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

HIST 2430 Digital Histories of Ethnic Boston
Professor Nick Brown
Spring 2023

Taught By: Chris McNulty & Benjamin Grey

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 is 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 data reflects and reinforces cultural, social, and political biases.
- Explore ways of interpreting and effectively utilizing data-based evidence in written arguments.

Slides available at: **bit.ly/sp23-brown-dataethics**

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!

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.

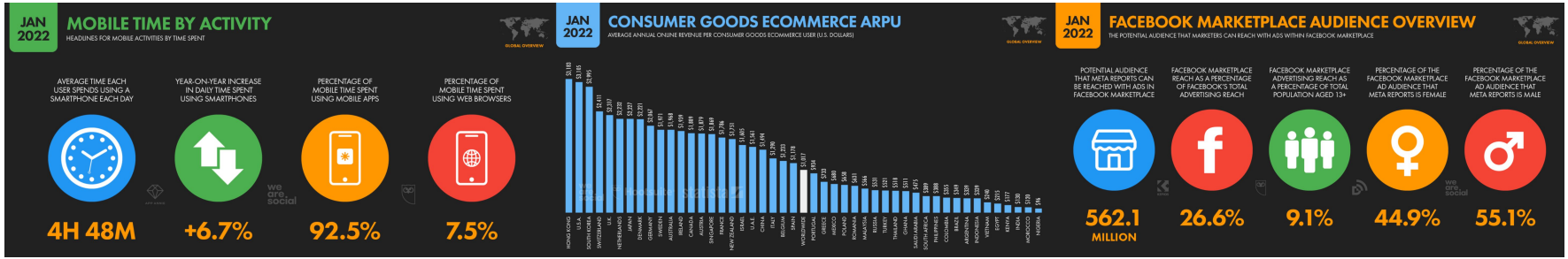
The goal of big data is **to predict individual user behavior based on patterns from the user as well as patterns from "similar" users** (based on demographic information, behavioral patterns, etc).

We're living in an era of "surveillance capitalism"**—our information is a valuable *product*.**

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

Big Data is getting bigger

- Internet usage is constantly increasing: 62.5% of the world's population regularly uses the internet, and 58.4% of those users use **social media**.
    - 35% of all time spent on the internet is spent on social media.
- All sorts of data is collected about online audiences and their activity.
- The big data collected allows advertisers to **target** users.
    - **Spending** on social media advertising is also increasing: $154bn was spent globally in 2021.



Source: DataReportal

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.
- Some websites **monetize** data in a "data exploitation market," selling their users' personal information.
- Big data can be used to (inadvertently or purposefully) **entrench stereotypes** or **reproduce results** that may harm certain communities.

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?**

Big Data, Online Presence, & Data Privacy

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**
- **News and information**
- **Social media**
- **Travel and transportation**
- **Education and employment**
- **Public policy and safety**

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

How does Big Data impact our daily lives?

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

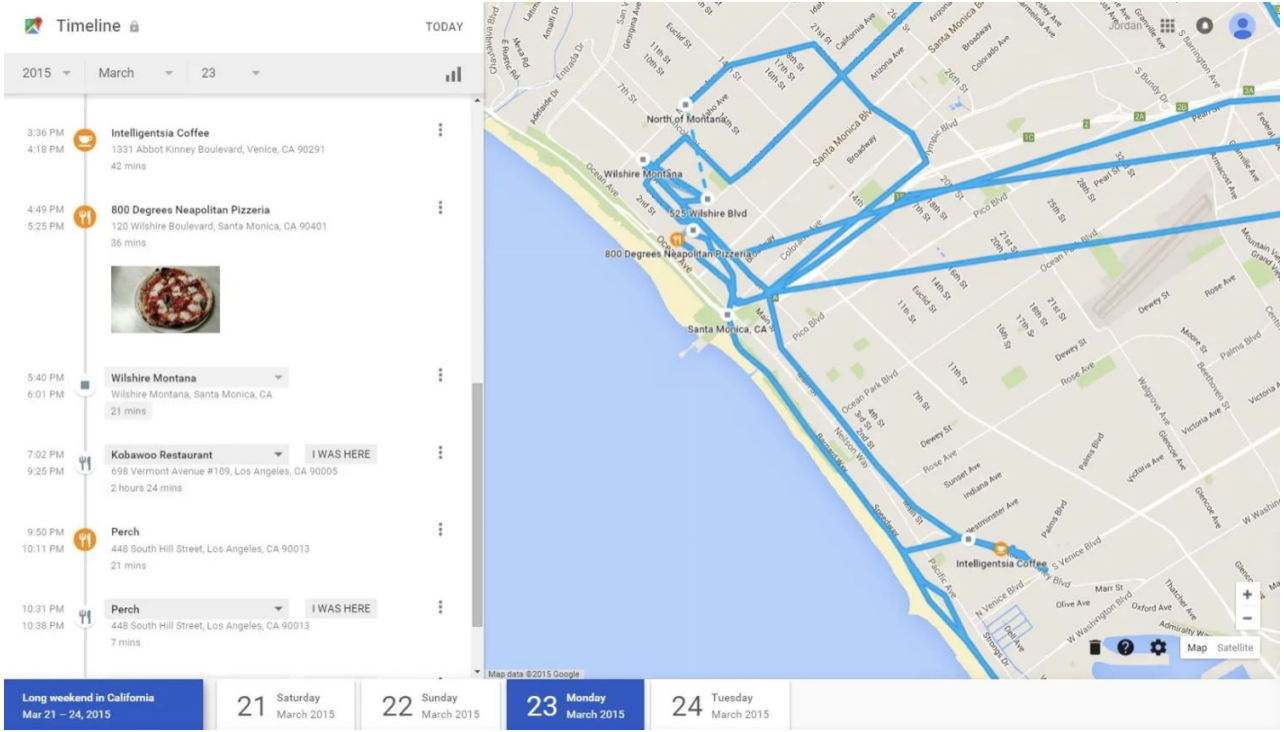
Google, Amazon, Apple, and Microsoft are all central players in "surveillance capitalism" and prey on our data.

THE ANTI-MEDIA

If you have **location services** turned on for Google (for instance, 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 Maps 'Your Timeline' feature here.

How does Big Data impact our daily lives?

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!

How Do We Know How We Are Being Tracked?

There are ways to identify what information websites collect about their users.

Blacklight is a "real-time website privacy inspector" 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.

You can try Blacklight 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

TikTok: Profile > 3-line icon (top right) > Settings & Privacy > Privacy > Download your data

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

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”?

Technology is Not Neutral

Information systems like Google as well as data collection, data analysis, and algorithms are **not neutral.**

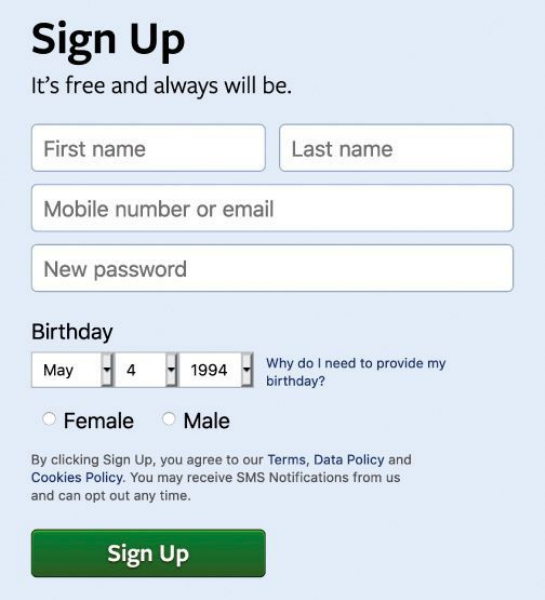
They can **reinforce** systemic, political, and cultural **biases.**

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.

Collecting Data

Which of these is likely to collect more accurate and representative data about users' gender?



Source: Facebook's new account creation page circa 2018, published in D'Ignazio & Klein, *Data Feminism*, 2020

Source: Positive Voices survey from Public Health England, published in D'Ignazio & Klein, *Data Feminism*, 2020

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

Algorithms & Big Data: *What gets counted counts*

D'Ignazio and Klein identify problematic data practises that cause harm:

- Lack of quantitative research on maternal mortality masks systemic problems.
- Undocumented immigrants are often (sometimes voluntarily) absent from census data, which determines levels of federal funding: a "paradox of exposure."
- TSA scanning machines binarize bodies to attempt to uncover concealments, but can thereby mistakenly assign risk alerts.

"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..."

Catherine D'Ignazio & Lauren Klein, *Data Feminism*, 2020

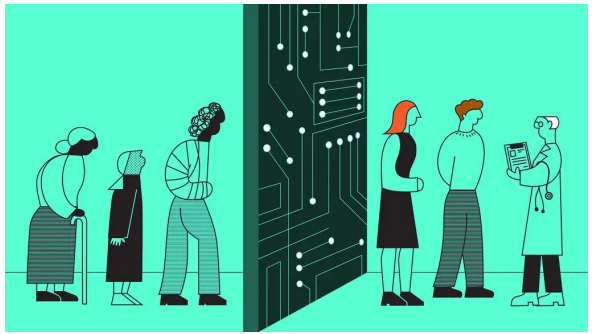
Algorithmic Bias

- Algorithms are *not neutral*. **People create algorithms**.
  - Algorithmic processes—and even the data itself—reflect societal biases.
- When an algorithm is written or trained using data that misrepresents the actual population, this produces **algorithmic bias.**
- Similarly, **when data reflects biased realities**, the algorithm will continue to reproduce outcomes if those outcomes are desirable (despite their harm to—or erasure of—other groups).
- Algorithms reflect social inequalities, and can serve to exacerbate them.
- Read this Vox article for more information on algorithmic bias.

Algorithmic Injustice: Healthcare

- Algorithms are used in public health systems to **inform decisions** of who should receive preventative care and medical treatment.
- They **predict** the likelihood of specific conditions, e.g. cardiovascular risk, or of general ill-health among different demographic groups.
- But the data used to make these predictions is often collected from **white patients**, which makes risk scores far less accurate for African American or other non-white patients.
    - The Harvard School of Public Health estimates that **Caucasians make up 80 percent** of collected data of in the field of genomics and genetics.

"We found that a category of algorithms that influences health care decisions for over a hundred million Americans shows significant racial bias."
Sendhil Mullainathan, Chicago Booth University



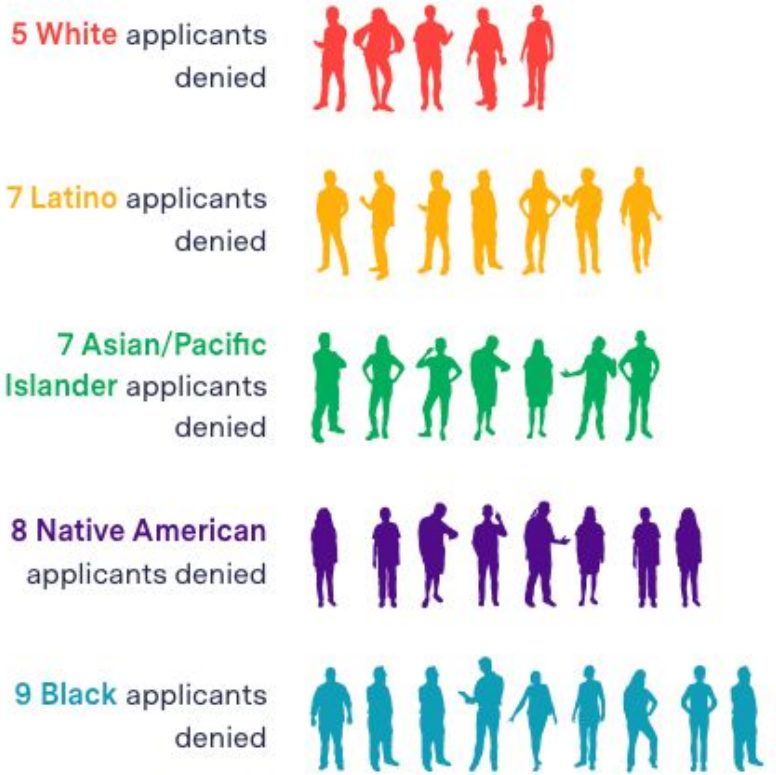
Source: Jenice Kim, *The New York Times*

Source: Katherine Igoe, Harvard TH Chan School of Public Health, 2021

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

Algorithmic Injustice: Mortgages

- Mortgage approval algorithms can gather and use data in ways that express a racial bias.
- On Fannie & Freddie, which buys 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."



Source: Emmanuel Martinez & Lauren Kirchner, *The Markup*, 2021

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

Alleviating Injustice

- 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?

"Algorithms by themselves are neither good nor bad. It is merely a question of taking care in how they are built."
Sendhil Mullainathan, Chicago Booth University

"Counting and measuring do not always have to be tools of oppression. We can also use them to hold power accountable, to reclaim overlooked histories, and to build collectivity and solidarity."
Catherine D'Ignazio & Lauren Klein, *Data Feminism*, 2020

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

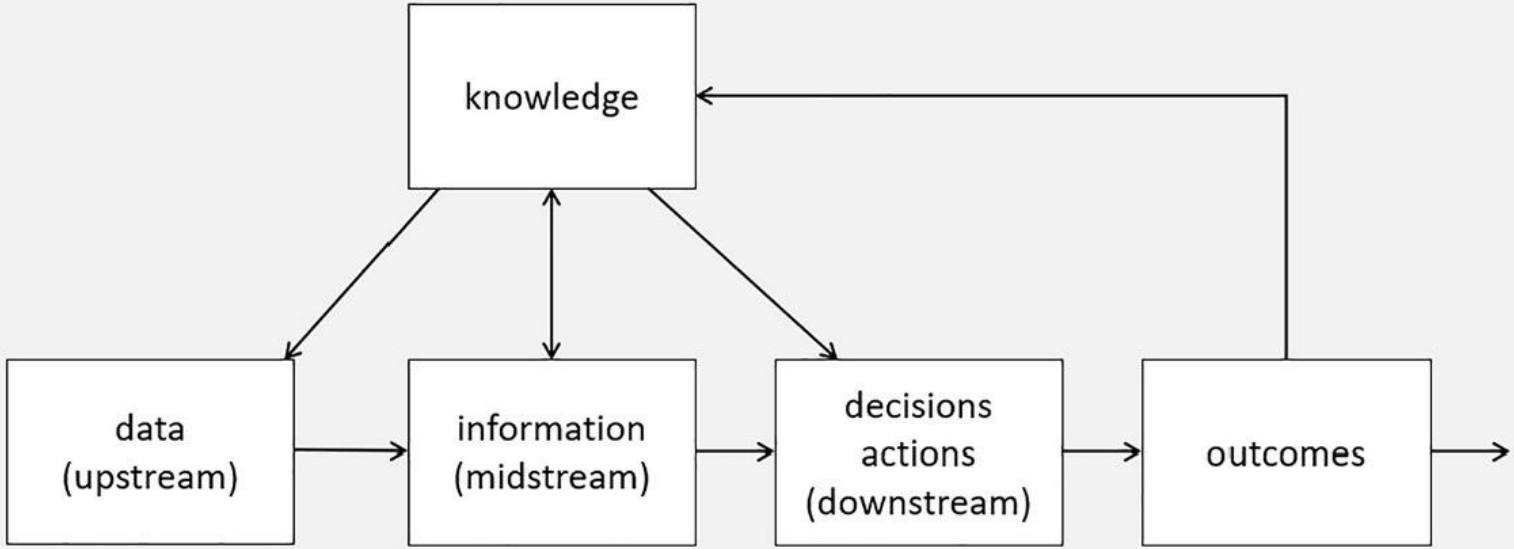
Data Justice

Northeastern University
NULab for Texts, Maps, and Networks

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

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

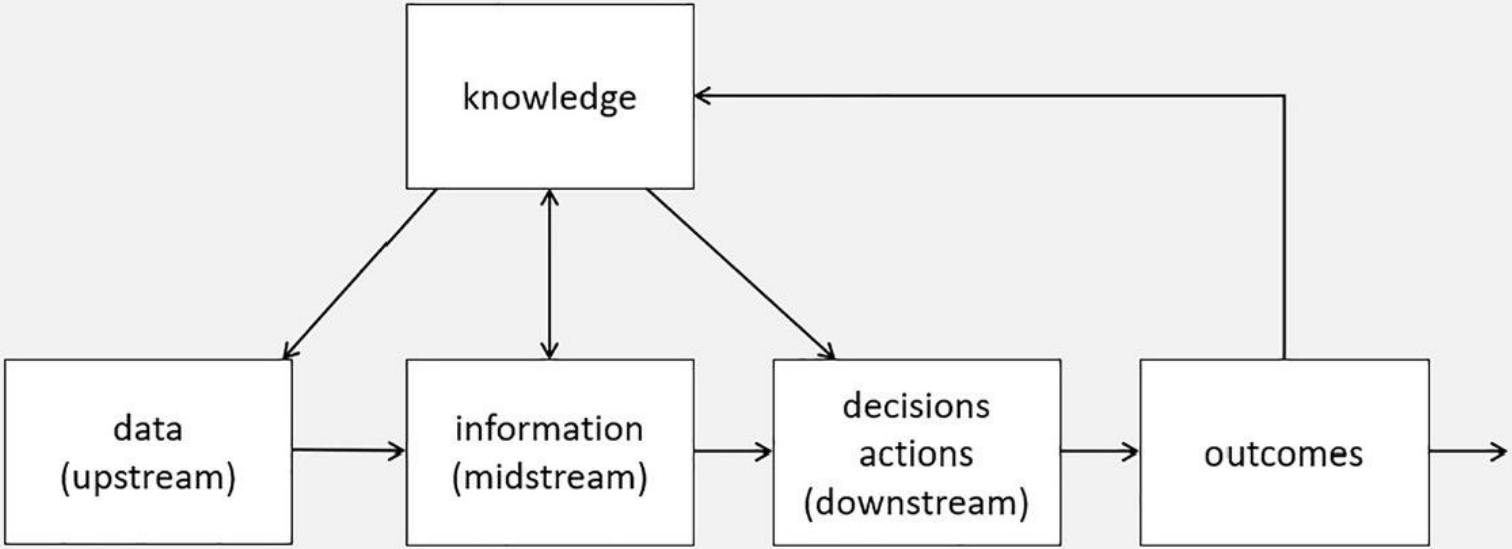
ACCESS: who (is able to) contribute?
ACCESS: who gains? (e.g. skills and connections)

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

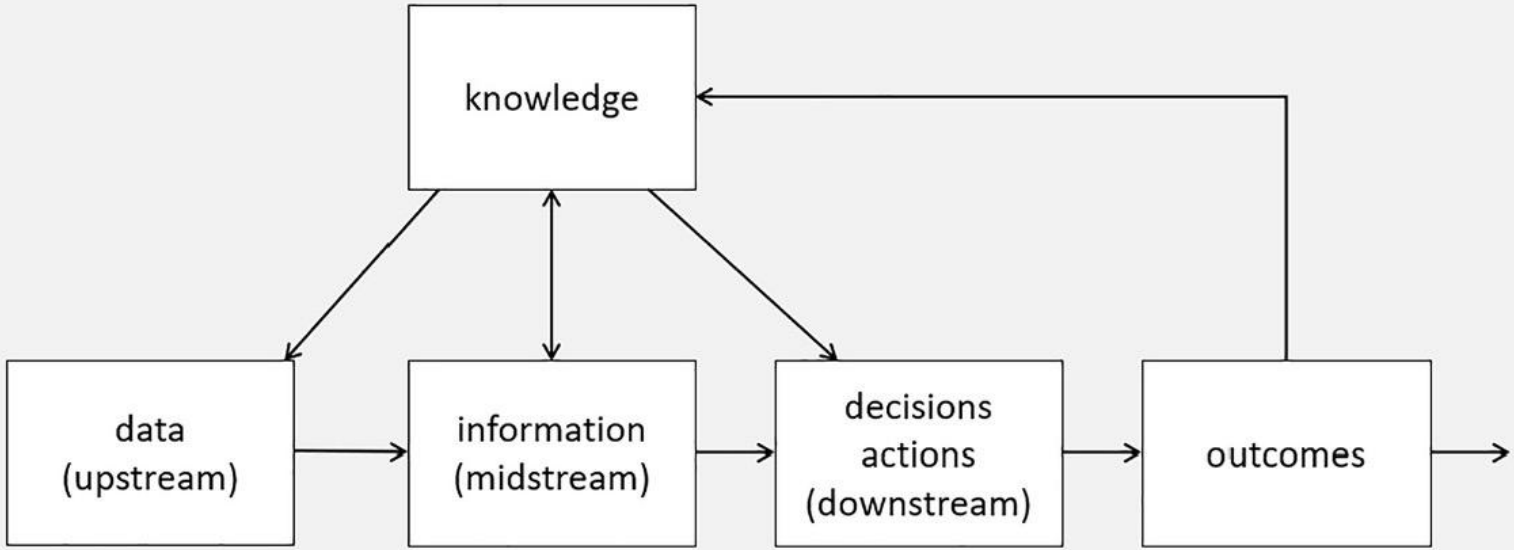
ACCESS: who (is able to) contribute?
ACCESS: who gains? (e.g. skills and connections)

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)

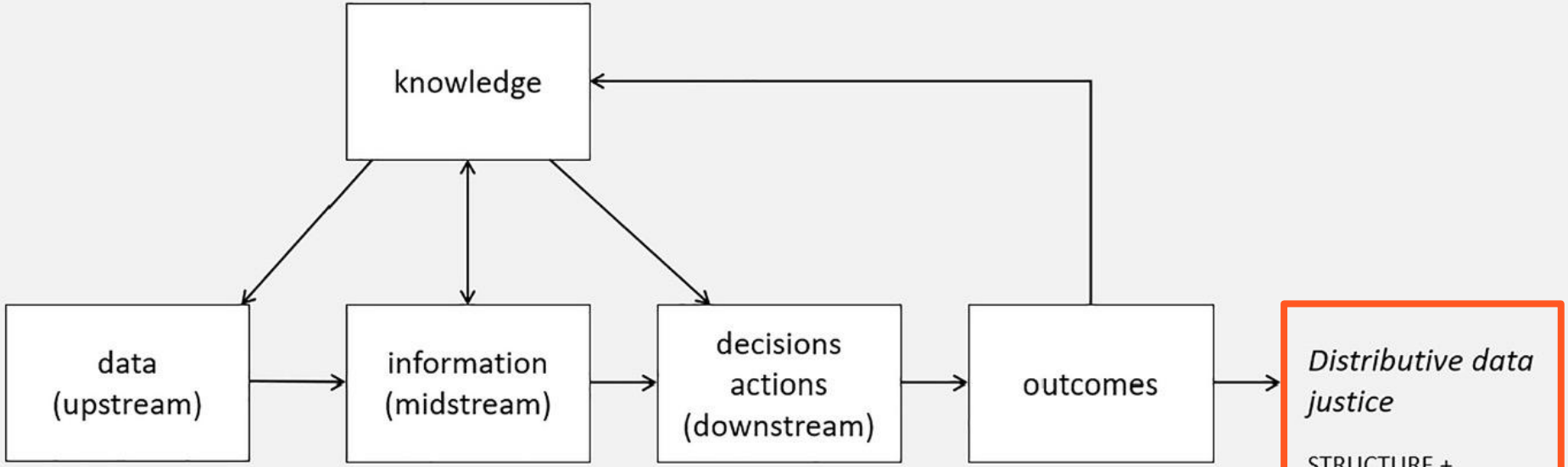
Instrumental data justice

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: Facebook) 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** 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!

Biases in Scholarship and Archival Silences

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?

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?

Moving Forward -
How can we use data in our work responsibly?

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

Questions Researchers Must Ask

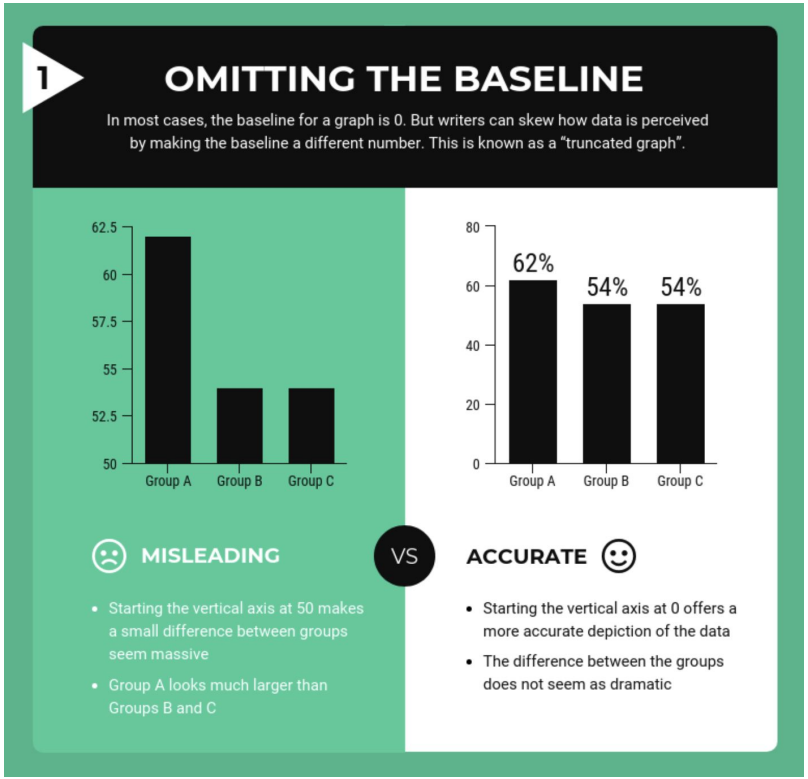
- What **information** is being collected and from where? To whom does this data **belong**?
- How is it being **collected**? Do **participants** know that it is collected, how it will be collected, and how will it be used?
- **How** will the data be analyzed? What **biases** and **ideologies** may be implicit in this analysis?
- Who will this research impact? Who will it **benefit**? Who will it potentially **harm**?

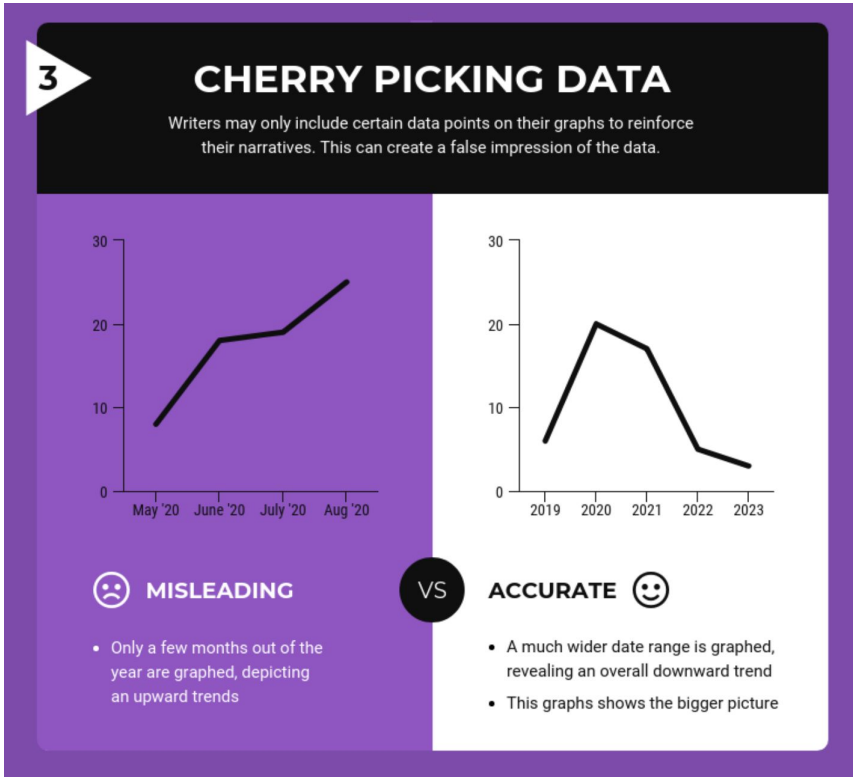
Feel free to ask questions at any point during the presentation!

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***?

Be Mindful of Infographics and Data Visualizations





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!

We love feedback! **Please fill out our 2-min survey:** bit.ly/diti-feedback

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

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

Developed by DITI Research Fellows: Claire Tratnyek, Vaishali Kushwaha, Yana Mommadova, Colleen Nugent, Tieanna Graphenreed, Javier Rosario, Ana Abraham & Chris McNulty

Slides & handout available at: bit.ly/sp23-brown-dataethics

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