plots





Seaborn Tutorial



Design Exercise

Hands-On Exercise

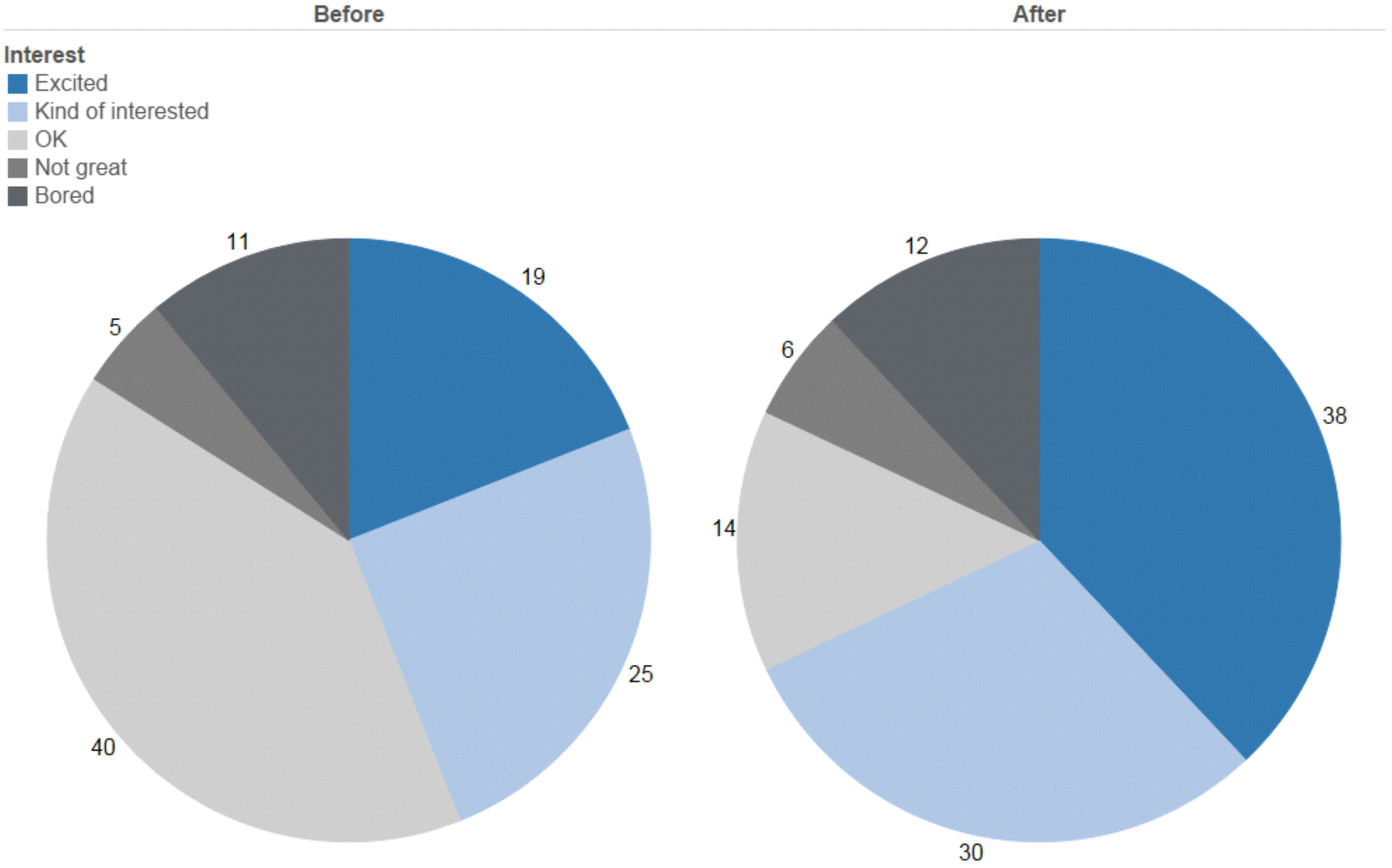
How do you feel about doing science?

Table

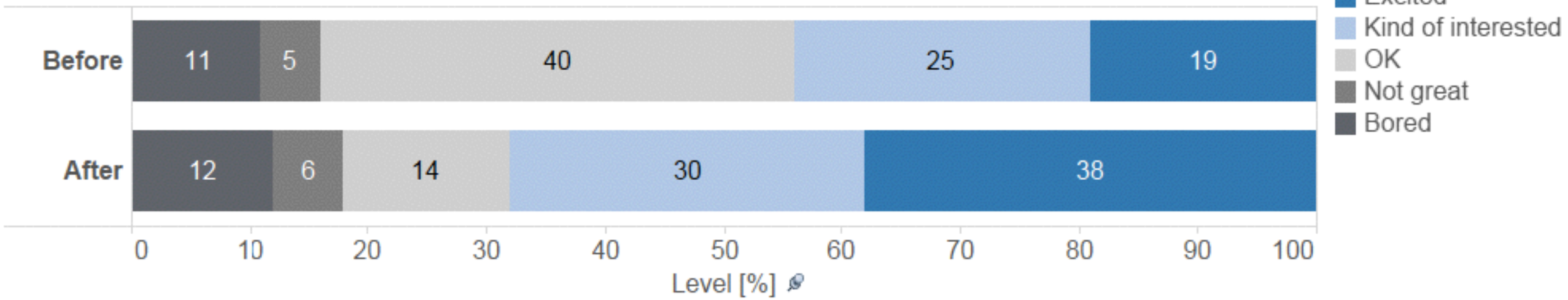
Interest	Before	After
Excited	19	38
Kind of interested	25	30
OK	40	14
Not great	5	6
Bored	11	12

Data courtesy of Cole Nussbaumer

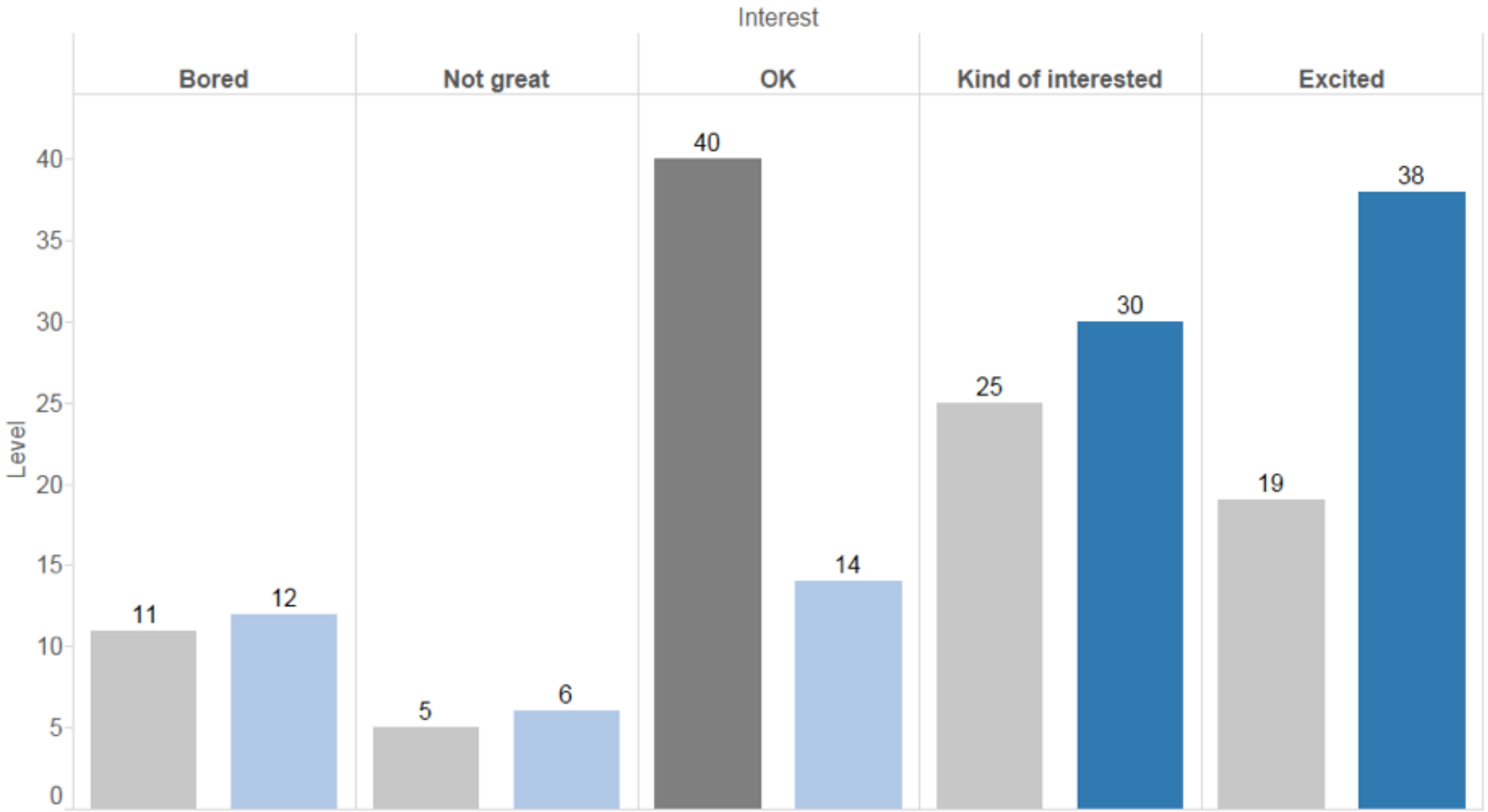
How do you feel about doing science?



How do you feel about doing science?

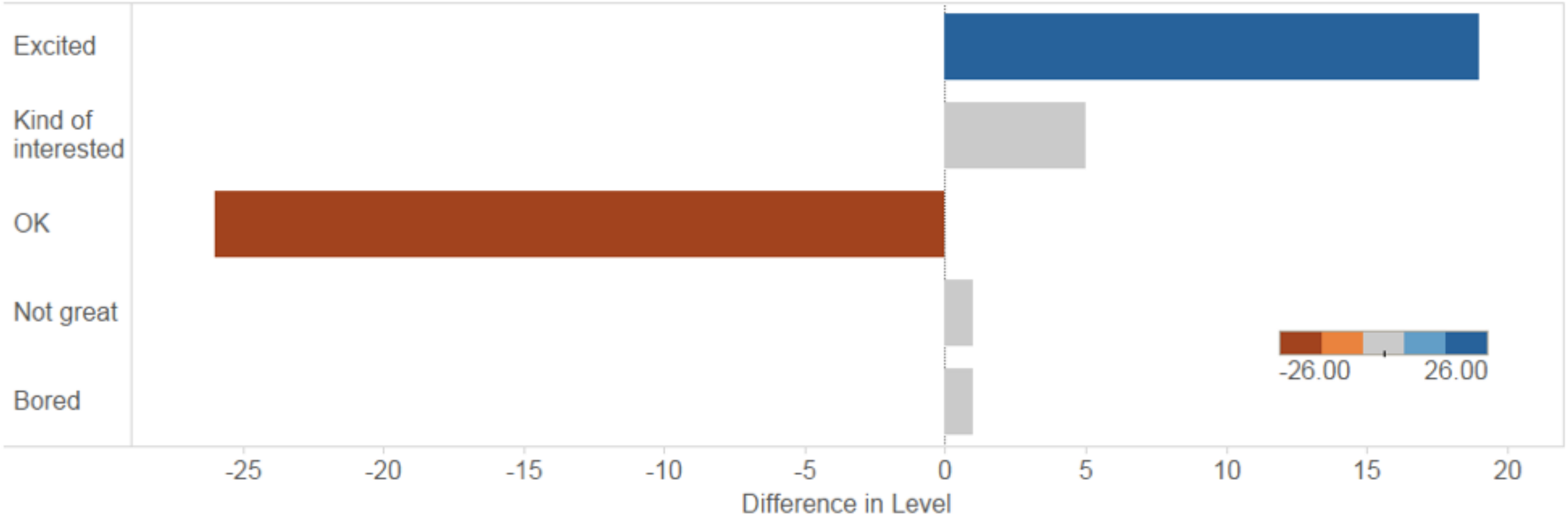


How do you feel about doing science?

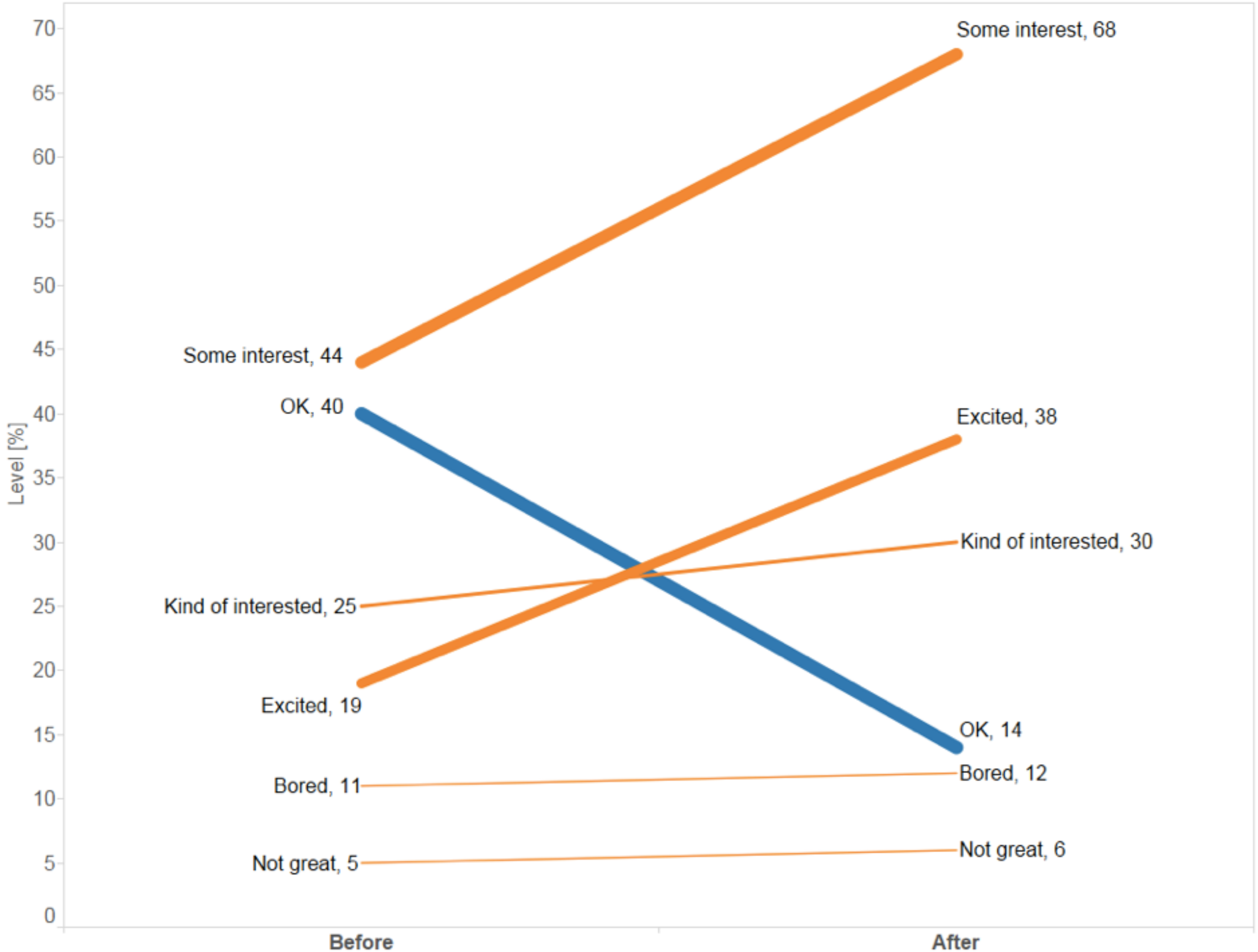


Before the program, the majority of children felt just ***OK*** about science. After the program, more children were ***Kind of interested*** and ***Excited*** about science.

Opinion change to the question: How do you feel about doing science?



How do you feel about doing science?

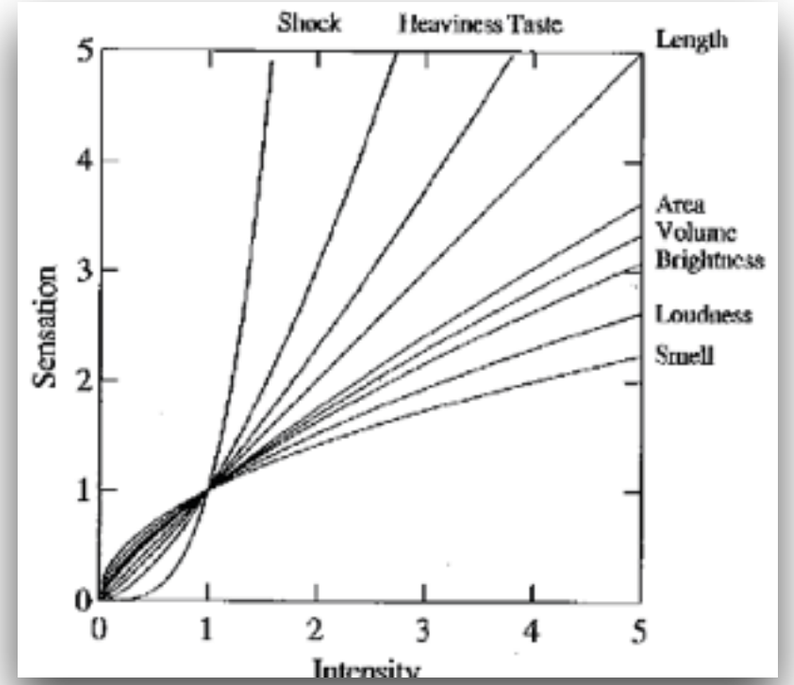


After the pilot program,

68%

of kids expressed interest towards science,
compared to 44% going into the program.

Perceptual Effectiveness



Stephen's Power Law, 1961

	Nominal	Ordinal	Quantitative
Position	✔	✔	✔
Size	✔	✔	~
(Grey)Value	✔	✔	~
Texture	✔	~	✖
Color	✔	✖	✖
Orientation	✔	✖	✖
Shape	✔	✖	✖

✔ = Good
~ = OK
✖ = Bad

J. Bertin, 1967

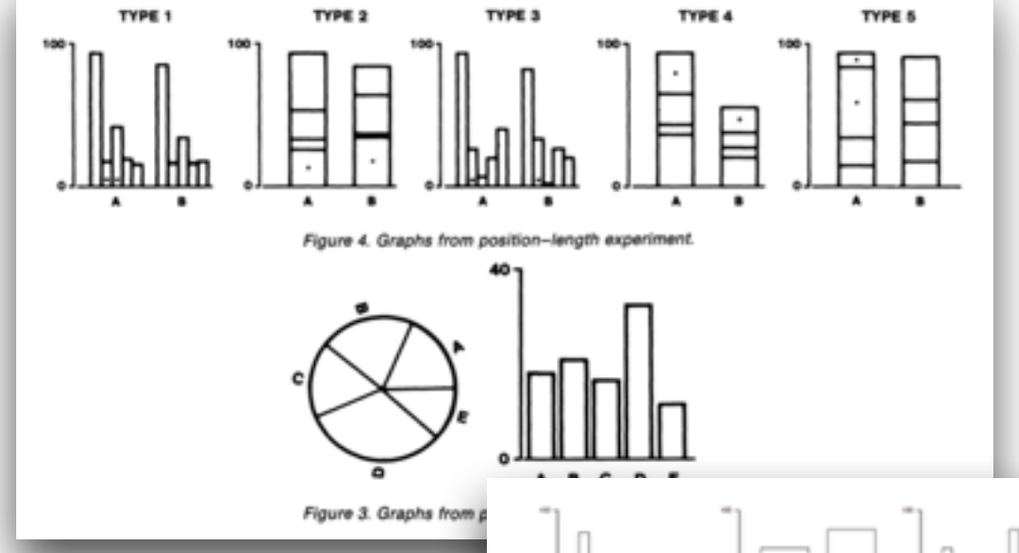
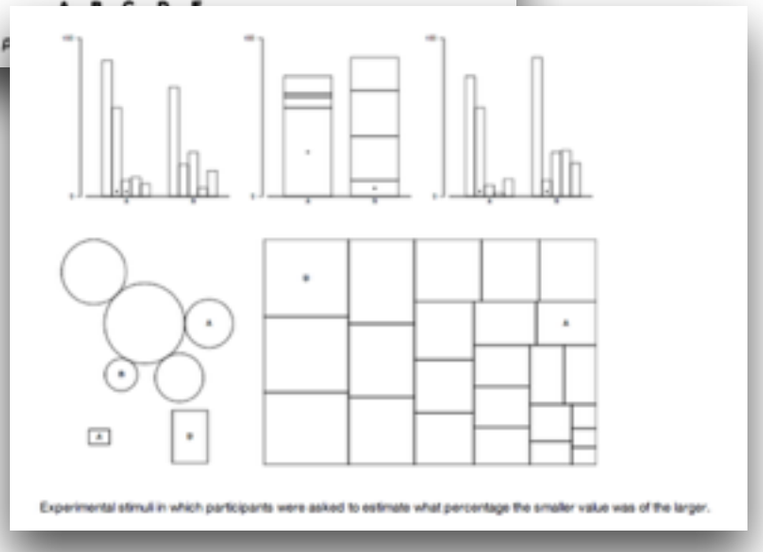


Figure 4. Graphs from position–length experiment.

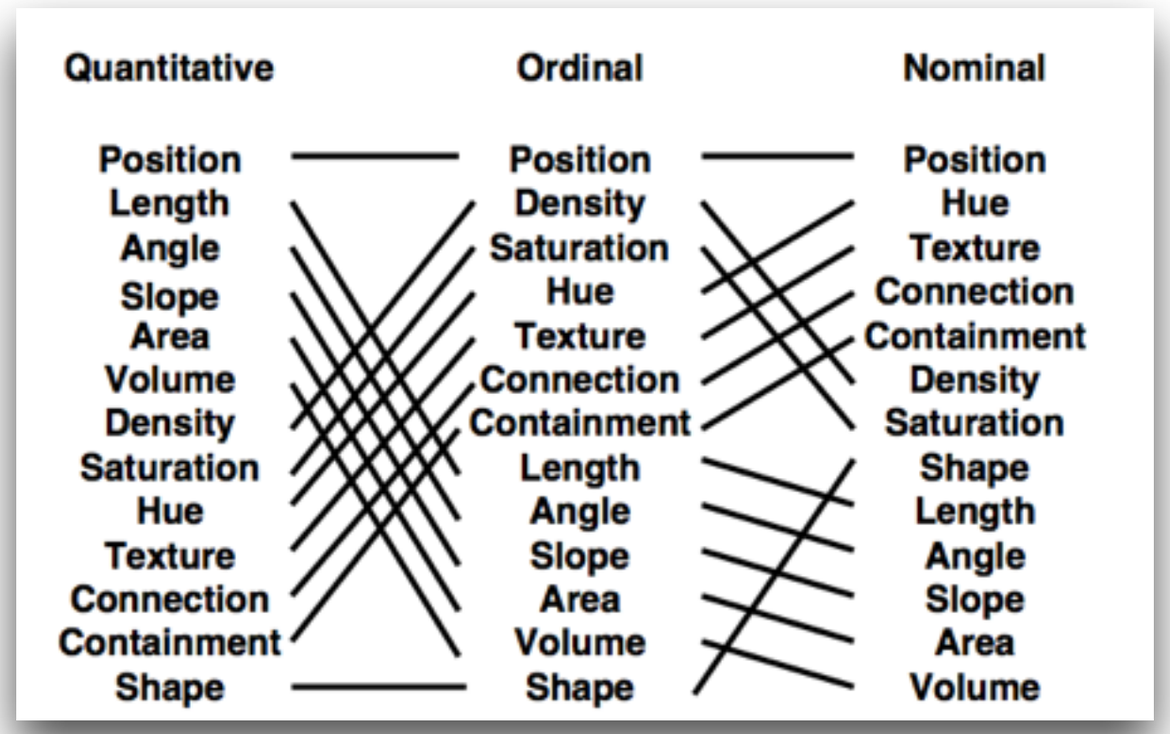
Figure 3. Graphs from p

Cleveland / McGill, 1984



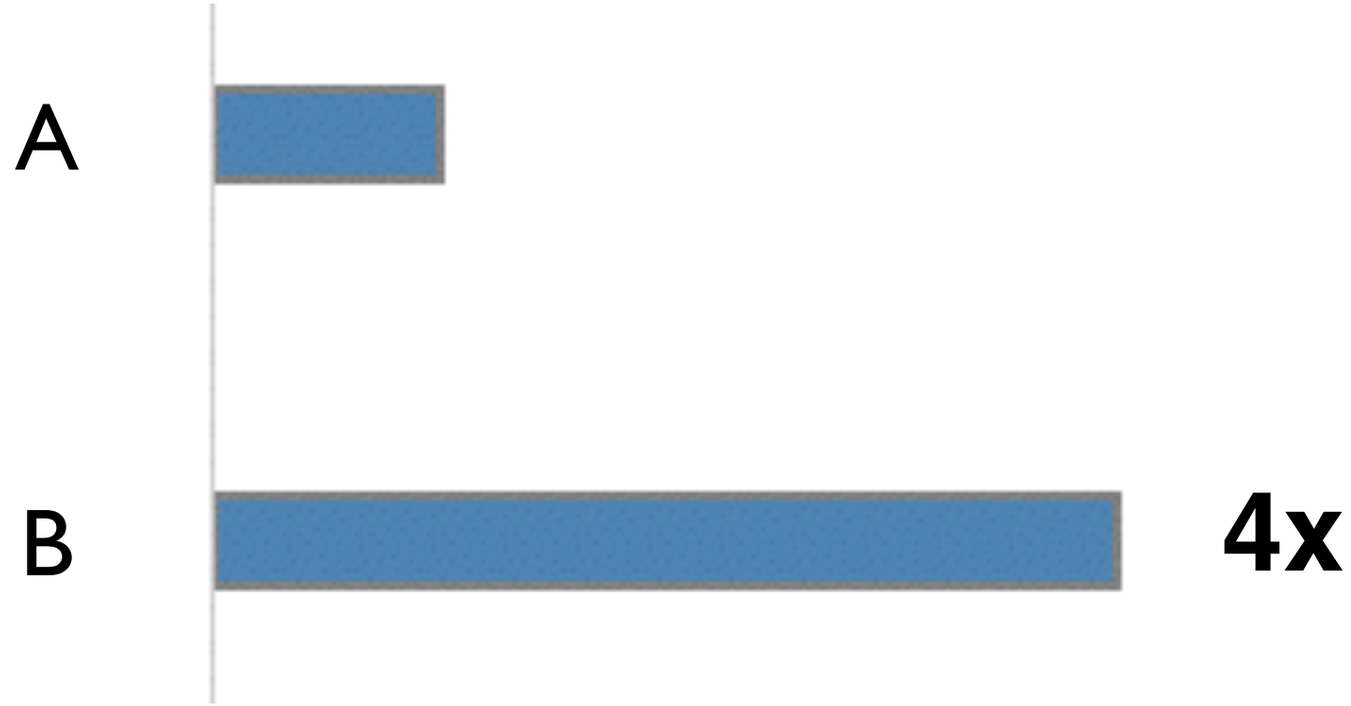
Experimental stimuli in which participants were asked to estimate what percentage the smaller value was of the larger.

Heer / Bostock, 2010

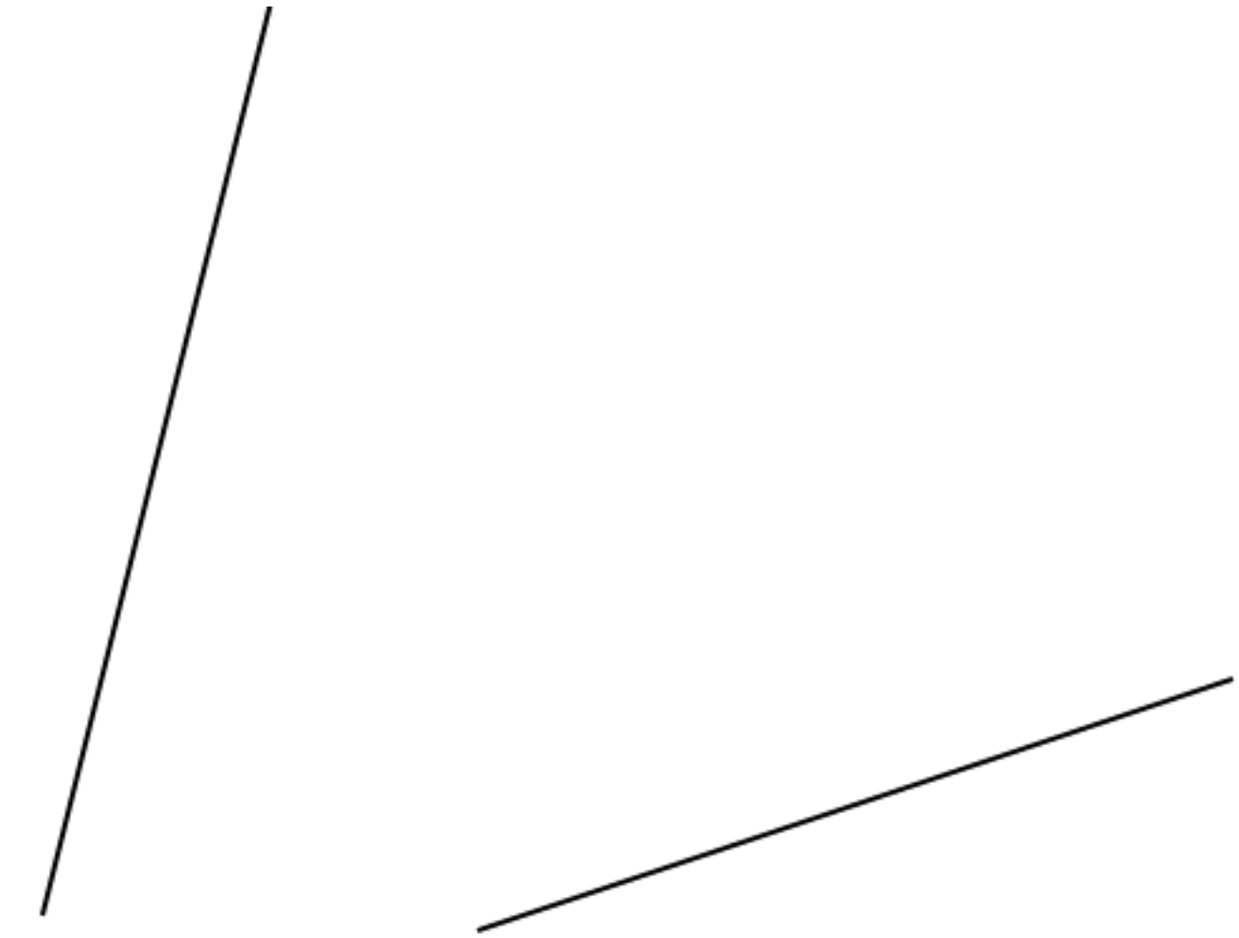


J. Mackinlay, 1986

How much longer?



How much steeper slope?

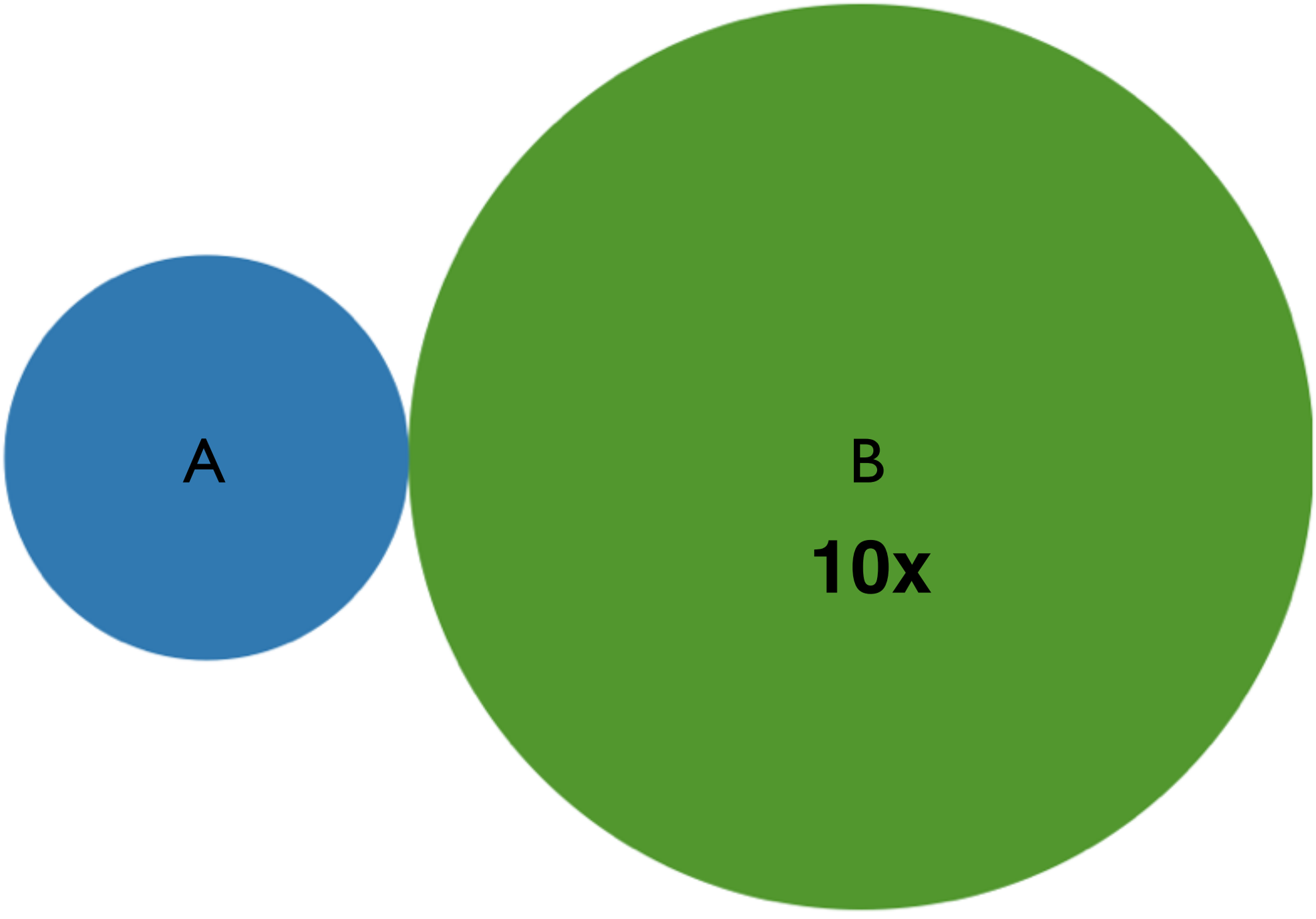


A

B

4x

How much larger area?



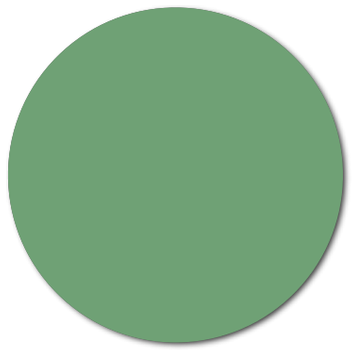
How much darker?

A

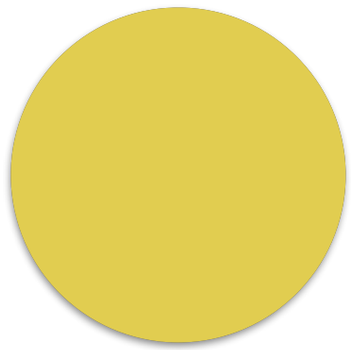
B

2x

How much bigger value?



A



B

4x

2

16

Most Efficient

Position

Length

Slope

Angle

Quantitative

Area

Intensity

Ordered

Least Efficient

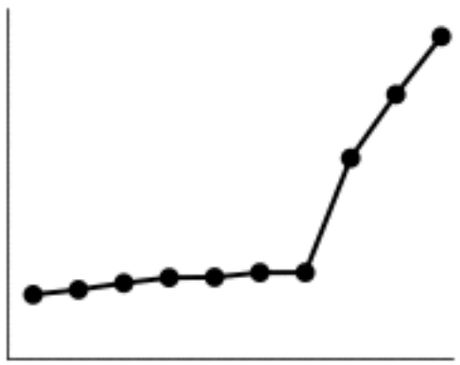
Color

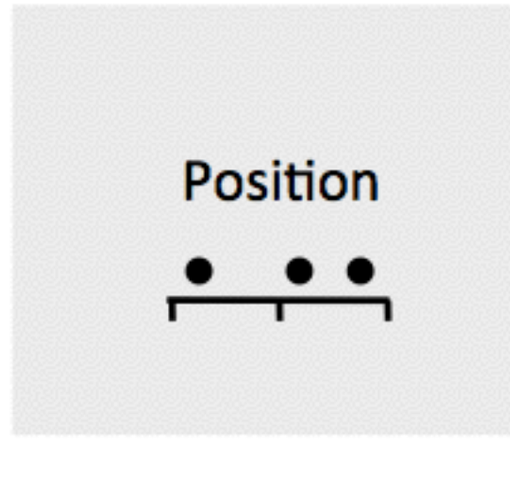
Shape

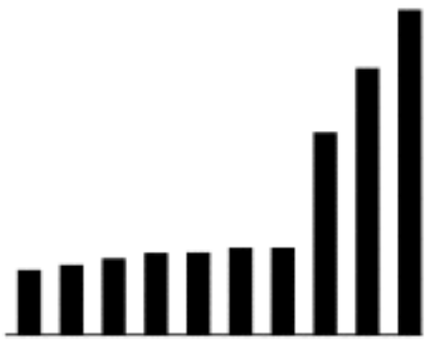
Categories

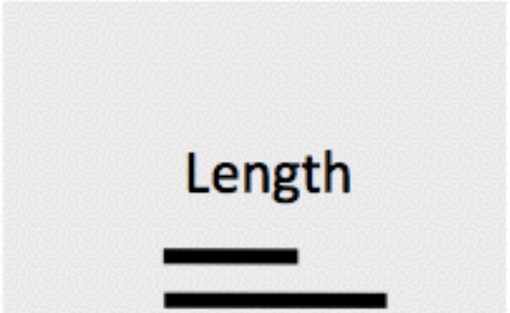
C. Mulbrandon
VisualizingEconomics.com

Most Effective

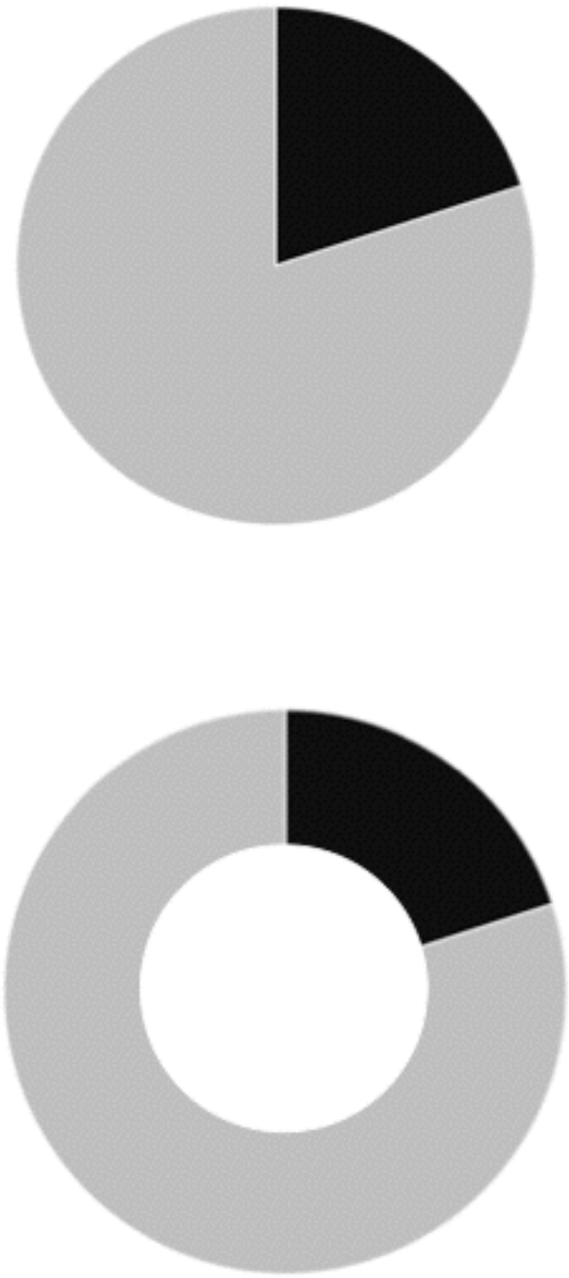


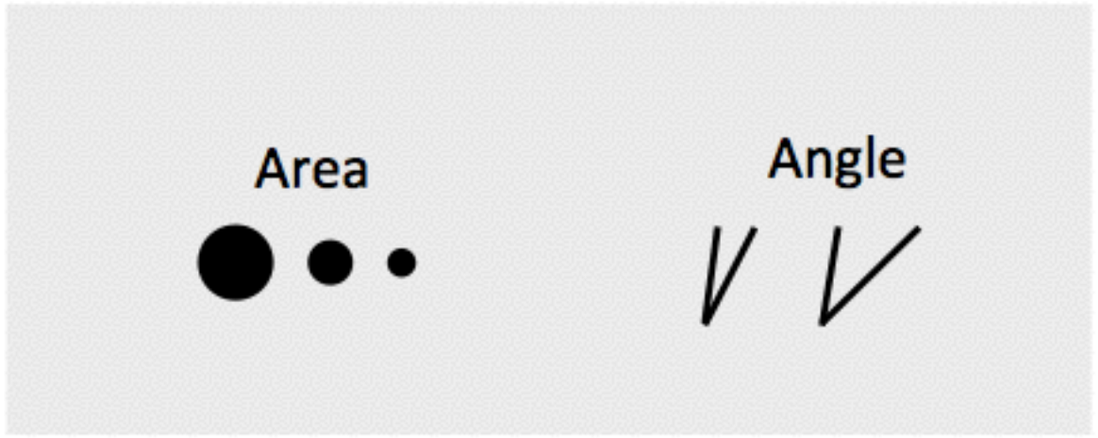


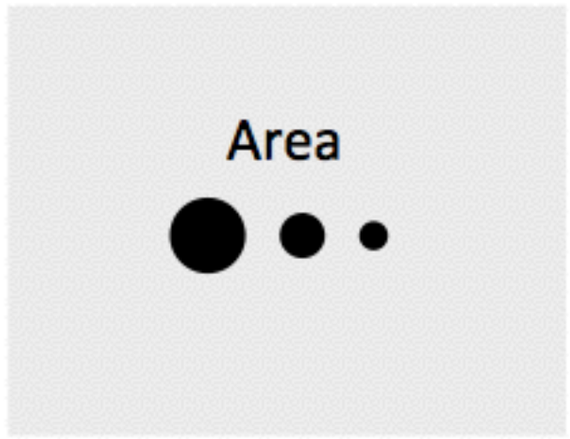




Less Effective

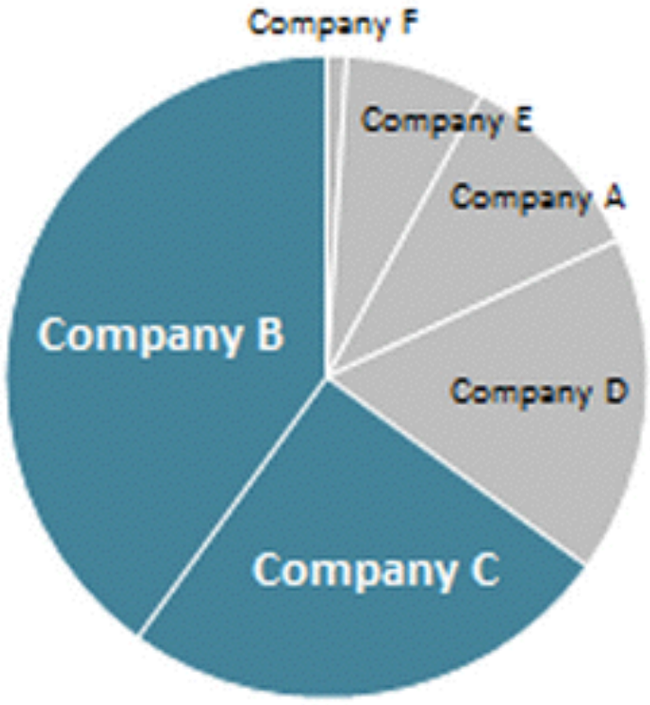


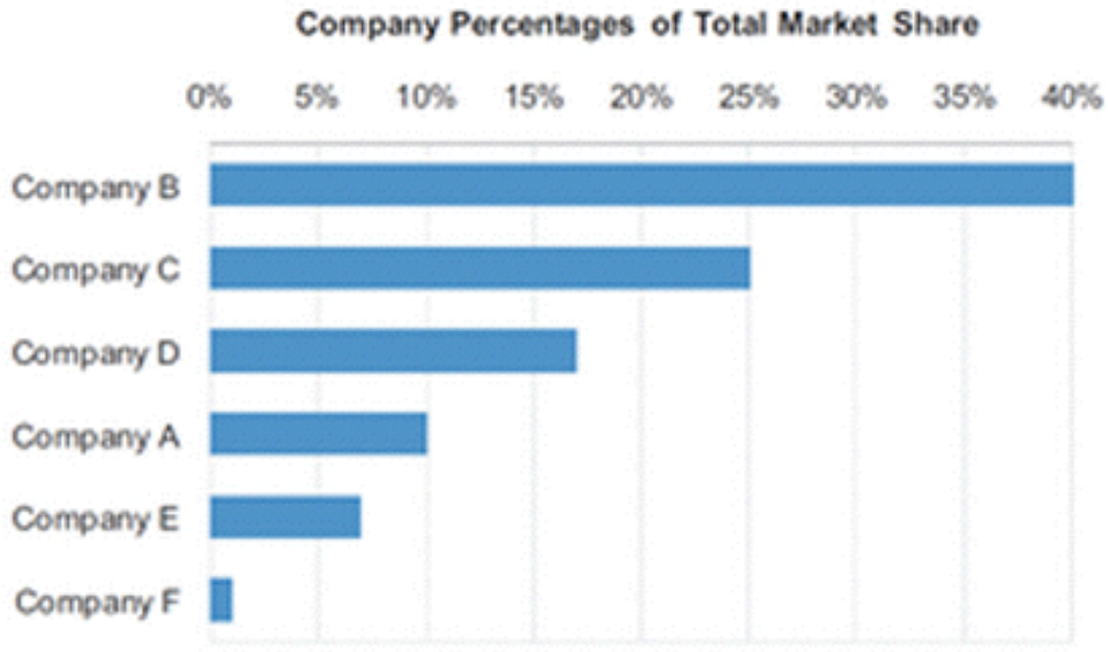




Pie vs. Bar Charts

65% of the market is controlled by companies B and C





Least Effective

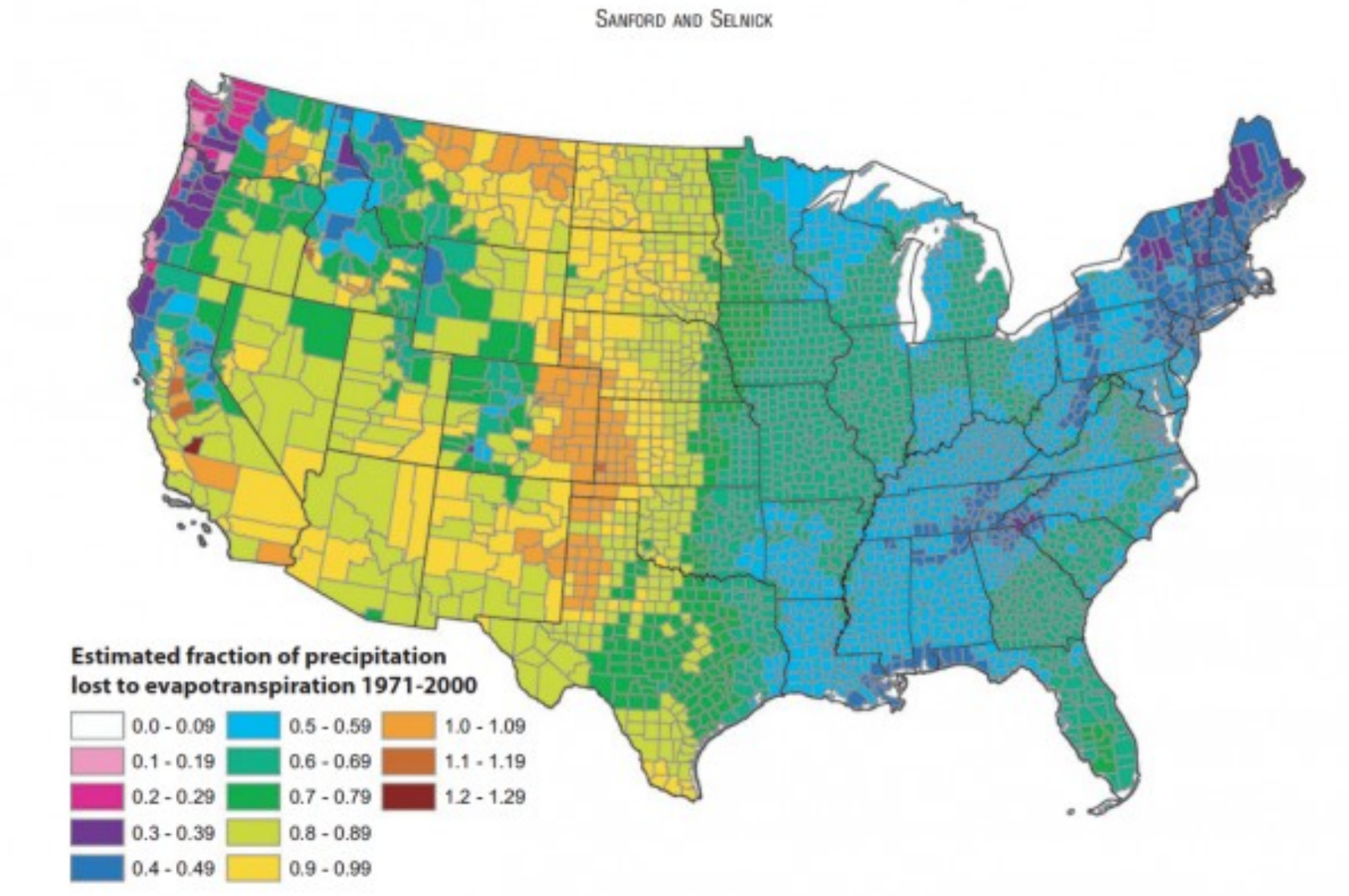
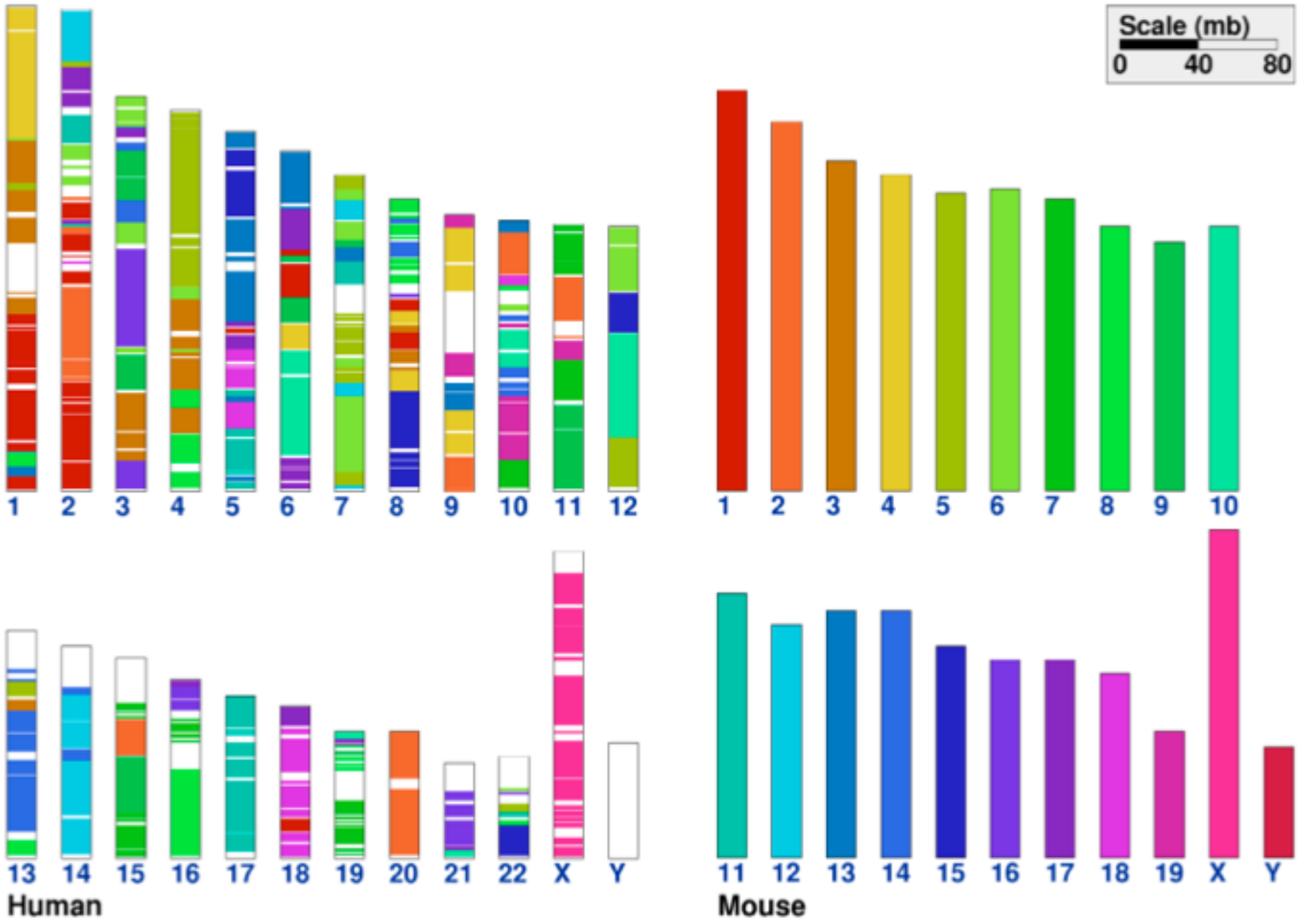


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (P) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/P were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.

Cliff Mass

Use Color Strategically

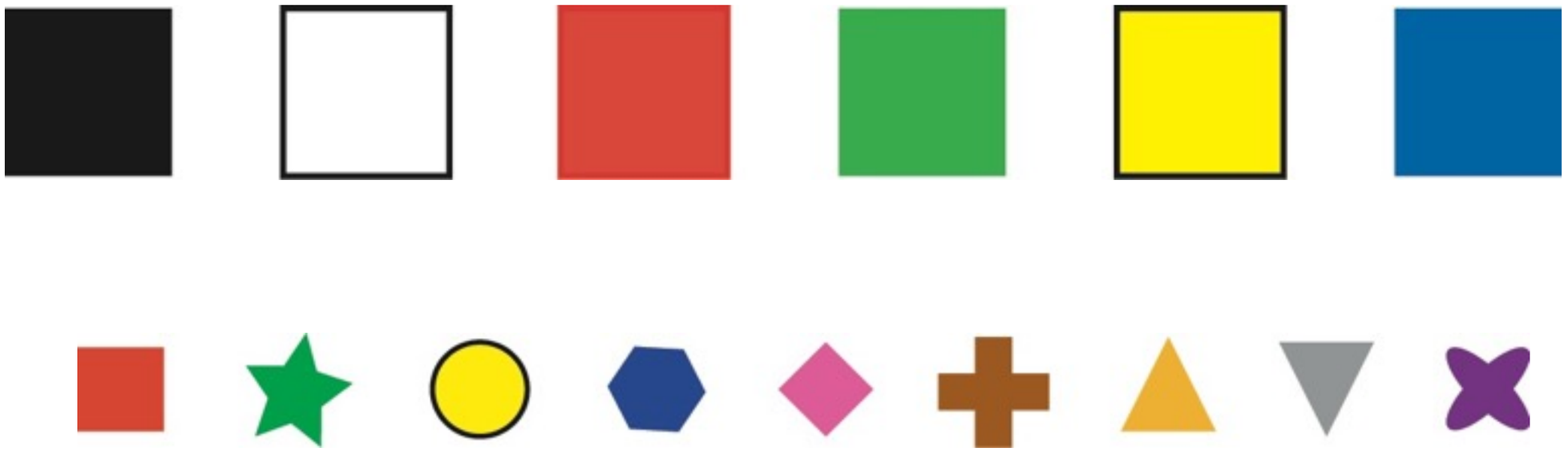
Color Discriminability



Sinha 2007

Colors for Categories

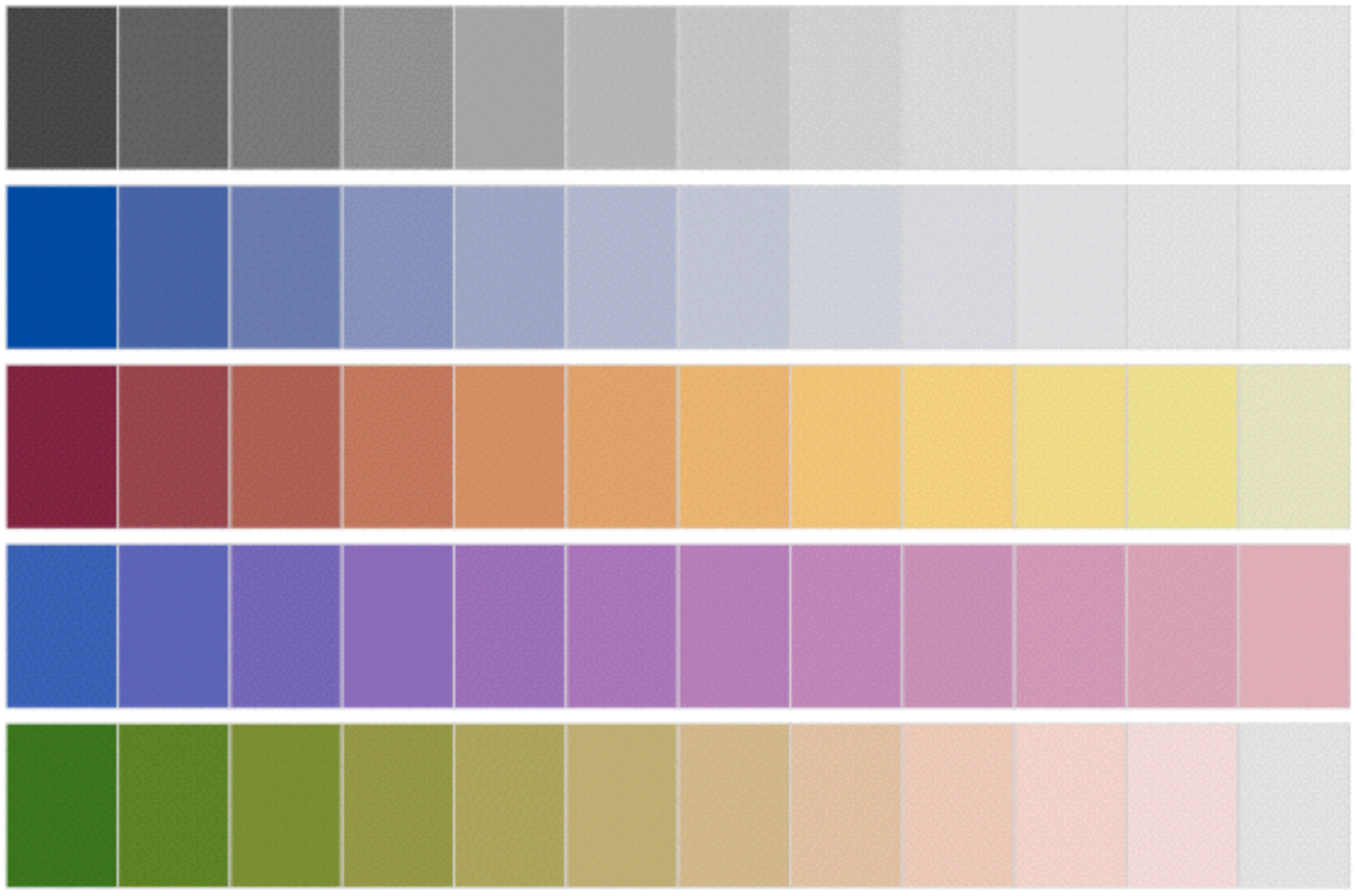
Do not use more than 5-8 colors at once



Ware, "Information Visualization"

Colors for Ordinal Data

Vary luminance and saturation

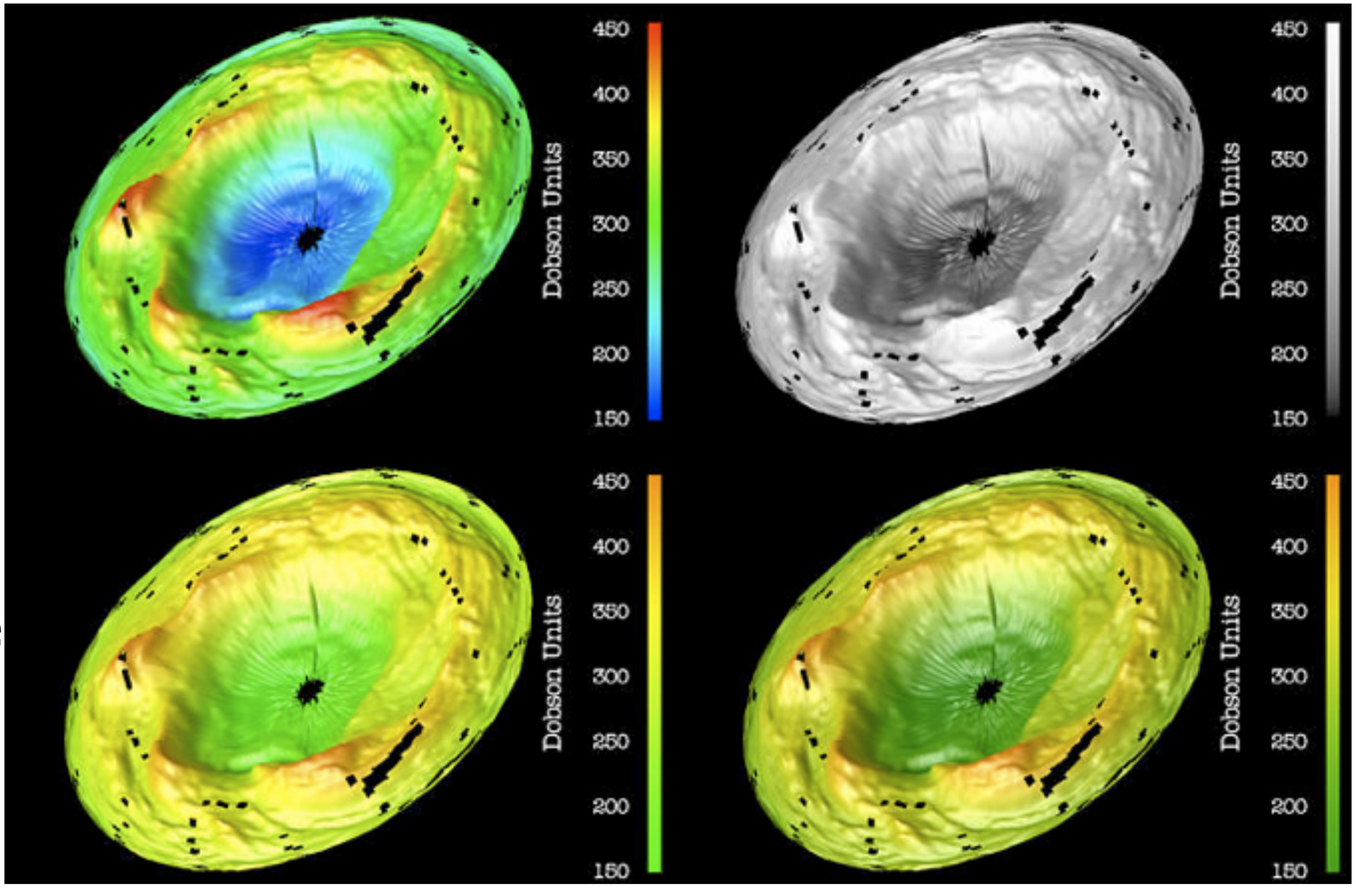


Zeilis et al, 2009, "Escaping RGBland: Selecting Colors for Statistical Graphics"

Colors for Quantitative Data

Hue
(Rainbow)

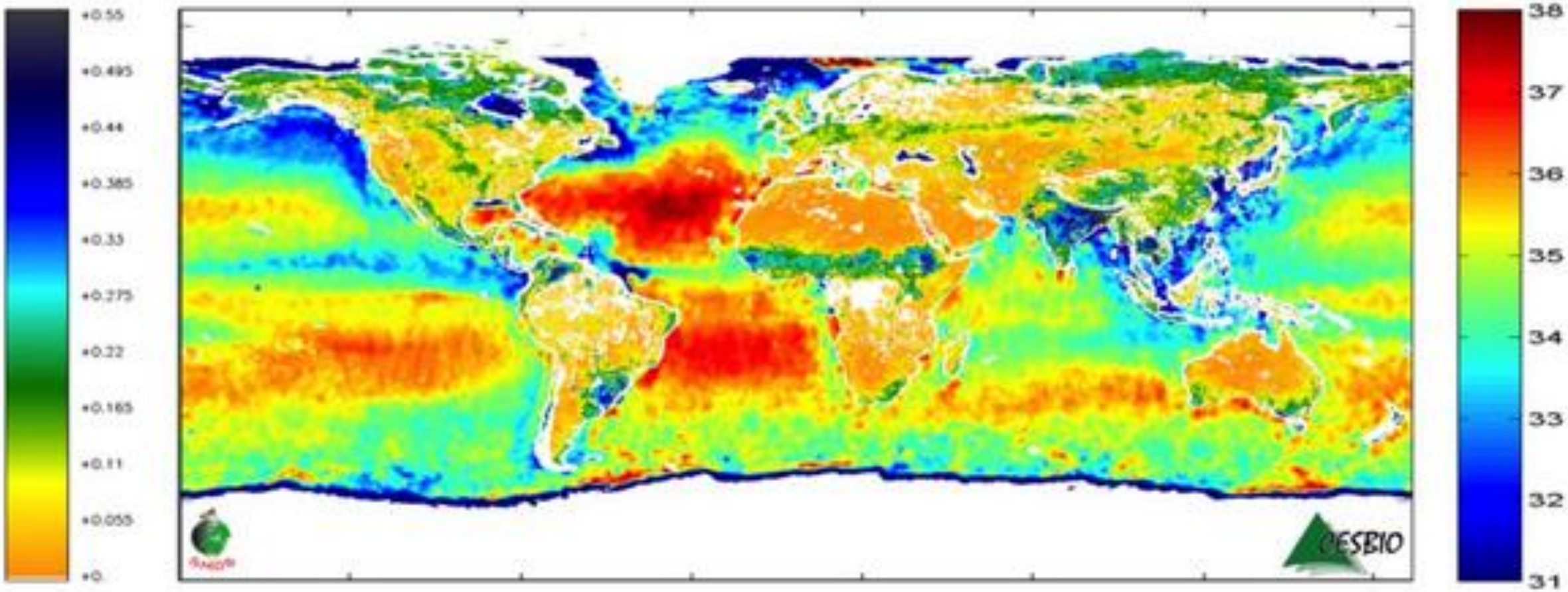
Luminance
& Hue



Luminance

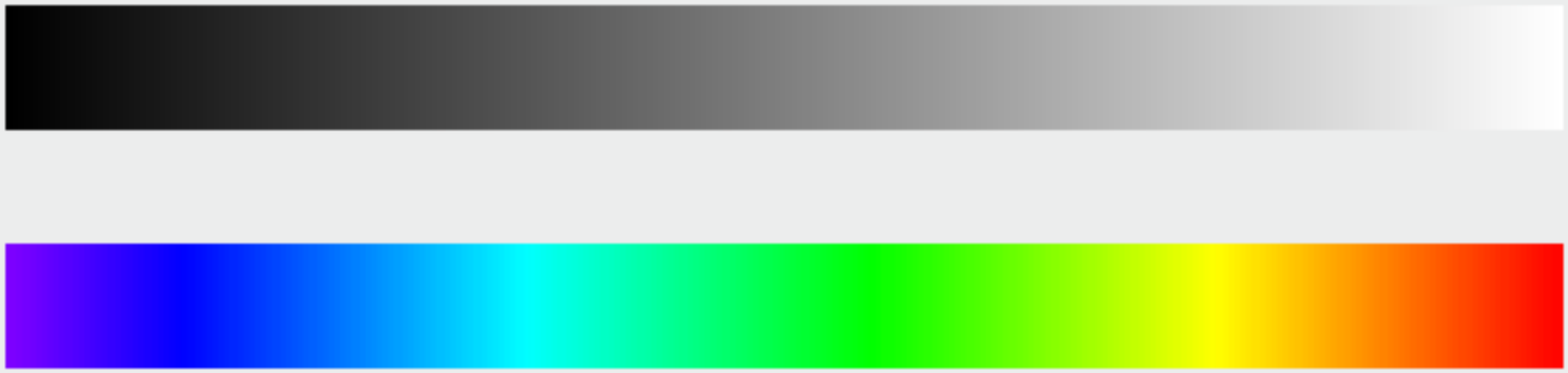
Rogowitz and Treinish, Why should engineers and scientists be worried about color?

Rainbow Colormap



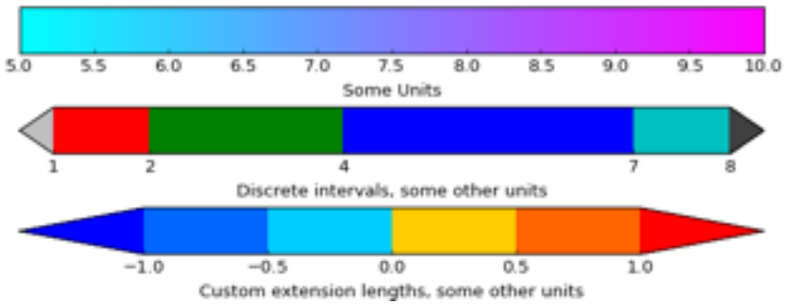
Rainbow Colormap

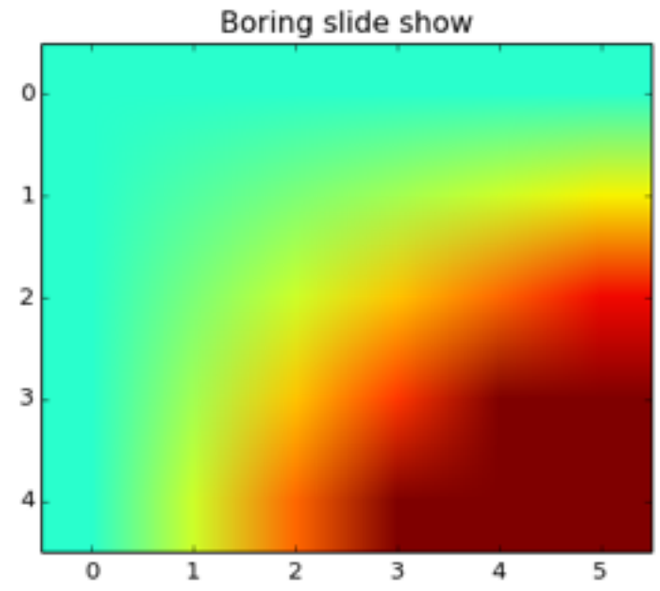
Perceptually nonlinear

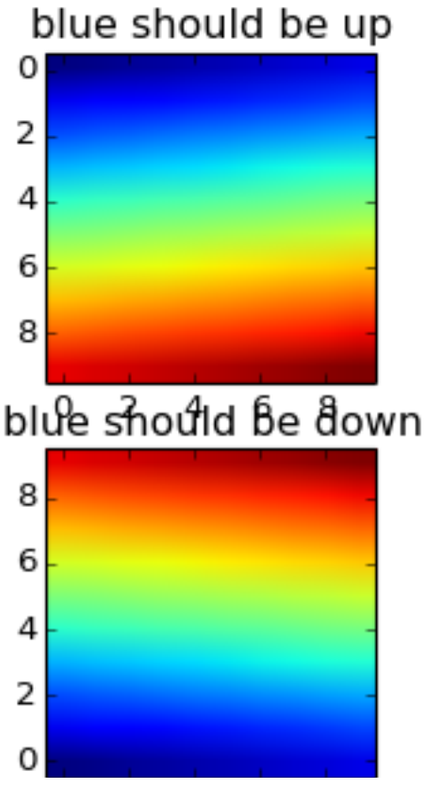


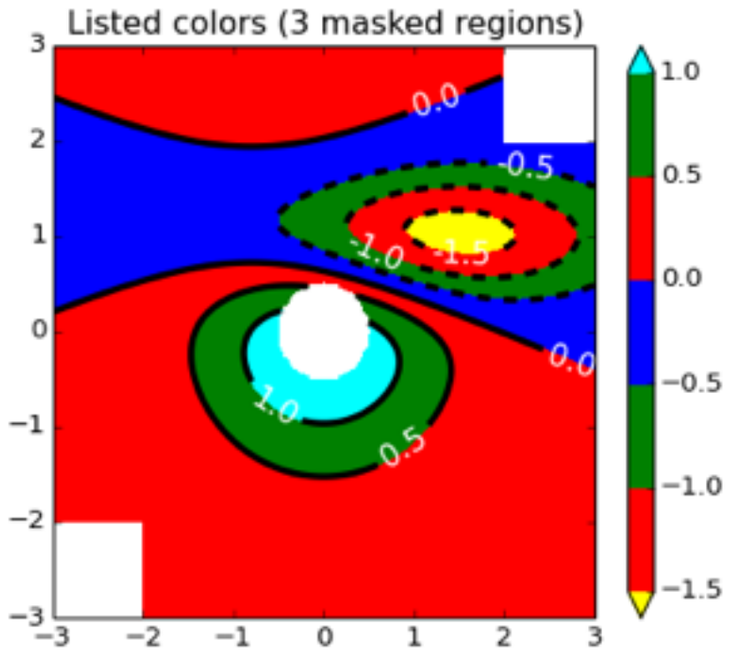
R. Simmon

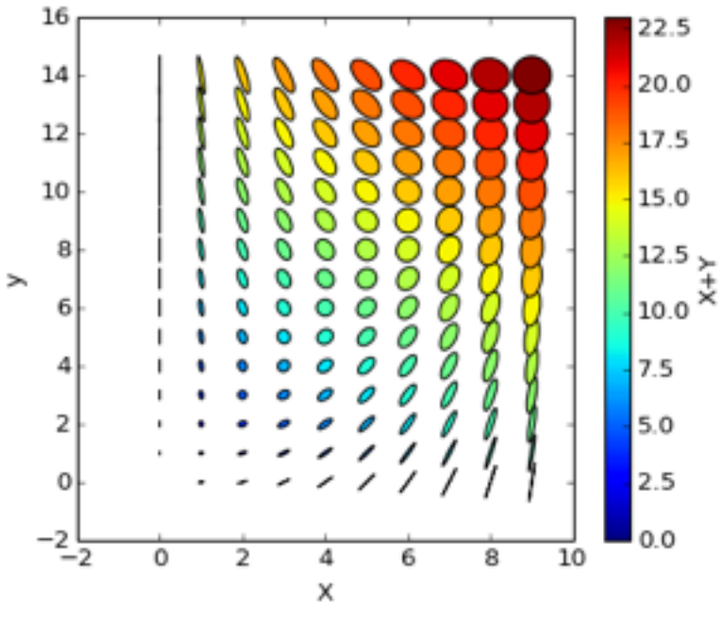
Avoid Rainbow Colors!

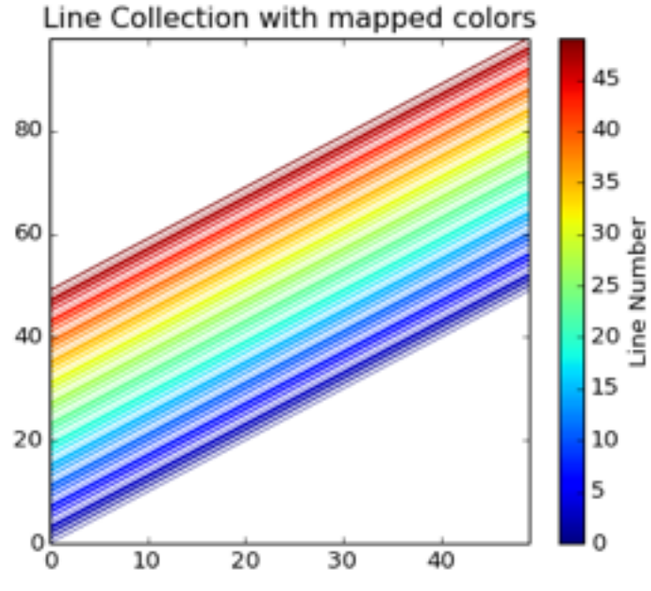






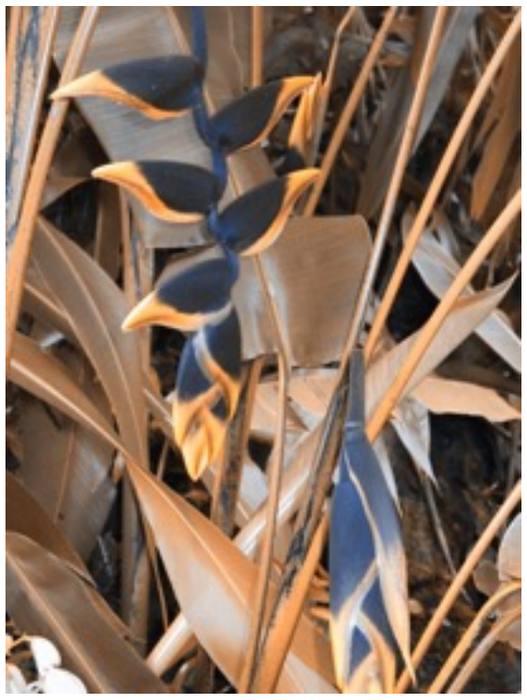




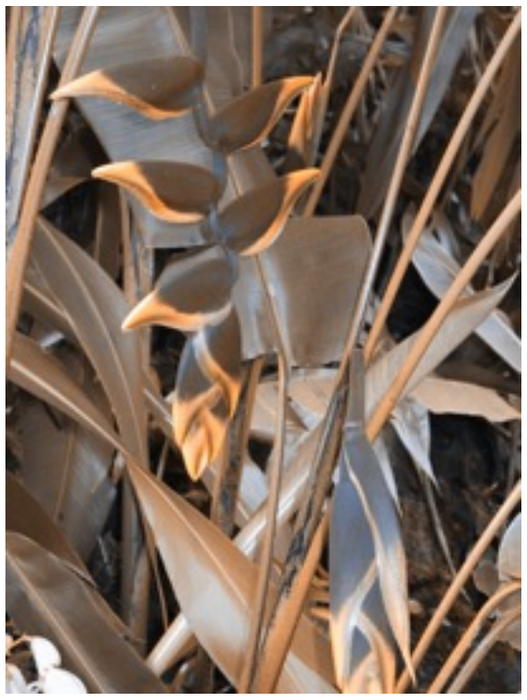


matplotlib gallery

Color Blindness



Protanope



Deuteranope

Red / green deficiencies

Tritanope

Blue / Yellow deficiency

Based on slide from Stone

Color Blindness

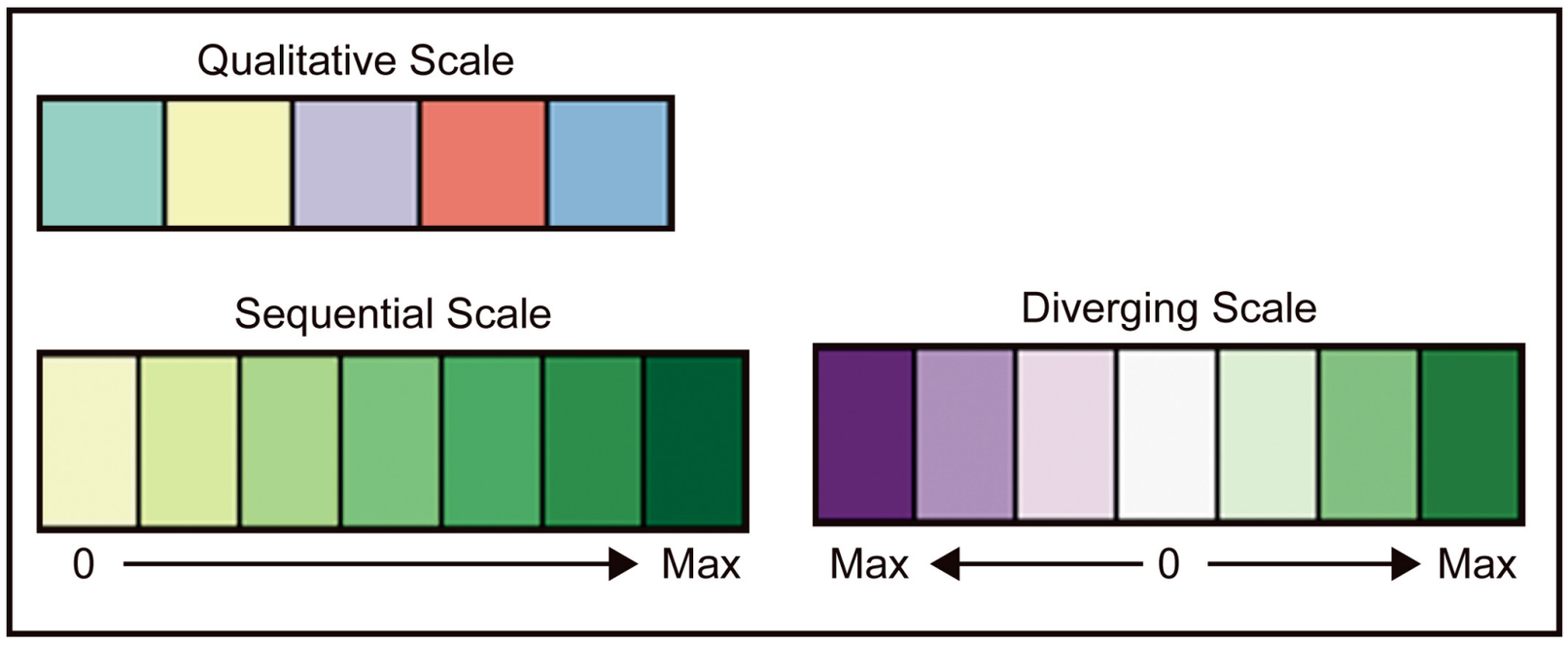
Normal Protanope Deuteranope Lightness

Based on slide from Stone

Color Brewer

Nominal

Ordinal



Cynthia Brewer, Color Use Guidelines for Data Representation

how to use | updates | credits

COLORBREWER 2.0

color advice for cartography

number of data classes on your map

3 | learn more >

the nature of your data

sequential | learn more >

pick a color scheme: GnBu

multihue | single hue

(optional) only show schemes that are:

- [x] colorblind safe
- [ ] print friendly
- [ ] photocopy-able

learn more >

pick a color system

224, 243, 219
168, 221, 181
67, 162, 202

(•) RGB ( ) CMYK ( ) HEX

adjust map context

- [ ] roads
- [ ] cities
- [x] borders

select a background

(•) solid color
( ) terrain

color transparency

learn more >

EXPORT YOUR COLORS >>

SCORE CARD

Support

Back to ColorBrewer 1.0

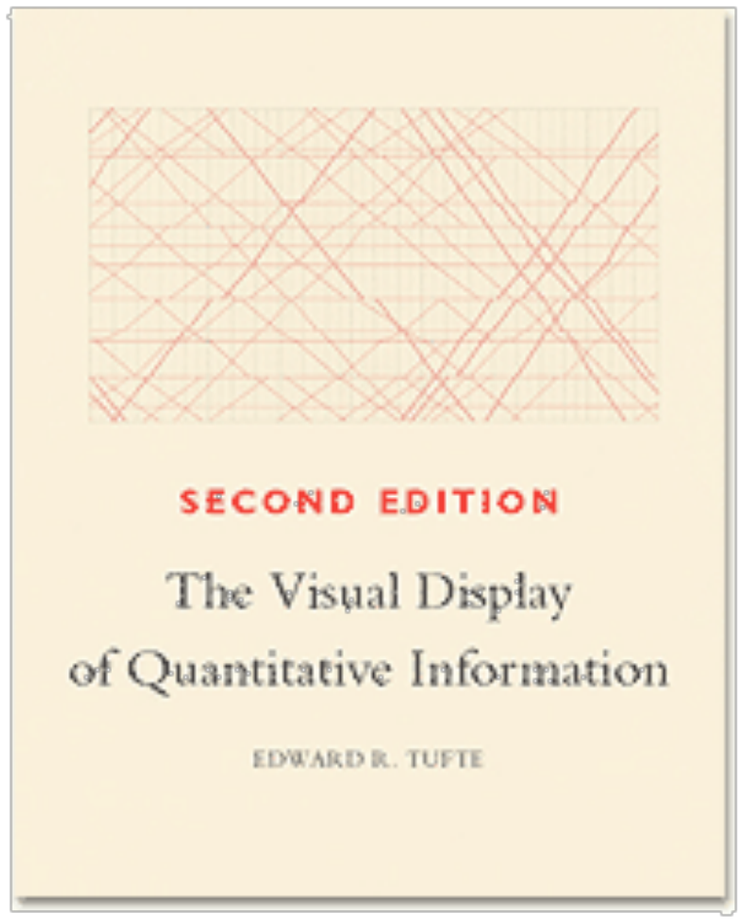
axm

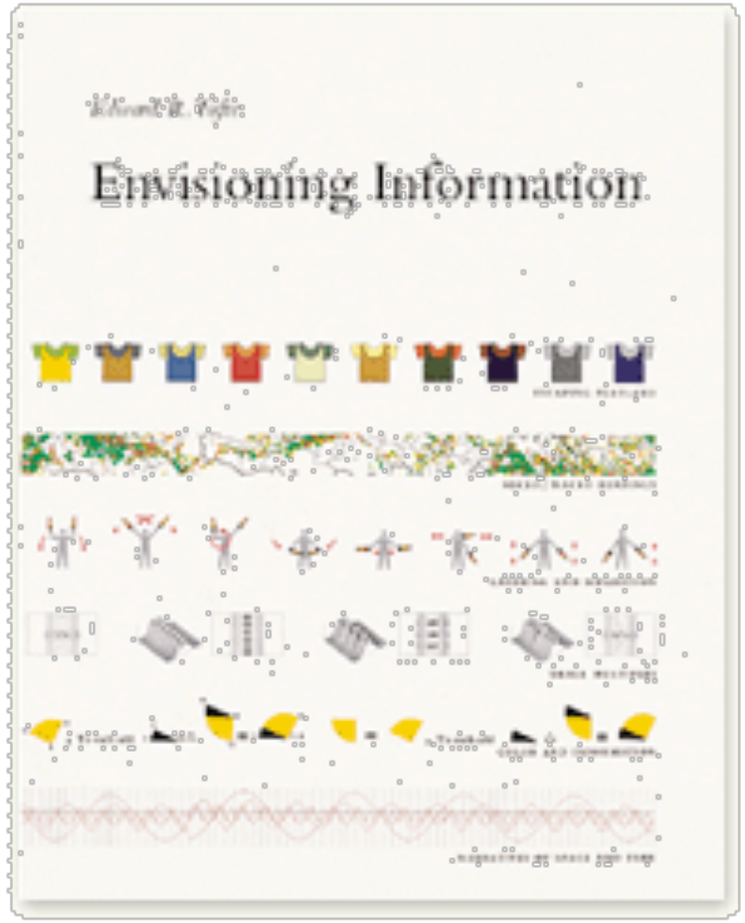
Effective Visualizations

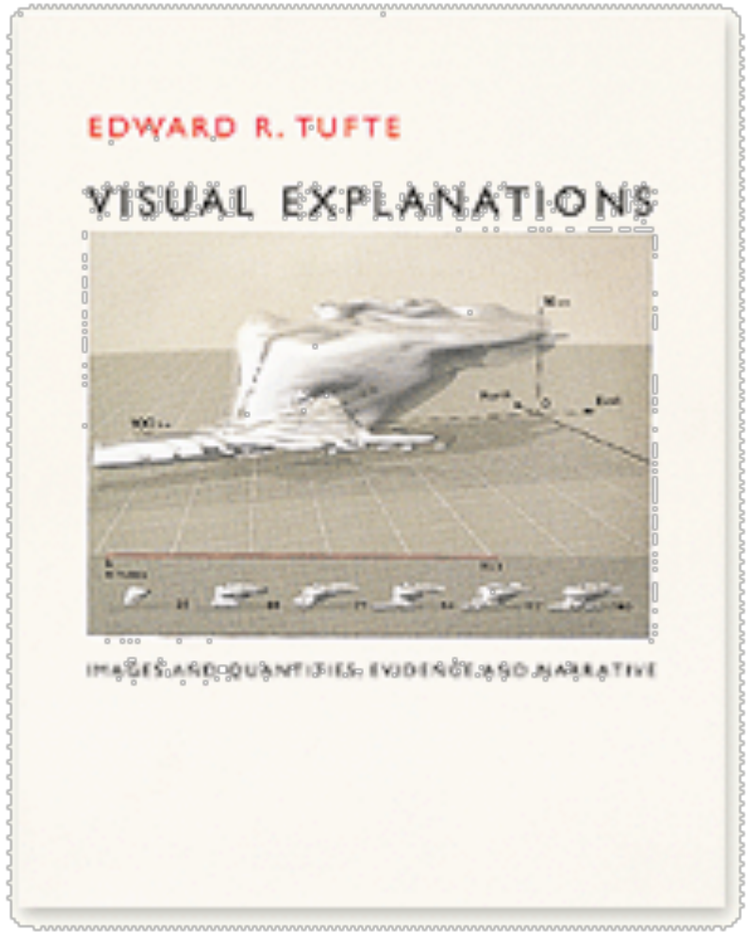
1. Have graphical integrity
2. Keep it simple
3. Use the right display
4. Use color strategically
5. Tell a story with data

Further Reading

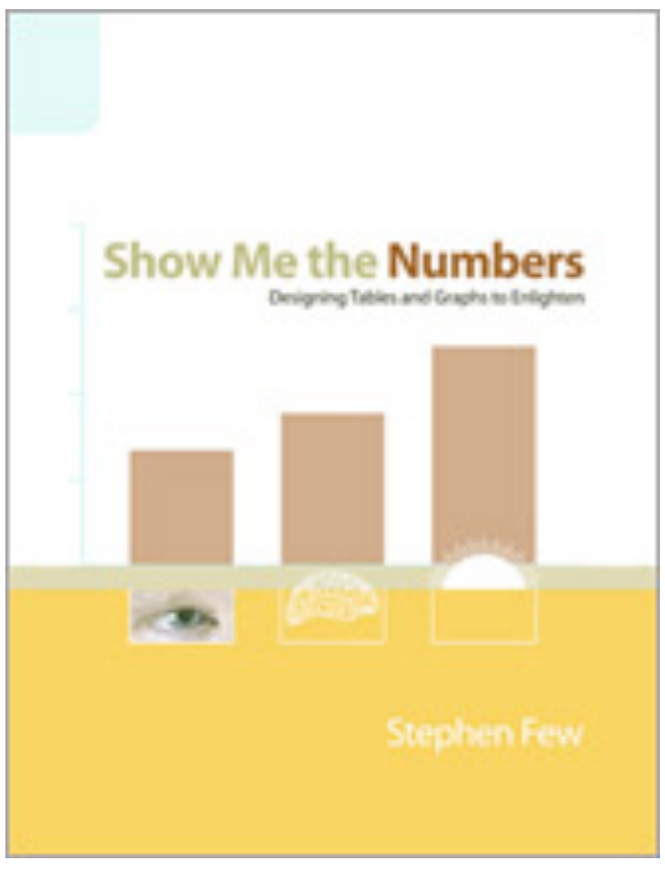
Edward Tufte

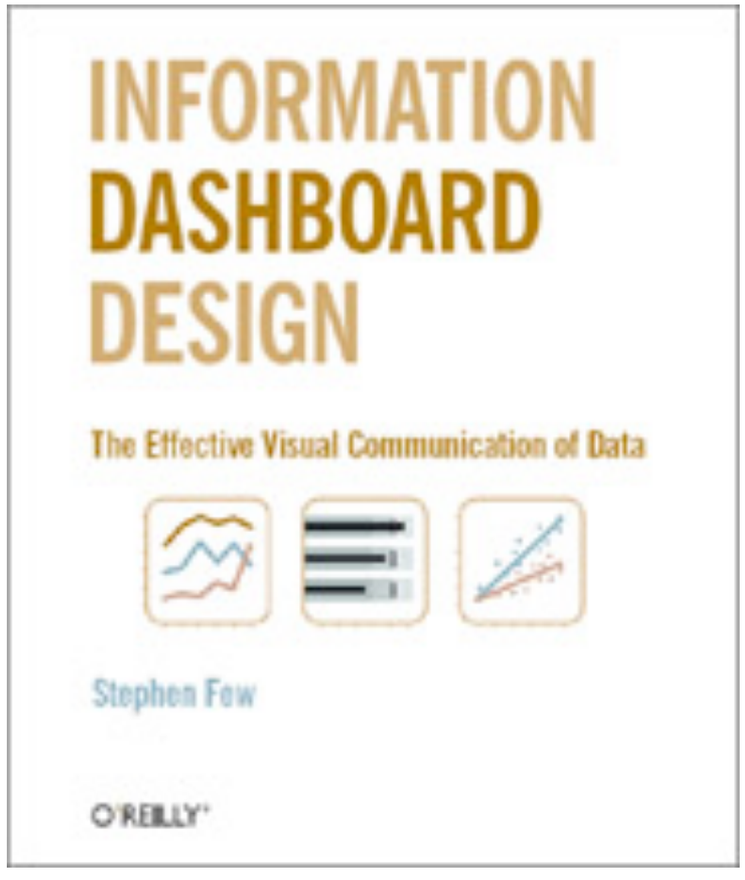


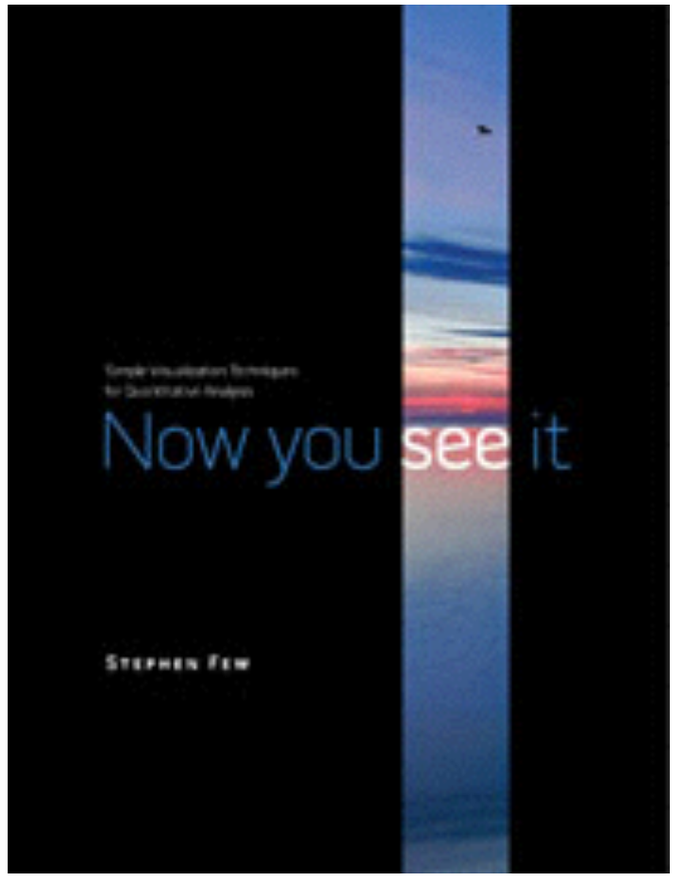




Stephen Few


Show Me the Numbers
Designing Tables and Graphs to Enlighten
Stephen Few


INFORMATION
DASHBOARD
DESIGN
The Effective Visual Communication of Data
Stephen Few
O'REILLY


Now you see it
Stephen Few