# Unit 6: Data Collection

# Sampling and Bias

Lesson 50

Derek Ruths

# Overview of unit

Objectives:

- Understand why data collection is a time-consuming part of a project
- Get experience working with API collection
- Become familiar with standard data storage solutions: JSON, csv, and sqlite
- Understand how bias arises and why it's hard to detect

1. Overview
2. API-based collection
3. Reddit data collection
4. JSON
5. Scraping
6. Sampling

# Lesson overview

## Objectives

- Understand when sampling is needed
- Understand the problems that come along with sampling

## Outline

- Why take samples?
- What is bias?
- Strategies for avoiding bias
- The paradox of sample design

# Why take samples?

Sampling happens when you can't look at every item you want to study.

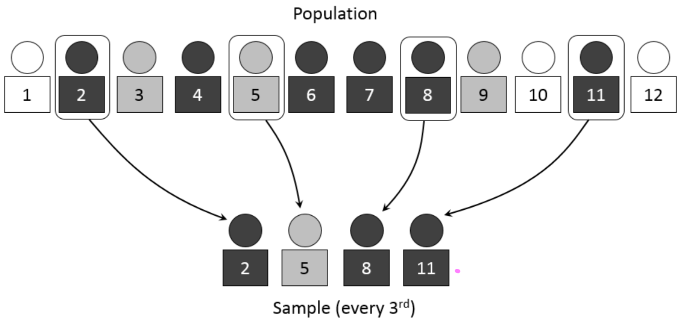
- How many friends do Twitter users have? ← cant collect all the data needed for thi

- What percentage of news articles posted this week mentioned the Marvels movie?

- How many bots posted to Reddit last month? ← eg what even is a bot

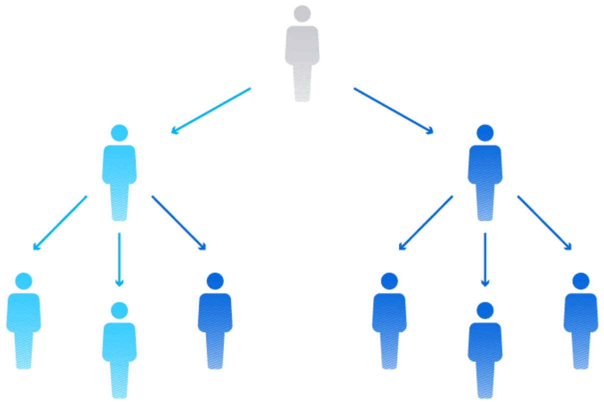
# Sampling Twitter users

Systematic sampling by ID



no control, (who are they, where are they/ are they active)

Snowball sampling



lack of diversity

explodes # of account

can do a random walk instead

# Sampling can produce “wrong” measurements

Bias is when we disproportionately weight factors that impact our measurement.

- How many friends do Twitter users have?
- What percentage of news articles posted this week mentioned the Marvels movie?

can't remove all biases from data
↳ can pick and avoid some

# Handling bias

↳ pick some biases to focus on

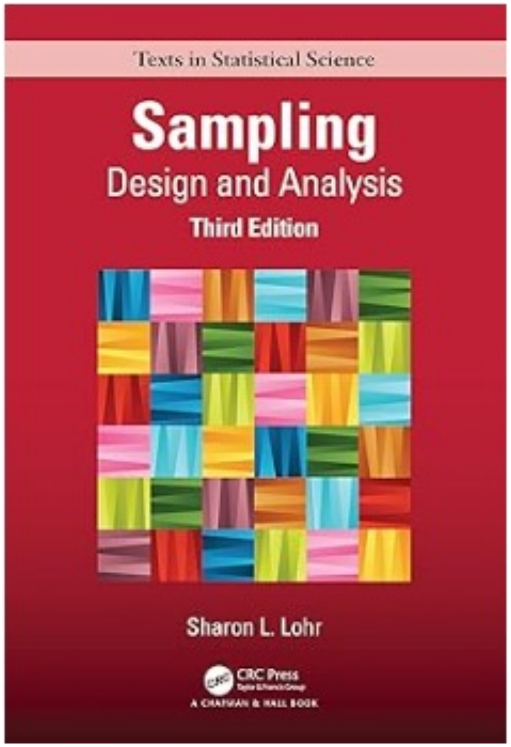
- What are the factors that matter within your context?
- Make sure your sampling method (and annotation and analysis) engages those factors and eliminate all others.

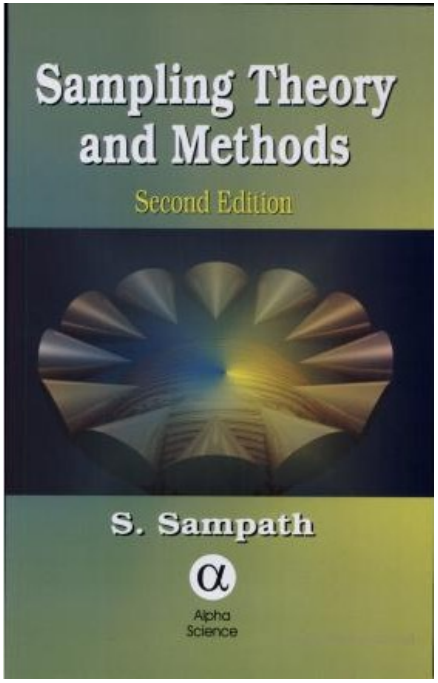
* be aware

* can you design it out

Sampling involves statistics!

# Statistics has a lot to offer





**MATH 525 Sampling Theory and Applications (4 credits)**

Overview
Mathematics & Statistics (Sci) : Simple random sampling, domains, ratio and regression estimators, superpopulation models, stratified sampling, optimal stratification, cluster sampling, sampling with unequal probabilities, multistage sampling, complex surveys, nonresponse.

what biases are we o/c with?

# Sampling hate speech on Twitter

need random controls

Statistically, it's very rare

often socially structured
↳ appears in pockets

method

biased toward chosen community

- problem accounts
↳ snowball sampling
↳ social similarity
↳ miss discourse communities
↳ committing accounts

- using keywords (eg slurs) → problems
↳ words w/ multiple meanings
↳ hate speech that doesn't use slurs
↳ miss hate in pictures
↳ code switching
↳ self-censoring
↳ misstypos

bias towards certain kinds of mistakes (not actual hate speech) vs biases against hate speech that doesn't use keywords giving a sampling method, what are we biasing towards

# The paradox of sample design

We're sampling X to know more about it.

To sample properly, we need to know something about X.

designing a collection method implies we know something about X

**Spend time studying the characteristics of your data!**

more perspective ⇒ better results

# Lesson wrap-up

## Takeaways

- All measurements are biased.
- We can limit certain kinds of bias.
- Always, we should be aware of the bias we want to avoid.
- Spending time with your data is necessary to understand potential sources of bias.

Need diversity!