Introduction to Data Analysis

Data Analysis for Journalism and Political Communication
(Fall 2025)

August 25, 2025

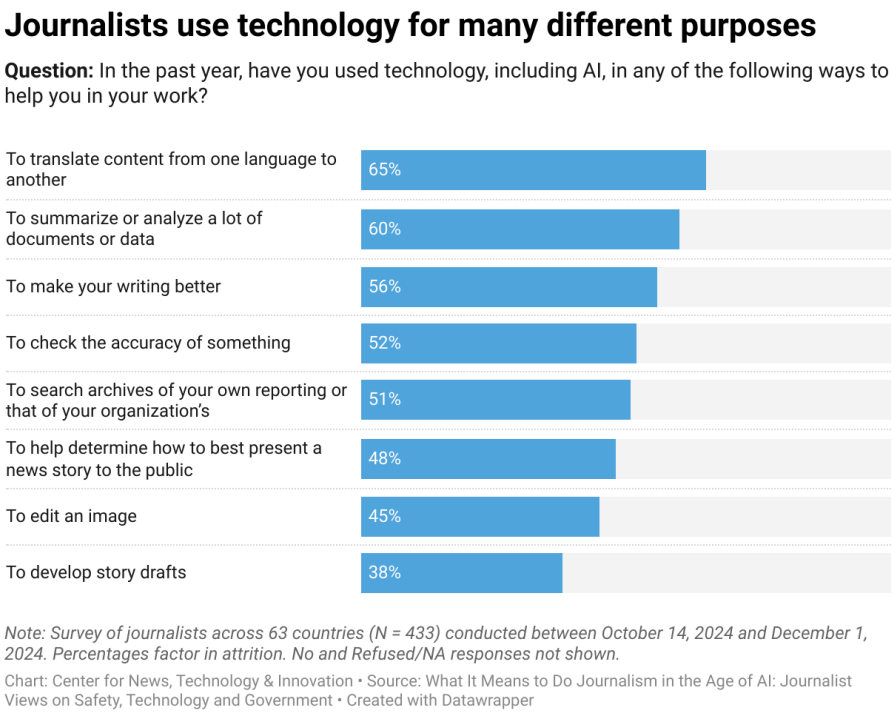
Key Details

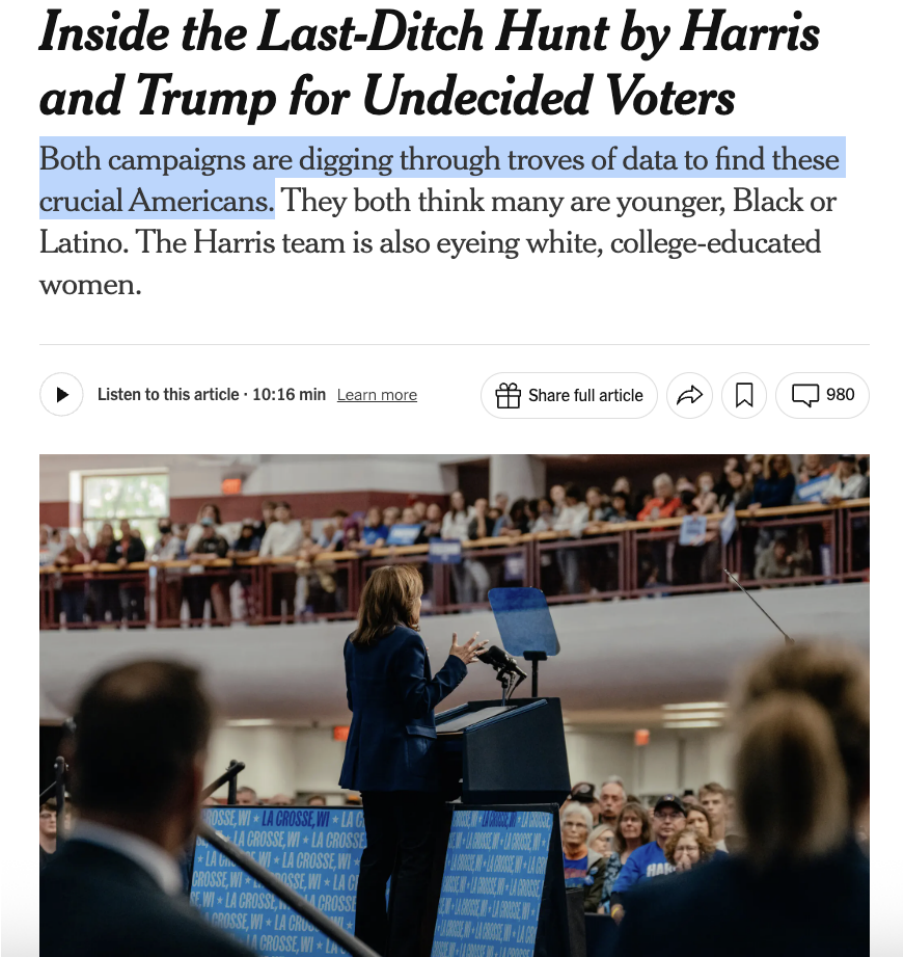
Professor: Nicholas Bell, Ph.D. (he/him)
nicholasbell@gwu.edu

Office Hours: Tuesdays 5:30 - 6:30pm
MPA 425
Appointments are recommended but not required: http://bit.ly/3V4riqR

I prefer to meet during office hours or by appointment. However, I am available by email, and I try to respond to emails by the end of the next business day (M-F).

Why Is This Course Required?





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 the uncertainty inherent in statistics

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 the uncertainty inherent in statistics
- 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 the uncertainty inherent in statistics
- Humans do not do well with probability

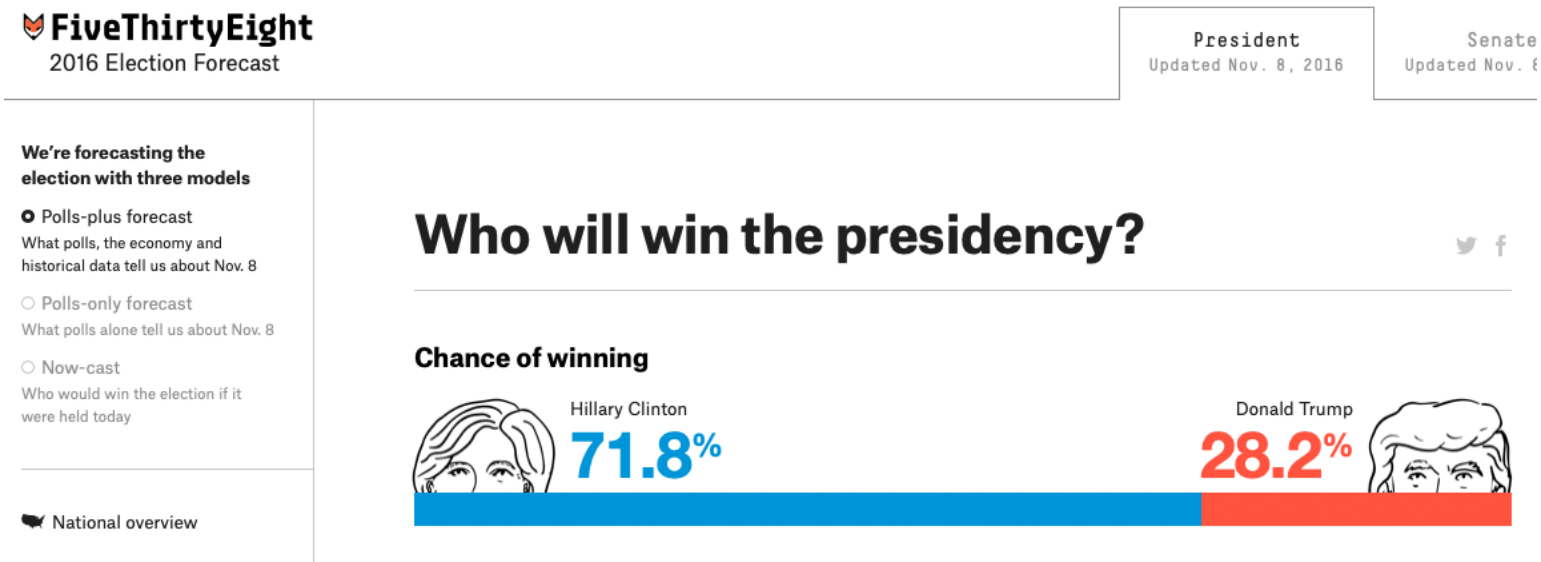
> "There are only five probabilities the average human can handle: 99 percent, one percent, 100 percent, zero, and 50-50. That's it."
> - Richard Thaler, Nobel Laureate in Economics

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 the uncertainty inherent in statistics
- 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 the uncertainty inherent in statistics
- Humans do not do well with probability



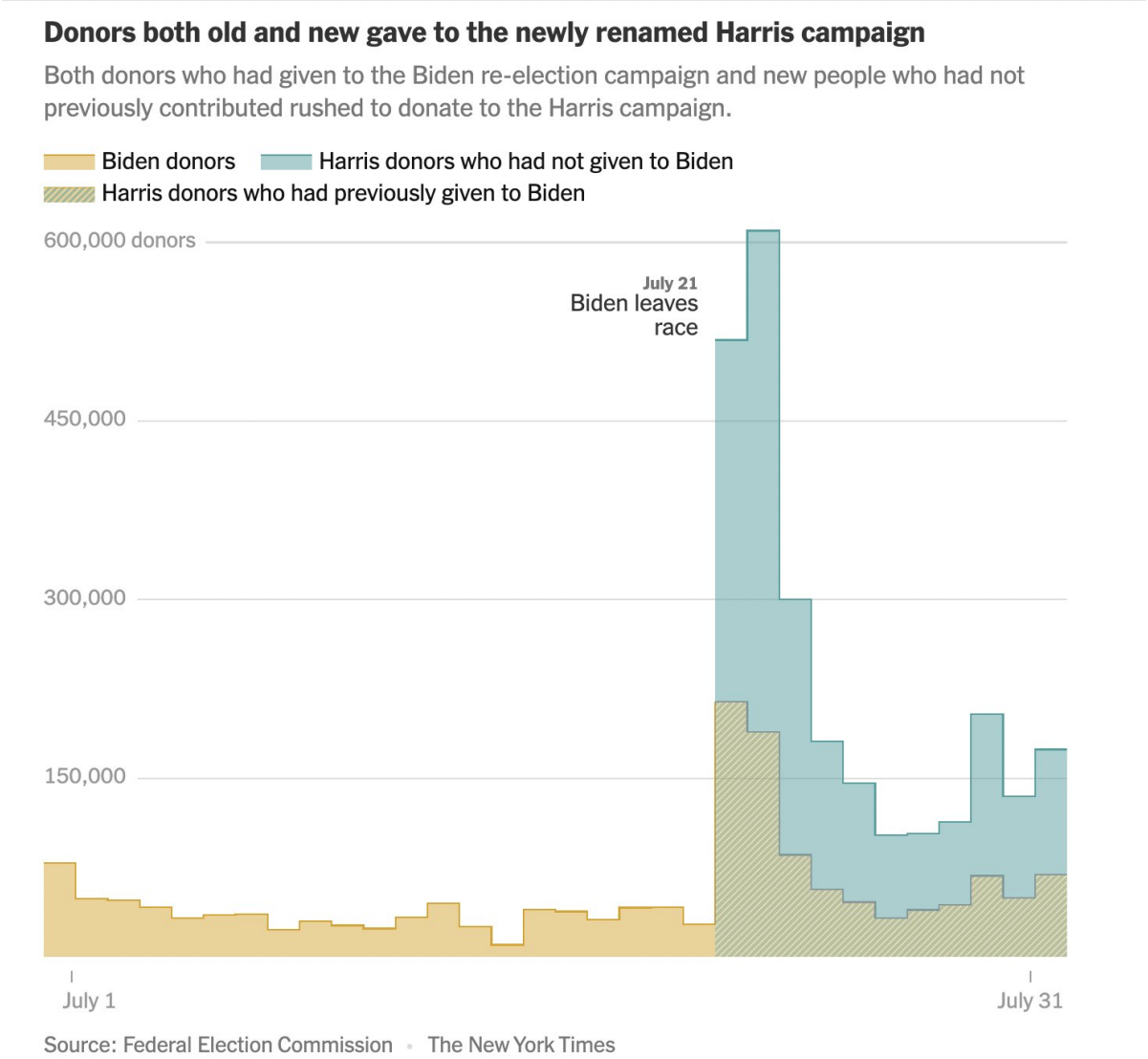
Can we look beyond the averages?

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

Can we look beyond the averages?

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

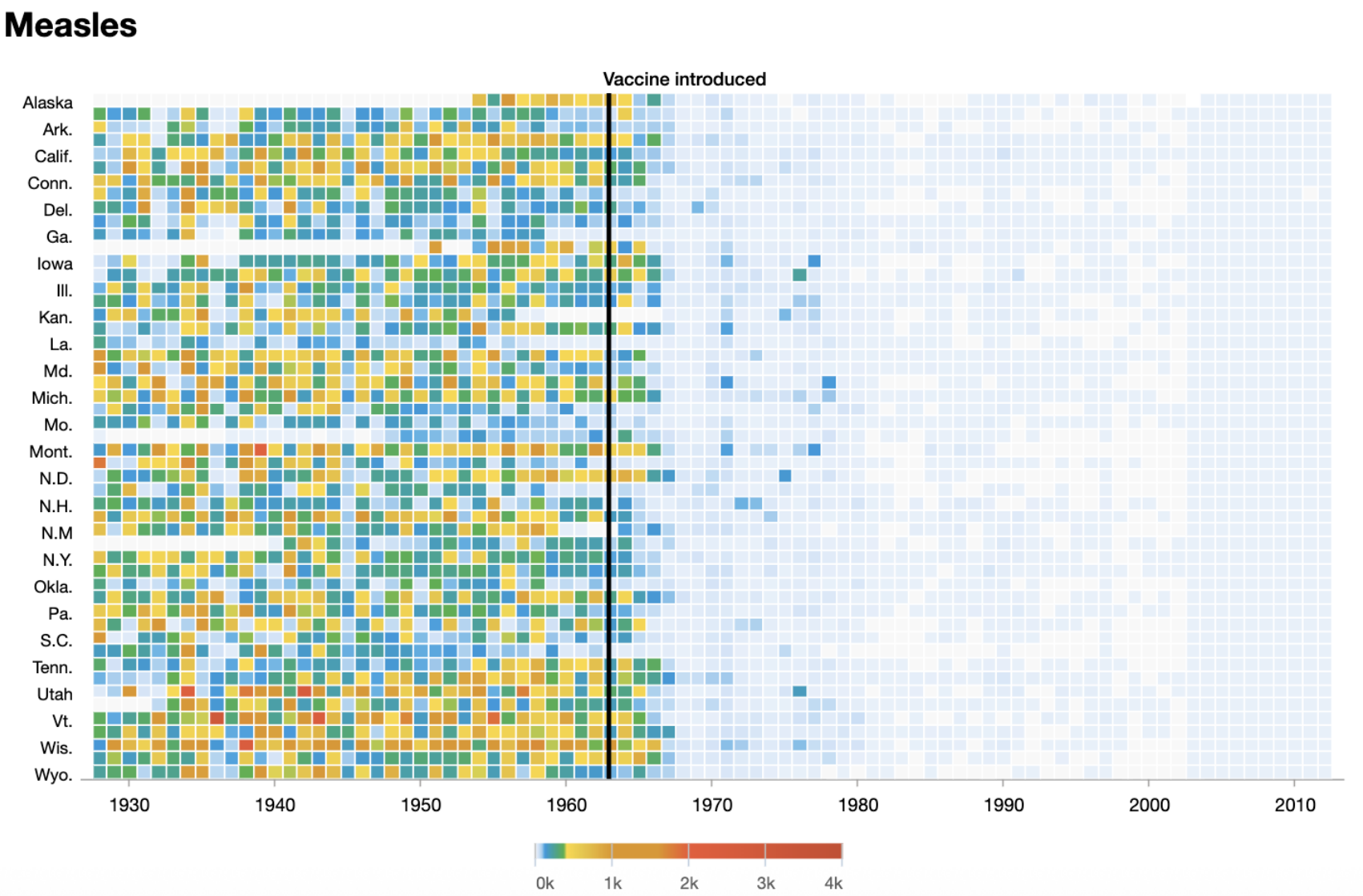
Can we look beyond the averages?



Can we look beyond the averages?

- There is always a trade-off between simplicity and precision when working with data
- We summarize 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

Can we look beyond the averages?



Can we look beyond the averages?

- There is always a trade-off between simplicity and precision when working with data
- We summarize 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
- 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 ◆

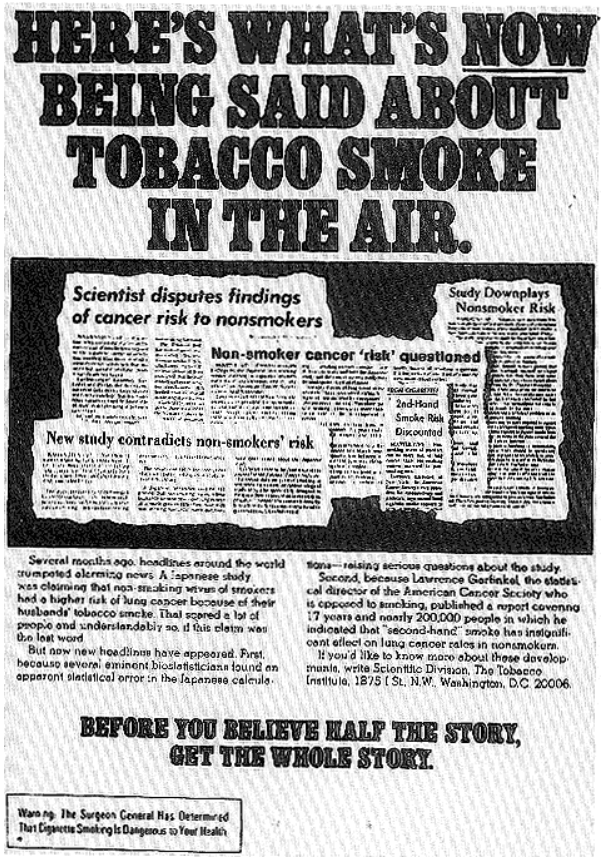
Introductory statistics textbooks routinely warn of the dangers of confusing correlation with causation, pointing out that while a high correlation coefficient is indicative of (linear) association, it cannot be taken as a measure of causation. Such association between storks and the concept of women 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
- 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

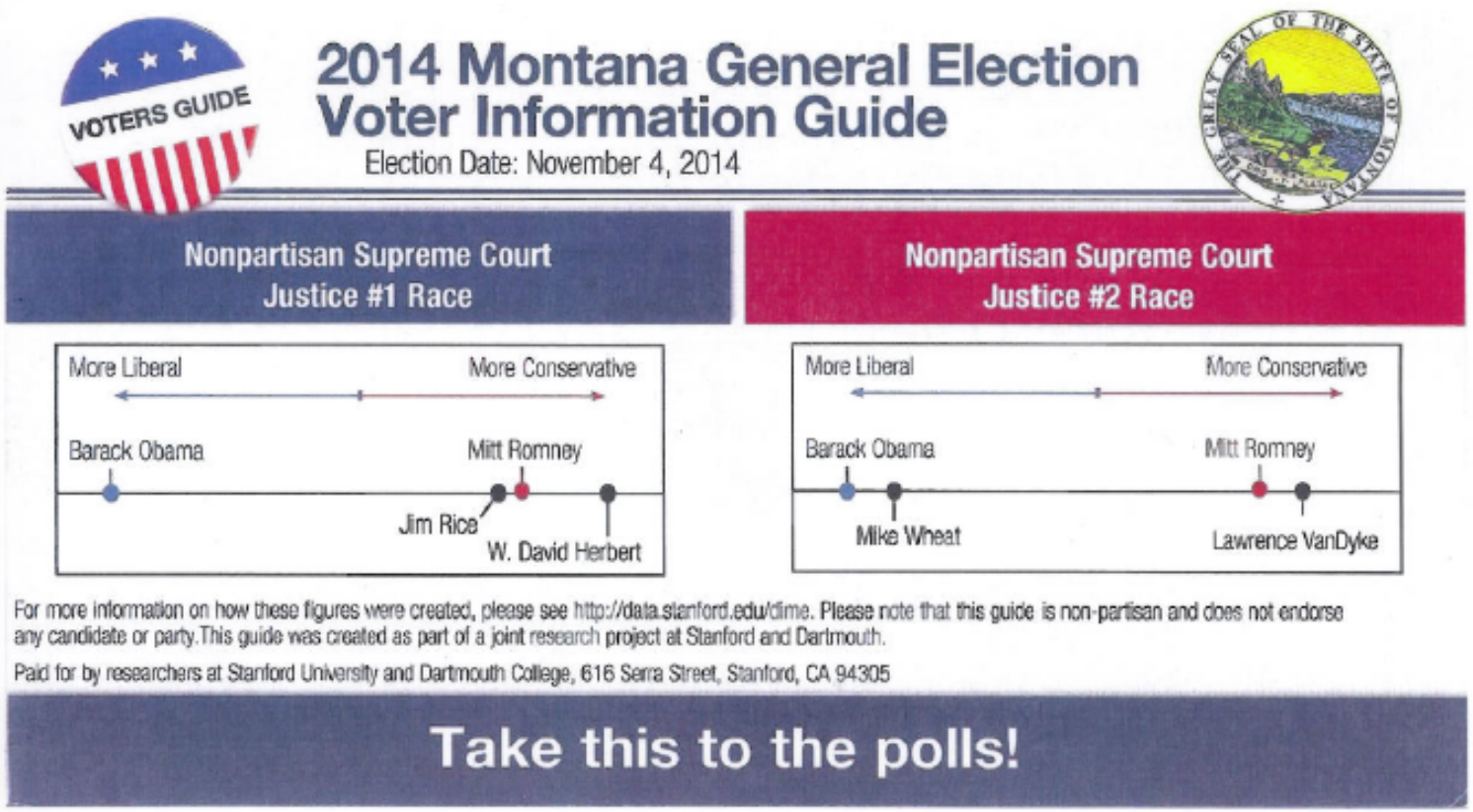
A sample of different questions asked about deportation this year show significant variation in levels of support — sometimes, even within the same survey

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?



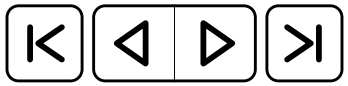
Group Discussion

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?

Group Discussion

Controls:



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?

How the Course Will Work

Your course grade is calculated as your grade on each of the following course components weighted by:

Attendance	10%
Lab assignments	35%
Class project	20%
Final exam	35%

Lab Assignments

- You will receive a very basic introduction to the R statistical programming language

Lab Assignments

- You will receive a very basic introduction to the R statistical programming language
- There are seven in-class Lab Days where you will work on an assignment applying the lessons learned in class to real-world problems using R

Lab Assignments

- You will receive a very basic introduction to the R statistical programming language
- There are seven in-class Lab Days where you will work on an assignment applying the lessons learned in class to real-world problems using R
- You must watch the video lectures on R prior to the Lab Day class.

Lab Assignments

- You will receive a very basic introduction to the R statistical programming language
- There are seven in-class Lab Days where you will work on an assignment applying the lessons learned in class to real-world problems using R
- You must watch the video lectures on R prior to the Lab Day class.
- You must turn in your completed lab assignment by 11:59pm the next day.

Lab Assignments

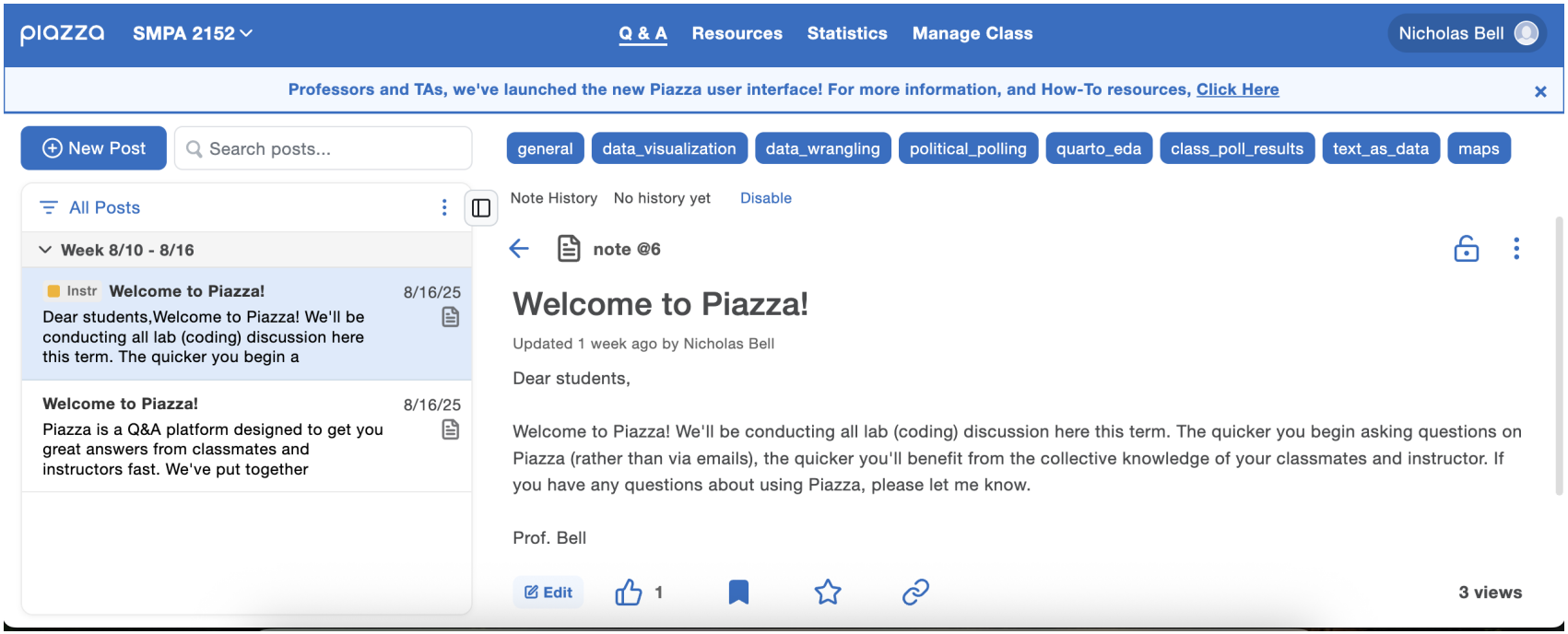
- You will receive a very basic introduction to the R statistical programming language
- There are seven in-class Lab Days where you will work on an assignment applying the lessons learned in class to real-world problems using R
- You must watch the video lectures on R prior to the Lab Day class.
- You must turn in your completed lab assignment by 11:59pm the next day.
- You may complete these assignments on your own or in collaboration with other students. This means that you may work together to write code and/or solve problems. **Do not split up the questions or combine independent work.** Each student must submit an assignment on Blackboard.

Lab Assignments

There is a software requirement for this course. Students may select from either option below.

1. **Option 1: Positron**. Positron is a free, open source development environment for R (and other coding languages) developed by the same creators of Posit Cloud.
2. **Option 2: Posit Cloud**. Posit Cloud is a web-based version of the popular R development environment RStudio. At the time of this writing, a Student account for Posit Cloud costs $5 per month.

Piazza


piazza
SMPA 2152
Q & A
Resources
Statistics
Manage Class
Nicholas Bell
Professors and TAs, we've launched the new Piazza user interface! For more information, and How-To resources, Click Here
New Post
Search posts...
general
data_visualization
data_wrangling
political_polling
quarto_eda
class_poll_results
text_as_data
maps
All Posts
Week 8/10 - 8/16
Instr Welcome to Piazza!
8/16/25
Dear students,Welcome to Piazza! We'll be conducting all lab (coding) discussion here this term. The quicker you begin a
Welcome to Piazza!
8/16/25
Piazza is a Q&A platform designed to get you great answers from classmates and instructors fast. We've put together
Note History
No history yet
Disable
note @6
Welcome to Piazza!
Updated 1 week ago by Nicholas Bell
Dear students,
Welcome to Piazza! We'll be conducting all lab (coding) discussion here this term. The quicker you begin asking questions on Piazza (rather than via emails), the quicker you'll benefit from the collective knowledge of your classmates and instructor. If you have any questions about using Piazza, please let me know.
Prof. Bell
Edit
1
3 views

Course Policy on Generative AI

This course permits the use of Generative AI on **code** submitted for evaluation without restriction. However, the use of GAI tools for **written text** (e.g., exposition, analysis, etc.) is not permitted.

Class Project

- You will conduct a survey of your fellow GW students

Class Project

- You will conduct a survey of your fellow GW students
- You will have an opportunity to indicate your preference for different roles in the survey process

Class Project

- You will conduct a survey of your fellow GW students
- You will have an opportunity to indicate your preference for different roles in the survey process
- Your grade consists of your successful participation in the process plus a Lab Day assignment which uses the survey data you collect.

Class Project

- You will conduct a survey of your fellow GW students
- You will have an opportunity to indicate your preference for different roles in the survey process
- Your grade consists of your successful participation in the process plus a Lab Day assignment which uses the survey data you collect.
- You must complete the assigned CITI Ethics training to participate in the class project.

Questions?

Reminder: There is no class this Wednesday (prof. at a conference) or next Monday (Labor Day).

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