

# ML and Society

May 3, 2022

# Reminder: The case study

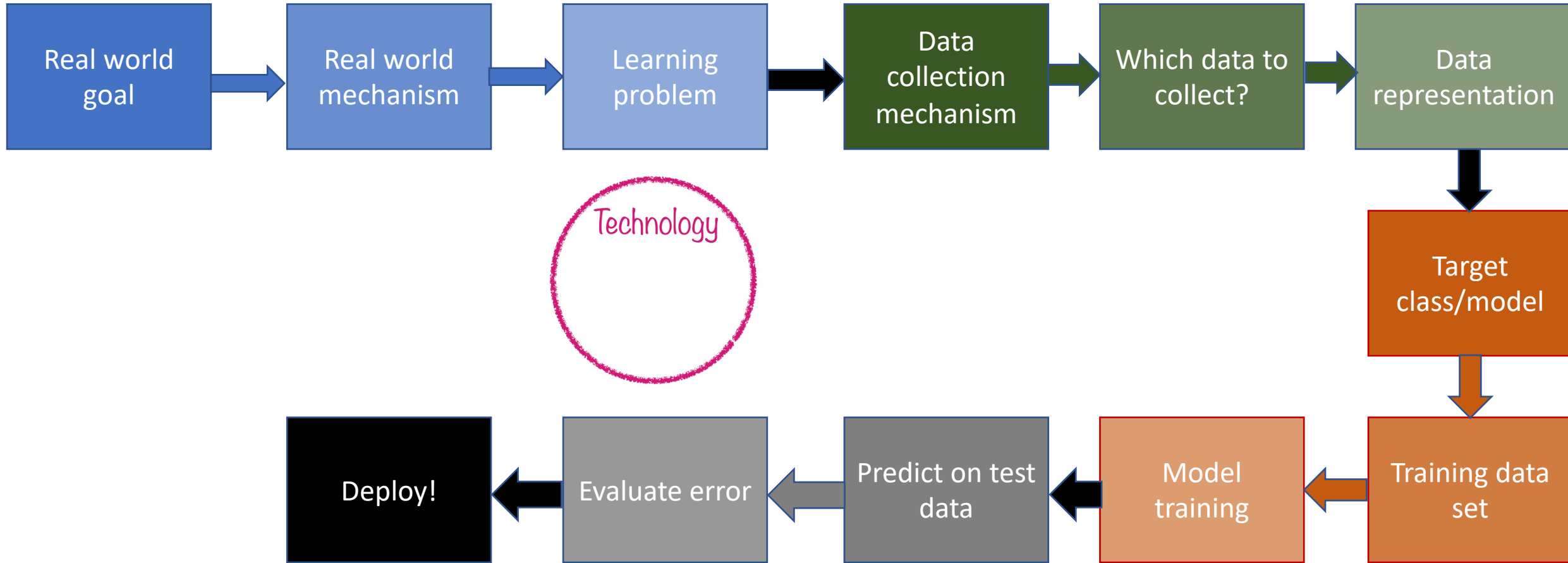


**Why** (were there racial disparities in the algorithm's recommendation)?

A: Because the **algorithm** developers chose a proxy variable that had racial bias.

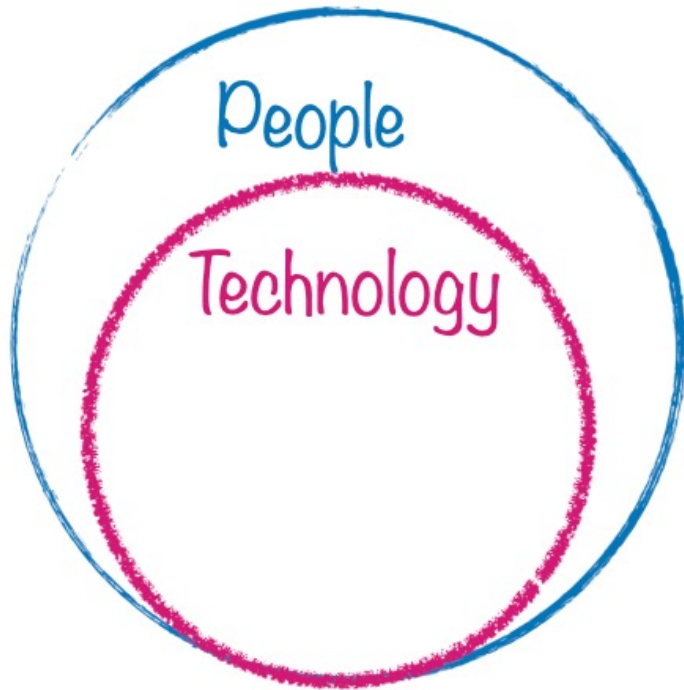


# In this course think beyond proxy variable



**Why** was the underlying proxy variable racially biased?

A: Because the **people** who made the underlying decisions did so in a racially biased way



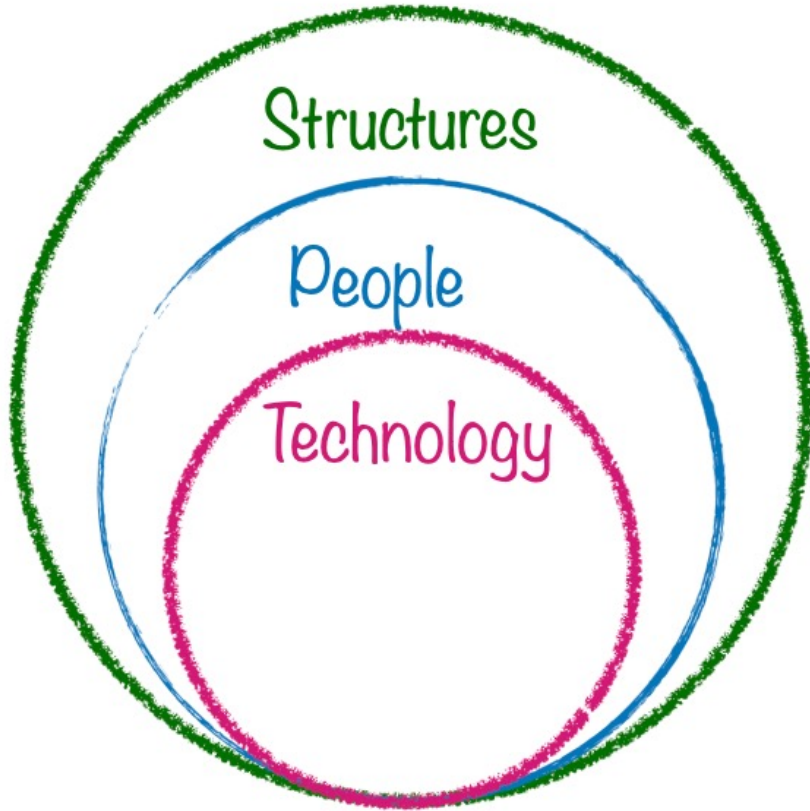
It is a widely known fact that there exists racial inequality within the health care system. Specifically, black people are often barred from health services due to its expensive cost, which artificially lowers the cost of care for black patients. Since the algorithm uses cost as the proxy variable, it creates this illusion that black patients do not need as much medical care.

I understand the underlying decisions that people made as a reference to the doctors and nurses whose decisions about their patients were used as the input data to compute the algorithm. Since their decisions were racially biased, the same translates into the algorithm.

Clearly there is vested self interest and a conflict of interest when united healthcare designs an algorithm that predicts cost.

This quote from the news story covering the original publication highlights the long history of racially disparate healthcare. This is further exacerbated by inequality in access created by the lower socioeconomic status and opportunities afforded to people of color as a result of structural racism. While the original creators of the algorithm may not have had racist intentions, their failure to consider historical context led to deeply racist outcomes.

**Why** did these people make racially biased decisions?



A: Because racial biases and discrimination are built into our **society** at a structural level

Starting from birth, till getting a job and beyond that, a white person is given more priority than a same aged black person with same qualifications. Why? **Because it is built into our society at a structural level.**

**Due to income inequality, black families cannot afford health insurance.** This goes to show how the fundamental institutions are racially biased, and it perpetuates across other institutions, even if it is through indirect means.

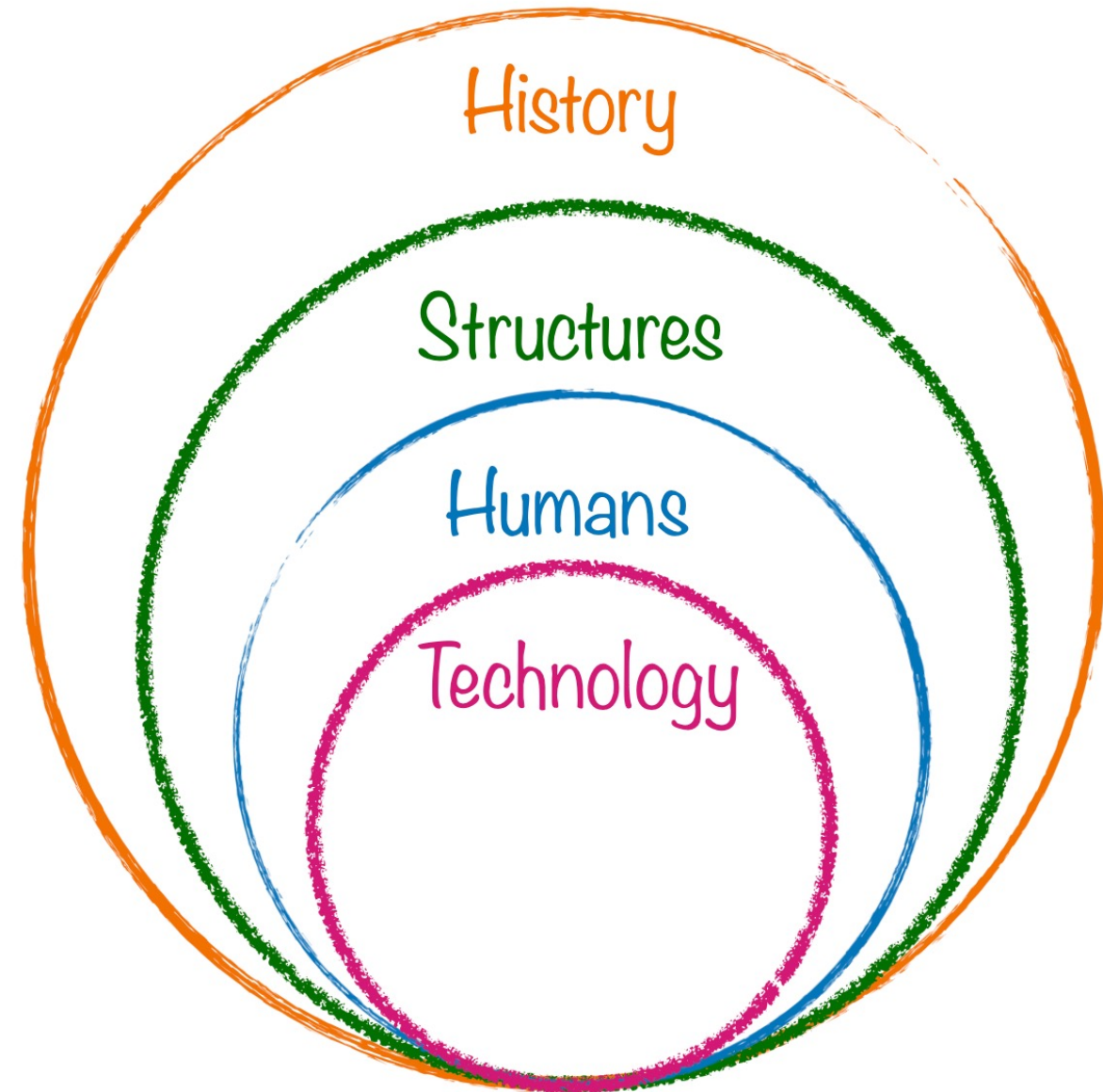
This textbook case is an excellent example of how racism is structural, rather than an individual failure. **Individual doctors and nurses may not be personally bigoted, but inequality is deeply engrained in the entire practice of medicine.**

Only awareness of that knowledge and active efforts to change it can be considered anti-racist, interpersonal equality on an individual level does nothing to change this status quo.



**Why** are racial bias and discrimination built into our society at a structural level?

A: Because American society has been **historically** unequal and discriminatory, and structural racism is reproduced over time



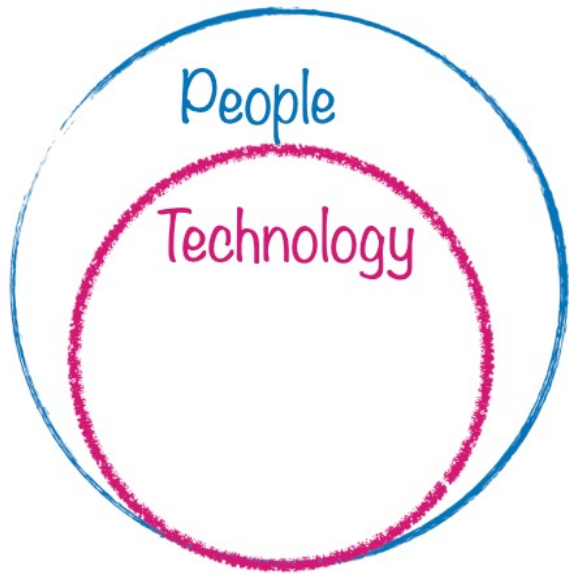
Time by time we see that blacks were neglected, given a lower stature enslaved by whites, forced to live in redlining areas, purposefully were denied the rights to social security or citizenship or were not allowed to own a property. These structural inequality and structural racism made the young white children believe that blacks were always inferior and were treated like an animal. These reasons say that due to bad historical decisions made by American society built racial discrimination into the society at a structural level.

These movements were largely target at Black Americans, and created this system of distrust against the medical system for the Black folks. This trickled down to modern society where many Black people are wary of the medical system, and would refrain from seeking medical help, such as with the COVID vaccine.

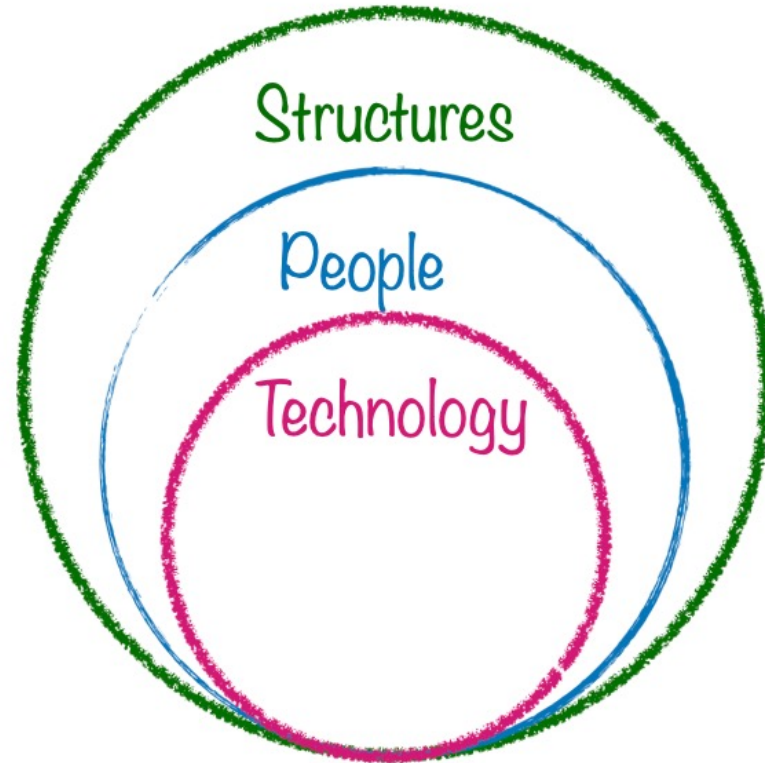
This brief quote is one of many examples from the Vox video of how the medical establishment is largely built on a foundation of discrimination and racism. Many medical discoveries and advancements were made at the expense of enslaved people, and conversely medicine was used to rationalize slavery. This deep intertwining of racism and healthcare still greatly effects modern day outcomes across racial lines, and is necessary context to build more equitable algorithms in healthcare

Because at that time segregation existed and the prices for the neighborhood varied depending how populated they were with black people or white people. That's why HOLC was created to grade the neighborhood when there was segregation existed and white people paid more to live in a better neighborhood for the better future and the redlining created and black people started to suffer in the different way that they never imagined.

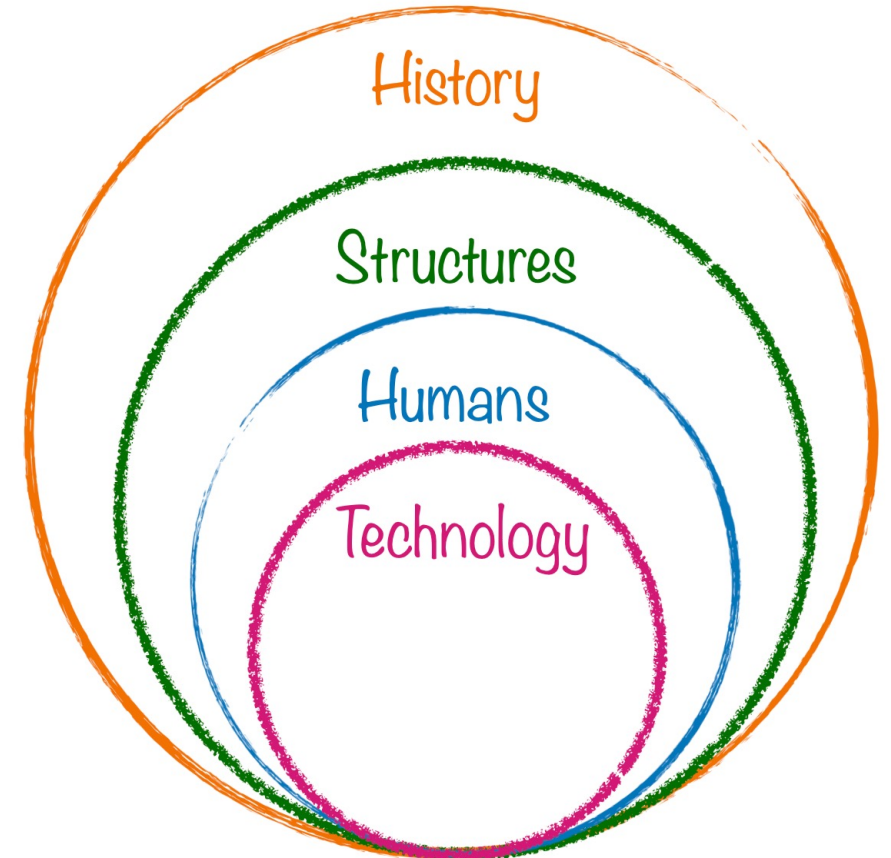
# People vs. Structure vs. History



How do individuals contribute to racist outcomes?

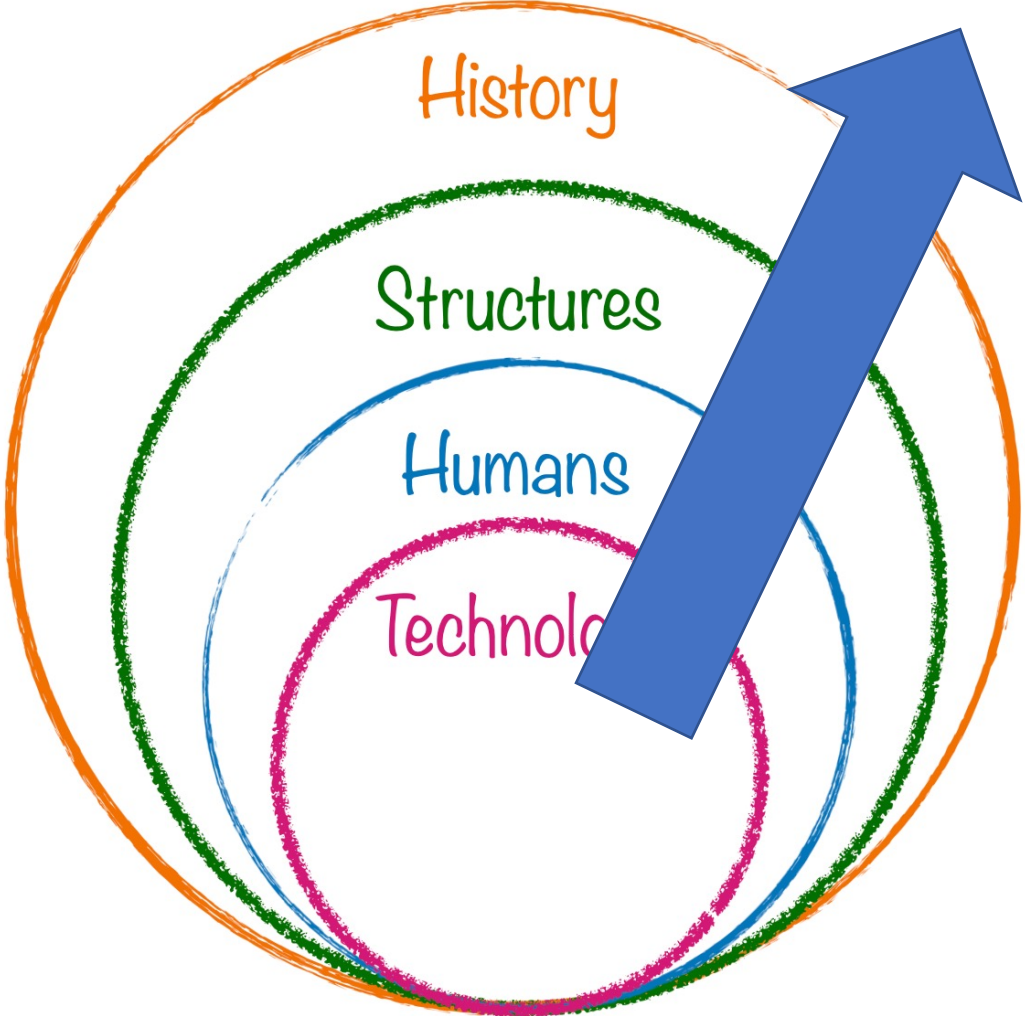


What are the *current* structures in society that lead to racist outcomes?

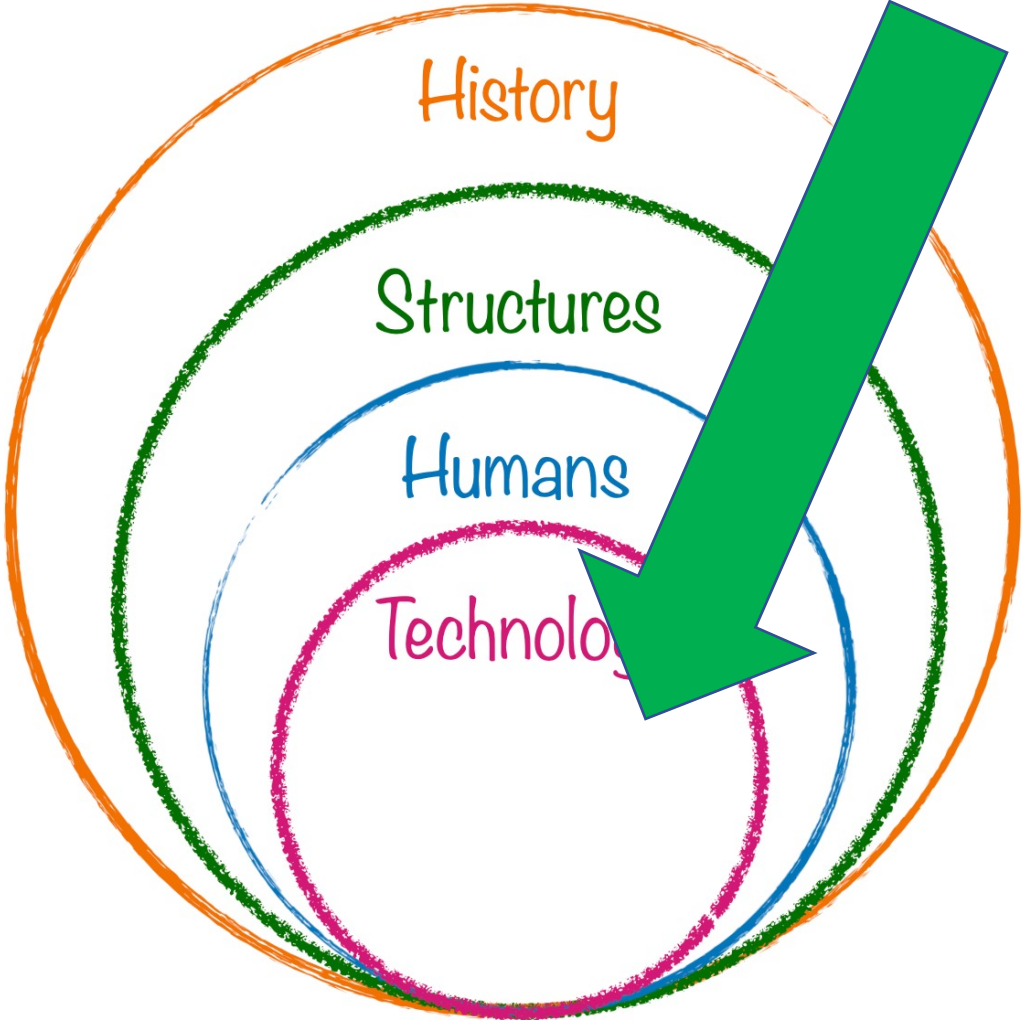


How does history *explain* the *current* racist structures in place?

# The WHY worksheet



# The HOW worksheet



# Discuss!

## Questions (4 × 15 = 60 points)

Keep the [instructions](#) and well as the [suggested resources](#) in mind when looking through the following questions.

- Question:** **How** do you address/change the fact that “*American society has been historically unequal and discriminatory, and structural racism is reproduced over time*”?
  - Answer:**
    - (First answer)
    - (Second answer)
  - List the Reference(s) You Used:
- Question:** **How** do you address/change the fact that “*racial biases and discrimination are built into our society at a structural level*”?
  - Answer:**
    - (First answer)
    - (Second answer)
  - List the Reference(s) You Used:
- Question:** **How** do you address/change the fact that “*people who made the underlying decisions did so in a racially biased way*”?
  - Answer:**
    - (First answer)
    - (Second answer)
  - List the Reference(s) You Used:
- Question:** **How** do you address/change the fact that “*the algorithm developers chose a proxy variable that had racial bias*”?
  - Answer:**
    - (First answer)
    - (Second answer)
  - List the Reference(s) You Used:

# HOW worksheet is out

 note @43    

stop following

7 views

## HOW worksheets are now out

Following up on @35 and @36, the HOW worksheet is now out:

<http://www-student.cse.buffalo.edu/~atri/ml-and-soc/spr22/ip/how.html>

Note that this is due last week of class. I'm putting this out now so that y'all know what is expected but y'all probably be in a better place to fill in this worksheet after we go over the stuff between now and May.

ip



edit

· good note | 0

Updated 1 day ago by Atri Rudra

# Passphrase for today: **Kristian Lum**

← **Kristian Lum**  
5,257 Tweets



**Kristian Lum**  
@KLdivergence

Responsible ML Researcher at Twitter META | Statistician at ❤️ | @FAccTConference  
OG | Past @PennEngineers CS Faculty & @hrdag | she/her

📍 Chicago, IL 🗓️ Joined April 2009

1,344 Following 16.3K Followers

Follow



# Some examples on how to do it “right”

Rest of the slides are from Kenny

# How might you change the algorithm?

Change the data so that includes less biased information

To change the algorithm, new data that is not racially biased would need to be used to improve the algorithm to make it not specifically discriminate against minorities.

We would not change the algorithm itself, rather the biases within the data.

Help to reprogram it, and change the data we give it. Eliminate bias from the data we are giving it. Also try to incorporate hard facts instead of human bias.

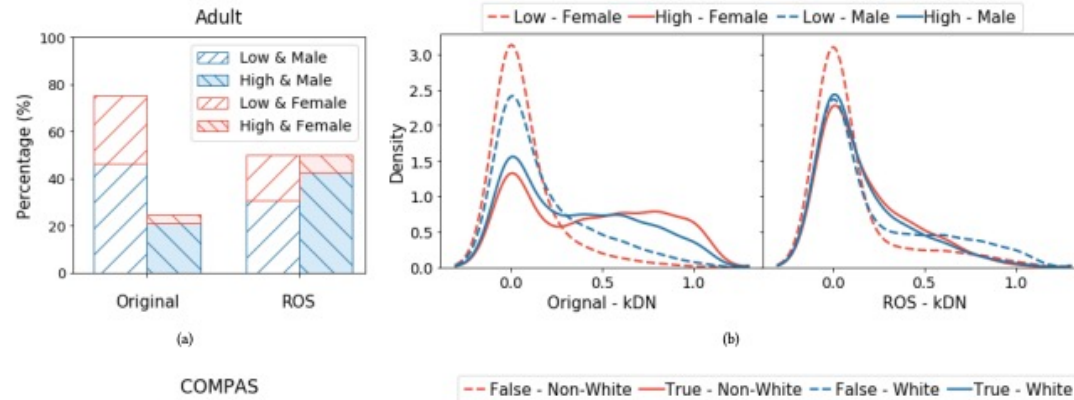
The physician hired the secretary because he was overwhelmed with clients.



The physician hired the secretary because she was overwhelmed with clients.



Feeding more information  
on people of color rather  
than focusing on white  
people.



# Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes

Shen Yan  
shenyan@usc.edu

University of Southern California

Hsien-te Kao  
hsientek@usc.edu

University of Southern California

Emilio Ferrara  
emiliofe@usc.edu

University of Southern California

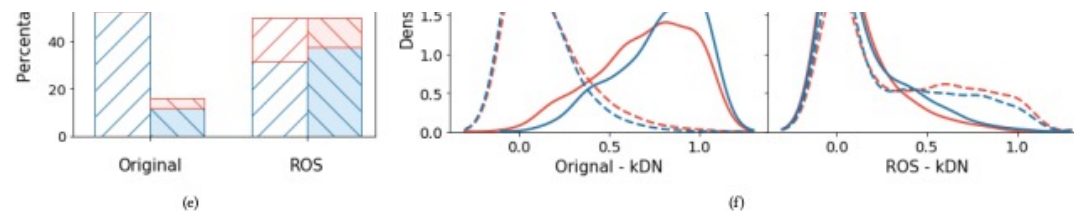
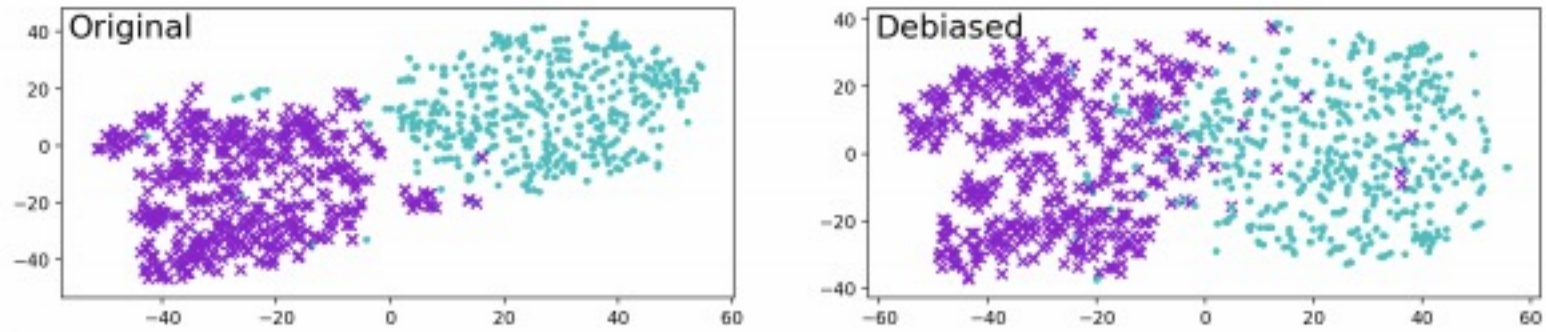


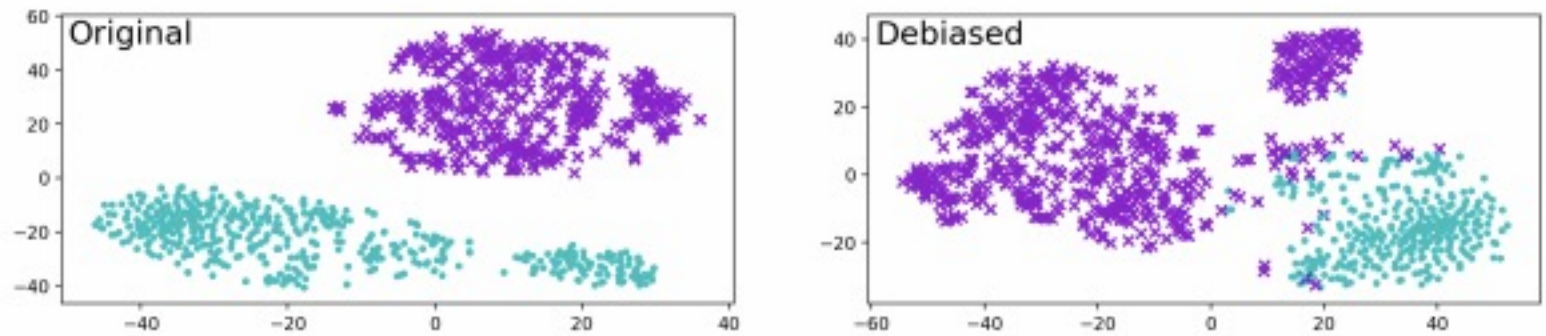
Figure 1: Examples of the bias changes before and after class balancing. (a), (c) and (e) illustrate the change of *distribution bias*. (b), (d) and (f) compare the *hardness bias* before and after class balancing for *Adult*, *COMPAS* and *Violent Crime* data.

# How might you change the algorithm?

Completely collect new data  
and **rewrite the entire**  
**algorithm** without the bias.



**(a) Clustering for HARD-DEBIASED embedding, before (left hand-side) and after (right hand-side) debiasing.**



**(b) Clustering for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.**

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of NAACL-HLT* (pp. 609-614).

# How might you change the algorithm?

Group consensus for the systems to maintain neutrality. **Third-party reviewing**

The steps we would take to change the algorithm is to **have a more diverse group recreate it from scratch**. To allocate a certain group of people to look after what is wrong with algorithms that are giving racist outcomes.



# New York City Proposes Regulating Algorithms Used in Hiring

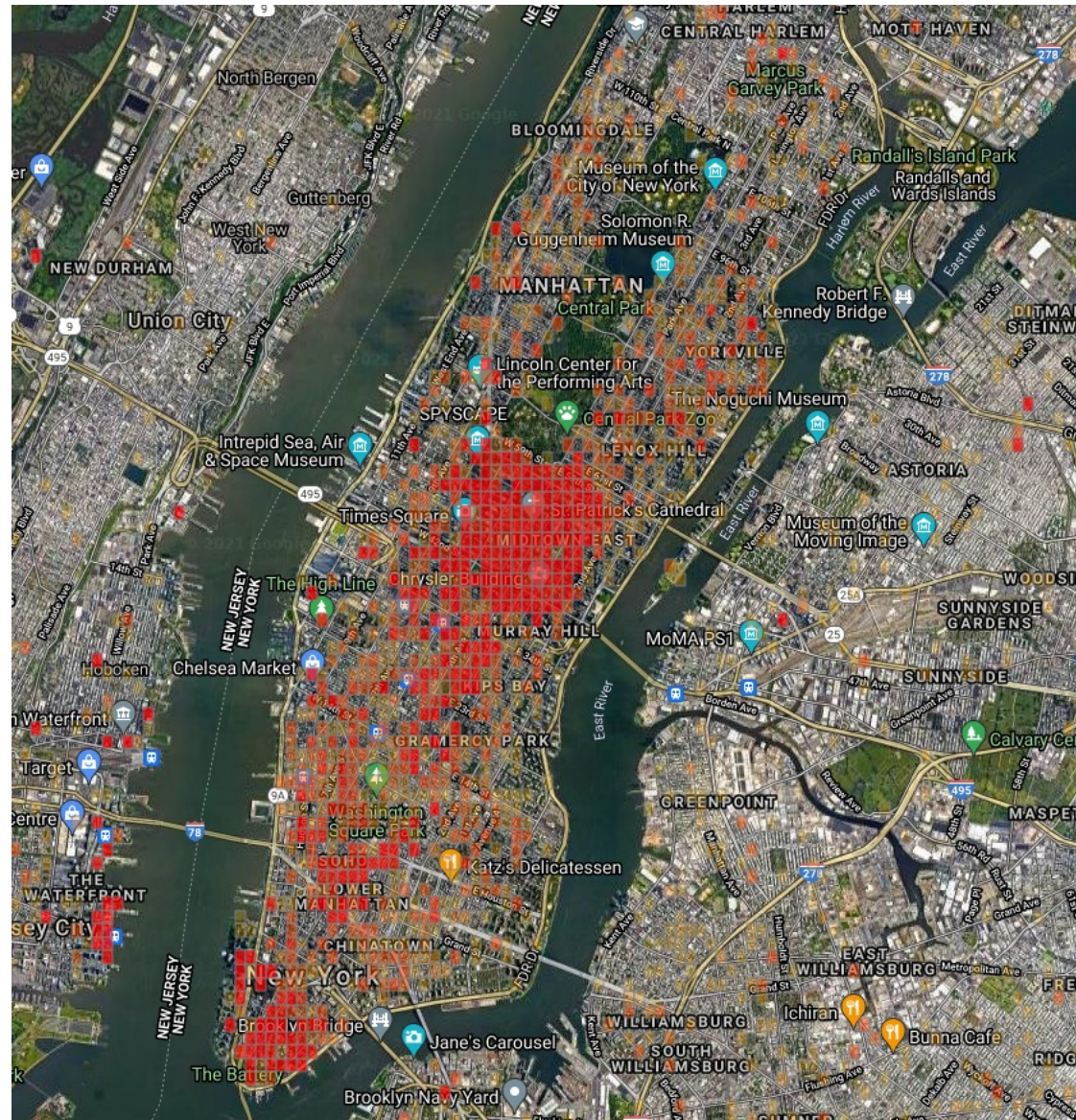
A bill would require firms to disclose when they use software to assess candidates, and vendors would have to ensure that their tech doesn't discriminate.



<https://www.wired.com/story/new-york-city-proposes-regulating-algorithms-hiring/>

# How might you change the algorithm?

Don't let the algorithm identify so specific of areas to be criminal, but broaden the range that the program identifies with a high criminal rate.



<https://whitecollar.thenewinquiry.com/>

# How might you change the algorithm?

What about giving each computer scientist a sheet about what they need to write in the algorithm? Like for example having in their heads that they have to include everyone in the algorithm?

# Datasheets for Datasets

TIMNIT GEBRU, Google

JAMIE MORGENSTERN, Georgia Institute of Technology

BRIANA VECCHIONE, Cornell University

JENNIFER WORTMAN VAUGHAN, Microsoft Research

HANNA WALLACH, Microsoft Research

HAL DAUMÉ III, Microsoft Research; University of Maryland

KATE CRAWFORD, Microsoft Research; AI Now Institute

## Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.<sup>1</sup>

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset's release in 2007.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The construction of the LFW database was supported by a United States National Science Foundation CAREER Award.

Any other comments?

## Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as “background”.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

<sup>1</sup>All information in this datasheet is taken from one of five sources. Any errors that were introduced from these sources are our fault.

Original paper: <http://www.cs.cornell.edu/people/pabo/movie-review-data/>; LFW survey: <http://vis-www.cs.umass.edu/lfw/lfw.pdf>; Paper measuring LFW demographic characteristics: <http://biometrics.cse.msu.edu/Publications/Face/HanJain.UnconstrainedAgeGenderRaceEstimation.MSU.TechReport2014.pdf>; LFW website: <http://vis-www.cs.umass.edu/lfw/>.

The dataset does not contain all possible instances. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format.

Is there a label or target associated with each instance? If so, please provide a description.

Each image is accompanied by a label indicating the name of the person in the image.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included in the dataset.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The dataset comes with specified train/test splits such that none of the people in the training split are in the test split and vice versa. The data is split into two views, View 1 and View 2. View 1 consists of a training subset (pairsDevTrain.txt) with 1100 pairs of matched and 1100 pairs of mismatched images, and a test subset (pairsDevTest.txt) with 500 pairs of matched and mismatched images. Practitioners can train an algorithm on the training set and test on the test set, repeating as often as necessary. Final performance results should be reported on View 2 which consists of 10 subsets of the dataset. View 2 should only be used to test the performance of the final model. We recommend reporting performance on View 2 by using leave-one-out cross validation, performing 10 experiments. That is, in each experiment, 9 subsets should be used as a training set and the 10<sup>th</sup> subset should be used for testing. At a minimum, we recommend reporting the estimated mean accuracy,  $\hat{\mu}$  and the standard error of the mean:  $S_E$  for View 2.

$\hat{\mu}$  is given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{10} p_i}{10} \quad (1)$$

where  $p_i$  is the percentage of correct classifications on View 2 using subset  $i$  for testing.  $S_E$  is given as:

$$S_E = \frac{\hat{\sigma}}{\sqrt{10}} \quad (2)$$

Fig. 1. Example datasheet for Labeled Faces in the Wild [14], page 1.

# How might you change the algorithm?

Incorporate a more diverse group into the machine learning process.

try building these new algorithms with a diverse group of people to get equal perspectives.

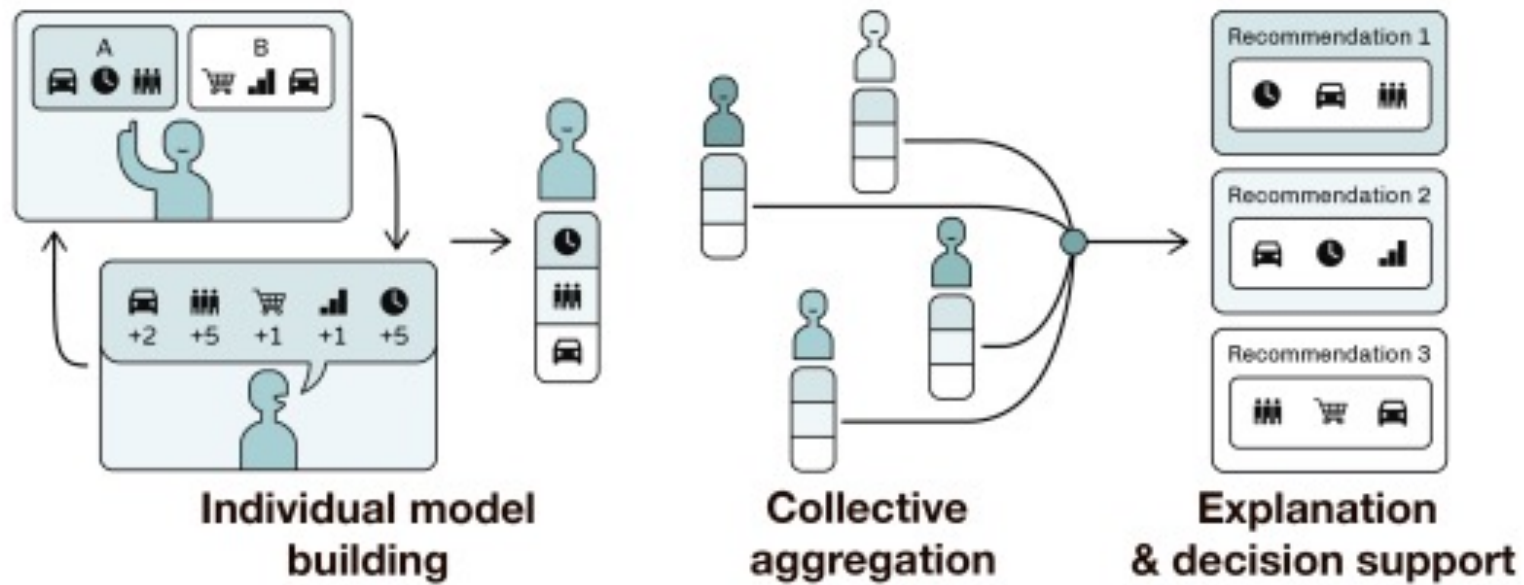


Fig. 1. The WeBuildAI framework allows people to participate in designing algorithmic governance policy. A key aspect of this framework is that individuals create computational models that embody their beliefs on the algorithmic policy in question and vote on the individual's behalf.

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., ... & Procaccia, A. D. (2019). WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1-35.

# How might you change the algorithm?

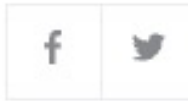
It would be best if police **stop using**  
**the algorithm** in general,



# Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ



---

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

The algorithm tries to give people a risk factor based on their age, criminal record, environment, etc. If the algorithm focused more on the nature of the crime that occurred and if the motive for committing another crime is still present rather than trying to characterize the criminal, it might be more accurate.

Instead of just passing through data as is with all the relevant labels. A stage should be added to allow the model to better understand how different labels influence each other.

Humans can form an understanding of how different hidden factors can affect other factors but that is an ability that machine learning models currently lack. This may be related to bias.

The algorithm should not take into consideration elements that do not impact the person themselves. Someone may live in a crime-filled area but that does not mean that that person will commit a crime

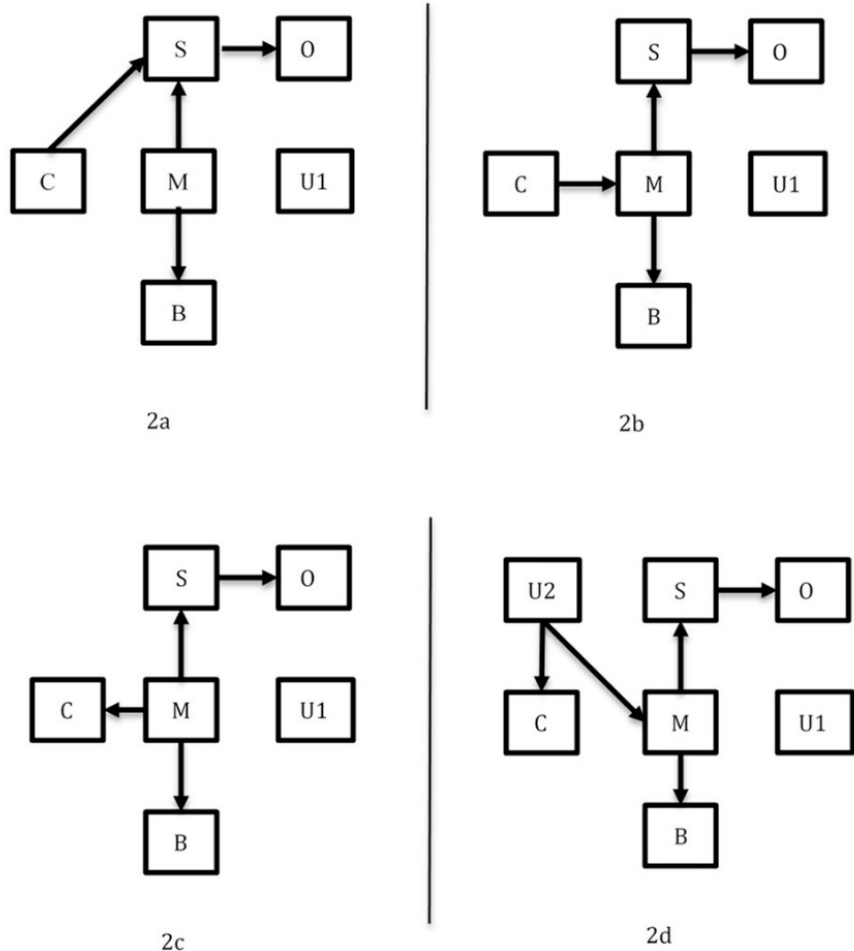


Figure 2: Causal Structures Inducing Associations Between C and S

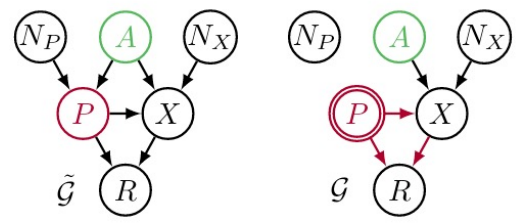


Figure 3: A template graph  $\tilde{\mathcal{G}}$  for proxy discrimination (left) with its intervened version  $\mathcal{G}$  (right). While from the benevolent viewpoint we do not generically prohibit any influence from  $A$  on  $R$ , we want to guarantee that the proxy  $P$  has no overall influence on the prediction, by adjusting  $P \rightarrow R$  to cancel the influence along  $P \rightarrow X \rightarrow R$  in the intervened graph.

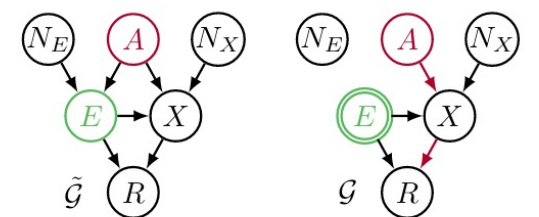


Figure 4: A template graph  $\tilde{\mathcal{G}}$  for unresolved discrimination (left) with its intervened version  $\mathcal{G}$  (right). While from the skeptical viewpoint we generically do not want  $A$  to influence  $R$ , we first intervene on  $E$  interrupting all paths through  $E$  and only cancel the remaining influence on  $A$  to  $R$ .

**Change the  
question you're  
asking**

<b>Officer 1</b>	High number of rule of conduct violations in last 15 years	Officer was suspended in last 15 years	High number of counselling interventions after special investigations	High number of sustained complaints in the last 15 years
<b>Officer 2</b>	High number of counselling interventions after special investigations	High number of rule of conduct violations in last 15 years	High number of prior adverse incidents in last 15 years	High number of special investigations correctives written in last 15 years
<b>Officer 3</b>	High number of complaints against officer in last 15 years	High number of rule of conduct violations in last 15 years	Officer was suspended in last 15 years	High number of counselling interventions after special investigations
<b>Officer 4</b>	Officer has dealt with high number of domestic violence incidents	High number of special investigations correctives written in last 15 years	Officer was suspended in last year	High number of accidents in last 1 year
<b>Officer 5</b>	Officer has dealt with high number of suicide incidents	High number of preventable accidents in last 1 year	Officer uses weapons often	Officer was suspended in last 15 years

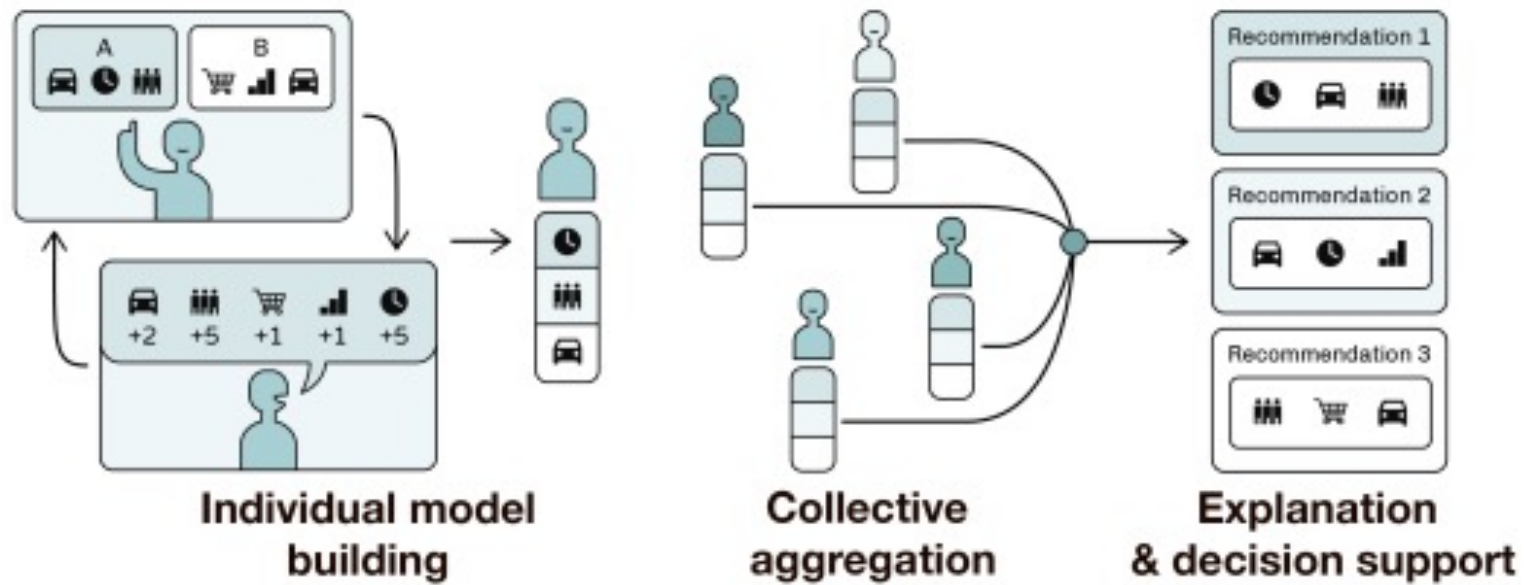


Fig. 1. The WeBuildAI framework allows people to participate in designing algorithmic governance policy. A key aspect of this framework is that individuals create computational models that embody their beliefs on the algorithmic policy in question and vote on the individual's behalf.

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., ... & Procaccia, A. D. (2019). WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1-35.

# Predicting poverty

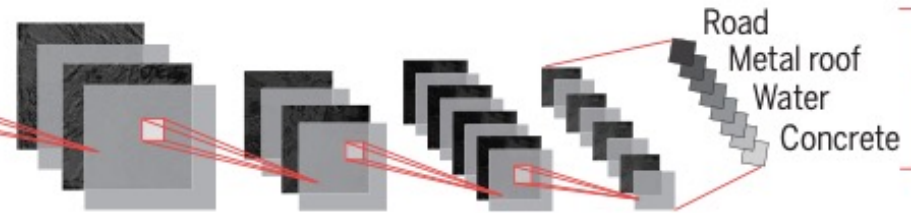
Satellite images can be used to estimate wealth in remote regions.

## Neural network learns features in satellite images that correlate with economic activity

Daytime satellite photos capture details of the landscape



Convolutional Neural Network (CNN) associates features from daytime photos with nightlight intensity



Satellite nightlights are a proxy for economic activity

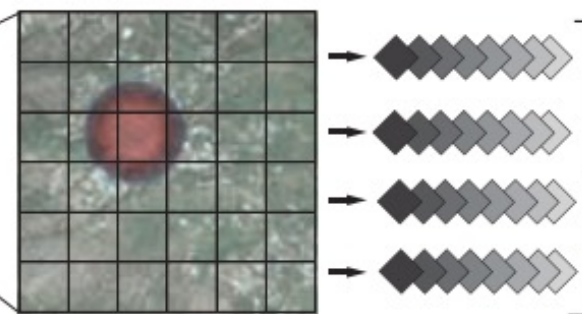


## Daytime satellite images can be used to predict regional wealth

Household survey locations



CNN processes satellite photos of each survey site



Features from multiple photos are averaged



Ridge regression model reconstructs ground truth estimates of poverty

**Fight back**



# Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies

Nicholas Vincent  
Northwestern University  
nickvincent@u.northwestern.edu

Hanlin Li  
Northwestern University  
lihanlin@u.northwestern.edu

Nicole Tilly  
Northwestern University  
nicoletilly2023@u.northwestern.edu

Stevie Chancellor\*  
University of Minnesota  
steviec@umn.edu

Brent Hecht  
Northwestern University  
bhecht@northwestern.edu

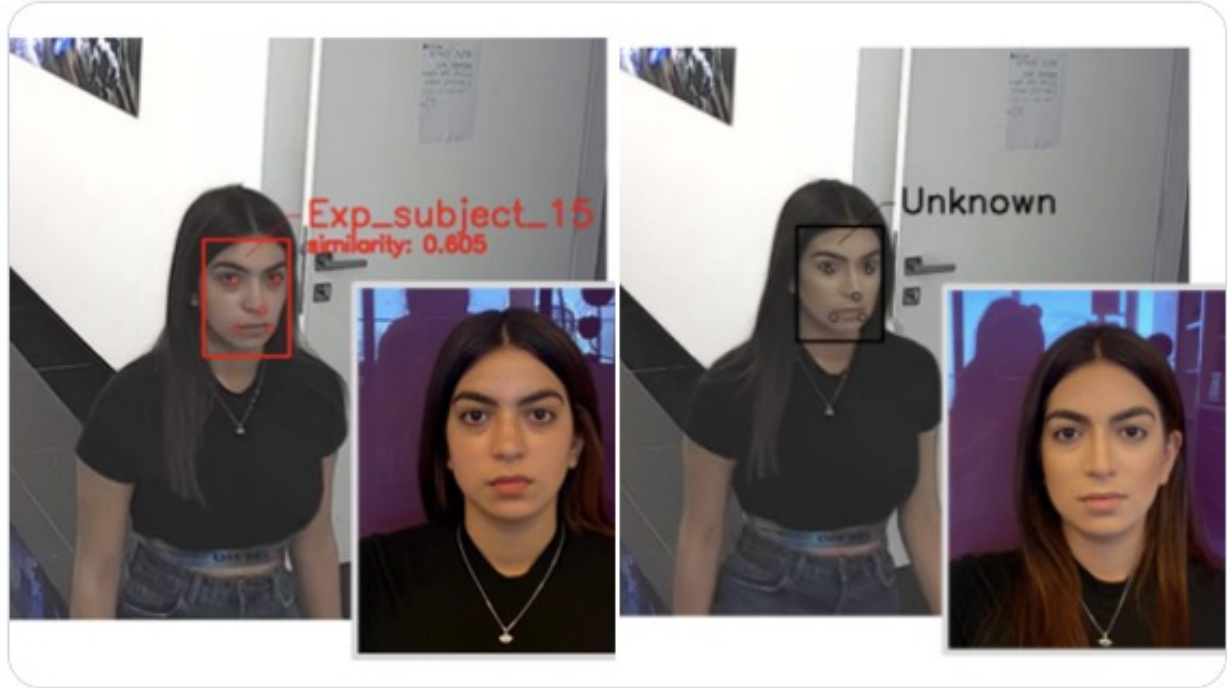


James Vincent ✓

@jjvincent



the secret to defeating facial recognition with make-up is ... contouring? left and right are before and after applying adversarial makeup; an advance over more obvious alterations, though only tested against the ArcFace model [arxiv.org/pdf/2109.06467...](https://arxiv.org/pdf/2109.06467...)



<https://www.wsj.com/articles/the-hong-kong-protesters-toolkit-for-a-cat-and-mouse-game-with-authorities-11568628648>