GENERAL ASSEMBLY

STORYTELLING WITH DATA

Joseph Nelson, Data Science Immersive

AGENDA

- Introduction
- Math and stats
- Visual Design
- Communication Tips

AGENDA

> “All models are wrong. Some are useful.”
>
> — George Box, 1978

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

USDA

GW

YOU

- Your background:

- Name
- How do you currently or need to communicate with data? (Brevity counts)

- Rank these three:
- 1. Mathematics and statistics
- 2. Visual design
- 3. Communication strength

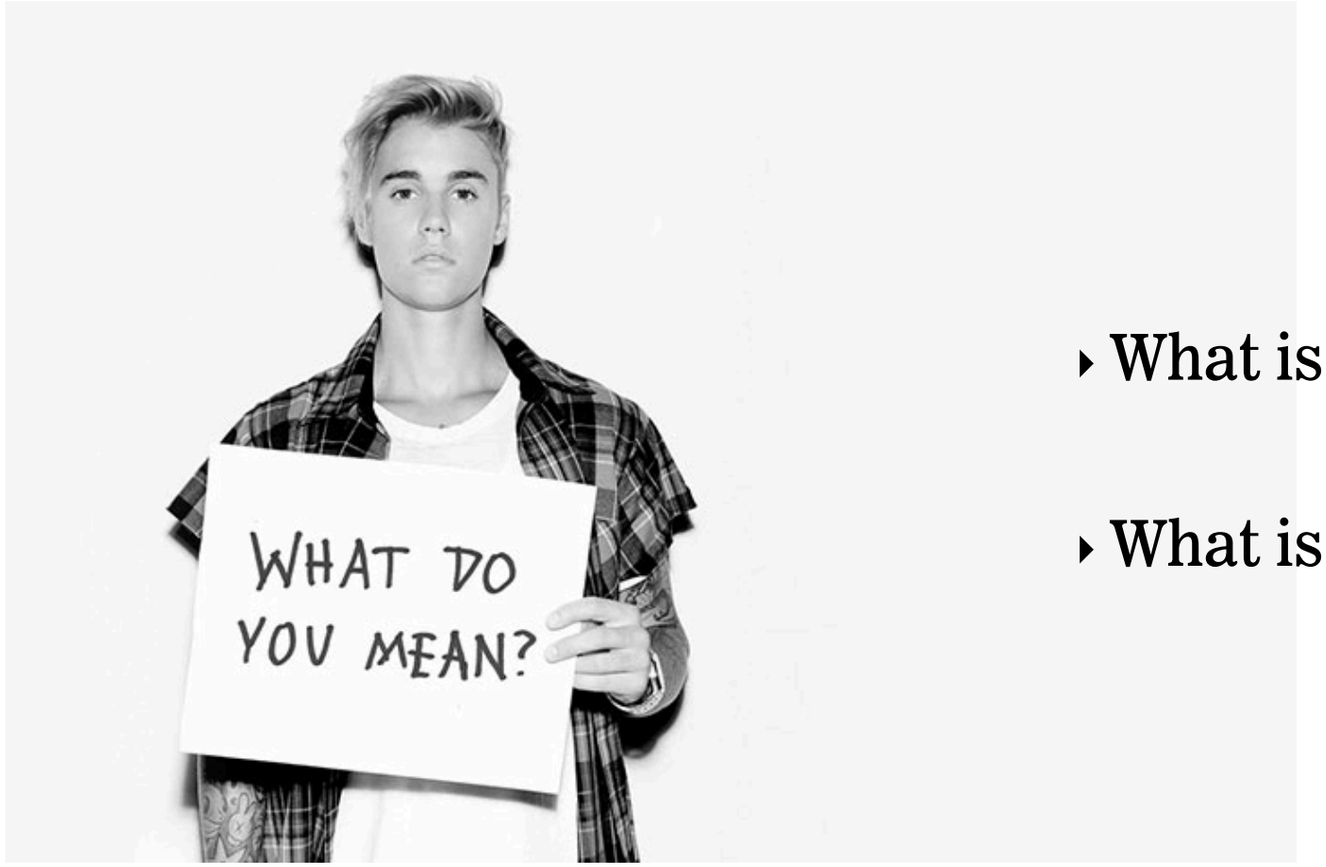
PART 1: MATH AND STATS REVIEW

PART 1: MATH AND STATS REVIEW

BASIC DESCRIPTIVE STATISTICS

- Mean
- Median
- Mode
- Max
- Min
- Quartile
- Inter-quartile Range
- Variance
- 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.

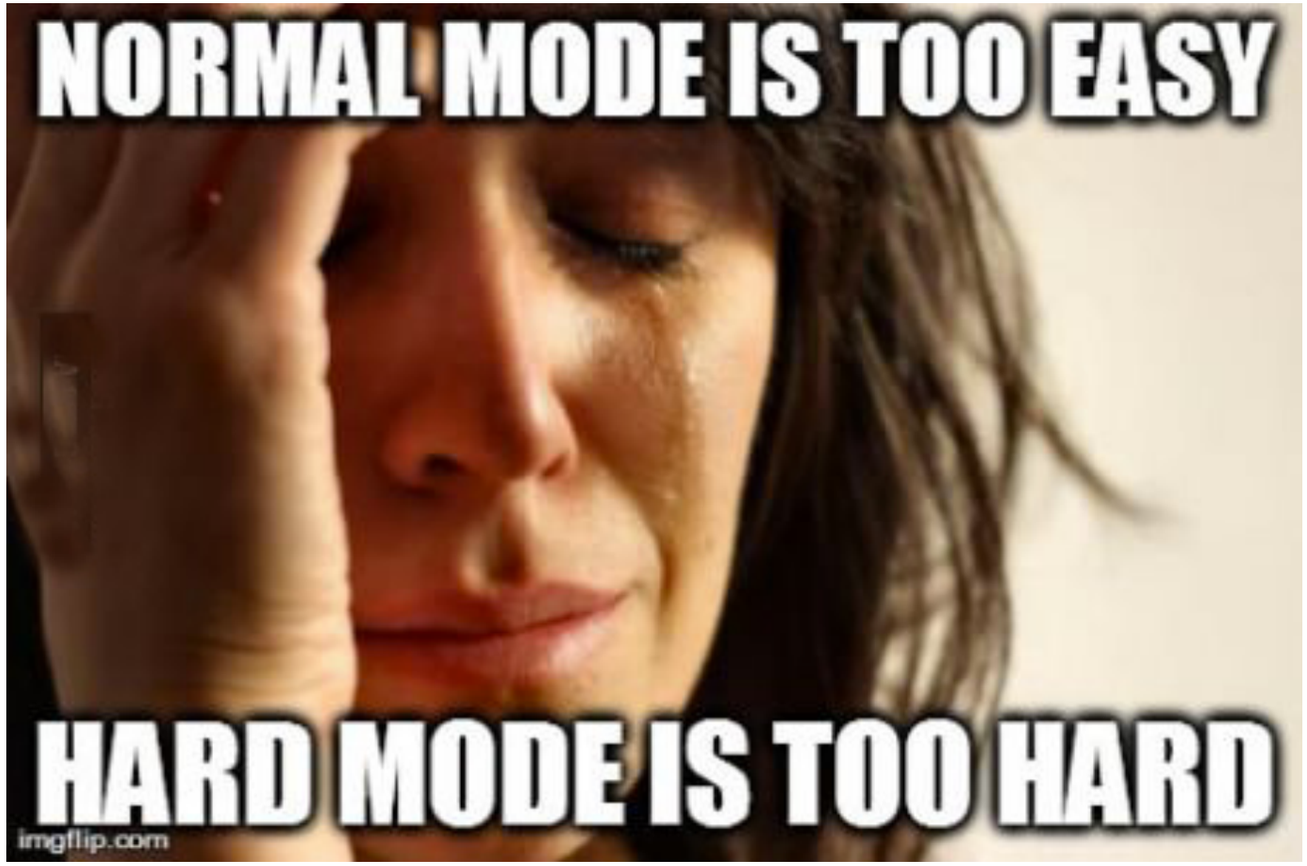
- To find the median, arrange the numbers in order from smallest to largest. If there is an odd number of values, the middle value is the median. If there is an even number of values, the average of the two middle values is the median.

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

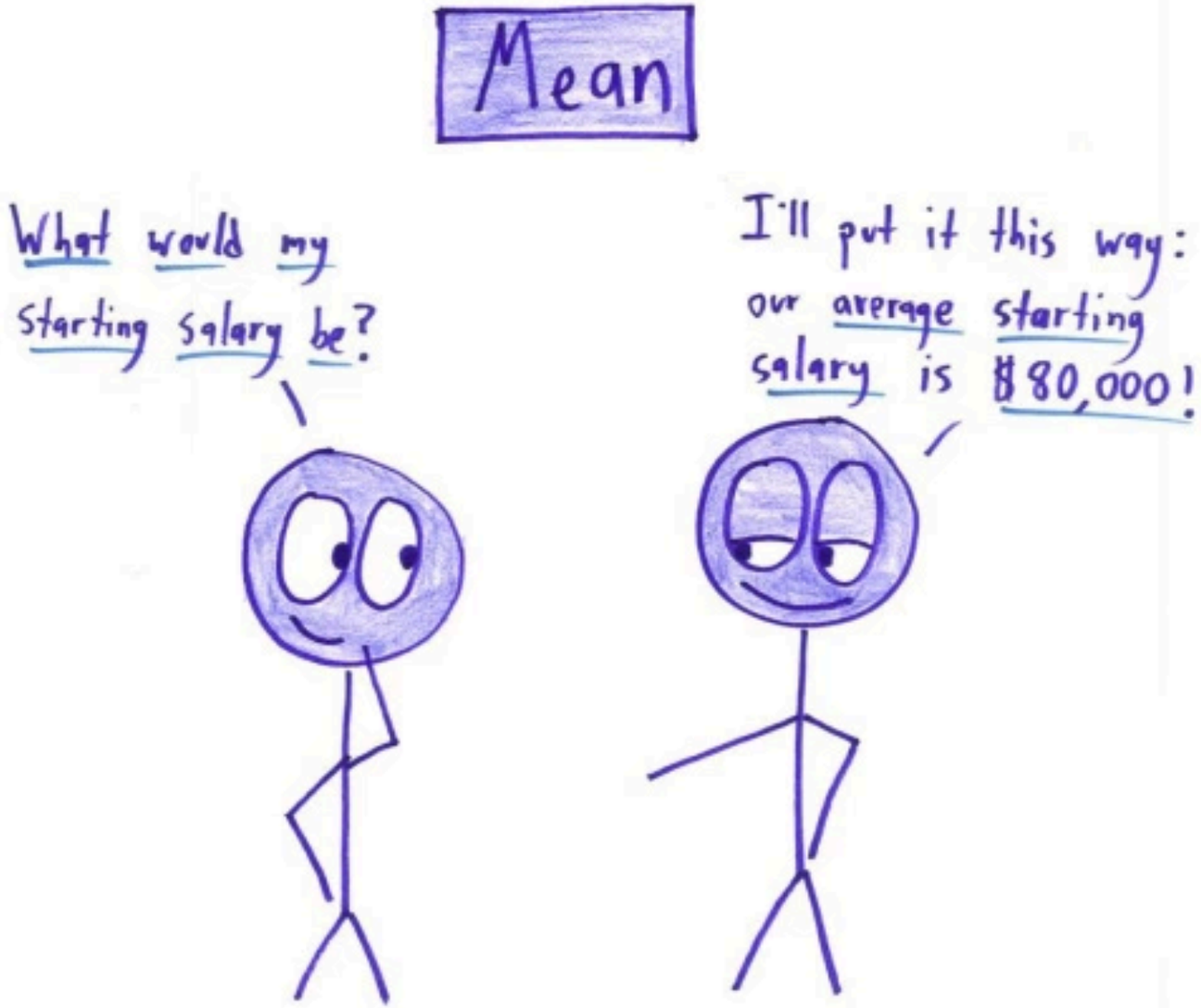
- C. Mean = 51.125 Median = 51 Mode = none Max = 66 Min = 38

HOW TO LIE WITH STATISTICS: EXERCISE 1

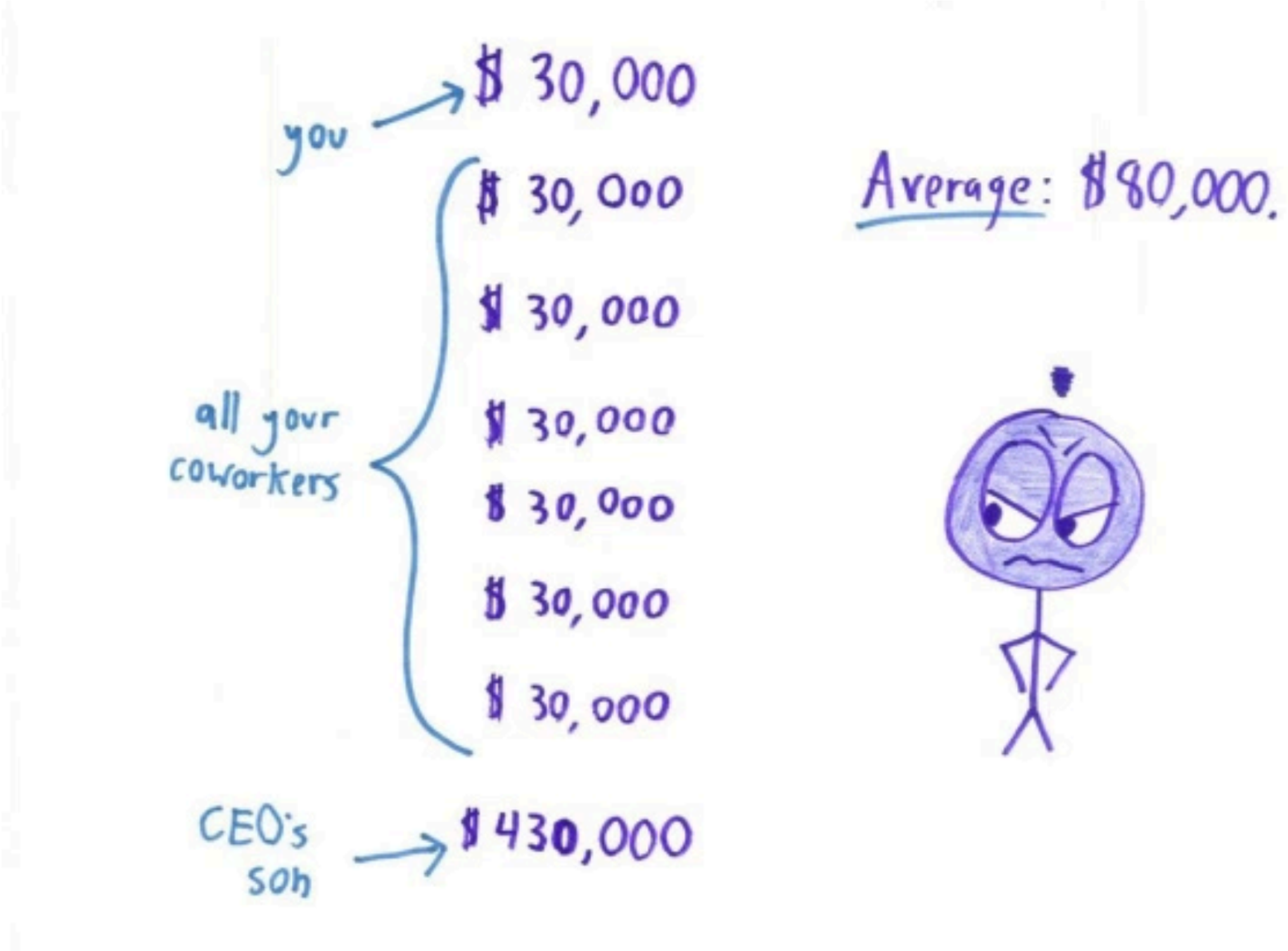
- For each picture:
- 1) What could go wrong
- 2) How to fix it

- Work in pairs!

HOW TO LIE WITH STATISTICS

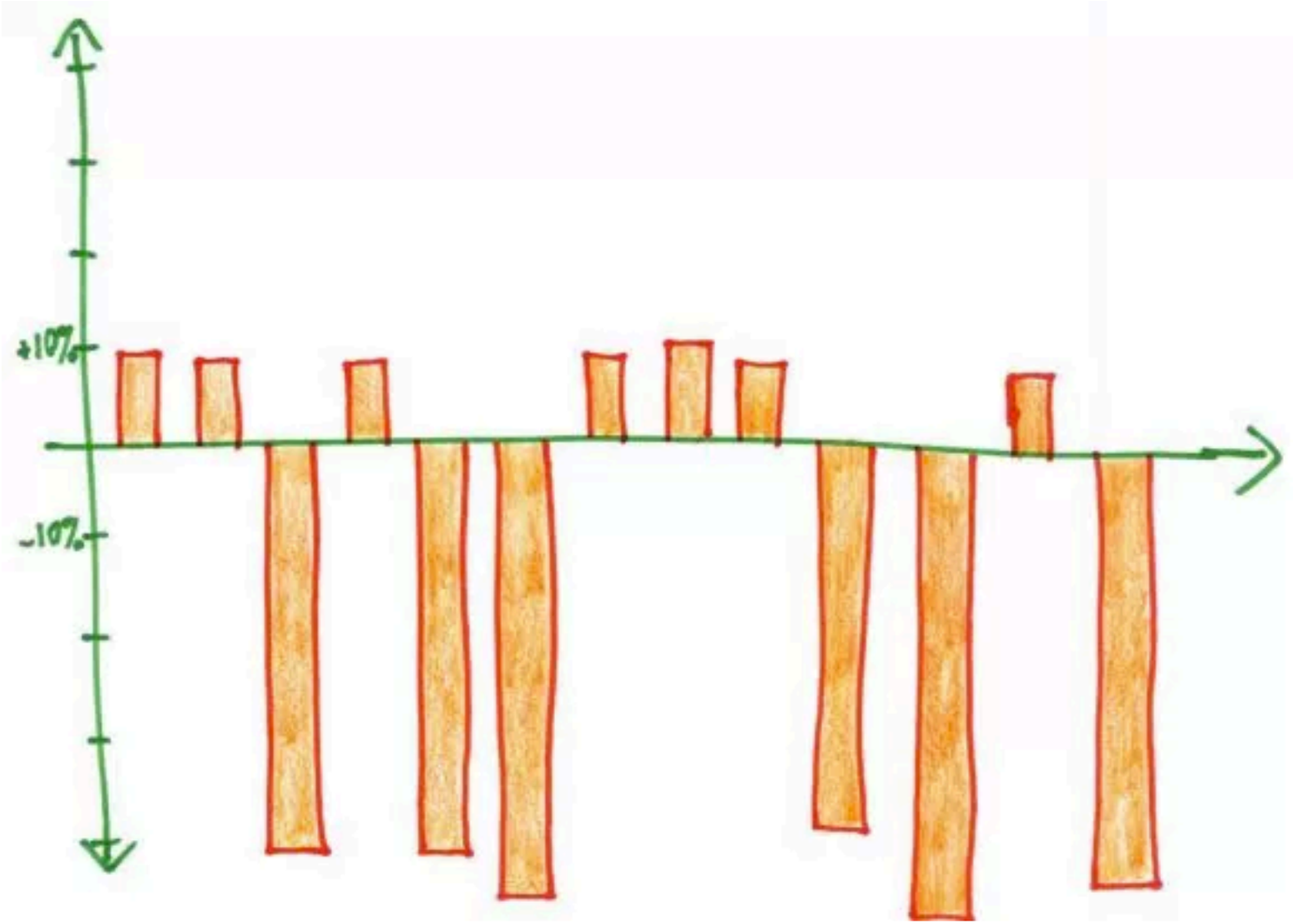


HOW TO LIE WITH STATISTICS

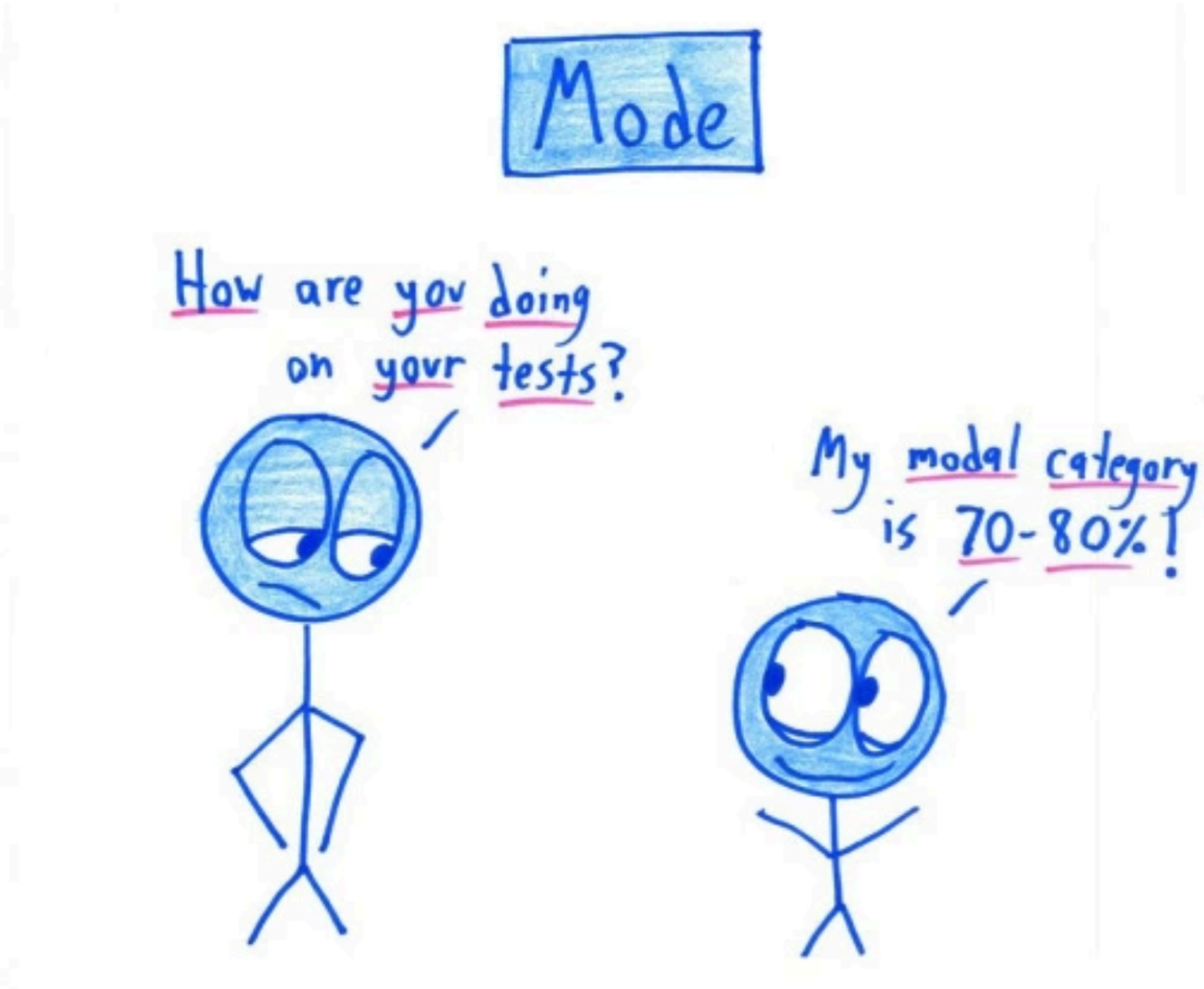


HOW TO LIE WITH STATISTICS

HOW TO LIE WITH STATISTICS



HOW TO LIE WITH STATISTICS

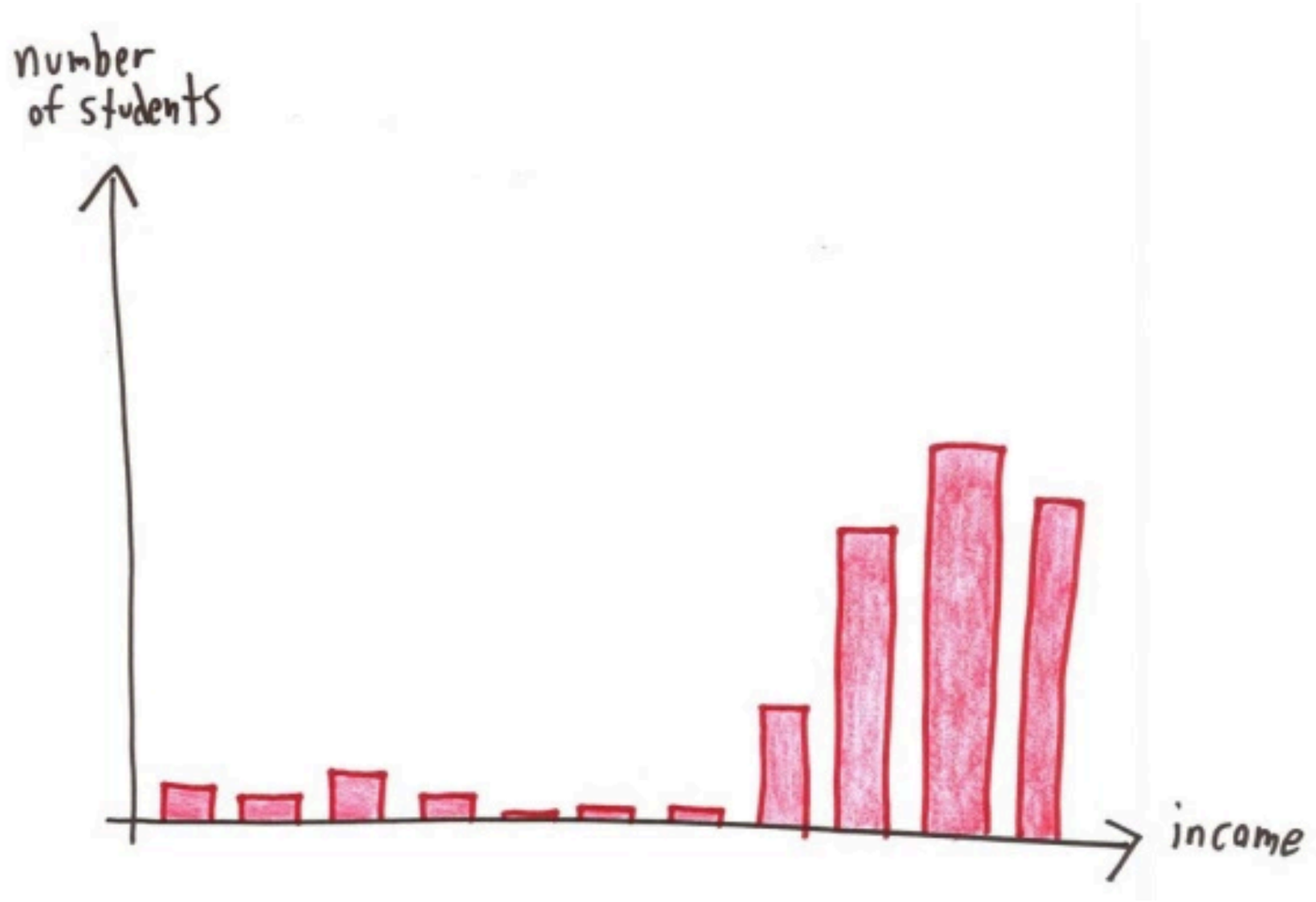


HOW TO LIE WITH STATISTICS

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

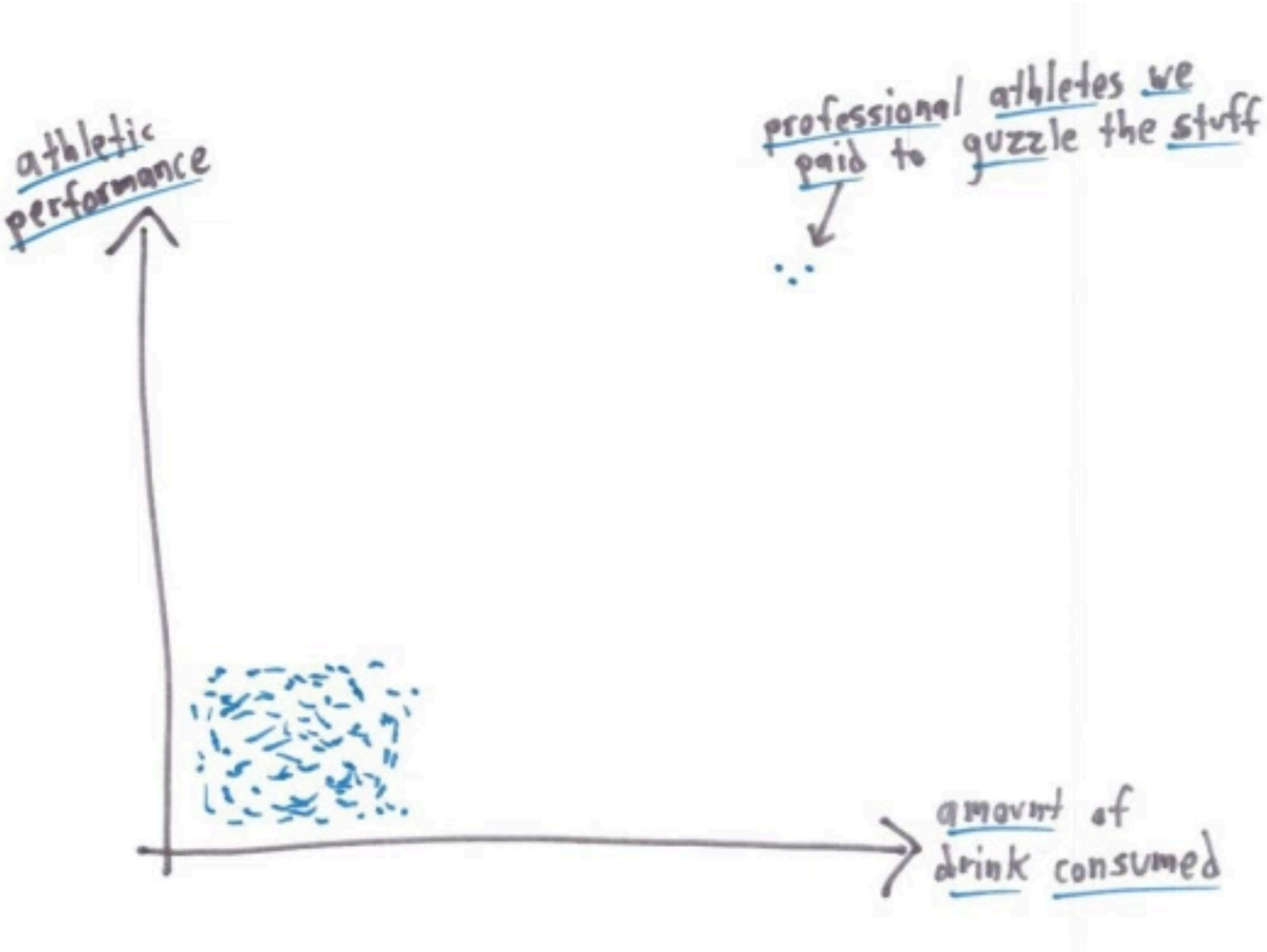
HOW TO LIE WITH STATISTICS

HOW TO LIE WITH STATISTICS

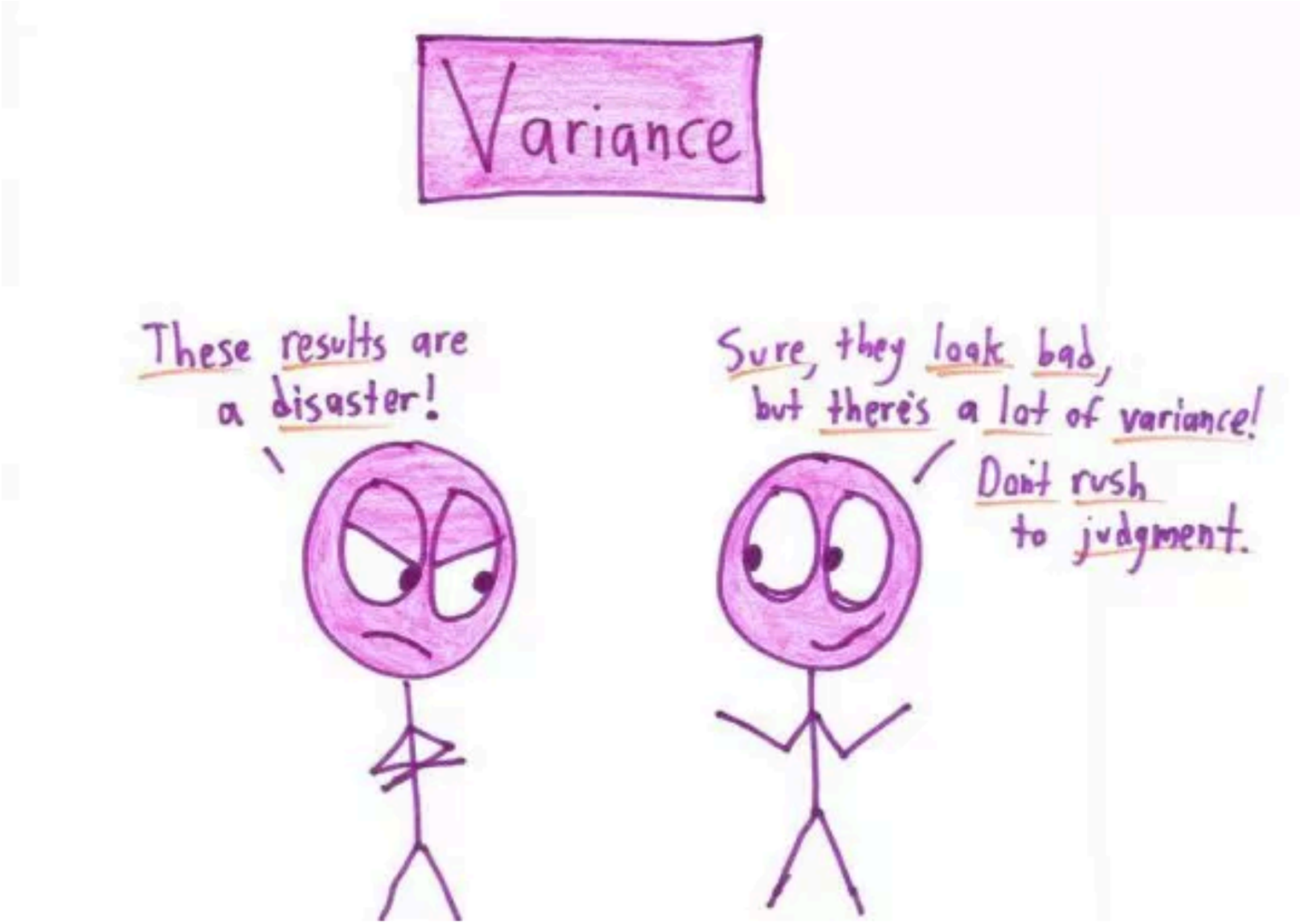


HOW TO LIE WITH STATISTICS

HOW TO LIE WITH STATISTICS



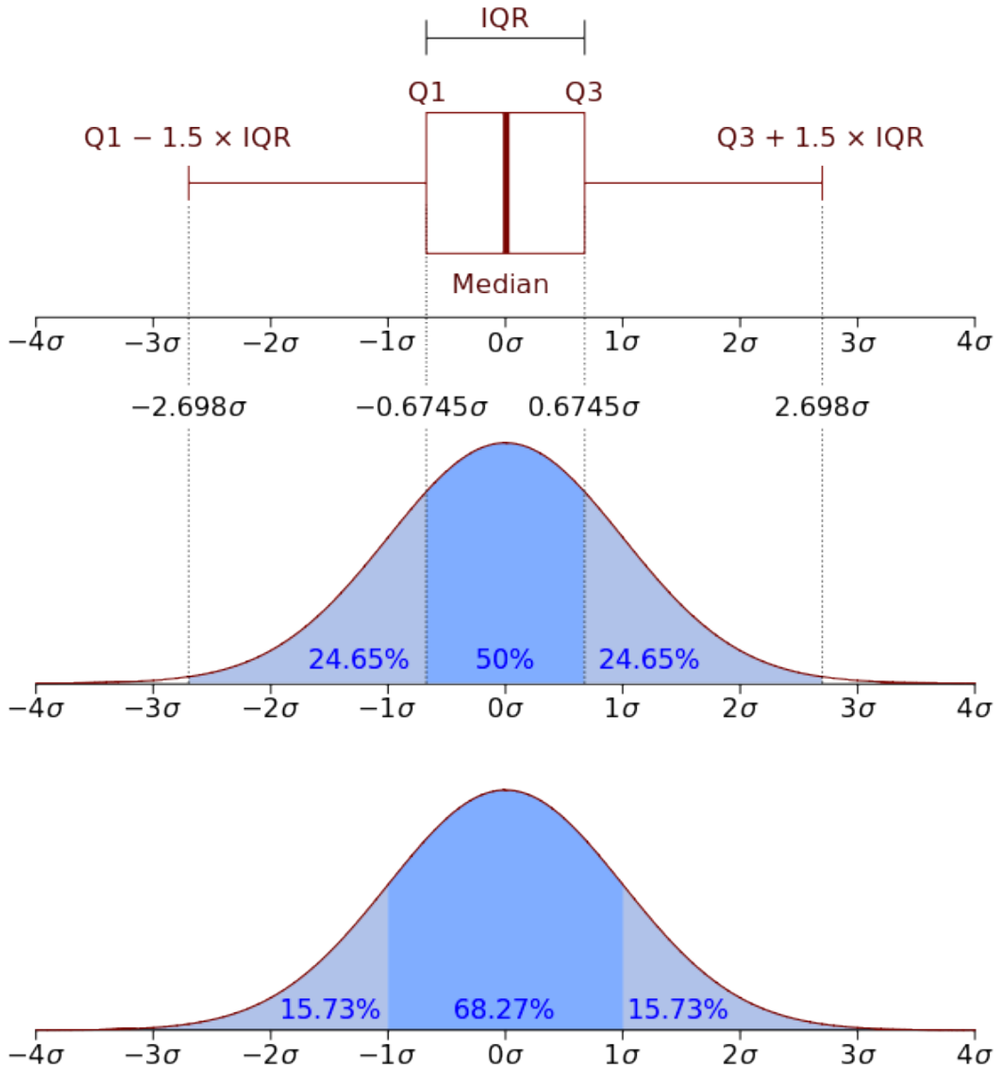
HOW TO LIE WITH STATISTICS



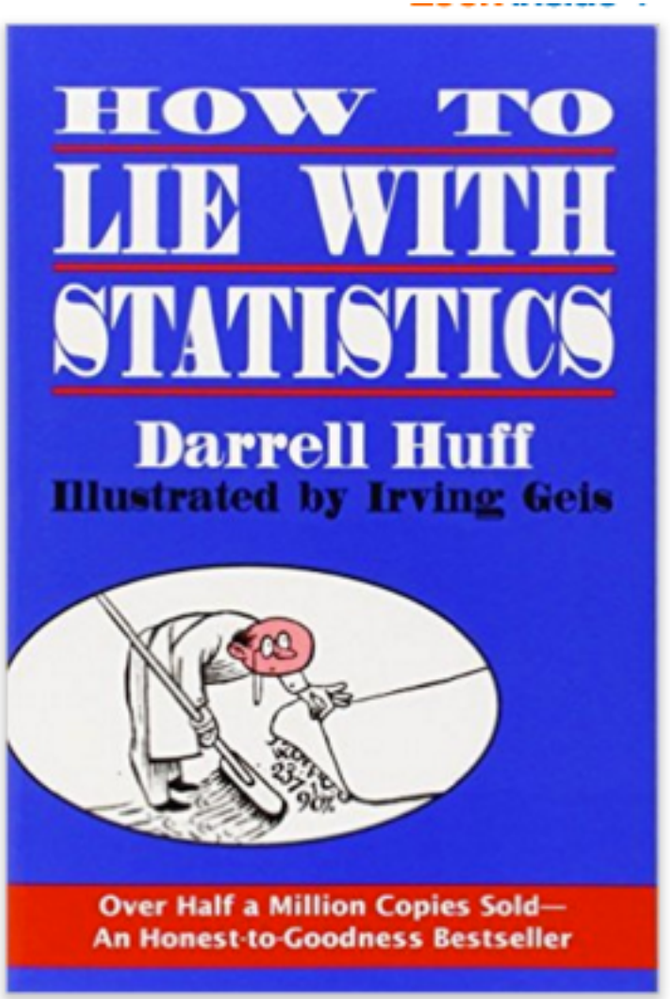
HOW TO LIE WITH STATISTICS

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 Q1, Q2, and Q3, 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 = Q3-Q1
- Outliers: Q1 - 1.5(IQR), Q3 + 1.5(IQR)



CRITERIA FOR GOOD VISUALIZATION



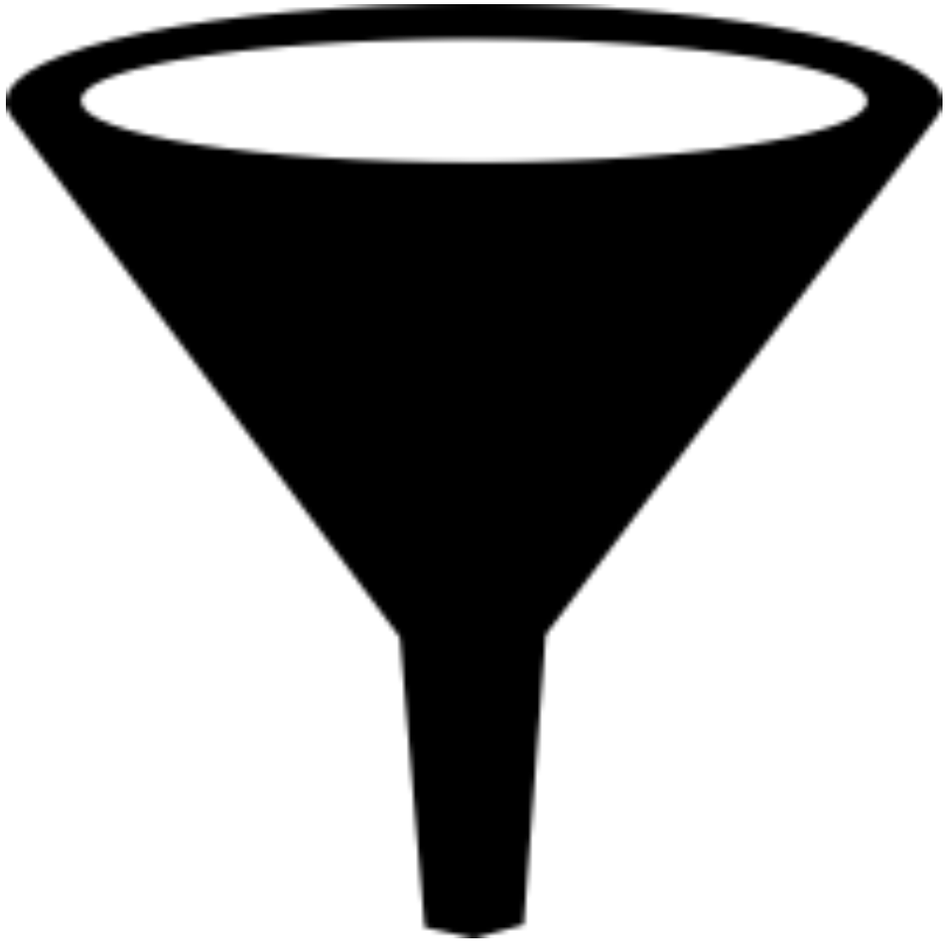
A CONVERSATION ON METRICS...

- Choosing a metric is (almost) everything.
- What are you optimizing?
- Why are you optimizing for that?
- Does that capture the full picture?

YOU RUN A MOBILE APPLICATION COMPANY CALLED SMACEBOOK

- What metric do you choose to prove your success?

You run a mobile application company called smacebook



- Site visit
- App download
- Profile creation
- Usage
- Referral

YOU RUN A MOBILE APPLICATION COMPANY CALLED SMACEBOOK

- What metric do you choose to prove your success?

YOU RUN A MOBILE APPLICATION COMPANY CALLED SMACEBOOK

- What metric do you choose to prove your success?
- Monthly active users (MAU) is a top choice.
- Facebook waited for Harvard University to demonstrate 50% of the campus to have DAILY active usage before launching to Stanford. This was controversial.
- Source: http://a16z.com/2016/07/16/network-effects-event/

A CONVERSATION ON METRICS...

- Another example

Date	Users		
1/1/16	3		
1/2/16	5		
1/3/16	9		
1/4/16	10		
1/5/16	17		
1/6/16	19		
1/7/16	22		
1/8/16	26		
1/9/16	30		

A CONVERSATION ON METRICS...

- Percentage growth is sensitive to small base rate changes.

Date	Users	Percent Growth	
1/1/16	3	0	
1/2/16	5	0.666666667	
1/3/16	9	0.8	
1/4/16	10	0.111111111	
1/5/16	17	0.7	
1/6/16	19	0.117647059	
1/7/16	22	0.157894737	
1/8/16	26	0.181818182	
1/9/16	30	0.153846154	

A CONVERSATION ON METRICS...

- Percent growth of percent growth tells an even different story

Date	Users	Percent Growth	Percent Growth Change
1/1/16	3	0	0
1/2/16	5	0.666666667	0
1/3/16	9	0.8	0.2
1/4/16	10	0.111111111	-0.861111111
1/5/16	17	0.7	5.3
1/6/16	19	0.117647059	-0.831932773
1/7/16	22	0.157894737	**0.342105263**
1/8/16	26	0.181818182	0.151515152
1/9/16	30	0.153846154	-0.153846154

YOU RUN A MOBILE APPLICATION COMPANY CALLED SMACEBOOK

- To summarize:
- Metrics are situational specific.
- We can tell different (and misleading) stories with each.
- Growth is new and retained. Be mindful.

PART 2: EFFECTIVE VISUAL DESIGN

CRITERIA FOR GOOD VISUALIZATION

- We'll break this section into parts:

- 1.) Examining best examples
- 2.) Distilling those examples
- 3.) Discussing tools to achieve those examples

CRITERIA FOR GOOD VISUALIZATION

- We'll break this section into parts:

- Read: http://fivethirtyeight.com/features/swing-voters-and-elastic-states/
- Read: http://fivethirtyeight.com/features/lionel-messi-is-impossible/
- Read: http://fivethirtyeight.com/features/lionel-messi-is-impossible/

CRITERIA FOR GOOD VISUALIZATION

- 1. Simplified
- 2. Easy to Interpret
- 3. Clearly Labeled

- Bonus: 4. Interactivity

CRITERIA FOR GOOD VISUALIZATION

- 1. Simplified
- 2. Easy to Interpret
- 3. 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

- 1. Simplified
- 2. Easy to Interpret
- 3. Clearly Labeled

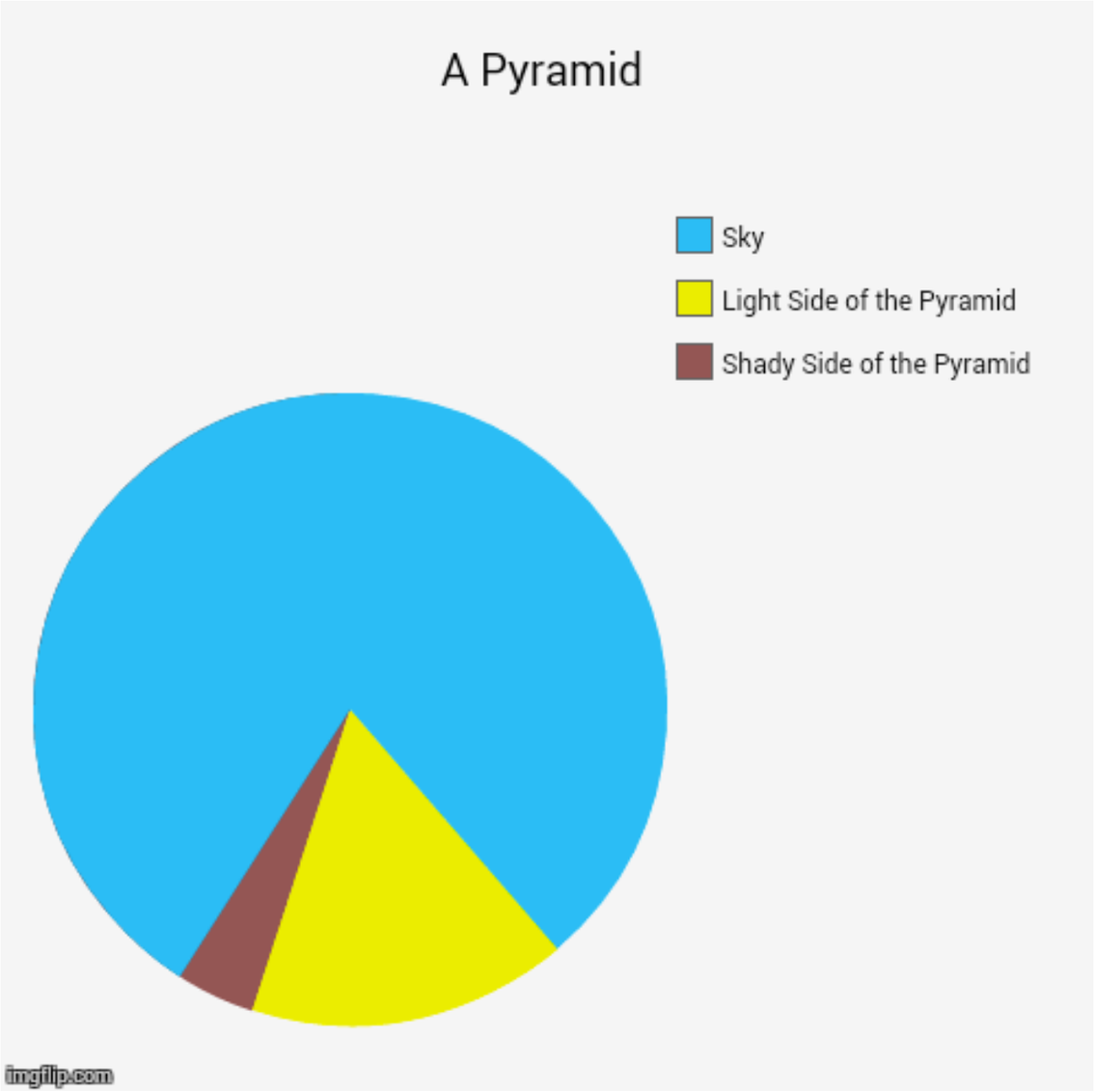
- Ask yourself:
- Who is my target audience?
- What do they already know, and what do they need to know?
- How does my project affect this audience? How might they interpret (or misinterpret) the data?

TOOLS

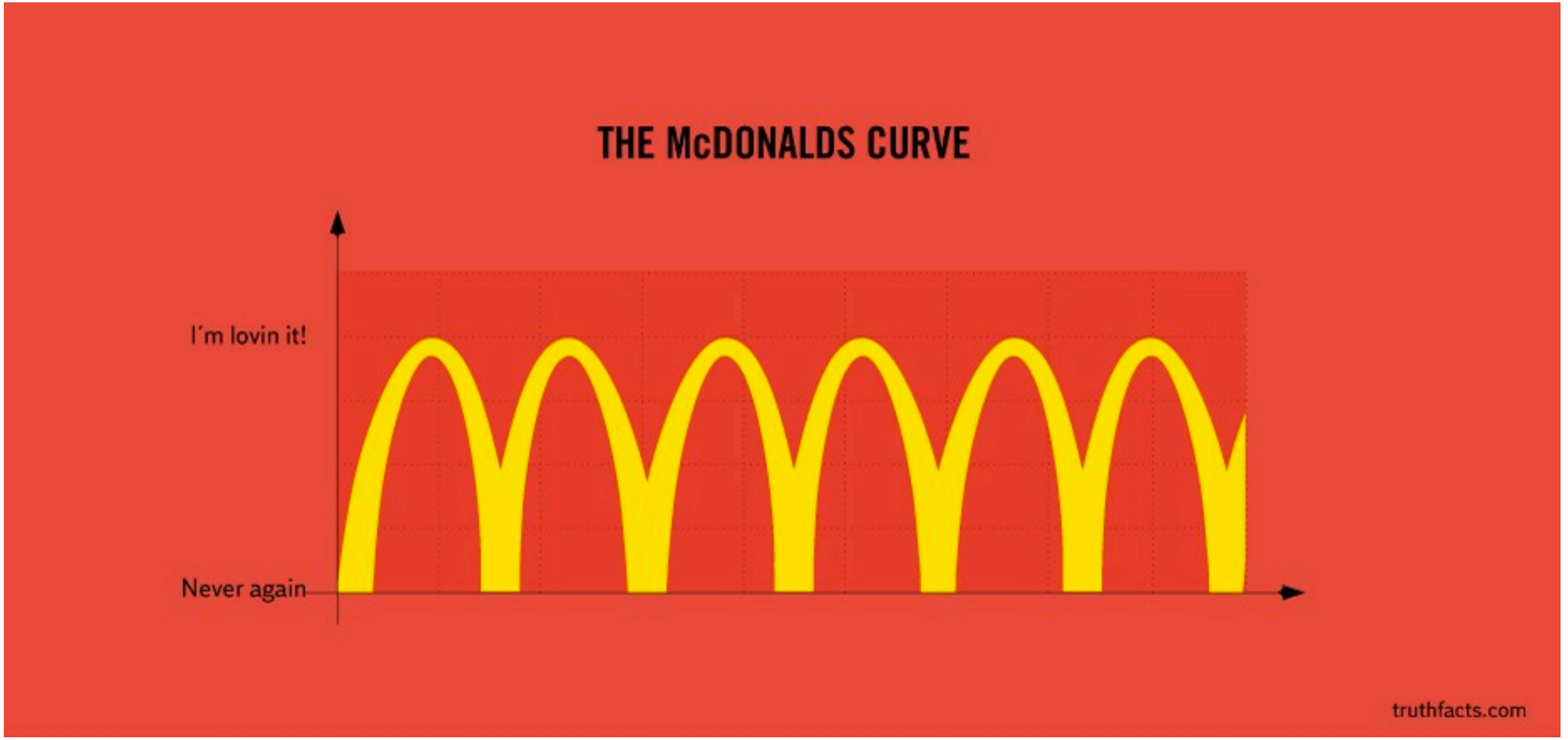
- D3.js: https://flowingdata.com/2015/12/15/a-day-in-the-life-of-americans/
- Tableau: https://public.tableau.com/en-us/s/blog/2015/07/analyzing-airbnb-data

- Excel
- Python (Bokeh)
- R

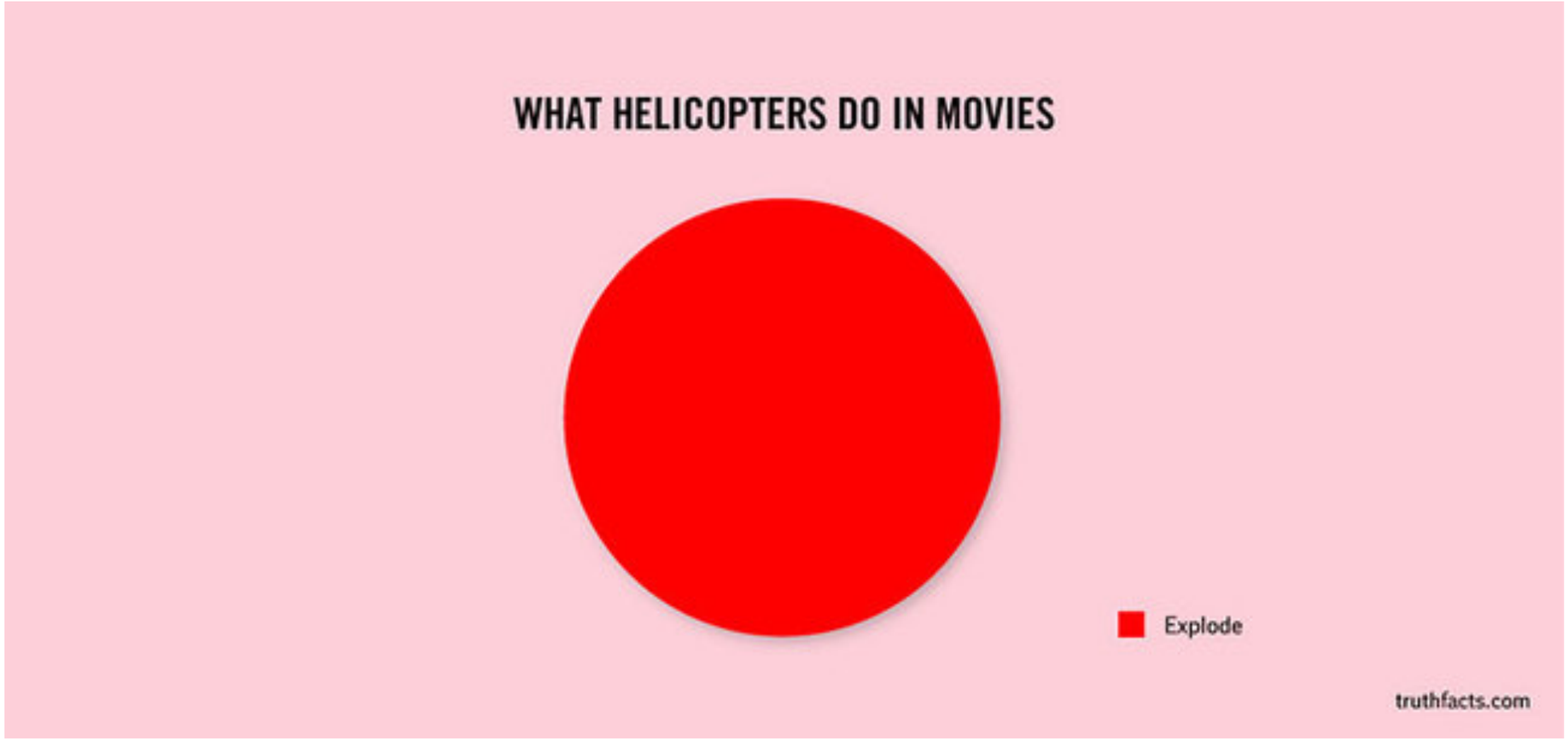
CHARTJUNK


A Pyramid
Sky
Light Side of the Pyramid
Shady Side of the Pyramid
imgflip.com

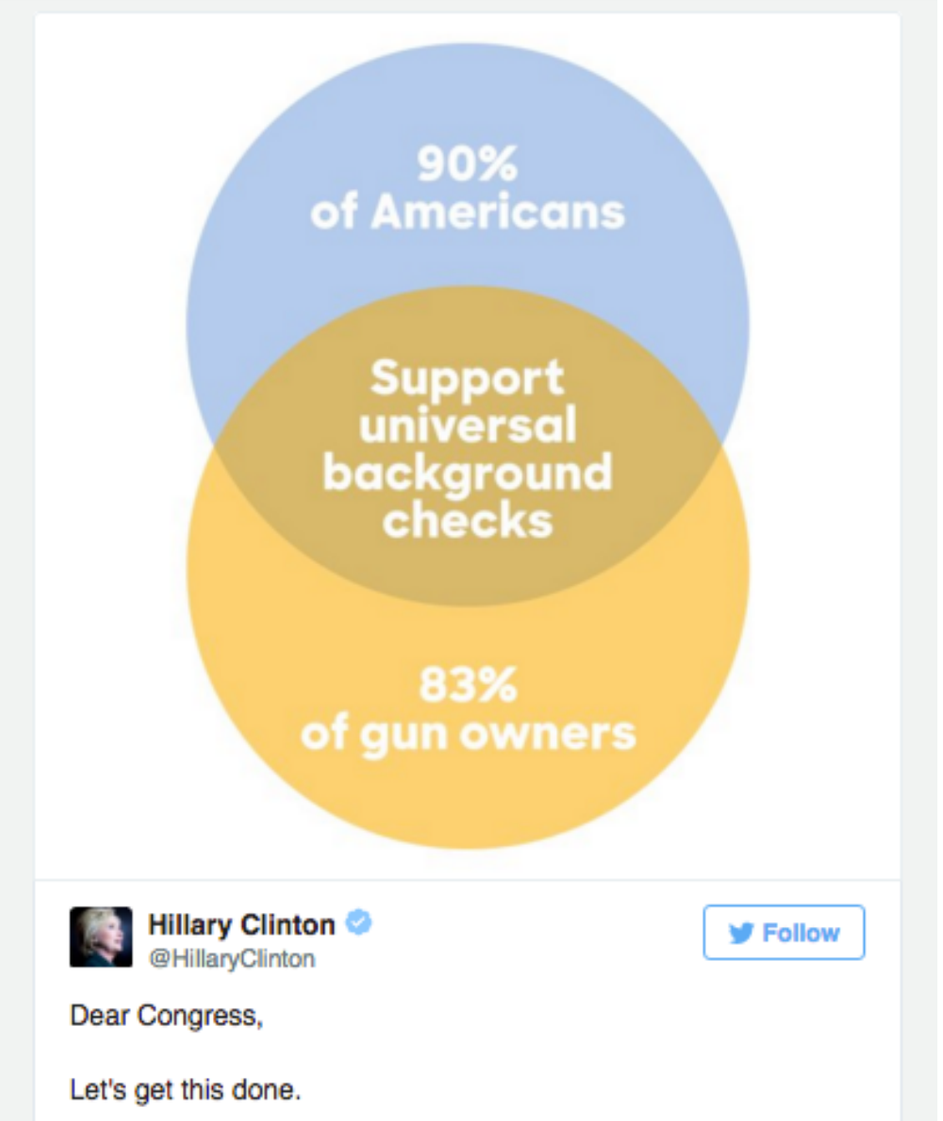
CHARTJUNK



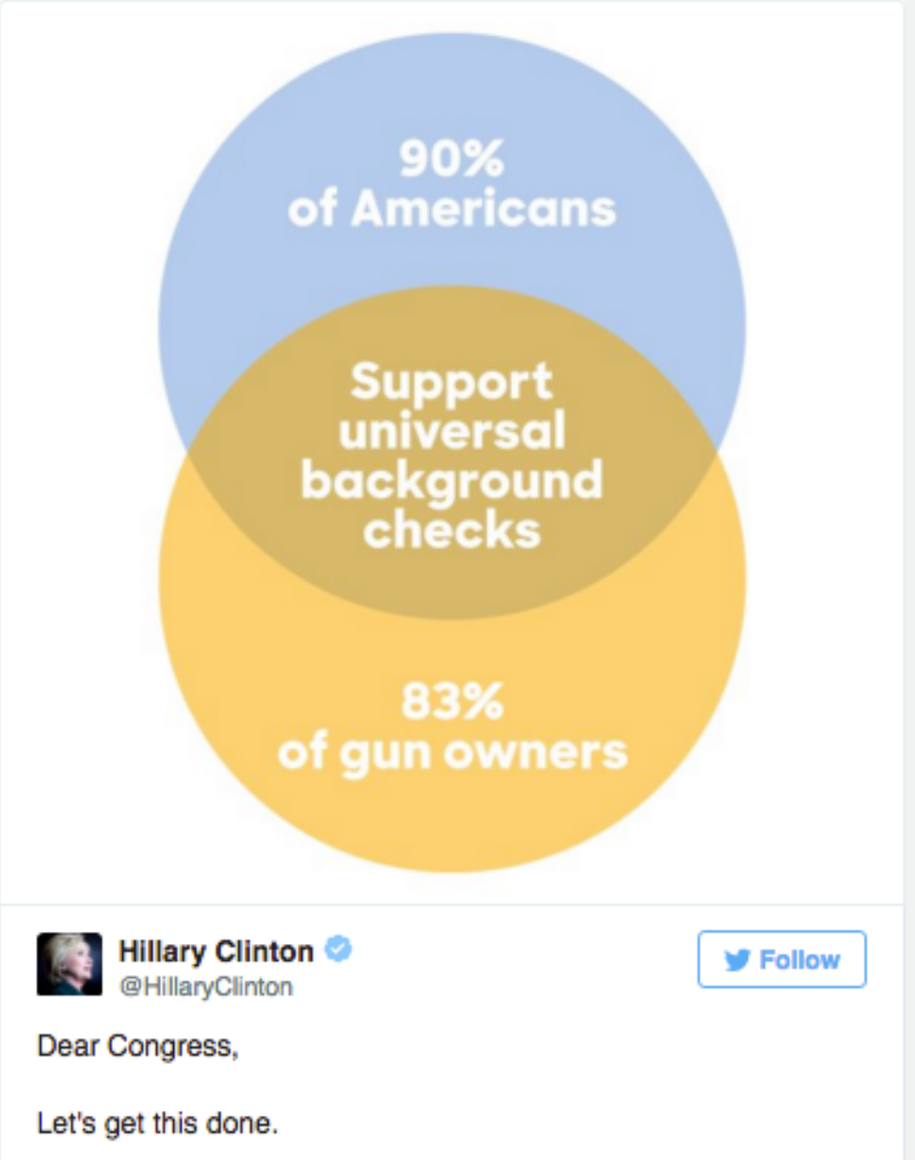
CHARTJUNK



GRAPHJUNK


90%
of Americans
Support
universal
background
checks
83%
of gun owners
Hillary Clinton
@HillaryClinton
Follow
Dear Congress,
Let's get this done.

GRAPHJUNK


90%
of Americans
Support
universal
background
checks
83%
of gun owners
Hillary Clinton
@HillaryClinton
Follow
Dear Congress,
Let's get this done.


People who know
how to make
Venn Diagrams
Hillary's graphic
design staff
In reply to Hillary Clinton
Michael Deppisch @deppisch · May 20
.@hillaryclinton
3.7K
5.4K

IT'S A BIPARTISAN PROBLEM

PART 3: EFFECTIVE COMMUNICATION

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!

@josephofiowa

josephnelson@generalassemb.ly

- If you like data, checkout:
- Python for Data Analysis Workshop
- Intermediate Python Workshop
- Part-time Data Science
- Data Science Immersive