

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)
- HWI - 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

IT'S A GAME CHANGER

THE YUKON DENALI OFFERS REFINEMENT INSIDE AND OUT



GMC
WE ARE PROFESSIONAL GRADE

EXPLORE YUKON >

 FiveThirtyEight

≡ MENU



NFL PREVIEW
2015

■ FOOTBALL | 6:

2015 NFL
Brady, Tom
The AFC

By HARRY ENTEN

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

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

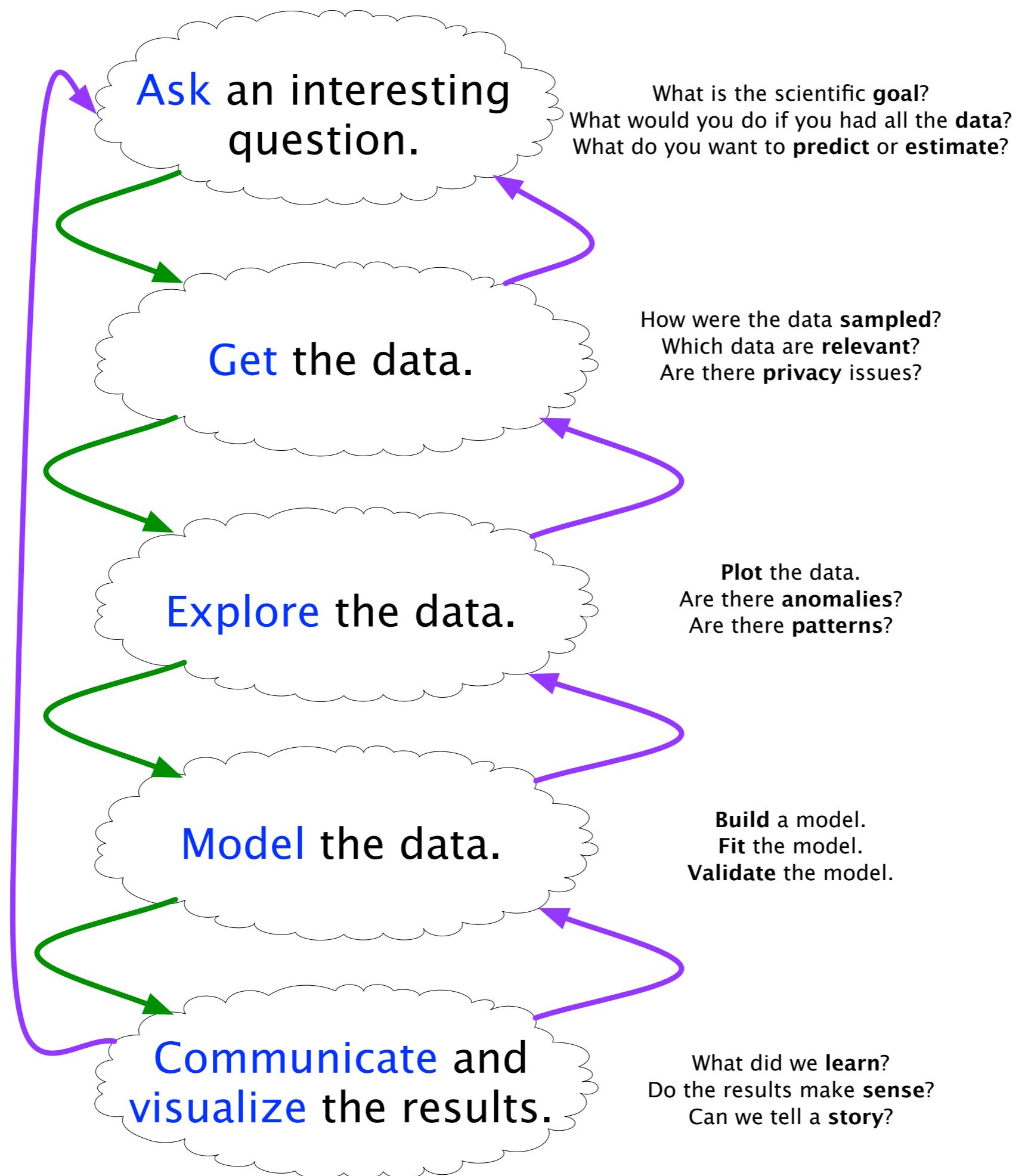


SPORTS

sement Primary



FiveThirtyEight Blog



Data Exploration

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



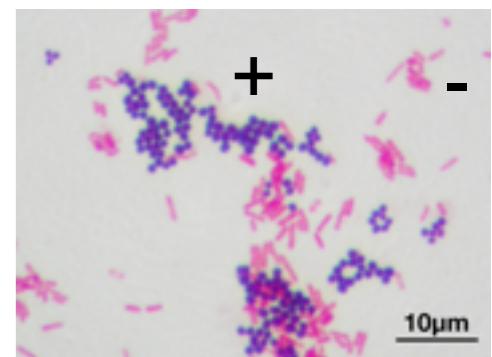
Example: Antibiotics
Will Burtin, 1951

Data

Genus, Species

Table 1: Burtin's data.

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



What Questions?

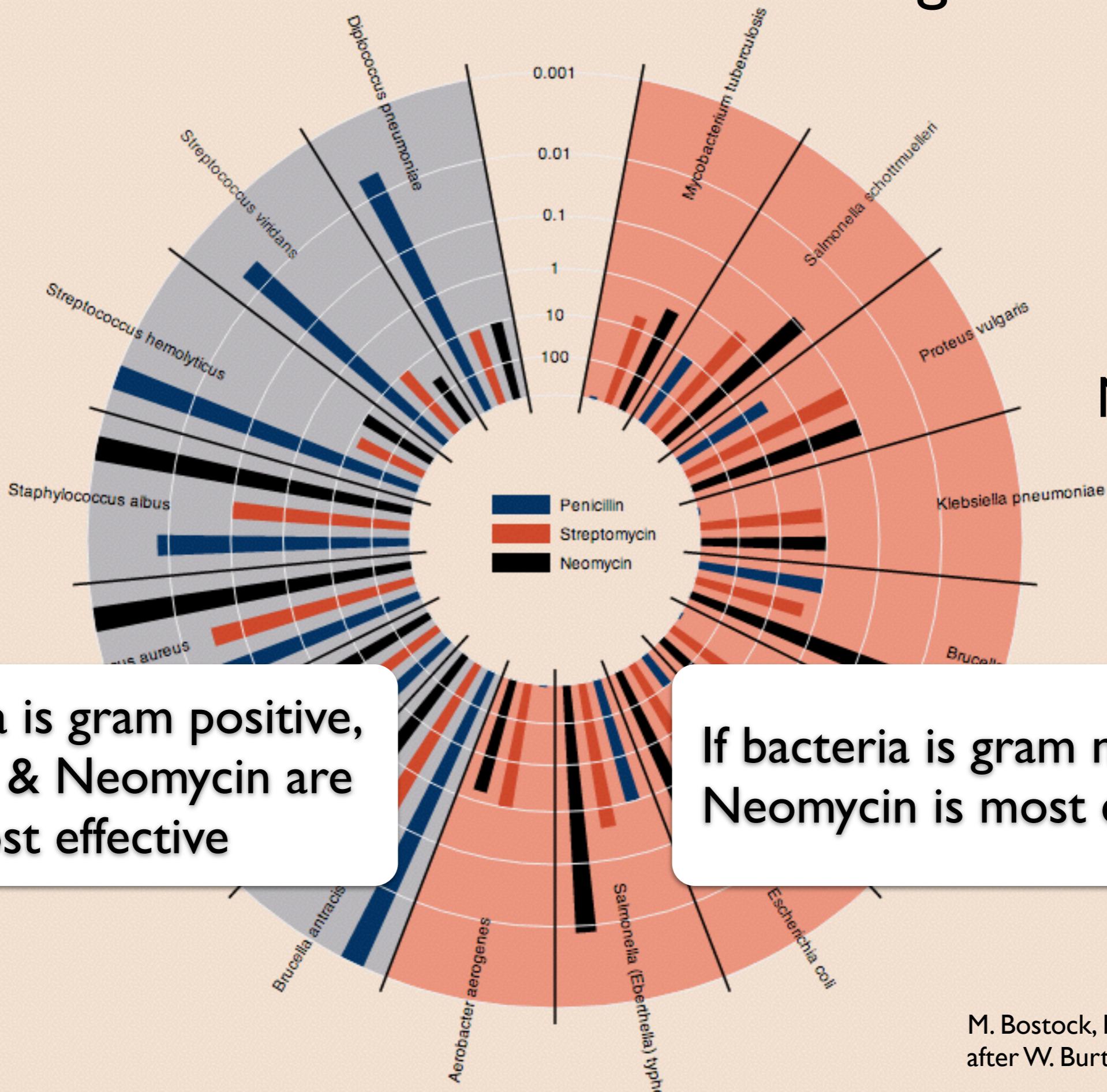
Table 1: Burtin's data.

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

How effective are the drugs?

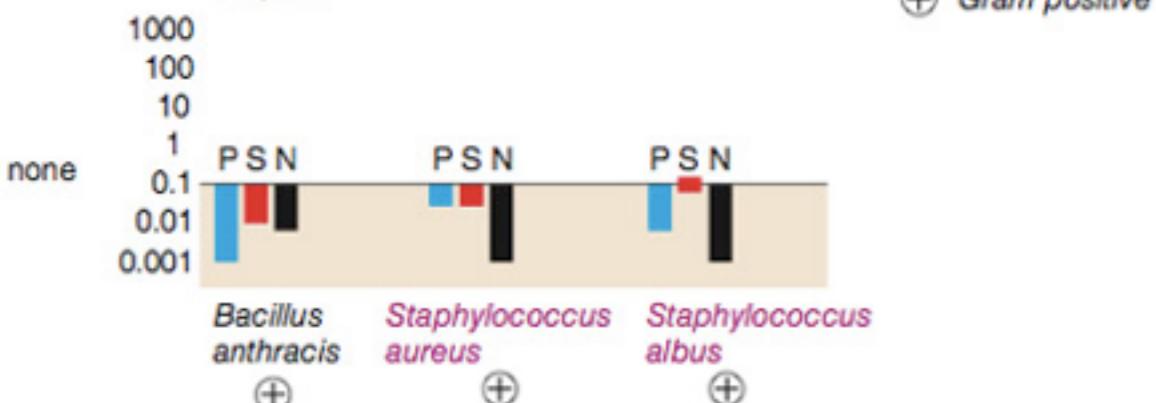
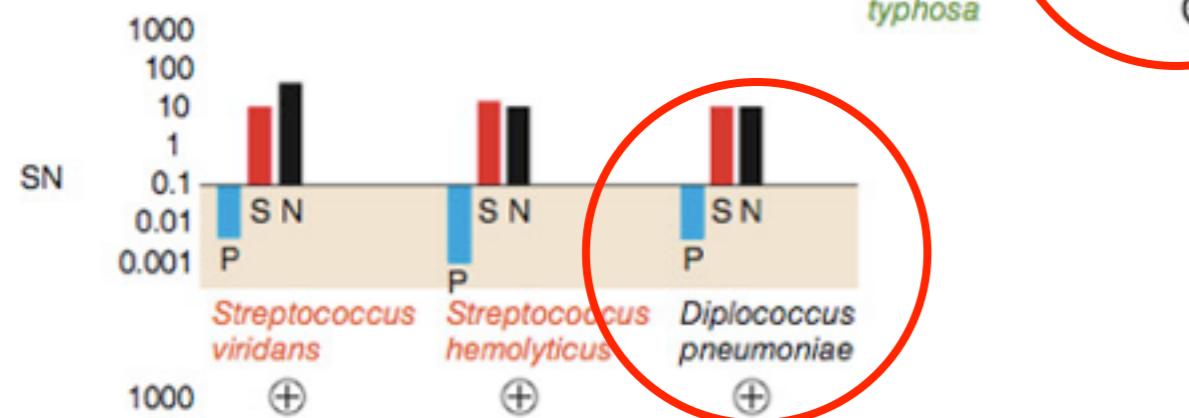
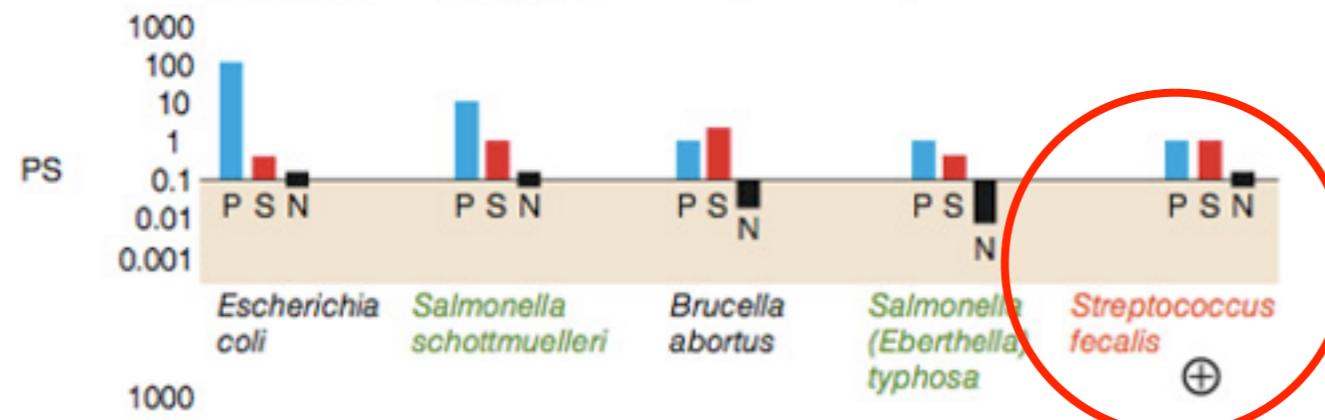
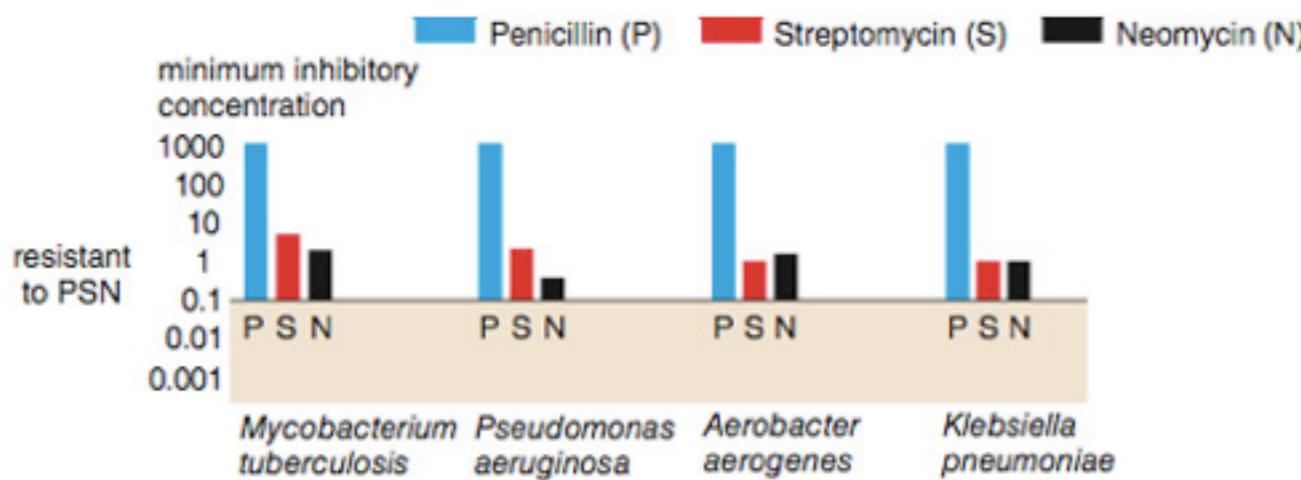
Gram
Positive

Gram
Negative



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

If bacteria is gram negative,
Neomycin is most effective



How do the bacteria compare?

Not a streptococcus!
(realized ~30 years later)

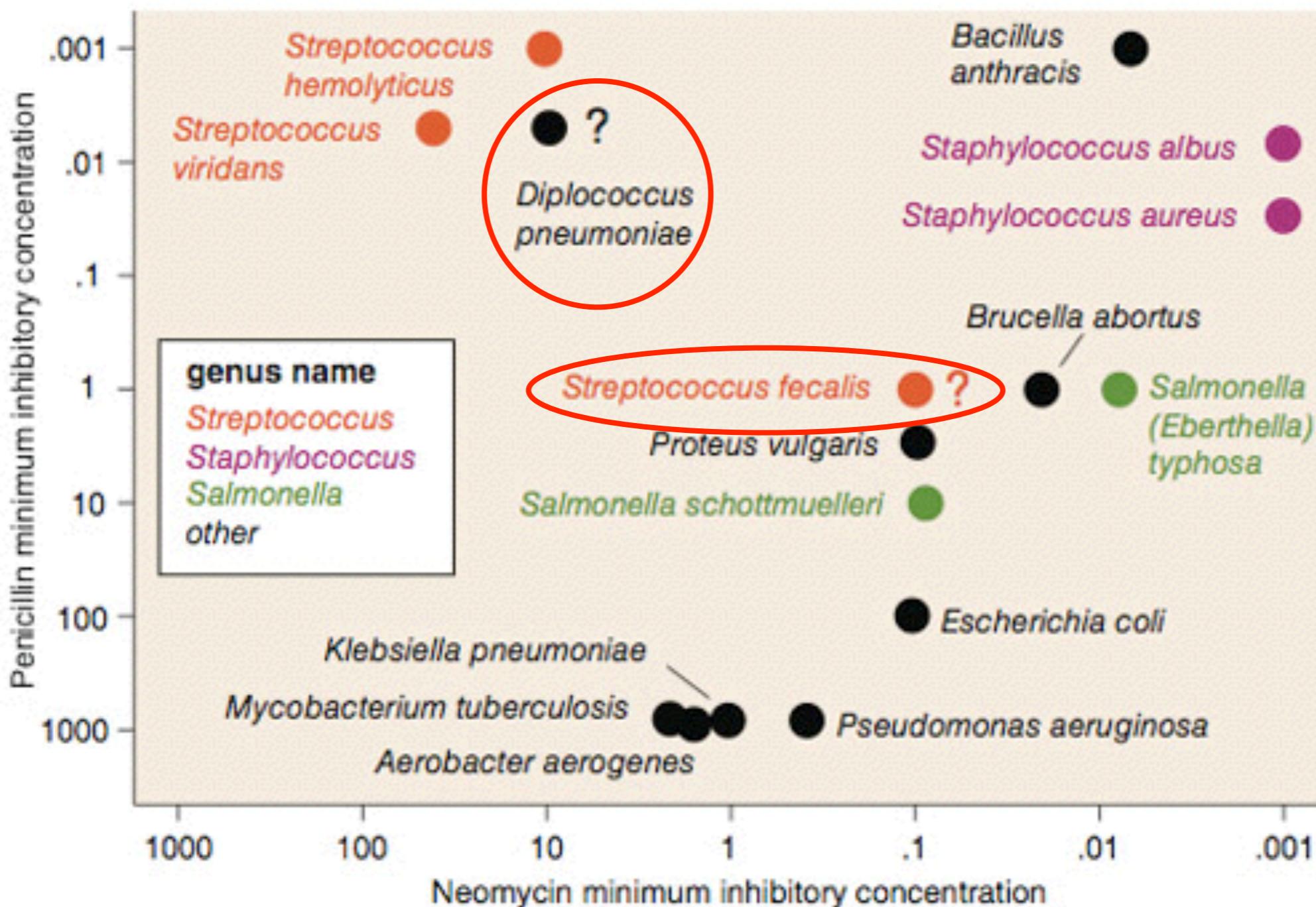
Really a streptococcus!
(realized ~20 years later)

Streptococcus
Staphylococcus
Salmonella
Other

⊕ Gram positive

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

How do the bacteria compare?



Exploratory Data Analysis

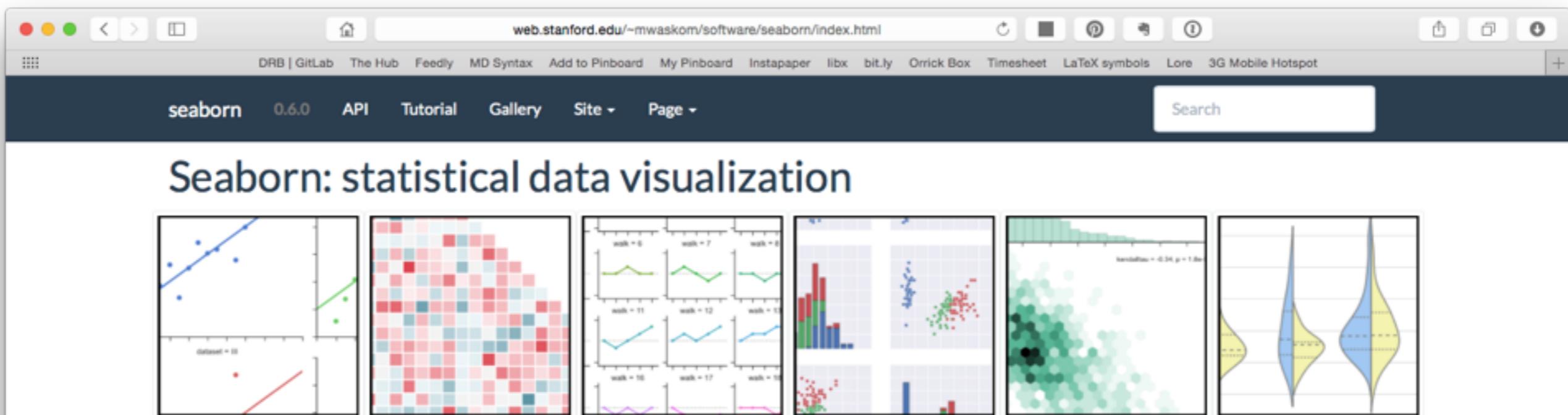
“The greatest value of a picture is when it forces us to notice what we never expected to see.”



John Tukey

Visualization

To convey information through graphical representations of data



The screenshot shows a web browser displaying the official Seaborn library website at web.stanford.edu/~mwaskom/software/seaborn/index.html. The page title is "Seaborn: statistical data visualization". Below the title, there is a grid of six small plots demonstrating various features of the library:

- A scatter plot with a regression line.
- A heatmap with a color gradient.
- A facet grid of line plots showing trends over time.
- A faceted scatter plot with multiple data series.
- A density plot with marginal histograms.
- A violin plot showing the distribution of data.

Below the grid, there is a brief introduction to Seaborn and links to documentation and features.

Seaborn is a Python visualization library based on `matplotlib`. It provides a high-level interface for drawing attractive statistical graphics.

For a brief introduction to the ideas behind the package, you can read the [introductory notes](#). More practical information is on the [installation page](#). You may also want to browse the [example gallery](#) to get a sense for what you can do with Seaborn and then check out the [tutorial](#) and [API reference](#) to find out how.

To see the code or report a bug, please visit the [github repository](#). General support issues are most at home on [stackoverflow](#), where there is a `seaborn` tag.

Documentation

- [An introduction to seaborn](#)
- [What's new in the package](#)
- [Installing and getting started](#)
- [Example gallery](#)
- [API reference](#)
- [Seaborn tutorial](#)

Features

- Style functions: [API](#) | [Tutorial](#)
- Color palettes: [API](#) | [Tutorial](#)
- Distribution plots: [API](#) | [Tutorial](#)
- Regression plots: [API](#) | [Tutorial](#)
- Categorical plots: [API](#) | [Tutorial](#)
- Axis grid objects: [API](#) | [Tutorial](#)

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

755



Steroids or Not, the Pursuit Is On

Babe Ruth is taking aim at the career home run record. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.

Lines are cumulative home runs

Hank Aaron
755 homers
23 seasons



Babe Ruth
714 homers
22 seasons



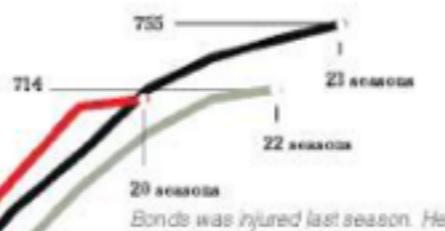
Barry Bonds
708 homers
20 seasons

Bonds takes lead
Home runs
after 16 seasons
Bonds 567
Aaron 554
Ruth 516



14th season

According to allegations
in a book about Bonds,
he began taking
steroids before the 1999
season, his 14th in the
league. Two seasons
later, he hit 73 home
runs, surpassing Aaron's
career pace.



Homer Pace After Age 34

If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.

PROJECTED PACE BASED ON
AVERAGE OF PREVIOUS FIVE SEASONS

Aaron

Actual homers
slightly
outpace
projected
homers for five
seasons.

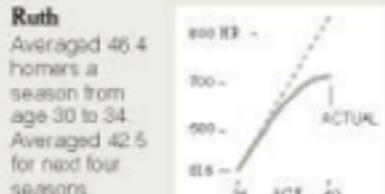
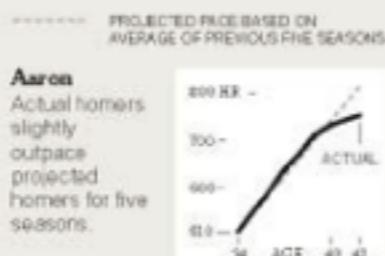
Ruth

Averaged 46.4
homers a
season from
age 30 to 34.
Averaged 42.5
for next four
seasons.

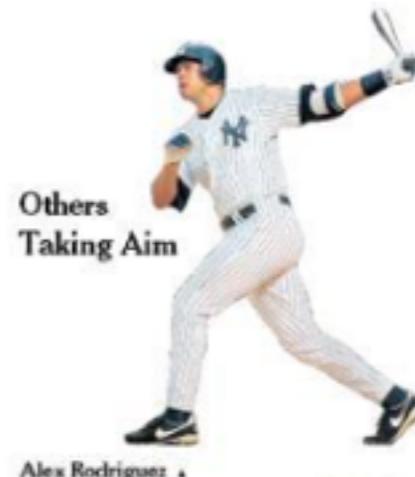
Bonds

From age 35
to 39, he
averaged 14
more homers
a season than
projected.

Note: Ages as of July 1 of each season.



Others Taking Aim



Alex Rodriguez

Is ahead of
the pace set
by all three
home run
leaders.

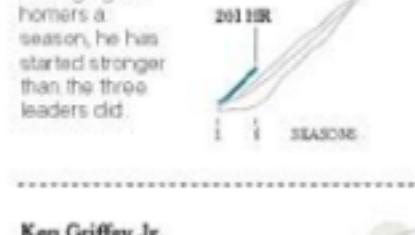
Albert Pujols

Averaging 40
homers a
season, he has
started stronger
than the three
leaders did.

Ken Griffey Jr.

Many thought he
would be the first
to catch Ruth
and Aaron until
injuries limited
his output.

SEASONS



SEASONS

SEASONS

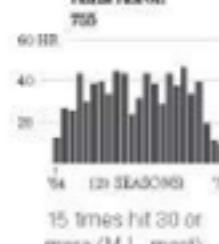
SEASONS

SEASONS

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)

Hank Aaron
755



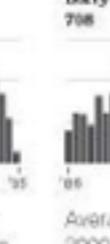
15 times hit 30 or
more (M.L. most)

Babe Ruth
714



Hit only 20 over
first five seasons.

Barry Bonds
708



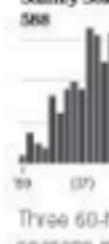
Averaged 52 from
2000 to 2004.

Willie Mays
660



No one hit more
from 1950-69.

Sammy Sosa
588



Three 60-homer
seasons is record

Frank Robinson
586



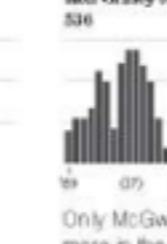
Triple Crown in '66
(49, 122, 316)

Mark McGwire
583



First to hit 70 in
a season.

Ken Griffey Jr.
536



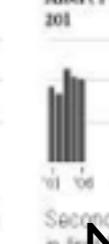
Only McGwire had
more in the 90's.

Alex Rodriguez
429



Youngest to reach
400 homers.

Albert Pujols
261



Second most ever
in first 5 years.

Atlanta Cor and Joe Ward/The New York Times

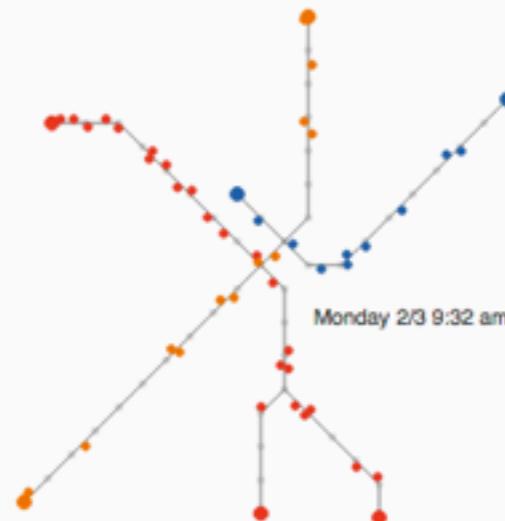
New York Times

Explore

Visualizing MBTA Data

An interactive exploration of Boston's subway system

Mike Barry and Brian Card - June 10, 2014



[Share](#) 37 [Tweet](#) 29 [G+](#) 4 [Share](#)

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.

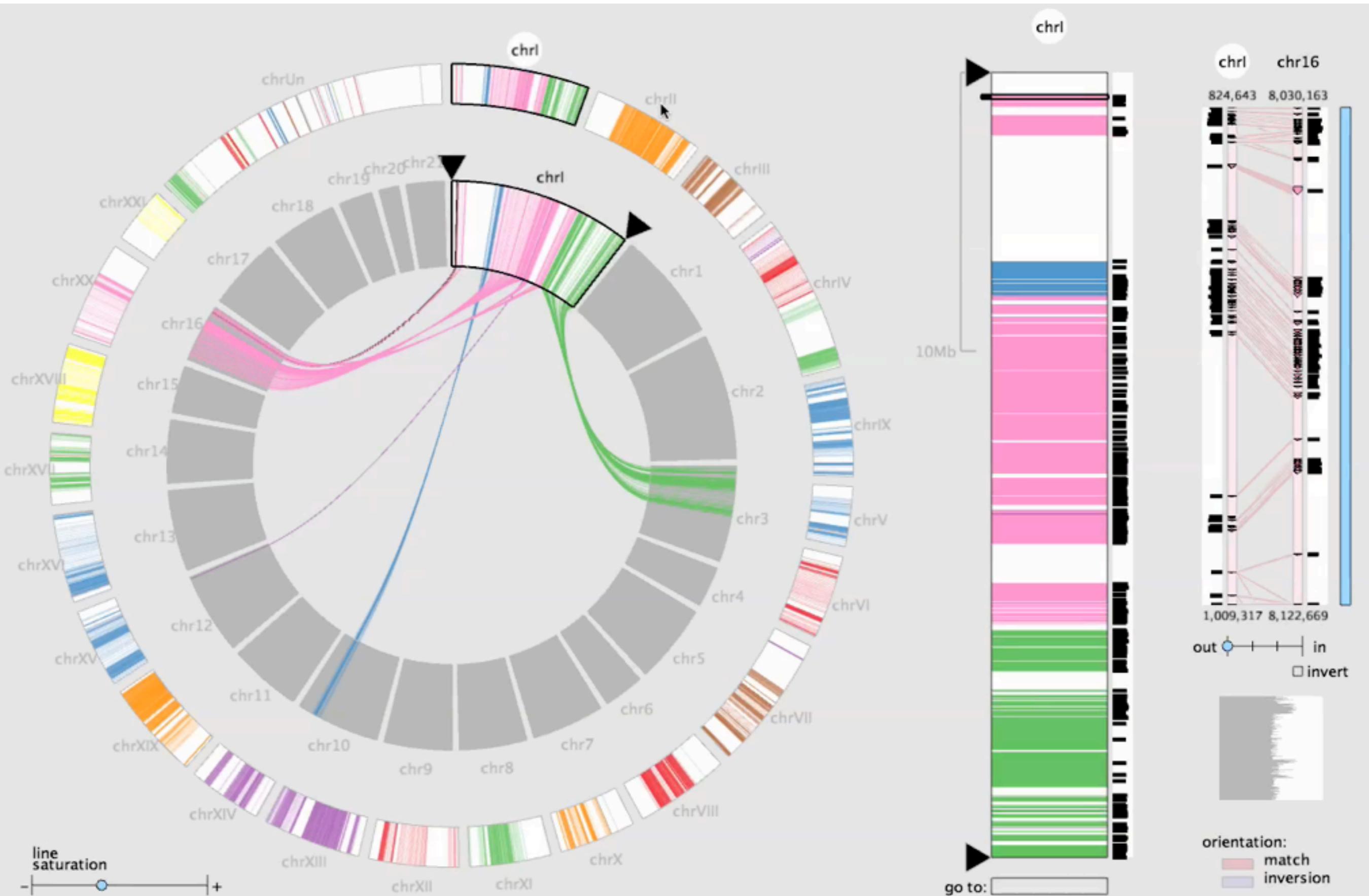
To better understand how the trains operate on a typical day, below are all trips that trains took on the red, orange, and blue lines on Monday February 3 2014. Each vertical line represents a station, and time extends from top to bottom. Steeper lines indicate slower trains. This visualization was first 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

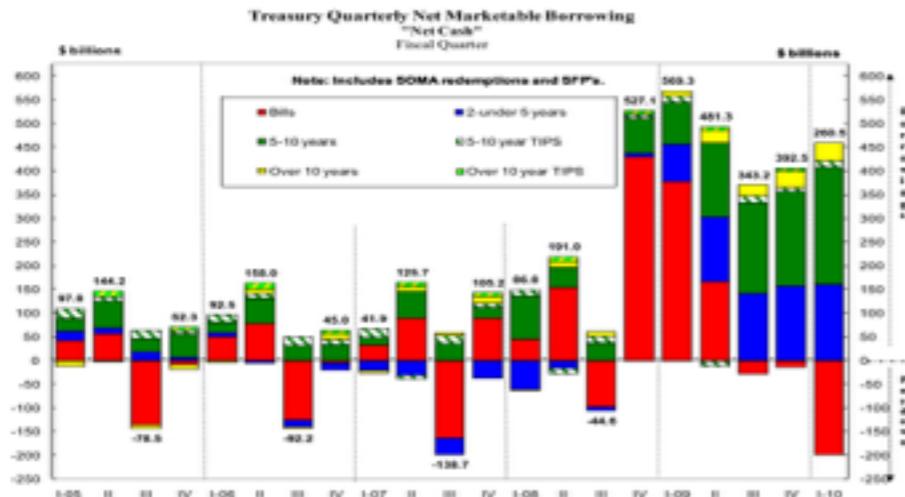
MizBee

[Meyer et al. 2009]



Effective Visualizations

Not Effective...



<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](#).

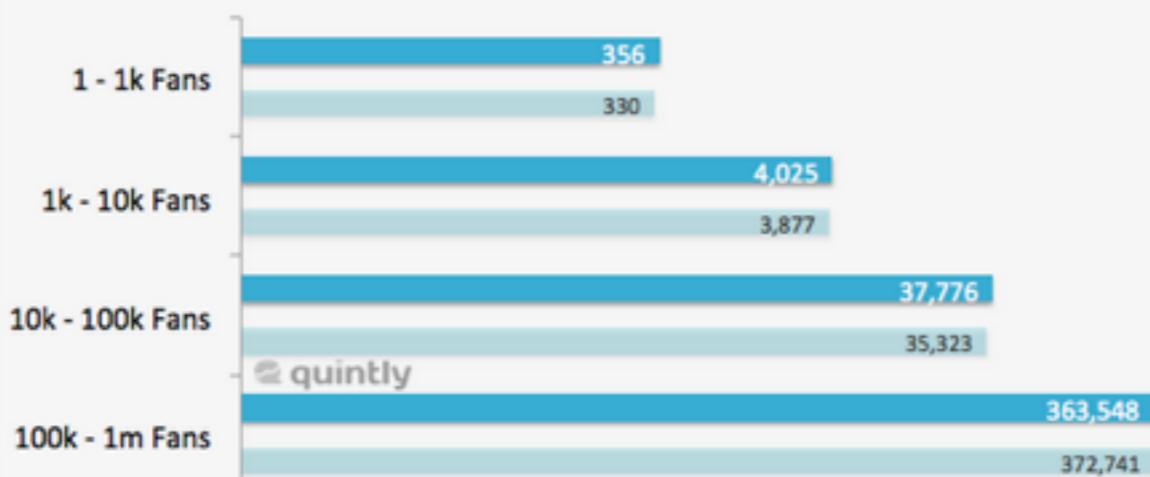
Check out our friends [Thumbs Up Viz](#) and [accidental aRt](#), or [submit](#).



Average Number Of Facebook Fans

The total number of fans is still one of the most important metrics for Facebook marketers.

Here you can see if your total number of fans is above the average.

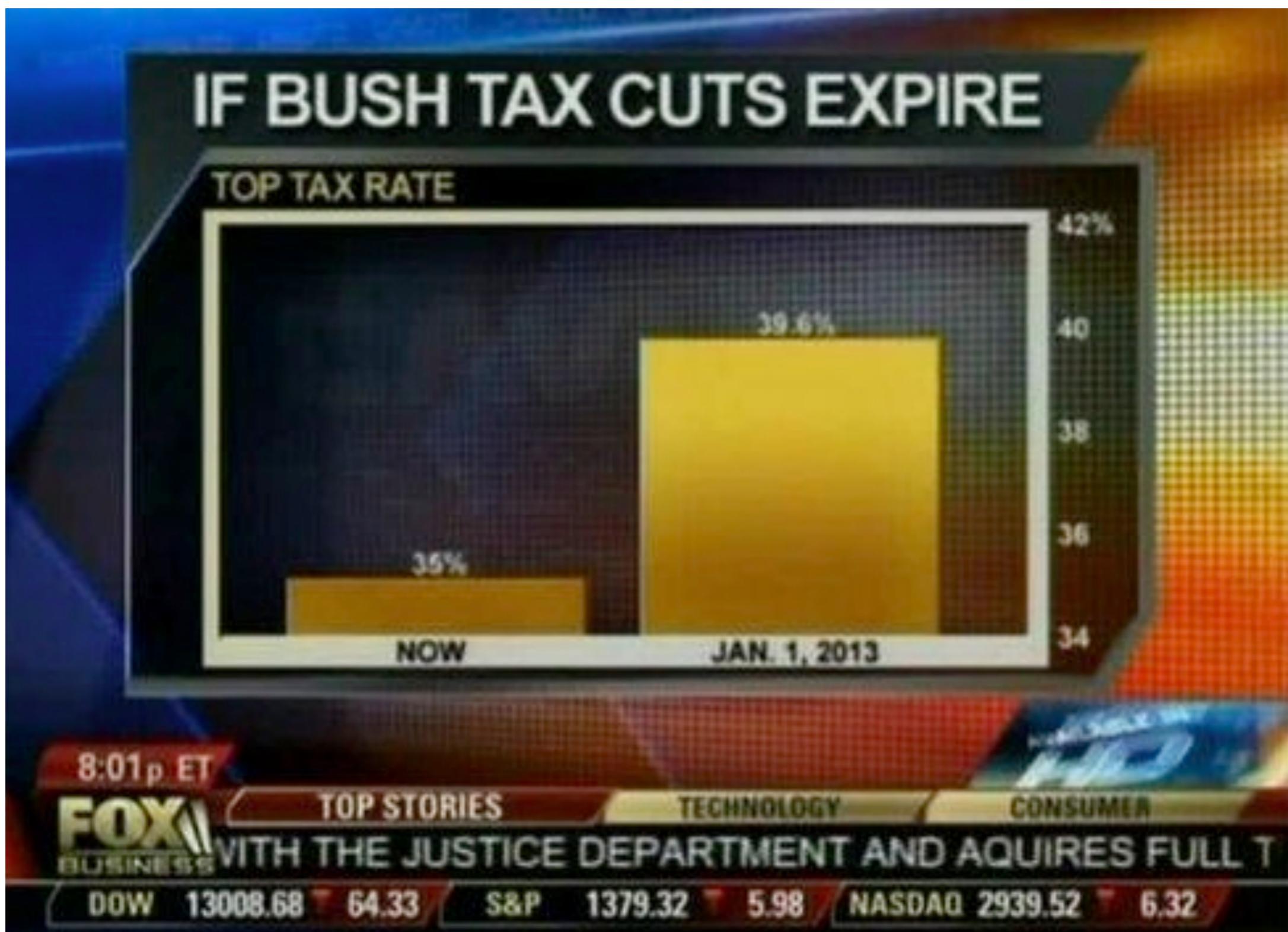


Effective Visualizations

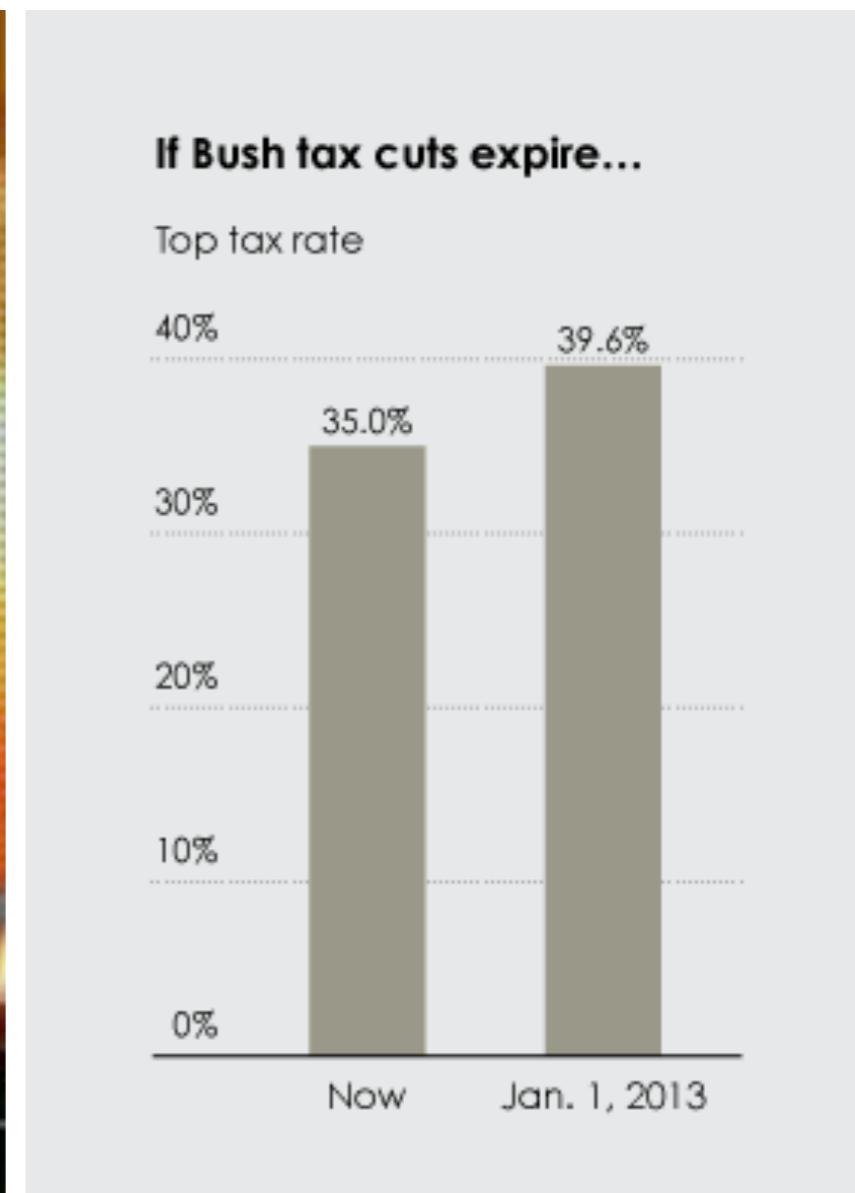
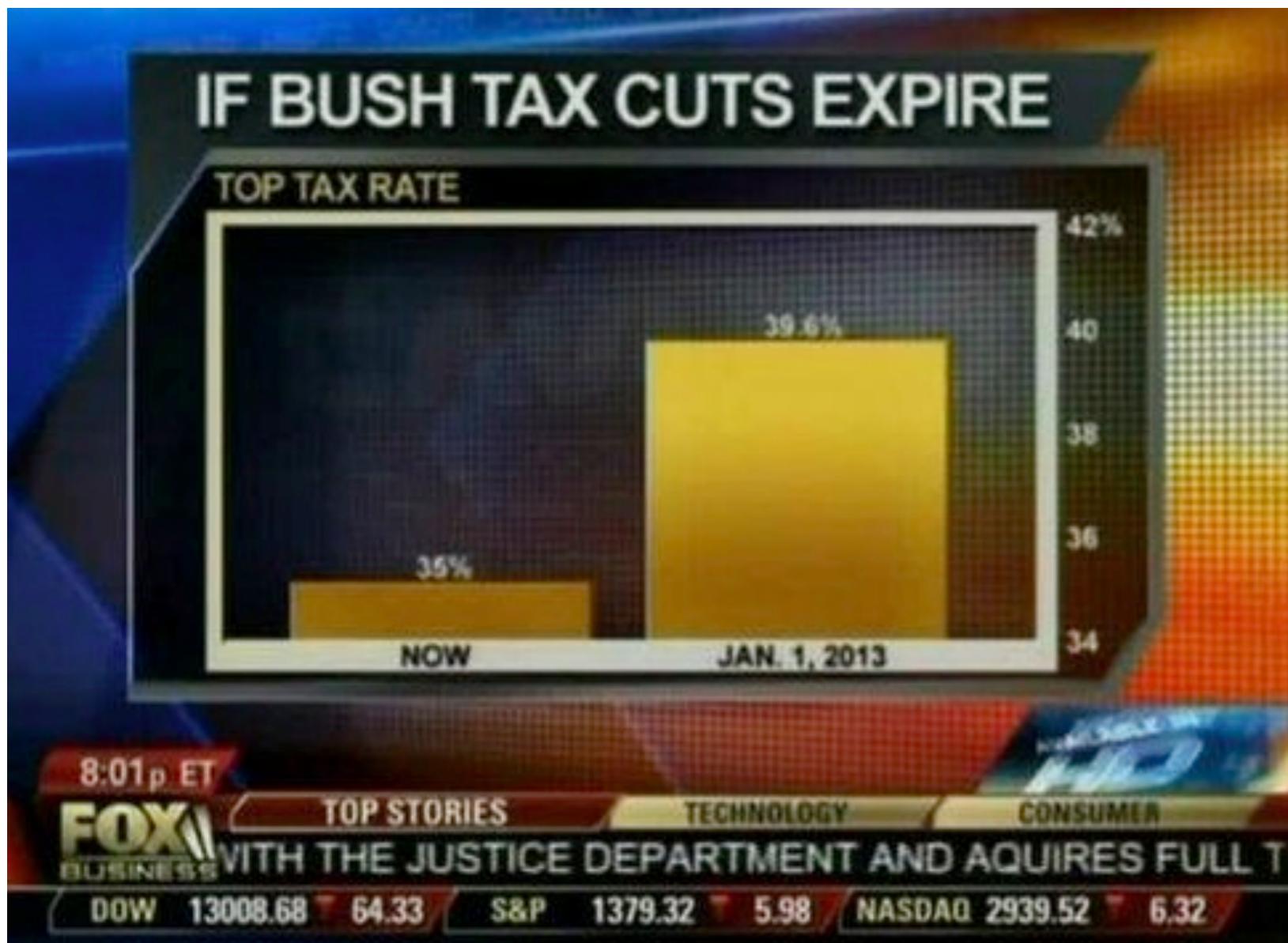
1. Have graphical integrity
2. Keep it simple
3. Use the right display
4. Use color strategically
5. Tell a story with data

Graphical Integrity

Graphical Integrity



Scale Distortions



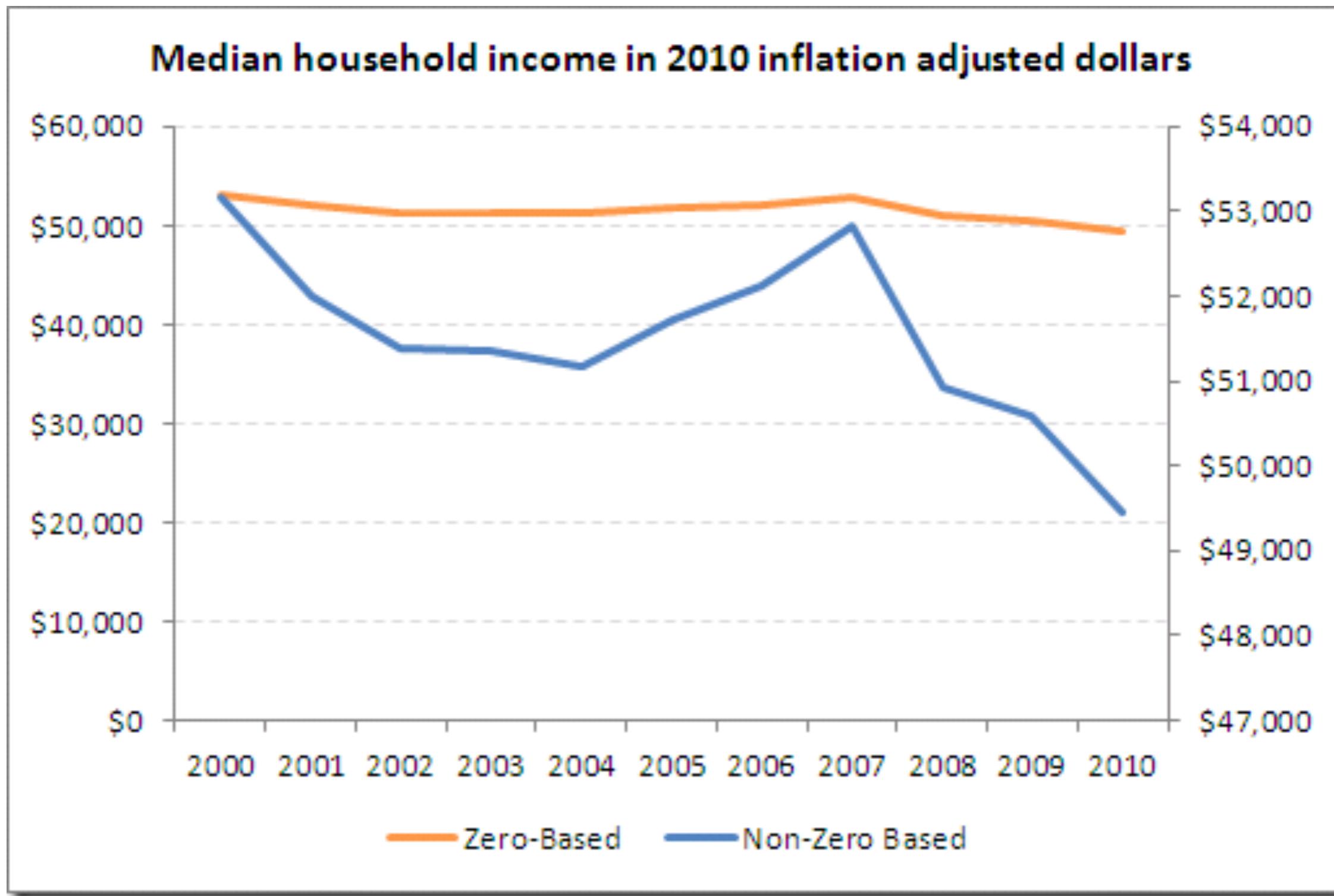
JOB LOSS BY QUARTER



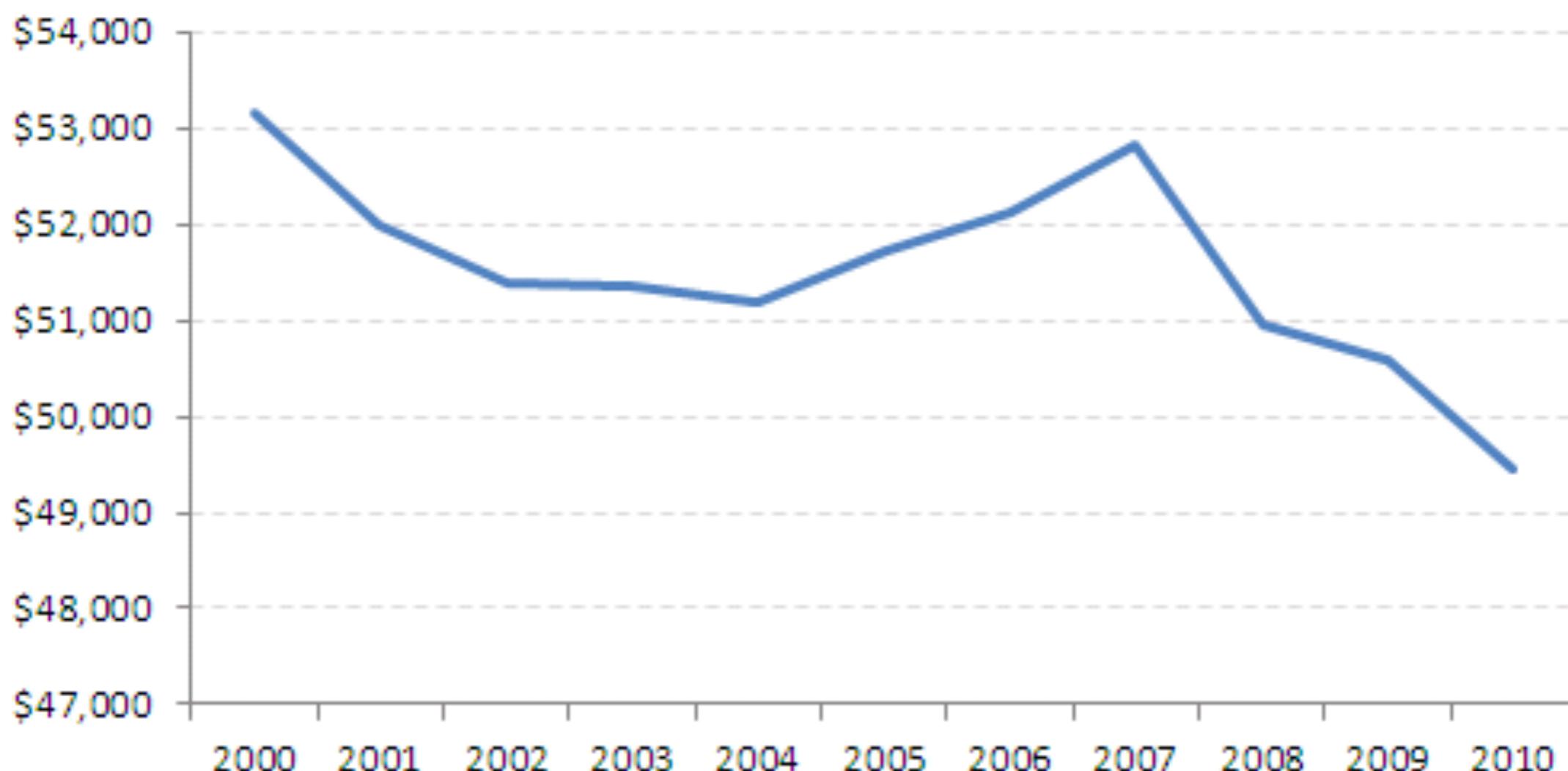
Scale Distortions



Scale Distortions



Median household income in 2010 inflation adjusted dollars



Attention: The dollar scale along the vertical axis is narrow to reveal the subtle, yet consistent declines in median household income since 2007.

*Same Veritas. More Lux.*

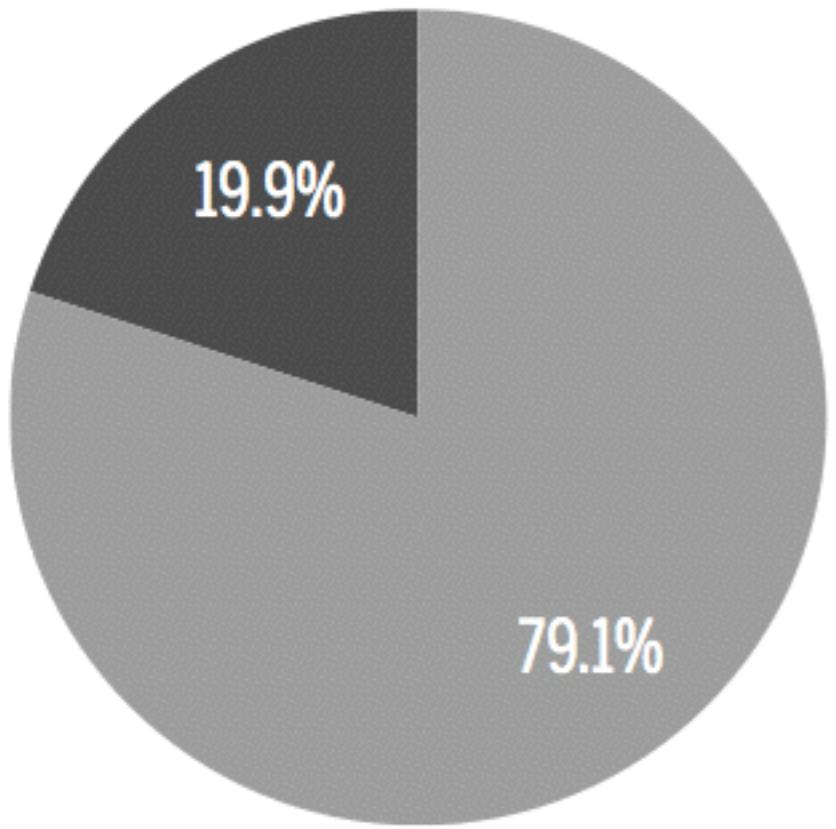
Yale Summer Session

Over 200 full-credit courses.

June 4 – July 6 , July 9 – Aug 10

2012 *experience Yale*

CHART YALE GRADUATES' MAJORS, CLASS OF 2011

Science, technology, engineering
and math degrees

Non-STEM degrees

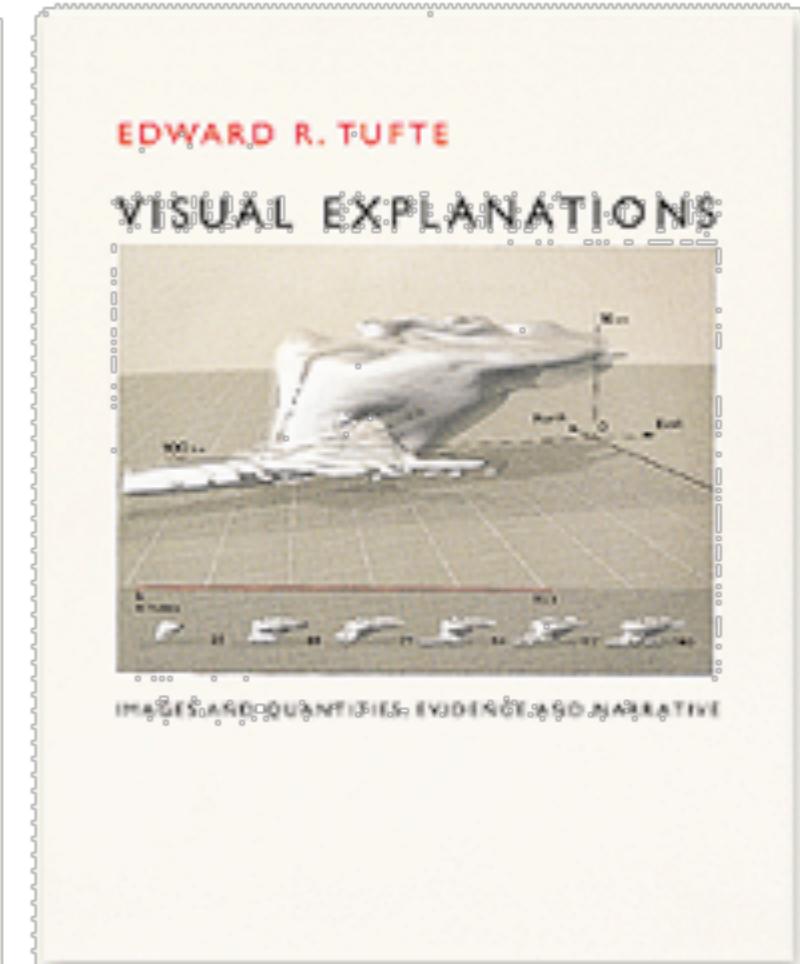
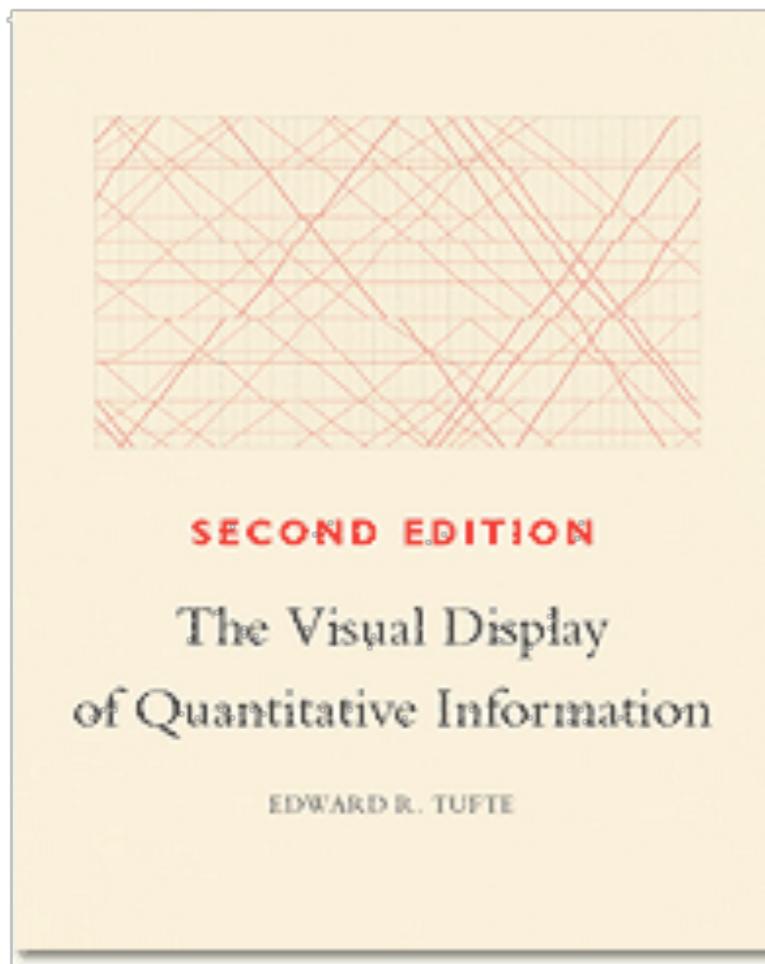
Facebook Recommendations

[Shake Shack to open in New Haven](#)
277 people recommend this.[Popular anti-religion creates false dichotomy](#)
15 people recommend this.[Friends remember Foucher LAW '14](#)
10 people recommend this.[AIDS activist speaks about documentary film](#)
8 people recommend this.[Panel outlines changes in hip-hop](#)
30 people recommend this.

Facebook social plugin

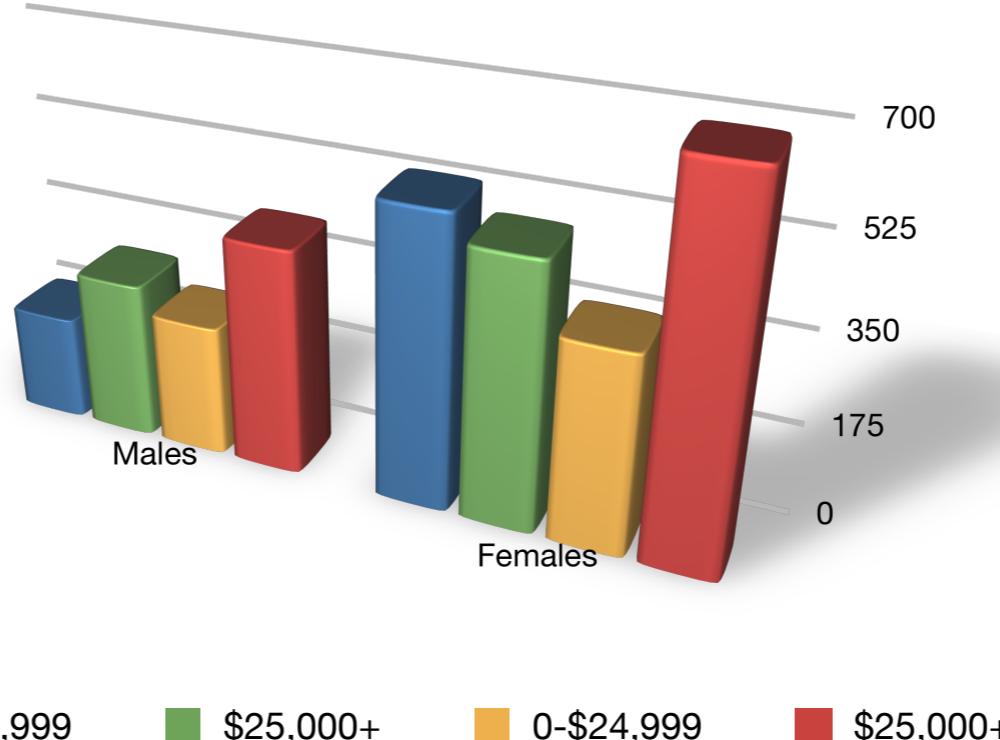
Keep It Simple

Edward Tufte



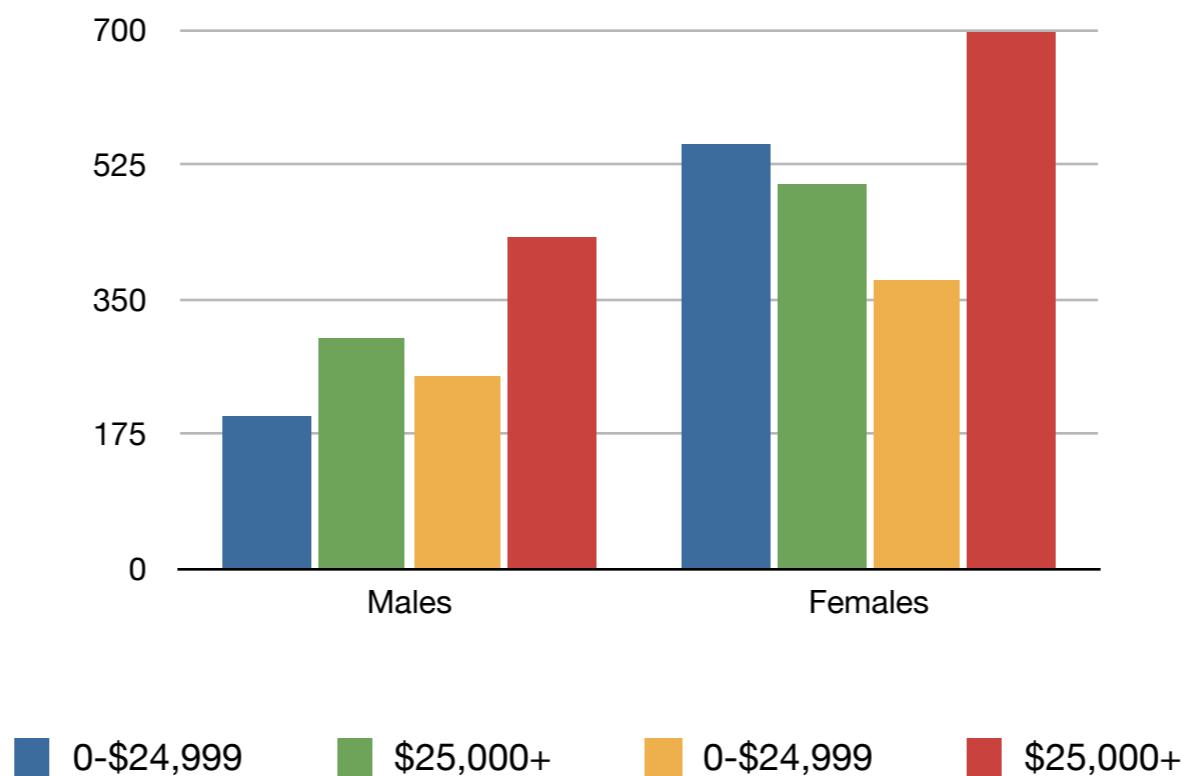
Maximize Data-Ink Ratio

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

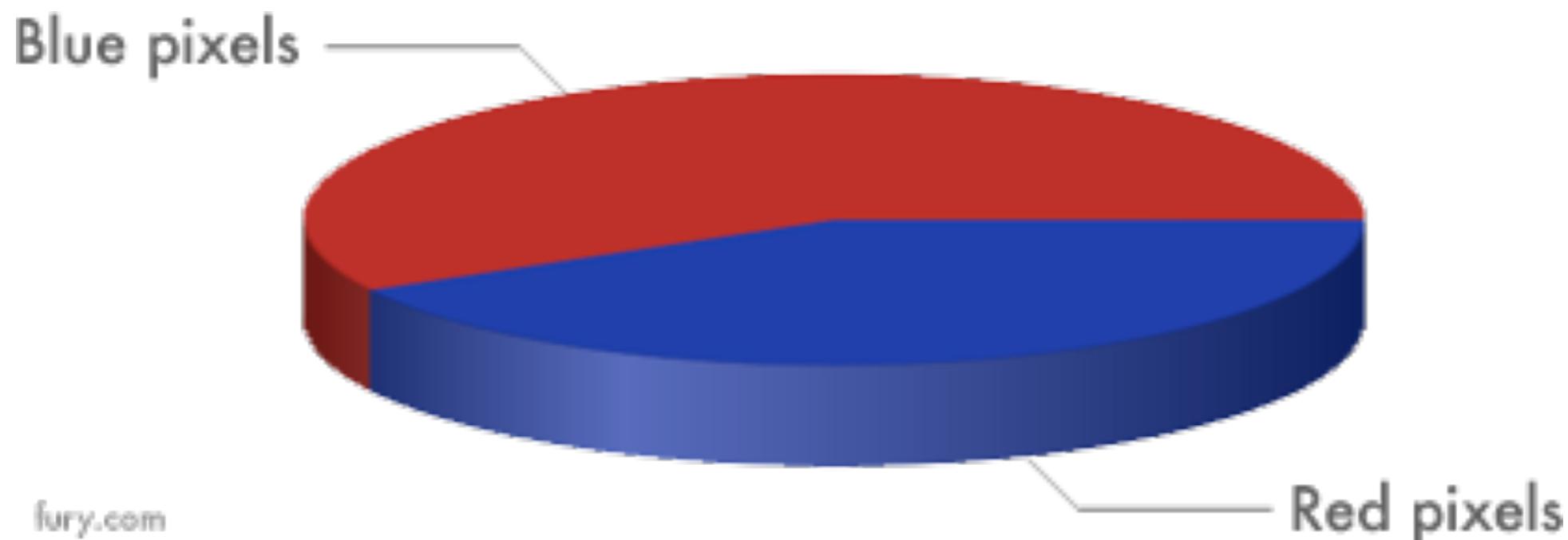


Maximize Data-Ink Ratio

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



Why 3D pie charts are bad

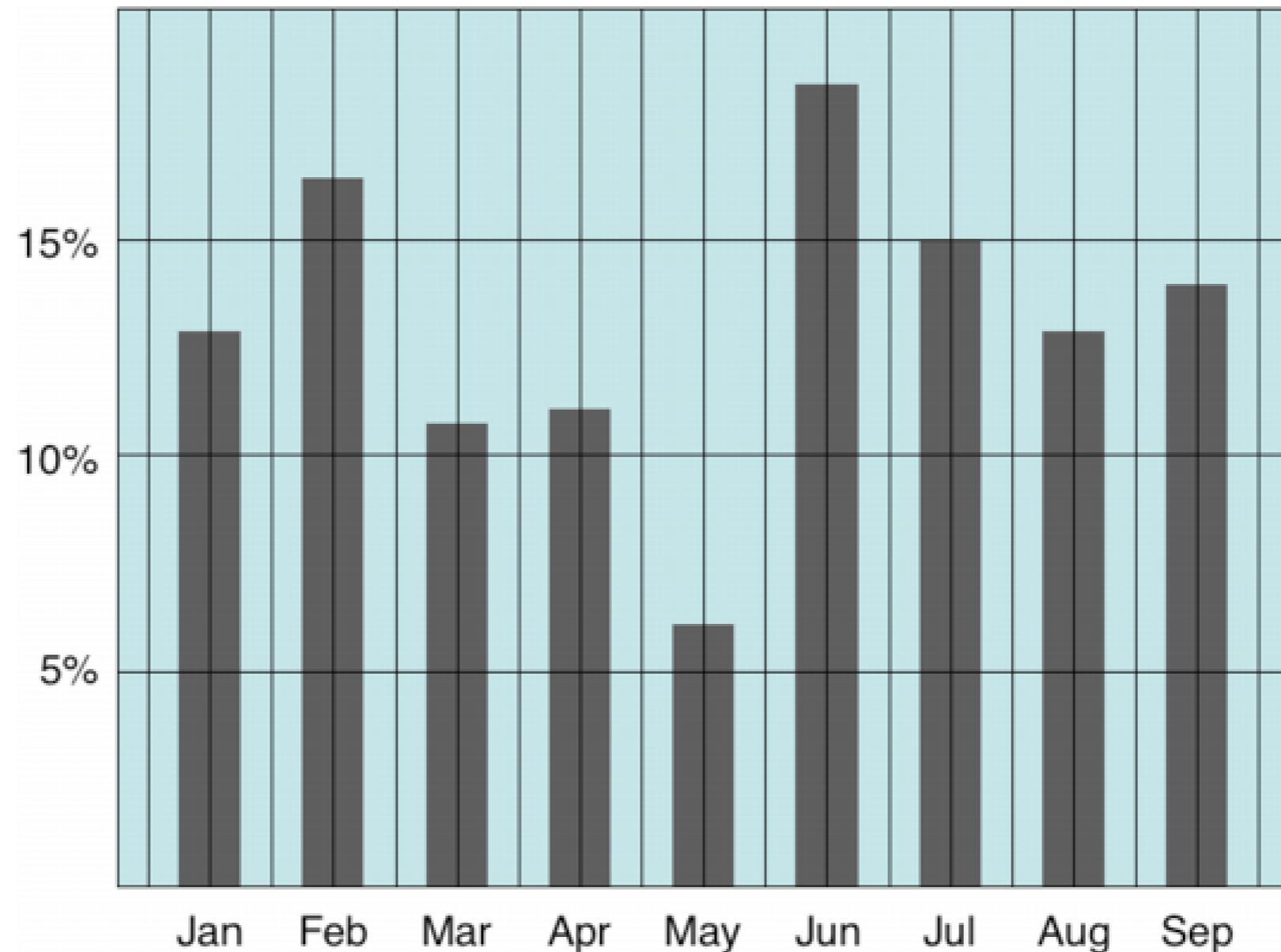


fury.com

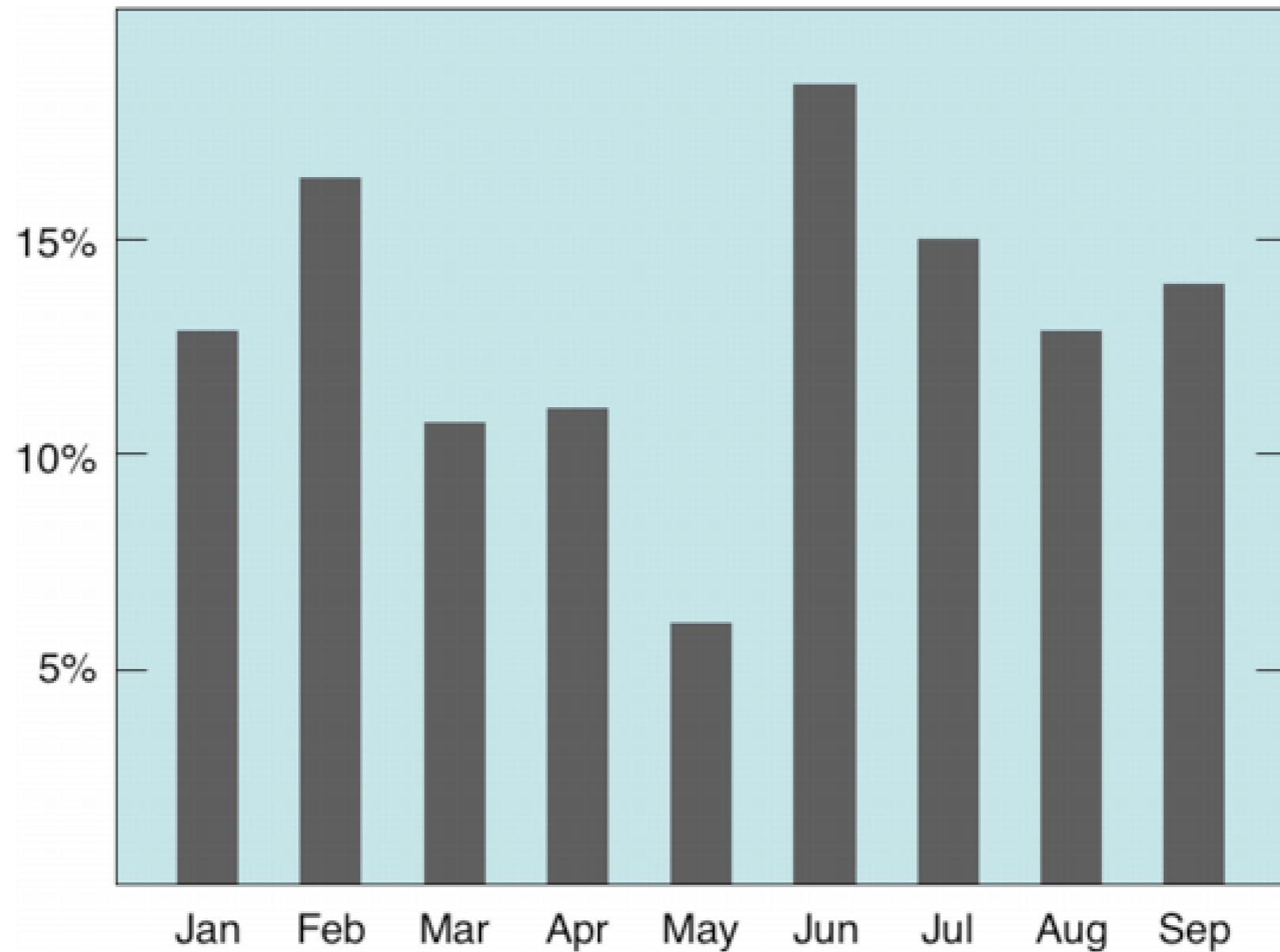
Kevin Fox

Avoid Chartjunk

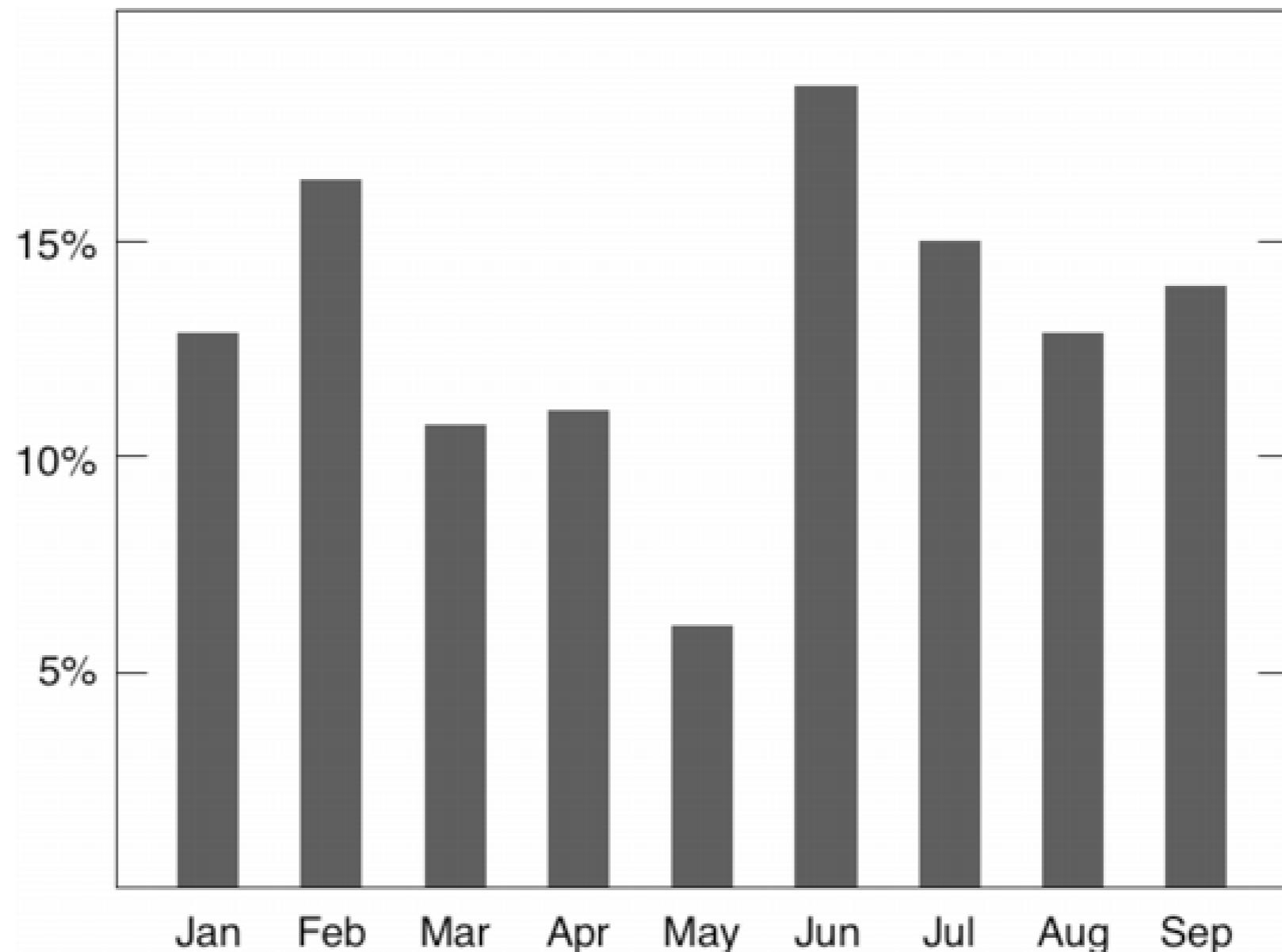
Extraneous visual elements that distract from the message



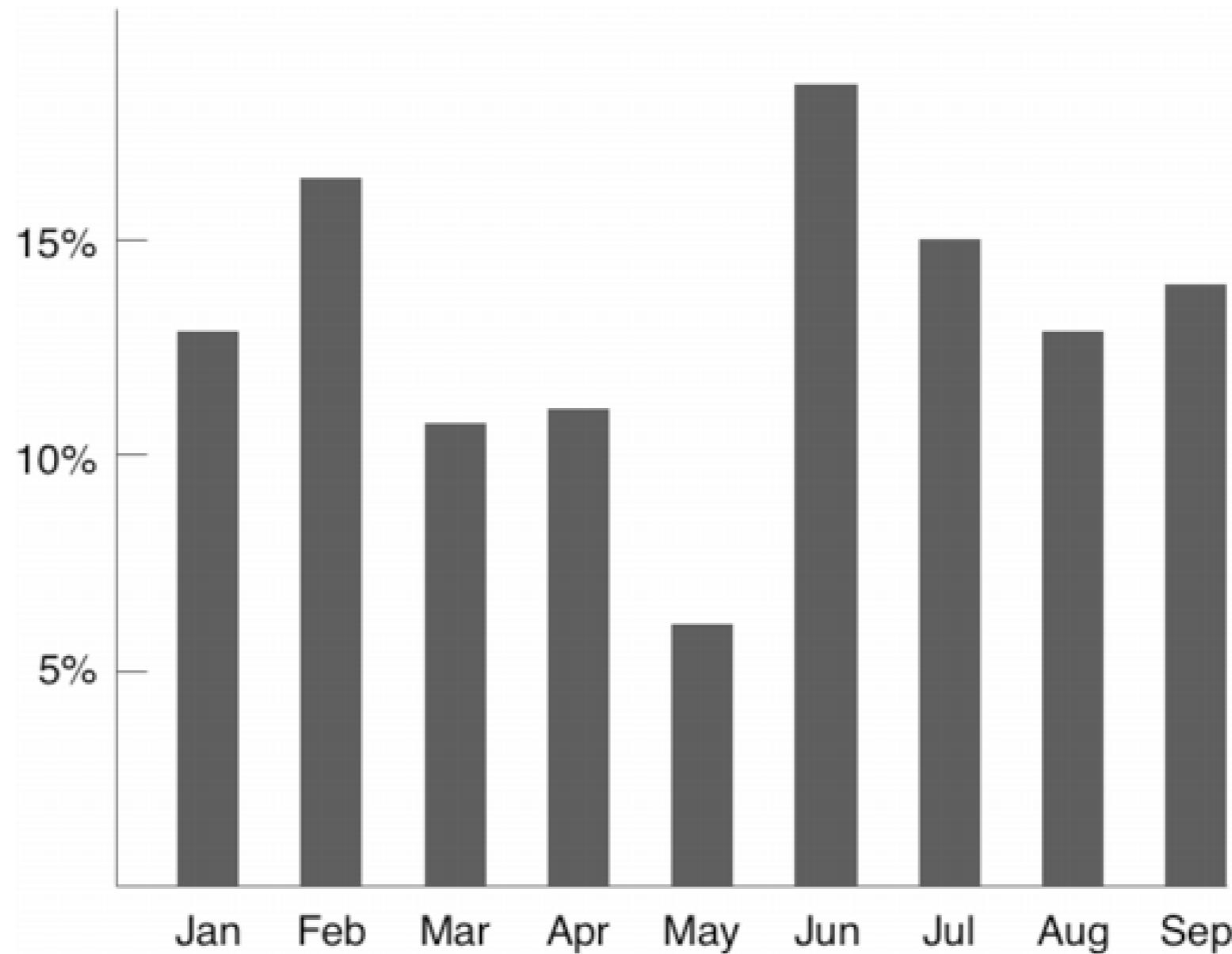
Avoid Chartjunk



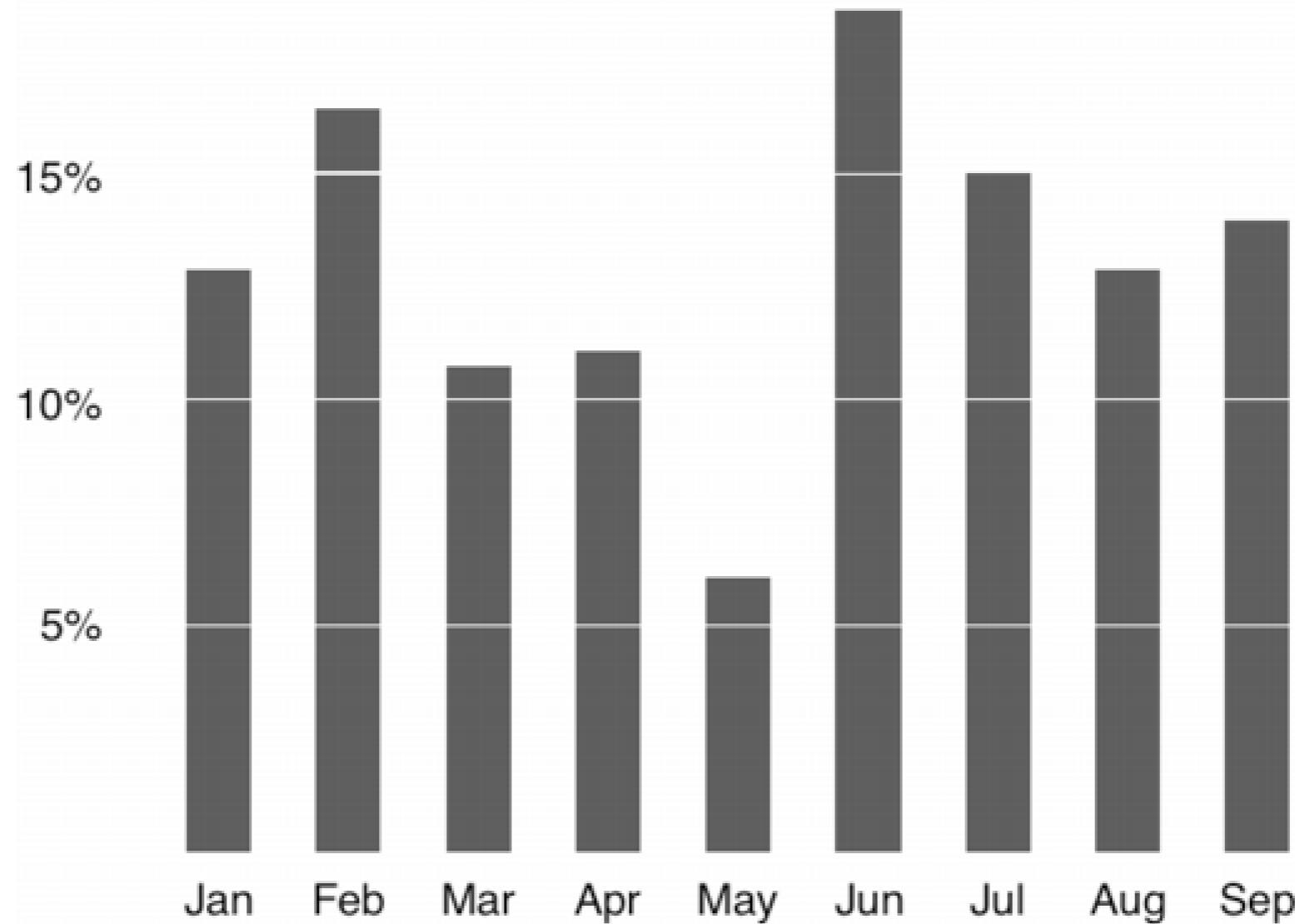
Avoid Chartjunk



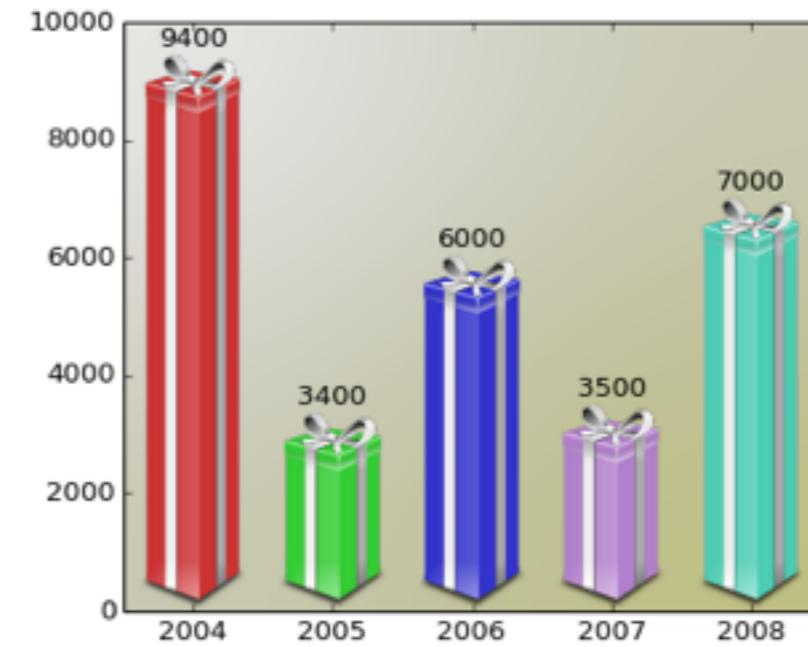
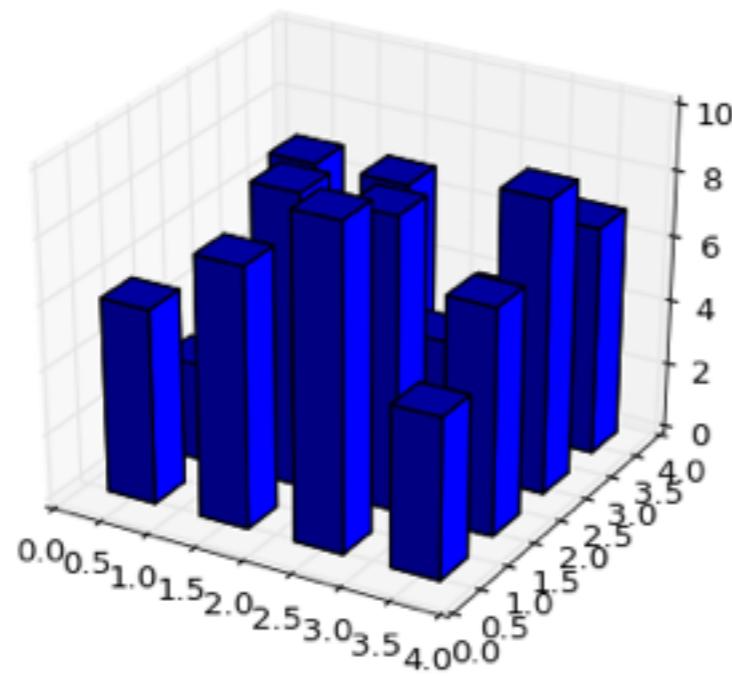
Avoid Chartjunk



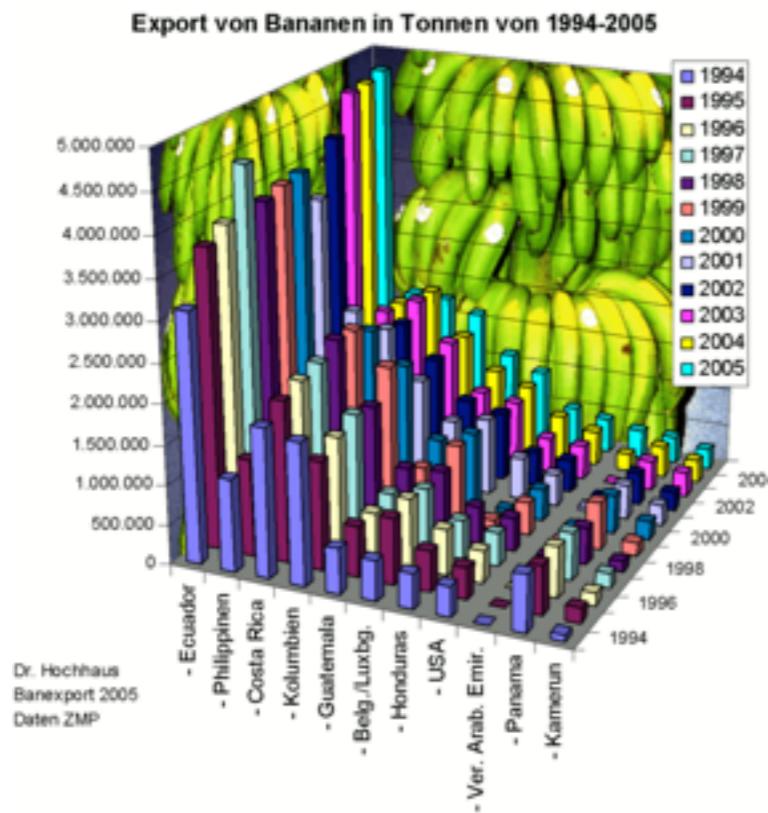
Avoid Chartjunk



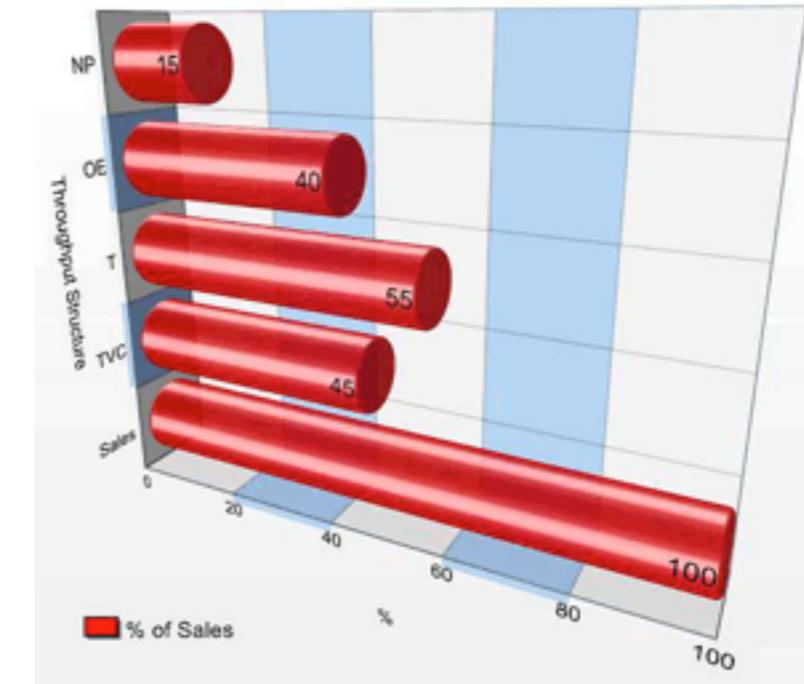
Don't!



matplotlib gallery



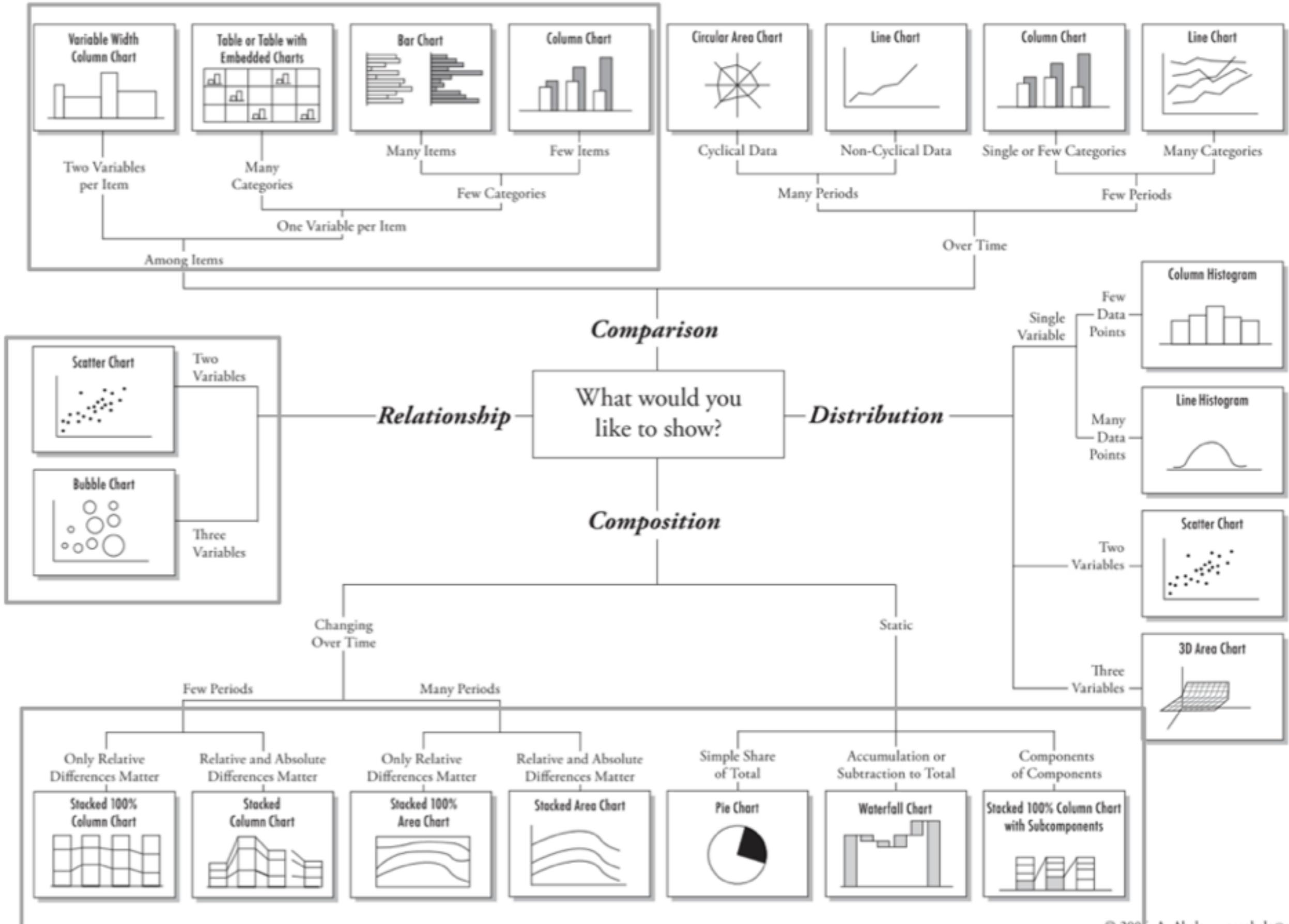
Dr. Hochhaus
Banlexport 2005
Daten ZMP



Excel Charts Blog

Use The Right Display

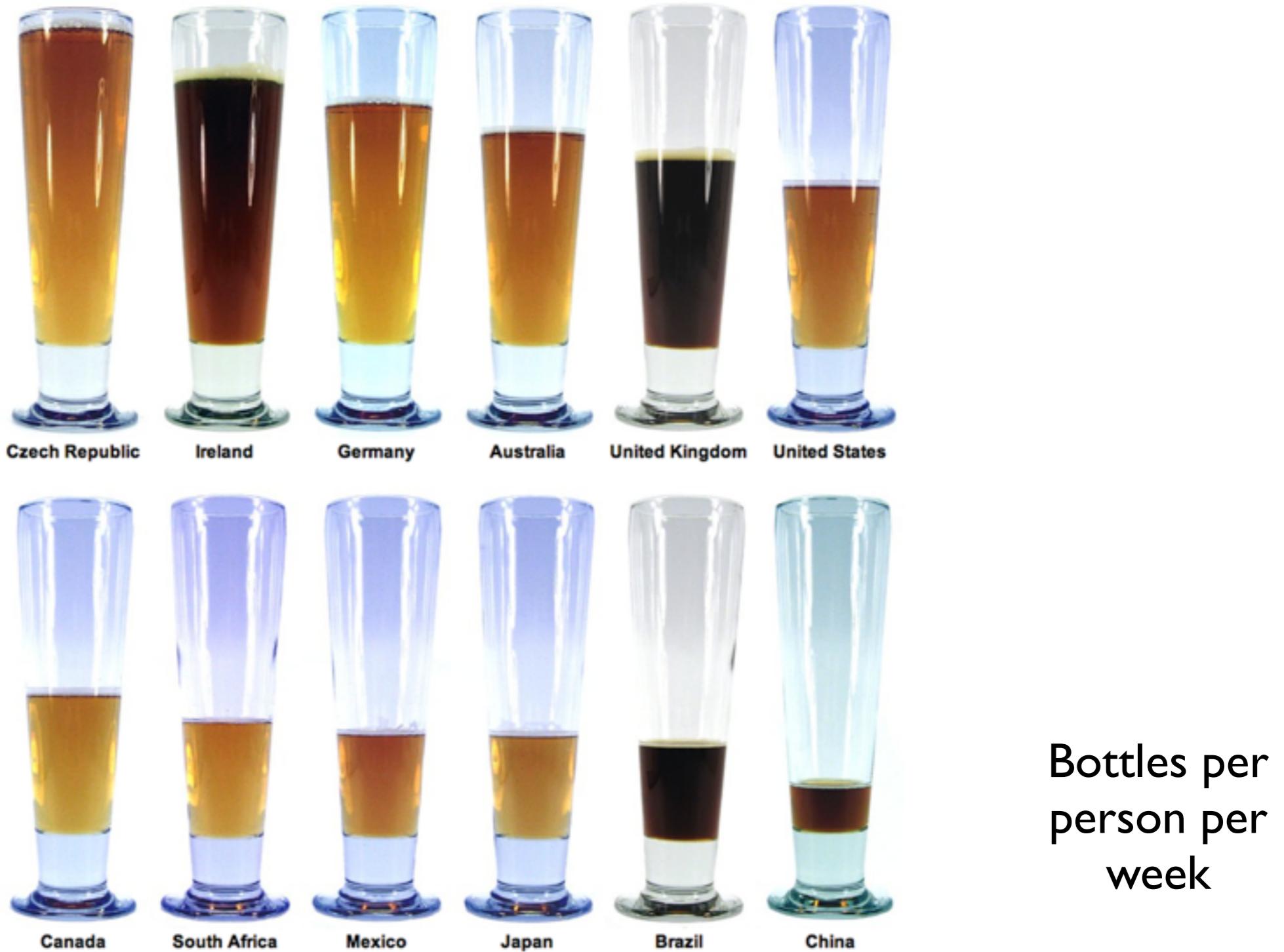
Chart Suggestions—A Thought-Starter



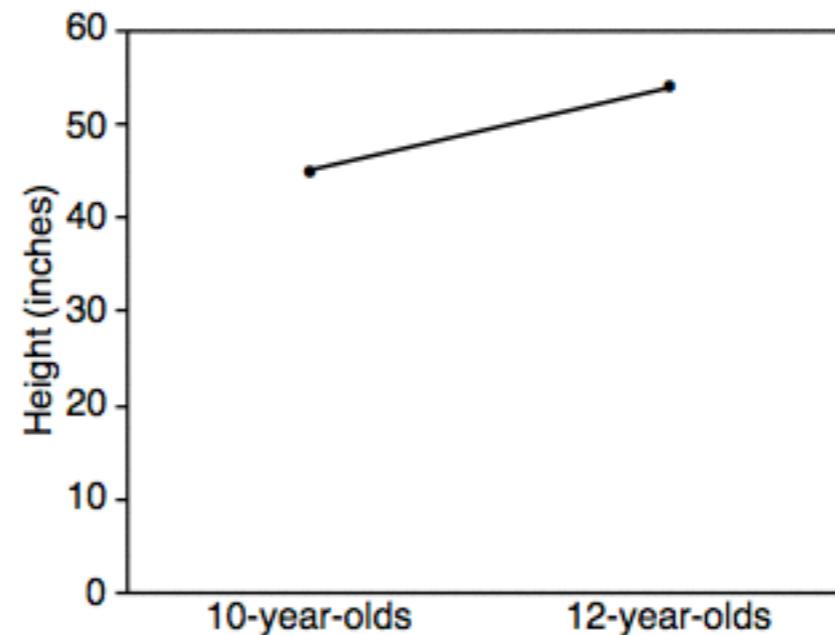
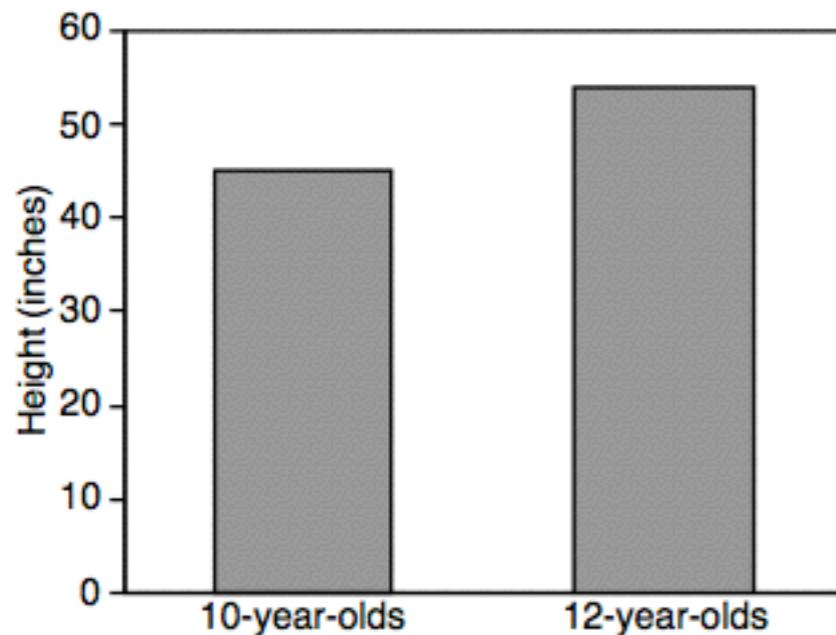
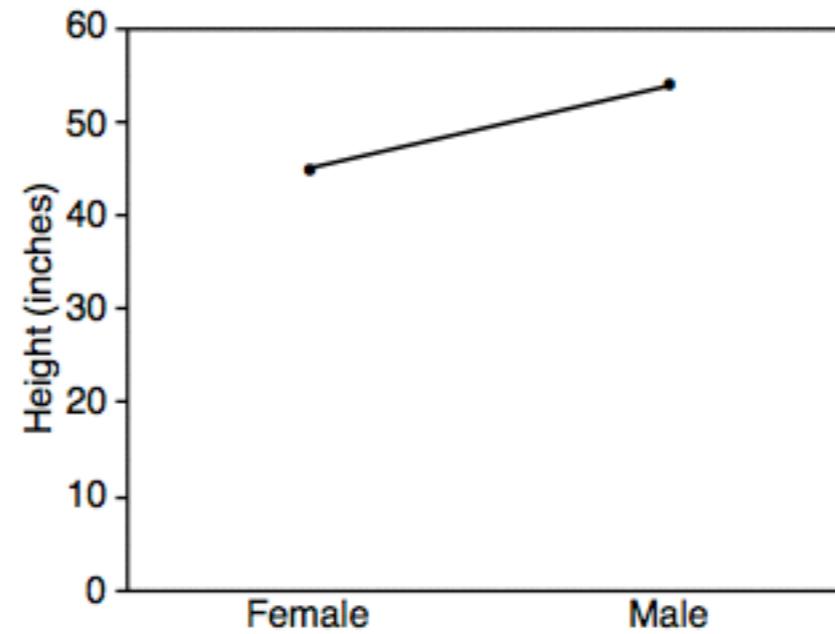
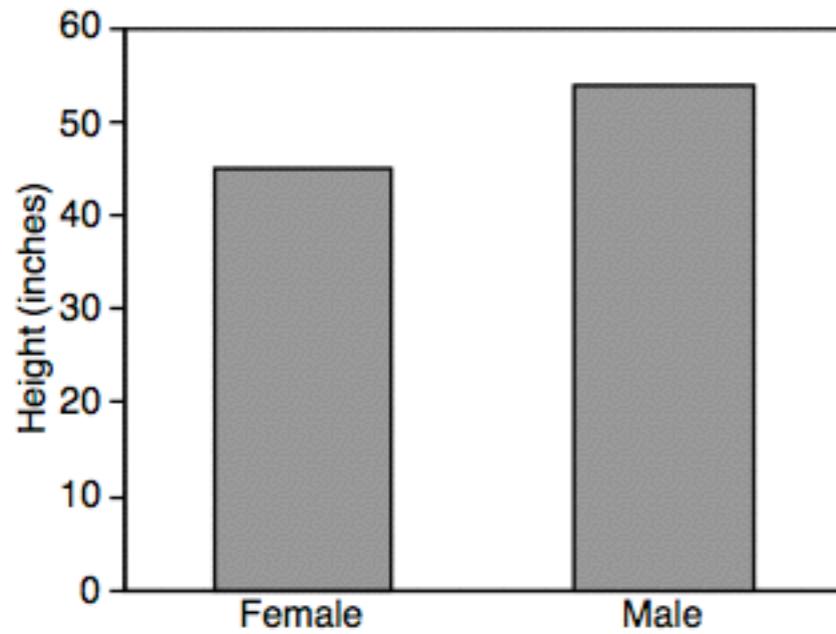
Comparisons

Bar Chart

How Much Does Beer Consumption Vary by Country?



Bars vs. Lines



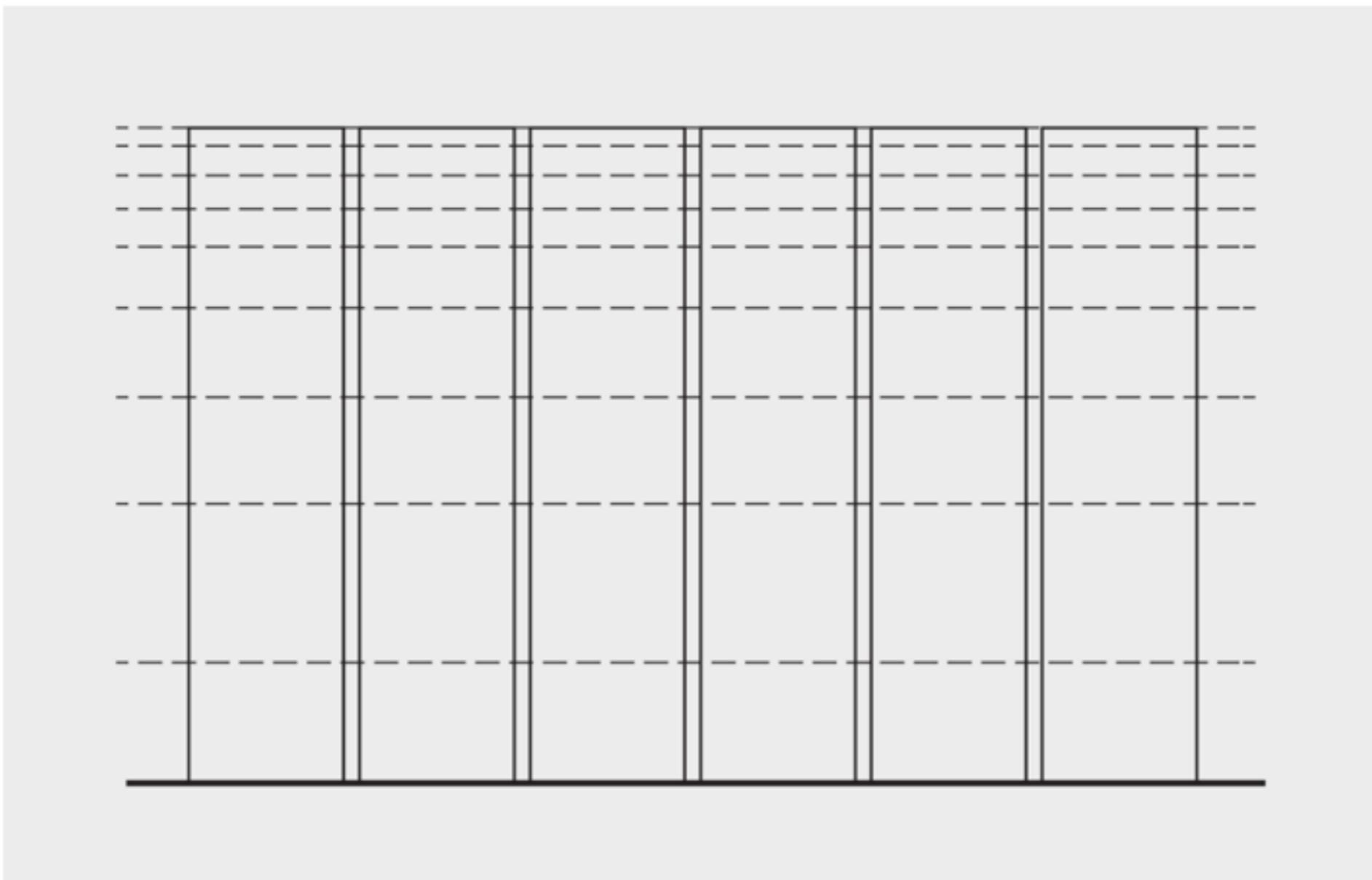
flowingdata.com/2015/08/31/bar-chart-baselines-start-at-zero

DRB | GitLab The Hub Feedly MD Syntax Add to Pinboard My Pinboard Instapaper libx bit.ly Orrick Box Timesheet LaTeX symbols Lore 3G Mobile Hotspot

FLOWINGDATA MEMBERSHIP TUTORIALS GUIDES BOOKS FEATURES [BECOME A MEMBER](#) | [LOG IN](#)

Bar Chart Baselines Start at Zero

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



There are visualization rules and there are visualization

Display a menu

Nathan Yau

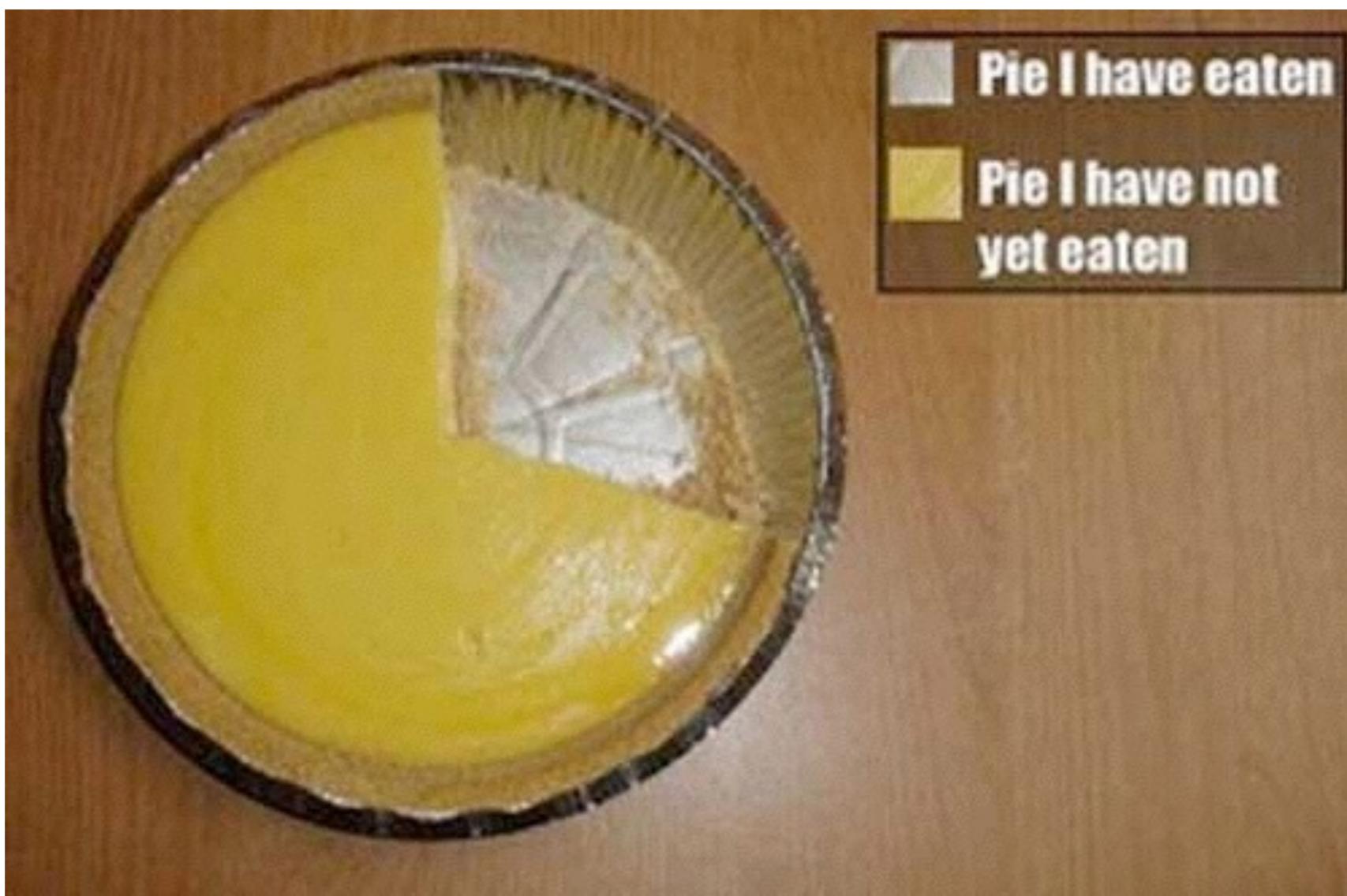
Trends

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

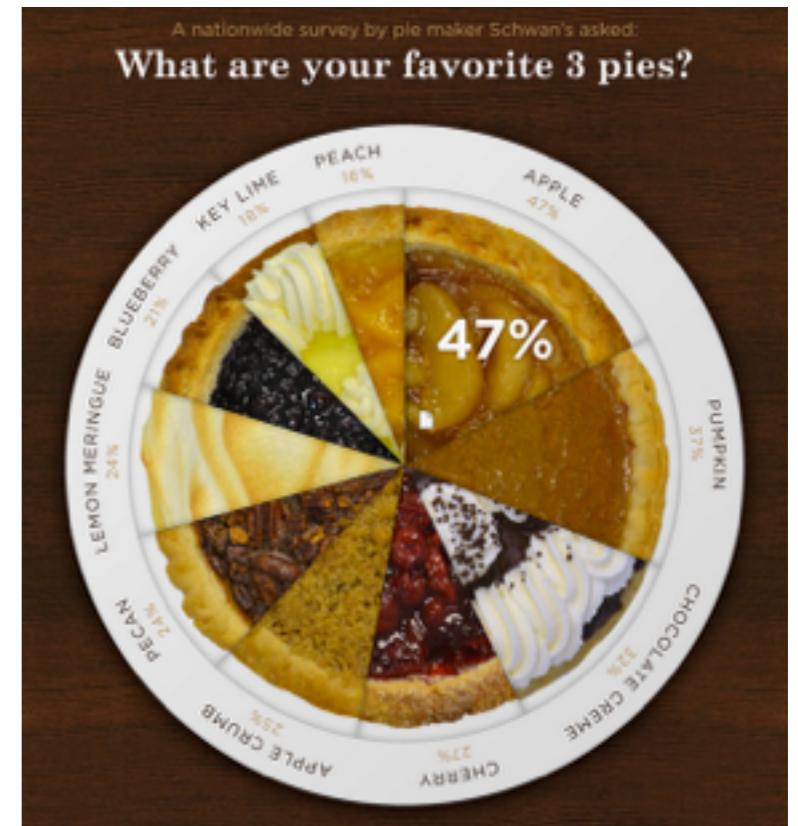
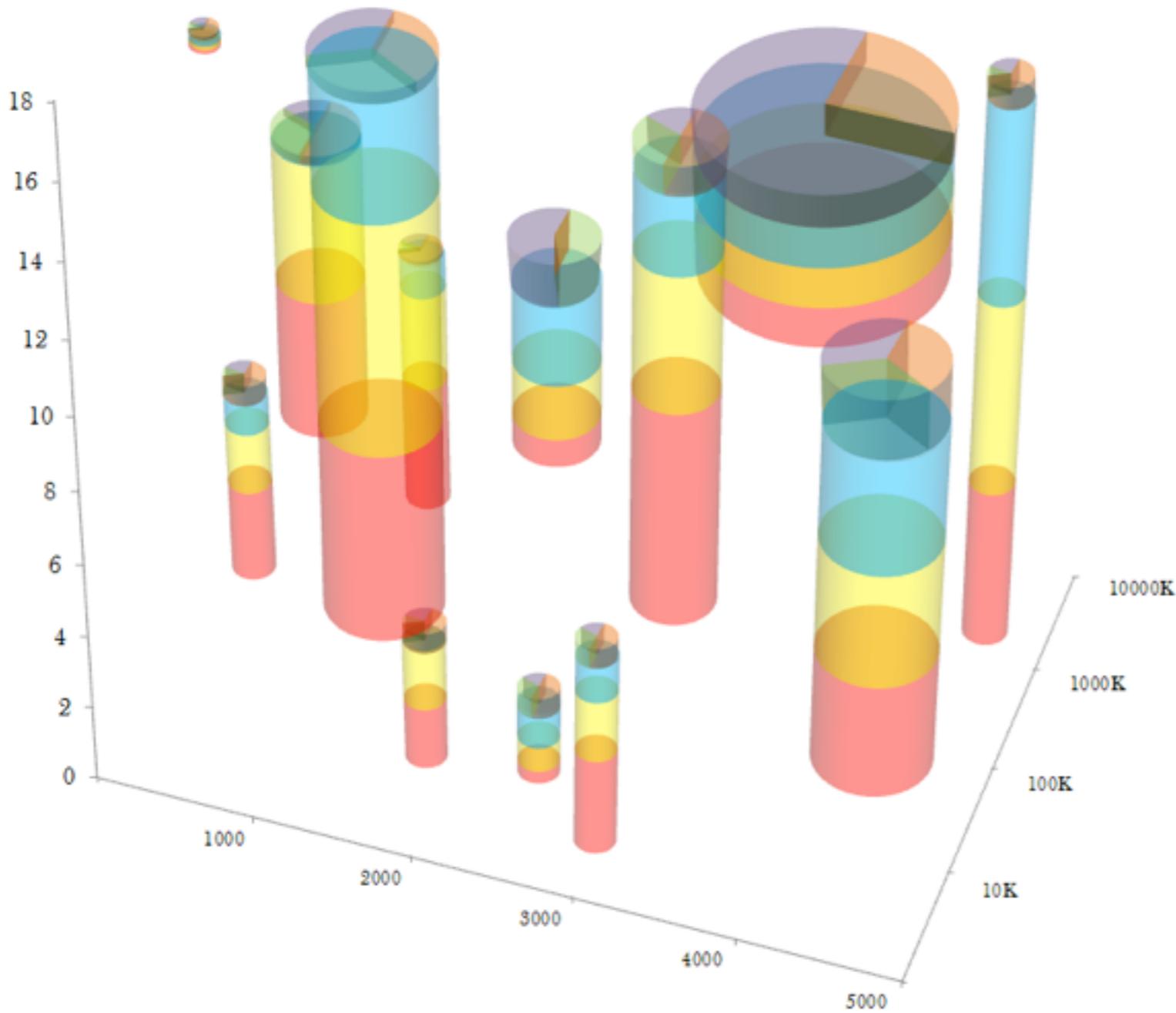
[GET CHART](#)[COMPARE](#)[EVENTS](#) ▾[TECHNICAL INDICATORS](#) ▾[CHART SETTINGS](#) ▾[RESET](#)

Proportions

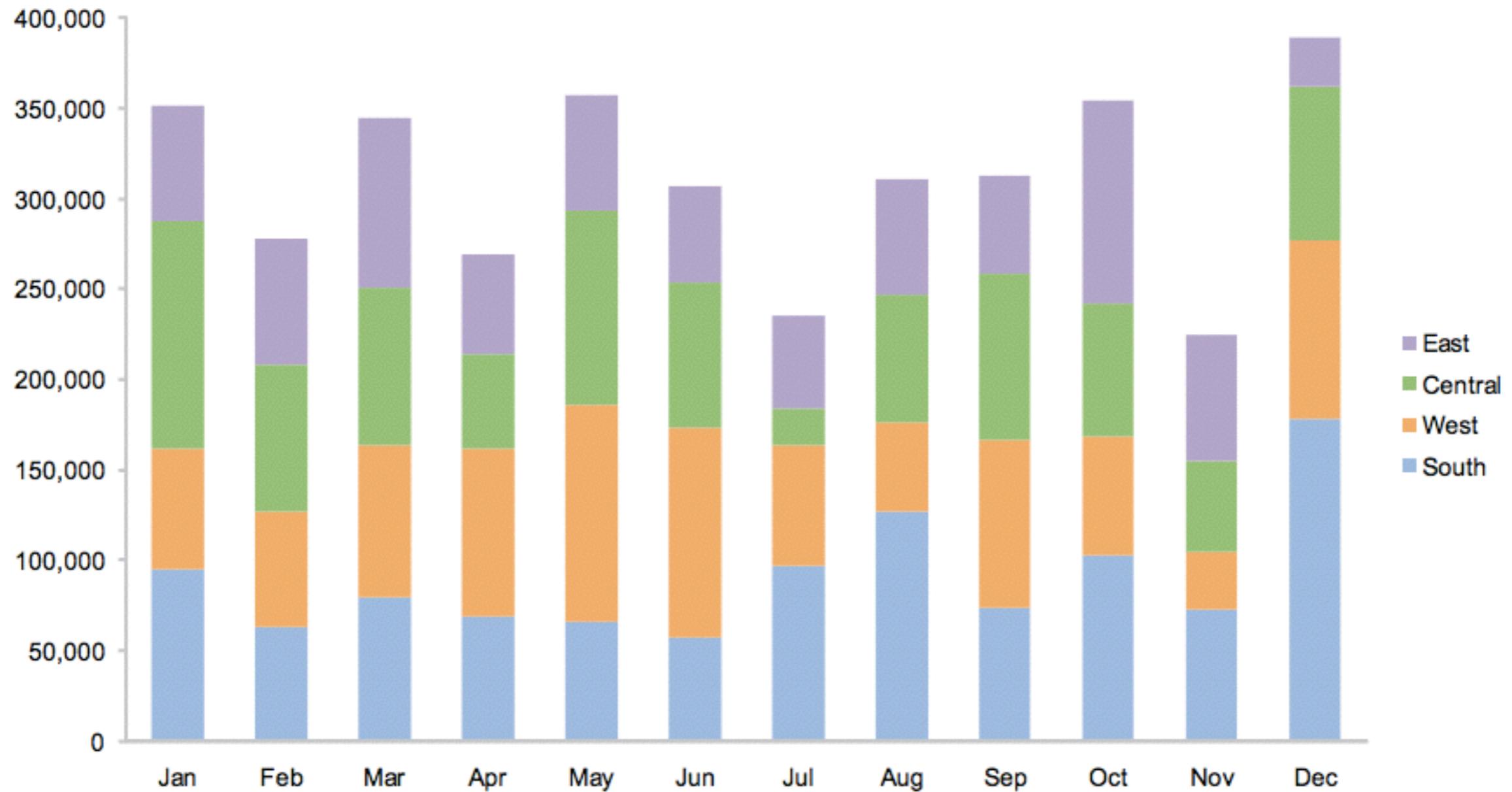
Pie Charts



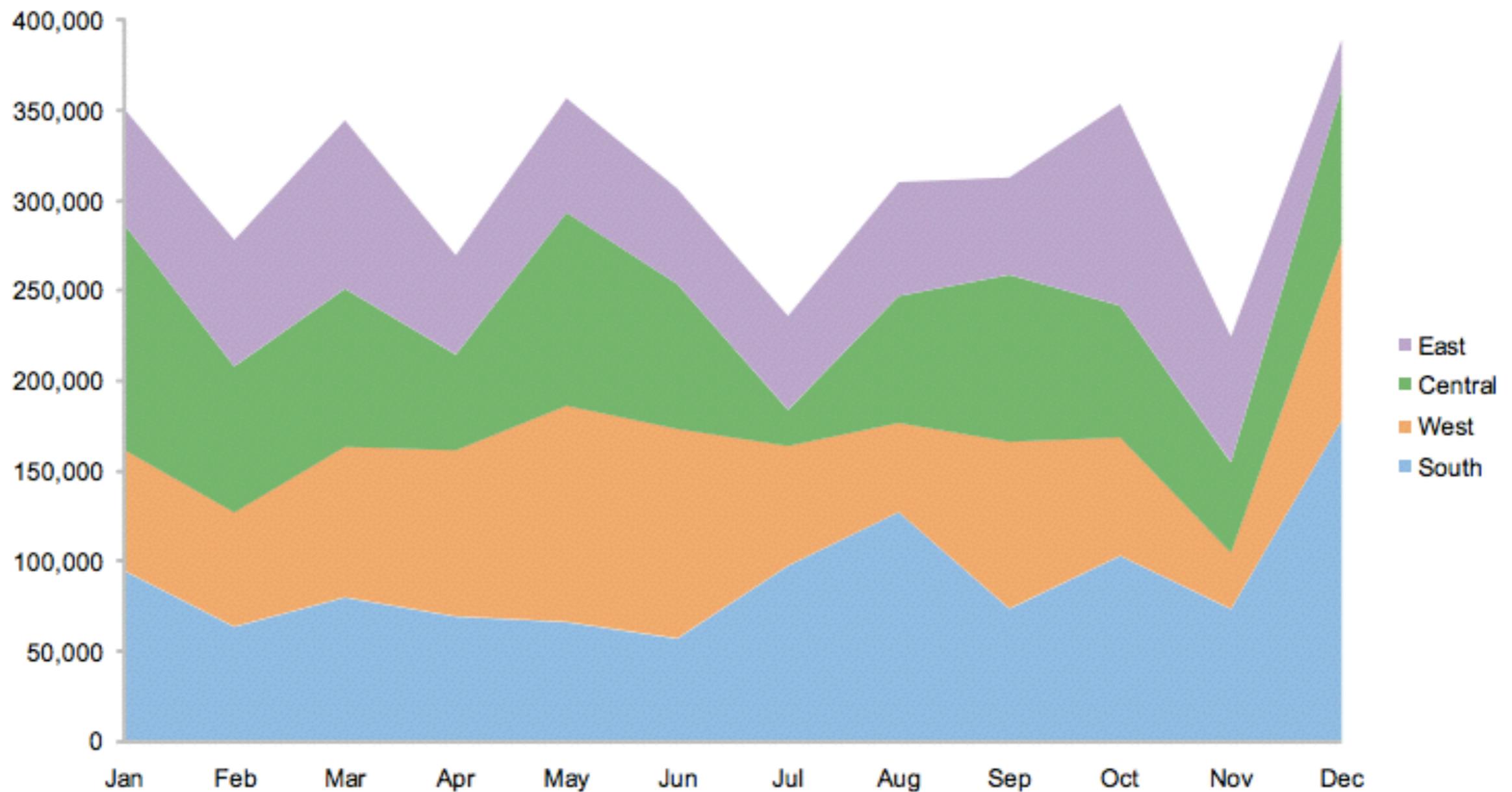
eagerpies.com



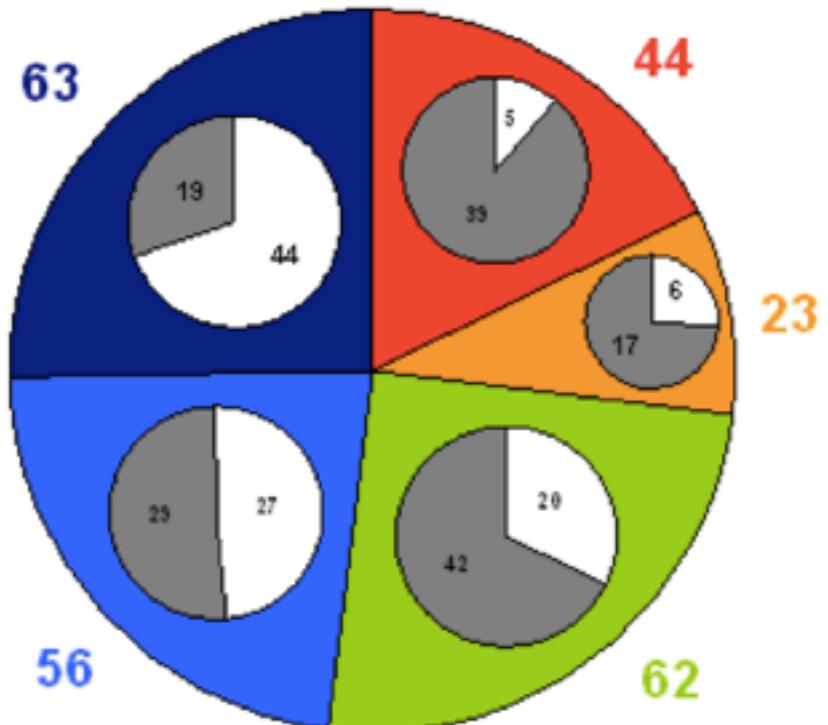
Stacked Bar Chart



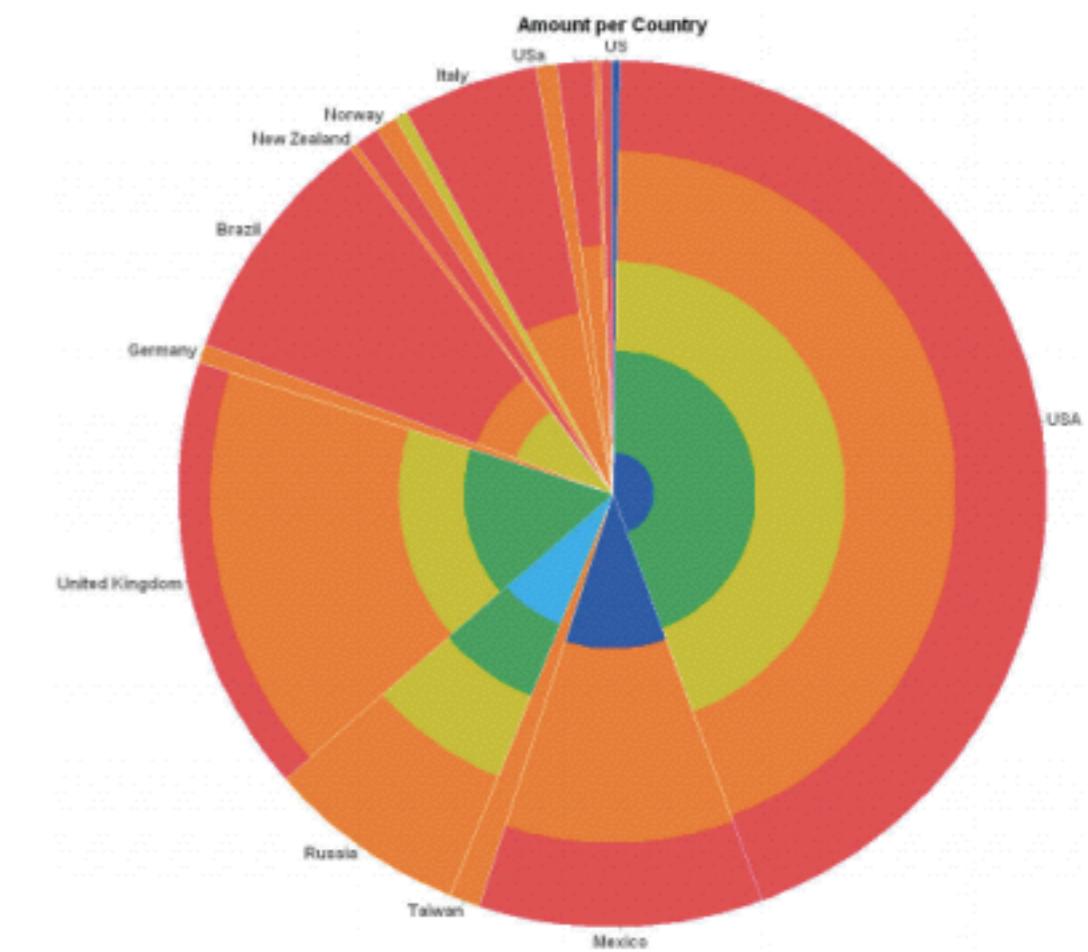
Stacked Area Chart



Don't!

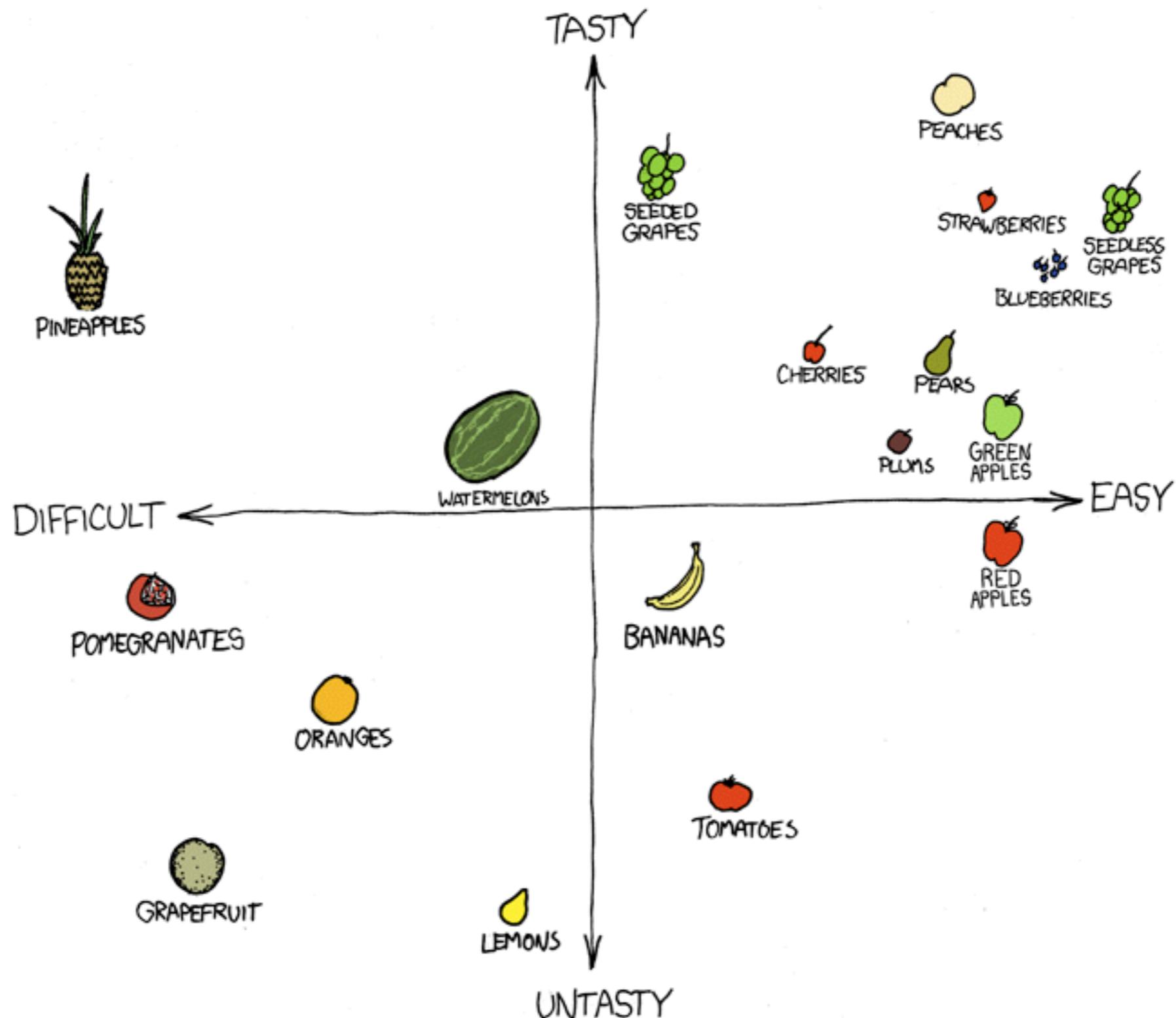


- new folds
- new folds, partial similarity
- putative analogs
- putative homologs
- recognizable homologs
- hypothesis about function
- no hypothesis about function

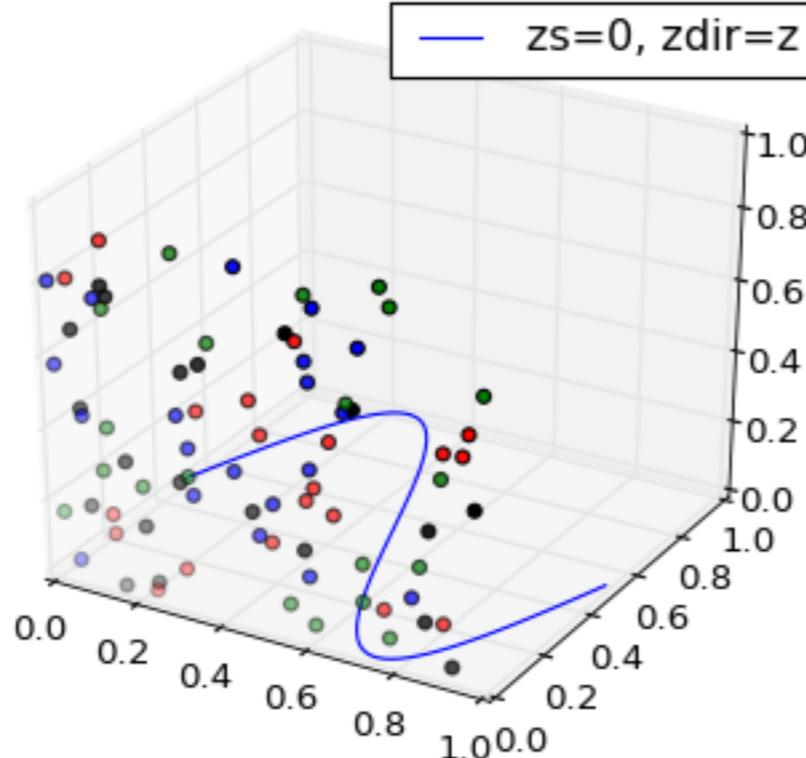
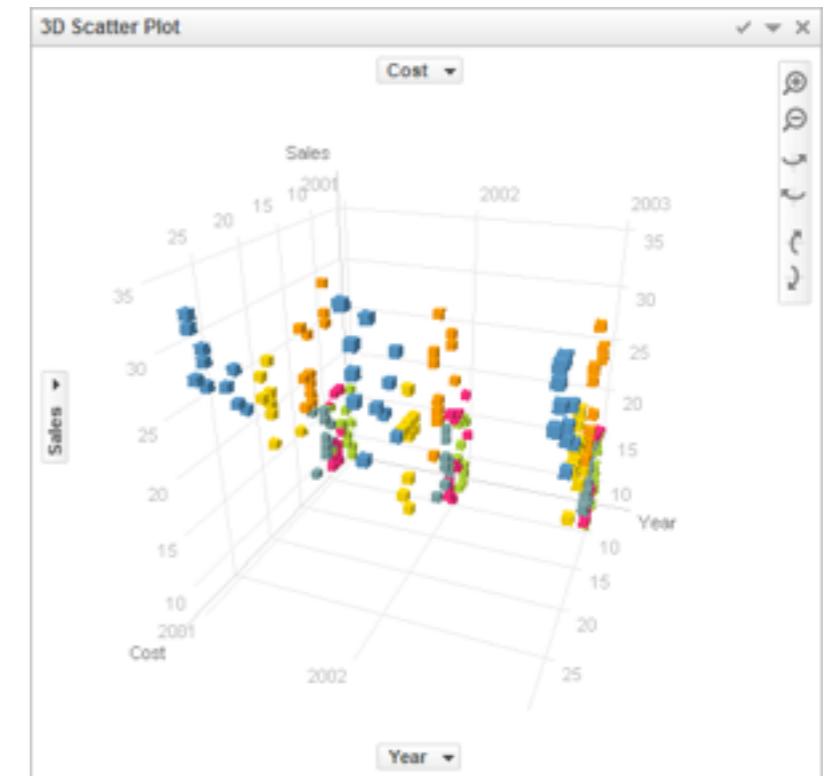
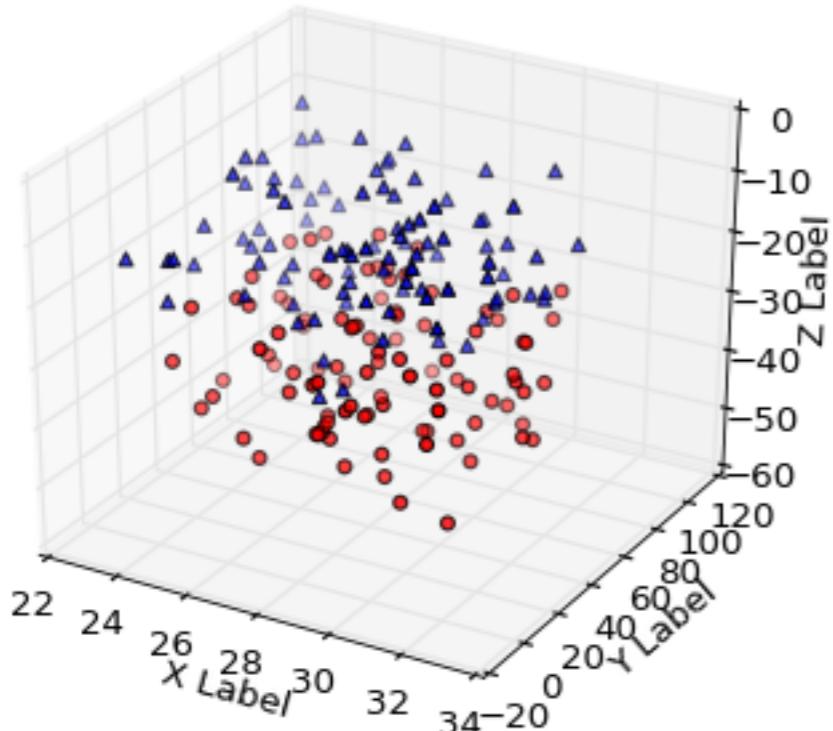


Correlations

Scatterplots

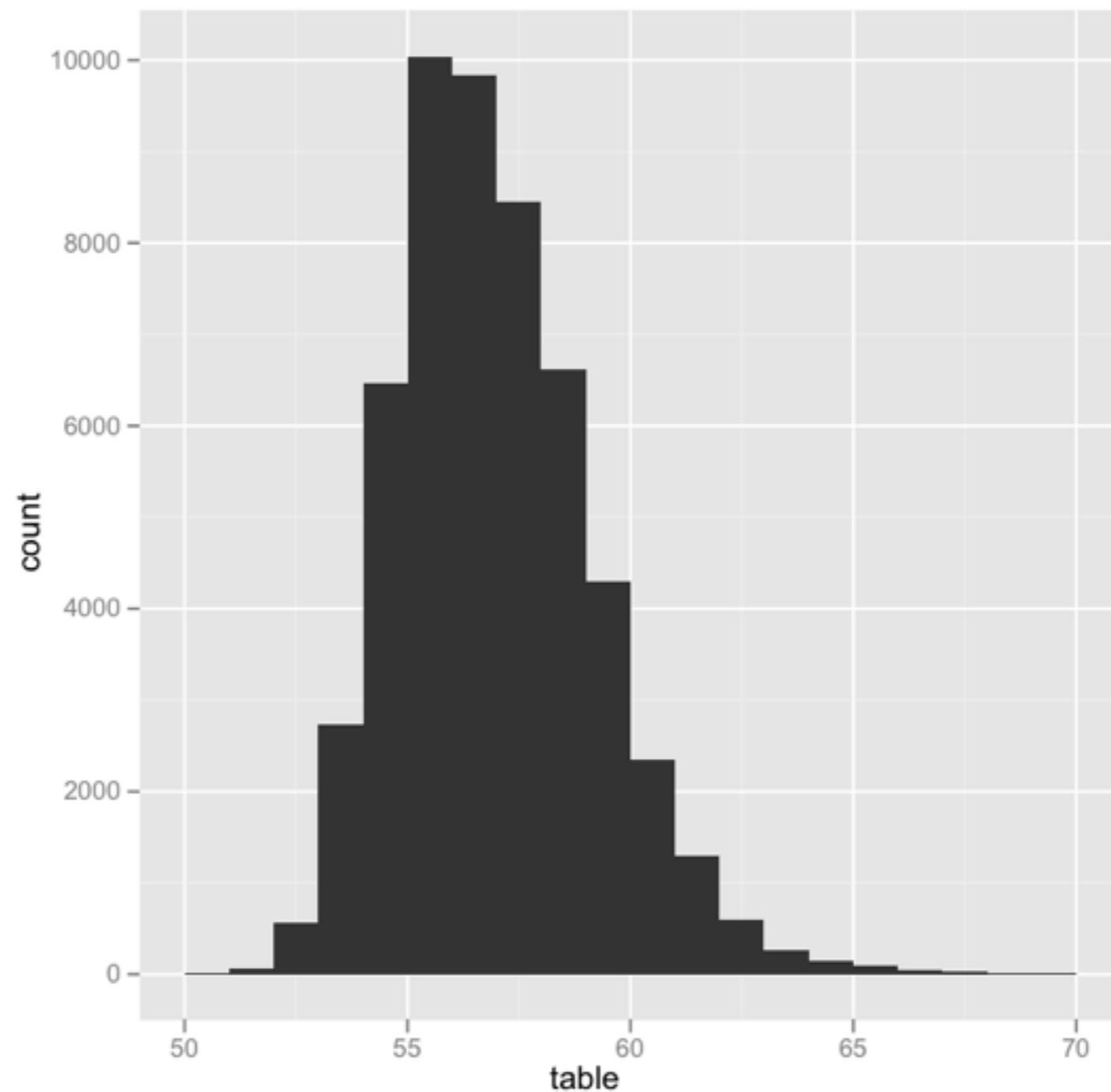


Don't!



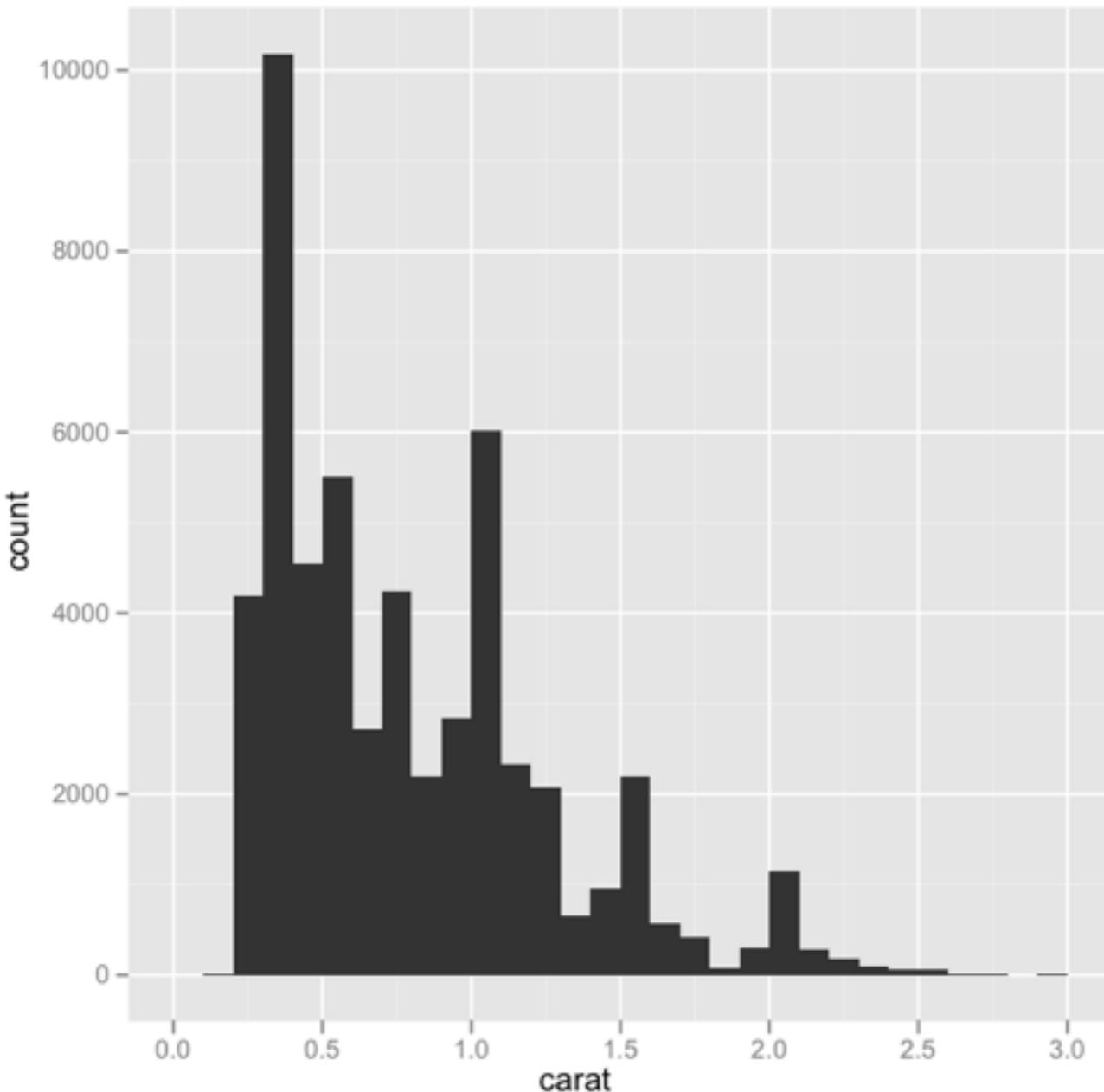
Distributions

Histogram

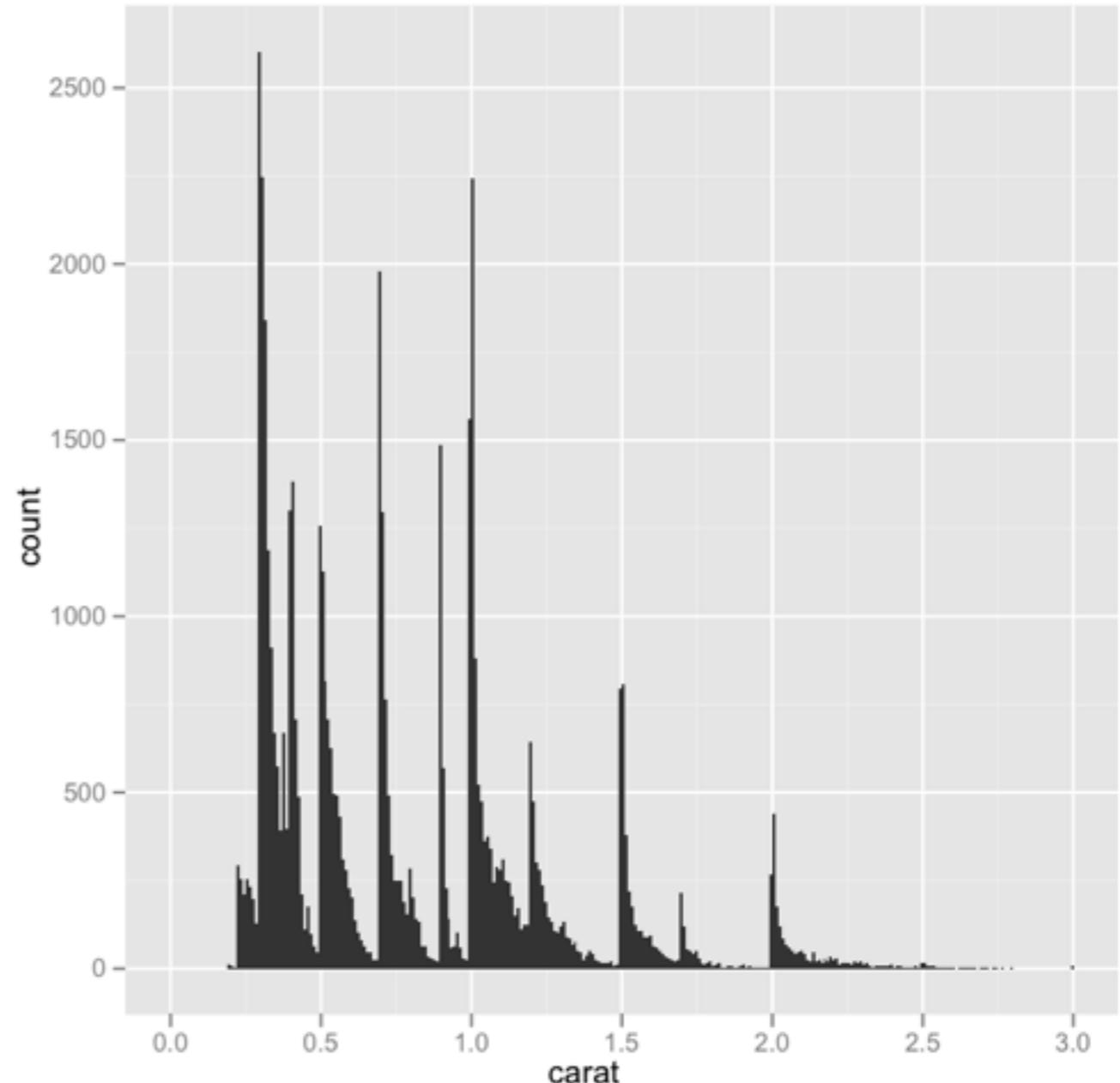


ggplot2

Bin Width

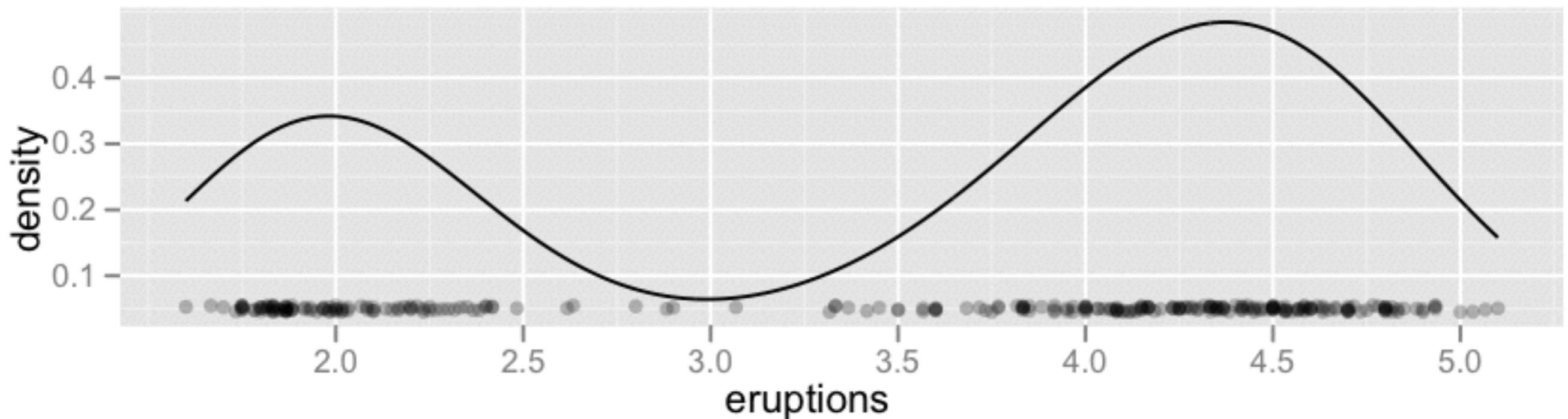


binwidth = 0.1

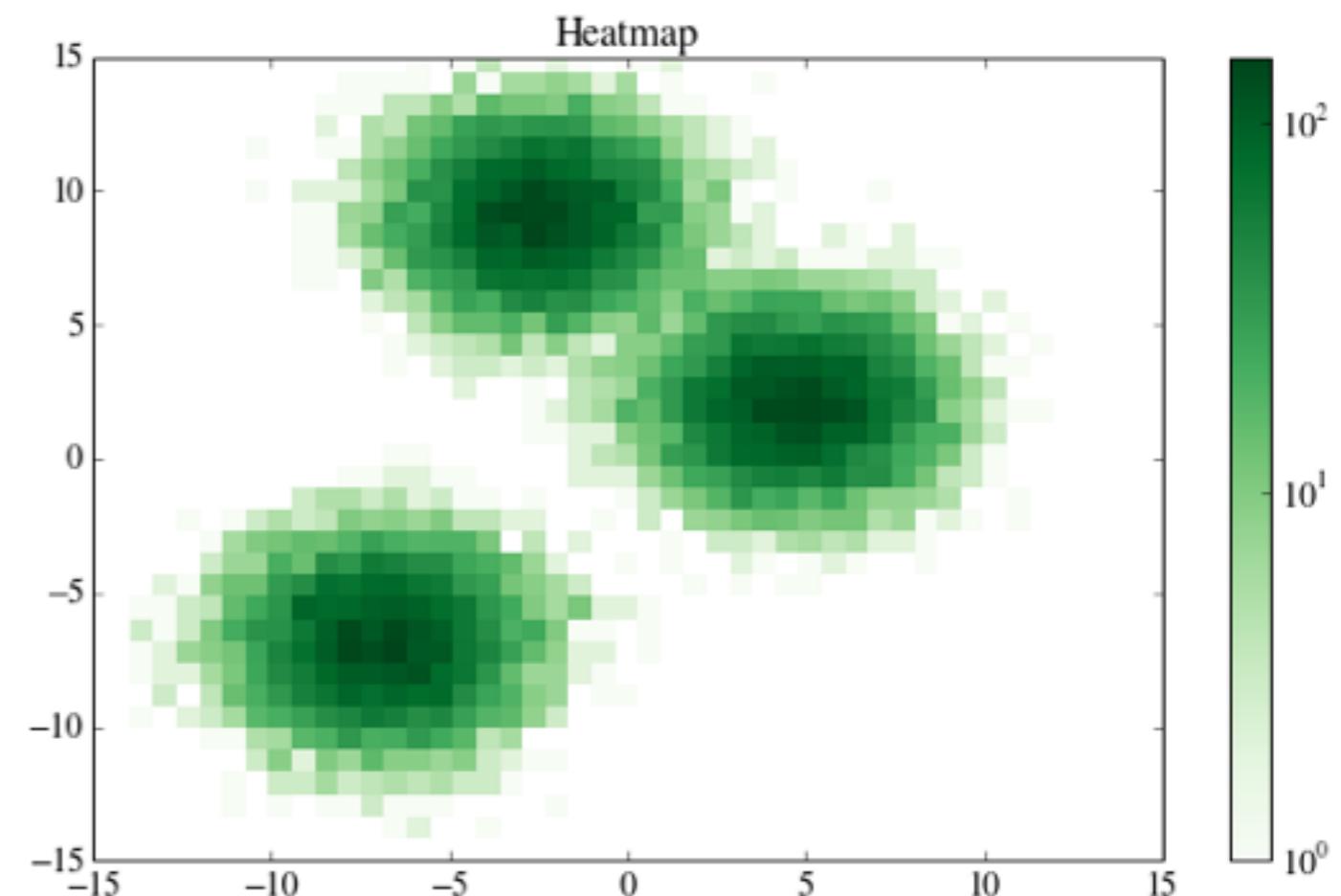
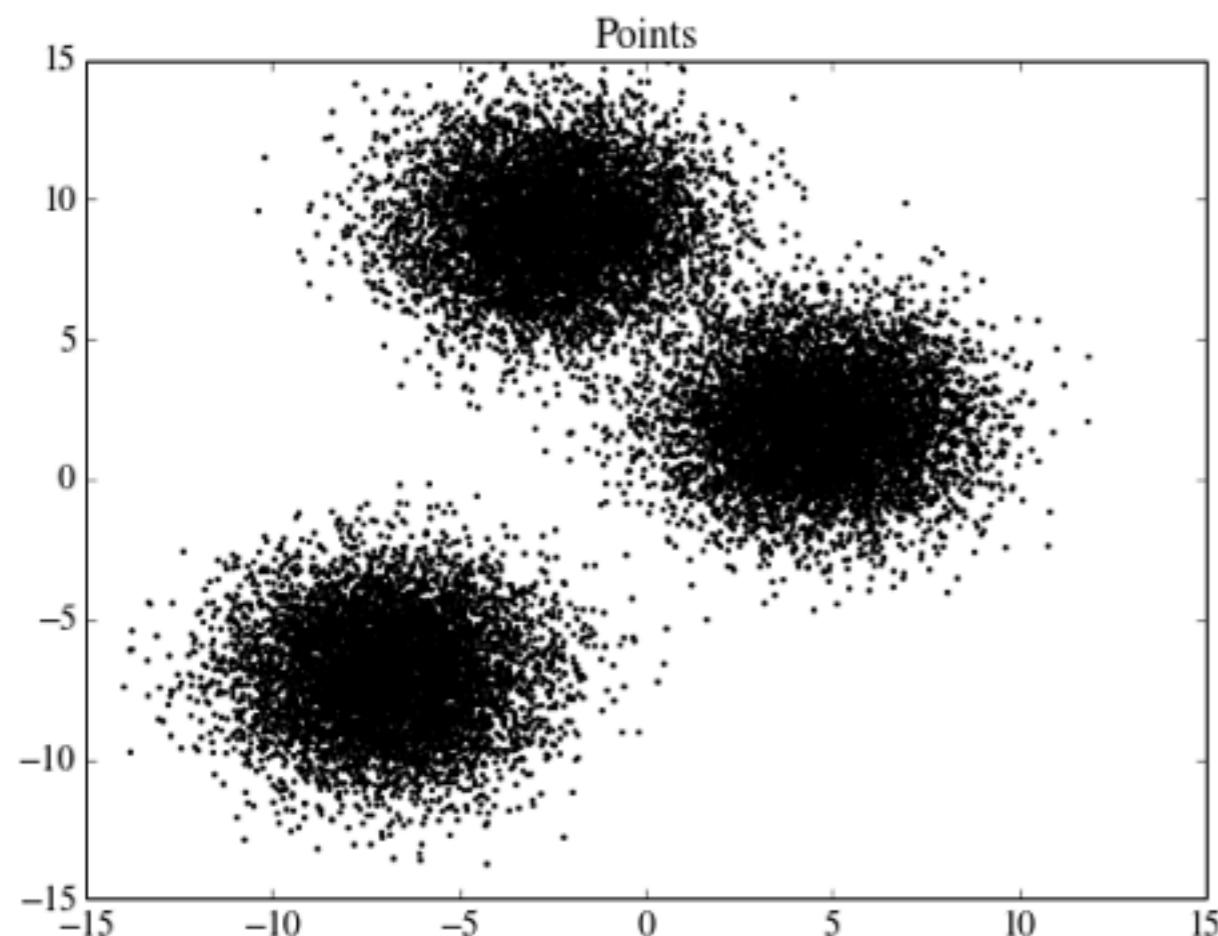


binwidth = 0.01

Density Plots



2D Density Plots



Seaborn Tutorial

DRB | GitLab The Hub Feedly MD Syntax Add to Pinboard My Pinboard Instapaper libx bit.ly Orrick Box Timesheet LaTeX symbols Lore 3G Mobile Hotspot

seaborn 0.6.0 API Tutorial Gallery Site ▾ Page ▾ Search

```
np.random.seed(sum(map(ord, "distributions")))
```

Plotting univariate distributions

The most convenient way to take a quick look at a univariate distribution in seaborn is the `distplot()` function. By default, this will draw a `histogram` and fit a `kernel density estimate` (KDE).

```
x = np.random.normal(size=100)
sns.distplot(x);
```

Histograms

Histograms are likely familiar, and a `hist` function already exists in matplotlib. A histogram represents the distribution of data by forming bins along the range of the data and then drawing bars to show the number of observations that fall in each bin.

To illustrate this, let's remove the density curve and add a rug plot, which draws a small vertical tick at each observation. You can make the rug plot itself with the `rugplot()` function, but it is also available in `distplot()`:

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

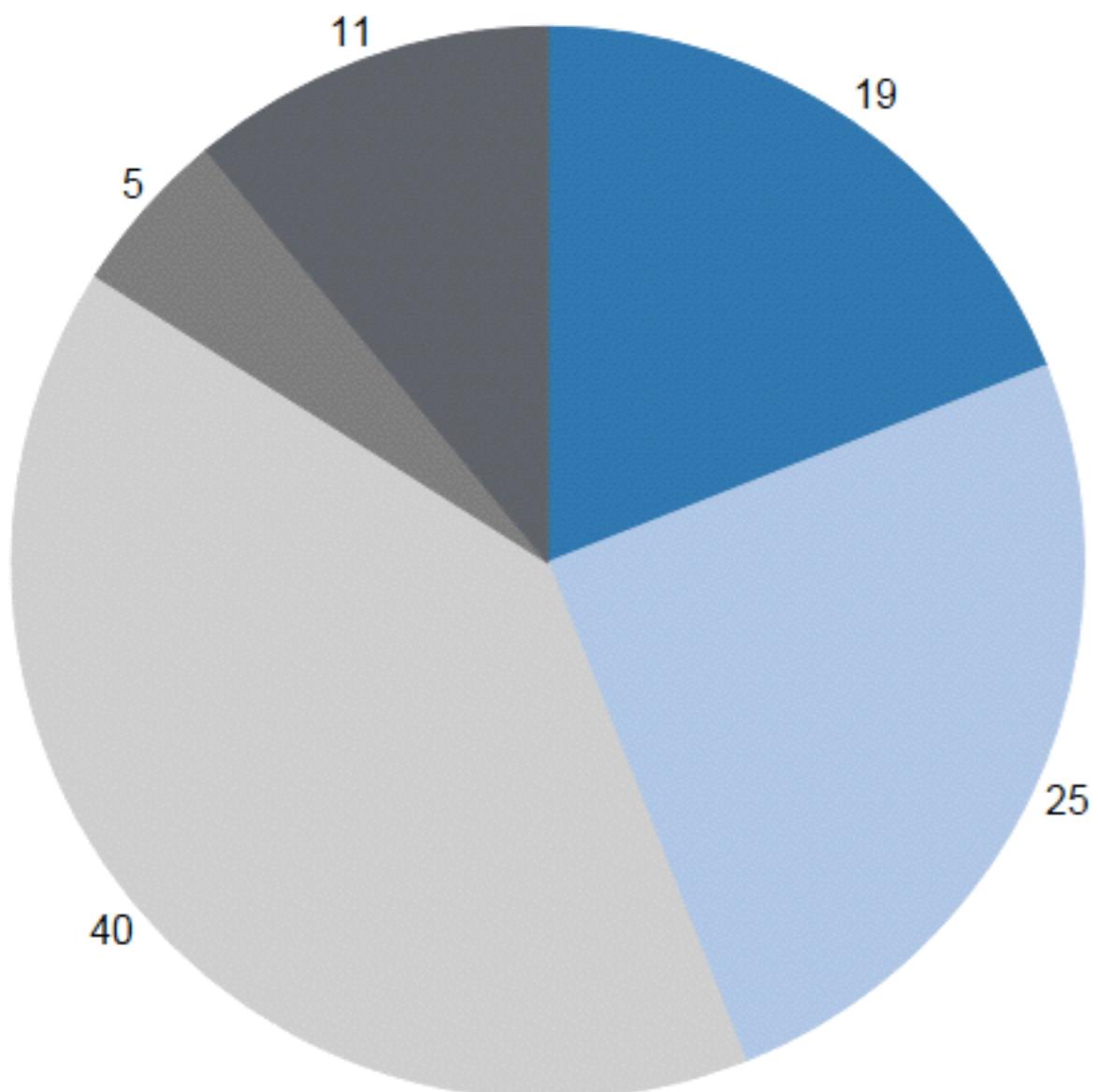
Data courtesy of Cole Nussbaumer

How do you feel about doing science?

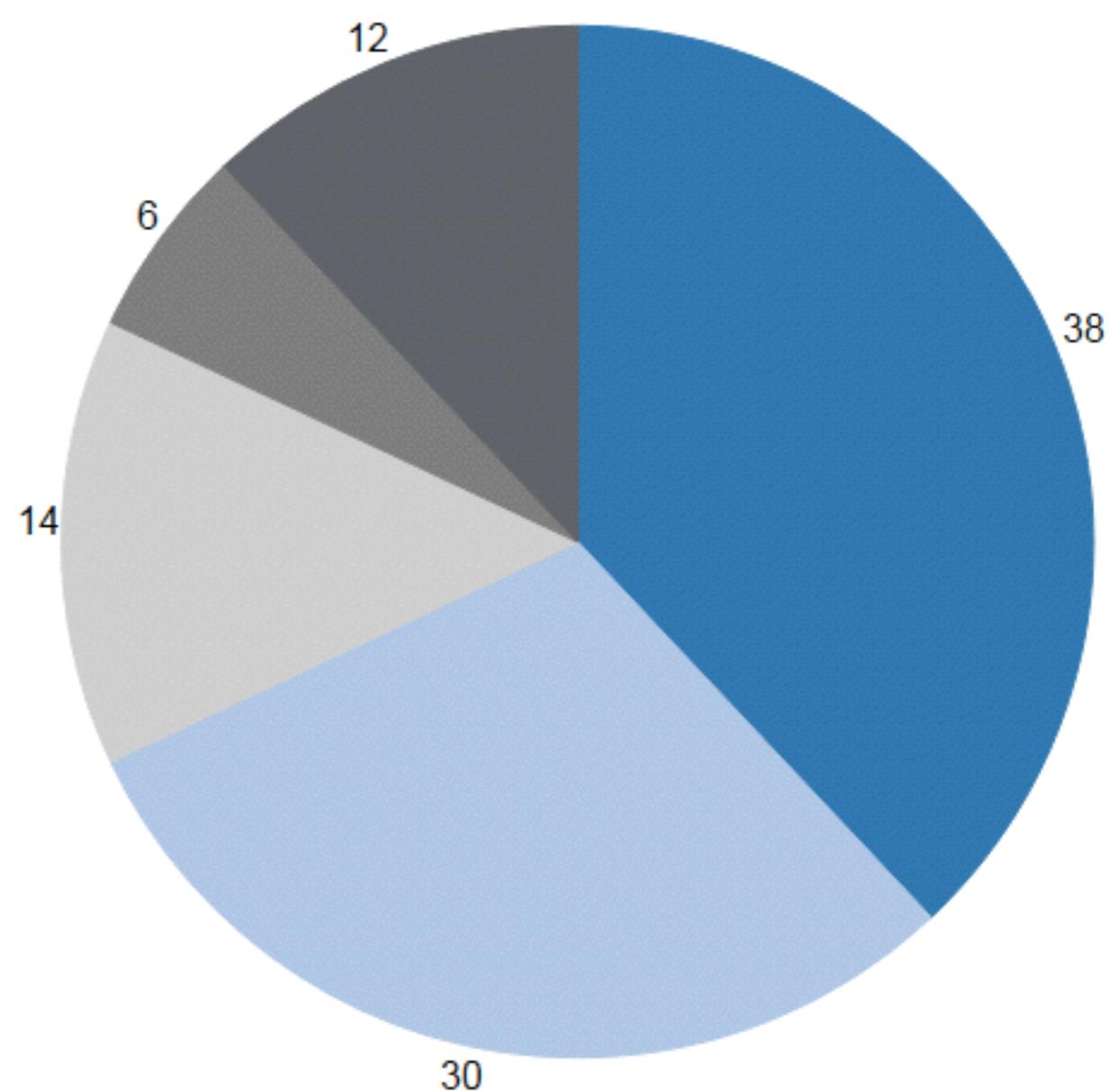
Before

Interest

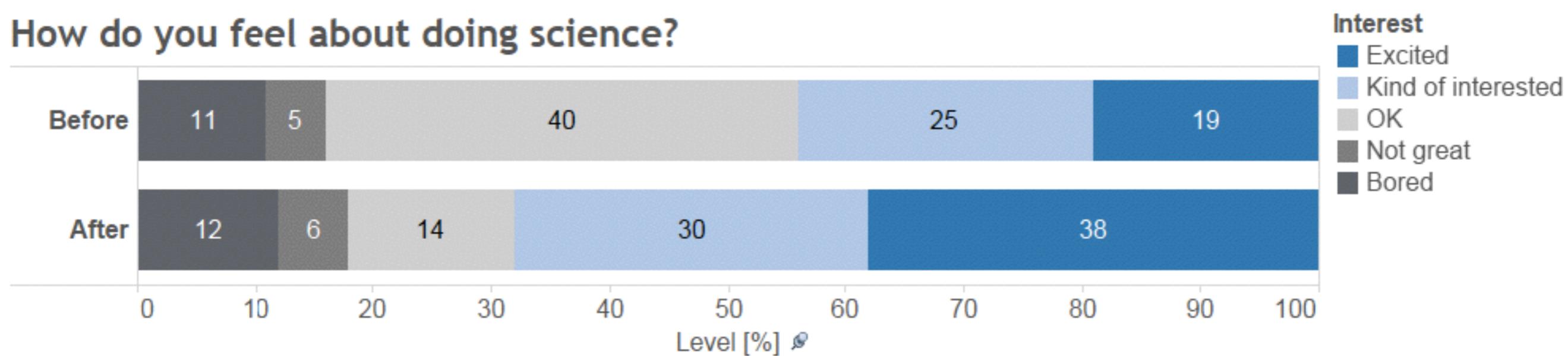
- Excited
- Kind of interested
- OK
- Not great
- Bored



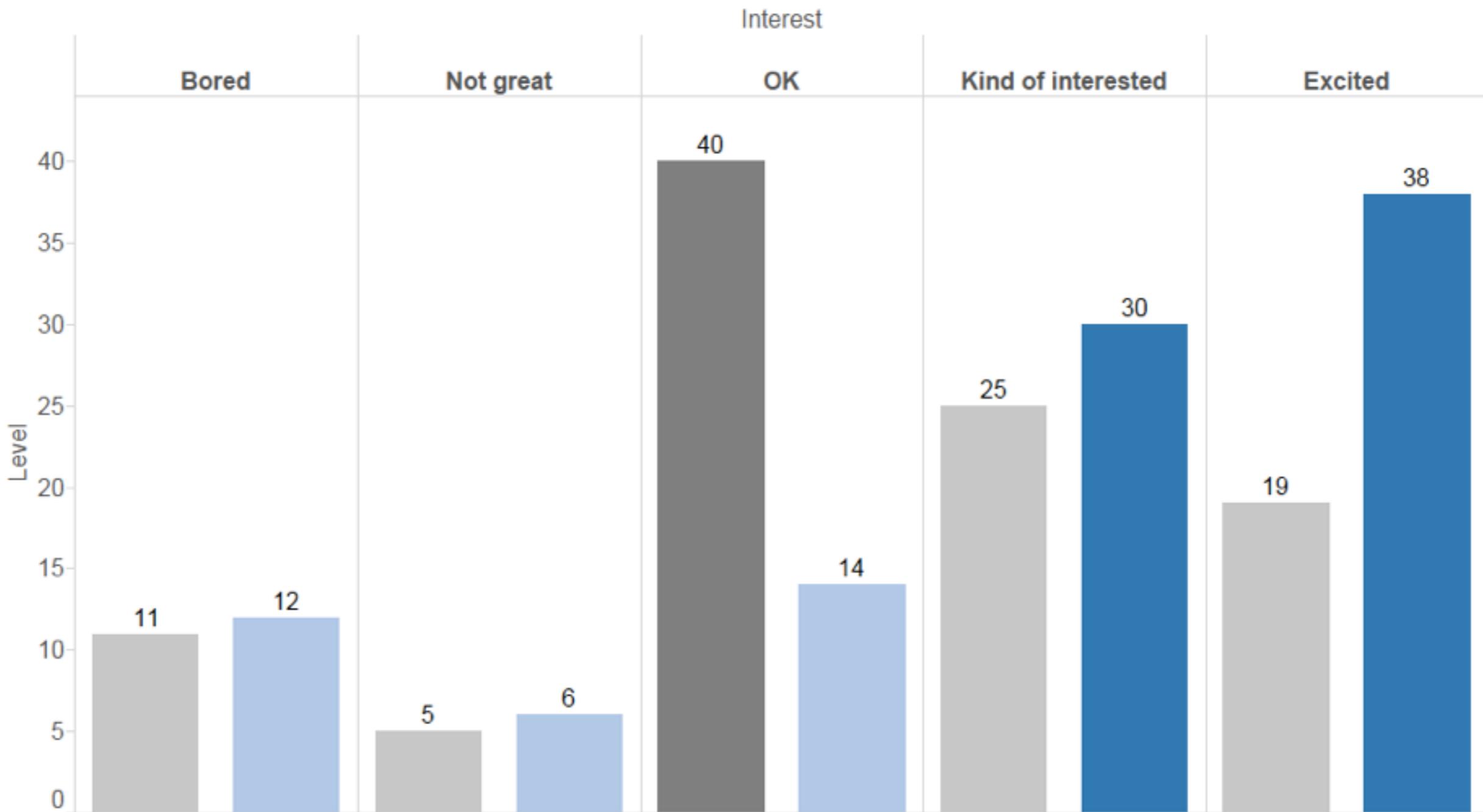
After



How do you feel about doing science?

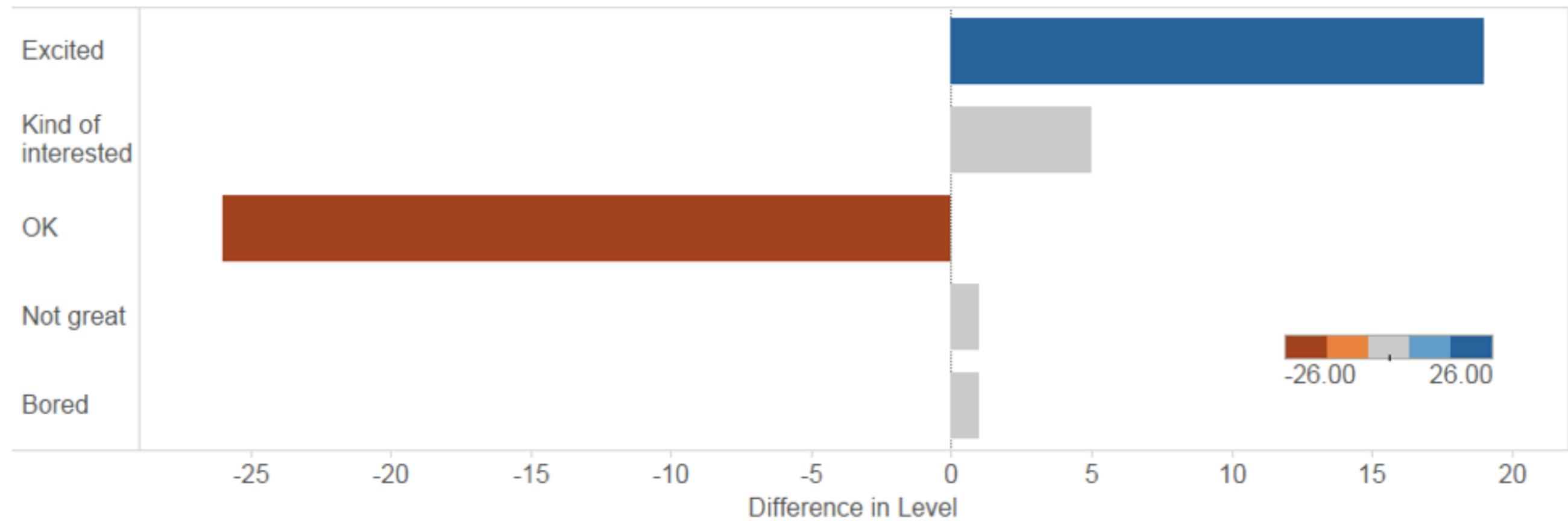


How do you feel about doing science?

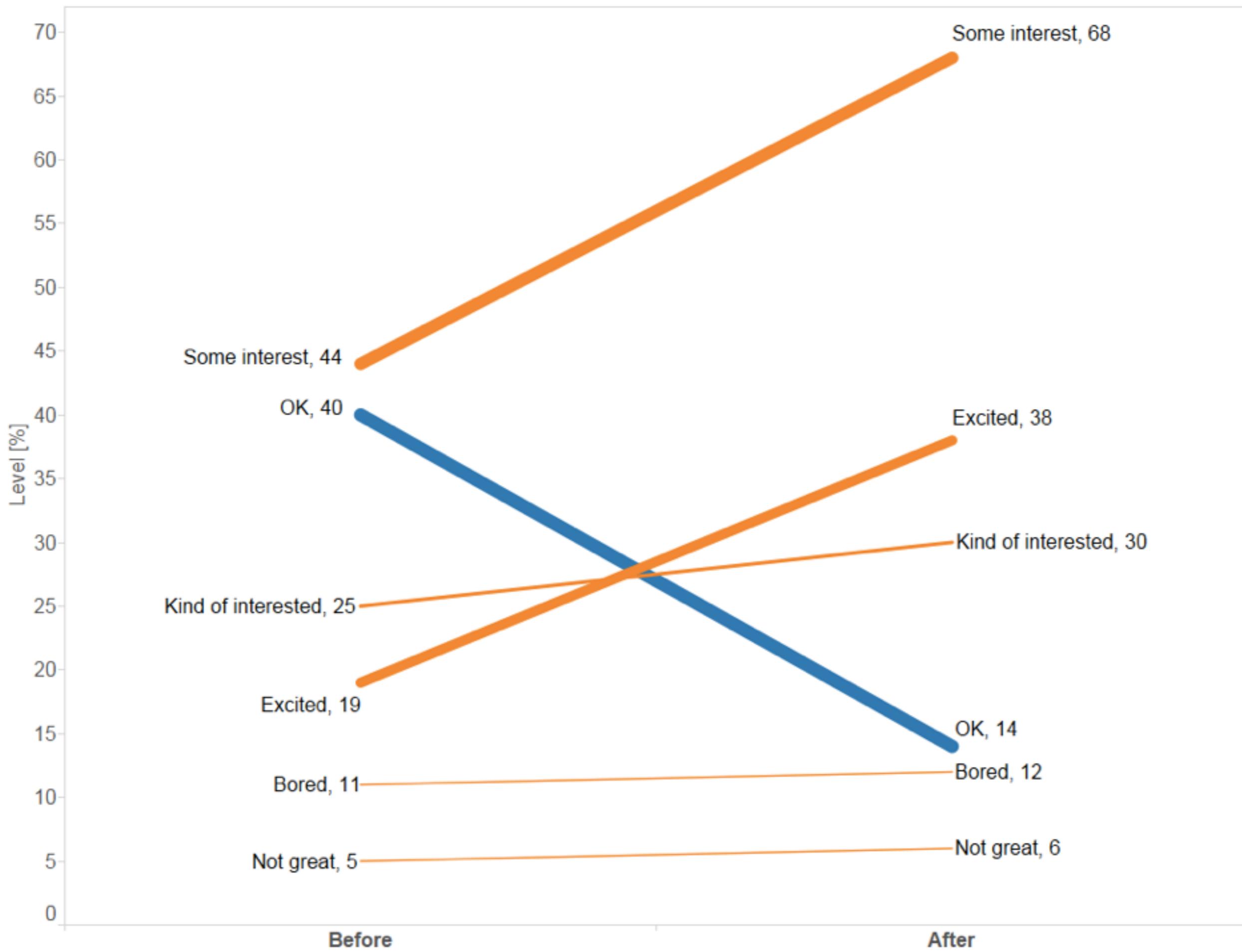


Before the program, the majority of children felt just *OK* about science. After the program, more children were *Kind of interested* and *Excited* about science.

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



How do you feel about doing science?

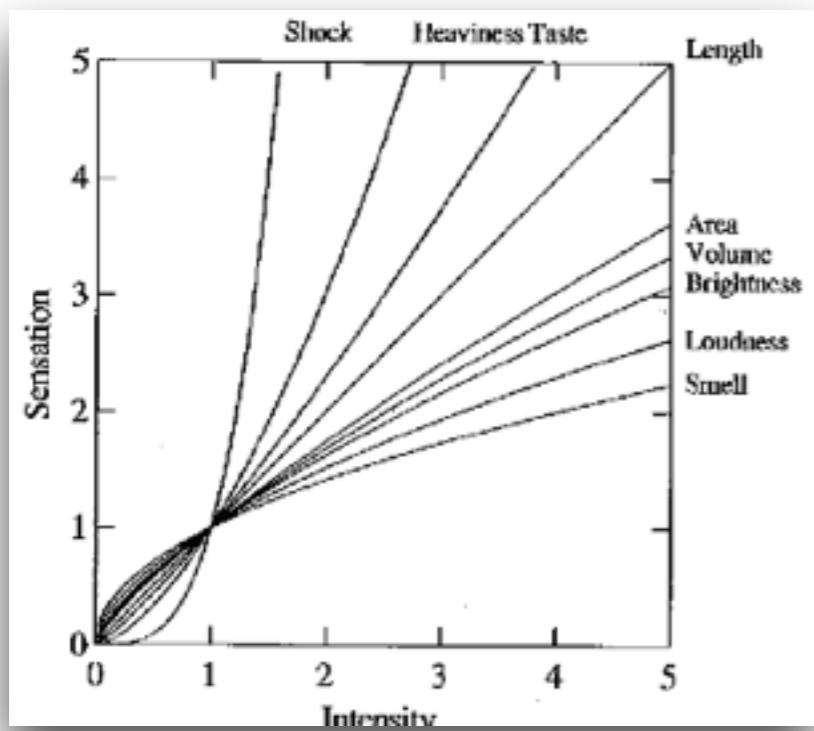


After the pilot program,

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

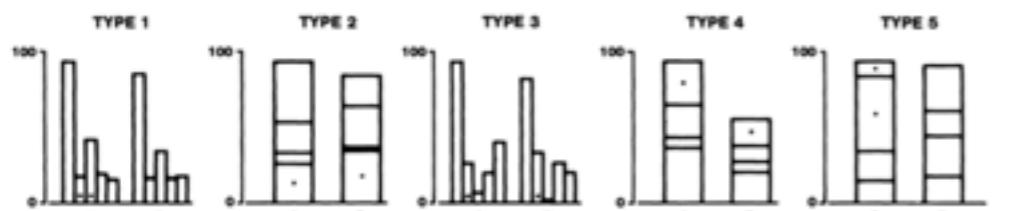


Figure 4. Graphs from position-length experiment.

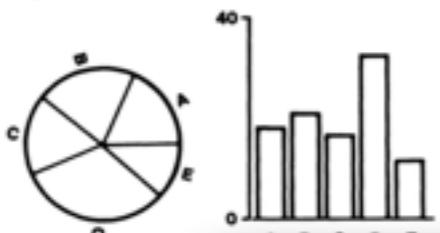
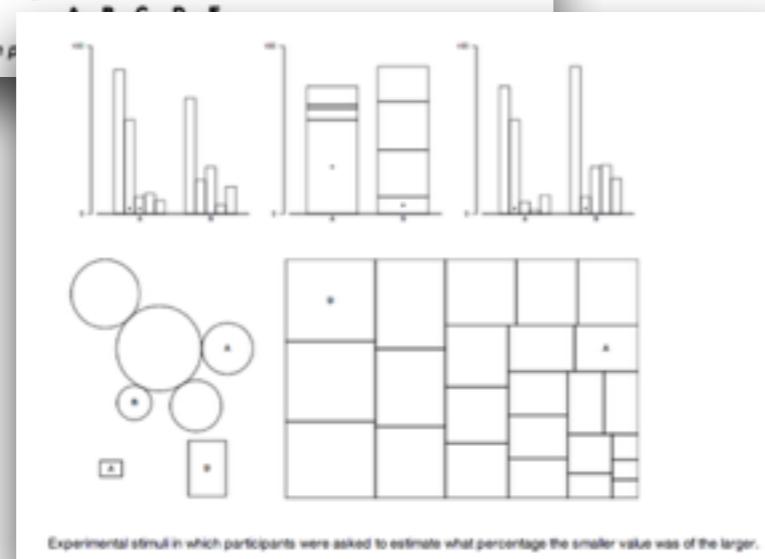
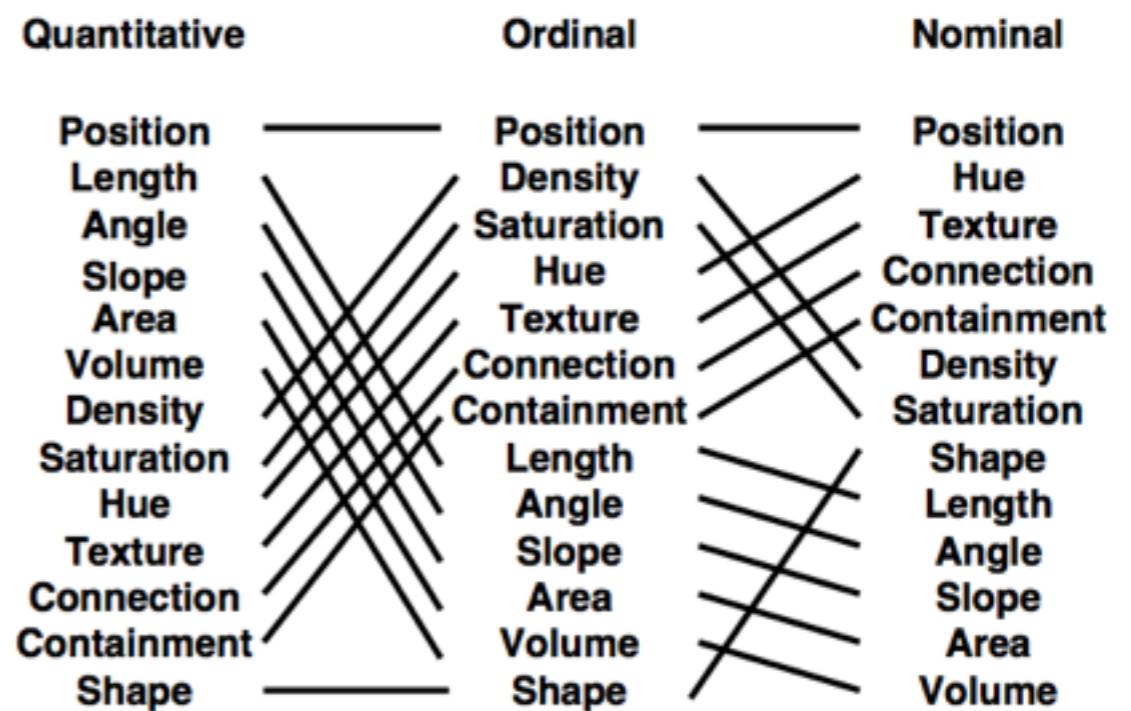


Figure 3. Graphs from position-angle experiment.

Cleveland / McGill, 1984

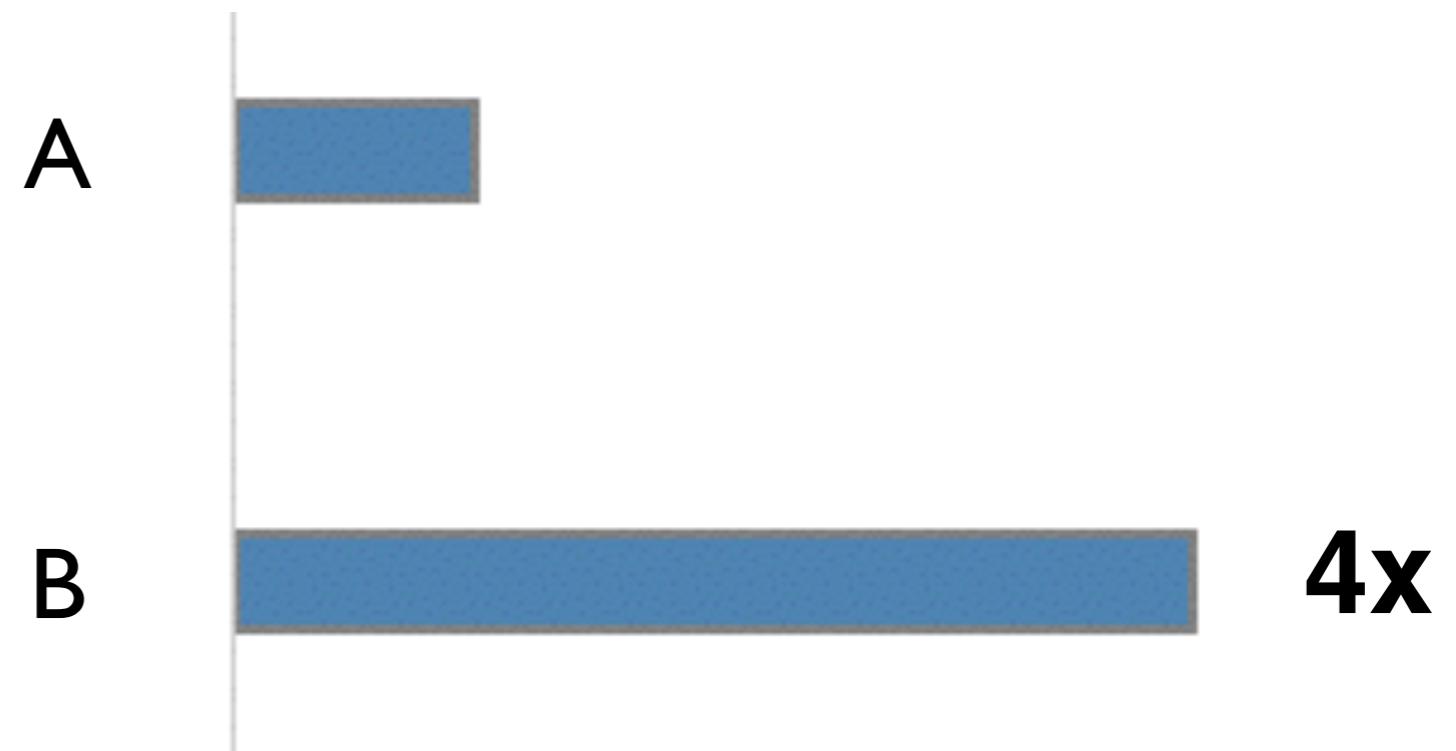


Experimental stimuli in which participants were asked to estimate what percentage the smaller value was of the larger.

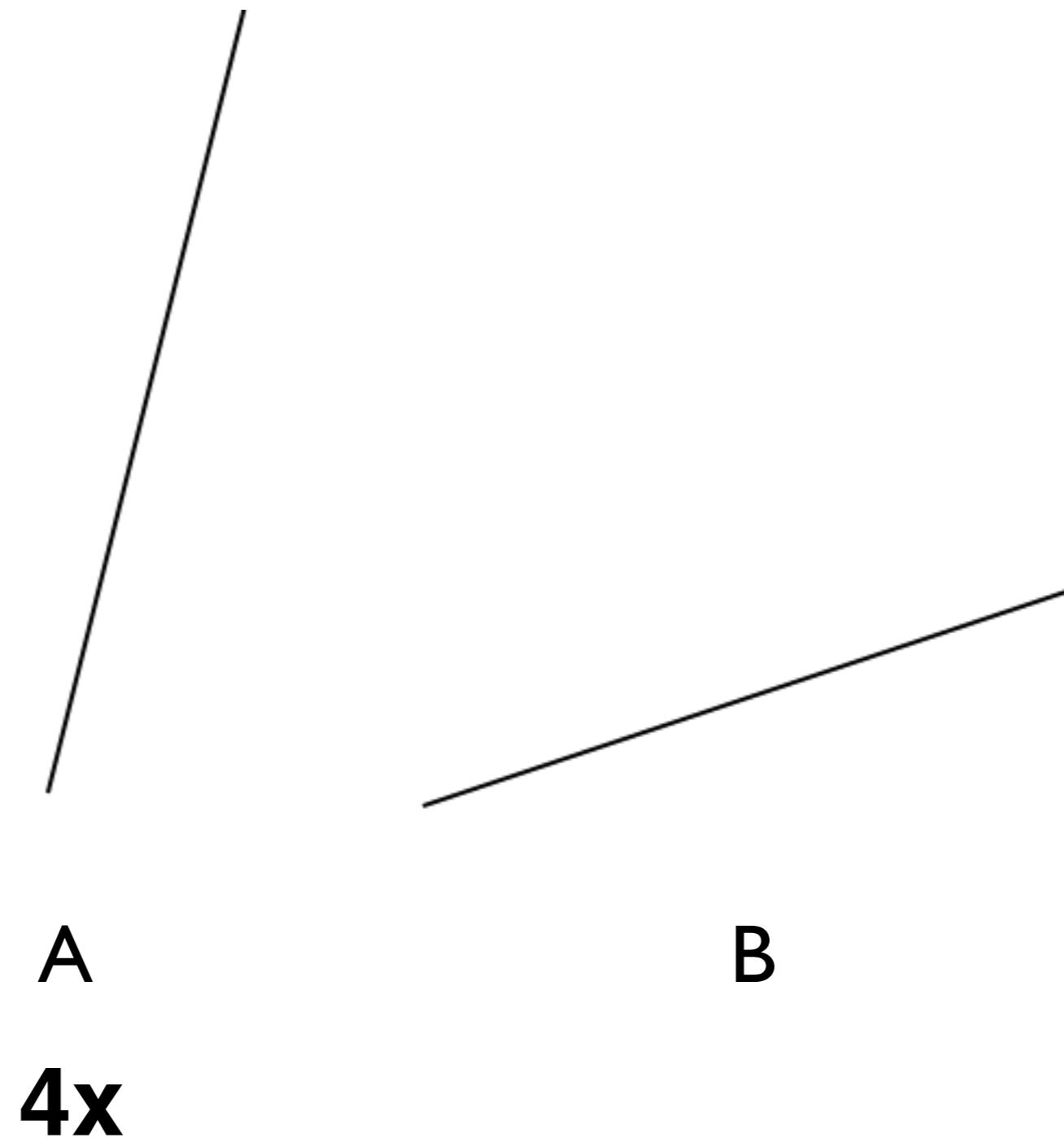


J. Mackinlay, 1986

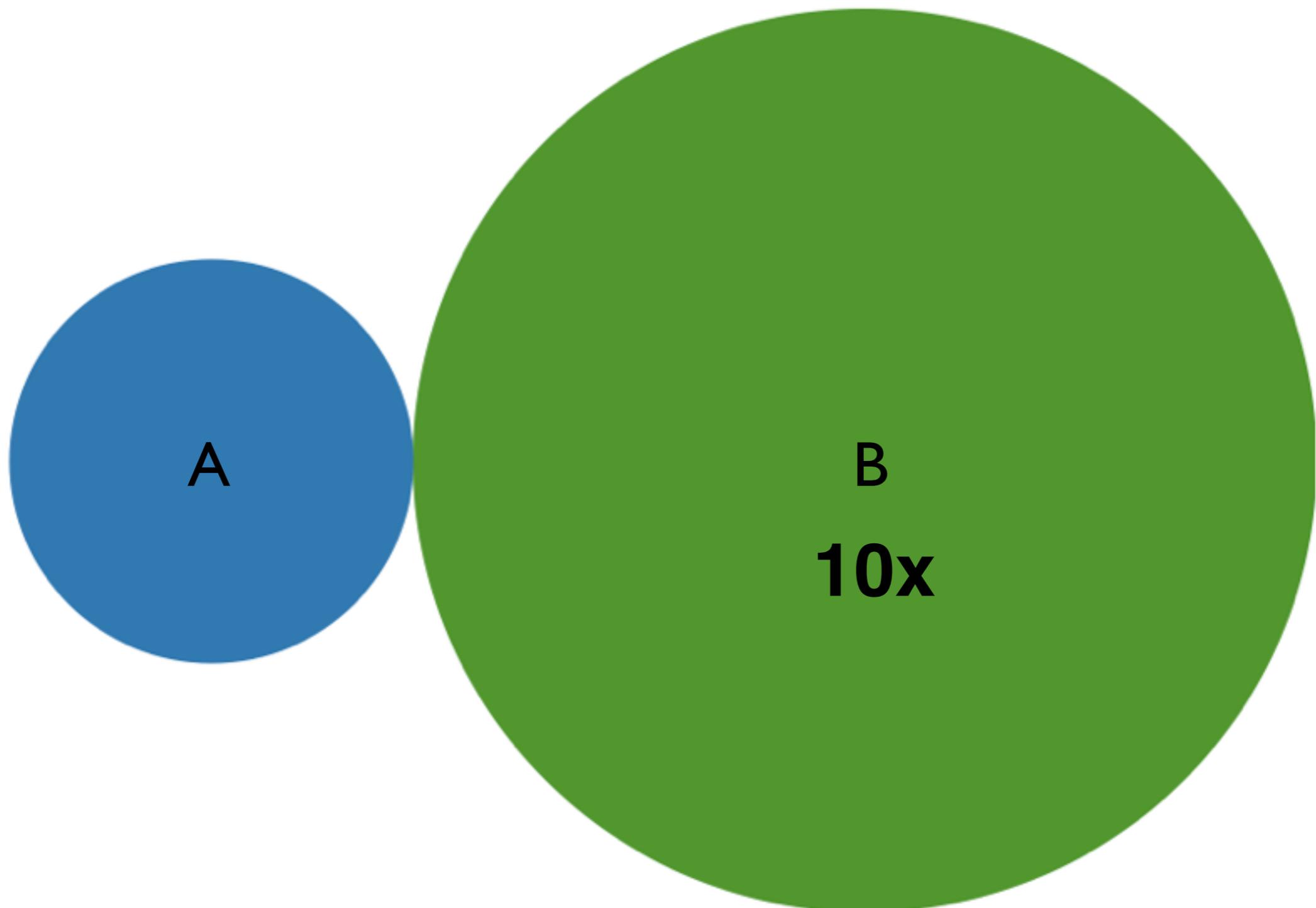
How much longer?



How much steeper slope?



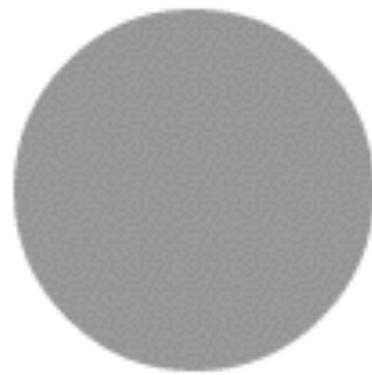
How much larger area?



How much darker?



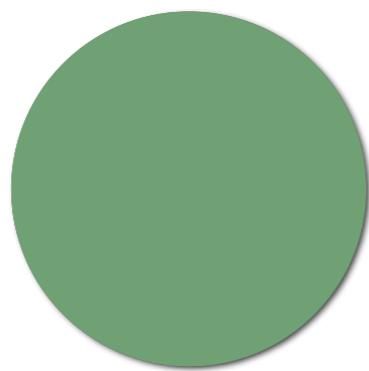
A



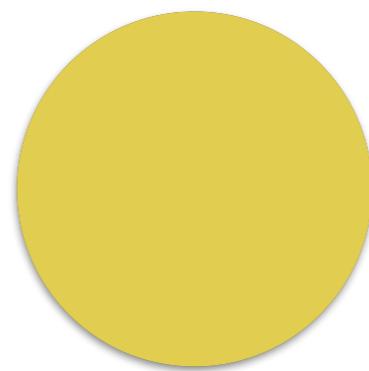
B

2x

How much bigger value?

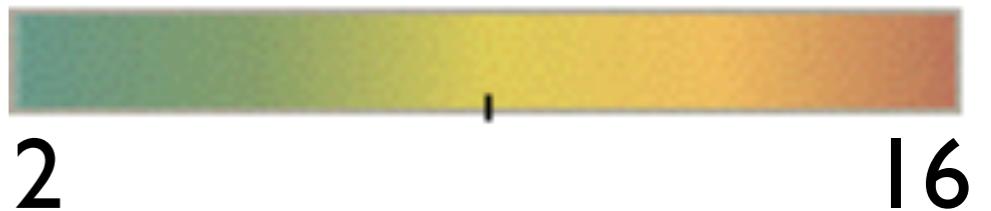


A



B

4x



Most
Efficient



Position



Length



Slope



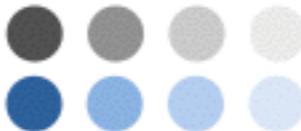
Angle



Area



Intensity



Least
Efficient

Color



Shape

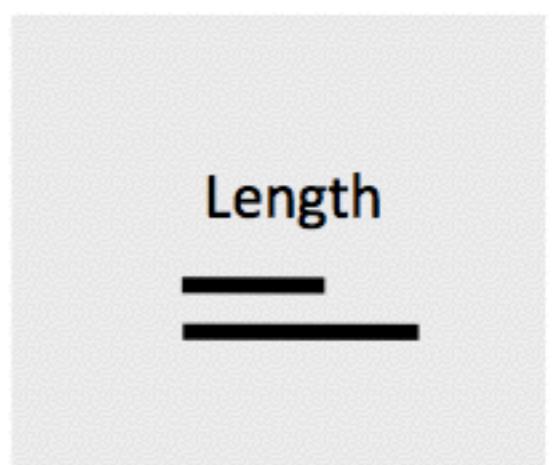
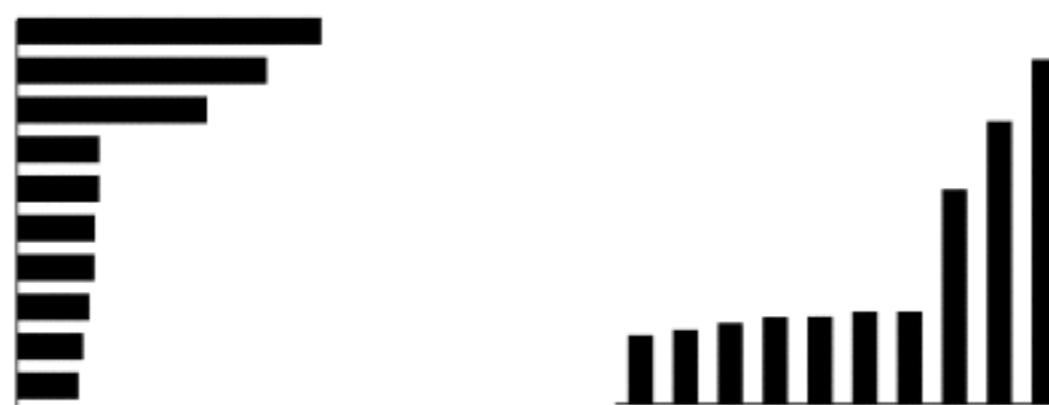
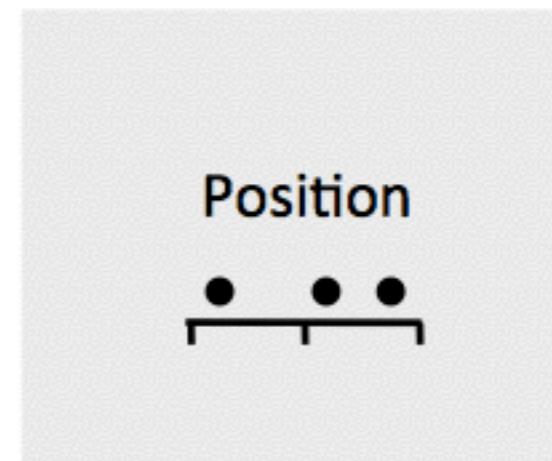
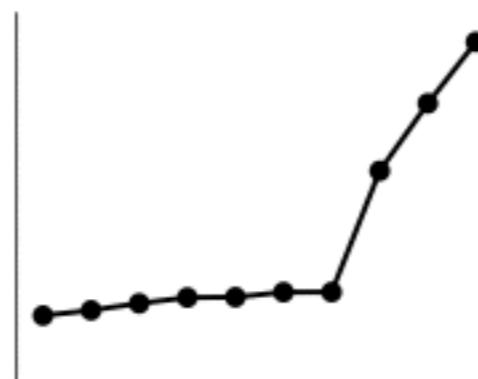


Quantitative

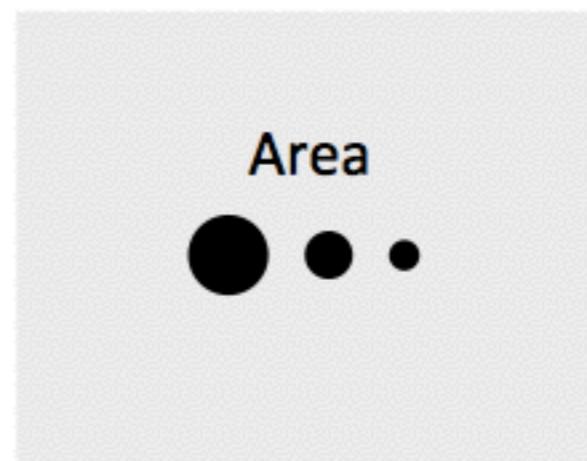
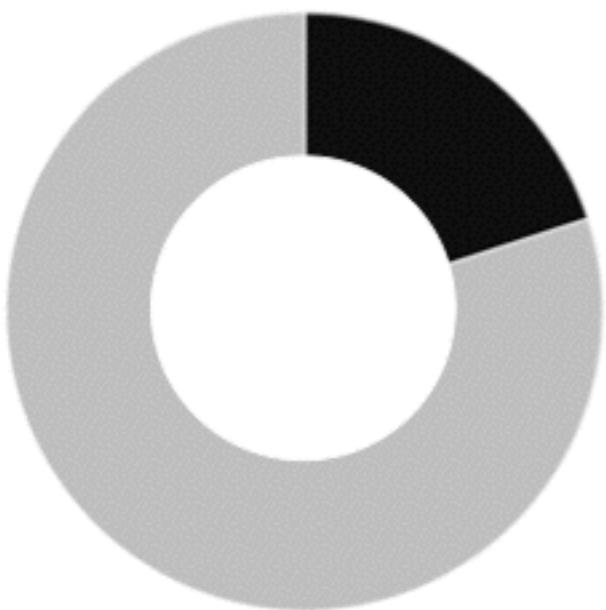
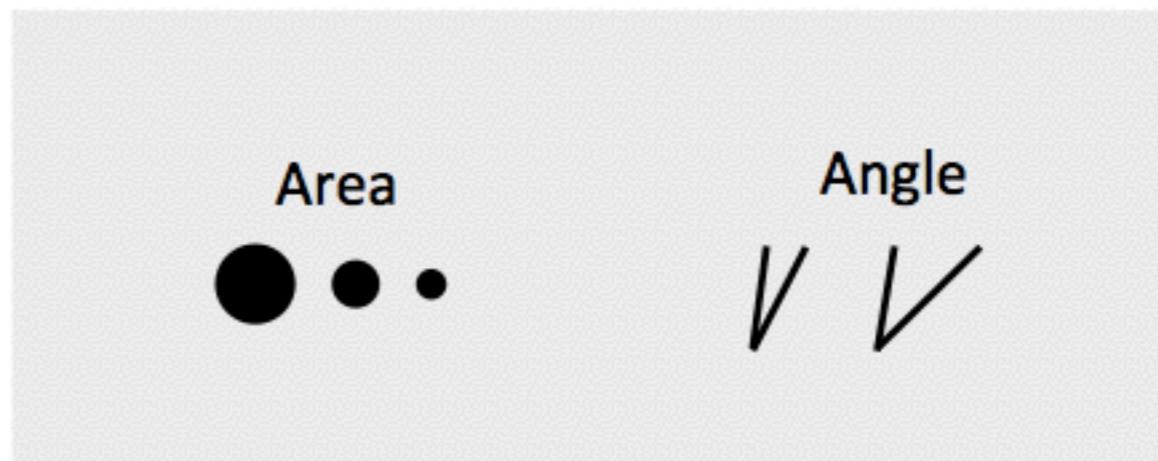
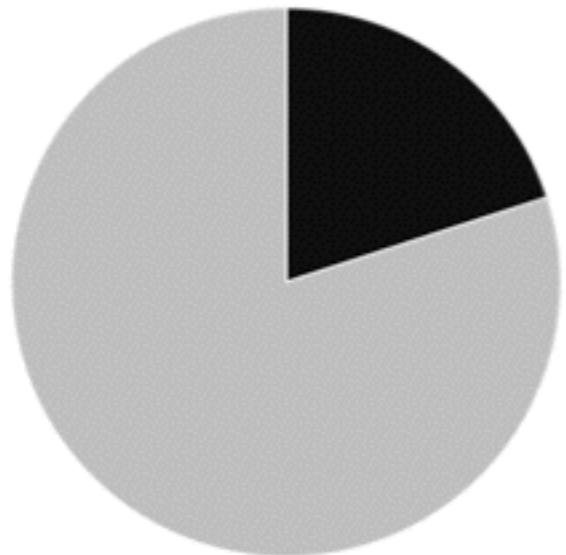
Ordered

Categories

Most Effective

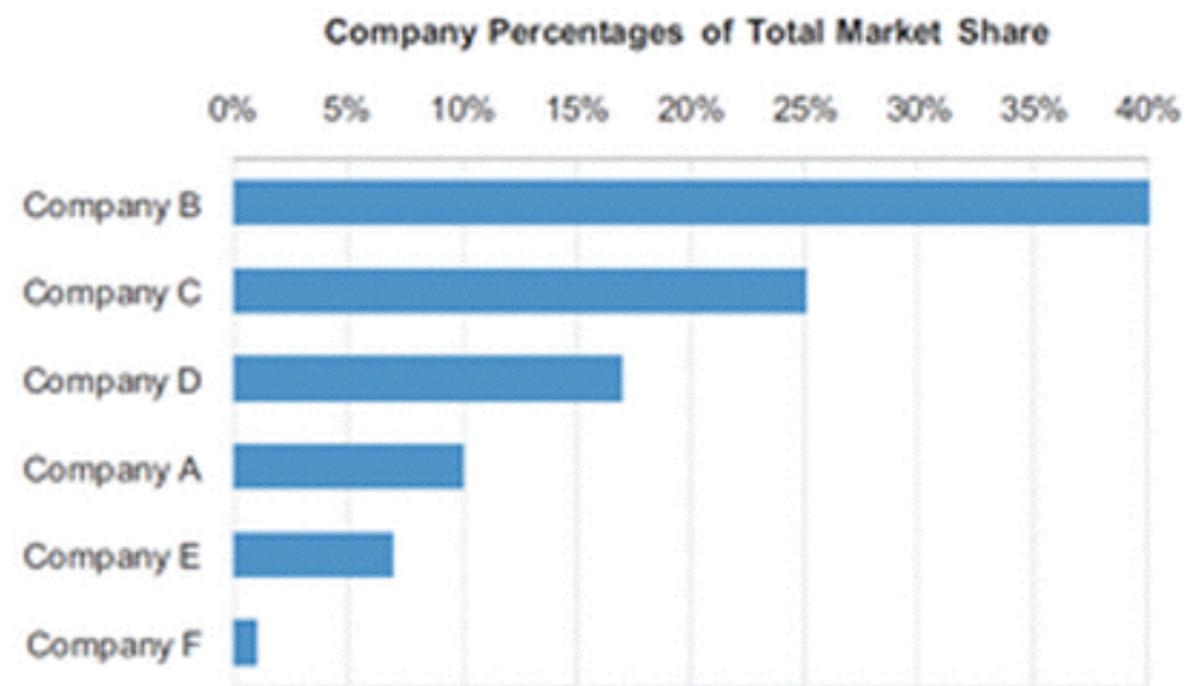
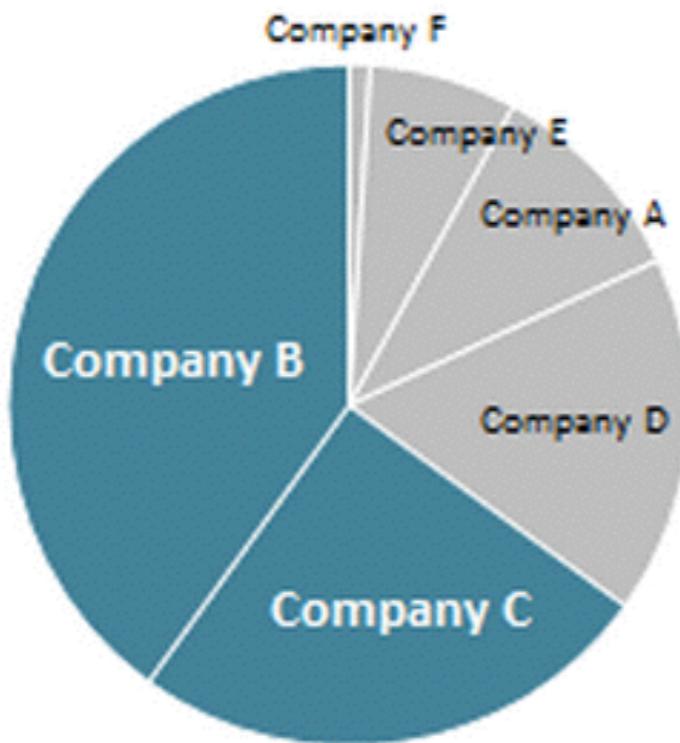


Less Effective



Pie vs. Bar Charts

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



Least Effective

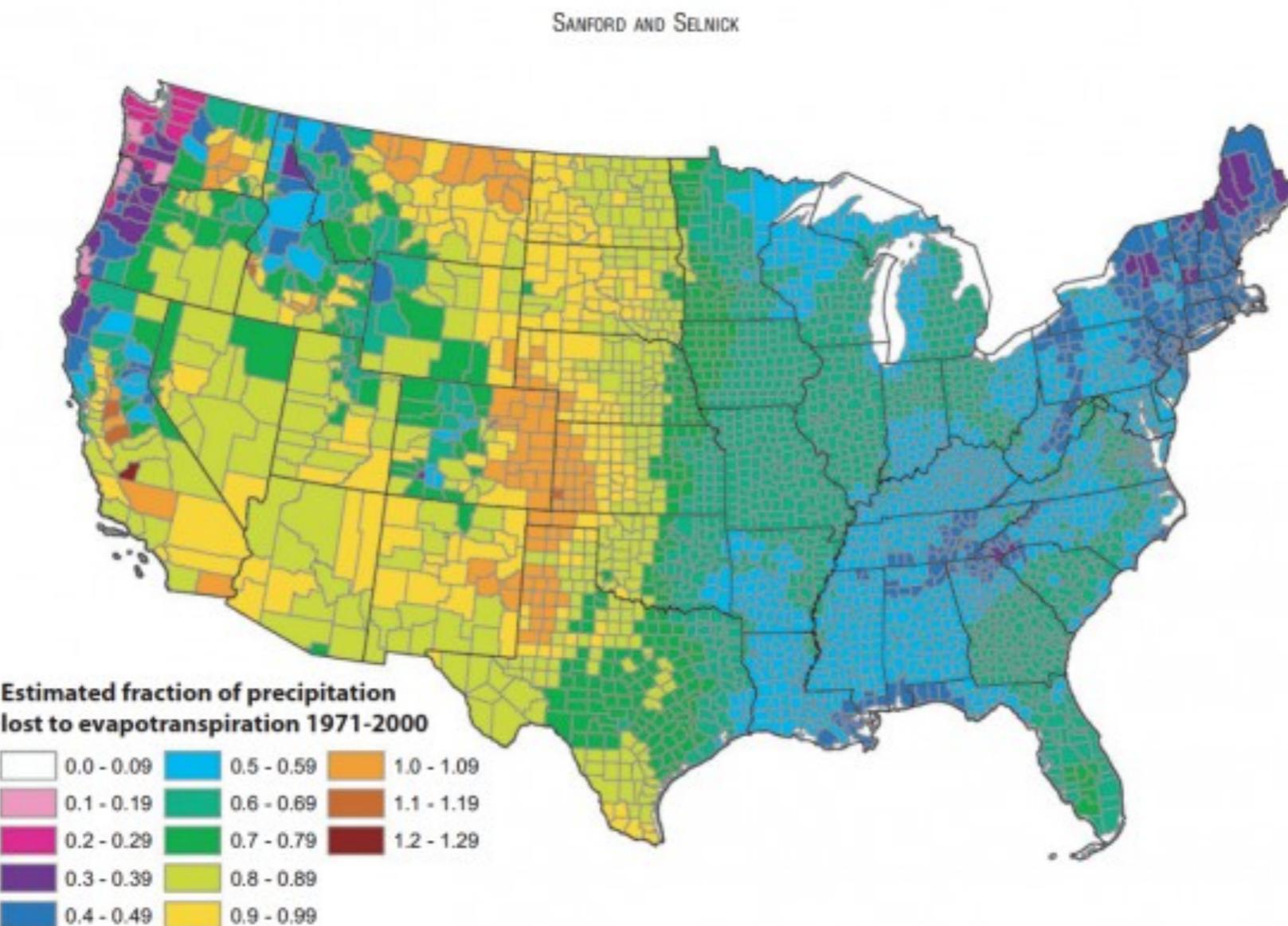
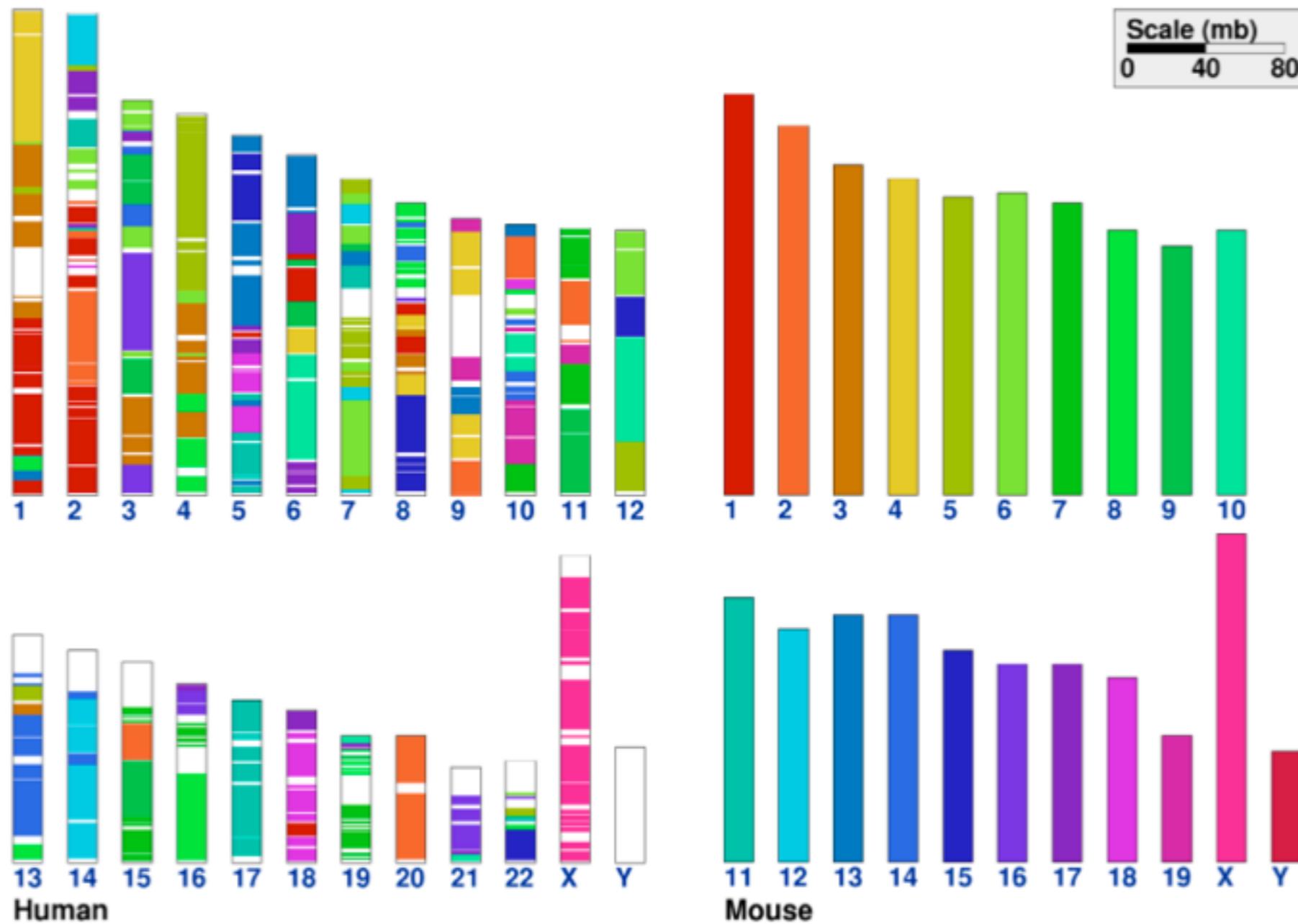


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (P) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/P were made first at the 800-m resolution of the PRISM 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.

Use Color Strategically

Color Discriminability



Colors for Categories

Do not use more than 5-8 colors at once



Colors for Ordinal Data

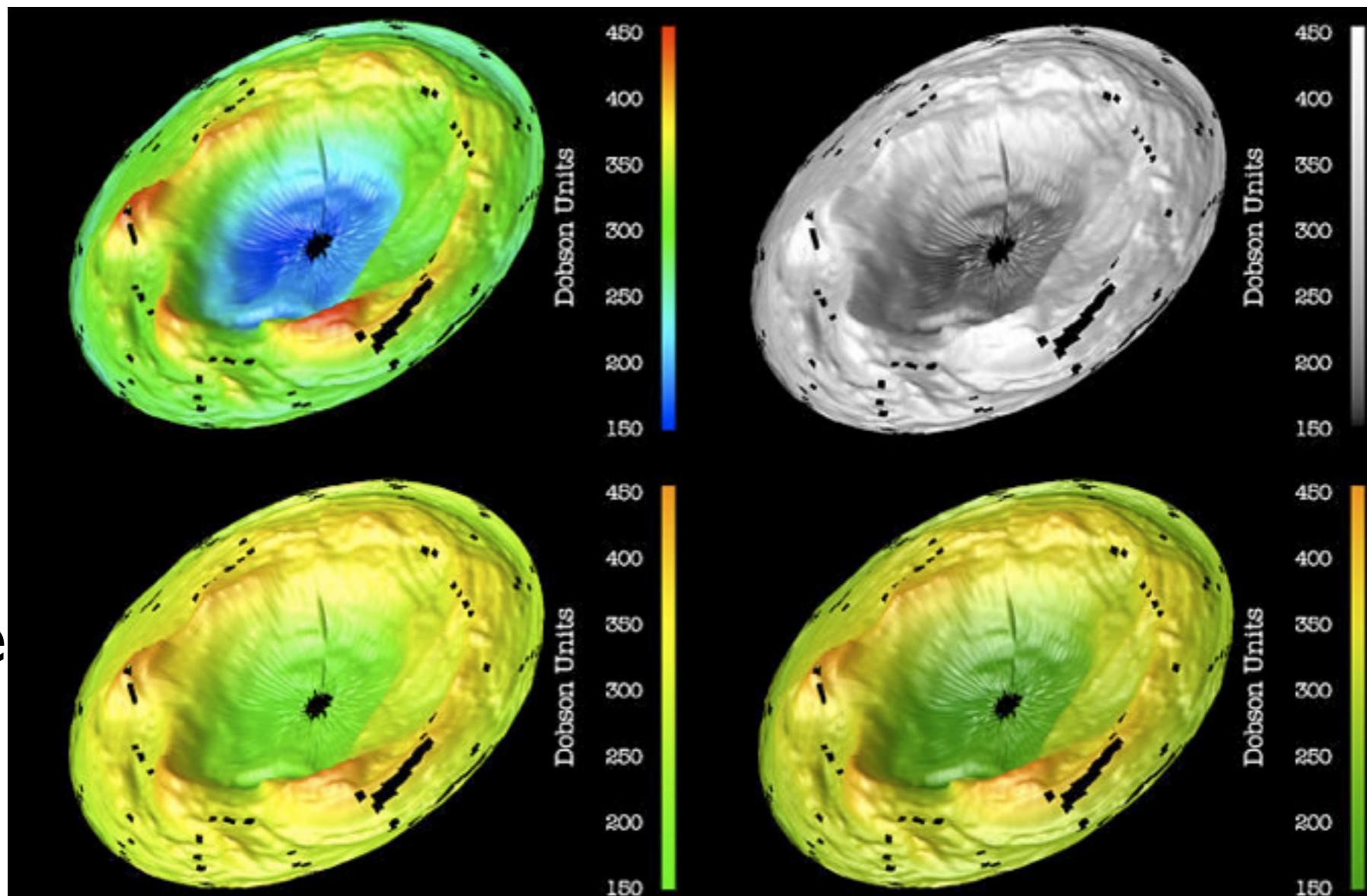
Vary luminance and saturation



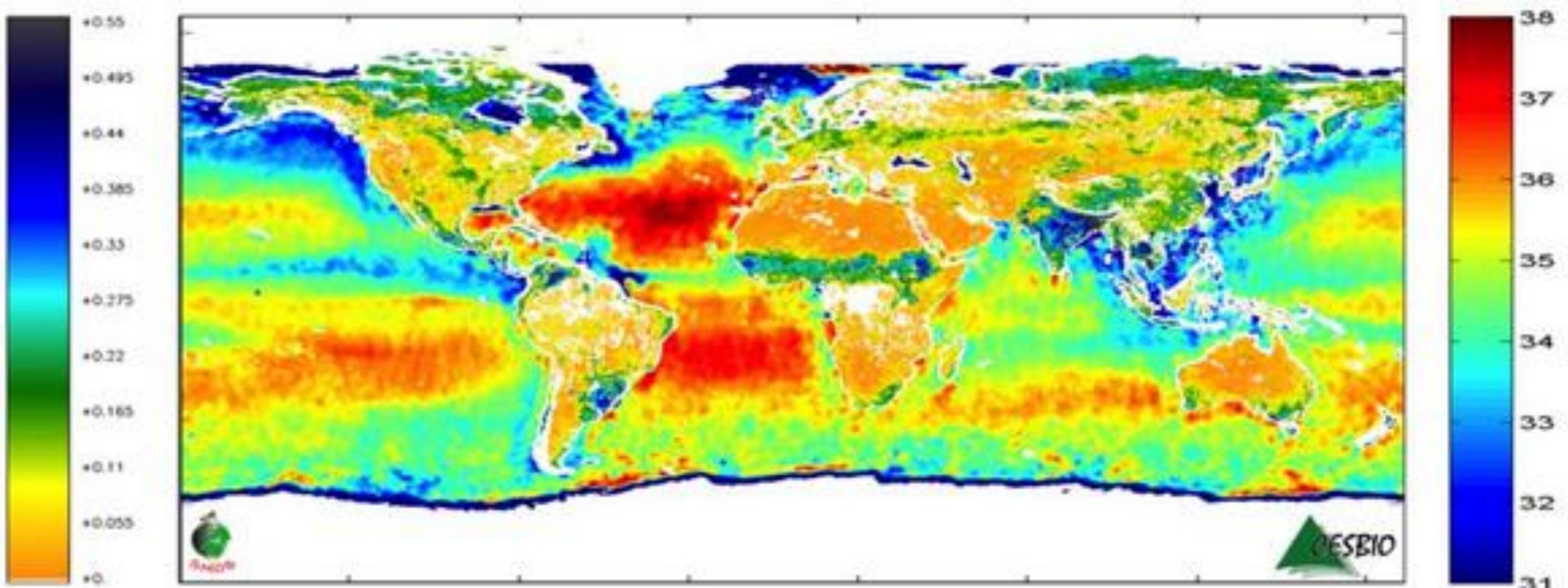
Zeilis et al, 2009, "Escaping RGBland: Selecting
Colors for Statistical Graphics"

Colors for Quantitative Data

Hue
(Rainbow)

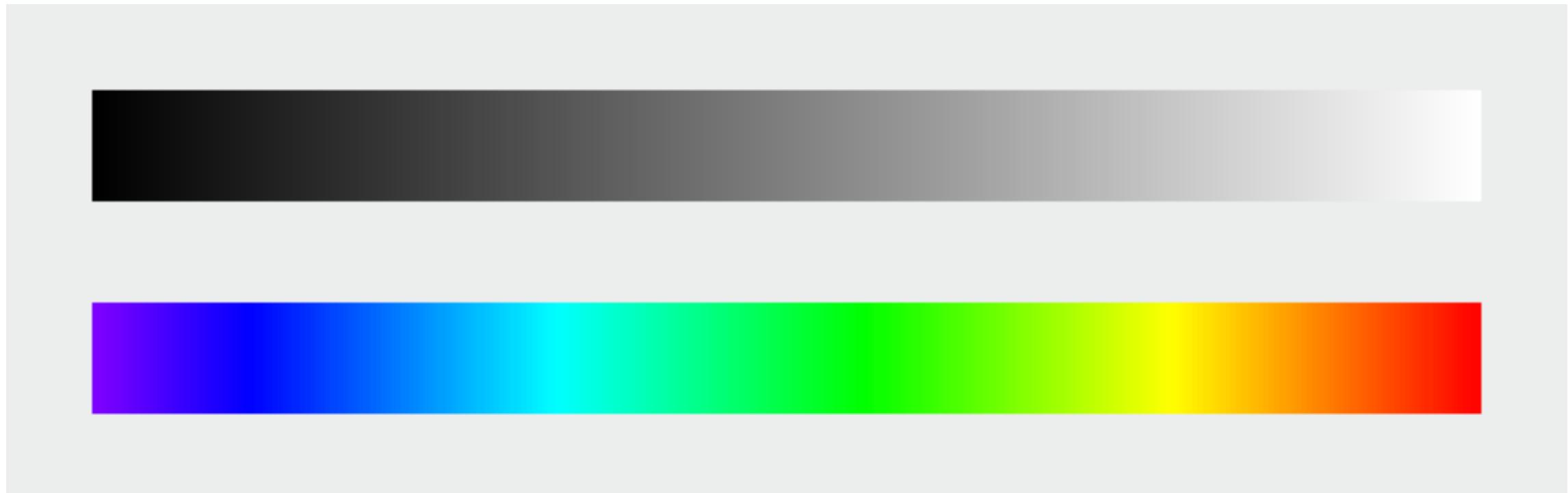


Rainbow Colormap

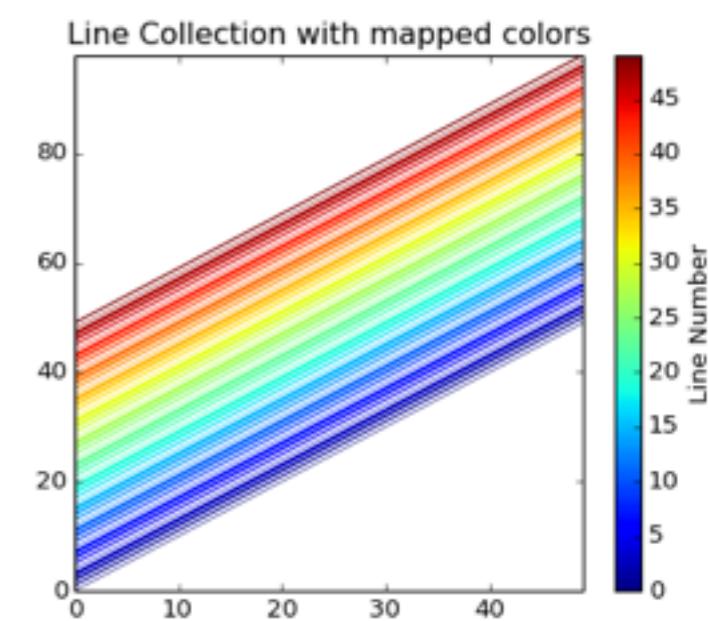
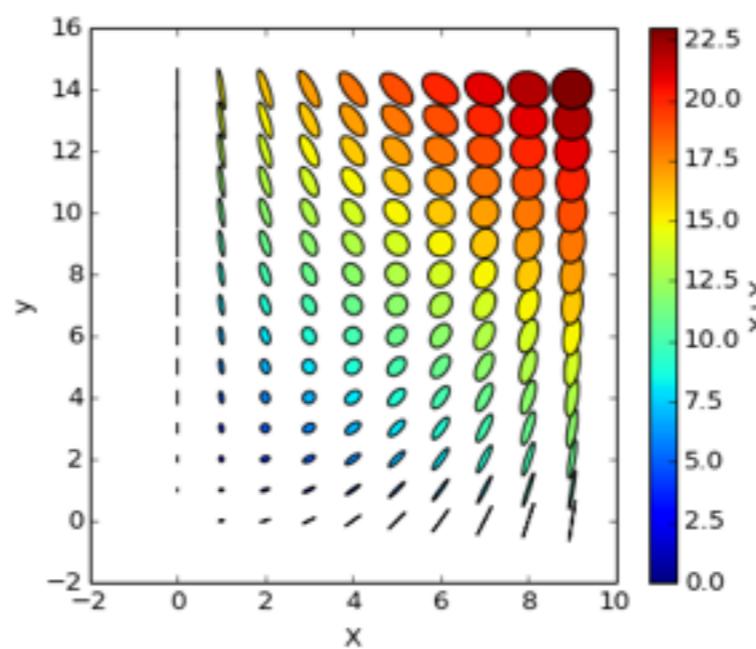
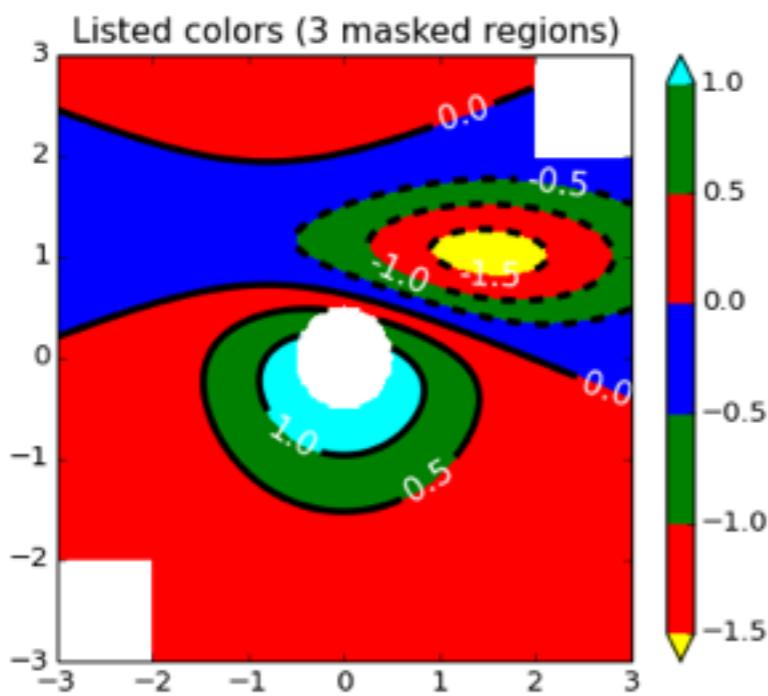
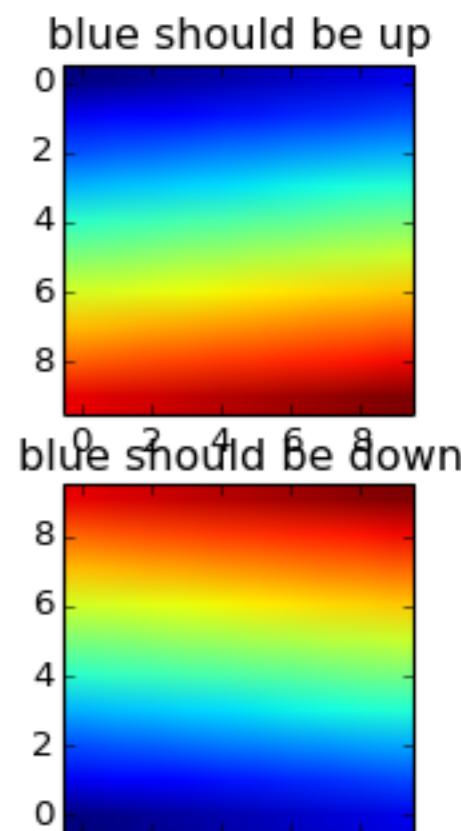
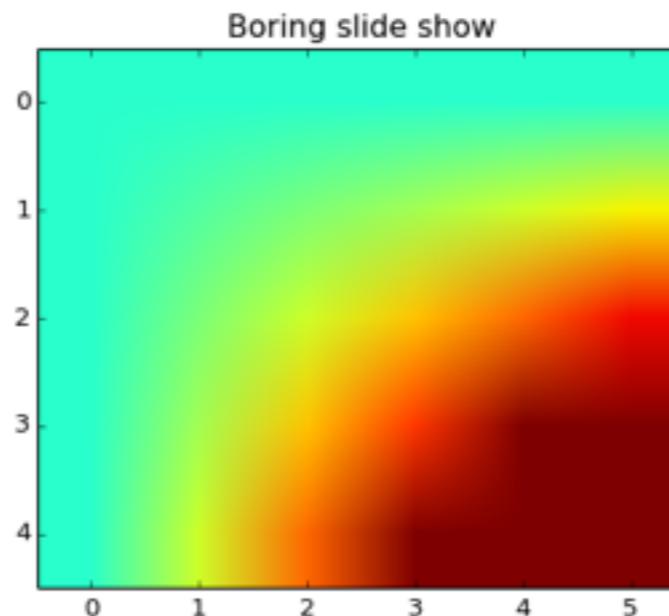
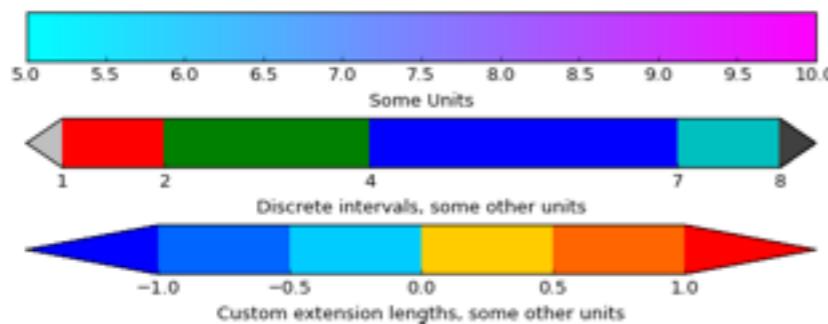


Rainbow Colormap

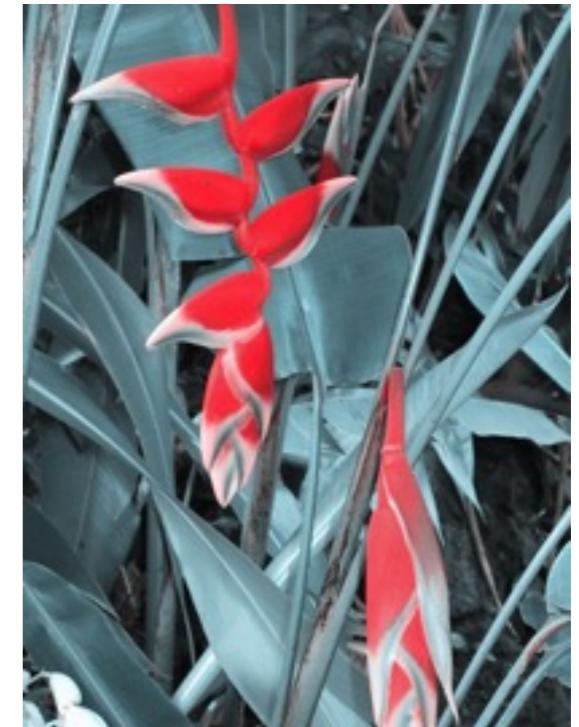
Perceptually nonlinear



Avoid Rainbow Colors!



Color Blindness



Protanope

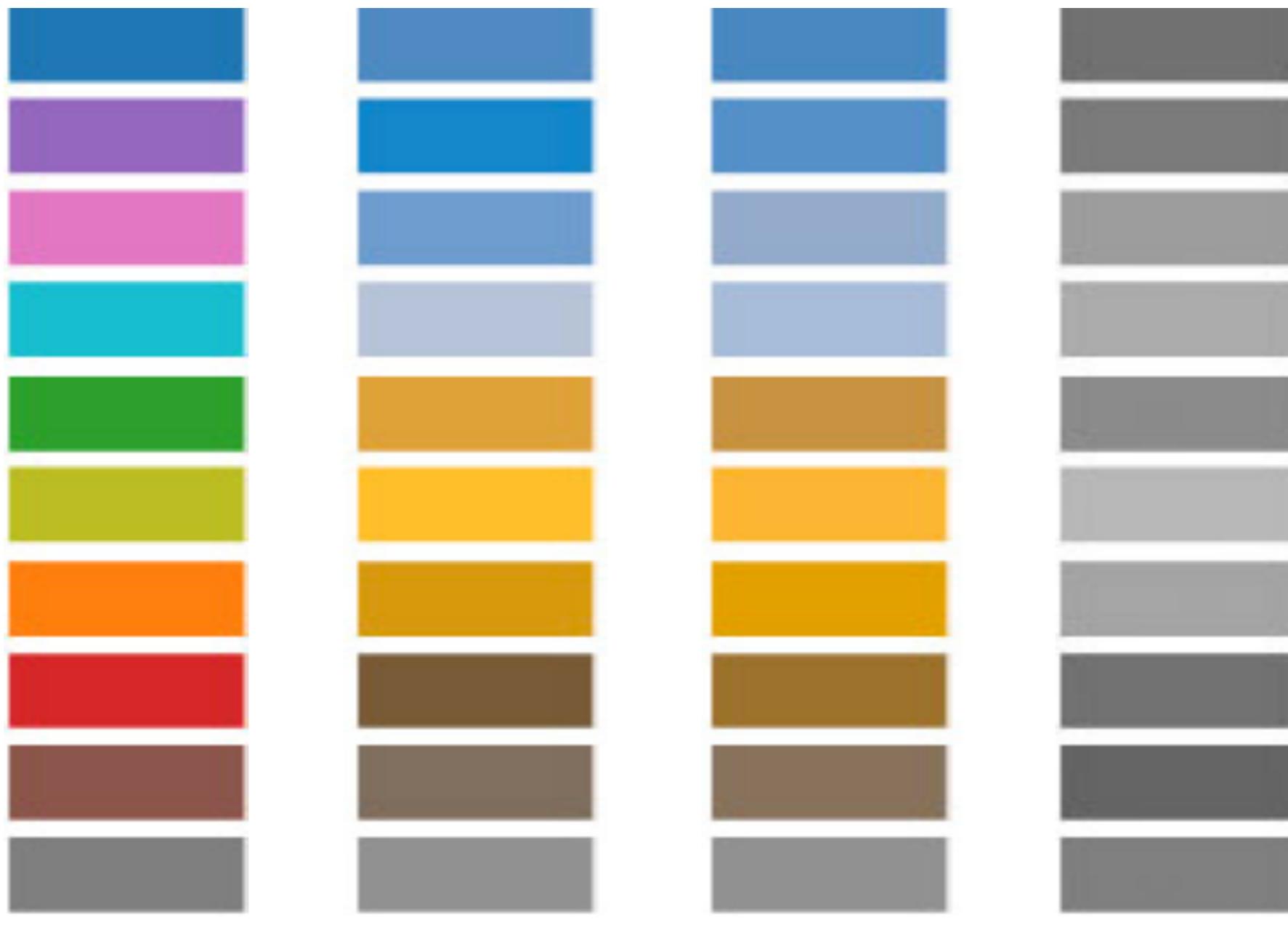
Deuteranope

Tritanope

Red / green
deficiencies

Blue / Yellow
deficiency

Color Blindness



Normal

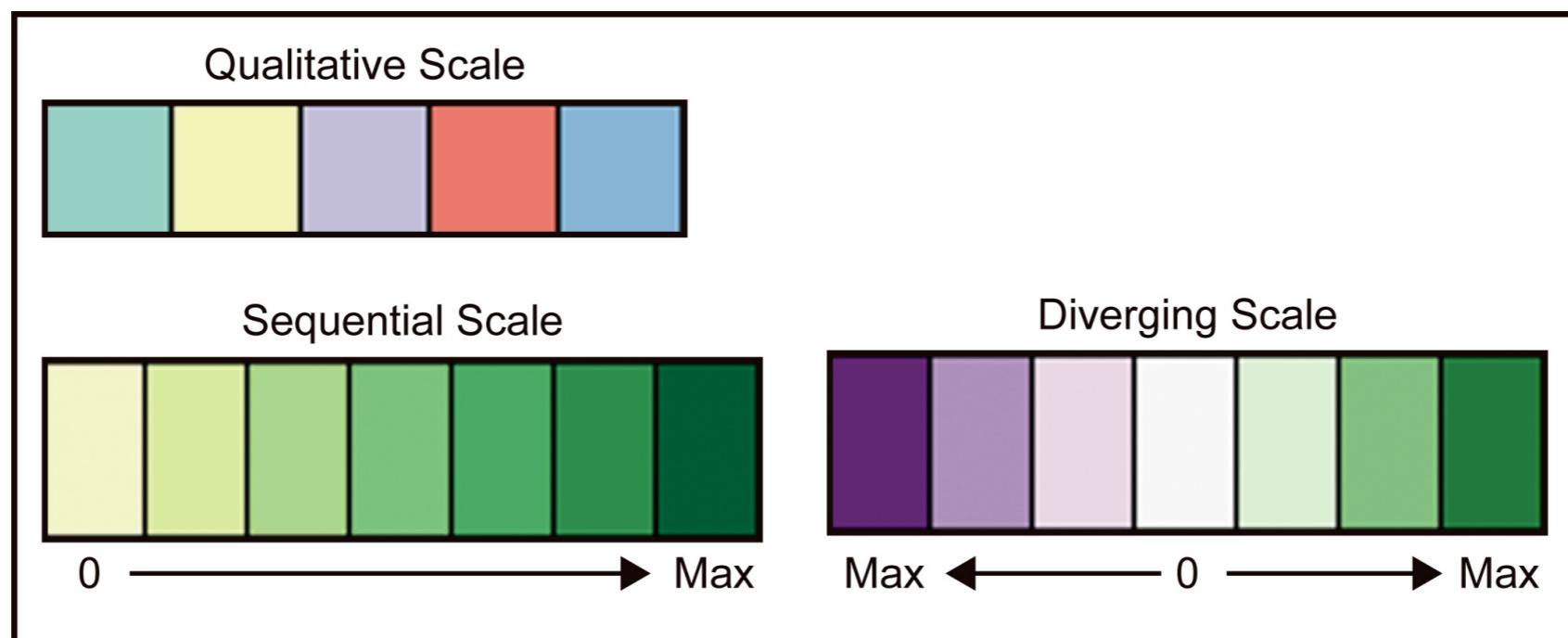
Protanope

Deuteranope

Lightness

Color Brewer

Nominal



Ordinal

number of data classes on your map

3 | [learn more >](#)

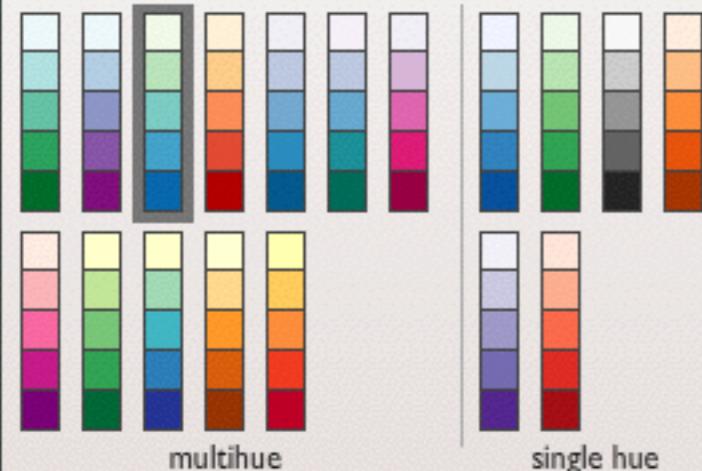
[how to use](#) | [updates](#) | [credits](#)

COLORBREWER 2.0
color advice for cartography

the nature of your data

sequential | [learn more >](#)

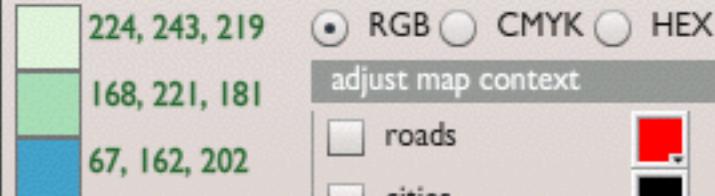
pick a color scheme: GnBu



(optional) only show schemes that are:

- colorblind safe print friendly
 photocopy-able [learn more >](#)

pick a color system



RGB CMYK HEX

adjust map context

- roads 
 cities 
 borders 

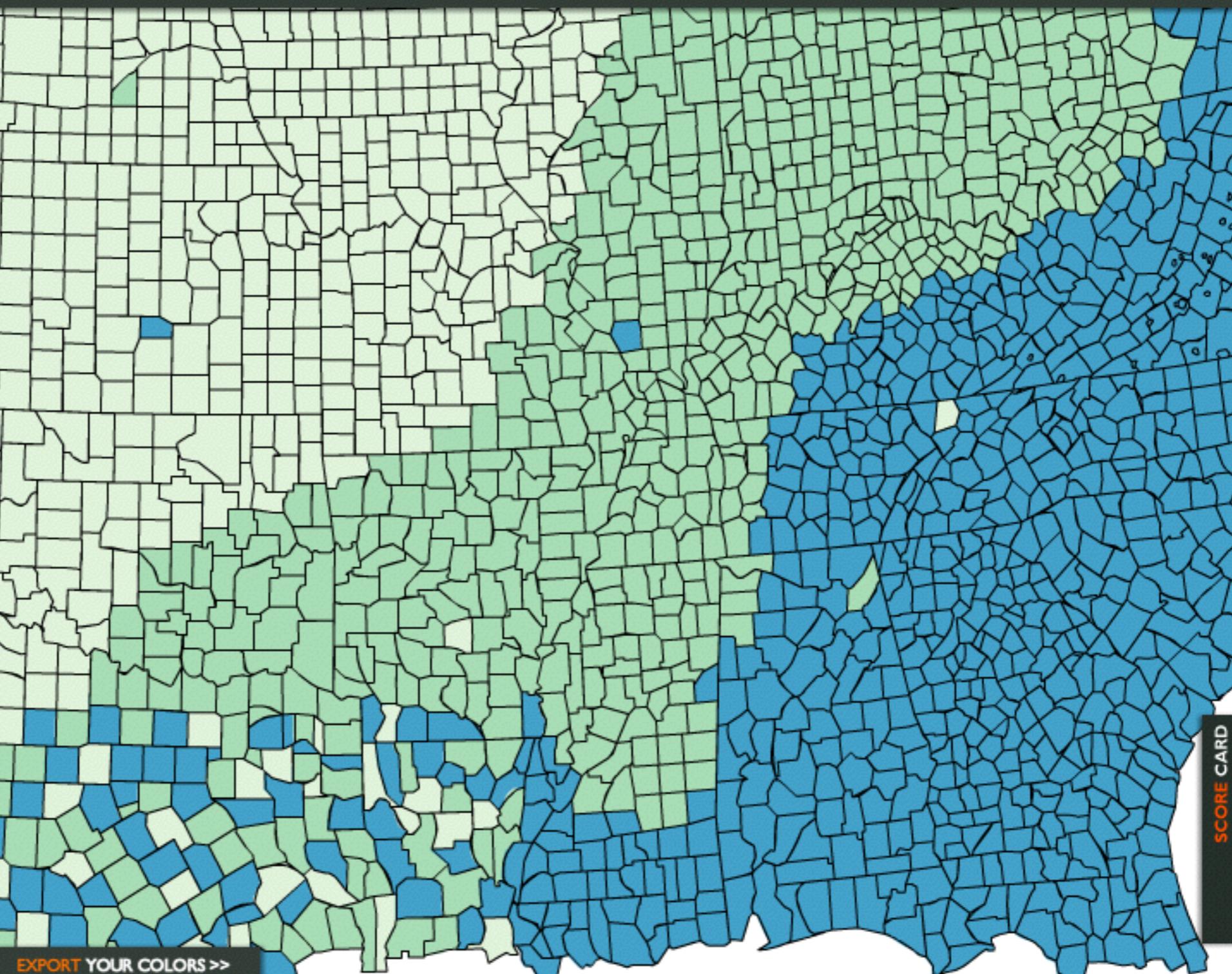
select a background

- solid color 
 terrain 

color transparency


[learn more >](#)

[EXPORT YOUR COLORS >>](#)

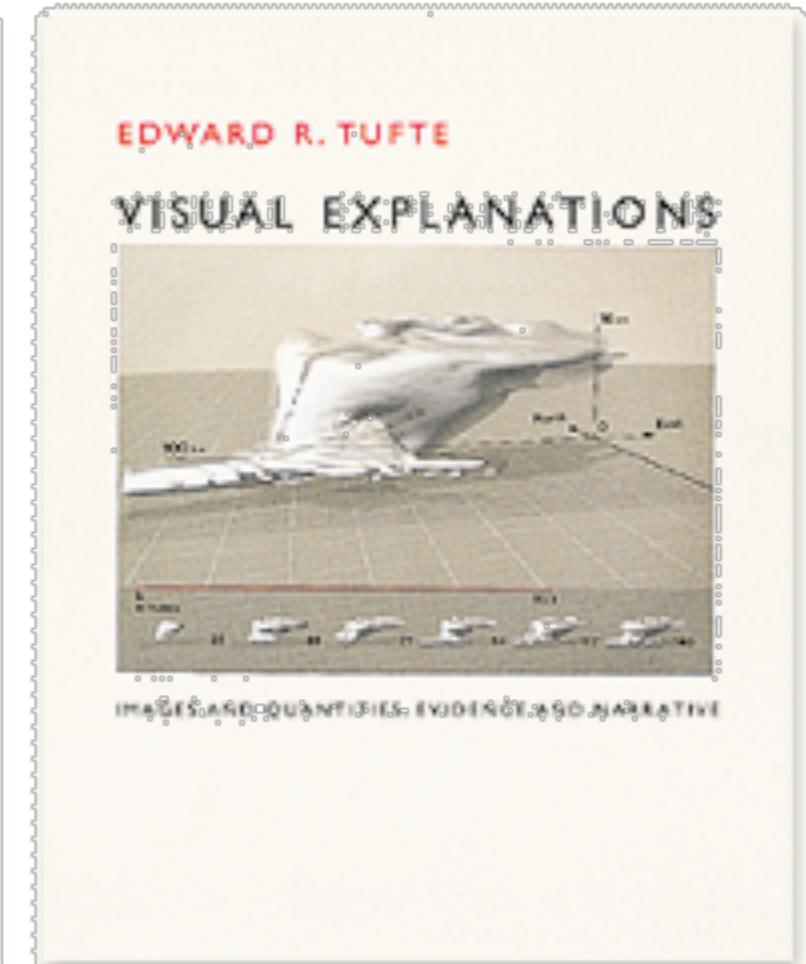
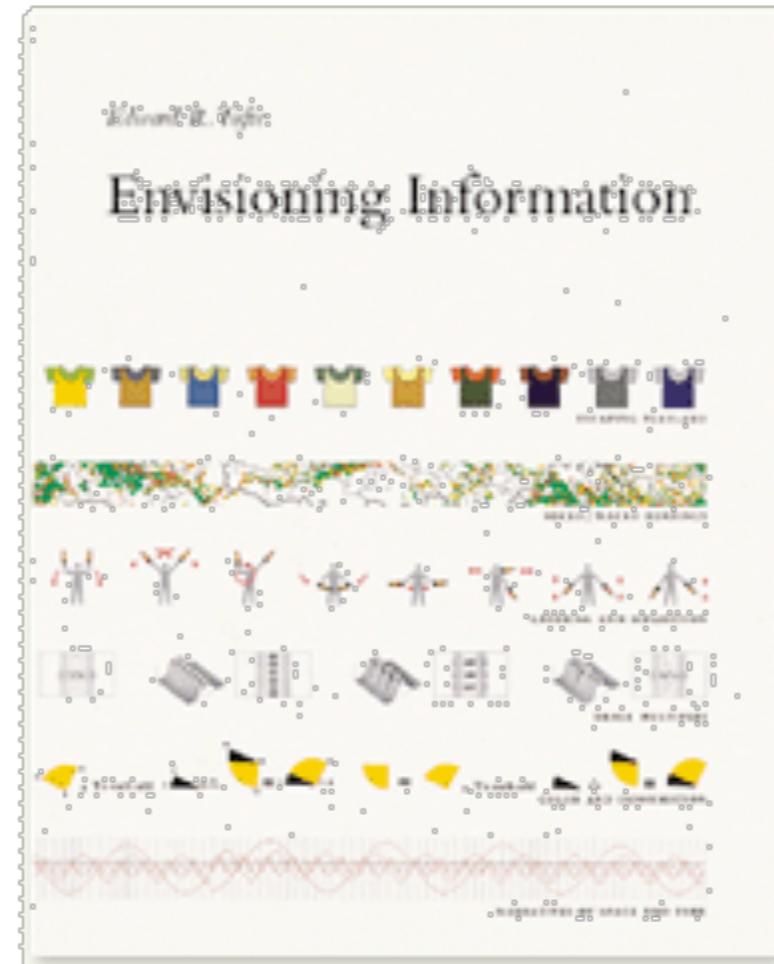
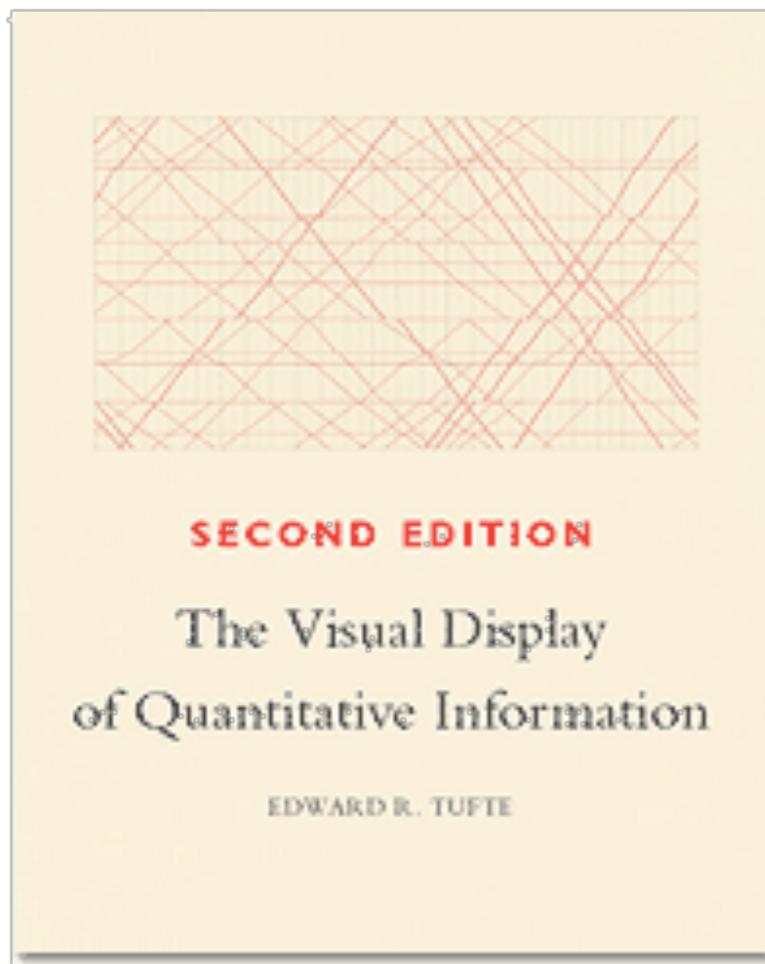


Effective Visualizations

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

