Visual Analytics– Visualization 101

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

Plan for Today

- Visualization overview
- Graphical primitives
- Visual dimensions
- Common visualization techniques
- *Time permitting*: Tableau demo

Reminder

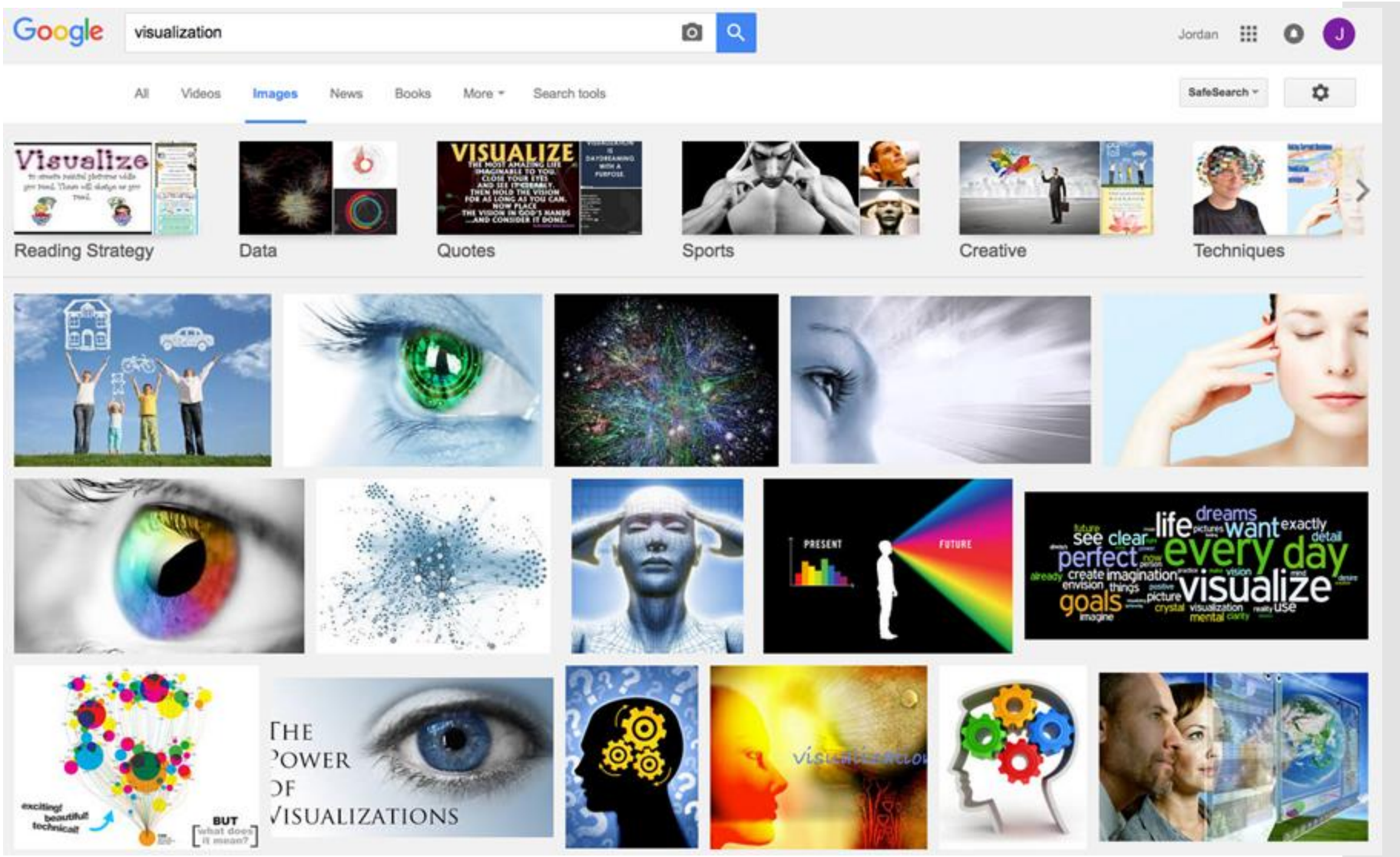
- hw01 is released today! (https://amosca01.github.io/SDS-CS235/ > Homework > hw01)
- Due next Thursday (09/19) at midnight
  - You have extensions if you need them, but you MUST tell me if you're taking one
  - Revise and resubmit also exists!
- Work with a small group (3-4) – I recommend finding people with *complementary skillsets* to work with

**I won't always remind you there is a homework released/due, make sure you stay up to date with the course schedule! **

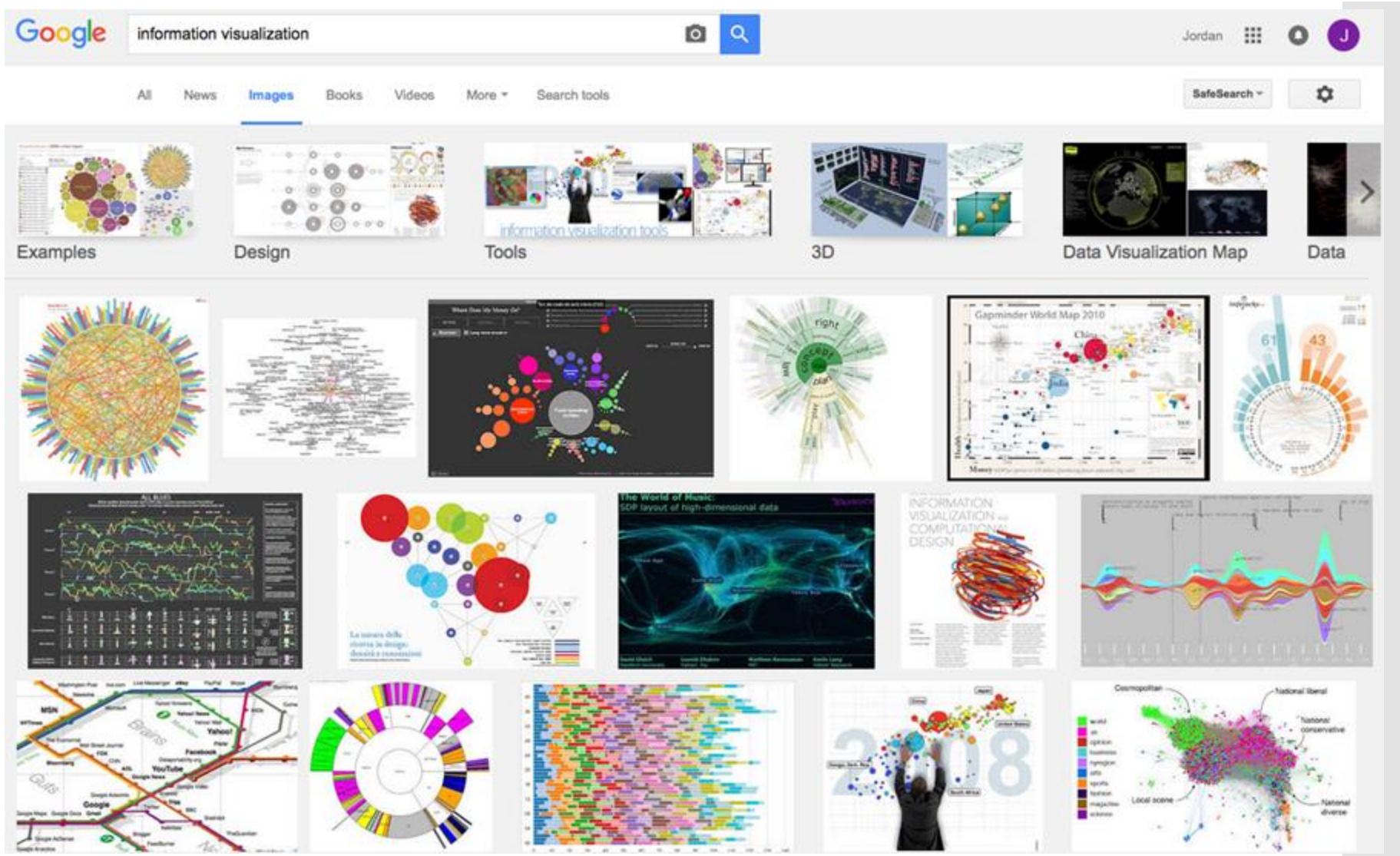
Visual Analytics

Visualization ∩ Analysis

What is visualization?



What is visualization?



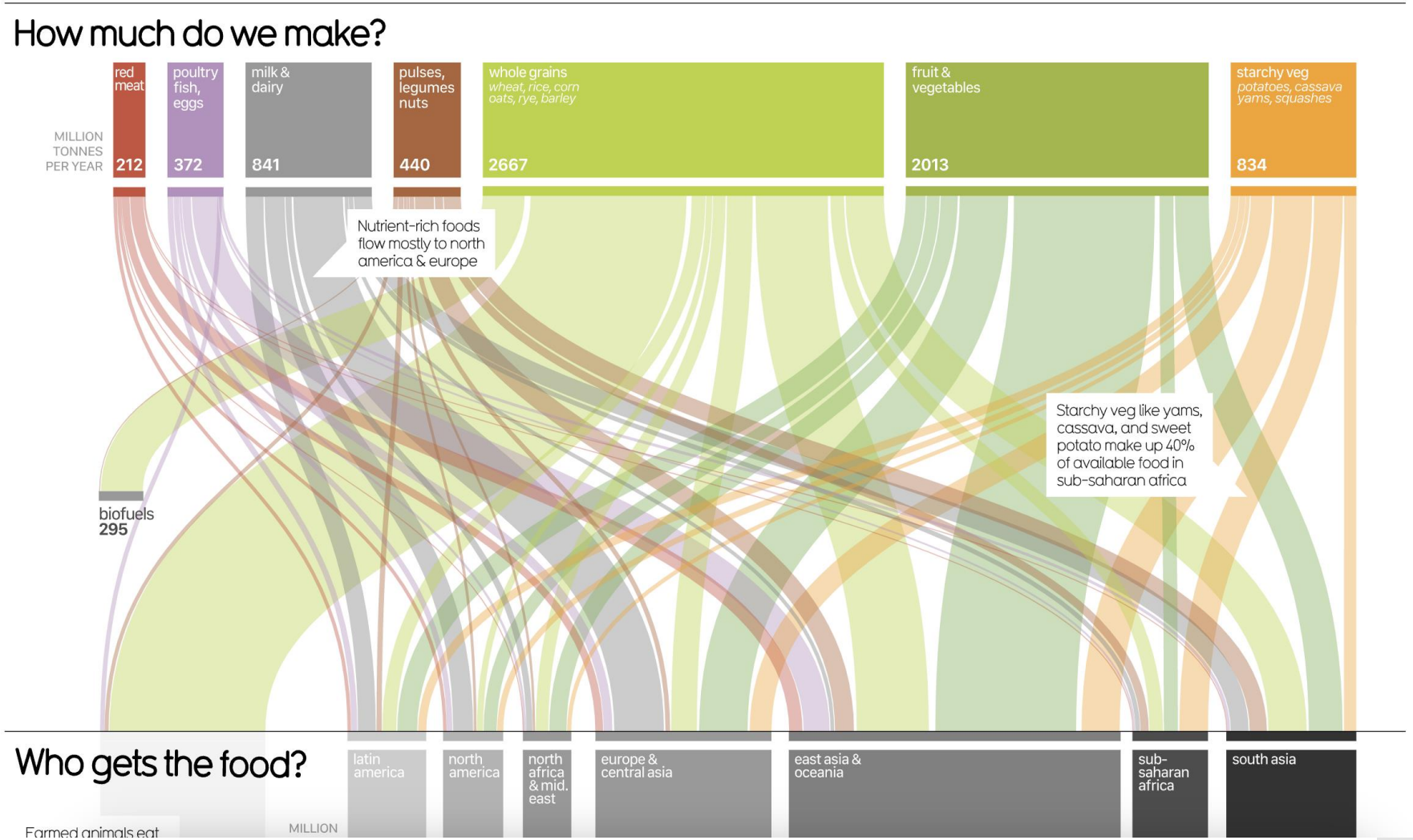
Perhaps a more helpful question:

What are some ways
a “visualization” can be **useful**?

Does it help you spot trends?

More info here: http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Does it help you explore?



https://informationisbeautiful.net/visualizations/global-food-supply-where-does-all-the-worlds-food-go/

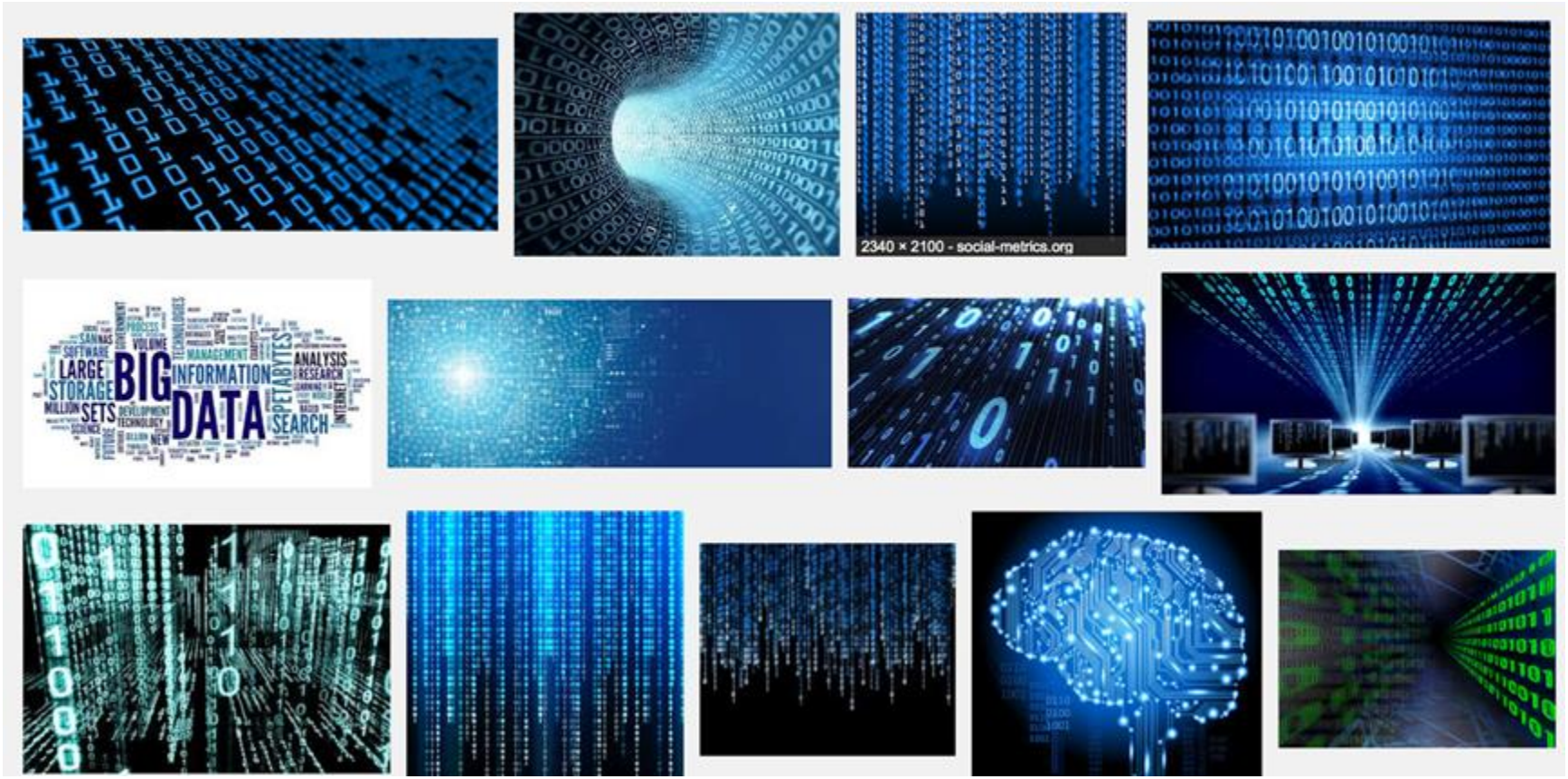
Does it tell a story?

Hans Rosling's 200 Countries, 200 Years, 4 Minutes – The Joy of Stats – BBC Four
https://www.youtube.com/watch?v=jbkSRLYSojo

Visualization (def.)

Visual **representations** of data that reinforce human **cognition**

Wait… what is “data”?



Data: a definition

Data is a set of *variables* that capture various aspects of the world:

Tuition rates, enrollment numbers, public vs. private, etc.

Data: a definition

A dataset also contains a set of *observations* (also called *records)* over these variables. For example:

tuition = $46,288, *enrollment* = 2,563, *private, etc.*

Data: a definition

A dataset also contains a set of *observations* (also called *records)* over these variables. For example:

tuition = $16,115, *enrollment* = 28,635, *public, etc.*

One way to think about this:

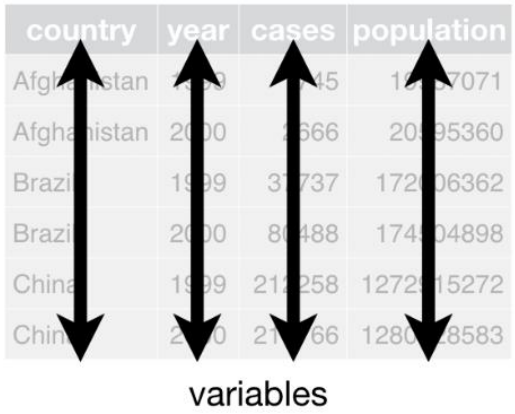
VARIABLES

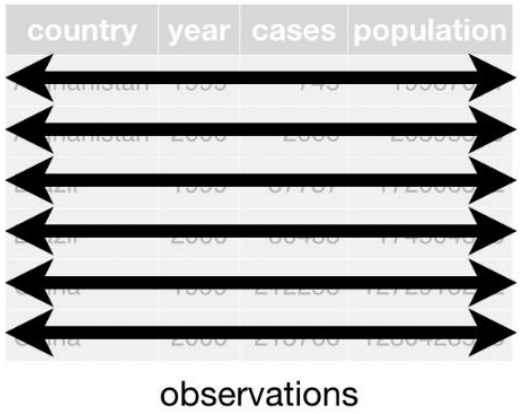
OBSERVATIONS

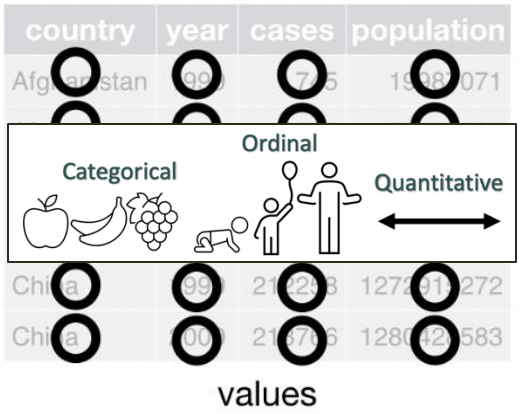
	Tuition	Enrollment	Public vs. Private	…
Smith College	$46,288	2,563	private	
UMass Amherst	$16,115	28,635	public	
Hampshire College	$48,065	1,400	private	
Mount Holyoke College	$43,886	2,189	private	
Amherst College	$50,562	1,792	private	
⋮				

Data

- Remember…

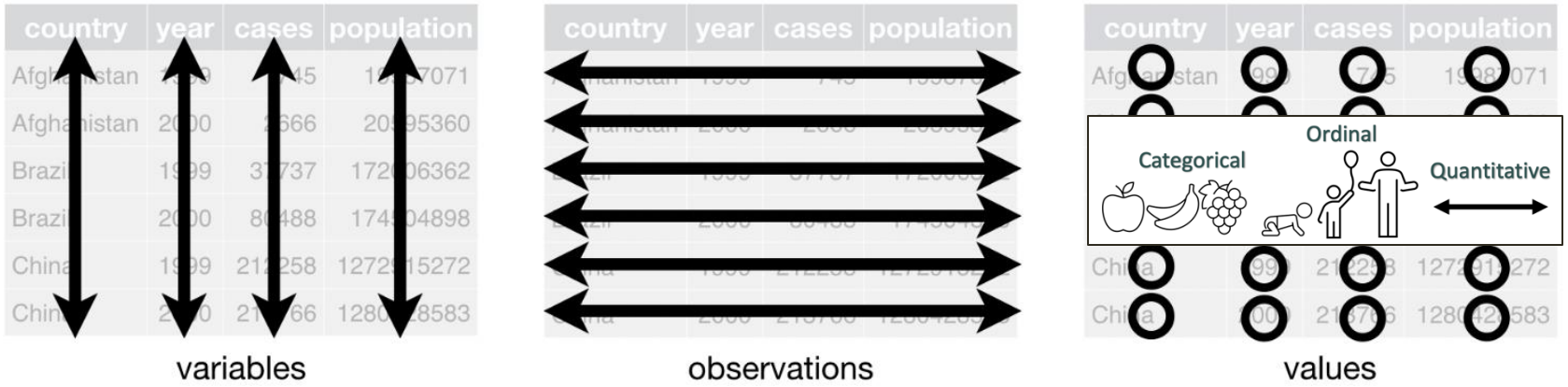






Data → Visuals

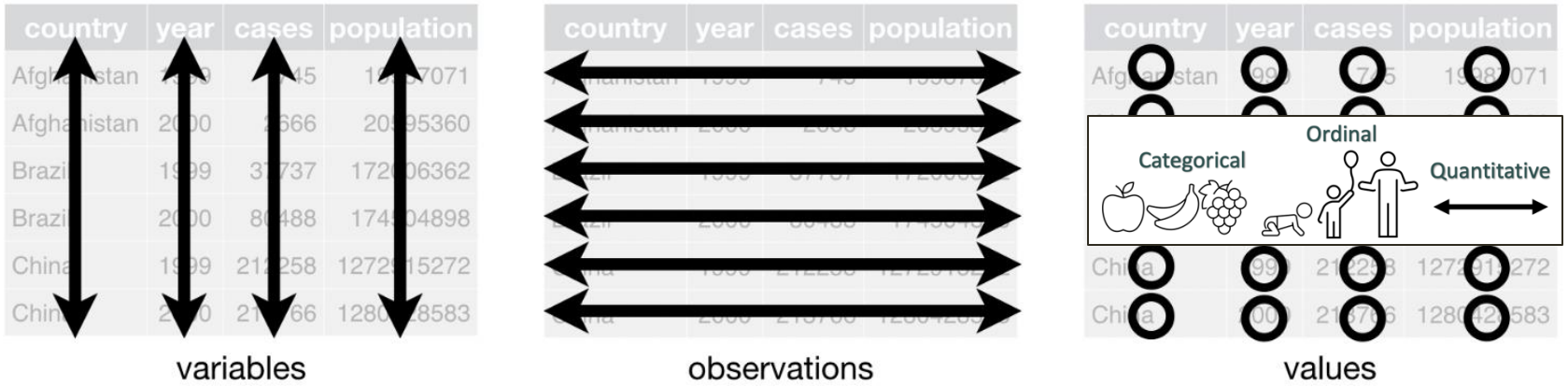
- Remember…



- **Big idea behind visualization**
  - Data have dimensions
  - Visualizations have dimensions, too
  - To build good visualizations, we need to **map data dimensions to visual dimensions** in a principled way

Data → Visuals

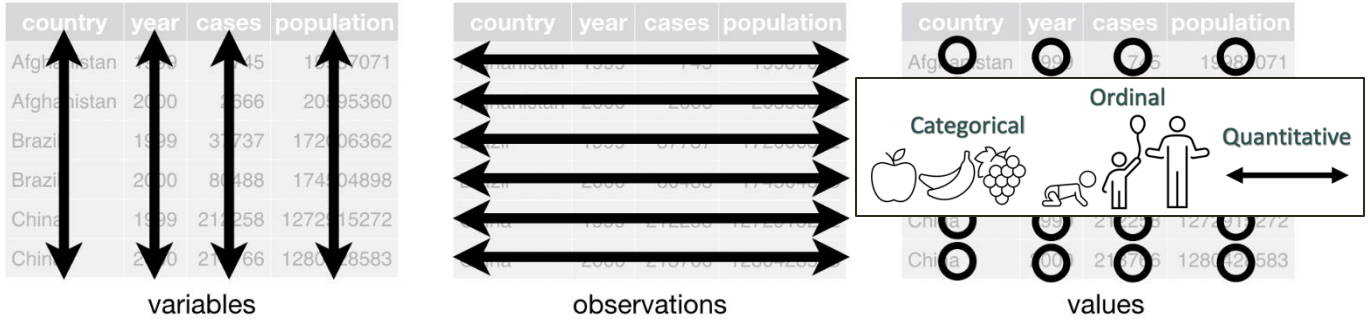
- Remember…



- **Big idea behind visualization**
  - Data have dimensions
  - Visualizations have dimensions, too
  - To build good visualizations, we need to **map data dimensions to visual dimensions** in a principled way

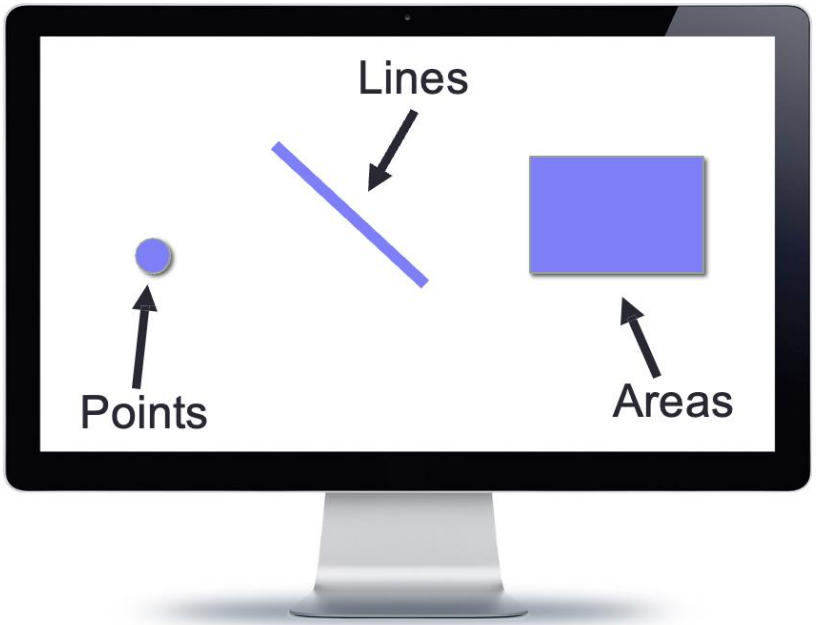
Data → Visuals

Data



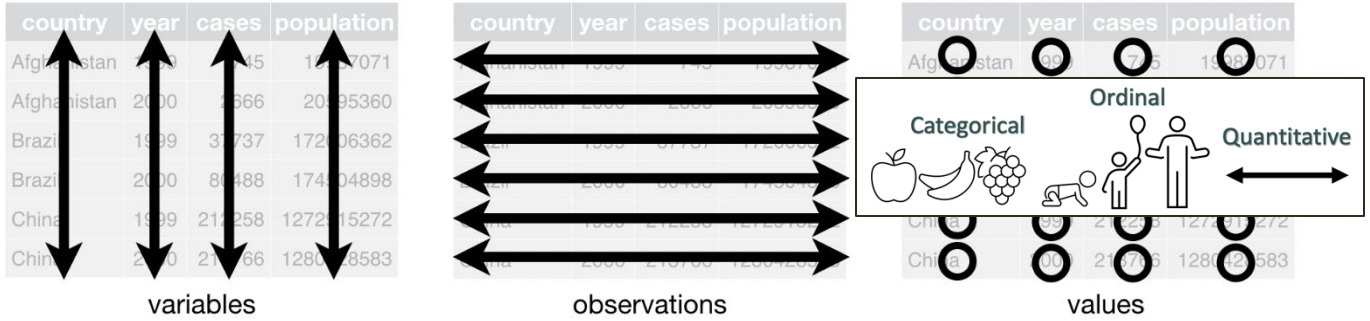
Visuals

- **Marks**
  - The "ink"



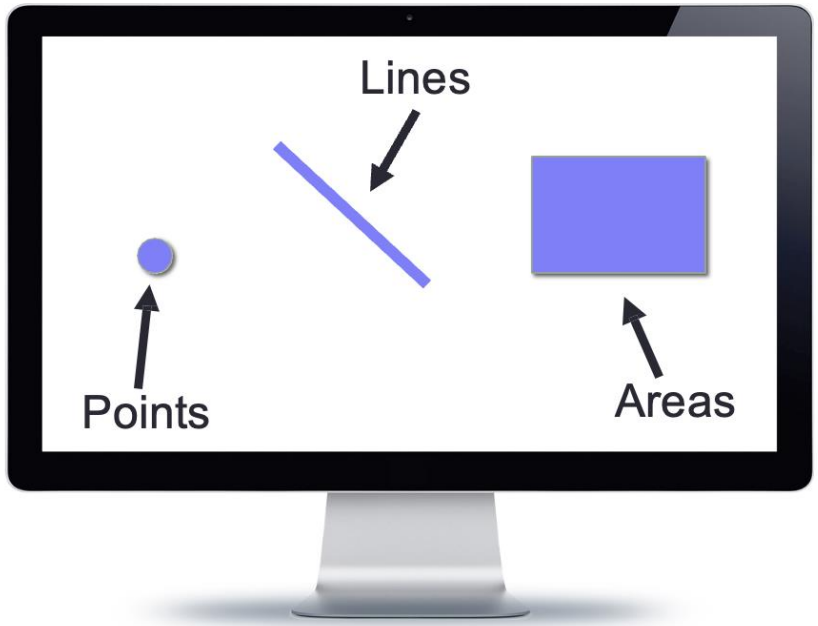
Data → Visuals

Data



Visuals

- **Marks**
  - The "ink"

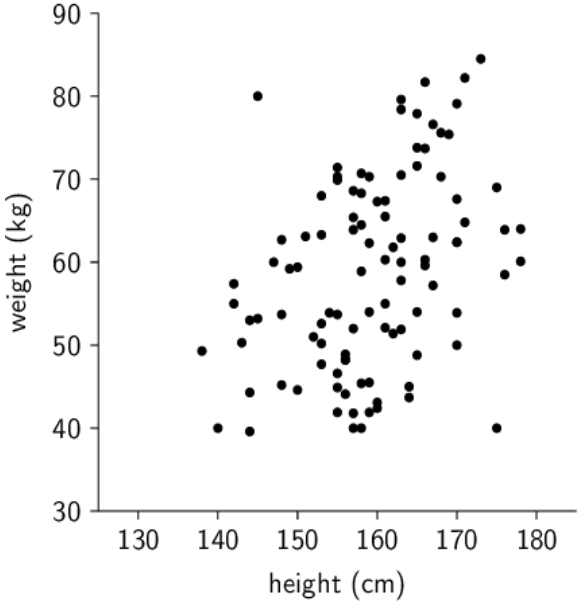


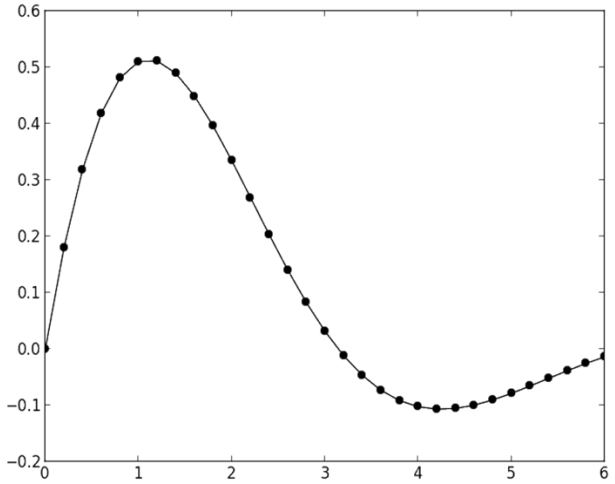
- **Channels or dimensions**
  - *How* the marks show up on the page

Visual Channels / Dimensions

Position

- Encode information using *where* mark is drawn
- Ex.

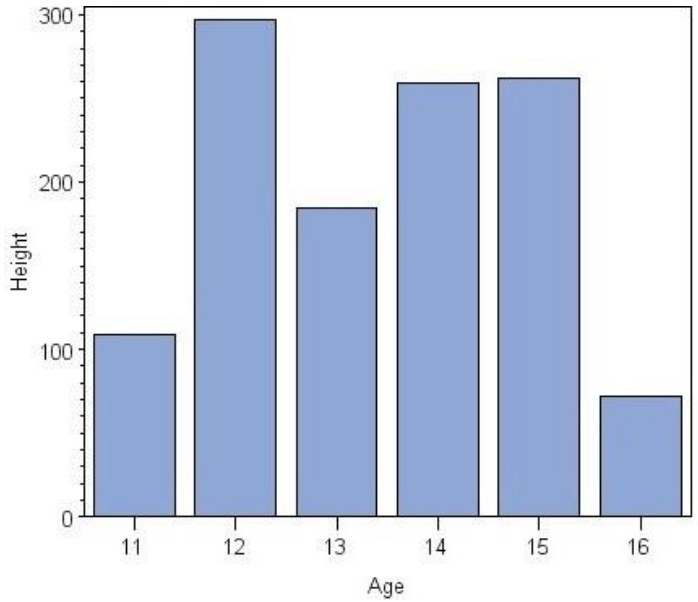


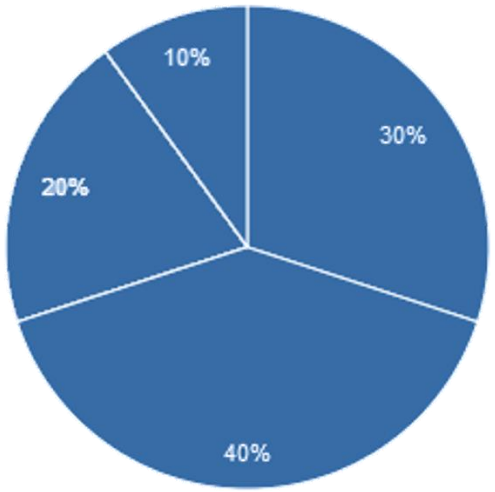


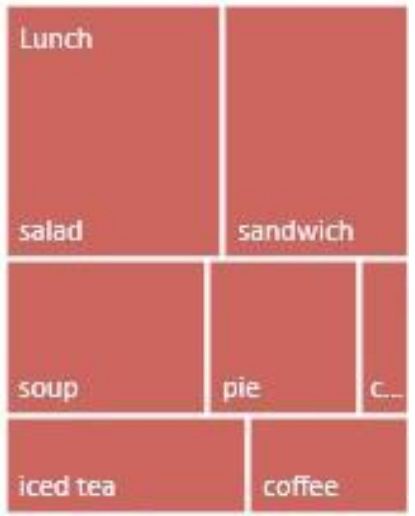
Visual Channels / Dimensions

Size

- Encode information using *how big* mark is drawn
- Ex.



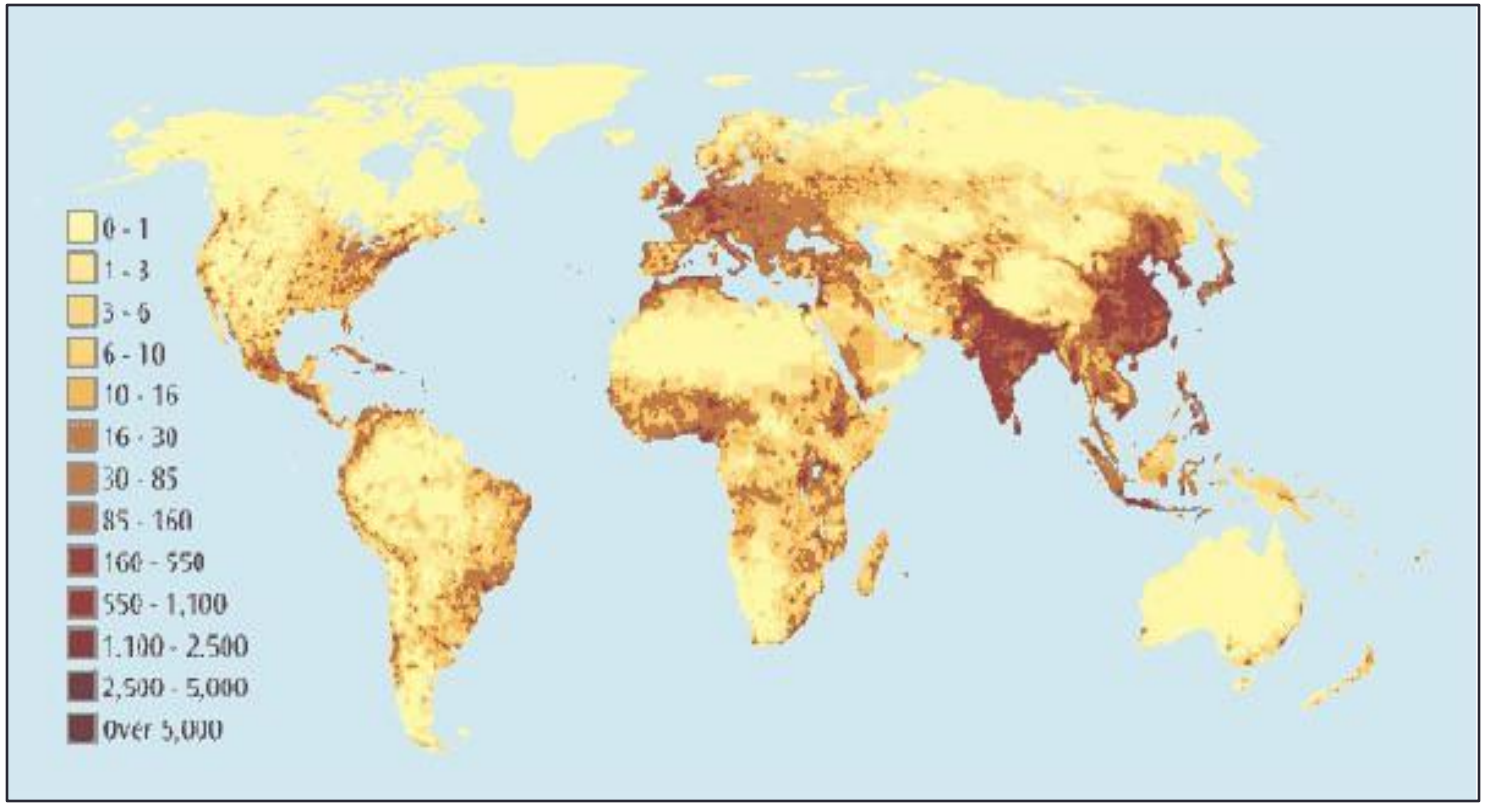




Visual Channels / Dimensions

Value

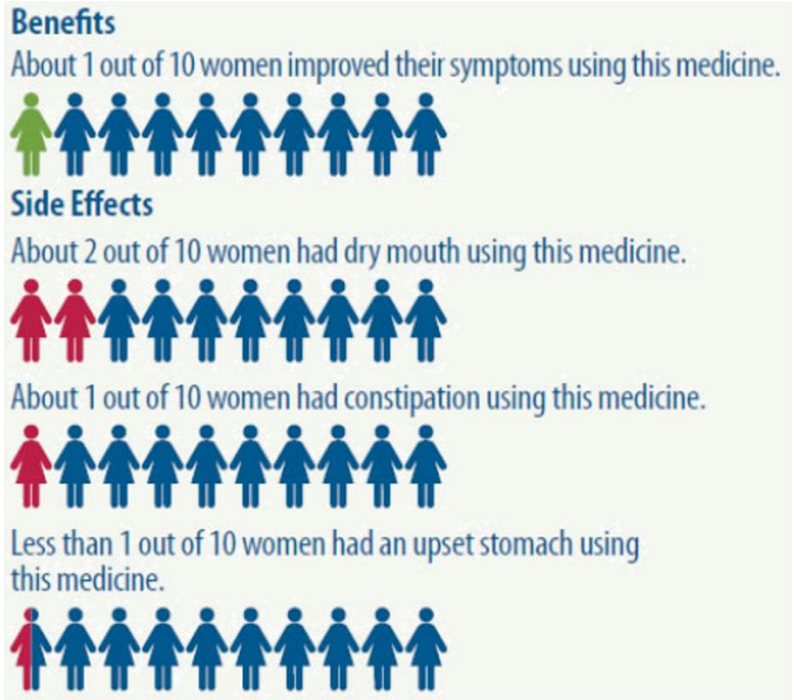
- Encode information using *how dark* mark is drawn
- Ex.

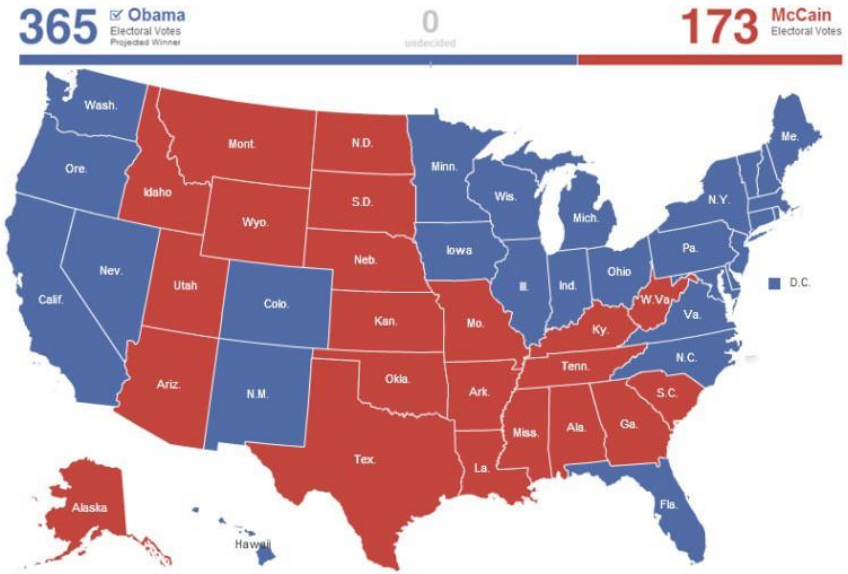


Visual Channels / Dimensions

Color

- Encode information using *hue* of mark
- Ex.

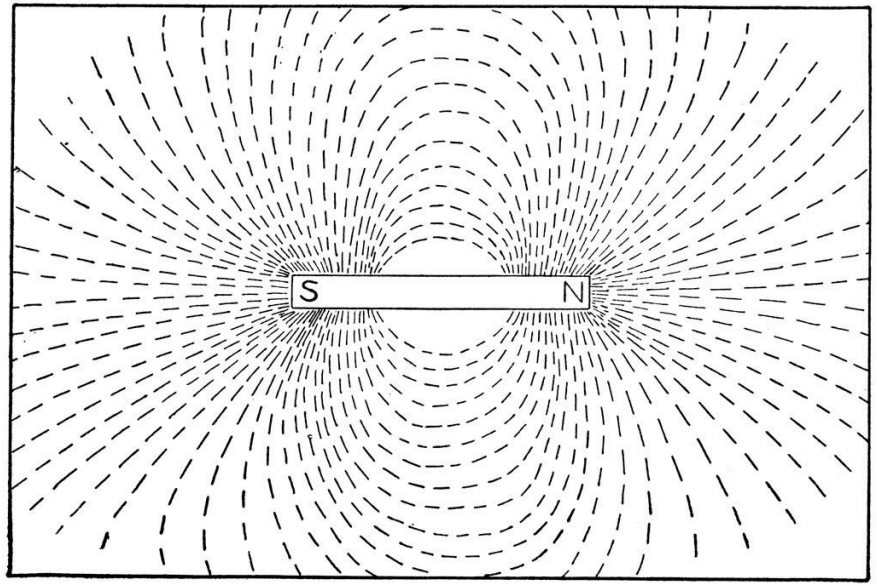




Visual Channels / Dimensions

Orientation

- Encode information using how mark is *rotated*
- Ex.



Visual Channels / Dimensions

Shape

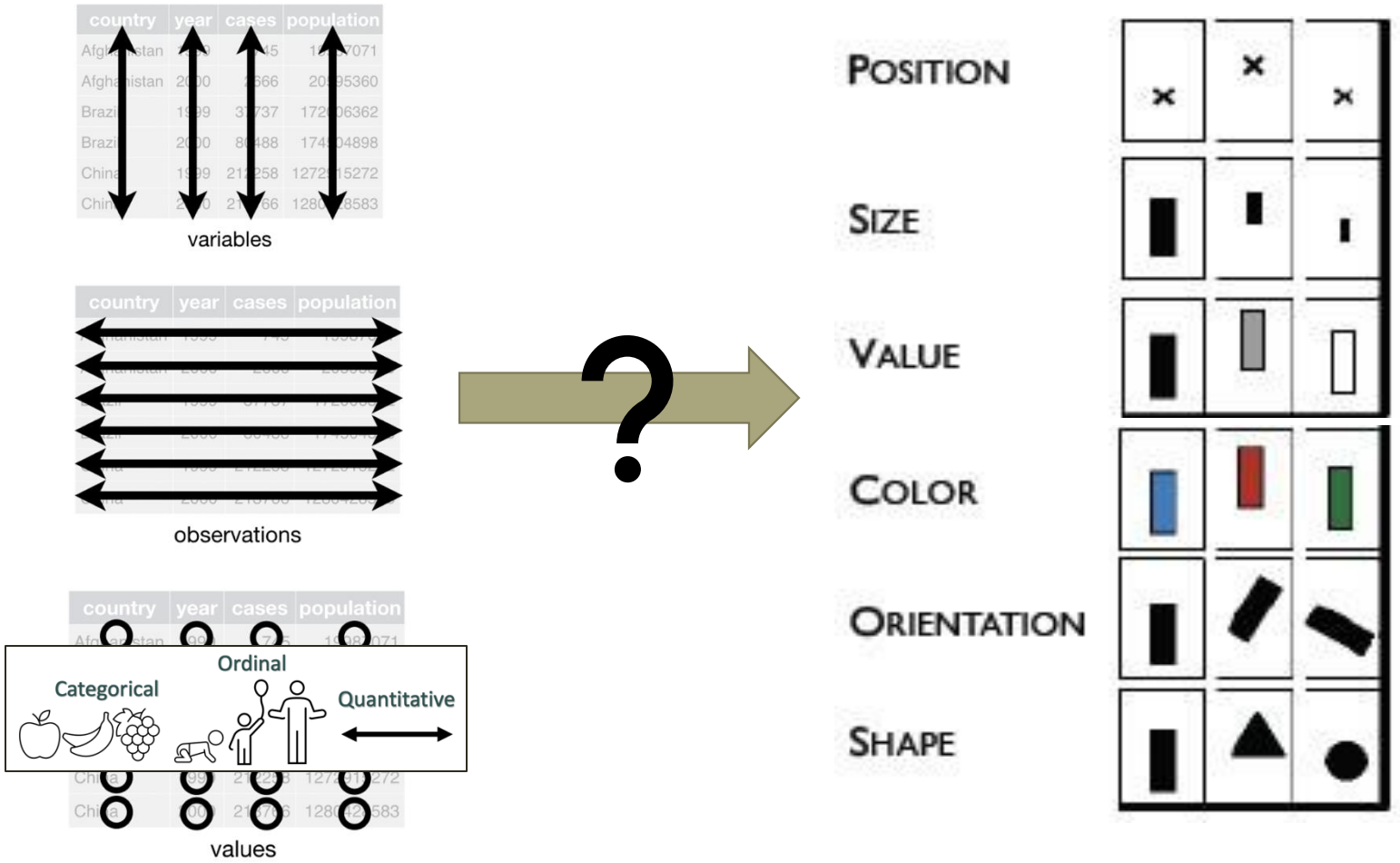
- Encode information using how mark is *shaped*
- Ex.

Data → Visuals

- Remember… **Big idea behind visualization**
  - Map data dimensions to visual dimensions in a principled way

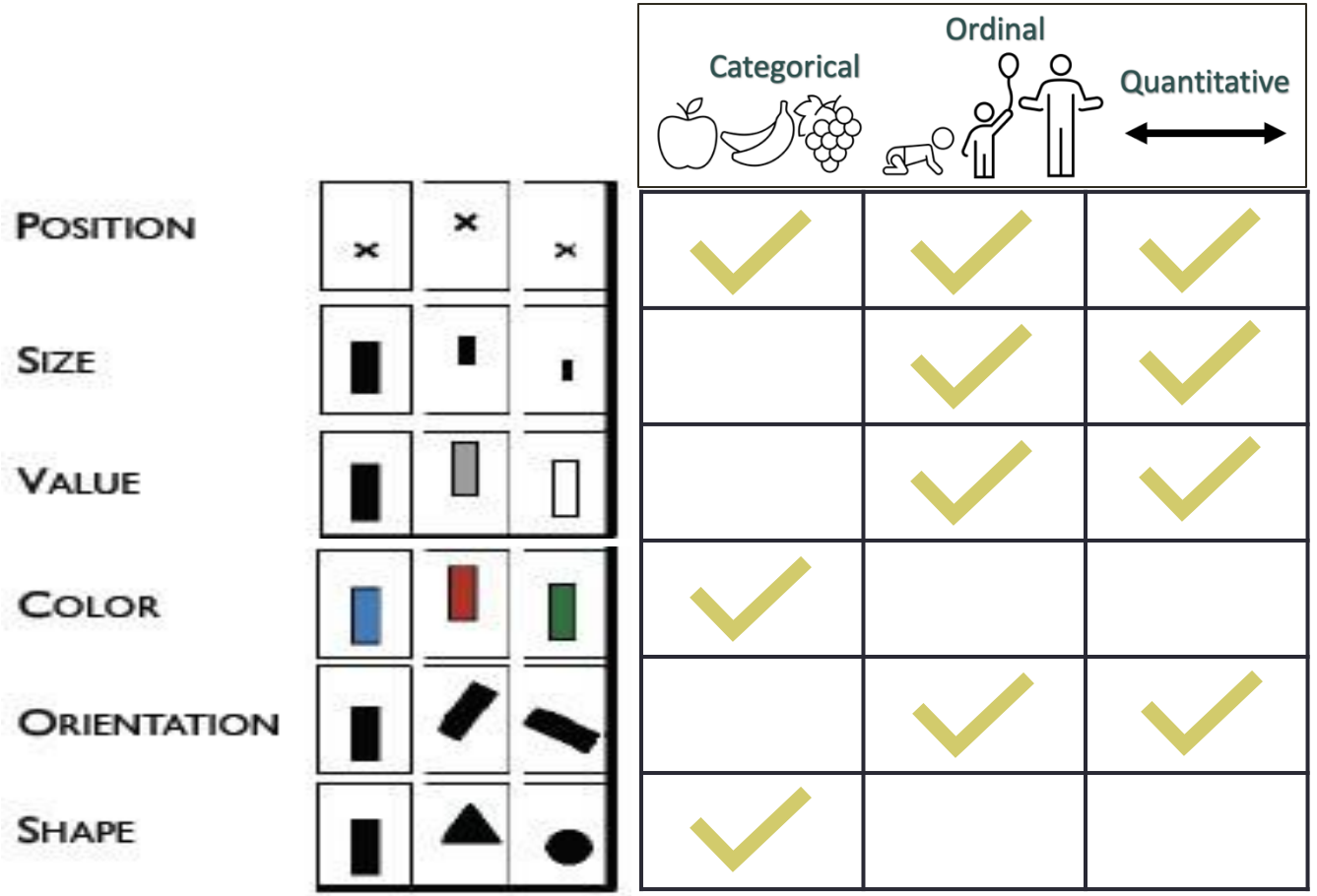
Data → Visuals

- Remember… **Big idea behind visualization**
  - Map data dimensions to visual dimensions in a principled way



Data → Visuals

- Remember… **Big idea behind visualization**
  - Map data dimensions to visual dimensions in a principled way
  - Not all visual dimensions can represent all data types

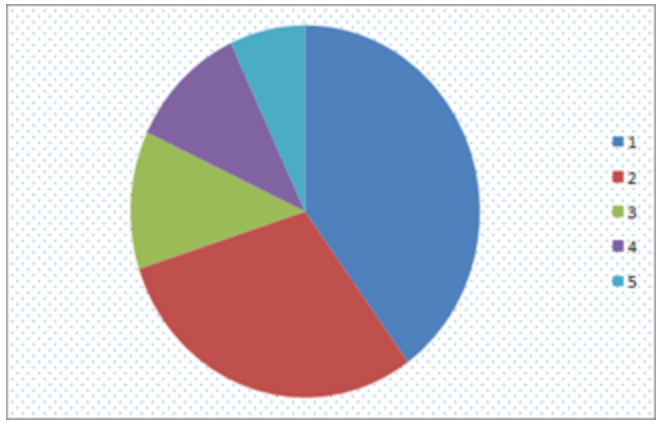


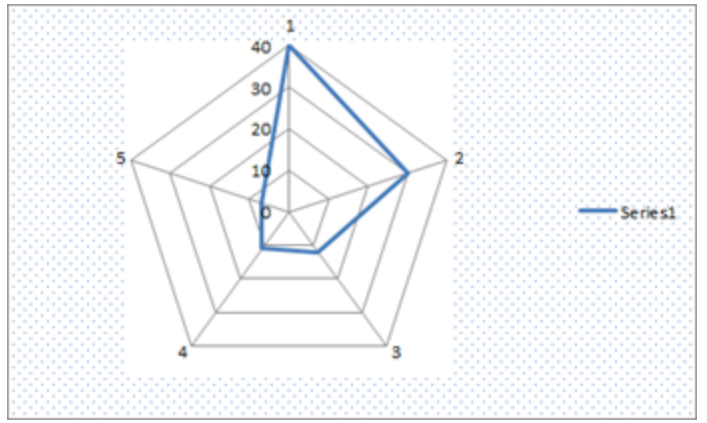
Jacques Bertin, *Semiologie Graphique* (Semiology of Graphics), 1967.

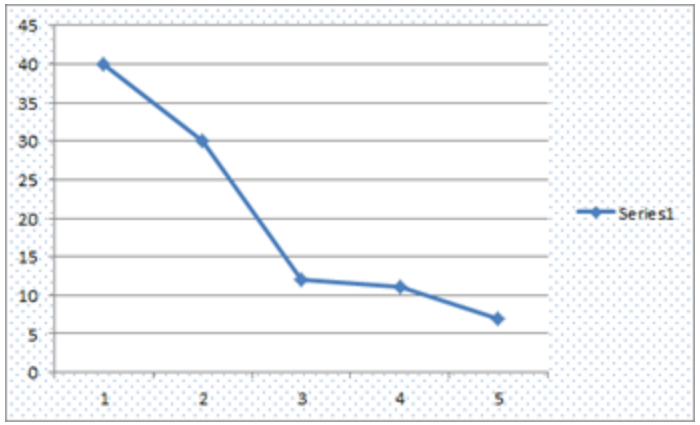
Key question for this course

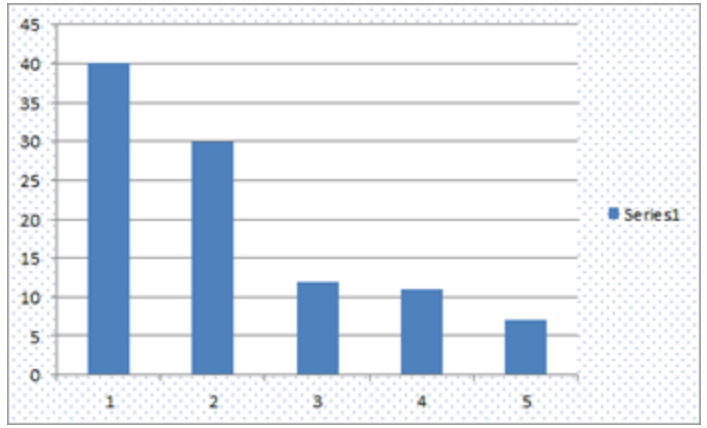
Which **data dimension** should be mapped

to which **visual dimension?**

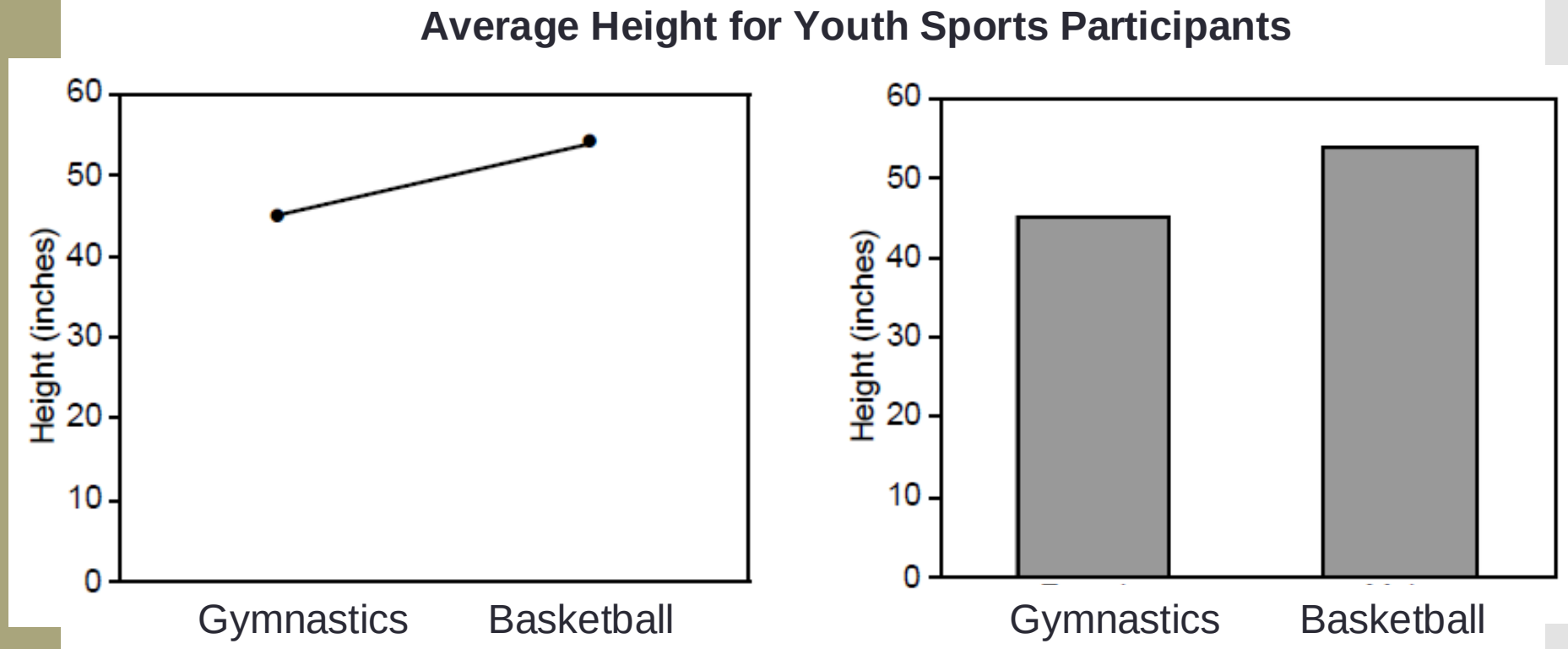








Answer: it depends

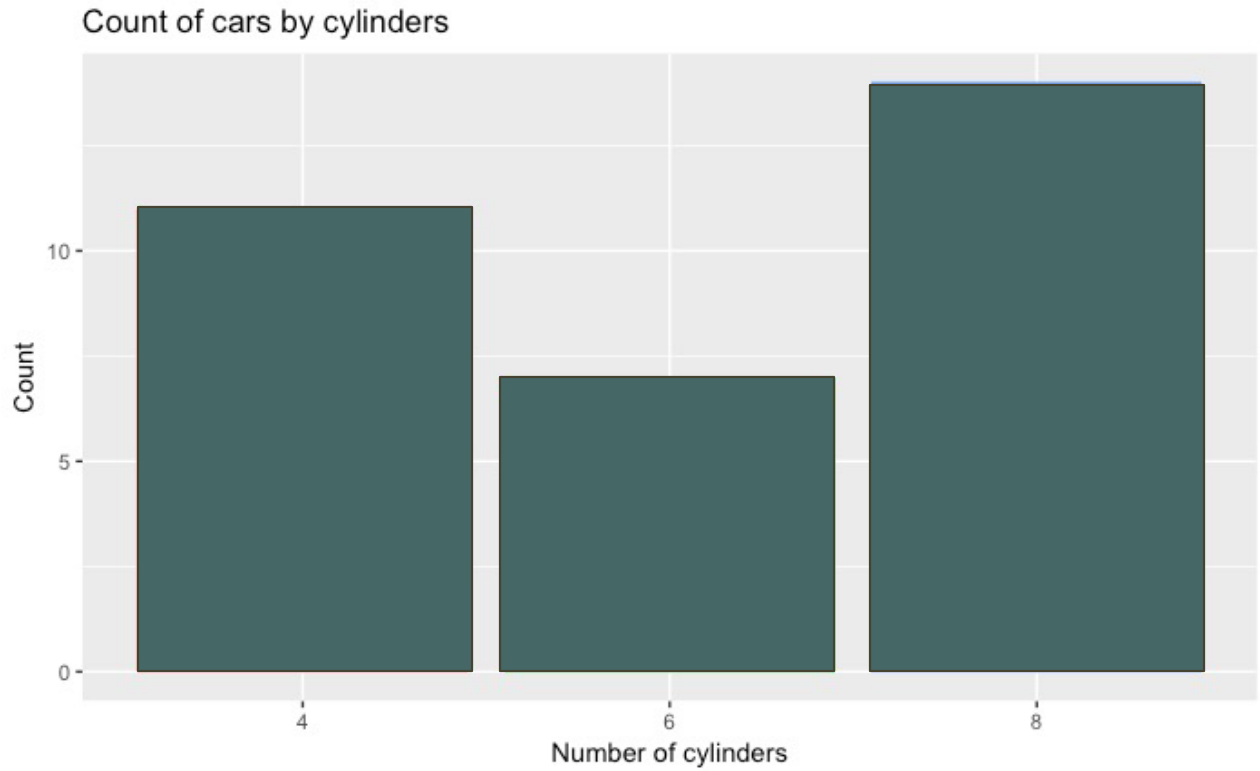


A sampling of visualization techniques

Bar chart

- Used to **ariable**
- **Compa** ies
- Highlig
- Groupe o sub-groups

What is the data visual mapping in this chart?



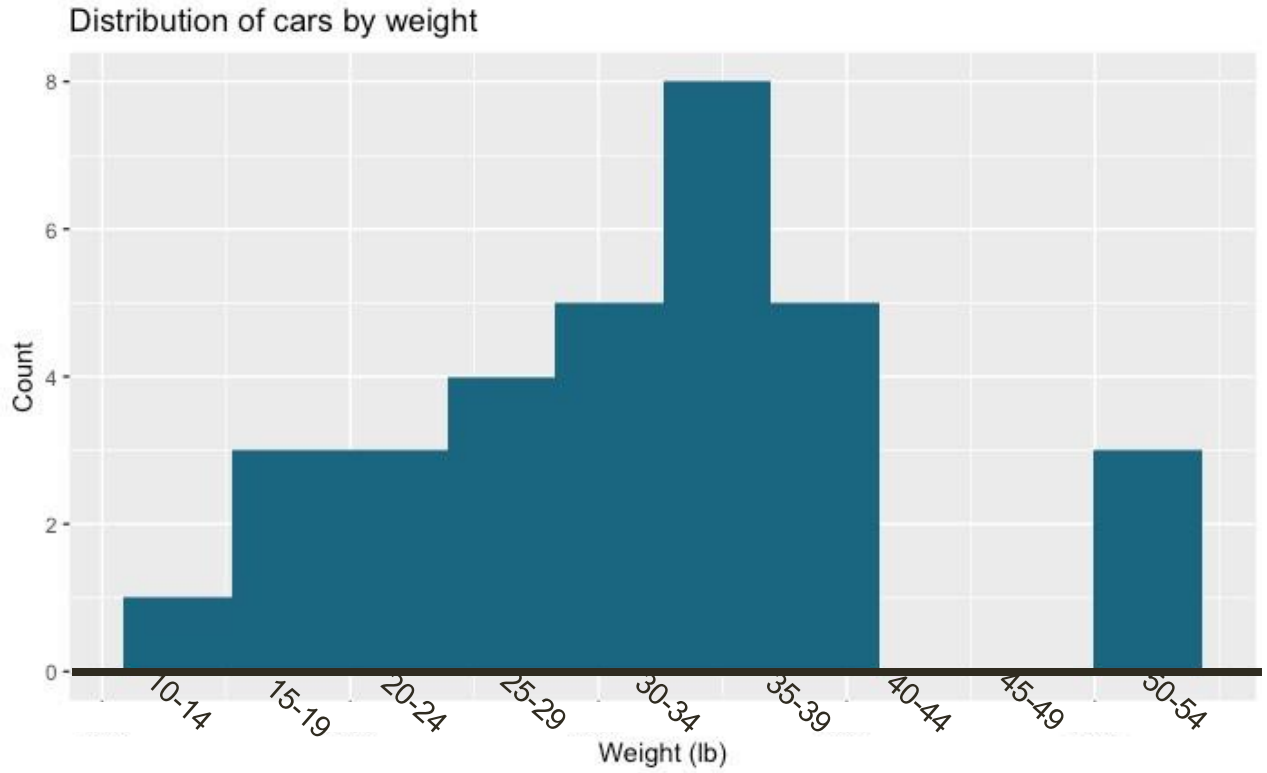
Histogram

- Used f
- Looks
- Y-axis
- Highlights **distribution**
- Note: bin size makes a big difference!

qual

size ra

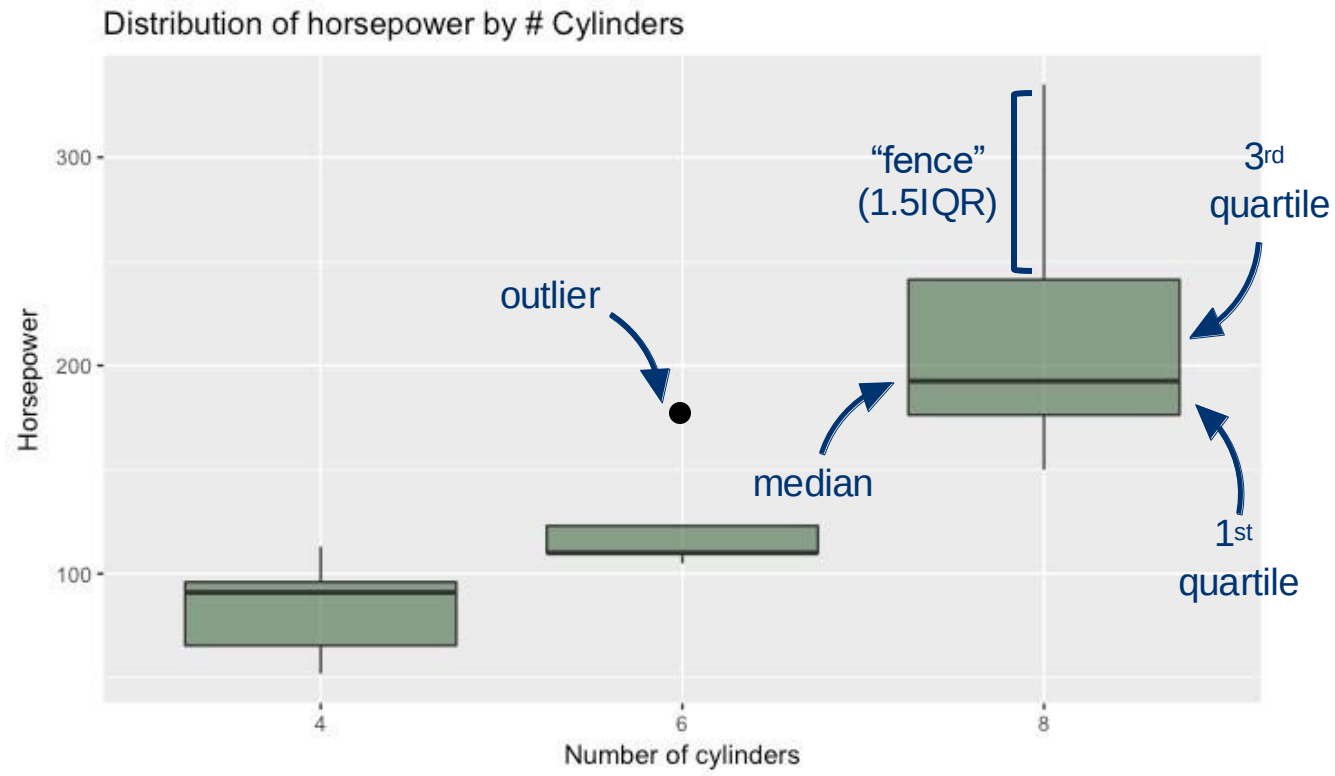
What is the data visual mapping in this chart?



Boxplot

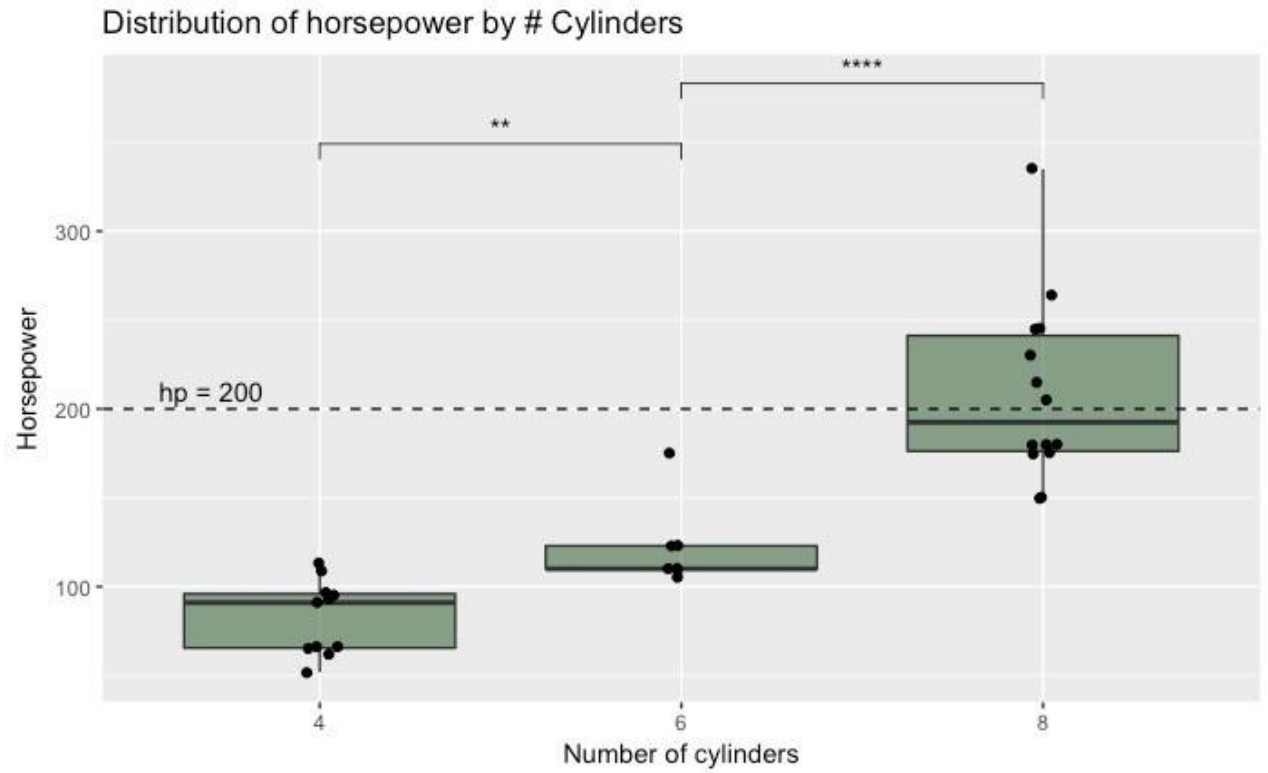
- Used f
- Also u
- Calls o
  - mec
  - 1st &
  - "fences"
  - outliers

What is the data visual mapping in this chart?



Boxplot

- Use "jitter" to show actual values
- Reference lines can help provide context
- Can use annotations to show statistical significance

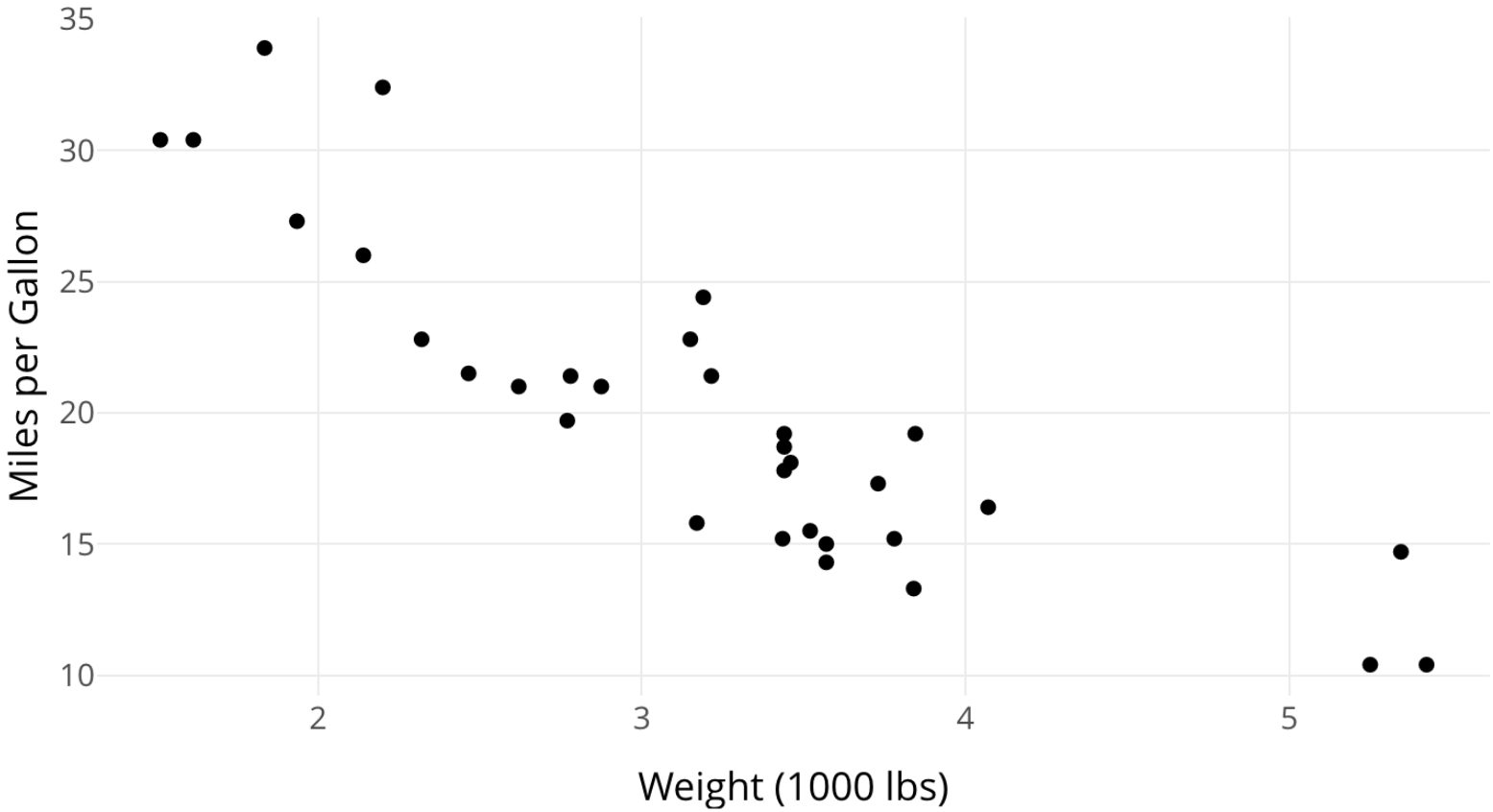


Scatterplot

- Used to
- Shows
- Each p

antitative variabl

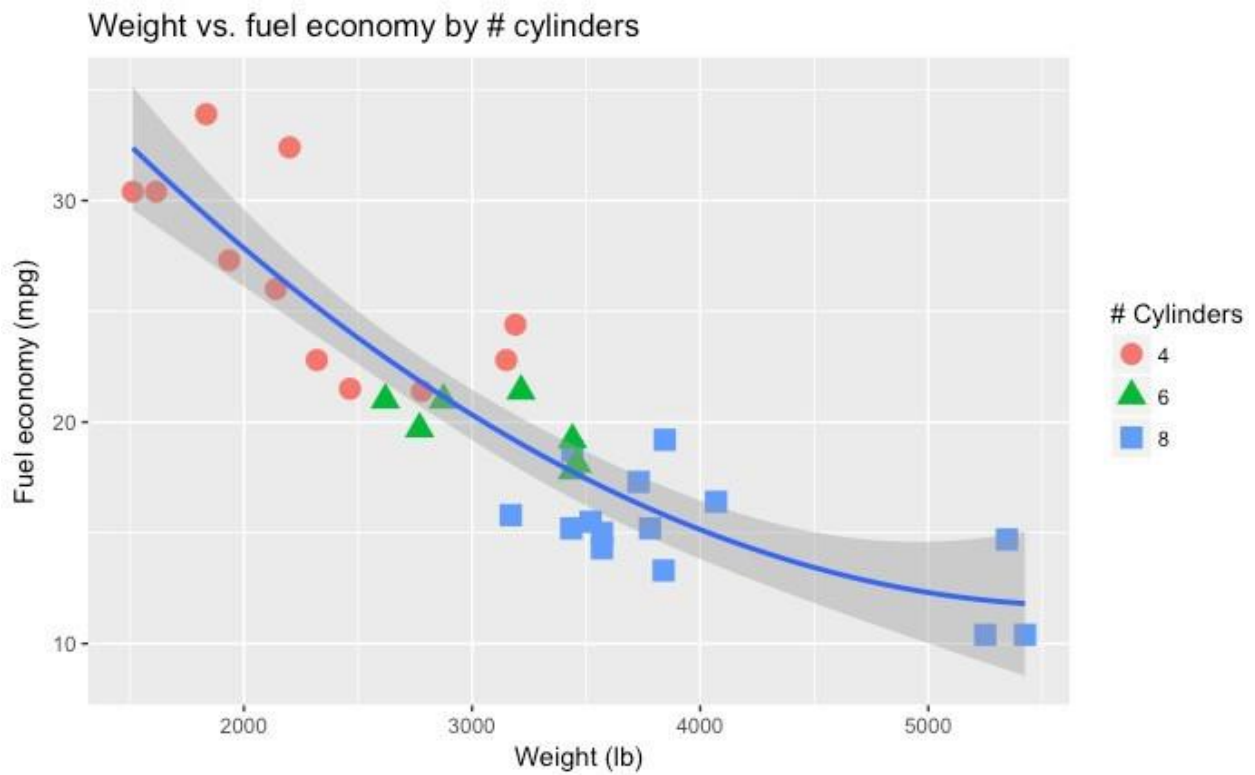
What is the data visual mapping in this chart?



Scatterplot

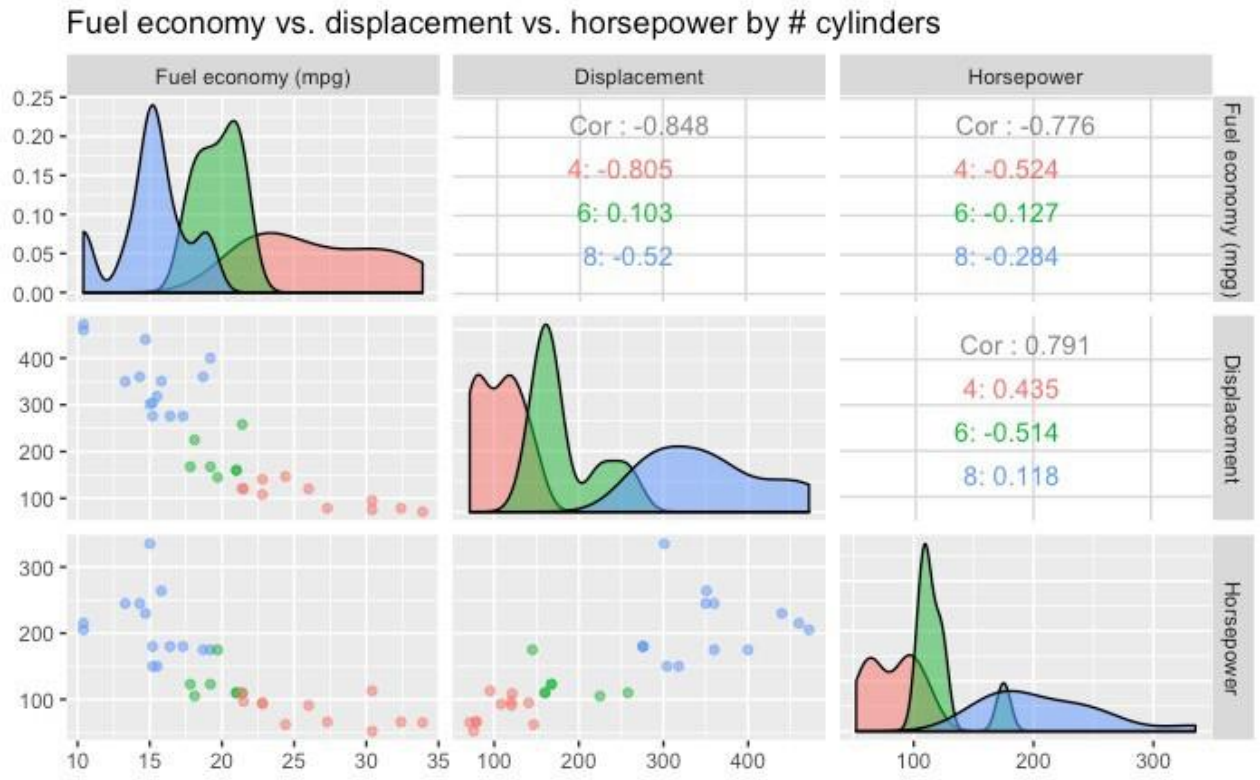
- Can use **variable X one quanti**
- This hig
- Someti

What is the data visual mapping in this chart?



Scatterplot matrix (SPLOM)

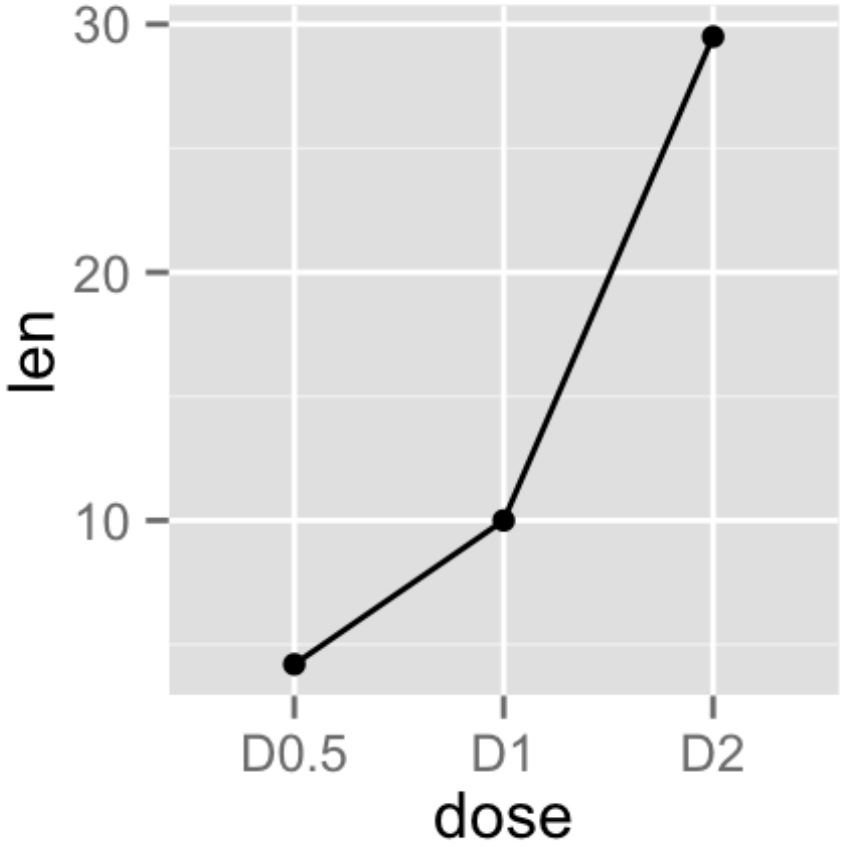
- Can use to show **many combinations of one quantitative variable X one quantitative variable**
- Combines multiple scatterplots into a matrix to show **additional relationships**



Line chart

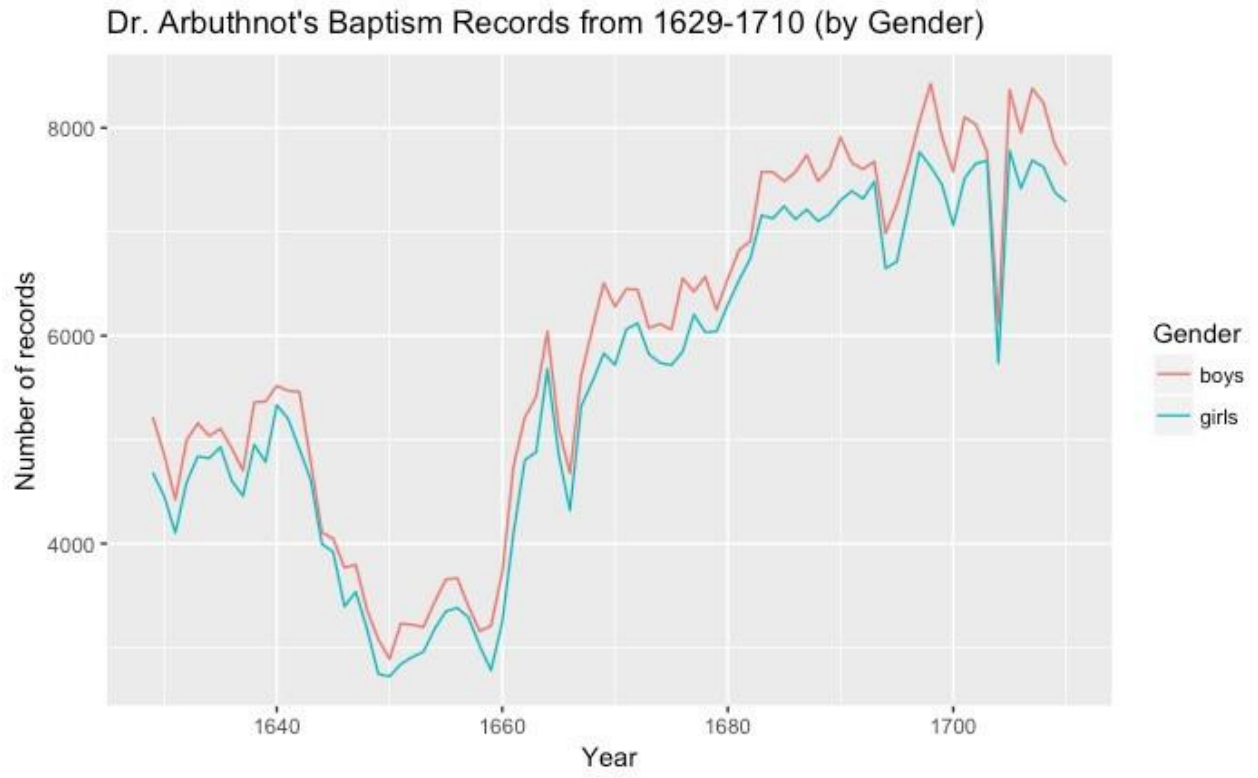
- Shows t **time** **over**

What is the data visual mapping in this chart?

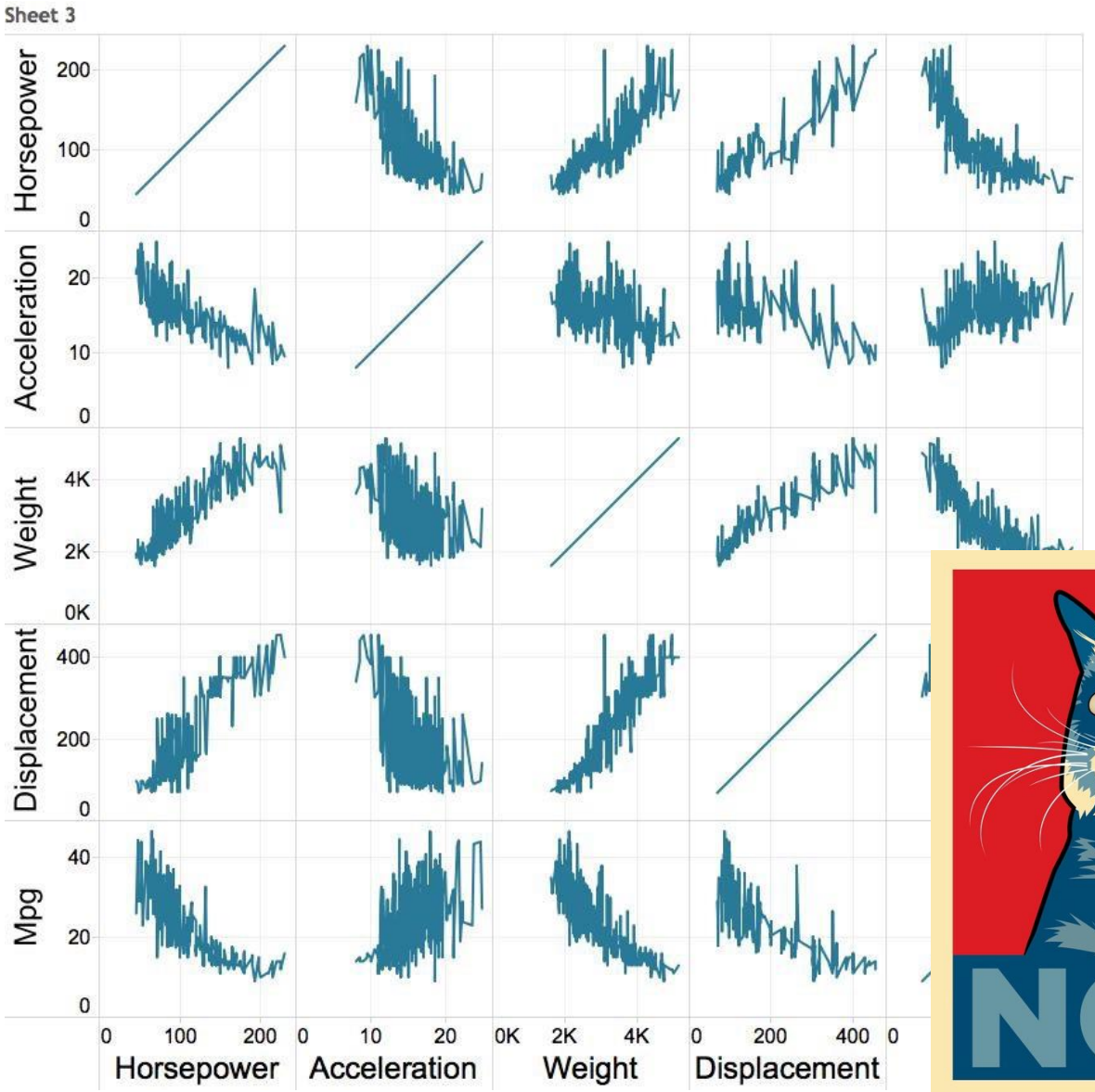


Line chart

- Multiple lines allow **comparison of trends**
- Can show one quantitative variable across groups, or multiple quantitative variables (if they have the same scale)
- Highlights "position switches"

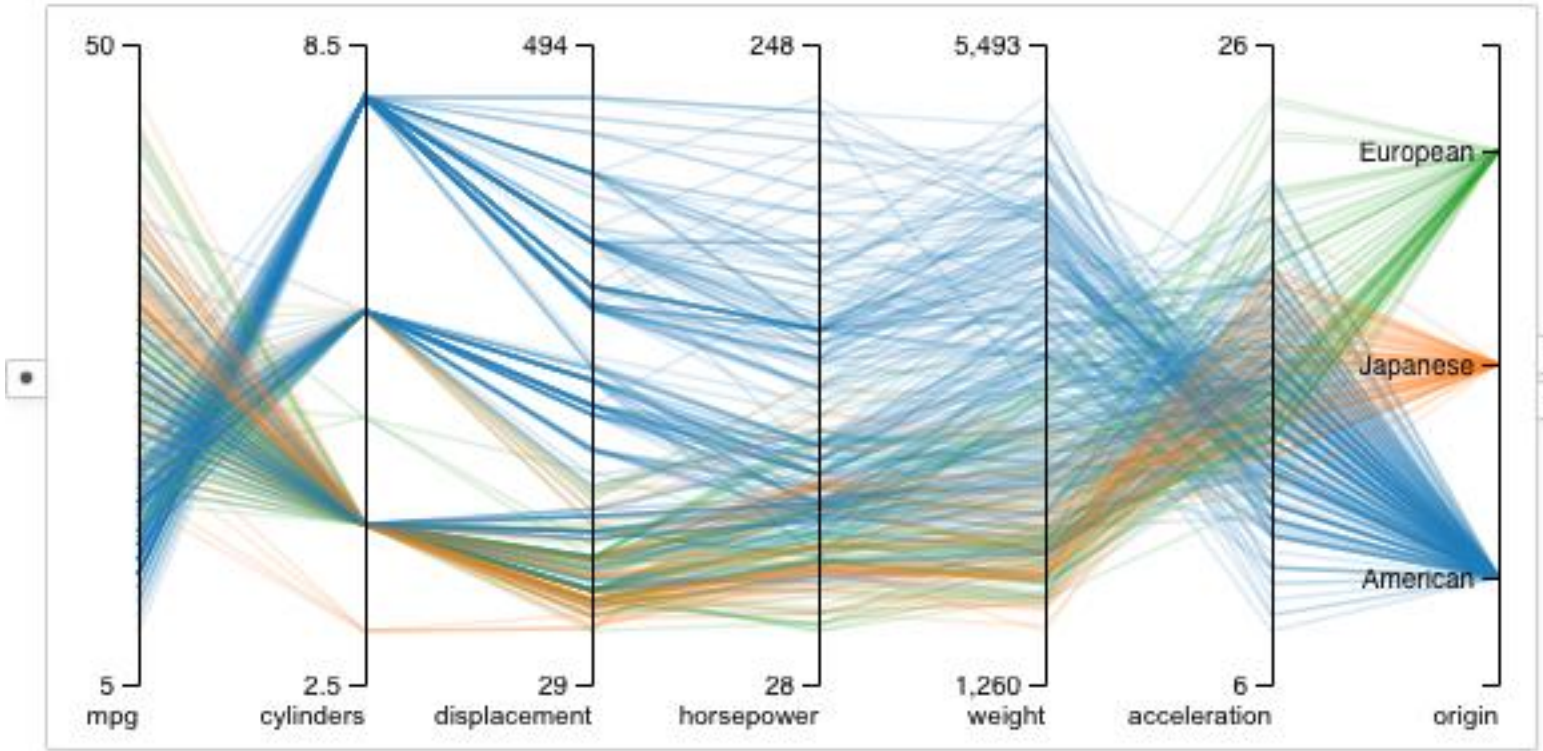


Multiple variables: line chart matrix?



Parallel Coordinates Plot

- Supports (pairwise) **comparison of a collection of quantitative variables**
- Each axis represents one variable
  - They may have different scales, typically you normalize them
- Each line represents one observation (connecting the associated values along each axis)
- Axis order matters!



https://visflow.org/node/visualization/parallel-coordinates.html

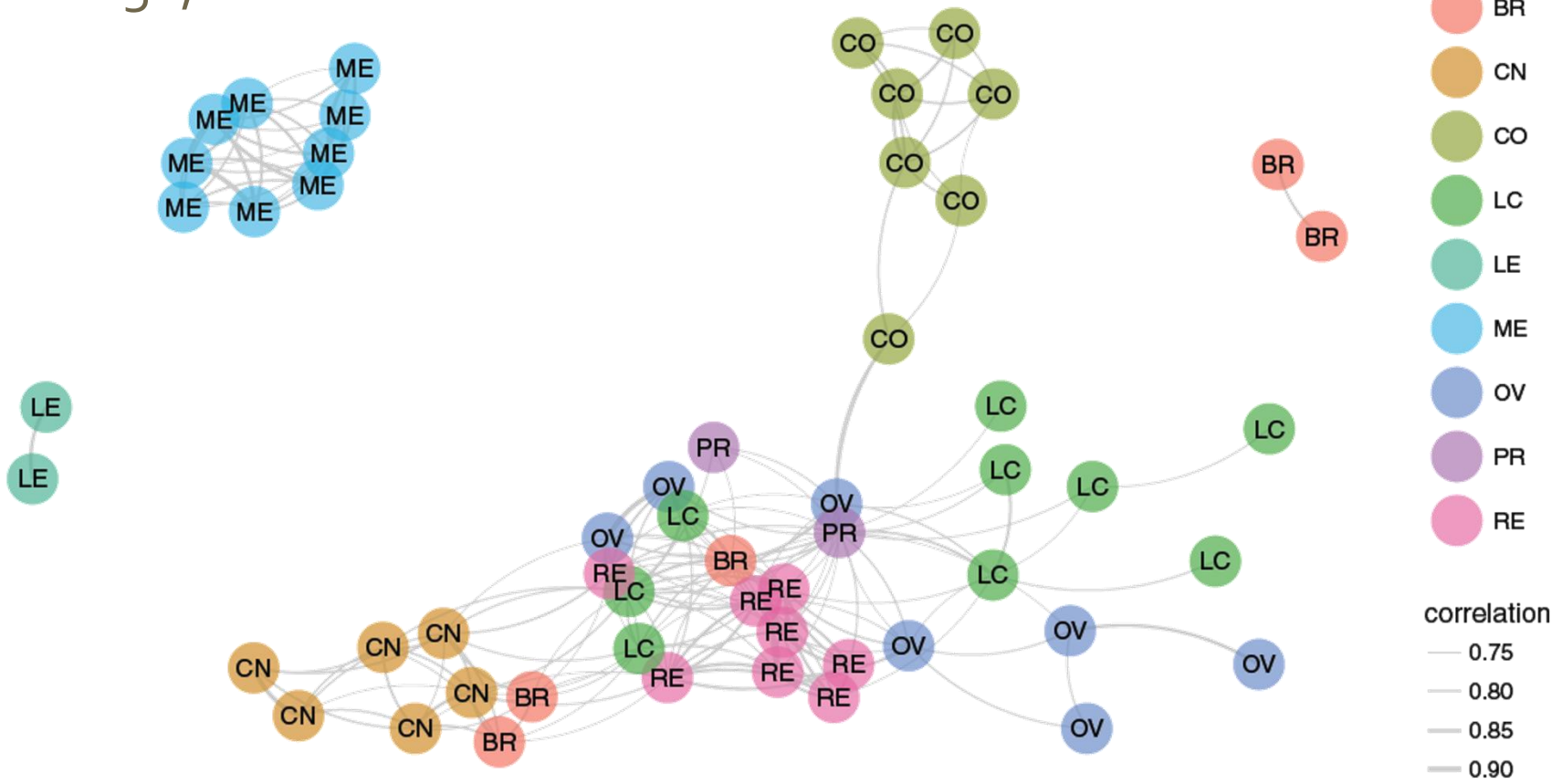
ta-to-
parallel.html

Network

- Shows
- Useful
- Can us
- Caveat

ities
nation
s for large, dense data

What is the data visual mapping in this chart?



Map

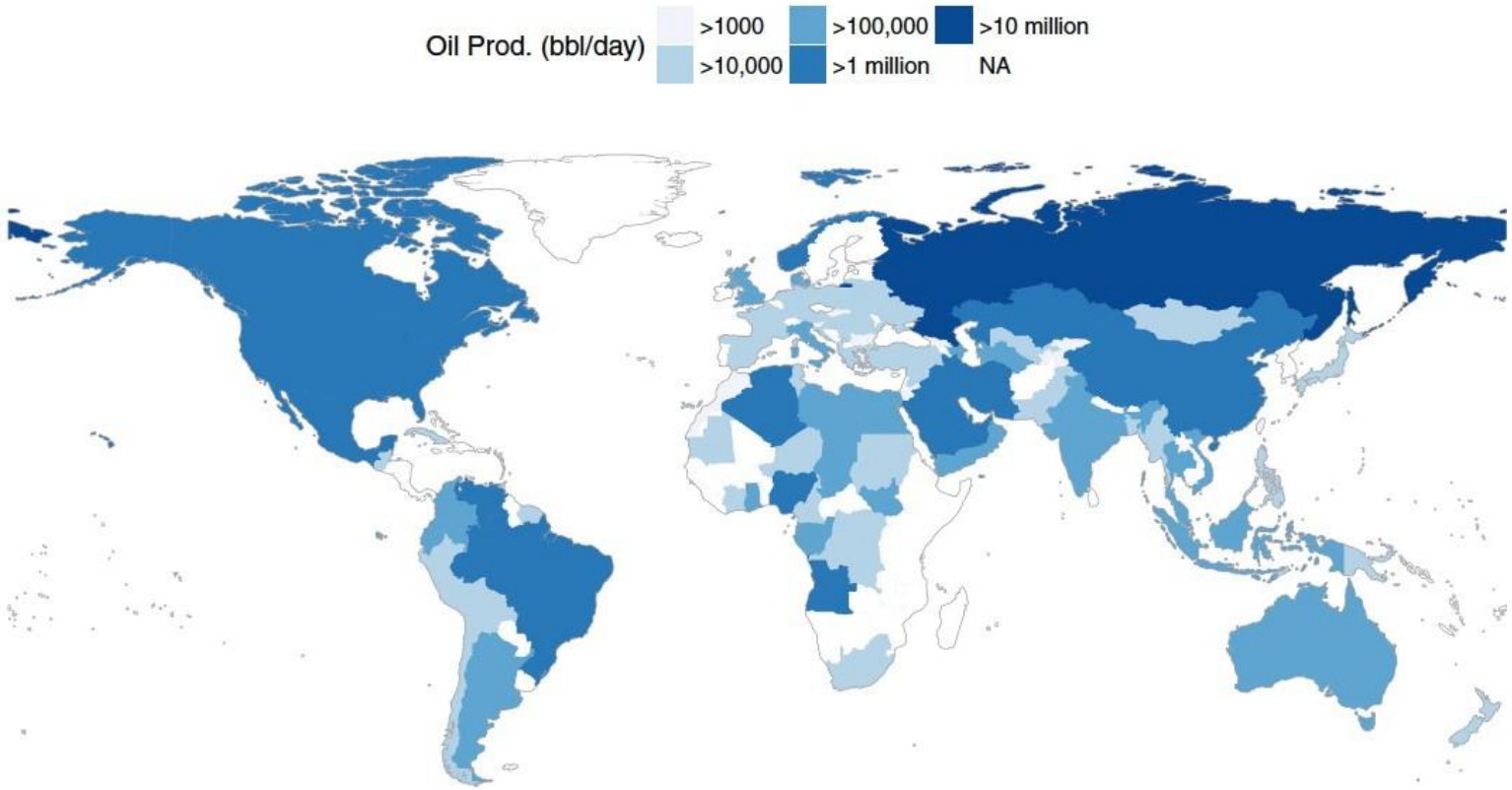
- Shows
- Useful
- Filled r
- Remen

onent

humans

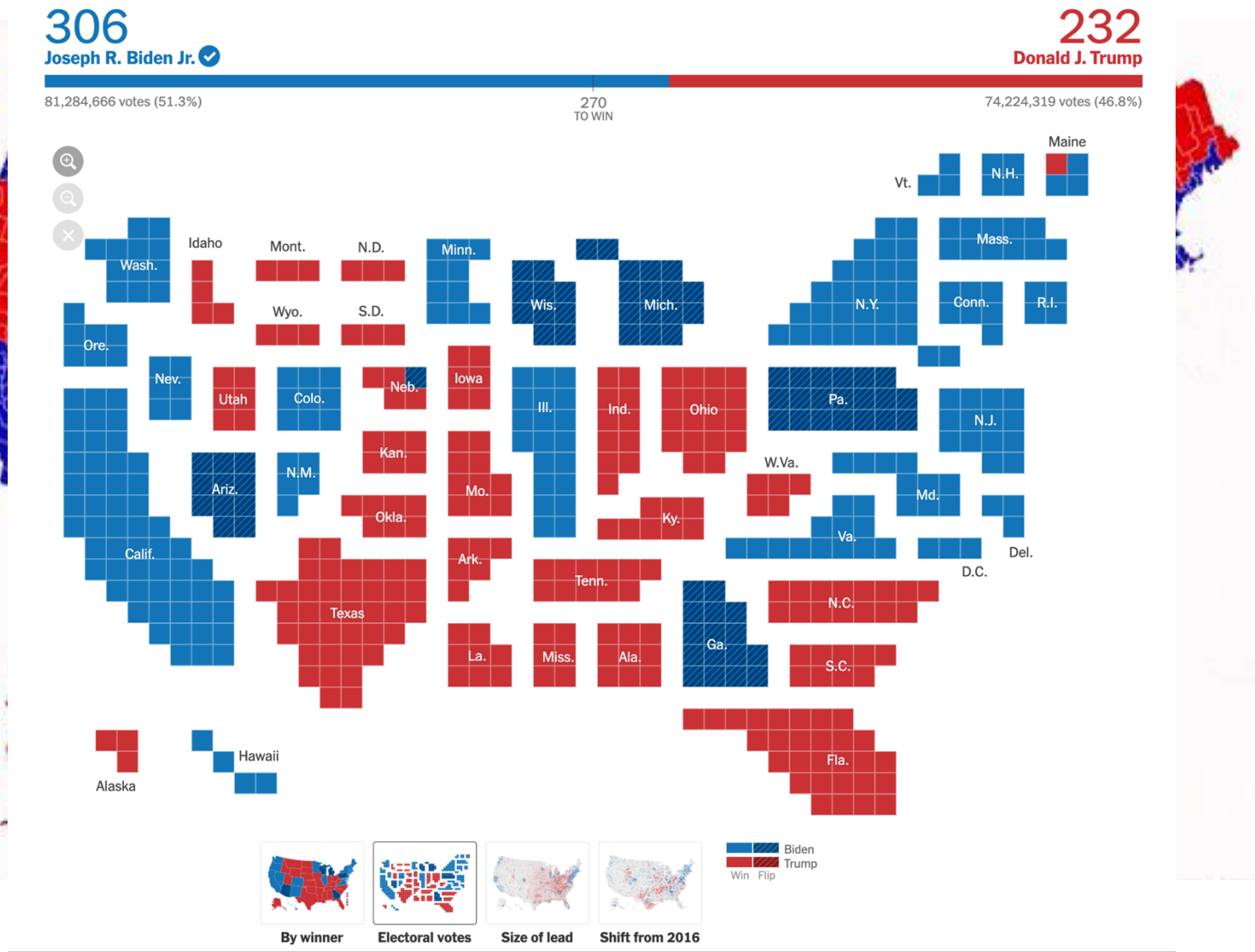
than size comparisons.

What is the data visual mapping in this chart?



Map

- Remember to map the correct data to your visual channels



Your turn

- Work with 1-2 classmates
- Download `nyc_trees.csv` from the course website.
  - Data is a subset of the dataset available here: https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh/about_data
- Using R or Python (you pick!) and `nyc_trees.csv` generate a:
  - Bar chart
  - Histogram
  - Scatterplot
  - If you have time, try a Boxplot and Linechart

- R Plotting Resources:
  https://r4ds.hadley.nz/layers
  https://r-graphics.org/
  https://r-graph-gallery.com/ggplot2-package.html
- Python Plotting Resources:
  https://matplotlib.org/
  https://plot.ly/python/
  https://seaborn.pydata.org/tutorial.html