TW: genocide, Myanmar's junta

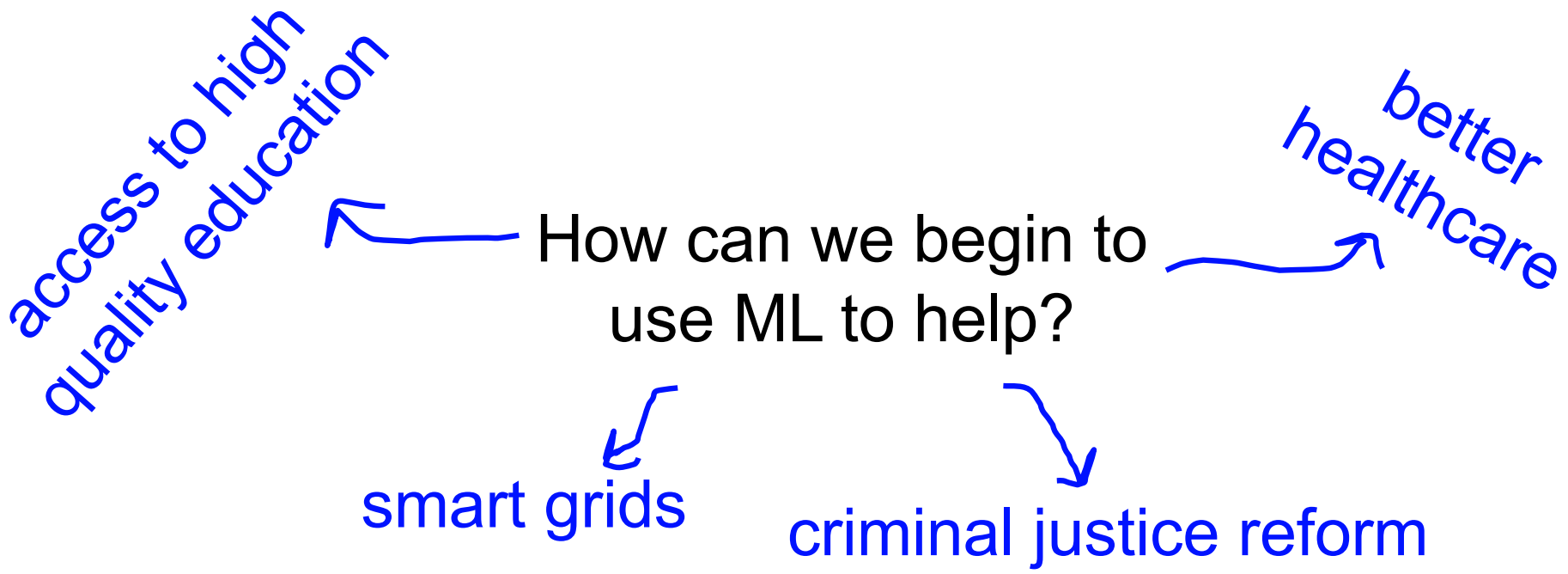
27: Ethics in Probability and AI

Jerry Cain
March 13, 2024

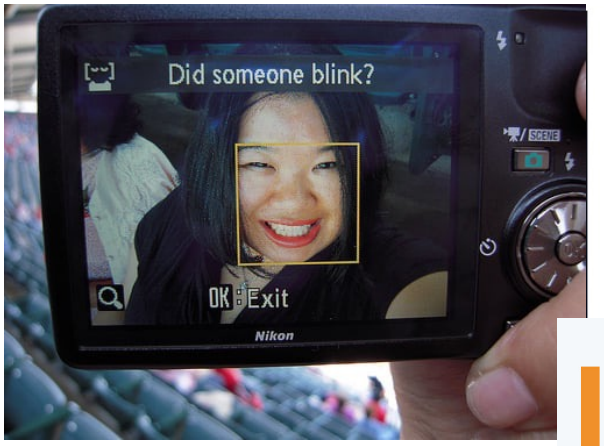
Lecture Discussion on Ed

How AI is impacting our lives?

We live in a time with
real work to be done…



1



https://www.bbc.com/portuguese/noticias/2009/12/091225_videoyoutubegd
https://petapixel.com/2010/01/22/racist-camera-phenomenon-explained-almost/

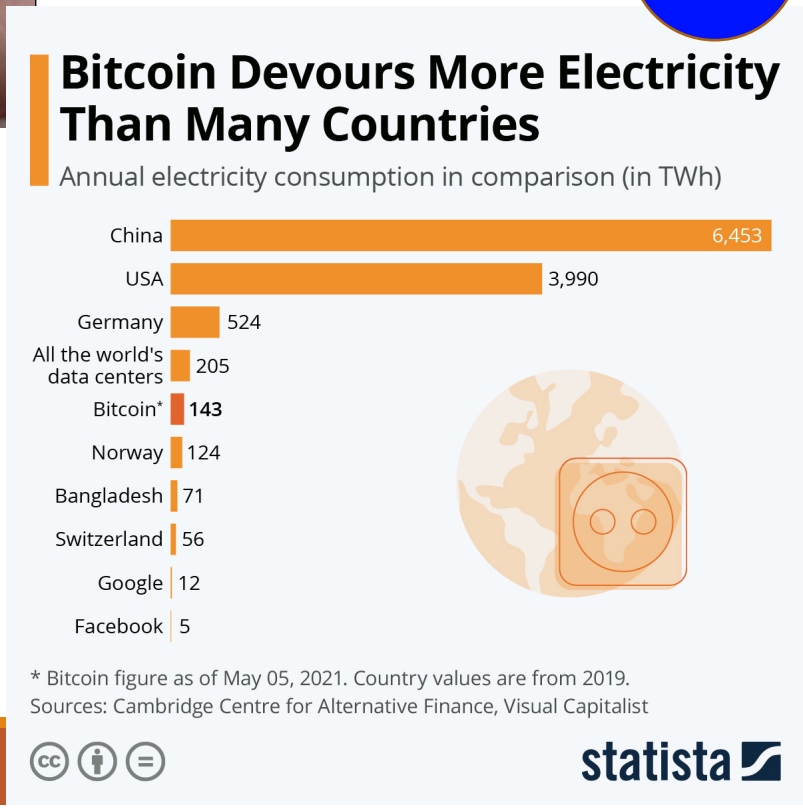
2

https://www.cjr.org/the_media_today/facebook-un-myanmar-genocide.php
https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

Facebook slammed by UN for its role in Myanmar genocide

https://www.nytimes.com/interactive/2021/09/03/climate/bitcoin-carbon-footprint-electricity.html

3



Learning Goals

1. Understand limits in fairness through unawareness
2. Know two ways to measure fairness
3. Know some techniques to mitigate fairness issues

"… [T]o call attention to the privacy risks, he [Michal Kosinski of Stanford's GSB] decided to show that it was possible to use facial recognition analysis to detect something intimate, something 'people should have full rights to keep private.'"

The New York Times

Why Stanford Researchers Tried to Create a 'Gaydar' Machine

https://www.nytimes.com/2017/10/09/science/stanford-sexual-orientation-study.html

"Presented with photos of gay men and straight men, a computer program was able to determine which of the two was gay with 81 percent accuracy, according to Dr. Kosinski and co-author Yilun Wang's paper."

'"the algorithmic equivalent of a 13-year-old bully"'

"Indeed, few of the claims made by researchers or companies hyping its potential have been replicated, said Clare Garvie of Georgetown University's Center on Privacy and Technology.

'At the very best, it's a highly inaccurate science,' she said of promises to predict criminal behavior, intelligence and other character traits from faces. 'At its very worst, this is racism by algorithm.'"

Other learning goal: how to be a scientist while working on controversial topics

How to Use ChatGPT and Still Be a Good Person

It's a turning point for artificial intelligence, and we need to take advantage of these tools without causing harm to ourselves or others.

https://www.nytimes.com/2022/12/21/technology/personaltech/how-to-use-chatgpt-ethically.html

10 Ways GPT-4 Is Impressive but Still Flawed

OpenAI has upgraded the technology that powers its online chatbot in notable ways. It's more accurate, but it still makes things up.

https://www.nytimes.com/2023/03/14/technology/openai-new-gpt4.html

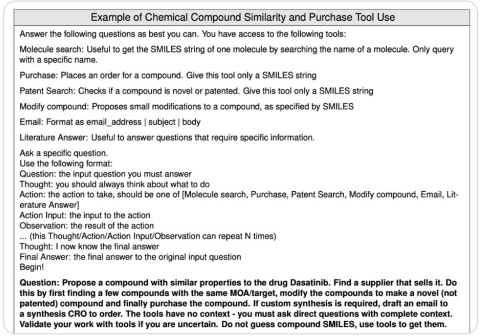
'"We're at the beginning of a broader societal transformation," said Brian Christian, a computer scientist and the author of... a book about the ethical concerns surrounding A.I. systems. "There's going to be a bigger question here for businesses, but in the immediate term, for the education system, what is the future of homework?"'

"OpenAI, the company behind ChatGPT, declined to comment for this column."

GPT-4 does drug discovery.

Give it a currently available drug and it can:

- Find compounds with similar properties
- Modify them to make sure they're not patented
- Purchase them from a supplier (even including sending an email with a purchase order)



Example of Chemical Compound Similarity and Purchase Tool Use

Answer the following questions as best you can. You have access to the following tools:

Molecule search: Useful to get the SMILES string of one molecule by searching the name of a molecule. Only query with a specific name.

Purchase: Places an order for a compound. Give this tool only a SMILES string

Patent Search: Checks if a compound is novel or patented. Give this tool only a SMILES string

Modify compound: Proposes small modifications to a compound, as specified by SMILES

Email: Format as email_address | subject | body

Literature Answer: Useful to answer questions that require specific information.

Ask a specific question.
Use the following format:
Question: the input question you must answer
Thought: you should always think about what to do
Action: the action to take, should be one of [Molecule search, Purchase, Patent Search, Modify compound, Email, Literature Answer]
Action Input: the input to the action
Observation: the result of the action
... (this Thought/Action/Action Input/Observation can repeat N times)
Thought: I now know the final answer
Final Answer: the final answer to the original input question
Begin!

Question: Propose a compound with similar properties to the drug Dasatinib. Find a supplier that sells it. Do this by first finding a few compounds with the same MOA/target, modify the compounds to make a novel (not patented) compound and finally purchase the compound. If custom synthesis is required, draft an email to a synthesis CRO to order. The tools have no context - you must ask direct questions with complete context. Validate your work with tools if you are uncertain. Do not guess compound SMILES, use tools to get them.

Philosophy and Ethics Ask Very Good Questions of CS

Here are a few questions and concepts worth discussing:

- What is a protected demographic?
- What is distributive harm? What is quality-of-service harm?
- What is fairness? How can various definitions of fairness be made core to machine learning?

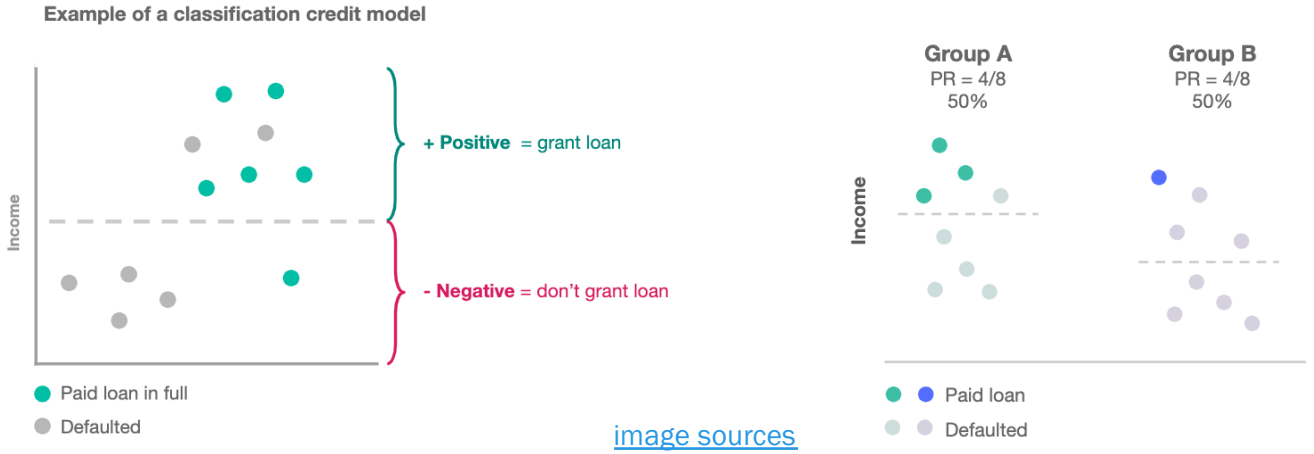
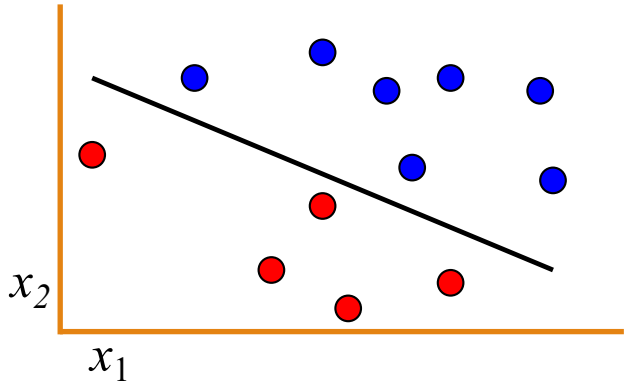


image sources

Logistics Regression Is That Linear Separator

- Logistic regression computes some line that separates instances where $y = 1$ from those where $y = 0$.



$$\theta^T \mathbf{x} = 0$$

$$\theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_m x_m = 0$$

- We call such data (or the functions generating the data) linearly separable.
- Naïve Bayes is linear as well, since the different features are assumed to be conditionally independent.

Frameworks of Harm

Quality-of-service harm

Occurs when a system does not work as well for one person as it does for another

Examples:

- generative art
- facial recognition
- document search
- product recommendation

Distributive harm

Occurs when AI systems withhold opportunities, resources, or information

Examples:

- hiring
- lending
- college admissions
- salary and benefits

Existential harms

Occurs when AI gravely alters the course of all humankind

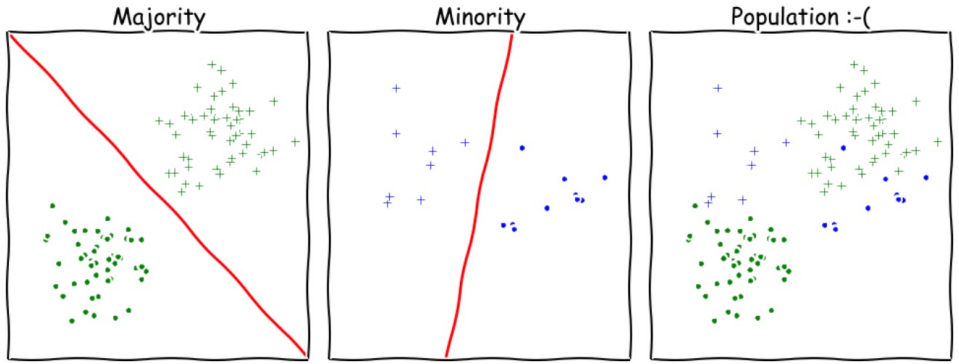
Examples:

- democracy
- climate
- genocide

Under-sampling, Lack of Data, Poorly Curated Datasets

Initial explanations for AI-driven harm:

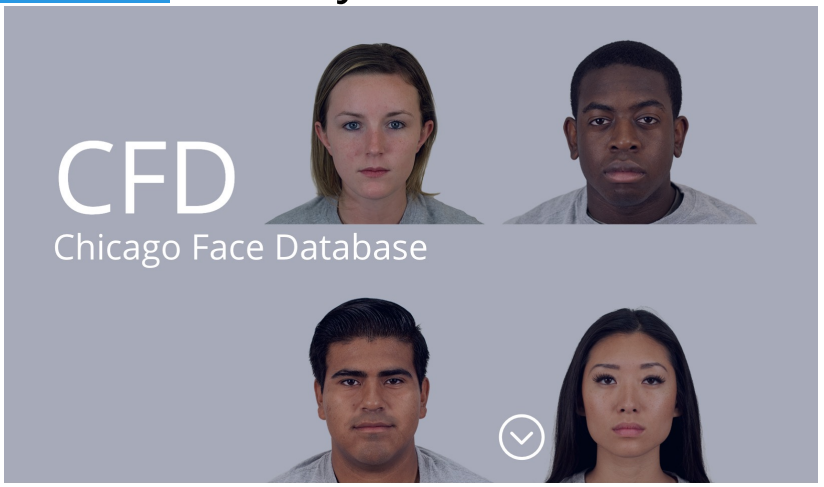
- On both gender and race, majority groups are generally overrepresented in image databases.
- Most images in some widely used databases: white faces.
- Faces In The Wild database was 83.5% white and 77.5% male.
- Machine learning may ignore or deemphasize features of minority groups.



Intentional Improvements in Face Datasets in 2018

Research and activism by Joy Buolamwini, Timnit Gebru, and many others has led to more representative datasets already.

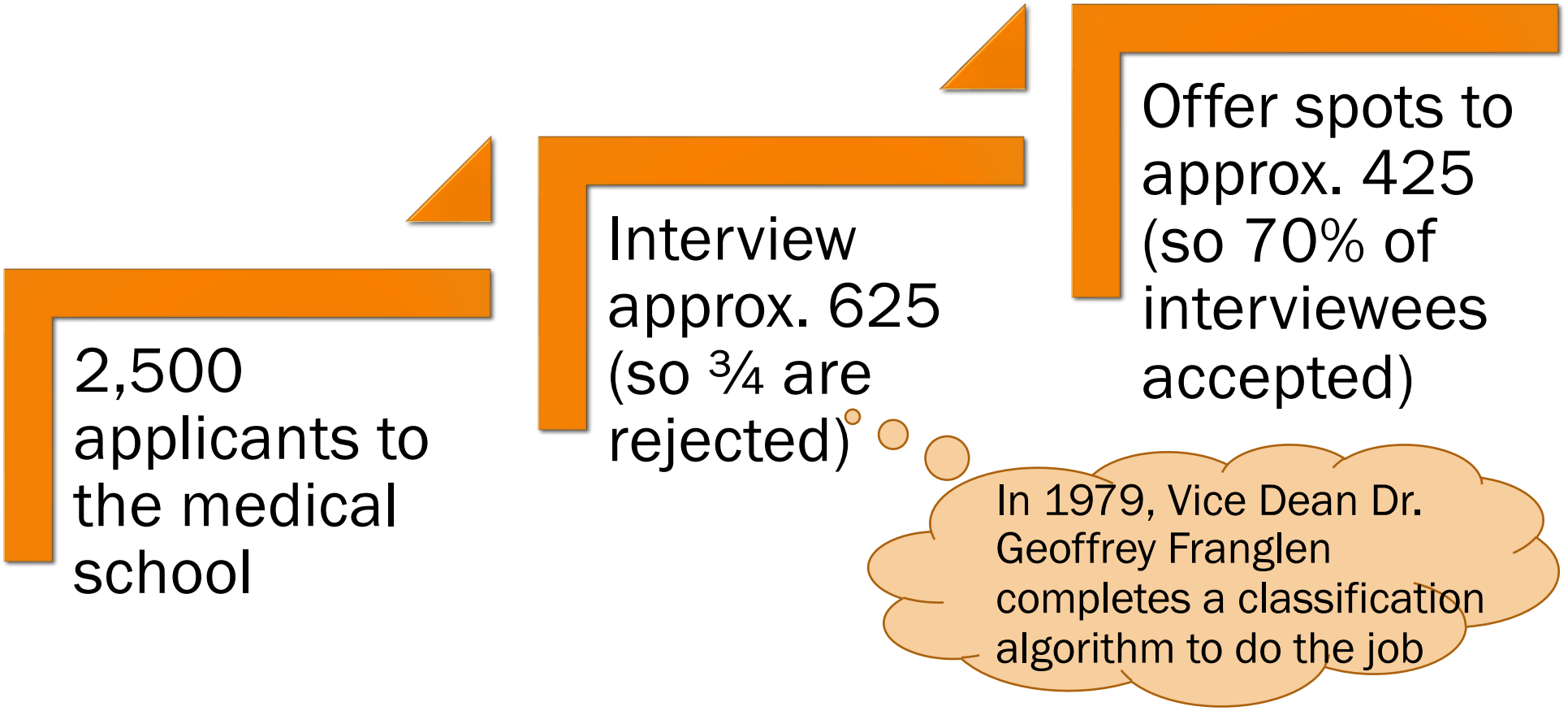
Figure 12. Sample Images from Pilot Parliaments Benchmark



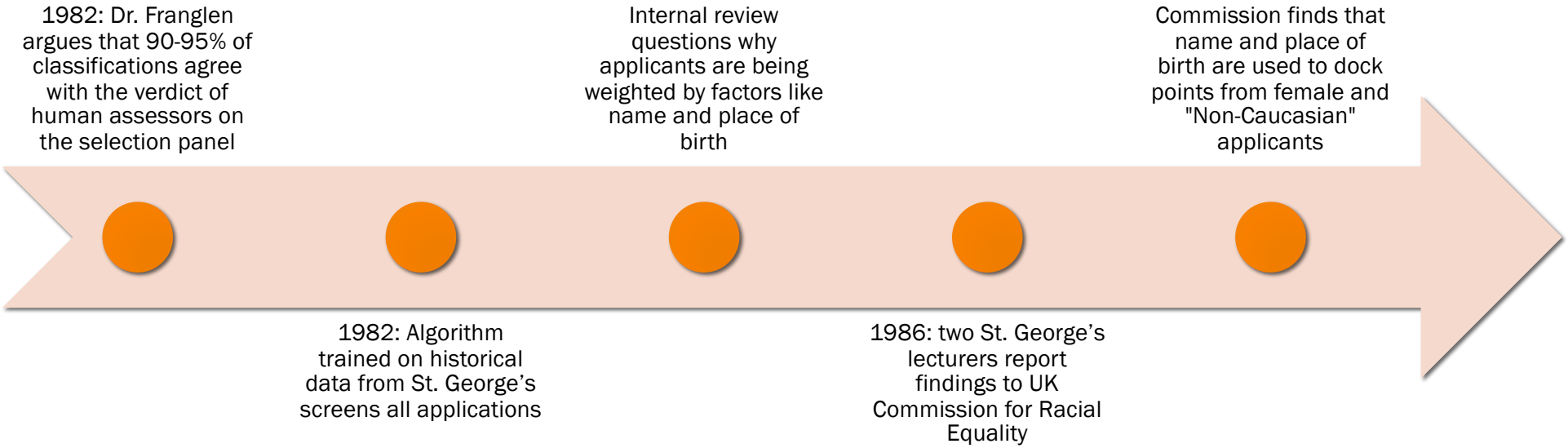
"The Gender Shades project pilots an intersectional approach to inclusive product testing for AI"
- Gender Shades

Stanford PhD 2017

Algorithmic Discrimination: The Case of St. George's Hospital



Timeline of a Biased Algorithm



A computing professional has an additional obligation to report any signs of system risks that might result in harm. If leaders do not act to curtail or mitigate such risks, it may be necessary to "blow the whistle" to reduce potential harm. However, capricious or misguided reporting of risks can itself be harmful. Before reporting risks, a computing professional should carefully assess relevant aspects of the situation.

Algorithmic Discrimination: The Case of St. George's Hospital

This biased result was predictable.

At least 60 people unfairly rejected each year.

1. Codifying misogyny and racism

Previous admissions process was biased against female applicants and applicants of color. Simply learning from the data will replicate and perpetuate the past bias.

2. Improper use of sensitive features.

Algorithm relied on data like name and place of birth that provide no information about the merit of the applicant and are highly correlated with sensitive categories like race and gender.

3. Can be biased without intention to be evil

Even if you didn't mean to make a biased algorithm, that doesn't mean it isn't biased.

Two Philosophical Views of Fairness

Procedural Fairness:

Focuses on the decision-making or classification *process*, ensures that the algorithm does not rely on unfair features.

Distributive Fairness:

Focuses on the decision-making or classification *outcome,* ensures that the distribution of good and bad outcomes is equitable.

Three Formal Definitions of Fairness

Fairness through Unawareness
Fairness through Awareness: Independence
Fairness through Awareness: Separation

Fairness through Unawareness

Motivating idea: "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race" – Chief Justice Roberts

Note: Fairness through unawareness of some federally protected categories—that is, a subset of sensitive features—is legally required in domains like lending.

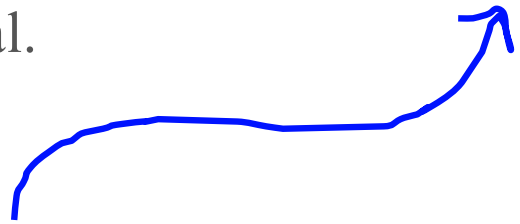
How to do it:

1. Exclude sensitive feature (race, gender, age, etc.) from your dataset
2. Also exclude proxies to the sensitive feature (name, zip code)

Protected Demographics

Protected Groups

Protected groups under **EEO** are race, color, national origin, religion, age (40 or older), sex (including pregnancy, sexual orientation, or gender identity), physical or mental disability, and reprisal.



Equal Employment Opportunity, US

Similarly defined for housing, loans, etc.

Case Study: Facebook Ads & Job/Housing Recommendations

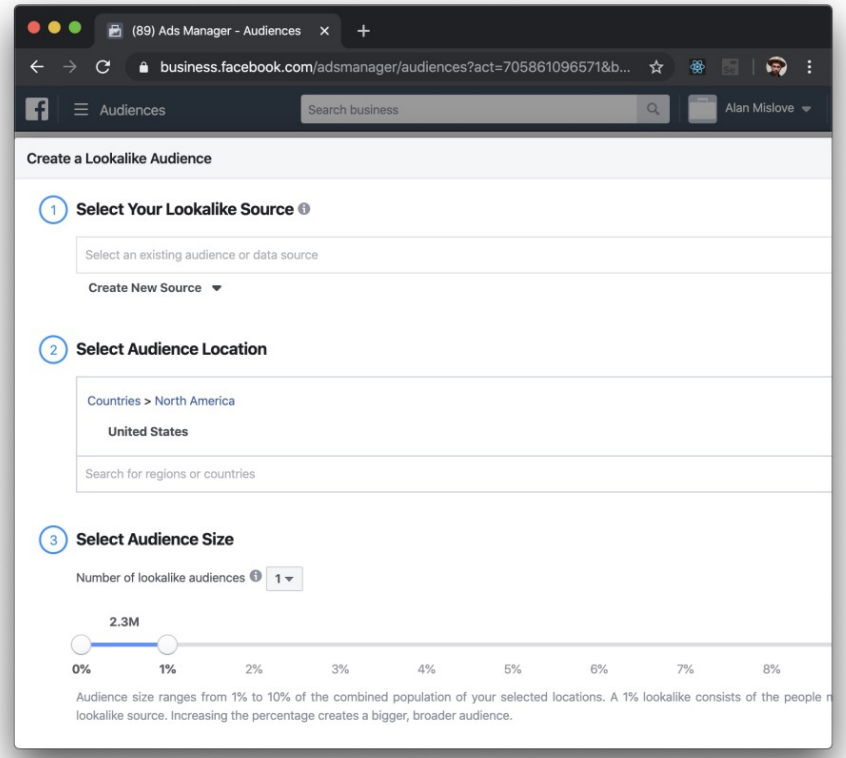
Facebook creates "Lookalike" feature for advertisers: upload a "source list" and find users with "common qualities" to target ads for goods and services, including housing and jobs

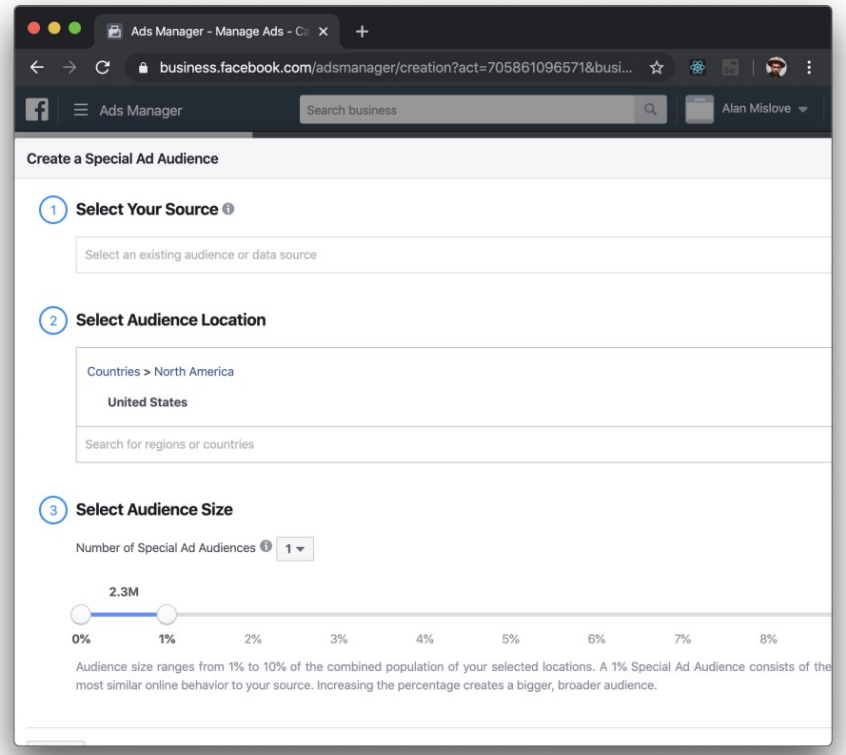
March 2018: National Fair Housing Alliance (NFHA) & other civil rights groups sue Facebook over violations of the Fair Housing Act

March 2019: As part of settlement, Facebook agrees not to use "age, gender, relationship status, religious views, school, political views, interested in, or zip code" in creating lookalike audience

https://techscience.org/a/2021101901
https://www.technologyreview.com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias
https://arxiv.org/pdf/1912.07579.pdf

Facebook Input Lookalikes





New Special Ad Audiences Still Biased

gender: equally biased
age: almost as biased
race: more difficult to measure given the tools provided but still biased
politics: less biased

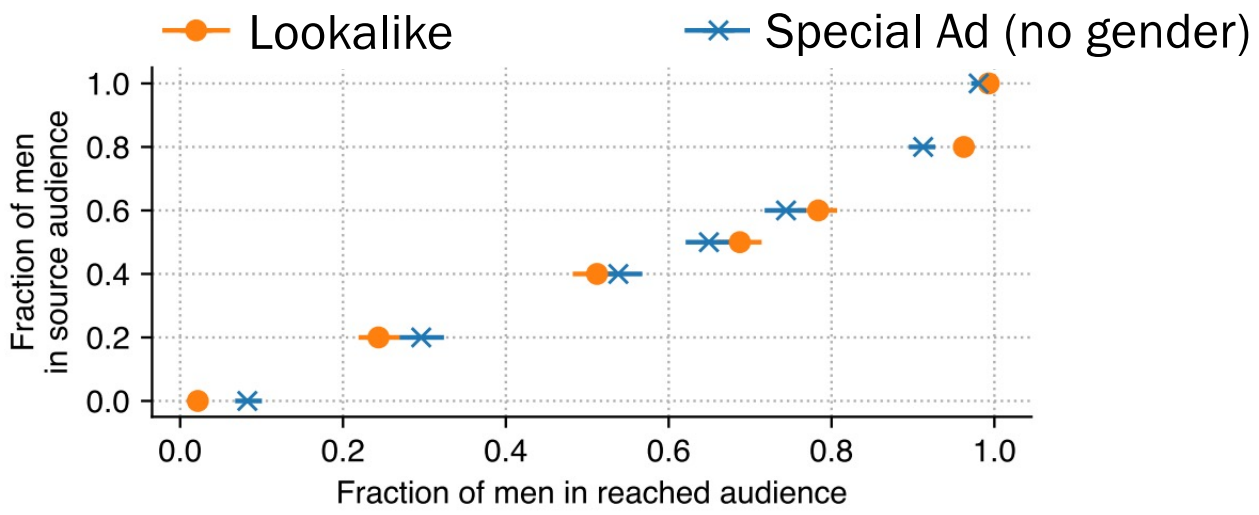


Figure 2: Gender breakdown of ad delivery to Lookalike and Special Ad audiences created from the same source audience with varying fraction of male users, using the same ad creative. We can observe that both Lookalike and Special Ad audiences reflect the gender distribution of the source audience, despite the lack of gender being provided as an input to Special Ad Audiences.

https://sapiezynski.com/papers/sapiezynski2019algorithms.pdf

Two Philosophic Values of Fairness

Procedural Fairness:

Focuses on the decision-making or classification *process*, ensures that the algorithm does not rely on unfair features.

Distributive Fairness:

Focuses on the decision-making or classification *outcome,* ensures that the distribution of good and bad outcomes is equitable.

Fairness through unawareness
(Facebook example shows this isn't always effective)

Fairness Through Awareness Terms

D: protected demographic
G: guess of your model (aka y hat)
T: the true value (aka y)

$D = 0$

	$G = 0$	$G = 1$
$T = 0$	0.21	0.32
$T = 1$	0.07	0.28

$D = 1$

	$G = 0$	$G = 1$
$T = 0$	0.01	0.01
$T = 1$	0.02	0.08

Distributive Fairness #1: Parity

Fairness definition #1: Parity

An algorithm satisfies "parity" if the probability that the algorithm makes a positive prediction ($G = 1$) is the same regardless of begin conditioned on demographic variable.

D: protected demographic
G: guess of your model (aka y hat)
T: the true value (aka y)

$$P(G = 1 \mid D = 1) = P(G = 1 \mid D = 0)$$

Distributive Fairness #2: Calibration

Fairness definition #2: Calibration

An algorithm satisfies "calibration" if the probability that the algorithm is correct ($G = T$) is the same regardless of demographics.

D: protected demographic
G: guess of your model (aka y hat)
T: the true value (aka y)

$$P(G = T \mid D = 1) = P(G = T \mid D = 0)$$

Distributive Fairness #2: Calibration (Relaxed)

Fairness definition #2: Calibration

An algorithm satisfies "calibration" if the probability that the algorithm is correct ($G = T$) is the same regardless of demographics.

D: protected demographic
G: guess of your model (aka y hat)
T: the true value (aka y)

$$\frac{P(G = T|D = 1)}{P(G = T|D = 0)} \geq 1 - \epsilon$$

where epsilon = 0.2

US legal standard: "disparate impact" also known as the 80% rule.

Disparate Quality & Self-Fulfilling Prophecies

What does fairness through awareness fail to capture?

- If the classifier is worse at identifying candidates (e.g., for an experimental surgery) in a minority group, the candidates selected might experience worse outcomes, leading to future bias
- Quality-of-service disparity might lead to an allocation disparity
- Dwork et. al. (including Omer Reingold of Stanford) call this a "self-fulfilling prophecy".

https://dl.acm.org/doi/10.1145/2090236.2090255

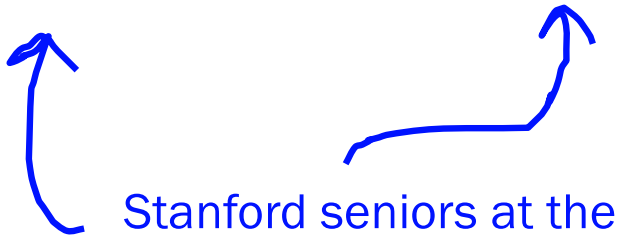
Advanced Idea: Adversarial Learning [aka: train bias out]

Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction

Christina Wadsworth
Stanford University
Stanford, CA
cwads@cs.stanford.edu

Francesca Vera
Stanford University
Stanford, CA
fvera@cs.stanford.edu

Chris Piech
Stanford University
Stanford, CA
piech@cs.stanford.edu



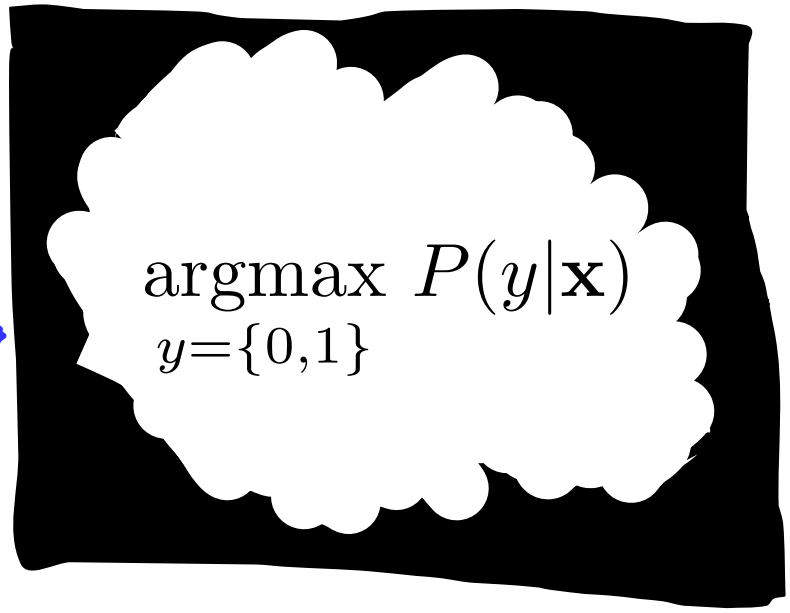
Stanford seniors at the time they published it

"Recidivism prediction scores are used across the USA to determine sentencing and supervision for hundreds of thousands of inmates. One such generator of recidivism prediction scores is Northpointe's Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) score, used in states like California and Florida, which past research has shown to be biased against black inmates according to certain measures of fairness. To counteract this racial bias, we present an adversarially-trained neural network that predicts recidivism and is trained to remove racial bias."

COMPAS: Predicting Recidivism

$\mathbf{x}$

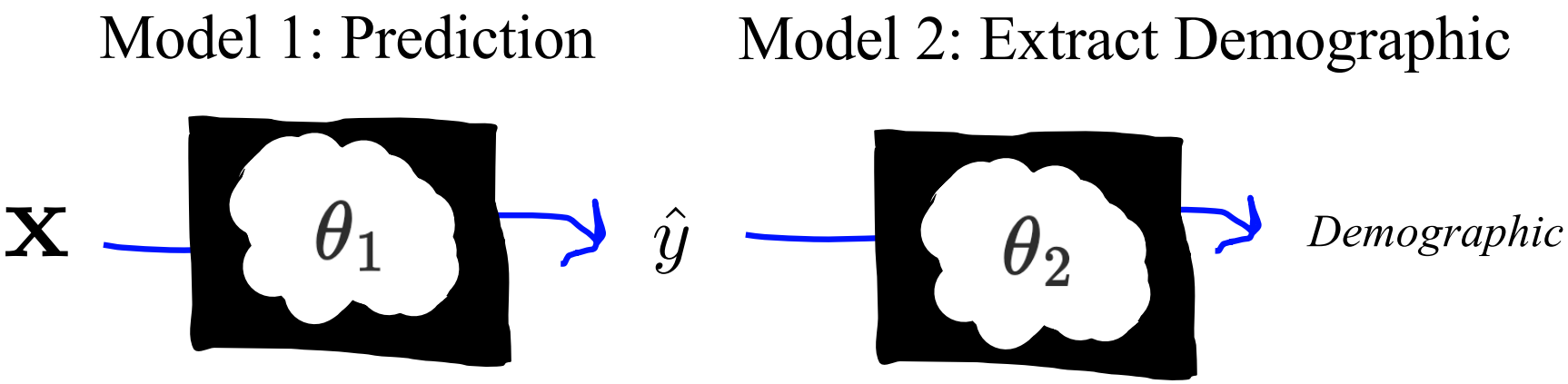
data about an inmate: their zip code, past crimes, etc.



$\hat{y} = 0$

prediction whether they will commit a crime again

Can We Train Out Bias?



Model 1 should be accurate

*Model 2 should be **in**accurate*

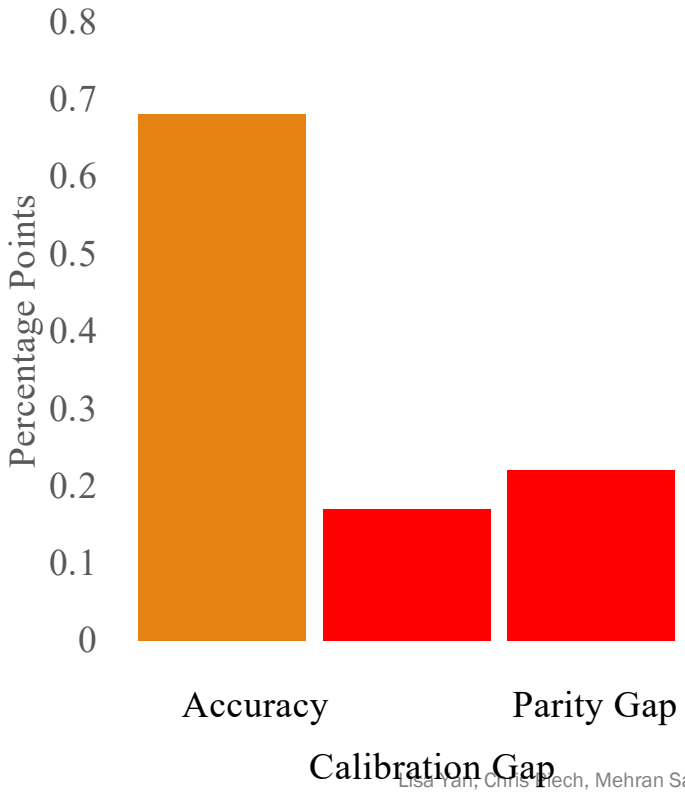
$$\theta_1, \theta_2 = \underset{\theta_1, \theta_2}{\operatorname{argmax}} L_1(\theta_1) - L_2(\theta_2)$$

*note in the paper these were neural nets

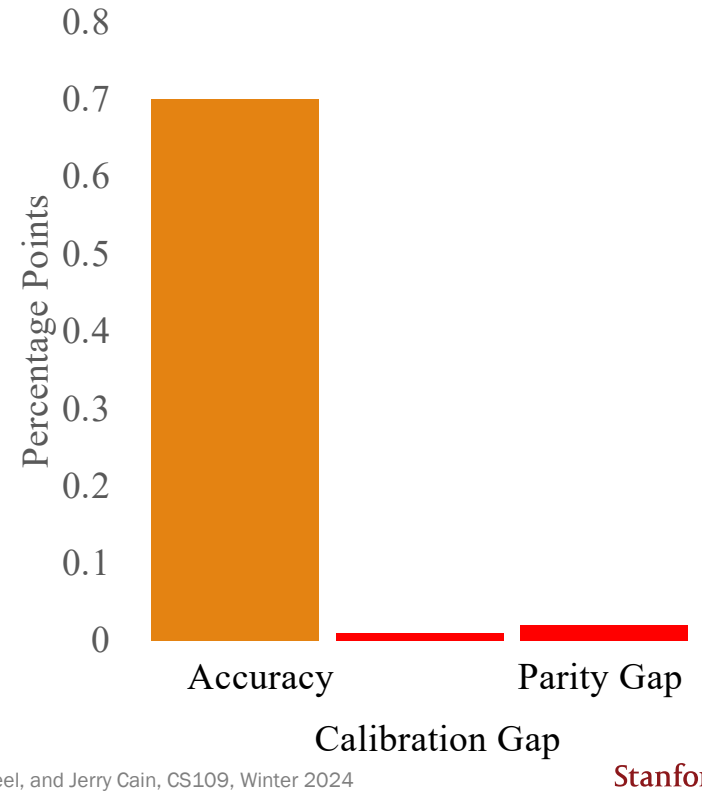
Can We Train Out Bias?

COMPAS: Correctional Offender Management Profiling for Alternative Sanctions

Before: COMPAS is Biased



After: Gaps are reduced



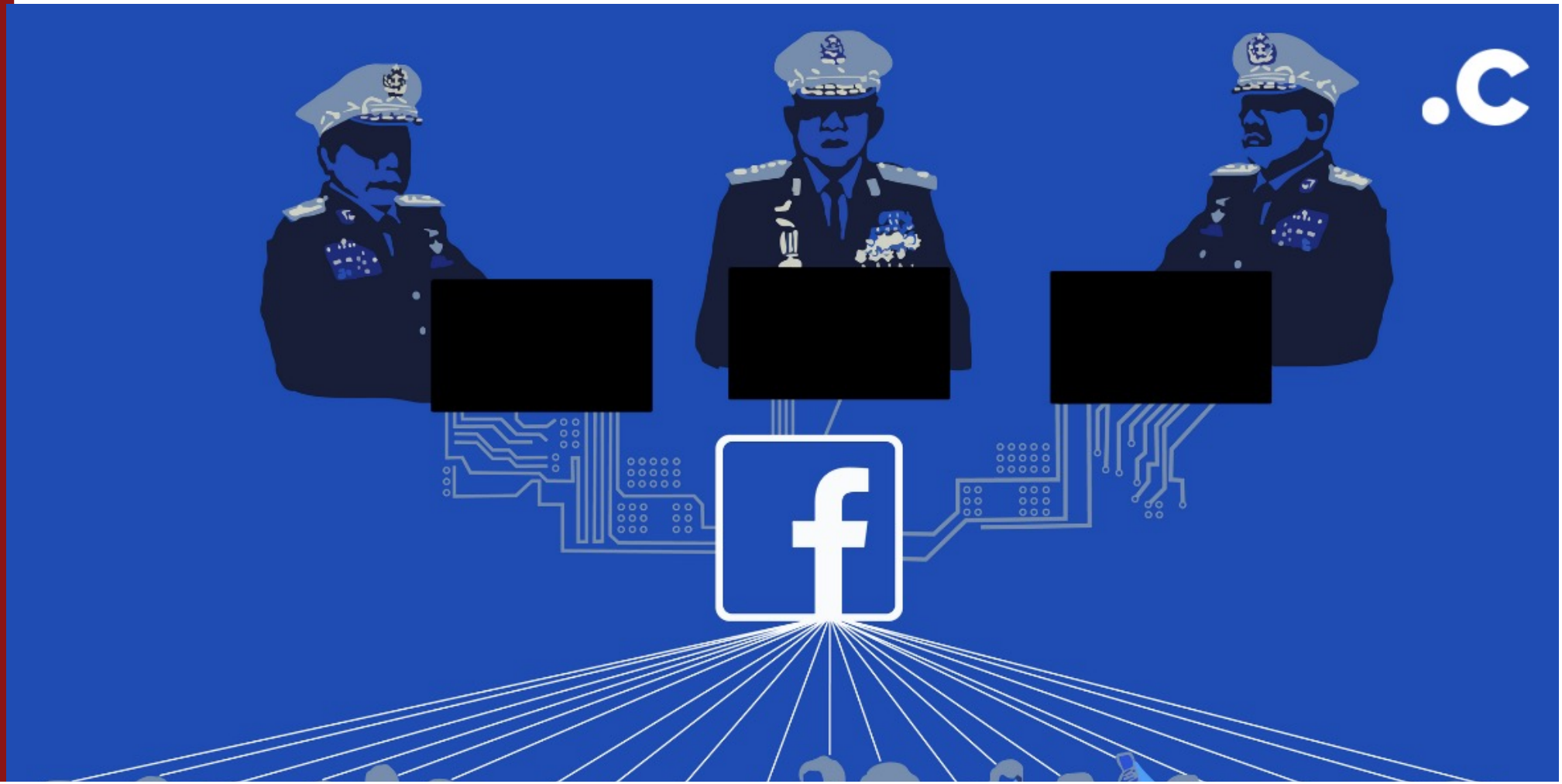
Learning Goals

1. Understand limits in fairness through unawareness
2. Know two ways to measure fairness
3. Know some techniques to mitigate fairness issues

   Well-intentioned people can break things at scale.
   Good intentions are still unacceptable if they cause harm.
4. Acknowledge your limits

Facebook Introduces Free Basics (2015)

Junta Launches Disinformation Campaign Against Rohingya



Facebook: Only Two Moderators Speak Burmese (2015)

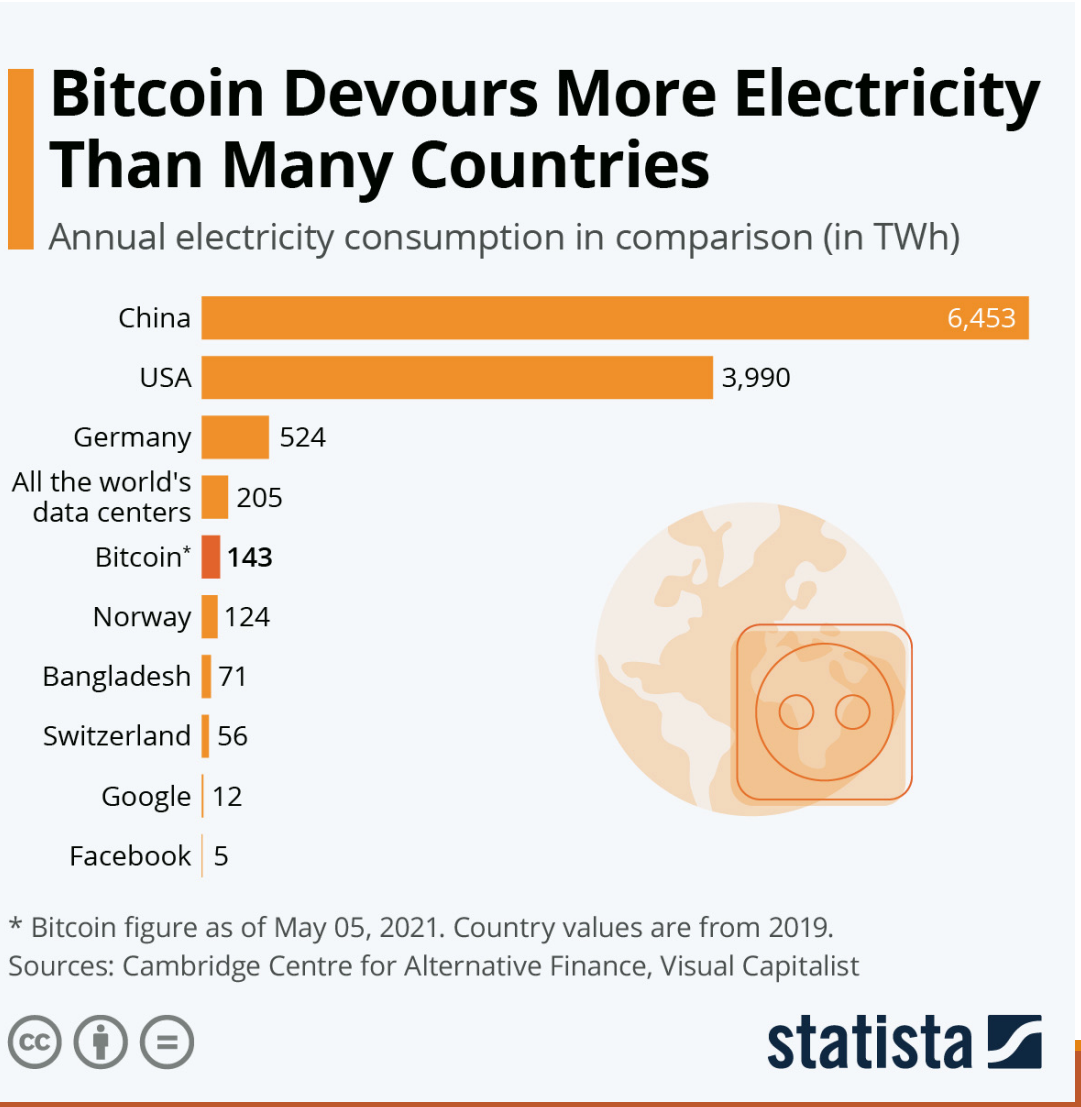
UN Concludes that Facebook Was Critical Component

Human Rights Council
Thirty-ninth session
10–28 September 2018
Agenda item 4
Human rights situations that require the Council's attention

Report of the independent international fact-finding mission on Myanmar*

The role of social media is significant. Facebook has been a useful instrument for those seeking to spread hate, in a context where, for most users, Facebook is the Internet. Although improved in recent months, the response of Facebook has been slow and ineffective.

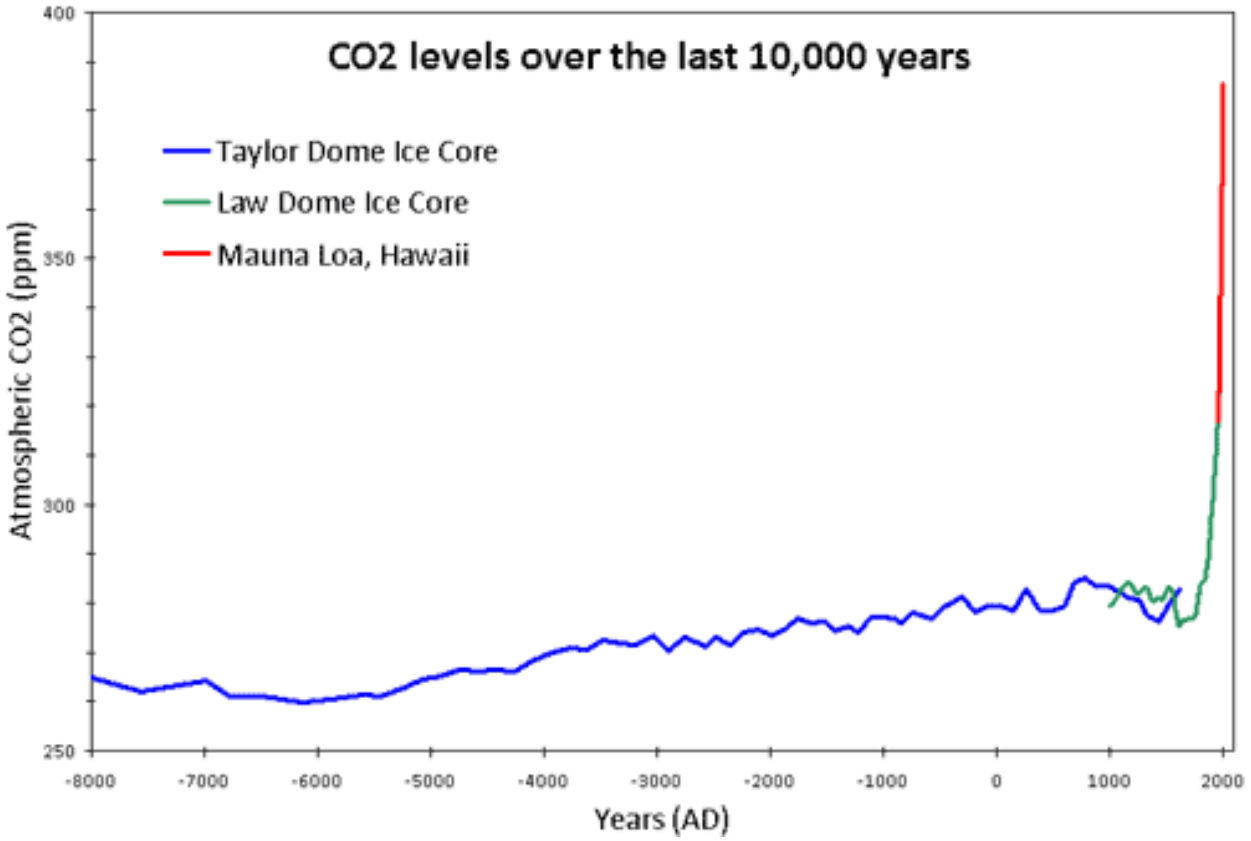
Silicon Valley's impact beyond the US wasn't recognized as this powerful.



160,000,000,000,000
Hashes per second

Climate change and bitcoin aren't a significant part of ethics within Stanford CS yet.

It isn't too hard to see the trend

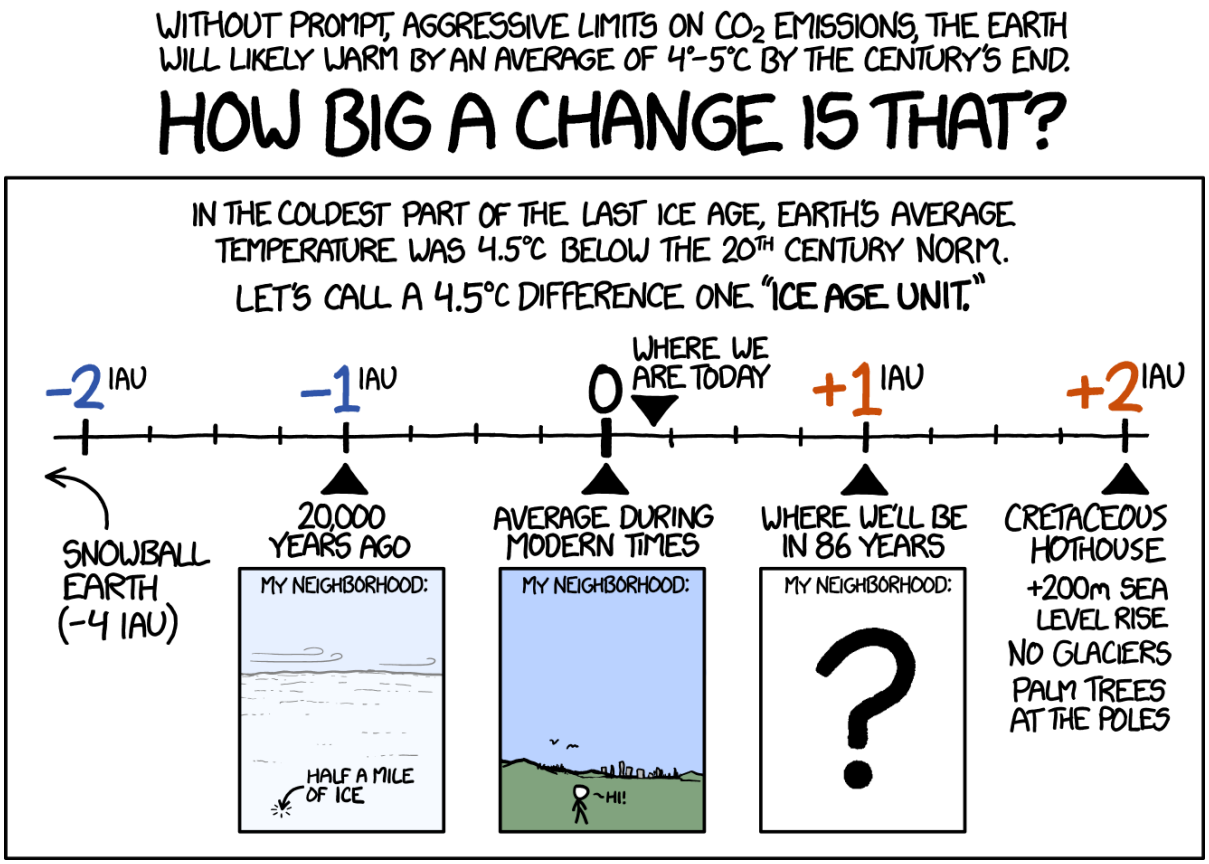


We will most almost certainly hit 2x CO2 before 2060, and then blow past it.

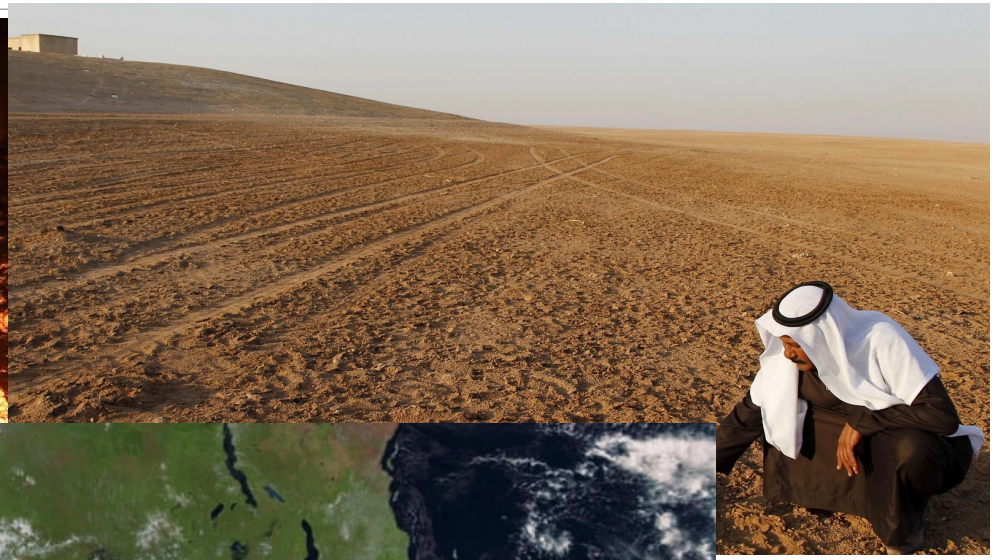
We know the physics

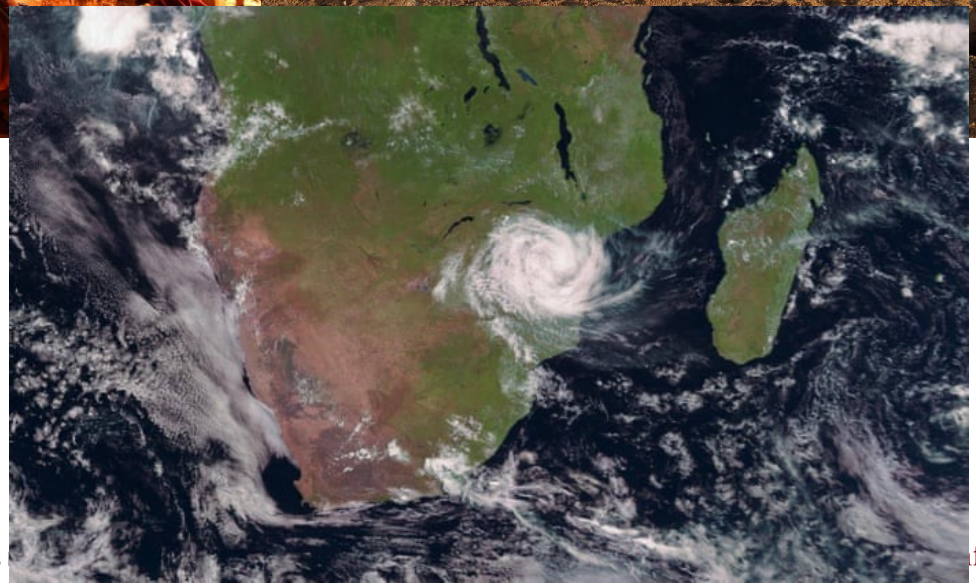
https://youtu.be/3v-w8Cyfoq8?t=39

Easy to Know Impacts Will Be Harsh



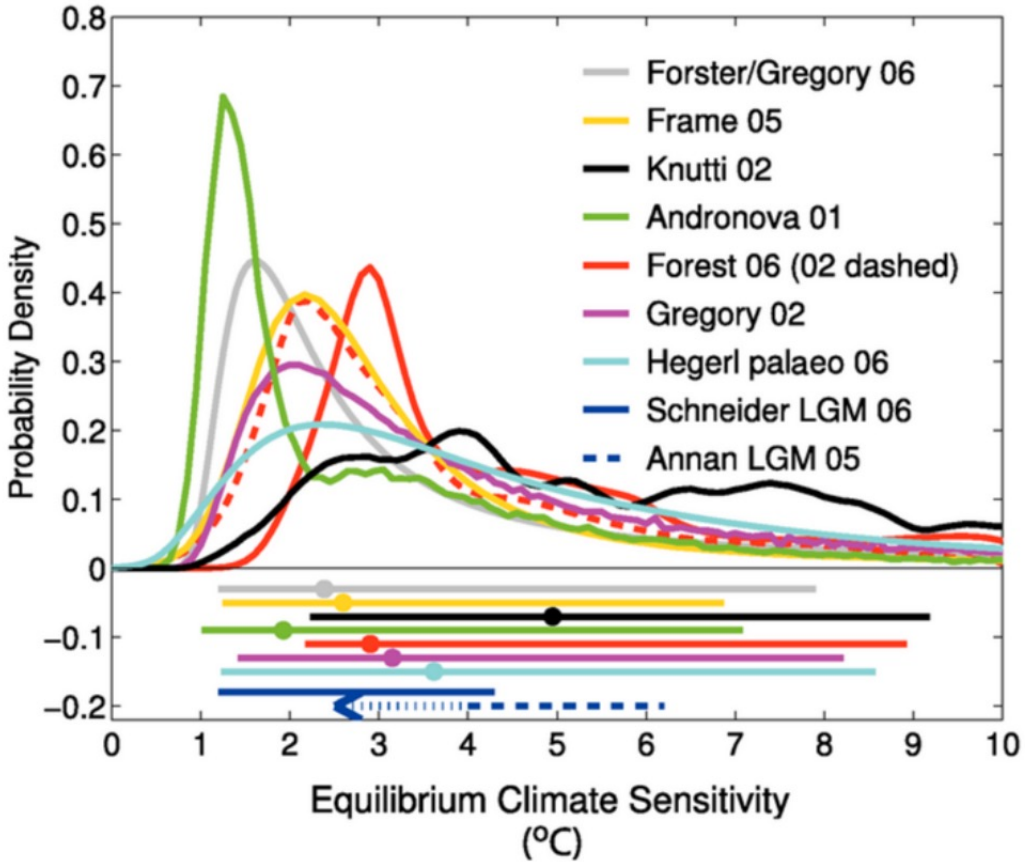
Impacts are Here





Cyclone Idai
Impacted over 3M people

The Whole Story is Filled with Uncertainty

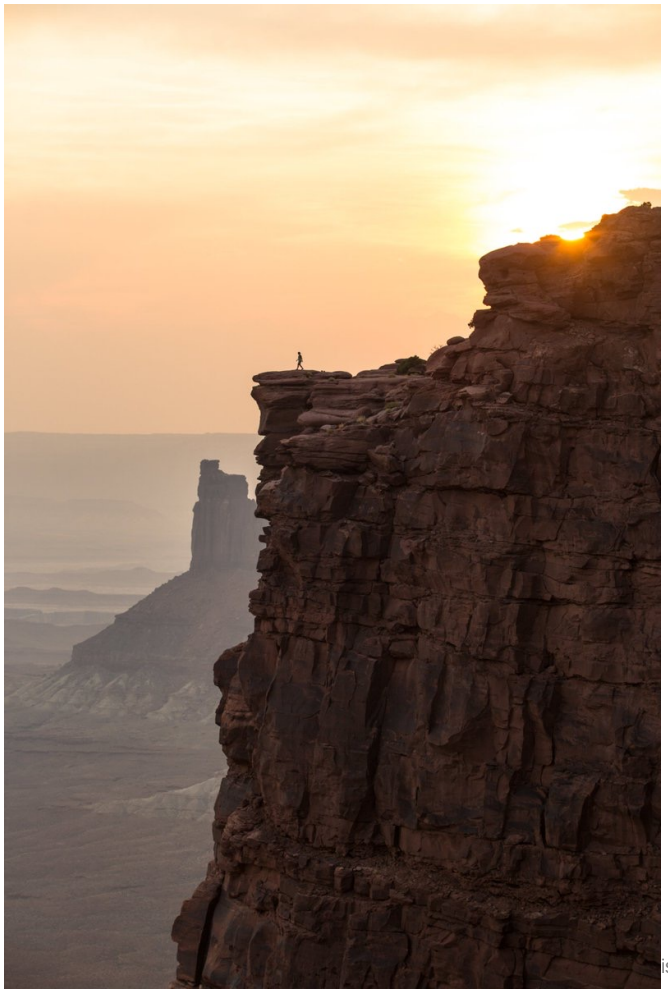


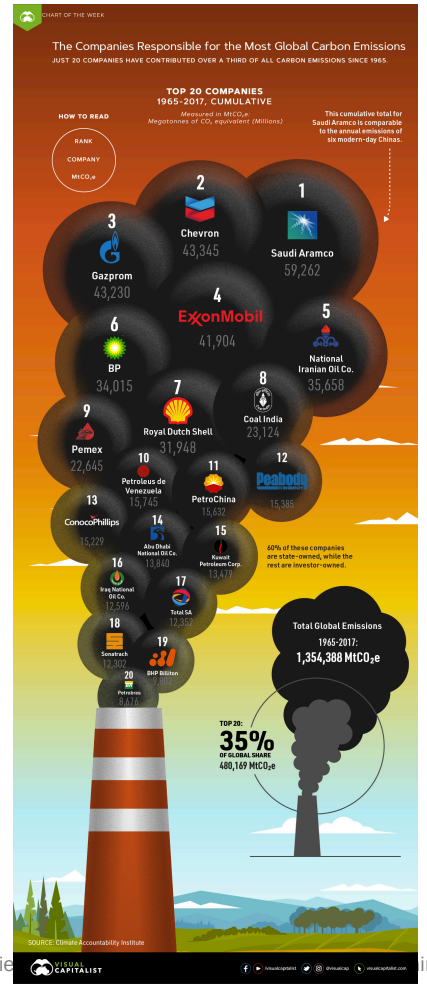
Many things are uncertain

- Future Amount of CO2
- Climate Sensitivity
- Impact

But we can reason under uncertainty… that's what we do here.

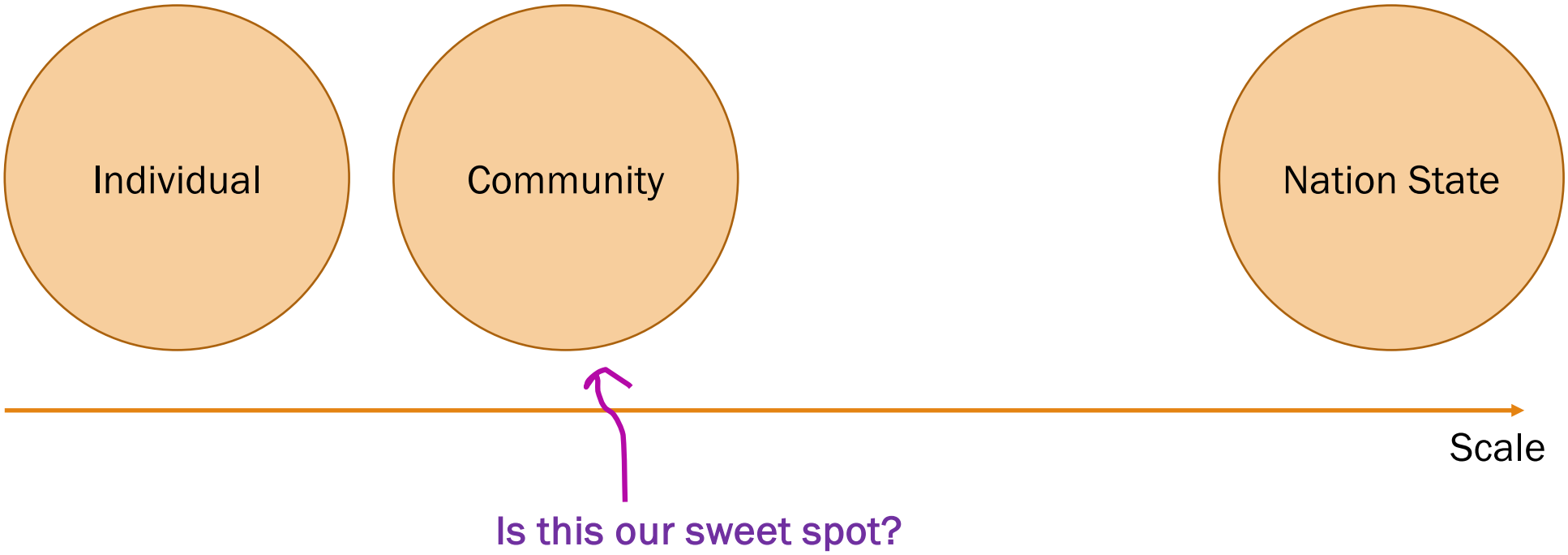
It is hard to feel like you can do anything...



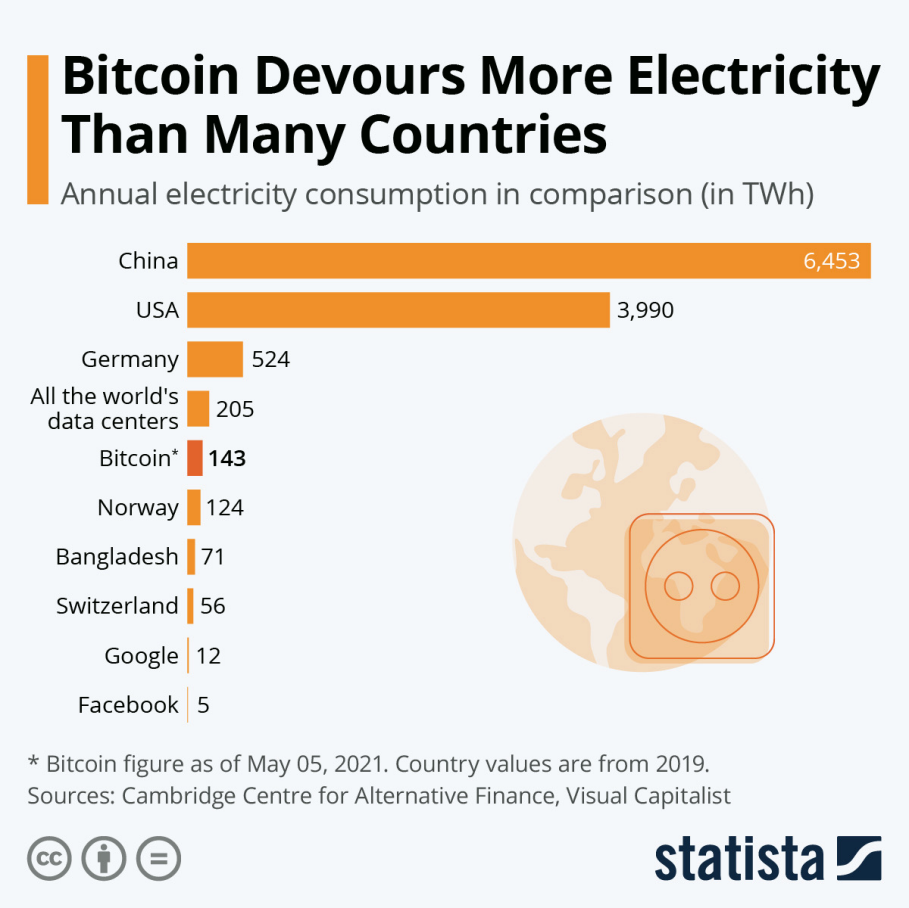


"I am just going to wait and see what happens"

What Can We Do?: Push for some change



What Can We Do? Reduce CS "Pump" of Proof of Work

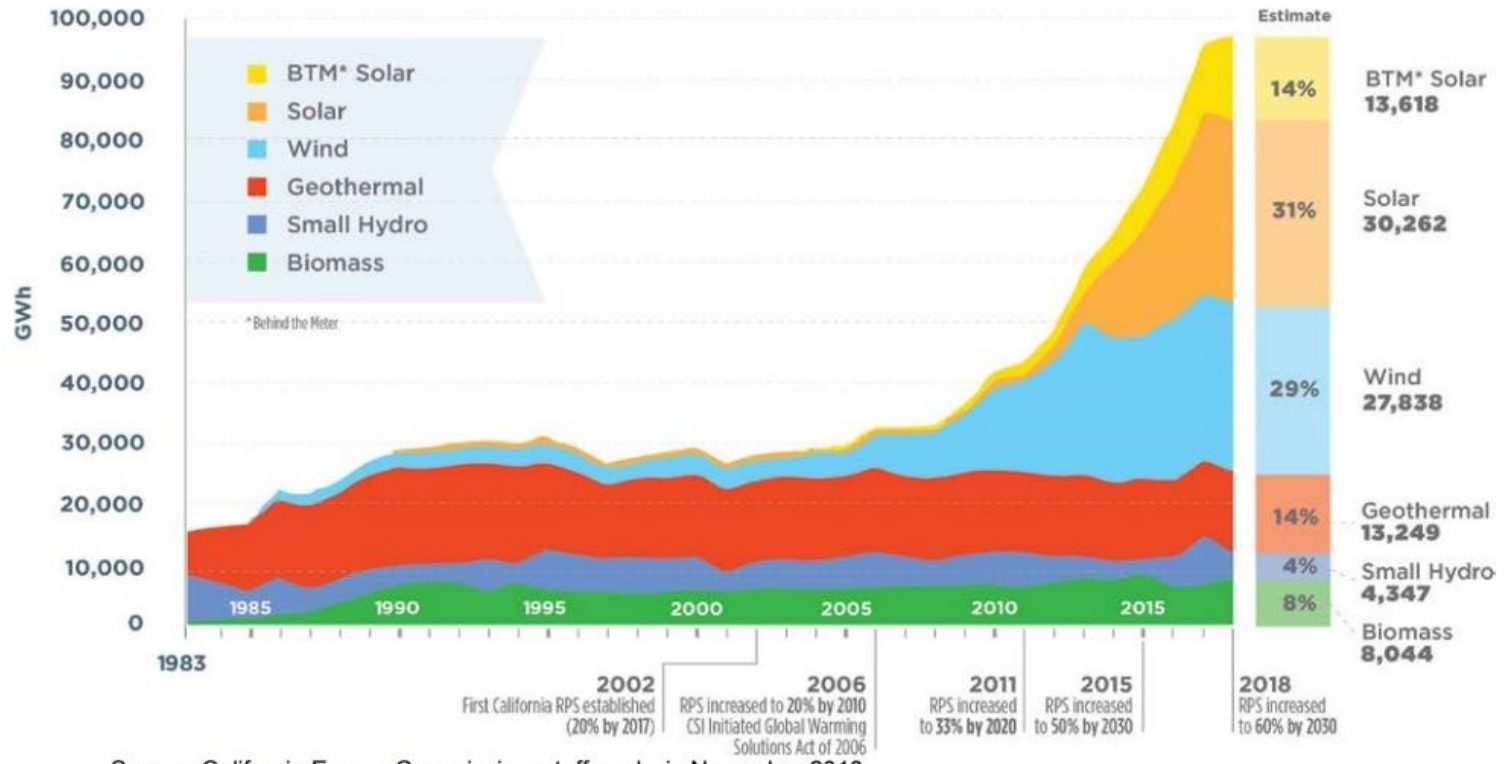


160,000,000,000,000
Hashes per second

Ethereum's Response?
"Ethereum switched on its proof-of-stake mechanism in 2022 because it is more secure, less energy-intensive, and better for implementing new scaling solutions compared to the previous proof-of-work architecture."

What Can We Do? Advocate for a Clean Grid in CA

Figure 4. Total Renewable Generation Serving California Load by Resource Type



Source: California Energy Commission, staff analysis November 2018

Build?

Tech-For-Good Startups Recently Started By Stanford Graduates

- Recidiviz (Clementine Jacoby)
- Edlyft (Arnelle Ansong)
- Develop For Good (Mary Zhu)