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

ENGW 1111 First Year Writing
Emily Avery Miller
Spring 2022

Taught By: Claire Tratnyek and Colleen Nugent

Northeastern University
NULab for Texts, Maps, and Networks

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

Workshop Objectives:

- 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 the use of big data, algorithms, and policy
- Understand the ways in which technology reflects cultural, social, and political biases.
- Explore ways of interpreting and effectively utilizing data-based evidence in written arguments

Slides available at

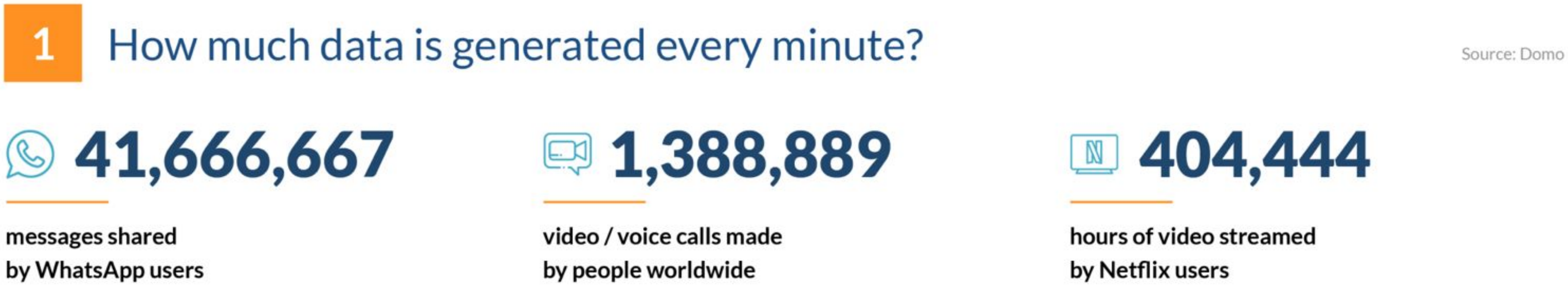
https://bit.ly/diti-spring2022-avery-miller-data-ethics

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*)



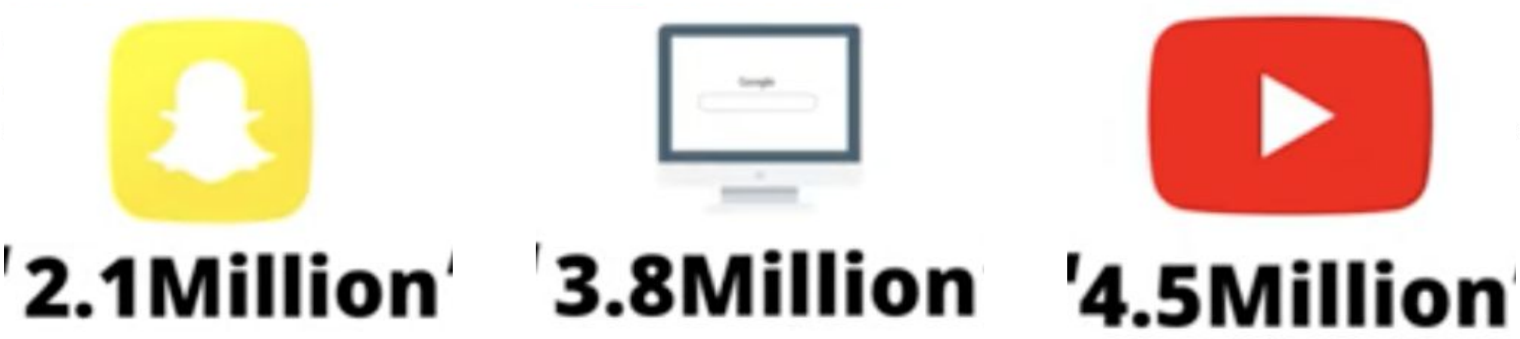
1 How much data is generated every minute?

Source: Domo

41,666,667 messages shared by WhatsApp users

1,388,889 video / voice calls made by people worldwide

404,444 hours of video streamed by Netflix users



Defining ‘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).

We’re living in an era of “surveillance capitalism” **- our information can be considered to be a valuable *product*.**

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

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

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]

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

30 BILLION PIECES OF CONTENT

are shared on Facebook every 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

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

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

IBM

Why should we care about Big Data?

- 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.

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?
- **Where** is data about our lives coming from, and how is it being **collected**?
- **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

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

Who Collects Our Data? How is it Used?

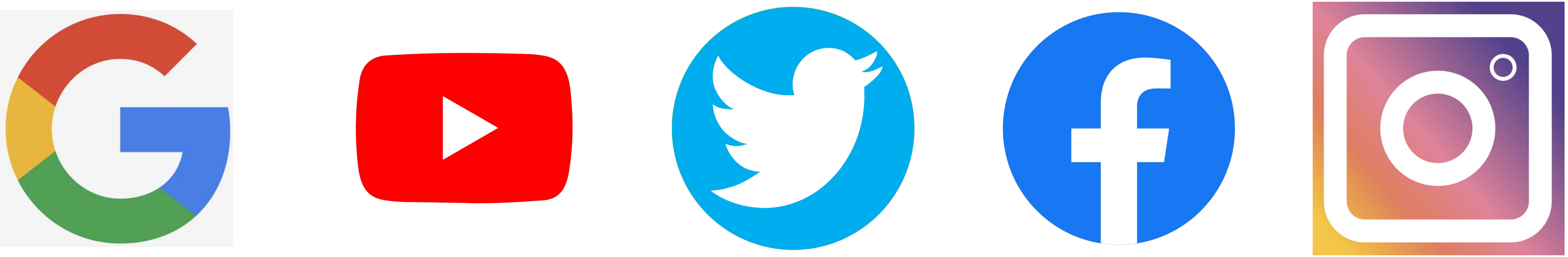
If our digital lives constantly (and silently) produce data, how is that data used, and how can we stay aware of it?

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.

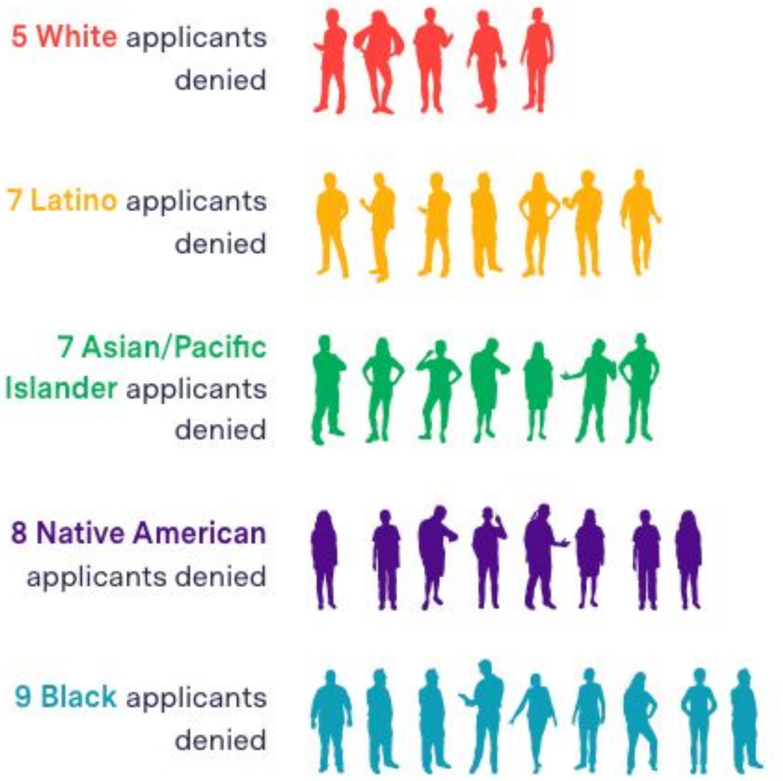
Social media sites collect, store, and sell information about you, so that you get better targeted ads and your newsfeed is tailored to your categories. **Some social media sites that do this:**



Algorithmic Injustice

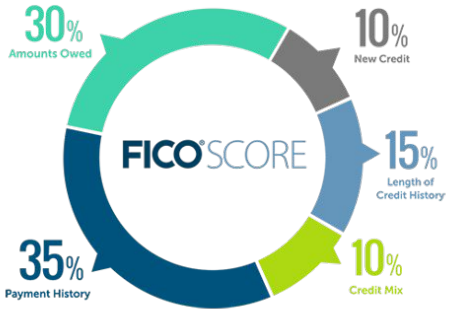
Mortgage approval algorithms can gather and use data in ways that express a racial bias.

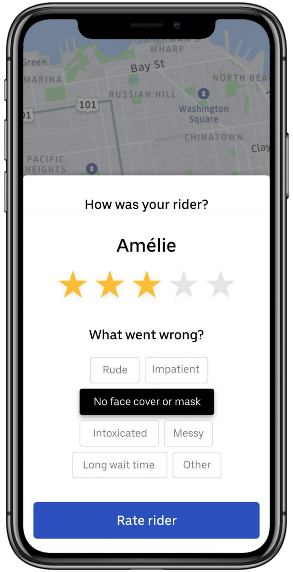
On Fannie & Freddie, who buy about half of all mortgages in America: “This algorithm was developed from data from the 1990s and is more than 15 years old. It’s widely considered detrimental to people of color because it rewards traditional credit, to which White Americans have more access.”



Discussion: America's Social Credit System

What social credit systems do we already see in the United States today? How do they impact our lives?





How was your rider?

Amélie

★★★☆☆

What went wrong?

Rude | Impatient

No face cover or mask

Intoxicated | Messy

Long wait time | Other

Rate rider

The bouncer that
never forgets a face

Spot trouble from 50,000+ individuals known for assaults, chargebacks, drugs and property damage.

Reduce nightlife incidents by as much as 97% by spotting trouble before it becomes a problem. Receive alerts when troublemakers scan their ID including details on why they've been flagged.

Book Demo

Northeastern University
NULab for Texts, Maps, and Networks

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

Social Credit Systems fueled by “big data”

★★★★★ Loved it

★★★★☆ Liked it

★★★☆☆ It was ok

★★☆☆☆ Disliked it

★☆☆☆☆ Hated it

Black Mirror: *Nosedive* (2016)

Northeastern University
NULab for Texts, Maps, and Networks

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

How Are We Being Tracked?

Most websites collect data on their visitors. Some monetize that data in a “data exploitation market,” monetizing their users’ personal information.

Blacklight is a website privacy investigation tool developed by *The Markup*, a nonprofit publication that investigates data misconduct. You can use it to scan and reveal the specific user-tracking technologies on any site.

Use Blacklight now!

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

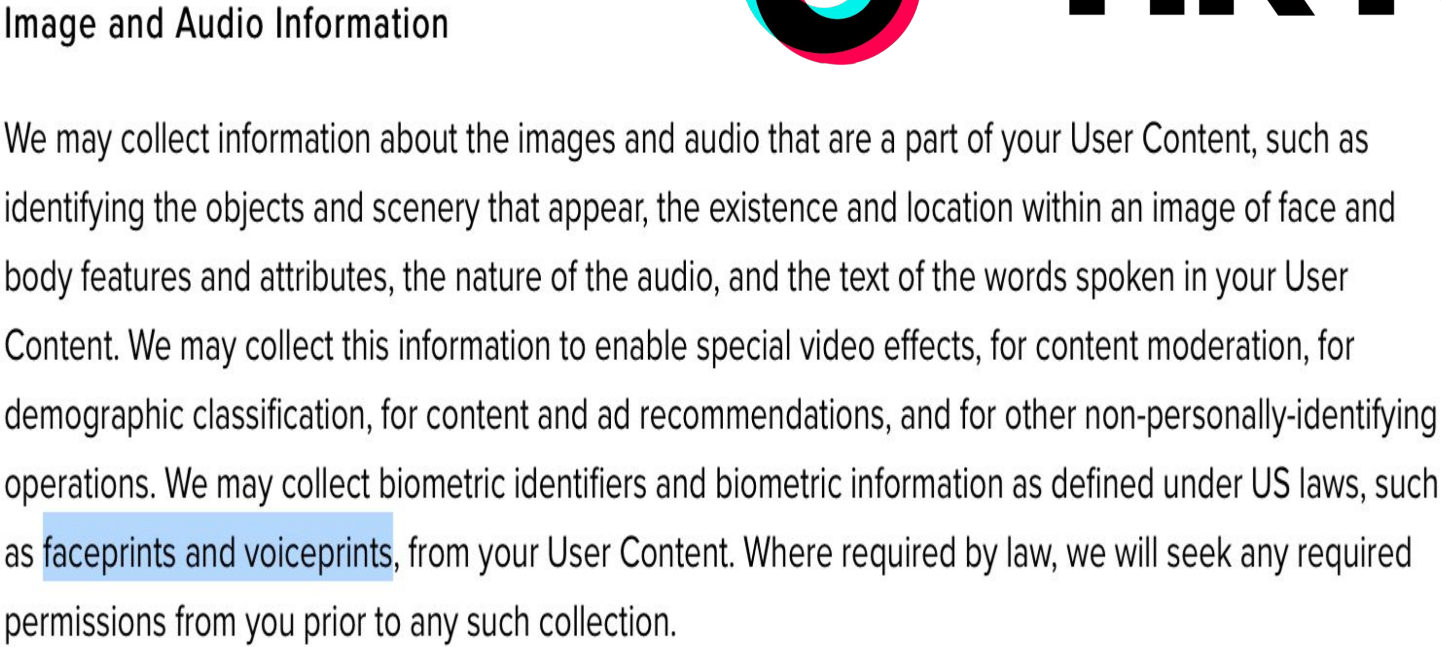
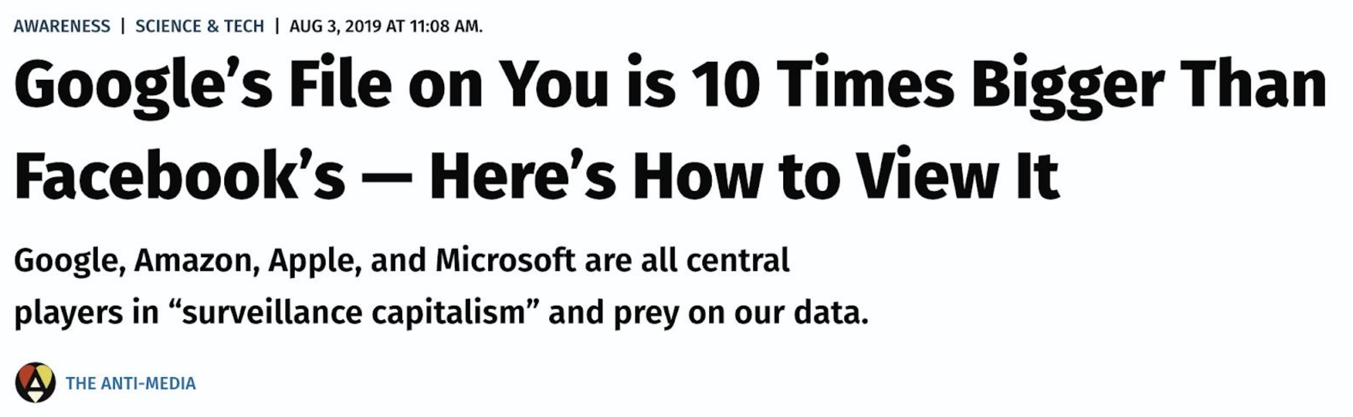


Image and Audio Information

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.

TikTok

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

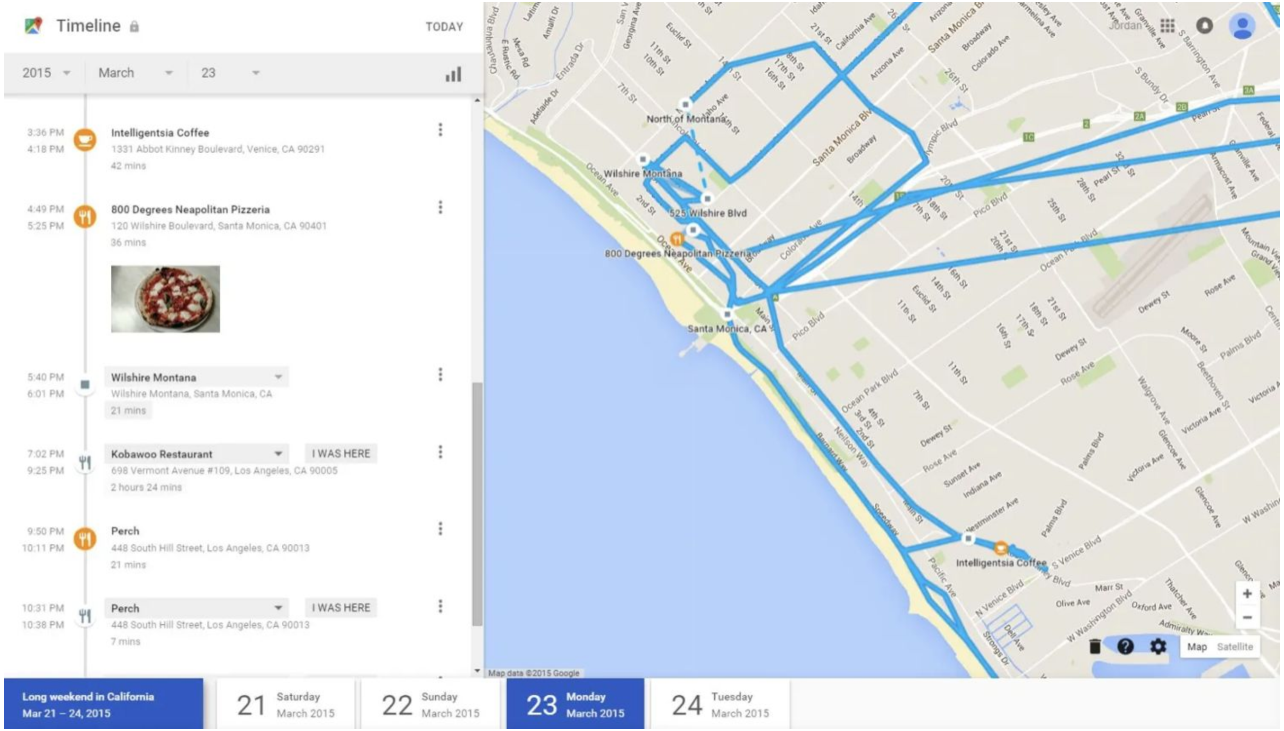
Google, Amazon, Apple, and Microsoft are all central players in "surveillance capitalism" and prey on our data.

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

Example of Google Maps' Timeline:



Timeline — TODAY

2015 | March | 23

3:36 PM – 4:18 PM — **Intelligentsia Coffee**
1331 Abbot Kinney Boulevard, Venice, CA 90291
42 mins

4:49 PM – 5:25 PM — **800 Degrees Neapolitan Pizzeria**
120 Wilshire Boulevard, Santa Monica, CA 90401
36 mins

5:40 PM – 6:01 PM — **Wilshire Montana**
Wilshire Montana, Santa Monica, CA
21 mins

7:02 PM – 9:25 PM — **Kobawoo Restaurant** — I WAS HERE
698 Vermont Avenue #109, Los Angeles, CA 90005
2 hours 24 mins

9:50 PM – 10:11 PM — **Perch**
448 South Hill Street, Los Angeles, CA 90013
21 mins

10:31 PM – 10:38 PM — **Perch** — I WAS HERE
448 South Hill Street, Los Angeles, CA 90013
7 mins

Long weekend in California
Mar 21 – 24, 2015

21 Saturday March 2015 | 22 Sunday March 2015 | 23 Monday March 2015 | 24 Tuesday March 2015

Check out an early (2015) Venturebeat article about "freaky" Google Map 'Your Timeline' feature here

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/

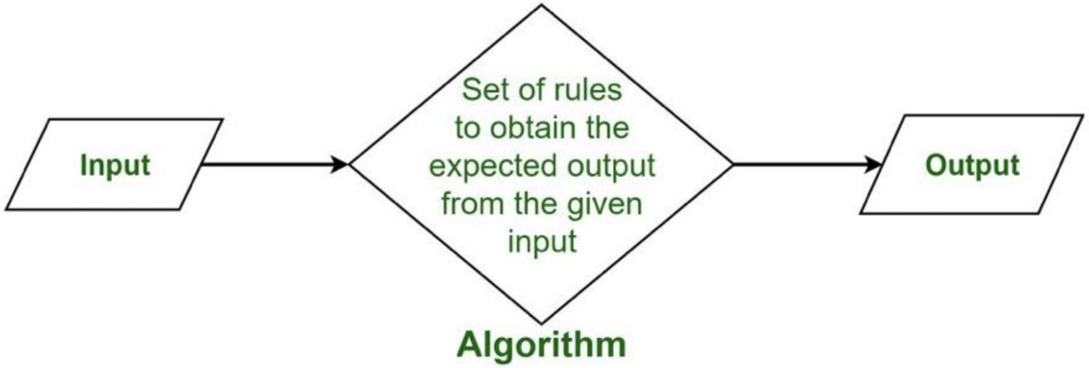
Issues in Big Data: Ethics and Algorithmic Bias

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.



Input

Set of rules to obtain the expected output from the given input

Output

Algorithm

- "**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.

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).

Check out this Vox article for more information on algorithmic bias

“Big Data” Unbounded — Ethical Issues

Some (relatively) recent controversies:

- **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**
  - Clearview AI: sells facial recognition “services”
  - Case of Robert Williams: wrongfully arrested
  - Machine Bias: Software used to predict future criminals, biased against Black men
  - Stanford study creates AI that can predict sexual orientation based on a photo with up to 91% accuracy

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

“Big Data” can also inform solutions to complex problems:

- Prof. Lazar and NetSI researchers at Northeastern have been working on COVID-19 research using big data
- Scientists have also created algorithms that can predict the likelihood of cancer (Breast cancer, Prostate cancer)
- An example from the social sciences: Allegheny County PA “family screening tool” to support human screeners in the Department of Children, Youth, and Families

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

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

- When we look at the data used to train an algorithm, we must ask **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 various populations, especially marginalized groups?
- Will the technology and big data-driven solution **eliminate** human bias or **amplify** it?

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

Search Engine Bias: Example and Discussion

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

“Greatest Authors of All Time”

Open Google’s search engine and type in “Greatest authors of all time.”

- What are some of the results? What do you notice about these results?
- Where do you think these results came from?
- How many authors on this list have you read? Do you agree with the list?
- What do these results suggest to you in terms of defining “greatest” and “authors”?

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

“Greatest ______ Authors of All Time”

Now try these results:

- Greatest women authors
- Greatest Black women authors
- Greatest Black authors
- Greatest white authors

“Black” leads to substantial results, while “white” does not.

Why do you think this might be?

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

Technology is Not Neutral

Information systems like Google as well as data collection, data analysis, and algorithms 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.

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

Biases in Scholarship and Archival Silences

Northeastern University
NULab for Texts, Maps, and Networks

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

Bias in Scholarship

Questions to consider:

- **Whose voices** and expertise **are valued** and heard?
- **What kinds of data are prioritized** in scholarship, and how/how often are they used?
- **Whose voices** and experiences and bodies can we easily find in the historical record, and whose **are missing**?
- **What other sources** of information might help **fill in gaps** in the 'official' records found in archives and academic discourse?

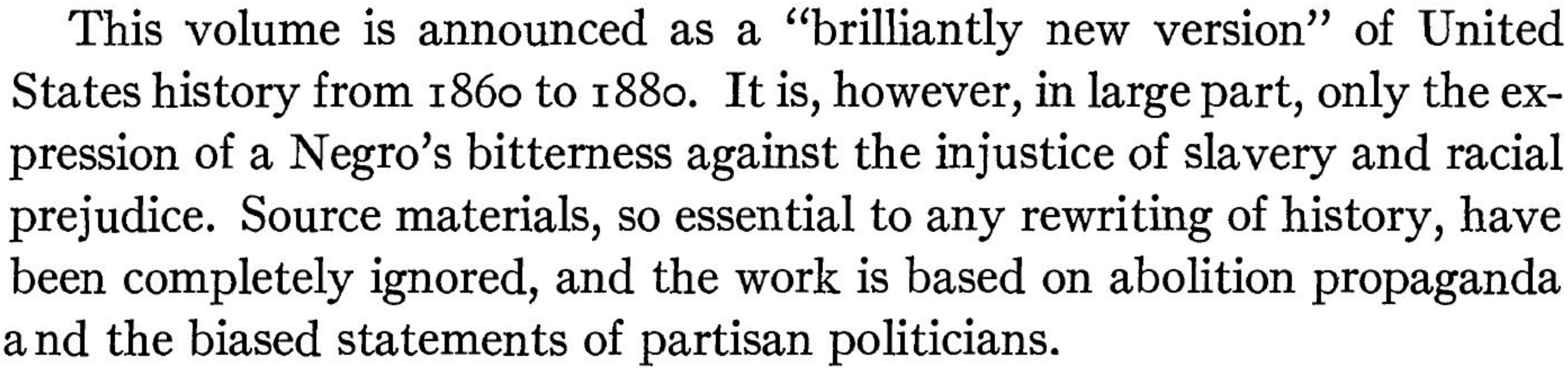
Feel free to ask questions at any point during the presentation!

Bias in Scholarship

- **W. E. B. DuBois**, b. 1868 d. 1963 (NAACP founder, scholar, sociologist, writer, activist)
- Published "***Black Reconstruction in America**: An Essay Toward a History of the Part Which Black Folk Played in the Attempt to Reconstruct Democracy in America, 1860–1880*" in 1935
- **Emphasized the role and agency of African Americans during the Civil War** and Reconstruction and framed it as a period that held promise for a worker-ruled democracy to replace a slavery-based plantation economy.

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

A review of DuBois' scholarship by a prominent academic at the time:



This volume is announced as a "brilliantly new version" of United States history from 1860 to 1880. It is, however, in large part, only the expression of a Negro's bitterness against the injustice of slavery and racial prejudice. Source materials, so essential to any rewriting of history, have been completely ignored, and the work is based on abolition propaganda and the biased statements of partisan politicians.

Archives and the “Historical Record”

- **Archives**
  - What comprises the historical record?
  - What information gets saved, and what doesn’t?
  - Who makes the decisions about what *can and cannot be included* in “official” records?

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

Archives and Archival Silences

- **Archival silences**
  - Whose **voices**, **bodies**, and **experiences** are missing from the historical record?
  - How can we **mitigate** archival silences in our work?
  - How can we think of our work as a **response** to or a **disruption** of these silences?

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

Moving Forward - How can we be cognizant of ‘big data,’ algorithms, and silences in our 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**?

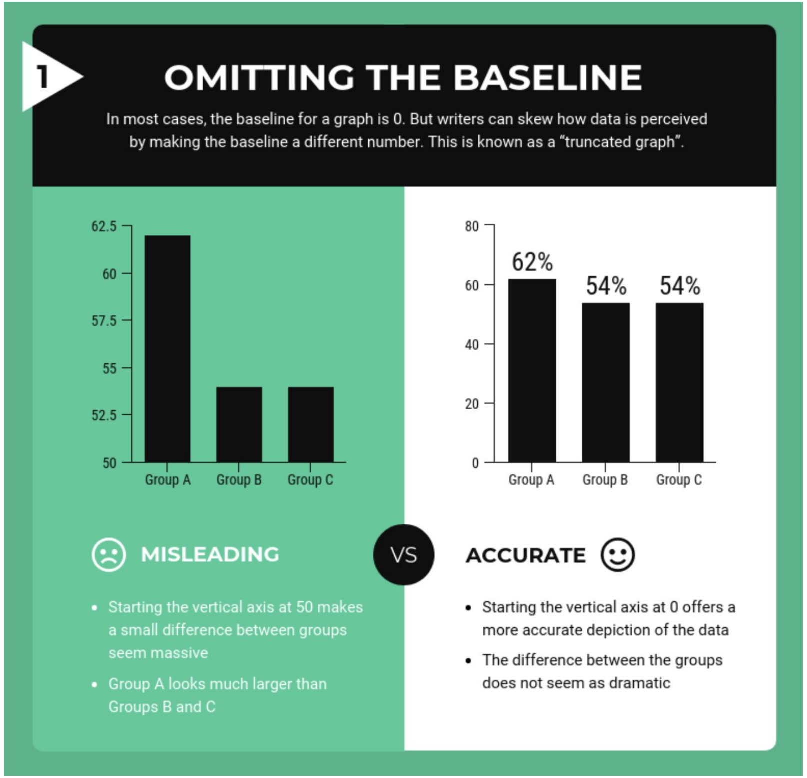
Feel free to ask questions at any point during the presentation!

Responsibly Using Big (or *any* kind of) Data

- Be **Data-Literate** - turn a critical eye to studies that use big data, evaluate the sources of that data, and carefully examine the conclusions authors draw from their sources
- Be **thoughtful** and **intentional** as you incorporate big data or conclusions drawn from big data sources in your work - think:
  - Could this evidence be interpreted in a different way?
  - Is this the strongest evidence I could use to support my claim?
  - Is the way I'm presenting this information accurate, or could it be considered in any way ***misleading***?

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

Be Mindful of Infographics and Data Visualizations



1 OMITTING THE BASELINE

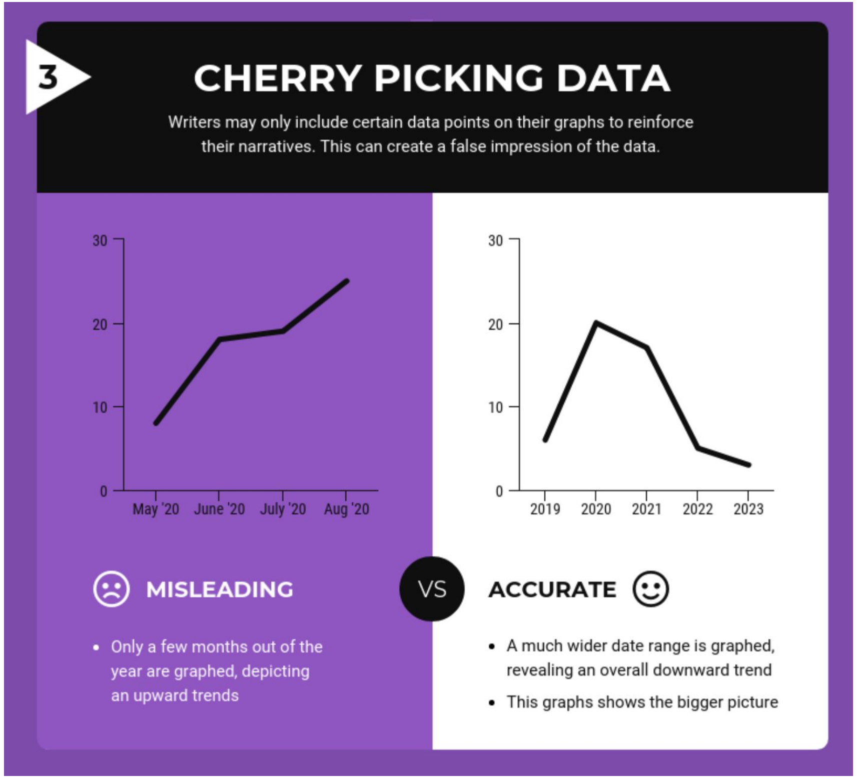
In most cases, the baseline for a graph is 0. But writers can skew how data is perceived by making the baseline a different number. This is known as a "truncated graph".

MISLEADING

- Starting the vertical axis at 50 makes a small difference between groups seem massive
- Group A looks much larger than Groups B and C

ACCURATE

- Starting the vertical axis at 0 offers a more accurate depiction of the data
- The difference between the groups does not seem as dramatic



3 CHERRY PICKING DATA

Writers may only include certain data points on their graphs to reinforce their narratives. This can create a false impression of the data.

MISLEADING

- Only a few months out of the year are graphed, depicting an upward trends

ACCURATE

- A much wider date range is graphed, revealing an overall downward trend
- This graphs shows the bigger picture

Northeastern University
NULab for Texts, Maps, and Networks

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

Finding and Using Non-Traditional Sources

Some kinds of non-traditional and/or non-academic sources:

- Public Media (written/broadcast journalism)
- Crowdsourced projects (including wikipedia, aggregate reviews, etc.)
- Multimedia sources (including social media and blog posts)
  - Using Twitter for academic research
  - Prof. Eunsong Kim's *The Politics of Trending*
- Oral histories and interviews
- Indigenous forms of knowledge

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

Vetting and Citing Non-Traditional Sources

Regardless of the type of source you're using, but *especially* if it isn't coming from an academic publication, you should always...

1) Try to **verify the information** presented in the source by finding other (independent) sources that support it
2) Be clear in your writing about what kind of source it is, where you found it, and how you're using it (be explicit about your **process** and the source's **purpose**)
3) **Cite your source** appropriately so that any reader can find it

Citing non-traditional sources correctly: Purdue Online Writing Lab (OWL)

Thank you!

If you have any questions, contact us at: nulab.info@gmail.com

Developed by DITI Research Fellows Claire Tratnyek, Vaishali Kushwaha, Yana Mommadova, Colleen Nugent, Tieanna Graphenreed, Javier Rosario

Slides & handout available at:

https://bit.ly/diti-spring2022-avery-miller-data-ethics

Sign up for office hours at: http://calendly.com/diti-nu/

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