Ethics in ML

Chris Piech and Katie Creel
CS109, Stanford University

Announcements

1. Contest due next Monday. Submission will be via Gradescope
2. Don’t use demographic data when training your algorithms (see this lecture for details)

Mencius Philosophy on Ethics

Mencius holds that all humans have innate but incipient tendencies toward benevolence, righteousness, wisdom, and propriety. Employing an agricultural metaphor, he refers to these tendencies as “sprouts” (2A6). The sprouts are manifested in cognitive and emotional reactions characteristic of the virtues.

How AI is impacting our lives?


REMEMBER, WITH GREAT POWER
COMES GREAT RESPONSIBILITY
memegenerator.net

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

Access to High
quality education

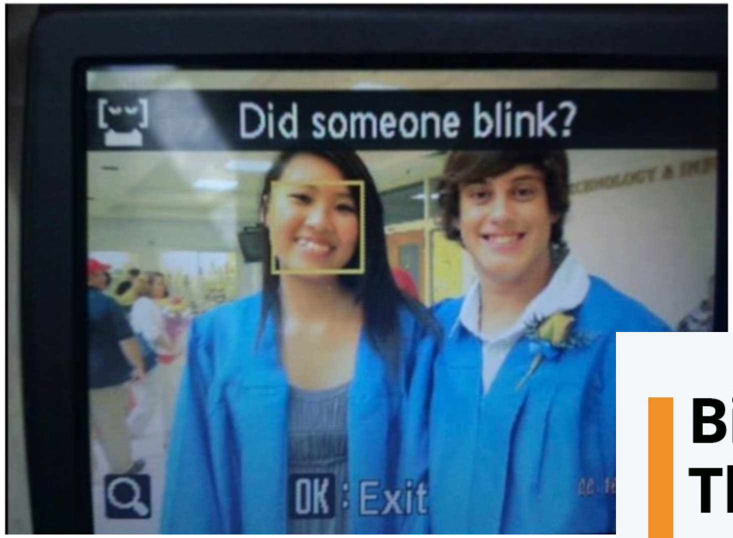
Better
healthcare

Can we use the
affordances of ML to help?

Smart grids

Story telling

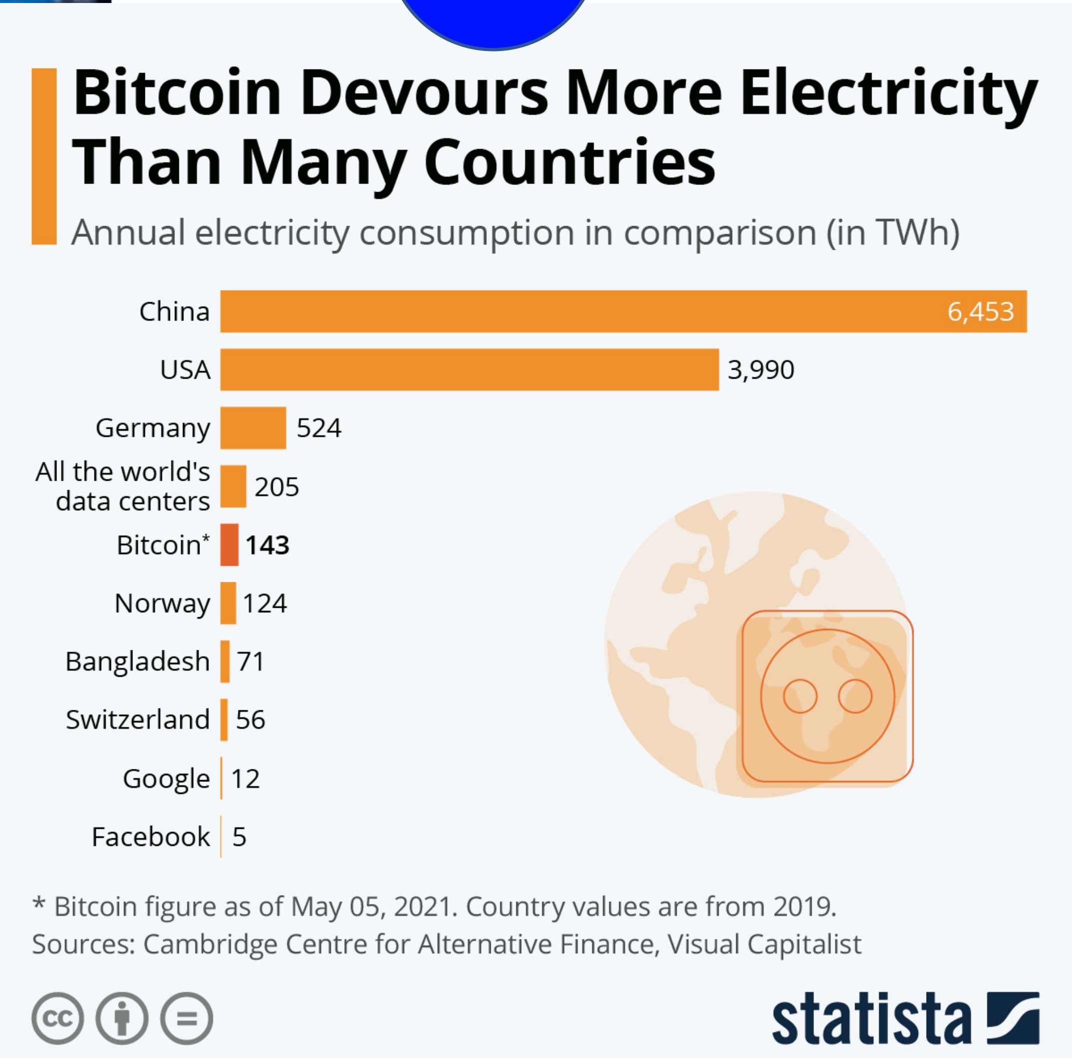
1



2

Facebook slammed by UN for its role in Myanmar genocide

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

The New York Times

Why Stanford Researchers Tried to Create a 'Gaydar' Machine

Other learning goal: how to be a responsible scientist (and not show up in the news in a negative way)

New Concepts from philosophy / ethics

- What is a Protected demographic?
- Distributive Harm vs Quality of Service Harm
- What is fairness?
  - Philosophy of procedural vs distributive
  - Different definitions of fairness

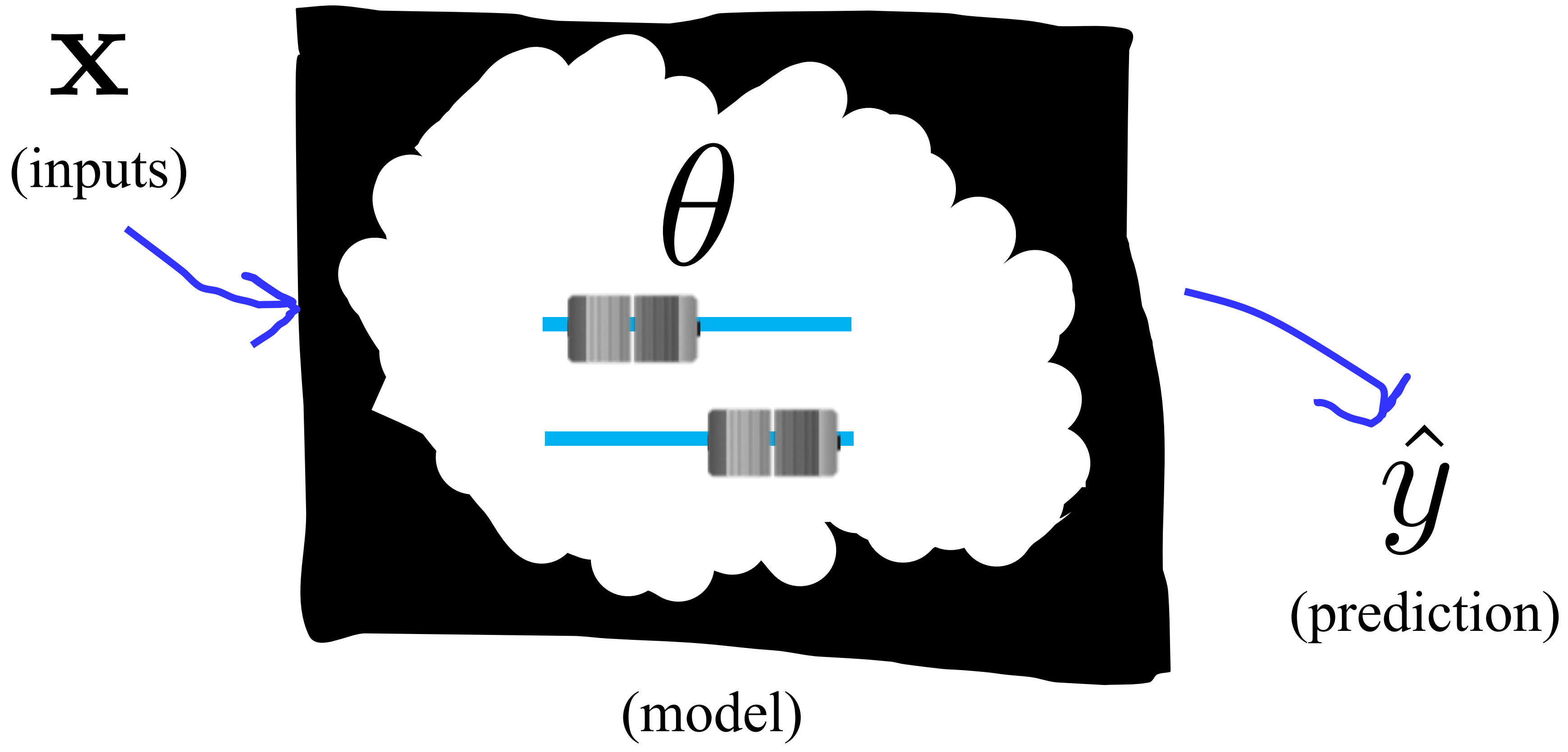
Warning: Many of the examples in this lecture come from the US. Much of ML fairness research has been done in the US (AFAIK)

Is that a meta bias?

Every community has axis of discrimination – framing goes beyond the US and beyond racism, sexism.

Part 0: Review

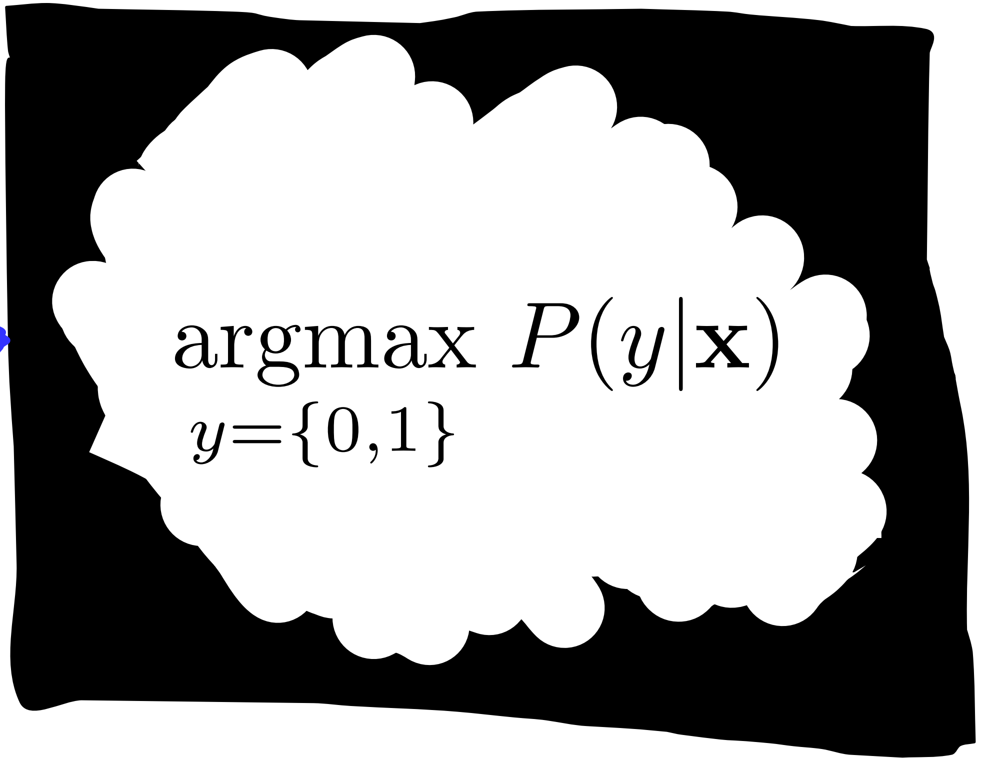
Machine Learning



Classification Algorithms

$\mathbf{x}$

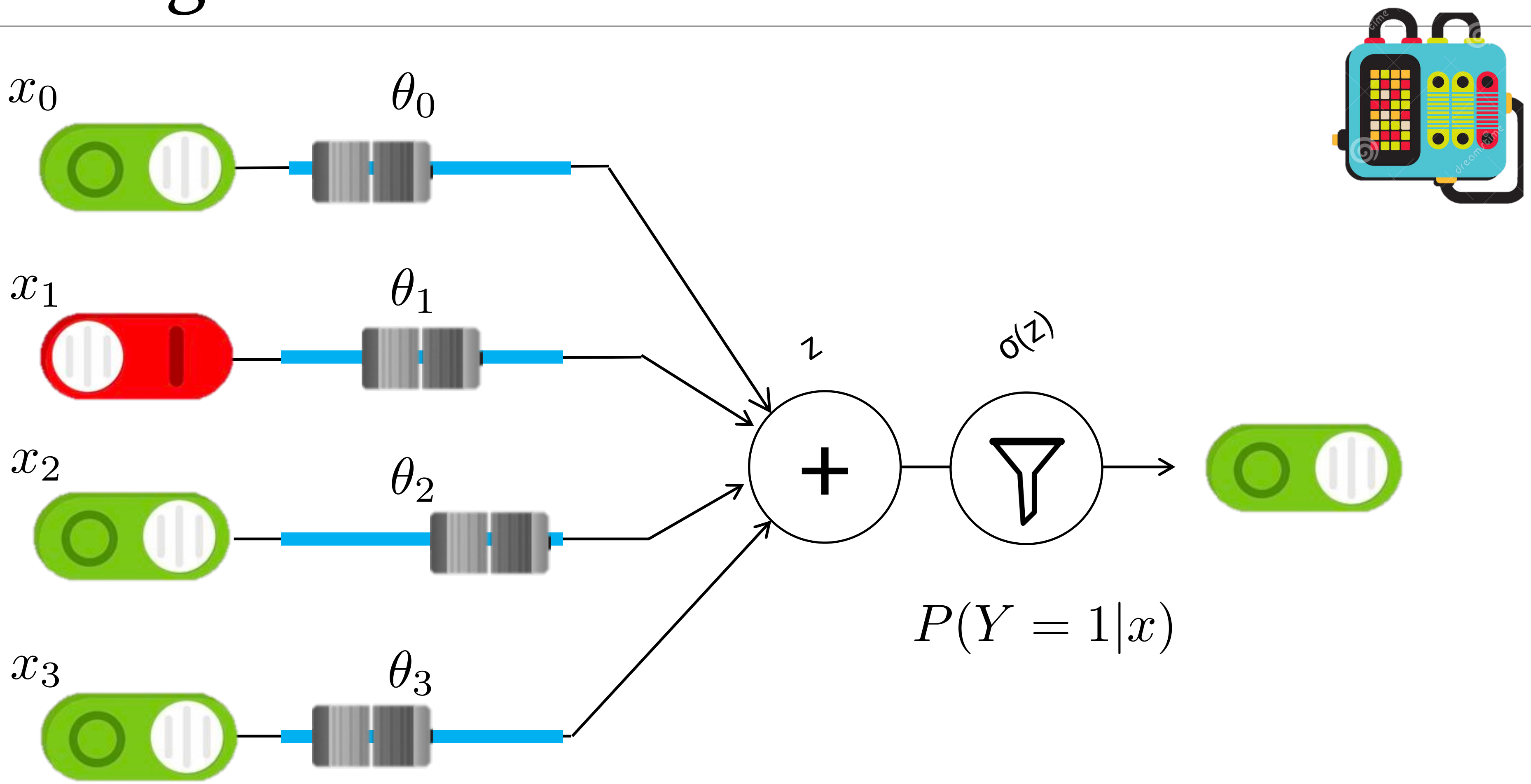
[0, 1, 1, 0]



$\hat{y} = 0$

Making a prediction…

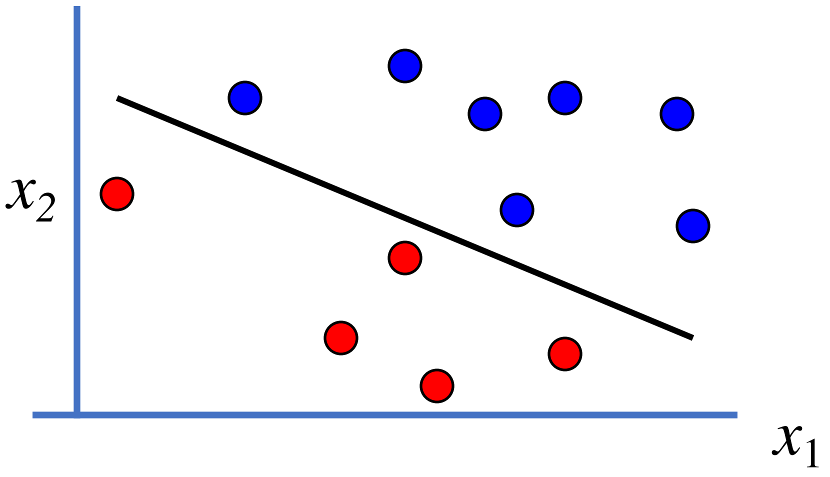
Logistic Regression



$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \sigma\left(\sum_i \theta_i x_i\right)$$

Turns out Logistic Regression is a Linear Classifier

- Logistic regression is trying to fit a **line** that separates data instances where y = 1 from those where y = 0



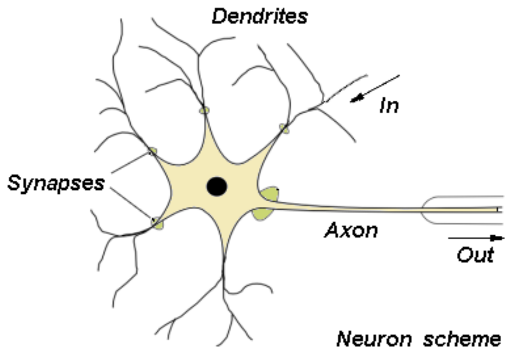
$$\theta^T \mathbf{x} = 0$$

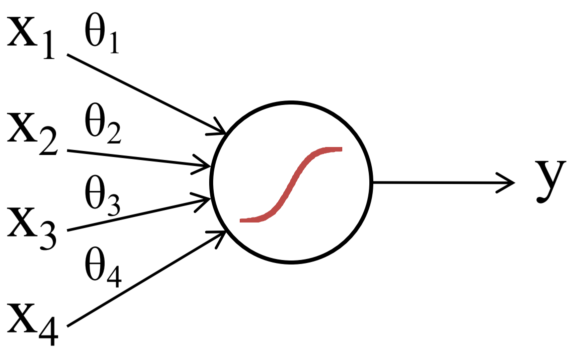
$$\theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_m x_m = 0$$

- We call such data (or the functions generating the data) "**linearly separable**"
- **Naïve bayes is linear too** as there is no interaction between different features.

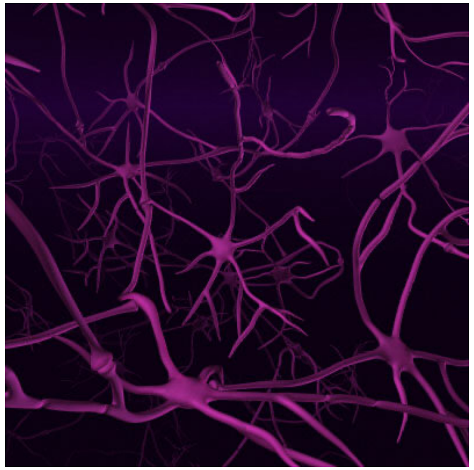
Deep Learning: Logistic Regression Can Be Stacked

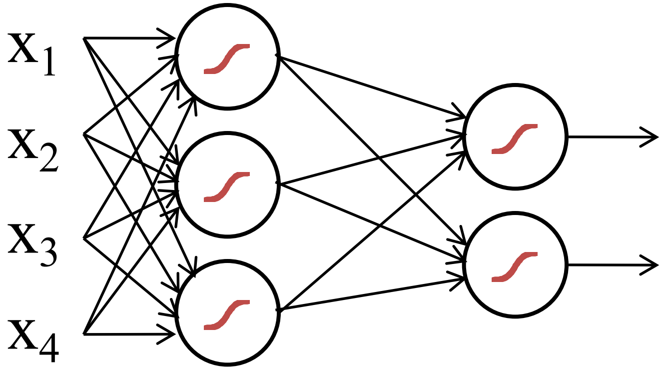
A neuron





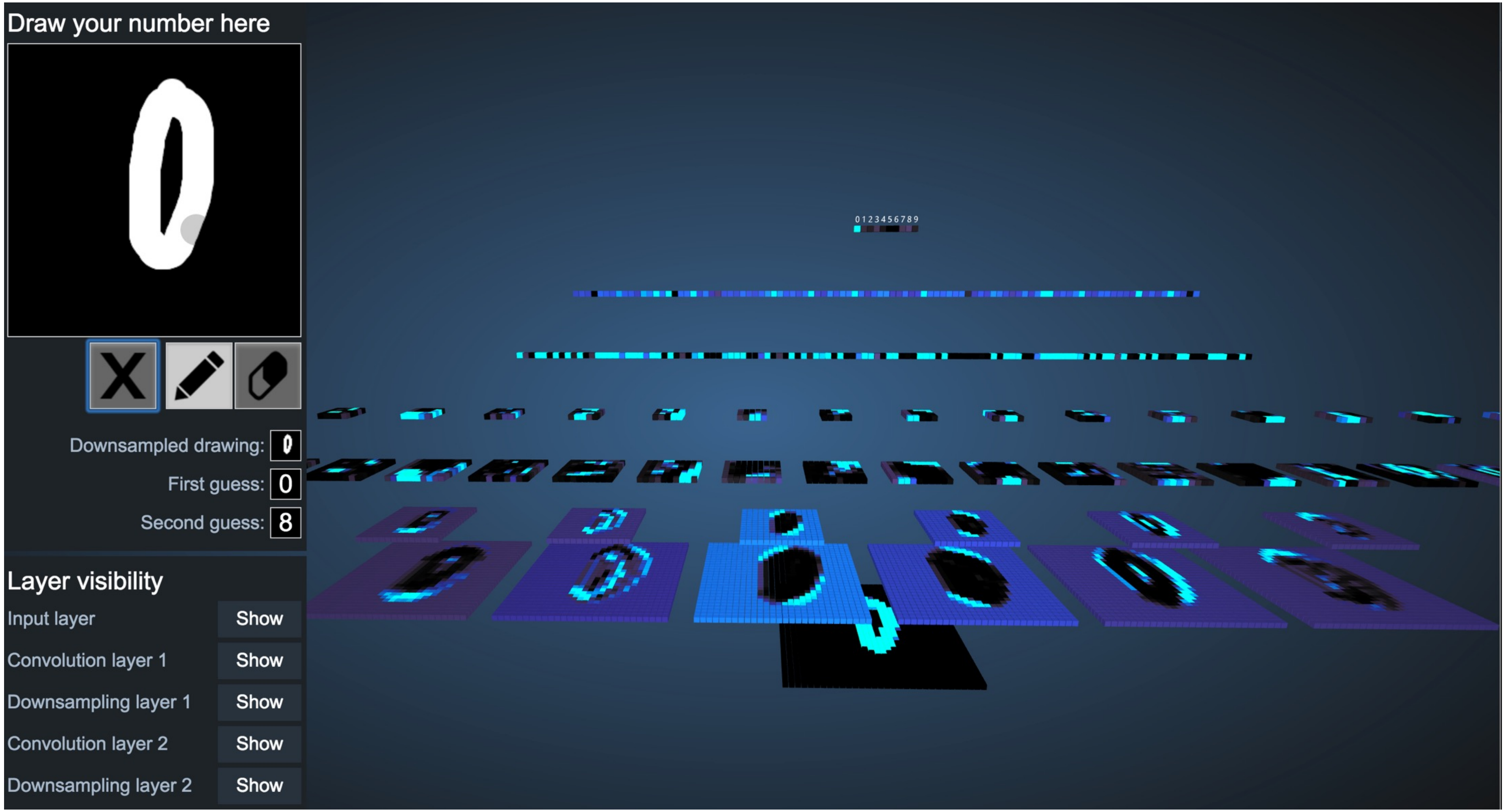
Your brain





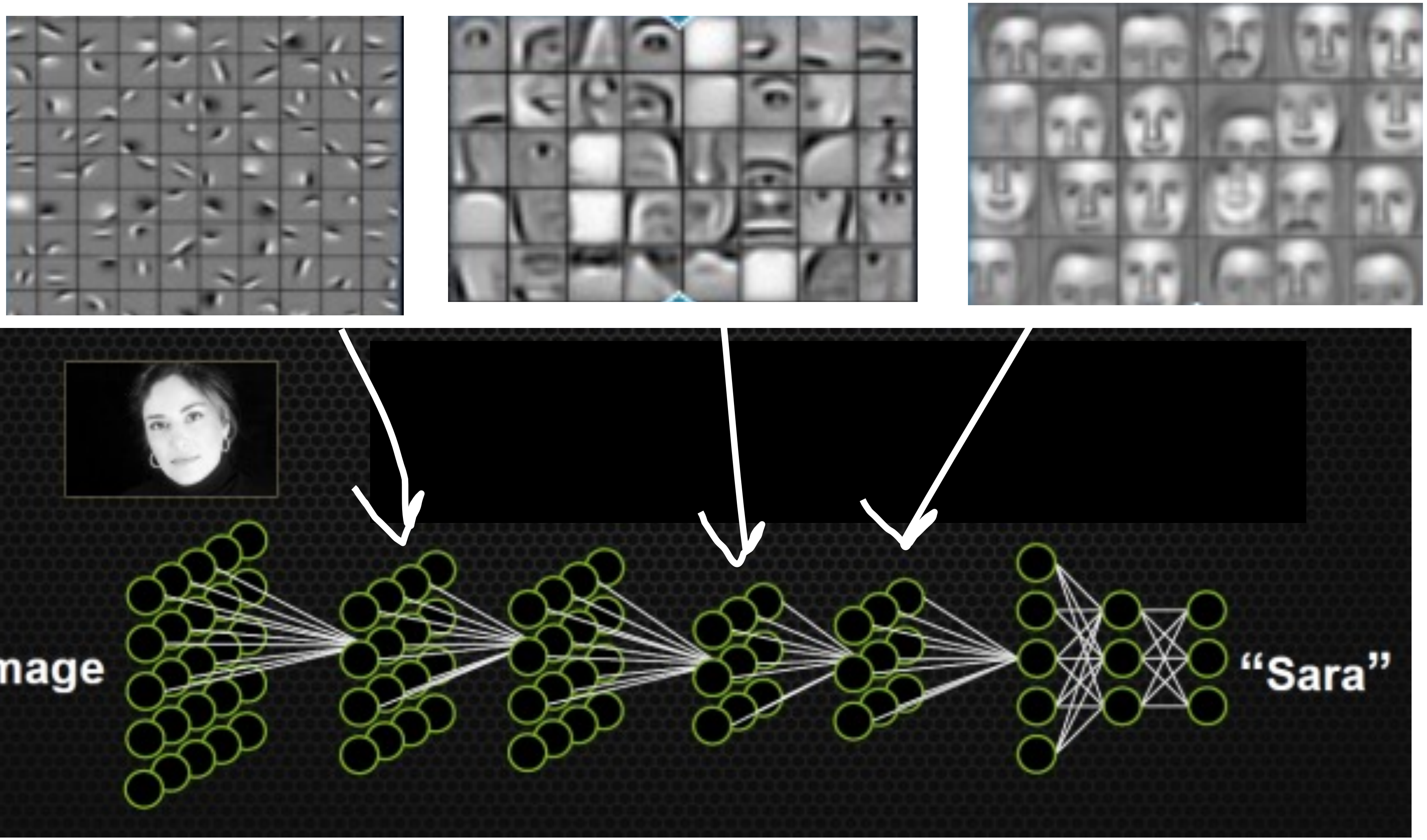
Actually, it's probably someone else's brain

Logistic Regression for Image Classification



http://scs.ryerson.ca/~aharley/vis/conv/

Works for any number of layers



$\hat{y}$

Part 1: Framework of Harm

Quality of Service Harms

Quality-of-service harms

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

Examples:

- Generative Art
- Face Recognition
- Document Search
- Product Recommendation

Distributive Harms

Quality-of-service harms

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

Examples:

- Generative Art
- Face Recognition
- Document Search
- Product Recommendation

Distributive harms

Occur when AI systems extend or withhold opportunities, resources, or information

Examples:

- Hiring
- Lending
- School admissions

Existential Harms?

Quality-of-service harms

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

Examples:

- Generative Art
- Face Recognition
- Document Search
- Product Recommendation

Distributive harms

Occur when AI systems extend or withhold opportunities, resources, or information

Examples:

- Hiring
- Lending
- School admissions

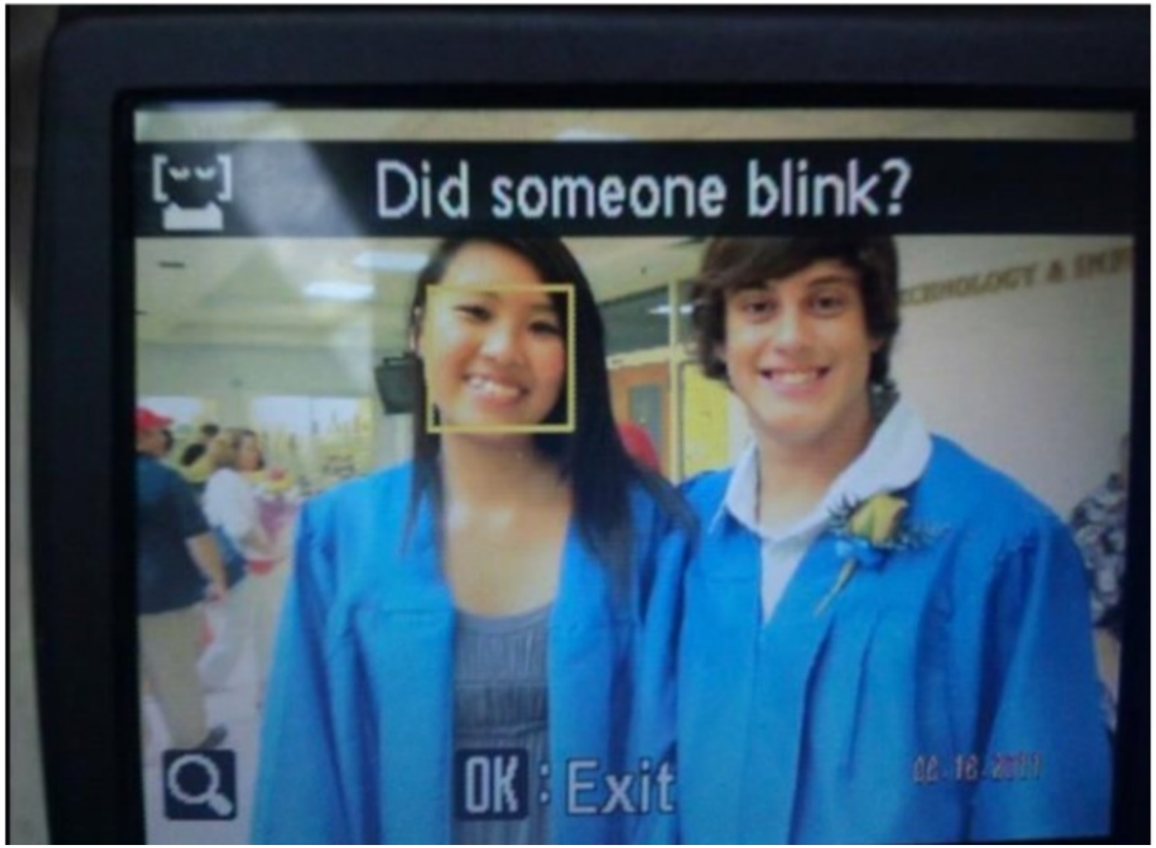
Existential harms

Maybe you will just break the whole damn thing

Examples:

- Genocide?
- Democracy?
- Climate?
- AI Supremacy?

Quality of Service Harm example



How did this happen?

ImageNet classification

22,000 categories

14,000,000 images

Hand-engineered features (SIFT, HOG, LBP), Spatial pyramid, SparseCoding/Compression

...
smoothhound, smoothhound shark, Mustelus mustelus
American smooth dogfish, Mustelus canis
Florida smoothhound, Mustelus norrisi
whitetip shark, reef whitetip shark, Triaenodon obseus
Atlantic spiny dogfish, Squalus acanthias
Pacific spiny dogfish, Squalus suckleyi
hammerhead, hammerhead shark
smooth hammerhead, Sphyrna zygaena
smalleye hammerhead, Sphyrna tudes
shovelhead, bonnethead, bonnet shark, S
angel shark, angelfish, Squatina squatina, monkfish
electric ray, crampfish, numbfish, torpedo
smalltooth sawfish, Pristis pectinatus
guitarfish
roughtail stingray, Dasyatis centroura
butterfly ray
eagle ray
spotted eagle ray, spotted ray, Aetobatus narinari
cownose ray, cow-nosed ray, Rhinoptera bonasus
manta, manta ray, devilfish
Atlantic manta, Manta birostris
devil ray, Mobula hypostoma
grey skate, gray skate, Raja batis
little skate, Raja erinacea
...

Stingray

Mantaray

Le, et al., Building high-level features using large-scale unsupervised learning. ICML 2012

ImageNet classification challenge

~~22,000 categories~~

1000 categories

14,000,000 images

1,200,000 images in train set

200,000 images in test set

Hand-engineered features (SIFT, HOG, LBP),
Spatial pyramid,
SparseCoding/Compression

moothhound shark, Mustelus mustelus
n dogfish, Mustelus canis
Florida smoothhound, Mustelus norrisi
odon obseus
smooth hammerhead, Sphyrna zygaena
smalleye hammerhead, Sphyrna tudes
shovelhead, bonnethead, bonnet shark, Sphyrna tiburo
angel shark, angelfish, Squatina squatina, monkfish
electric ray, crampfish, numbfish, torpedo
smalltooth sawfish, Pristis pectinatus
guitarfish
roughtail stingray, Dasyatis centroura
butterfly ray
eagle ray
spotted eagle ray, spotted ray, Aetobatus narinari
cownose ray, cow-nosed ray, Rhinoptera bonasus
manta, manta ray, devilfish
Atlantic manta, Manta birostris
devil ray, Mobula hypostoma
grey skate, gray skate, Raja batis
little skate, Raja erinacea
...

Le, et al., Building high-level features using large-scale unsupervised learning. ICML 2012

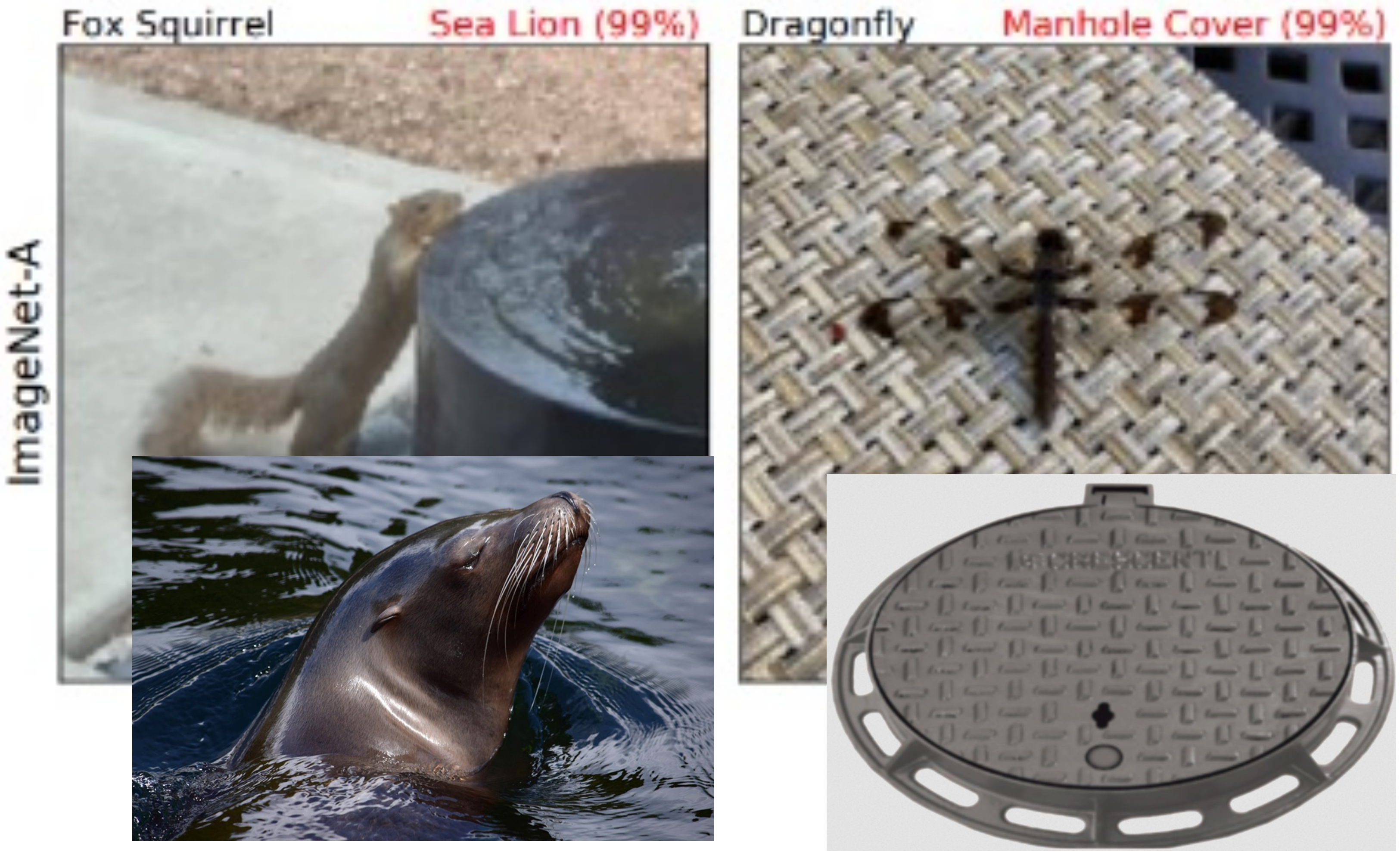
Biases in ImageNet

Imagenet is biased (in a neutral sense) towards texture ...

Hendrycks et. al. 202

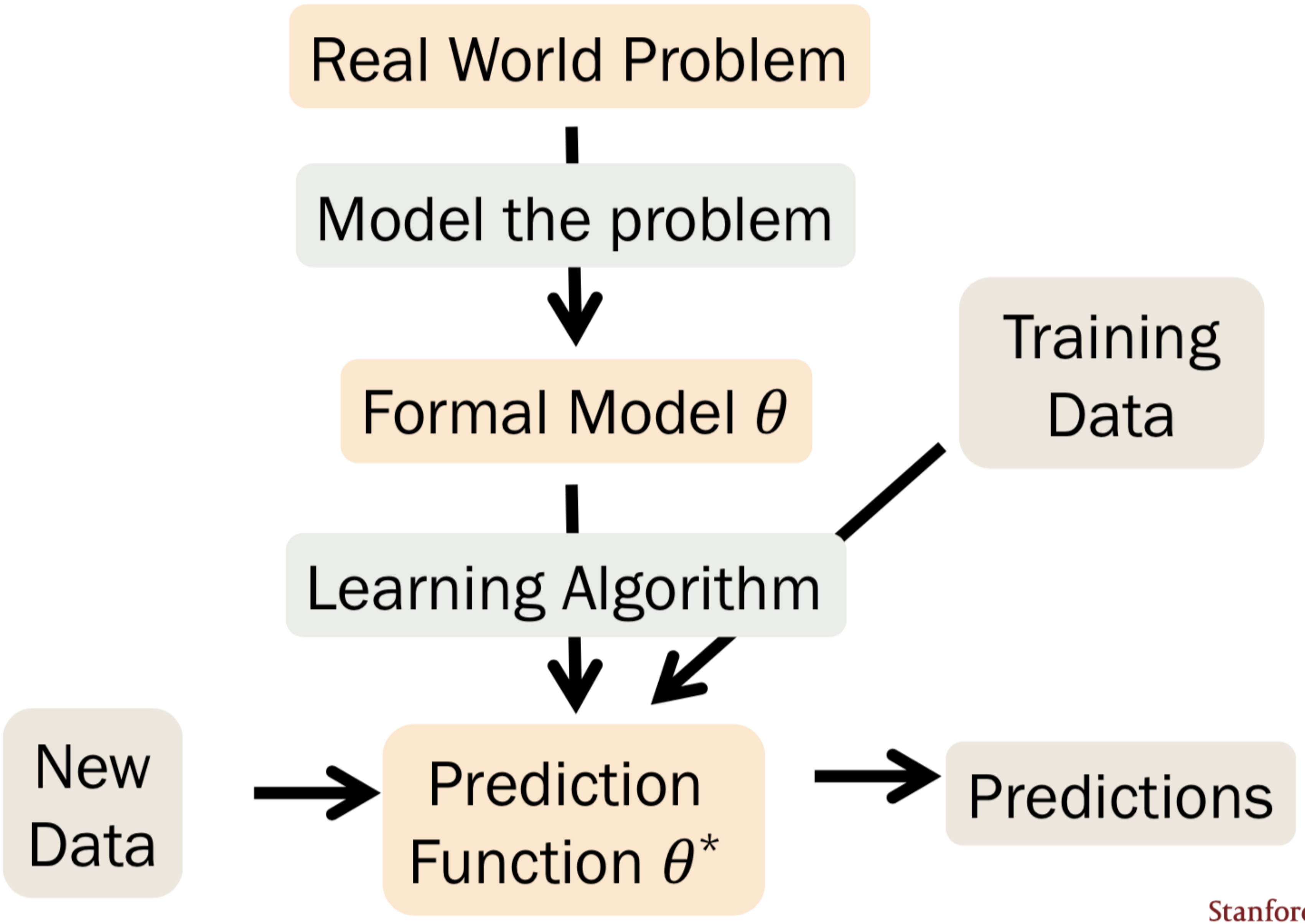
Biases in ImageNet

Imagenet is biased (in a neutral sense) towards texture …

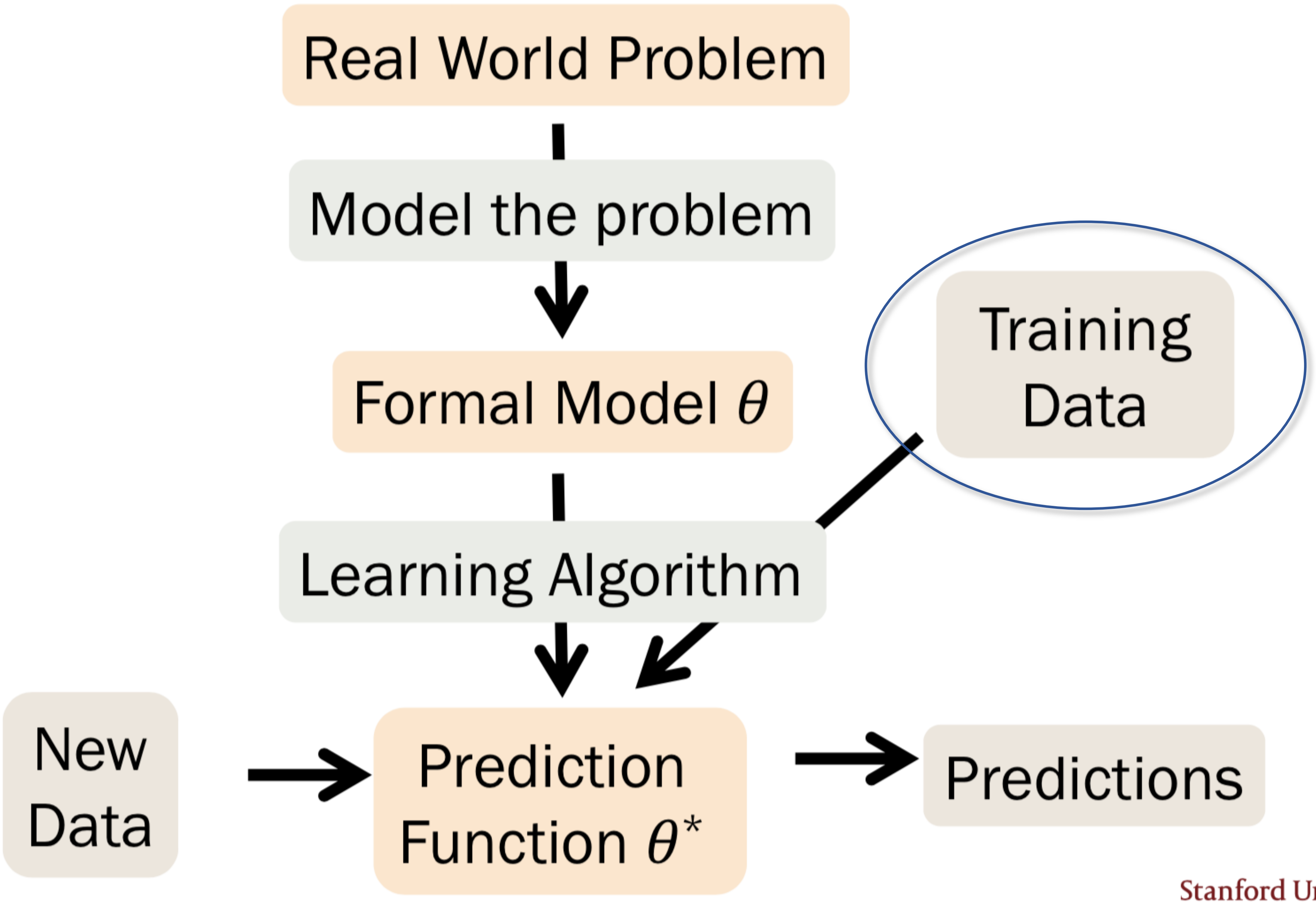


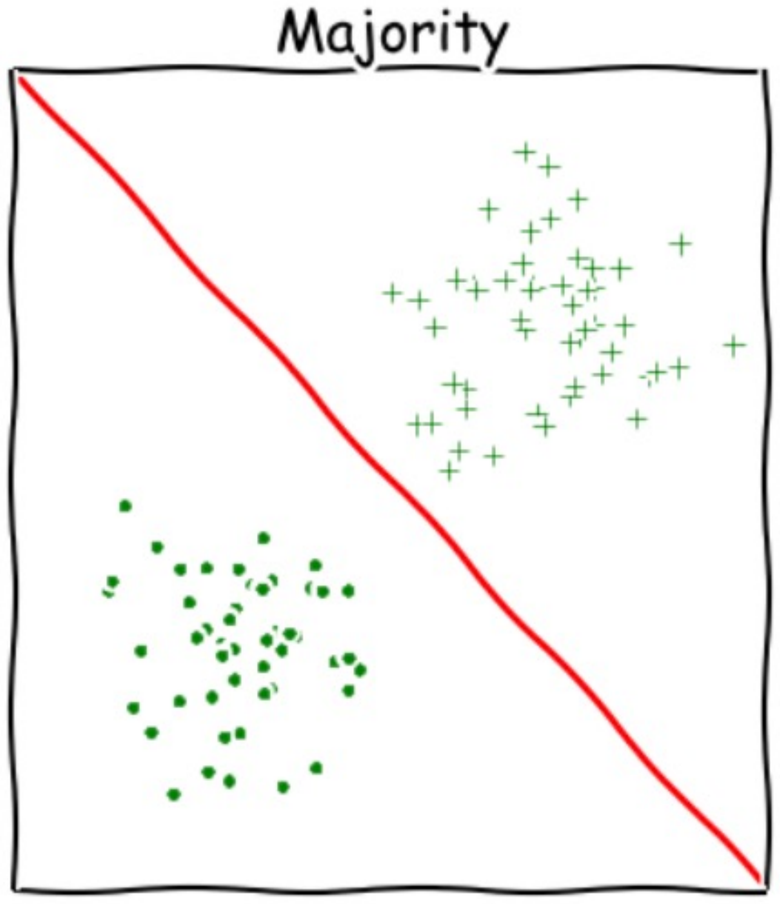
Hendrycks et. al. 202

Machine Learning

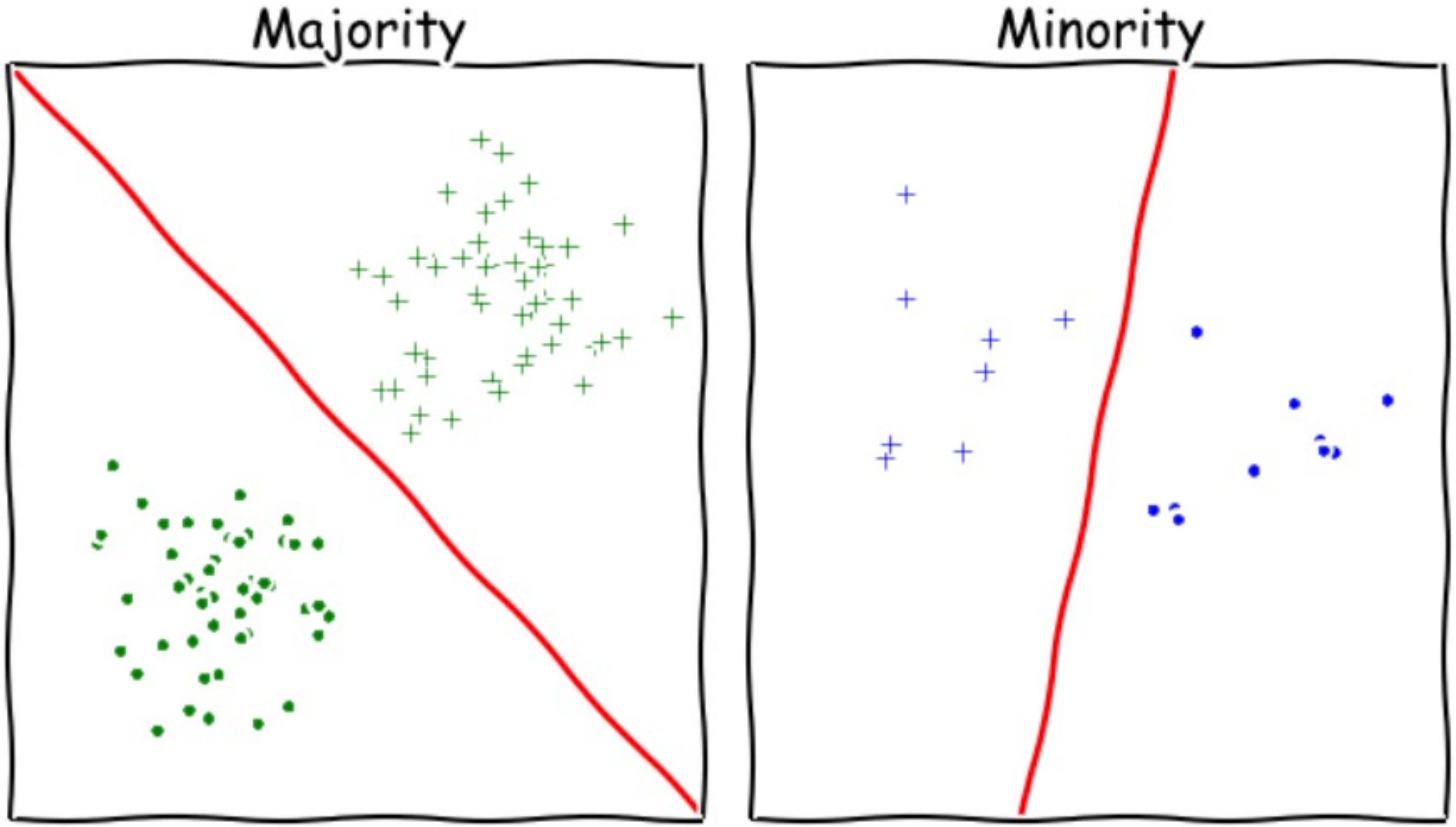


Machine Learning

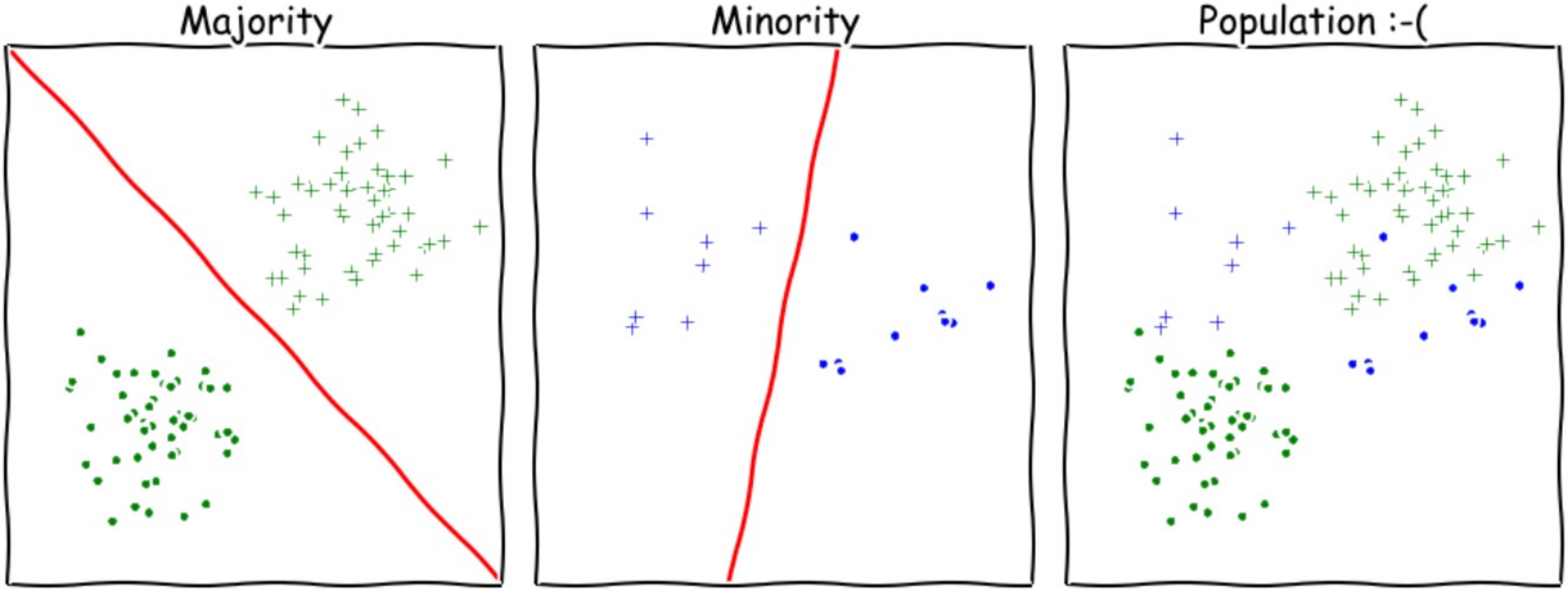




Classification of the minority group may be worse.



Classification of the minority group may be worse.



Classification of the minority group may be worse …
even with “awareness” or “stereotyping.”

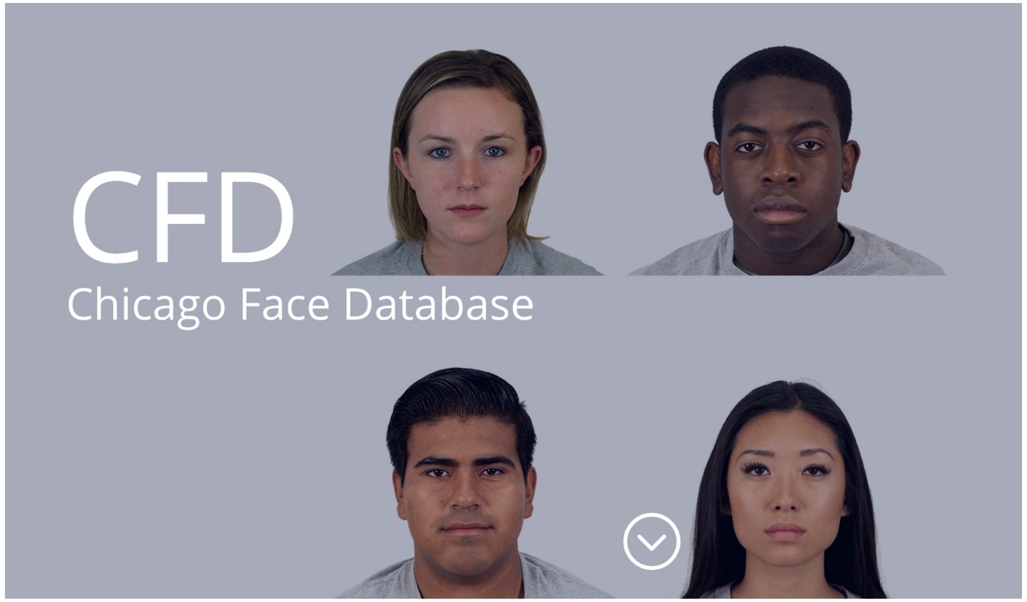
Problem 1: Undersampling & Lack of Data

- For both gender and race, the majority groups are often undersampled in image databases.
- Majority of images in some databases of faces are of white faces.
- Faces In The Wild database was 83.5% white and 77.5% male.

Huge Improvement in Face Datasets in 2018

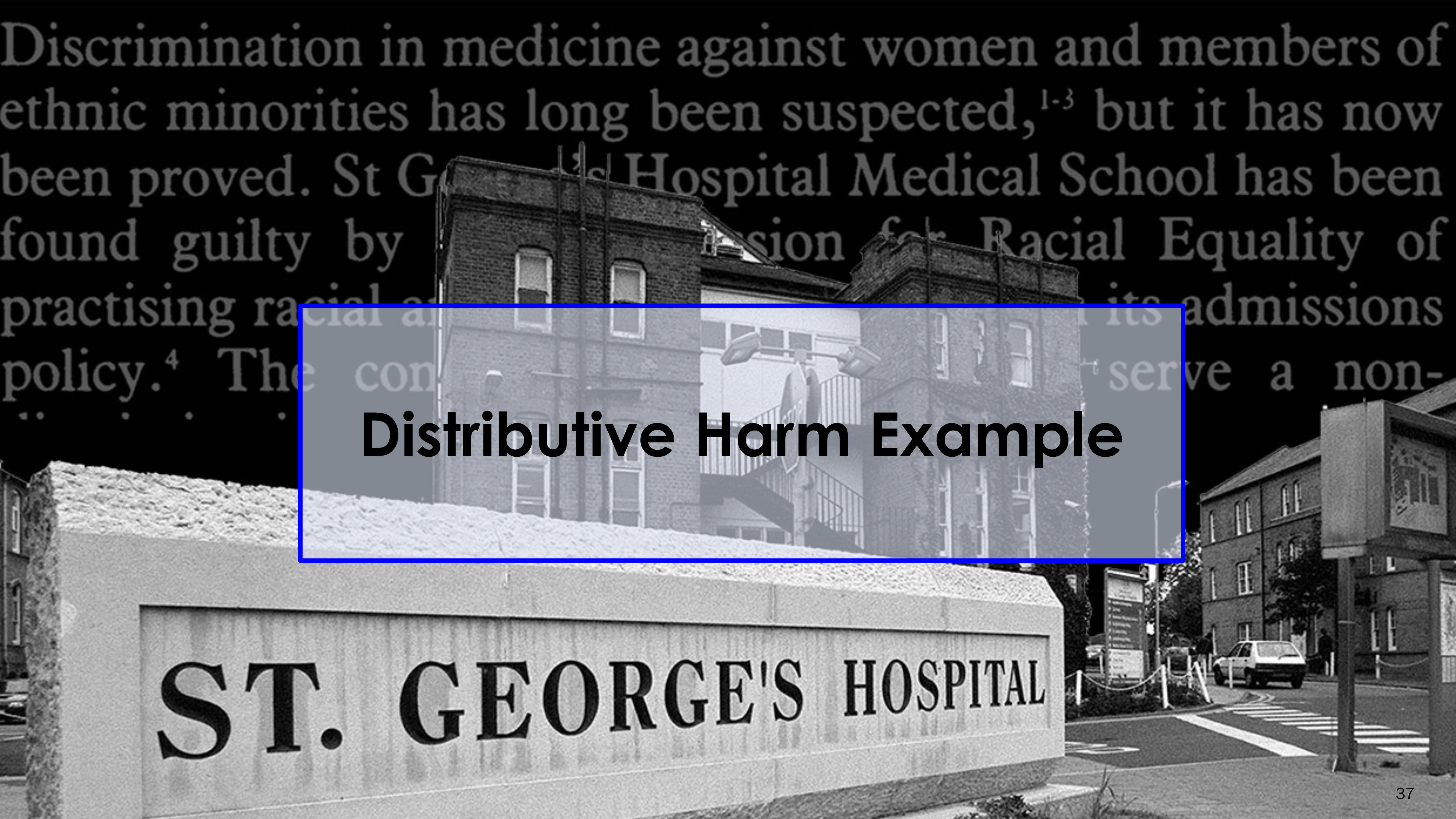
Research and activism by Joy Buolamwini, Timnit Gebru, and many others has led to more representative datasets already.

Figure 12. Sample Images from Pilot Parliaments Benchmark



Stanford PhD 2017

Distributive Harm Example



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

2,500 applicants to the medical school

Interview approx. 625 (so ¾ are rejected)

Offer spots to approx. 425 (so 70% of interviewees accepted)

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

2,500 applicants to the medical school

Interview approx. 625 (so ¾ are rejected)

Offer spots to approx. 425 (so 70% of interviewees accepted)

In 1979, Vice Dean Dr. Geoffrey Franglen finishes a classification algorithm to do the job

Timeline of a Biased Algorithm

- 1982: Dr. Franglen argues that 90-95% of classifications agree with the verdict of human assessors on the selection panel
- 1982: Algorithm trained on historical data from St. George’s screens all applications
- Internal review questions why applicants are being weighted by factors like name and place of birth
- 1986: two St. George’s lecturers report findings to UK Commission for Racial Equality
- Commission finds that name and place of birth are used to dock points from female and “Non-Caucasian” applicants

Timeline of a Biased Algorithm

1982: Dr. Franglen argues that 90-95% of classifications agree with the verdict of human assessors on the selection panel

1982: Algorithm trained on historical data from St. George's

Internal review questions why applicants are being weighted by factors like name and place of birth

1986: two St. George's lecturers report findings to UK Commission

Commission finds that name and place of birth are used to dock points from female and "Non-Caucasian" applicants

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.

This biased result was predictable

Costs: At least 60 people wrongly rejected each year.

1. Garbage In, Garbage Out.

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.

Can we get formal about fairness?

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.

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” (subset of sensitive features) is legally required in domains like lending.

How to do it:

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

Protected Demographics

Protected Groups

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

Equal Employment Opportunity, USA

Similarly defined for housing, loans, etc

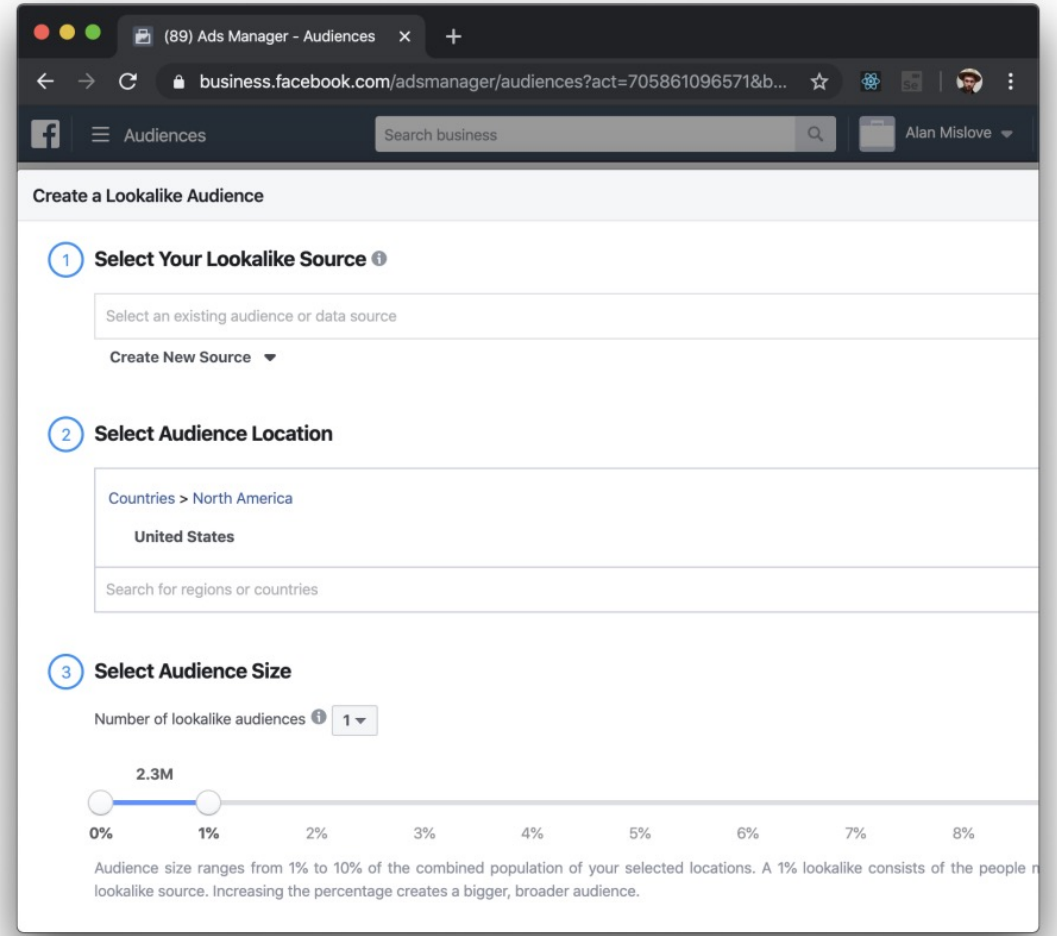
Case Study: Facebook Ads & Job/Housing Recommendations

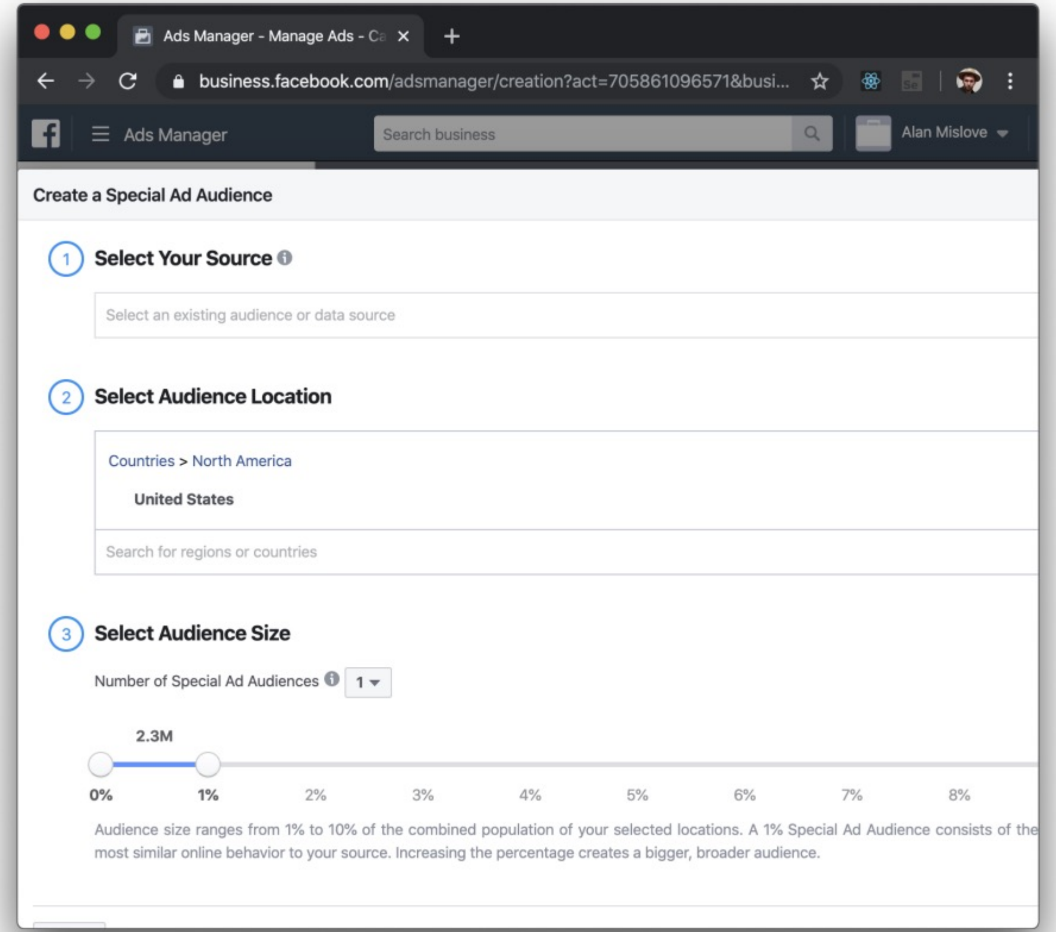
Facebook creates “Lookalike” feature for advertisers: upload a “source list” and find users with “common qualities” to target ads, including for housing and jobs

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

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 somewhat biased

Political Views: Less Biased

Sapiezynski et. al 2019,

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

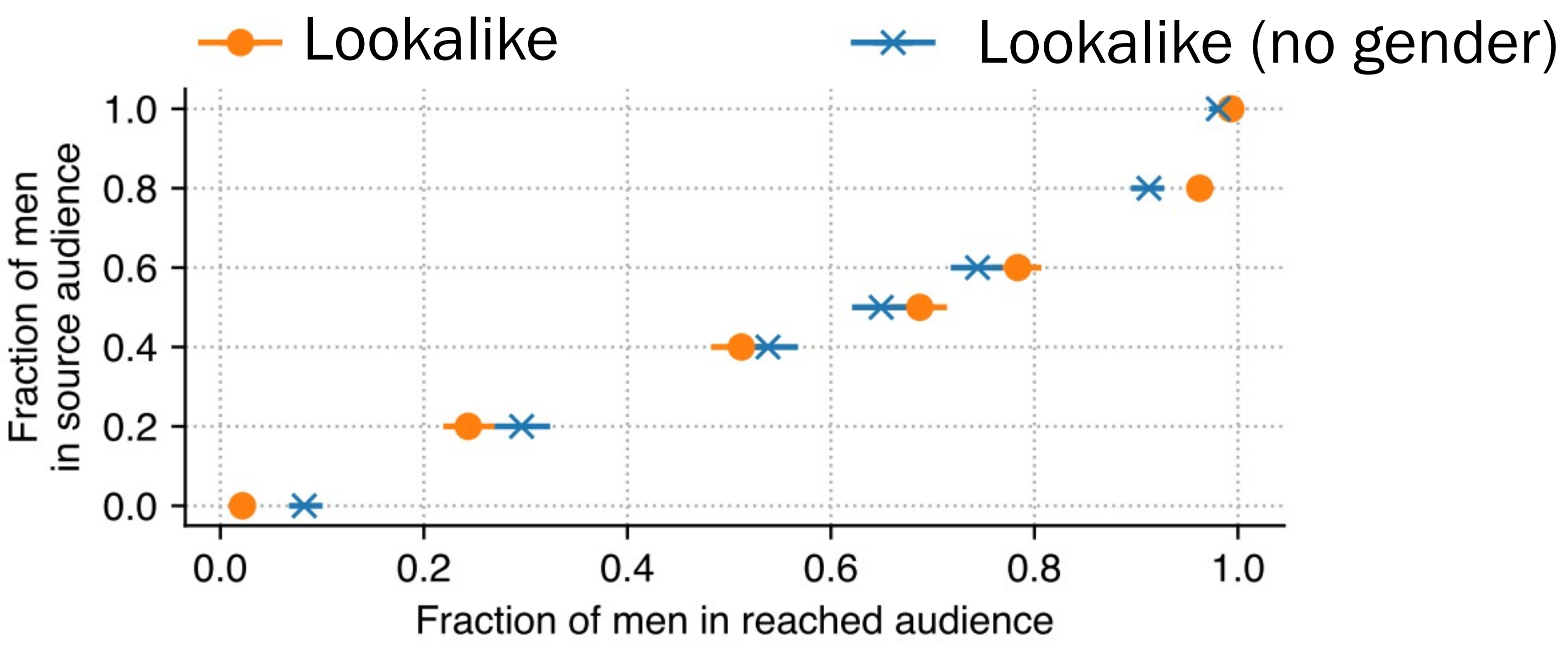
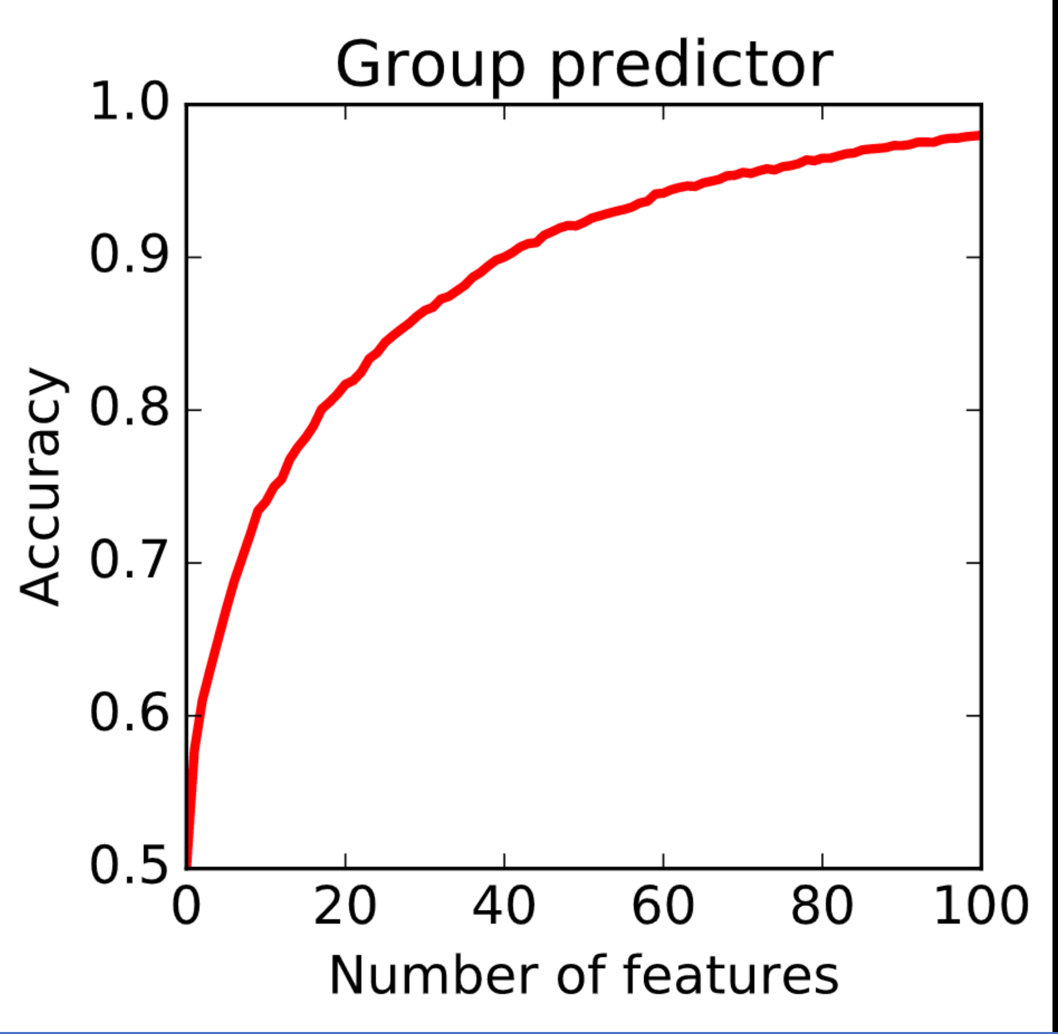


Figure 2: Gender breakdown of ad delivery to Lookalike and Special Ad audiences created from the same source audience with varying fraction of male users, using the same ad creative. We can observe that both Lookalike and Special Ad audiences reflect the gender distribution of the source audience, despite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

Yo, Piotr, you got your axis backwards ☺



Many Features = Accurate Group Prediction

Sensitive attributes are often “redundantly encoded” in the dataset

Many of the features or datapoints are correlated with the sensitive attribute

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 is hard)

Let’s Try Fairness Through Awareness!

Awareness of what?

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)

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

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)

$$P(G = T | D = 0) = P(G = T | D = 1)$$

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)

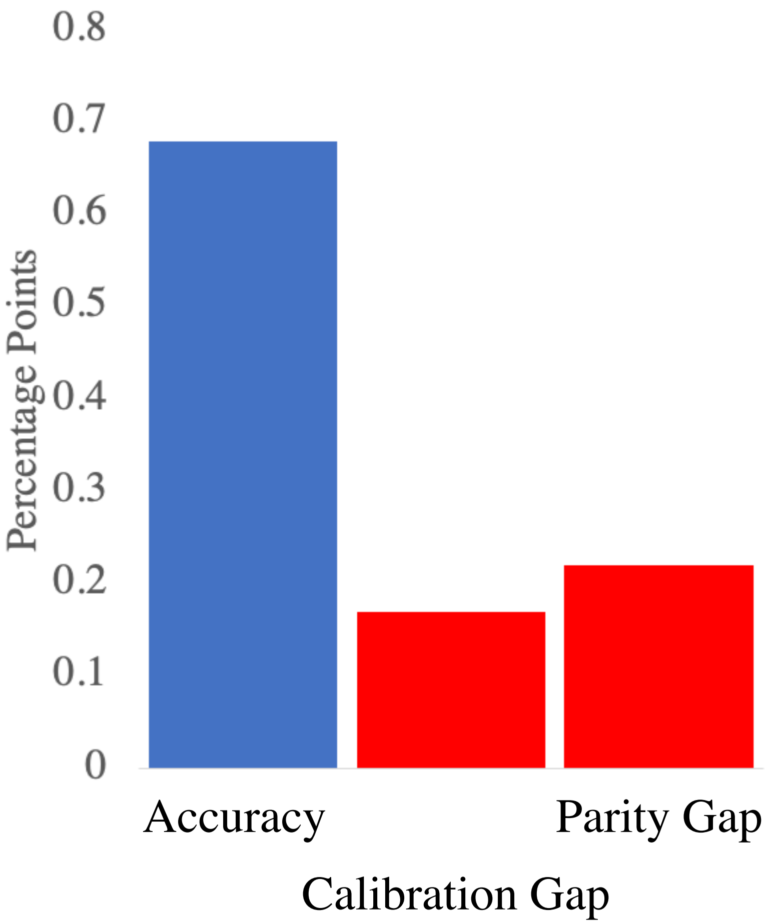
$$\frac{P(G = T | D = 1)}{P(G = T | D = 0)} \geq 1 - \epsilon$$

Where epsilon = 0.2

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

COMPAS: Biased Against Black Inmates

Before: Compas is Biased



Disparate Quality & Self-Fulfilling Properties

Dwork et. al. 2012, “Fairness Through Awareness” https://dl.acm.org/doi/10.1145/2090236.2090255

What does fairness through awareness fail to capture?

- If the classifier is significantly less good at identifying candidates e.g. for a surgery in a minority group (relative to the data), the candidates accepted might have worse outcomes, leading to future bias
- Quality of Service Disparity might then lead to an Allocation Disparity.
- Dwork et. al. (including Omer Reingold!) call this a “self-fulfilling prophecy.”

Part 3: What are you going to do about it?

Balanced Training Data

Transparent Reporting

Model Cards: A systematic checklist for investigating your model and sharing the results with others (Mitchell et. al. 2019)

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Train bias out

Advanced Idea: Adversarial Learning

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

Seniors at the time they wrote it

COMPAS: Predicting “Recidivism”

$\mathbf{x}$

Data about an inmate:
Their zip code, past crimes, etc

$$\underset{y=\{0,1\}}{\operatorname{argmax}} \; P(y|\mathbf{x})$$

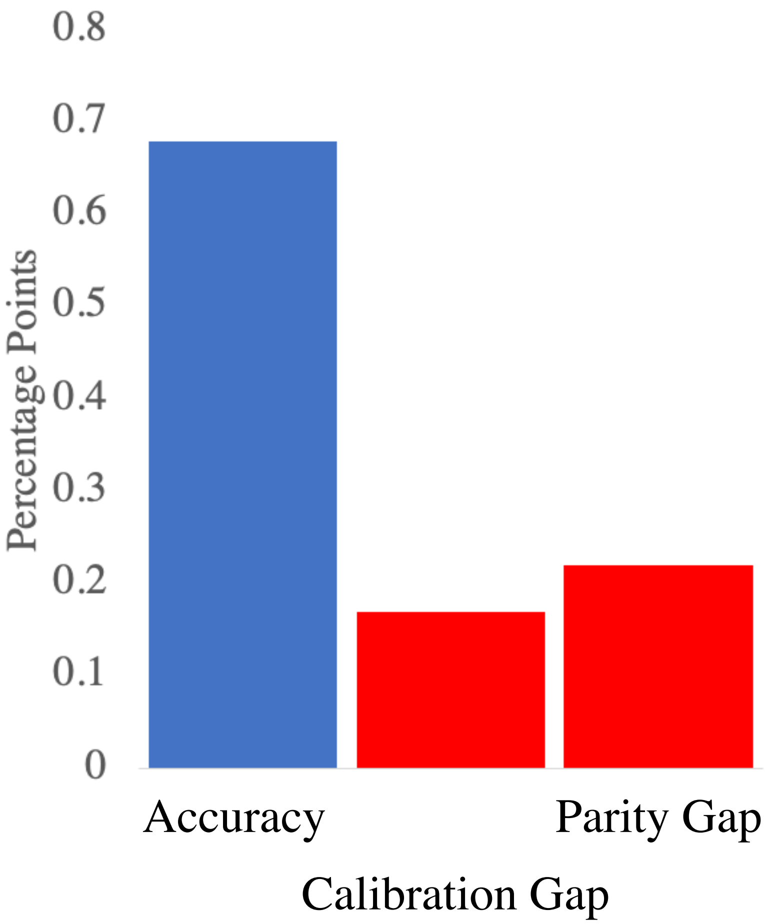
$\hat{y} = 0$

Will they commit a crime again

Was in use in California and Florida

COMPAS: Biased Against Black Inmates

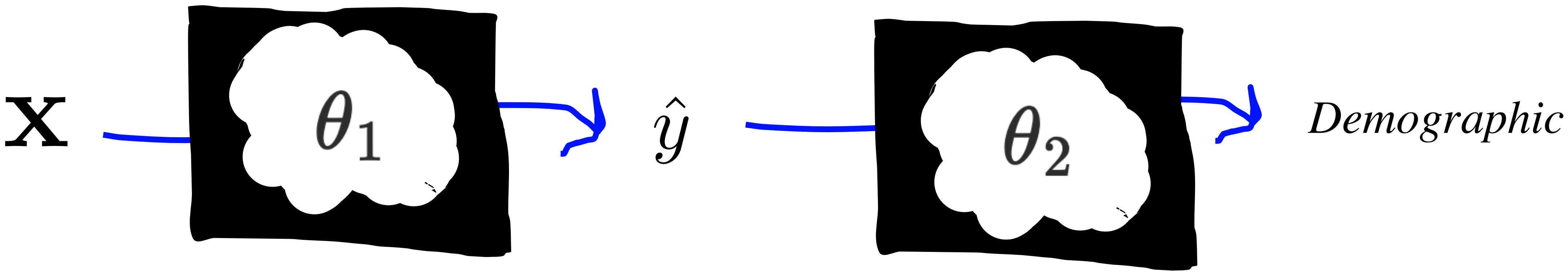
Before: Compas is Biased



Can We Train Out Bias?

Model 1: Prediction

Model 1: Extract Demographic



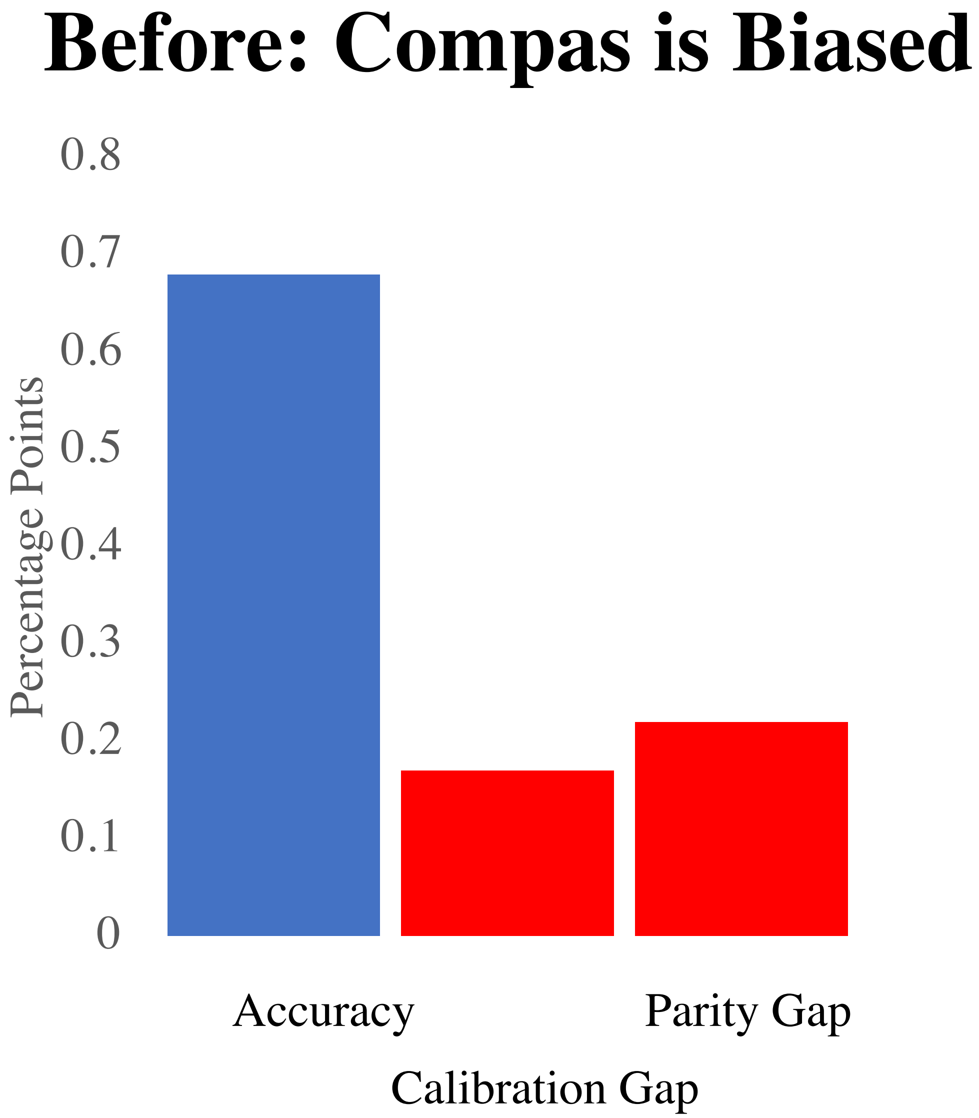
Model 1 should be accurate

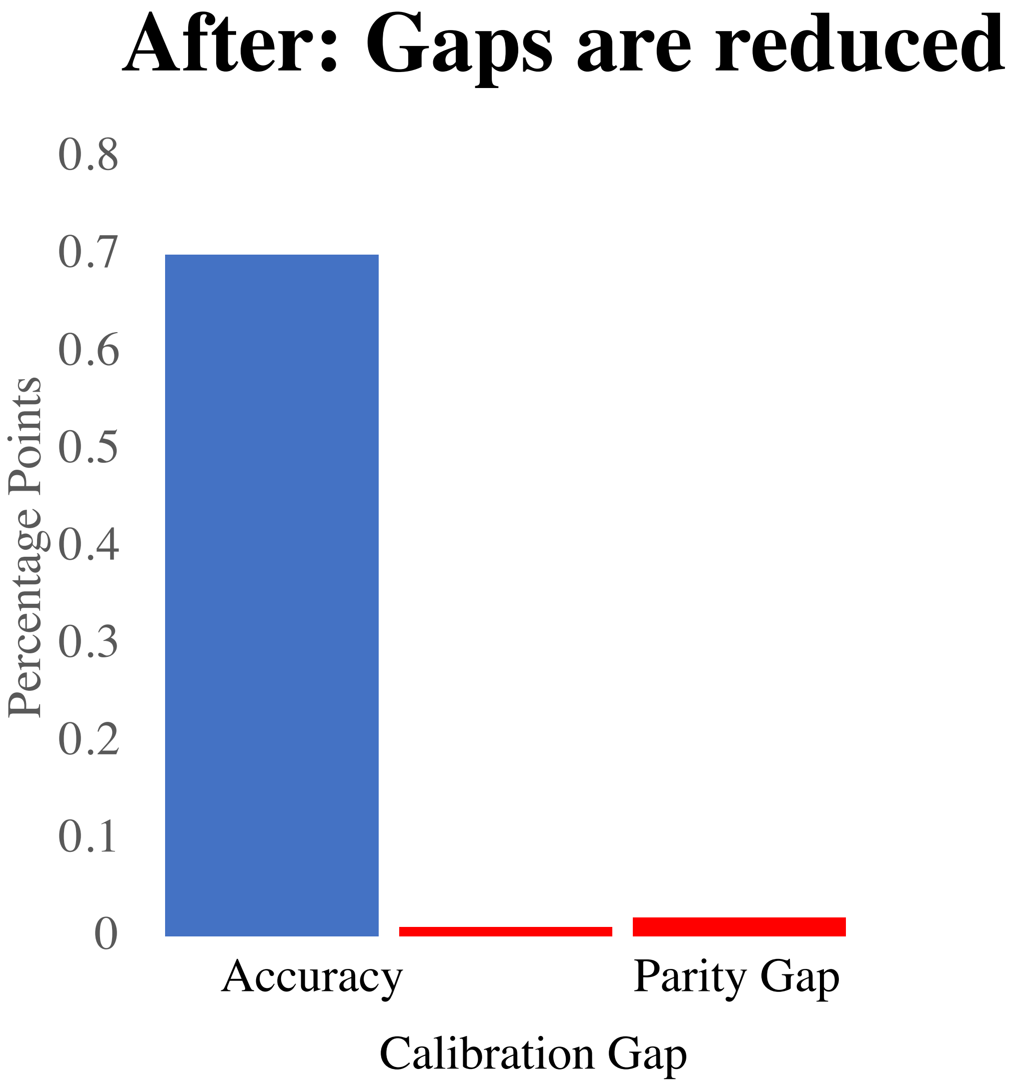
Model 2 should be ***in****accurate*

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

*note in the paper these were neural nets

Can We Train Out Bias?



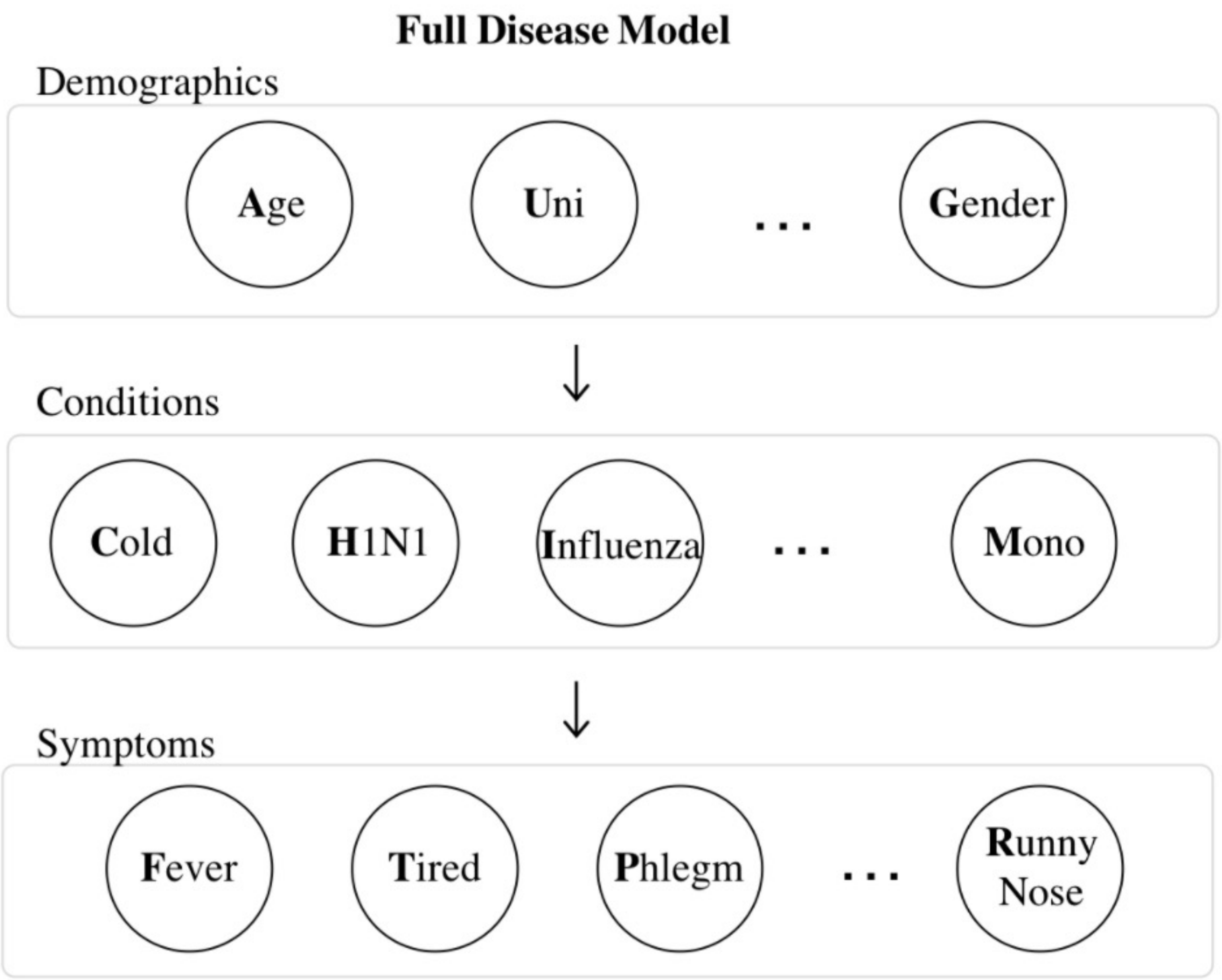


Their Conclusion

DON'T USE BLACK BOX ALGORITHMS TO MAKE RECIDIVISM PREDICTIONS

Use A Bayes Net?

Bayes Nets > Black Box?



Activism by Computer Scientists

Before #TechWontBuildIt

Retail Polaroid cameras had only one flash button, but the ID-2, sold to the South African government, had a second “boost” flash which increased the illumination by 42% to better capture Black skin tones.

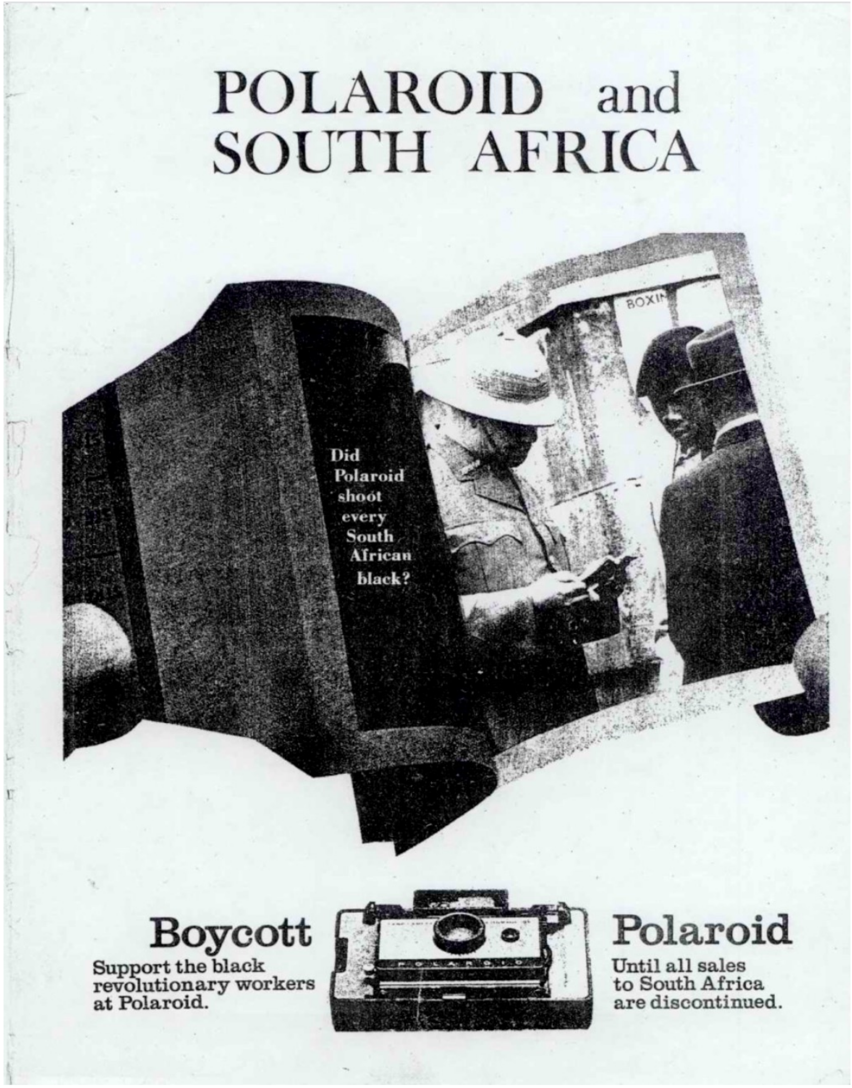
This was used to create passbook photographs for the Apartheid government.

http://physical-electrical-digital.nyufasedtech.com/items/show/46

Workers at Polaroid Whistleblowing

Caroline Hunter: “I worked at Polaroid as a research chemist and my late husband Ken Williams was in the photo department producing advertisements for Polaroid, and one day I went to pick him up for lunch and we discovered an ID badge with a mockup of a black guy that we knew from Polaroid saying ‘Union of South Africa Department of the Mines’”

“We discovered that Polaroid was in South Africa and that they’d been there for quite some time, since 1938, and that they were actually the producers of the notorious passbook photographs which South Africans, black South Africans called their ‘handcuffs.'”



(Pedagogic Pause)

Learning Goals

1. Understand limits in fairness through unawareness

2. Know two ways to measure fairness

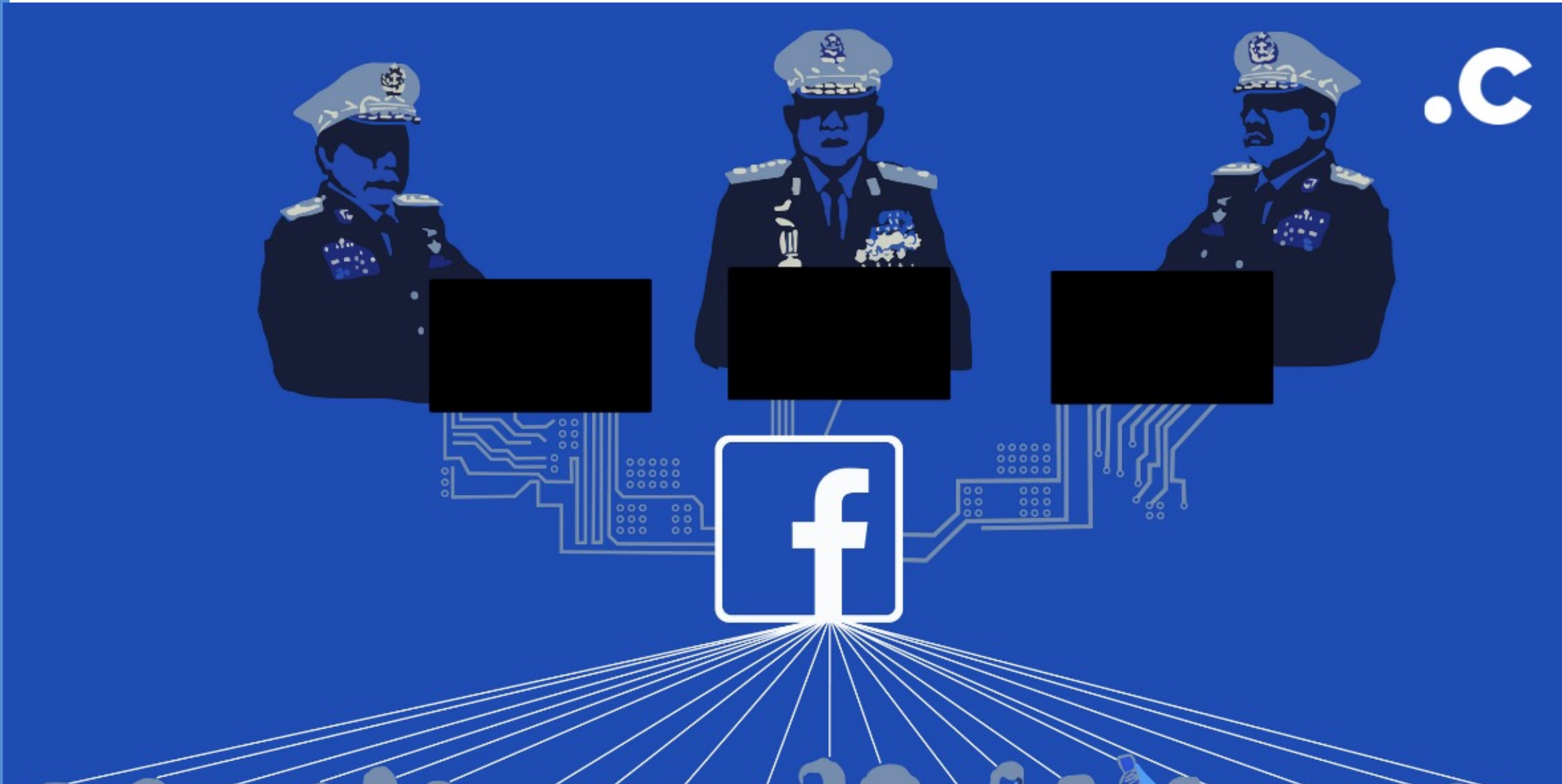
3. Know some techniques to mitigate fairness issues

Part 4: The Blind Spots

Well intentioned people can break things at scale (especially while moving fast)

Facebook Introduces Free Basics (2015)

Junta Starts a Misinformation Campaign Against Rohingya



Facebook: Two Moderators Who Speak Burmese (2015)

Genocide Against Rohingya Starts (2016)

Almost 1M Displaced

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

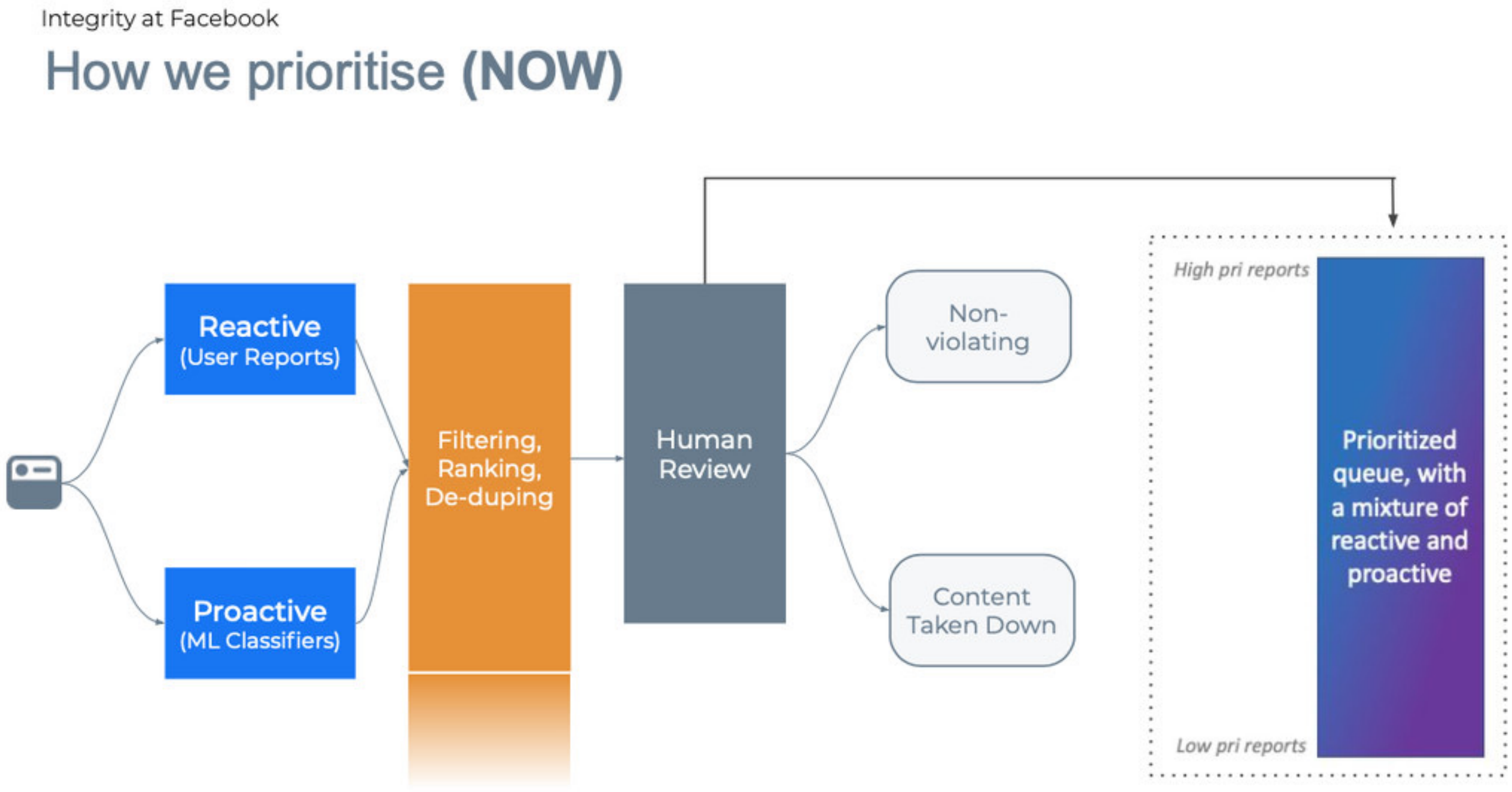
Report of the independent international fact-finding mission on Myanmar*

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.

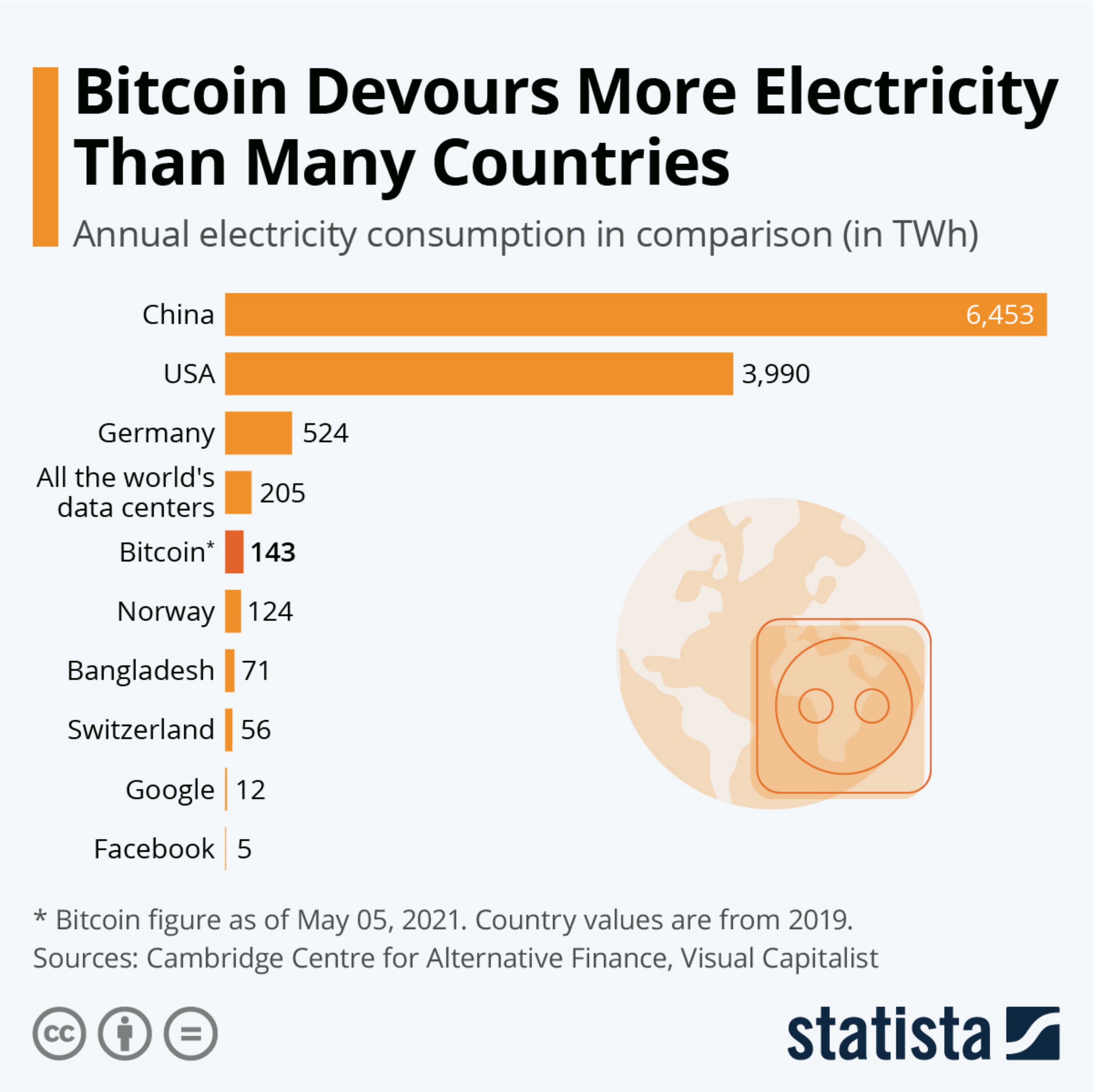
Silicon Valley's impact beyond the US was a major blind spot


MOVE
FAST AND
BREAK
THINGS

Aside: Facebook Says the Answer is Better ML



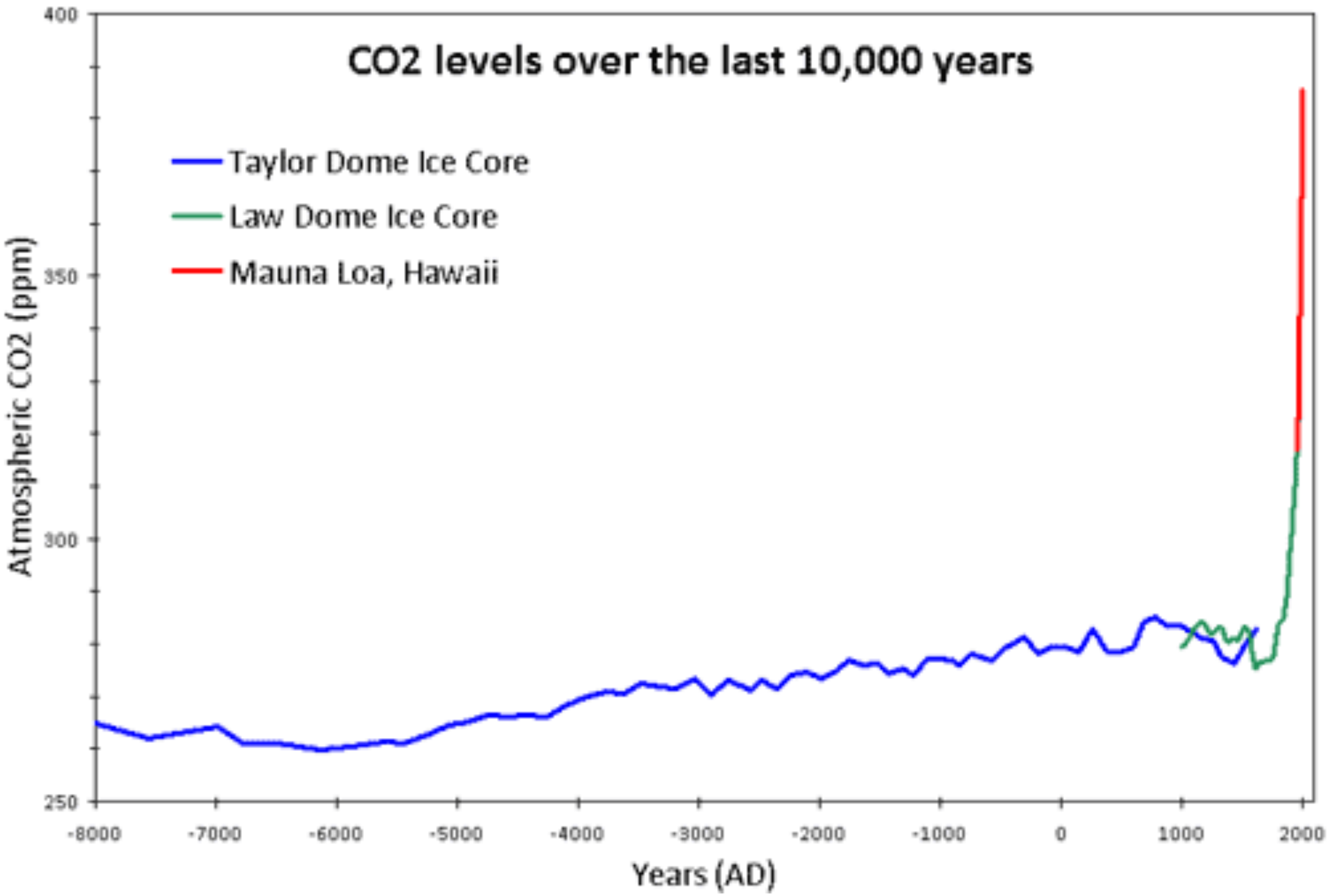
One Blind Spot I Want to Highlight



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

But climate change and bitcoin isn't even part of ethics at Stanford CS (I will update this slide once that changes)

It isn’t too hard to see the trend

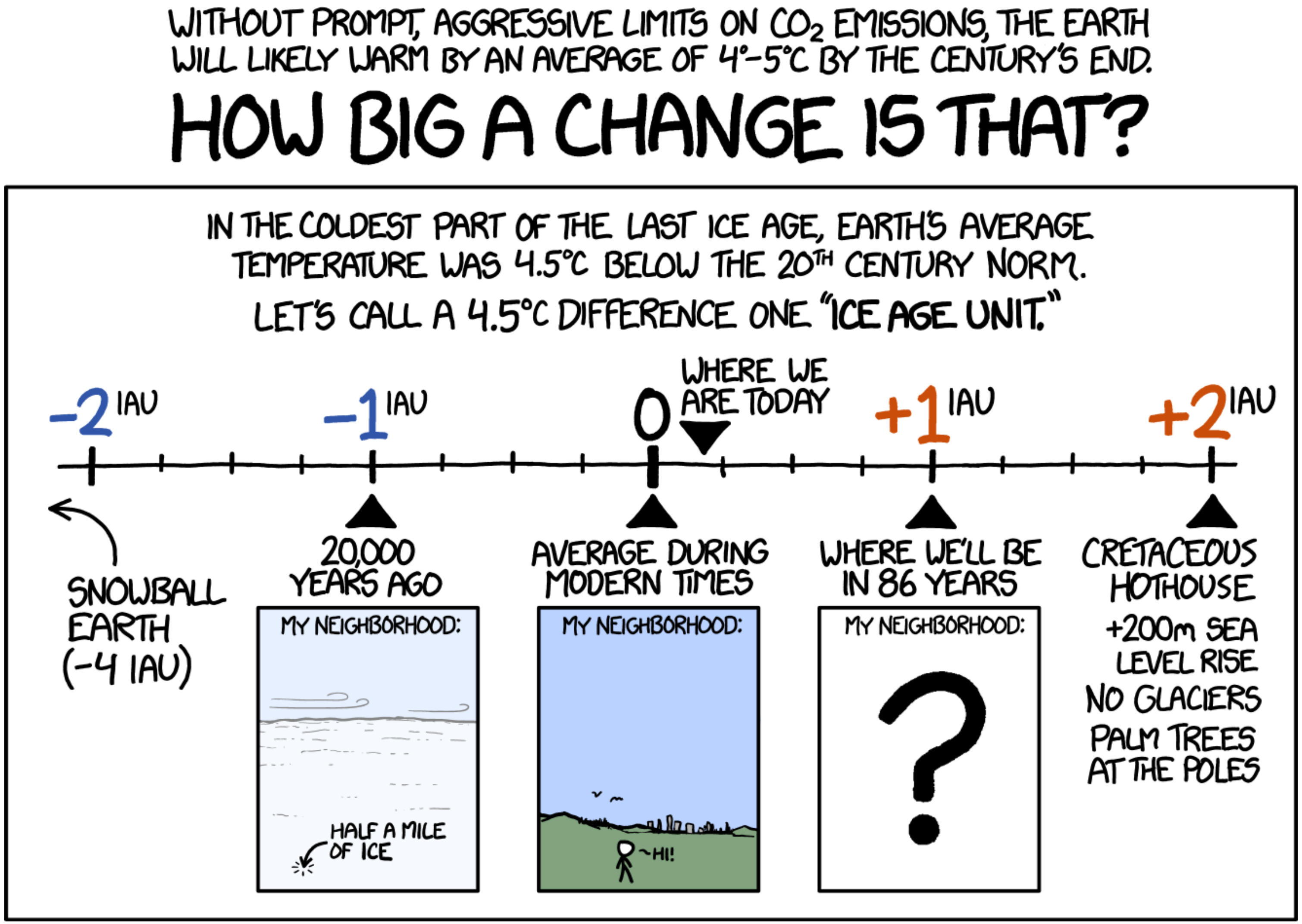


We will most almost certainly hit 2x CO2 before 2060, and then blow past it.

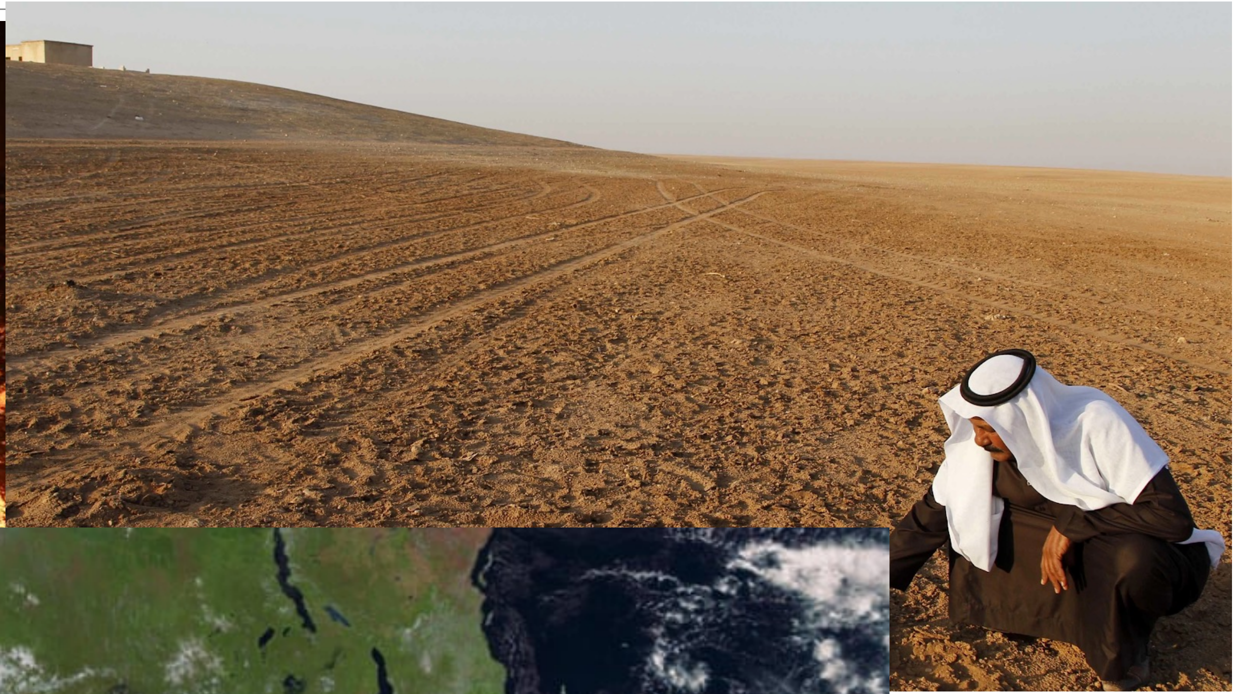
We know the physics

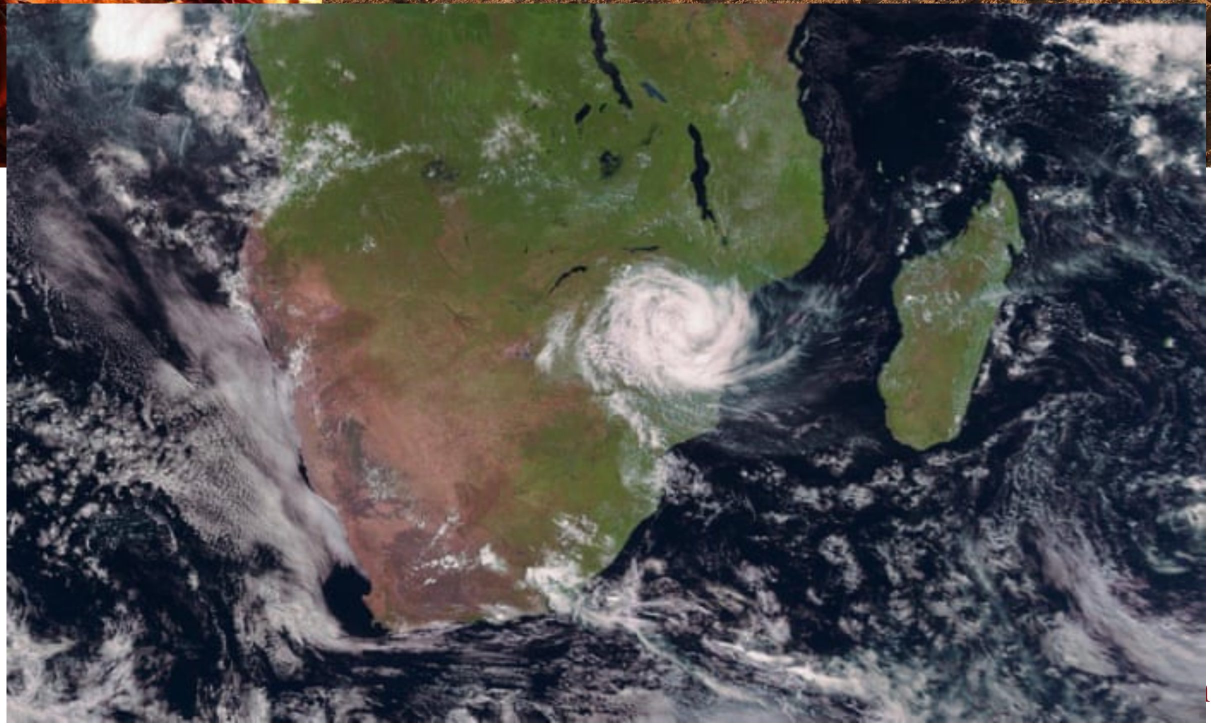
https://youtu.be/3v-w8Cyfoq8?t=39

Easy to Know Impacts Will Be Harsh



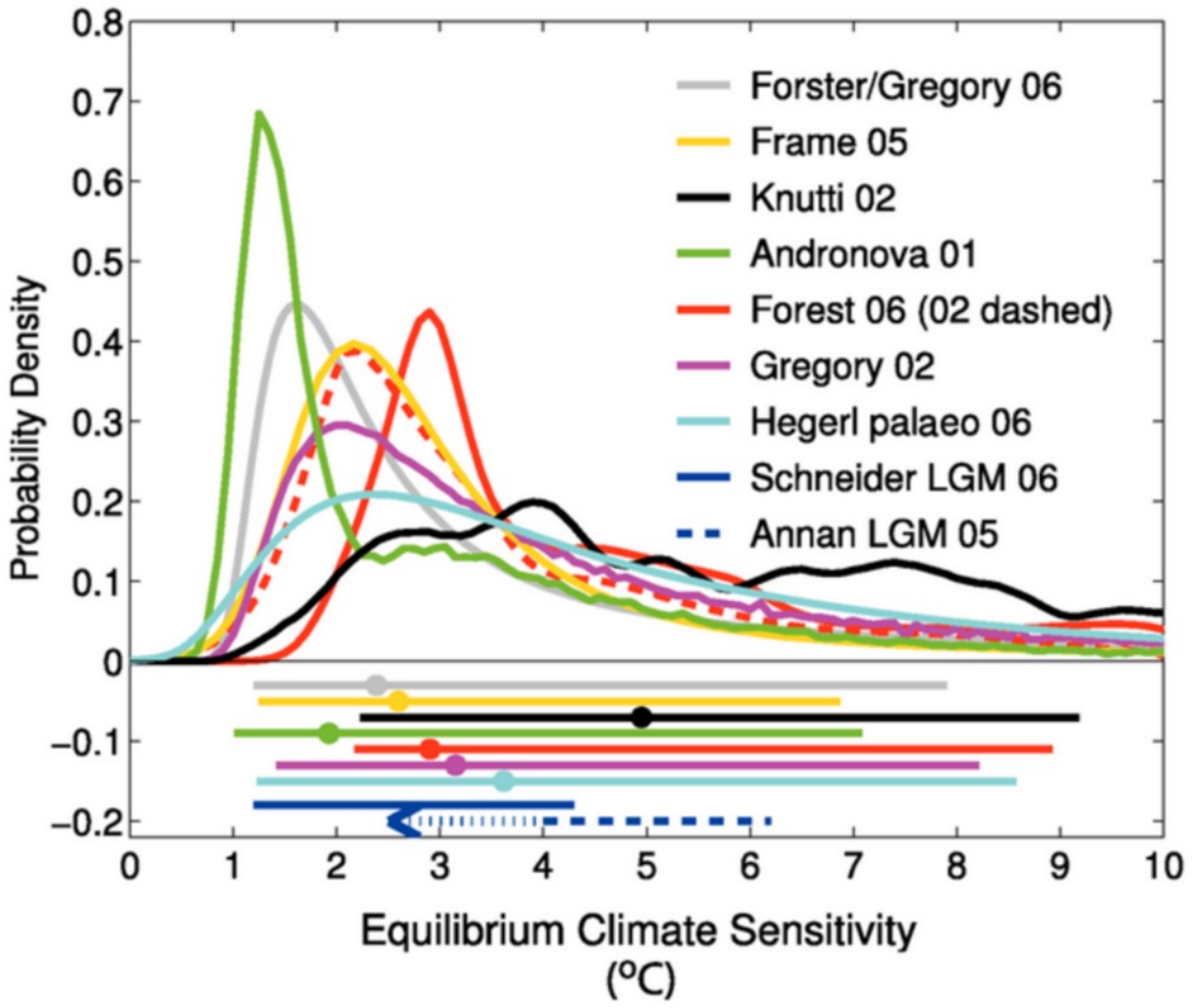
Impacts are Here





Cyclone Idai
Impacted over 3M people

The Whole Story is Filled with Uncertanties

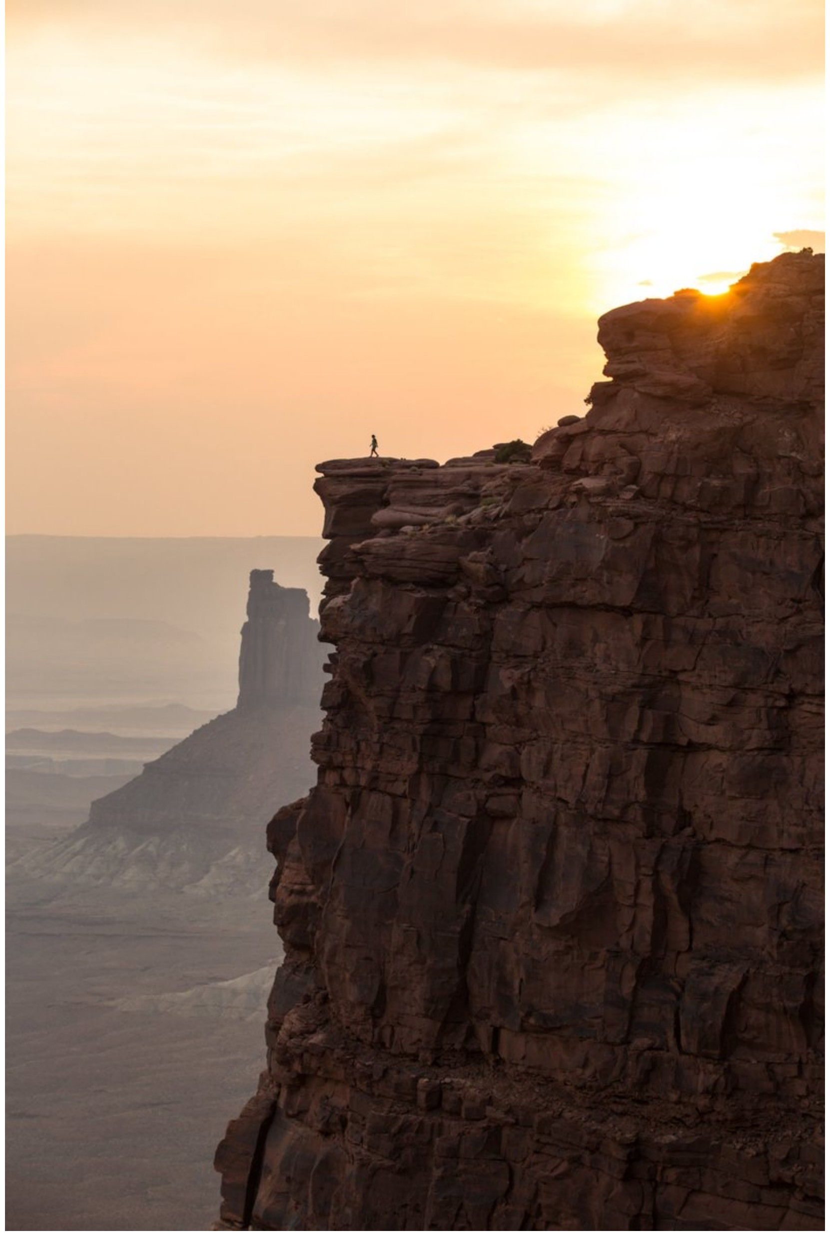


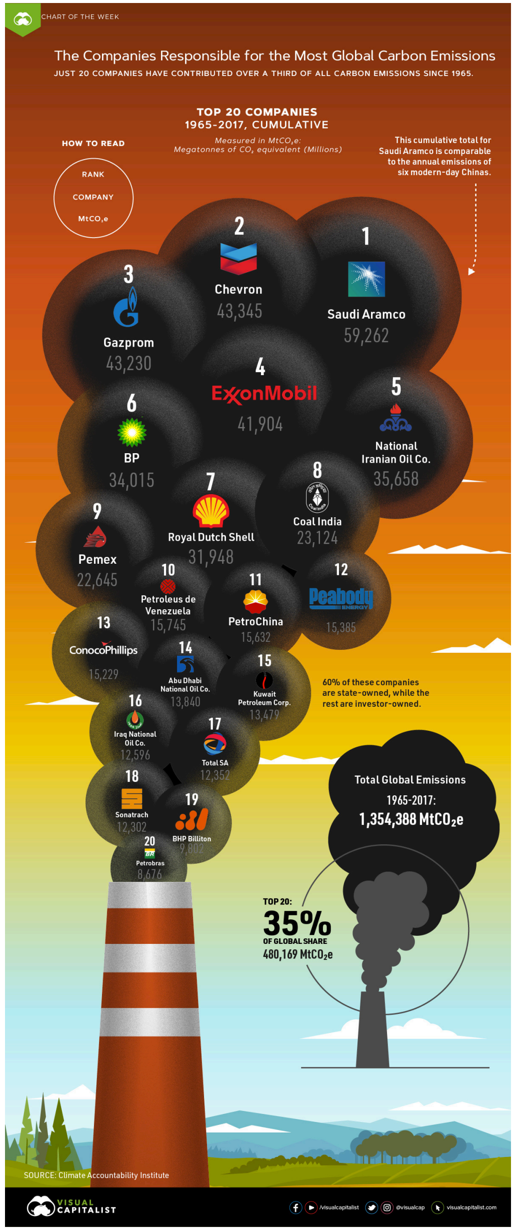
Many things are uncertain

- Future Amount of CO2
- Climate Sensitivity
- Impact

But we can reason under uncertainty

It is hard to feel like you can do anything…





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

Not really an ethical stance

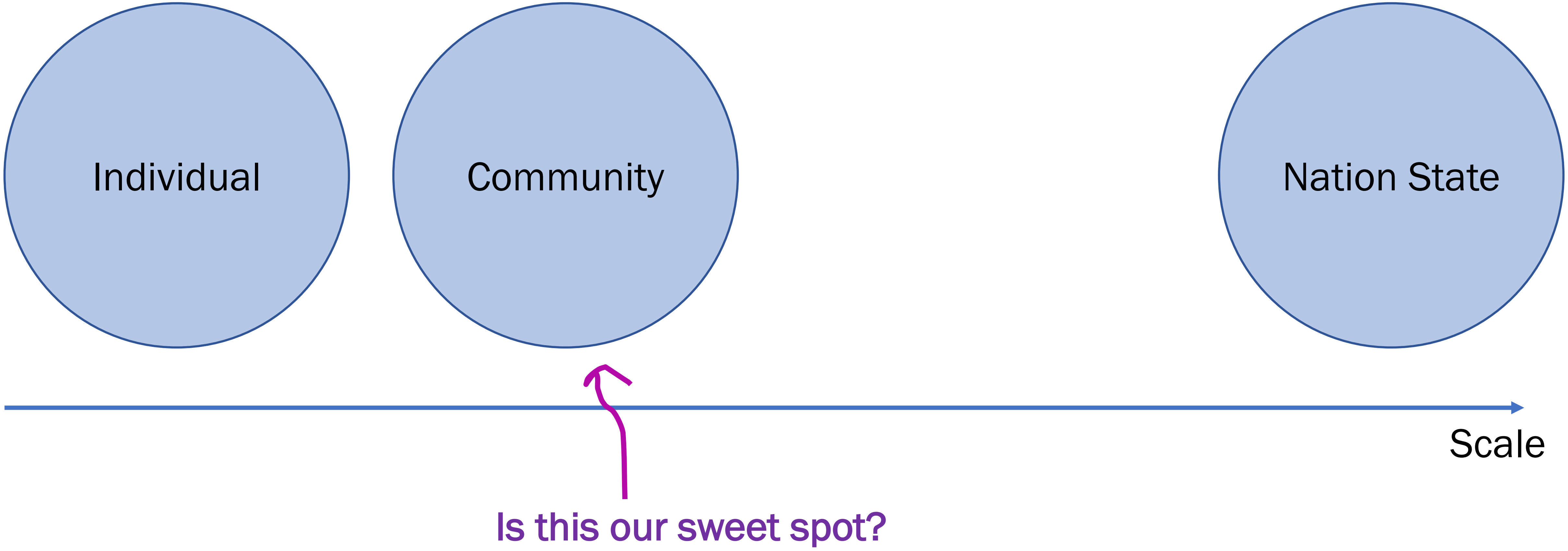
"I am just going to wait and see what happens"

Hannah Arendt what is the problem with bureaucrats of Hitler's empire?

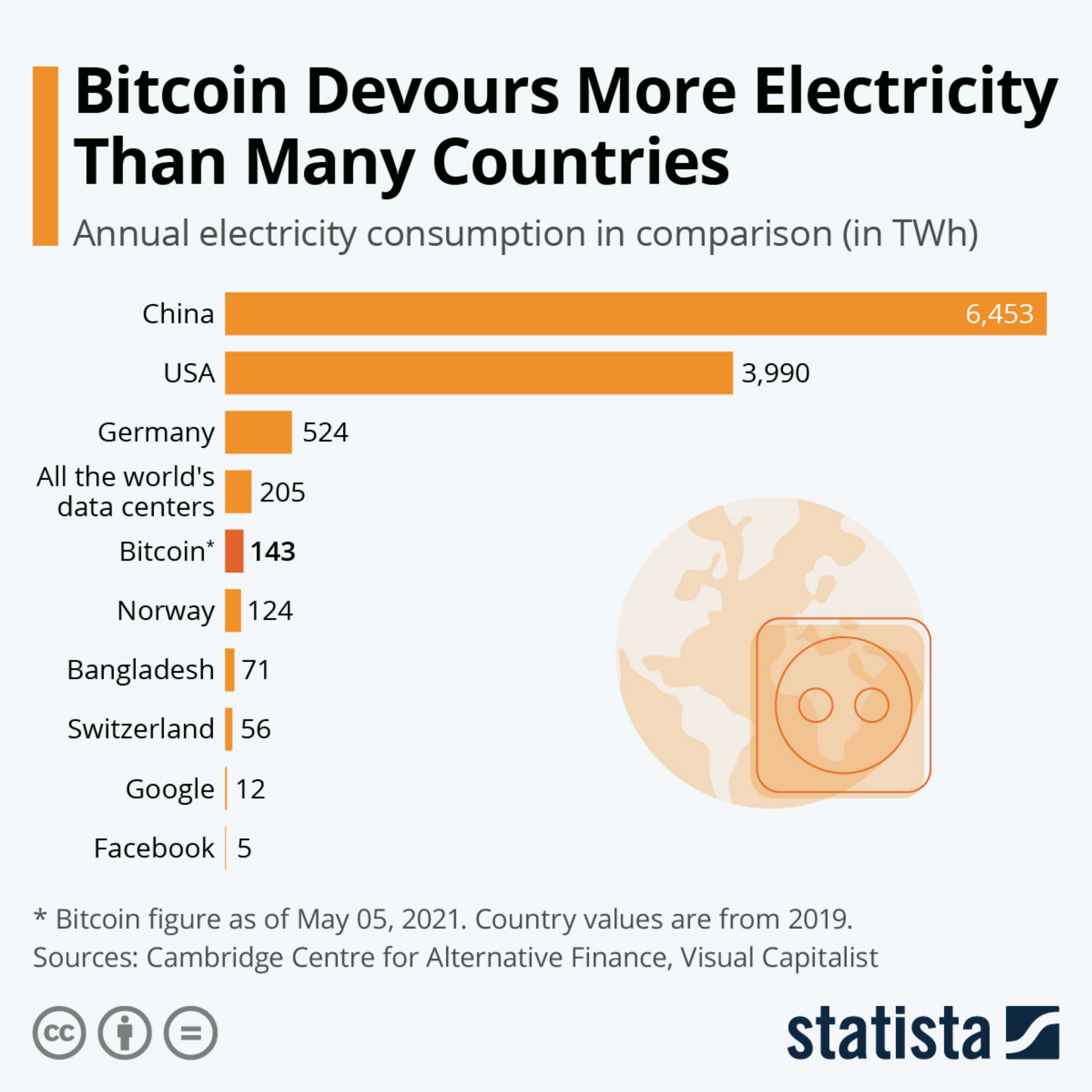
Is this an ethical policy?

What can we do?

Push for some change



Reduce CS "Pump" of Proof of Work

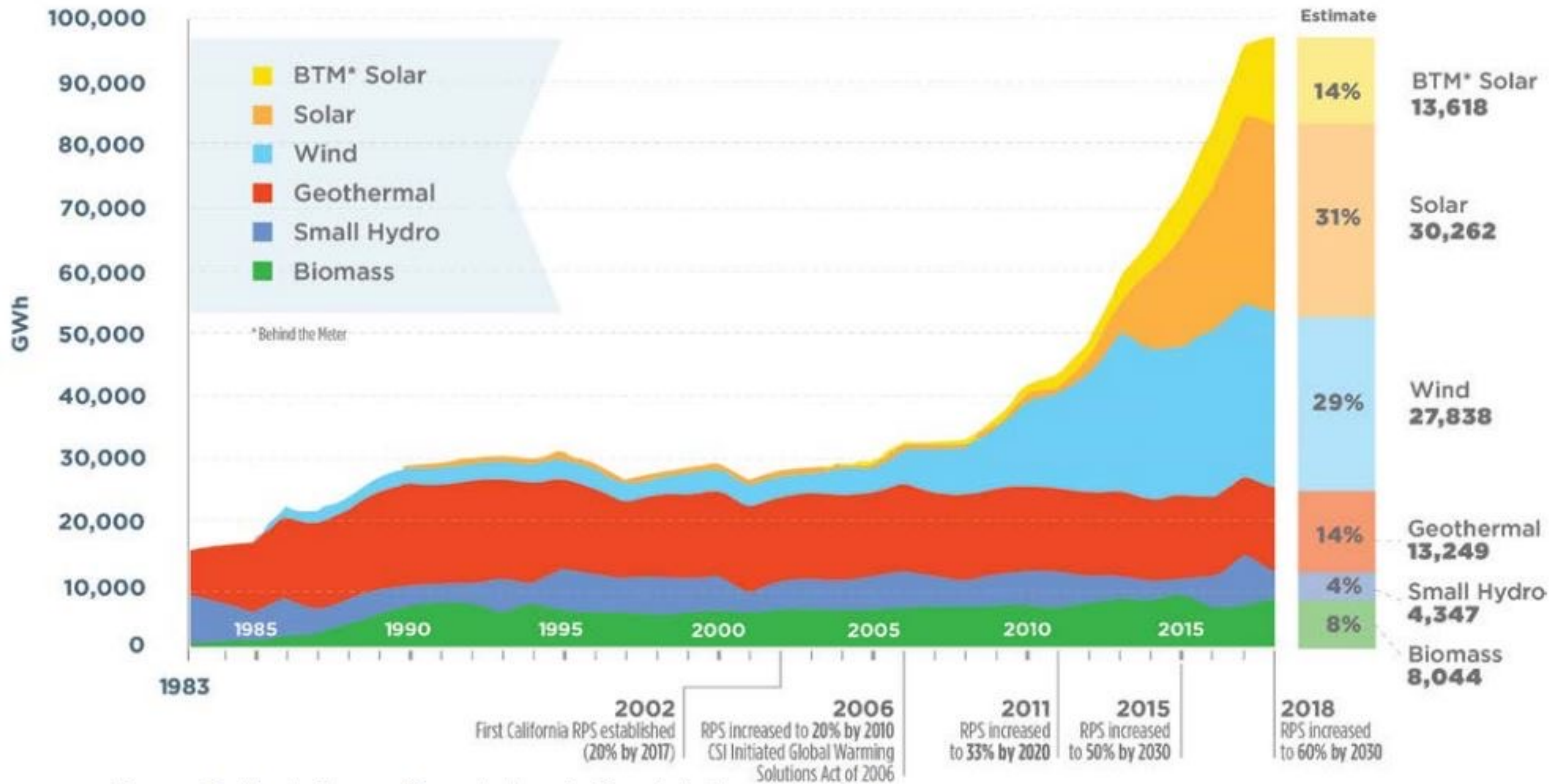


160,000,000,000,000
Hashes per second

But climate change and bitcoin isn't even part of ethics at Stanford CS (I will update this slide once that changes)

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?

Your Homework

Give yourself space to reflect on your own sense of what is right. And what you want for your own life’s work

Thank you!

Lecture made with Katie Creel

kcreel@stanford.edu