

The image features a central graphic of a scale of justice, where the pans and the central pillar are constructed from a complex network of interconnected nodes and thin lines. The scale is positioned on a light gray surface against a dark gray background. The text is overlaid on the central part of the scale.

Justice, Fairness, Bias

The Big Three

Justice, Fairness and Bias

- You will remember that Kant was emphasizing the importance of **human dignity**.
- Individuals expect to be treated fairly; the violation of human dignity leads to **discrimination**.
- Discrimination is the **unjust treatment** of people based on the groups or classes they belong to. Discrimination may stem from biases.
- We often talk about algorithmic fairness, since algorithms may amplify existing **economic** and **societal bias**.

Discrimination and Biases

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

Tolga Bolukbasi¹, Kai-Wei Chang², James Zou², Venkatesh Saligrama^{1,2}, Adam Kalai²

¹Boston University, 8 Saint Mary's Street, Boston, MA

²Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

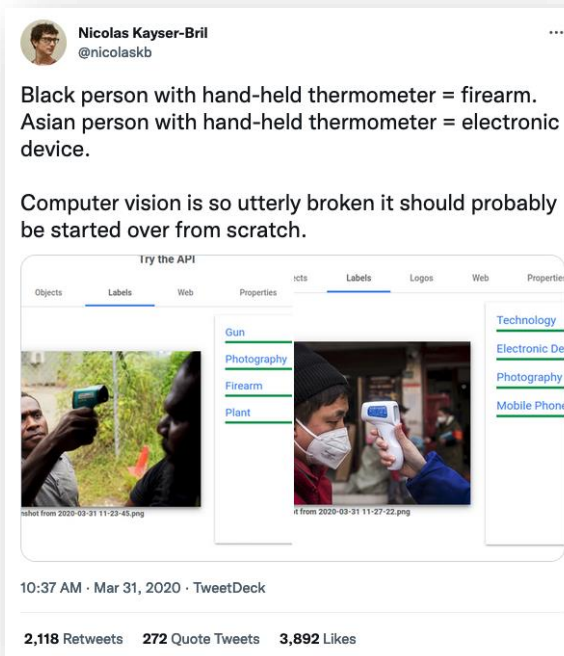
The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent.

This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words *receptionist* and *female*, while maintaining desired associations such as between the words *queen* and *female*. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

Nicolas Kayser-Bril
@nicolaskb

Black person with hand-held thermometer = firearm.
Asian person with hand-held thermometer = electronic device.

Computer vision is so utterly broken it should probably be started over from scratch.



Try the API

Objects Labels Web Properties

Labels

Gun
Photography
Firearm
Plant

Technology
Electronic De
Photography
Mobile Phone

10:37 AM · Mar 31, 2020 · TweetDeck

2,118 Retweets 272 Quote Tweets 3,892 Likes

Discrimination from the perspective of harms

- Allocative Harms

- Harm is defined in terms of available **resources** (e.g., women being paid less, risks for people of color created by harmful algorithms)
- Easier to spot, hard to mitigate

We are trying to optimize utilities to allocate goods in a society full of inequalities

- Representational Harms

- Harm is defined in terms of **the representation of groups and individuals**.
- They affect the narrative (e.g., word embeddings such as secretary-woman, manager-man; search for CEO resulting in a bunch of white-male men)
- More difficult to spot, hard to mitigate

Hidden Bias: Gendering AI Technologies

- Unesco Report – I'd Blush if I could
- The report offers **guidance** for education and steps to address the issues to push for equality.

"The more that culture teaches people to **equate women with assistants**, the more real women will be seen as assistants — and penalized for not being assistant-like."

<https://www.nytimes.com/2019/05/22/world/siri-alexa-ai-gender-bias.html>
<https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1>



Bias in AI Systems

- Awareness
 - Understanding bias in AI systems
- Detection
 - Measuring bias in AI systems
- Mitigation
 - Fixing bias in AI systems



2. Check bias metrics

Dataset: German credit scoring

Mitigation: none

Protected Attribute: Sex

Privileged Group: **Male**, Unprivileged Group: **Female**

Accuracy with no mitigation applied is 75%

With default thresholds, bias against unprivileged group detected in 0 out of 5 metrics

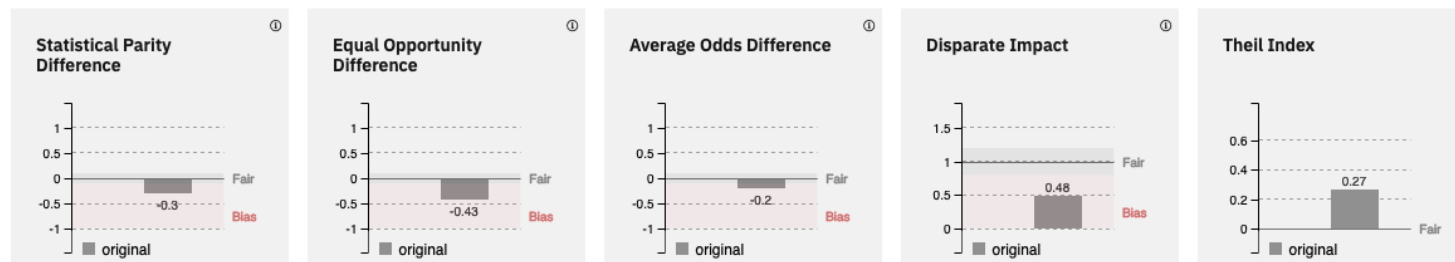


Protected Attribute: Age

Privileged Group: **Old**, Unprivileged Group: **Young**

Accuracy with no mitigation applied is 75%

With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics



2. Check bias metrics

Dataset: German credit scoring

Mitigation: none

Protected Attribute: Sex

Privileged Group: **Male**, Unprivileged Group: **Female**

Accuracy with no mitigation applied is 75%

With default thresholds, bias against unprivileged group detected in 0 out of 5 metrics

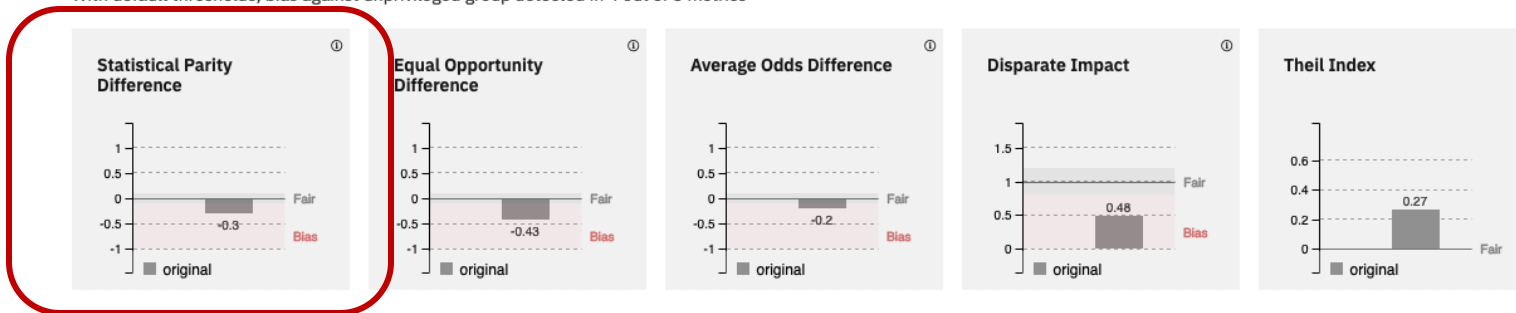


Protected Attribute: Age

Privileged Group: **Old**, Unprivileged Group: **Young**

Accuracy with no mitigation applied is 75%

With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics



A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

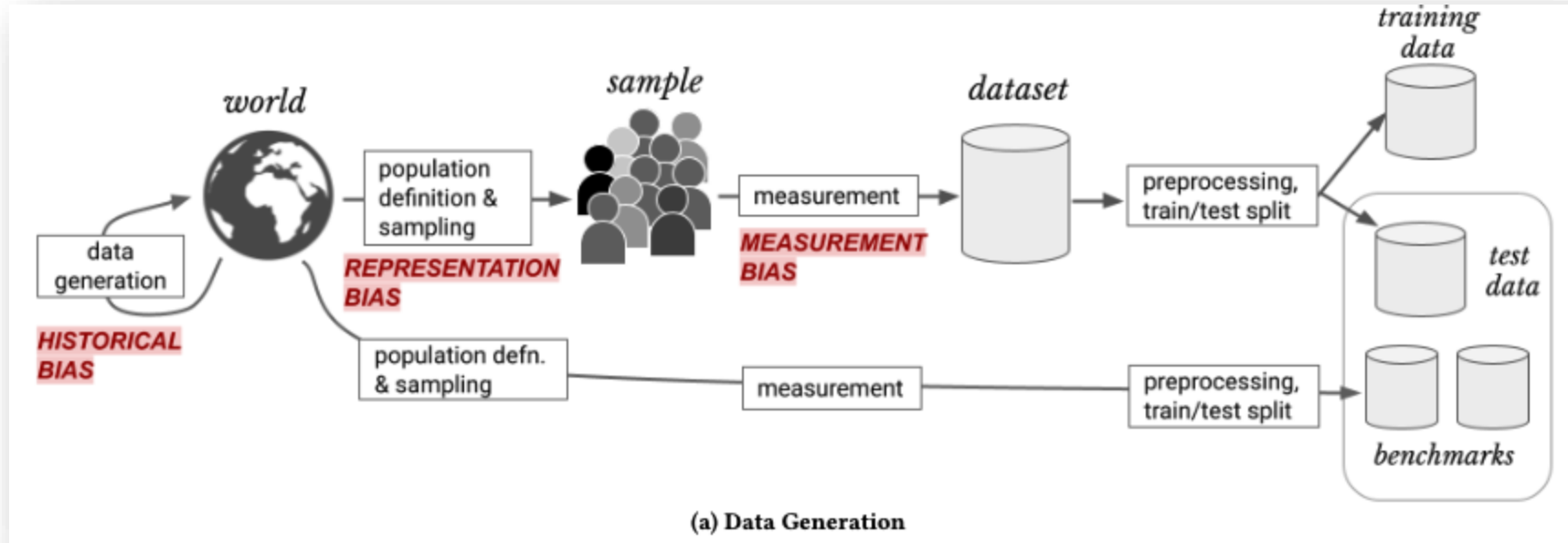
Harini Suresh
John Guttag
hsuresh@mit.edu
guttag@mit.edu

ABSTRACT

As machine learning (ML) increasingly affects people and society, awareness of its potential unwanted consequences has also grown. To anticipate, prevent, and mitigate undesirable downstream consequences, it is critical that we understand when and how harm might be introduced throughout the ML life cycle. In this paper, we provide a framework that identifies seven distinct potential sources of downstream harm in machine learning, spanning data collection, development, and deployment. In doing so, we aim to facilitate more productive and precise communication around these issues, as well as more direct, application-grounded ways to mitigate them.

necessarily because the statement “data is biased” is *false*, but because it treats data as a static artifact divorced from the process that produced it. This process is long and complex, grounded in historical context and driven by human choices and norms. Understanding the implications of each stage in the data generation process can reveal more direct and meaningful ways to prevent or address harmful downstream consequences that overly broad terms like “biased data” can mask.

Moreover, it is important to acknowledge that not all problems should be blamed on the data. The ML pipeline involves a series of choices and practices, from model definition to user interfaces used upon deployment. Each stage involves decisions that can lead



Historical Bias

- It arises if the world **as it is** or **was** leads to a model that produces harmful outcomes.
- Garg et al. show that:
 - Word embeddings reflect real-world biases about women and ethnic minorities;
 - Specific adjectives and occupations become more closely associated with certain populations over time.

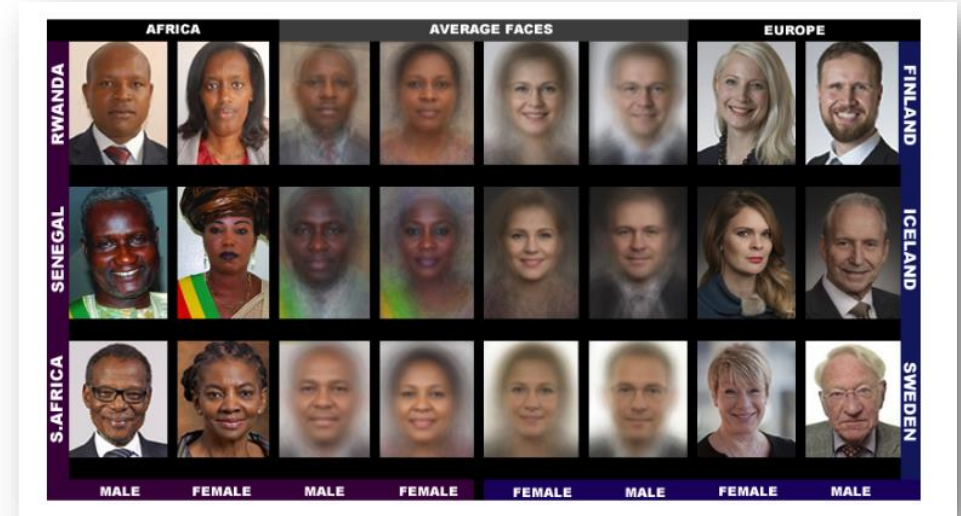
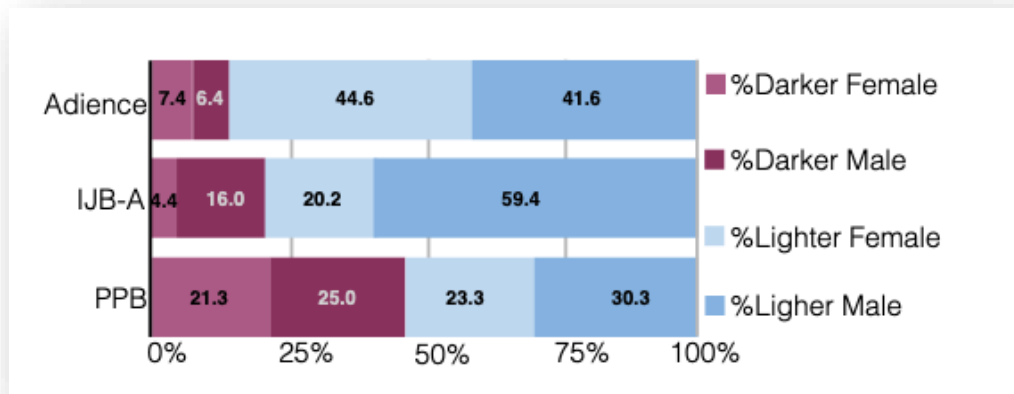


Representation Bias

- Target population **does not reflect** the use population
 - Model is trained on population X and applied to population Y
 - Model is trained on the same population in different time frames
- Target population contains **under-represented groups**
 - For example, some age groups may not be represented well in the data
- Sampling method is limited (**sampling bias**)
 - Target population is set to X, but the data available is only a small subset of X

Gender Shades

- Buolamwini and Gebru analyze two benchmarks to report gender and skin type distribution.



Measurement Bias

- It is difficult to **choose correct proxies** to measure constructs, which can be quite complex. What features measure construct X the best?
- The method of measurement can be different. (e.g., COVID tests)
- The accuracy of measurement can be different.
 - For example, systematically higher rates of misdiagnosis/underdiagnosis are reported for certain groups.

Why Aren't We Talking About Schizophrenia in Black People?

✓ Medically Reviewed

Schizophrenia in Black People: Racial Disparities Explained

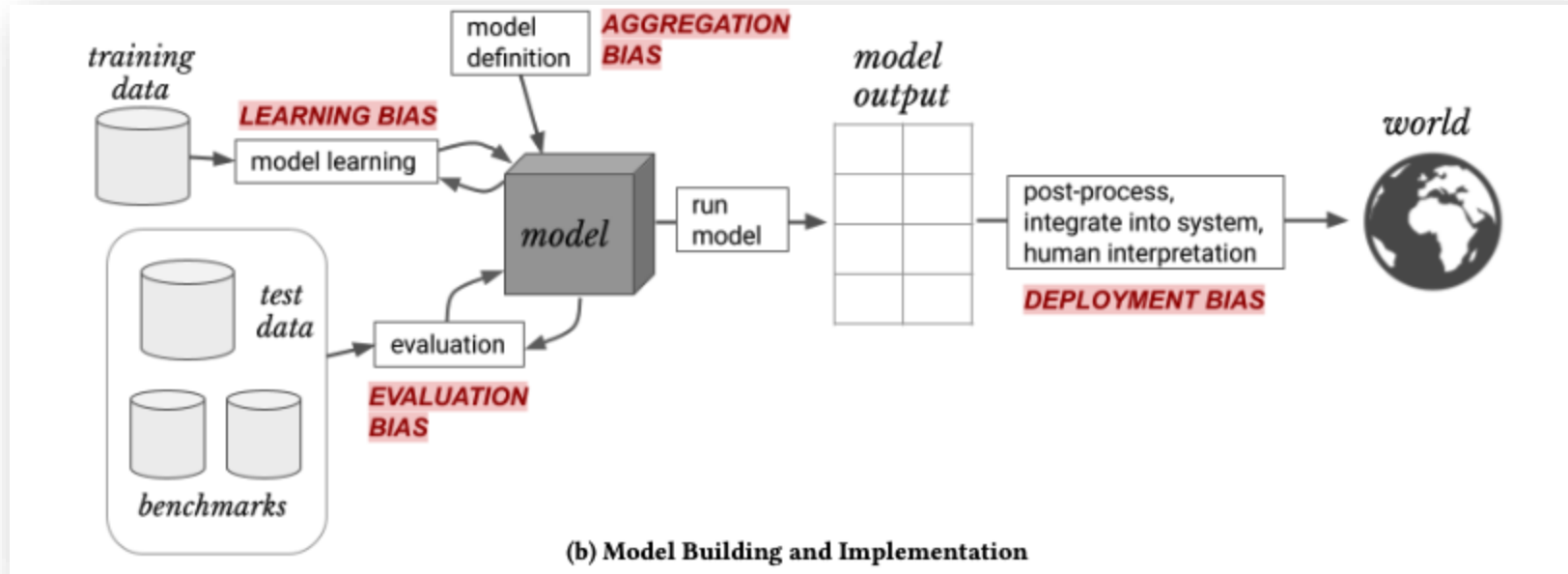
Top schizophrenia experts and researchers shed light and offer solutions to the issue of racial disparities in schizophrenia diagnoses.

Updated Nov 2, 2022

By: Diane Herbst

Medical Reviewer: Stacia Alexander, PhD, LPC-S



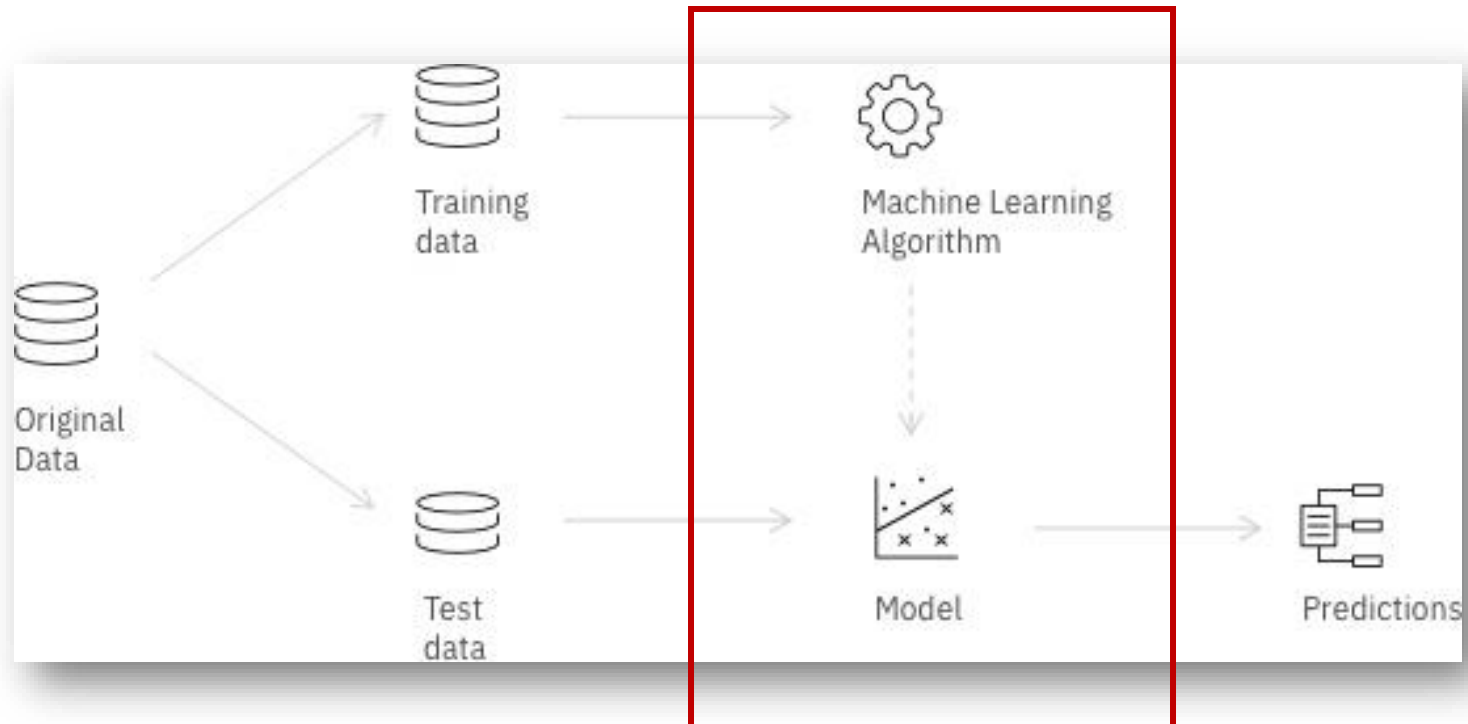


Aggregation Bias

- Aggregate level findings **may not** also be observed at an individual level.
- Exploratory data analysis is important to find out any **group-based trends**.
- One should be very careful to conduct any analysis based on aggregated data which does not capture **true relationships** (e.g., correlations between variables).

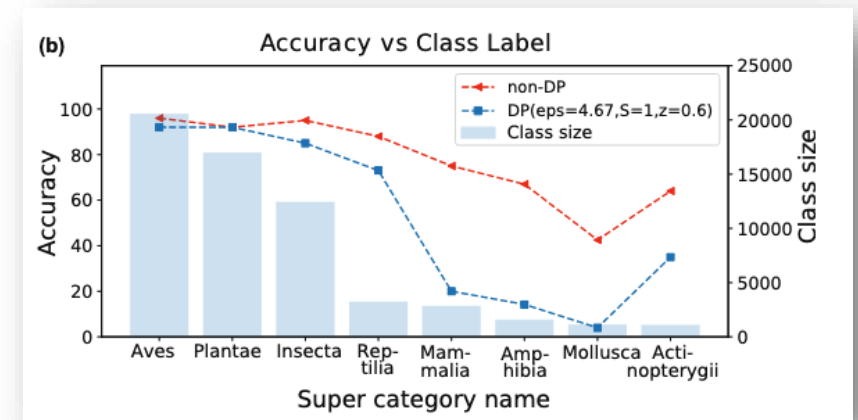
Learning Bias

- **Learning bias** happens when modeling choices amplify performance disparities.



Disparate Impact on Model Accuracy

- **Differential privacy (DP)** comes with a cost, which is a reduction in the model's accuracy.
- Bagdasaryan *et al.* show that accuracy of models, trained with DP stochastic gradient descent, drops much more for the underrepresented classes and subgroups.
- This gap is **bigger** in the DP model than in the non-DP model.
- The results are reported from the sentiment analysis of text and image classification.



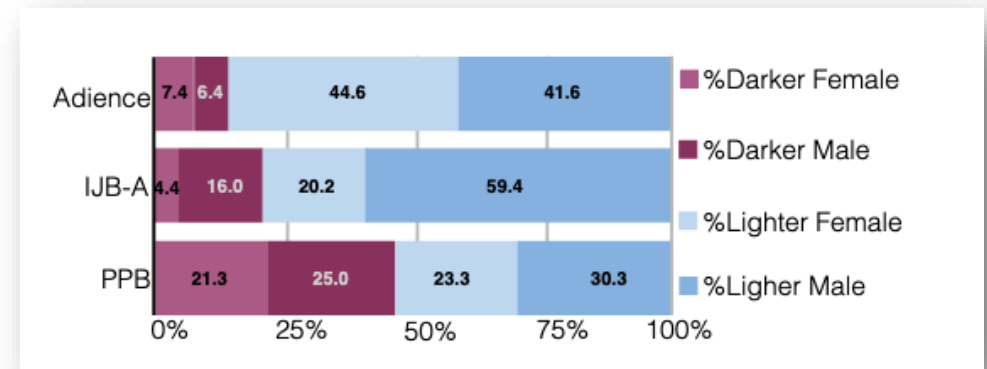
Evaluation Bias

What is Evaluation Bias?

- **Evaluation bias** occurs when the benchmark datasets (e.g., ImageNet) do not represent the use population.
- The choice of **metrics** can also result in evaluation bias (e.g., aggregate results, reporting one type of metric)

Gender Shades

- Buolamwini and Gebru analyze two benchmarks to report gender and skin type distribution.



Gender Shades (Evaluation/Learning Bias example)

- They use their dataset (PPB) to evaluate three commercial gender classification systems (Microsoft, IBM, Face++):

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	100
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	20.8	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	100	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	16.3	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	99.3	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	34.5	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	98.9	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	23.4	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	99.7
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	34.7	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	99.6	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	25.2	17.7	5.20	0.4

Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability and transparency*, pp. 77-91. PMLR, 2018.

Deployment Bias

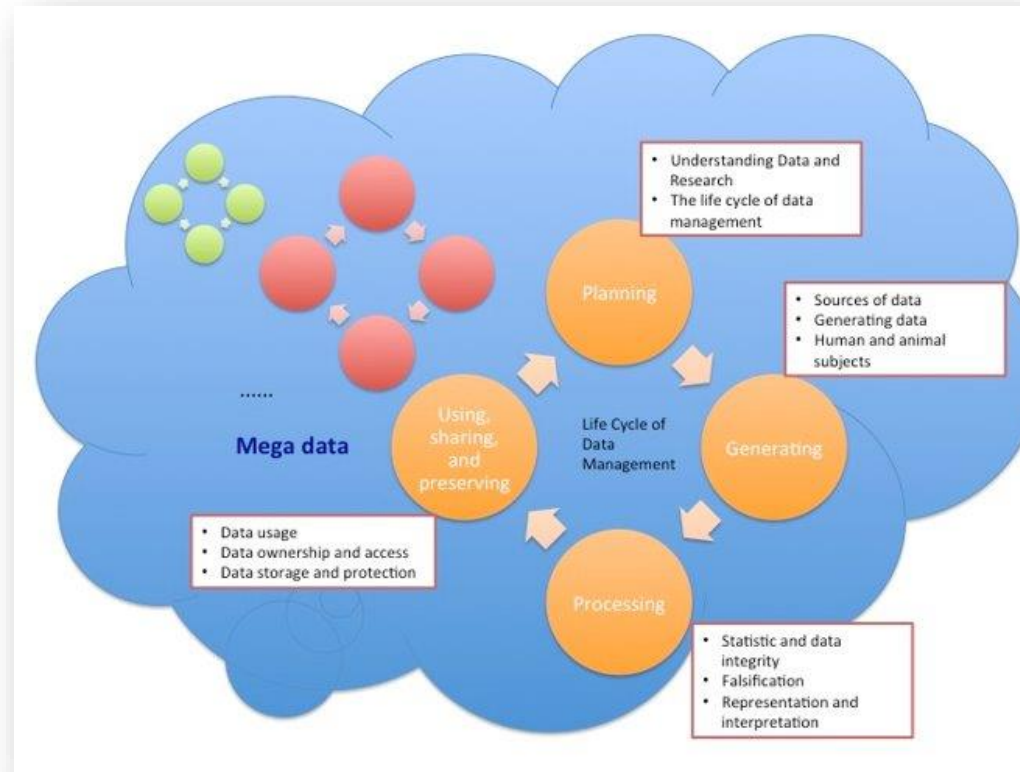
- **Deployment bias** arises when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used.
- One good example is **framing trap** in the sociotechnical context.

entirely. Failure to account for how judges respond to scores creates a problem for risk assessment tools that come with fairness guarantees. Such a tool might present a guarantee of the form "if you use these thresholds to determine low, medium and high risks, then your system will not have a racial disparity in treatment of more than X%". But if a judge only adopts the tool's recommendation some of the time, the claimed guarantee might be incorrect, because a "shifted" threshold caused by judicial modification comes with a much poorer effective guarantee of fairness. Moreover, if the judge demonstrates a bias in the types of cases on which she alters the recommendation, there might be no validity to the guarantee at all. In other words, a frame that does not incorporate a model of the judge's decisions cannot provide the end-to-end guarantees that this frame requires.

Other: Label Bias

- Labels reflect **interpretations** about data.
- Labelling can be done by humans:
 - The use of crowdsourcing platforms in order to get a labelled dataset.
 - Domain experts annotating data (e.g., medical domain)
- Weak-ML techniques may be used to annotate data automatically.
 - Based on defined functions, heuristics etc.

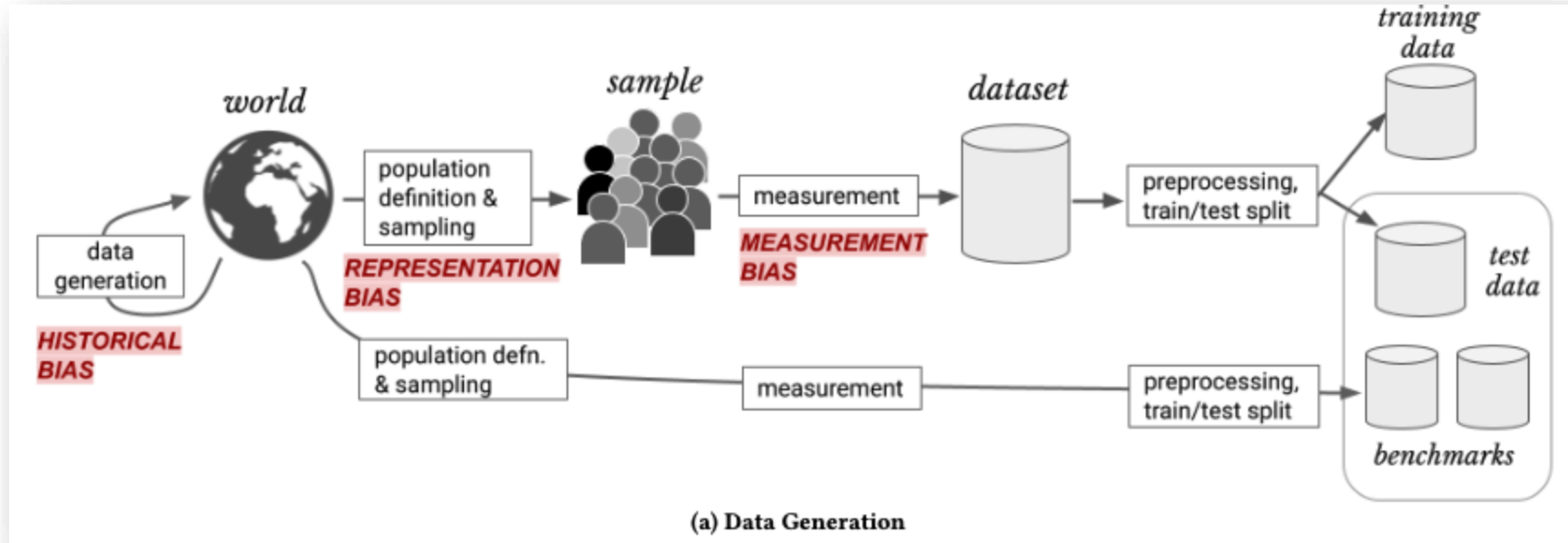
Other: Human Bias / Team Bias



Removing Bias

- Anti-classification
 - Removal of protected attributes (gender, ethnicity...) and their proxies from data
- Resampling
 - Goal: Having a **good distribution of examples across groups**
 - If the dataset is large, a small portion of the dataset could be used to achieve this.
 - If the dataset is small, removing examples from data is expensive. One approach is to increase the size of samples from underrepresented groups in the training set.

Summary



Summary

