Data Ethics: Understanding Big Data, Algorithmic Bias, and Research Ethics

SOCL 1101 Introduction to Sociology
Ineke Marshall
Fall 2020

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Discussion: China's Social Credit System

- What is China's Social Credit system? How does it work?
- In what ways might America have similar or different technological infrastructures?

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Workshop Agenda

- Objectives

- Introduce 'Big Data' Concepts

- Discuss data, privacy, and algorithms

- Activity: Adopt or Not?

- Discuss ethical implications of big data and lessons for (digital) research

Slides, handouts, and data available at

http://bit.ly/diti_fall2020-marshall1

Feel free to ask questions at any point during the presentation!

Workshop Goals

- Understand the ways data is being used in society as well as how algorithms impact and shape our daily lives
- Explore the ways in which privacy and security are being reshaped and redefined through the use of big data, algorithms, and policy
- Understand the ways in which technology reflects cultural, social, and political biases.
- Explore the ways in which these questions and methods are influencing how social scientists do research and practice their craft

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

What is "Big Data"?

Companies, governments, and other groups collect vast amounts of data ("big data") from vast amounts of users and analyze that data quickly for particular purposes (advertising, surveillance, search results, etc).

The goal of collecting and processing these data is to predict individual user behavior based on patterns from the user as well as patterns from "similar" users (based on demographic information, behavioral patterns, etc).

Northeastern University
NULab for Texts, Maps, and Networks

Why should we care?

- Big data is characterized by its **scale**
- Big data **sources** include: digitized records, social media/internet activity, or sensors from the physical environment.
- Big data is often **privately owned**
  - Example: an insurance company purchasing social media activity from Facebook in order to make insurance sales decisions.

Northeastern University
NULab for Texts, Maps, and Networks

Questions to consider

- How are we being represented online?
- How is our data being used?
- Who is using it and for what purposes?
- How might it be used in the future?

Feel free to ask questions at any point during the presentation!
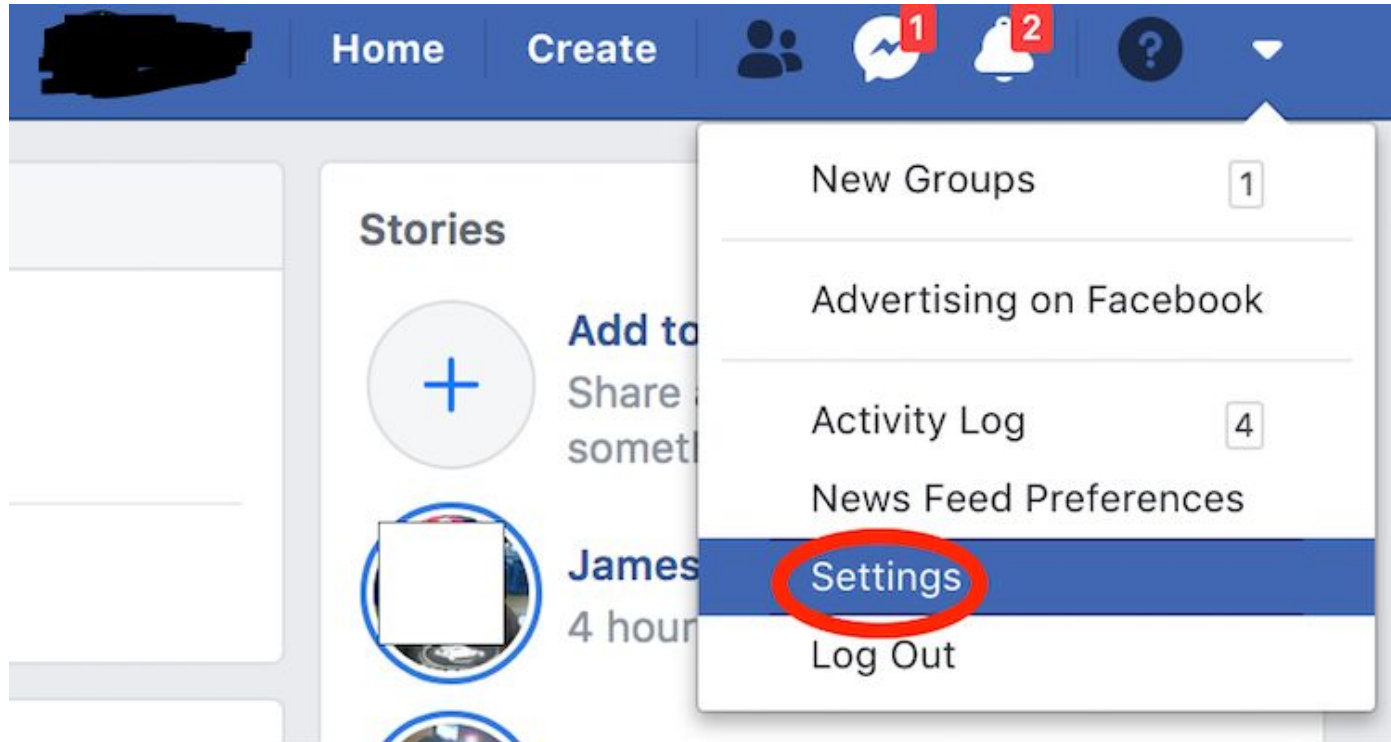
Facebook Preferences

Facebook collects, stores, and sells information about you so you get more targeted ads and your newsfeed is tailored to your categories.

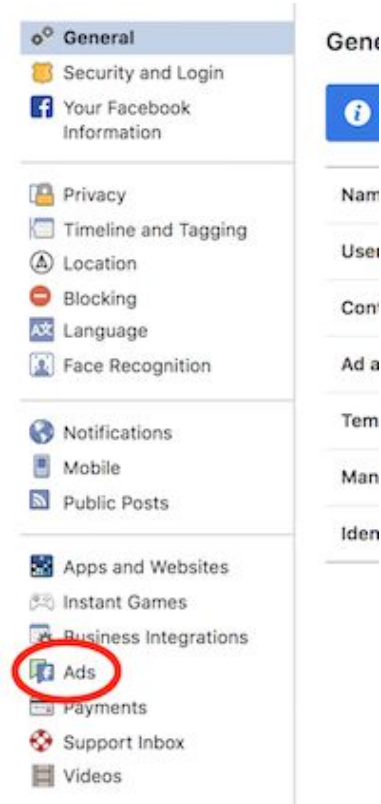Other social media sites that do this:

- Instagram (owned by Facebook)
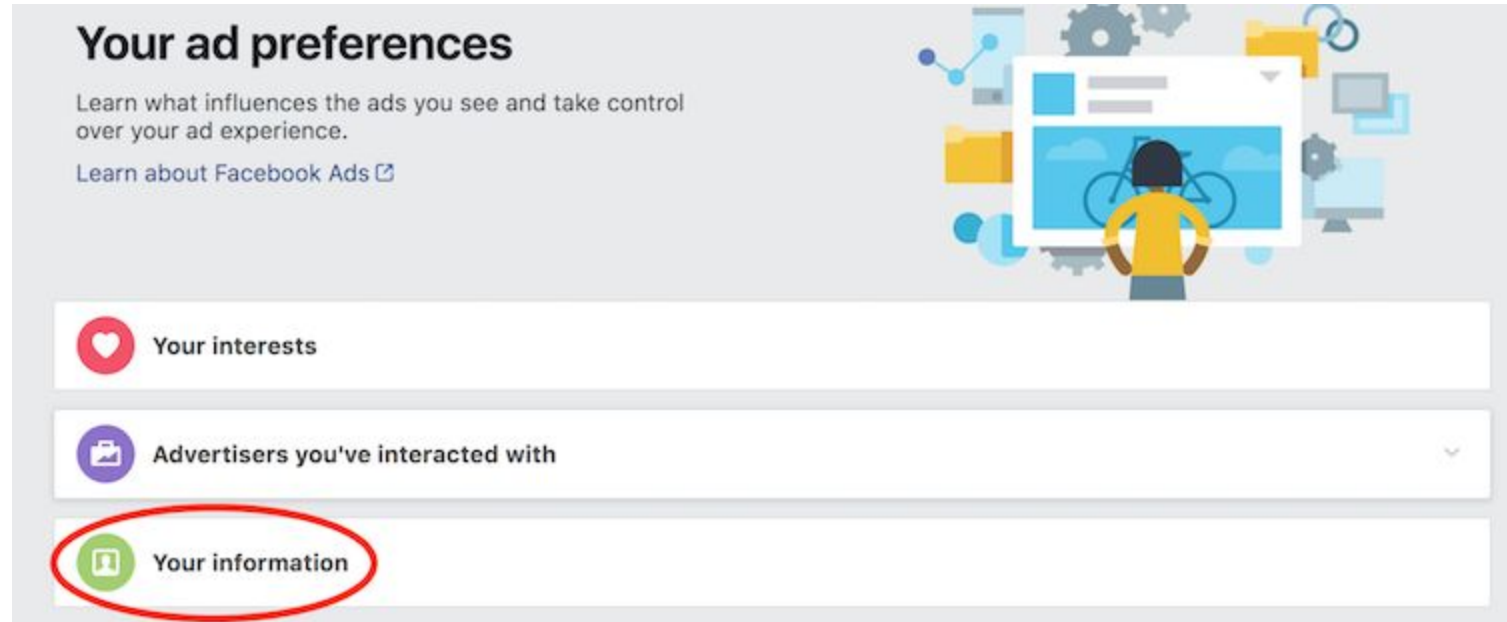- Google
- YouTube (owned by Google)
- Twitter
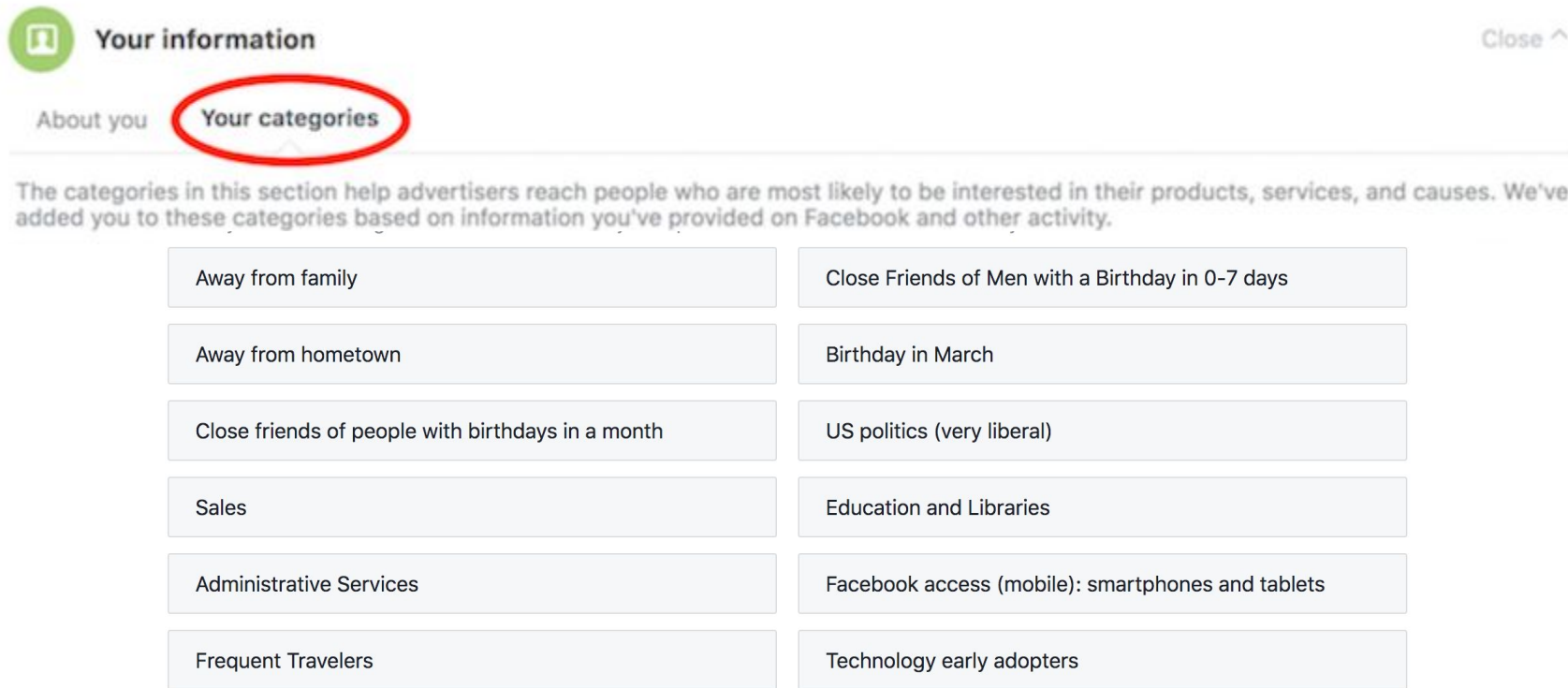
Northeastern University
NULab for Texts, Maps, and Networks

Settings > Ads > Your information > Categories

Feel free to ask questions at any point during the presentation!

Settings > Ads > Your information > Categories

Feel free to ask questions at any point during the presentation!

Settings > Ads > Your information > Categories

Feel free to ask questions at any point during the presentation!

Settings > Ads > Your information > Categories

Your information   Close ⌃

About you   **Your categories**

The categories in this section help advertisers reach people who are most likely to be interested in their products, services, and causes. We've added you to these categories based on information you've provided on Facebook and other activity.

Away from family	Close Friends of Men with a Birthday in 0-7 days
Away from hometown	Birthday in March
Close friends of people with birthdays in a month	US politics (very liberal)
Sales	Education and Libraries
Administrative Services	Facebook access (mobile): smartphones and tablets
Frequent Travelers	Technology early adopters

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Google's File on You is 10 Times Bigger Than Facebook's — Here's How to View It

Google, Amazon, Apple, and Microsoft are all central players in "surveillance capitalism" and prey on our data.

THE ANTI-MEDIA

Example: If you have **location services** turned on for Google (like if you use Google maps), Google can track your every move. Go to:

https://www.google.com/maps/timeline

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Downloading Your Data

Facebook: Settings > Your Facebook Information > Download your Information

Google:

https://support.google.com/accounts/answer/3024190?hl=en

Instagram app: Settings > Security > Access Data/Download data

Feel free to ask questions at any point during the presentation!

DIY Cybersecurity and Tightening your Privacy

Want to make your life more private? Follow this "DIY Guide to Feminist Cybersecurity"

https://hackblossom.org/cybersecurity/

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

"Big Data" Unbounded - Ethical Issues

- Controversies in the recent years:
  - Cambridge Analytica 2016 elections [controversy](#)
  - [Clearview AI](#): facial recognition "services" in 2020
  - General [use of facial recognition in policing](#) in recent years
  - Place of algorithms in [racially differential health outcomes](#)
  - Using algorithms in [grading in the UK](#) in 2020
  - And many many more all across the world…
- "Big data" raises questions of power, autonomy, anonymity, privacy, discrimination, and bias.

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Algorithms, race, and health

- Source: [Obermeyer et al. 2019](#).
- Healthcare systems in the US use an algorithm that tries to predict which patients have needs that warrant special care programs
- Black patients ended up having to have greater number of chronic health issues to qualify for special care
- Algorithm explicitly does not take race as input
- The things it takes as input are effectively using the healthcare costs as a proxy for the health needs - this leads to algorithm predicting what will be the costs of each patient
- However, since Black people historically generate less costs for the same amount of health issues, it leads to the outcome outlined above

Feel free to ask questions at any point during the presentation!

# Algorithms and Bias - Activity

Feel free to ask questions at any point during the presentation!

Algorithms

Do you rely on algorithms in your everyday life? Any examples?

An algorithm is a process of instructions provided, usually for computers to interpret and follow. There is usually an **input**, which is determined by the programmer; then there is a set of rules (the algorithm) that help lead to the **output**, or the results of the program following instructions.

Algorithms can be fairly simple, but they can also be much more complex.

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!
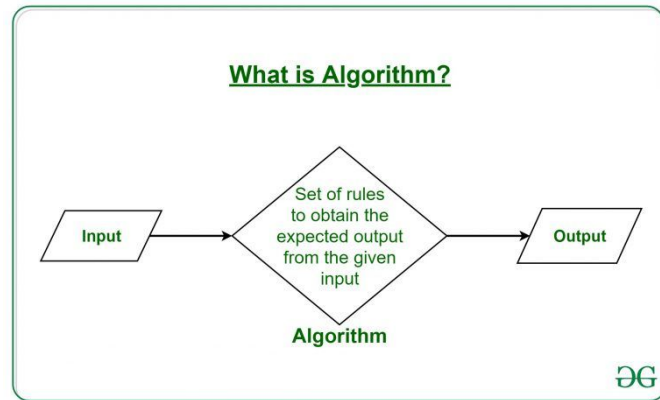
Activity: Adopt or Not?

Small Group: You will be assigned into small groups. You work for an adoption agency and have to decide if someone can adopt a dog. On your handouts, please read the four previous adoption applications and decide if the new adoption applicant can adopt or not.

Do you think this new applicant should be allowed to adopt a dog? Why or why not?

Class Discussion: Adopt or Not?

- Would you ACCEPT or REJECT their application? Why?
- What questions from the application did you weigh more? Why?
- What might be some implicit biases in this application form, the process, and in your choices?

Feel free to ask questions at any point during the presentation!

Adopt or Not? Algorithm

Algorithms "read" through data such as these applications, and often help us make decisions. Here are some questions to think about when assessing algorithms:

- Where might you see these algorithms being used to make decisions? Why are they being used? What are they replacing or adding on to?
- What biases may be ingrained in the data collected for the algorithms? What biases may be ingrained in the actual process of using the algorithm?
- In what ways might the algorithm prevent or reinscribe human bias?

Feel free to ask questions at any point during the presentation!

Want to learn more about accountability and best practices when creating algorithms?

Visit https://www.fatml.org/, or Fairness, Accountability, and Transparency in Machine Learning

Feel free to ask questions at any point during the presentation!

So what do 'big data' & algorithms have to do with research?

Feel free to ask questions at any point during the presentation!

Questions Researchers Must Ask

- What **information** is being collected and from where? To whom does this data **belong**?
- How is it being **collected**? Do **participants** know that it is collected, how it will be collected, and how will it be used?
- **How** will the data be analyzed? What **biases** and **ideologies** may be implicit in this analysis?
- Who will this research impact? Who will it **benefit**? Who will it potentially **harm**?

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Discussion

- What are some benefits and what are some risks coming with the increased focus on "big data" in research and policy?

- Do you see any commonalities among the risks and how would you address them?

- Finally, big data is a vague term and refers to many different phenomena at once. What are some other terms you would use in order to think about these issues more precisely?

Feel free to ask questions at any point during the presentation!

Thank you!

If you have any questions, contact DITI at nulab.info@gmail.com

Developed by Cara Messina and Garrett Morrow
Digital Integration Teaching Initiative
DITI Research Fellow

Taught by Milan Skobic
Digital Integration Teaching Initiative
DITI Research Fellow

Slides, handouts, and data available at **http://bit.ly/diti_fall2020-marshall1**

Schedule an appointment with us! **https://calendly.com/diti-nu**

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!