Ethics in ML

Chris Piech and Katie Creel

CS109, Stanford University

How AI is impacting our lives?

Smartphones

Social Media Platforms

E-Commerce

Autonomous Vehicles

Security & Surveillance

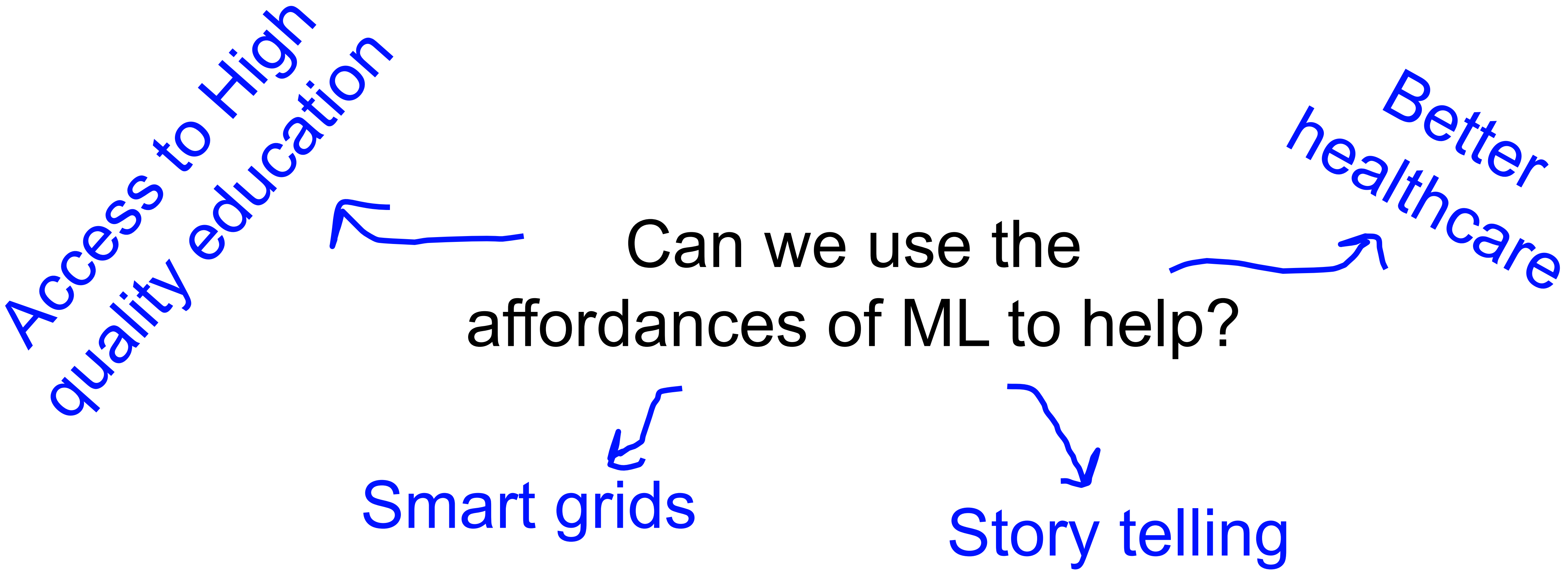
Navigation

Banking & Finance Sector

Smart Home

REMEMBER, WITH GREAT POWER
COMES GREAT RESPONSIBILITY
memegenerator.net

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



Can we use the
affordances of ML to help?

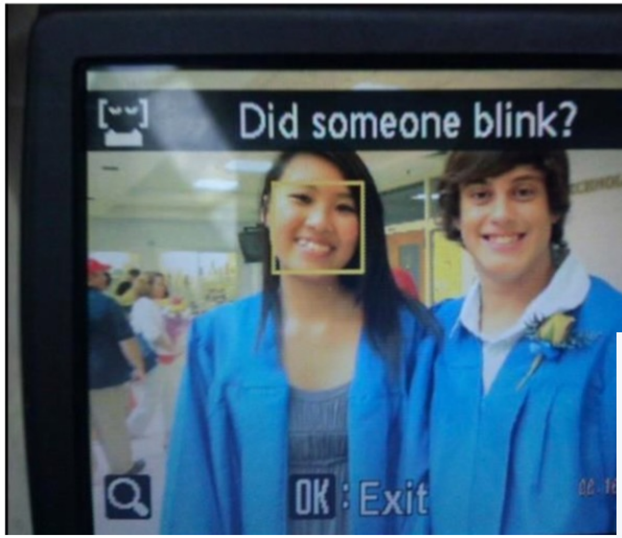
Access to High
quality education

Better
healthcare

Smart grids

Story telling

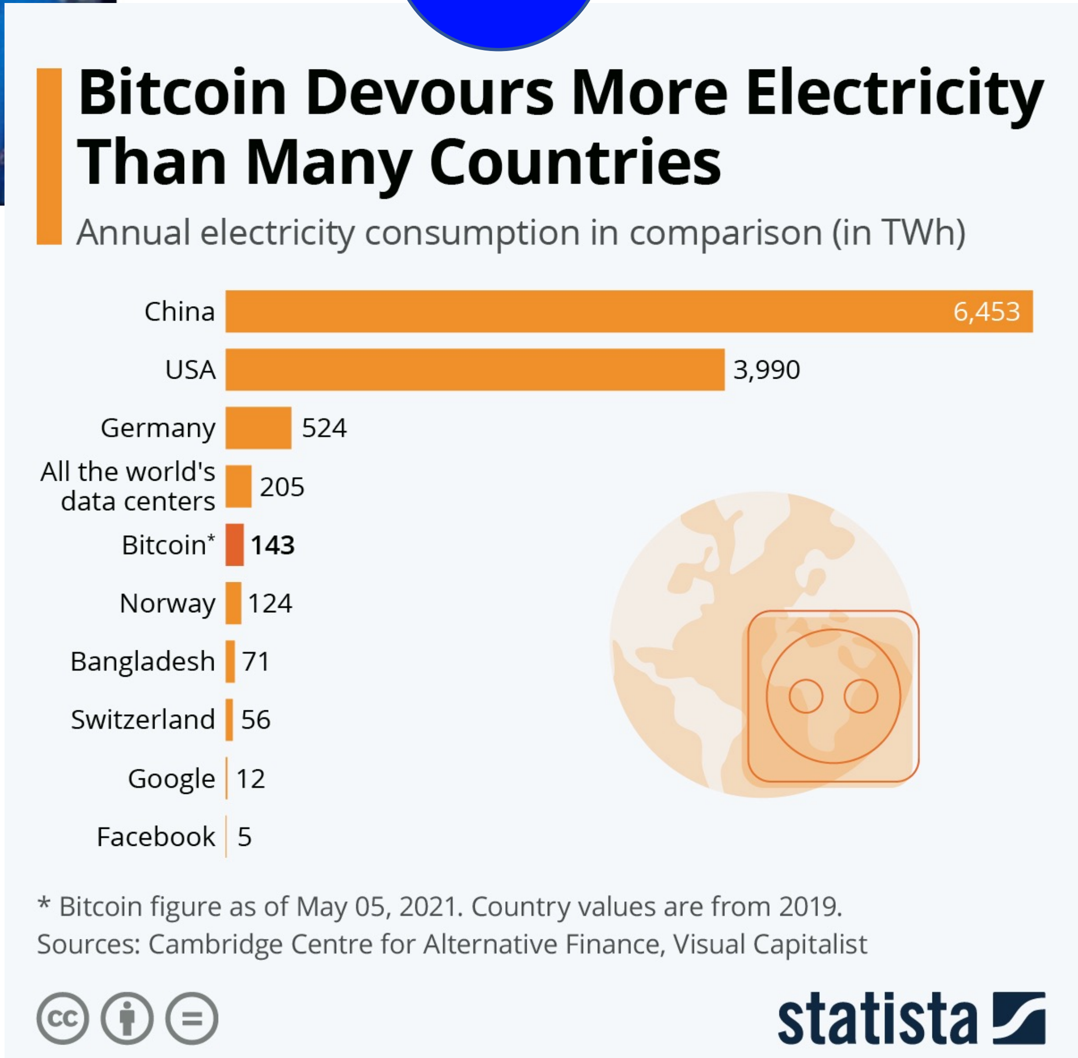
1



2

Facebook slammed by UN for its role in Myanmar genocide

3



Bitcoin Devours More Electricity Than Many Countries

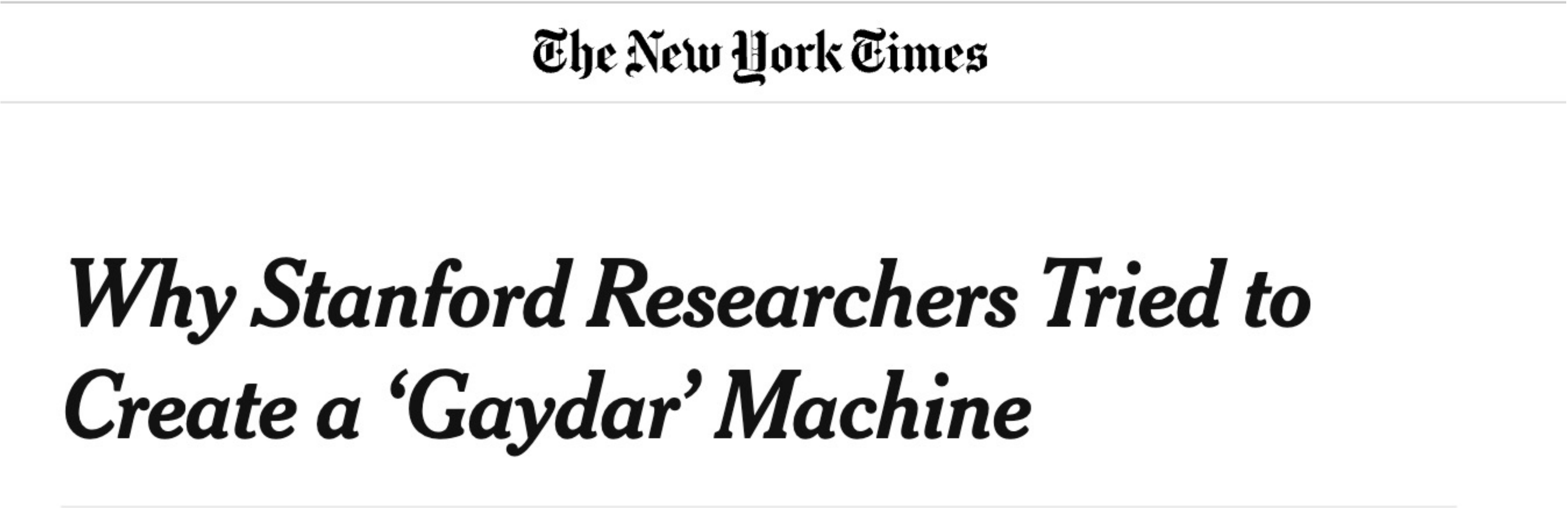
Annual electricity consumption in comparison (in TWh)

* Bitcoin figure as of May 05, 2021. Country values are from 2019.
Sources: Cambridge Centre for Alternative Finance, Visual Capitalist

Learning Goals

1. Recognize a hidden ethics issue in ML with respect to protected demographics (and how to solve them)

2. Discuss ways to address them



The New York Times

Why Stanford Researchers Tried to Create a 'Gaydar' Machine

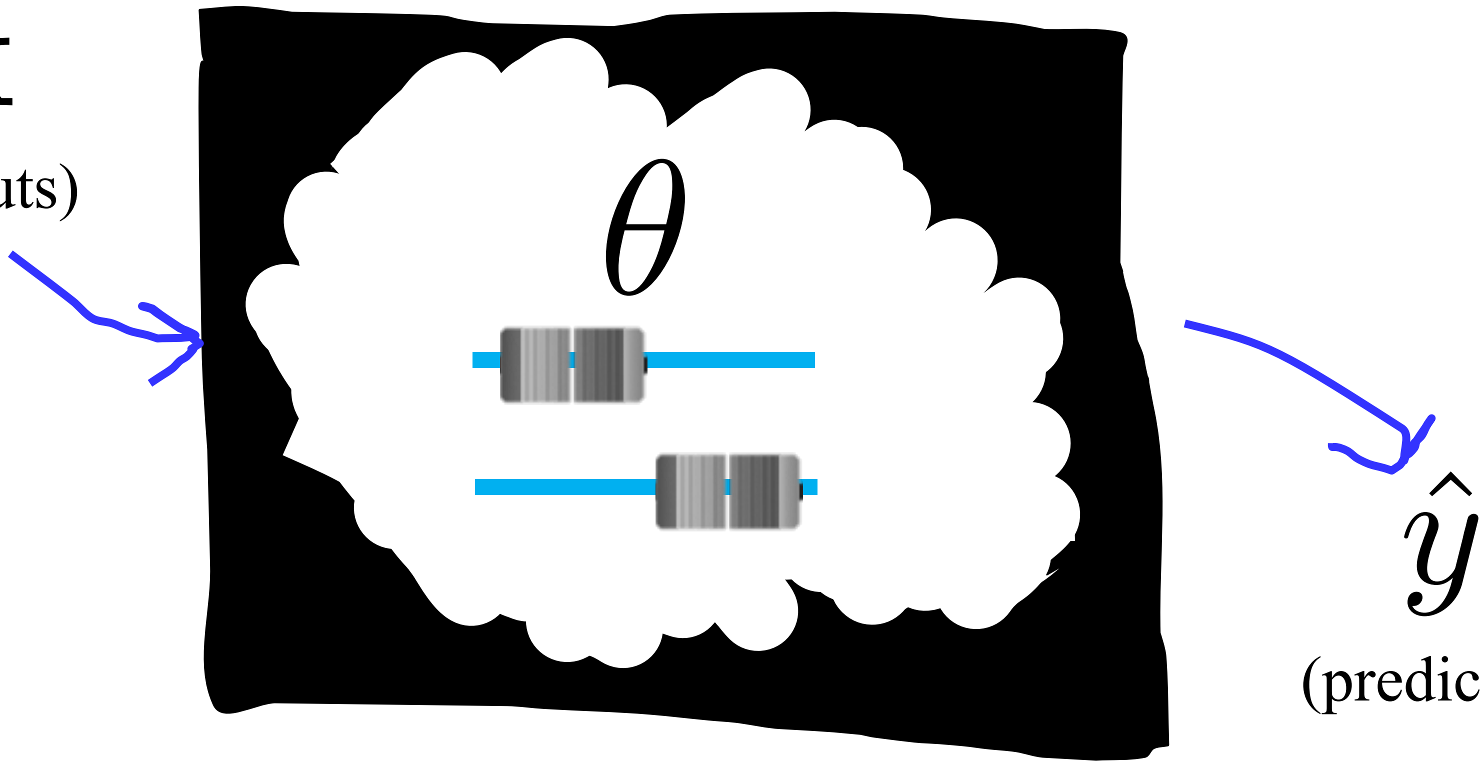
Other learning goal: how not to show up in a negative piece on the internet

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

Part 0: Review

Machine Learning



$\mathbf{x}$

(inputs)

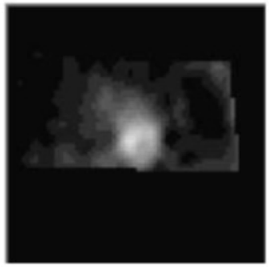
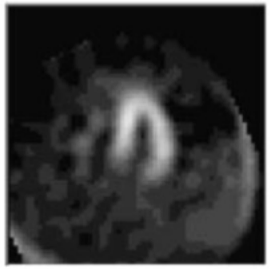
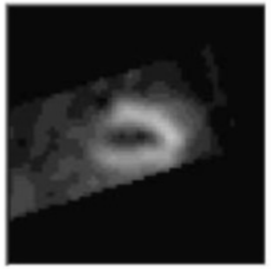
θ

$\hat{y}$

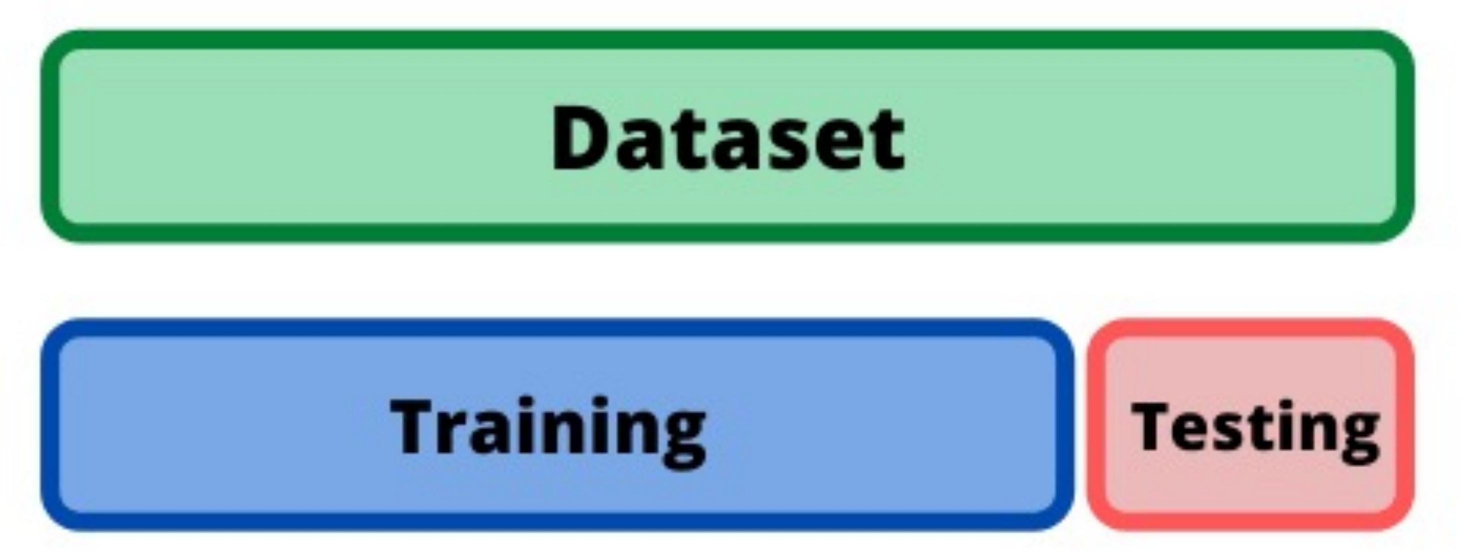
(prediction)

(model)

Classification Algorithms

	ROI 1	ROI 2		ROI *m*	Output
			…		
Heart 1	0	1		1	0
Heart 2	1	1		1	0
			⋮		⋮
Heart *n*	0	0		0	1

The Training / Testing Paradigm



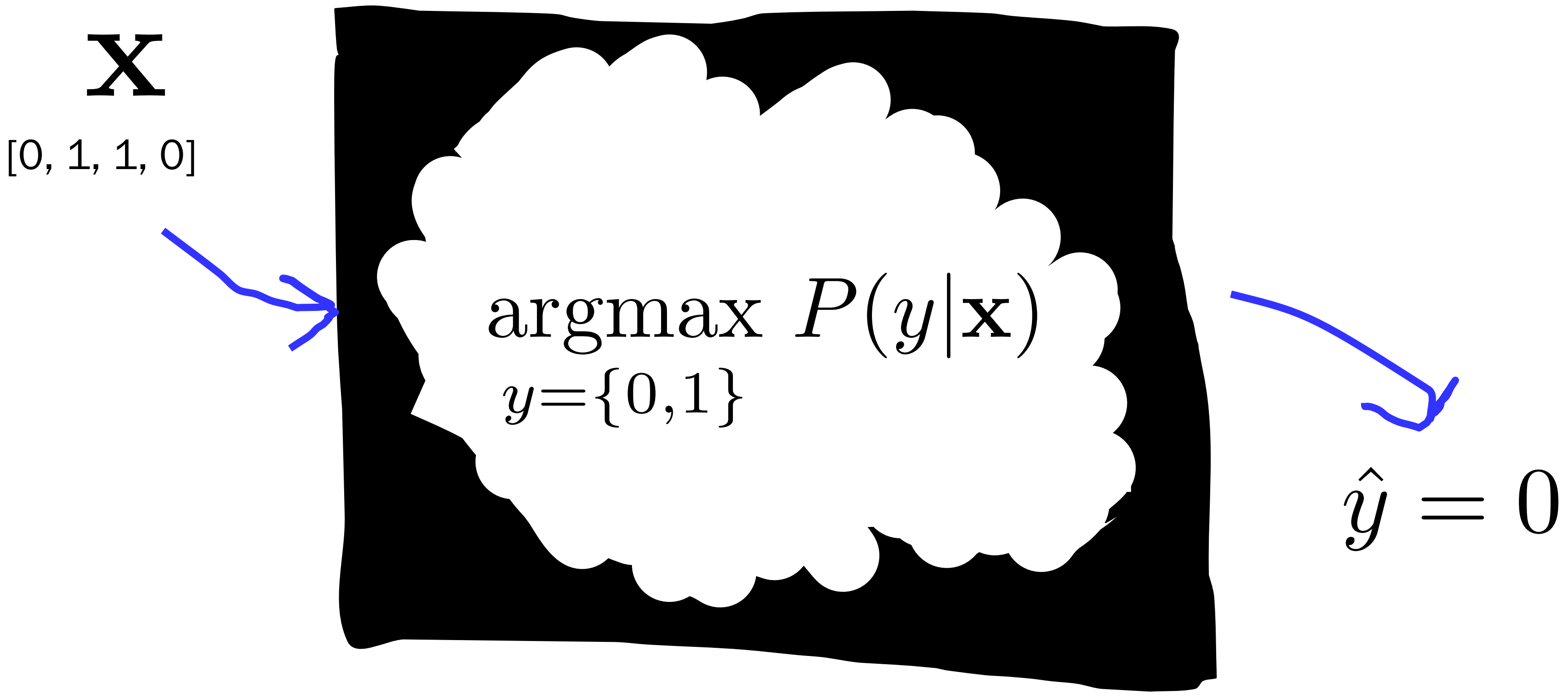
Learn your parameters

Make sure that they work

If your model passes testing...

Deployment

Classification Algorithms



$\mathbf{x}$

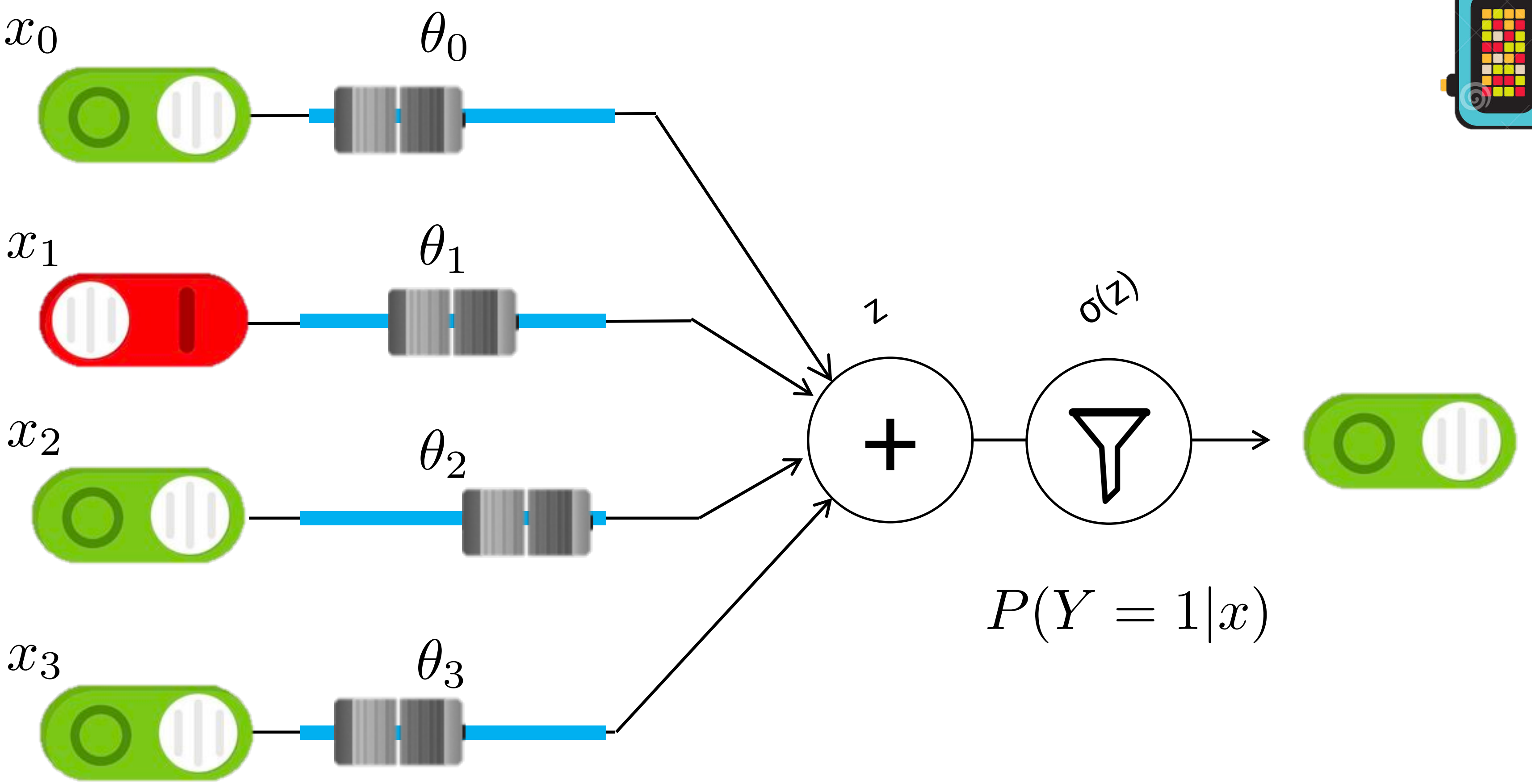
[0, 1, 1, 0]

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

$\hat{y} = 0$

Making a prediction…

Logistic Regression

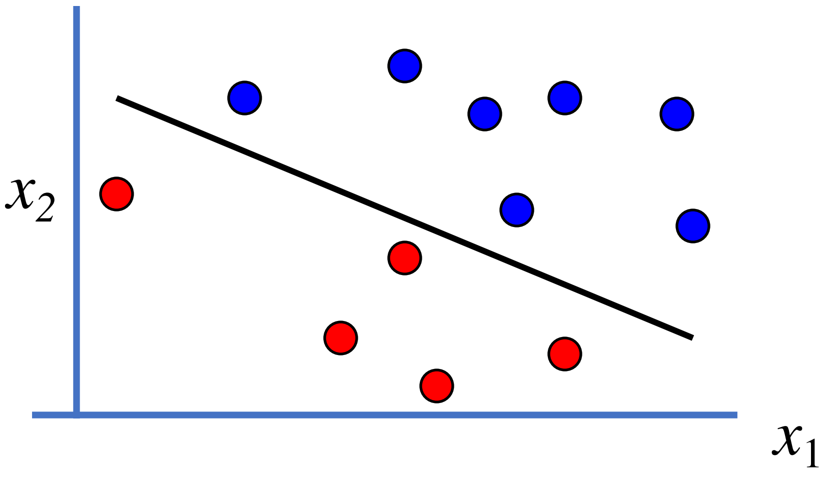


$P(Y = 1|x)$

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

Single 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



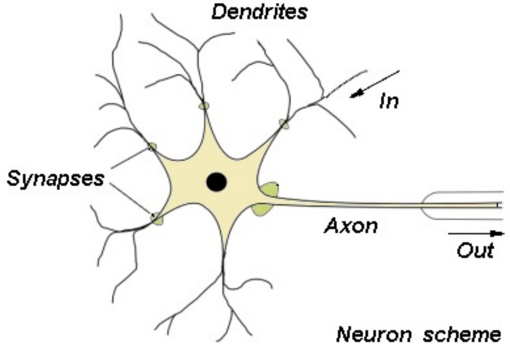
$$\theta^T \mathbf{x} = 0$$

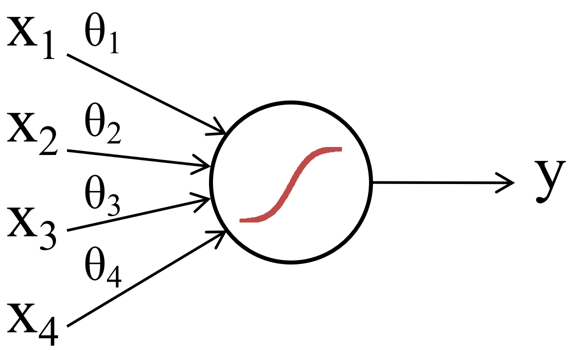
$$\theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_m x_m = 0$$

- 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.

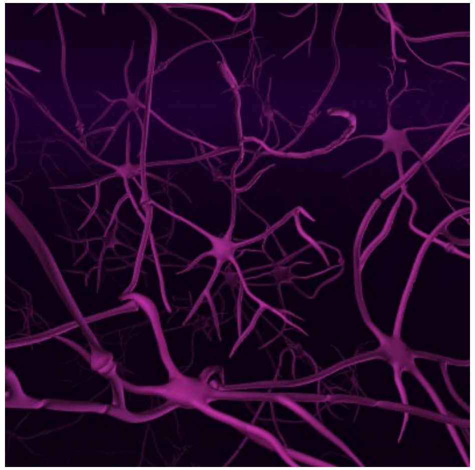
Deep Learning: Logistic Regression Can Be Stacked

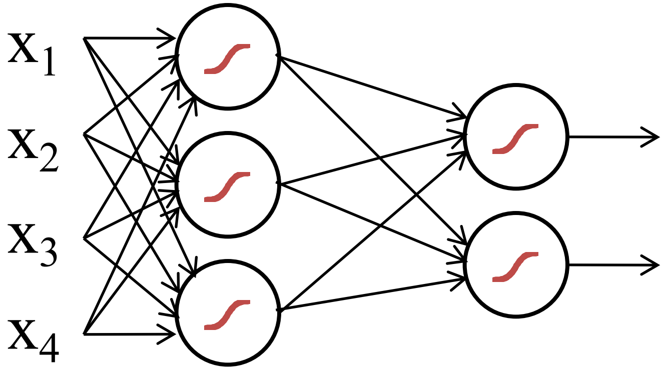
A neuron





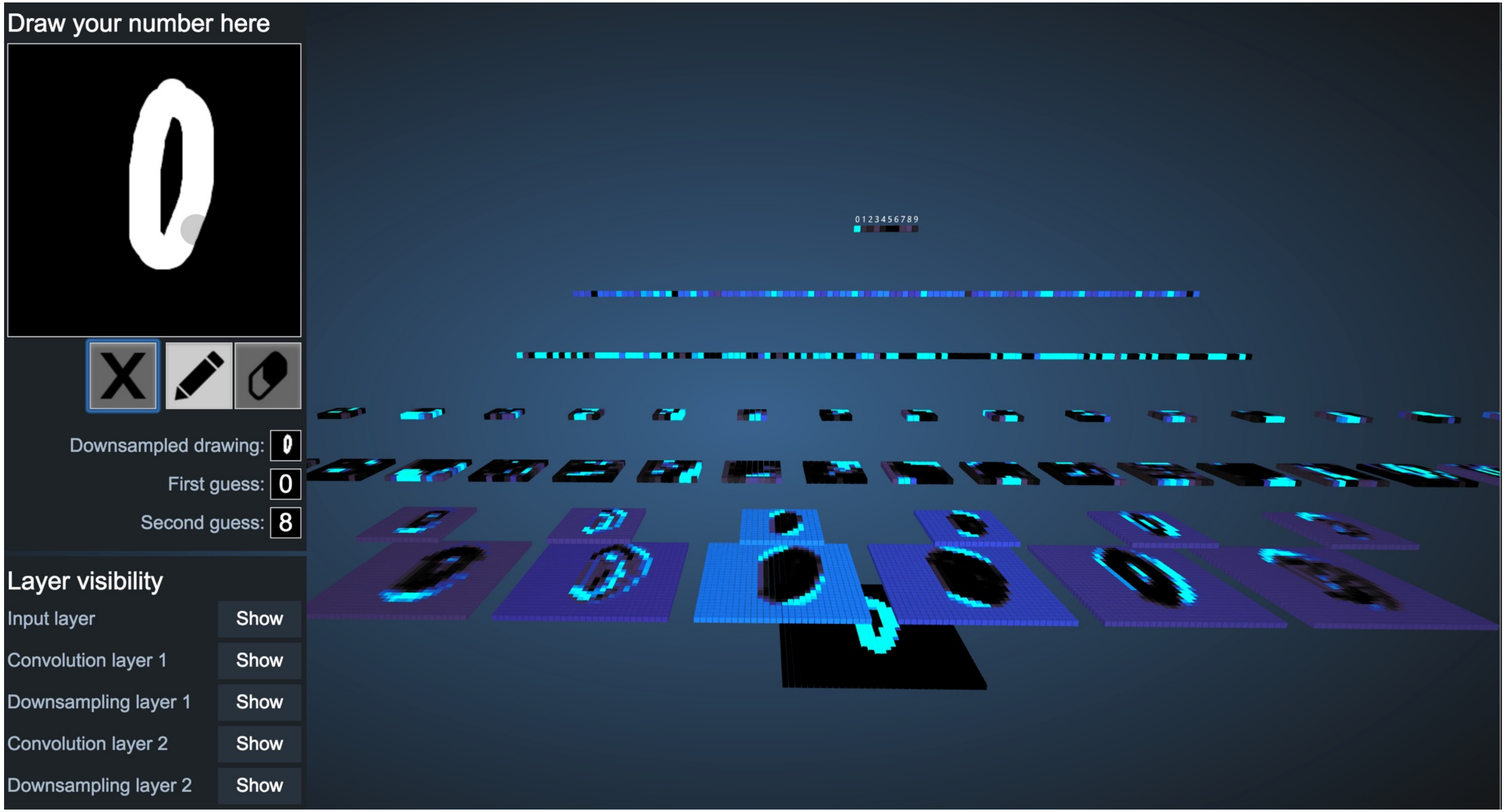
Your brain





Actually, it’s probably someone else’s brain

Logistic Regression for Image Classification



Draw your number here

Downsampled drawing: 0

First guess: 0

Second guess: 8

Layer visibility

Input layer Show

Convolution layer 1 Show

Downsampling layer 1 Show

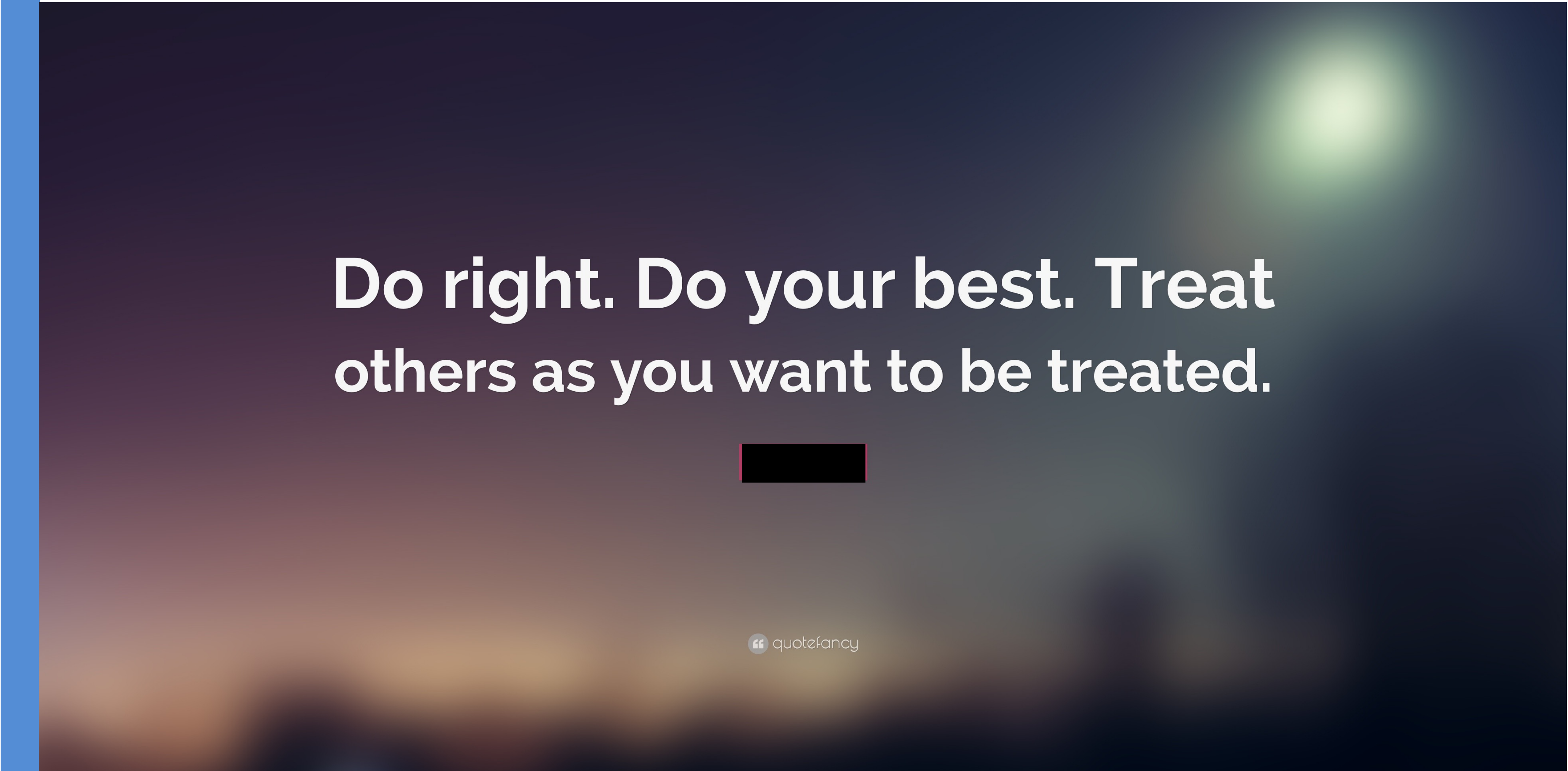
Convolution layer 2 Show

Downsampling layer 2 Show

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

Part 1: Framework of Harm

A Value System



Do right. Do your best. Treat others as you want to be treated.

[redacted]

quotefancy

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

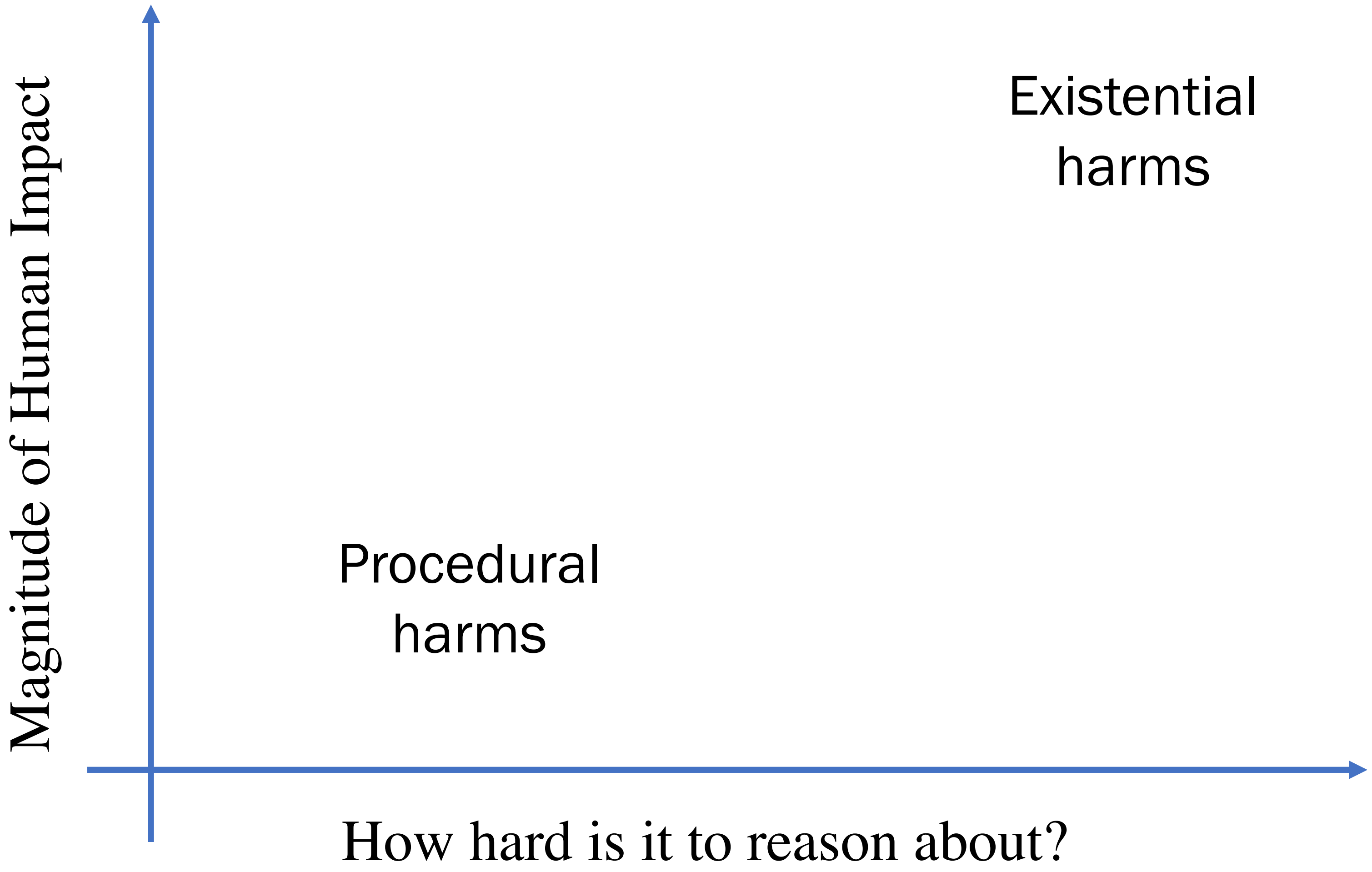
Existential harms

Maybe you will just break the whole damn thing

Examples:

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

Sticky Challenges



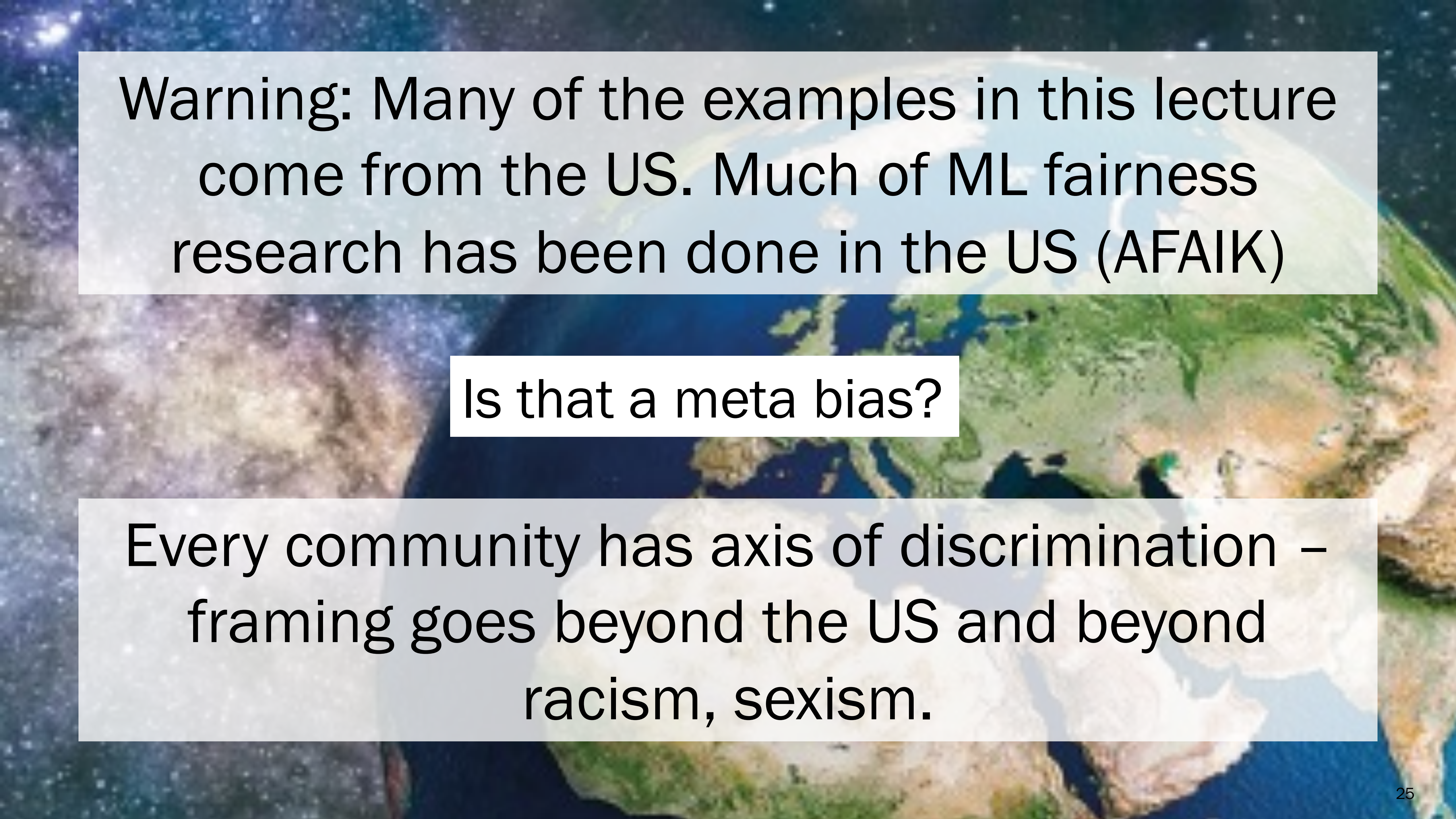
Magnitude of Human Impact

Existential harms

Procedural harms

How hard is it to reason about?

Part 2: Detecting Hidden Bias

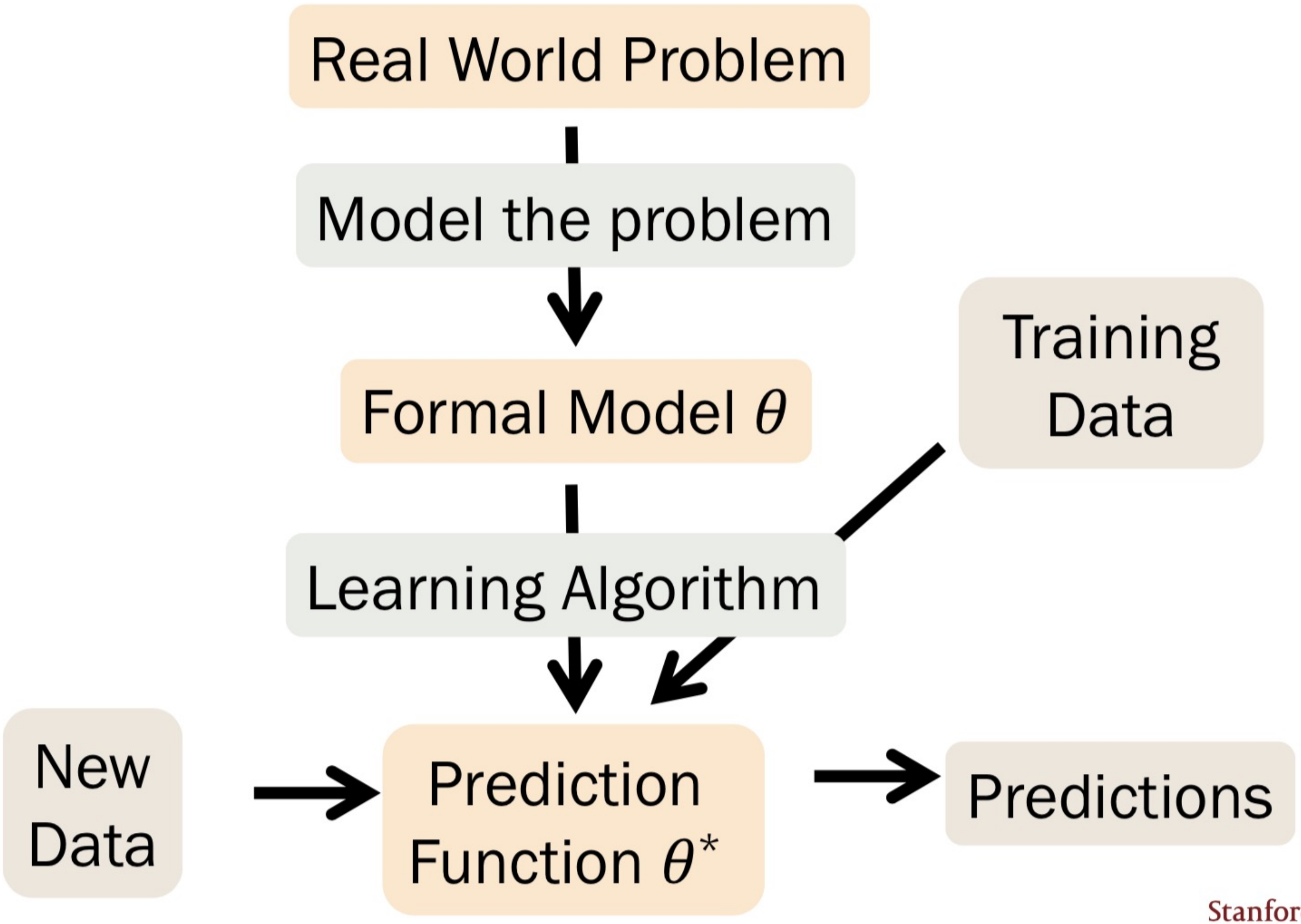


Warning: Many of the examples in this lecture come from the US. Much of ML fairness research has been done in the US (AFAIK)

Is that a meta bias?

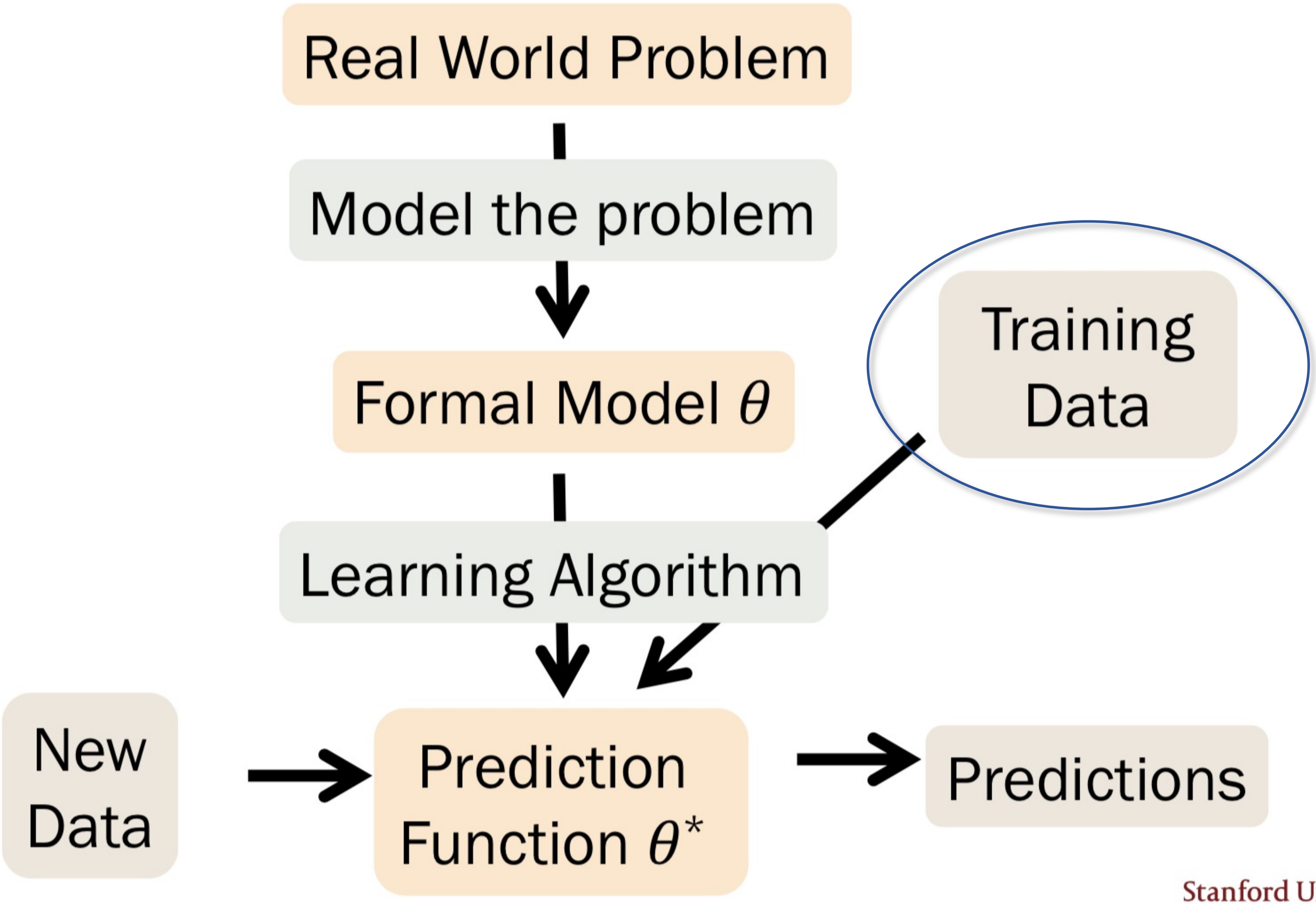
Every community has axis of discrimination – framing goes beyond the US and beyond racism, sexism.

Machine Learning



Stanford University

Machine Learning



Real World Problem

Model the problem

Formal Model θ

Training Data

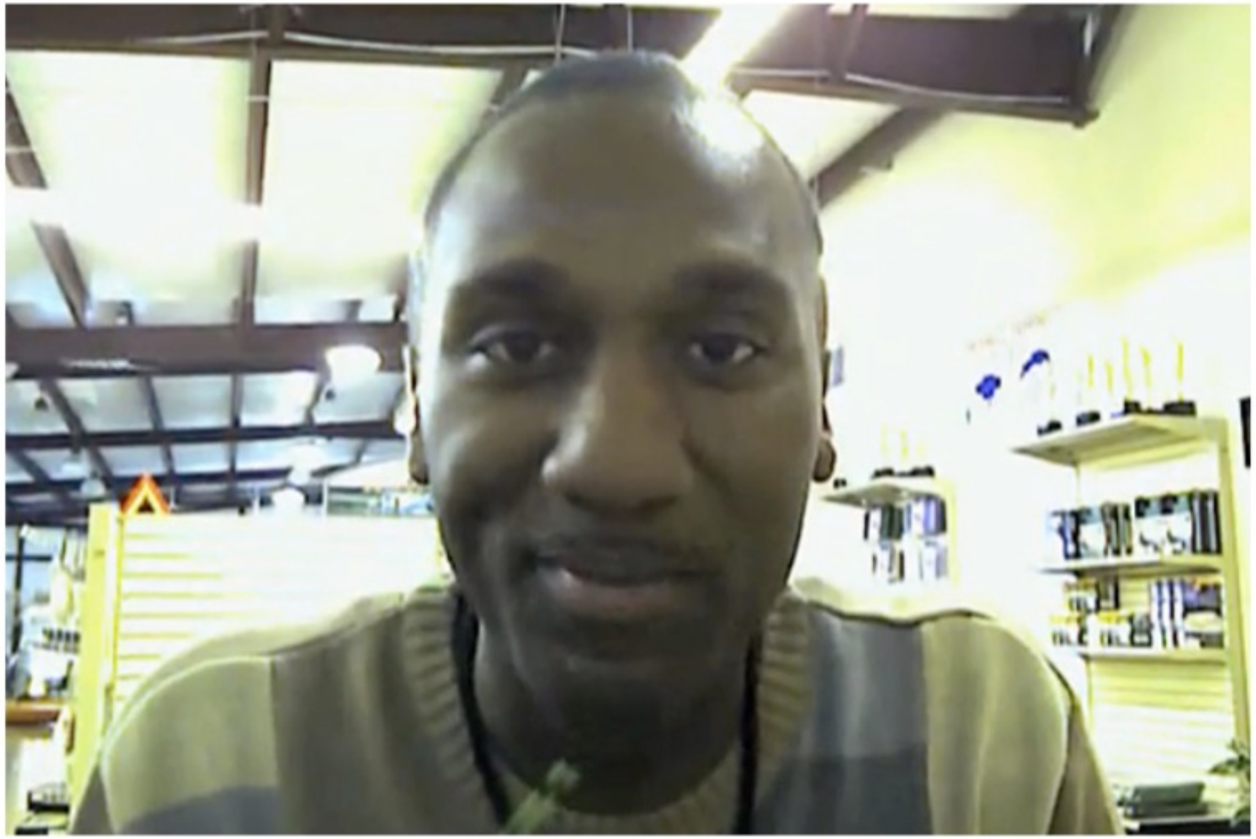
Learning Algorithm

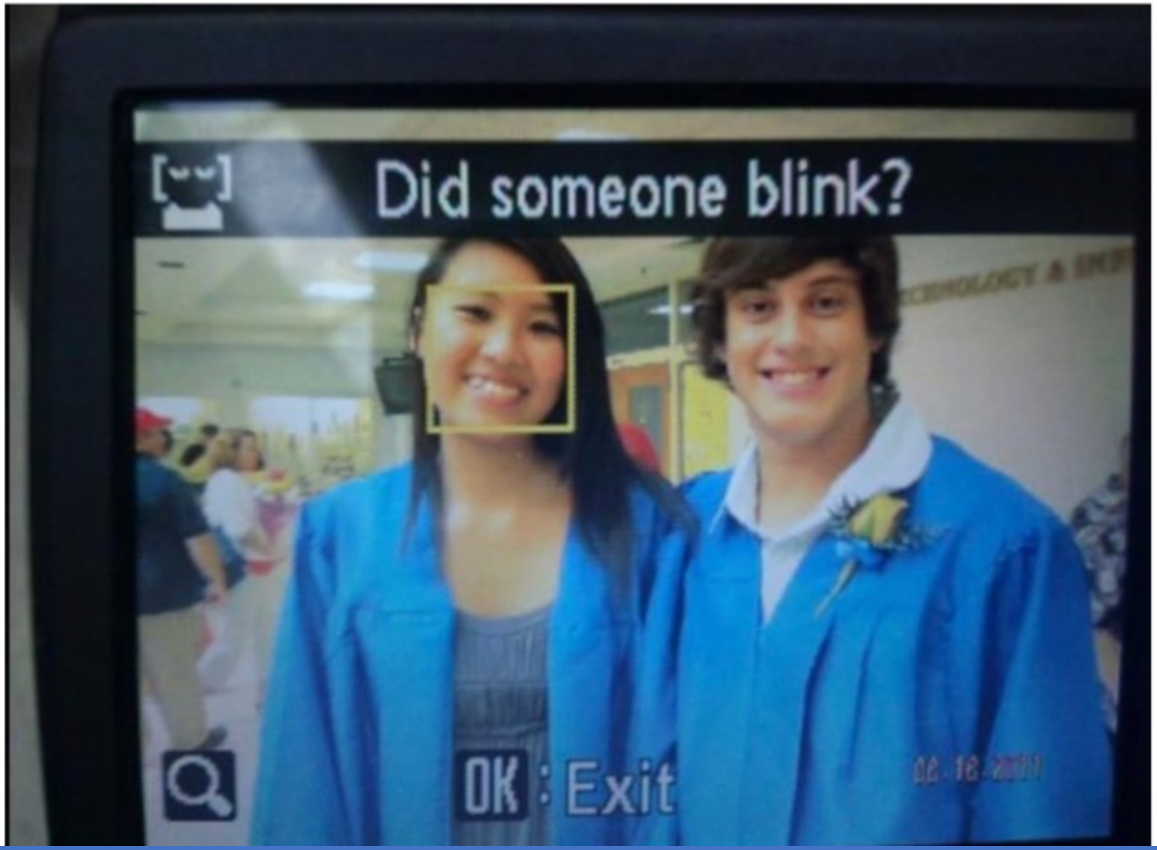
New Data

Prediction Function θ^*

Predictions

Ethics and Datasets?

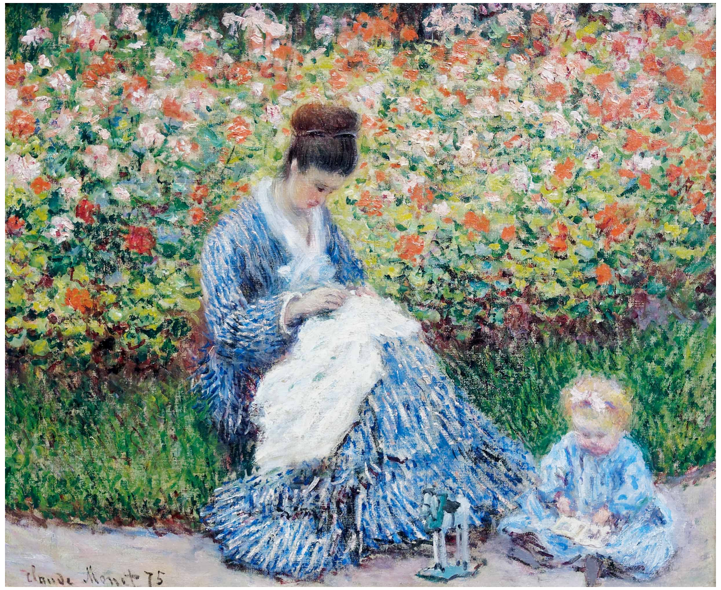


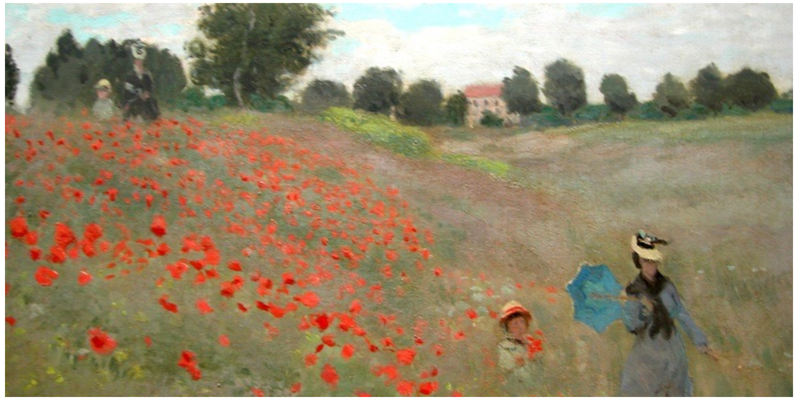


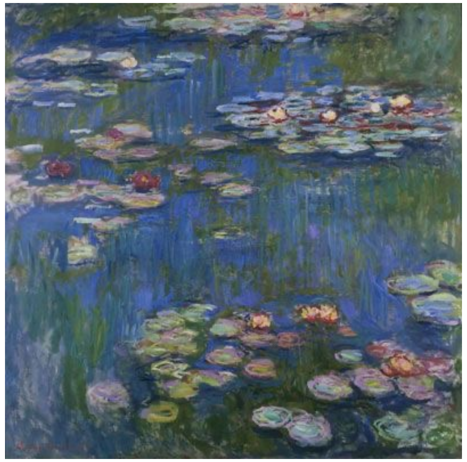
Theme #1:
Building Responsible Datasets

How is training data created and why is it often biased?

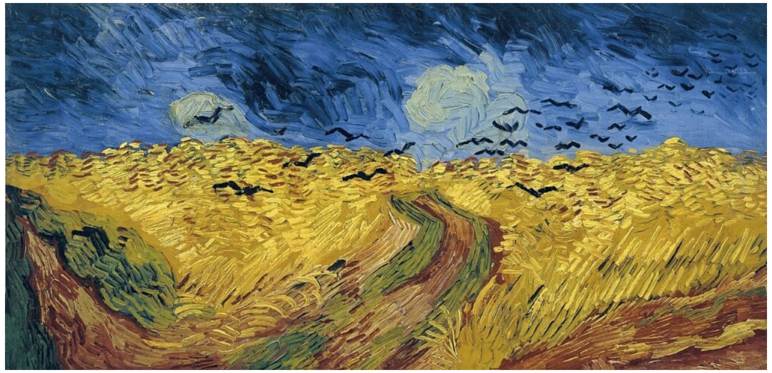
Classify the Artist

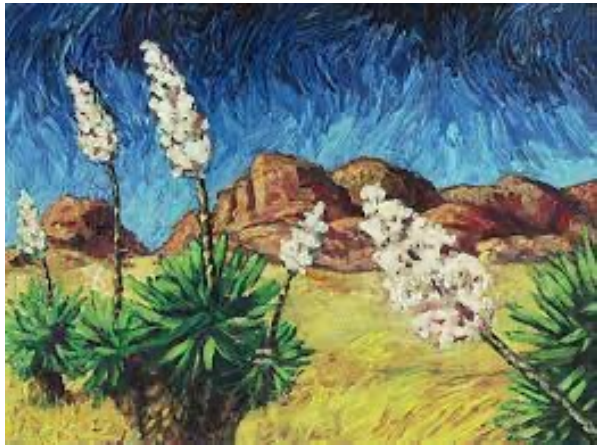


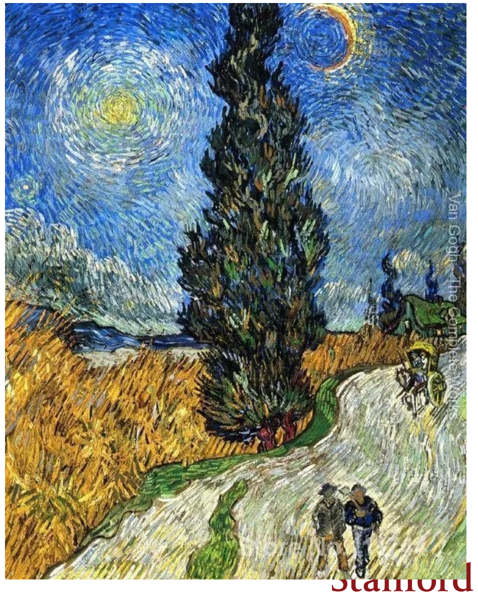




Monet

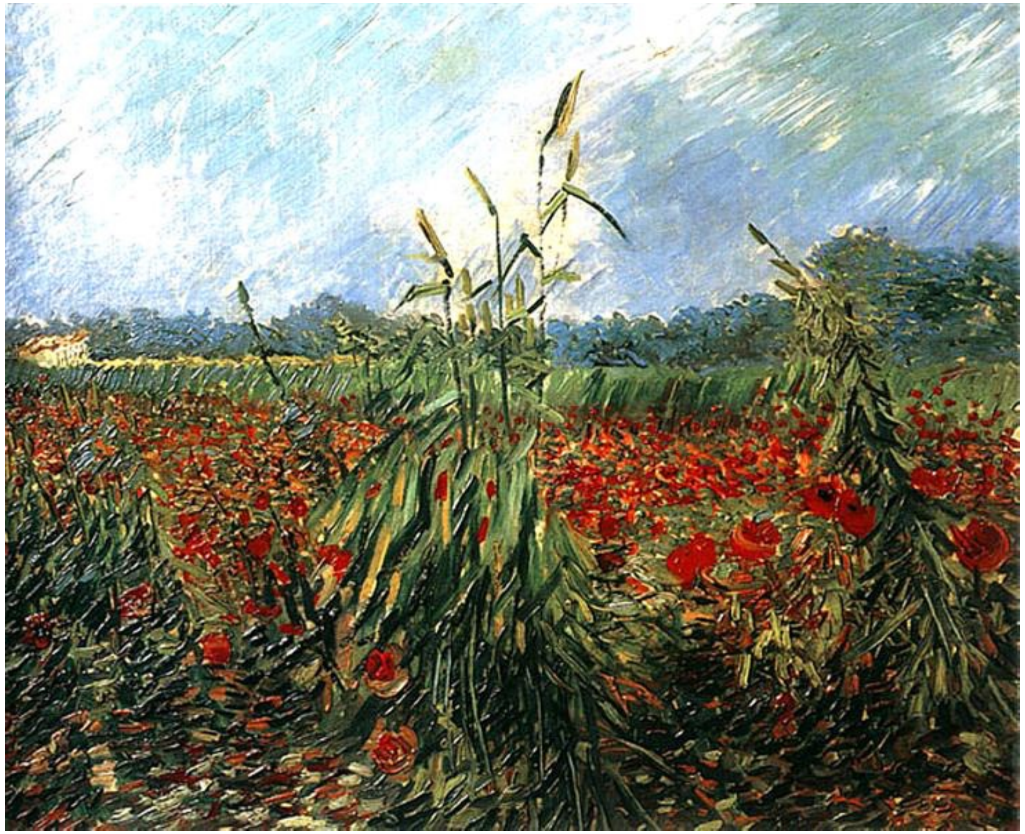






Van Gough

Classify the Artist

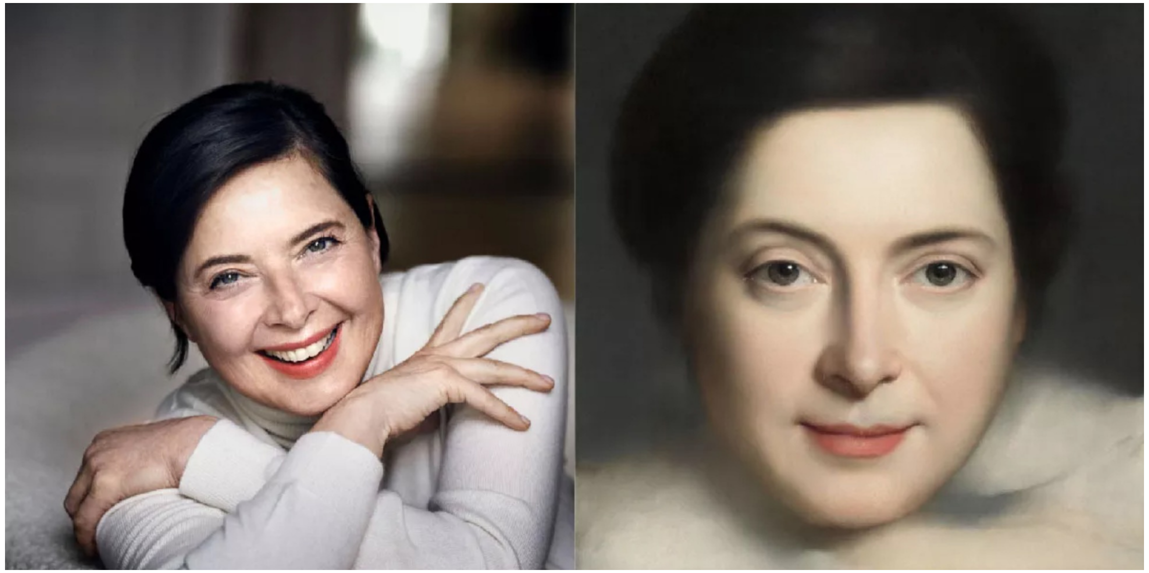


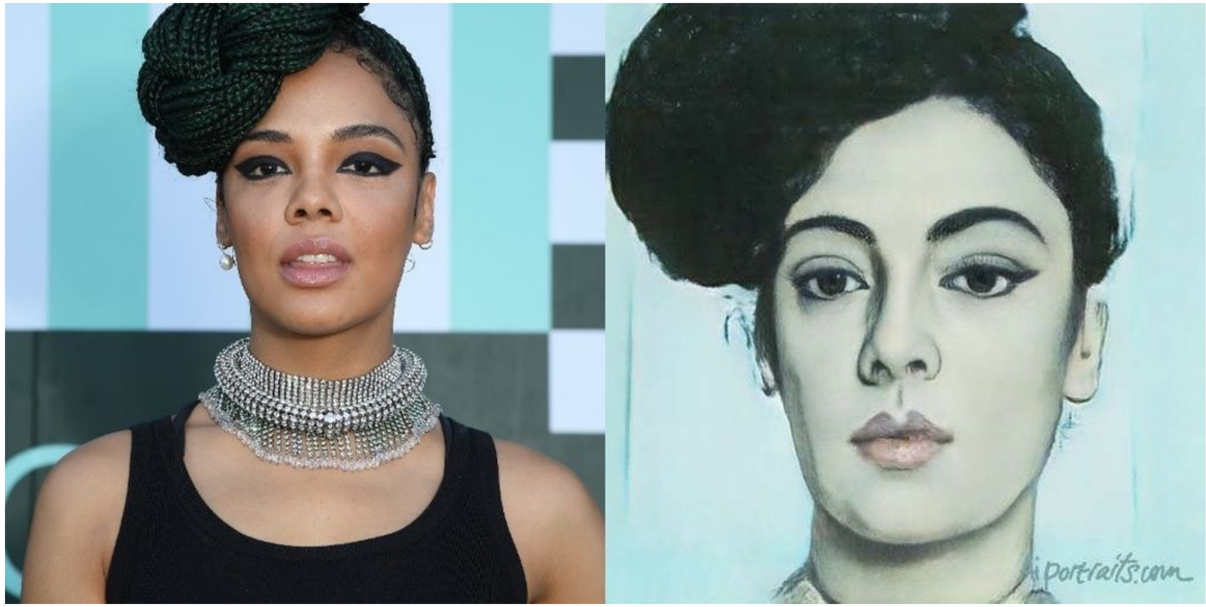
$\hat{y}$ = Monet

In the training dataset there are few examples of money drawing red flowers

Op. cite and Srinivasan & Uchino 2021
dl.acm.org/doi/10.1145/3442188.3445869

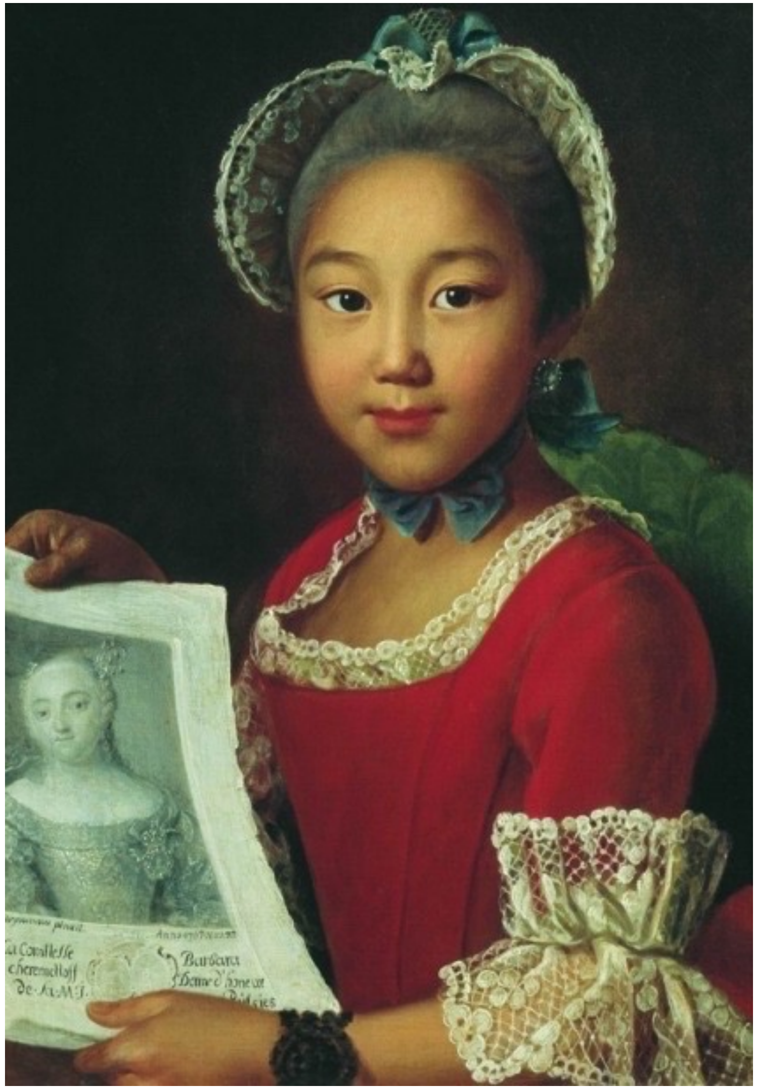
Skin lightening & feature whitening in generative art

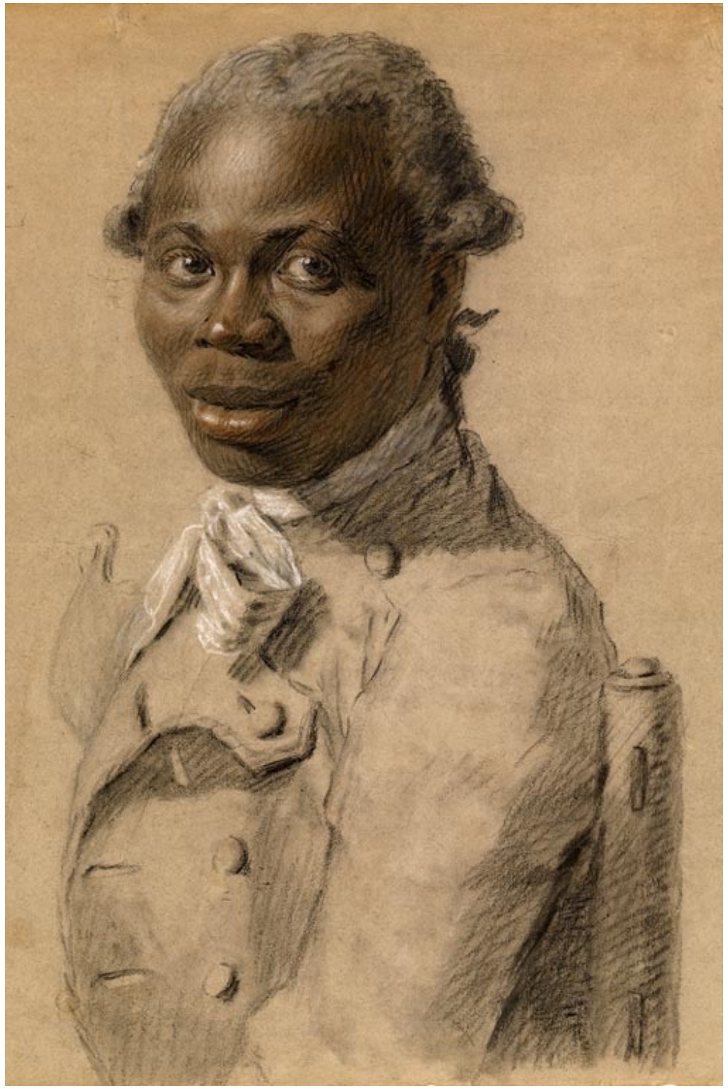


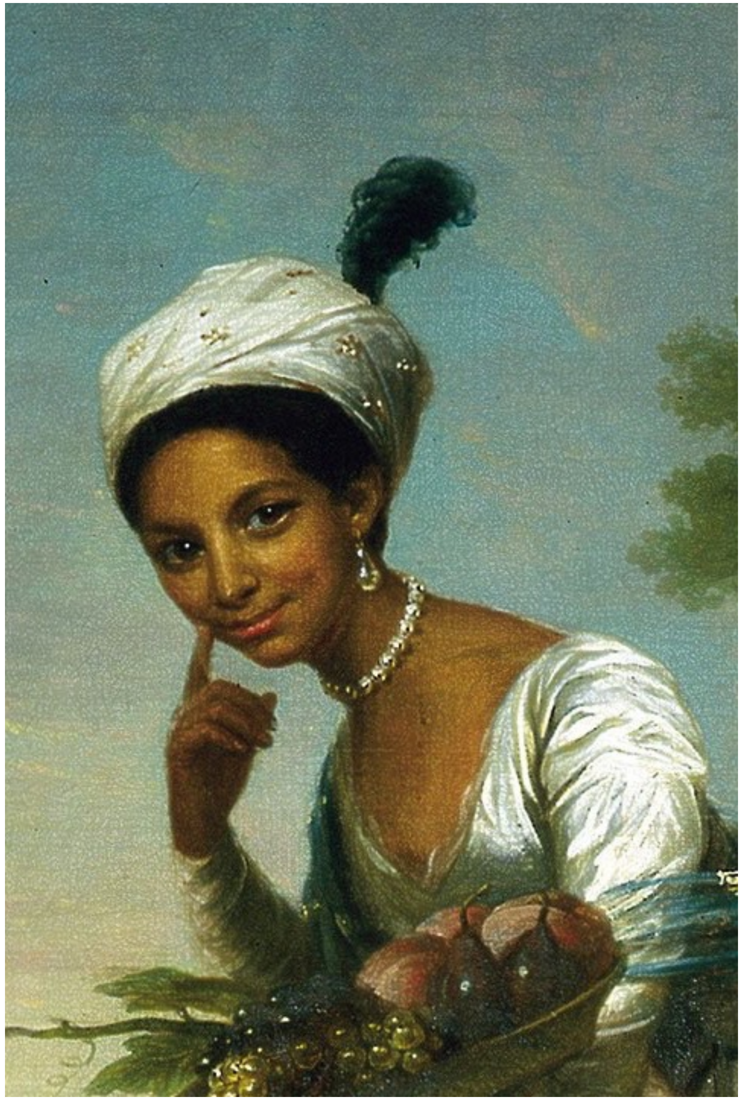


Images generated by AI Portrait Ars (now offline)

Better generative art is possible … if we train on datasets more representative of human population (but not of the European art archive)



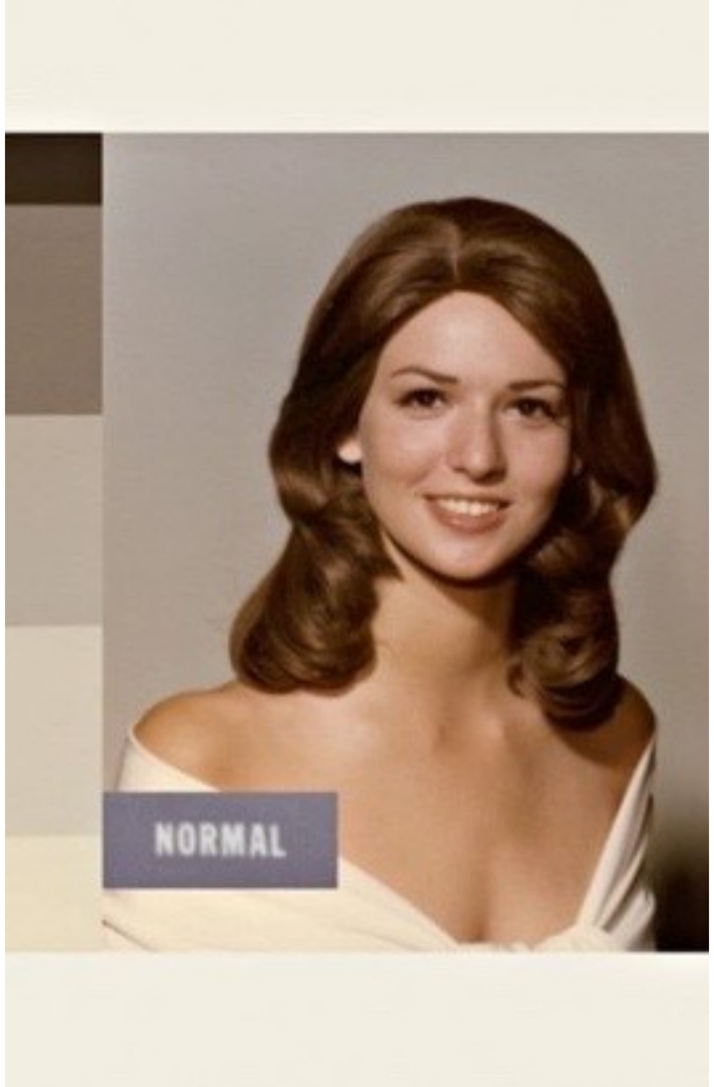




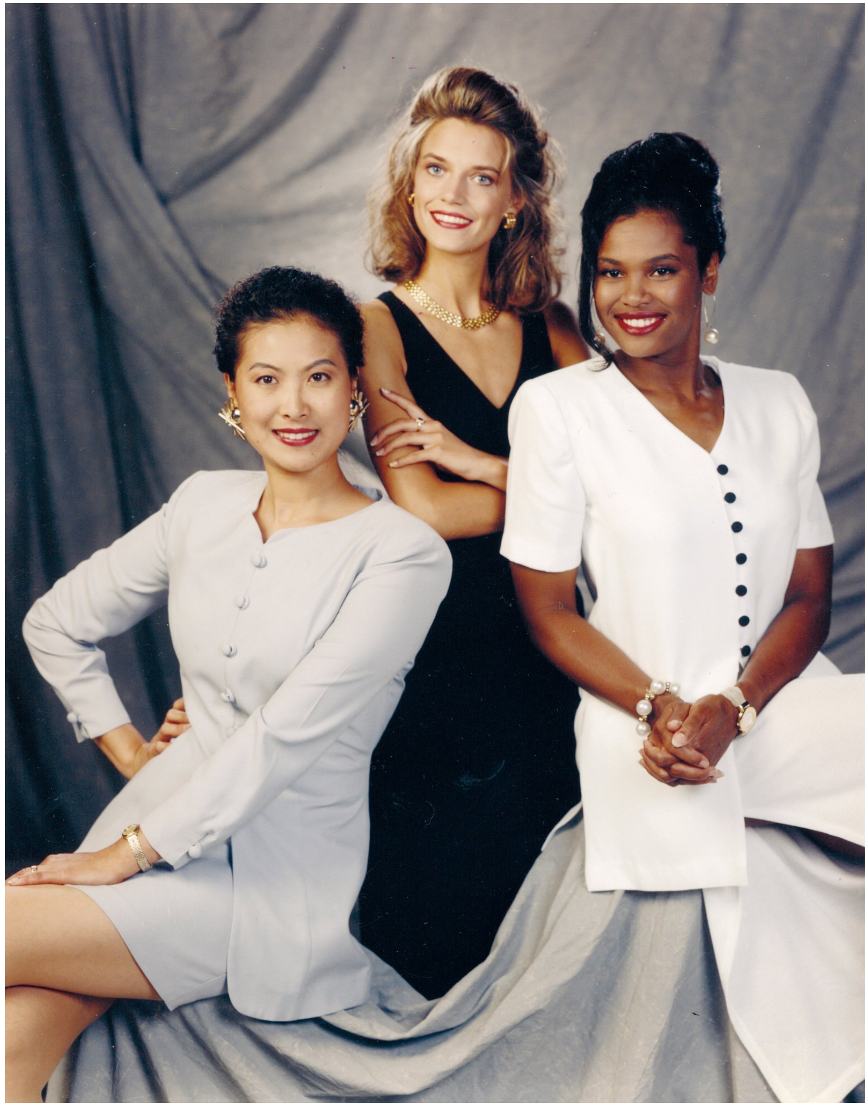
Biases in Image Benchmarks … A very brief history.

Tools used for benchmarks or calibration often are biased towards majority or dominant social groups. The “Shirley Card” film developers used as the test image original showed a white woman and only later included darker skintones.

(source: work of Sarah Lewis & Lorna Roth)



Shirley Card, 1944



Shirley Card, 1995

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

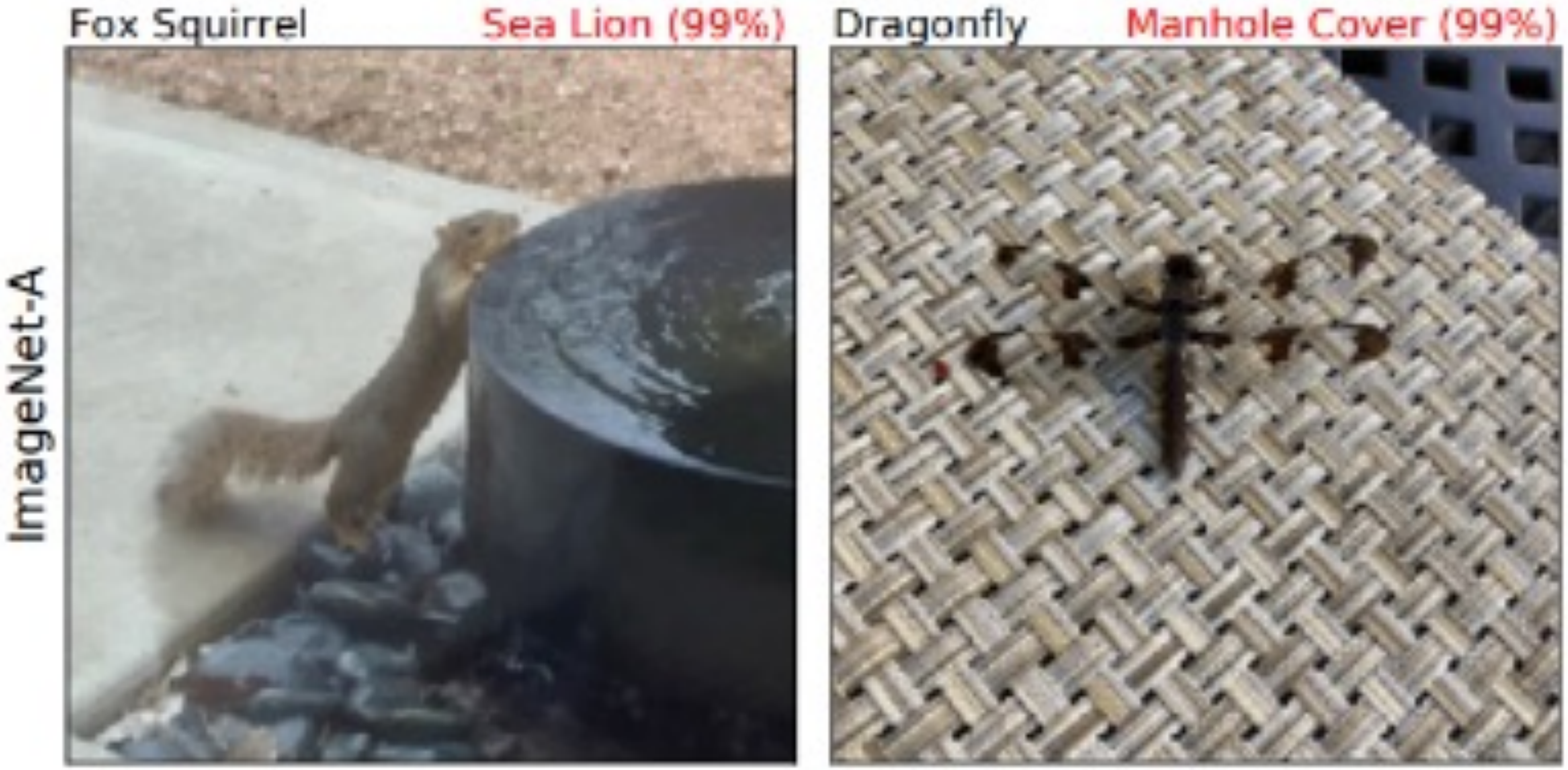
Hand-engineered features (SIFT, HOG, LBP), Spatial pyramid, SparseCoding/Compression

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

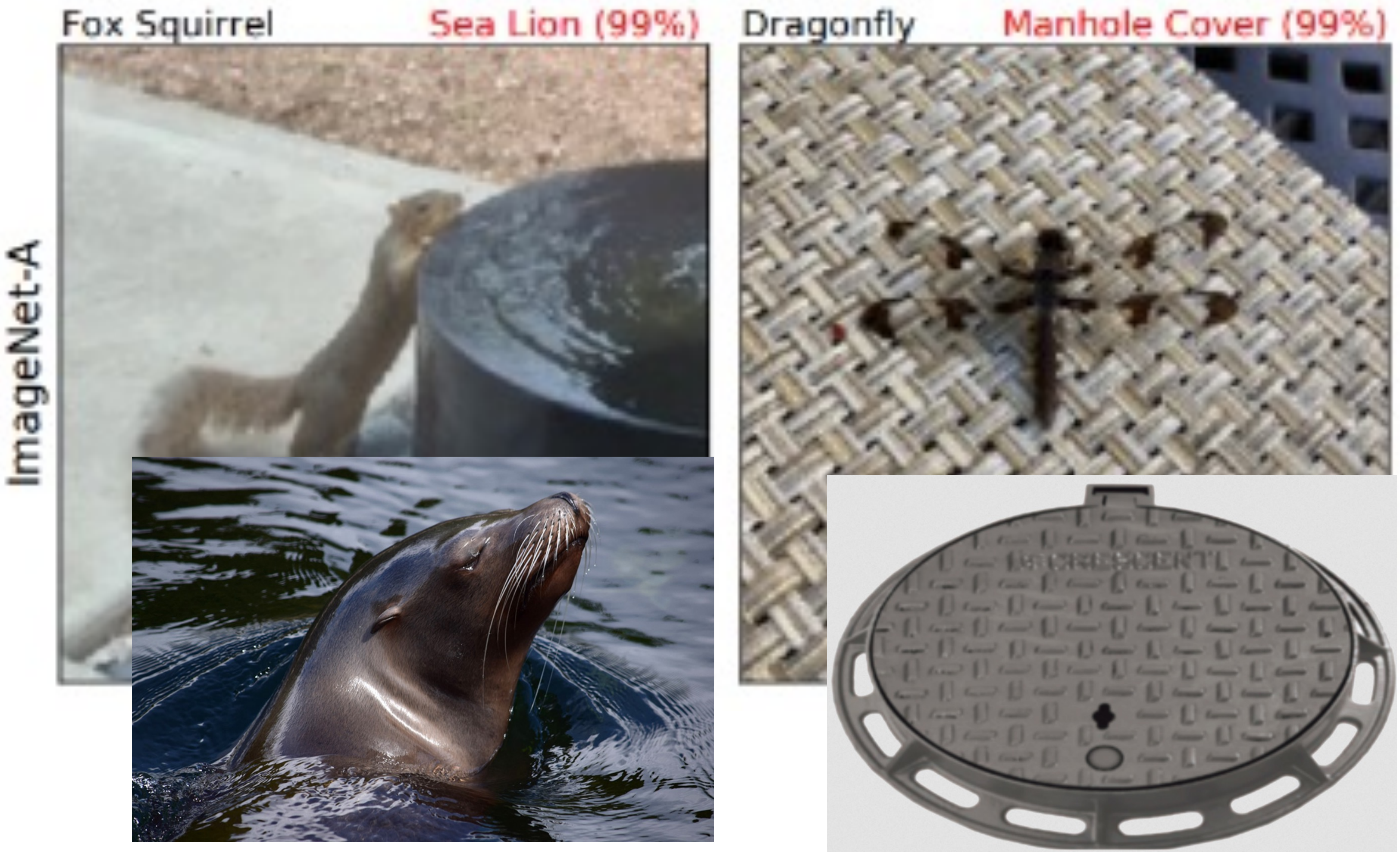
Imagenet is biased (in a neutral sense) towards texture ...



Hendrycks et. al. 202

Biases in ImageNet

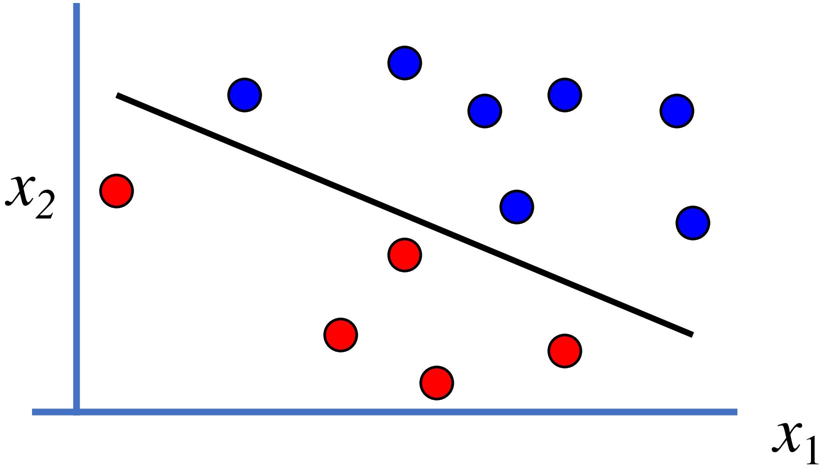
Imagenet is biased (in a neutral sense) towards texture ...



Hendrycks et. al. 202

Recall this visualization of ML prediction

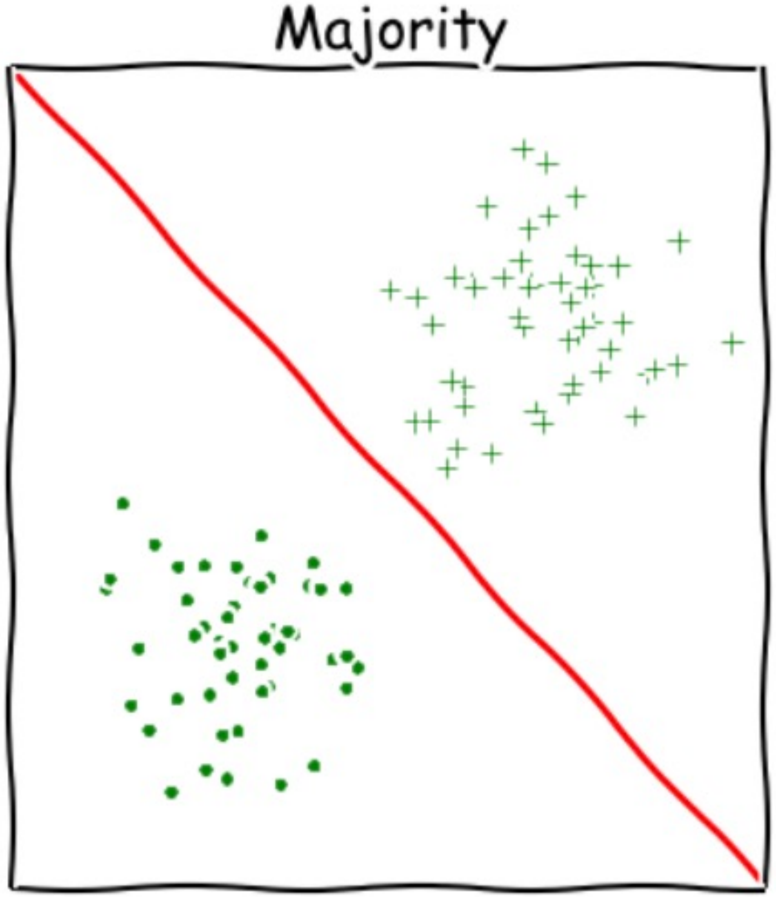
- Logistic regression is trying to fit a **line** that separates data instances where *y* = 1 from those where *y* = 0



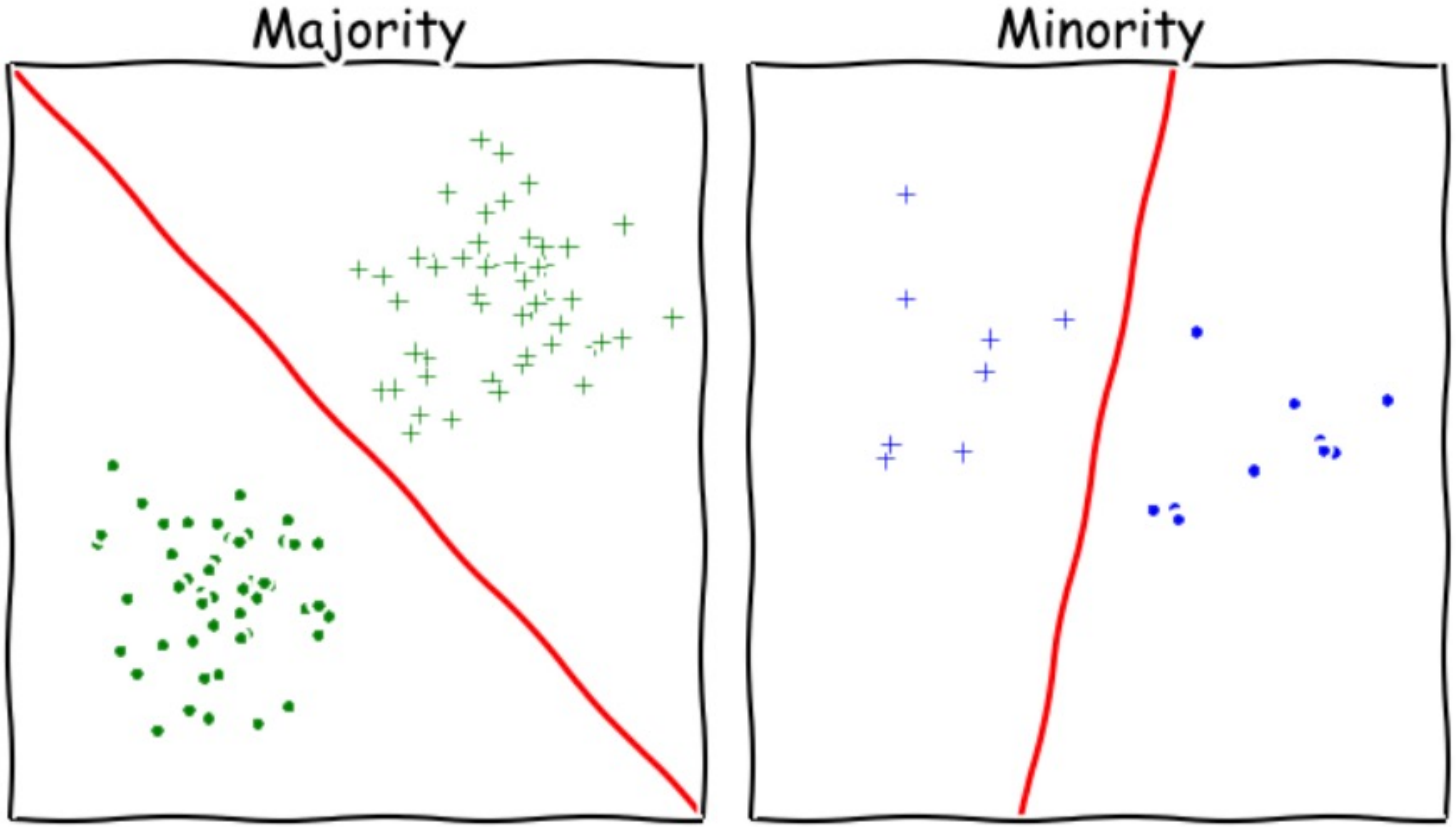
$$\theta^T \mathbf{x} = 0$$

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

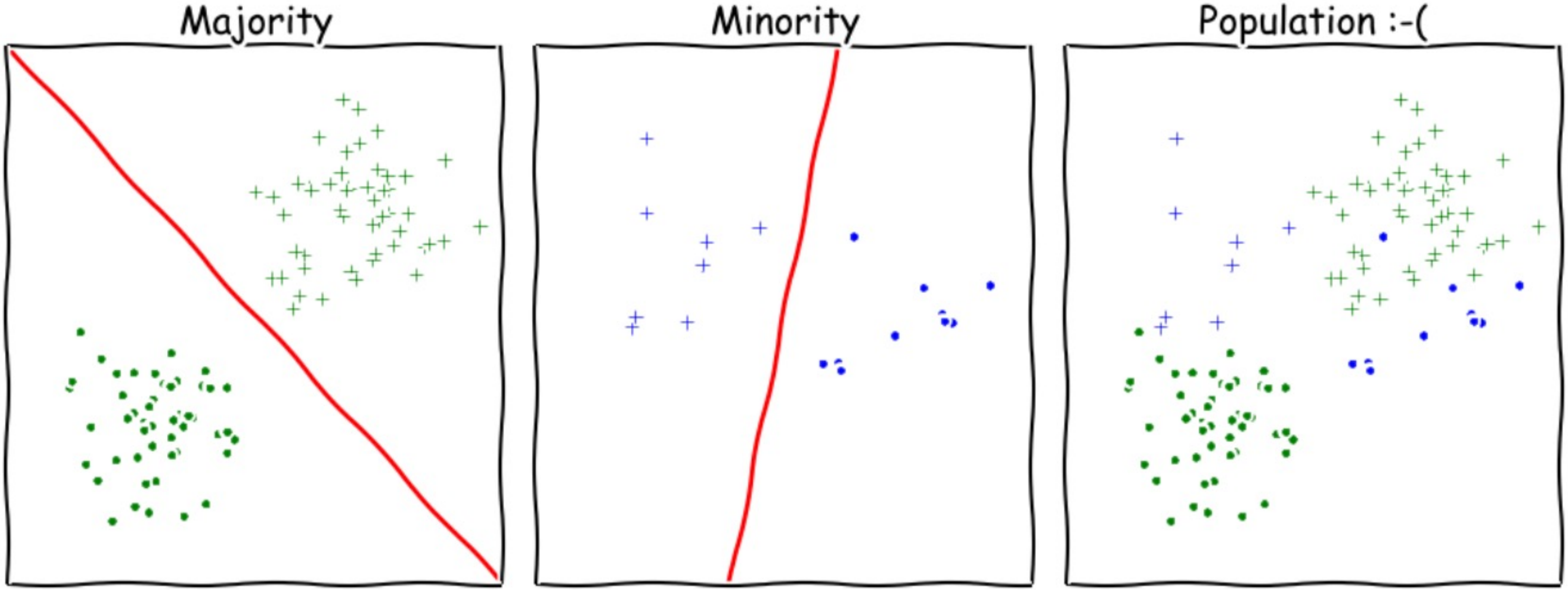
- We call such data (or the functions generating the data) "**linearly separable**"
- **Naïve bayes is linear too** as there is no interaction between different features.



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.”

Biases in ImageNet

... but the dataset also overrepresents males, light-skinned people, and adults between the ages of 18 & 40.

Yang et. al 2020
https://dl.acm.org/doi/10.1145/3351095.3375709

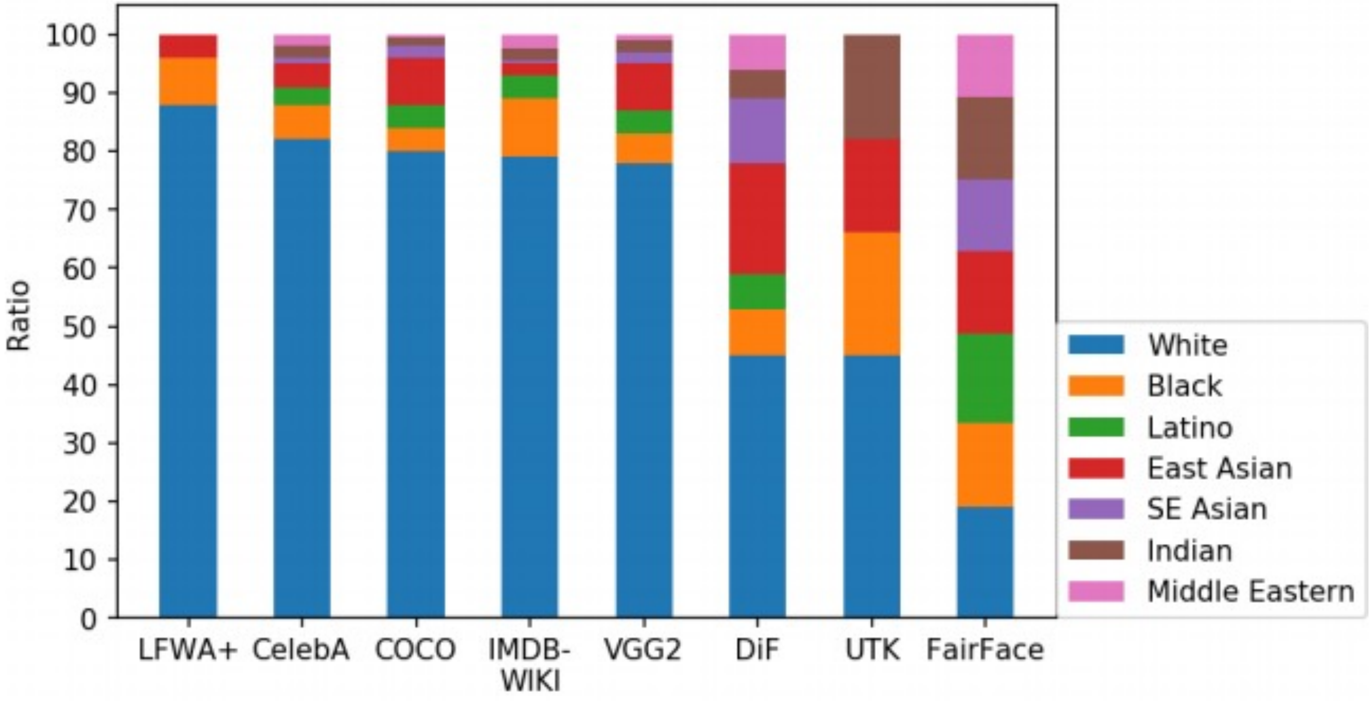


Figure 2: Racial compositions in face datasets.

Kärkkäinen & Joo 2019
https://arxiv.org/pdf/1908.04913.pdf

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 since 2014

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

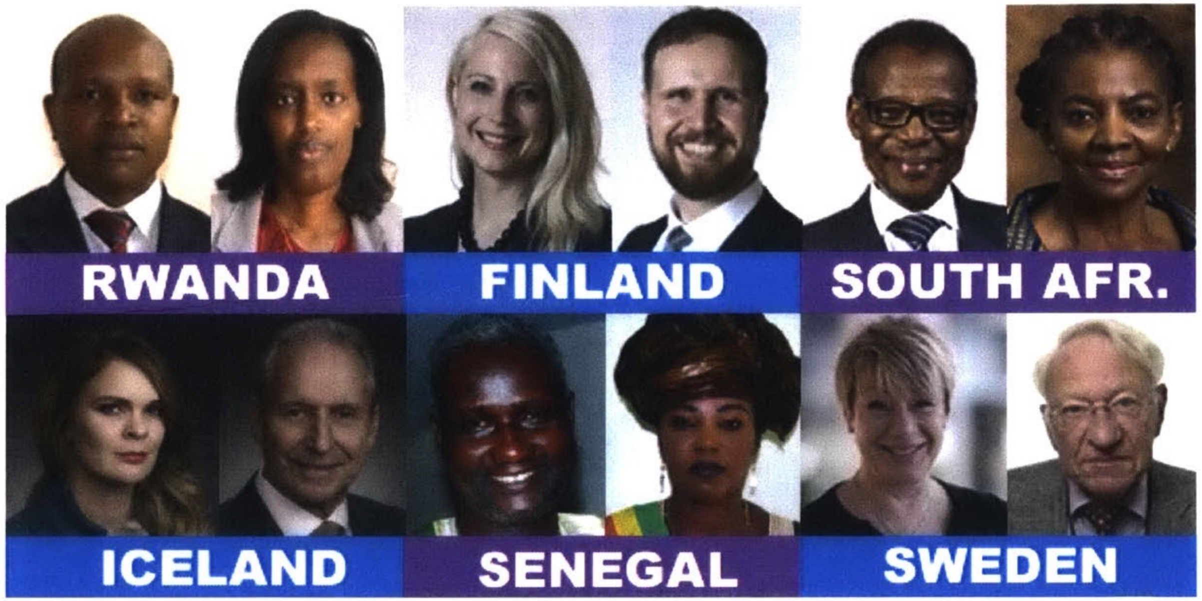
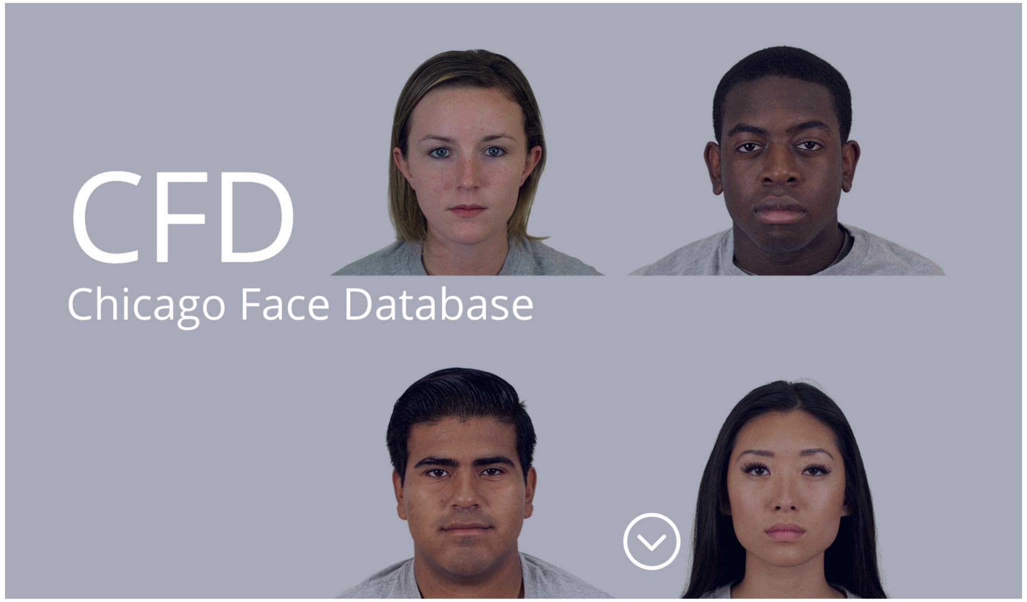
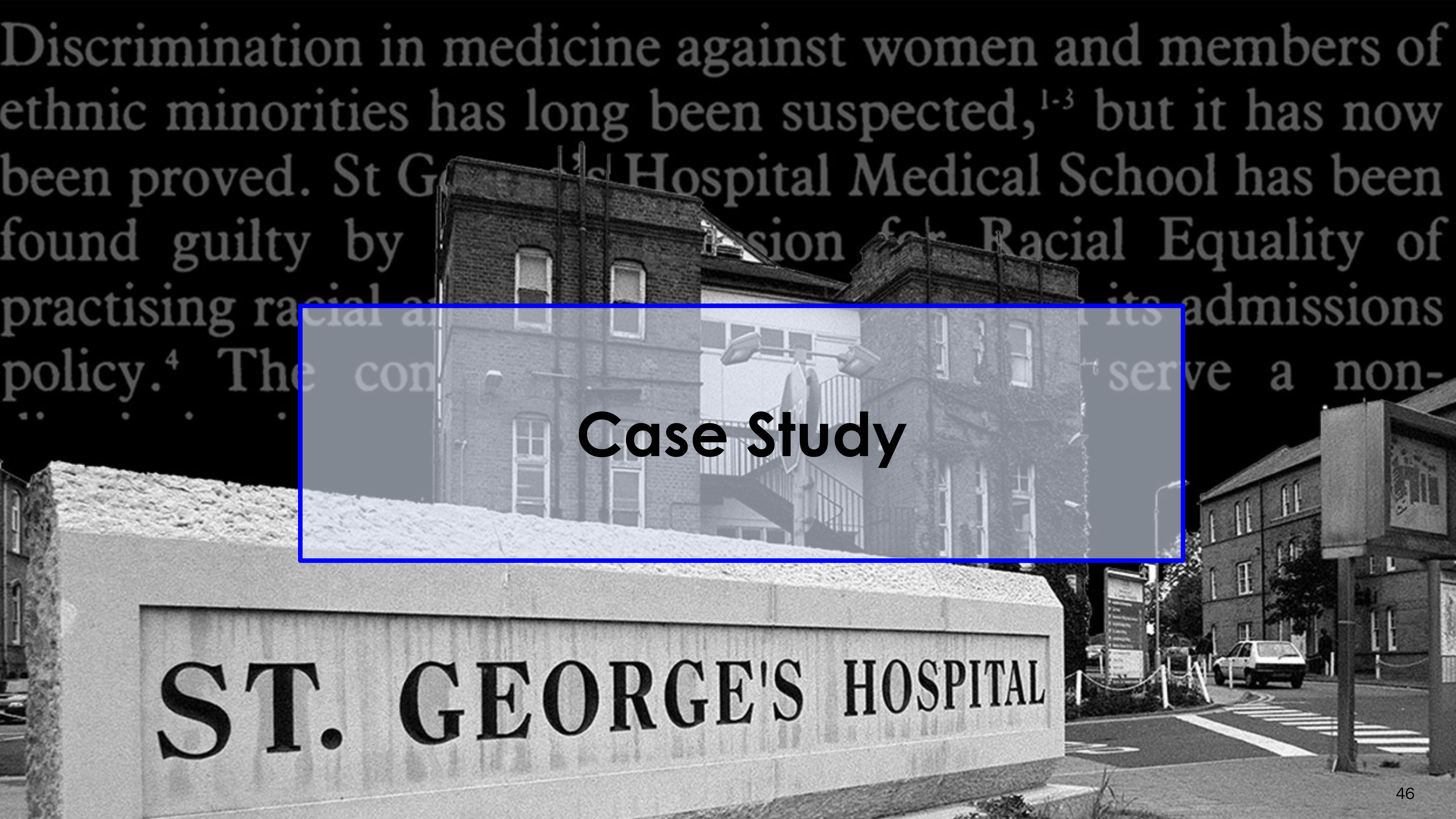


Figure 12. Sample Images from Pilot Parliaments Benchmark

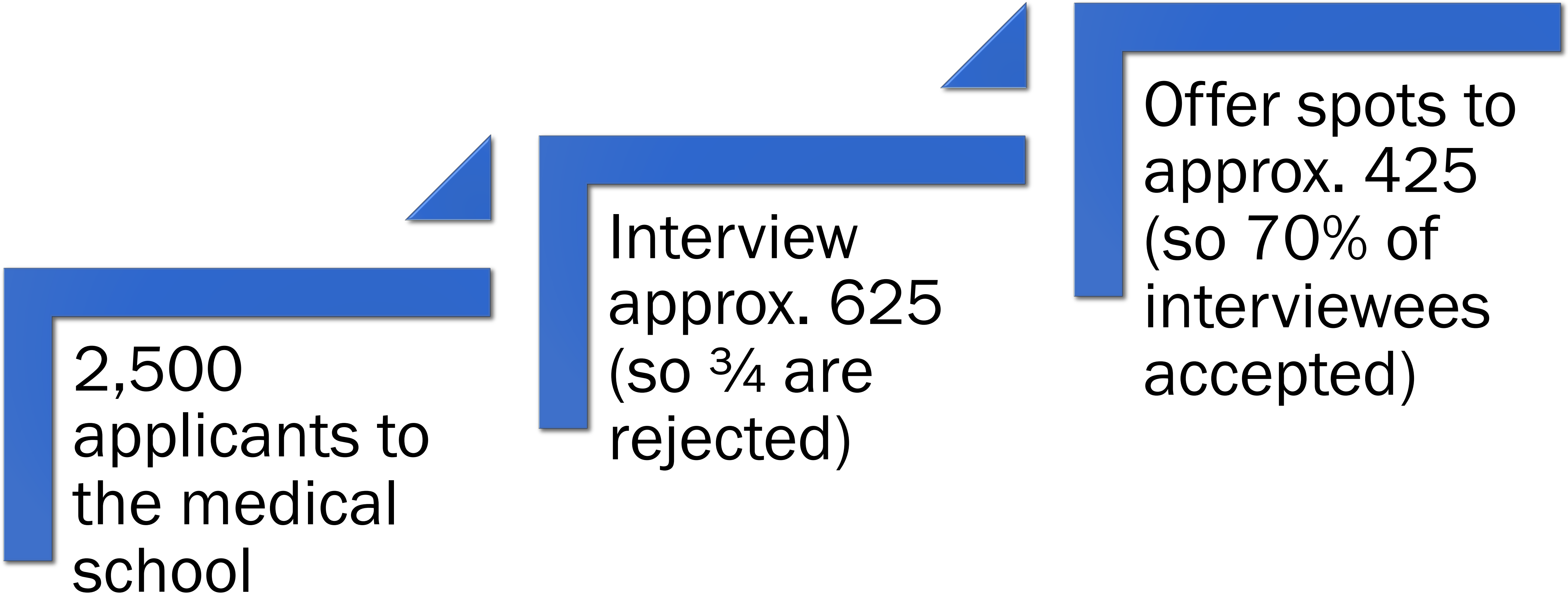


Case Study



Discrimination in medicine against women and members of ethnic minorities has long been suspected,[1-3] but it has now been proved. St G[illegible]'s Hospital Medical School has been found guilty by [illegible]sion for Racial Equality of practising racial a[illegible] its admissions policy.[4] The con[illegible] serve a non-

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

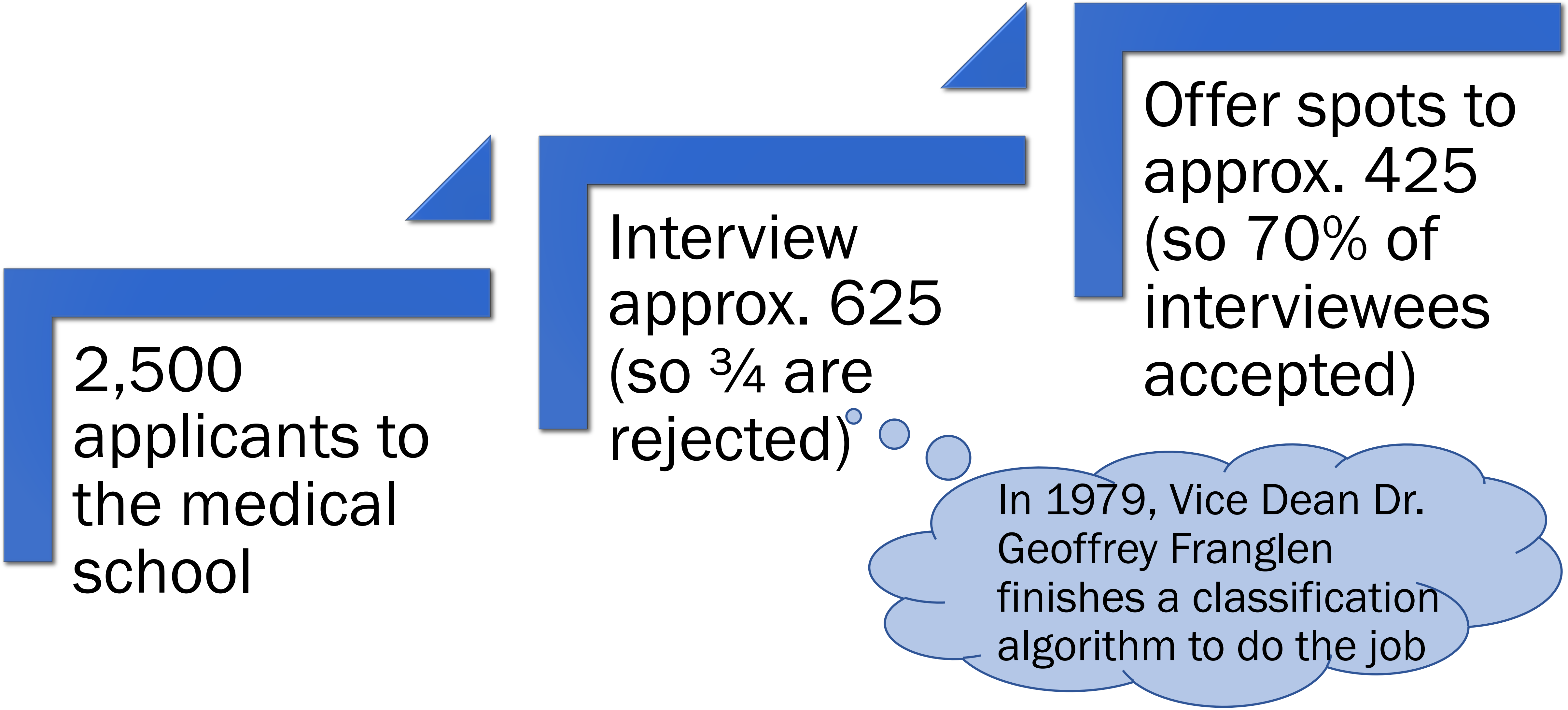


2,500 applicants to the medical school

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



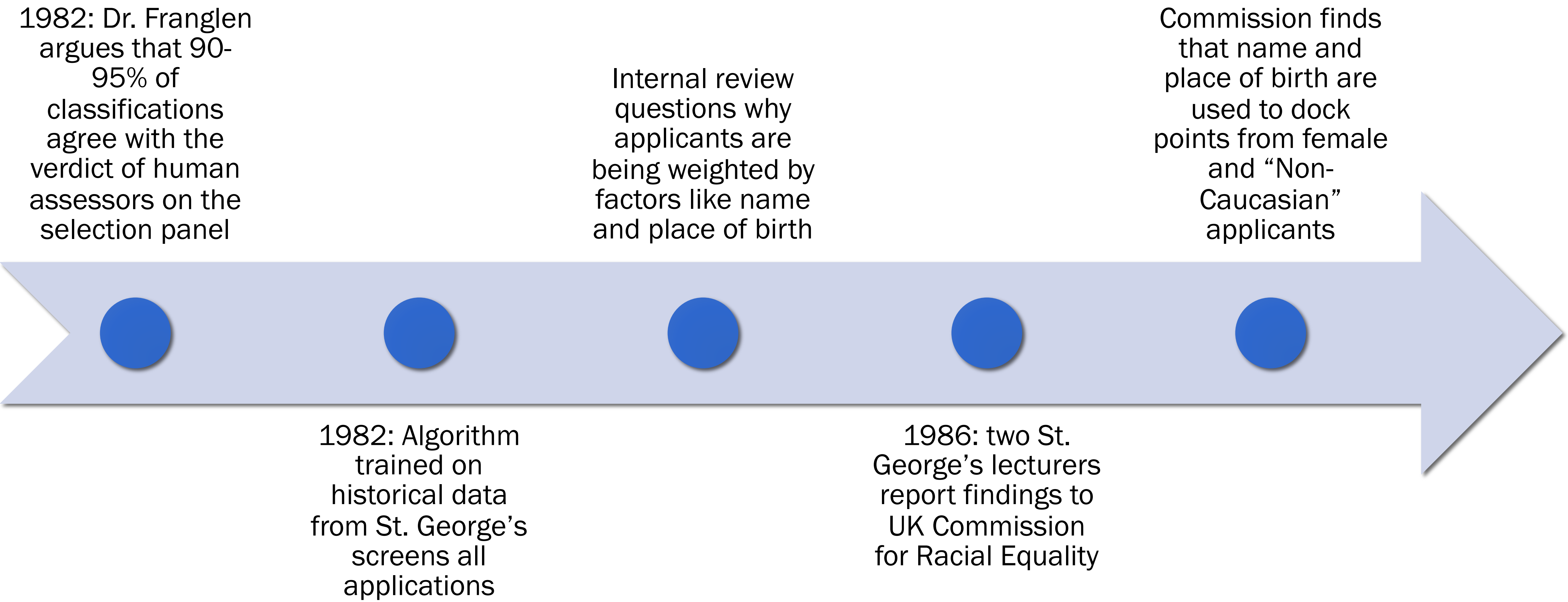
2,500 applicants to the medical school

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

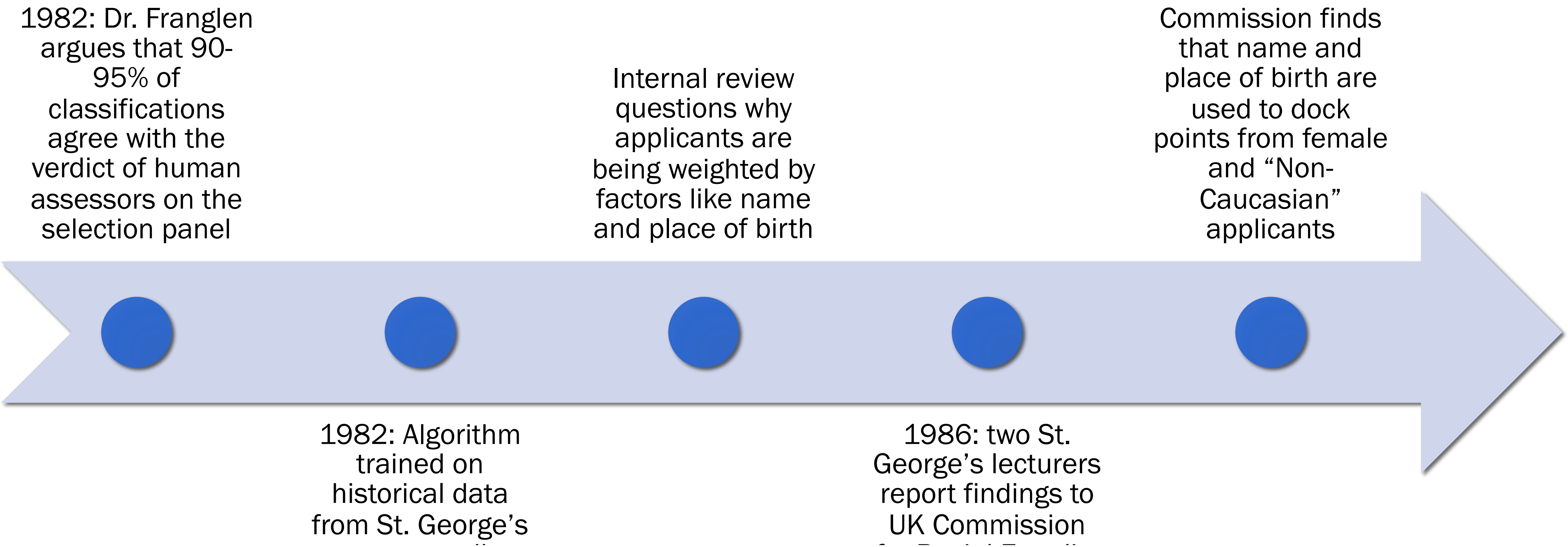
1982: Algorithm trained on historical data from St. George’s screens all applications

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



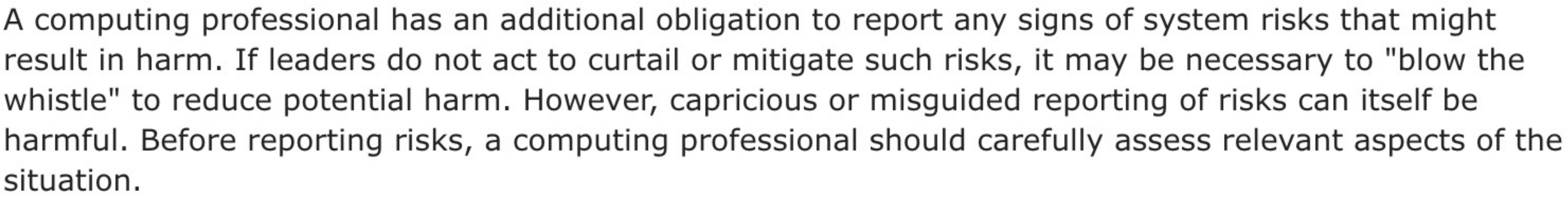
1982: Dr. Franglen argues that 90-95% of classifications agree with the verdict of human assessors on the selection panel

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.

Definitions of Bias

Nissenbaum: we will use “bias to refer to computer systems that **systematically and unfairly discriminate** against certain individuals or groups of individuals in favor of others.

A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate”

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

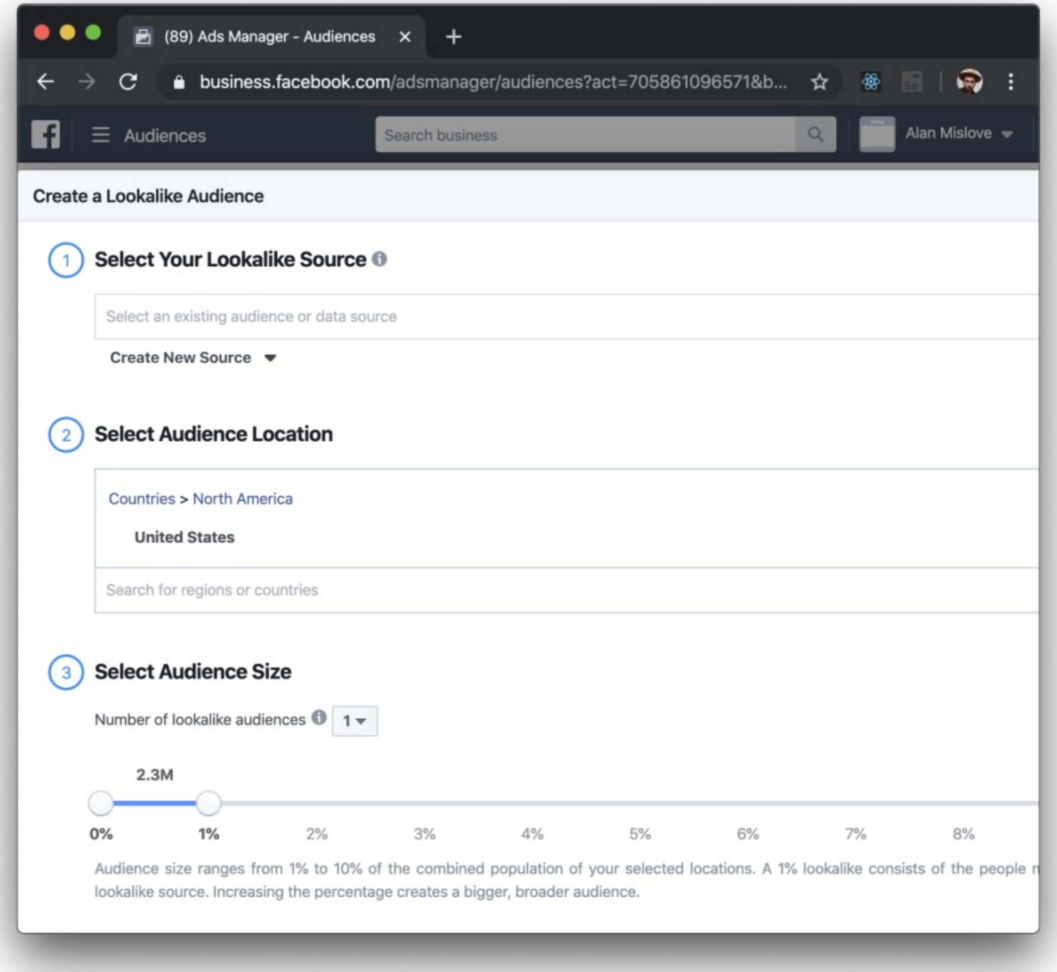
Case Study: Facebook Ads & Job/Housing Recommendations

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


(89) Ads Manager - Audiences
business.facebook.com/adsmanager/audiences?act=705861096571&b...
Audiences
Search business
Alan Mislove

Create a Lookalike Audience

1 Select Your Lookalike Source

Select an existing audience or data source

Create New Source

2 Select Audience Location

Countries > North America

United States

Search for regions or countries

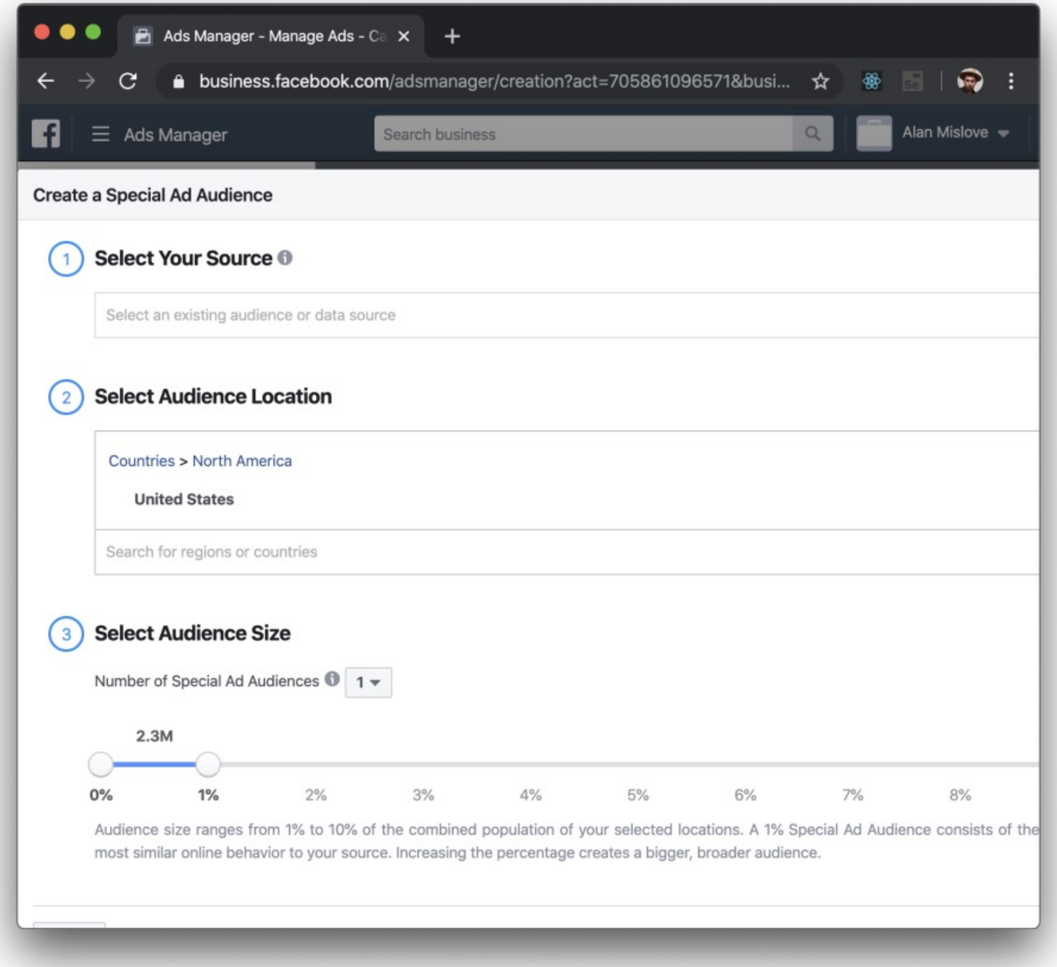
3 Select Audience Size

Number of lookalike audiences 1

2.3M

0% 1% 2% 3% 4% 5% 6% 7% 8%

Audience size ranges from 1% to 10% of the combined population of your selected locations. A 1% lookalike consists of the people r lookalike source. Increasing the percentage creates a bigger, broader audience.


Ads Manager - Manage Ads - Ca
business.facebook.com/adsmanager/creation?act=705861096571&busi...
Ads Manager
Search business
Alan Mislove

Create a Special Ad Audience

1 Select Your Source

Select an existing audience or data source

2 Select Audience Location

Countries > North America

United States

Search for regions or countries

3 Select Audience Size

Number of Special Ad Audiences 1

2.3M

0% 1% 2% 3% 4% 5% 6% 7% 8%

Audience size ranges from 1% to 10% of the combined population of your selected locations. A 1% Special Ad Audience consists of the most similar online behavior to your source. Increasing the percentage creates a bigger, broader audience.

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

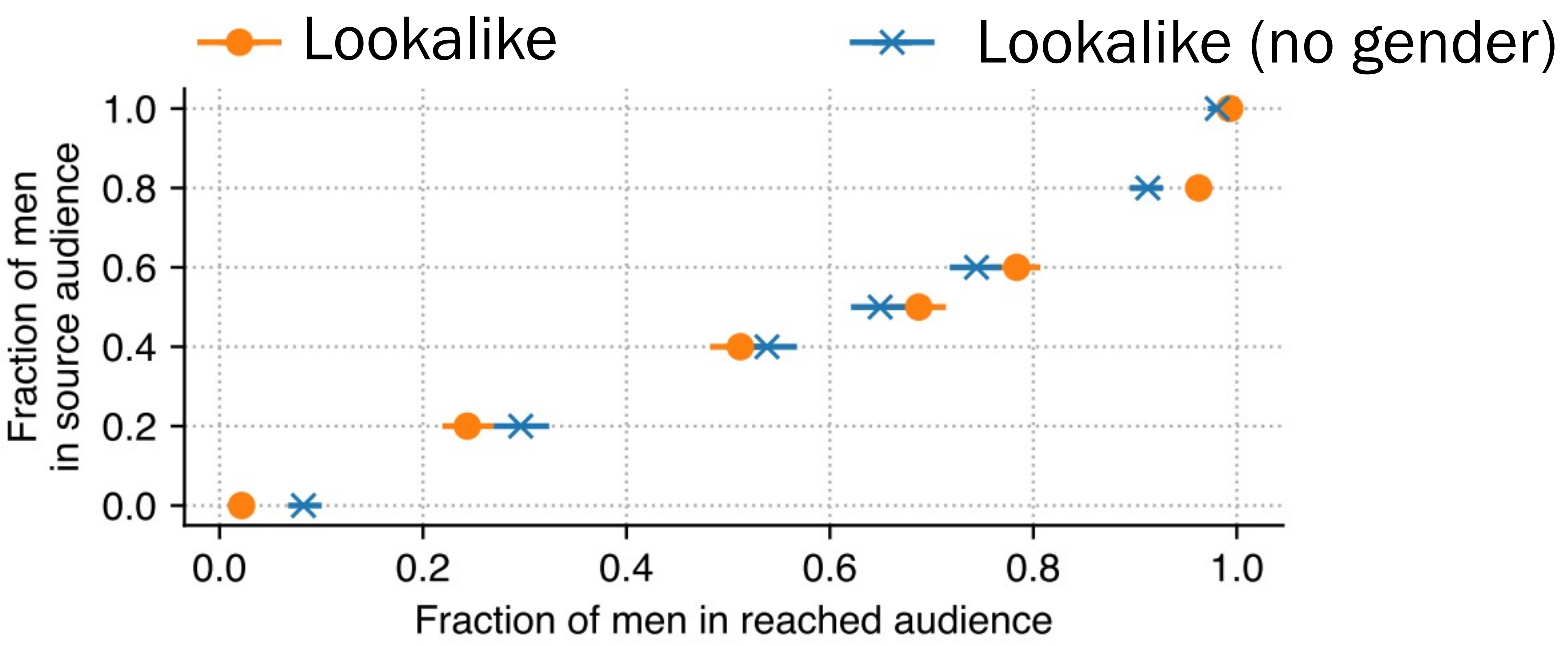
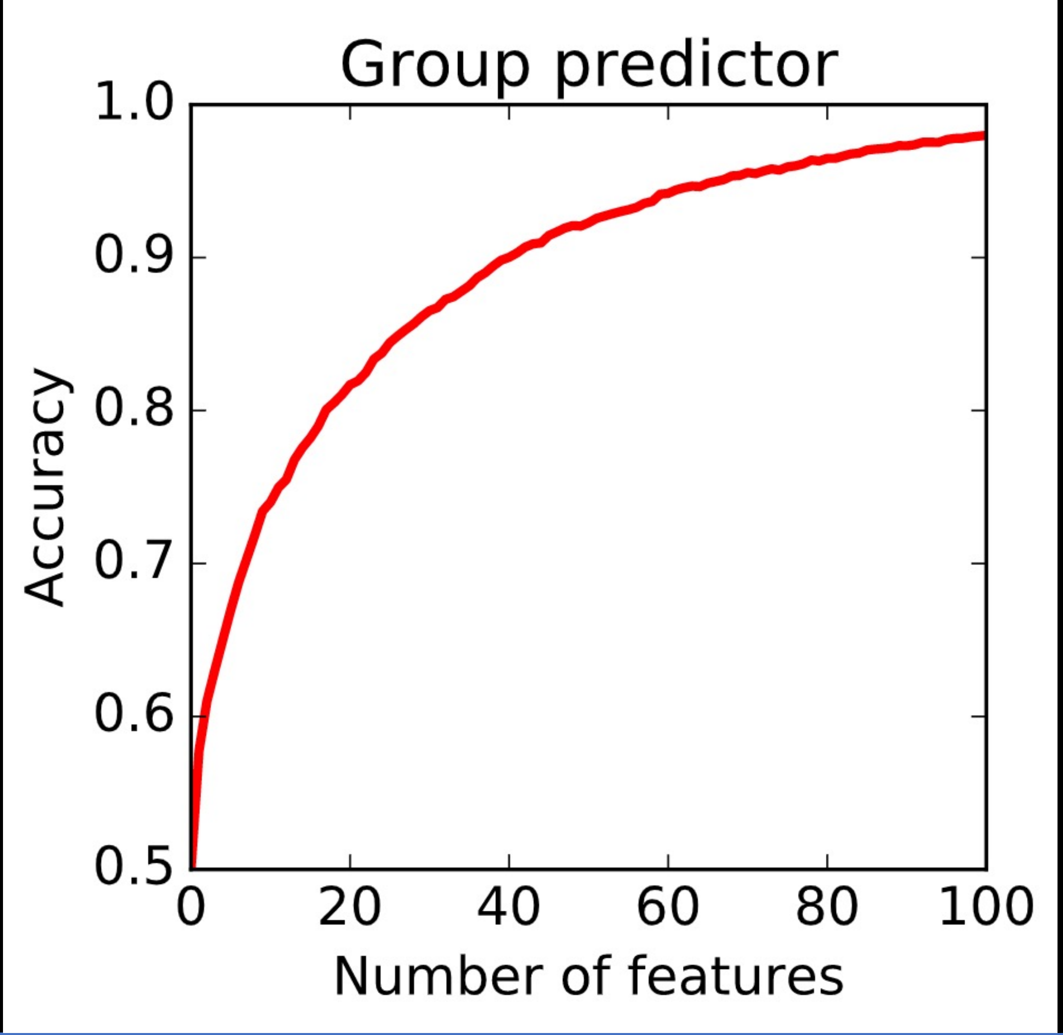


Figure 2: Gender breakdown of ad delivery to Lookalike and Special Ad audiences created from the same source audience with varying fraction of male users, using the same ad creative. We can observe that both Lookalike and Special Ad audiences reflect the gender distribution of the source audience, despite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

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.

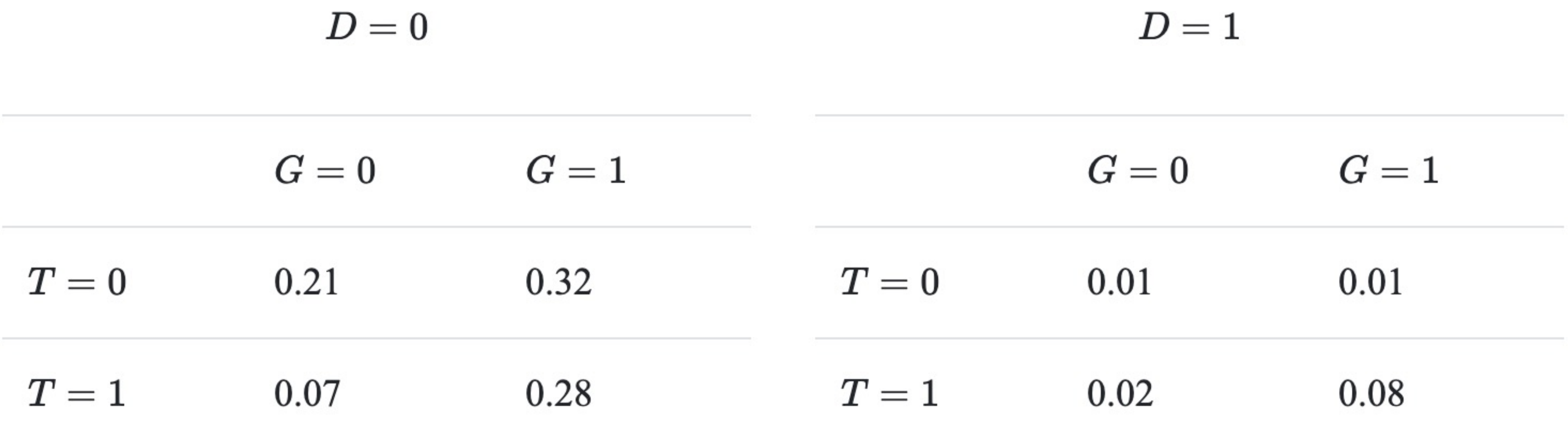
Fairness through unawareness
(facebook example helps with this)

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

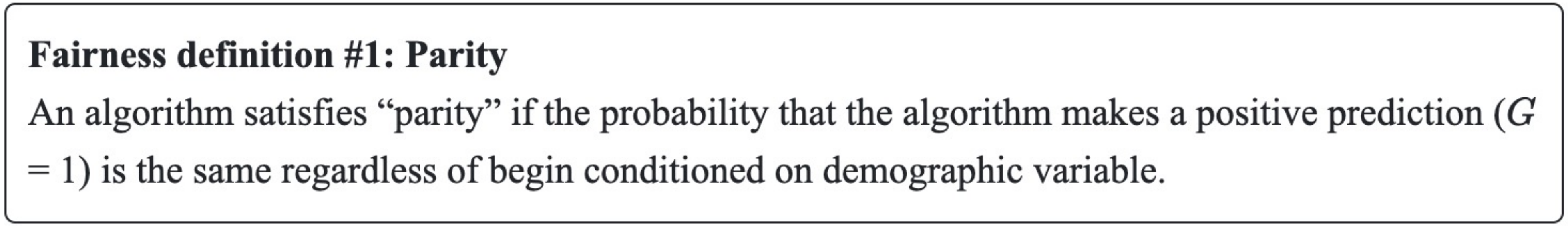
False Positives and False Negatives

	Condition $y = 1$	Condition $y = 0$
Event $\hat{y} = 1$	True Positive	False Positive
Event $\hat{y} = 0$	False Negative	True Negative

This table is sometimes called a “confusion matrix”

Errata: The labels on this matrix were fixed during class

Parity



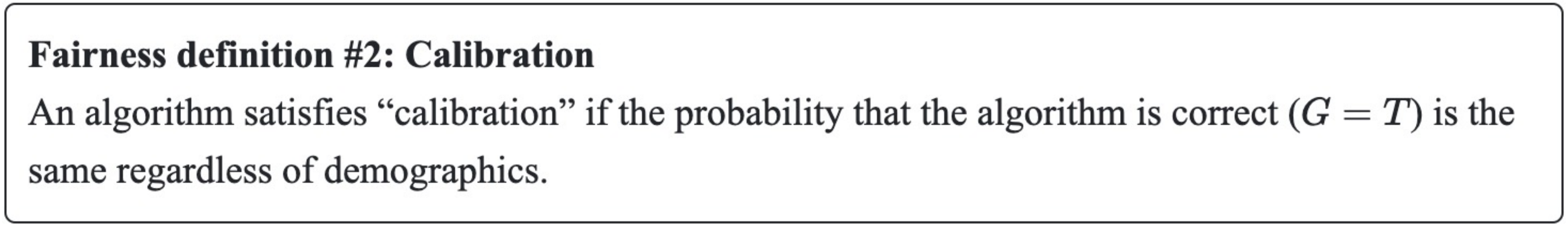
Fairness definition #1: Parity

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)

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

Calibration



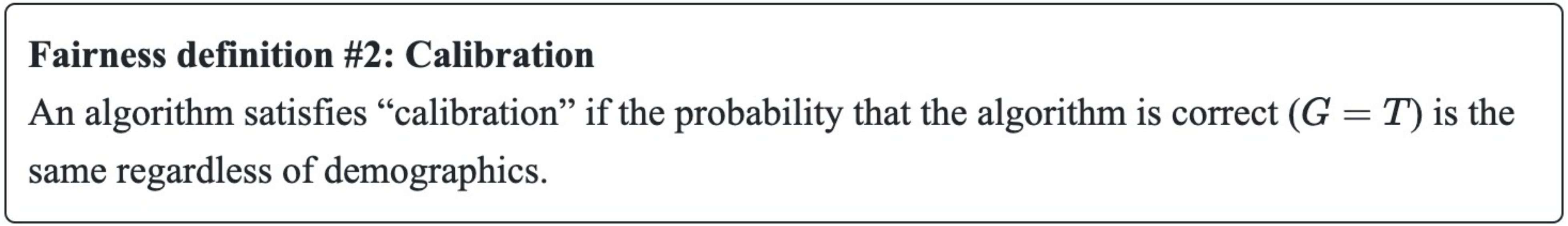
Fairness definition #2: Calibration

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.

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?

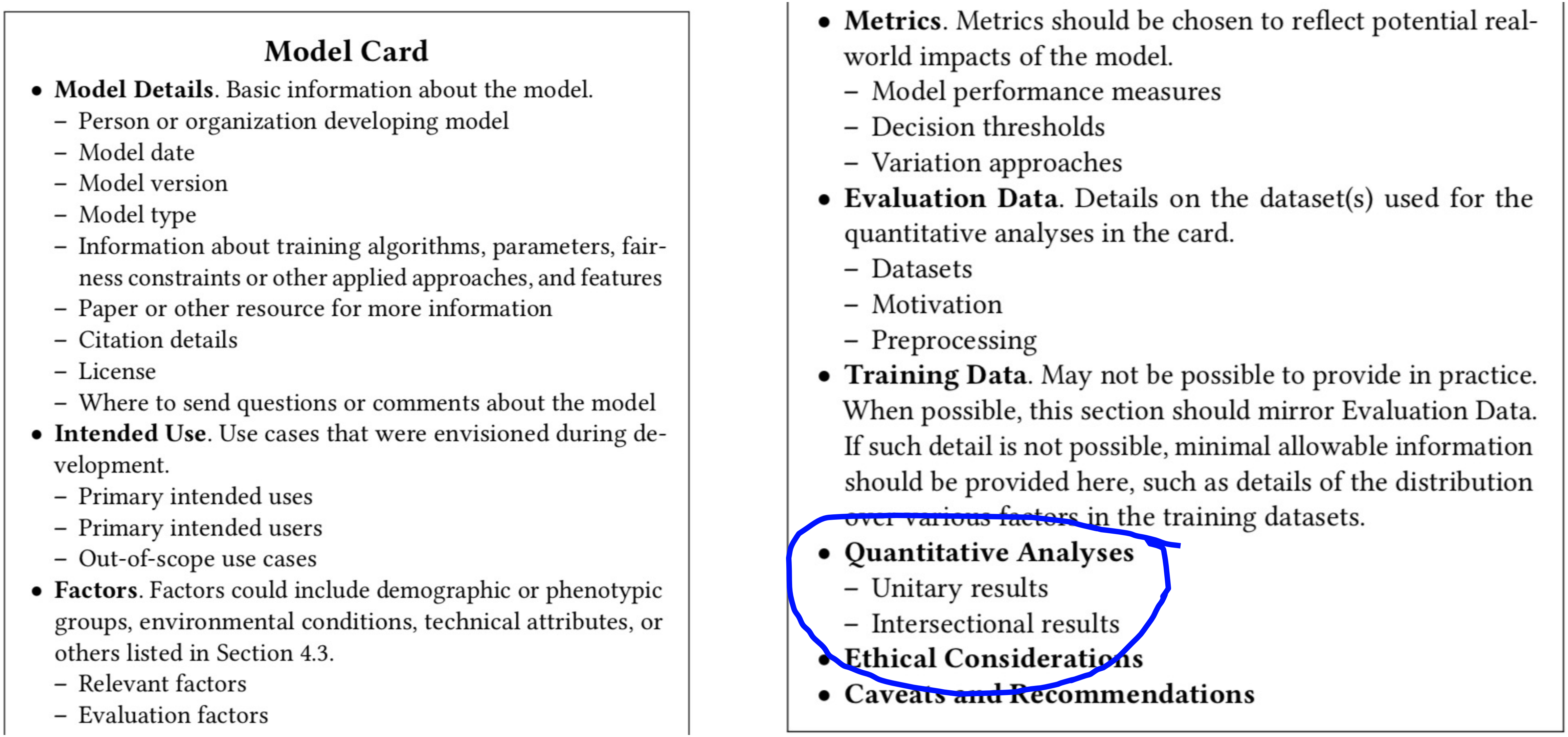
- If the classifier is significantly less good at identifying candidates e.g. for a surgery in a minority group (relative to the data), the candidates accepted might have worse outcomes, leading to future bias & over or under treatment.
- Quality of Service Disparity might then lead to an Allocation 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)

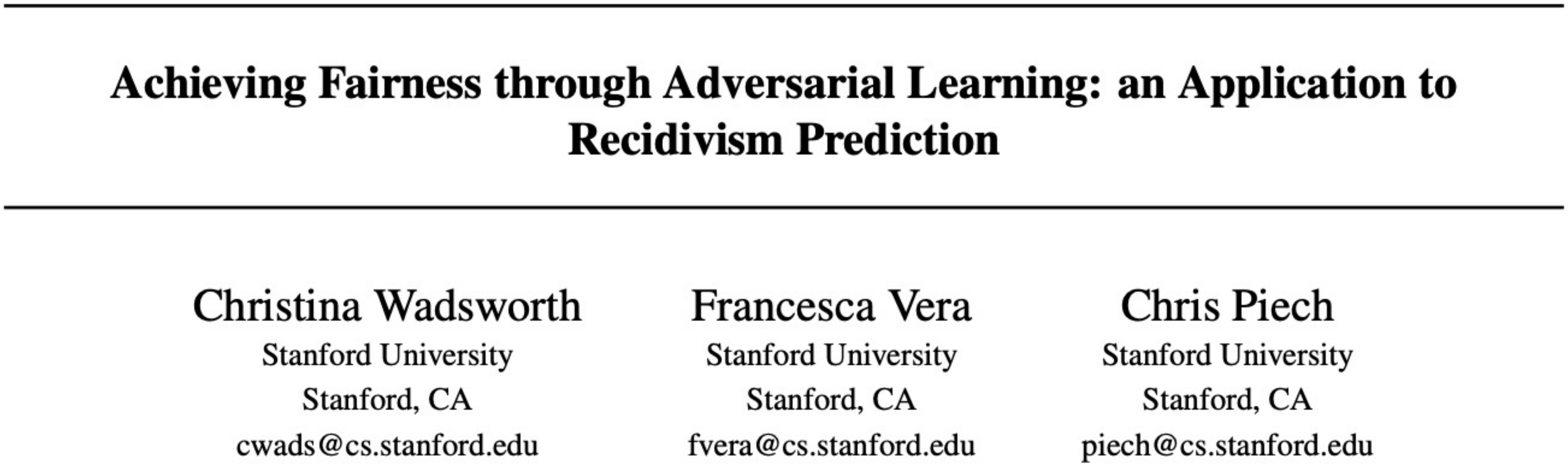


Model Card

- **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

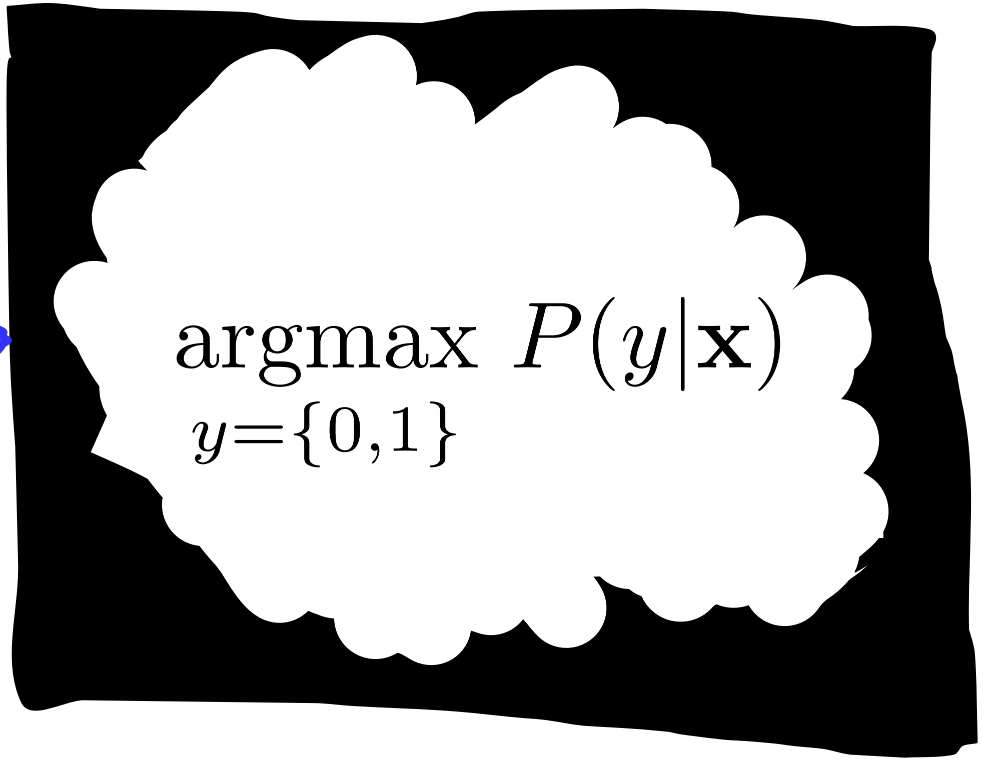
Chris Piech
Stanford University
Stanford, CA
piech@cs.stanford.edu

Seniors at the time
they wrote it

COMPAS: Predicting "Recidivism"

$\mathbf{x}$

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



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

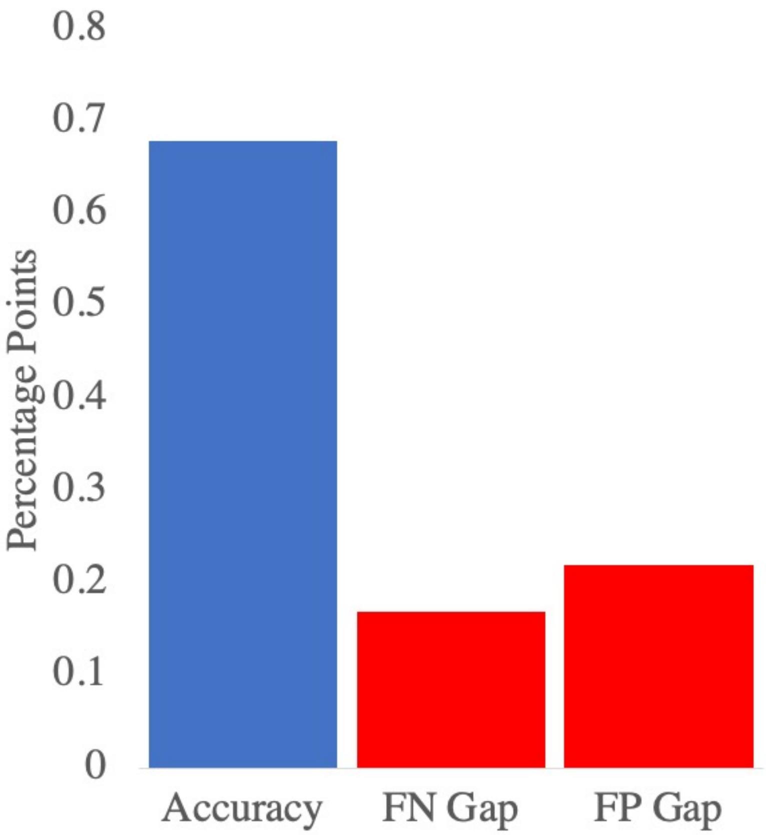
$\hat{y} = 0$

Will they commit a crime again

Was in use in California and Florida

COMPAS: Biased Against Black Inmates

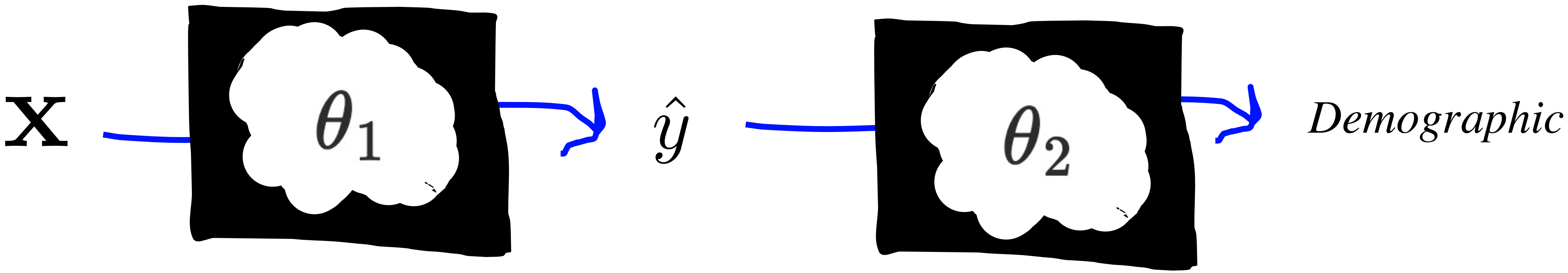
Before: Compas is Biased



Can We Train Out Bias?

Model 1: Prediction

Model 1: Extract Demographic



Model 1 should be accurate

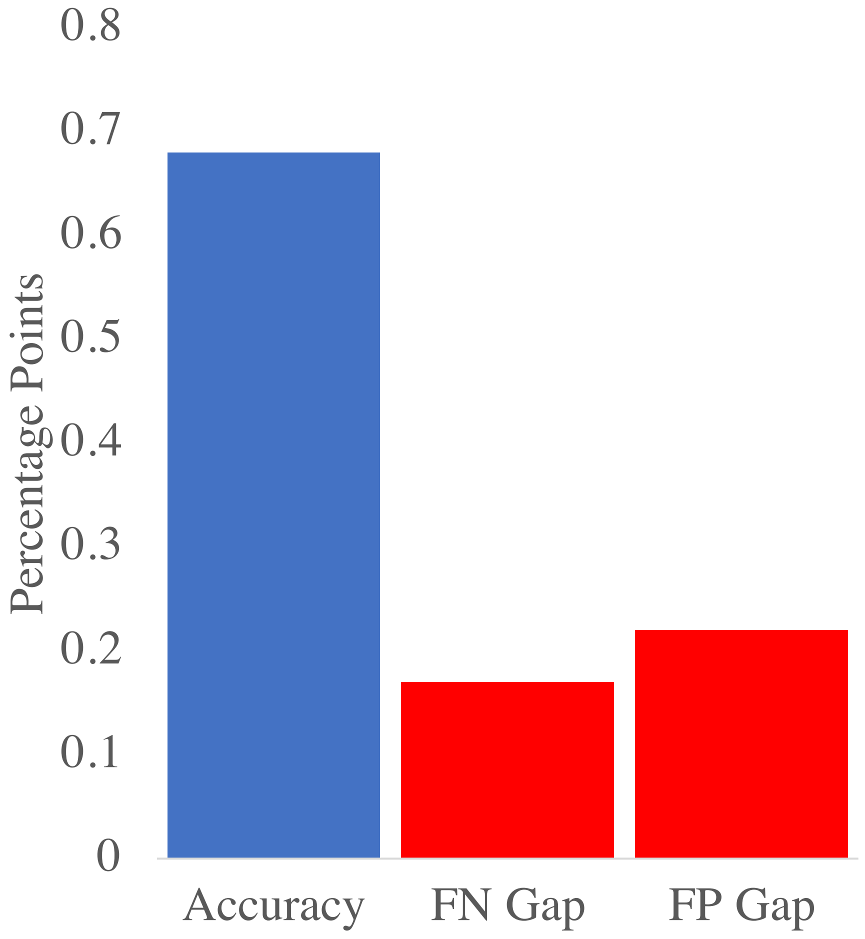
*Model 2 should be **in**accurate*

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

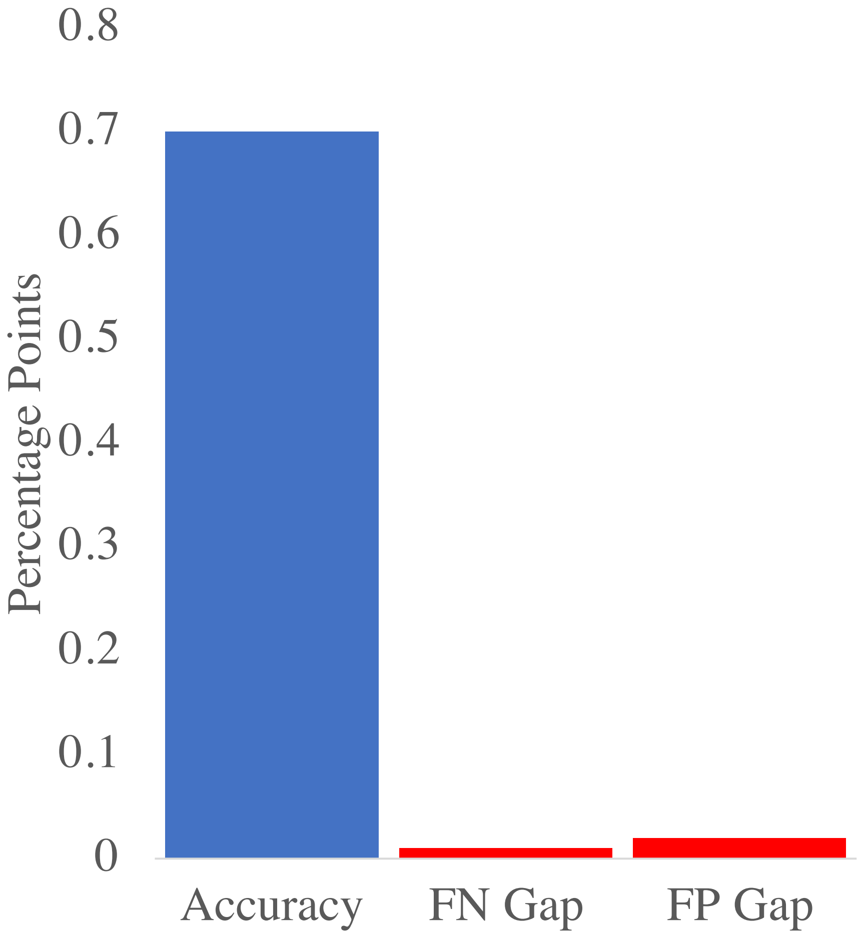
*note in the paper these were neural nets

Can We Train Out Bias?

Before: Compas is Biased



After: FN Gap is reduced

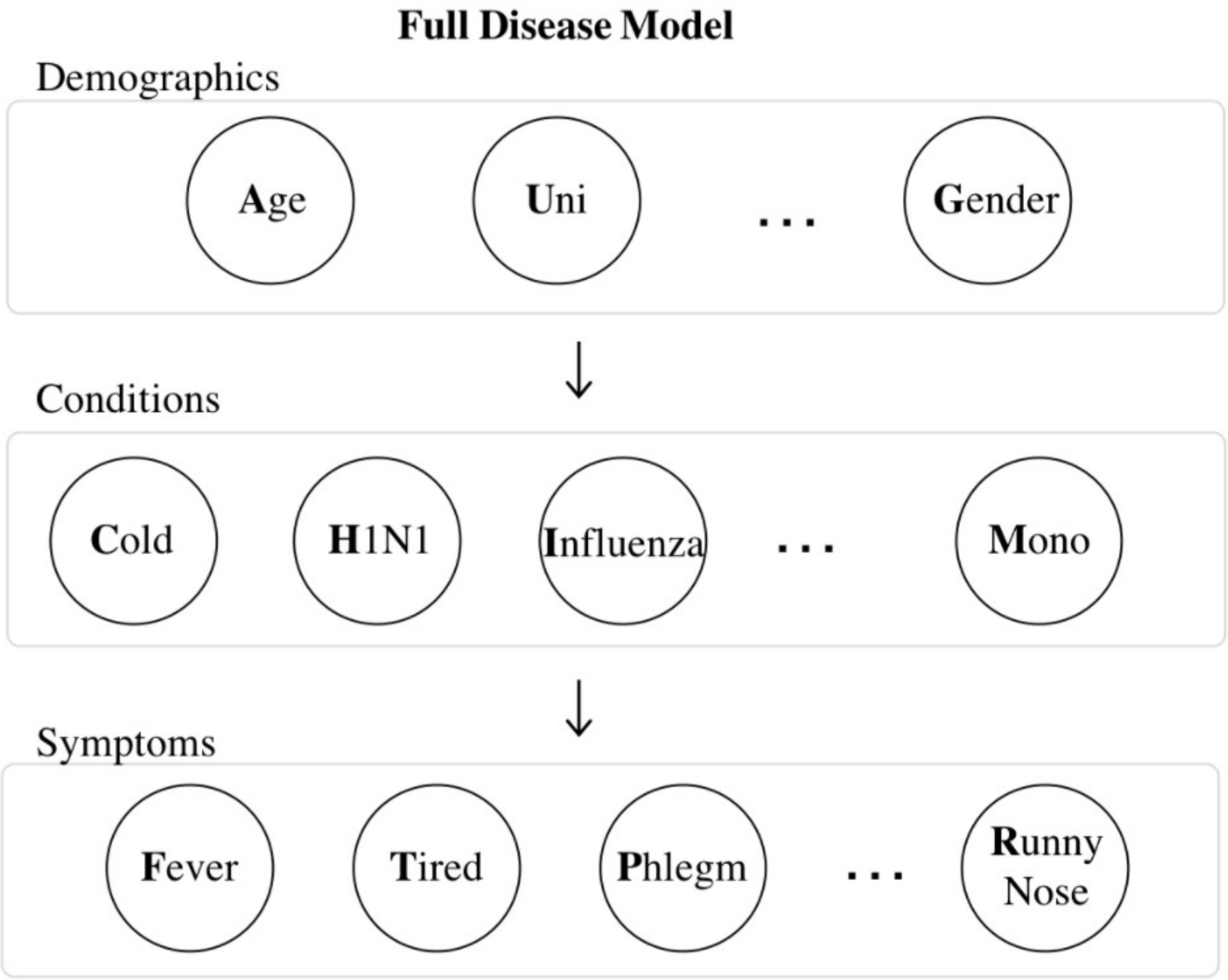


Their Conclusion

DON’T USE BLACK BOX ALGORITHMS TO MAKE RECIDIVISM PREDICTIONS

Use A Bayes Net?

Bayes Nets > Black Box?



Justice Beyond Distribution

Justice beyond Distribution

Zero-sum:

Resources and outcomes are fixed: the only task of justice is to fairly distribute them between individuals and groups. Improving the outcomes of the least-well-off group means worse outcomes for the best-off group (although in many cases only slightly worse).

Leveling Up & Expanding the Pie:

Outcomes and Resources are not fixed: justice means distributing outcomes fairly *and* increasing the number of good outcomes. Improving outcomes of the least-well-off group need not come at the expense of any other group.

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.

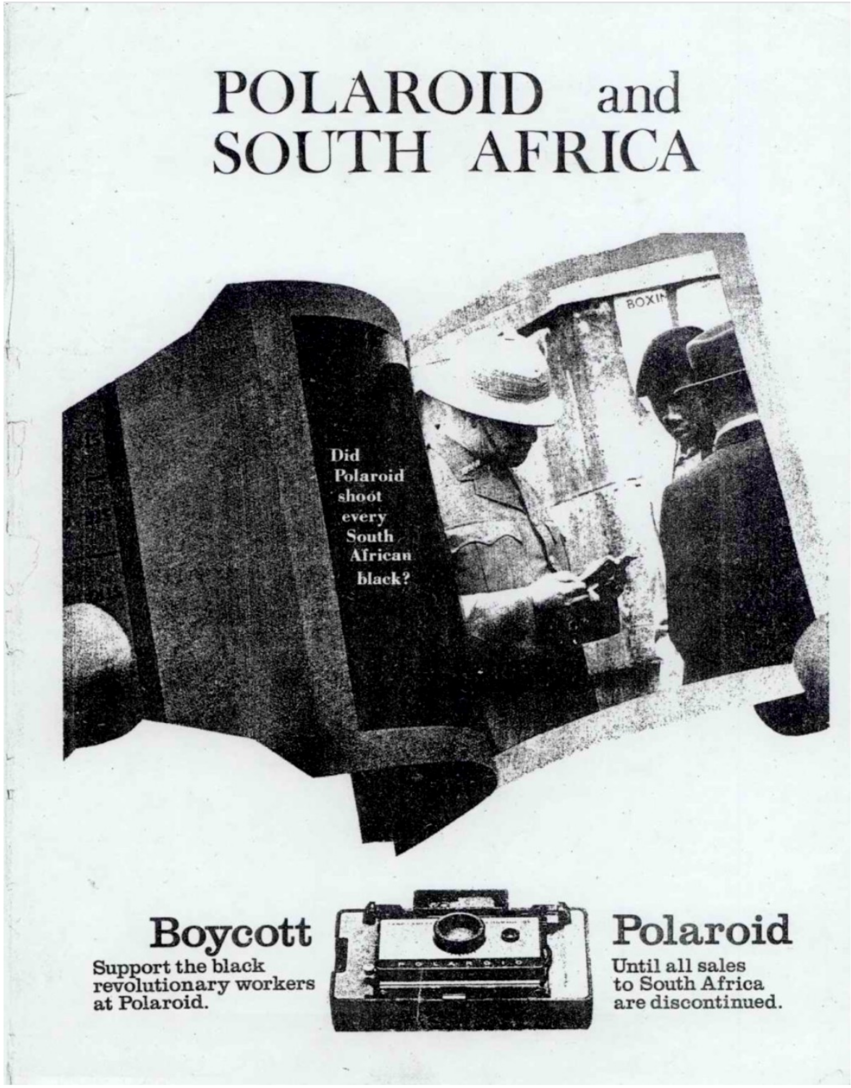
This was used to create passbook photographs for the Apartheid government.

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.'”



POLAROID and SOUTH AFRICA

Boycott
Support the black revolutionary workers at Polaroid.

Polaroid
Until all sales to South Africa are discontinued.

Support internal & external efforts to honestly evaluate models

Do your own analysis of the systems you are making.

Ensure that they line up with your values and function for the "greater good."

Work with others inside and outside your company to hold machine learning to the highest standards of fairness.

Timnit Gebru & Margaret Mitchell, recently of Google's Ethical AI team

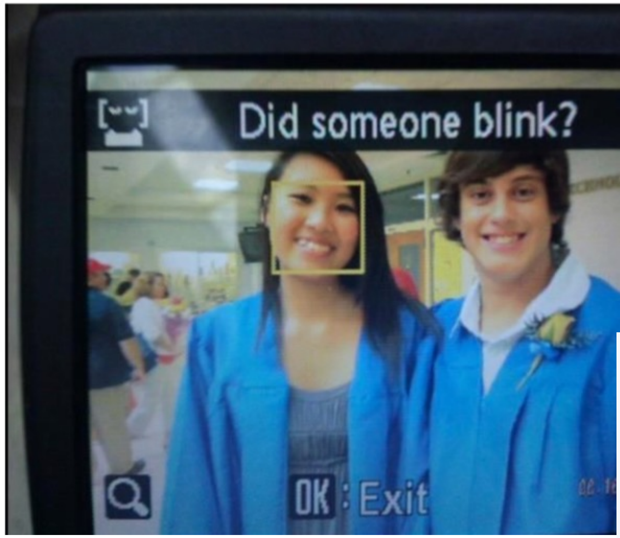
(Pedagogic Pause)

Learning Goals

1. Recognize a hidden ethics issue in ML with respect to protected demographics (and how to solve them)

2. Discuss ways to address them

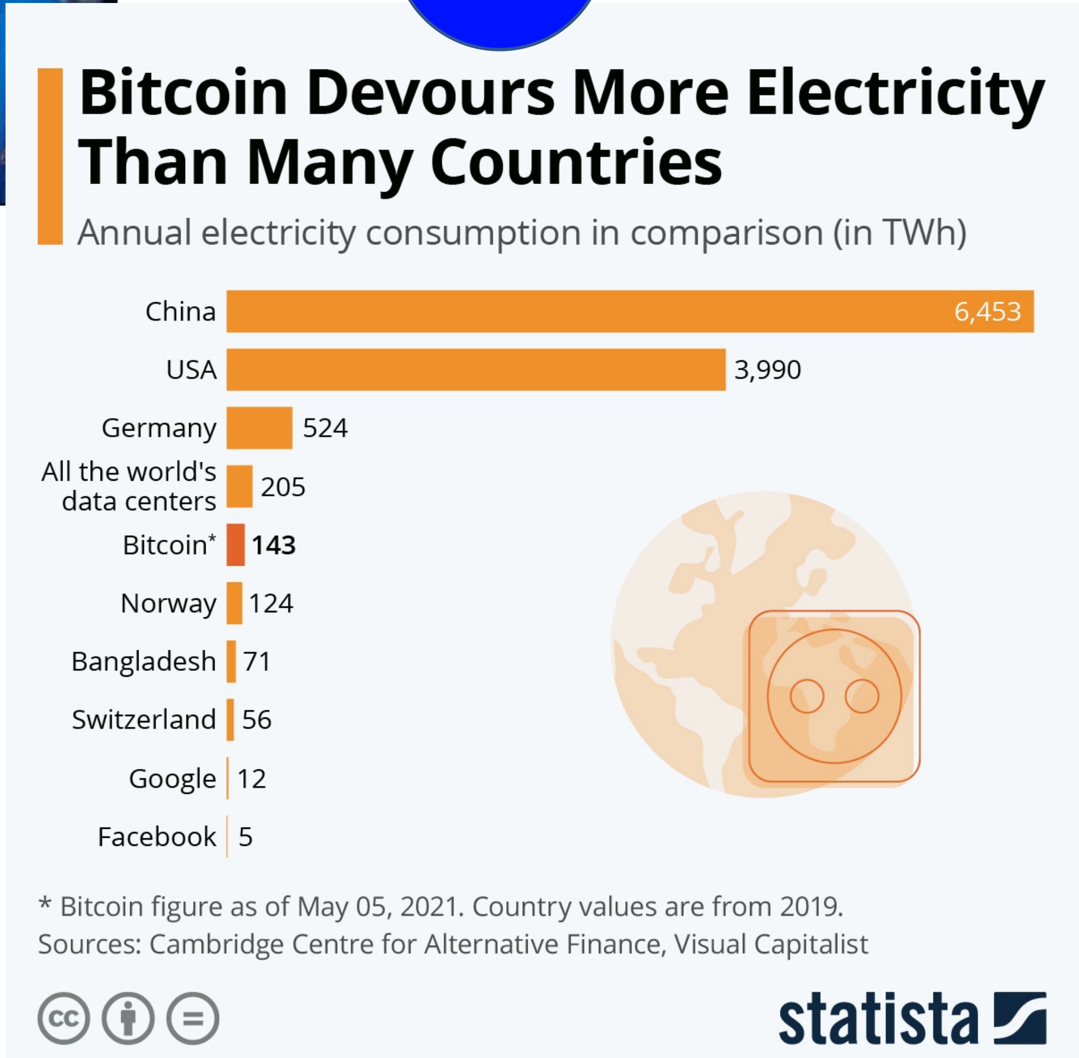
1



2

Facebook slammed by UN for its role in Myanmar genocide

3



Bitcoin Devours More Electricity Than Many Countries

Annual electricity consumption in comparison (in TWh)

* Bitcoin figure as of May 05, 2021. Country values are from 2019.

Sources: Cambridge Centre for Alternative Finance, Visual Capitalist

statista

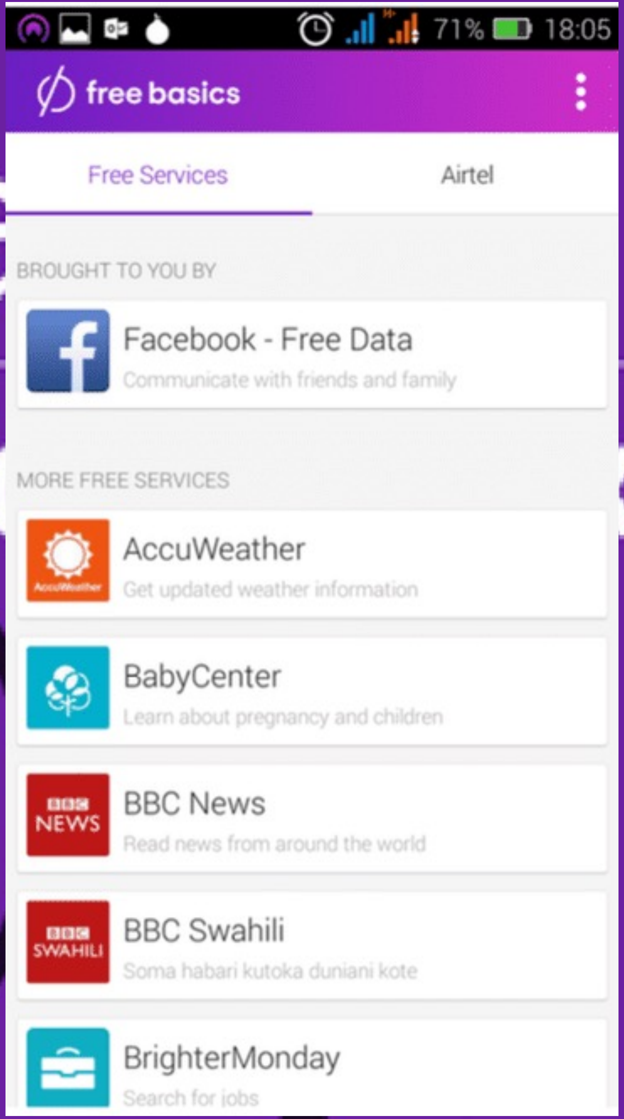
Part 4: The Blind Spots

What are our current blind spots?

(Chris Opinion)

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

Facebook Introduces Free Basics (2015)



BROUGHT TO YOU BY

Facebook - Free Data
Communicate with friends and family

MORE FREE SERVICES

AccuWeather
Get updated weather information

BabyCenter
Learn about pregnancy and children

BBC News
Read news from around the world

BBC Swahili
Soma habari kutoka duniani kote

BrighterMonday
Search for jobs

Junta Starts a Misinformation Campaign Against Rohingya

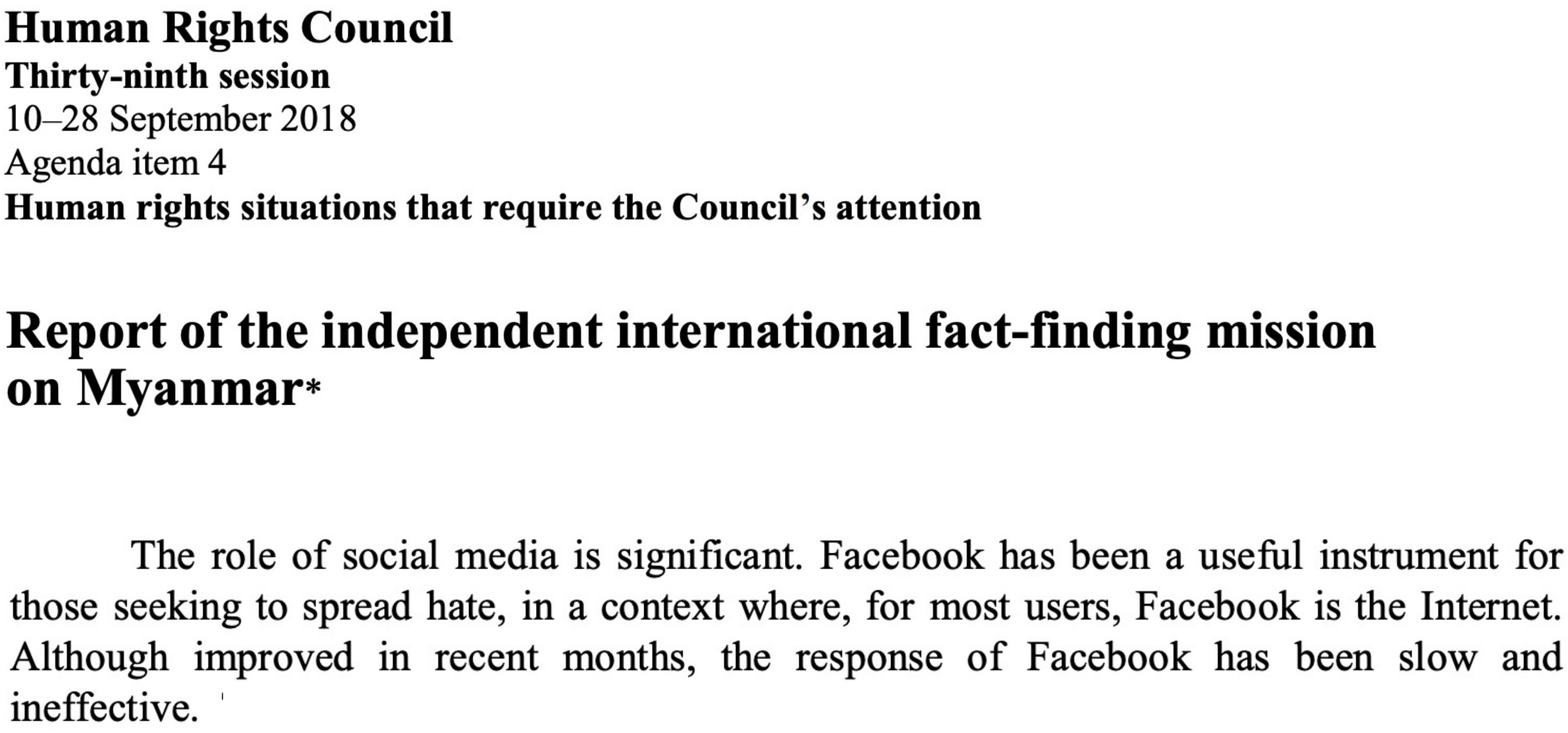
[redacted] [redacted] [redacted]

Facebook: Two Moderators Who Speak Burmese (2015)

Genocide Against Rohingya Starts (2016)

Almost 1M Displaced

UN Concludes that Facebook Was Critical Component

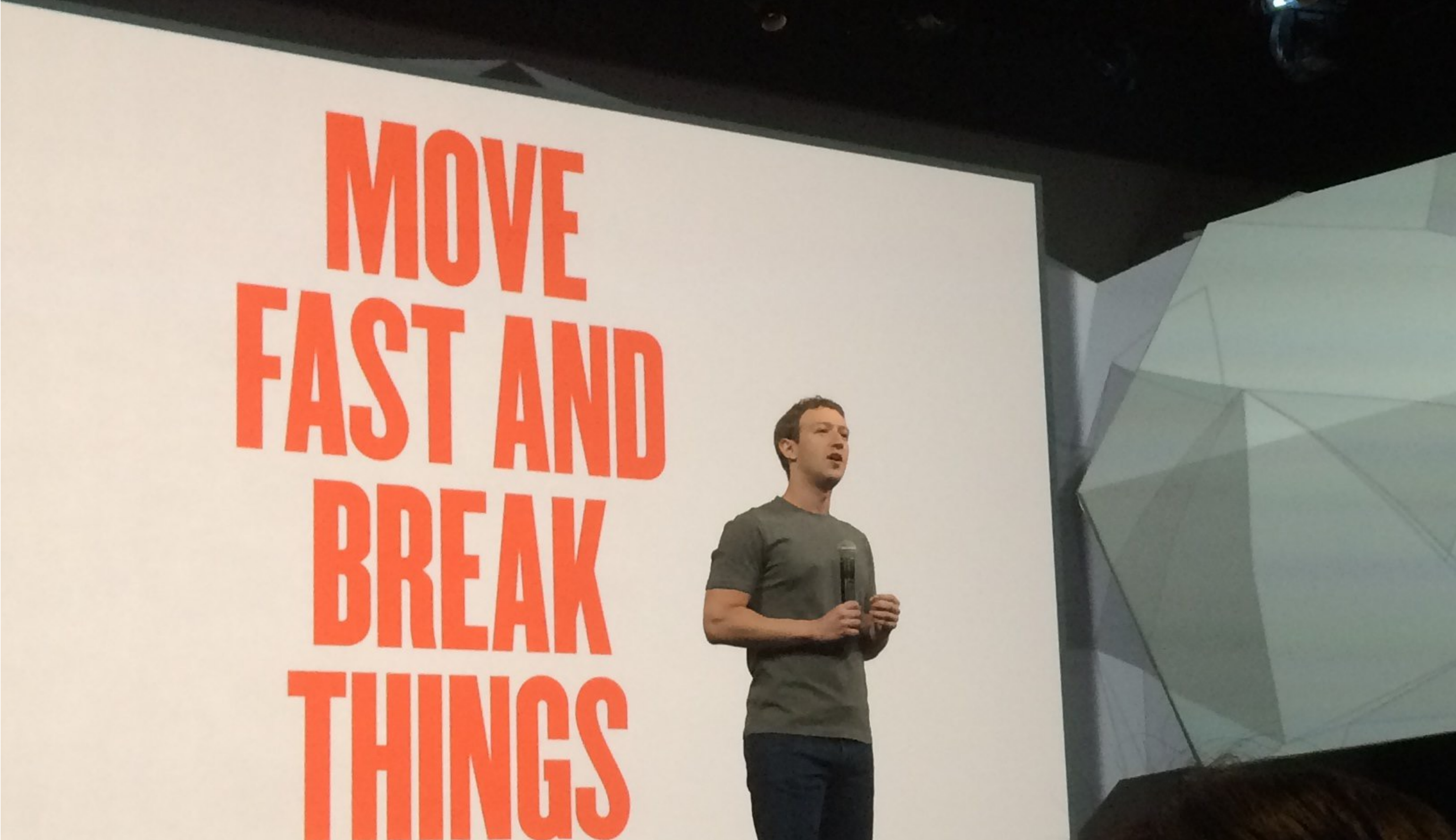


Human Rights Council
Thirty-ninth session
10–28 September 2018
Agenda item 4
Human rights situations that require the Council's attention

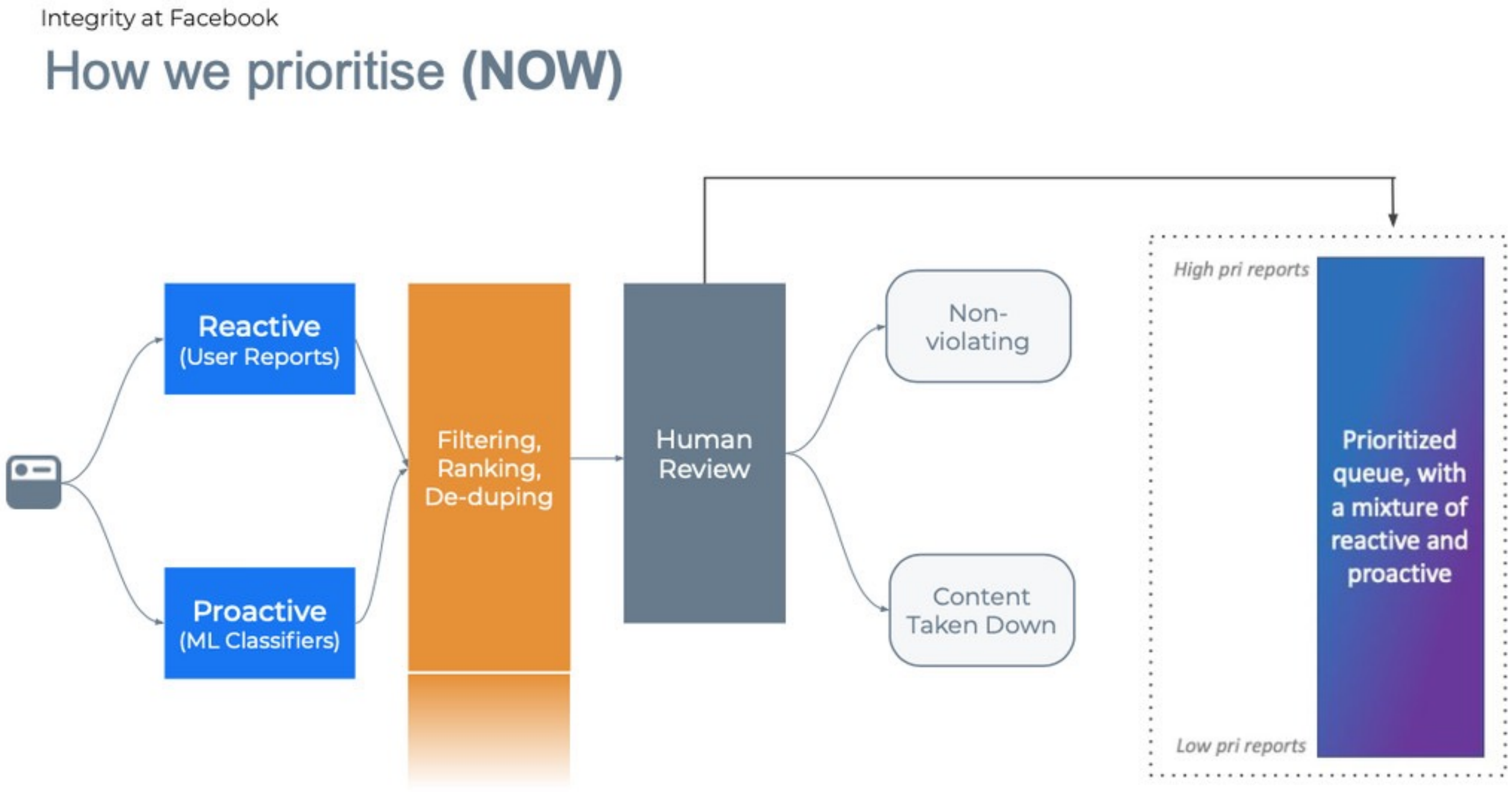
Report of the independent international fact-finding mission on Myanmar*

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

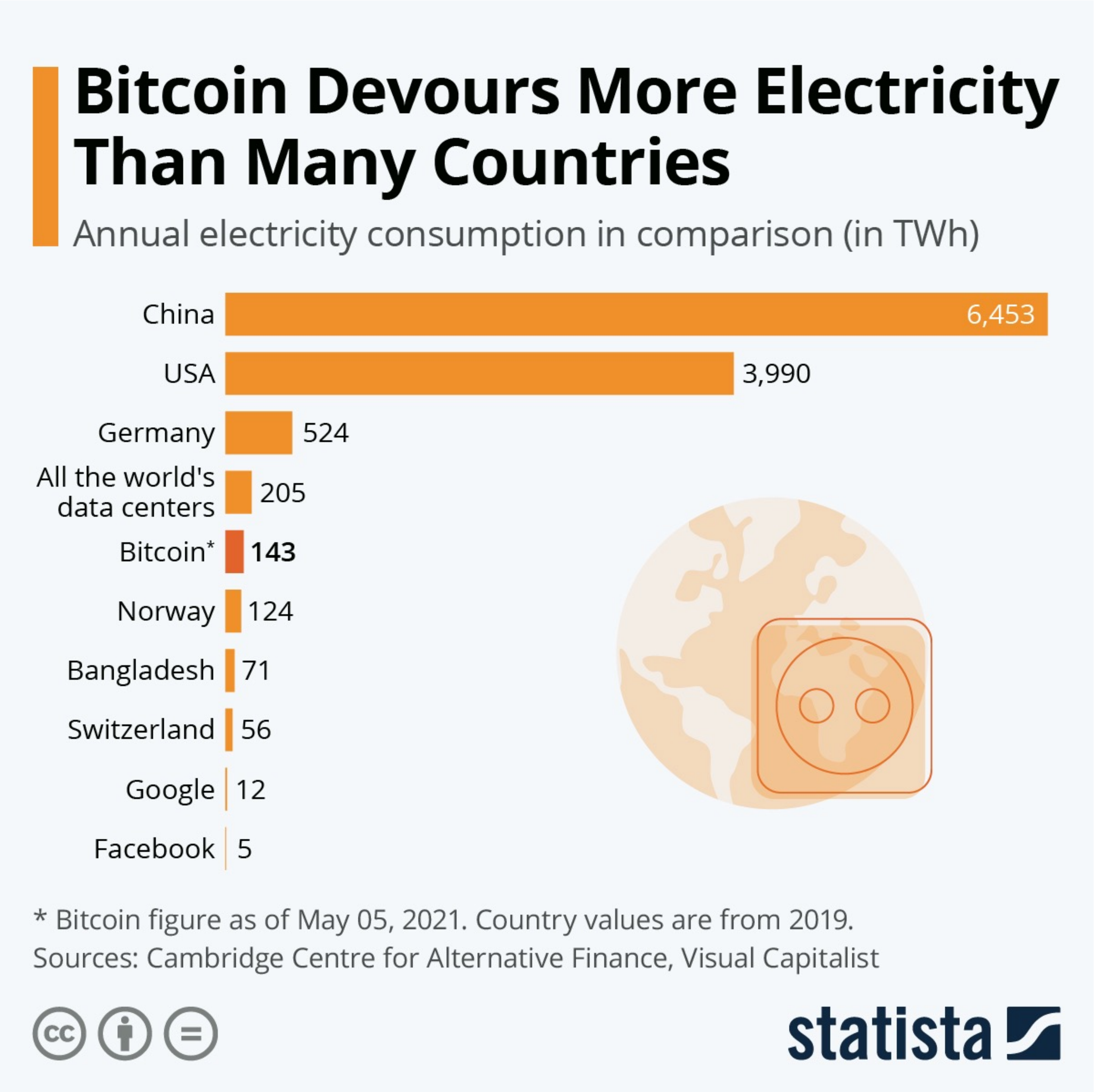
Aside: Facebook Says the Answer is Better ML



Integrity at Facebook

How we prioritise **(NOW)**

One Blind Spot I Want to Highlight



Bitcoin Devours More Electricity Than Many Countries

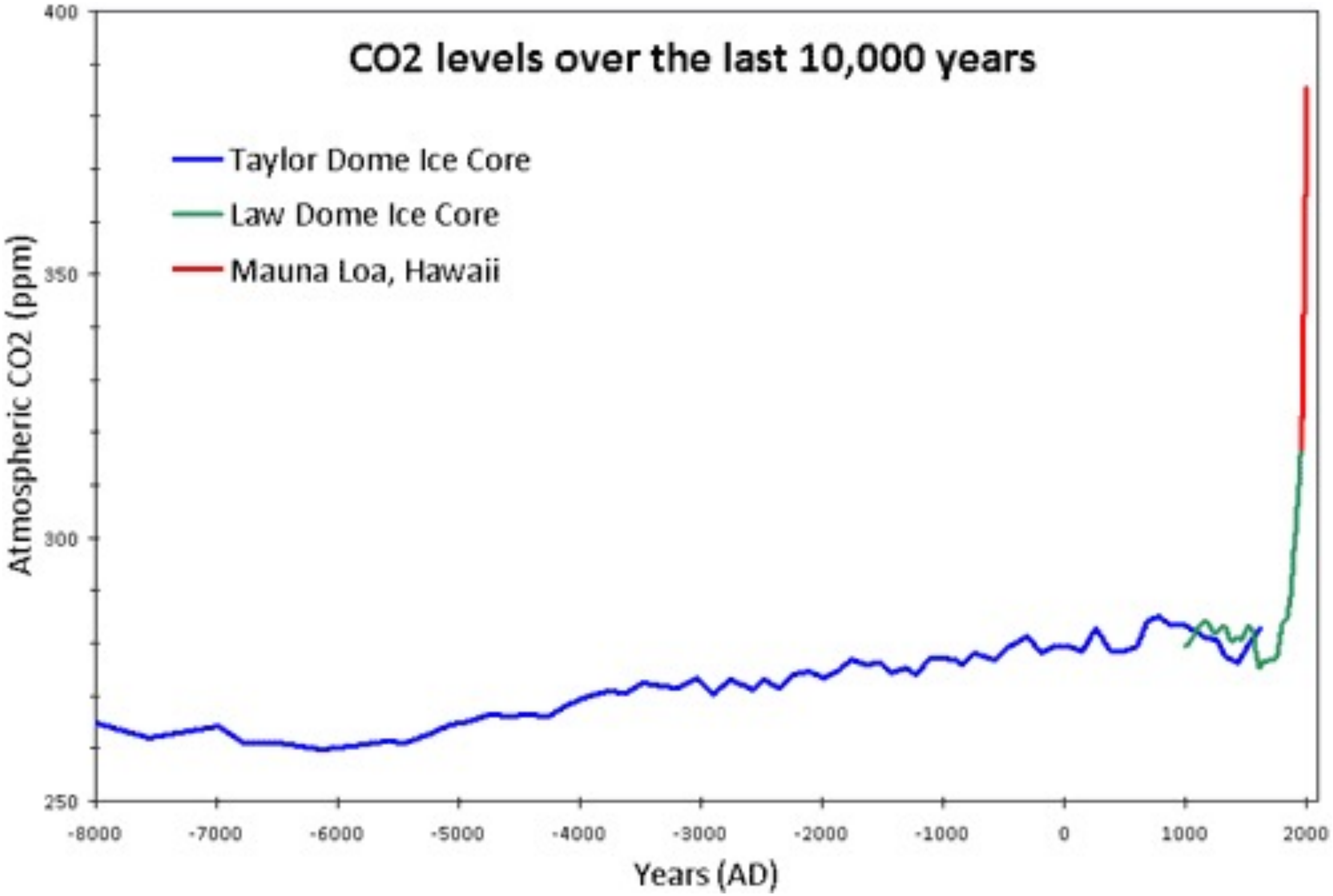
Annual electricity consumption in comparison (in TWh)

* Bitcoin figure as of May 05, 2021. Country values are from 2019.
Sources: Cambridge Centre for Alternative Finance, Visual Capitalist

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

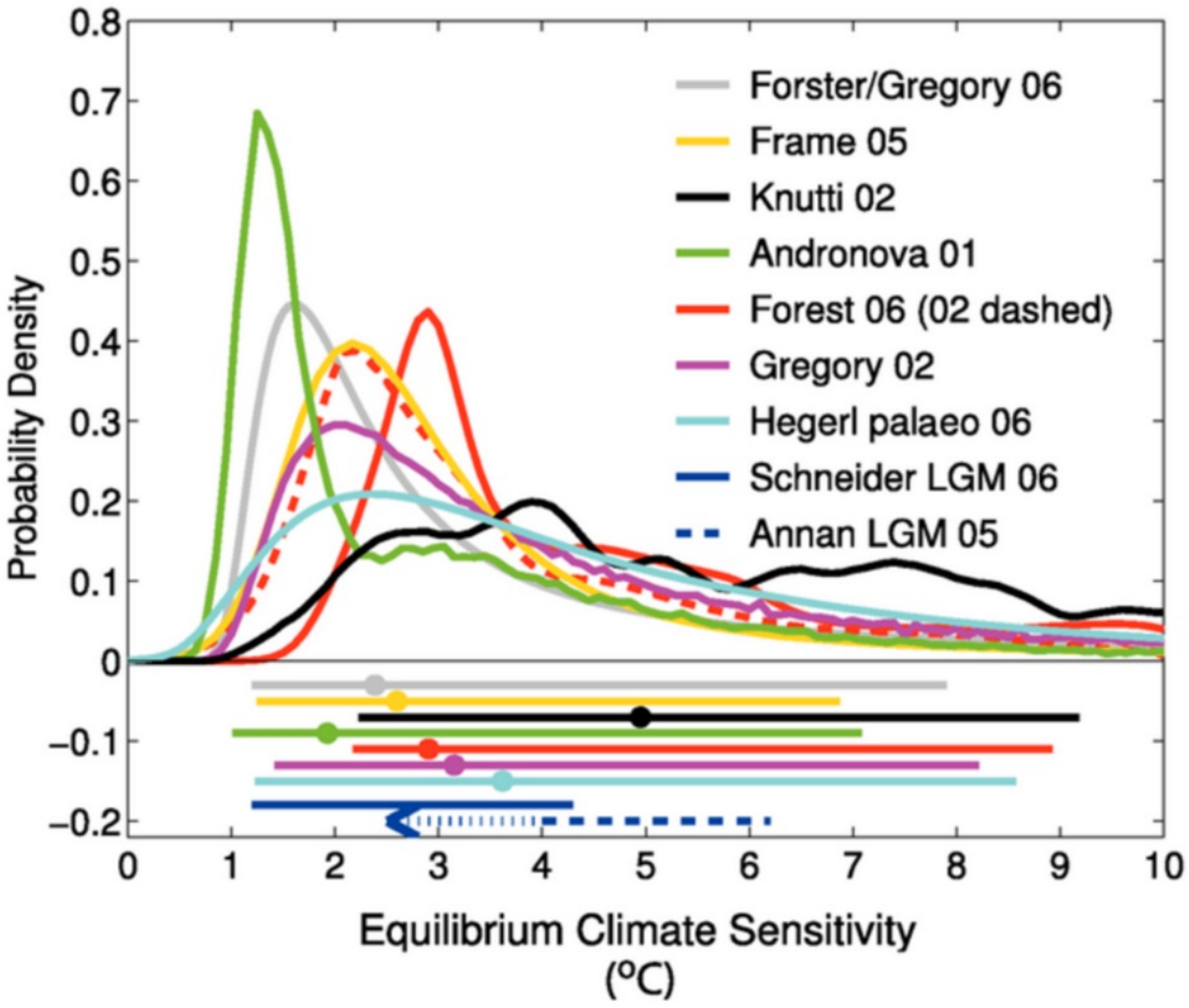
But climate change and bitcoin isn’t even part of ethics at Stanford CS (I will update this slide once that changes)

It isn’t too hard to see the trend



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

The Whole Story is Filled with Uncertanties



Many things are uncertain

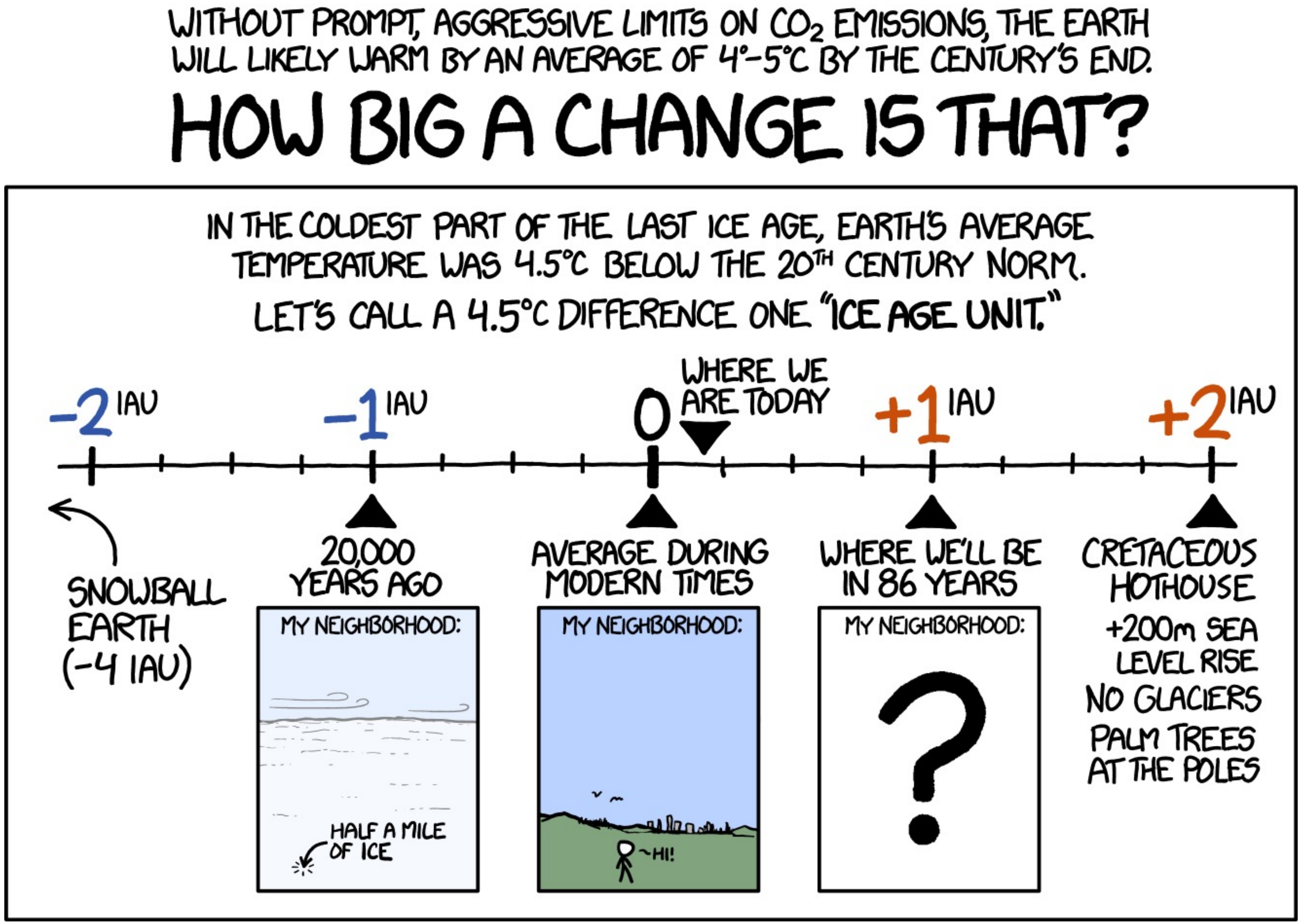
- Future Amount of CO2
- Climate Sensitivity
- Impact

But we can reason under uncertainty

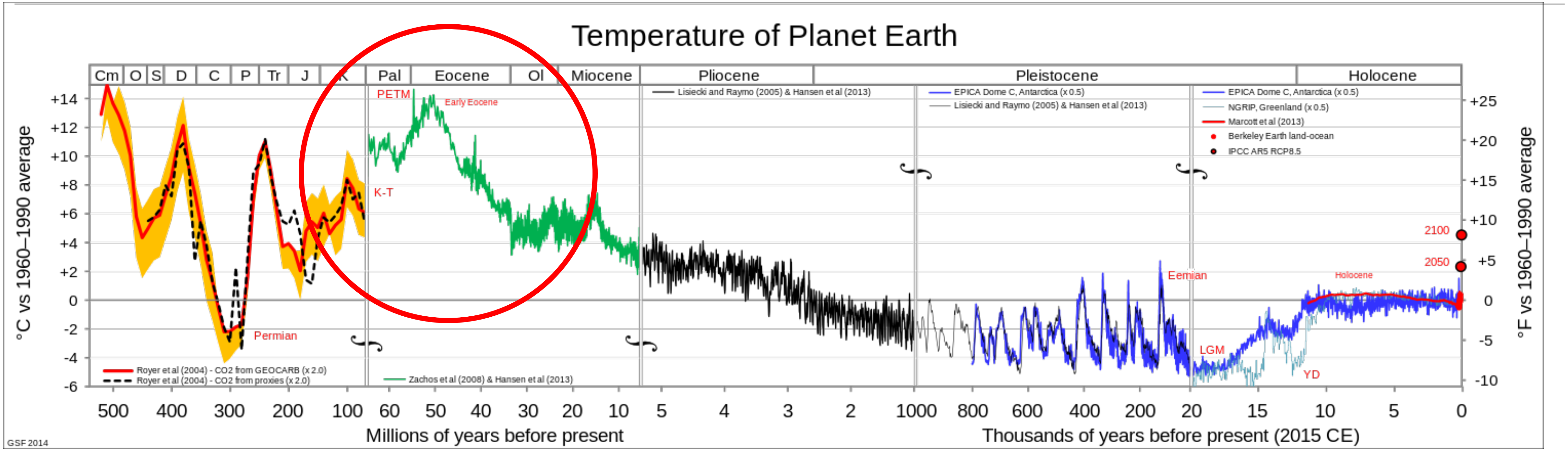
We know the physics

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

Easy to Know Impacts Will Be Harsh


WITHOUT PROMPT, AGGRESSIVE LIMITS ON CO₂ EMISSIONS, THE EARTH WILL LIKELY WARM BY AN AVERAGE OF 4°–5°C BY THE CENTURY'S END.
HOW BIG A CHANGE IS THAT?
IN THE COLDEST PART OF THE LAST ICE AGE, EARTH'S AVERAGE TEMPERATURE WAS 4.5°C BELOW THE 20TH CENTURY NORM.
LET'S CALL A 4.5°C DIFFERENCE ONE "ICE AGE UNIT."
−2 IAU
−1 IAU
0
WHERE WE ARE TODAY
+1 IAU
+2 IAU
SNOWBALL EARTH (−4 IAU)
20,000 YEARS AGO
MY NEIGHBORHOOD:
HALF A MILE OF ICE
AVERAGE DURING MODERN TIMES
MY NEIGHBORHOOD:
~HI!
WHERE WE'LL BE IN 86 YEARS
MY NEIGHBORHOOD:
?
CRETACEOUS HOTHOUSE
+200m SEA LEVEL RISE
NO GLACIERS
PALM TREES AT THE POLES

Paleoclimate Gives us a Clue

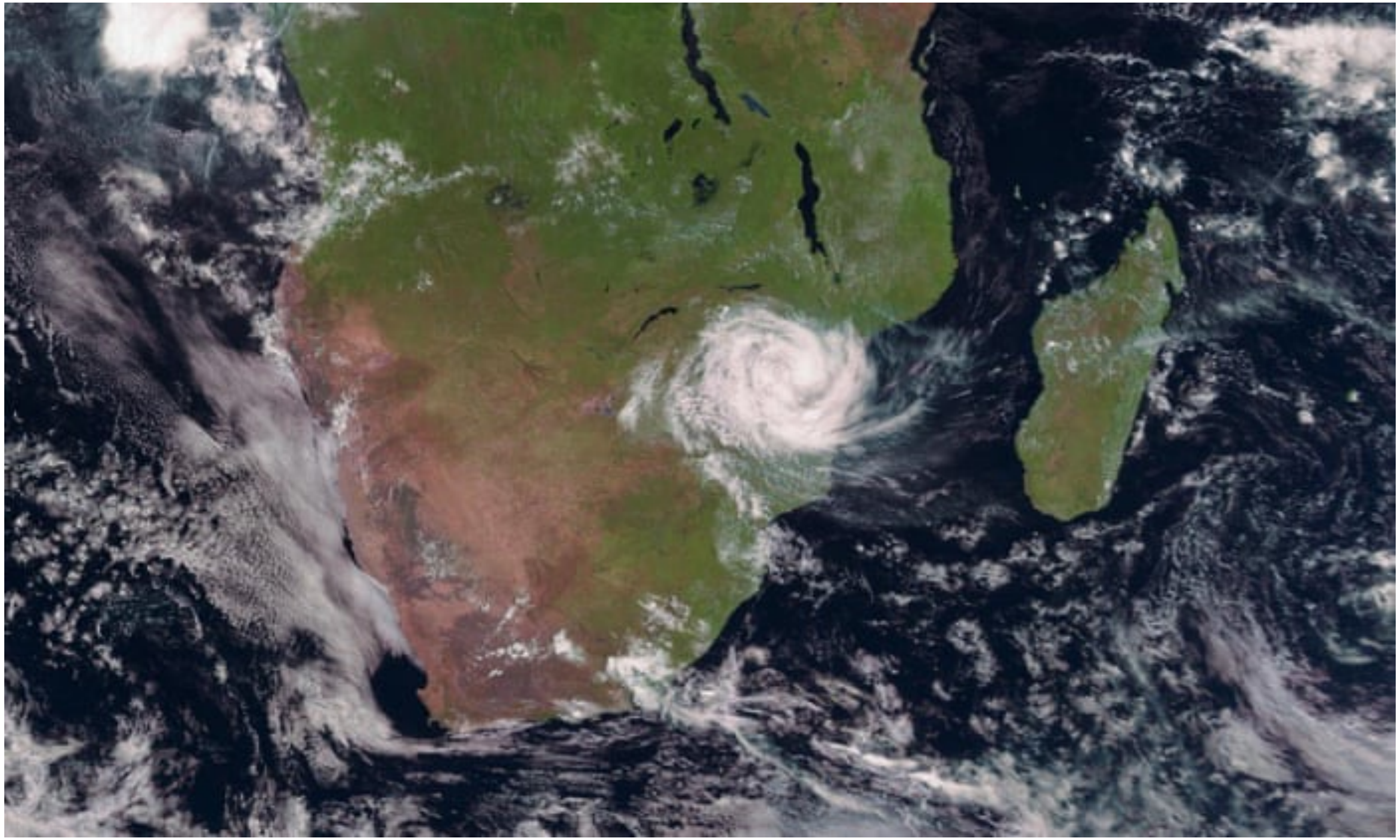


PETM Video: https://www.youtube.com/watch?v=IdLBoErAhz4

Impacts are Here

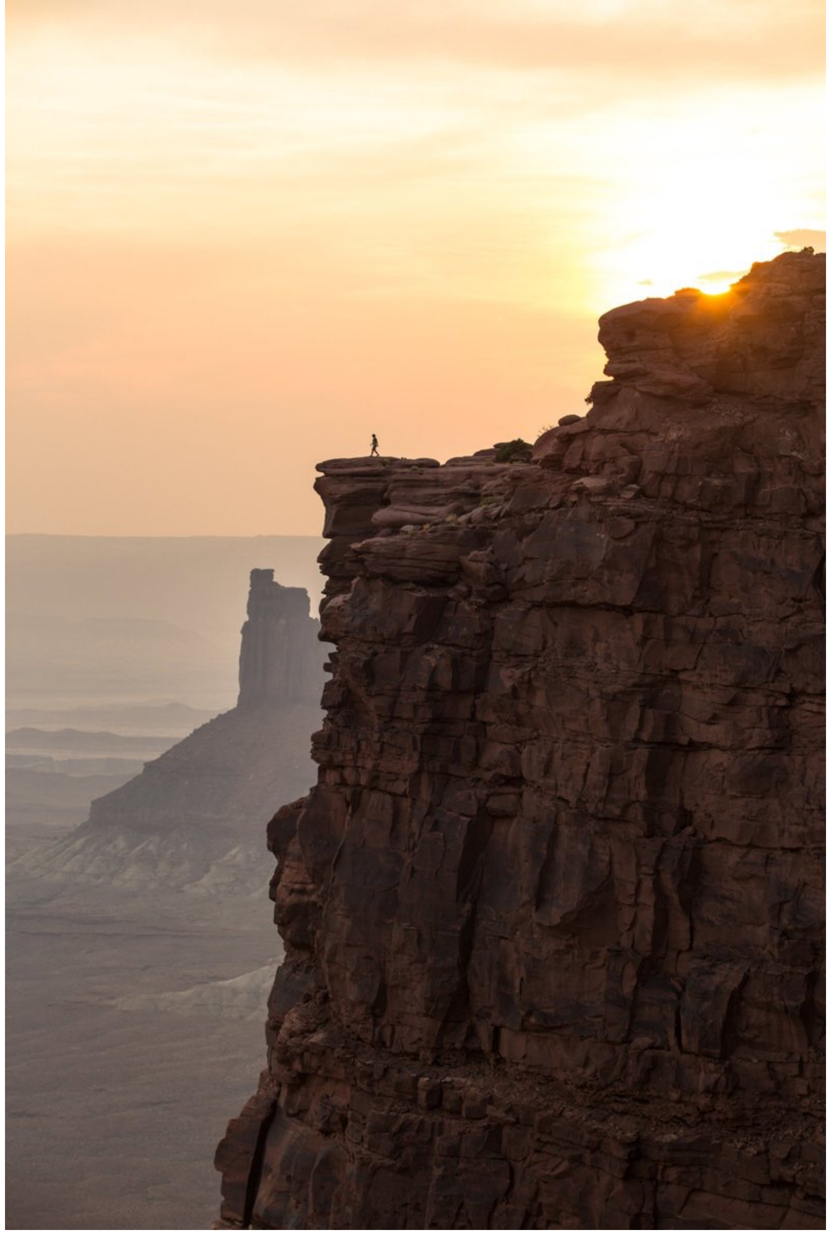
Cyclone Idai
Impacted over 3M people

But Most Impacts are Far in Time and Space



Cyclone Idai, impacted over 3M people

It is hard to feel like you can do anything…



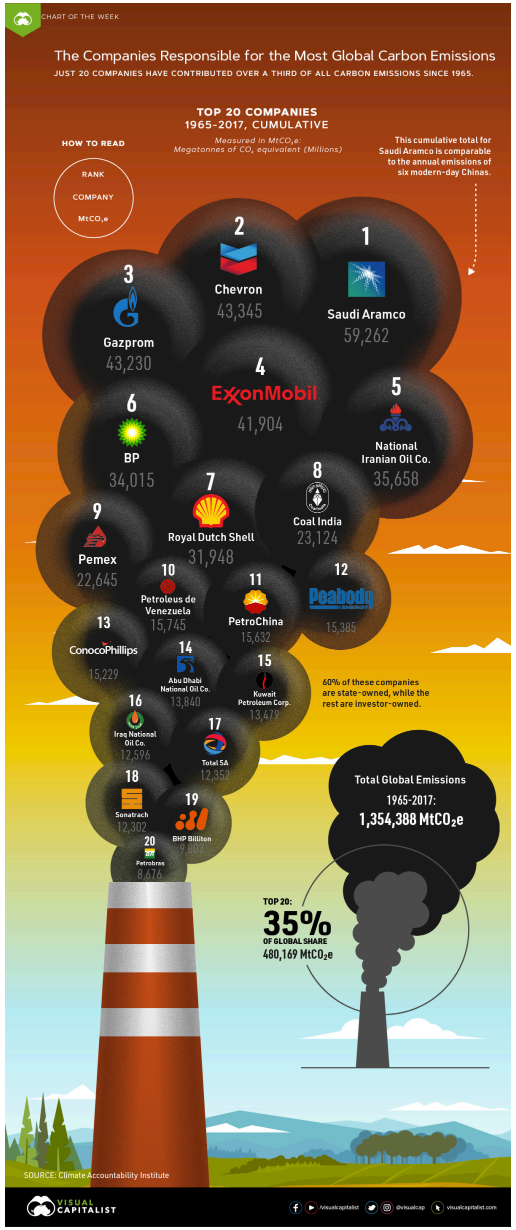


CHART OF THE WEEK
The Companies Responsible for the Most Global Carbon Emissions
JUST 20 COMPANIES HAVE CONTRIBUTED OVER A THIRD OF ALL CARBON EMISSIONS SINCE 1965.
TOP 20 COMPANIES
1965-2017, CUMULATIVE
Measured in MtCO₂e:
Megatonnes of CO₂ equivalent (Millions)
HOW TO READ
RANK
COMPANY
MtCO₂e
This cumulative total for Saudi Aramco is comparable to the annual emissions of six modern-day Chinas.
1 Saudi Aramco 59,262
2 Chevron 43,345
3 Gazprom 43,230
4 ExxonMobil 41,904
5 National Iranian Oil Co. 35,658
6 BP 34,015
7 Royal Dutch Shell 31,948
8 Coal India 23,124
9 Pemex 22,645
10 Petroleus de Venezuela 15,745
11 PetroChina 15,632
12 Peabody 15,385
13 ConocoPhillips 15,229
14 Abu Dhabi National Oil Co. 13,840
15 Kuwait Petroleum Corp. 13,479
16 Iraq National Oil Co. 12,596
17 Total SA 12,352
18 Sonatrach 12,302
19 BHP Billiton 9,802
20 Petrobras 8,676
60% of these companies are state-owned, while the rest are investor-owned.
Total Global Emissions 1965-2017: 1,354,388 MtCO₂e
TOP 20: 35% OF GLOBAL SHARE 480,169 MtCO₂e
VISUAL CAPITALIST

SOURCE: Climate Accountability Institute

“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”

Is this an ethical policy?

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

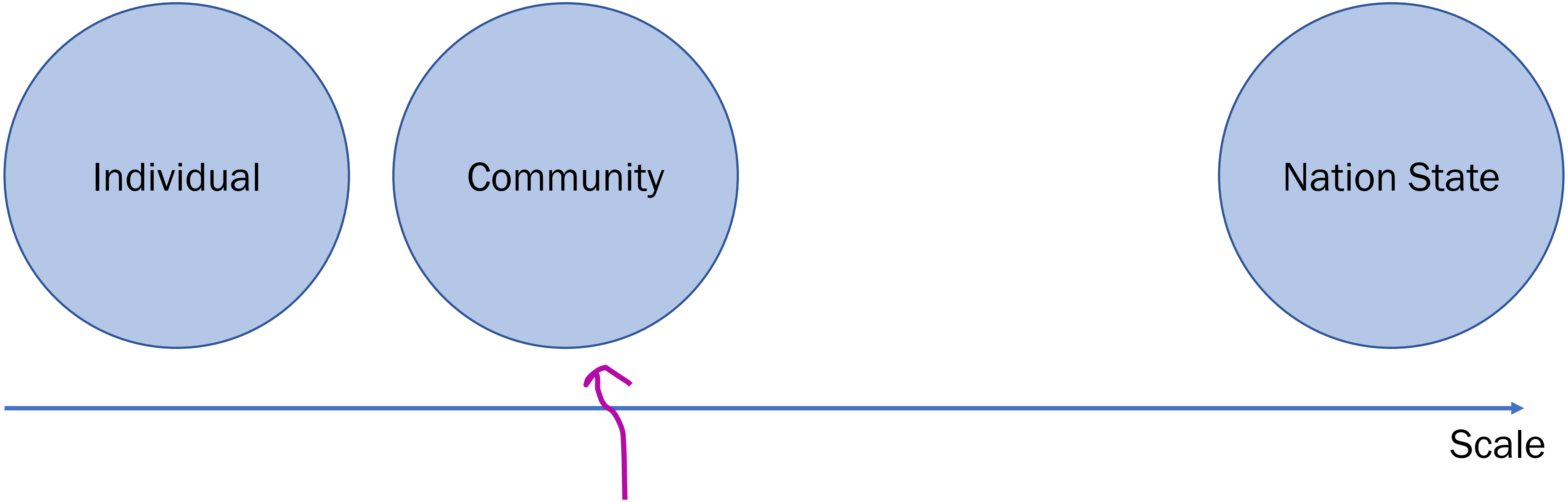
What Value System Can You Use?

Concept: aesthetics / mencious

Concept: awareness
was that they were not thoughtful

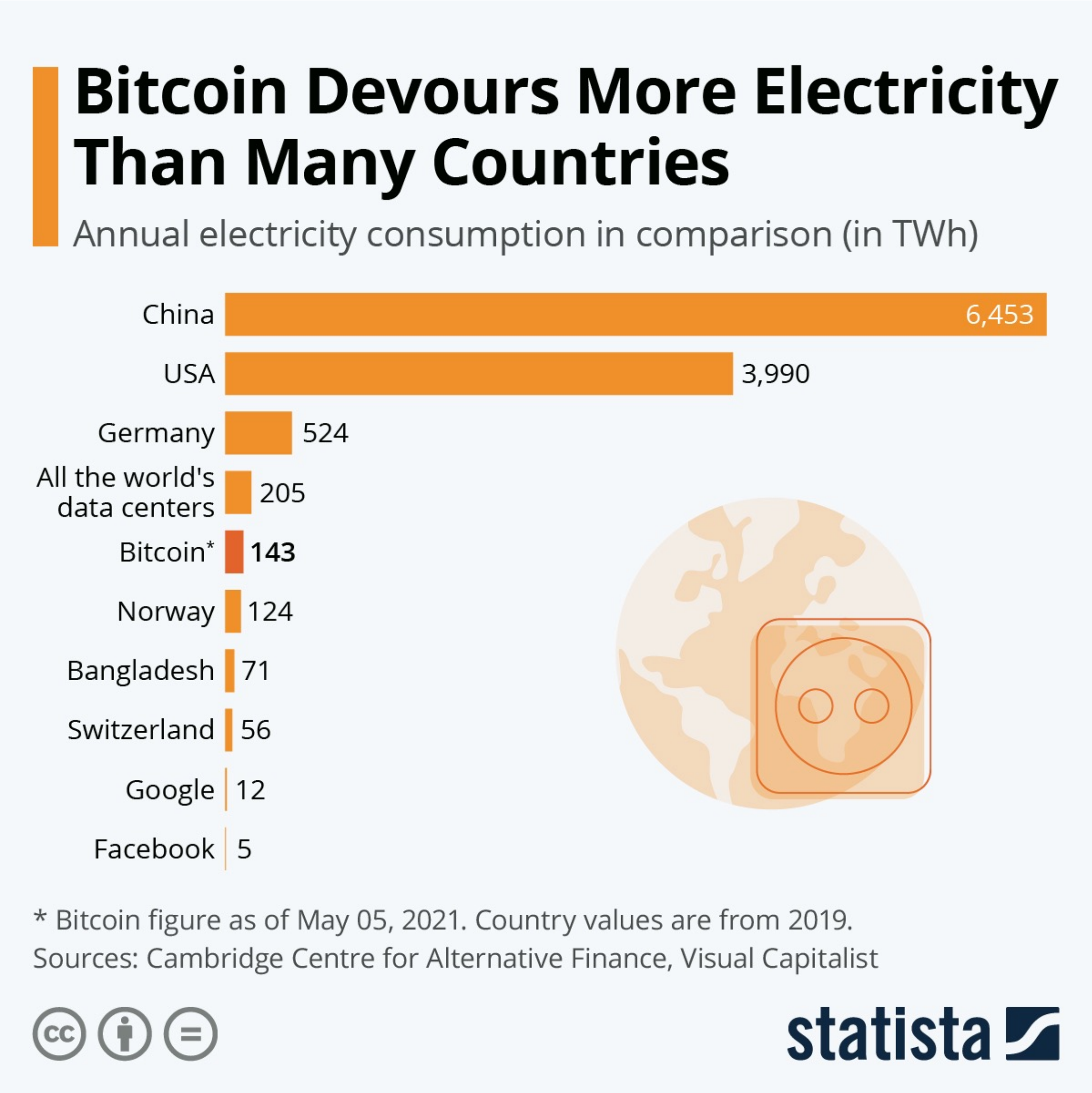
What can we do?

Push for some change



Is this our sweet spot?

Reduce CS "Pump" of Proof of Work



Bitcoin Devours More Electricity Than Many Countries

Annual electricity consumption in comparison (in TWh)

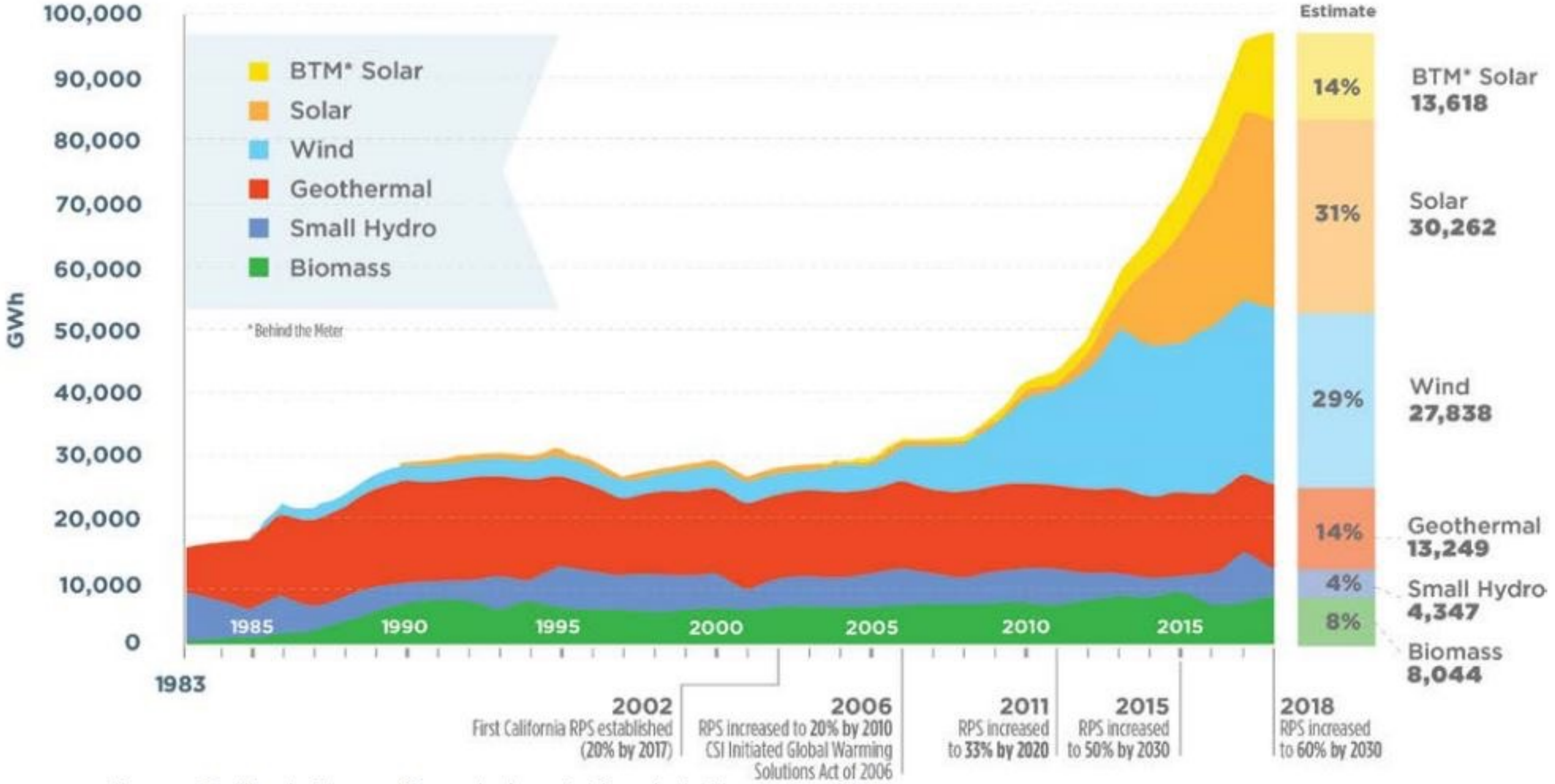
* Bitcoin figure as of May 05, 2021. Country values are from 2019.
Sources: Cambridge Centre for Alternative Finance, Visual Capitalist

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



2002
First California RPS established
(20% by 2017)

2006
RPS increased to 20% by 2010
CSI Initiated Global Warming
Solutions Act of 2006

2011
RPS increased
to 33% by 2020

2015
RPS increased
to 50% by 2030

2018
RPS increased
to 60% by 2030

Source: California Energy Commission, staff analysis November 2018

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

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.

Thank you!

Feel free to chat about this with Chris or with our Embedded Ethics instructor, Katie Creel
kcreel@stanford.edu