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

Introduction to Contemporary Moral Issues
Professor Katy Shorey
Fall 2021

Northeastern University
NULab for Texts, Maps, and Networks

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

Workshop Agenda

- Introduce ‘Big Data’ Concepts - what, exactly, is “Big Data”?
  - What makes it “Big,” where does it come from, how is it being used?
- Discuss data privacy and data ethics in the context of social media (and doxing)
- Explore machine learning and algorithmic bias
- Think critically about the Skid Row ethics case study through the lenses of data collection and representation

Slides, handouts, and data available at **https://bit.ly/diti-fa21-shorey**

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!

Why should we care?

- Big data is **omnipresent**—its **sources** include: digitized records, internet activity, and even sensors from the physical environment
- Big data is often **privately owned** and it is hard to ensure oversight over how it is developed, used, and controlled
- The **scale** of big data enables those who use, develop, and control it to magnify their influence
- Big data can be used to (inadvertently or purposefully) **entrench stereotypes** or **reproduce results** that may harm certain communities

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

Big Data:
Sources and Uses

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 is data about us being used?
- Who is using our data and for what purposes?
- How might our data be used in the future?
- **How does Big Data impact our daily lives?**

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

How does Big Data impact our daily lives?

Entertainment media (music, shows, movies)

Healthcare and medical services

Shopping and marketing **Travel and transportation**

Education and Employment **News and Information**

Public policy and safety

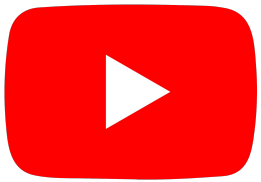
Northeastern University
NULab for Texts, Maps, and Networks

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

Social Media Preferences & Targeted Ads

You are categorized by your series of behaviors and identity markers.

Social media sites collect, store, and sell information about you, so that you get better targeted ads and your newsfeed is tailored to your categories. **Some social media sites that do this:**


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.

Doxing and Social Media

- **The ubiquity of images and videos on the Internet make it virtually impossible to not leave digital footprints.**
  - Problems? Privacy? Safety?
- Doxing might provide a way to challenge those who are in power but…
- … it can also be used as a tool for online harassment and bullying
  - **Where do we draw the line between freedom of speech (and privacy rights) and social activism?**

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

Downloading Your Data & Tightening your Privacy

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

Want to make your life more private? Follow this "DIY Guide to Feminist Cybersecurity" https://hackblossom.org/cybersecurity/

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

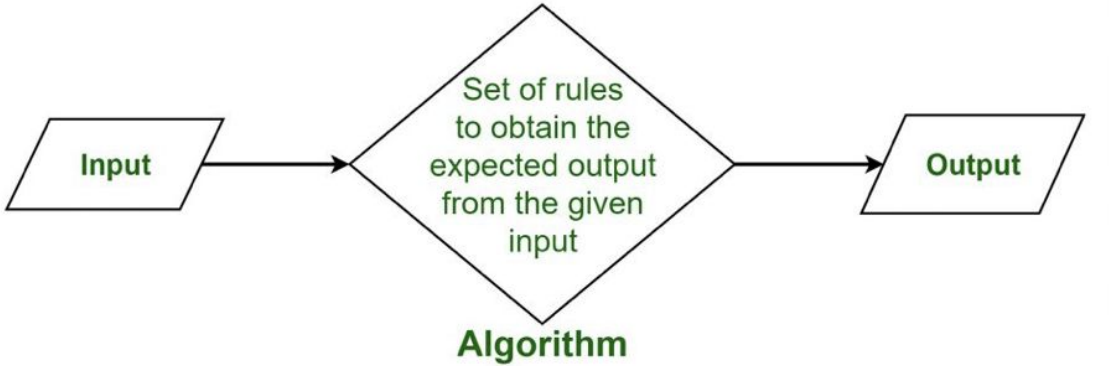
Big Data Ethics and Algorithmic Bias

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.



- "**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!

Algorithmic Bias

Algorithms are *not neutral*. **People create algorithms.**

The algorithmic processes, and even the data itself, reflect societal biases.

When an algorithm is written or trained using data that does not adequately represent/reflect the actual population (because the sample only captures a particular demographic, and other groups are under- or unrepresented), this creates **Algorithmic bias.**

Similarly, **when data reflects biased realities**, the algorithm will continue to **reproduce and reinforce outcomes** if those outcomes are desirable (despite their harm to - or erasure of - other groups).

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

Questions to consider:

- What are some **benefits** and what are some **risks** coming with the increased focus on ‘big data’ in research and policy?
- Are technology- and big data-driven solutions more likely to **eliminate** human bias or **amplify** it?
- In any case study, where can we find **data-driven** analyses, possible solutions, or policy arguments?
    - How can we critically analyze these to determine whether the **data is being used ethically**?

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

Data-Informed Policies: Approaches to Skid Row

Northeastern University
NULab for Texts, Maps, and Networks

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

Los Angeles' Skid Row

What are the **moral issues** involved with unhoused people living in cities?

In what ways could **data inform** Skid Row policy decisions and implementation?

Entire blocks are packed with homeless encampments on skid row in downtown Los Angeles. (Luis Sinco / Los Angeles Times)

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

Discussion: Skid Row case study

- What **data** could be helpful in informing policy decisions re: skid row?
- How might this data be **collected**? **verified**? **used**?
- How should we go about **evaluating** the solutions that big data 'gives' us?
- What **challenges** do we see in undertaking data collection and analysis in this case?

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

Collecting Data on Unhoused Populations

- **Measuring the extent of homelessness is essential to combating it**, and efforts to count the homeless population have evolved significantly since the early 1980s.
- A combination of Homeless Management Information Systems, Point-in-Time counts, and Housing Inventory Counts inform policymakers and advocates on demographics, trends, and the availability and usage of services among America's homeless population.
- **Improved accuracy and detail of homeless data** have influenced all aspects of HUD's policies as well as those of its partner agencies.

SOURCE: **Using Data to Understand and End Homelessness (HUD,2012)**

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

“What gets counted counts”

- **“What is counted—like being a man or a woman—often becomes the basis for policymaking and resource allocation. By contrast, what is not counted—like being nonbinary—becomes invisible…”**
- Again, what kinds of data are being counted, and what kinds of data are being overlooked, ignored, excluded?
  - What are the consequences of counting and not counting different kinds of data on populations experiencing homelessness?
- The two approaches to homelessness as a moral issue we’ll discuss next both may each use data-driven arguments…
  - what kinds of data might support each approach?
  - are these kinds of data accurate? complete? appropriate? valid?

SOURCE: **“What Gets Counted Counts” Principle #4 of Data Feminism (mitpress, 2020)**

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

Homelessness as a Moral Issue

- The Liberal Approach to Homelessness
- The Care Approach to Homelessness
- Financial and Political Implications of Both

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

The Liberal Approach

- Are we willing to tolerate an economic system in which large numbers of people are homeless?
- Yes… because this economic system has brought the prosperity **for most of "us,"** we take as a given the fact that **there is some unavoidable collateral damage** in the form of those who, for some reason, cannot keep up, and fall through the cracks of the system.
- All we can try to do is mitigate the damage...

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

The Care Approach

- Not having a place that one can call one's home is a **violation of fundamental human needs** and hence unacceptable.
- Homelessness is when a subclass of citizens is concerned that for some reason they are trapped in a situation of dependency, need and suffering.
- It is our moral imperative to extend the liberal focus on autonomy with a care perspective in the case of co-citizens who are dependent and cannot be considered sufficiently autonomous.

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

Discussion

- **It is almost natural to side with the care approach and yet…**

- … whose moral responsibility? Individual? Government?

- How do we address the question within the confines of a liberal-capitalist political-economic system?

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, Colleen Nugent, Jeff Sternberg, and Claire Tratnyek

Slides, handouts, and data available at **https://bit.ly/diti-fa21-shorey**

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

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