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

Vaishali Kushwaha & Yana Mommadova
Digital Integration Teaching Initiative

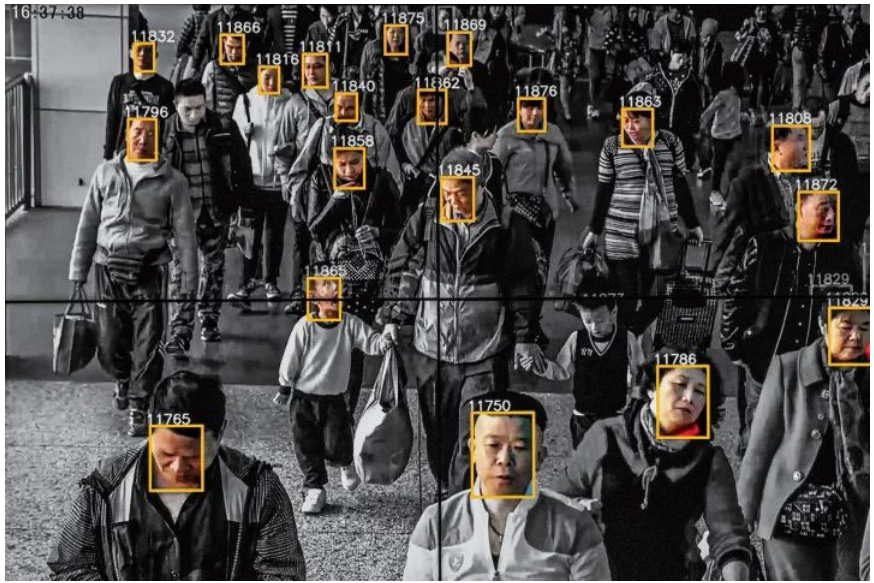
CRIM 3600 Research Methods
Megan Denver
Fall 2021

Northeastern University
NULab for Texts, Maps, and Networks

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

Discussion: China's Social Credit System

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

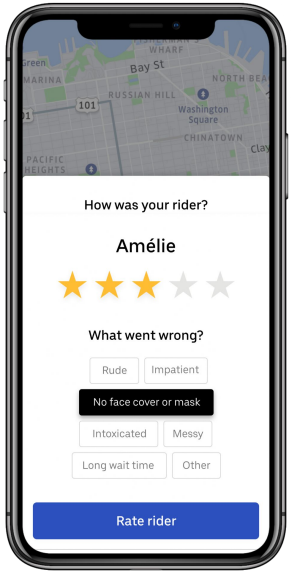


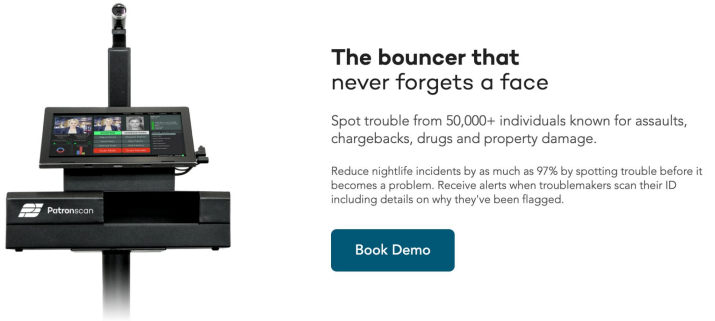
(Photo-Illustration by TIME: Source Photo: Gilles Sabrié—The New York Times/Redux)

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!

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-denver-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 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!

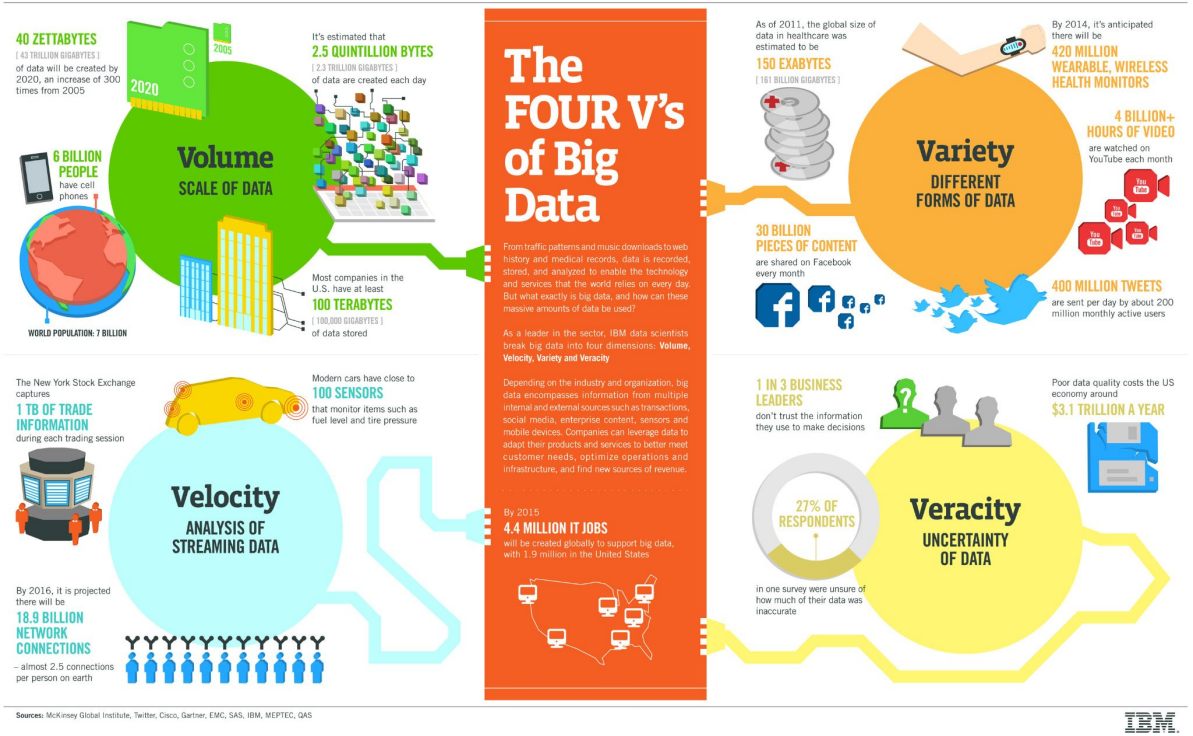
What is “Big Data”?

Companies, governments, and other groups collect vast amounts of data (“big data”) from vast amounts 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!

Big Data: 4 V's


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
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
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
Variety
DIFFERENT FORMS OF DATA
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
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
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
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 **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!

Facebook Preferences

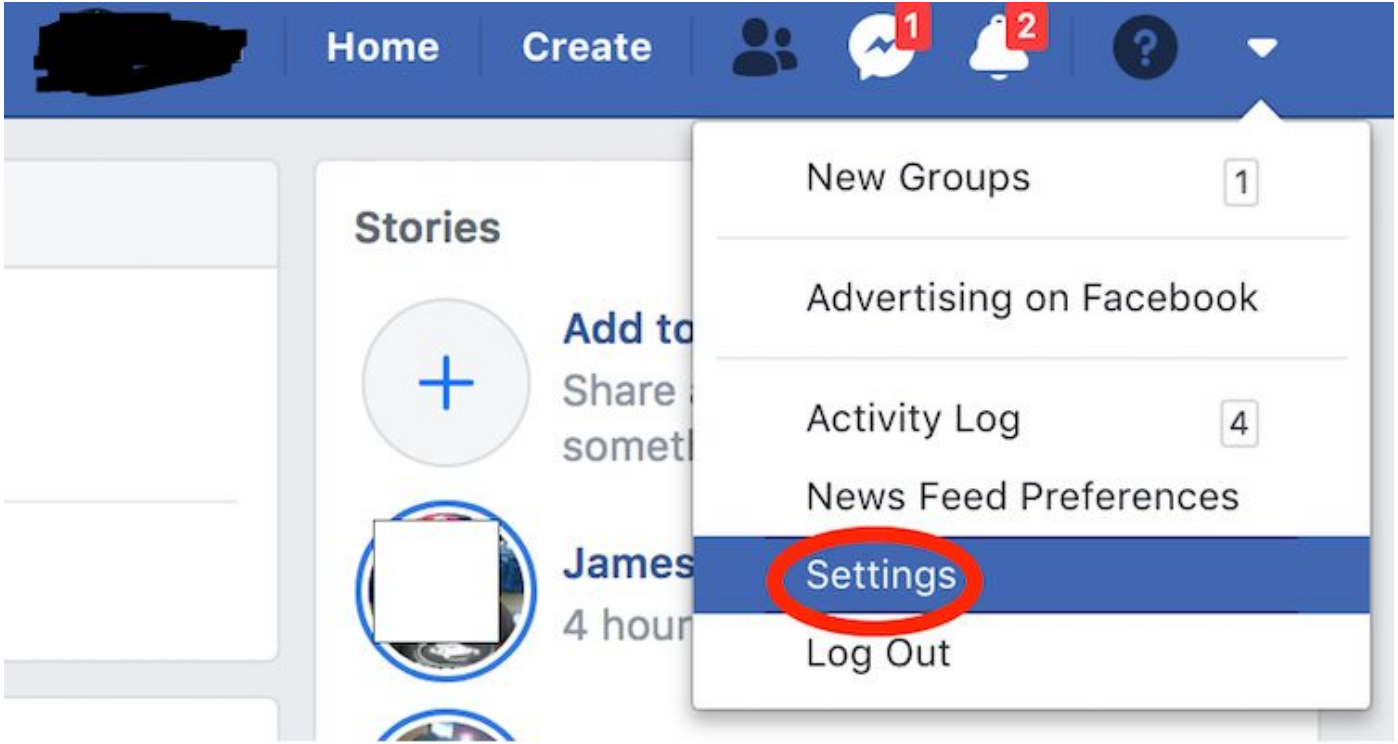
Facebook collects, stores, and sells information about you so you get more targeted ads and your newsfeed is tailored to your categories.

Other social media sites that do this:

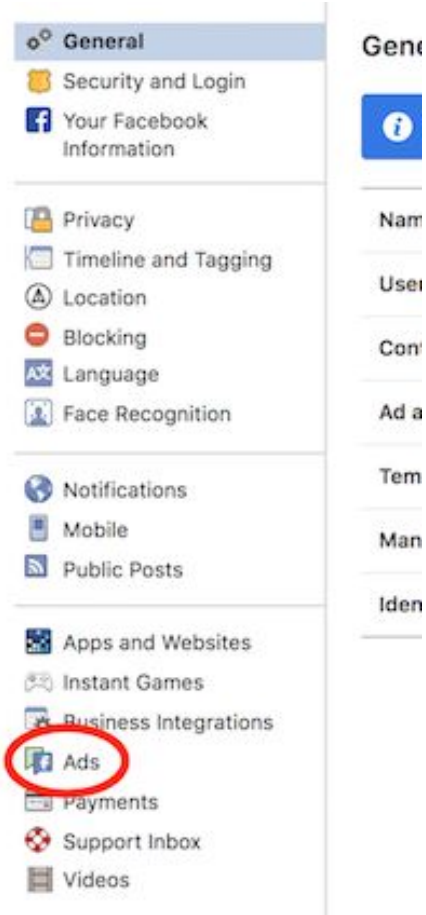
- Instagram (owned by Facebook)
- Google
- TikTok
- YouTube (owned by Google)
- Twitter

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

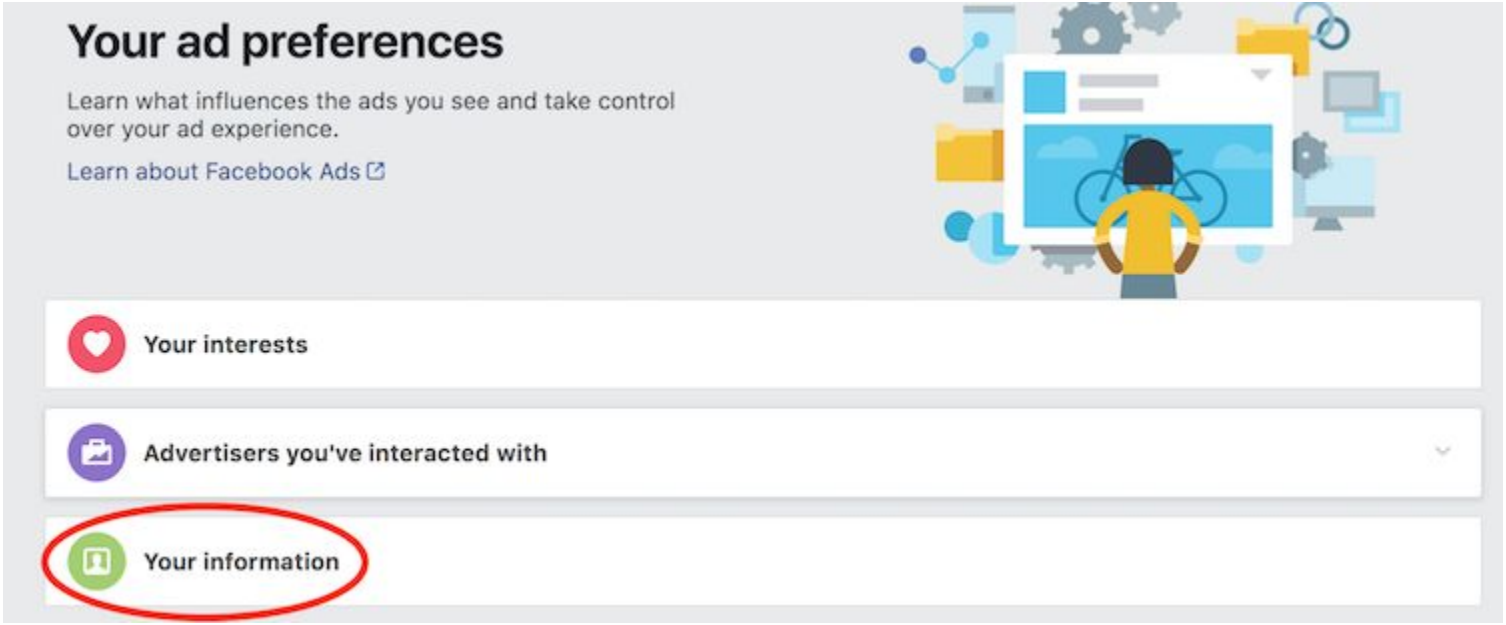
Settings > Ads > Your information > Categories



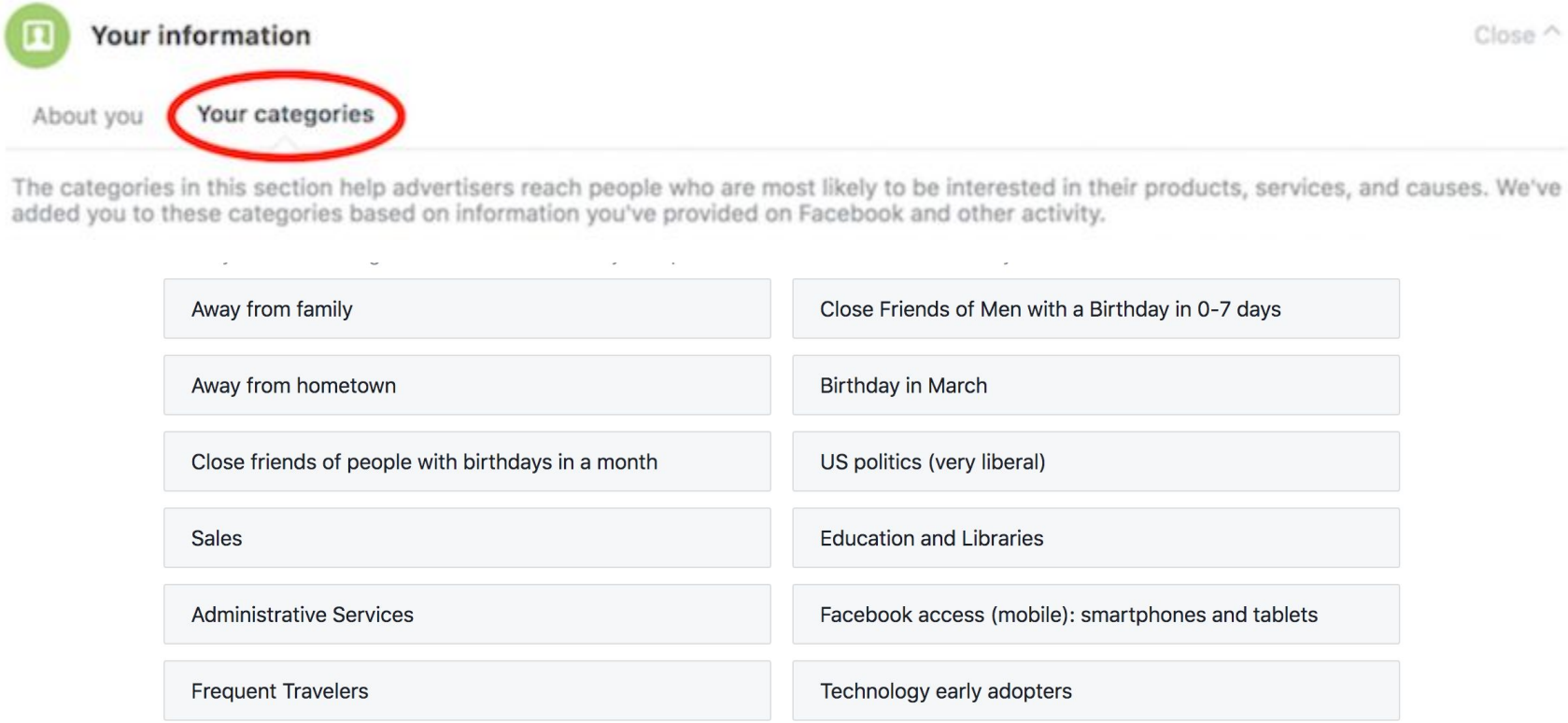
Feel free to ask questions at any point during the presentation!



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



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



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

TikTok

Information we collect automatically

We automatically collect certain information from you when you use the Platform, including internet or other network activity information such as your IP address, geolocation-related data (as described below), unique device identifiers, browsing and search history (including content you have viewed in the Platform), and Cookies (as defined below).

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

TikTok

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!

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!

Downloading Your Data

Facebook: Settings > Your Facebook Information > Download your Information

Google: https://support.google.com/accounts/answer/3024190?hl=en

Instagram app: 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/

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

Explore: Ethical Issues and Biases

Northeastern University
NULab for Texts, Maps, and Networks

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

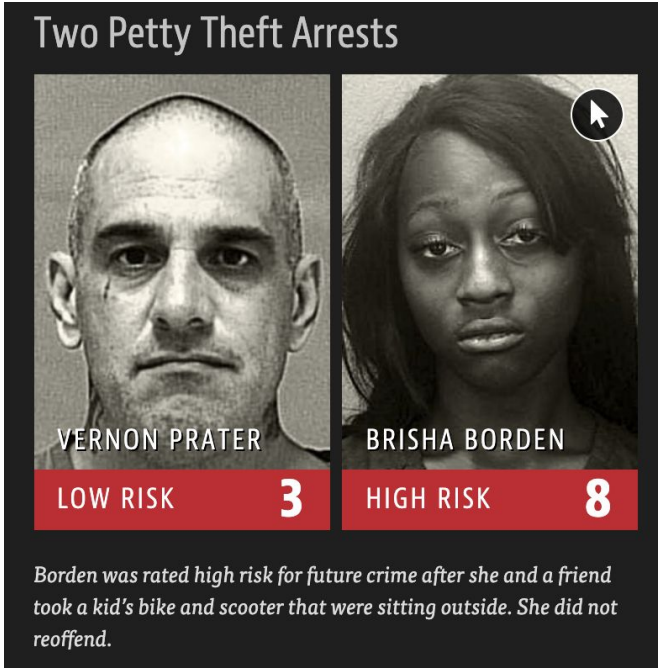
“Big Data” Unbounded — Ethical Issues

- Controversies in recent years:
  - 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 in policing
    - Clearview AI: sells facial recognition “services”
    - Case of Robert Williams: wrongfully arrested
    - Machine Bias: Software used to predict future criminals, biased against Black men

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

Facial recognition in policing and beyond

- Why is the software used for policing 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?



ProPublica

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

Algorithms and Applications

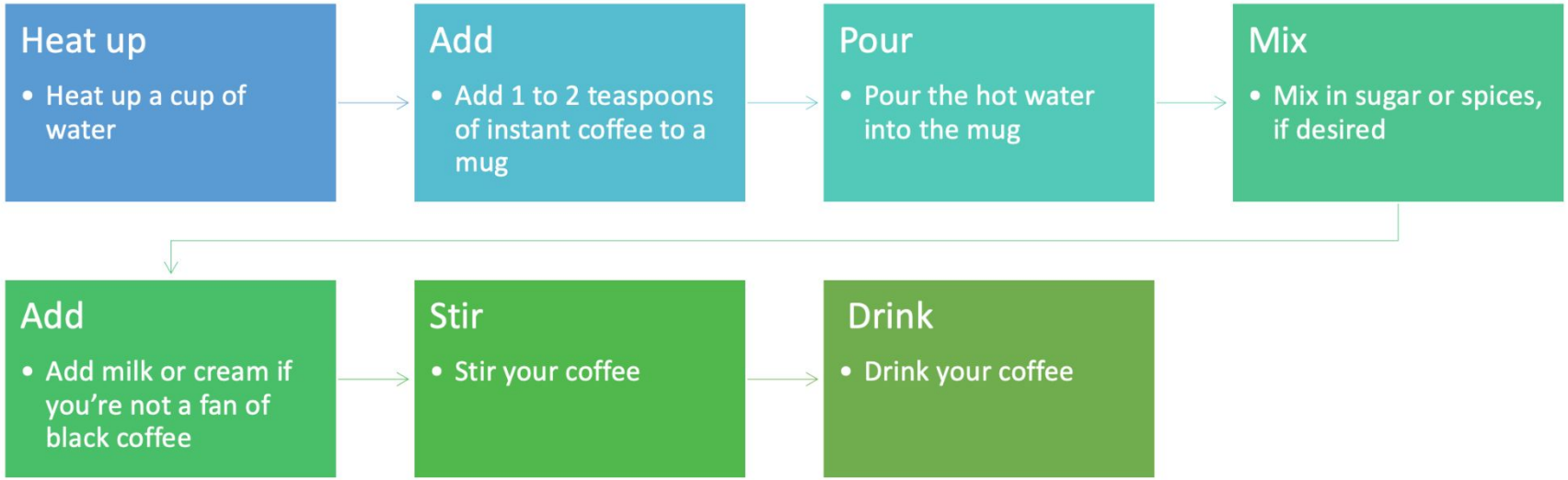
Does the problem lie inherently only in the **algorithm** or also its **application**? What do you think?

- Prof. Lazar and NetSI researchers, at Northeastern, working on COVID-19
- Algorithms predicting the likelihood of cancer (Breast cancer, Prostate cancer)
- Stanford study builds an algorithms/AI that can predict sexual orientation based on a photo with up to 91% accuracy; critics accuse authors of entering an ethically gray area, potential use for anti-LGBTQ purposes, violation of privacy etc.

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

Algorithms

Have you used algorithms in your everyday life? Any examples?



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!

Class Activity: Algorithms and Bias

Northeastern University
NULab for Texts, Maps, and Networks

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

Activity: Dog Adoption

You all work for an adoption agency. You have access to four previous adoption applications and their outcome. You will use those to decide if the new applicant can adopt a dog or not.

You will be assigned into small groups for this exercise. You will discuss within the group and try to come with a unified decision.

Outcome: Team's decision! **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!

Discussion

- 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!

Algorithms and Applicants: Machine Learning

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?

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

Watch this PBS video, if you want to learn about the five common types of algorithmic biases that we should pay attention to and ways to reduce them.

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

So what do ‘big data’ & algorithms have to do with 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!

Discussion

- What are some benefits and what are some risks coming with the increased focus on “big data” in research and policy?
- Do you see any commonalities among the risks and how would you address them?
- Finally, “big data” is a vague term and refers to many different phenomena at once. What are some other terms you would use in order to think about these issues more precisely?

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

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

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

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