# Introduction to Data Science

## Data Ethics - Misrepresentation and Data Privacy

Joanna Bieri DATA101

# Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours: Duke 209 Click Here for Joanna's Schedule

# Introduction

# Data Science Ethics

Today we will start our week long focus on Data Ethics.

Data Science in a Box Unit 3 Decks 1 and 2

# Data Science Ethics

Today we will consider:

- Misrepresentation
- Data Privacy

# Small Group Discussion

As a group discuss the following:

Each Individual:

1. What book or articles are you reading?
2. What so far have been the most interesting ethical points?

As a group:

1. Define misrepresentation and data privacy in your own words.
2. Can you draw some examples from your individual reading that might pertain to our discussion today.

# Misrepresentation

- Misrepresentation can happen intentionally or unintentionally.
- It can arise because of lack of knowledge of competence in data science.
- It is important to be aware of misrepresentation and be able to spot it a mile away in your work and in other peoples work!

# Causality

# Causality

Our human brains are looking for connections between things:

"If I do X then Y will happen"

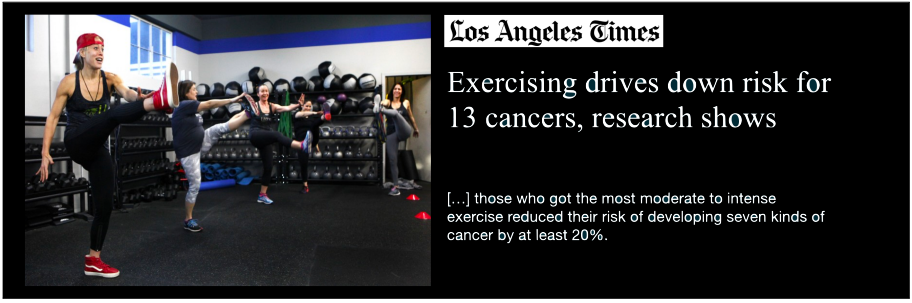
# Example Study - Time Magazine


Alice Park. Exercise Can Lower Risk of Some Cancers By 20%. Time Magazine. 16 May 2016.

# Example Study - Time Magazine

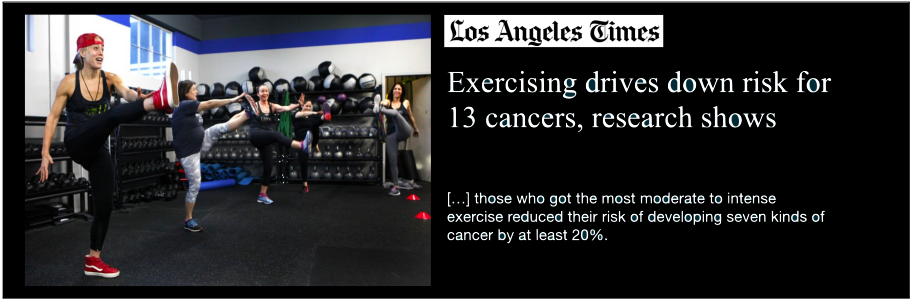

- Are there confounding variables?
- This was not a randomized study.
- Can they claim exercise causes this reduction in cancer?

# Example Study - Los Angeles Times




Melissa Healy. Exercising drives down risk for 13 cancers, research shows.
Los Angeles Times. 16 May 2016.

# Example Study - Los Angeles Times




[...] those who got the most moderate to intense exercise reduced their risk of developing seven kinds of cancer by at least 20%

- This is a VERY causal statement!
- This makes it sound like if I started exercising today my risk for cancer would go down - **Risk was reduced?**

# Original study

Moore, Steven C., et al. **"Association of leisure-time physical activity with risk of 26 types of cancer in 1.44 million adults."** JAMA internal medicine 176.6 (2016): 816-825.

- There were a HUGE number of volunteers!
- **Volunteers** were **asked** about their physical activity level over the preceding year. (Survey!)
- Half exercised less than about 150 minutes per week, half exercised more.
- Compared to the bottom 10% of exercisers, the top 10% had lower rates of esophageal, liver, lung, endometrial, colon, and breast cancer.
- Researchers found no association between exercising and 13 other cancers (e.g. pancreatic, ovarian, and brain).

Carl Bergstrom and Jevin West. Calling Bullshit: The art of skepticism in a data-driven world.
Random House, 2020.

Sharon Begley. "Does exercise prevent cancer?". StatNews. 16 May 2016.

# Original study

**So what can we actually say?**

Exercise was associated with lower cancer rates.

**What is the harm here?**

# Data Visualization

# Data Visualization

There are many ways that visualizations can be (and have been) created to be misleading.

You want to make sure that you are communicating your point HONESTLY!

# Axes and scale

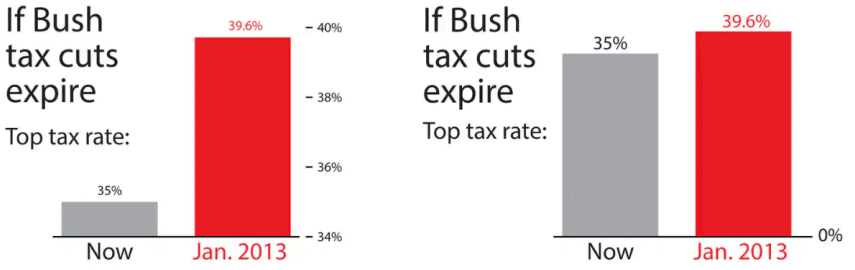



Figure 1: Tax cuts plot

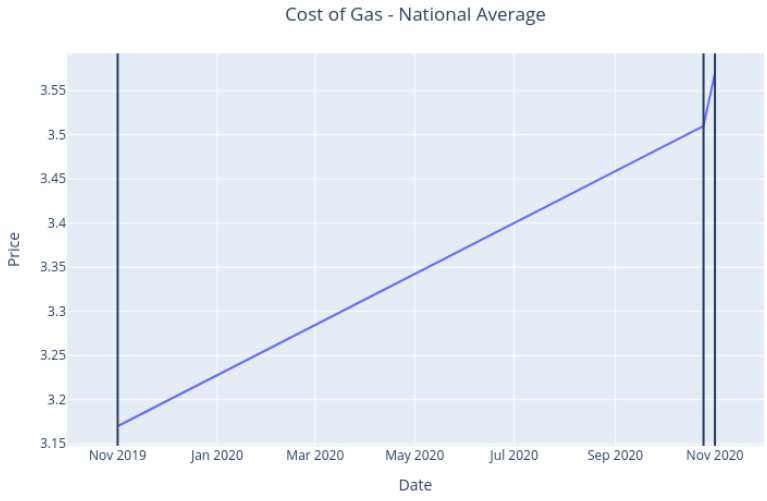
Christopher Ingraham. "You've been reading charts wrong. Here's how a pro does it.".
The Washington Post. 14 October 2019.

# Axes and scale

- The numbers at the top of the bars are the same and represent the top tax rate.
- Notice that the minimum for the y-axis is different!
- These tell a very different visual story.

Your axis should **usually** start at zero, unless there is a good, honest, data visualization reason that you are not starting a zero.

# Axes and scale



# Axes and scale

What do you see here? What is not quite right?

- Notice that the x-axis is evenly spaced even though it represents very different time steps!
- This makes it look like gas prices are increasing only a little bit in the last week.

# Axes and scale

How could we fix this?

1. Read the data off of the plot.
2. Put it into a data frame.
3. Make a plot of our own.

# Axes and scale

```
date = ["2019-11-01", "2020-10-25", "2020-11-01"]
cost = [3.17, 3.51, 3.57]
text = ['Last year','Last week','Current']

DF = pd.DataFrame()
DF['date'] = date
DF['cost']=cost
```

| | date | cost |
|---|---|---|
| 0 | 2019-11-01 | 3.17 |
| 1 | 2020-10-25 | 3.51 |
| 2 | 2020-11-01 | 3.57 |

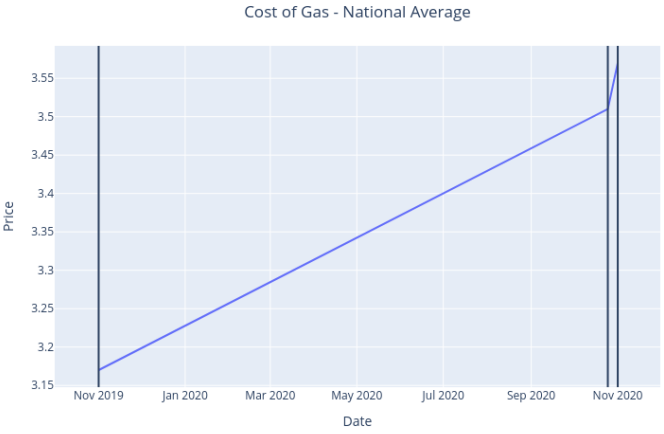
# Axes and scale

```
fig = px.line(DF,x='date',y='cost')
fig.update_layout(title='Cost of Gas - National Average',
                  title_x=0.5,
                  xaxis_title="Date",
                  xaxis_range=["2019-10-01", "2020-12-01"],
                  yaxis_title="Price",
                  autosize=False,
                  width=800,
                  height=500)

fig.add_vline(x="2019-11-01")
fig.add_vline(x="2020-10-25")
fig.add_vline(x="2020-11-01")

fig.show()
```

# Axes and scale

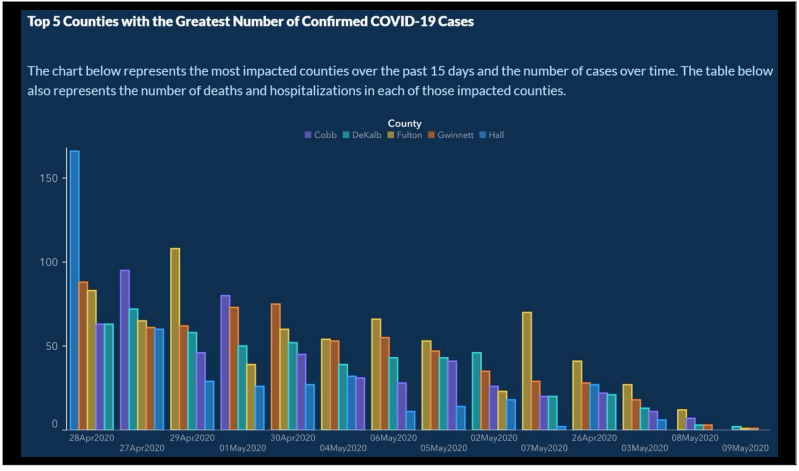


# Axes and scale

The picture here is a lot different!

**Your axis should always be consistent in scale!!!**

# Axes and scale



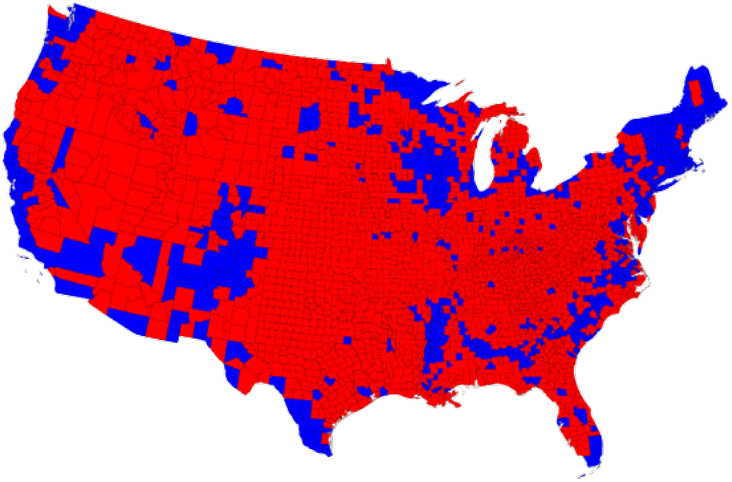
# Maps and areas

# Maps and areas

This is a special type of data visualization.

One common pitfall in visualizing data is mixing geographic area data with data about quantities.

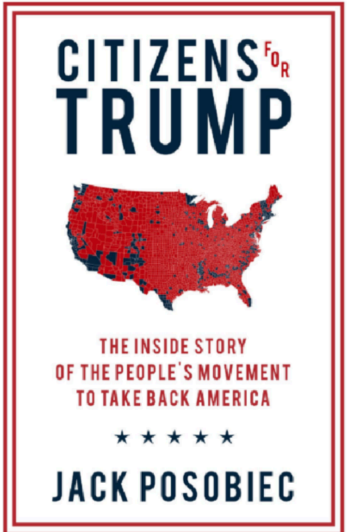
# Maps and areas



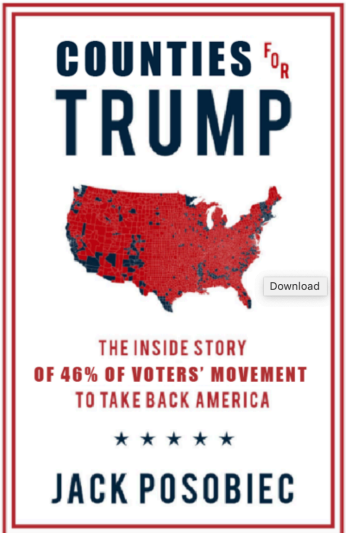
Lazaro Gamio. "Election maps are telling you big lies about small

# Maps and areas

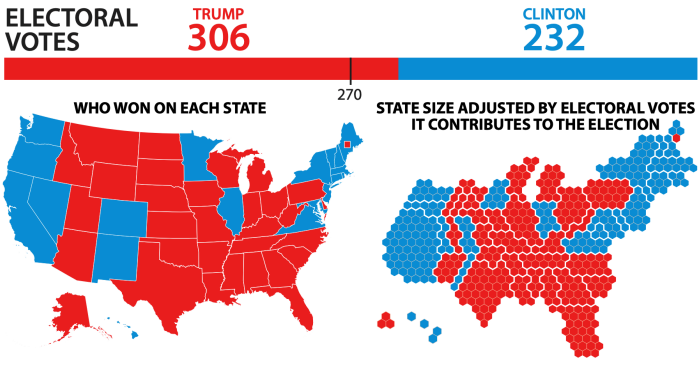
**WRONG**



# Maps and areas



# Maps and areas




Alberto Cairo. Visual Trumpery talk.

# Visualizing uncertainty

# Visualizing uncertainty

On December 19, 2014, the front page of Spanish national newspaper El País read *“Catalan public opinion swings toward ‘no’ for independence, says survey”.

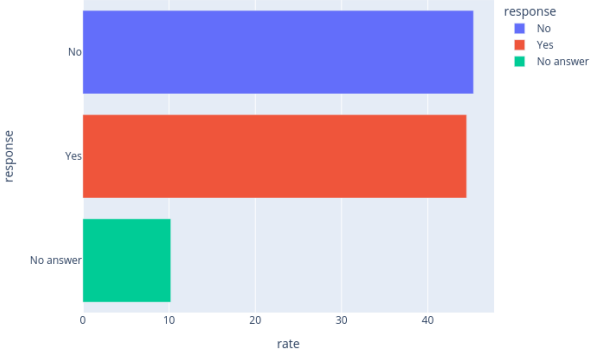
Here is the data for that study:

| | response | rate | error |
|---|---|---|---|
| 0 | No | 45.3 | 2.95 |
| 1 | Yes | 44.5 | 2.95 |
| 2 | No answer | 10.2 | 2.95 |

# Visualizing uncertainty

Lets use a bar plot

```
fig = px.bar(DF,x='rate',y='response',color='response')
fig.show()
```



# Visualizing uncertainty

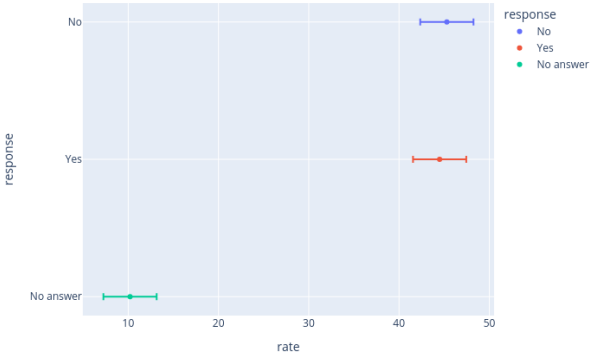
This representation of the data is misleading

Margin of error is $+/-2.95\%$ at 95% confidence level

Alberto Cairo. The truthful art: Data, charts, and maps for communication. New Riders, 2016.

# Visualizing uncertainty

```
fig = px.scatter(DF,x='rate',y='response',color='response',
                 error_x='error')
fig.show()
```

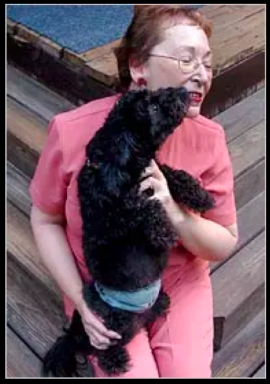


# Data Privacy

# Data Privacy

The question of data privacy is complicated and rests on the idea of what amount of expected privacy are we entitled to when we put our data online.

# Data Privacy - Case study:



The New York Times

## A Face Is Exposed for AOL Searcher No. 4417749

Ms. [Thelma] Arnold, who agreed to discuss her searches with a reporter, said she was shocked to hear that AOL had saved and published three months' worth of them. "My goodness, it's my whole personal life," she said. "I had no idea somebody was looking over my shoulder."

In the privacy of her four-bedroom home, Ms. Arnold searched for the answers to scores of life's questions, big and small. How could she buy "school supplies for Iraq children"? What is the "safest place to live"? What is "the best season to visit Italy"?

# Data Privacy - Case study:

AOL saved and published three months of search data. In the data leak the names were not included in the data, but it was very easy to connect the names.

Even today our search data is being saved, unless you opt out.

How would you feel if your data was leaked?

Michael Barbaro and Tom Zeller Jr. A Face Is Exposed for AOL Searcher No. 4417749. New York Times. 9 August 2006.

# Data Privacy - Case study:

- You should be very critical of where your data is coming from.
- Make sure it was sourced ethically!

# Data Privacy - Case study:

- In 2016, researchers published data of 70,000 OkCupid users—including usernames, political leanings, drug usage, and intimate sexual details
- Researchers didn't release the real names and pictures of OKCupid users, but their identities could easily be uncovered from the details provided, e.g. usernames
- Usernames were often either real names or reused across platforms that were easy to connect to a person.

# Data Privacy - Case study:

*Some may object to the ethics of gathering and releasing this data. However, all the data found in the dataset are or were already publicly available, so releasing this dataset merely presents it in a more useful form.* - Researchers Emil Kirkegaard and Julius Daugbjerg Bjerrekær

- When users gave this information was there an expectation of privacy?

# Data Privacy - Case study:

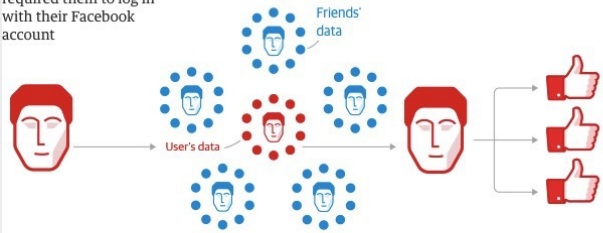
## Cambridge Analytica: how 50m Facebook records were hijacked

**1** Approx. 320,000 US voters ('seeders') were **paid $2-5 to take a detailed personality/ political test** that required them to log in with their Facebook account

**2** The app also **collected data such as likes and personal information** from the test-taker's Facebook account ...

**3** The **personality quiz results** were paired with their Facebook data - such as **likes** - to seek out psychological patterns

**4** Algorithms combined the data with other sources such as voter records to **create a superior set of records (initially 2m people in 11 key states*)**, with hundreds of data points per person




... as well their **friends'** data, amounting to over 50m people's raw Facebook data

These individuals could then be targeted with **highly personalised advertising** based on their personality data

Guardian graphic. *Arkansas, Colorado, Florida, Iowa, Louisiana, Nevada, New Hampshire, North Carolina, Oregon, South Carolina, West Virginia

# Data Privacy - Case study:

- About 320,000 US voters were paid to take a personality politics test
- The app also collected data about likes and personal information
- It also grabbed their friends data
- Even if you did not take the survey, your friends access to the survey might mean your data was included
- Algorithms combined the data to target people with highly personalized advertising based on their personality data.

# Data Privacy - Case study:

People did not realize that this is how their data was going to be used.

All around there are ethical issues around this type of data collection and use!

Carole Cadwalladr and Emma Graham-Harrison. How Cambridge Analytica turned Facebook 'likes' into a lucrative political tool. The Guardian. 17 March 2018.