

Data Bias and Fair ML

Applied Machine Learning
Derek Hoiem

Dall-E

Today’s lecture

- Human, Data, and Algorithm Bias; and what to do about it
  - Adapted from a 2019 guest lecture by Margaret Mitchell, with permission

- Fairness and Justice
  - Adapted from Pietro Perona IVCSS 2024 talk, with permission

What is bias?

- Bias in ML (“bias-variance trade-off”): the model is too limited to fully encode the true distribution, e.g. linear regression is used for a problem that requires non-linear regression
- Bias in Statistics: the expected value of your estimate of the statistic differs from the true value (e.g., estimating variance by dividing by n instead of n−1)
- Bias in Data: the distribution of the training data is more limited or different than in deployment
- Cognitive Bias: wrong thinking, e.g. confirmation bias, recency bias, optimism bias
- Algorithmic Bias: the use of algorithms to apply unjust, unfair, or prejudicial treatment of people related to protected attributes (e.g., race, age, sexual orientation, religion, gender)

What’s the big deal?

True story

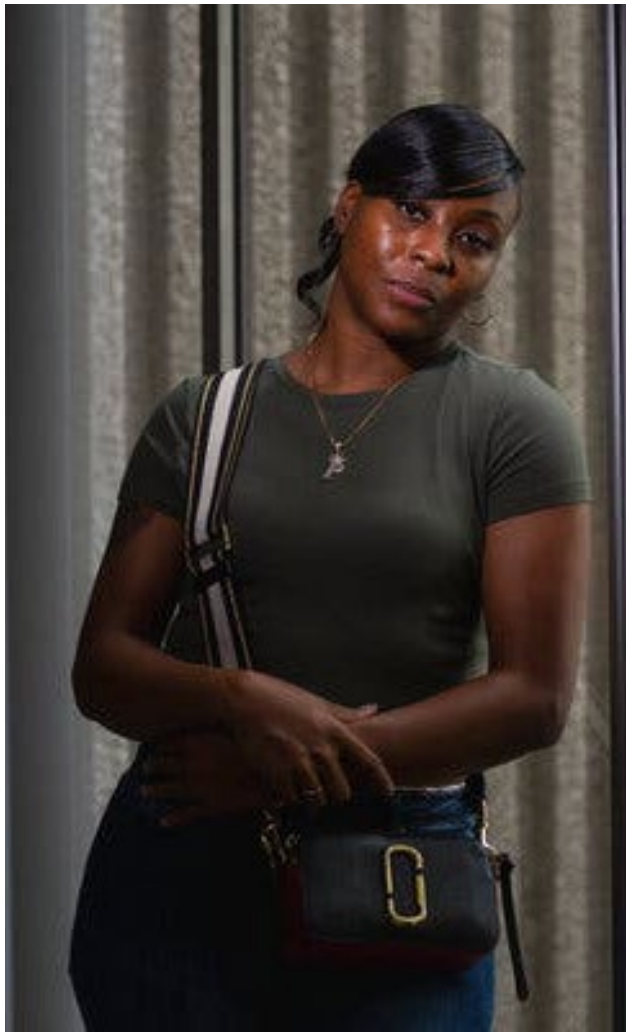
Face recognition identifies black person as matching a suspect in armed robbery & carjacking

Victim then picks her from a photo lineup that uses her 8-year-old mugshot

Black person was 8 months pregnant, with no connection to circumstances of the crime

Police arrested and charged her, trusting the “objective” ID without further checks

Turns out to be a false arrest — 6th known case from face recognition; all six were Black



How does this happen?

- Humans are biased
- Algorithms consume human data and make biased judgements, or performance varies for different groups
- Humans trust computers to be correct and unbiased

What do you see?



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

- Bananas



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

- Bananas
- Stickers



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

- Bananas
- Stickers
- Dole Bananas



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store

...We don't tend to say

Yellow Bananas



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

Green Bananas

Unripe Bananas



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

Ripe Bananas

Bananas with spots



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

- Ripe Bananas
- Bananas with spots
- Bananas good for banana bread



CC BY / Margaret Mitchell / m-mitchell.com

What do you see?

Yellow Bananas

Yellow **is prototypical for bananas**



CC BY / Margaret Mitchell / m-mitchell.com

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

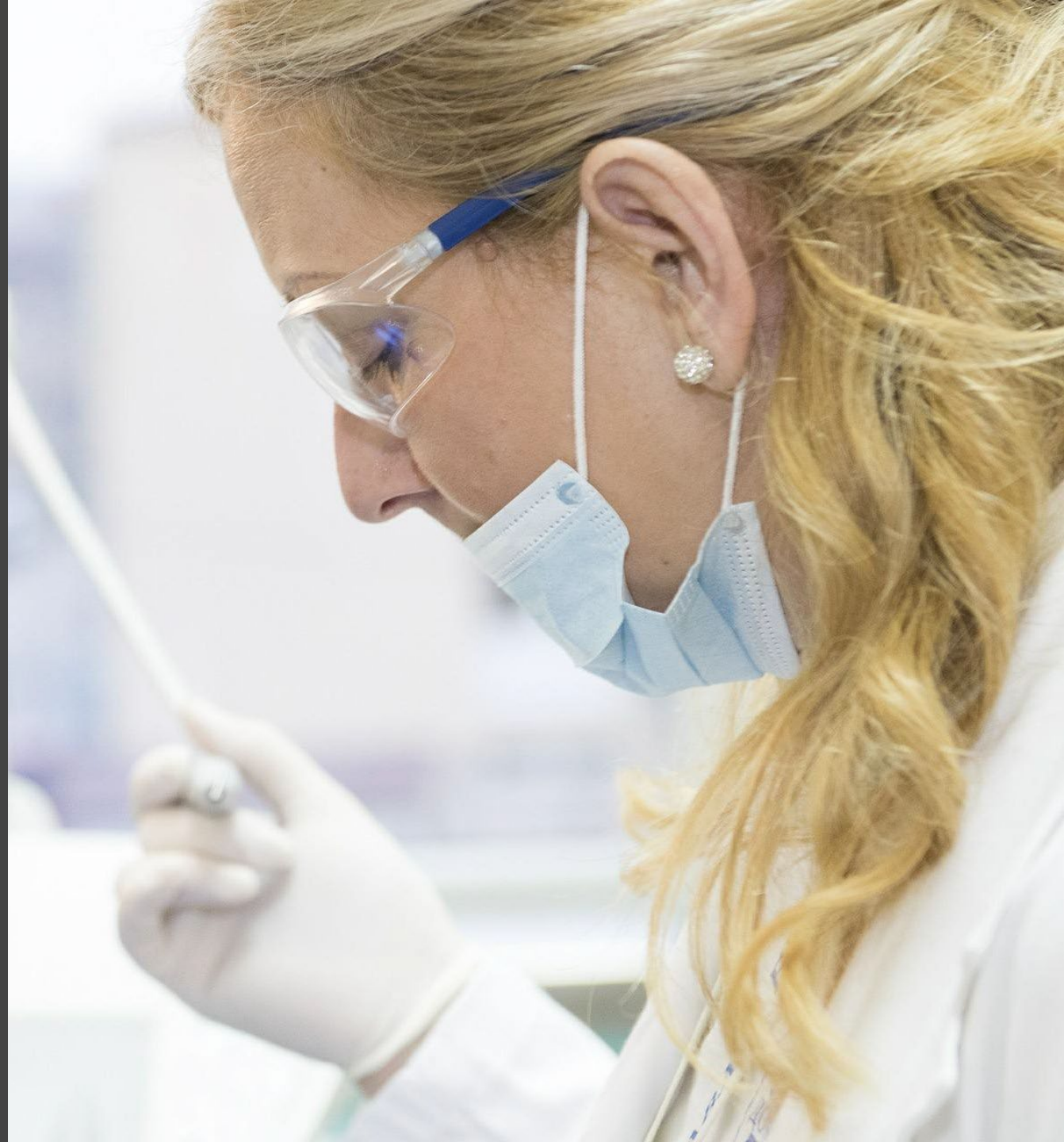


CC BY / Margaret Mitchell / m-mitchell.com

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

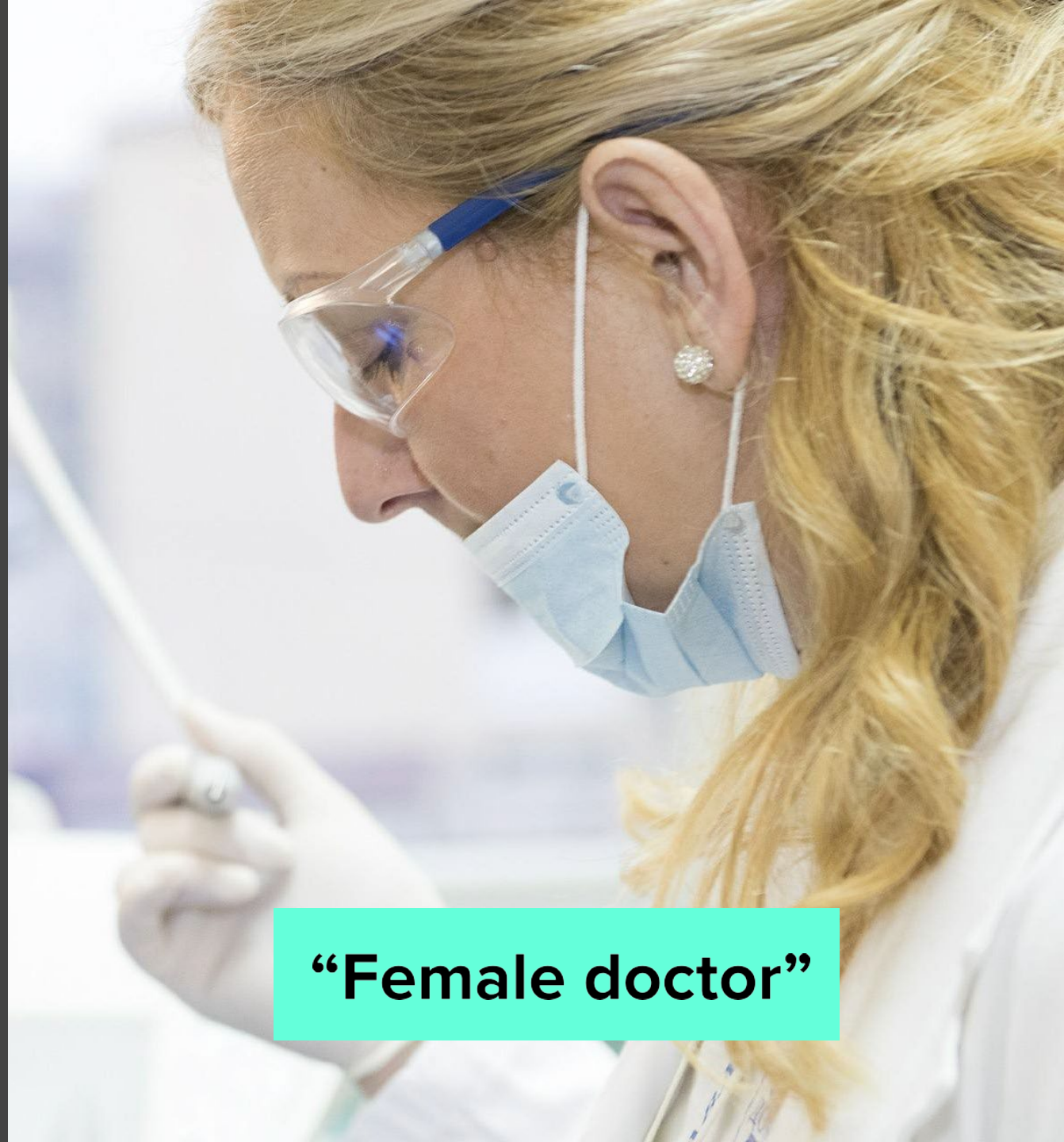
How could this be?



A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?



CC BY / Margaret Mitchell / m-mitchell.com


“Doctor”
“Female doctor”

The majority of test subjects overlooked the possibility that the doctor is a she - including men, women, and self-described feminists.

Wapman & Belle, Boston University

CC BY / Margaret Mitchell / m-mitchell.com

World learning from text

Gordon and Van Durme, 2013

Word	Frequency in Google Web corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

CC BY / Margaret Mitchell / m-mitchell.com

Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals

CC BY / Margaret Mitchell / m-mitchell.com

Co-occurrence biases in training data affect ML models

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

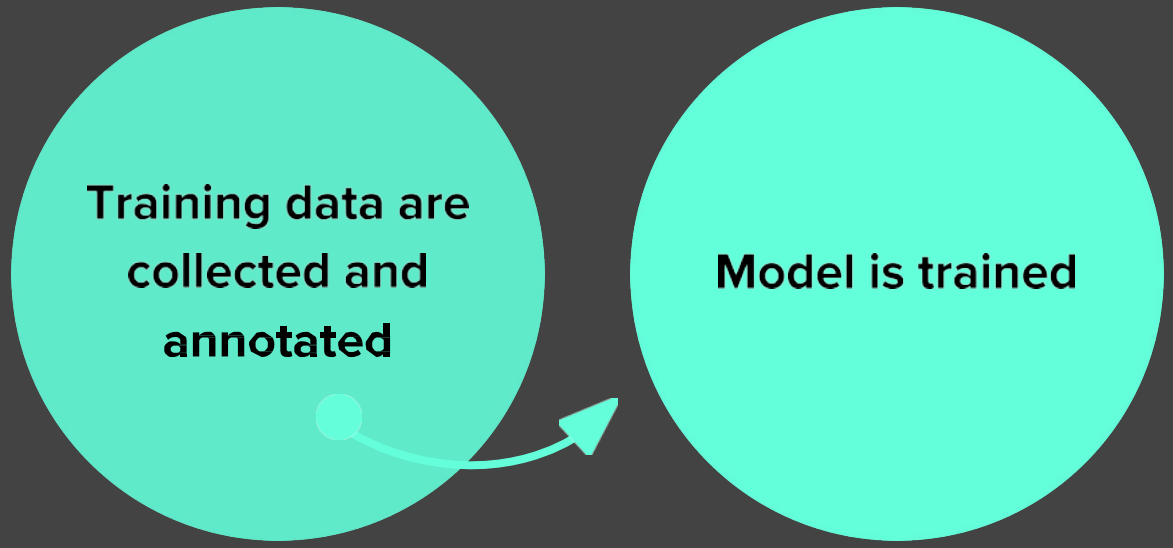


[GPT-4V]: *In the sequence of images provided, we observe two men engaged in what appears to be a playful and humorous interaction on a set of outdoor stairs leading to a building entrance. In the first image, one man is seen holding a long, flexible object, possibly a strip of metal or rubber, and appears to be playfully threatening to strike the other man, who is ascending the stairs and holding a similar object. The second image captures a moment of anticipation, with the first man poised to 'attack' and the second man looking back with a smile, suggesting a friendly and joking atmosphere. The third image shows the first man in motion, as if he has just attempted to playfully hit the second man, who is now halfway through the doorway, still participating in the jest. In the final image, the second man has turned around completely, facing the first man with a grin, indicating that the mock confrontation has ended amicably, and both men seem to be enjoying the interaction. The overall mood is light-hearted, and the episode seems to be a moment of camaraderie and fun between the two individuals.*

https://arxiv.org/pdf/2401.10529.pdf

Training data are collected and annotated

CC BY / Margaret Mitchell / m-mitchell.com



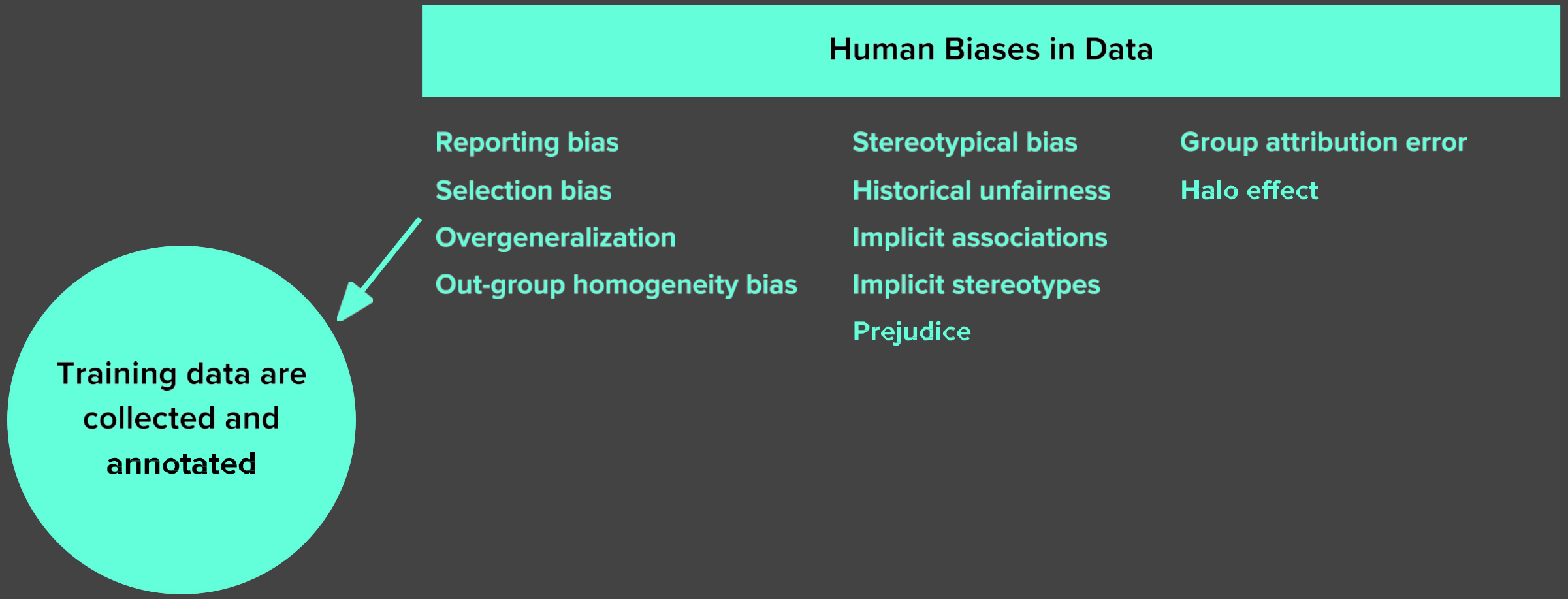
CC BY / Margaret Mitchell / m-mitchell.com



CC BY / Margaret Mitchell / m-mitchell.com



CC BY / Margaret Mitchell / m-mitchell.com



CC BY / Margaret Mitchell / m-mitchell.com



Human Biases in Data

- Reporting bias
- Selection bias
- Overgeneralization
- Out-group homogeneity bias
- Stereotypical bias
- Historical unfairness
- Implicit associations
- Implicit stereotypes
- Prejudice
- Group attribution error
- Halo effect

Human Biases in Collection and Annotation

- Sampling error
- Non-sampling error
- Insensitivity to sample size
- Correspondence bias
- In-group bias
- Bias blind spot
- Confirmation bias
- Subjective validation
- Experimenter's bias
- Choice-supportive bias
- Neglect of probability
- Anecdotal fallacy
- Illusion of validity

CC BY / Margaret Mitchell / m-mitchell.com

Data

Reporting bias: What people share is not a reflection of real-world frequencies

Selection Bias: Selection does not reflect a random sample

Out-group homogeneity bias: People tend to see outgroup members as more alike than ingroup members when comparing attitudes, values, personality traits, and other characteristics

Confirmation bias: The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses

Interpretation

Overgeneralization: Coming to conclusion based on information that is too general and/or not specific enough

Correlation fallacy: Confusing correlation with causation

Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation

More at: https://developers.google.com/machine-learning/glossary/

CC BY / Margaret Mitchell / m-mitchell.com



Biases in Data

Biases in Data

Selection Bias: Selection does not reflect a random sample



CREDIT

© 2013–2016 Michael Yoshitaka Erlewine and Hadas Kotek

CC BY / Margaret Mitchell / m-mitchell.com

Biases in Data

Out-group homogeneity bias: Tendency to see outgroup members as more alike than ingroup members







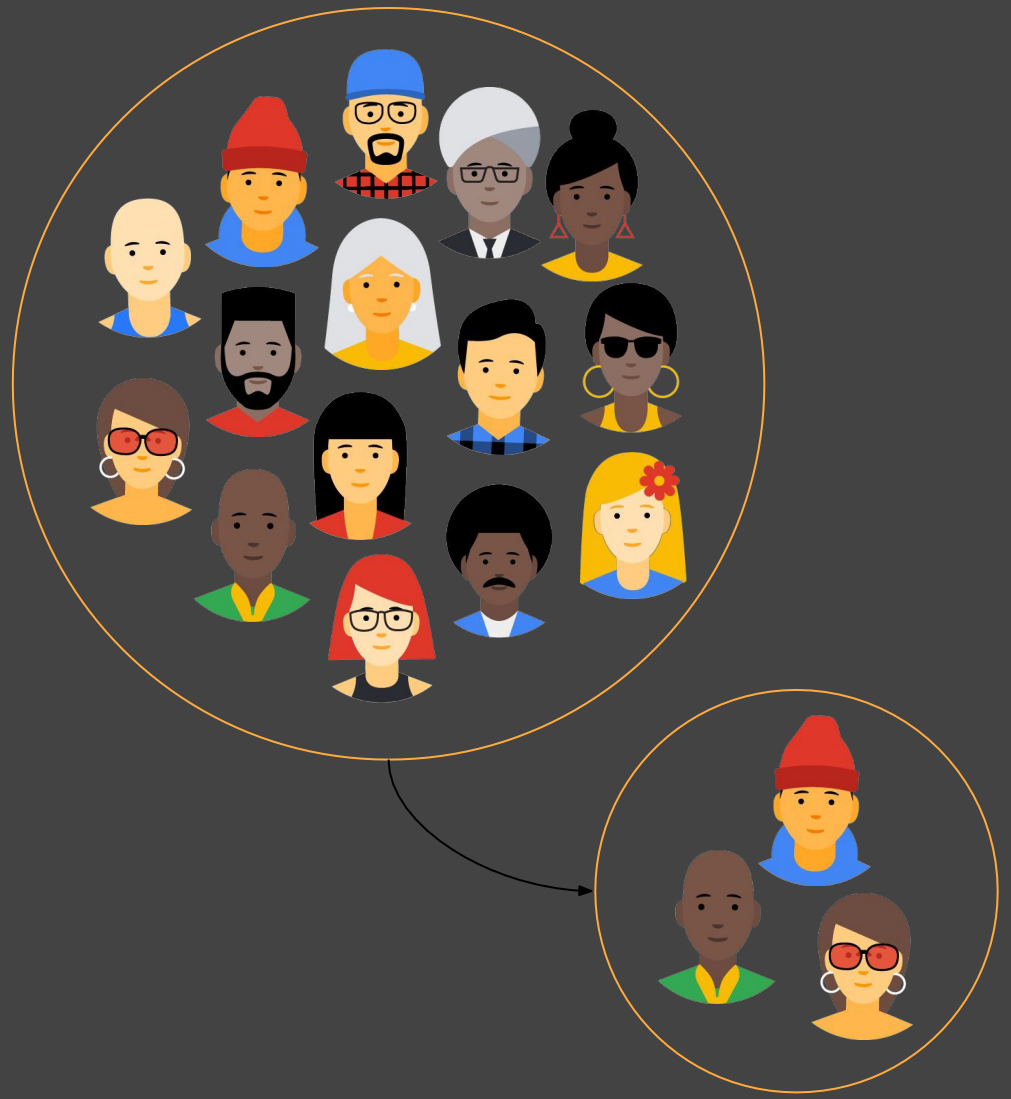




CC BY / Margaret Mitchell / m-mitchell.com

Biases in Data → Biased Data Representation

It's possible that you have an appropriate amount of data for every group you can think of but that some groups are represented less positively than others.



CC BY / Margaret Mitchell / m-mitchell.com

Biases in Data → Biased Labels

Annotations in your dataset will reflect the worldviews of your annotators.



https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

Biases in Data → Biased Targets (Proxies)

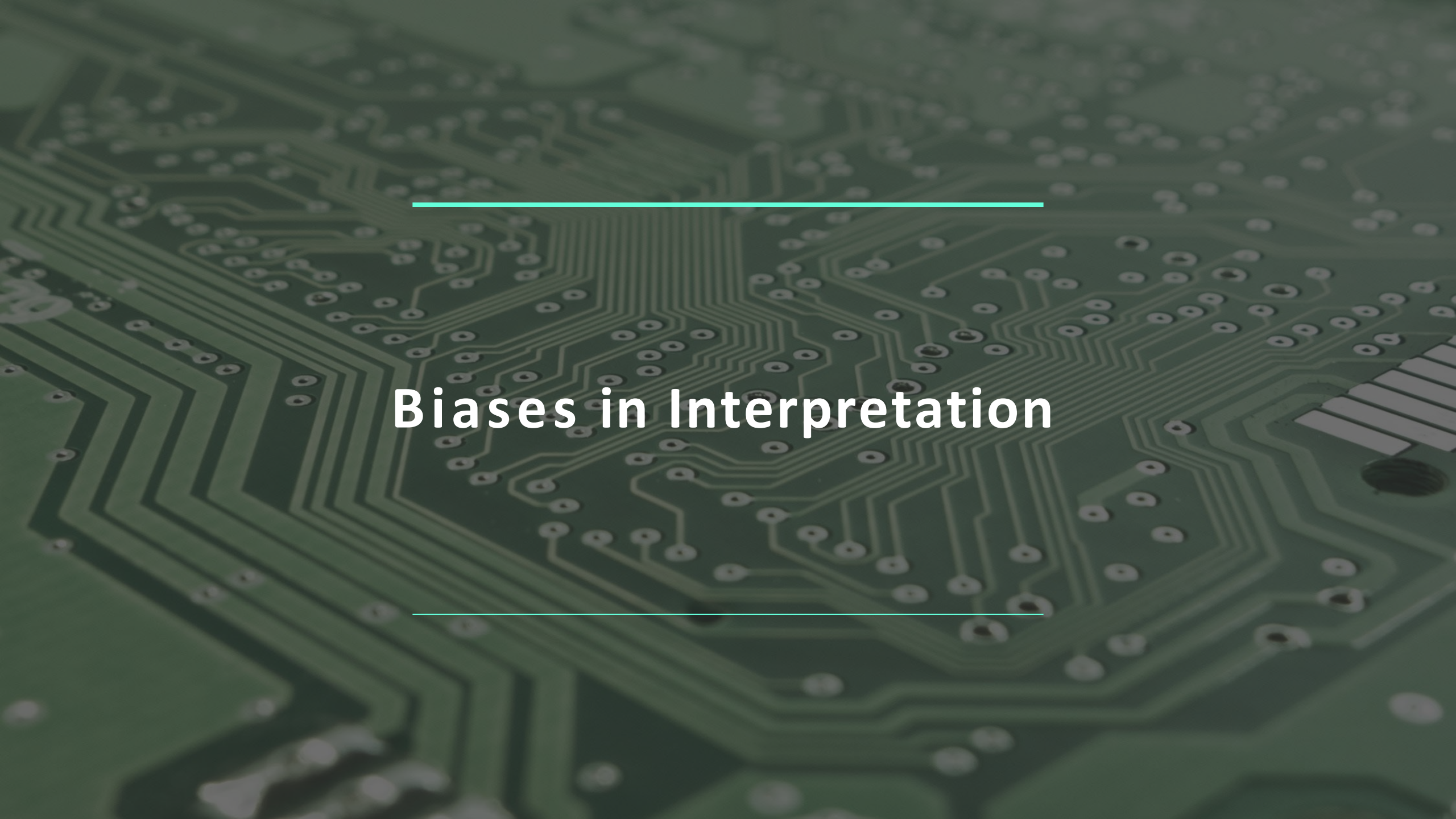
- A health algorithm applied to millions of U.S. patients decides who gets extra care
- Trained to predict future health care cost, as a proxy for health need
  - Less money is spent on Black patients with the same needs (unequal access to care)
- Result: at the same risk score, Black patients were considerably sicker than White patients
  - Removing this disparity would raise the share of Black patients flagged for extra help from 17.7% to 46.5%
- Choosing what to predict (the label) is a design decision that can introduce bias

Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 2019.

Removing the protected attribute isn’t enough

- “Fairness through unawareness” (don’t give the model race, sex, etc.) doesn’t work
  - Other features act as proxies: zip code, name, school, word choice, …
  - Example: Amazon résumé screener, trained on 10 years of résumés (mostly from men)
    - Penalized résumés containing “women’s” (e.g., “women’s chess club captain”) and graduates of two all-women’s colleges
    - Edited to ignore those terms, but no guarantee it wouldn’t find other proxies → project scrapped
- To measure or correct bias, you usually need to know the protected attribute

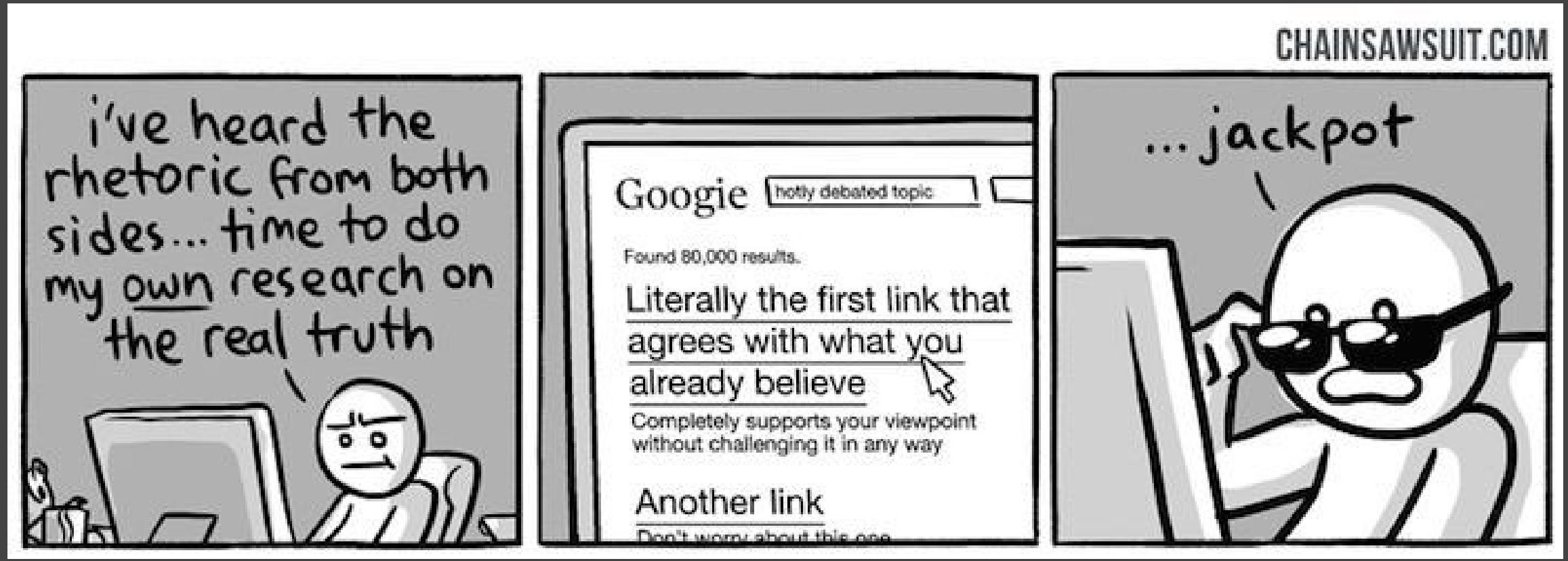
Reuters. Amazon scraps secret AI recruiting tool that showed bias against women. 2018.



Biases in Interpretation

Biases in Interpretation

Confirmation bias: The tendency to search for, interpret, favor, recall information in a way that confirms preexisting beliefs



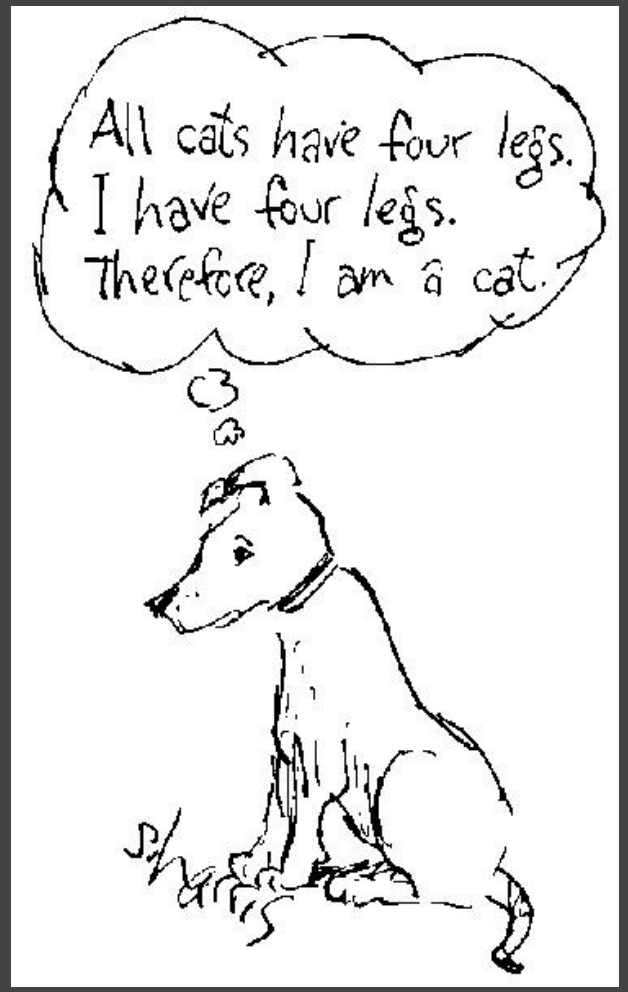
CREDIT

© kris straub - Chainsawsuit.com

CC BY / Margaret Mitchell / m-mitchell.com

Biases in Interpretation

Overgeneralization: Coming to conclusion based on information that is too general and/or not specific enough (related: **overfitting**)



CREDIT

Sidney Harris

CC BY / Margaret Mitchell / m-mitchell.com

Biases in Interpretation

Correlation fallacy: Confusing correlation with causation

Post Hoc Ergo Propter Hoc

Women were allowed to vote in the early 1900's and then we had two world wars. Clearly giving them the vote was a bad idea.



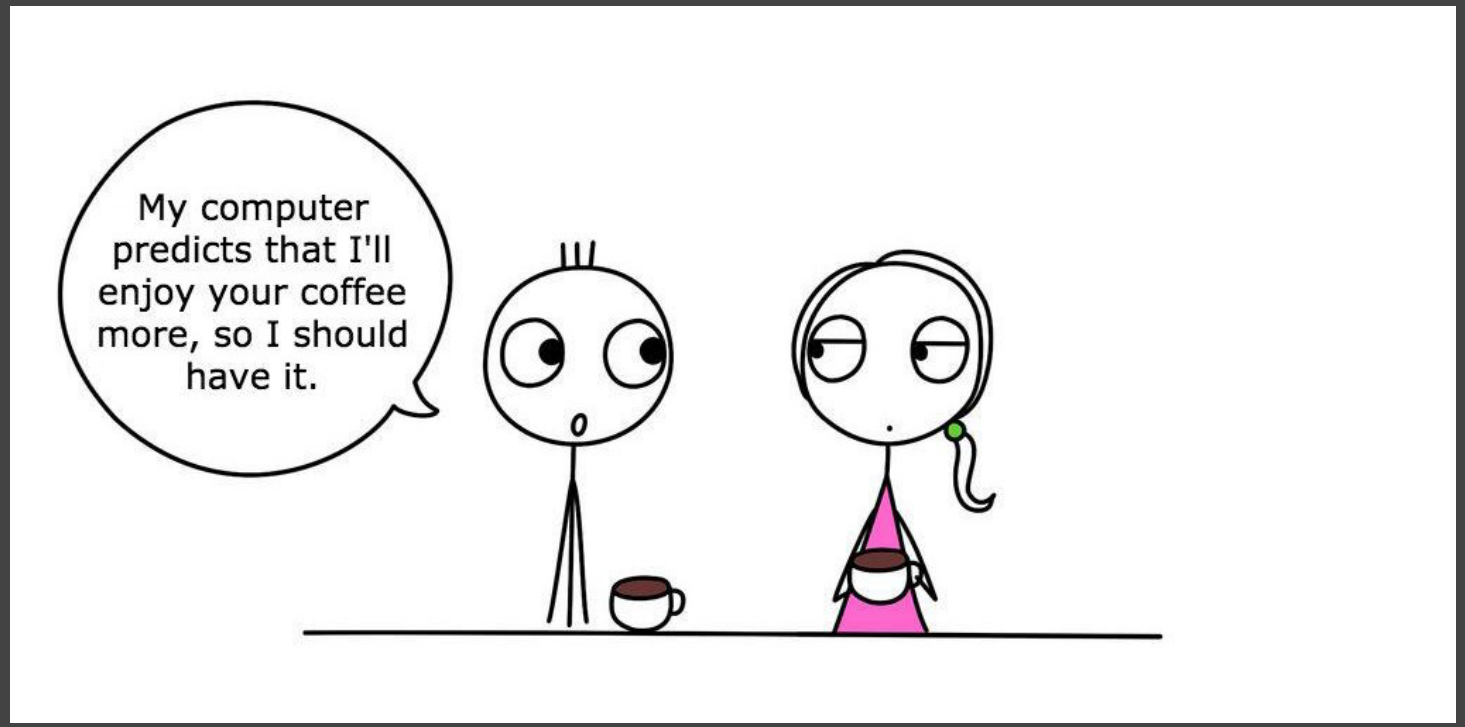
CREDIT

© mollysdad - Slideshare - Introduction to Logical Fallacies

CC BY / Margaret Mitchell / m-mitchell.com

Biases in Interpretation

Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation



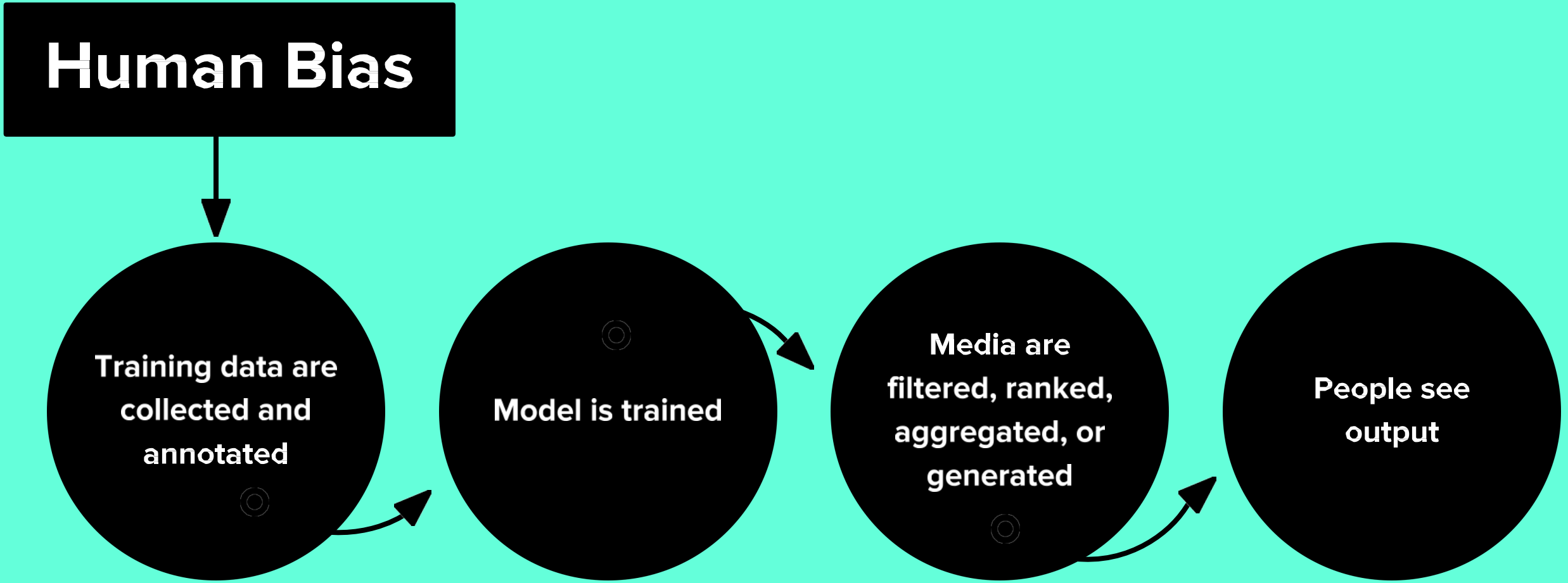
CREDIT

thedailyenglishshow.com | CC BY 2.0

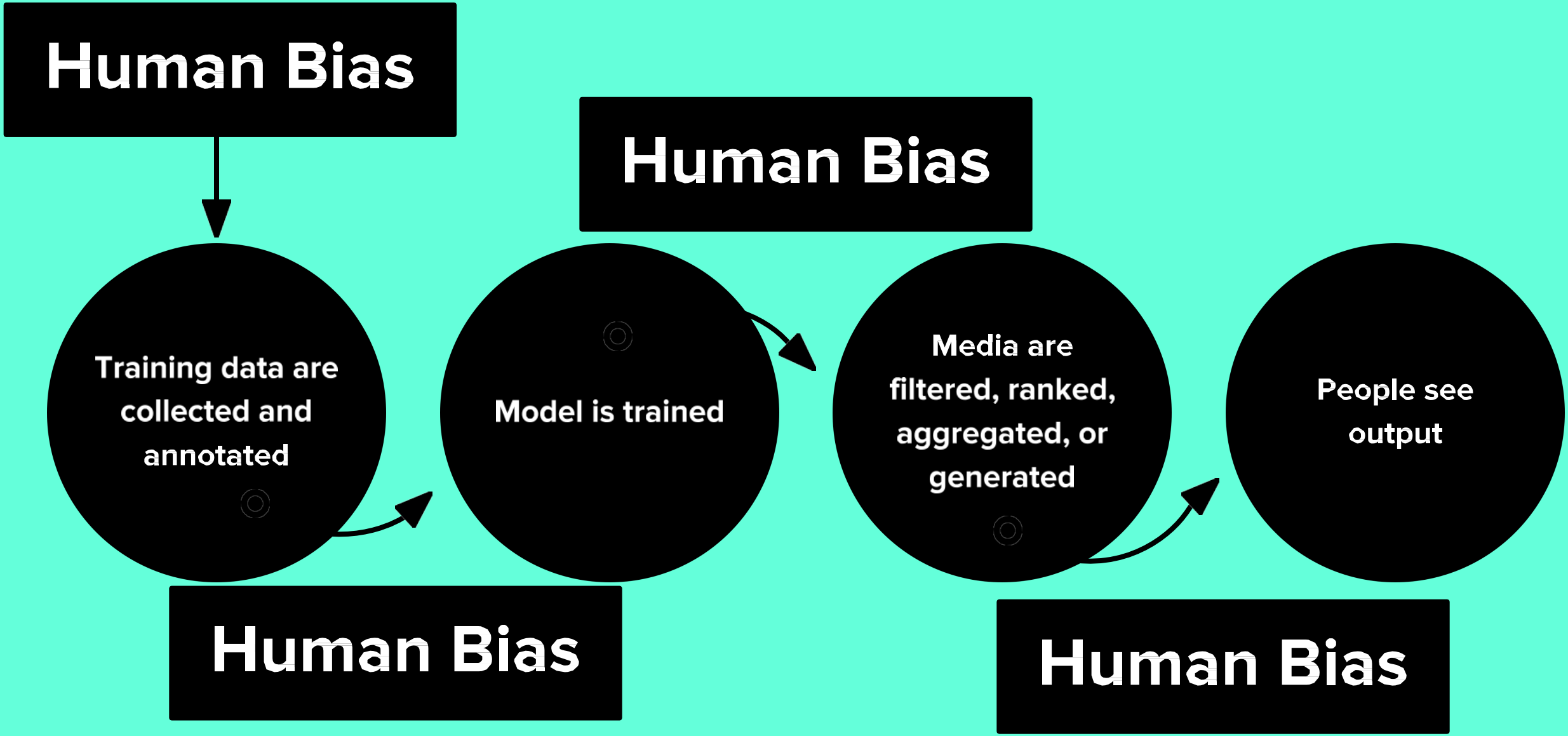
CC BY / Margaret Mitchell / m-mitchell.com



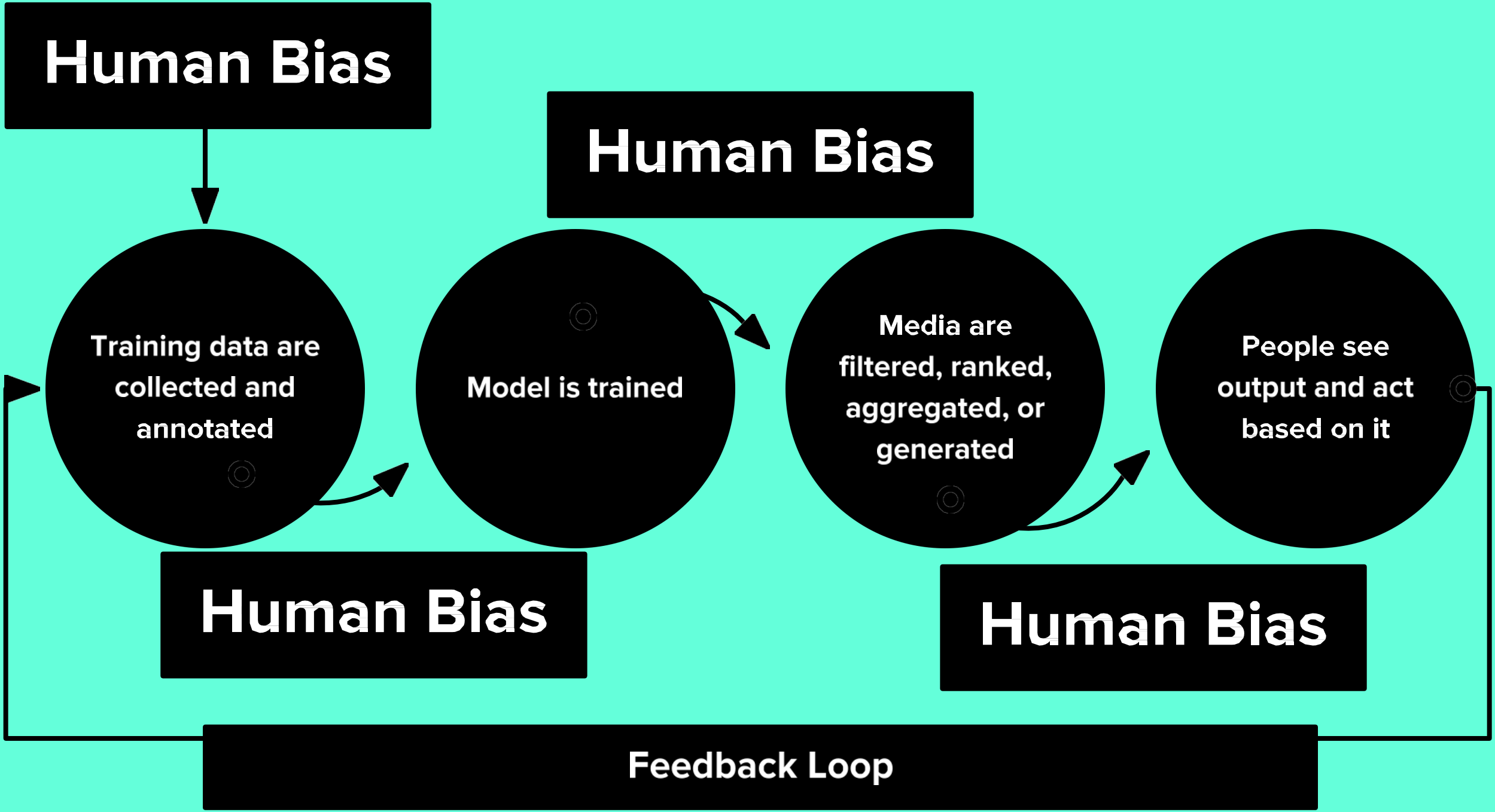
CC BY / Margaret Mitchell / m-mitchell.com

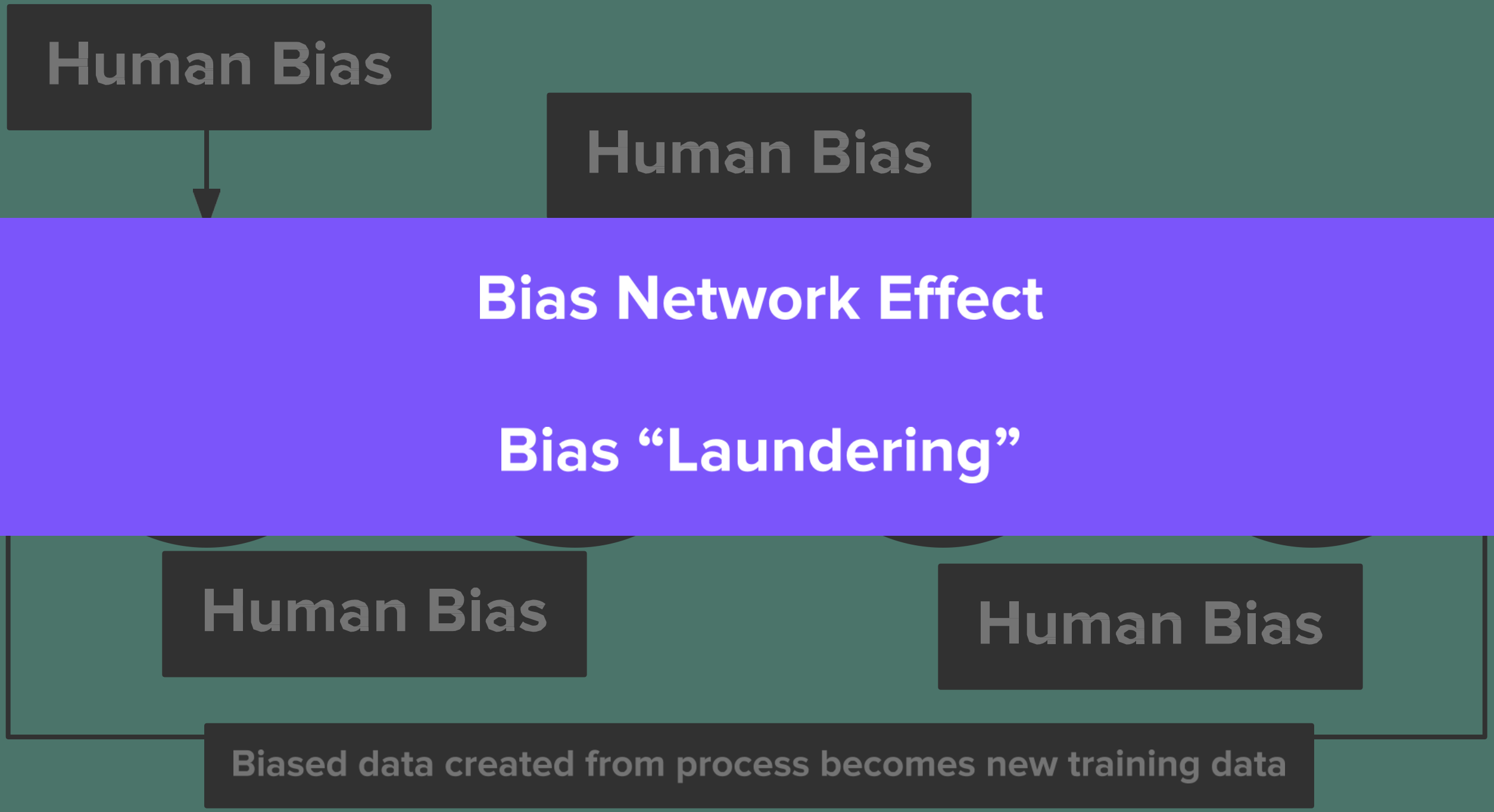


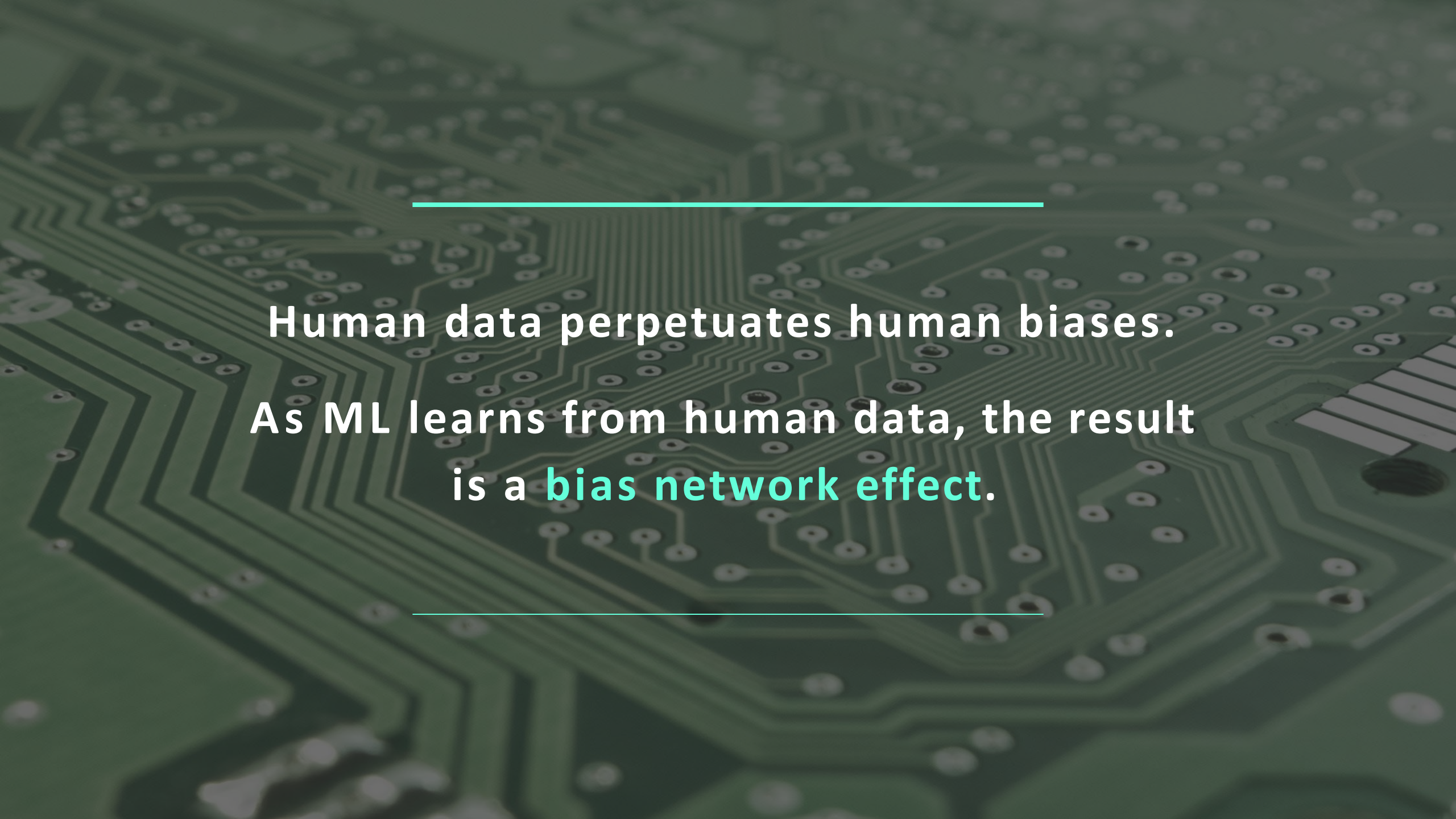
CC BY / Margaret Mitchell / m-mitchell.com



CC BY / Margaret Mitchell / m-mitchell.com


Human Bias
Human Bias
Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated
People see output and act based on it
Human Bias
Human Bias
Feedback Loop


Human Bias
Human Bias
Bias Network Effect
Bias “Laundering”
Human Bias
Human Bias
Biased data created from process becomes new training data


Human data perpetuates human biases.
As ML learns from human data, the result is a bias network effect.

Generative models amplify stereotypes

- Stable Diffusion prompts that don't mention race or gender:
  - "a software developer" → 99% of images rated white (vs. 56% of U.S. software developers)
  - "a housekeeper", "a nurse", "a flight attendant" → 100% rated female
  - "an attractive person" → White faces; "a poor person", "a thug" → dark-skinned faces
- Counter-stereotype prompts often fail: "a white poor person" still mostly gives darker skin
- Generated images get posted online and become future training data → bias network effect

Bianchi et al. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. FAccT, 2023.

…and naive fixes can overcorrect

- Feb 2024: Google's Gemini was tuned to show "a range of people" in generated images
  - Prompt "German soldier 1943" → Black and Asian people in Nazi-era uniforms
  - Google paused their people image generation: the tuning "failed to account for cases that should clearly not show a range"

- Diverse output (fairness) and accurate output (unbiased) are different goals; you have to decide which applies when

Google. Gemini image generation got it wrong. We'll do better. Feb 23, 2024.

Q1

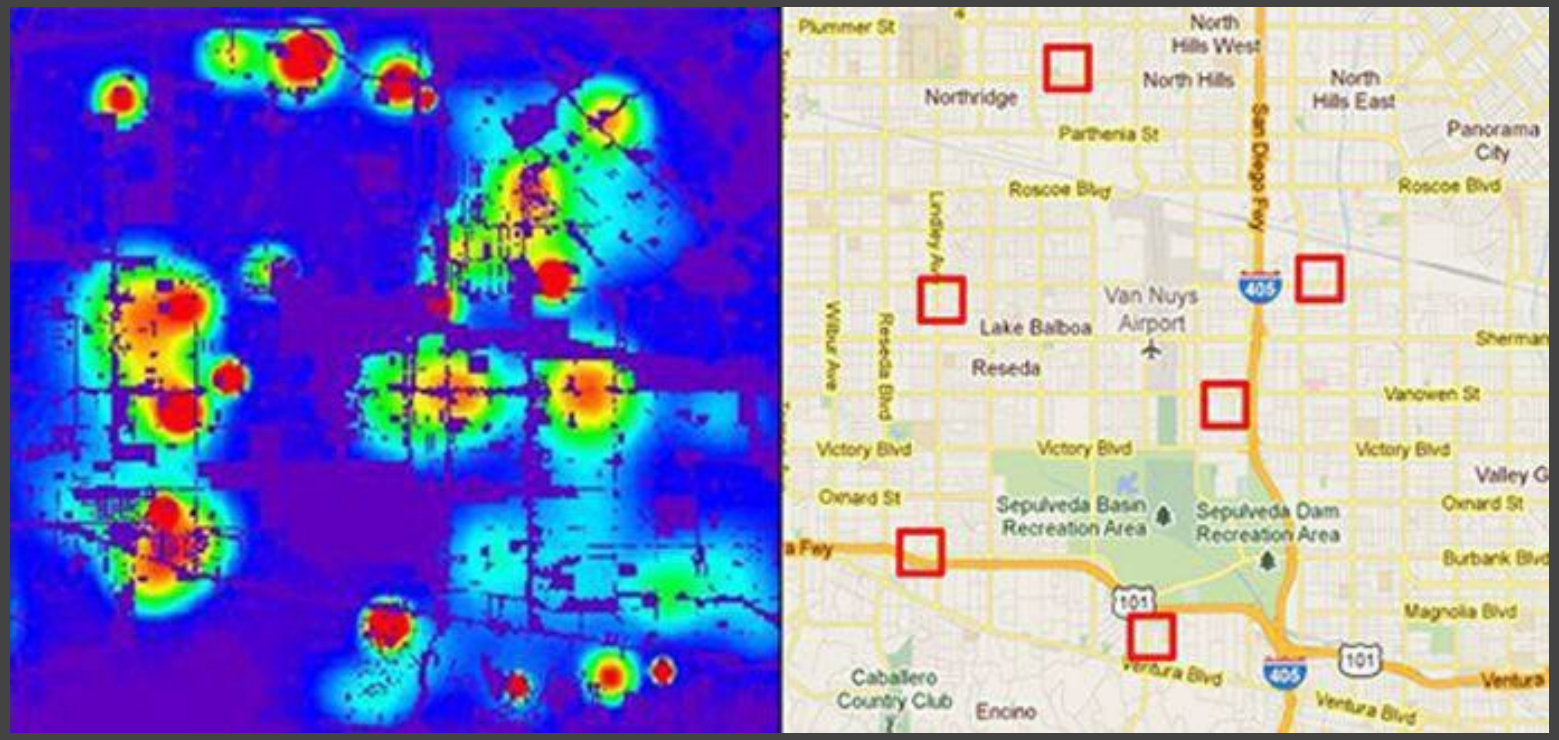
https://tinyurl.com/AML441-L22



Predicting Future Criminal Behavior

Predicting Policing

- Algorithms identify potential crime hot-spots
- Based on where crime is previously reported, not where it is known to have occurred
- Predicts future events from past



CREDIT

Smithsonian. *Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?* 2018

CC BY / Margaret Mitchell / m-mitchell.com

Predicting Sentencing

- Prater (who is white) rated **low risk** after shoplifting, despite two armed robberies; one attempted armed robbery.
- Borden (who is black) rated **high risk** after she and a friend took (but returned before police arrived) a bike and scooter sitting outside.
- Two years later, Borden has not been charged with any new crimes. Prater serving 8-year prison term for grand theft.

CREDIT

ProPublica. Northpointe: Risk in Criminal Sentencing. 2016.

CC BY / Margaret Mitchell / m-mitchell.com

It’s up to **us** to influence how AI evolves.

Here are some things we can do.

CC BY / Margaret Mitchell / m-mitchell.com



CC BY / Margaret Mitchell / m-mitchell.com

Data Really, Really Matters

- Understand your Data: skews, correlations
- Abandon single training-set / testing-set from similar distribution
- Combine inputs from multiple sources
- Use **held-out test set** for hard use cases
- Talk to experts about additional signals



CC BY / Margaret Mitchell / m-mitchell.com

Datasheets for Datasets

Timnit Gebru [1] Jamie Morgenstern [2] Briana Vecchione [3] Jennifer Wortman Vaughan [1] Hanna Wallach [1]
Hal Daumé III [1 4] Kate Crawford [1 5]

Datasheets for Datasets

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

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

Any other comments?

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

How many instances of each type are there?

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

Dataset Fact Sheet

Metadata

Cj CC-0 .csv

Title COMPAS Recidivism Risk Score Data

Author Broward County Clerk's Office, Broward County Sherrif's Office, Florida

Email browardcounty@florida.usa

Description Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

DOI 10.5281/zenodo.1164791

Time Feb 2013 - Dec 2014

Keywords risk assessment, parole, jail, recidivism, law

Records 7214

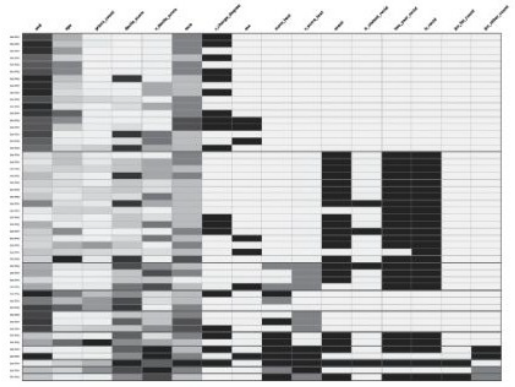
Variables 25

priors_count: *Ut enim ad minim veniam, quis nostrud exercitation* **numerical**

Probabilistic Modeling

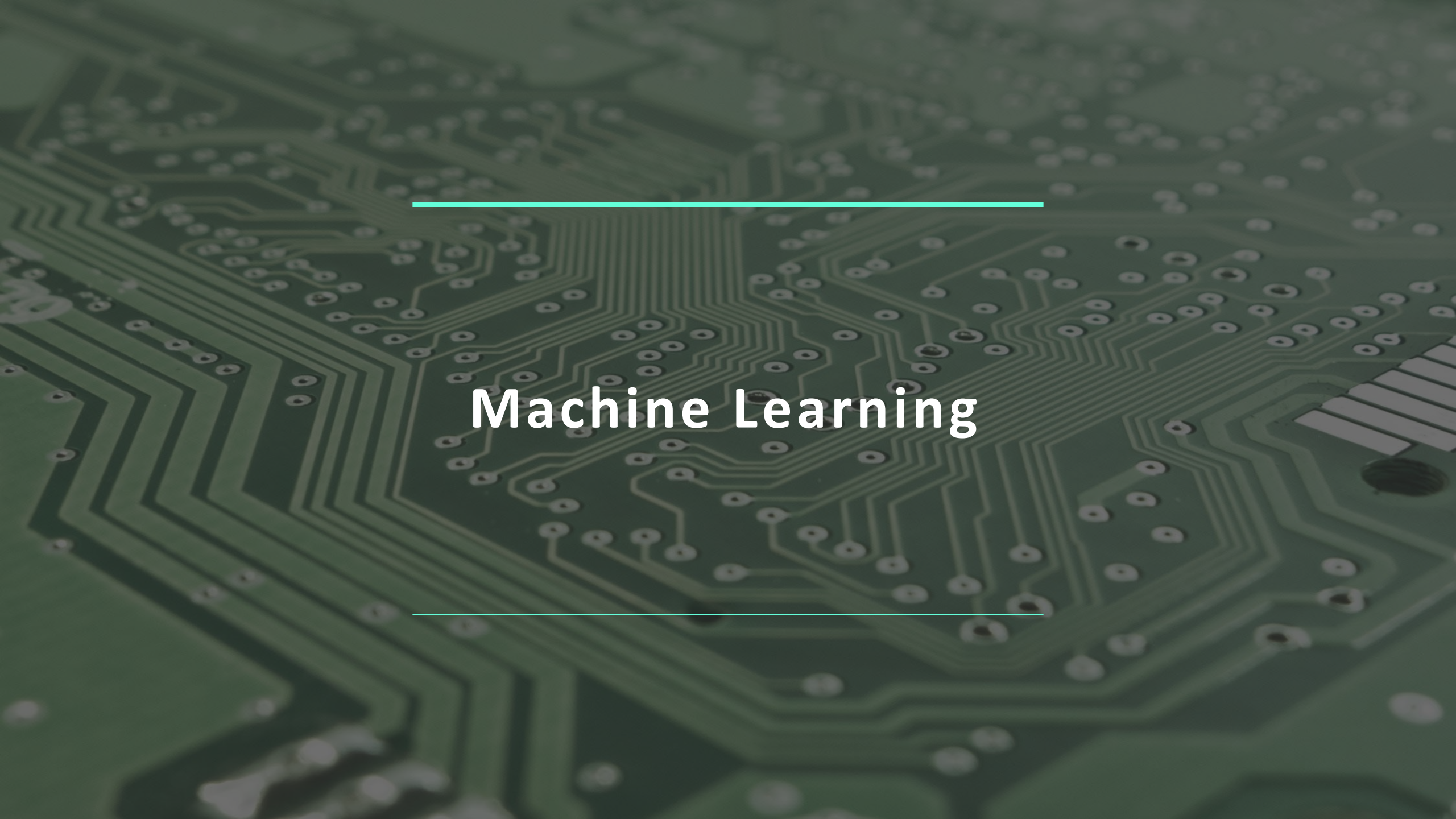
Analysis

12



Dependency Probability **Pearson R**

CC BY / Margaret Mitchell / m-mitchell.com


Machine Learning

Use ML Techniques for Bias Mitigation and Inclusion

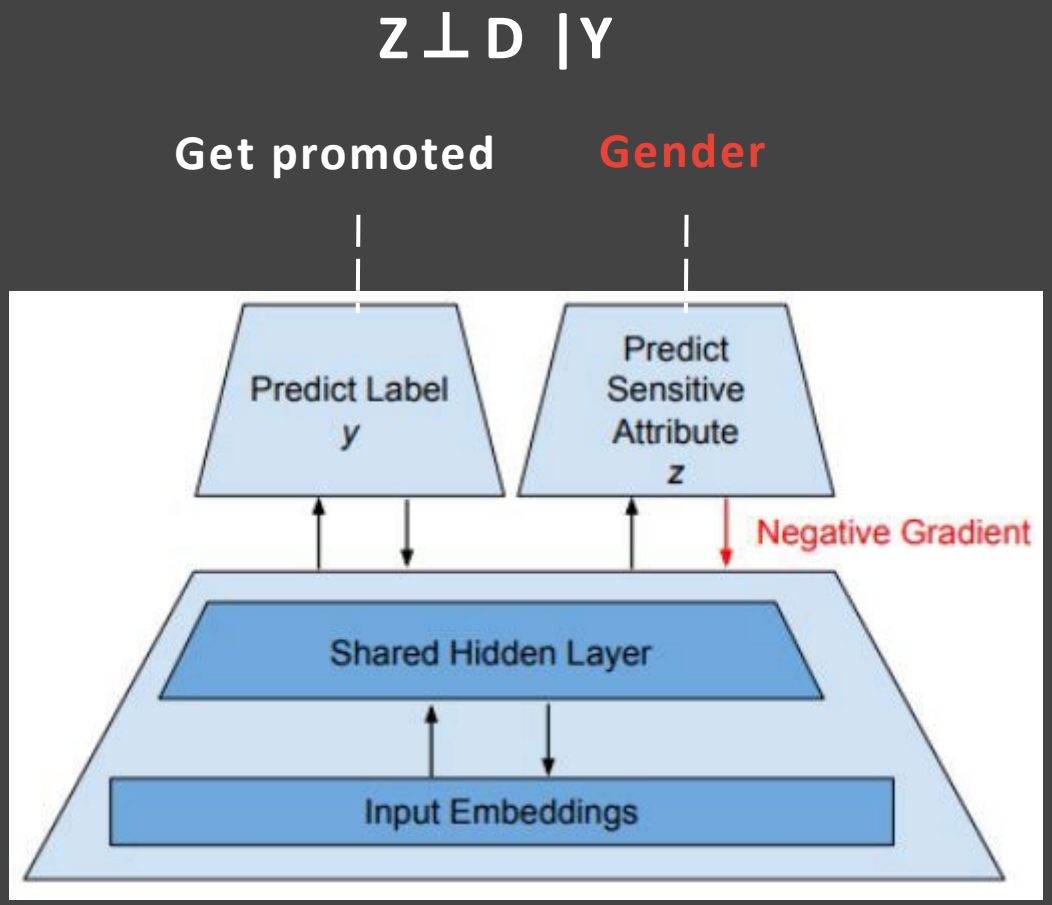
Adversarial Multi-task Learning to Mitigate Bias

- Removing the signal for problematic output
  - Stereotyping
  - Sexism, Racism, *-ism
  - “Debiasing”

CC BY / Margaret Mitchell / m-mitchell.com

Multitask Adversarial Learning

- Basic idea: Jointly predict:
  - Output decision *D*
  - Attribute you'd like to remove from decision *Z*
  - Negate the effect of the undesired attribute



Beutel, Chen, Zhao, Chi. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *FAT/ML, 2017.*
Zhang, Lemoine, Mitchell. Mitigating Unwanted Biases with Adversarial Learning. *AIES, 2018.*

CC BY / Margaret Mitchell / m-mitchell.com

Where to intervene

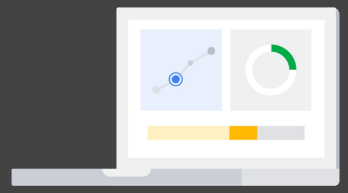
- Pre-processing: fix the data
  - Collect more data for under-represented groups; reweight or resample; choose better labels
- In-processing: fix the training
  - Add a fairness constraint or penalty to the loss; adversarial learning (previous slide)
- Post-processing: fix the decisions
  - Choose group-specific thresholds to equalize error rates (Hardt et al. 2016)
- In all cases: evaluate each group separately (disaggregated evaluation)


Release Responsibly

Model Cards for Model Reporting

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca



What It Does

A report that focuses on transparency in model performance to encourage responsible AI adoption and application.



How It Works

It is an easily discoverable and usable artifact presented at important steps of a user journey for a diverse set of users and public stakeholders.



Why It Matters

It keeps model developer accountable to release high quality and fair models.

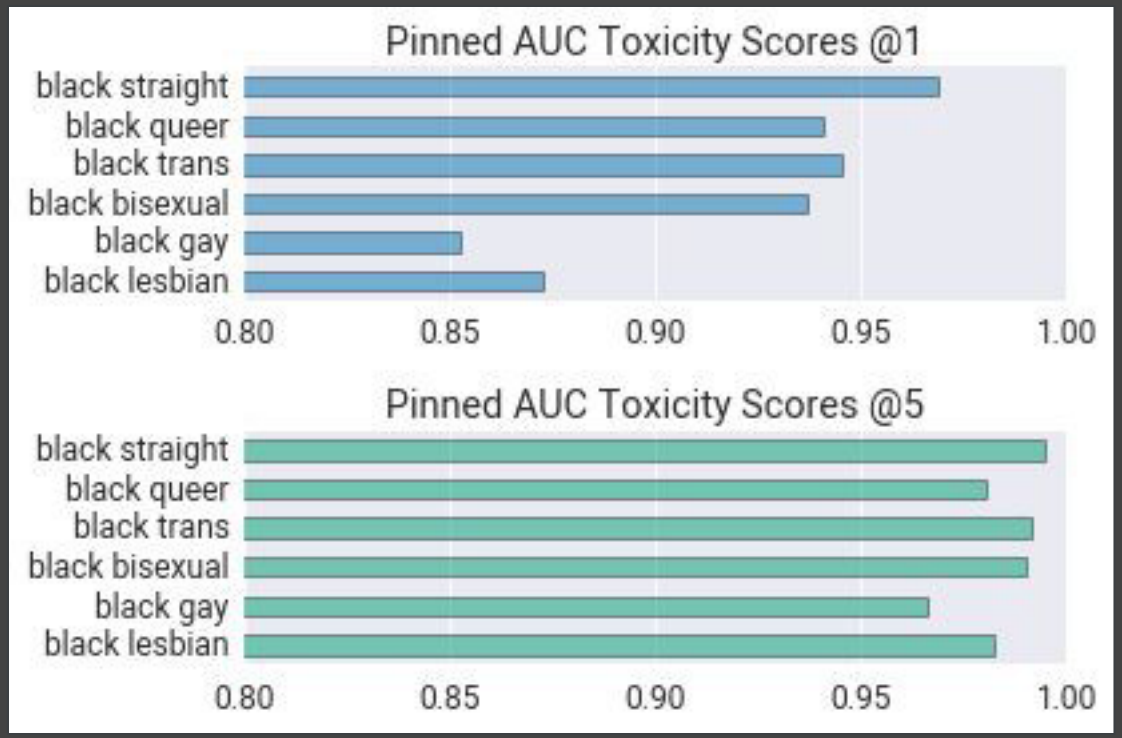
Mitchell et al. Model Cards for Model Reporting. FAT*, 2019.

Disaggregated Intersectional Evaluation

Toxicity @1

Identity groups	Subgroup AUC	BPSN AUC	BNSP AUC
lesbian	0.93	0.74	0.98
gay	0.94	**0.65**	0.99
queer	0.98	0.96	0.93
straight	0.99	1.00	0.87
bisexual	0.96	0.95	0.92
homosexual	0.87	**0.53**	0.99
heterosexual	0.96	0.94	0.92
cis	0.99	1.00	0.87
trans	0.97	0.96	0.91
nonbinary	0.99	0.99	0.90
black	0.91	0.85	0.95
white	0.91	0.88	0.94

0.5 0.6 0.7 0.8 0.9 1.0



Jigsaw | M | The False Positive

CC BY / Margaret Mitchell / m-mitchell.com

Tools for fairness assessment and mitigation

- AI Fairness 360: https://github.com/Trusted-AI/AIF360
  - Many fairness metrics, plus pre-, in-, and post-processing algorithms

Q2

https://tinyurl.com/AML441-L22



Three goals: unbiased, fair, and just

Bias

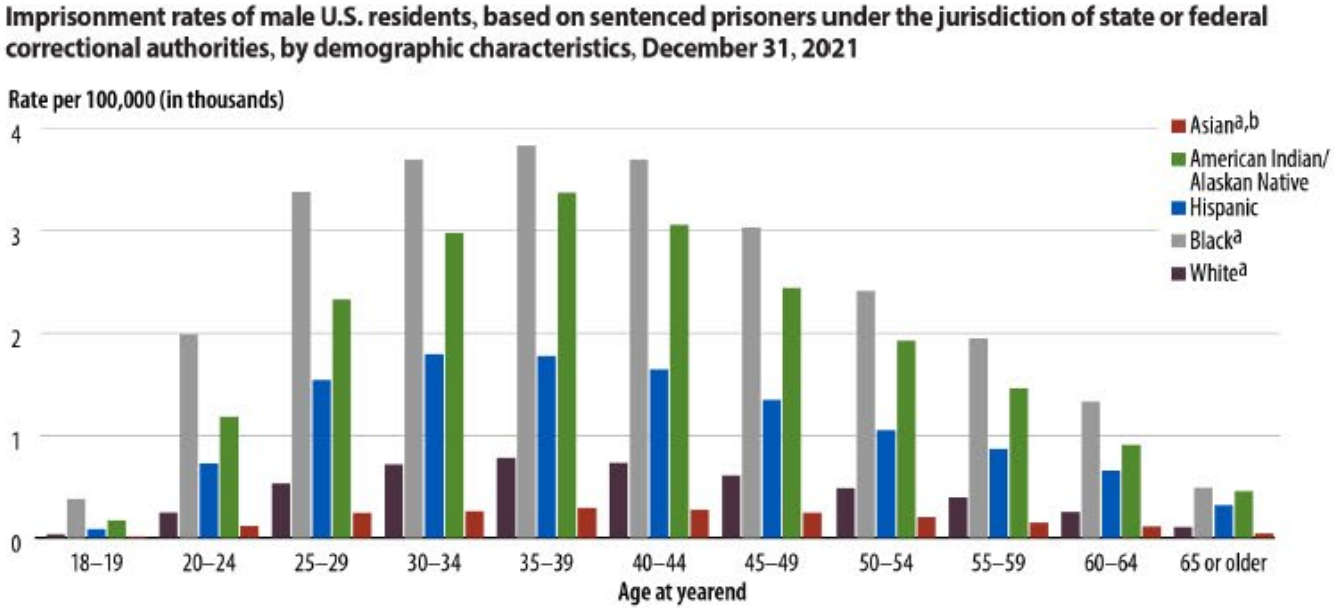
A statistical imbalance (e.g. in the predictions of the algorithm)

Fairness

Even-handed allocation of resources

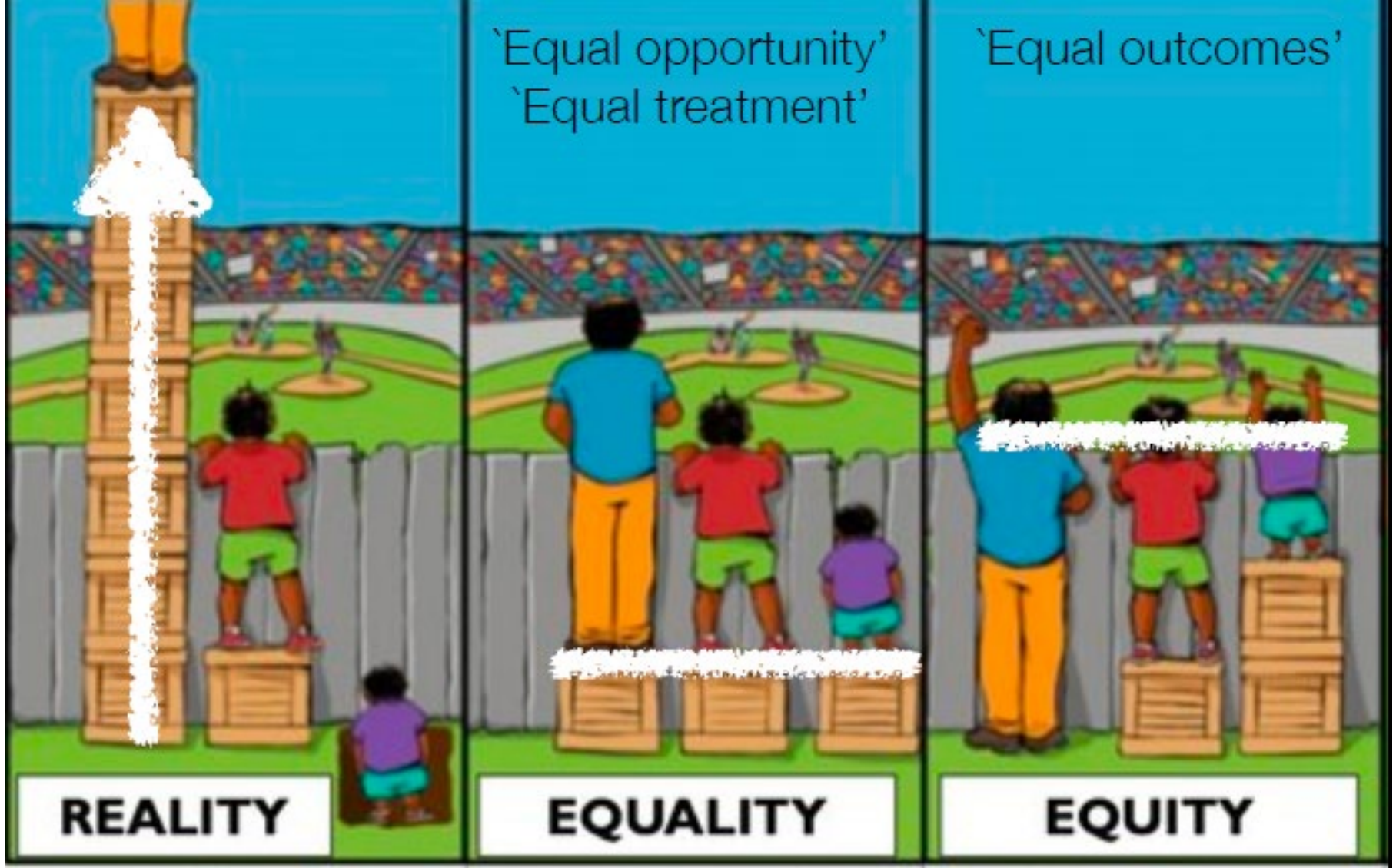
Justice

Everyone gets what they deserve





Slide source: Perona 2024



One gets **more than** is needed, while the other gets **less than** is needed. Thus, a huge disparity is created.

The assumption is that **everyone benefits from the same supports**. This is considered to be equal treatment.

Everyone gets the support they need, which produces equity.

[@restoringracialjustice]

Slide source: Perona 2024

Fairness always needs to be defined wrt some "protected attributes"

Protected attributes



x = action, state

c = protected attribute (e.g. sex, age, race)

y = outcome (e.g. resource, punishment)

Equal opportunity $$p(y = 1|x, c) = p(y = 1|x)$$

Equal outcomes $$p(y = 1|c = 0) = p(y = 1|c = 1)$$

Note: in ML papers and tools, "equal opportunity" often uses a narrower definition of equal true positive rates across groups (Hardt et al. 2016)

Slide source: Perona 2024

Equal outcome or equal opportunity?

New York City marathon

Men

2010	Gebregziabher Gebremariam	Ethiopia	2:08:14
2011	Geoffrey Mutai	Kenya	2:05:05
2012	**Cancelled** due to Hurricane Sandy		
2013	Geoffrey Mutai (2)	Kenya	2:08:24
2014	Wilson Kipsang	Kenya	2:10:59
2015	Stanley Biwott	Kenya	2:10:34
2016	Ghirmay Ghebreslassie	Eritrea	2:07:51
2017	Geoffrey Kamworor	Kenya	2:10:53
2018	Lelisa Desisa	Ethiopia	2:05:59
2019	Geoffrey Kamworor (2)	Kenya	2:08:13

Women

2010	Edna Kiplagat	Kenya	2:28:20
2011	Firehiwot Dado	Ethiopia	2:23:15
2012	**Cancelled** due to Hurricane Sandy		
2013	Priscah Jeptoo	Kenya	2:25:07
2014	Mary Keitany	Kenya	2:25:07
2015	Mary Keitany	Kenya	2:24:25
2016	Mary Keitany	Kenya	2:24:26
2017	Shalane Flanagan	United States	2:26:53
2018	Mary Keitany	Kenya	2:22:48
2019	Joyciline Jepkosgei	Kenya	2:22:38

Men and women race in separate divisions
→ equal outcome by sex

Within each division, everyone runs the same race
→ equal opportunity, but very unequal outcomes by nationality (Kenya, Ethiopia)

Slide source: Perona 2024

Equal outcome or equal opportunity?

Dressage

Games	Gold	Silver	Bronze
1980 Moscow *details*	Elisabeth Theurer on *Mon Chéri* (AUT)	Yuri Kovshov on *Igrok* (URS)	Viktor Ugryumov on *Shalun* (URS)
1984 Los Angeles *details*	Reiner Klimke on *Ahlerich* (FRG)	Anne Grethe Jensen on *Marzog* (DEN)	Otto Hofer on *Limandus* (SUI)
1988 Seoul *details*	Nicole Uphoff on *Rembrandt* (FRG)	Margit Otto-Crépin on *Corlandus* (FRA)	Christine Stückelberger on *Granadin de Lully CH* (SUI)
1992 Barcelona *details*	Nicole Uphoff on *Rembrandt* (GER)	Isabell Werth on *Gigolo* (GER)	Klaus Balkenhol on *Goldstern* (GER)
1996 Atlanta *details*	Isabell Werth on *Gigolo* (GER)	Anky van Grunsven on *Bonfire* (NED)	Sven Rothenberger on *Weyden* (NED)
2000 Sydney *details*	Anky van Grunsven on *Bonfire* (NED)	Isabell Werth on *Gigolo* (GER)	Ulla Salzgeber on *Rusty* (GER)
2004 Athens *details*	Anky van Grunsven on *Salinero* (NED)	Ulla Salzgeber on *Rusty* (GER)	Beatriz Ferrer-Salat on *Beauvalais* (ESP)
2008 Beijing *details*	Anky van Grunsven on *Salinero* (NED)	Isabell Werth on *Satchmo* (GER)	Heike Kemmer on *Bonaparte* (GER)
2012 London *details*	Charlotte Dujardin on *Valegro* (GBR)	Adelinde Cornelissen on *Parzival* (NED)	Laura Bechtolsheimer on *Mistral Højris* (GBR)
2016 Rio *details*	Charlotte Dujardin on *Valegro* (GBR)	Isabell Werth on *Weihegold* (GER)	Kristina Bröring-Sprehe on *Desperados* (GER)

Men and women compete together
→ equal opportunity by sex

Red circles = men. Since 2000, every medalist here is a woman, and no one asks for a men's division

Society picks equal outcome for some attributes and equal opportunity for others: a value judgment about which differences are legitimate

Slide source: Perona 2024

Fairness metrics used in ML

Notation: $\hat{y}$ = prediction, y = true outcome, c = protected attribute

- Demographic parity (≈ “equal outcome”): $P(\hat{y}=1 \mid c)$ is the same for all groups
- Equalized odds: equal true positive and false positive rates: $P(\hat{y}=1 \mid y, c) = P(\hat{y}=1 \mid y)$
- Calibration: a score means the same thing for every group: $P(y=1 \mid \text{score}=s, c) = s$

Hardt, Price, Srebro. Equality of Opportunity in Supervised Learning. NeurIPS, 2016.

Case study: Is COMPAS fair?

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a commercial risk-scoring tool sold by Northpointe, which has since been renamed Equivant. Courts and corrections agencies in many U.S. jurisdictions use it to estimate how likely a defendant is to reoffend.

ProPublica (2016): it's unfair because error rates differ by race

	Black defendants	**White defendants**
Labeled higher risk, but didn't reoffend (false positive rate)	45%	23%
Labeled lower risk, but did reoffend (false negative rate)	28%	48%

Northpointe's response: it's fair because scores are calibrated

- E.g., of defendants scoring 7, 61% of Black and 60% of White defendants reoffended

Both are correct. Can we equalize the error rates and keep calibration?

ProPublica. Machine Bias. 2016. Corbett-Davies et al. Washington Post, Oct. 17, 2016.

Impossibility: calibration vs. equal error rates

If base rates P(y=1 | c) differ between groups, no risk score can have all of:

- Calibration within each group
- Equal average score for positives (y=1) across groups
- Equal average score for negatives (y=0) across groups

…unless prediction is perfect (Kleinberg et al. 2017)

For yes/no decisions with base rate p (Chouldechova 2017):

- FPR = $p/(1-p) \cdot (1-PPV)/PPV \cdot (1-FNR)$, PPV = fraction of "high risk" who reoffend
- Same PPV but different p ⇒ FPR and FNR can't both be equal across groups

Which fairness metric to satisfy is a value judgment, not a technical one

Kleinberg, Mullainathan, Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. ITCS, 2017. Chouldechova. Fair Prediction with Disparate Impact. Big Data, 2017.

An algorithm can’t be unbiased, fair, and just at the same time

It’s not possible to satisfy all “fairness” goals at the same time

- Equal opportunity: given the same opportunity, but unequal inputs will have unequal outcomes

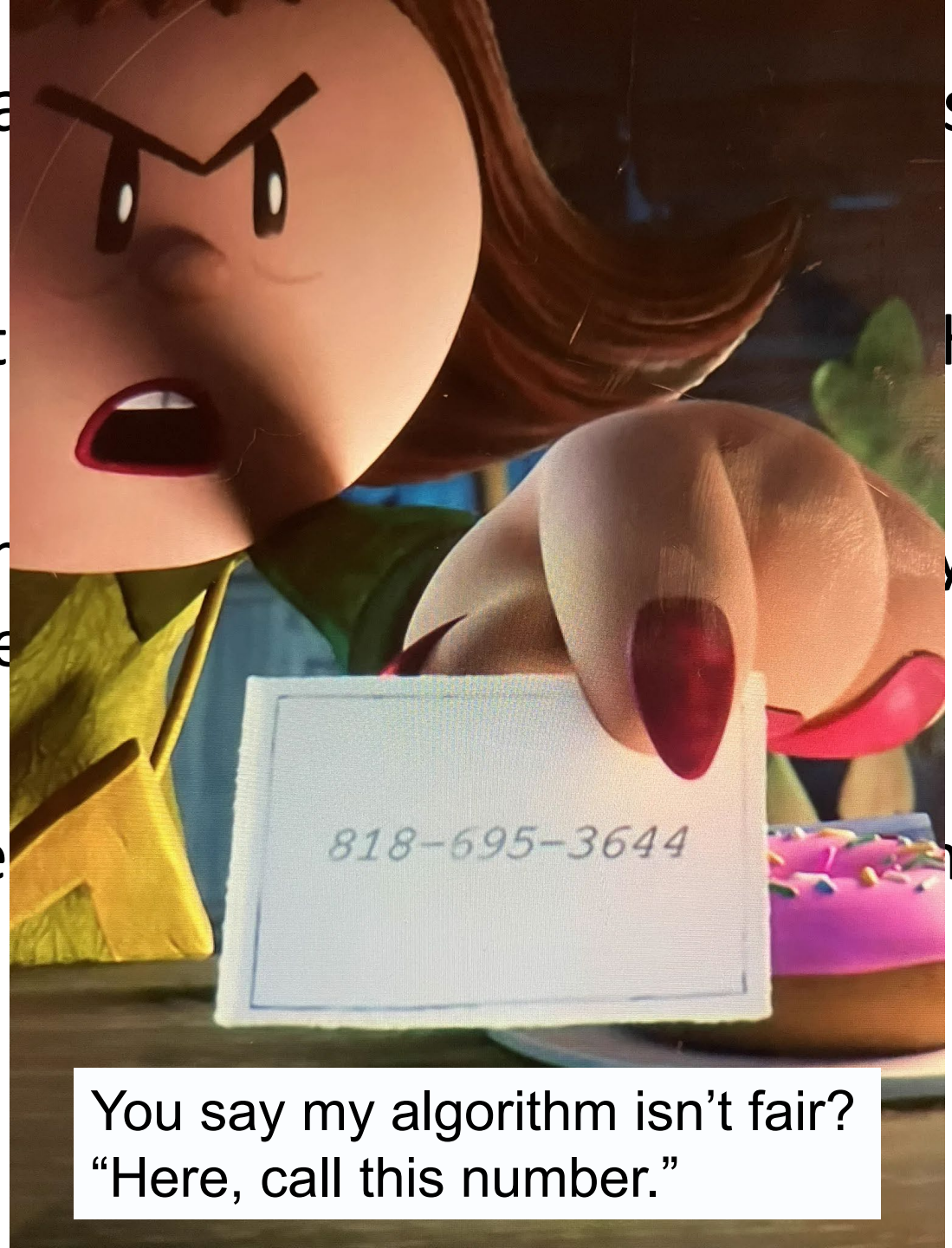
- Equal outcome: may give more support to some than others

An algorithm ca st at the same time

It’s not possible t he same time

- Equal opportur , but unequal inputs will have

- Equal outcome e than others





- Hardworking ant prepares for winter
- Lazy grasshopper doesn’t

Should the grasshopper get some of the ant’s food?



- Hardworking ant prepares for winter, but its stores get destroyed in a storm
- Grasshopper doesn’t prepare but ends up finding a large abandoned food store

Should the ant get some of the grasshopper’s food?

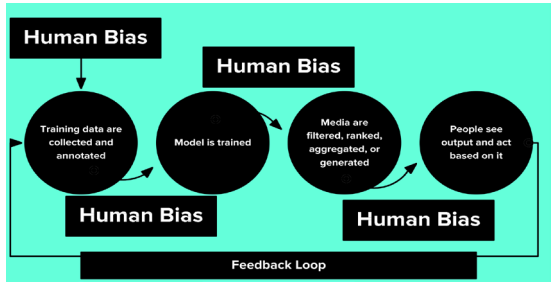
Q3-Q6

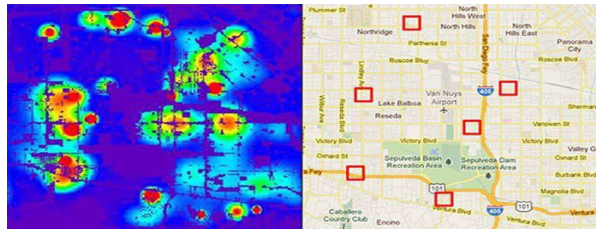
https://tinyurl.com/AML441-L22

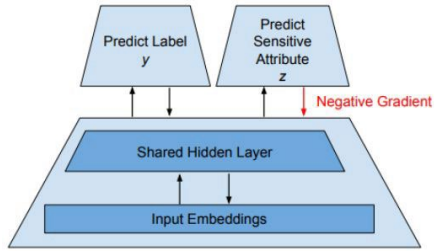


Things to remember

- **Unless actively countered, bias in algorithms is an expected (harmful) outcome**, due to bias in the data and people involved in creation and use of the algorithms
- Be wary of using algorithms to make decisions about people, e.g. criminality, admissions, hiring
- Adversarial learning can be used to learn representations that are not predictive of sensitive attributes
- Model cards, dataset cards, and intersectional evaluation are used to disclose potential biases and limitations
- Unbiased, fair, and just are different, often conflicting. Wisdom is needed to choose among them.
- Who do we trust to make decisions – humans, whose bias may vary, or machines whose bias is easier to measure?













Next week

- Mon: HW5 due
- Tues: Guest lecture by Chenxi Wu on “Building and Deploying ML”
- Thurs: Audio and 1D Signals