Exploring Fairness and Socio-demographic Bias in Machine Learning

Theodora Chaspari, Ph.D.
HUman Bio-Behavioral Signals (HUBBS) Lab
Computer Science & Engineering, Texas A&M University

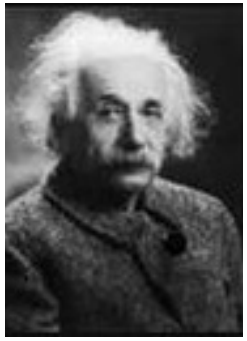
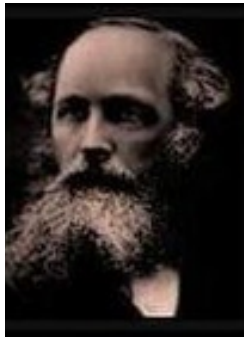
TAMU Data Science Institute
4/23/2021

Close your eyes and picture a shoe



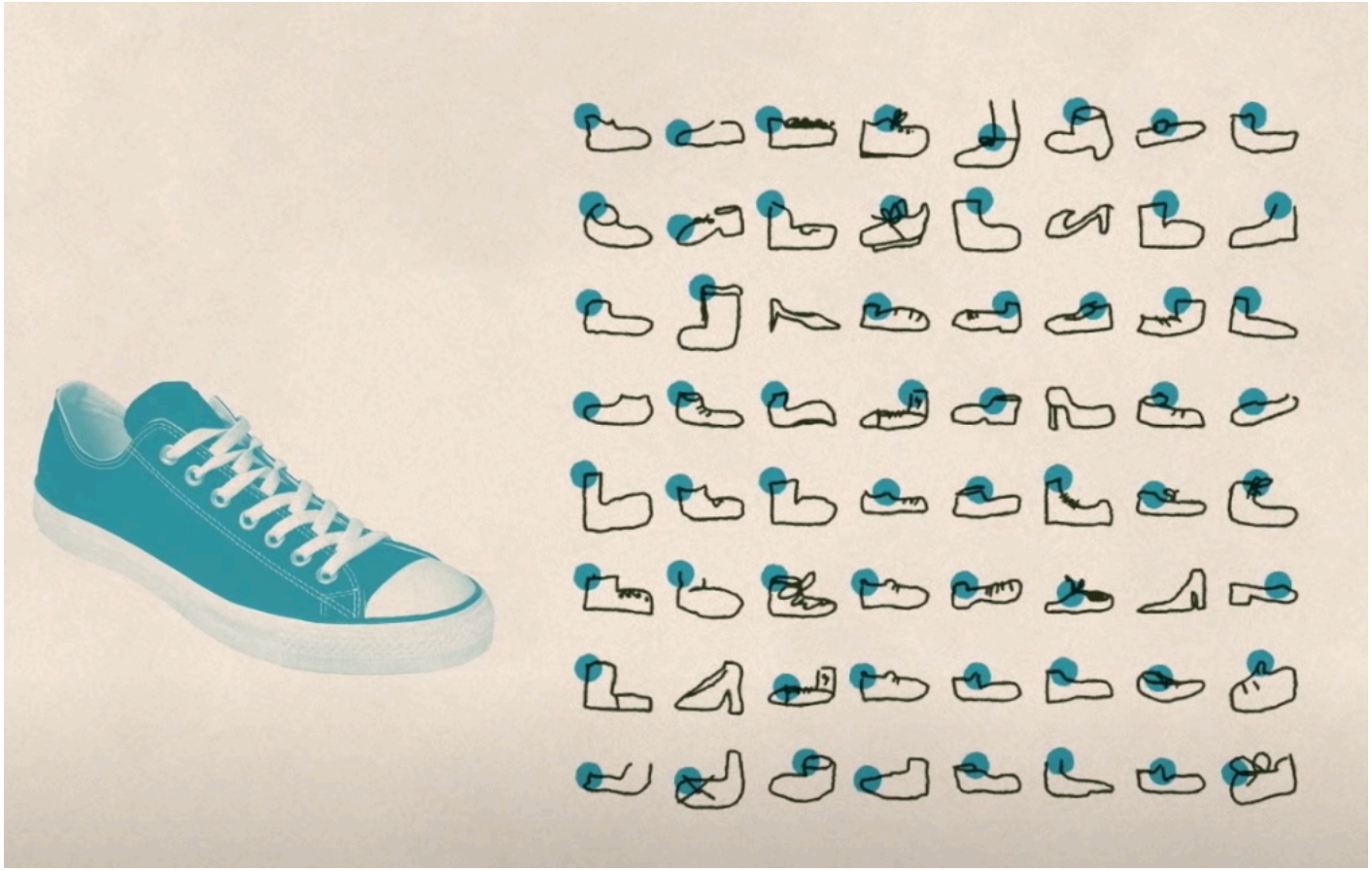

https://www.youtube.com/watch?v=59bMh59JQDo

How about a physicist?

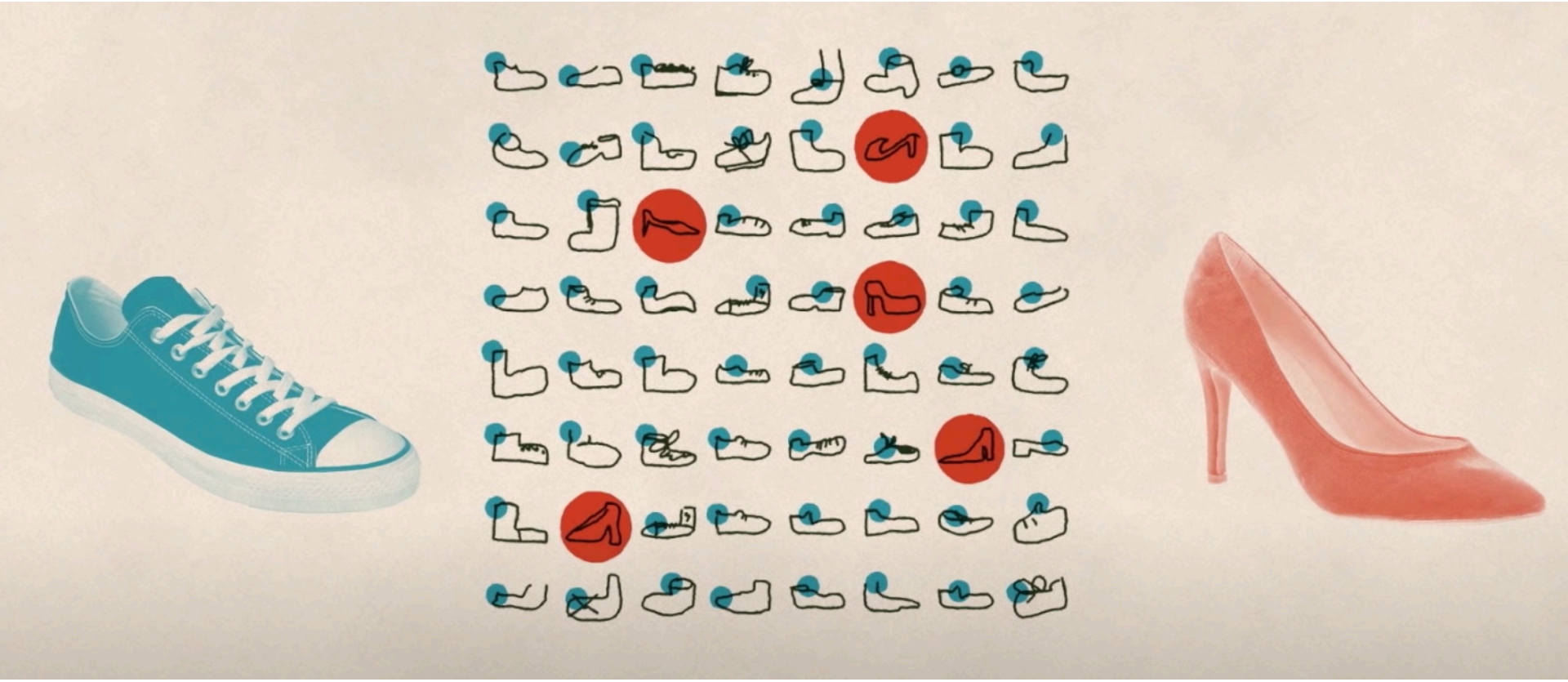
https://www.youtube.com/watch?v=59bMh59JQDo

Machine learning



https://www.youtube.com/watch?v=59bMh59JQDo

Machine learning



https://www.youtube.com/watch?v=59bMh59JQDo

TECH

The problem with AI? Study says it's too white and male, calls for more women, minorities

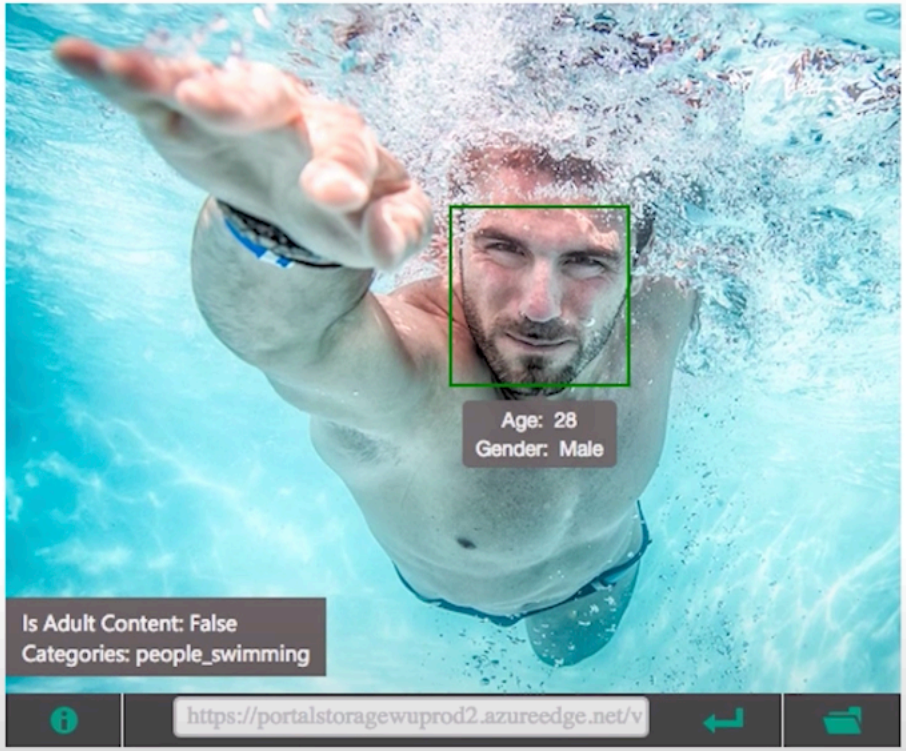
Jessica Guynn USA TODAY

Published 8:00 p.m. ET Apr. 16, 2019 | **Updated 11:37 a.m. ET Apr. 17, 2019**

Two out of four leading face recognition platforms do not reliably detect African American users.

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. AI Now.
https://www.youtube.com/watch?v=TWWsW1w-BVo&t=45s

Evidence of racial bias in computer vision



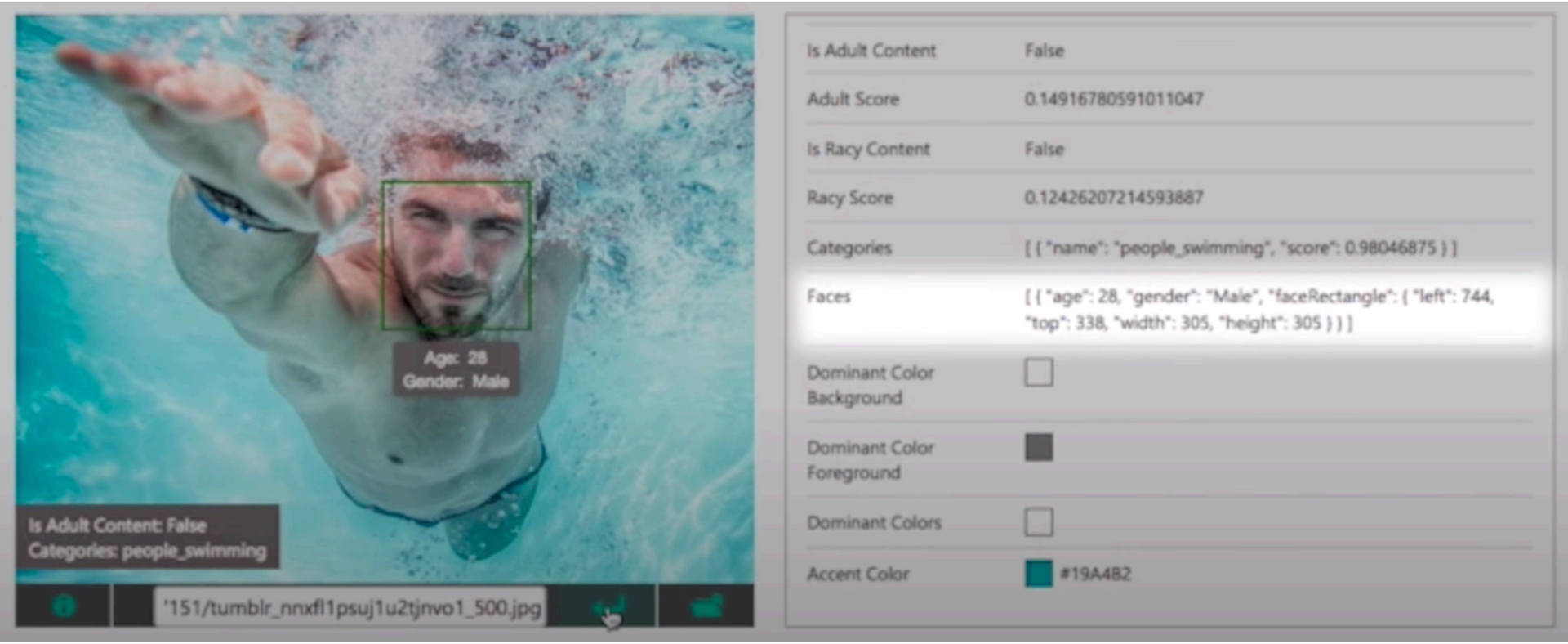
Features:

Feature Name	Value
Description	{ "type": 0, "captions": [{ "text": "a man swimming in a pool of water", "confidence": 0.7850108693093019 }] }
Tags	[{ "name": "water", "confidence": 0.9996442794799805 }, { "name": "sport", "confidence": 0.9504992365837097 }, { "name": "swimming", "confidence": 0.9062818288803101, "hint": "sport" }, { "name": "pool", "confidence": 0.8787588477134705 }, { "name": "water sport", "confidence": 0.631849467754364, "hint": "sport" }]
Image Format	jpeg
Image Dimensions	1500 x 1155
Clip Art Type	0 Non-clipart
Line Drawing Type	0 Non-LineDrawing
Black & White Image	False

Microsoft Computer Vision API

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. AI Now.
Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

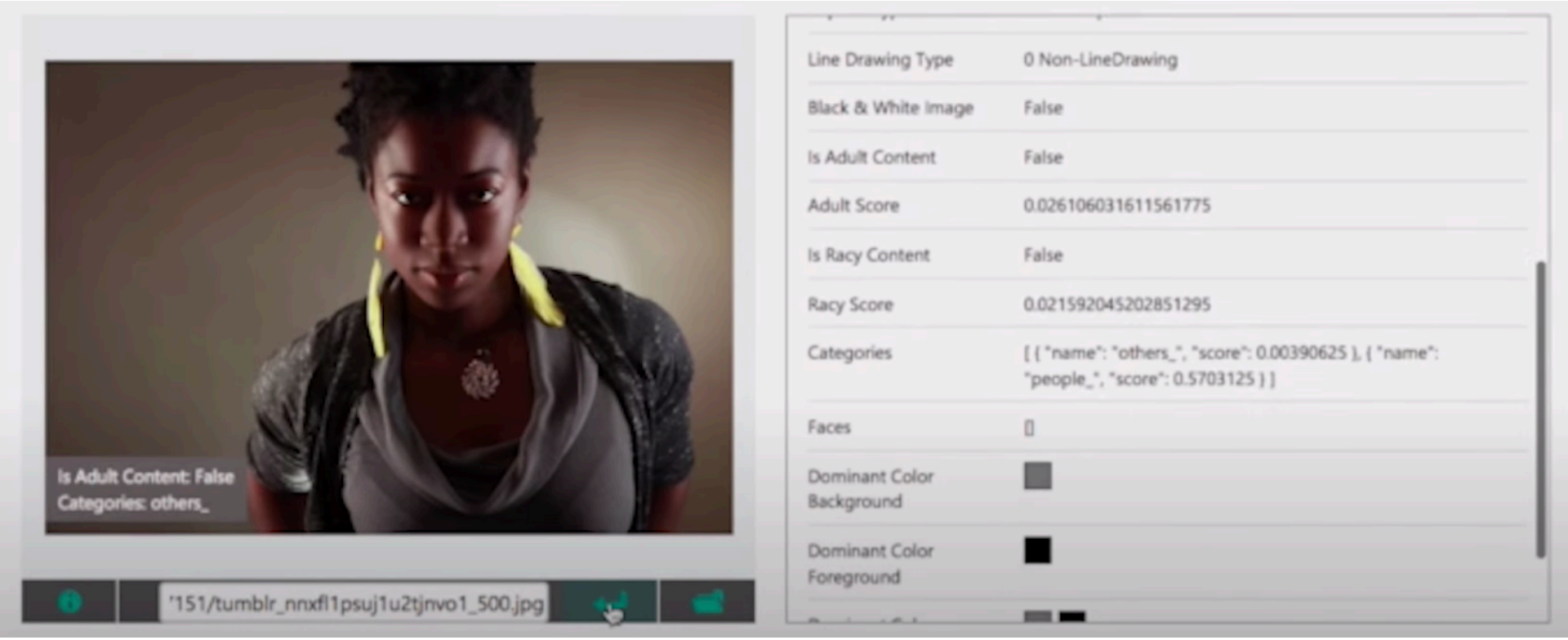
Evidence of racial bias in computer vision



Microsoft Computer Vision API

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. AI Now.
Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

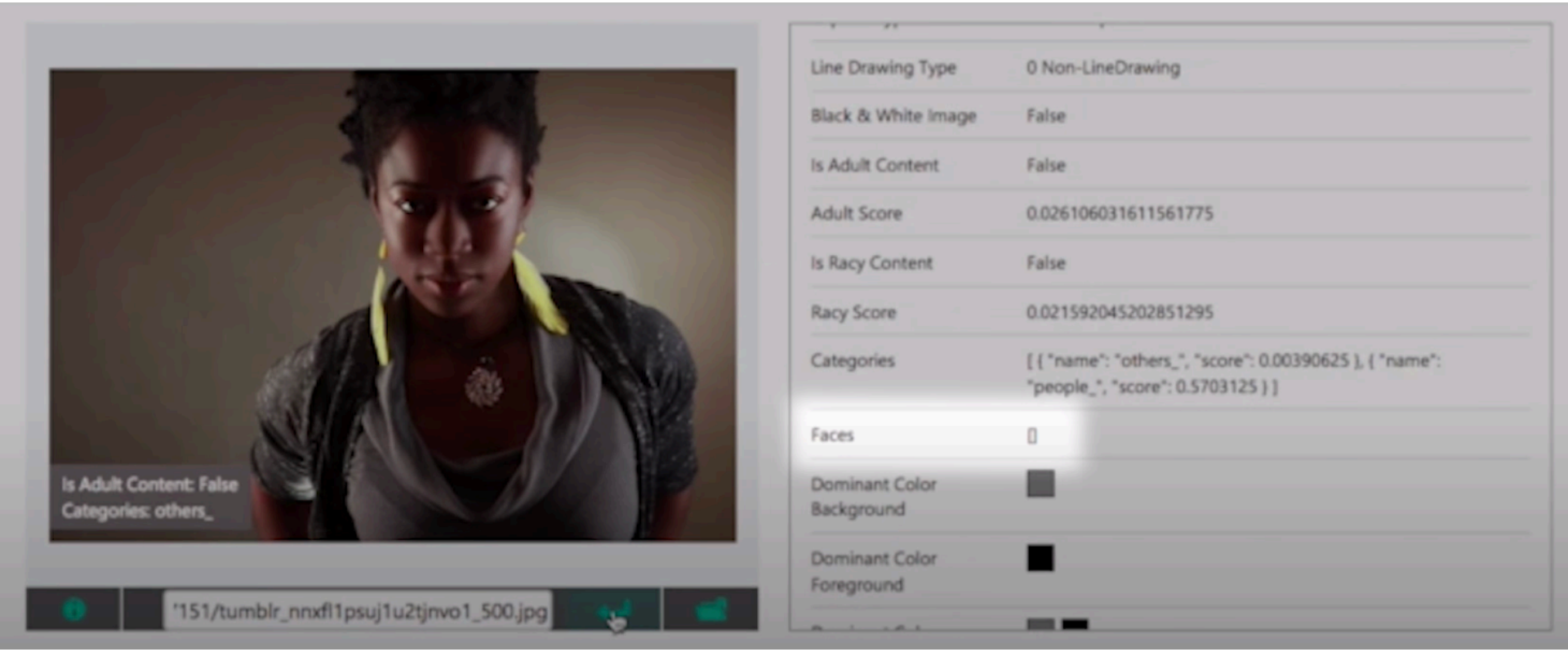
Evidence of racial bias in computer vision



Microsoft Computer Vision API

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. AI Now.
Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

Evidence of racial bias in computer vision



Microsoft Computer Vision API

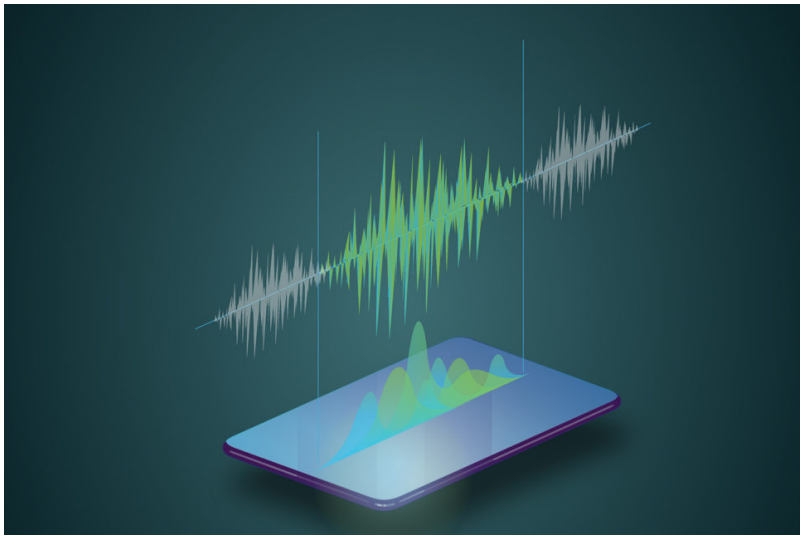
West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. AI Now.
Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

MARCH 23, 2020

Stanford researchers find that automated speech recognition is more likely to misinterpret black speakers

The disparity likely occurs because such technologies are based on machine learning systems that rely heavily on databases of English as spoken by white Americans.

Five leading speech recognition programs make twice as many errors with African American speakers as with Whites



Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684-7689.

Bias related to socioeconomic status (SES)

Visual disparities between communities of different socioeconomic (SES) status

Tutorial Overview

- Sources of sociodemographic bias in machine learning (ML)
- Examples of studies examining sociodemographic bias in ML
  - Health - Electronic Health Records
  - Family well-being - Speech, language, physiology
- Approaches for mitigating sociodemographic bias in ML
  - Adversarial learning
  - Fairness regularization
  - Explainability

Tutorial Overview

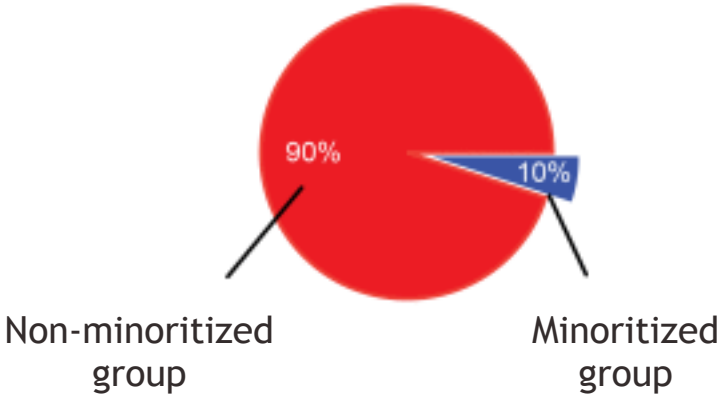
- Sources of sociodemographic bias in machine learning (ML)
- Examples of studies examining sociodemographic bias in ML
  - Health - Electronic Health Records
  - Family well-being - Speech, language, physiology
- Approaches for mitigating sociodemographic bias in ML
  - Adversarial learning
  - Fairness regularization
  - Explainability

Sources of bias in machine learning

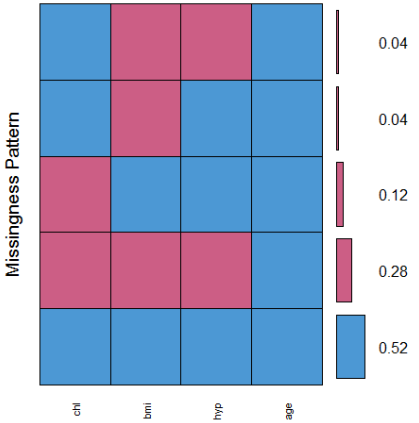
Bias in training data

- Minority bias: minoritized groups might have insufficient number of samples
- Missing data bias: minoritized groups may have missing data in a non-random fashion (e.g., lower quality sensor devices)
- Confounding factors: socio-demographic factors influencing both input and output variables (e.g., gender influences both resting heart rate and heart disease risk at early age)

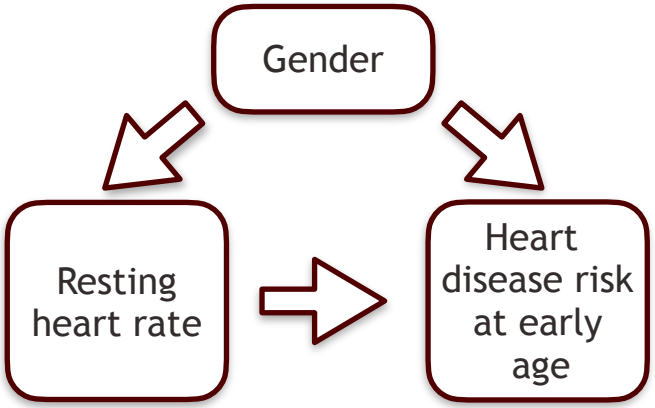
Minority bias



Missing data bias



Confounding factors



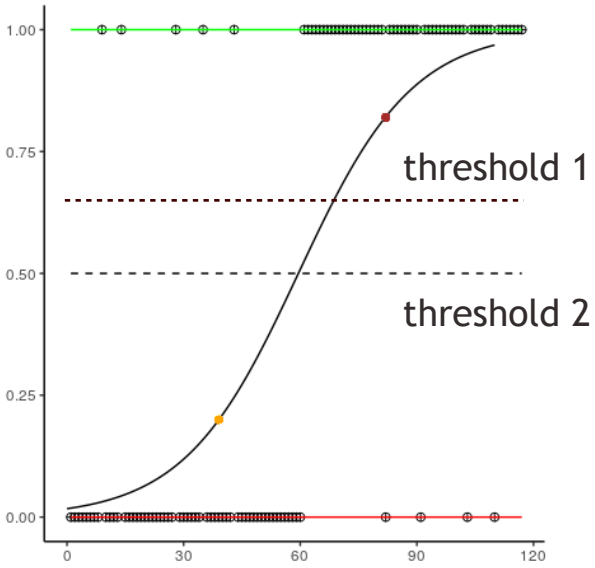
Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872.

Sources of bias in machine learning

Bias in model design

- Label bias: the same outcome might not mean the same for all individuals
- Cohort bias: considering traditional groups (e.g., male/female) without considering other protected groups (e.g., LGTBQ) and levels of granularity
- Proprietary algorithms, making it difficult to dissect them

Label bias



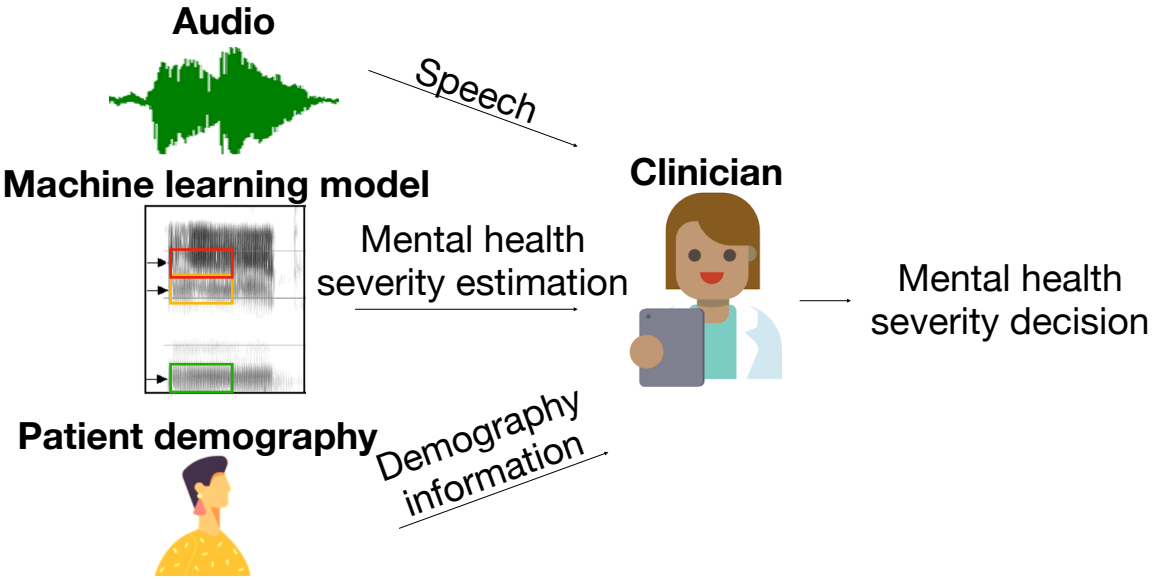
Cohort bias

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872.

Sources of bias in machine learning

Bias in interaction with experts

- Automation bias: experts are unaware that a model is underperforming for a certain group
- Feedback loops: If the clinician accepts incorrect model outputs, the mistake is propagated next time the model is trained
- Dismissal bias: Desensitization to alerts that are systematically incorrect for a specific group



Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872.

Sources of bias in machine learning

Bias in interaction with users

- Privilege bias: ML models might be unavailable in places where specific groups receive care (e.g., devices with low computational resources, poor internet connectivity)
- Informed mistrust: Users might believe that a model is biased against them due to historical exploitation practices

Privilege bias

Informed mistrust

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872.

Tutorial Overview

- Sources of sociodemographic bias in machine learning (ML)
- Examples of studies examining sociodemographic bias in ML
  - Health - Electronic Health Records
  - Family well-being - Speech, language, physiology
- Approaches for mitigating sociodemographic bias in ML
  - Adversarial learning
  - Fairness regularization
  - Explainability

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

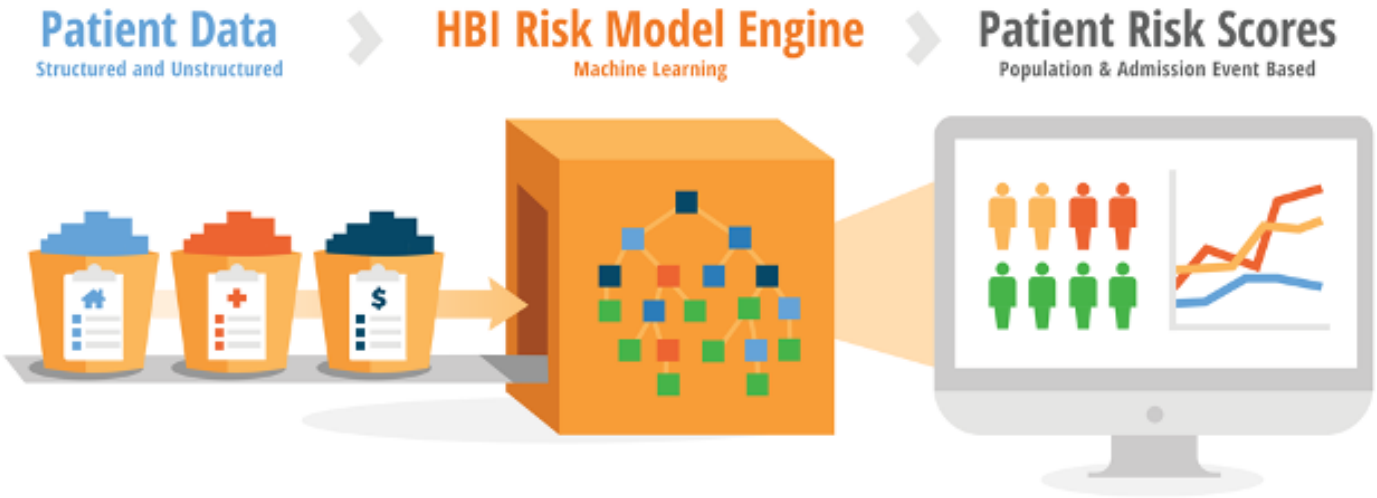
Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Dissecting racial bias in a health management algorithm

- Commercial risk-prediction algorithm applied to ~ 200 million people in the U.S.
  - Improving patient care by providing additional resources
  - Including greater attention from trained providers
  - Contributing toward ensuring well-coordinated care
- Primary care patients enrolled in risk-based contracts from 2013 to 2015
- 6,079 Black patients and 43,539 White patients
- 71.2% in commercial insurance and 28.8% in Medicare
- 50.9 years old on average and 63% female



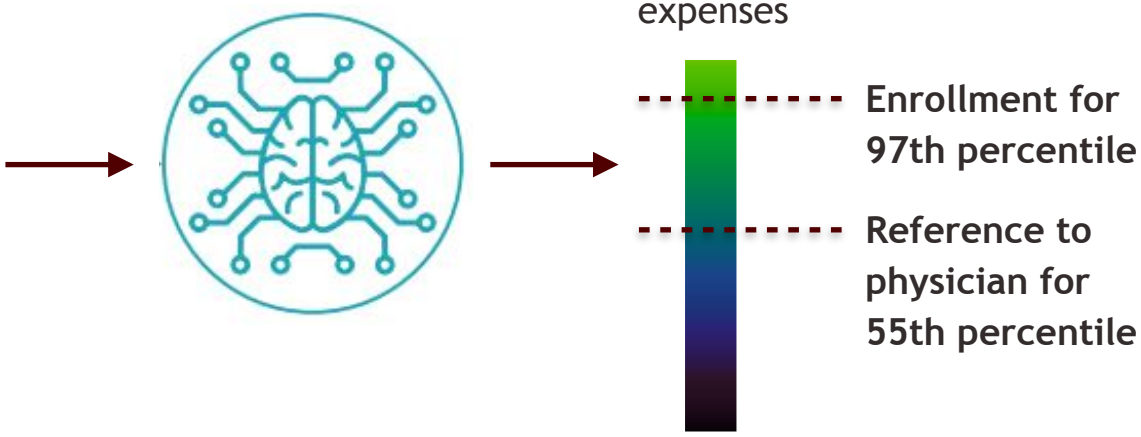
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Dissecting racial bias in a health management algorithm

Algorithmic outline

Input features at time (t-1)

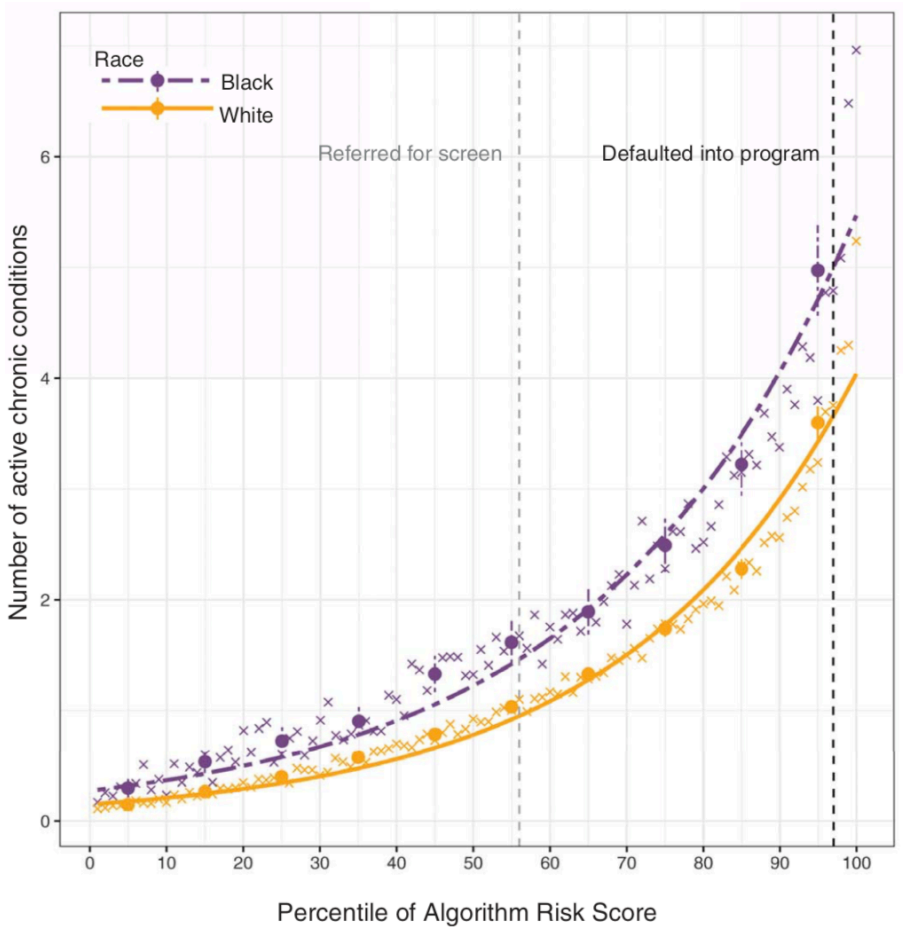
1. Demographics, excluding race (e.g., biological sex, age)
2. Insurance type
3. IDC-9 codes (International Statistical Classification of Diseases)
4. Prescribed medications
5. Medical service encounters (e.g., surgical, radiology),
6. Billed amounts, categorized by type (e.g., outpatient specialists, dialysis)



Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Dissecting racial bias in a health management algorithm

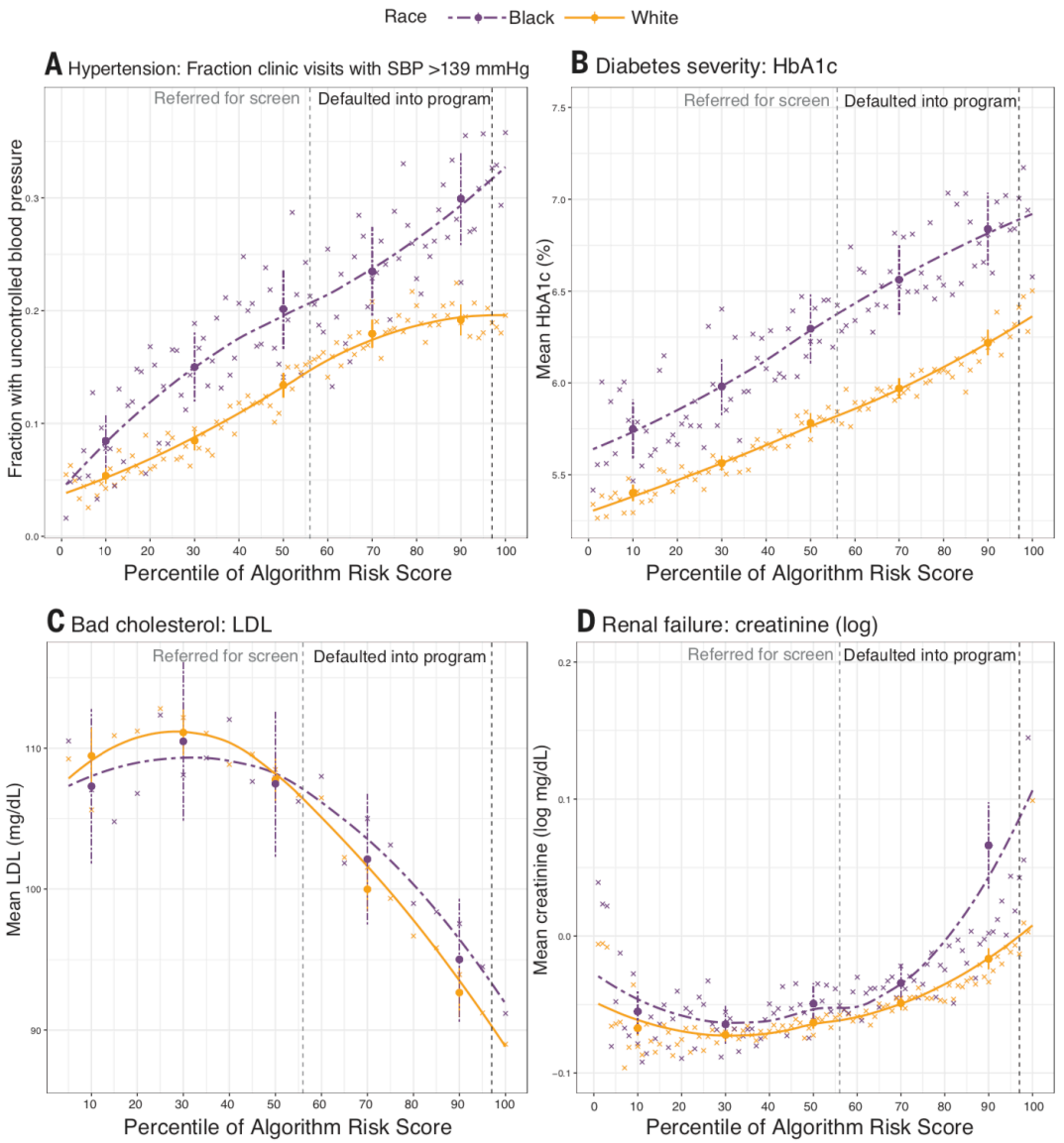
- Mean number of chronic conditions by race, plotted against algorithmic score
- For the same level of algorithm-predicted risk, Black patients found to depict significantly more illness burden compared to White patients



Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Dissecting racial bias in a health management algorithm

- For the same level of algorithm-predicted risk, Black patients found to depict significantly more illness burden across significant health markers compared to White patients

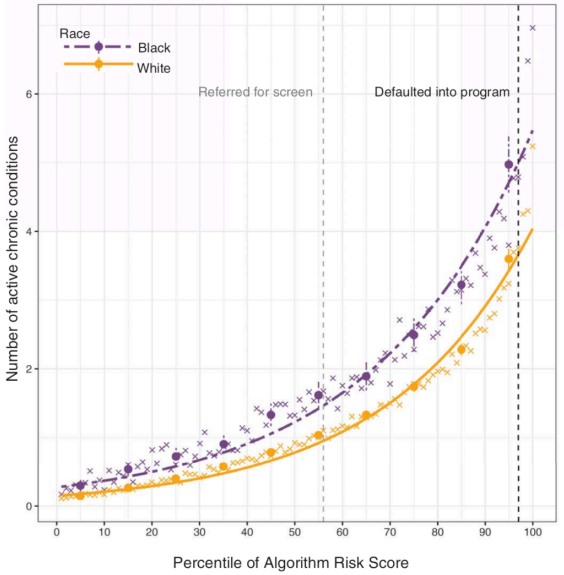


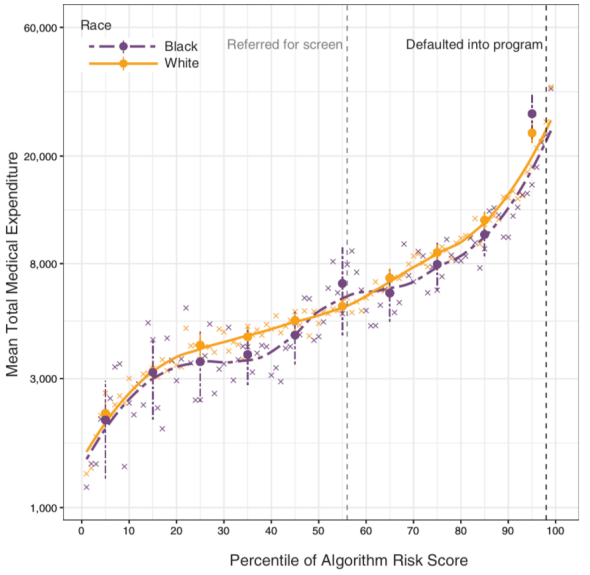
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Dissecting racial bias in a health management algorithm

- For the same level of algorithm-predicted risk
  - Blacks depict significantly more illness burden compared to Whites
  - Blacks and Whites have (roughly) the same costs the following year

- Substantial disparities in health burden
- Little disparity in costs



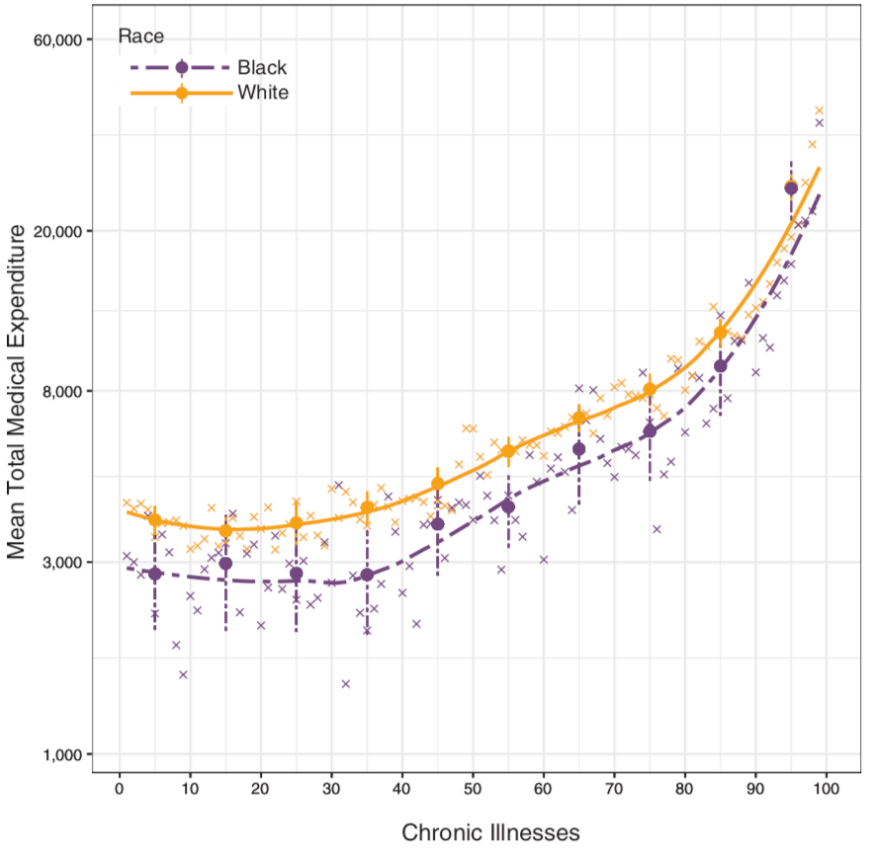


Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Dissecting racial bias in a health management algorithm

- At a given level of health, Black patients generate lower costs than White patients

- Potentially the driving force behind this algorithmic disparity is that Black patients generate less medical expenses, therefore they are considered as lower risk by the algorithm



Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Dissecting racial bias in a health management algorithm

Examining algorithmic risk predictions with respect to label choice

Input features at time (t-1)

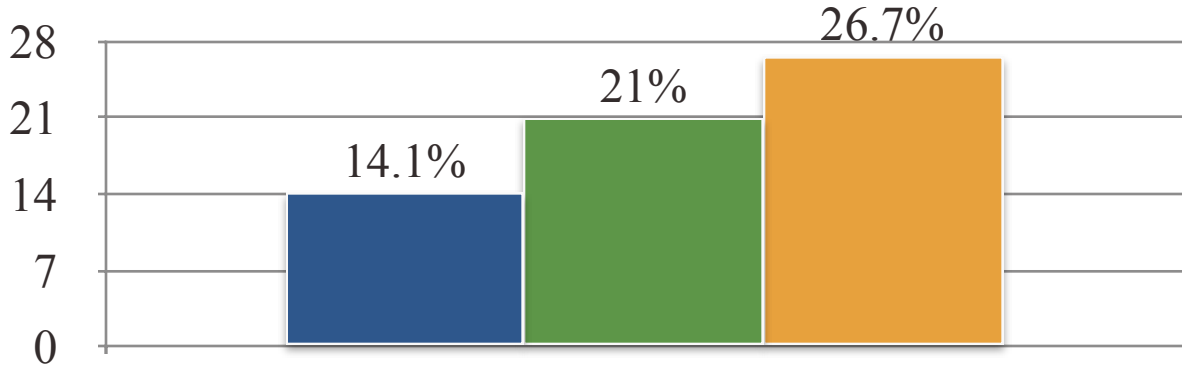
1. Demographics, excluding race (e.g., biological sex, age)
2. Insurance type
3. IDC-9 codes (International Statistical Classification of Diseases)
4. Prescribed medications
5. Medical service encounters (e.g., surgical, radiology),
6. Billed amounts, categorized by type (e.g., outpatient specialists, dialysis)

Output 1:
Total medical costs

Output 2:
Avoidable medical costs

Output 3:
Active chronic conditions

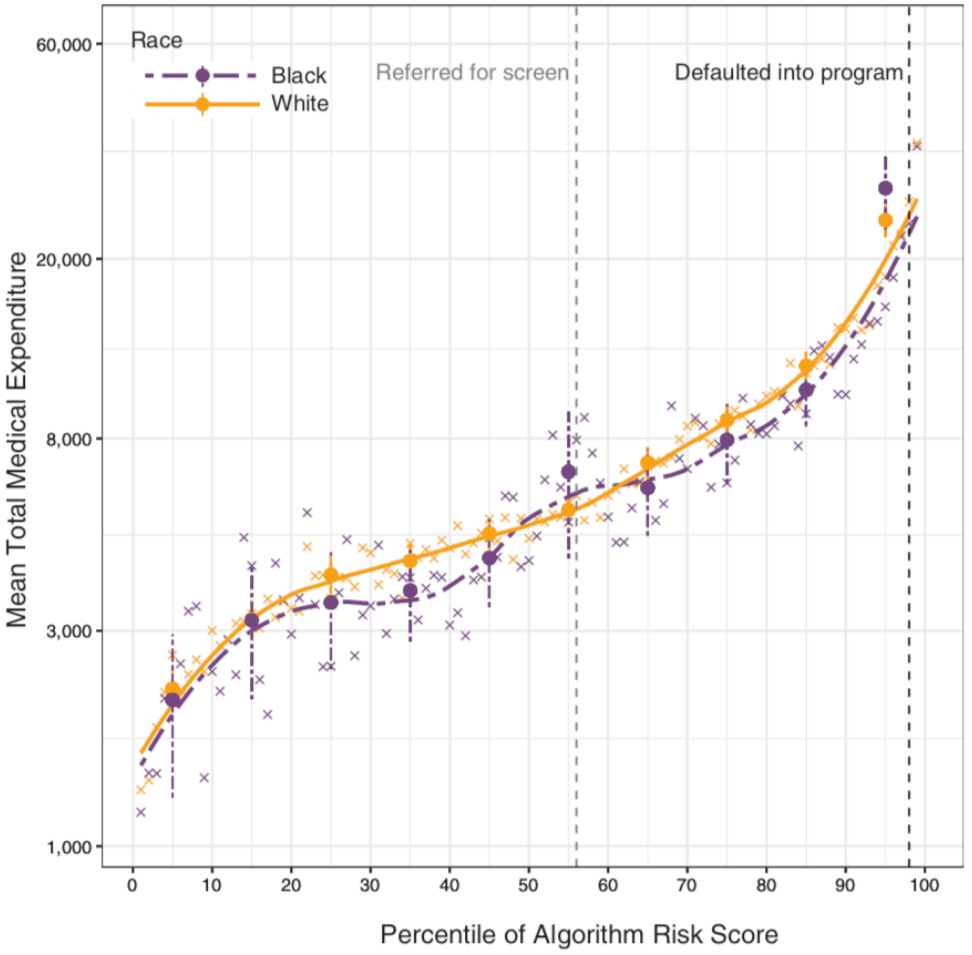
Percentage of Black patients in group with highest predicted risk



Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Dissecting racial bias in a health management algorithm

- At every level of algorithm-predicted risk, Blacks and Whites have (roughly) the same costs the following year
- we find substantial disparities in health conditional on risk but little disparity in costs.



Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Tutorial Overview

- Sources of sociodemographic bias in machine learning (ML)
- Examples of studies examining sociodemographic bias in ML
  - Health - Electronic Health Records
  - Family well-being - Speech, language, physiology
- Approaches for mitigating sociodemographic bias in ML
  - Adversarial learning
  - Fairness regularization
  - Explainability

Demographic bias in ambulatory-based machine learning models of interpersonal conflict

self-reports (8)
- ✓ mood
- ✓ quality of interactions

context and interaction (11)
- ✓ GPS
- ✓ activity count
- ✓ body temperature
- ✓ alcohol/caffeine/drugs

electrodermal activity (2)
- ✓ skin conductance level
- ✓ skin conductance response

THE USC COUPLE MOBILE SENSING PROJECT

http://homedata.github.io/

language use
- ✓ linguistic constructs (25)
- ✓ psychological factors (32)
- ✓ personal concern (7)
- ✓ paralinguistic (3)

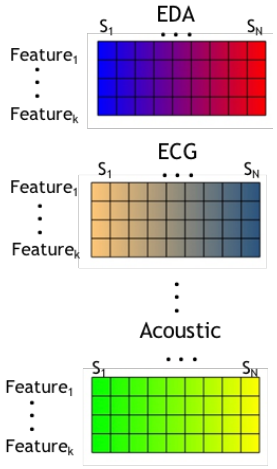
electrocardiogram (2)
- ✓ heart rate
- ✓ heart rate variability

physiological synchrony (2)
- ✓ joint sparse representation
- ✓ multiple time scales

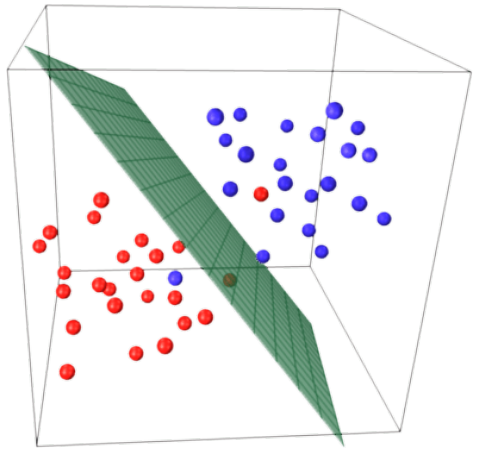
acoustic analysis (8)
- ✓ pitch (F0)
- ✓ intensity

Demographic bias in ambulatory-based machine learning models of interpersonal conflict

Input:
122 acoustic, linguistic, physiological features



Support vector machine classifier



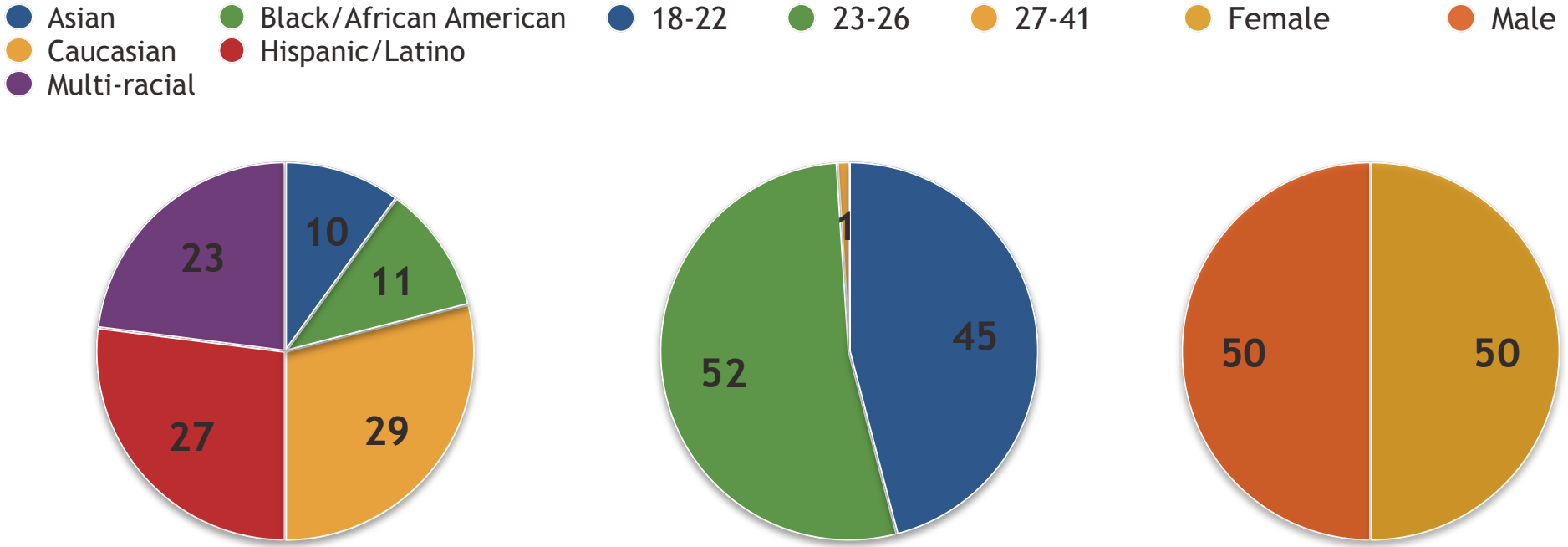
Output:
Conflict/Non-Conflict

- 50 couples, 1438 samples (548 conflict)
- 5-fold stratified cross-validation
- 90.7% precision for non-conflict, 19.5% precision of conflict, 55.15% balanced accuracy

Gujral, Chaspari, et al., Proc. *ACM ICMI*, 2018 (18.5% acceptance)
Gupta, Gujral, Chaspari, et al., *ACM Transactions on Internet Technology*, 2020

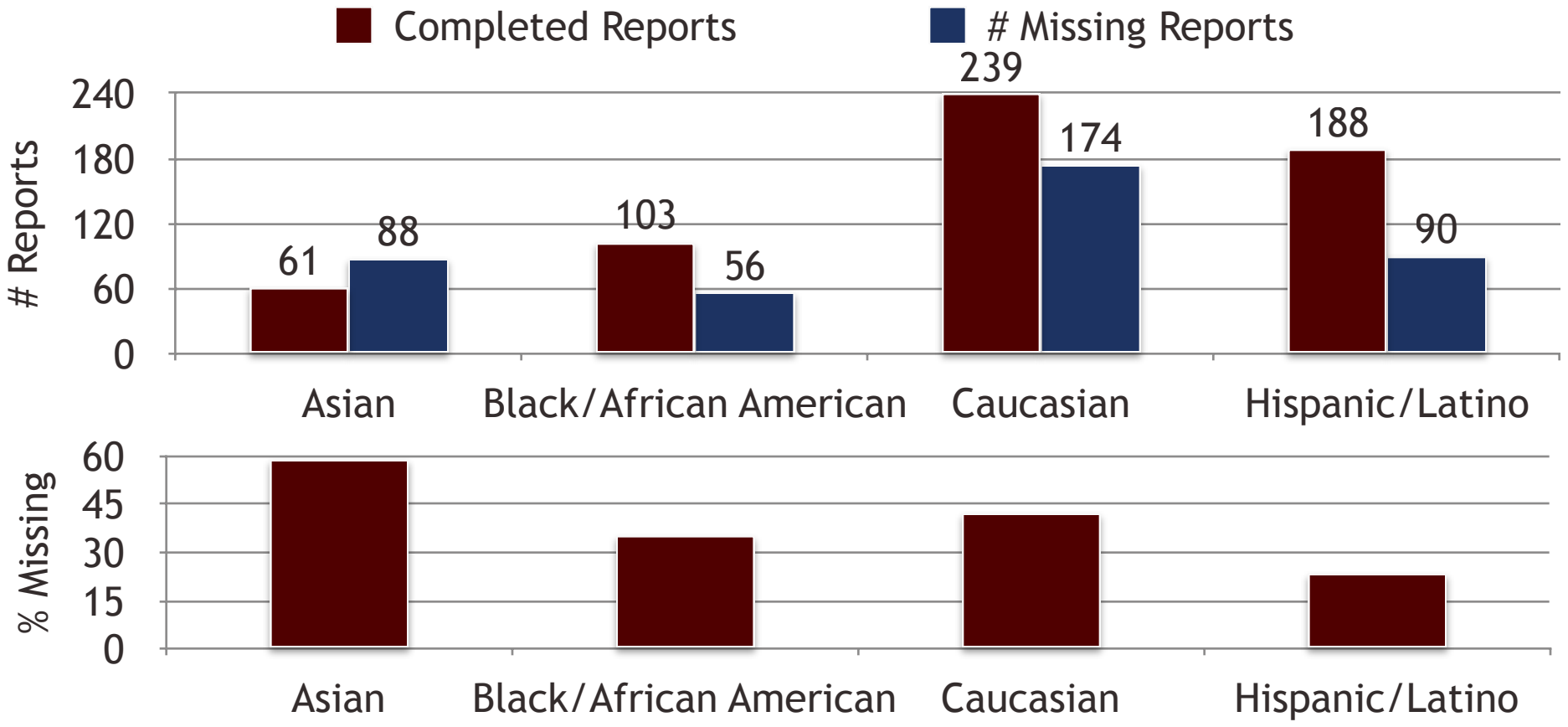
Demographic bias in ambulatory-based machine learning models of interpersonal conflict

Participants' demographic distribution

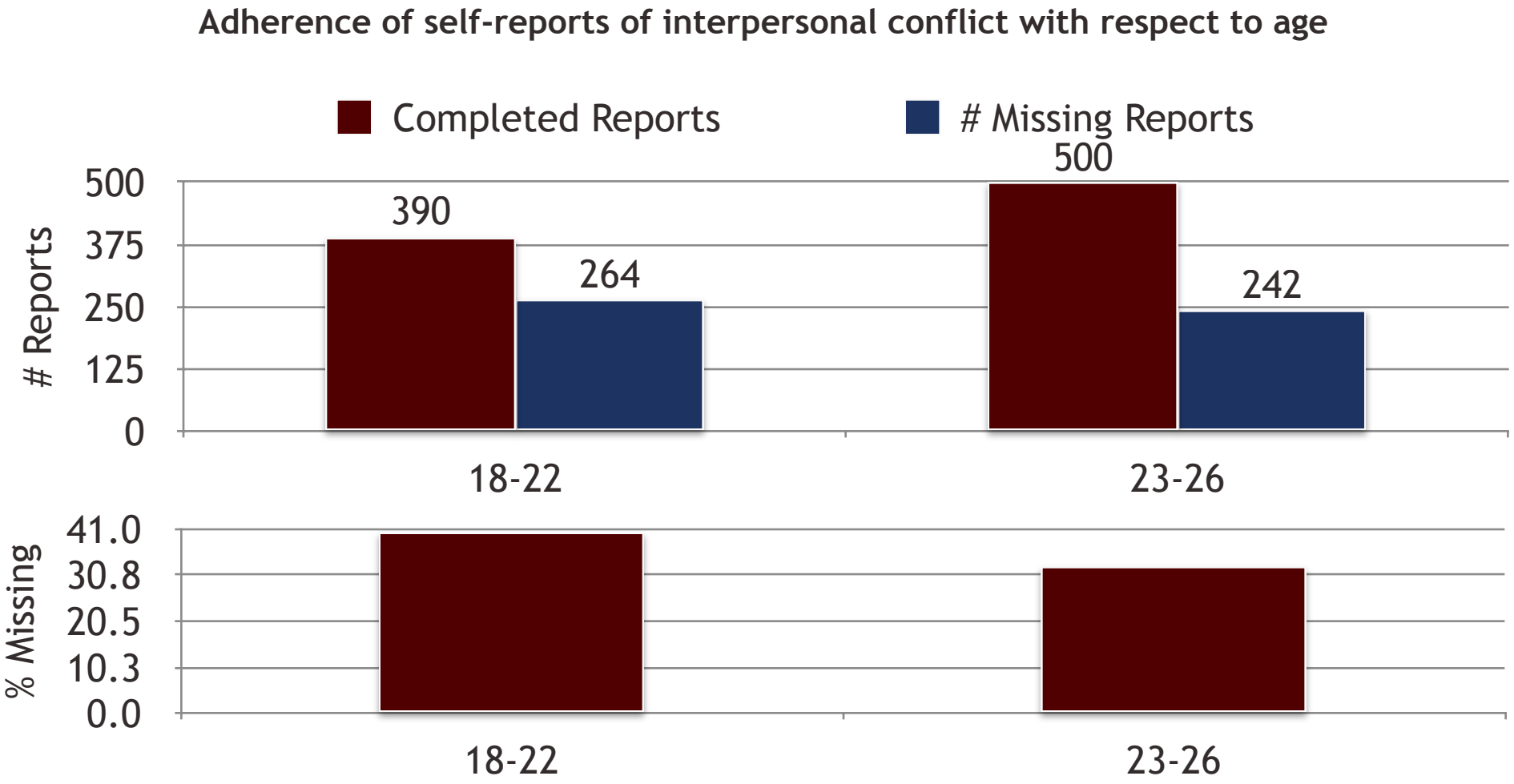


Demographic bias in ambulatory-based machine learning models of interpersonal conflict

Adherence of self-reports of interpersonal conflict with respect to race

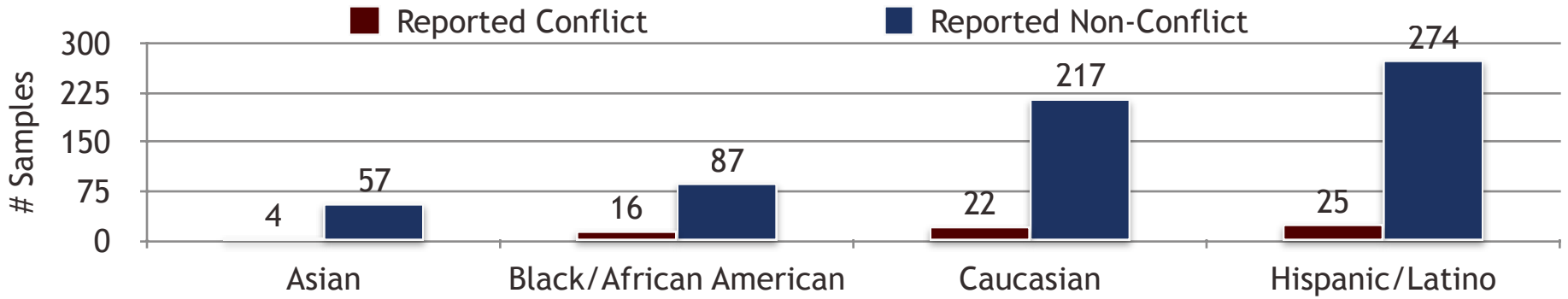


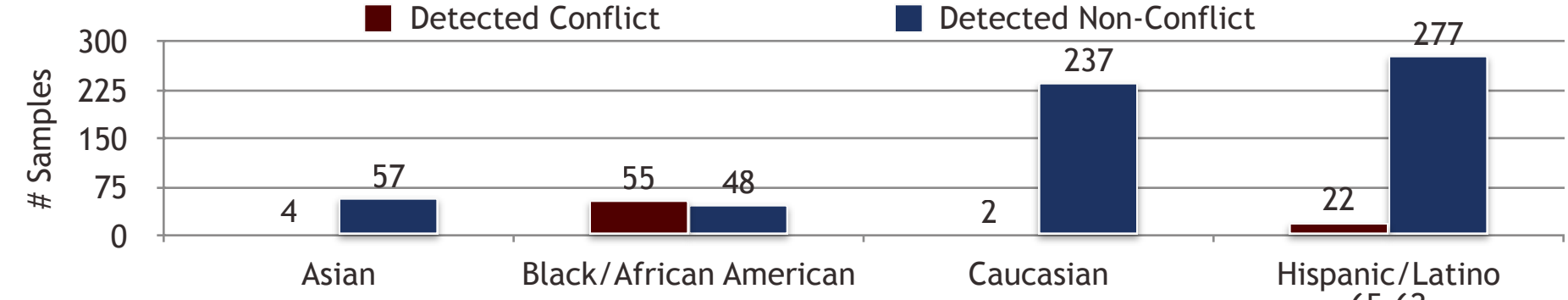
Demographic bias in ambulatory-based machine learning models of interpersonal conflict

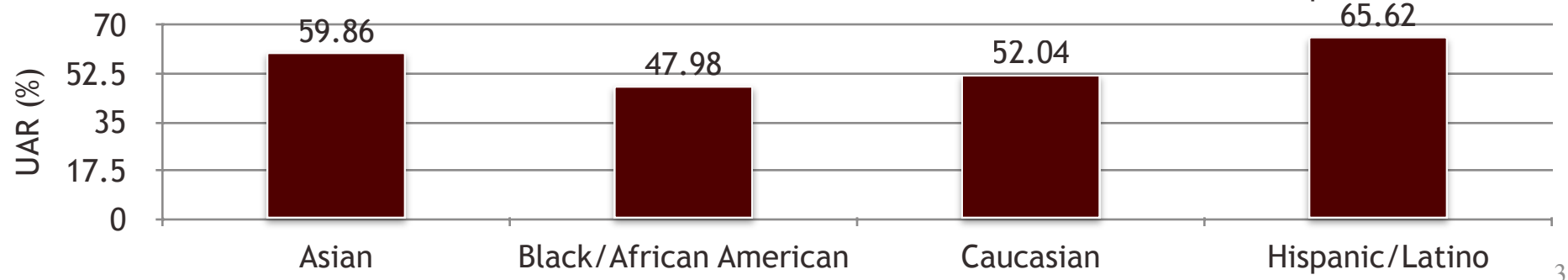


Demographic bias in ambulatory-based machine learning models of interpersonal conflict

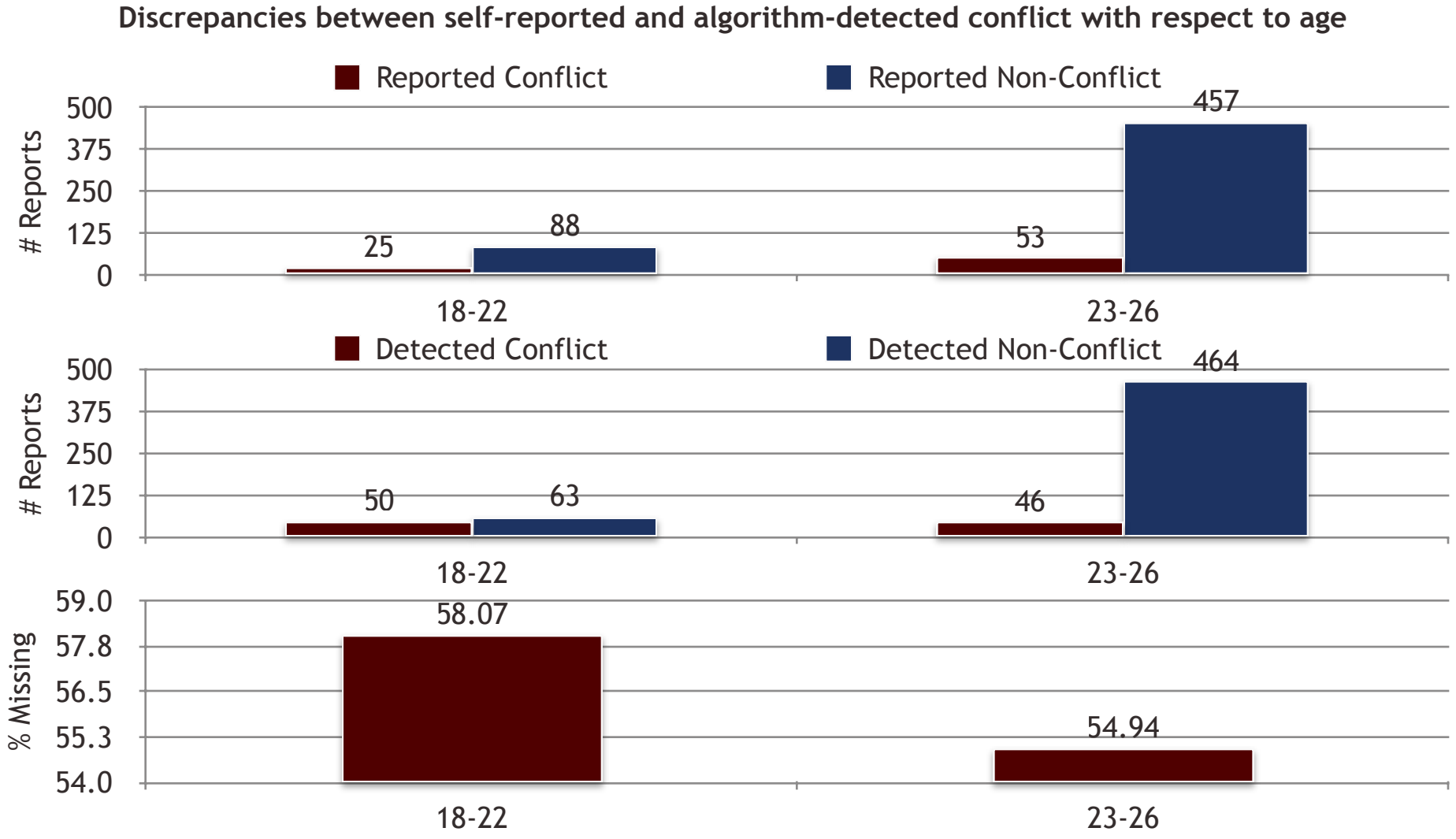
Discrepancies between self-reported and algorithm-detected conflict with respect to race







Demographic bias in ambulatory-based machine learning models of interpersonal conflict



Tutorial Overview

- Sources of sociodemographic bias in machine learning (ML)
- Examples of studies examining sociodemographic bias in ML
  - Health - Electronic Health Records
  - Family well-being - Speech, language, physiology
- Approaches for mitigating sociodemographic bias in ML
  - Adversarial learning
  - Fairness regularization
  - Explainability

Tutorial Overview

- Sources of sociodemographic bias in machine learning (ML)
- Examples of studies examining sociodemographic bias in ML
  - Health - Electronic Health Records
  - Family well-being - Speech, language, physiology
- Approaches for mitigating sociodemographic bias in ML
  - Adversarial learning
  - Fairness regularization
  - Explainability

User anonymization through adversarial learning

Overarching research questions

- Can we characterize user re-identification risk in human behavior recognition models?
- Can we learn anonymized signal transformations that preserve behavioral information?

Case study

- Anonymized models of facial emotion recognition

Challenges

- Necessary to capture the subtlety of emotional expression
- Images captured in close proximity to user's face
  - Iterative adversarial learning with alternate training between minimizing emotion classification cost and maximizing user classification cost

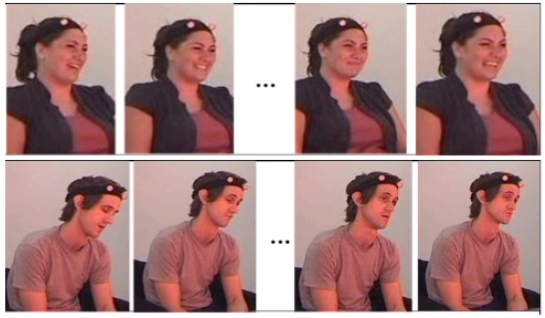
Arora & Chaspari, ACM ICMI 2018
Narula & Chaspari, ACM ICMI, 2020

User anonymization through adversarial learning

Japanese Female Facial Expression (JAFFE)
- 213 images
- 7 emotions

YALE Face Dataset
- 60 images
- 4 emotions

IEMOCAP
- 32,902 images
- 4 emotions

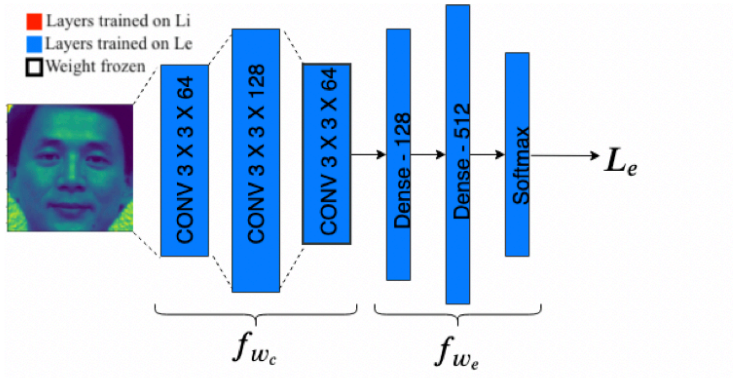


- Data augmentation through rotation, horizontal flip, and noise for JAFFE and YALE

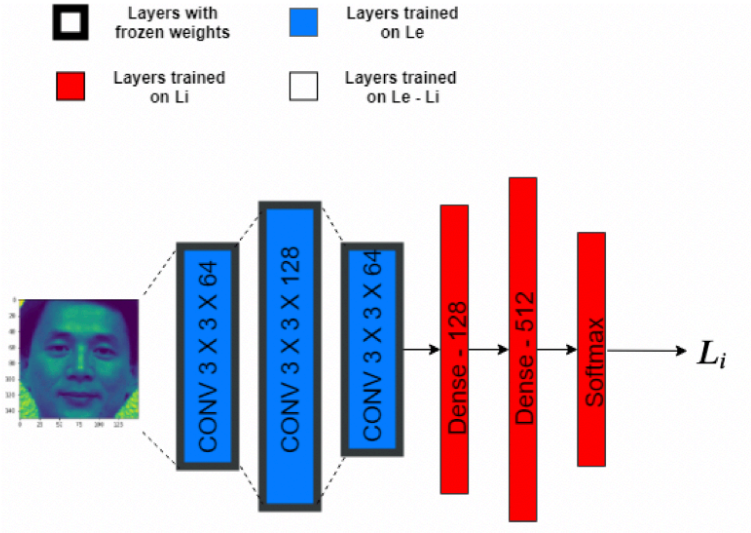
User anonymization through adversarial learning

RQ1: Can we characterize user re-identification risk?

- **Emotion:** CNN trained only on emotion classification



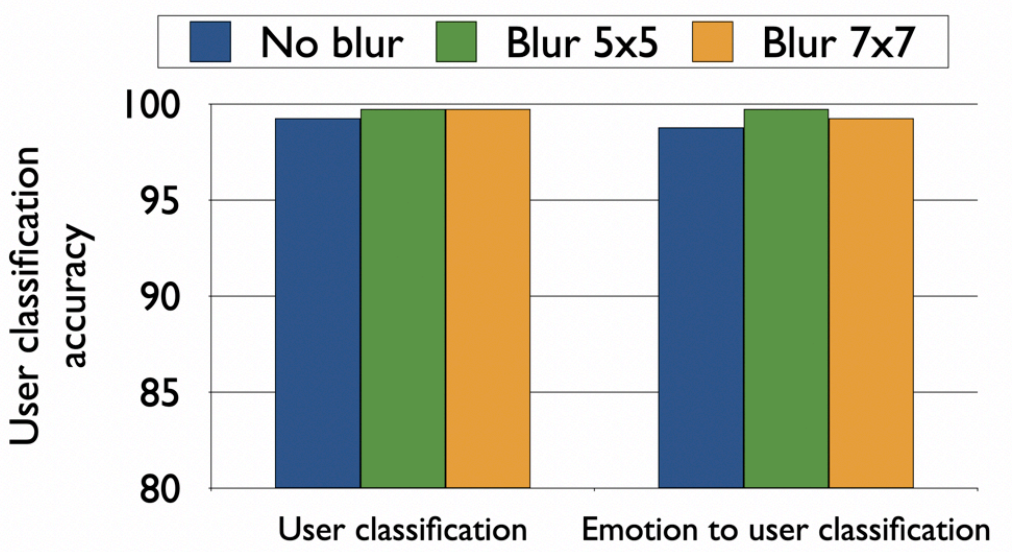
- **Emotion to Face:** CNN trained on emotion classification and fine tuned on user classification



User anonymization through adversarial learning

RQ1: Can we characterize user re-identification risk?

- **User classification:** CNN trained on user classification only
- **Emotion to user classification:** CNN trained on emotion classification and fine tuned on user classification

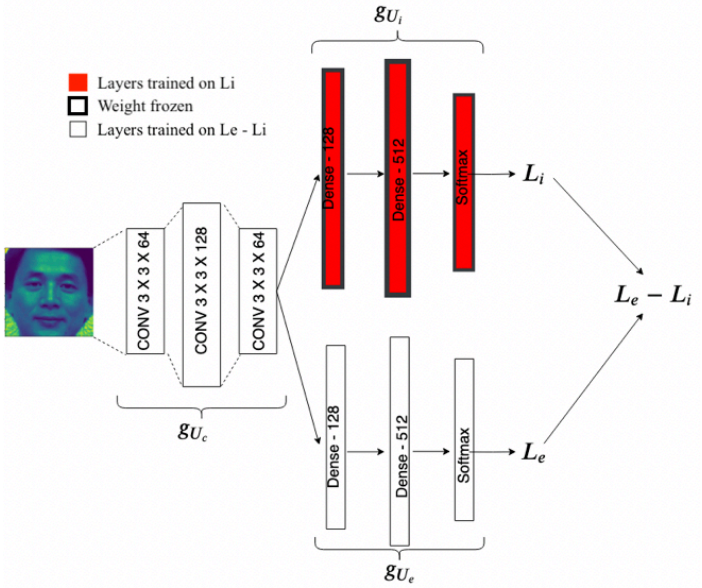


Significant amount of user-dependent information preserved in facial emotion classification models (even after blurring)

User anonymization through adversarial learning

RQ2: Can we learn anonymized signal transformations that preserve behavioral information?

- Anonymizing the convolutional transformation so that user re-identification is not possible
- Alternate training between emotion and user classification losses

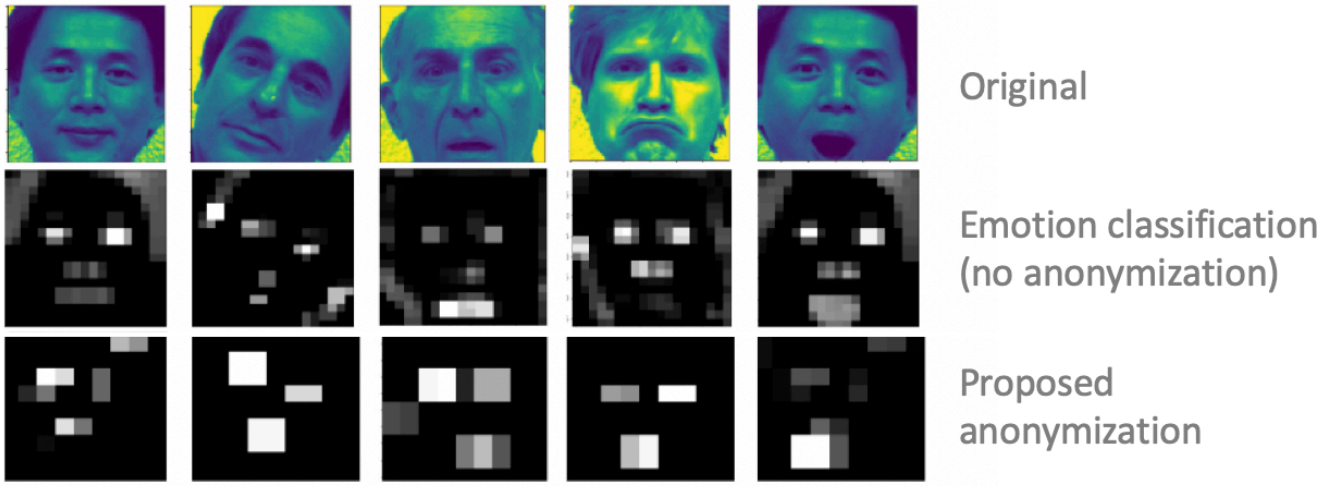


$$\min_{\{\mathbf{U_c},\mathbf{U_e},\mathbf{U_i}\}} \{L_e\left(g_{\mathbf{U_e}}(g_{\mathbf{U_c}}(\mathbf{x})), y_e\right) - \alpha L_i\left(g_{\mathbf{U_i}}(g_{\mathbf{U_c}}(\mathbf{x})), y_i\right)\}$$

User anonymization through adversarial learning

RQ2: Can we learn anonymized signal transformations that preserve behavioral information?

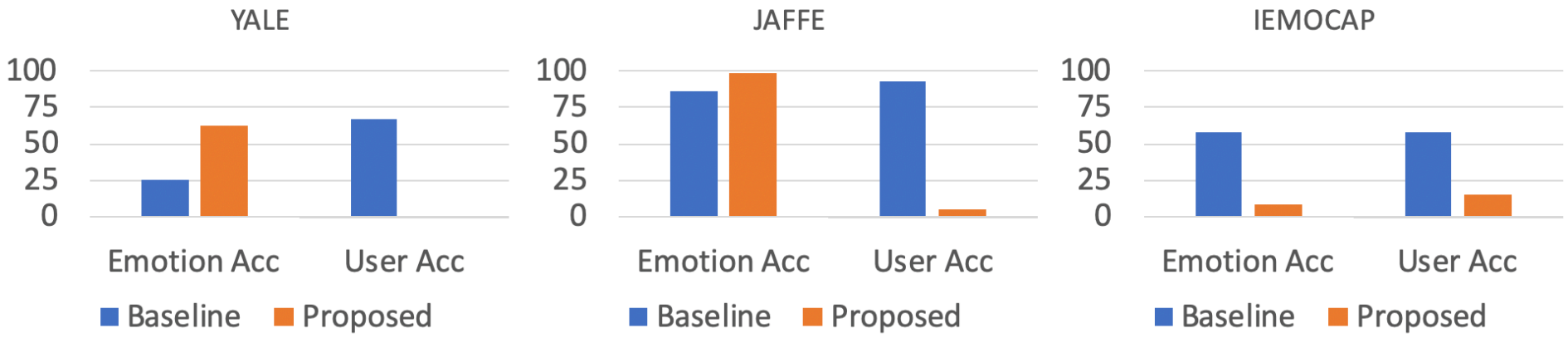
- **Original:** Original images
- **Emotion classification:** Images transformed by the emotion classification models
- **Proposed:** Images transformed by the proposed anonymization approach



User anonymization through adversarial learning

RQ2: Can we learn anonymized signal transformations that preserve behavioral information?

- Baseline: Adversarial learning without alternate training
- Proposed: Adversarial learning with alternate training



Tutorial Overview

- Sources of sociodemographic bias in machine learning (ML)
- Examples of studies examining sociodemographic bias in ML
  - Health - Electronic Health Records
  - Family well-being - Speech, language, physiology
- Approaches for mitigating sociodemographic bias in ML
  - Adversarial learning
  - Fairness regularization
  - Explainability
- General recommendations for fair machine learning

Fairness-aware learning through regularization

- Discrimination score

$$CV = Pr[Y = HighRisk \mid S = SensitiveGroup] - Pr[Y = HighRisk \mid S = NonSensitiveGroup]$$

Predicted risk of sensitive group

Predicted risk of non-sensitive group

- Fairness-aware regularization

$$L = -l(\mathcal{D} \mid \boldsymbol{\Theta}) + nR(\mathcal{D} \mid \boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Loss function

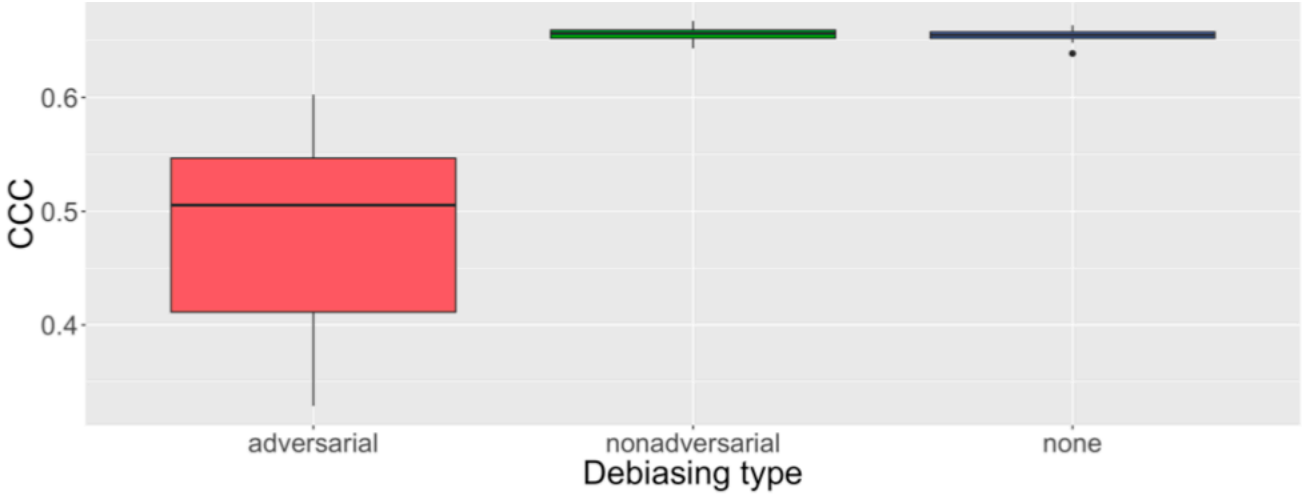
Classifier regularization

Discrimination score
Iteratively estimated after each learning iteration

Kamishima, T., Akaho, S., & Sakuma, J. (2011, December). Fairness-aware learning through regularization approach. In 2011 IEEE 11th International Conference on Data Mining Workshops (pp. 643-650). IEEE.

Comparison between adversarial learning and fairness regularization

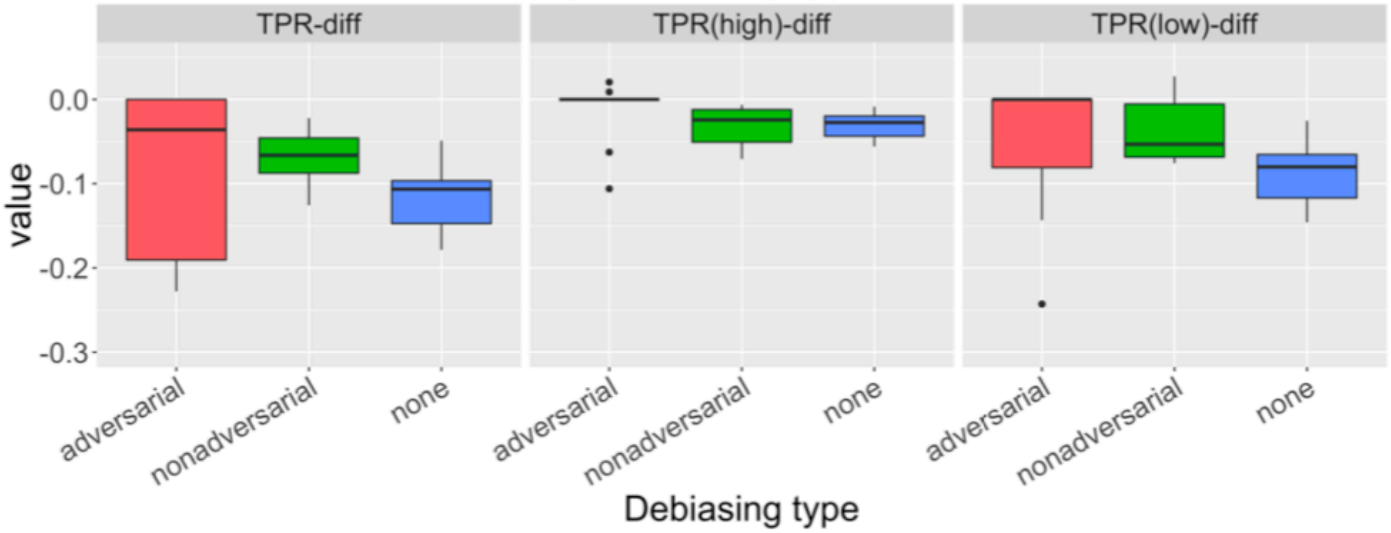
- Gender de-biasing in speech emotion recognition
- The non-adversarial approach maintains equality of odds (CCC) levels similar to no bias mitigation
- The adversarial approach yields lower CCC



Gorrostieta, C., Lotfian, R., Taylor, K., Brutti, R., & Kane, J. (2019). Gender De-Biasing in Speech Emotion Recognition. In INTERSPEECH (pp. 2823-2827).

Comparison between adversarial learning and fairness regularization

- The non-adversarial approach, on the other hand, achieves much better consistency with all metrics



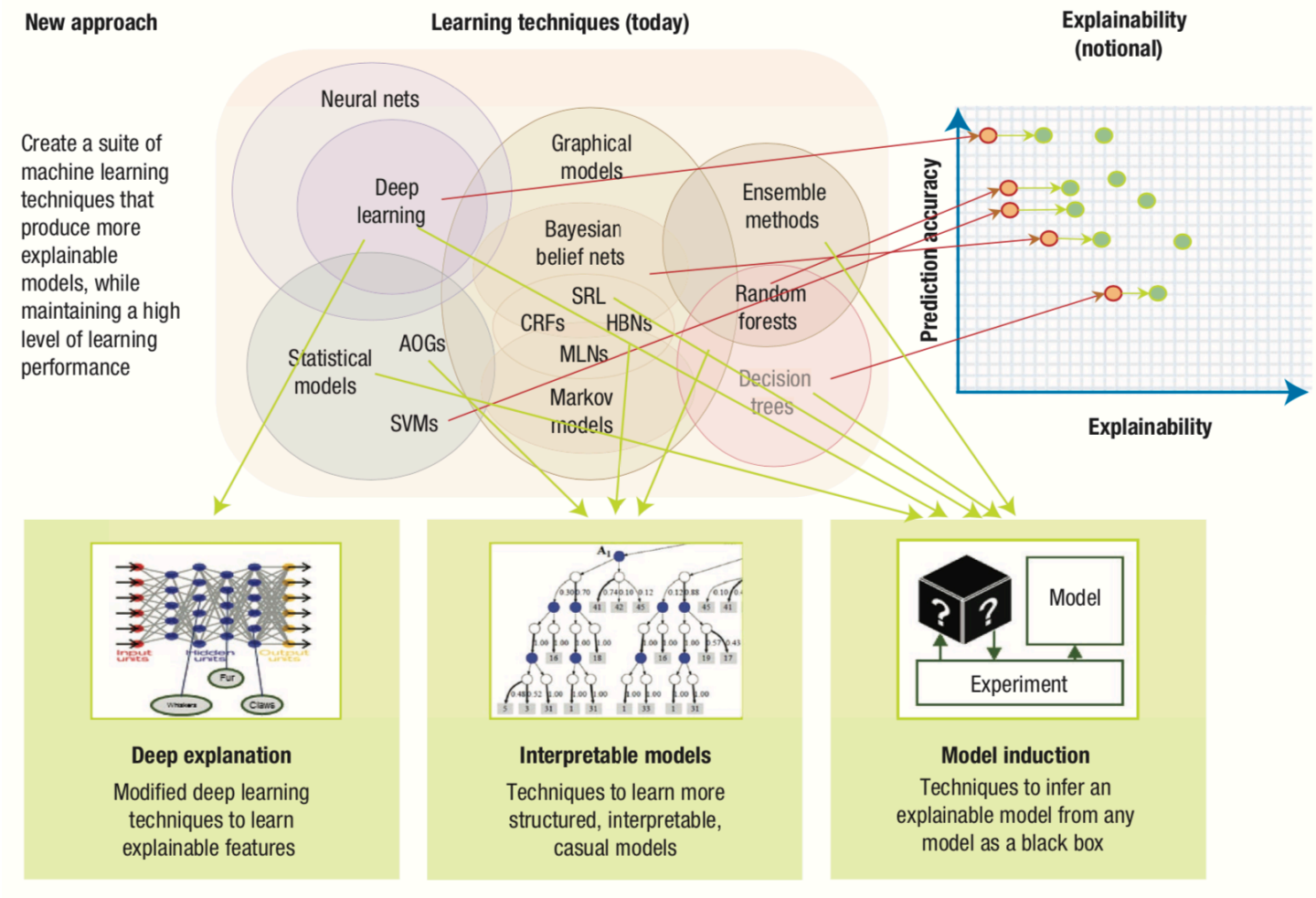
Gorrostieta, C., Lotfian, R., Taylor, K., Brutti, R., & Kane, J. (2019). Gender De-Biasing in Speech Emotion Recognition. In INTERSPEECH (pp. 2823-2827).

Tutorial Overview

- Sources of sociodemographic bias in machine learning (ML)
- Examples of studies examining sociodemographic bias in ML
  - Health - Electronic Health Records
  - Family well-being - Speech, language, physiology
- Approaches for mitigating sociodemographic bias in ML
  - Adversarial learning
  - Fairness regularization
  - Explainability
- General recommendations for fair machine learning

Improving fairness via explainability

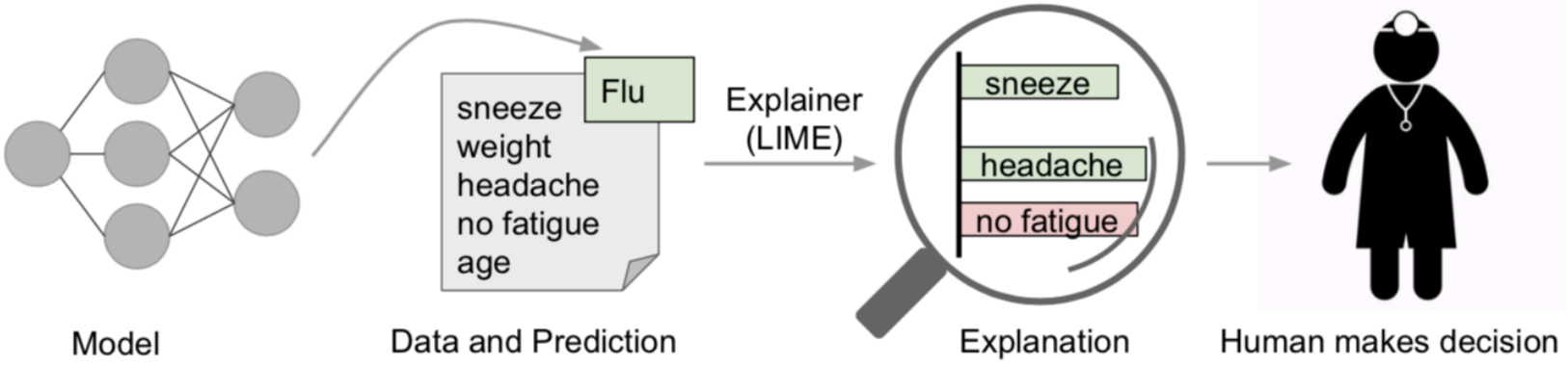
- Understanding inner mechanisms of the model via explainable methods



Improving fairness via explainability

Local Interpretable Model-agnostic Explanations (LIME)

- Presenting artifacts that provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction via explainable methods
- Quantify trust in individual predictions and entire model



Tutorial Overview

- Sources of sociodemographic bias in machine learning (ML)
- Examples of studies examining sociodemographic bias in ML
  - Health - Electronic Health Records
  - Family well-being - Speech, language, physiology
- Approaches for mitigating sociodemographic bias in ML
  - Adversarial learning
  - Fairness regularization
  - Explainability
- General recommendations for fair machine learning

General recommendations for fair machine learning

Design

- Define the goal of the machine learning model and review with diverse stakeholders
- Discuss ethical concerns of how the model could be used and what are the protective groups, also informed by historical data

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872.

General recommendations for fair machine learning

Data collection

- Collect and document training data
- Ensure that participants in the protected group can be identified
- Make sure that the protected group is adequately represented in terms of numbers and features

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872.

General recommendations for fair machine learning

Training & Evaluation

- Train the model to take into account fairness goals
- Measure algorithmic output differences between sensitive and non-sensitive groups
- Assess model output with diverse stakeholders

Deployment

- Systematically review data and continuously evaluate metrics
- Collect feedback from participants and stakeholders

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872.