

# STORYTELLING WITH DATA

Joseph Nelson, Data Science Immersive

---

# AGENDA

---

- ▶ Introduction
  - ▶ Statistics
  - ▶ Visual Design
  - ▶ Excel Tips
  - ▶ Communicating
  - ▶ “Presentations”
- 
- ▶ (Work time and break spliced in between)

---

**ME**  
**AGENDA**

---

“ ”  
All models are wrong. Some are useful.

— George Box, 1978

# ME AGENDA

---

- Data Science Immersive Instructor
- **From:** Des Moines, Iowa
- **Influences:** Marc Andreessen & Ben Horowitz, Zuckerberg, Andrew Ng, Yann LeCun, Jürgen Schmidhuber
- **Likes:** Hockey, SaaS, bad data science puns, running



FLEISHMANHILLARD



---

# BASIC DESCRIPTIVE STATISTICS

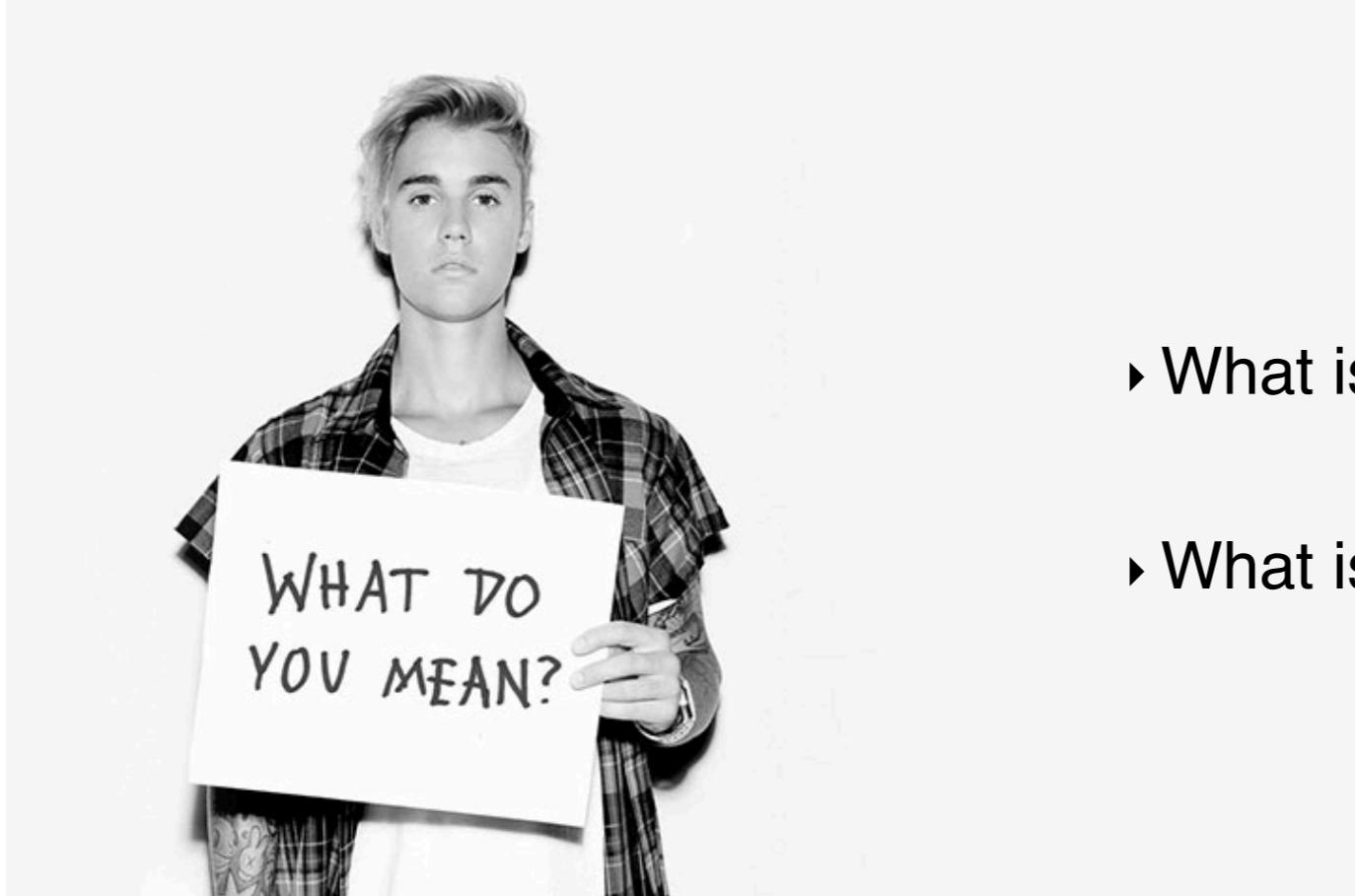
---

- ▶ Mean
- ▶ Median
- ▶ Mode
- ▶ Max
- ▶ Min
- ▶ Quartile
- ▶ Inter-quartile Range
- ▶ Variance
- ▶ Standard Deviation
- ▶ Correlation



# MEAN

---



- ▶ What is the mean?
- ▶ What is another name for the mean?

# MEAN

---



- ▶ What is the mean?
- ▶ The mean of a set of values is the sum of the values divided by the number of values. It is also called the average.
- ▶ It is also known as the average.
- ▶ Example: Find the mean of 19, 13, 15, 25, and 18

# MEDIAN

---

- ▶ What is the median?
- ▶ How do you find the median?



# MEDIAN

---

- ▶ What is the median?
- ▶ How do you find the median?
- ▶ Bonus: Why might the median be advantageous instead of the mean? When does this condition NOT hold?



# MEDIAN

---

- ▶ The median refers to the midpoint in a series of numbers.
- ▶ To find the median, arrange the numbers in order from smallest to largest. If there is an odd number of values, the middle value is the median. If there is an even number of values, the average of the two middle values is the median.



# MEDIAN

---

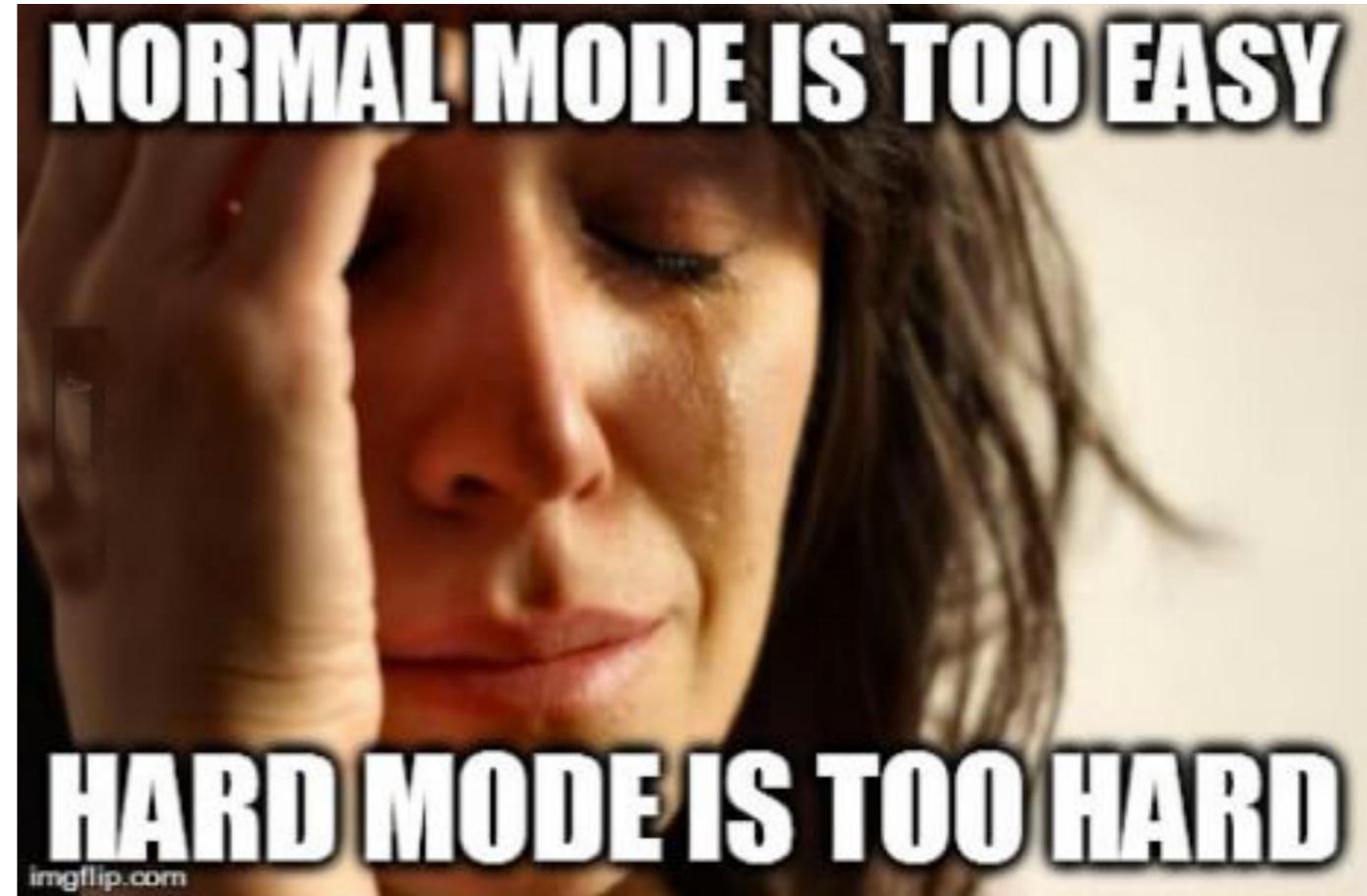
- ▶ The median refers to the midpoint in a series of numbers.
- ▶ Example #1: Find the median of 19, 29, 36, 15, and 20
- ▶ Example #2: Find the median of 67, 28, 92, 37, 81, 75
- ▶ Bonus: Median may be more useful than average in a highly skewed population.



## **MODE**

---

- ▶ What is the mode?
- ▶ What is the mode in the following: 1, 2, 3, 4, 5



# MODE

---

- ▶ What is the mode?
- ▶ The mode of a set of values is the value that occurs most often.
- ▶ A set of values may have more than one mode or no mode.



## CHECK FOR UNDERSTANDING

---

- ▶ For the following groups of numbers, calculate the mean, median and mode by hand:
- ▶ A. 18, 24, 17, 21, 24, 16, 29, 18
- ▶ B. 75, 87, 49, 68, 75, 84, 98, 92
- ▶ C. 55, 47, 38, 66, 56, 64, 44, 39



## CHECK FOR UNDERSTANDING

---

- ▶ Answers:
  - ▶ A. Mean = 20.875 Median = 19.5 Mode = 18,  
24 Max = 29 Min = 16
  - ▶ B. Mean = 78.5 Median = 79.5 Mode = 75 Max  
= 98 Min = 49
  - ▶ C. Mean = 51.125 Median = 51 Mode = none  
Max = 66 Min = 38



---

# HOW TO LIE WITH STATISTICS

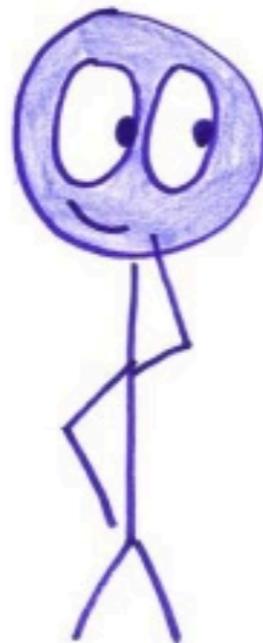
---

- ▶ For each picture:
  - ▶ 1) What could go wrong
  - ▶ 2) How to fix it

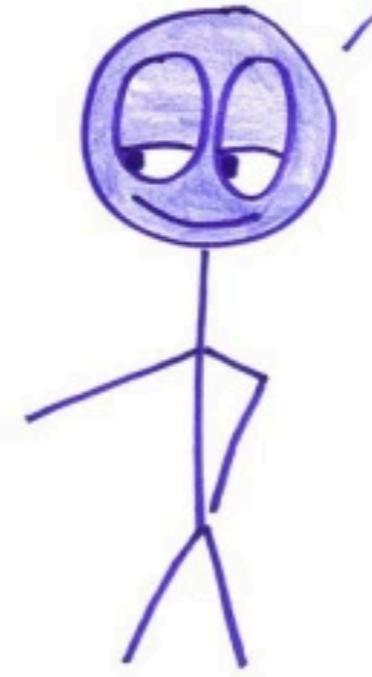
# HOW TO LIE WITH STATISTICS

Mean

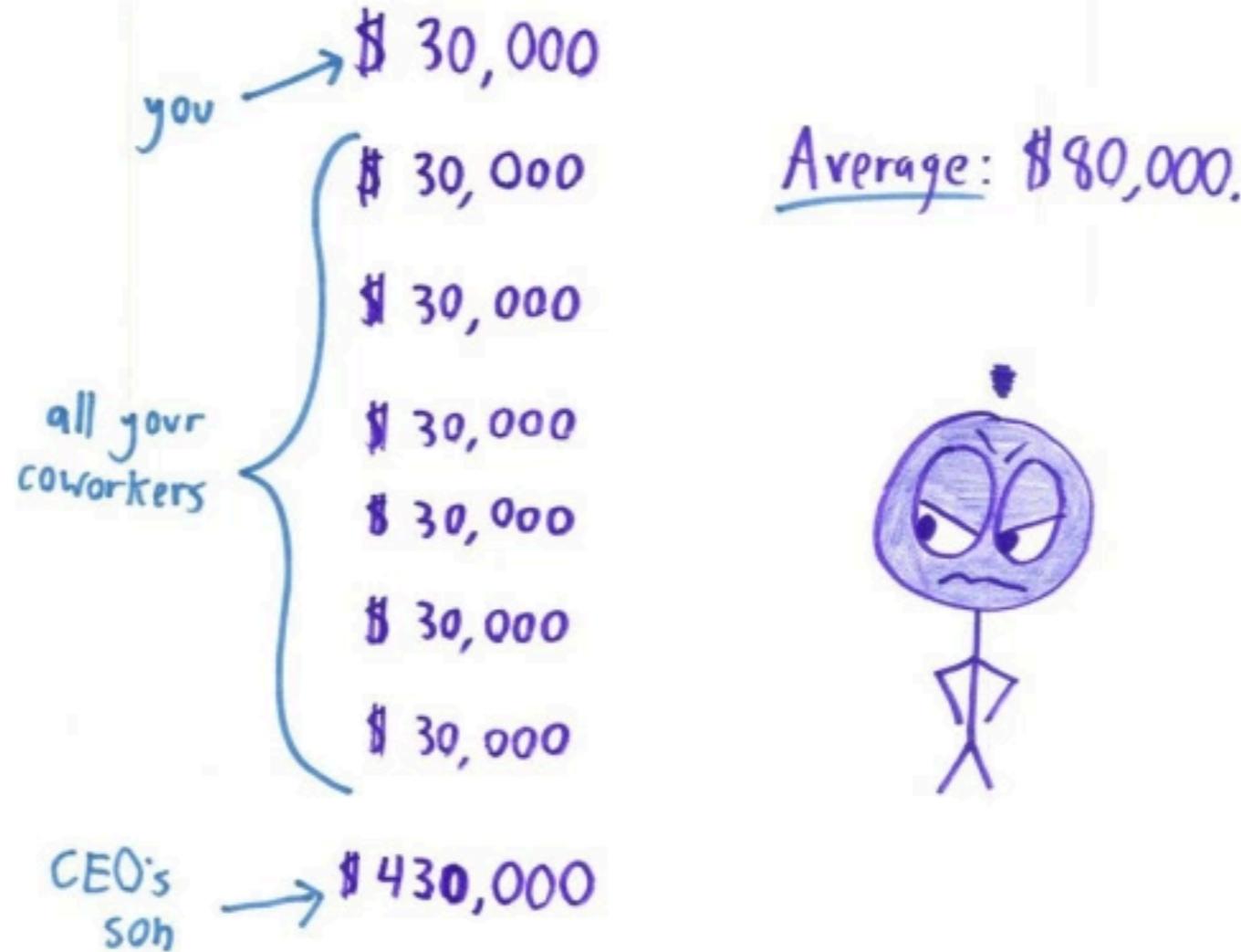
What would my  
starting salary be?



I'll put it this way:  
our average starting  
salary is \$80,000!



# HOW TO LIE WITH STATISTICS



# HOW TO LIE WITH STATISTICS

Median

So, why should I  
invest with you?

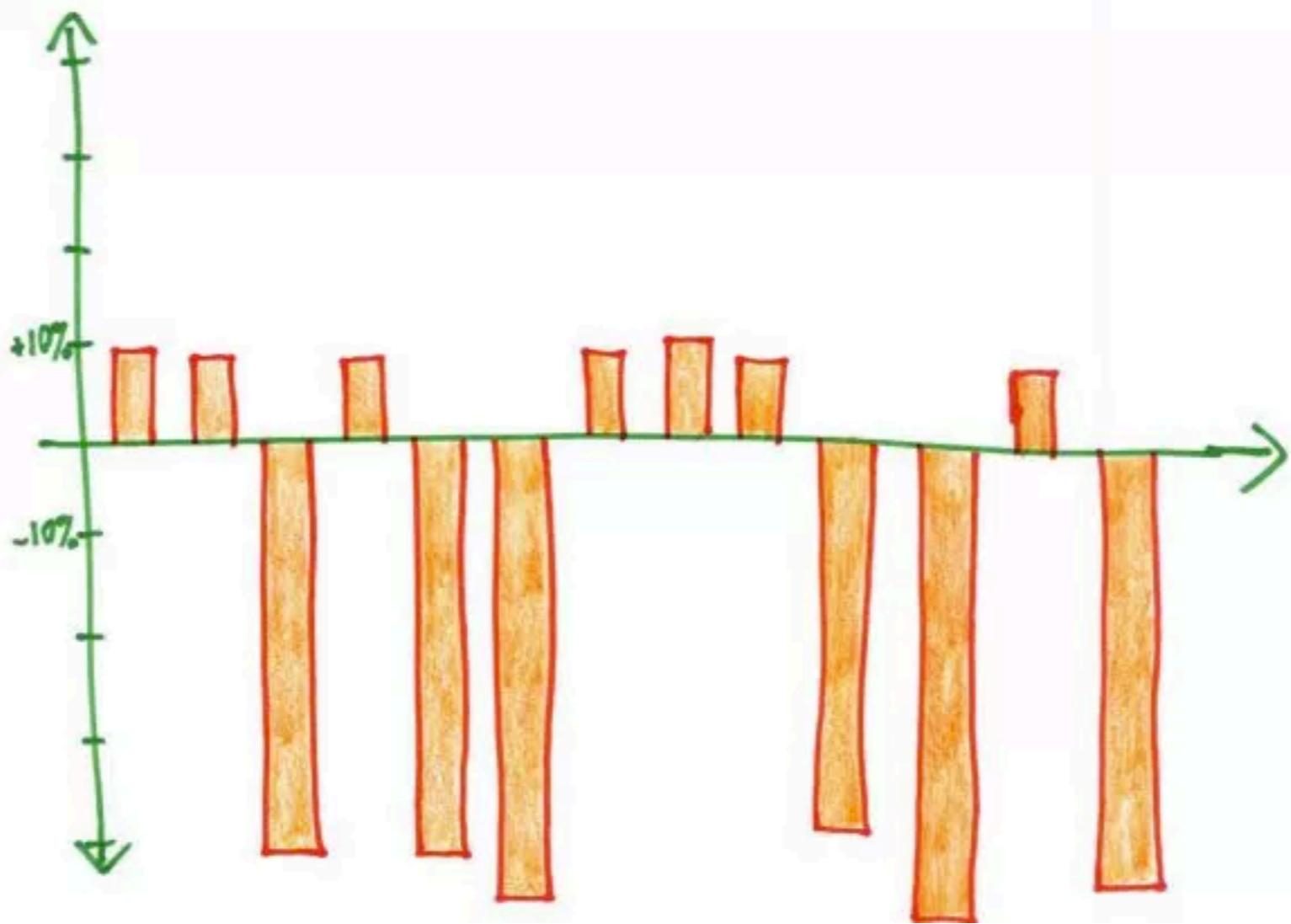


Well, not to brag, but  
my fund has a median  
gain of 8% per year!



# HOW TO LIE WITH STATISTICS

---



# HOW TO LIE WITH STATISTICS

Mode

How are you doing  
on your tests?



My modal category  
is 70-80%!



# HOW TO LIE WITH STATISTICS



Score Category	Number of Tests
90s	0
80s	0
70s	2
60s	1
50s	1
40s	1
30s	1
20s	1

# HOW TO LIE WITH STATISTICS

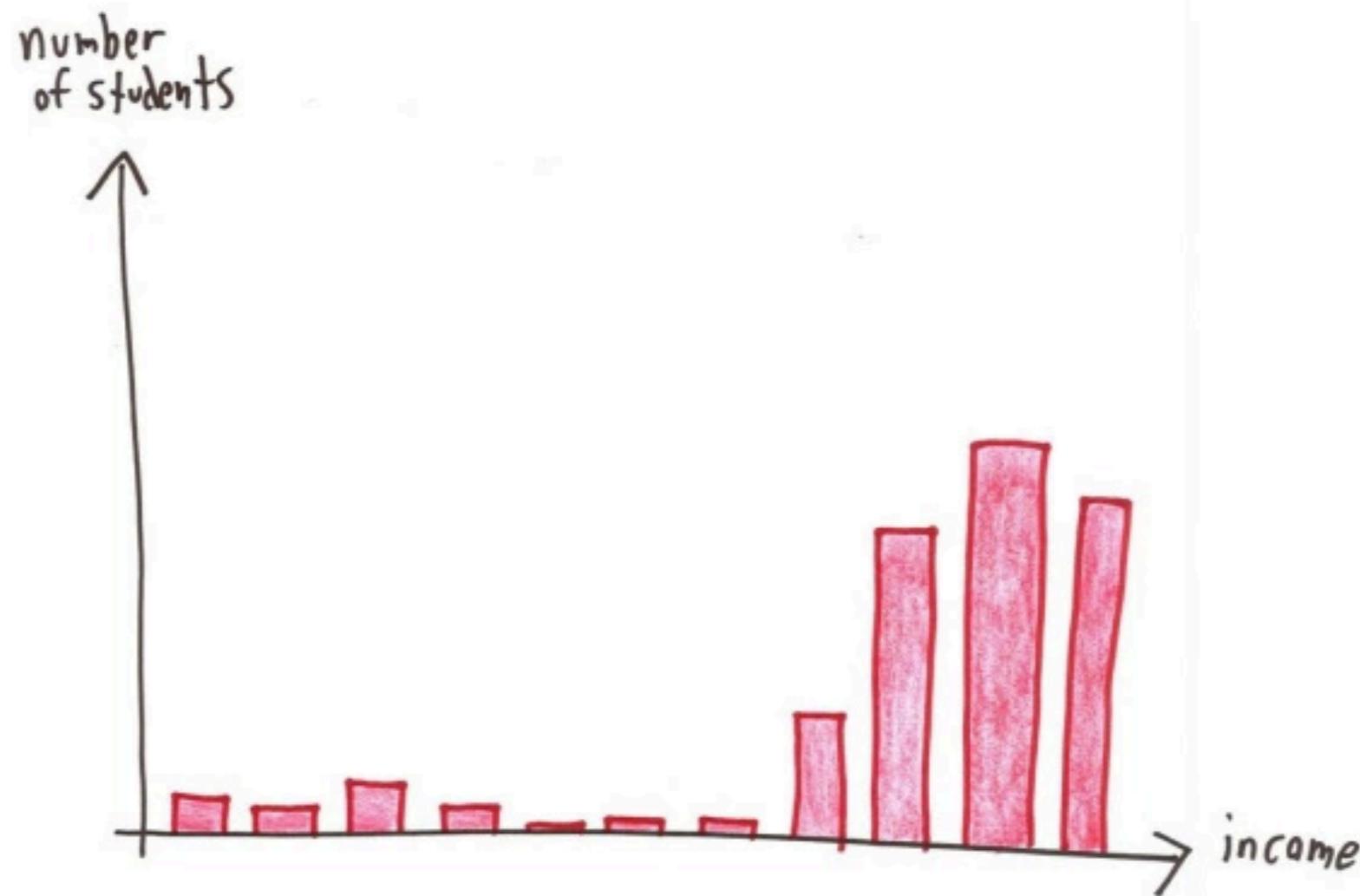
Range

Our students come from a  
wide range of  
Socioeconomic  
backgrounds...

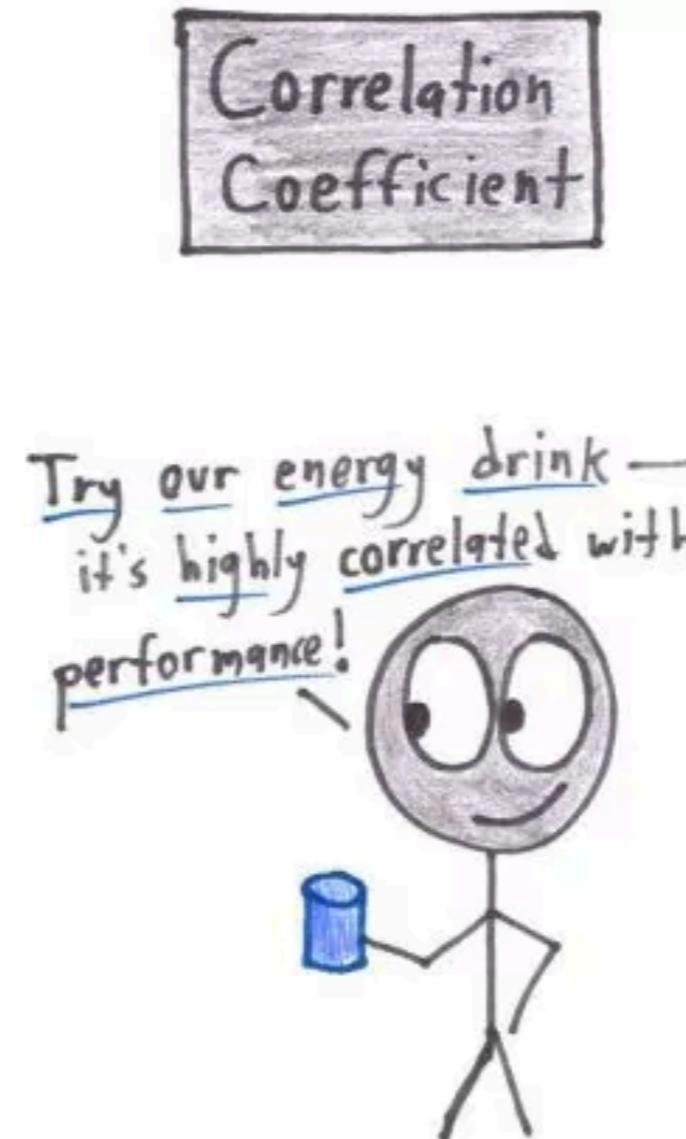


# HOW TO LIE WITH STATISTICS

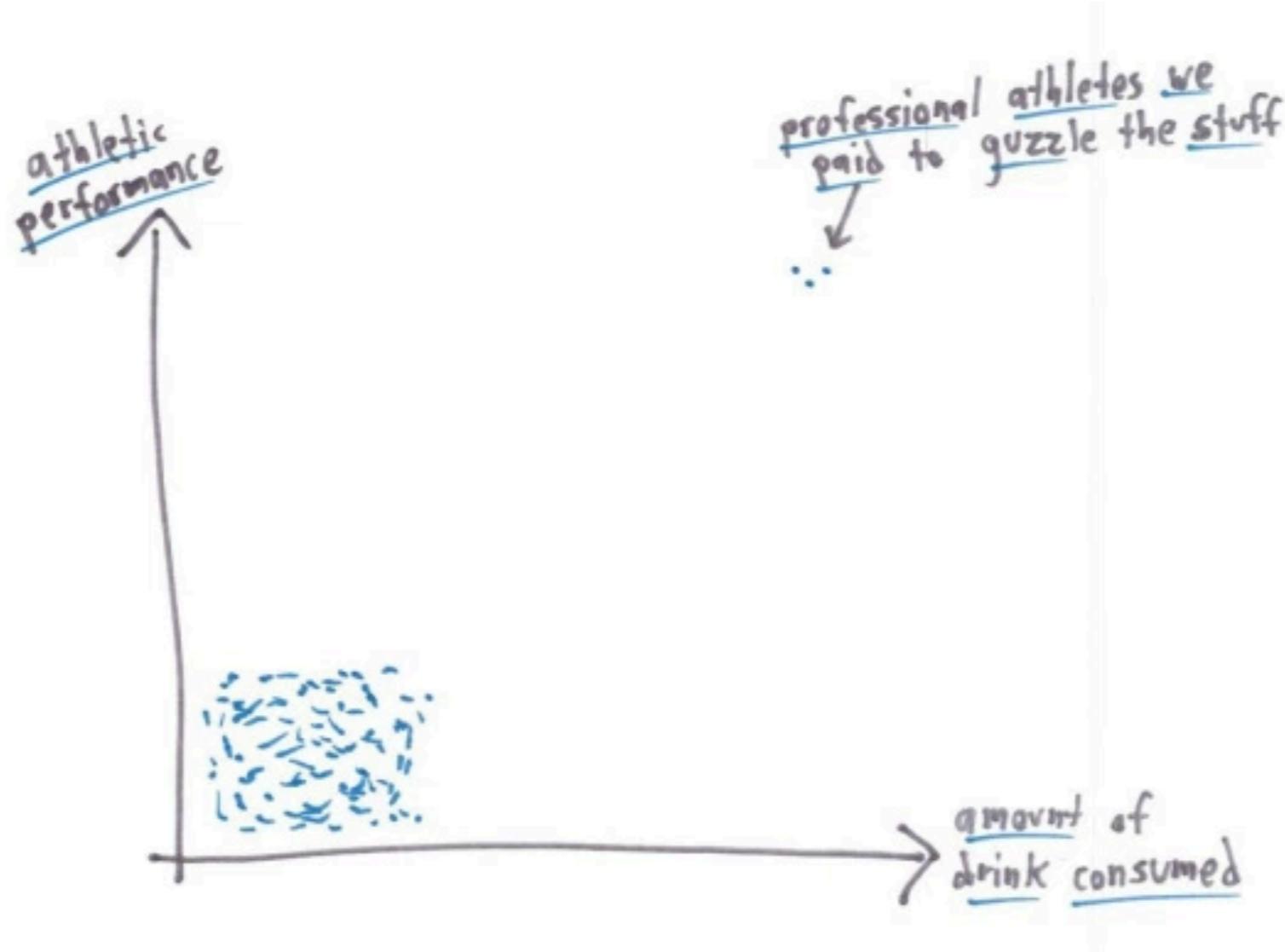
---



# HOW TO LIE WITH STATISTICS



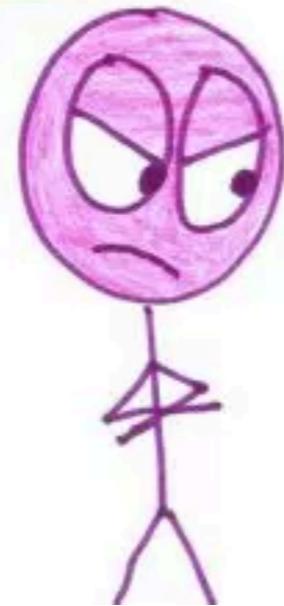
# HOW TO LIE WITH STATISTICS



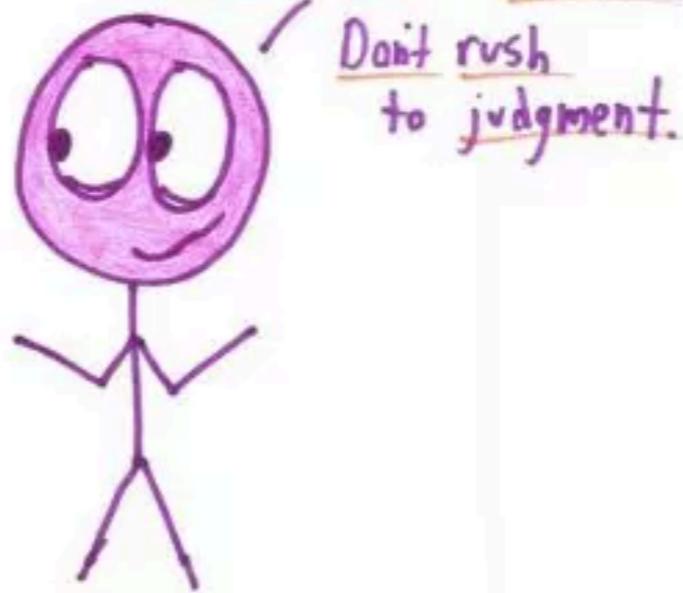
# HOW TO LIE WITH STATISTICS

## Variance

These results are  
a disaster!

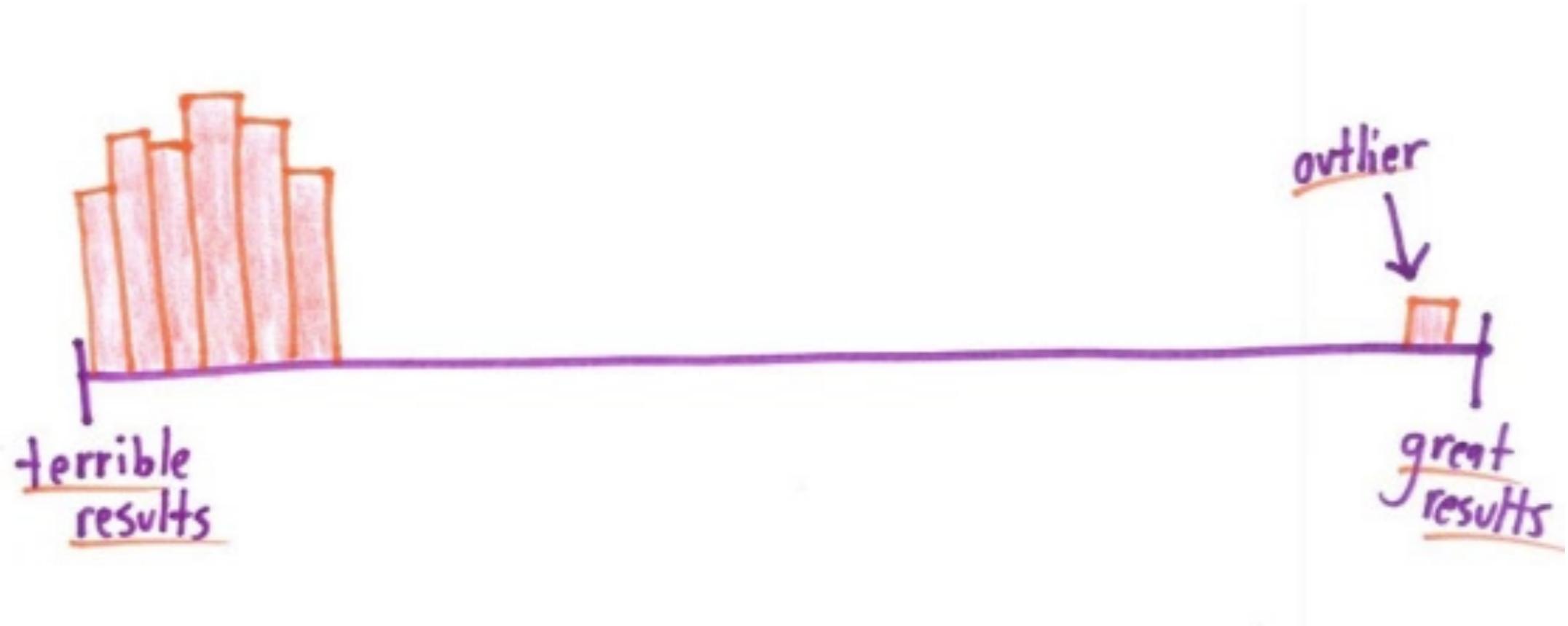


Sure, they look bad,  
but there's a lot of variance!



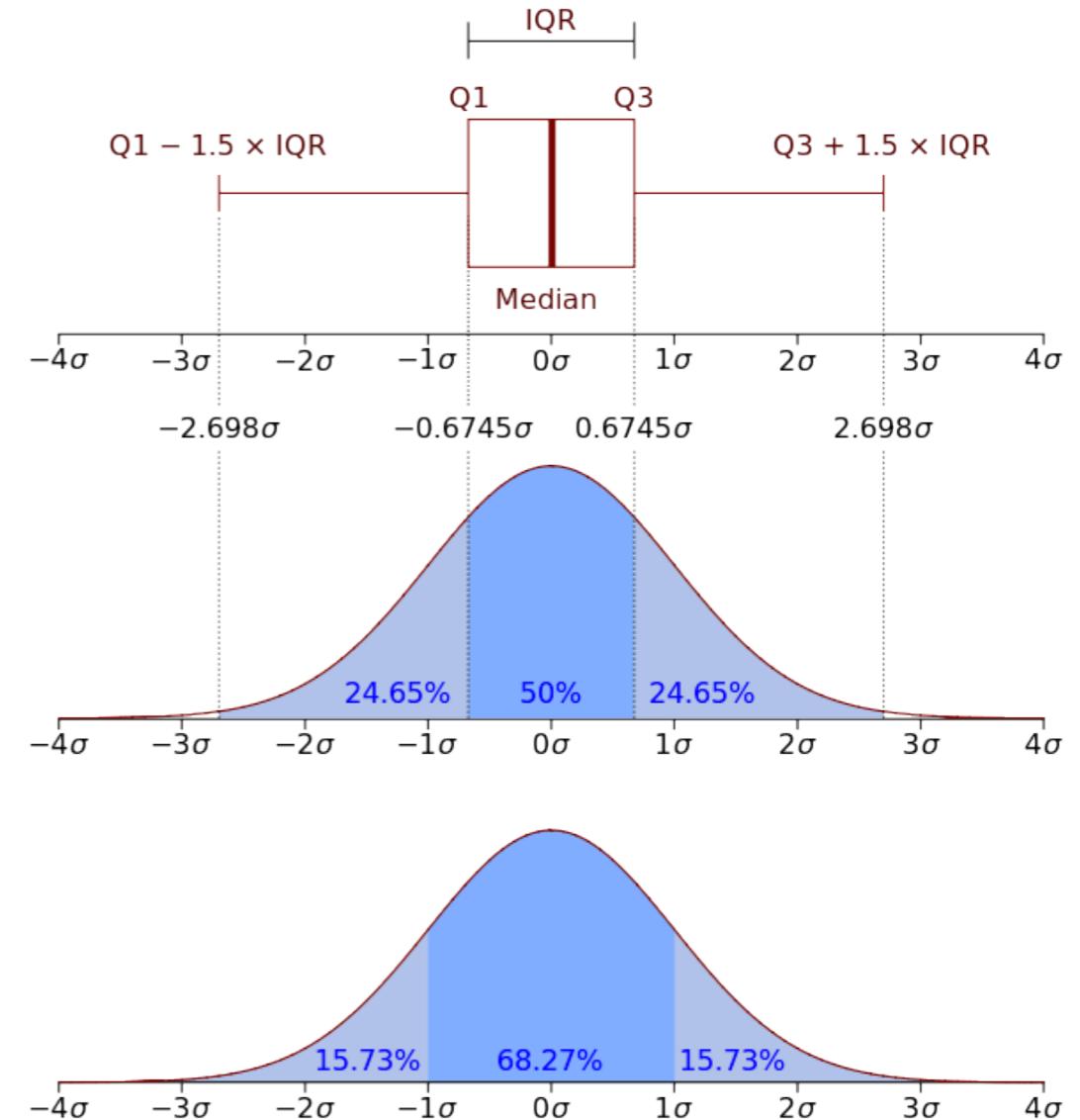
Don't rush  
to judgment.

# HOW TO LIE WITH STATISTICS



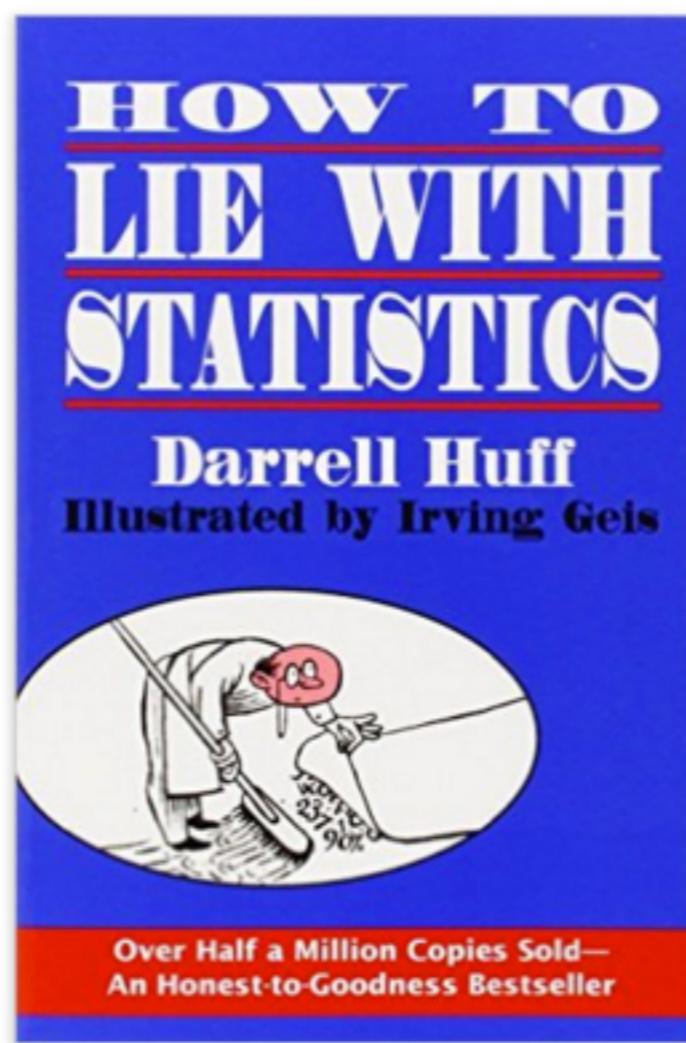
# QUARTILES AND THE INTER QUARTILE RANGE

- ▶ Quartiles divide a rank-ordered data set into four equal parts.
- ▶ The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q<sub>1</sub>, Q<sub>2</sub>, and Q<sub>3</sub>, respectively.
- ▶ The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. It is the “middle 50” of your data. Also called the H-spread.  
 $IQR = Q_3 - Q_1$
- ▶ Outliers:  $Q_1 - 1.5(IQR)$ ,  $Q_3 + 1.5(IQR)$



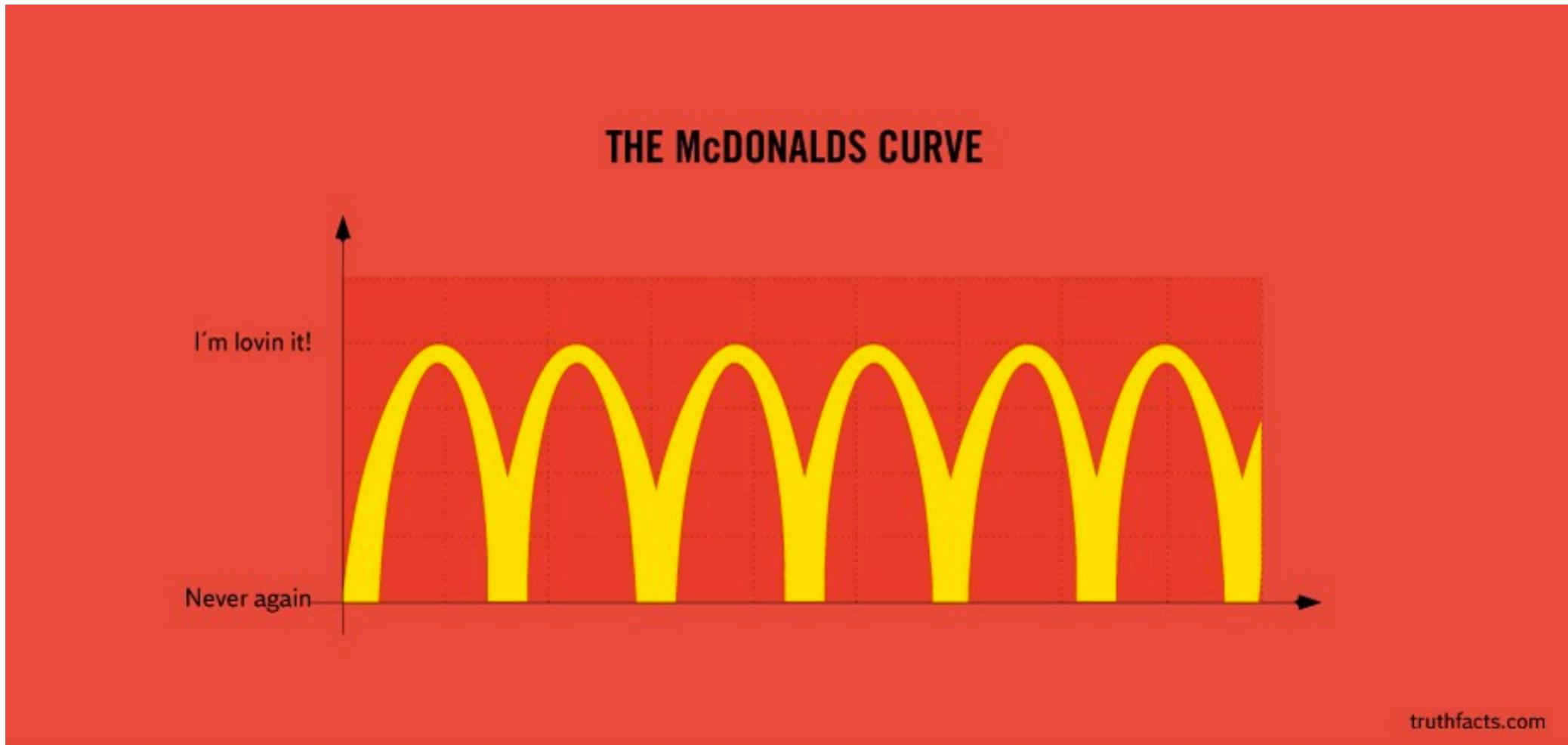
# CRITERIA FOR GOOD VISUALIZATION

---



## COMMUNICATING RESULTS TO NON-TECHNICAL AUDIENCES

- ▶ Should you present results differently for ensemble models vs non-ensemble models?



## REVIEWING MODEL QUALITY

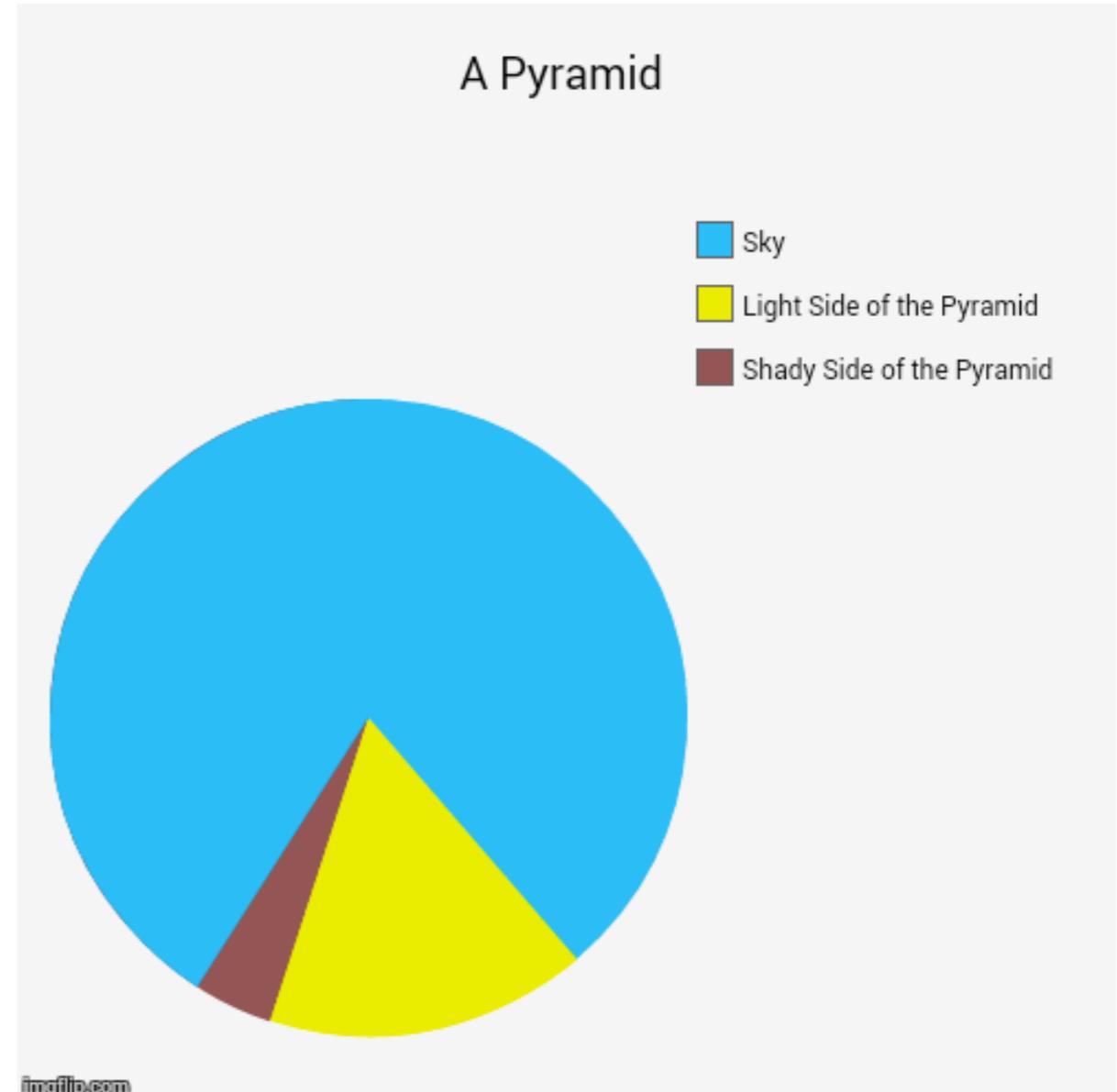
---

- ▶ High level review question: what are the two types of models we've discussed?  
How do they differ?

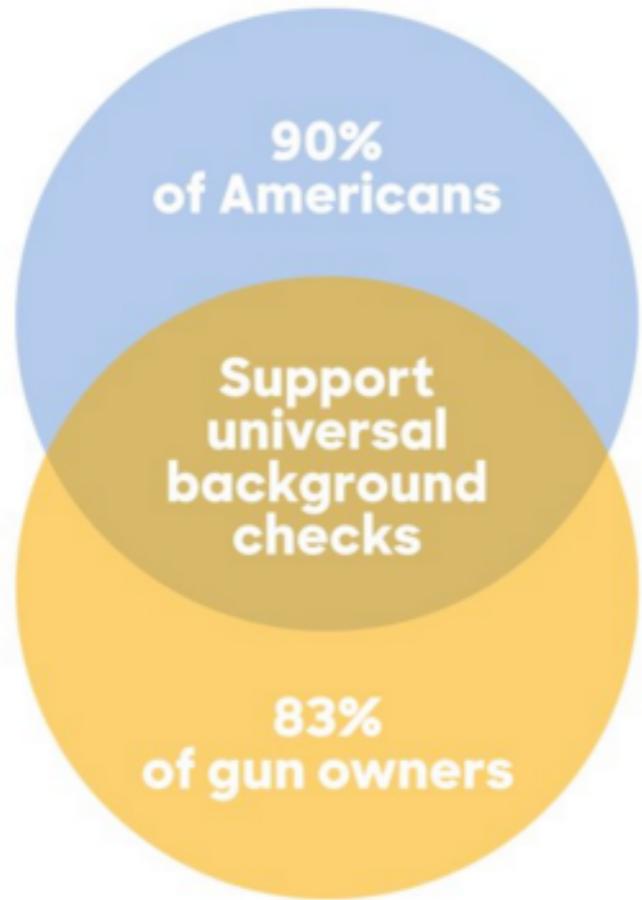


# CRITERIA FOR GOOD VISUALIZATION

- ▶ Simplified
- ▶ Easy to Interpret
- ▶ Clearly Labeled
- ▶ Ask yourself:
  - ▶ Who is my target audience?
  - ▶ What do they already know, and what do they need to know?
  - ▶ How does my project affect this audience? How might they interpret (or misinterpret) the data?



# CRITERIA FOR GOOD VISUALIZATION



Hillary Clinton

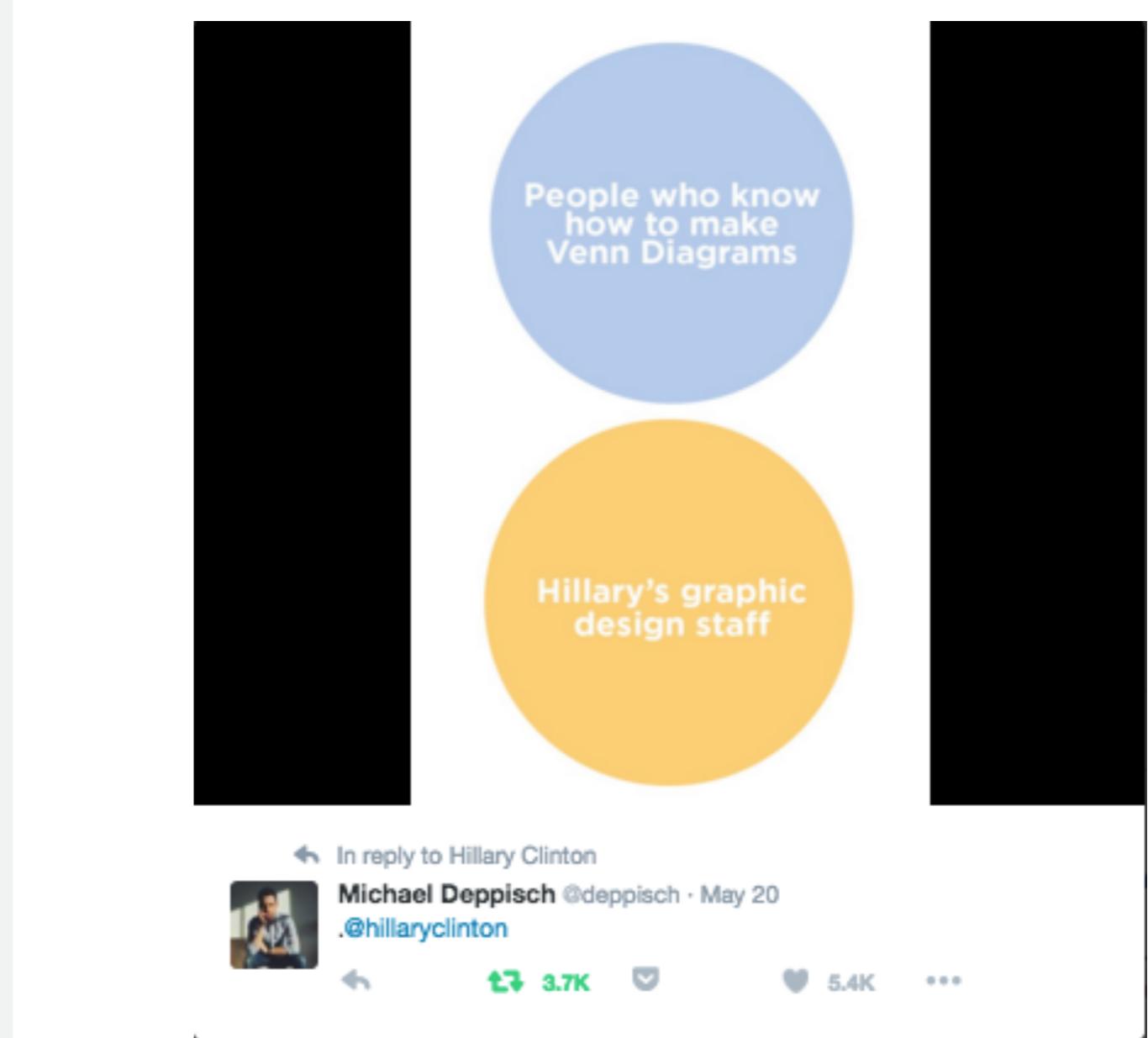
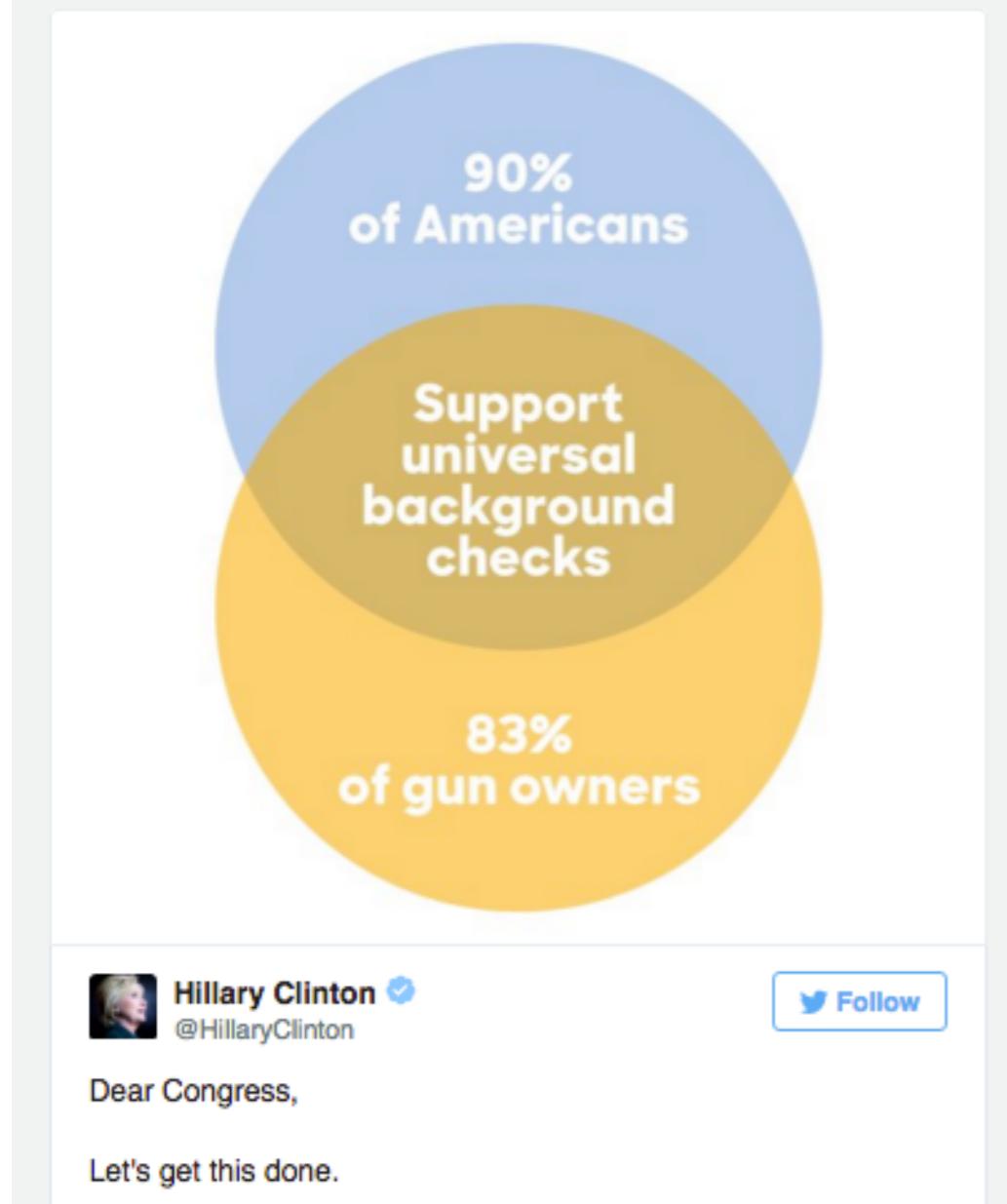
@HillaryClinton

Follow

Dear Congress,

Let's get this done.

# CRITERIA FOR GOOD VISUALIZATION



# IT'S A BIPARTISAN PROBLEM

 **GOP**   
@GOP

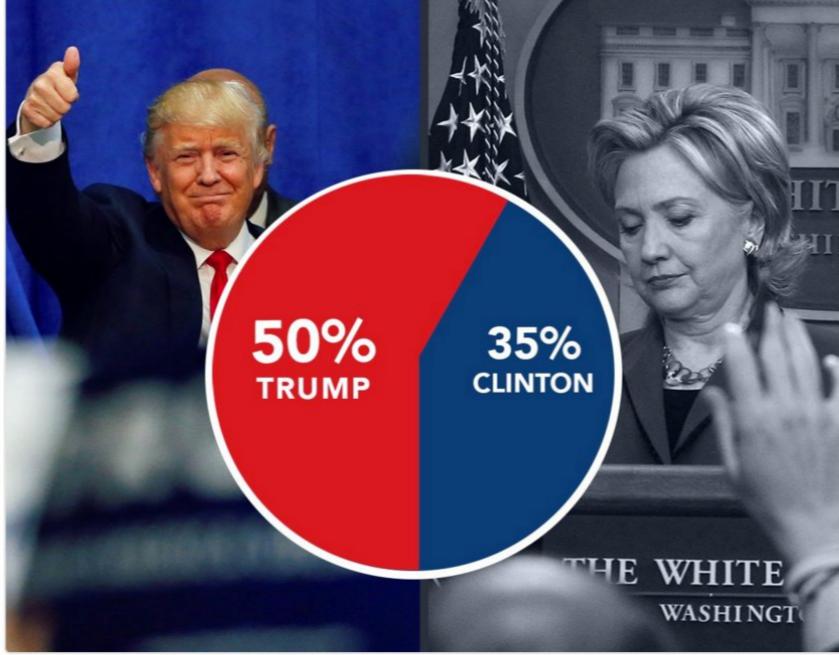
 **Following**

.@HillaryClinton's persistent lies have lost her the trust of the American people 

**VOTERS TRUST TRUMP OVER CLINTON**

— VOTERS SAY TRUMP IS MORE HONEST AND TRUSTWORTHY —

(CNN/ORC, 9/6/16)



**50%**  
**TRUMP**

**35%**  
**CLINTON**

THE WHITE  
WASHINGTON

RETWEETS   LIKES  
**39**   **65**



3:26 PM - 7 Sep 2016

---

## **COMMUNICATION - CONTENT**

---

- ▶ Establish a central thesis
- ▶ Create a narrative arc: problem, addressing, result
- ▶ “Completed X by doing Y as measured by Z”
- ▶ Follow the SAR principles: Situation, Action, Result

---

## **COMMUNICATION - DELIVERY**

---

- ▶ Establish a confident stance
- ▶ Engage your audience: fill the room and ensure your audience knows you're speaking to every one of them
- ▶ Speak as though you're writing: you begin with a thesis, topic sentence, and example
- ▶ Offer a “nugget” and reflect humility
- ▶ A pause is more powerful than “Um”

## NEXT STEPS

---

- ▶ Keep in touch!
- ▶ If you like data, checkout:
- ▶ Python for Data Analysis Workshop
- ▶ Intermediate Python Workshop
- ▶ Part-time Data Science
- ▶ Data Science Immersive



@josephofiowa



josephnelson@generalassemb.ly