



Data Bias and Fair ML

Applied Machine Learning
Derek Hoiem

Today's lecture

- Human, Data, and Algorithm Bias; and what to do about it
 - Adopted from a 2019 guest lecture by Margaret Mitchell, with permission
- Fairness and Justice
 - Adopted 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: even with infinite samples, your estimate of the statistic will not converge to the true value
- 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, income, sexual orientation, religion, gender)

What's the big deal?

True story

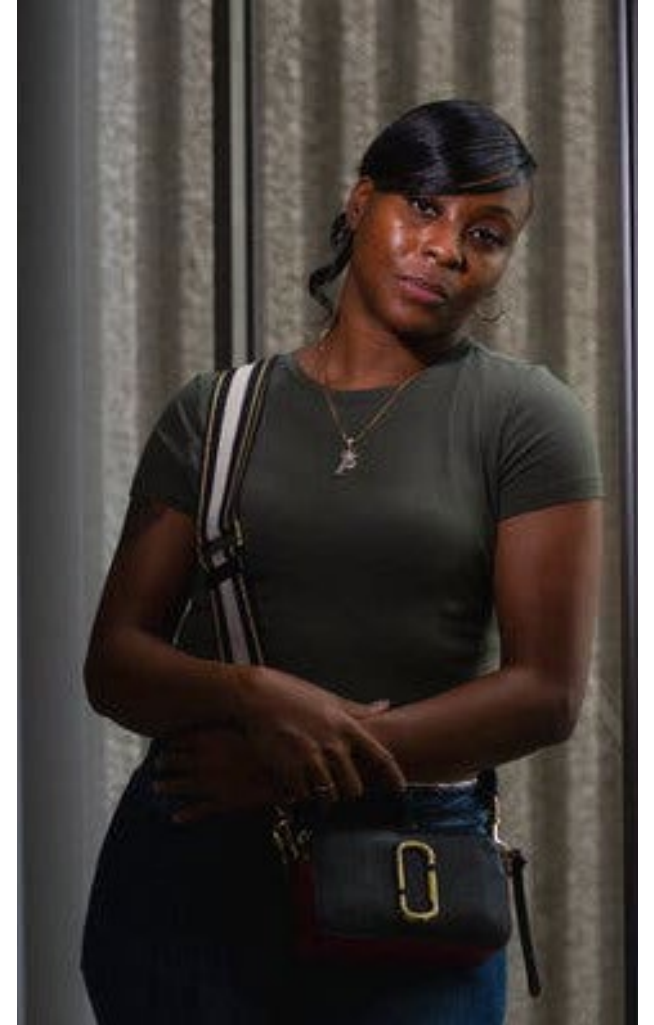
Face recognition trained on predominantly white people

Identifies black person as matching a suspect in armed robbery & carjacking

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

Police arrested and charged her, due to the high confidence “objective” ID

Turns out to be a false arrest



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



Conversation AI

Bias in the Vision and Language of Artificial Intelligence



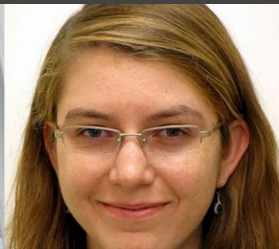
Margaret Mitchell
Senior Research Scientist
Google AI



Andrew
Zaldivar



Me



Simone
Wu



Parker
Barnes



Lucy
Vasserman



Ben
Hutchinson



Elena
Spitzer



Deb
Raji



Timnit Gebru



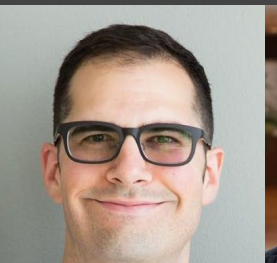
Adrian
Benton



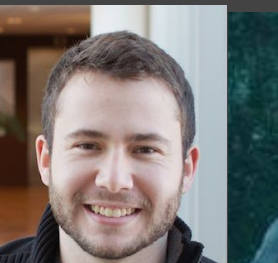
Brian
Zhang



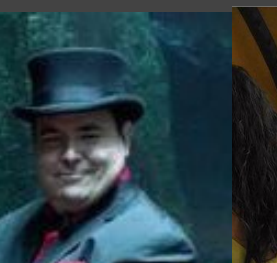
Dirk
Hovy



Josh
Lovejoy



Alex
Beutel



Blake
Lemoine



Hee Jung
Ryu



Hartwig
Adam



Blaise
Agüera y
Arcas

What do you see?



What do you see?

- Bananas



What do you see?

- Bananas
- Stickers



What do you see?

- Bananas
- Stickers
- Dole Bananas



What do you see?

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



What do you see?

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



What do you see?

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



What do you see?

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



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



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



What do you see?

Green Bananas

Unripe Bananas



What do you see?

Ripe Bananas

Bananas with **spots**



What do you see?

Ripe Bananas

Bananas with **spots**

Bananas good for **banana
bread**



What do you see?

Yellow Bananas

Yellow is prototypical for
bananas



Prototype Theory

One purpose of categorization is to **reduce the infinite differences** among stimuli **to** behaviourally and **cognitively usable proportions**

There may be some central, prototypical notions of items that arise from stored typical properties for an object category (Rosch, 1975)

May also store exemplars (Wu & Barsalou, 2009)



Fruit



Bananas
“Basic Level”



**Unripe Bananas,
Cavendish Bananas**

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!"

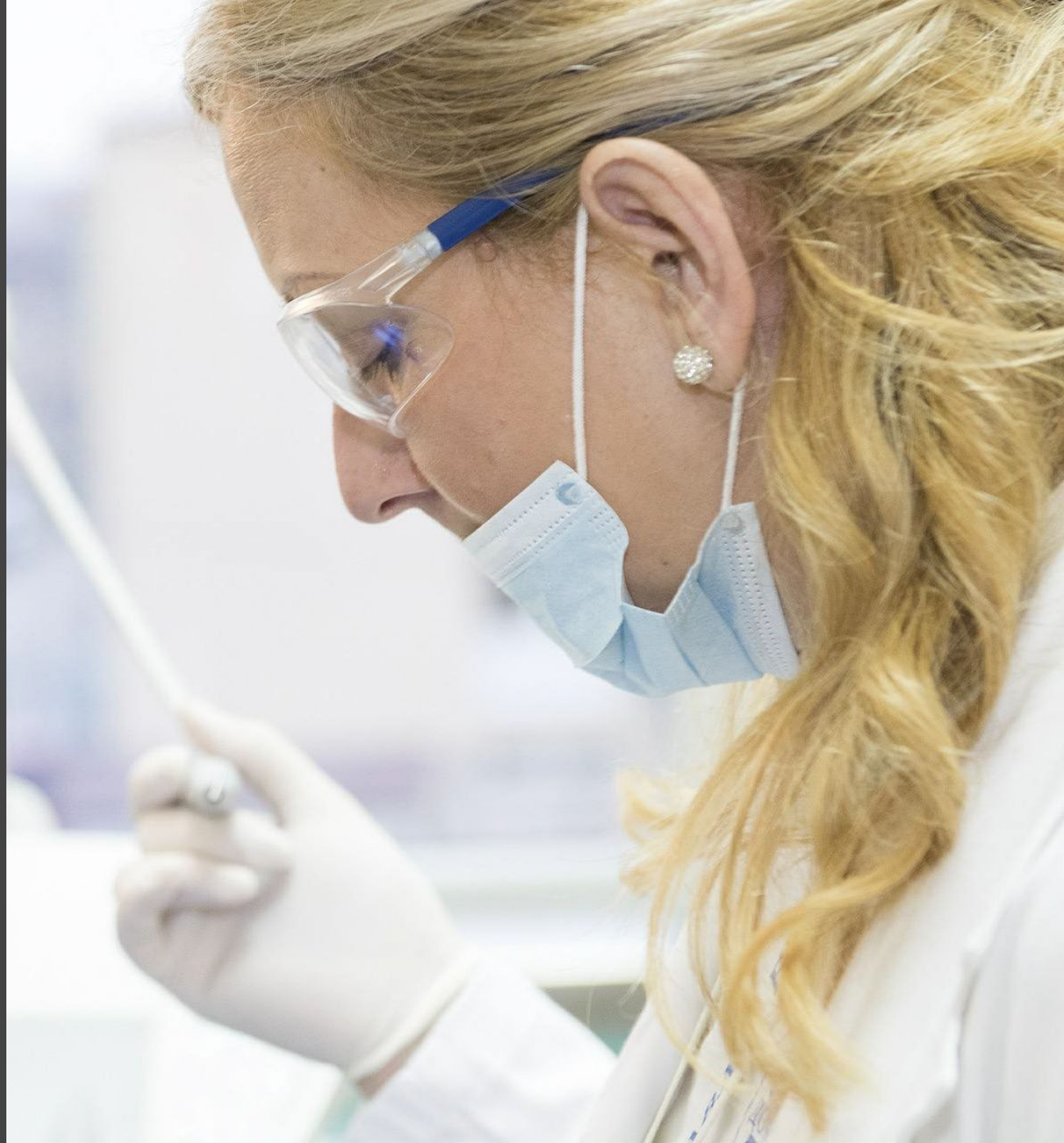
How could this be?



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!"

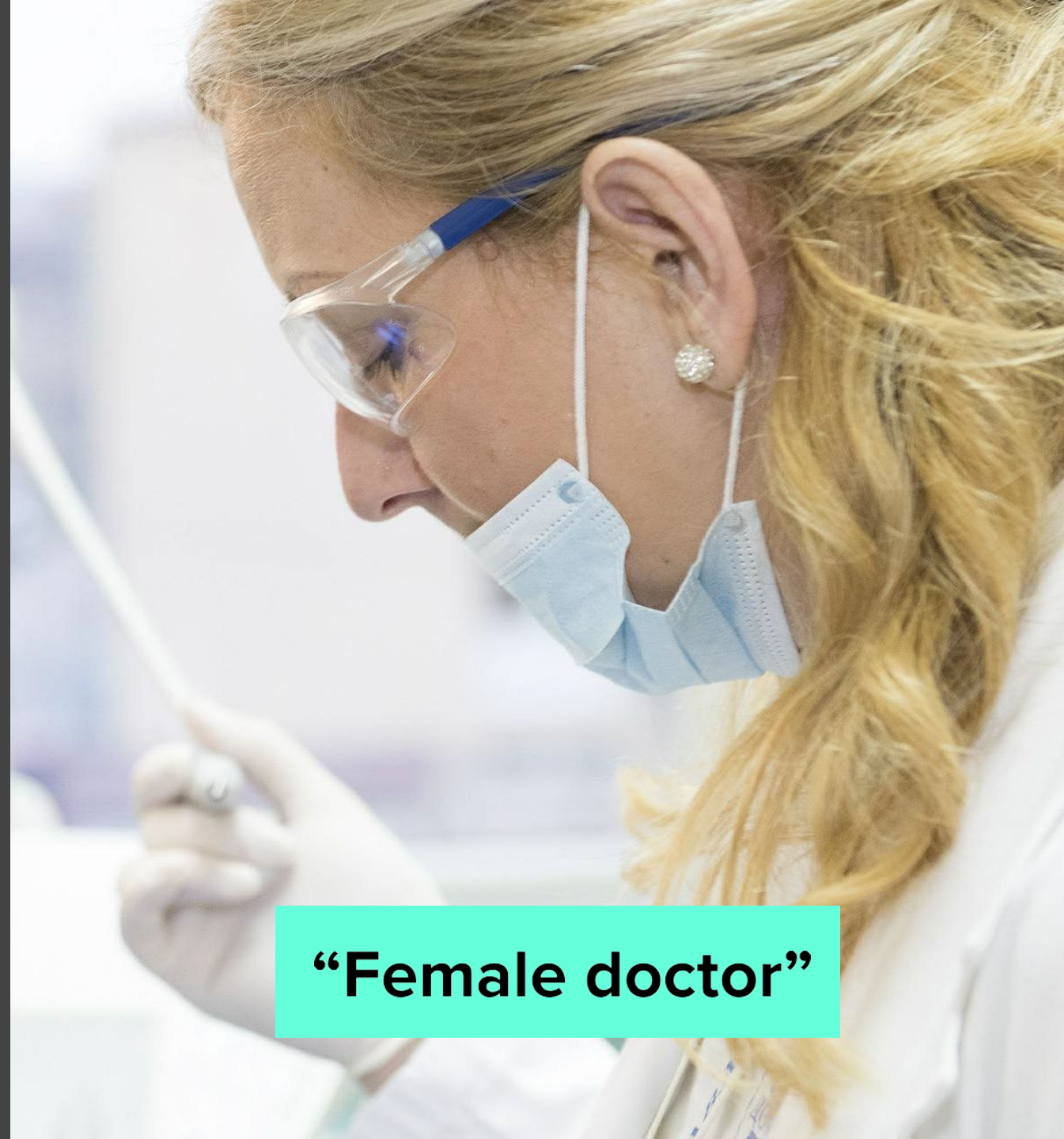
How could this be?



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?



“Female doctor”



“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

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

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

Human reporting bias affects ML models

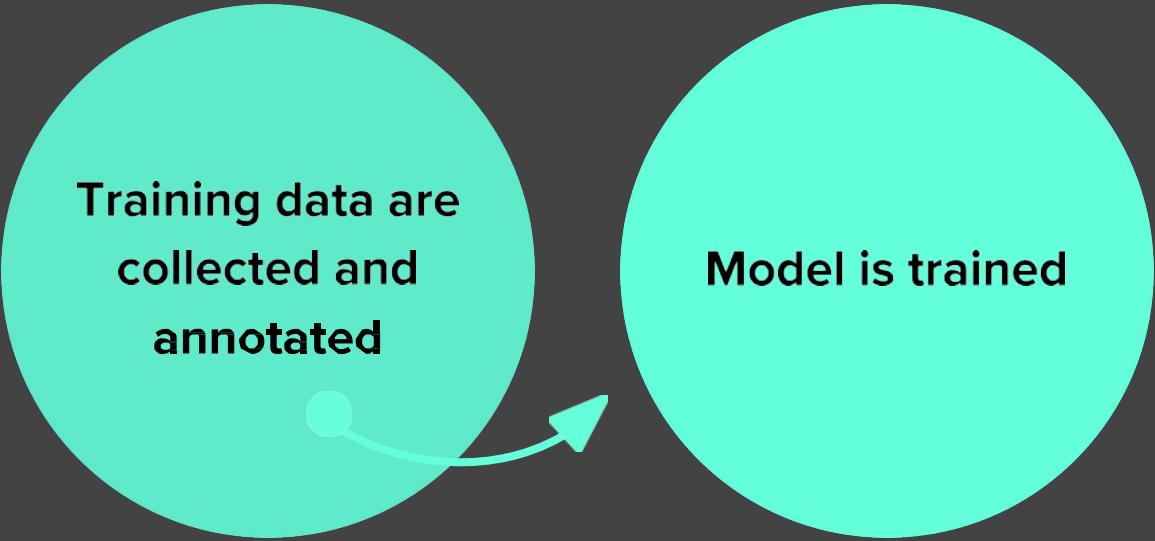
[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?



[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.



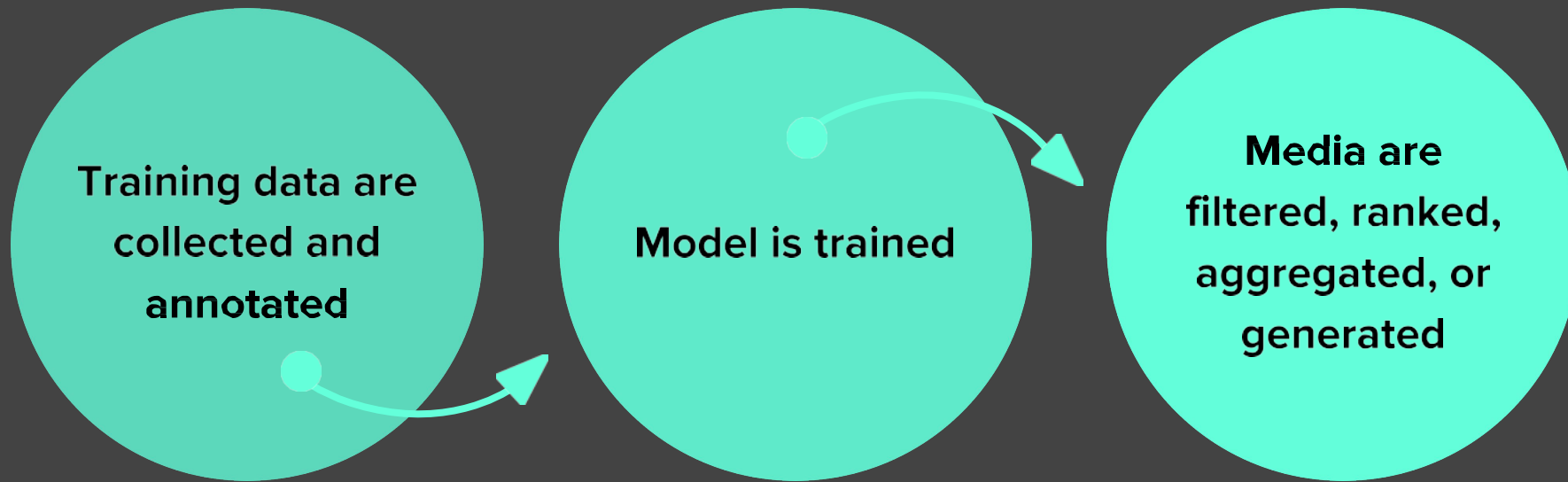
**Training data are
collected and
annotated**

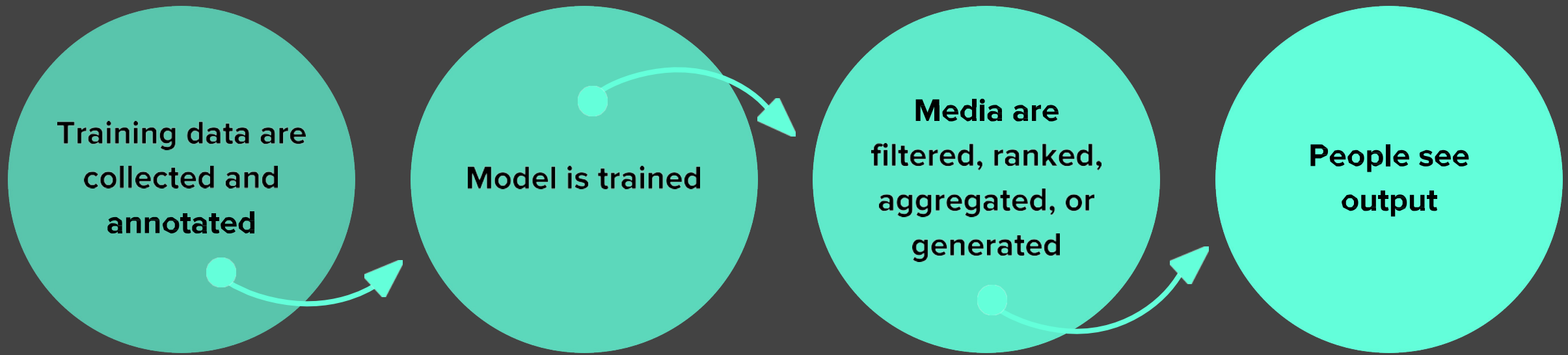


```
graph LR; A((Training data are collected and annotated)) --> B((Model is trained))
```

**Training data are
collected and
annotated**

Model is trained





Human Biases in Data

Reporting bias

Stereotypical bias

Group attribution error

Selection bias

Historical unfairness

Halo effect

Overgeneralization

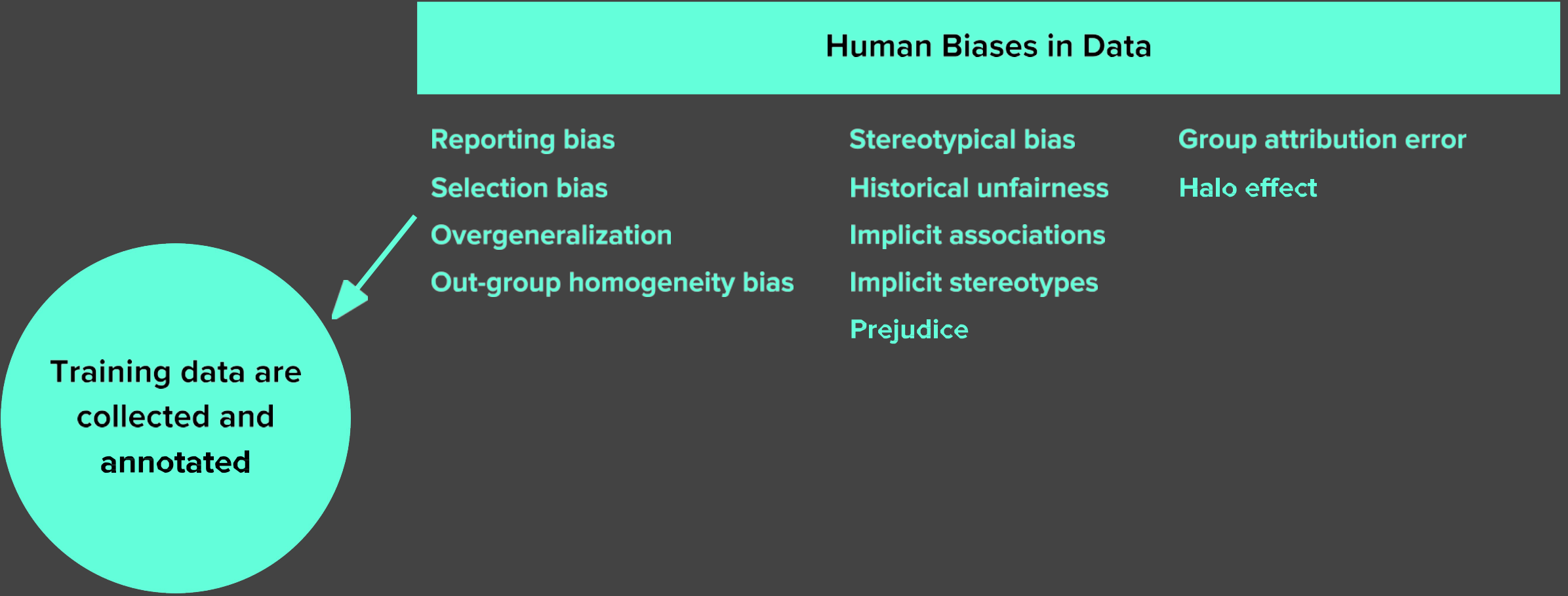
Implicit associations

Out-group homogeneity bias

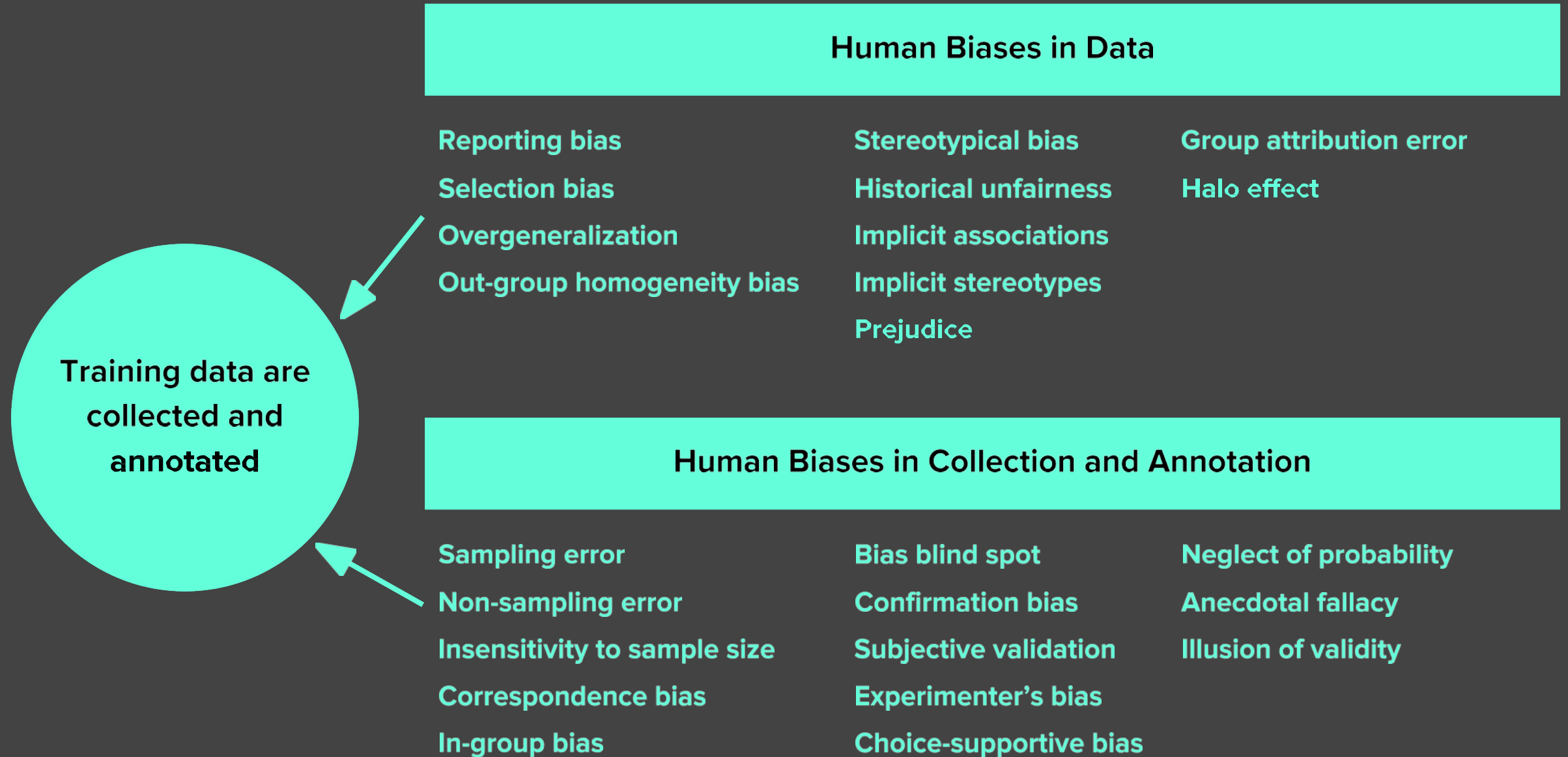
Implicit stereotypes

Prejudice

Training data are
collected and
annotated



The diagram illustrates the concept of human biases in data. On the left, a large blue circle contains the text 'Training data are collected and annotated'. An arrow points from this circle towards a list of biases on the right. The biases are organized into three columns under the heading 'Human Biases in Data'. The first column lists 'Reporting bias', 'Selection bias', 'Overgeneralization', and 'Out-group homogeneity bias'. The second column lists 'Stereotypical bias', 'Historical unfairness', 'Implicit associations', 'Implicit stereotypes', and 'Prejudice'. The third column lists 'Group attribution error' and 'Halo effect'.



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

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/>



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

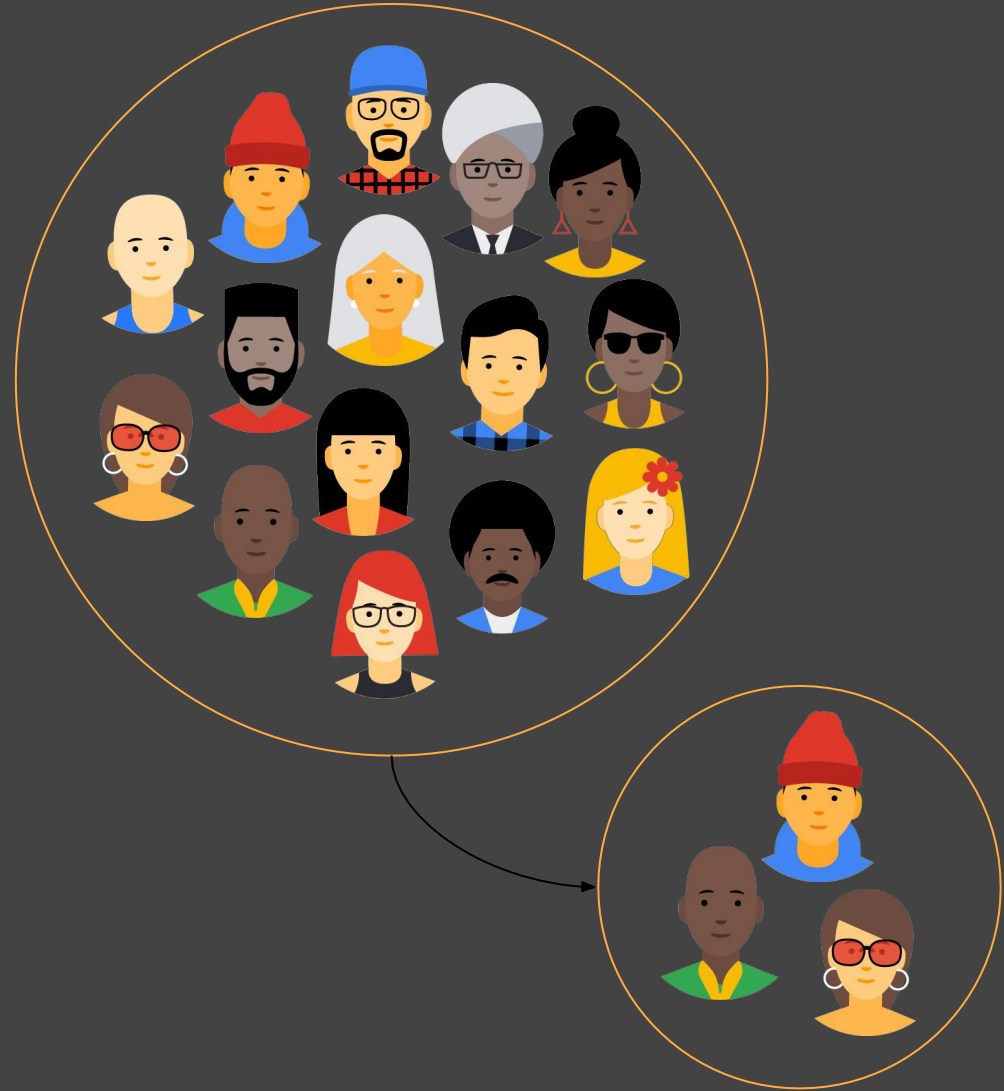
Biases in Data

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



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.

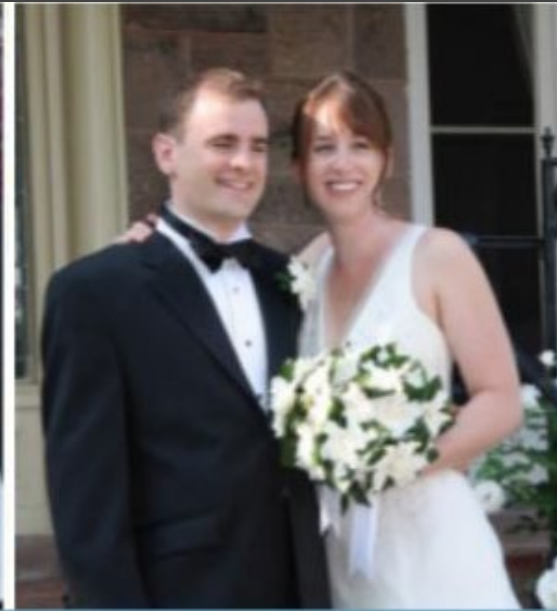


Biases in Data → Biased Labels

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



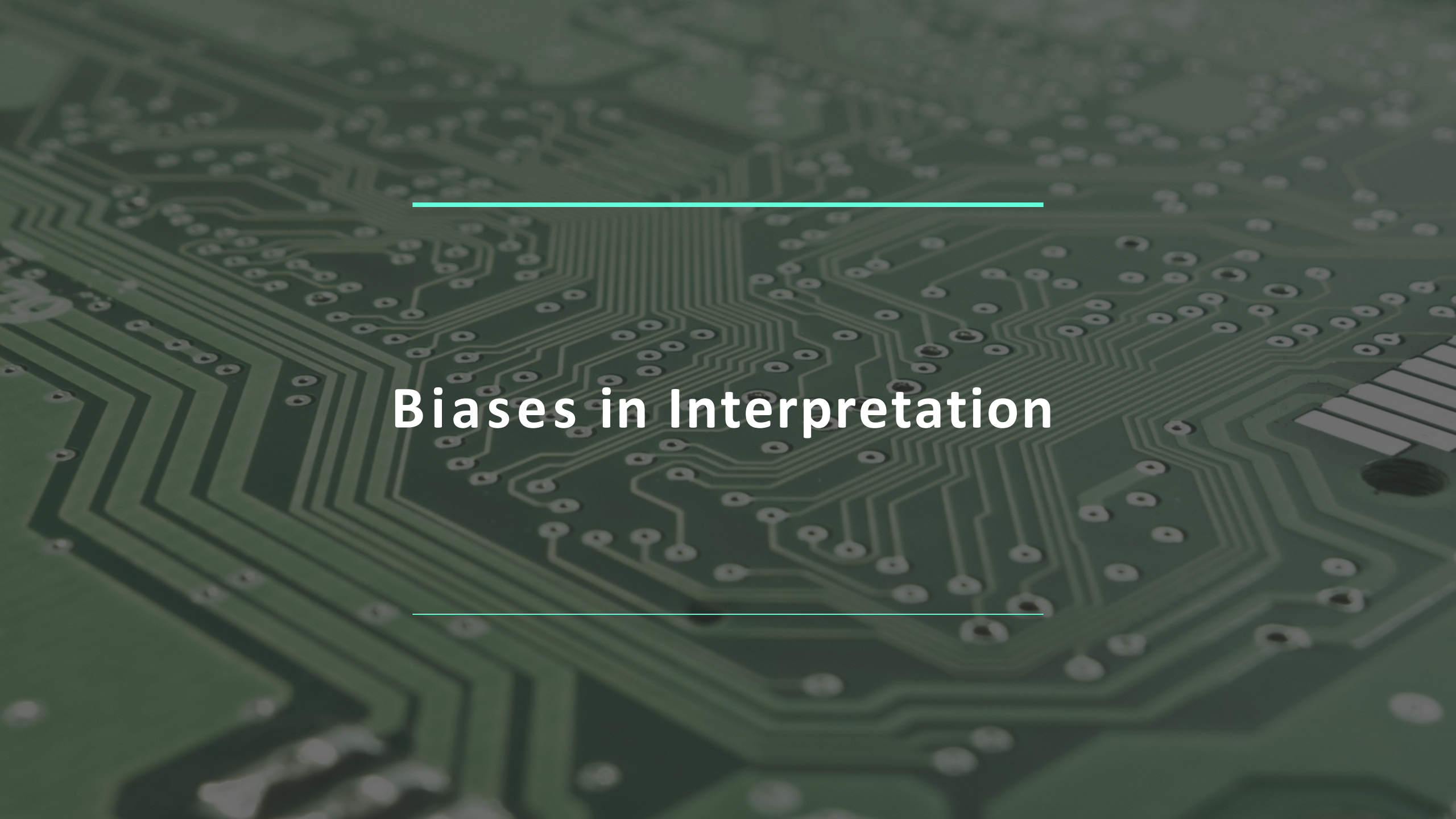
*ceremony,
wedding, bride,
man, groom,
woman, dress*



*ceremony,
bride, wedding,
man, groom,
woman, dress*



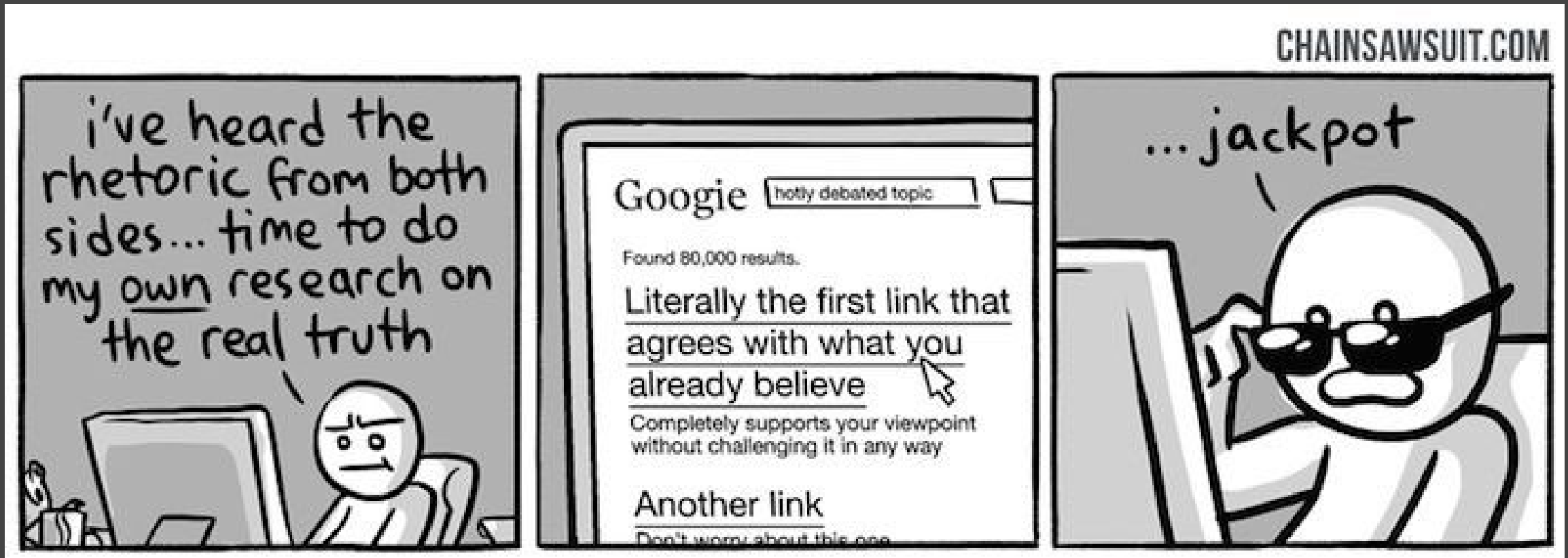
person, people



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

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.



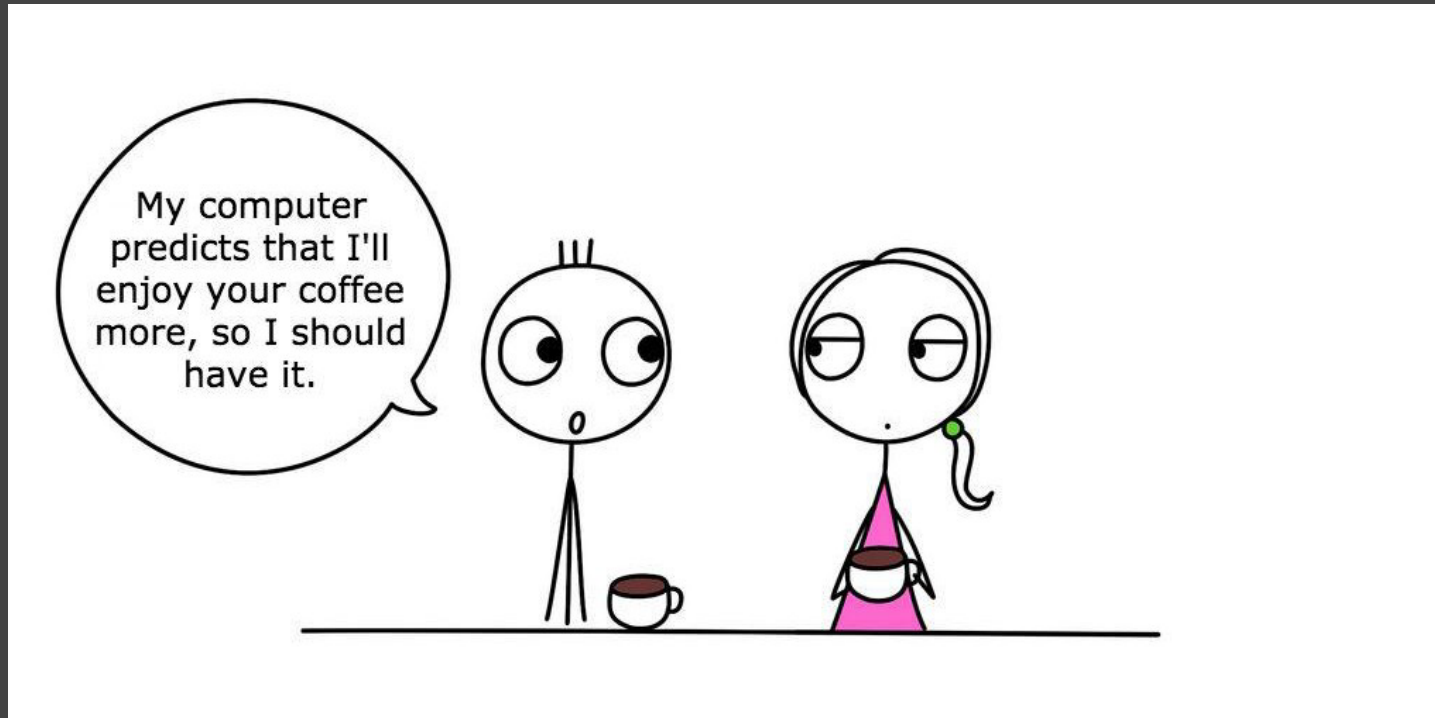
CREDIT

© mollysdad - Slideshare - Introduction to Logical Fallacies

CC BY / Margaret Mitchell / m-mitchell.com

Biases in Interpretation

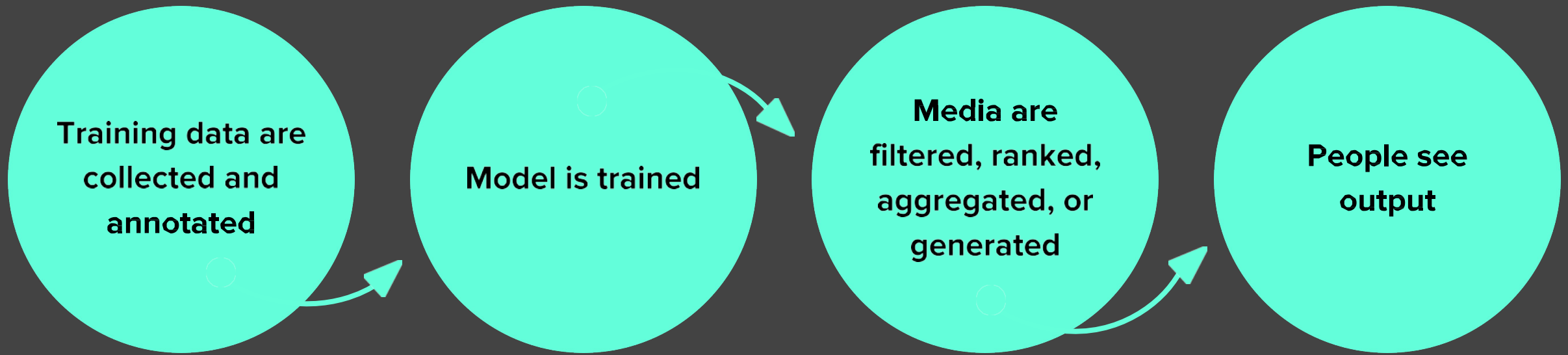
Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation



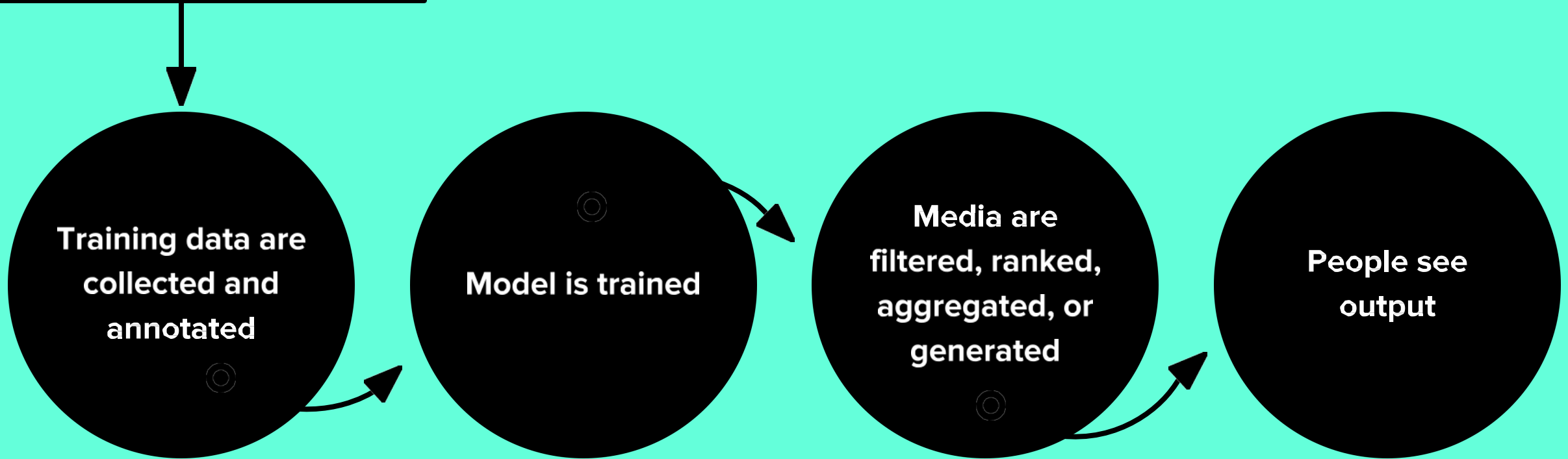
CREDIT

thedailyenglishshow.com | CC BY 2.0

CC BY / Margaret Mitchell / m-mitchell.com



Human Bias



Human Bias



Training data are
collected and
annotated



Human Bias



Model is trained



Human Bias



Media are
filtered, ranked,
aggregated, or
generated



Human Bias



People see
output

Human Bias

Human Bias

Training data are
collected and
annotated

Model is trained

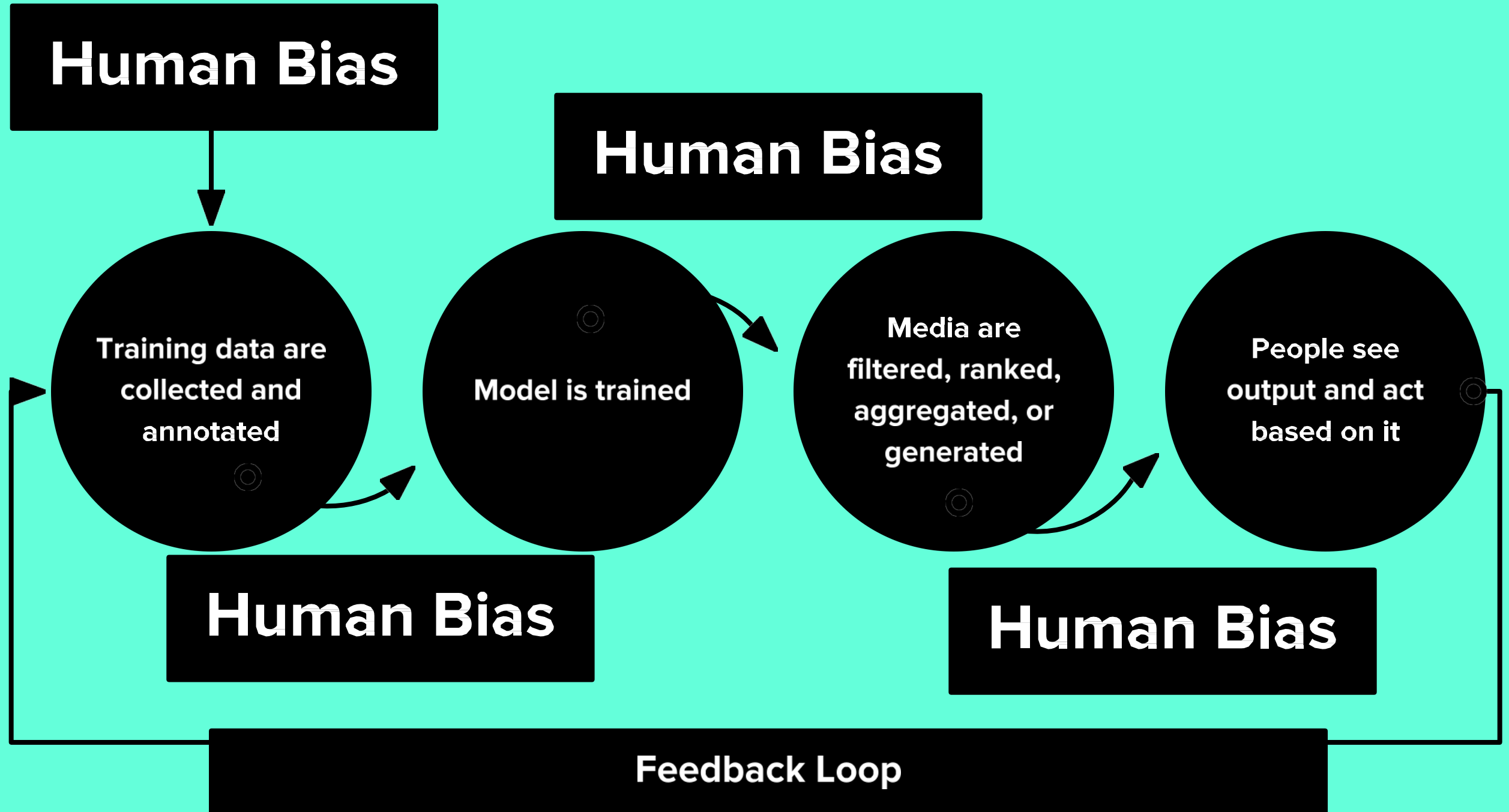
Media are
filtered, ranked,
aggregated, or
generated

People see
output and act
based on it

Human Bias

Human Bias

Feedback Loop



Human Bias



Human Bias

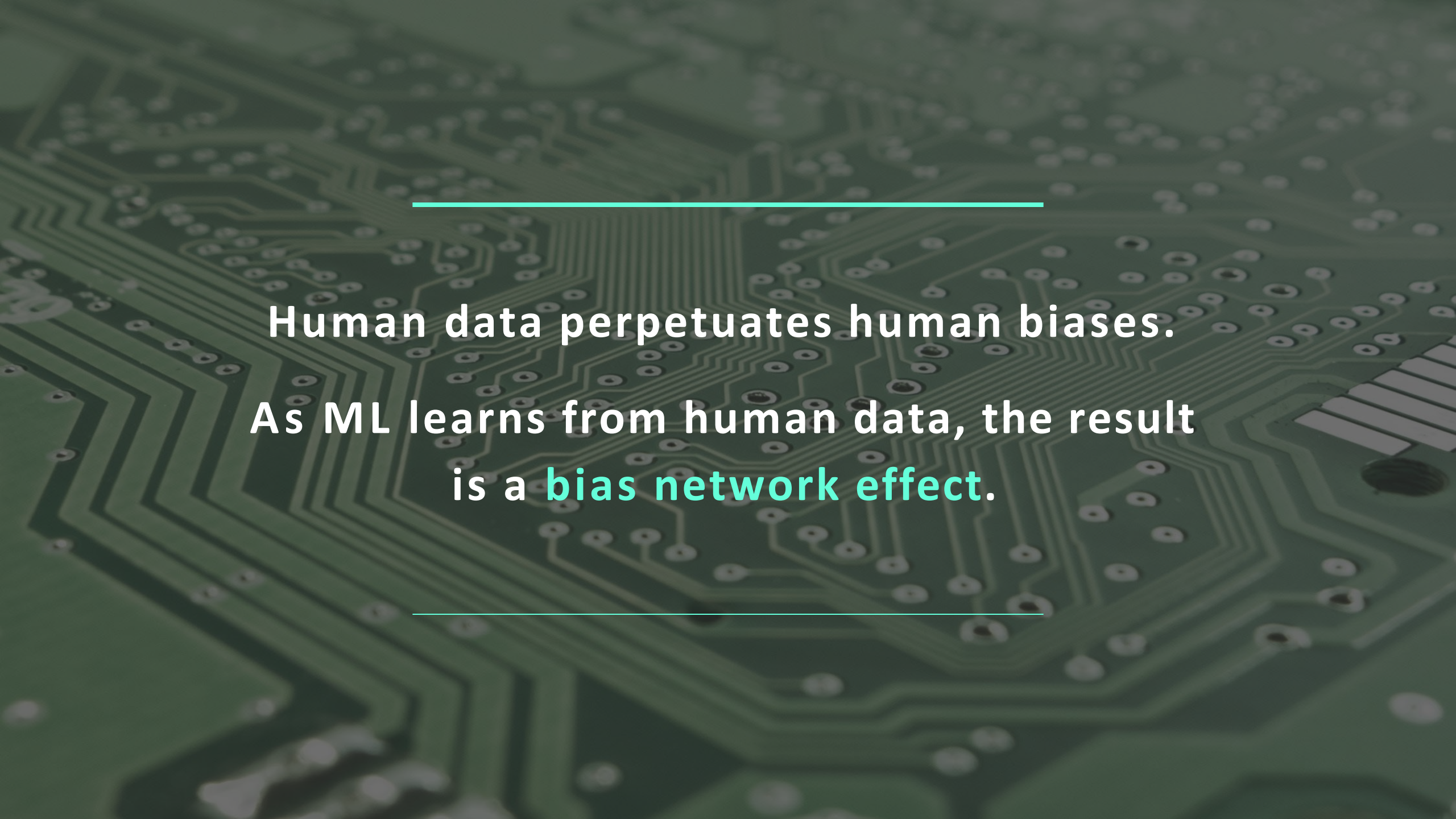
Bias Network Effect

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**.

Q1

<https://tinyurl.com/AML441-L22>

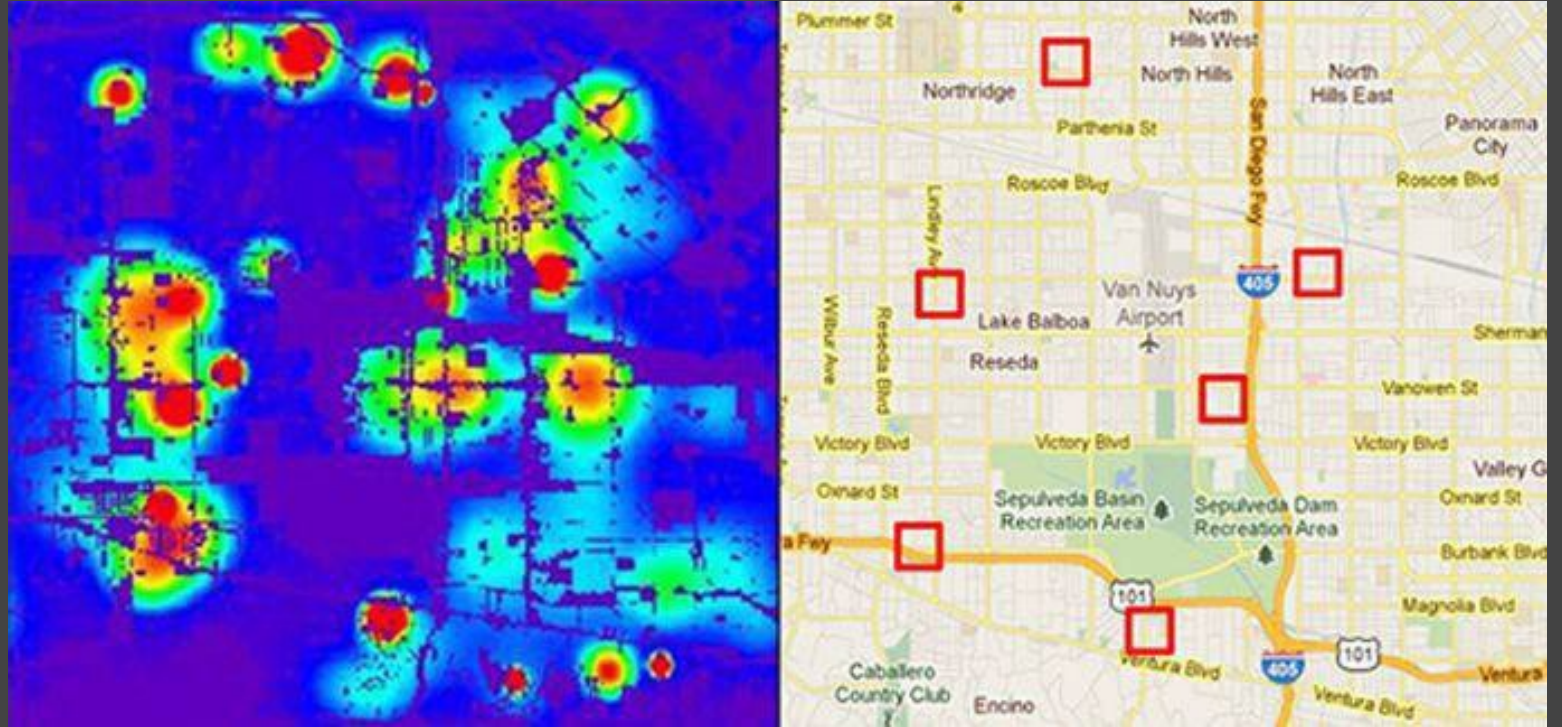




Predicting Future Criminal Behavior

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

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.

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

Here are some things we can do.



Data

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



Tools for data exploration

<https://pair.withgoogle.com/>

Facets: feature visualization libraries

<https://pair-code.github.io/facets/>

Know your data: dataset visualization

<https://knowyourdata.withgoogle.com/>

Datasheets for Datasets

Timnit Gebru¹ Jamie Morgenstern² Briana Vecchione³ Jennifer Wortman Vaughan¹ Hanna Wallach¹
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



Title COMPAS Recidivism Risk Score Data

Author Broward County Clerk's Office, Broward County Sheriff'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

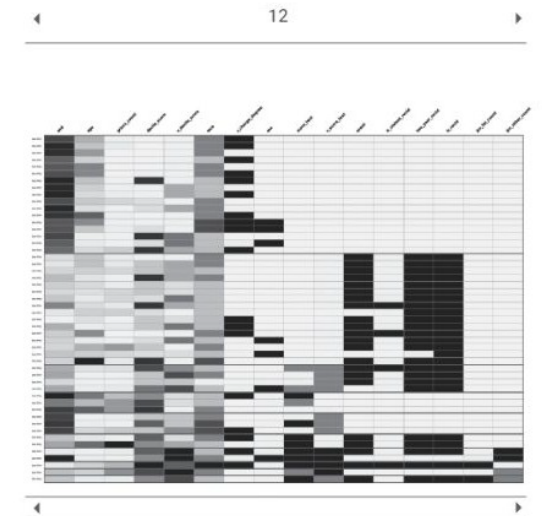
Variables 25

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

two-year-risk: *Lorem ipsum dolor sit amet, consectetur*

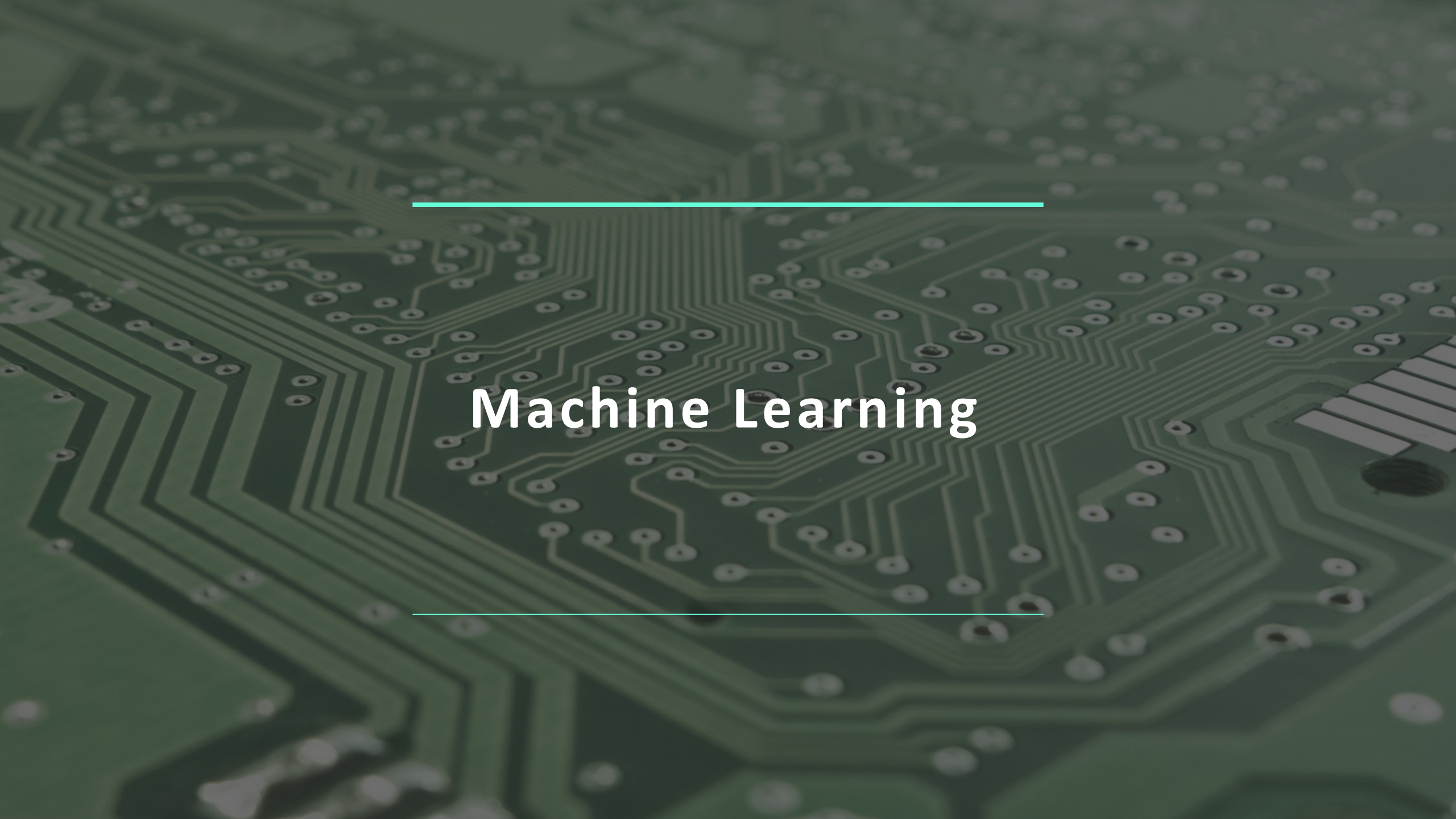
Probabilistic Modeling

Analysis



Dependency Probability Pearson R



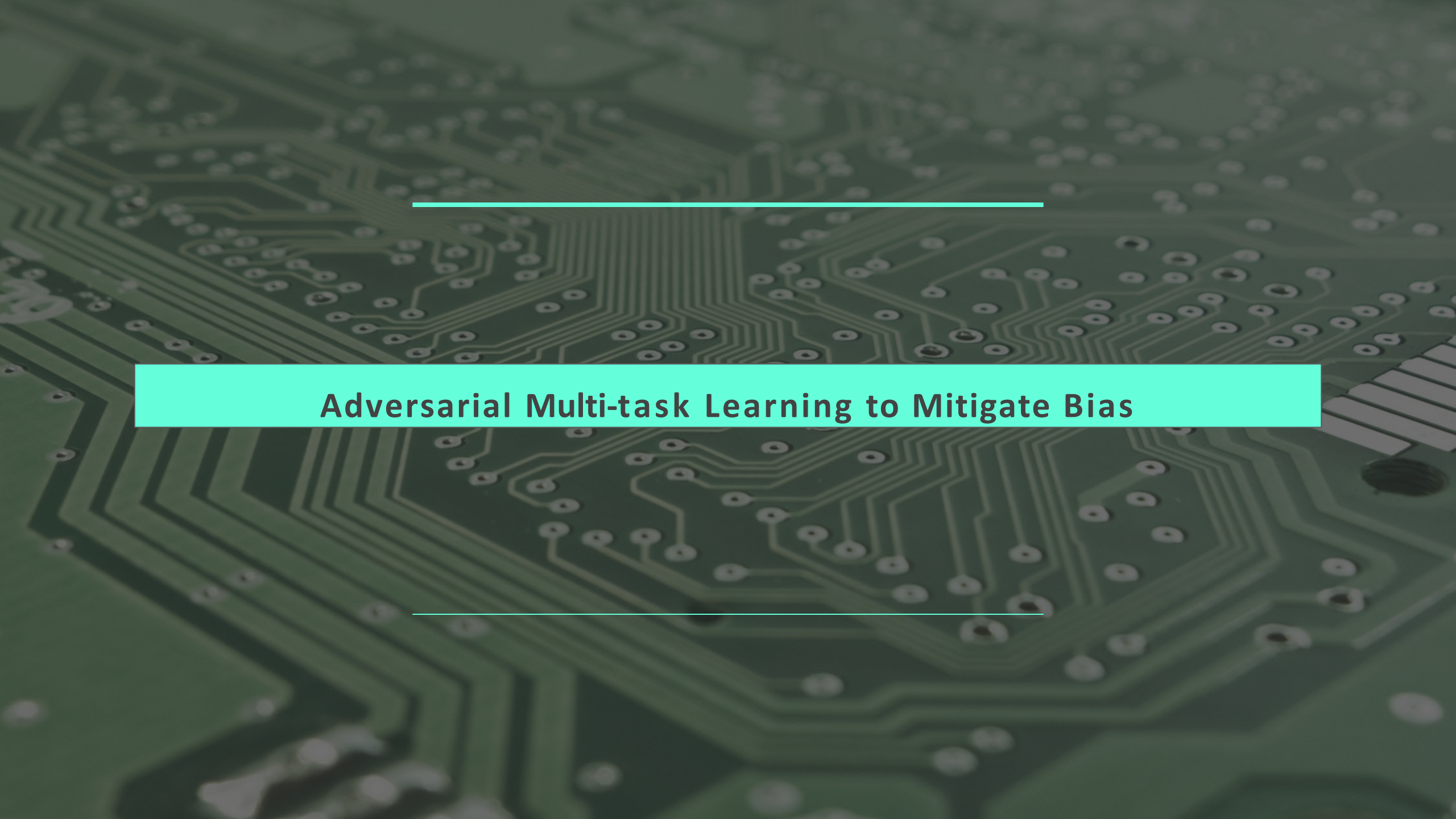


Machine Learning

Use ML Techniques for Bias Mitigation and Inclusion

Adversarial Multi-task Learning to Mitigate Bias

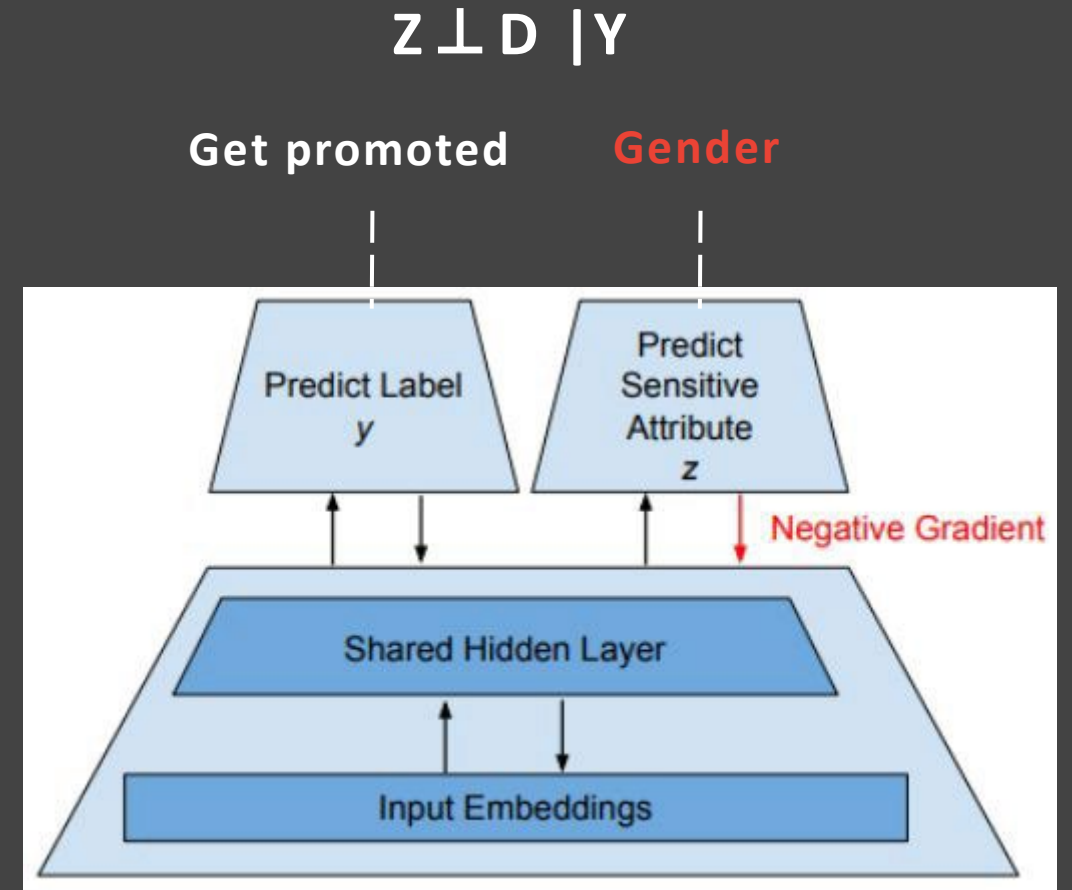
- Removing the signal for problematic output
 - Stereotyping
 - Sexism, Racism, *-ism
 - “Debiasing”



Adversarial Multi-task Learning to Mitigate Bias

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.

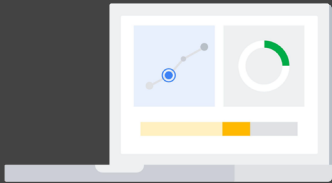


Release Responsibly

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.

Intended Use, Factors and Subgroups

Example Model Card - Toxicity in Text	
Model Details	Developed by Jigsaw in 2017 as a convolutional neural network trained to predict the likelihood that a comment will be perceived as toxic.
Intended Use	Supporting human moderation, providing feedback to comment authors, and allowing comment viewers to control their experience.
Factors	Identity terms referencing frequently attacked groups focusing on the categories of sexual orientation, gender identity and race.

Metrics and Data

Metrics	<i>Pinned AUC</i> , which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.
Evaluation Data	A synthetic test set generated using a template-based approach, where identity terms are swapped into a variety of template sentences.
Training Data	Includes comments from a variety of online forums with crowdsourced labels of whether the comment is “toxic”. “Toxic” is defined as, “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion”.

Considerations, Recommendations

Ethical Considerations	A set of values around community, transparency, inclusivity, privacy and topic-neutrality to guide their work.
Caveats & Recommendations	Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

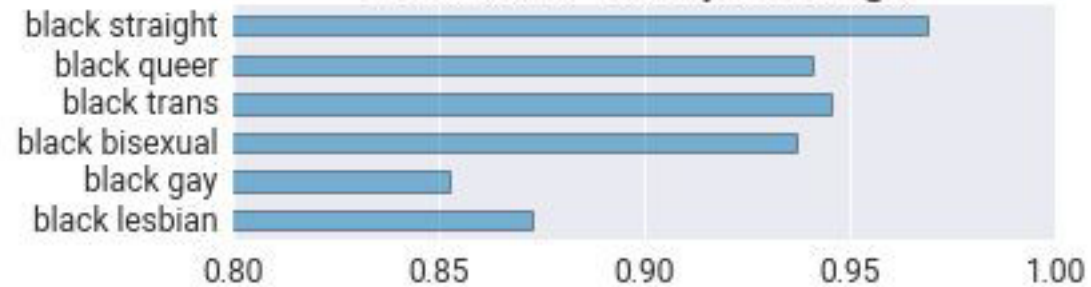
Disaggregated Intersectional Evaluation

Toxicity @1

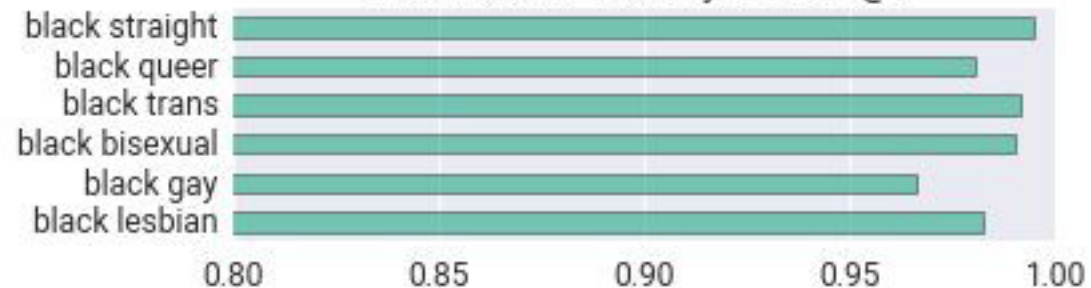
Identity groups	Subgroup AUC	BPSN AUC	BNSP AUC
lesbian	0.93	0.74	0.98
gay	0.94	0.65	0.99
queer	0.98	0.96	0.93
straight	0.99	1.00	0.87
bisexual	0.96	0.95	0.92
homosexual	0.87	0.53	0.99
heterosexual	0.96	0.94	0.92
cis	0.99	1.00	0.87
trans	0.97	0.96	0.91
nonbinary	0.99	0.99	0.90
black	0.91	0.85	0.95
white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



Pinned AUC Toxicity Scores @1



Pinned AUC Toxicity Scores @5



Jigsaw



The False Positive

Q2

<https://tinyurl.com/AML441-L22>



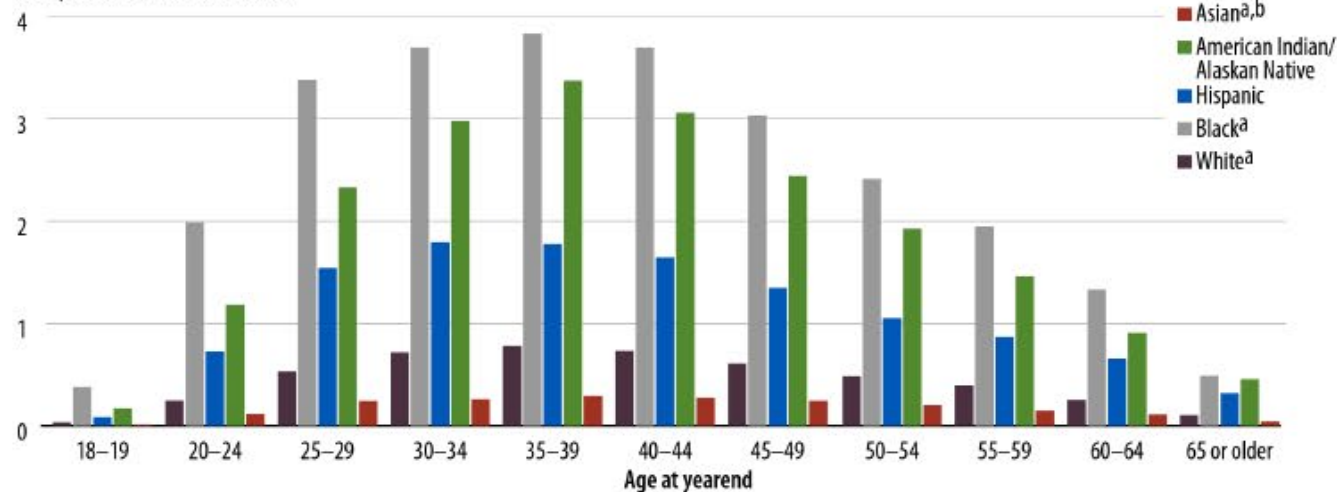
Three goals: unbiased, fair, and just

Bias

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

Imprisonment rates of male U.S. residents, based on sentenced prisoners under the jurisdiction of state or federal correctional authorities, by demographic characteristics, December 31, 2021

Rate per 100,000 (in thousands)



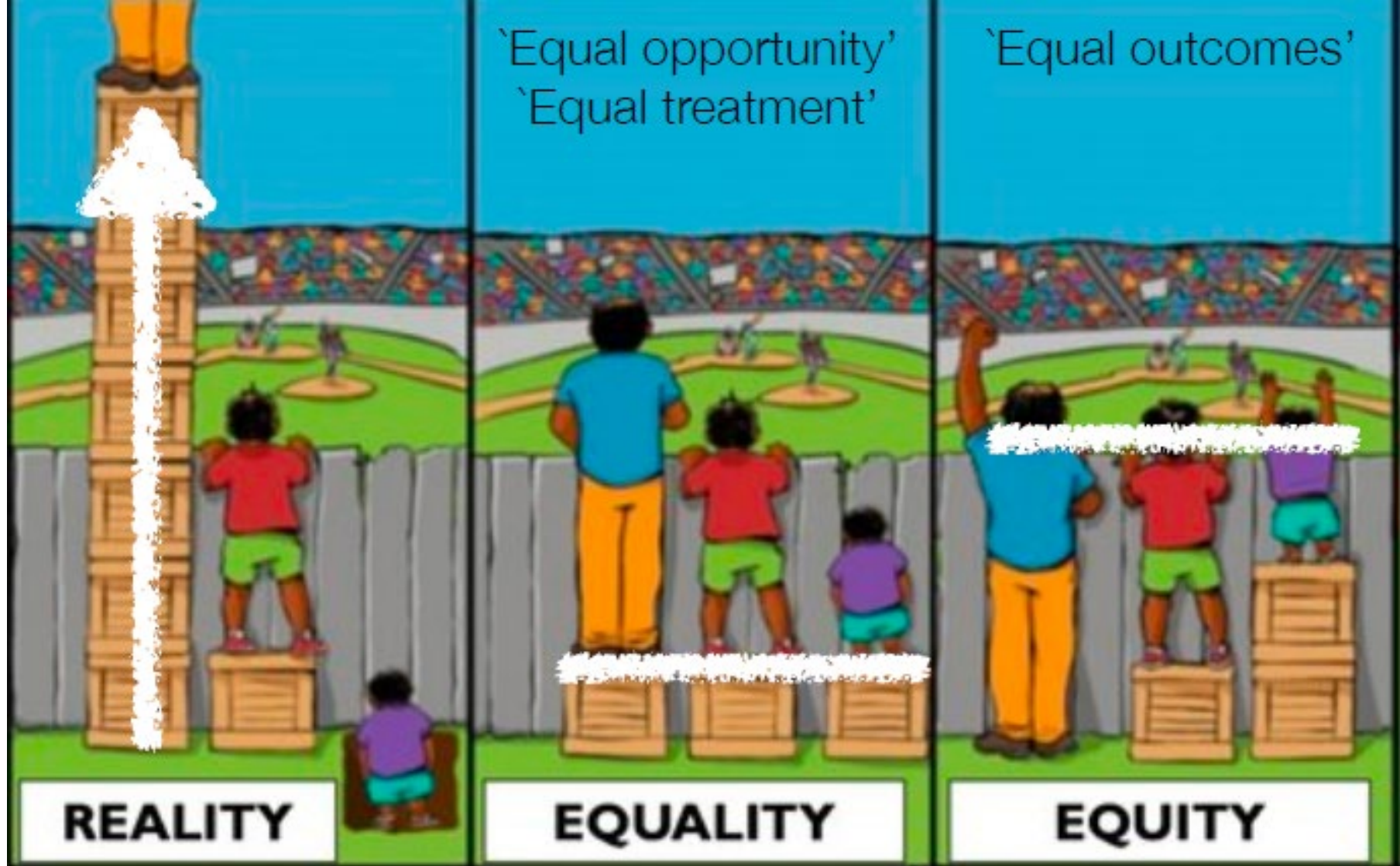
Fairness

*Even-handed allocation
of resources*

Justice

*Everyone gets what they
deserve*





'Equal opportunity'
'Equal treatment'

'Equal outcomes'

REALITY

EQUALITY

EQUITY

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

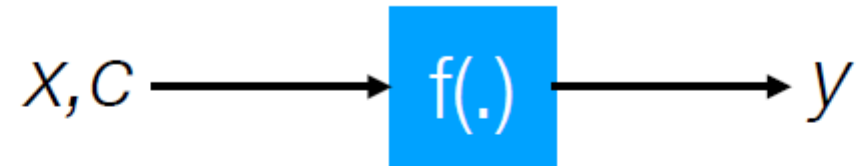
The assumption is that **everyone benefits from the same supports**. This is considered to be equal treatment.

Everyone gets the support they need, which produces equity.

[@restoringracialjustice]

Fairness always needs to be defined wrt some “protected attributes”

Protected attributes



$x = \text{action, state}$

$c = \text{protected attribute (e.g. sex, age, race)}$

$y = \text{outcome (e.g. resource, punishment)}$

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

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

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

Equal outcome or equal opportunity?

Dressage

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

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



- 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 gets 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?

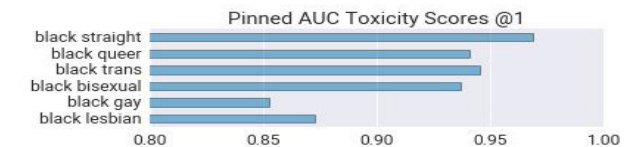
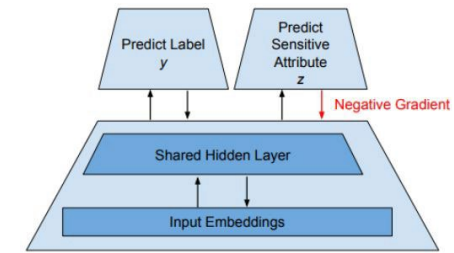
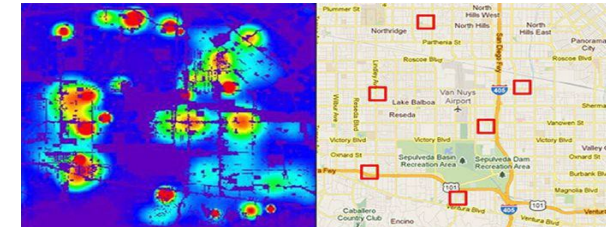
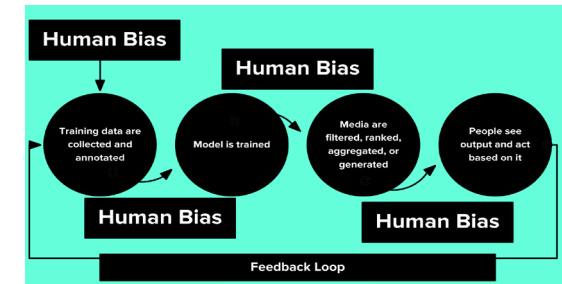
Q3-Q6

<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: Audio and 1D Signals
- Thurs: Guest lecture by Chenxi Wu on “Building and Deploying ML”