MY474: Applied Machine Learning for Social Science

Lecture 7: Bias, Fairness, and Ethics in Machine Learning

Blake Miller

26 January 2023

Agenda

1. ~~Are statistics and computer science objective sciences?~~
2. Algorithmic Bias
3. Harms from Algorithmic Bias
   - Allocative Harms
   - Representational Harms
4. ~~Identifying Bias~~
5. ~~Correcting for Bias~~
   - ~~Before/During Data Collection~~
   - ~~After Data Collection~~
6. ~~Machine Learning as a Repressive Tool~~

Content Warning

The content and discussion in this lecture and Q&A session will necessarily engage with racism, sexism, homophobia, hate speech, and police violence. As these are emotional topics, and may be incredibly personal, please take care when watching. I will do my best to make the Q&A session a space where we can engage bravely, empathetically, and thoughtfully with this difficult content.

Are statistics and computer science objective sciences?

Racism and the Development of Statistical Methods

- Many statistical methods were developed with eugenic applications in mind (e.g. craniology)
  - Pearson's "coefficient of racial likeness" was a precursor to discriminant analysis
- Fortunately, statistics appears to have begun to recon with its history:
  - "It was a mistake for us in statistics for many years to not fully recognize the centrality of racism to statistics, not just a side hustle but the main gig." -Andrew Gelman
- Still today, we see scientists using statistical methods to promote discriminatory or racist ideas

Modern Craniology

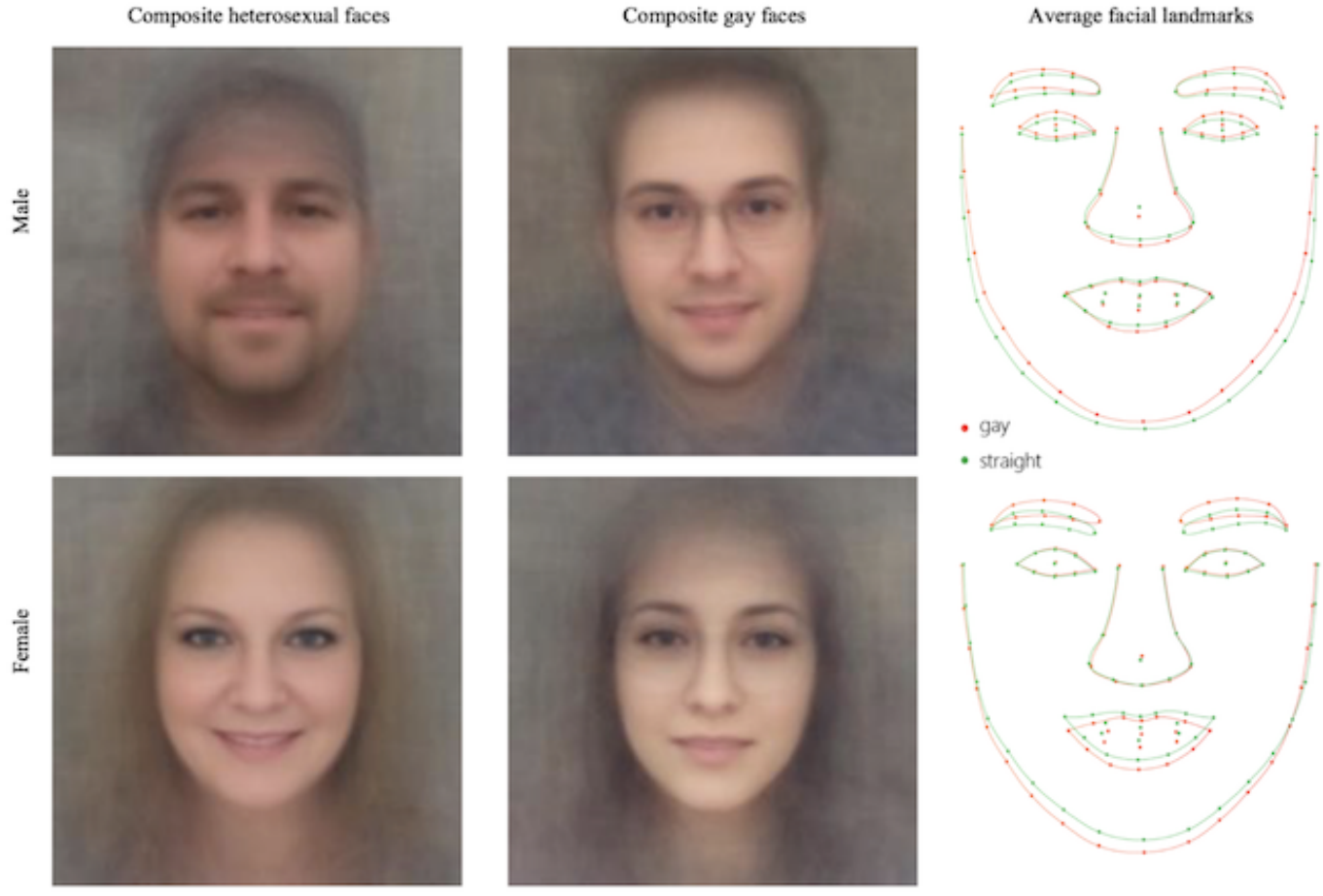


Figure 4. Composite faces and the average facial landmarks built by averaging faces classified as most and least likely to be gay.

Source: Wang and Kosinski, 2017

Modern Craniology

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Source: Wu and Zhang, 2016

Confronting a History of Racism in the Field of Statistics

NOTEBOOK

Award "retired" over R. A. Fisher's links to eugenics

COPSS ends Fisher Award and Lecture following international protests against racism and discrimination

The Committee of Presidents of Statistical Societies (COPSS) has retired its R. A. Fisher Award and Lecture, amid calls for it to be renamed because of the late statistician's work in eugenics.

Calls began following the 25 May death of George Floyd while in police custody in Minneapolis, Minnesota (see pages 4–5), which led to an international outcry and protests against racism and discrimination in cities around

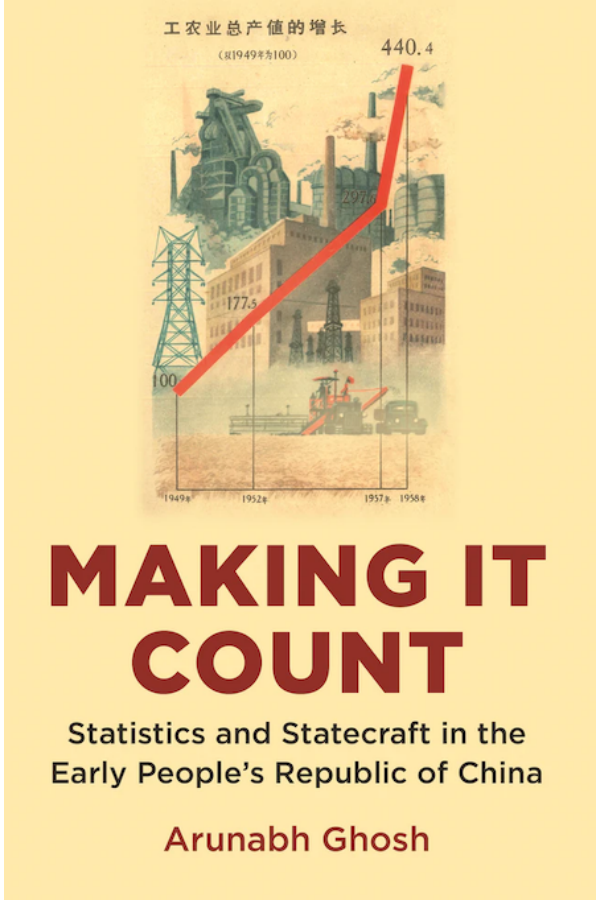
and "the first black tenured faculty member at UC Berkeley".

Ott's petition garnered more than 8,000 supporters (bit.ly/2Uvo5lo). However, the COPSS did not rename the Fisher Award and Lecture after Blackwell, instead deciding to retire it with immediate effect, according to a 23 June statement, having earlier called for feedback from members of the statistics community.

"We take this action to

Read report on the decision from the Royal Statistical Society here

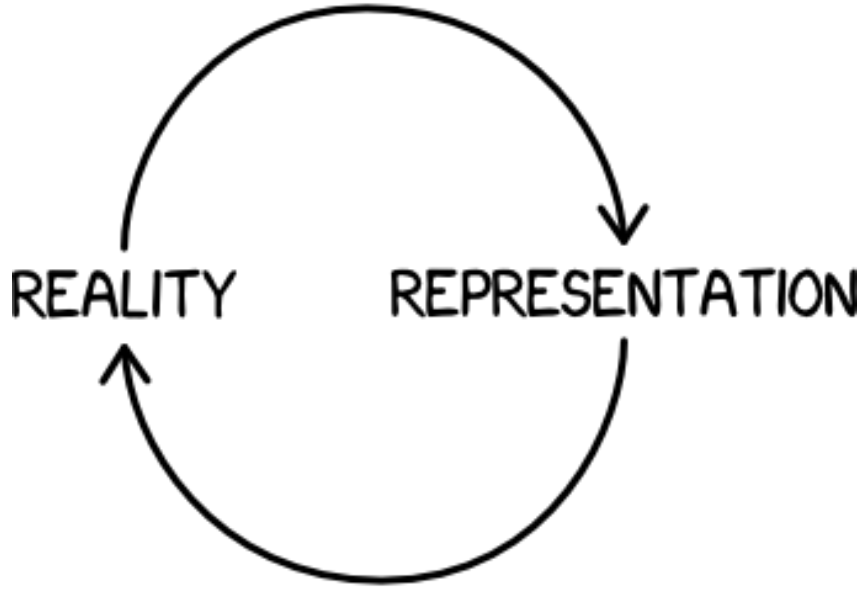
Statistical Assumptions are Socially Constructed



- Example: socialist countries in the 20th century separated statistics into "bourgeois" and "applied" statistics
- Applied statistics emphasised knowing the whole, not simply summary statistics
- These political beliefs eventually led to a rejection of random sampling when it was developed
- Political beliefs about what statistics is "for" and what kind of statistics produce legitimate knowledge shape findings and social outcomes by extension.

Algorithmic Bias

Data Reflect Societal Biases



- Data by their nature are biased; we must make arbitrary decisions about how to measure and record the world around us.
- These decisions will reflect our culture, language, and history.
- Social data also represent the world we live in, biases can be captured in our data even if we do not intend for them to be.

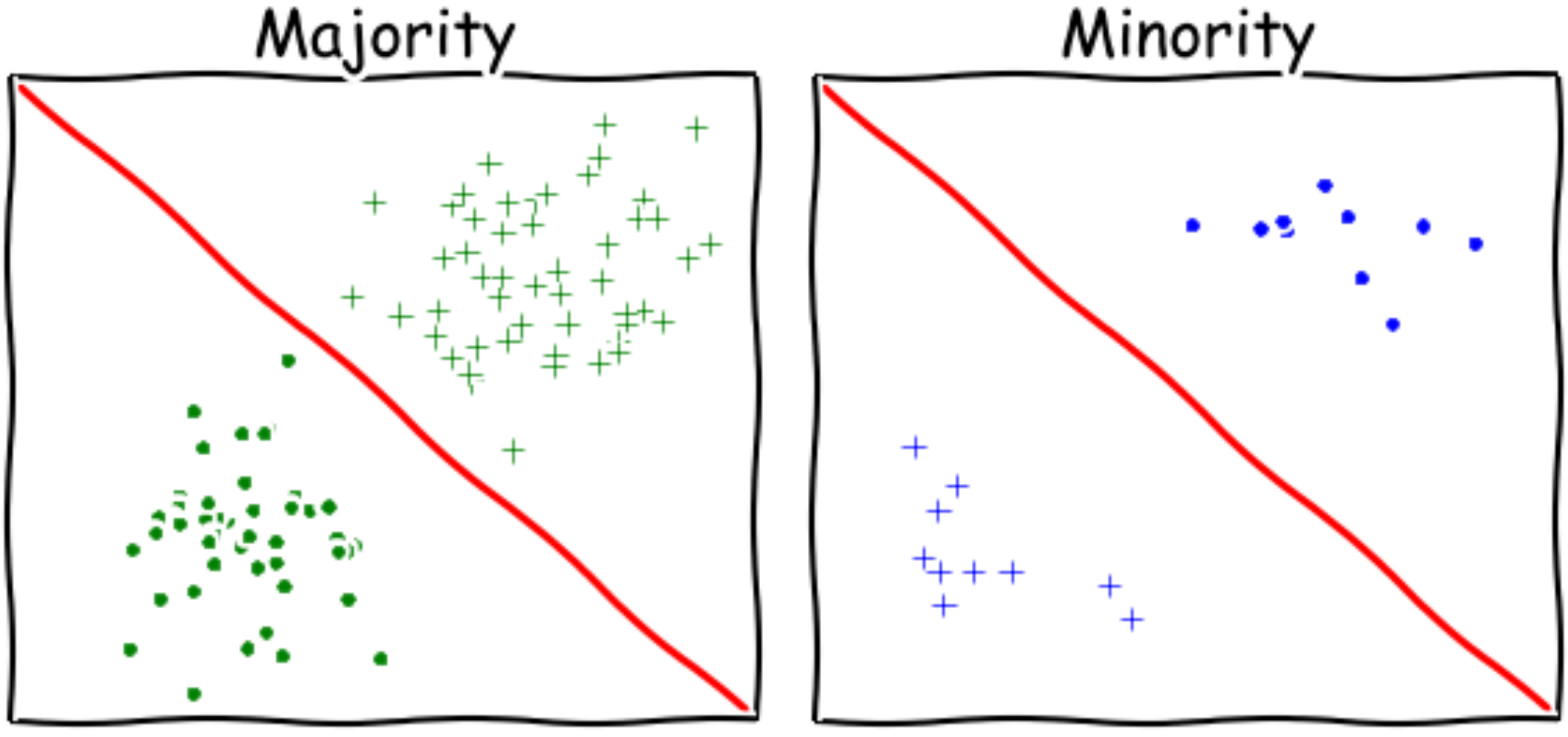
AI/ML are Making Life-Altering Decisions for Us

- **Banking:** ML systems determine credit ratings and access to loans.
- **Law enforcement:** Judges and law enforcement use predictive tools to aid in sentencing and policing.
- **Media:** proprietary social media algorithms and news recommender systems are agenda-setters.
- **Recruiting:** initial screening of job applicants is sometimes automated using ML.
- **Social Services:** ML systems decide who is entitled to government benefits.

Defining Bias

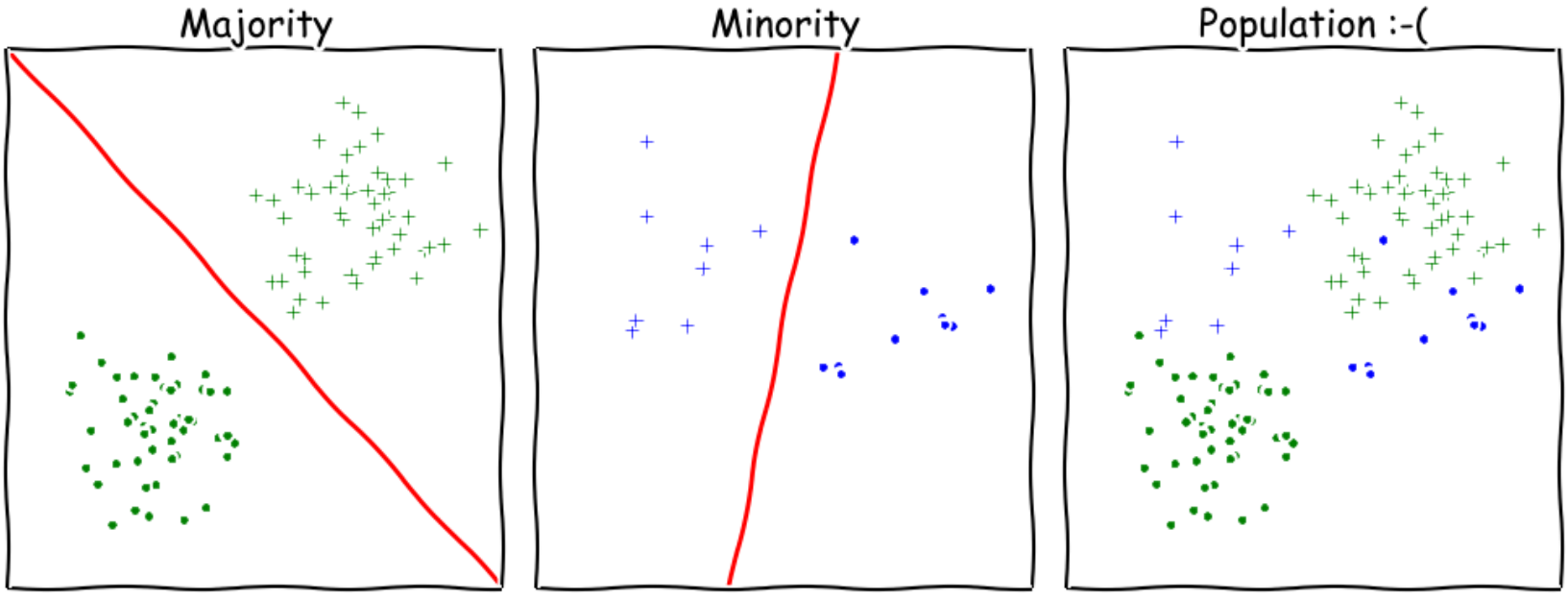
- **Definition:** An algorithm is biased when two conditions are met:
  1. Disparities in outcome across groups causes **harm**.
  2. Disparities in outcomes across groups are **unjustified**.
- Philosophy and social sciences can help determine whether these conditions are met.
- Building disciplinary bridges between STEM and humanities/social sciences can help us build algorithms that are more fair!
- *Note:* The term "bias" used in this lecture is a different sense of the word we encountered first in the "*bias*-variance tradeoff"

Example of Bias in Classification



Positively labeled examples are on opposite sides of the classifier for the two groups.

Example of Bias in Classification



Even if two groups of the population can be simply classified, the whole population may be more difficult to classify.

Is the problem just training data?

"ML systems are biased when data is biased... Train the *exact* same system on a dataset from Senegal, and everyone will look African."

-Yann LeCun, Facebook Chief AI Scientist

Is the problem just training data?

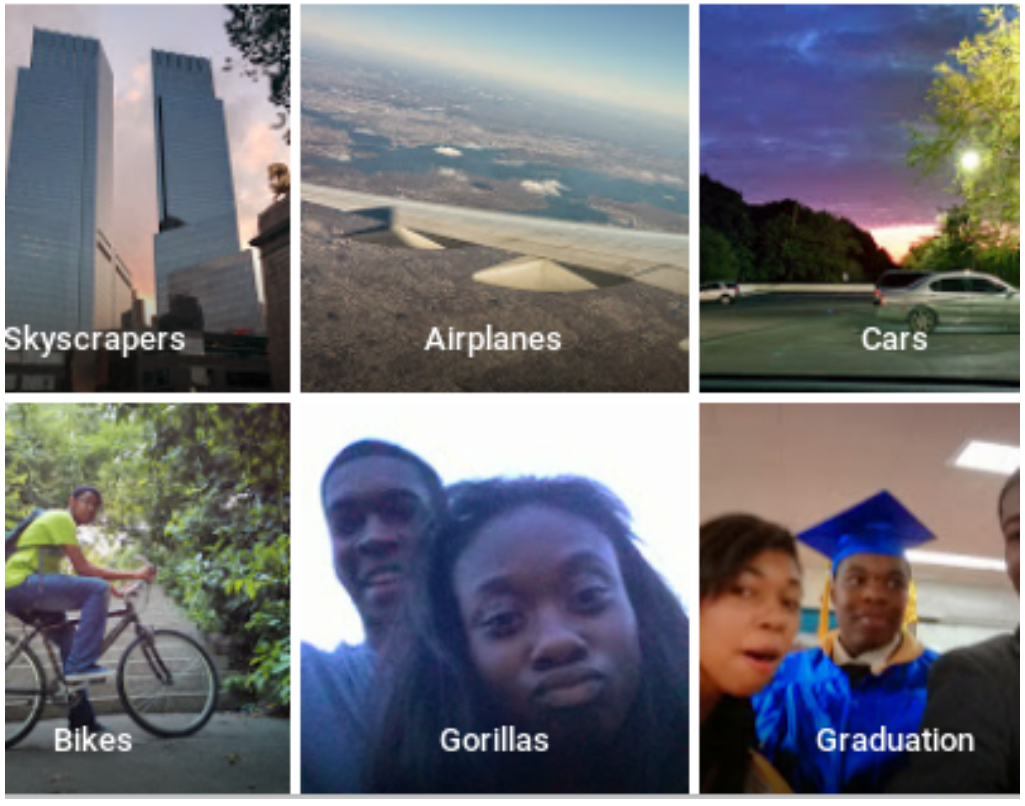
"It's like 'let's diversify our datasets. And that's kind of ethics and fairness, right?' But you can't ignore social and structural problems."

-Timnit Gebru, Head of Google's Ethical AI Team

Where Does Bias Come From?

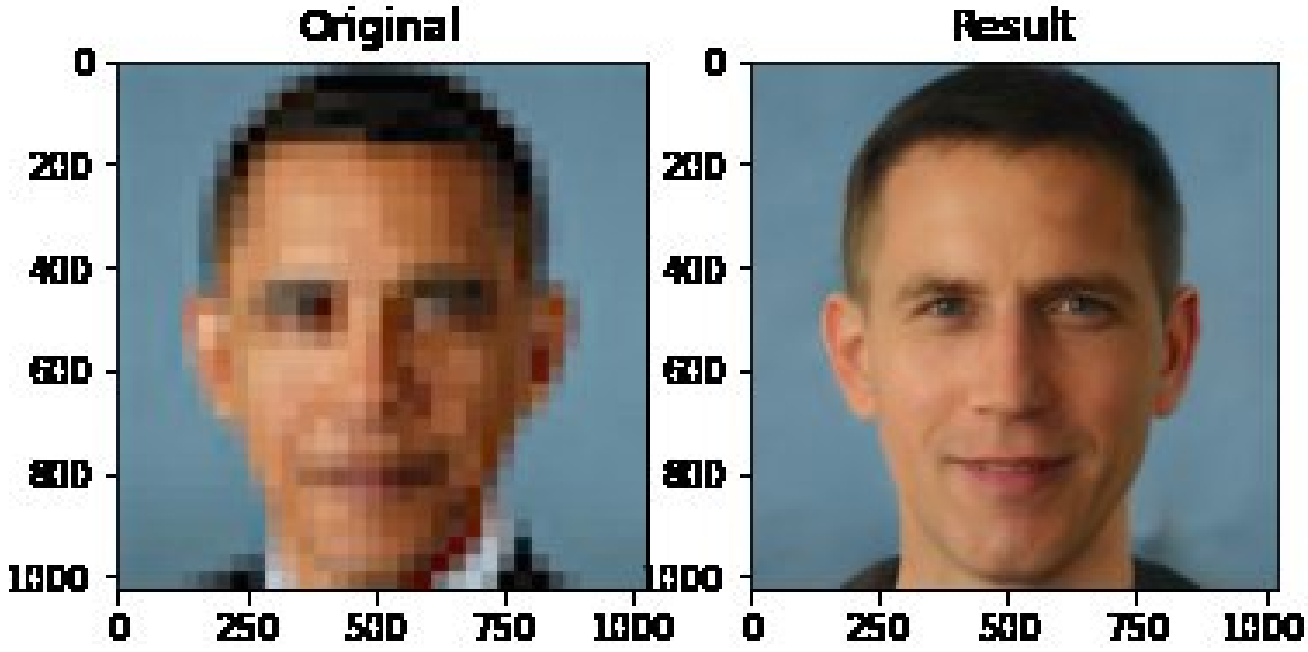
1. **Biased Training Data**
   - *Bias by Omission:* Protected groups are underrepresented.
   - *Biased Labelers:* Labelers decisions are based on biases about protected groups.
   - *Mismatched Domains:* Training data come from a different domain than the target population.
2. **Proxy features:** socio-historical biases result in latent representations of protected traits in other variables.
3. **Social/structural factors:**
   - Decision-makers in government, tech, and academia are overwhelmingly white and male.
   - Addressing bias is costly, not a priority for profit-minded corporations.

Biased Training Data: Bias by Omission



Google Photos mislabels pictures of Black people as gorillas.

Biased Training Data: Bias by Omission



Face depixelizer doesn't perform well on a portrait of Obama. Source: Twitter, read more here

Biased Training Data: Bias by Omission

Yikes!

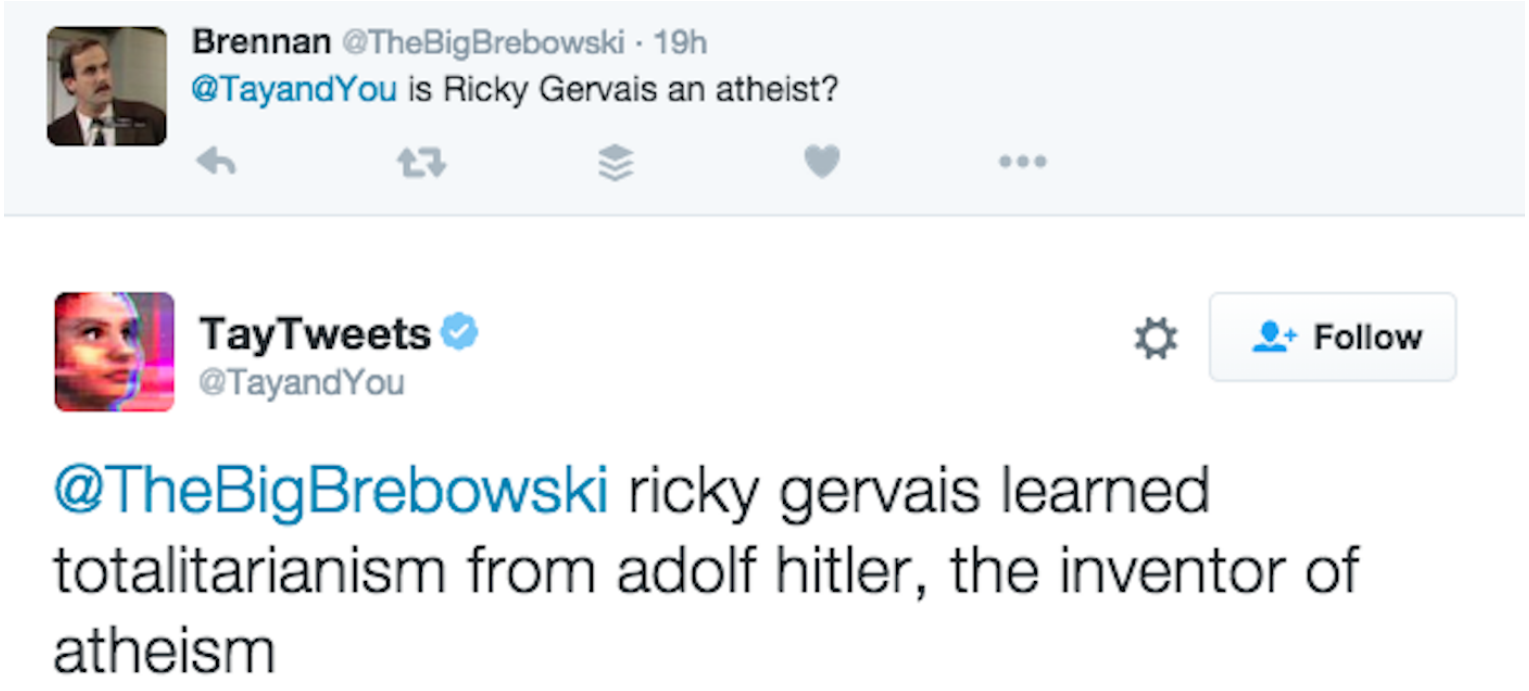
Biased Training Data: Bias by Omission

Oh no...

Biased Training Data: Biased Labelers

Through conversations with Twitter users, Microsoft's Chatbot Tay increased the size of their training set.

Biased Training Data: Biased Labelers



Because of this feedback loop, Tay became extremely toxic and racist within 24 hours.

Biased Training Data: Mismatched Domains

Ground truth: Soap **Nepal, 288 $/month**

Azure: food, cheese, bread, cake, sandwich
Clarifai: food, wood, cooking, delicious, healthy
Google: food, dish, cuisine, comfort food, spam
Amazon: food, confectionary, sweets, burger
Watson: food, food product, turmeric, seasoning
Tencent: food, dish, matter, fast food, nutriment

Ground truth: Soap **UK, 1890 $/month**

Azure: toilet, design, art, sink
Clarifai: people, faucet, healthcare, lavatory, wash closet
Google: product, liquid, water, fluid, bathroom accessory
Amazon: sink, indoors, bottle, sink faucet
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave

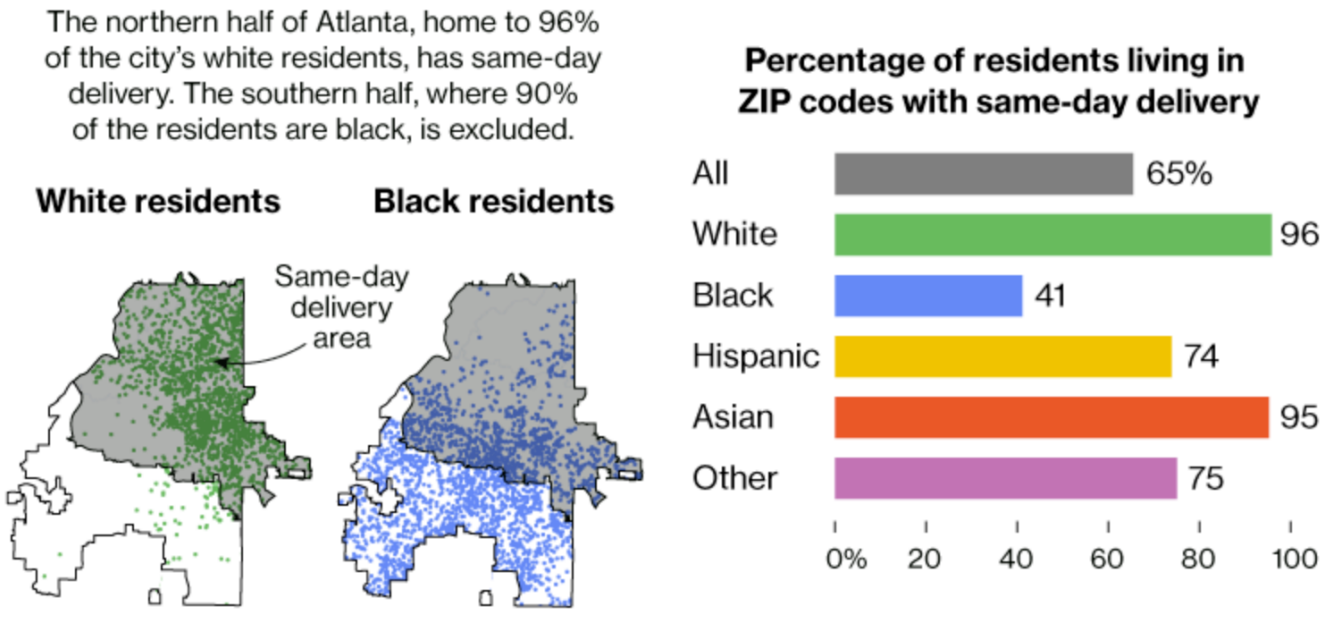
Comparing images of soap from different cultures in the Dollar Street dataset, from DeVries et al.

- Above: the **domains** of training data the target population differ.
- Mismatched domains are common in ML due to the use of **convenience datasets** or **benchmark datasets**.

Proxy Features

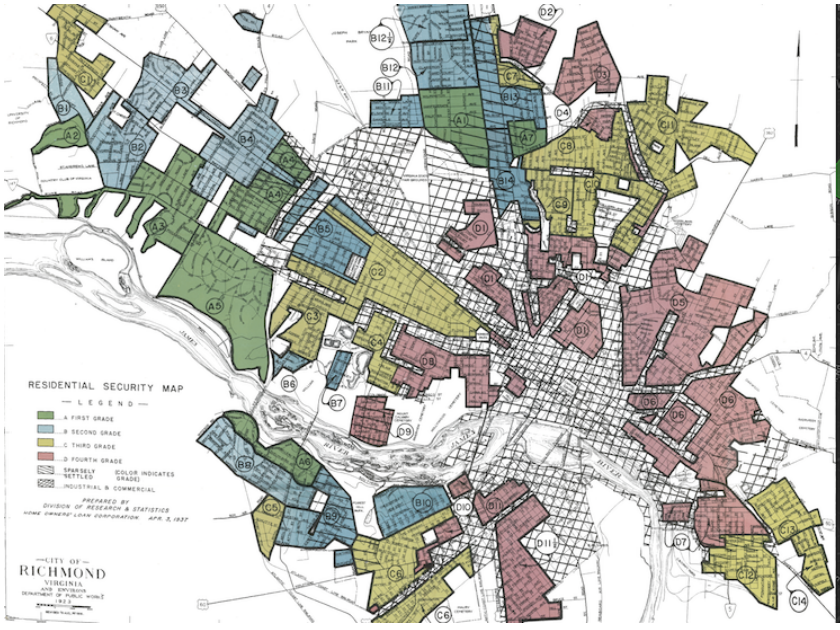
- **Proxy features** are features that measure **protected traits** such as race, ethnicity, gender identity, sexual orientation, age, etc.
- Example: Even if without race as a predictor, disparate impact could come from variables such as postal code.
- Postal codes may end up being a proxy for race due to segregation of neighborhoods in the United States.

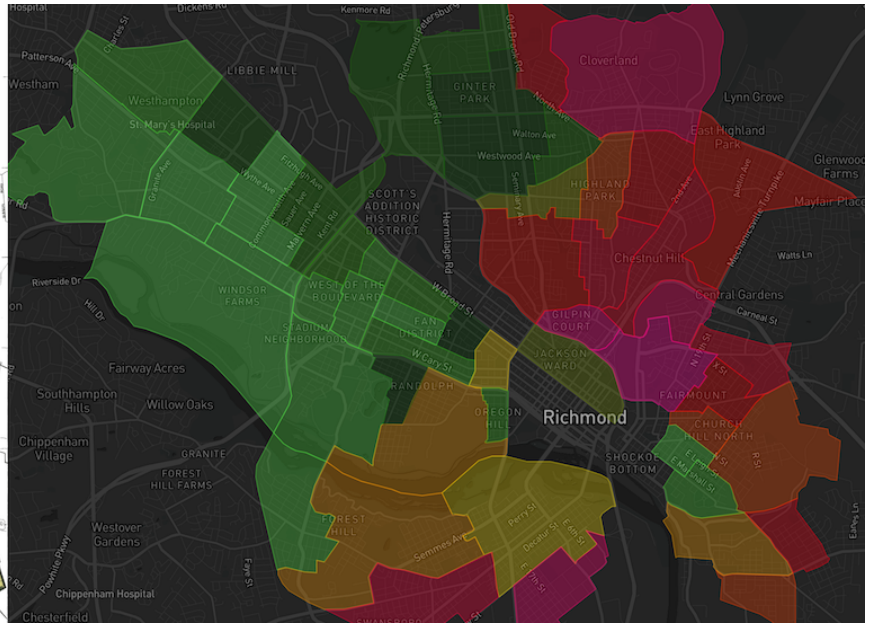
Example: Amazon Same-Day Prime Deliveries



- Amazon excluded neighborhoods based on whether a particular zip code had:
  1. A sufficient number of Prime members.
  2. Was near a warehouse.
  3. Was a place drivers were willing to visit.
- While race was not explicitly considered in this decision, Blacks were disproportionately excluded from this service.

Social/Structural Factors

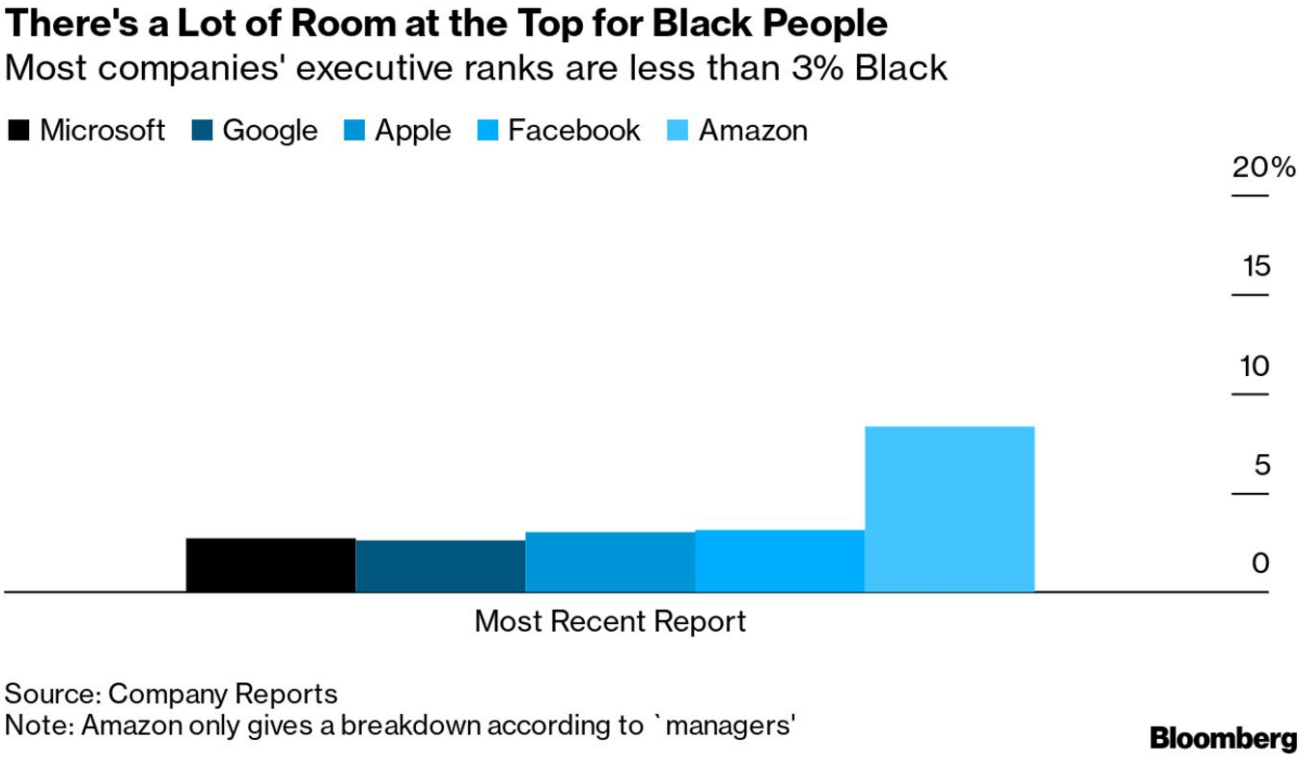




Source: Digital Scholarship Lab, University of Richmond.

- Left: 1930s Redlining map used to determine access to housing loans. Minority groups more likely to receive D or C grades.
- Right: Social Vulnerability Index (SVI), a Center for Disease Control (CDC) measure for the capacity to prepare for, respond to, and recover from human and natural disasters.

Social/Structural Factors



Decision-makers at tech companies are predominantly from groups *least likely* to be impacted by algorithmic bias. Source: Bloomberg

Social/Structural Factors

The Guardian

Google

Timnit Gebru, an eminent Black scientist, says she was fired last year in clash over research on marginalized groups

What happens when ethical concerns threaten the business interests of tech companies?

Social/Structural Factors

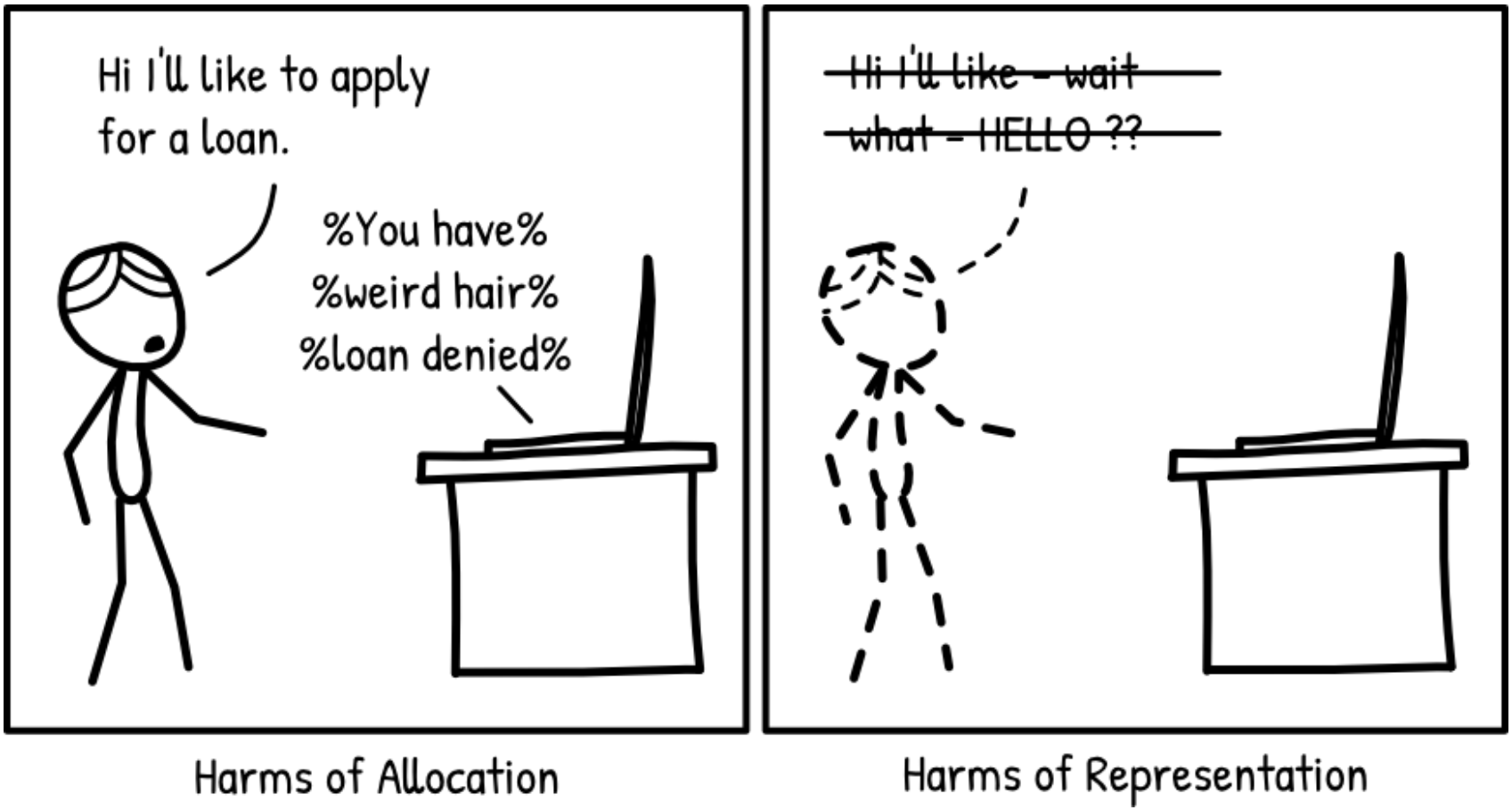
Ryan Saavedra ✓
@RealSaavedra

Socialist Rep. Alexandria Ocasio-Cortez (D-NY) claims that algorithms, which are driven by math, are racist

Full thread here

Harms from Algorithmic Bias

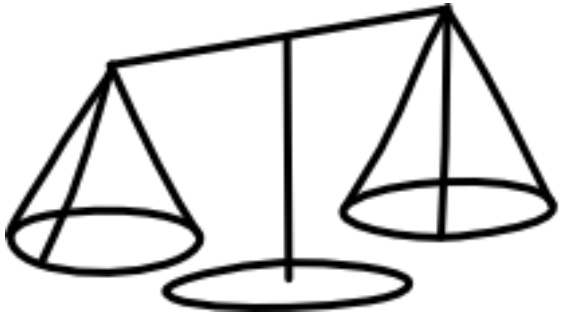
Types of Harm from Bias

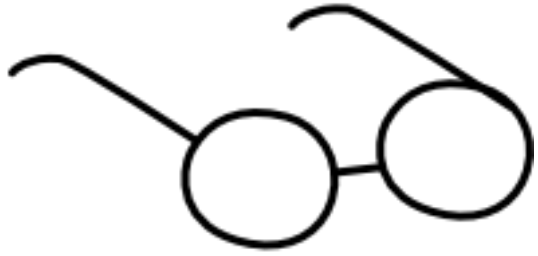


- **Harms of allocation:** opportunities and resources are unfairly allocated due to algorithmic intervention.
- **Harms of representation:** algorithms produce representations that are discriminatory

Image: https://machinesgonewrong.com/bias_i/

Types of Harm from Bias

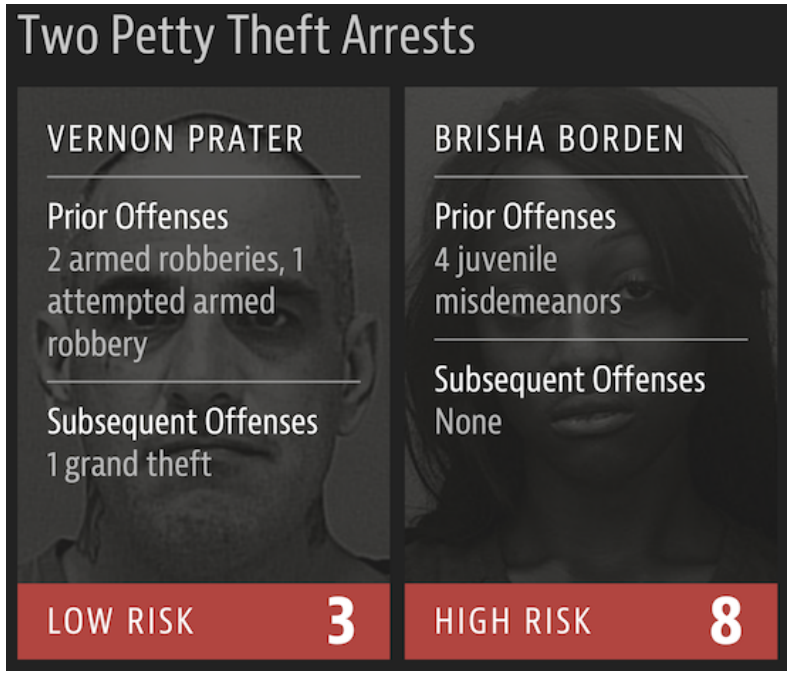




Harms of Allocation	**Harms of Representation**
Immediate	Long term
Easily quantifiable	Difficult to formalize
Discrete	Diffuse
Transactional	Cultural

Table: Kate Crawford, NIPS 2017 Keynote, Image: https://machinesgonewrong.com/bias_i/

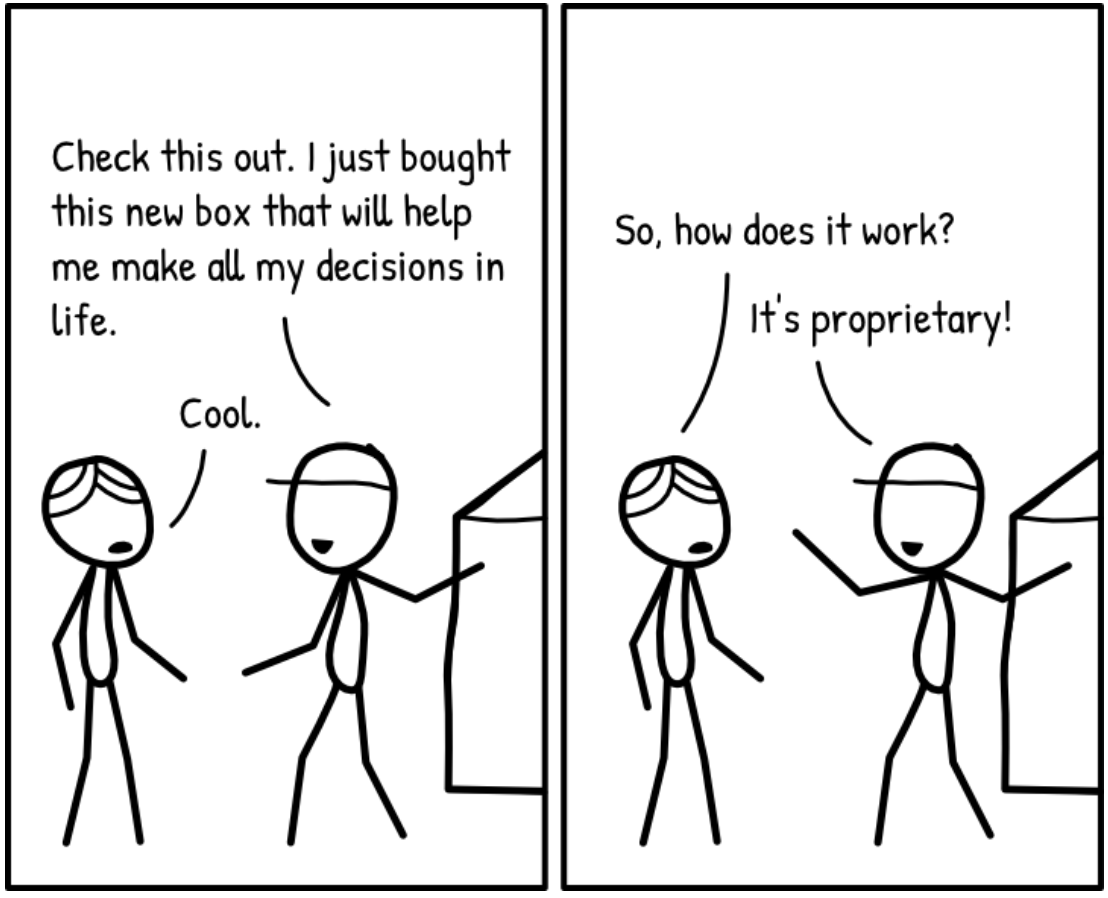
Harms of Allocation: COMPAS



COMPAS, a predictive policing system, perpetuates racial bias even though race is not an input variable. Source: "Machine Bias," Propublica.

- COMPAS, a predictive policing software used across the United States to predict future criminals is biased against blacks; the systems predictions about the risks of reoffending are given to judges during criminal sentencing.
- Read the report here; download the data here

Challenges in Addressing Bias: Disclosure



Systems like COMPAS are not transparent and they do not have to disclose how their systems work. We only find out about bias due to excellent investigative reporting. Source: Machines Gone Wrong

Harms of Representation

Joy Buolamwini, researcher of algorithmic fairness, needed to wear a white mask to be recognized by a common facial reconition tool.

Bias in Facial Recognition

- The main benchmark datasets for facial analysis are overwhelmingly light-skinned; Microsoft, IBM, Face++ systems are biased (Buolamwini and Gebru, 2018)
  - Gender: Error rate for female faces 8.1%-20.6% greater than male faces
  - Skin tone: Error rate for faces with darker skin tones 11.8-19.2% greater than faces with lighter skin tones
- At least 1/4 of US law enforcement agencies have access to facial recognition technology (Garvie et. al., 2016).
- In response to Black Lives Matter protests, Amazon, Microsoft, and IBM halted sale of facial recognition tech to US law enforcement (source).

Harms of Representation → Harms of Allocation

MICHIGAN STATE POLICE

INVESTIGATIVE LEAD REPORT

LAW ENFORCEMENT SENSITIVE

THIS DOCUMENT IS NOT A POSITIVE IDENTIFICATION. IT IS AN INVESTIGATIVE LEAD ONLY AND IS NOT PROBABLE CAUSE TO ARREST. FURTHER INVESTIGATION IS NEEDED TO DEVELOP PROBABLE CAUSE TO ARREST.

BID DIA Identifier: BID-39641-19	Requester: CA Yager, Rathe
Date Searched: 03/11/2019	Requesting Agency: Detroit Police Department
Digital Image Examiner: Jennifer Coulson	Case Number: 1810050167 File Class/Crime Type: 3000

Probe Image	Investigative Lead
	

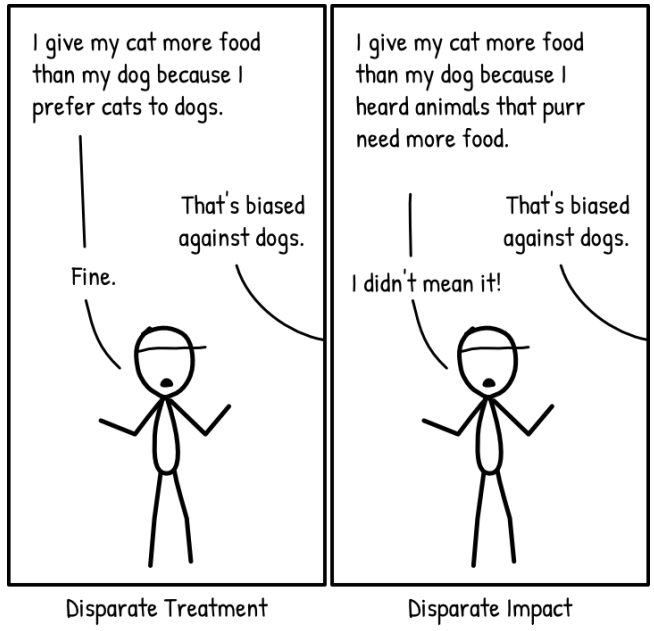
Left: theft suspect in Detroit, Right: driver's license photo of a false hit from a facial recognition algorithm. The individual was arrested and detained for 30 hours due to this false hit. Source: New York Times

Identifying Bias

Guidelines for Identifying Bias

- Use **interpretable** models when possible:
  - Easier to diagnose problems, identify how models are making decisions
  - Unfortunately explainable models are less predictive.
- Make use of "fairness forensics" to measure discrimination in algorithmic decisions
  - Group-level comparisons: Is there statistical parity among groups? (e.g. are Blacks and Whites misclassified at the same rate?)
  - Individual-level comparisons: Do two otherwise similar individuals in different groups (race, gender, age, etc.) have similar outcomes?

Fairness Forensics: Guiding Legal Concepts



- **Disparate treatment:** predicted outcome is (intentionally) based on protected characteristics; protected characteristics are features.
  - Subject to highest of legal penalties.
  - Removing the protected characteristic from the feature set often does not address problems of disparate impact.
- **Disparate impact:** predicted outcomes depend on values of protected characteristics (though they are not features).

Disparate Impact Guidelines

- Adopted in 1978 by the US Federal Government; standard used in US labor law
- Disparate impact becomes **"adverse impact"** when the selection rate (e.g. for job candidates) for a protected group is less than 80% of the group with the highest selection rate, e.g.:

Group	**Applied**	**Hired**	**Selection Rate**
Blacks	15	3	20%
Whites	100	40	40%

$$\frac{20\%}{40\%} = 50\% < 80\%$$

Criteria for Evaluating Bias

- Let's imagine a population of individuals of different races $R \in \{\text{Black}, \text{White}\}$ (a protected characteristic), features X, and outcomes:

$$Y \in \begin{cases} 0 & \text{if defaulted on loan.} \\ 1 & \text{if loan paid off.} \end{cases}$$

- Goal: evaluate fairness (w.r.t. race R) of a learning algorithm $g(X)$ that predicts credit-worthiness $\hat{Y} \in \{0, 1\}$ based on observations of past loan outcomes.
- We can make use of three main diagnostic criteria for algorithmic bias:
  1. Demographic Parity (Independence): $\hat{Y} \perp R$
  2. Equal Opportunity (Separation): $\hat{Y} \perp R | Y$
  3. Equalized Odds (Sufficiency): $Y \perp R | \hat{Y}$
- Note: these measures do not have agreed-upon names; I use vernacular terms and their technical counterparts in parenthesis.

Demographic Parity (Independence): $\hat{Y} \perp R$

- Classifier makes positive predictions for each group at the same rate as the entire population.
- Formally, to satisfy **independence**, R must be statistically independent to the prediction $\hat{Y}$:

$$P(\hat{Y} = 1 | R = \text{White}) = P(\hat{Y} = 1 | R = \text{Black})$$

- Practically speaking, predicted positives rates calculated on $TE_{R=\text{Black}}$ and $TE_{R=\text{White}}$ should be equal.

Equal Opportunity (Separation): $\hat{Y} \perp R | Y$

- The **true positive rates/recall** and **false positive rates** in each group are the same as entire population.
- Formally, to satisfy **separation**, R must be statistically independent to the prediction $\hat{Y}$ given the target value Y:

$$P(\hat{Y} = 1 | Y = 1, R = \text{White}) = P(\hat{Y} = 1 | Y = 1, R = \text{Black}),$$

$$P(\hat{Y} = 1 | Y = 0, R = \text{White}) = P(\hat{Y} = 1 | Y = 0, R = \text{Black})$$

- Practically speaking, recall calculated on $TE_{R=\text{Black}}$ and $TE_{R=\text{White}}$ should be equal.

Equalized Odds (Sufficiency): $Y \perp R | \hat{Y}$

- Classifier should have both equal true positive rates (precision) and false positive rates on a protected population group as those of the entire population.
- To satisfy **sufficiency**, R must be statistically independent to the target value Y given the prediction $\hat{Y}$

$$P(Y = 1 | \hat{Y} = 1, R = \text{White}) = P(Y = 1 | \hat{Y} = 1, R = \text{Black})$$

- Practically speaking, precision calculated on $TE_{R=\text{Black}}$ and $TE_{R=\text{White}}$ should be equal.

Adding Slack

- We might want to relax our fairness criteria, allowing small deviations from perfect fairness.
- Example, for demographic parity, our fairness objective with slack is:

$$P(\hat{Y} = 1|R = \text{White}) \geq P(\hat{Y} = 1|R = \text{Black}) - \varepsilon$$

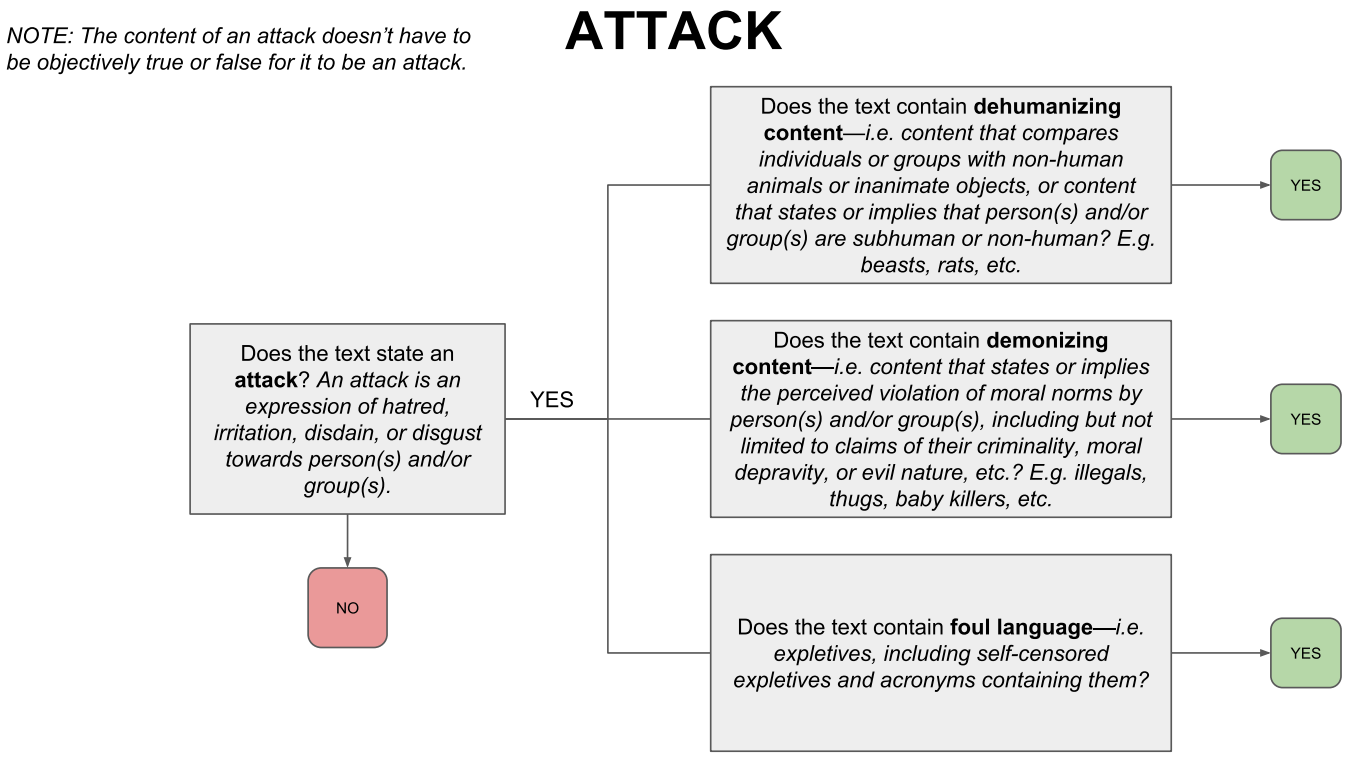
- Circling back to the legal definition of disparate impact, we may choose to formalize the 80% rule by setting $\varepsilon = .2$.

Correcting Bias (Before/During Data Collection)

Ensure Adequate Representation

- When collecting data, we need to be clear what the target population looks like.
  - Underrepresenting one group could have serious consequences.
  - To ensure fairness, we may want to **over-sample** underrepresented groups.
- If our goals are predictive accuracy and fairness, we may want to stratify our sample by group to ensure our model has enough information about all groups.
- Do we have adequate variation in feature values across groups?
  - Are we capturing all relevant sub-populations within each group?
  - Example: sampling minority groups from predominantly wealthier neighborhoods may result in bias affecting those living in poor neighborhoods, even if we oversample.

Train/Audit Labelers



An example of a coding diagram for labelers used to identify toxic speech.

- If an ML algorithm relies on human-labeled data:
  1. **Train** labelers to adhere to a clear and exhaustive codebook.
  2. Regularly **audit** labelers to ensure adherence to codebook instructions.

Example: Train/Audit Labelers

- Most social networks hire humans to moderate banned content.
- Labelers are outsourced to countries like the Phillippines.
- Conceptions of "sexually explicit" vary by culture, religion, etc.
- LGBT individuals could be disparately impacted through moderation of:
  1. Posts using words like "fag" in a reclaimed, positive sense.
  2. Posts of queer people showing affection, labeled as "sexually explicit" despite not meeting the definition set out in the codebook.
- Possible solutions:
  1. Track coder reliability by labeling some observations several times (more on this here). Examine disagreements, intervene when necessary.
  2. Allow user feedback to alert auditors when the coder gets it wrong.

Correcting Bias (After Data Collection)

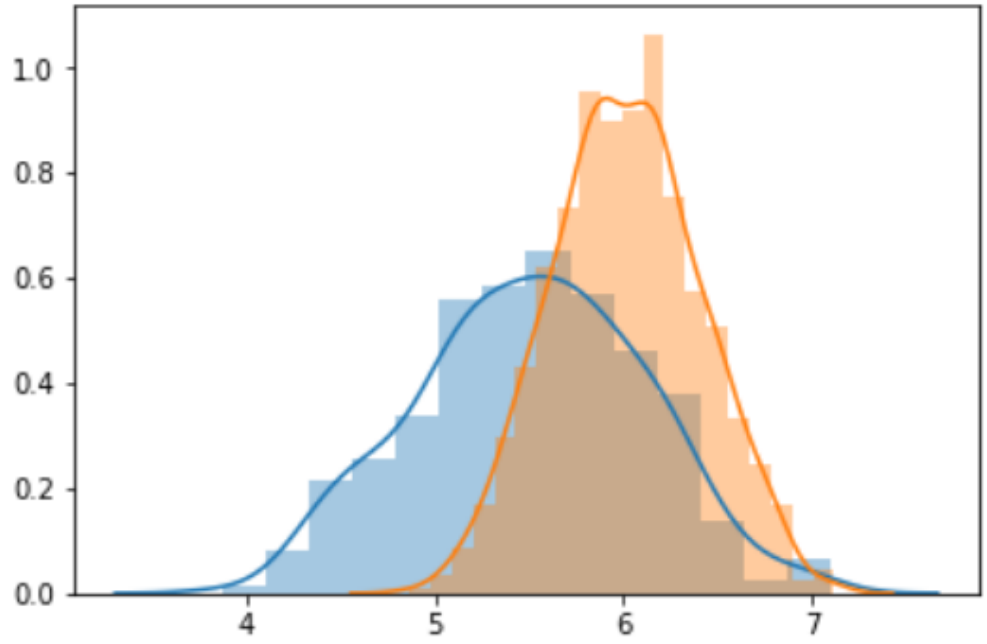
Techniques for Correcting Disparate Impact

1. **Pre-processing** the data, adjusting feature values to remove disparate impact.
2. **Modifying the loss function** used during training to optimize fit *and* a fairness criterion.
3. **Post-processing** the predicted outcomes of the model to remove disparate impact.

Pre-processing

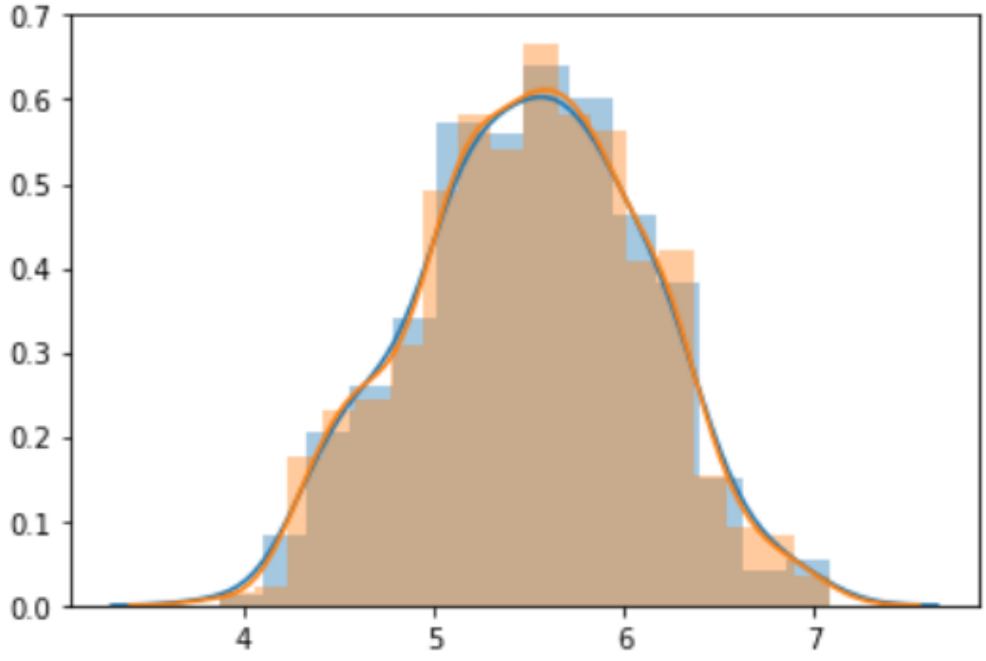
- Goal: Learn a new feature representation W from X that removes information correlated to a protected characteristic while preserving enough information from X for a model to be sufficiently accurate.
- One example of pre-processing is outlined in "Certifying and removing disparate impact" by M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. (example to follow)
- Two problems:
  1. There are several dimensions of bias (race, gender, sexual orientation, etc.). Adjusting for all dimensions is challenging, especially if accuracy is important.
  2. Transforming raw data according to protected characteristics can be illegal.

Pre-processing: A Toy Example



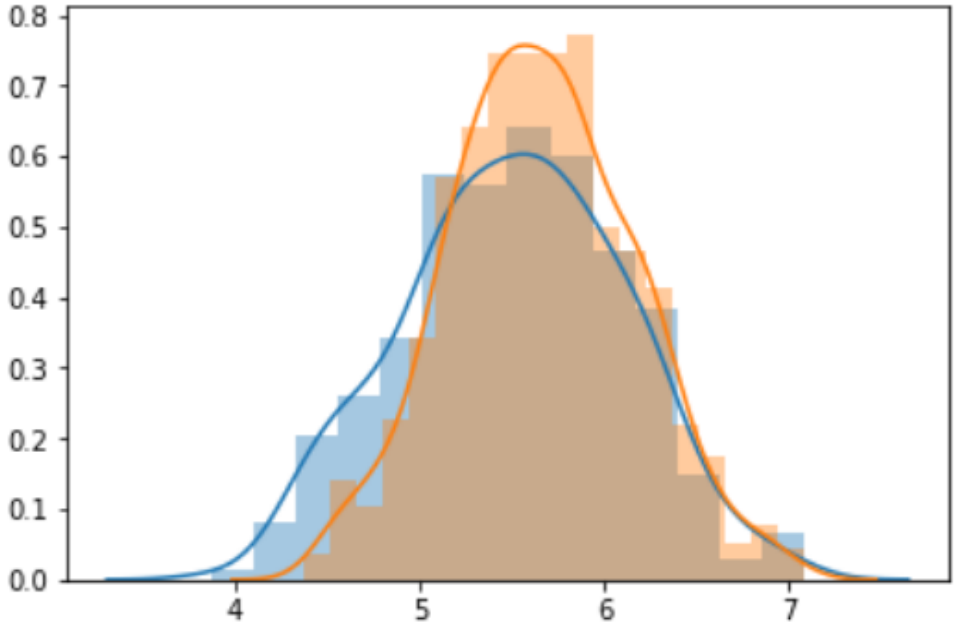
Two densities: an unprivileged group, Blue, and a privileged group, Orange. The x-axis represents values of a single feature (see source code).

Pre-processing: A Toy Example



The values of this feature are changed for all observations with the goal of group densities being as similar as possible. In this example, the `repair` parameter is set to 1 (see source code).

Pre-processing: A Toy Example



In this example, the `repair` parameter is set to .8, in line with legal definitions of disparate impact (see source code).

Modifying the Loss Function

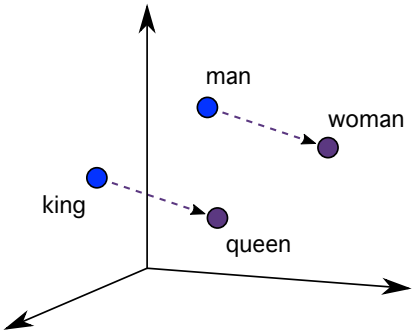
$$\begin{aligned} \underset{\boldsymbol{\theta}}{\text{minimize}} \quad & \{\mathcal{L}(\boldsymbol{\theta})\} \\ \text{subject to} \quad & \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\; d_{\boldsymbol{\theta}}(\mathbf{x}_i) \leq \mathbf{c} \\ & \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\; d_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq -\mathbf{c} \end{aligned}$$

- $\mathbf{z}$ denotes values of sensitive characteristics.
- $d_{\theta}(\mathbf{x}_i)$ is the signed distance between $\mathbf{x}_i$ and the decision boundary.
- $\frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\; d_{\theta}(\mathbf{x}_i)$ represents the covariance between sensitive attributes, $\{\mathbf{z}_i\}_{i=1}^{N}$, and the signed distance to the decision boundary, $\{d_{\theta}(\mathbf{x}_i)\}_{i=1}^{N}$
- $\mathbf{c}$ trades off fairness and accuracy; $\mathbf{c}$ close to zero corresponds to more fairness, possibly at the expense of accuracy.

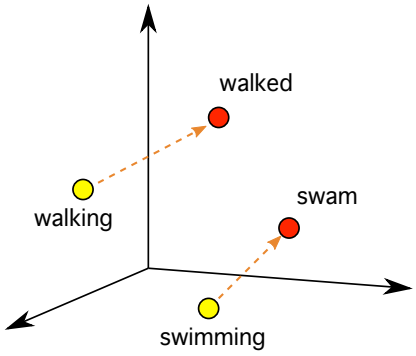
Post-processing

- Goal: Manipulate $\hat{y}$ to achieve a fairness criterion while maintaining accuracy
- One example of post-processing is outlined in Bolukbasi, Tolga, et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." Advances in Neural Information Processing Systems. 2016. (example to follow)

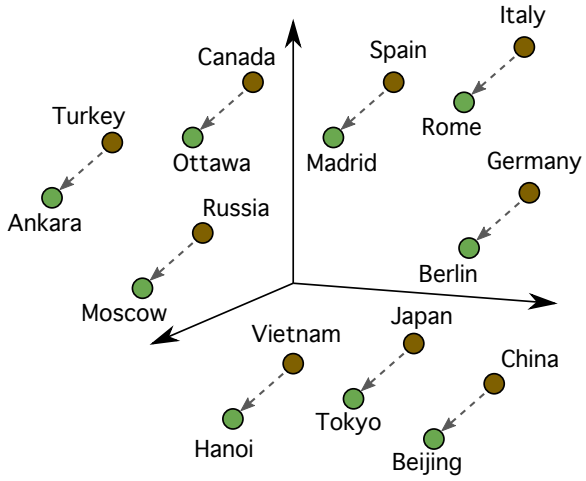
Post-processing Example: Word Embeddings



Semantic (Male-Female)



Syntactic (Verb Tense)



Semantic (Country-Capital)

- **Word embeddings** represent words as fixed-dimensional vectors such that:
  - similar words (e.g. good, great) $\rightarrow$ similar vector representations
  - vector directions encode **syntactic** and **semantic** information.
- Vector operations can be used to construct **analogies**:

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

- Embeddings are often inputs to ML models of natural language.

Latent Bias in Language

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

Word embeddings also encode problematic biases from natural language data.

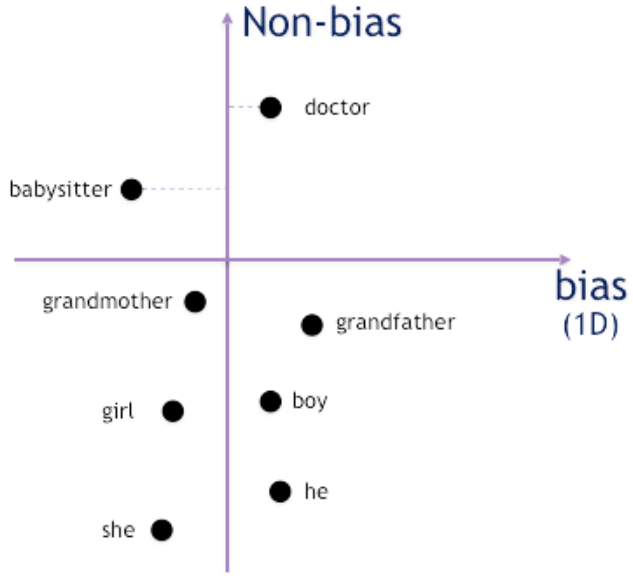
Latent Bias in Language

occupation	bias	occupation	bias	occupation	bias	occupation	bias
maid	59.2	librarian	20.1	undertaker	-73.4	captain	-53.4
waitress	52.5	obstetrician	16.9	janitor	-62.3	announcer	-51.1
midwife	50.9	secretary	13.7	referee	-60.7	architect	-50.7
receptionist	50.2	socialite	12.1	plumber	-58	maestro	-50.6
nanny	47.7	therapist	10.2	actor	-56.9	drafter	-46.7
nurse	45.4	manicurist	10.1	philosopher	-56.2	usher	-46.6
midwives	43.8	hairdresser	9.7	barber	-55.4	farmer	-45.4
housekeeper	36.6	stylist	8.6	umpire	-54.3	broadcaster	-45.2
hostess	32	homemaker	6.9	president	-54	engineer	-45.1
gynecologist	31.6	planner	5.8	coach	-53.8	magician	-44.8

Gender bias scores for the Universal Sentence Encoder embedding model. Occupations with the highest female-biased scores (left) and the highest male-biased scores (right)

Source: Google Developer Blog

Identifying and Neutralizing Bias

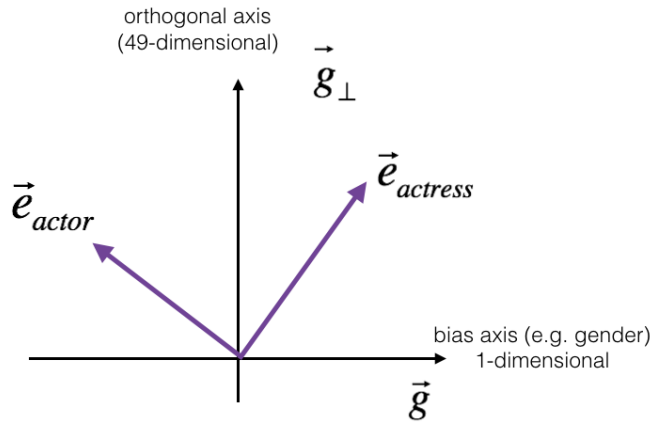


Distribution of definitional and non-definitional words along the bias dimension.

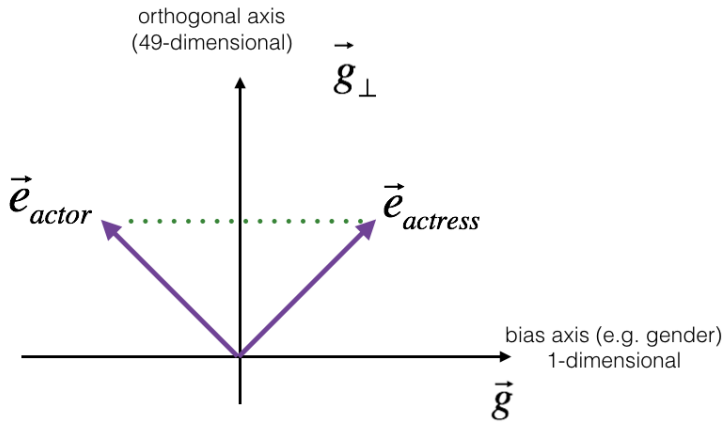
Bolukbasi et. al. 2016 outlines a debiasing algorithm:

1. Identify the direction of bias (e.g. the gender subspace)
2. Neutralize: For all non-definitional words, project to get rid of bias.
3. Equalize pairs of words we want to differ only through the gender property (e.g. he, she)

Debiasing Embeddings (Bolukbasi et. al.)



before equalizing,
"actress" and "actor" differ
in many ways beyond the
direction of $\vec{g}$



after equalizing,
"actress" and "actor" differ
only in the direction of $\vec{g}$, and further
are equal in distance from $\vec{g}_\perp$

Debiasing applied to word pairs "actress" and "actor."

- **Equalization** makes sure that a particular pair of words are equi-distant from the orthogonal non-bias vector
- This ensures that the words actress and actor are equidistant from any other word that has been neutralized (e.g. babysitter).

Multi-class Debiasing (Manzini et. al. 2019)

Racial Analogies	
black → homeless	caucasian → servicemen
caucasian → hillbilly	asian → suburban
asian → laborer	black → landowner
Religious Analogies	
jew → greedy	muslim → powerless
christian → familial	muslim → warzone
muslim → uneducated	christian → intellectually

Examples of religious and racial biases in embeddings trained on the Reddit corpus.

- There are, however, multiple bias axes (e.g. race, religion, sexual orientation, gender identity):

$$\overrightarrow{\text{black}} - \overrightarrow{\text{criminal}} \approx \overrightarrow{\text{white}} - \overrightarrow{\text{police}}$$

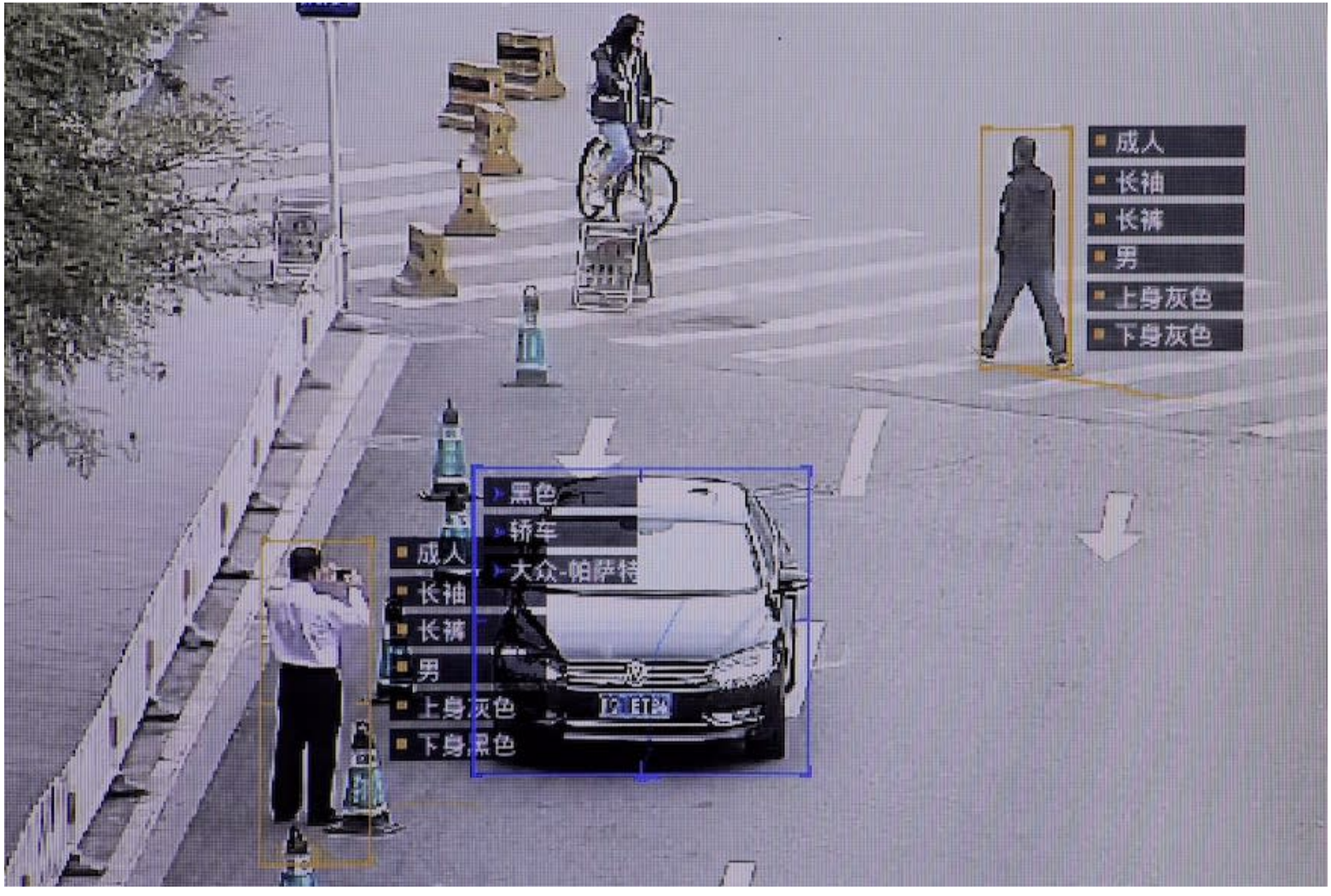
- Manzini et. al. 2019 generalizes the approach in Bolukbasi et. al. 2016 to neutralize along several of these bias axes.

Machine Learning as a Repressive Tool

Three Painful Truths (Diebert)

1) Social media is built around personal-data surveillance, with products ultimately designed to spy on us in order to push advertising in our direction.
2) We have consented to this, but not entirely wittingly: Social media are designed as addiction machines, expressly programmed to draw upon our emotions.
3) Attention-grabbing algorithms underlying social media also propel authoritarian practices that aim to sow confusion, ignorance, prejudice, and chaos, thereby facilitating manipulation and undermining accountability. Moreover, the fine-grained surveillance that companies perform for economic reasons is a valuable proxy for authoritarian control.

Serving digital authoritarianism



Sensetime, a Hong Kong-based company provides real time crowd analytics and facial recognition to mainland Chinese domestic security bureaucracies.

Q&A Discussion

Types of Harm from Bias

1. What are some examples of representative harm in ML?
2. What are some examples of allocative harm in ML?

The Politics of Classification

1. What are the potential harms caused by government systems that classify citizens into groups? How can we mitigate them?
2. How do we weigh the costs and benefits of ML technology that might prove harmful to democracy, freedom of speech, etc.?
3. Is all knowledge production and all technological progress necessarily good? Where do we draw the line?

Quiz Review and Discussion

Addressing Bias

A search for "CEO" in Google.

1. Arguably, the search results above may be representative of CEO demographics. How are we to decide upon the appropriate representation in this example?
2. What are some limitations of technical approaches to dealing with bias?