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

Sociology of the Family
Linda Blum
Fall 2020

Northeastern University
NULab for Texts, Maps, and Networks

Workshop Agenda

- Objectives and Goals
- Introducing 'Big Data' Concepts
- Algorithmic bias
- Algorithm activity
- Implications for research
- Discussion of the readings: algorithms and family policies in the US

Slides available at **http://bit.ly/diti_fall2020-blum2**

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

Workshop Goals

- Understand the ways that data is being used in society as well as how algorithms impact and shape our daily lives
- Understand the ways in which technology reflects cultural, social, and political biases
- Explore the ways in which privacy and security are being reshaped and redefined through big data, algorithms, and policies related to various aspects of lives - public, social, economic, etc.
- Explore the ways in which these questions and methods are influencing how social scientists do research

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

Big Data

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

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 (advertizing, 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).

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

Why should we care?

- Big data is characterized by its **scale**
- Big data **sources** include: digitized records, social media/internet activity, and 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

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!

Another example: Algorithms, race, and health

- Source: [Obermeyer et al. 2019](#).
- Healthcare systems use an algorithm that tries to predict which patients have needs that warrant special care program
- 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

Northeastern University
NULab for Texts, Maps, and Networks

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

Algorithms and Bias

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.

What is Algorithm?

Input → Set of rules to obtain the expected output from the given input → Output

Algorithm

GeeksforGeeks

Northeastern University
NULab for Texts, Maps, and Networks

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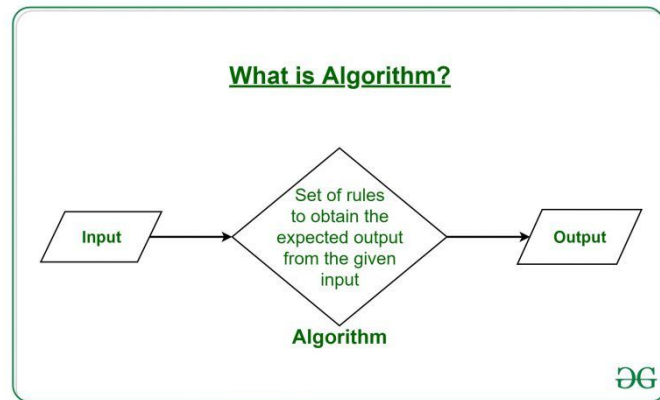
Activity: Adopt or Not?

Small Group: Find a partner or two. 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 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!

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?

Northeastern University
NULab for Texts, Maps, and Networks

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

Algorithmic Bias

Algorithms are *not neutral*. They are created and used by people and therefore reflect the biases of their creators and users. Those biases are often implicit.

For example, Amazon attempted to create an algorithm to analyze potential hires' resumes. Their input data was people who had been hired at tech companies and people who were not hired. Because tech companies are male-dominated, the input data reflected this. Many things the algorithm was concerned with, such as hobbies or the language use, are correlated with gender. The company valued things that are historically correlated with masculinity, ending up disproportionately rejecting applications by women.

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

So what can we do?

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!

Initiatives for Justice

Code for America is an organization that works with the mass amount of undigitized, unorganized government documents to help previously incarcerated people.

One project they have, titled "Clear My Record," attempts to parse through the mass data of governmental records to help clear criminal records, particularly for people who were arrested for marijuana use/distribution.

https://www.codeforamerica.org/programs/clear-my-record

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

Readings discussion

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

Reminder: Racism in foster care systems (Roberts 2017)

- What are the features of the foster care system Roberts criticizes?
- How does that system relate to racial stereotypes?
- How does it relate to political institutions?
- What are the effects of the system on the families and neighborhoods?

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

Discussion: Pittsburgh's CYF algorithm (Hurley 2018)

- What does the CYF office use the Family Screening Tool for? What are the stated goals?
- What changes does this algorithm bring to the officers?,
- What impact does the algorithm have on the families?
- Does the algorithm remove the institutional bias or shifts it onto another mechanism?
- How does the introduction of predictive algorithm relate to the institutional racism Roberts wrote about?
- In what other ways could you see similar or different technological infrastructures used in the US?

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

Wrap-up discussion

- What are the promises brought by increasing reliance on data and algorithms?
- What are the dangers?
- Do you see yourself relying on big data and algorithms in your research? How would you deal with the ethical issues?

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, Jeff Sternberg, Colleen Nugent, and Milan Skobic
Digital Integration Teaching Initiative
DITI Research Fellows

Taught by
Milan Skobic
Digital Integration Teaching Initiative
DITI Assistant Director
Talia Brenner
NULab Co-coordinator

Slides, handouts, and data available at **http://bit.ly/diti_fall2020-blum2**
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!