CS109:
Ethics & Machine Learning

ACM Code of Ethics and Professional Conduct

ACM Code of Ethics and Professional Conduct

Preamble

Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good. The ACM Code of Ethics and Professional Conduct ("the Code") expresses the conscience of the profession.

The Code is designed to inspire and guide the ethical conduct of all computing professionals, including current and aspiring practitioners, instructors, students, influencers, and anyone who uses computing technology in an impactful way. Additionally, the Code serves as a basis for remediation when violations occur. The Code includes principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration. Each principle is supplemented by guidelines, which provide explanations to assist computing professionals in understanding and applying the principle.

Section 1 outlines fundamental ethical principles that form the basis for the remainder of the Code. Section 2 addresses additional, more specific considerations of professional responsibility. Section 3 guides individuals who have a leadership role, whether in the workplace or in a volunteer professional capacity. Commitment to ethical conduct is required of every ACM member, and principles involving compliance with the Code are given in Section 4.

The Code as a whole is concerned with how fundamental ethical principles apply to a computing professional's conduct. The Code is not an algorithm for solving ethical problems; rather it serves as a

On This Page

1.2 Avoid harm.

In this document, "harm" means negative consequences, especially when those consequences are significant and unjust. Examples of harm include unjustified physical or mental injury, unjustified destruction or disclosure of information, and unjustified damage to property, reputation, and the environment. This list is not exhaustive.

Well-intended actions, including those that accomplish assigned duties, may lead to harm. When that harm is unintended, those responsible are obliged to undo or mitigate the harm as much as possible. Avoiding harm begins with careful consideration of potential impacts on all those affected by decisions. When harm is an intentional part of the system, those responsible are obligated to ensure that the harm is ethically justified. In either case, ensure that all harm is minimized.

To minimize the possibility of indirectly or unintentionally harming others, computing professionals should follow generally accepted best practices unless there is a compelling ethical reason to do otherwise. Additionally, the consequences of data aggregation and emergent properties of systems should be carefully analyzed. Those involved with pervasive or infrastructure systems should also consider Principle 3.7.

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.

1.2 Avoid harm.

In this document, "harm" means negative consequences, especially when those consequences are significant and unjust. Examples of harm include unjustified physical or mental injury, unjustified destruction or disclosure of information, and unjustified damage to property, reputation, and the environment. This list is not exhaustive.

Well-intended actions, including those that accomplish assigned duties, may lead to harm. When that harm is unintended, those responsible are obliged to undo or mitigate the harm as much as possible. Avoiding harm begins with careful consideration of potential impacts on all those affected by decisions. When harm is an intentional part of the system, those responsible are obligated to ensure that the harm is ethically justified. In either case, ensure that all harm is minimized.

To minimize the possibility of indirectly or unintentionally harming others, computing professionals should follow generally accepted best practices unless there is a compelling ethical reason to do otherwise. Additionally, the consequences of data aggregation and emergent properties of systems should be carefully analyzed. Those involved with pervasive or infrastructure systems should also consider Principle 3.7.

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.

1.4 Be fair and take action not to discriminate.

The values of equality, tolerance, respect for others, and justice govern this principle. Fairness requires that even careful decision processes provide some avenue for redress of grievances.

Computing professionals should foster fair participation of all people, including those of underrepresented groups. Prejudicial discrimination on the basis of age, color, disability, ethnicity, family status, gender identity, labor union membership, military status, nationality, race, religion or belief, sex, sexual orientation, or any other inappropriate factor is an explicit violation of the Code. Harassment, including sexual harassment, bullying, and other abuses of power and authority, is a form of discrimination that, amongst other harms, limits fair access to the virtual and physical spaces where such harassment takes place.

The use of information and technology may cause new, or enhance existing, inequities. Technologies and practices should be as inclusive and accessible as possible and computing professionals should take action to avoid creating systems or technologies that disenfranchise or oppress people. Failure to design for inclusiveness and accessibility may constitute unfair discrimination.

1.4 Be fair and take action not to discriminate.

The values of equality, tolerance, respect for others, and justice govern this principle. Fairness requires that even careful decision processes provide some avenue for redress of grievances.

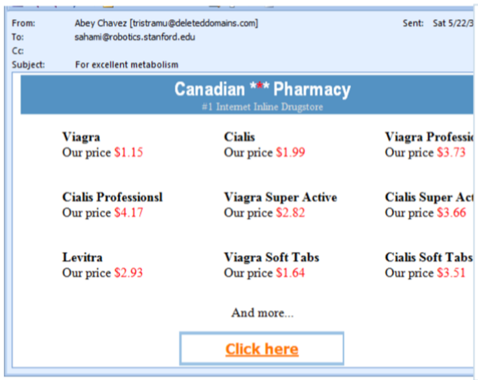
Computing professionals should foster fair participation of all people, including those of underrepresented groups. Prejudicial discrimination on the basis of age, color, disability, ethnicity, family status, gender identity, labor union membership, military status, nationality, race, religion or belief, sex, sexual orientation, or any other inappropriate factor is an explicit violation of the Code. Harassment, including sexual harassment, bullying, and other abuses of power and authority, is a form of discrimination that, amongst other harms, limits fair access to the virtual and physical spaces where such harassment takes place.

The use of information and technology may cause new, or enhance existing, inequities. Technologies and practices should be as inclusive and accessible as possible and computing professionals should take action to avoid creating systems or technologies that disenfranchise or oppress people. Failure to design for inclusiveness and accessibility may constitute unfair discrimination.

(Relatively) Easy Cases: Spam Detection & OCR

What is Bayes Doing in my Mail Server

This is spam:



Let's get Bayesian on your spam:

```
Content analysis details:   (49.5 hits, 7.0 required)
 0.9 RCVD_IN_PBL            RBL: Received via a relay in Spamhaus PBL
                            [93.40.189.29 listed in zen.spamhaus.org]
 1.5 URIBL_WS_SURBL         Contains an URL listed in the WS SURBL blocklist
                            [URIs: recragas.cn]
 5.0 URIBL_JP_SURBL         Contains an URL listed in the JP SURBL blocklist
                            [URIs: recragas.cn]
 5.0 URIBL_OB_SURBL         Contains an URL listed in the OB SURBL blocklist
                            [URIs: recragas.cn]
 5.0 URIBL_SC_SURBL         Contains an URL listed in the SC SURBL blocklist
                            [URIs: recragas.cn]
 2.0 URIBL_BLACK            Contains an URL listed in the URIBL blacklist
                            [URIs: recragas.cn]
 8.0 BAYES_99               BODY: Bayesian spam probability is 99 to 100%
                            [score: 1.0000]
```

A Bayesian Approach to Filtering Junk E-Mail

Mehran Sahami* Susan Dumais† David Heckerman† Eric Horvitz†

*Gates Building 1A
Computer Science Department
Stanford University
Stanford, CA 94305-9010
sahami@cs.stanford.edu

†Microsoft Research
Redmond, WA 98052-6399
{sdumais, heckerma, horvitz}@microsoft.com

Abstract

In addressing the growing problem of junk E-mail on the Internet, we examine methods for the automated

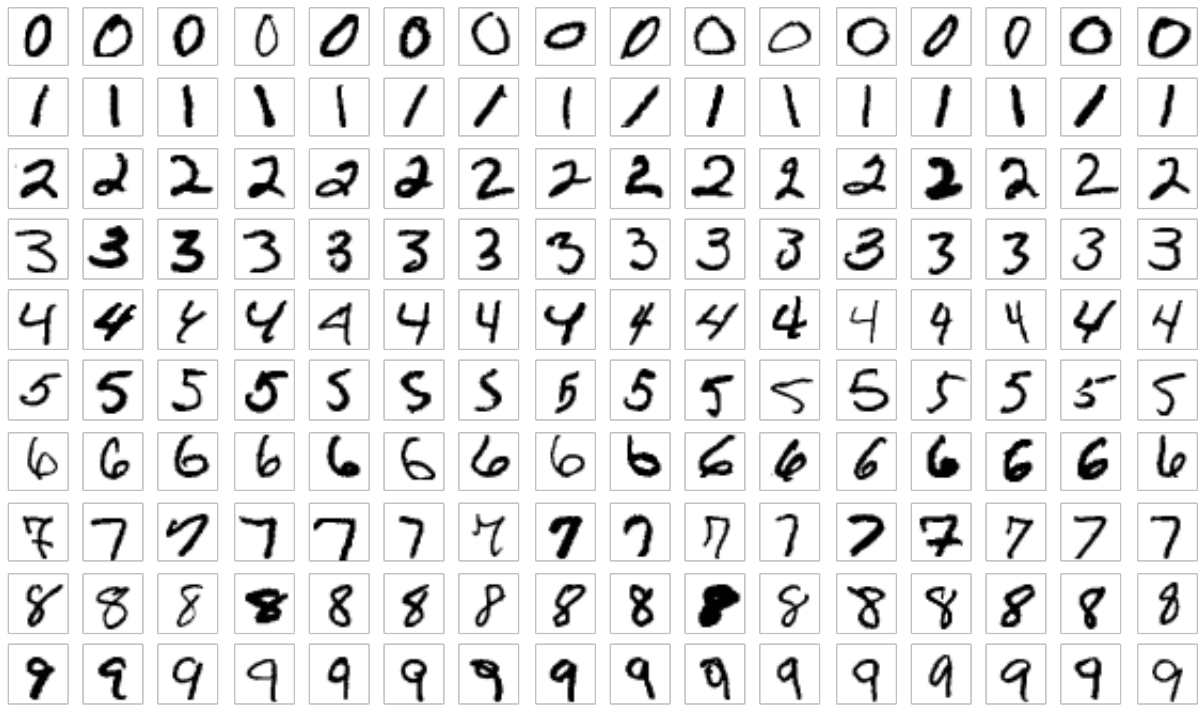
contain offensive material (such as graphic pornography), there is often a higher cost to users of actually viewing this mail than simply the time to sort out the junk. Lastly, junk mail not only wastes user time, but

Digit recognition example

Input image	Input feature vector	Output label
	$\boldsymbol{x}^{(i)} = [0,0,0,0, \dots, 1,0,0,1, \dots, 0,0,1,0]$	$y^{(i)} = 0$
	$\boldsymbol{x}^{(i)} = [0,0,1,1, \dots, 0,1,1,0, \dots, 0,1,0,0]$	$y^{(i)} = 1$

We make feature vectors from (digitized) pictures of numbers.

MNIST Database



USPS Mail Sorting using Optical Character Recognition (OCR)

43% of the world’s mail

161.4 million domestic addresses

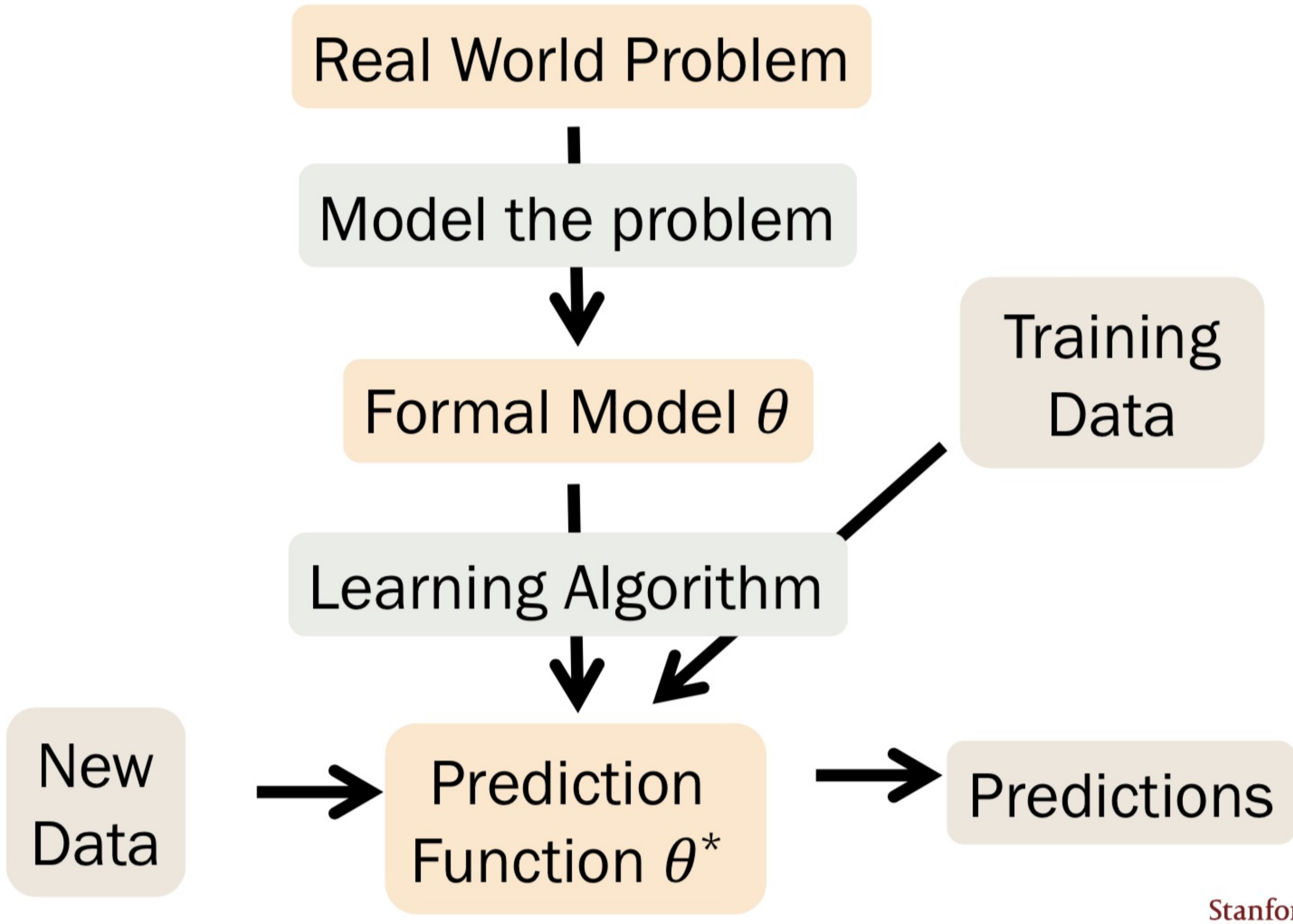
Mail Sorting & OCR

Ah, one of the relatively uncomplicated cases ...

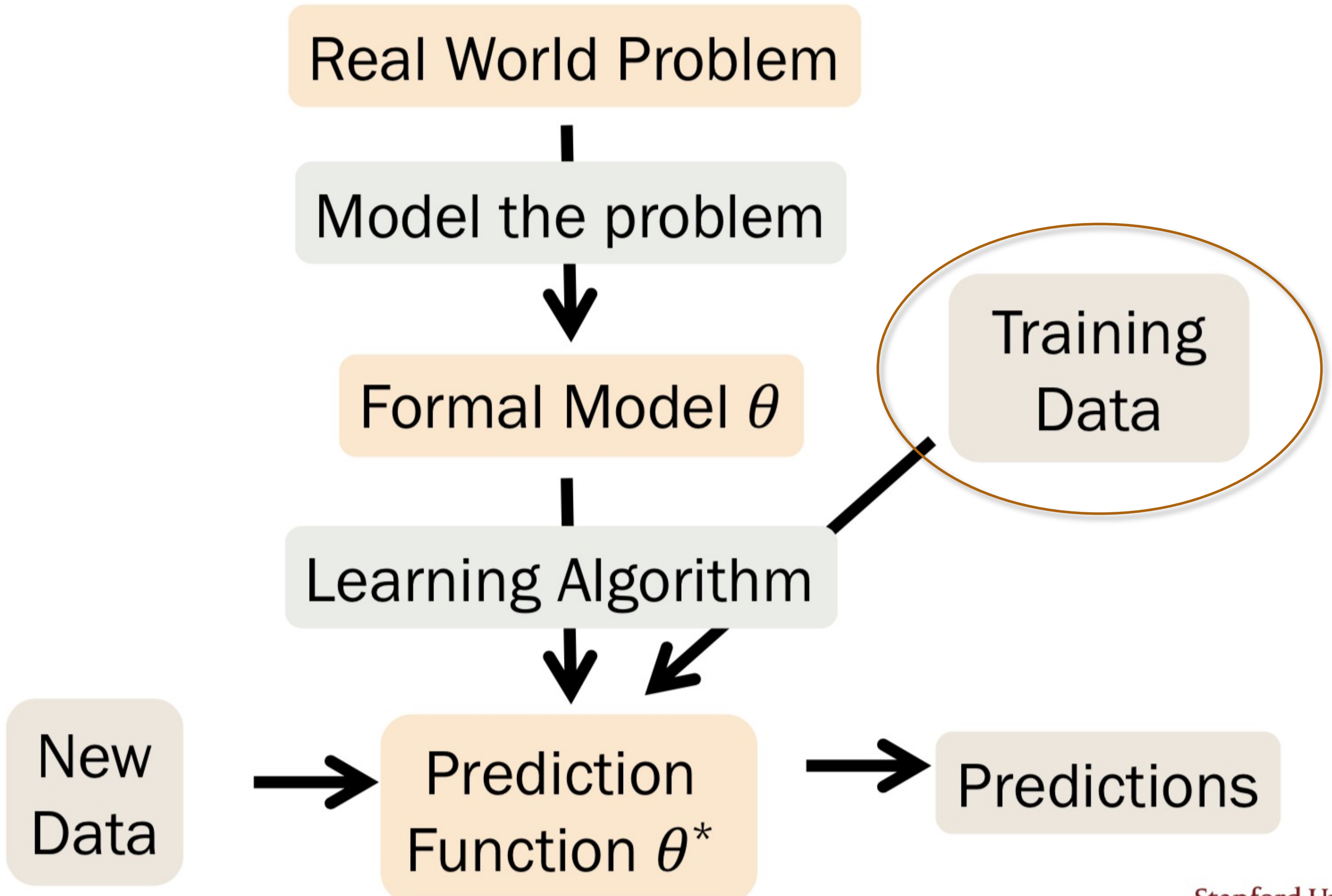
- Training Database (MNIST)
- Reference Standard/Benchmark
- Deployed everywhere

Responsible Machine Learning using Data about People

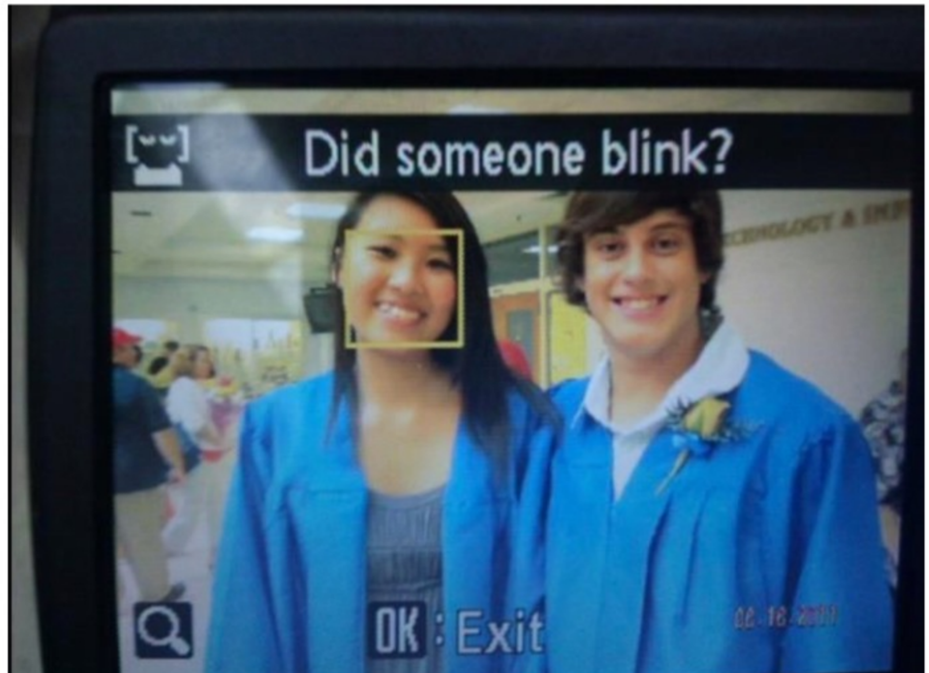
Machine Learning



Machine Learning

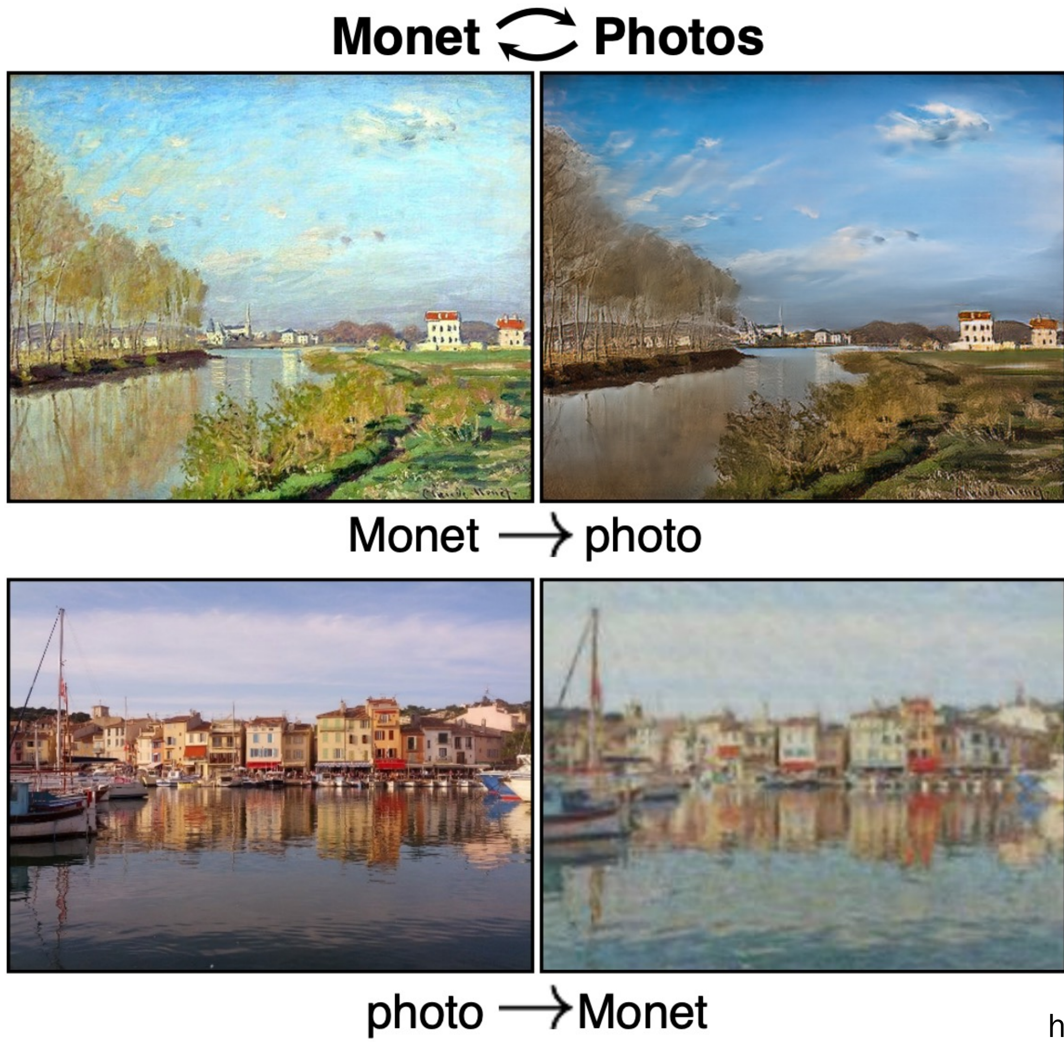


Ethics and Datasets?



Theme #1:
Building Responsible Datasets

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



Zhu et al 2017
https://arxiv.org/abs/1703.10593

“Van Gogh” is biased towards a yellow/green/blue palette …

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

.. But real van Gogh painted red poppies.

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

Skin lightening & feature whitening in generative art

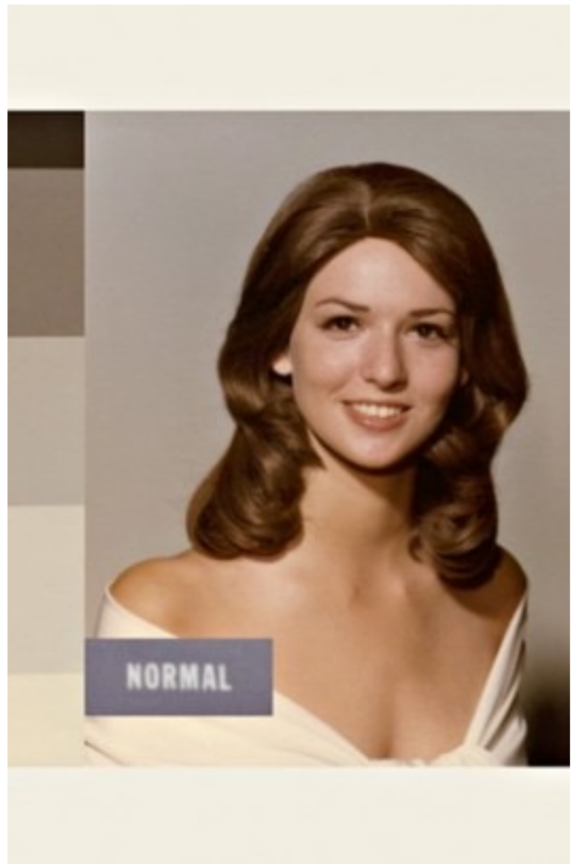
Images generated by AI Portrait Ars (now offline)

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

REVIEW

~~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

noothhound 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

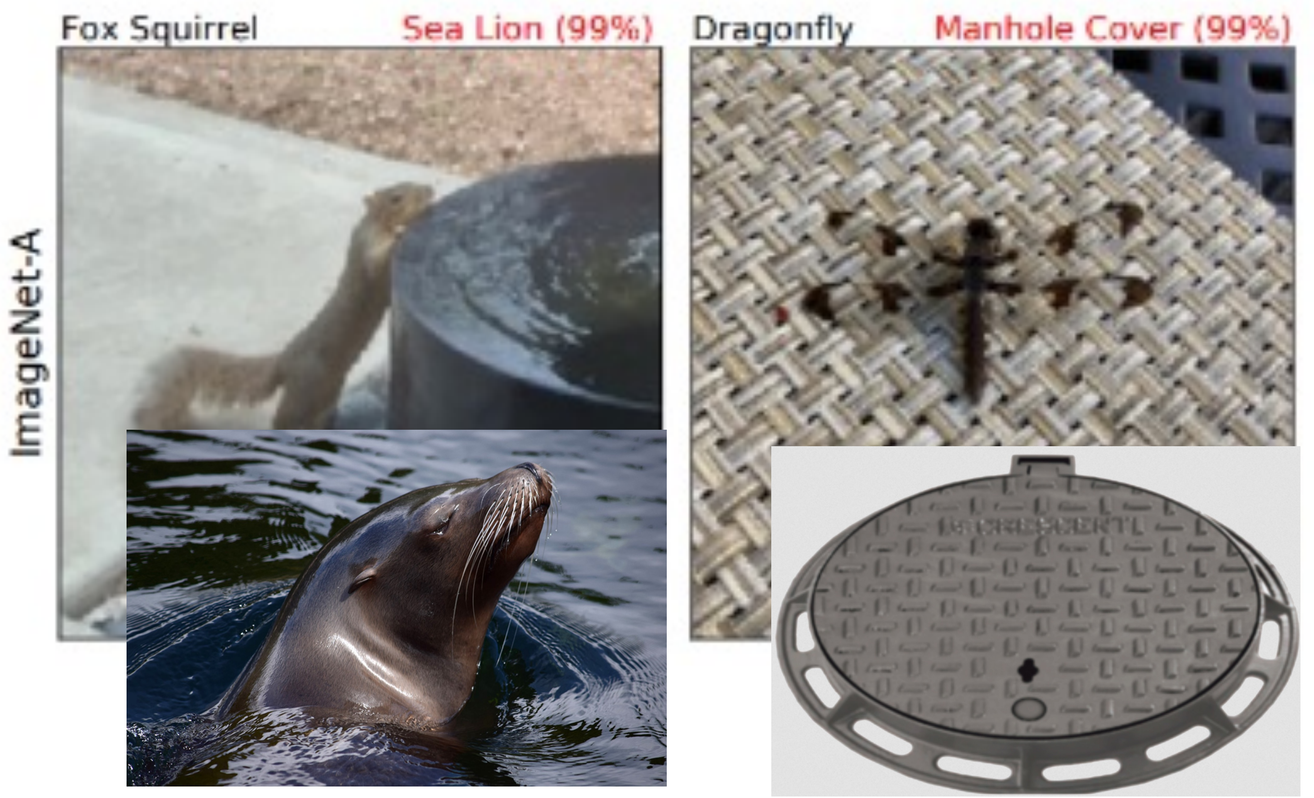
Biases in ImageNet

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. 2020

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

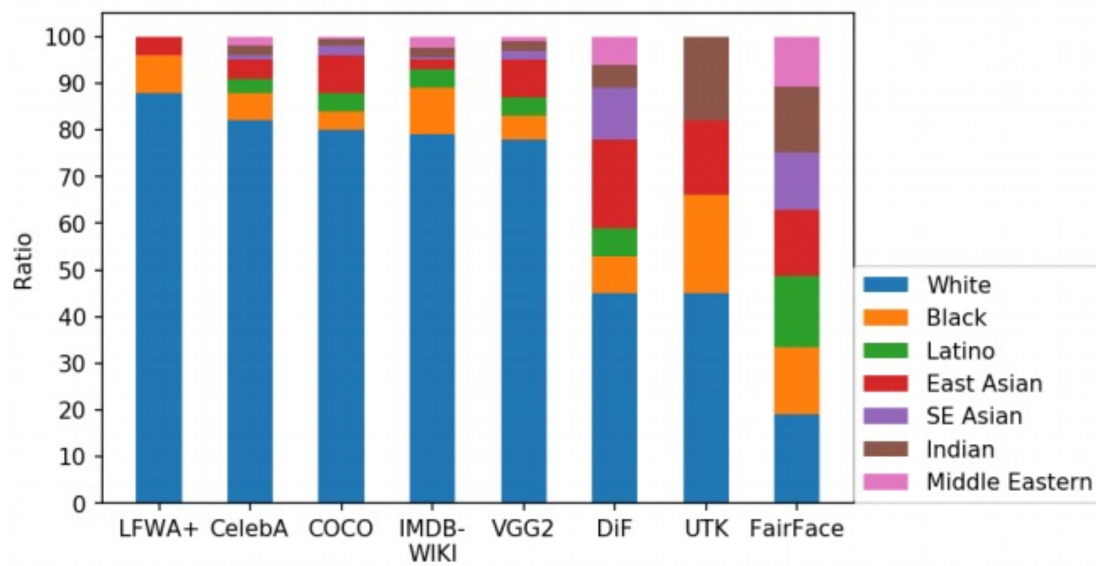


Figure 2: Racial compositions in face datasets.

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.

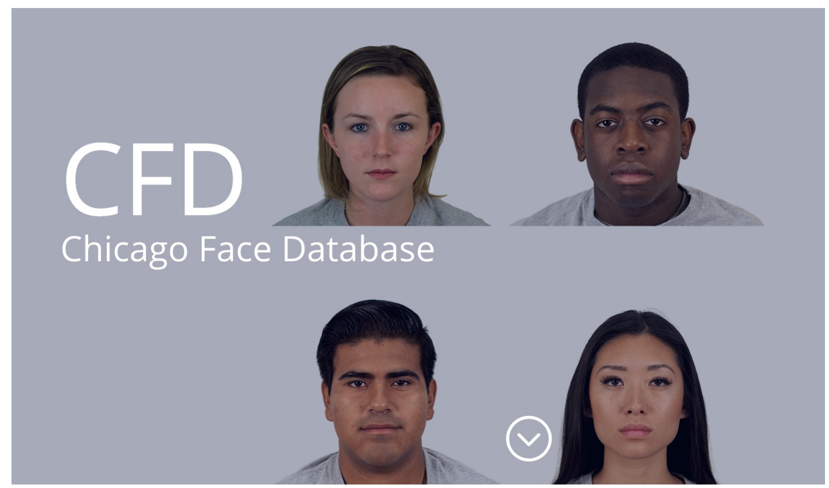
- Majority of images in some databases of faces are of white faces.

- 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



“Quality of Service” Harm

“Quality-of-service harms can occur when a system does not work as well for one person as it does for another, even if no opportunities, resources, or information are extended or withheld.” (Crawford)

Examples:

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

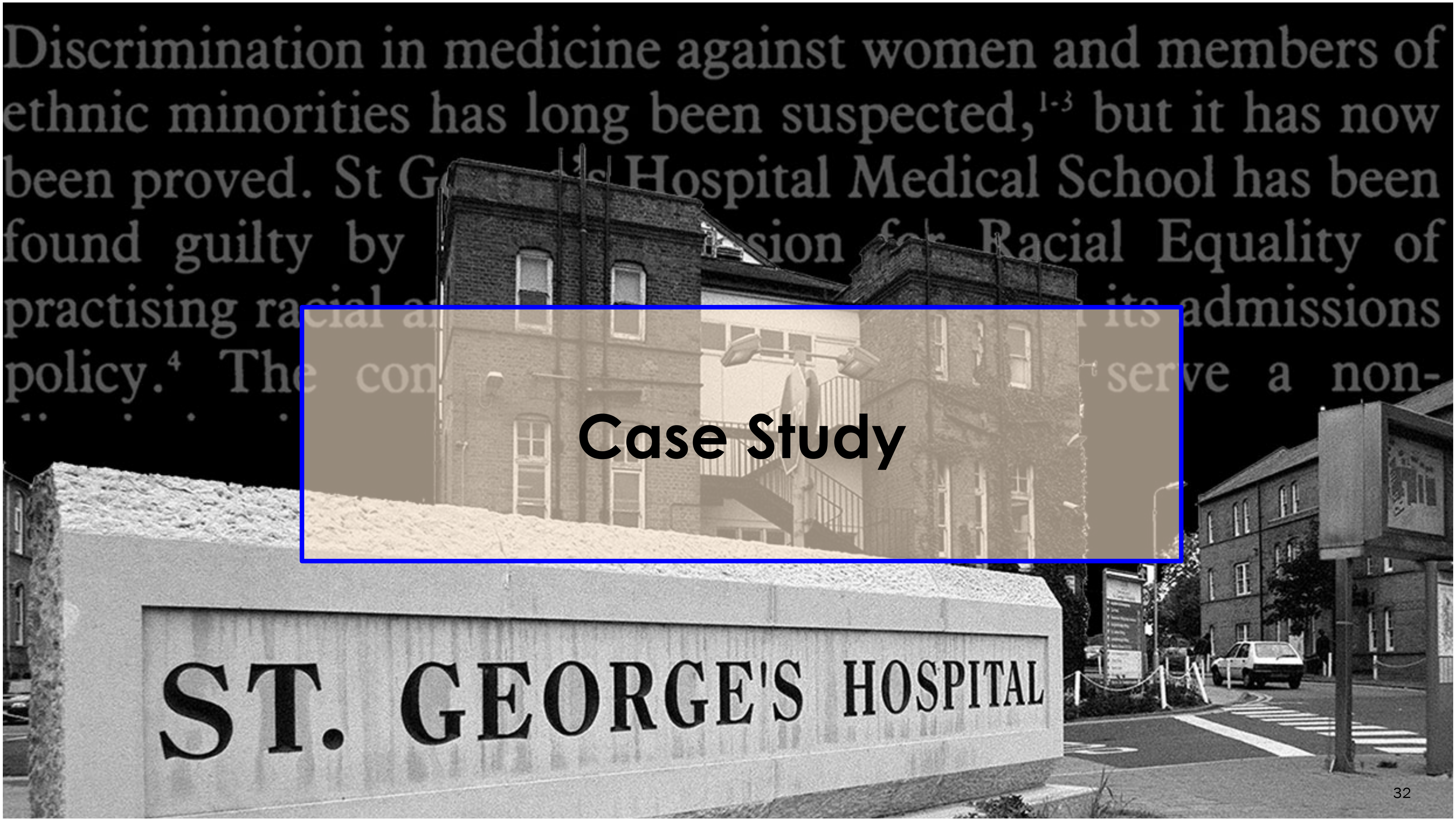
Allocation Harms

Allocation harms can occur when AI systems extend or withhold opportunities, resources, or information

What is a just distribution of outcomes for:

- Hiring
- Lending
- School admissions

Case Study



Discrimination in medicine against women and members of
ethnic minorities has long been suspected,1-3 but it has now
been proved.
Hospital Medical School has been
found guilty by
Racial Equality of
practising
admissions
policy.4
a non-
ST. GEORGE'S HOSPITAL

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)

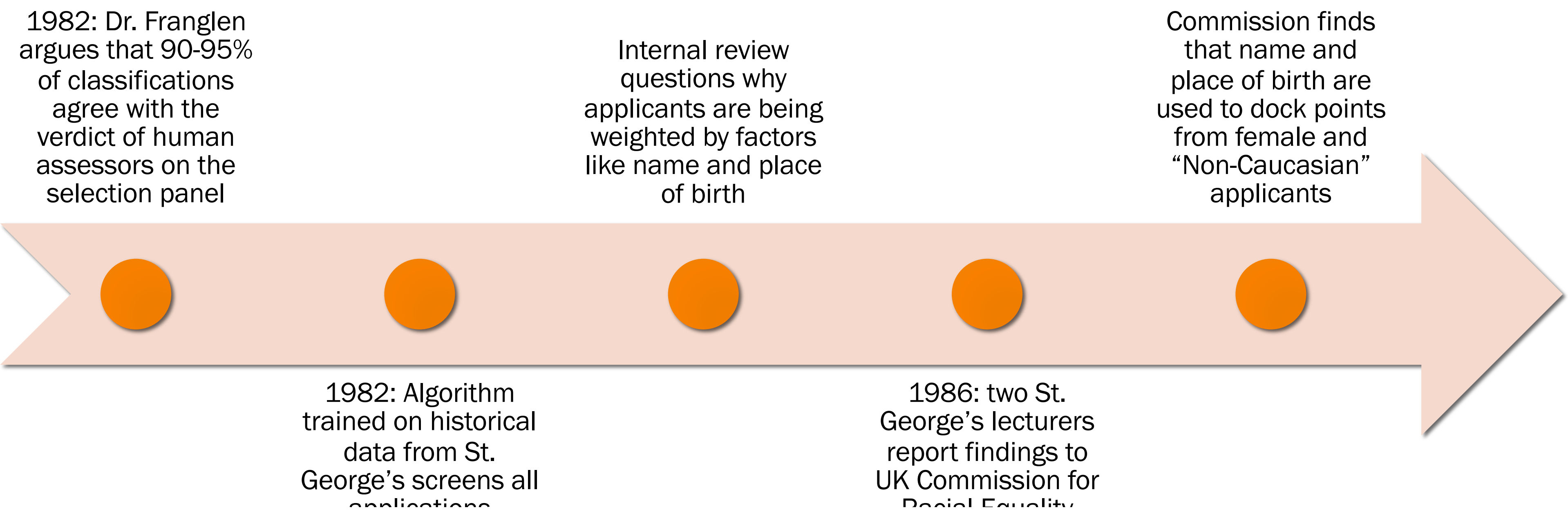
Offer spots to approx. 425 (so 70% of interviewees accepted)

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



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.

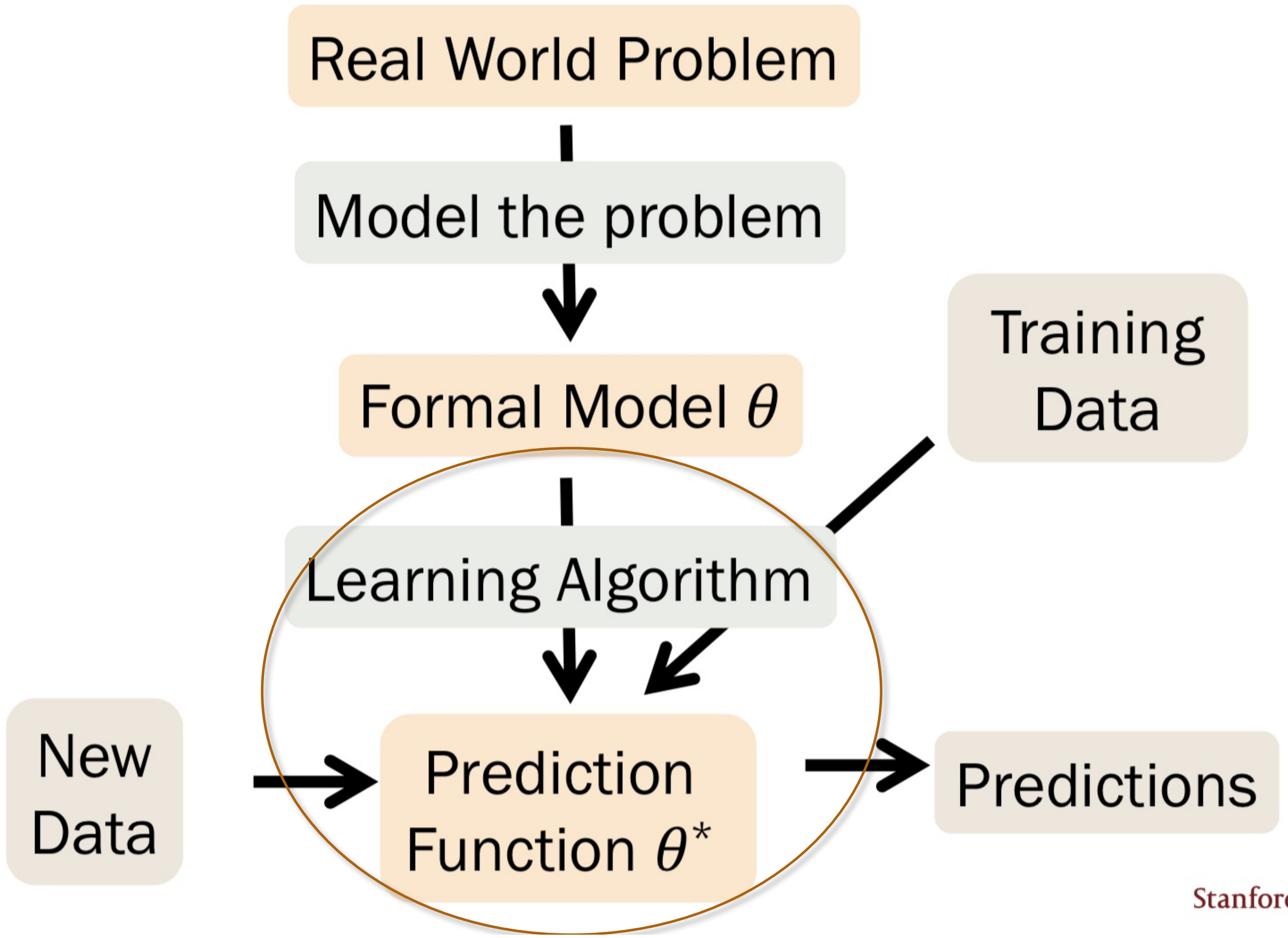
1. Garbage In, Garbage Out.

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.

Machine Learning



Overcoming Ossified Biases In Training Data

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)

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

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

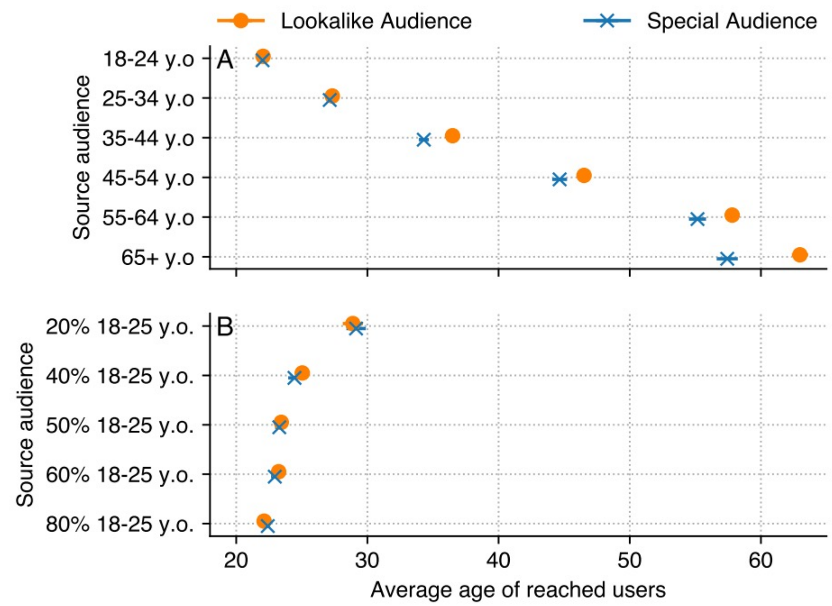
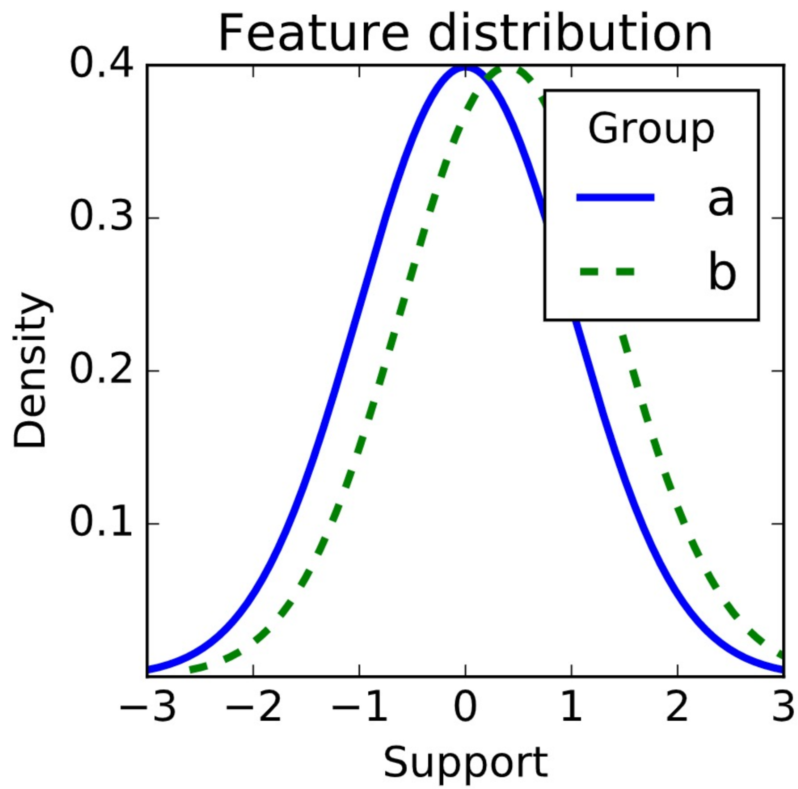
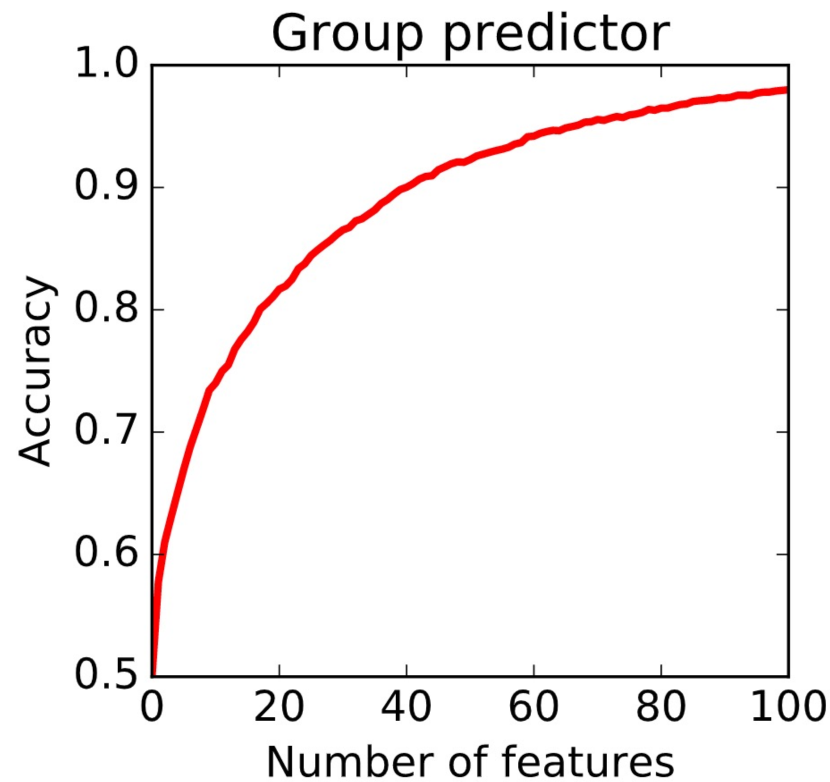


Figure 3: Age breakdown of ad delivery to Lookalike and Special Ad audiences created from the same source audience, using the same ad creative. We can observe extremely similar levels of bias, despite the lack of age as an input to Special Ad audiences. Panel A shows the results for source audiences consisting only of users in one age bracket. Panel B shows the results of mixing the youngest and the oldest users in different proportions.





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

In what way is Fairness through Unawareness Fair?

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.

In what way is “Fairness through Unawareness” Fair?

Procedural Fairness:

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

In our case, Facebook increases procedural fairness by removing “age, gender, relationship status, religious views, school, political views, interested in, zip code” from algorithm that creates Lookalike/SpecialAd audiences.

Distributive Fairness:

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

In our case, little increase in distributive fairness because the outcome does not change very much.

Let’s Try Fairness Through Awareness!

Awareness of what?

Independence & Demographic Parity

Sensitive Attribute = A

Other group membership = B

Classifier or Score = R

The random variables (A,R) satisfy independence if $A \perp R$

For binary classification (our jam!)

- $P\{R=1 | A=a\} = P\{R=1 | A=b\}$
- E.g. acceptance rate should be the same for all groups

Relaxed Independence Condition

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

- Imagine people from group A and group B apply to a job.
- The percentage accepted from group B must be at least 80% of the percentage from group A accepted.

$$\frac{\mathbb{P}\{R=1 \mid A=a\}}{\mathbb{P}\{R=1 \mid A=b\}} \geq 1-\epsilon \,. \quad \text{where } \varepsilon = 0.2.$$

Disparate Quality & Self-Fulfilling Properties

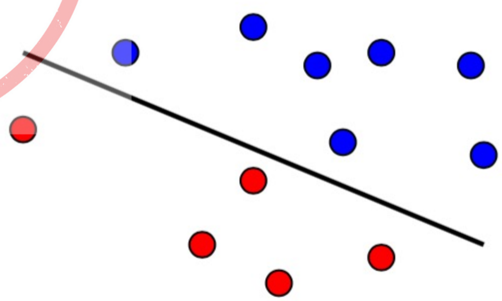
Dwork et. al. 2012, “Fairness Through Awareness” https://dl.acm.org/doi/10.1145/2090236.2090255

What does fairness through awareness fail to capture?

- If the classifier is significantly less good at identifying quality candidates in a minority group (relative to the data), the candidates accepted might be evaluated as worse, 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.”

Discrimination Intuition

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



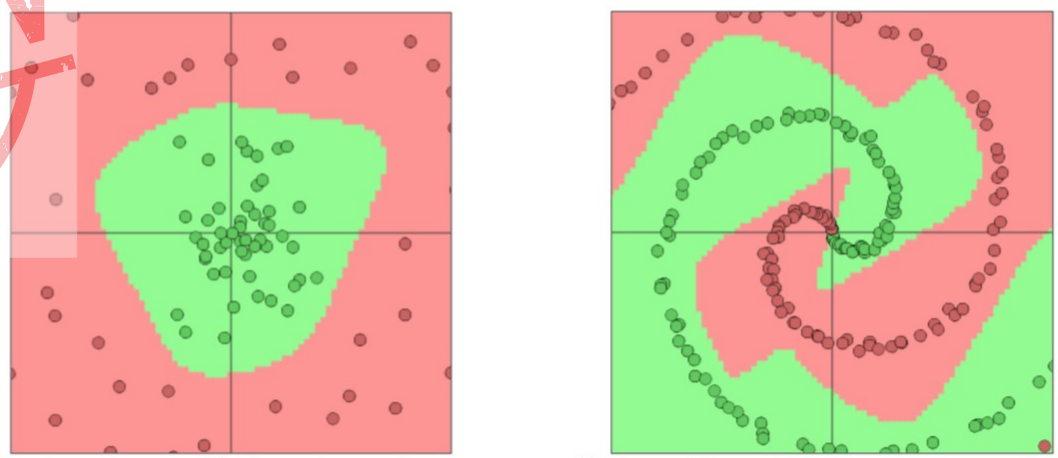
$$\theta^T \mathbf{x} = 0$$

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

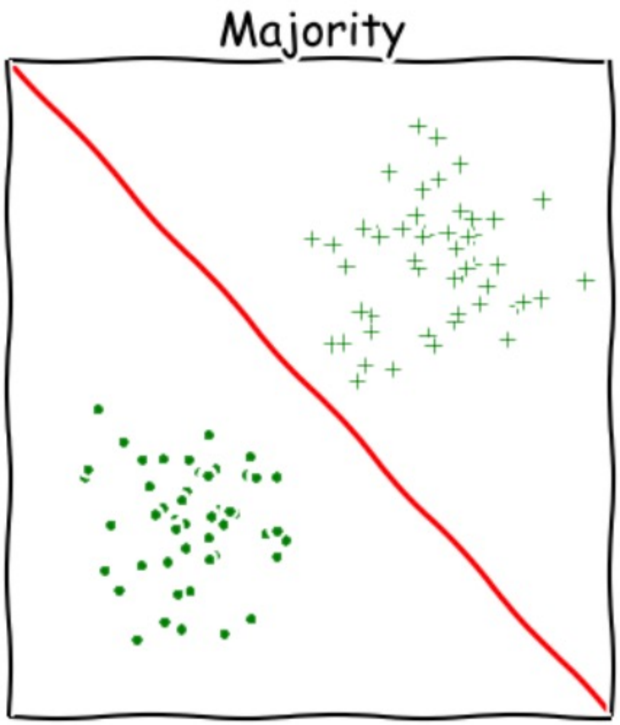
- We call such data (or the functions generating the data) "**linearly separable**"
- **Naïve bayes is linear too** as there is no interaction between different features.

Some Data Not Linearly Seperable

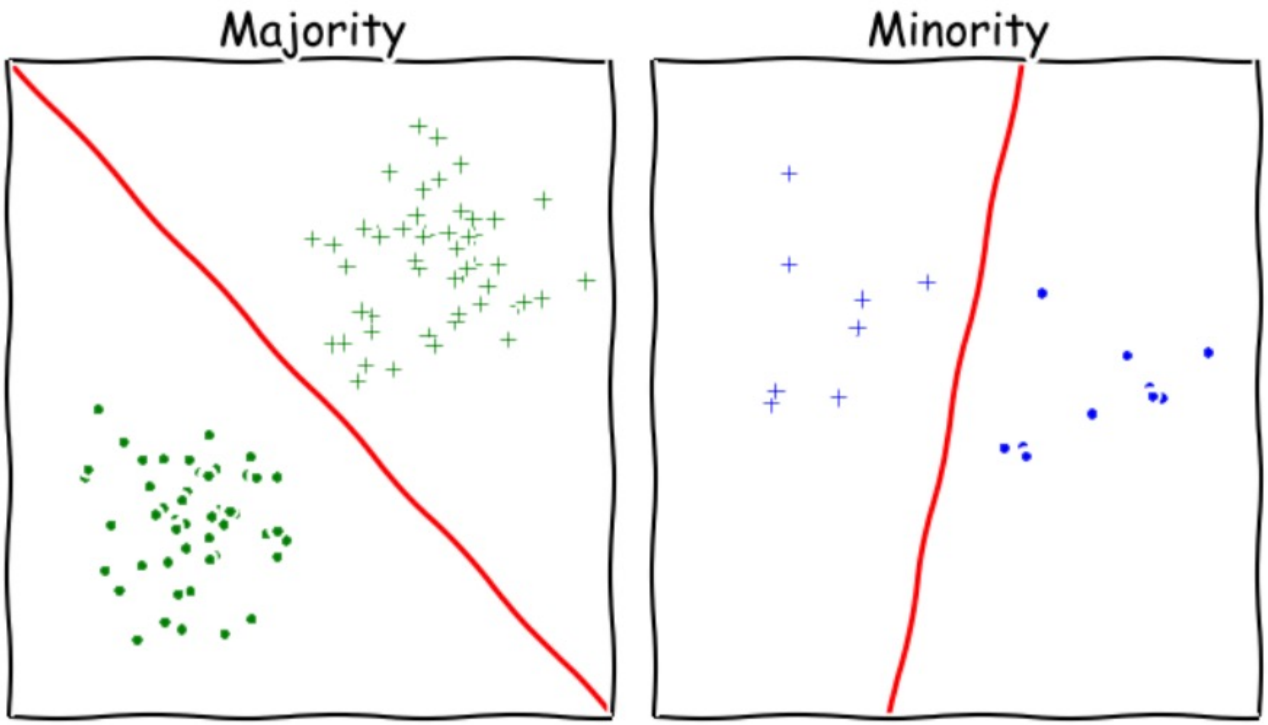
Some data sets/functions are not separable



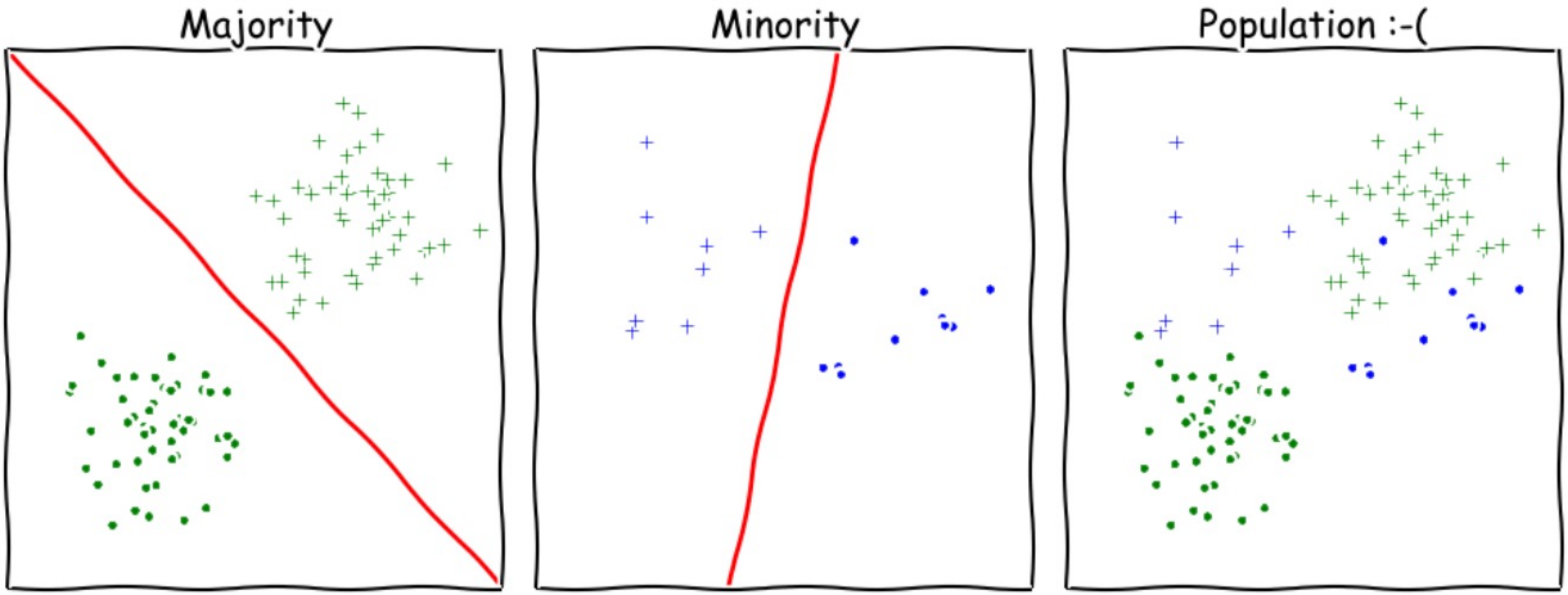
- Not possible to draw a line that successfully separates all the $y = 1$ points (green) from the $y = 0$ points (red)
- Despite this fact, logistic regression and Naive Bayes still often work well in practice



Classification of the minority group may be worse.



Classification of the minority group may be worse.



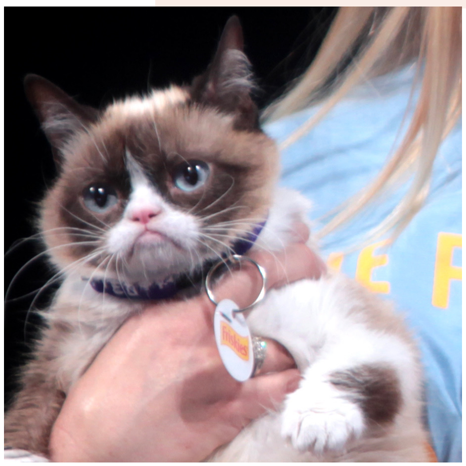
Classification of the minority group may be worse even with awareness.

False Positives and False Negatives

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

False Positives and False Negatives

	Condition y = 1	Condition y = 0
Event ŷ = 1	True positive	False positive
Event ŷ = 0	False Positive	False Negative



= CAT! (True positive)

False Positives and False Negatives

	Condition y = 1	Condition y = 0
Event ŷ = 1	True positive	False positive
Event ŷ = 0	False Positive	False Negative

= CAT! (False Positive)

Fairness through Separation

Motivating idea: in some cases, a sensitive attribute is correlated with the target. Separation criterion allows correlation between the score and the sensitive attribute to the *extent that it is justified* by the target variable.

Definition: Random variables (R,A,Y) satisfy separation if $R \perp A | Y$

Separation means that the true positive and false positive rates for both groups will be equal.

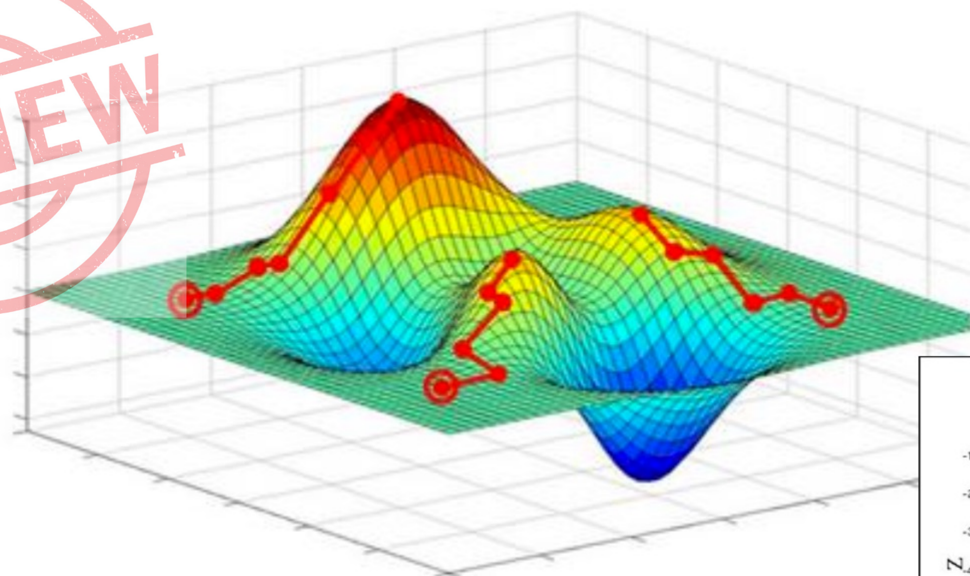
$$\mathbb{P}\{R=1 \mid Y=1, A=a\} = \mathbb{P}\{R=1 \mid Y=1, A=b\}$$
$$\mathbb{P}\{R=1 \mid Y=0, A=a\} = \mathbb{P}\{R=1 \mid Y=0, A=b\}$$

How do we address bias in machine learning?

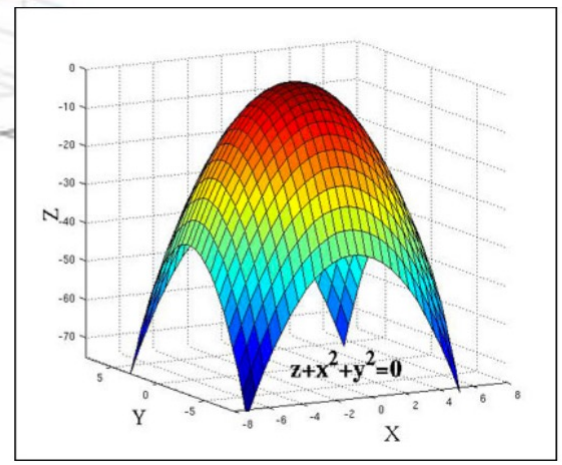
Algorithmic Auditing!

Gradient Ascent

REVIEW



Logistic regression LL function is convex



Walk uphill and you will find a local maxima
(if your step size is small enough)

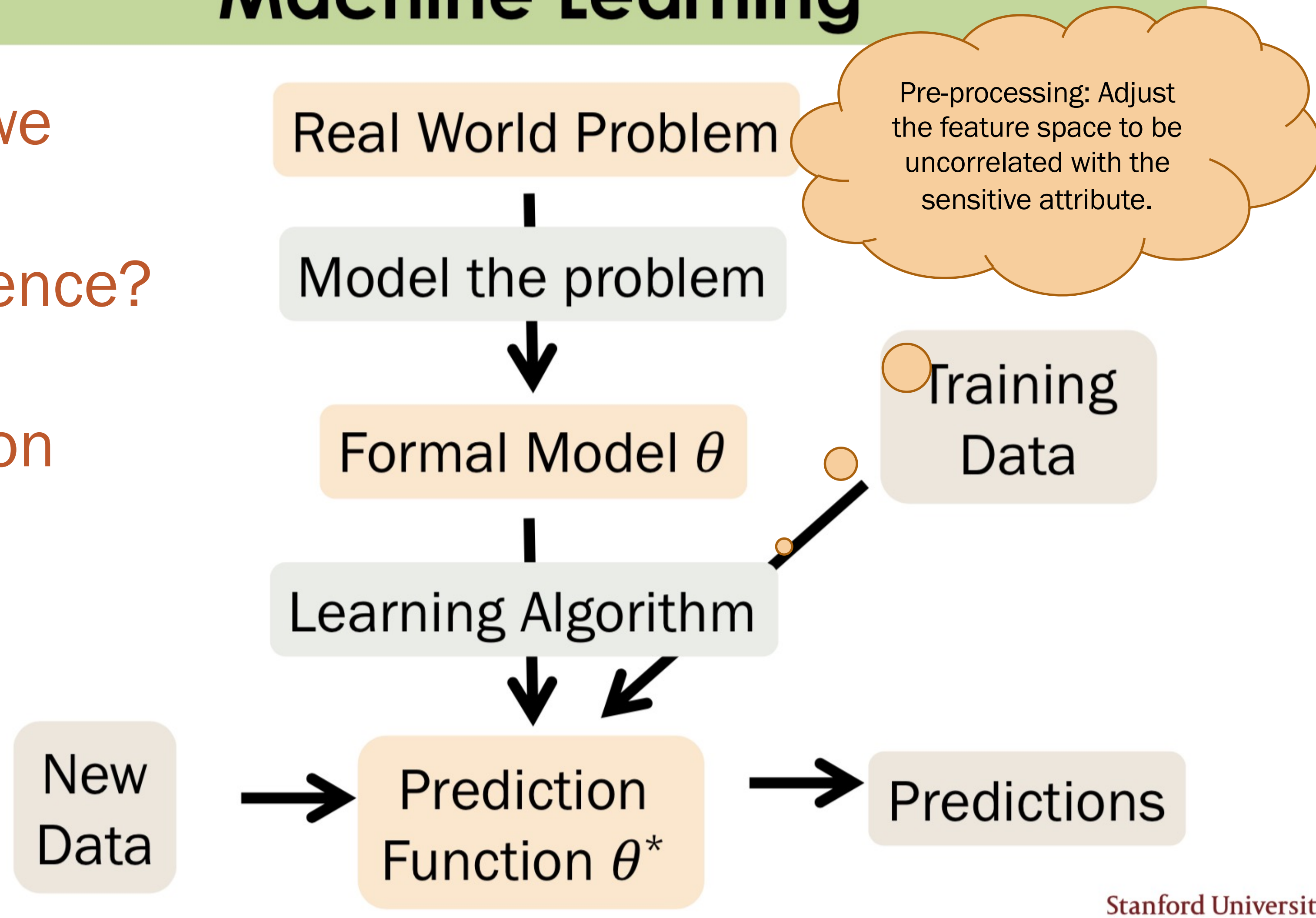
Intersectionality & Subgroup Analysis

- Audits often only focus on a federally protected categories (race, religion, national origin, age, sex, disability, veteran status).
- Exclusion can also correlate with subgroup or intersectional categories within axes of existing discrimination
- Audits for “single-axis” discrimination will miss it, and legal standards do not require audits for multi-axis discrimination

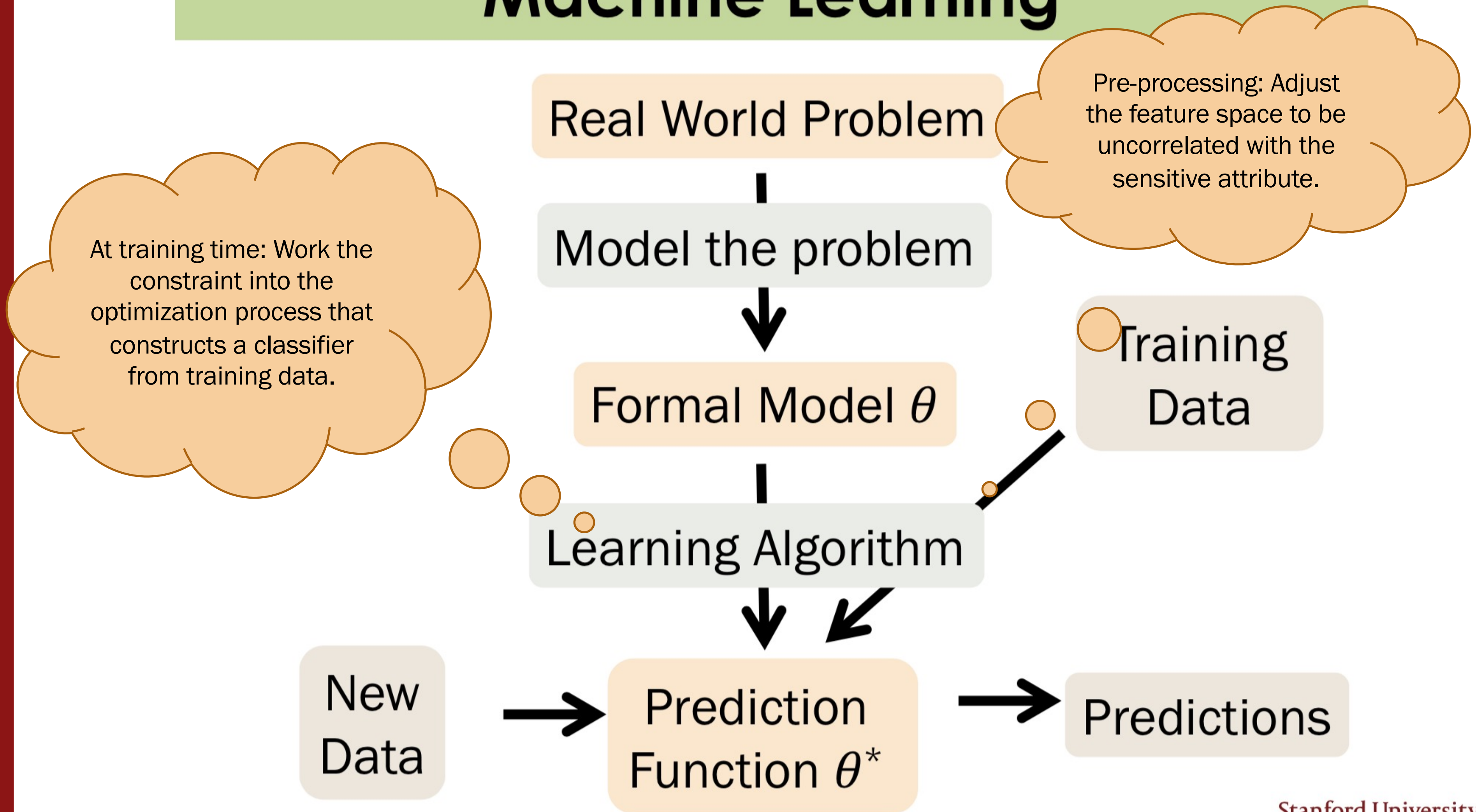
(see Crenshaw 1989, 140; Raji and Buolamwini 2019; Wilson et. al 2021)

Machine Learning

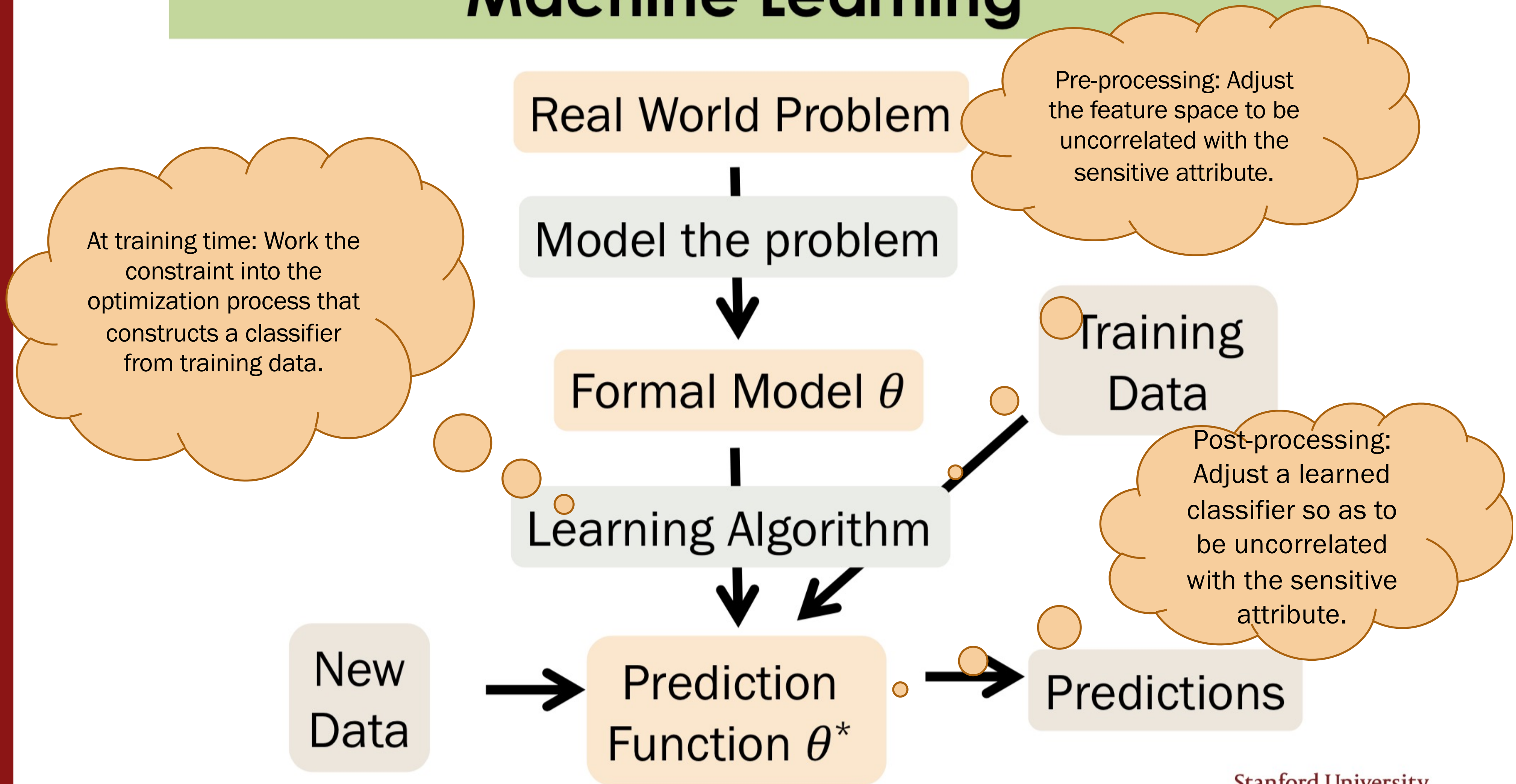
How can we achieve independence? A Formal Intervention



Machine Learning



Machine Learning



Stanford University

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**


Leveling Up & Leveling Down:
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.

Create Your Own Representations

You are 109 graduates now – you have the power!

IN OUR GLORY PHOTOGRAPHY AND BLACK LIFE 55

Snapshot of Veodis Watkins. *1949. Courtesy of bell hooks. Photographer unknown.*

Photographic representation as a site of subversion

bell hooks, “In Our Glory: Photography and Life”

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.

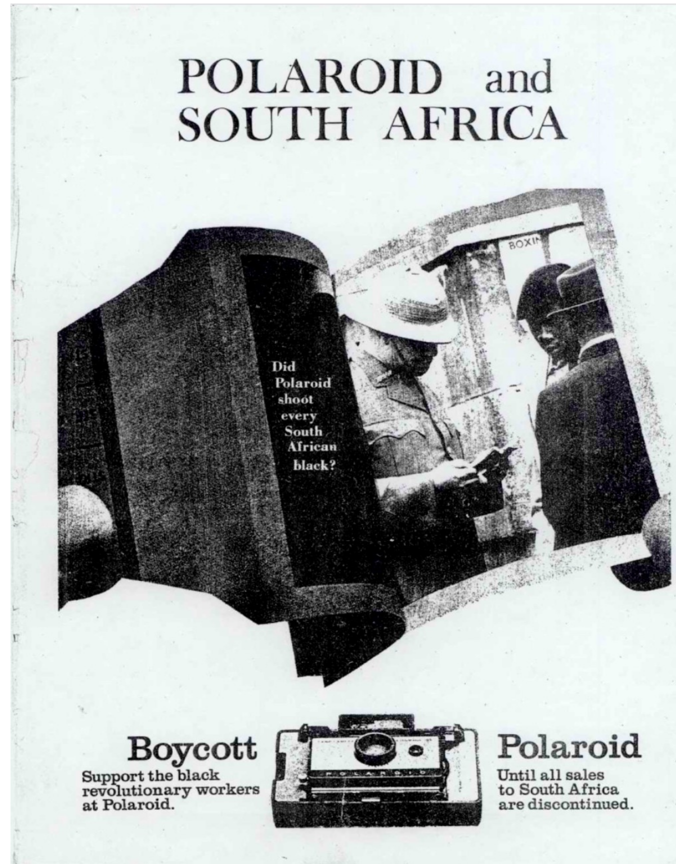
This was used to create passbook photographs for the Apartheid government.

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.'"



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

Thank you!

Office Hours: https://calendly.com/kathleencreel
Email: kcreel@stanford.edu