# Data science for everyone

Prof. Jones-Rooy & Prof. Policastro

Feb. 19, 2020

4.1: Working with data

# ANNOUNCEMENTS

1. Lab 1 due today, Wed., Feb. 19, 8p
2. Lab 2 out today, Wed., Feb. 19, 8p
3. Homework 2 out Mon., Feb. 24, 8p
4. Homework 2 due Mon., March 9, 8p

**Gradescope:** It's slow! Sometimes it takes awhile. We were lenient for Homework 1. We will not be in the future. Also note the timestamp reflects when it is **uploaded**, but if you haven't selected pages by the time the TAs grade, it counts as late/you don't get credit.

Pay attention to this feedback on formatting!

**Lab 0**: Grades are released. Everyone who submitted correctly (e.g., nb-grader and Gradescope) got 1/1. You grade is on Gradescope, the feedback is on nb-grader. See feedback for correctness of answers (no points off, but look for that feedback). Your TAs will walk you through this in section this week.

You will be held to this once this feedback is delivered!

# Outline

1. Samples v. populations
2. Measurement
3. Evaluating data

# POPULATIONS

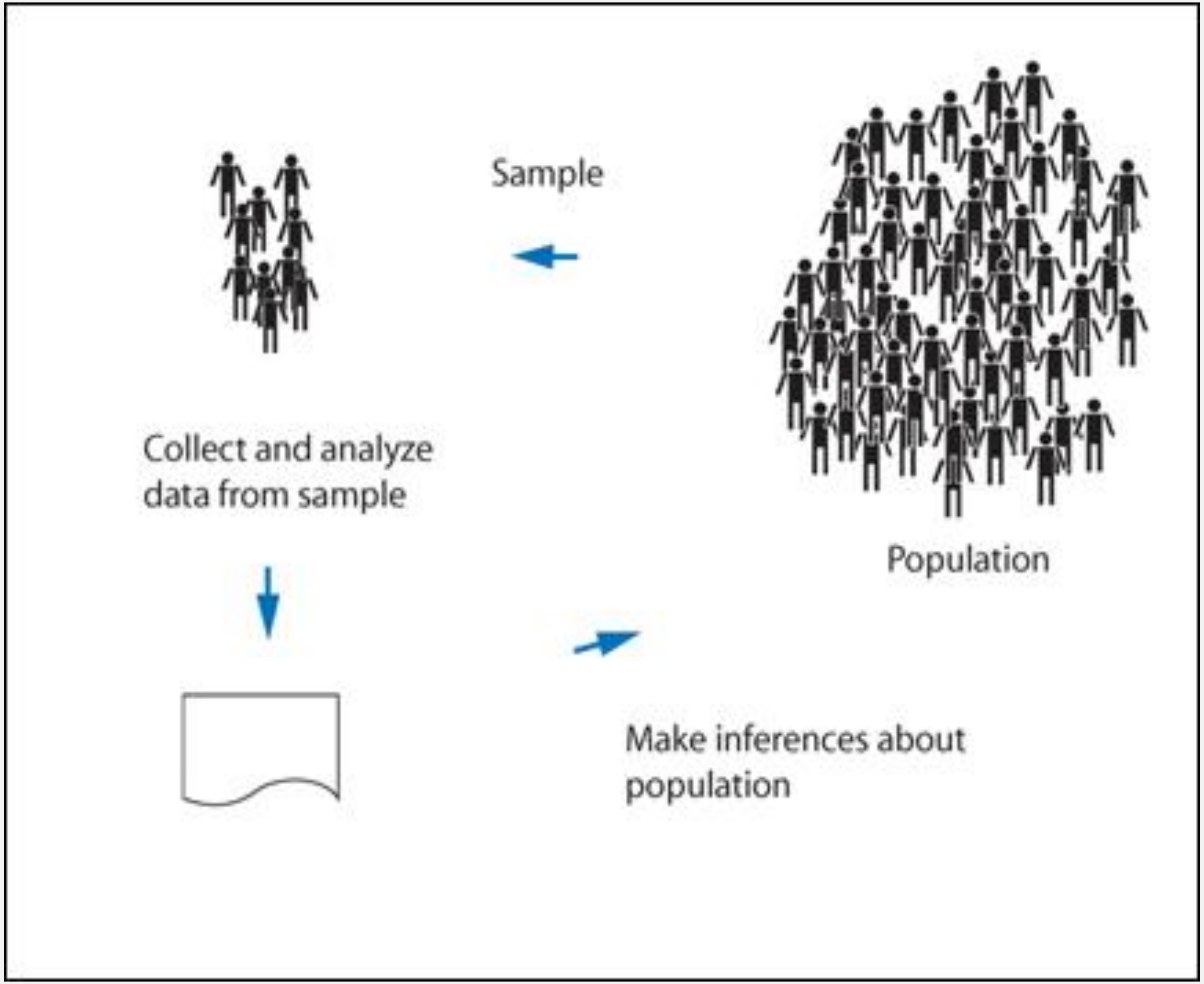
- Population = the universe of cases we want to describe, understand, or predict
- For example:
  - How will every voter in the US vote (and will they vote) in the 2020 presidential election?
  - What is the true average income for all people in India?
- The thing we care about (voter behavior, income) is the **population parameter**

# POPULATIONS

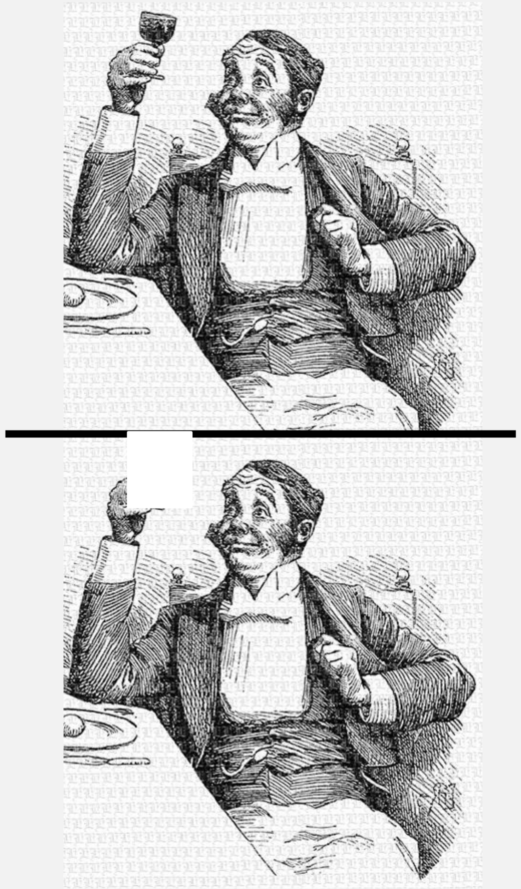
- Unfortunately, we can rarely study the population directly
  - Even a national census, which attempts to study the entire population, will not (likely) get the **entire** population
  - Some people may be systematically, or predictably, likely to not respond or complete the census
  - Does this mean we ignore the results? Or can we use our expectations about these systematic deviations to improve our interpretation of the results?
  - Hint: If it weren't the latter, data science would not exist!

# SAMPLES

- Because we can rarely observe the entire population, we use **samples** from the population, and use those to **estimate** the parameters of the population
- This means we can talk about how well the sample likely captures the true population
  - After the midterm: statistical uncertainty around our estimates once we have and analyze the data
  - Today: errors introduced, either systematically or randomly, during the measurement stage – i.e., as we turn the world into data in the first place
- As with randomized, controlled experiments, the gold standard in observational studies is a **random** sample that is large
  - Why large? Minimize the possibility that we've accidentally picked a few unusual observations (often we will call them **outliers**) that don't accurately reflect the population
  - How large is large enough? Stay tuned for the second half of the course!


Sample
Collect and analyze
data from sample
Population
Make inferences about
population

# RANDOMNESS



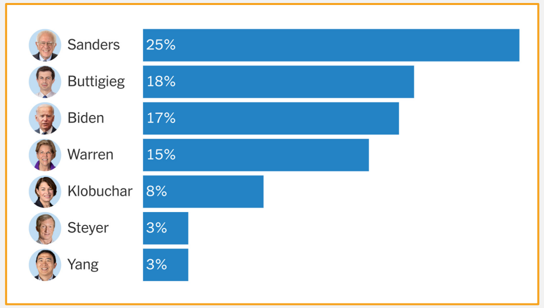
- Randomness: Units are chosen in a **non-deterministic** way, by **chance**
- In experiments, we saw that we want people to be randomly assigned to treatment and control groups
  - We don’t want them to self-select into these groups!
  - Largely because we expect there will be some differentiating factor that is **relevant** to our research that drives their preferences
  - For example, people who are more seriously ill are more likely to sign up to be in the treatment group for a drug trial
  - Why is that problematic for evaluating the effectiveness of the drug?
- **Computers** are very good at helping us generate random numbers and random samples

# SAMPLING FROM A POPULATION: EXPERIMENTAL

- We want to select a (sufficiently large) random sample from a population for experimental and observational studies
- For experimental studies, this can mean simply randomly assignment units into treatment and control (computers can help us do this so we don’t have to do it manually)
- If we cannot randomly assign units to groups, we hope there is a natural experiment we can conduct – where units are sorted into treatment and control for reasons we think are **unrelated** or **orthogonal** to what we are interested in
  - Example: John Snow argues people live in different parts of Soho for reasons unrelated to likelihood of contracting cholera
  - Example: People live in communities in a region of India for reasons unrelated to access to a hospital; sudden access puts them in treatment and control

# SAMPLING FROM A POPULATION: OBSERVATIONAL

- For observational studies, we also hope we can randomly select units, but this isn't always possible
  - Example: For a sample of today's class, I could randomly select a subset by assigning everyone a number then writing a program to draw 30 random numbers between 1-180
- Or, sometimes we think we are selecting randomly, but are not
  - What if for today's class I selected students with an "a" in their first name?
- As with experimental studies, if we can't randomize, we try to select based on things we hope are **orthogonal** to things we care about, but even this is not perfect
  - Example: For political polling in the US, often respondents are selected based on randomization of phone numbers
  - What's a problem with this method? Is it truly random?
  - What if I randomly selected people by last name? What would that do instead?
  - How would these two different methods ultimately affect any conclusions I draw about likely voting behaviors of voters in the US




This is the kind of rigorous thinking (and eventually intuition) we want to develop when working with data

# SAMPLES & POPULATIONS

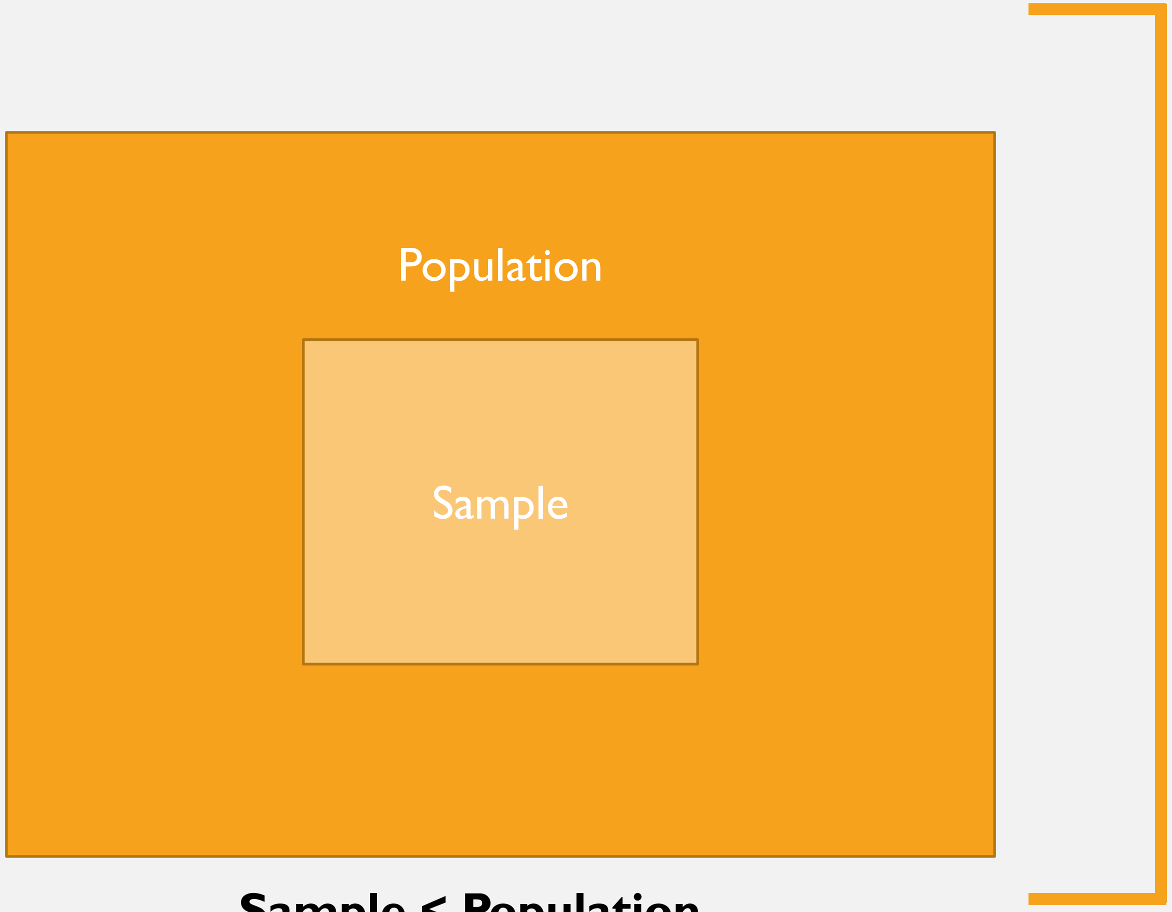
- Data is an imperfect snapshot of the real world
- In time and space

**Population**

The full collection of things (people, animals, plants, countries, etc.) you wish to study

**Sample**

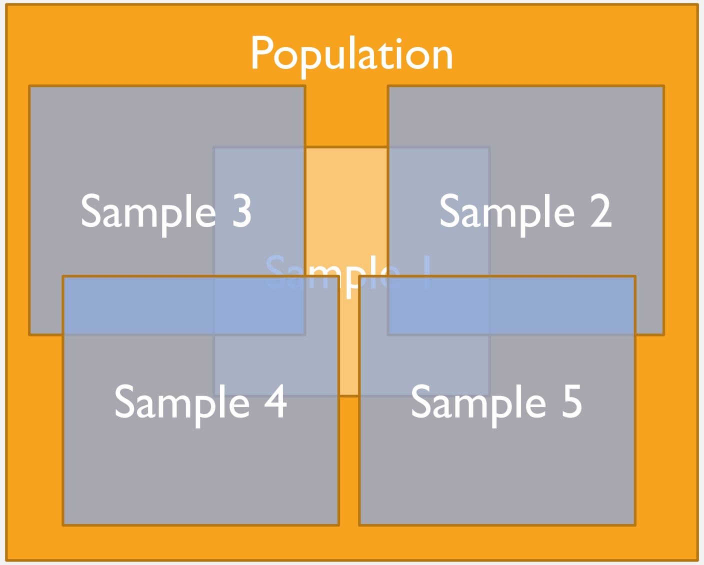
The subset of the population you actually study, due to time, resources, feasibility, other constraints




The precise moment the sample was taken

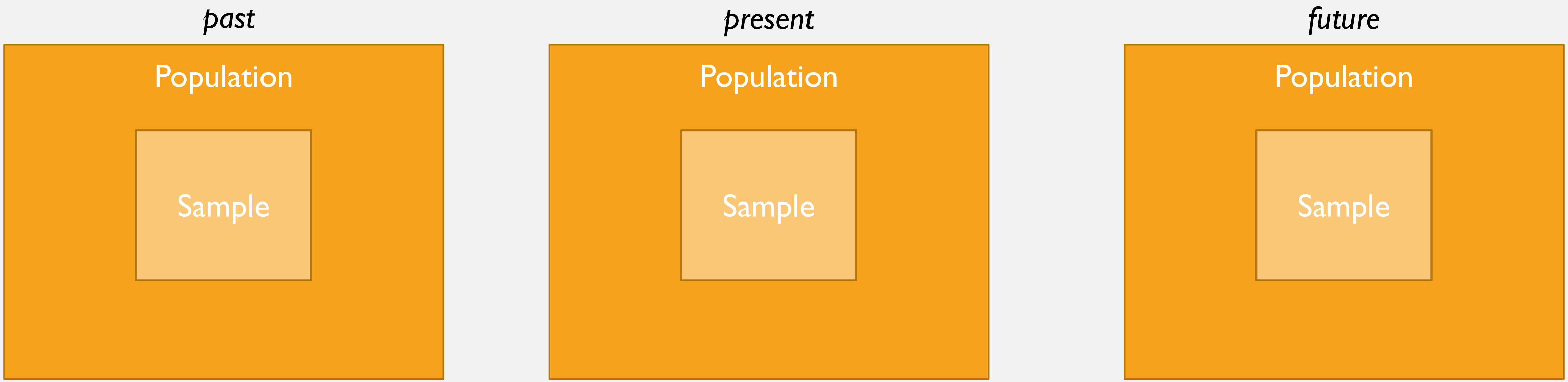
**Sample < Population**
**Sample = Population**

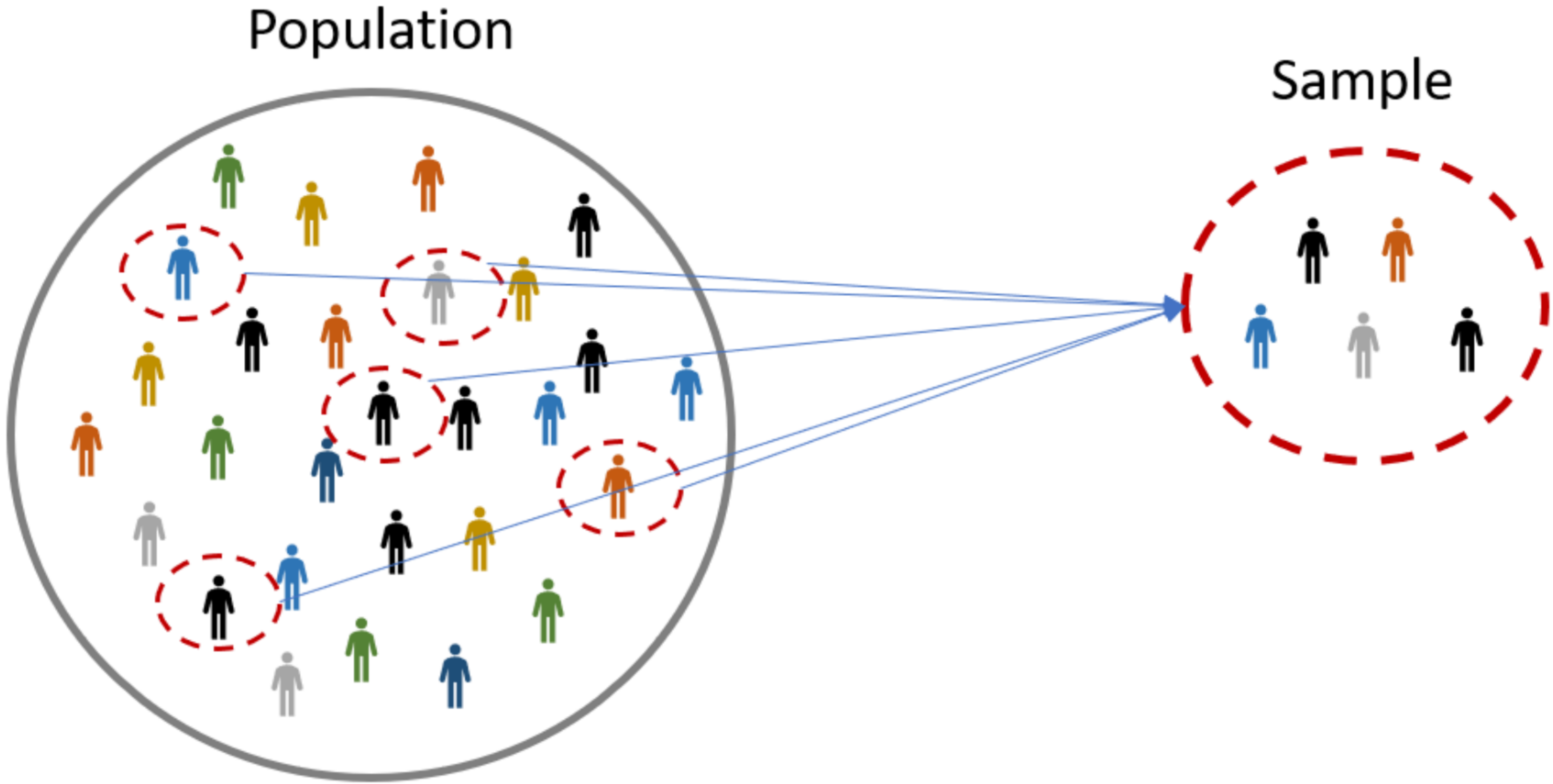
# Two questions we ask when working with data




1. Is my sample representative of the population of interest?

2. How robust is this sample over time?




Population
Sample

# Outline

1. Samples v. populations
2. $\boxed{\text{Measurement}}$
3. Evaluating data

# DS PROCESS

1. Observe the world
   - Be interested in something
   - Think something is worth studying
2. Turn it into data
   - Yourself
   - Find existing data
   - Combination
3. Turn data into discovery
   - Hypotheses & statistical tests
4. Turn discovery into insights
   - Share & communicate findings

**Ultimate goal: Turning the world into insights**

# TURNING THE WORLD INTO INSIGHTS

*Fundamentally human exercise*

Real world → Tables → Analysis → Sharing

**Real world**

Where to focus

Program for collecting

**Tables**

Measurement

Program for cleaning

**Analysis**

Hypotheses
Models

Confidence
Certainty

Program for statistics

**Sharing**

Clarity
Aesthetics

Program for visualizations

# MEASUREMENT

- Measurement = the act of turning the world (Truth) into data
- Humans decide what to study and how to study it, and whether it's worth studying in the first place
- All of these inject bias, subjectivity, and normative ideas into data, plus the chance for (hopefully random) accidental errors during data collection or processing
- This is what we are referring to when we talk about data not equaling Truth

# Truth → Measurement → Data


Only some of it ends up as data

The question is: What is that "some"? Why is it in there rather than other data, and what does that mean for our inferences?

# TWO STEPS IN MEASUREMENT

## Conceptualization

What you mean by the thing you're interested in

What do I **mean** by the **concept** that I'm trying to understand?

## Operationalization

How you're going to measure the concept

How am I going to **count** and **record** my concept of interest?

# EXAMPLE: DEMOCRACY

| Conceptualization | Operationalization |
|---|---|
| A country with regular elections | • A country must have at least two fair, free, and competitive elections in a row<br>• Some % of the population is eligible to vote, some % of the eligible pop. actually votes<br>• Ex ante uncertainty, ex post irreversibility, repeatability<br>• Opposition is allowed, multiple parties are legal, more than one candidate competes |
| A country with a free press | • How would you operationalize a free press? |

Exercise:
Try it yourself for something you care about!

# HOW DO YOU KNOW IF YOUR MEASUREMENT IS **CORRECT**?

- Generally, there is no universally **correct** measure, especially as things become more abstract
- E.g., Ok, measuring height comes down to more of an operationalization question, like specifics on metric vs. English system, or whether we count while people are wearing shoes, or round up, etc.
  - Though we might wonder why someone is interested in height – they are probably trying to conceptualize something!
- But lots of things we care about are not obvious in terms of either conceptualization or operationalization
  - Success, diversity, health, good environment, maturity, wealth
- To evaluate whether your measure is any good, we look for lots of indicators that the data is "correct"; i.e., it is capturing what we think it's capturing, it's a representative and hopefully random sample, we're not missing something relevant to our inferences

# Outline

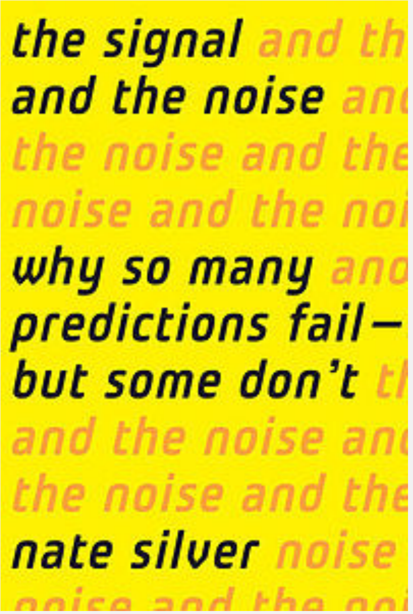
1. Samples v. populations
2. Measurement
3. Evaluating data

Optional further reading here!

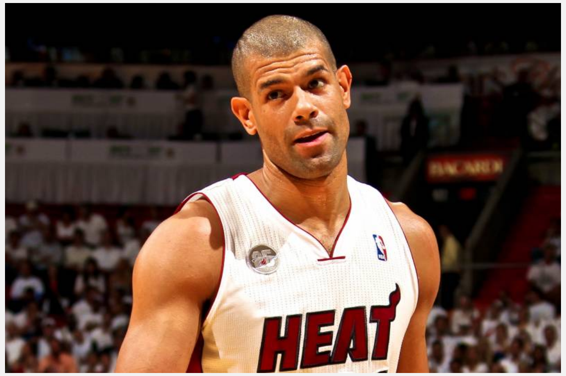
# EVALUATING DATA

| Random errors | Systematic errors | Errors of validity | Errors of exclusion |
| --- | --- | --- | --- |
| “noise" | 90% of the time this is just called selection bias | You may also hear about reliability | “Invisibility bias” |

# RANDOM ERRORS




- Measurement errors that are (we hope) are random, or **orthogonal**, i.e., unrelated to what we care about
- Here we go with sports!!!!
  - We want to measure how good a basketball player is
  - We conceptualize as contributions to the team
  - We operationalize as points scored in a game
  - Sometimes the player will have great games, sometimes horrible games, but we expect that that on average the good and bad games cancel out and the mean number of points per game represents something meaningful
- BUT! What's a problem with this **operationalization?**




Shane Battier

Here's a great [long read ](#) about this for you!

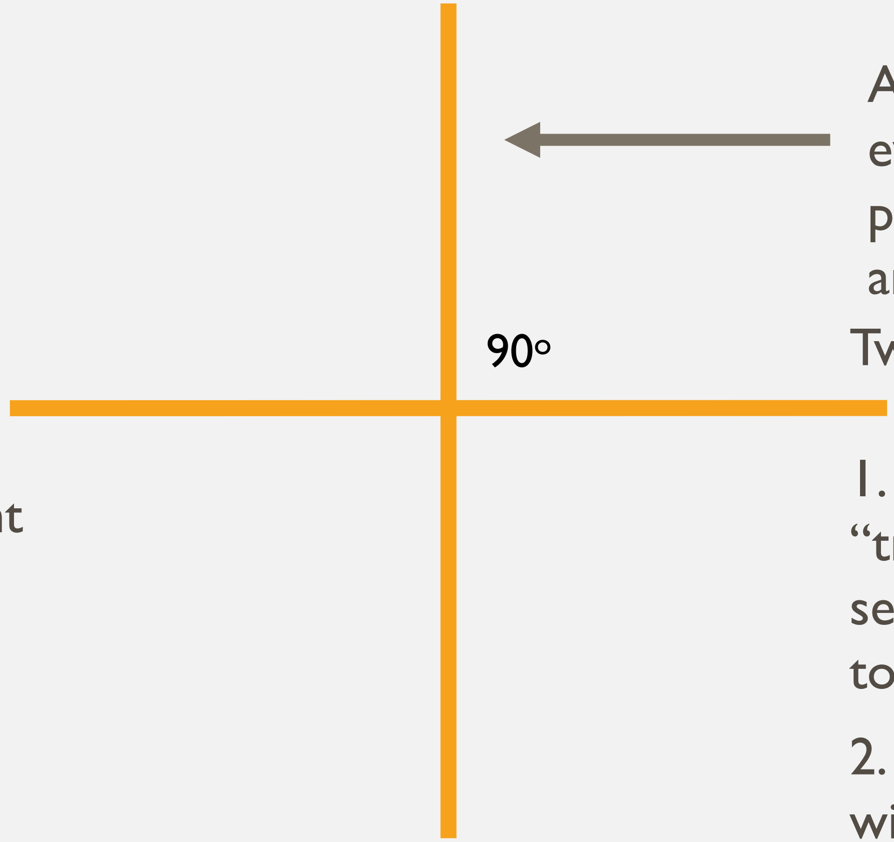
# Orthogonal

*Mathematics:*
Perpendicular

*Statistics:*
Independent

e.g., two variables are statistically independent

*Casual:*
Unrelated




A change in one does not even a little bit affect or predict a change in another

Two common uses in statistics:

1. Sampling! When you can't get "true" randomness, can you select on something *orthogonal* to variable(s) of interest?

2. Instrumental variables: To work with endogeneity in regression

# SYSTEMATIC ERRORS

- Selection bias
- Any time the sample is collected in a way that isn't random, or at least isn't **orthogonal** to relevant variables
  - If people in Soho had chosen to live near the Broad Street Pump because of anything actually related to their probability of becoming infected with cholera, John Snow's research would have suffered from selection bias
  - When pollsters in the US conduct polls and randomize by phone number, this is a good start, but there are still possible selection biases:
    - It rules out anyone who doesn't have a phone (in many cases they still use land lines only!)
    - It only picks up responses from people willing to **pick up** the phone and **answer** the question
    - People may lie about who they support

This is a great show about inference in data science!

# SELECTION BIAS

- In practice, one of the first things I look for when evaluating a study or a dataset
- Who ultimately ended up being included in the sample, how does that relate to what I'm trying to study, and **how does this bias my inferences**?
- Example:
  - Companies voluntarily publicly disclose diversity data
  - Those who do are probably more diverse than those who don't
  - This means the diversity numbers we see in the news are likely biased upwards; e.g., the sample of companies looks more diverse than the population truly is

# ERRORS OF VALIDITY

- Am I measuring what I think I'm measuring?
- Often this is related to conceptualization, but not limited to this step
- What am I ruling in or ruling out with this measure?
- For example:
  - If I measure democracy in terms of elections, am I accidentally including a bunch of authoritarian countries that happen to have rigged elections?
  - Or am I ruling out countries that I think are democracies but they aren't making the cut as I've defined it?
- Another example:
  - Companies look for top candidates, and aim to recruit people with good grades from good schools
  - Is this picking up **talent** or is it picking up opportunities to go to certain schools and a knack for test-taking?

# ERRORS OF EXCLUSION

- Invisibility bias
- Missing variables or members of a population due to lack of interest, lack of perceived importance, or lack of metrics
  - Usually these three are related
- Examples:
  - Socially excluded groups (Dalit, Romani, incarcerated populations)
  - Certain diseases (HIV/AIDS)
  - Informal laborers
  - Less tangible skills that contribute to success in a workplace
  - Aspects of diversity, like disability or chronic illness, or religion

# WHAT TO DO ABOUT ALL THESE?

1. Be aware of them
2. Evaluate every data set you work with in terms of how it performs through these lenses
3. Ask yourself: How and in what direction would this bias my results?
4. When to throw out data: When problems are so pervasive that you conclude you cannot trust the results
5. ART and SCIENCE!!

# Outline

1. Samples v. populations
2. Measurement
3. Evaluating data


Done for now, and happy lifetime of measuring!