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

SOCL 1101 Introduction to Sociology
Ineke Marshall
Fall 2021

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
https://bit.ly/diti-fa21-marshall-data-ethics

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

Workshop Goals

- Understand the ways data are 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

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

What is “Big Data”?

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

Big Data is here (and it's getting *bigger*)

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 from vast numbers of users** and analyze that data quickly for a variety of purposes, including advertising, marketing, surveillance, building profiles, etc.

The goal of big 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).

If we’re living in an era of “surveillance capitalism,” **our information can be considered to be a valuable *product*.**

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

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

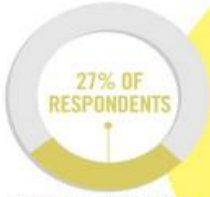
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

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.

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

Online Presence
&
Data Privacy

Northeastern University
NULab for Texts, Maps, and Networks

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

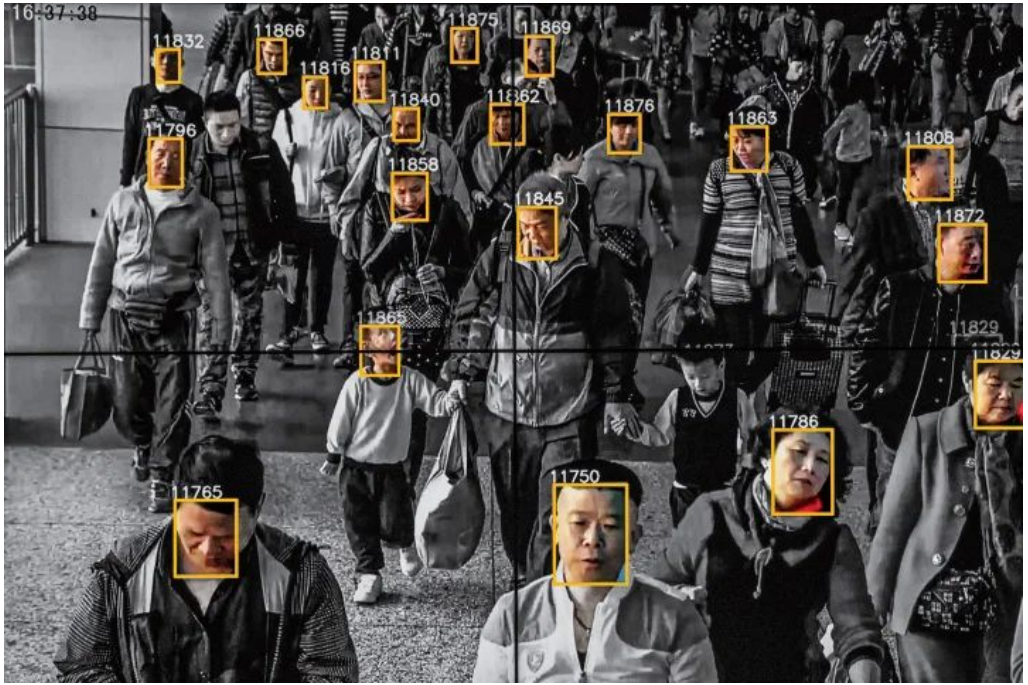
Questions to consider

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

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

An Example: China's Social Credit System

- What is China's Social Credit system? How does it work?



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

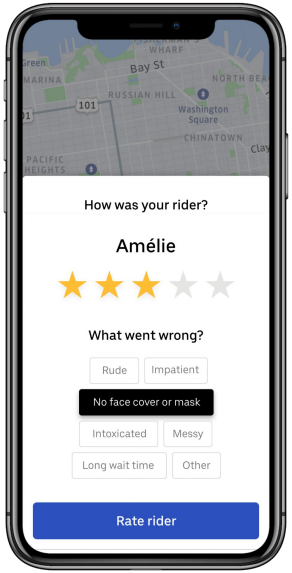
Black Mirror: *Nosedive* (2016)

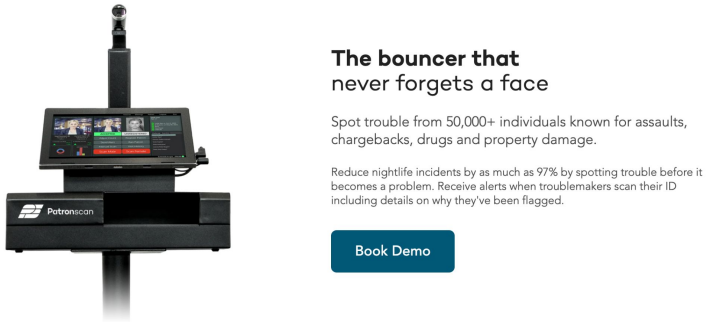
Northeastern University
NULab for Texts, Maps, and Networks

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

Discussion: America's Social Credit System

In what ways might America have similar or different technological infrastructures when compared with China?





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

What about Social Media?

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

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

AWARENESS | SCIENCE & TECH | AUG 3, 2019 AT 11:08 AM.

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

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

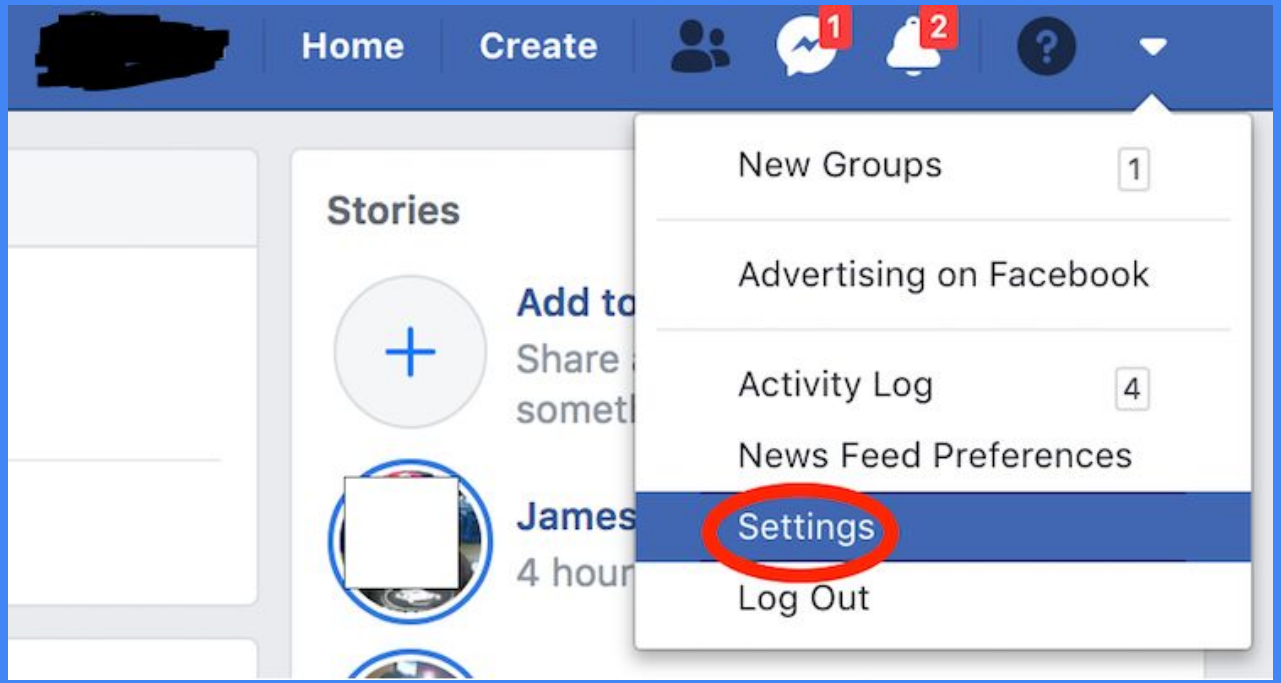
Image and Audio Information

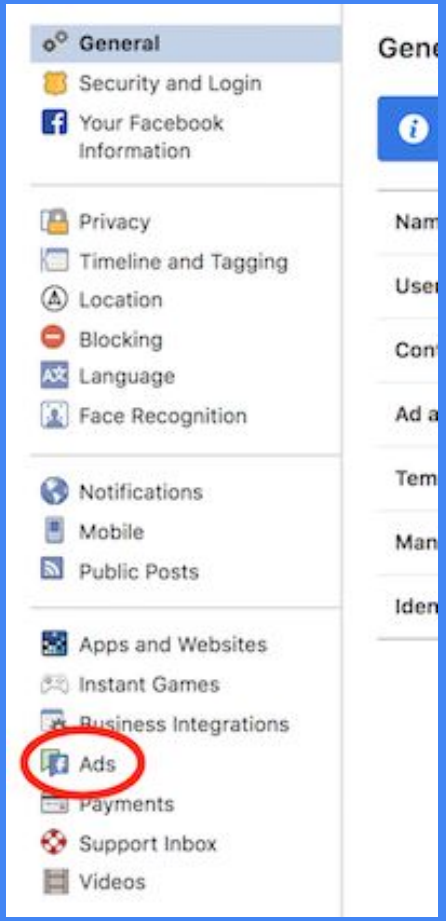
We may collect information about the images and audio that are a part of your User Content, such as identifying the objects and scenery that appear, the existence and location within an image of face and body features and attributes, the nature of the audio, and the text of the words spoken in your User Content. We may collect this information to enable special video effects, for content moderation, for demographic classification, for content and ad recommendations, and for other non-personally-identifying operations. We may collect biometric identifiers and biometric information as defined under US laws, such as faceprints and voiceprints, from your User Content. Where required by law, we will seek any required permissions from you prior to any such collection.

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

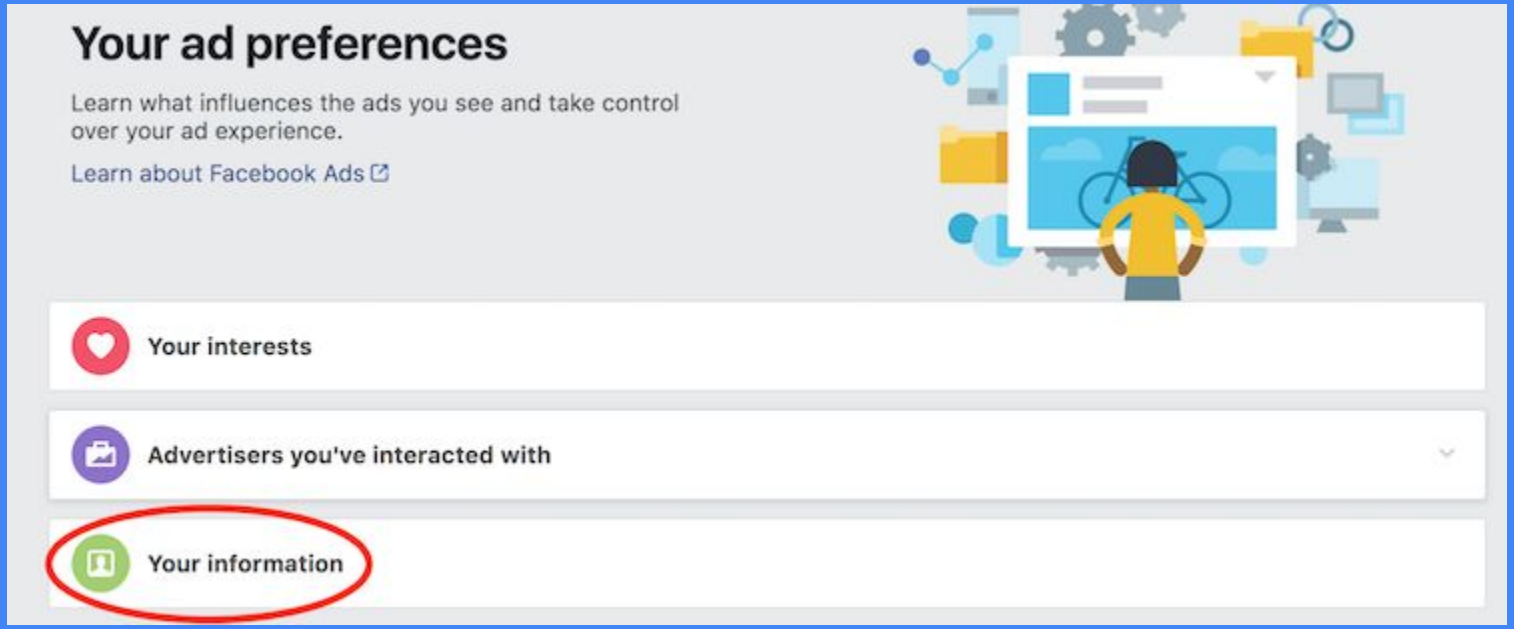
Your Privacy on Facebook

<u>Find Out</u>:
Settings > Ads > Your information > Categories

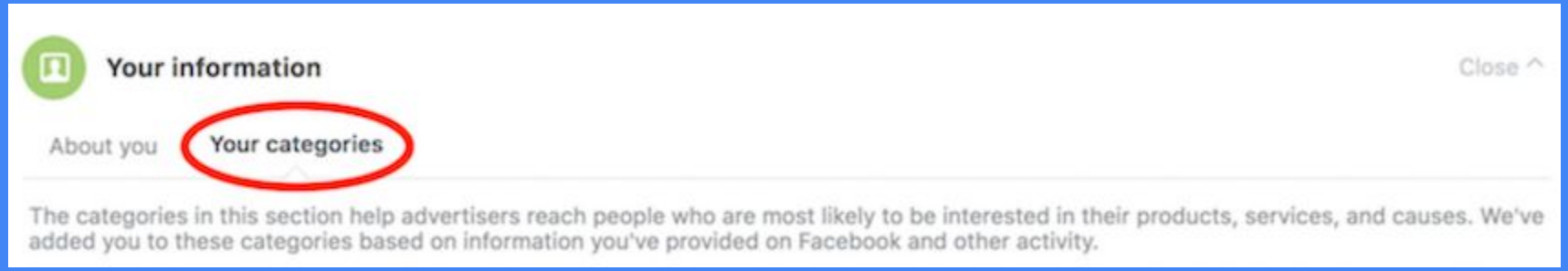




Feel free to ask questions at any point during the presentation!



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



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

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: Settings > Privacy and Security > Data download/Request Download

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!

Issues in Big Data: Ethics and Algorithmic Bias

Northeastern University
NULab for Texts, Maps, and Networks

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

Algorithms

- 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.
  - Algorithms can be fairly simple, but they can also be much more complex.
- "**Machine learning**" happens when an algorithm tells a computer to make decisions based on a set of patterns derived from data, instead of following specific predetermined instructions.

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

“Big Data” Unbounded — Ethical Issues

Some recent controversies:

- **Cambridge Analytica controversy**: psychological profiles of American voters
- **Racial bias in health algorithms**: results in reduced access to care for Black people
- **Use of facial recognition**
  - Clearview AI: sells facial recognition “services”
  - Case of Robert Williams: wrongfully arrested
  - Machine Bias: Software used to predict future criminals, biased against Black men
  - Stanford study creates AI that can predict sexual orientation based on a photo with up to 91% accuracy

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

It’s not all bad . . .

- Prof. Lazar and NetSI researchers, at Northeastern, working on COVID-19
- Algorithms predicting the likelihood of cancer (Breast cancer, Prostate cancer)
- Allegheny County PA “family screening tool” to support human screeners in the Department of Children, Youth, and Families

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

Identifying Bias: Some Guiding Questions

- In what way(s) is the software used in each scenario biased?
- Do technology and big data-driven solutions **eliminate** human bias or **amplify** it?
- What can be done to decrease bias and improve data-driven decision-making software?

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

The takeaway?

Algorithms are NOT neutral!

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

Algorithms and Bias Activity

Northeastern University
NULab for Texts, Maps, and Networks

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

Activity: Data deciding dog adoption

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?

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

Let's Discuss

Please elect one representative from your group to explain your group's responses to the following questions:

- 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!

Moving Forward

How can we be cognizant of 'big data' & algorithms in our 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**?

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 DITI Research Fellows:
Tieanna Graphenreed, Vaishali Kushwaha, Cara Messina, Yana Mommadova, Garrett Morrow, and Claire Tratnyek

Slides, handouts, and data available at

https://bit.ly/diti-fa21-marshall-data-ethics

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

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