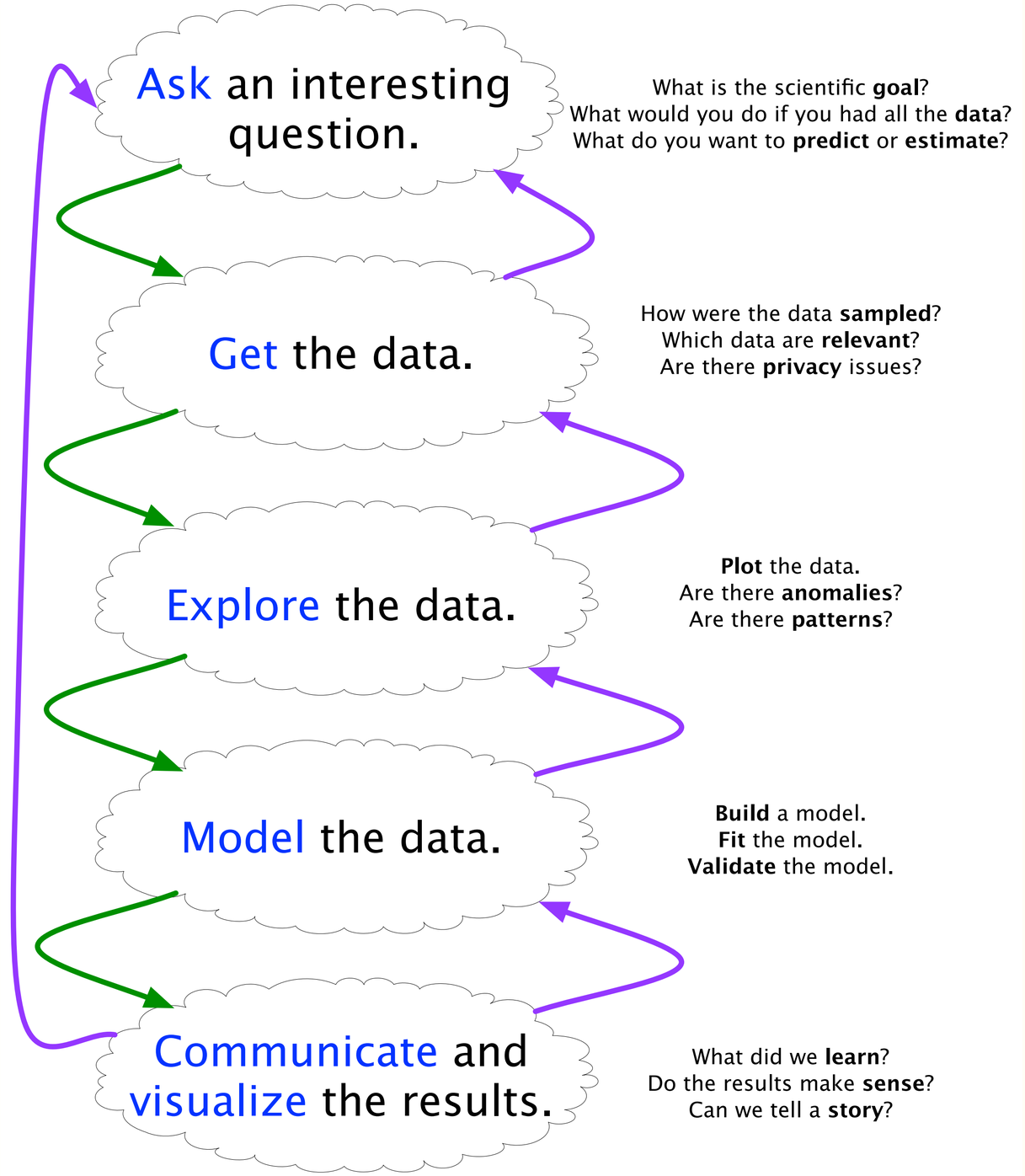
A bit about Visualization

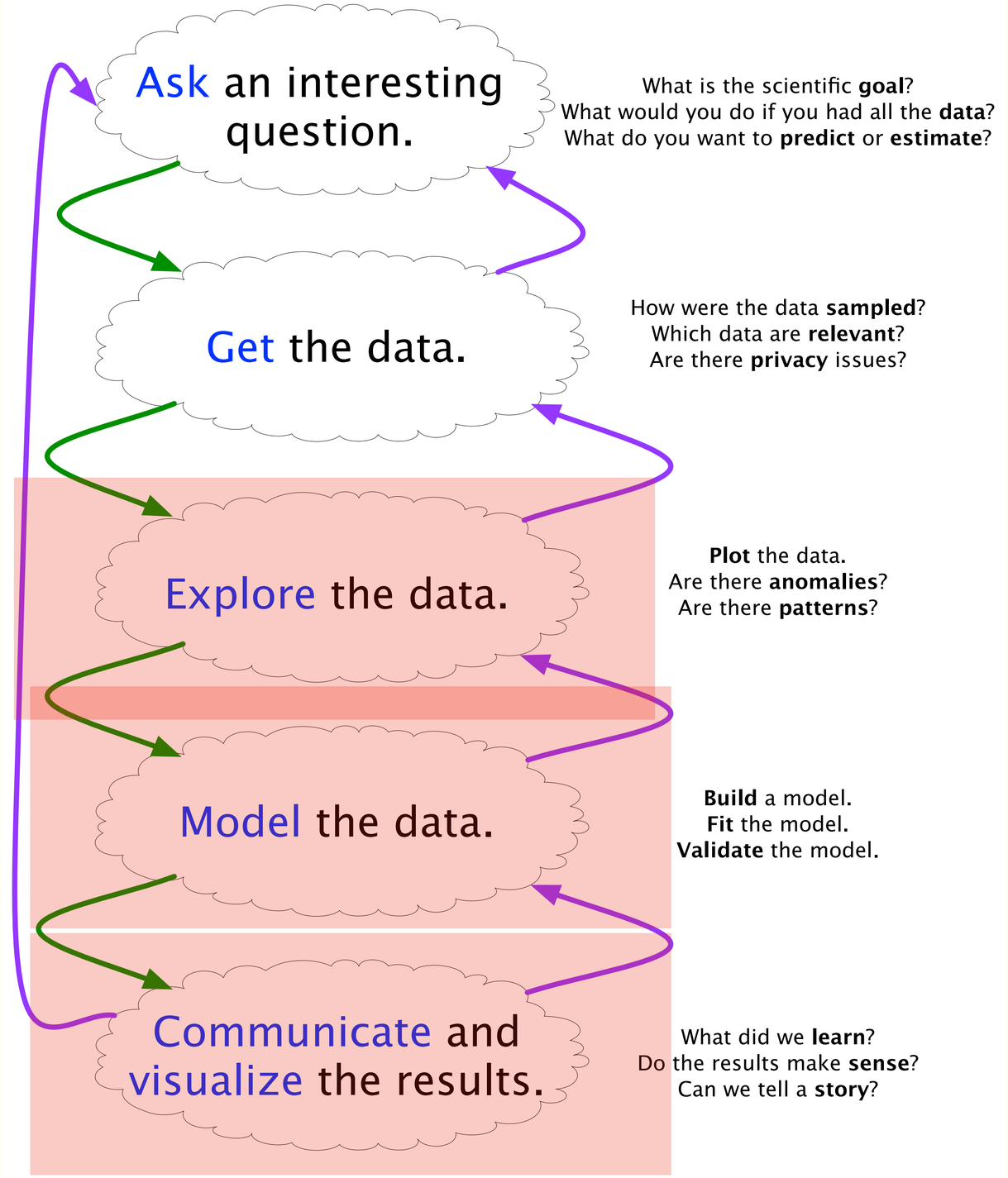
Rahul Dave

Based on cs109a: Data Science with lots of slides from Hanspeter Pfister (Harvard) and Alberto Cairo (University of Miami).

Why Visualize?

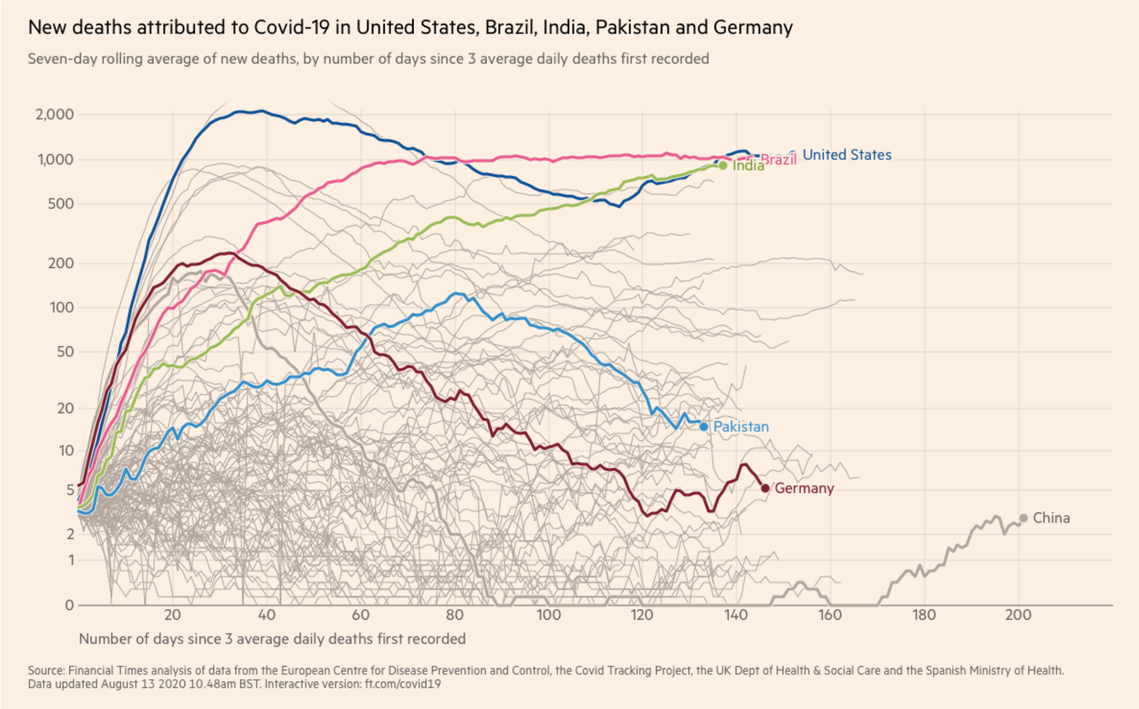
From cs109, Harvard

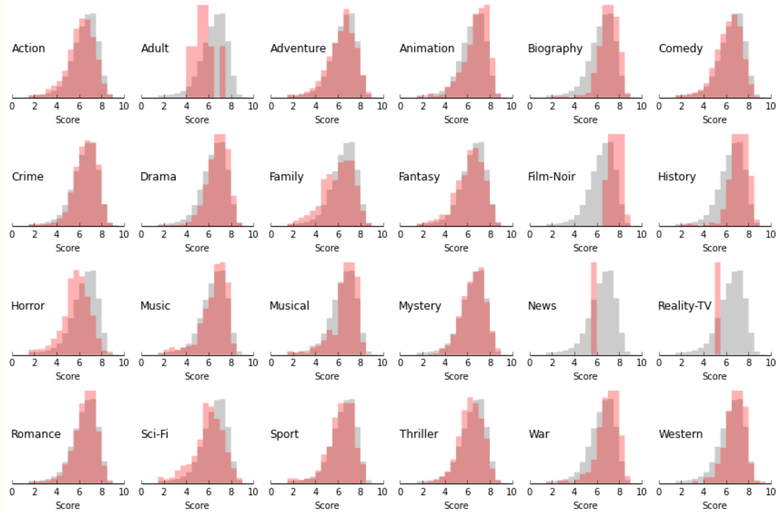




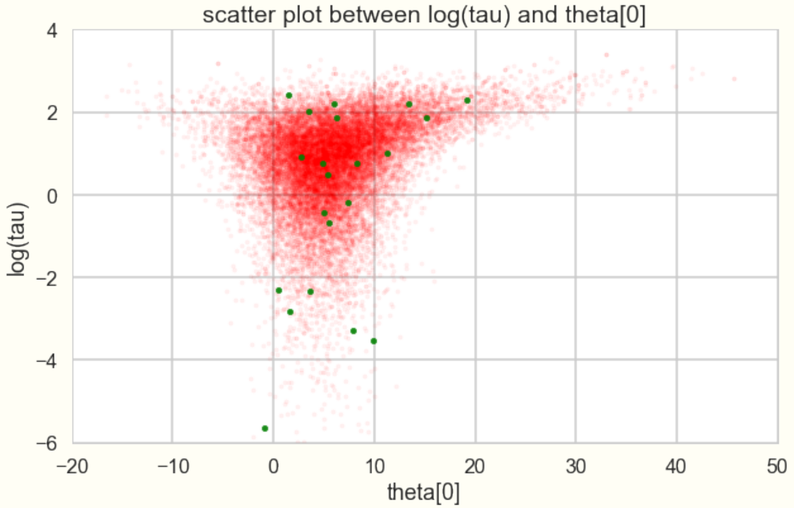
From cs109, Harvard

Communicate Results





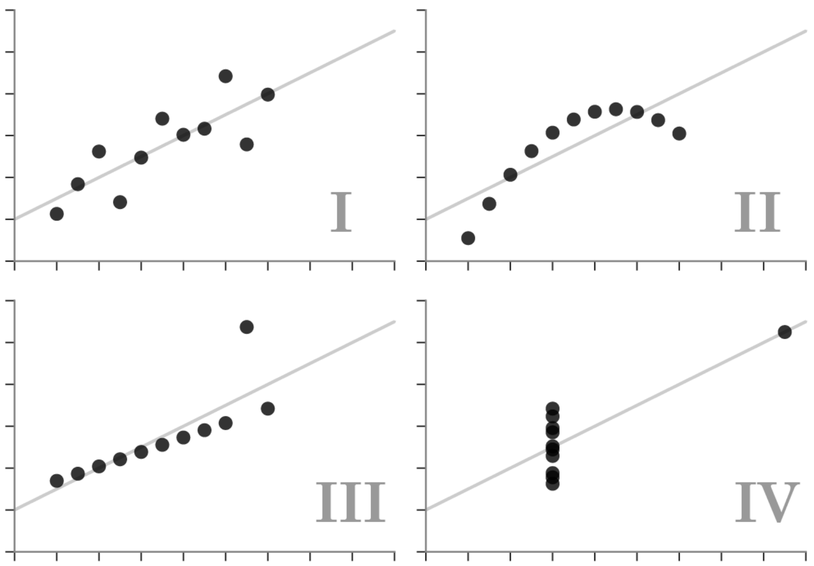
Explore data.
Explore models.



Make sure statistics are not fooling you

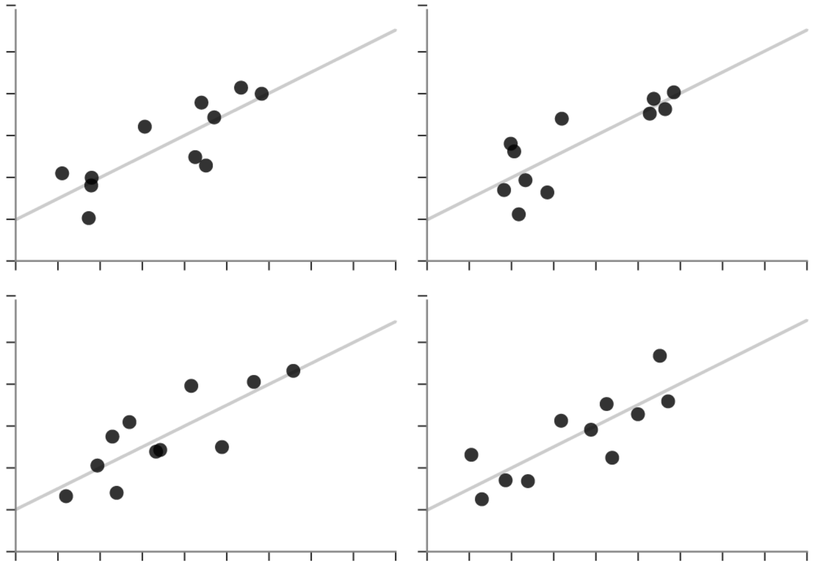
Anscombe's Quartet

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



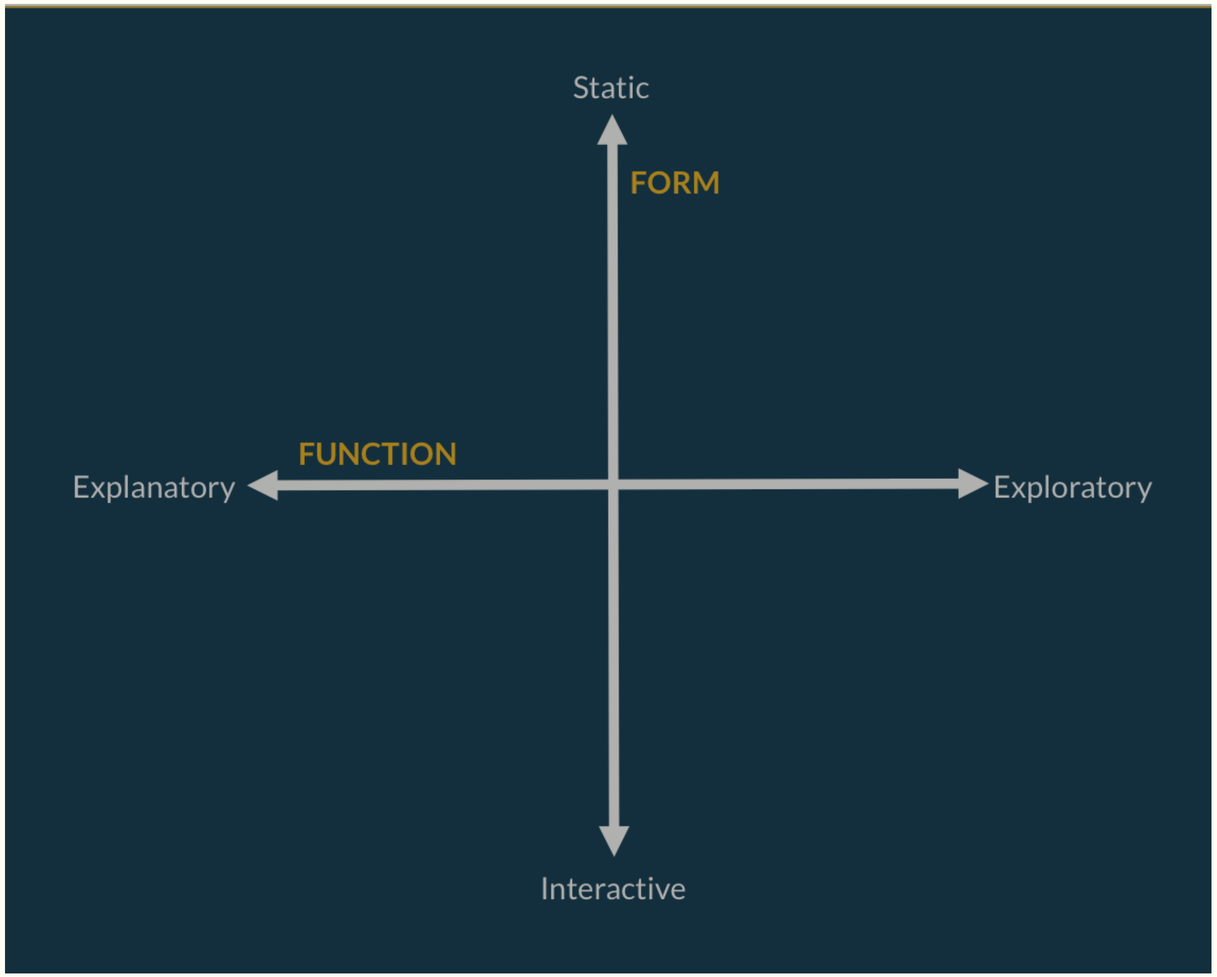
Unstructured Quartet

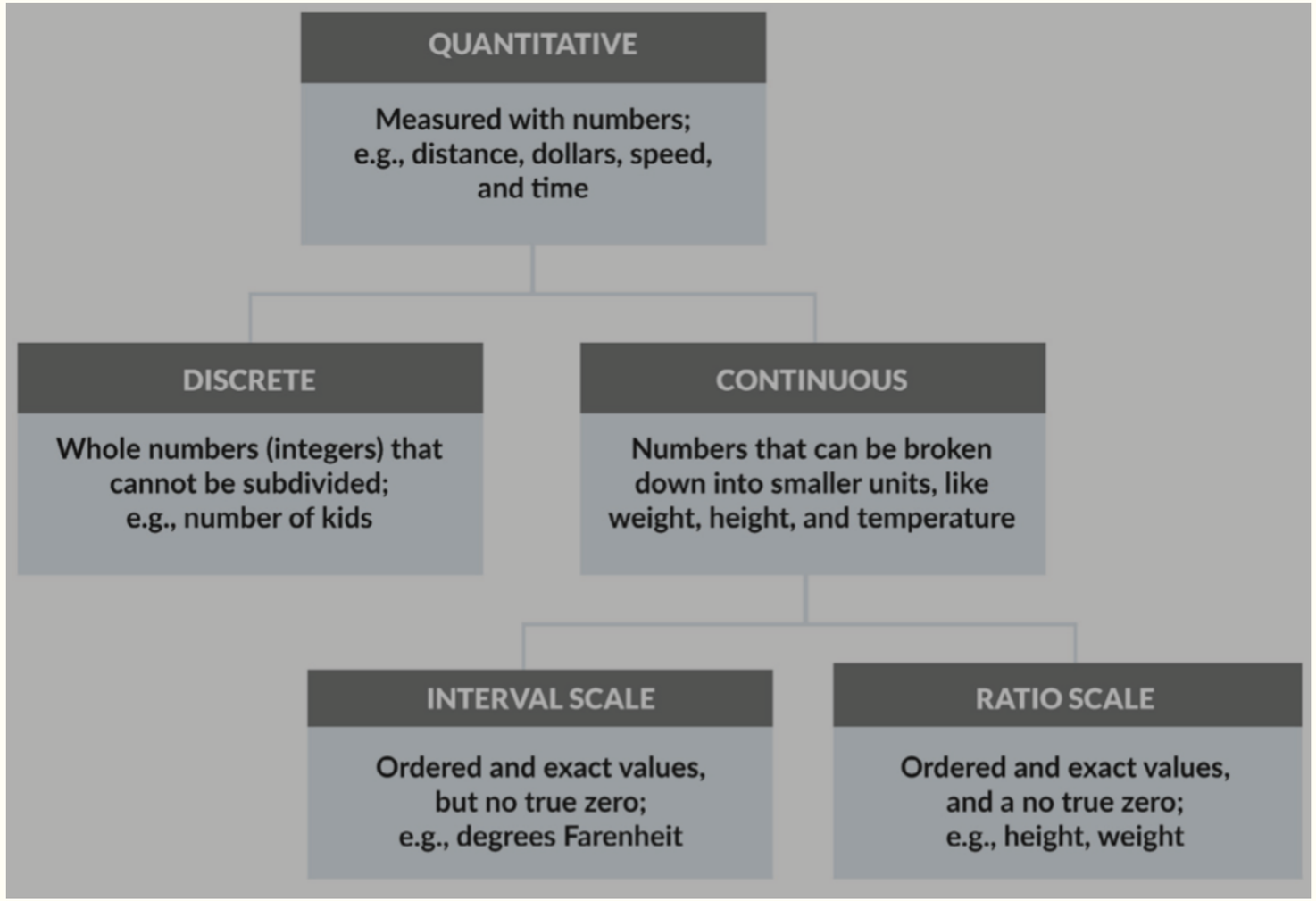
Each dataset here also has the same summary statistics. However, they are not *clearly different* or *visually distinct*.



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

This talk: Static Exploratory



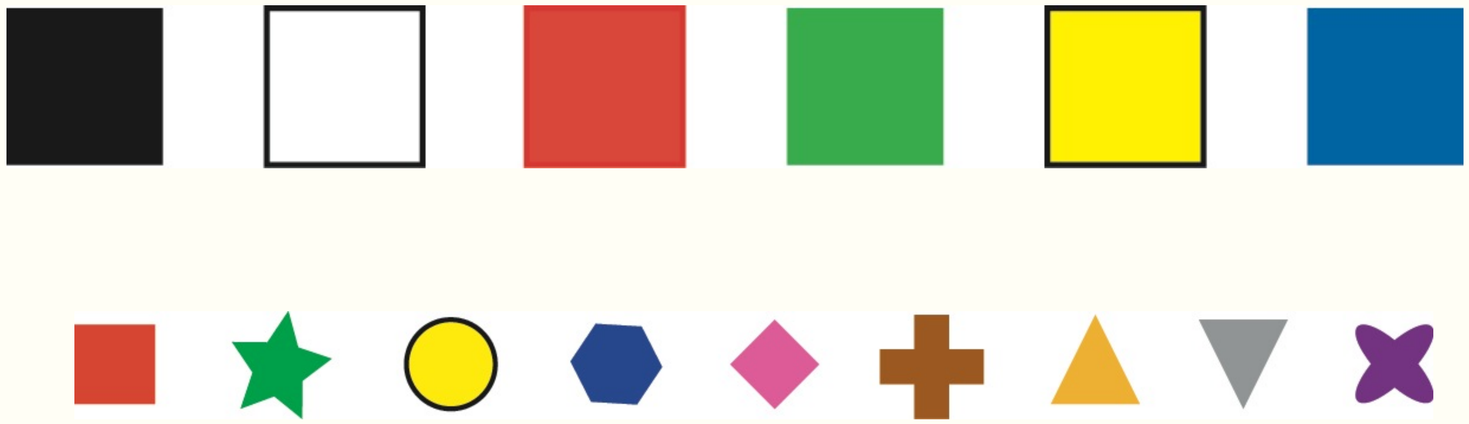

QUANTITATIVE
Measured with numbers; e.g., distance, dollars, speed, and time
DISCRETE
Whole numbers (integers) that cannot be subdivided; e.g., number of kids
CONTINUOUS
Numbers that can be broken down into smaller units, like weight, height, and temperature
INTERVAL SCALE
Ordered and exact values, but no true zero; e.g., degrees Farenheit
RATIO SCALE
Ordered and exact values, and a no true zero; e.g., height, weight

Data Types

Use Color Sensibly

Colors for Categories

Do not use more than 5-8 colors at once



Ware, “Information Visualization”

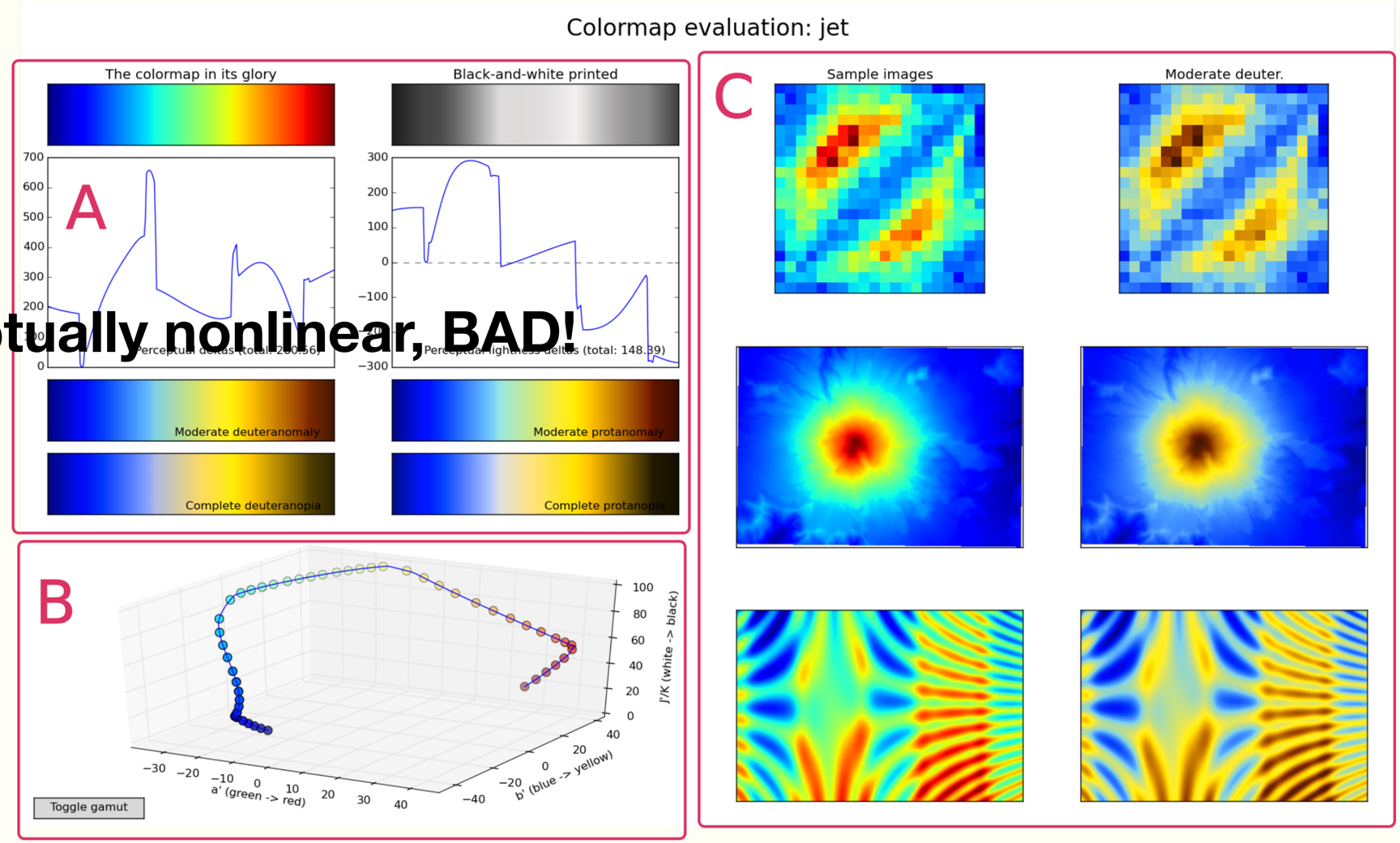
Colors for Ordinal Data

Vary luminance and saturation (see HSL Scale)

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

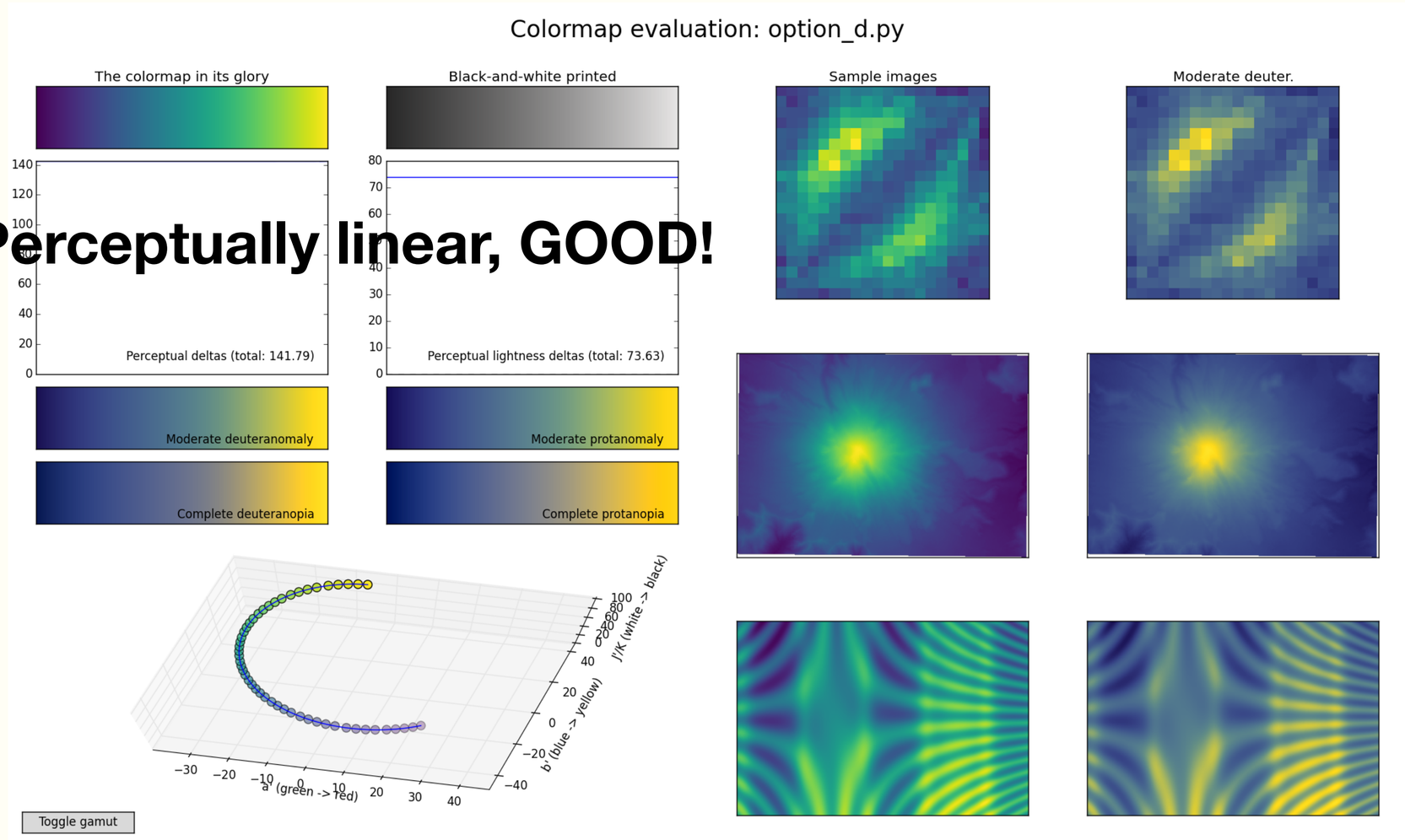
Quantitative data colors: Rainbow Colormap

Perceptually nonlinear, BAD!



R. Simmon

Viridis



Perceptually linear, GOOD!

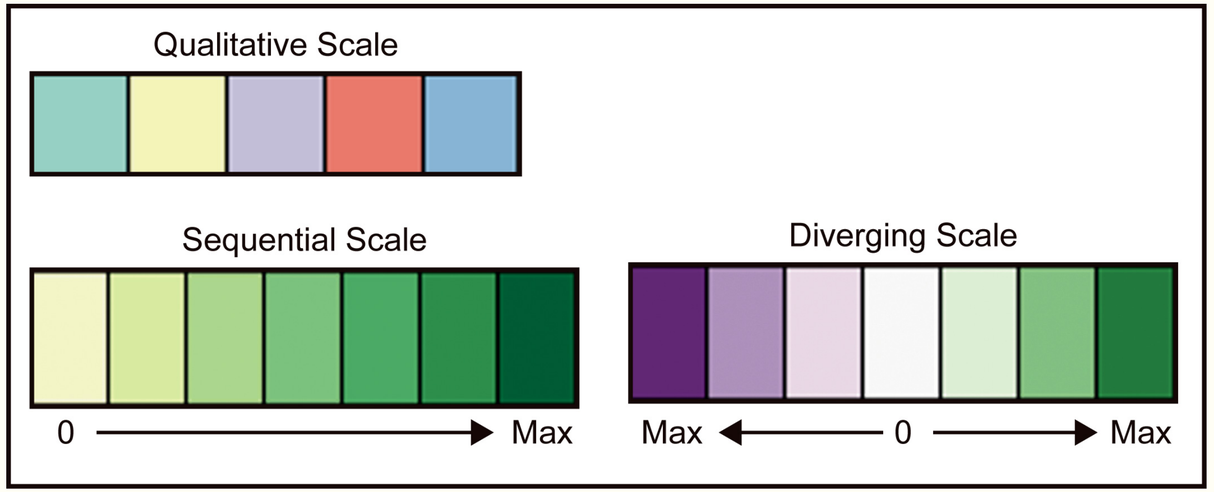
Color Scales

See colorbrewer2.org

Qualitative palettes are used for multiple curves in a plot or categoricals

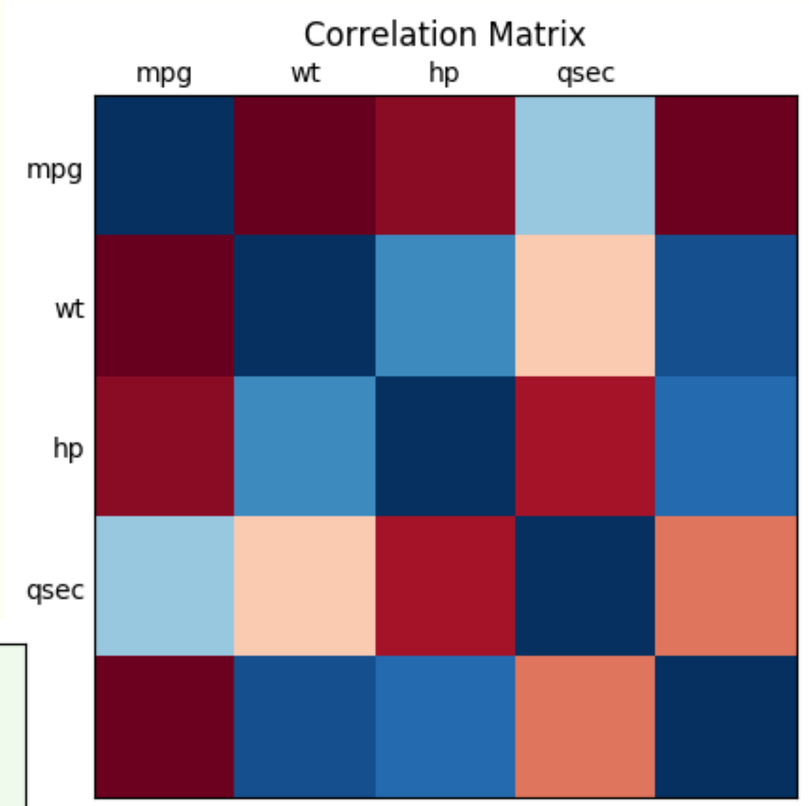
Nominal

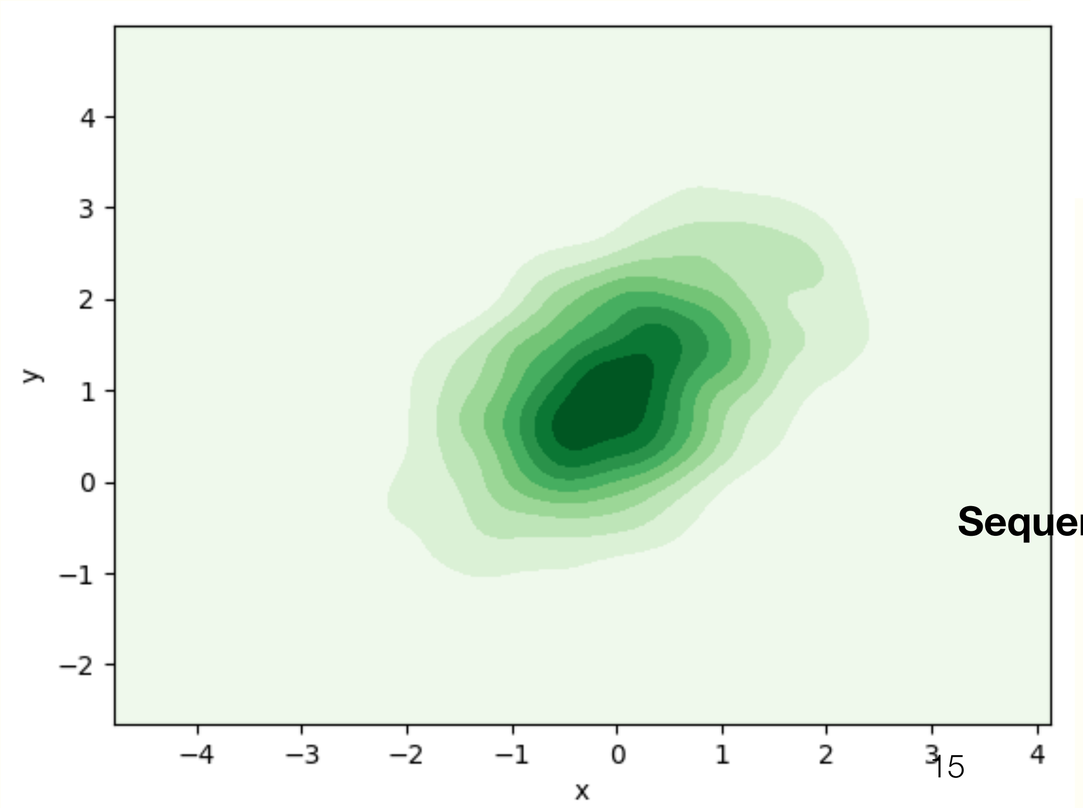
Ordinal or Quantitative



Cynthia Brewer, Color Use Guidelines for Data Representation

Diverging Palette for Correlations





Sequential Palette for Densities (all positive)

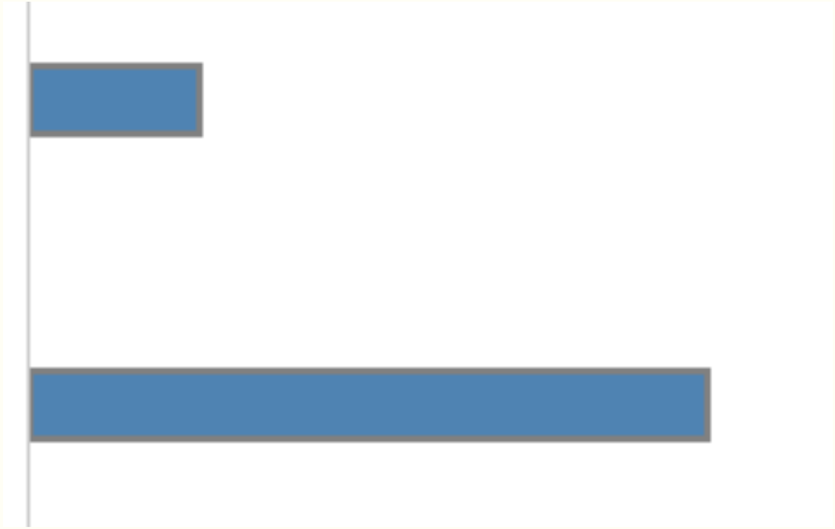
Perceptual Effectiveness

from cs109

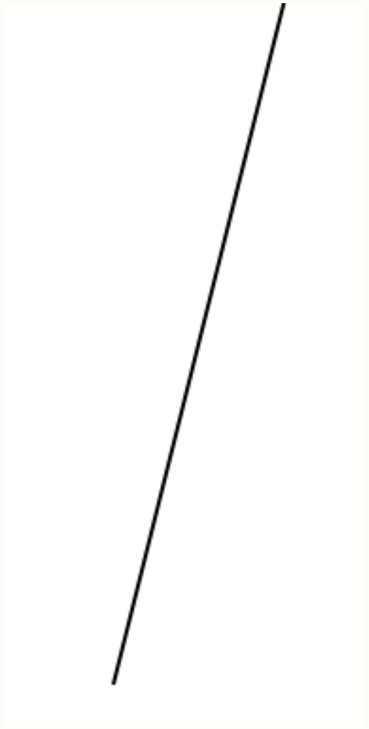
How much longer?

A

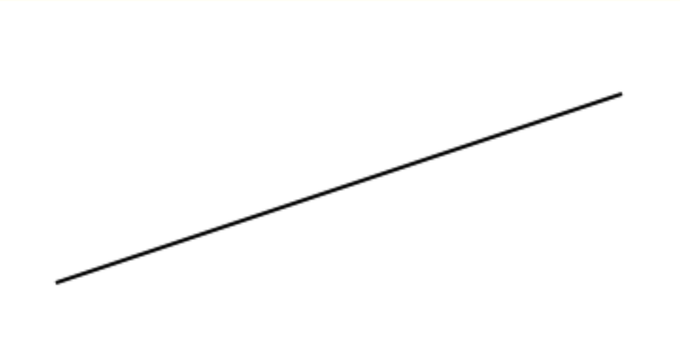
B



How much steeper the slope?

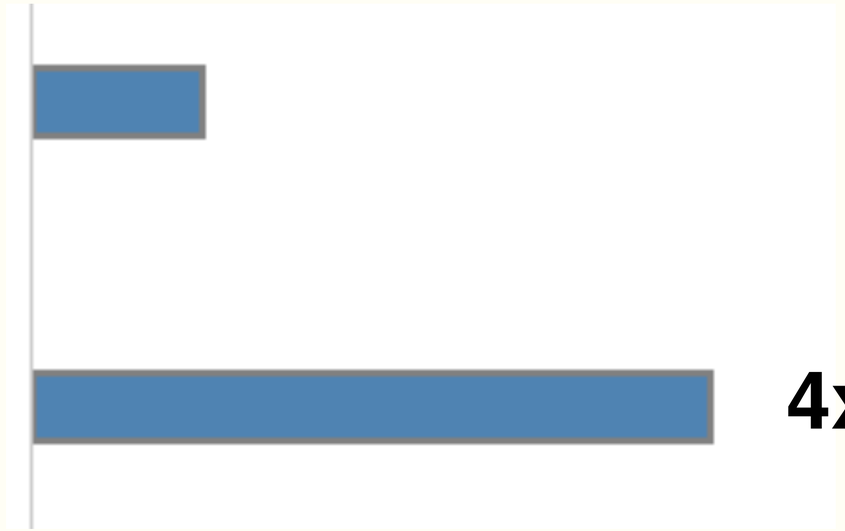


A



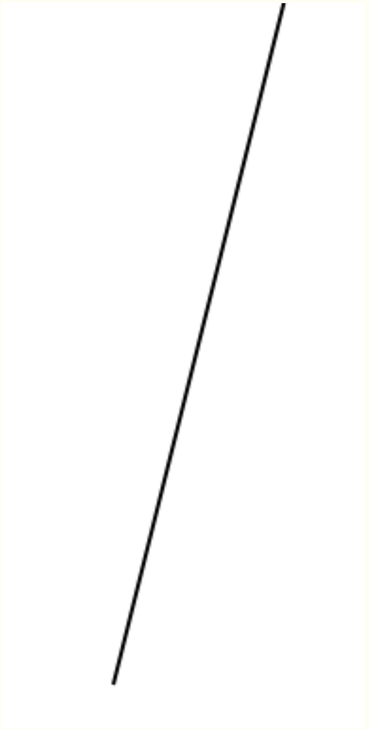
B

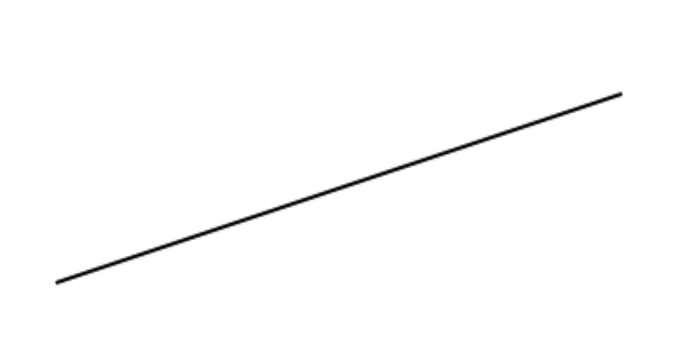
How much longer?



4x

How much steeper the slope?



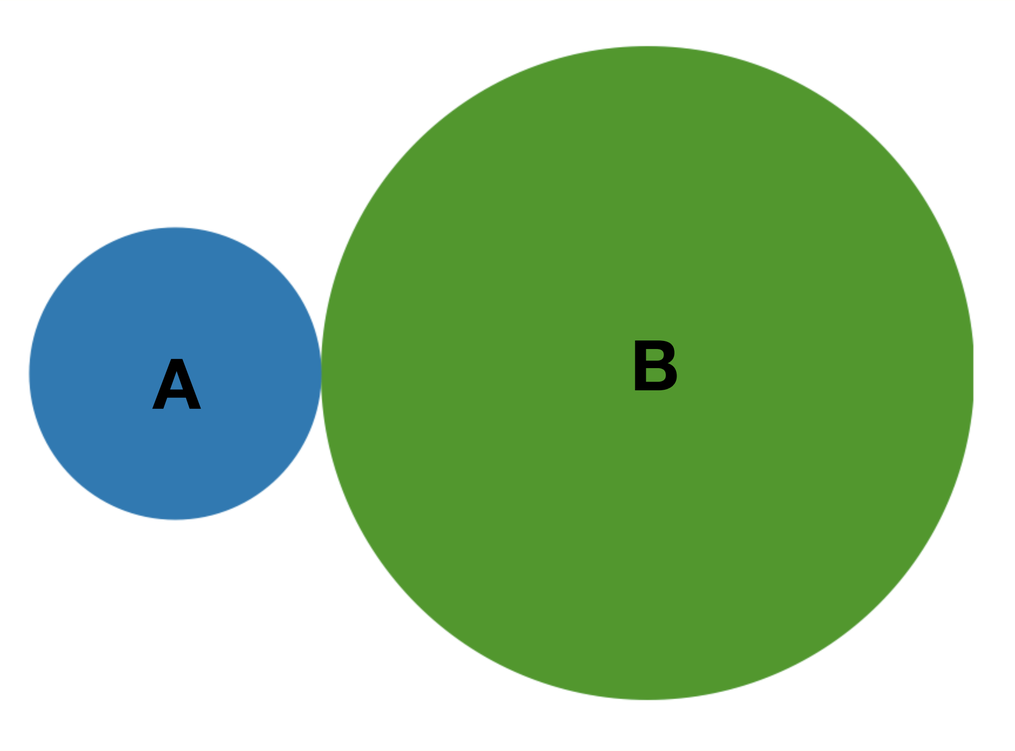


A

B

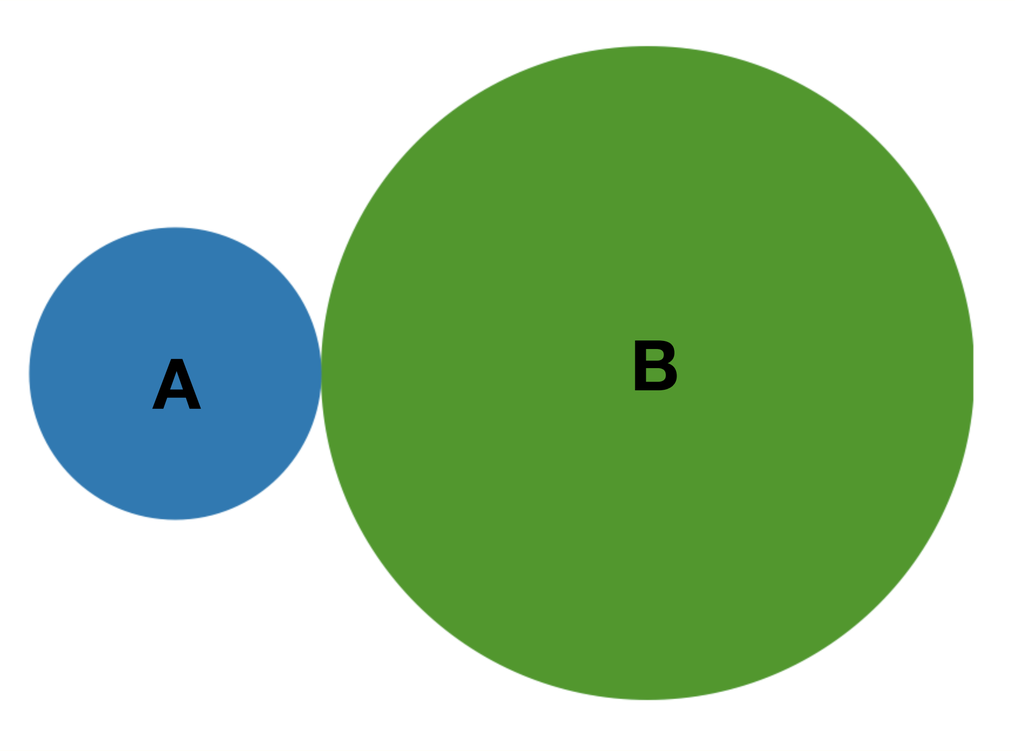
4x

How much larger area?

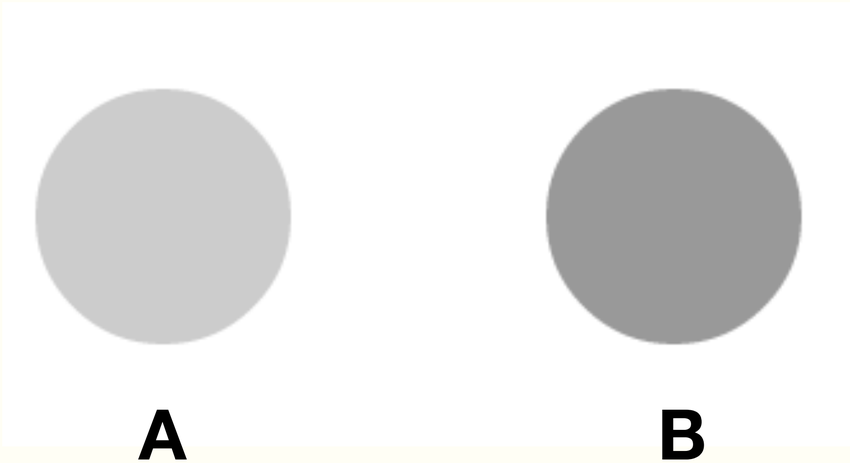


How much Darker?

How much larger area?



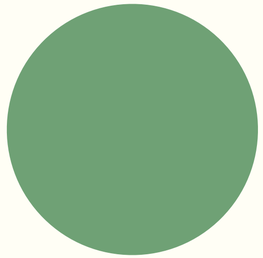
10x



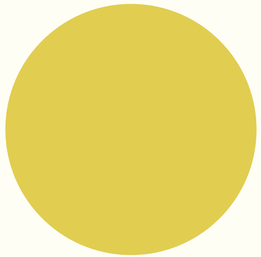
2x

How much Darker?

How much bigger value?



A



B

How much bigger value?

A

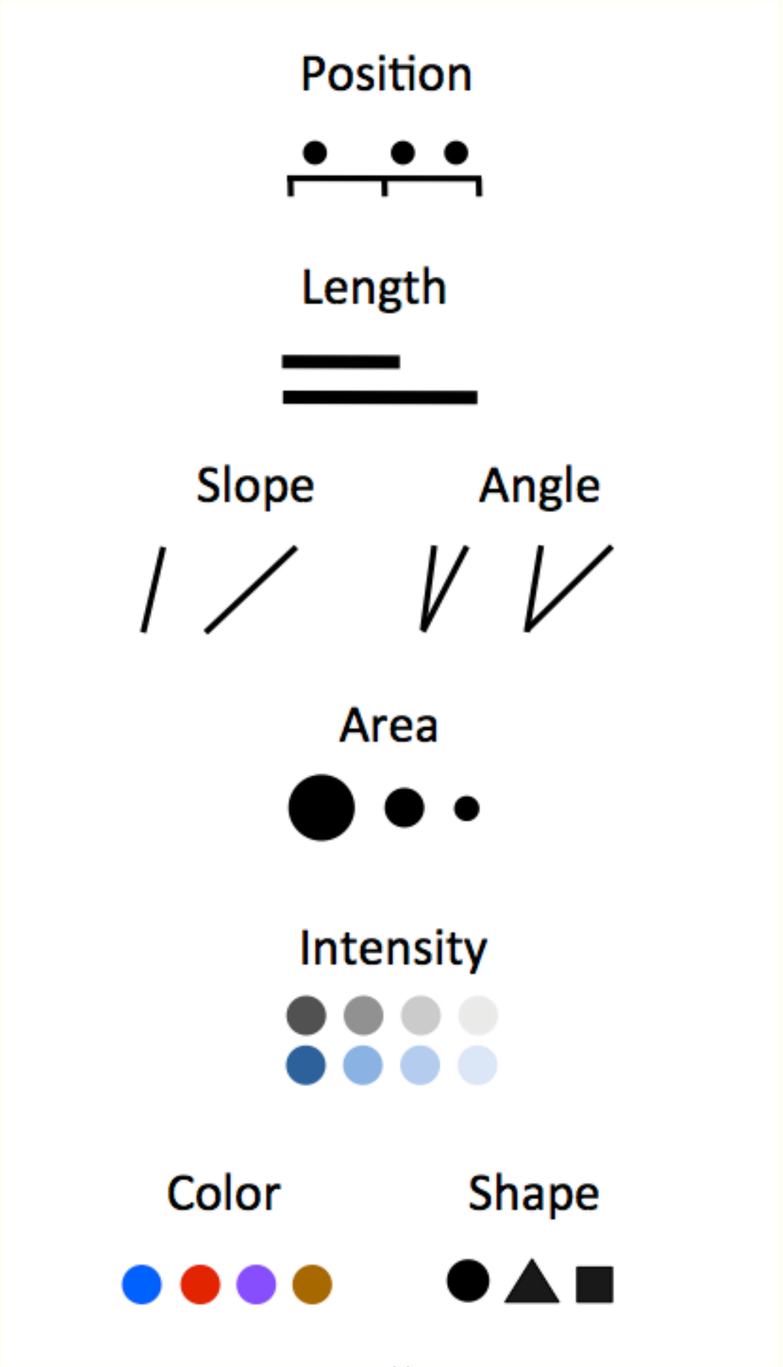
B

4x

2

16

Most Efficient



Least Efficient

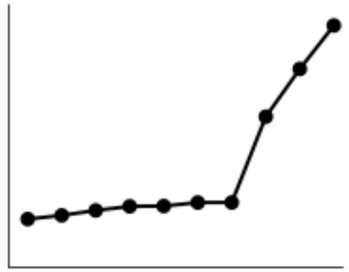
Quantitative

Ordered

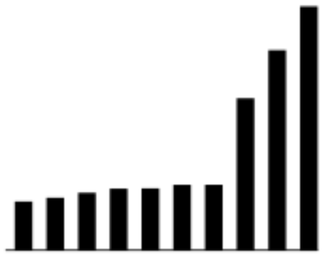
Categories

C. Mulbrandon
VisualizingEconomics.com

Most Effective

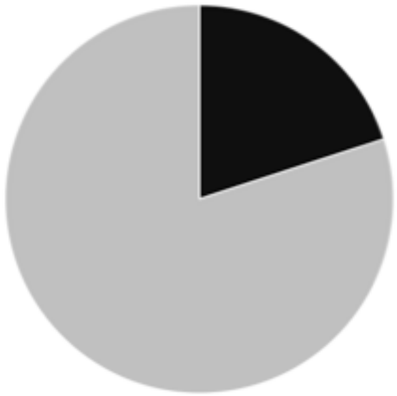


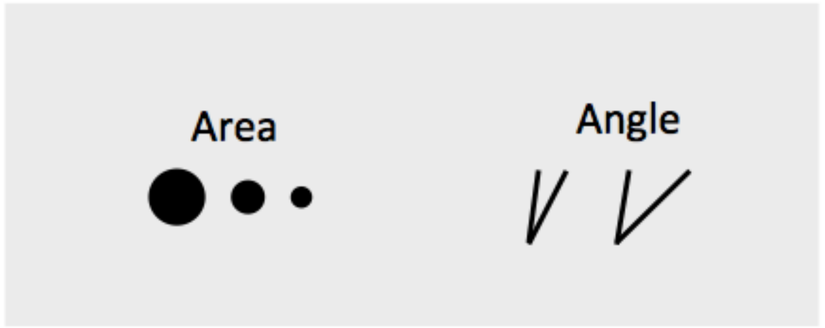
Position



Length

Less Effective





EDA: Which plots when?

EDA Process

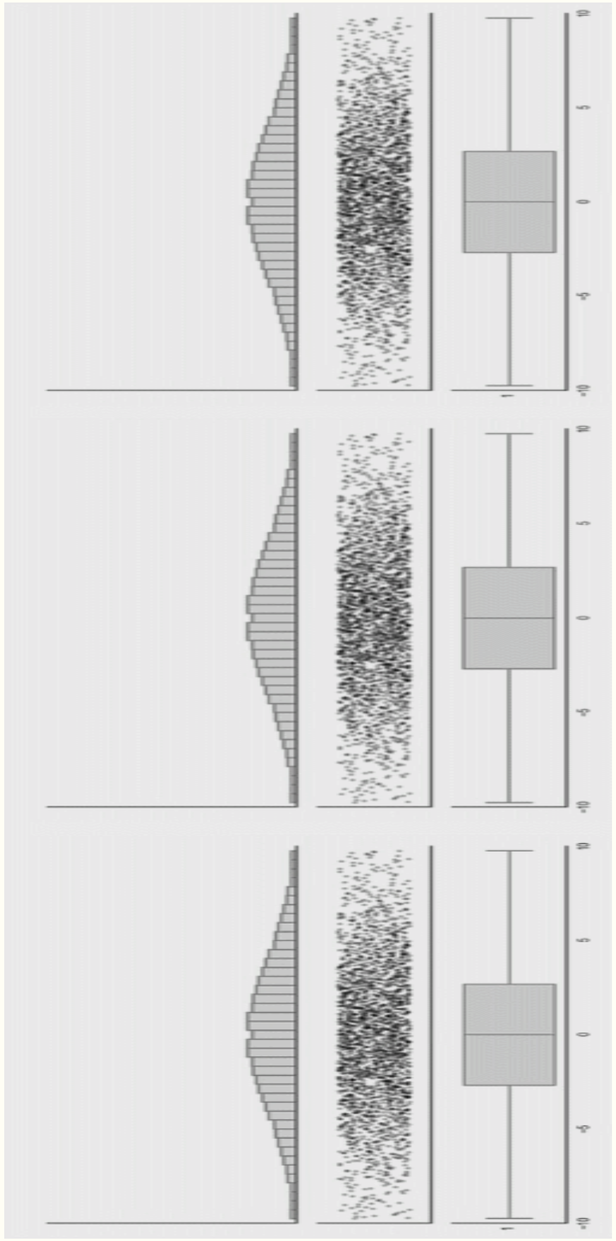
1. Visualize and Summarize individual features
2. Viz and summarize combination feature and target distributions
3. Transform targets and features to make them more amenable to modeling.

EDA Rubric

1. Explore **global properties**. Use histograms, scatter plots, parallelism, and aggregation functions to summarize the data.

2. Explore **group properties**. Use groupby and faceting and small multiples to compare subsets of the data.

3. Transform targets and features to make them more amenable to modeling.

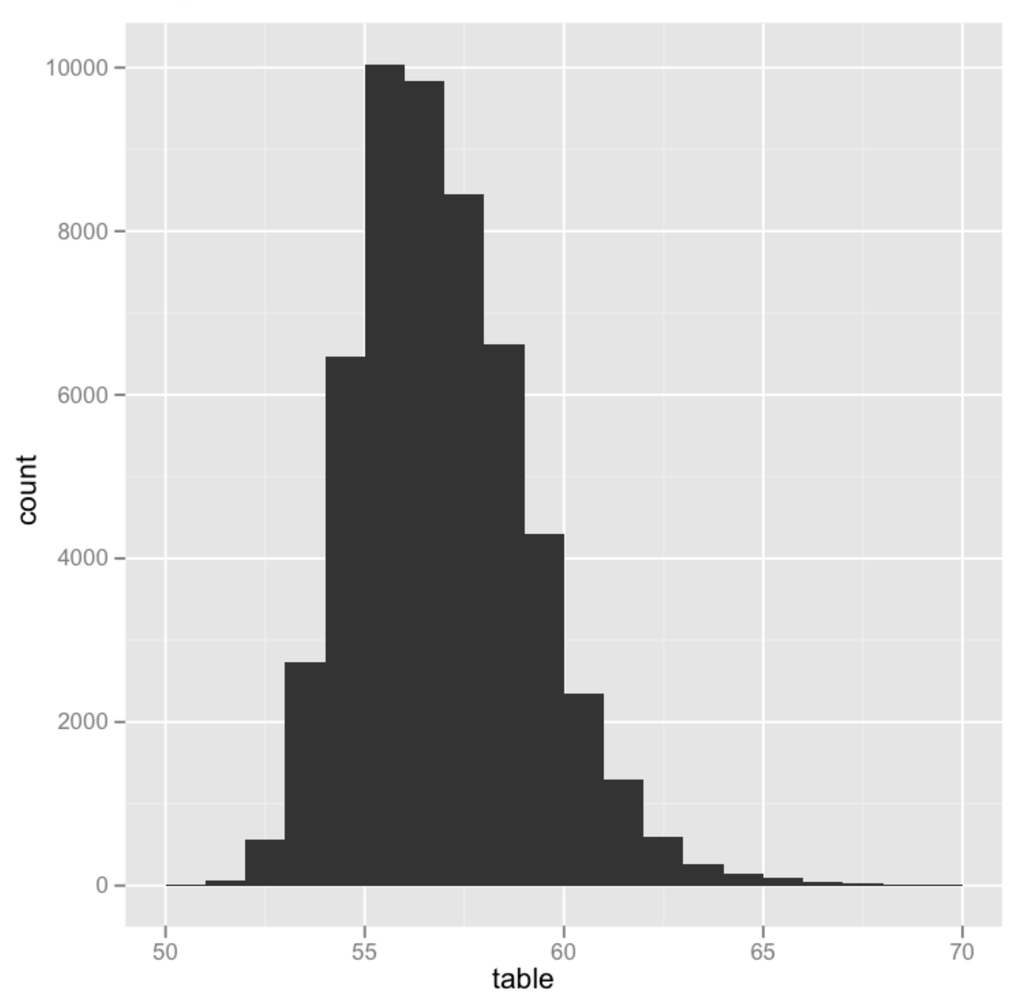


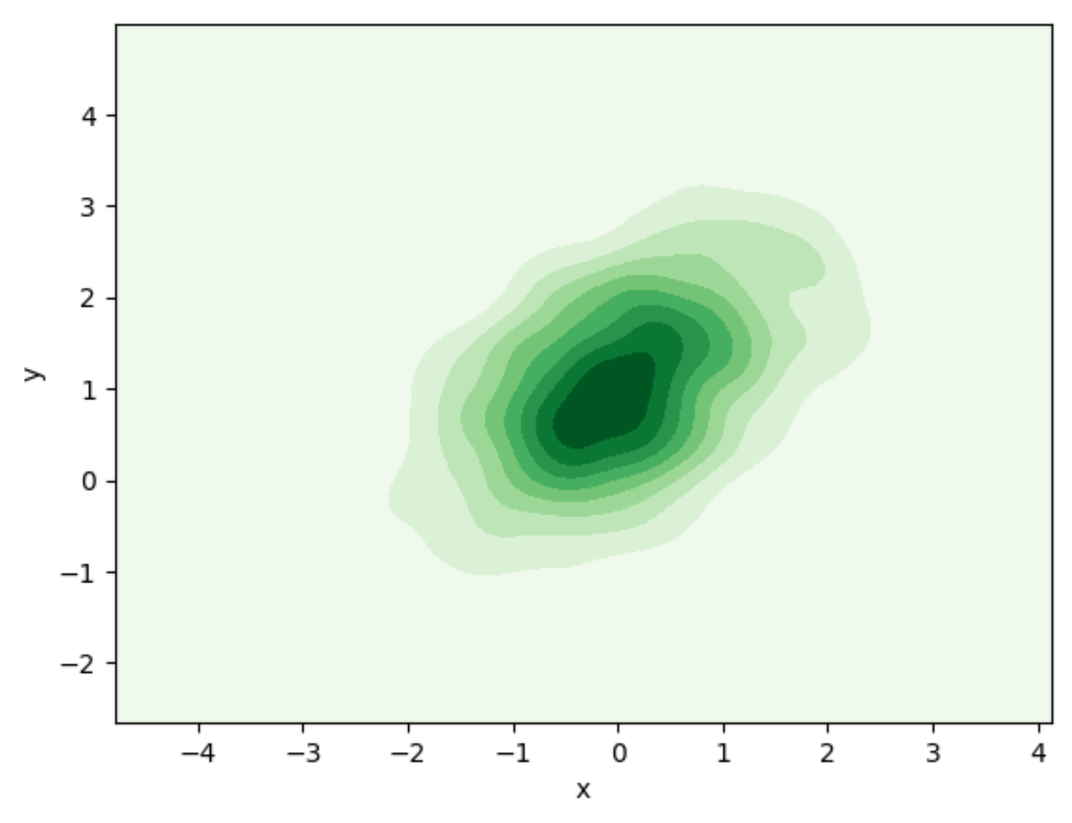
Simple Summaries: Box Plots

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

From Summaries to Distributions

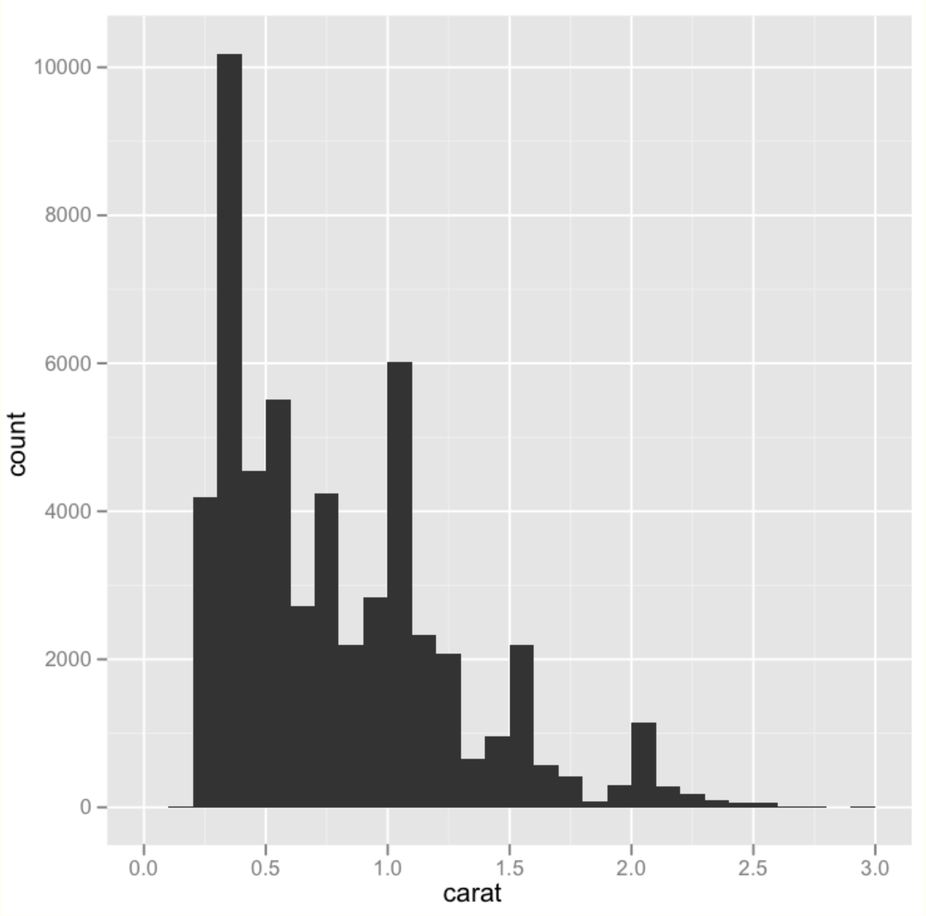
Histogram



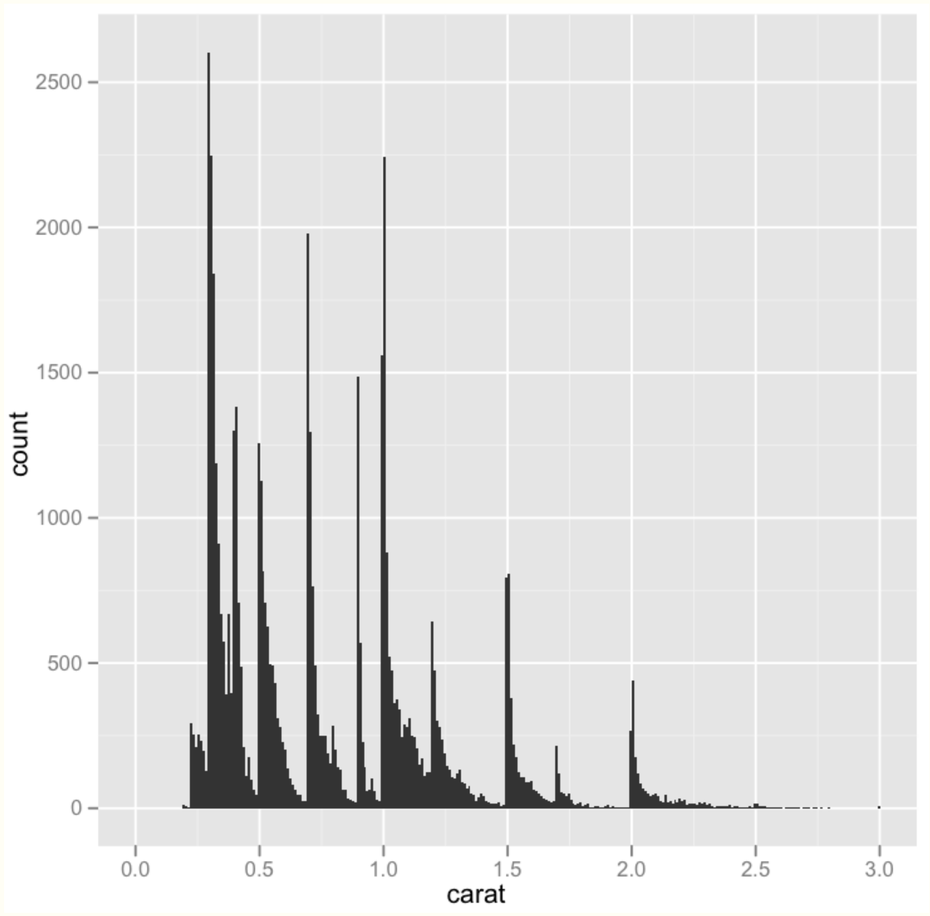


2-D Kernel Density Estimate (KDE)

Bin Width is Important

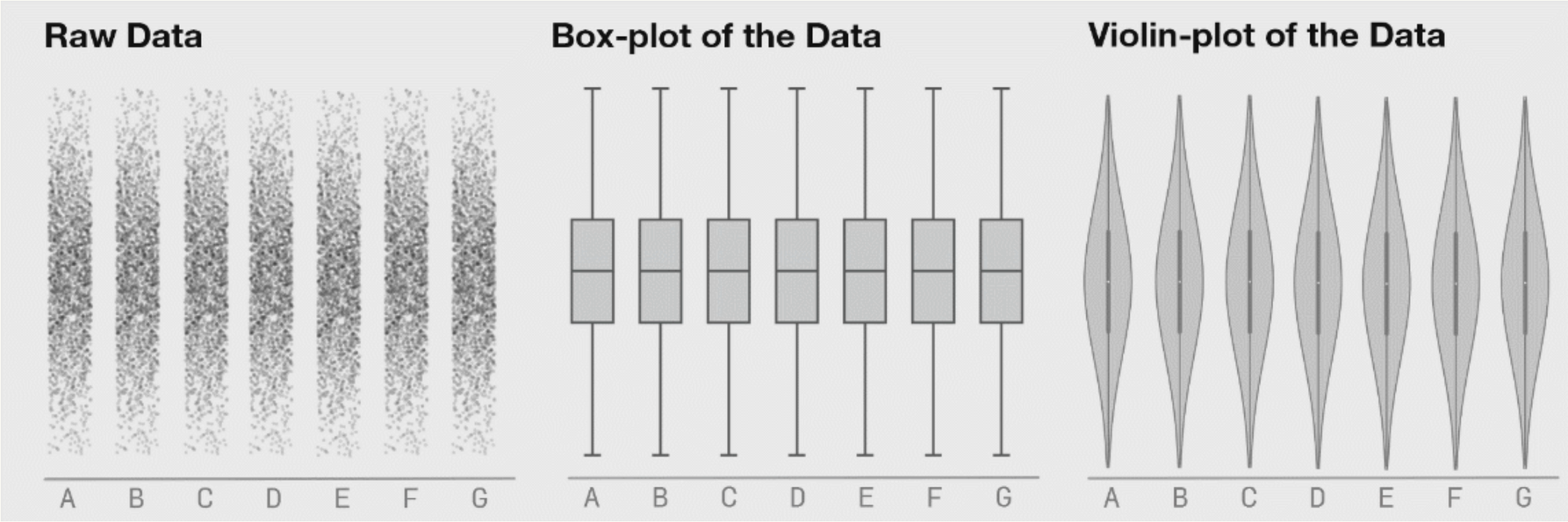


binwidth = 0.1



binwidth = 0.01

Summary Distributions: Density Plots



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

Design Exercise

Hands-On Exercise

from cs109

How do you feel about doing science?

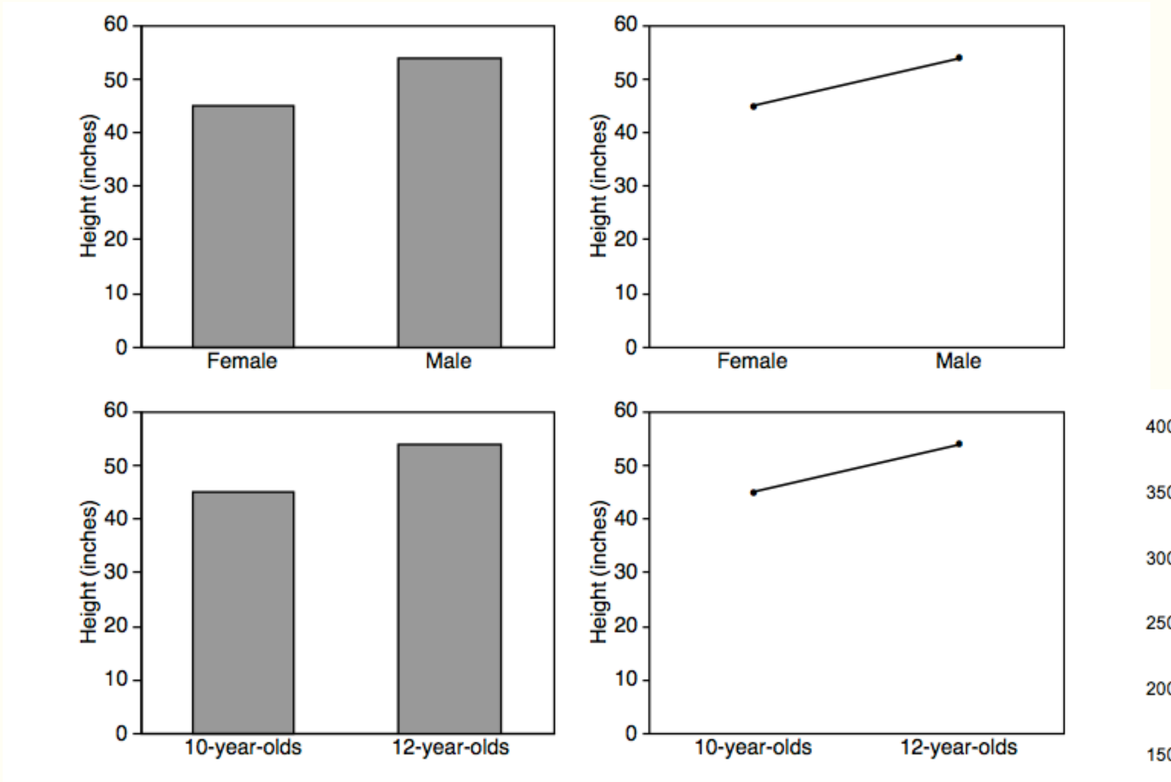
Interest	Before	After
Excited	19	38
Kind of Interested	25	30
OK	40	14
Not Great	5	6
Bored	11	12

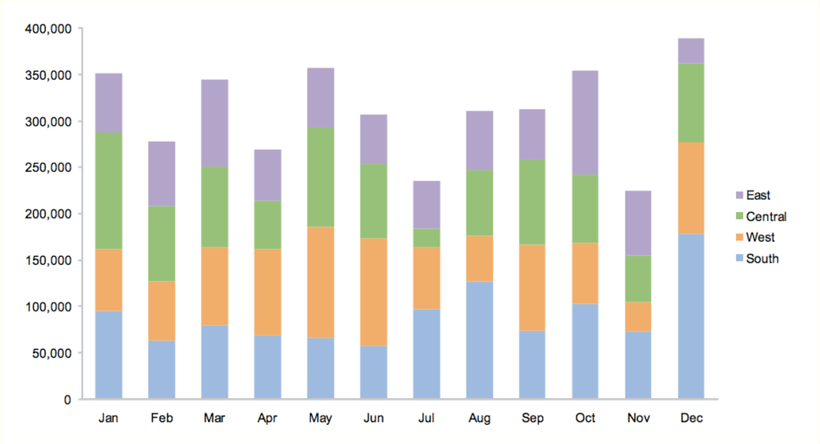
Data courtesy of Cole Nussbaumer

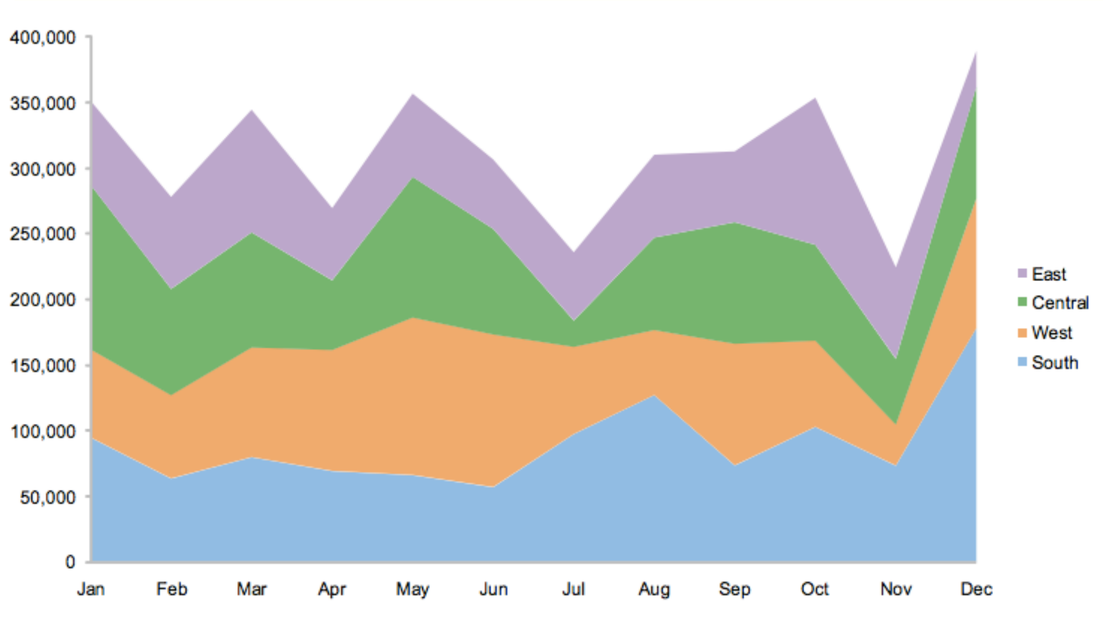
Come up with multiple visualizations.

Pen and Paper Only. Pies, Bars, What else?

Grouping

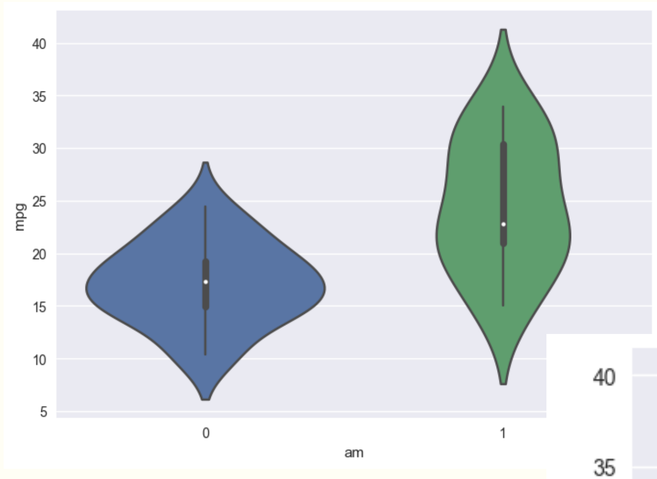


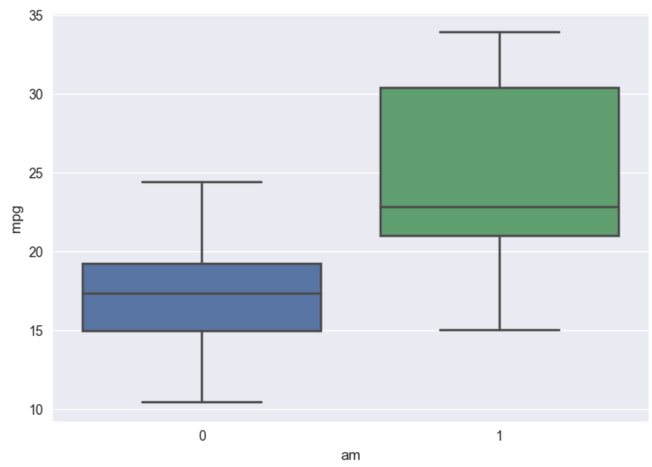




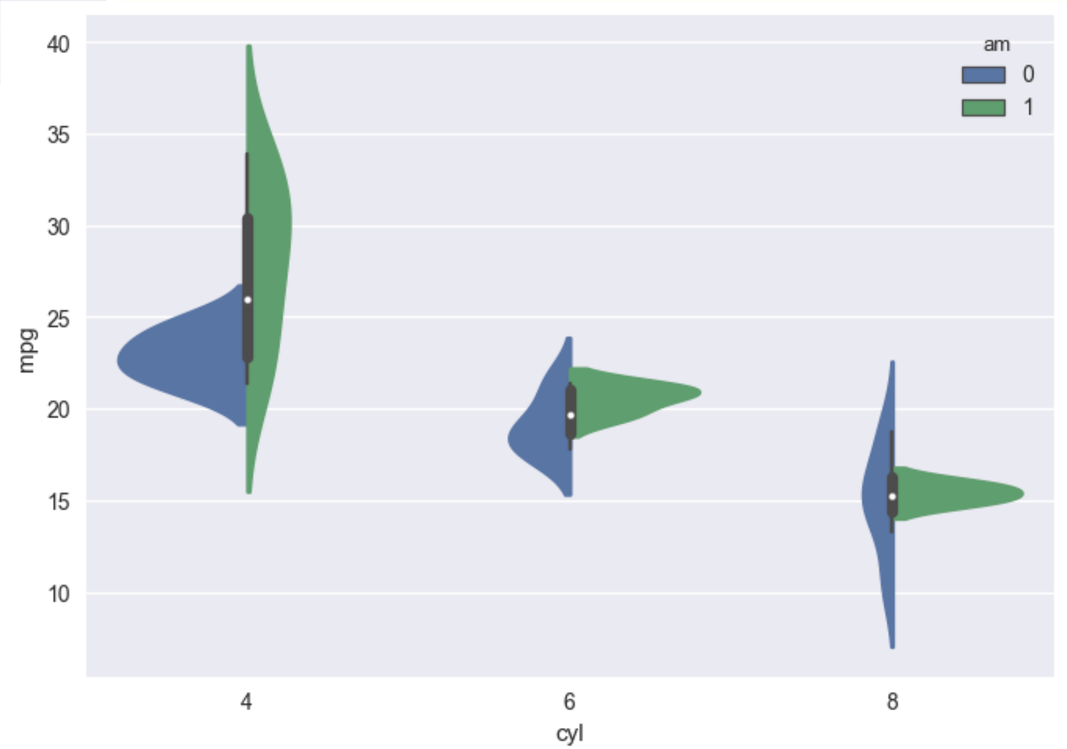
from cs109

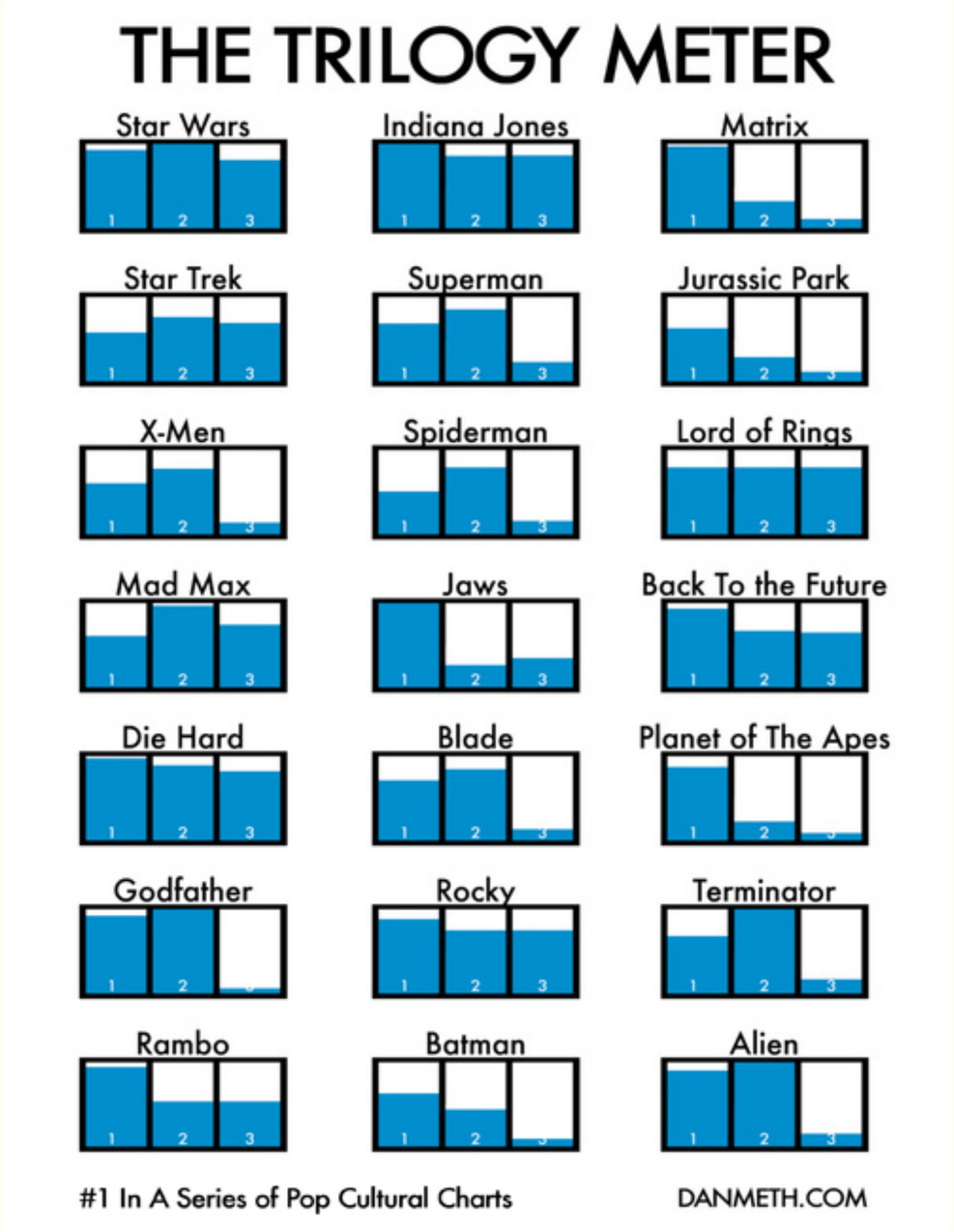
GROUP





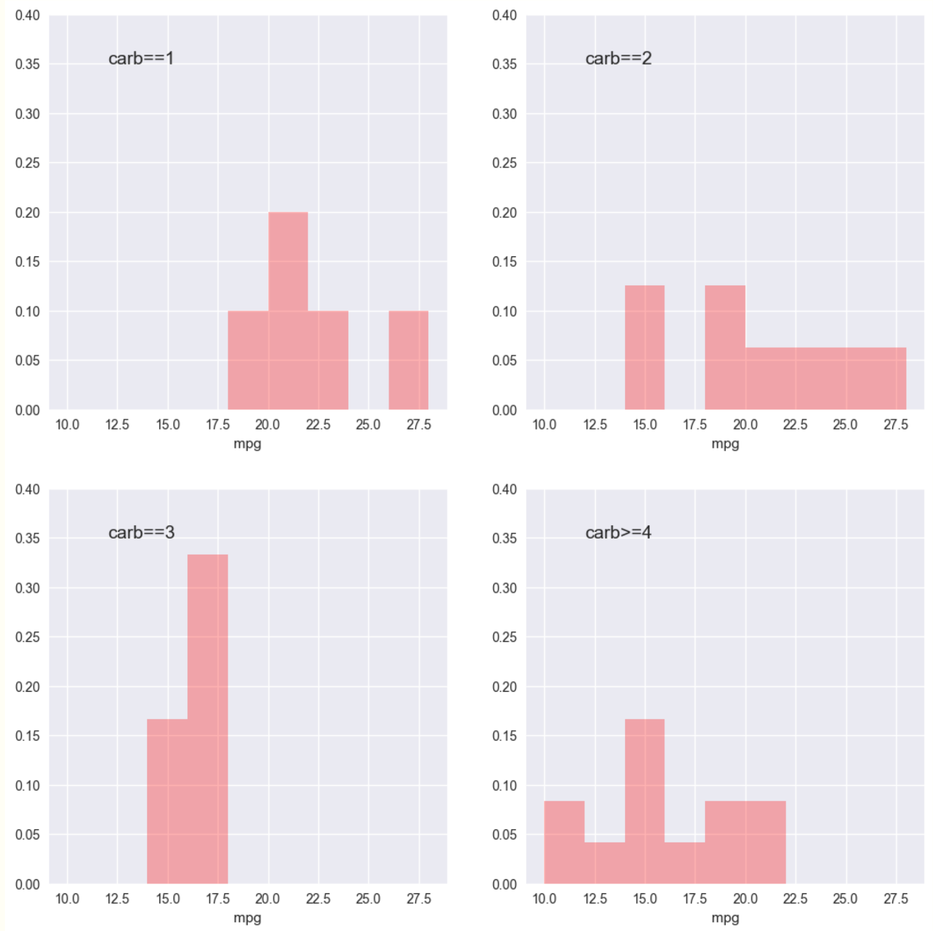
getting complex…

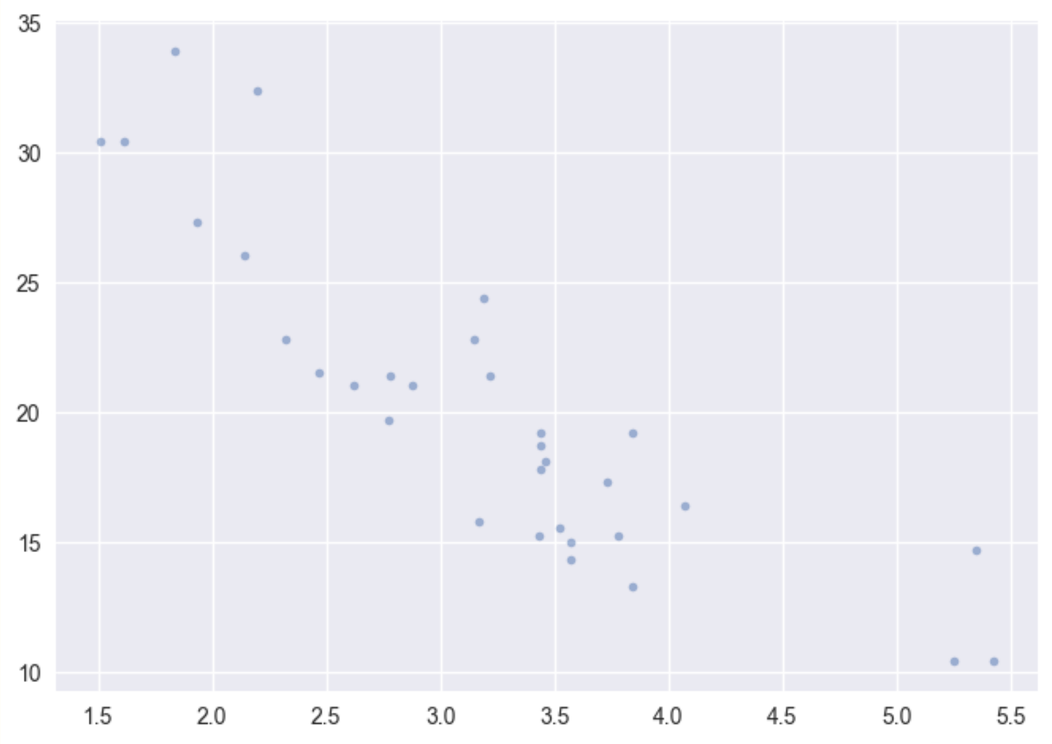




Small multiples

Use seaborn or
multiple plots in matplotlib

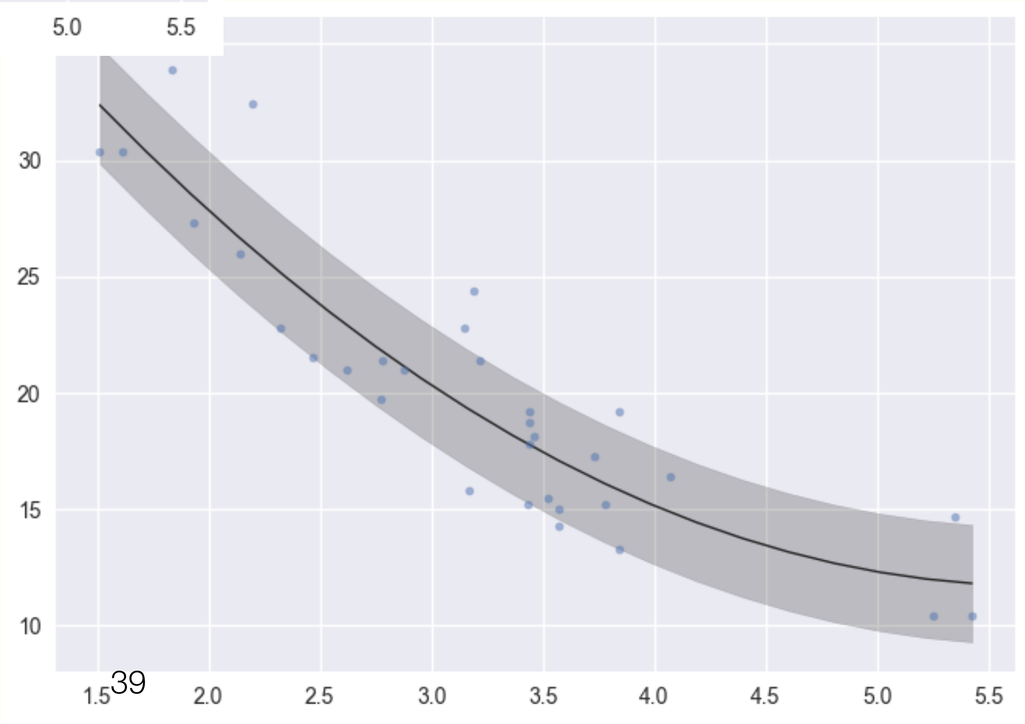




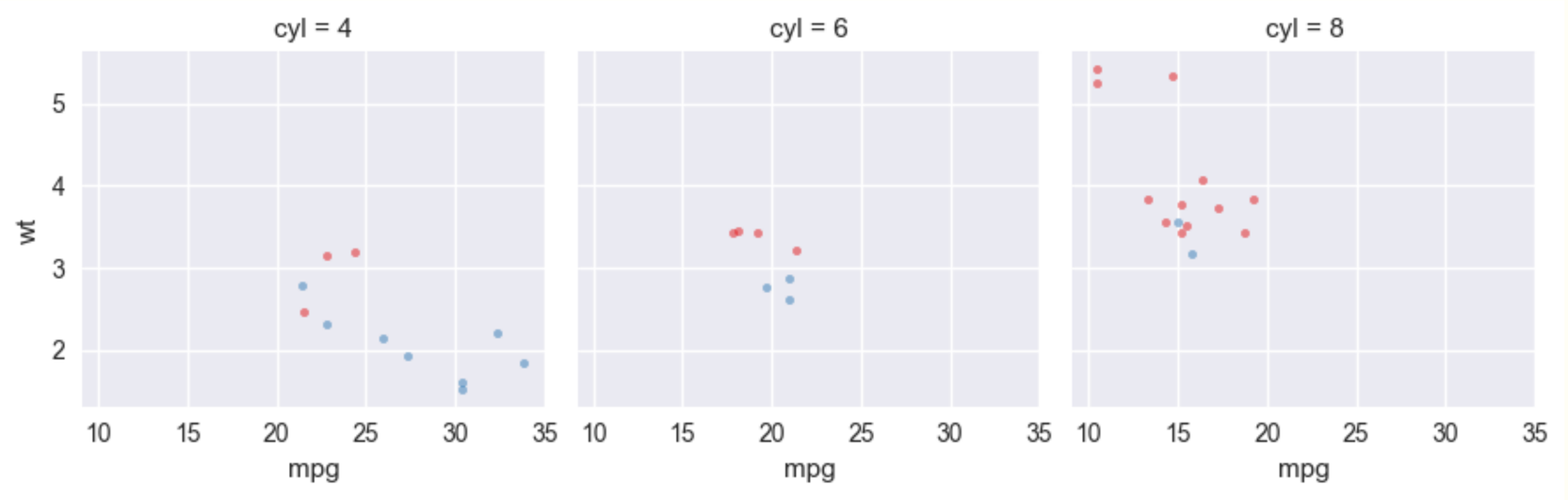
Relationships

Use lines when adjoining points involve some continuity e.g. time

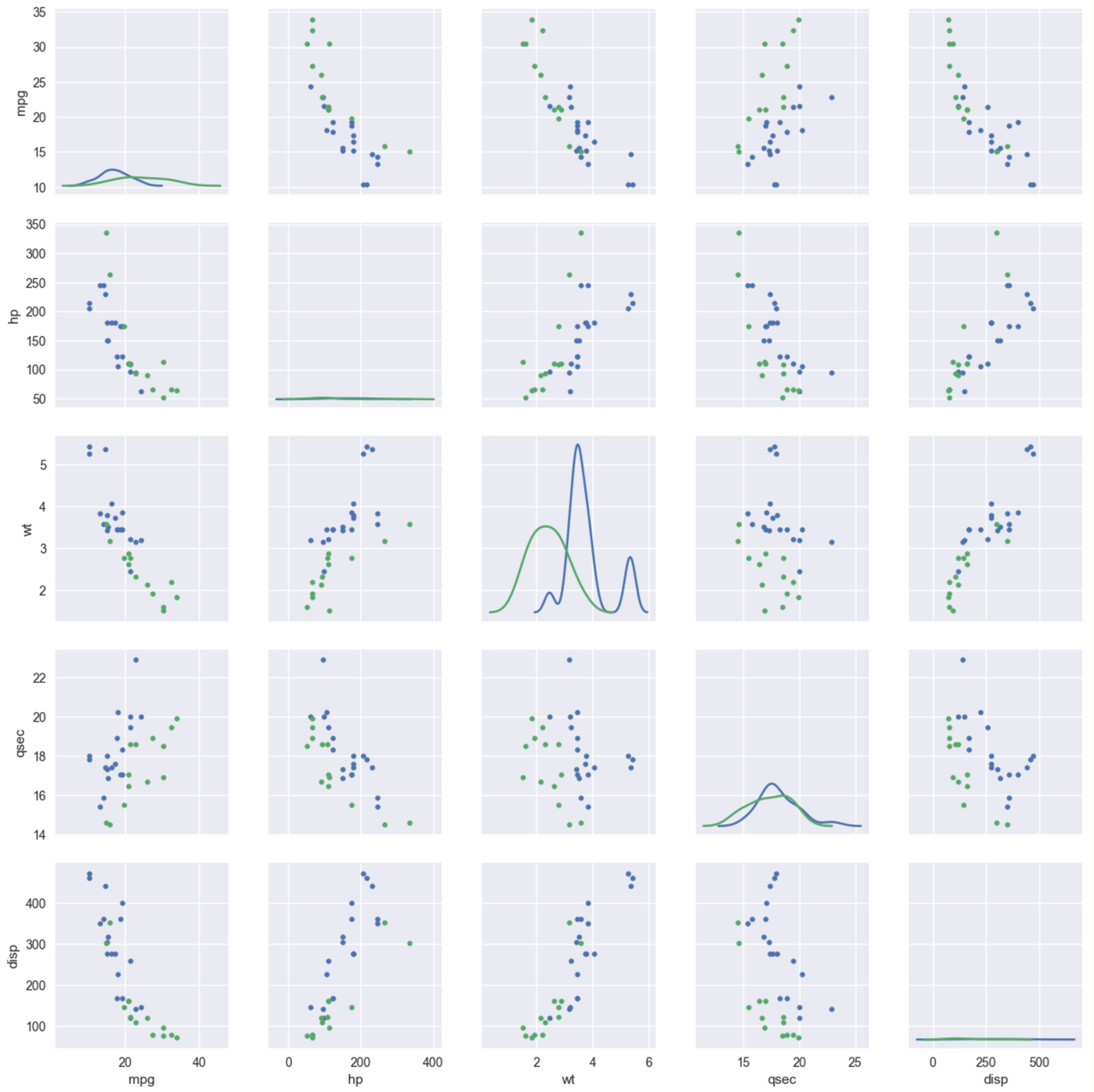
Scatter Plots to discover relationships.



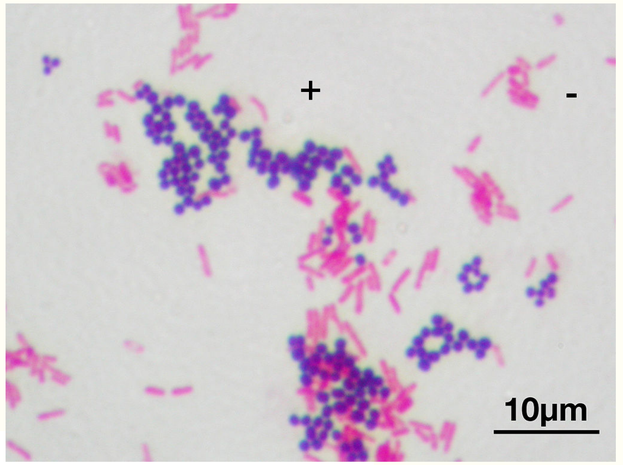
Relationships with grouping



SPLOM


mpg
hp
wt
qsec
disp
35
30
25
20
15
10
350
300
250
200
150
100
50
5
4
3
2
22
20
18
16
14
400
300
200
100
20
40
0
200
400
2
4
6
15
20
25
0
250
500

Example: Antibiotics Will Burtin, 1951



Genus, *Species*

Min. Inhibitory Concentration [ml/g]

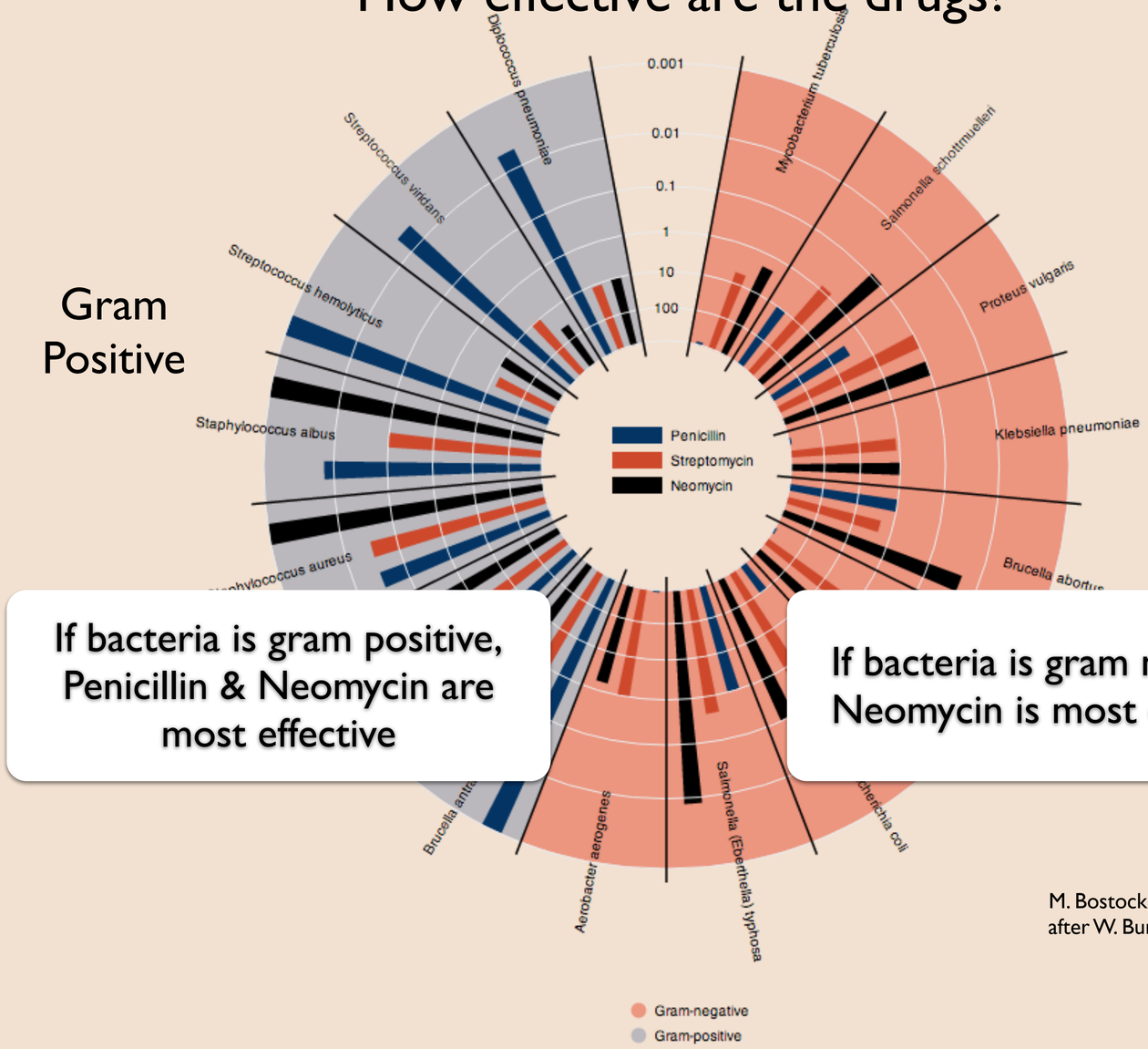
Table 1. Burtin's data.

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

How effective are the drugs?

Gram Positive

Gram Negative



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

If bacteria is gram negative, Neomycin is most effective

M. Bostock, Protovis
after W. Burtin, 1951