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

Cara Marta Messina and Jeff Sternberg
Katy Shorey
Spring 2020

Northeastern University
NULab for Texts, Maps, and Networks

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

Discussion: China's DNA Surveillance

- What is DNA surveillance? How is China's government trying to implement it?
- Who is being targeted with this surveillance?
- In what ways might America have similar or different technological infrastructures and forms of surveillance?

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

Workshop Agenda

- Introduce important concepts such as big data, algorithms, and algorithmic bias
- Discuss data, privacy, and data categorization
- Discuss ethical implications of big data and more generally digital research
- SAIL reflection

Slides available at **https://bit.ly/diti-spring2020-shorey**

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 policy
- Explore the ways in which these questions and methods are influencing how humanists and social scientists do research

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

Big Data and Surveillance

Northeastern University
NULab for Texts, Maps, and Networks

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

Big Data

Big data collects vast amounts of data from vast amounts of users and analyzes that data quickly for particular purposes (advertizing, surveillance, search results, 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).

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

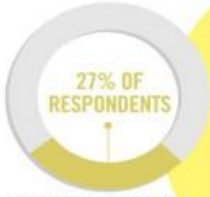
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

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

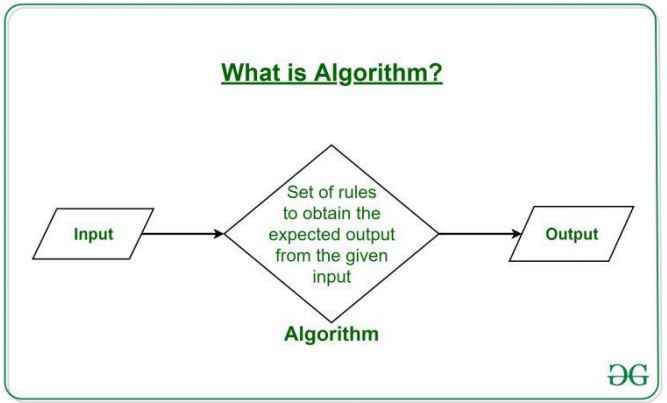
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

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

IBM

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 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!

Algorithmic Bias

Algorithms are *not neutral*. While they do not have minds of their own, people create these algorithms. The processes and data, itself, may reflect particularly biases about society.

For example, Amazon attempted to create an algorithm 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 known to be a male-dominating field, the input data reflected this. The algorithm interpreted 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!

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.

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

Social Media Preferences

Social media sites collect, store, and sell information about you so you get better targeted ads and your newsfeed is tailored to your categories. **What are the targeted ads you see and why do you think you receive these ads?**

Some social media sites that do this:

- Facebook
- Instagram (owned by Facebook)
- Google
- YouTube (owned by Google)
- Twitter

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

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

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

Ethical Implications

- Cambridge Analytica controversy
- Big data also raises questions of autonomy, anonymity, privacy, discrimination, and bias.
- Disparate impact
- 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/

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

Is it all gloom and doom?

Are the ethics around algorithms, surveillance, and big data all negative or controversial?

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!

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!

SAIL: Self-Authored Integrated Learning

SAIL is Northeastern's new platform that helps you holistically track what you have learned in courses, extracurricular events, and coops. SAIL values all types of learning, from interpersonal skills to cultural and/or technical knowledge.

Why SAIL? With SAIL, you can keep track of everything you have learned through your experience at Northeastern and produce portfolios to share with others or remind yourself while building your resume.

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

DITI Partners with SAIL

In order to help you remember and capture what you have learned with DITI, please take a few minutes to log into SAIL and reflect.

- Login to SAIL
- Click the + at the bottom of your timeline and add a “Moment”
- Fill out all the information.
    - Take a few minutes to reflect on what you learned today, what you all did, and how you may use it in the future.
- When you click “Next,” it will ask you to connect to a “Learning Opportunity”
- Connect it to both your course (**PHIL 1102** or **PHIL 1112**) and the shorter opportunity **PHIL 1102 and 1112: Digital Proficiency Module**

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

Thank you!

If you have any questions, contact us at:

Cara Marta Messina
Digital Teaching Integration
Assistant Director
messina.c@husky.neu.edu

Jeff Sternberg
Digital Teaching Integration
Research Fellow
sternberg.je@husky.neu.edu

Slides, handouts, and data available at **http://bit.ly/diti-spring2020-shorey**

DTI office hours: **http://bit.ly/diti-office-hours**

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