Data Ethics: Understanding Big Data and Algorithmic Bias

Milan Skobic and Vaishali Kushwaha
Advanced Writing in the Disciplines - ENGW 3307
Cecelia Musselman
Spring 2021

Northeastern University
NULab for Texts, Maps, and Networks

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

Workshop Agenda

- Discuss data, privacy, and data categorization examples
- Introduce ‘Big Data’ concepts and algorithmic bias
- Activity: Adopt or Not?
- Further examination of data-related practices in our everyday lives

Slides, handouts, and data available at

http://bit.ly/diti-spring2021-musselman

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

Workshop Objectives

- Understand the ways data is collected and analyzed 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 policy.
- Engage with critical rethinking of everyday practices related to data collection and use

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

Technology, data and COVID-19 pandemic

- **Example:** Philly Fighting COVID's privacy policy
- This start-up ran a vaccination operation and collected personal data through registration for weeks without a privacy policy, although one was eventually added
- There are indications that the policy text was generated by algorithm

> "You, the USER, agrees to allow the storage within any person for which you are legally entitled to for such purposes as may be necessary to provide you with services and information through COVIDReadi."
>
> - From "Philly Fighting COVID" privacy policy

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

Some questions

- What was the start-up doing with people's data before and after the introduction of their privacy policy?
- What was the role of the Philadelphia Health Department in this situation?
- Does this privacy policy address all of the potentially important concerns?
- What essential points should privacy policies address?
- Should organizations state clearly if they are using data for commercial purposes, or any other purposes?

More info on this story here

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 numbers of users and analyze these 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).

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

Why should we care?

- Omnipresence: Big data **sources** may include: digitized records, social media/internet activity, and sensors from the physical environment.
- Ownership and control: Big data is often **privately owned or beyond oversight**
  - Example: an insurance company purchasing social media activity from Facebook in order to make insurance sales decisions.
- Bias: Big data can often reproduce results that may **harm** certain communities.

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, Google can track your every move. Go to:

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

Technology does not bring neutrality

Algorithms, information systems, and systems for data collection and analysis are **not neutral**. They can reinforce and make explicit systemic, political, and cultural biases. They are affected by input data, the way that data is presented, how the data is interpreted by machines, and more. This means we also have the ability to challenge these biases, norms, and forms of discrimination.

For example, Amazon's algorithm for hiring employees in 2018 reflected the historical male domination in tech companies. The algorithm taught itself to interpret any mention of "women" in the new resumes as negative and rejected these applications.

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

Ethical Implications

- Big data also raises questions of power, autonomy, anonymity, privacy, discrimination, and bias.
- 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!

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!

Algorithms and Bias

Northeastern University
NULab for Texts, Maps, and Networks

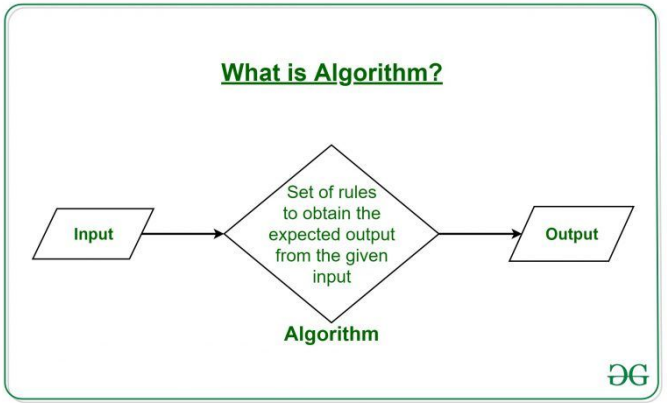
Feel free to ask questions at any point during the presentation!

Algorithms

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

An algorithm is a set of instructions, 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.



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

Facial recognition in policing and beyond

- The case of Robert Williams, who was wrongfully arrested in 2019
- Most algorithms have been found to contain gender and racial bias when tested on the accuracy of their face recognition
- These issues were present from the onset of the implementation of these technologies, but they still got introduced
- These biases are reproduced both in the programming of algorithms, and in collection of datasets from which algorithms are trained

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

Activity: Adopt or Not?

Small Group: You will be separated into break-out rooms with a few colleagues. In this scenario, you all 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 can “read” through data such as these applications, and 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 algorithms?
- In what ways might the algorithms prevent or reinscribe human biases?

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

More examples of introducing bias

- "Falsehoods programmers believe about names" - Patrick McKenzie
- How does a program recognize a "name"?
- Criteria that programmers insert in order to enable a program to recognize a "name" are informed by programmers' assumptions and biases

Some examples of the assumptions from the text:

- People's names fit within a certain defined amount of space.
- People's names do not change.
- People's names are written in ASCII.
- People's names are written in any single character set.
- People's names are case sensitive.
- People's names are case insensitive.

Can you think of any examples?

You can find some examples here as well

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 data ethics have to do with our lives?

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

Important Questions

These are questions *all* professionals working in education, tech & tech-related industries must be thinking about:

- What **information** is being collected and from where? To whom does this data **belong**? **Who** is doing the collecting?
- 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** is analyzing the data?
- 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!

COVID-19 and higher education: The case of Northeastern

- Northeastern runs a testing program, which it tries to connect with research initiatives
- In order to collect data for research, they have provided a consent form - let's examine some bits of it, and then discuss!

e-Informed Consent Form

Northeastern University COVID-19 Testing Research Registry & Repository: Consent to Donate Biospecimens to the Repository and Authorization to Contribute Identifiable Private Information to the Registry

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

Some selections from the consent form

- The project for which the consent is given: “Northeastern University COVID-19 Testing - Research Registry & Repository”
- "We do not know in advance what specific research projects will request use of your information or biospecimens. Future research studies may involve testing, diagnosis, prediction, prevention, treatments; genetic and genomic studies, including whole genome sequencing of SARS-CoV2; analyses for public health officials and social, scientific and medical research."
- "Your materials and data may be used by Northeastern faculty and investigators and their research collaborators; Northeastern University research collaborators may include investigators from other universities, research institutions, industry, non-profit foundations and public health agencies."
- Your consent to participate in the registry and repository allows Northeastern to share your data and samples with researchers anywhere, including those in other countries or working for other academic, medical or research institutions, companies, non-profit foundations and public health agencies."
- "Your consent to participate in the registry and repository allows researchers to use your samples and data to study any research question"

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

Northeastern and data - discussion

- How specific are these points on the future usage of data?
- What are the limitations on what they can do with this data, based on these points?
- How clear is it how this data will be used? How clear do you think it should be, and why?
- How clear is it with whom this data will be shared? What could be the purposes of sharing it with companies?
- What other Northeastern's data-related practices would you like to critique, and how?

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

Thank you!

If you have any questions, contact us at:

Milan Skobic
Digital Integration Teaching Initiative
Assistant Director
skobic.m@northeastern.edu

Vaishali Kushwaha
Digital Integration Teaching Initiative
Teaching fellow
v.kushwaha@northeastern.edu

Slides, handouts, and data available at

http://bit.ly/diti-spring2021-musselman

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

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