

AP Statistics



Homework

Read Chpt 1

Do pg 10 8, 10, 14, 17, 23



Chapter 1

What is "Statistics"



There is no knowledge without statistics

- Statistics may be the least popular subject in high school and college and are rarely selected as electives, but ...
- The probability is high that you will be required to take statistics in college.
- Statistics is life explained. Everything truly known in your life is known via statistics.
- Lies, damned lies, and statistics. People often believe statistics can be manipulated to suggest any conclusion. That is not reality. Statistics tell us what they tell us. Some use the general lack of sophistication about statistics to suggest misleading conclusions.

What Is (Are?) Statistics?

- Statistics (the discipline) is a way of reasoning, a collection of tools and methods, designed to allow us to understand our world.
- Data are values **with context**.
- Statistics (plural) are particular calculations derived from data, and the conclusions that are drawn from those calculations.

This is how I view life and statistics.

Statistics is the study of variability.

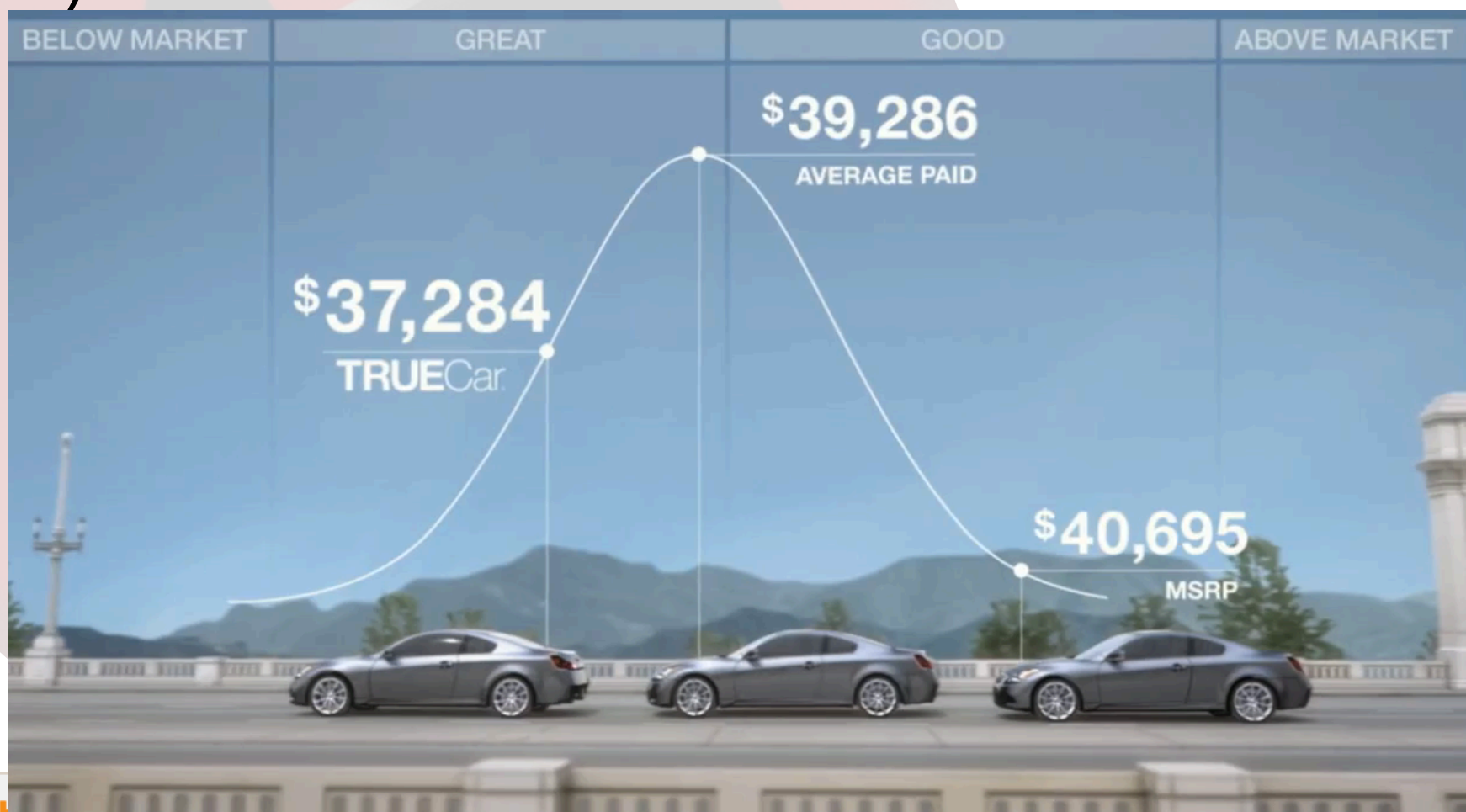
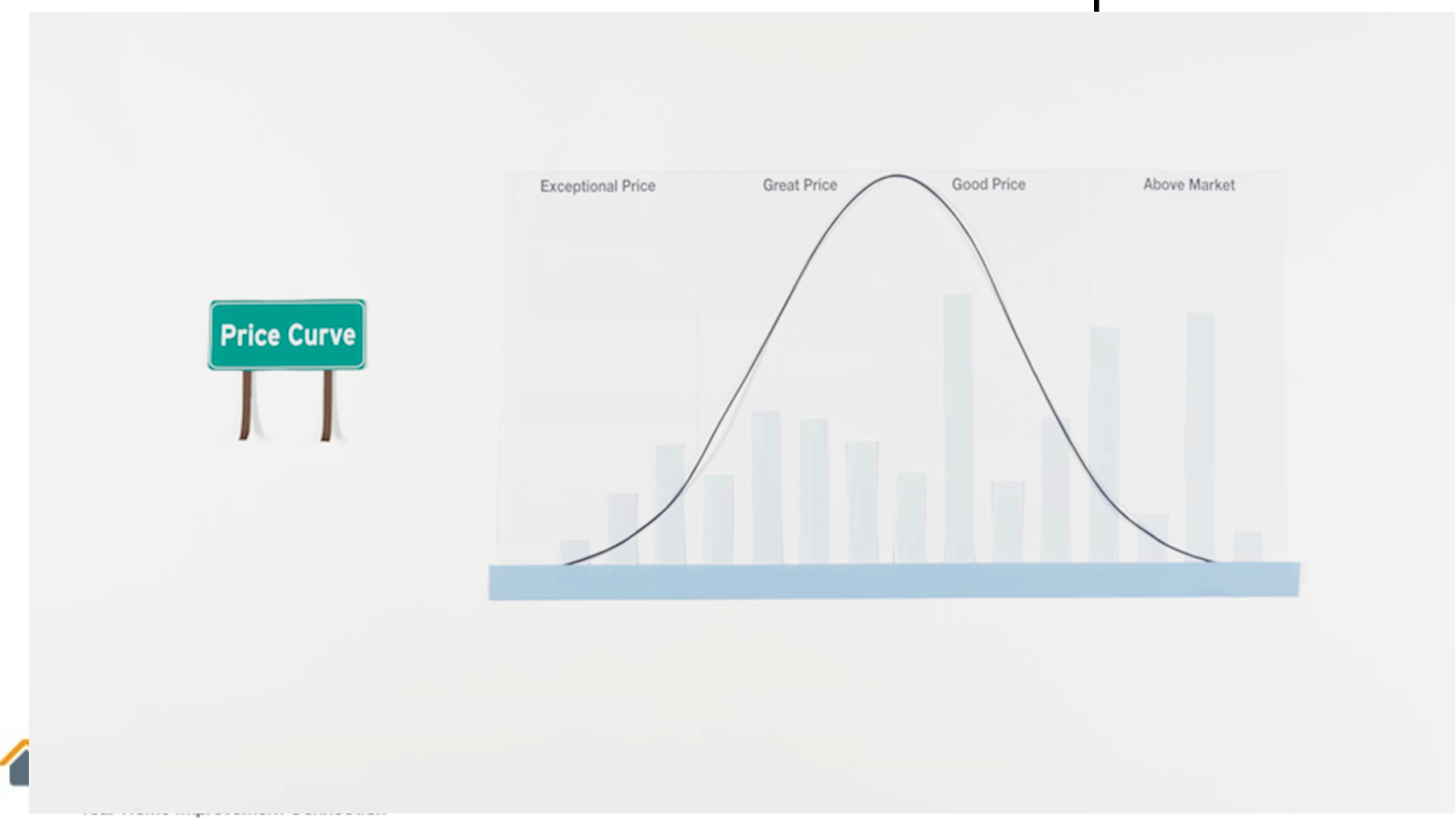
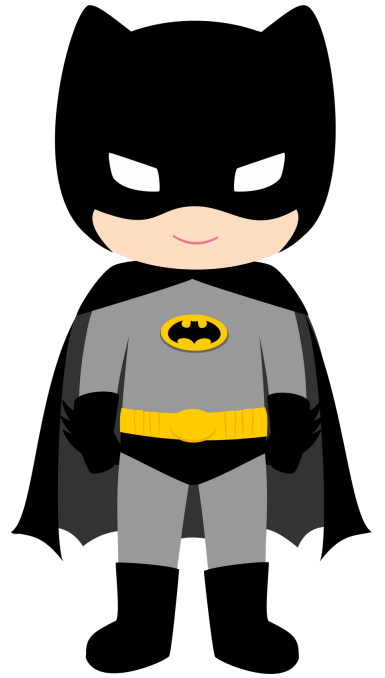
What is Statistics Really About?

- Statistics is about variation.
- All measurements are imperfect since those measurement have variation that is unavoidable or not immediately known to us. You are not the **exact** same height and weight every single day.
- Statistics is our best attempt to understand the real, imperfect, **varying, diverse** world in which we live.

Why Study Statistics

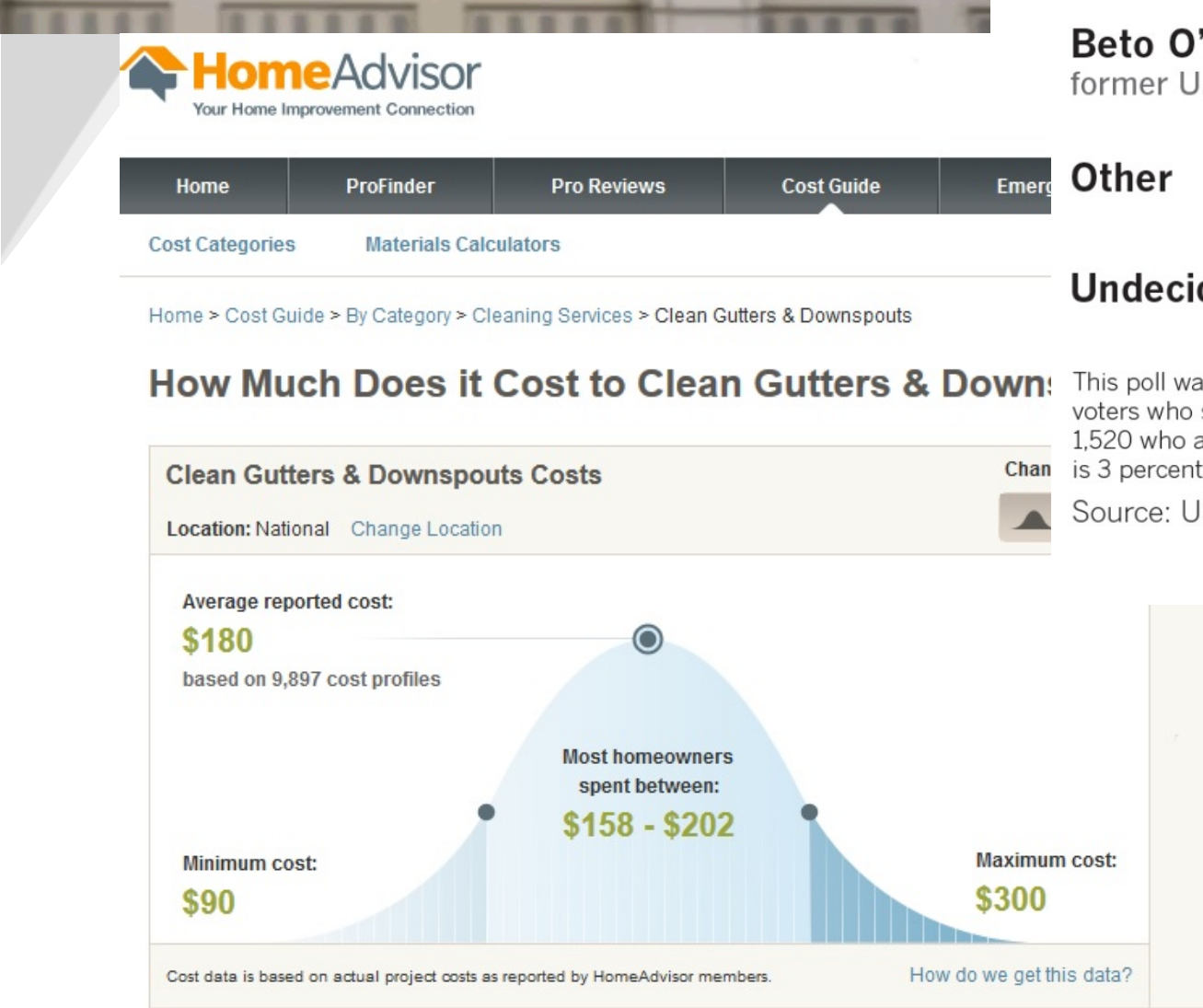
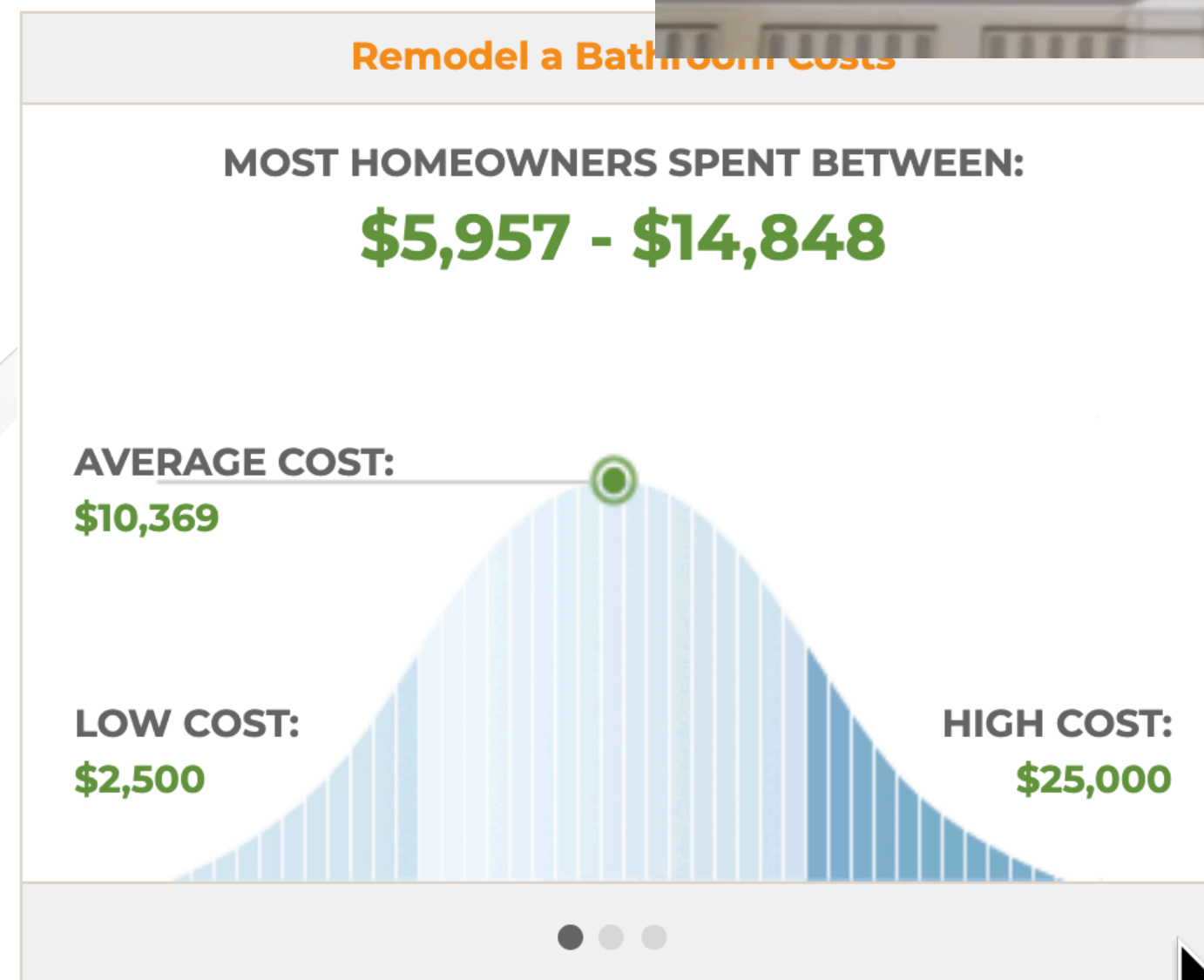
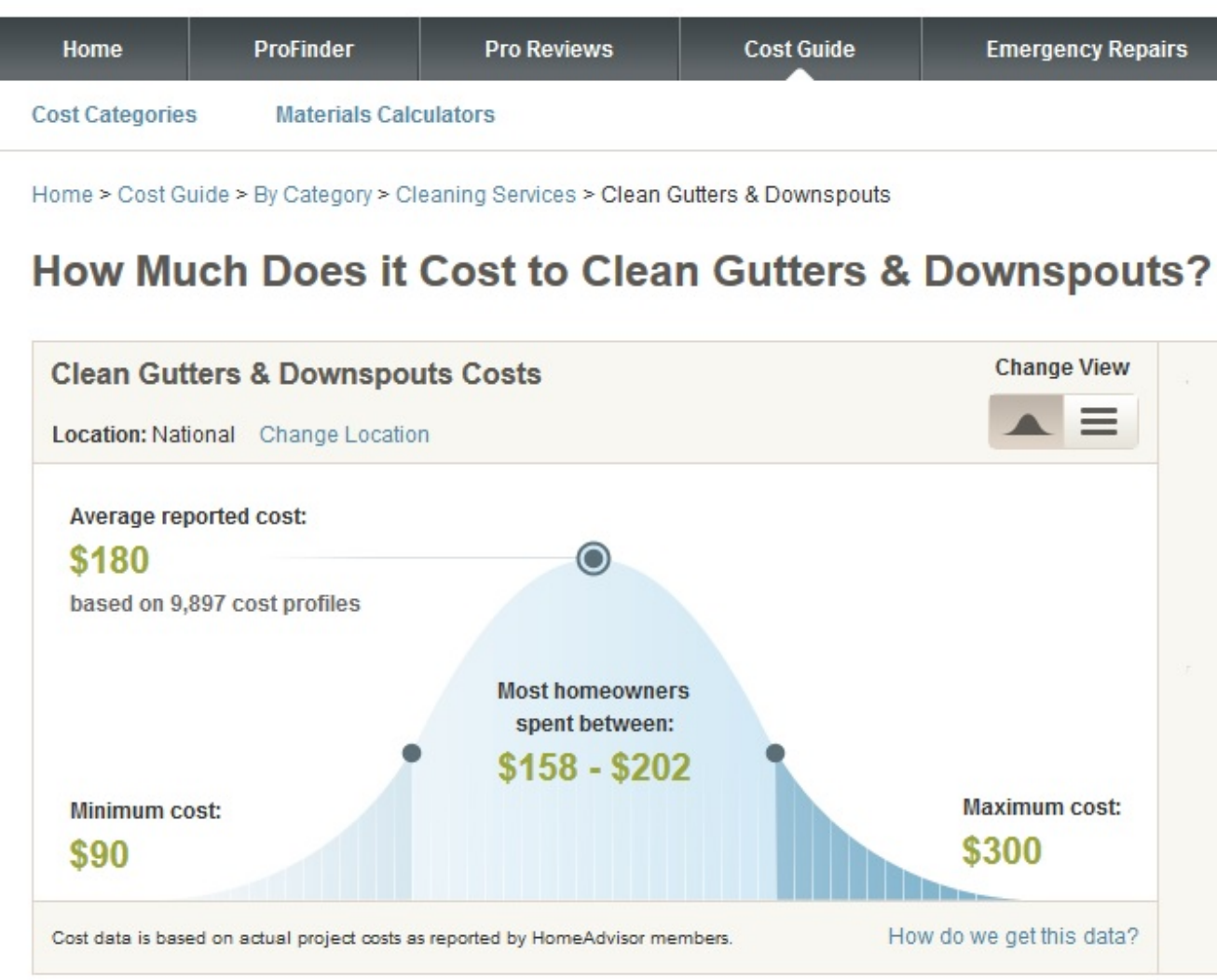
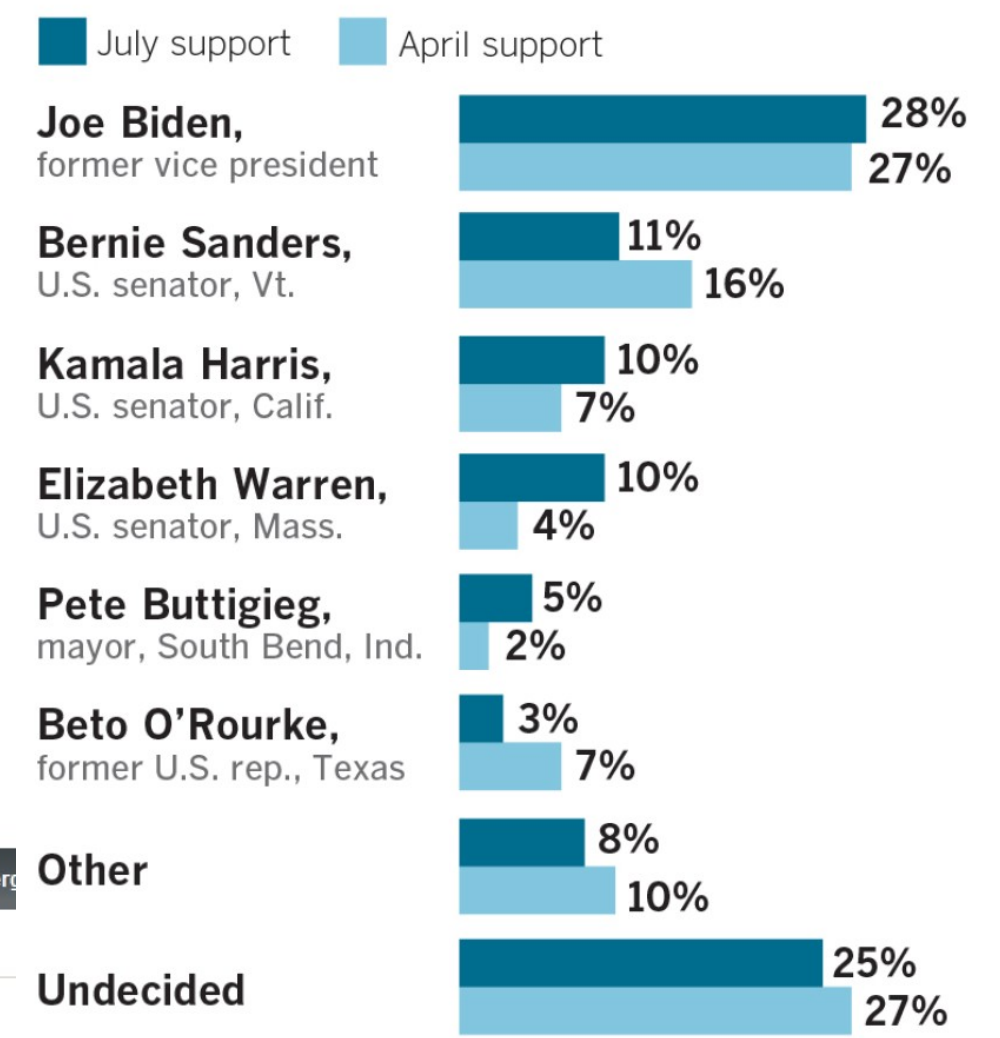


Even though you are unlikely to pursue statistics as a field of study, you will be exposed to statistics throughout your life. After this class you will recognize and be better able to interpret the statistics you see.



Biden in front while Sanders, Harris and Warren are close second, poll says

Since April, Sens. Kamala Harris and Elizabeth Warren have gained ground, while Sen. Bernie Sanders and former Rep. Beto O'Rourke have faded. Only six candidates received more than 1% in the poll.



This poll was conducted from July 12 to 25 from a national sample of 1,827 voters who said they expect to cast ballots in a Democratic primary, including 1,520 who also participated in the April survey. The margin of sampling error is 3 percentage points in either direction.

Source: USC Dornsife/Los Angeles Times national poll
Chris Keller / Los Angeles Times

Think, Show, Tell

🦇 Your book suggests there are three not so simple steps to doing Statistics right:

Think first. Know where you're headed and why.

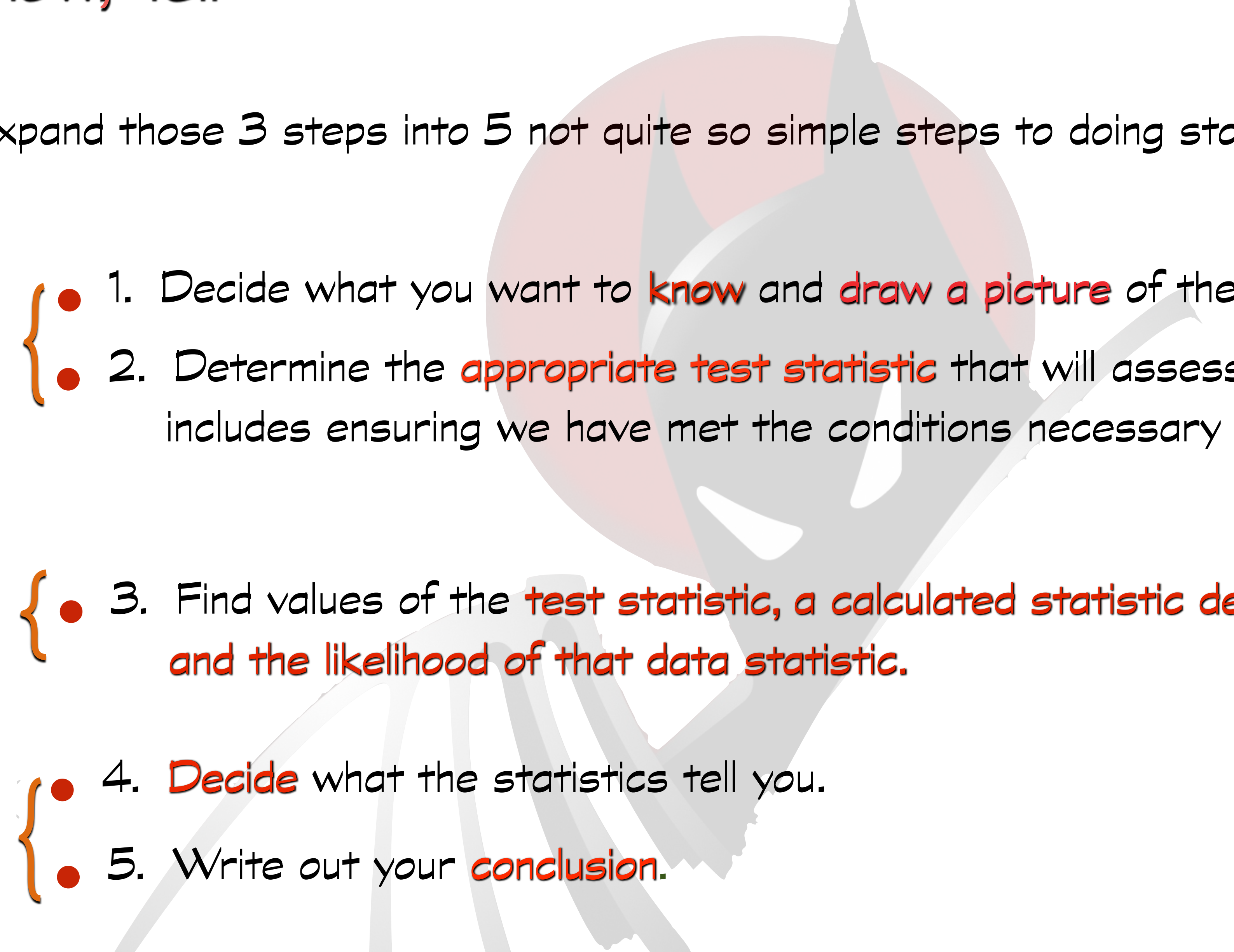
Show is about the mechanics of calculating statistics and making graphical displays of data.

Tell what you've learned. You must explain your results so that someone else (like your Nana from the old country) can understand your conclusions.



Think, Show, Tell

🦇 I will expand those 3 steps into 5 not quite so simple steps to doing statistics:

- 
- Think* {
- 1. Decide what you want to **know** and **draw a picture** of the data.
 - 2. Determine the **appropriate test statistic** that will assess the data. That includes ensuring we have met the conditions necessary for that test statistic.
- Show* {
- 3. Find values of the **test statistic, a calculated statistic determined by the data, and the likelihood of that data statistic.**
- Tell* {
- 4. **Decide** what the statistics tell you.
 - 5. Write out your **conclusion.**

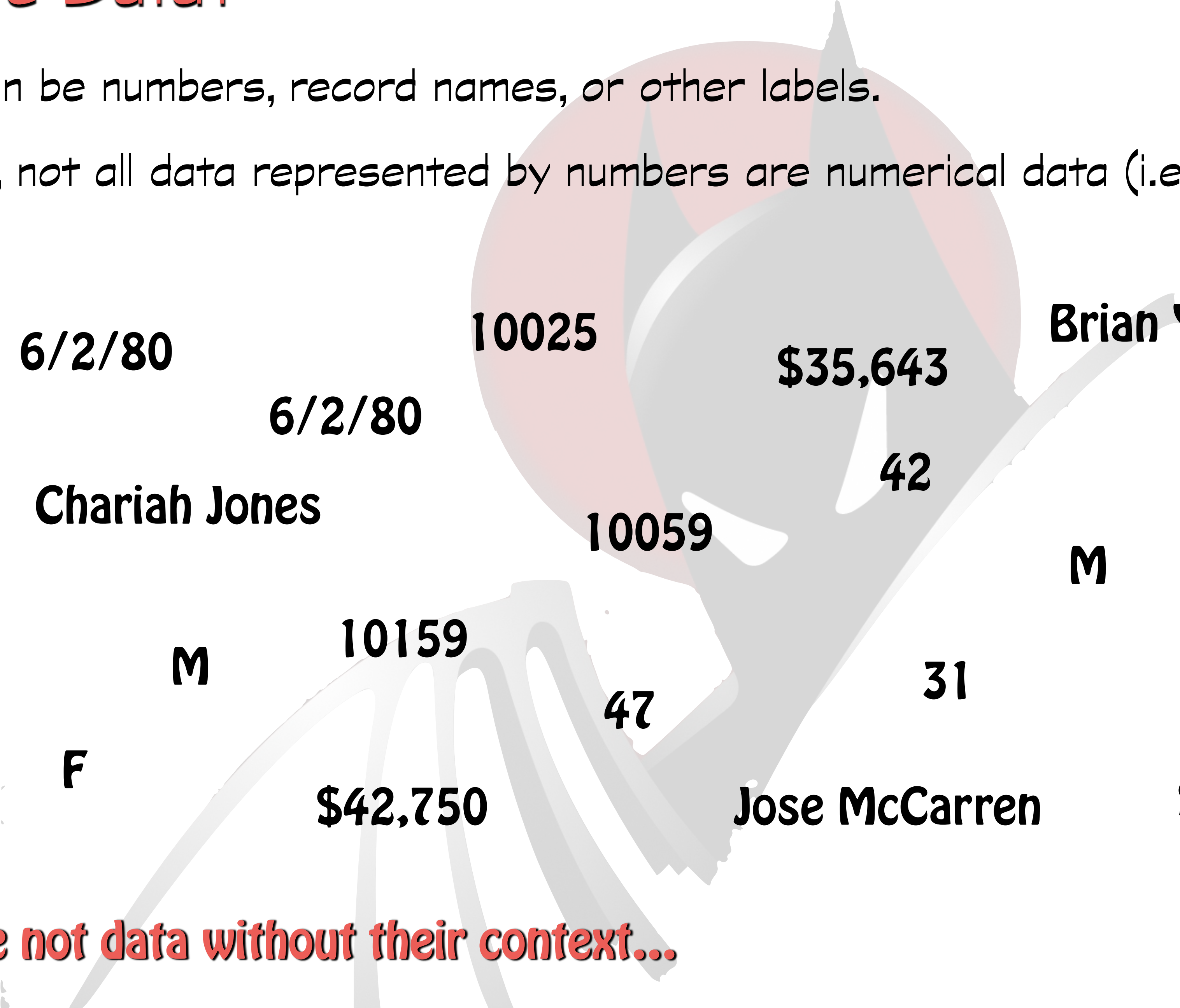
So let us get started

- 🦇 As you can see, statistics is much more than simple calculations
- 🦇 You will be asked to do a lot of writing for every question you are required to answer.
- 🦇 Statistics starts with **data**. So we start with a discussion of **data**.

■ *Welcome to AP Statistics*

What Are Data?

- 🦇 Data can be numbers, record names, or other labels.
- 🦇 Careful, not all data represented by numbers are numerical data (i.e., 1 = male, 2 = female).



6/2/80	10025	\$35,643	Brian Yu	4/16/86
6/2/80	10059	42		10/13/92
Chariah Jones			M	
M	10159			10101
F	\$42,750	47	31	\$54,875
		Jose McCarren		

Data are not data without their context...

Data

... howsomever, if we organize the information into a table, we can make sense of that information.

Employees of StatCo					
Name	Hire Date	PIN	Gender	Age	Salary
Brian Yu	6/2/80	10025			\$35,643
Chariah Jones		10059	M	42	10/13/92
	M			31	10101
	F	47			\$42,750
Jose McCarren					\$54,875

Now we have data!

Context and the Ws

- 🦇 If I ask you to “give me five” do you know what I mean?
- 🦇 Context gives us information that helps us understand the data.
- 🦇 The context is provided by the **W's**

- 🦇 **Who**

- 🦇 **What** (and in what units)

- 🦇 **When**

- 🦇 **Where**

- 🦇 **Why** (if possible)

- 🦇 and **How**

- 🦇 ... of the data.

🦇 Note: the answers to “who” and “what” are essential, but the most important question is most likely “why”.



Who

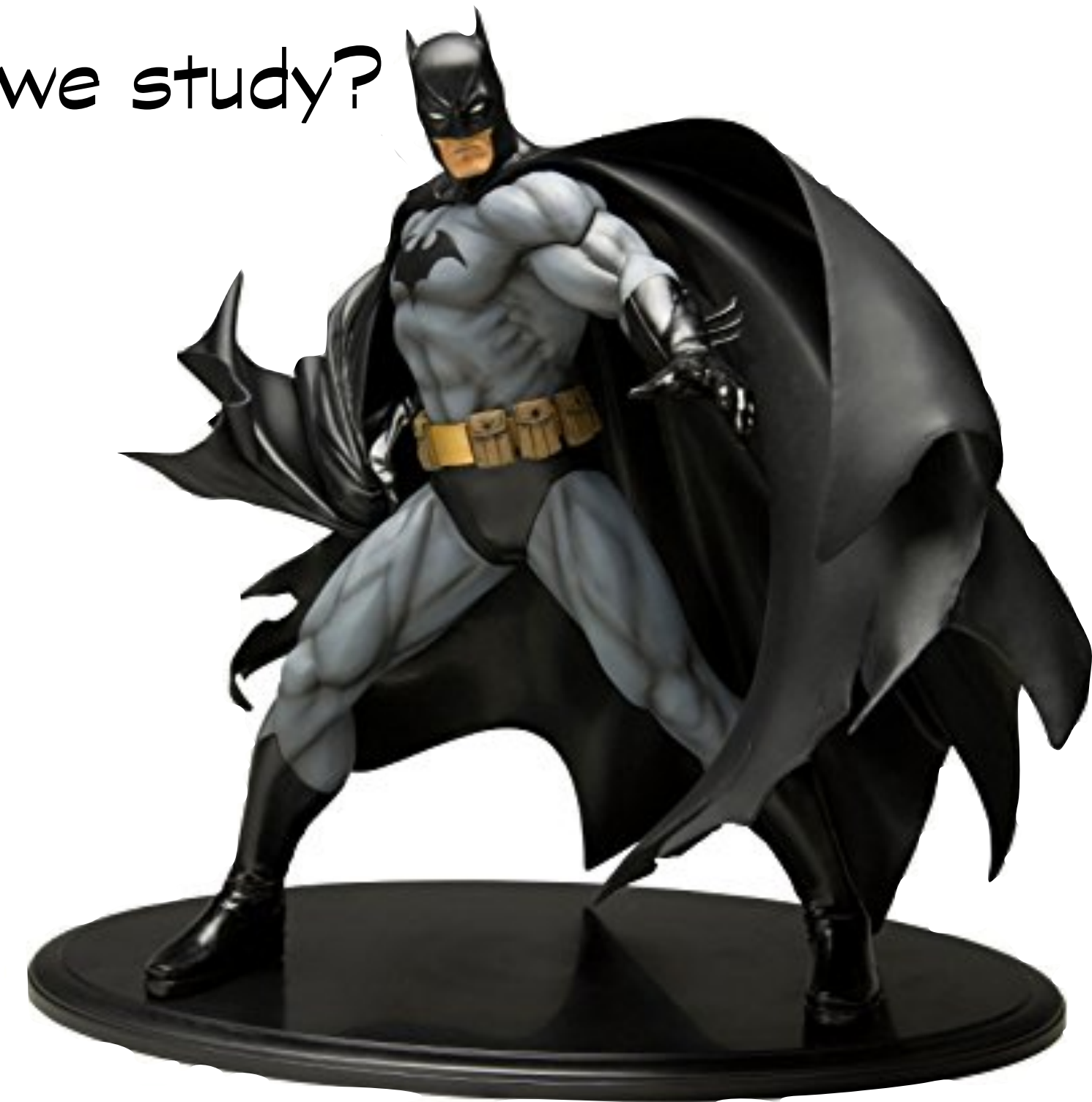
🦇 What is the population of interest? From whom was the data collected?

🦇 If we are interested in tax reform, does it matter whom we ask?

🦇 Corporate Executives? Plumbers? Men? Women?

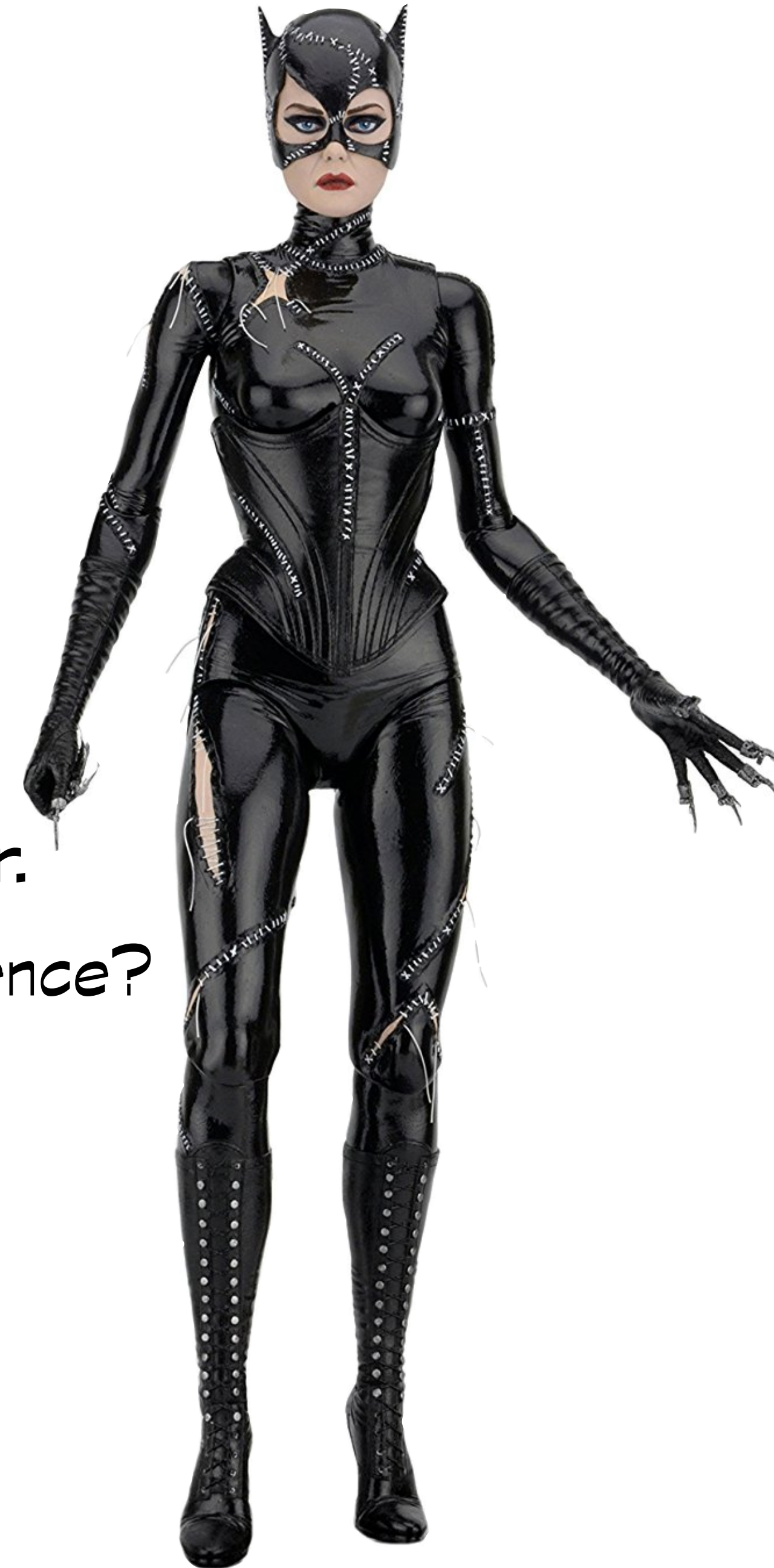
🦇 If we are studying the spread of a virus, does it matter whom we study?

🦇 Mountain Goats? Ducks? Tuna?



What

- What, exactly, are you measuring? What do you want to know?
- If we are looking at incidence of cancer.
 - Size of tumor? Metastasis? Rate of Occurrence?
- Suppose we want to investigate population growth of the Polar Bear.
 - Number of Births? Survival Rate of Cubs? Rate of Occurrence?



When

- When was the data collected? Month? Day? Year? Time?
- If we are interested in popularity of a beverage, when we ask might influence results.
 - If the beverage is hot, is summer a good time to ask?
- If we are trying to determine the political climate of an area, when is the best time to ask?
 - During a presidential primary? During Christmas Holiday?



Where

- ☛ Where was the data collected? Country? Region? State? City?
- ☛ Opinions on gun control.
 - ☛ Arizona? At a gun show? Los Angeles? Cody, Wyoming?
- ☛ Population decline of whales?
 - ☛ Pacific Ocean? Atlantic Ocean? Japan? United States?



🦇 How, exactly, was the data collected? In person? Phone? Internet?

🦇 Opinions on internet companies.

🦇 Phone? Internet? Los Angeles? Cody, Wyoming?

🦇 How do people get their music?

🦇 At a record store? Online Poll? At Comic-Con?



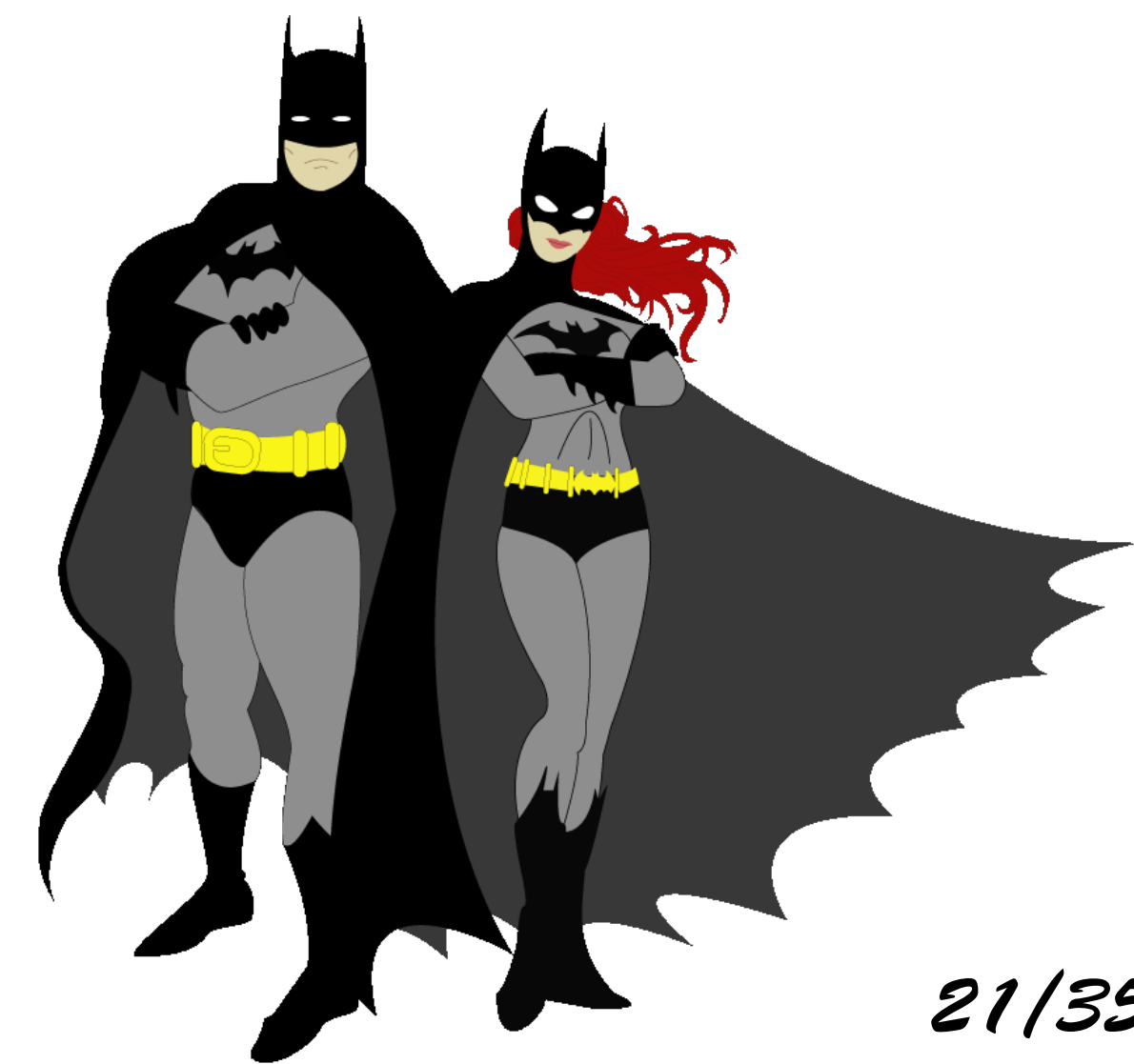
Why

- 🦇 Why is the data being gathered. What question is being asked?
- 🦇 In this artificial classroom environment, you may not know why. But if you are doing your own research, why will inform what, when, where, and how the data will be collected.



Who

- 🦇 The **Who** of the data tells us the individual **cases** for which (or whom) we have collected data.
- 🦇 Individuals who answer a **survey** are called **respondents**.
- 🦇 People on whom we **experiment** are called **subjects** or **participants**.
- 🦇 Animals, plants, and inanimate subjects are called (my personal favorite) ...
 - 🦇 **experimental units**.
- 🦇 We need to know the **Who** of the data so we can understand what the data say.
 - 🦇 Is running the 100 meters in 10 seconds fast?
 - 🦇 Not if you are a cheetah.

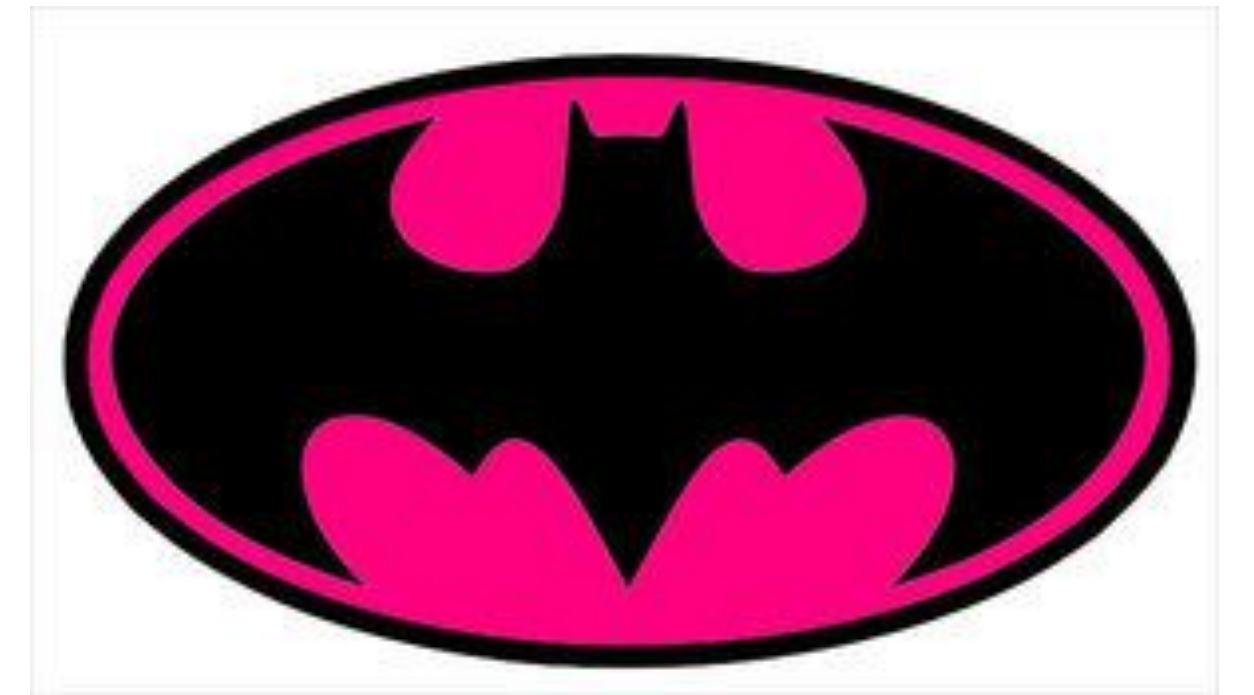


Data Table

- 🦇 We will spend a lot of time looking at data. Often the data will be organized into a table.
- 🦇 A **data table** of subsidies provides by New York to spur job growth shows the context of the data presented:
- 🦇 This data table tells us the What (column) and Who (row) for these data.
- 🦇 There is no information about where, when, why, or how.

Table 1: JCRP grants recaptured.			
Company Name	Grant Distributed	Recapture	Reason for Default
Board of Trade of the City of New York, Inc. (Aka ICE Futures)	\$23,300,000.00	\$8,737,500	Termination
AON Service Corporation	\$4,235,000.00	\$1,450,000	Jobs Default
American Stock Exchange	\$3,000,000.00	\$1,125,000	Merger
Refco Group, Ltd.	\$2,000,000.00	\$447,214.70	Bankruptcy
BearingPoint, Inc.	\$1,941,504.00	\$38,565.74	Bankruptcy
Thacher Proffitt & Wood	\$1,500,000.00	\$100,000	Dissolved
Dow Jones & Company, Inc.	\$1,402,250.00	\$329,388.52	Relocated
T-Systems, North America, Inc.	\$1,200,000.00	\$342,500	Relocated
Holland & Knight LLP	\$800,000.00	\$197,580	Relocated
Country-Wide Insurance Company	\$450,000.00	\$224,689.14	Jobs Default
Embassy Suites New York City	\$267,000.00	\$114,810	Relocated
Georgeson Shareholder Communications, Inc	\$235,000.00	\$62,837.36	Jobs Default
Banco Popular North America	\$185,000.00	\$138,750	Jobs Default
The Regent Wall Street Hotel	\$100,000.00	\$75,000	Out of Business
Total Recaptured		\$13,383,835	

Variables



- 🦇 **Variables** are characteristics recorded about each individual.
- 🦇 **Variables** are holistic, a variable will have multiple possible values.
- 🦇 **Variables** have names identifying **What** has been measured.
 - 🦇 To determine appropriate **variables**, you **Think** about **what you want to know**.
- 🦇 **Variables** have values, but the values are **not** the variable.
 - 🦇 Be careful not to confuse variable names with variable values.
- 🦇 **Variables** should include units defining scale and how each value has been measured.
 - 🦇 **Distance** (variable) might be measured in values of miles, kilometers, ft, cm, etc.
 - 🦇 **Personality** (variable) might be measured in MMPI Scale (hypochondriasis, depression, hysteria, psychopathic deviance (social deviance), masculinity versus femininity, paranoia, psychasthenia (obsessive/compulsive qualities), schizophrenia, hypomania, and social introversion) or Meyers-Briggs measures (Extrovert/Introvert, Sensing/Intuition, Thinking/Feeling, Judging/Perceiving).

Variable vs. Value

🦇 In the previous examples be careful about confusing the **variable name** with the **variable values**.

variable name	variable values
Distance	miles
Personality	Extrovert/Introvert, Sensing/Intuition, Thinking/Feeling, Judging/Perceiving
Gender	male/female
Size	XL, L, M, S, XS



Who, What, and Why

- Levi Strauss & Co showed a list of clothing to a sample of students, and asked which clothes would be most popular.

Context

Who?	Jeans	Levi Strauss & Co	Students
What?	Jeans	Clothes popularity	Students
Why?	Advertising	Marketing	Not Specified



Variable Type

- ⚡ A **categorical (or qualitative)** variable has a **counted** value, **names** categories, and provides information about how cases fall into those categories.
 - ⚡ Categorical examples: sex, race, ethnicity, personality type
- ⚡ A **quantitative** variable is a **measured** variable (with units) that informs about the quantity of what is being measured.
 - ⚡ Quantitative examples: income (\$), SAT score, weight (pounds)
- ⚡ The questions we ask of a variable (Why) informs how we interpret the data and how we treat the variable.

Types of Variables

Employees of StatCo					
Name	Hire Date	PIN	Gender	Age	Salary
Chariah Jones	6/2/80	10025	F	31	\$35,643
Jose McCarren	4/16/86	10059	M	42	\$42,750
Brian Yu	10/13/92	10101	M	47	\$54,875

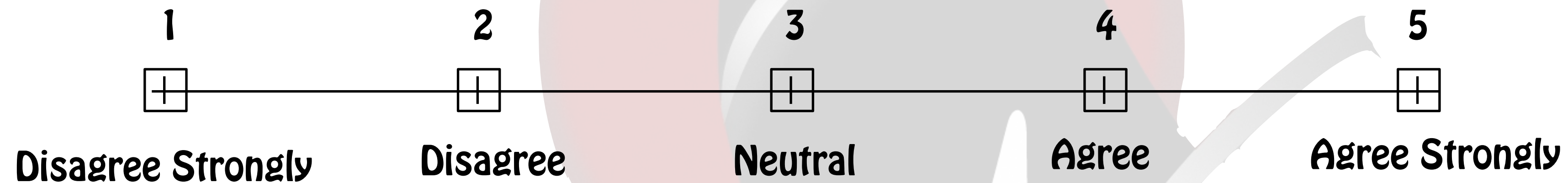
Categorical Variables

Quantitative Variables

Be careful with identifiers. Identifiers are not variables.

Variable Types

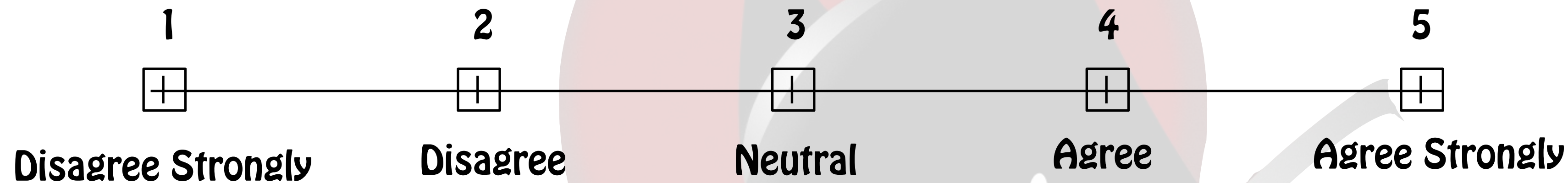
- 🦇 In a student evaluation of instruction at a university, one question asks students to evaluate the statement “The instructor was generally interested in teaching” on the following scale:



- 🦇 This is known as a “Likert Scale”
- 🦇 Variable? 🦇 **interest in teaching**
- 🦇 Values? 🦇 **1, 2, 3, 4, 5 or Disagree Strongly, Disagree, ...**
- 🦇 Question: Is **interest in teaching** categorical or quantitative?
 - 🦇 That depends on how we intend to use the data (Why).

Variables

🦇 Question: Is **interest in teaching** categorical or quantitative?



🦇 You know there is an “**order**” to the responses, but there are no natural units for the variable **interest in teaching**.

🦇 **interest in teaching** is an **ordinal variable**.

🦇 With an ordinal variable, look at the **Why** of the study to decide whether to treat it as categorical or quantitative. What you plan to do with the data will also determine the type. Are you going to find a mean value? or are you simply determining a ranking?

Sometimes a count is just a count.

- When we count the cases in each category of a categorical variable, the **counts** are not the **data or values**, but something we summarize about the variable categories.
- The category labels are the **What** (variable), and
- the individuals counted are the **Who**.



Not variables, simply counts.

Variable

Degree

Number

Rel Freq

Percent

Values for
variable

H.S.

2

0.05

5

Bachelor'

7

0.175

17.5

Master's

23

0.575

57.5

Law

4

0.1

10

PhD

4

0.1

10

Frequencies

Sometimes a count is a variable value.

🦇 When we focus on the amount of something, those counts then become values of a quantitative variable. Tracking the number of fatalities from mass shootings (4 or more killed or injured by gunfire) in the U.S. since 2000.

🦇 The Who is Year

🦇 The What is fatalities from mass shootings

🦇 The number of fatalities is value of the quantitative variable **number of fatalities**.

🦇 In this case, the count is the value of a variable.

Year	Number
2016	63
2015	46
2014	17
2013	31
2012	67
2011	18
2010	8
2009	38
2008	16
2007	51
2006	18
2005	16
2004	4
2003	6
2002	6
2001	4
2000	7

Identifiers

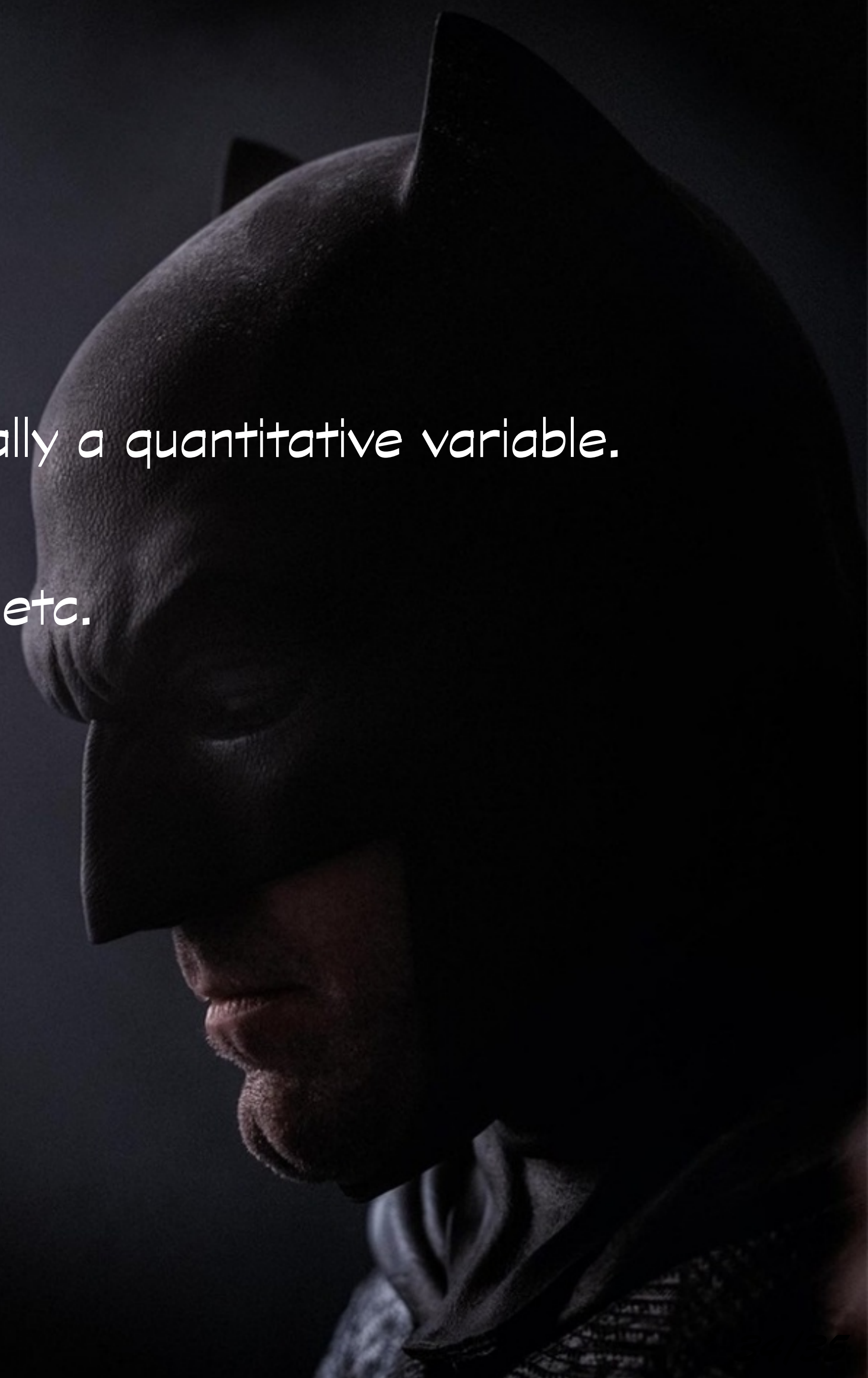
- 🦇 **Identifier variables** are categorical variables with exactly one individual in each category.
- 🦇 Examples: Name, Social Security Number (SSN), ISBN, UPS Tracking #
- 🦇 Don't be tempted to analyze identifier variables.
- 🦇 Be careful not to consider all variables with one case per category, like year, as identifier variables.
- 🦇 The Why will help you decide how to treat identifier variables.
- 🦇 You are not just a number, you are several numbers. In this class you are your class id, school id, permanent id, and state student id. Those are just for this class.

The Ws

- 🦇 We need the Who, What, and Why to analyze data. But, the more we know, the more we understand and the better we can interpret our results.
- 🦇 When and Where give us some nice information about the context.
 - 🦇 Example: Values recorded at a large public university may mean something different than similar values recorded at a small private college.
 - 🦇 Asking about your favorite meal may return very different results at 6:00 AM vs. 6:00 PM.
- 🦇 How data are collected can spell the difference between knowledge and garbage.
 - 🦇 Example: results from Internet surveys are nearly always garbage. Asking students for their GPA is garbage data, searching school records for GPA is most assuredly knowledge, not garbage.
 - 🦇 So, before you even begin to collect and analyze data get clear about the **Ws**. Being certain about **W**hy, **W**ho, **W**hat, **W**here, ho**W**, and **W**hen will help insure your data is useful and your conclusions are valid.

Warnings

- ⚠ Do not assume a variable with number values is automatically a quantitative variable.
ID numbers are not variables but identifiers.
- ⚠ Categories can be labeled with numbers; Class 1, Class 2, etc.



- 🦇 Know how to enter and edit data in a list.
- 🦇 The information is given starting on page 8 of the text.

Stat

1:Edit

Select List "L1"

Enter first value "20"

Enter

Enter 2nd value "30"

Enter

Repeat to end of list

