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

ENGW 1111 First Year Writing
Professor Emily Avery-Miller
Fall 2022

Taught By: Chris McNulty & Javier Rosario

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

Slides available at:

https://bit.ly/fa22-avery-miller-data-ethics

Northeastern University
NULab for Texts, Maps, and Networks

What is "Big Data"?

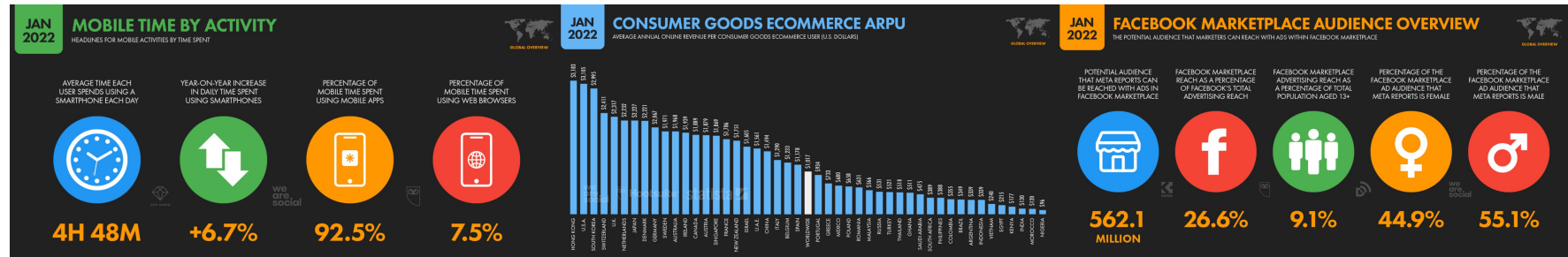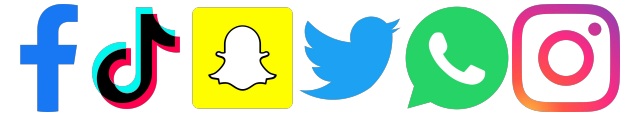*Feel free to ask questions at any point during the presentation!*

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% 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.

JAN 2022	MOBILE TIME BY ACTIVITY			
	AVERAGE TIME EACH USER SPENDS USING A SMARTPHONE EACH DAY	YEAR-ON-YEAR INCREASE IN DAILY TIME SPENT USING SMARTPHONES	PERCENTAGE OF MOBILE TIME SPENT USING MOBILE APPS	PERCENTAGE OF MOBILE TIME SPENT USING WEB BROWSERS
	4H 48M	+6.7%	92.5%	7.5%

CONSUMER GOODS ECOMMERCE ARPU — JAN 2022

JAN 2022	FACEBOOK MARKETPLACE AUDIENCE OVERVIEW			
	POTENTIAL AUDIENCE THAT META REPORTS CAN BE REACHED WITH ADS IN FACEBOOK MARKETPLACE	FACEBOOK MARKETPLACE REACH AS A PERCENTAGE OF FACEBOOK'S TOTAL ADVERTISING REACH	FACEBOOK MARKETPLACE ADVERTISING REACH AS A PERCENTAGE OF TOTAL POPULATION AGED 13+	PERCENTAGE OF THE FACEBOOK MARKETPLACE AD AUDIENCE THAT META REPORTS IS FEMALE / MALE
	562.1 MILLION	26.6%	9.1%	44.9% / 55.1%

Source: DataReportal

Northeastern University
NULab for Texts, Maps, and Networks

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

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.

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?**

Northeastern University
NULab for Texts, Maps, and Networks

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!

How does Big Data impact our daily lives?

Image and Audio Information 🎵 **TikTok**

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 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:
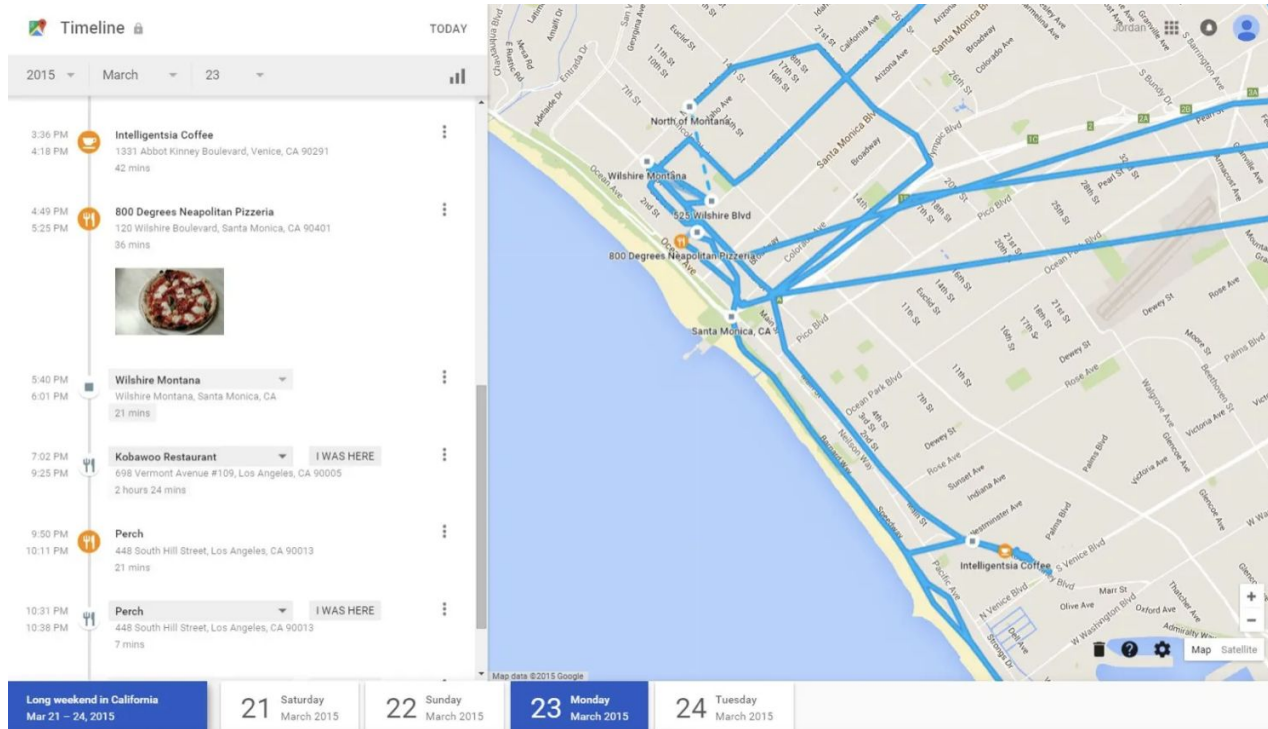
https://www.google.com/maps/timeline

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

Example of Google Maps' Timeline:

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

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.

Northeastern University
NULab for Texts, Maps, and Networks

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/

Northeastern University
NULab for Texts, Maps, and Networks

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

Issues in Big Data: Ethics and Algorithmic Bias

Northeastern University
NULab for Texts, Maps, and Networks

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

"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** systemic, political, and cultural **biases**.

They are **affected by input data**, the way that data is presented, how the data is interpreted by machines, and more.

This means **we also have the ability to challenge these biases**, norms, and forms of discrimination.

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

Algorithms

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

Input → Set of rules to obtain the expected output from the given input → Output
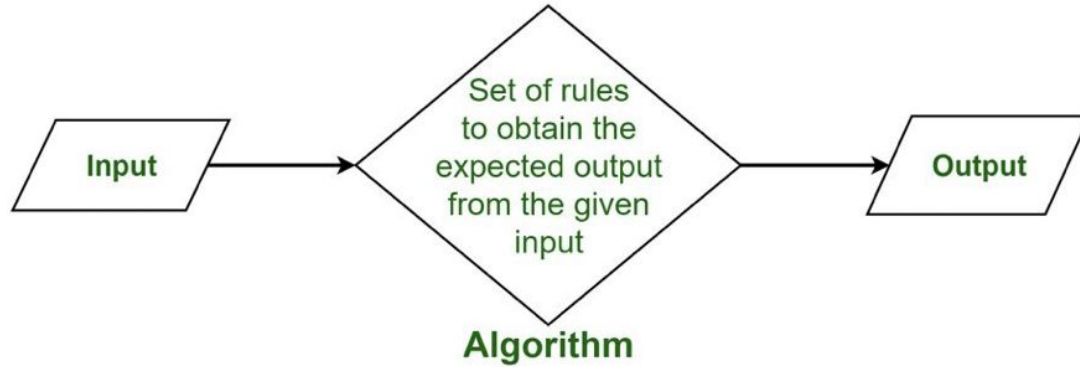
Algorithm

- "**Machine learning**" happens when an algorithm tells a computer to make decisions based on a set of patterns derived from data, instead of following specific predetermined instructions.

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

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

Northeastern University
NULab for Texts, Maps, and Networks

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

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.

Northeastern University
NULab for Texts, Maps, and Networks

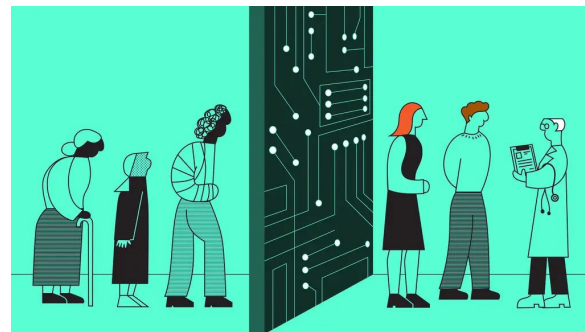*Feel free to ask questions at any point during the presentation!*

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."

5 White applicants denied

7 Latino applicants denied

7 Asian/Pacific Islander applicants denied

8 Native American applicants denied

9 Black applicants denied

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

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

Alleviating Algorithmic 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

Northeastern University
NULab for Texts, Maps, and Networks

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!

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?

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.

Northeastern University
NULab for Texts, Maps, and Networks

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!

Moving Forward -

How can we be cognizant of 'big data,' algorithms, and structural silences in our research?

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

Questions Researchers Must Ask

- What **information** is being collected and from where? To whom does this data **belong**?

- How is it being **collected**? Do **participants** know that it is collected, how it will be collected, and how will it be used?

- **How** will the data be analyzed? What **biases** and **ideologies** may be implicit in this analysis?

- Who will this research impact? Who will it **benefit**? Who will it potentially **harm**?

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

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

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

Northeastern University
NULab for Texts, Maps, and Networks

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

Be Mindful of Infographics and Data Visualizations
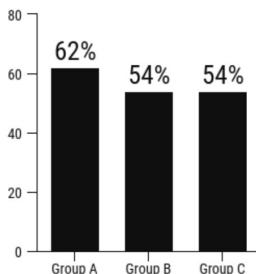

1 OMITTING THE BASELINE

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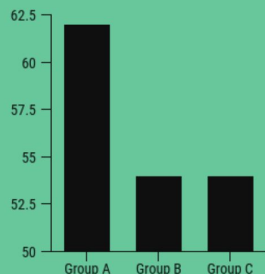
☹ **MISLEADING** VS **ACCURATE** ☺

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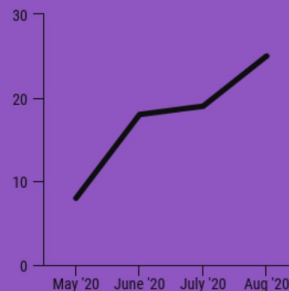- Group A looks much larger than Groups B and C

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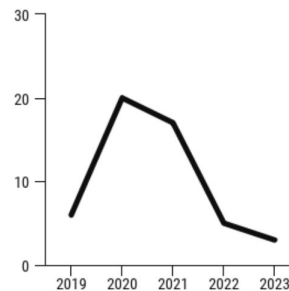

3 CHERRY PICKING DATA

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

☹ **MISLEADING** VS **ACCURATE** ☺

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

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

Northeastern University
NULab for Texts, Maps, and Networks

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

Finding and Using Non-Traditional Sources

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

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

Northeastern University
NULab for Texts, Maps, and Networks

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)

Northeastern University
NULab for Texts, Maps, and Networks

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

Thank you!

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

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

Slides & handout available at:

https://bit.ly/fa22-avery-miller-data-ethics

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

Northeastern University
NULab for Texts, Maps, and Networks

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