Introduction to Data Analysis

Data Analysis for Journalism and Political Communication
(Spring 2025)

Prof. Bell

Expectations

- **An A effort**: Starts homework early and asks questions/attends office hours; reviews the professor's feedback; reads assignment instructions and grading rubrics carefully; attends class; being perfect at coding is not an expectation of an A effort.
- **A B effort**: Completes all homework assignments; does not carefully read assignment instructions; uses outside resources to complete the homework; attends class; a B effort meets minimum expectations.
- **A C effort or lower**: Does not complete all homework assignments; does not carefully read assignment instructions; does not communicate with the professor; does not consistently attend class; I will reach out if you meet these criteria.

When You Get Stuck

1. Look back at the lecture materials. Everything required to complete the homework is in the lecture materials.
2. Google the error (this is different than Googling “how do I do X?”).
3. Email the professor. Be sure to include enough code in your email that I can recreate the problem.

Chalabi: 3 Ways to Spot a Bad Statistic

Chalabi: 3 Ways to Spot a Bad Statistic

1. Can you see uncertainty?
2. Can we look beyond the averages?
3. How was the data collected?

Can you see the uncertainty?

- We rarely get a complete count of everything

Can you see the uncertainty?

- We rarely get a complete count of everything
- How do we know that the **sample** we choose is a good representative of the whole **population**?

Can you see the uncertainty?

- We rarely get a complete count of everything
- How do we know that the **sample** we choose is a good representative of the whole **population**?
- We will learn how to measure and communicate about uncertainty, which is both art and science

Can you see the uncertainty?

- We rarely get a complete count of everything
- How do we know that the **sample** we choose is a good representative of the whole **population**?
- We will learn how to measure and communicate about uncertainty, which is both art and science
- Humans do not do well with probability

Can you see the uncertainty?

- We rarely get a complete count of everything
- How do we know that the **sample** we choose is a good representative of the whole **population**?
- We will learn how to measure and communicate about uncertainty, which is both art and science
- Humans do not do well with probability

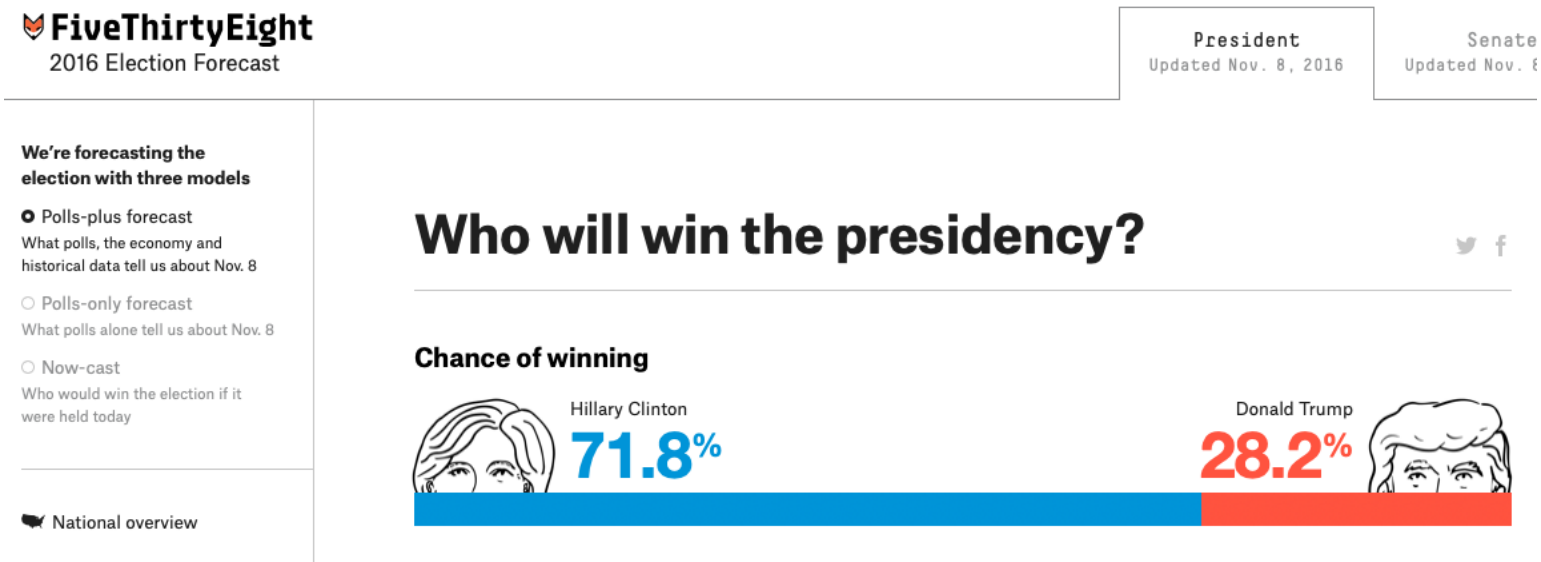
> "There are only five probabilities the average human can handle: 99 percent, one percent, 100 percent, zero, and 50-50. That's it."
> - Richard Thaler, Nobel Laureate in Economics

Can you see the uncertainty?

- We rarely get a complete count of everything
- How do we know that the **sample** we choose is a good representative of the whole **population**?
- We will learn how to measure and communicate about uncertainty, which is both art and science
- Humans do not do well with probability

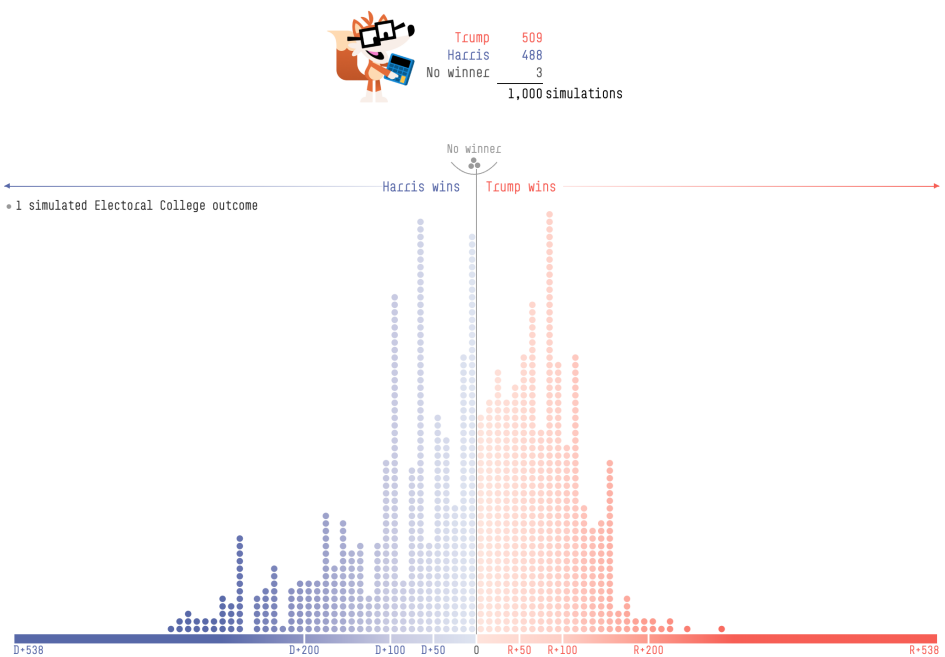
Can you see the uncertainty?

- We rarely get a complete count of everything
- How do we know that the **sample** we choose is a good representative of the whole **population**?
- We will learn how to measure and communicate about uncertainty, which is both art and science
- Humans do not do well with probability



Can you see the uncertainty?

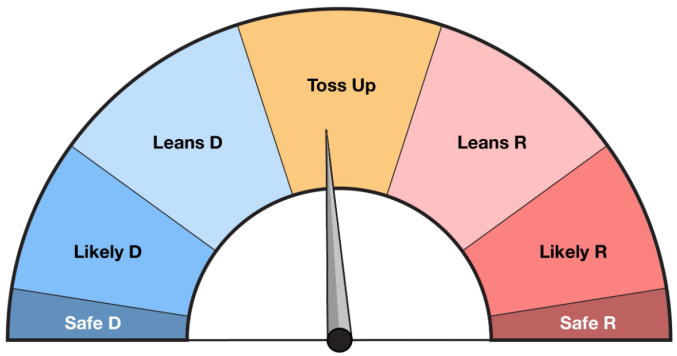
- We rarely get a complete count of everything
- How do we know that the **sample** we choose is a good representative of the whole **population**?
- We will learn how to measure and communicate about uncertainty, which is both art and science
- Humans do not do well with probability



Can you see the uncertainty?

- We rarely get a complete count of everything
- How do we know that the **sample** we choose is a good representative of the whole **population**?
- We will learn how to measure and communicate about uncertainty, which is both art and science
- Humans do not do well with probability

2024 Presidential Forecast



President

Harris 52%

241 Harris, 246 Trump, 51 Toss Up

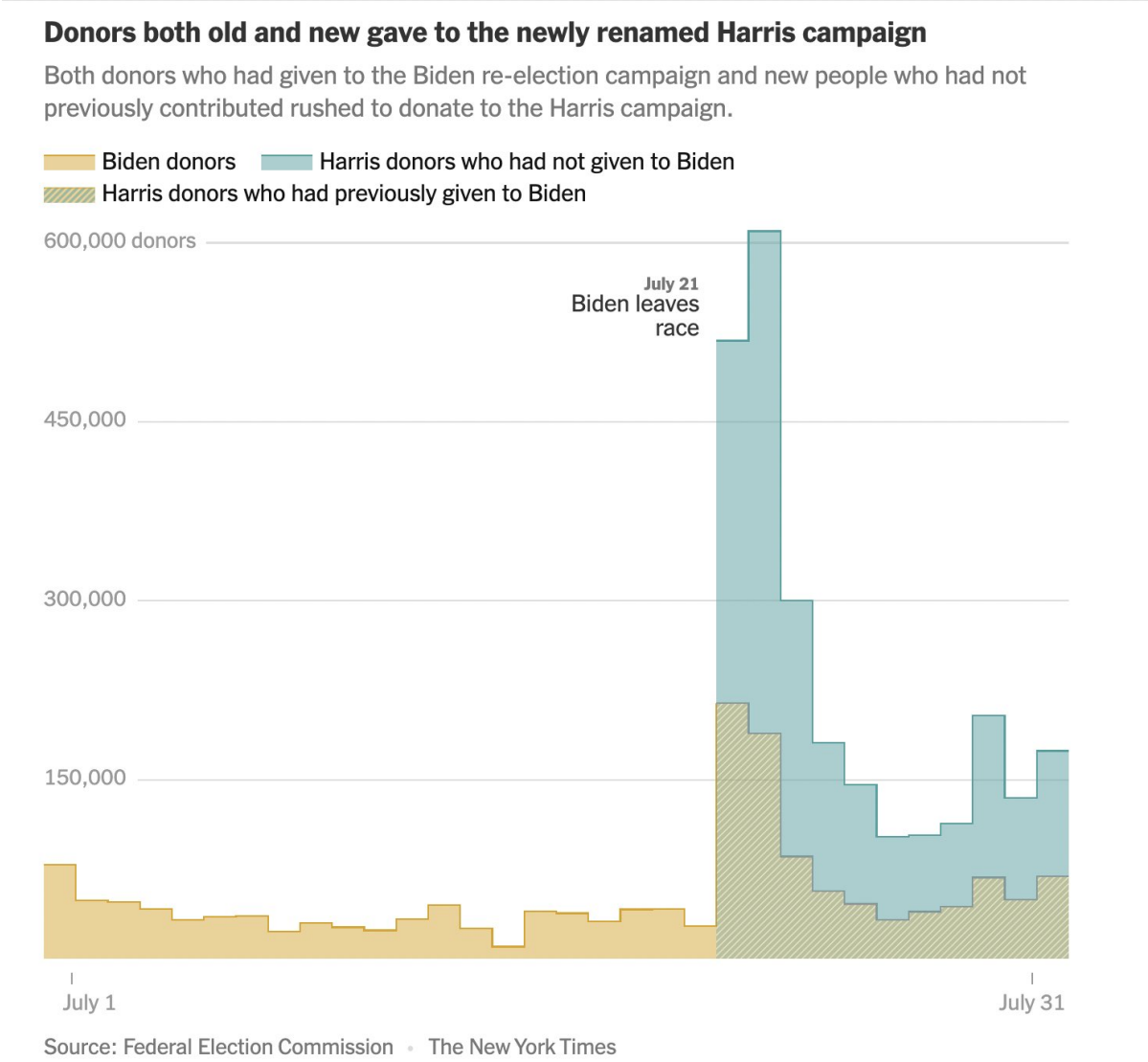
Can we look beyond the averages?

- There is always a trade-off between accuracy and simplicity when working with data

Can we look beyond the averages?

- There is always a trade-off between accuracy and simplicity when working with data
- We aggregate data to make it easier to comprehend, but we may also lose important context

Can we look beyond the averages?

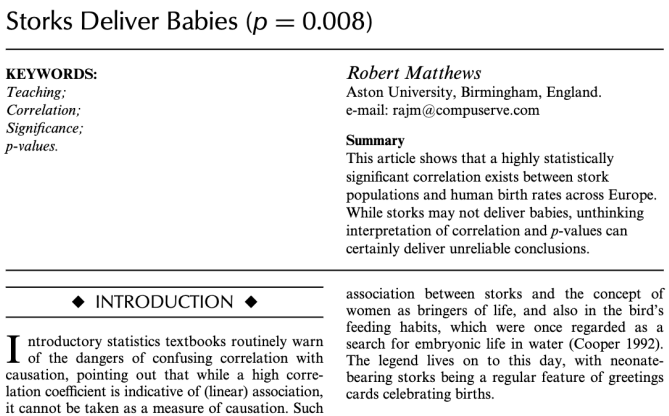


Can we look beyond the averages?

- There is always a trade-off between accuracy and simplicity when working with data
- We aggregate data to make it easier to comprehend, but we may also lose important context
- We will talk about using data visualization to communicate about data, as well as researcher choices and biases

Can we look beyond the averages?

- There is always a trade-off between accuracy and simplicity when working with data
- We aggregate data to make it easier to comprehend, but we may also lose important context
- We will talk about using data visualization to communicate about data, as well as researcher choices and biases
- We will also talk about the importance of theory in understanding data, especially correlation vs. causation



Storks Deliver Babies ($p = 0.008$)

KEYWORDS:
Teaching;
Correlation;
Significance;
p-values.

Robert Matthews
Aston University, Birmingham, England.
e-mail: rajm@compuserve.com

Summary
This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and *p*-values can certainly deliver unreliable conclusions.

◆ INTRODUCTION ◆

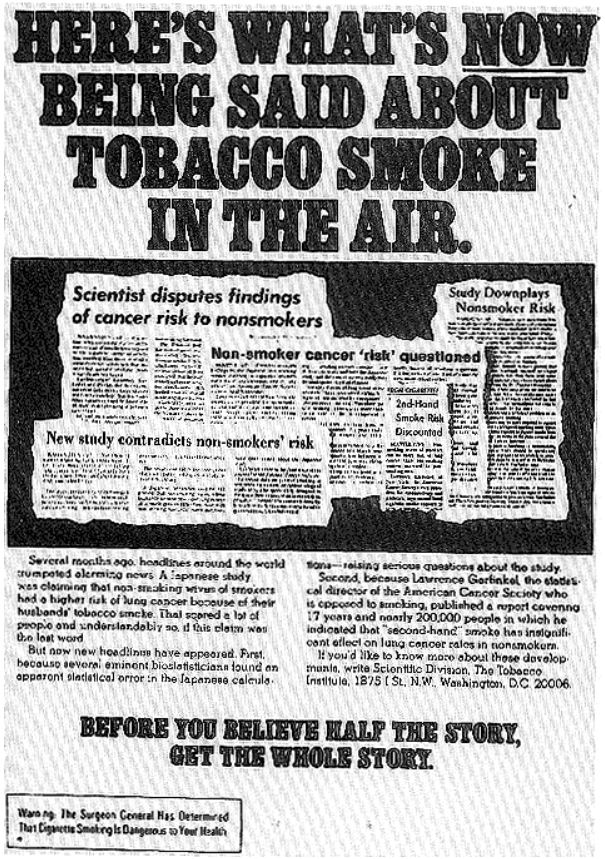
Introductory statistics textbooks routinely warn of the dangers of confusing correlation with causation, pointing out that while a high correlation coefficient is indicative of (linear) association, it cannot be taken as a measure of causation. Such association between storks and the concept of women as bringers of life, and also in the bird's feeding habits, which were once regarded as a search for embryonic life in water (Cooper 1992). The legend lives on to this day, with neonate-bearing storks being a regular feature of greetings cards celebrating births.

How was the data collected?

- Data is not objective – it is generated by humans

How was the data collected?

- Data is not objective – it is generated by humans
- Some data is produced by unscrupulous actors



HERE'S WHAT'S NOW BEING SAID ABOUT TOBACCO SMOKE IN THE AIR.

Scientist disputes findings of cancer risk to nonsmokers

Study Downplays Nonsmoker Risk

Non-smoker cancer 'risk' questioned

2nd-Hand Smoke Risk Discounted

New study contradicts non-smokers' risk

Several months ago, headlines around the world trumpeted alarming news. A Japanese study was claiming that non-smoking wives of smokers had a higher risk of lung cancer because of their husbands' tobacco smoke. That scared a lot of people and understandably so, if this claim was the last word.

But now new headlines have appeared. First, because several eminent biostatisticians found an apparent statistical error in the Japanese calculations—raising serious questions about the study.

Second, because Lawrence Garfinkel, the statistical director of the American Cancer Society who is opposed to smoking, published a report covering 17 years and nearly 200,000 people in which he indicated that "second-hand" smoke has insignificant effect on lung cancer rates in nonsmokers.

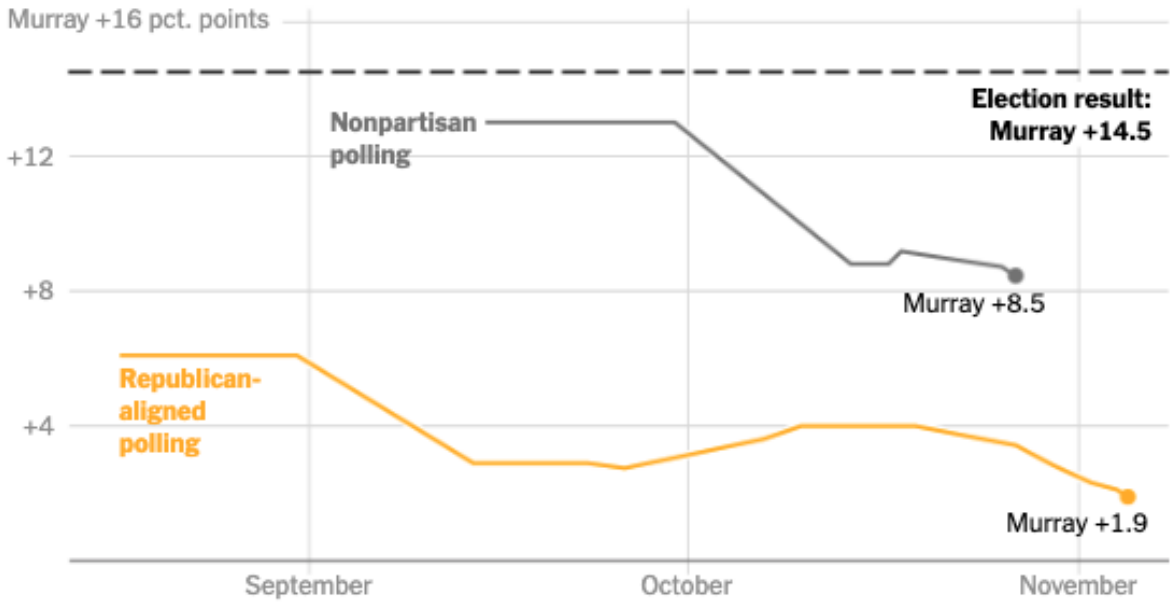
If you'd like to know more about these developments, write Scientific Division, The Tobacco Institute, 1875 I St., N.W., Washington, D.C. 20006.

BEFORE YOU BELIEVE HALF THE STORY, GET THE WHOLE STORY.

Warning: The Surgeon General Has Determined That Cigarette Smoking Is Dangerous to Your Health

How was the data collected?

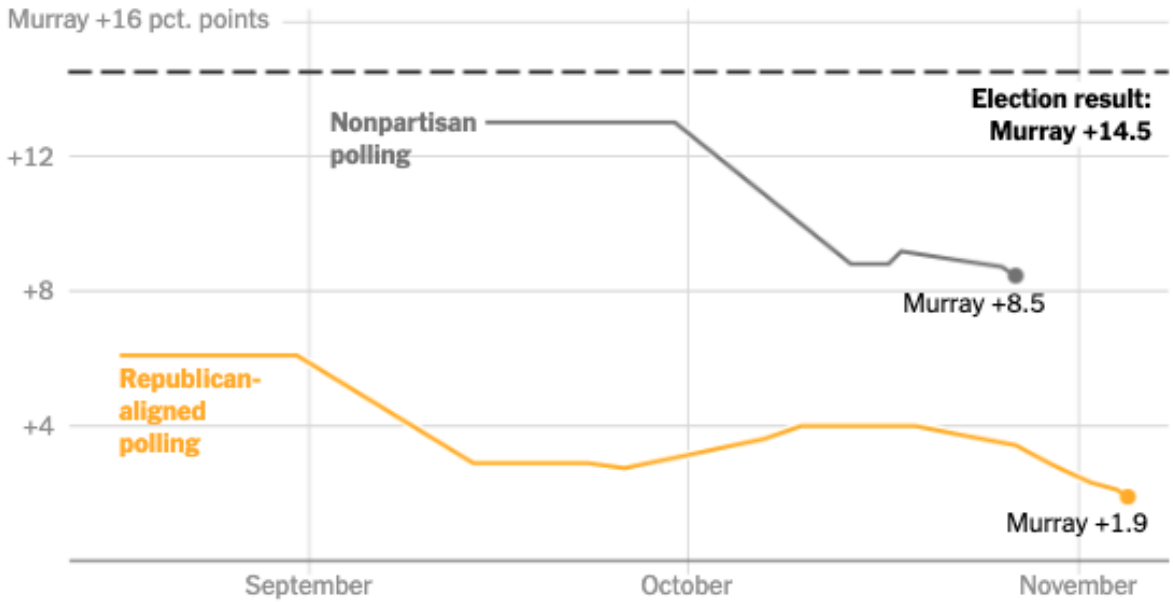
- Data is not objective – it is generated by humans
- Some data is produced by unscrupulous actors



Source: New York Times analysis of Washington Senate race polls aggregated by FiveThirtyEight • Notes: Trends are calculated with a 14-day average. Polling groups considered Republican-aligned include those identified by The New York Times and FiveThirtyEight. Polling groups considered nonpartisan are those not known to be aligned with or funded by a political party. • By Jason Kao

How was the data collected?

- Data is not objective – it is generated by humans
- Some data is produced by unscrupulous actors



Source: New York Times analysis of Washington Senate race polls aggregated by FiveThirtyEight • Notes: Trends are calculated with a 14-day average. Polling groups considered Republican-aligned include those identified by The New York Times and FiveThirtyEight. Polling groups considered nonpartisan are those not known to be aligned with or funded by a political party. • By Jason Kao

- But most of the time, poor analysis is not nefarious – humans are imperfect

How was the data collected?

- Garbage in = garbage out: no amount of statistical wizardry can compensate for bad data

How was the data collected?

- Garbage in = garbage out: no amount of statistical wizardry can compensate for bad data
- We will spend a lot of time thinking about the **data generating process** and how it can bias our results

How was the data collected?

Support for mass deportation varies depending on how the question is asked

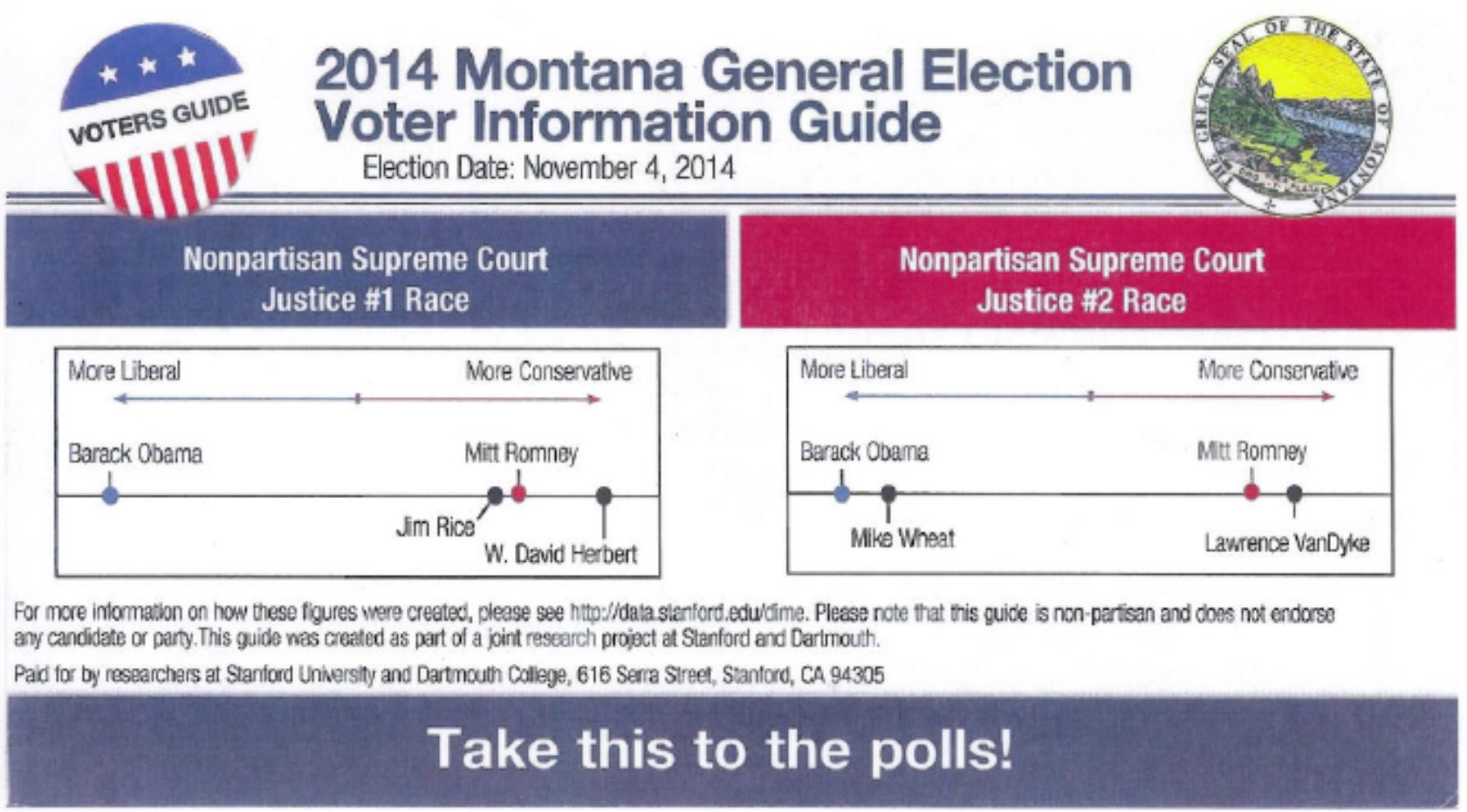
A sample of different questions asked about deportation this year show significant variation in levels of support — sometimes, even within the same survey

Pollster	Wording	Support deportations ▼	Do not support deportations
CBS News/YouGov *Registered voters, June 5-7*	Would you favor or oppose the U.S. government starting a new national program to deport all undocumented immigrants currently living in the U.S. illegally?	62%	38%
Marquette Law School *Registered voters, October 1-10*	Do you favor or oppose deporting immigrants who are living in the United States illegally back to their home countries?	58%	42%
ABC/Ipsos *US adults, October 4-8*	There are at least 11 million undocumented immigrants living in the United States. Would you support or oppose an effort by the federal government to deport all these undocumented immigrants and send them back to their home countries?	56%	43%
CNN *US adults, January 25-30*	If Donald Trump becomes president again, would you favor or oppose him trying to...detain and deport millions of undocumented immigrants?	48%	52%
Gallup *US adults, June 3-23*	Please tell me whether you strongly favor, favor, oppose, or strongly oppose each of the following proposals....Deporting all immigrants who are living in the United States illegally back to their home country	47%	51%
Marquette Law School *Registered voters, October 1-10*	Do you favor or oppose deporting immigrants who are living in the United States illegally back to their home countries even if they have lived here for a number of years, have jobs and no criminal record?	40%	60%
Pew Research *US adults, April 8-14*	Which comes closer to your view about how to handle undocumented immigrants who are now living in the U.S.? (They should not be allowed to stay in the country legally/There should be a way for them to stay in the country legally, if certain requirements are met) If "not be allowed: *"Do you think there should be a national law enforcement effort to deport all immigrants who are now living in the U.S. illegally?"* *("Support" percentage includes those who say there should be a national deportation effort, "do not support" includes all others)*	33%	67%

How was the data collected?

- Garbage in = garbage out: no amount of statistical wizardry can compensate for bad data
- We will spend a lot of time thinking about the **data generating process** and how it can bias our results
- We will also discuss our ethical responsibilities around data

How was the data collected?



Group Discussion

Introduce yourself to your neighbor(s) and take a few minutes to review these additional graphs from Mona Chalabi. Do any of these stand out to you as being good (or bad) examples of our three questions for spotting a bad statistic?

1. Can you see uncertainty?
2. Can we look beyond the averages?
3. How was the data collected?

On your notecard, please write:

1. Preferred name
2. Preferred pronouns
3. Year in school and major
4. Your background in coding and/or statistics
5. One thing you hope to get out of this class