# Communicating with Data – Understanding Data

Dr. Ab Mosca (they/them)

# Plan for today

- Recap/Last bits from last class
- Power structures in data science
- Importance of context & documentation

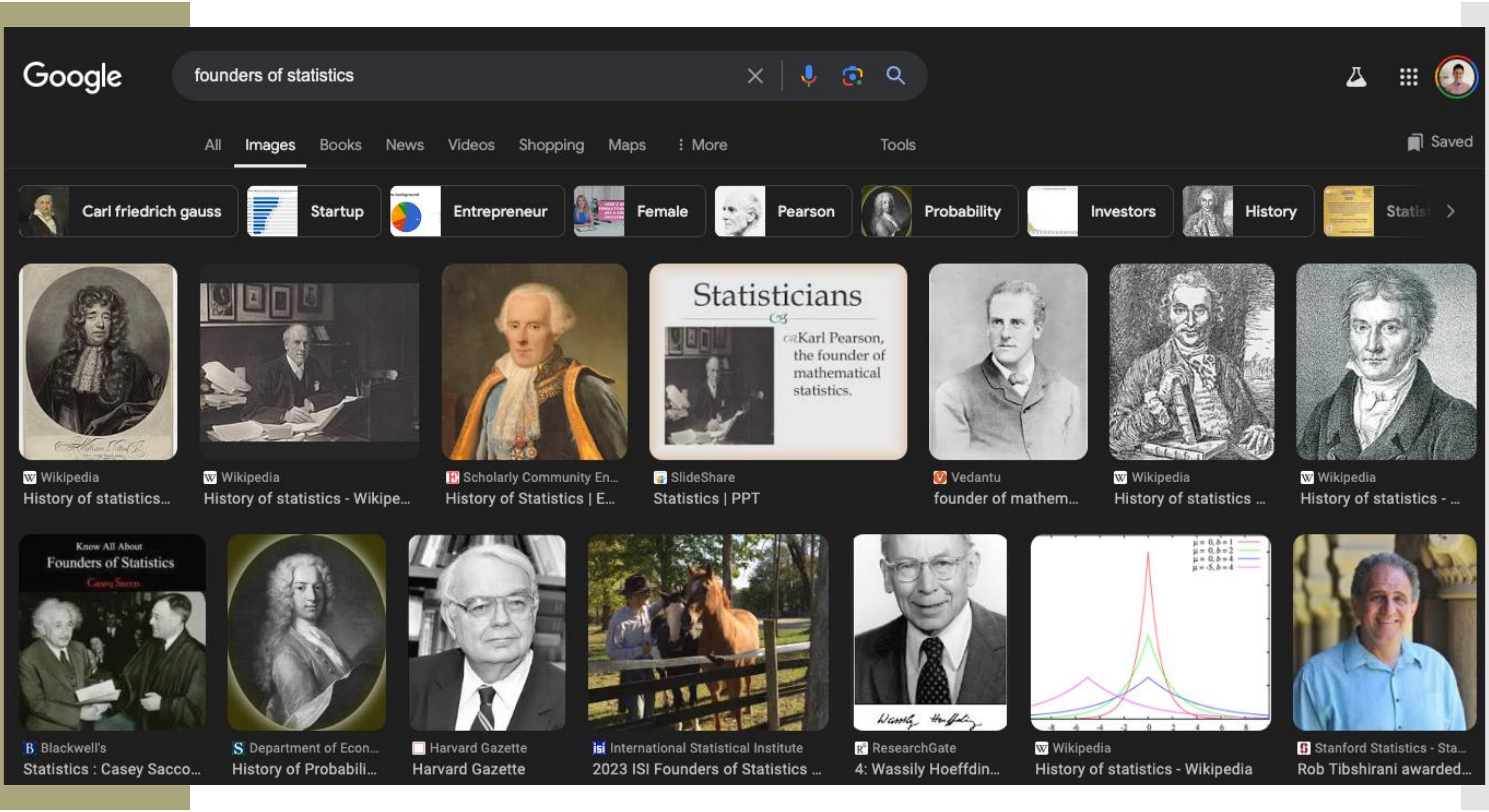
# Discussion

What is the origin of data science?

Where does it come from?

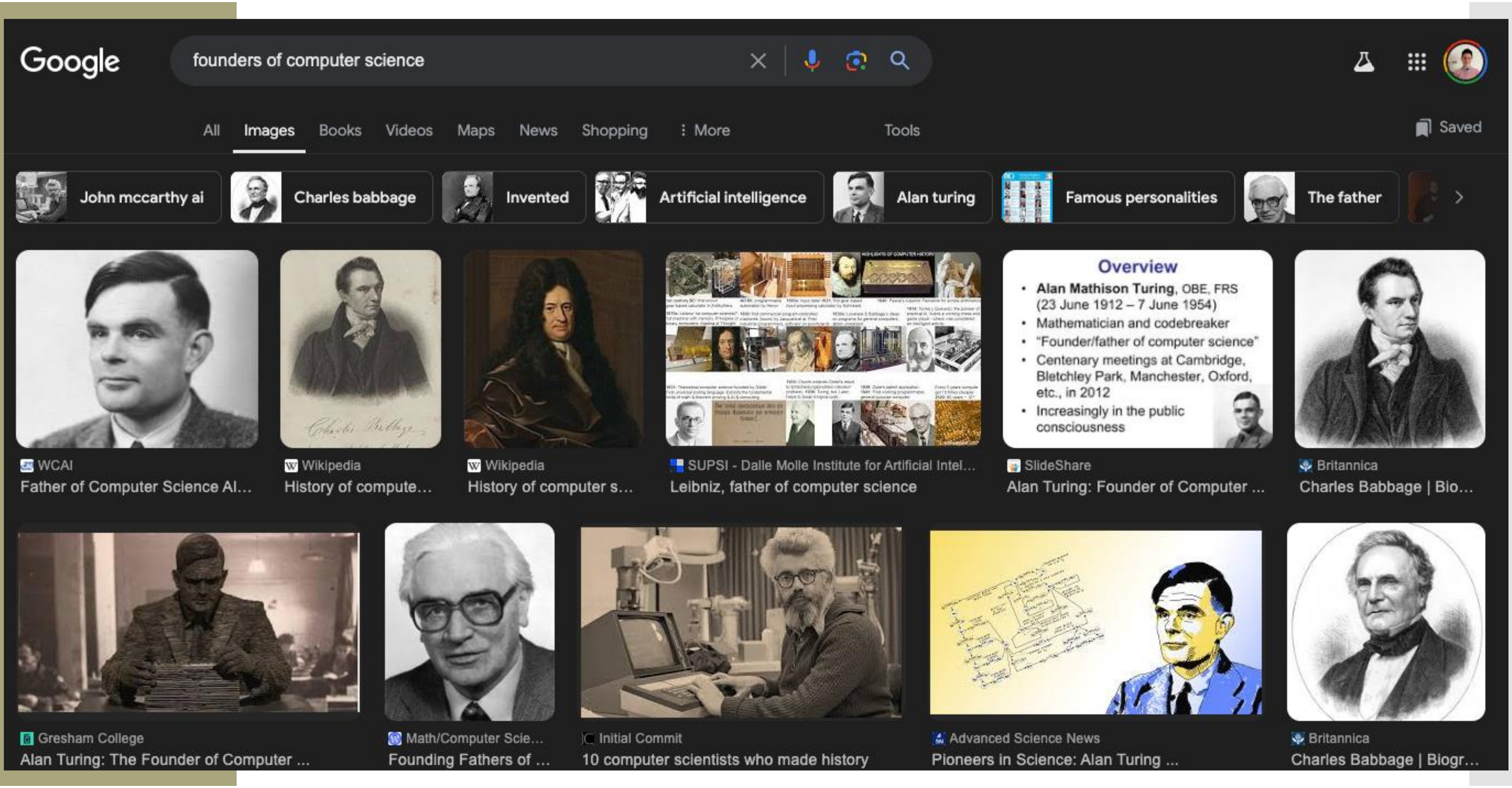
When did it start?

Who started it?



founders of statistics

https://data-feminism.mitpress.mit.edu/



founders of computer science

WCAI
Father of Computer Science AI...

Wikipedia
History of compute...

Wikipedia
History of computer s...

SUPSI - Dalle Molle Institute for Artificial Intel...
Leibniz, father of computer science

Overview
- **Alan Mathison Turing**, OBE, FRS (23 June 1912 – 7 June 1954)
- Mathematician and codebreaker
- "Founder/father of computer science"
- Centenary meetings at Cambridge, Bletchley Park, Manchester, Oxford, etc., in 2012
- Increasingly in the public consciousness

SlideShare
Alan Turing: Founder of Computer ...

Britannica
Charles Babbage | Bio...

Gresham College
Alan Turing: The Founder of Computer ...

Math/Computer Scie...
Founding Fathers of ...

Initial Commit
10 computer scientists who made history

Advanced Science News
Pioneers in Science: Alan Turing ...

Britannica
Charles Babbage | Biogr...

https://data-feminism.mitpress.mit.edu/

# Data Science & Power

- Hmmm….. Seems like data science is born from white cis males
    - Fun fact: this is not actually true if you dig deeper! But that’s for another class

https://data-feminism.mitpress.mit.edu/

# Data Science & Power

## Data Capitalism (Meyers West 2019)

- History is full of examples of data being used to control
- West argues, data as a commodity "enables an asymmetric redistribution of power that is weighted toward the actors who have access and the capability to make sense of information."

# Data Science & Power

## Data as Power

- In South American Andean cultures, Khipus are elaborate assemblages of knotted string used for millennia to record extracted numerical data such as tax records and military obligations of the populace (Medrano & Urton, 2018).
- From 2500BC the ancient Egyptian cultures were creating census datasets in order to determine how much labor force could be conscripted into the construction of pyramids for their pharaohs (Census-Taking in the Ancient World, 2016)

Bhargava, R. (2023). Teaching Data That Matters: History and Practice

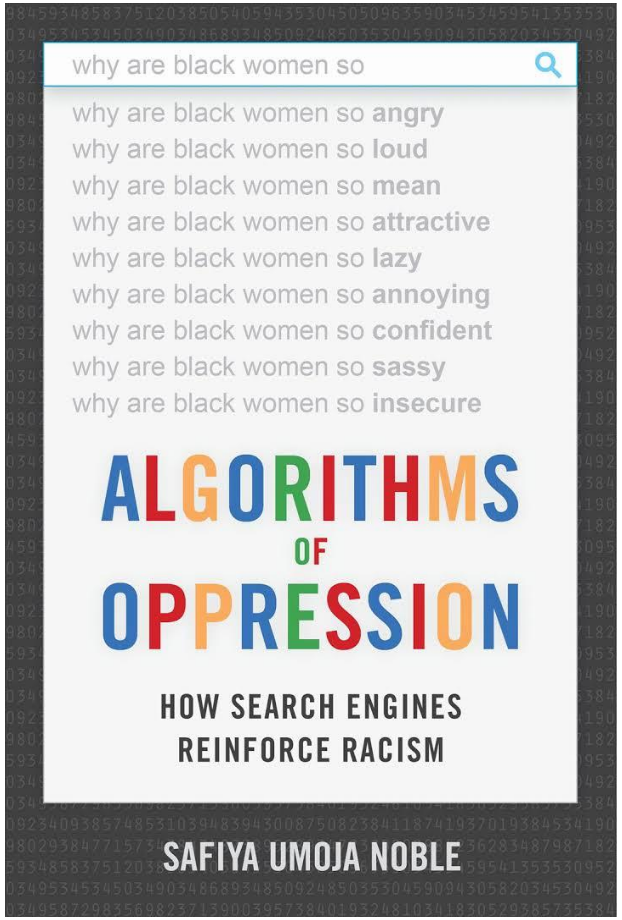
# Data Science & Power

- The “Big Data” revolution argues that with enough data we can make unbiased decisions
- However, data science:
  - Lacks transparency
  - Employs extractive collection
  - Leverages technological complexity
  - Controls impact

Bhargava, R. (2023). Teaching Data That Matters: History and Practice

# Example: Search Engines

Dr. Safiya Noble



why are black women so

why are black women so **angry**
why are black women so **loud**
why are black women so **mean**
why are black women so **attractive**
why are black women so **lazy**
why are black women so **annoying**
why are black women so **confident**
why are black women so **sassy**
why are black women so **insecure**

ALGORITHMS OF OPPRESSION

HOW SEARCH ENGINES REINFORCE RACISM

SAFIYA UMOJA NOBLE

https://youtu.be/iRVZozEEWlE?si=qzRtPmQzxl9KDxR2

# Example: Search Engines

- Search engine algorithms are largely based on:
  - Profit
  - Historical data
  - Predictive analytics

What are downstream real-life impacts of this data bias?

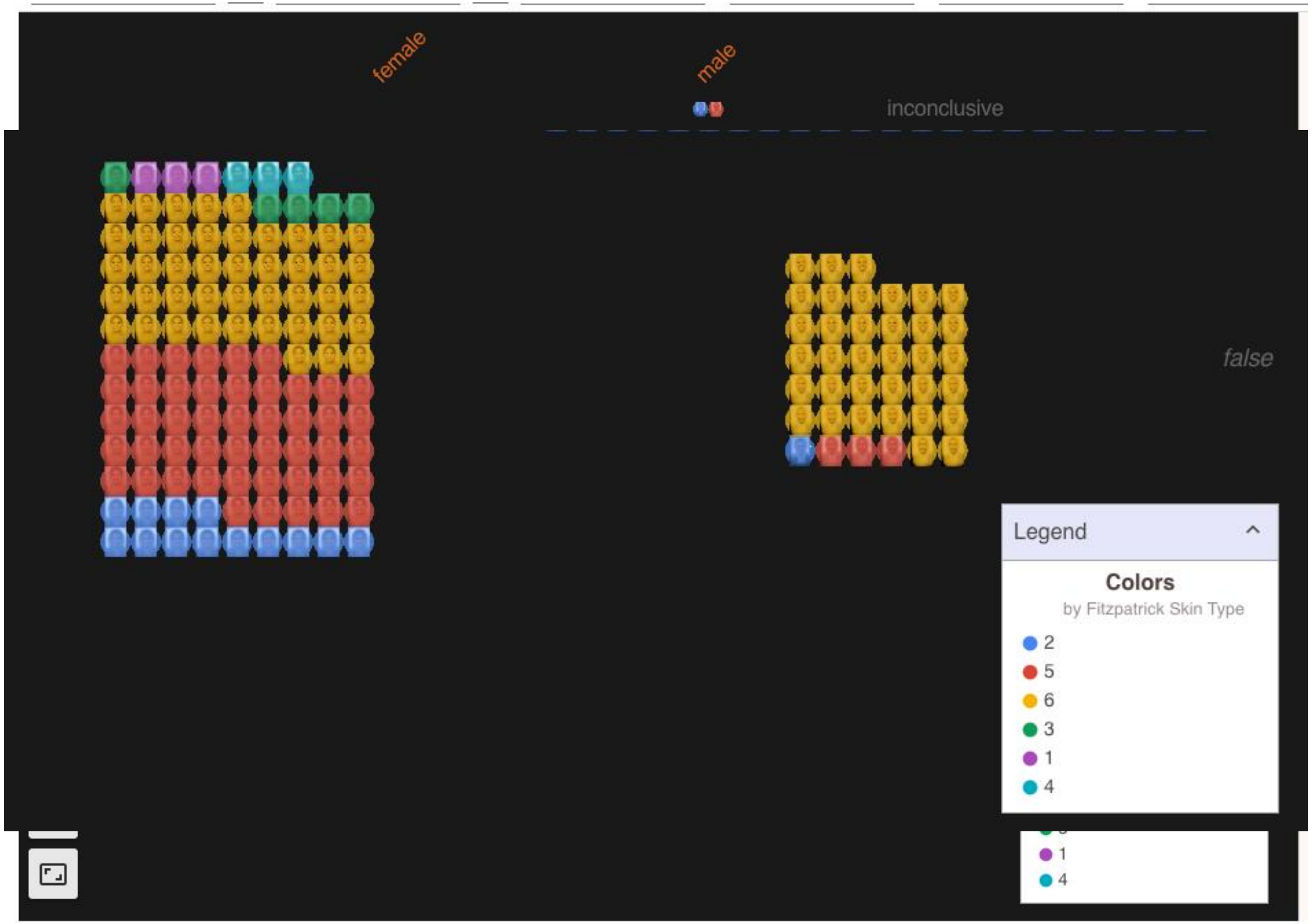
# Example: Facial Recognition

Dr. Joy Buolamwini

https://youtu.be/UG_X_7g63rY?si=qDMmUX5VjpaJYURe

https://data-feminism.mitpress.mit.edu/

# Example: Facial Recognition



https://data-feminism.mitpress.mit.edu/

# Example: Facial Recognition

- Training dataset used for most facial recognition systems contains
  - 78% male faces
  - 84% white faces
  - Only 4% were women and dark-skinned

What are downstream real-life impacts of this data bias?

https://data-feminism.mitpress.mit.edu/

# Example: Research Funding

NIH NATIONAL CANCER INSTITUTE

- 2022 prostate cancer research funding: 280.5 million
- 2022 uterine cancer research funding: 15.0 million

How does this monetary breakdown perpetuate systems of power?

https://www.cancer.gov/about-nci/budget/fact-book/data/research-funding

https://data-feminism.mitpress.mit.edu/

# Example: Research Funding

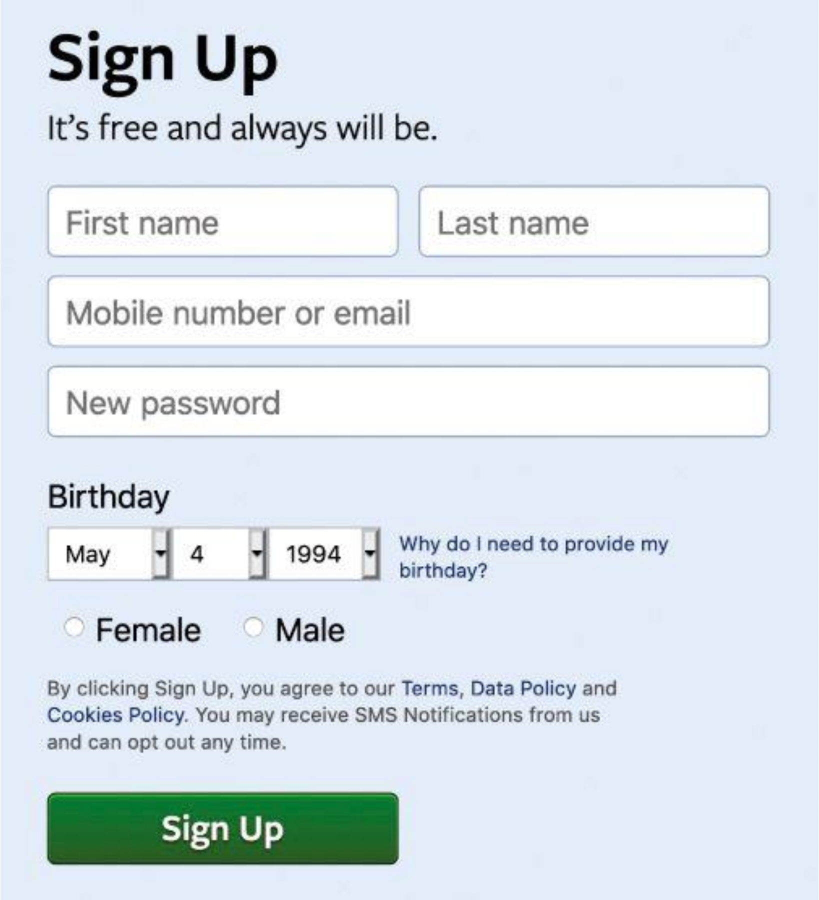
In U.S. dollars per 100 incident cases, **prostate cancer received an average of $1,821,000** per person-years of life lost, while **ovarian cancer received $97,000**, **cervical cancer $87,000**, and **uterine cancer $57,000**.

Ovarian and cervical cancers had lower average Funding to Lethality scores compared to nine other cancers, while uterine cancer was lower than 13 other cancers ($p<0.01$ for all comparisons).

Spencer et al. Disparities in the allocation of research funding to gynecologic cancers by Funding to Lethality scores

# "What gets counted counts"

- Data is often used to inform policy and allocate resources
- What is not counted in that data collection can become invisible
  - Ex. Expansive gender



## Sign Up

It's free and always will be.

First name

Last name

Mobile number or email

New password

Birthday

May 4 1994

Why do I need to provide my birthday?

○ Female ○ Male

By clicking Sign Up, you agree to our Terms, Data Policy and Cookies Policy. You may receive SMS Notifications from us and can opt out any time.

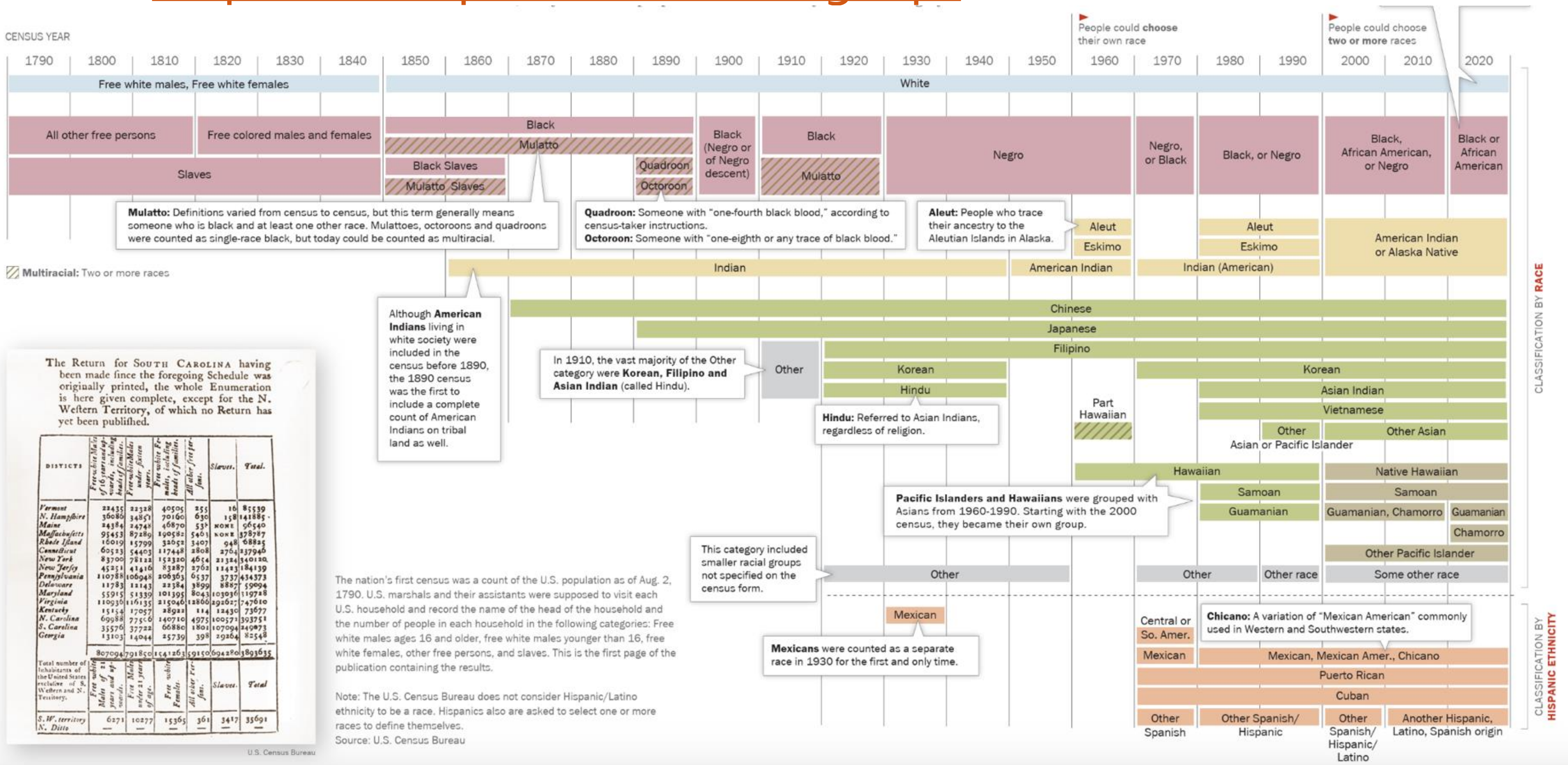
https://data-feminism.mitpress.mit.edu/

# “What gets counted counts”

- Data is often used to inform policy and allocate resources
- What is not counted in that data collection can become invisible
  - Ex. US Census

# “What gets counted counts”

- Data is often used to inform policy and allocate resources
- What *is* counted is considered important
  - Ex. US Census & Race
  - https://www.pewresearch.org/social-trends/feature/what-census-calls-us/
  - https://www.pewresearch.org/wp-



**Mulatto:** Definitions varied from census to census, but this term generally means someone who is black and at least one other race. Mulattoes, octoroons and quadroons were counted as single-race black, but today could be counted as multiracial.

**Quadroon:** Someone with “one-fourth black blood,” according to census-taker instructions.
**Octoroon:** Someone with “one-eighth or any trace of black blood.”

**Aleut:** People who trace their ancestry to the Aleutian Islands in Alaska.

Although **American Indians** living in white society were included in the census before 1890, the 1890 census was the first to include a complete count of American Indians on tribal land as well.

In 1910, the vast majority of the Other category were **Korean, Filipino and Asian Indian** (called Hindu).

**Hindu:** Referred to Asian Indians, regardless of religion.

**Pacific Islanders and Hawaiians** were grouped with Asians from 1960-1990. Starting with the 2000 census, they became their own group.

This category included smaller racial groups not specified on the census form.

**Mexicans** were counted as a separate race in 1930 for the first and only time.

**Chicano:** A variation of “Mexican American” commonly used in Western and Southwestern states.

The Return for South Carolina having been made ſince the foregoing Schedule was originally printed, the whole Enumeration is here given complete, except for the N. Weſtern Territory, of which no Return has yet been publiſhed.

| DISTICTS | Free white Males of 16 years and upwards, including heads of families. | Free white Males under ſixteen years. | Free white Females, including heads of families. | All other free perſons. | Slaves. | Total. |
|---|---|---|---|---|---|---|
| Vermont | 22435 | 22328 | 40505 | 255 | 16 | 85539 |
| N. Hampſhire | 36086 | 34851 | 70160 | 630 | 158 | 141885 |
| Maine | 24384 | 24748 | 46870 | 538 | NONE | 96540 |
| Maſſachuſetts | 95453 | 87289 | 190582 | 5463 | NONE | 378787 |
| Rhode Iſland | 16019 | 15799 | 32652 | 3407 | 948 | 68825 |
| Connecticut | 60523 | 54403 | 117448 | 2808 | 2764 | 237946 |
| New York | 83700 | 78122 | 152320 | 4654 | 21324 | 340120 |
| New Jerſey | 45251 | 41416 | 83287 | 2762 | 11423 | 184139 |
| Pennſylvania | 110788 | 106948 | 206363 | 6537 | 3737 | 434373 |
| Delaware | 11783 | 12143 | 22384 | 3899 | 8887 | 59094 |
| Maryland | 55915 | 51339 | 101395 | 8043 | 103036 | 319728 |
| Virginia | 110936 | 116135 | 215046 | 12866 | 292627 | 747610 |
| Kentucky | 15154 | 17057 | 28922 | 114 | 12430 | 73677 |
| N. Carolina | 69988 | 77506 | 140710 | 4975 | 100572 | 393751 |
| S. Carolina | 35576 | 37722 | 66880 | 1801 | 107094 | 249073 |
| Georgia | 13103 | 14044 | 25739 | 398 | 29264 | 82548 |
|  | 807094 | 791850 | 1541263 | 59150 | 694280 | 3893635 |

| Total number of Inhabitants of the United States excluſive of S. Weſtern and N. Territory. | Free white Males of 21 years and upwards. | Free Males under 21 years of age. | Free white Females. | All other Perſons. | Slaves. | Total |
|---|---|---|---|---|---|---|
| S. W. territory | 6271 | 10277 | 15365 | 361 | 3417 | 35691 |
| N. Ditto | — | — | — | — | — | — |

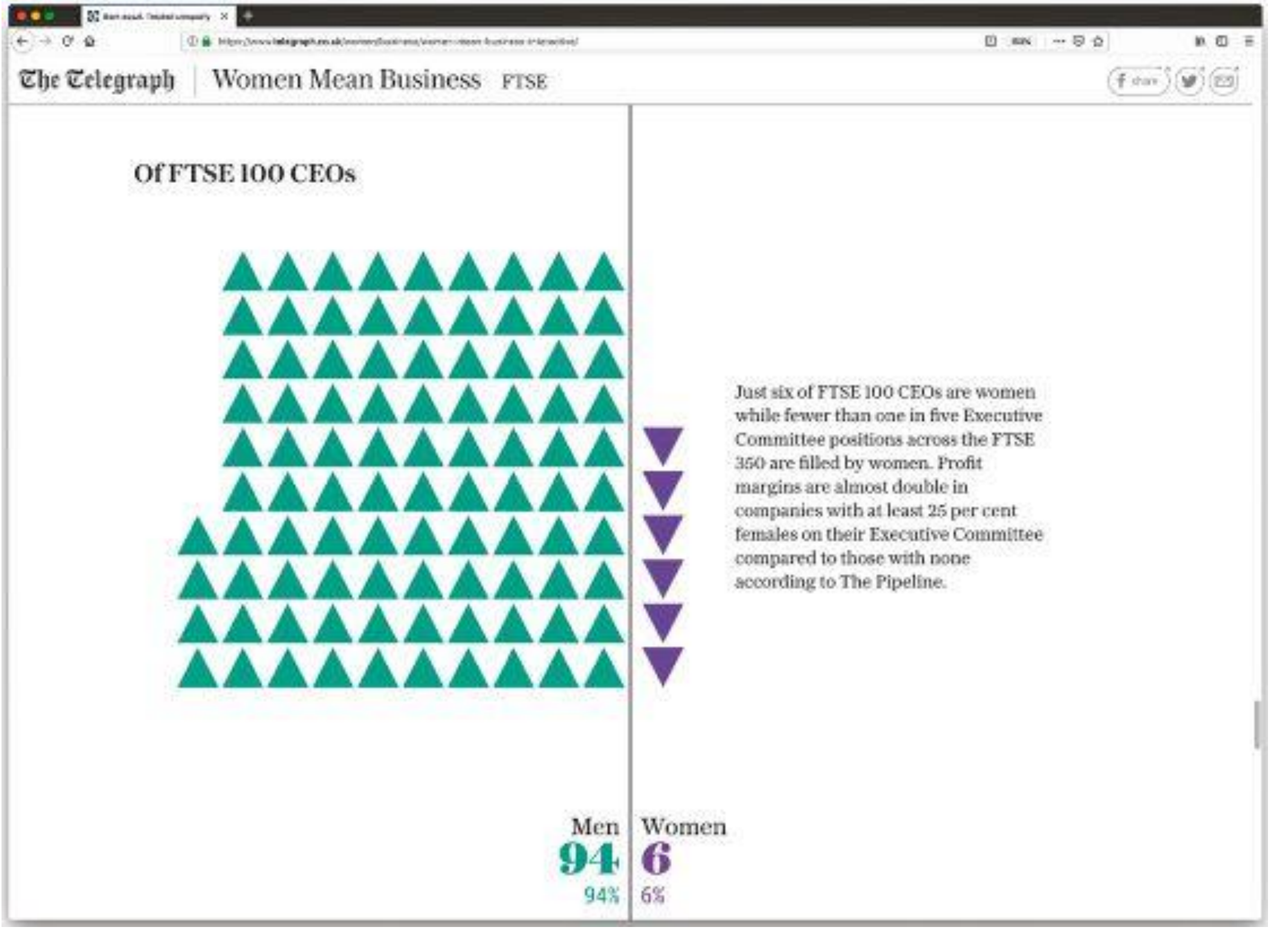
U.S. Census Bureau

The nation’s first census was a count of the U.S. population as of Aug. 2, 1790. U.S. marshals and their assistants were supposed to visit each U.S. household and record the name of the head of the household and the number of people in each household in the following categories: Free white males ages 16 and older, free white males younger than 16, free white females, other free persons, and slaves. This is the first page of the publication containing the results.

Note: The U.S. Census Bureau does not consider Hispanic/Latino ethnicity to be a race. Hispanics also are asked to select one or more races to define themselves.
Source: U.S. Census Bureau

# Subverting Power

- Subvert norms

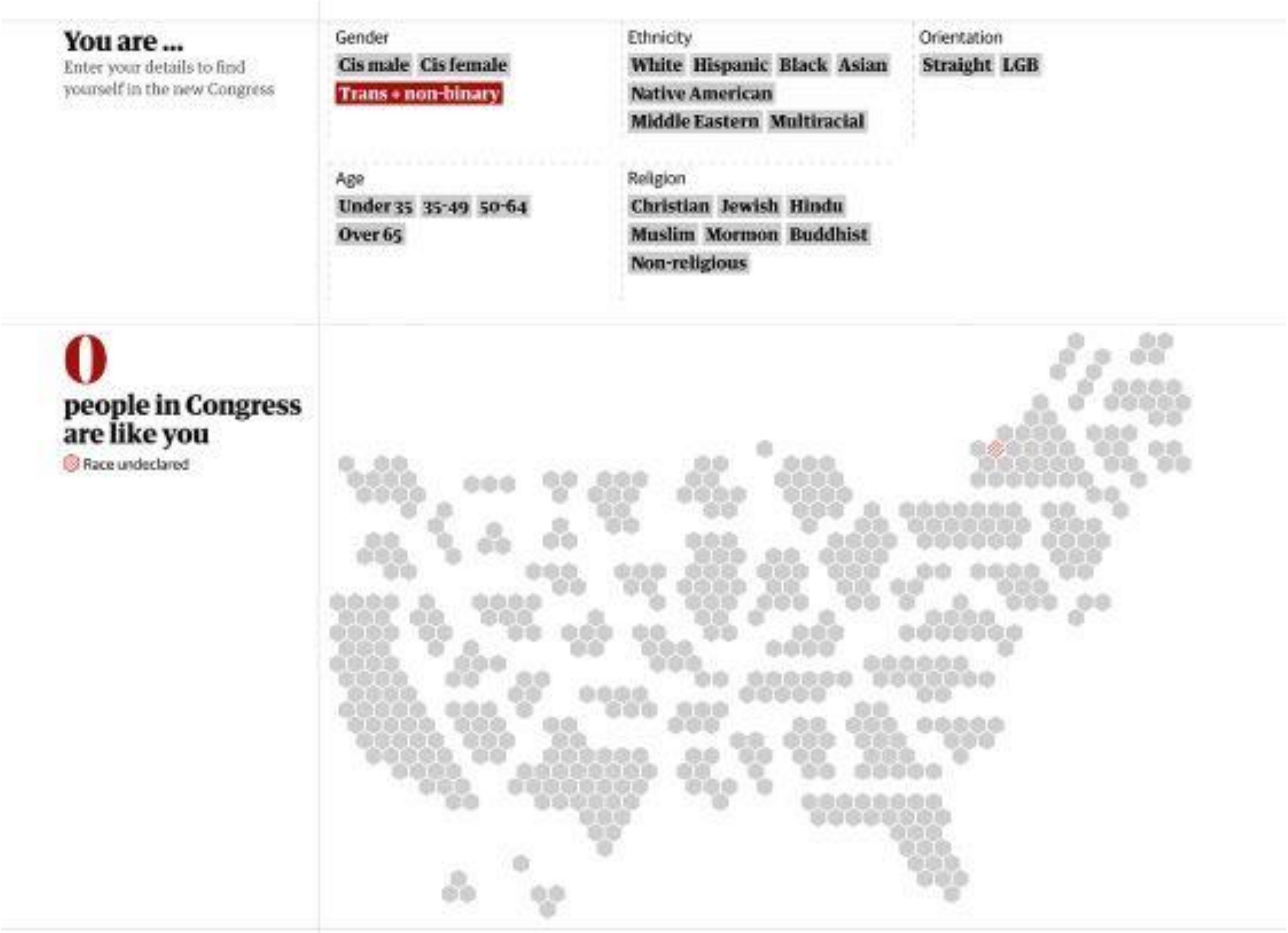


Just six of FTSE 100 CEOs are women while fewer than one in five Executive Committee positions across the FTSE 350 are filled by women. Profit margins are almost double in companies with at least 25 per cent females on their Executive Committee compared to those with none according to The Pipeline.

https://data-feminism.mitpress.mit.edu/

# Subverting Power

- Highlight missing categories



**You are ...**

Enter your details to find yourself in the new Congress

Gender

Cis male Cis female

Trans + non-binary

Ethnicity

White Hispanic Black Asian

Native American

Middle Eastern Multiracial

Orientation

Straight LGB

Age

Under 35 35-49 50-64

Over 65

Religion

Christian Jewish Hindu

Muslim Mormon Buddhist

Non-religious

**0**

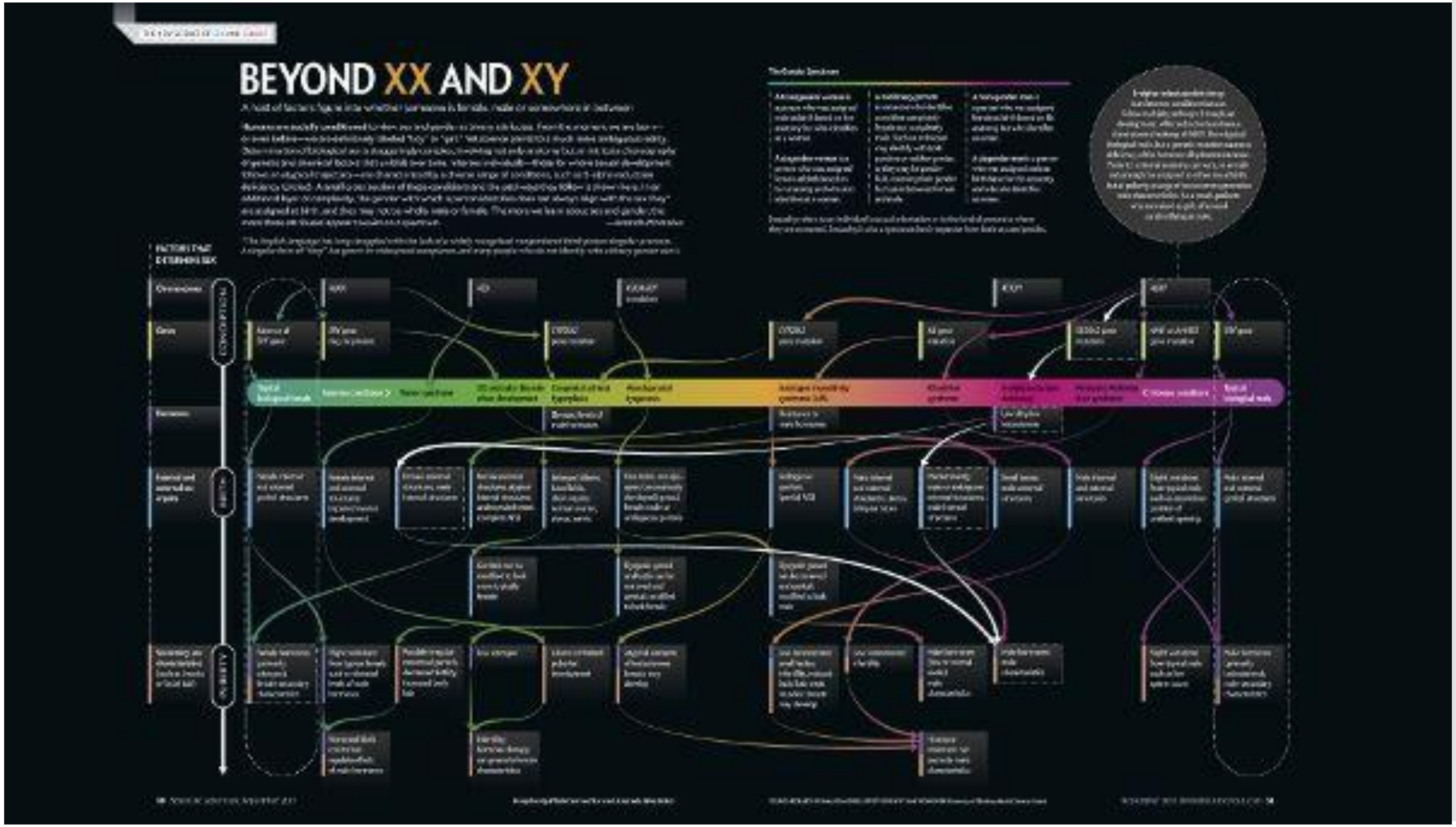
**people in Congress are like you**

Race undeclared

https://data-feminism.mitpress.mit.edu/

# Subverting Power

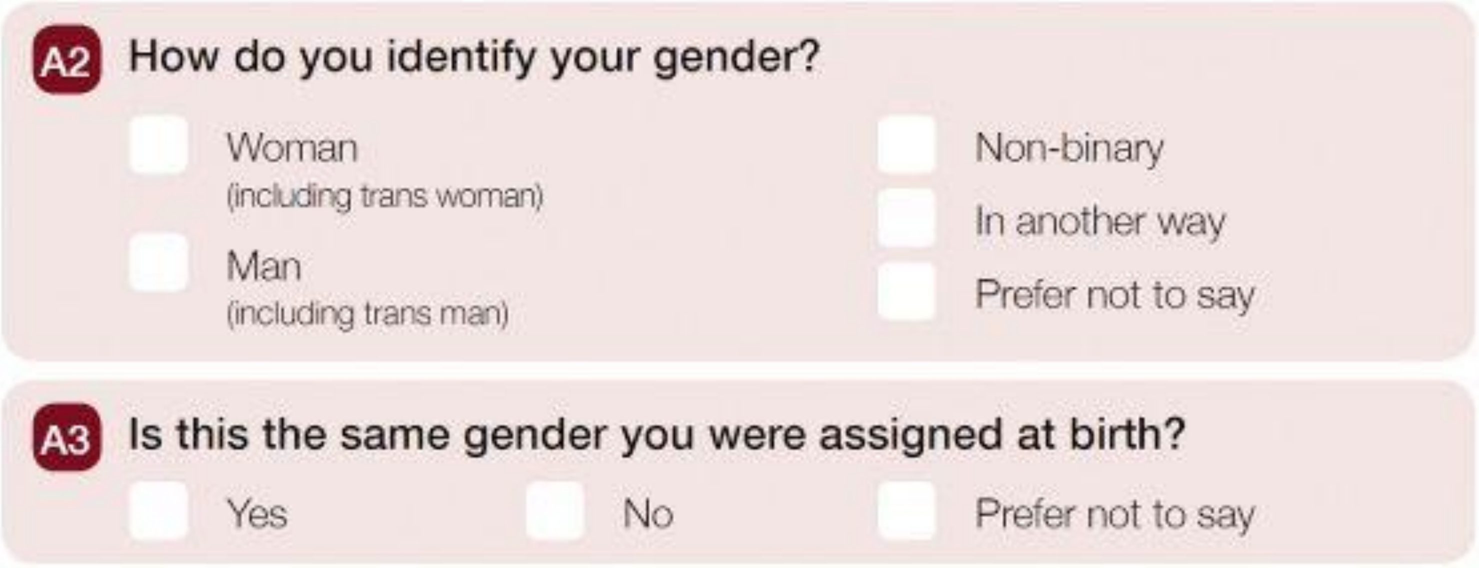
- Educate



https://data-feminism.mitpress.mit.edu/

# Subverting Power

- Rethink data collection



**A2** **How do you identify your gender?**

□ Woman (including trans woman)

□ Man (including trans man)

□ Non-binary

□ In another way

□ Prefer not to say

**A3** **Is this the same gender you were assigned at birth?**

□ Yes □ No □ Prefer not to say

https://data-feminism.mitpress.mit.edu/

# Subverting Power

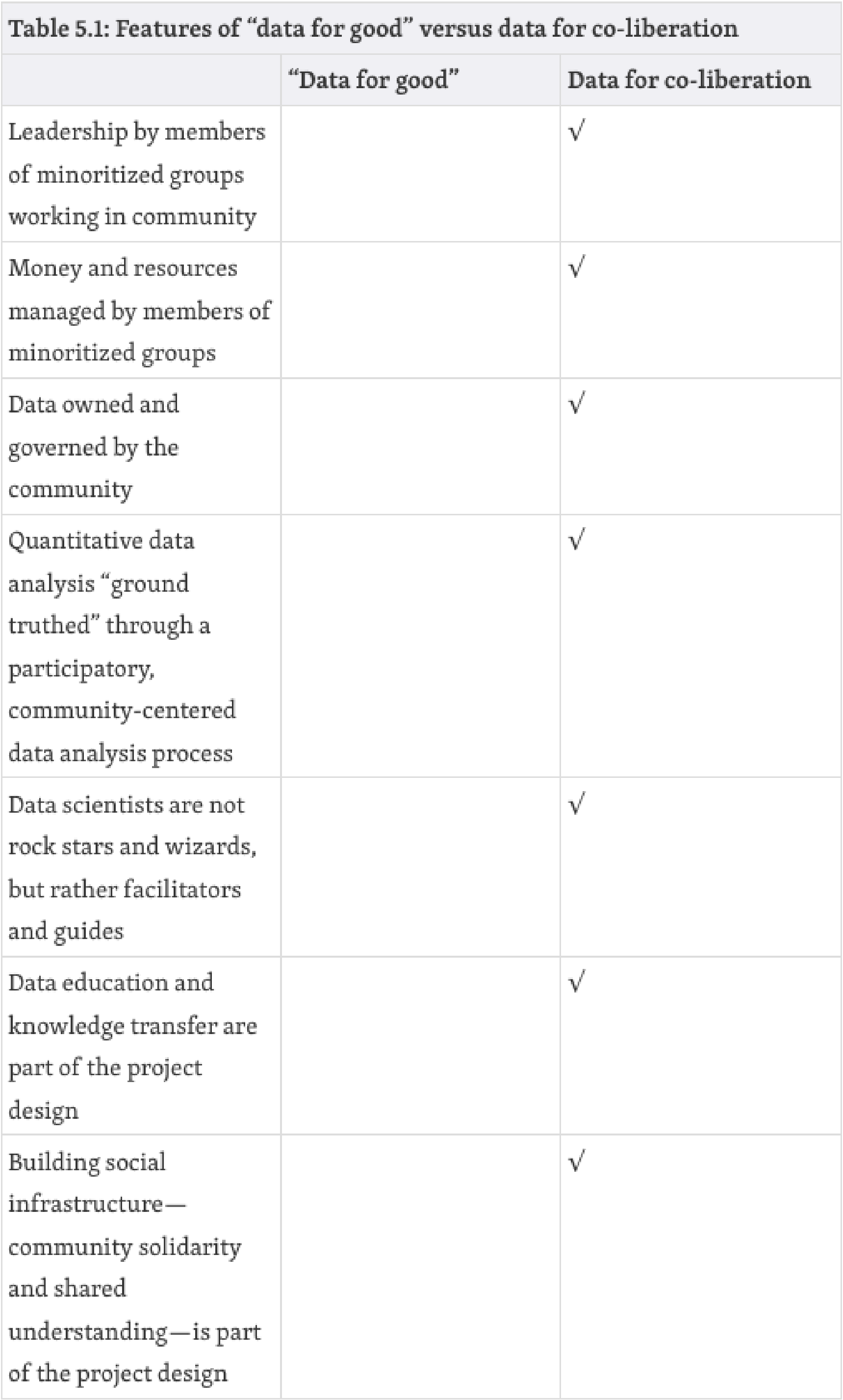


Table 5.1: Features of "data for good" versus data for co-liberation

| | "Data for good" | Data for co-liberation |
|---|---|---|
| Leadership by members of minoritized groups working in community | | √ |
| Money and resources managed by members of minoritized groups | | √ |
| Data owned and governed by the community | | √ |
| Quantitative data analysis "ground truthed" through a participatory, community-centered data analysis process | | √ |
| Data scientists are not rock stars and wizards, but rather facilitators and guides | | √ |
| Data education and knowledge transfer are part of the project design | | √ |
| Building social infrastructure—community solidarity and shared understanding—is part of the project design | | √ |

- Add transparency
- Avoid extractive approaches
- Follow the lead of the community

https://data-feminism.mitpress.mit.edu/

# Subverting Power

## Acknowledge context

RESPONSIBLE DATASETS IN CONTEXT

https://www.responsible-datasets-in-context.com/datasets.html

- What is the historical context of the data?
- Where did the data come from? Who collected it?
- Why was the data collected?
- How was the data collected?
- How is the data used?
- What’s in the data?
- What “counts” as a data point?
- What data is missing?
- How is uncertainty handled?

What biases or ethical issues could the answers to these questions reveal that would otherwise be hidden?

# Lab 1

- Go to the course website to find instructions for today’s lab
- Be prepared to share your findings!