

MassMutual DSDP 2018:

# INTRODUCTION TO DATA VISUALIZATION

---

June 15, 2018

R. Jordan Crouser

Assistant Professor of Computer Science

Smith College

# Introductions & background

Jordan

(he / him)



- **2017 on: Asst. Prof. in CS (Smith)**
- 2015 to 2017: Visiting Asst. Prof. in SDS (Smith)
- 2013 – 2015: Research Scientist (MITLL)
- 2010 – 2013: PhD in Visual Analytics (Tufts)
- 2008 – 2010: MSc in Educational Tech. (Tufts)
- 2004 – 2008: BA in CS and Math (Smith)

**For more info, visit:**

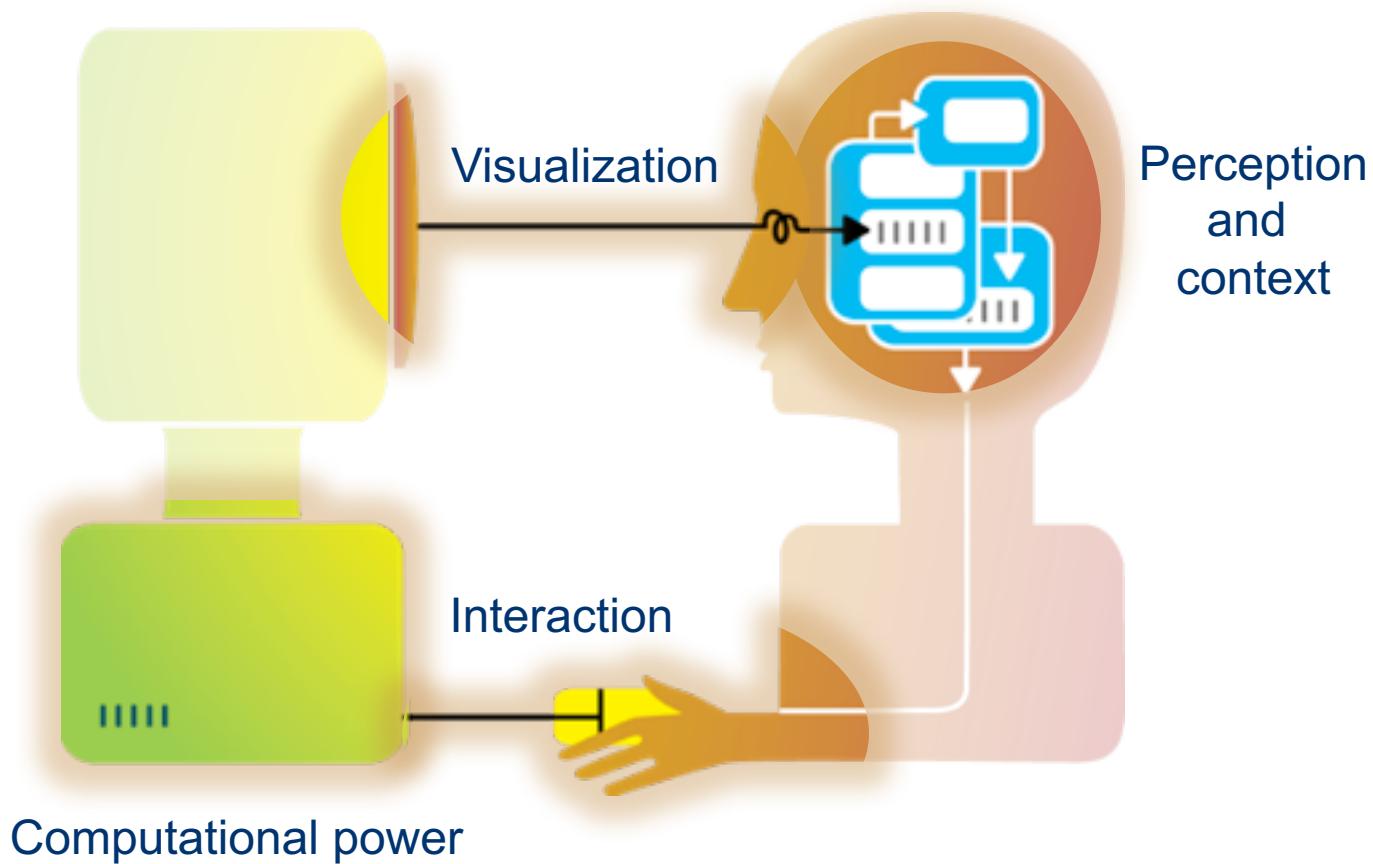
[www.science.smith.edu/~jcrouser](http://www.science.smith.edu/~jcrouser)

# What I do: analytical tools for messy data



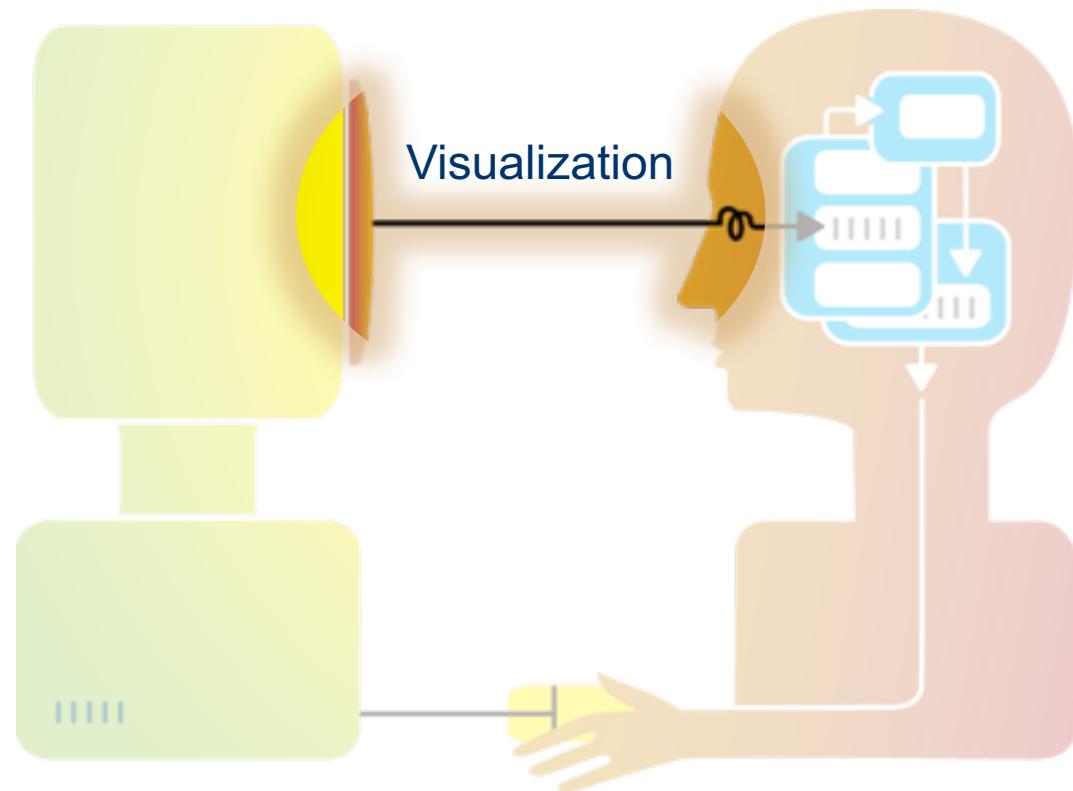
# Big idea behind my research

Humans and machines have **complimentary strengths**



# Focus of this workshop

How do we build **visualizations** that help humans understand patterns in data?



# Interesting (non-CS) thing to know about me



I'm a big  
board game  
nerd  
(we have a closet  
in our house  
dedicated to them)

# Way more interesting thing

Logan  
(10 months old)



“...I might come to visit  
later this summer,  
I have lots of thoughts  
on **data vis!**”

# 3-minute biographies

## About you:

- Your name and pronouns
- Your alma mater
- Your major / area of focus

## 3 questions:

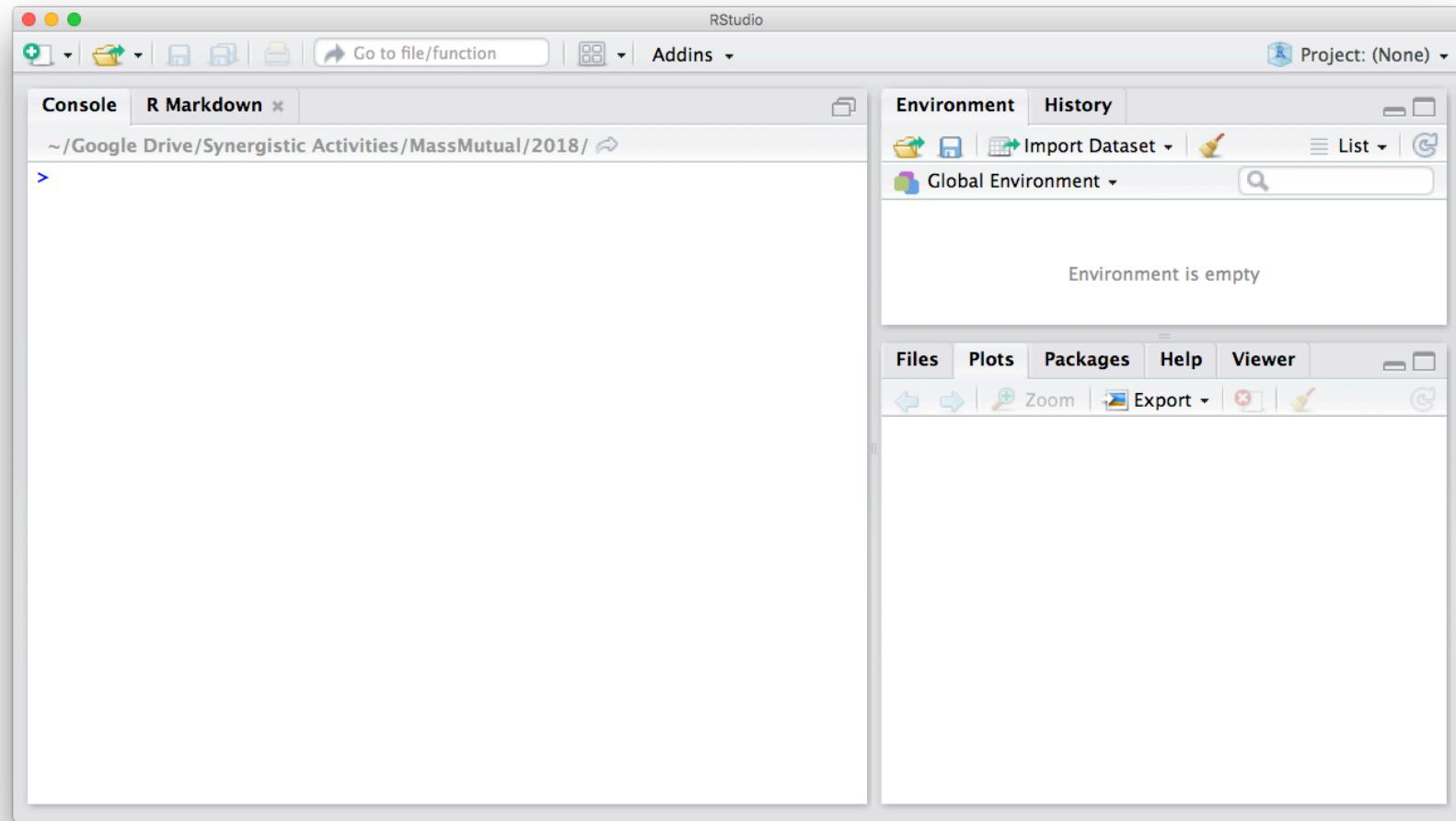
1. What brought you to **DSDP**?
2. What's one **big thing** you hope to get out of it?
3. What's one thing about life after graduation that you find particularly **challenging / anxiety-provoking**?  
-OR-  
3. What's one thing about you that would probably **surprise** us?

# Outline

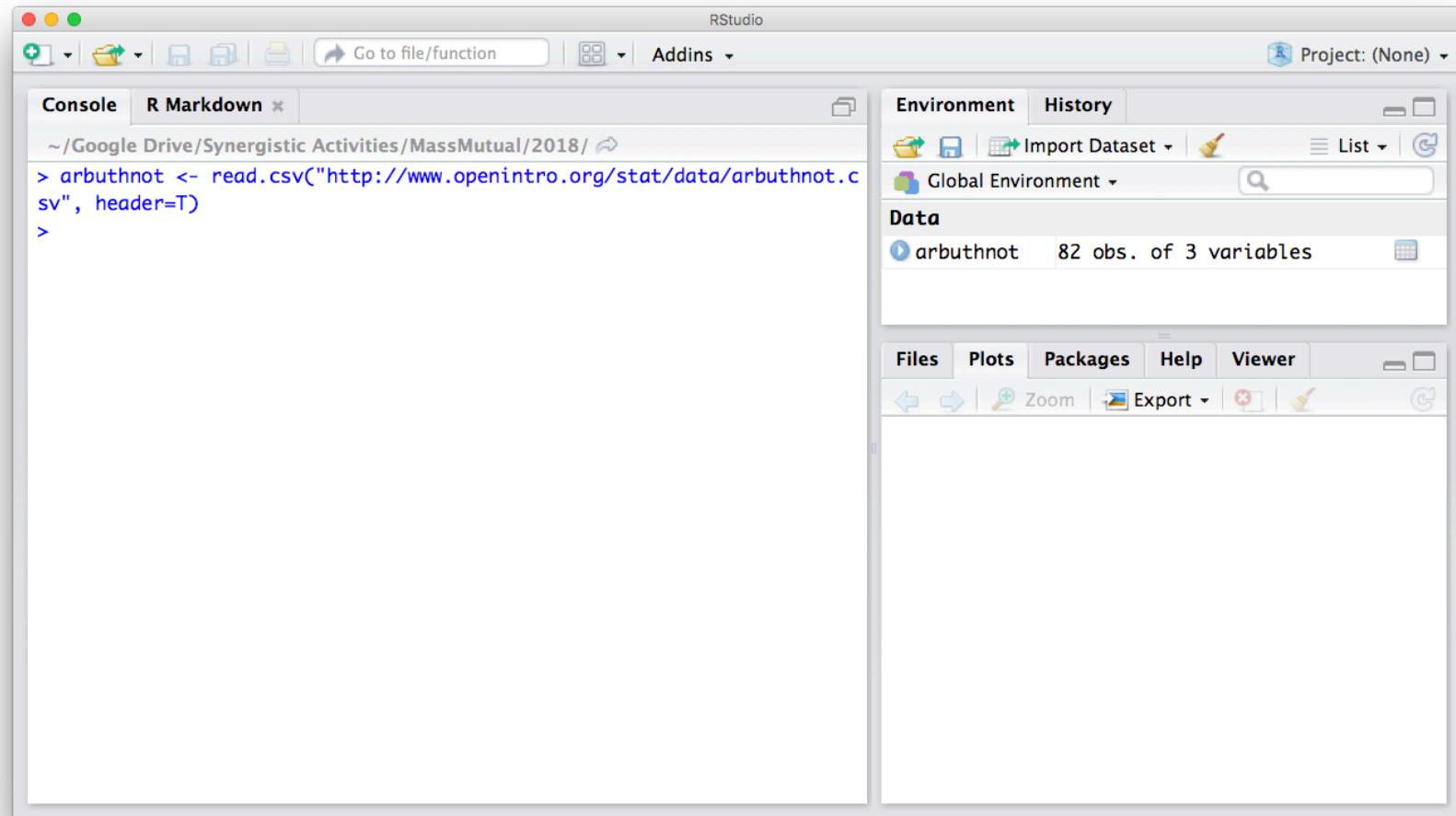
✓ Introductions

- Visualization overview
  - Flashback to early experiences in data wrangling
  - Visualization (def.)
  - Data (def.)
- Graphical primitives
- Visual dimensions
- Pre-lunch activity: mapping visual to data dimensions
- After lunch: ggplot2 crash course

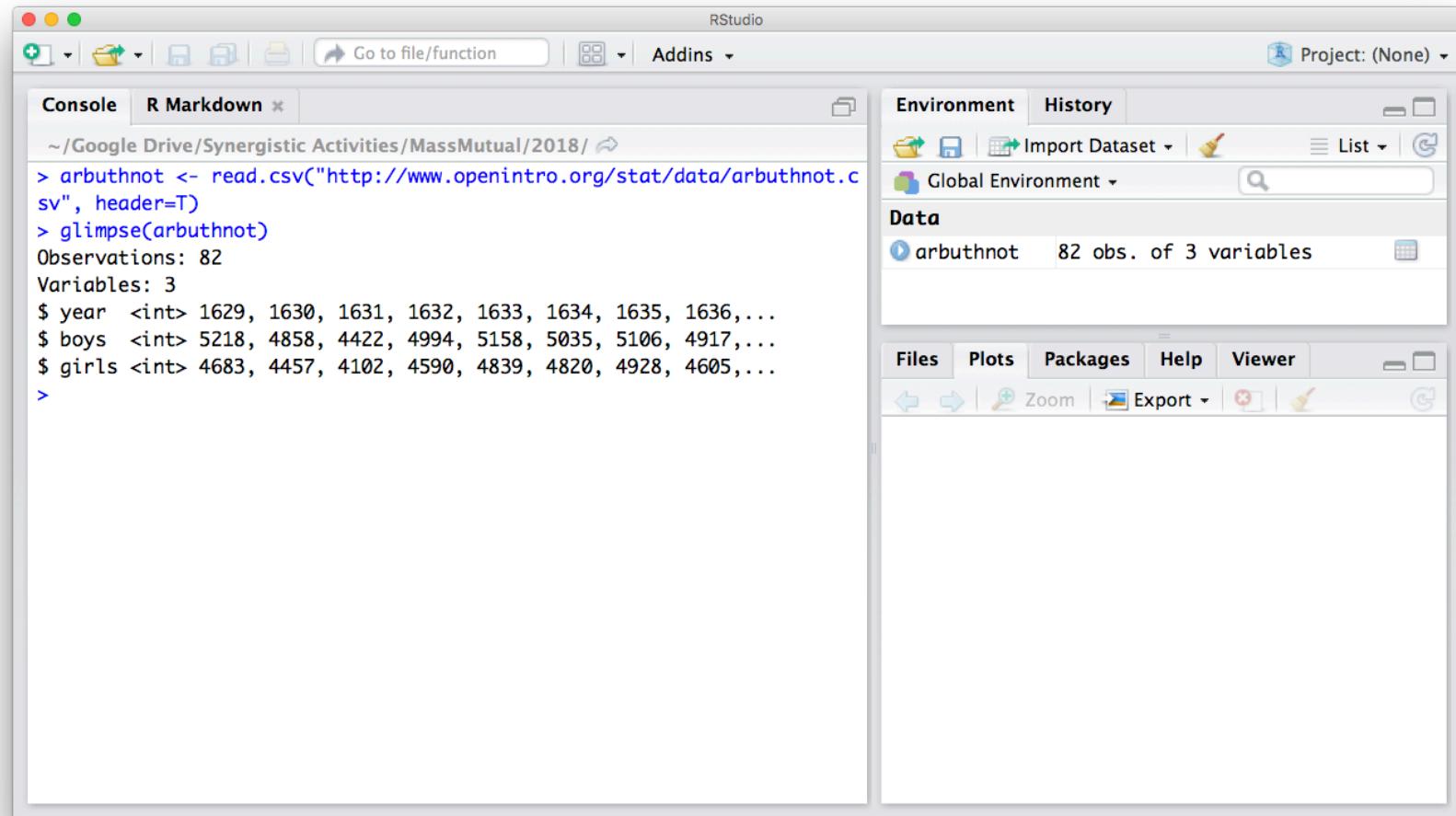
# Flashback...



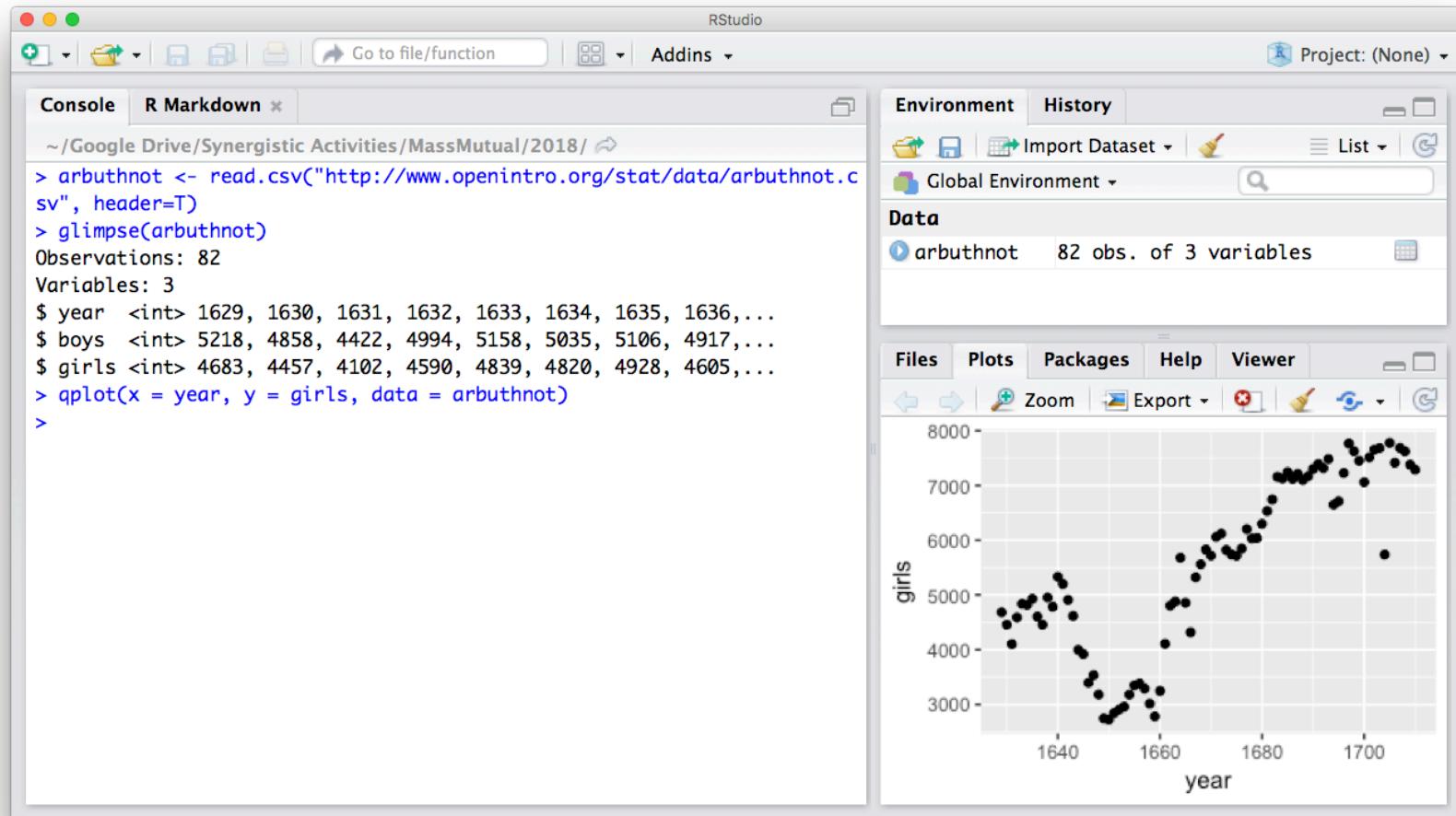
# Flashback...



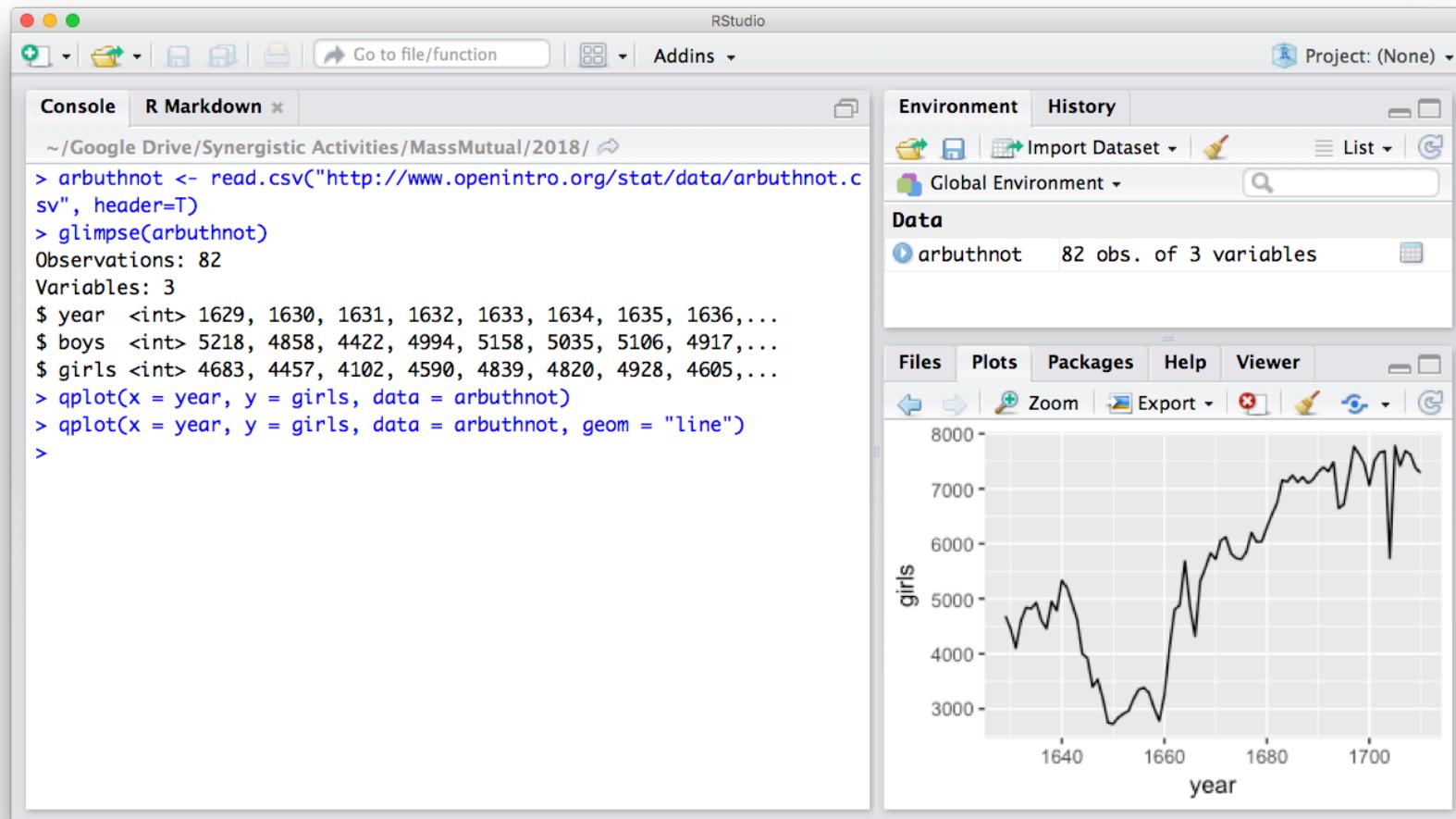
# Flashback...



# Flashback...



# Flashback...



# Question

What makes these  
“visualizations” **useful?**

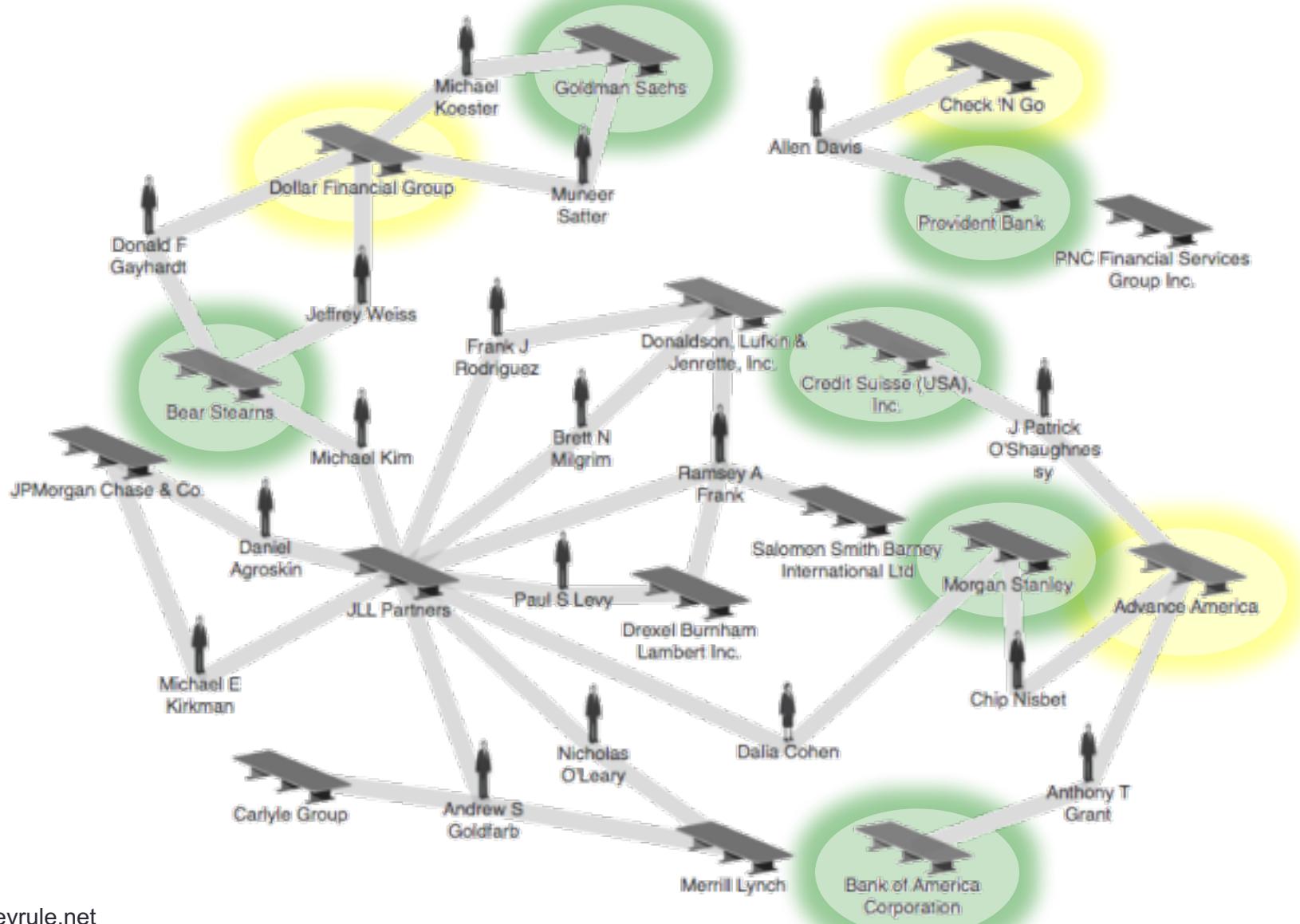


# Do they help you spot trends?



More info here: [http://en.wikipedia.org/wiki/1854\\_Broad\\_Street\\_cholera\\_outbreak](http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak)

# Do they help you explore?



# Do they tell a story?



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

# Visualization (def.)

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



# Data (def.)

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



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

# Data (def.)

and a corresponding set of *observations* (a.k.a. *records*) over these variables. For example:



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

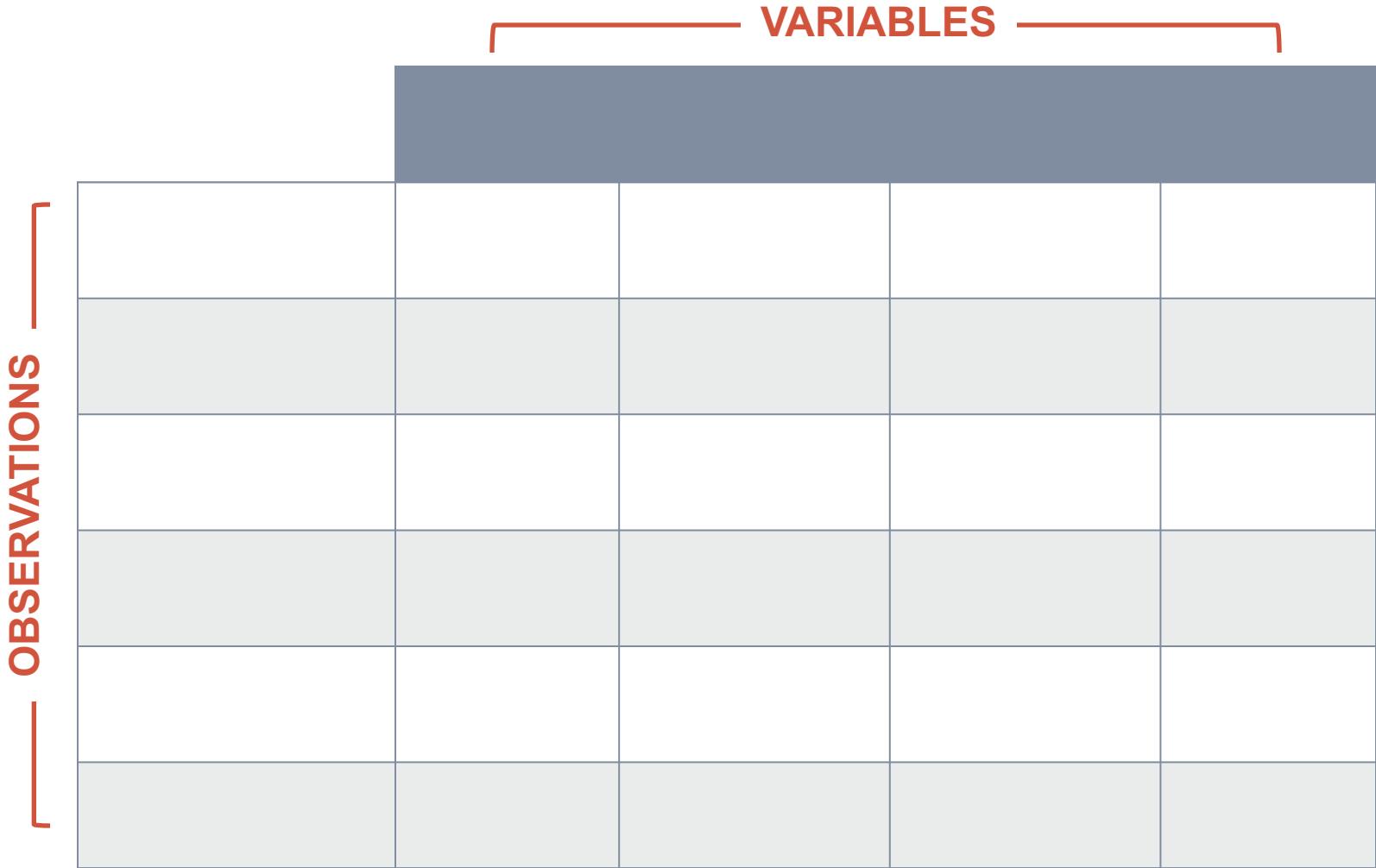
# Data (def.)

and a corresponding set of *observations* (a.k.a. *records*) over these variables. For example:



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

# One way to think about this:



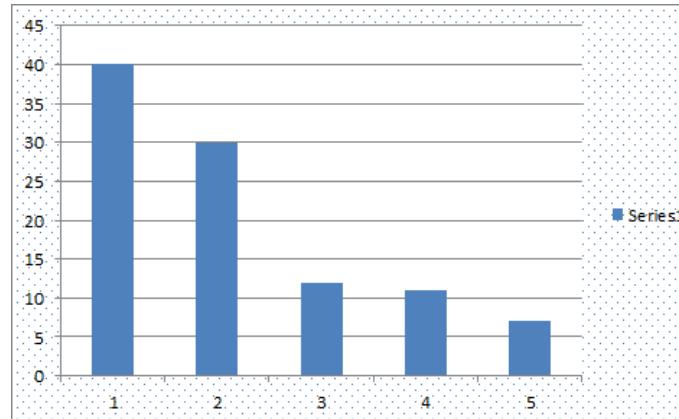
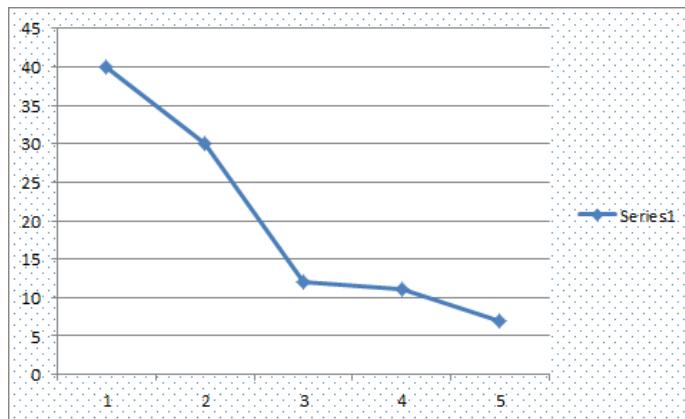
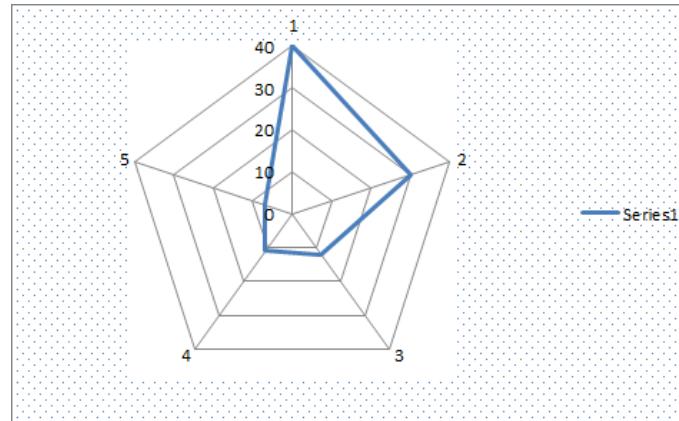
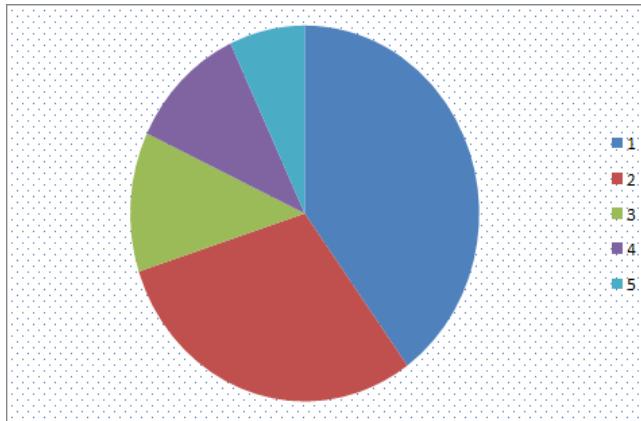
# Why is this important?

- Data have dimensions
- Visualizations have dimensions, too
- To build visualizations, we need to **map** data dimensions to visual dimensions

# Key question for this workshop

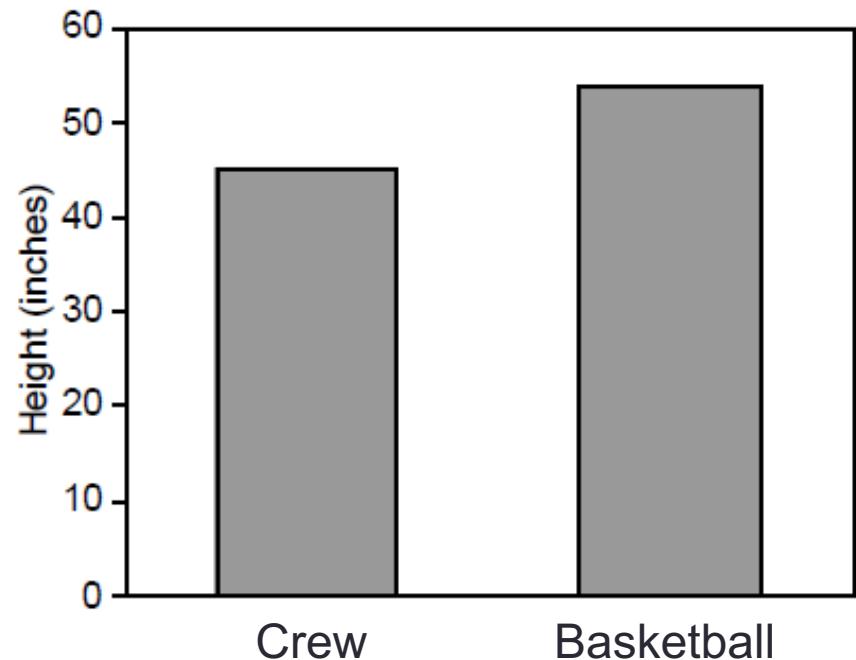
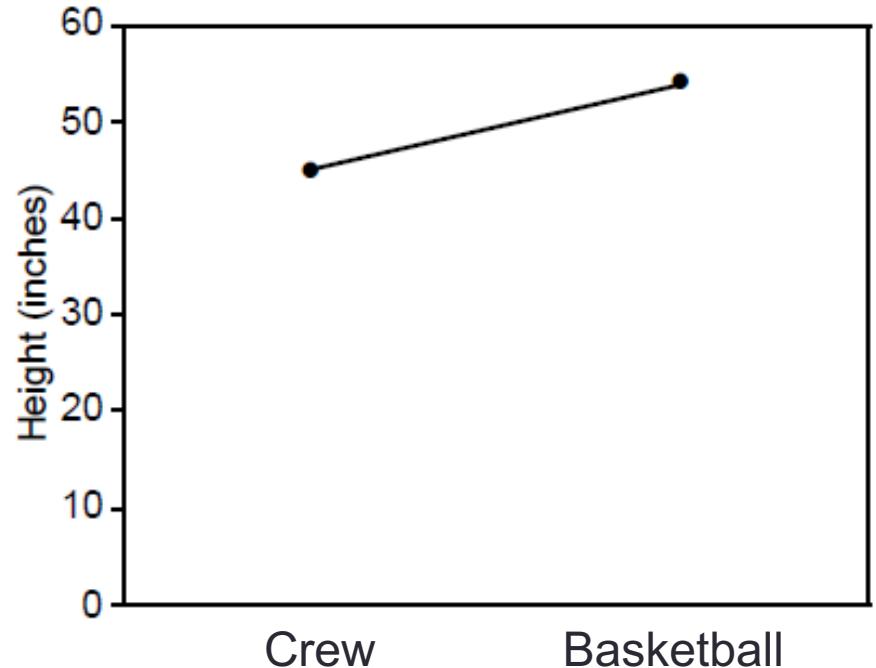
Which **data dimension** should be mapped

to which **visual dimension?**

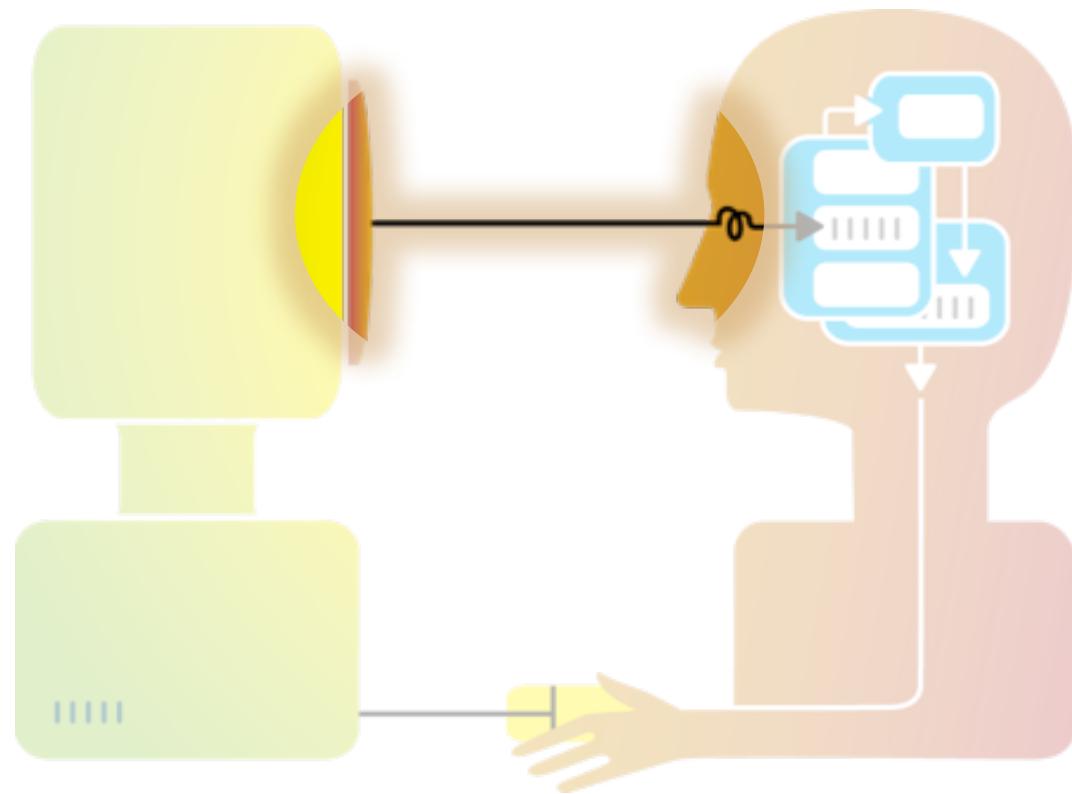


# Answer: it depends

Average Height for Youth Sports Participants

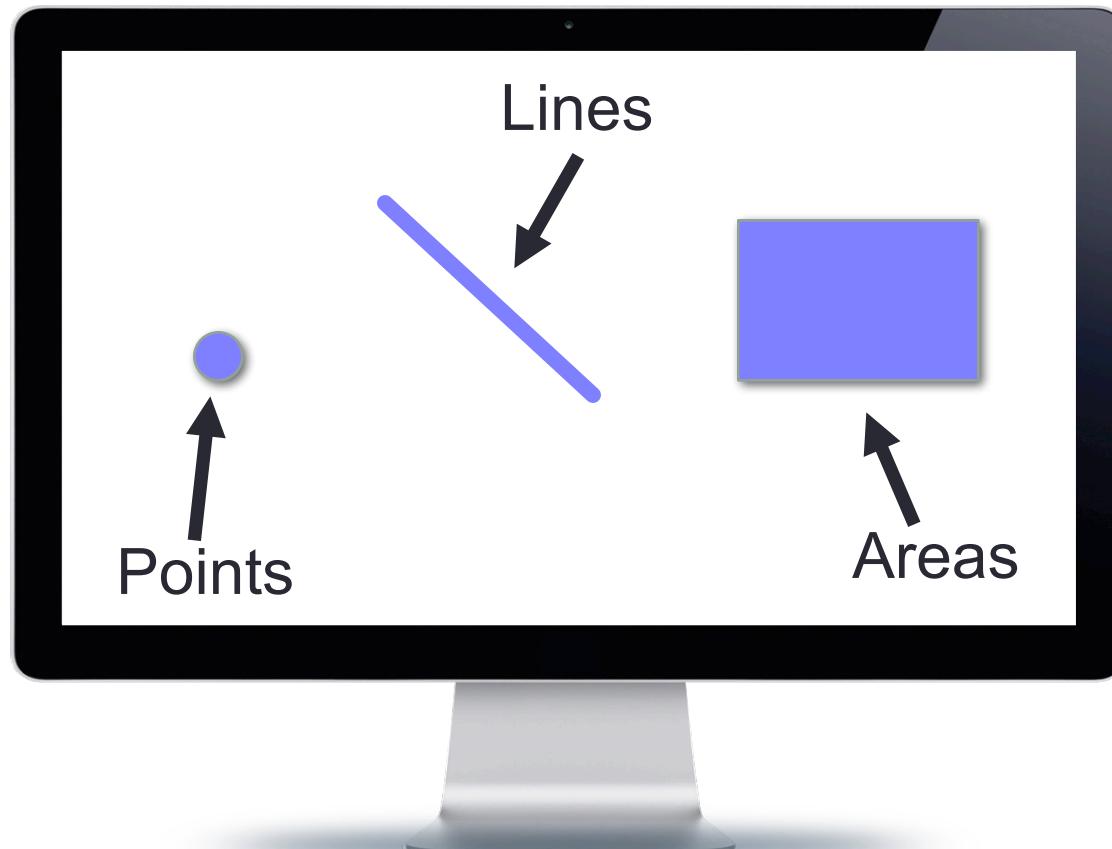


# So what do we have to work with?



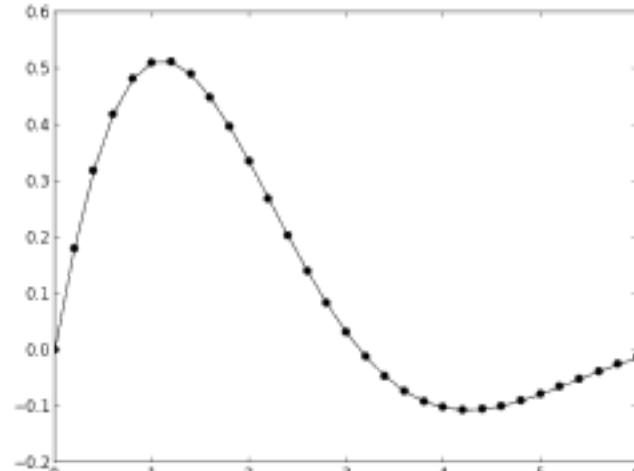
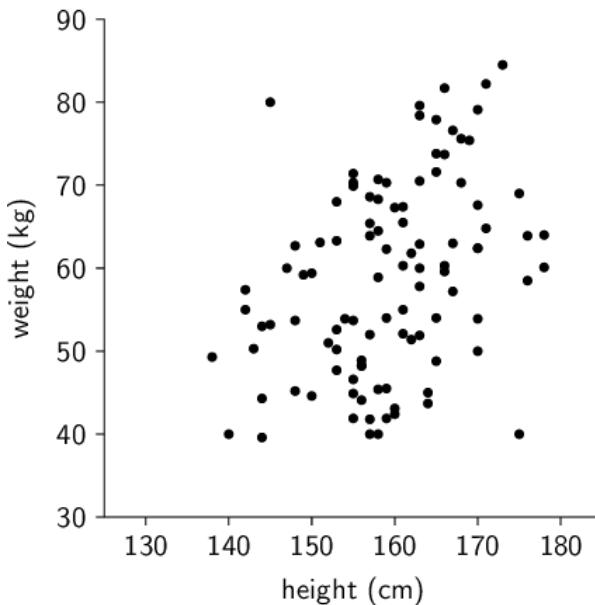
# Graphical primitives

The images we draw are composed of marks: like ink



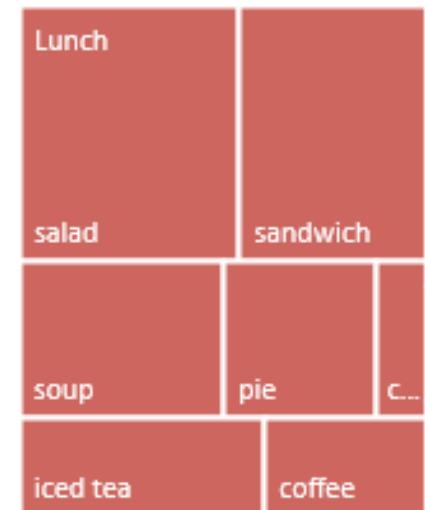
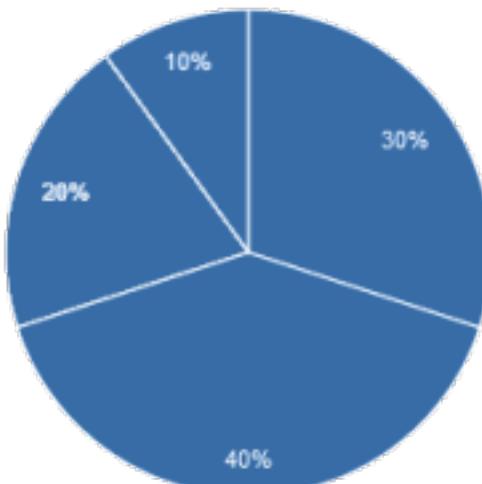
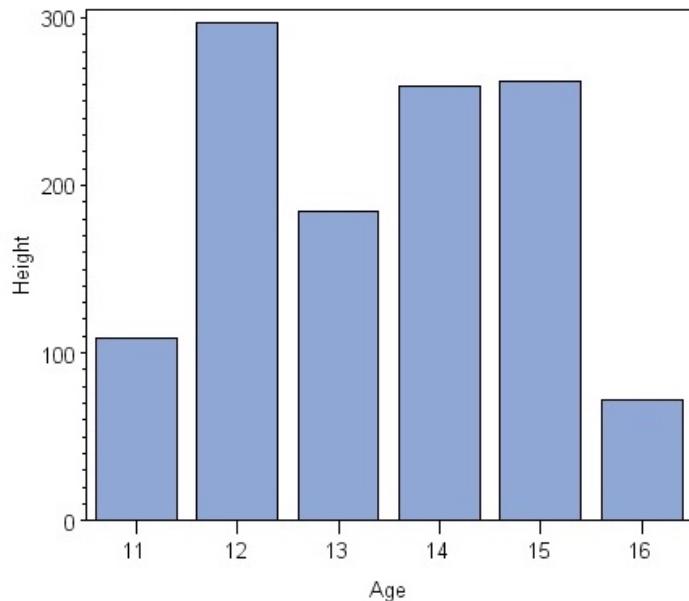
# Visual dimension: position

- Encode information using **where** the mark is drawn
- Some examples:



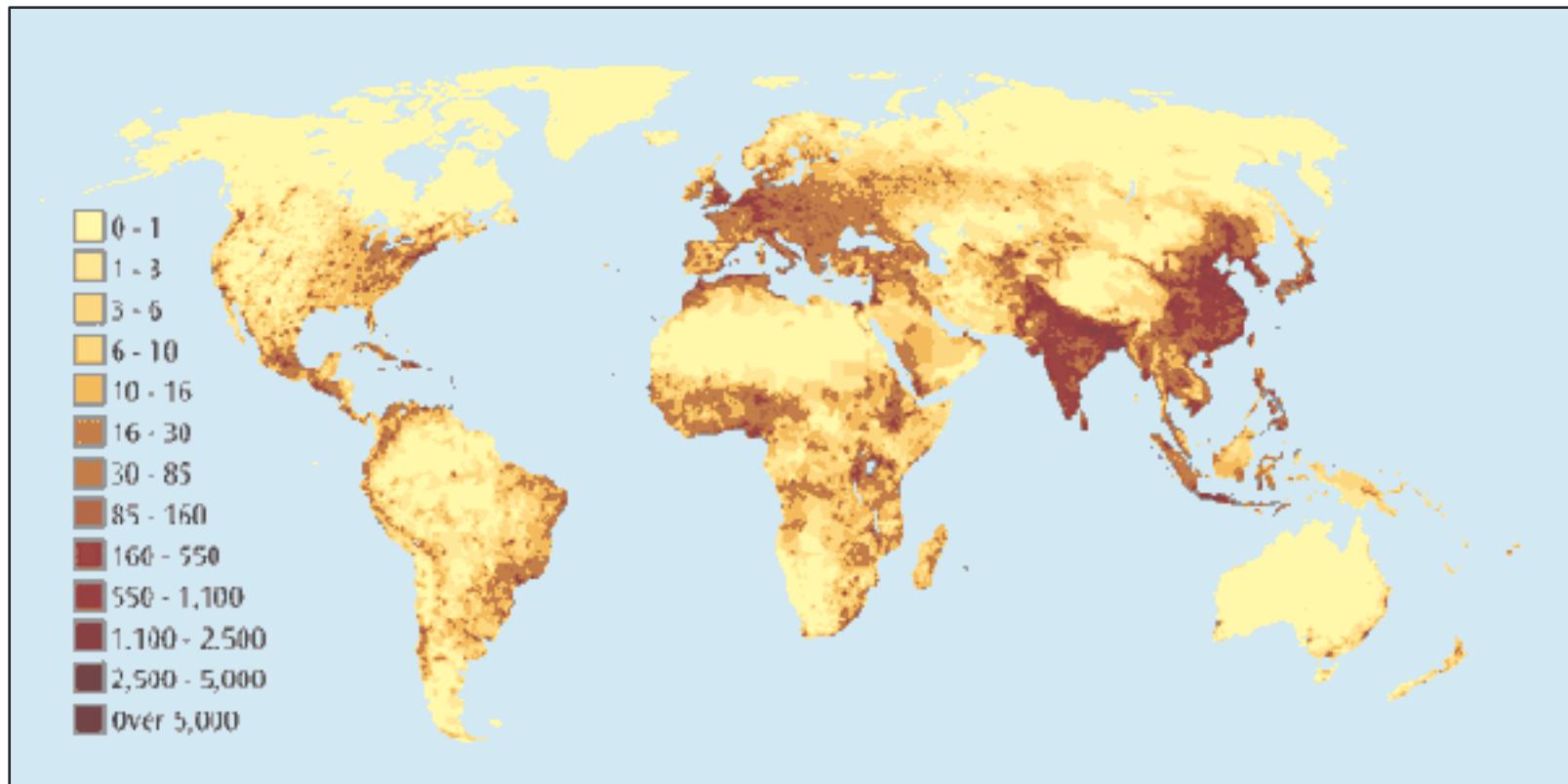
# Visual dimension: size

- Encode information using **how big** the mark is drawn
- Examples:



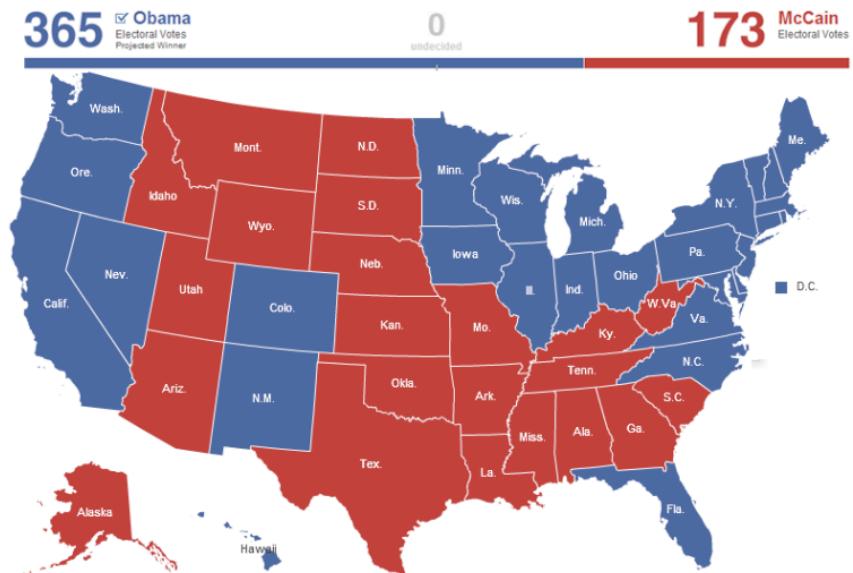
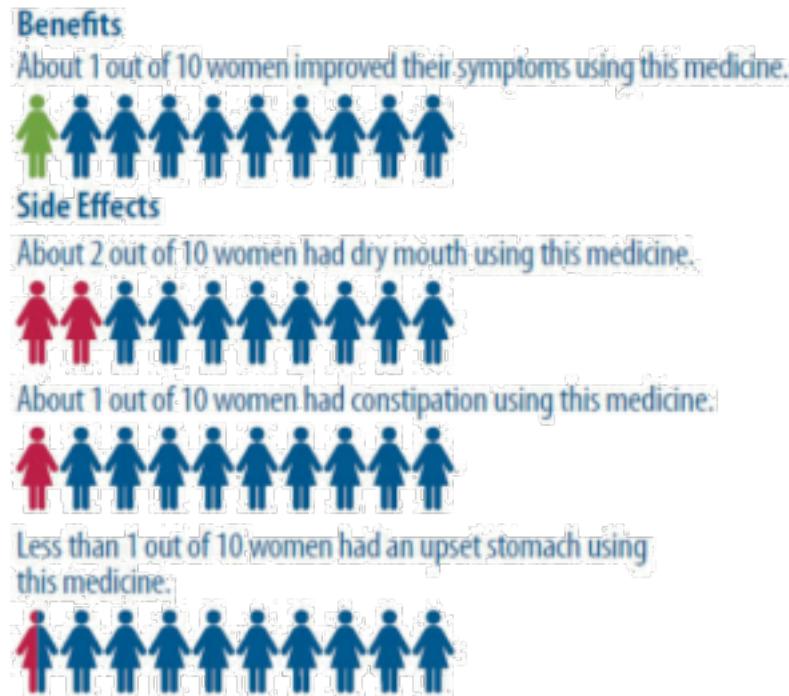
# Visual dimension: value

- Encode information using **how dark** the mark is drawn
- Example:



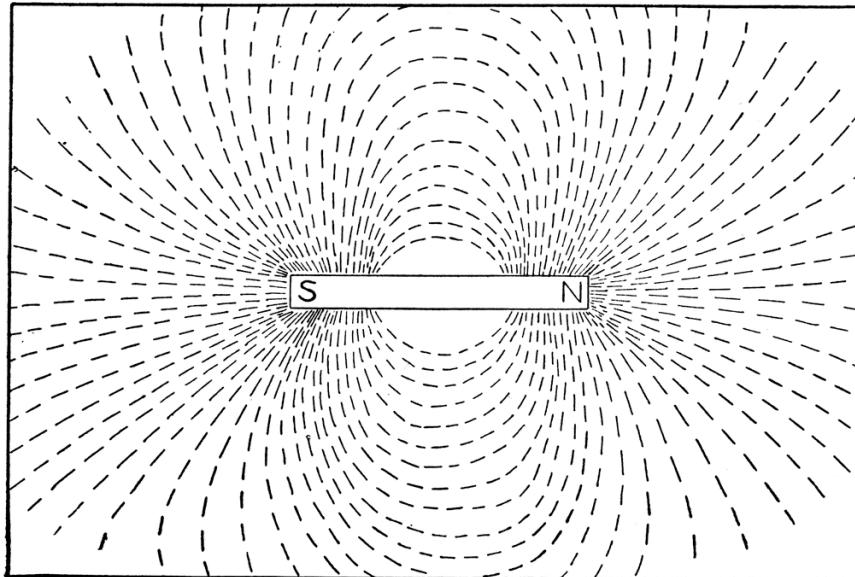
# Visual dimension: color

- Encode information using the **hue** of the mark
- Examples:



# Visual dimension: orientation

- Encode information using how the mark is **rotated**
- Examples:



# Visual dimension: shape

- Encode information using how the mark is **shaped**
- Examples:



# Discussion: visual dimensions & data type

	Categorical	Numerical
POSITION		
SIZE		
VALUE		
COLOR		
ORIENTATION		
SHAPE		



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

# Pre-lunch activity: deconstructing graphics

1. Find a data visualization you think is interesting
  - Some ideas: NYTimes, VisualisingData.com, Visual.ly
  - Remember to cite your source!
  
2. Identify the following:
  - What is the **data** that's being visualized? Where did it come from?
  - Which **data dimensions** are mapped to which **visual dimensions**?
  - How does this **shape your understanding** of the data?
  - If you **liked** the visualization: what is it doing **well**?
  - If you **disliked** the visualization: what would you **change**?

# After lunch



- Mini-lecture and lab: building data graphics with ggplot2
- **TODO** (if you haven't already):  
    > `install.packages('ggplot2')`