Security and Privacy of ML

Fairness in ML

Shang-Tse Chen

Department of Computer Science
& Information Engineering
National Taiwan University

Many slides adapted from Moritz Hardt's NeurIPS'17 tutorial

HW 1 Score Released

- Phase 1:
  - 5 models: wrn16, preresnet20, rir_pgd, densenet_fgsm, ror3_pgd
  - You get 1 point if the accuracy of each model <= 250/500
- Phase 2:
  - Model: Ensemble of resnet56_fgsm, nin_fgsm, resnet110_pgd
  - Accuracy <=100/500: 5 points, (101-200)/500: 4 points, …, (401-500)/500: 1 points

(Un)Fairness in The Real World

(Un)Fairness in The Real World

> **DHH**
> @dhh
>
> The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.
>
> 2:34 PM · Nov 7, 2019 · Twitter for iPhone

Source: Tweet by DAVID HEINEMEIER HANSSON

> **Steve Wozniak**
> @stevewoz
>
> Replying to @dedwards93 @dhh and @AppleCard
>
> I'm a current Apple employee and founder of the company and the same thing happened to us (10x) despite not having any separate assets or accounts. Some say the blame is on Goldman Sachs but the way Apple is attached, they should share responsibility.
>
> 1:06 AM · Nov 10, 2019 · Twitter Web App

Source: Tweet by Steve Woznaik

(Un)Fairness in The Real World

Clarification from the company

We wanted to address some recent questions regarding the Apple Card credit decision process.

With Apple Card, your account is individual to you; your credit line is yours and you establish your own direct credit history. Customers do not share a credit line under the account of a family member or another person by getting a supplemental card.

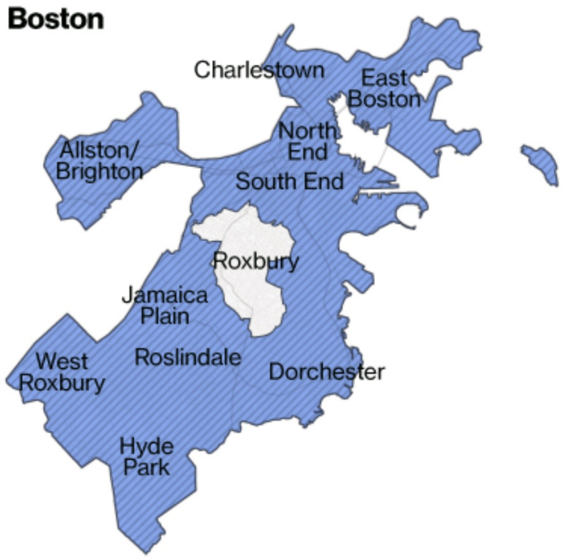
As with any other individual credit card, your application is evaluated independently. We look at an individual's income and an individual's creditworthiness, which includes factors like personal credit scores, how much debt you have, and how that debt has been managed. Based on these factors, it is possible for two family members to receive significantly different credit decisions.

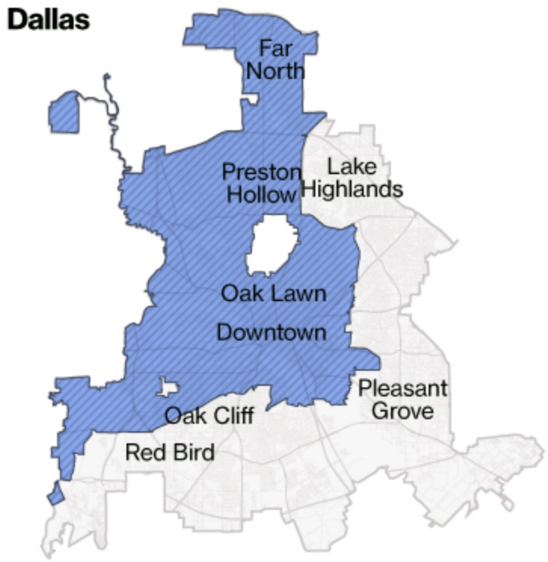
In all cases, we have not and will not make decisions based on factors like gender.

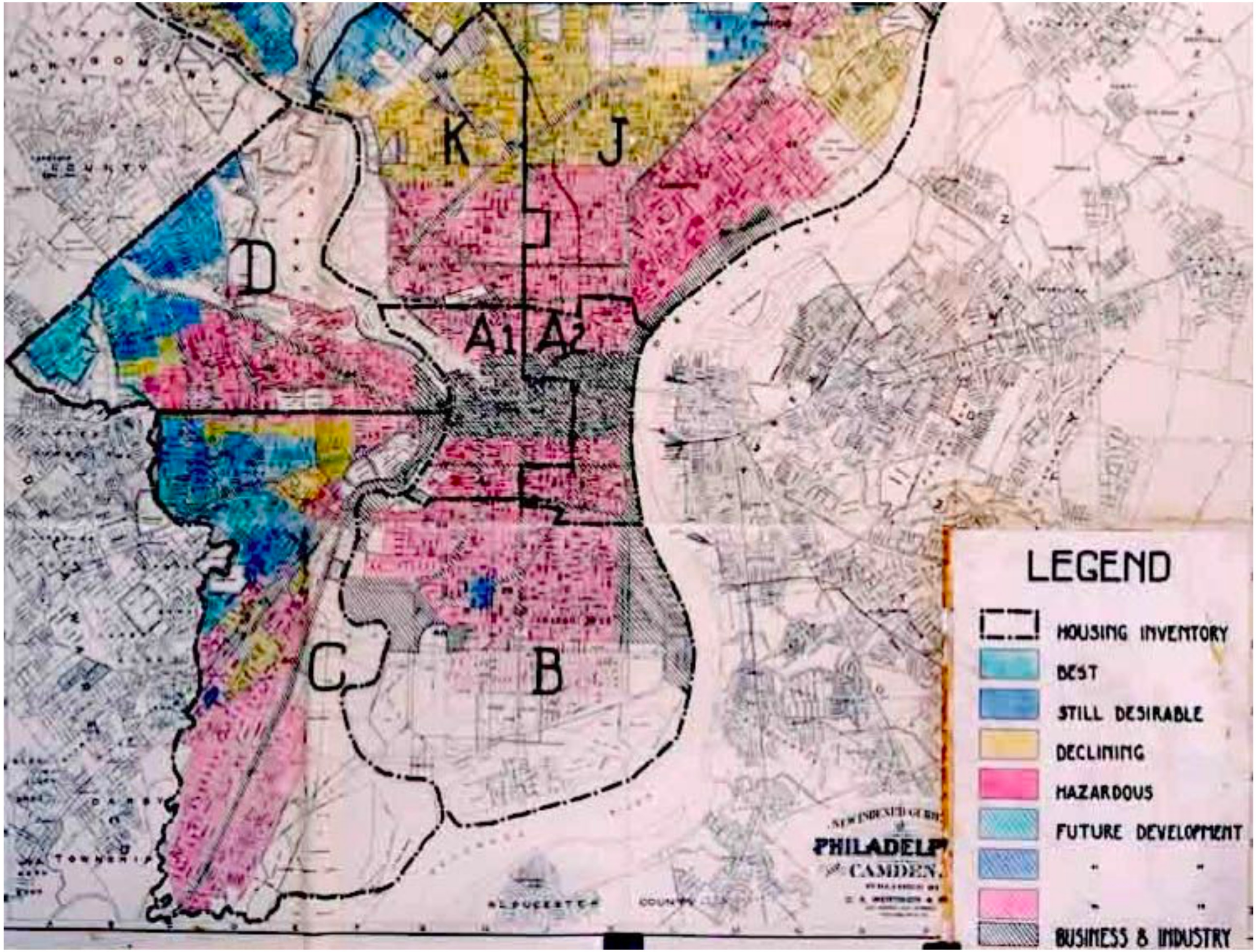
Finally, we hear frequently from our customers that they would like to share their Apple Card with other members of their families. We are looking to enable this in the future.

- Andrew Williams, Goldman Sachs Spokesperson

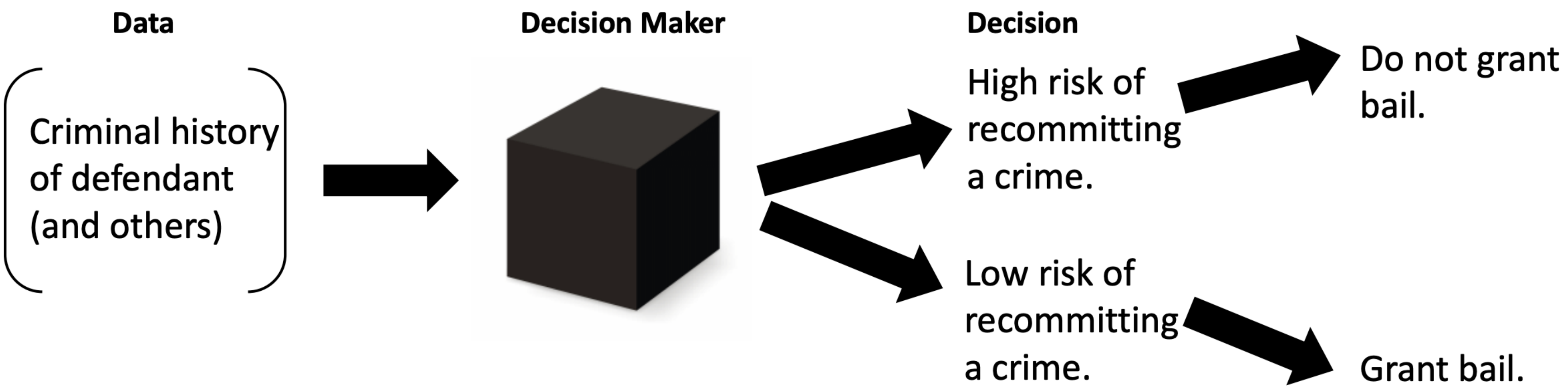
Amazon Same-Day Delivery Coverage






K
J
D
A1
A2
C
B
LEGEND
HOUSING INVENTORY
BEST
STILL DESIRABLE
DECLINING
HAZARDOUS
FUTURE DEVELOPMENT
BUSINESS & INDUSTRY

Recidivism Prediction with ML

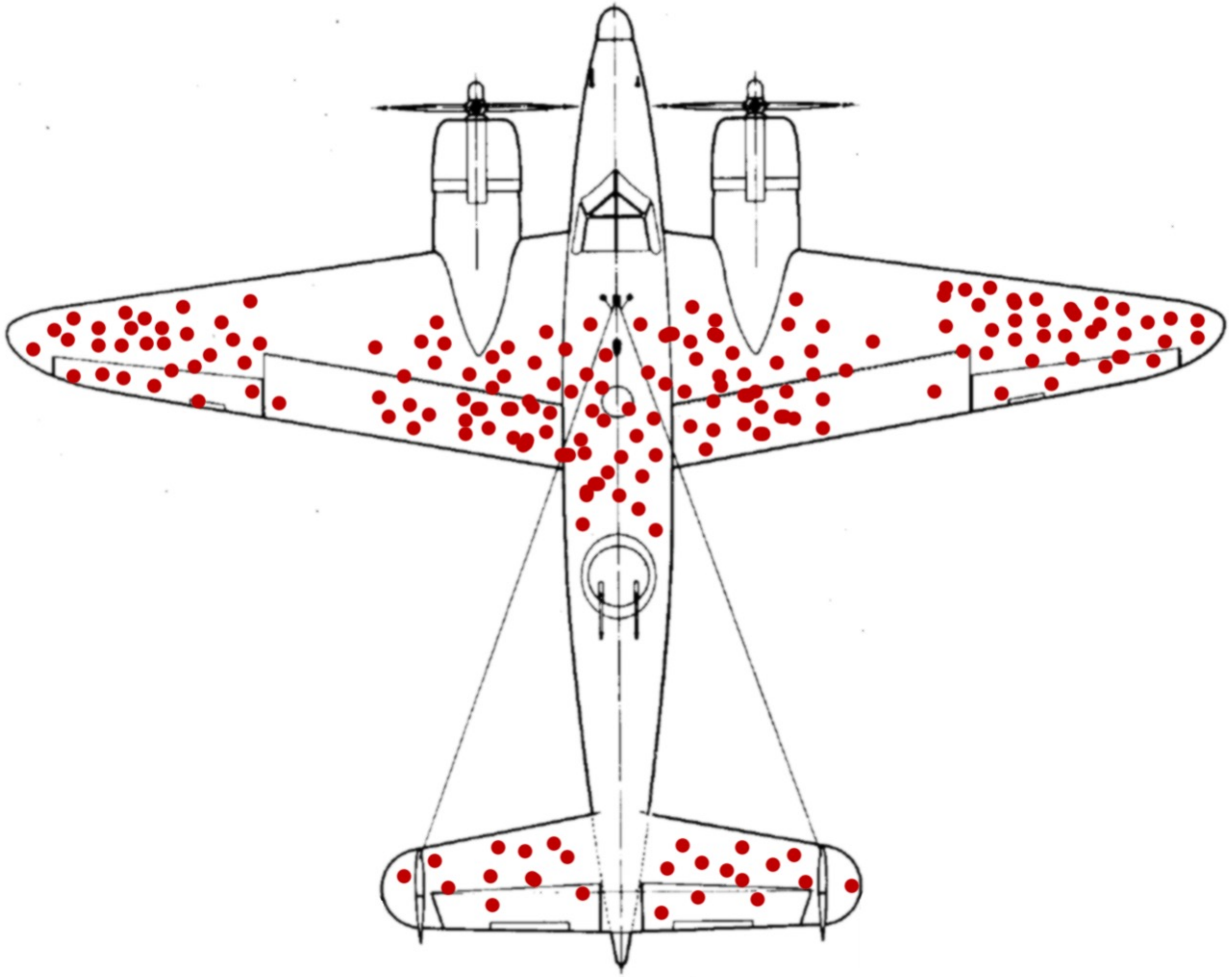


Machine (and Human) Bias

There's software used across the U.S. to predict future criminals, and it's biased against blacks. [Angwin et al., 2016]

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Survival Bias



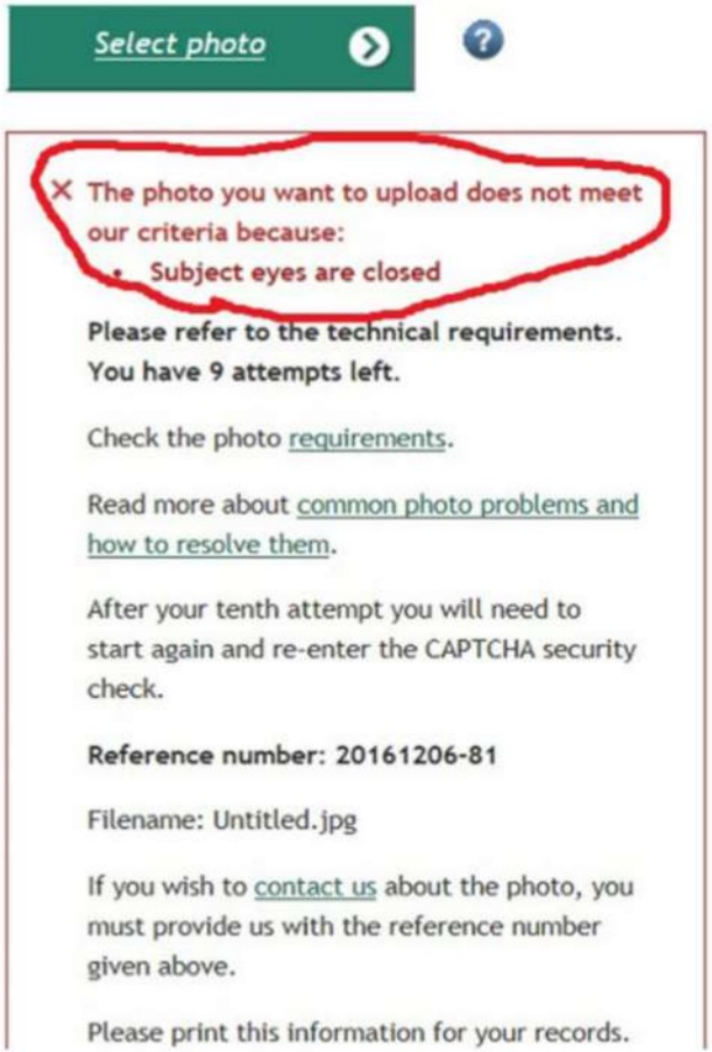
Fairness in Computer Vision

Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Fairness in Computer Vision

Fairness in Computer Vision



Select photo

✕ The photo you want to upload does not meet our criteria because:

- Subject eyes are closed

Please refer to the technical requirements.
You have 9 attempts left.

Check the photo requirements.

Read more about common photo problems and how to resolve them.

After your tenth attempt you will need to start again and re-enter the CAPTCHA security check.

Reference number: 20161206-81

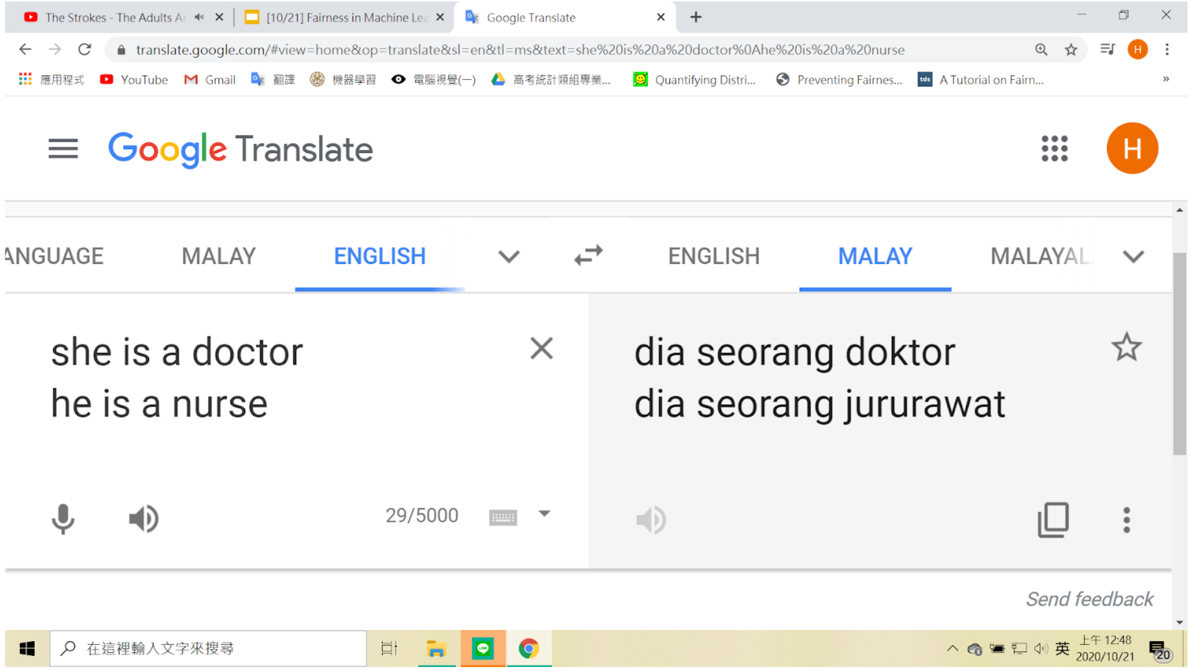
Filename: Untitled.jpg

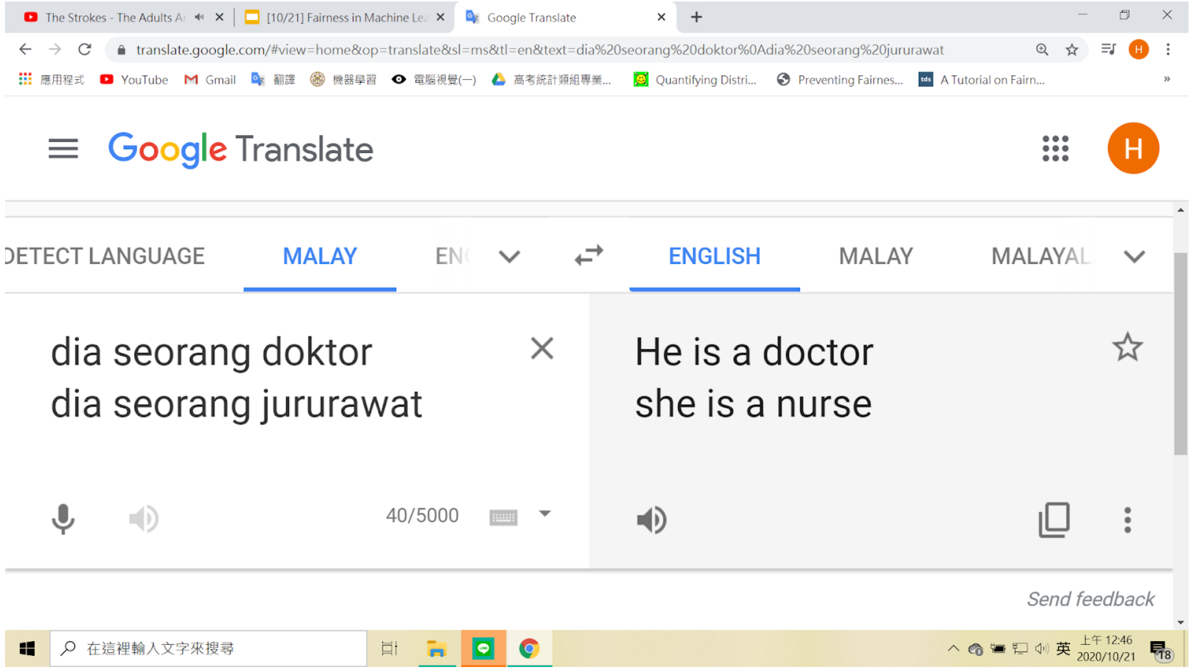
If you wish to contact us about the photo, you must provide us with the reference number given above.

Please print this information for your records.

A screenshot of New Zealand man Richard Lee's passport photo rejection notice, supplied to Reuters December 7, 2016. Richard Lee/Handout via REUTERS

Fairness in NLP





Fairness in NLP

- Word embeddings may contain bias from data

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

KIMBERLY WHITE / STRINGER

Tech policy / AI Ethics

A leading AI ethics researcher says she's been fired from Google

Timnit Gebru says she's facing retaliation for conducting research that was critical of Google and sending an email "inconsistent with the expectations of a Google manager."

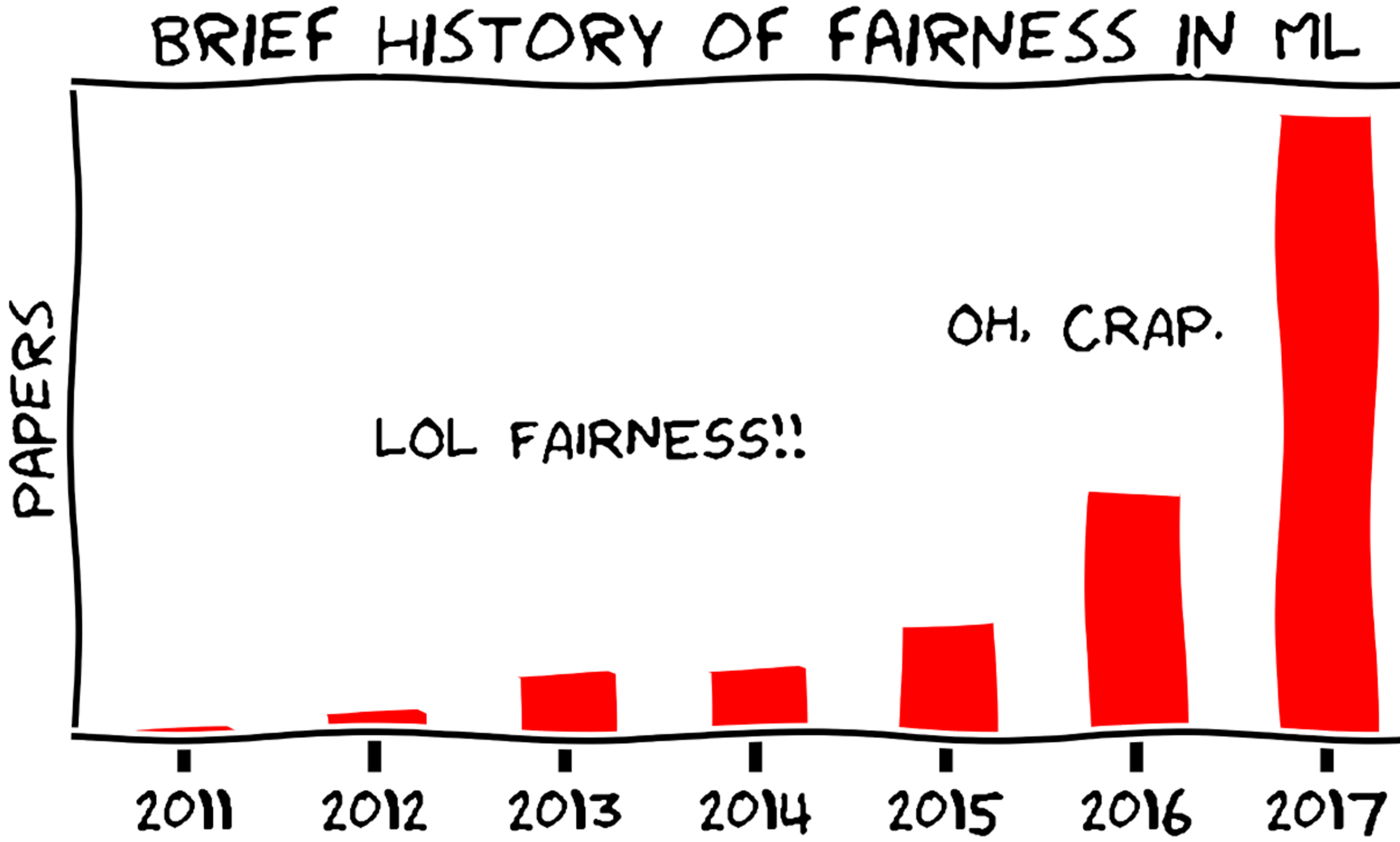
Definition of Fairness

- Discrimination refers to unfavorable treatment of people due to the membership to certain demographic groups
- Illegal to distinguish based on attributes protected by law
- Legally protected domains/classes: Race, Color, Sex, Religion
- Fairness in a decision making implies the designing algorithms that make fair predictions devoid of discrimination

Why Important in ML

- ML increasingly being used in high-impact domains such as credit, employment, education and criminal justice
- Sources of errors: sample size disparity, biases in data
- Decisions made by unfair ML models will increase bias in future data, making a **vicious cycle**

Number of Papers on ML Fairness



Source:
https://fairmlclass.github.io/1.html#/4

Mathematical Formulation

- X: set of individuals
- A: set of protected attributes (those protected by law)
- Z: set of remaining attributes
- Y: set of the outcomes
- Individual Predictor: $\mathcal{H} : X \to Y$
- Group-conditional predictor consists of a set of mappings, one for each group of population $\mathcal{H} = \{\mathcal{H}_S\} \forall S \subset X$

Mathematical Formulation

X (features)

A (protected attribute)

Y (label)

X1	…	…	…	…	Race	Bail
0	…	0	1	…	1	Y
1	…	1	0	…	1	N
1	…	1	0	…	0	N
..	…	…	…	…	…	…

$$\mathbb{P}_a\{E\} = \mathbb{P}\{E \mid A = a\}.$$

What is Fair?

- Many definitions
- There is no single best definition
- We will introduce and discuss some popular definitions

Demographic parity (group fairness)

Definition. Classifier C satisfies *demographic parity* if C is independent of A.

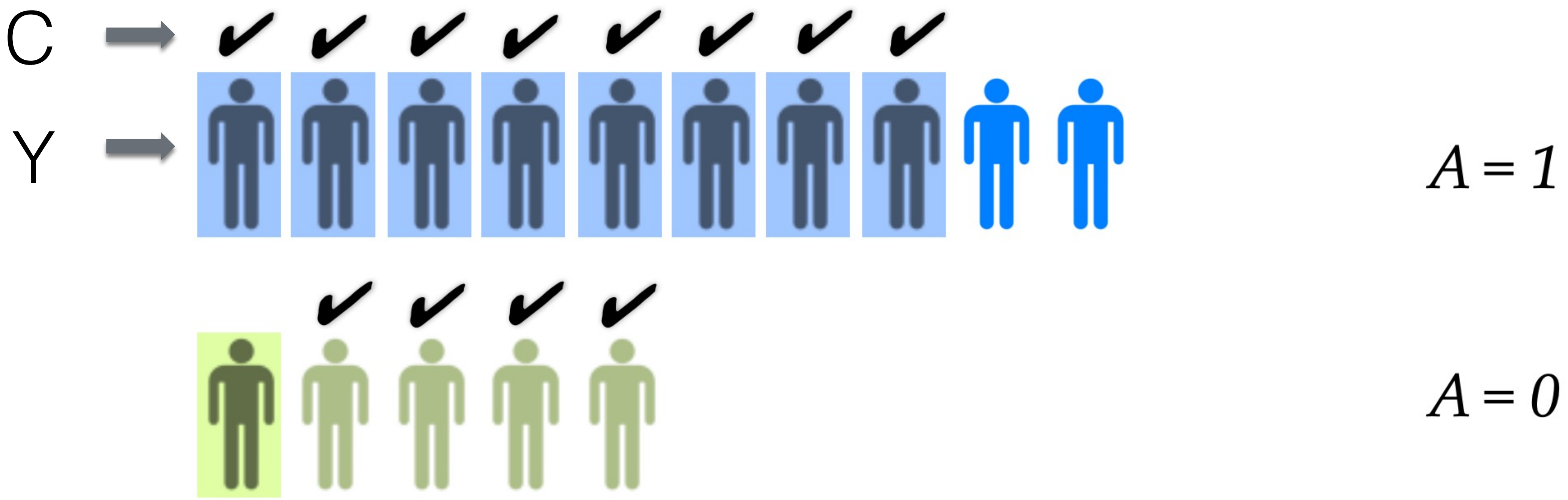
When C is binary $0/1$-variables, this means
$\mathbb{P}_a\{C = 1\} = \mathbb{P}_b\{C = 1\}$ for all groups a, b.

Approximate versions:

$$\frac{\mathbb{P}_a\{C = 1\}}{\mathbb{P}_b\{C = 1\}} \geq 1 - \epsilon$$

$$|\mathbb{P}_a\{C = 1\} - \mathbb{P}_b\{C = 1\}| \leq \epsilon$$

Demographic parity Issues



- Does not seem “fair” to allow random performance on A = 0
- Perfect classification is impossible

Accuracy Parity

Definition. Classifier C satisfies *accuracy parity* if $\mathbb{P}_a\{C = Y\} = \mathbb{P}_b\{C = Y\}$ for all groups a, b.

- Pros:
  - Random guessing doesn't work
  - Allows perfect classifier
- Cons:
  - Error types matter!
  - Allows you to make up for rejecting qualified women by accepting unqualified men

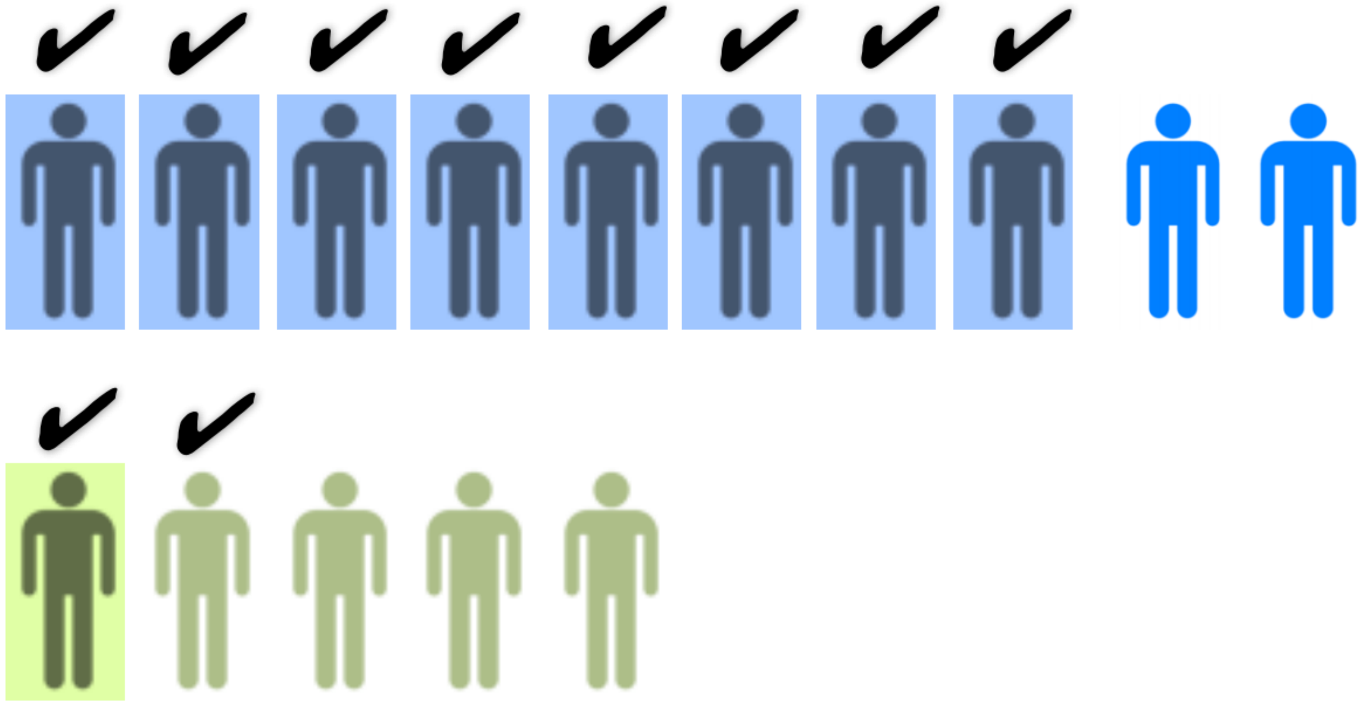
True Positive Parity (TPP)
(or equal opportunity)

Assume C and Y are binary $0/1$-variables.

Definition. Classifier C satisfies *true positive parity* if
$\mathbb{P}_a\{C = 1 \mid Y = 1\} = \mathbb{P}_b\{C = 1 \mid Y = 1\}$ for all groups a, b.

- When positive outcome (1) is desirable
- Equivalently, primary harm is due to false negatives
  - Deny bail when person will not recidivate

True Positive Parity (TPP)



$A = 1$

$A = 0$

Forces similar performance on $Y = 1$

False Positive Parity (FPP)

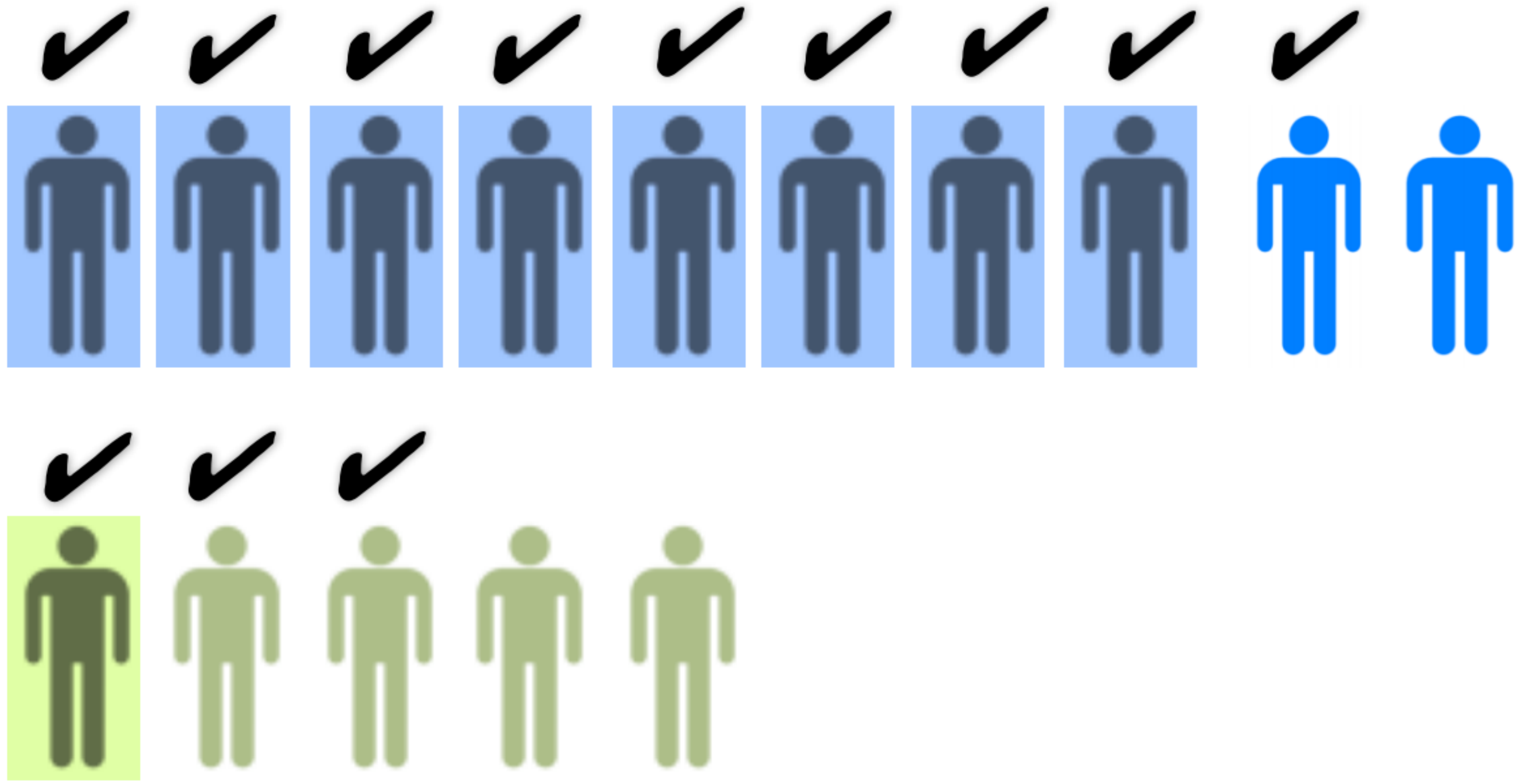
Assume C and Y are binary $0/1$-variables.

Definition. Classifier C satisfies *false positive parity* if $\mathbb{P}_a\{C = 1 \mid Y = 0\} = \mathbb{P}_b\{C = 1 \mid Y = 0\}$ for all groups a, b.

- TPP + FPP: Equalized Odds, or
  Positive Rate Parity

R satisfies equalized odds if R is conditionally independent of A given Y.

Positive Rate Parity



$A = 1$

$A = 0$

Predictive Value Parity

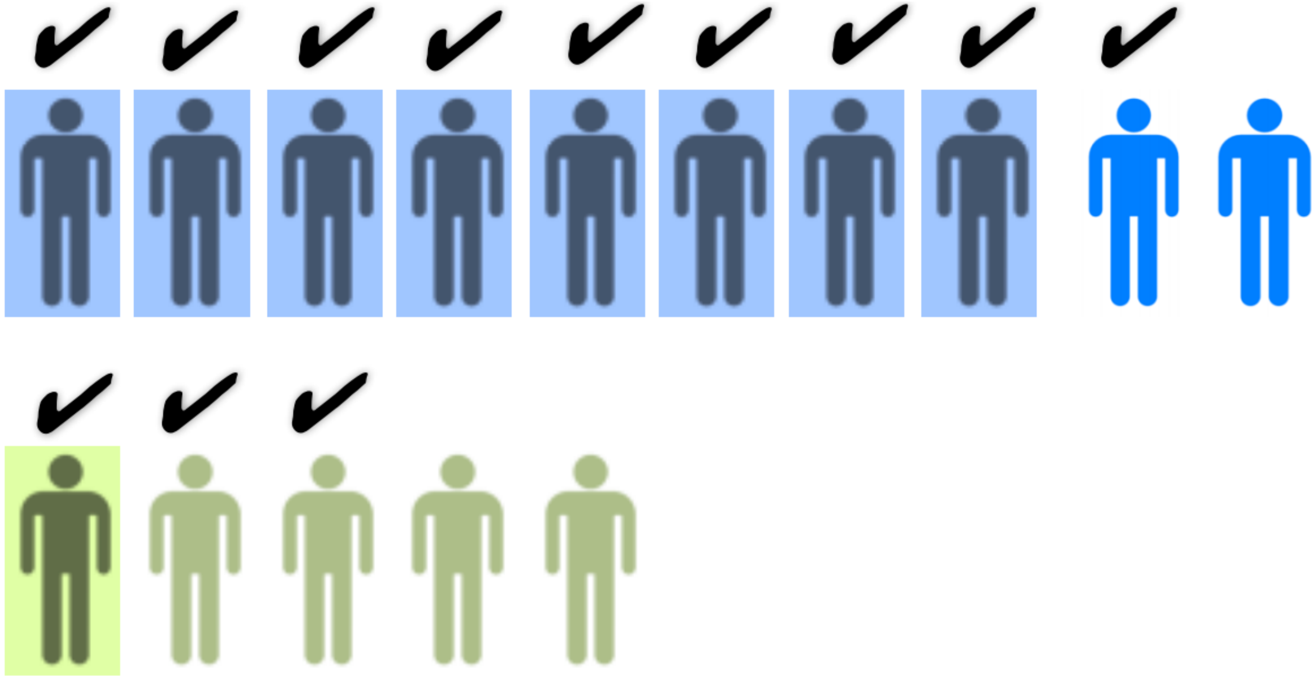
Assume C and Y are binary $0/1$-variables.

Definition. Classifier C satisfies

- *positive predictive value parity* if for all groups a, b:
  $\mathbb{P}_a\{Y = 1 \mid C = 1\} = \mathbb{P}_b\{Y = 1 \mid C = 1\}$
- *negative predictive value parity* if for all groups a, b:
  $\mathbb{P}_a\{Y = 1 \mid C = 0\} = \mathbb{P}_b\{Y = 1 \mid C = 0\}$
- *predictive value parity* if it satisfies both of the above.

Equalized chance of success given acceptance

Predictive Value Parity



$A = 1$

$A = 0$

$$P_1[Y = 1 \mid C = 1] = \ 8/9 \qquad P_1[Y = 1 \mid C = 0] = \ 0$$

$$P_0[Y = 1 \mid C = 1] = \ 1/3 \qquad P_0[Y = 1 \mid C = 0] = \ 0$$

Trade-off

$$\mathbb{P}_a\{C = 1\} \neq \mathbb{P}_b\{C = 1\}$$

Proposition. Assume differing base rates and an imperfect classifier $C \neq Y$. Then, either

- positive rate parity fails, or
- predictive value parity fails.

Fairness through Blindness

- Ignore all protected attributes
- Issue: other non-protected attributes might correlate with the protected attributes
  - E.g., Guess gender by name

Counterfactual Measures

Predictor $\mathcal{H}$ is counterfactually fair, if

$$\mathcal{P}(\mathcal{H}_{A=a} = y|Z = z) = \mathcal{P}(\mathcal{H}_{A=a'} = y|Z = z)$$

- A predictor is fair if its output remains the same when the protected attribute is flipped to its counterfactual value.
- **Issue**: susceptible to hindsight bias and outcome bias (i.e. evaluating the quality of a decision when its outcome is already known)

Individual Fairness

Treat *similar* individuals *similarly*

Similar for the purpose of the classification task

Similar distribution over outcomes

Examples of Individual Fairness

- Financial/insurance risk metrics
- IBM's AALIM (Advanced Analytics for Information Management) system: treating similar patients similarly

Individual Fairness

Definition 4 (Individual fairness) *A predictor achieves individual fairness iff* $\mathcal{H}(x_i) \approx \mathcal{H}(x_j) \mid d(x_i, x_i) \approx 0$ *where* $d : X \times X \to \mathbb{R}$ *is a distance metric for individuals.*

- Captured by (D, d)-Lipschitz property: $D(\mathcal{H}(x_i)_Y, \mathcal{H}(x_j)_Y) \leq d(x_i, x_j)$

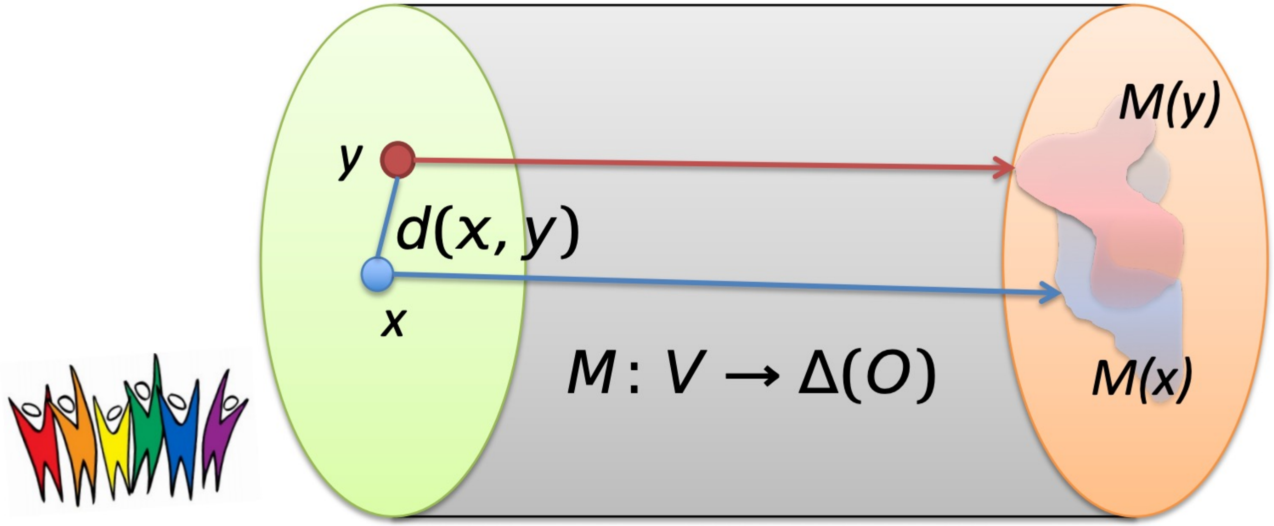
ISSUES: This notion delegates the responsibility of ensuring fairness from the predictor to its distance metric. If the distance metric uses the protected attributes directly (or indirectly), the predictor (satisfying above)could still be discriminatory

Individual Fairness: Definition

Metric $d : V \times V \rightarrow \mathbb{R}$

Lipschitz condition $\|M(x) - M(y)\| \leq d(x, y)$

This talk: Statistical distance in [0,1]



V: Individuals O: outcomes

Connection to Differential Privacy

- Close connection between individual fairness and **differential privacy** [Dwork-McSherry-Nissim-Smith'06]

  DP: Lipschitz condition on set of databases

  IF: Lipschitz condition on set of individuals

	Differential Privacy	Individual Fairness
Objects	Databases	Individuals
Outcomes	Output of statistical analysis	Classification outcome
Similarity	General purpose metric	Task-specific metric

Fairness through Privacy?

- Fairness: Avoid using certain attributes
- Privacy: protect certain attributes from being inferred

"At worst, **privacy solutions can hinder efforts to identify classifications that unintentionally produce objectionable outcomes**—for example, differential treatment that tracks race or gender—by limiting the availability of data about such attributes."

-- Dwork & Mulligan

Why So Many Definitions

- Different context and applications
- Different Stakeholders
- Impossibility theorems
  - Any overarching definitions will inevitably be vacuous

Goal is to build algorithmic systems that further human values, which can't be reduced to a formula

Fair Robust Active Learning by Joint Inconsistency

Tsung-Han Wu, Hung-Ting Su, **Shang-Tse Chen**, Winston H. Hsu

ICCV AROW 2023

國立臺灣大學 National Taiwan University

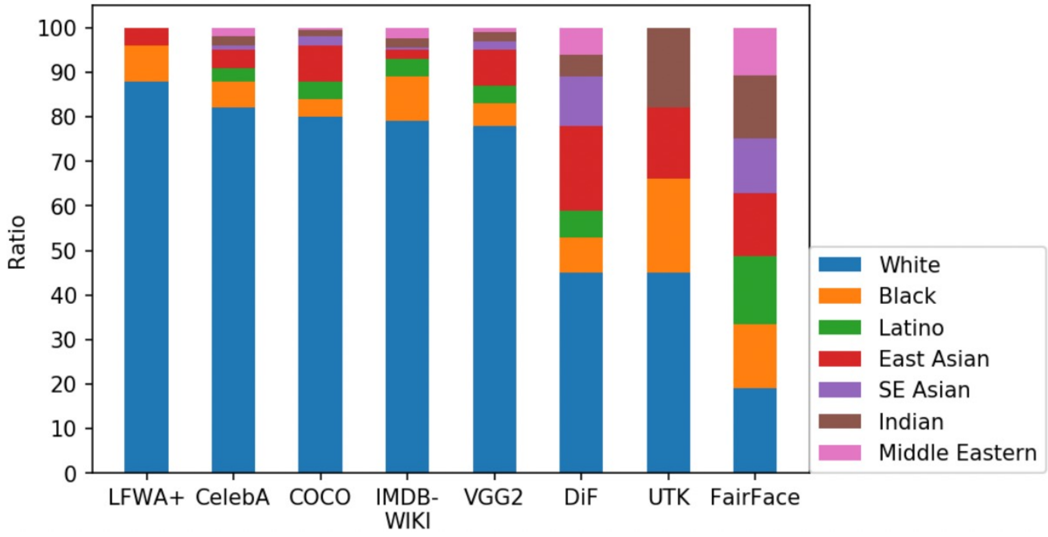
MobileDrive

NARLabs 財團法人國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing

NSTC 國家科學及技術委員會
National Science and Technology Council

Relations Between Data and Trustworthy AI

Fairness : Addressing Data Imbalance

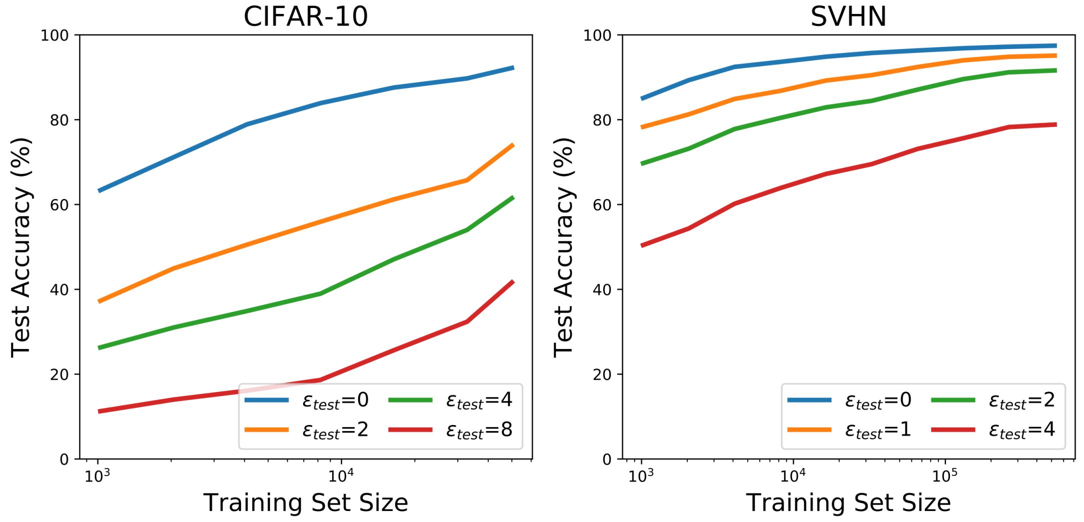


Facial recognition tool leads to mistaken-identity arrest of a Georgian black man

The arrest brings new attention to the use of a technology that results in a higher rate of misidentification of people of color.

By **Sahil Pawar** January 3, 2023

Robustness: Requiring More Labeled Data



Adversarially Robust Generalization Requires More Data

Ludwig Schmidt, MIT
Shibani Santurkar, MIT
Dimitris Tsipras, MIT
Kunal Talwar, Google Brain
Aleksander Mądry, MIT

[1] Kärkkäinen et al. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age". WACV 2021.
[2] Schmidt et al, "Adversarially Robust Generalization Requires More Data". NeurIPS 2018.

Trustworthy Applications

- Both Fairness and Robustness Requirements
  - Fairness among genders, ages, ethnicity
  - Robustness against adversarial attacks
- Examples: Medical Imaging, Facial Biometric Systems

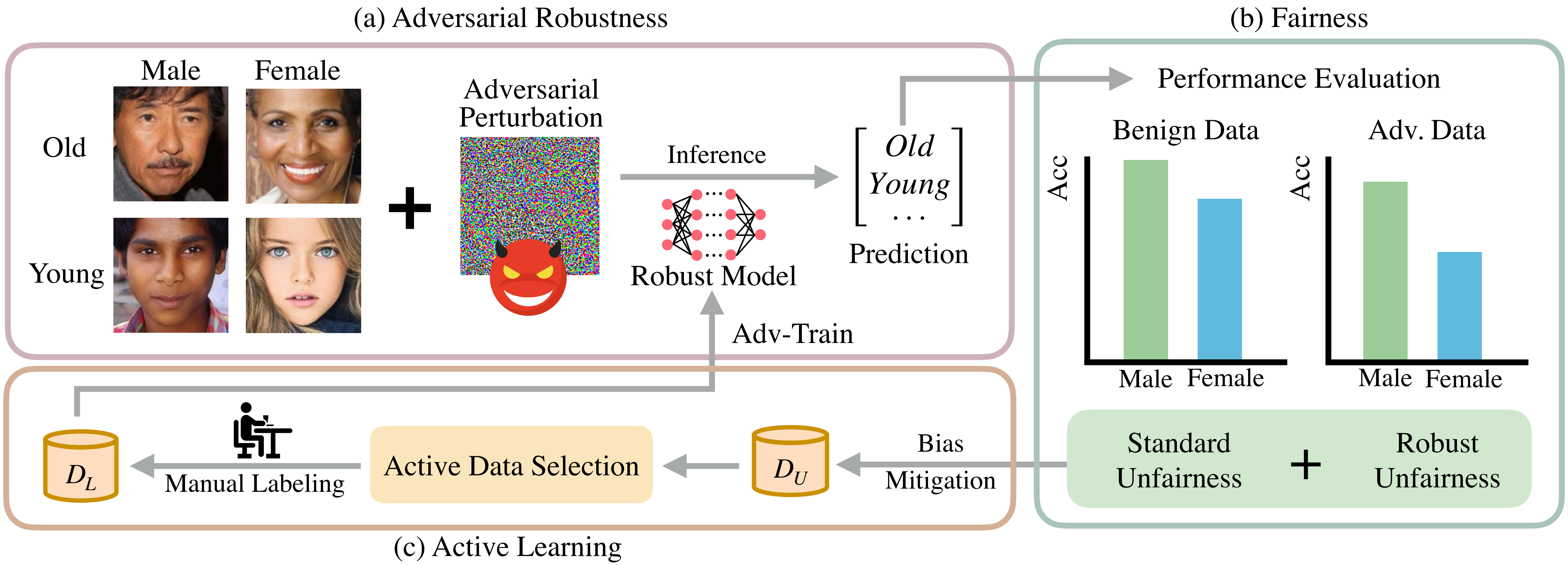
✓ Fairness
✓ Adversarial Robustness
✓ Costly Annotation Process

Our Motivation

Towards fair and robust visual apps with limited labeled data

Fair Robust Active Learning (FRAL)

First framework for annotation-expensive and safety-critical applications



Active Data Selection in FRAL

Existing "Standard" fairness-aware methods

- Randomly draw data from the worst-group for labeling
- Estimate expected unfairness reduction for each sample

Challenge under AT: (1) Amplified performance disparity (2) Unaffordable computational burdens

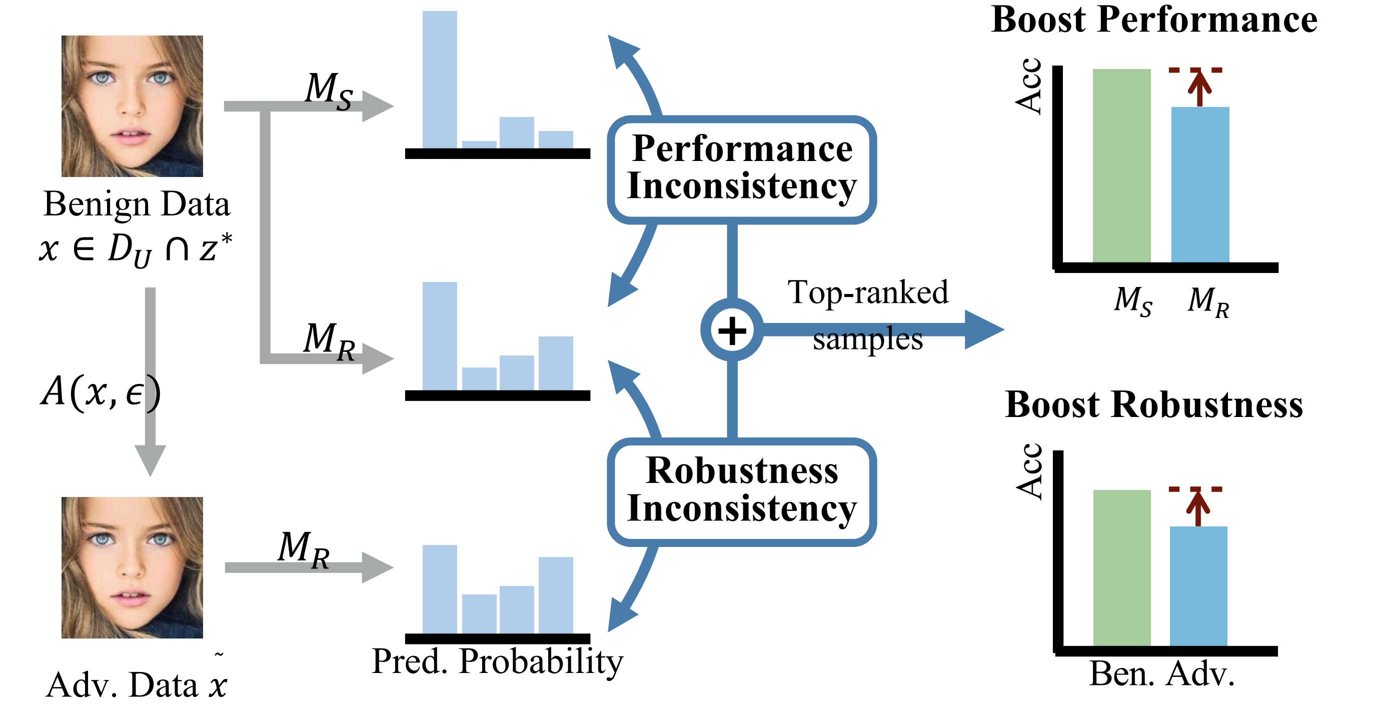
HAM-10000 Skin Lesion Identification (sensitive groups: {Male, Female})						
Methods	Standard F1-score (%)			Robust F1-score (%)		
	Worst (↑)	Disp (↓)	Avg (↑)	Worst (↑)	Disp (↓)	Avg (↑)
Init. AT	37.37±0.76	3.62±0.51	39.18±0.76	15.84±0.22	1.92±0.49	16.80±0.31
G-RAND	36.15±1.37	3.46±0.61	37.88±1.67	16.65±0.36	2.85±0.89	18.07±0.66
MinMax	37.21±1.21	3.59±0.86	39.00±1.46	16.68±0.83	2.17±0.80	17.77±0.72
OPT	35.53±1.45	4.88±1.68	37.98±0.64	17.32±0.38	**1.88±0.38**	18.26±0.39
FairAL	43.65±0.99	3.53±0.77	45.42±0.68	19.64±0.54	2.44±0.81	20.86±0.83
JIN	**44.98±1.41**	**2.96±0.58**	46.46±1.48	**21.95±0.91**	2.28±0.66	**23.09±1.16**

Methods	UTKFace	CINIC-10	HAM-10000
Init. AT	1h 4m 26s	1h 9m 31s	1h 22m 7s
ENT	14s	45s	12s
G-RAND	1m 5s	2m 17s	18s
FairAL	39m 47s	2h 21m 29s	19m 55s
JIN	10m 29s	19m 46s	15m 40s

Our Goal: Effective and Efficient Active Data Selection

Joint Inconsistency (JIN) Data Selection

Concept Figure of our Joint Inconsistency (JIN) Method



$$I_x^{per} = D_{\mathrm{KL}}(p(x, M_S) \,||\, p(x, M_R))$$

$$I_x^{rob} = D_{\mathrm{KL}}(p(x, M_R) \,||\, p(\mathcal{A}(x, \epsilon), M_R))$$

$$I_x = N(I_x^{per}) + N(I_x^{rob})$$

M_s: Auxiliary standard-trained model

<u>Our Algorithm</u>

1. Initialed the robust model
2. Estimate the "worst group"
3. Calculate JIN score on data in that group
4. Select top-ranked samples for labeling
5. Retrain/Fine-tune the model
6. Loop Back to Step #2

- **Effectiveness**: Our method is grounded in fundamental properties of adversarial training.
- **Efficiency**: We conduct selection based on two easily calculable prediction softmax inconsistencies.

Experimental Results (1) — Main Experiments

Methods	UTKFace 4-Race Classification (sensitive groups: {Young, Old})						CINIC-10 Classification (sensitive groups: {CIFAR-10, ImageNet})					
	Standard Accuracy (%)			Robust Accuracy (%)			Standard Accuracy (%)			Robust Accuracy (%)		
	Worst (↑)	Disp (↓)	Avg (↑)	Worst (↑)	Disp (↓)	Avg (↑)	Worst (↑)	Disp (↓)	Avg (↑)	Worst (↑)	Disp (↓)	Avg (↑)
Init. AT	67.58±0.30	5.38±0.25	70.27±0.31	52.98±0.08	7.26±0.31	56.61±0.06	52.53±0.17	12.48±0.21	58.77±0.40	31.29±0.11	10.64±0.23	36.61±0.03
RAND	70.57±0.21	4.32±0.03	72.73±0.21	55.63±0.06	7.71±0.02	59.49±0.07	55.53±0.53	12.14±0.55	61.60±0.61	37.01±0.43	11.43±0.37	42.73±0.60
ENT	74.10±0.79	2.45±0.48	75.33±0.56	56.94±0.64	6.60±0.33	60.25±0.56	56.23±0.52	11.30±0.39	61.88±0.64	36.29±0.40	10.52±0.42	41.55±0.51
CSET	71.44±0.46	3.47±0.52	73.31±0.21	56.55±0.19	6.42±0.49	59.76±0.05	55.28±0.44	12.94±0.51	61.75±0.52	36.73±0.27	12.22±0.52	**42.74±0.39**
BADGE	72.63±0.20	3.53±0.23	74.31±0.13	56.94±0.40	6.07±0.20	59.98±0.50	55.86±0.38	11.96±0.44	61.84±0.37	36.66±0.30	11.04±0.38	42.18±0.29
G-RAND	72.37±0.32	2.15±0.26	73.45±0.23	56.60±0.04	6.07±0.33	59.63±0.13	55.56±0.43	**10.76±0.61**	60.94±0.66	36.71±0.35	10.02±0.41	41.72±0.59
MinMax	71.35±0.24	3.27±0.28	72.98±0.20	56.95±0.22	6.59±0.12	60.25±0.21	55.52±0.49	11.32±0.63	61.22±0.60	36.69±0.46	10.52±0.53	41.95±0.47
OPT	71.99±0.31	2.76±0.23	73.37±0.20	57.09±0.33	6.11±0.19	60.15±0.24	55.78±0.33	10.90±0.37	61.23±0.49	36.90±0.29	9.96±0.36	41.88±0.50
FairAL	74.74±0.31	2.20±0.13	**75.84±0.25**	56.94±0.16	6.64±0.17	**60.47±0.07**	56.35±0.45	10.98±0.44	61.84±0.58	36.25±0.29	10.40±0.33	41.45±0.37
JIN	**75.07±0.53**	**1.35±0.09**	75.74±0.49	**57.39±0.10**	**5.69±0.30**	60.10±0.25	**57.37±0.67**	11.16±0.52	**62.95±0.68**	**37.10±0.45**	**9.84±0.45**	42.02±0.48

- **Surpassing more than 1 standard deviation on most fairness metrics**
- **Limited fairness-accuracy tradeoffs**

Experimental Results (2) — Analyses on UTKFace

Effectiveness of two inconsistency scores

	STD. Acc. (%)		Rob. Acc. (%)	
	Worst (↑)	Avg (↑)	Worst (↑)	Avg (↑)
P	**75.18**	**75.84**	56.53	59.30
R	72.89	74.31	56.89	59.94
P+R	75.07	75.74	**57.39**	**60.10**

(+) Combining the two performs the best!

Experiments on ResNet-18

	STD. Acc. (%)		Rob. Acc. (%)	
	Worst (↑)	Avg (↑)	Worst (↑)	Avg (↑)
Init. AT	64.80±1.79	67.46±1.39	51.48±0.41	56.42±0.23
RAND	70.86±1.46	72.83±1.01	55.40±1.36	59.52±0.81
ENT	73.30±1.07	74.67±0.93	56.03±0.80	60.30±0.40
G-RAND	72.71±0.78	73.47±0.54	56.69±0.67	59.71±0.36
FairAL	74.28±0.60	75.41±0.35	56.80±0.46	**60.68±0.35**
JIN	**75.38±0.66**	**75.58±0.61**	**57.75±0.69**	60.42±0.25

(+) General under Various Network Architectures!

Methods selecting from only the worst-group

	STD. Acc. (%)		Rob. Acc. (%)	
	Worst (↑)	Avg (↑)	Worst (↑)	Avg (↑)
ENT	74.10±0.79	75.33±0.56	56.94±0.64	**60.25±0.56**
G-ENT	68.14±0.62	70.56±0.44	54.89±0.32	58.85±0.37
JIN	**75.07±0.53**	**75.74±0.49**	**57.39±0.10**	60.10±0.25

(+) Without our method, directly modifying conventional AL methods yields poor results!

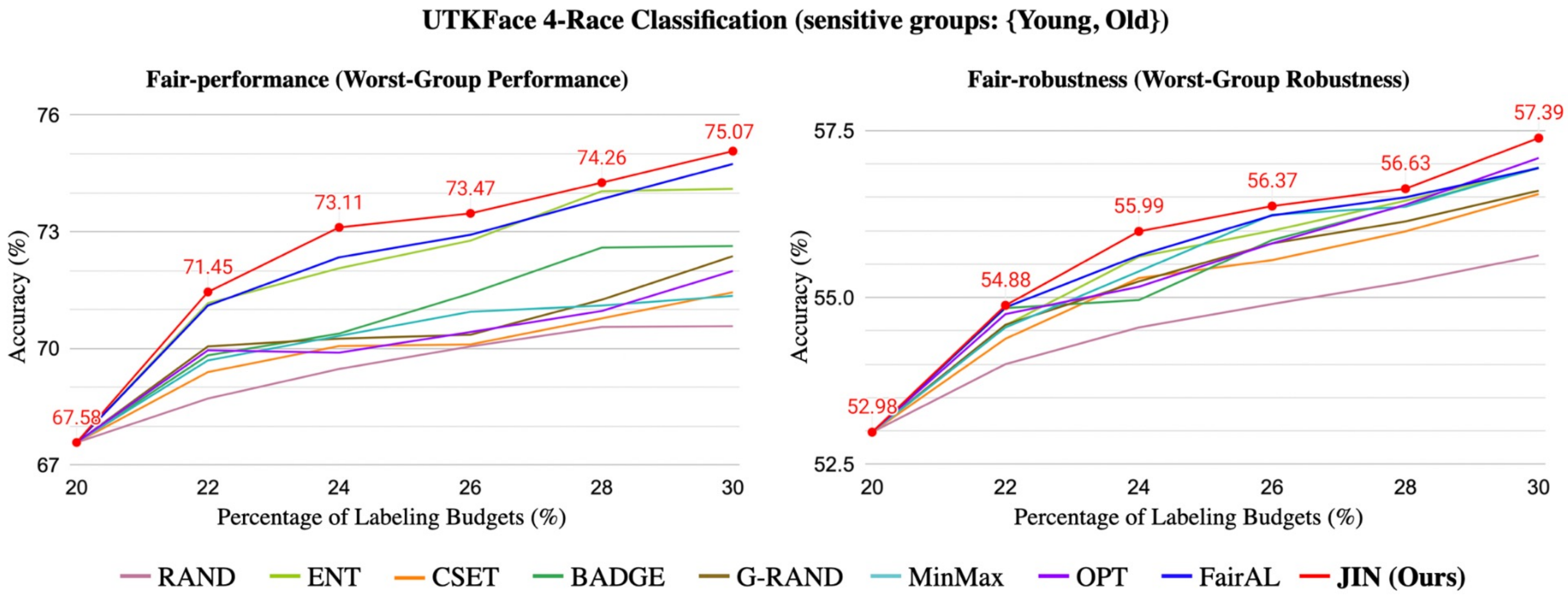
Gender Classification Tasks {4 Races}

	STD. Acc. (%)		Rob. Acc. (%)	
	Worst (↑)	Avg (↑)	Worst (↑)	Avg (↑)
Init. AT	77.74±0.66	81.03±0.33	67.61±0.21	70.84±0.36
RAND	78.57±0.31	82.34±0.10	69.14±0.30	72.34±0.14
ENT	80.70±0.59	84.01±0.10	69.79±0.14	72.67±0.21
G-RAND	81.08±0.22	82.95±0.31	70.38±0.23	**73.39±0.29**
FairAL	80.41±0.60	83.78±0.32	69.78±0.30	72.56±0.16
JIN	**82.77±0.27**	**84.96±0.25**	**70.56±0.13**	73.11±0.11

(+) Support Multiple Sensitive Groups!

Our Contributions

- **Novel framework**: First practice for annotation-expensive and safety-critical apps.
- **Inconsistency-based strategy**: Elegant, efficient, and effective
- **Experimental results**: SOTA results on three different tasks



Open Research Problems

- Metric
  - Social aspects, who will define them?
  - generate metric (semi-)automatically?
- Explore connection to Differential Privacy
- Connection to Economics literature/problems
- Trade-offs of fairness, privacy, accuracy, and robustness