NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 2: Lecture 1
Collecting Data – Observations and Surveys

NYU Center for Data Science

How can we summarize data? How can we visualize data to quantify information?

DS-UA 112
Introduction to Data Science

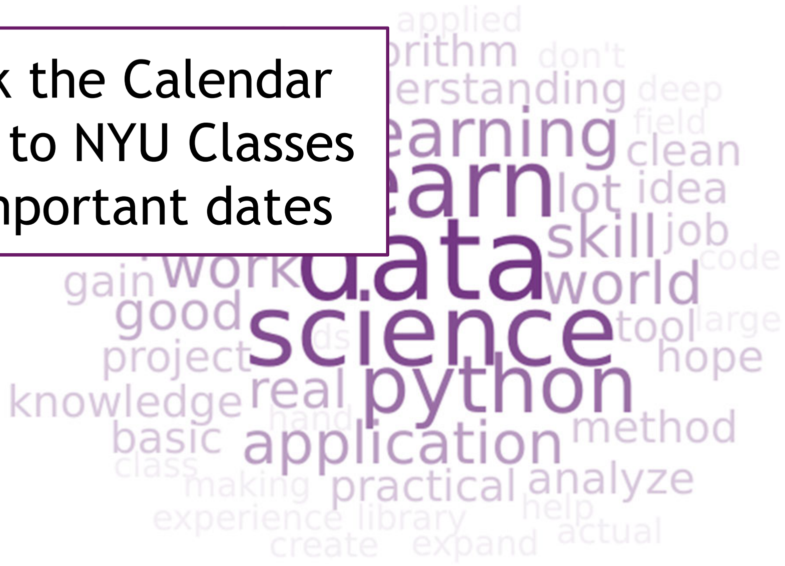
Week 2: Lecture 1
Collecting Data – Observations and Surveys

Adapted from Nolan, Lau, and Salganik

Announcements

- Please check Week 2 agenda on NYU Classes
  - Homework 1
  - Lab 2
  - Survey 1
- Remember to post to Piazza

Check the Calendar linked to NYU Classes for important dates



Review: What characterizes data?

- Good Properties
  - Volume
  - Velocity
  - Variety
    - Not Reactive

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

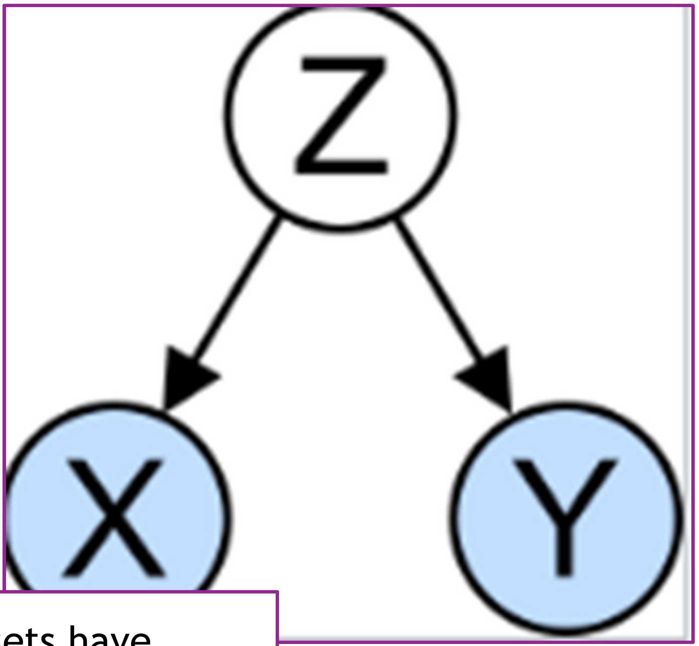
May 2011 | Report

By James Manyika,
Brown, Jacques Bug
Charles Roxburgh, a

Big data will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus—as long as the right policies and enablers are in place.

Review: What characterizes data?

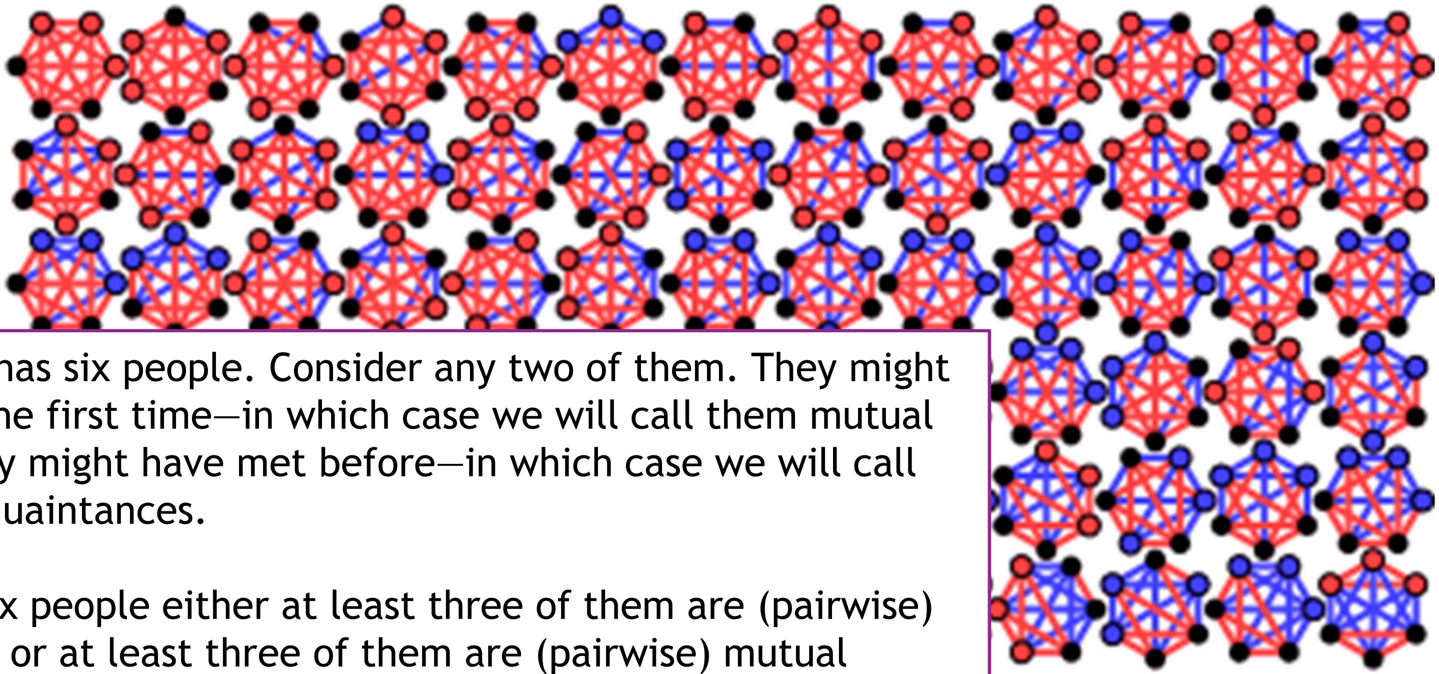
- Bad Properties
  - Messy
  - Drifting
  - Biased
    - Not Representative
  - Confounded



Confounded datasets have **lurking variables** that influences the patterns of cause and effect. Often lurking variables get in the way of data collection.

Review: Exercise

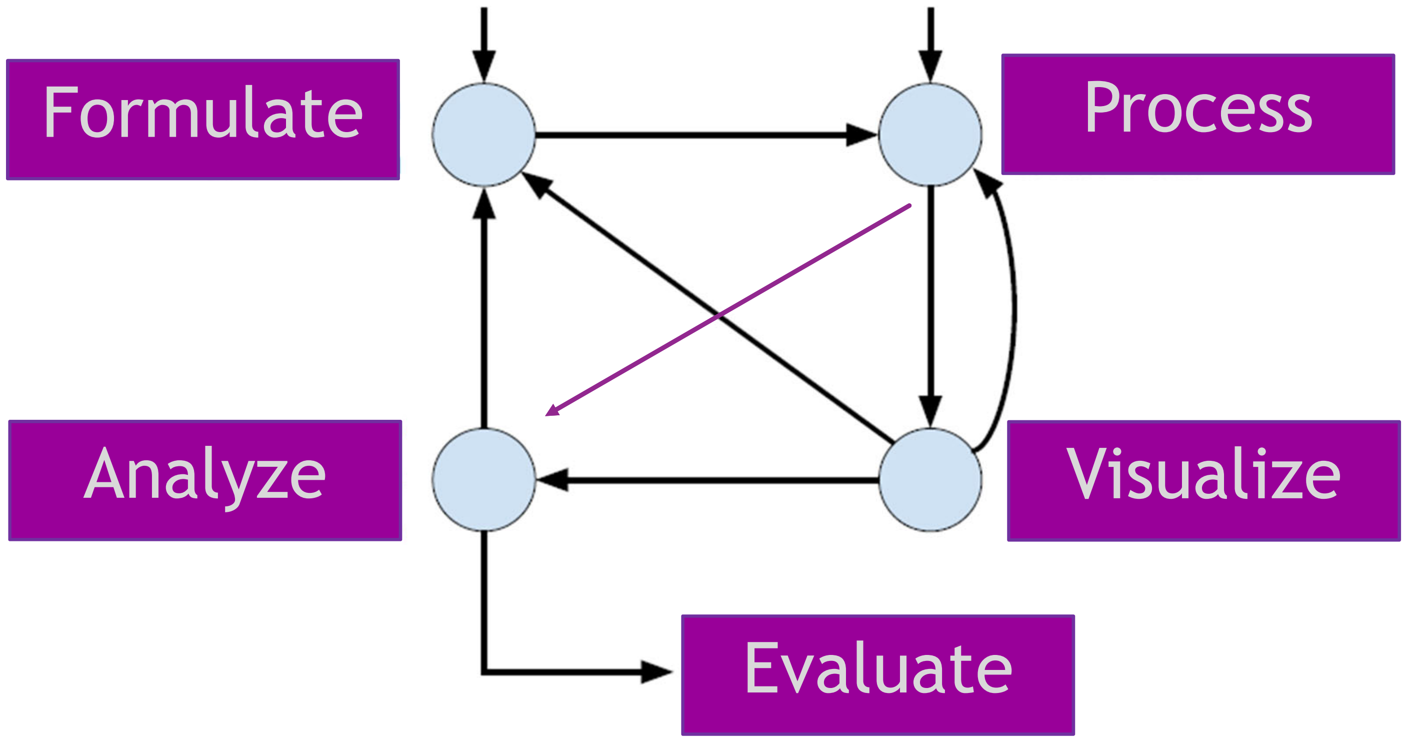
- Using Graphs with Colored Edges to Detect Patterns in Networks



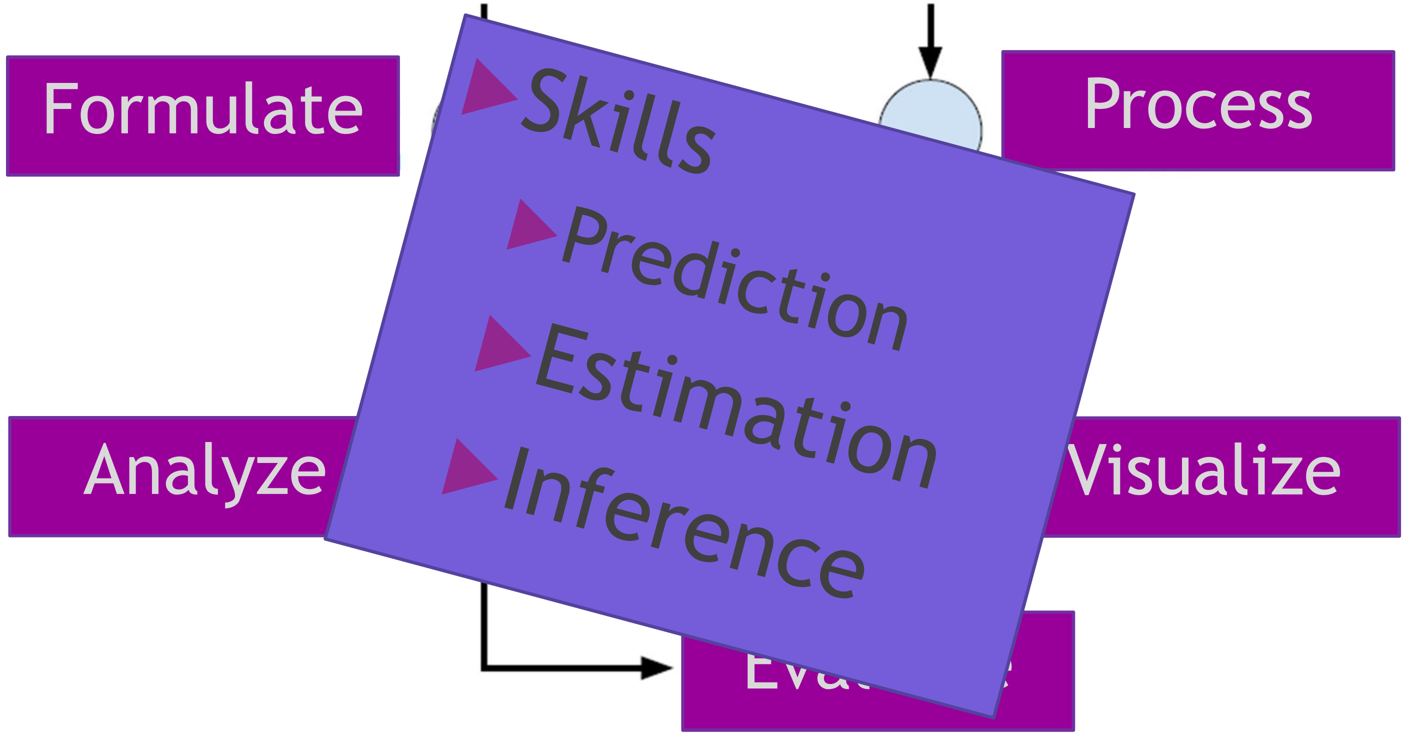
Suppose a party has six people. Consider any two of them. They might be meeting for the first time—in which case we will call them mutual strangers; or they might have met before—in which case we will call them mutual acquaintances.

In any party of six people either at least three of them are (pairwise) mutual strangers or at least three of them are (pairwise) mutual acquaintances.

Review: What is data science?



Review: What is data science?



- Skills
  - Prediction
- Estimation
- Inference

Agenda

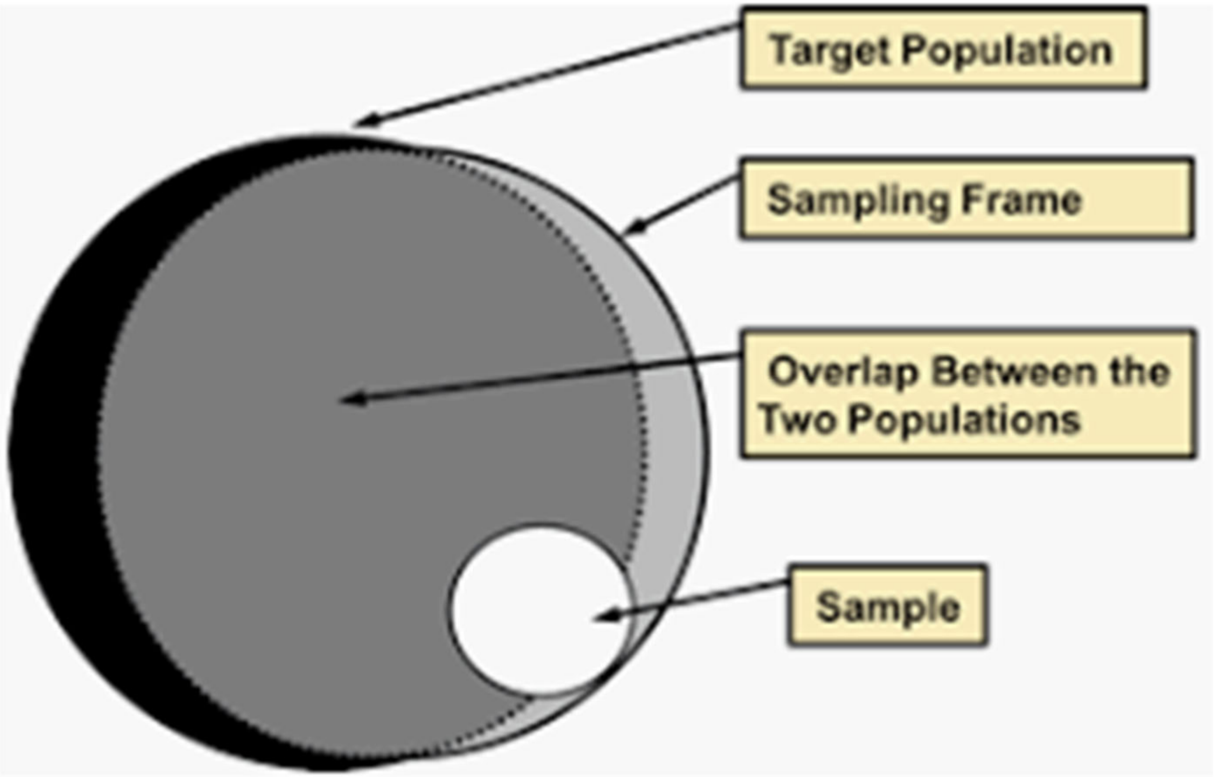
- Step 1: Formulate
  - Name, Age, Gender
  - US Elections
- Step 2: Process
  - Samples from Population
- Step 3:
  - Histogram
  - Line Chart

References

- Nolan, Lau, Gonzalez (Chapter 2)
  - https://cp71.github.io/textbook/intro
- Salganik (Chapter 3)
- https://developers.google.com/edu/python/exercises/baby-names

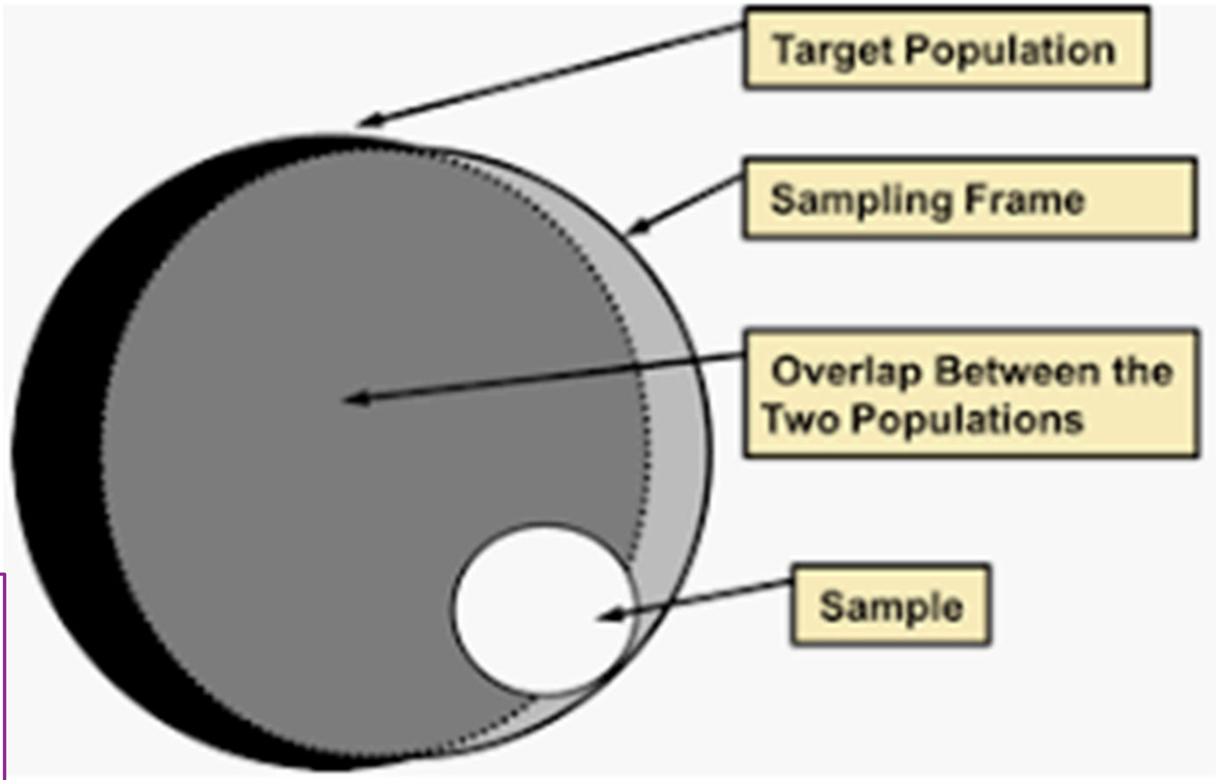
Population and Sample

- Population
  - Group of Interest for Study
- Sample
  - Subset of the Population observed or surveyed by the Researchers



Population and Sample

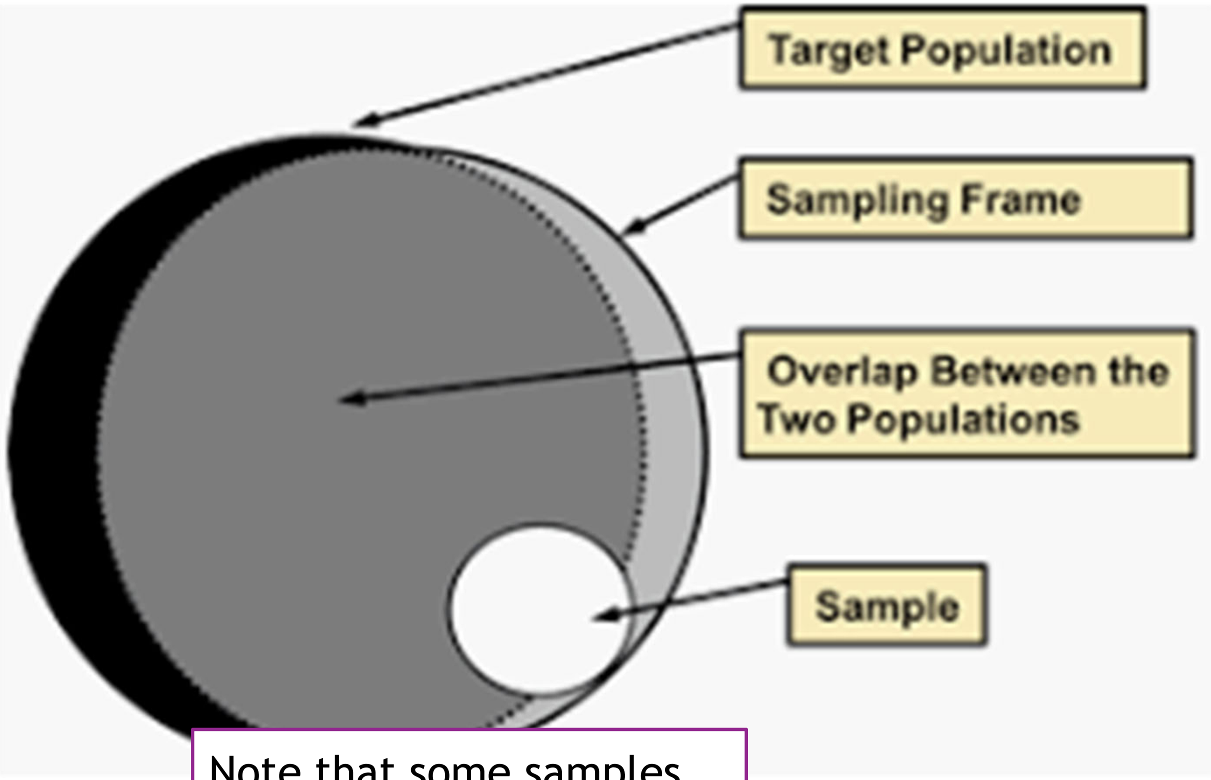
- Population
  - Group of Interest for Study
- Sample
  - Subset of the Population observed or surveyed by the Researchers



Not that non-representative is different from **incomplete**. Sometimes we cannot access data for the entire **population**. Instead we have a **sample**.

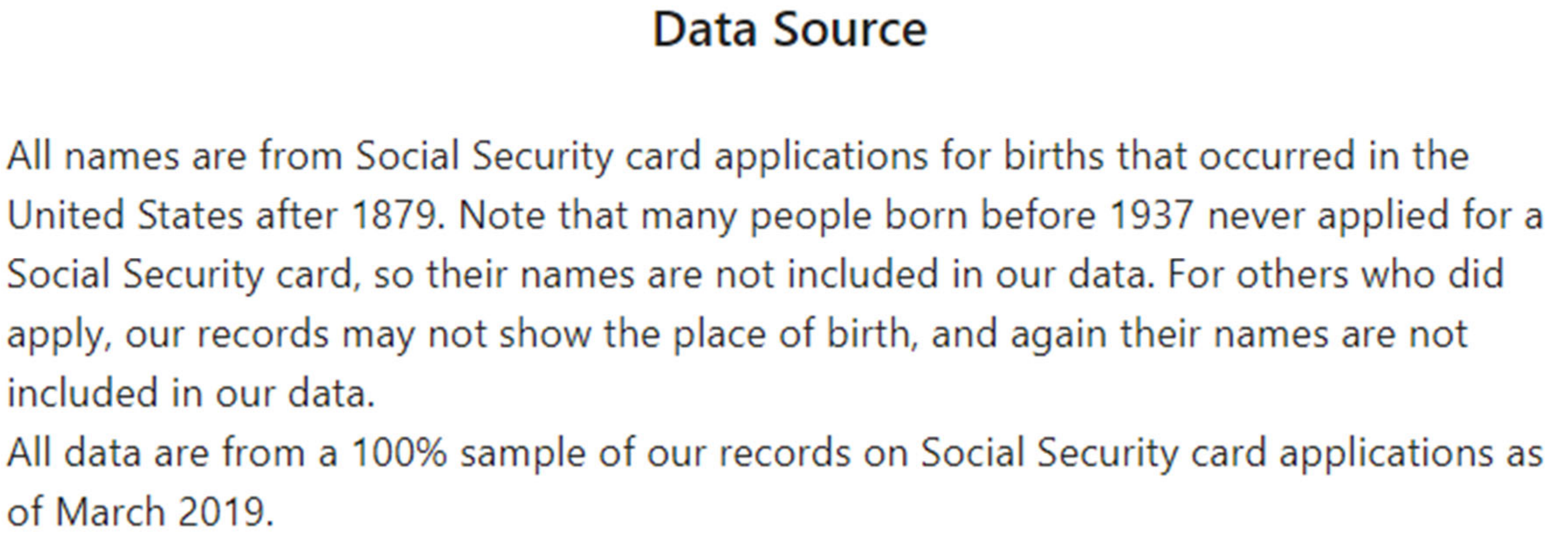
Population and Sample

- Sampling Frame
  - Members of Population Eligible for Inclusion in Sample
- Types of Samples
  - Census
  - Administrative Dataset
  - Self-selected
  - Convenience
  - Judgement



Note that some samples are random and some are non-random

Exercise

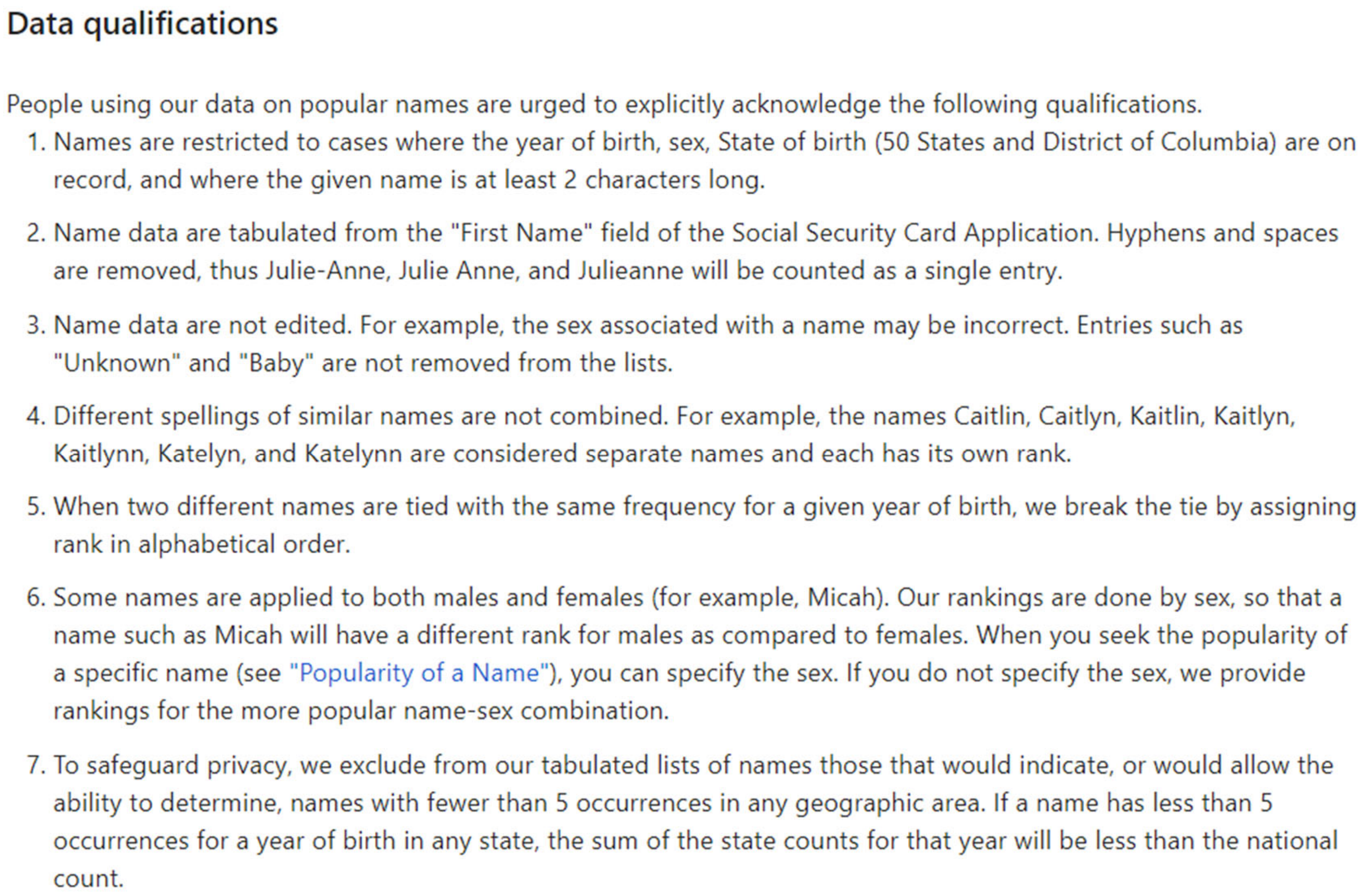


Data Source

All names are from Social Security card applications for births that occurred in the United States after 1879. Note that many people born before 1937 never applied for a Social Security card, so their names are not included in our data. For others who did apply, our records may not show the place of birth, and again their names are not included in our data.

All data are from a 100% sample of our records on Social Security card applications as of March 2019.

Exercise

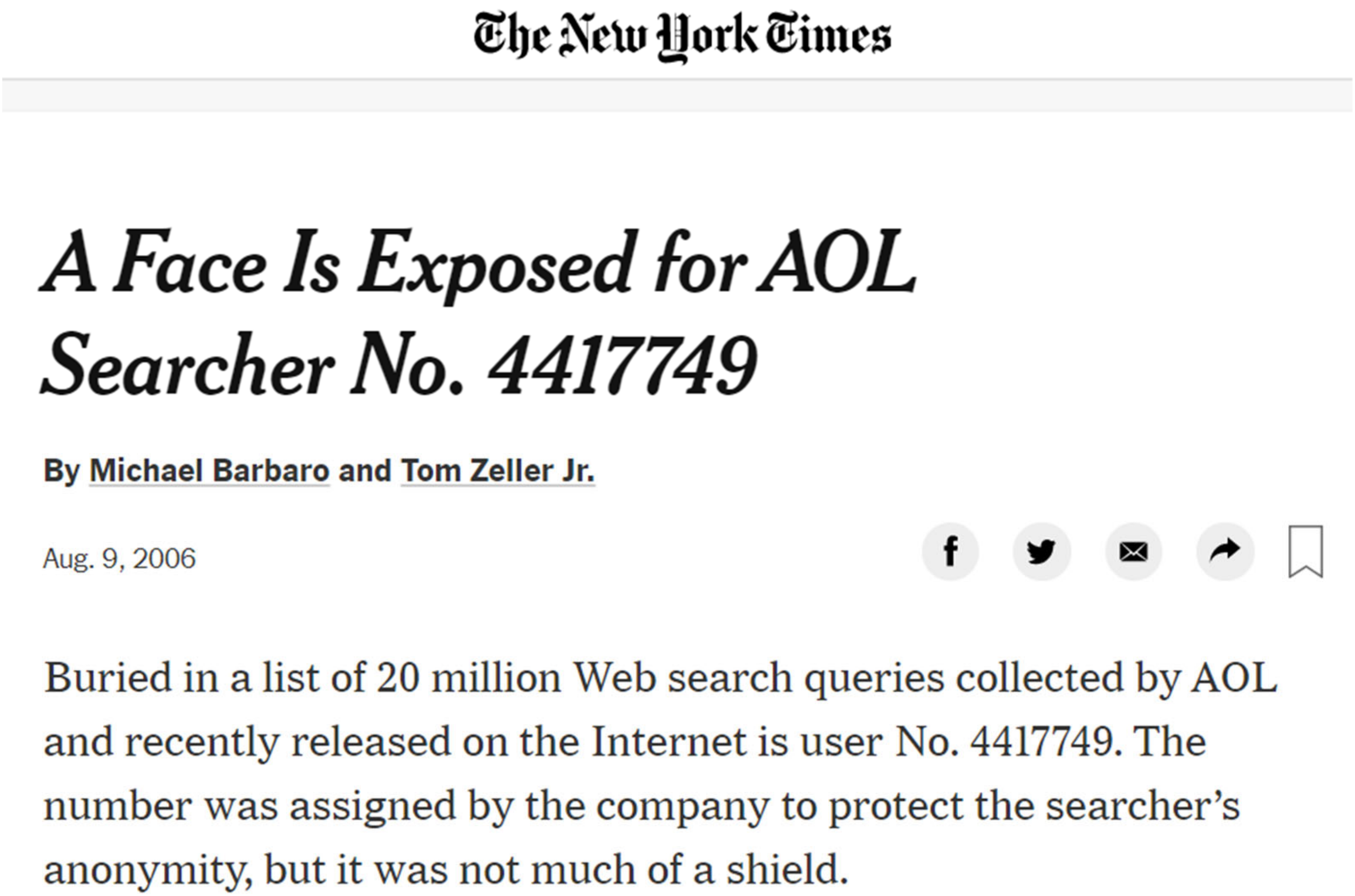


Data qualifications

People using our data on popular names are urged to explicitly acknowledge the following qualifications.

1. Names are restricted to cases where the year of birth, sex, State of birth (50 States and District of Columbia) are on record, and where the given name is at least 2 characters long.
2. Name data are tabulated from the "First Name" field of the Social Security Card Application. Hyphens and spaces are removed, thus Julie-Anne, Julie Anne, and Julieanne will be counted as a single entry.
3. Name data are not edited. For example, the sex associated with a name may be incorrect. Entries such as "Unknown" and "Baby" are not removed from the lists.
4. Different spellings of similar names are not combined. For example, the names Caitlin, Caitlyn, Kaitlin, Kaitlyn, Kaitlynn, Katelyn, and Katelynn are considered separate names and each has its own rank.
5. When two different names are tied with the same frequency for a given year of birth, we break the tie by assigning rank in alphabetical order.
6. Some names are applied to both males and females (for example, Micah). Our rankings are done by sex, so that a name such as Micah will have a different rank for males as compared to females. When you seek the popularity of a specific name (see "Popularity of a Name"), you can specify the sex. If you do not specify the sex, we provide rankings for the more popular name-sex combination.
7. To safeguard privacy, we exclude from our tabulated lists of names those that would indicate, or would allow the ability to determine, names with fewer than 5 occurrences in any geographic area. If a name has less than 5 occurrences for a year of birth in any state, the sum of the state counts for that year will be less than the national count.

Exercise



The New York Times

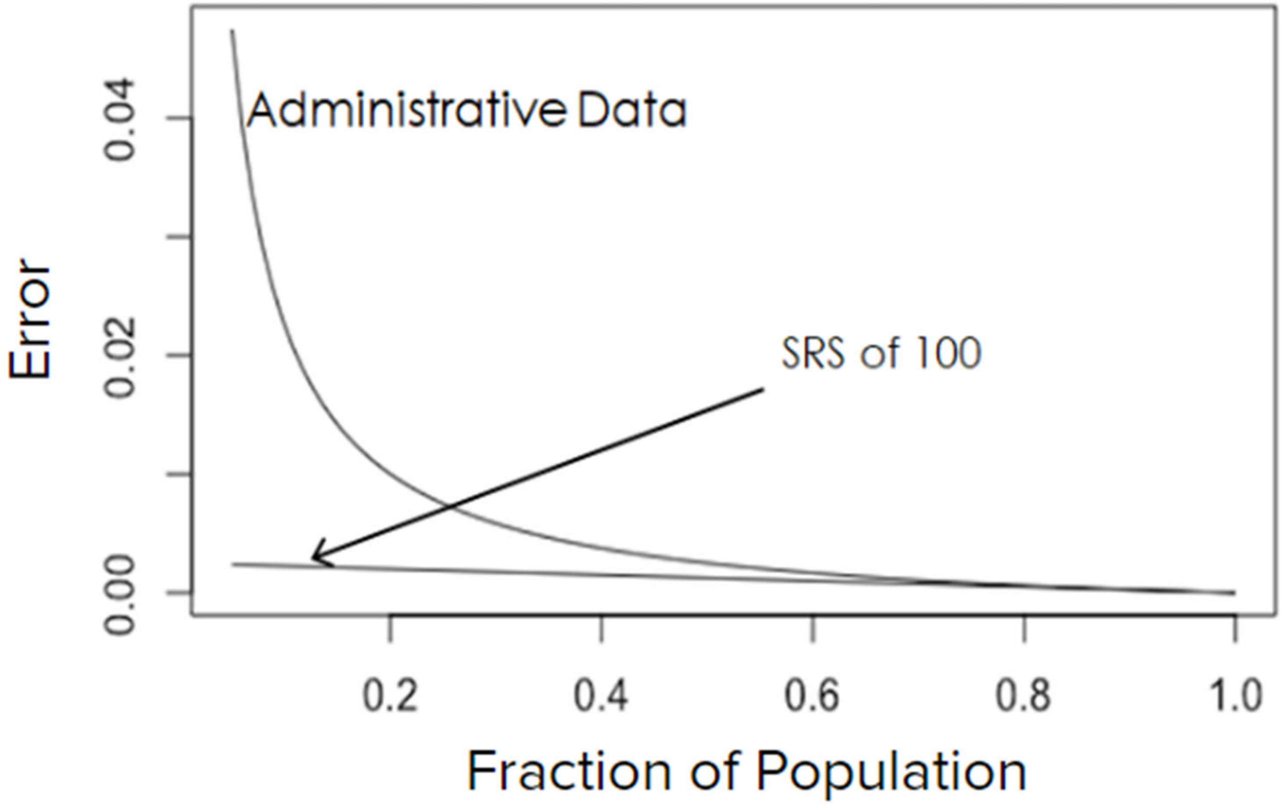
A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

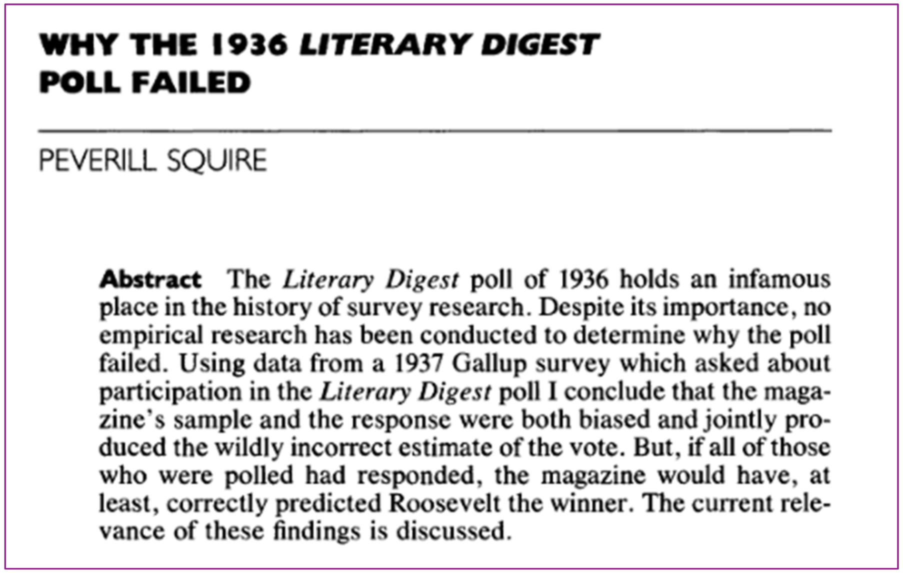
Aug. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

Population and Sample



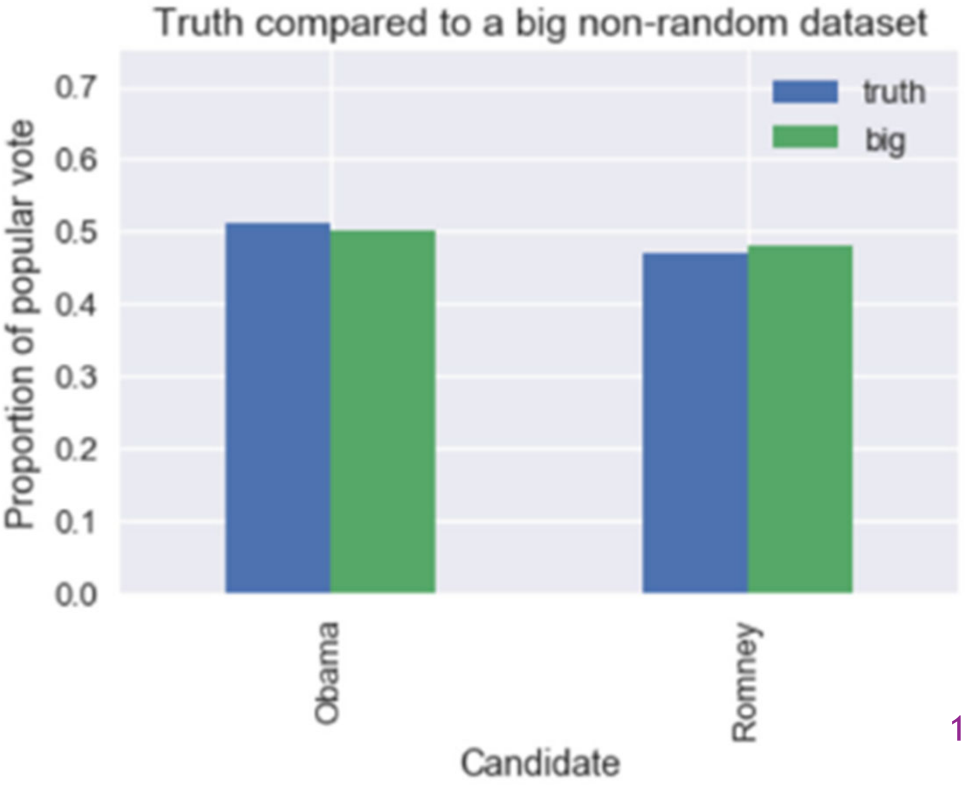
Exercise



WHY THE 1936 *LITERARY DIGEST* POLL FAILED

PEVERILL SQUIRE

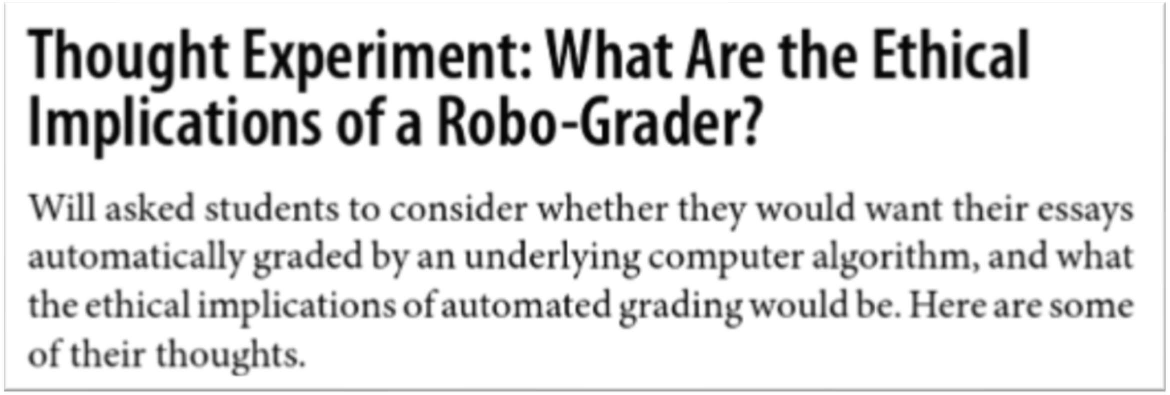
Abstract The *Literary Digest* poll of 1936 holds an infamous place in the history of survey research. Despite its importance, no empirical research has been conducted to determine why the poll failed. Using data from a 1937 Gallup survey which asked about participation in the *Literary Digest* poll I conclude that the magazine's sample and the response were both biased and jointly produced the wildly incorrect estimate of the vote. But, if all of those who were polled had responded, the magazine would have, at least, correctly predicted Roosevelt the winner. The current relevance of these findings is discussed.



Questions

- Questions on Piazza?
- Question for You!

What are the advantages and drawbacks of automated grading?



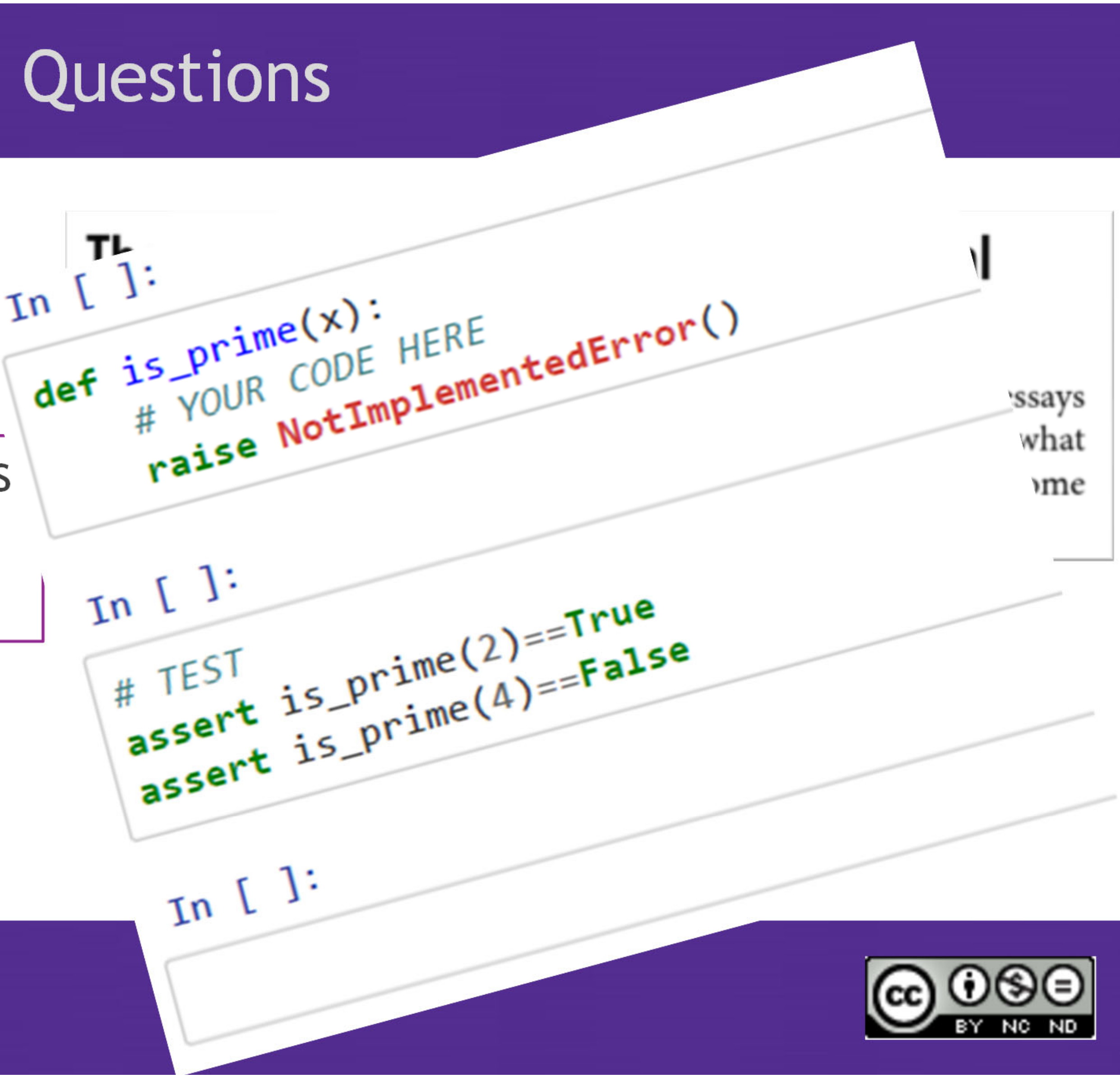
Thought Experiment: What Are the Ethical Implications of a Robo-Grader?

Will asked students to consider whether they would want their essays automatically graded by an underlying computer algorithm, and what the ethical implications of automated grading would be. Here are some of their thoughts.

Questions

- Questions on Piazza?
- Question for You!

What are the advantages and drawbacks of automated grading?



In [ ]:

```
def is_prime(x):
    # YOUR CODE HERE
    raise NotImplementedError()
```

In [ ]:

```
# TEST
assert is_prime(2)==True
assert is_prime(4)==False
```

In [ ]: