

What is a Dataset?

Part 2: Collecting Data

INFO-1301, Quantitative Reasoning 1
University of Colorado Boulder

September 7, 2016

Prof. Michael Paul

Prof. William Aspray

Administrivia

Quiz 1 on Friday

- Covers everything up to and including today
- Problems will be similar to homework
- Review lecture slides online, plus readings
 - Be comfortable with the exercises in the book
- Office hours this week (ENVD 207):
 - Wed. 11am-noon: Prof. Aspray
 - Thurs. 10am-noon: Prof. Paul

Overview

This lecture will...

- get you thinking about where data comes from,
- and introduce concepts of populations and sampling.

How to collect data is a huge topic – you could take an entire course on it. This is just a starting point.

Data collection: an example

‘Spanish flu’ of 1918

- 20-50 million deaths worldwide
 - Precise numbers are unknown (due to lack of data)
- Not much known at the time about how to control epidemics
 - We know more now
 - ... thanks to years of data to aid our understanding

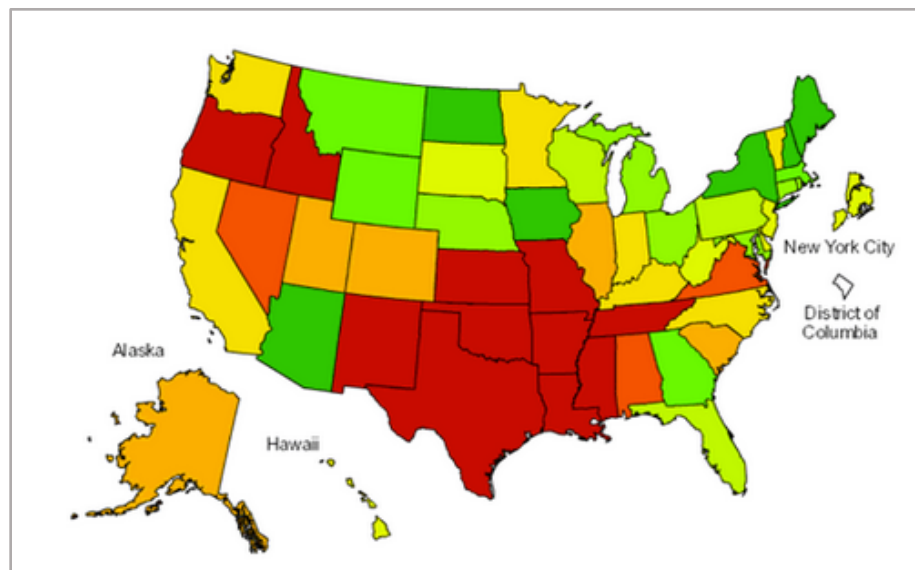
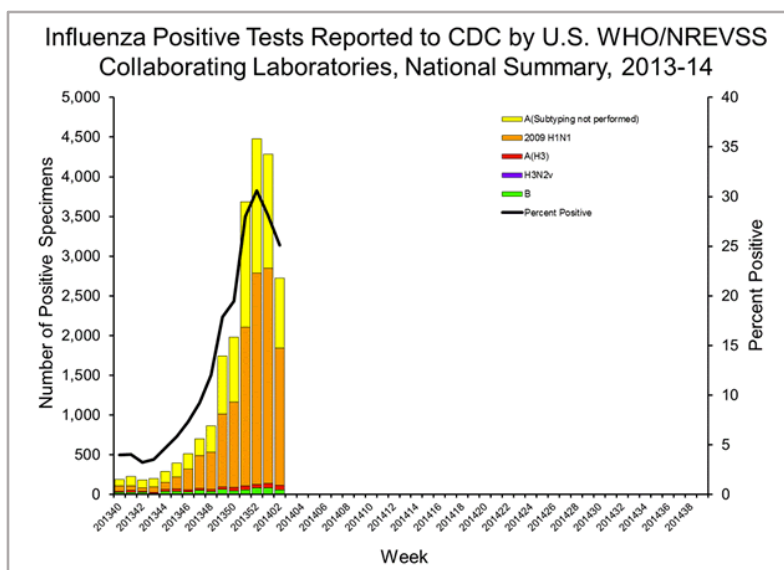
Data collection: 1918



Image from: <http://nyamcenterforhistory.org/tag/spanish-flu/>

This type of data is called **anecdotal evidence**

Data collection: 1980s-Present



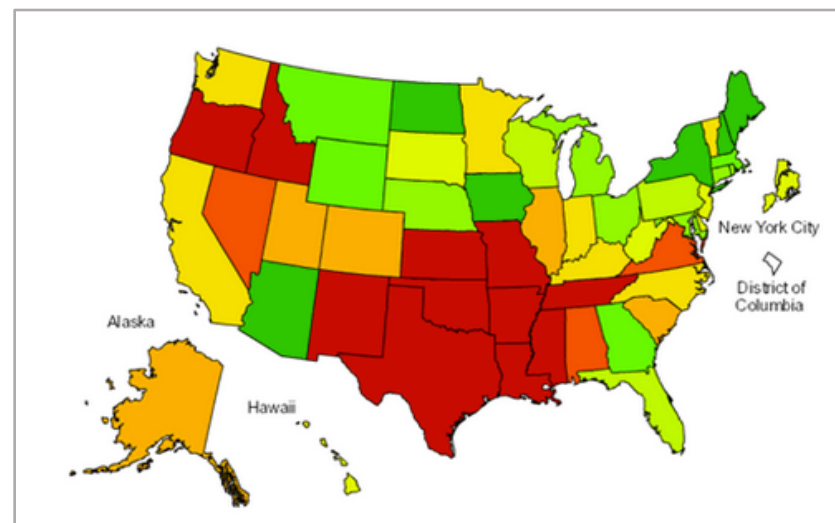
Flu cases monitored in depth by the federal government

- Data from the Centers for Disease Control and Prevention (CDC)

Data collection: 1980s-Present

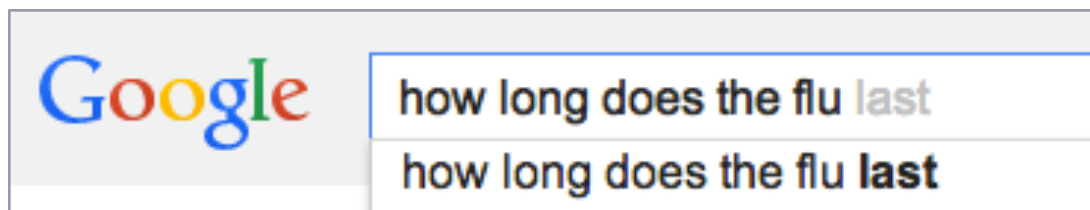
How does the CDC get this data?

- A number of healthcare providers across the country report numbers to the CDC each week
 - Approximately 50 clinics per state
- The CDC then has a snapshot of influenza in the US from the past week

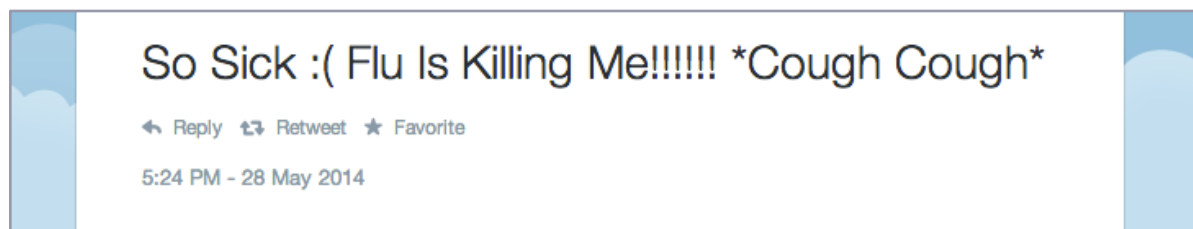


Data collection: 2010s-Present

Search queries:



Twitter posts:

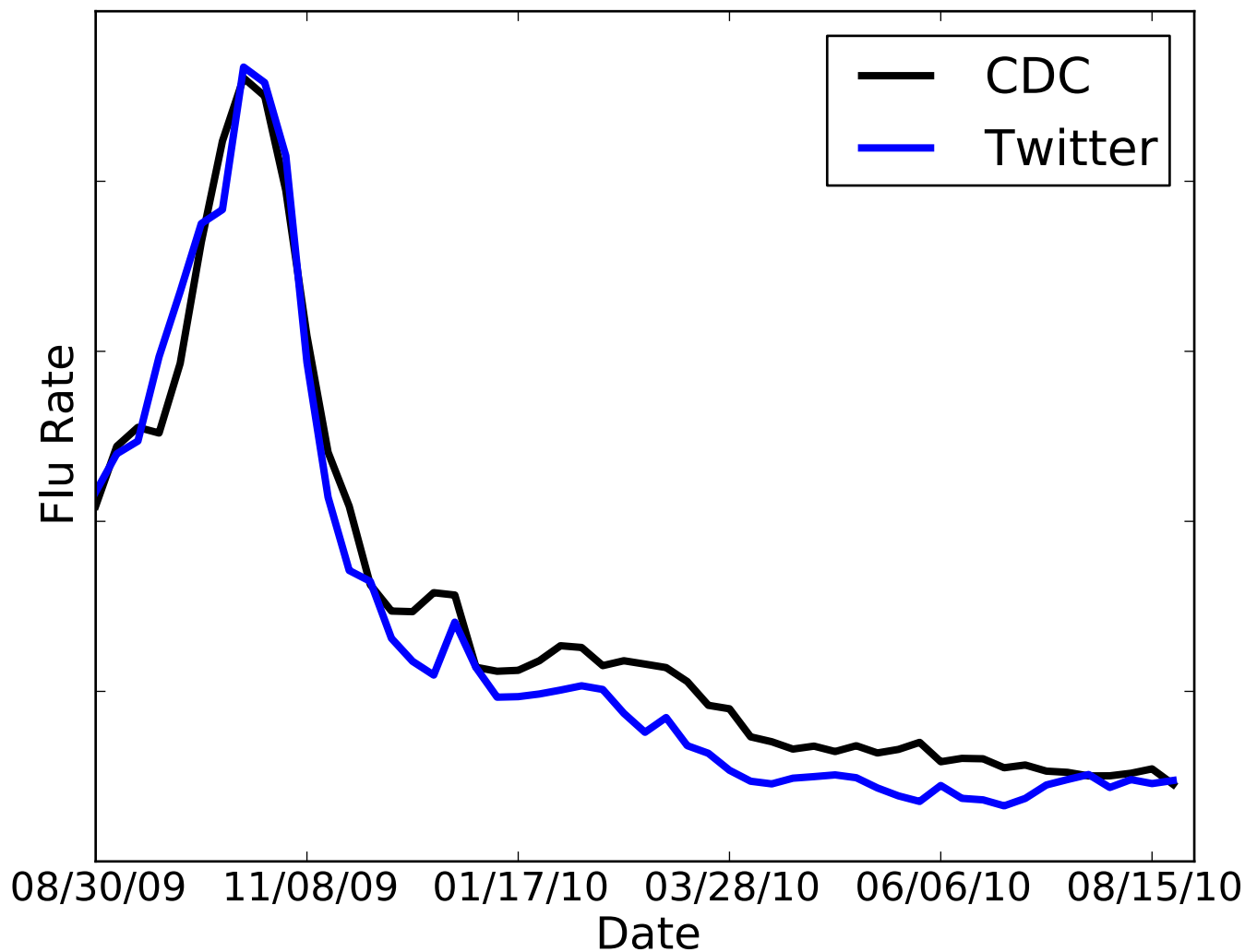


A recent innovation:

Internet data as an alternative to hospital data

- We know when someone has the flu because they said so online

Data collection: 2010s-Present



CDC vs Twitter

- Which is more accurate?
 - The CDC is accepted as the gold standard
- What does it mean to be accurate?
 - What we observe vs what is true
- Which is “better”?
 - Speed/cost vs accuracy

Using both data sources together is actually more accurate than only one of them alone

- Why? We'll think about it again later.

Populations

A **population** is a **set** of potential observations/cases

A **target** population is the population that is needed to answer a particular question

Example:

- Question: What is the average income of Colorado residents?
 - Target population: Set of all Colorado residents

Populations

Populations don't have to be people

More examples:

- What percentage of HP computers are defective?
 - Target population: set of all HP computers
- What is the average level of mercury in salmon?
 - Target population: set of all salmon



Samples

Sometimes it is impossible or impractical to collect data from an entire population

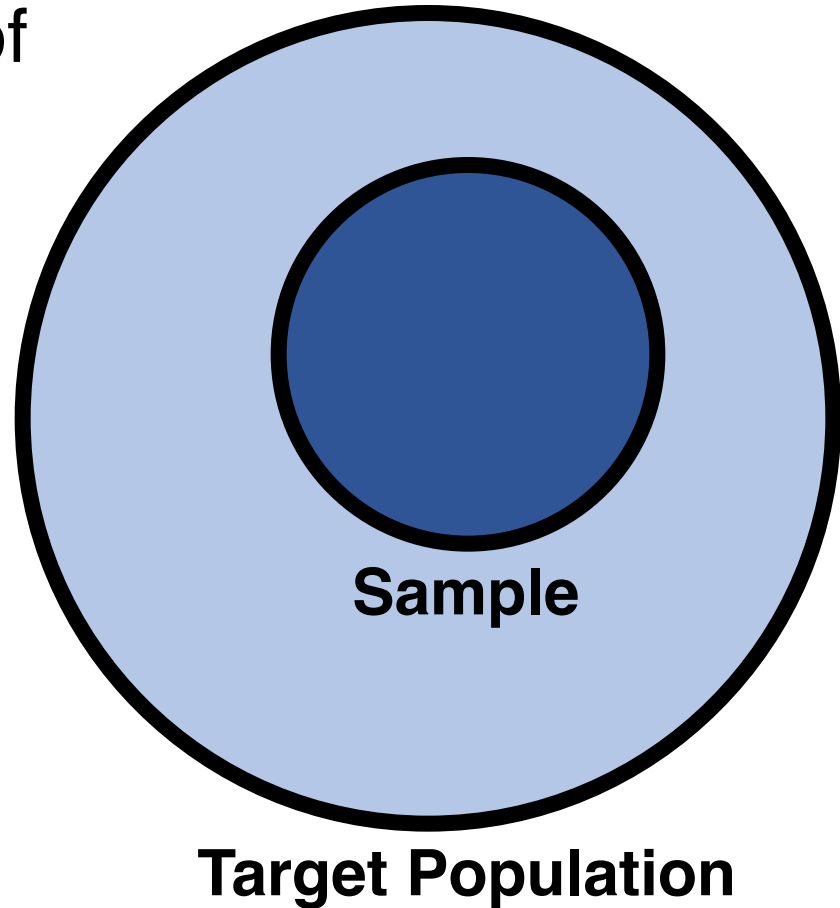
A **sample** is a **subset** of a population

Example:

- Question: What is the average income of Colorado residents?
 - Target population: Set of all Colorado residents
 - Sample: 1,000 randomly selected Colorado residents

Samples

A sample is a subset of the target population



Samples

Most datasets are samples

Common examples:

- Being randomly selected to give feedback to a company on a recent purchase
- Phone questionnaires from polling companies (e.g., to collect political opinions)
- Estimates of TV viewership or radio listenership

The process of collecting data about an entire population (no sampling) is called a **census**

Samples

Simple random sampling from the target population produces an **unbiased** sample of that population

A unbiased sample is considered **representative** of the target population

Statistics computed from unbiased samples are expected to be “close” to the population statistics

- We'll explain this more rigorously later in the course

Samples

The **sampling frame** is the set from which you sample

- It is a subset (or equal to) the target population
- Example: If you randomly sample residents from Colorado, the sampling frame is the set of Coloradans

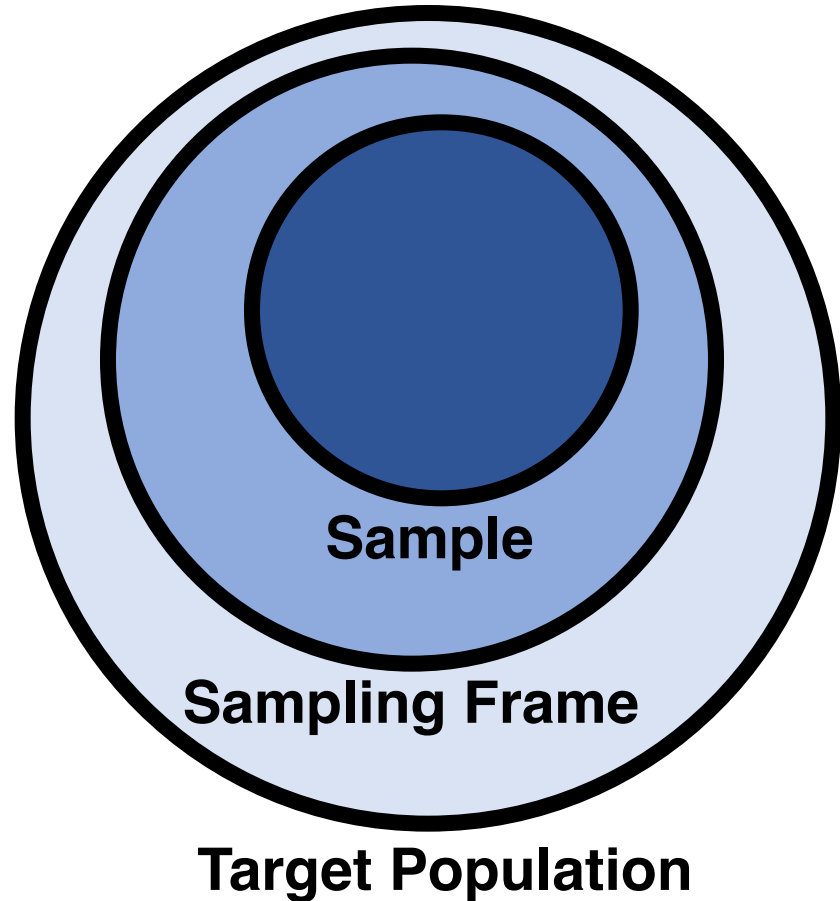
If the sampling frame is different from the target population, then the sample will be biased

- Example: You want to measure the average income of Americans, but you only sample people from Colorado

Samples

The sampling frame is a subset of the target population

A sample is a subset of the sampling frame



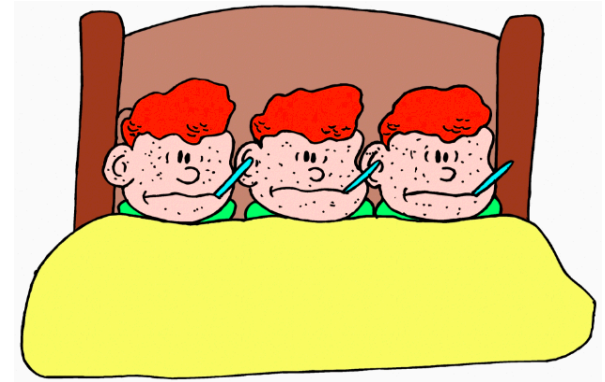
Returning to flu...

Research question:

- What percentage of Americans are currently infected with the flu?

Target population?

- Set of all Americans



Flu data: CDC

Recall: how does the CDC collect their data?

- A number of healthcare providers across the country report numbers to the CDC each week
 - Approximately 50 clinics per state

What is the sampling frame?

- People who have visited a U.S. healthcare clinic in the past week
 - Not exactly the same as the target population – not everyone with flu goes to see a doctor

Flu data: Tweets

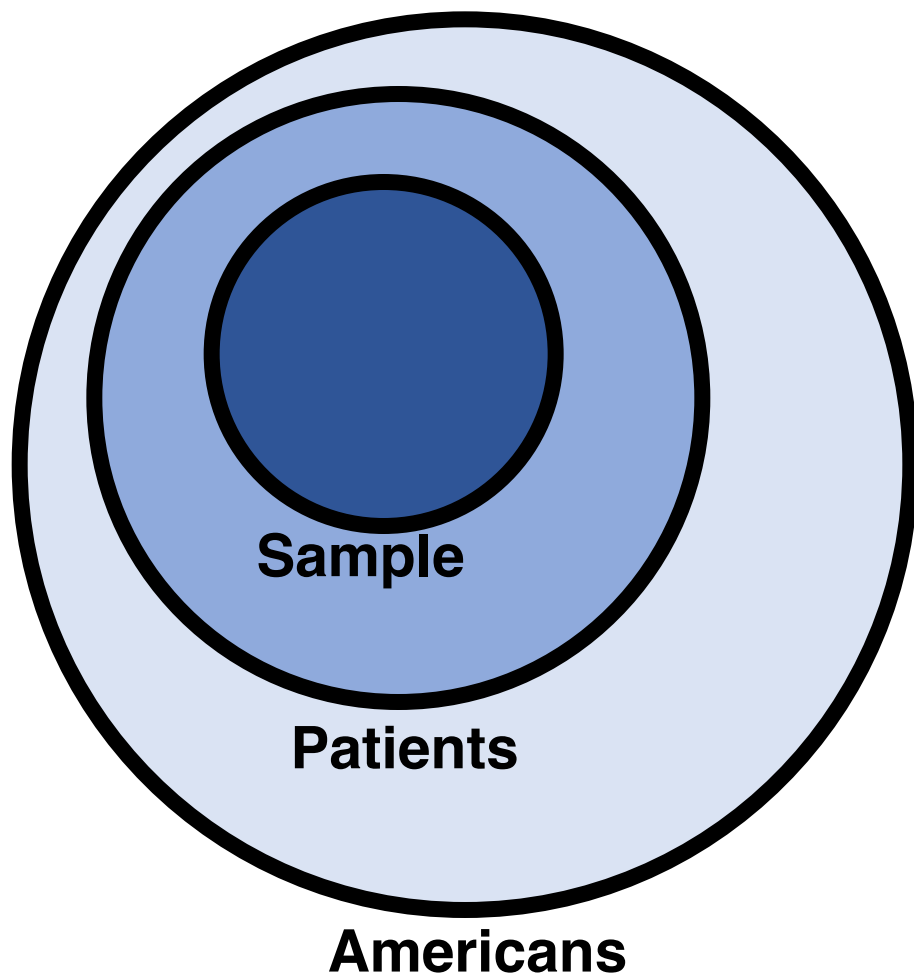
Where does the Twitter data come from?

- People tweet that they are sick

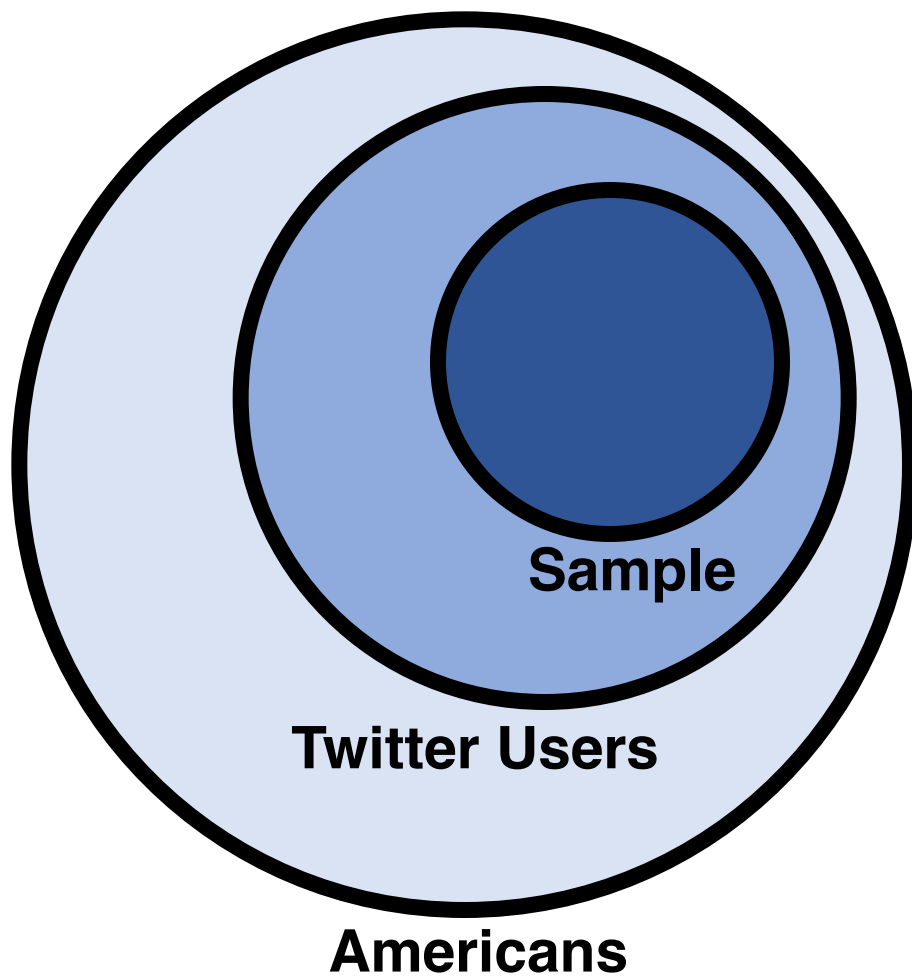
What is the sampling frame?

- People who use Twitter and choose to tweet about their health status
 - Clearly not the same as the target population, since many people are not included

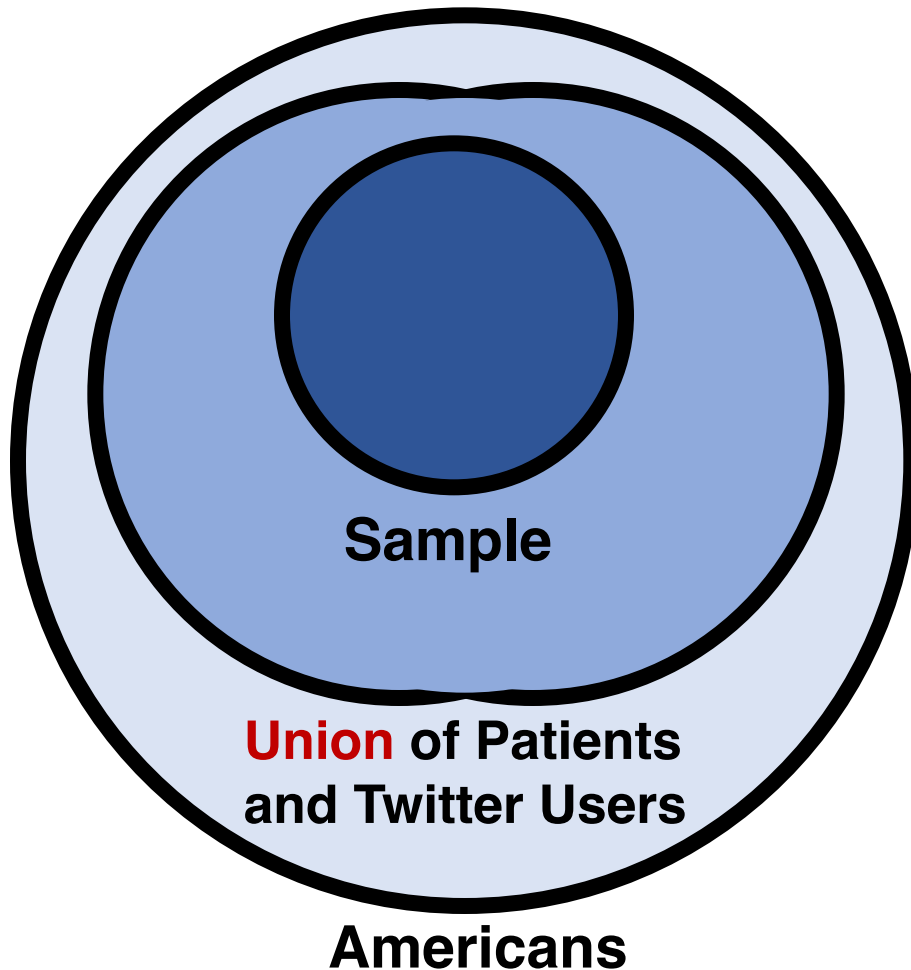
CDC data:



Tweet data:



Combined data:



The sampling frame is closer to the target population

- Less biased

This is why sampling from multiple data sources can be better than just one