NUS | Computing
National University of Singapore

CS5562: Trustworthy Machine Learning

Part III Lecture 1: Fairness $\rightarrow$ Bias in Machine Learning

Reza Shokri[a]

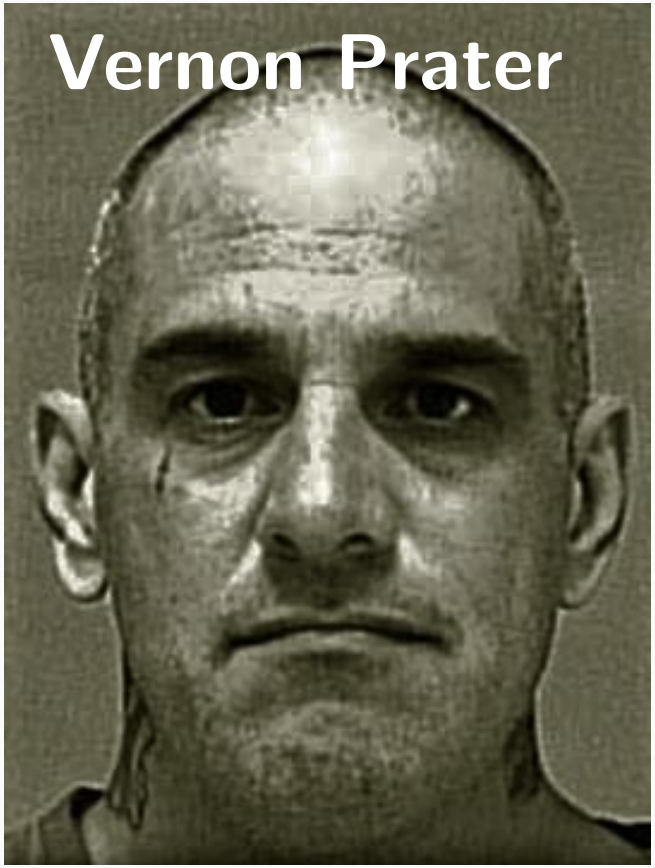
Aug 2023

[a]Acknowledgment. The wonderful teaching assistants: Hongyan Chang, Martin Strobel, Jiashu Tao, Yao Tong, Jiayuan Ye

Contents

Machine Bias

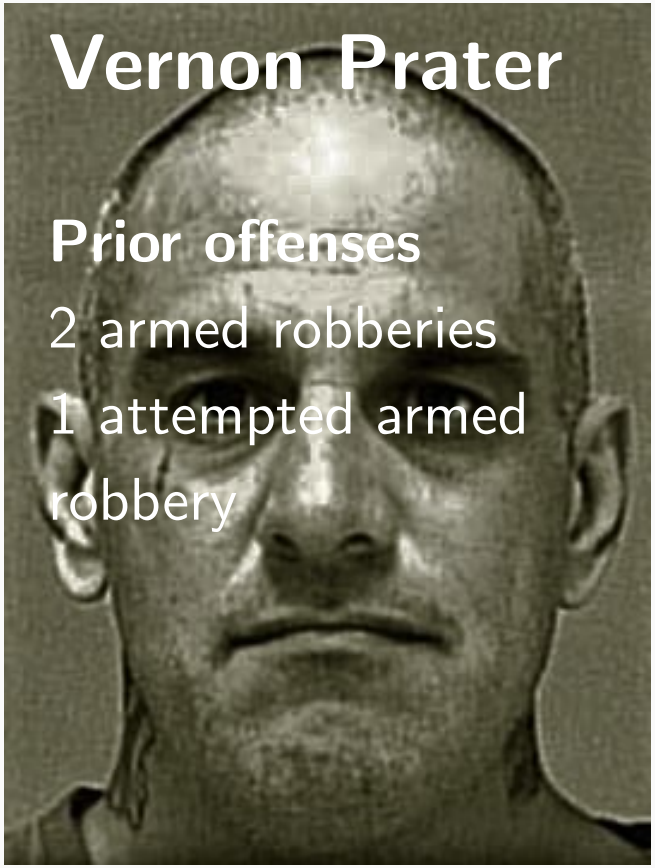
A software was used to predict the likelihood of any defendant committing a future crime (in Broward County, Florida).



Source: [Angwin et al., 2016]

Machine Bias

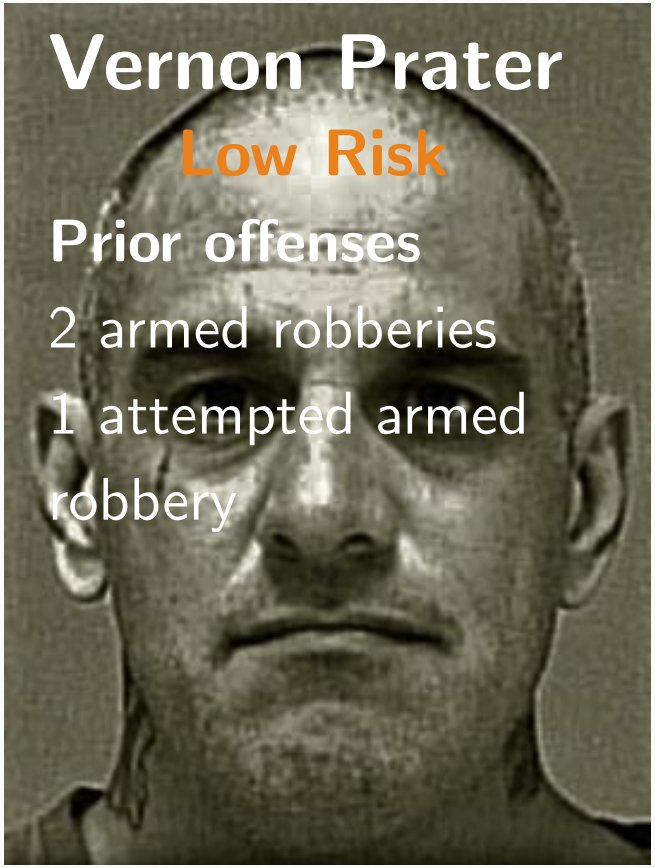
A software was used to predict the likelihood of any defendant committing a future crime (in Broward County, Florida).



Source: [Angwin et al., 2016]

Machine Bias

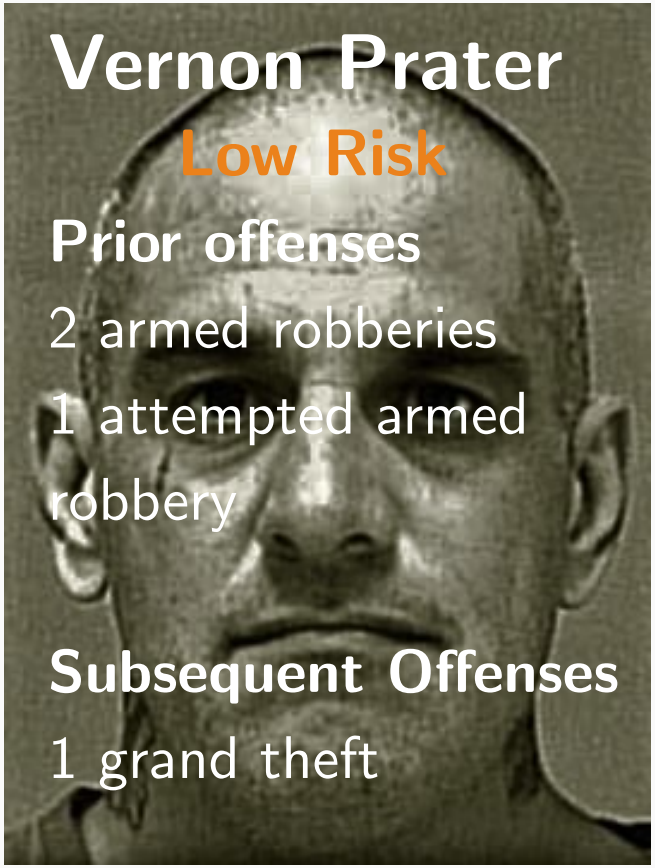
A software was used to predict the likelihood of any defendant committing a future crime (in Broward County, Florida).



Source: [Angwin et al., 2016]

Machine Bias

A software was used to predict the likelihood of any defendant committing a future crime (in Broward County, Florida).



Source: [Angwin et al., 2016]

The COMPAS dataset

This is not limited to a small number of cases.

Source: [Angwin et al., 2016]

The COMPAS dataset

This is not limited to a small number of cases.

	White	**African-American**
Accuracy	67.0%	63.8%

Source: [Angwin et al., 2016]

The COMPAS dataset

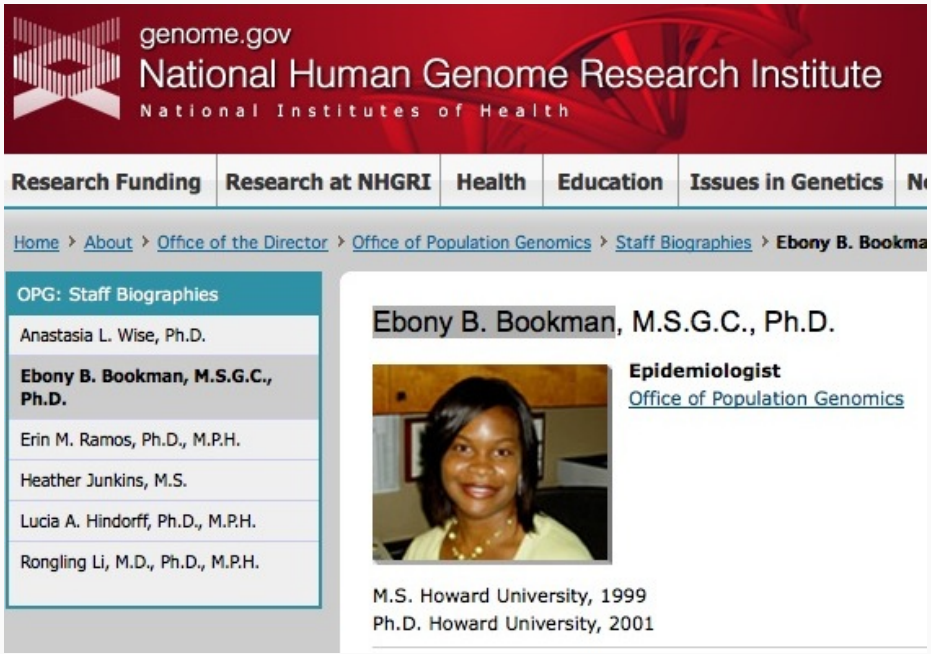
This is not limited to a small number of cases.

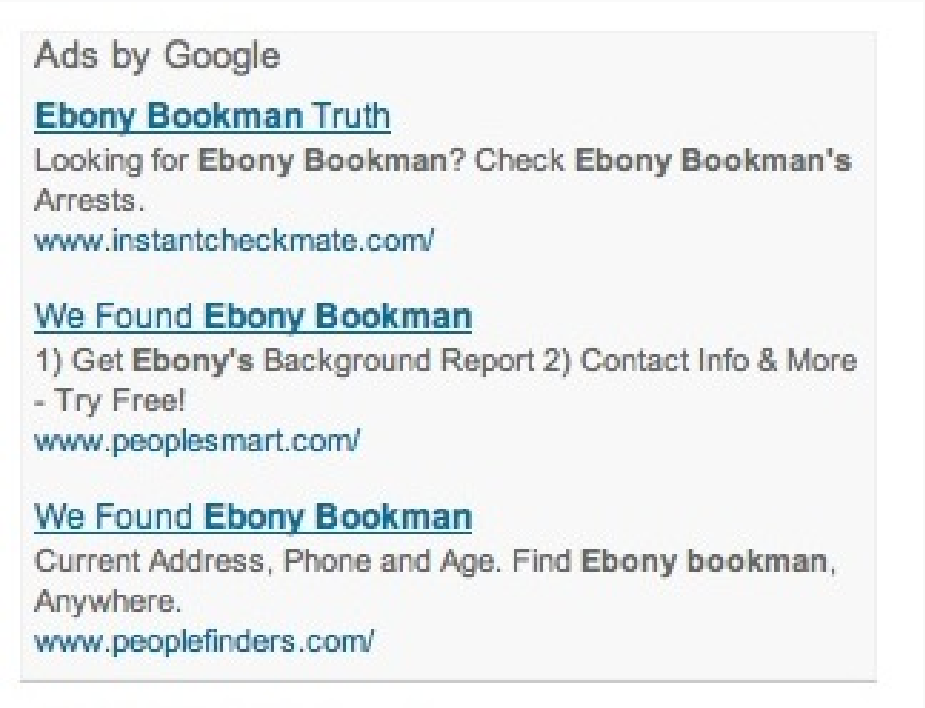
	White	**African-American**
Accuracy	67.0%	63.8%
Labeled High Risk, Didn't Re-Offend	23.5%	44.9%
Labeled Low Risk, Did Re-Offend	47.7%	28.0%

Source: [Angwin et al., 2016]

Online Ad Delivery

Google searches for predominately "black names" are more likely to show ads containing the word **arrest**.





Source: [Sweeney, 2013]

Sources and Types of Bias in ML

What is Bias?

For us

$$\text{Bias} = \text{Source of harm in an ML system}$$

Avoiding bias when making decisions

Practical irrelevance
Society believes the group membership is practically irrelevant for the decision.

Example
For most jobs, a candidate's religious beliefs are practically irrelevant.

Source: M. Hardt [2020]

Avoiding bias when making decisions

Practical irrelevance
Society believes the group membership is practically irrelevant for the decision.

Example
For most jobs, a candidate's religious beliefs are practically irrelevant.

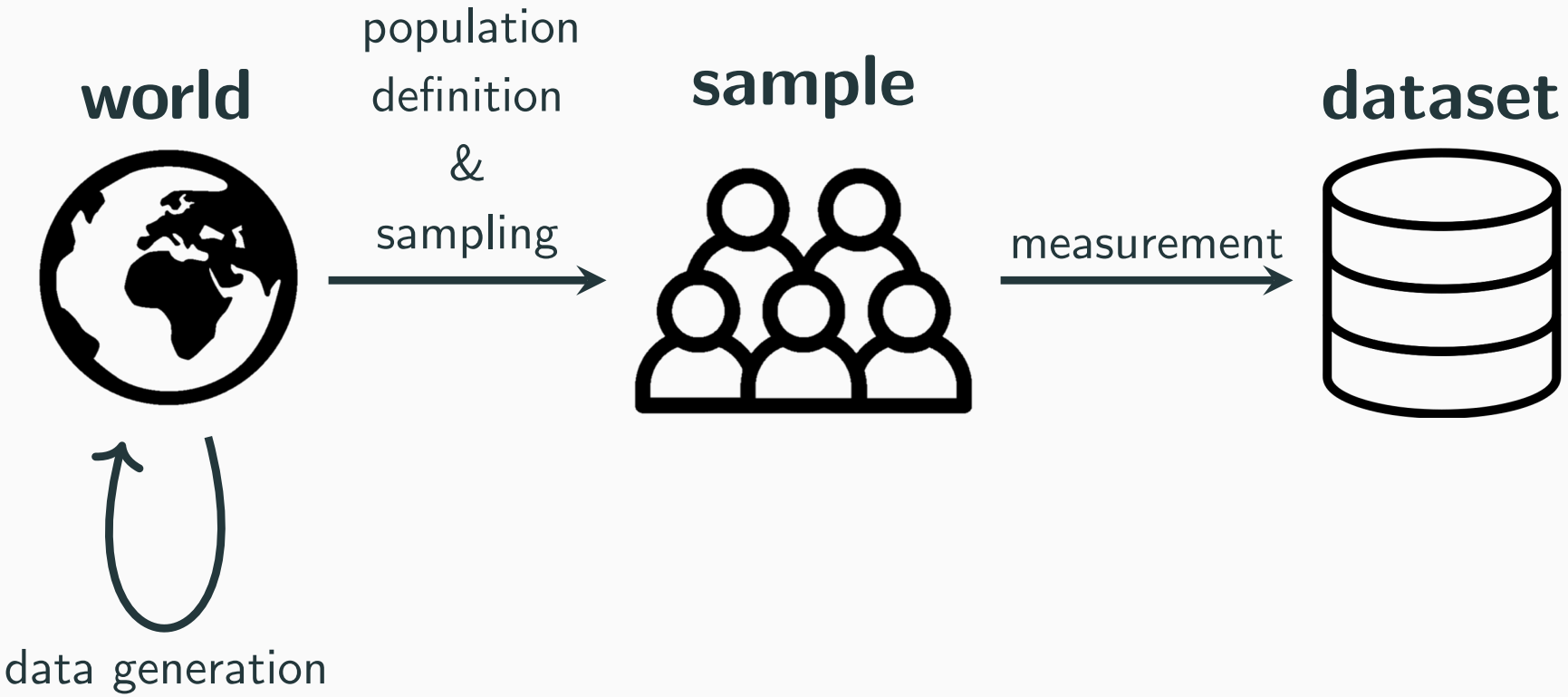
Moral irrelevance
Society believes the group membership is morally irrelevant for the decision even if it has predictive value.

Example
Even if a disability leads to higher costs for a company, it shouldn't be considered during the application process.

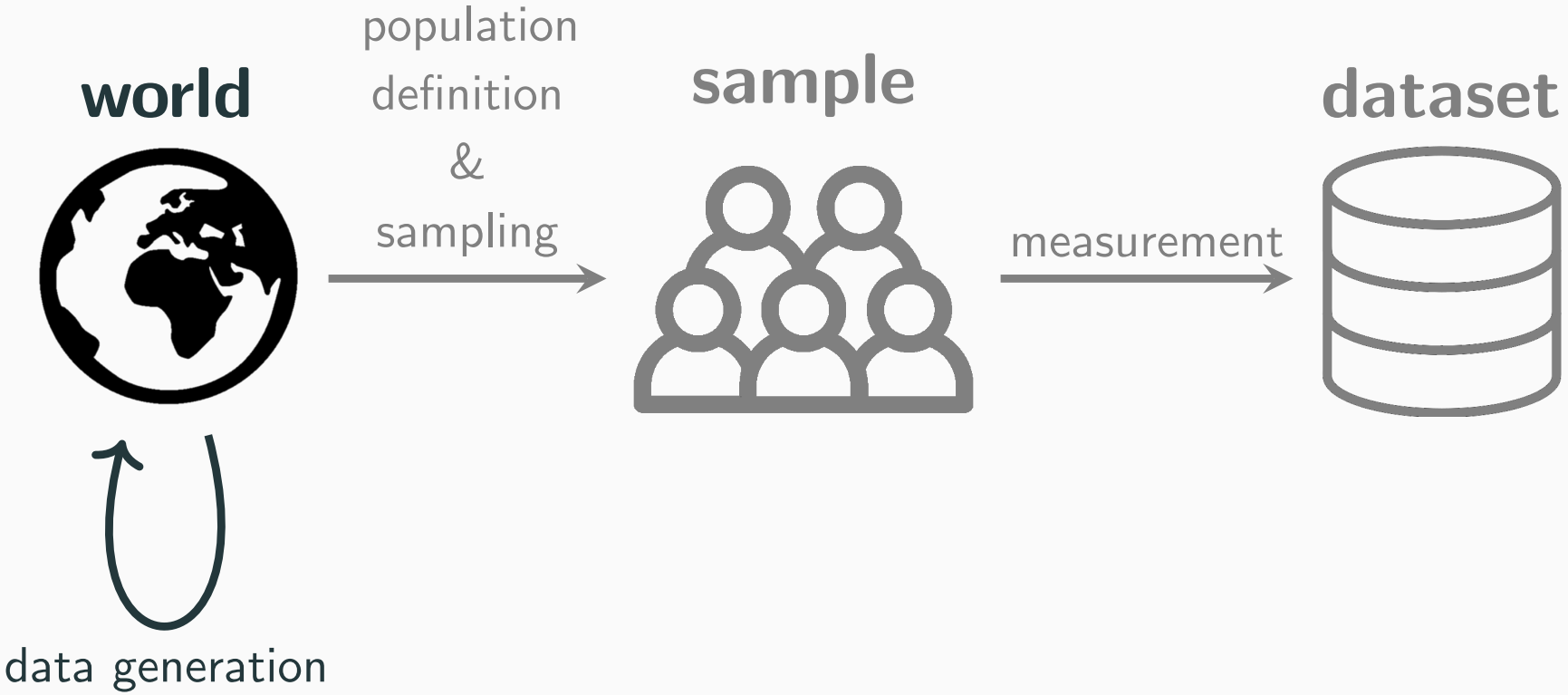
Source: M. Hardt [2020]

Data generation pipeline

To understand some origins of bias we need to look at the data generation pipeline.



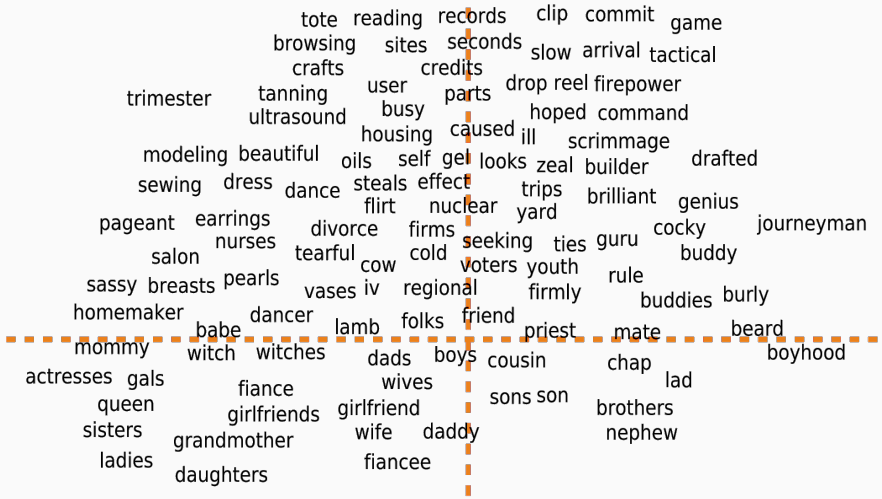
Data generation pipeline



Historical bias

Historical bias
arises when the world as is or was leads to a model that produces harmful outcomes.

Word embeddings
reflect association of words to men and women (and ethnic groups)

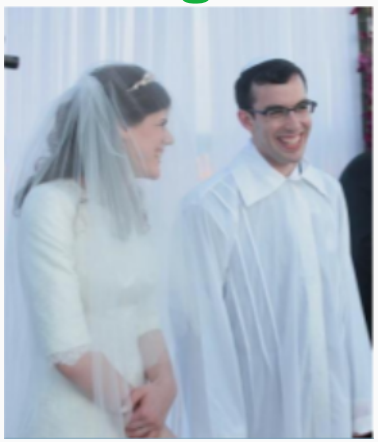


x is a projection onto the difference between the embeddings of the words he and she

Source: [Bolukbasi et al., 2016]

So what if there is no historical bias?

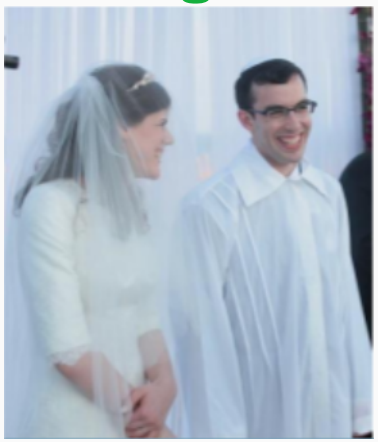
Weddings




Source: ai.googleblog.com

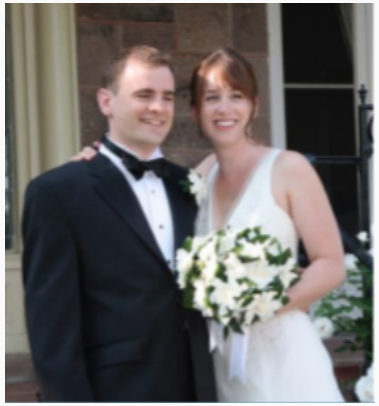
So what if there is no historical bias?

Weddings



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

bride, ceremony, wedding, dress, woman

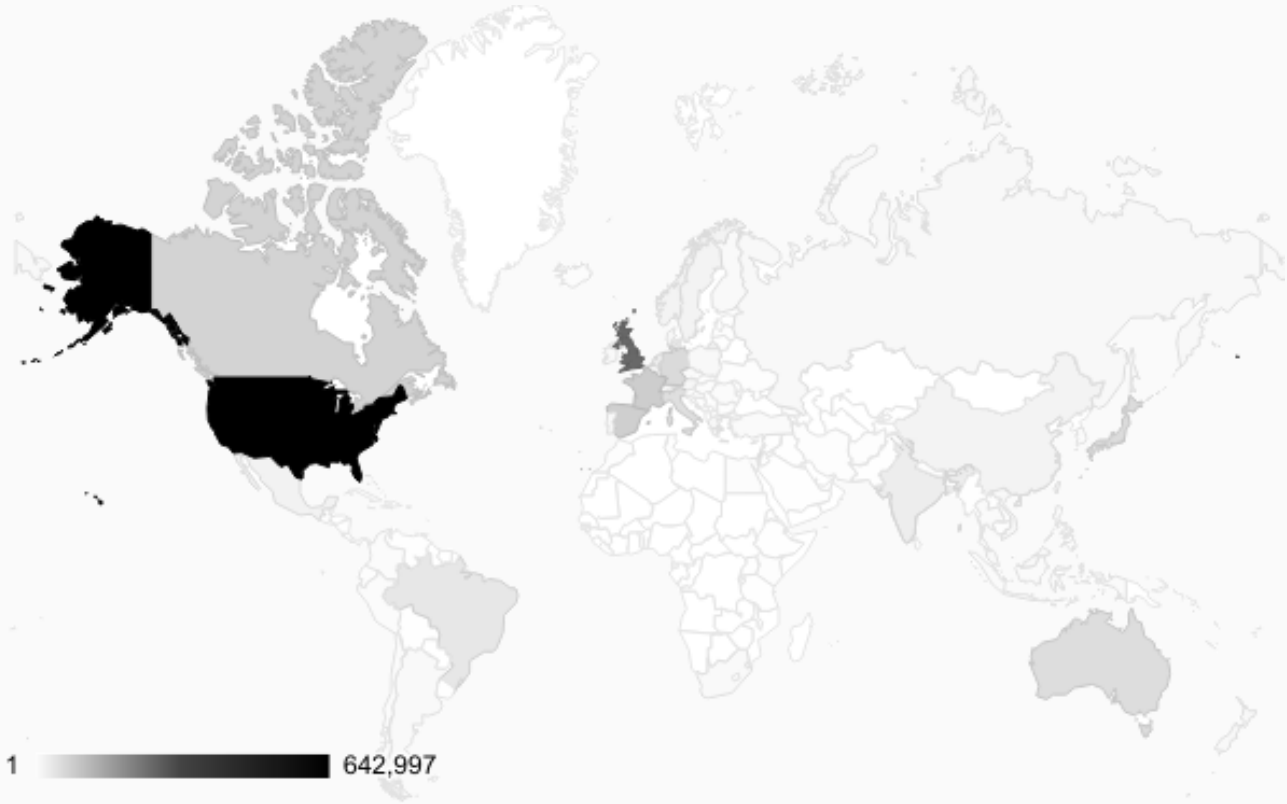


ceremony, bride, wedding, man, groom, woman, dress

person, people

Source: ai.googleblog.com

Open Images Distribution

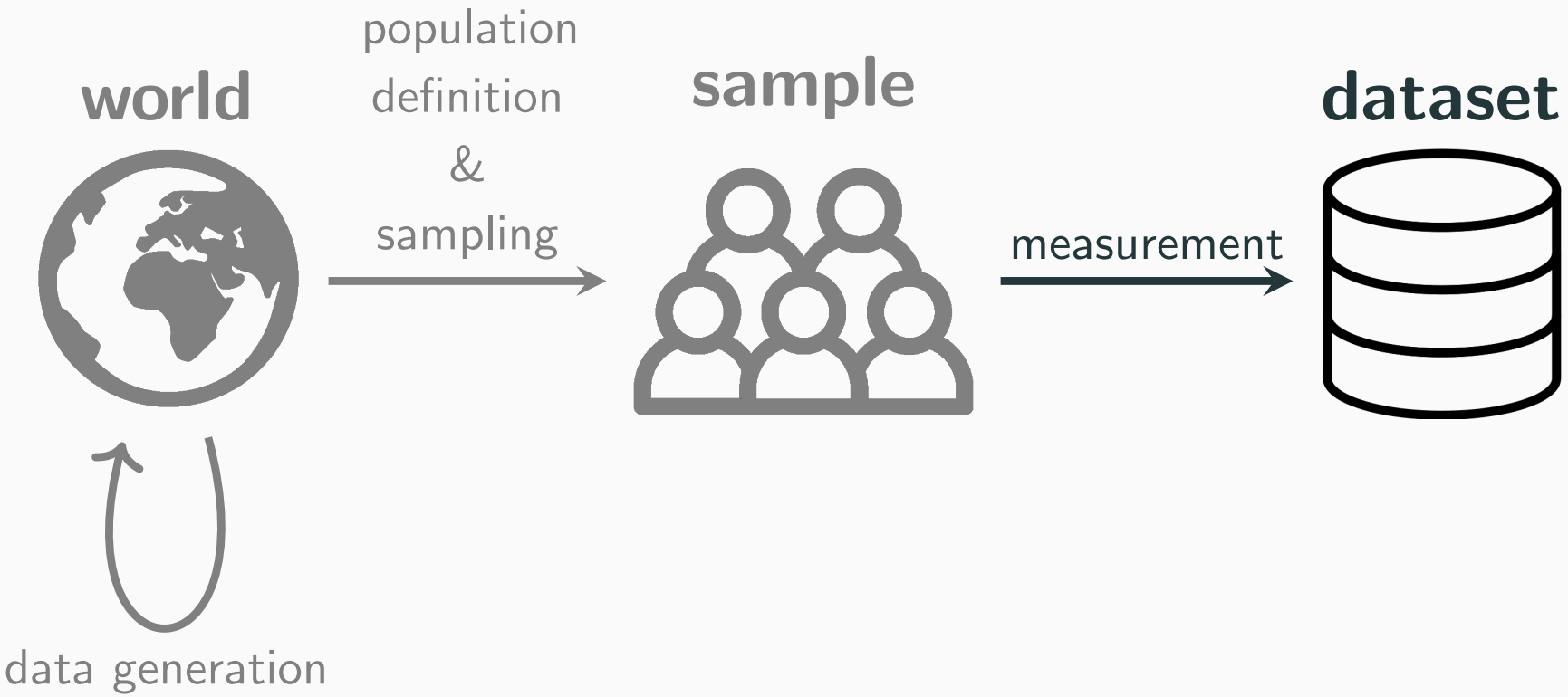


Source: [Shankar et al., 2017]

Representation bias

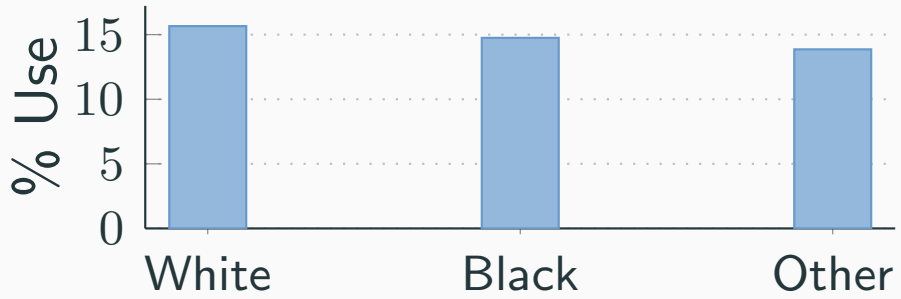
Representation bias
a sample underrepresents some part of the population, and subsequently fails to generalize well.

Data generation pipeline



Sampling from the right distribution

The percent of **drug use** in the population is estimated to be very similar across racial groups in Oakland:

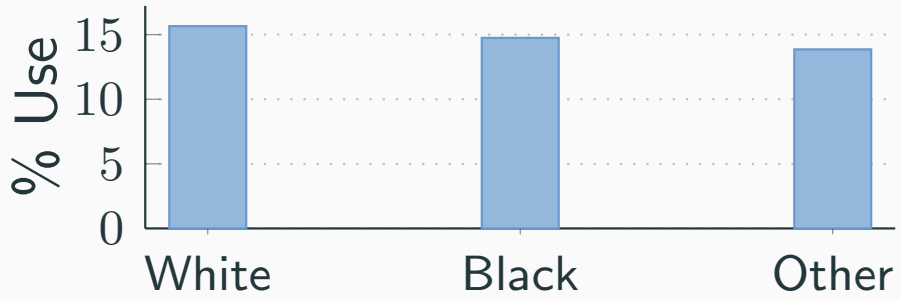


How would a predictive policing algorithm trained on historic **drug crime** data police the city?

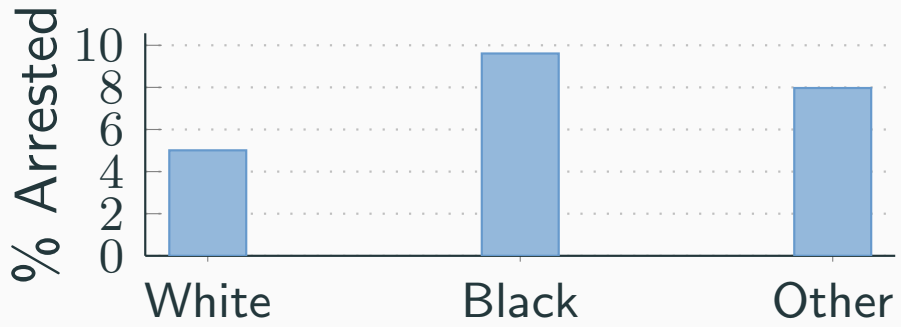
Source: [Lum and Isaac, 2016]

Sampling from the right distribution

The percent of **drug use** in the population is estimated to be very similar across racial groups in Oakland:



How would a predictive policing algorithm trained on historic **drug crime** data police the city?



Source: [Lum and Isaac, 2016]

Measurement bias

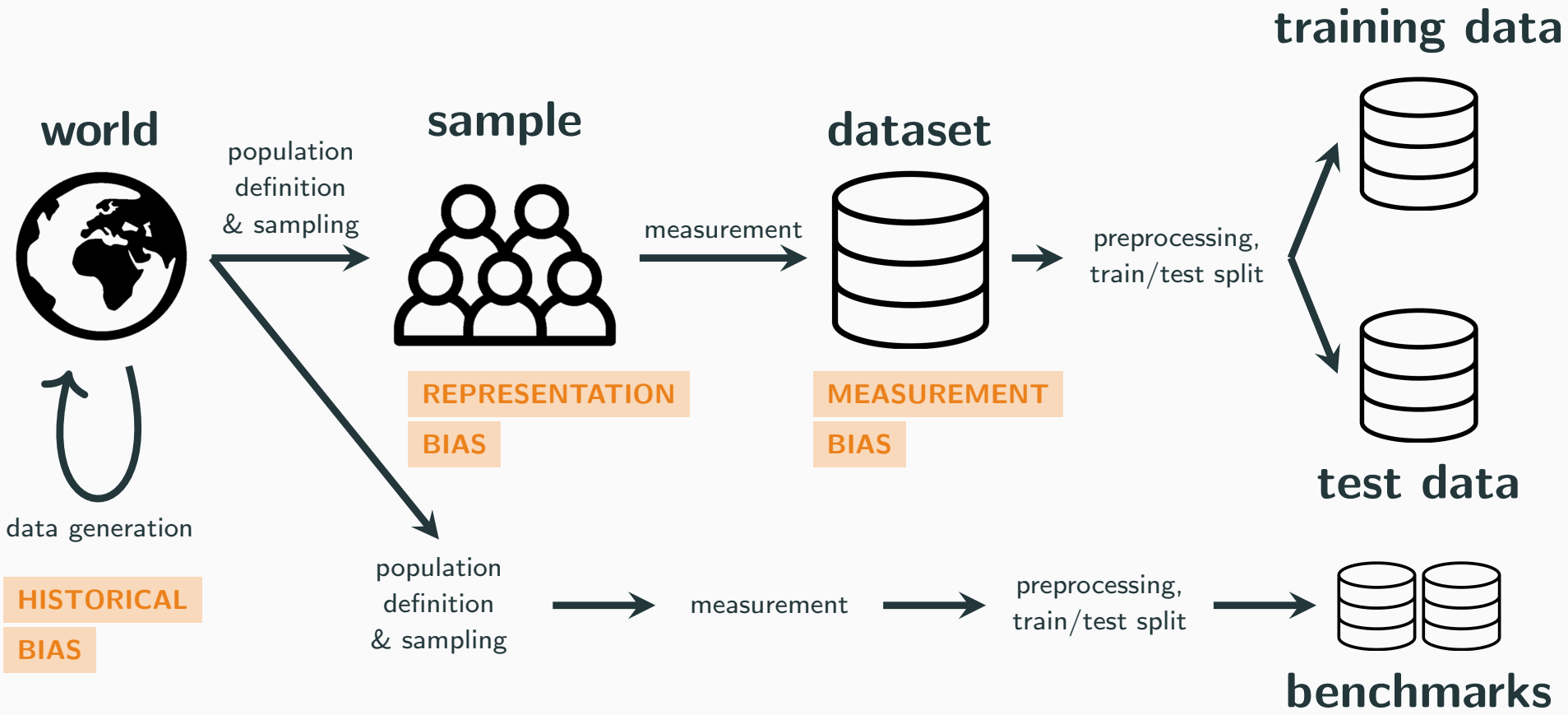
Measurement bias
occurs when choosing, collecting, or computing features and labels to use in a prediction problem.

Proxies become problematic when:

- The proxy is an oversimplification of a more complex construct
- The method of measurements varies across groups
- The accuracy of measurement varies across groups

Biased arrest data $\rightarrow$ Biased crime data

Data generation

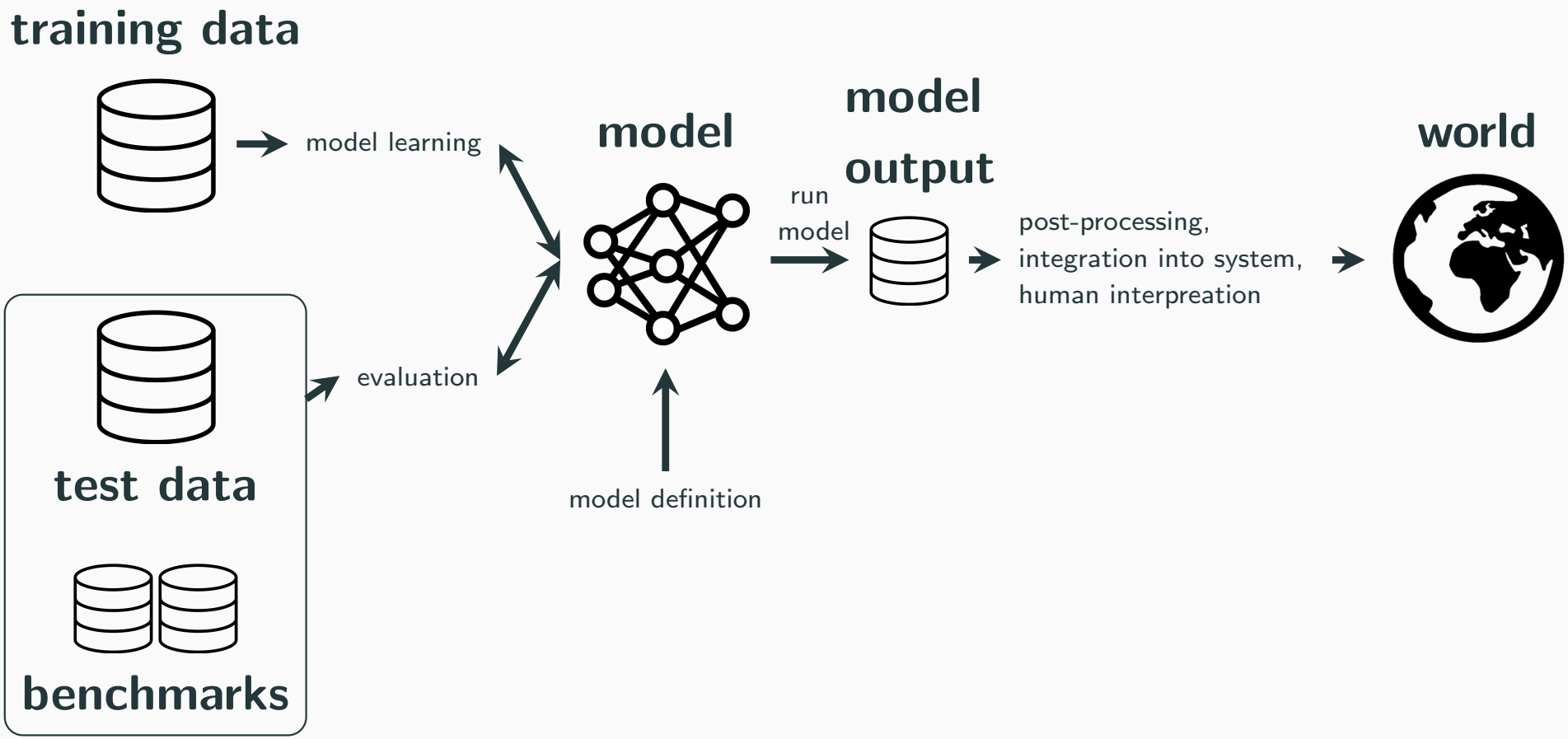


Source: [Suresh and Guttag, 2021]

Biased data = Biased classifier?

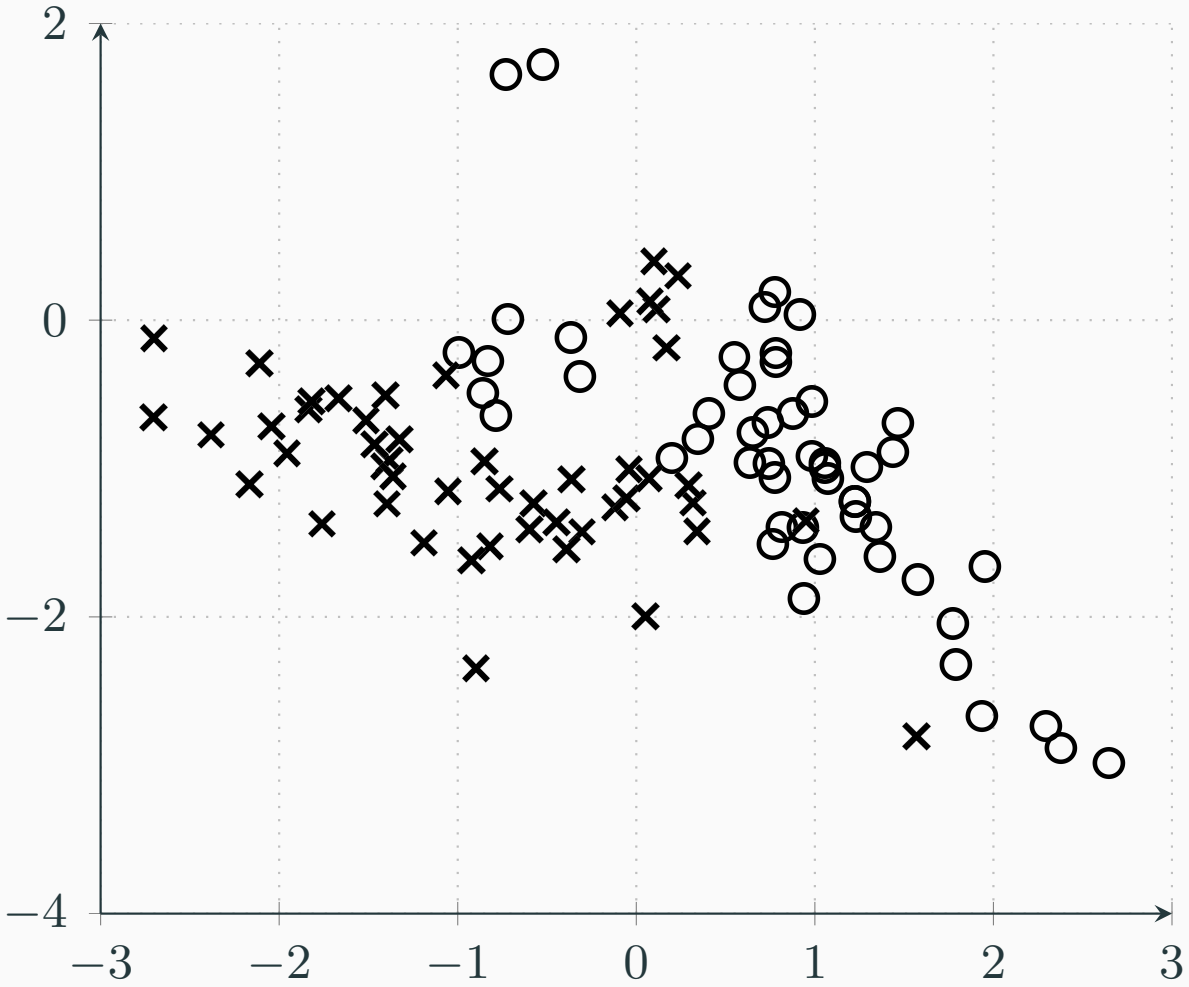
Source: xkcd

Model building & Implementation

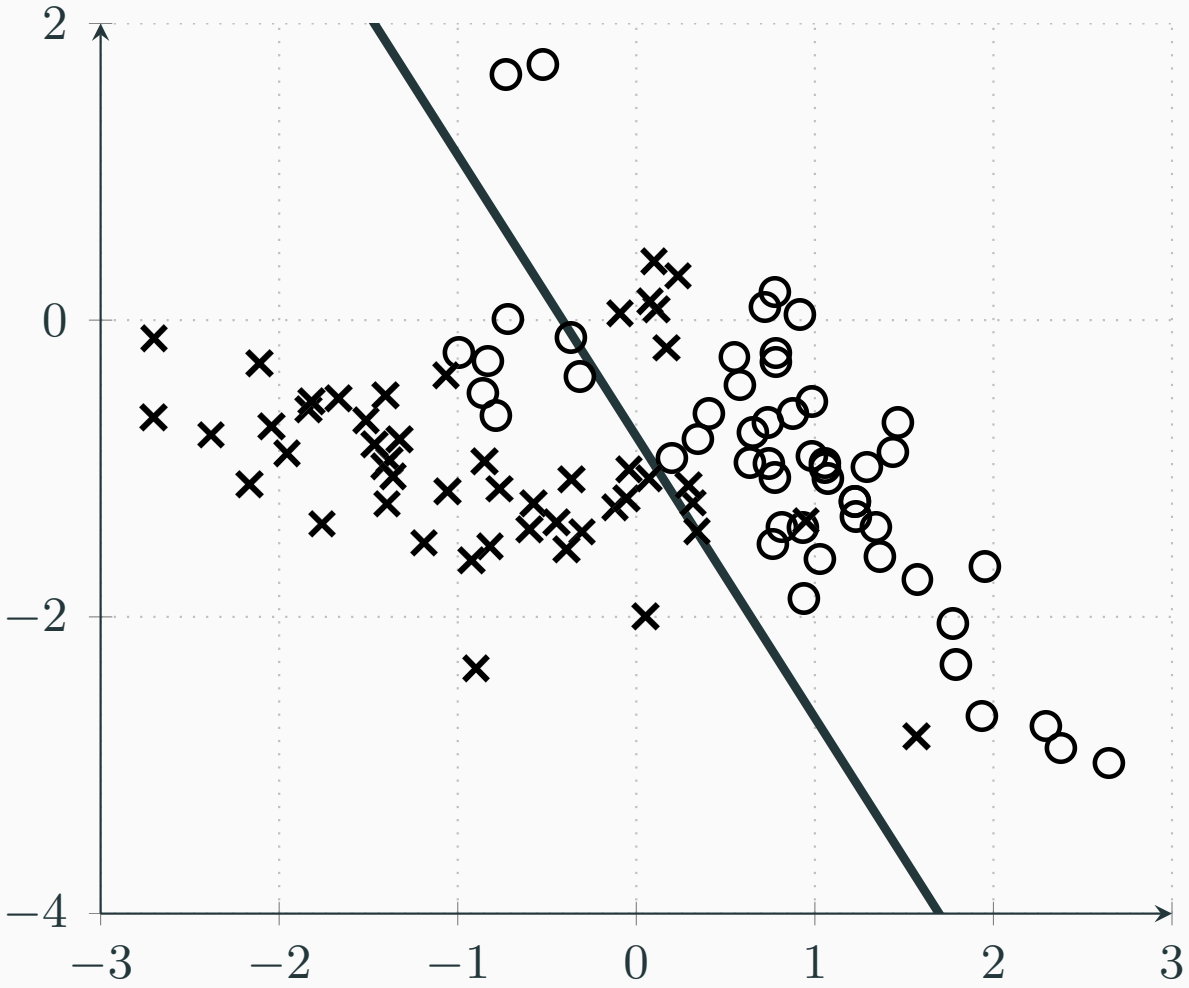


Source: [Suresh and Guttag, 2021]

A simple classification

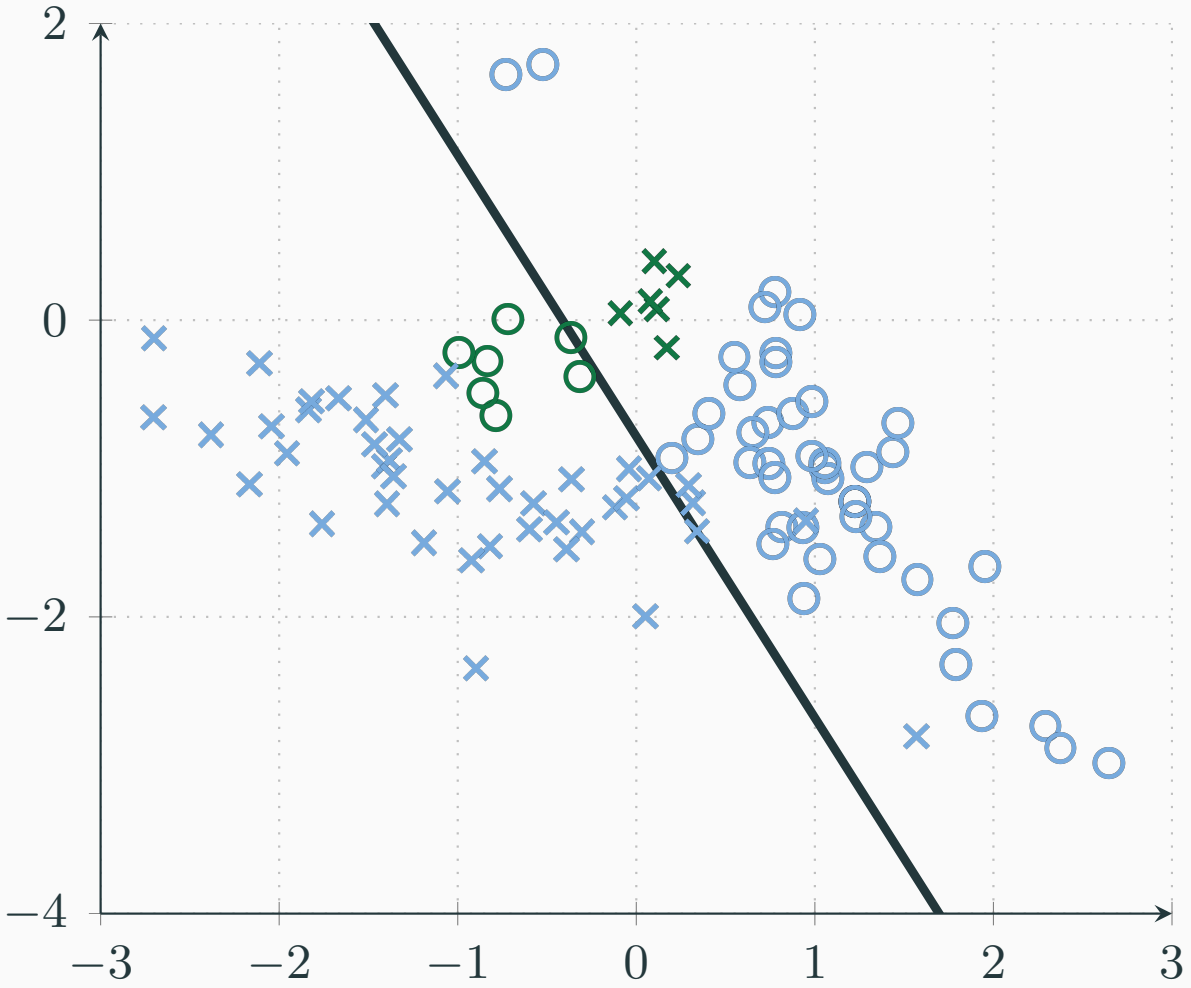


A simple classification



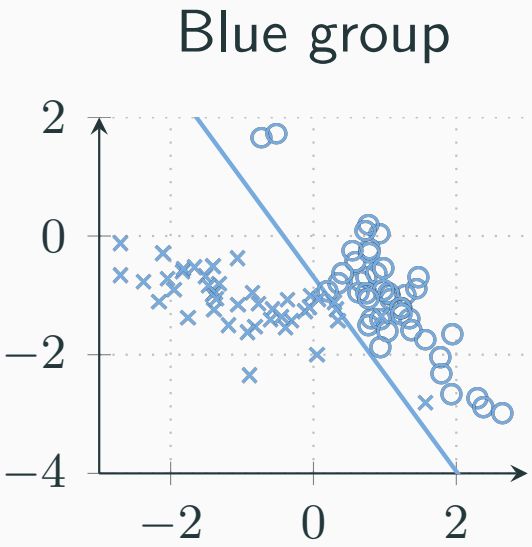
Overall 82%

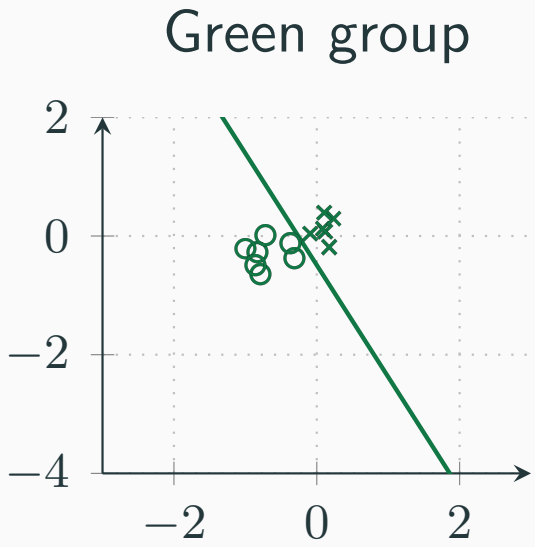
A simple classification



Overall	82%
Green	0%
Blue	94%

Two simple classifications





Overall	97%
Green	100%
Blue	97%

Aggregation bias

Aggregation bias
rises when a one-size-fits-all model is used for data in which there are underlying groups or types of examples that should be considered differently.

Social media analysis
Non-context aware classifiers may harmfully misclassify tweets written in certain dialects.

Source: [Patton et al., 2020]

Learning bias

Learning bias
arises when modeling choices amplify performance disparities across different examples in the data.

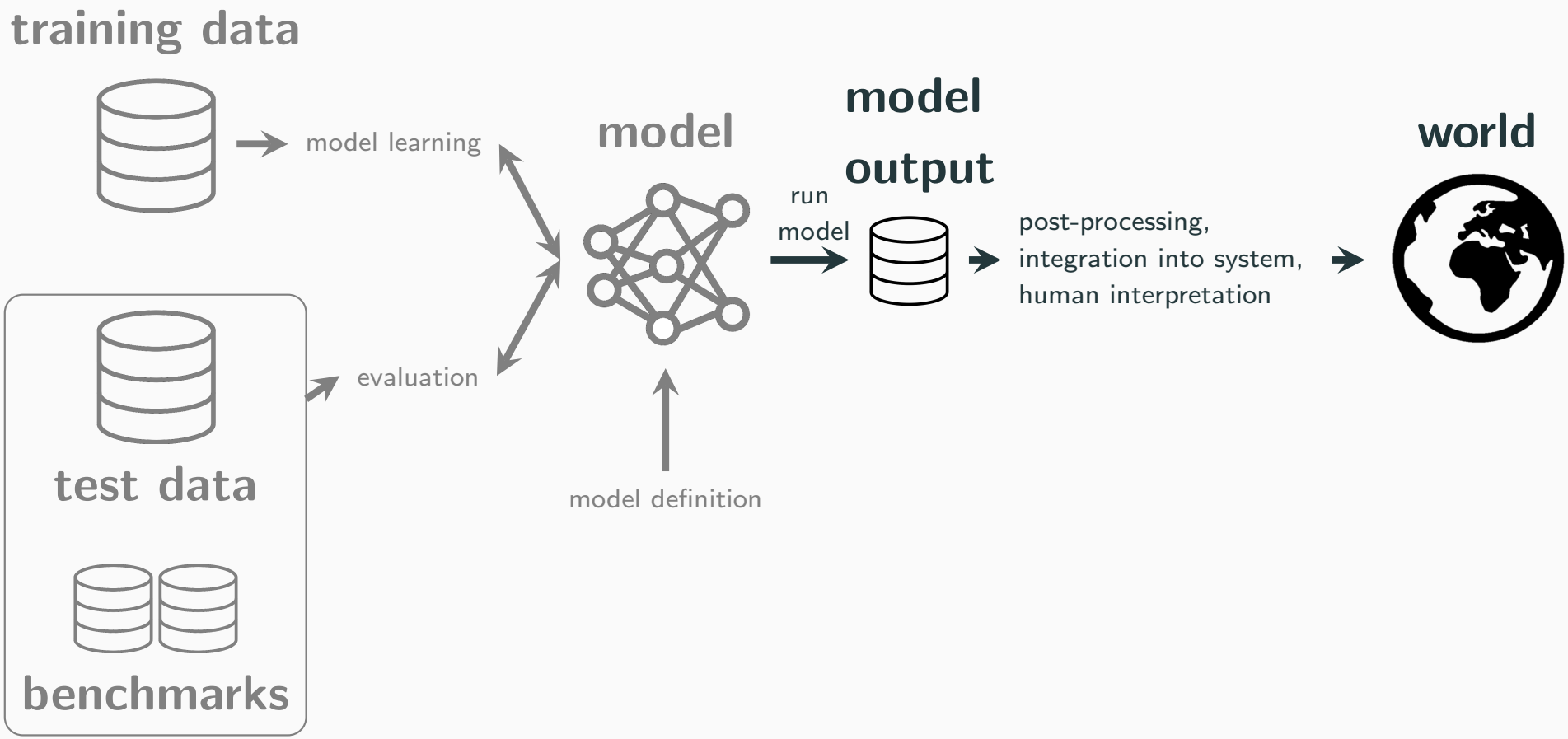
Empirical risk minimization
Every data point has the same influence; majority rules

Differentially private training
Only generic patterns are allowed

Compact models
Generic patterns can be compressed

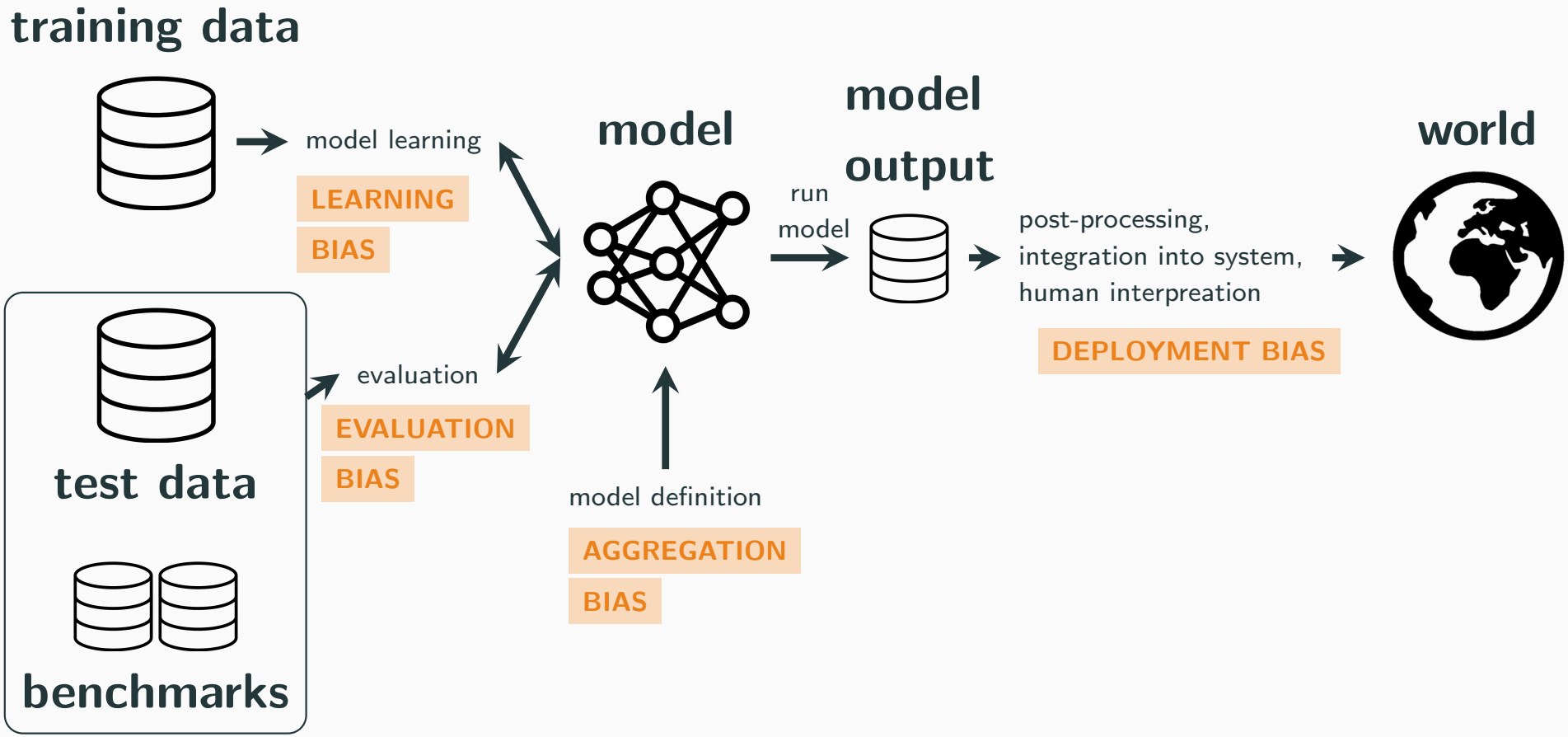
Evaluation bias
occurs when the benchmark data used for a particular task does not represent the user population.

Model building & Implementation



Source: [Suresh and Guttag, 2021]

Model Building and implementation



Takeaways

1. Bias is highly contextual
2. Bias may emerge at any stage of model building
3. Historical social biases are reflected in ML models
4. Formalizing bias requires knowledge about model use context
5. This is not a complete list

Group Fairness

Formal prediction and decision making setting

- Data described by covariates X

Source: M. Hardt [2020]

Formal prediction and decision making setting

- Data described by covariates X
- Ground truth Y (often binary, sometimes called *target variable*)

Source: M. Hardt [2020]

Formal prediction and decision making setting

- Data described by covariates X
- Ground truth Y (often binary, sometimes called *target variable*)
- **Goal:** predict Y from X

Source: M. Hardt [2020]

Formal prediction and decision making setting

- Data described by covariates X
- Ground truth Y (often binary, sometimes called *target variable*)
- **Goal:** predict Y from X
- Use ML to learn a score function $R = r(x)$

Source: M. Hardt [2020]

Formal prediction and decision making setting

- Data described by covariates X
- Ground truth Y (often binary, sometimes called *target variable*)
- **Goal:** predict Y from X
- Use ML to learn a score function $R = r(x)$
- Make binary decisions according to threshold rule

$$D_t(R) = \begin{cases} 1 & R \geq t \\ 0 & R < t \end{cases}$$

Source: M. Hardt [2020]

Decision theory recap

Theory

Outcome Y	Decision D = 0	Decision D = 1
0	True negative	False positive
1	False negative	True positive

Example (COMPAS data)

	Decision D = High Risk	Decision D = Low risk
Recidivated	1874	993
Stayed clean	1154	2129

Decision theory recap

Theory

Outcome Y \ Decision D	0	1
0	True negative	False positive
1	False negative	True positive

Example (COMPAS data)

	Decision D: High Risk	Low risk
Recidivated	1874	993
Stayed clean	1154	2129

$$TPR = \mathbb{P}[D = 1|Y = 1] \qquad = 2129/(2129 + 1154) \qquad = 0.65$$

Decision theory recap

Theory

Outcome Y	Decision D = 0	Decision D = 1
0	True negative	False positive
1	False negative	True positive

Example (COMPAS data)

	Decision D = High Risk	Decision D = Low risk
Recidivated	1874	993
Stayed clean	1154	2129

$$TPR = \mathbb{P}[D = 1|Y = 1] = 2129/(2129 + 1154) = 0.65$$

$$FPR = \mathbb{P}[D = 1|Y = 0] = 993/(1874 + 993) = 0.35$$

Decision theory recap

Theory

Outcome Y	Decision D = 0	Decision D = 1
0	True negative	False positive
1	False negative	True positive

Example (COMPAS data)

	Decision D = High Risk	Decision D = Low risk
Recidivated	1874	993
Stayed clean	1154	2129

$$TPR = \mathbb{P}[D = 1|Y = 1] = 2129/(2129 + 1154) = 0.65$$

$$FPR = \mathbb{P}[D = 1|Y = 0] = 993/(1874 + 993) = 0.35$$

$$TNR = \mathbb{P}[D = 0|Y = 0] = 1874/(1874 + 993) = 0.65$$

Decision theory recap

Theory

Outcome Y	Decision D = 0	Decision D = 1
0	True negative	False positive
1	False negative	True positive

Example (COMPAS data)

	Decision D = High Risk	Decision D = Low risk
Recidivated	1874	993
Stayed clean	1154	2129

$$TPR = \mathbb{P}[D = 1|Y = 1] = 2129/(2129 + 1154) = 0.65$$
$$FPR = \mathbb{P}[D = 1|Y = 0] = 993/(1874 + 993) = 0.35$$
$$TNR = \mathbb{P}[D = 0|Y = 0] = 1874/(1874 + 993) = 0.65$$
$$FNR = \mathbb{P}[D = 0|Y = 1] = 1154/(2129 + 1154) = 0.35$$

Statistical fairness criteria

- Introduce additional (random) variable A encoding membership status in a protected class

Statistical fairness criteria

- Introduce additional (random) variable A encoding membership status in a protected class
- Equalize different statistical quantities involving group membership

What if we just ignore protected class?

- "Fairness through unawareness"

Source: M. Hardt [2020]

What if we just ignore protected class?

- "Fairness through unawareness"
- Other features serve as proxies for protected class and the resulting model is still biased

Source: M. Hardt [2020]

What if we just ignore protected class?

- "Fairness through unawareness"
- Other features serve as proxies for protected class and the resulting model is still biased
- Taking protected class membership into consideration can allow us to examine and mitigate bias in our models

Source: M. Hardt [2020]

What if we just ignore protected class?

- "Fairness through unawareness"
- Other features serve as proxies for protected class and the resulting model is still biased
- Taking protected class membership into consideration can allow us to examine and mitigate bias in our models
- "*'We don't consider that in our data'* is **never** a valid argument."

Source: M. Hardt [2020]

Equalizing acceptance rate (Independence)

Equal positive rate (Demographic parity)
For any two groups a, b, we require

$$\mathbb{P}[D = 1|A = a] = \mathbb{P}[D = 1|A = b]$$

Source: M. Hardt [2020]

Equalizing acceptance rate (Independence)

Equal positive rate (Demographic parity)
For any two groups a, b, we require

$$\mathbb{P}[D = 1|A = a] = \mathbb{P}[D = 1|A = b]$$

"Acceptance rate" equal in all groups

General Definition: Require D to be independent of A

Source: M. Hardt [2020]

Does COMPAS satisfy independence?

Black Defendants

	High Risk	Low risk
Recidivated	1369	532
Stayed clean	805	990

$$\mathbb{P}[D = 1 | A = \text{Black}] = (532 + 990)/3696 = 0.41$$

$$(\text{TPR} = 0.65,\ \text{TNR} = 0.63)$$

White Defendants

	High Risk	Low risk
Recidivated	505	461
Stayed clean	349	1139

$$\mathbb{P}[D = 1 | A = \text{White}] = (461 + 1139)/2454 = 0.65$$

$$(\text{TPR} = 0.71,\ \text{TNR} = 0.59)$$

Source: for data [Angwin et al., 2016]

Does COMPAS satisfy independence?

Black Defendants

	High Risk	Low risk
Recidivated	1369	532
Stayed clean	805	990

White Defendants

	High Risk	Low risk
Recidivated	505	461
Stayed clean	349	1139

$$\mathbb{P}[D = 1 | A = \text{Black}] = (532 + 990)/3696 = 0.41$$

$$(\text{TPR} = 0.65,\ \text{TNR} = 0.63)$$

$$\mathbb{P}[D = 1 | A = \text{White}] = (461 + 1139)/2454 = 0.65$$

$$(\text{TPR} = 0.71,\ \text{TNR} = 0.59)$$

COMPAS does not satisfy independence!

Source: for data [Angwin et al., 2016]

Is independence enough?

Black Defendants

	High Risk	Low risk
Recidivated	~~1369~~ **489**	~~532~~ **1412**
Stayed clean	805	990

White Defendants

	High risk	Low risk
Recidivated	505	461
Stayed clean	349	1139

$$\mathbb{P}[D = 1 | A = \text{Black}] = (1412 + 990)/3696 = 0.65$$

$$(\text{TPR} = 0.41,\ \text{TNR} = 0.38)$$

$$\mathbb{P}[D = 1 | A = \text{White}] = (461 + 1139)/2454 = 0.65$$

$$(\text{TPR} = 0.71,\ \text{TNR} = 0.59)$$

Does independence "solve fairness"?

Potential situation: Equalize the rate, yet, make good/informed decisions in one group, poor/arbitrary decisions in the others.

Potential reason: Malicious model or less/poor data for one group.

Intuition: You shouldn't get to match true positives in one group with false positive in another.

Source: M. Hardt [2020]

Equal error rates (Separation)

Equal error rates
For any two groups a, b, we require

$$\mathbb{P}[D = 1 | Y = 0, A = a] = \mathbb{P}[D = 1 | Y = 0, A = b] \qquad \text{(equal FPR)}$$

$$\mathbb{P}[D = 0 | Y = 1, A = a] = \mathbb{P}[D = 0 | Y = 1, A = b] \qquad \text{(equal FNR)}$$

Source: M. Hardt [2020]

Equal error rates (Separation)

Equal error rates
For any two groups a, b, we require

$$\mathbb{P}[D = 1 | Y = 0, A = a] = \mathbb{P}[D = 1 | Y = 0, A = b] \qquad \text{(equal FPR)}$$
$$\mathbb{P}[D = 0 | Y = 1, A = a] = \mathbb{P}[D = 0 | Y = 1, A = b] \qquad \text{(equal FNR)}$$

General Definition: Require D to be independent of A given Y

Source: M. Hardt [2020]

Equal error rates (Separation)

Equal error rates
For any two groups a, b, we require

$$\mathbb{P}[D = 1|Y = 0, A = a] = \mathbb{P}[D = 1|Y = 0, A = b] \qquad \text{(equal FPR)}$$
$$\mathbb{P}[D = 0|Y = 1, A = a] = \mathbb{P}[D = 0|Y = 1, A = b] \qquad \text{(equal FNR)}$$

General Definition: Require D to be independent of A given Y

For score functions: Require R to be independent of A given Y

Source: M. Hardt [2020]

Equal error rates (Separation)

Equal error rates
For any two groups a, b, we require

$$\mathbb{P}[D = 1|Y = 0, A = a] = \mathbb{P}[D = 1|Y = 0, A = b] \qquad \text{(equal FPR)}$$
$$\mathbb{P}[D = 0|Y = 1, A = a] = \mathbb{P}[D = 0|Y = 1, A = b] \qquad \text{(equal FNR)}$$

General Definition: Require D to be independent of A given Y

For score functions: Require R to be independent of A given Y

Source: M. Hardt [2020]

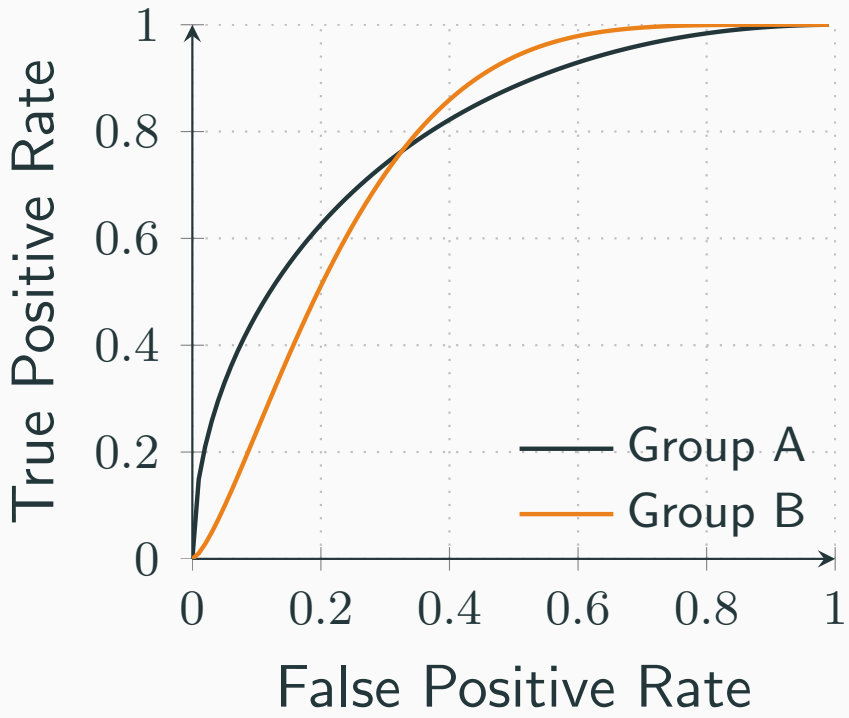
Pros and cons of separation

Pros: Alignment with human intuition. People usually *agree* that violating this measure reflects unfairness

Cons: We have to have the ground truth observations, so it's impossible to know if we are satisfying separation at decision time

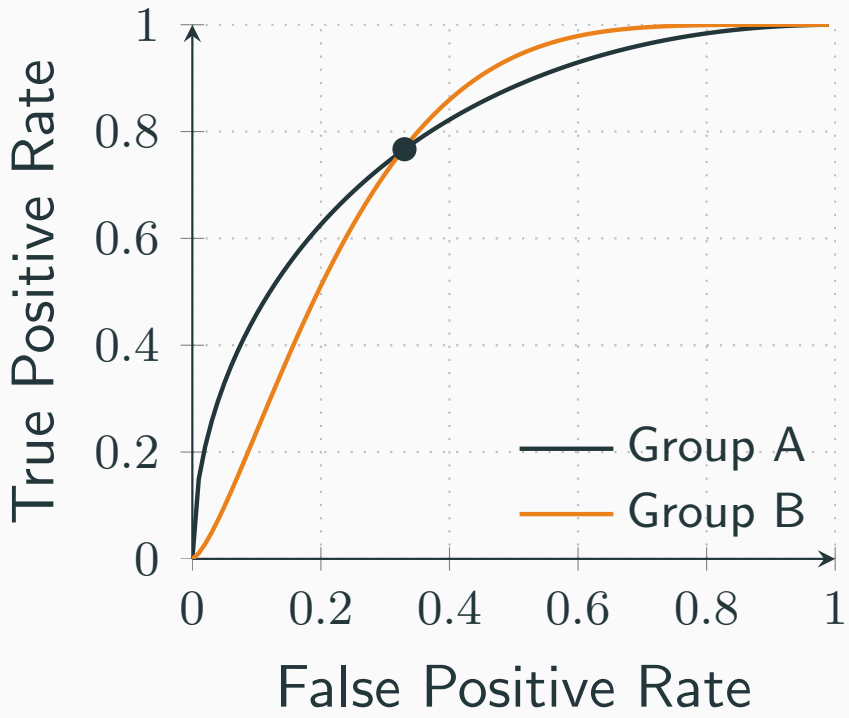
Equal error rates in terms of ROC

- Suppose for each group, D uses individual thresholds for score function R
- If we want to satisfy separation, which classifiers can we accept?



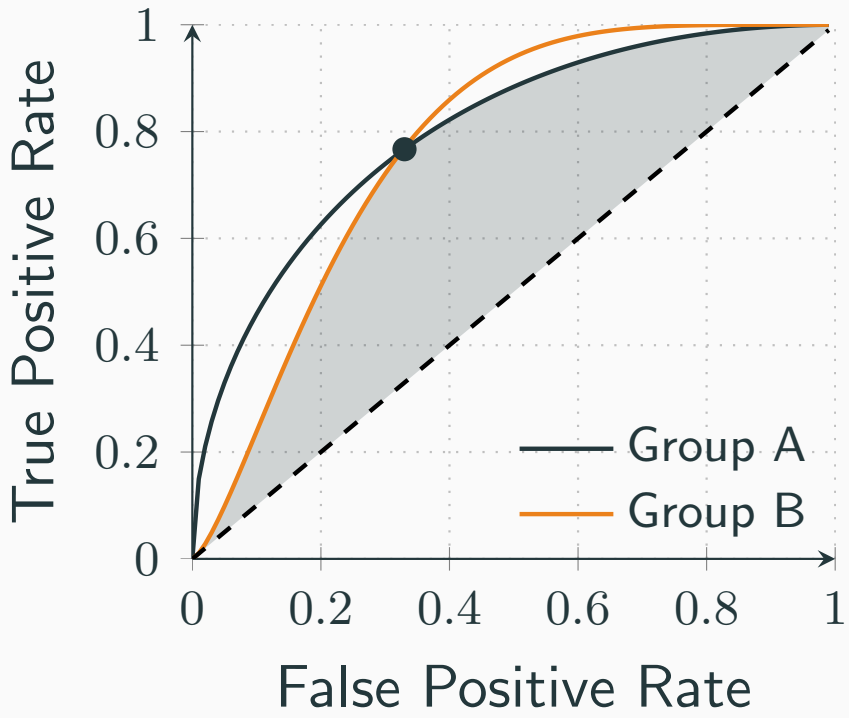
Equal error rates in terms of ROC

- Suppose for each group, D uses individual thresholds for score function R
- If we want to satisfy separation, which classifiers can we accept?
- One threshold setting perfectly satisfies equal error rates



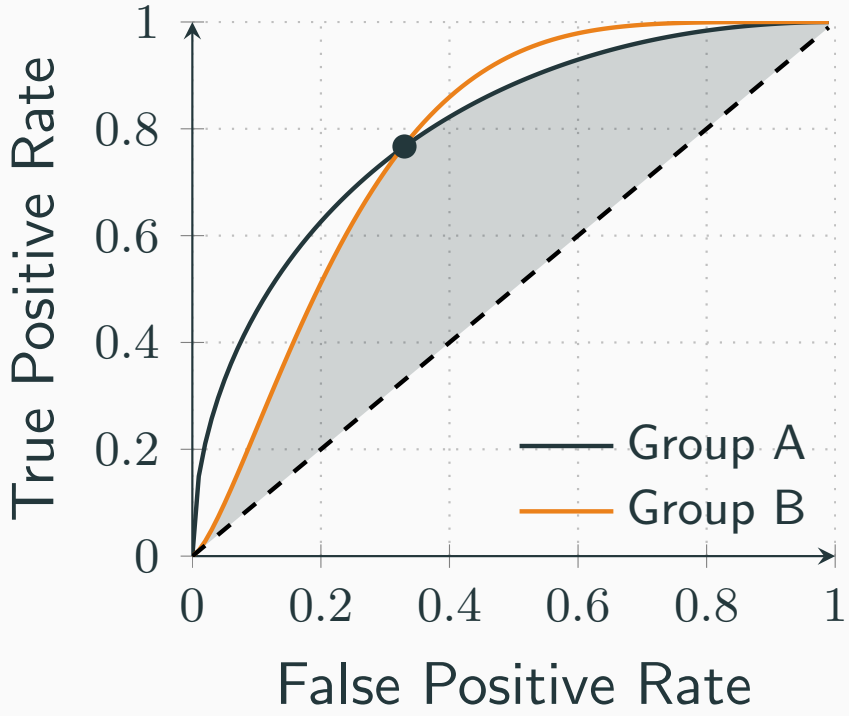
Equal error rates in terms of ROC

- Suppose for each group, D uses individual thresholds for score function R
- If we want to satisfy separation, which classifiers can we accept?
- One threshold setting perfectly satisfies equal error rates
- We can achieve any trade-off below both ROC curves via randomization



Equal error rates in terms of ROC

- Suppose for each group, D uses individual thresholds for score function R
- If we want to satisfy separation, which classifiers can we accept?
- One threshold setting perfectly satisfies equal error rates
- We can achieve any trade-off below both ROC curves via randomization



This might hurt performance of one group more than the other.

Can we satisfy independence and separation?

- We can make unequal errors and satisfy independence

Can we satisfy independence and separation?

- We can make unequal errors and satisfy independence
- Separation solves this, but may allow positive classifications to drop more for one group than another

Can we satisfy independence and separation?

- We can make unequal errors and satisfy independence
- Separation solves this, but may allow positive classifications to drop more for one group than another
- So why not just satisfy both at once?

Independence vs Separation

Theorem
Let X, Y, A be random variables with $Y \in \{0, 1\}$, if there exists a score function $R = r(x)$ such that

$$R \perp A \qquad \text{and} \qquad R \perp A|Y$$

then

$$A \perp Y \qquad \text{or} \qquad R \perp Y$$

In words: One can only satisfy **Independence** and **Separation** if the group membership is independent of the outcome or the score is independent of the outcome (i.e., useless).

Here: $\perp$ stands for independence of two variables.

Source: [Barocas et al., 2019]

Individual Fairness

Group Fairness Overview

So far
We measure unfairness for specific, predefined groups (**group fairness**).

Group Fairness Overview

So far
We measure unfairness for specific, predefined groups (**group fairness**).

Pros

- Often good indicators of unfairness
- Overlap with human intuitions for fairness
- Well aligned with legal definitions

Group Fairness Overview

So far
We measure unfairness for specific, predefined groups (**group fairness**).

Pros

- Often good indicators of unfairness
- Overlap with human intuitions for fairness
- Well aligned with legal definitions

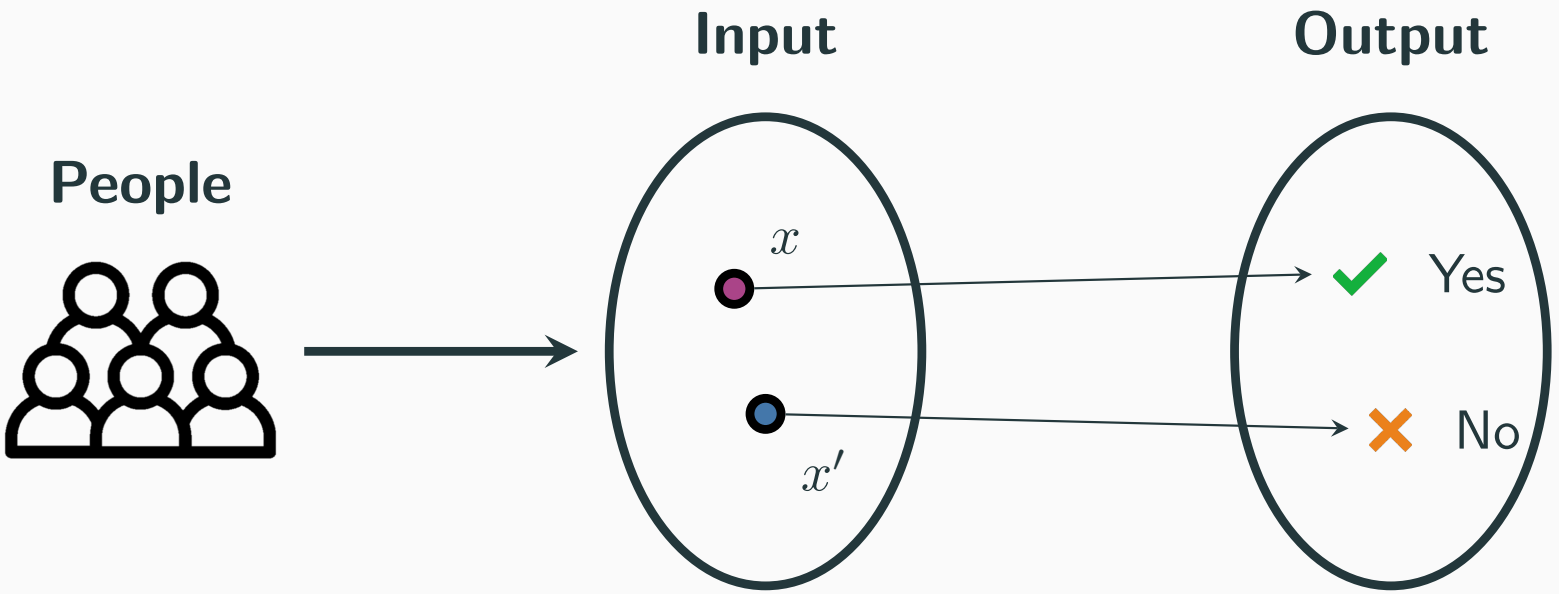
Cons

- All relevant groups may not be specified
- Very small subgroups pose statistical problems
- Fairness to groups doesn't guarantee fairness to individuals

"Similar people" are treated similarly

People who are similar wrt a classification task should be treated similarly.

$$||R(x) - R(x')|| \leq d(x, x')$$

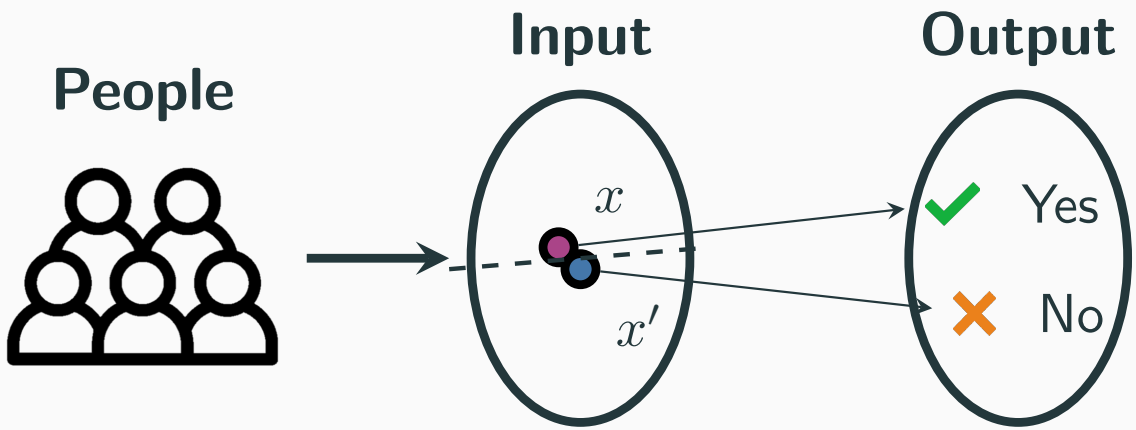


Challenge
Need right notion of dissimilarity $d(x, x')$ for the specific task

Source: Dwork [2022], [Dwork et al., 2012]

“Similar people” are treated similarly

Challenge
Given arbitrarily close input, no model with discrete output can satisfy the Lipschitz constraint. (You have to draw the line somewhere.)



Source: Dwork [2022], [Dwork et al., 2012]

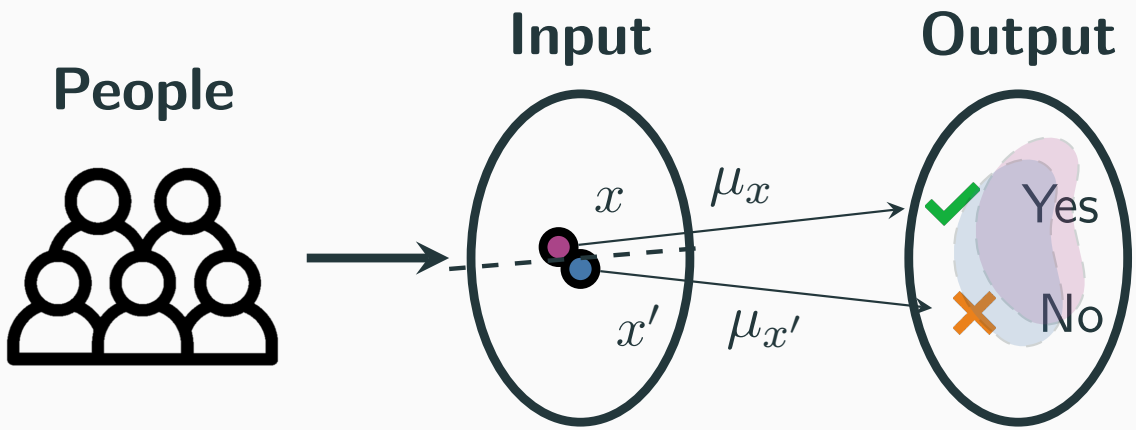
“Similar people” are treated similarly

Challenge

Given arbitrarily close input, no model with discrete output can satisfy the Lipschitz constraint. (You have to draw the line somewhere.)

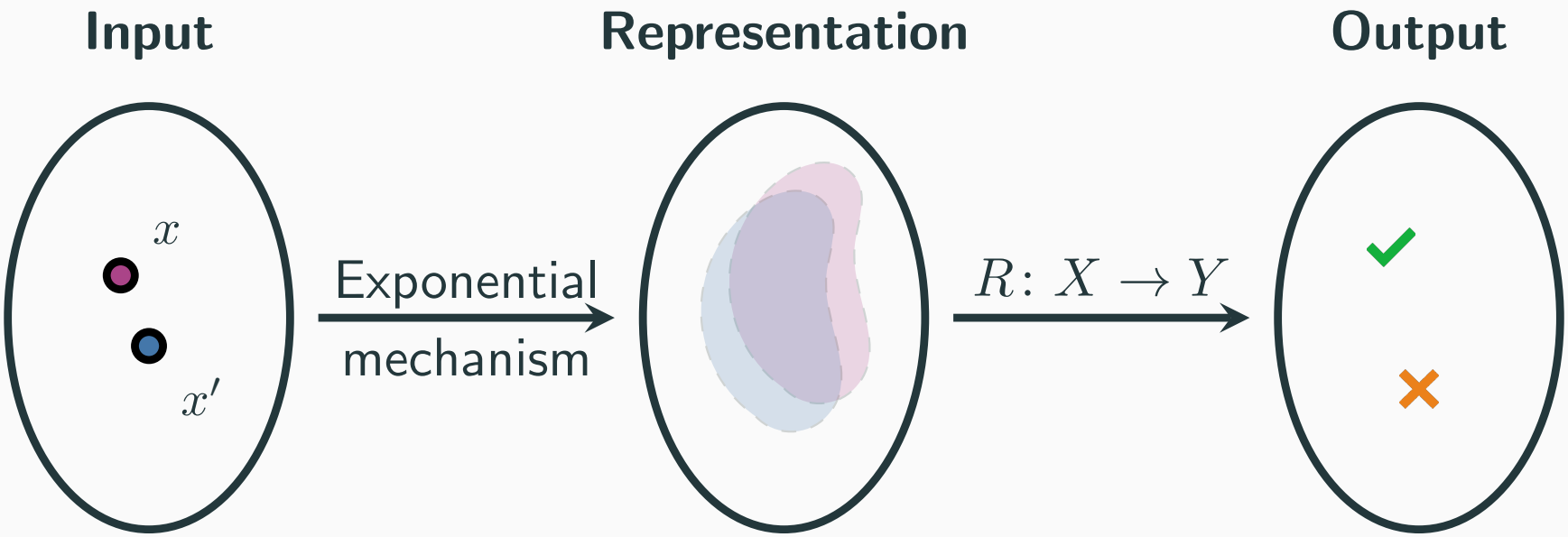
Solution

Instead of mapping to discrete output, map to continuous probability distributions over the output. (**Similar people** have similar **probability distributions** on outcomes.)



Source: Dwork [2022], [Dwork et al., 2012]

Individually fair classifiers



Takeaways

1. Individual fairness is the notion that similar people should be treated similarly.
2. Individual fairness can help overcome shortcomings of the statistical fairness notions discussed so far.
3. Given a similarity metric we can use tools from privacy research and linear optimization to create individually fair classifiers.

Main takeaways

1. Biased ML models have the potential to cause real harm
2. There are many possible sources of bias at every stage of model development
3. This course: Technical fairness definitions and interventions
4. Non-technical interventions can and should be considered

Things to think about

- How do we define protected groups and membership to those groups?
- What happens to fairness when an individual belongs to multiple protected groups? (intersectional fairness)
- How do we measure fairness when the task is generation, not classification?
- Can we achieve all the fairness guarantees we introduced at once?
- How does fairness interact with robustness and privacy?

References i


Angwin, J., larson, J., Mattu, S., and Krichner, L. (2016).
Machine bias: There's software used across the country to predict future criminals and it's biased against blacks.
ProPublica.


Barocas, S., Crawford, K., Chapiro, A., and Wallach, H. (2017).
The problem with bias: Allocative versus representational harms in machine learning.
In SIGCIS Conference.


Barocas, S., Hardt, M., and Narayanan, A. (2019).
Fairness and Machine Learning.
fairmlbook.org.
http://www.fairmlbook.org.

References ii

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016).

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.

In Advances in Neural Information Processing Systems, volume 29. Curran Associates, Inc.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012).

Fairness through awareness.

In Proceedings of the 3rd innovations in theoretical computer science conference, pages 214–226.

References iii

Lum, K. and Isaac, W. (2016).

To predict and serve?

Significance, 13(5):14–19.

Patton, D. U., Frey, W. R., McGregor, K. A., Lee, F.-T., McKeown, K., and Moss, E. (2020).

Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing.

In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pages 337–342.

References iv

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D. (2017).

No classification without representation: Assessing geodiversity issues in open data sets for the developing world.

arXiv preprint arXiv:1711.08536.

Suresh, H. and Guttag, J. (2021).

A framework for understanding sources of harm throughout the machine learning life cycle.

In Equity and Access in Algorithms, Mechanisms, and Optimization, pages 1–9.

References v

Sweeney, L. (2013).

Discrimination in online ad delivery.

Communications of the ACM, 56(5):44–54.

Icon credits

- Check by ainul muttaqin from NounProject.com
- Circle by Leinad Lehmko
- Cross by Three Six Five from Noun Project
- Data by shashank singh from NounProject.com
- World by Henry from NounProject.com
- People by Adrien Coquet from NounProject.com
- Question by Seochan from NounProject.com
- Robot by Oksana Latysheva from NounProject.com