

STAT 135

2. Critical thinking

Spring 2022

Lecturer: Dr Rebecca Barter (she/her)

Office hours: Tu 9:30-10:30, Th 1:00-2:00

Office: Evans 339

Email: rebeccabarter@berkeley.edu

Twitter: [@rbarter](https://twitter.com/rbarter)

GitHub: [rbarter](https://github.com/rbarter)

**Learning about the real world
from real data is hard**

Data is everywhere... and is of varying quality

Send us your feedback! X

What's your opinion on this web page?

Please select your feedback category below:

Suggestion Bug Compliment

Please leave your feedback below:

Please fill in your answer

Powered by mopinion



The Cornell Lab
of Ornithology



United States Census 2010 U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

This is the official form for all the people at this address.
It is quick and easy, and your answers are protected by law.

Use a blue or black pen.
Start here

The Census must count every person living in the United States on April 1, 2010.

Before you answer Question 1, count the people living in this house, apartment, or mobile home using our guidelines.

- Count all people, including babies, who live and sleep here most of the time.

The Census Bureau also conducts counts in institutions and other places, so:

- Do not count anyone living away either at college or in the Armed Forces.
- Do not count anyone in a nursing home, jail, prison, detention facility, etc., on April 1, 2010.
- Leave these people off your form, even if they will return to live here after they leave college, the nursing home, the military, jail, etc. Otherwise, they may be counted twice.

The Census must also include people without a permanent place to stay, so:

5. Please provide information for each person living here. Start with a person living here who owns or rents this house, apartment, or mobile home. If the owner or renter lives somewhere else, start with any adult living here. This will be Person 1.
What is Person 1's name? Print name below.

Last Name
First Name MI

6. What is Person 1's sex? Mark ONE box.
 Male Female

7. What is Person 1's age and what is Person 1's date of birth?
Please report babies as age 0 when the child is less than 1 year old.
Print numbers in boxes.
Age on April 1, 2010 Month Day Year of birth

→ NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.

8. Is Person 1 of Hispanic, Latino, or Spanish origin?
 No, not of Hispanic, Latino, or Spanish origin
 Yes, Mexican, Mexican Am., Chicano

Data is an imperfect reflection of reality



Photo by [Dasha Yukhymyuk](#) on [Unsplash](#)

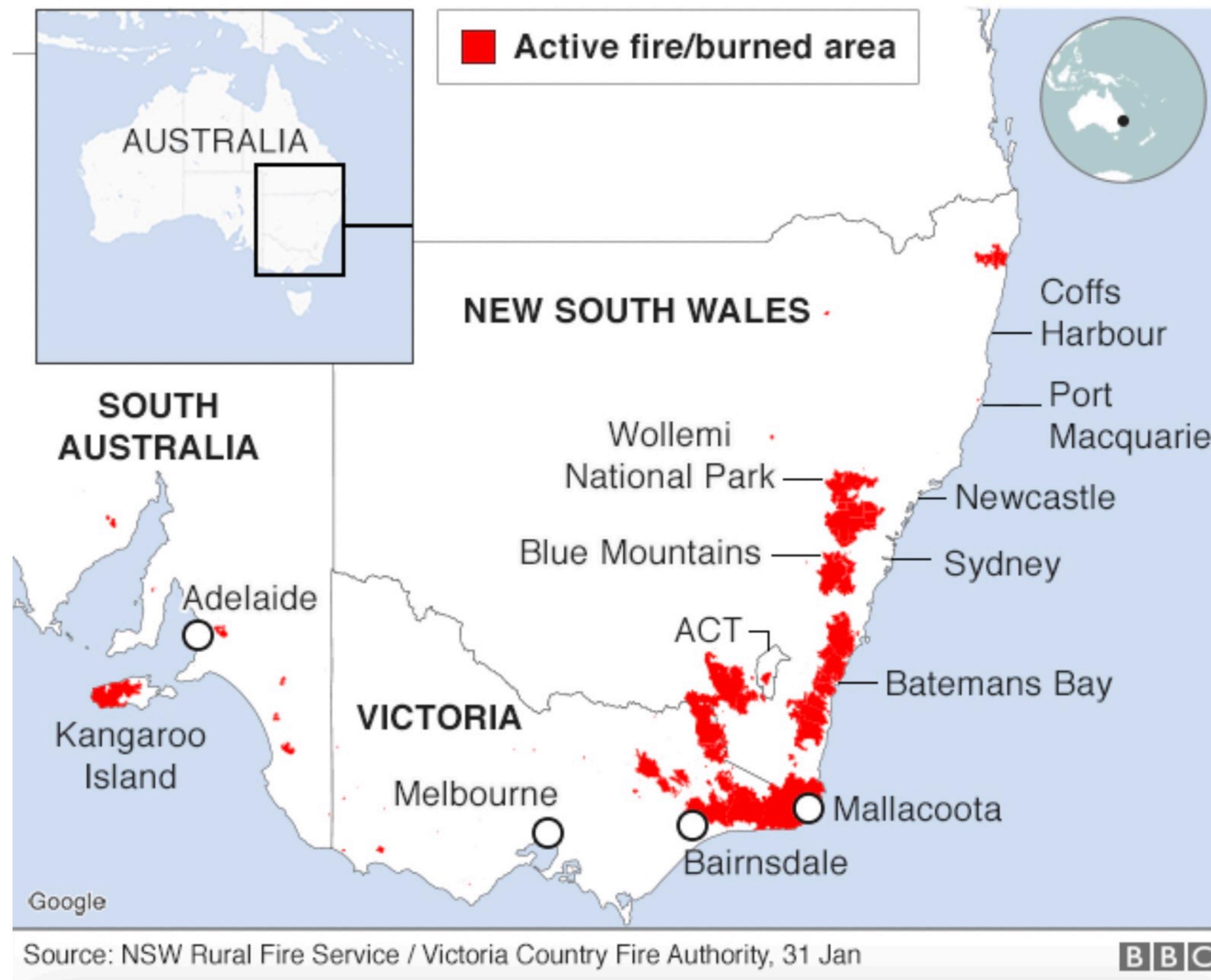
Can we learn from data that is an imperfect reflection of reality?

Sure we can!

We just have to be transparent about the limitations of the data and our analyses

An example

Remember the massive bushfires in Australia in early 2020?



Context is important

Remember the “dark day” (Sep 9 2020)?

...Those fires paled in comparison:

California burnt area: 4.6 million acres

Australia burnt area: 46 million acres



Reports started to emerge about the Australia fires....

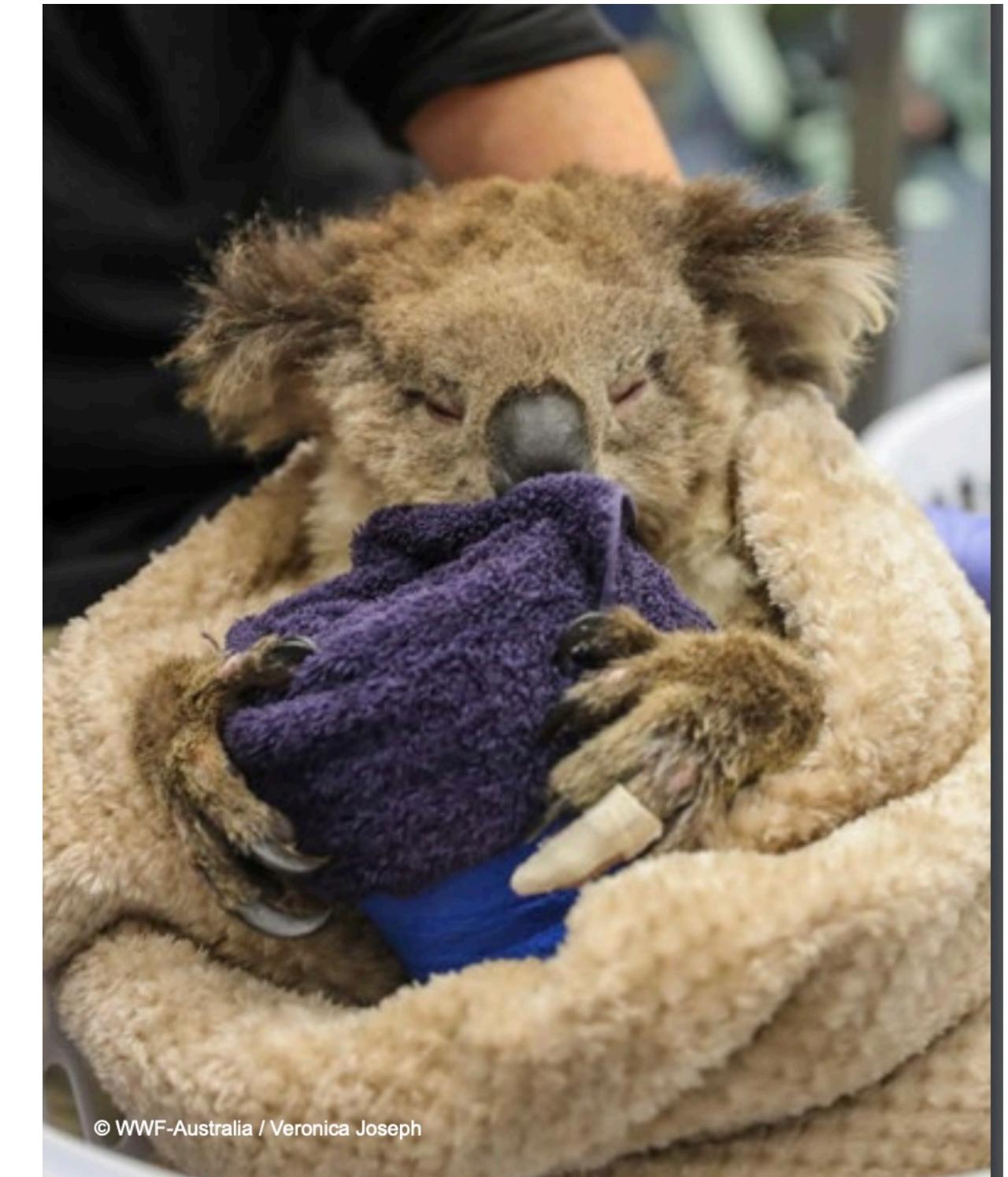
Over 1 billion animals killed



Discussion:

How do you think scientists came up with this number that “1 billion” animals were killed?

What are the real-world challenges involved in trying to identify the number of animals that were killed by the fires?



© WWF-Australia / Veronica Joseph

The World Wildlife Fund report:

(wwf_report.pdf on BCourses)

The estimation technique used:

Historical data of the densities of a few animal species in various Australian regions were extrapolated to a wide range of species across the burnt areas.

Discussion:

What are some limitations with this approach?

Would it be possible to collect “better” data?

The World Wildlife Fund report:

(wwf_report.pdf on BCourses)

Some of the limitations highlighted in the WWF Report:

- Not all animals who were in the area would have necessarily died
- The **geographic regions** on which the animal density data were based may not have been the same as the geographic region of the fires
- Much of the animal density data used in the study was based collected decades ago, and the densities may have changed dramatically since then.
- These historical studies are typically based on regions where these animals were *known* to have higher densities (extrapolating estimates based on this data might be **biased**)
- **Seasonality** was not taken into account
- The study did not take into account **fire intensity**: regions were treated as binary, where every geographic location was either not affected by fire or it was.
- There was no density data for some species (e.g., platypus and flying foxes), so they were not included in the total

The World Wildlife Fund report:

In total, we estimate that almost **3 billion** native vertebrates are likely to have been present within the 2019-20 bushfire areas. Our estimate comprises approximately:

- 143 million mammals
- 2.46 billion reptiles
- 181 million birds
- 51 million frogs



These are estimates of the numbers of individuals within selected taxon groups that were likely to have been present within the impact area of the 2019-20 bushfires. We will never know the true number of individuals killed by the fire, although future efforts to quantify this could be informed by our study in combination with research on other factors such as the ability of different taxa to survive fire.

6.1. THE DATA DEFICIT

There are many limitations in the data that prevent us from developing a more accurate estimate of the impacts of bushfires, especially during atypical intense and extensive fire seasons such as what occurred over 2019-20. These include data on densities of different taxa, an understanding of different species' abilities to survive and their responses to different fire regimes and

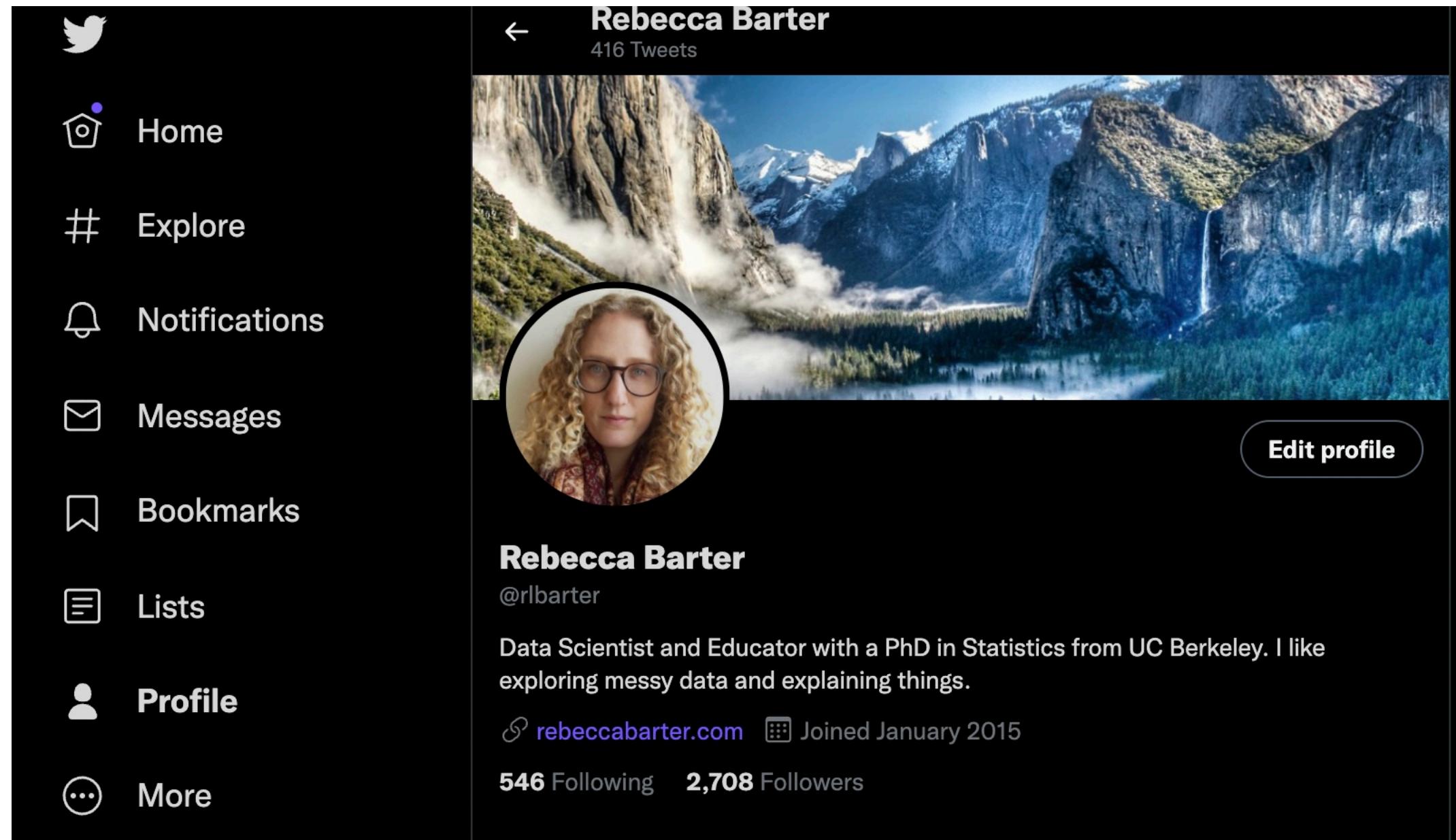
The best data there is is often the data that you have

Transparency is key

Always be transparent about the limitations of your data and your analyses

Defining “probability”

What is a probability?



Is my twitter account a bot?

Botometer[®]
An OSoMe project (bot•o•meter)



What is a probability?

Botometer[®]

An OSoMe project (bot•o•meter)



But what does it mean that the chance that I am a bot is 0.3/5?

What if I told you that the bot was trained based on the opinions of 10 graduate students?

What is a probability?

How would you interpret the following statements?

1. There is a 50% chance that when I flip a coin the outcome will be “heads”
2. There is an 80% chance that an earthquake of magnitude 6.0 or larger will occur in the Bay Area over the next 30 years
3. There is a 1% chance that Matthew McConaughey will be elected the governor of Texas
4. There is a 50% chance that it will rain tomorrow

In many situations, probabilities are not well-defined

Frequentist: the probability of an event is based on the relative frequency with which the event tends to occur across a repeatable process

Bayesian: the probability of an event is based on your *belief* (which is based on prior knowledge and data-based evidence) that it will occur

Neither perspective is necessarily the “correct” one

Sometimes one perspective will make more sense for a particular problem than the other

What is a probability?

Which of the following statements might make more sense in a Frequentist context vs a Bayesian one?

1. There is a 50% chance that when I flip a coin the outcome will be “heads” **Frequentist**
2. There is an 80% chance that an earthquake of magnitude 6.0 or larger will occur in the Bay Area over the next 30 years **Bayesian**
3. There is a 1% chance that Matthew McConaughey will be elected the governor of Texas **Bayesian**
4. There is a 50% chance that it will rain tomorrow **?**

In this class, we will *mostly* be Frequentists