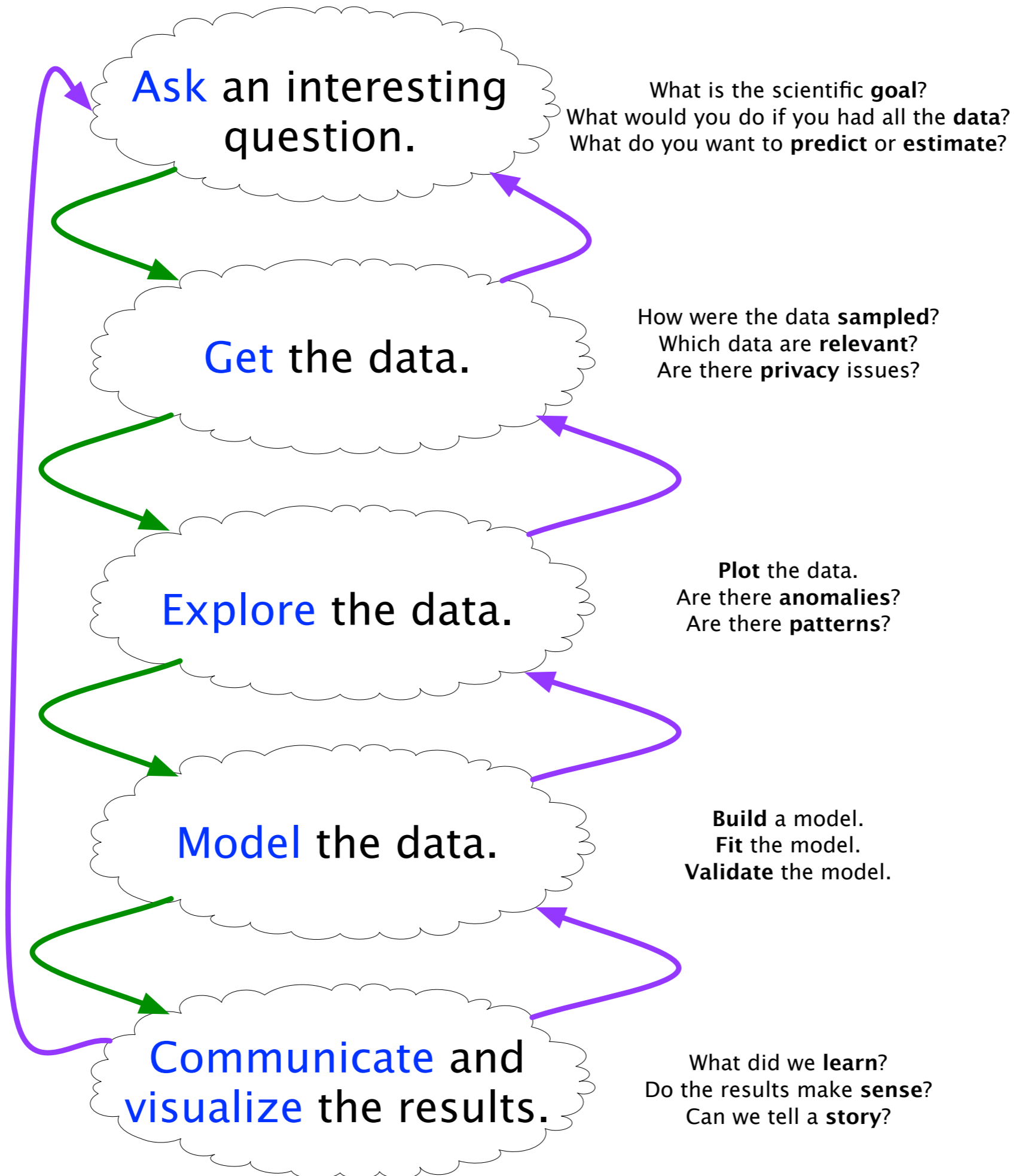
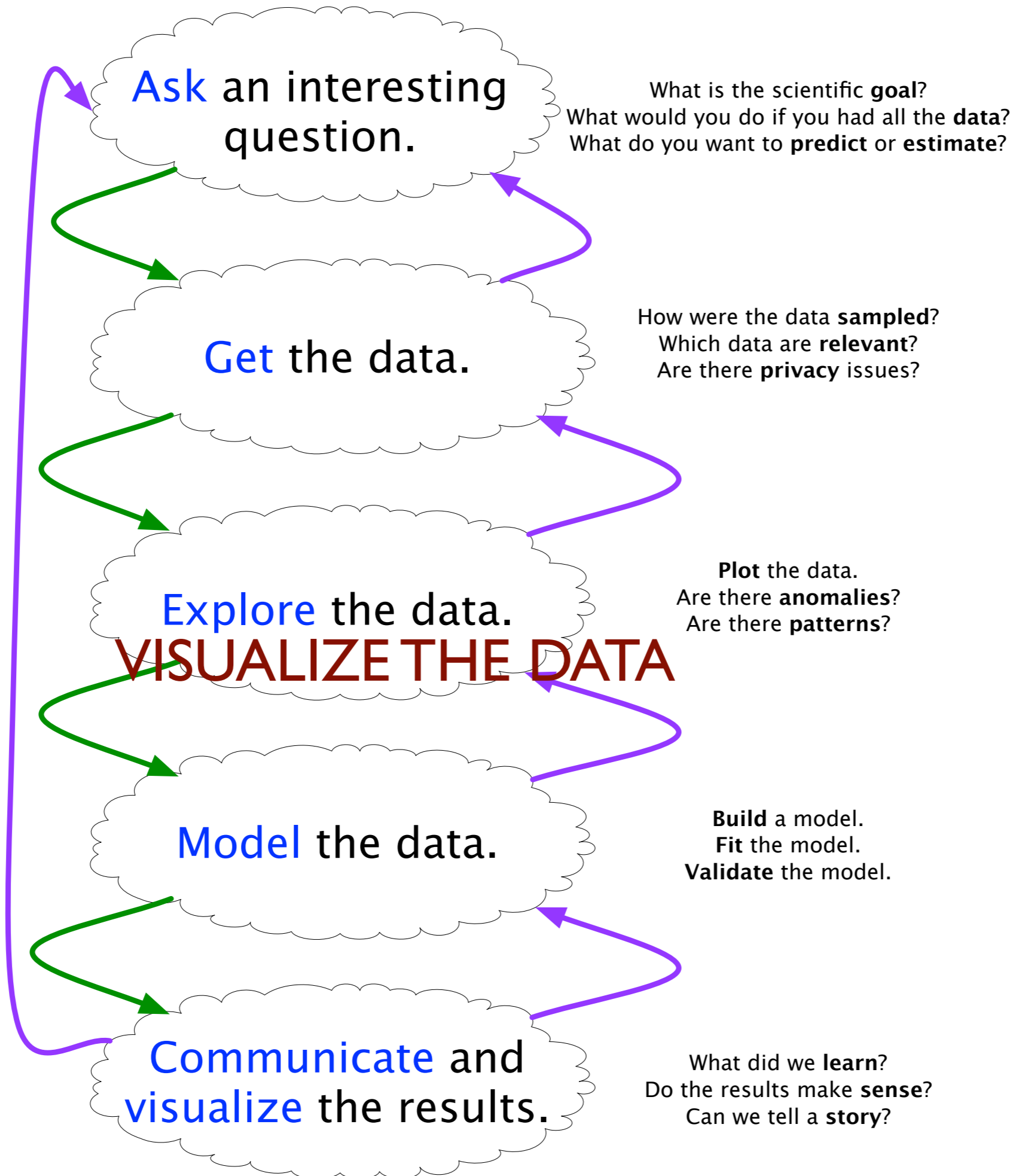


CS 109a: Data Science

Effective Exploratory Data Analysis and Visualization

Pavlos Protopapas, Kevin Rader, Rahul Dave, Margo Levine

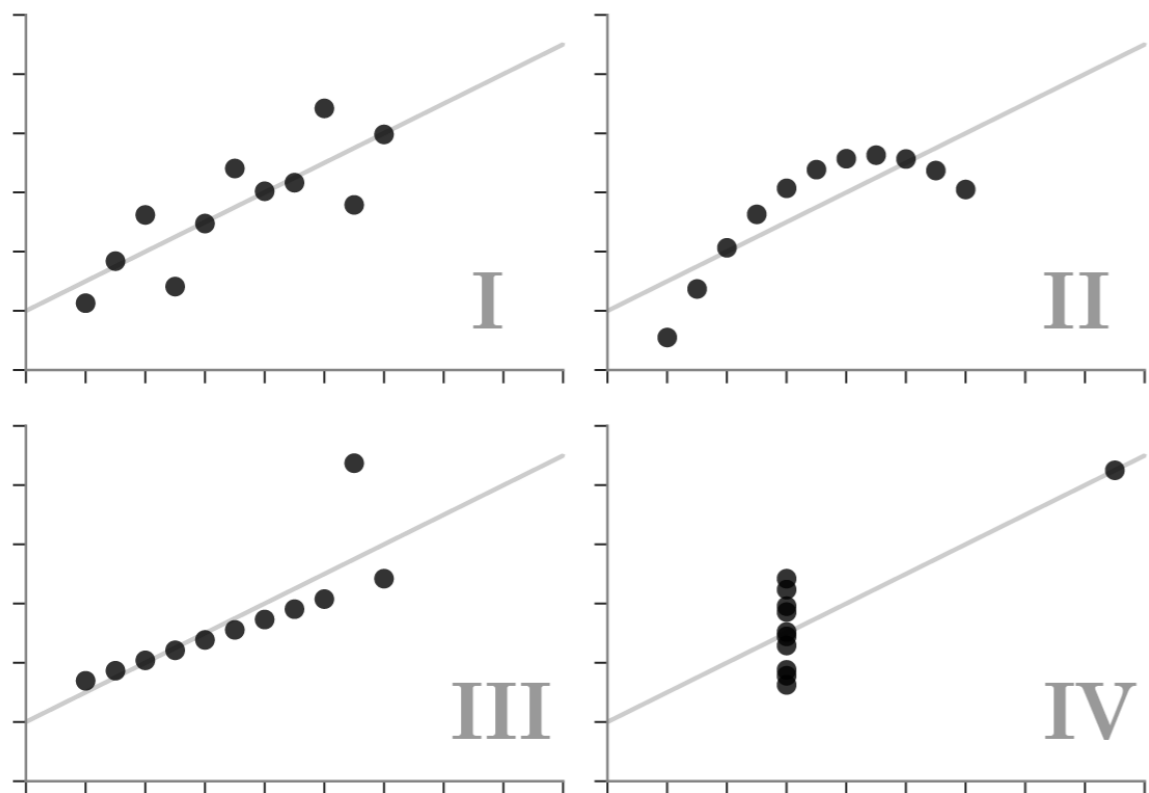






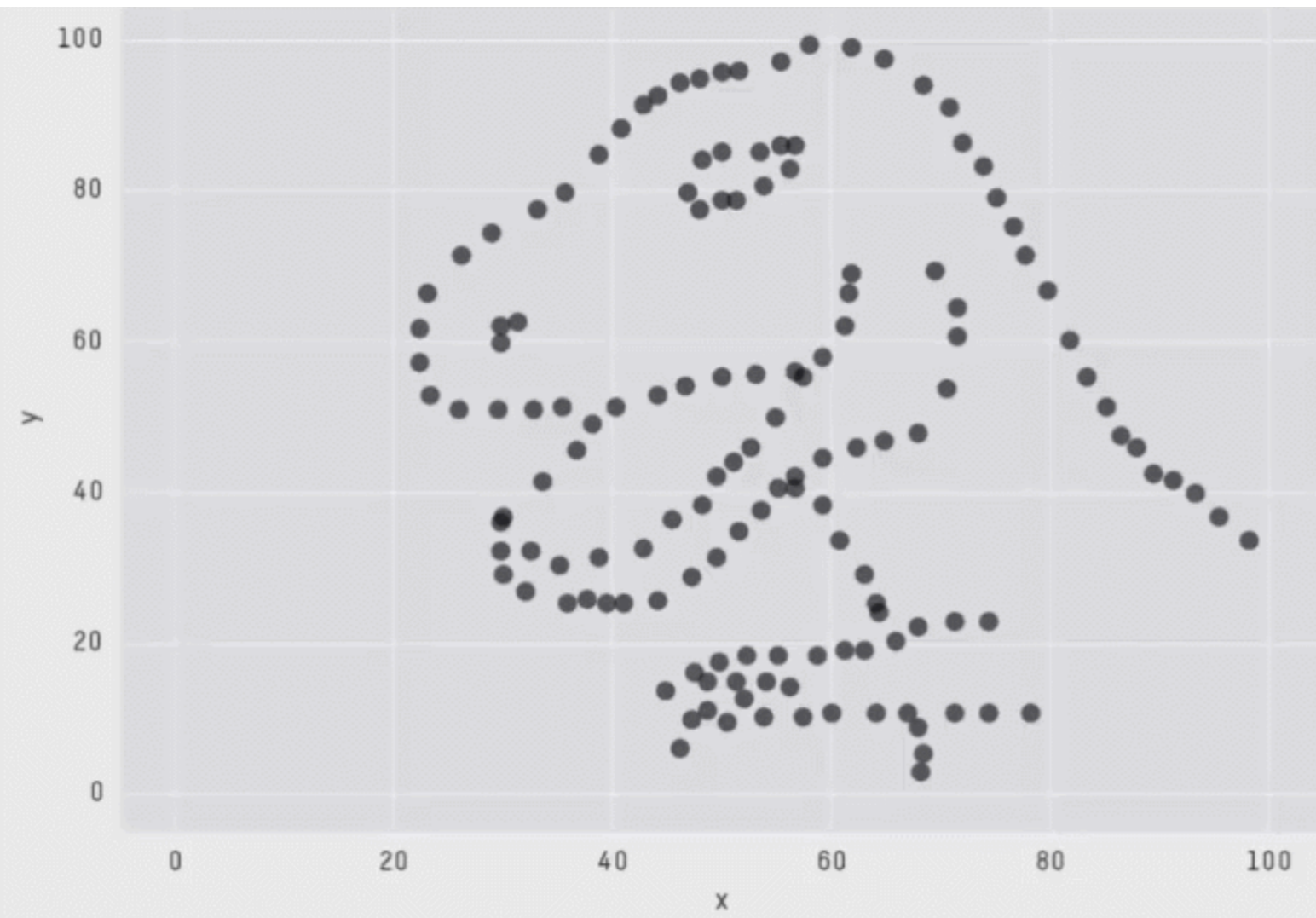
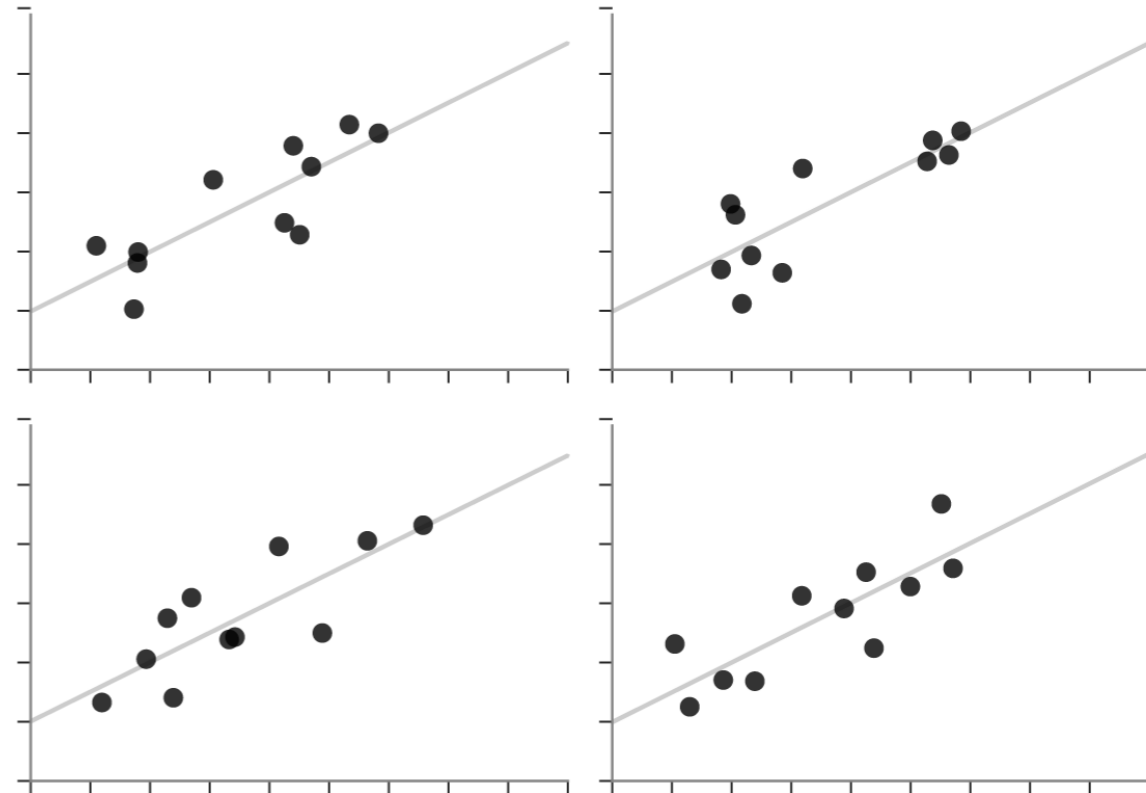
Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.



Unstructured Quartet

Each dataset here also has the same summary statistics. However, they are not *clearly different* or *visually distinct*.

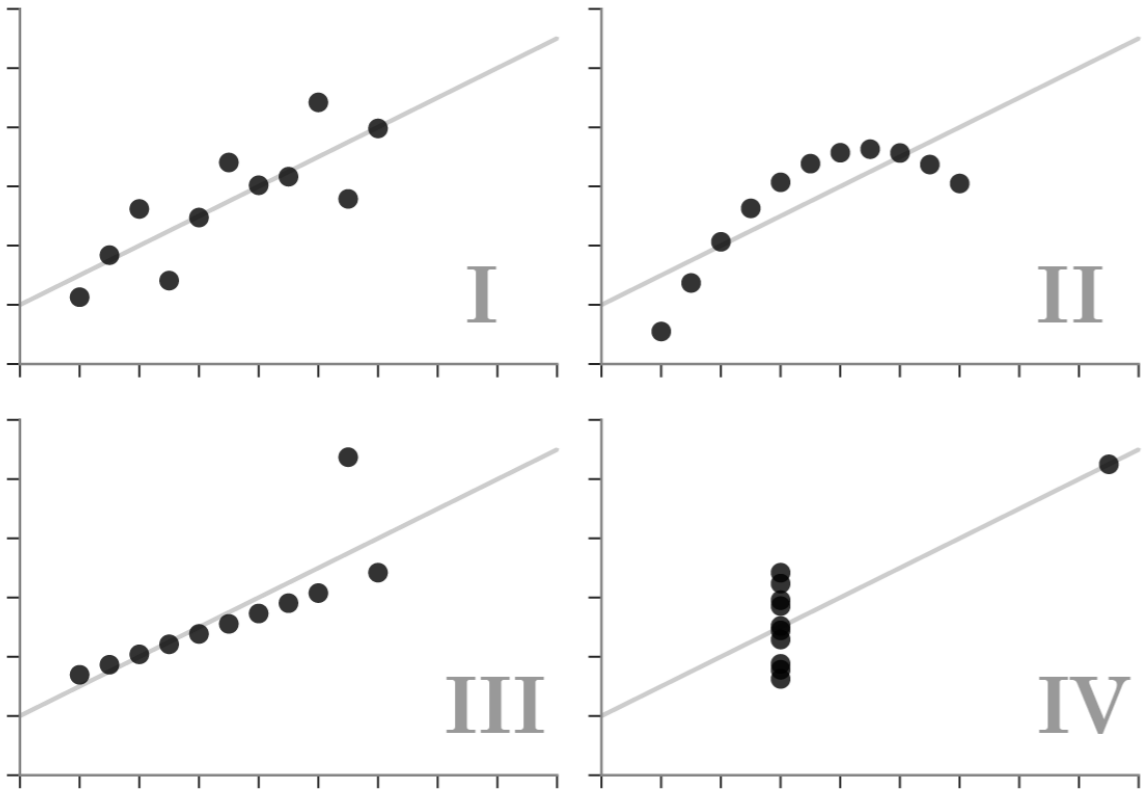


X Mean: 54.2659224
 Y Mean: 47.8313999
 X SD : 16.7649829
 Y SD : 26.9342120
 Corr. : -0.0642526



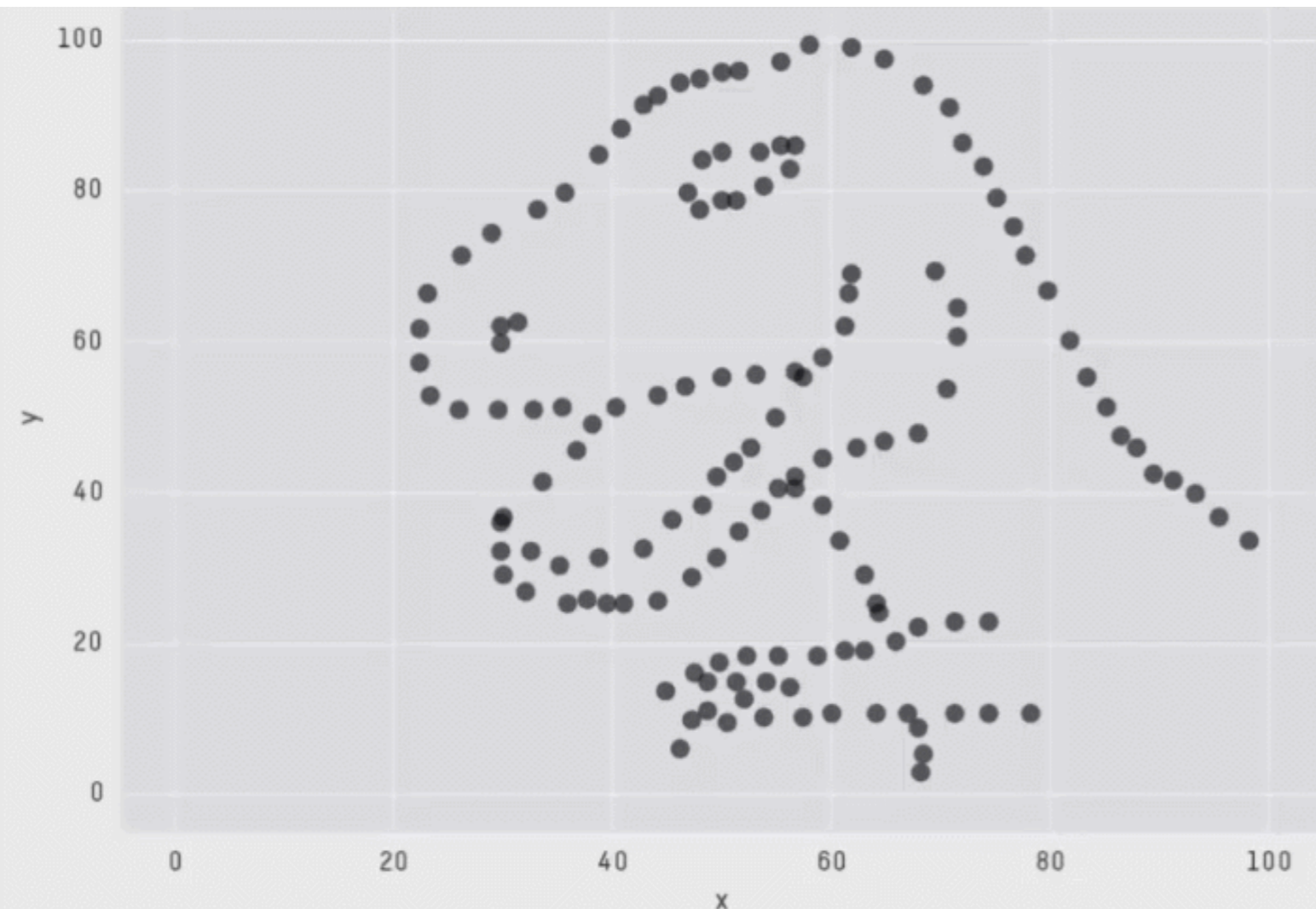
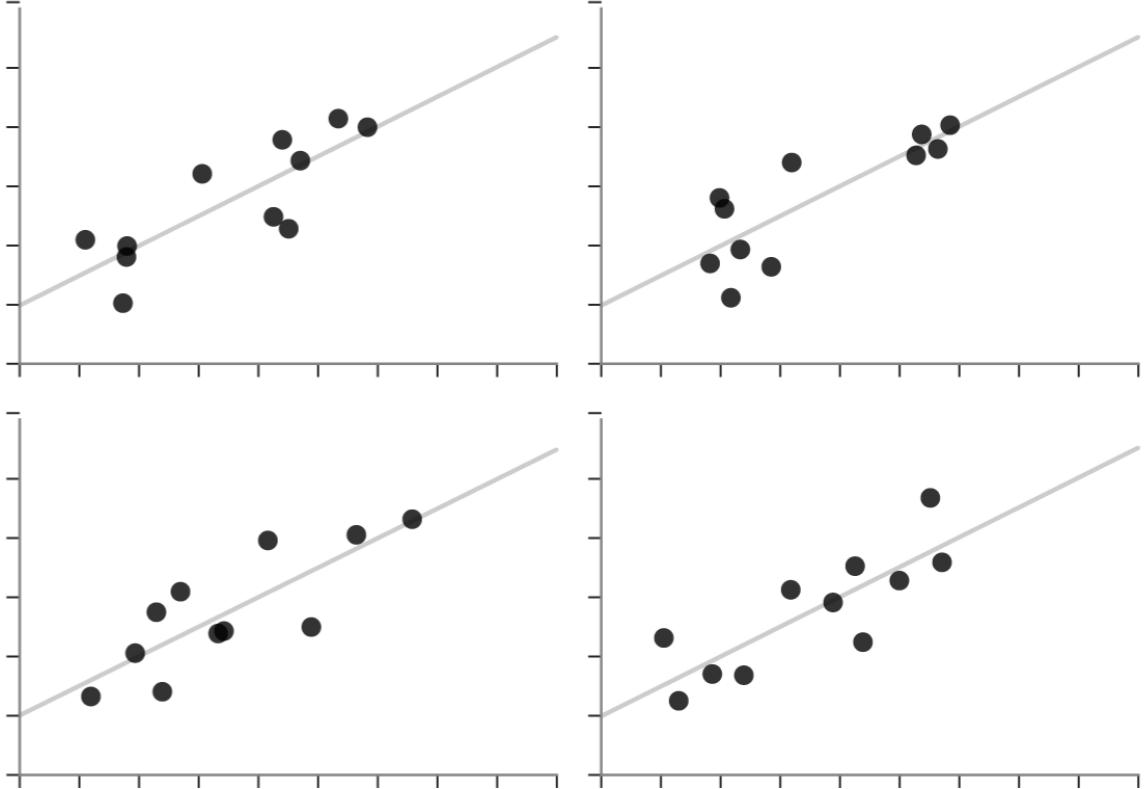
Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.



Unstructured Quartet

Each dataset here also has the same summary statistics. However, they are not *clearly different* or *visually distinct*.



X Mean: 54.2659224
 Y Mean: 47.8313999
 X SD : 16.7649829
 Y SD : 26.9342120
 Corr. : -0.0642526

Example: Antibiotics
Will Burtin, 1951

Data

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 fecalis</i>	1	1	0.1	positive
<i>Streptococcus hemolyticus</i>	0.001	14	10	positive
<i>Streptococcus viridans</i>	0.005	10	40	positive

Data

Genus, Species

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 fecalis</i>	1	1	0.1	positive
<i>Streptococcus hemolyticus</i>	0.001	14	10	positive
<i>Streptococcus viridans</i>	0.005	10	40	positive

Data

Genus, *Species*

Table 1: Burtin's data.

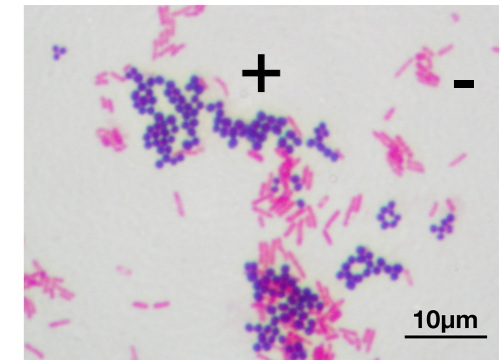
Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
<i>Aerobacter aerogenes</i>	870	1	1.6	negative
<i>Brucella abortus</i>	1	2	0.02	negative
<i>Brucella anthracis</i>	0.001	0.01	0.007	positive
<i>Diplococcus pneumoniae</i>	0.005	11	10	positive
<i>Escherichia coli</i>	100	0.4	0.1	negative
<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 fecalis</i>	1	1	0.1	positive
<i>Streptococcus hemolyticus</i>	0.001	14	10	positive
<i>Streptococcus viridans</i>	0.005	10	40	positive

Data

Genus, Species

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 fecalis</i>	1	1	0.1	positive
<i>Streptococcus hemolyticus</i>	0.001	14	10	positive
<i>Streptococcus viridans</i>	0.005	10	40	positive

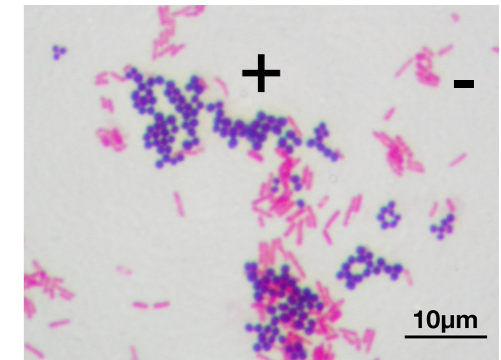


Data

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 fecalis</i>		1	1	0.1	positive
<i>Streptococcus hemolyticus</i>		0.001	14	10	positive
<i>Streptococcus viridans</i>		0.005	10	40	positive



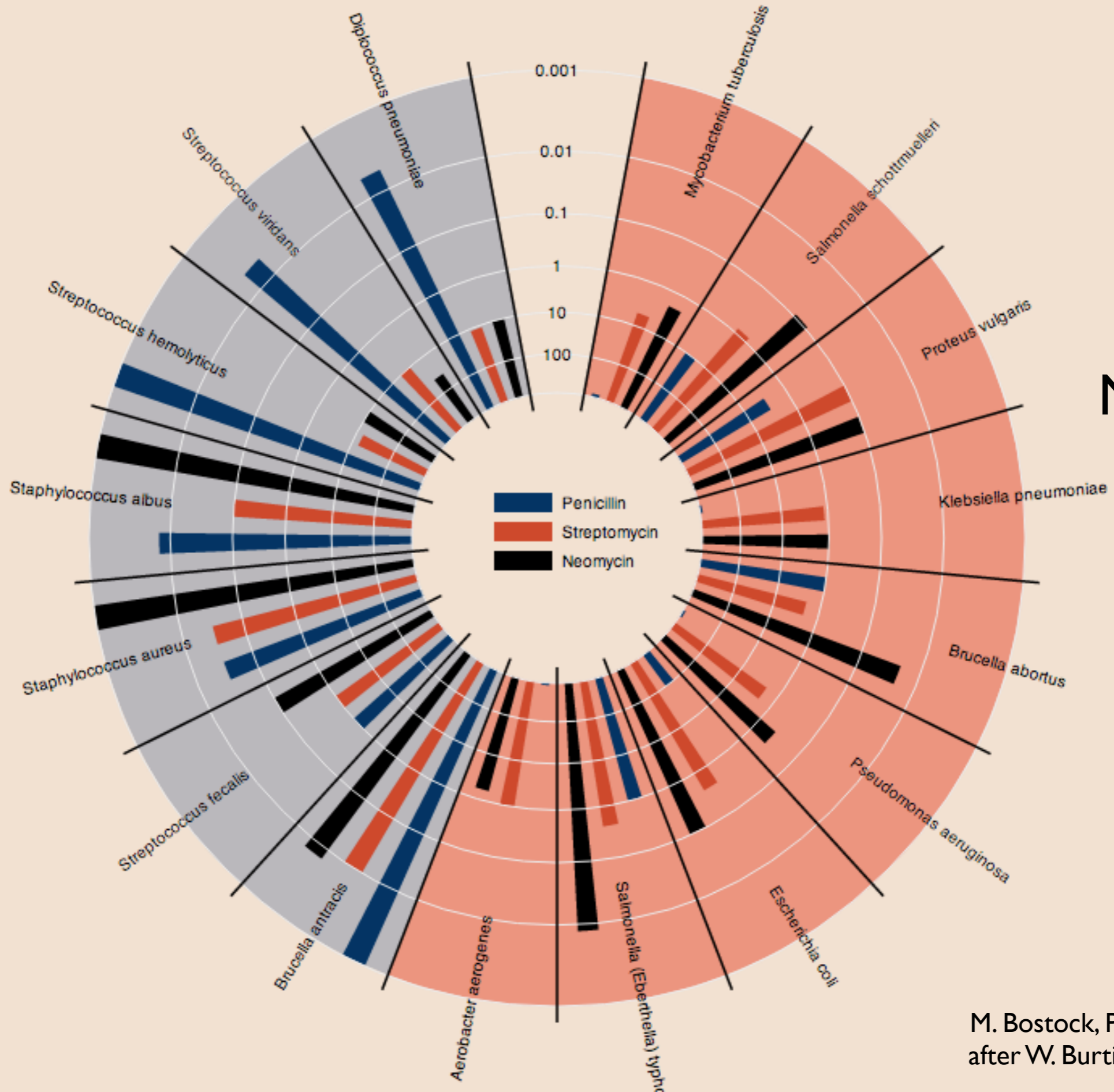
What Questions?

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 fecalis</i>	1	1	0.1	positive
<i>Streptococcus hemolyticus</i>	0.001	14	10	positive
<i>Streptococcus viridans</i>	0.005	10	40	positive

Gram
Positive

Gram
Negative

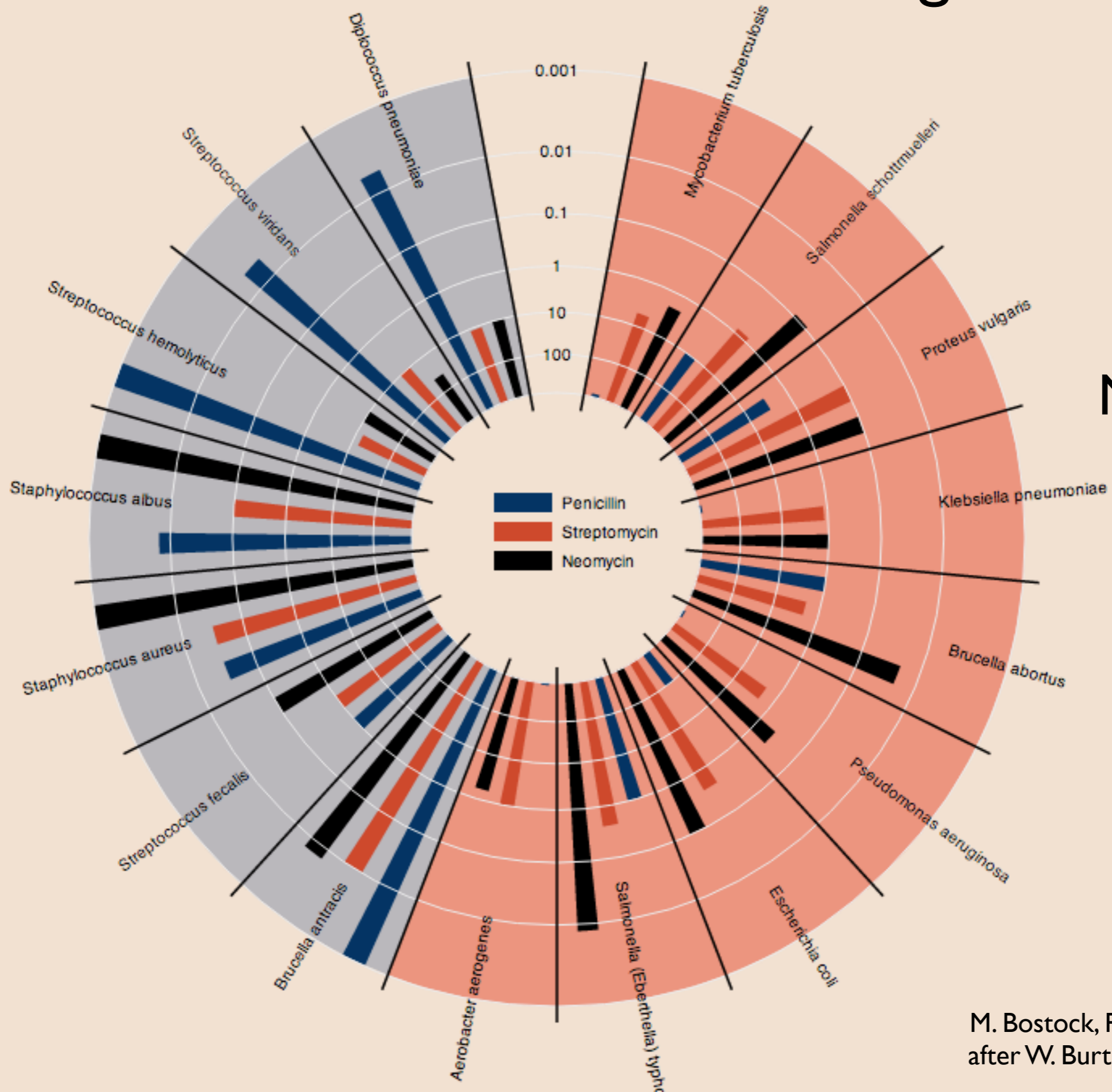


M. Bostock, Protovis
after W. Burtin, 1951

How effective are the drugs?

Gram
Positive

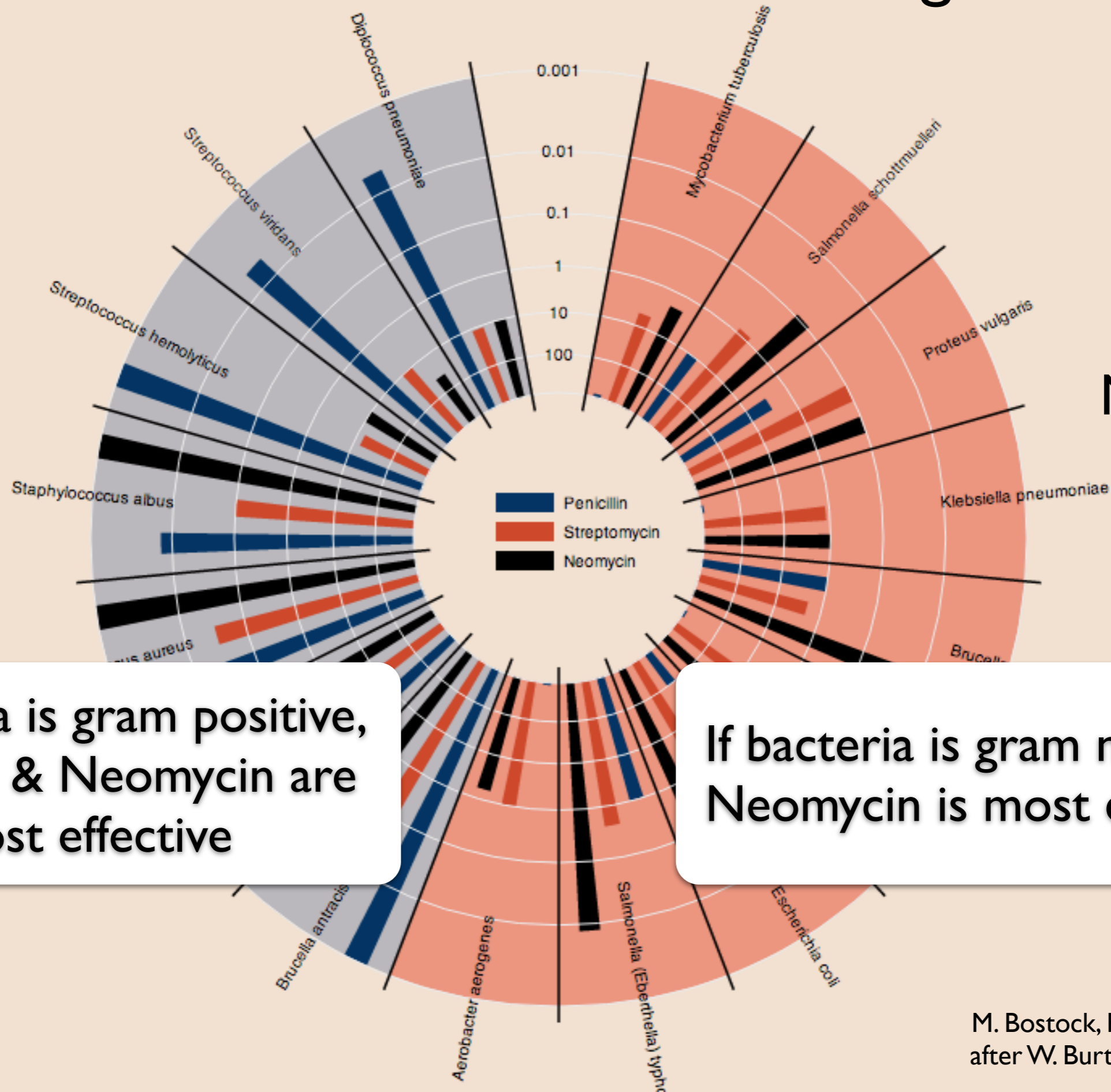
Gram
Negative



How effective are the drugs?

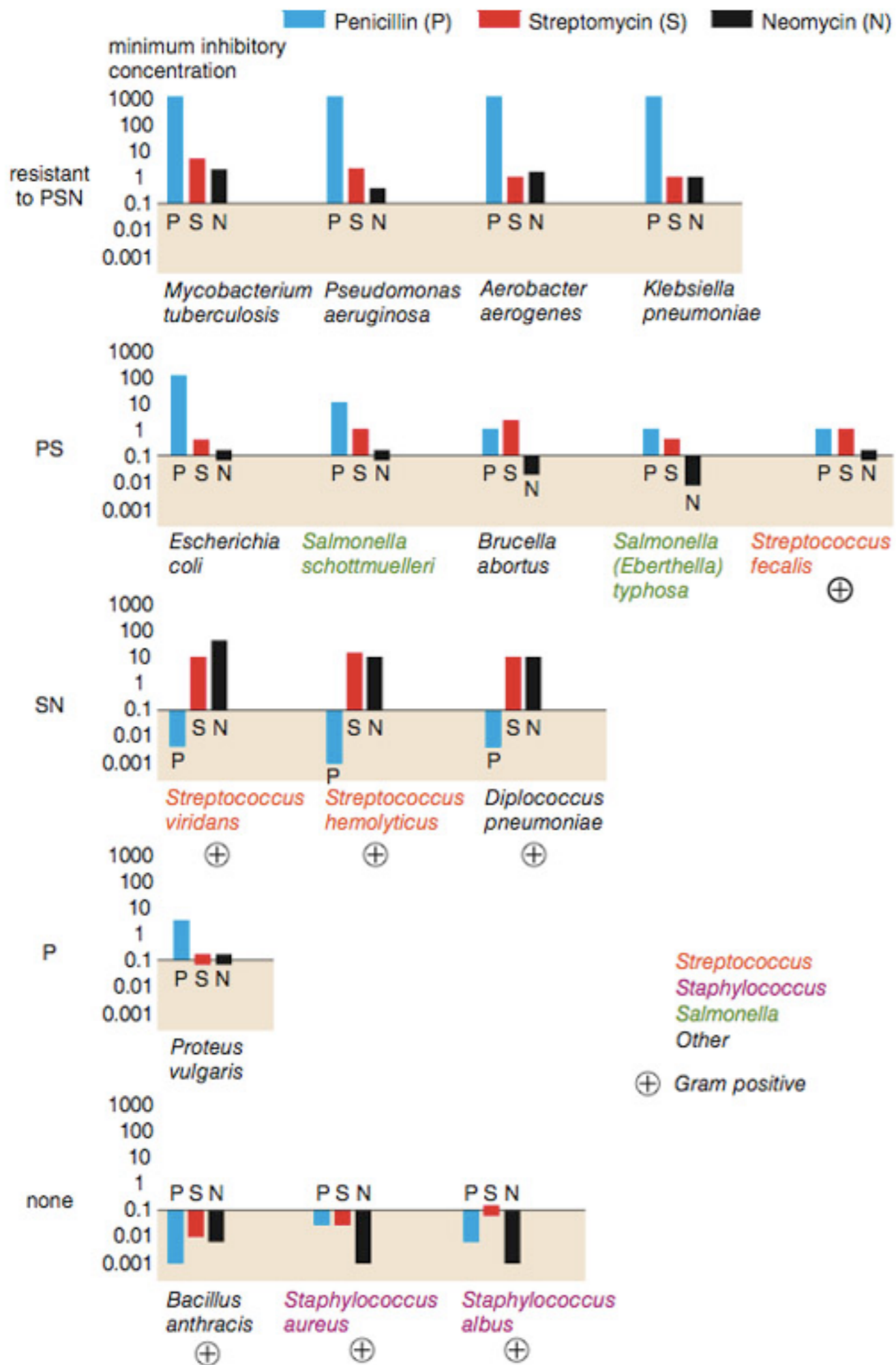
Gram Positive

Gram Negative



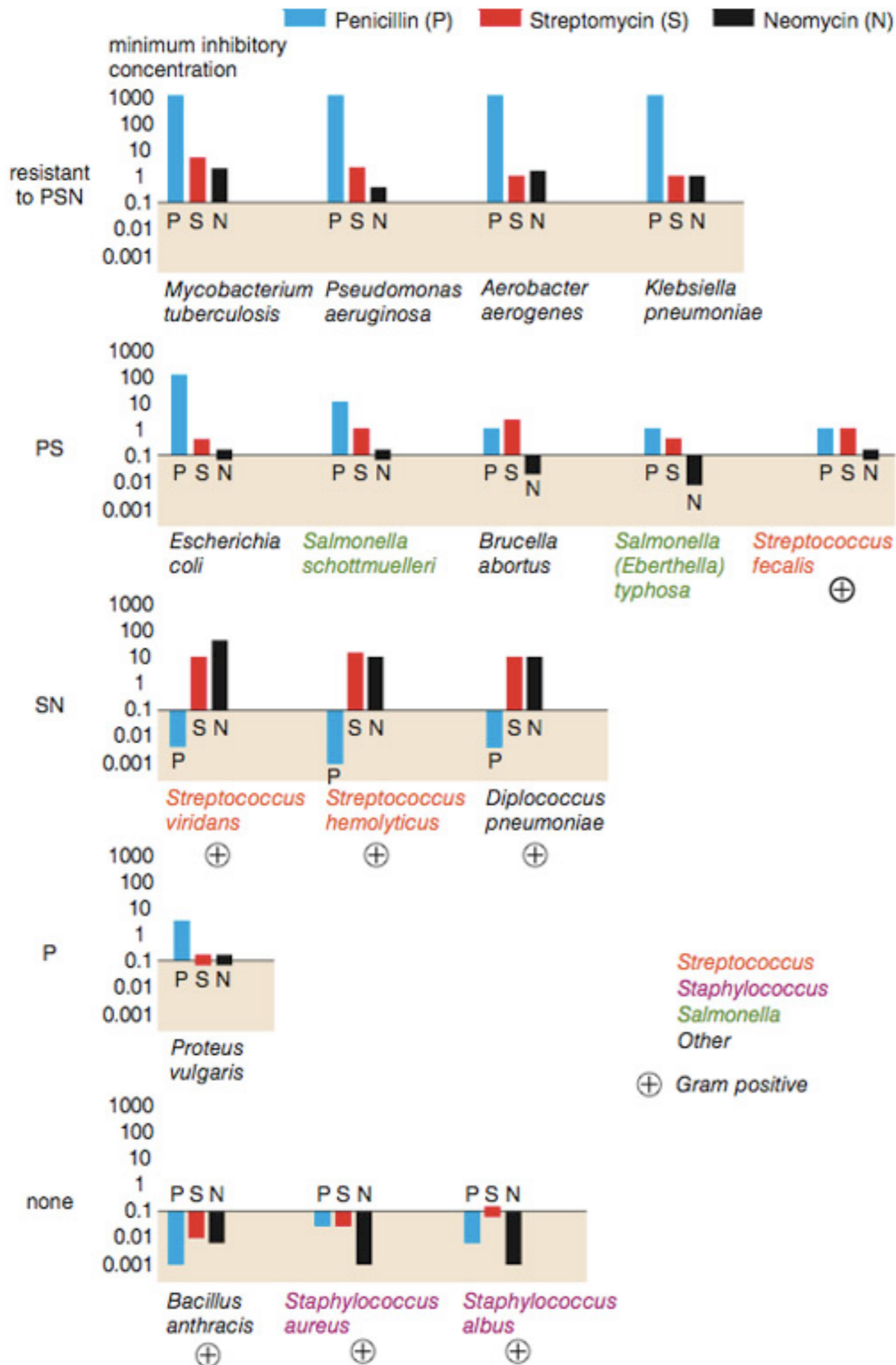
If bacteria is gram positive, Penicillin & Neomycin are most effective

If bacteria is gram negative, Neomycin is most effective



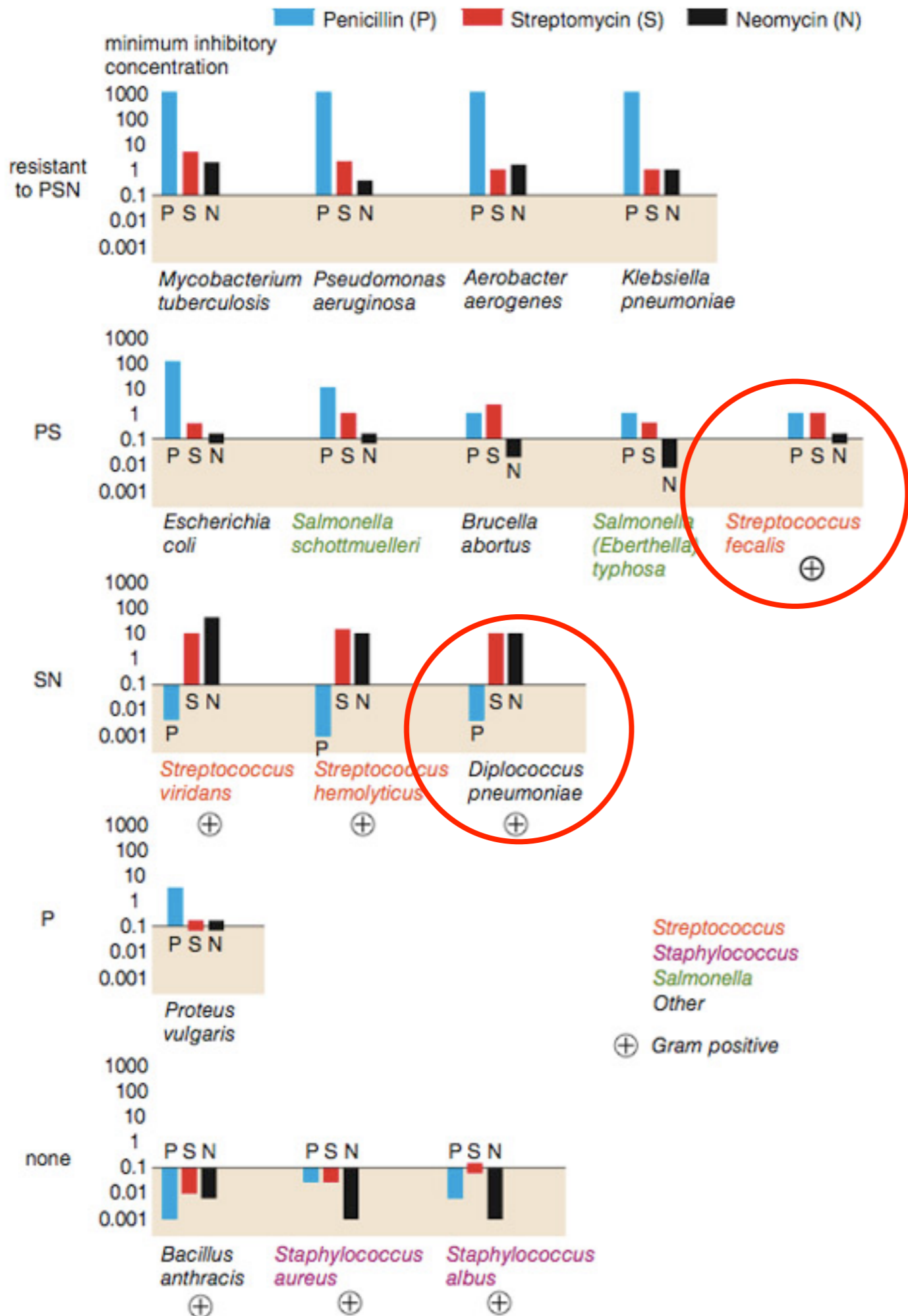
Wainer & Lysen, "That's funny..."
 American Scientist, 2009
 Adapted from Brian Schmotzer

How do the bacteria compare?

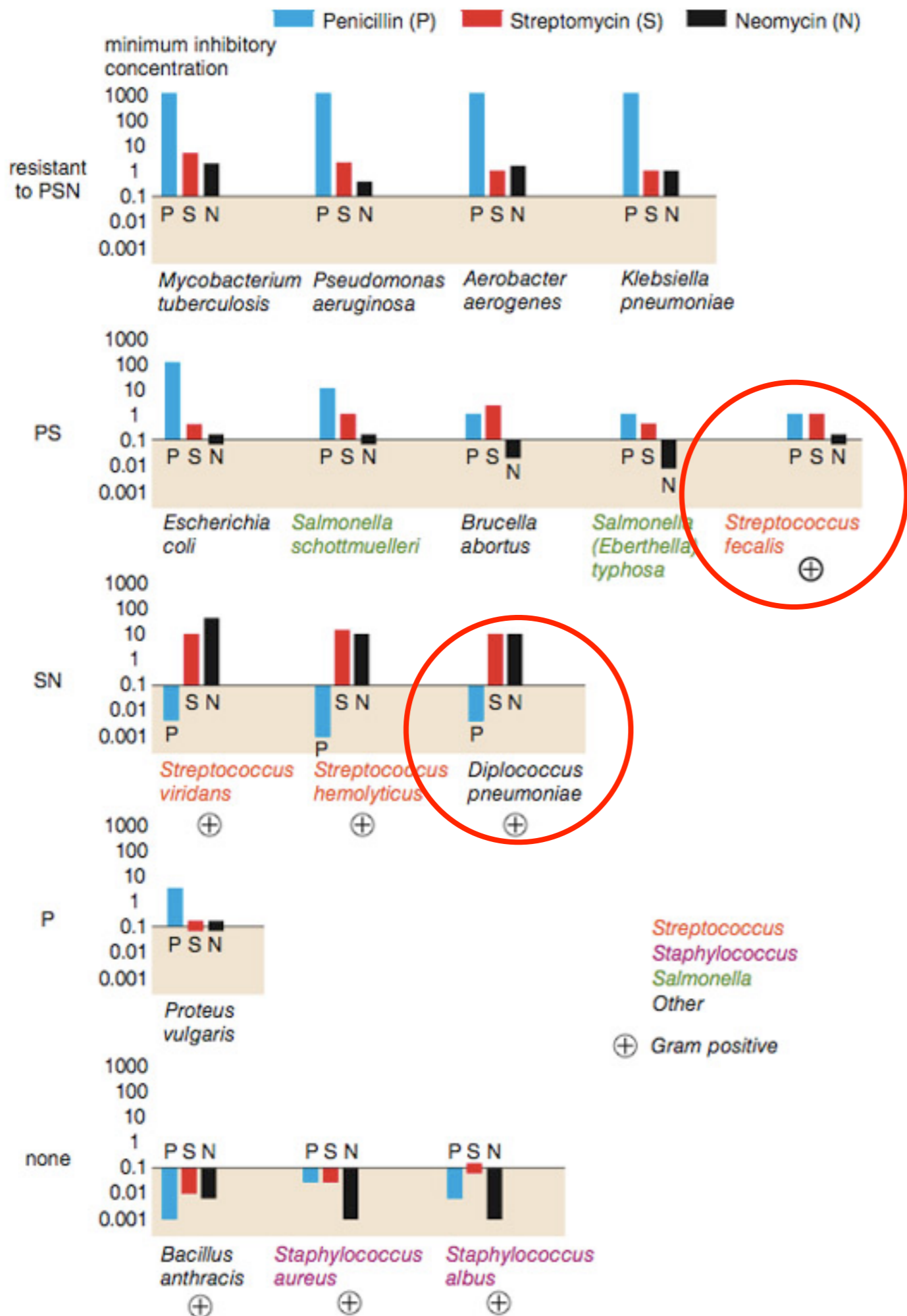


Wainer & Lysen, "That's funny..."
 American Scientist, 2009
 Adapted from Brian Schmotzer

How do the bacteria compare?



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

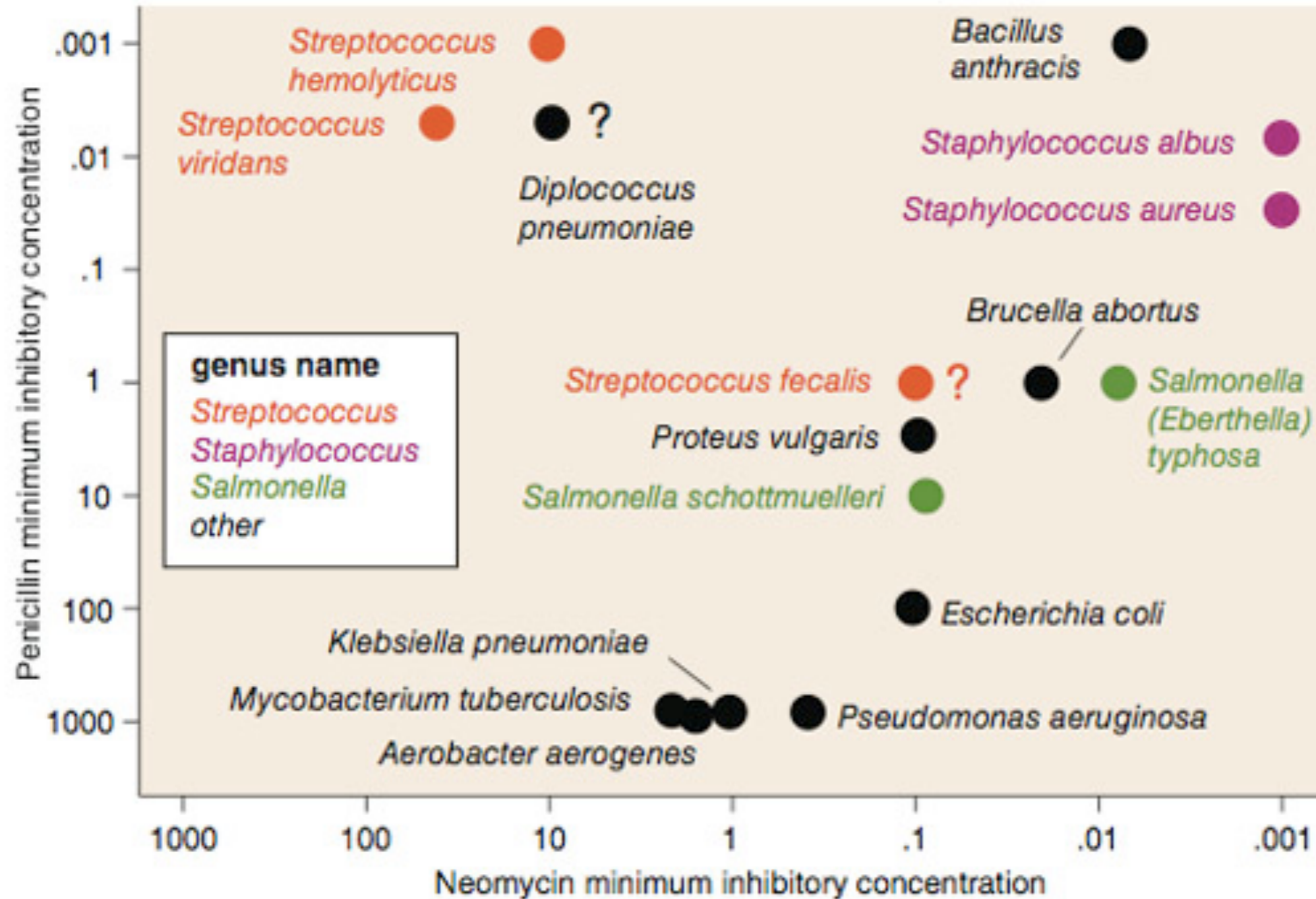


How do the bacteria compare?

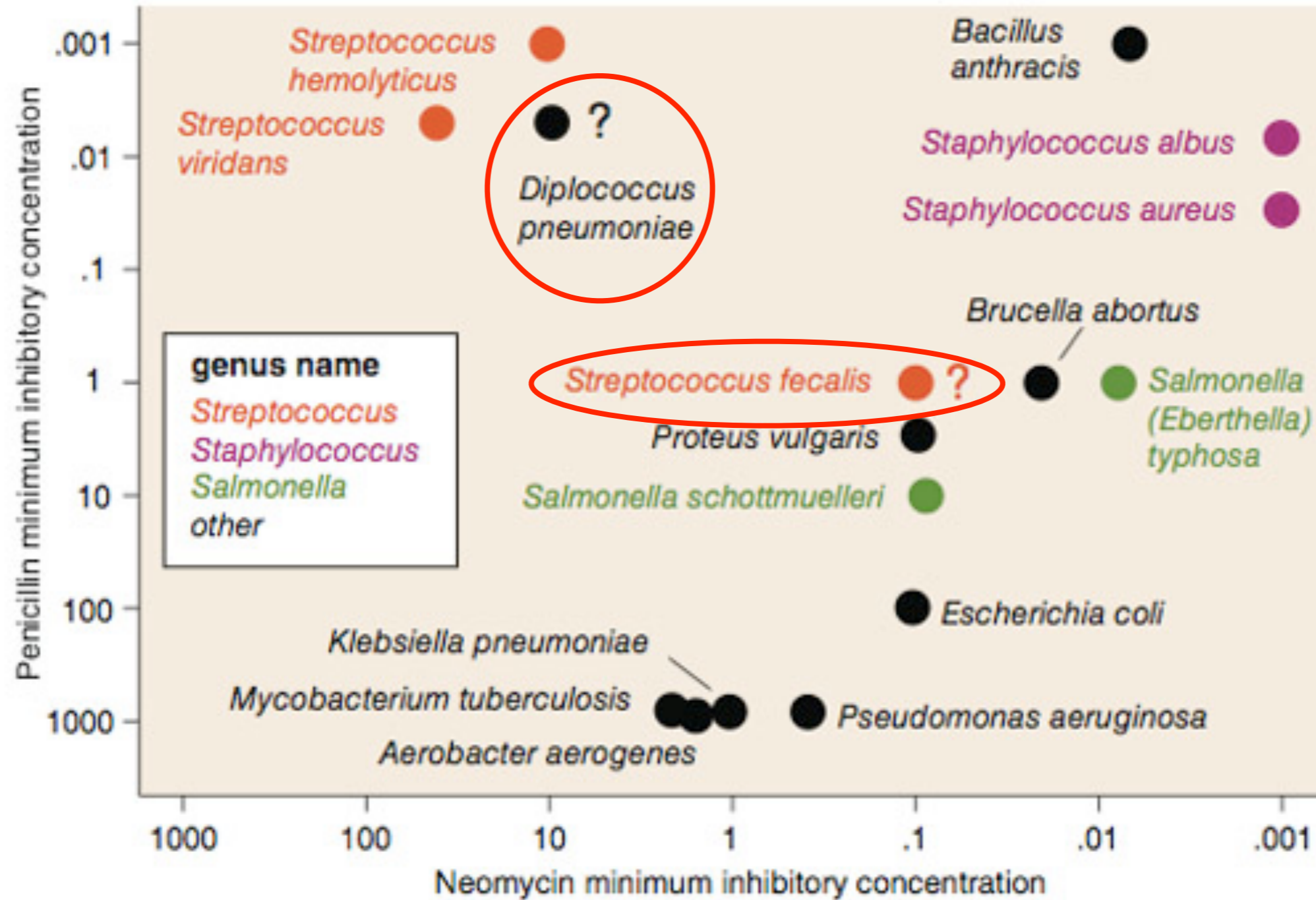
Not a streptococcus!
(realized ~30 years later)

Really a streptococcus!
(realized ~20 years later)

How do the bacteria compare?



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 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

Barry Bonds 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

755
23 seasons

714

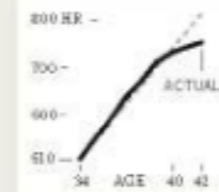
20 seasons
Bonds was injured last season. He played 14 games and hit 5 homers

Homer Pace After Age 34

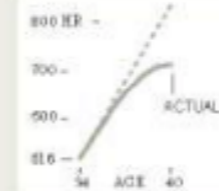
If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.

----- PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS

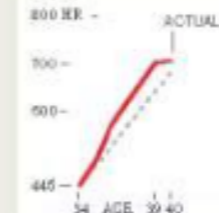
Aaron
Actual homers slightly outpace projected homers for five seasons.



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



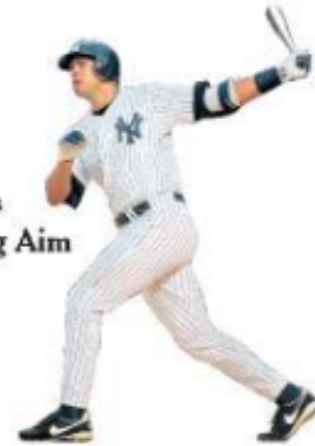
Bonds
From age 35 to 39, he averaged 14 more homers a season than projected.



Note: Ages as of July 1 of each season.

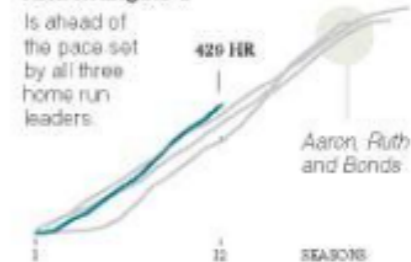
According to allegations in a book about Bonds, he began taking steroids before the 1999 season, his 14th in the league. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.

Others Taking Aim



Alex Rodriguez

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



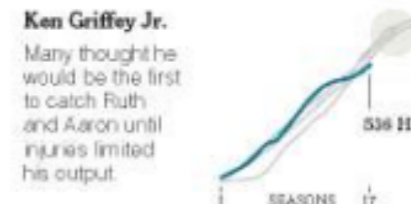
Albert Pujols

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.

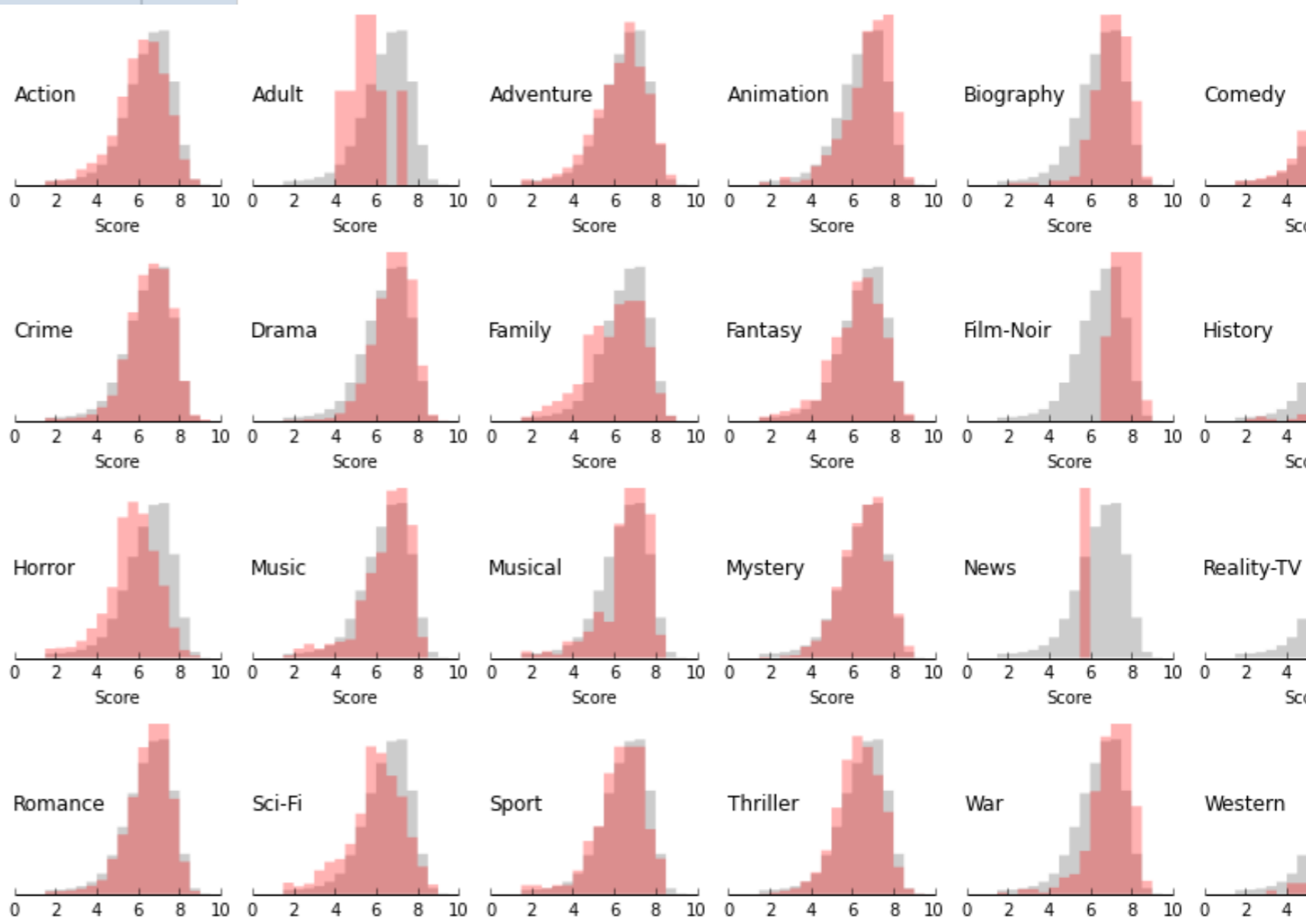
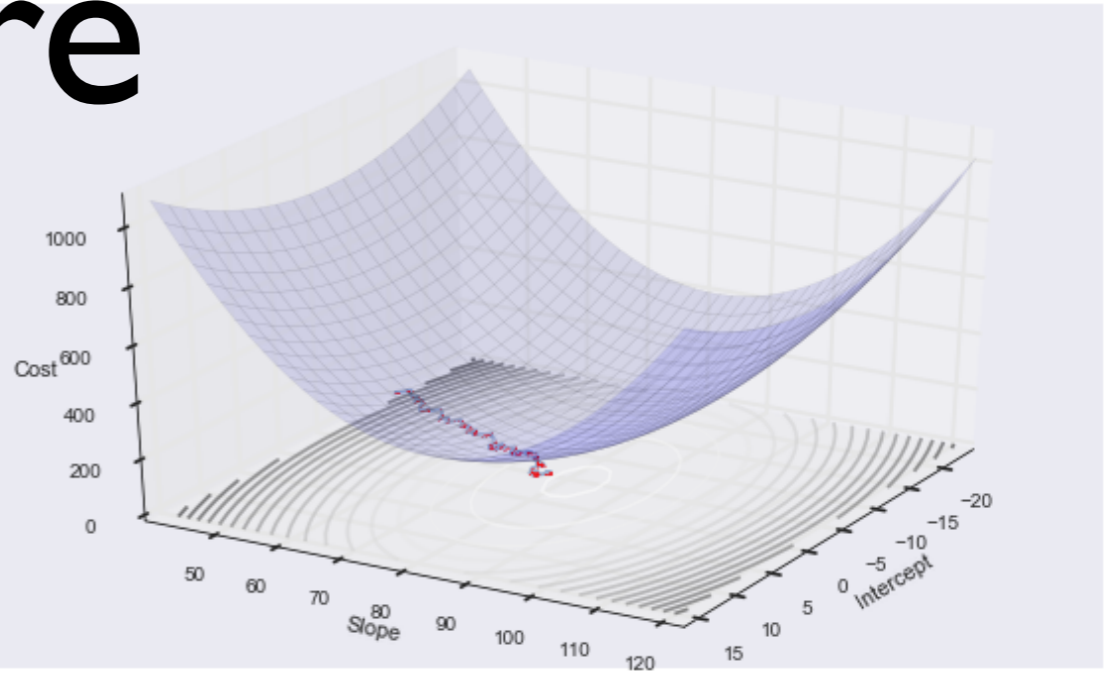
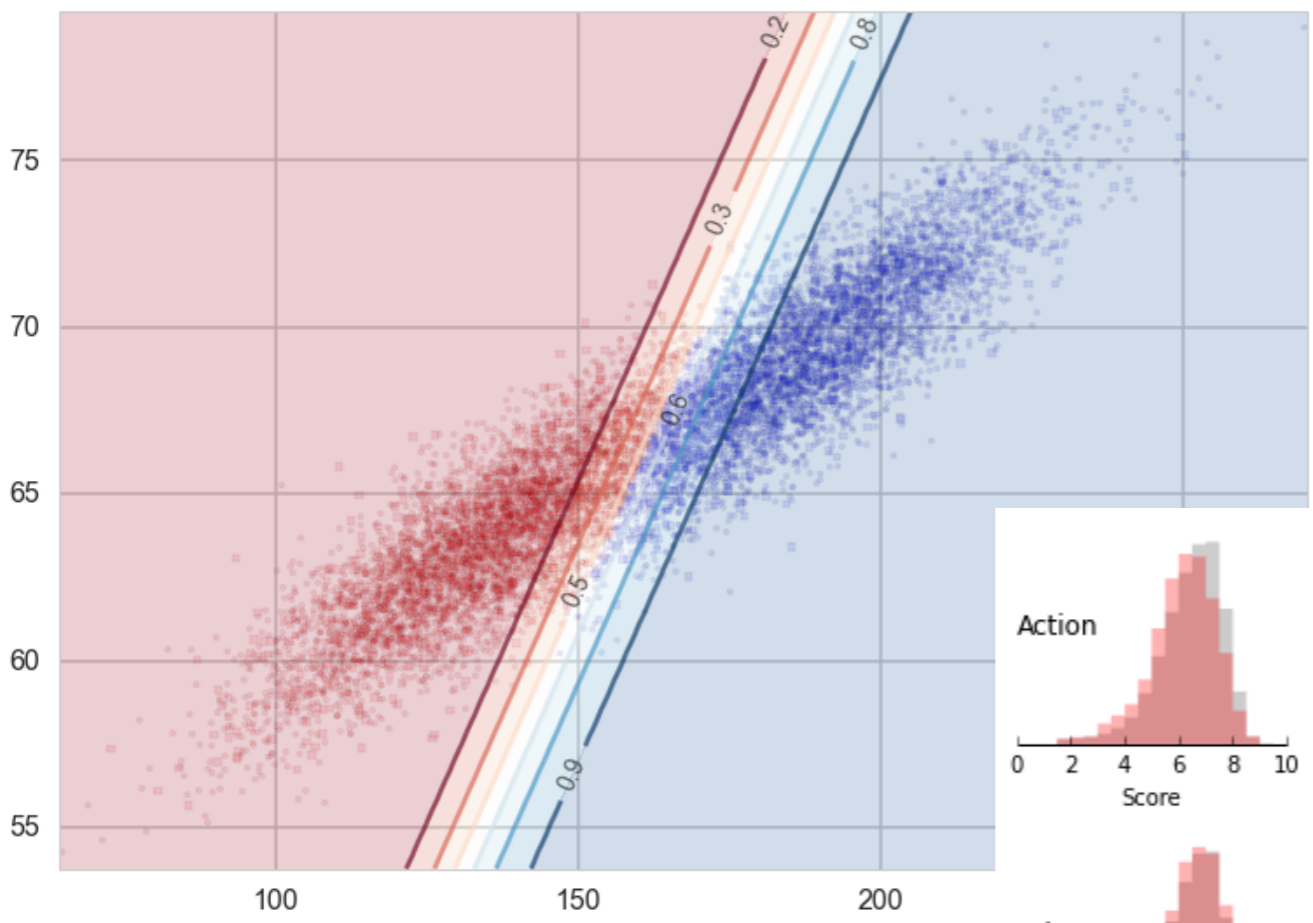


Differing Paths to the Top of the Charts

The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th)



Explore



EDA Workflow

1. **Build** a DataFrame from the data (ideally, put all data in this object)
2. **Clean** the DataFrame. It should have the following properties
 - Each row describes a single object
 - Each column describes a property of that object
 - Columns are numeric whenever appropriate
 - Columns contain atomic properties that cannot be further decomposed
3. Explore **global properties**. Use histograms, scatter plots, and aggregation functions to summarize the data.
4. Explore **group properties**. Use groupby and small multiples to compare subsets of the data.

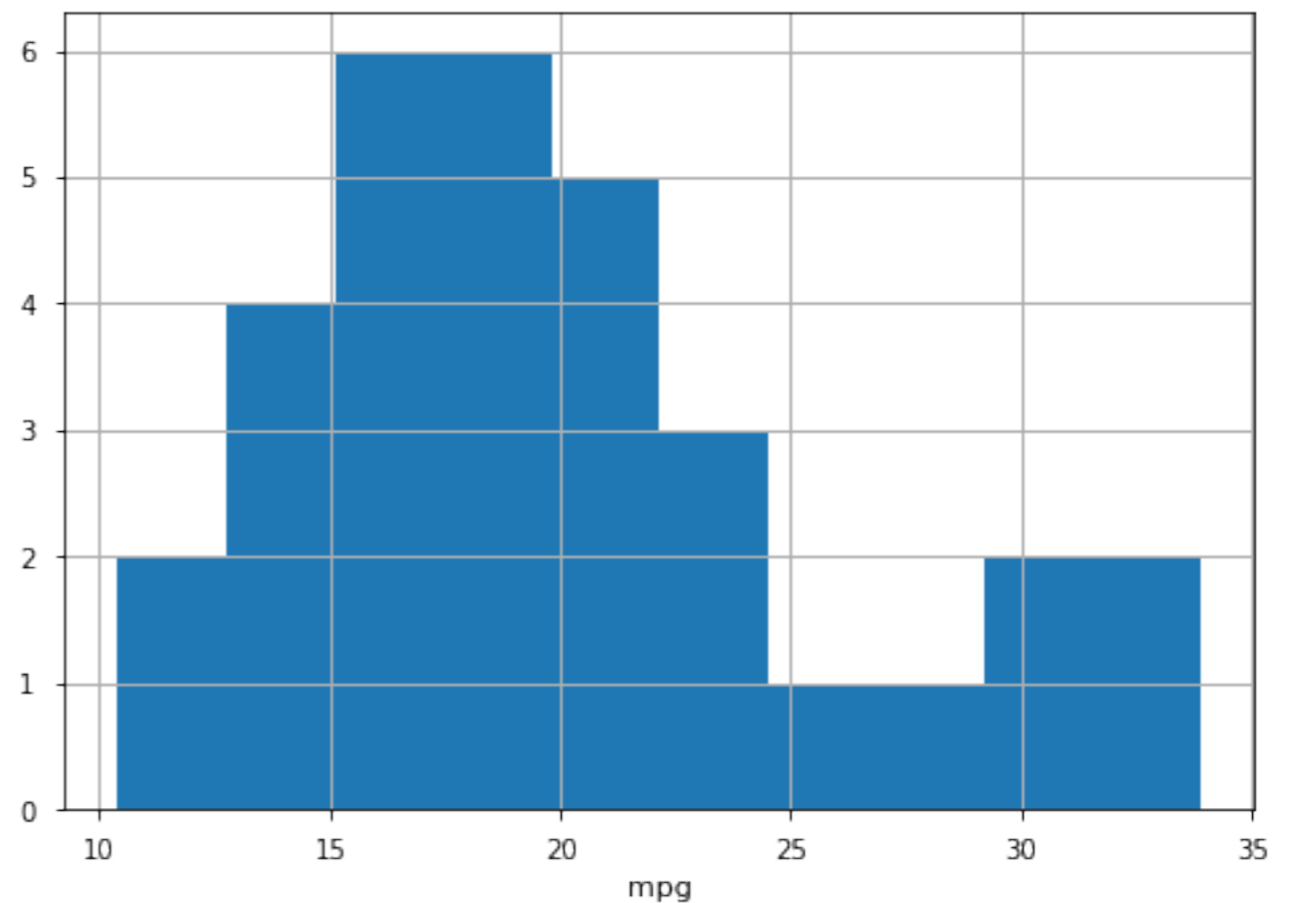
Viz options

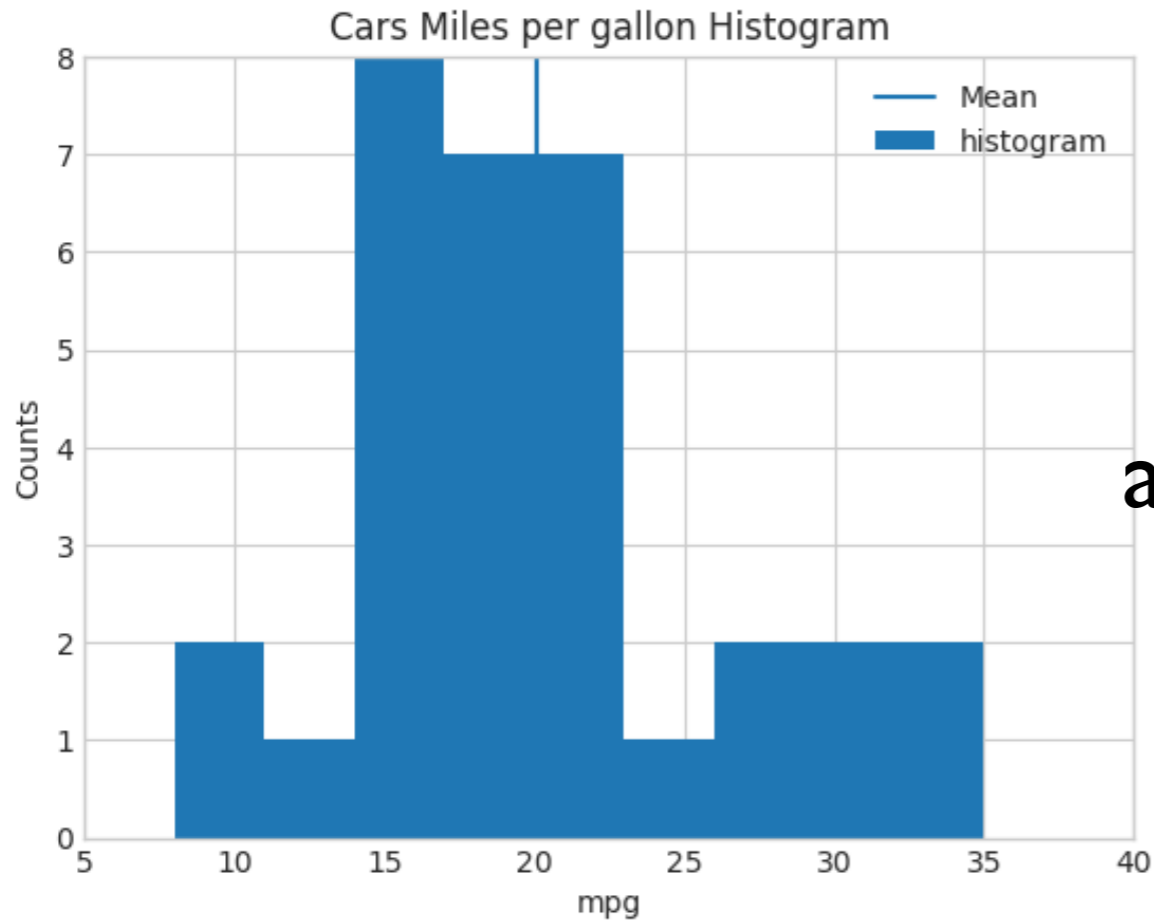
- Pandas Visualization module
- Matplotlib
- Seaborn
- Above 3 are inter-mixable
- Be lazy (to an extent...)
- Other options: Bokeh, Vega, Vincent, Altair

Cars Dataset

	name	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	maker
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4	Mazda
1	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4	Mazda
2	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1	Datsun
3	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1	Hornet
4	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2	Hornet

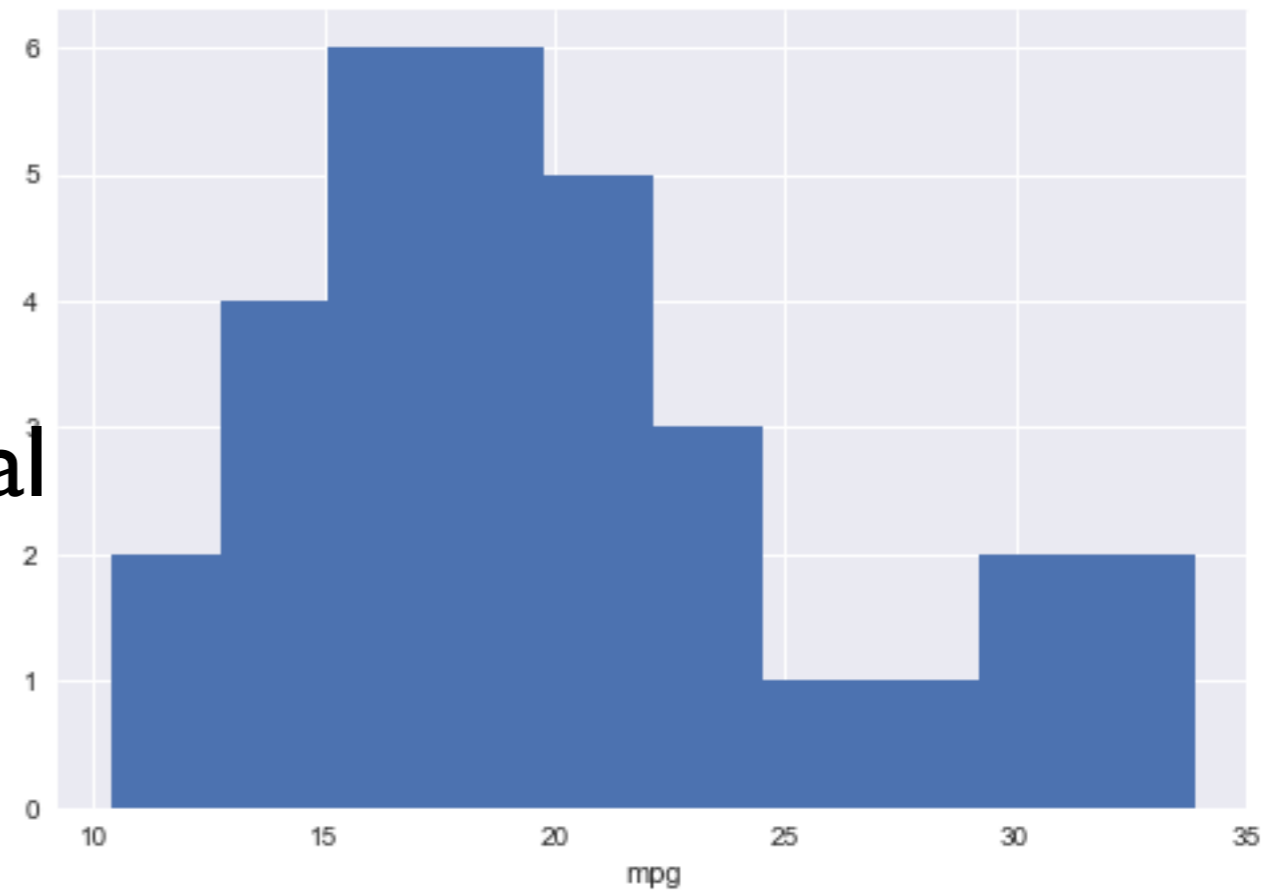
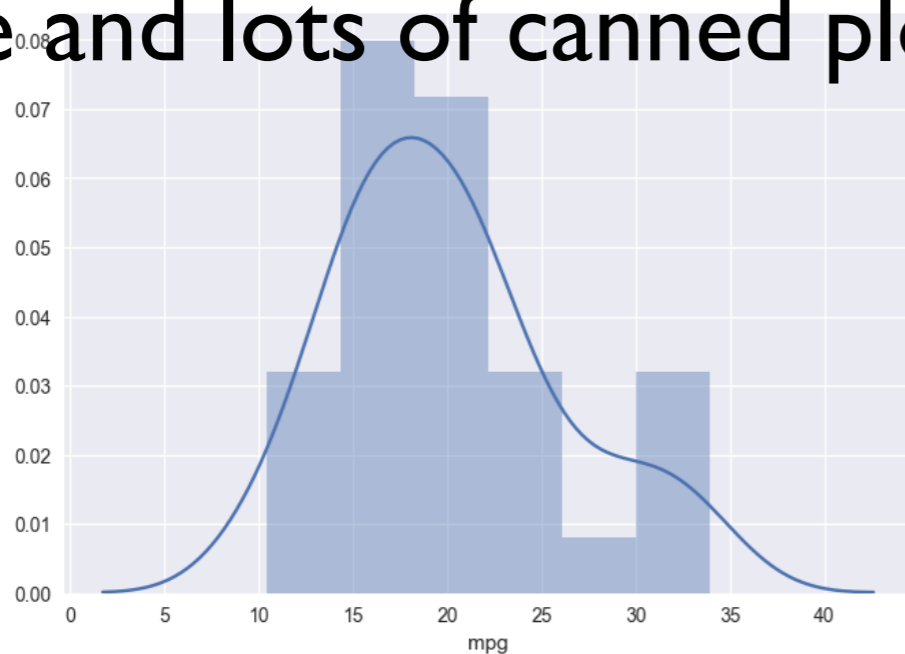
Basic Pandas/matplotlib





Can set limits, tick styles, scales, add lines, annotations, titles, legends

Seaborn provides a different visual style and lots of canned plots.



Effective Visualizations

Not Effective...

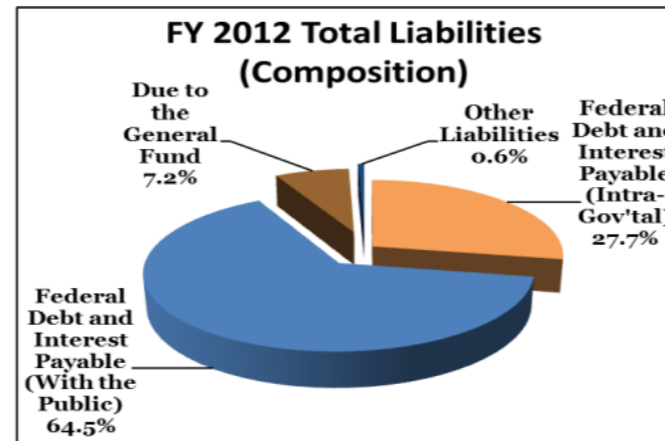
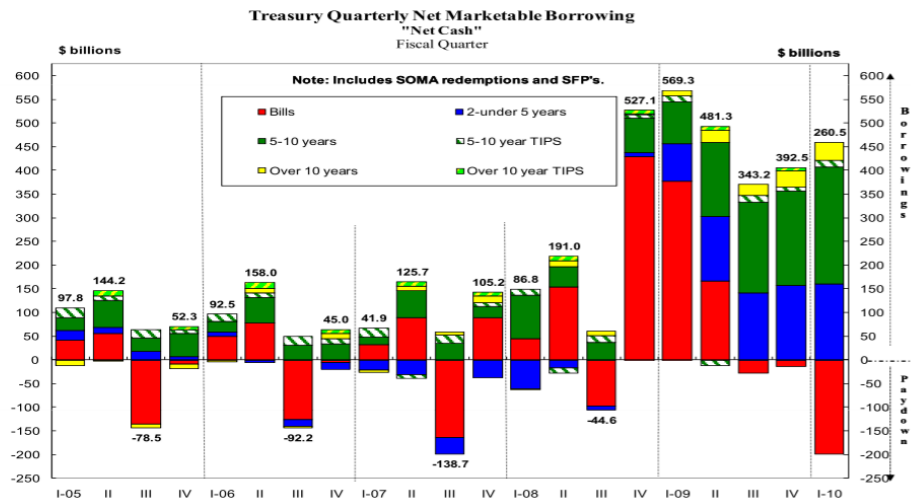


Figure 10

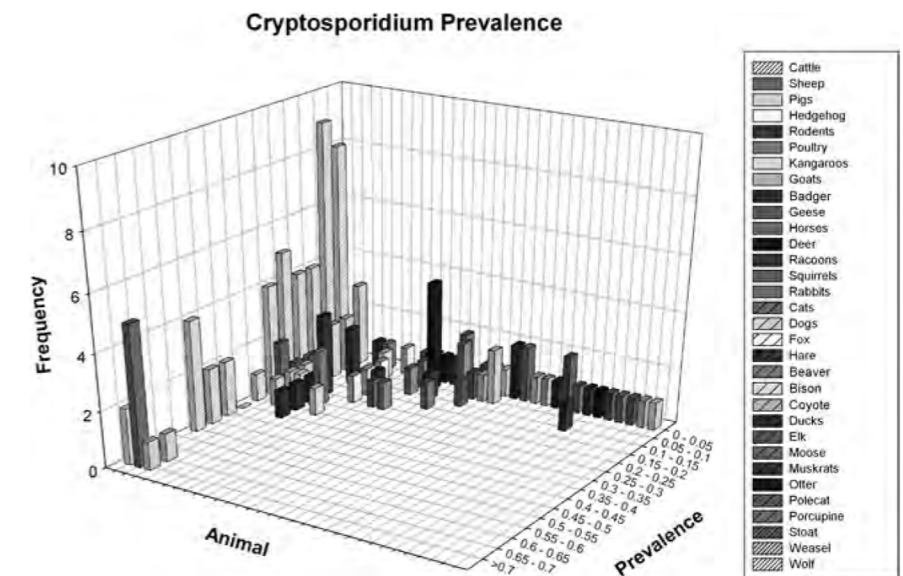
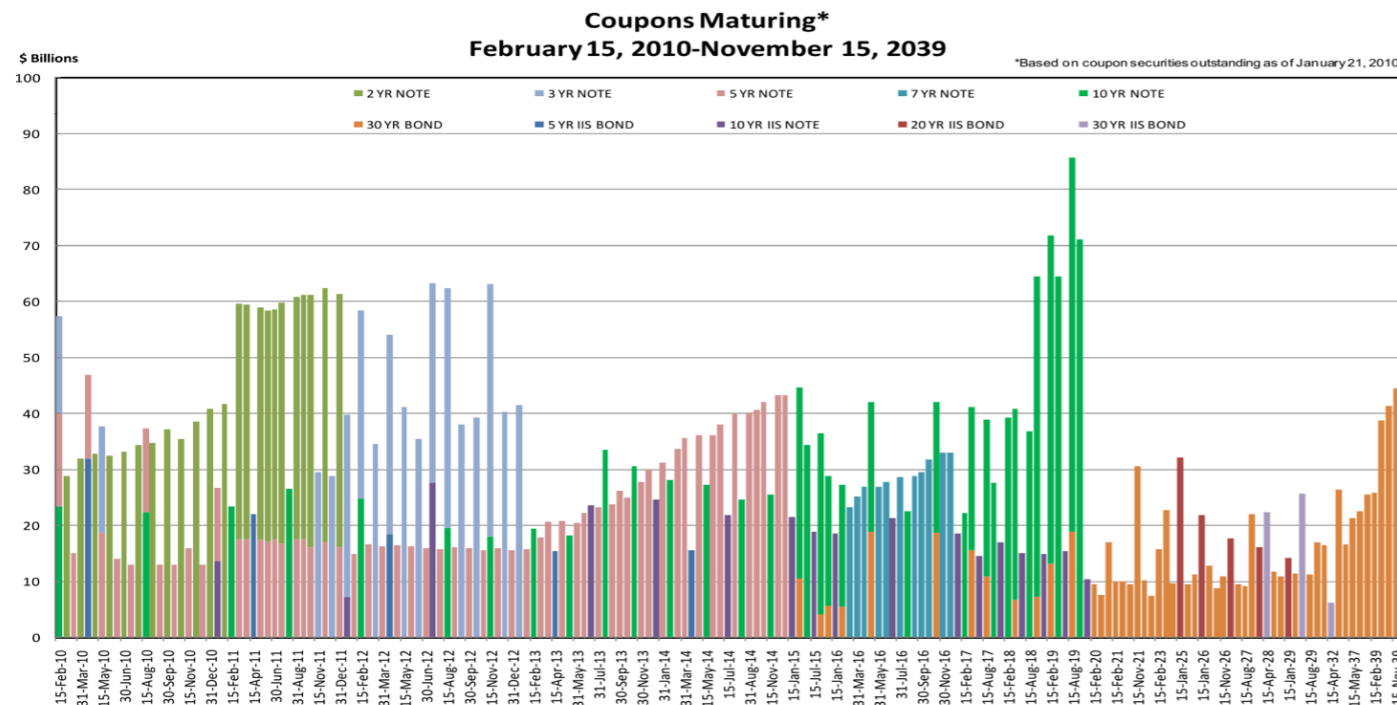
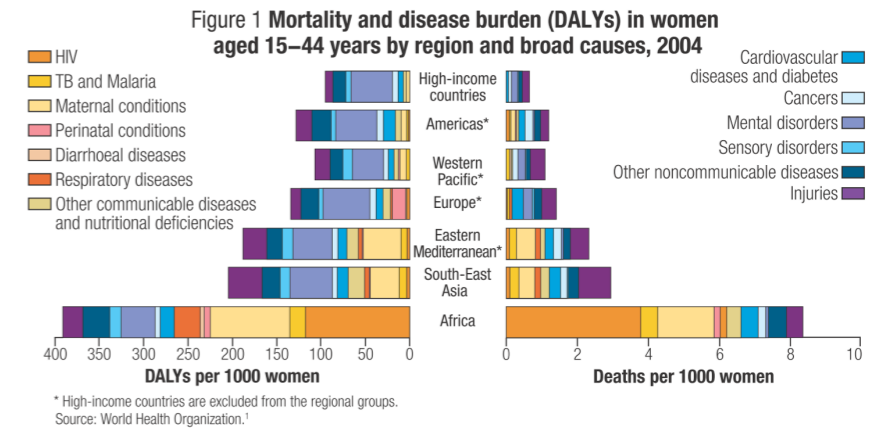


Figure 5.2 Mean prevalence rates of *Cryptosporidium* oocysts by animal species.

Effective EDA Viz

1. Have graphical integrity
2. Keep it simple
3. Use the right display
4. Use color sensibly

I. Graphical Integrity



THE OLDEST COLLEGE DAILY
ESTABLISHED 1878



Yale Summer Session

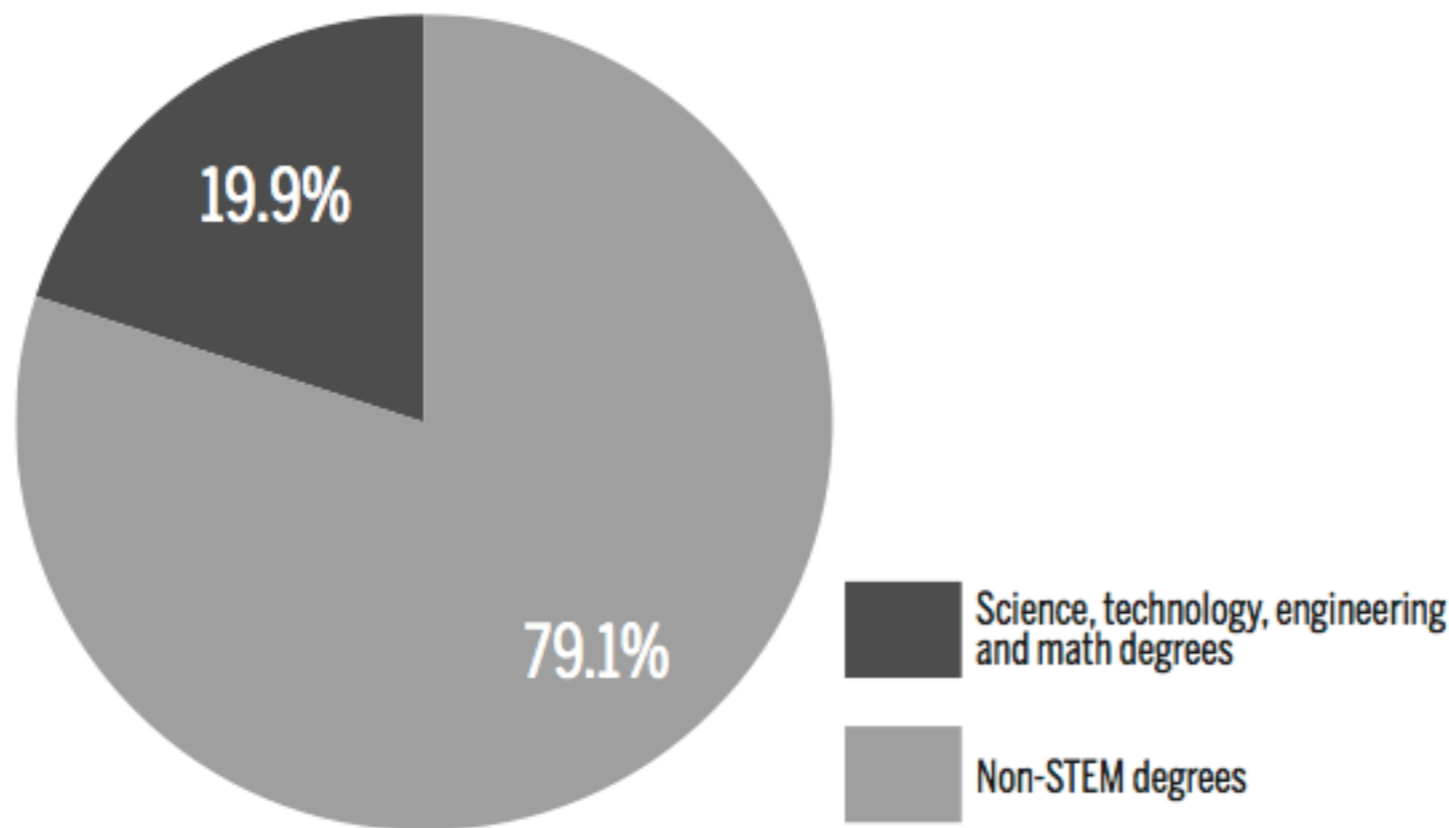
Over 200 full-credit courses.

June 4 – July 6 , July 9 – Aug 10

2012 *experience Yale*




CHART YALE GRADUATES' MAJORS, CLASS OF 2011



Facebook Recommendations

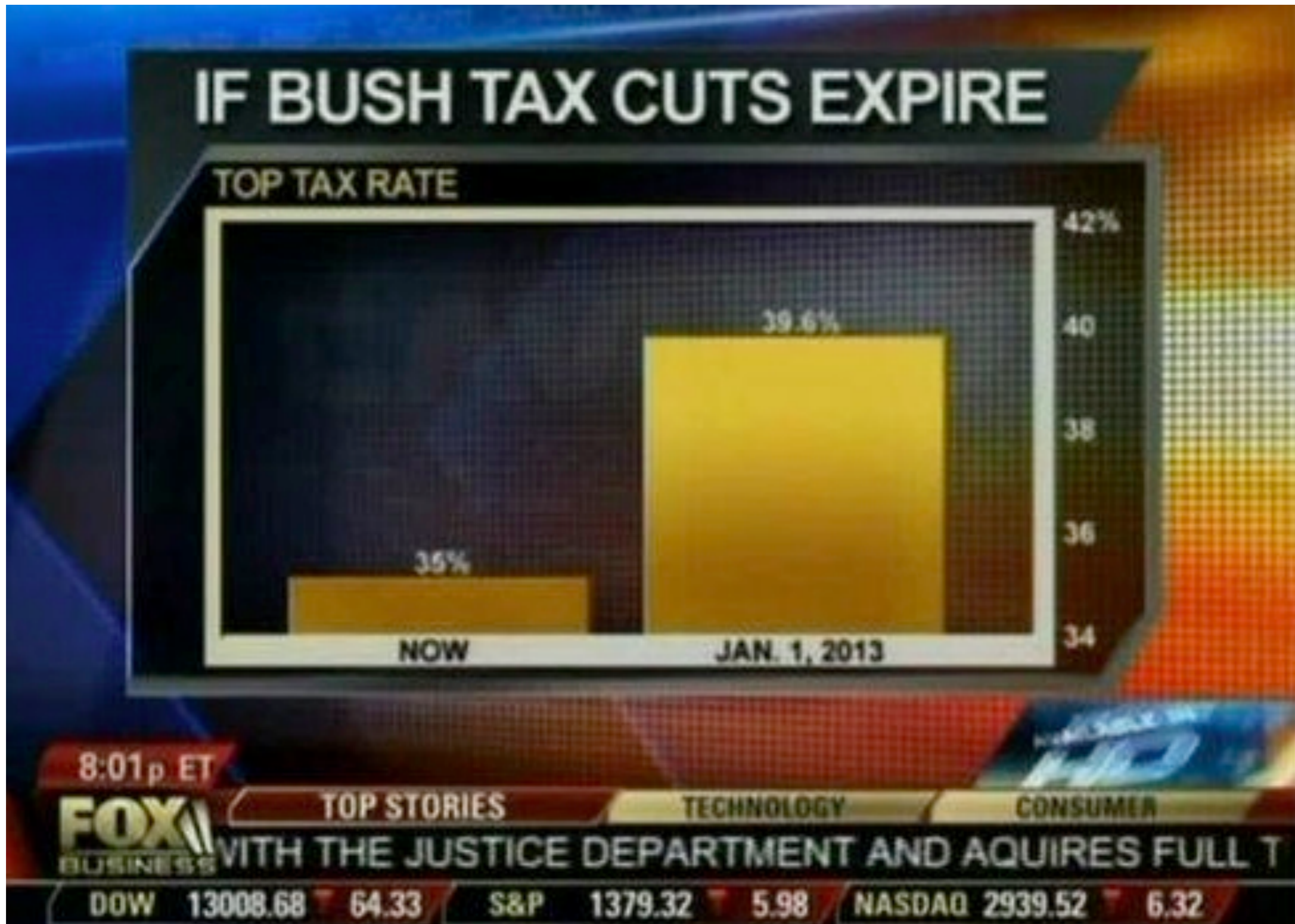
-  **Shake Shack to open in New Haven**
277 people recommend this.
-  **Popular anti-religion creates false dichotomy**
15 people recommend this.
-  **Friends remember Foucher LAW '14**
10 people recommend this.
-  **AIDS activist speaks about documentary film**
8 people recommend this.
-  **Panel outlines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

 Facebook social plugin

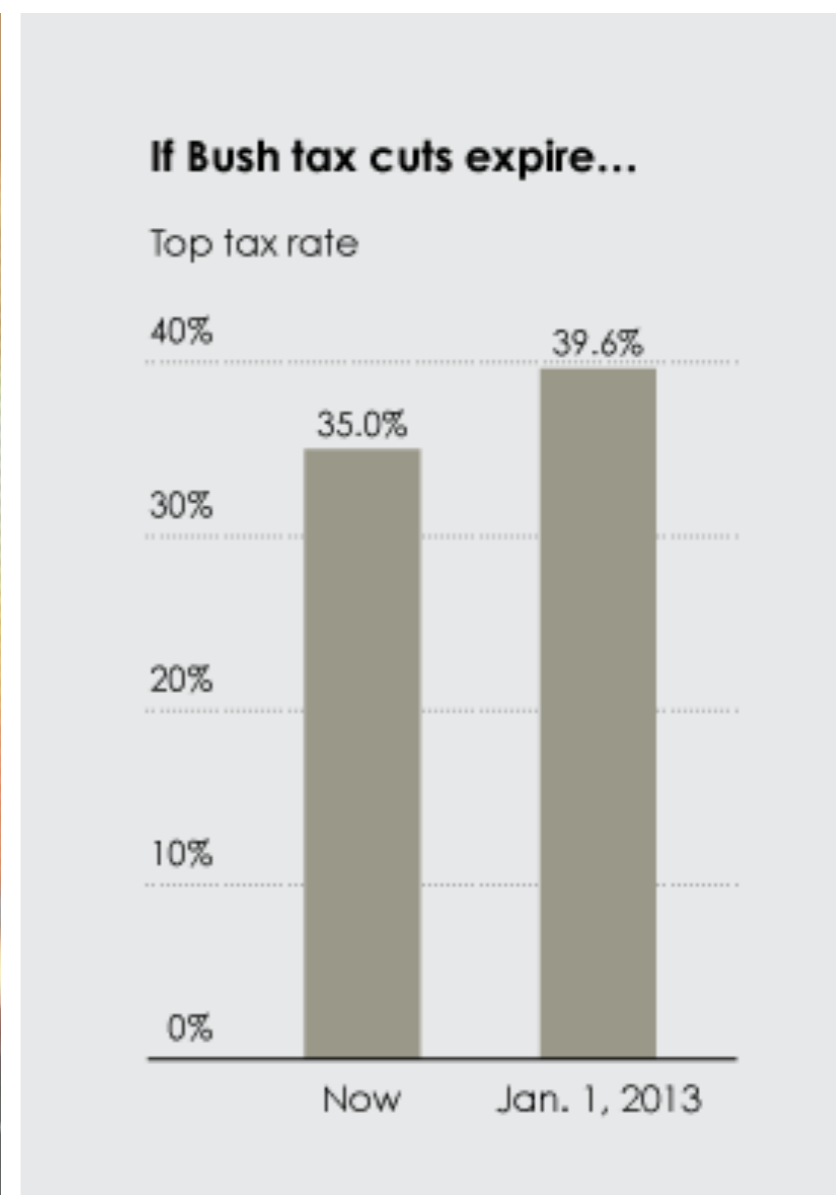
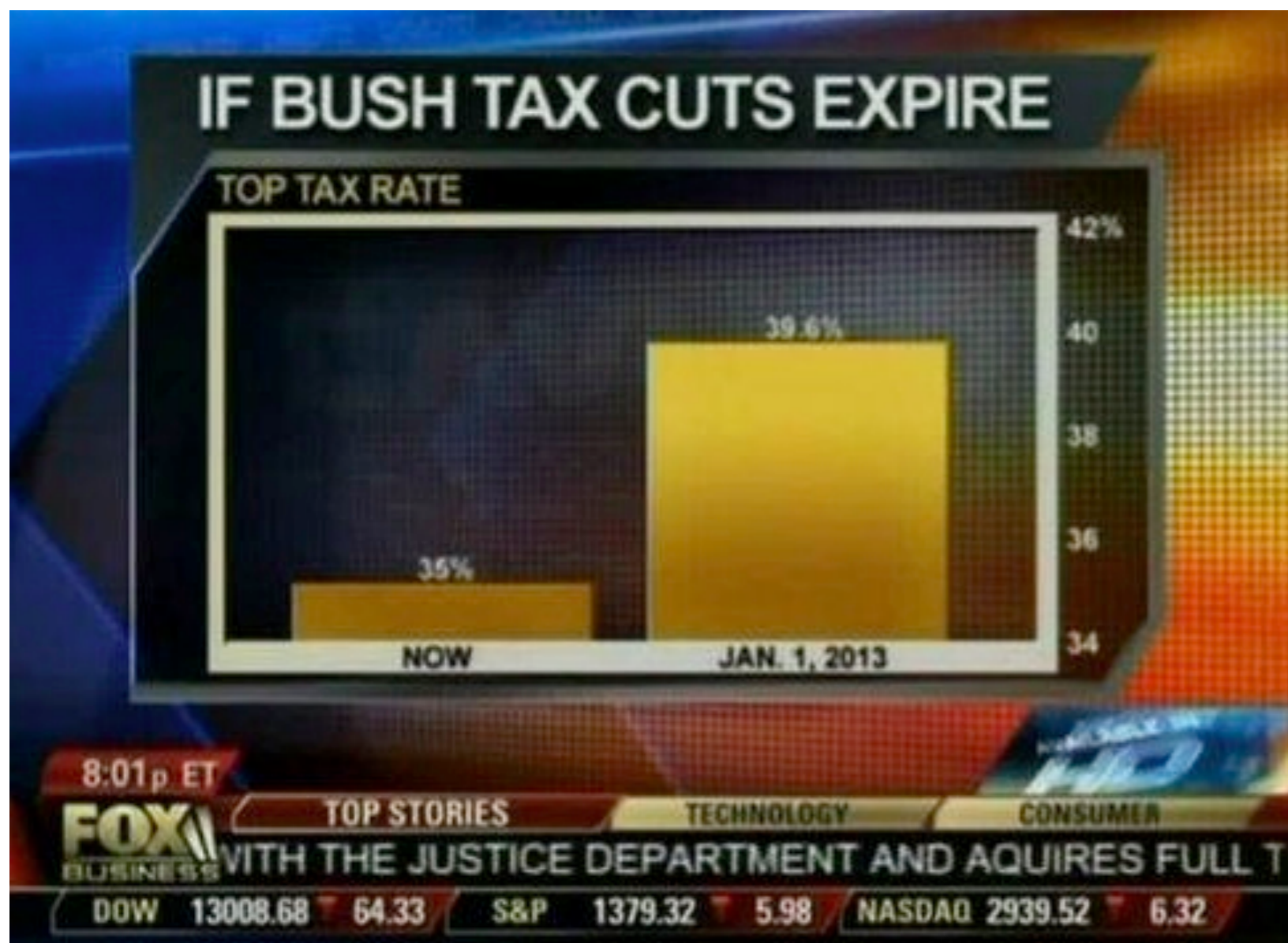
Advertisement



Graphical Integrity



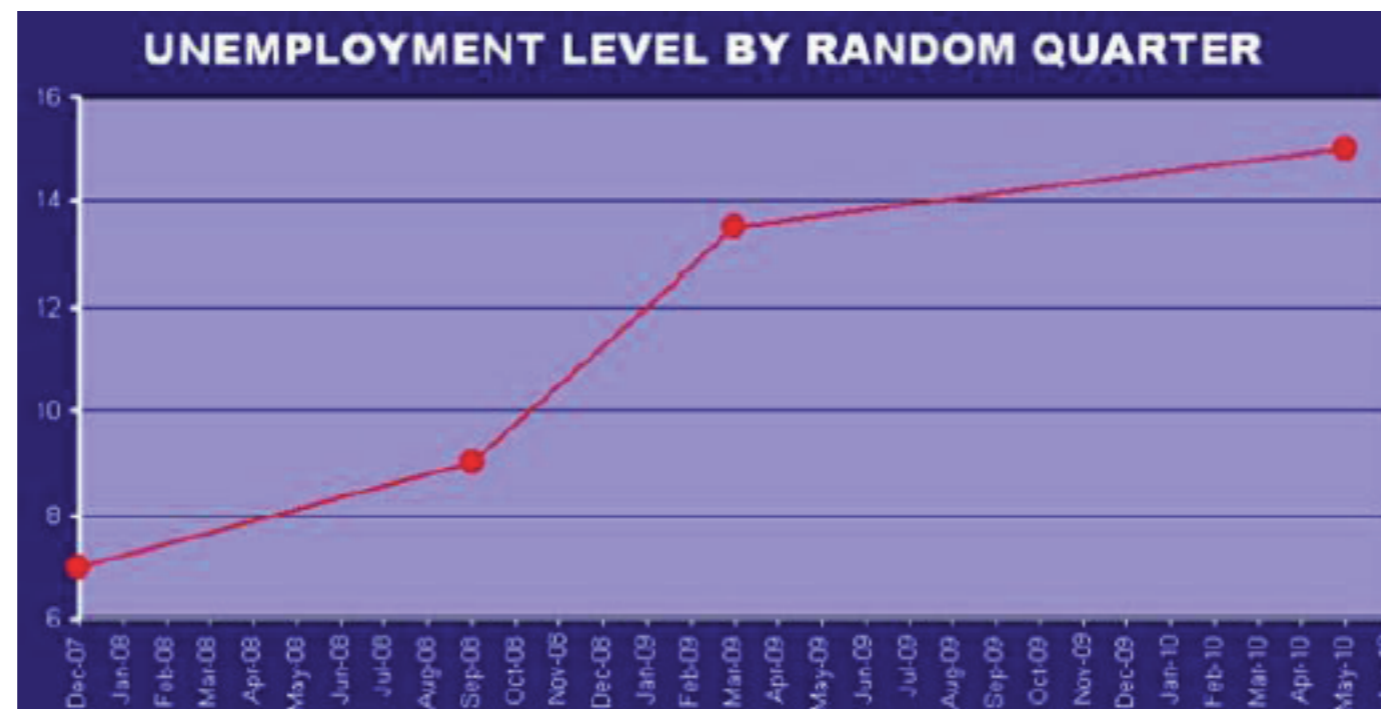
Scale Distortions



JOB LOSS BY QUARTER



Scale Distortions

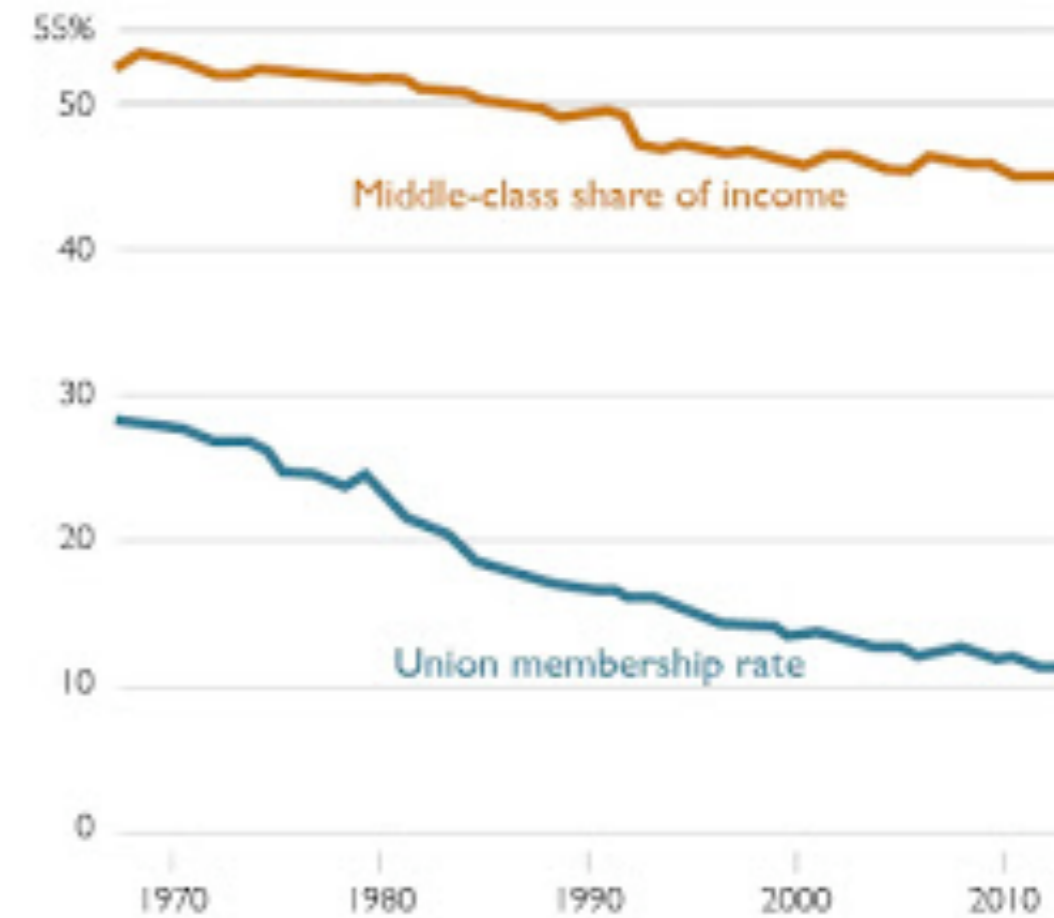


“Double the axes, double the mischief”
(Quote from Gary Smith’s *Standard Deviations*)

FIGURE 7. AS UNION MEMBERSHIP DECLINES, THE SHARE OF INCOME GOING TO THE MIDDLE CLASS SHRINKS



NEW VERSION



Graphic from Robert Reich’s *Saving Capitalism*

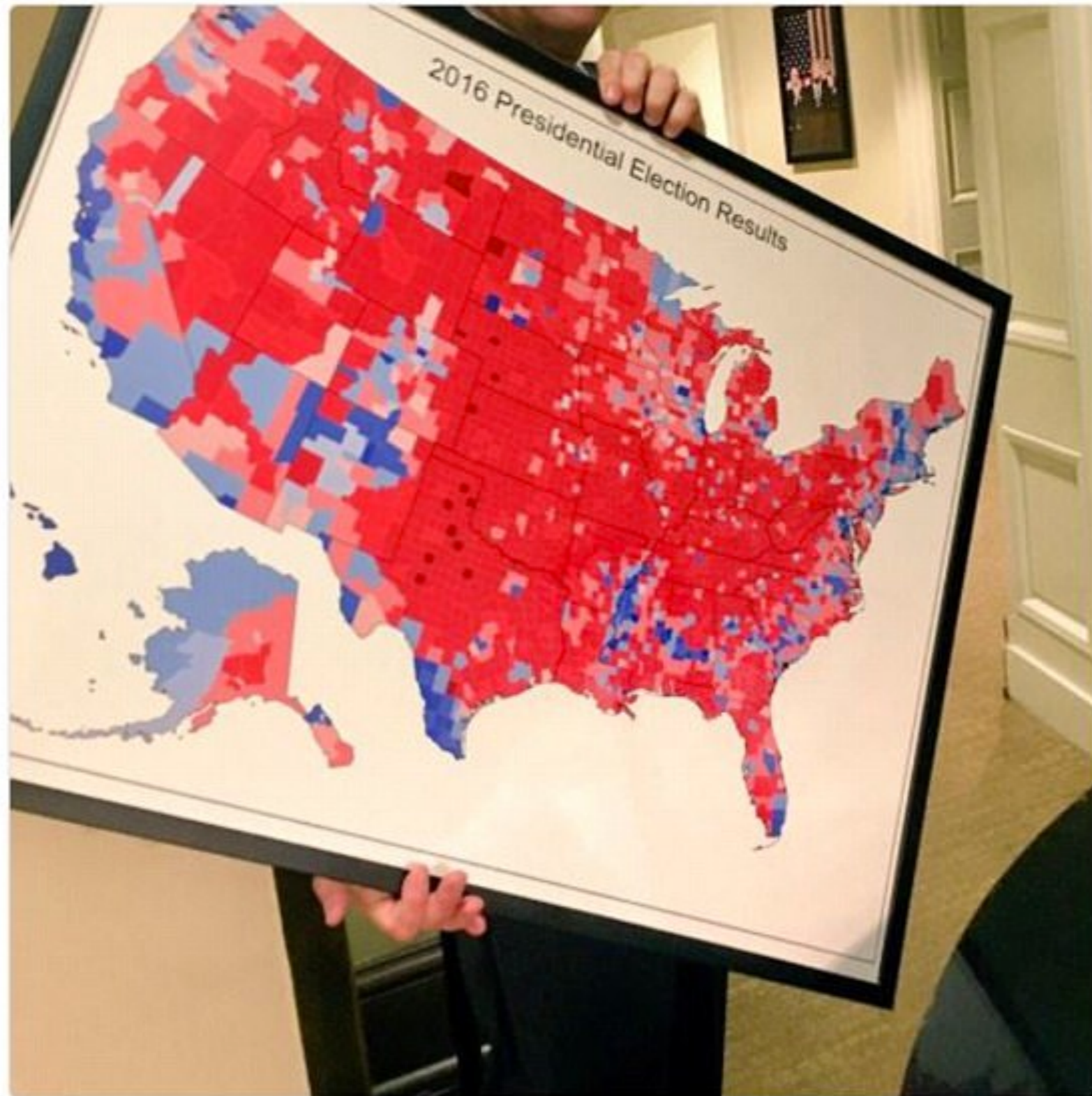
<http://www.thefunctionalart.com/2015/10/double-axes-double-mischief.html>

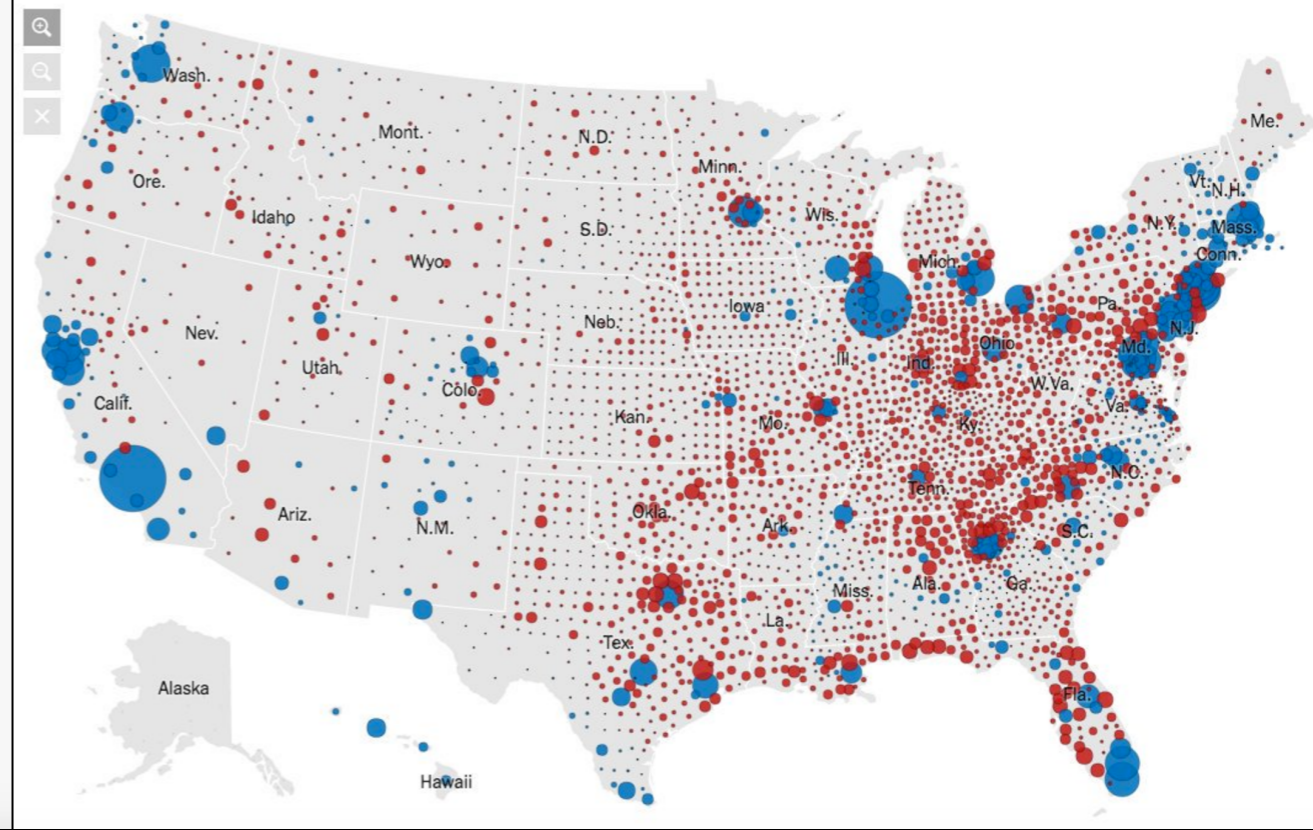
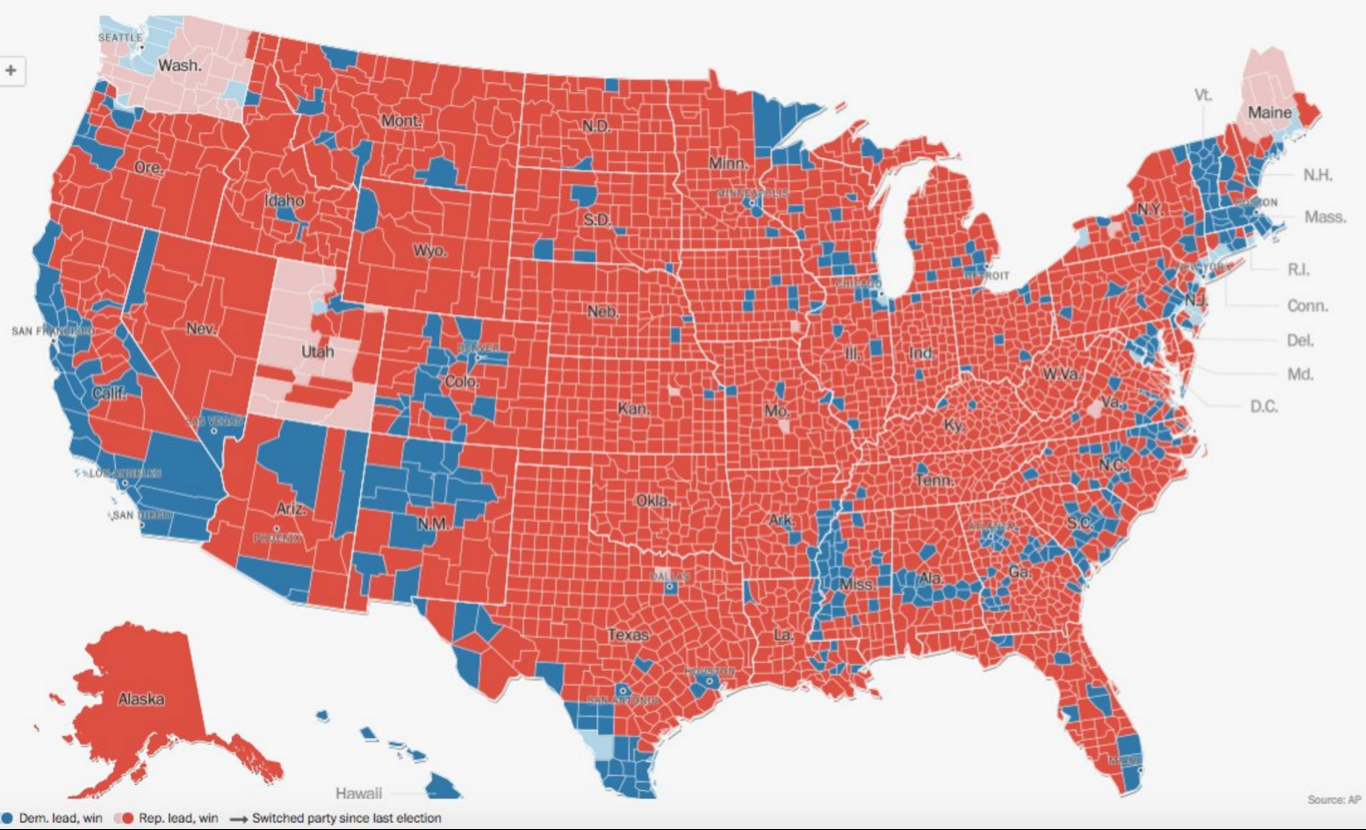
Be Proportional



Trey Yingst [@TreyYingst](#) · May 11

Spotted: A map to be hung somewhere in the West Wing





US Presidential Election 2016

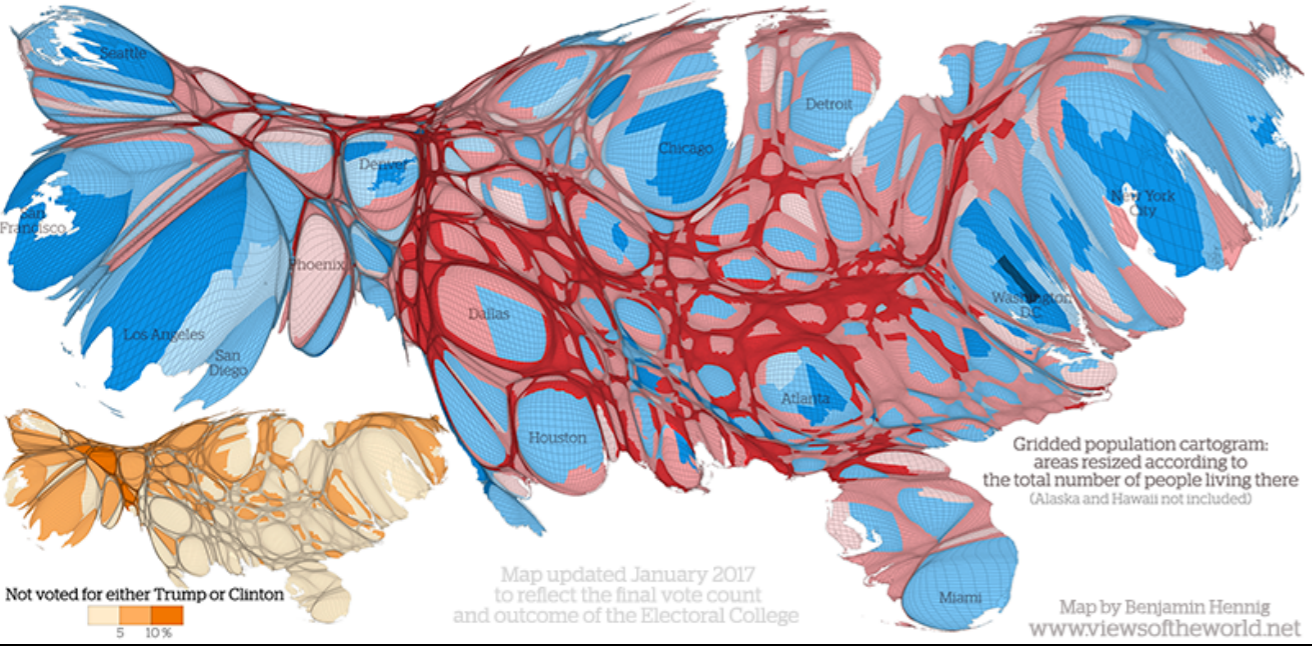
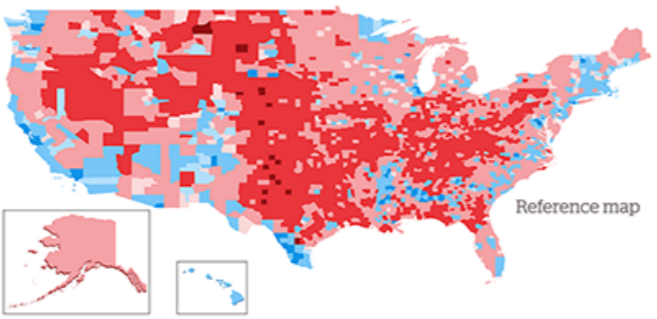
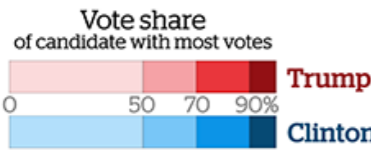
Results mapped at county level showing the candidate with the largest vote share in each area

Overall result:

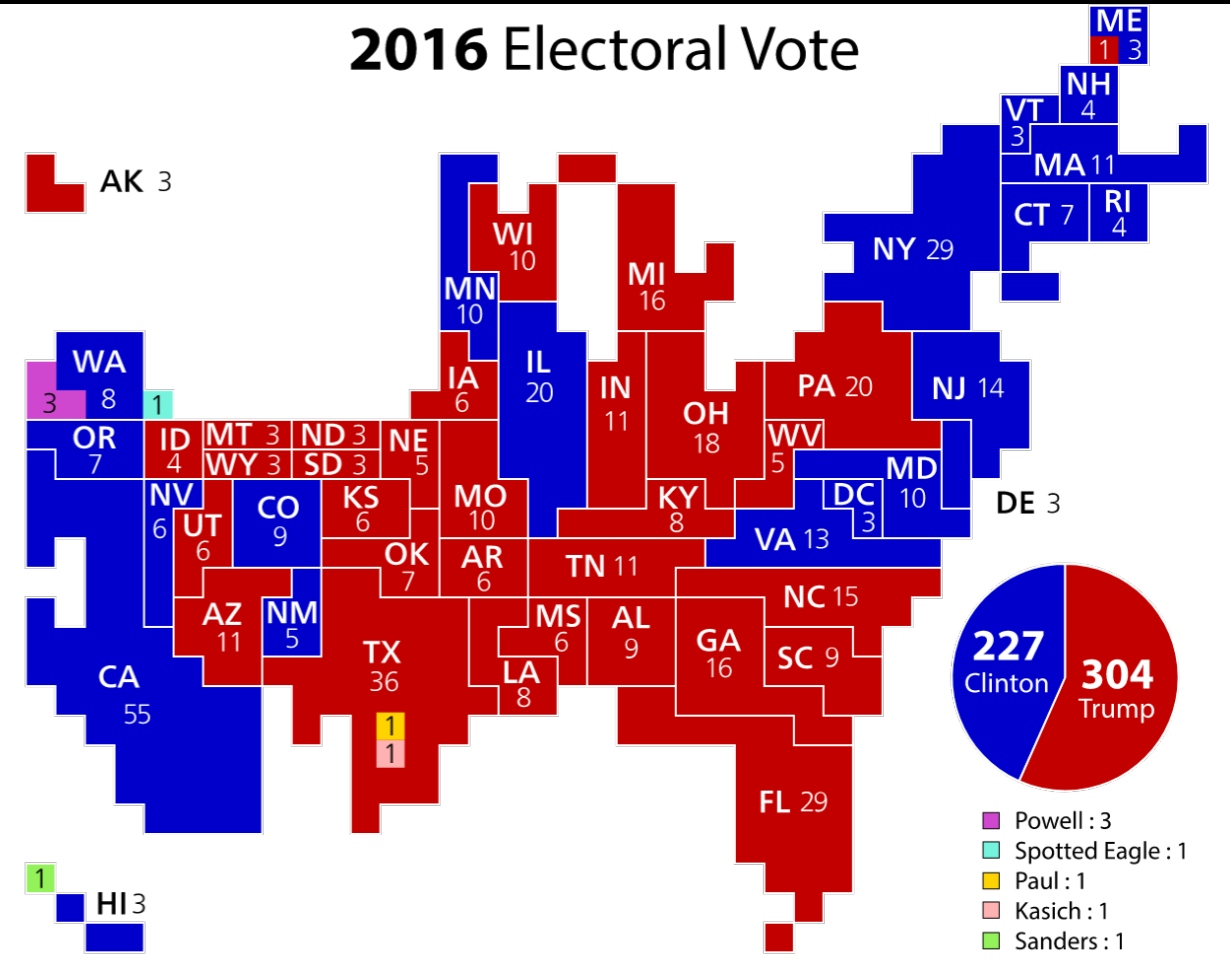
Trump
62,979,636 votes (46.1%)
306 electoral votes
(received 304 in the Electoral College)

Clinton
65,844,610 votes (48.2%)
232 electoral votes
(received 227 in the Electoral College)

Other candidates
7,804,213 votes (5.7%)



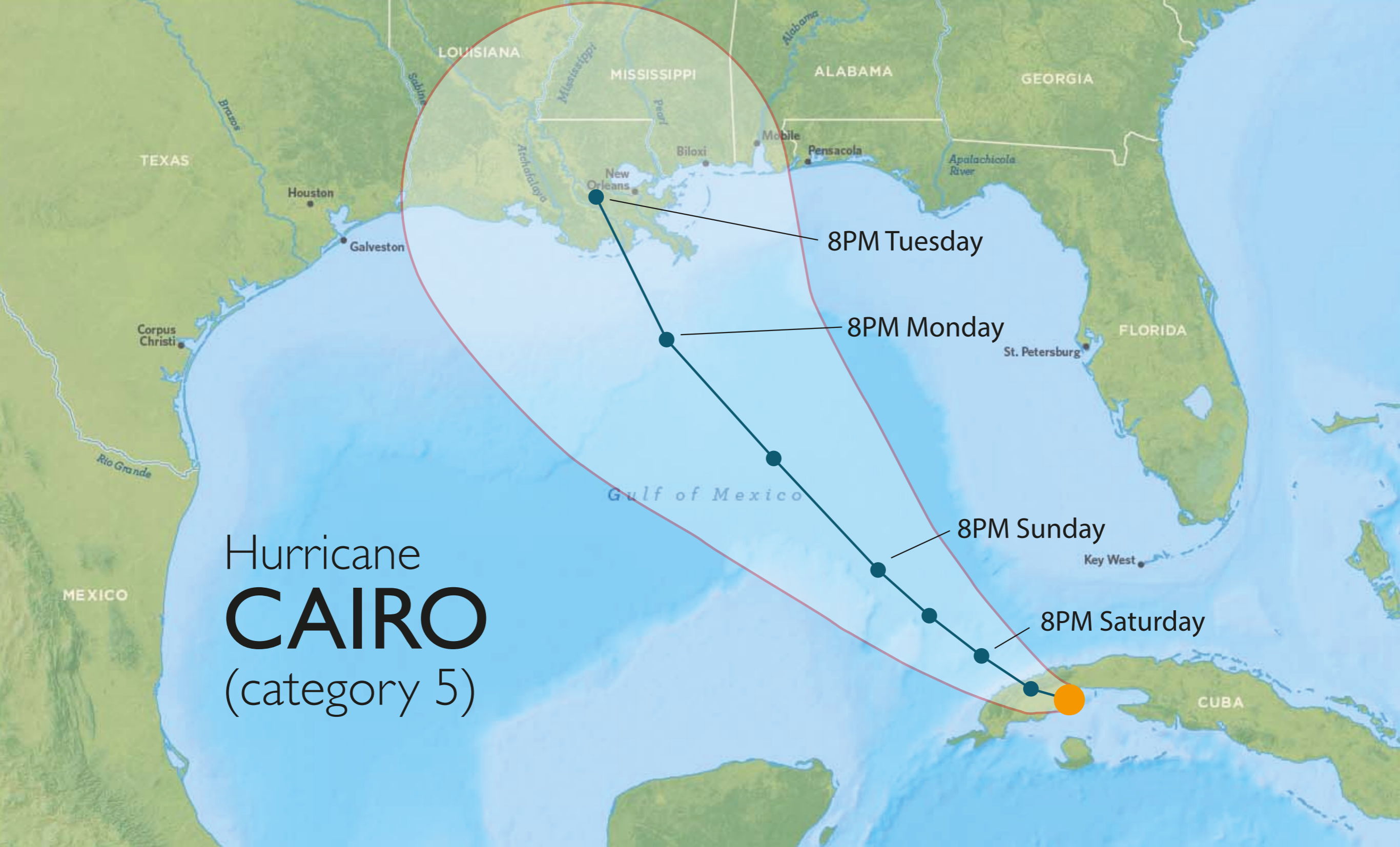
2016 Electoral Vote



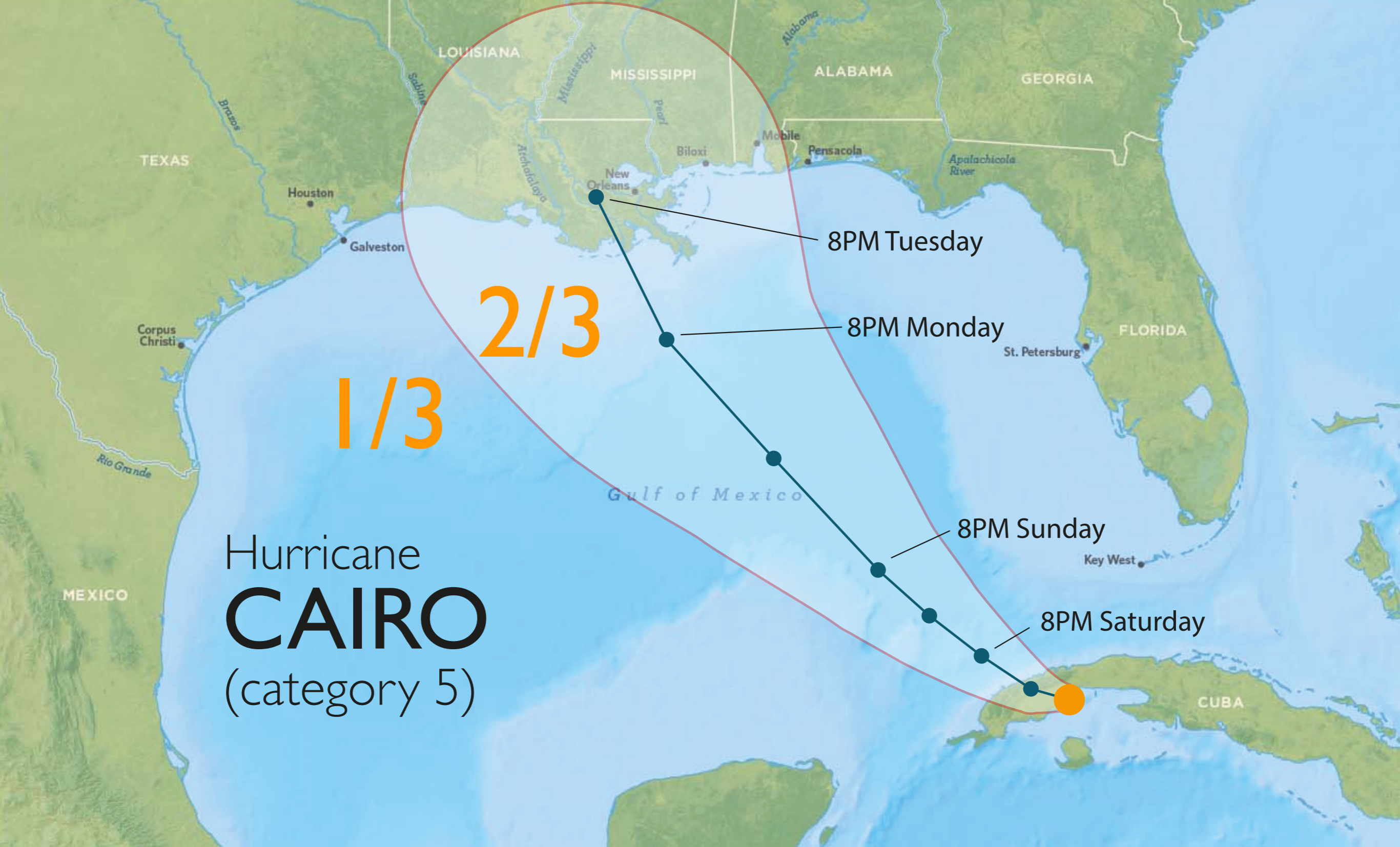
Include Uncertainty

Note: The cone contains the probable path of the storm center but does not show the size of the storm. Hazardous conditions can occur outside of the cone.





What you show



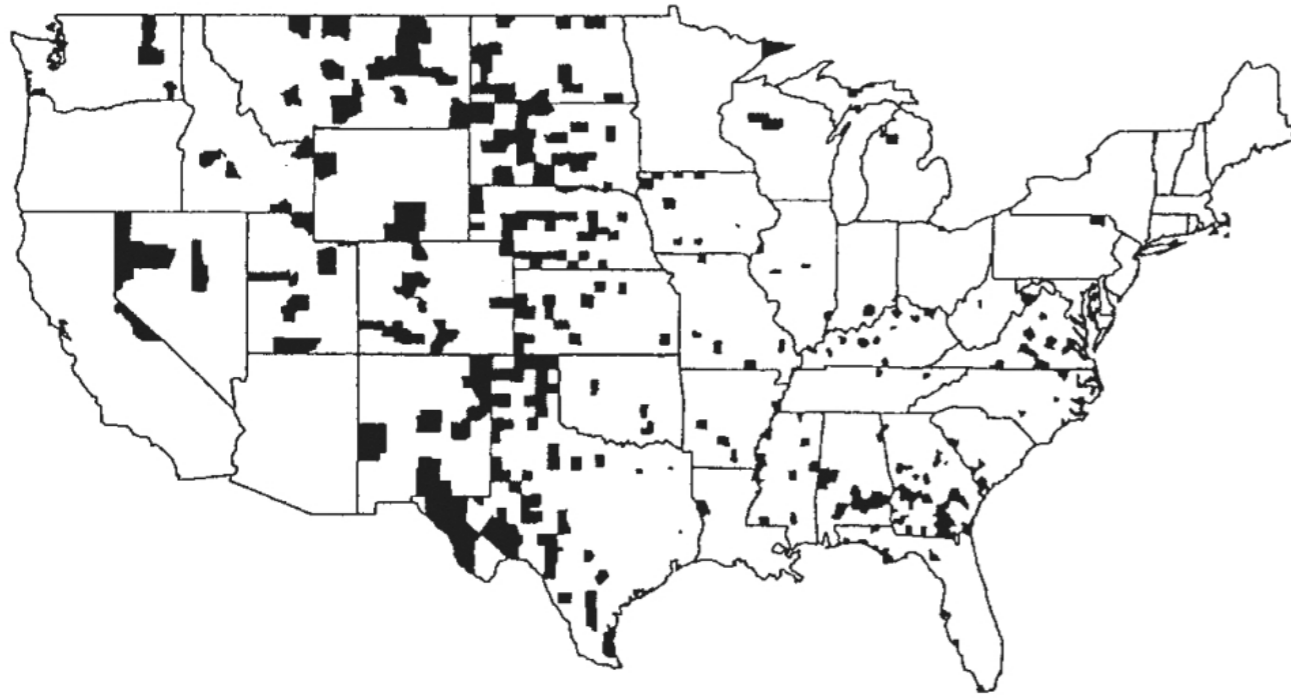
What non-scientists are not aware of (cone is just 66% probability)



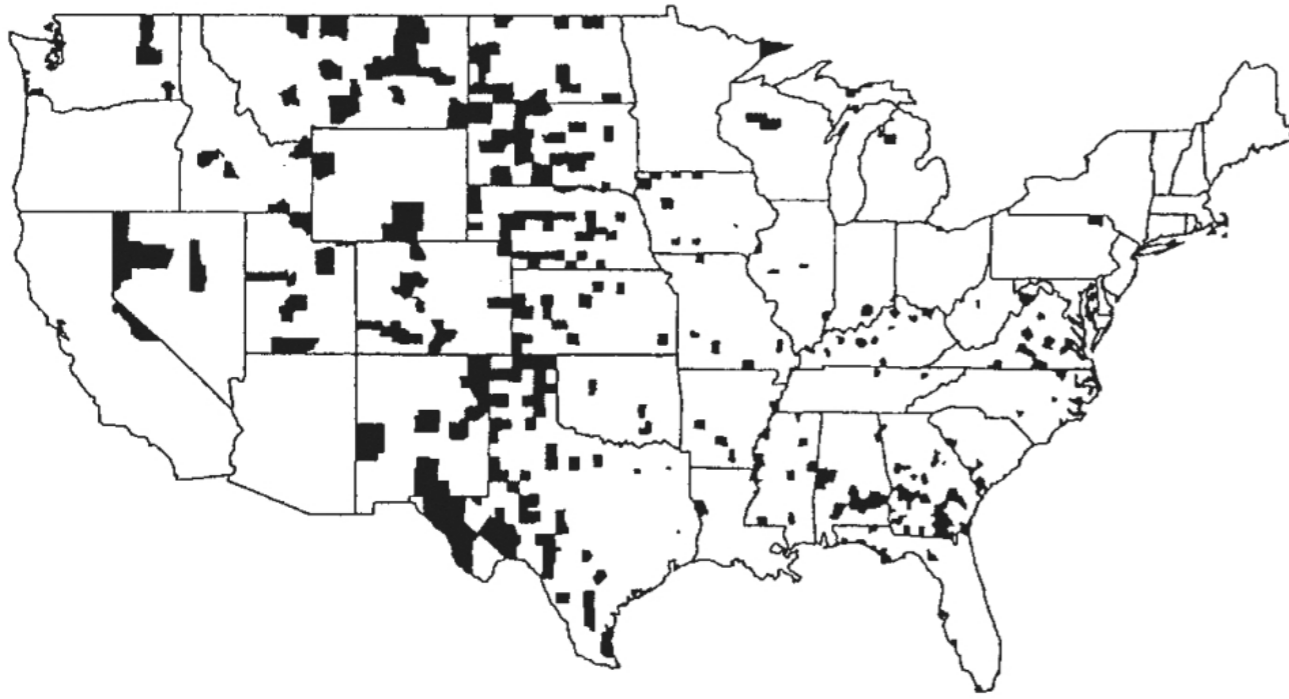
Hurricane
CAIRO
(category 5)

What we could be showing instead

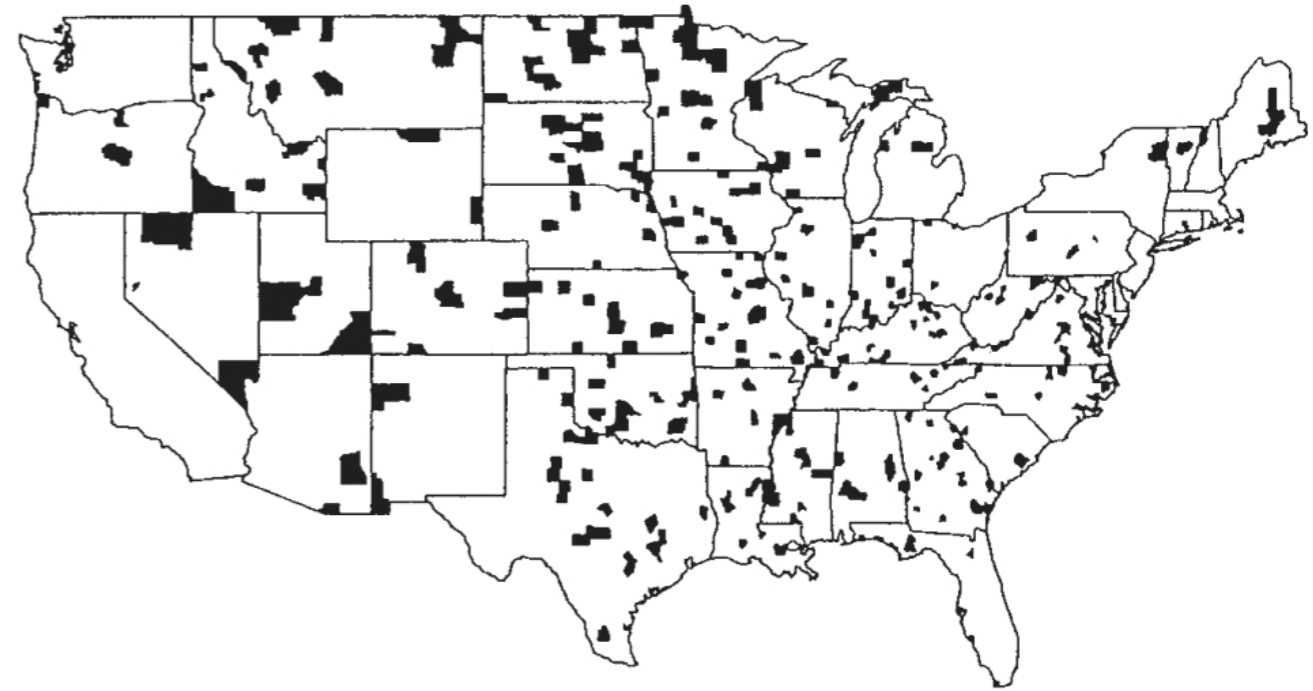
Plot all your data



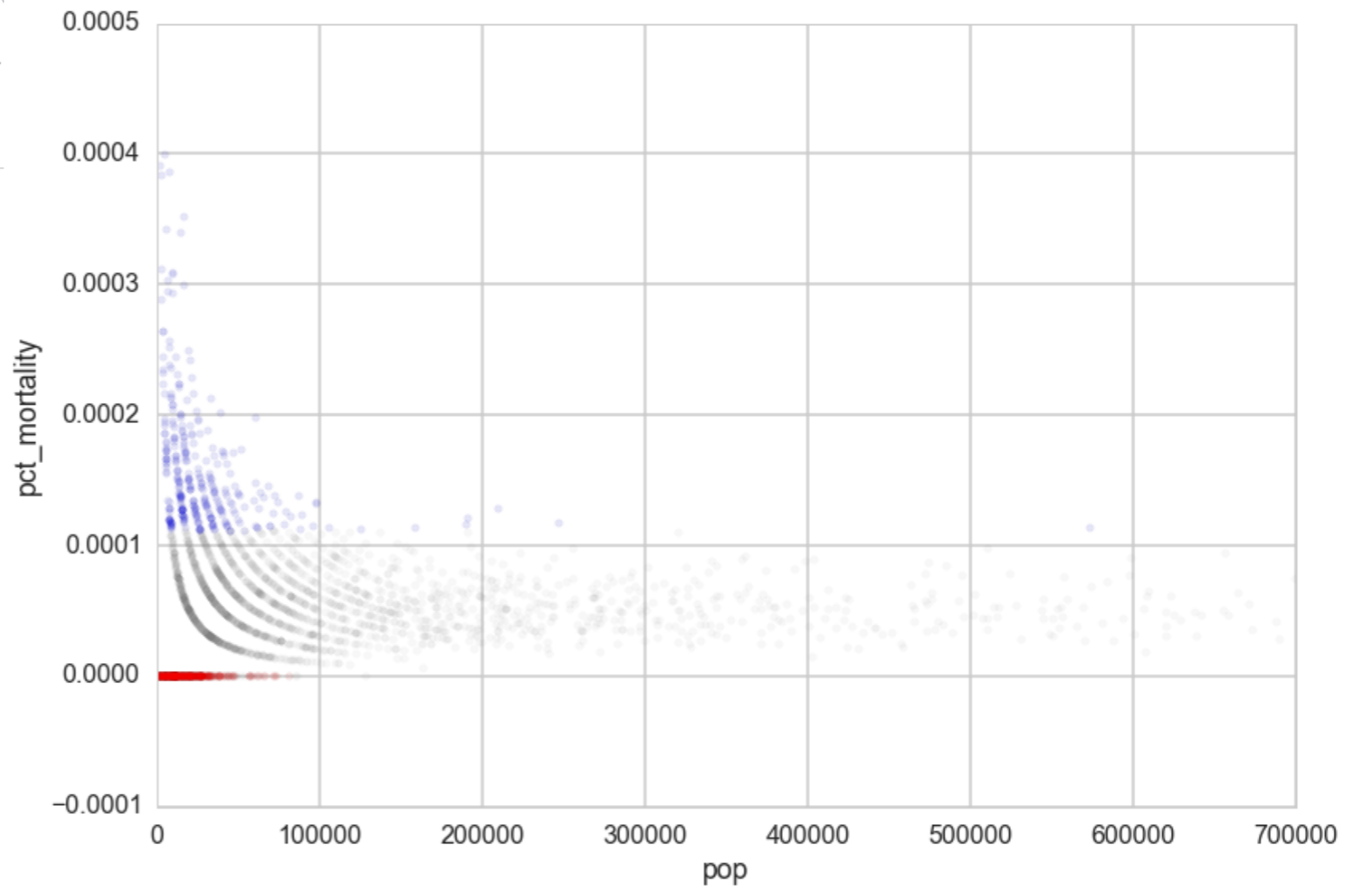
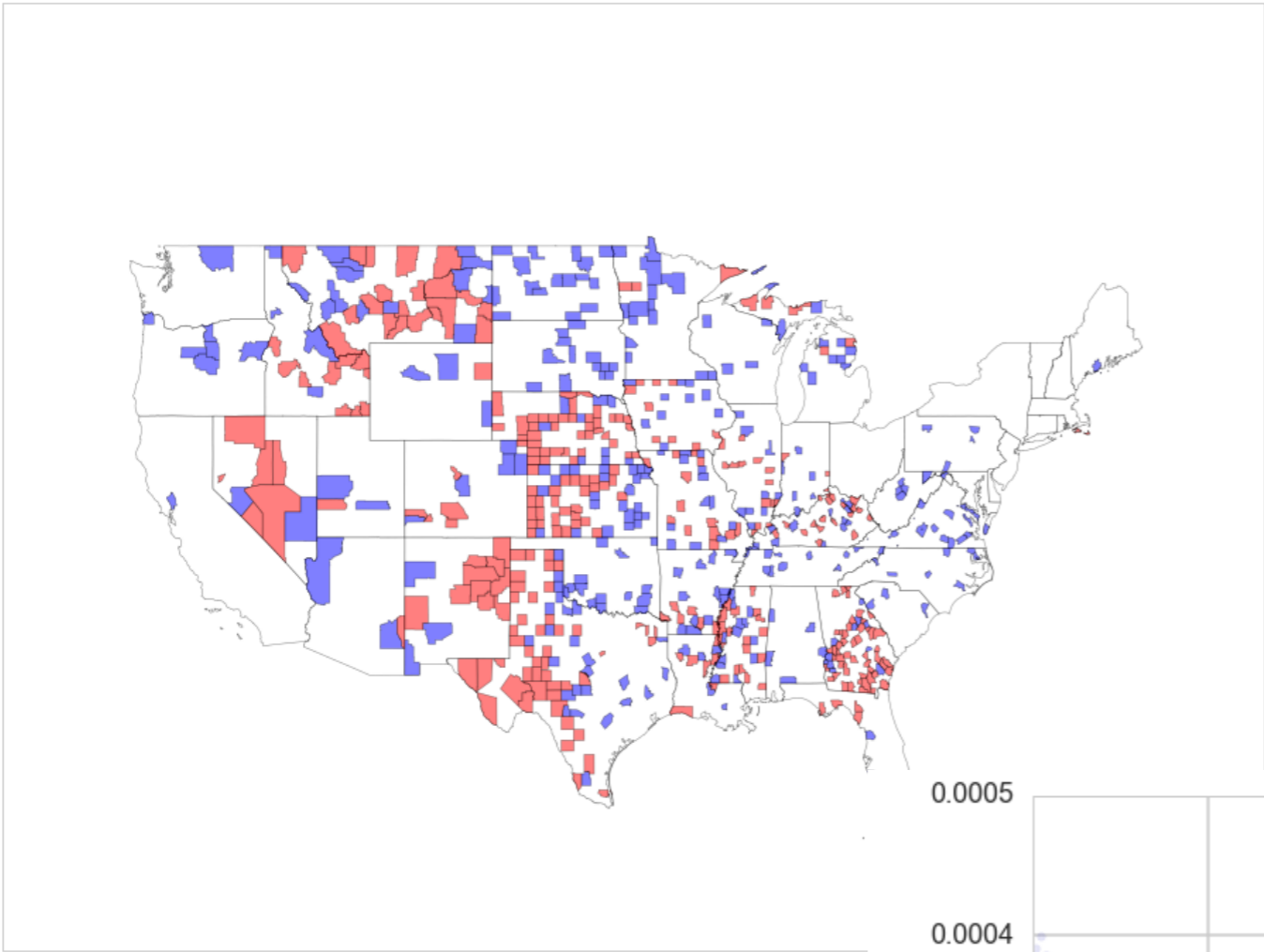
Counties with the LOWEST
kidney cancer death rates
(1980-1989)



Counties with the **LOWEST**
kidney cancer death rates
(1980-1989)



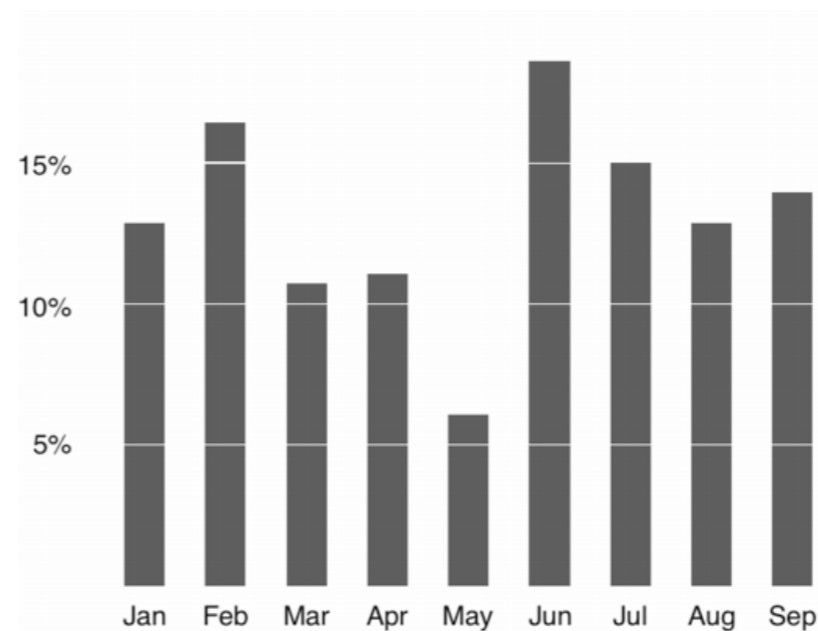
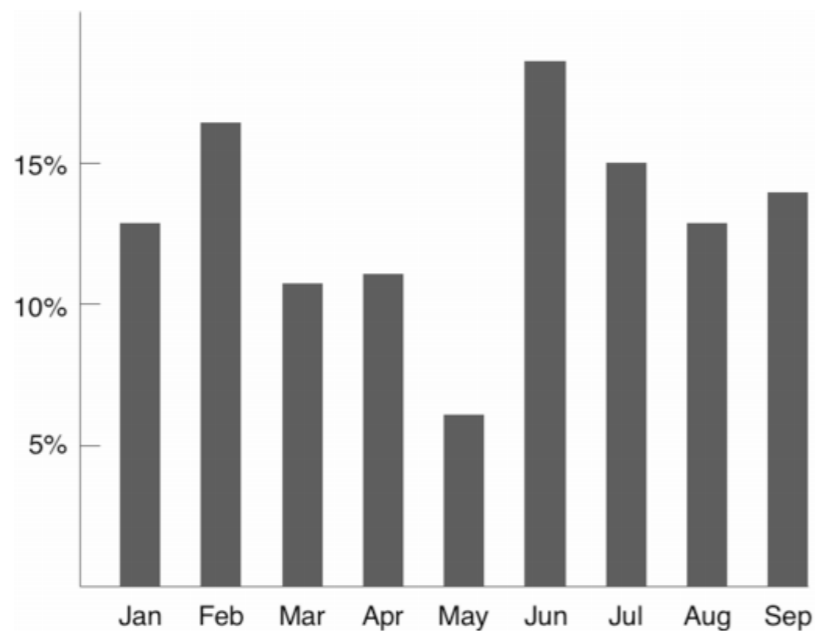
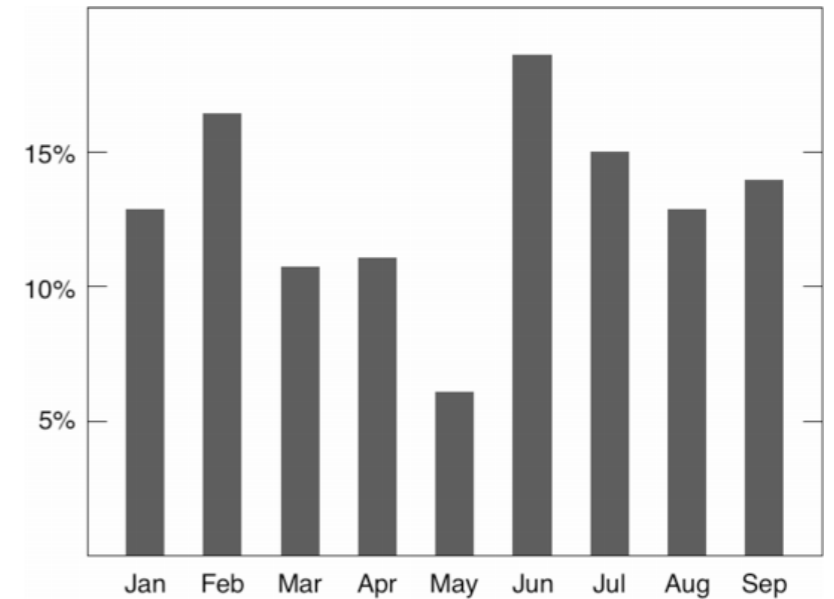
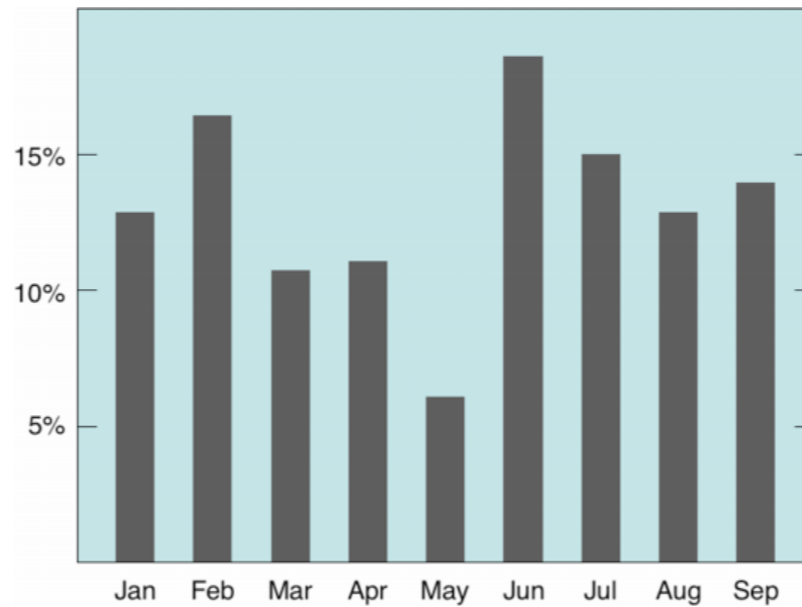
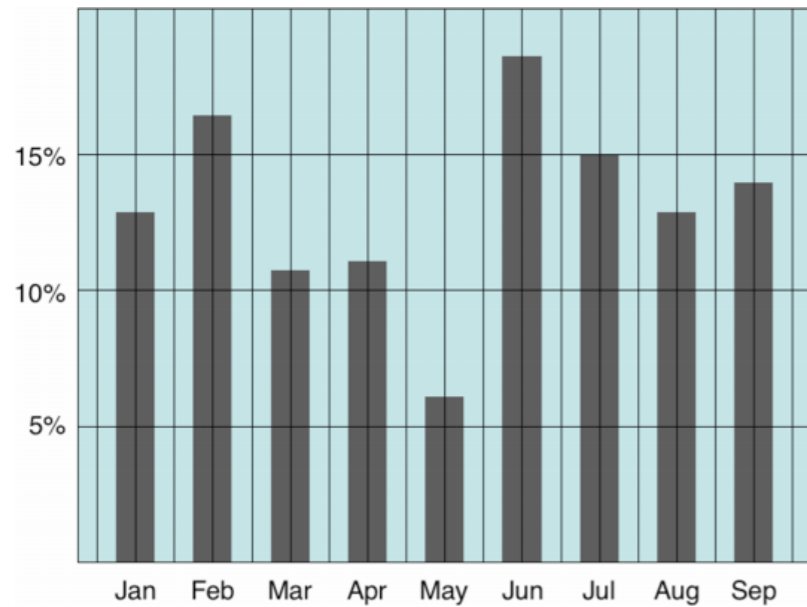
Counties with the **HIGHEST**
kidney cancer death rates
(1980-1989)

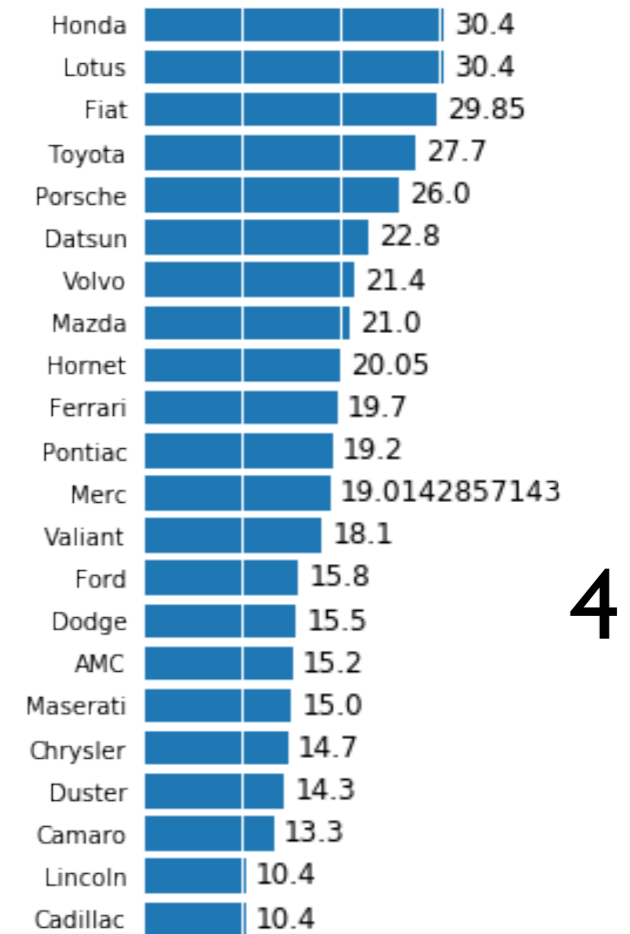
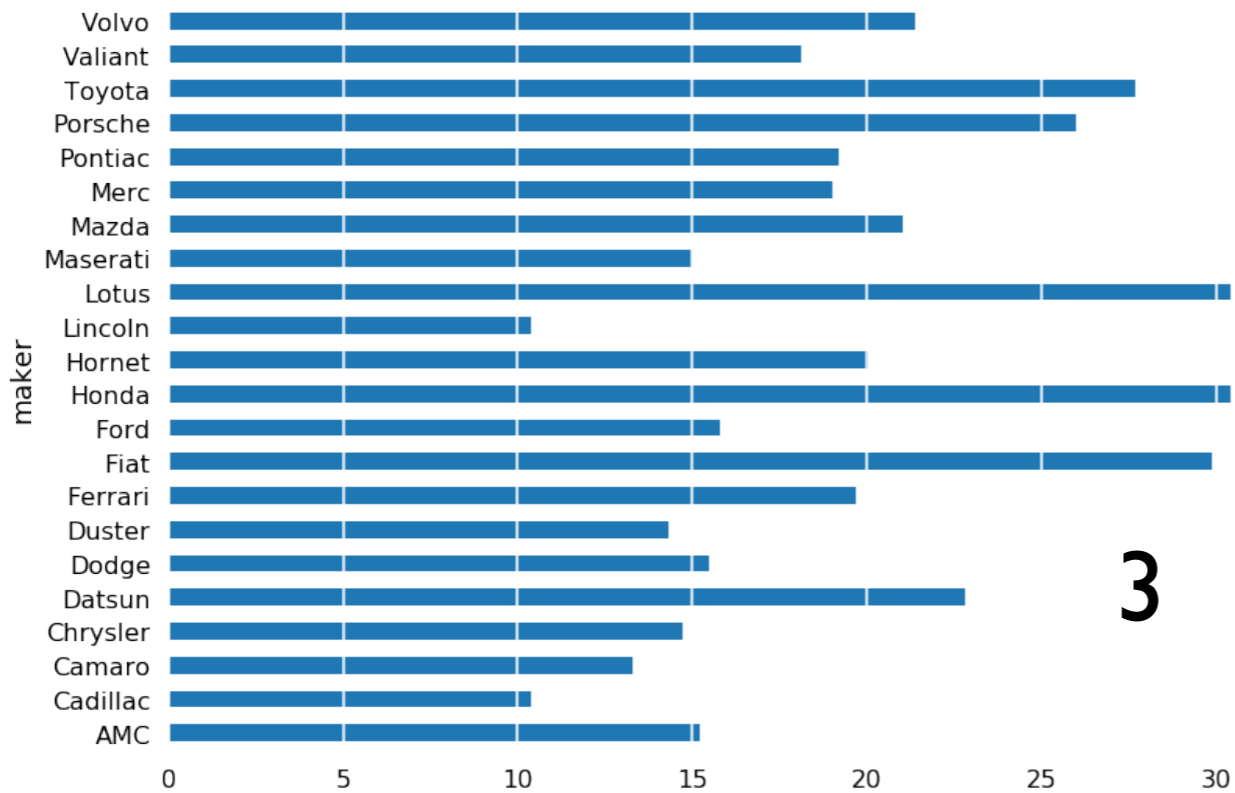
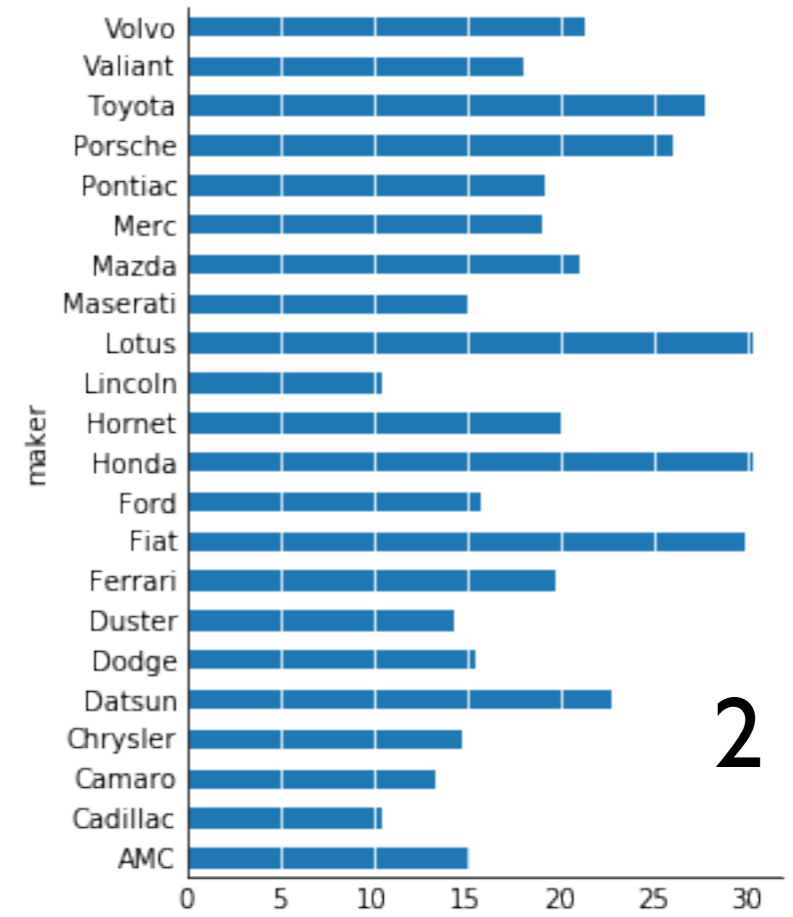
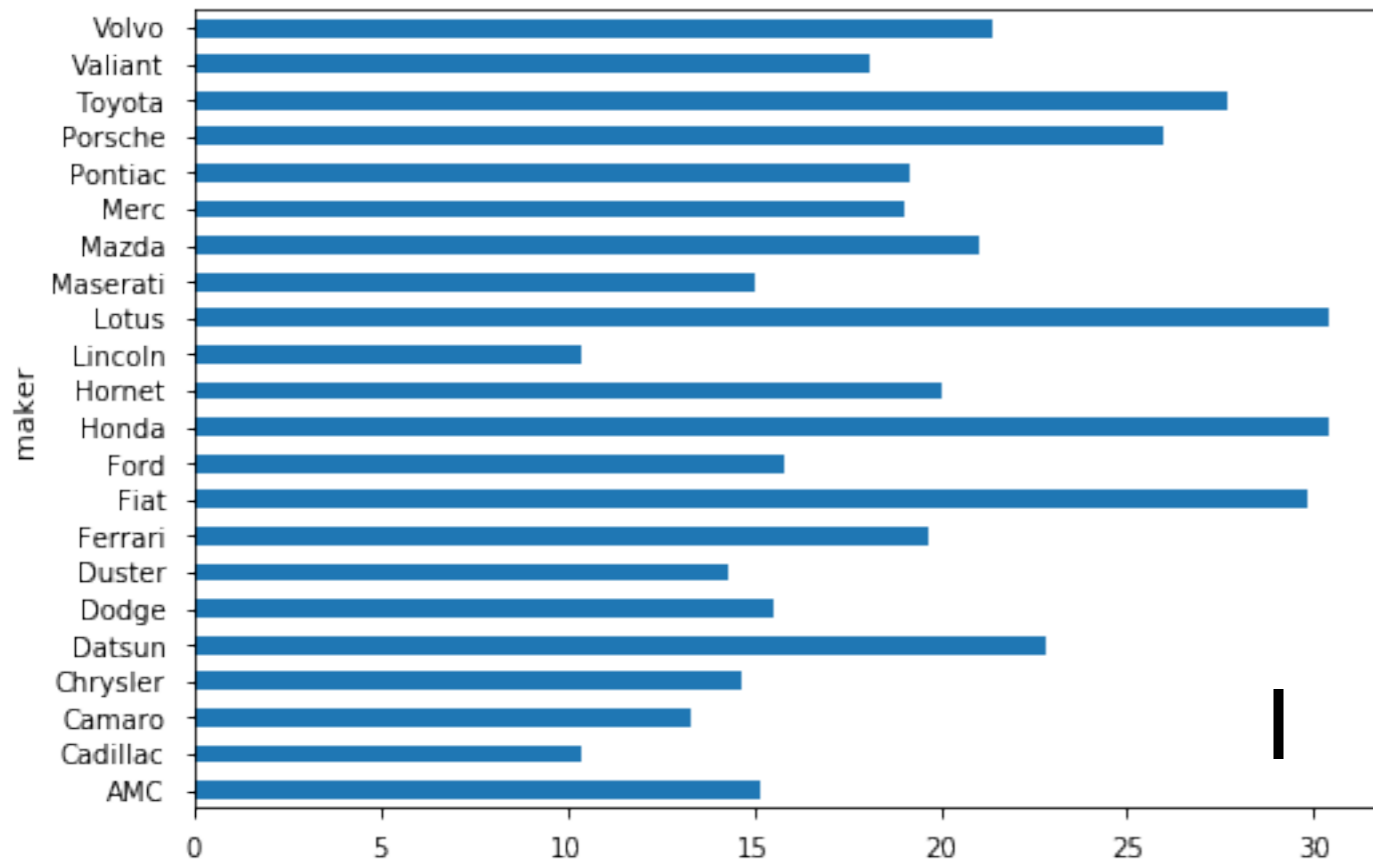


2. Keep It Simple

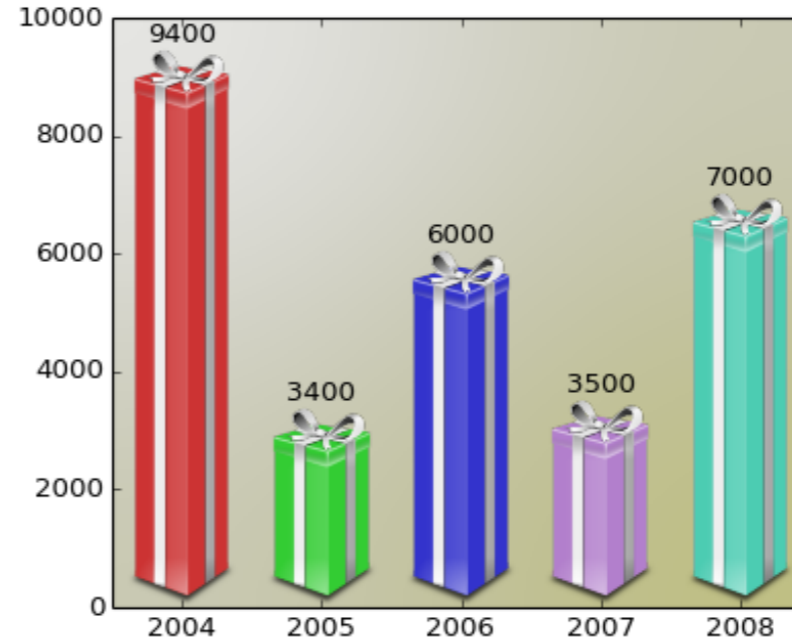
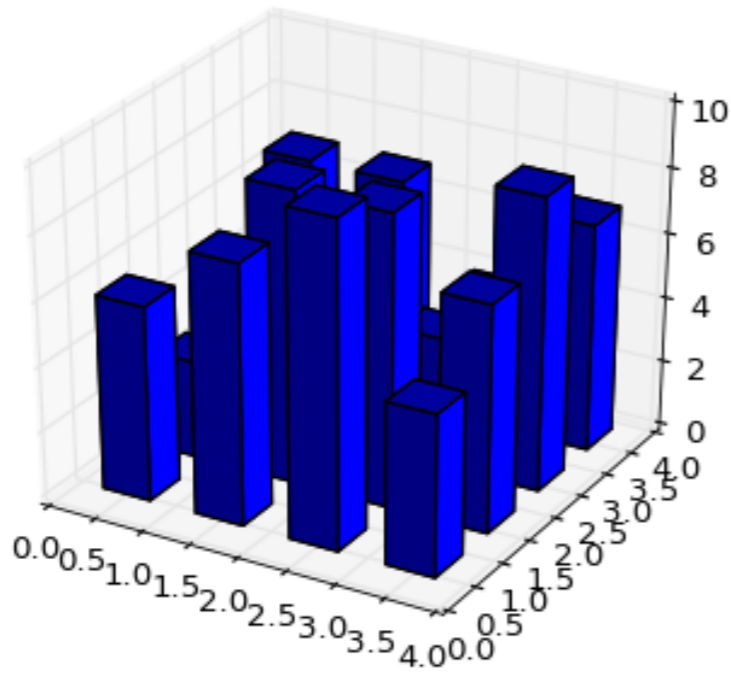
Avoid Chartjunk

Extraneous visual elements that distract from the message



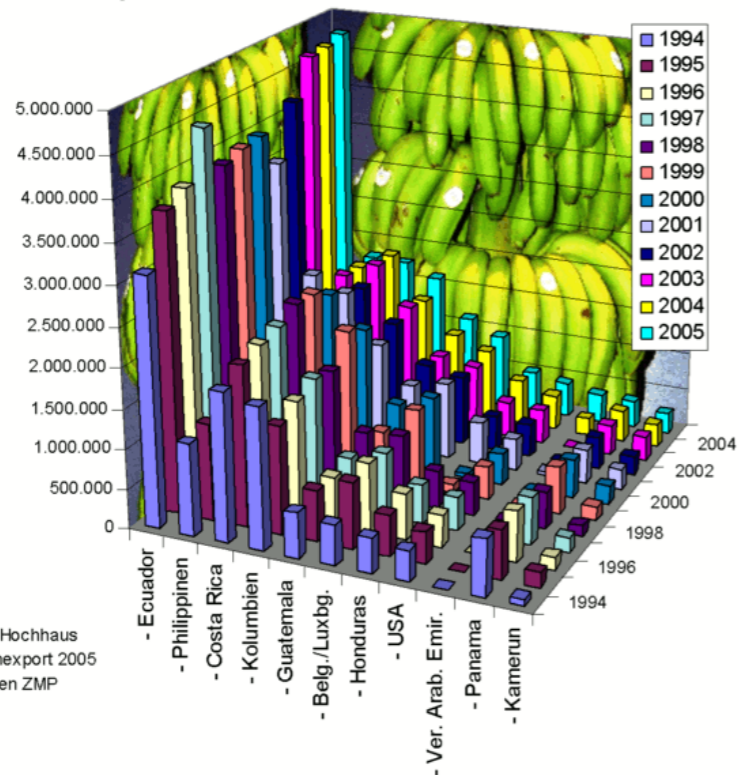


Don't!

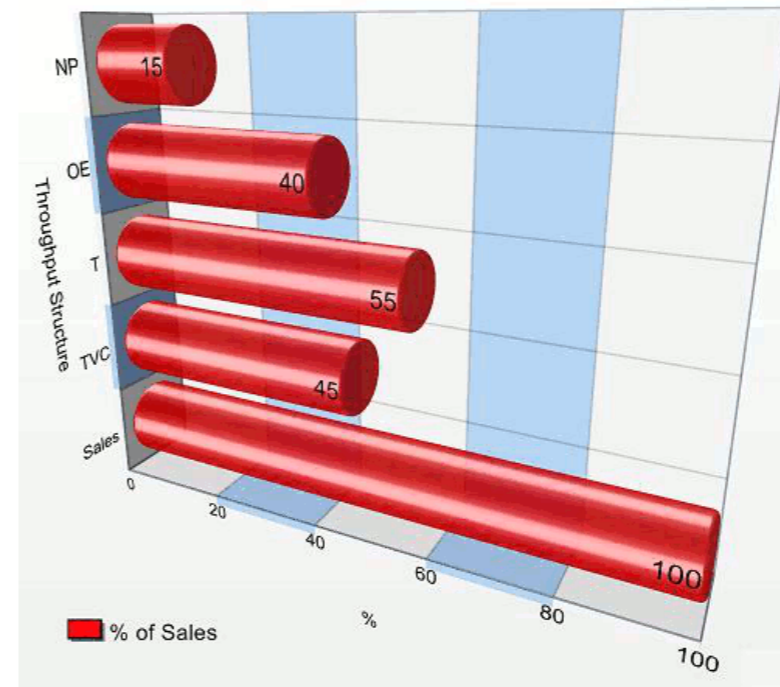


matplotlib gallery

Export von Bananen in Tonnen von 1994-2005



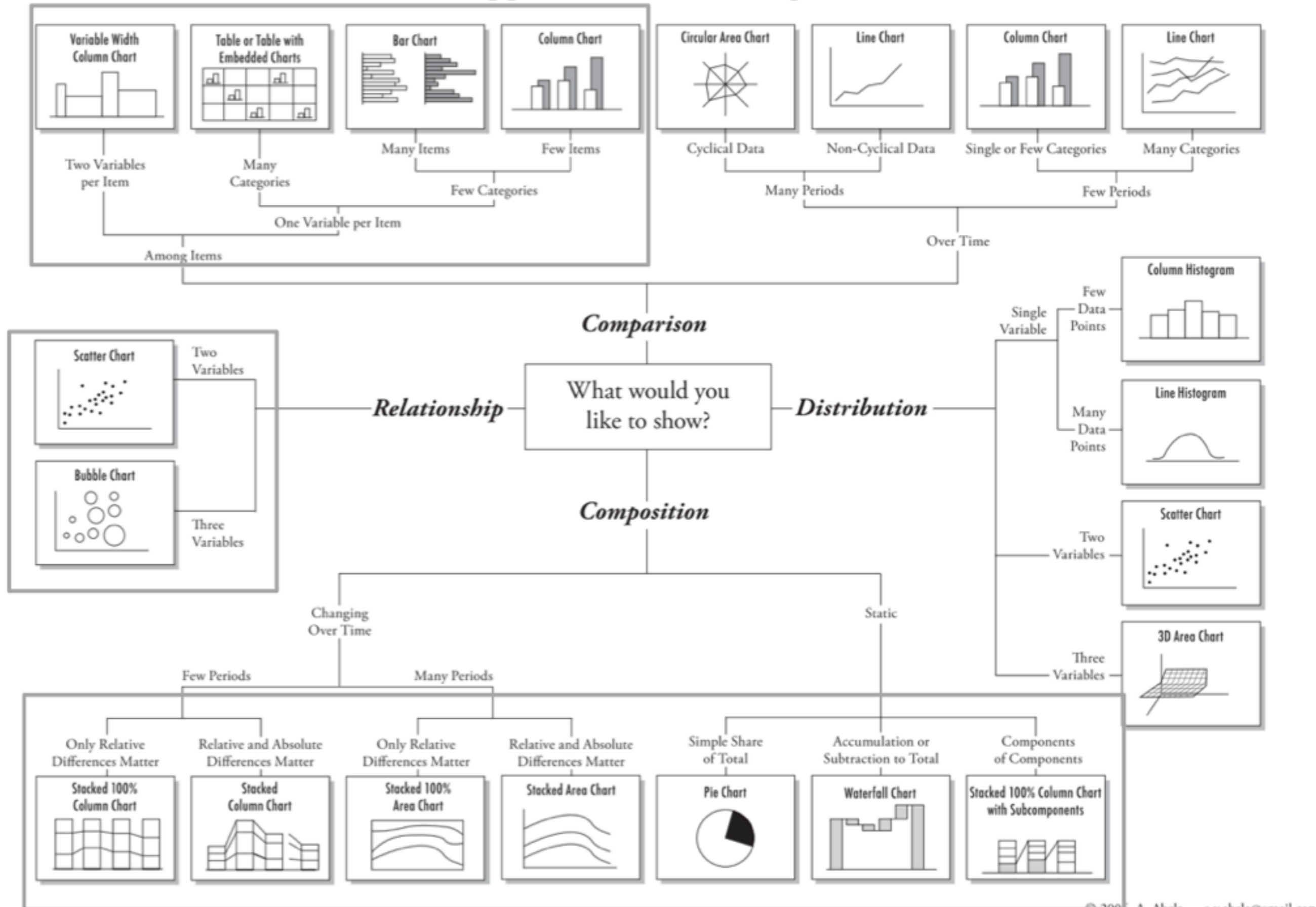
Dr. Hochhaus
Banexport 2005
Daten ZMP



Excel Charts Blog

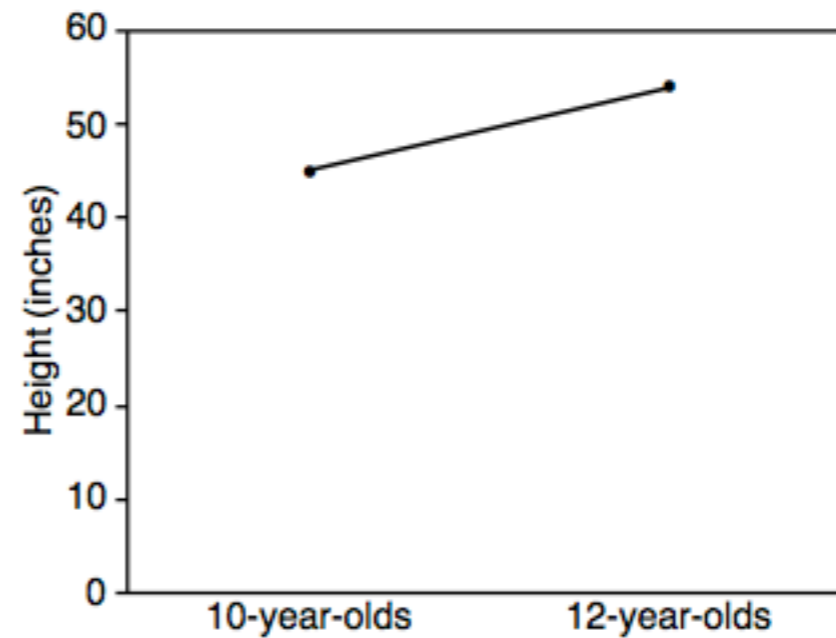
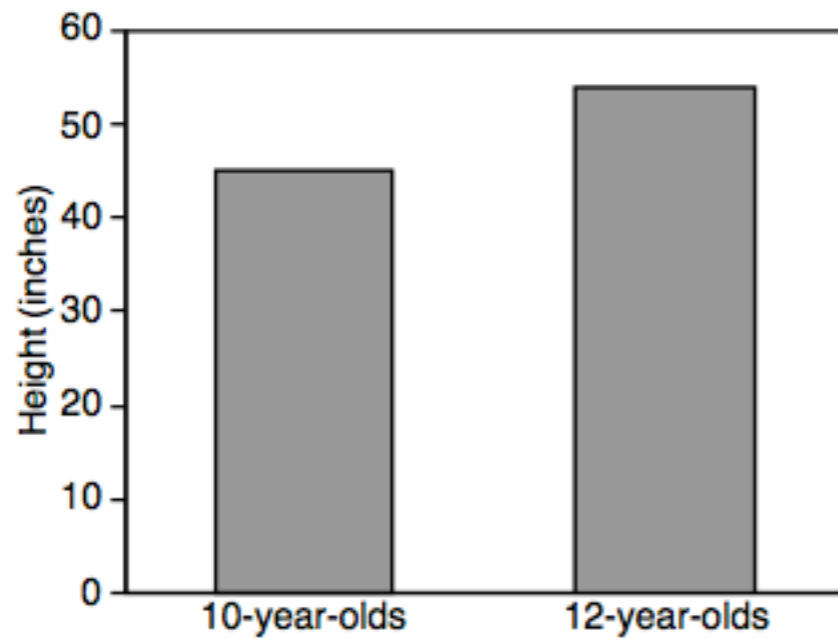
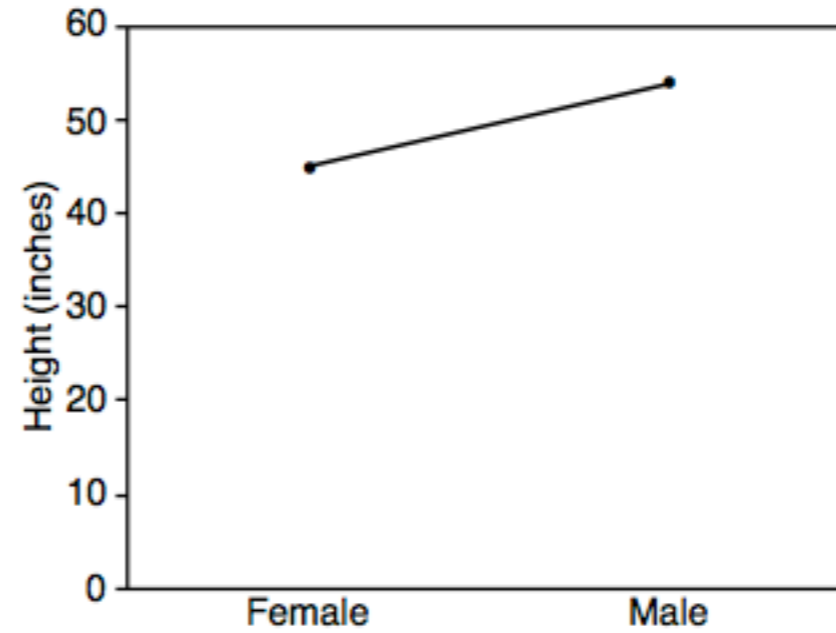
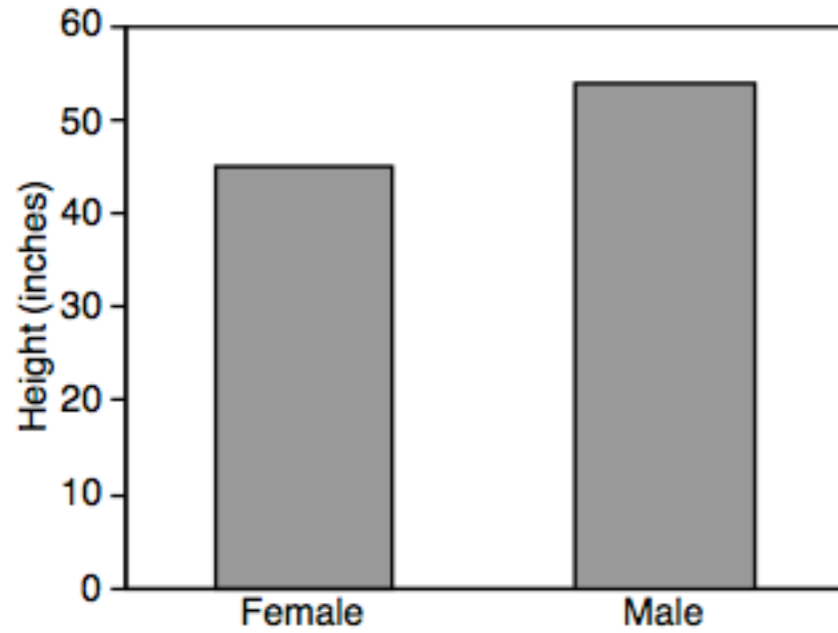
3. Use The Right Display

Chart Suggestions—A Thought-Starter



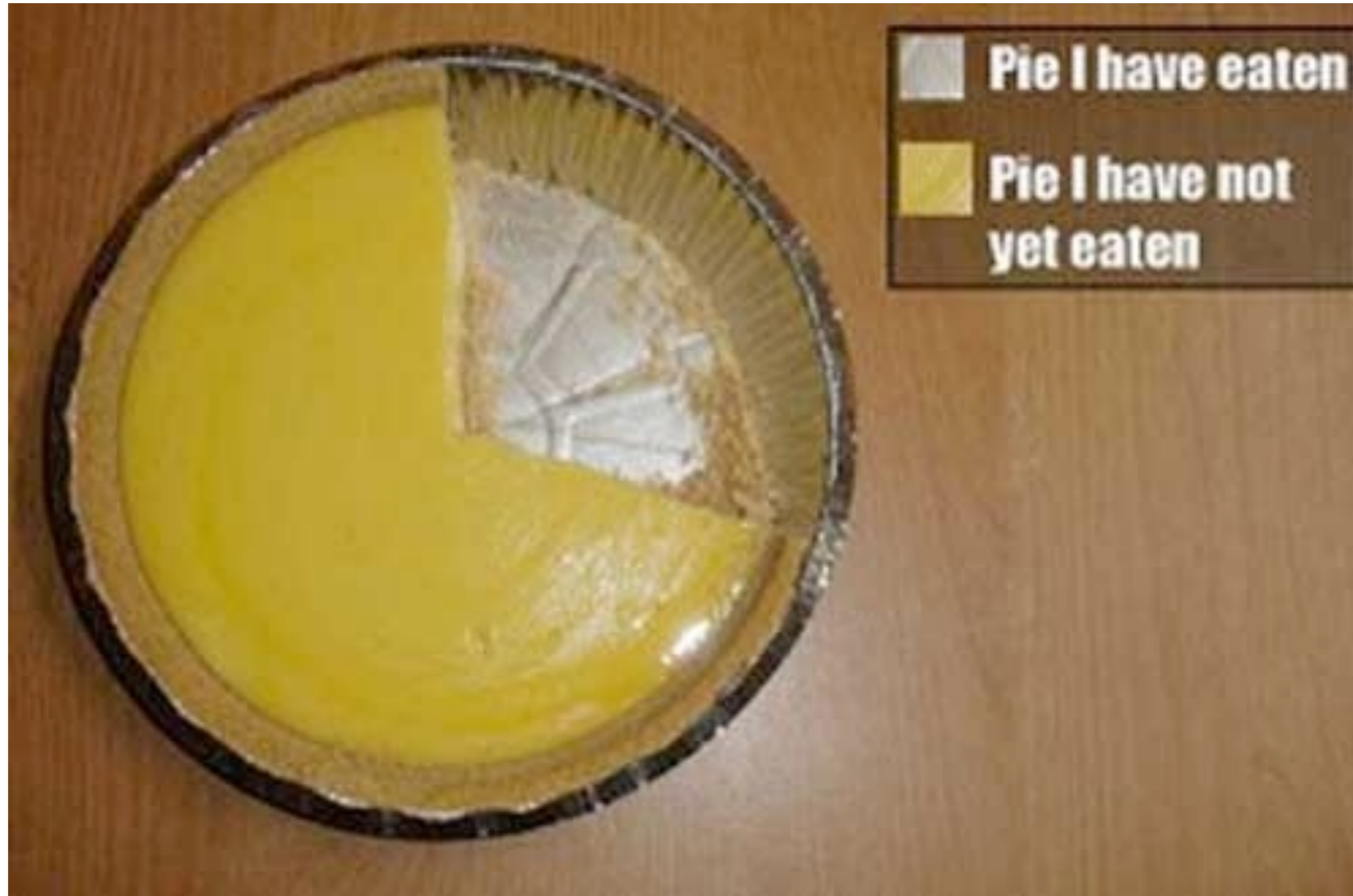
Comparisons

Bars vs. Lines

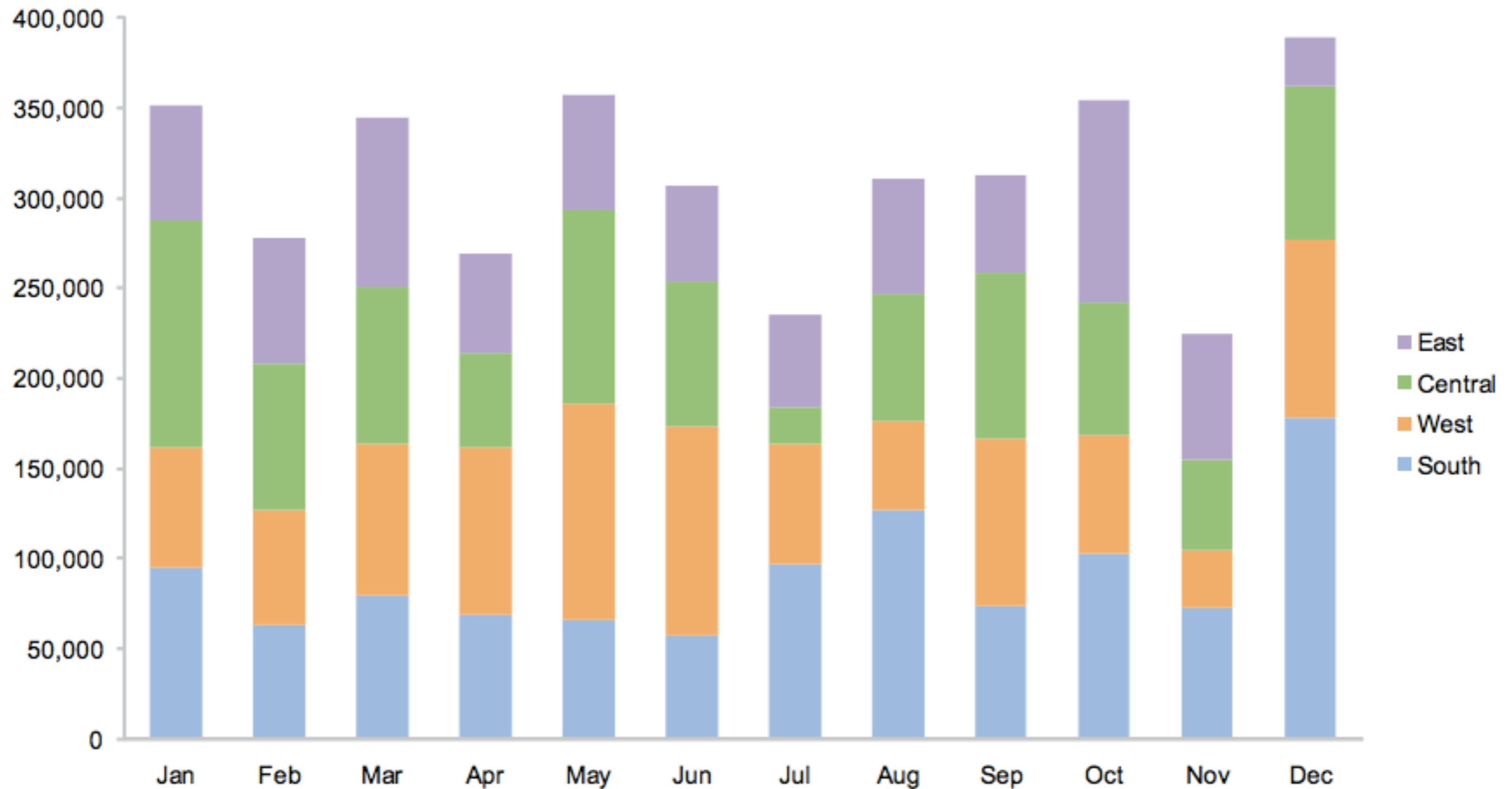


Proportions

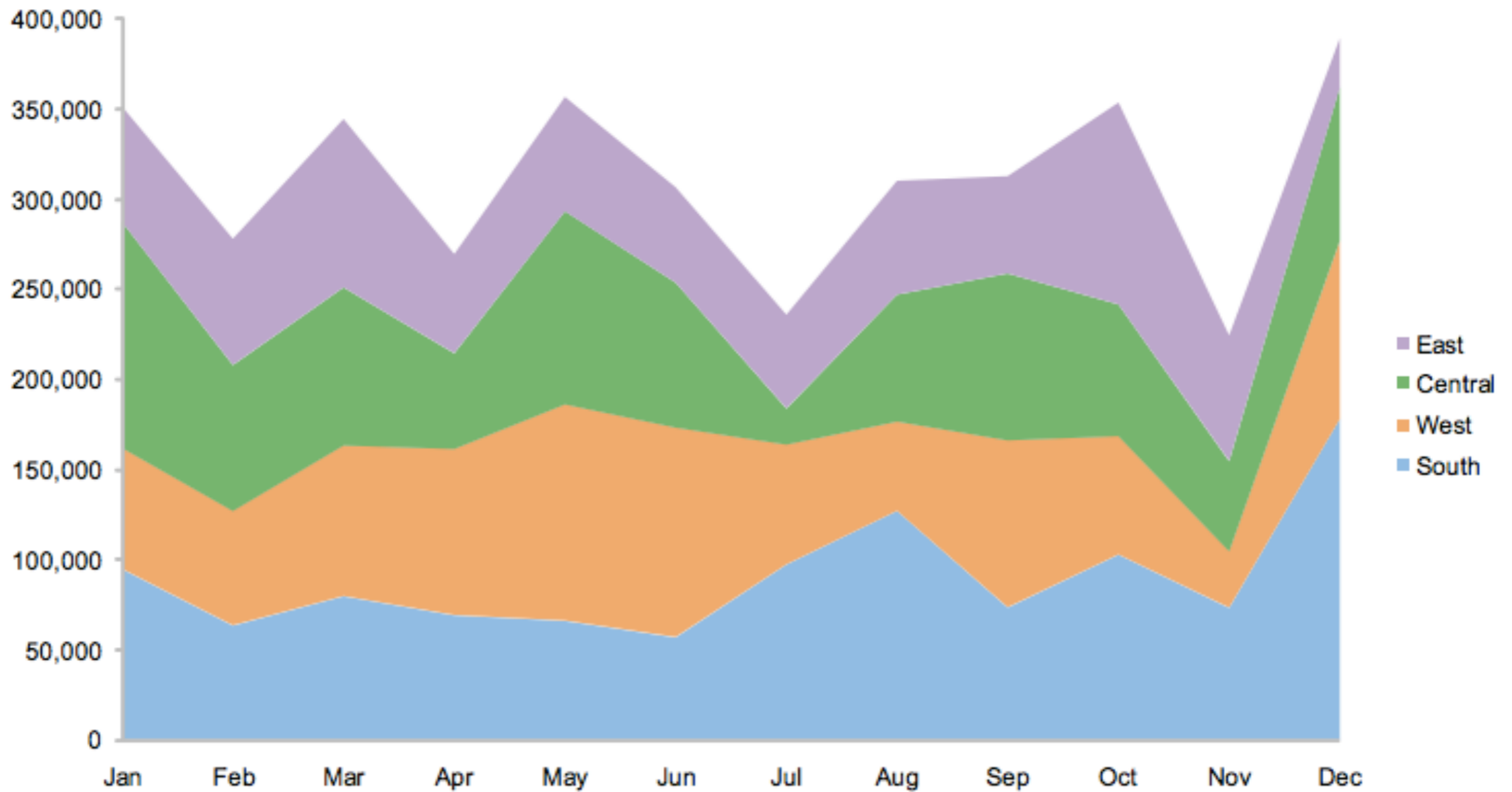
Pie Charts



Stacked Bar Chart

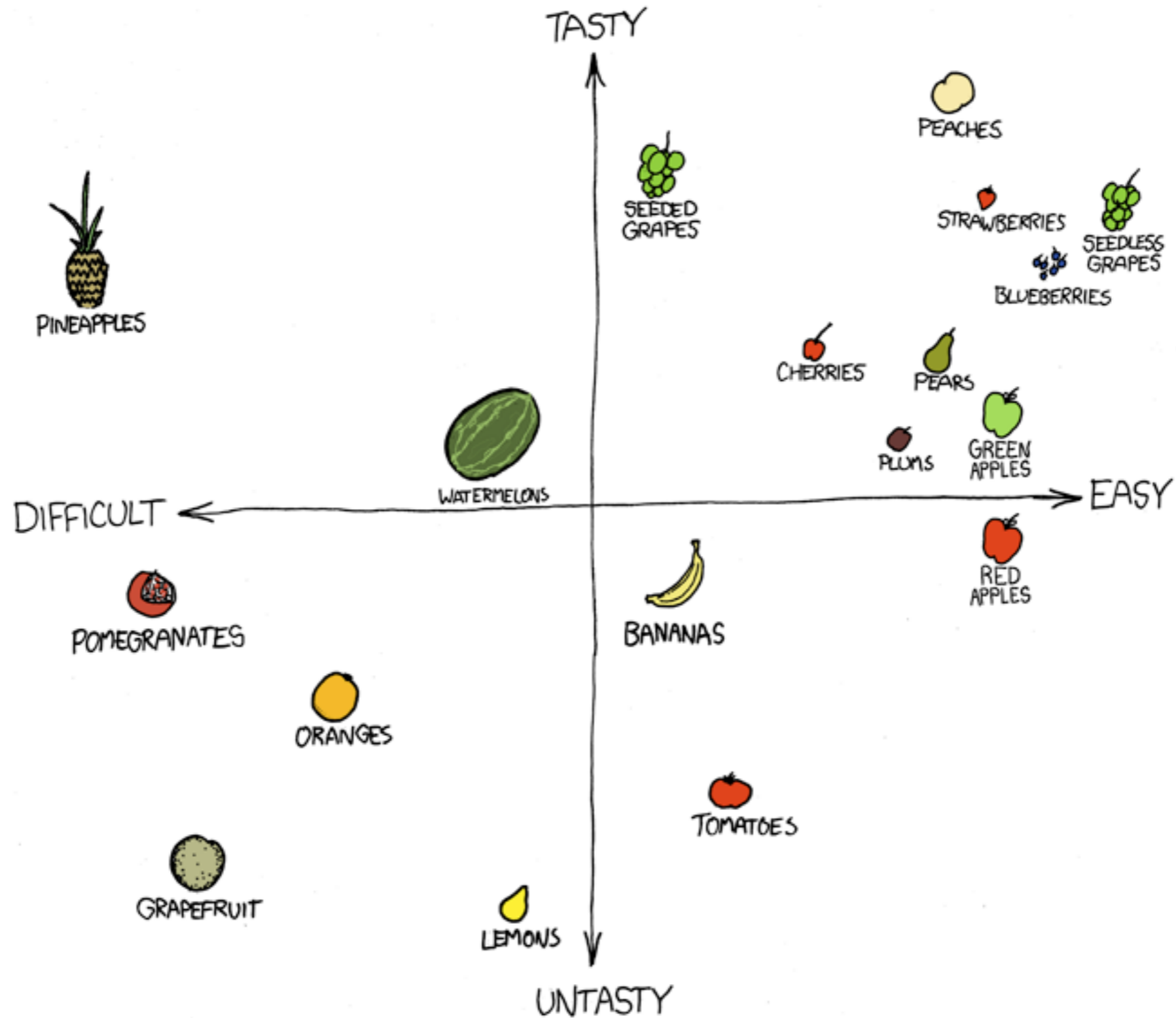


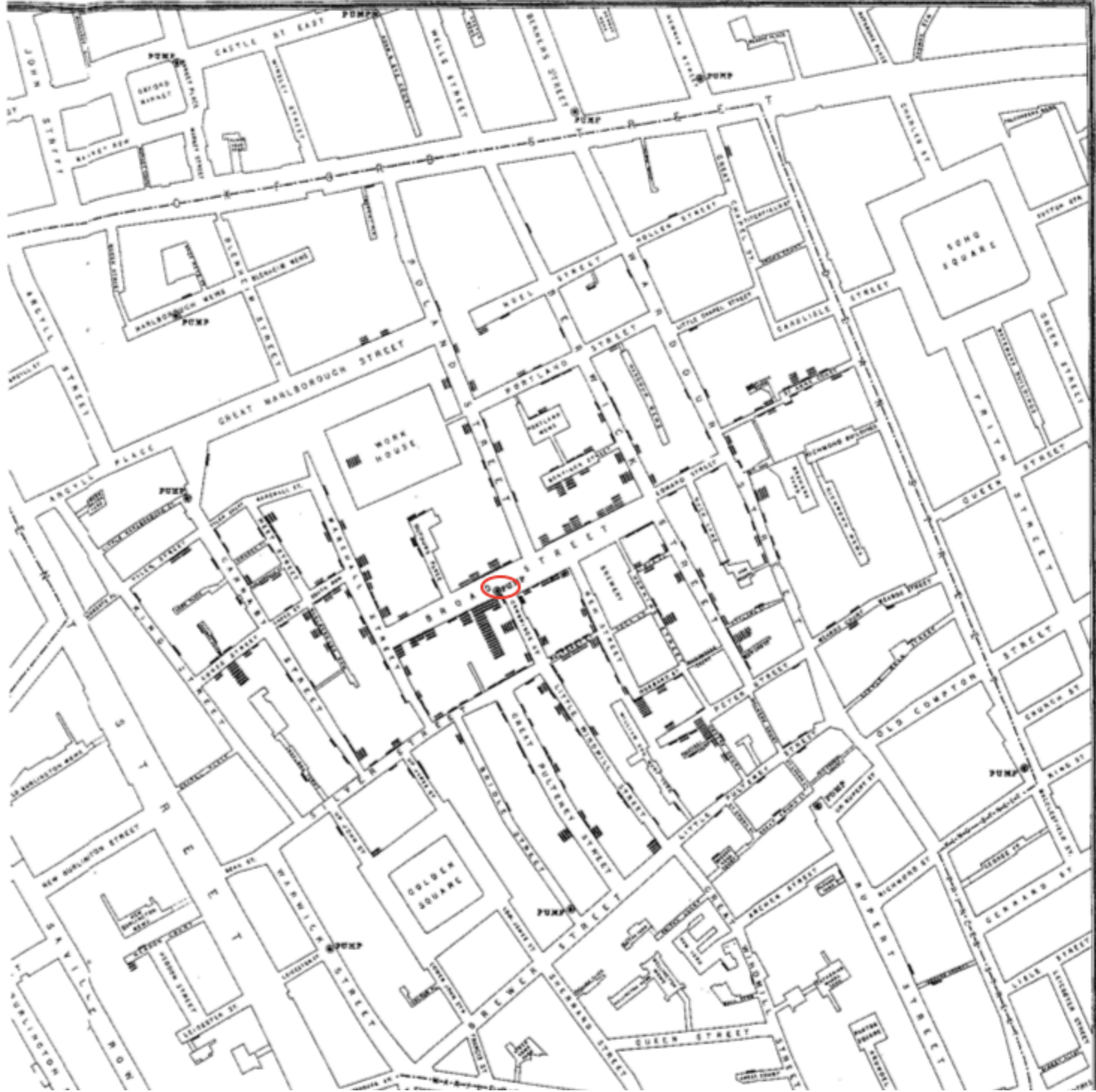
Stacked Area Chart



Correlations

Scatterplots

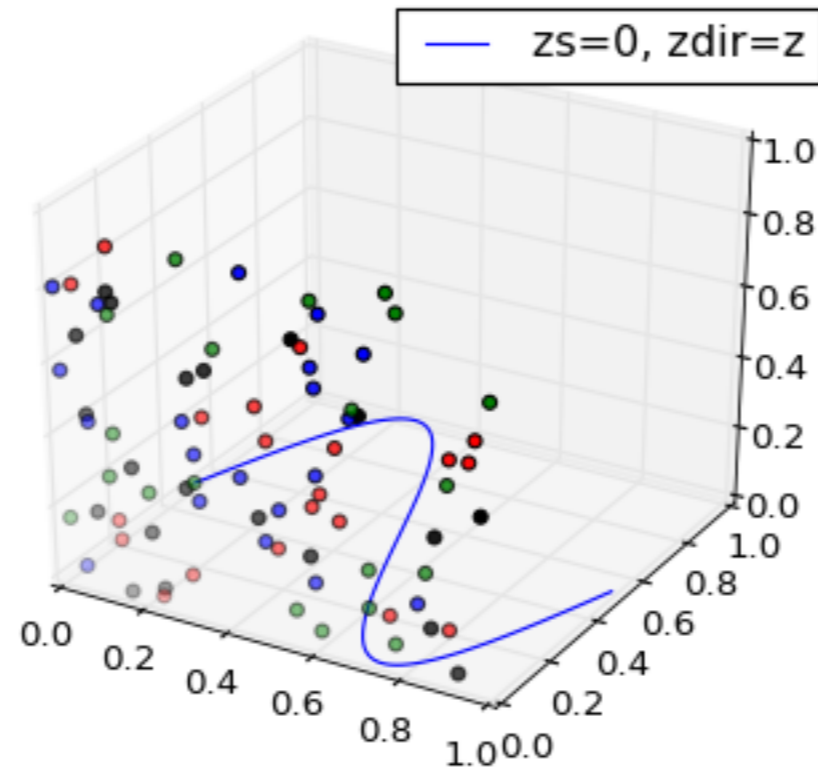
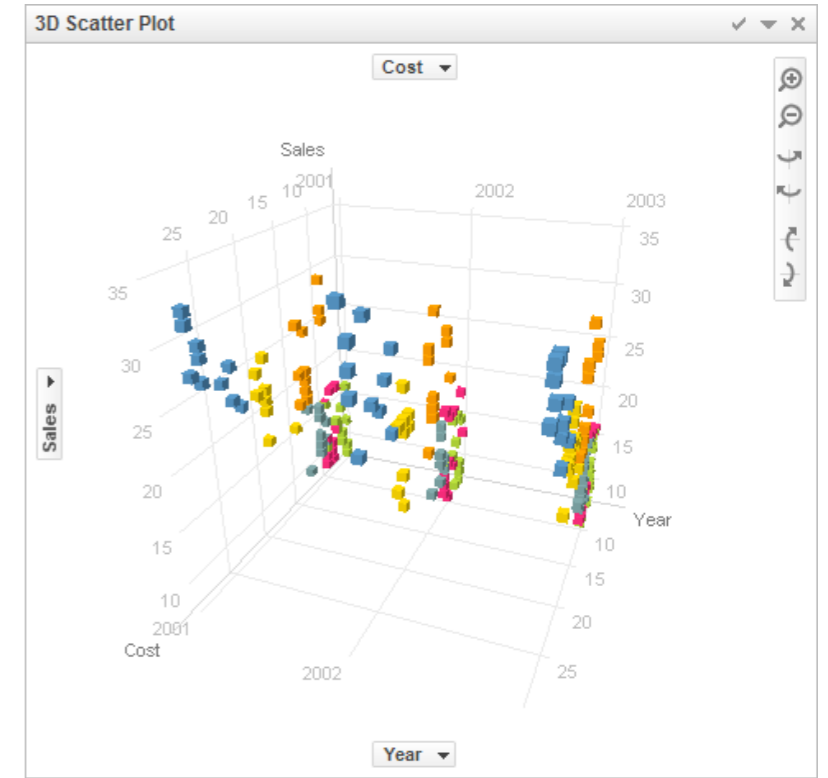
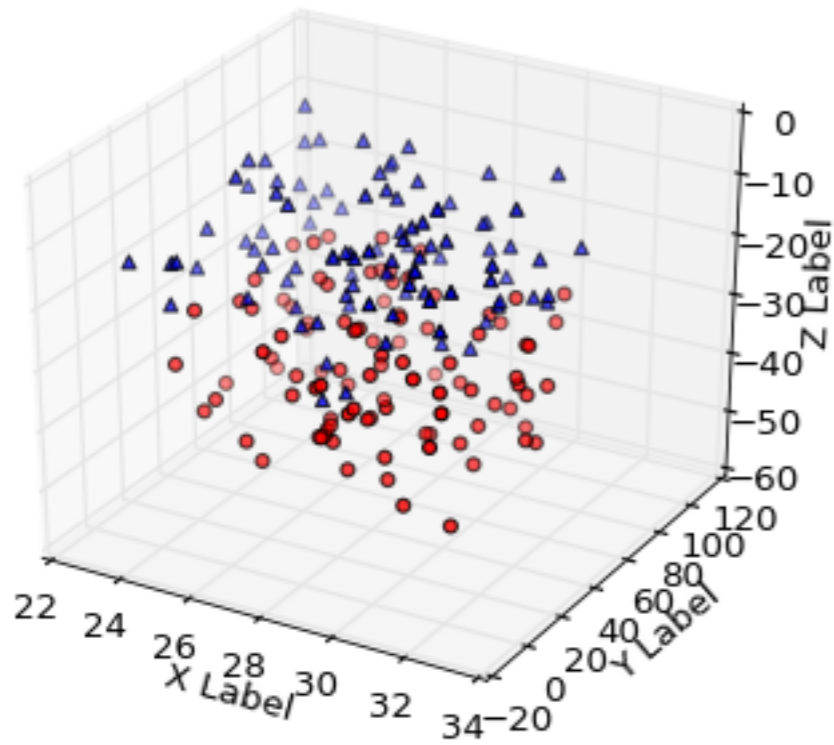




London Cholera Epidemic

From Edward Tufte, Visual and Statistical Thinking

Don't!



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

Feb 10, 2012 : ■ AAPL 493.42



■ Volume 22,523,900

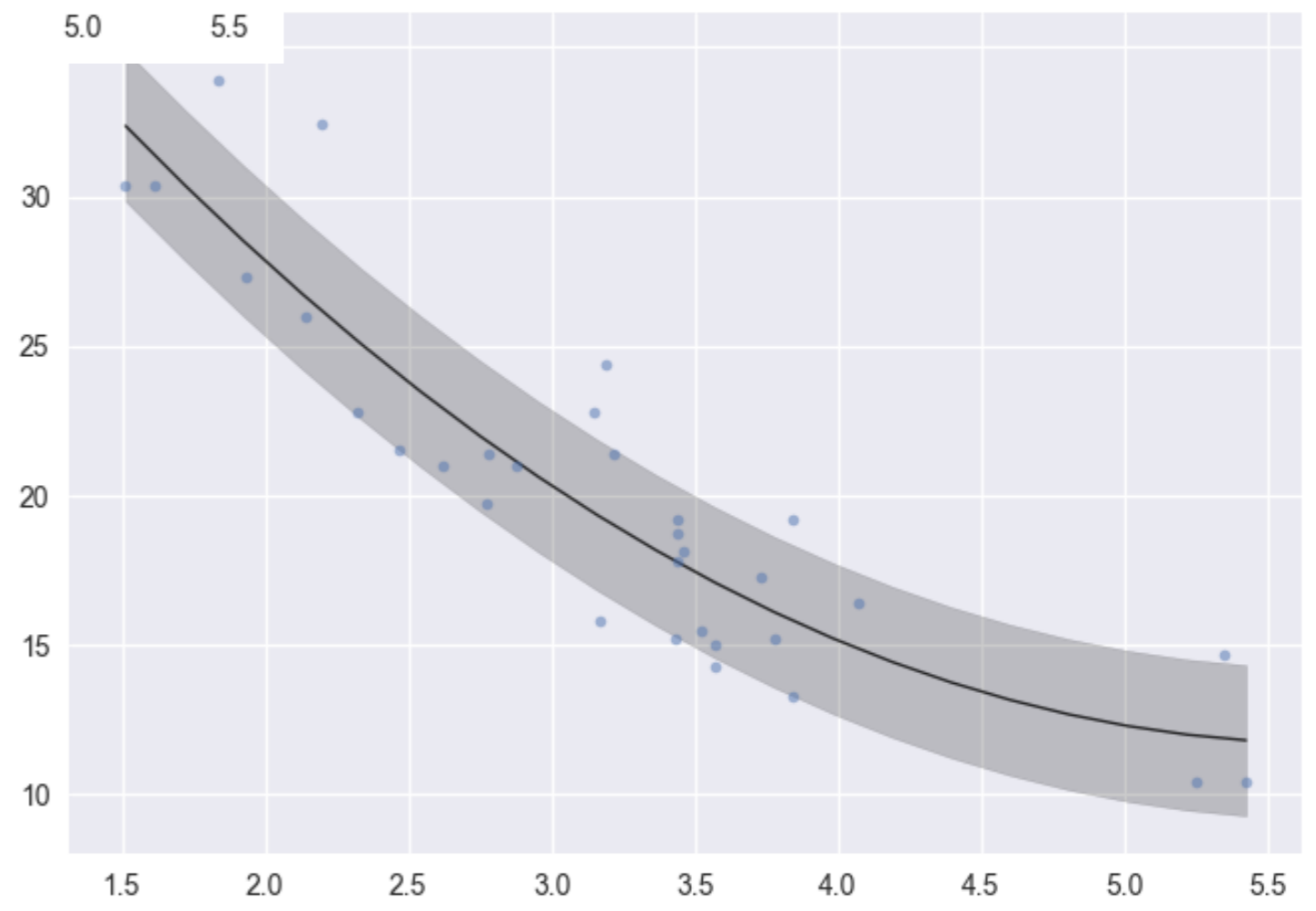
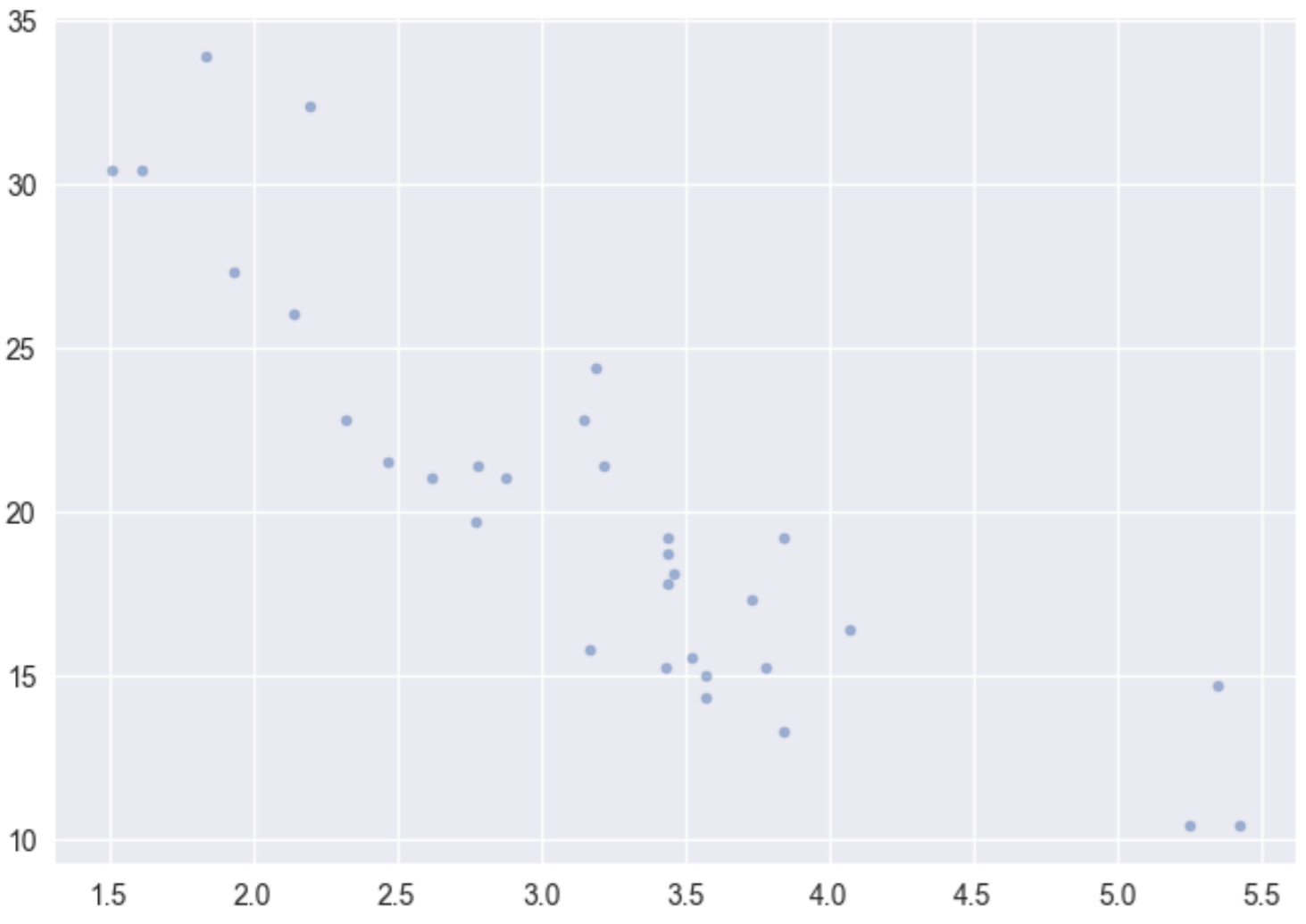


1D 5D 1M YTD 3M 6M 1Y 2Y 5Y Max

FROM: Mar 18 2011 TO: Mar 16 2012

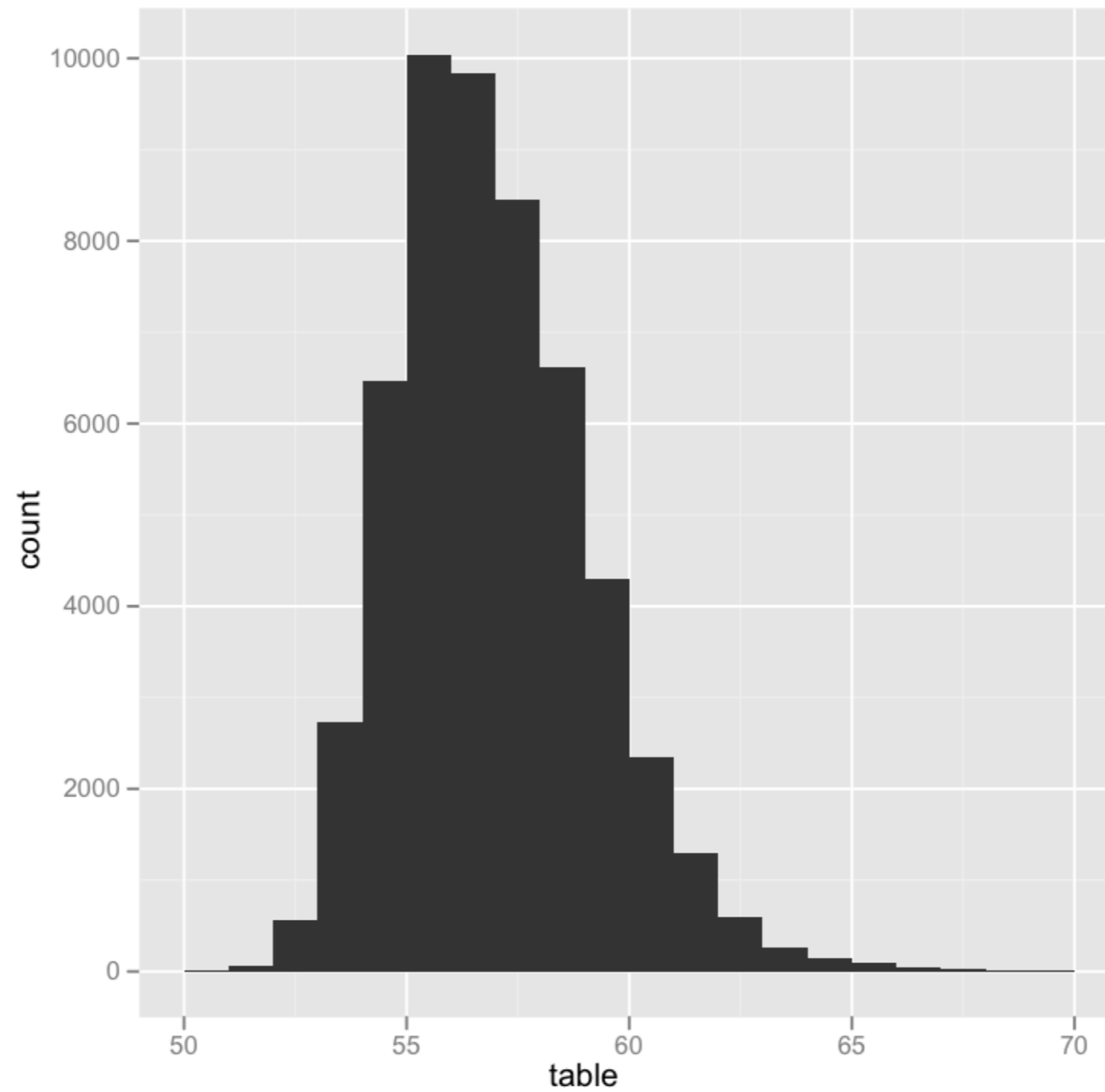


Basic Chart | Full Screen | Print | Share | Send Feedback

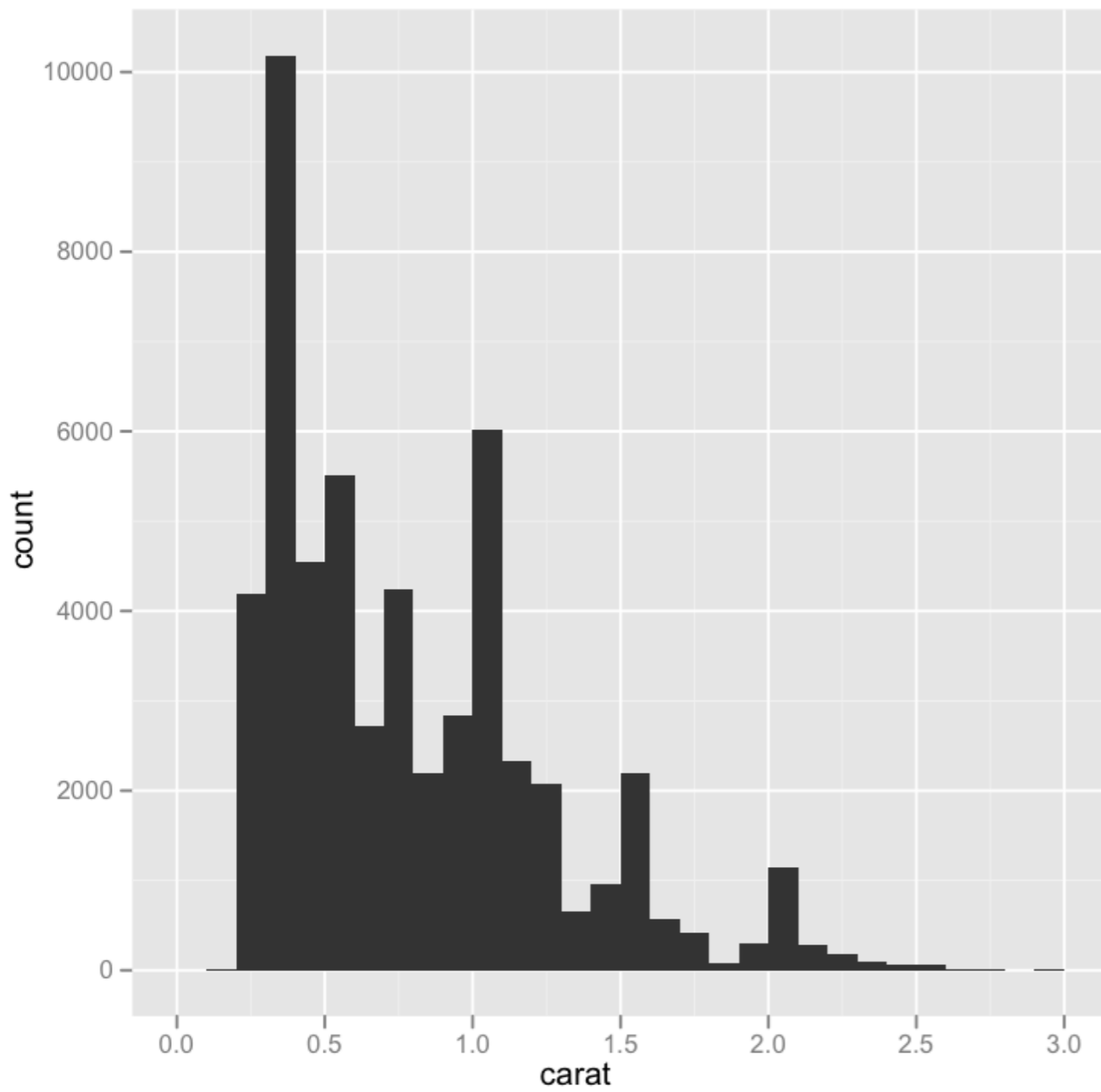


Distributions

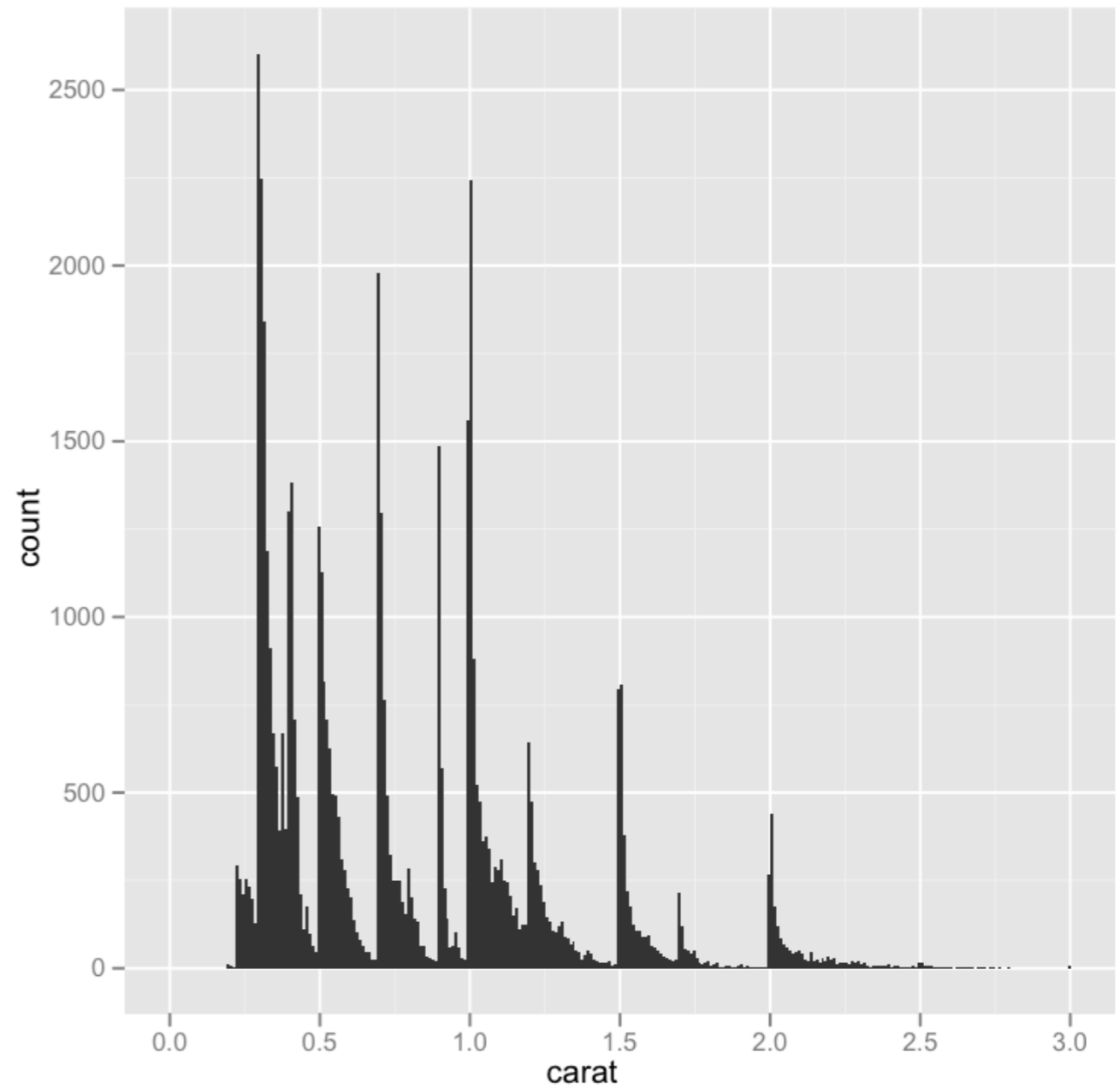
Histogram



Bin Width

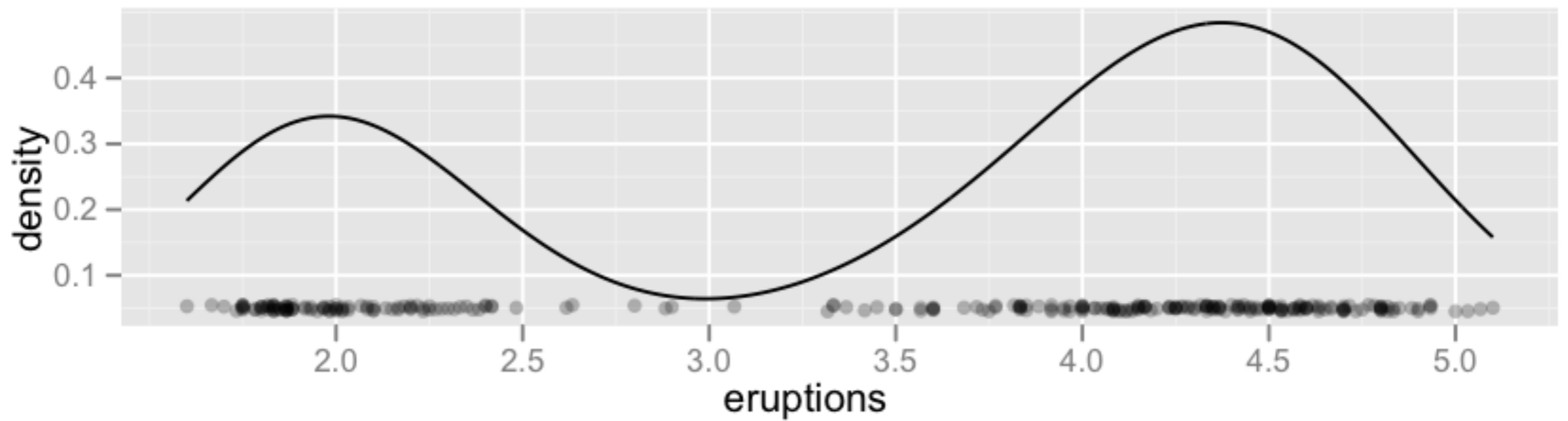


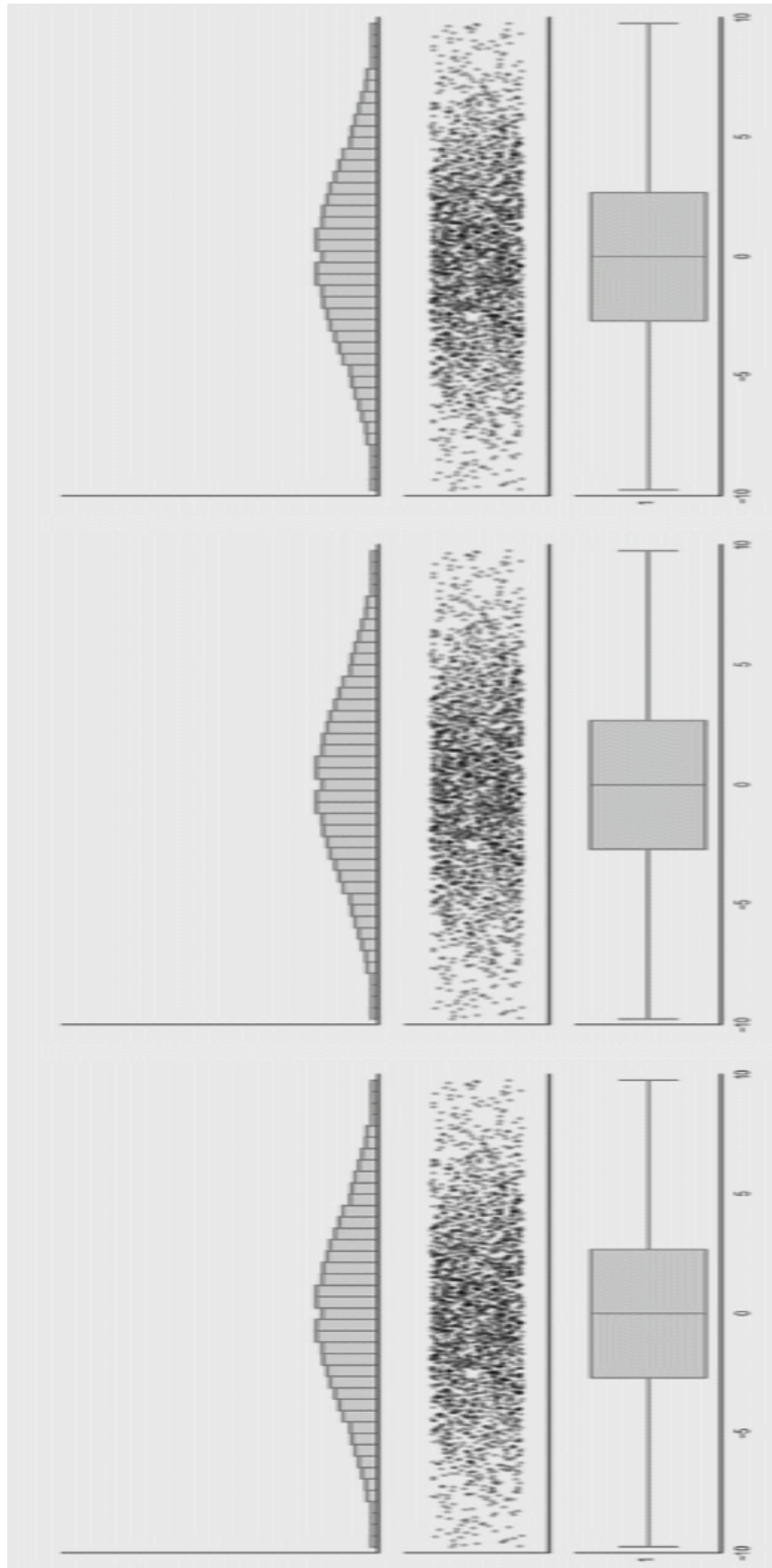
binwidth = 0.1



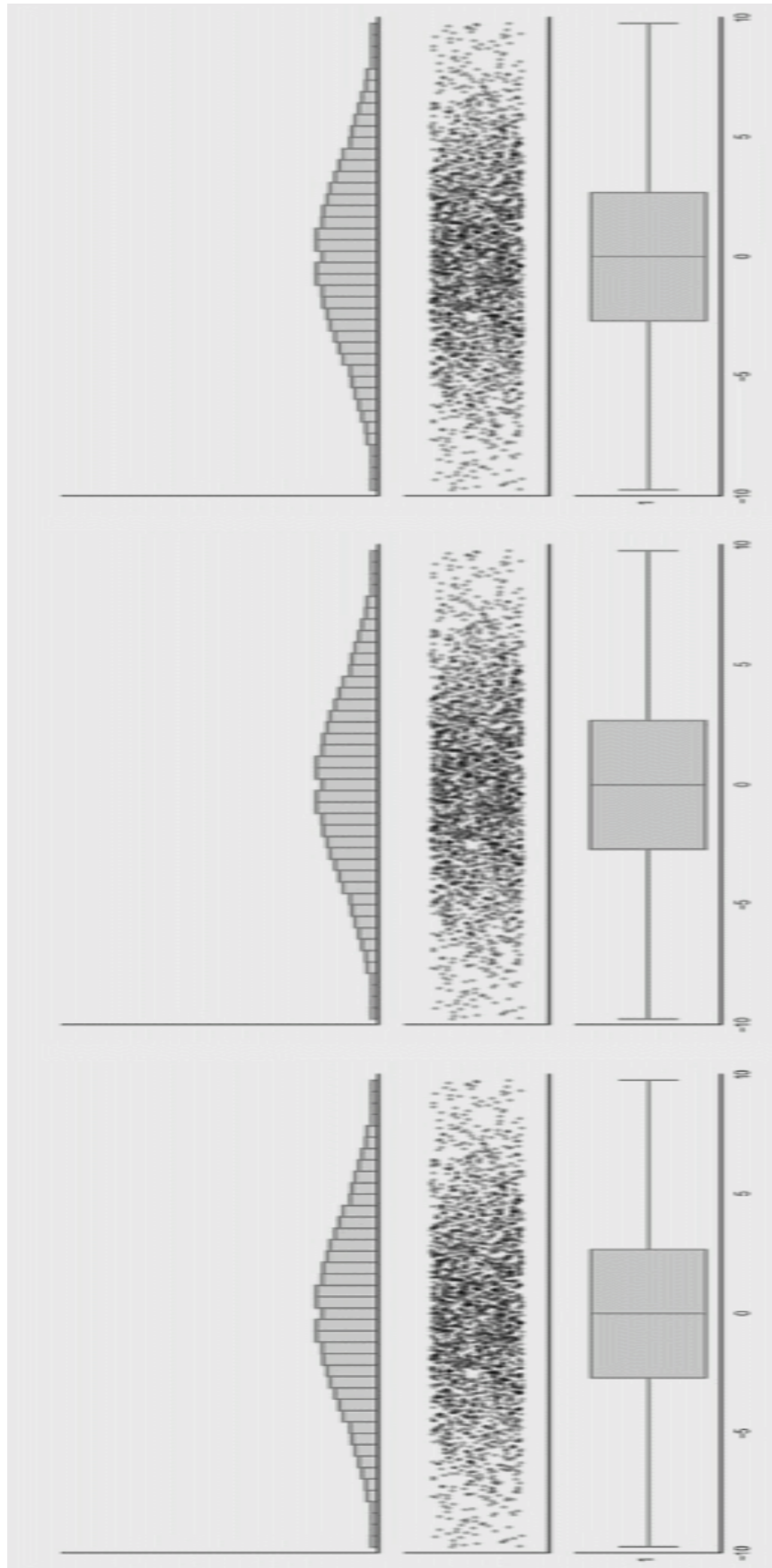
binwidth = 0.01

Density Plots



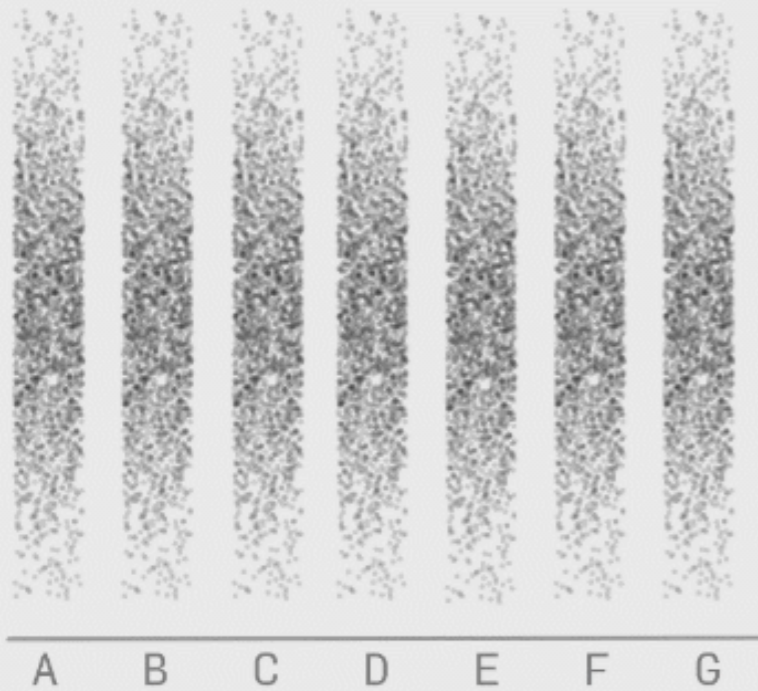


[https://
www.autodeskresearch.com/
publications/samestats](https://www.autodeskresearch.com/publications/samestats)

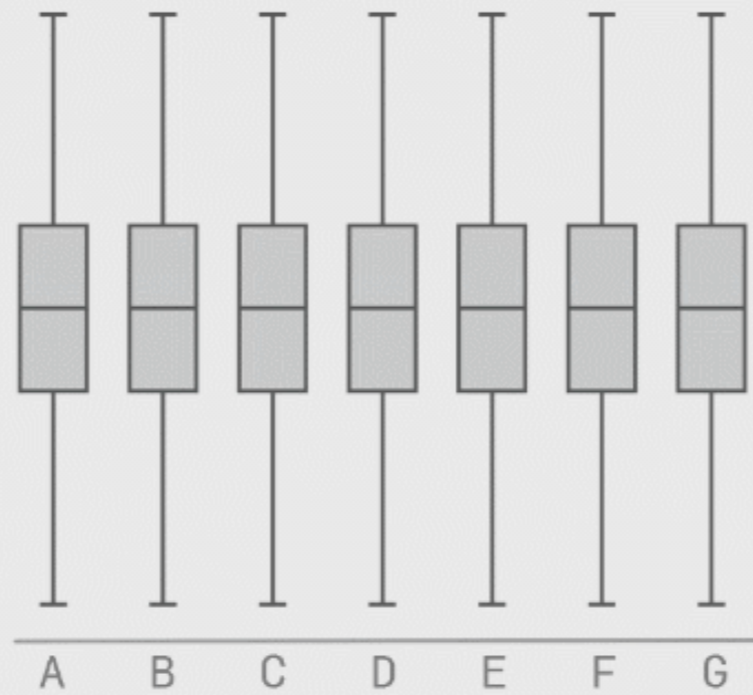


[https://
www.autodeskresearch.com/
publications/samestats](https://www.autodeskresearch.com/publications/samestats)

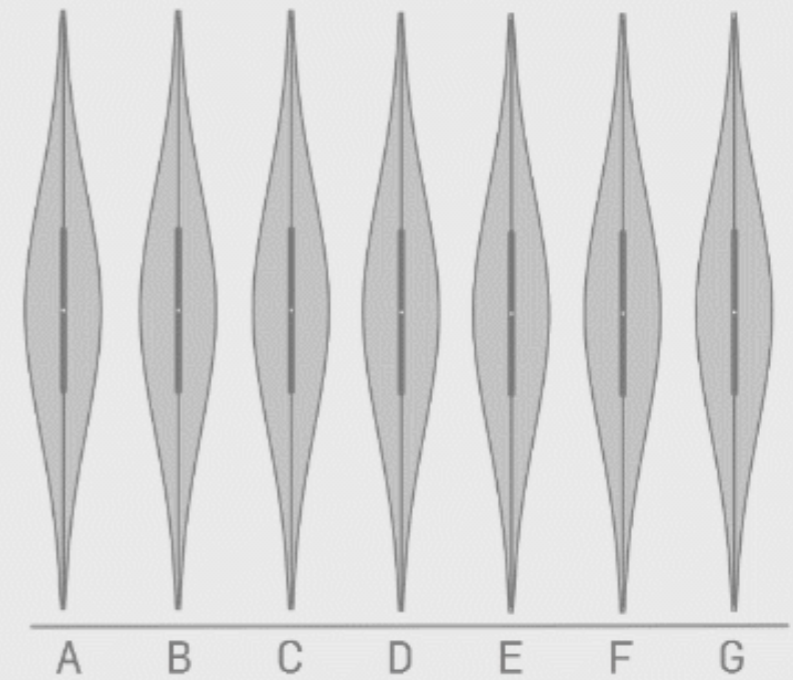
Raw Data



Box-plot of the Data

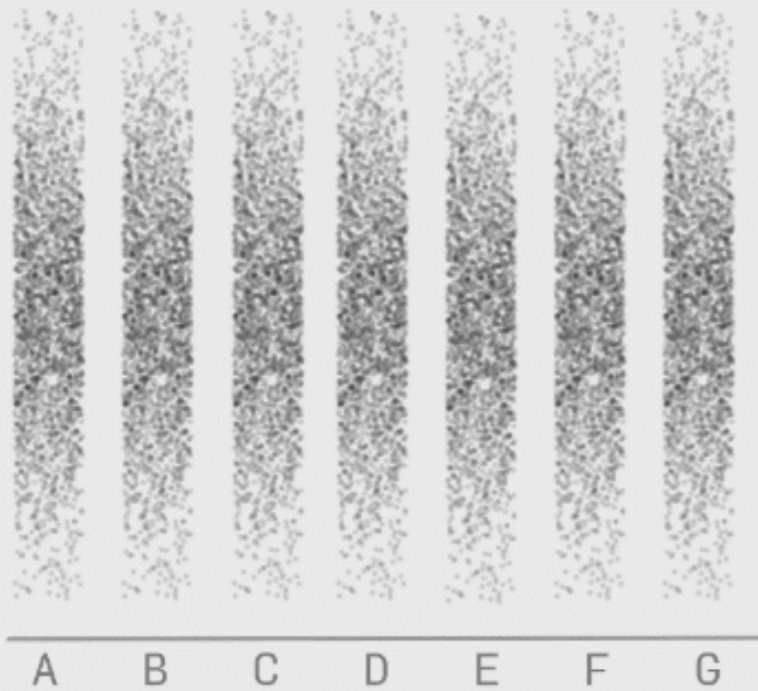


Violin-plot of the Data

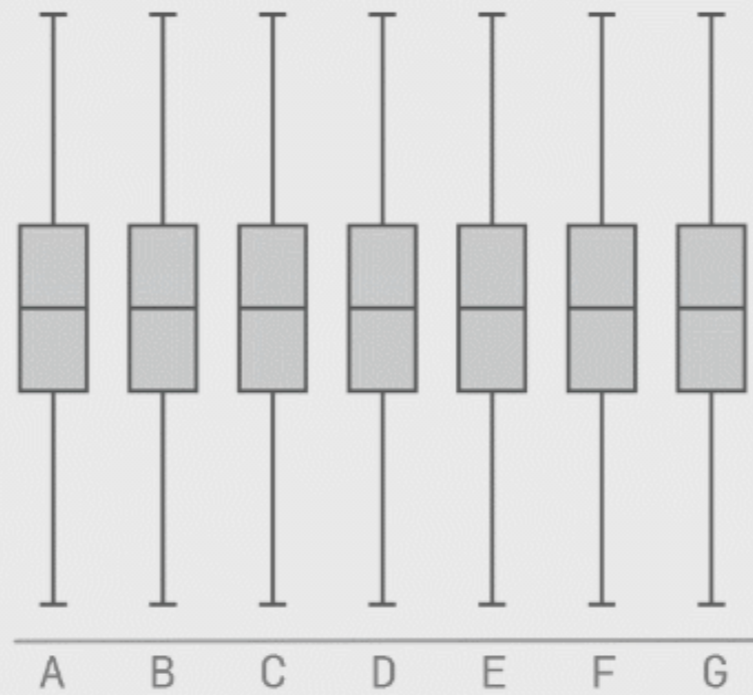


<https://www.autodeskresearch.com/publications/samestats>

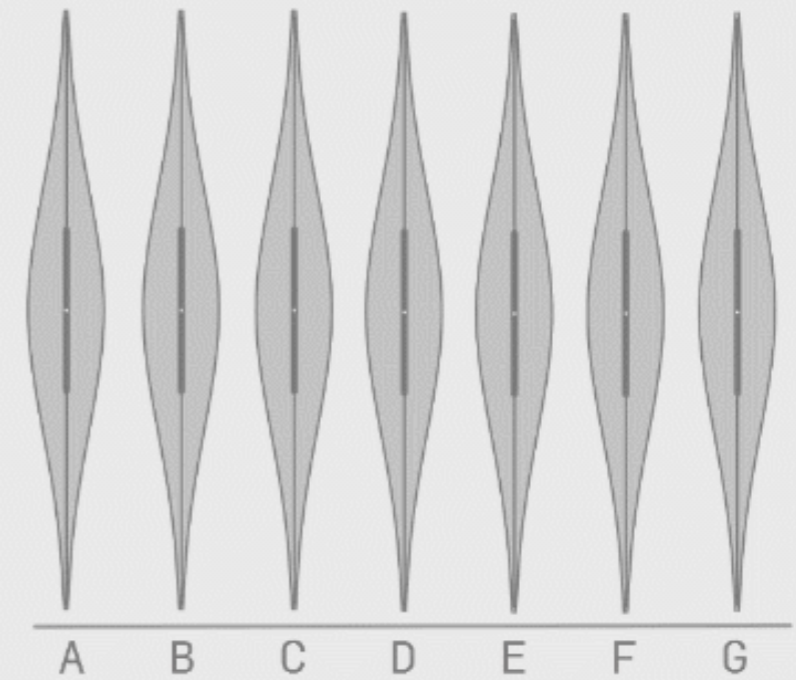
Raw Data



Box-plot of the Data

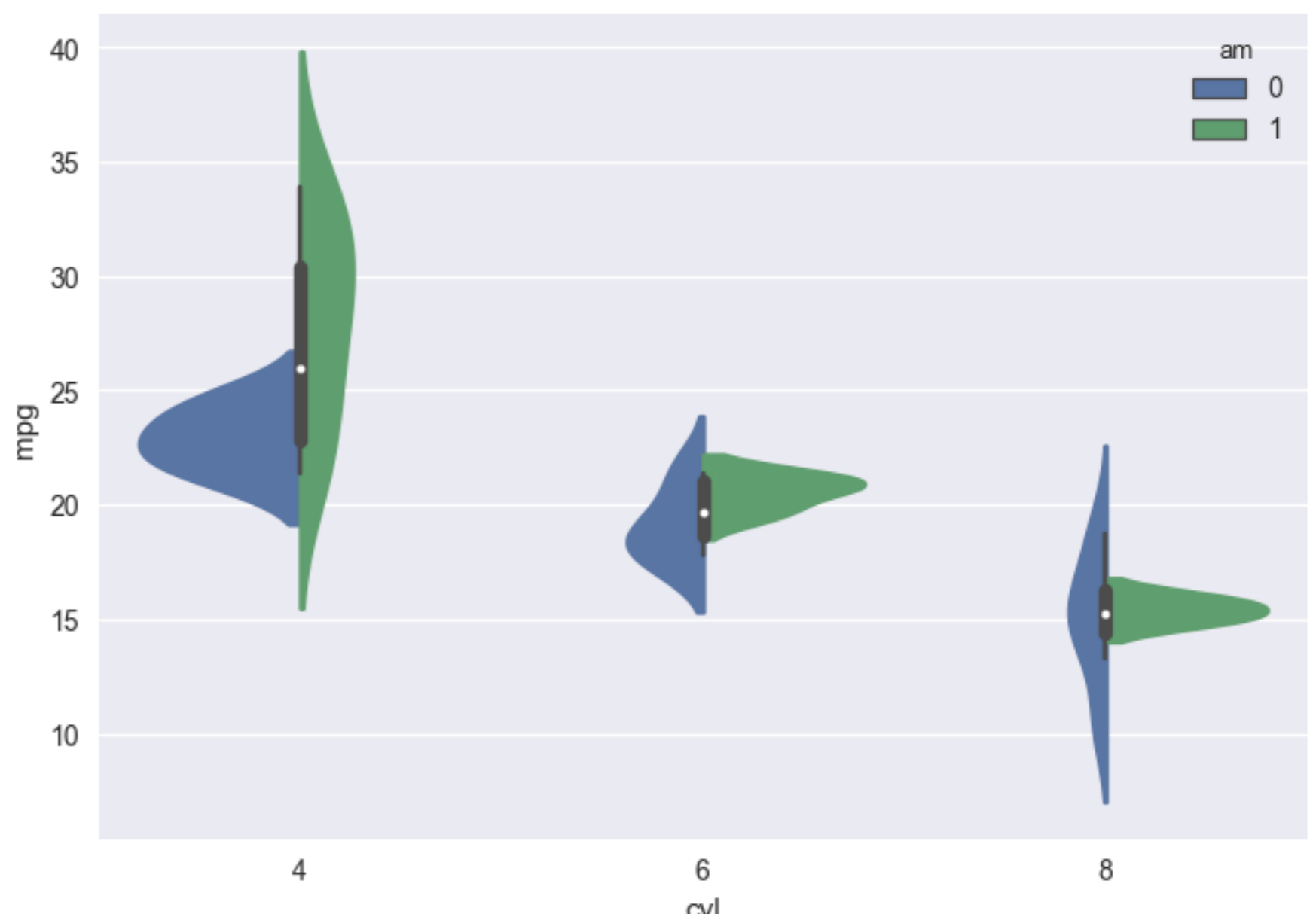
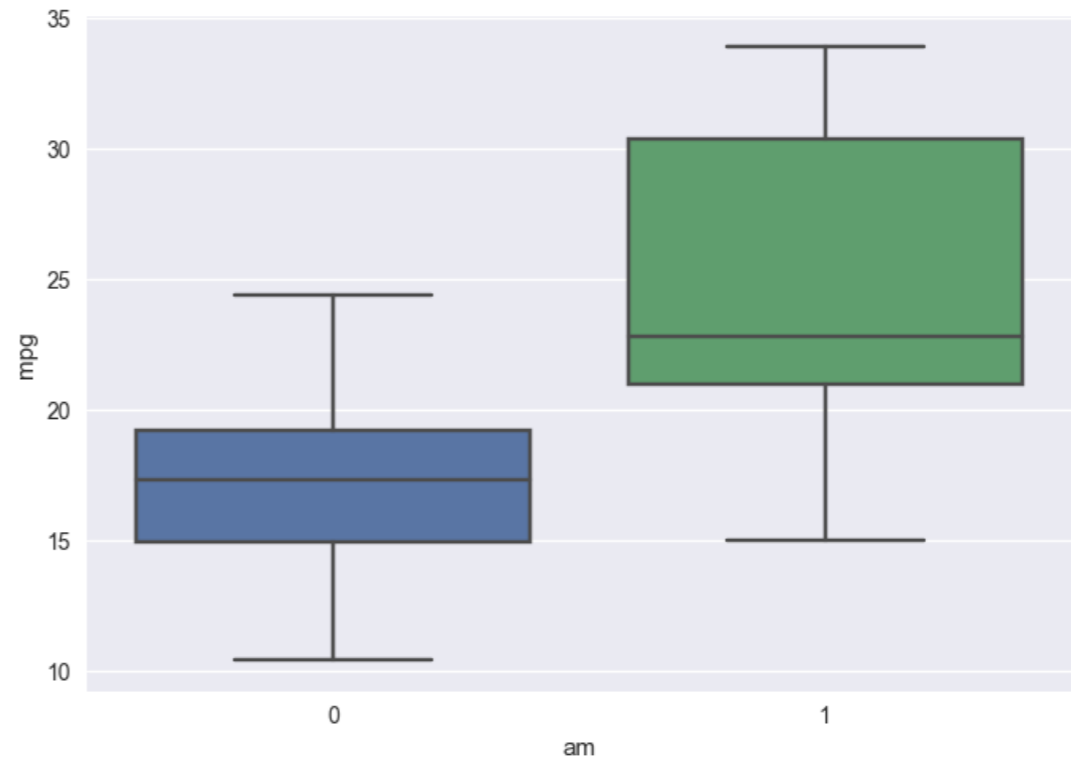
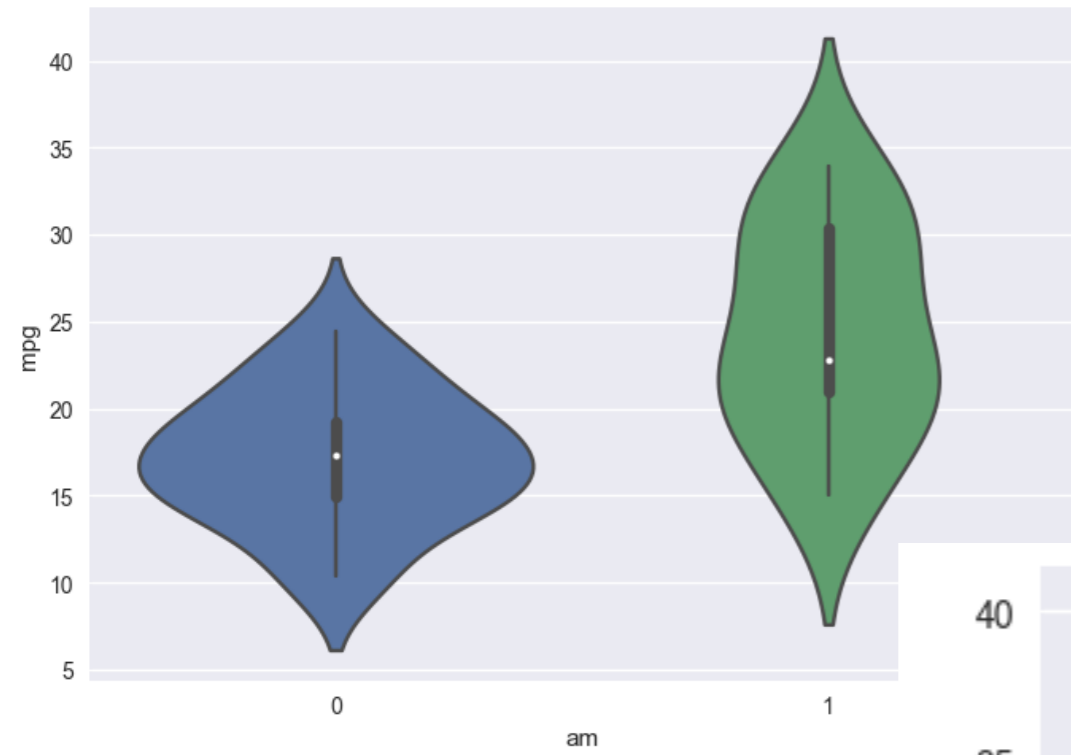


Violin-plot of the Data



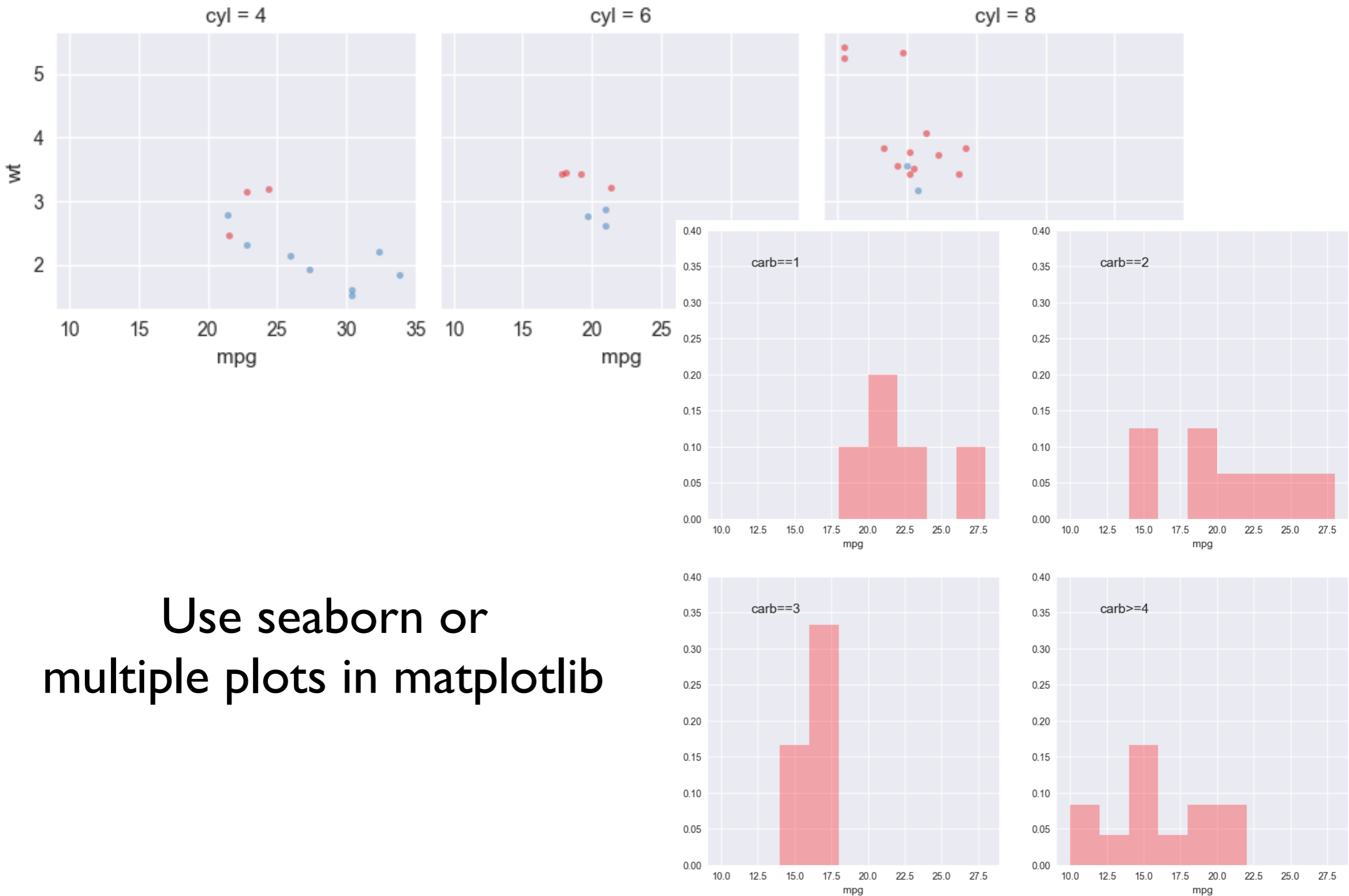
<https://www.autodeskresearch.com/publications/samestats>

GROUP



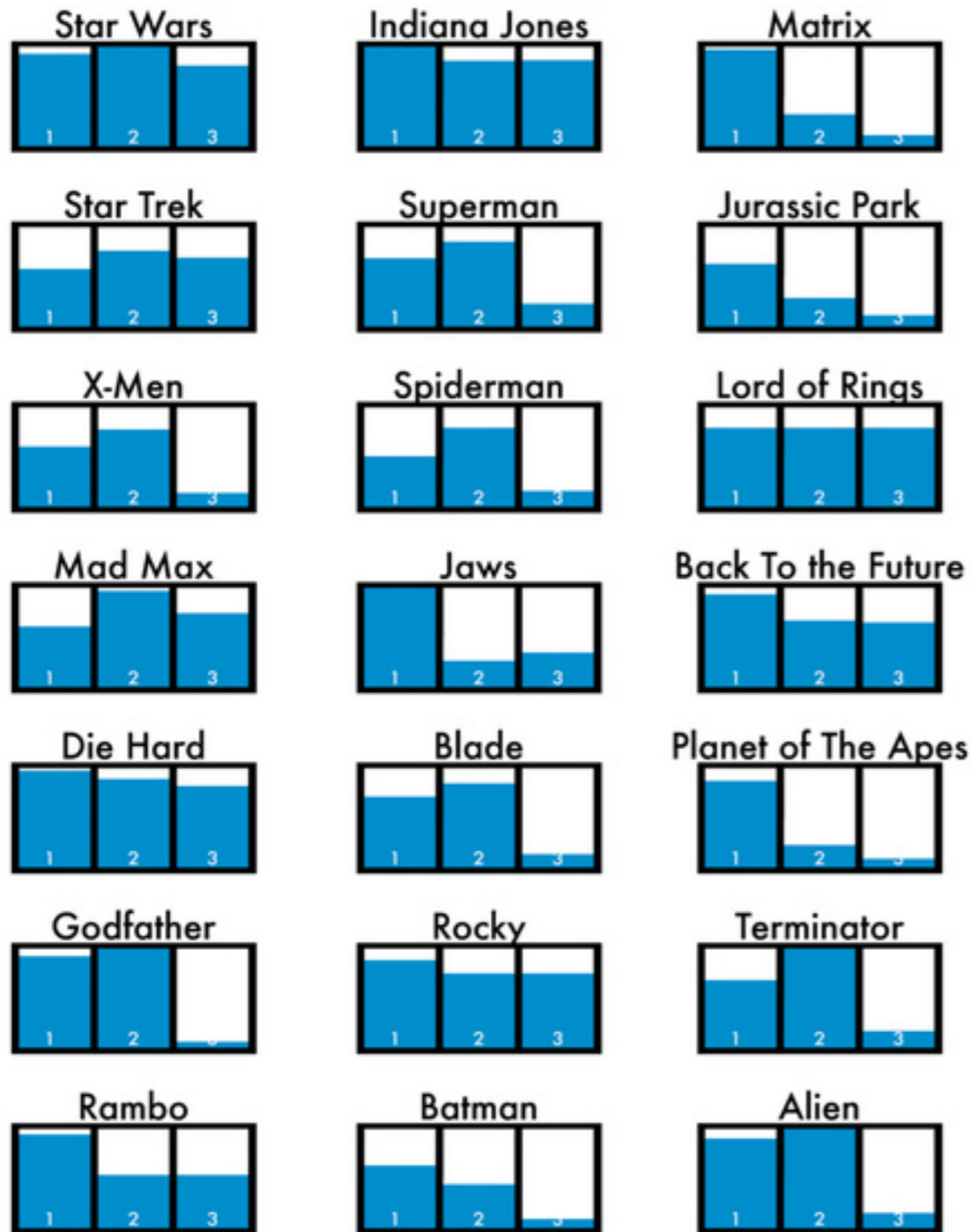
getting complex...

Faceting and Small Multiples



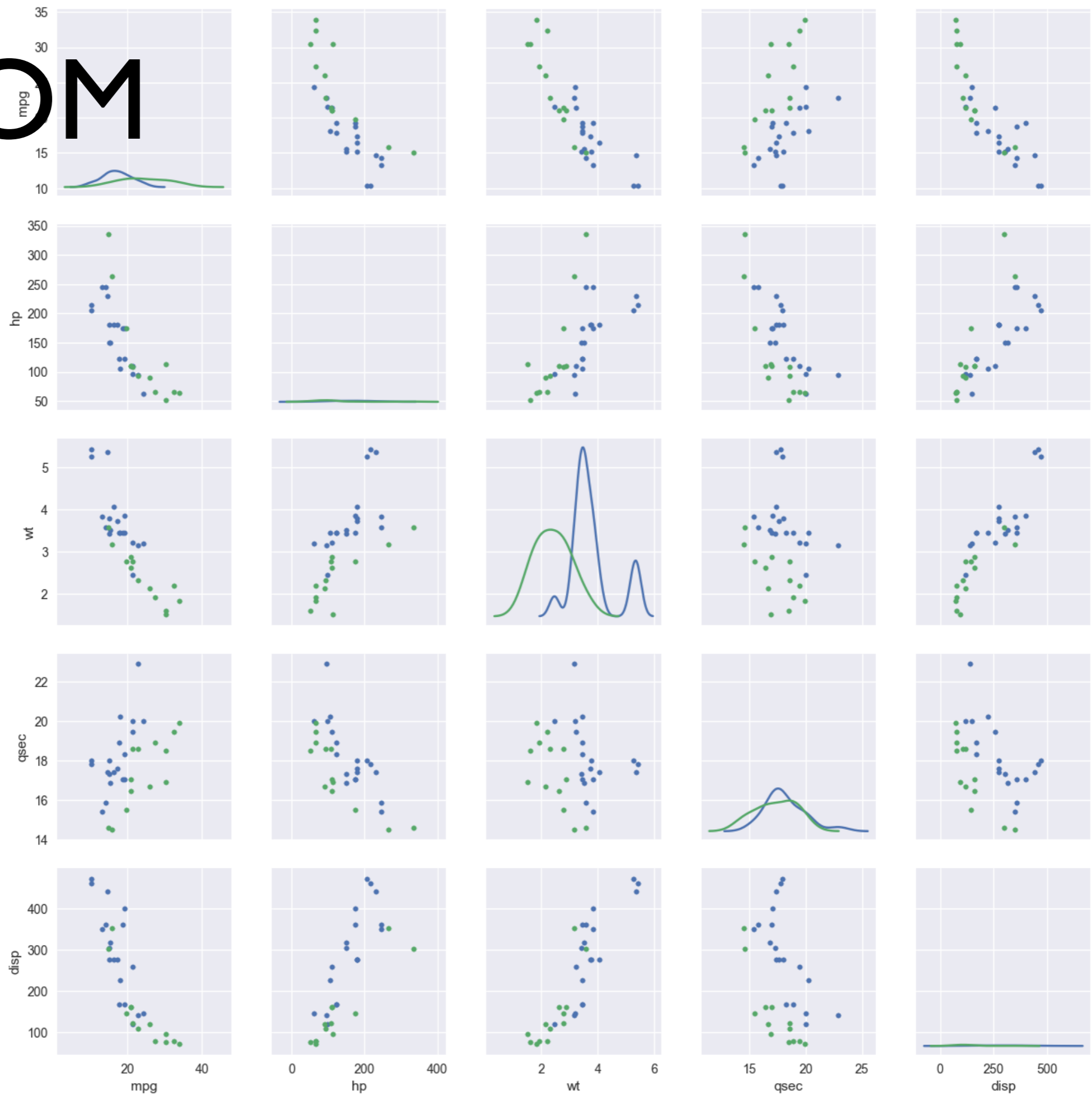
Use seaborn or
multiple plots in matplotlib

THE TRILOGY METER



Small multiples

SPLOM



Design Exercise

Hands-On Exercise

How do you feel about doing science?

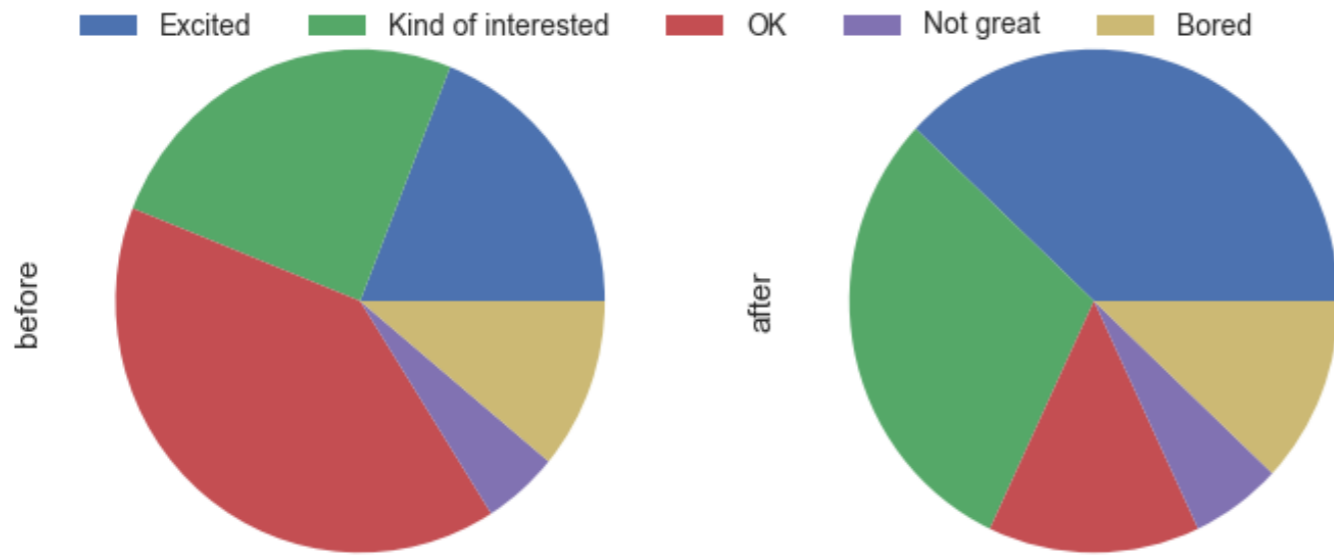
Table

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

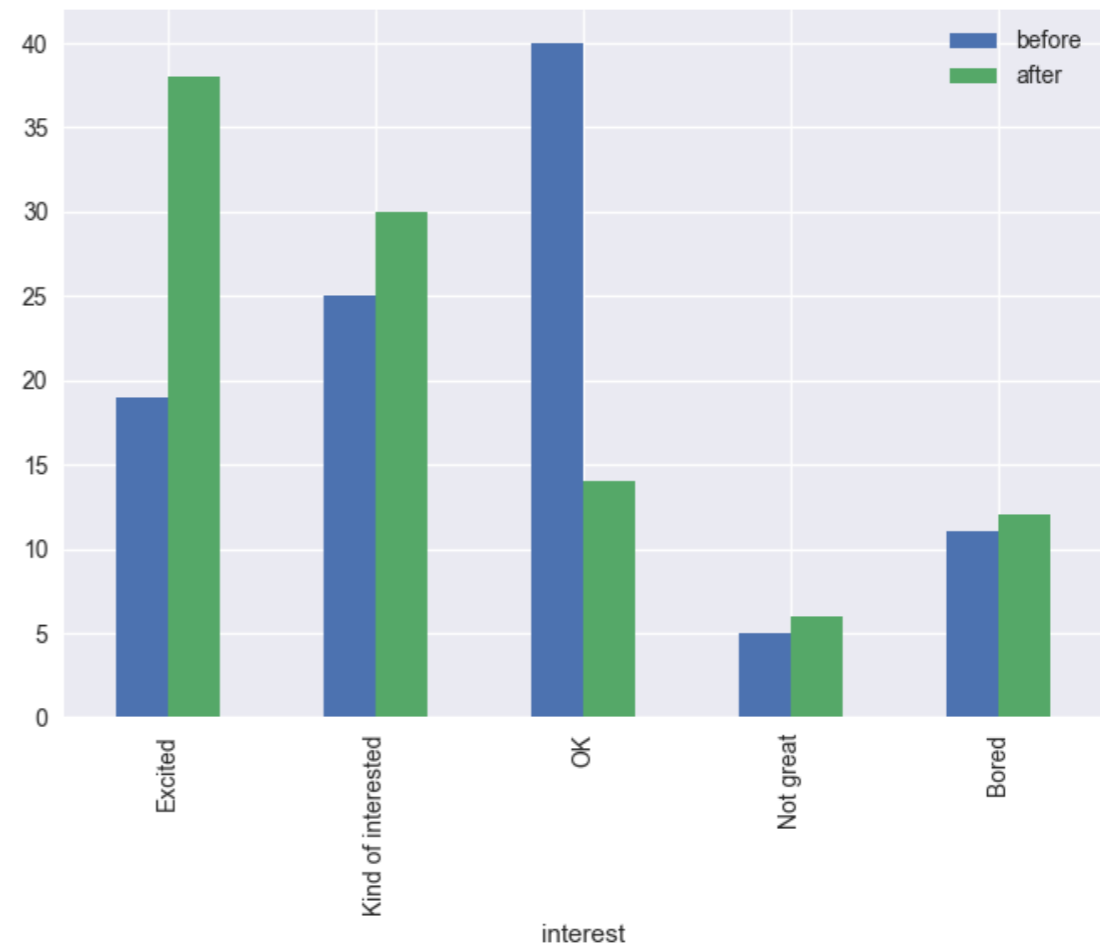
Come up with multiple visualizations.

Pen and Paper Only.

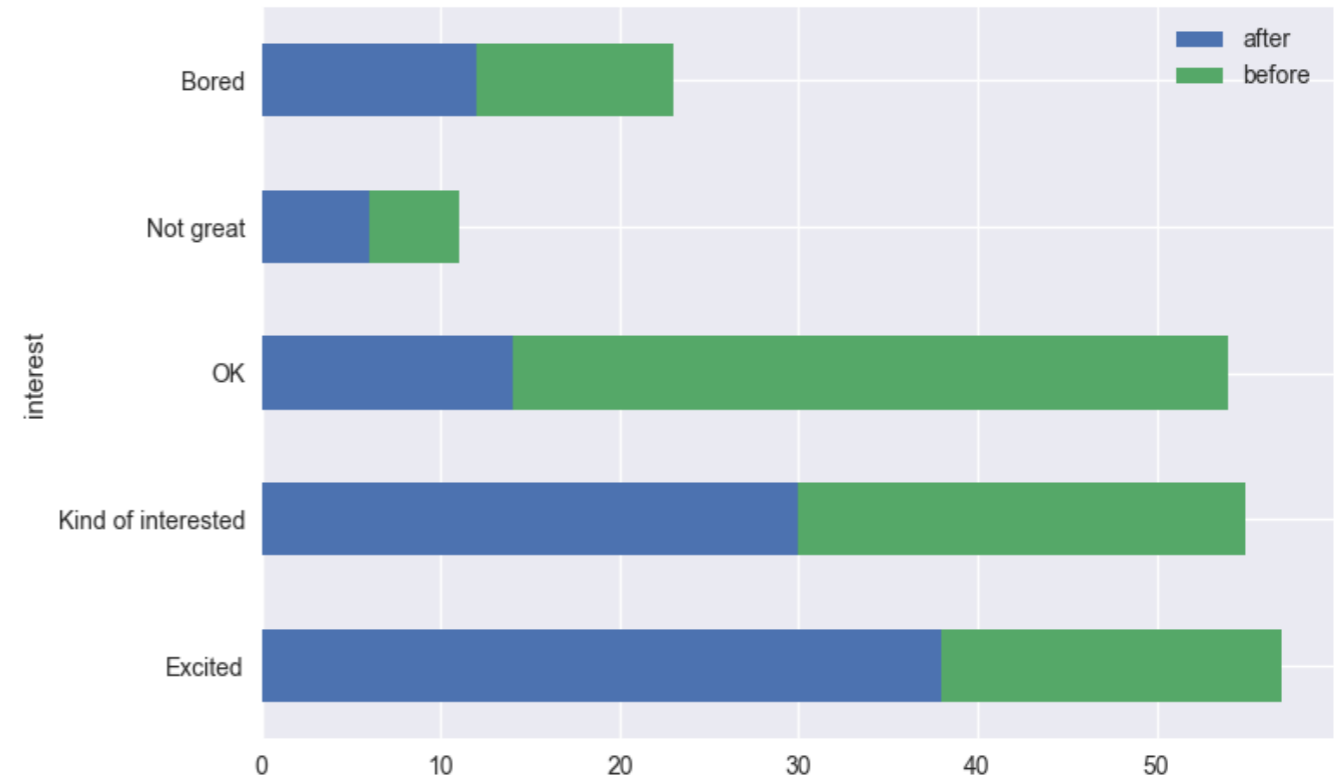
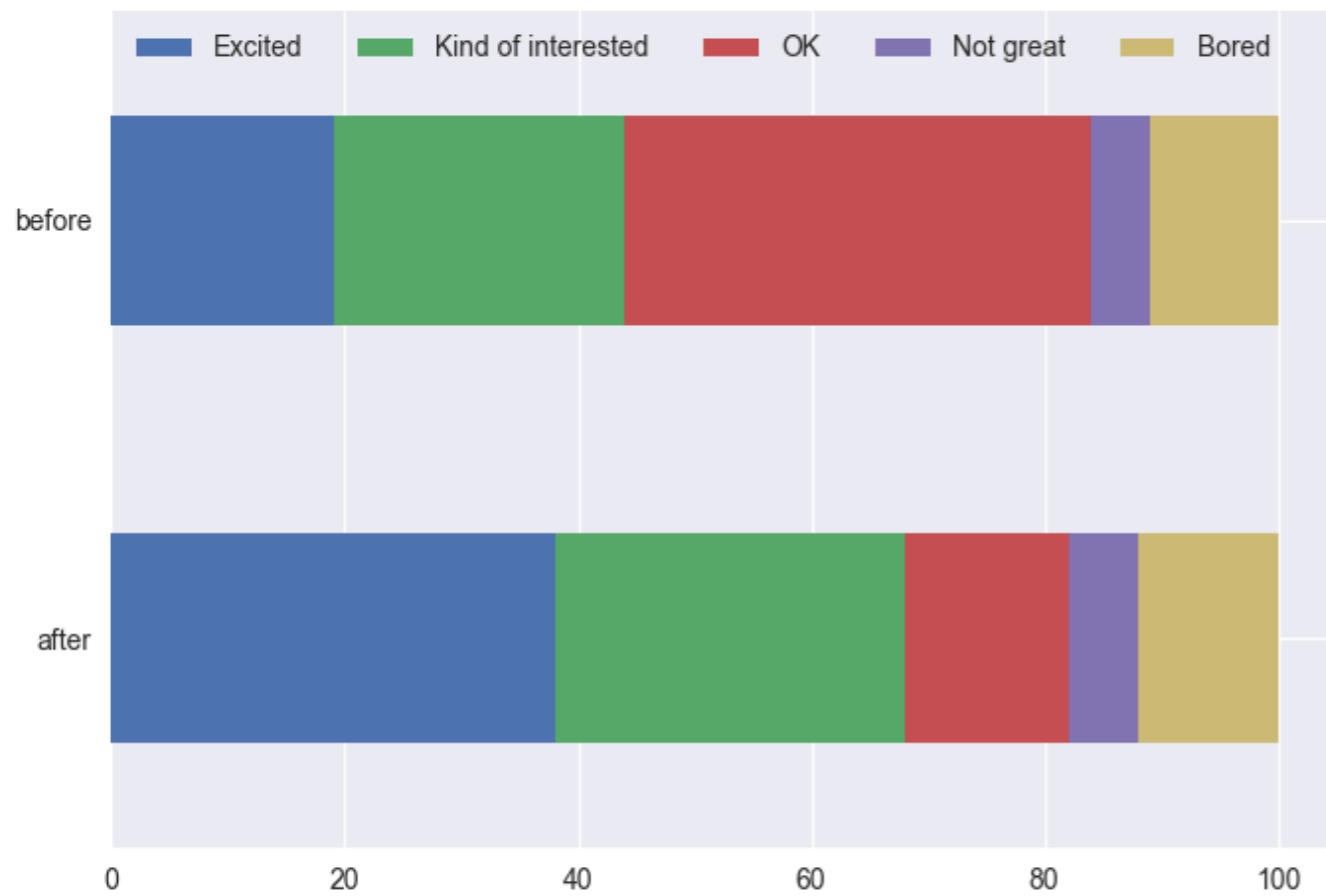


Pie

Side by side bar

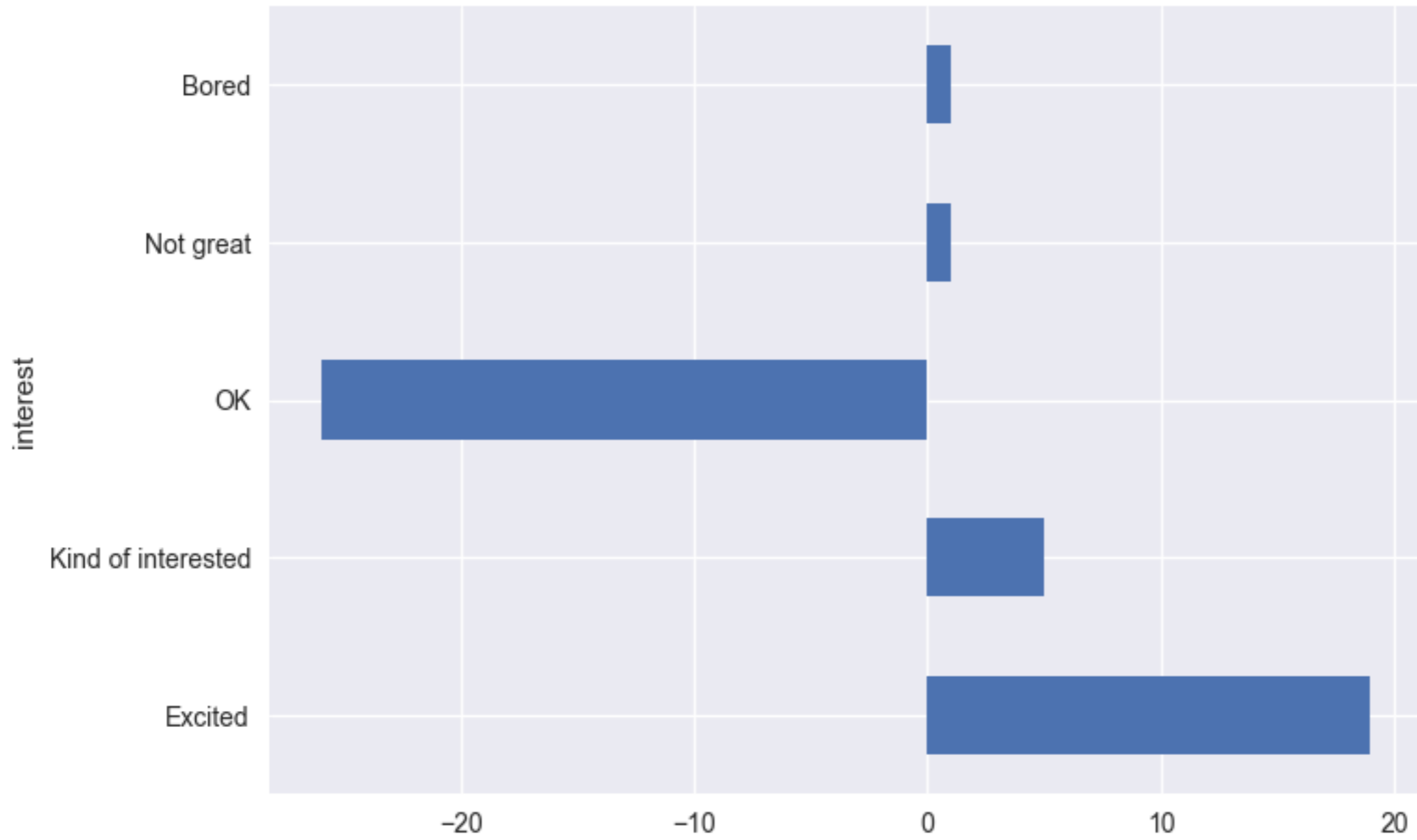


Stacked bar, not very useful

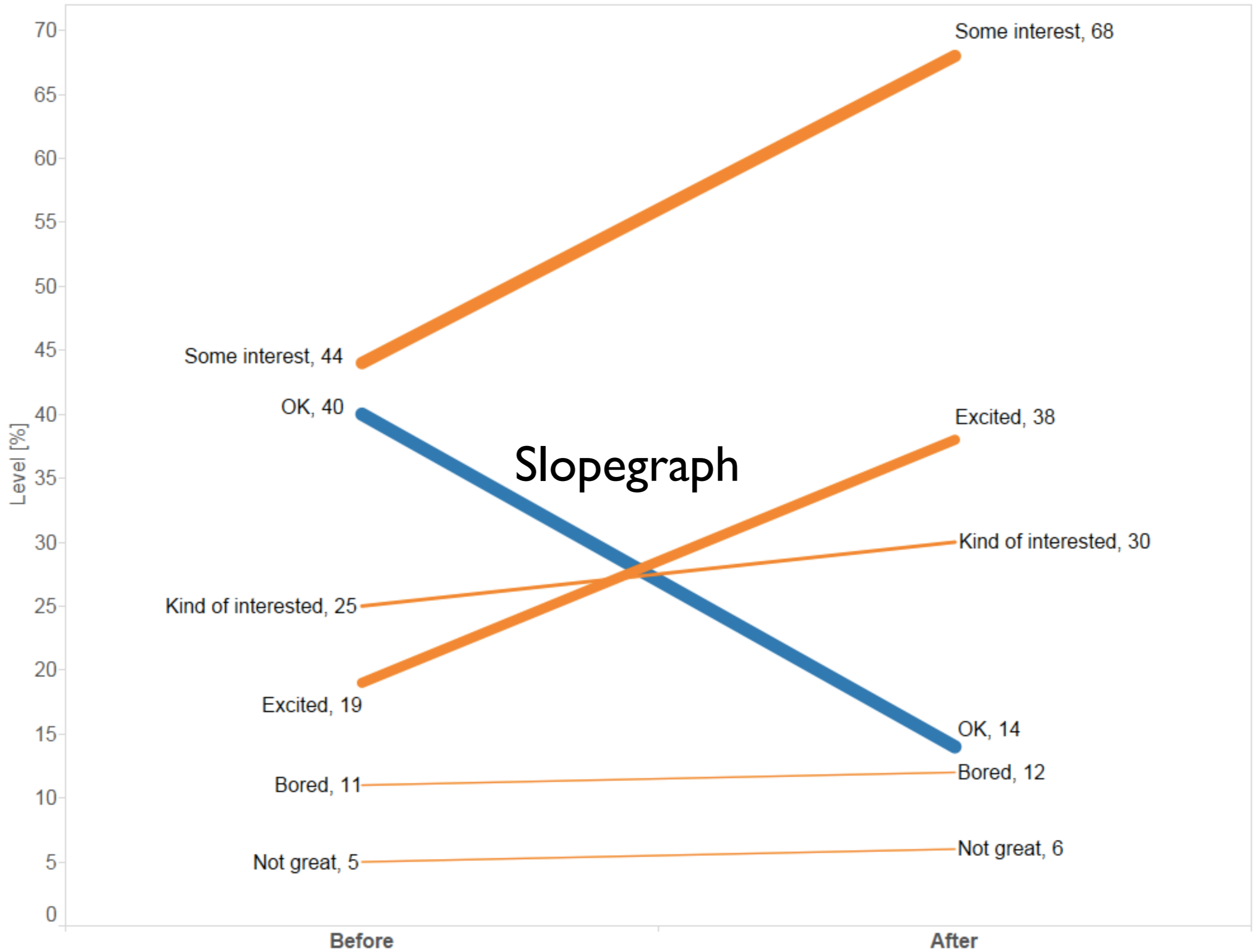


Data Transposed Bar Chart

Difference Bar Chart



How do you feel about doing science?

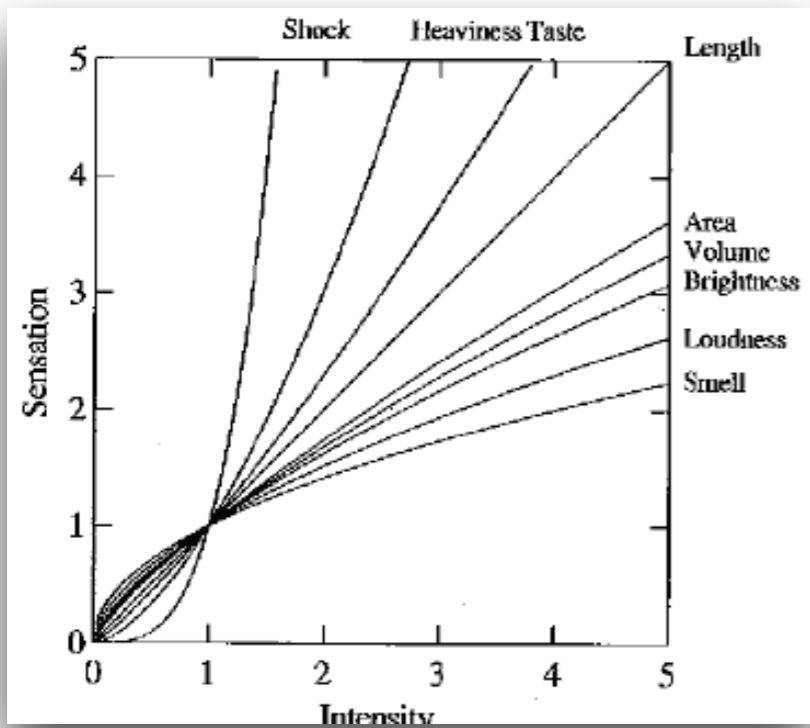


After the pilot program,

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

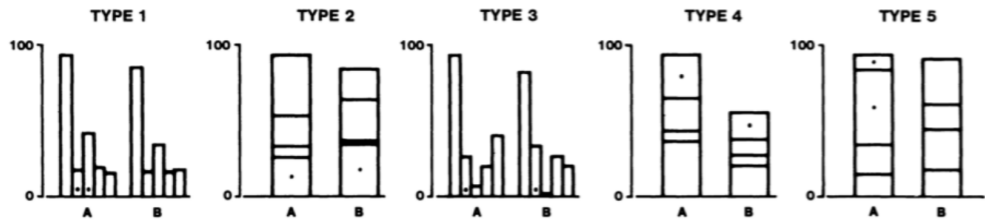


Figure 4. Graphs from position-length experiment.

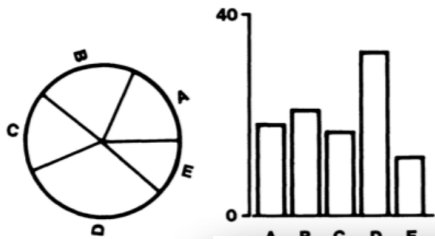
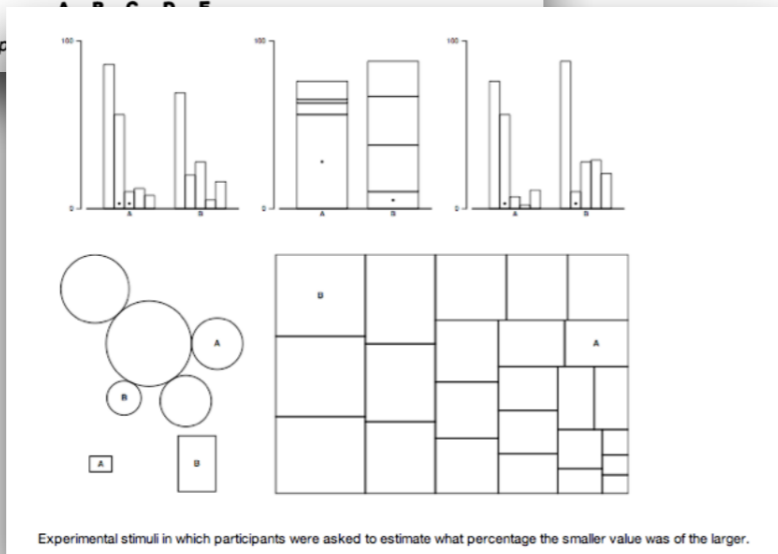


Figure 3. Graphs from p...

Cleveland / McGill, 1984

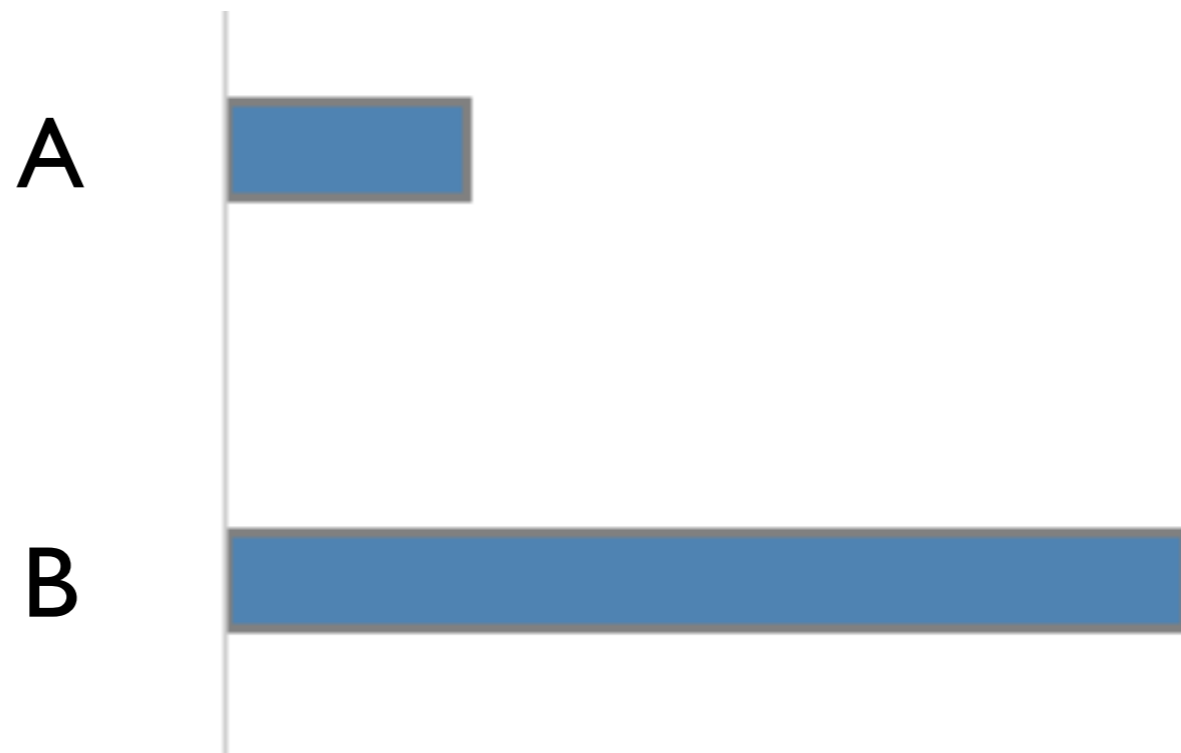


Heer / Bostock, 2010

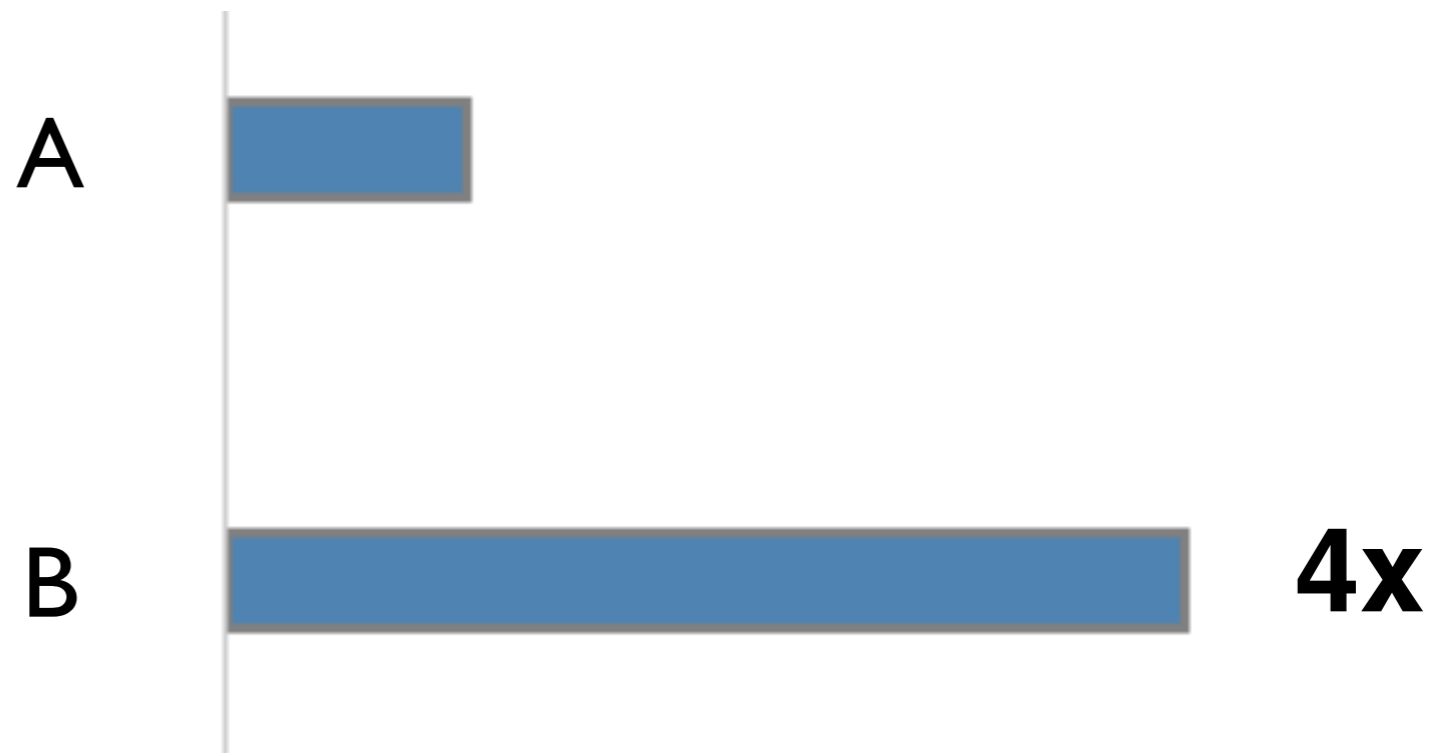
Quantitative	Ordinal	Nominal
Position	Position	Position
Length	Density	Hue
Angle	Saturation	Texture
Slope	Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Saturation
Saturation	Length	Shape
Hue	Angle	Length
Texture	Slope	Angle
Connection	Area	Slope
Containment	Volume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

J. Mackinlay, 1986

How much longer?



How much longer?



How much steeper slope?



A



B

How much steeper slope?

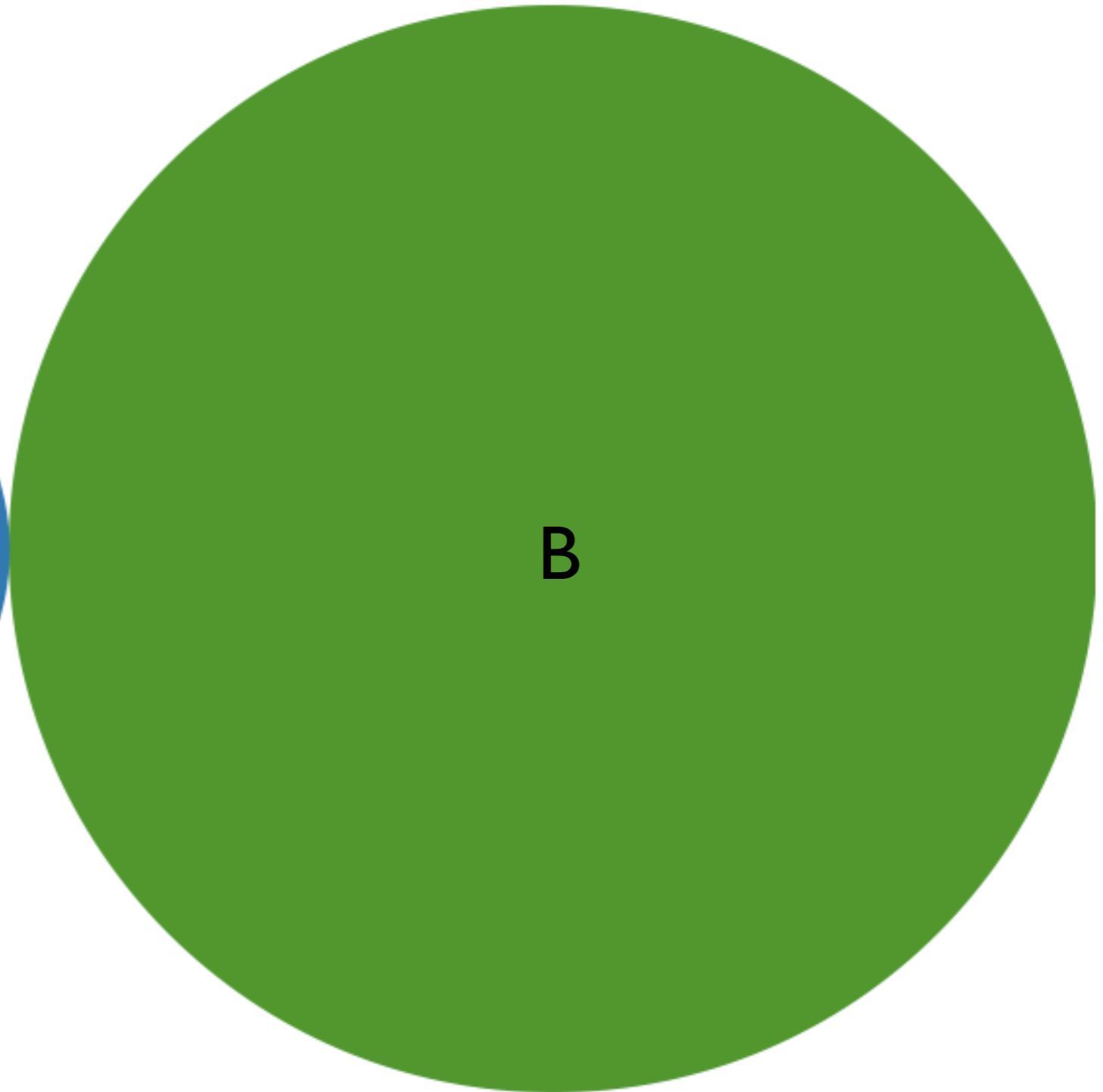
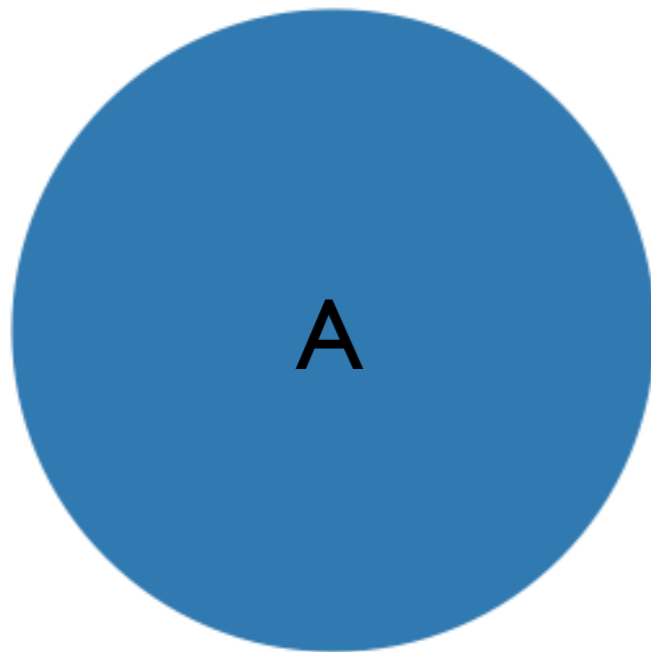


A
4x

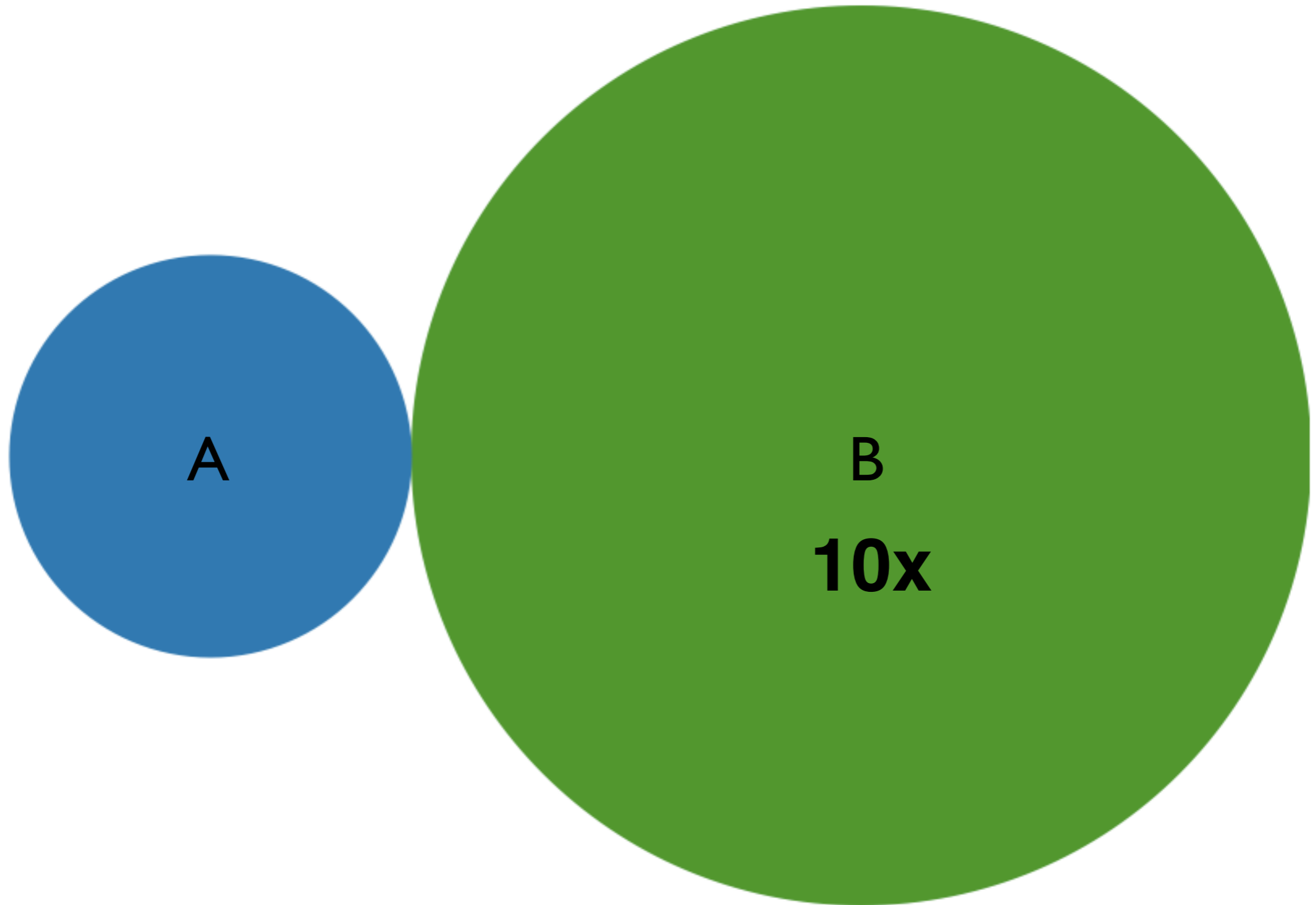


B

How much larger area?



How much larger area?



How much darker?



A



B

How much darker?



A



B

2x

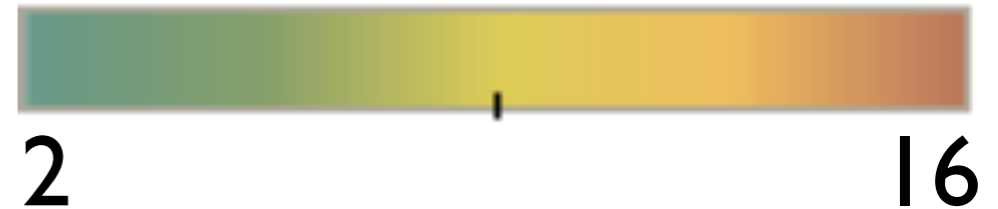
How much bigger value?



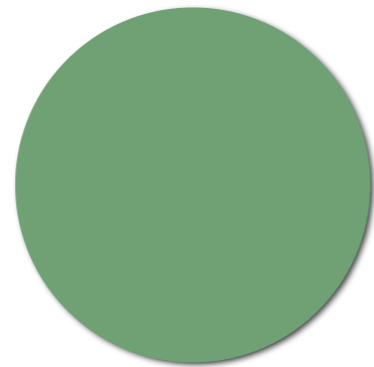
A



B



How much bigger value?



A



B

4x



2

16

Most
Efficient

Position



Length



Slope



Angle



Area



Intensity



Least
Efficient

Color



Shape



Most Efficient



Least Efficient

Position



Length



Slope



Angle



Area



Intensity



Color



Shape

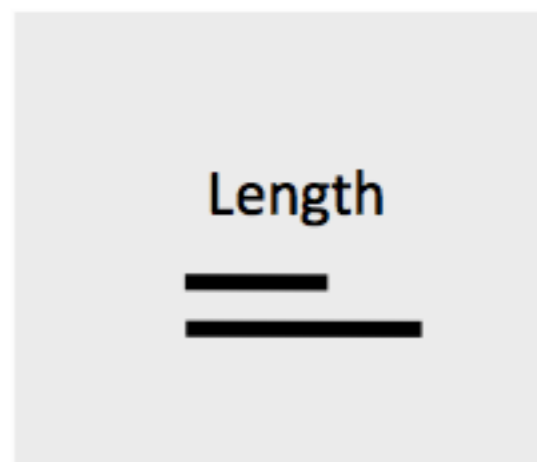
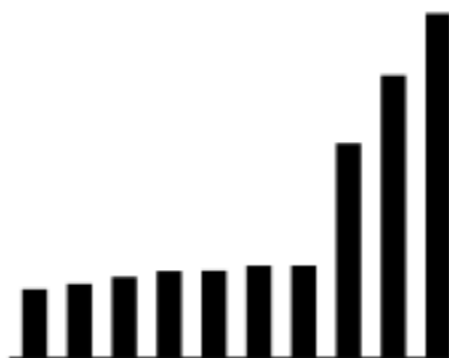
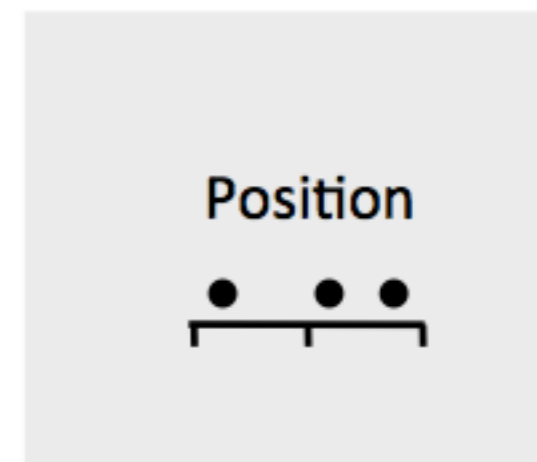
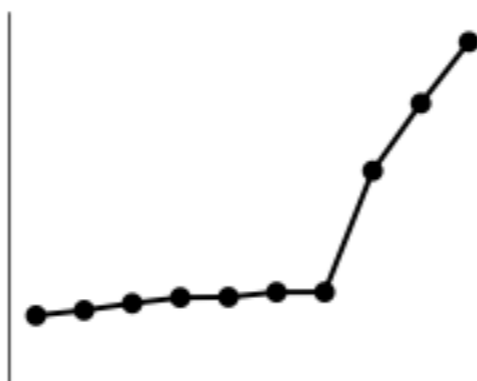


Quantitative

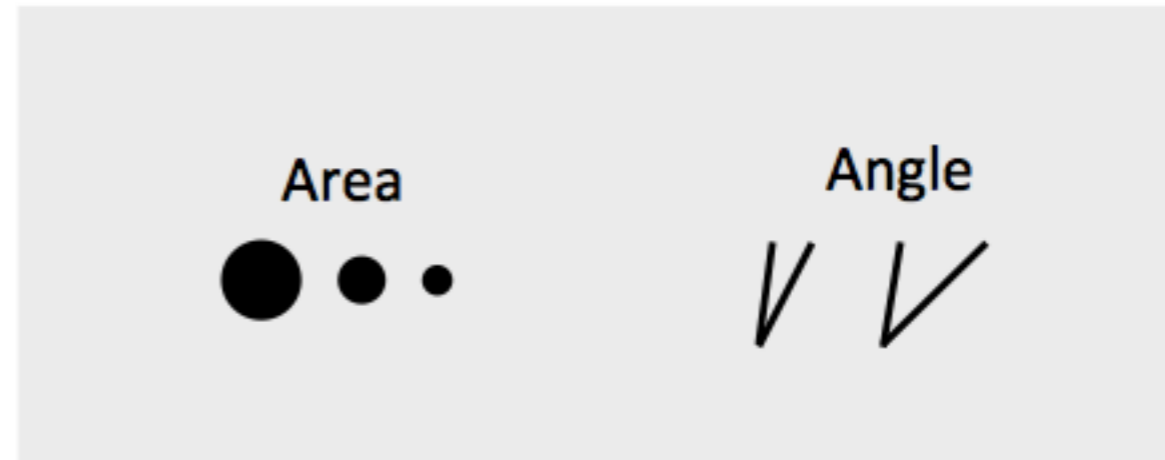
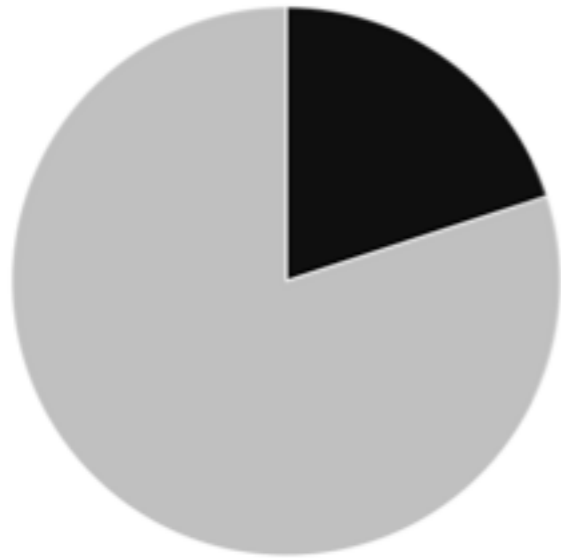
Ordered

Categories

Most Effective

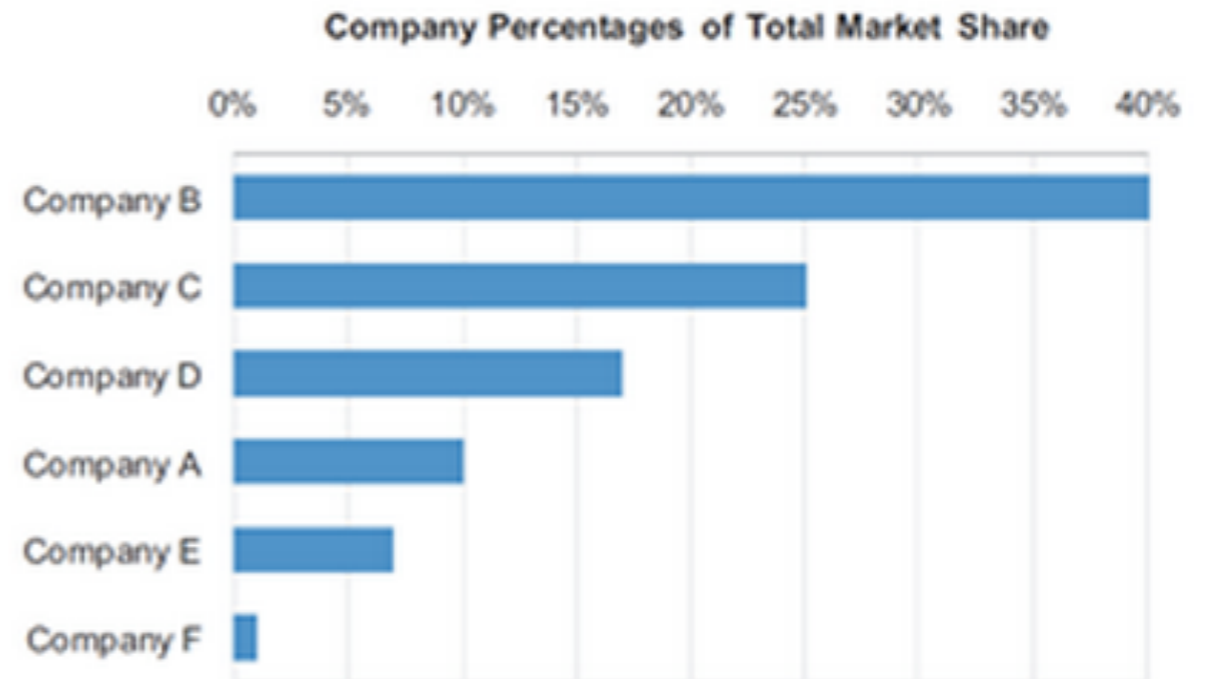
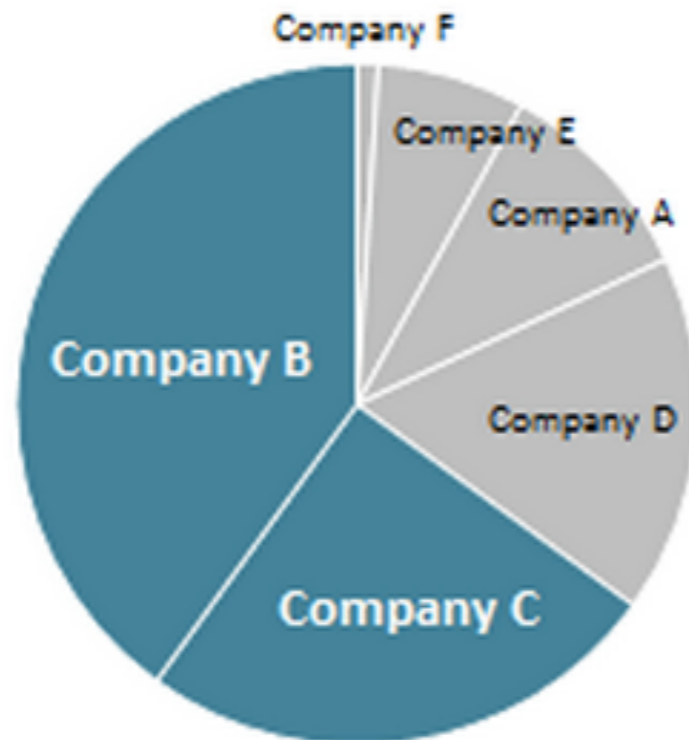


Less Effective



Pie vs. Bar Charts

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



Least Effective

SAMFORD AND SELNICK

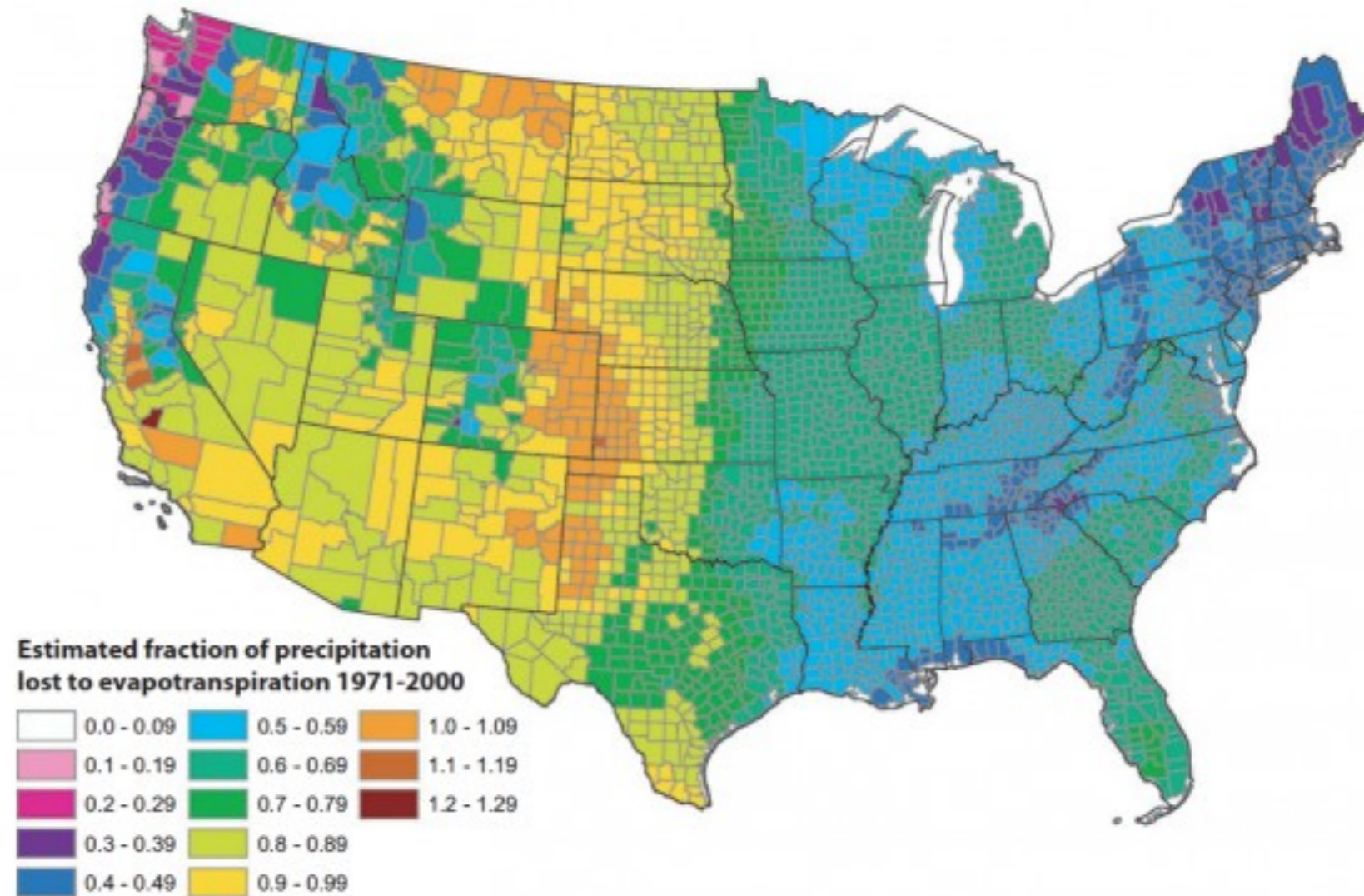
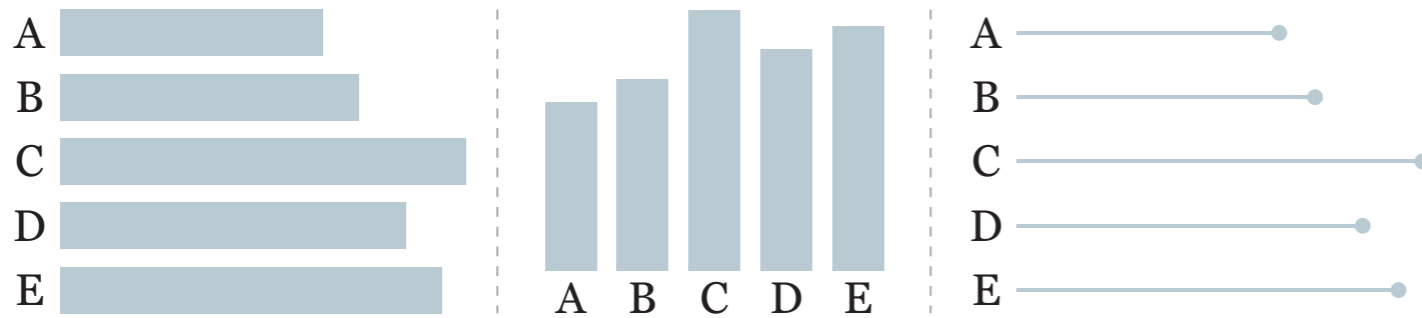
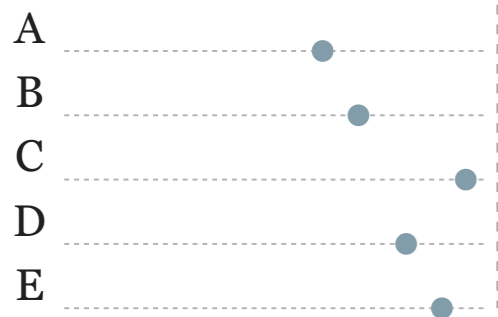


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (P) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/P were made first at the 800-m resolution of the PRISM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

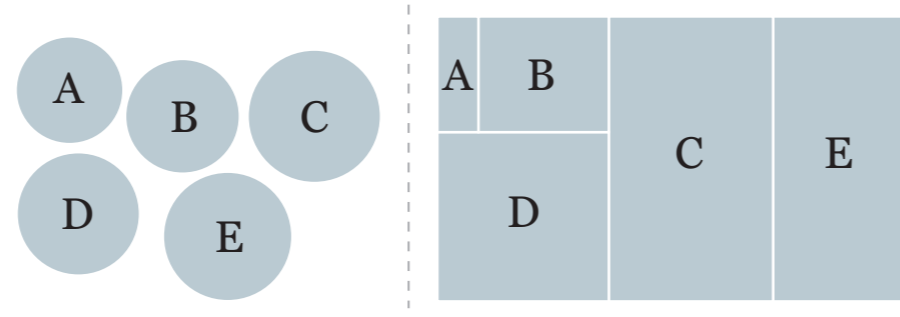
Length or height



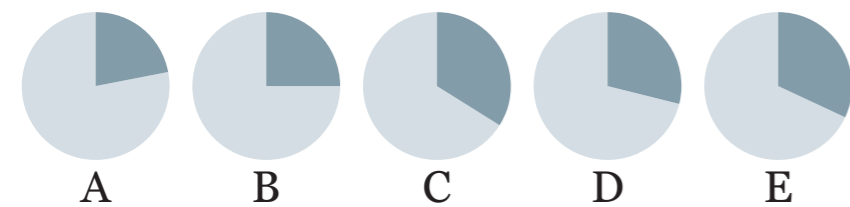
Position



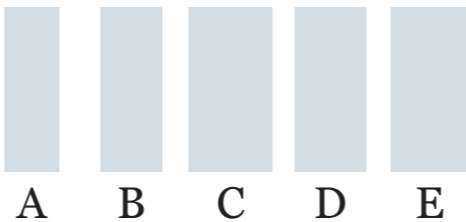
Area



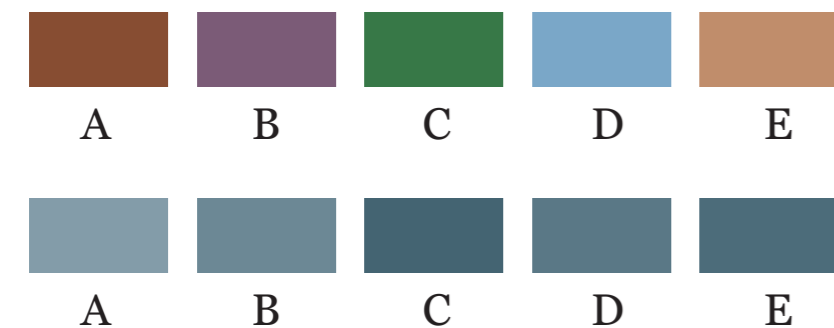
Angle/area



Line weight



Hue and shade



Figures represented
in all these graphics:
22%, 25%, 34%, 29%, 32%

Data visualization
and visual encoding

4. Use Color Strategically

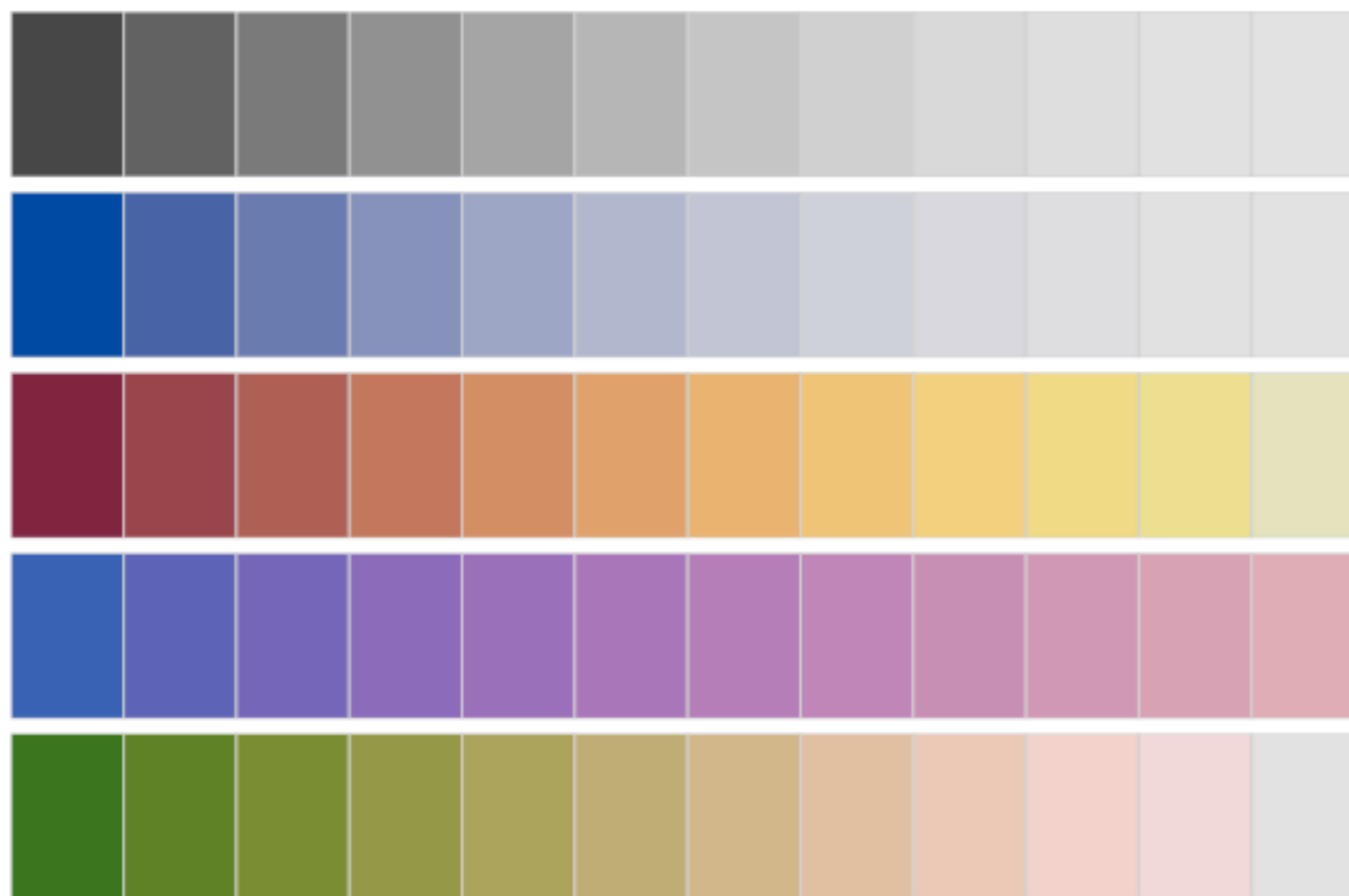
Colors for Categories

Do not use more than 5-8 colors at once



Colors for Ordinal Data

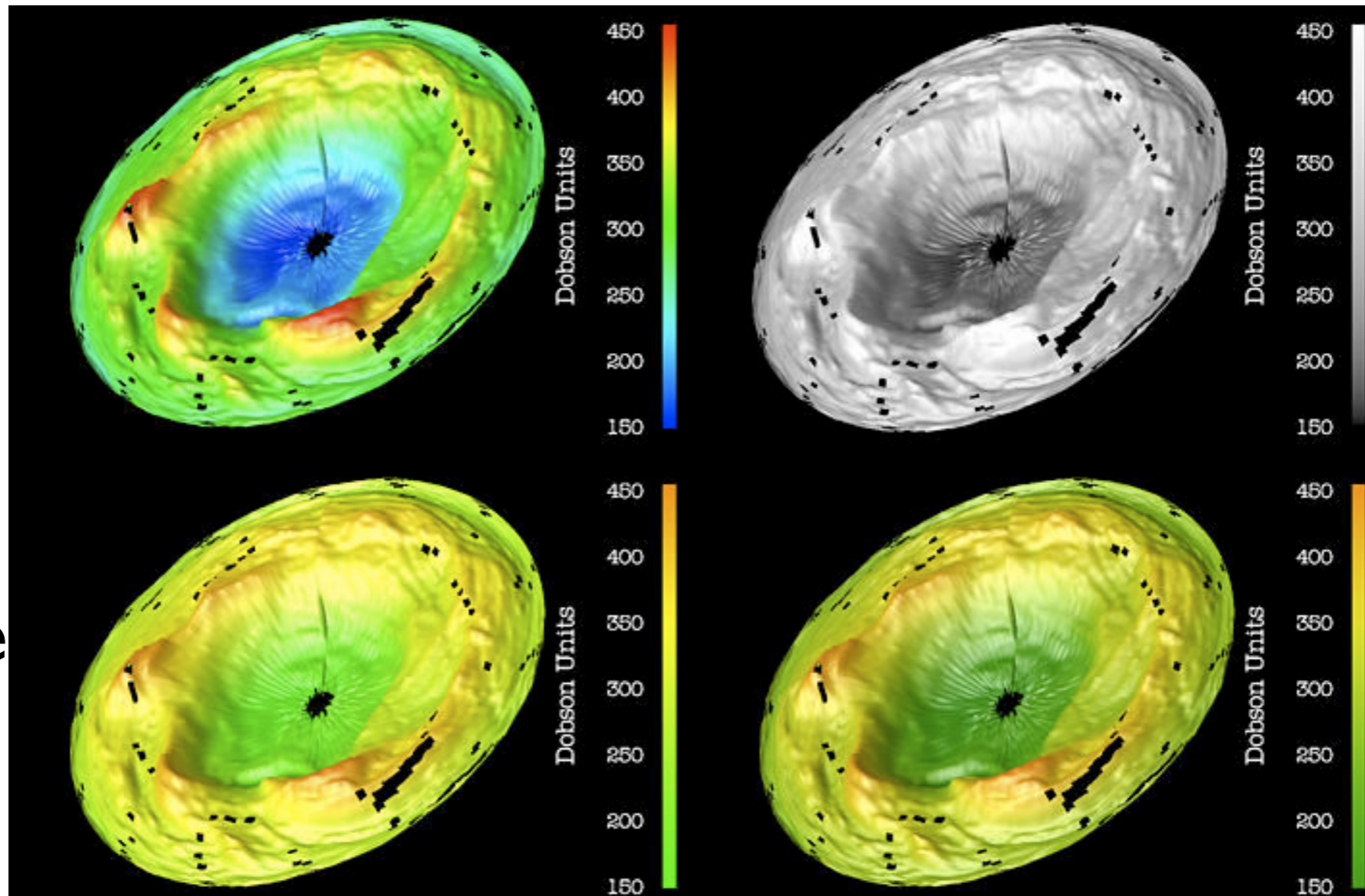
Vary luminance and saturation



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

Colors for Quantitative Data

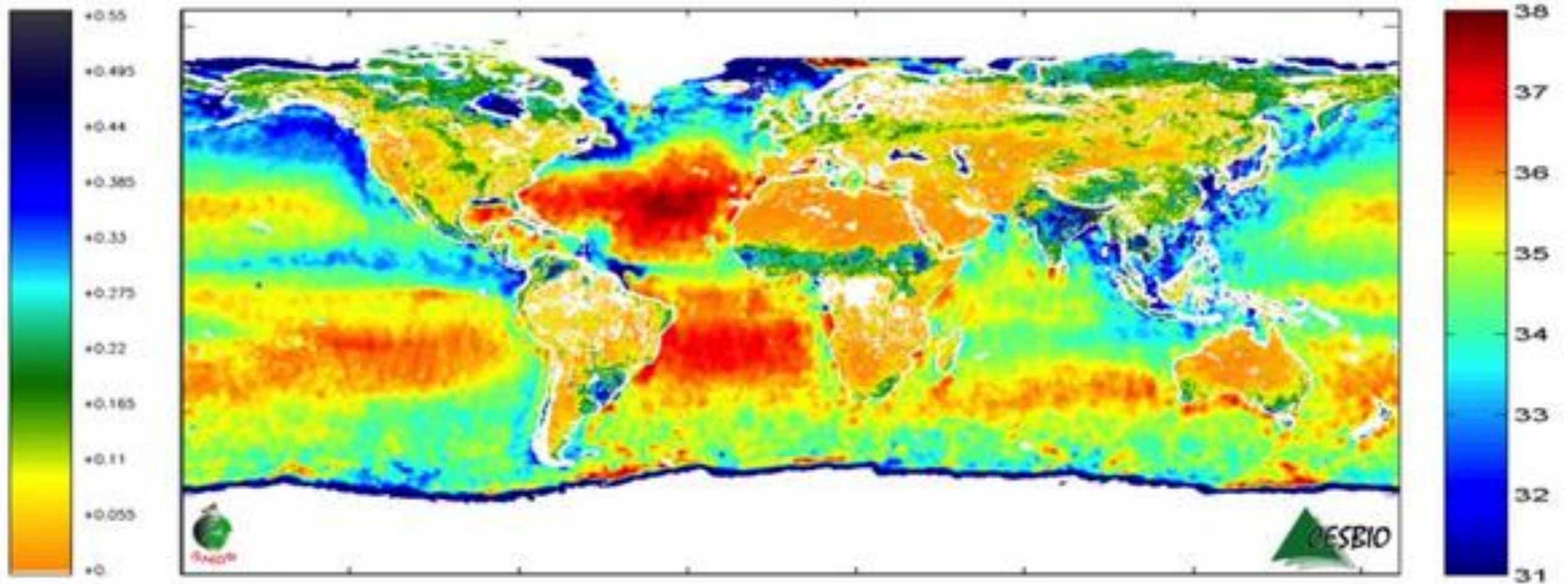
Hue
(Rainbow)



Luminance

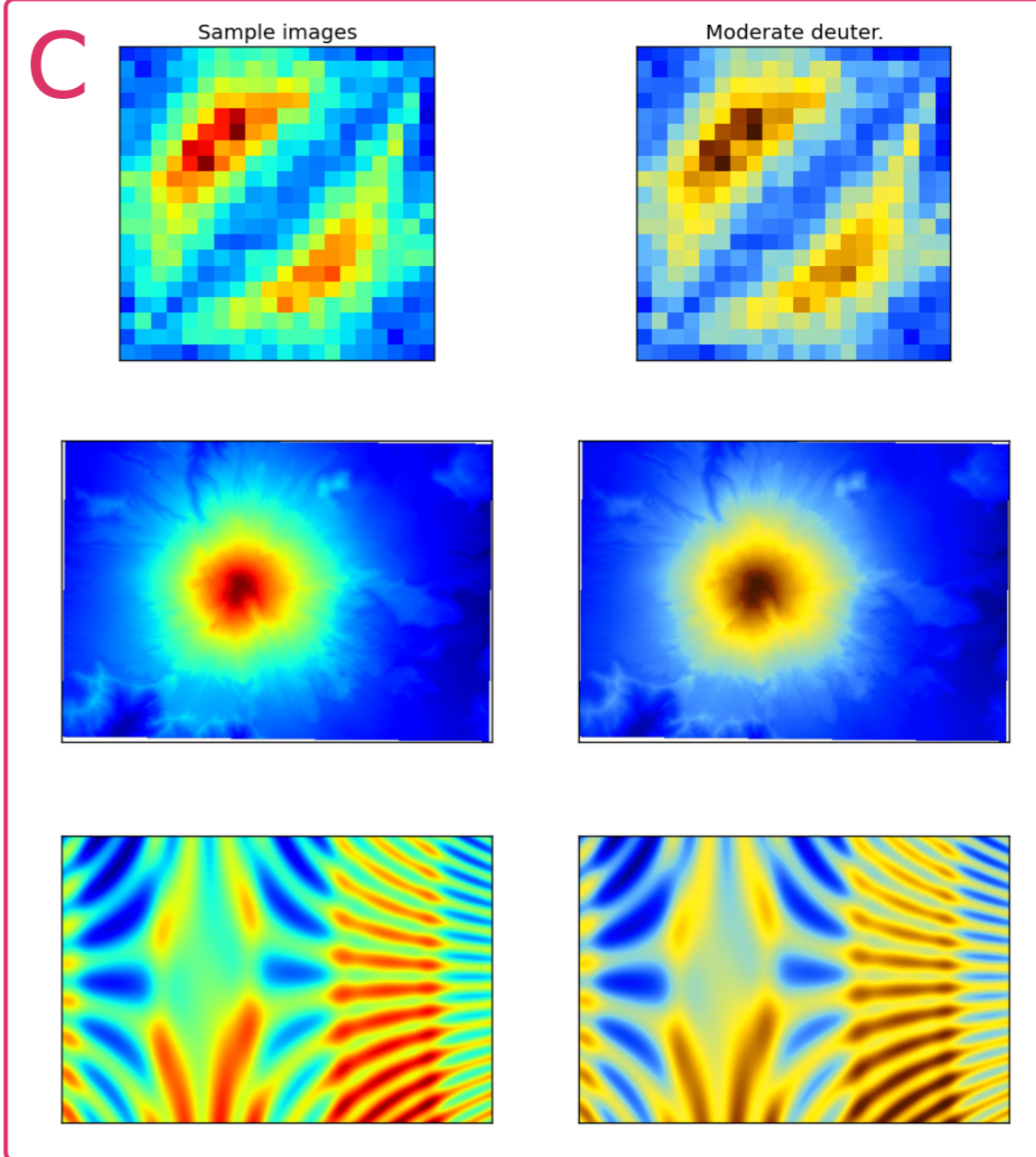
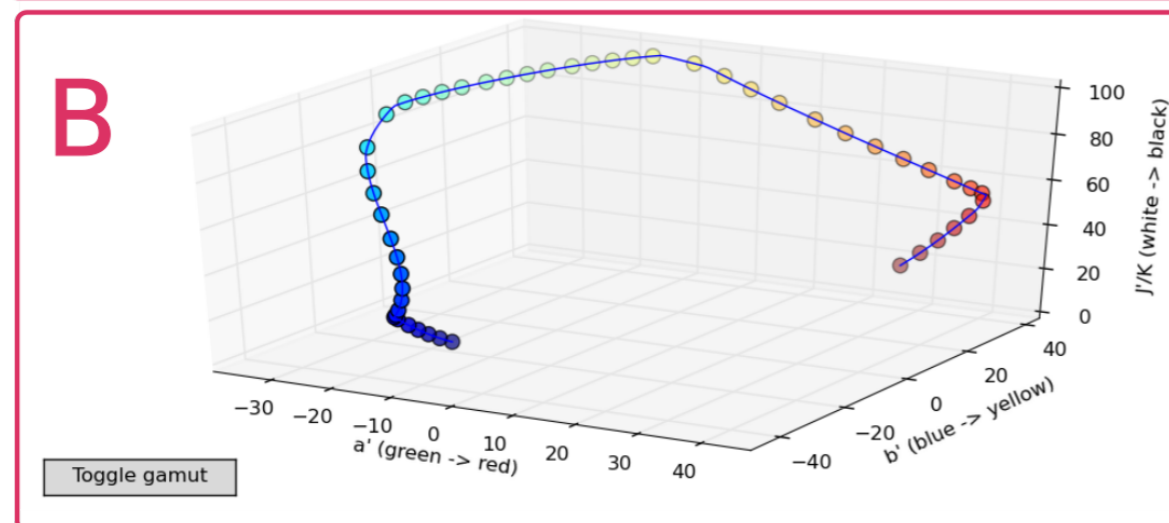
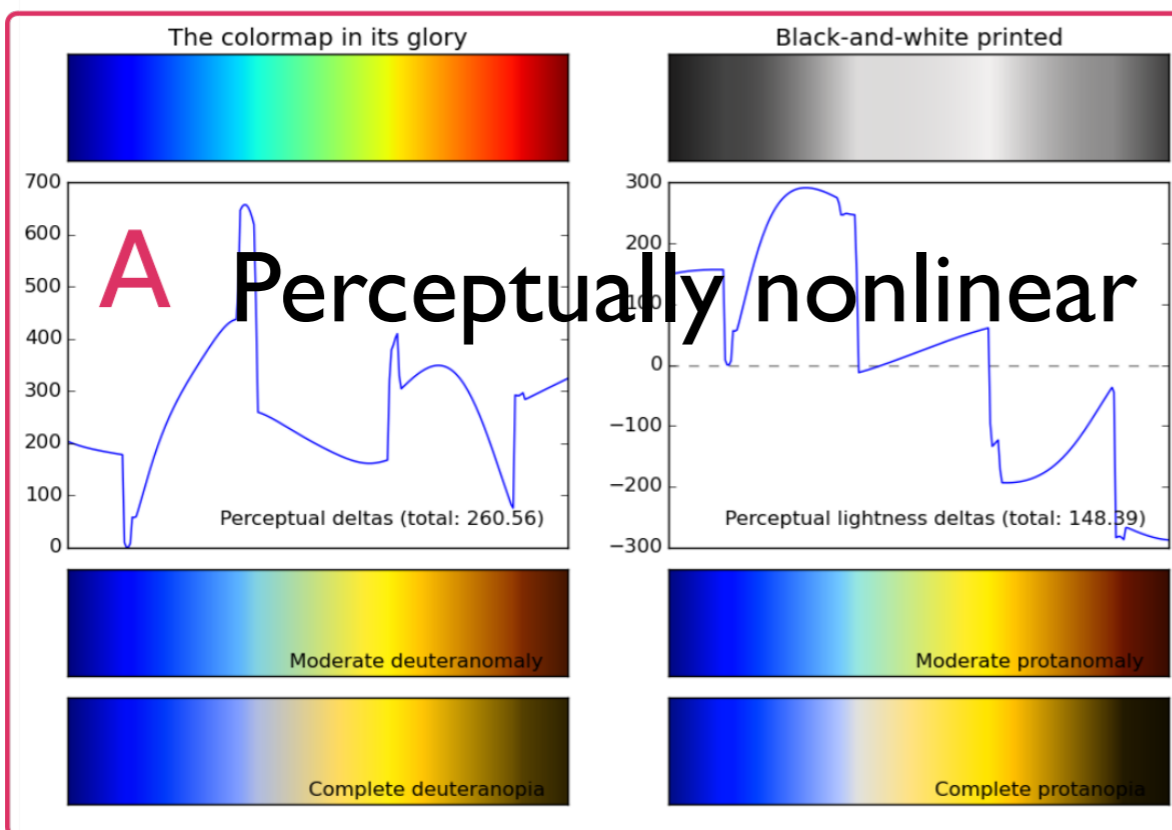
Luminance
& Hue

Rainbow Colormap

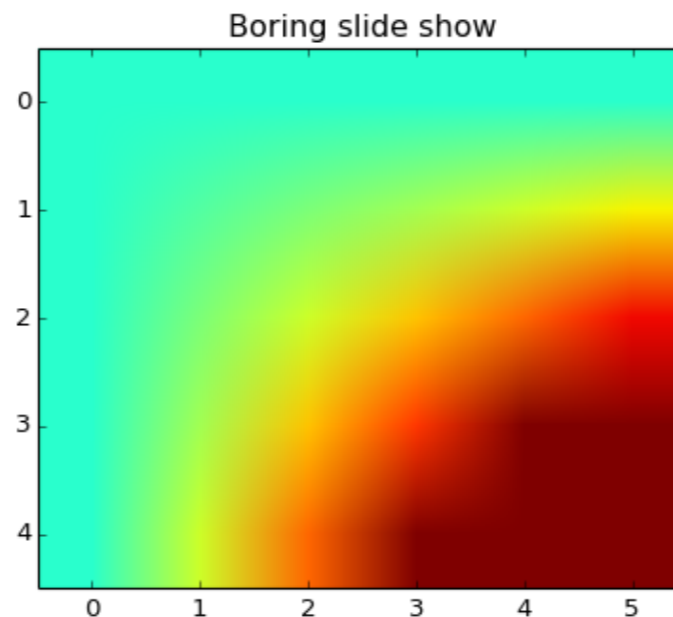
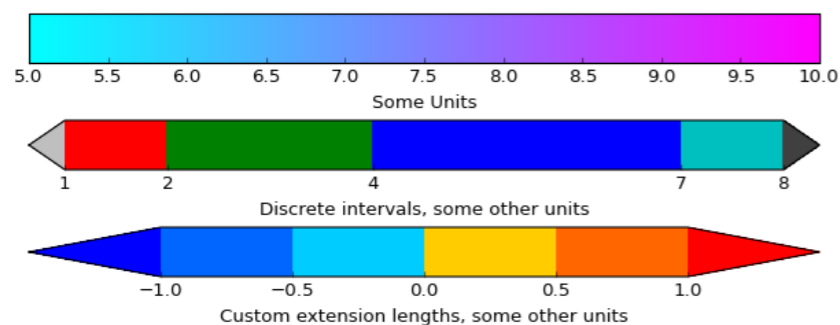


Rainbow Colormap

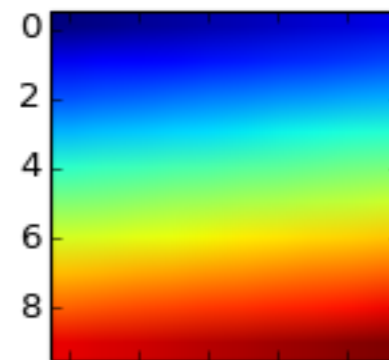
Colormap evaluation: jet



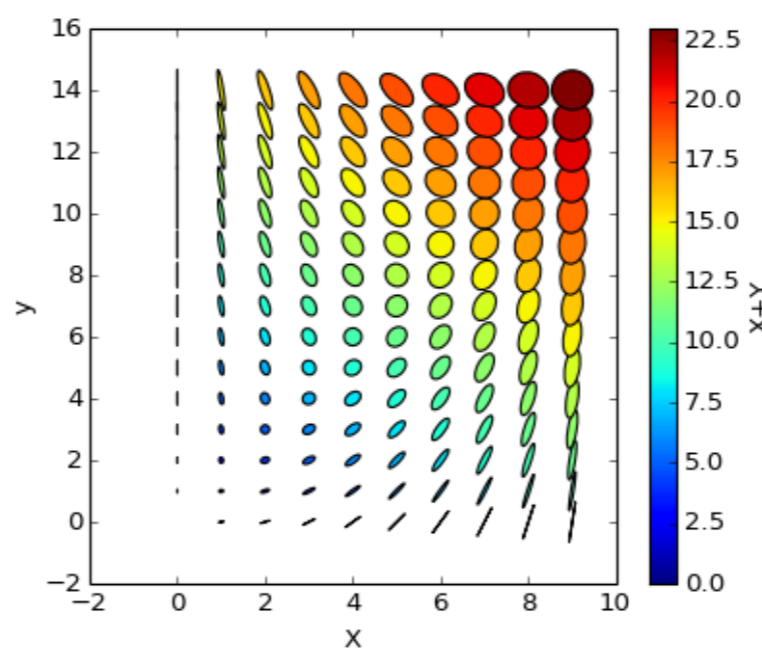
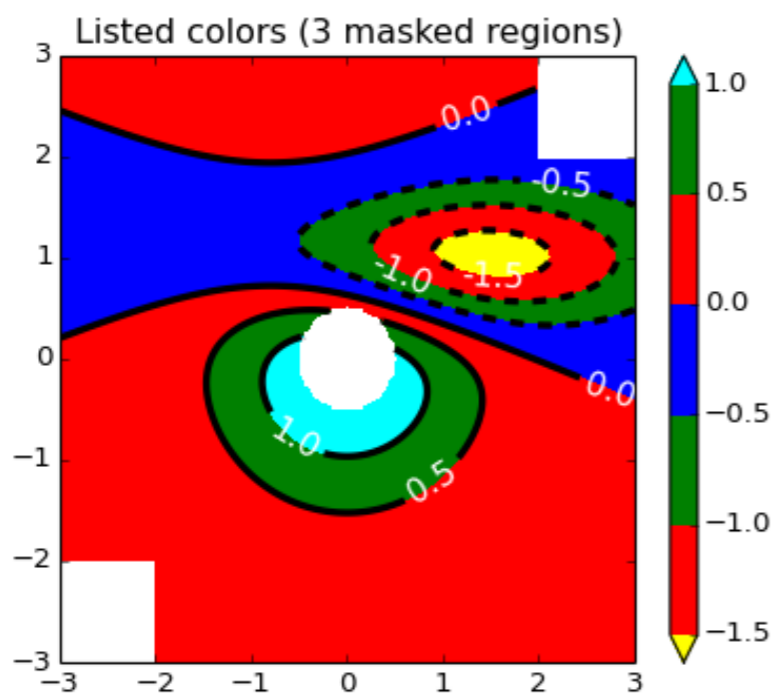
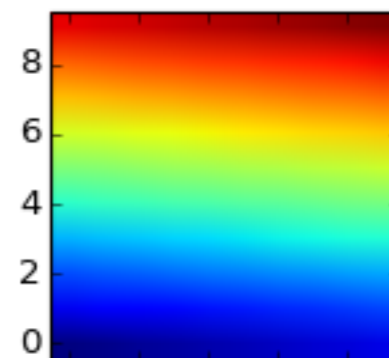
Avoid Rainbow Colors!



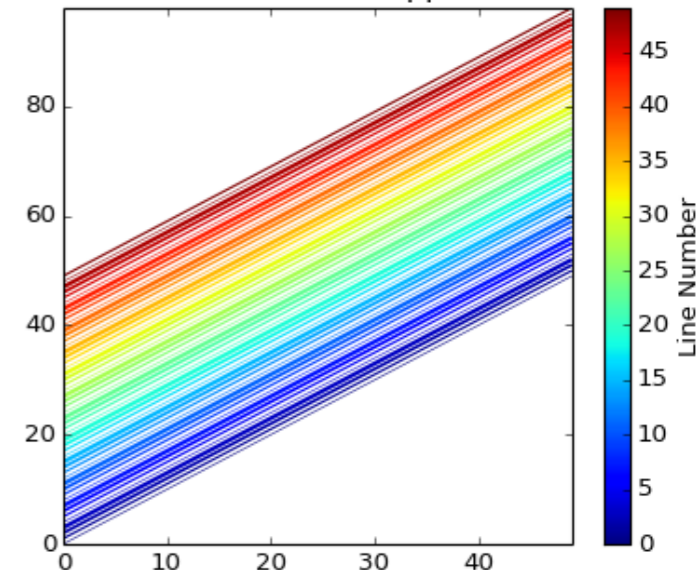
blue should be up



blue should be down

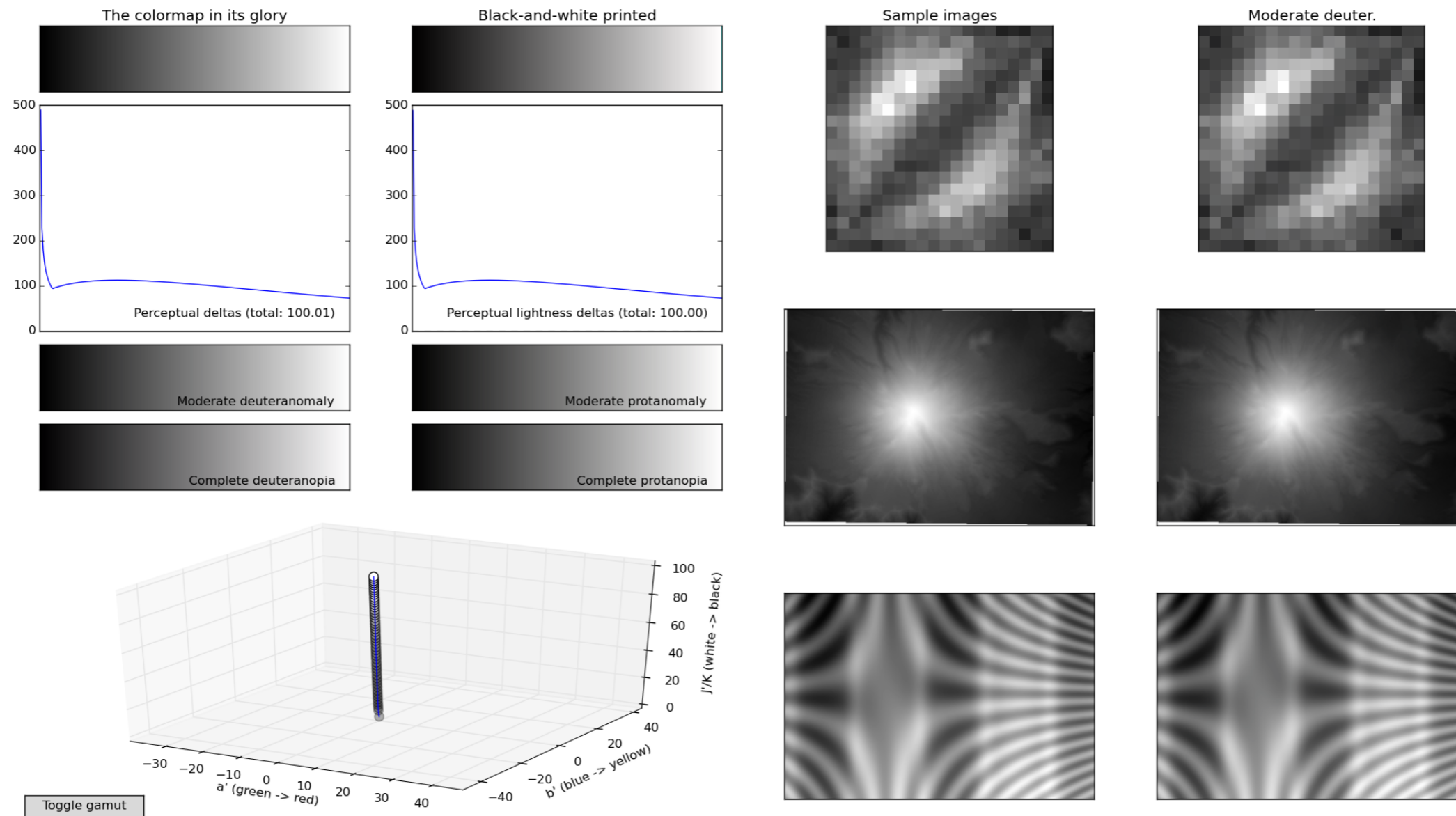


Line Collection with mapped colors



Gray

Colormap evaluation: gray



Color Blindness



Protanope

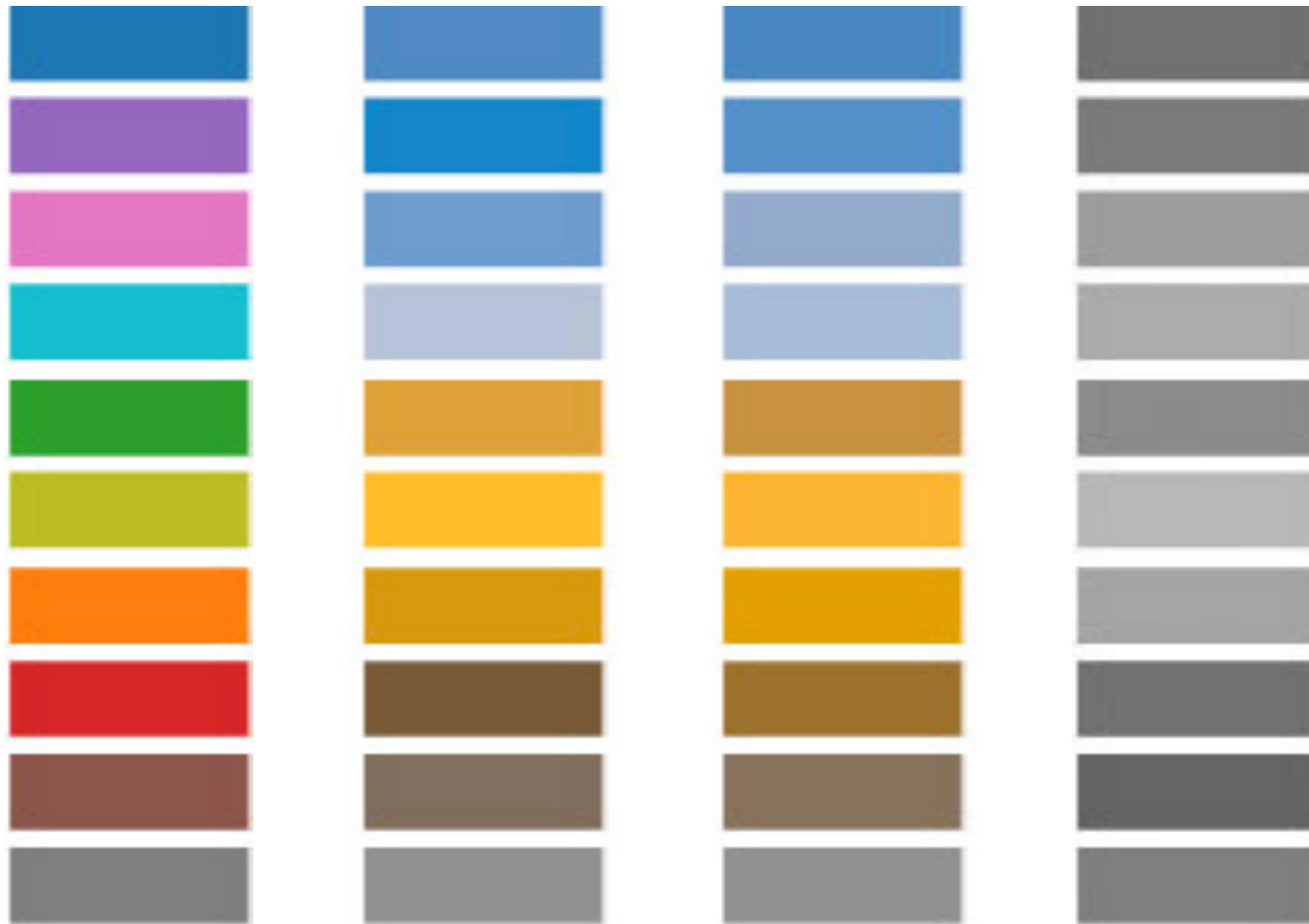
Deuteranope

Tritanope

Red / green
deficiencies

Blue / Yellow
deficiency

Color Blindness



Normal

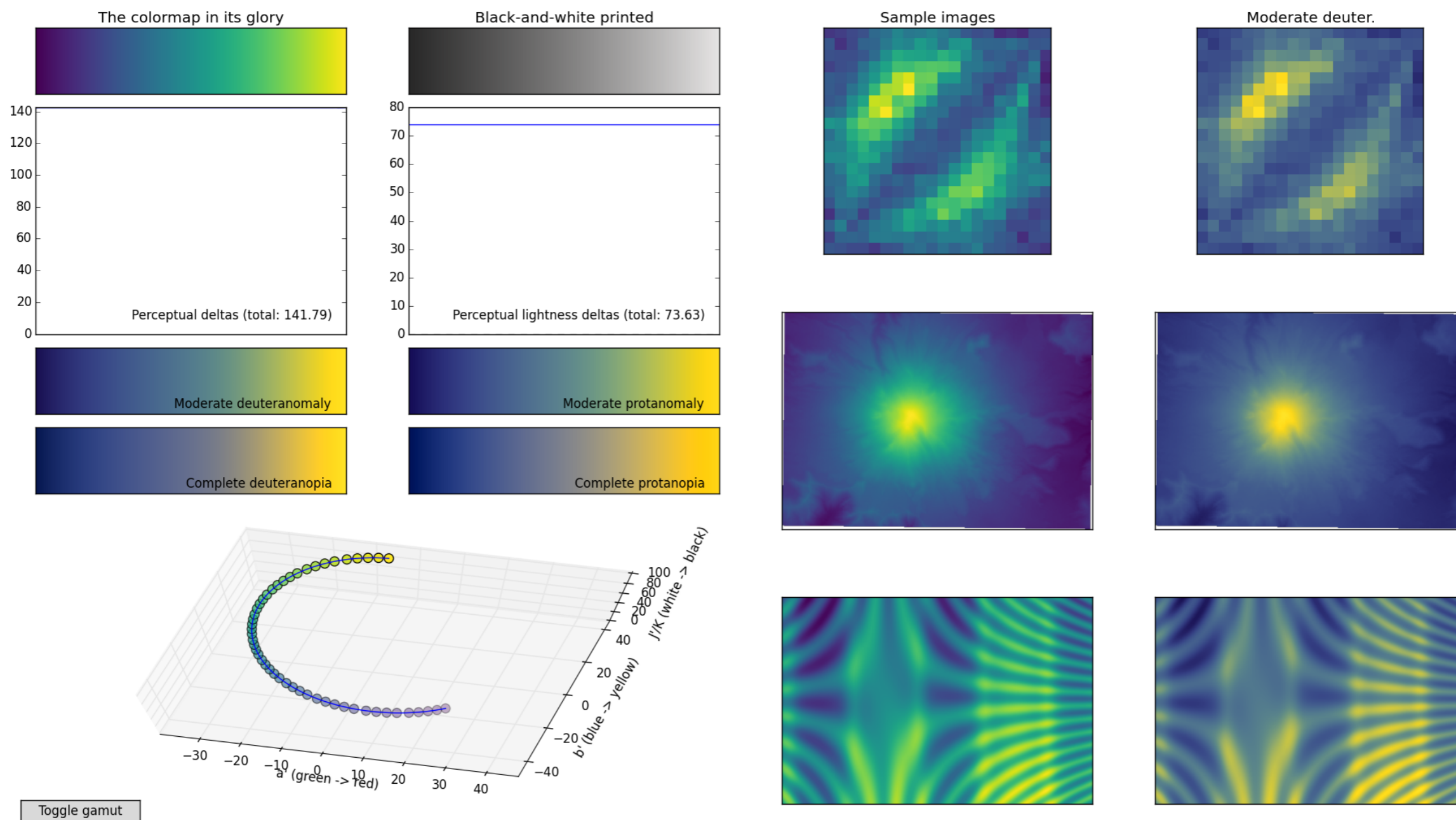
Protanope

Deuteranope

Lightness

Viridis

Colormap evaluation: option_d.py



Color Brewer

Nominal

Qualitative Scale



Ordinal

Sequential Scale



0 → Max

Diverging Scale



Max ← 0 → Max

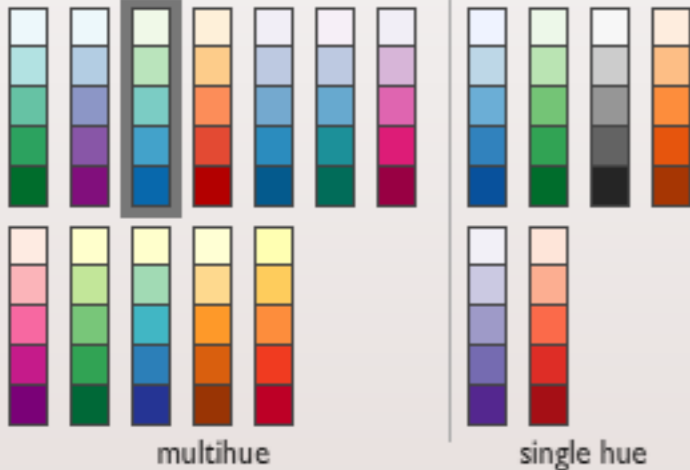
number of data classes on your map

3 [learn more >](#)

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

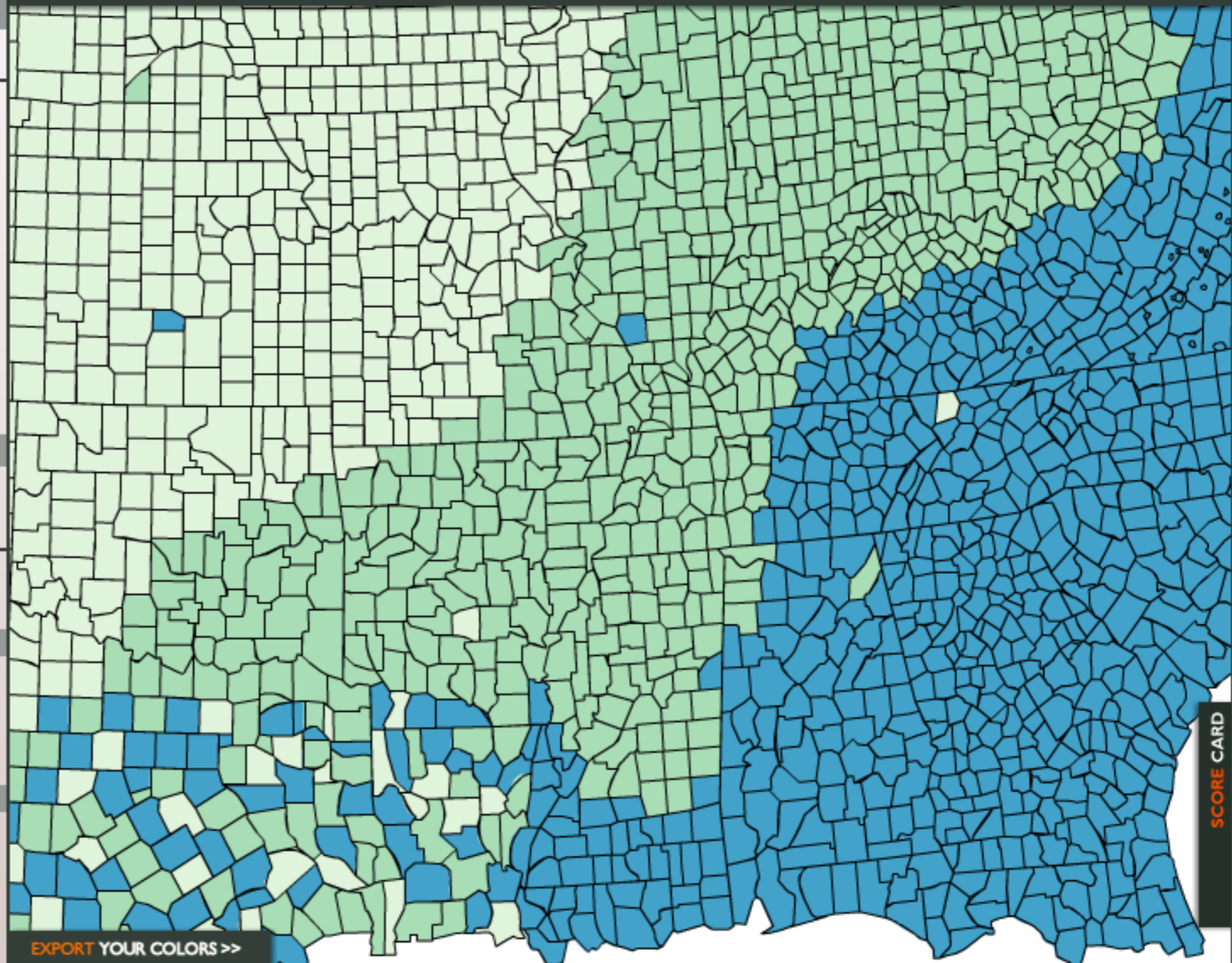
select a background

- solid color
 - terrain
- color transparency

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

COLORBREWER 2.0

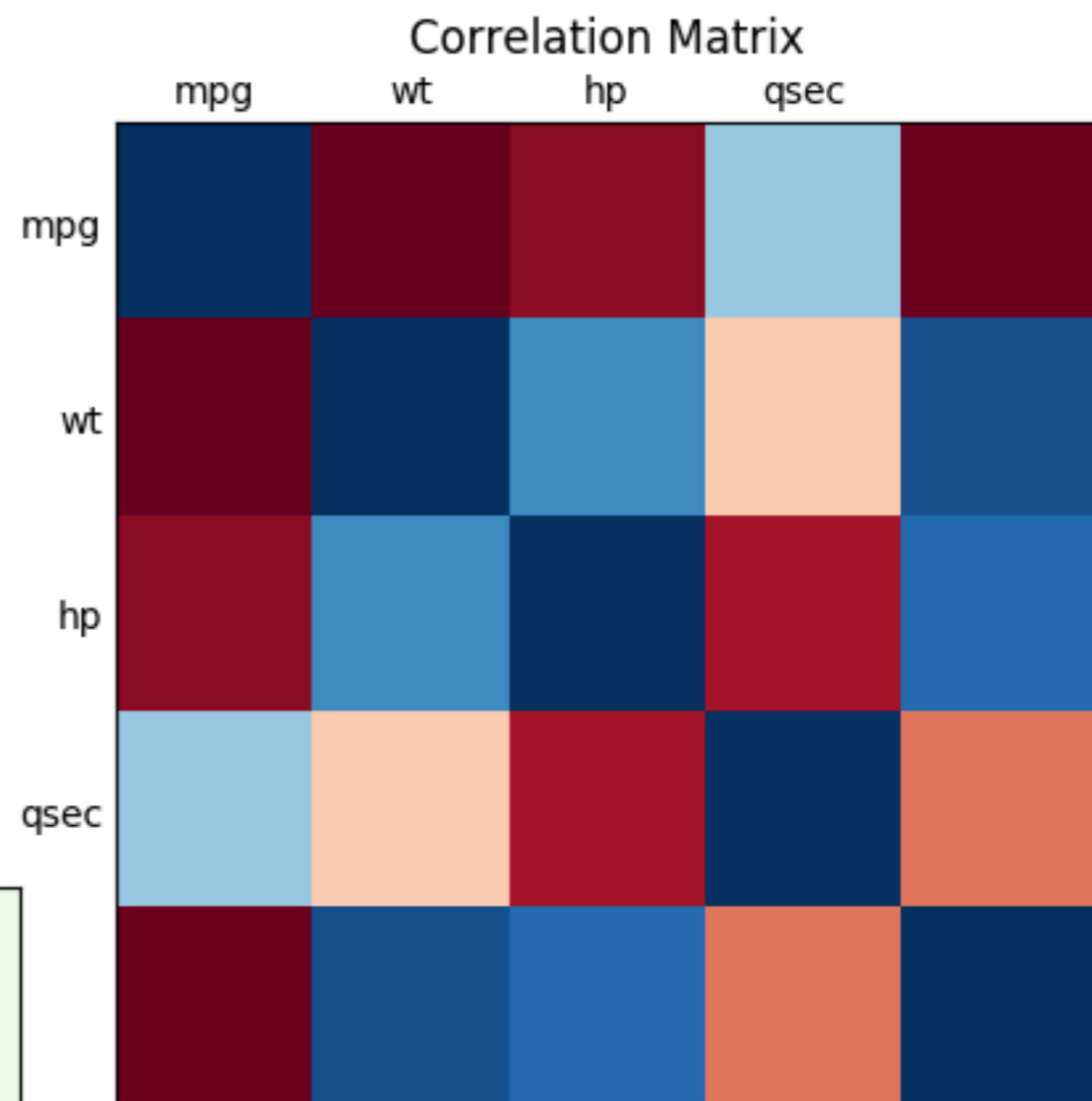
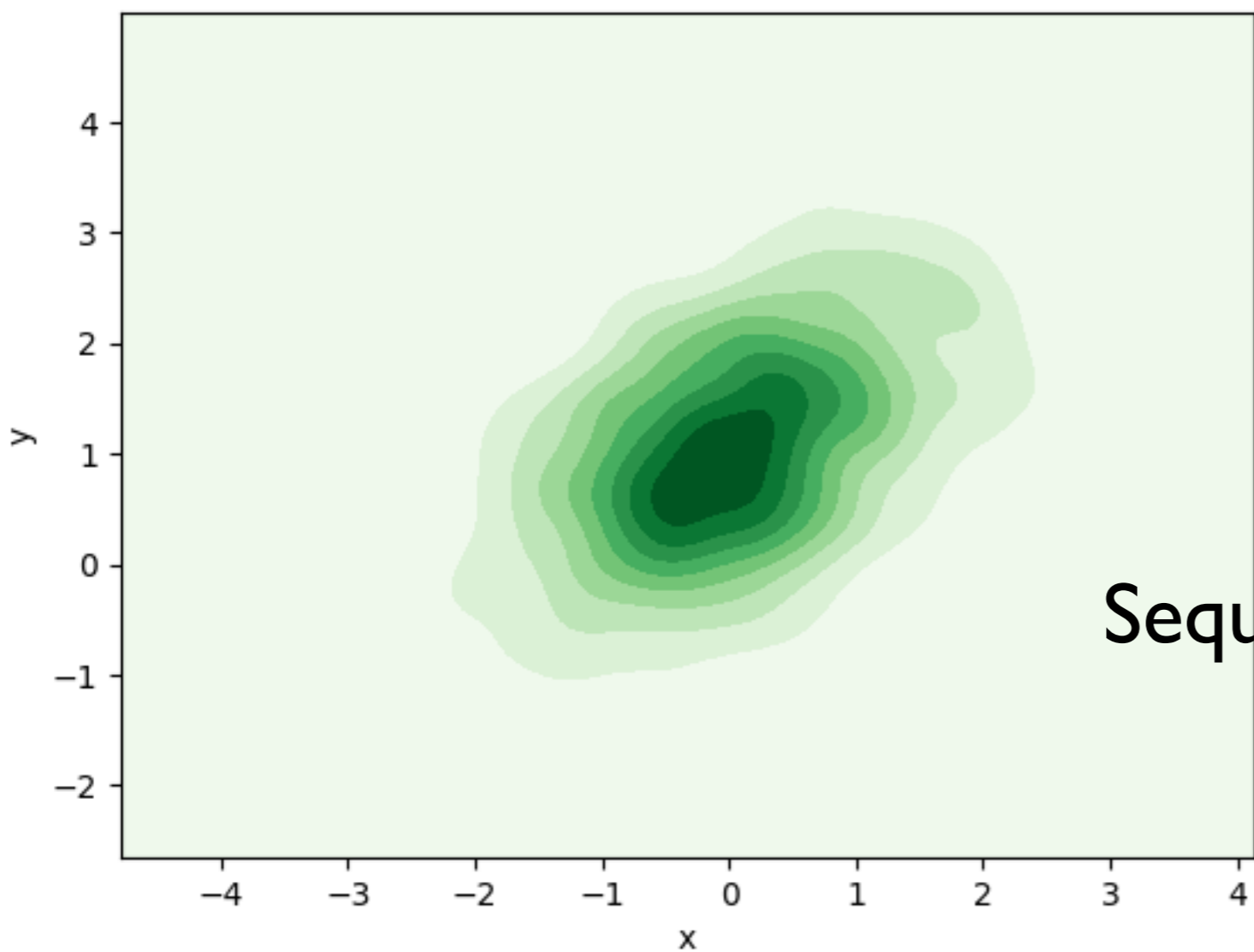
color advice for cartography



[EXPORT YOUR COLORS >>](#)

SCORE CARD

Diverging Palette for Quantitative or Ordinal



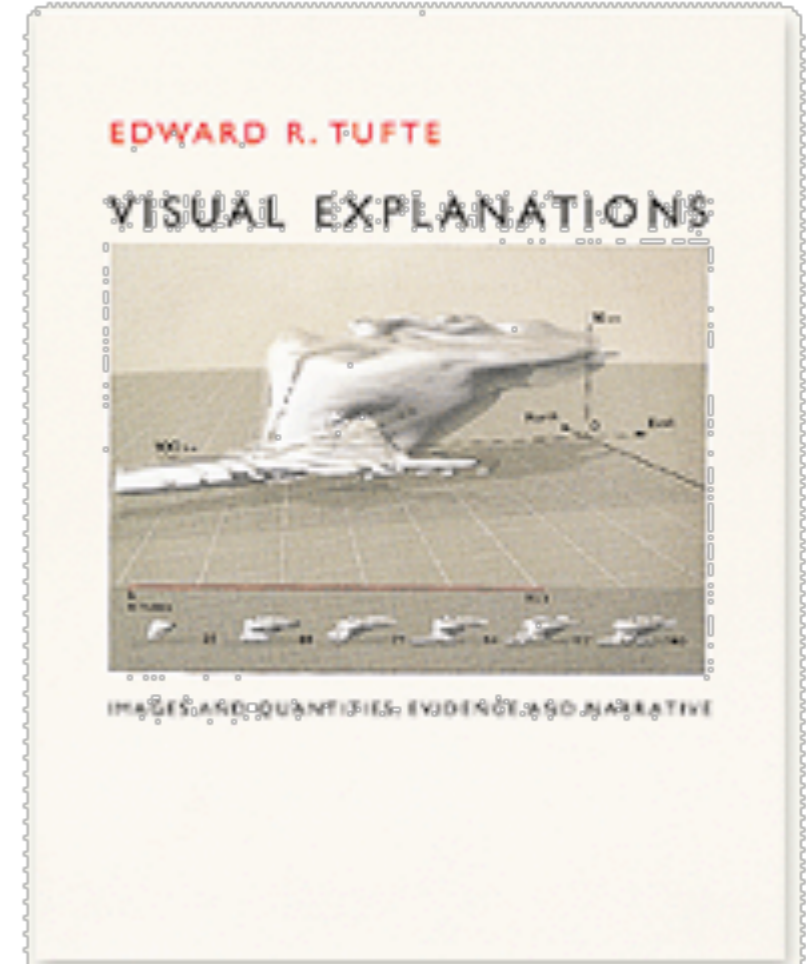
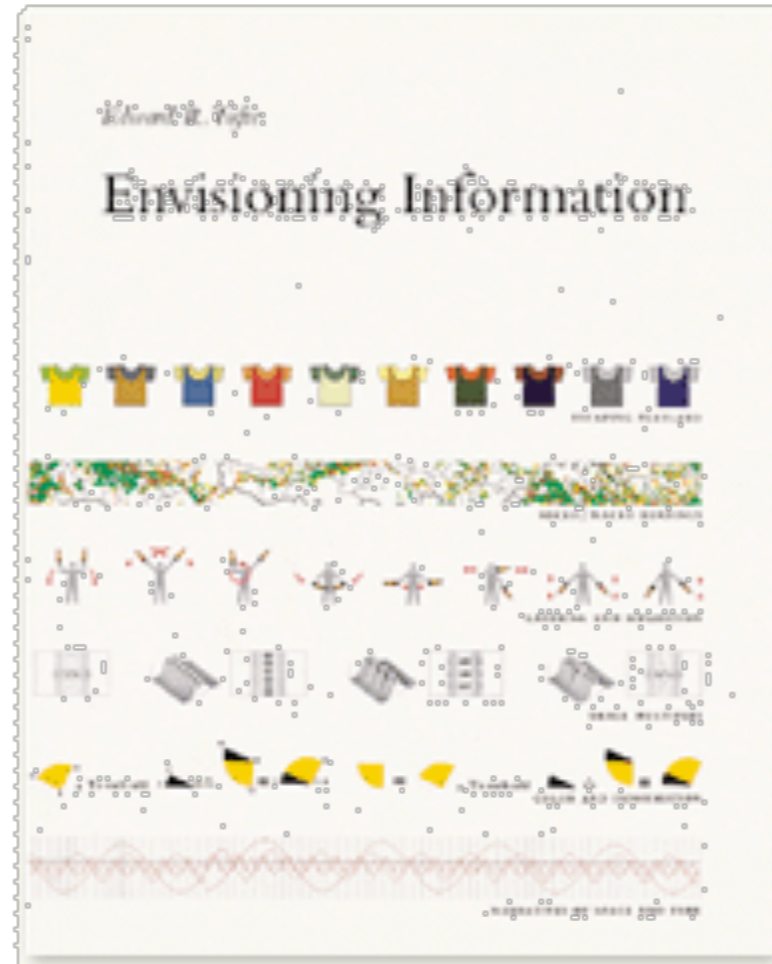
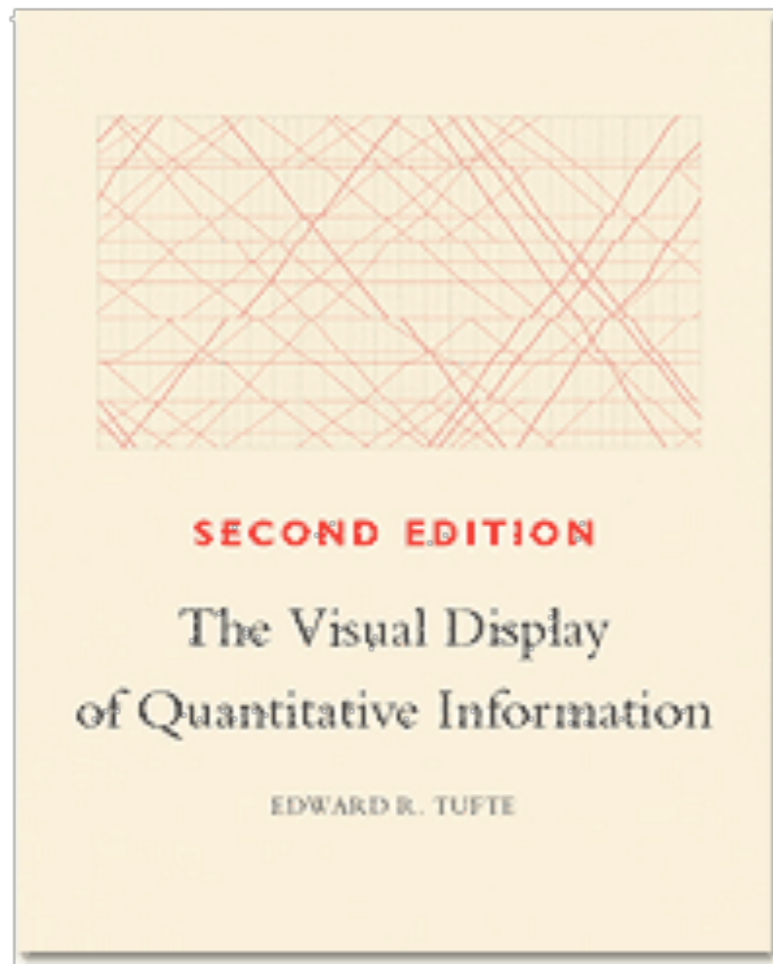
Sequential Palette for Densities

Effective Visualizations

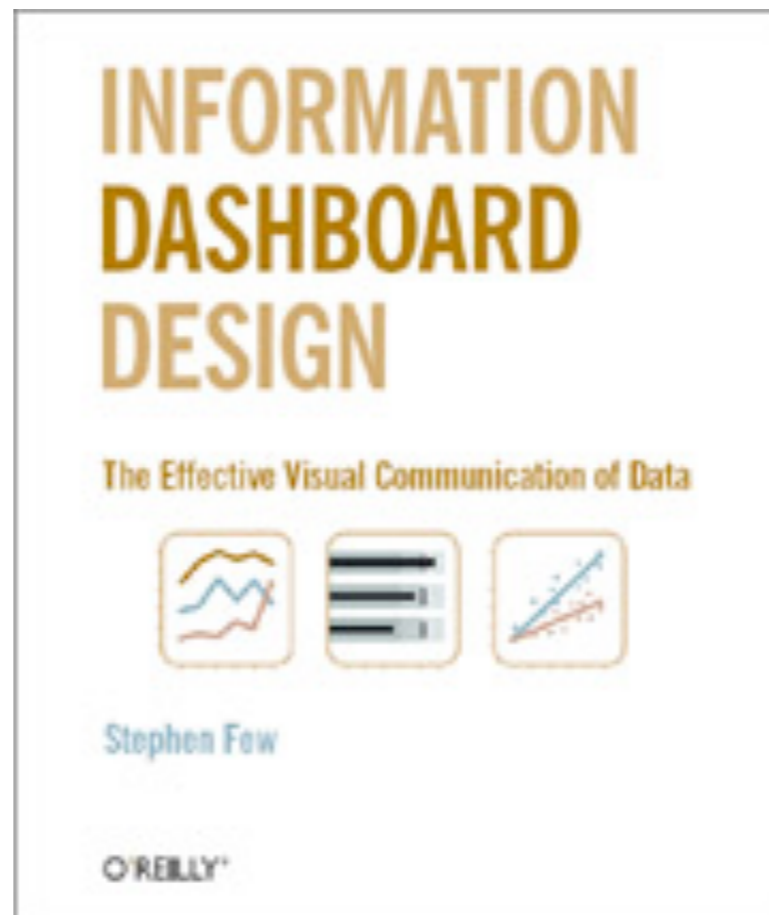
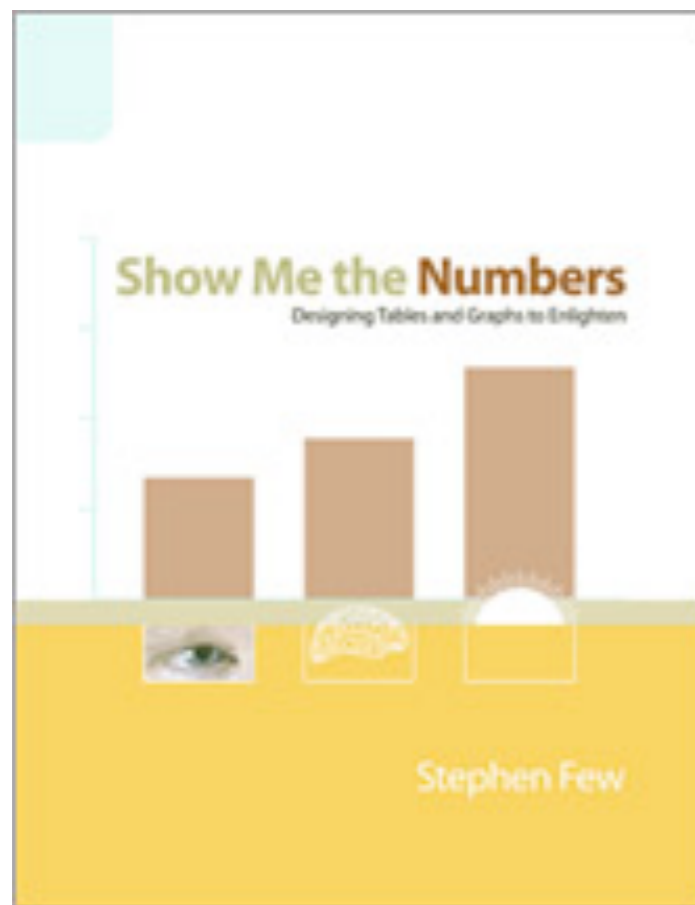
1. Have graphical integrity
2. Keep it simple
3. Use the right display
4. Use color strategically

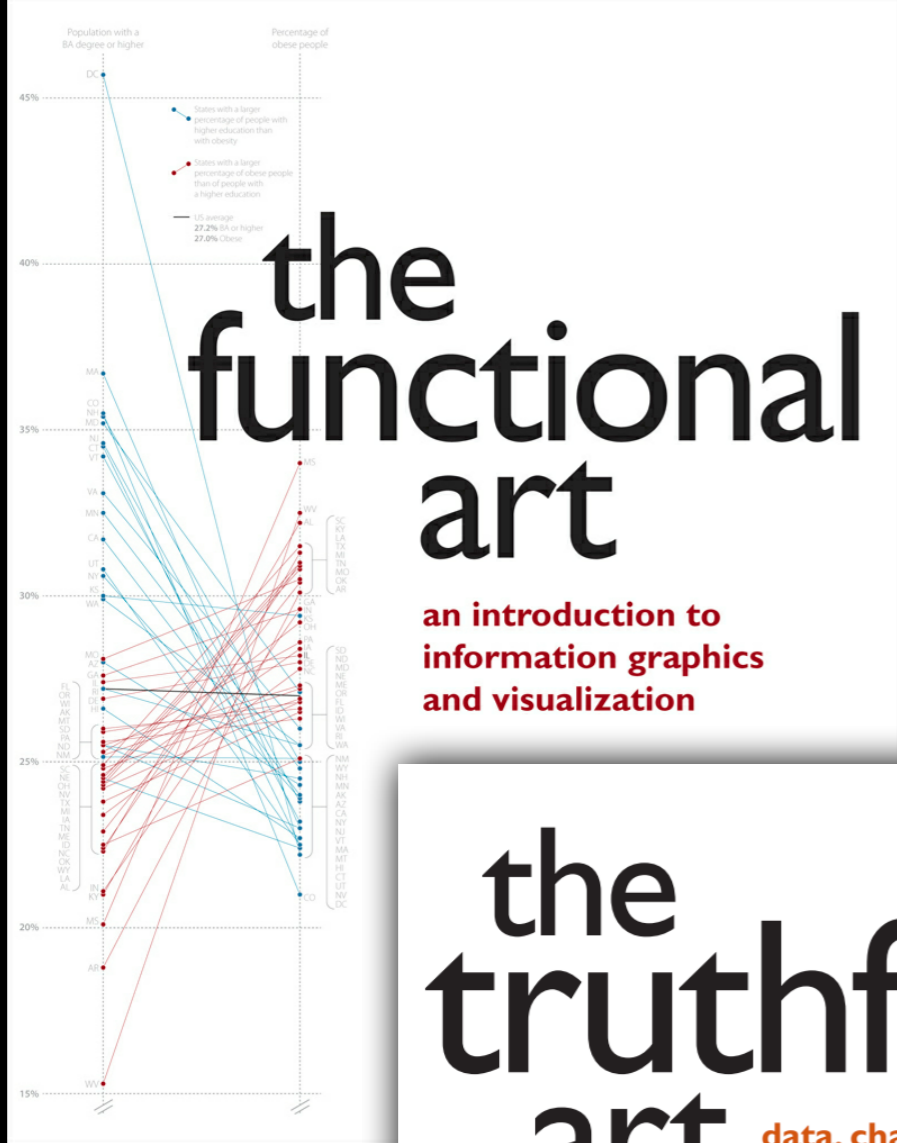
Further Reading

Edward Tufte

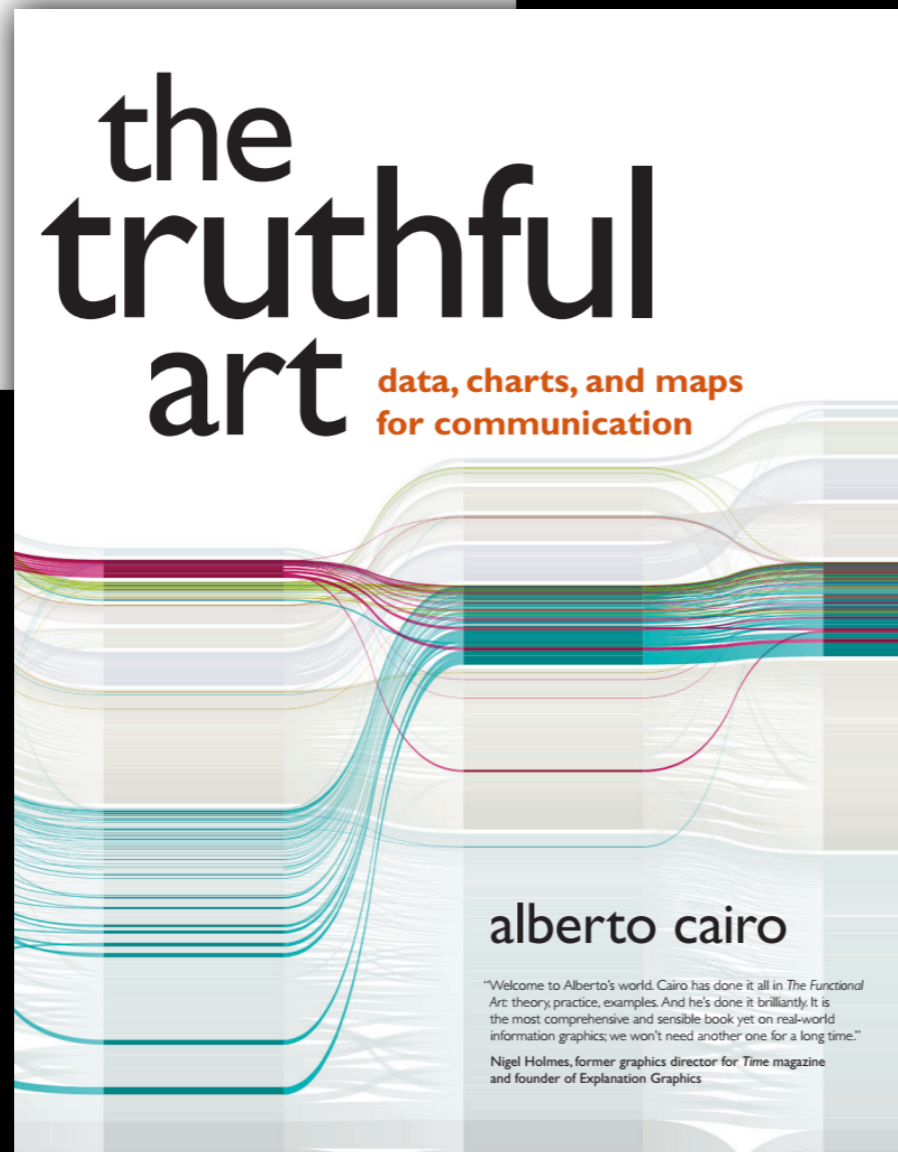


Stephen Few





2016



I've always believed in the power of data visualization (the representation of information by means of charts, diagrams, maps, etc.) to enable understanding