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

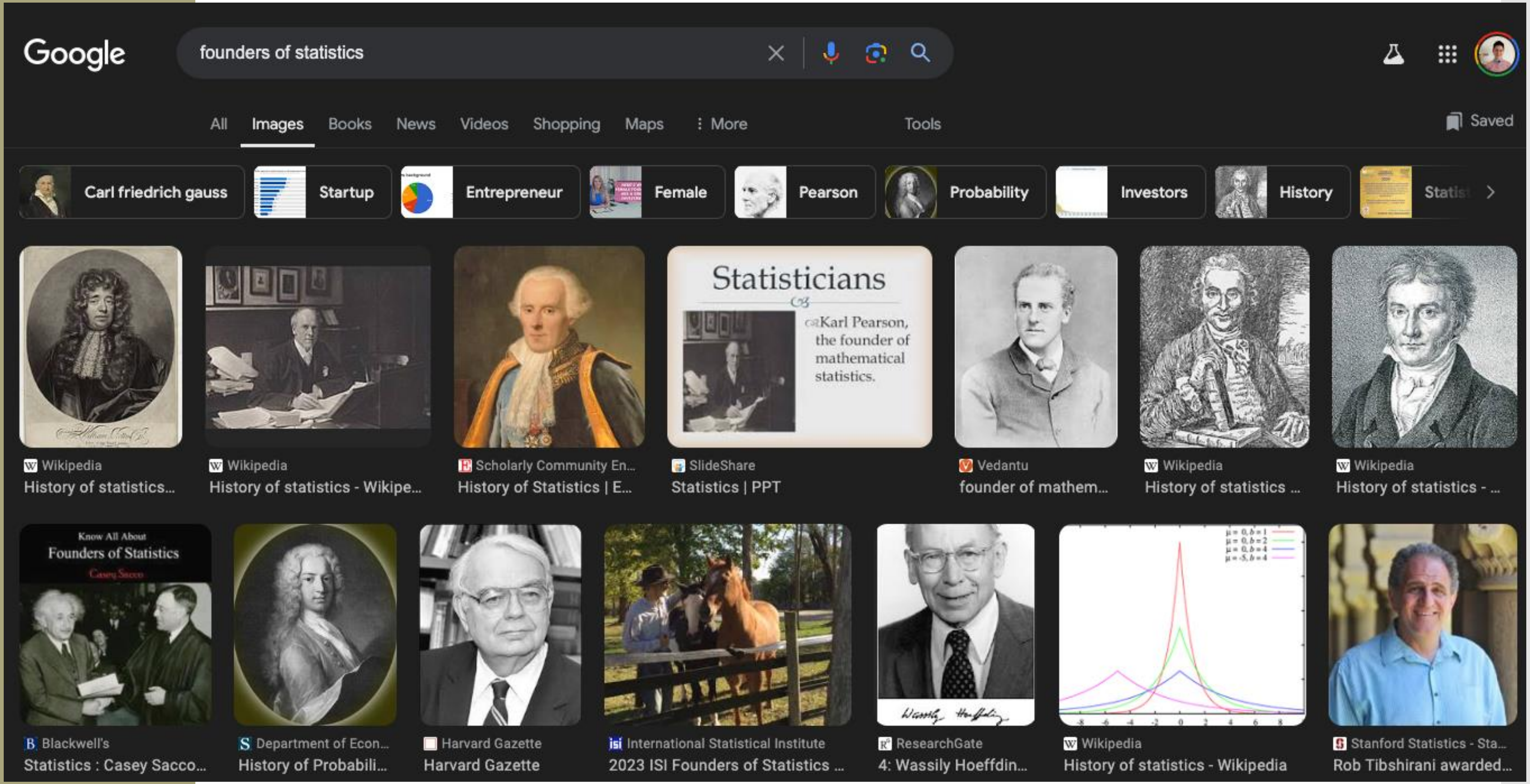
Discussion

What is the origin of data science?

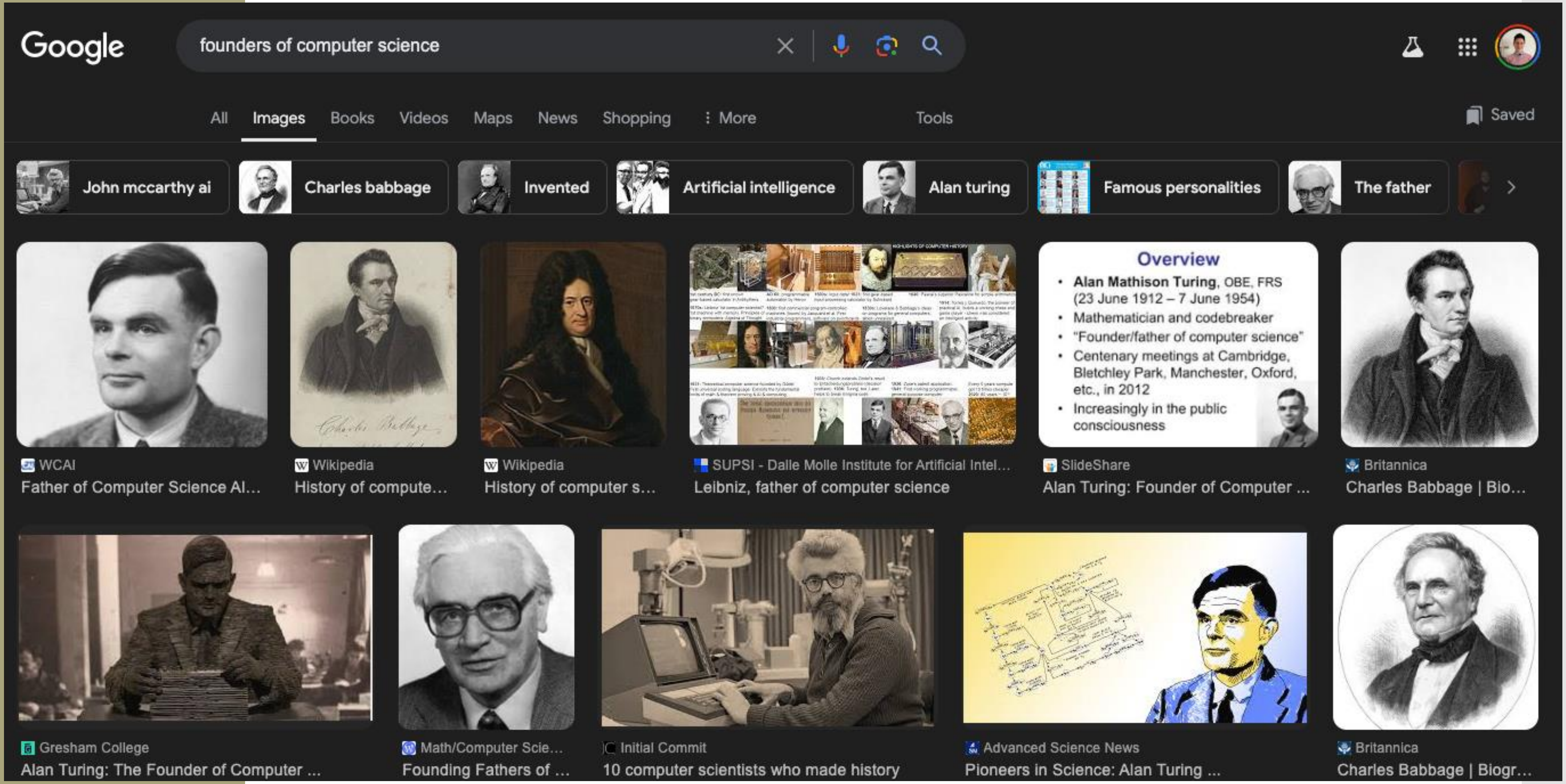
Where does it come from?

When did it start?

Who started it?



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



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

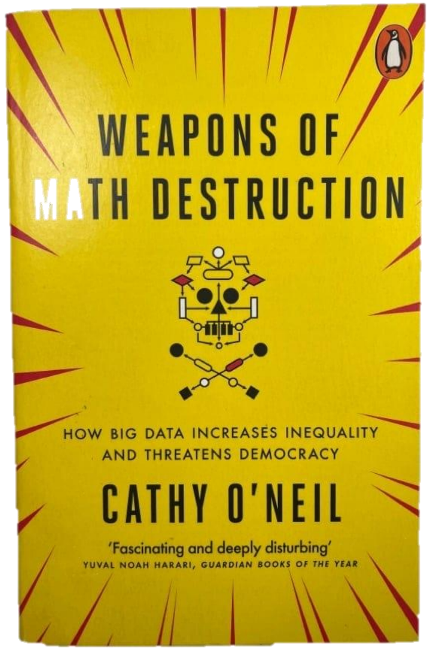
Data as Power

- 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

Algorithms as Power



Cathy O'Neil

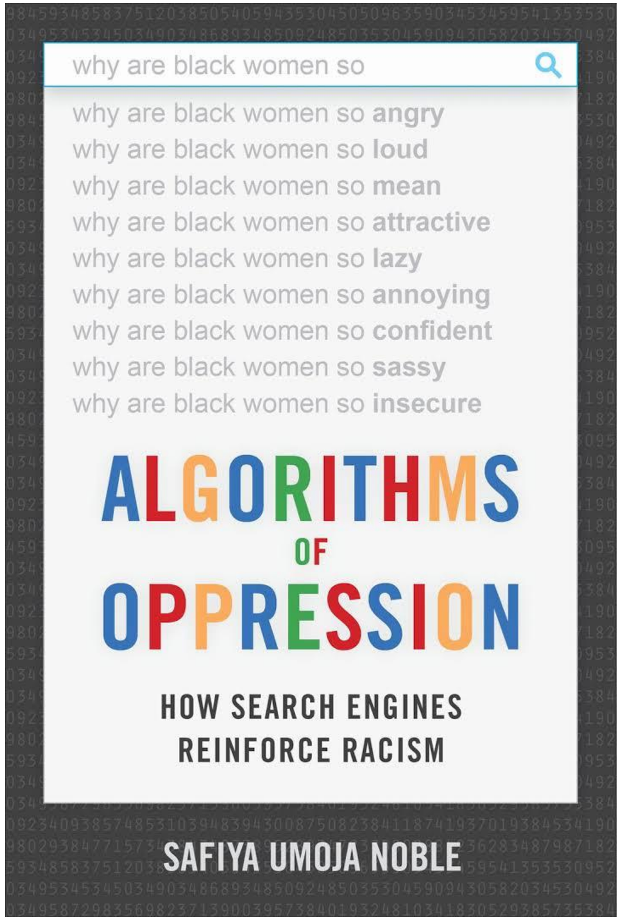
https://www.youtube.com/watch?v=heQzqX35c9A

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

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 search engine bias?

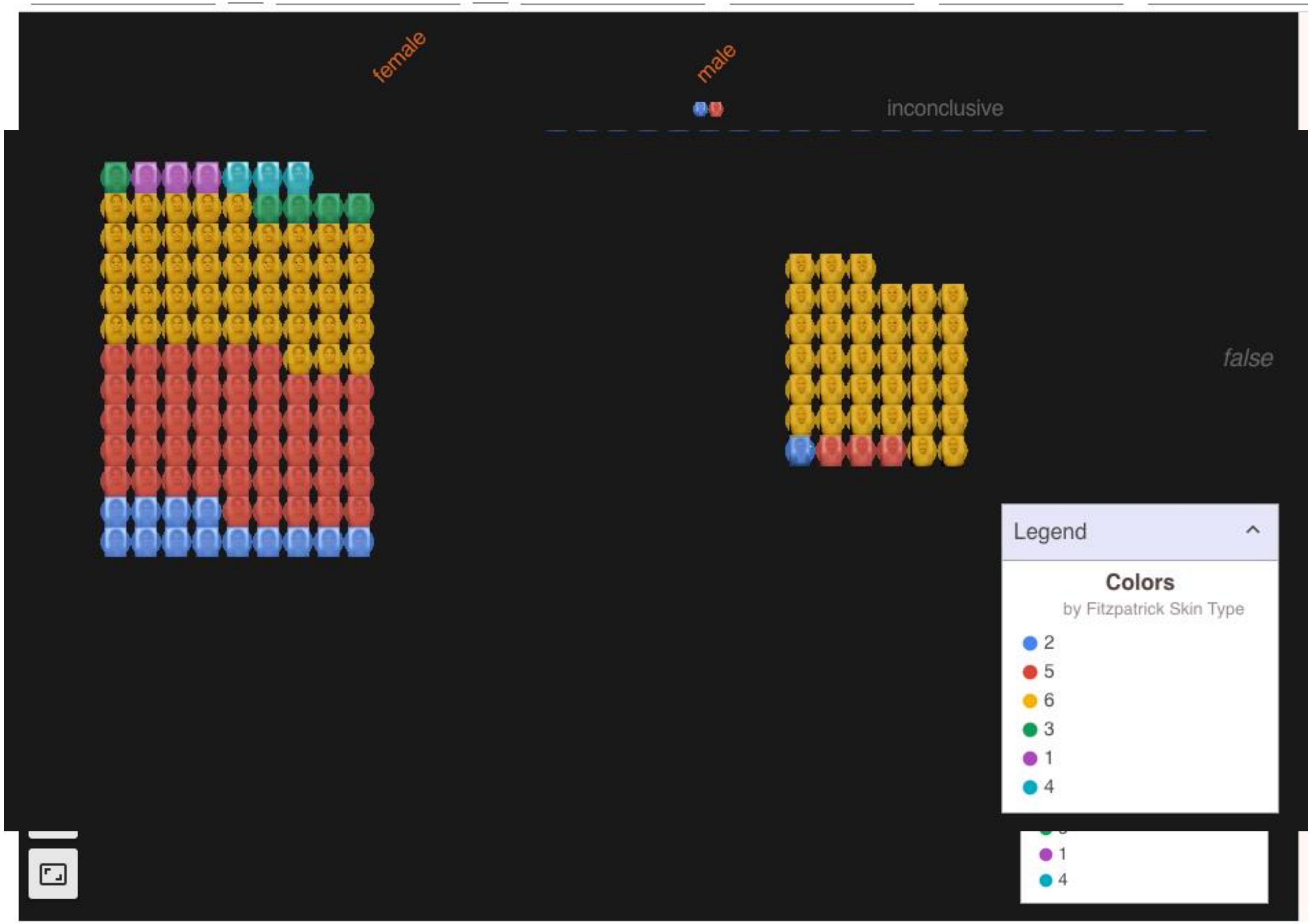
Example: Facial Recognition

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

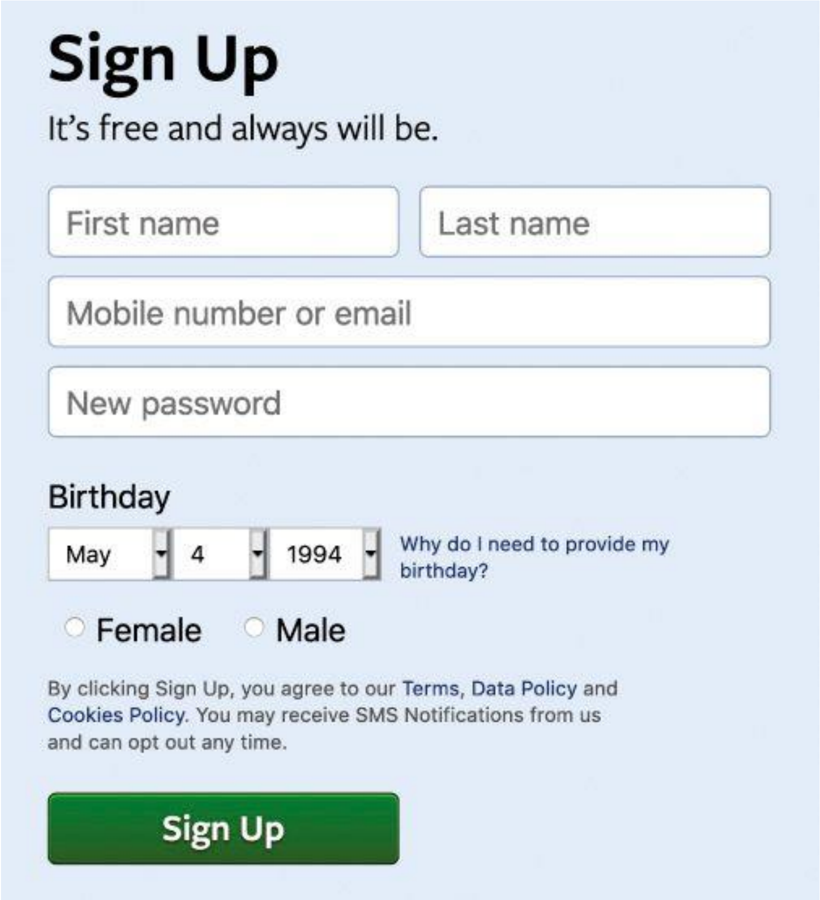
- Training dataset used for most facial recognition systems contains
  - 78% male faces
  - 84% white faces
  - Only 4% were women and dark-skinned

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

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. Expansive gender



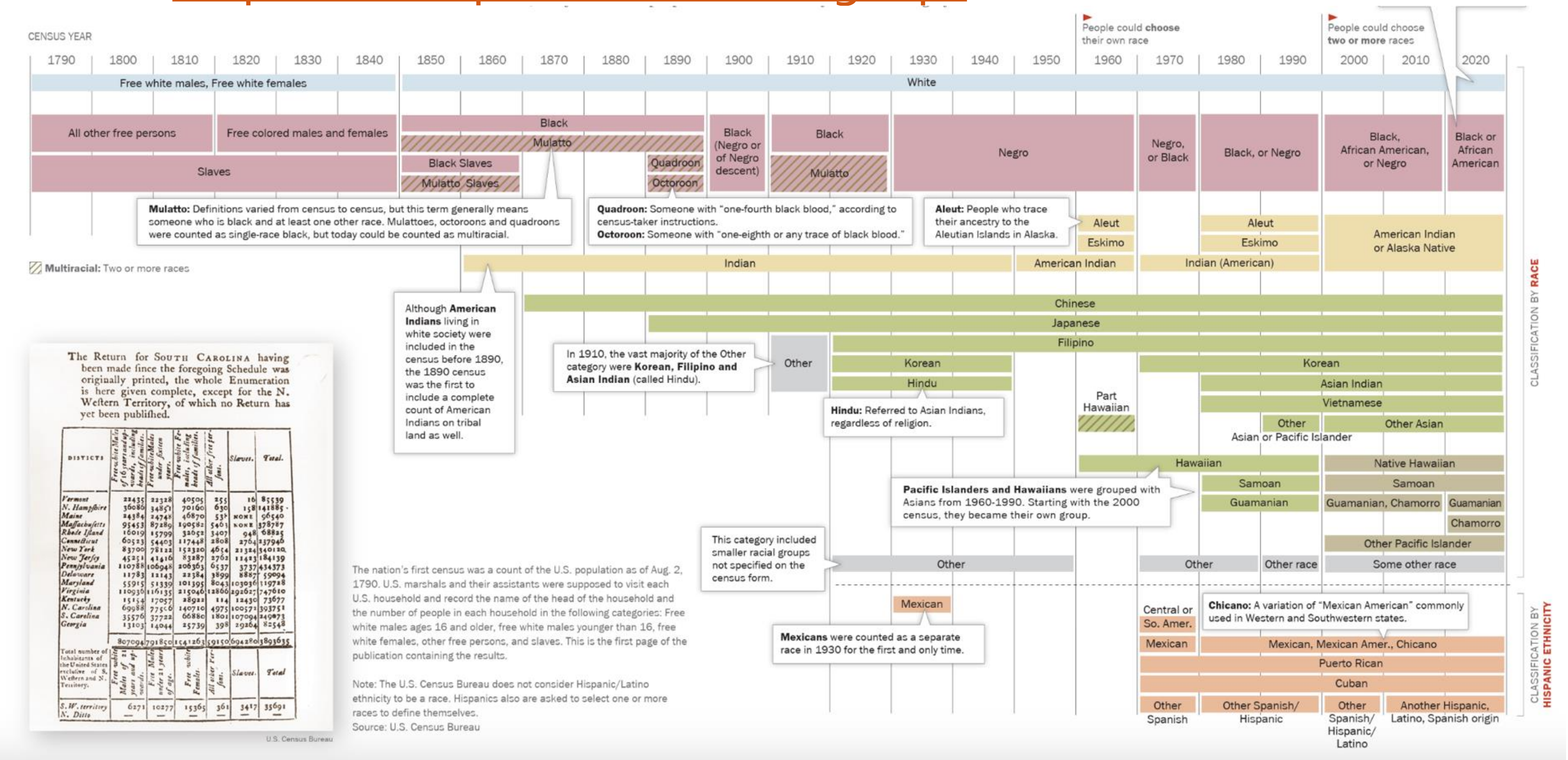
https://data-feminism.mitpress.mit.edu/

“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

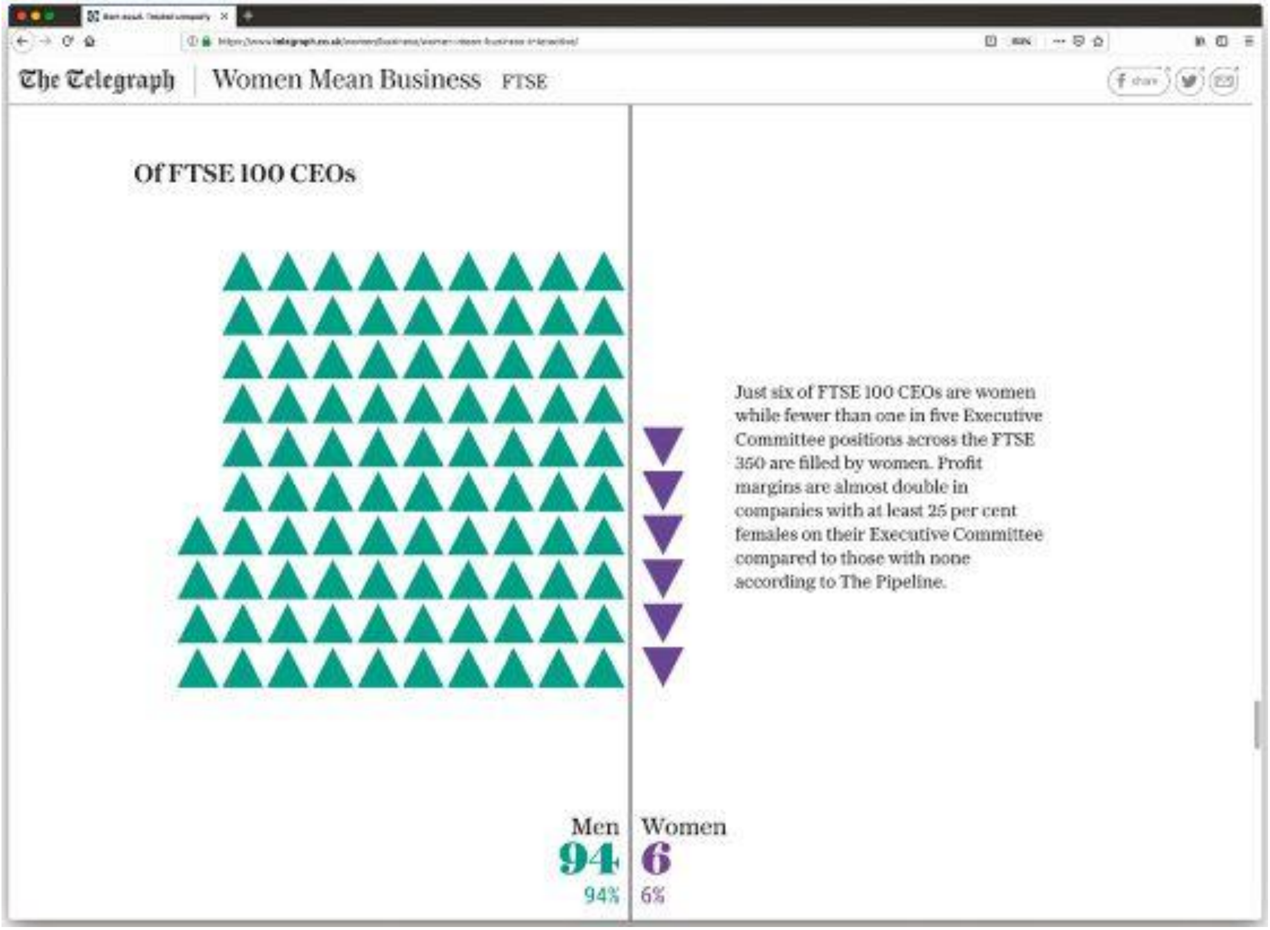
“What gets counted counts”

- 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-



Subverting Power

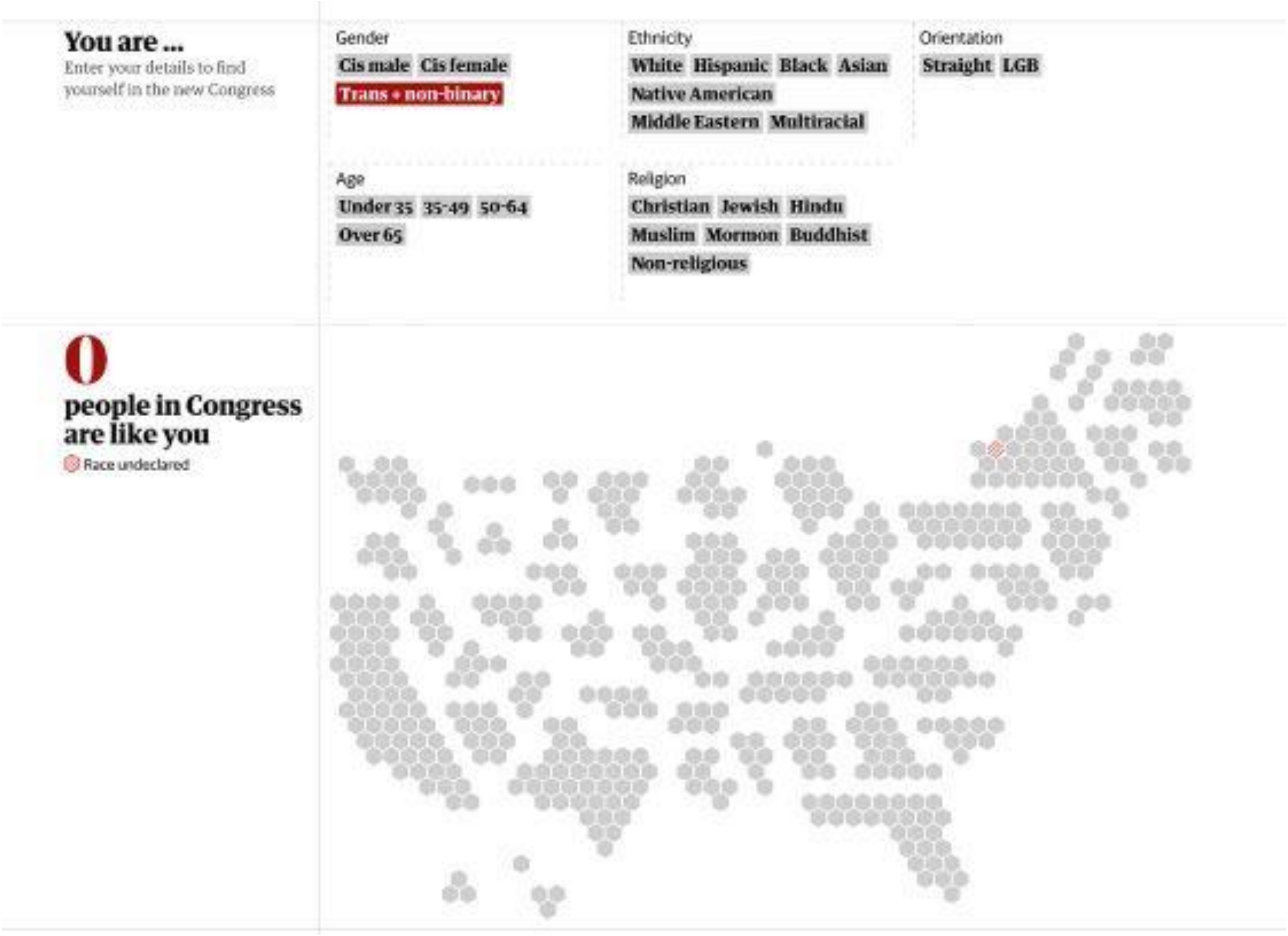
- Subvert norms



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

Subverting Power

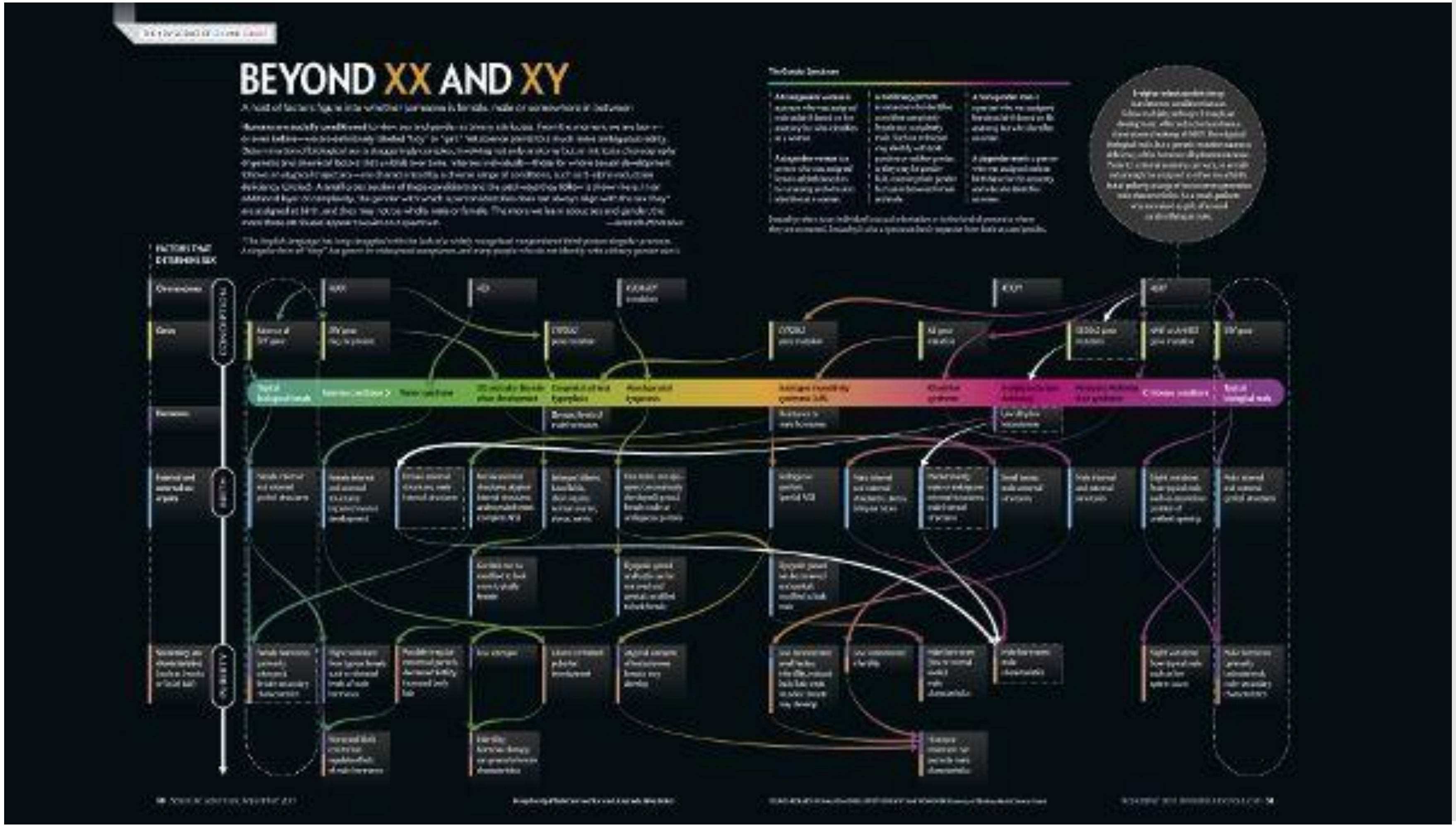
- Highlight missing categories



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?

In-class Activity:

- Go to the course website to find instructions for lab 01
- Be prepared to share your findings!