

# 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?

Google


founders of statistics

All Images Books News Videos Shopping Maps More Tools


Carl friedrich gauss Startup Entrepreneur Female Pearson Probability Investors History Statist >



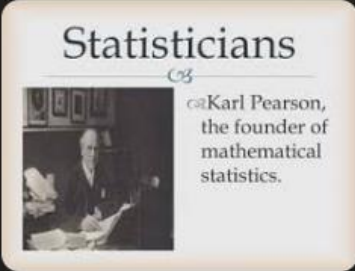
Wikipedia  
History of statistics...




Wikipedia  
History of statistics - Wikip...




Scholarly Community En...  
History of Statistics | E...




SlideShare  
Statistics | PPT



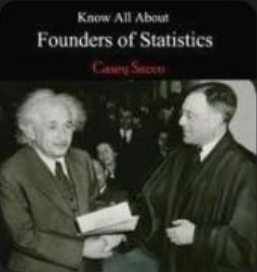
Vedantu  
founder of mathem...




Wikipedia  
History of statistics ...




Wikipedia  
History of statistics - ...




Blackwell's  
Statistics : Casey Sacco...




Department of Econ...  
History of Probabili...



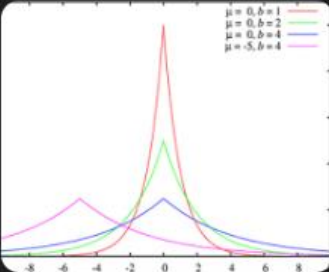
Harvard Gazette  
Harvard Gazette




International Statistical Institute  
2023 ISI Founders of Statistics ...



ResearchGate  
4: Wassily Hoeffdin...



Wikipedia  
History of statistics - Wikipedia



Stanford Statistics - Sta...  
Rob Tibshirani awarded...

Google

founders of computer science

All

Images

Books

Videos

Maps

News

Shopping

More

Tools

Saved

John mccarthy ai

Charles babbage


Invented

Artificial intelligence

Alan turing

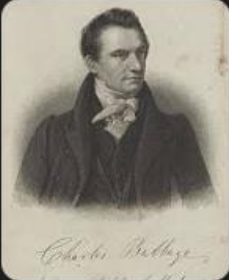
Famous personalities

The father




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




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

- Hmmmm..... 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

# 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)



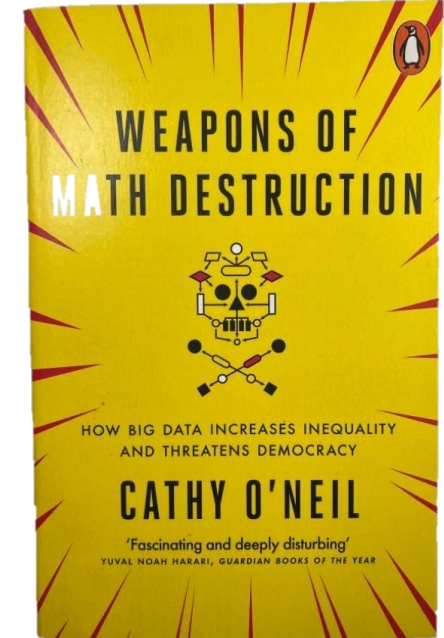
# Data Science & Power

## Algorithms as Power



Cathy O'Neil

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



# Data Science & Power

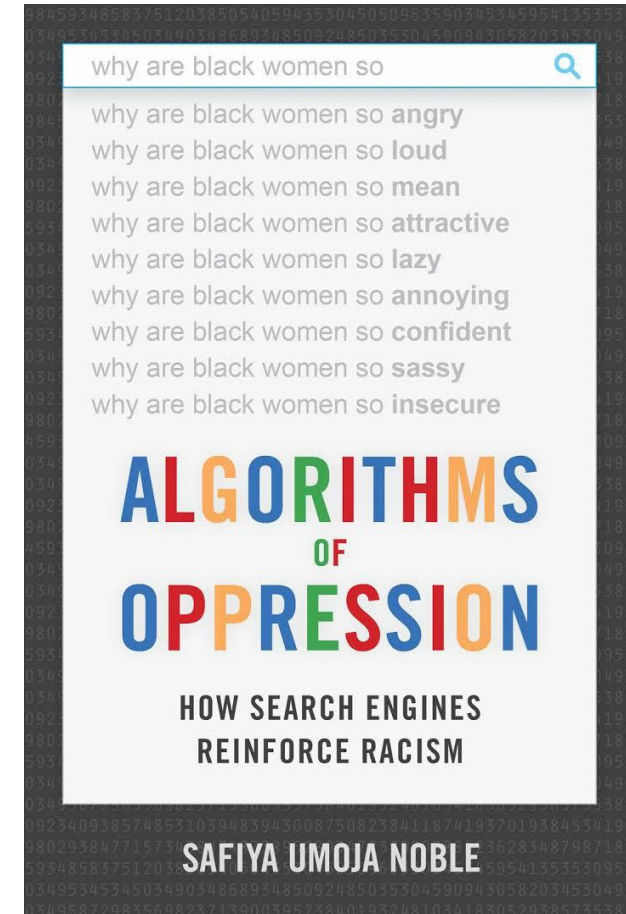
- The “Big Data” revolution argues that with enough data we can make unbiased decisions
- However, data science:
  - Lacks transparency
  - Employs extractive collection
  - Leverages technological complexity
  - Controls impact

# Example: Search Engines



Dr. Safiya Noble

<https://youtu.be/iRVZozEEWIE?si=qzRtPmQzxlgKDxR2>



## 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

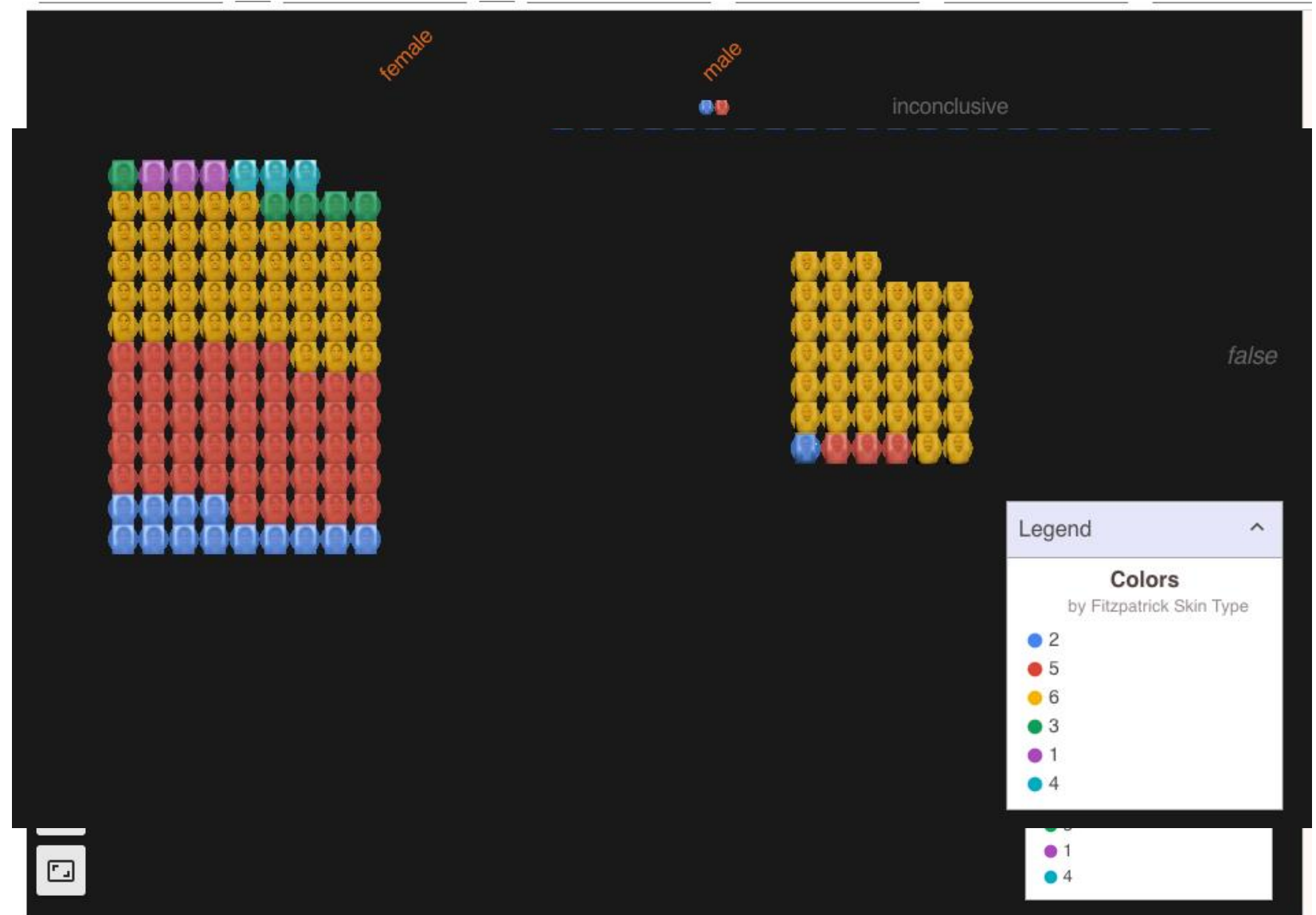


Dr. Joy Buolamwini

[https://youtu.be/UG\\_X\\_7g63rY?si=qDMmUX5VjpaJYURe](https://youtu.be/UG_X_7g63rY?si=qDMmUX5VjpaJYURe)

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

# Example: Facial Recognition



## 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?



“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.

**Sign Up**

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

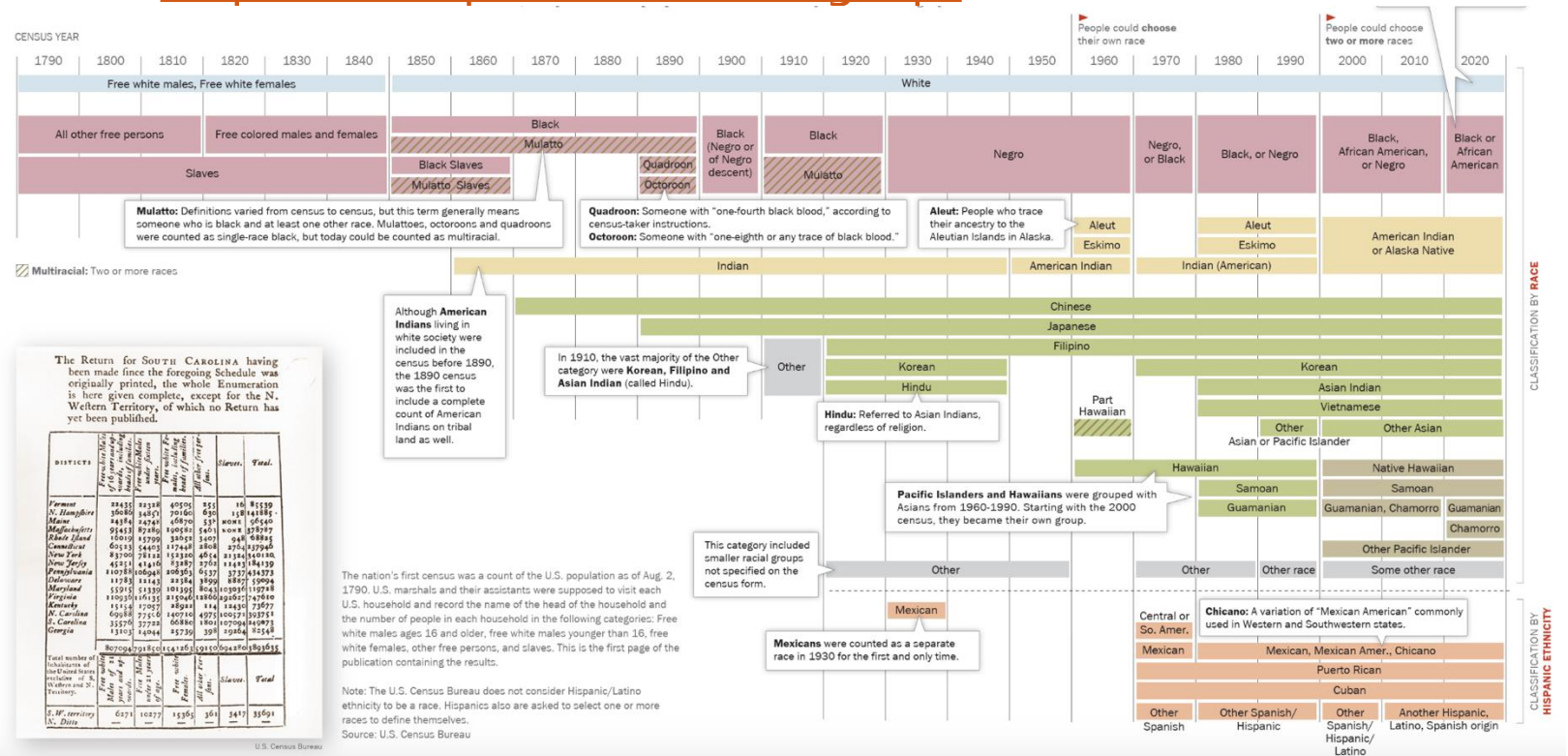
“What gets  
counted  
counts”

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



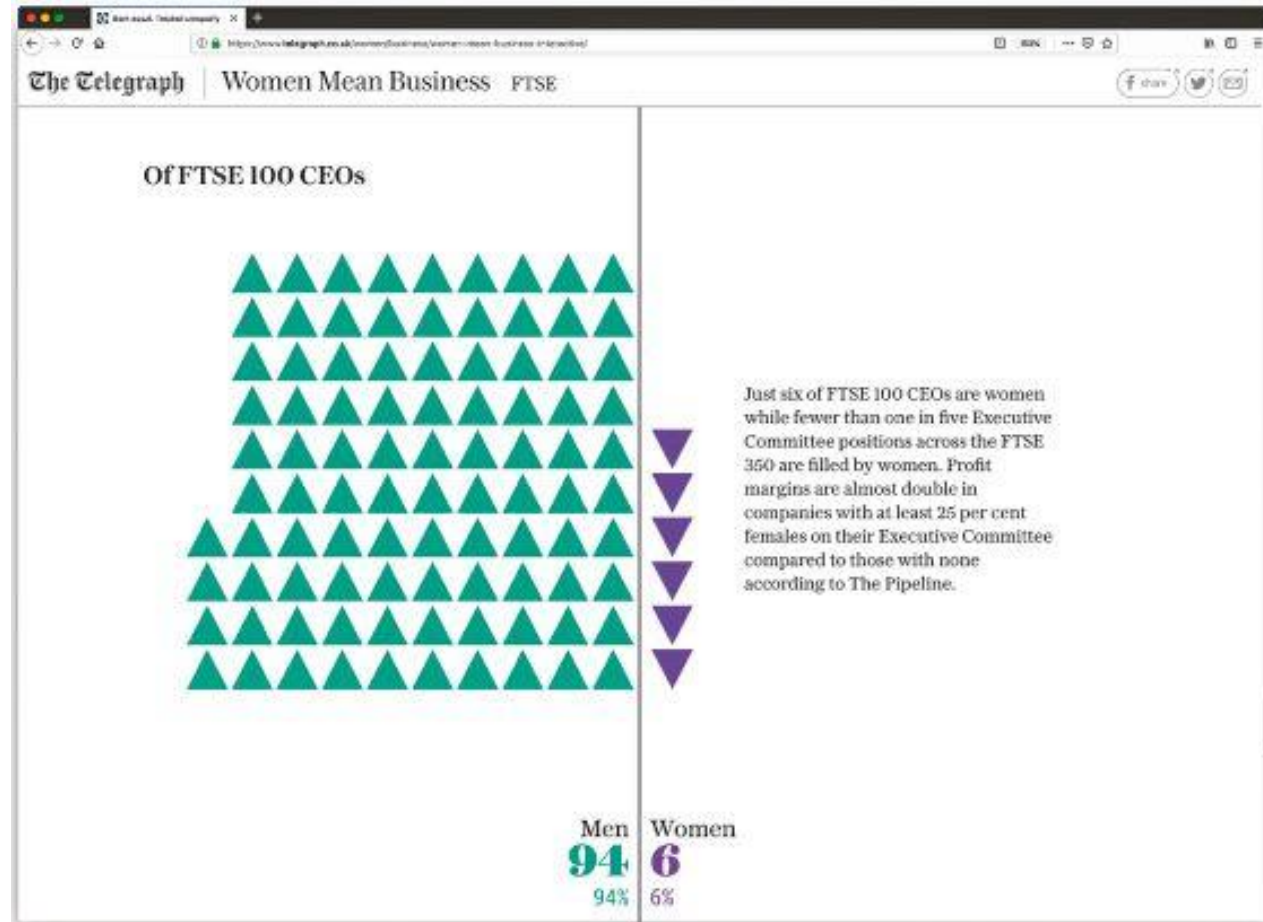
# "What gets counted counts"

- 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



# Subverting Power

- Highlight missing categories

<b>You are ...</b> Enter your details to find yourself in the new Congress	<b>Gender</b> <input type="checkbox"/> Cis male <input type="checkbox"/> Cis female <input checked="" type="checkbox"/> <b>Trans + non-binary</b>	<b>Ethnicity</b> <input type="checkbox"/> White <input type="checkbox"/> Hispanic <input type="checkbox"/> Black <input type="checkbox"/> Asian <input type="checkbox"/> Native American <input type="checkbox"/> Middle Eastern <input type="checkbox"/> Multiracial	<b>Orientation</b> <input type="checkbox"/> Straight <input type="checkbox"/> LGB
	<b>Age</b> <input type="checkbox"/> Under 35 <input type="checkbox"/> 35-49 <input type="checkbox"/> 50-64 <input type="checkbox"/> Over 65	<b>Religion</b> <input type="checkbox"/> Christian <input type="checkbox"/> Jewish <input type="checkbox"/> Hindu <input type="checkbox"/> Muslim <input type="checkbox"/> Mormon <input type="checkbox"/> Buddhist <input type="checkbox"/> Non-religious	

**people in Congress are like you**  
 Race undeclared







# Subverting Power

- Rethink data collection

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

☐

Woman  
(including trans woman)

☐

Non-binary

☐

In another way

☐

Man  
(including trans man)

☐

Prefer not to say

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

☐

Yes

☐

No

☐

Prefer not to say



# 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/>

## 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?

Subverting  
Power

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!