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

ENG 7360 Topics in Rhetoric

Prof. Mya Poe Fall 2022

Taught By:

Northeastern University

NULab for Texts, Maps, and Networks

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

Workshop Goals

- Recognize how data are being used in society as well as how algorithms impact and shape our daily lives.
- Explore how notions of privacy, security, and agency are being reshaped and redefined through big data, algorithms, and policy.
- Understand the ways in which technology reflects cultural, social, and political biases.
- Engage with critical rethinking of everyday **practices related to data collection and use.**
- Explore ways of interpreting and effectively **utilizing data-based evidence in written arguments.**
- Explore how these questions and methods are influencing how **[social scientists/policy-makers/etc.] do research and practice their craft.**

Slides, handouts, and data available at https://bit.ly/3Q991EG

Northeastern University
NULab for Texts, Maps, and Networks

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*)

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

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 consumer/political 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 a valuable *product*.**

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

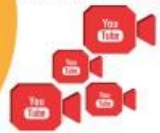
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

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

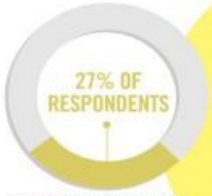
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

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



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

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.
- Big data also **raises ethics questions** about access, power, autonomy, anonymity, privacy, discrimination, and bias.

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?
- **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?

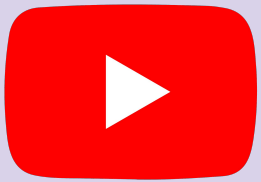
If our digital lives constantly (and silently) produce data, how is that data used, and how can we stay aware of it?

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:**

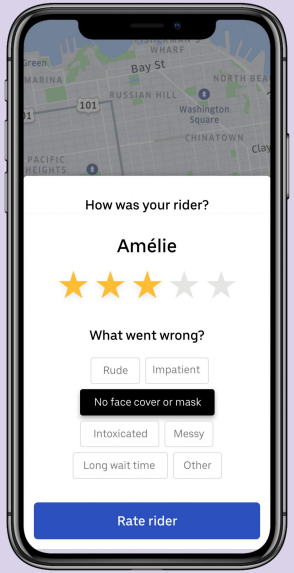



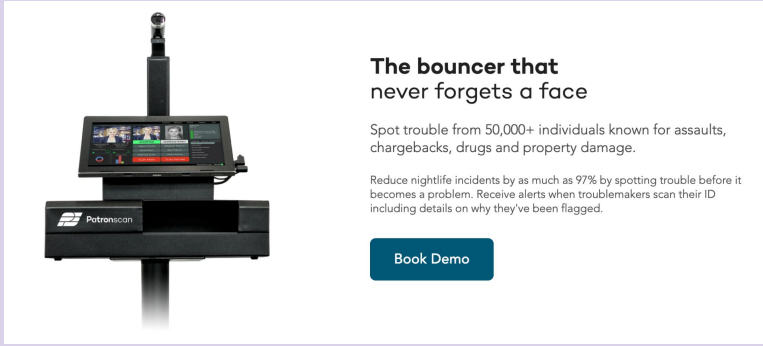

Feel free to ask questions at any point during the presentation!

Discussion: America's Social Credit Systems

What social credit systems exist in the United States today?

How do they impact our lives?





Northeastern University
NULab for Texts, Maps, and Networks

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

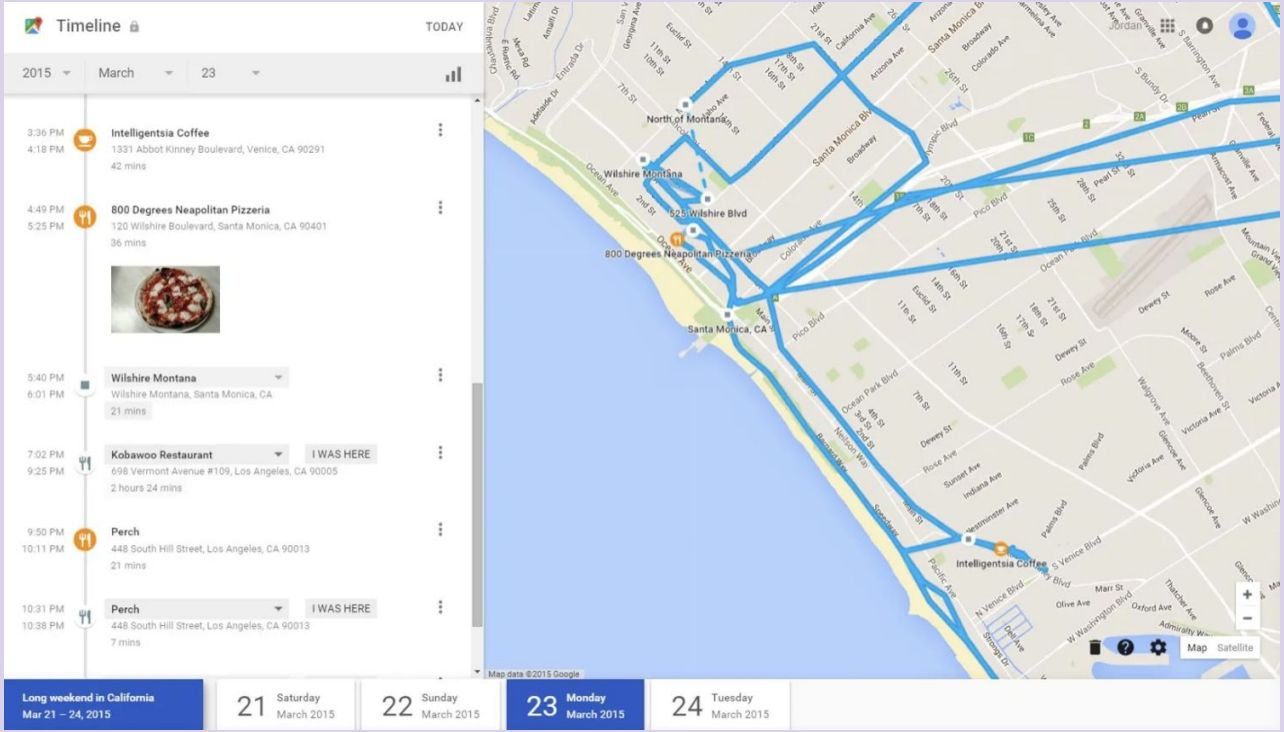
Google, Amazon, Apple, and Microsoft are all central players in "surveillance capitalism" and prey on our data.

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:



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

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.

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!

Northeastern University
NULab for Texts, Maps, and Networks

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

Online Presence & Data Privacy Discussion

Northeastern University
NULab for Texts, Maps, and Networks

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

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/

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

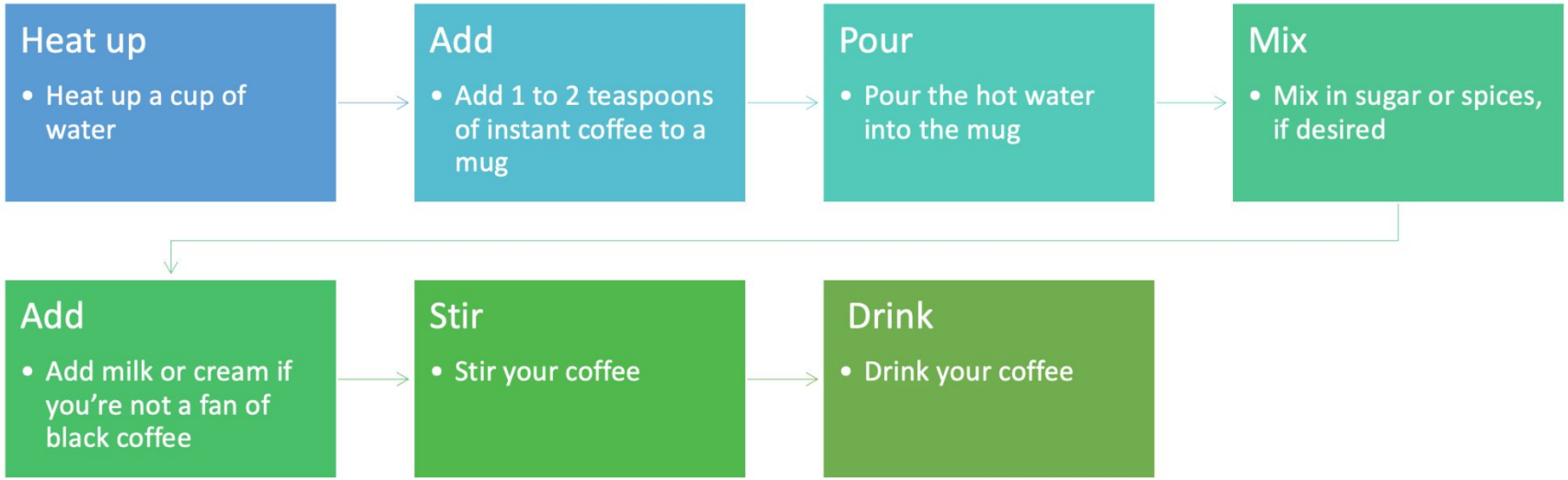
Algorithms and Bias

Northeastern University
NULab for Texts, Maps, and Networks

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

What is an **algorithm**?

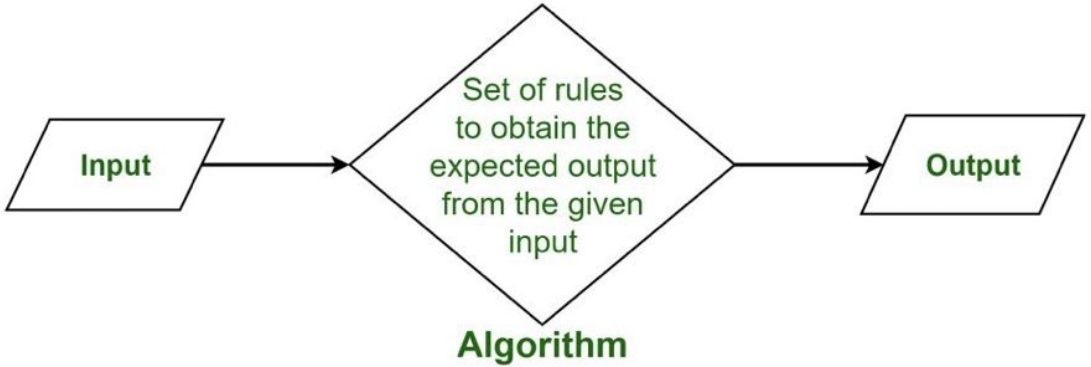
Have you used algorithms in your everyday life? Any examples?



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

Defining Algorithms

- An algorithm is a set of instructions, usually for computers to interpret and follow.



- "**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!

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).

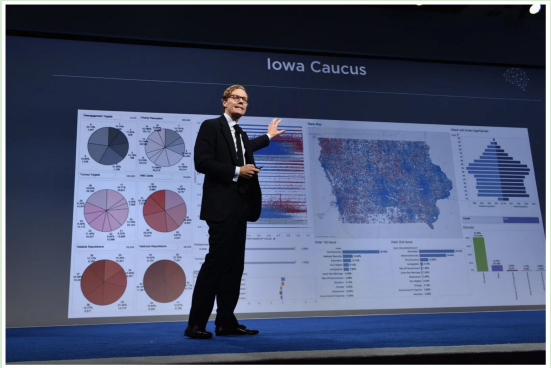
Check out this Vox article for more information on algorithmic bias!

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

“Big Data” Unbounded — Ethical Issues

Some (relatively) recent controversies:

Cambridge Analytica controversy: psychological profiles of American voters



Alexander Nix, the currently suspended CEO of Cambridge Analytica, speaks at a 2016 event in New York City. Nix and his firm are accused of misusing the personal data of 50 million people as part of their political consulting work, which included President Trump's 2016 campaign.

Bryan Bedder/Getty Images for Concordia Summit

Racial bias in health algorithms: results in reduced access to care for Black people

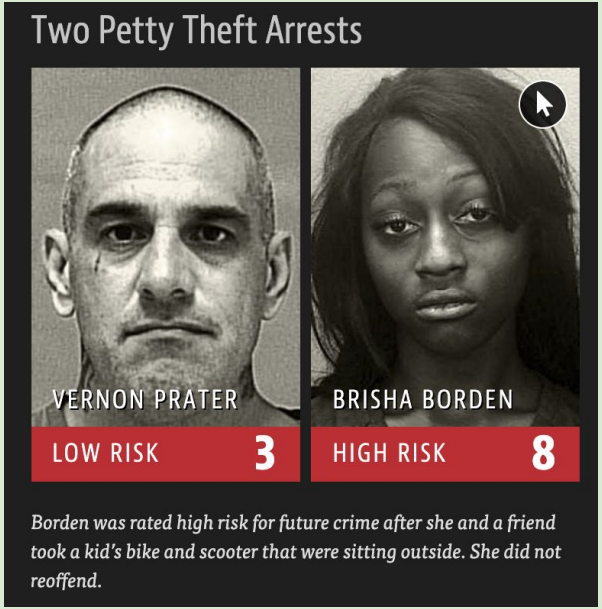
Black people with complex medical needs were less likely than equally ill white people to be referred to programmes that provide more personalized care. Credit: Ed Kashi/VII/Redux/eyevine

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

“Big Data” Unbounded — Ethical Issues

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
- 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

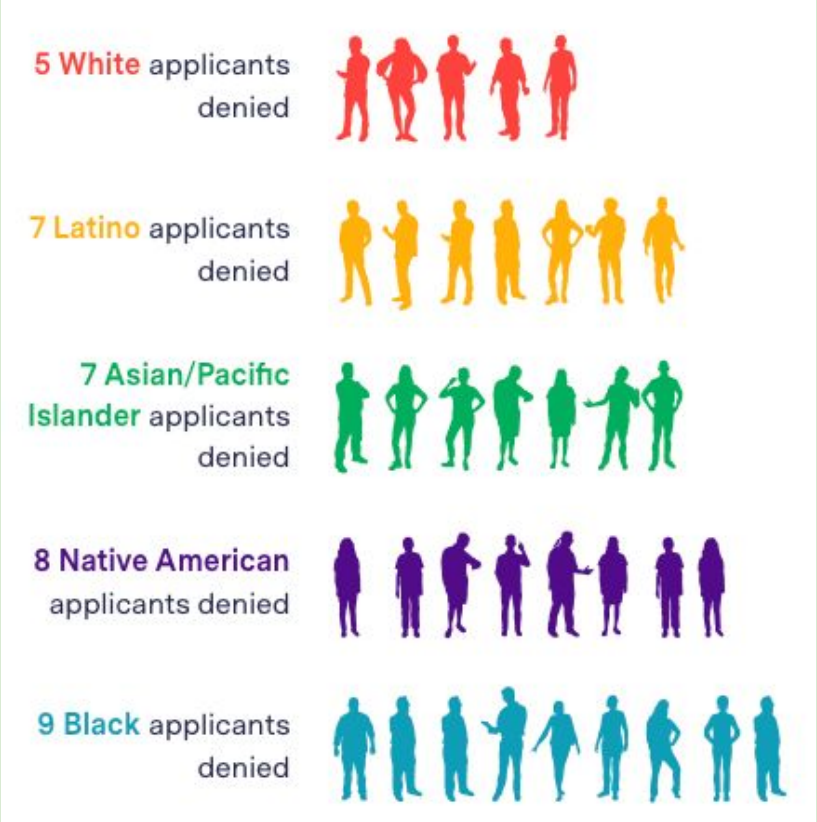


Feel free to ask questions at any point during the presentation!

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.”



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!

“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..."

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?

SOURCE: **"What Gets Counted Counts" Principle #4 of Data Feminism (mitpress, 2020)**

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

Questions to consider/Discussion:

- What are some **benefits** and what are some **risks** coming with the increased focus on 'big data' in research and policy?
- Are technology- and big data-driven solutions more likely to **eliminate** human bias or **amplify** it?
- Do problems lie inherently only in the **algorithm** or also its **application**?
- In any case study, where can we find **data-driven** analyses, possible solutions, or policy arguments?
    - How can we critically analyze these to determine whether the **data is being used ethically**?

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!

Data Justice

Northeastern University
NULab for Texts, Maps, and Networks

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

Definitions of Data Justice

Data justice is an approach that redresses ways of collecting and disseminating data that have invisibilized and harmed historically marginalized communities. For decades, if not centuries, data has been weaponized against BIPOC communities, in particular, to reinforce oppressive systems that result in divestment and often inappropriate and harmful policies.

From the Coalition of Communities of Color explanation of "Research Justice"

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

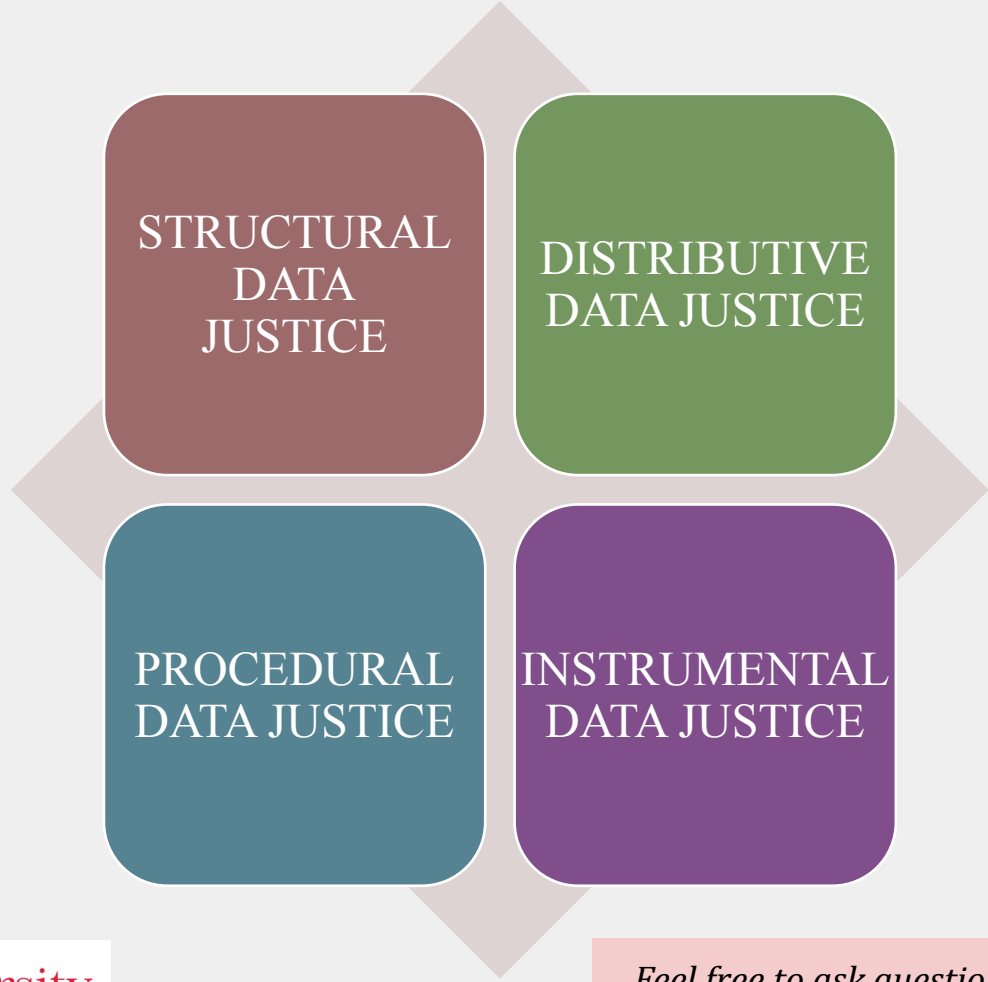
Definitions of Data Justice

Data justice aims to capture forms of knowledge and lived experiences that are community-centered and community-driven to counter the systemic erasure and harm perpetrated on BIPOC communities via oppressive data practices. **The fundamental premises of data justice are that data should: (1)** make visible community-driven needs, challenges, and strengths, **(2)** be representative of community; and **(3)** treat data in ways that promote community self-determination.

From the Coalition of Communities of Color explanation of "Research Justice"

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

Examining Data Justice Processes

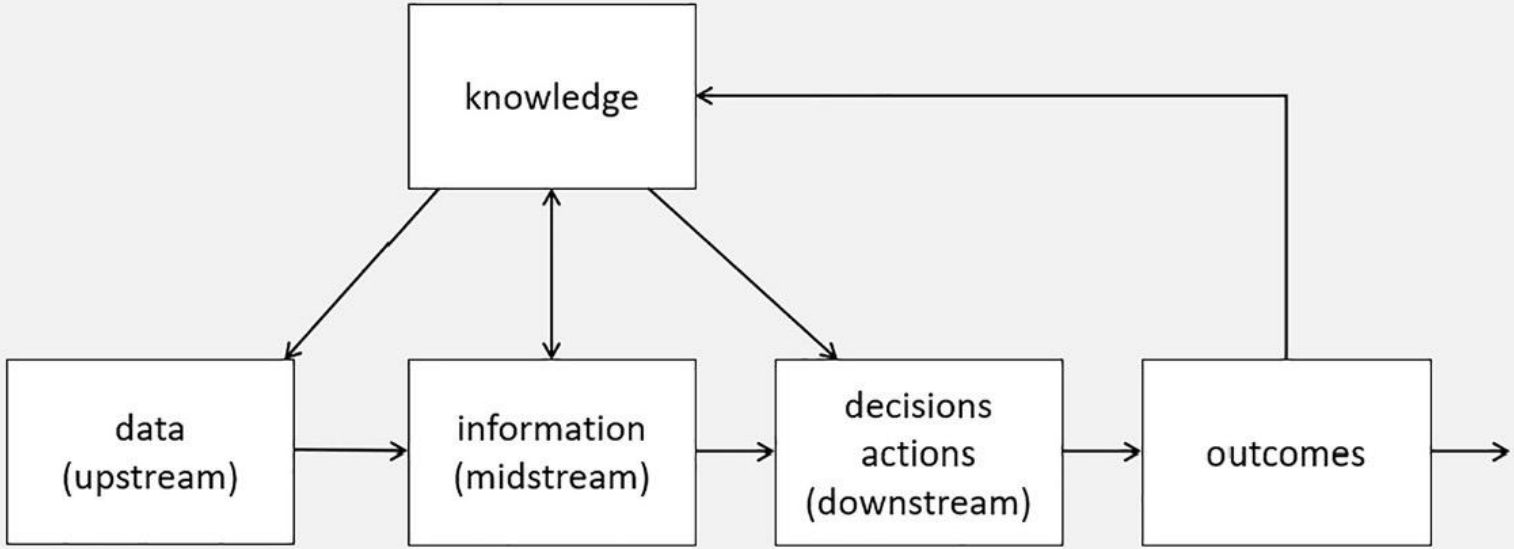


Northeastern University
NULab for Texts, Maps, and Networks

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

Structural data justice/Rights based data justice

STRUCTURE: Impact structure on data (e.g. impact 'digital and virtual divides' on data)
STRUCTURE: Impact data on structure (e.g. impact 'who and what is visible – and to whom?' on municipality)



Distributive data justice

STRUCTURE +
ACCESS +
MOBILIZATION:
who gets what?

Procedural data justice

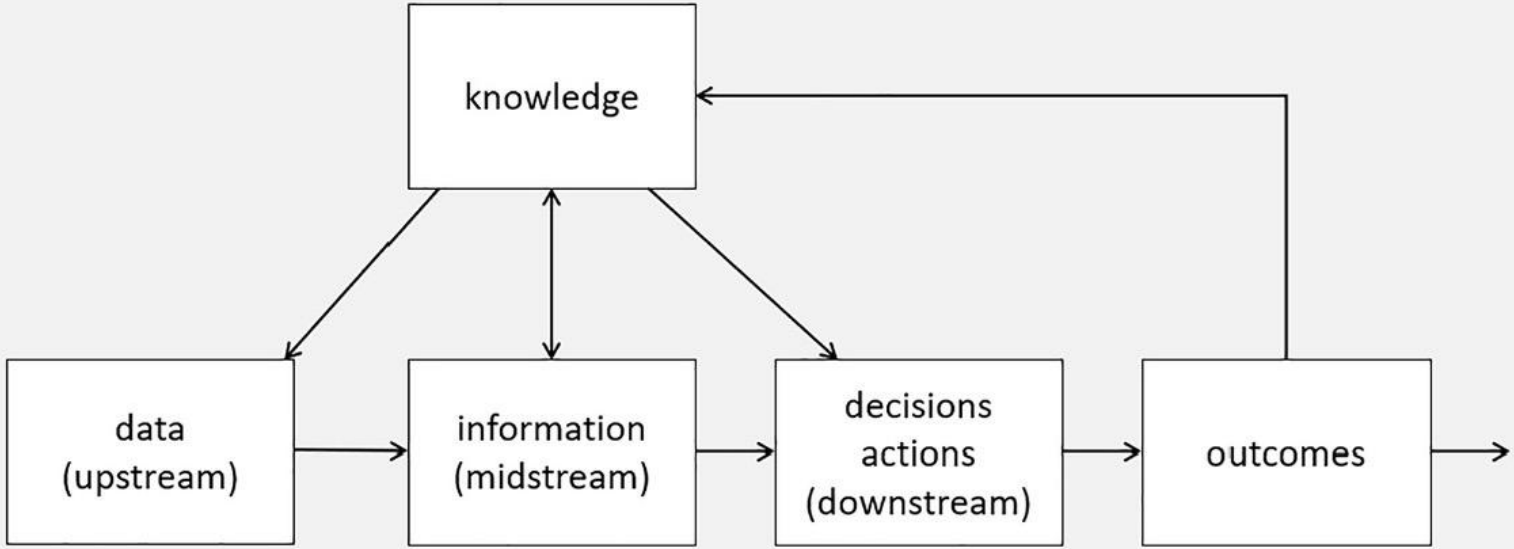
ACCESS: who (is able to) contribute?
ACCESS: who gains? (e.g. skills and connections)

Instrumental data justice

MOBILIZATION: who (is able to) use the data?
MOBILIZATION: how is the data used?

Structural data justice/Rights based data justice

STRUCTURE: Impact structure on data (e.g. impact 'digital and virtual divides' on data)
STRUCTURE: Impact data on structure (e.g. impact 'who and what is visible – and to whom?' on municipality)



Distributive data justice

STRUCTURE + ACCESS + MOBILIZATION: who gets what?

Procedural data justice

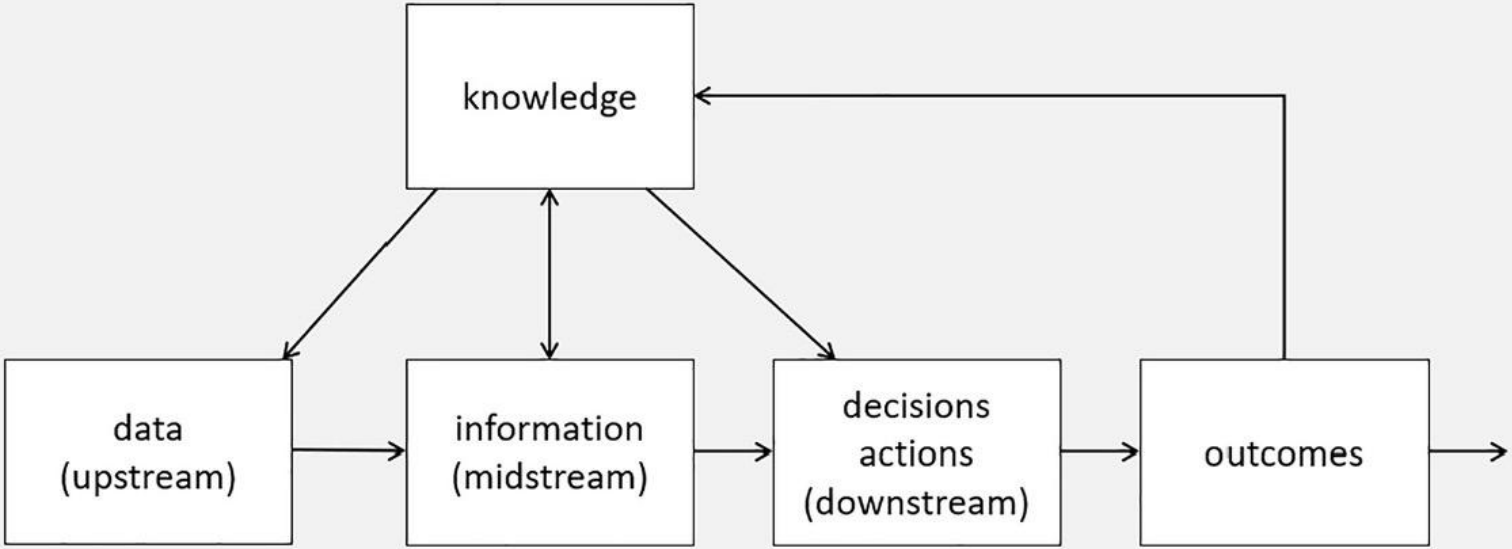
ACCESS: who (is able to) contribute?
ACCESS: who gains? (e.g. skills and connections)

Instrumental data justice

MOBILIZATION: who (is able to) use the data?
MOBILIZATION: how is the data used?

Structural data justice/Rights based data justice

STRUCTURE: Impact structure on data (e.g. impact 'digital and virtual divides' on data)
STRUCTURE: Impact data on structure (e.g. impact 'who and what is visible – and to whom?' on municipality)



Distributive data justice

STRUCTURE +
ACCESS +
MOBILIZATION:
who gets what?

Procedural data justice

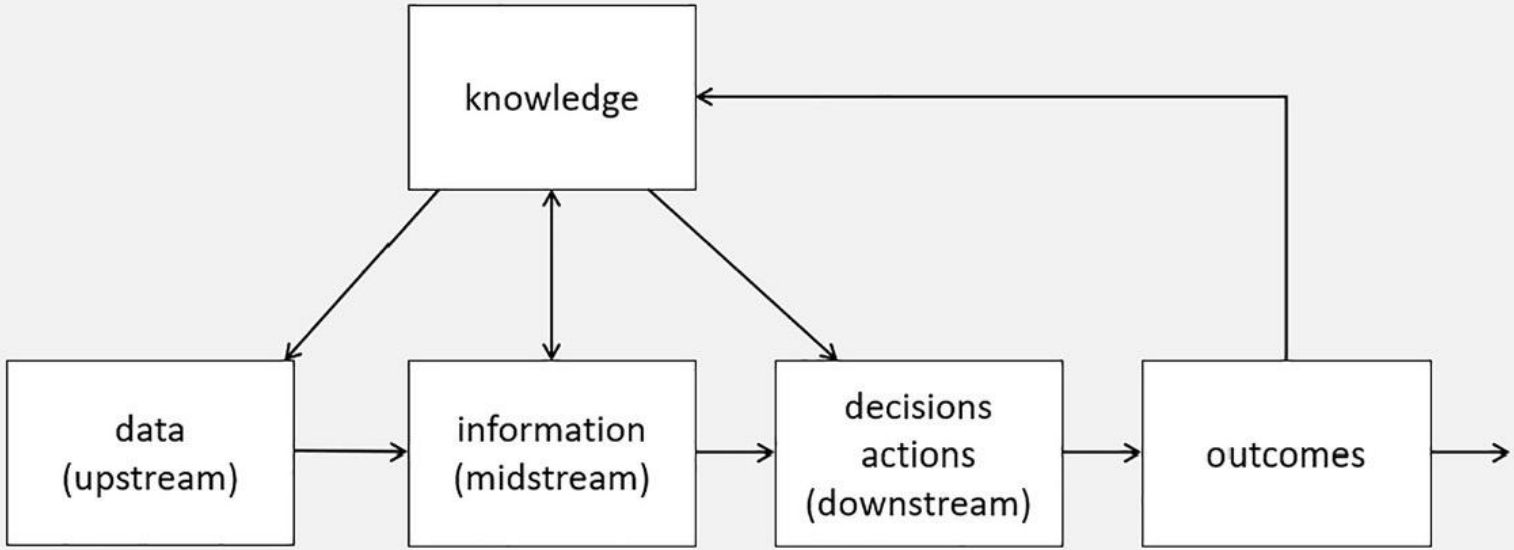
ACCESS: who (is able to) contribute?
ACCESS: who gains? (e.g. skills and connections)

Instrumental data justice

MOBILIZATION: who (is able to) use the data?
MOBILIZATION: how is the data used?

Structural data justice/Rights based data justice

STRUCTURE: Impact structure on data (e.g. impact 'digital and virtual divides' on data)
STRUCTURE: Impact data on structure (e.g. impact 'who and what is visible – and to whom?' on municipality)



Distributive data justice

STRUCTURE +
ACCESS +
MOBILIZATION:
who gets what?

Procedural data justice

ACCESS: who (is able to) contribute?
ACCESS: who gains? (e.g. skills and connections)

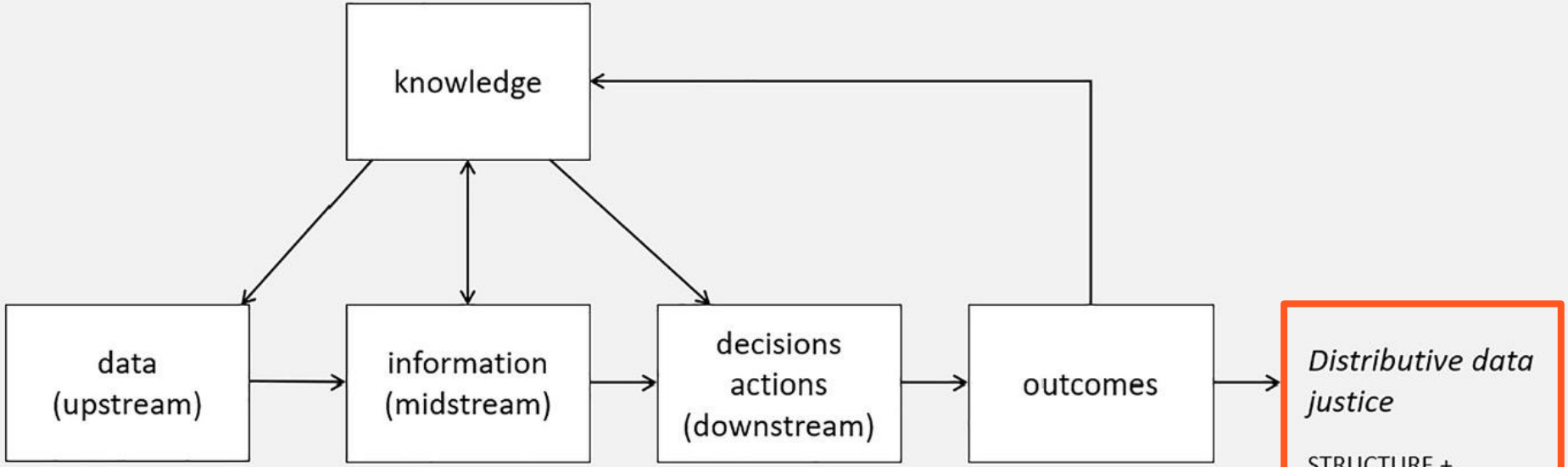
Instrumental data justice

MOBILIZATION: who (is able to) use the data?
MOBILIZATION: how is the data used?

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

Structural data justice/Rights based data justice

STRUCTURE: Impact structure on data (e.g. impact 'digital and virtual divides' on data)
STRUCTURE: Impact data on structure (e.g. impact 'who and what is visible – and to whom?' on municipality)



Procedural data justice

ACCESS: who (is able to) contribute?
ACCESS: who gains? (e.g. skills and connections)

Instrumental data justice

MOBILIZATION: who (is able to) use the data?
MOBILIZATION: how is the data used?

Some Examples

Which type of data justice could each example represent?

- Bureaucratic systems designed to assure that people are not misusing state welfare funds and other publicly funded support are part of the apparatus of governmentality
- Data-driven law enforcement focuses unequally on poor neighborhoods which experience certain types of criminality
- Undocumented migrants are tracked and acted upon by digital systems in more invasive ways than higher income travelers.
- Tech companies (eg: FB)' misuse of privacy datasets

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

Data and Precarity

> *"…problems of exclusion cannot be solved simply by including Black women within an already established analytical structure. Because the in- tersectional experience is greater than the sum of racism and sex- ism, any analysis that does not take intersectionality into account cannot sufficiently address the particular manner in which Black women are subordinated." - Crenshaw, "*Demarginalizing the Intersection of Race and Sex" (1989)

While specific to Black women, Crenshaw's theory also stands to explain the specific challenges that people experience **across multiple axes of identity** i.e., when they are Black or non-white, *and* non-male, *and/or* otherwise a member of a group facing identity-based discrimination.

The DITI borrows from legal scholar and theorist Kimberlé Williams Crenshaw's theory of *intersectionality* (1989) to understand the multiplying impact of data discrimination and algorithmic bias on oppressed persons, and to help consider possibilities for data justice.

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

Considering Intersectionality in Data Justice

A range of interacting and overlapping identity characteristics (e.g., *race, ethnicity, religion, gender, location, nationality, socio-economic status, etc.*) **determine how individuals are made into administrative (institutionally) and legal (as non/citizen) subjects through their data** and, consequently, **how data can be used to act upon and against them** by policymakers, commercial firms, and other entities.

Depending on the various identities a person inhabits–especially for with regard to race, gender, and sexuality–**the likelihood of and frequency by which someone identified as a target of surveillance multiplies.**

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

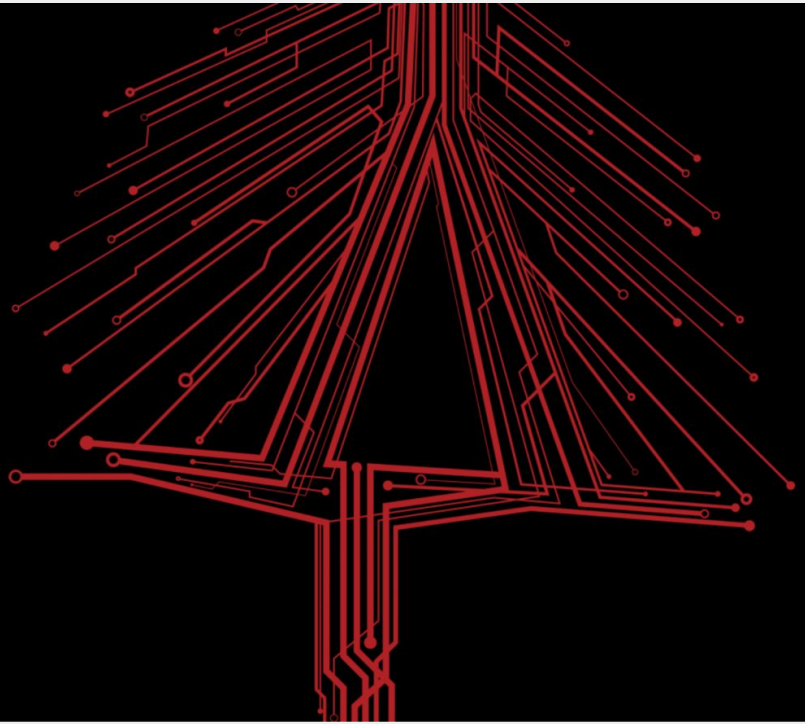
A **Data Justice** Example from the Boston Area Research Initiative (BARI)

Northeastern University
NULab for Texts, Maps, and Networks

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

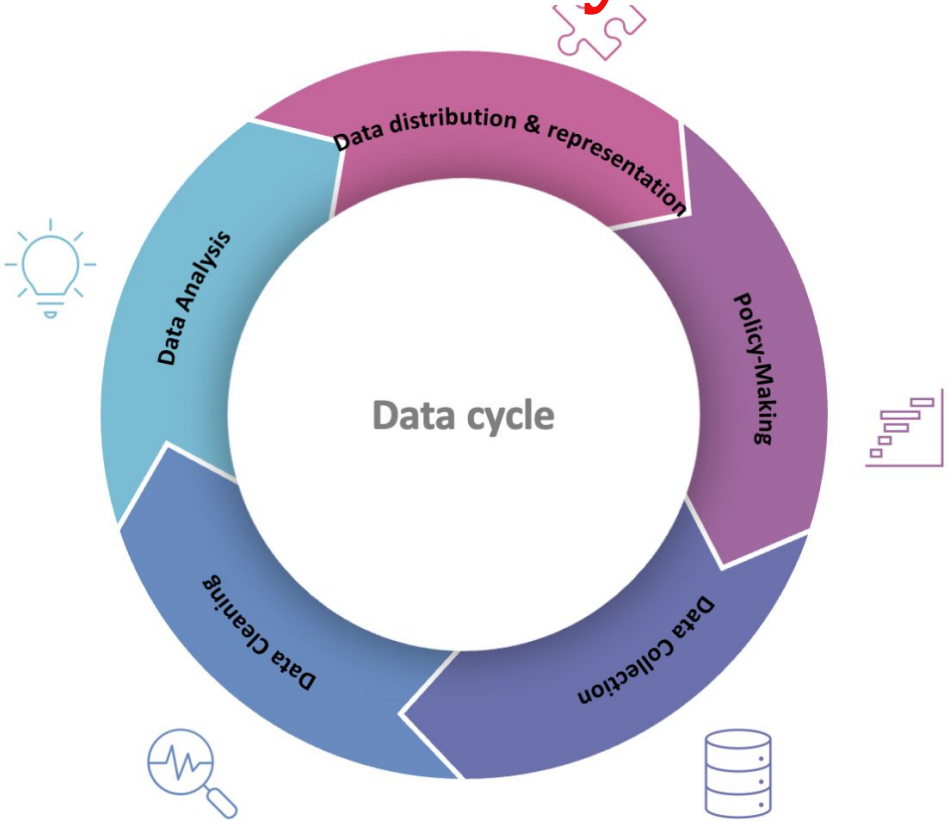
BARI's Statement on Data Justice:

BARI's Data Justice Statement



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

The Data Cycle:



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

BARI's Statement on Data Justice

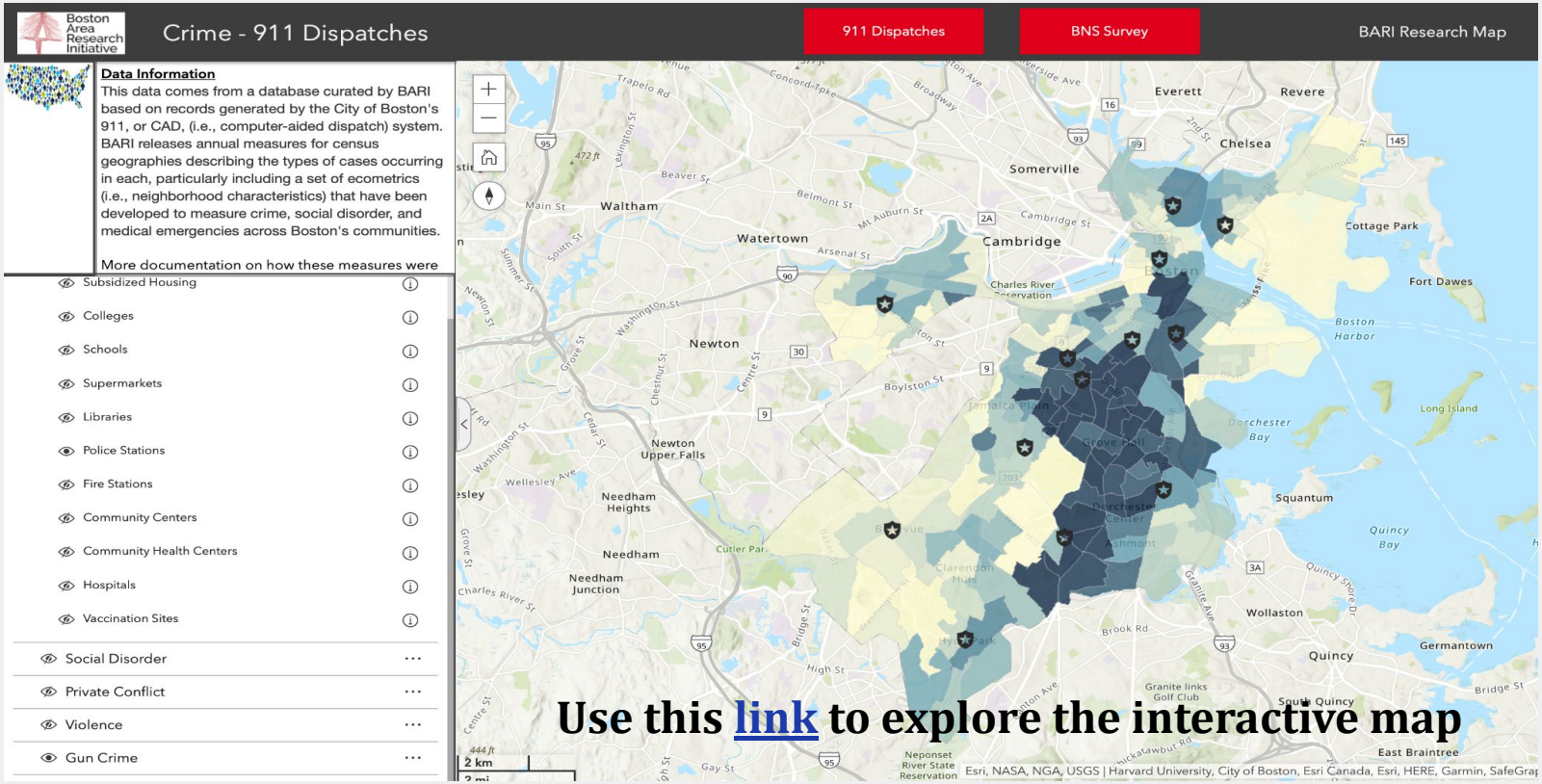
Examine the following excerpts from

- Data is for all, thus the data accessibility and employment should be inclusive, representative and serve for every voice.
  1) data training targeted contents: we provide training on how to engage with data (including data analysis and data visualization) for *all* potential users through an intersectional approach
- Transparency in acknowledging the strengths and weaknesses of naturally occurring data (seeing the holes) and communicating effectively about methods and limitations.

Discussion question:
Imagine, you just scraped Airbnb dataset from websites:
How will you clean the datasets?
How will you communicate the limitation of the research?

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

An Example from BARI: Boston 911 Data



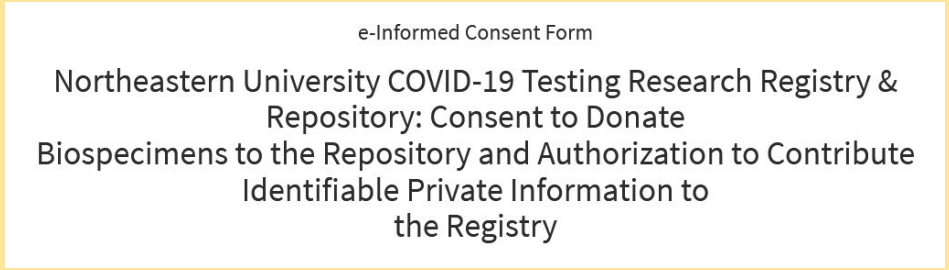
Class Activity:
Northeastern's COVID-19 Research - Data Collection and Consent

Northeastern University
NULab for Texts, Maps, and Networks

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

Northeastern's COVID-19 Research: Consent Form Analysis

- **Northeastern runs a testing program, which it tries to connect with research initiatives**
- The research project is called **"Northeastern University COVID-19 Testing Research Registry & Repository,"** and the lead researchers are Jared Auclair, PhD. and Spencer Pruitt, PhD. from the Department of Biotechnology & Bioinformatic, College of Science; ITS. The research is sponsored by Northeastern University.
- **In order to collect data for research, the researchers have provided a consent form—let's examine some language from the form, 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!

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: Adopt or Not?

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

You will be assigned into small groups for this exercise. Have a group discussion and try to come up with a unified 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!

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 onto**?
- 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!

Class Activity:
Search Engine Bias
Example and Discussion

Northeastern University
NULab for Texts, Maps, and Networks

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.

Northeastern University
NULab for Texts, Maps, and Networks

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?

Northeastern University
NULab for Texts, Maps, and Networks

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.

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

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!

Data Presentation: Considerations

Northeastern University
NULab for Texts, Maps, and Networks

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

Misrepresentation of Data

From D.B. Resnik, in International Encyclopedia of the Social & Behavioral Sciences, 2001:

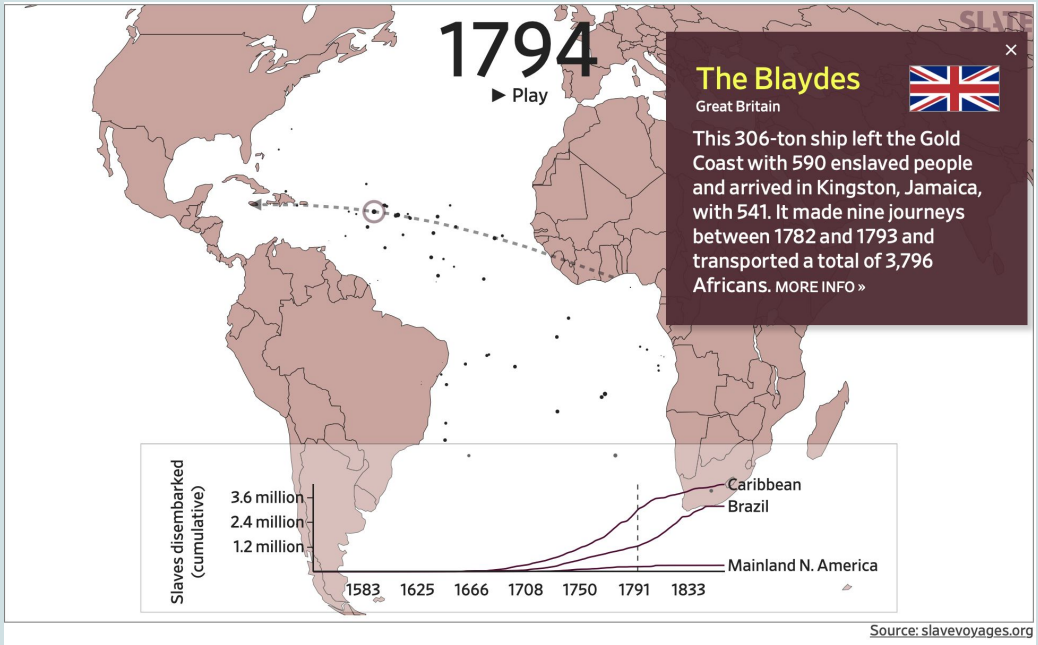
"The concept of 'misrepresentation,' unlike 'fabrication' and 'falsification,' is neither clear nor uncontroversial. Most scientists will agree that fabrication is making up data and falsification is changing data. ***But what does it mean to misrepresent data? As a minimal answer to this question, one can define 'misrepresentation of data' as 'communicating honestly reported data in a deceptive manner.'***"

- This **online book from The Data School** covers some common ways data could be misrepresented at multiple points in the process of gathering, analyzing, and presenting findings on data-based research.

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

Even when data isn't being willfully misrepresented, the way it's presented can still end up being ***reductive…***

This is a screenshot from [a digital history project from Slate] that visualizes information from the [Trans-Atlantic Slave-Trade Database] as an animated map. In the map, **each dot represents individual slave ships**, and the size of the dot corresponds to the number of enslaved passengers aboard. You can learn more about each ship's history by clicking on its respective dot.



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

In this case, the map is presenting humans as *objects* rather than *people*.

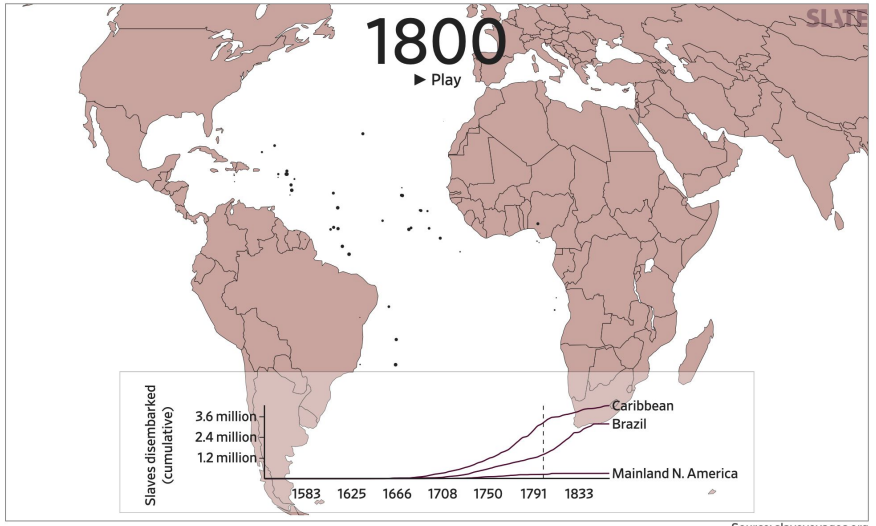
Reductive data can end up doing real **harm**.

- What happens when human lives become **reduced to data points**?
- **What is lost and what is gained** in visual representations of data like this?
- How can we represent data both ***accurately***, ***completely,*** and ***with care***?

The Atlantic Slave Trade in Two Minutes

315 years. 20,528 voyages. Millions of lives.

BY ANDREW KAHN AND JAMELLE BOUIE SEPT 16, 2021 • 4:18 PM



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

Limitations of Some Data Presentation Methods: Maps

- Viewers may have **limited knowledge** about the spaces depicted
- **Mapping technologies** may not accurately/completely show all relevant variables
- **Navigability** and **clarity** are concerns. Consider: how usable is the map?
- Maps may not have been **normalized** (normalizing refers to adjusting data that may have been collected at different scales into a common scale), so comparisons might be inaccurate or misleading
- Like any other type of rhetoric, **maps can be used to tell**—or obfuscate—**specific stories**

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

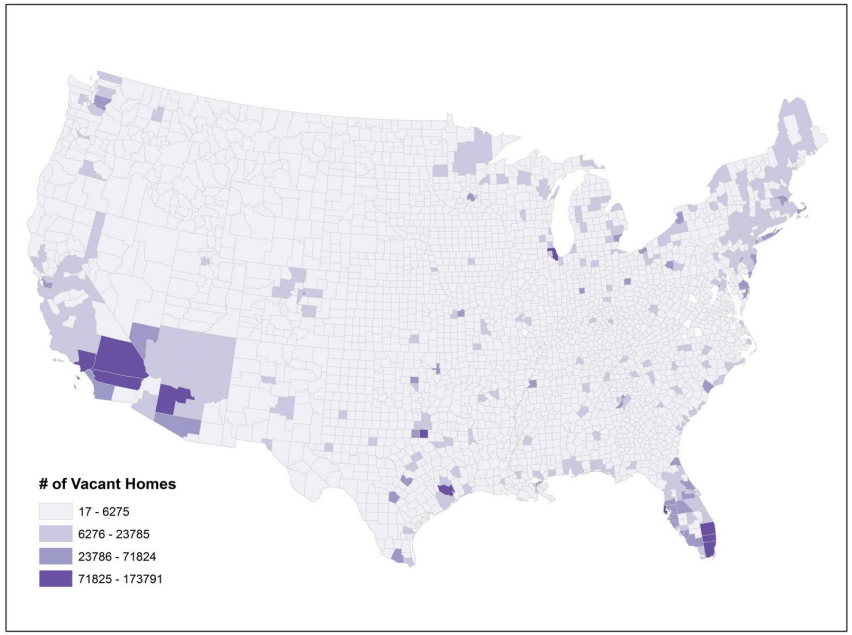
Which map of the T is more *navigable*?

or

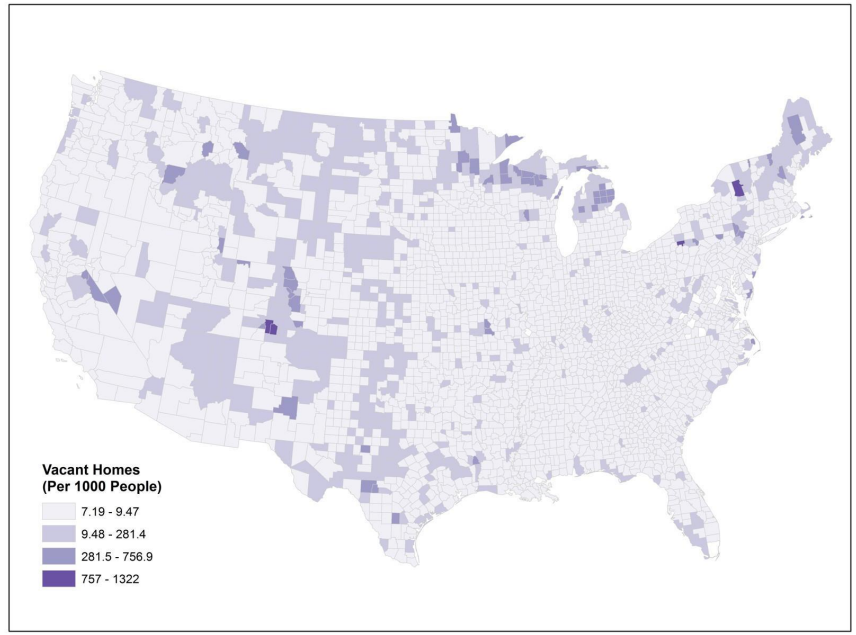
Northeastern University
NULab for Texts, Maps, and Networks

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

Example of Unnormalized vs. Normalized Maps



Unnormalized Map of Vacant Houses in the U.S.
Credit: U.S. Census website



Normalized Map of Vacant Houses in the U.S.
Credit: U.S. Census website

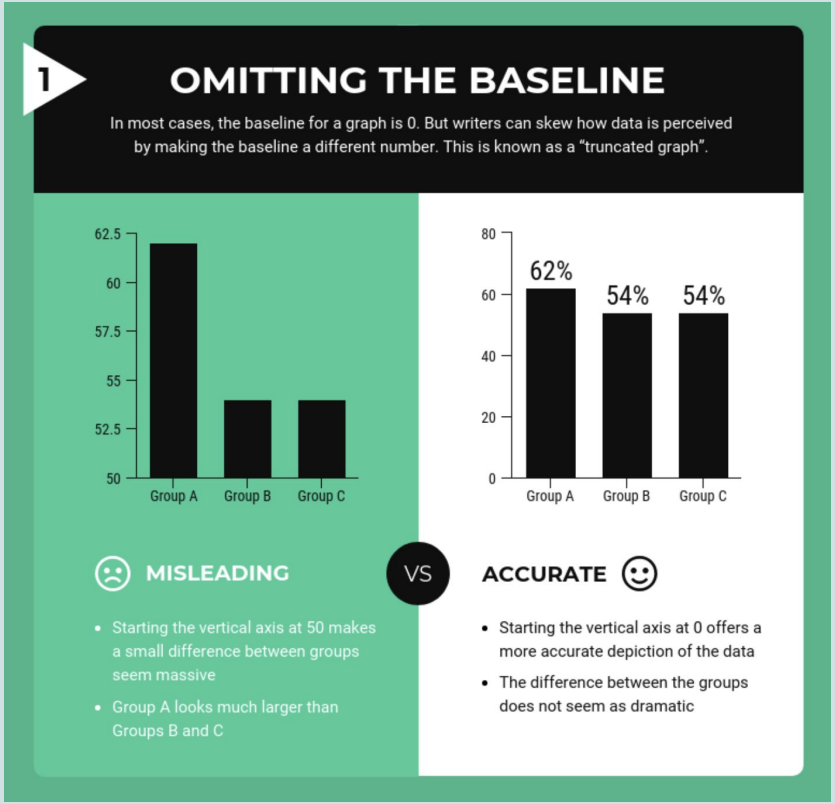
Feel free to ask questions at any point during the presentation!

Limitations of Some Data Presentation Methods: Charts and Diagrams

- The **structure** and **scale** of charts and graphs could be **manipulated** to amplify or diminish differences
- **Different types** of graphs and charts work better for some types of data presentation than others—for example, a pie chart and a line graph might not both be able to represent the same data accurately
- A chart with **too much information** will be difficult to understand, but **too little information** could be an indication that data has been cherry-picked to support an argument

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

Examples of Limitations using Charts and Diagrams

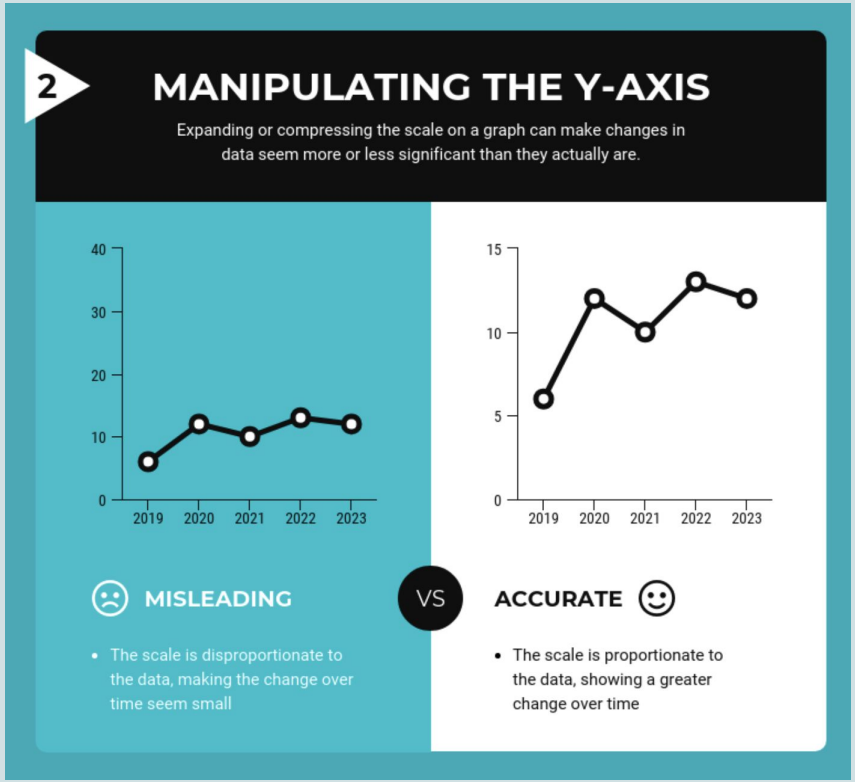


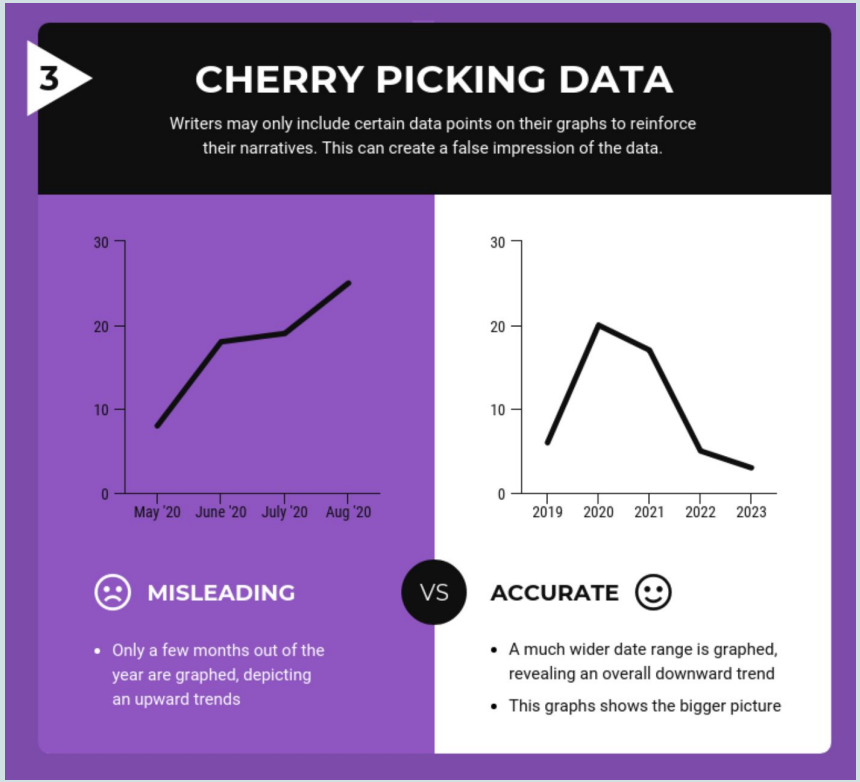
Discussion:

- What **commonalities** do you notice among the more misleading and more accurate versions of graphs and charts in these examples?
- How would you define "**accuracy**" in the context of data presentation? Why is that question essential to ask?
- In what **contexts** does it make the most sense to use these kinds of visuals to present data? Are there other times where they're inappropriate? How so?

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

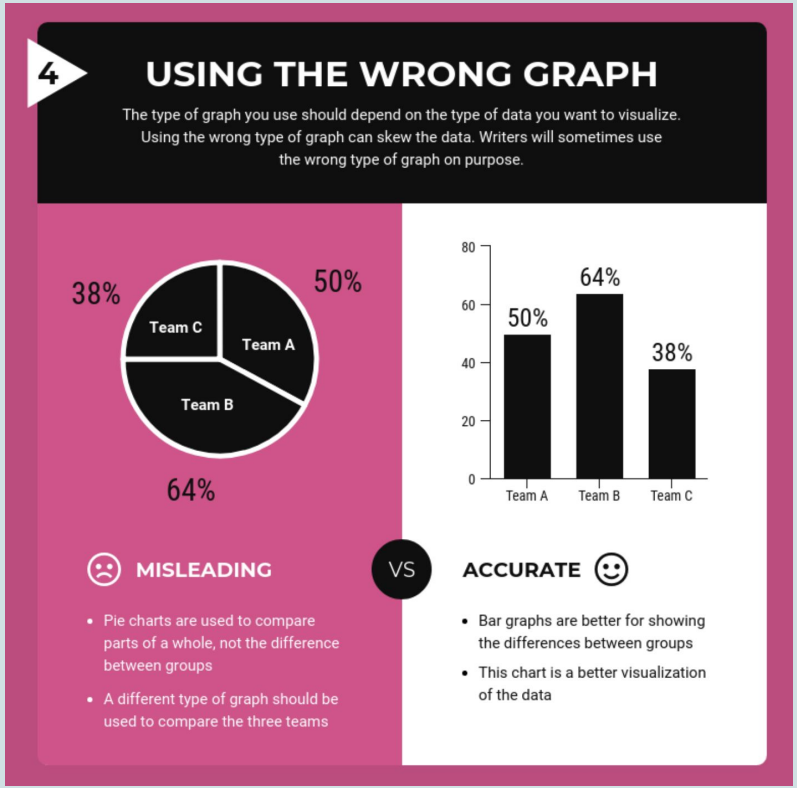
More limitations with presenting data using **CHARTS and DIAGRAMS:**

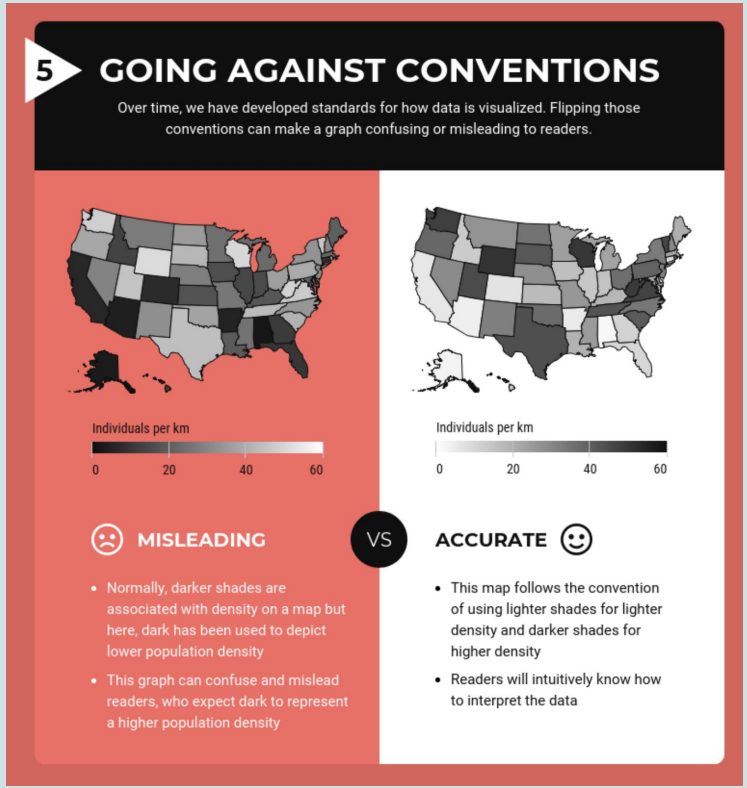




Feel free to ask questions at any point during the presentation!

More limitations with presenting data using **GRAPHS and MAPS:**





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?

Northeastern University
NULab for Texts, Maps, and Networks

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 **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!

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

When *reading, evaluating, and citing the work of others*, be **data-literate**:

- turn a critical eye to studies that use big data
- evaluate the sources of that data
- carefully examine the conclusions authors draw from their sources

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

Sources are important!

The inherent intertextuality of academic writing is important for us to remember as we begin to plan and execute research projects.

- So much academic work involves weaving together multiple texts, synthesizing the work of others and using it to frame or support your argument - the way you use and analyze the work of others is key!
- Much of your work will probably involve using sources published in academic books and journals, or primary sources from archives, but non-traditional and non-academic sources could pay an important role in your work, too.
- In your own work you can address silences and gaps in scholarship by seeking a variety of traditional and non-traditional sources to support your argument or narrative, and amplify voices that are heard less often.

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)

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 DITI Research Fellows:
Tieanna Graphenreed, Vaishali Kushwaha, Cara Messina, Yana Mommadova, Garrett Morrow, Colleen Nugent, Milan Scobic, and Claire Tratnyek, with help from BARI Data Specialist Shunan You

Slides, handouts, and data available at **https://bit.ly/3Q991EG**

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

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