Welcome to DATA 151

I’m so glad you’re here!

DATA 151: CLASS 2A
INTRODUCTION TO DATA SCIENCE (WITH R)

SAMPLING DESIGN

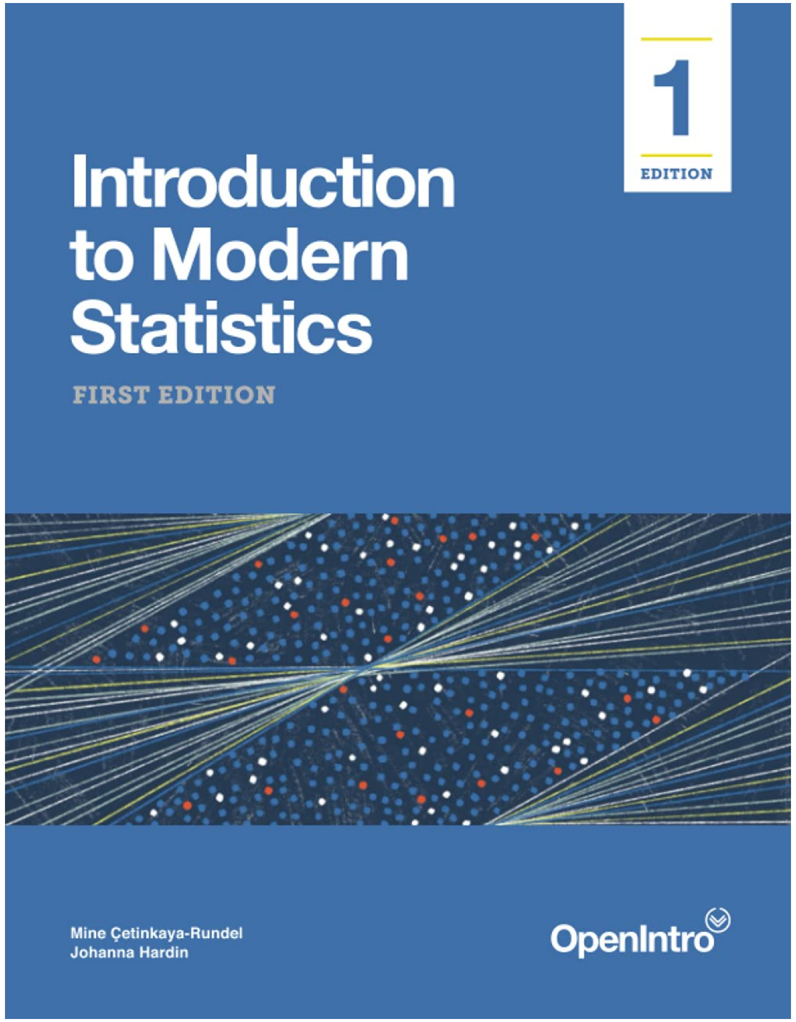
NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

AGENDA: DATA 151 – CLASS 2A

Time	Topics
2:30 – 2:45	Announcements Review and Warm-up
2:45 – 3:10	Principles of data collection Basics of Sampling Discussion about bias, errors, and ethics
3:10 – 3:15	*5 minute break*
3:15 – 3:45	Sampling activity with beads
3:45 – 4:00	Sampling Designs

ANNOUNCEMENTS

RELEVANT READING



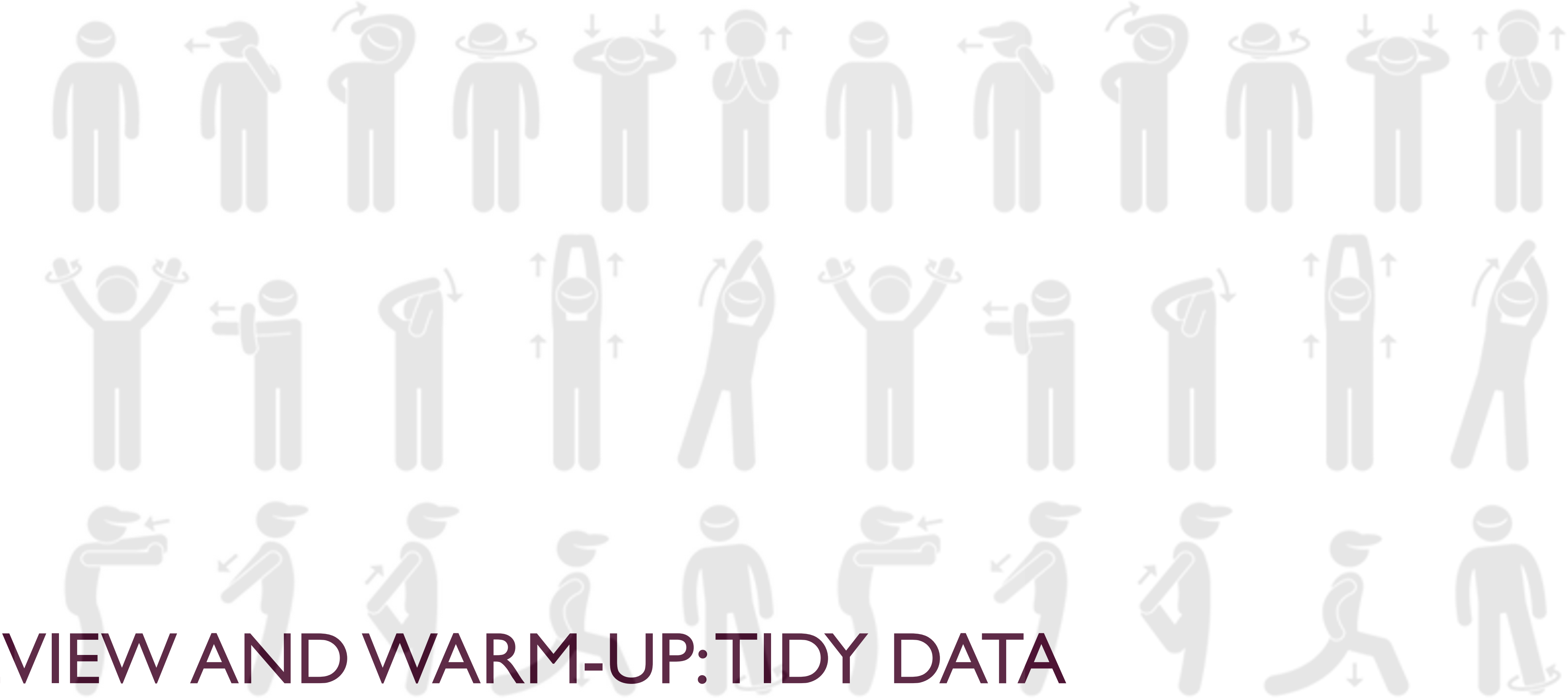
Introduction to Modern Statistics:

- Tuesday:
  - iMStat - Ch 2: Study Design Sections:
    - 2.1: Sampling Principles and Strategies
- Thursday:
  - iMStat - Ch 2: Study Design Sections:
    - 2.2: Experiments
    - 2.3 Observational Studies

HOMEWORK REMINDER

Due this week:

- *Student Info Survey*
  - *Please do before Wednesday so I can put you into project groups on Thursday*
- *Student Sample Survey*
  - *Fun data that we can use for examples in class (this will be de-identified)*
- *HW #1: Dear Data* ***(due in class 9/8)***

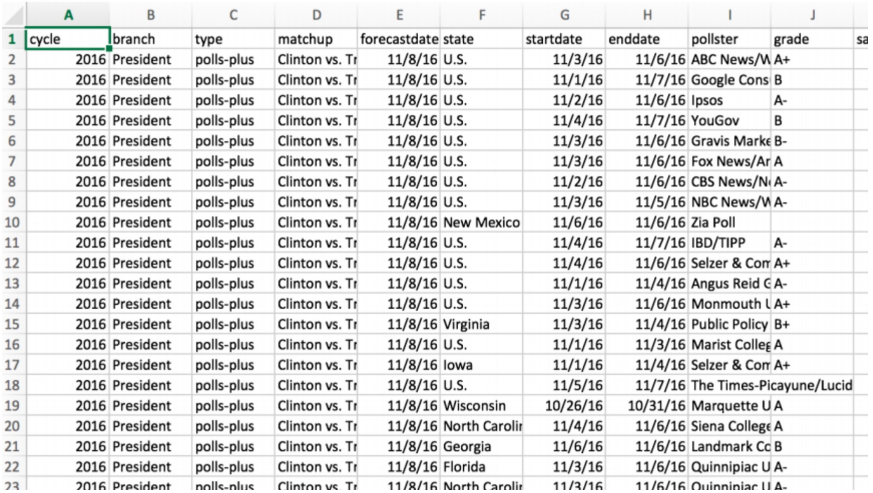


REVIEW AND WARM-UP: TIDY DATA

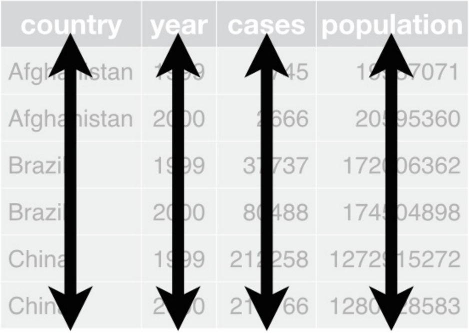
PRINCIPLES OF TIDY DATA

(Wickham 2014)

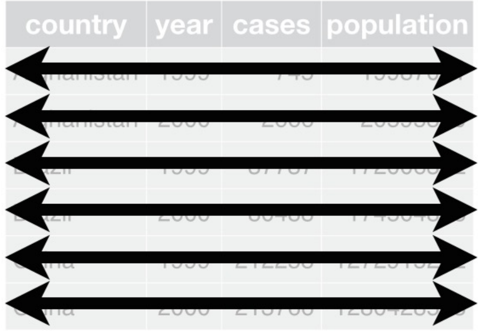
1) Each variable forms a column
2) Each observation forms a row
3) Each cell contains a single value



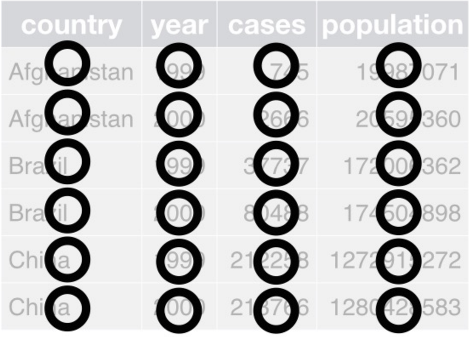
	A	B	C	D	E	F	G	H	I	J	
1	cycle	branch	type	matchup	forecastdate	state	startdate	enddate	pollster	grade	sa
2	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	ABC News/W	A+	
3	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/1/16	11/7/16	Google Cons	B	
4	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/2/16	11/6/16	Ipsos	A-	
5	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/4/16	11/7/16	YouGov	B	
6	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	Gravis Marke	B-	
7	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	Fox News/Ar	A	
8	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/2/16	11/6/16	CBS News/N	A-	
9	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/5/16	NBC News/W	A-	
10	2016	President	polls-plus	Clinton vs. Tr	11/8/16	New Mexico	11/6/16	11/6/16	Zia Poll		
11	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/4/16	11/7/16	IBD/TIPP	A-	
12	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/4/16	11/6/16	Selzer & Com	A+	
13	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/1/16	11/4/16	Angus Reid G	A-	
14	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/3/16	11/6/16	Monmouth U	A+	
15	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Virginia	11/3/16	11/4/16	Public Policy	B+	
16	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/1/16	11/3/16	Marist Colleg	A	
17	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Iowa	11/1/16	11/4/16	Selzer & Com	A+	
18	2016	President	polls-plus	Clinton vs. Tr	11/8/16	U.S.	11/5/16	11/7/16	The Times-Picayune/Lucid		
19	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Wisconsin	10/26/16	10/31/16	Marquette U	A	
20	2016	President	polls-plus	Clinton vs. Tr	11/8/16	North Carolir	11/4/16	11/6/16	Siena College	A	
21	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Georgia	11/6/16	11/6/16	Landmark Cc	B	
22	2016	President	polls-plus	Clinton vs. Tr	11/8/16	Florida	11/3/16	11/6/16	Quinnipiac U	A-	
23	2016	President	polls-plus	Clinton vs. Tr	11/8/16	North Carolir	11/3/16	11/6/16	Quinnipiac U	A-	



variables

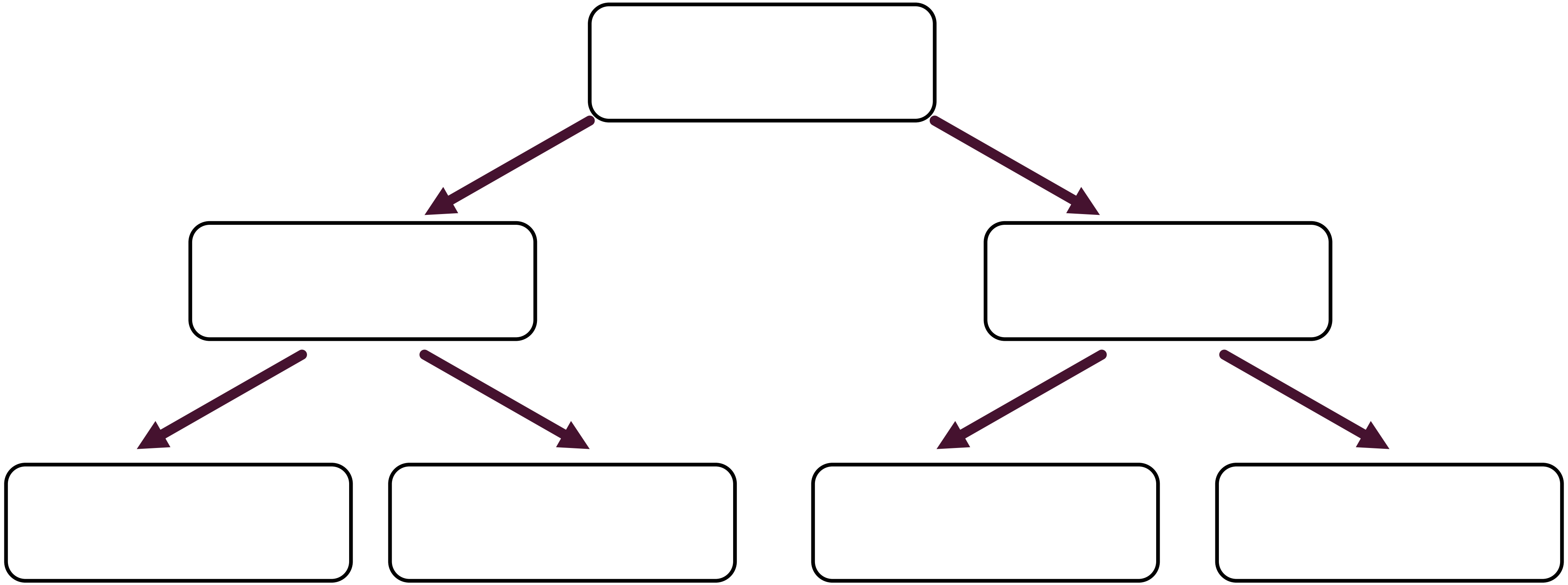


observations

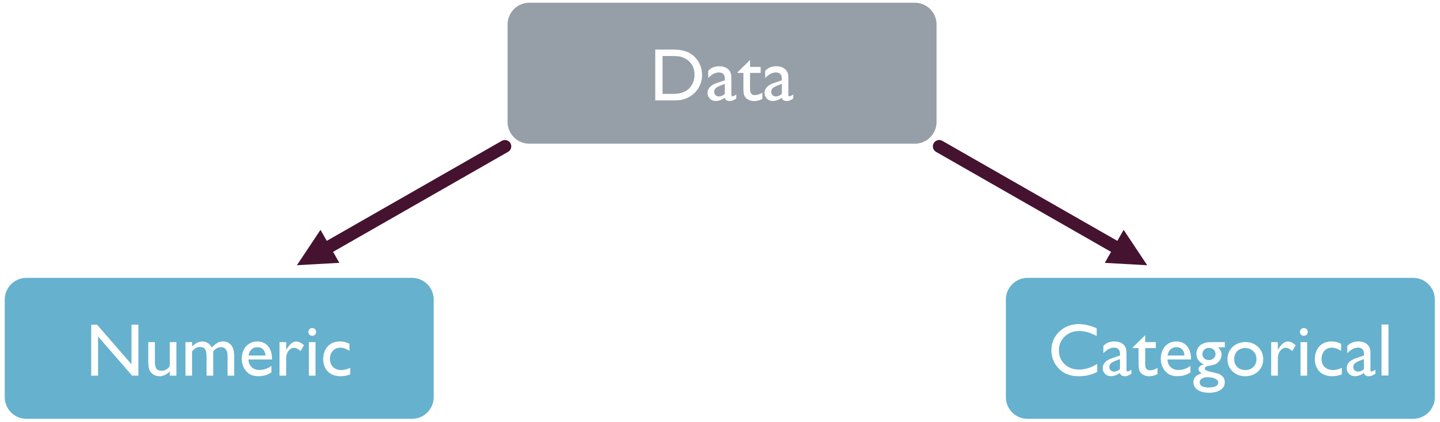


values

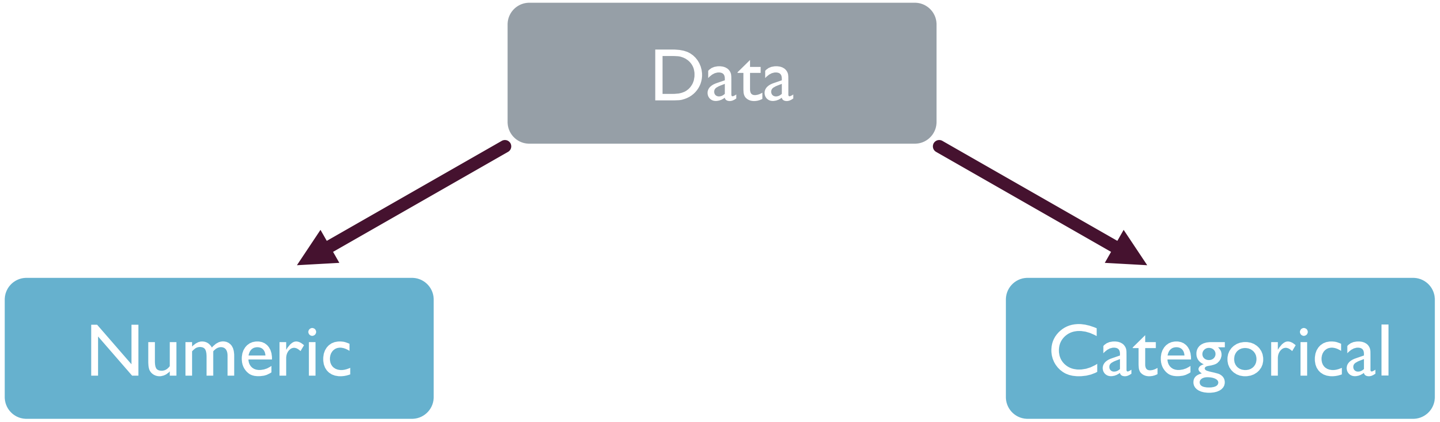
TYPES OF VARIABLES



TYPES OF VARIABLES



TYPES OF VARIABLES


Data

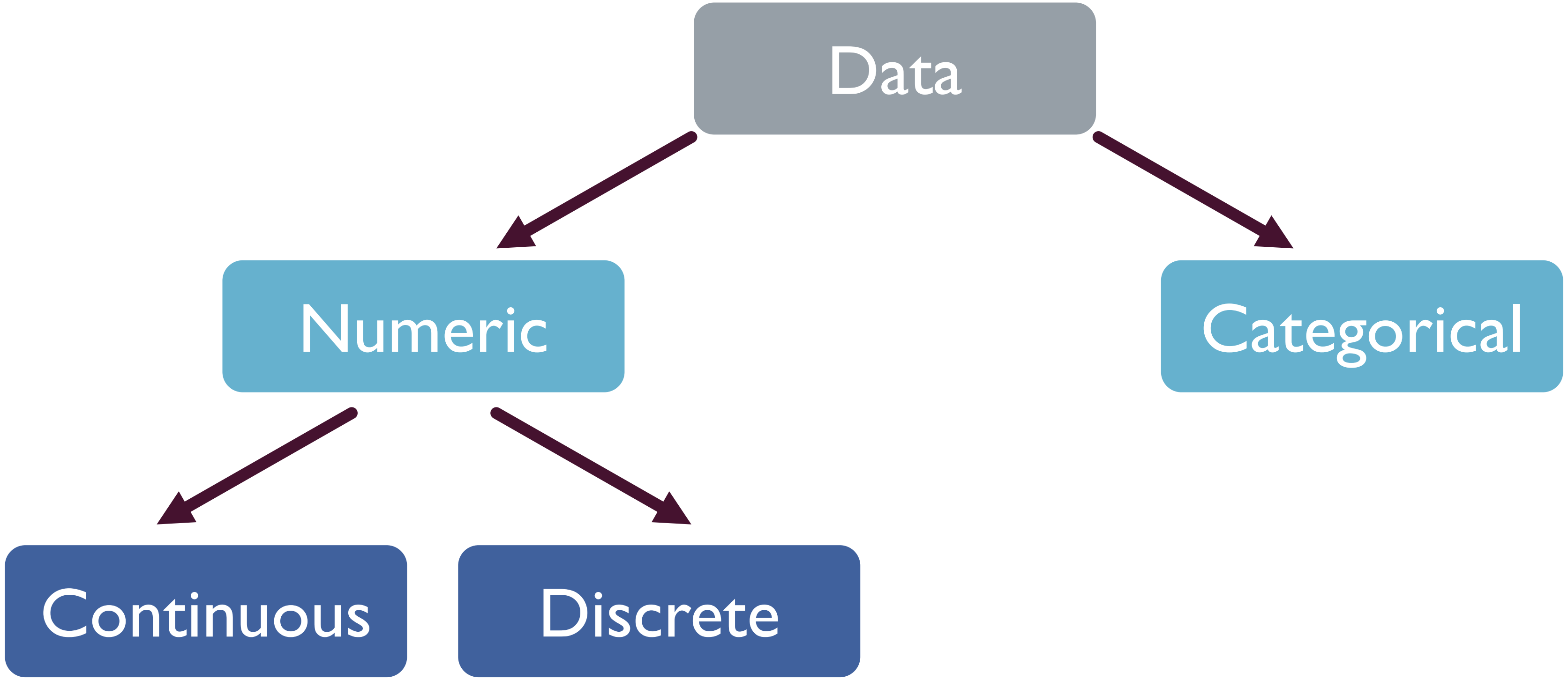
Numeric

Categorical

aka Quantitative

aka Qualitative

TYPES OF VARIABLES



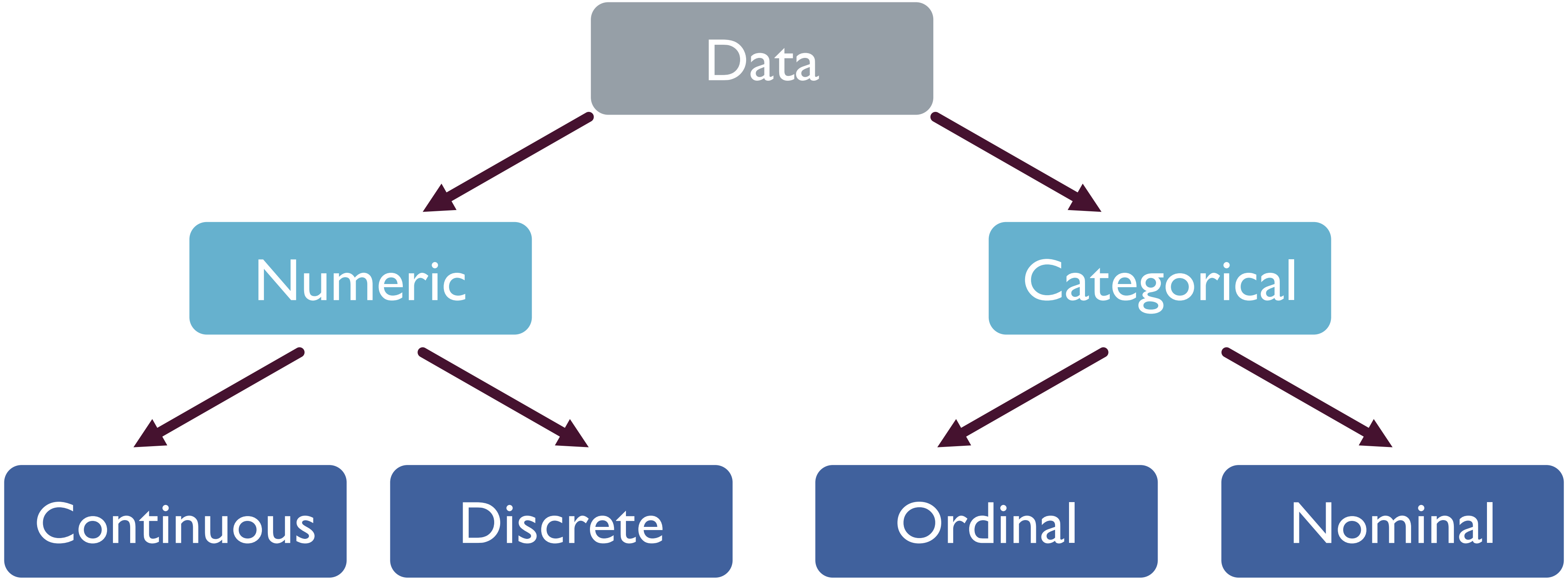
DISCRETE EXAMPLE

- Observations are restricted to a set of values
- A common example are counting numbers (integers)
- **Examples:**
  - Counting startfish

CONTINUOUS EXAMPLE

- Data can take any value within an interval (real numbers)
- **Examples:**
  - Height, weight, …

TYPES OF VARIABLES



Data

Numeric

Categorical

Continuous

Discrete

Ordinal

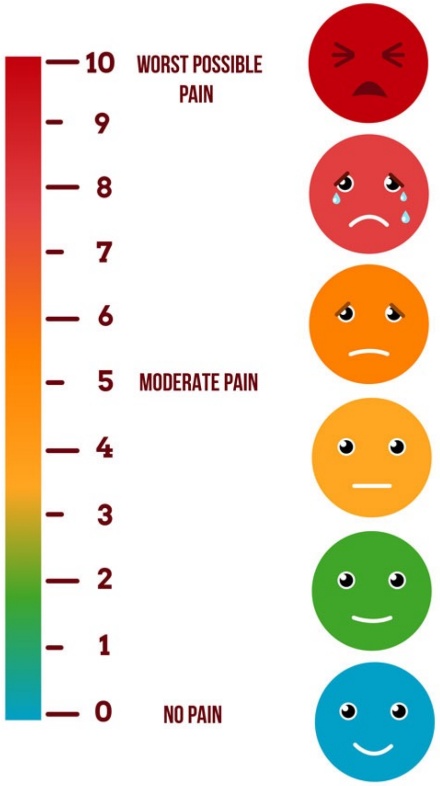
Nominal

(CATEGORICAL) NOMINAL DATA

- A nominal scale describes a variable with categories that do not have a natural order or ranking.
- You can code nominal variables with numbers if you want, but the order is arbitrary and any calculations, such as computing a mean, median, or standard deviation, would be meaningless.
- **Examples of nominal variables include:**
  - genotype, blood type, zip code, gender, race, eye color, political party

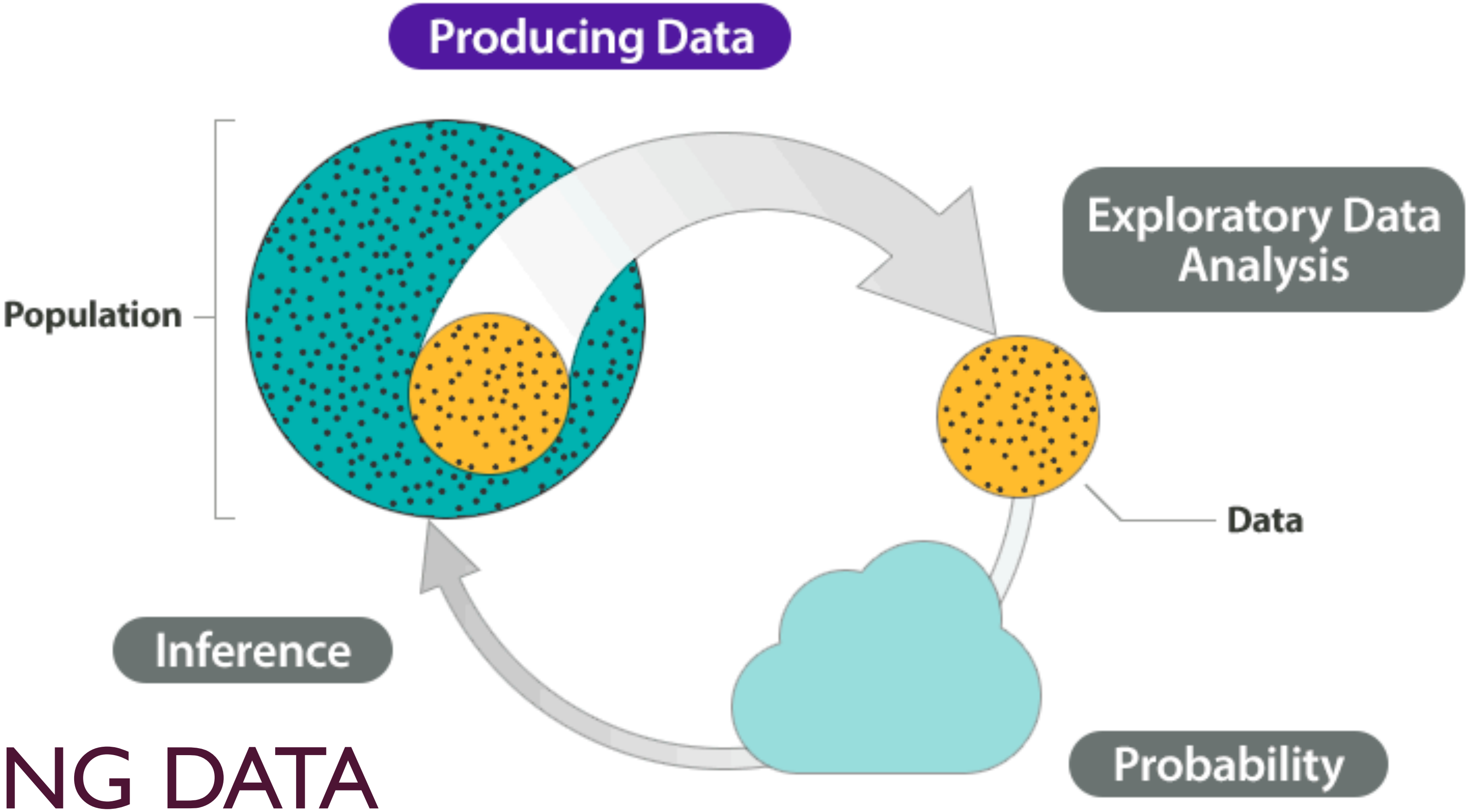
(CATEGORICAL) ORDINAL DATA

- An ordinal scale is one where the order matters but not the difference between values.
- **Likert scale** questions are common in surveys
- **Examples of ordinal variables include:**
  - socio economic status (“low income”,”middle income”,”high income”), education level (“high school”,”BS”,”MS”,”PhD”), income level (“less than 50K”, “50K-100K”, “over 100K”), satisfaction rating (“extremely dislike”, “dislike”, “neutral”, “like”, “extremely like”).



SMALL GROUP WARM-UP (10 MINUTES)

- **Groups of 4**
- Each group will get an example dataset and identify…
  - If the data is tidy
  - Examples of different types of variables



PRODUCING DATA

WHY?

Before digging into the details of working with data, we stop to think about **how data came to be**.

That is, if the data are to be used to make broad and complete conclusions, then it is important to **understand who or what the data represent**.

How are units selected? Can findings be generalized?

SOME KEY TERMS TO LOOK OUT FOR

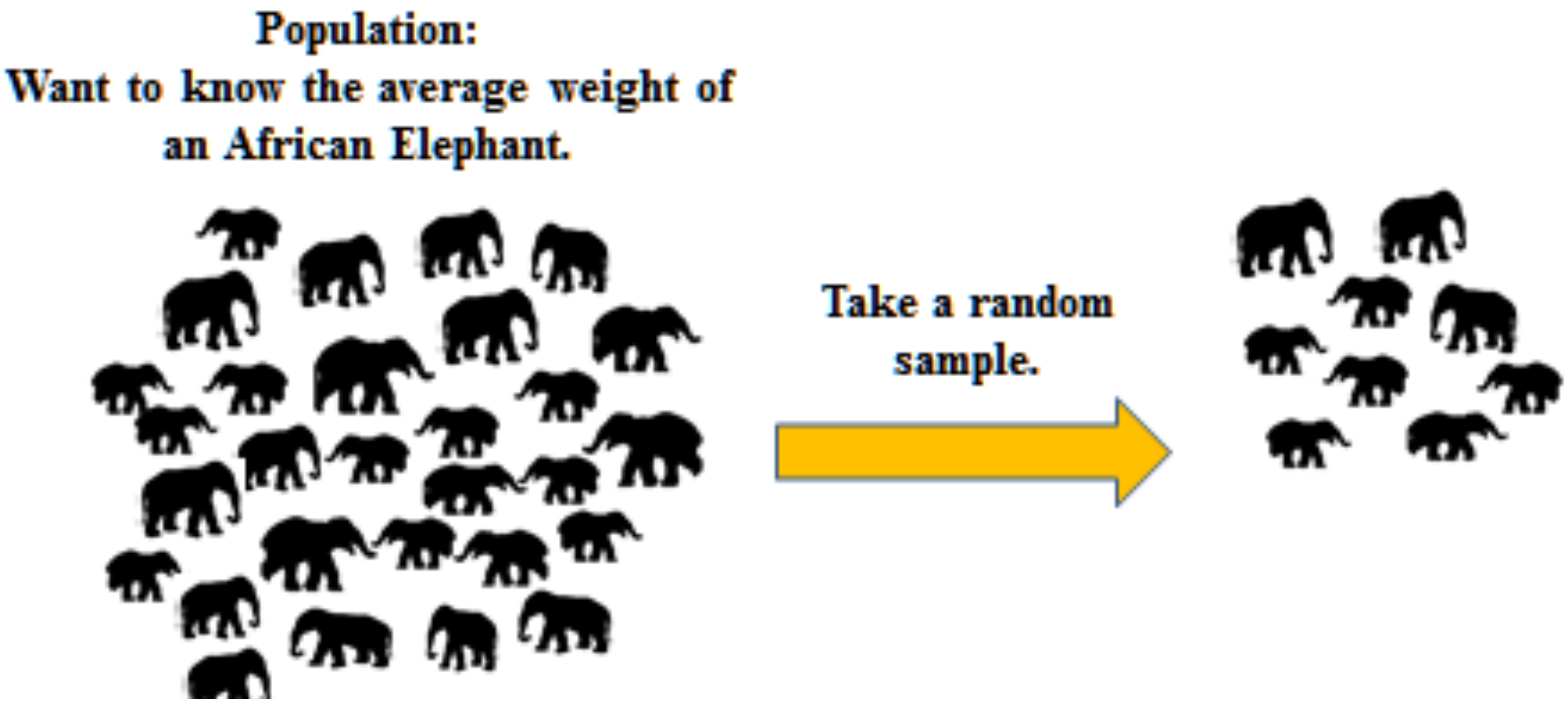
Principles of Data Collection

- Population
- (Sampling) Frame
- Census
- Population Parameters
- Sample
- Statistic

PRINCIPLES OF DATA COLLECTION

1. Well defined research question or topic of interest
2. Identify what subject(s) or cases should be studied
3. Consider what mechanisms of data collection are used
    1. Random sampling?
    2. Experimental design?

SAMPLING



Population:
Want to know the average weight of an African Elephant.

Take a random sample.

POPULATIONS AND SAMPLES

Definitions:

- **(Target) Population**: the entire group of individuals we want to make inference about
- **Sample:** a subset of the population

Note: Researchers often desire to obtain info about an entire group, but lack the resources to collect information on every individual in the population (this is called a census), therefore we collect a sample.

REAL WORLD EXAMPLE

REAL WORLD EXAMPLE

- The **United States census** is legally mandated by the US Constitution
- Takes place every 10 years
- The first census after the American Revolution was taken in 1790, under Secretary of State Thomas Jefferson

The 2020 Census is estimated to cost **approximately $15.6 billion** after adjusting for inflation.

REAL WORLD EXAMPLE

Happens every year and is random sample of the US population.

EXAMPLE 1: POPULATIONS VS SAMPLES

The Principal Financial Group conducted a survey of 1172 employees in the United States between July 28, 2010 and August 8, 2010, and asked if they were currently participating in the employer-sponsored automatic payroll deduction for a 401(k) plan to save for retirement.

- What is the POPULATION in the example?
- What in the SAMPLE in this example?

SAMPLING

1. What is sampling?

A procedure of selecting certain individuals from a population of interest to participate in a study.

SAMPLING

2. Why do we ”sample”?

1. Limited resources usually make it impossible to collect data on the entire population.
2. If done properly, inferences can be drawn about the entire population based on data collected from the sample

SAMPLING

3. What is the main goal of sampling?

To draw a sample from the population so that the sample is representative of the population.

SAMPLING

4. So, how can we obtain a sample that is representative of the population?

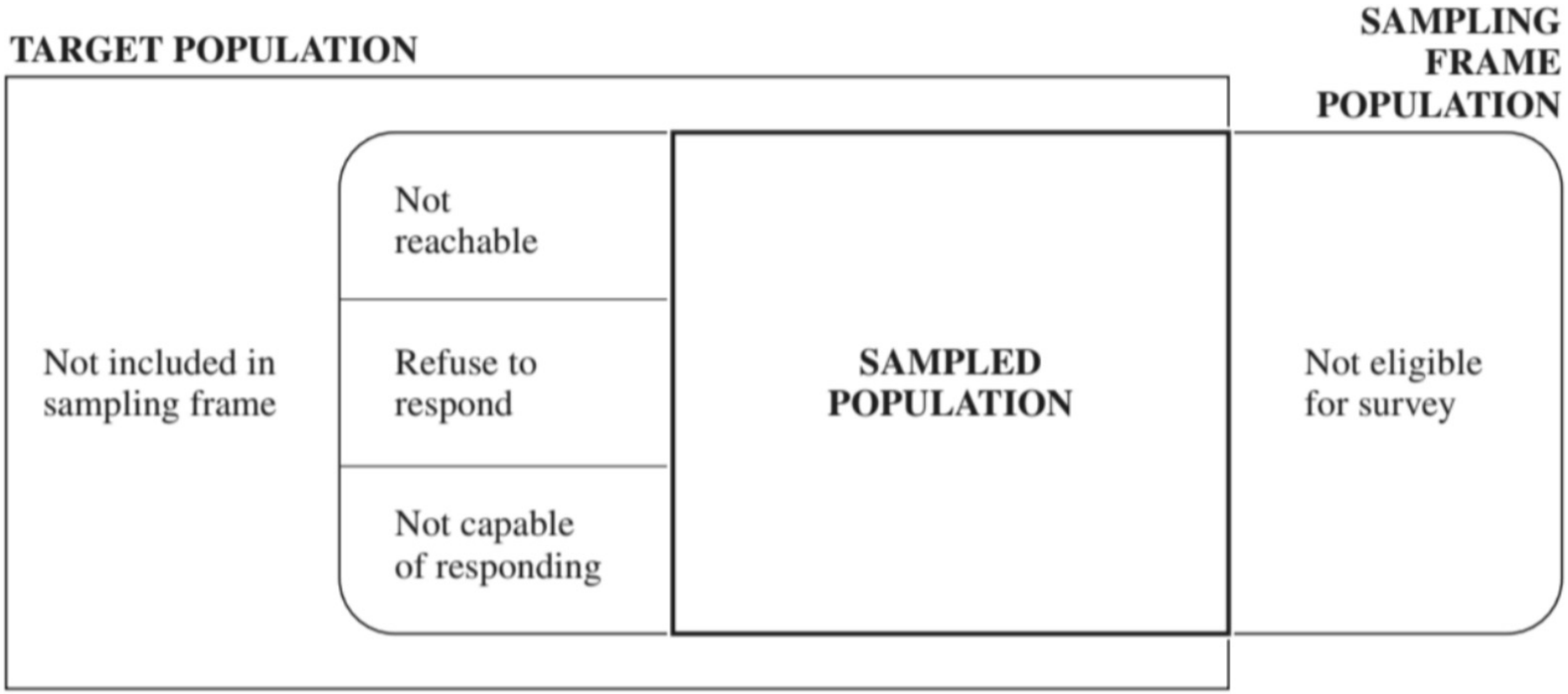
Randomly select individuals!

SAMPLING

Some more important definitions:

- **Sample:** a subset of a population
- **Sampled population:** the collection of all possible observation units that might have been chosen in a sample; the population from which the sample was taken
- **Sampling unit:** a unit that can be selected for a sample.
  - **Ex:** individuals or households
- **Sampling frame:** a list, map, or other specification of sampling units in the population from which a sample may be selected

SAMPLING



TARGET POPULATION

SAMPLING FRAME POPULATION

Not included in sampling frame

Not reachable

Refuse to respond

Not capable of responding

SAMPLED POPULATION

Not eligible for survey

EXAMPLE 2: ODOT

ODOT (Oregon Department of Transportation) performs the Transportation Needs and Issues Survey (TNIS) biennially. ODOT uses data from this survey to inform policy decisions on the state gas tax, road and bridge maintenance, etc. Since 2010 respondents have been randomly chosen from the delivery sequence file (DSF) from the United States Postal Service (USPS). Approximately 1200 Oregon households were selected to participate in the survey.

Identify the following:

- Target Population, sampling frame, sampling unit, and sample

BIAS IN SAMPLING

BIAS IN STATISTICS

Definition:

- **Biased:** The difference between the true value and the observed value

TRUE – OBSERVED = BIAS

SOURCES OF BIAS

Definition:

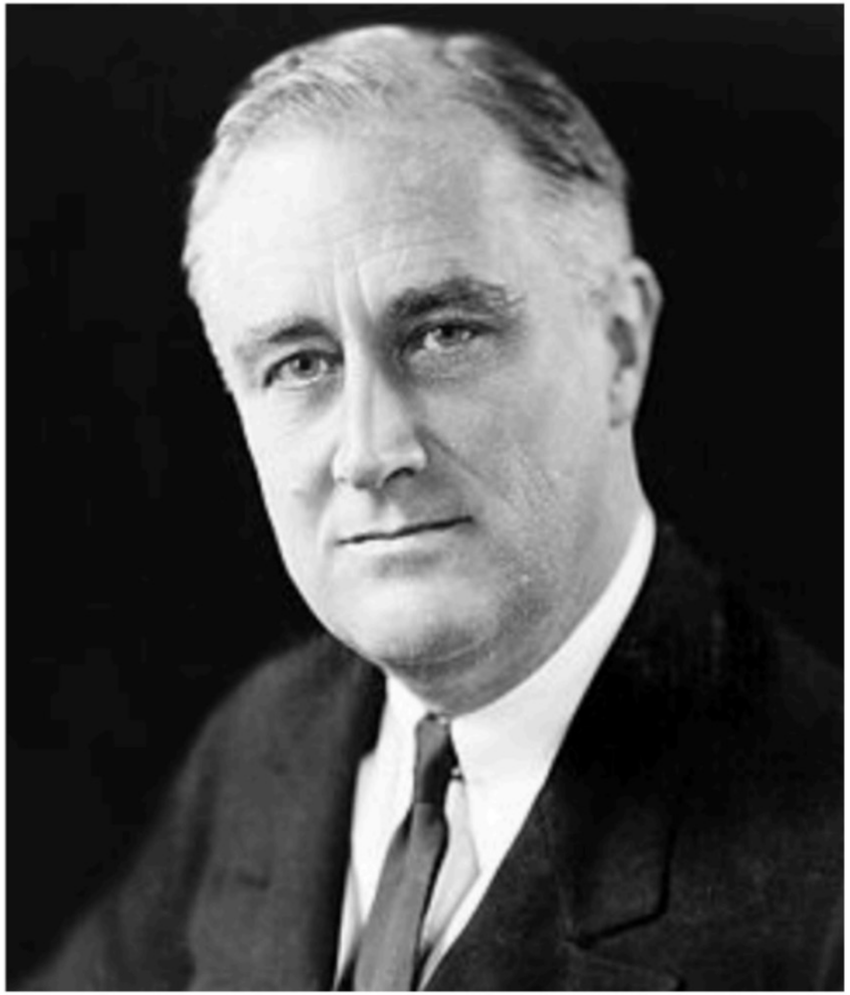
- **Biased sampling:** sampling that is likely to under or over represent groups in the population that tend to have different values under investigation

SOURCES OF BIAS

Types of Bias:

1. **Undercoverage Bias / Coverage Error:** when some groups in the population are left out of the process of choosing the sample.
2. **Nonresponse Bias / Nonresponse Error:** when an individual chosen for the sample can't be contacted or refuses to participate. It is the sampling bias that results when the distribution of a variable of interest is different for the non-responders and responders.
3. **Response Bias / Measurement Error:** When an individual does not answer accurately or honestly for any reason.

EXAMPLE 3: LANDON VS FDR



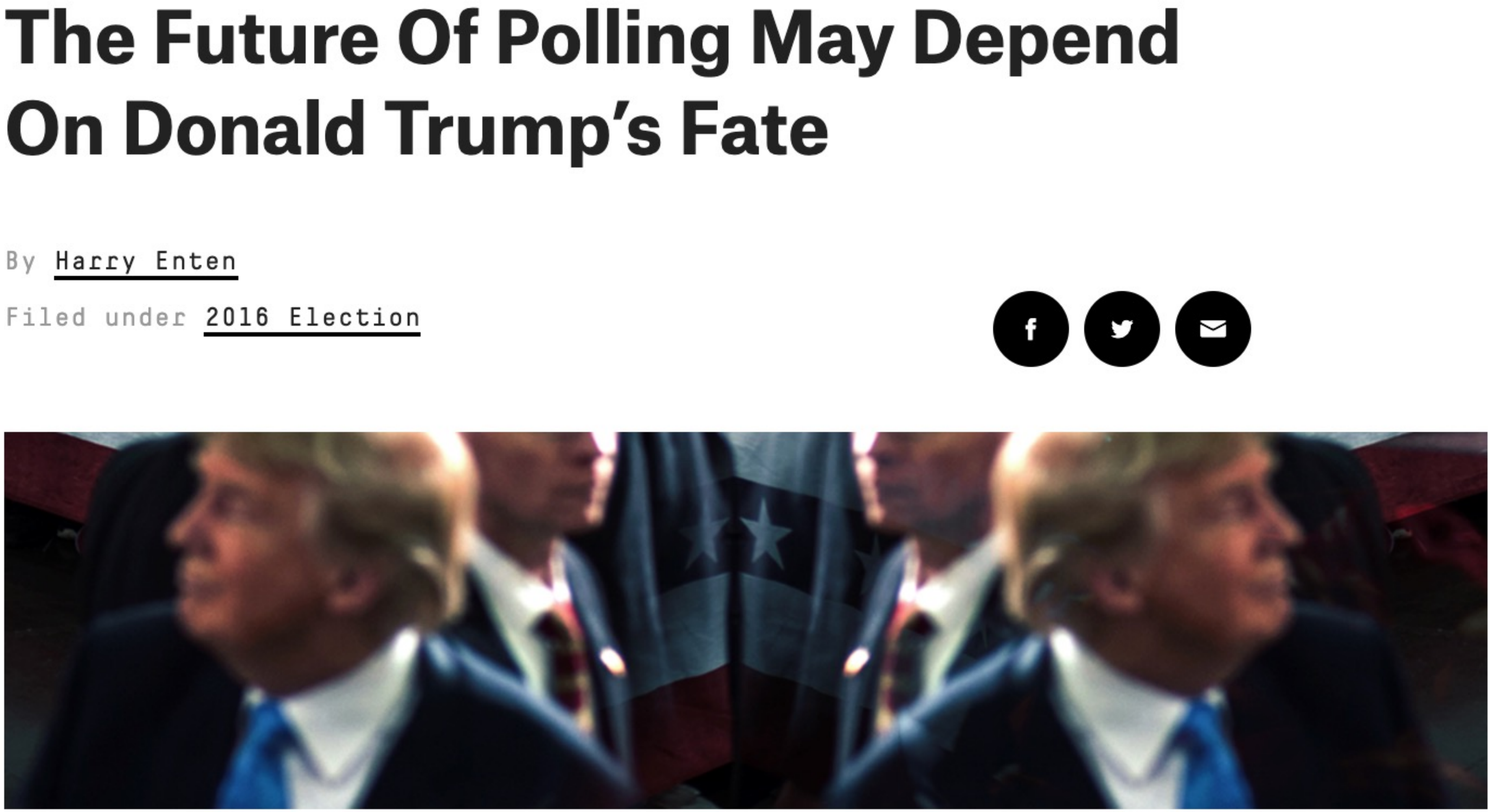
EXAMPLE 3: LANDON VS FDR

- It was 1936 and the Great Depression was 7 years old.
- Roosevelt was up for re-election in 1936 and faced Republican, Alf Landon.
- Literary Digest, an influential weekly magazine of the time had begun political polling
- They had polled a sample of 2.4 million people based upon telephone and car registrations. The population of the US at that time was 128 million.
  - Prediction: 43% for FDR
  - Result: 62% for FDR

Citation: Qualtrics

What do you think happened?!

EXAMPLE 4: SHY TRUMP VOTER EFFECT?

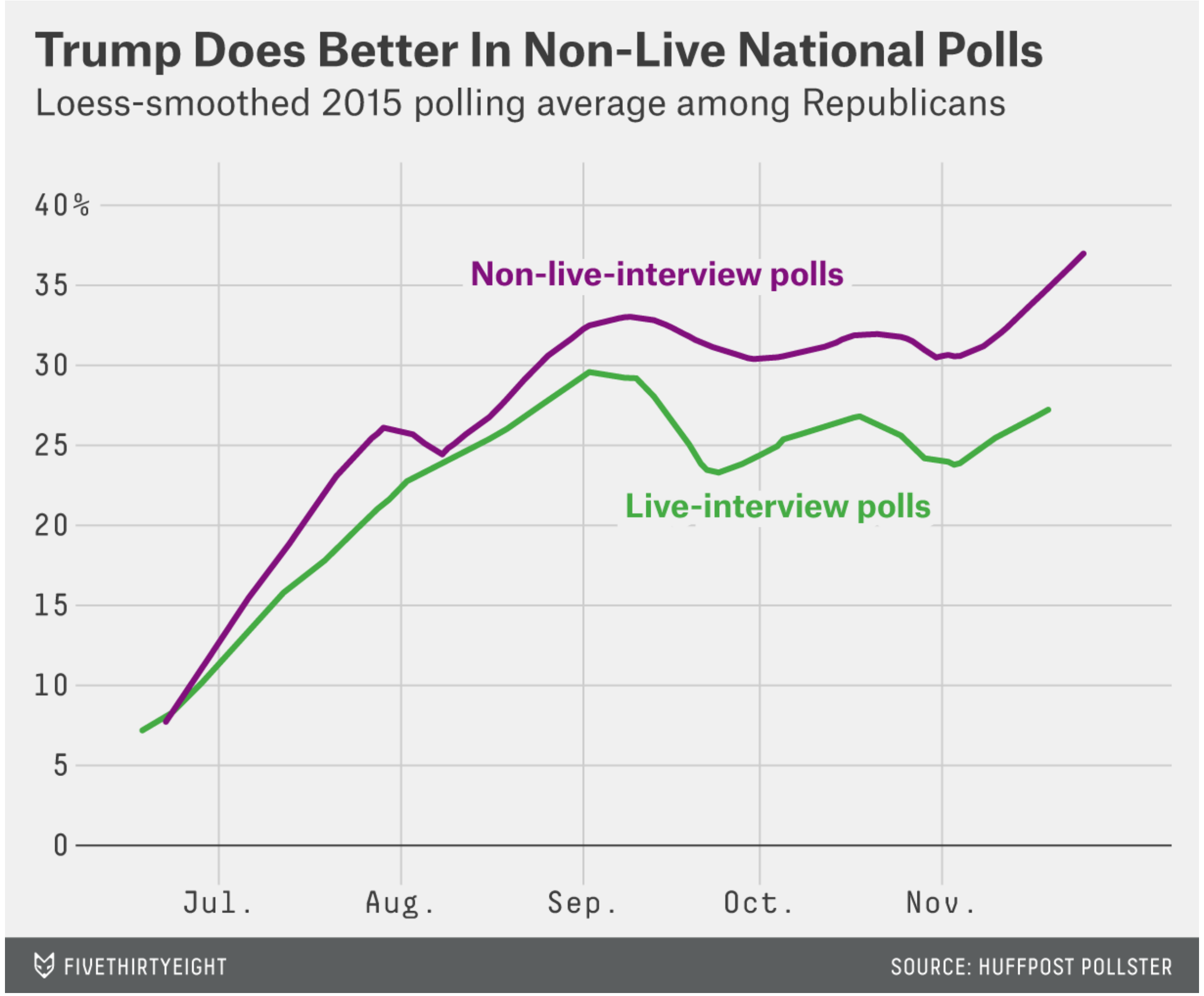


The Future Of Polling May Depend On Donald Trump's Fate

By Harry Enten

Filed under 2016 Election

EXAMPLE 4: SHY TRUMP VOTER EFFECT?



What do you think happened?!

PARAMETERS VS STATISTICS

PARAMETERS VS STATISTICS

Parameters vs Statistics

- A **parameter** is a number that describes some characteristic of the population
  - In practice, this value of a parameter is unknown
  - Usually denoted with Greek letter, such as μ

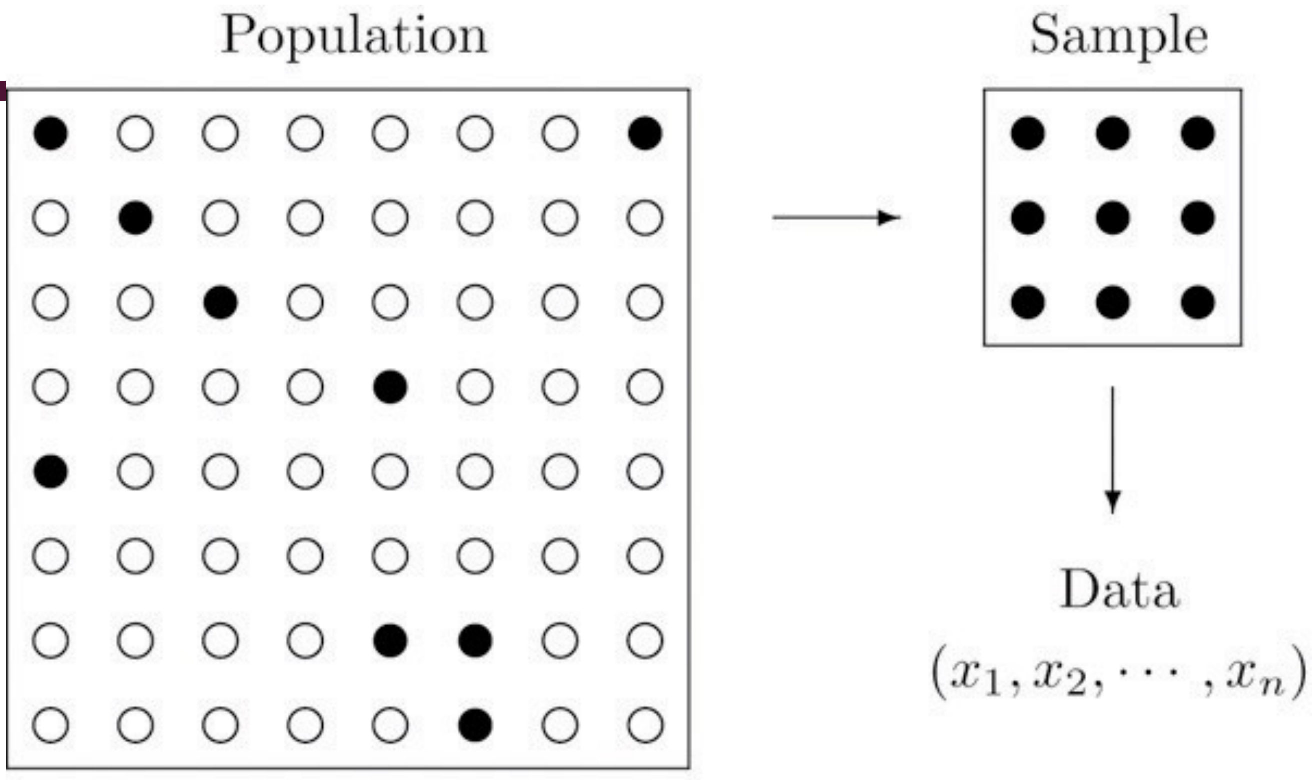
- A **statistic** is a number that describes some characteristic of a sample
  - The value of a statistic can be calculated directly from the sample data.
  - We often use a statistic to estimate the value of the unknown parameter.

STATISTICAL NOTATION

- n = Number of observations; sample size

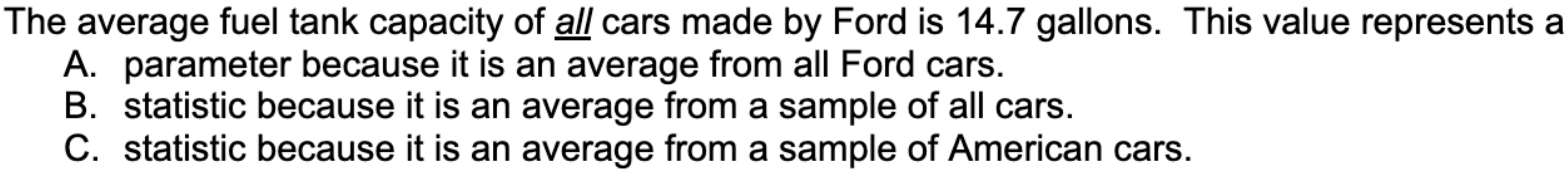
- μ = population mean
  - Considered fixed and unknown
  - What we are trying to estimate

- $\bar{x}$ = Sample mean
  - *This is the mean of some quantitative variable for a sample of size n.*
  - *The average (arithmetic mean) of the observed numbers*



	Parameters		Statistics	
Population mean	μ	←	$\bar{x}$	Sample mean
Population variance	σ^2	←	s^2	Sample variance
Population standard deviation	σ	←	s	Sample standard deviation
Population proportion	p	←	$\hat{p}$	Sample proportion

PRACTICE QUESTIONS



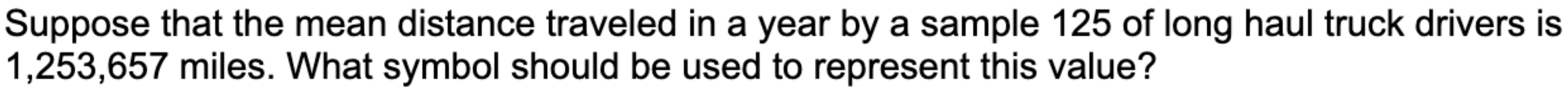
The average fuel tank capacity of ***all*** cars made by Ford is 14.7 gallons. This value represents a

A. parameter because it is an average from all Ford cars.

B. statistic because it is an average from a sample of all cars.

C. statistic because it is an average from a sample of American cars.

PRACTICE QUESTIONS



Suppose that the mean distance traveled in a year by a sample 125 of long haul truck drivers is 1,253,657 miles. What symbol should be used to represent this value?

ACTIVITY

Please form 6 groups of approximately 4 people

Supplies: Each group will need a container of 200 beads and a small cup to obtain their sample

Steps:

1. Make a hypothesis
2. Collect Data
3. Record the Data
4. Estimation
5. Resampling

Cousin
34734140
Fun Pack™
Acrylic Beads
Perles en acrylique
Bolitas de acrílico
700 pcs/pz
WARNING: CHOKING HAZARD Small parts: Not for children under 3 years.
AVERTISSEMENT: DANGER D'ÉTOUFFEMENT Petits morceaux : Tenir à l'écart des enfants de moins de 3 ans
ADVERTENCIA: PELIGRO DE ASFIXIA No es un Jugete: Prohibido para niños menores de 3 años.

ACTIVITY

THE BIG REVEAL

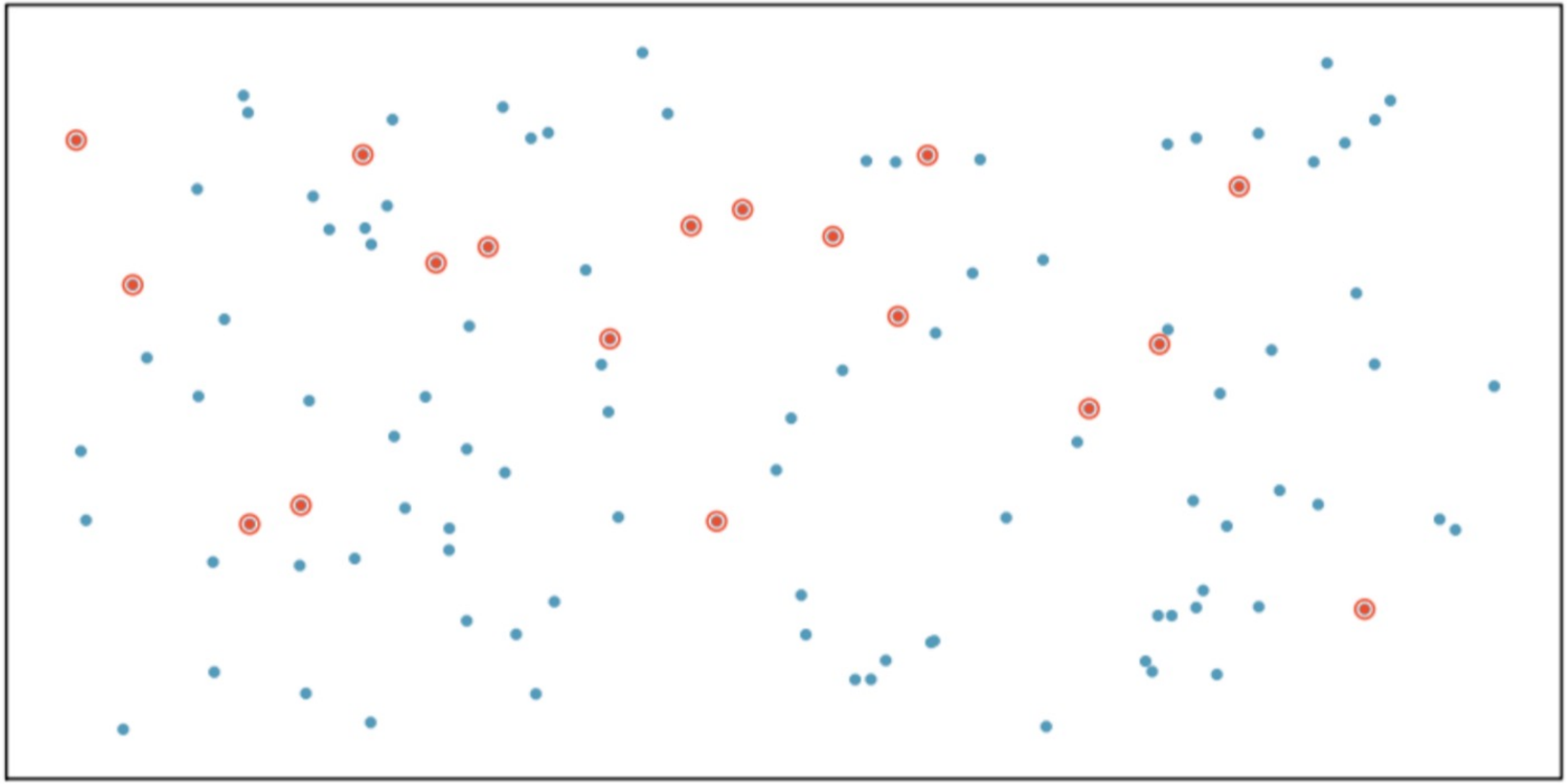
SAMPLING DESIGNS

PROBABILITY SAMPLING DESIGN

- **Simple Random Sample (SRS):** Consists of *n* individuals from the population chosen in such a way that every set of *n* individuals has an equal chance to be in the sample actually selected.

- **Steps:**
  1. Assign each individual a unique ID #
  2. Use a random number generator (there's one built into R!)

SIMPLE RANDOM SAMPLE

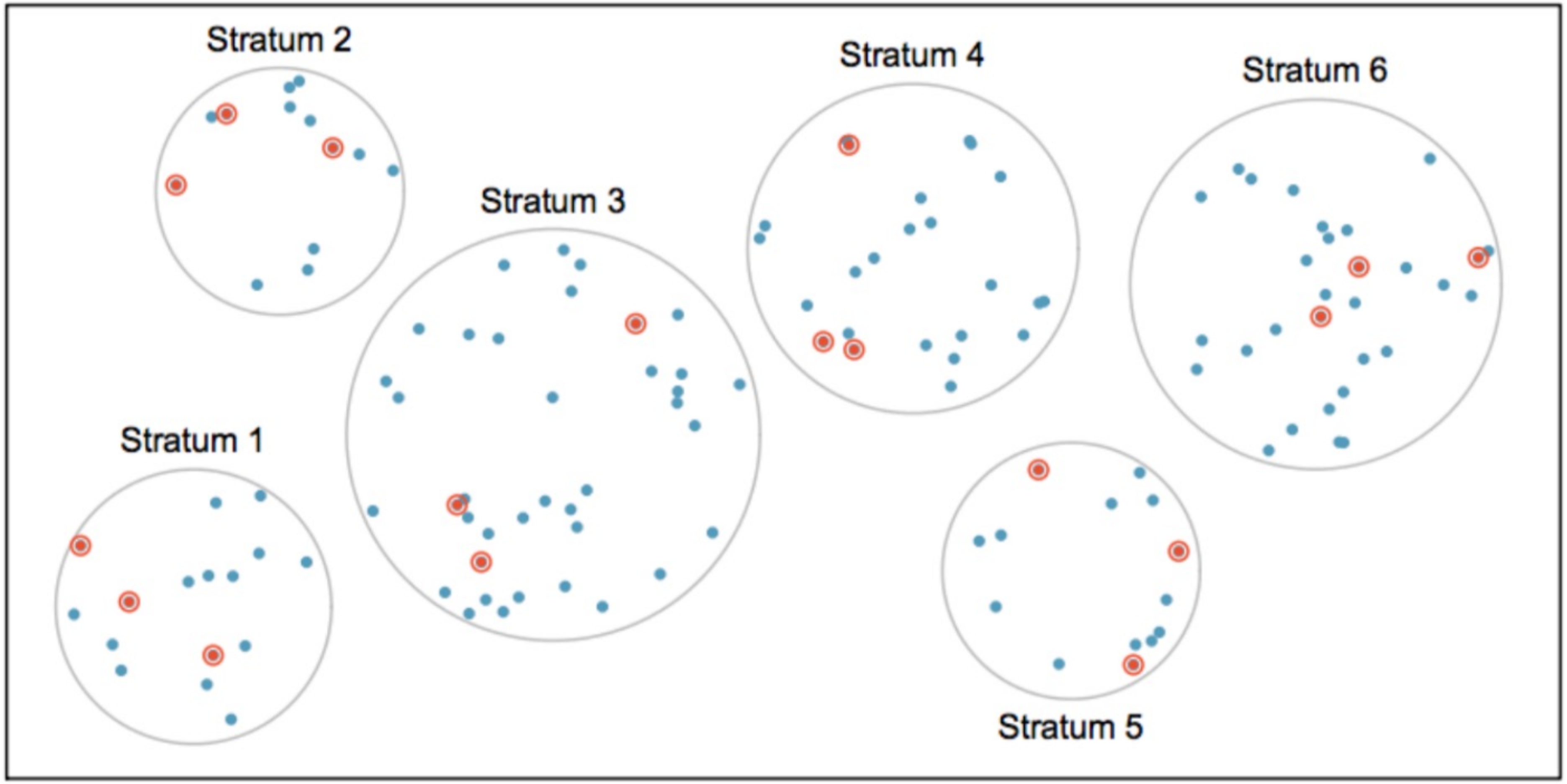


PROBABILITY SAMPLING DESIGN

- **Stratified Random Sample:** Individuals are first divided into groups of similar individuals (called strata) then perform an SRS within each strata. This ensures representation of each strata in the sample.

- **Steps:**
  1. Decided if there are similar groups such that groups may respond differently
  2. Perform an SRS within each strata

STRATIFIED RANDOM SAMPLE

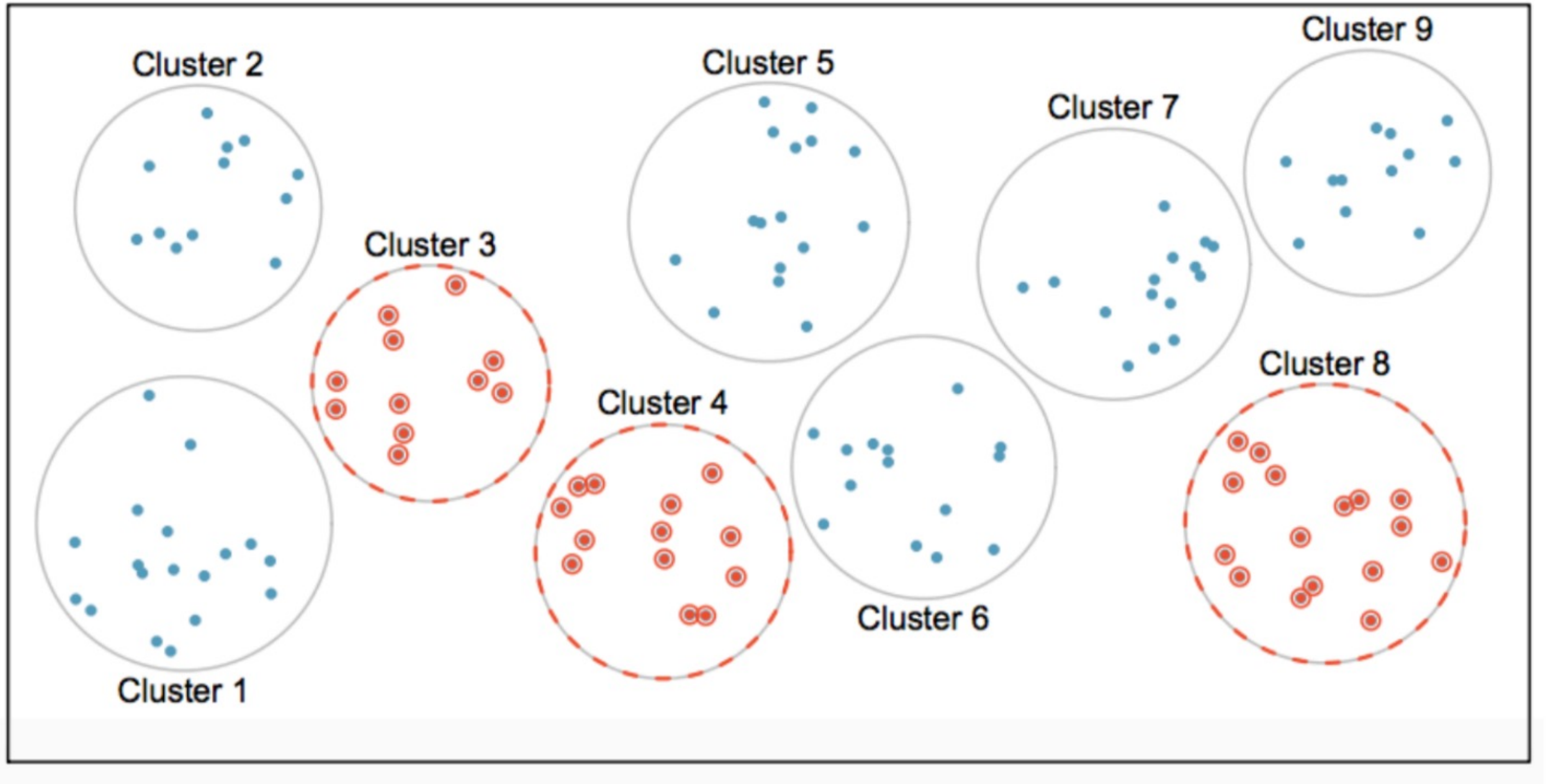


PROBABILITY SAMPLING DESIGN

- **Clustered Random Sample:** Sample clusters not individuals. Useful when there is variability within clusters but not between them.

- **Steps:**
  1. Assign each clusters a unique ID #
  2. Use a random number generator to select clusters
  3. Observe all individuals within a cluster

CLUSTERED RANDOM SAMPLE



EXAMPLE 4: BEAD ACTIVITY CONTINUED

Thinking back to the bead activity….

What kind of sampling design was used? Please explain

How might you use these values to obtain an estimate?

HOW TO SAMPLE POORLY

HOW TO SAMPLE POORLY: NON-RANDOM SAMPLING

- **Voluntary Sample:** a sample of units from a population that select themselves for inclusion.
  - Shows bias because people with strong opinions (often in the same direction) are most likely to respond.

- **Convenience Sample:** A sample of easily accessible units in a population

HOW TO SAMPLE POORLY: NON-RANDOM SAMPLING

- **Haphazard Sample:** A sample obtained when a researcher attempts to emulate a true chance mechanism or tries to pick a representative sample based on their idea of what the population looks like

- **Anecdotal Sample:** A very small sample of data collected based on life experience

EXAMPLE 5: STUDENT STUDIES

Imagine you're a senior psychology major conducting a study that examines procrastinating among Willamette students. How should you select a sample?

- **(A)**: Post a link to your survey on your Facebook page
- **(B)**: Get a list of Willamette student emails from the Registrar, take a simple random sample (SRS), and email that sample.

What sources of bias might affect each method?

GOALS OF STATISTICS

2 BRANCHES/TYPES OF STATISTICS

- **Descriptive Statistics (DATA 151)**
  - Summary statistics that quantitatively describe or summarize features of the data (such as mean, median, mode, range, variance, standard dev, frequency)
  - They do not allow us to make conclusions beyond the data we have analyzed
- **Inferential Statistics (DATA 152)**
  - Tries to draw conclusions (inferences) about populations based on (much smaller) samples
  - Data analysis used to deduce properties of an underlying probability distribution

DESCRIPTIVE STATISTICS

Questions to keep in mind:

- What is a typical value for the measurements?
- How much variation do the measurements possess?
- What is the shape or distribution of the measurements?
- Are there any extreme values in the measurements and, if so, what does that tell us?
- What is the relative position of a particular measurement in the group of data?
- What kind of relationship exists, if any, when there are two variables and how strong is the relationship?