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

HIST 1357 History of Information
Victoria Cain Spring 2022

Taught By: Claire Tratnyek & Colleen Nugent

Northeastern University
NULab for Texts, Maps, and Networks

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

Workshop Goals

- Engage with critical rethinking of everyday practices related to data collection and use, as well as how algorithms impact and shape our daily lives
- Explore ways of interpreting and effectively utilizing data-based evidence in written arguments
- Understand the ways in which technology reflects cultural, social, and political biases

Slides, handouts, and data available at **https://bit.ly/cain-dataethics-sp22**

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

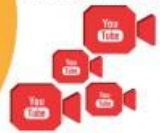
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

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

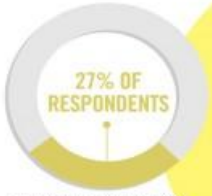
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

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**

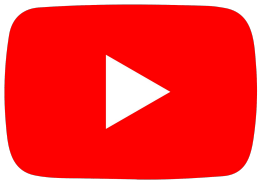
Public policy and safety

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

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!

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!

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!

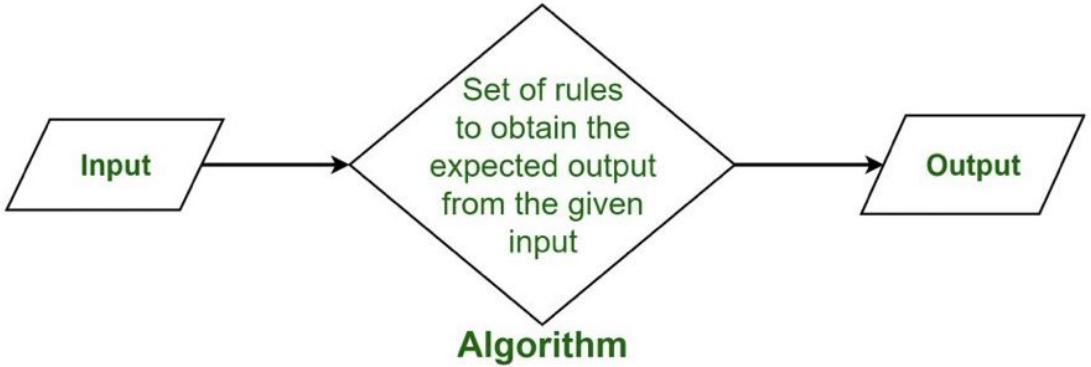
Algorithms and Bias

Northeastern University
NULab for Texts, Maps, and Networks

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

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.

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!

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?

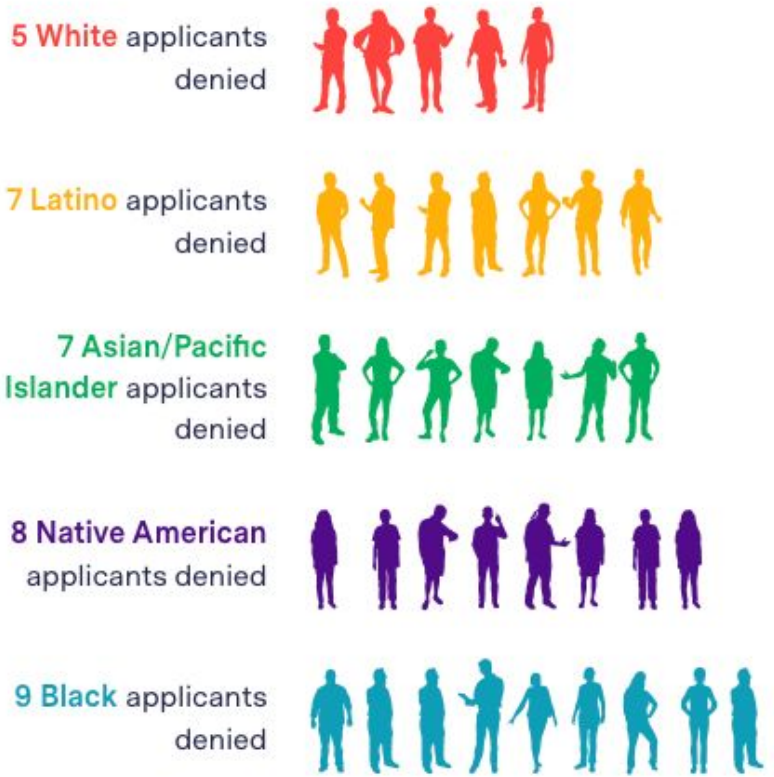
SOURCE: **"What Gets Counted Counts" Principle #4 of Data Feminism (mitpress, 2020)**

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

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!

Considering Intersectionality

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!

Working towards 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) be treated 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!

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 in 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!

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!

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:

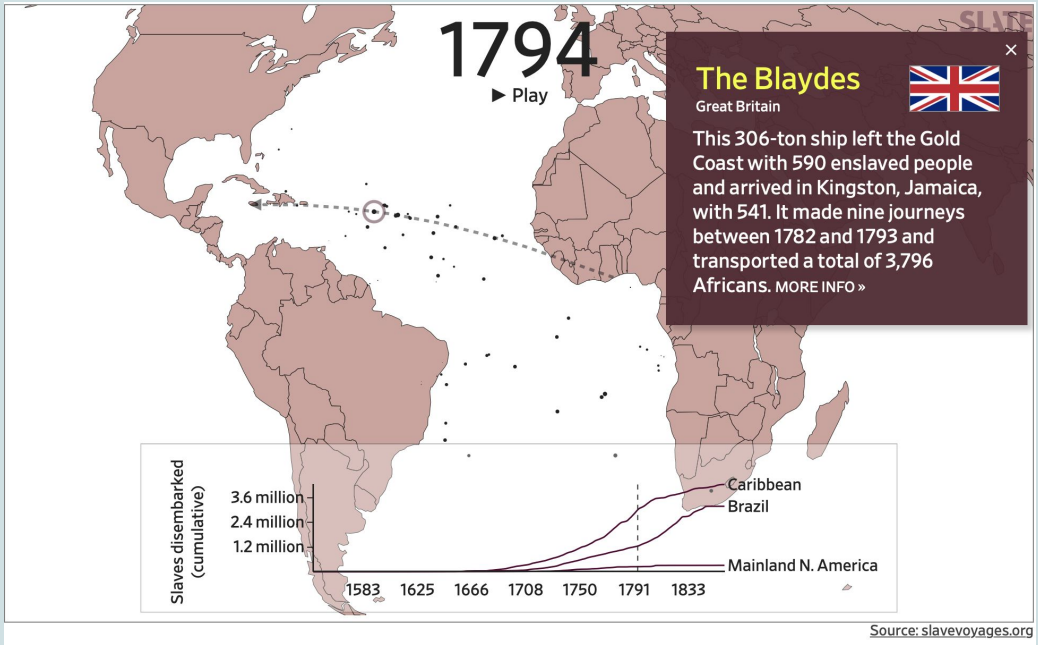
"The concept of 'misrepresentation,' unlike 'fabrication' and 'falsification,' is neither clear nor uncontroversial. Most scientists will agree that fabrication is making up data and falsification is changing data. ***But what does it mean to misrepresent data? As a minimal answer to this question, one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

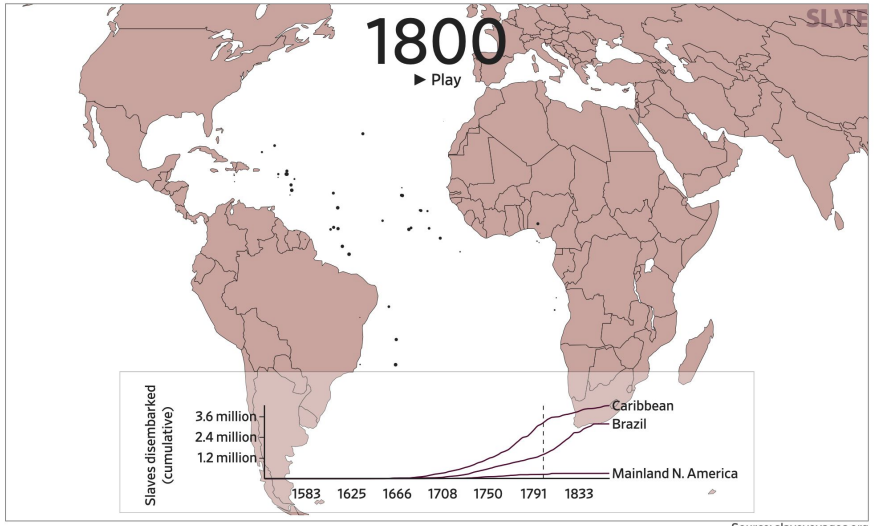
Reductive data can end up doing real **harm**.

- 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



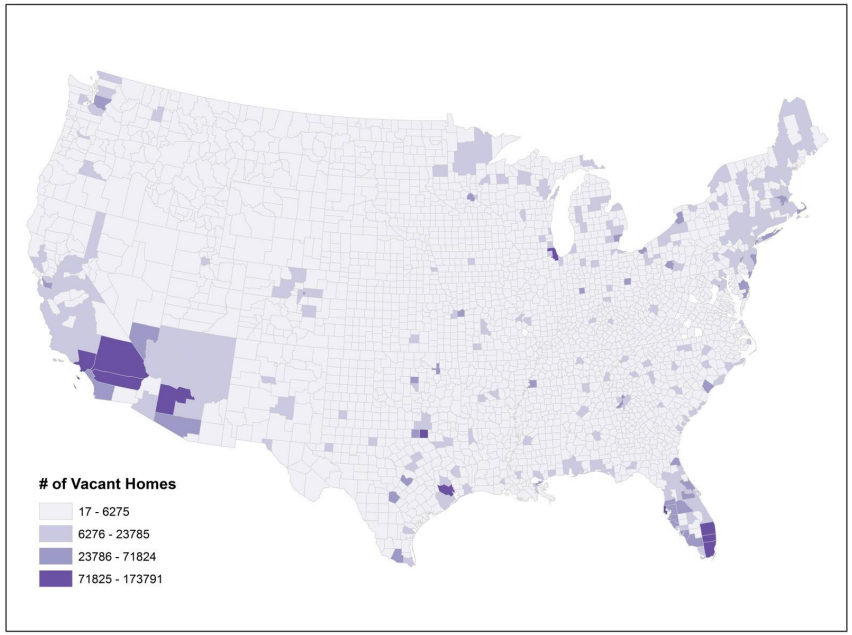
Feel free to ask questions at any point during the presentation!

Limitations of Some Data Presentation Methods: Maps

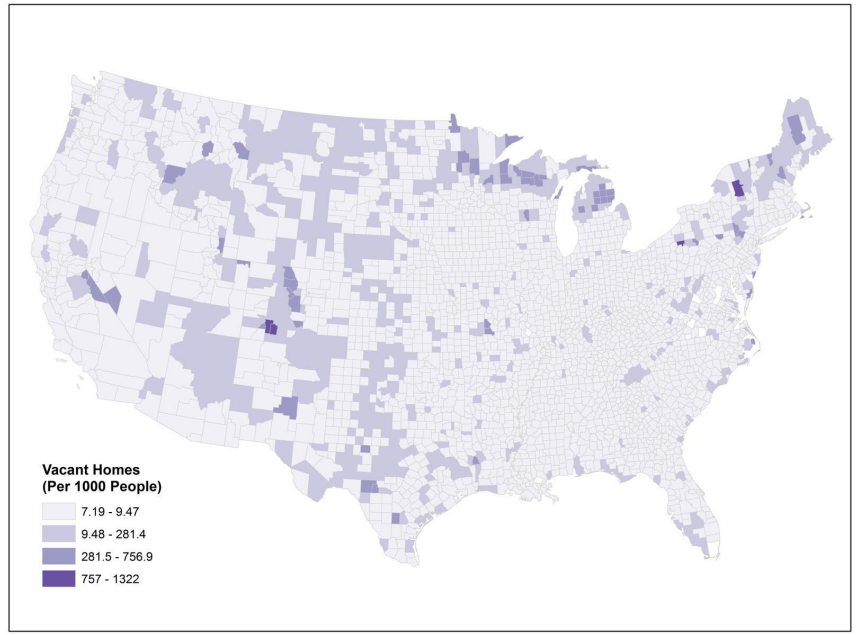
- Viewers may have **limited knowledge** about the spaces depicted
- **Mapping technologies** may not accurately/completely show all relevant variables
- **Navigability** and **clarity** are concerns. Consider: how usable is the map?
- Maps may not have been **normalized** (normalizing refers to adjusting 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!

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

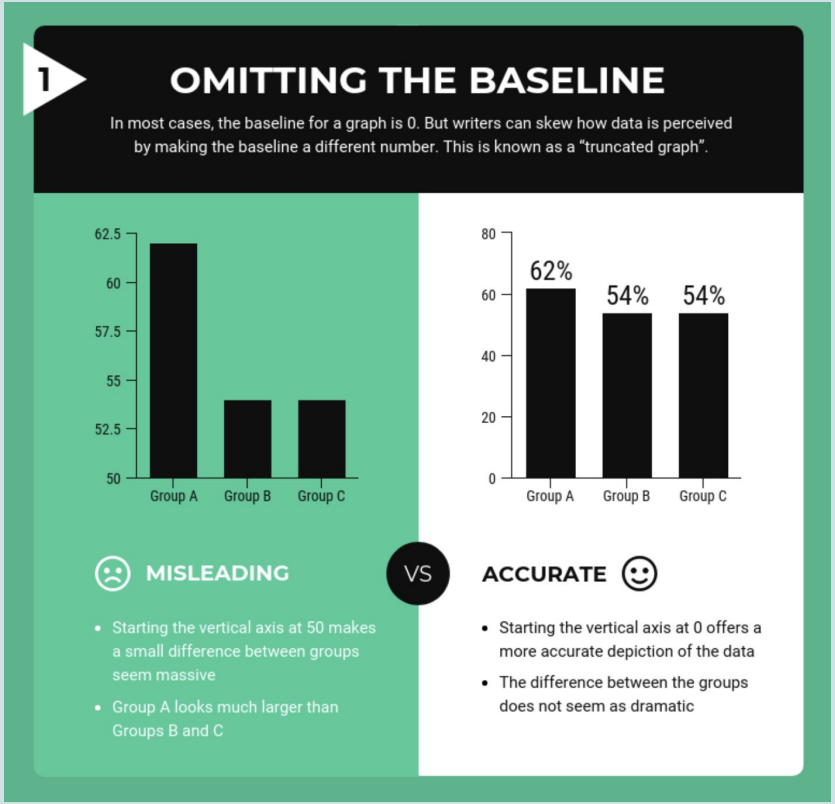
Feel free to ask questions at any point during the presentation!

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 Graphs, Charts, & Maps

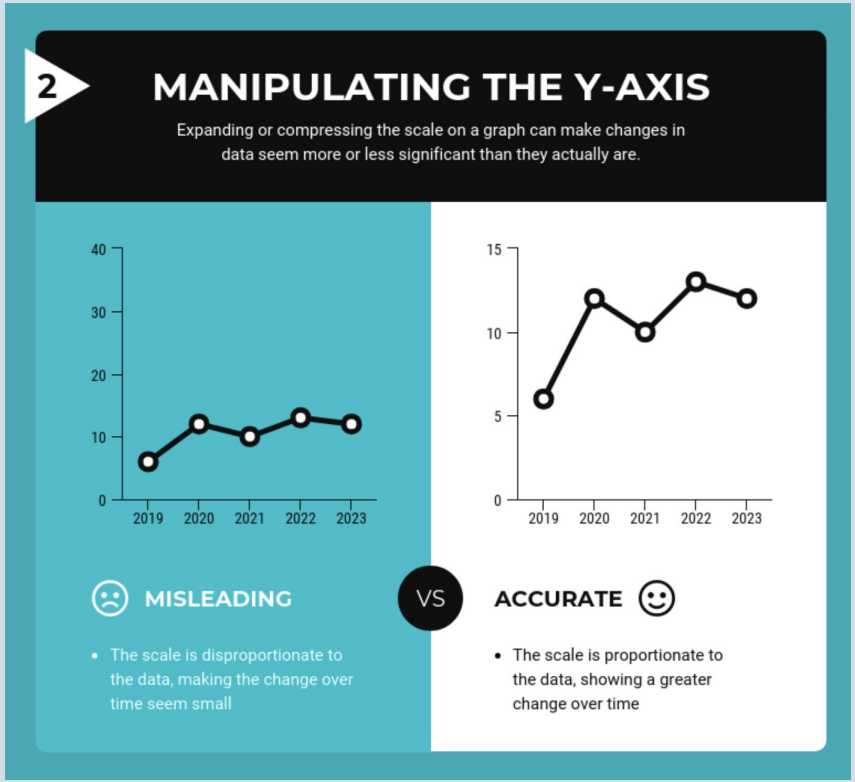


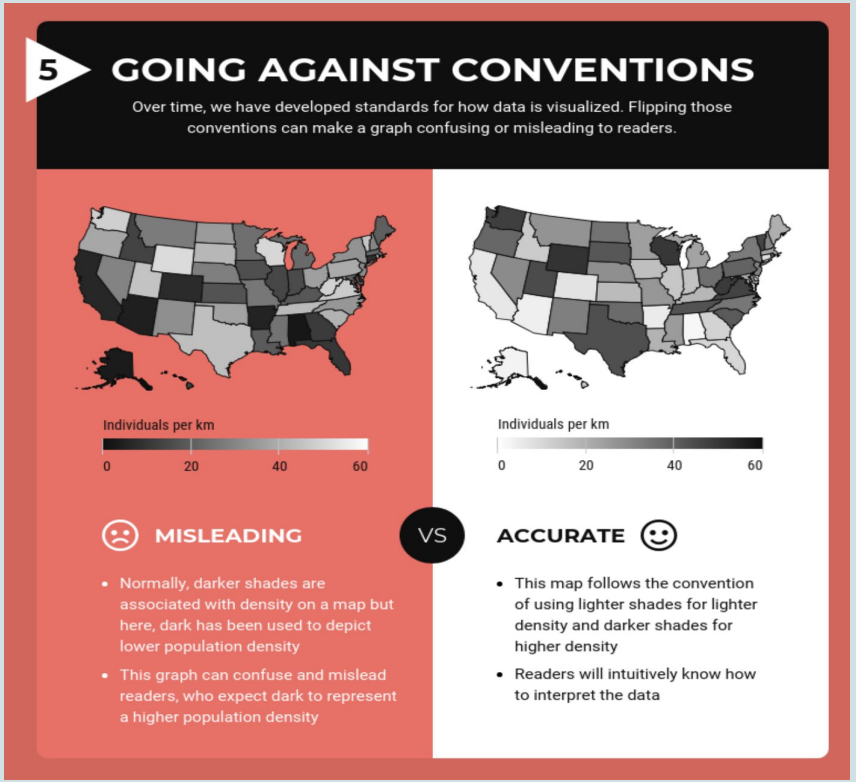
Discussion:

- 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 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!

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/cain-dataethics-sp22**

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

Northeastern University
NULab for Texts, Maps, and Networks

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