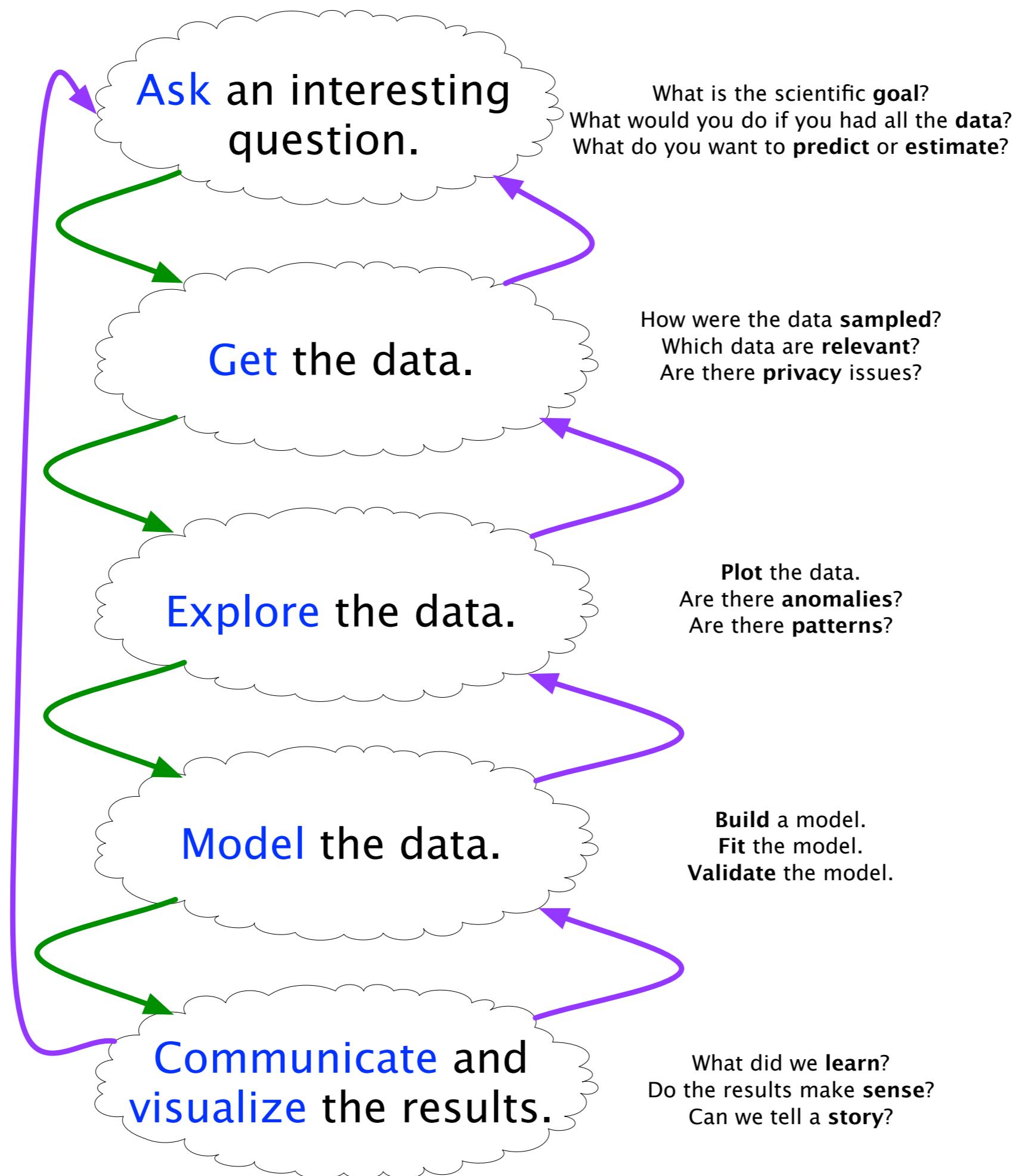
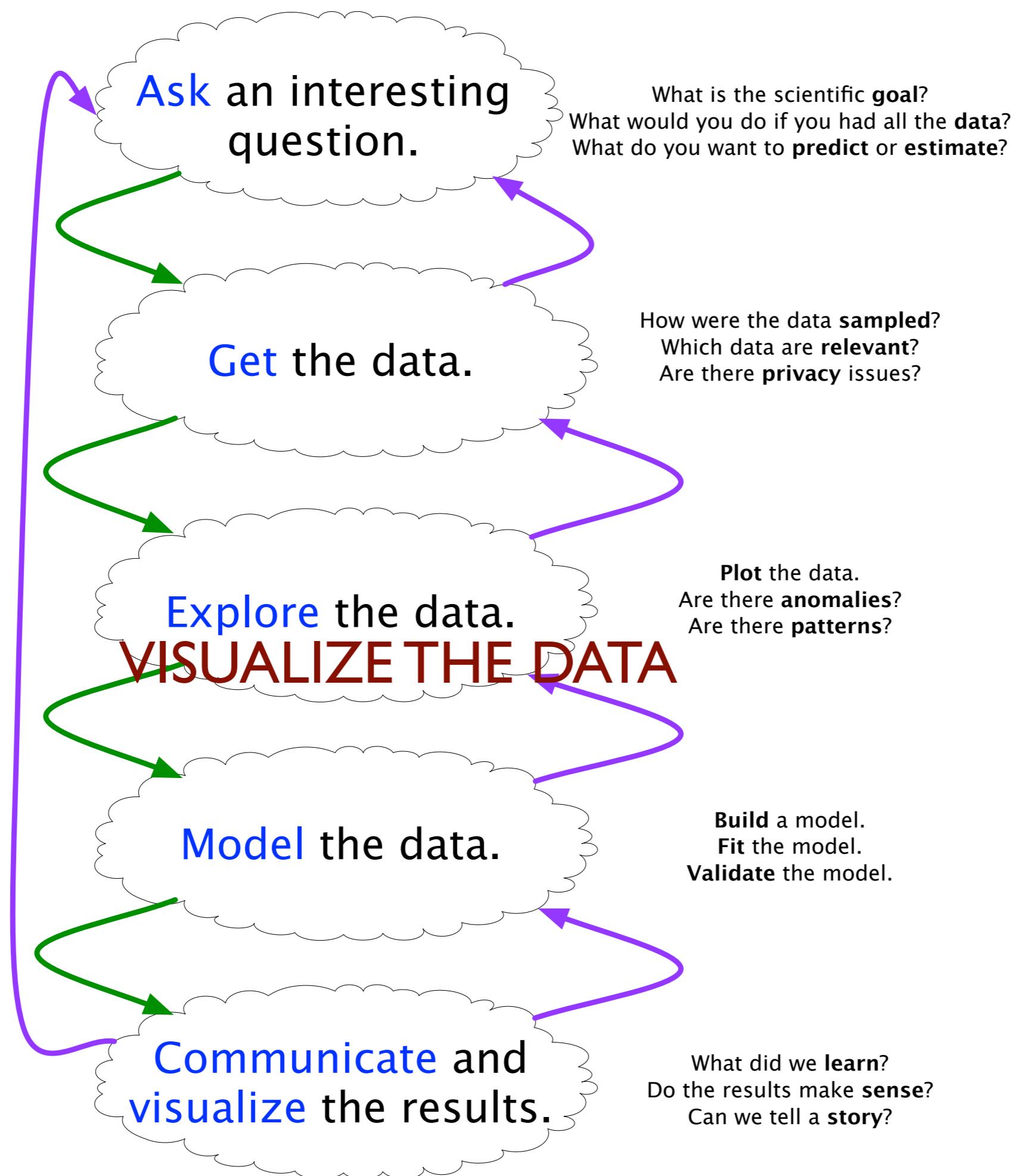


# CS 109a: Data Science

## Effective Exploratory Data Analysis and Visualization

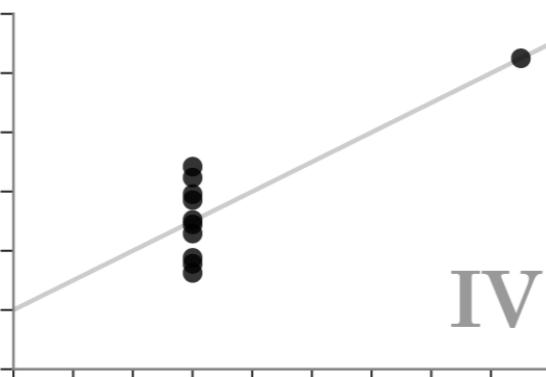
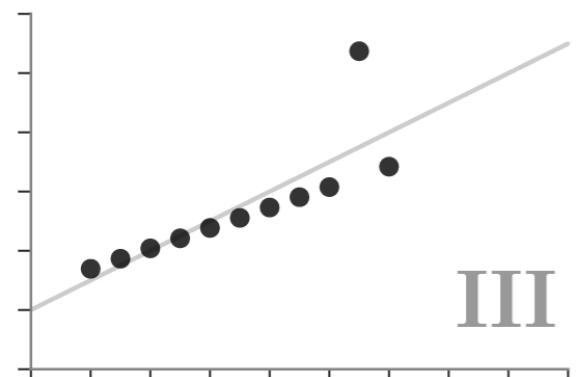
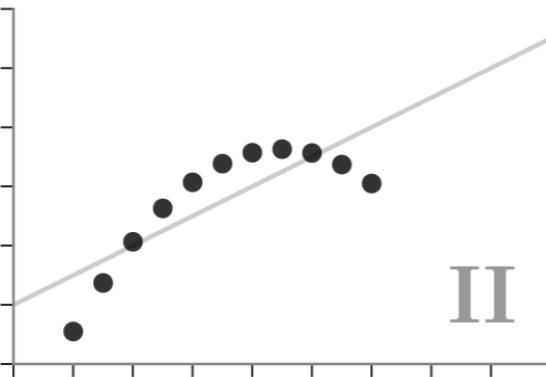
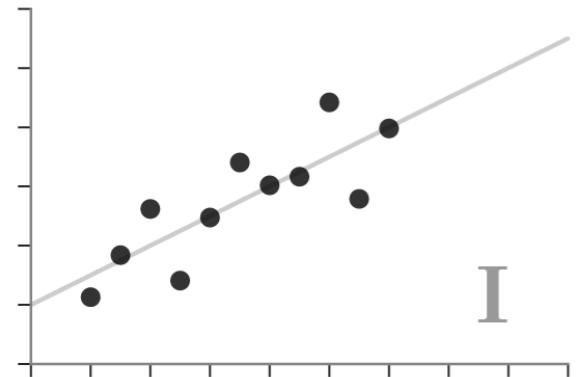
Pavlos Protopapas, Kevin Rader, Rahul Dave, Margo Levine





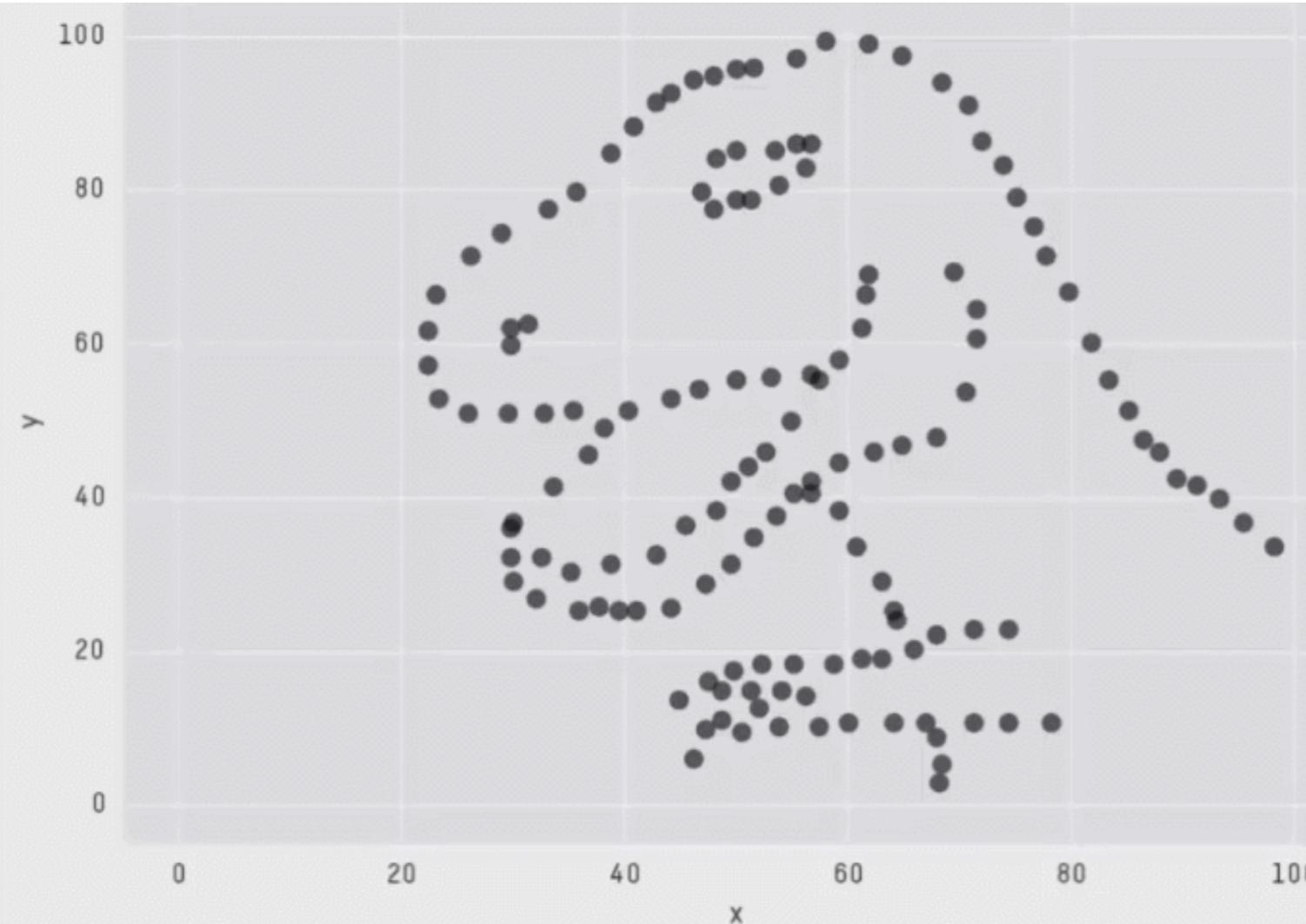
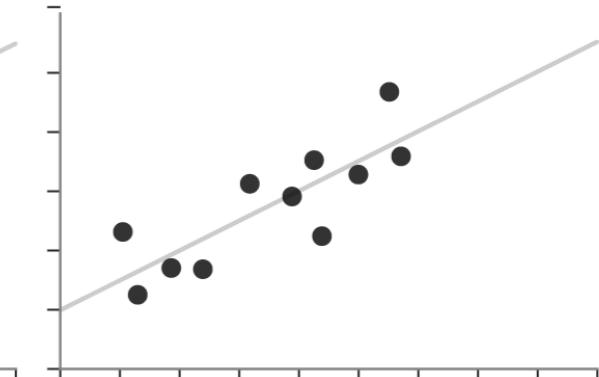
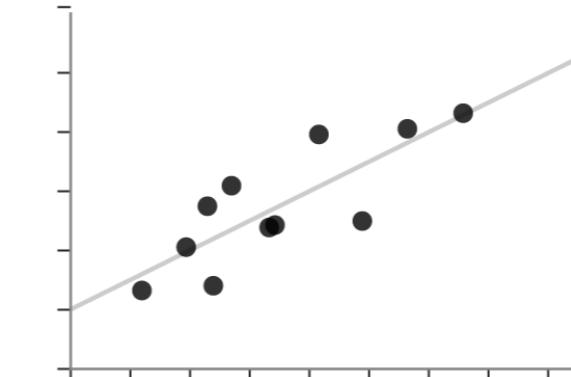
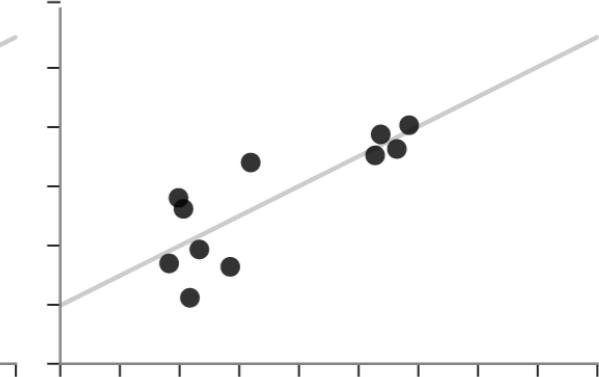
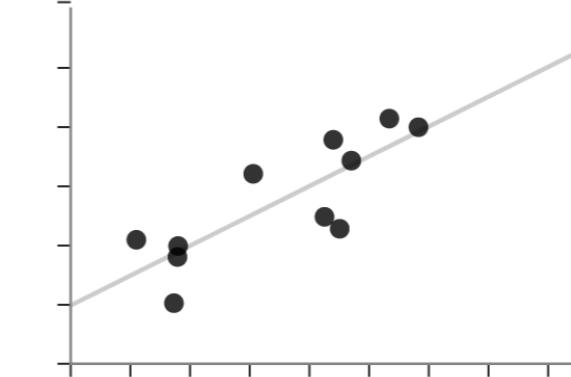
## ✓ Anscombe's Quartet

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



## ✗ Unstructured Quartet

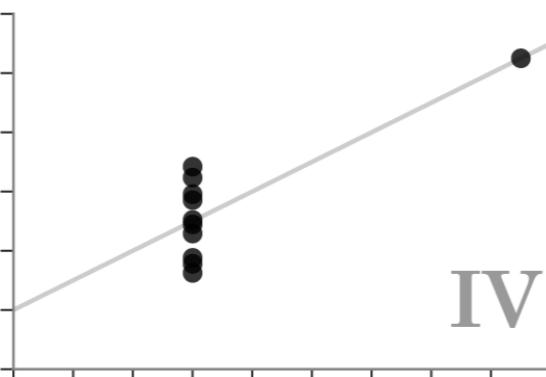
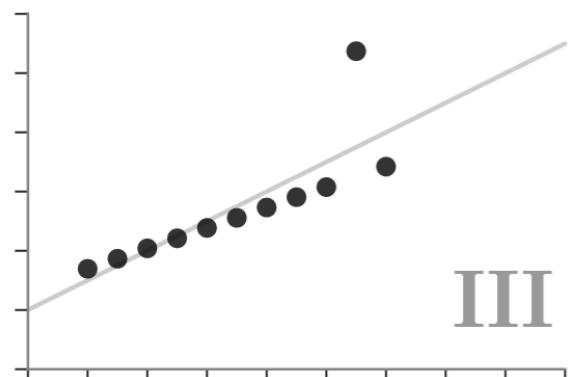
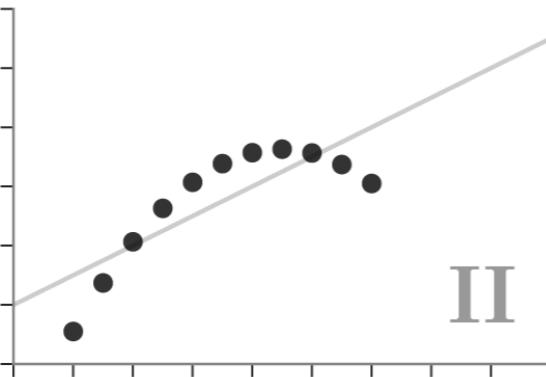
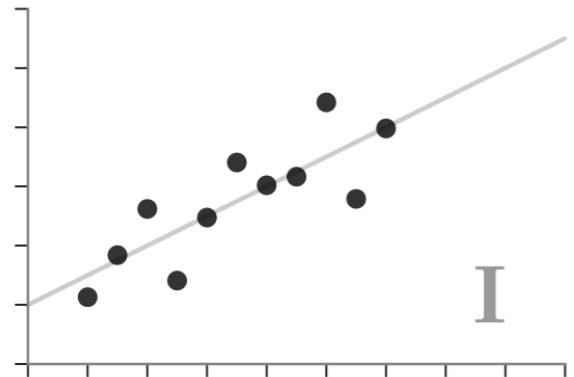
Each dataset here also has the same summary statistics. However, they are not *clearly different* or *visually distinct*.



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

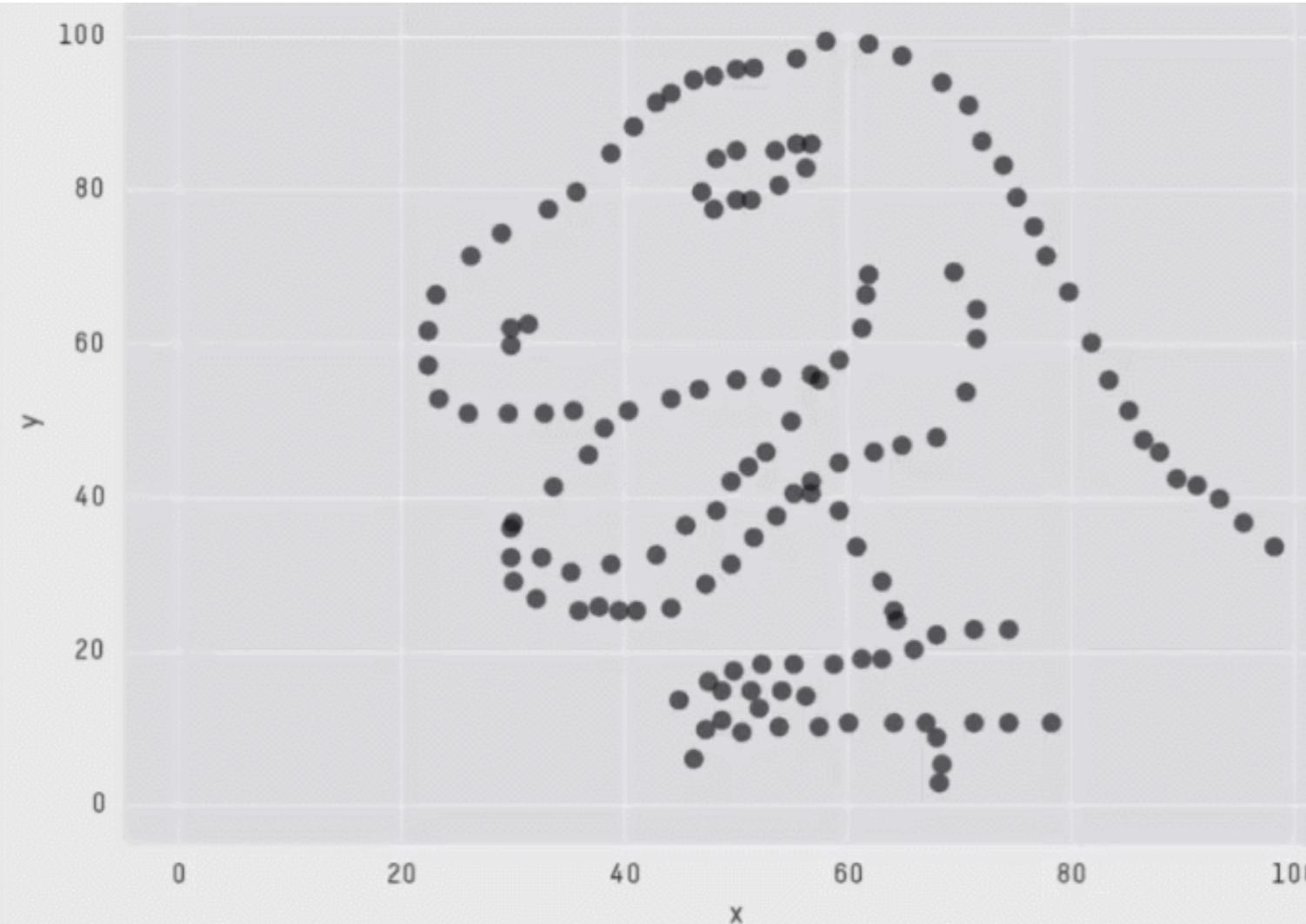
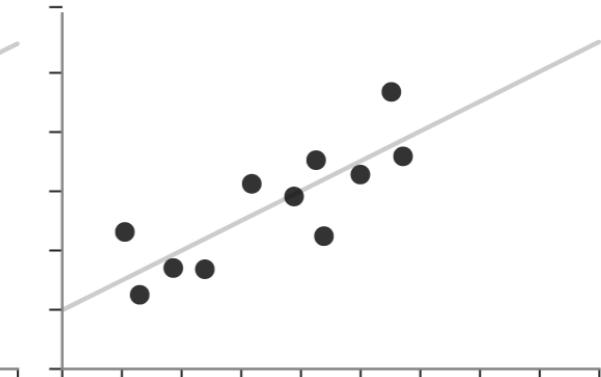
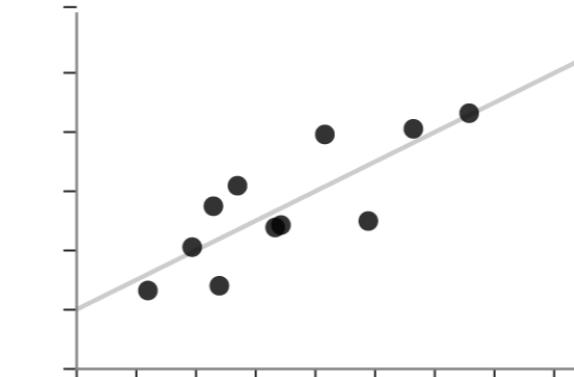
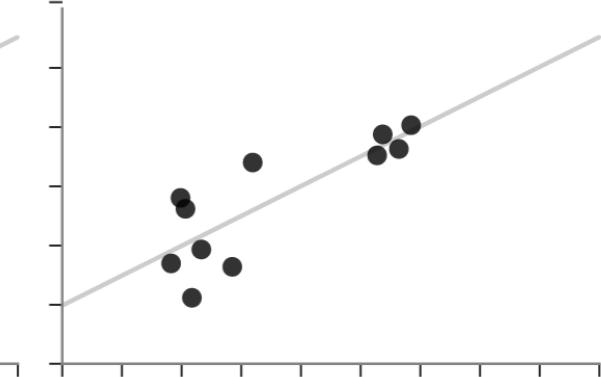
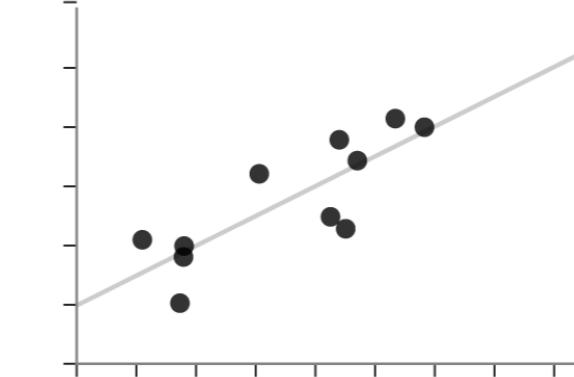
## ✓ Anscombe's Quartet

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 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

# 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 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

# Data

## Genus, Species

Table 1: Burtin's data.

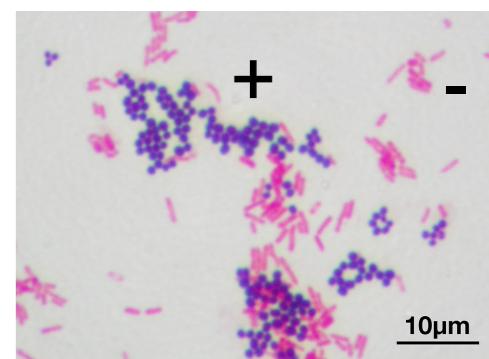
Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
<i>Aerobacter aerogenes</i>	870	1	1.6	negative
<i>Brucella abortus</i>	1	2	0.02	negative
<i>Brucella anthracis</i>	0.001	0.01	0.007	positive
<i>Diplococcus pneumoniae</i>	0.005	11	10	positive
<i>Escherichia coli</i>	100	0.4	0.1	negative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

# 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

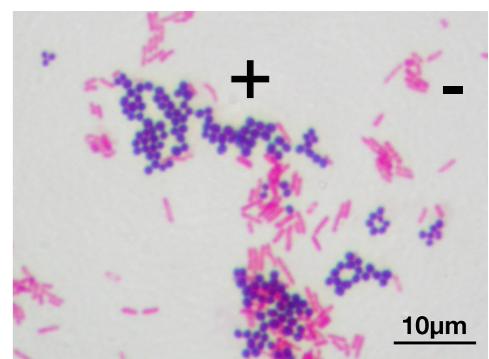


# Data

## 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 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

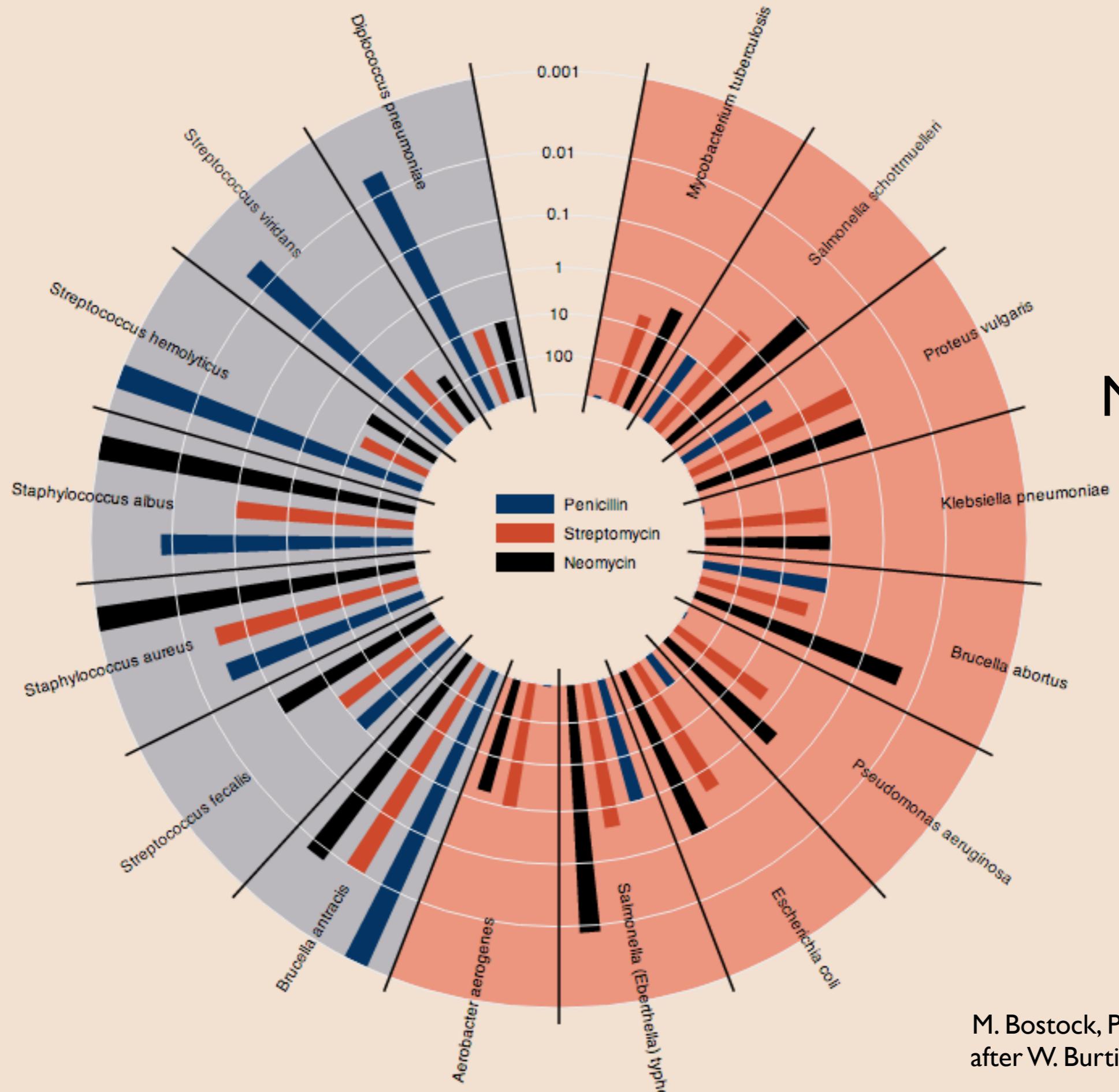


# What Questions?

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 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

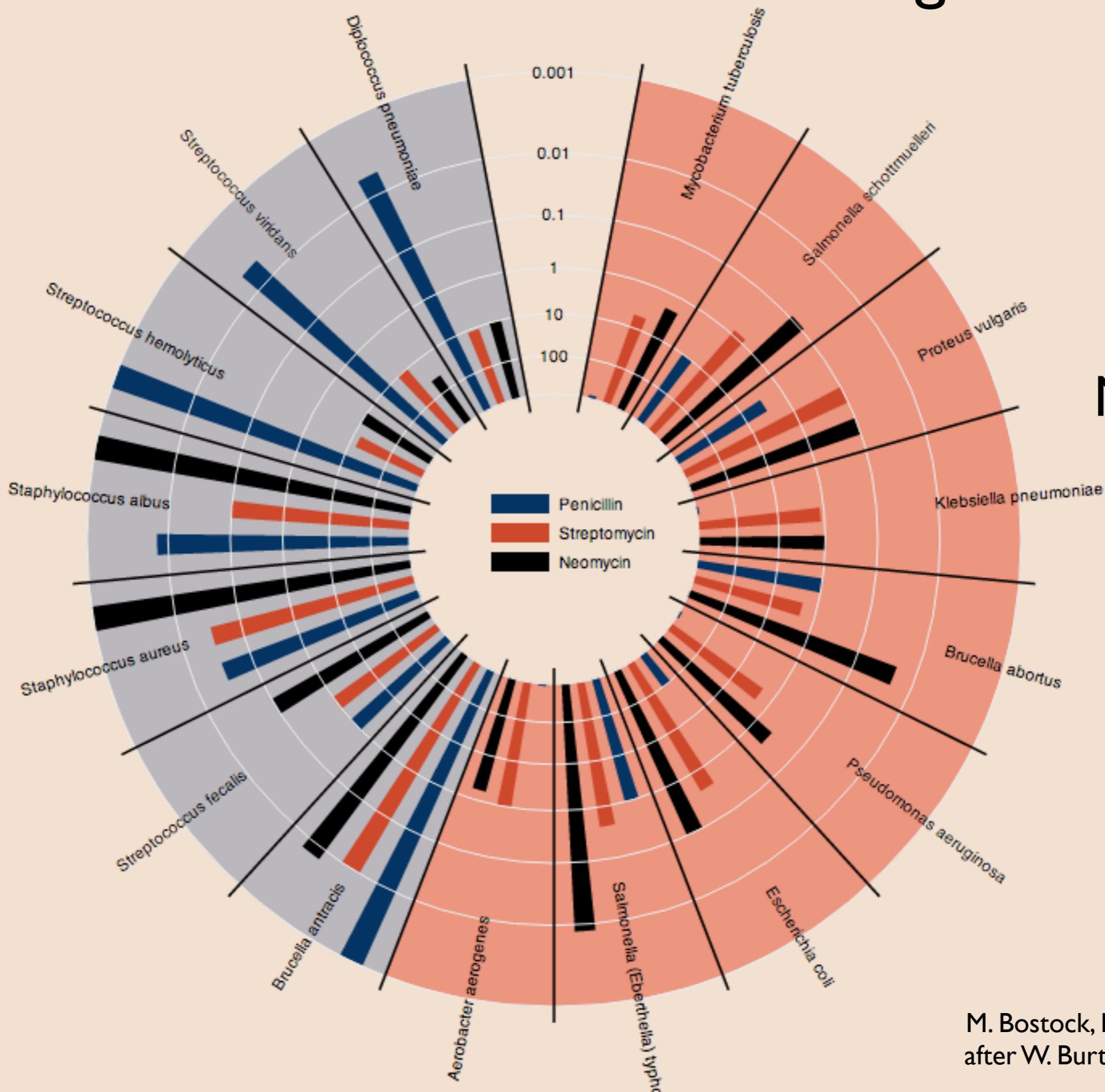


M. Bostock, Protopis  
after W. Burtin, 1951

# How effective are the drugs?

Gram  
Positive

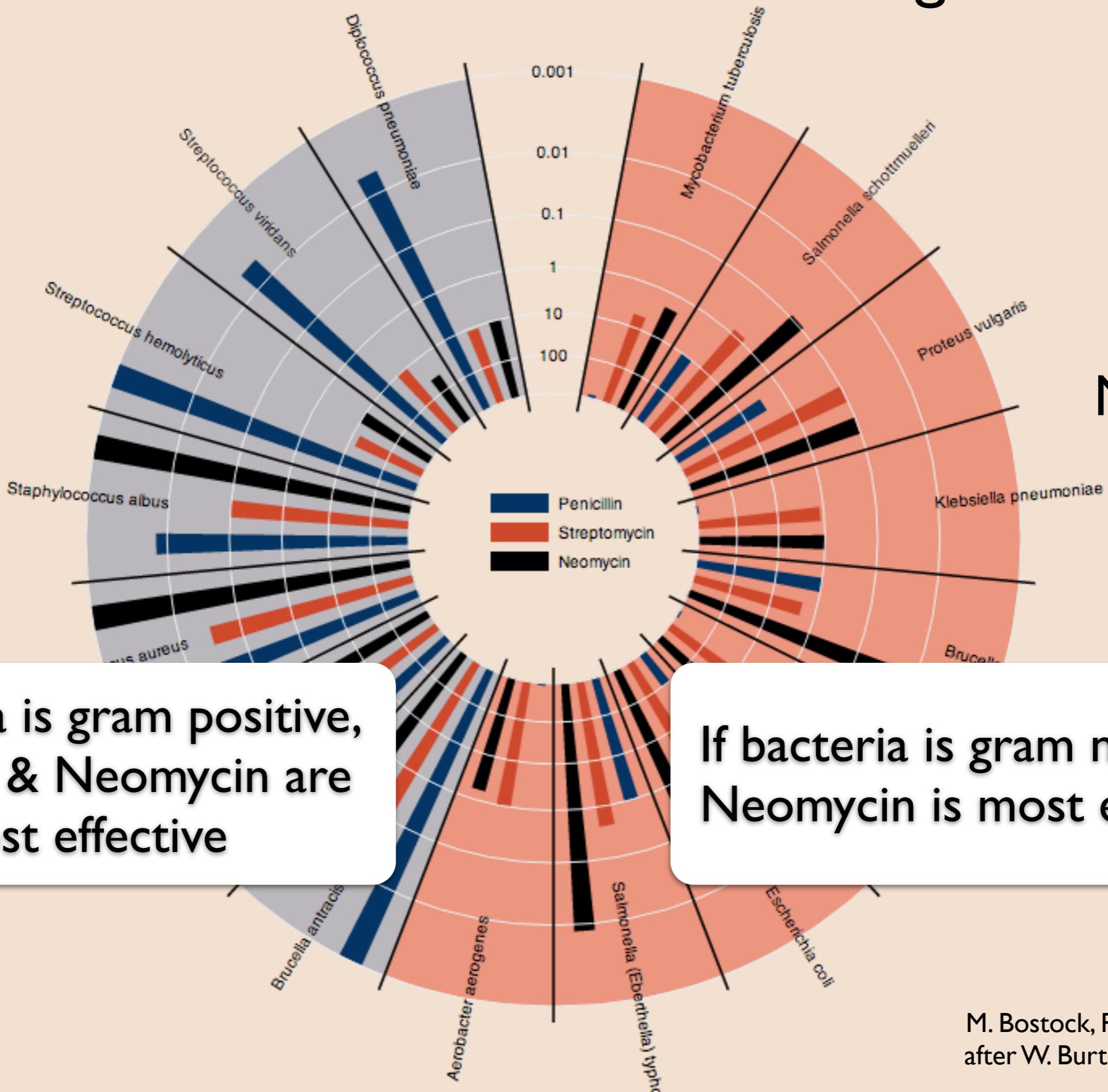
Gram  
Negative



# How effective are the drugs?

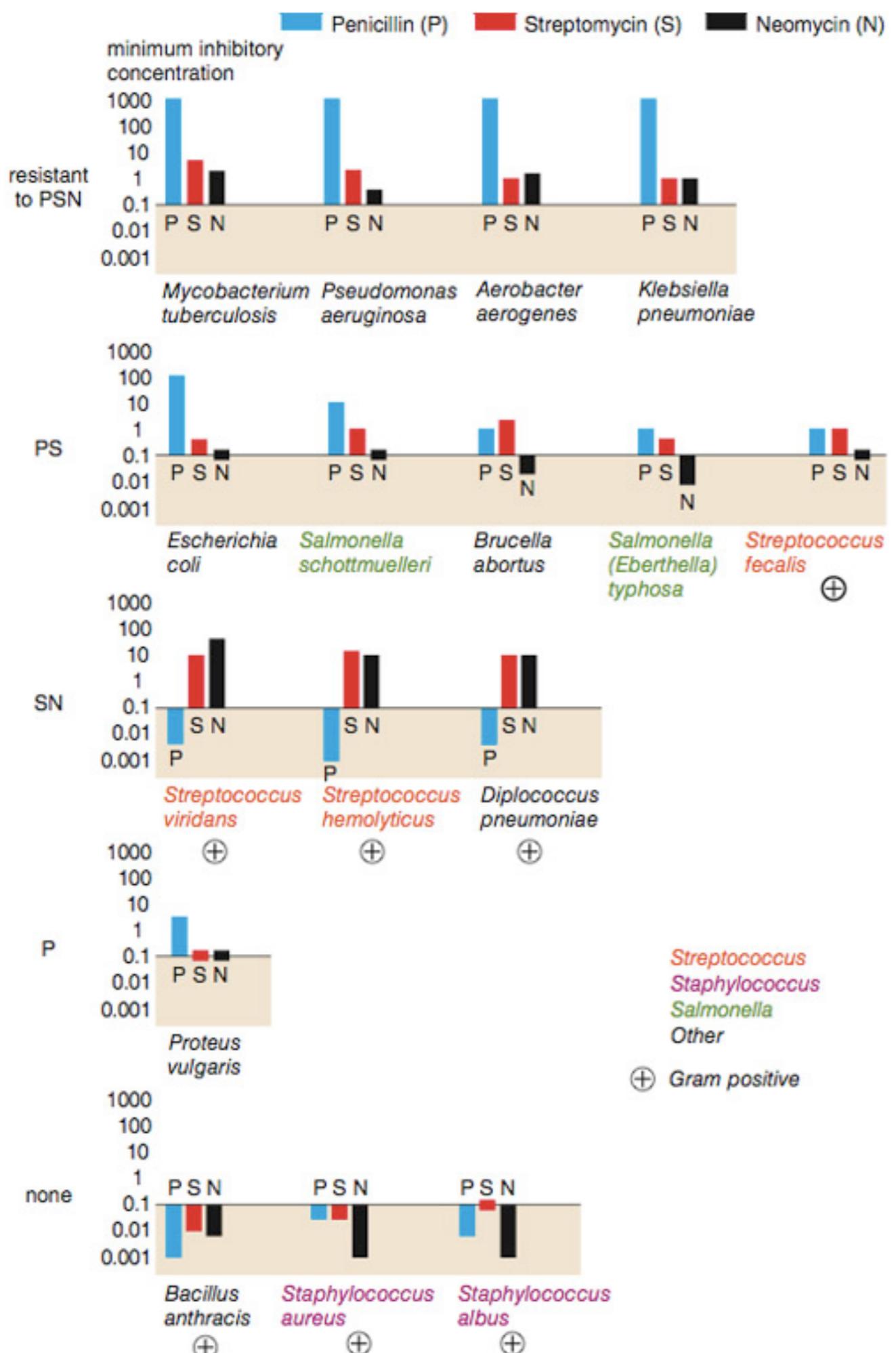
Gram  
Positive

Gram  
Negative

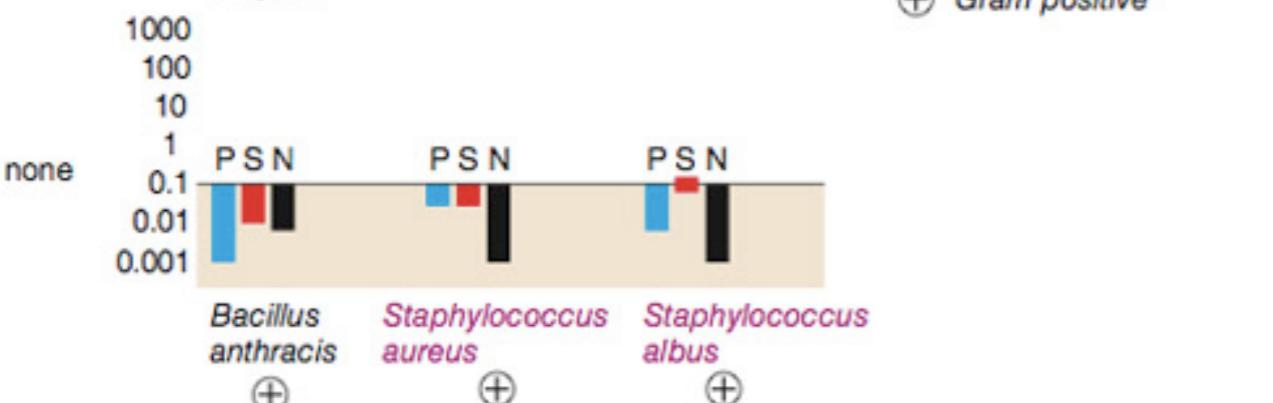
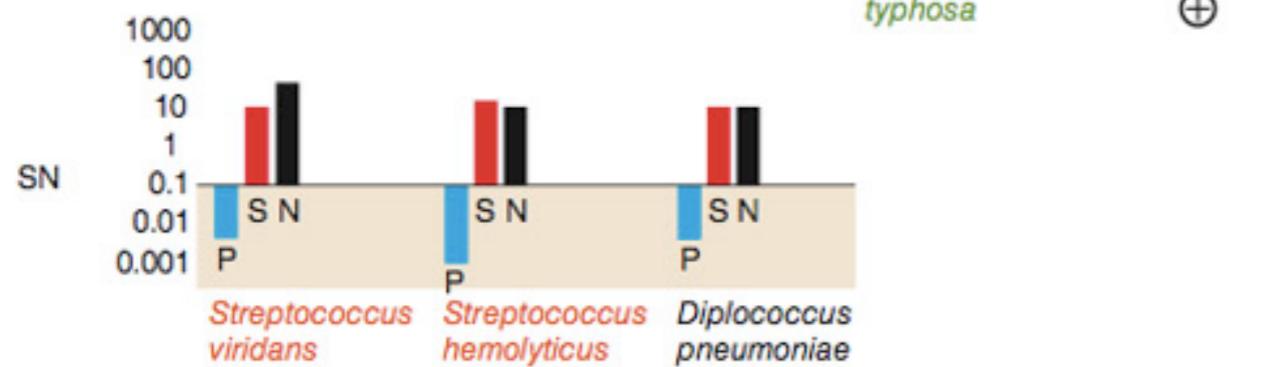
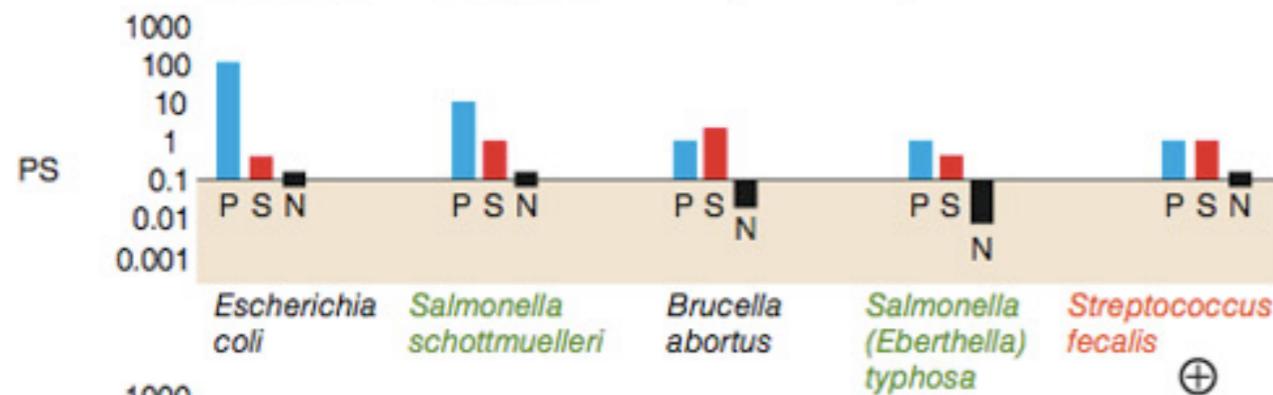
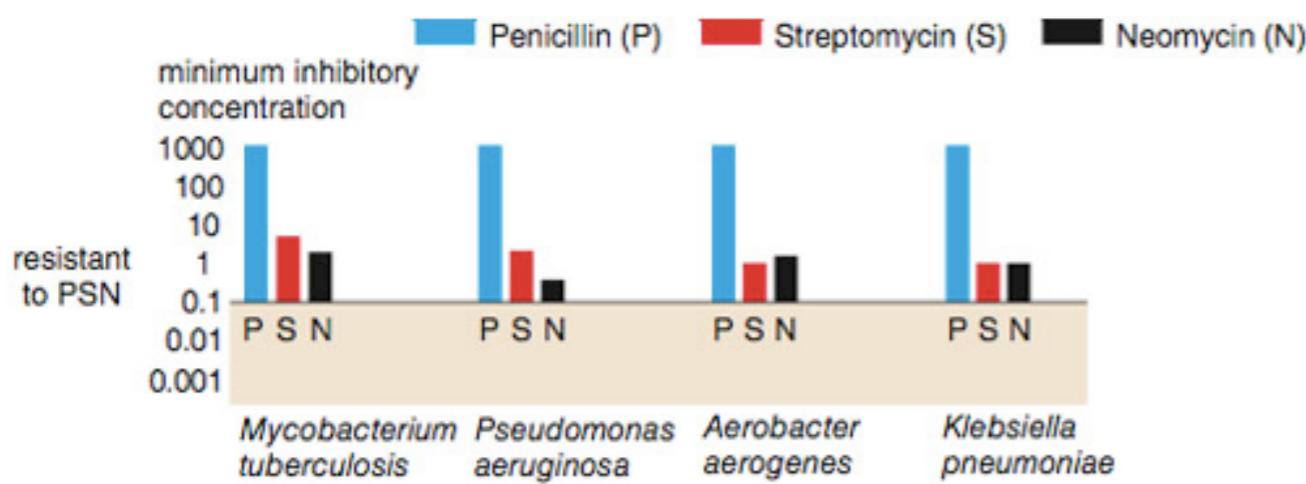


If bacteria is gram positive,  
Penicillin & Neomycin are  
most effective

If bacteria is gram negative,  
Neomycin is most effective

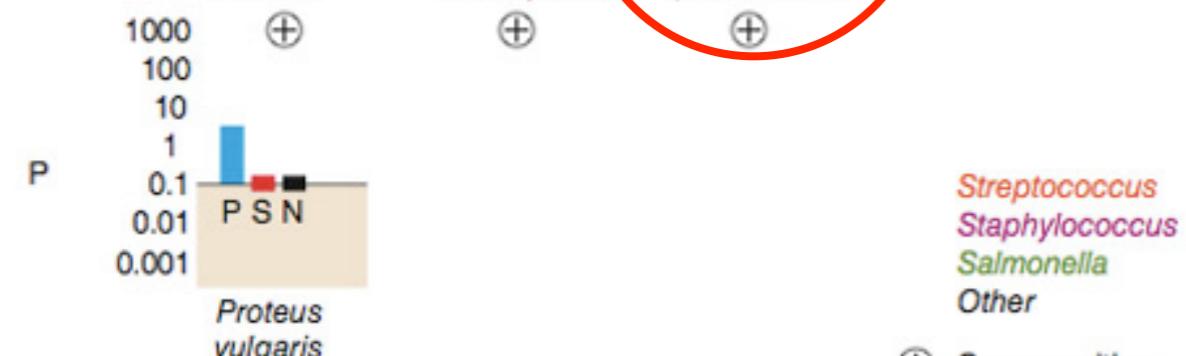
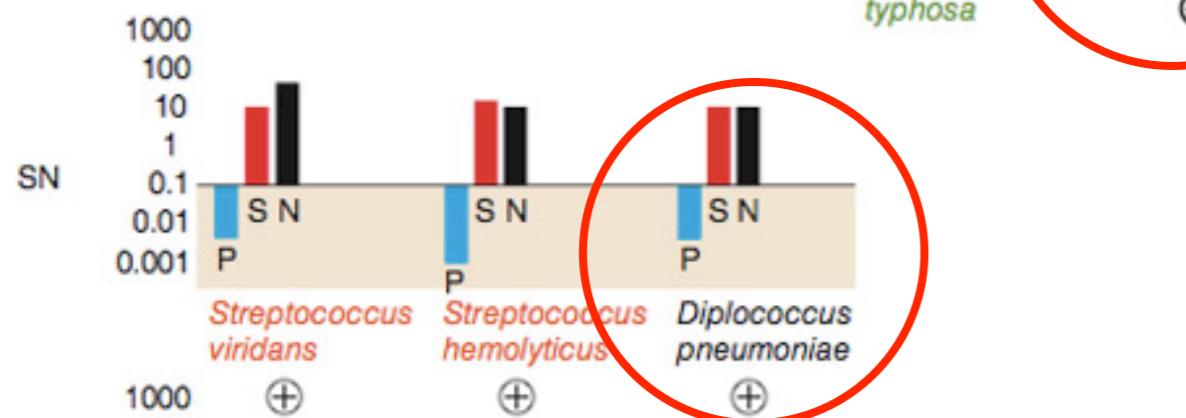
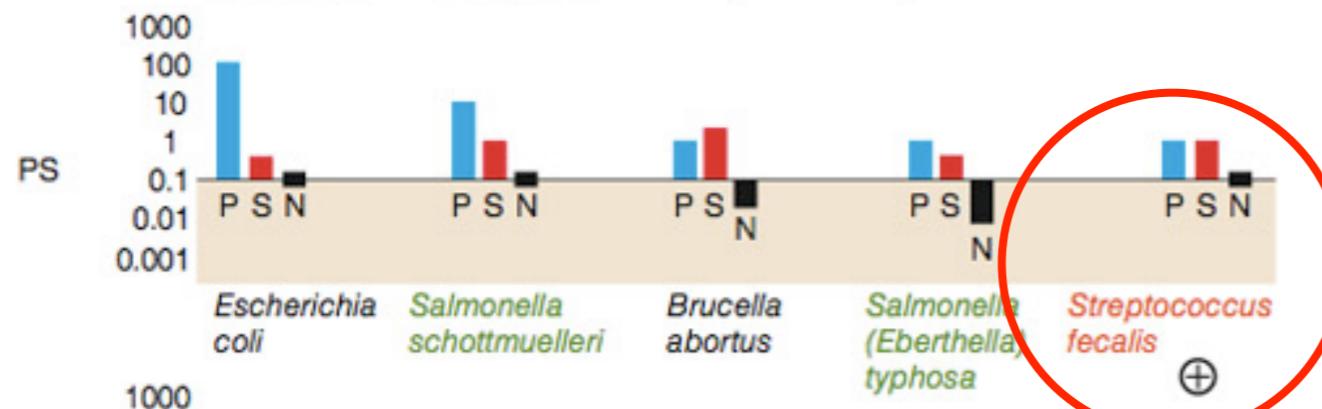
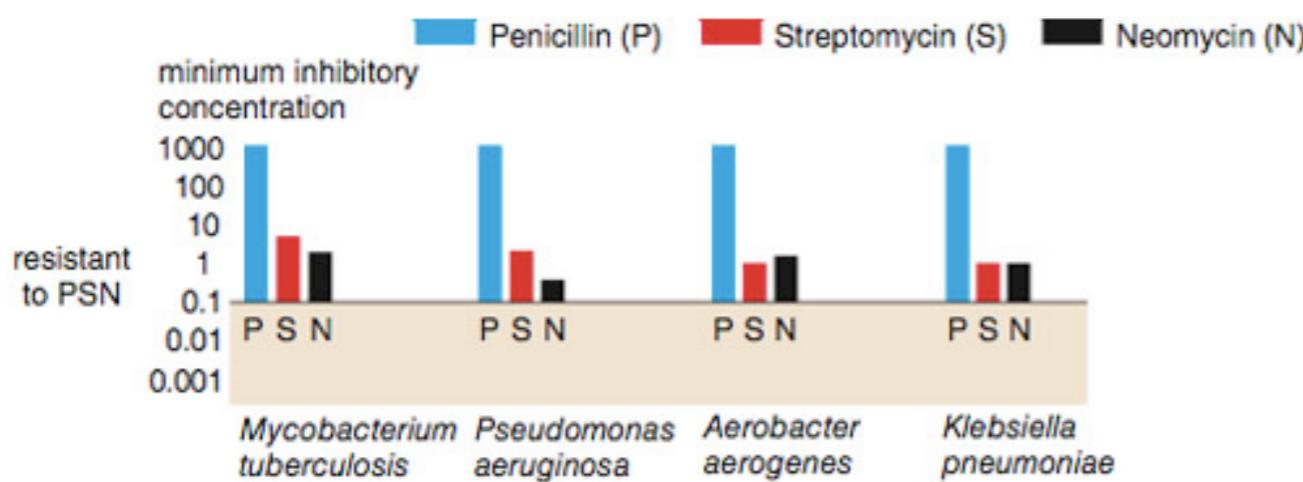


Wainer & Lysen, "That's funny..."  
American Scientist, 2009  
Adapted from Brian Schmotzer



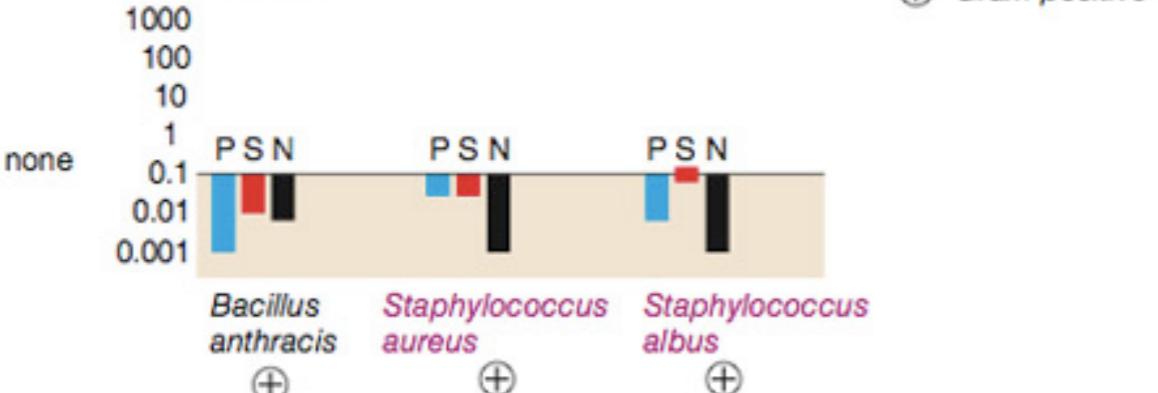
# How do the bacteria compare?

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



Streptococcus  
Staphylococcus  
Salmonella  
Other

(+) Gram positive



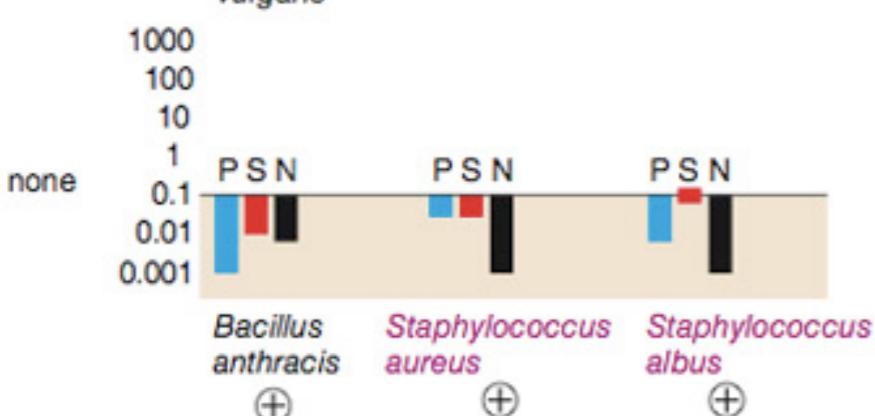
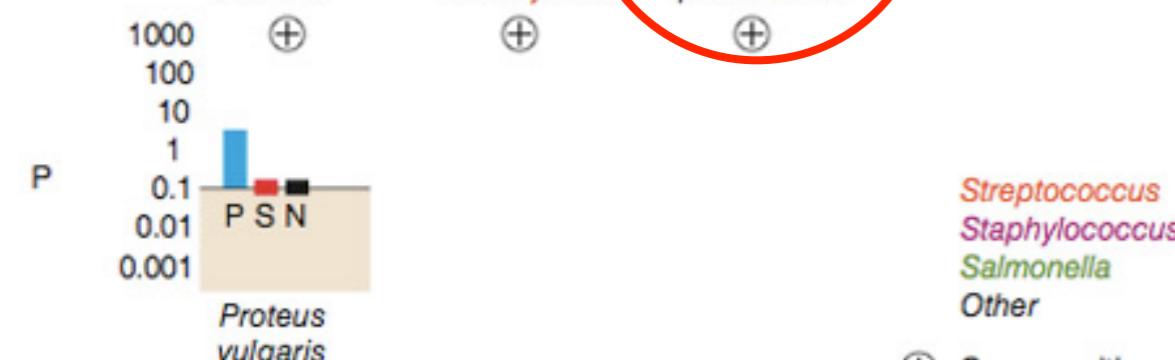
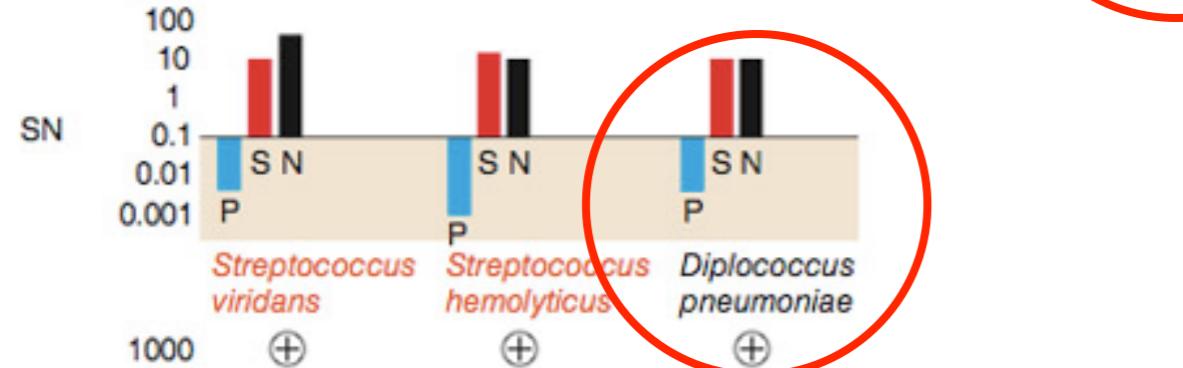
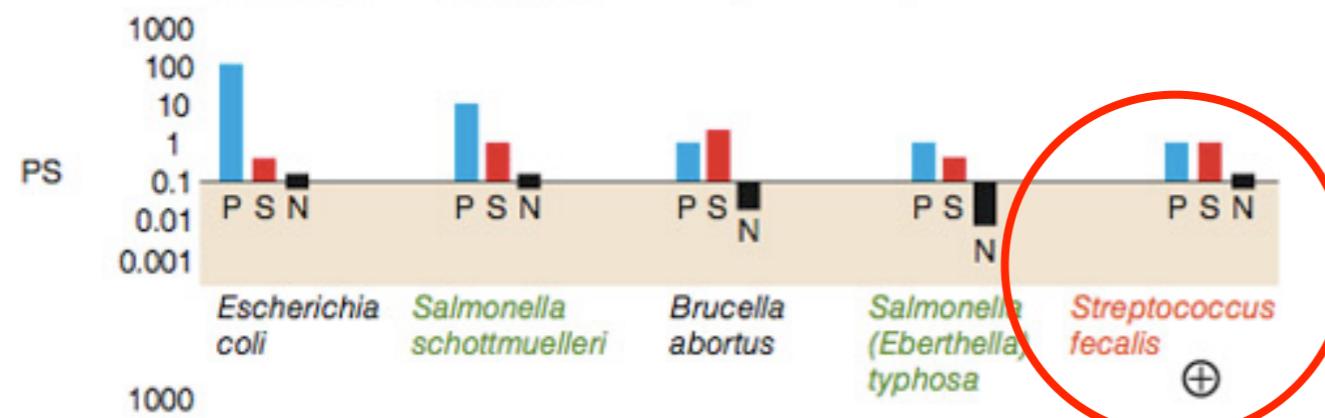
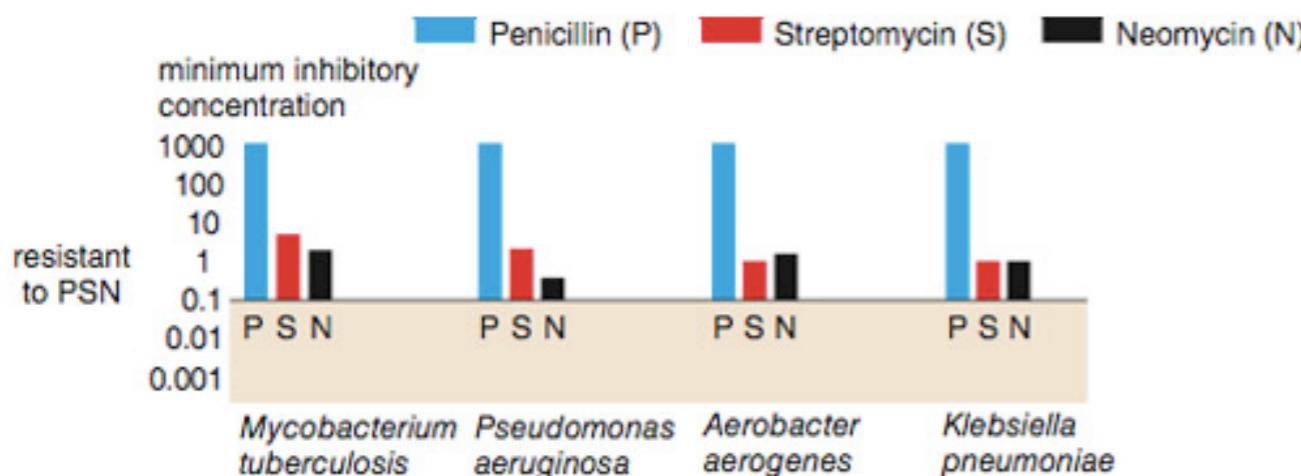
(+) Gram positive

# 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)

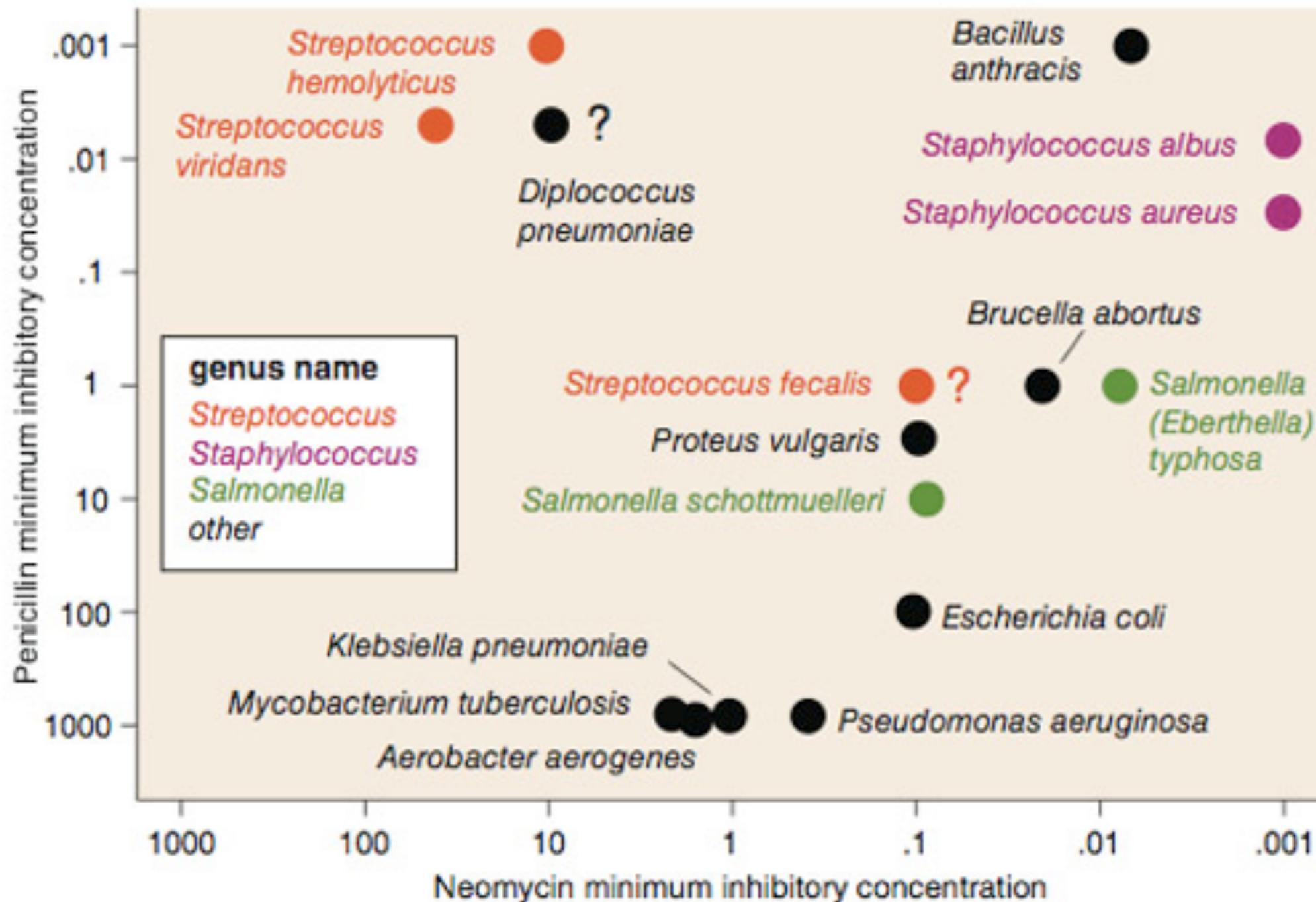
Really a streptococcus!  
(realized ~20 years later)

Streptococcus  
Staphylococcus  
Salmonella  
Other

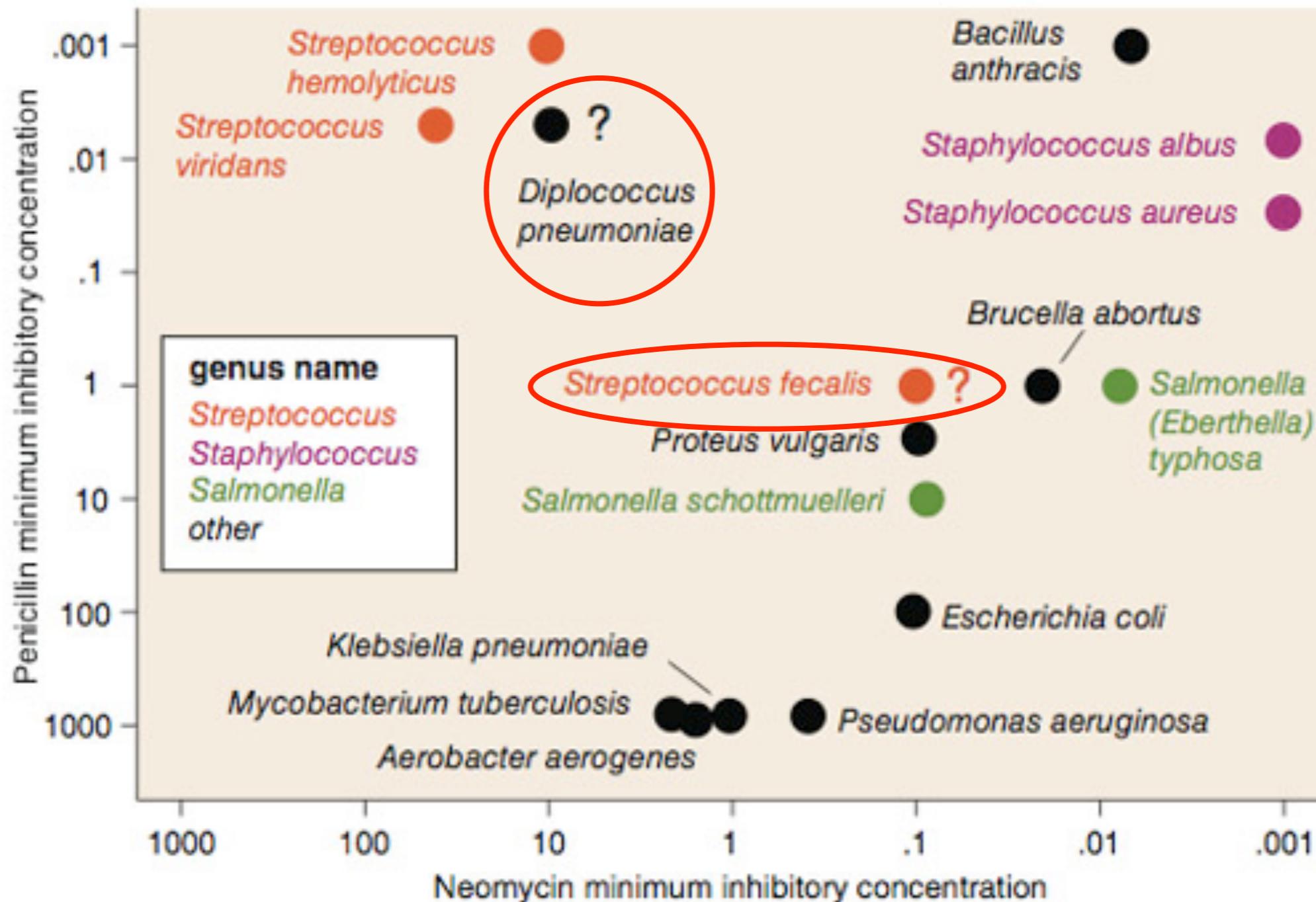
⊕ Gram positive

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

# 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

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

600

400

200

0

5 seasons

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

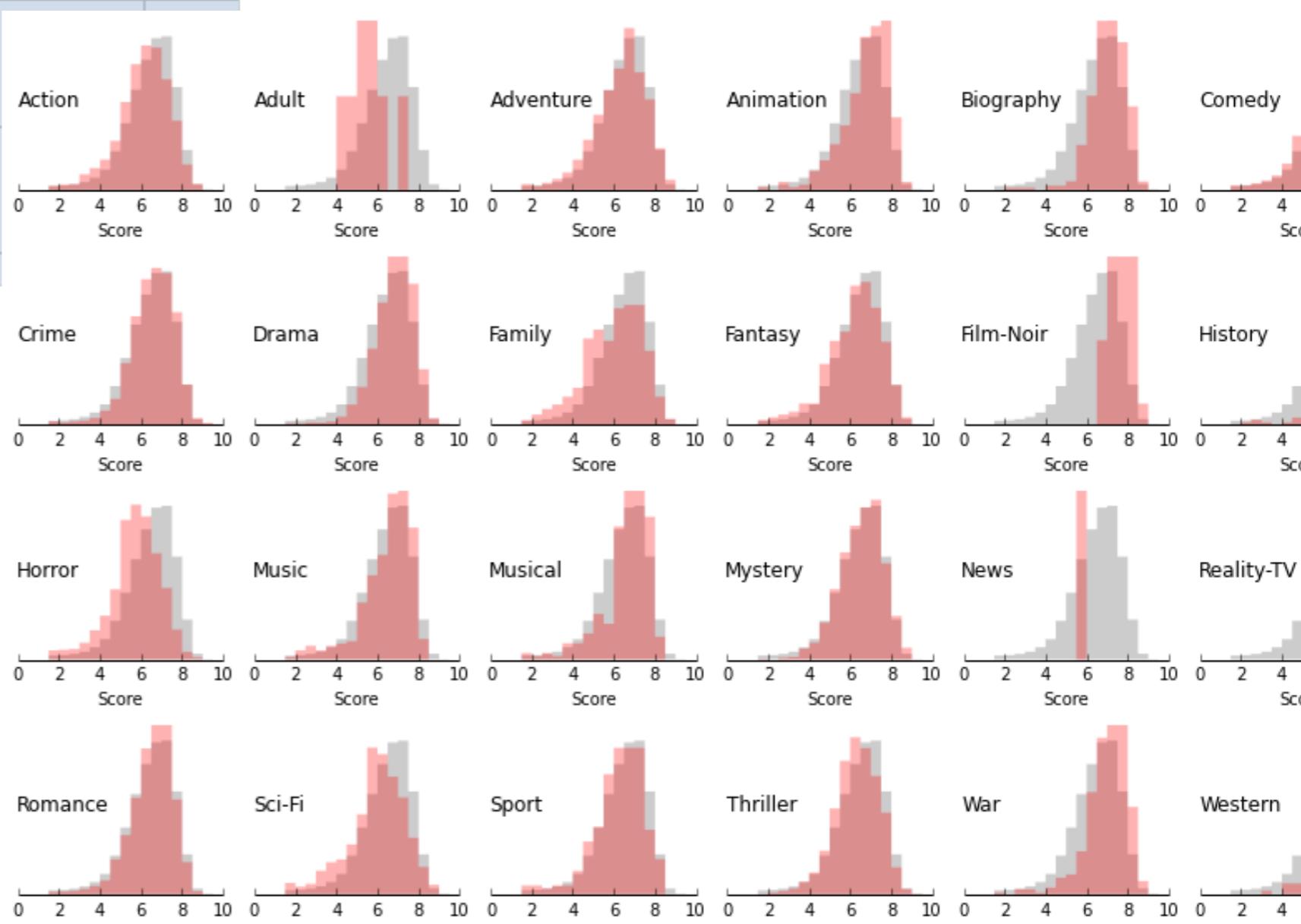
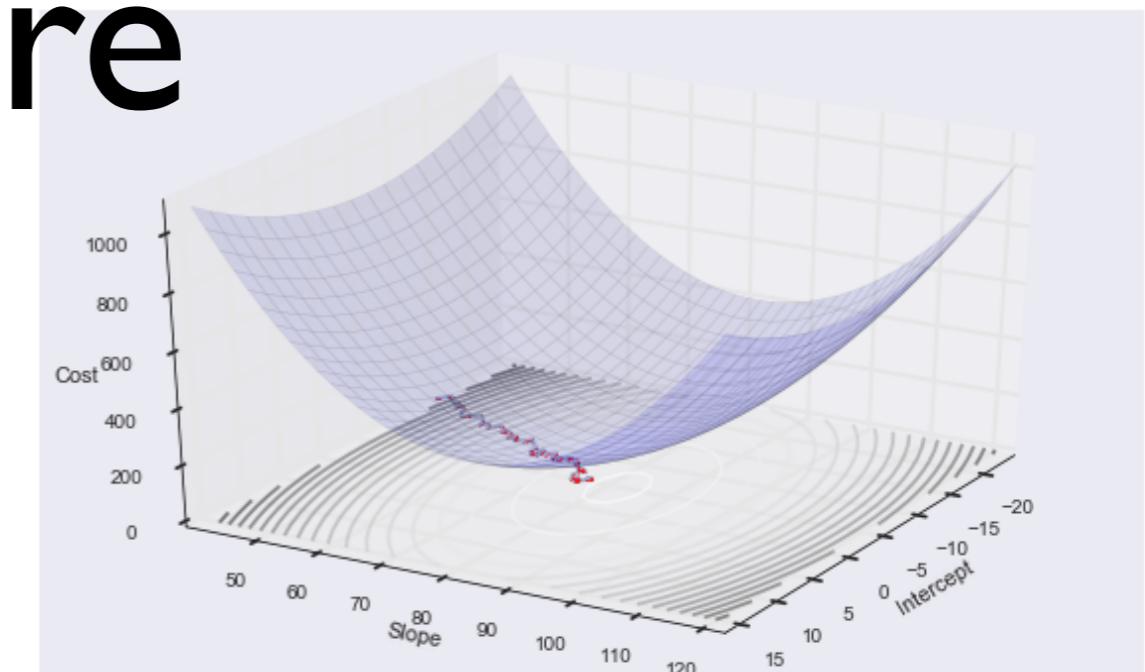
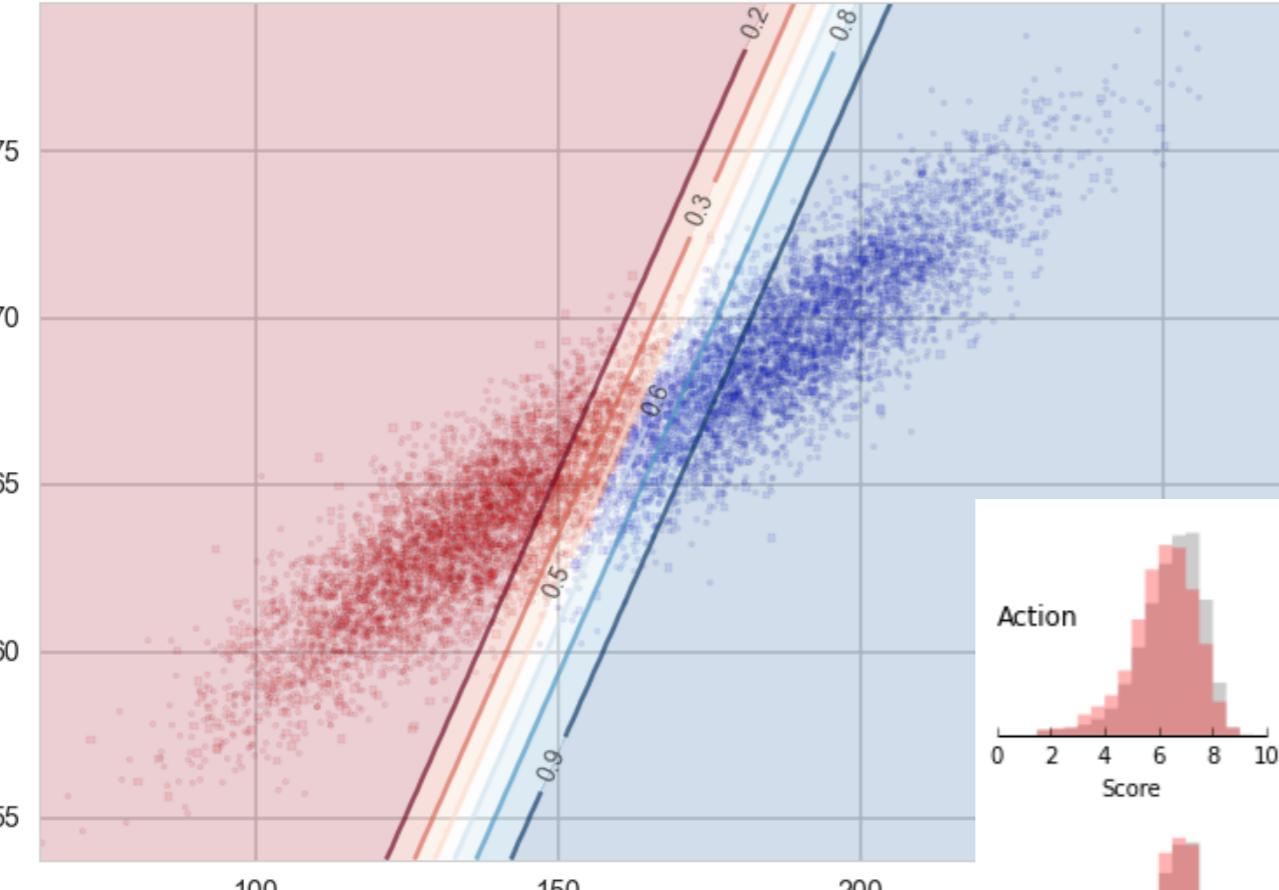
25

30

35

</div

# 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.

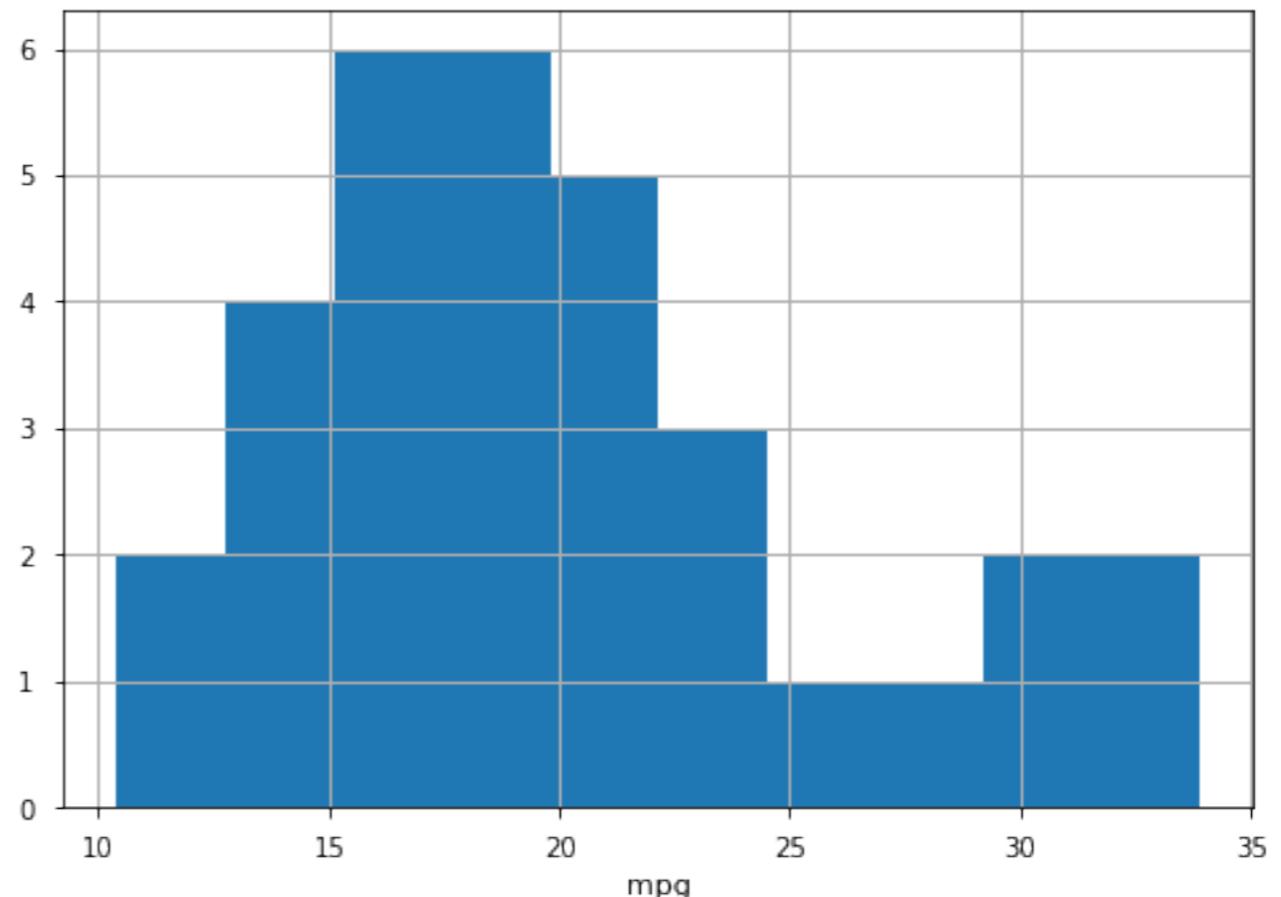
# Viz options

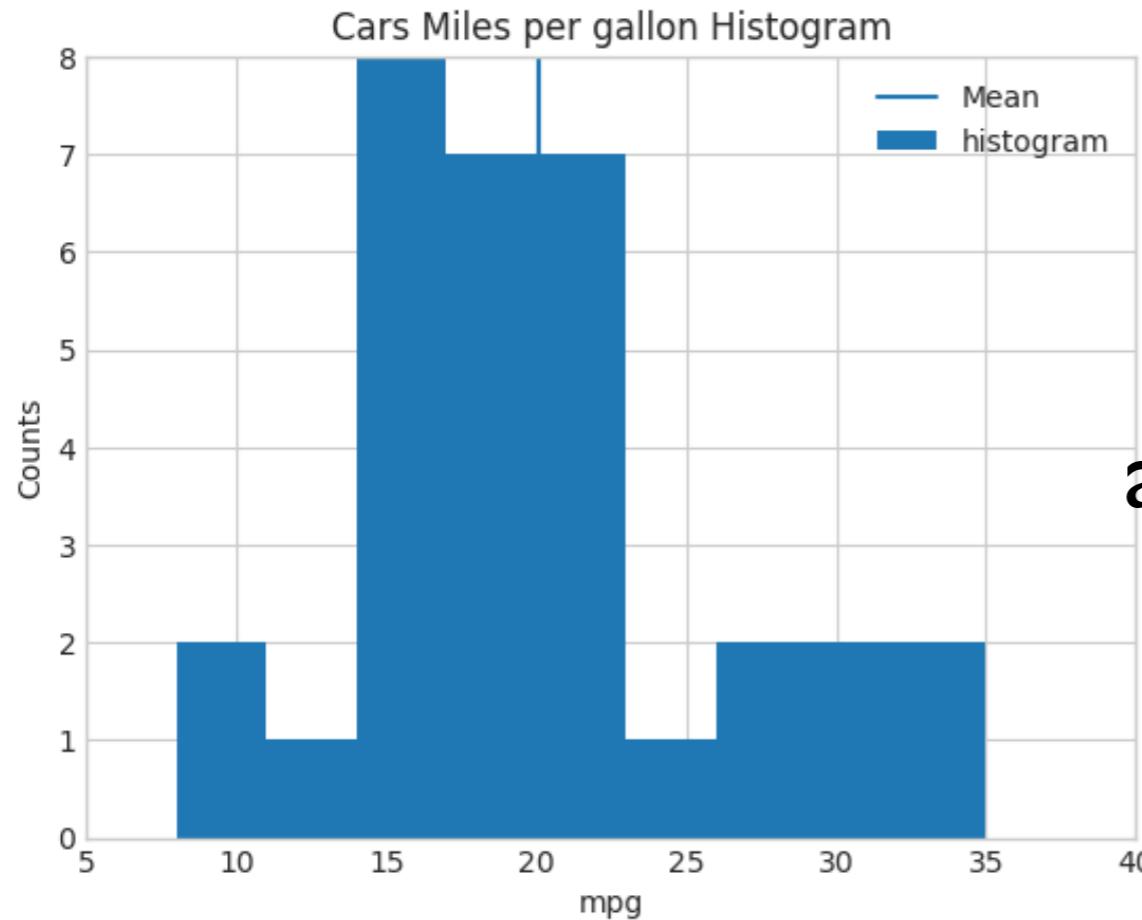
- Pandas Visualization module
- Matplotlib
- Seaborn
- Above 3 are inter-mixable
- Be lazy (to an extent...)
- Other options: Bokeh, Vega, Vincent, Altair

# 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

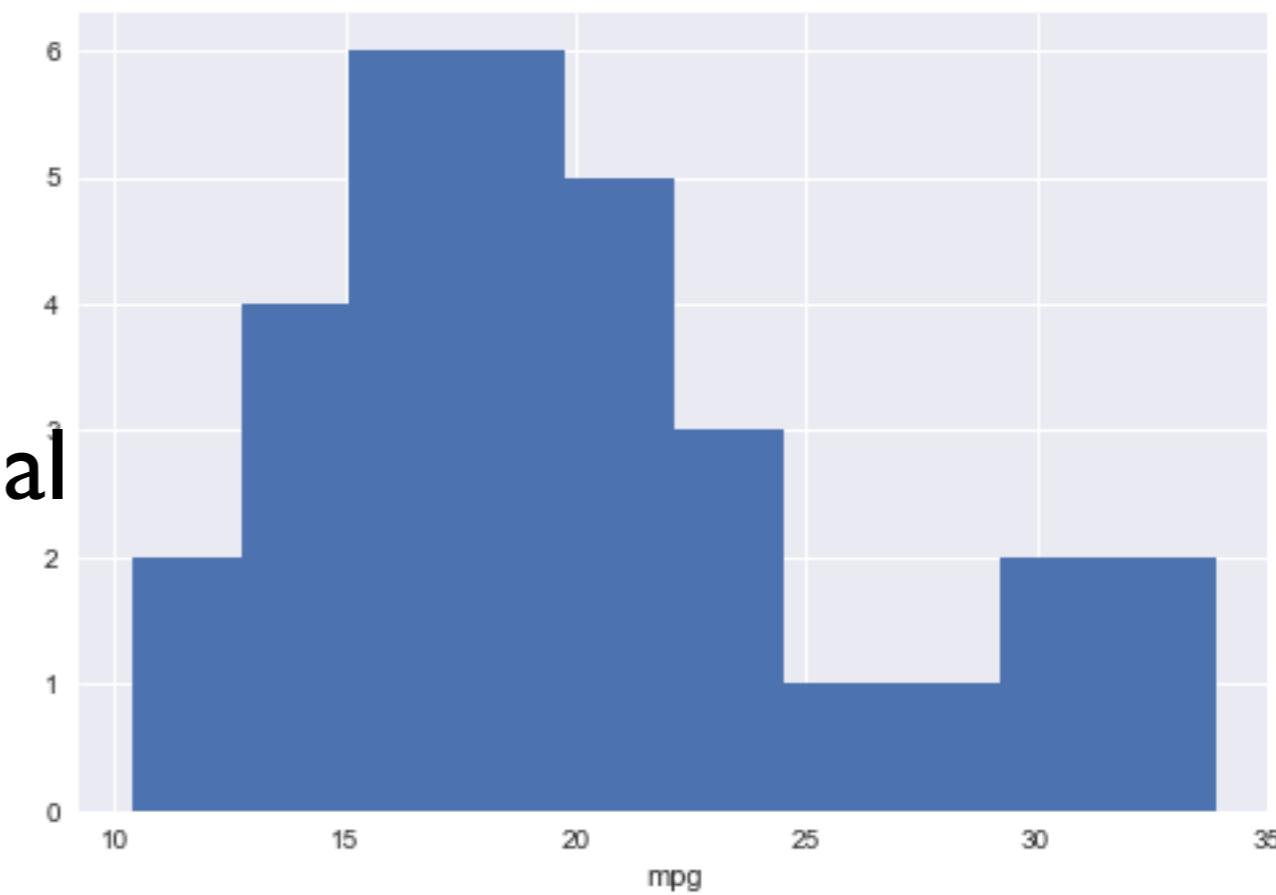
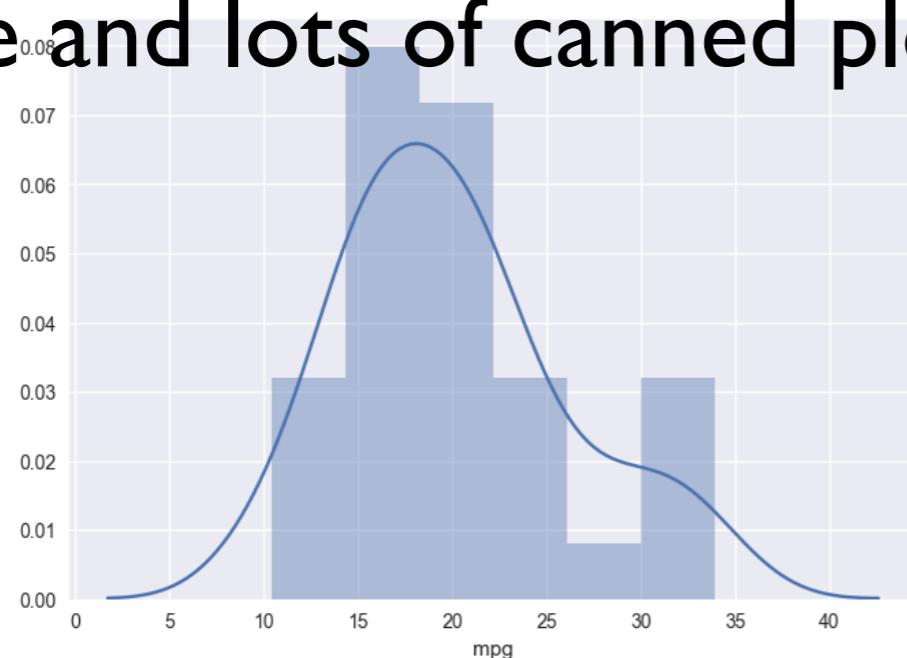
Basic Pandas/matplotlib





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...

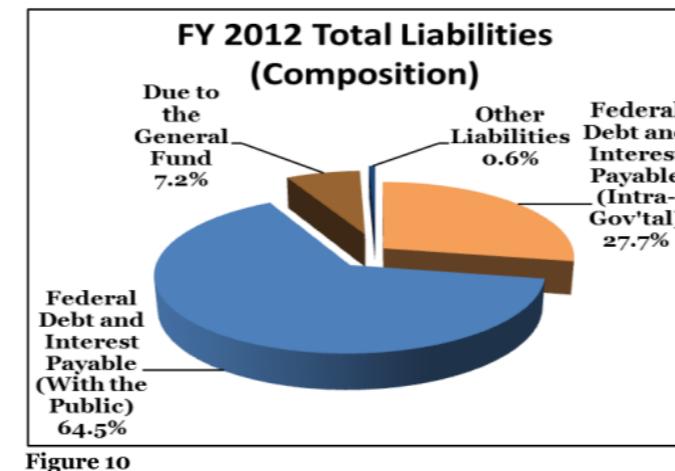
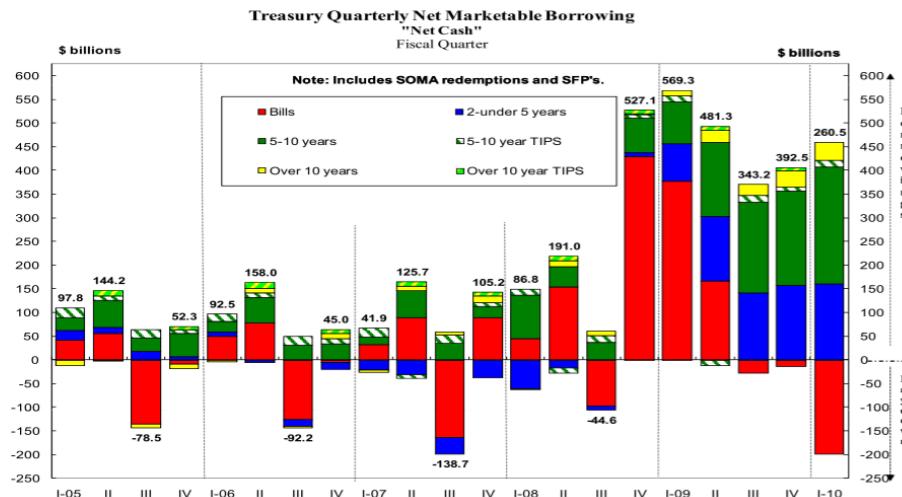


Figure 10

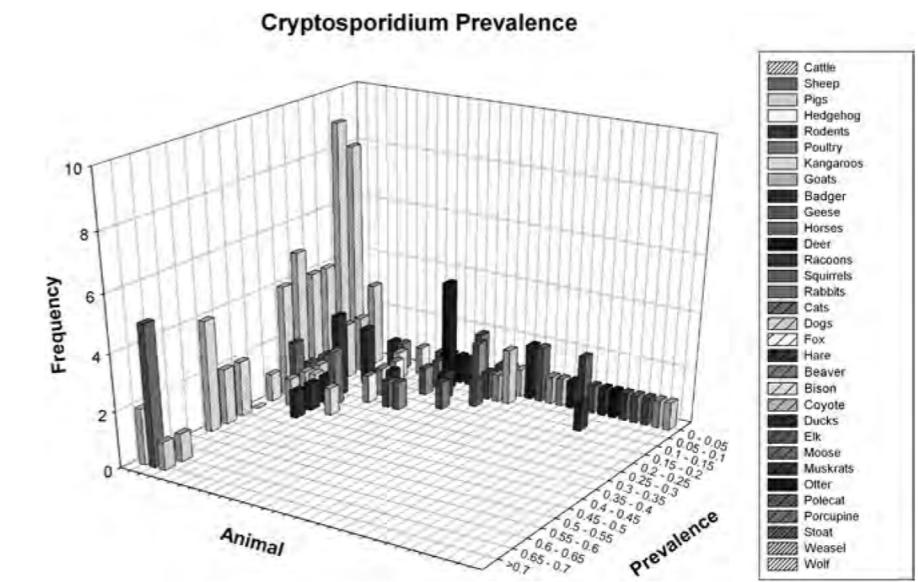
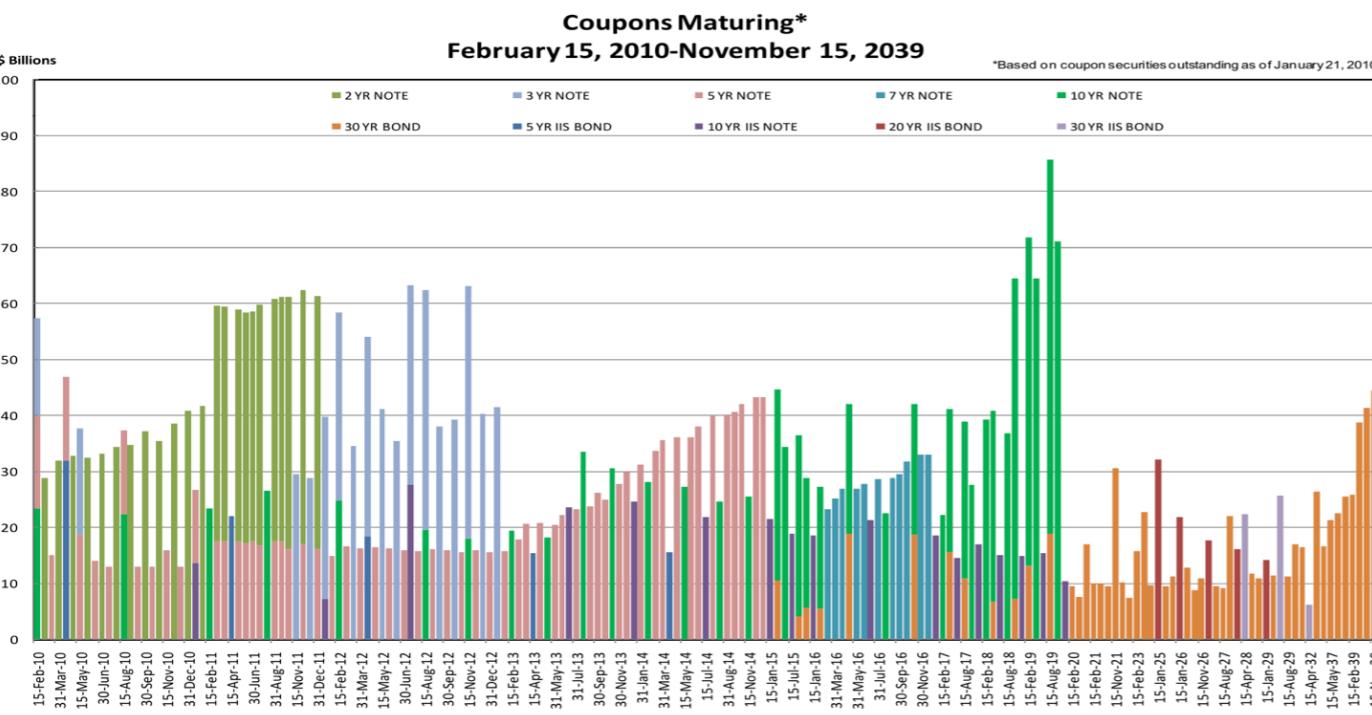
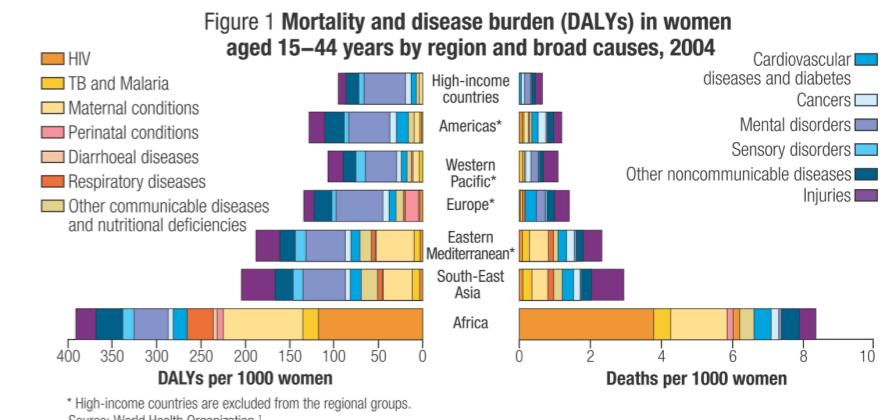


Figure 5.2 Mean prevalence rates of *Cryptosporidium* oocysts by animal species.

# Effective EDA Viz

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

# I. Graphical Integrity

*Same Veritas. More Lux.*

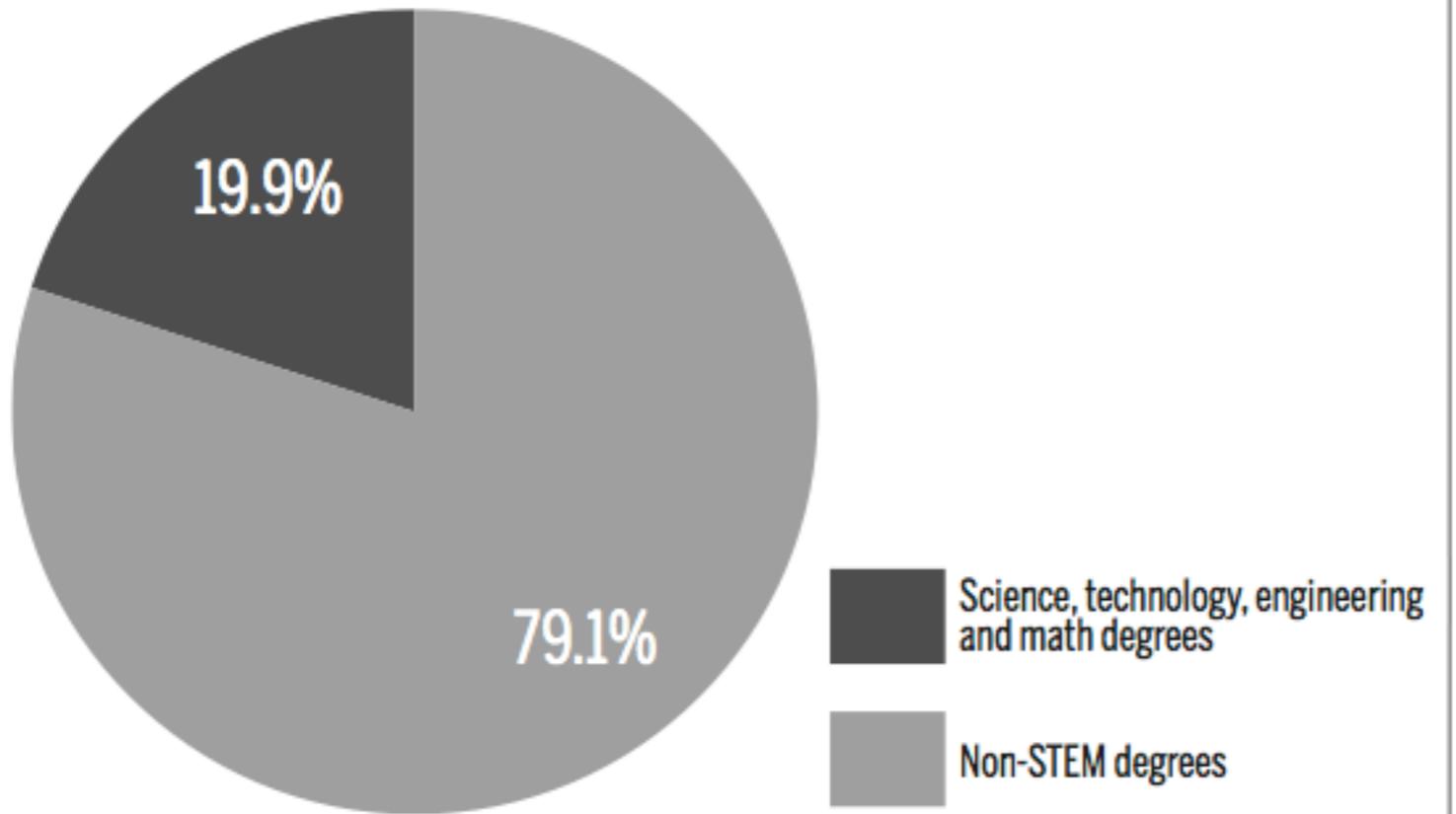
## Yale Summer Session

Over 200 full-credit courses.

June 4 – July 6 , July 9 – Aug 10

2012 *experience Yale*

### CHART YALE GRADUATES' MAJORS, CLASS OF 2011



### 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.

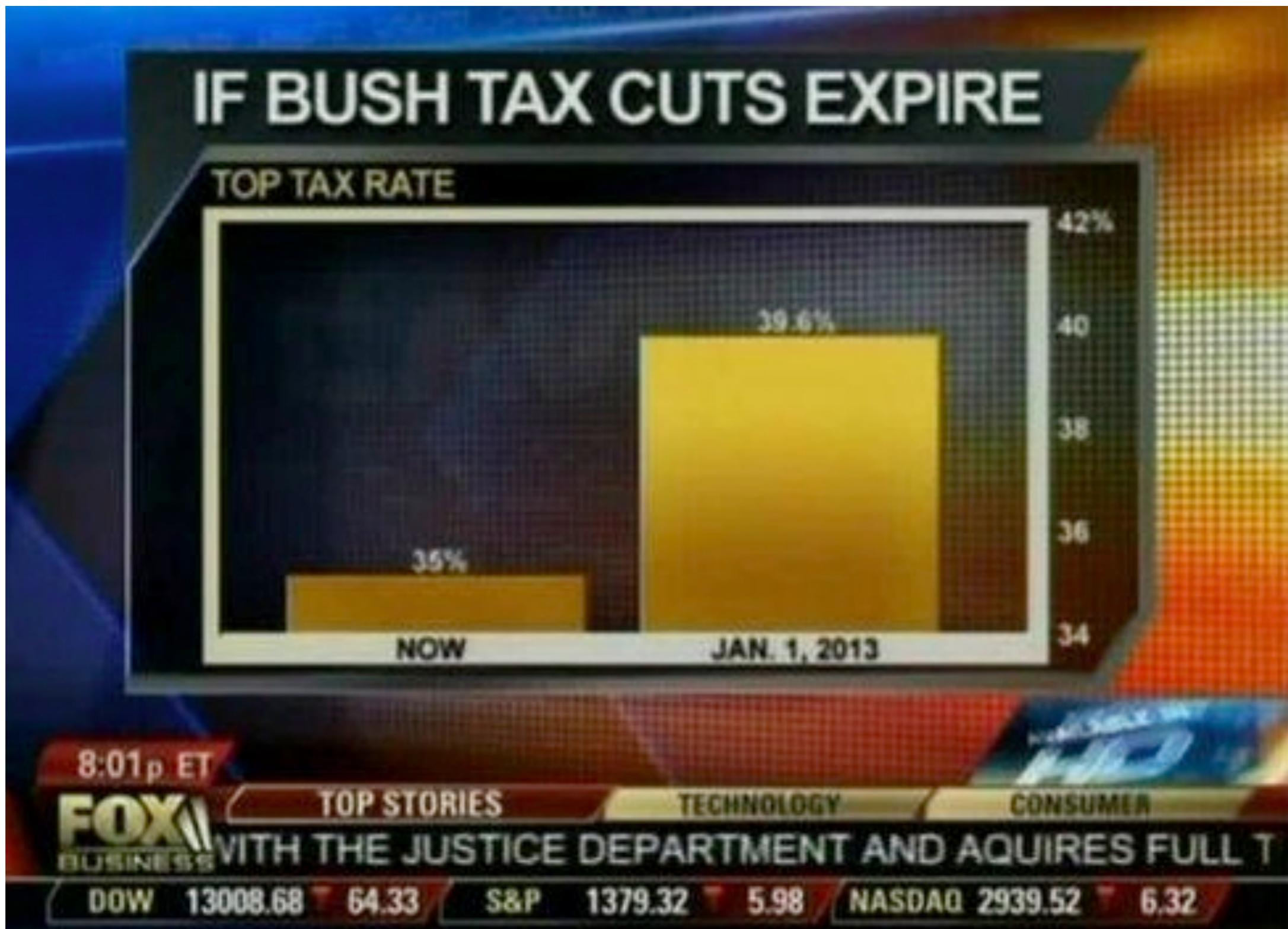


Facebook social plugin

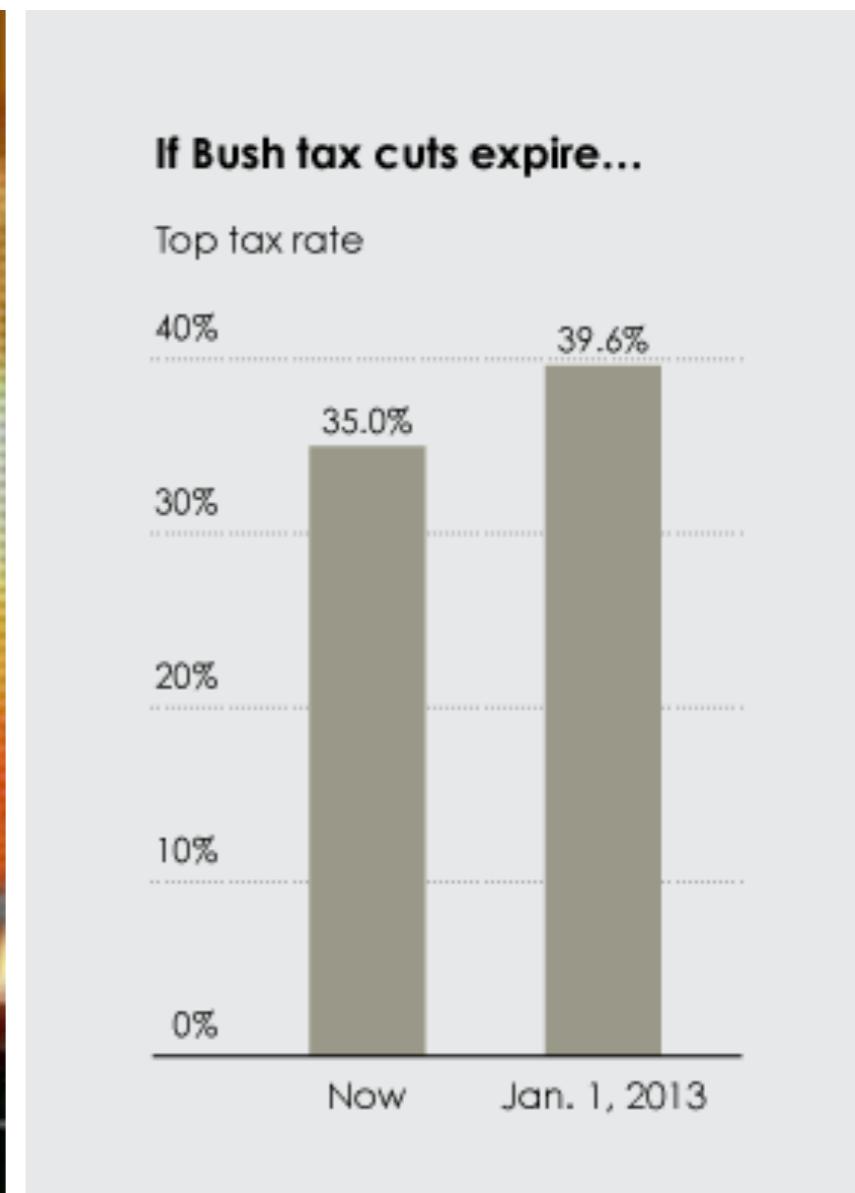
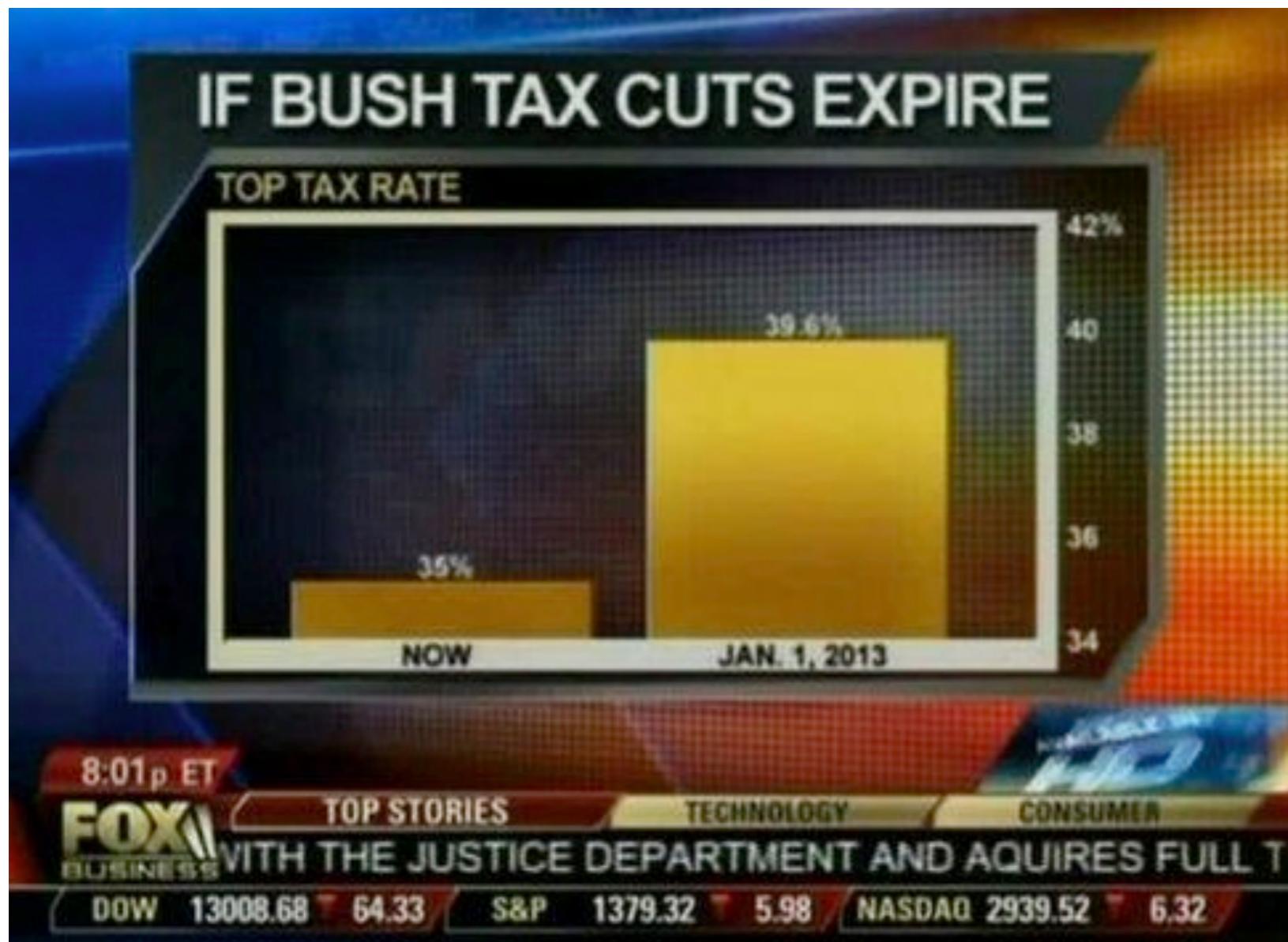
### Advertisement

Featured  
**Jobs**

# Graphical Integrity



# Scale Distortions



## JOB LOSS BY QUARTER



FOX NEWS .com

SOURCE: BLS

AMERICA'S  
NEWSROOM

# Scale Distortions



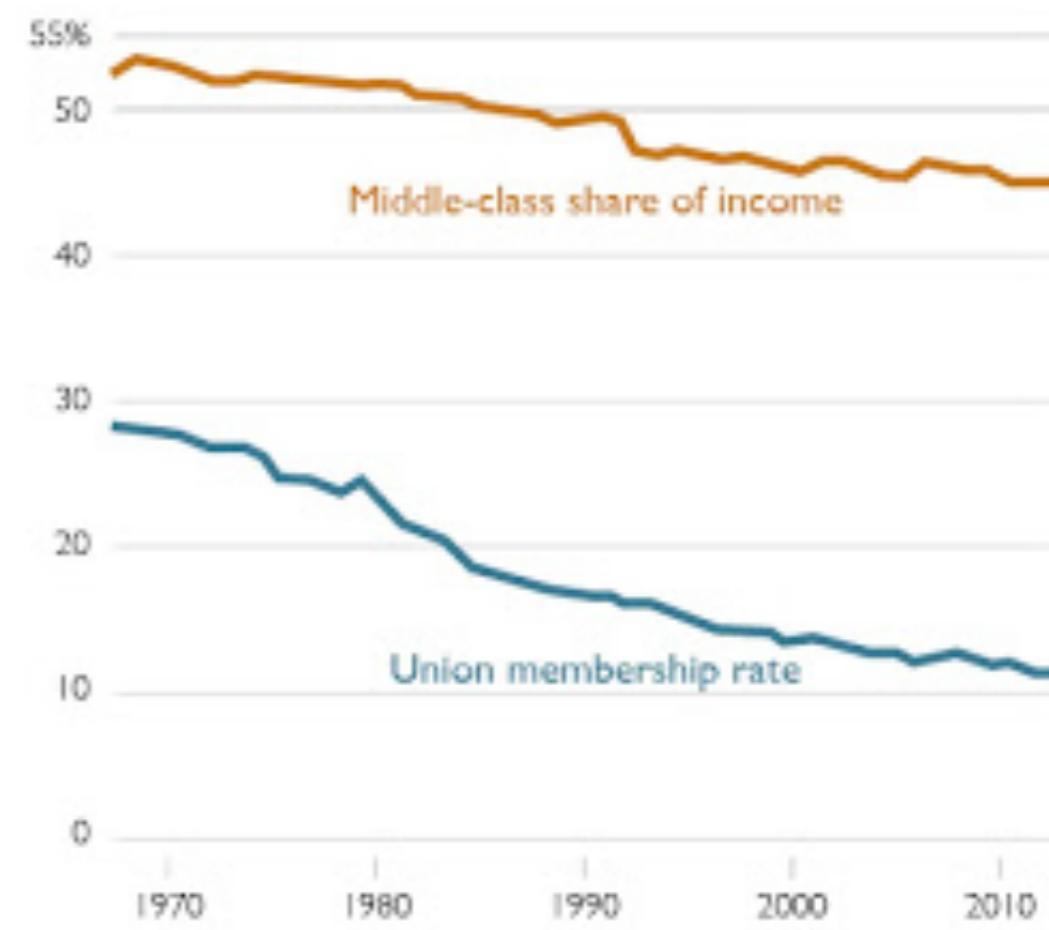
“Double the axes, double the mischief”

(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*

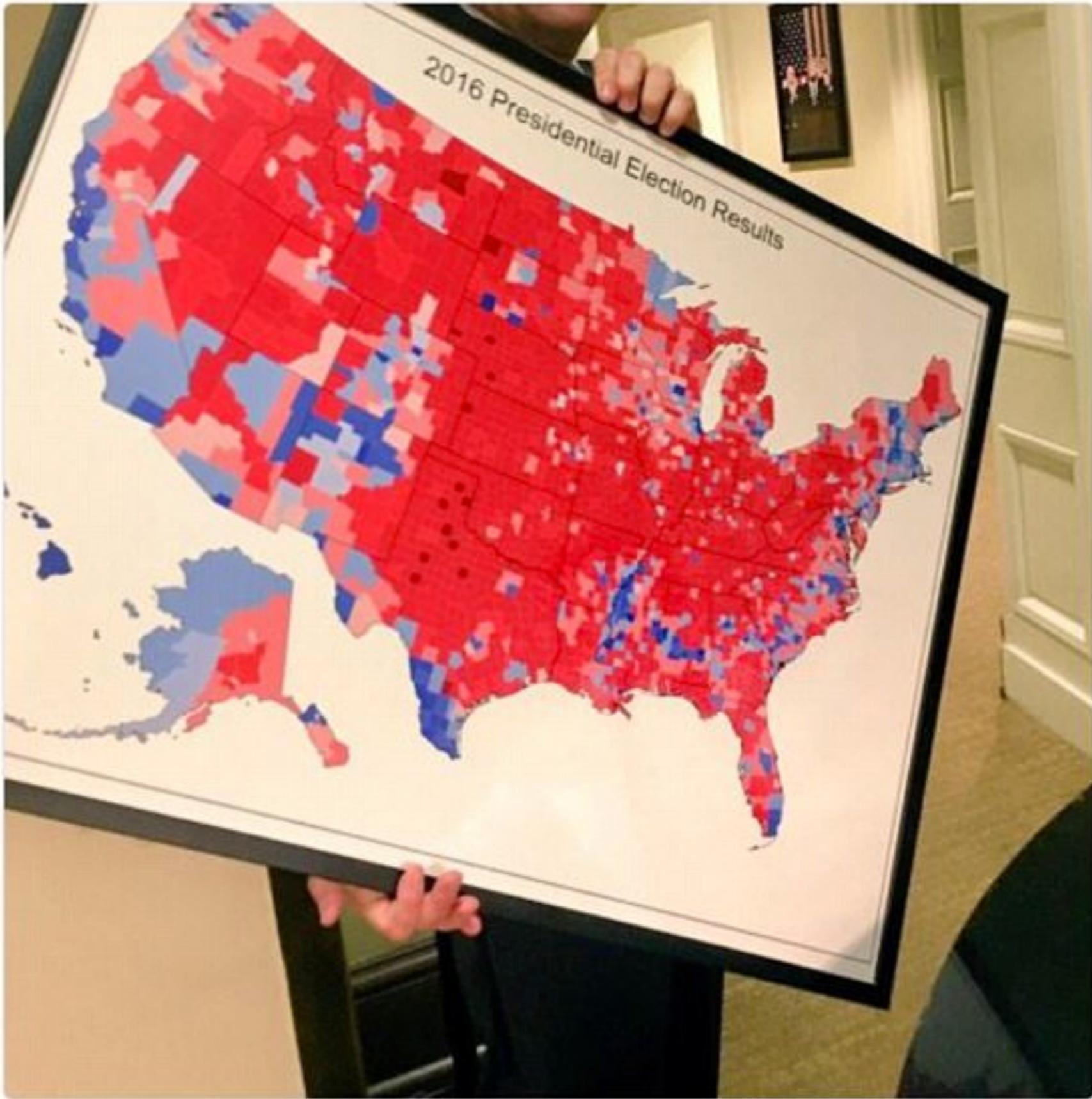
<http://www.thefunctionalart.com/2015/10/double-axes-double-mischief.html>

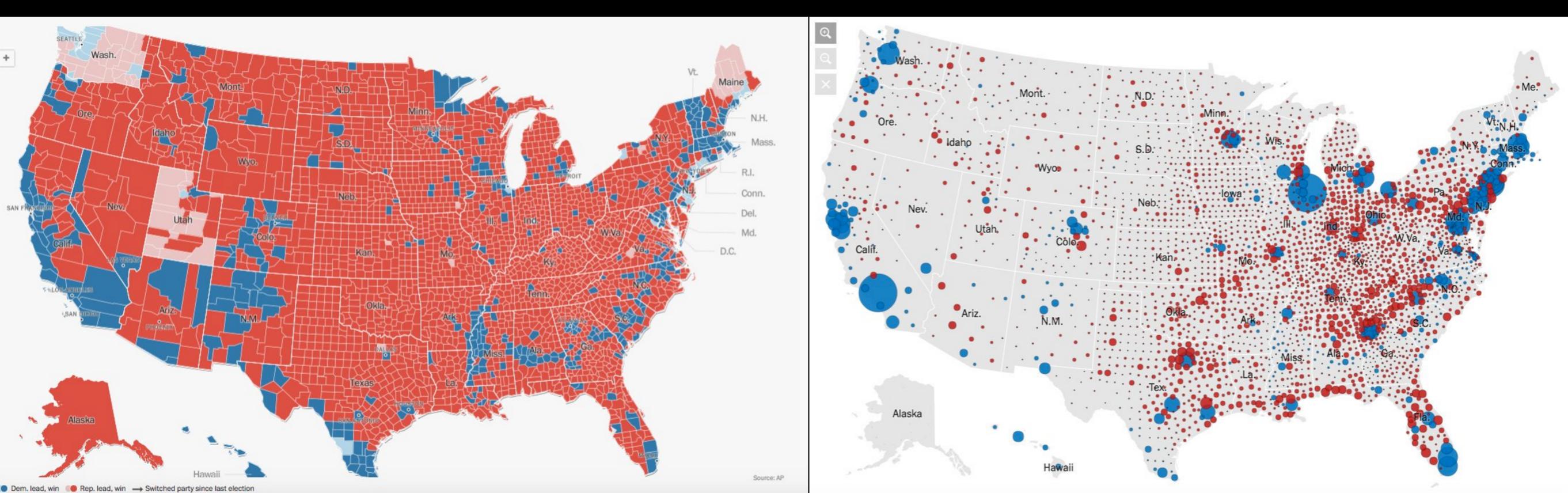
# Be Proportional



Trey Yingst @TreyYingst · May 11

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





## US Presidential Election 2016

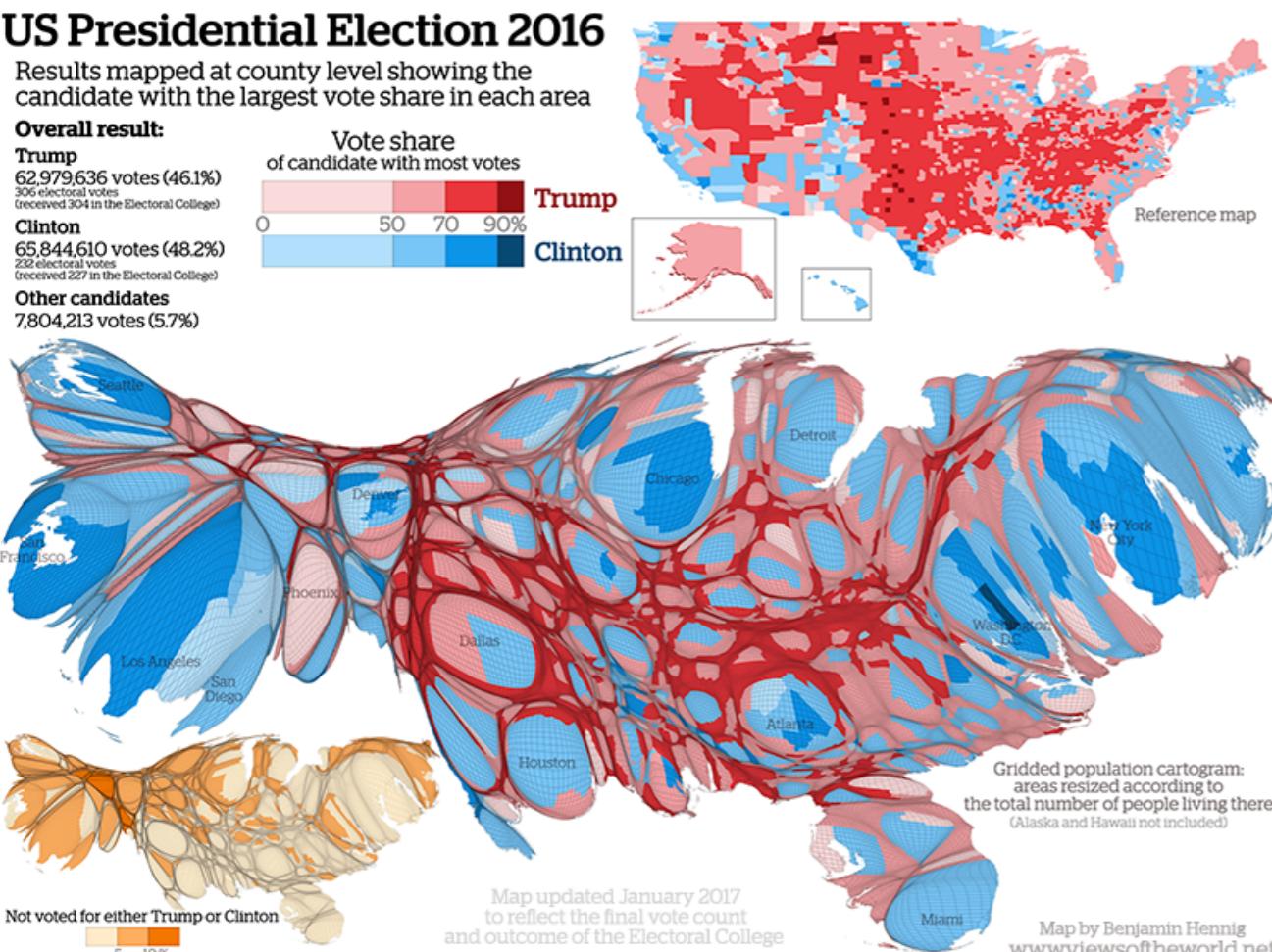
Results mapped at county level showing the candidate with the largest vote share in each area

### Overall result:

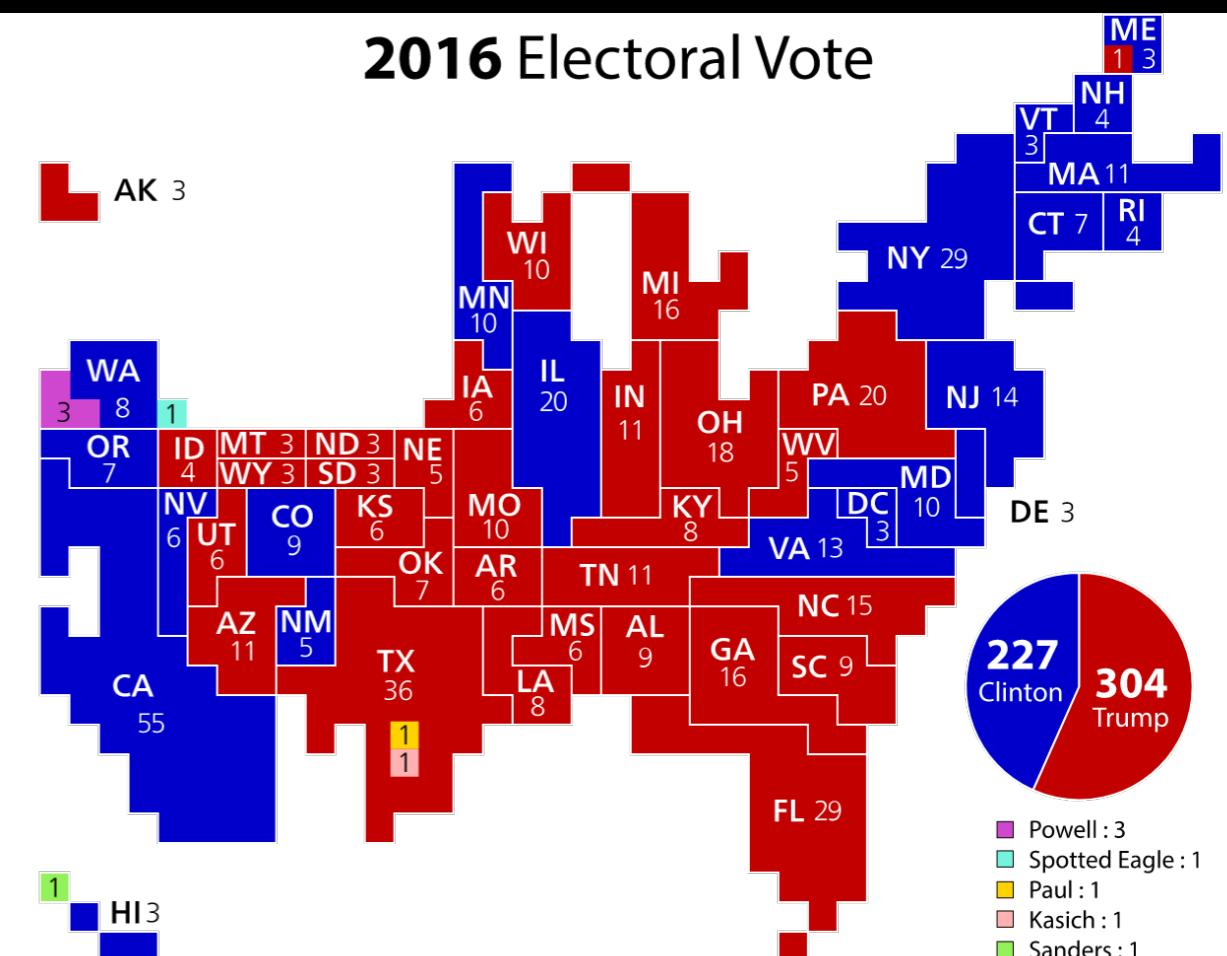
**Trump**  
62,979,636 votes (46.1%)  
(received 304 in the Electoral College)

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

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

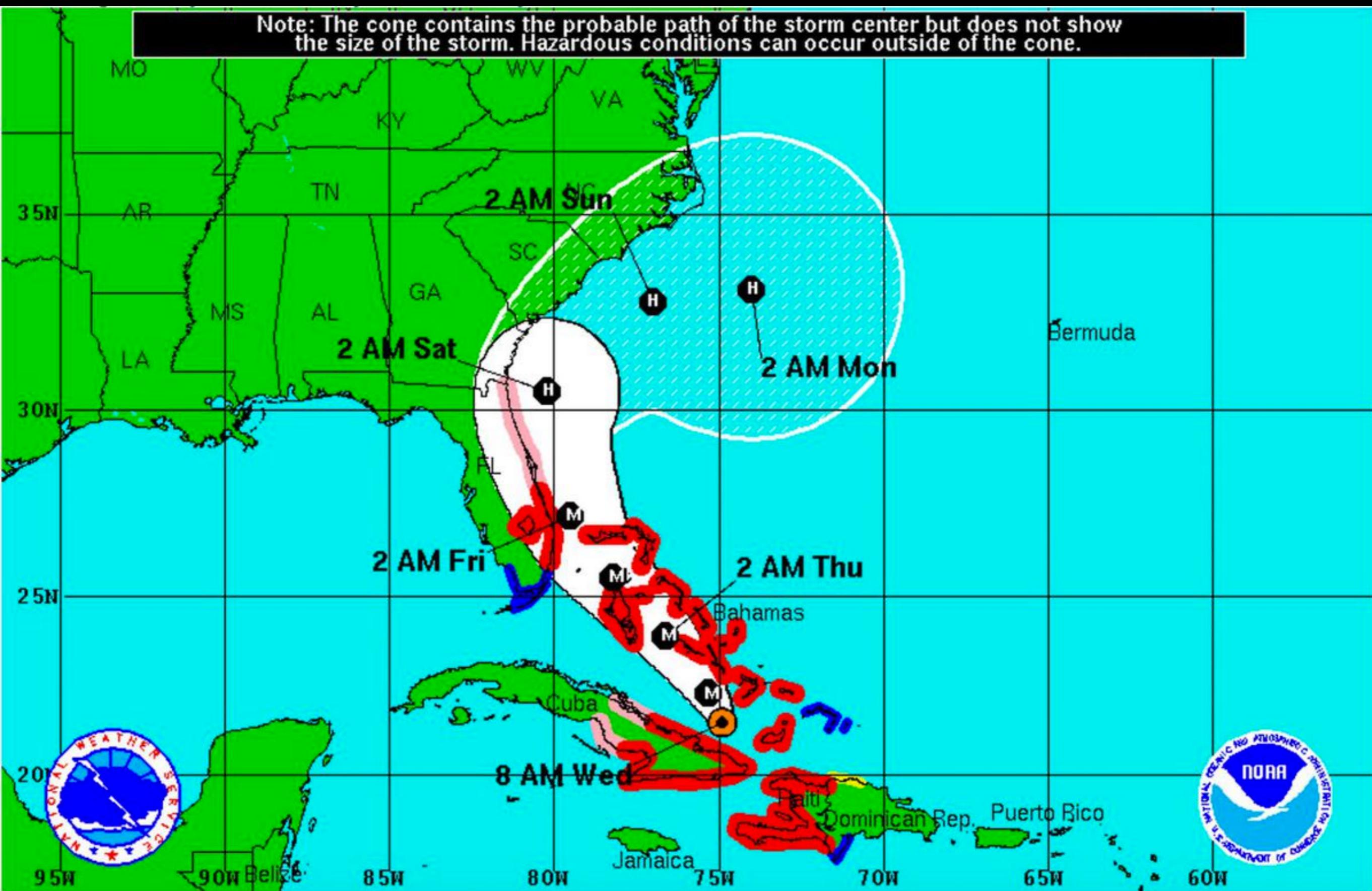


## 2016 Electoral Vote



# 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.



# Hurricane **CAIRO** (category 5)



What you show

# Hurricane **CAIRO** (category 5)

2/3

1/3

8PM Tuesday

8PM Monday

8PM Sunday

8PM Saturday

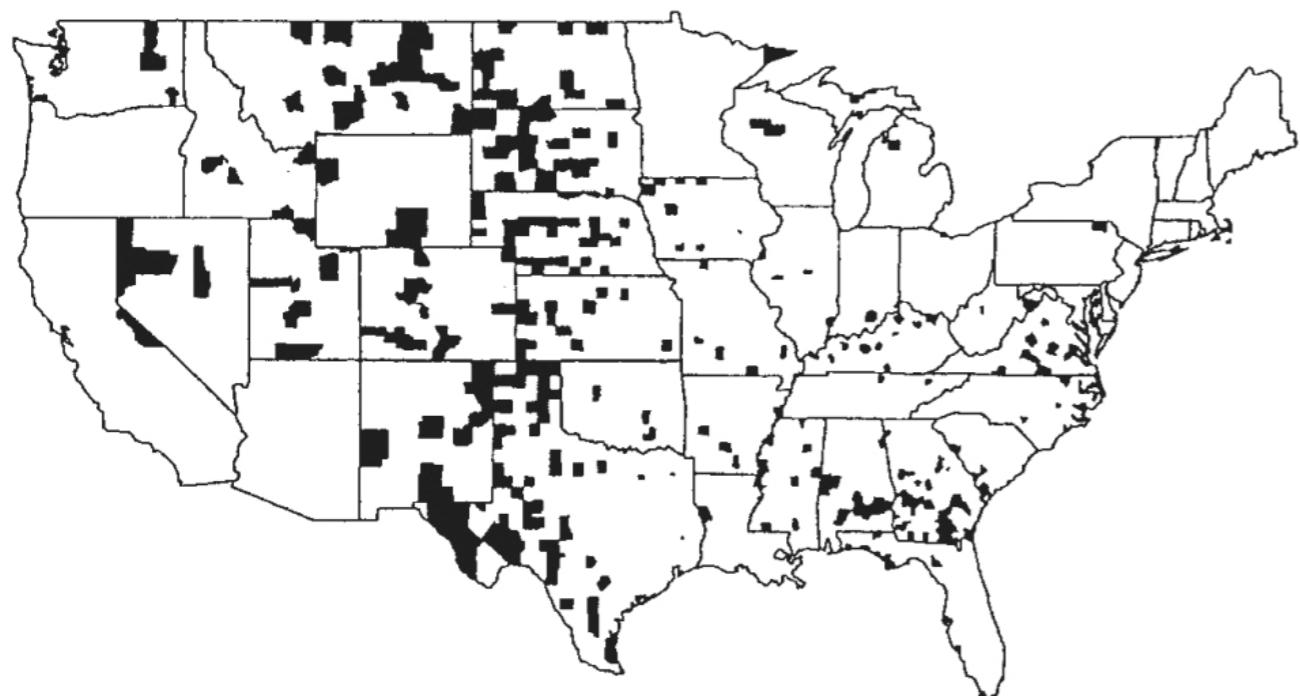
What non-scientists are not aware of (cone is just 66% probability)

# Hurricane **CAIRO** (category 5)

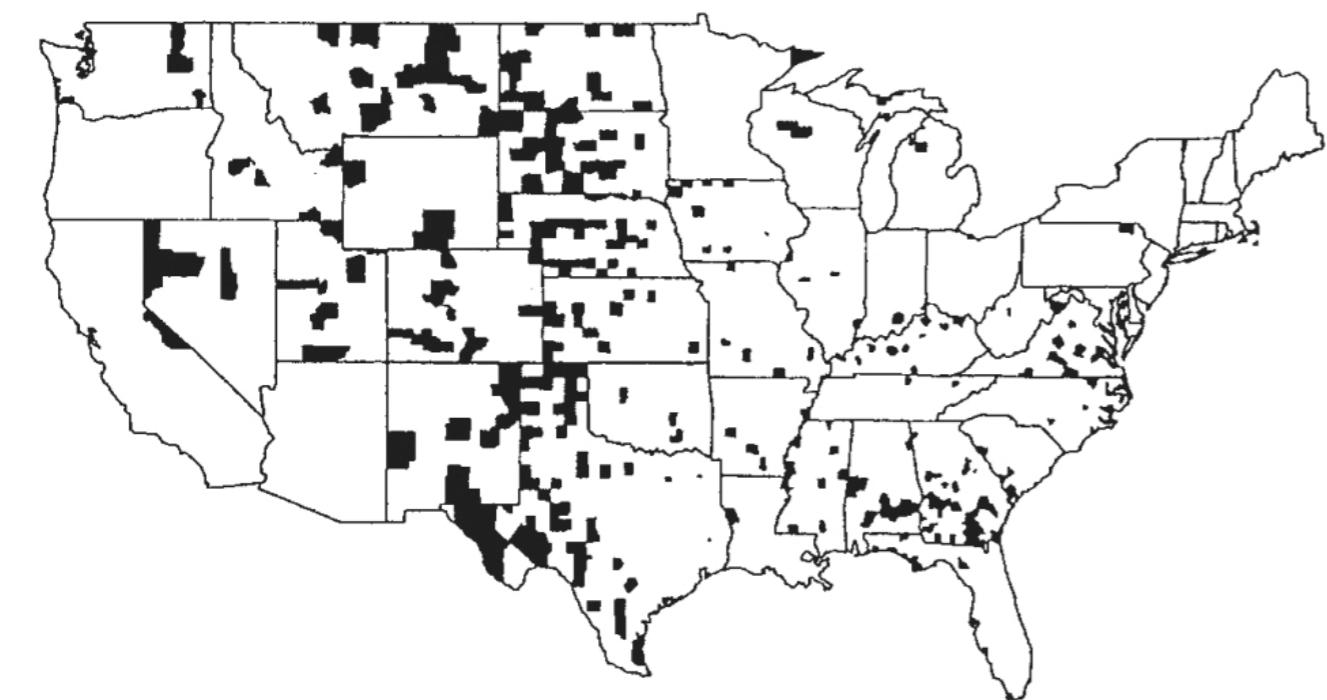


What we could be showing instead

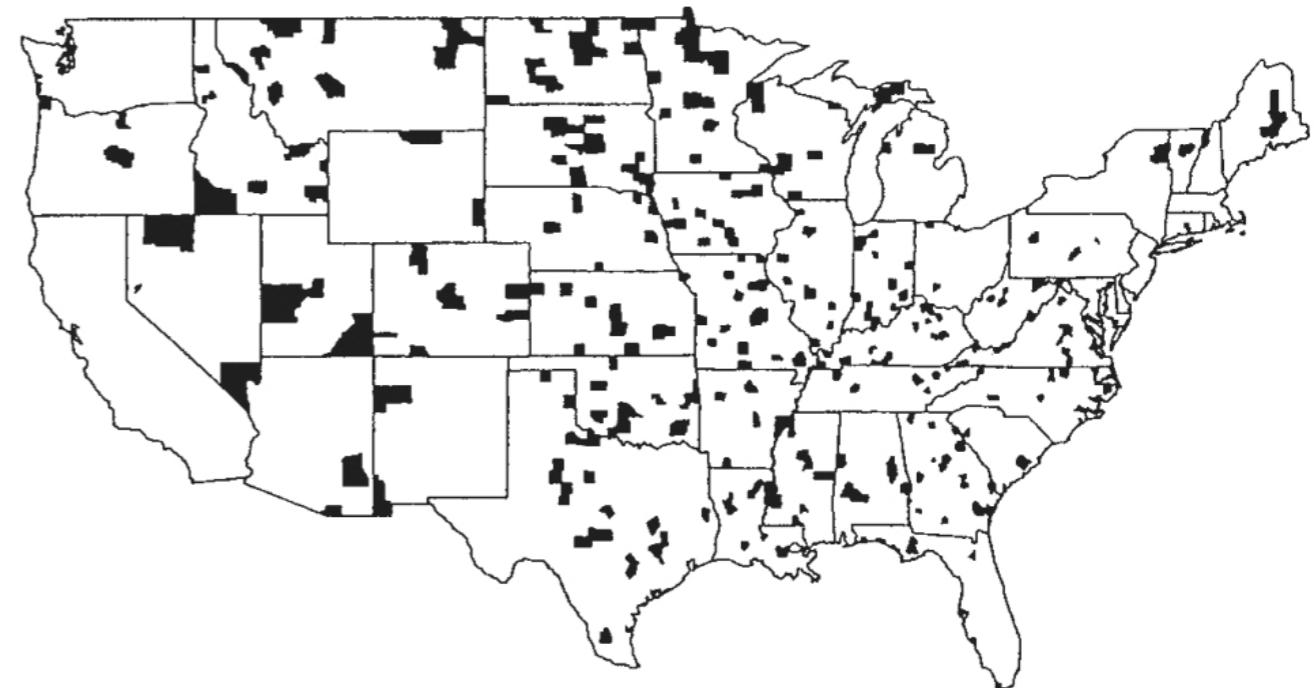
**Plot all your data**



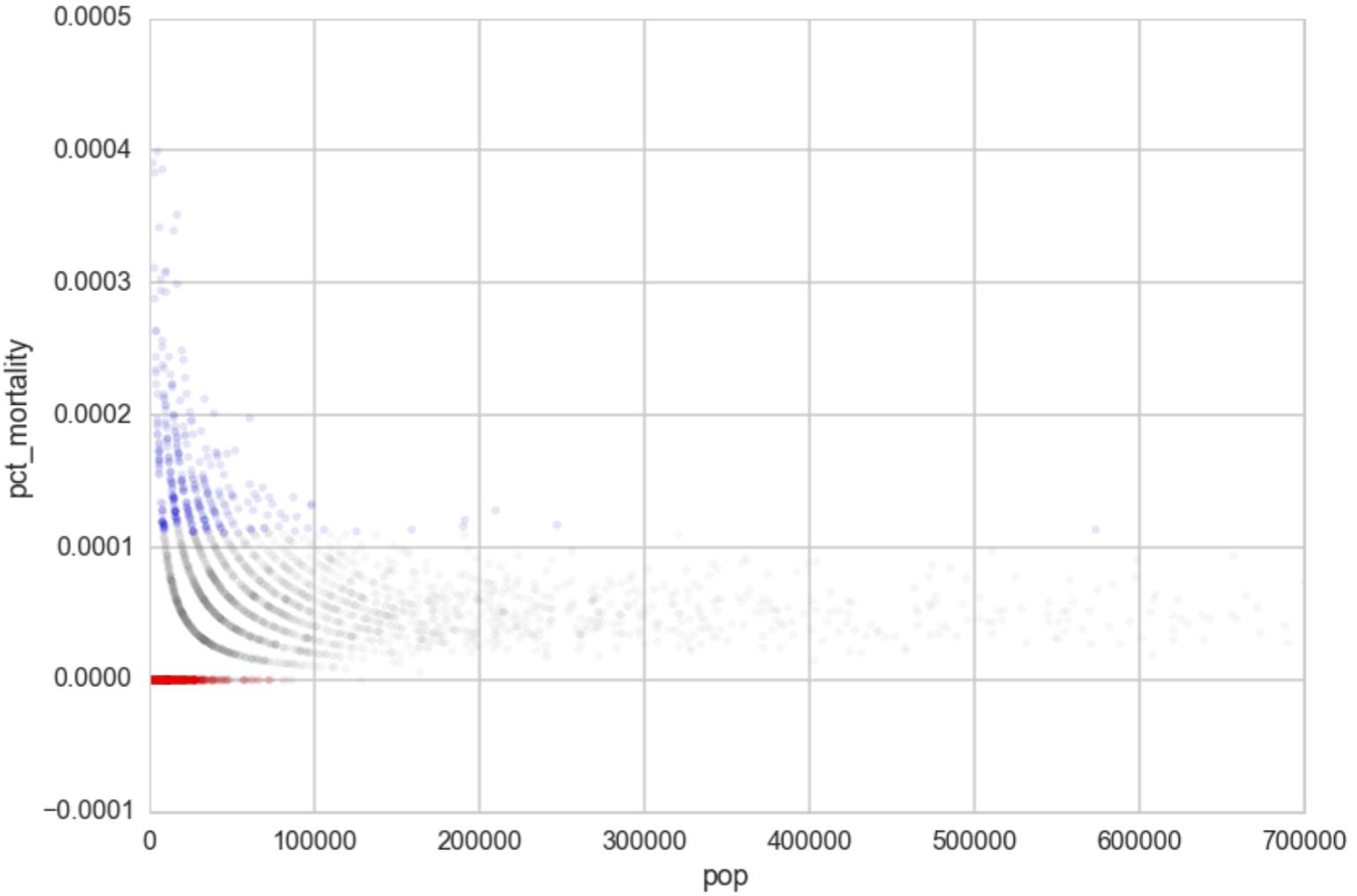
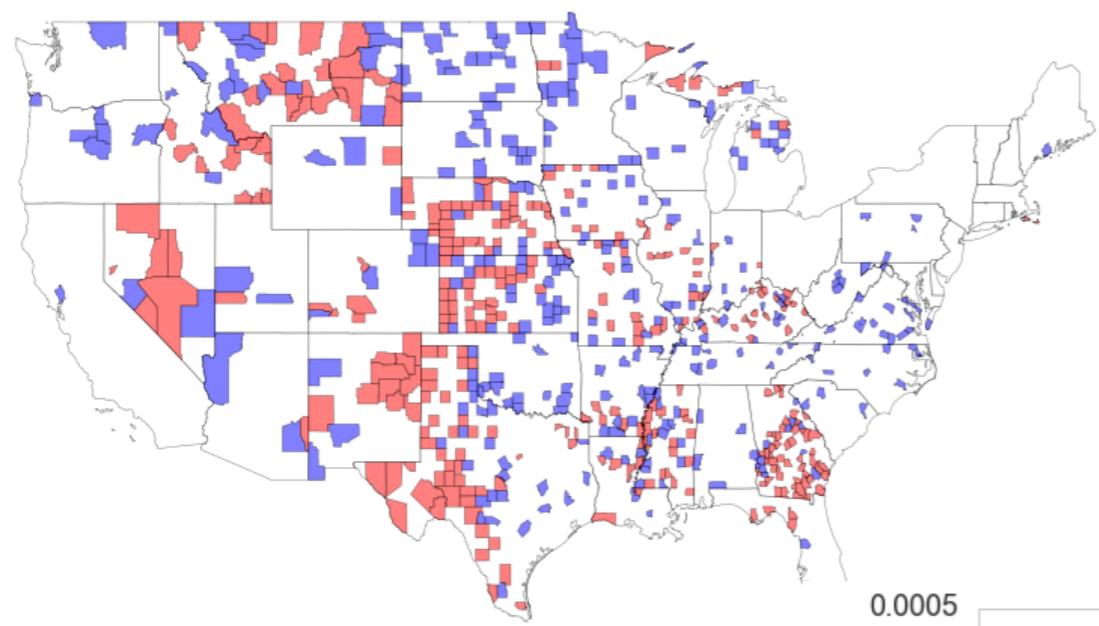
Counties with the LOWEST  
kidney cancer death rates  
(1980-1989)



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



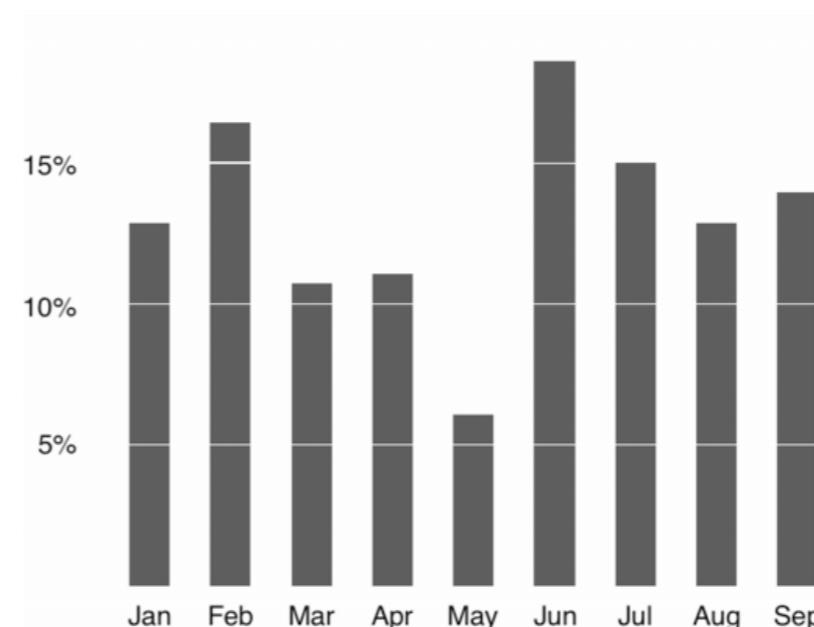
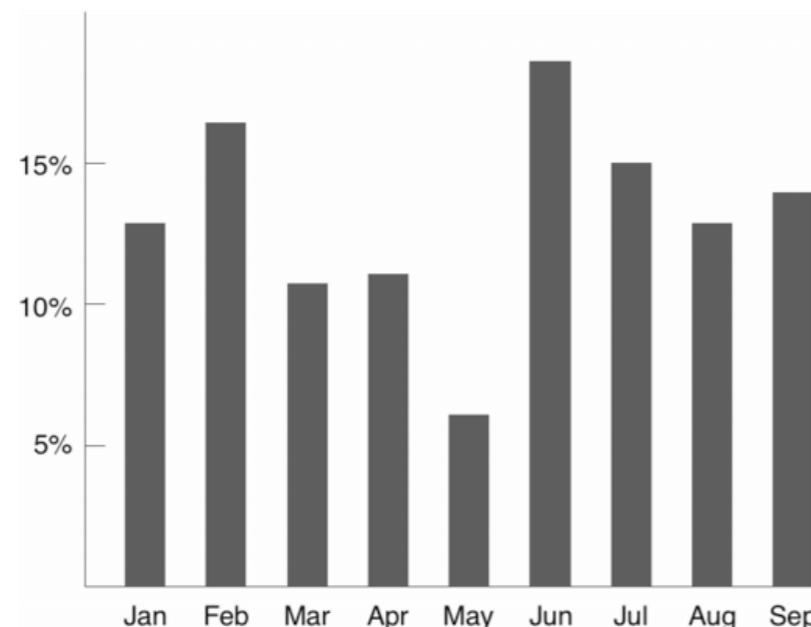
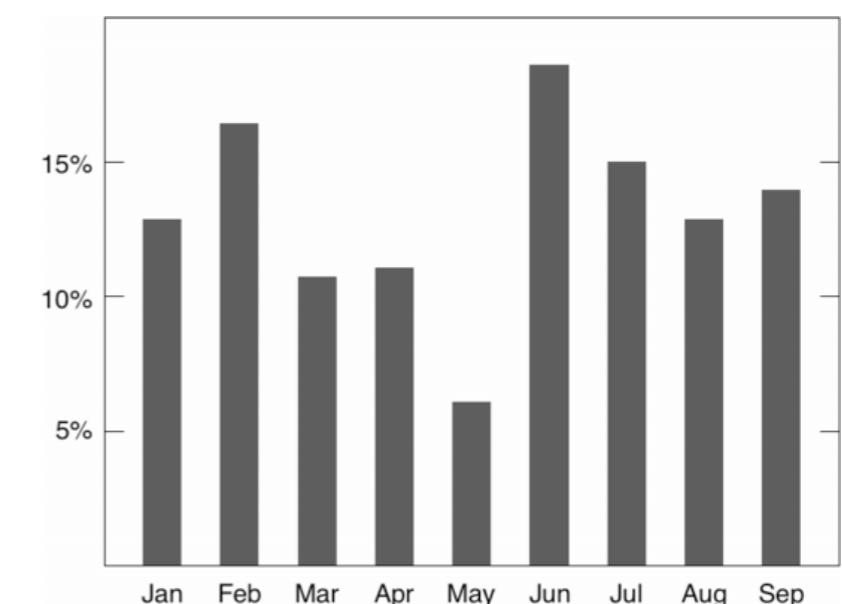
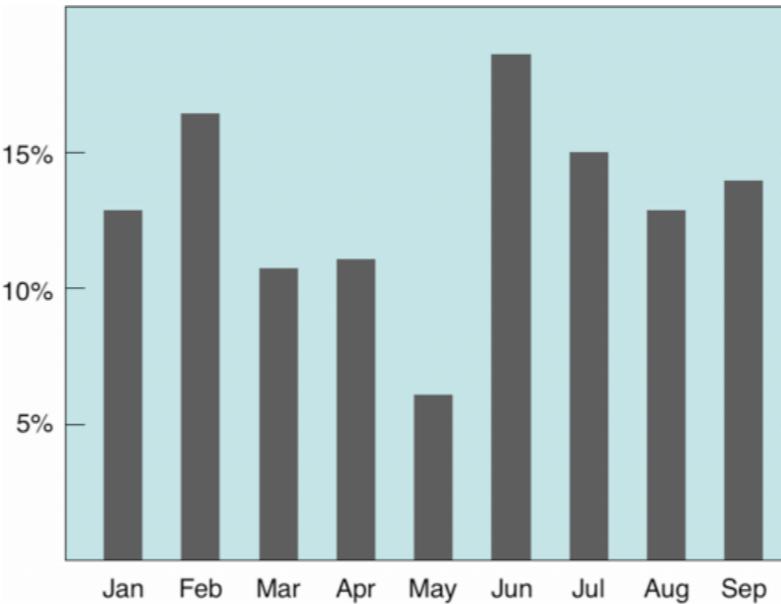
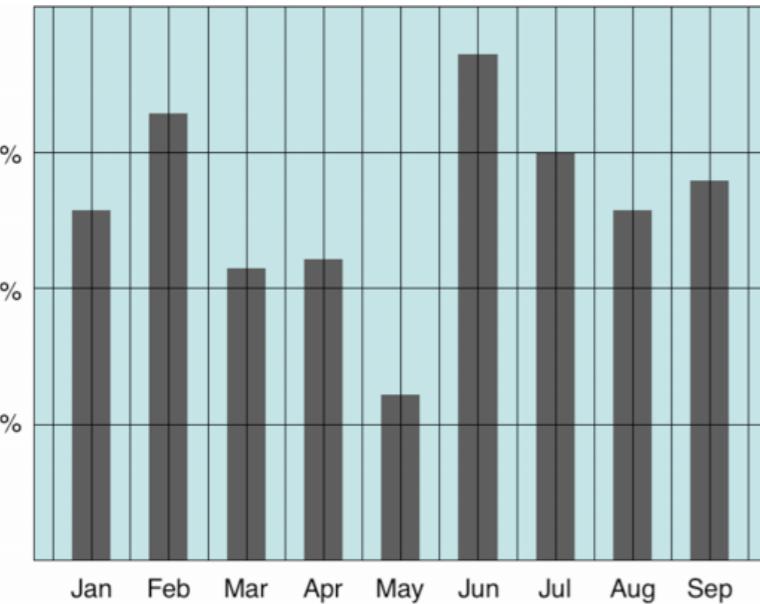
Counties with the HIGHEST  
kidney cancer death rates  
(1980-1989)

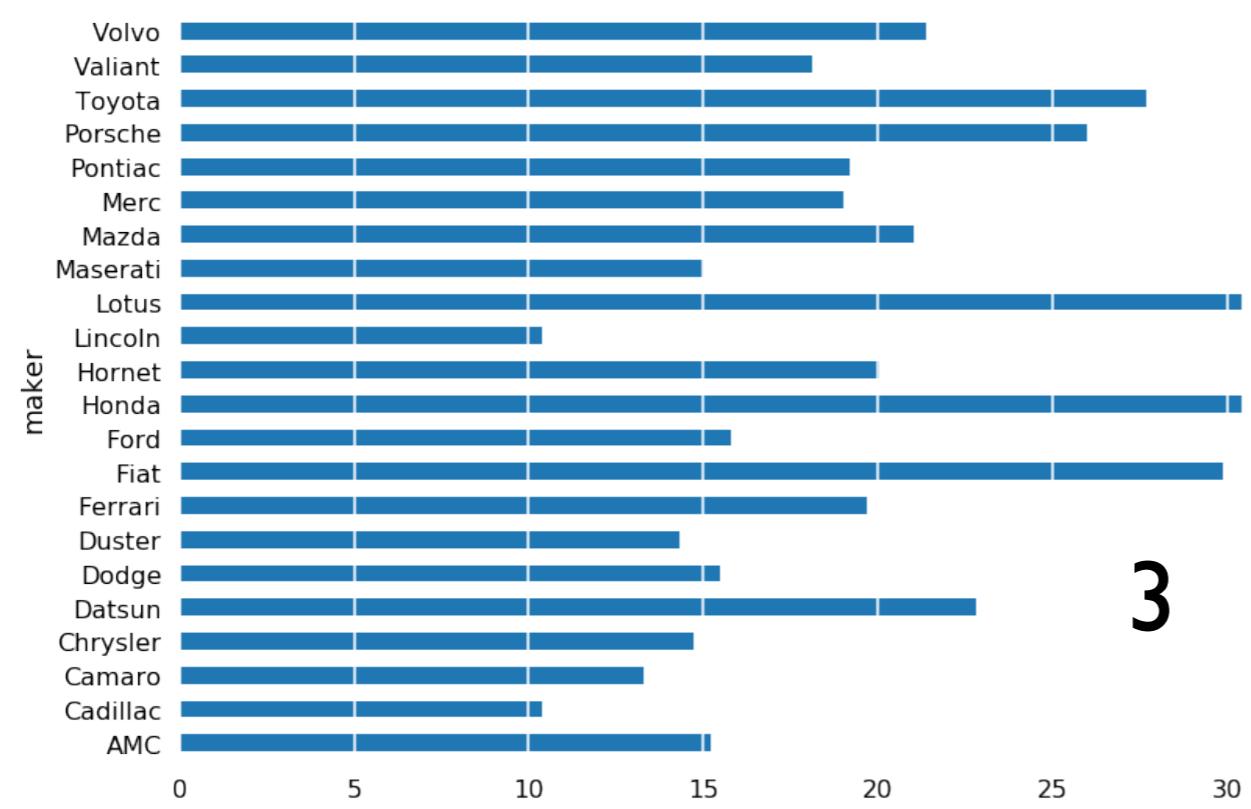
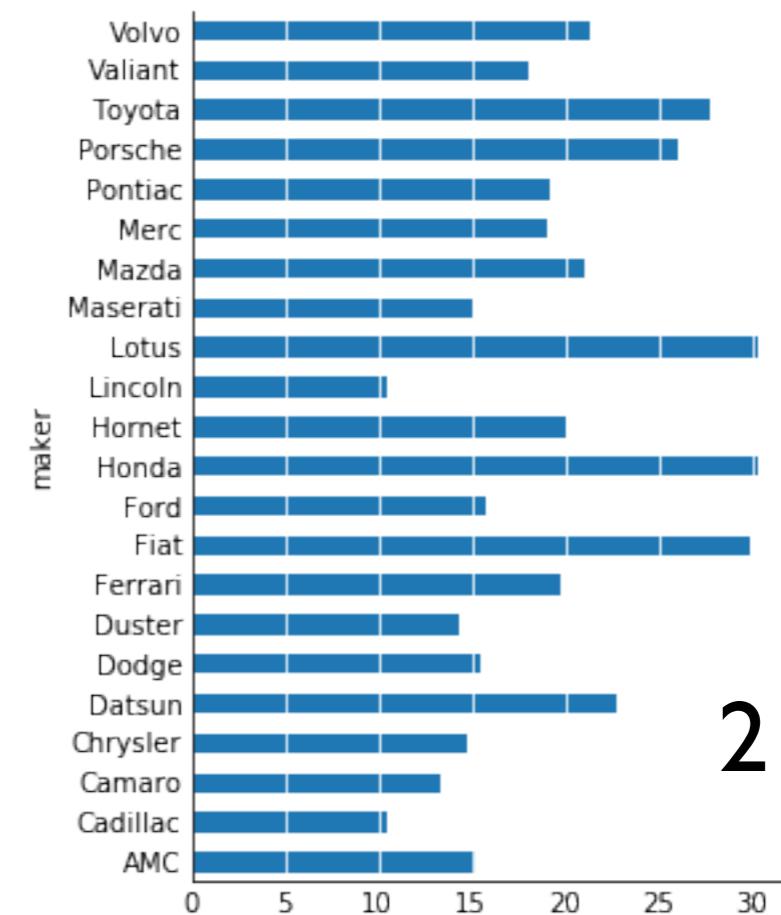
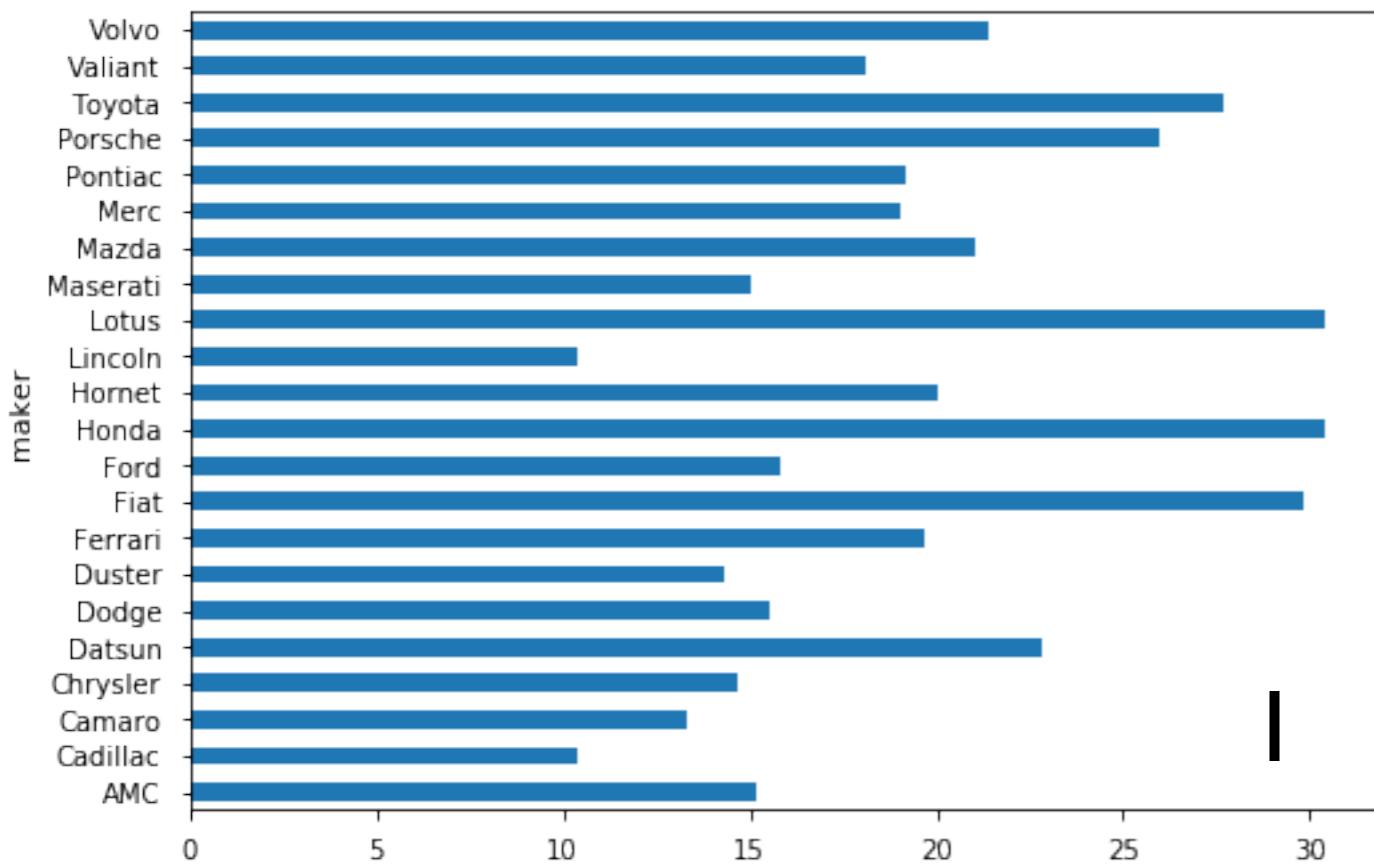


## 2. Keep It Simple

# Avoid Chartjunk

Extraneous visual elements that distract from the message



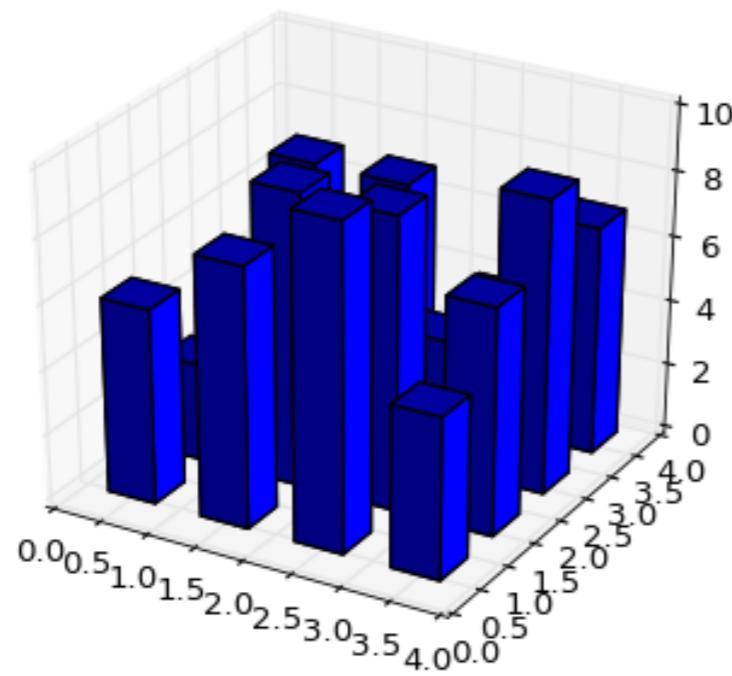


A vertical bar chart with 'maker' on the y-axis and an unlabeled numerical value on the x-axis (0 to 30). The bars are blue.

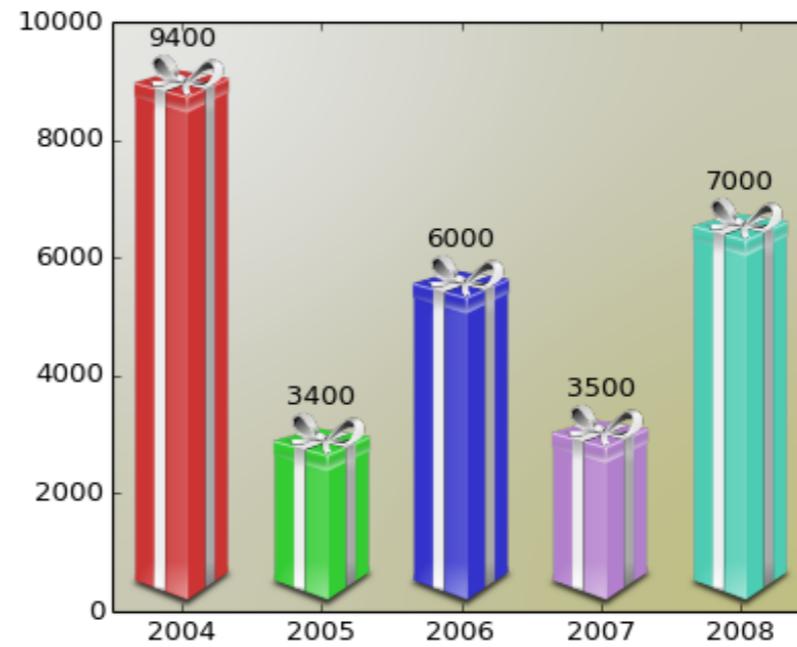
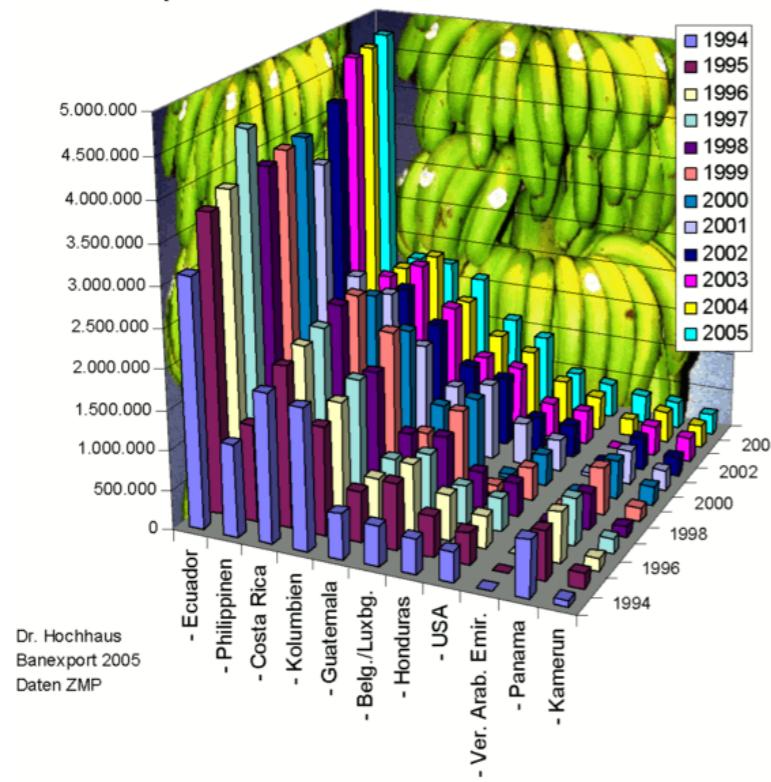
maker	Value
Honda	30.4
Lotus	30.4
Fiat	29.85
Toyota	27.7
Porsche	26.0
Datsun	22.8
Volvo	21.4
Mazda	21.0
Hornet	20.05
Ferrari	19.7
Pontiac	19.2
Merc	19.0142857143
Valiant	18.1
Ford	15.8
Dodge	15.5
AMC	15.2
Maserati	15.0
Chrysler	14.7
Duster	14.3
Camaro	13.3
Lincoln	10.4
Cadillac	10.4

4

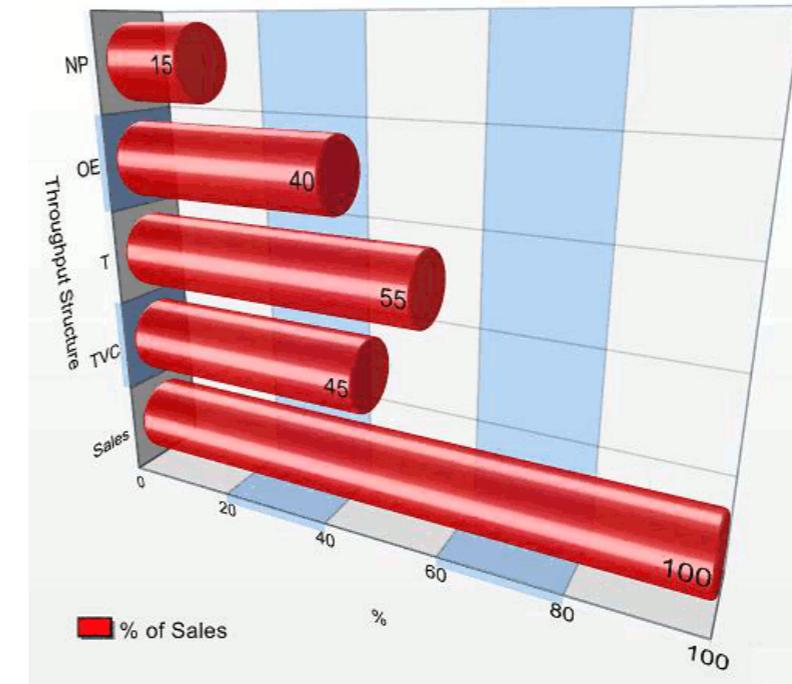
# Don't!



Export von Bananen in Tonnen von 1994-2005



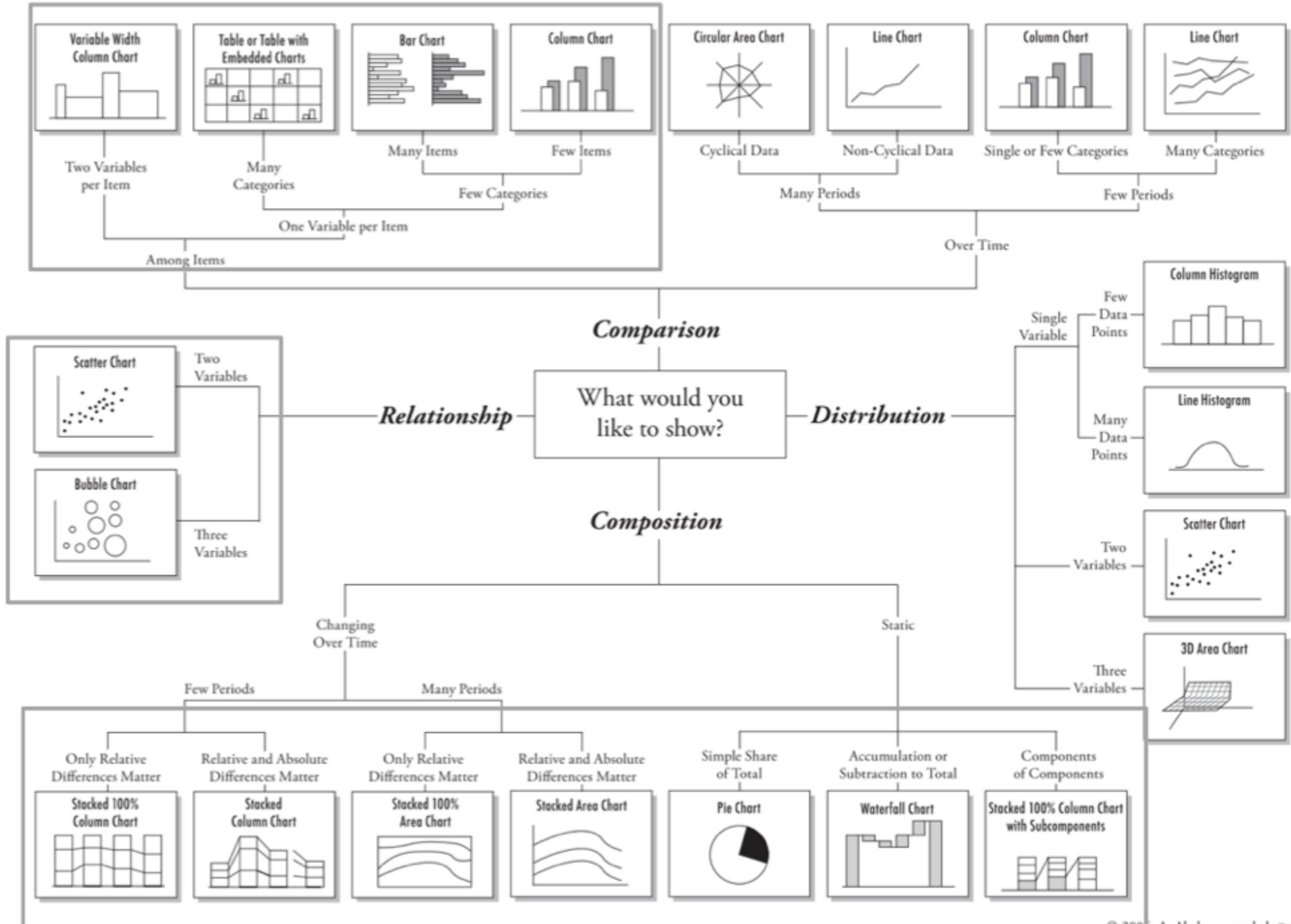
matplotlib gallery



Excel Charts Blog

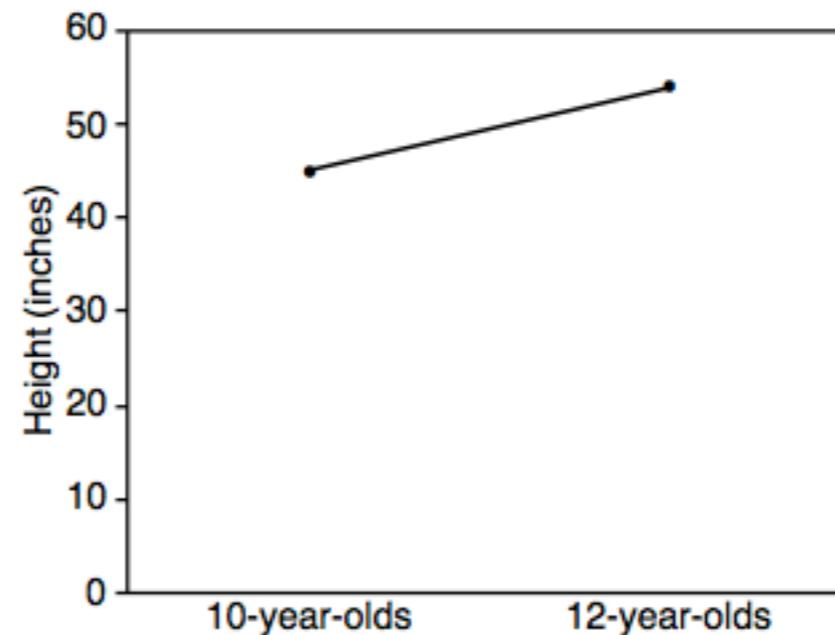
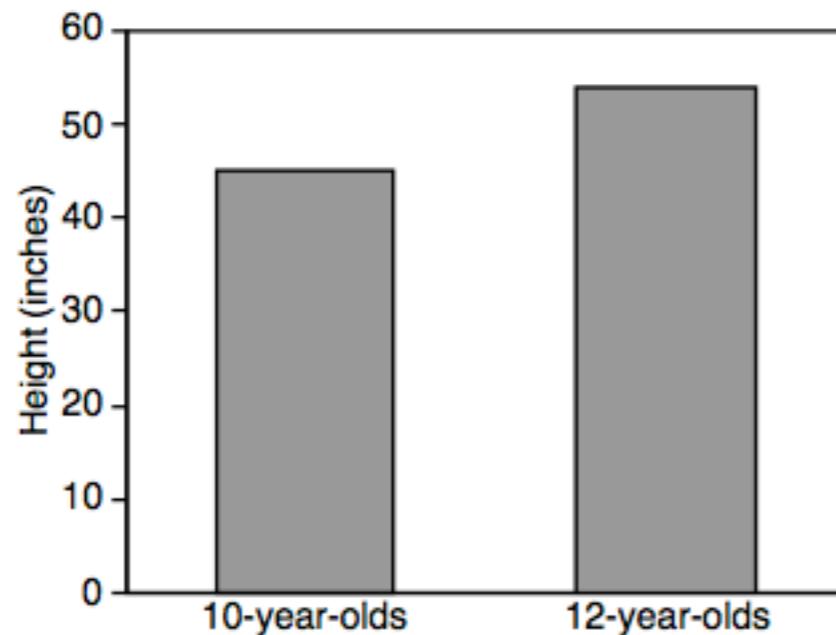
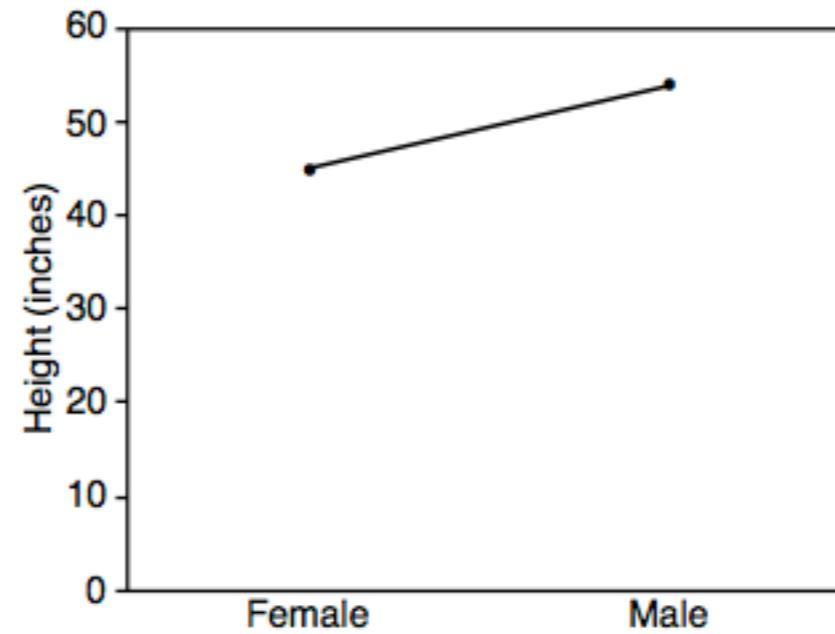
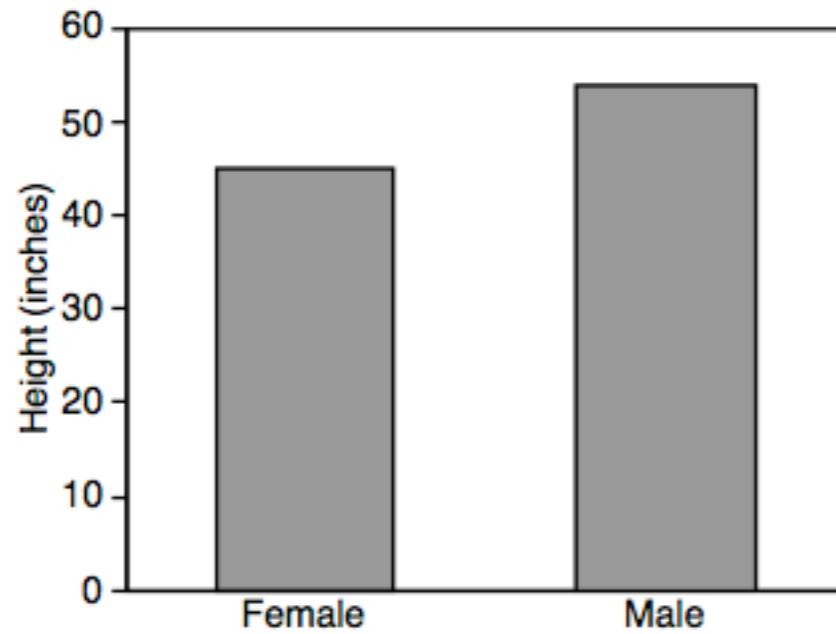
# **3. Use The Right Display**

# Chart Suggestions—A Thought-Starter



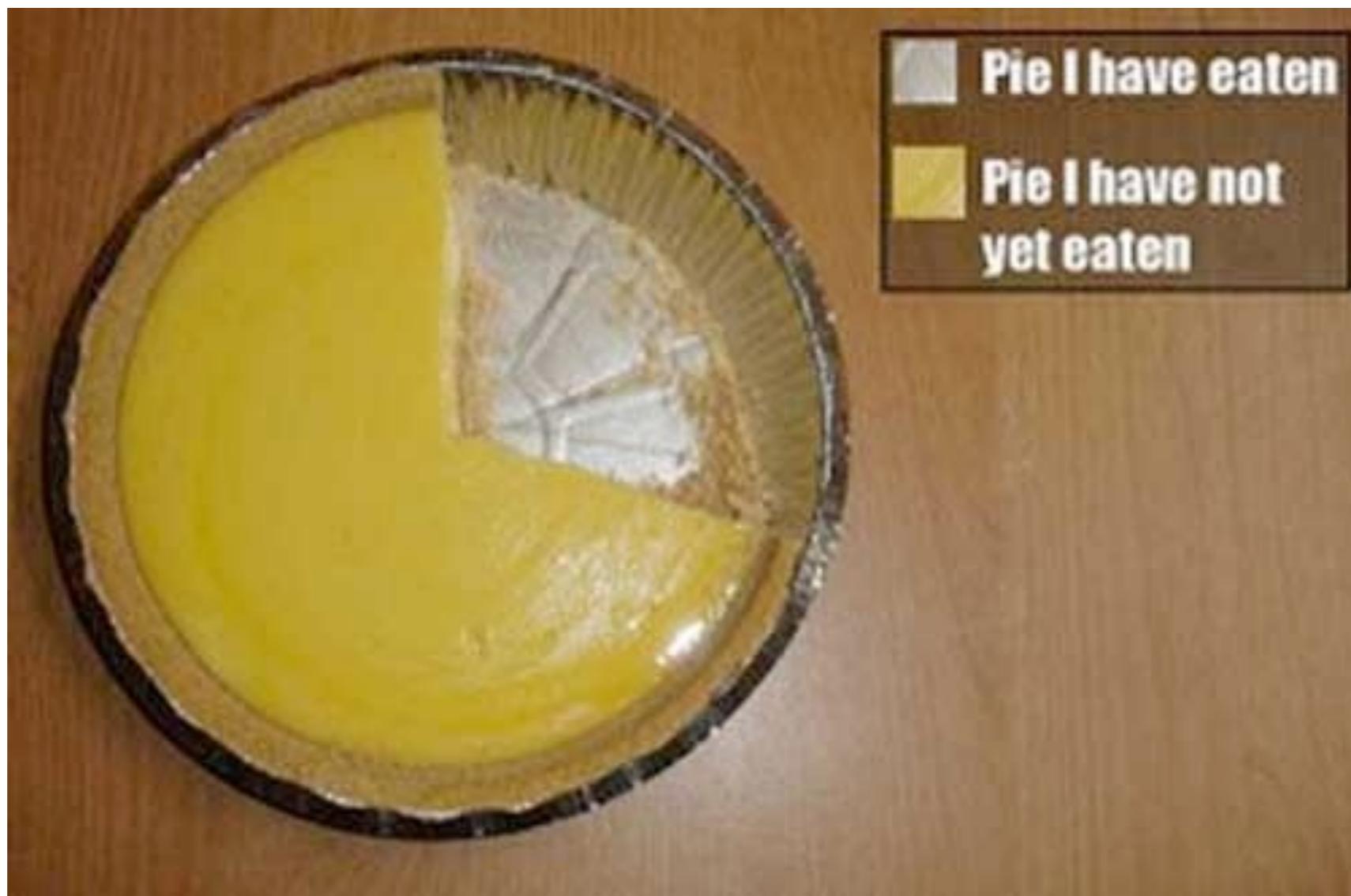
# Comparisons

# Bars vs. Lines

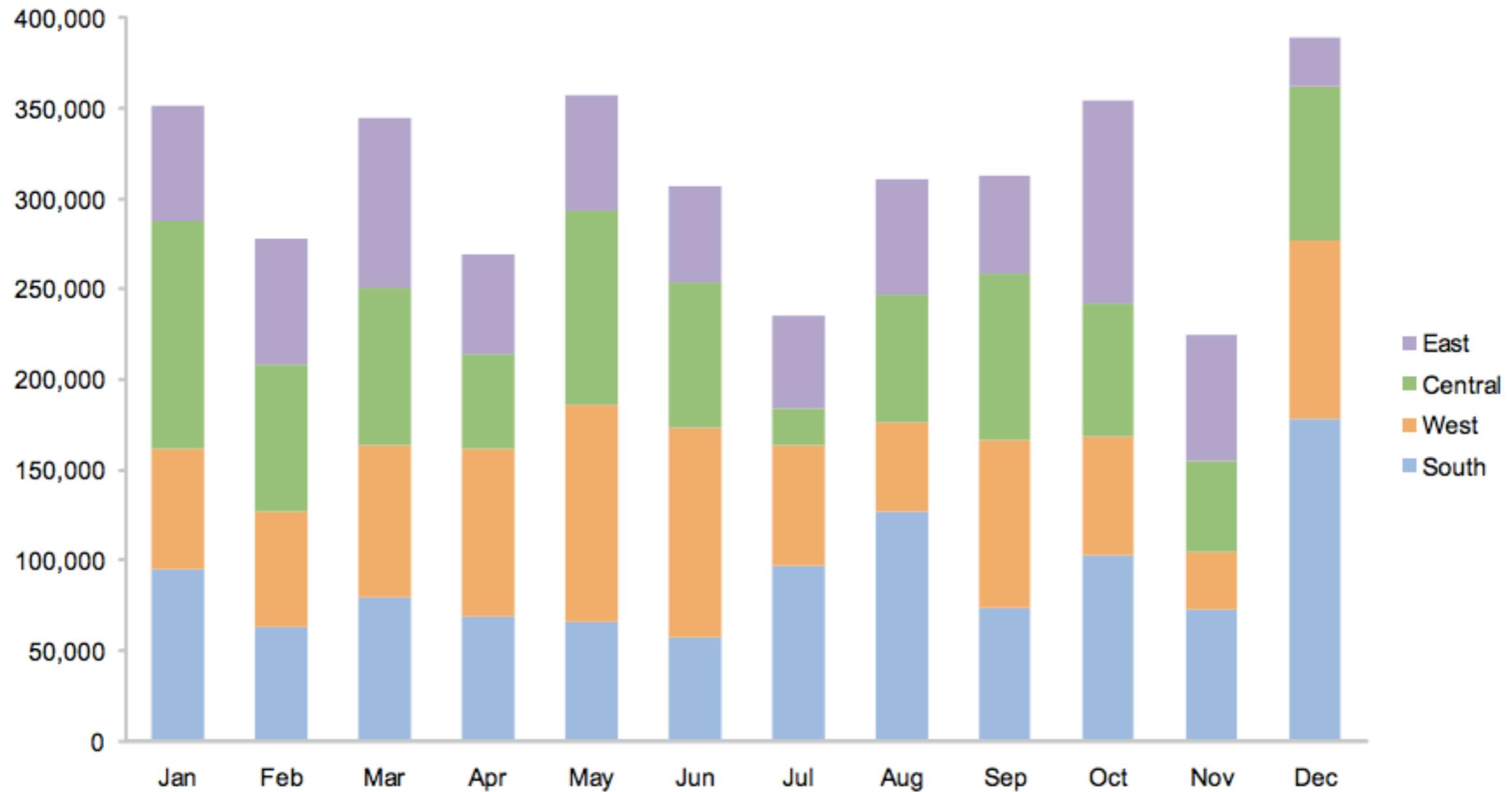


# Proportions

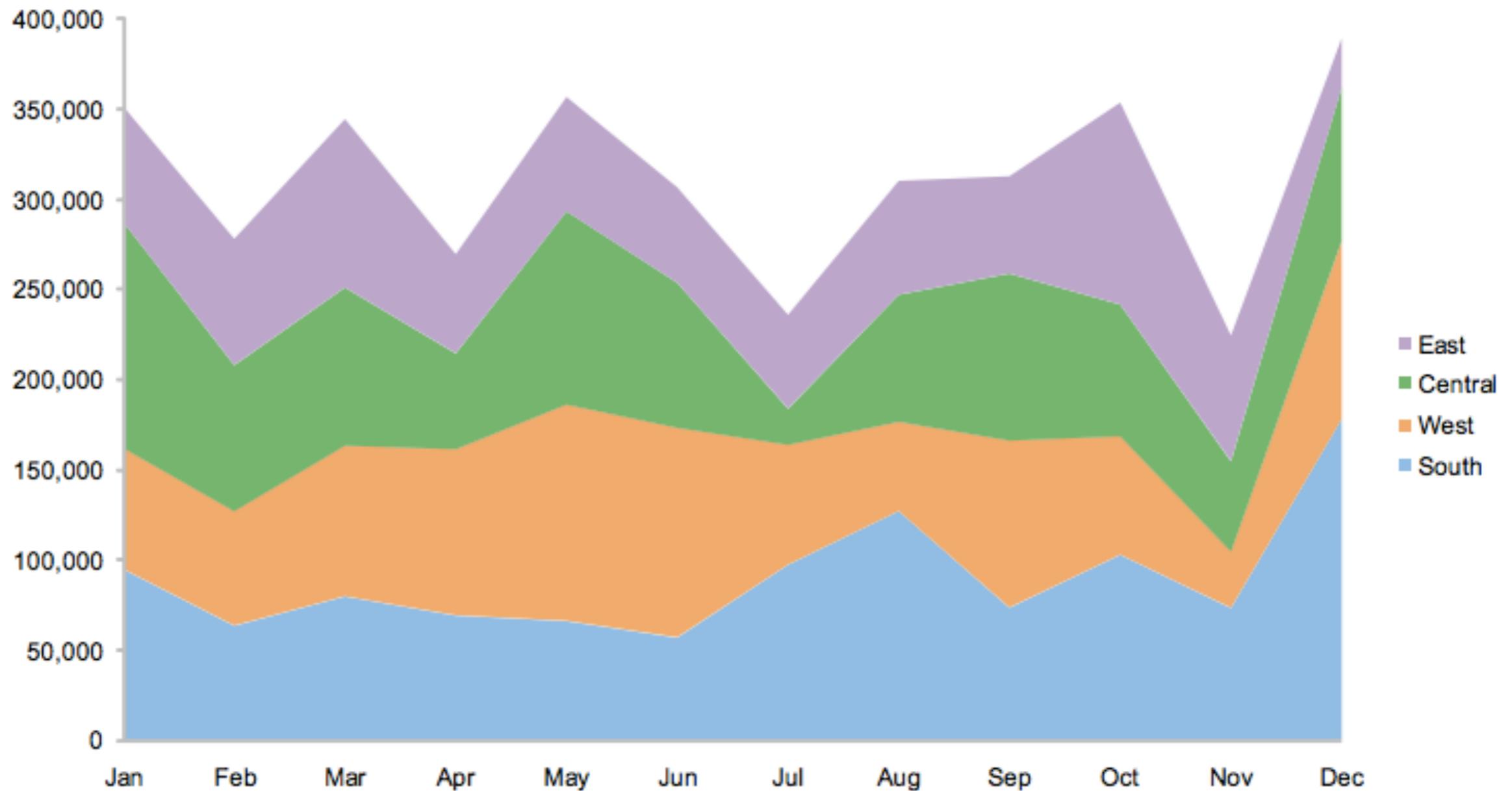
# Pie Charts



# Stacked Bar Chart

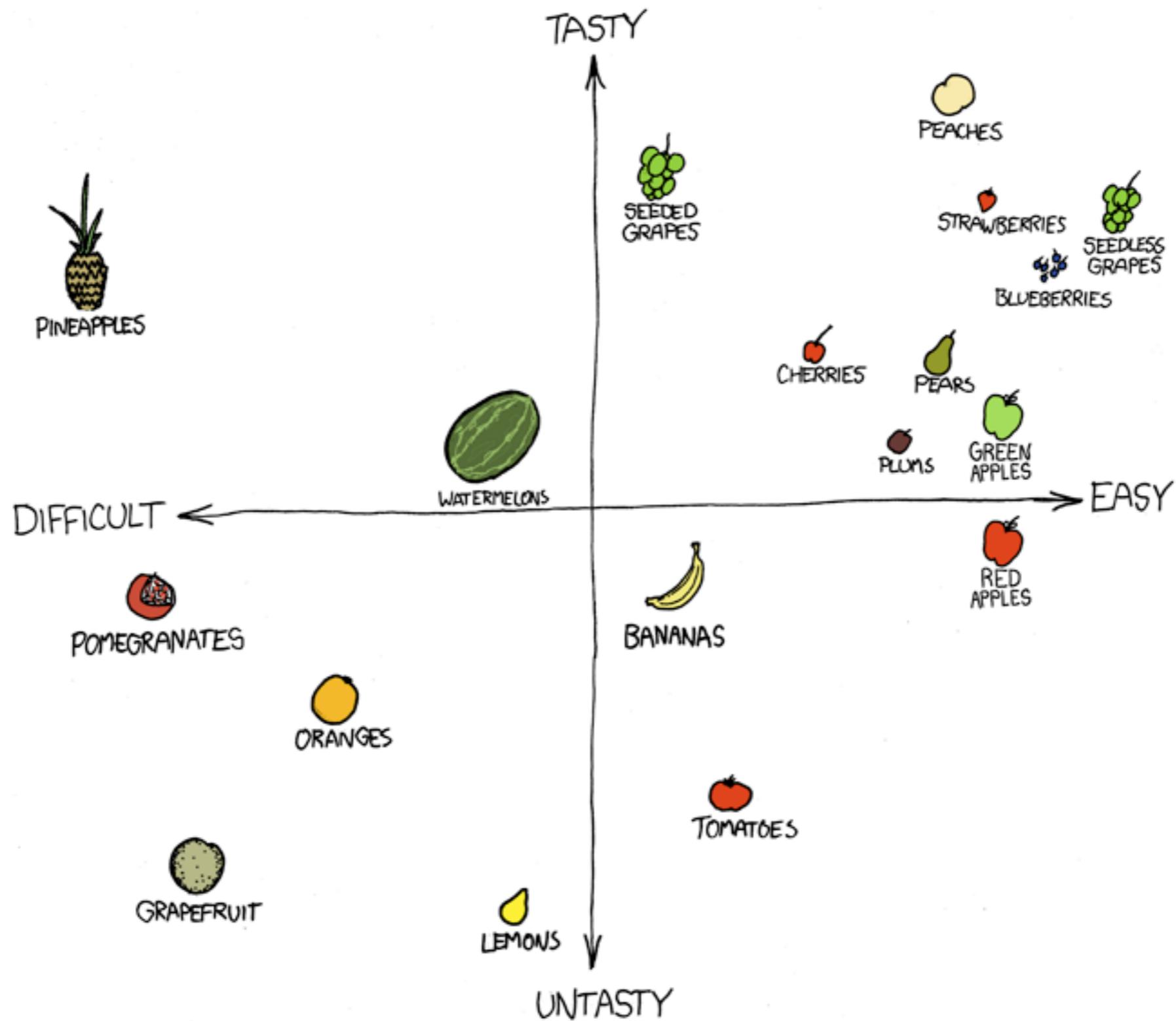


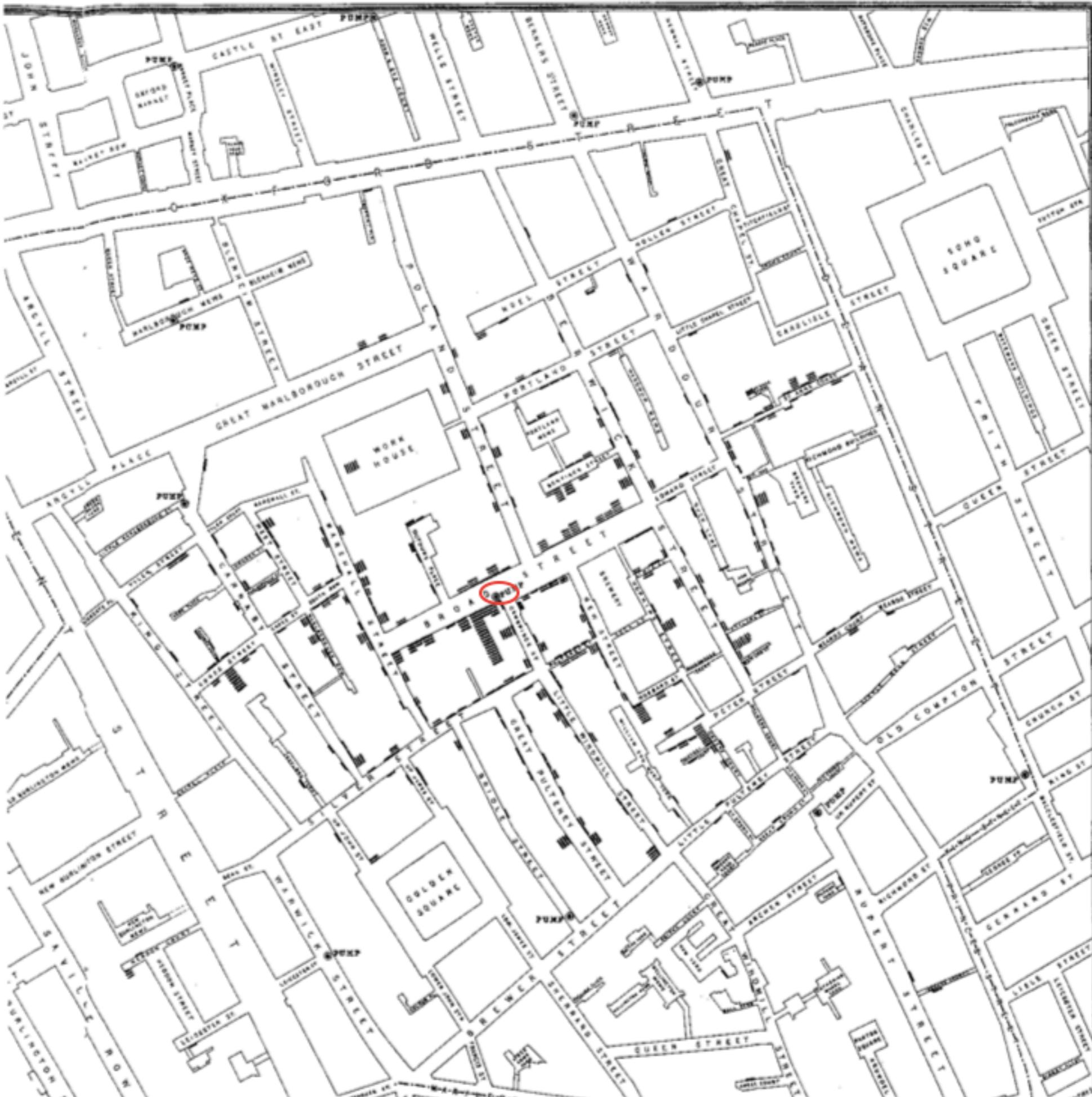
# Stacked Area Chart



# Correlations

# Scatterplots

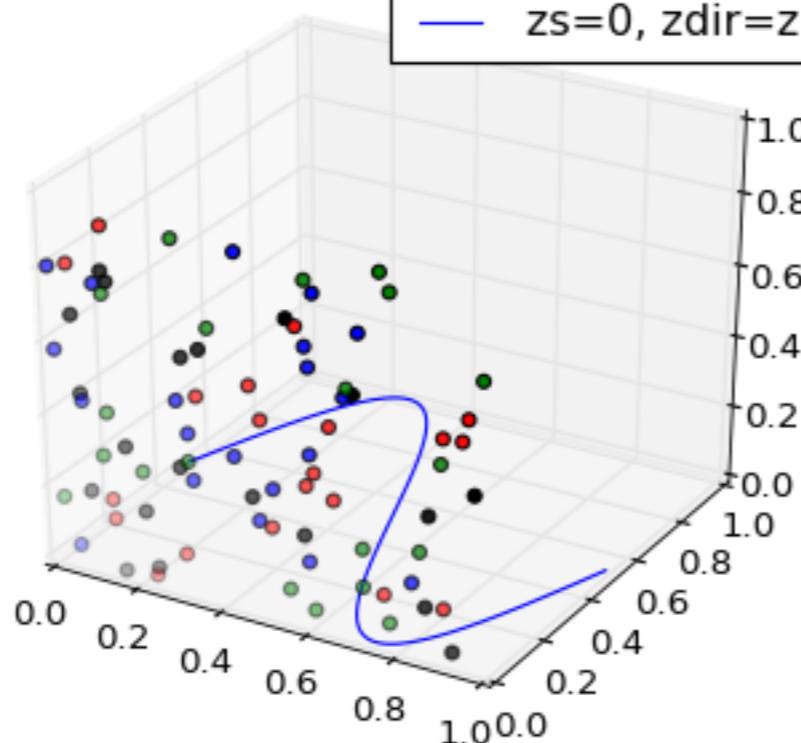
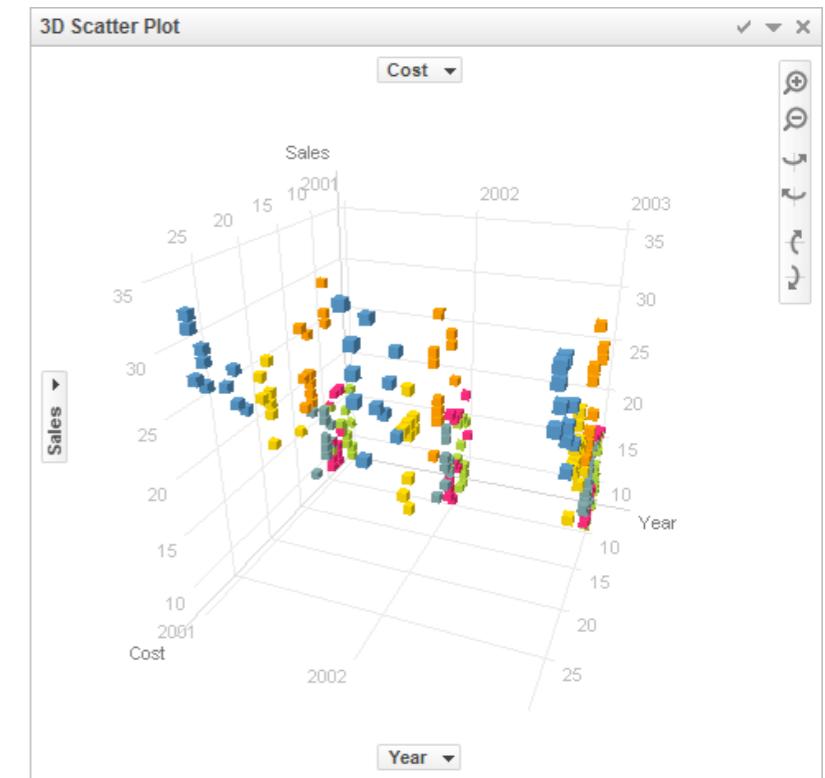
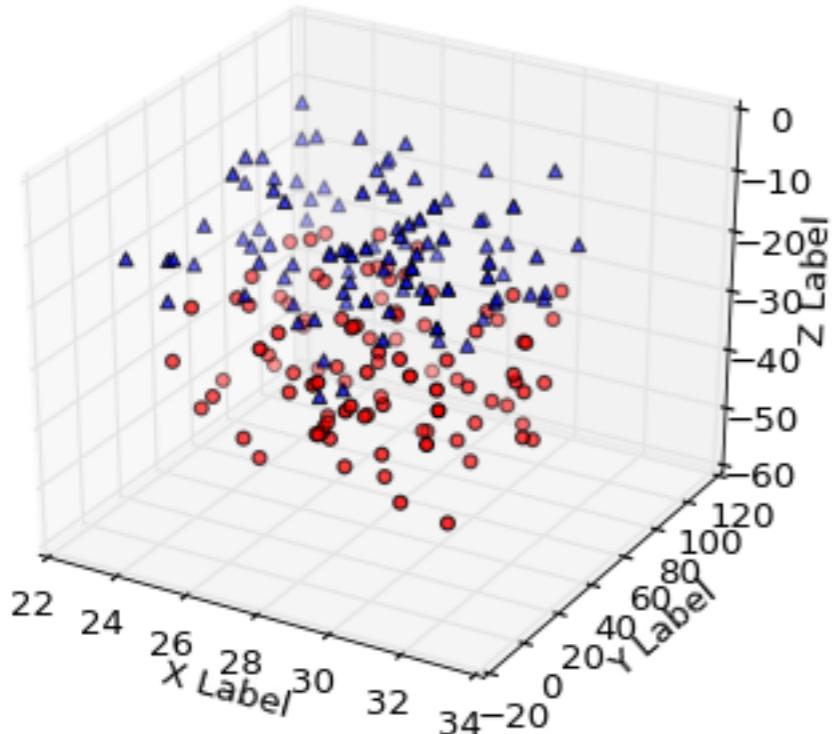




## London Cholera Epidemic

From Edward Tufte, Visual and Statistical Thinking

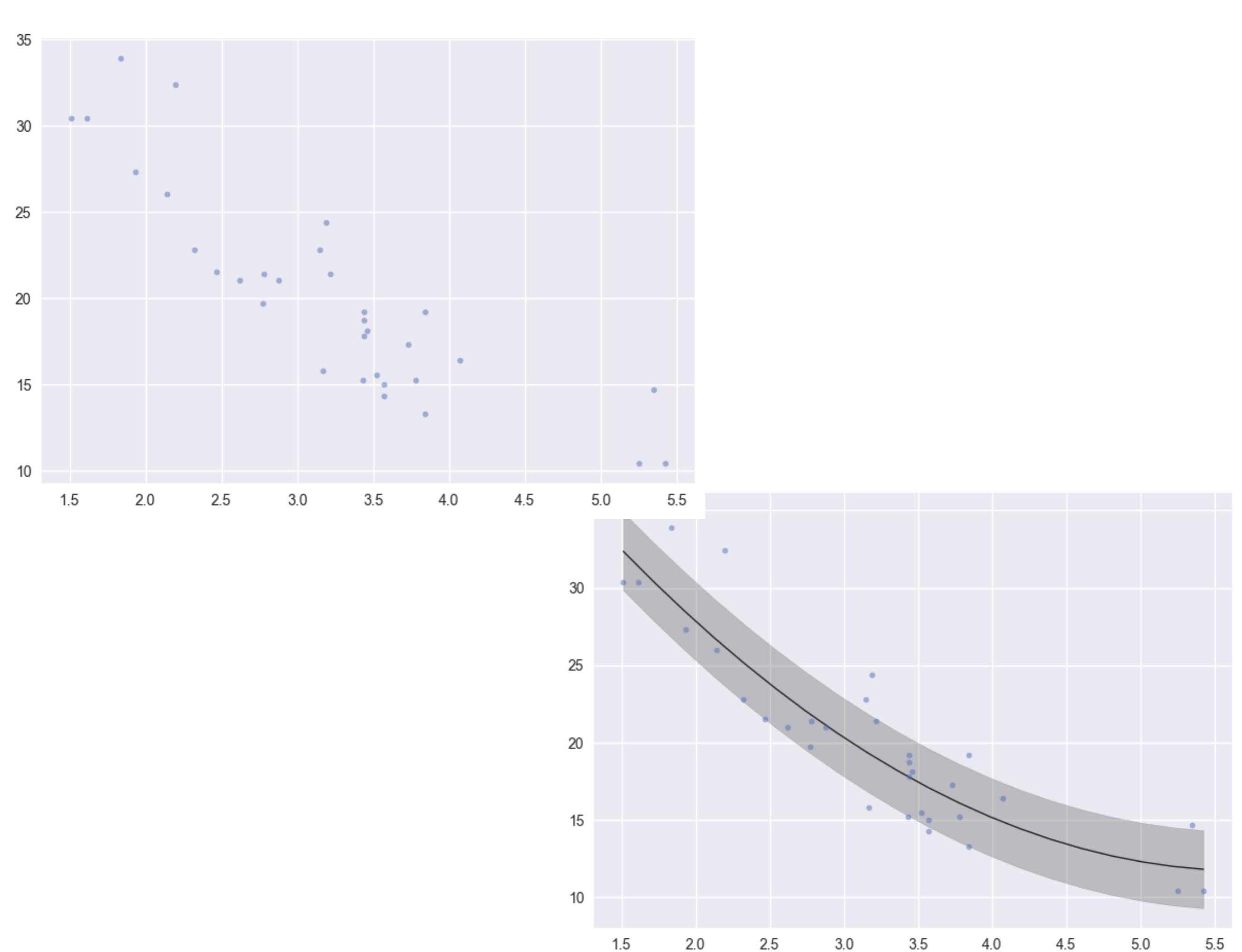
# Don't!



# Trends

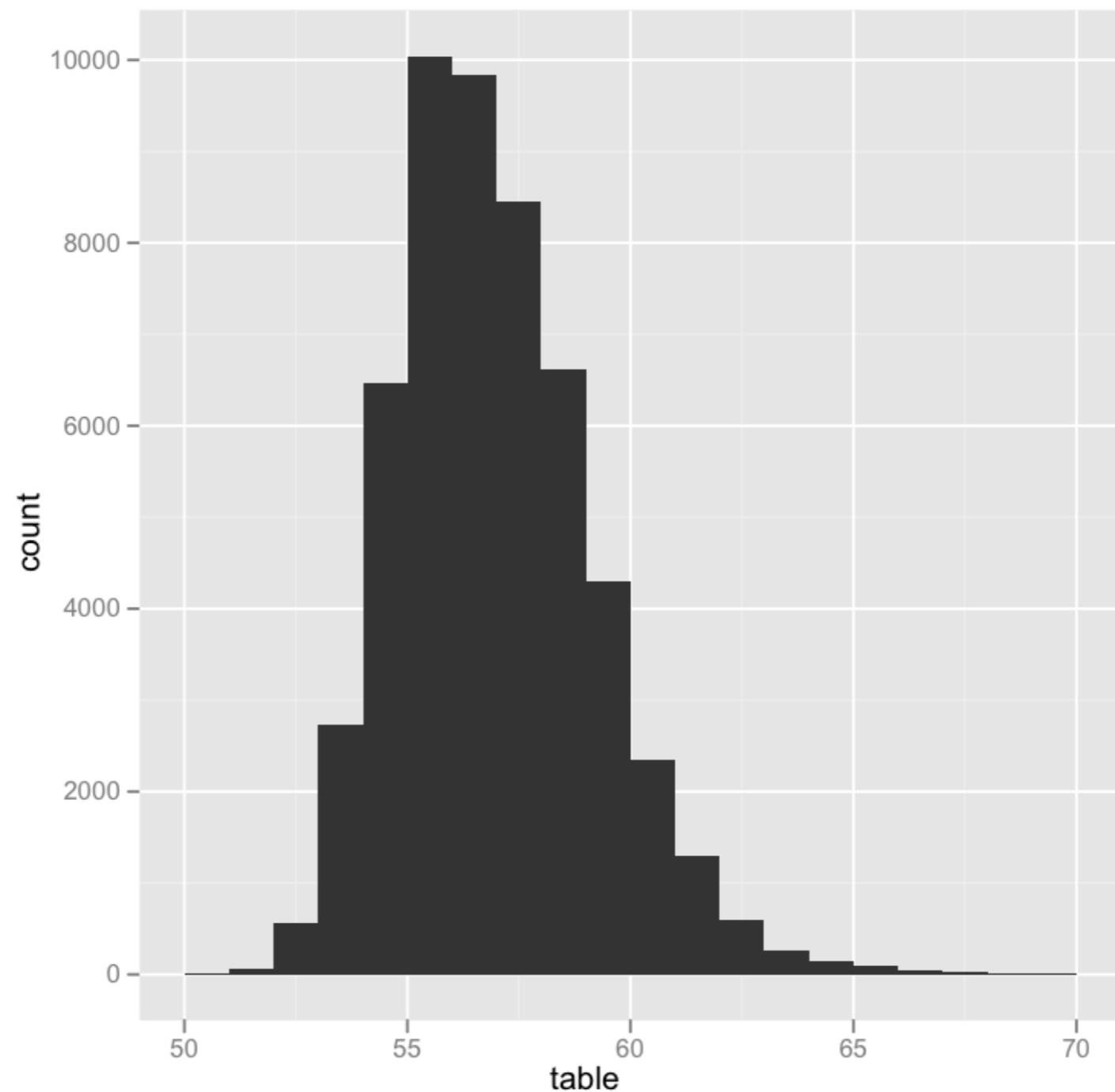
**601.10** ↑15.53(2.65%) 4:00PM EDT | After Hours: **604.60** ↑3.50 (0.58%) 7:15PM EDT - Nasdaq Real Time Price

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

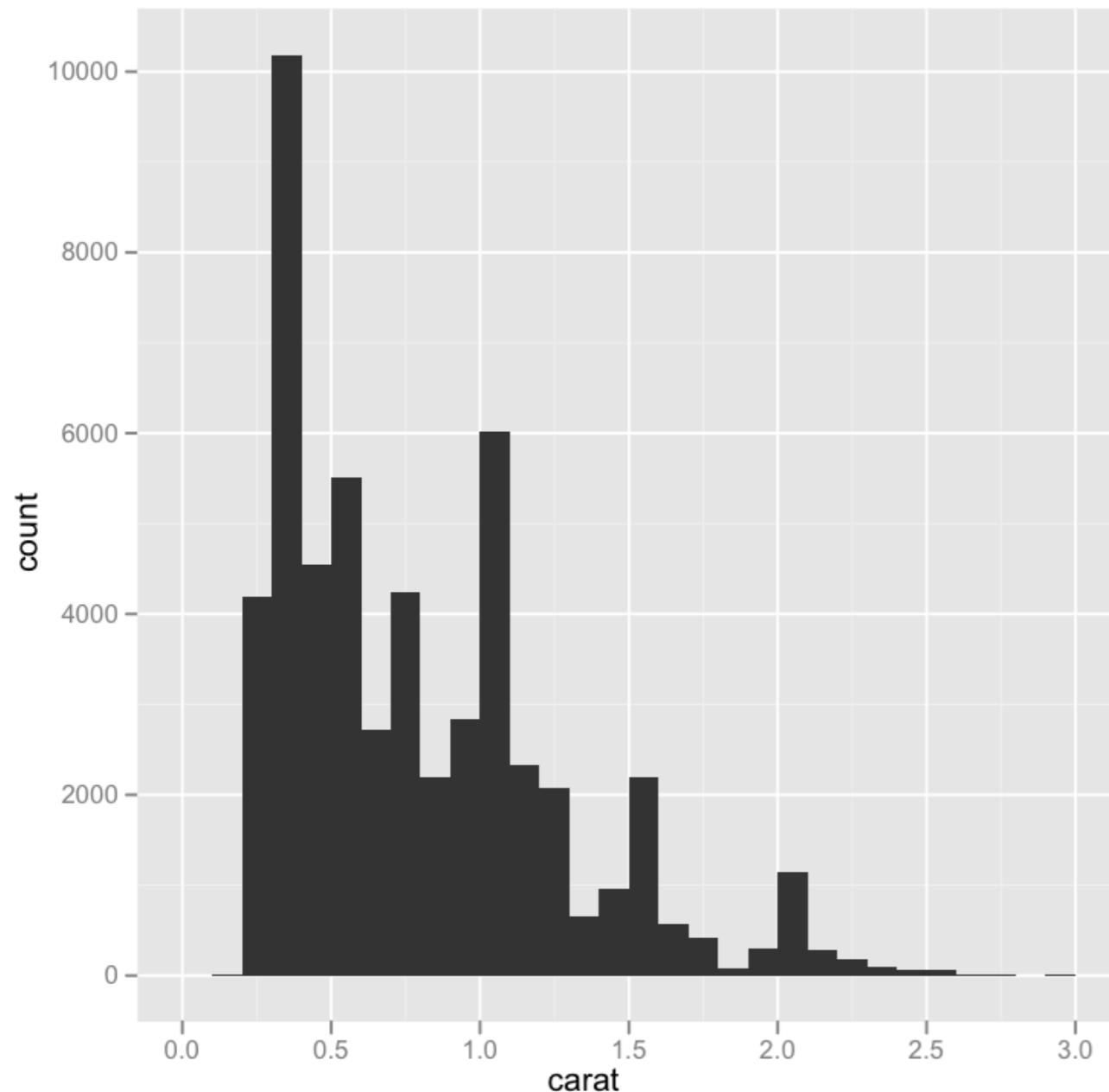


# Distributions

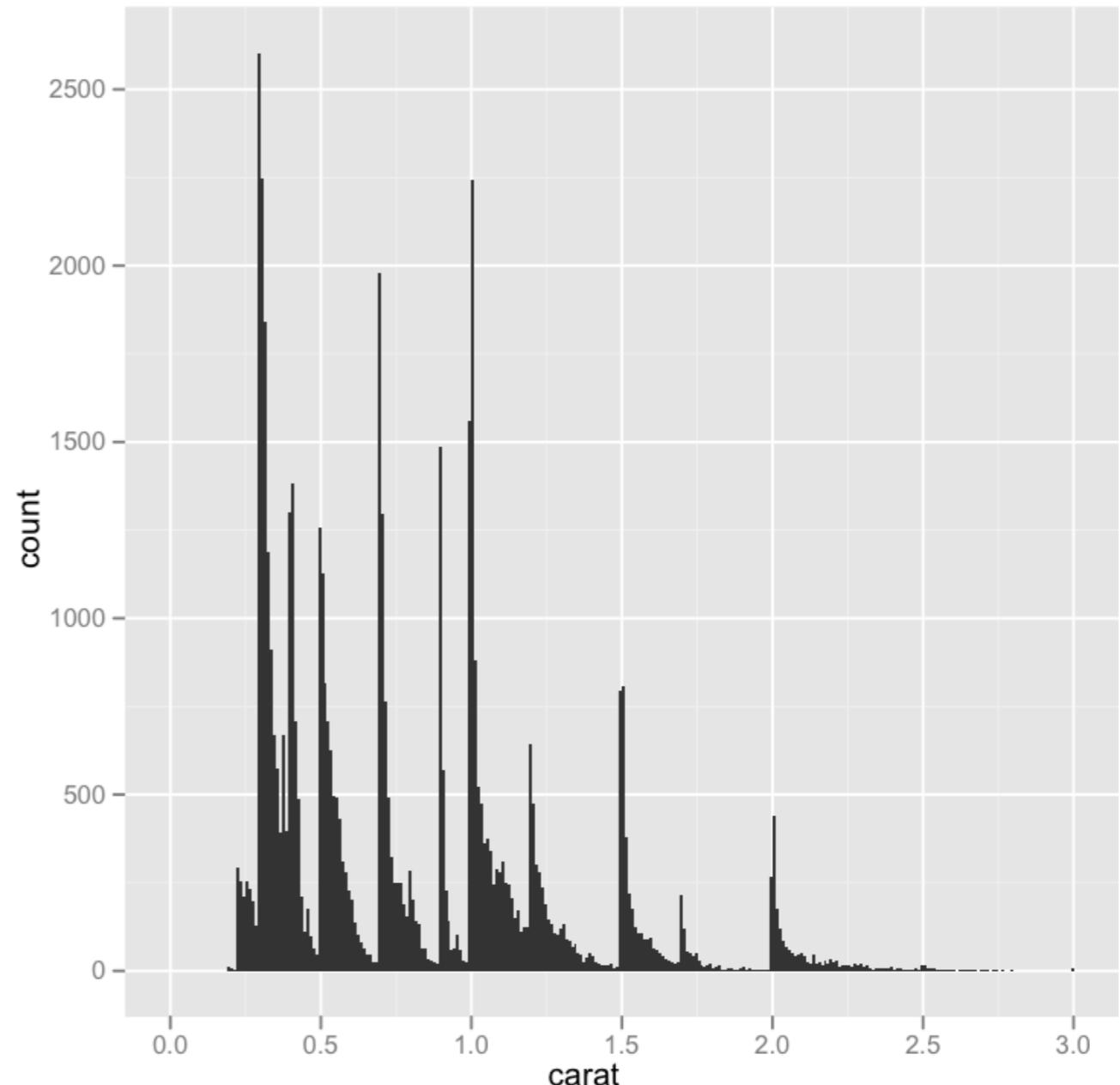
# Histogram



# Bin Width

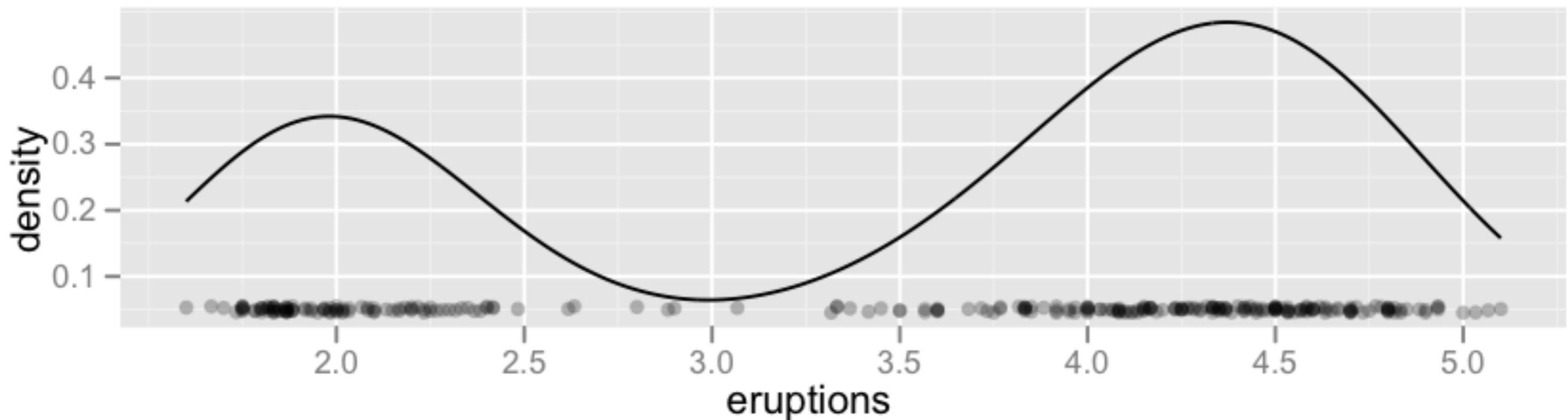


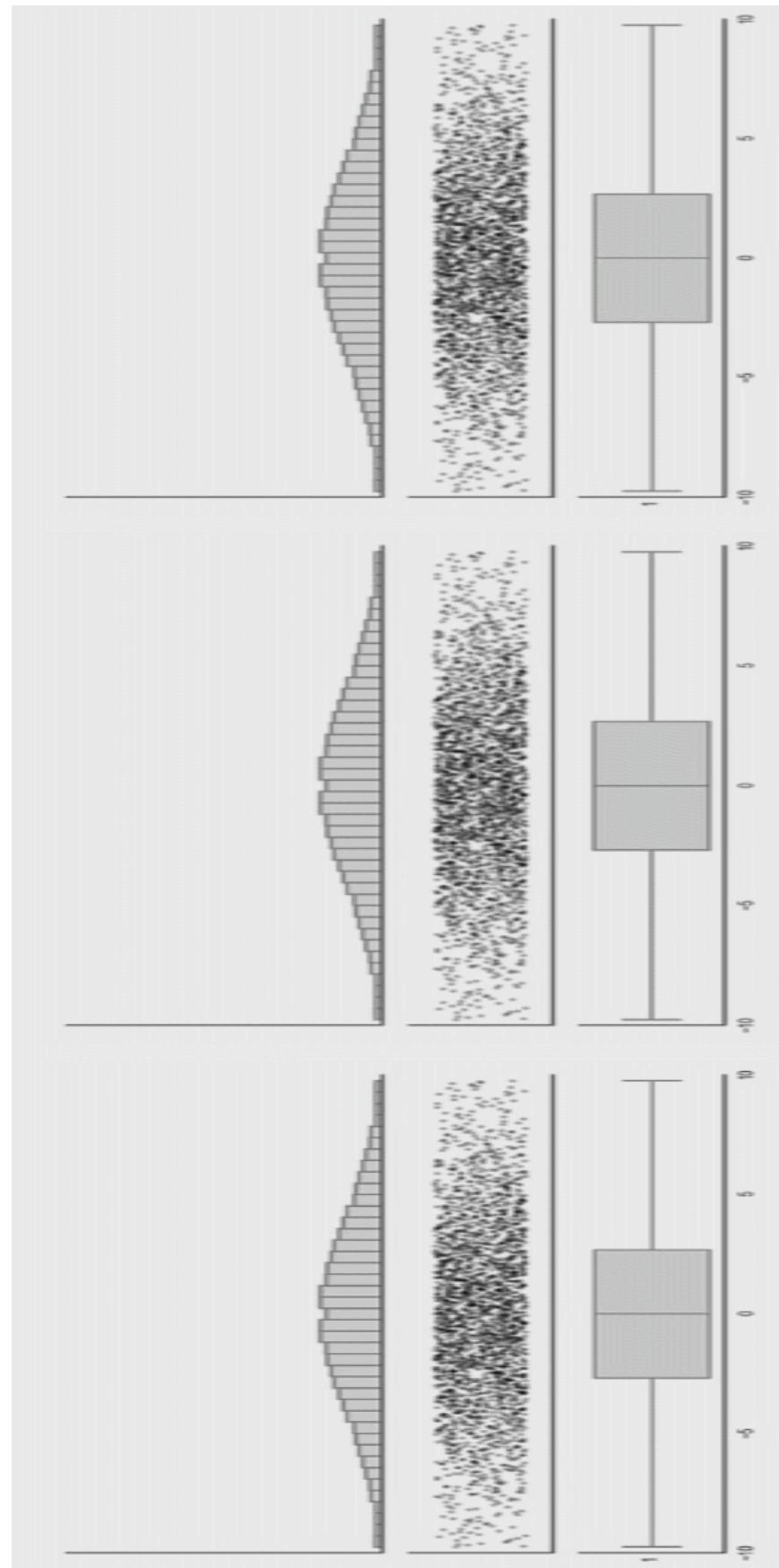
**binwidth = 0.1**



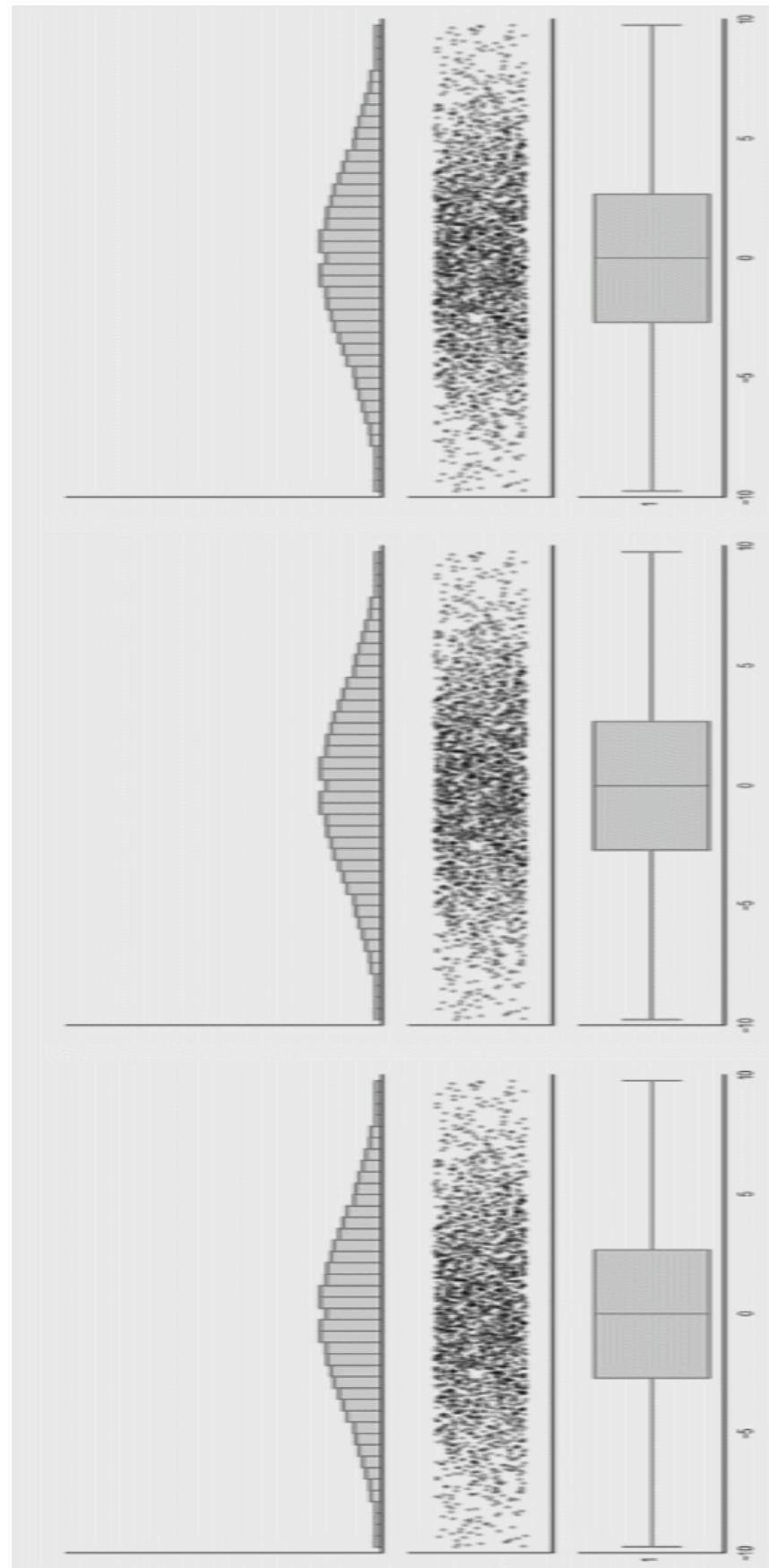
**binwidth = 0.01**

# Density Plots



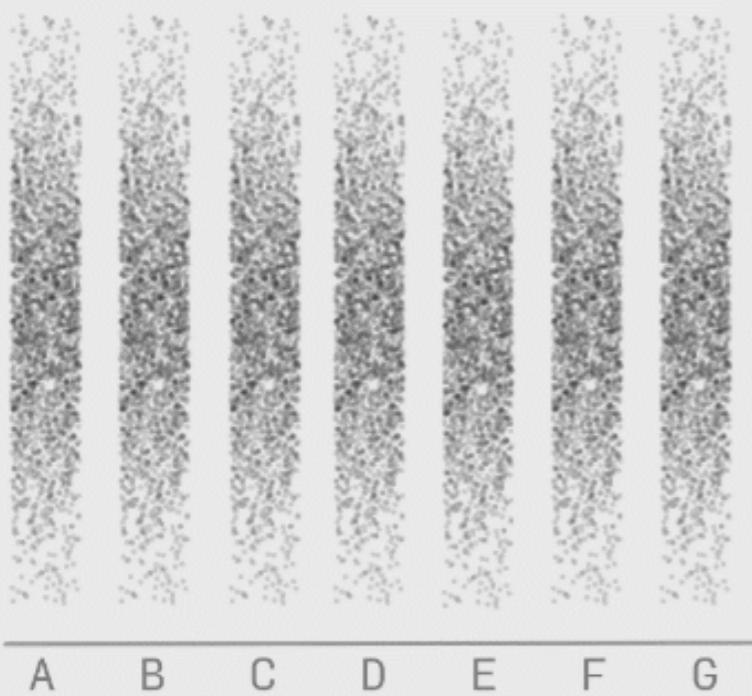


[https://  
www.autodeskresearch.com/  
publications/samestats](https://www.autodeskresearch.com/publications/samestats)

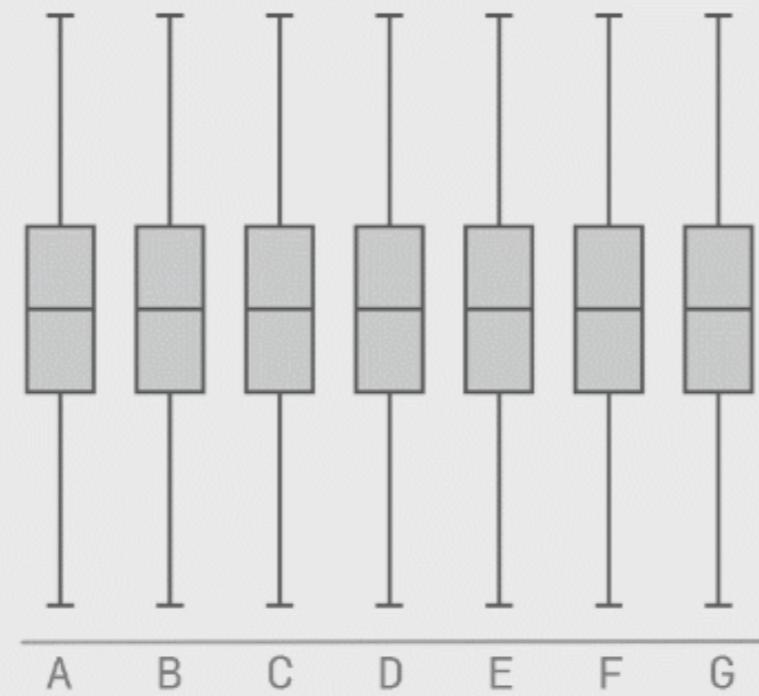


[https://  
www.autodeskresearch.com/  
publications/samestats](https://www.autodeskresearch.com/publications/samestats)

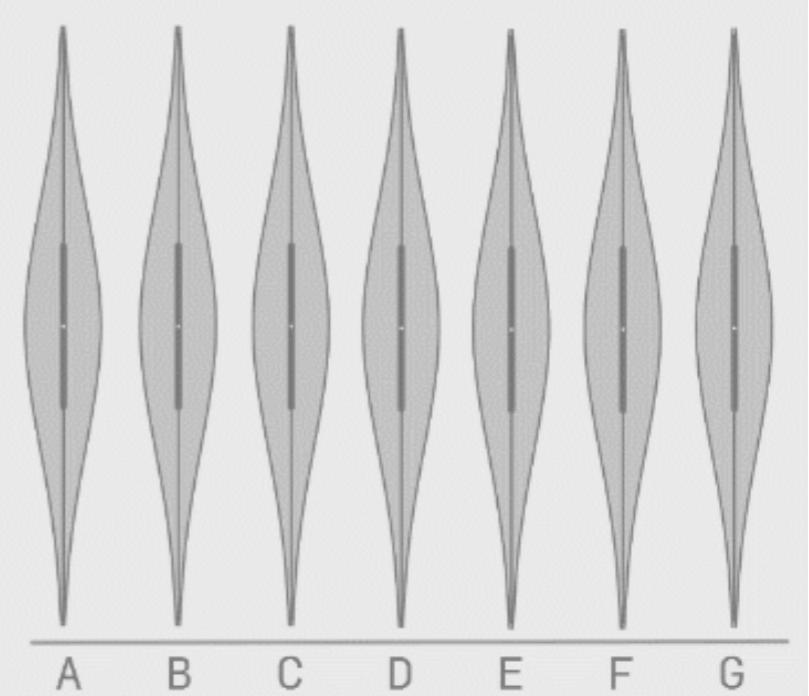
**Raw Data**



**Box-plot of the Data**

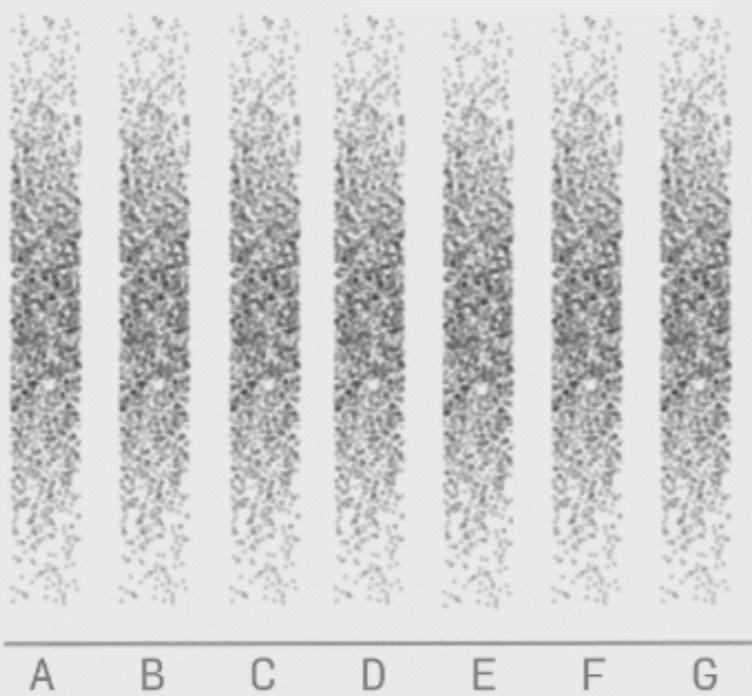


**Violin-plot of the Data**

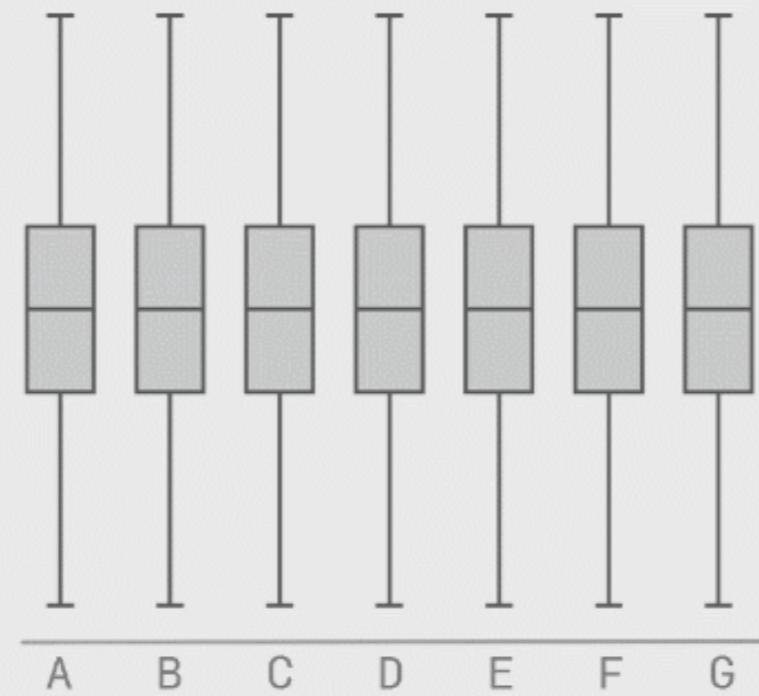


<https://www.autodeskresearch.com/publications/samestats>

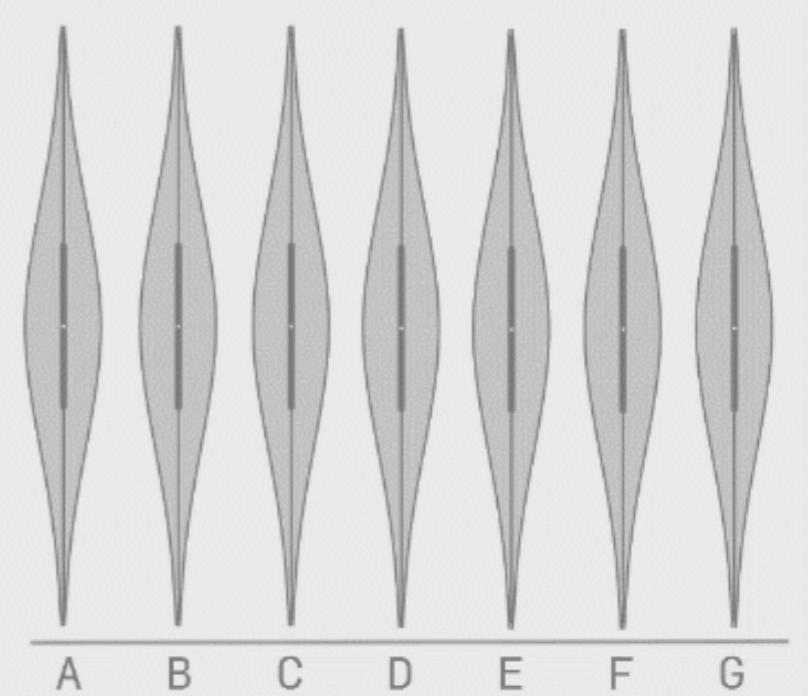
**Raw Data**



**Box-plot of the Data**

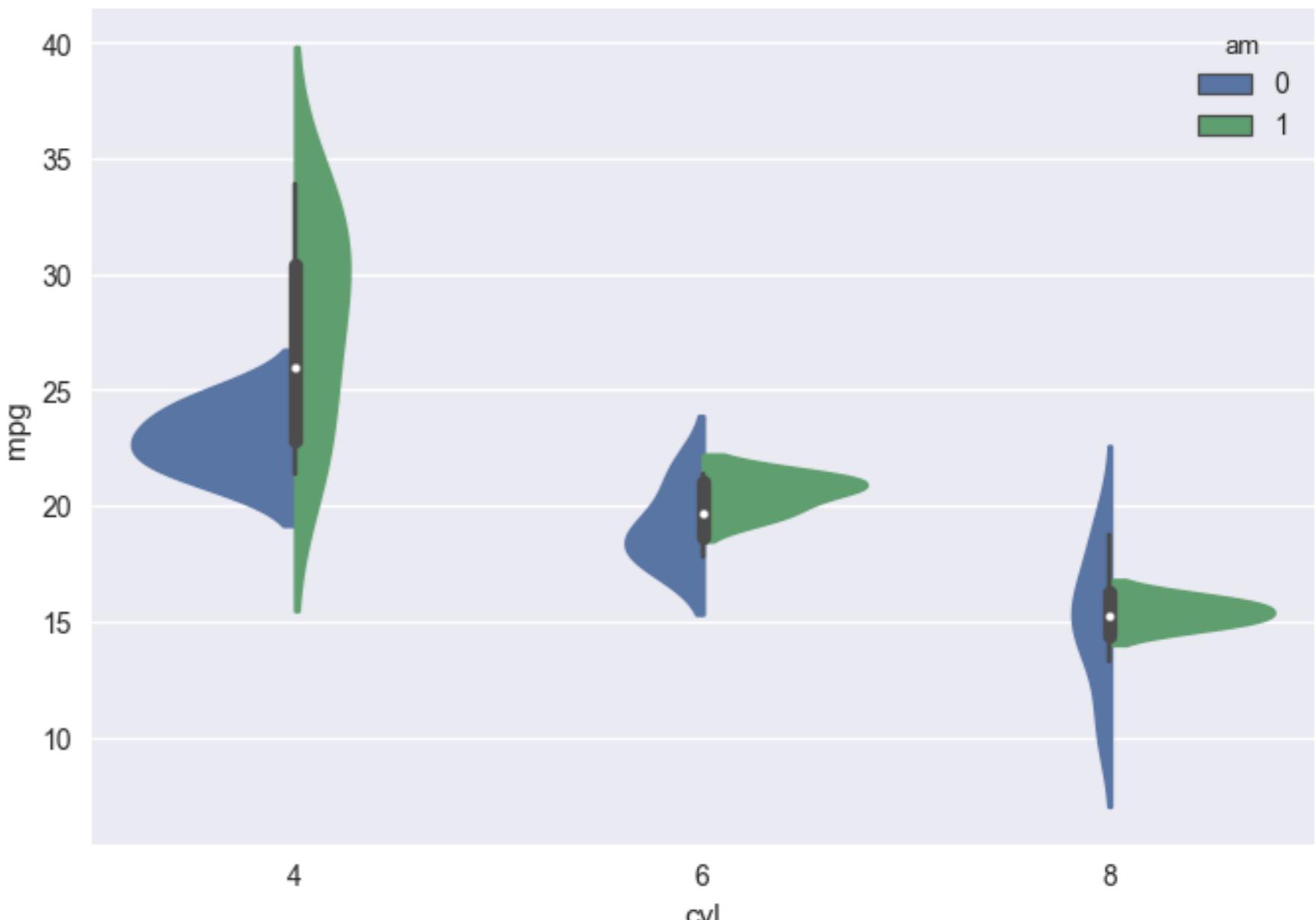
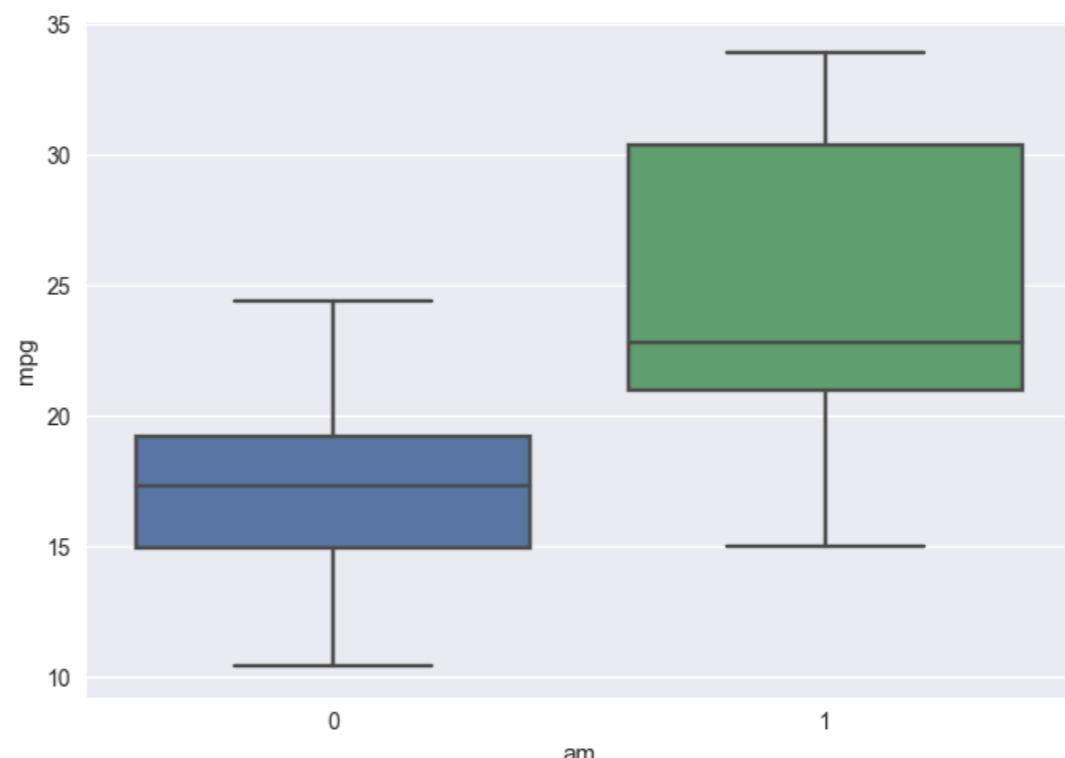
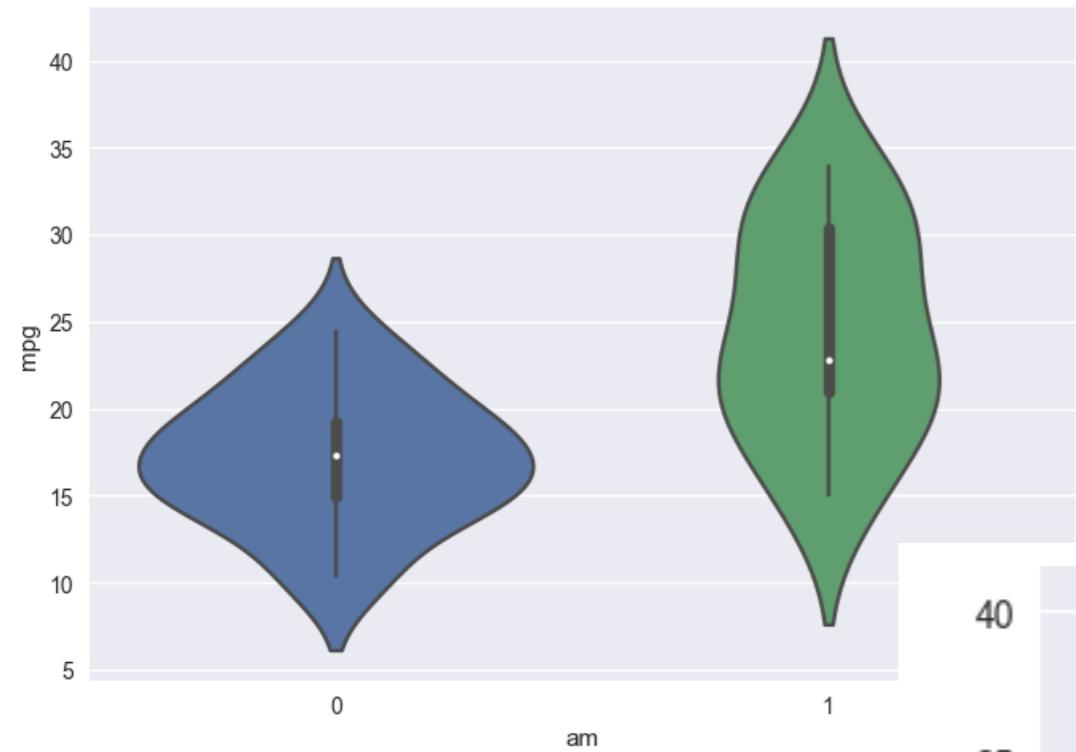


**Violin-plot of the Data**



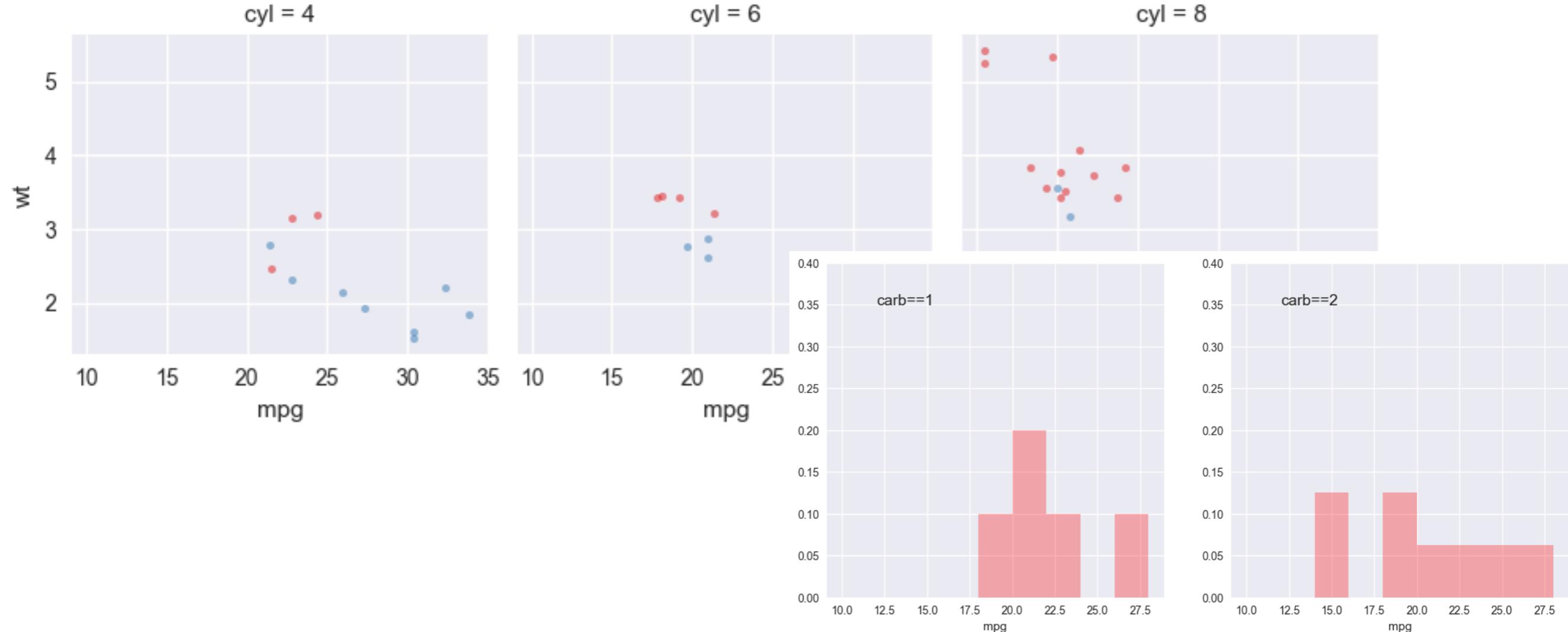
<https://www.autodeskresearch.com/publications/samestats>

# GROUP

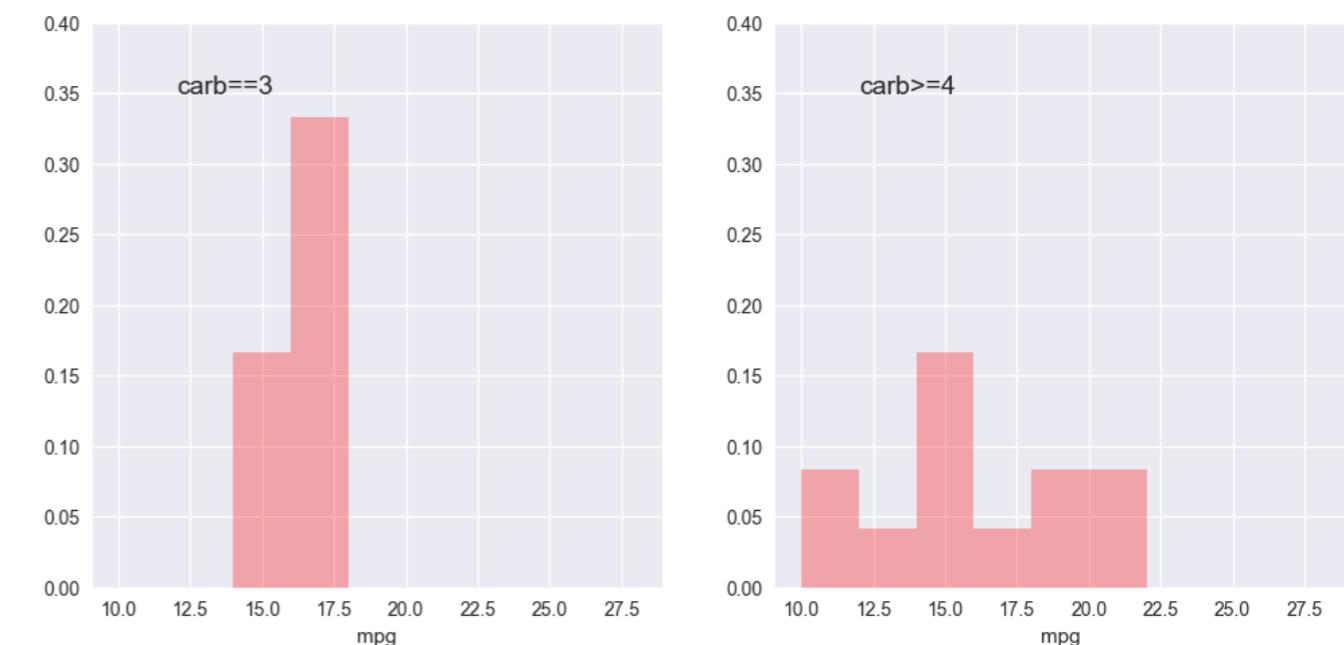


getting complex...

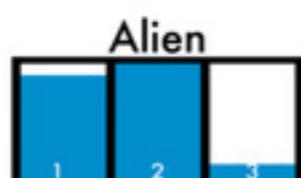
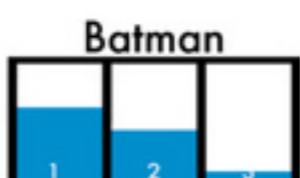
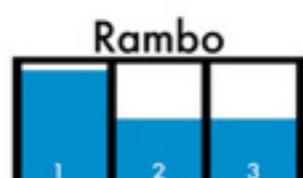
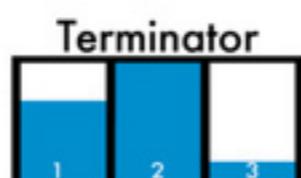
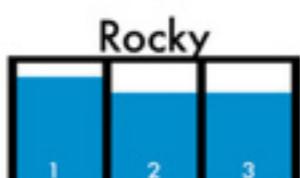
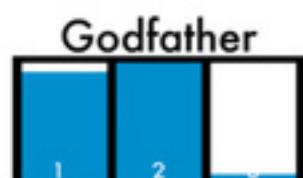
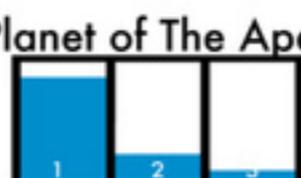
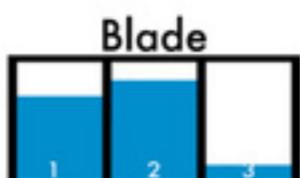
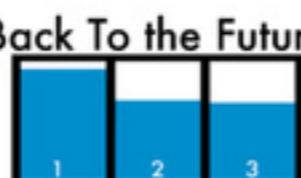
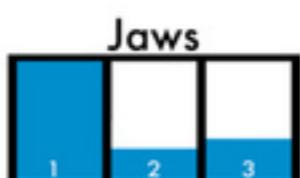
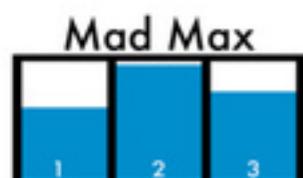
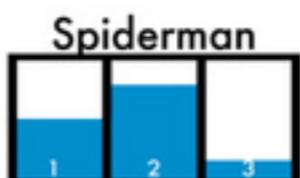
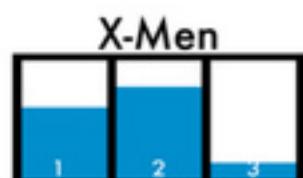
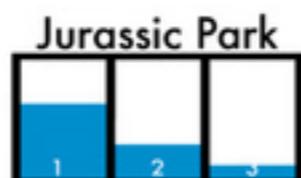
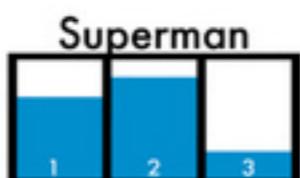
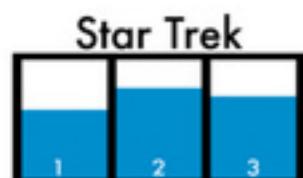
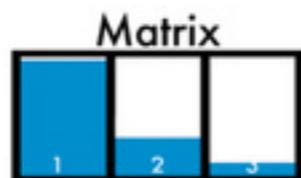
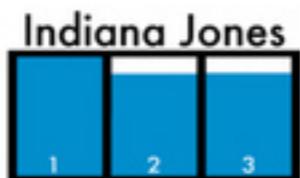
# Faceting and Small Multiples



Use seaborn or  
multiple plots in matplotlib

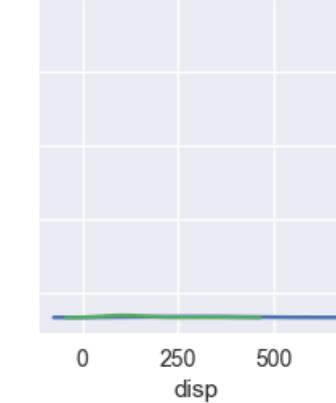
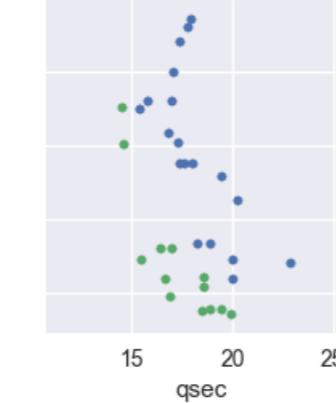
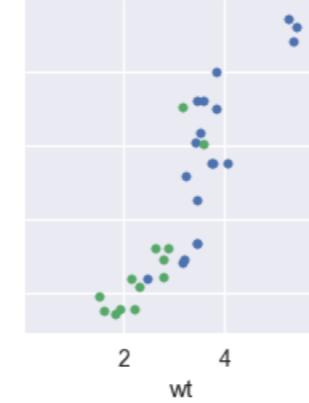
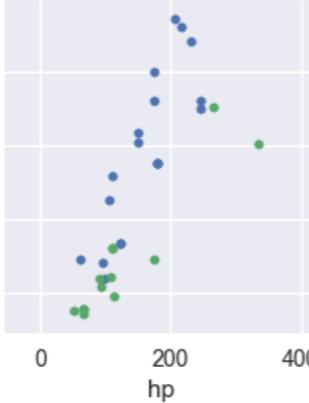
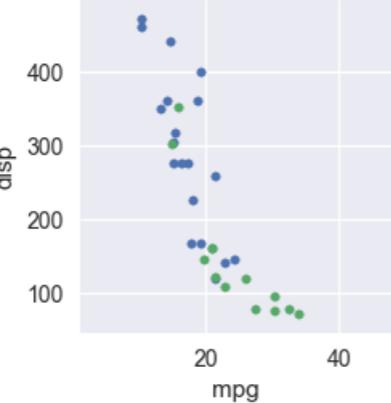
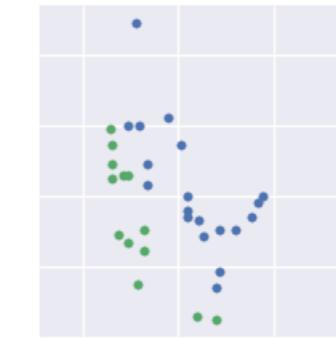
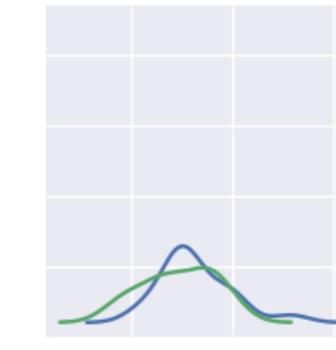
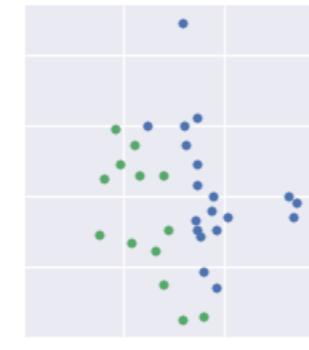
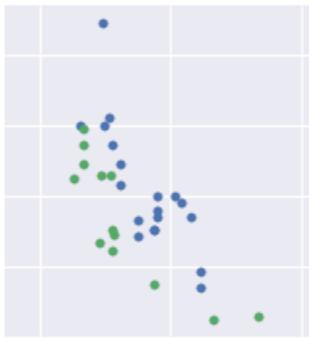
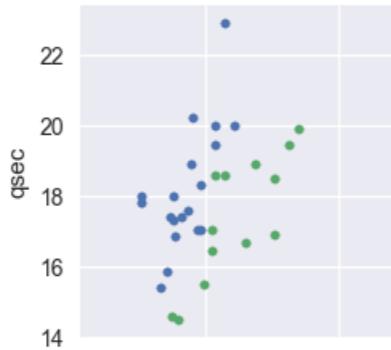
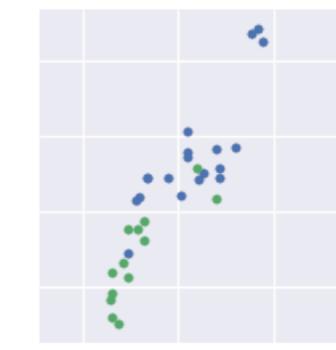
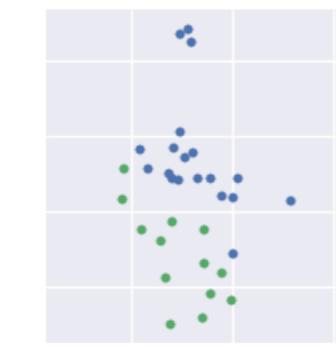
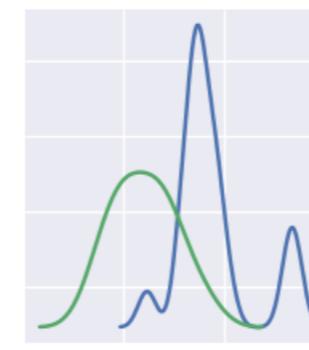
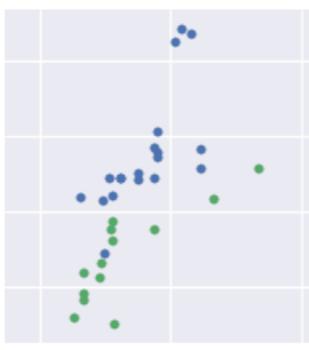
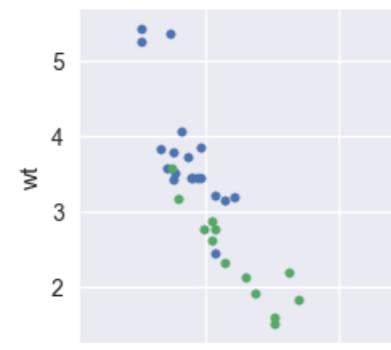
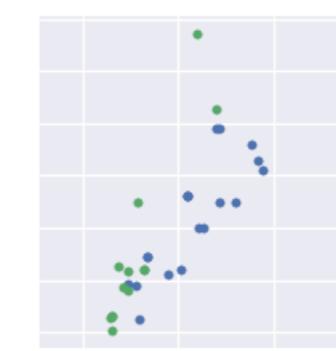
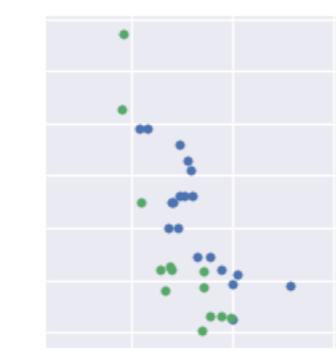
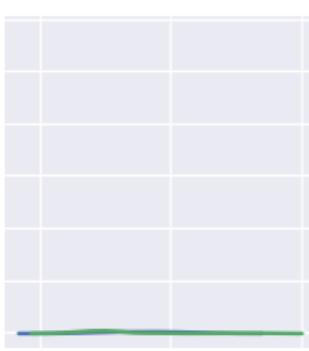
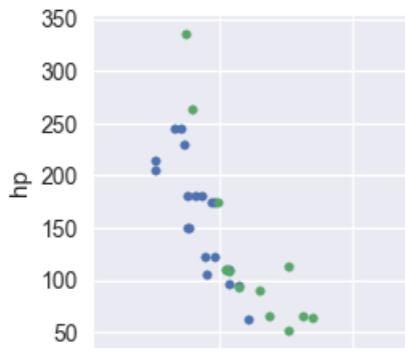
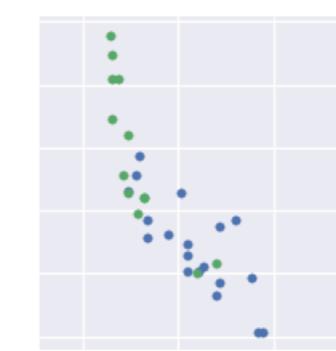
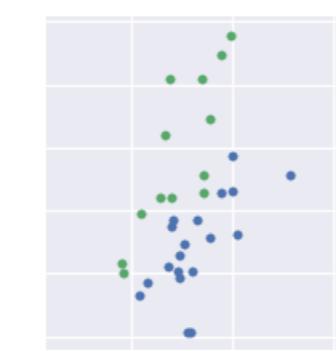
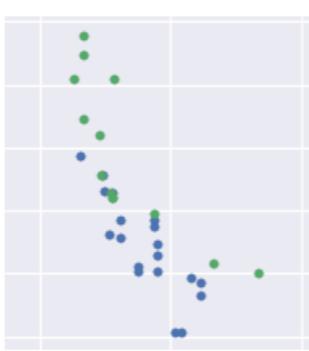
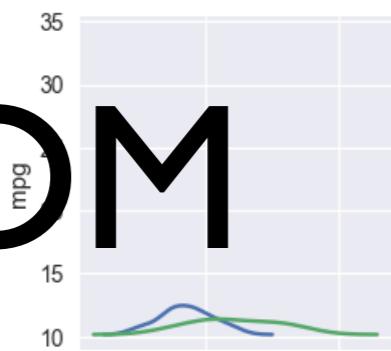


# THE TRILOGY METER



Small multiples

# SPLOM



# **Design Exercise**

## **Hands-On Exercise**

# How do you feel about doing science?

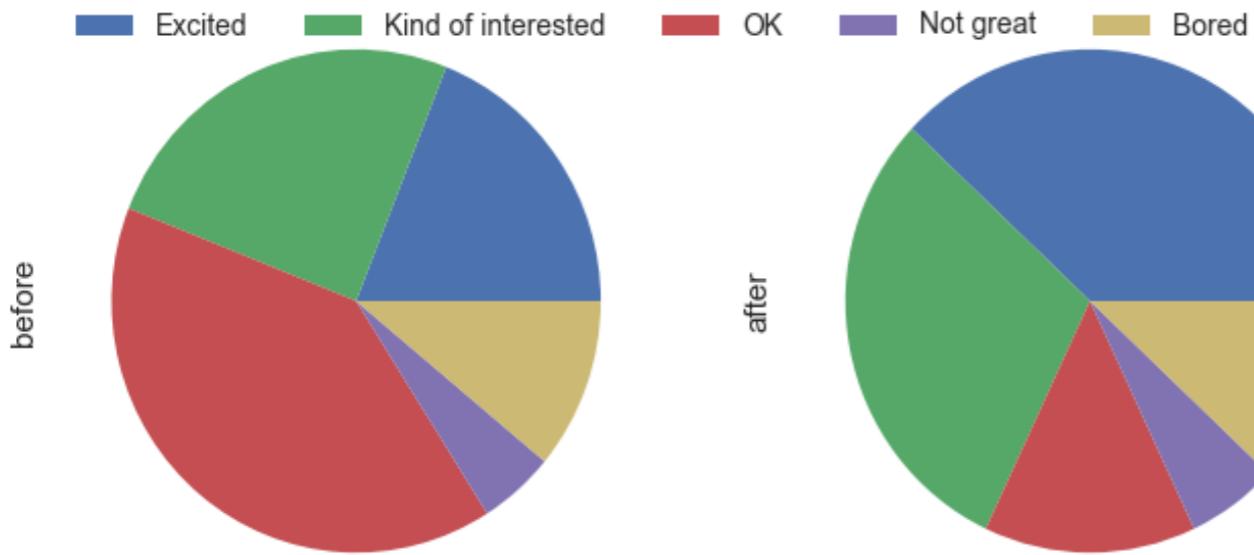
Table

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

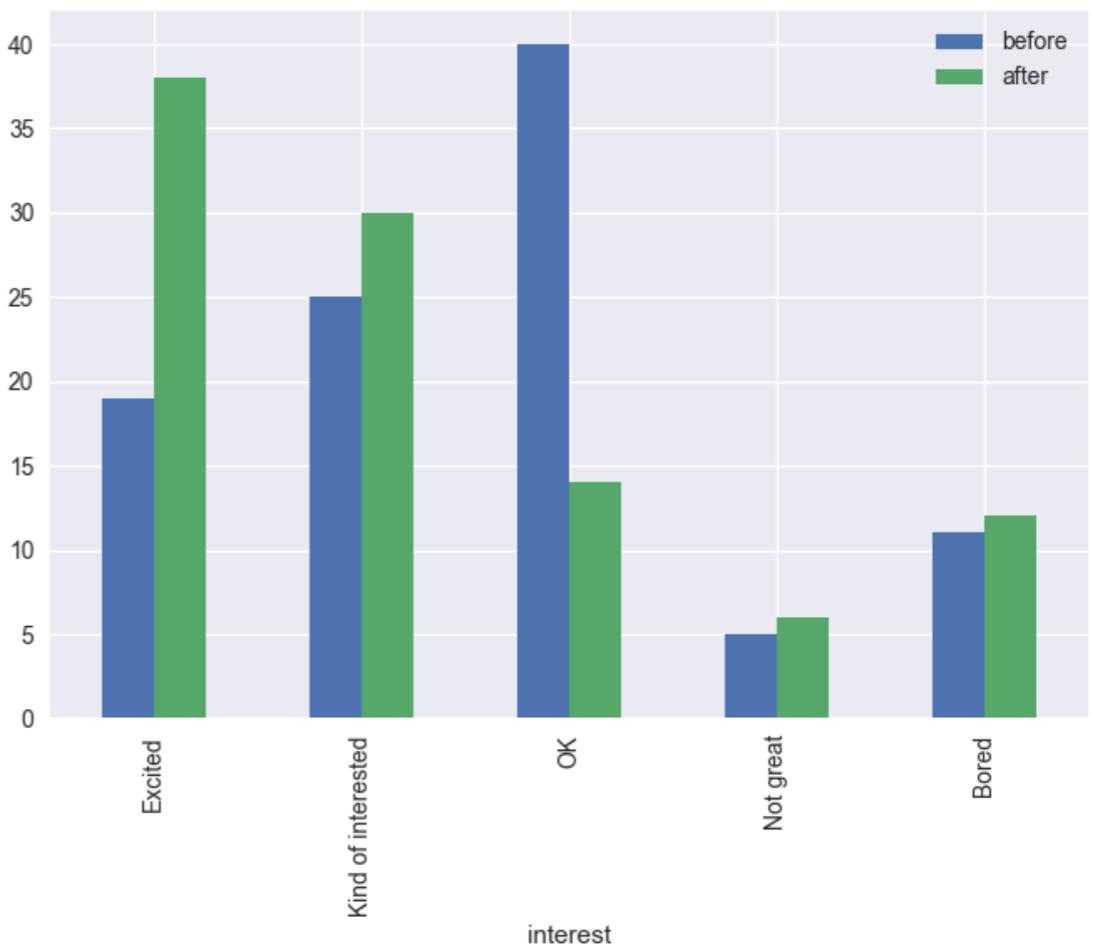
**Come up with multiple visualizations.**

**Pen and Paper Only.**

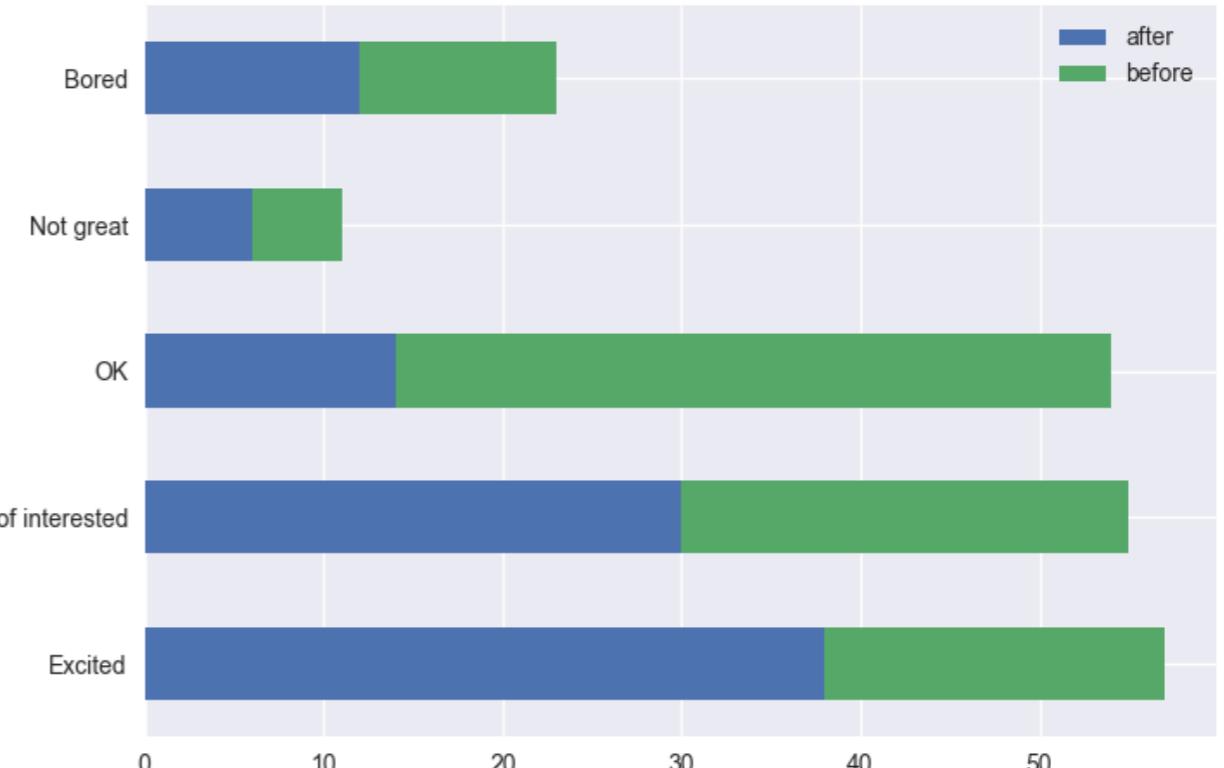
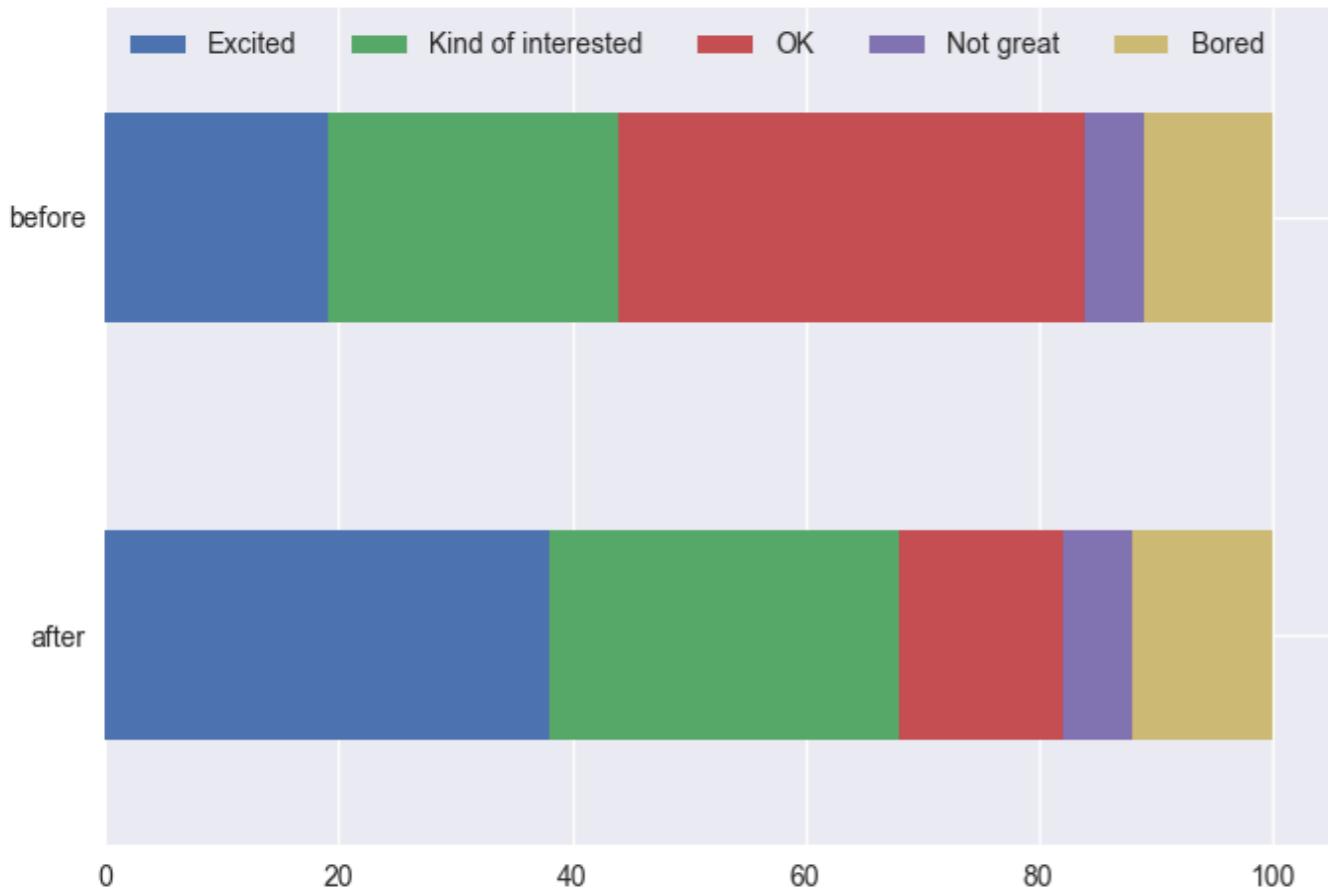


# Pie

## Side by side bar

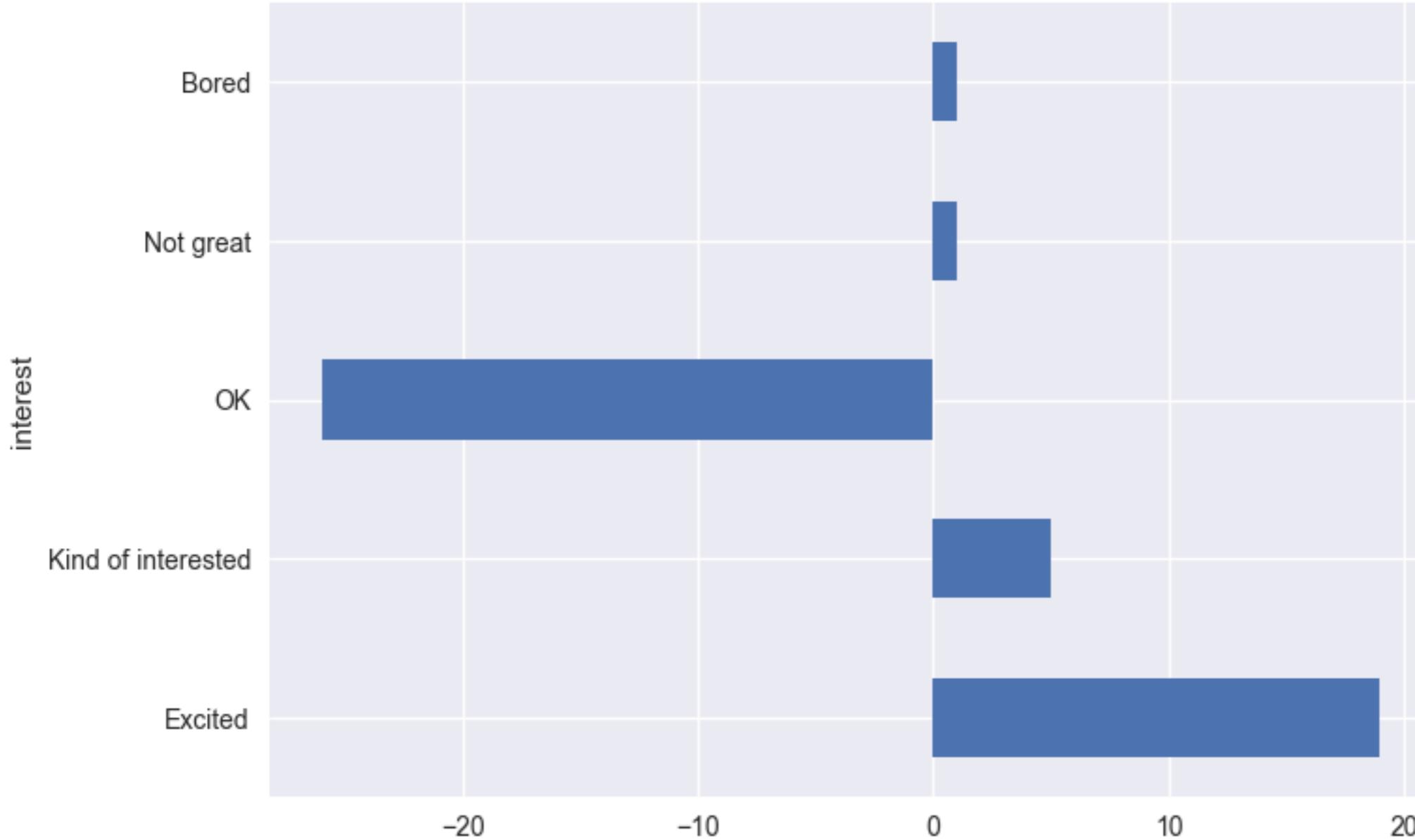


# Stacked bar, not very useful

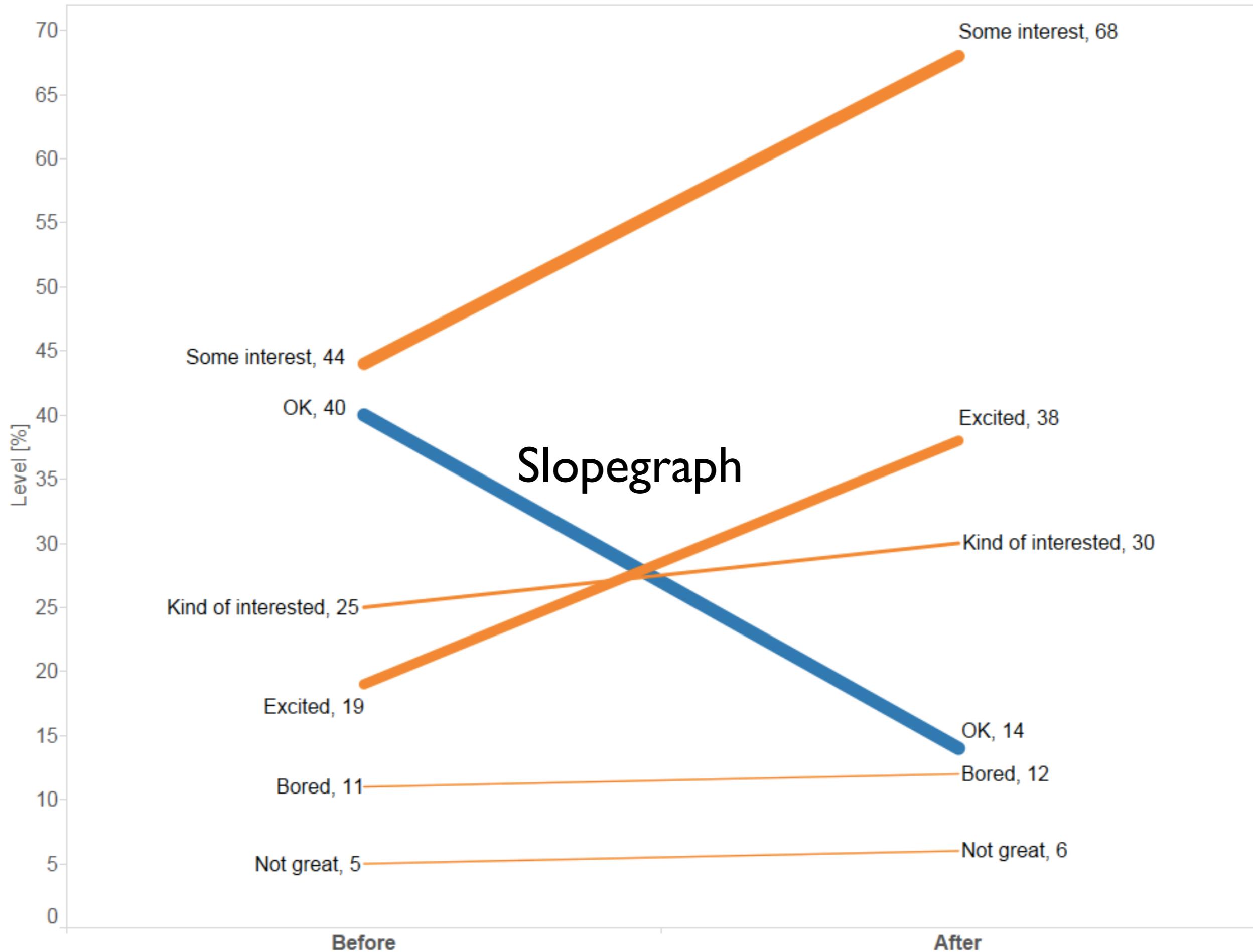


## Data Transposed Bar Chart

# Difference Bar Chart



## How do you feel about doing science?

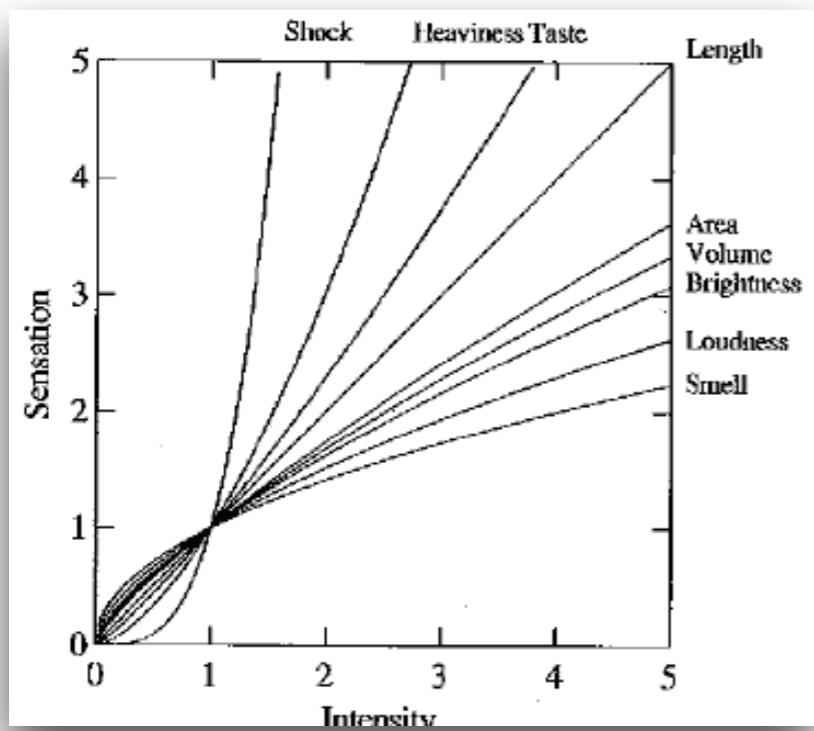


After the pilot program,

**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

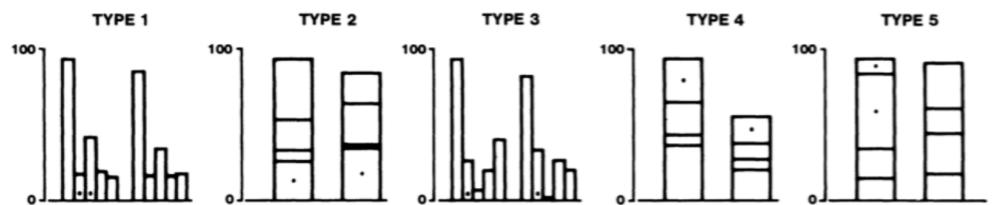


Figure 4. Graphs from position-length experiment.

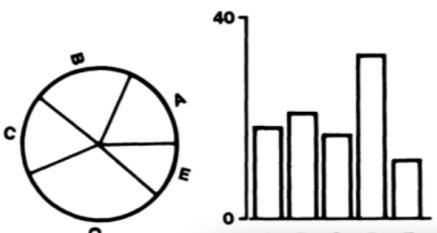
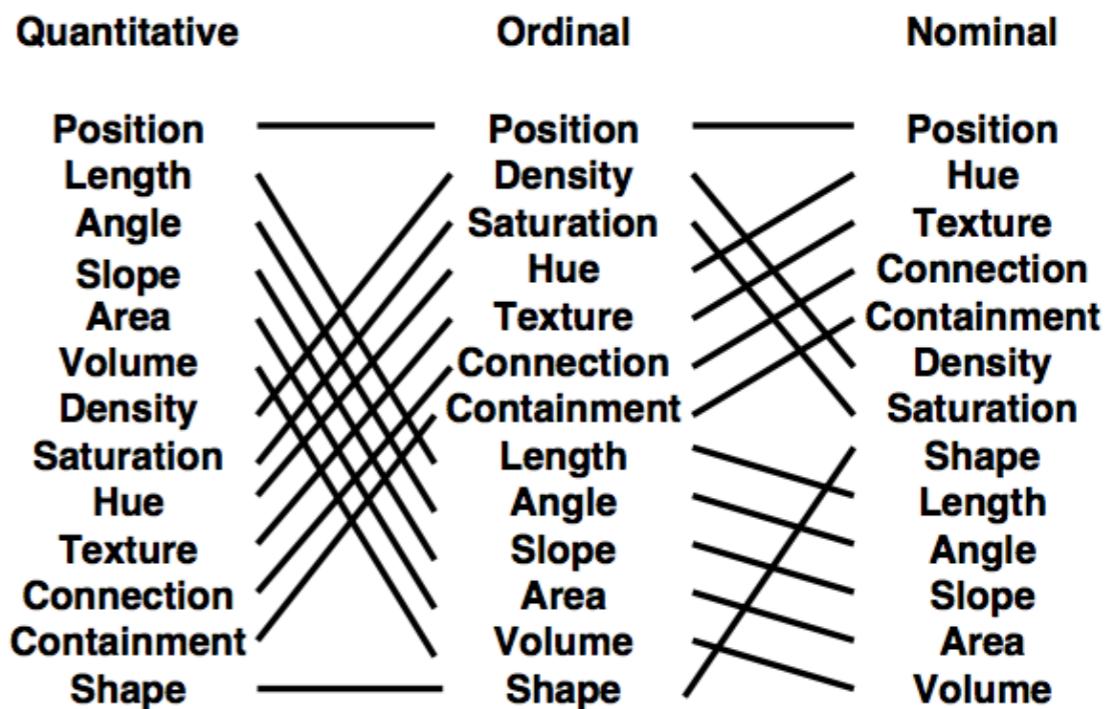
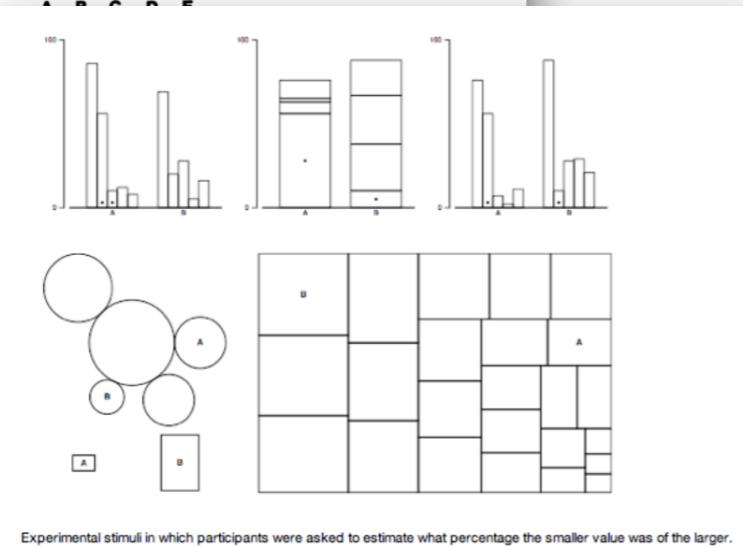


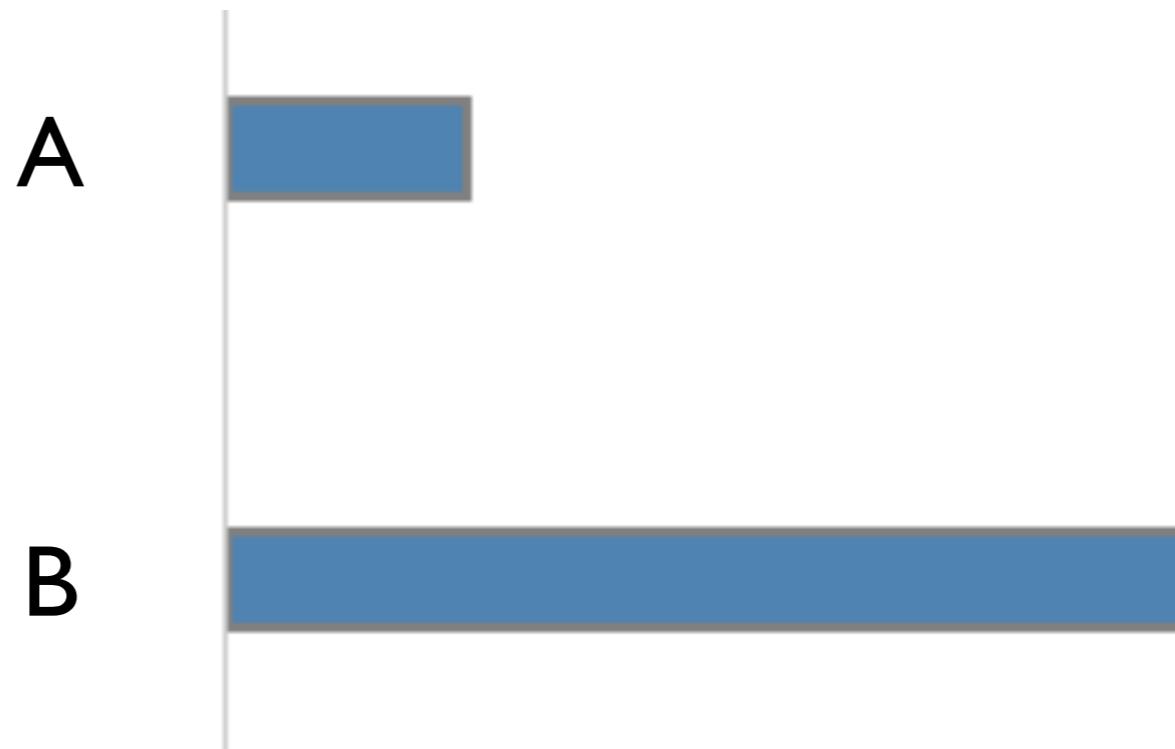
Figure 3. Graphs from position-length experiment.

Cleveland / McGill, 1984

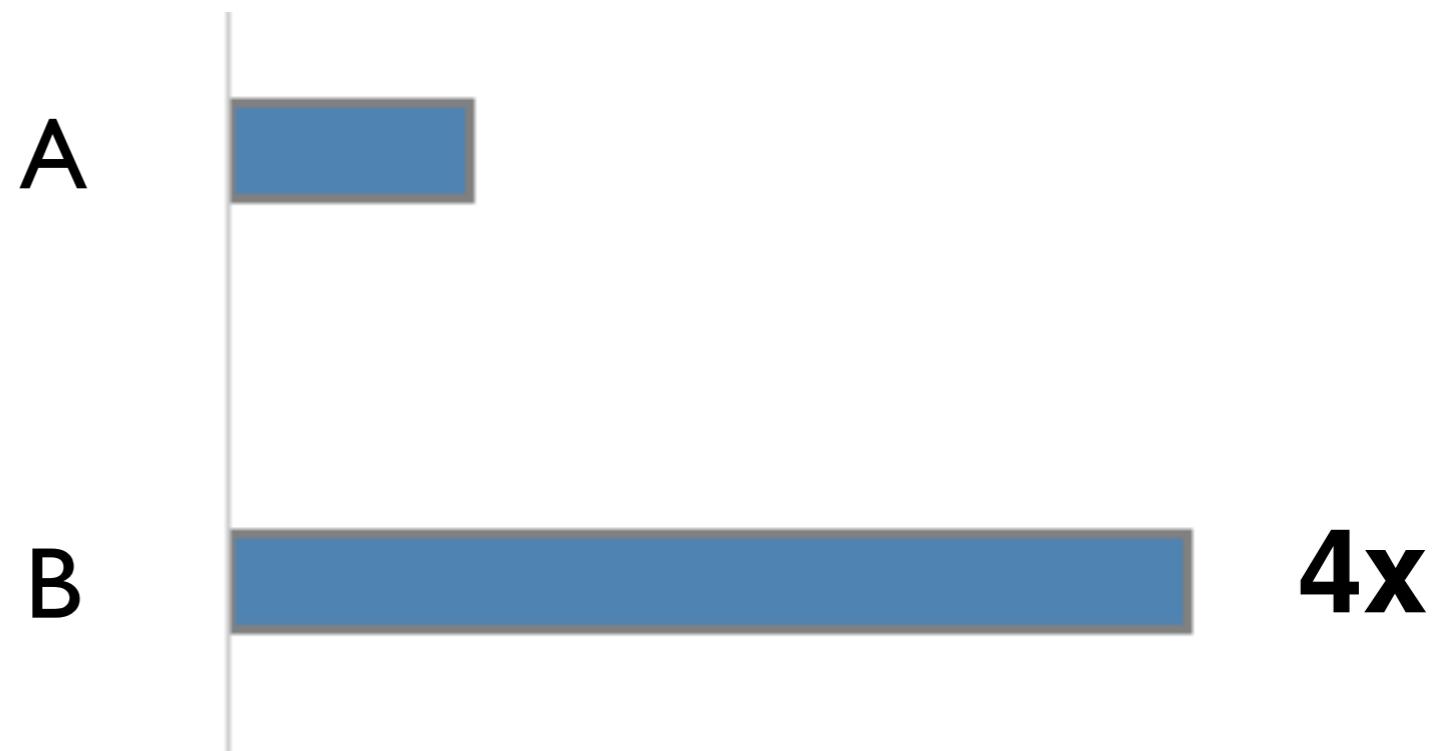


J. Mackinlay, 1986

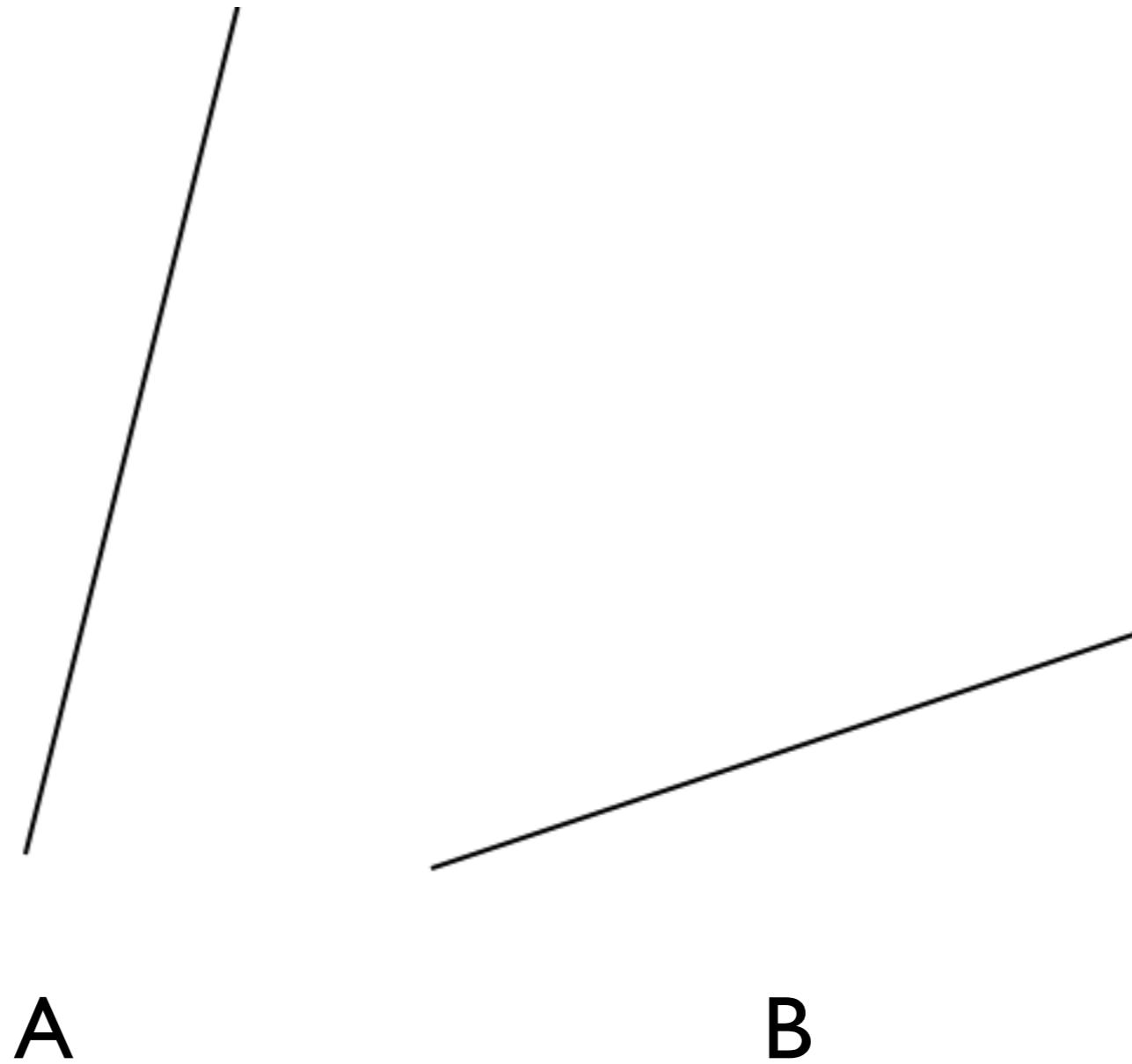
# How much longer?



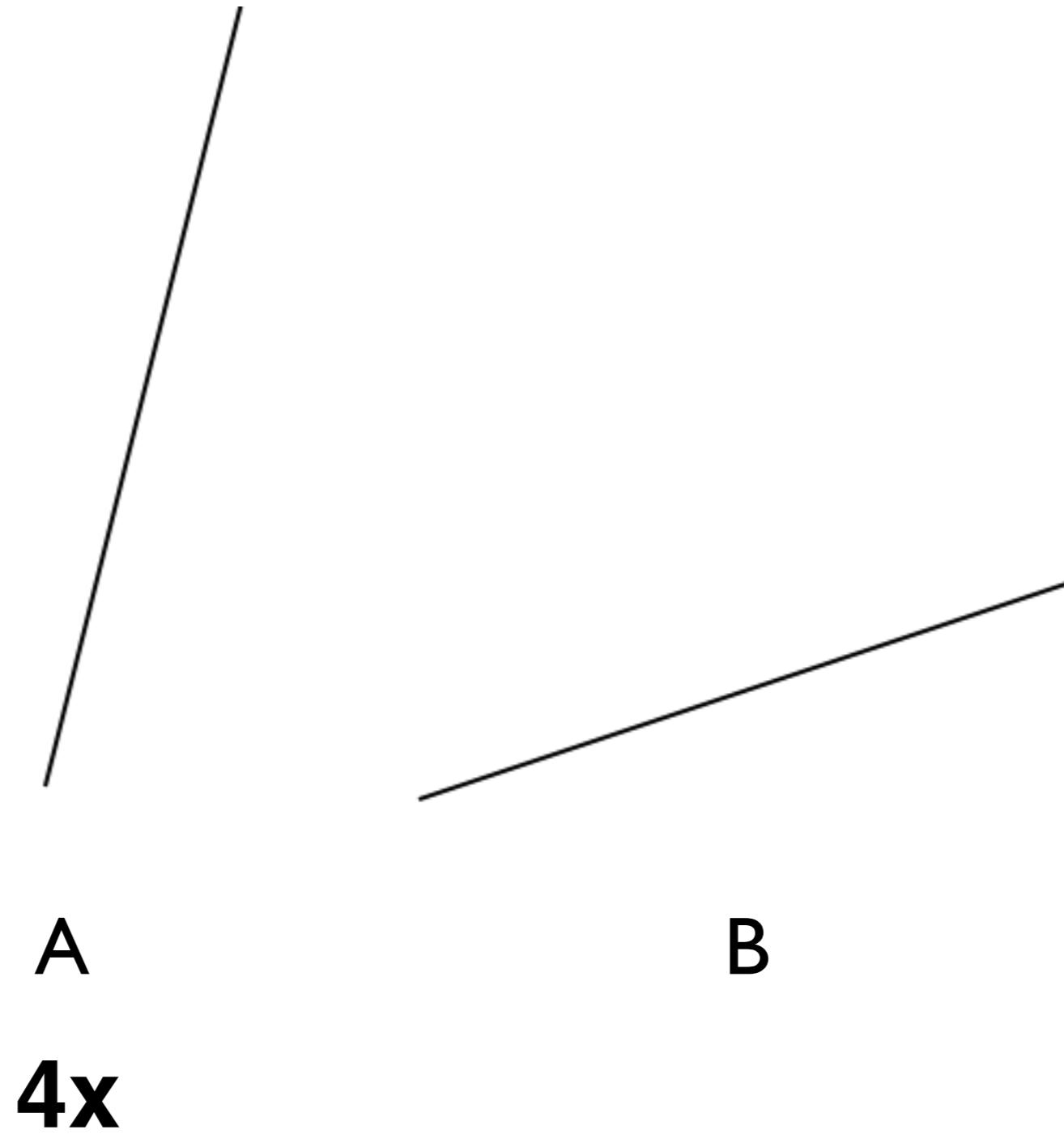
# How much longer?



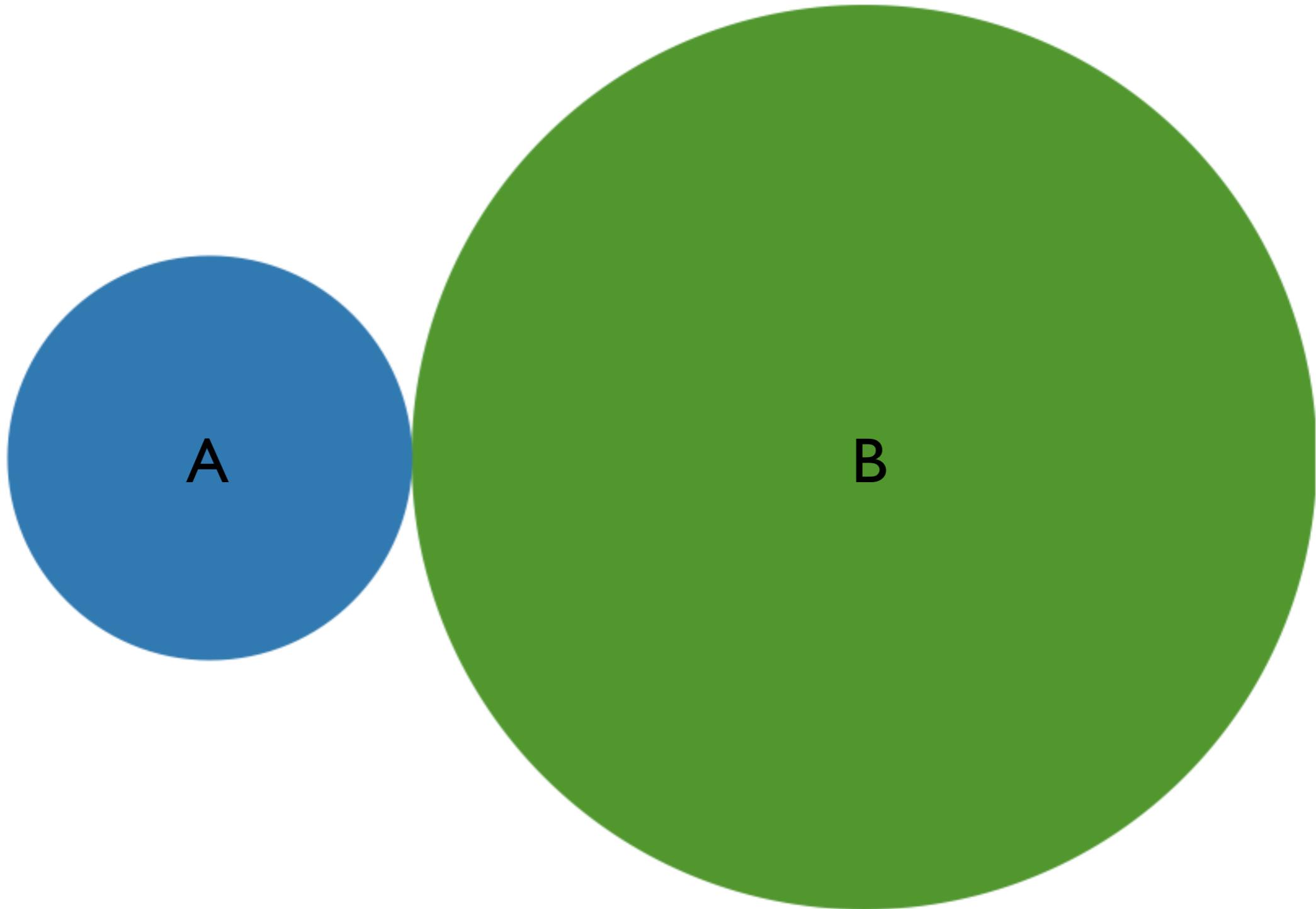
# How much steeper slope?



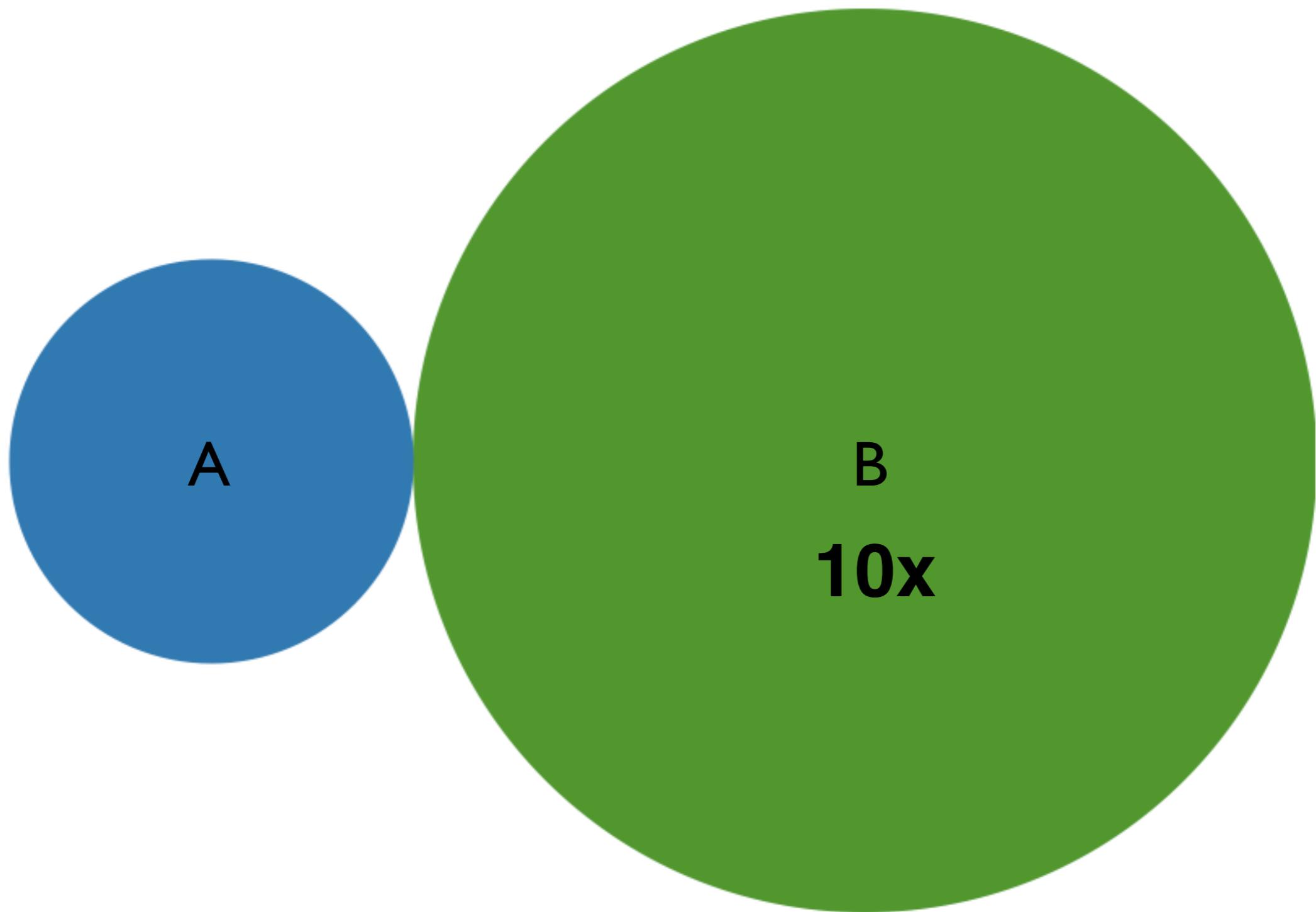
# How much steeper slope?



# How much larger area?



# How much larger area?



# How much darker?



A



B

# How much darker?



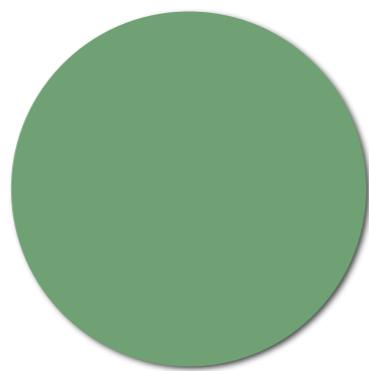
A



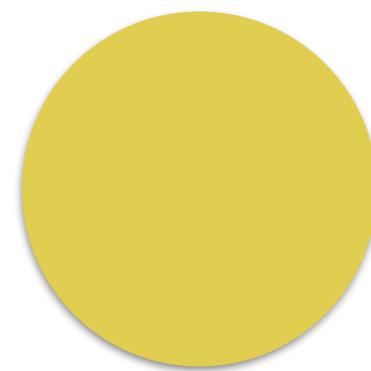
B

**2x**

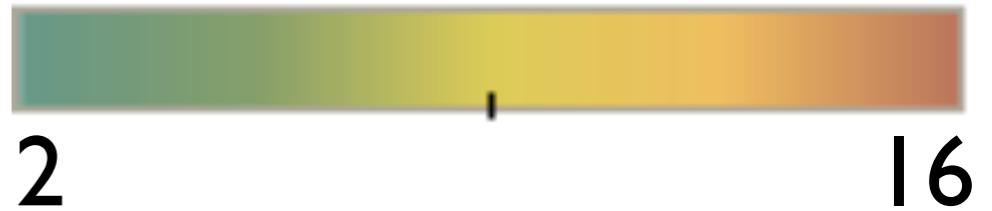
# How much bigger value?



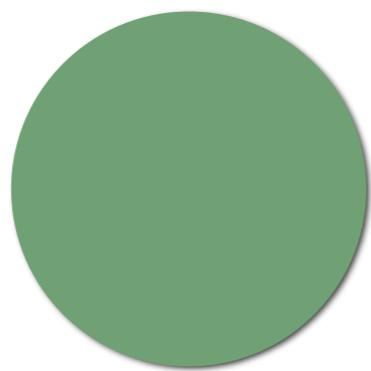
A



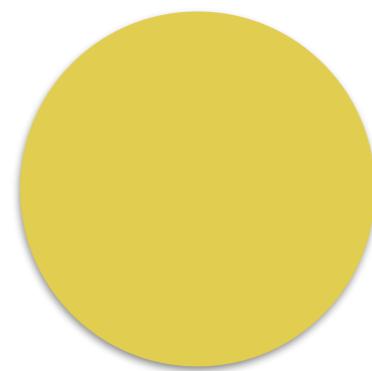
B



# How much bigger value?

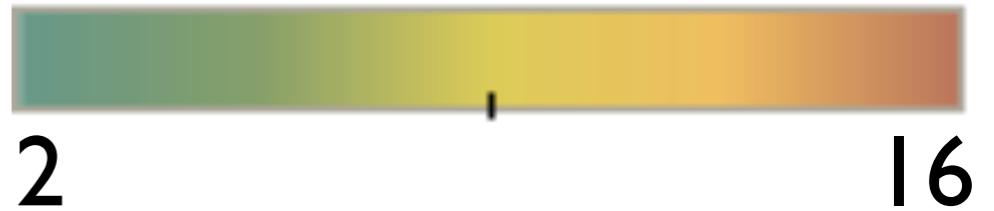


A



B

4x



Most  
Efficient

Position



Length



Slope



Angle

Area



Intensity



Least  
Efficient

Color



Shape



Most  
Efficient



Least  
Efficient

Position



Length



Slope



Angle



Area



Intensity



Color



Shape

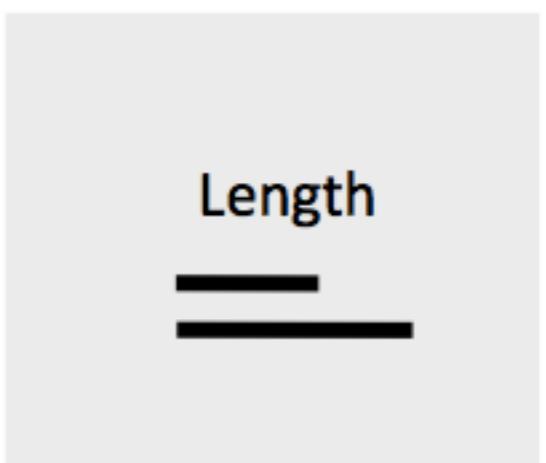
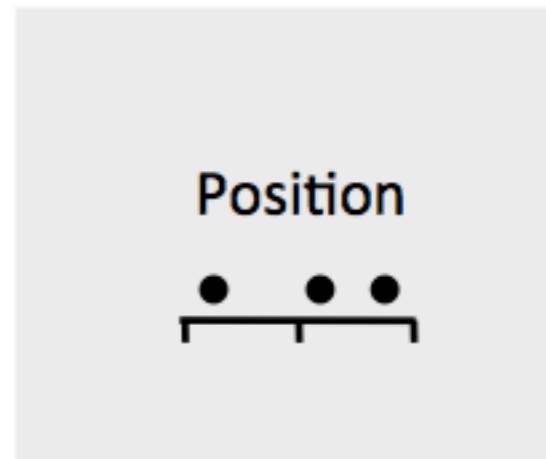
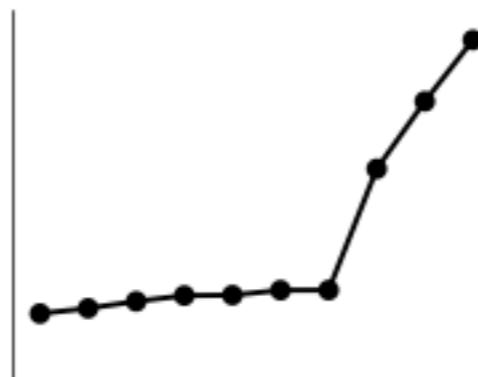


Quantitative

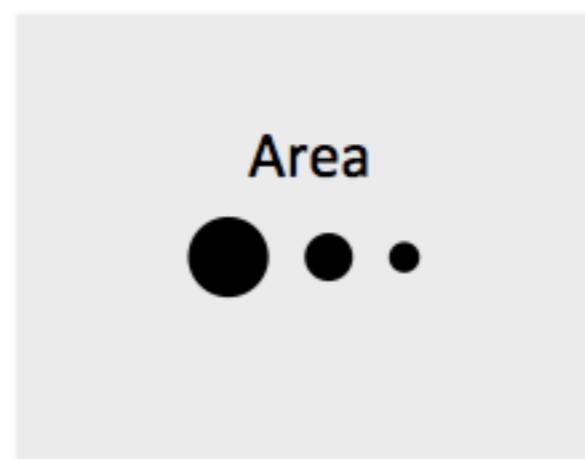
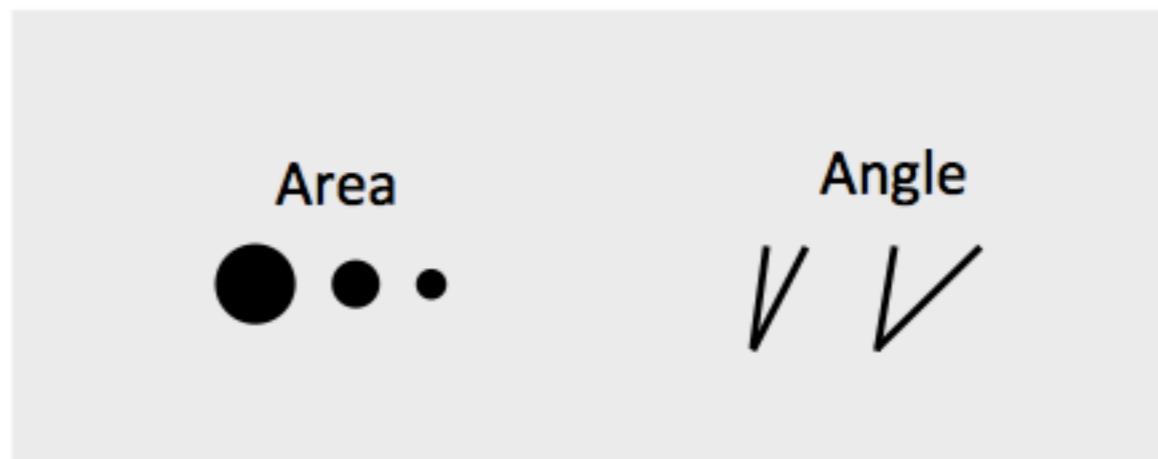
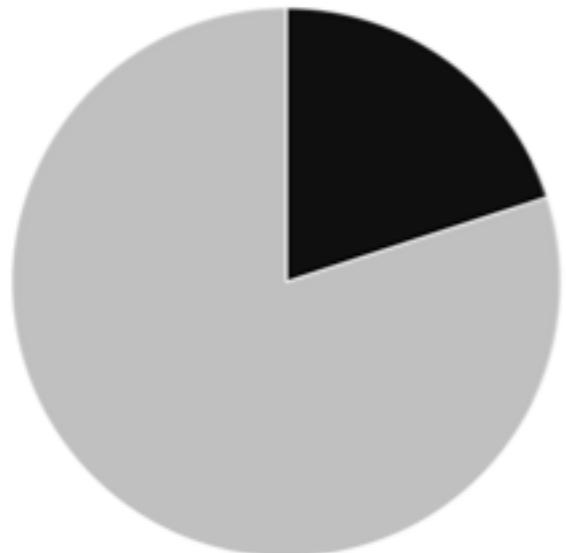
Ordered

Categories

# Most Effective

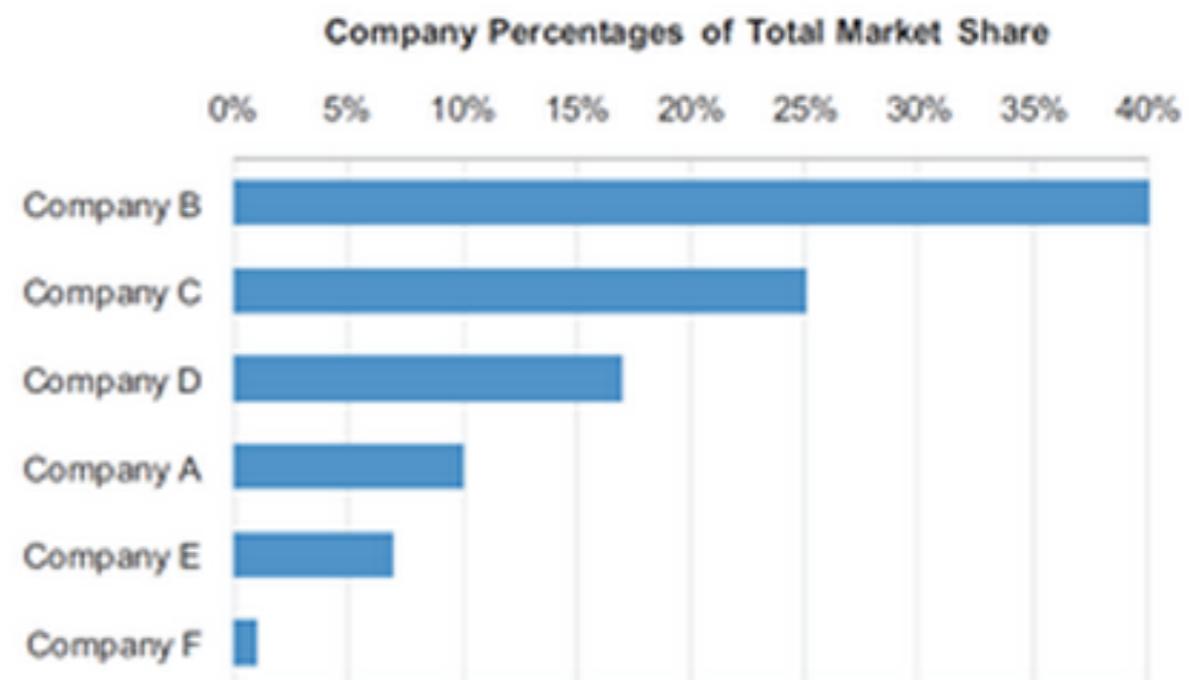
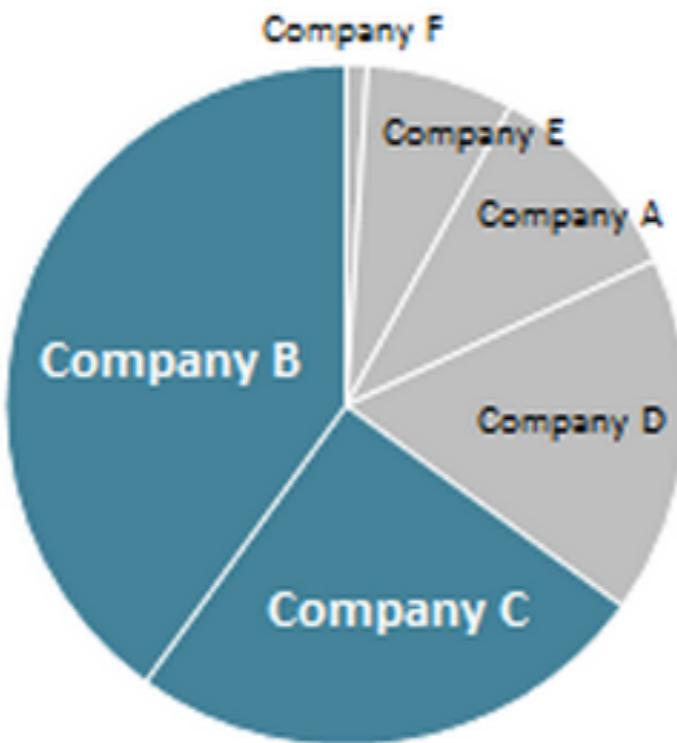


# Less Effective



# Pie vs. Bar Charts

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



# Least Effective

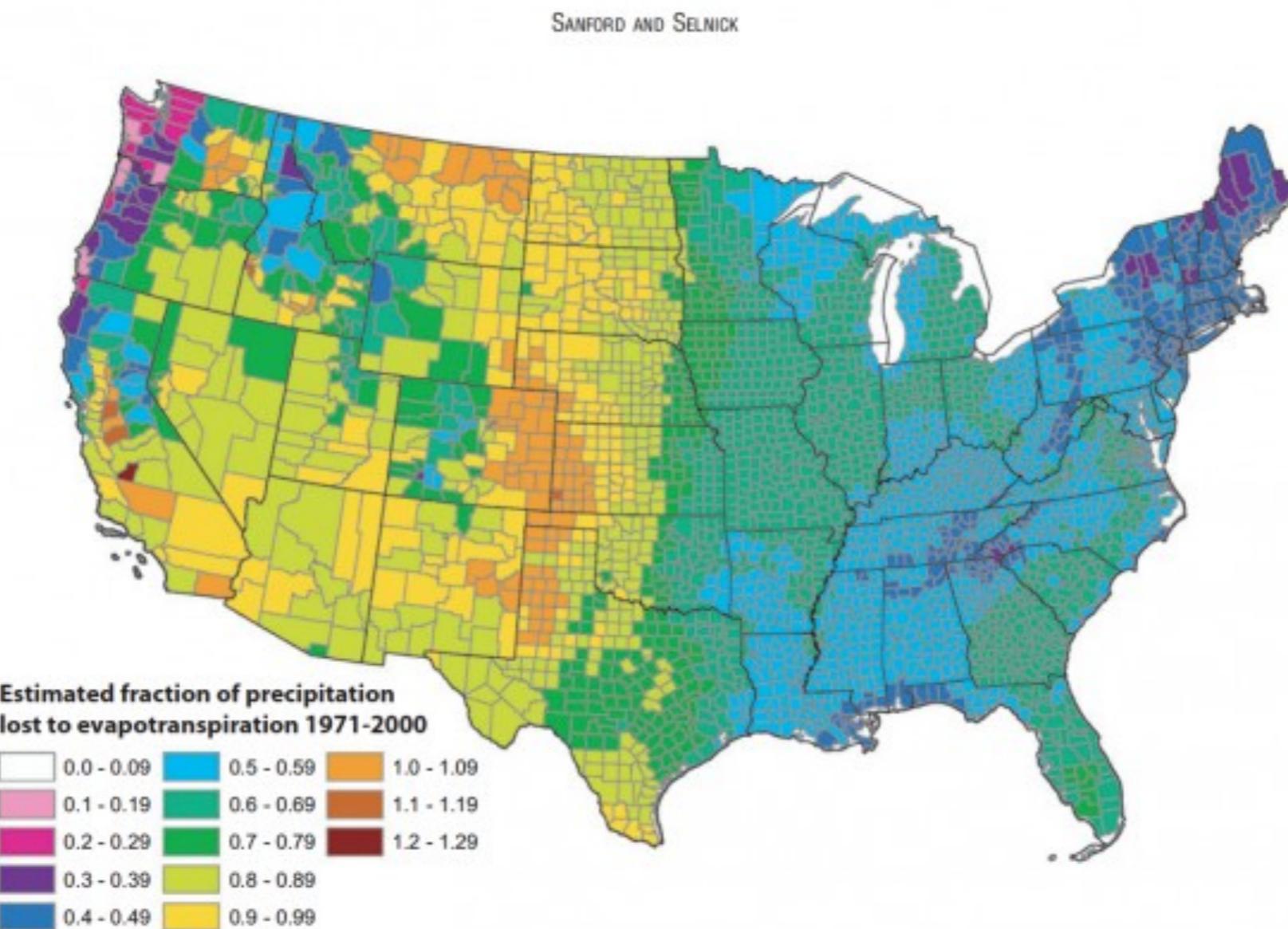
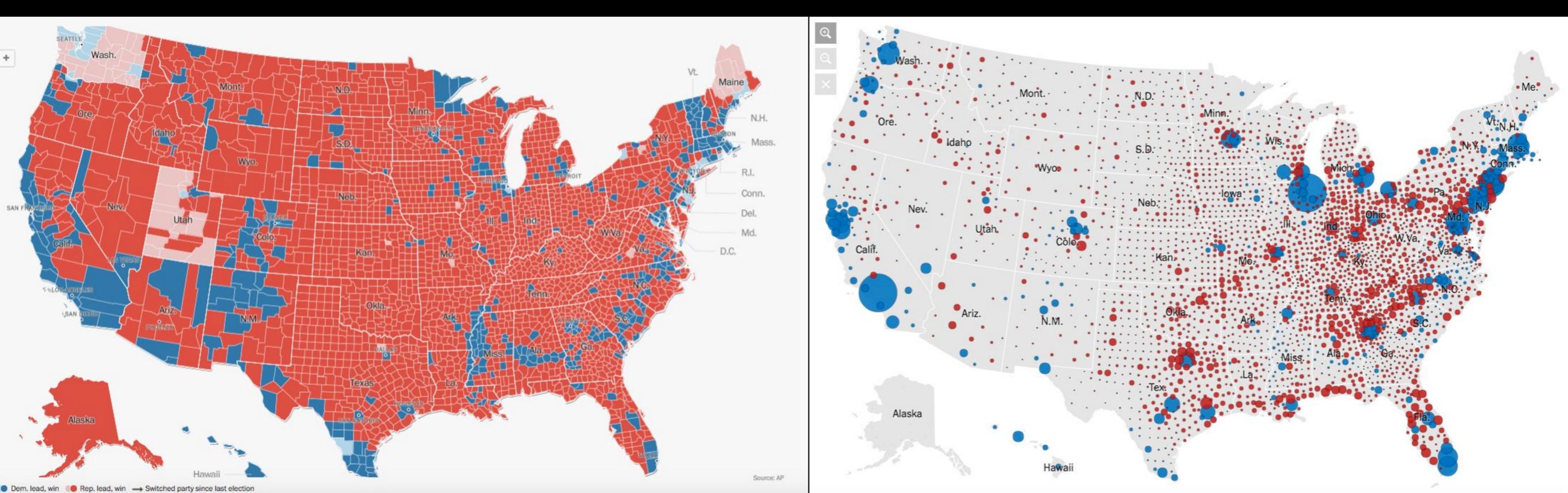


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation ( $P$ ) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of  $ET/P$  were made first at the 800-m resolution of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



## US Presidential Election 2016

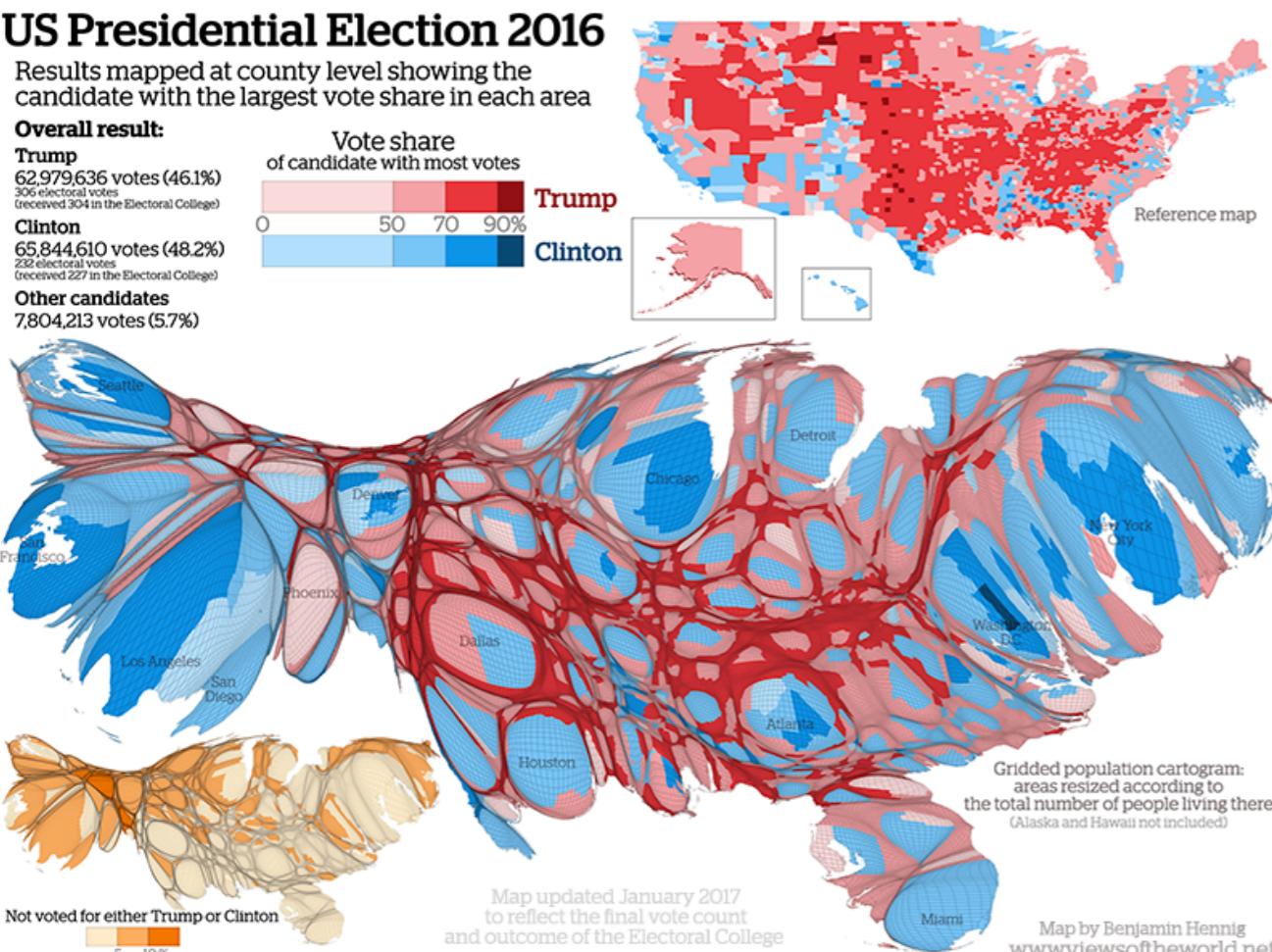
Results mapped at county level showing the candidate with the largest vote share in each area

### Overall result:

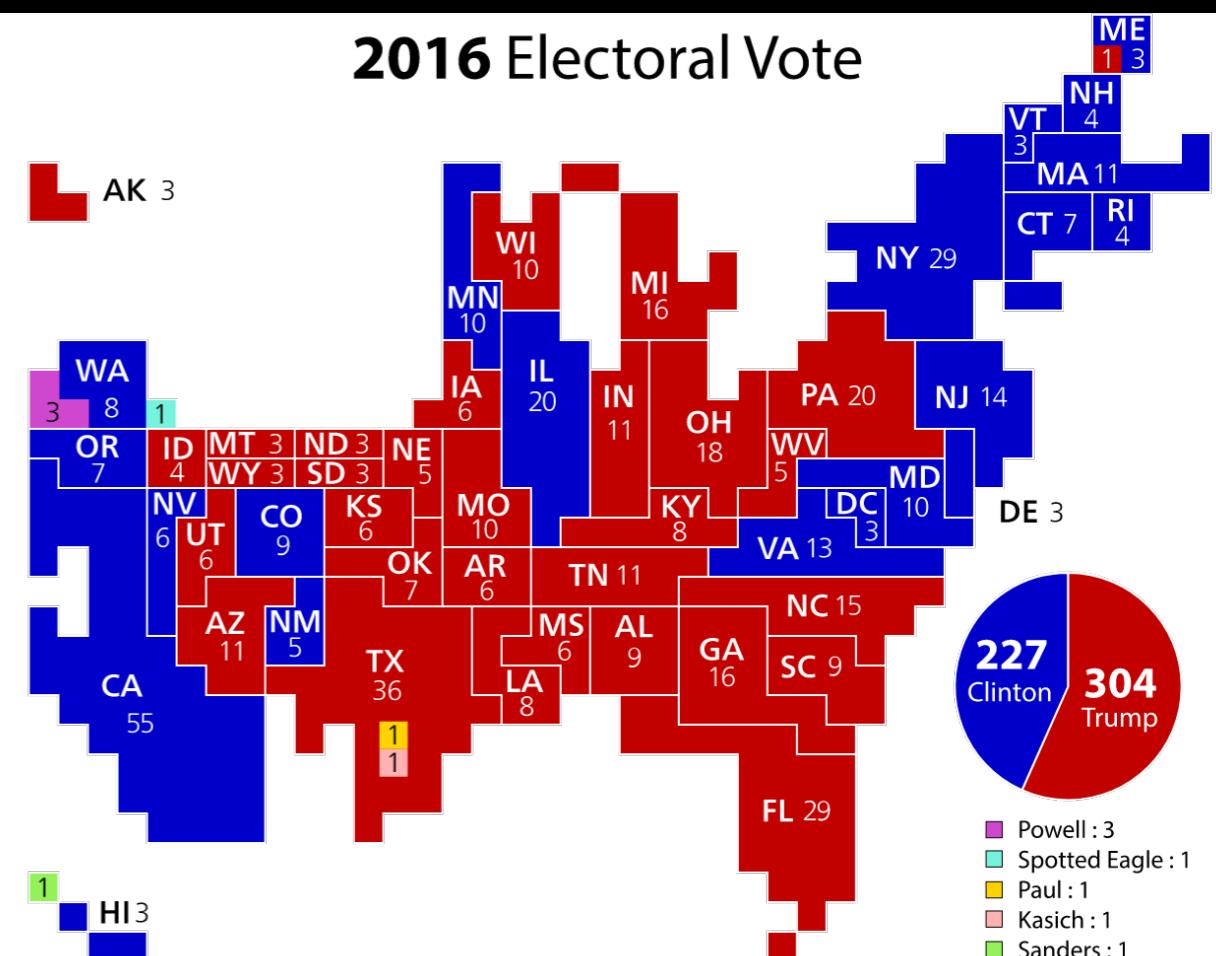
**Trump**  
62,979,636 votes (46.1%)  
(received 304 in the Electoral College)

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

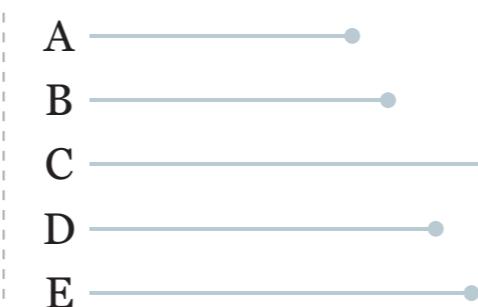
**Other candidates**  
7,804,213 votes (5.7%)



## 2016 Electoral Vote



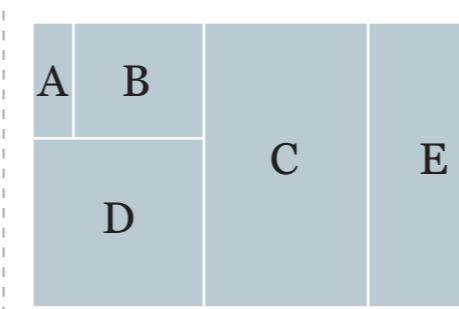
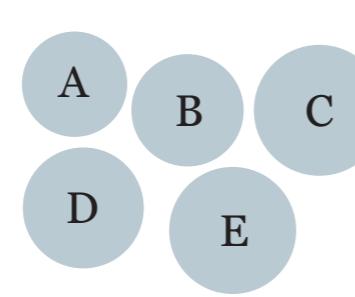
## *Length or height*



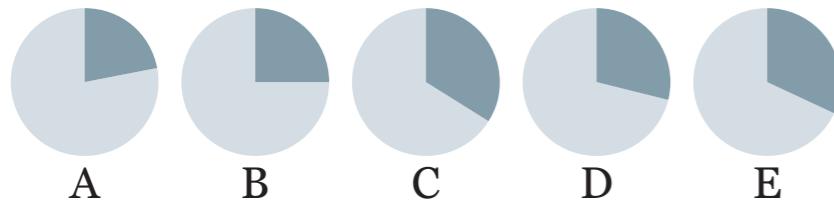
## *Position*



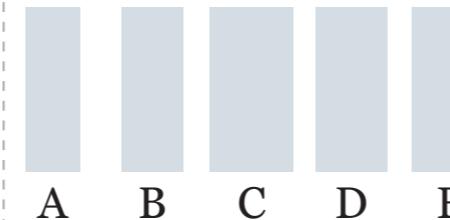
## *Area*



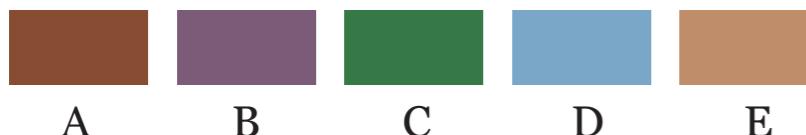
## *Angle/area*



## *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

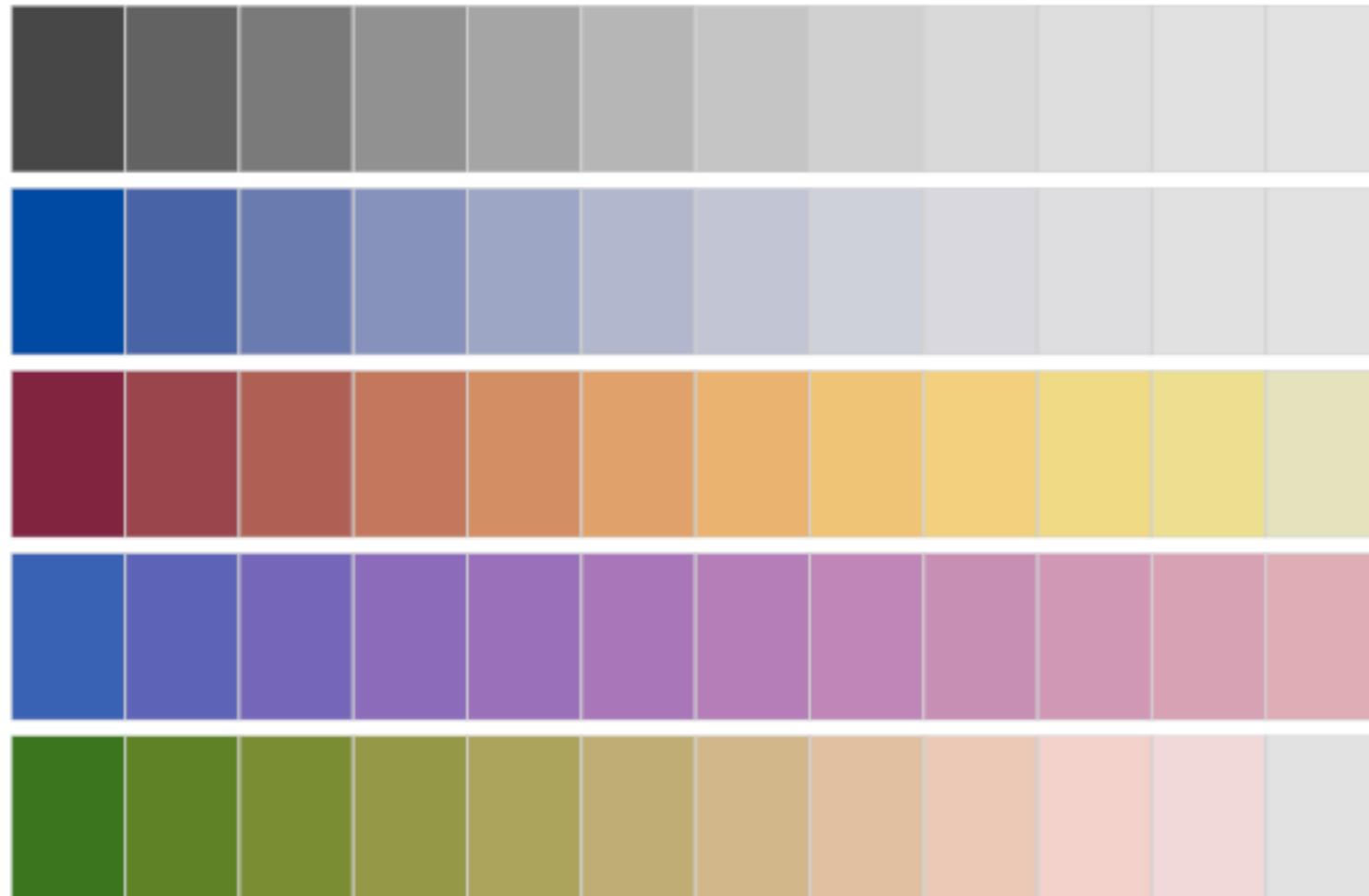
# Colors for Categories

Do not use more than 5-8 colors at once



# Colors for Ordinal Data

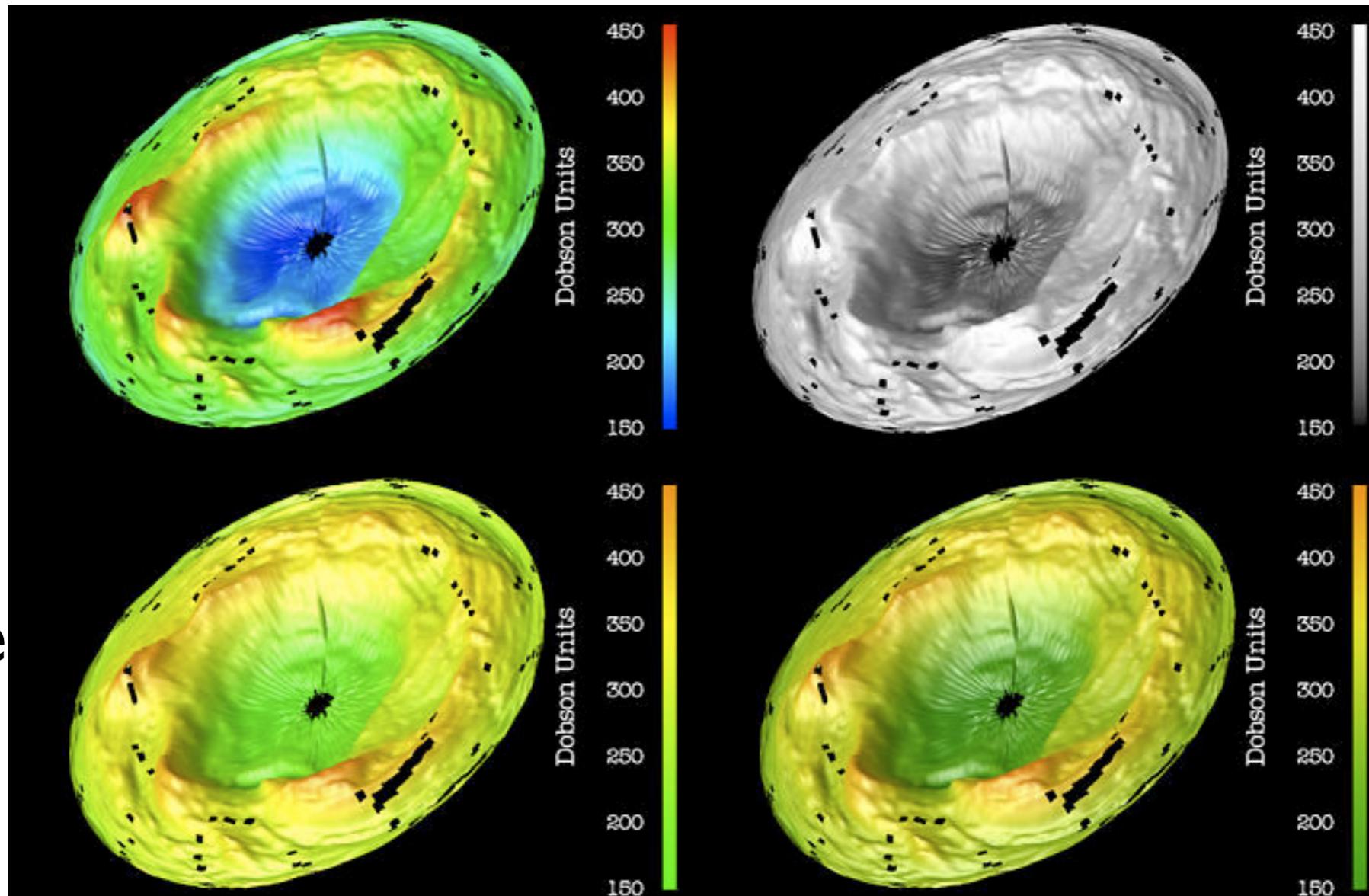
Vary luminance and saturation



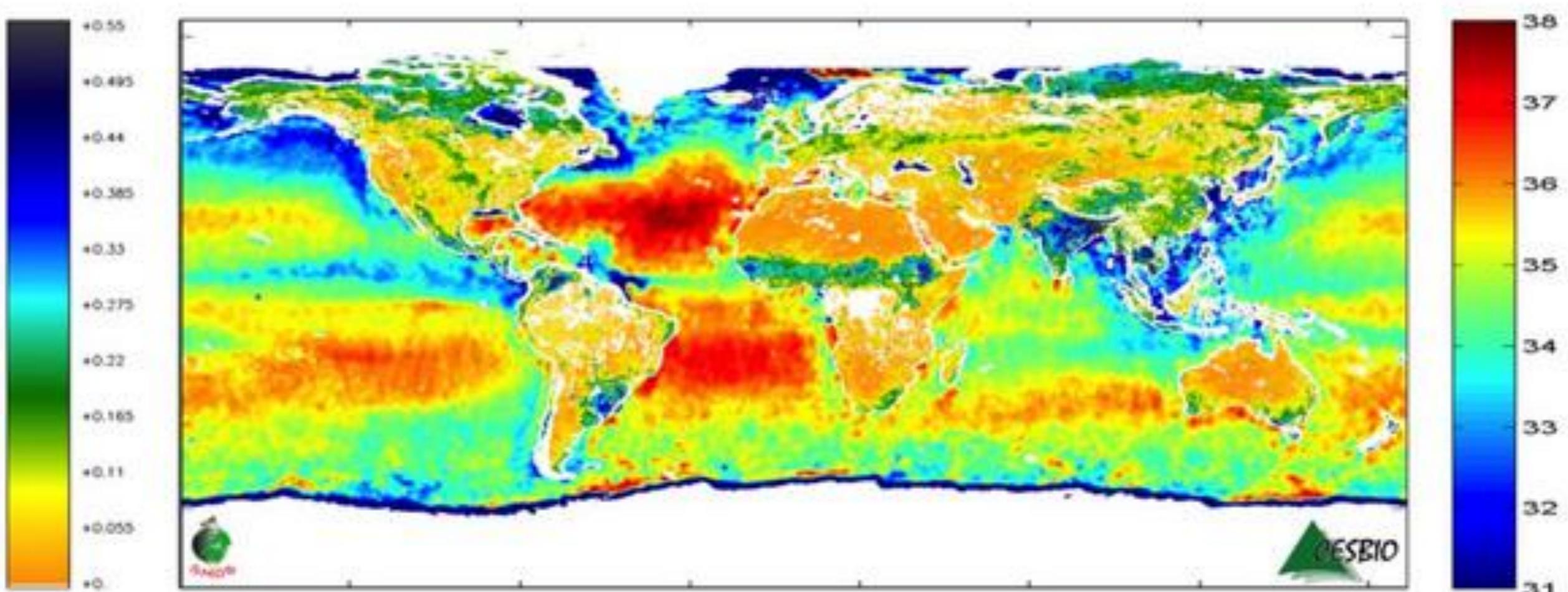
Zeilis et al, 2009, "Escaping RGBland: Selecting  
Colors for Statistical Graphics"

# Colors for Quantitative Data

Hue  
(Rainbow)

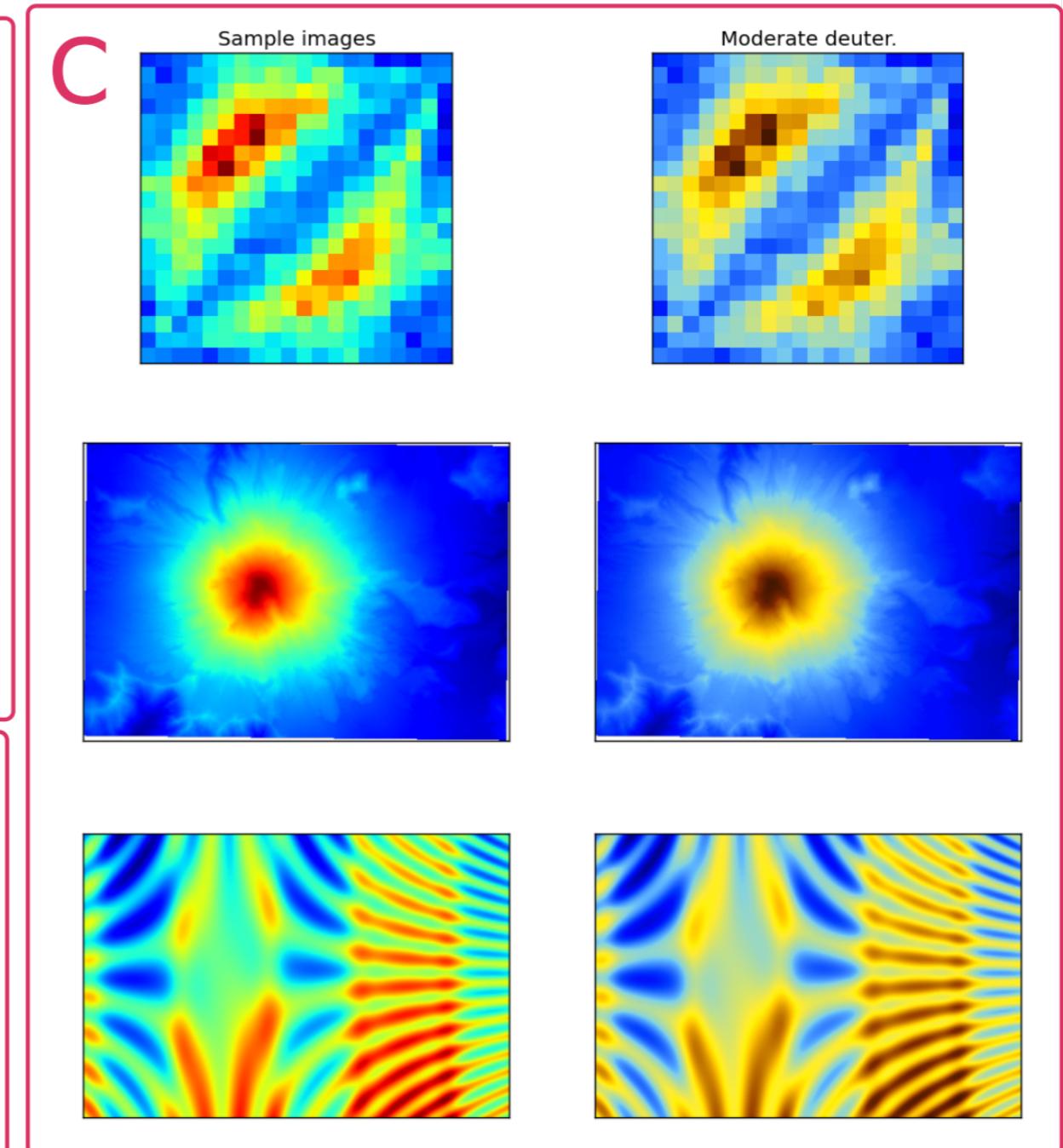
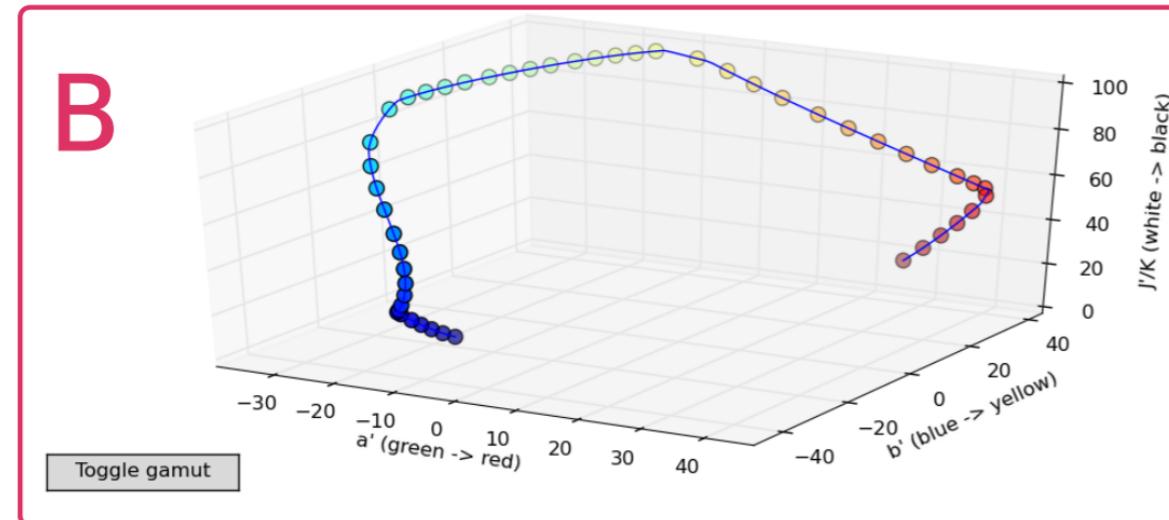
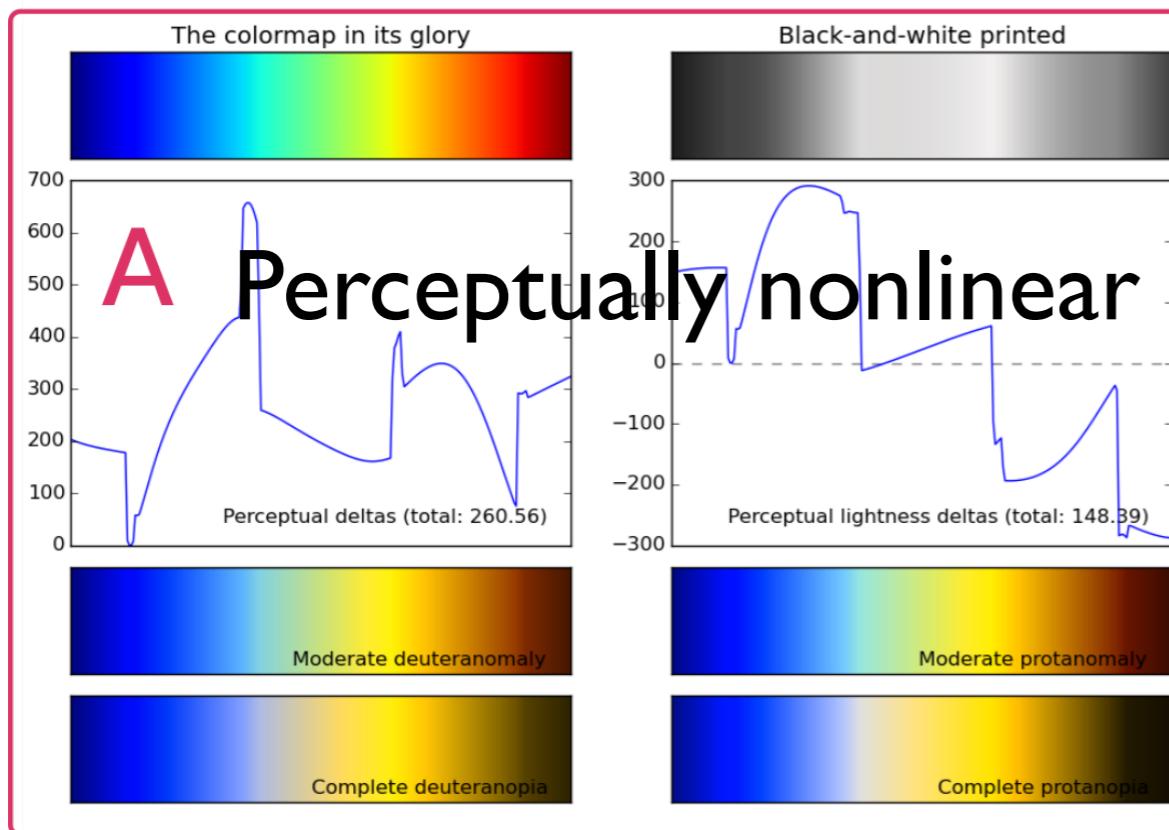


# Rainbow Colormap



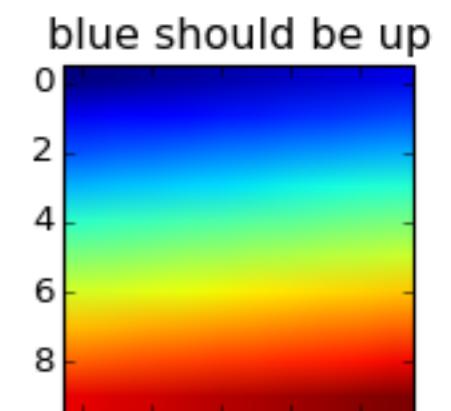
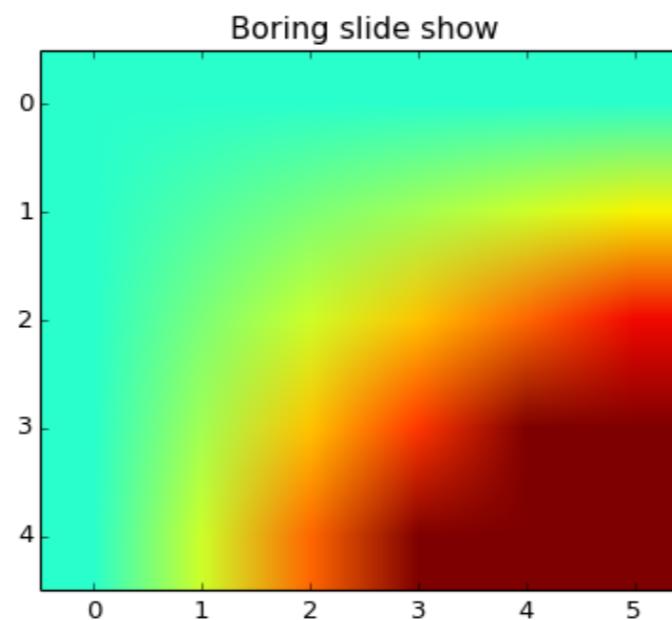
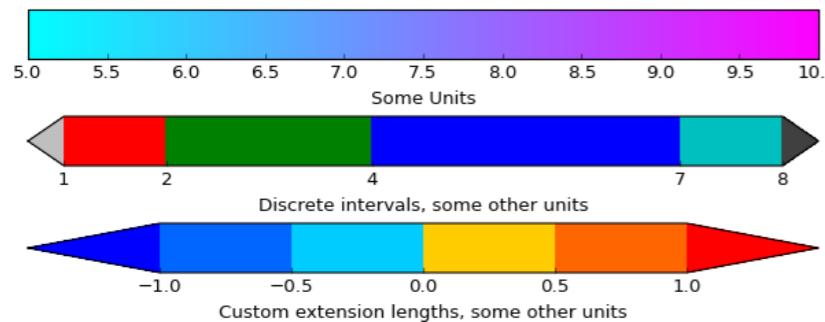
# Rainbow Colormap

Colormap evaluation: jet

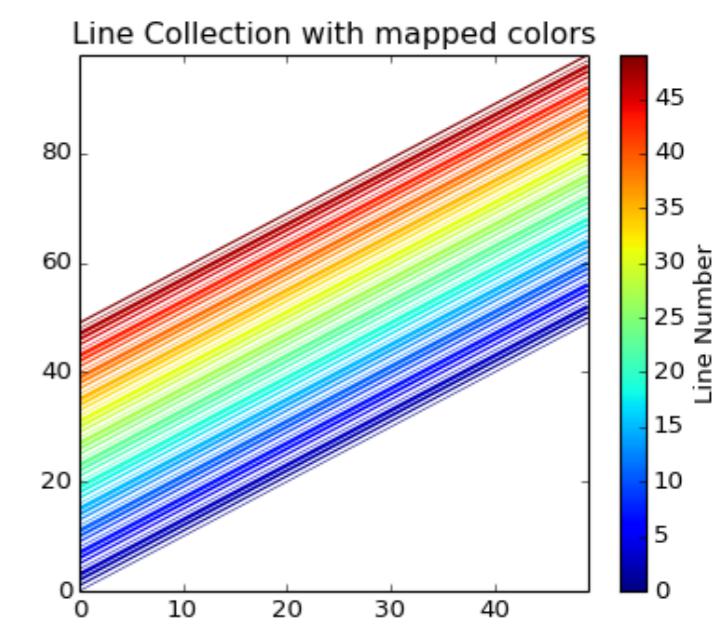
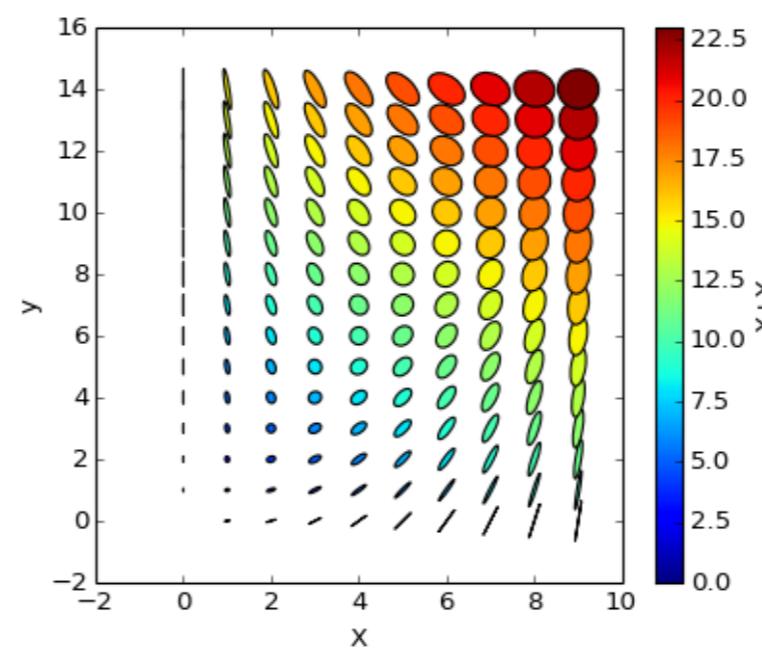
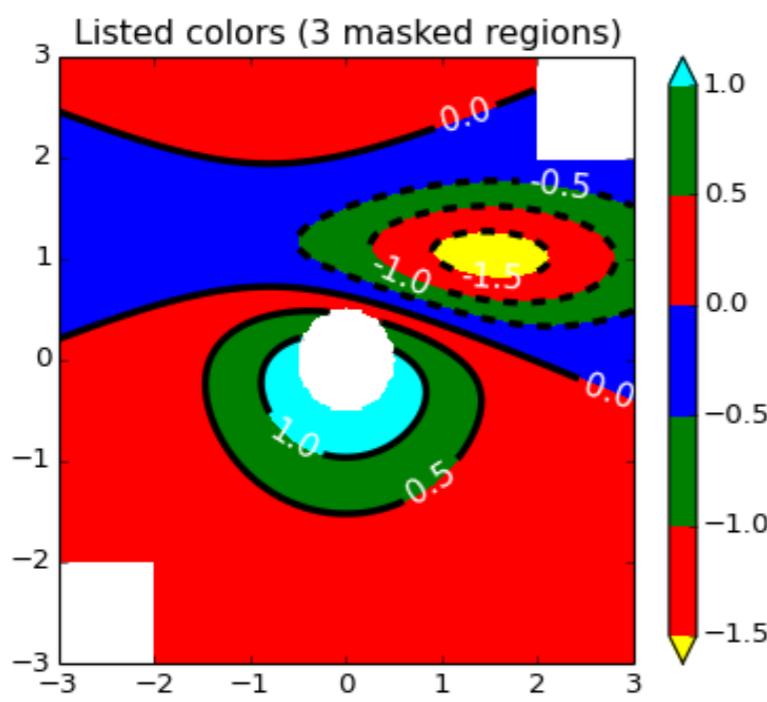


R. Simmon

# Avoid Rainbow Colors!

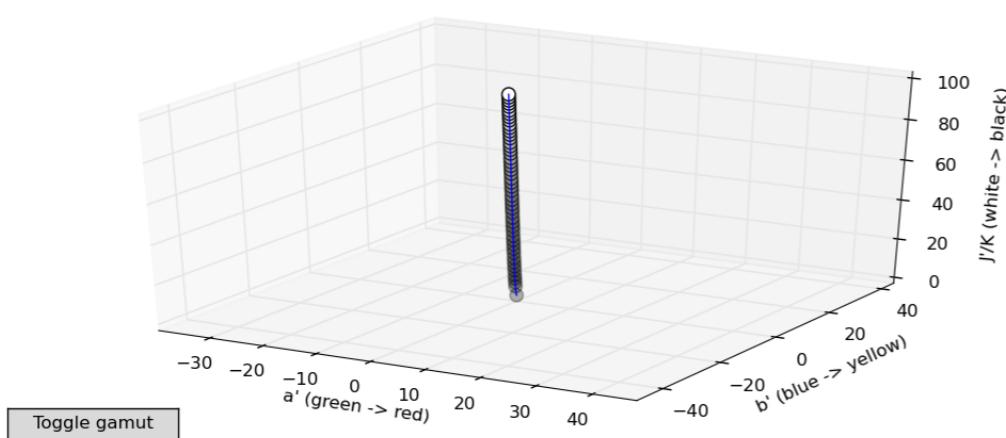
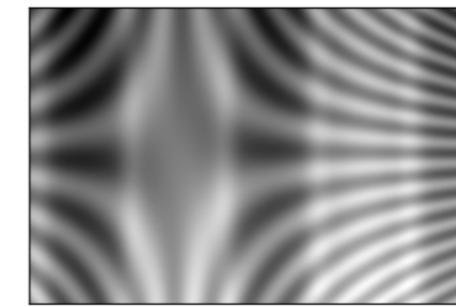
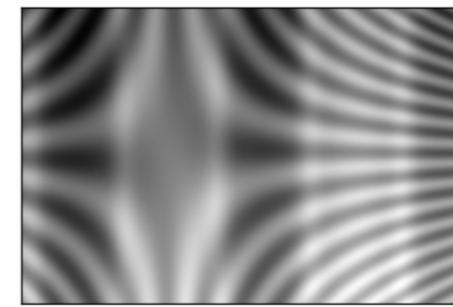
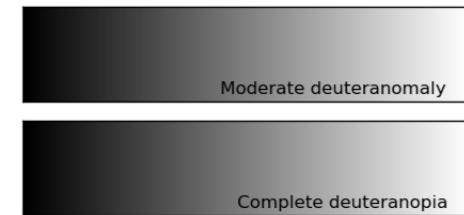
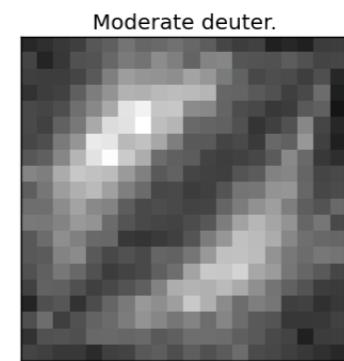
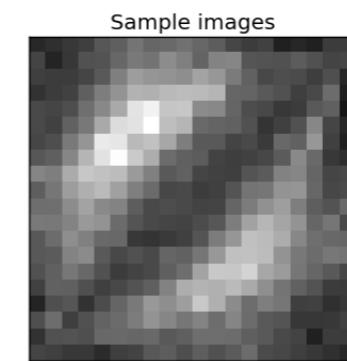
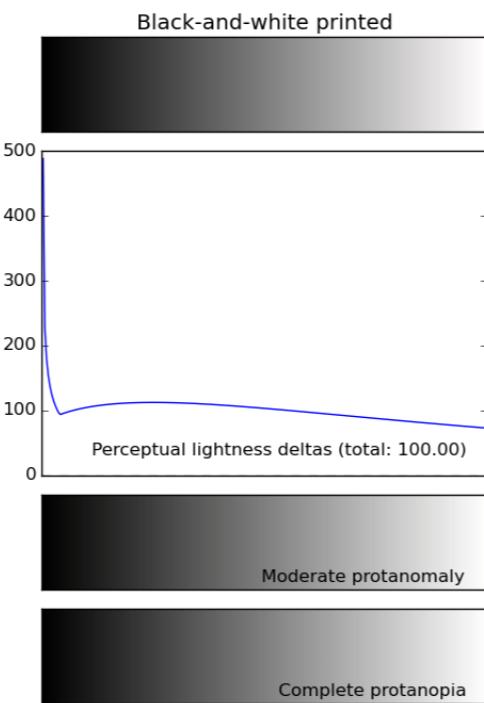
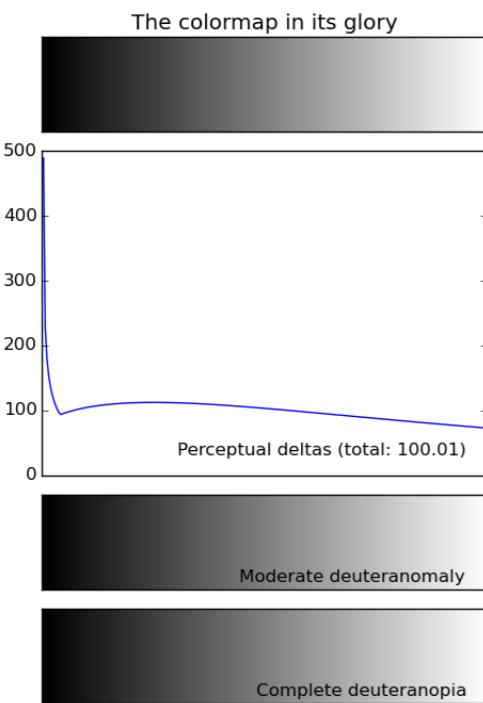


blue should be down



# Gray

Colormap evaluation: gray



# Color Blindness



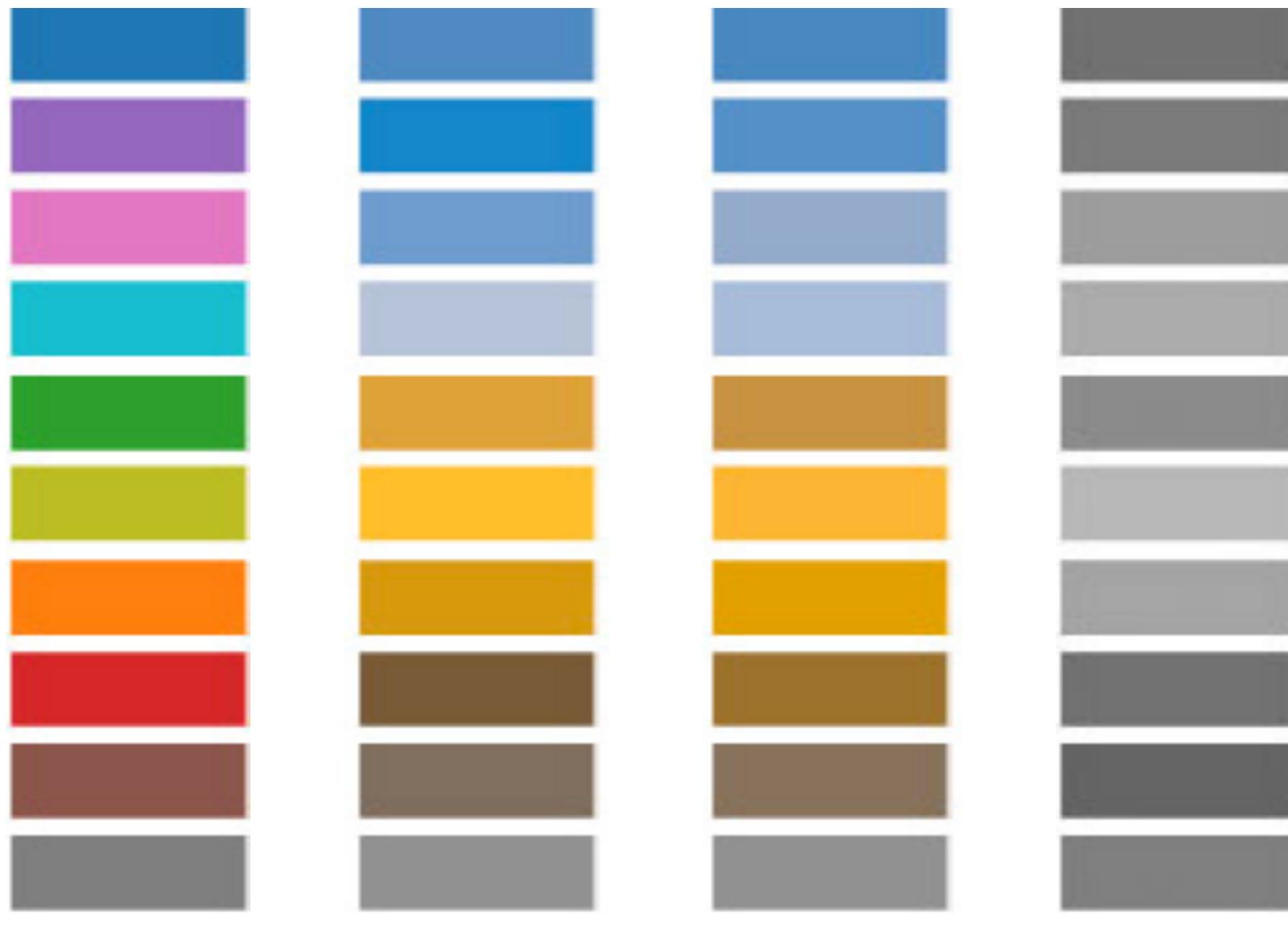
Protanope

Red / green  
deficiencies

Deuteranope

Blue / Yellow  
deficiency

# Color Blindness



Normal

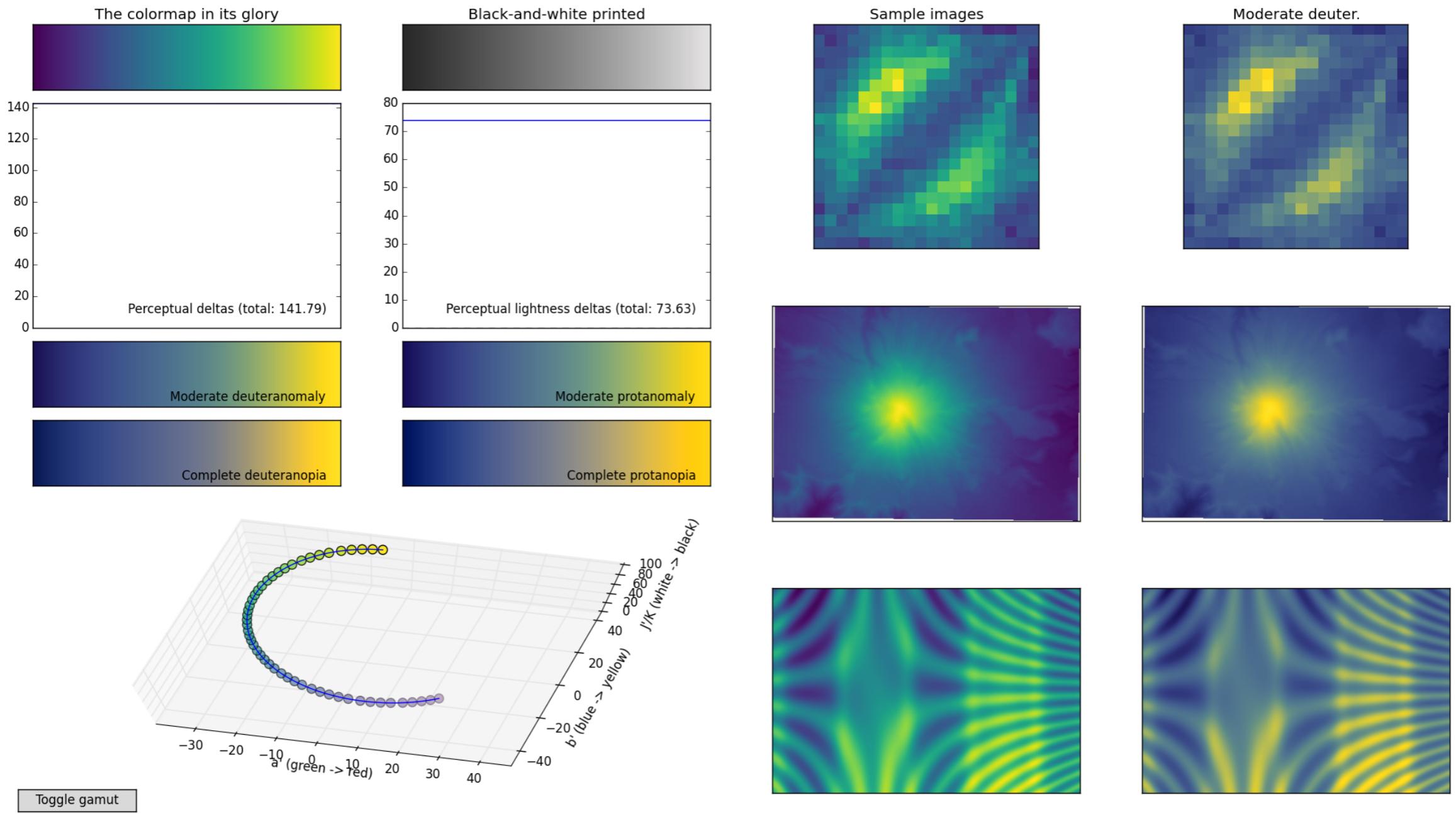
Protanope

Deuteranope

Lightness

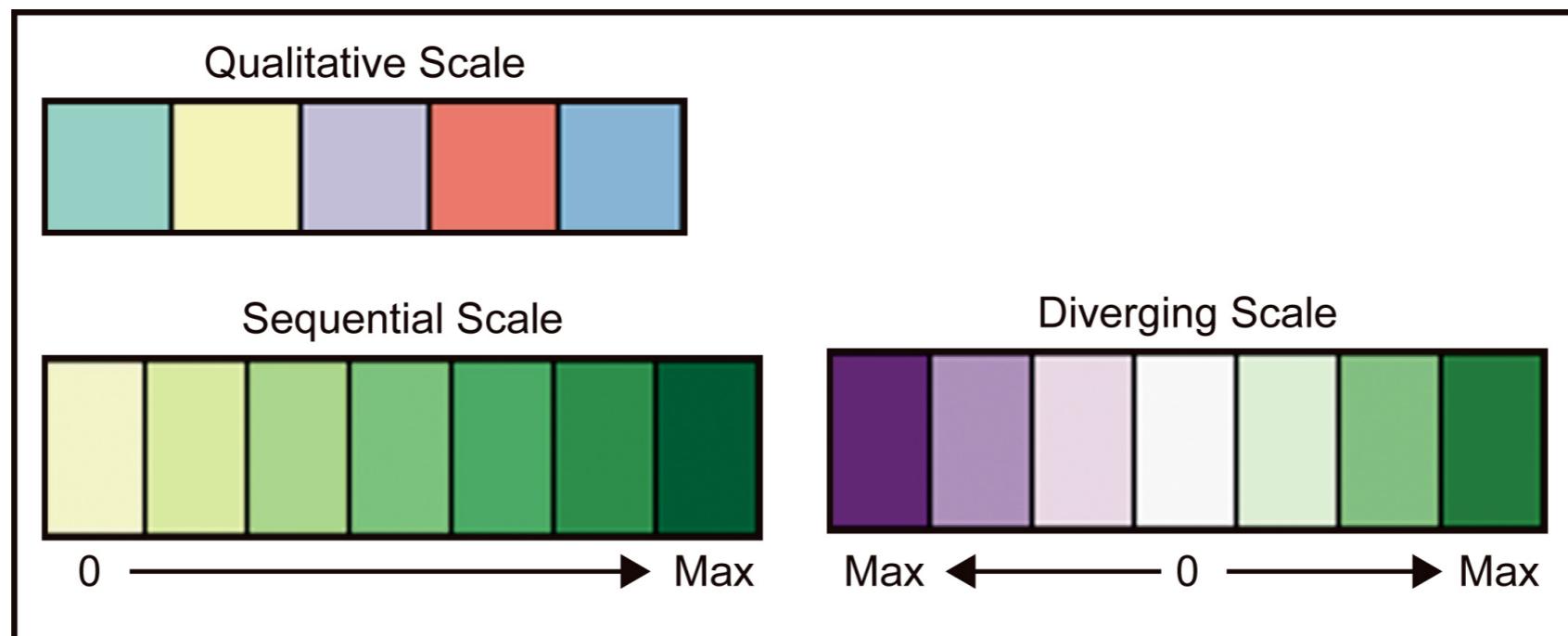
# Viridis

Colormap evaluation: option\_d.py



# Color Brewer

Nominal



Ordinal

number of data classes on your map

3

[learn more >](#)

how to use | updates | credits

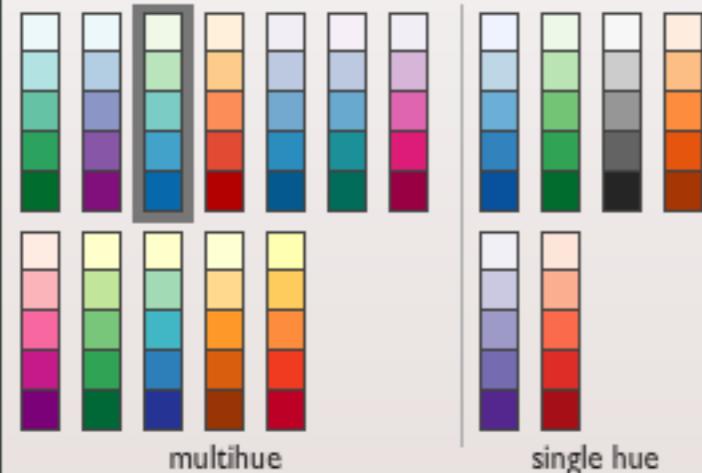
**COLORBREWER 2.0**  
color advice for cartography

the nature of your data

sequential

[learn more >](#)

pick a color scheme: GnBu



(optional) only show schemes that are:

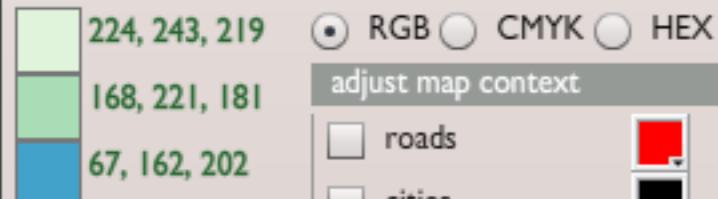
colorblind safe

print friendly

photocopy-able

[learn more >](#)

pick a color system



RGB  CMYK  HEX

adjust map context

roads

cities

borders

select a background

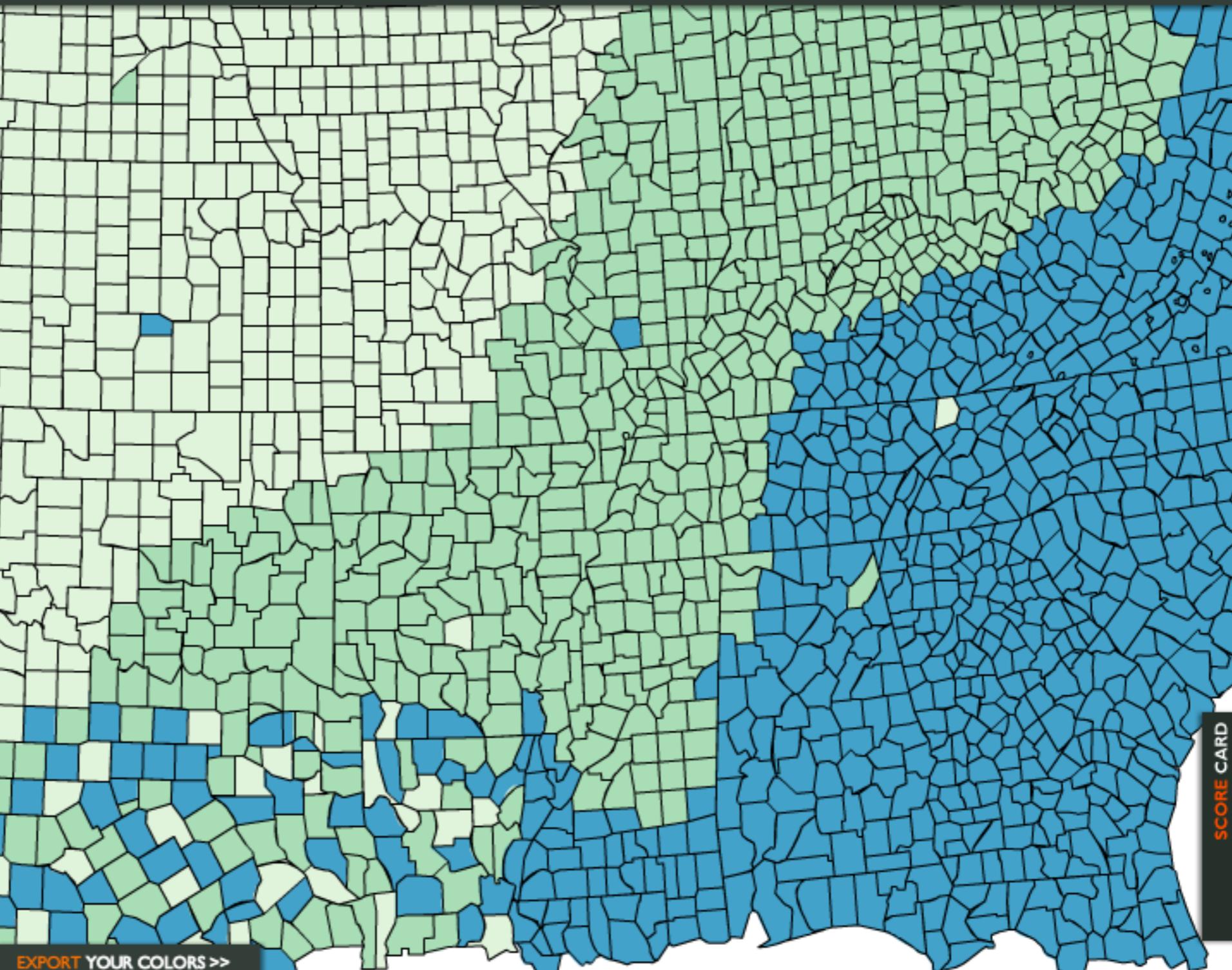
solid color

terrain

color transparency

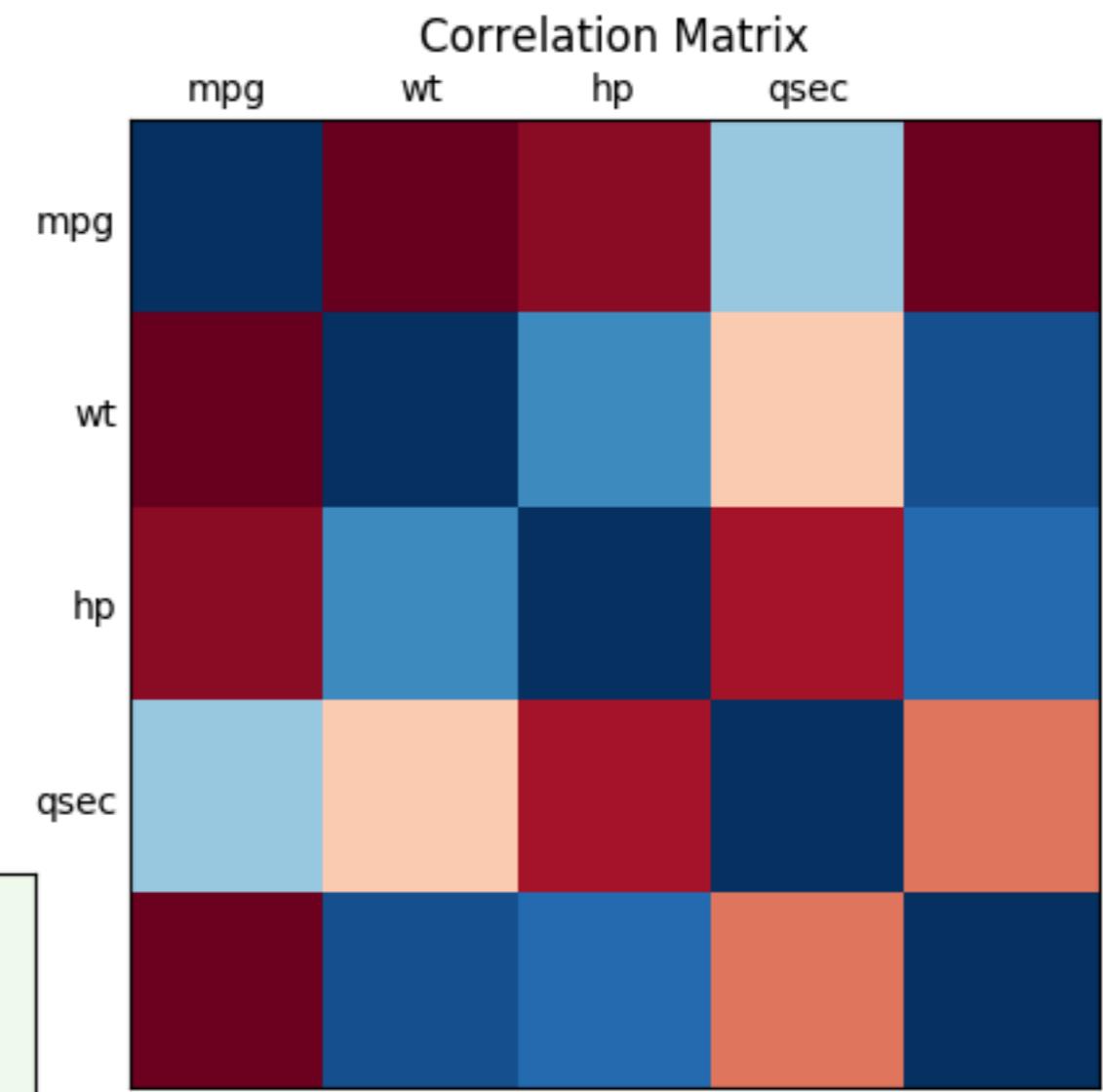
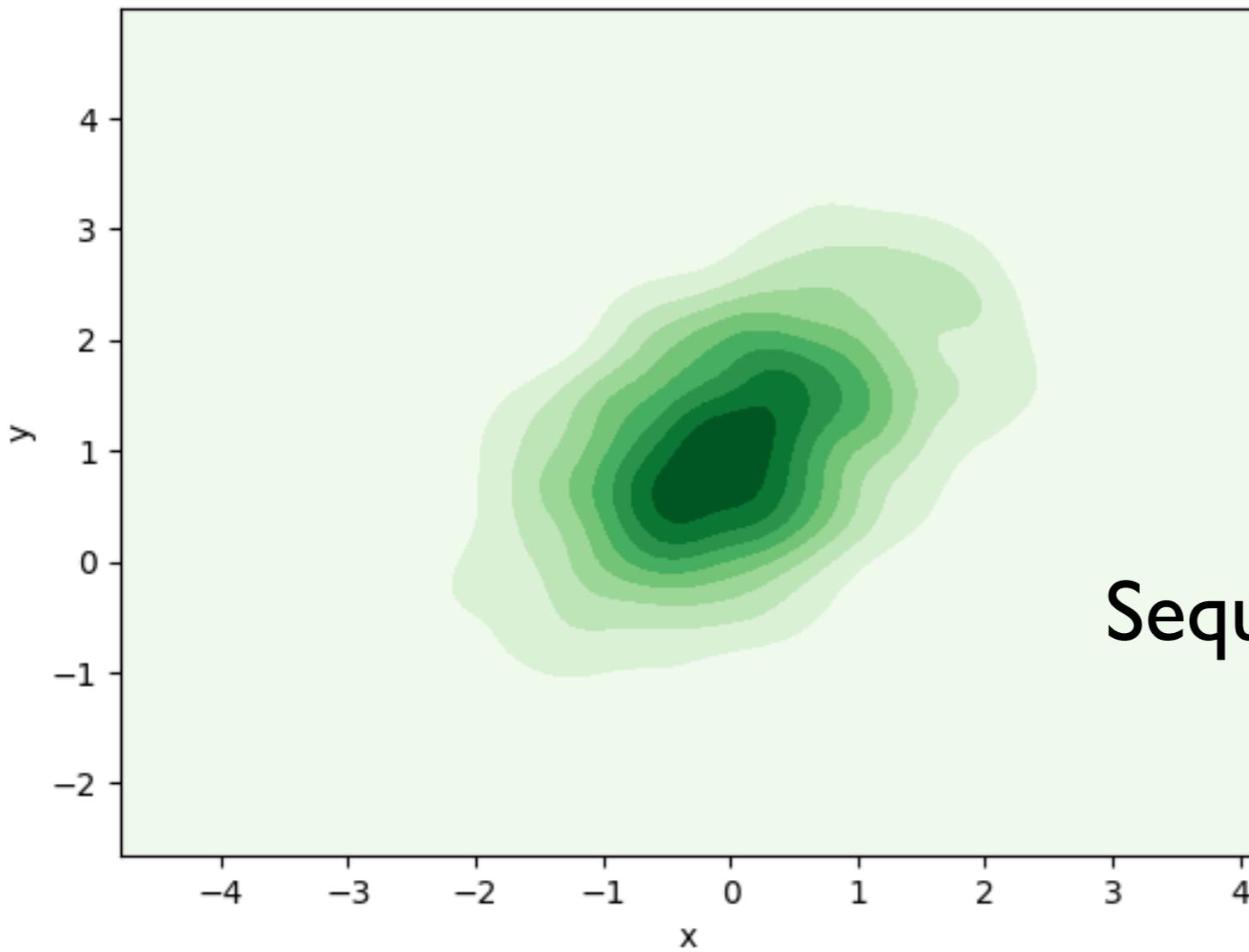
[learn more >](#)

[EXPORT YOUR COLORS >>](#)



SCORE CARD

# Diverging Palette for Quantitative or Ordinal



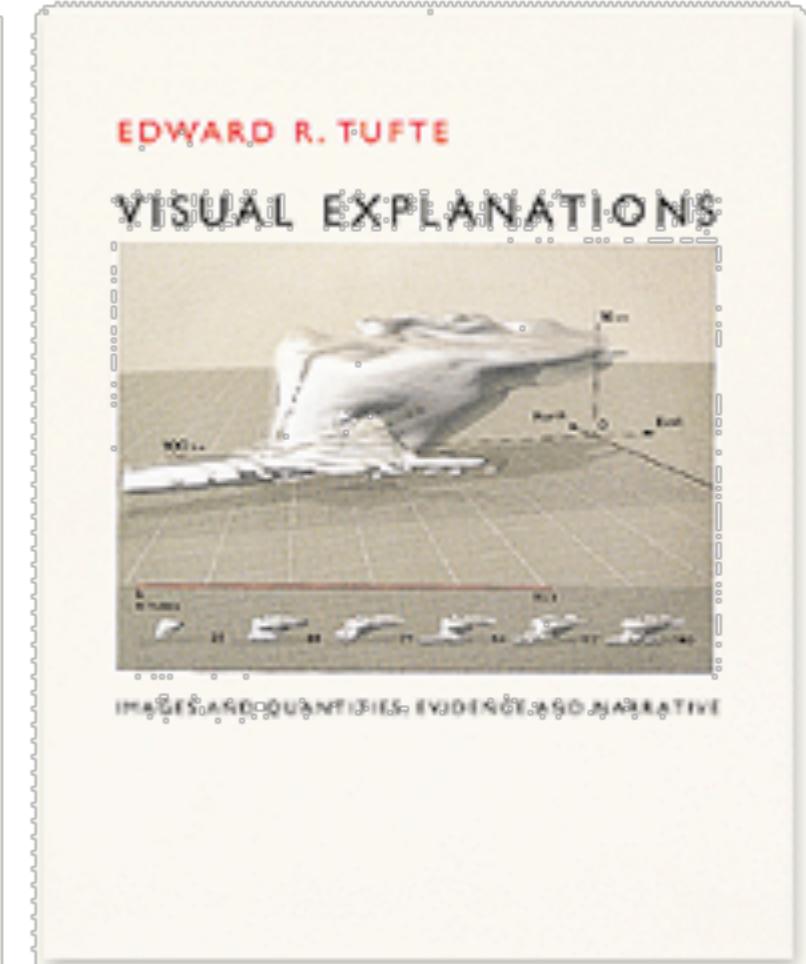
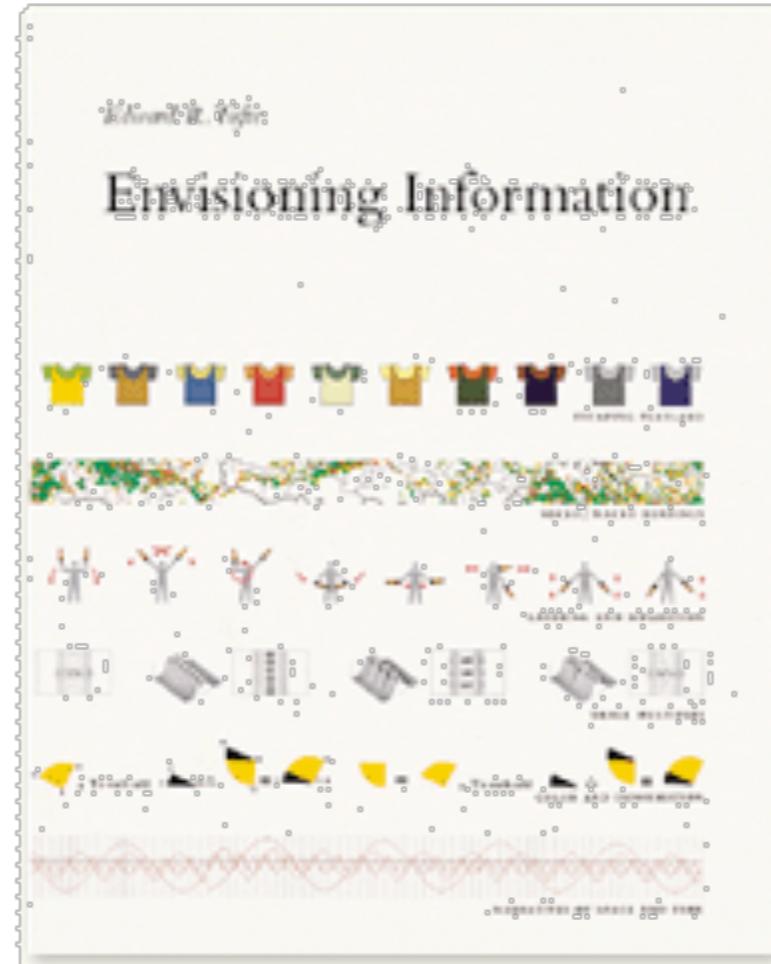
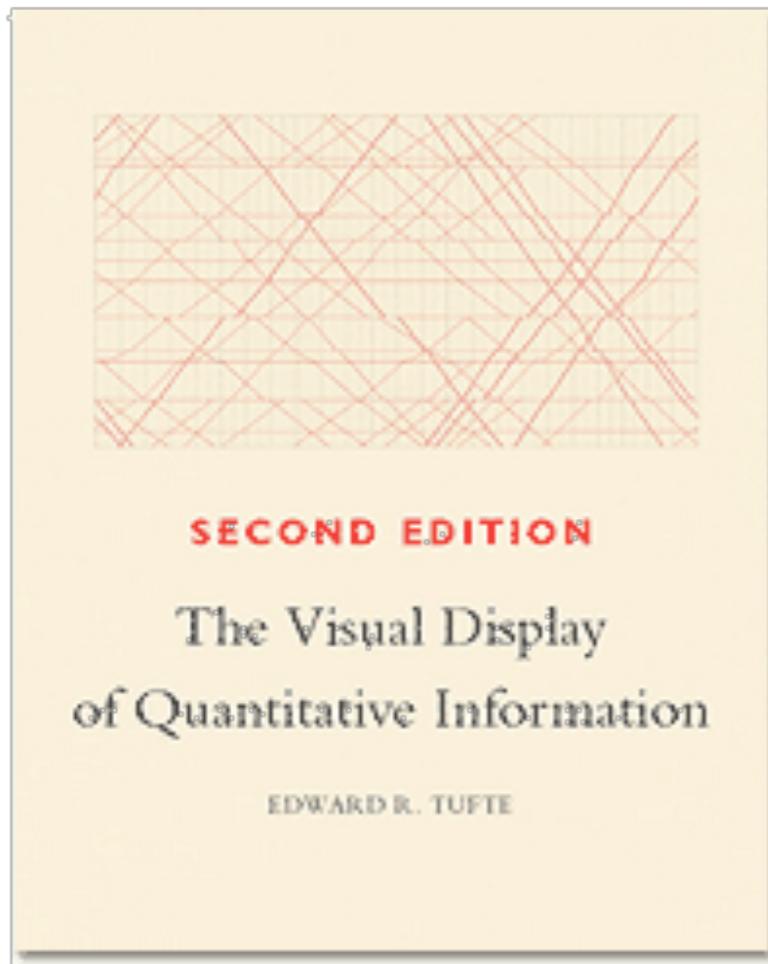
Sequential Palette for Densities

# Effective Visualizations

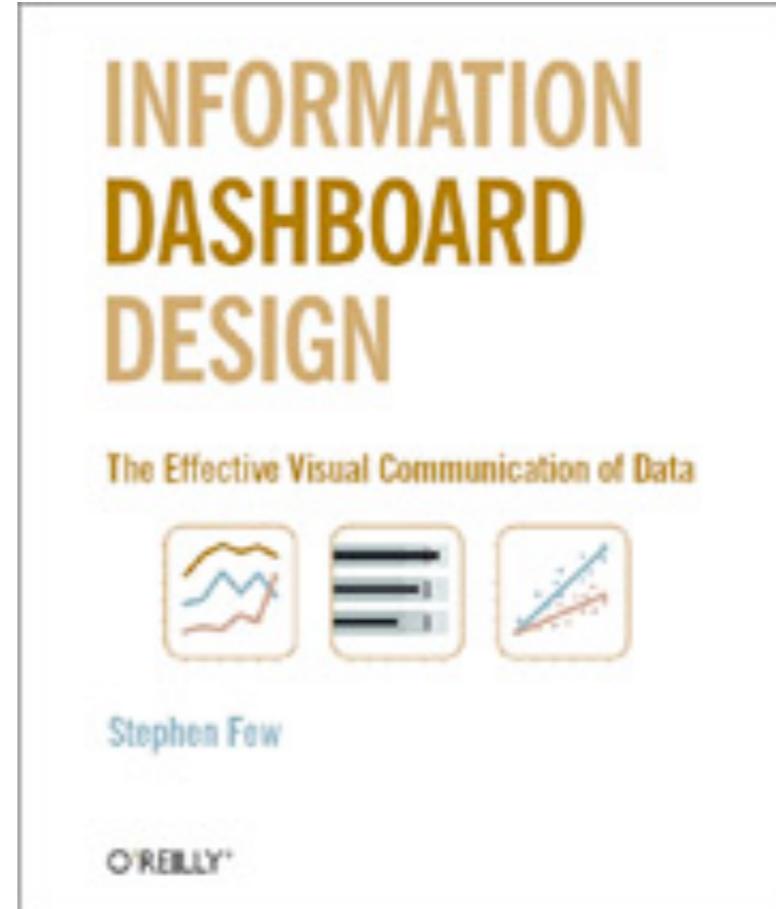
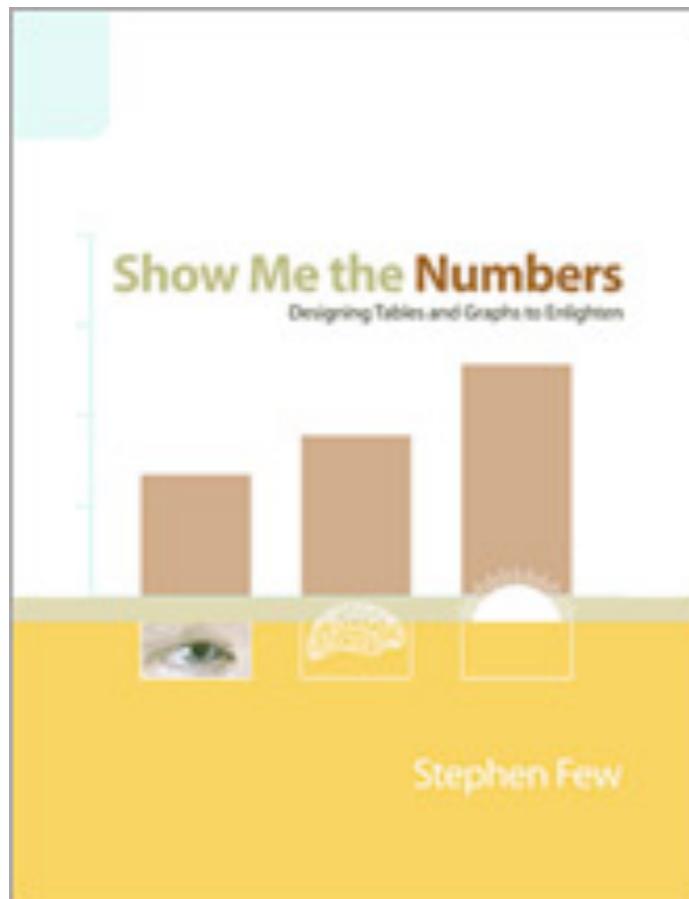
1. Have graphical integrity
2. Keep it simple
3. Use the right display
4. Use color strategically

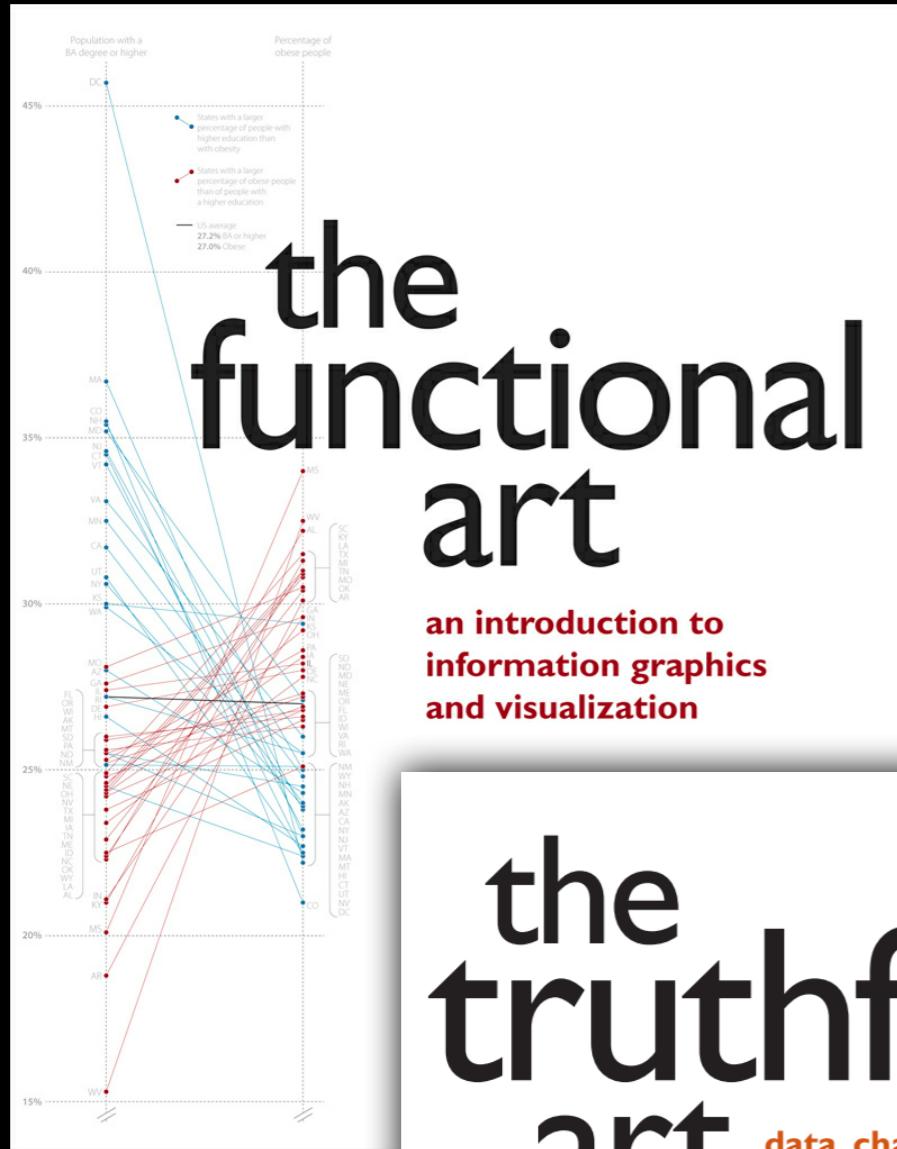
# Further Reading

# Edward Tufte



# Stephen Few

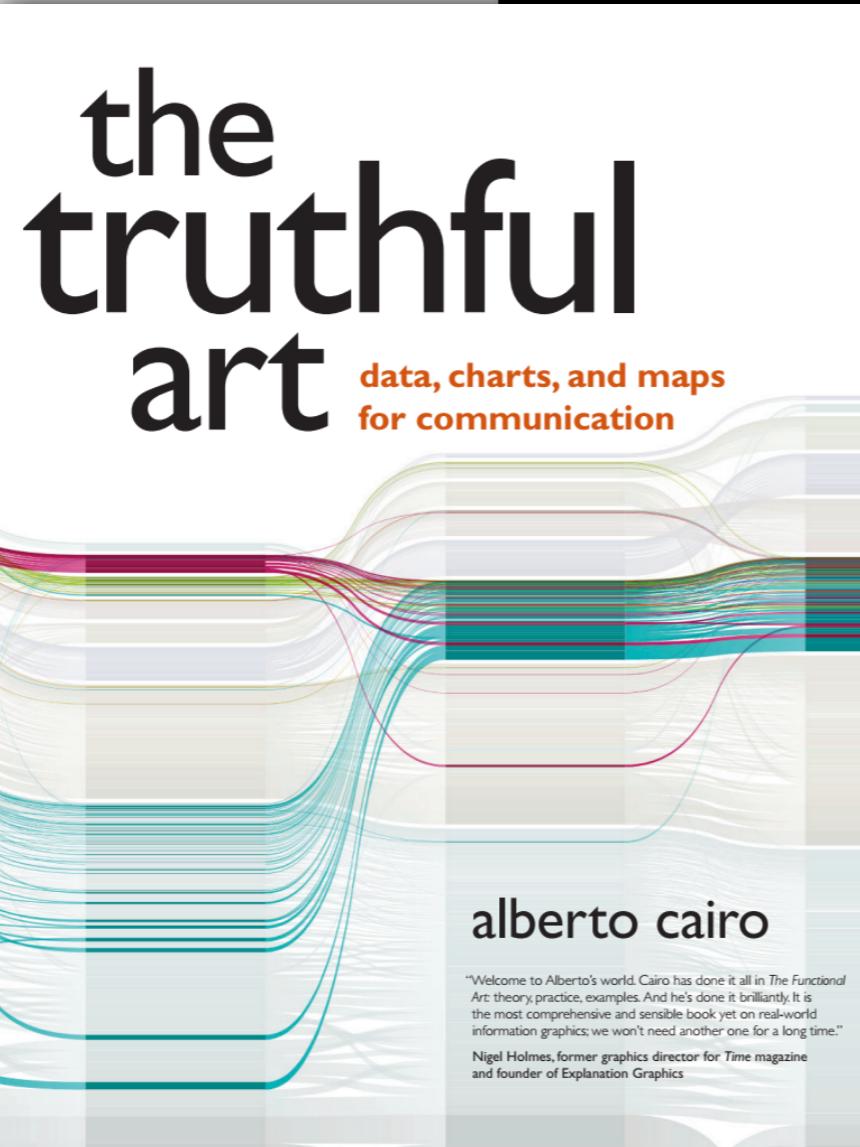




# the functional art

**an introduction to  
information graphics  
and visualization**

2012



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