Fairness and Bias
in Machine Learning

Surya Dutta

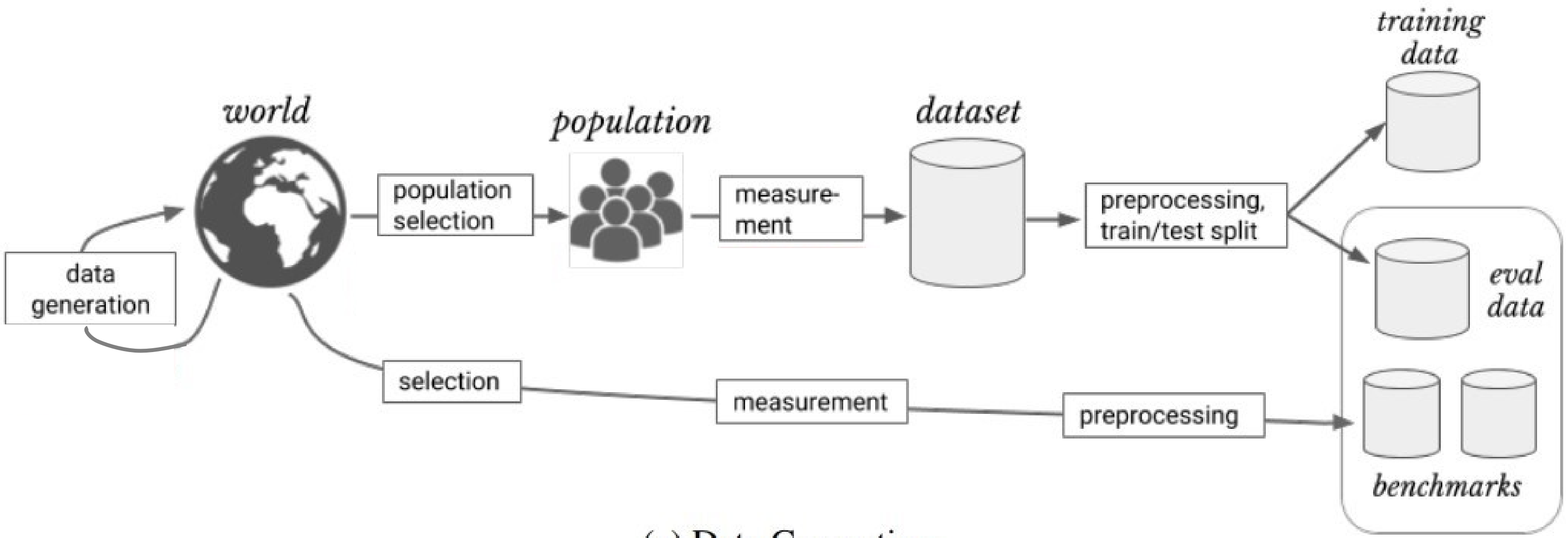
Today, we'll discuss:

- What does bias in machine learning look like?
- How does algorithmic bias get introduced & amplify?

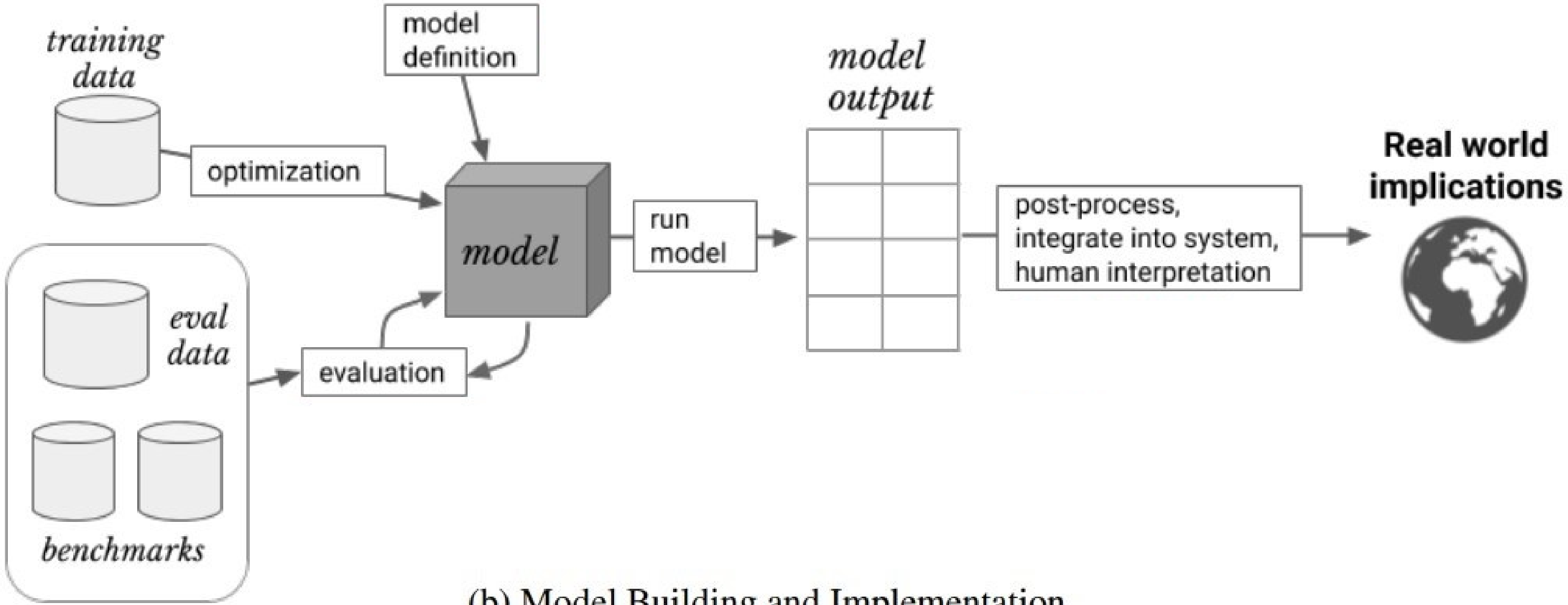
Today, we'll discuss:

- What does bias in machine learning look like?
- How does algorithmic bias get introduced & amplify?

- How can we quantify bias and fairness?
- How can we mitigate algorithmic bias?

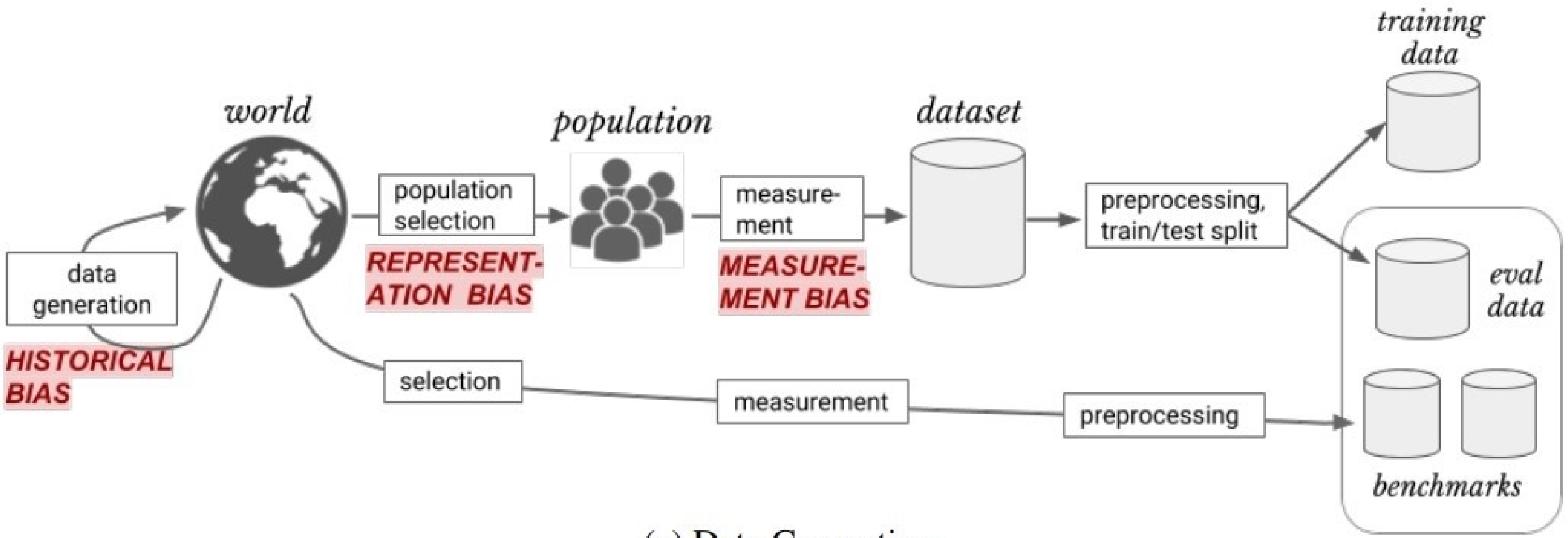


(a) Data Generation

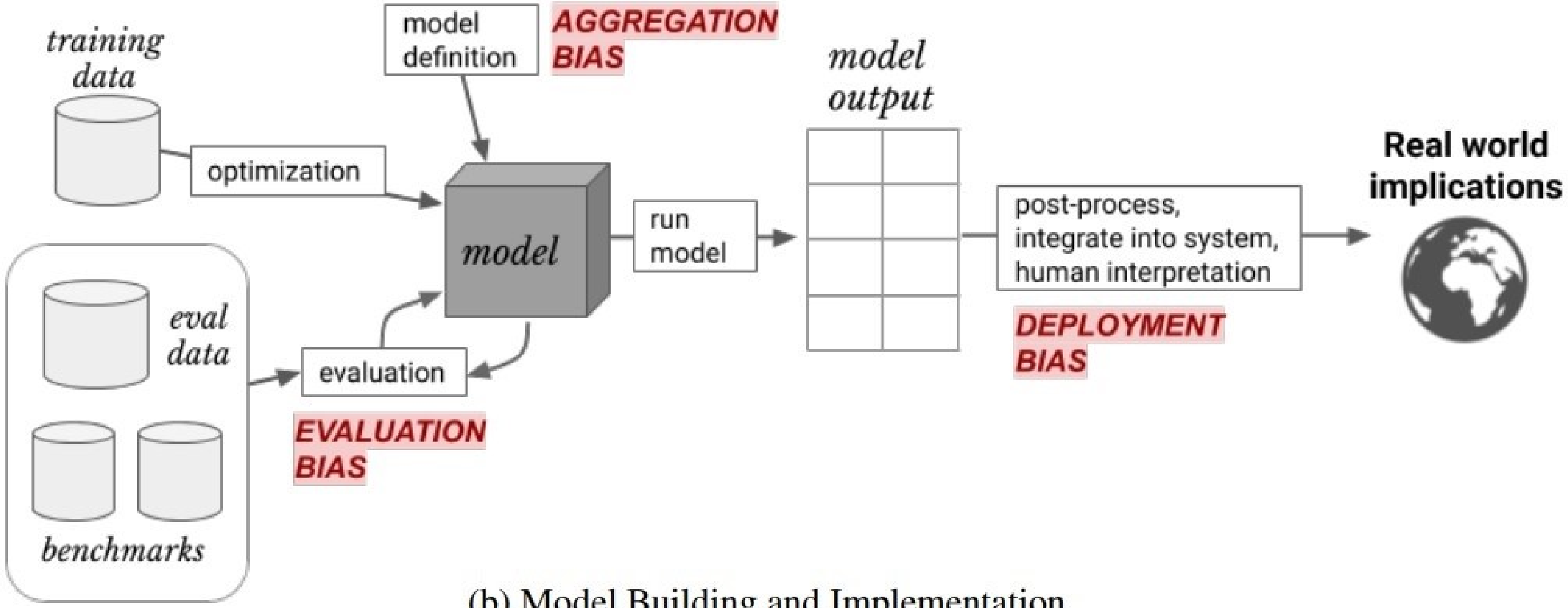


(b) Model Building and Implementation

[Source: Suresh, Guttag 2020]

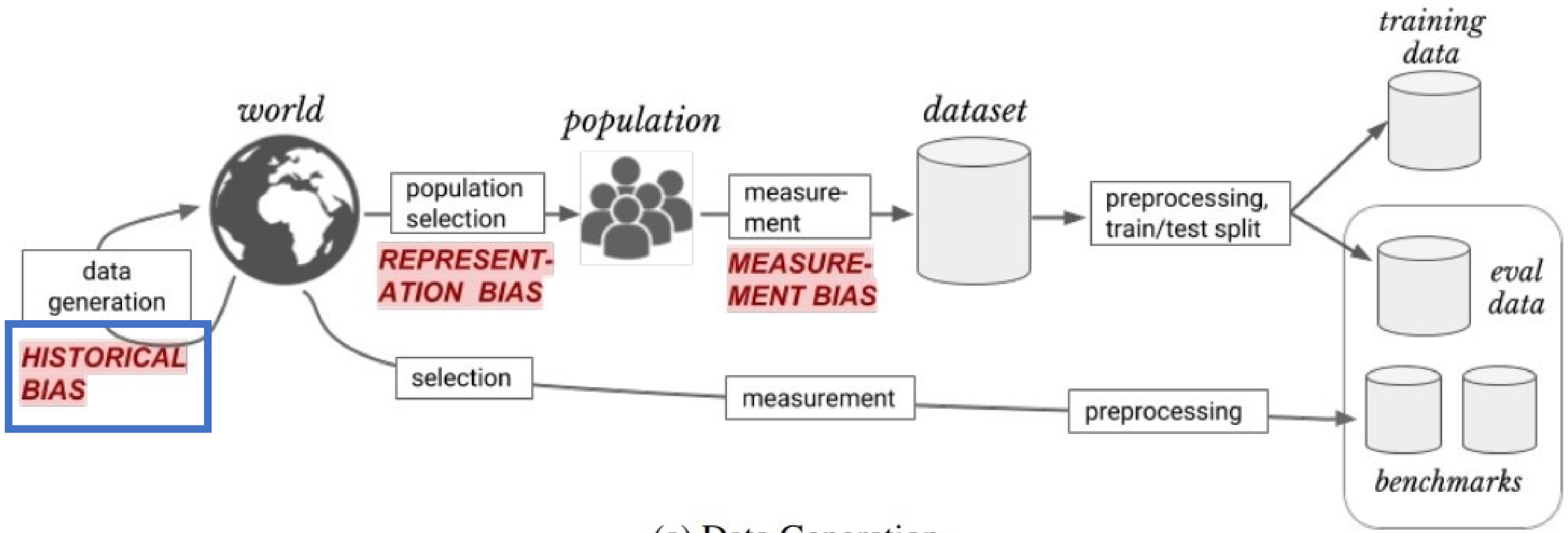


(a) Data Generation



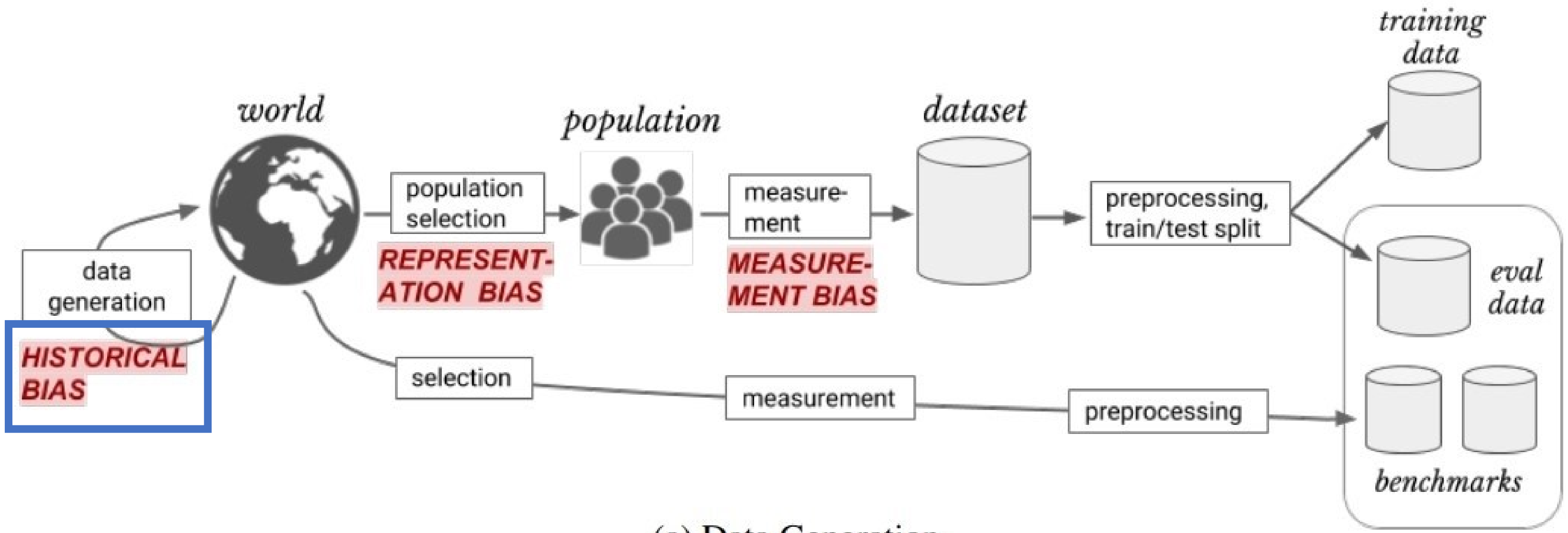
(b) Model Building and Implementation

[Source: Suresh, Guttag 2020]



(a) Data Generation

Historical Bias



(a) Data Generation

LAPD ditches predictive policing program accused of racial bias

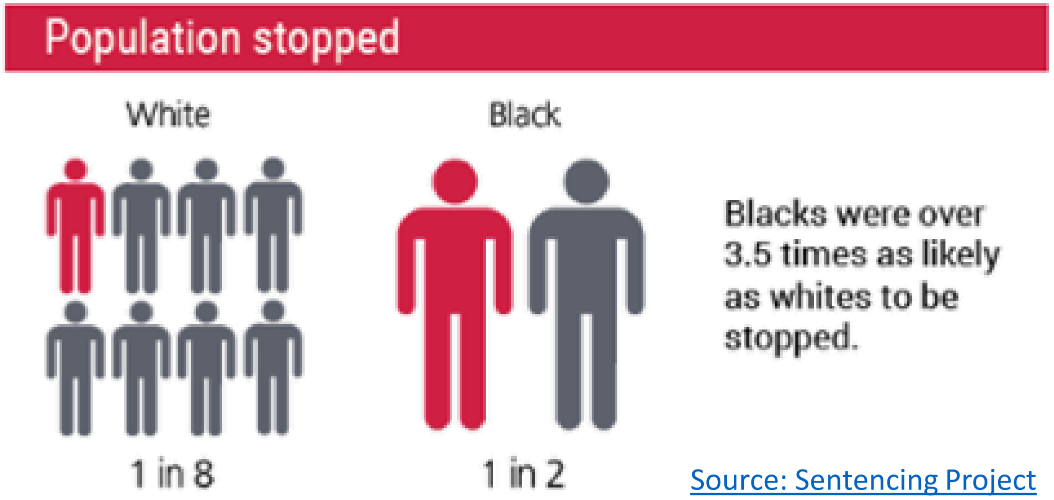
[Source: The Next Web]

Chicago's predictive policing tool just failed a major test

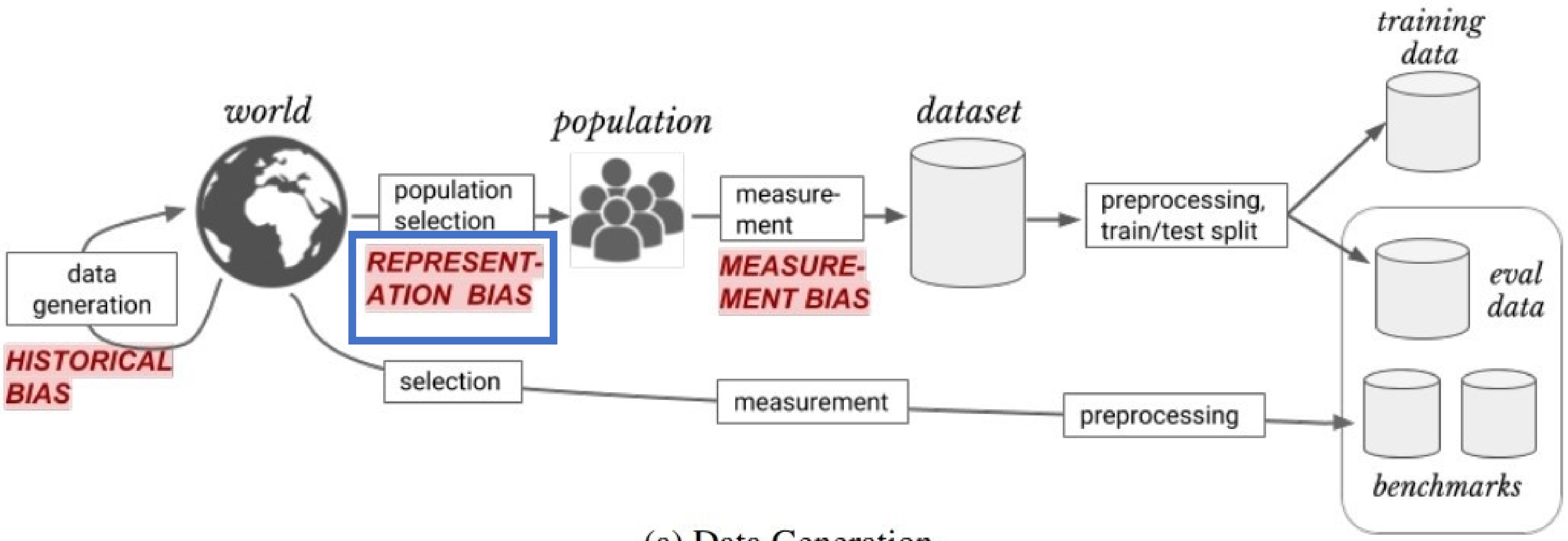
A RAND report shows that the 'Strategic Subject List' doesn't reduce homicides

[Source: The Verge]

Ferguson, Missouri 2013

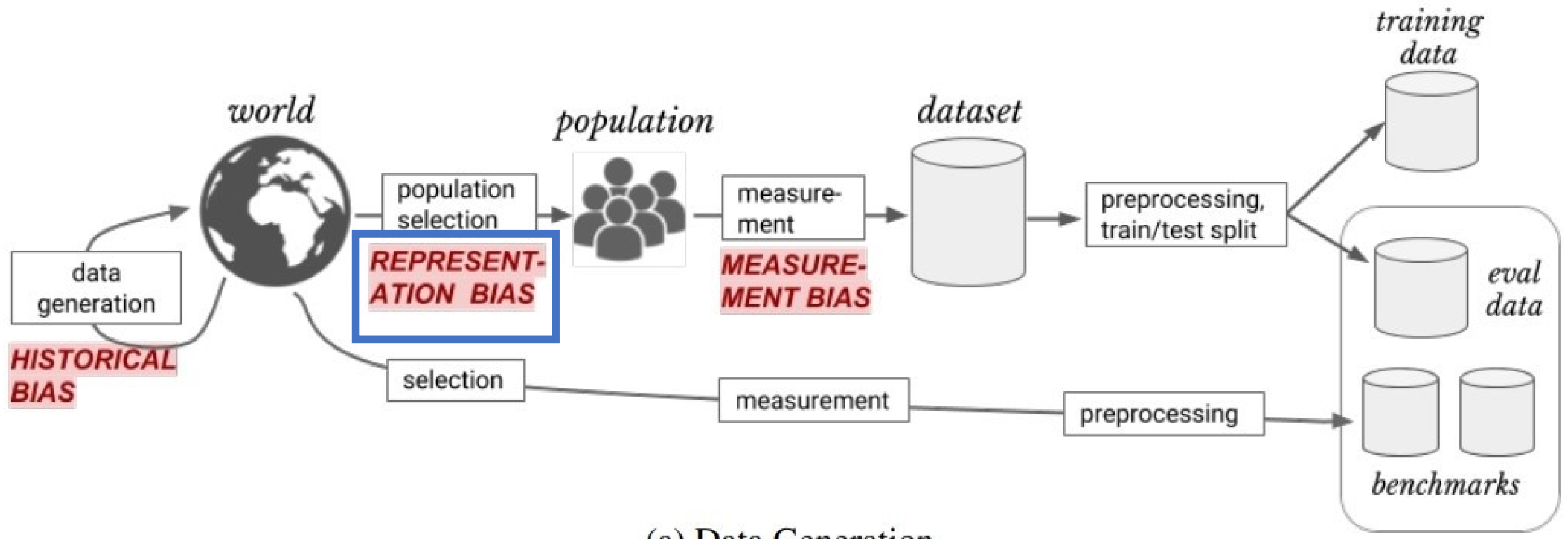


[Source: Sentencing Project]



(a) Data Generation

Representation Bias



(a) Data Generation

Crash Test Dummies Based on Men Pose Risks for Female Drivers

Source: Invisible Women

71% more likely to be **moderately injured**

47% more likely to be **seriously injured**

17% more likely to **die**

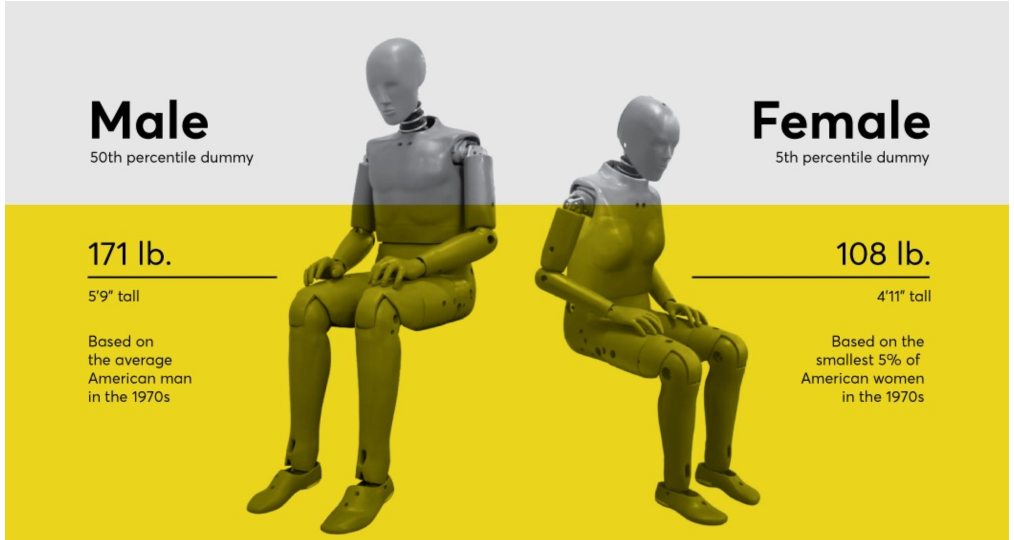
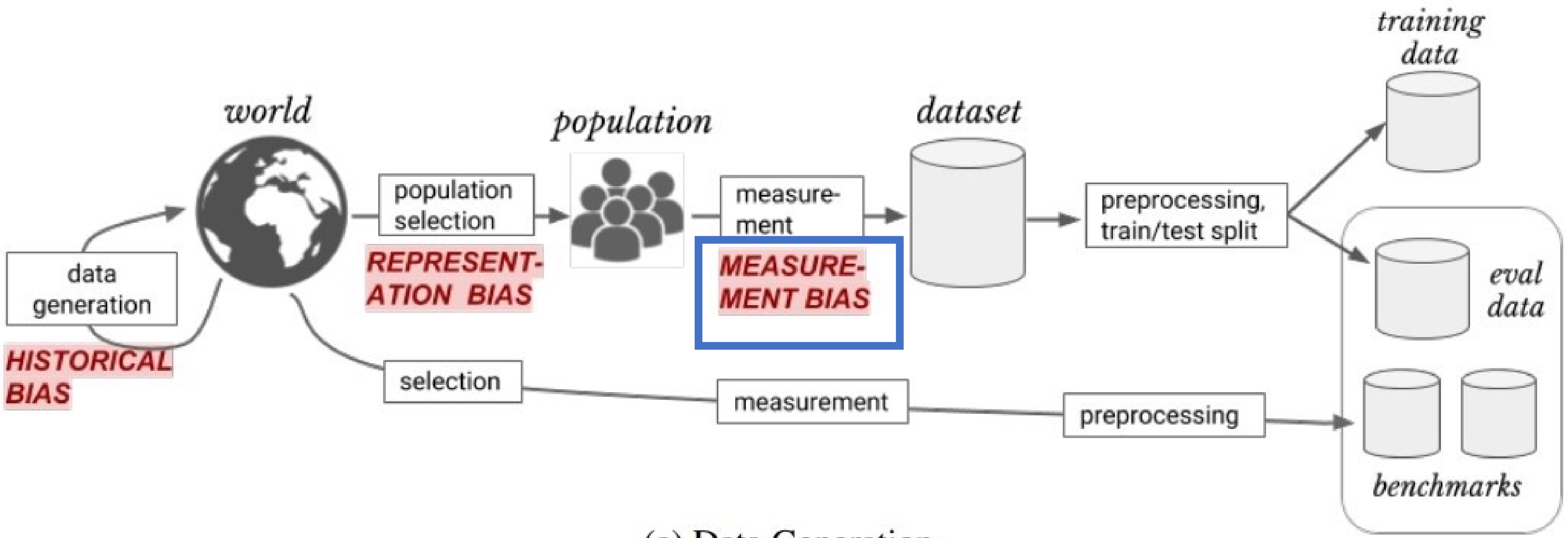
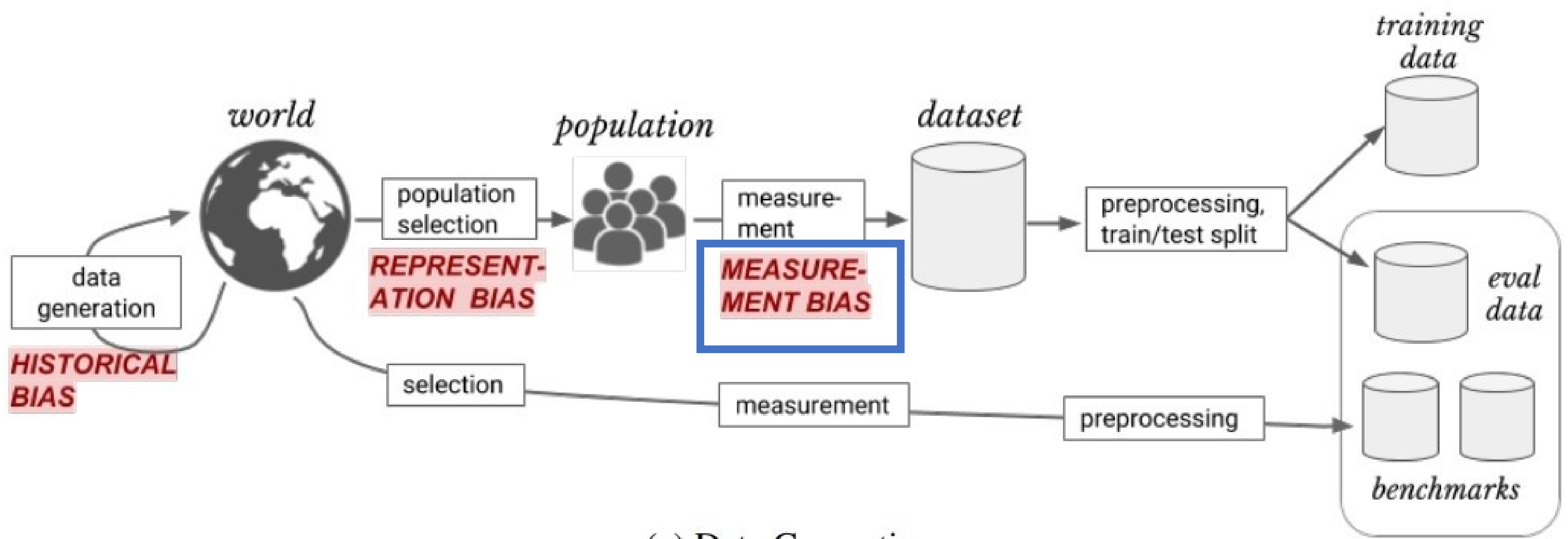


Image Source: Consumer Reports



(a) Data Generation

Measurement Bias



(a) Data Generation

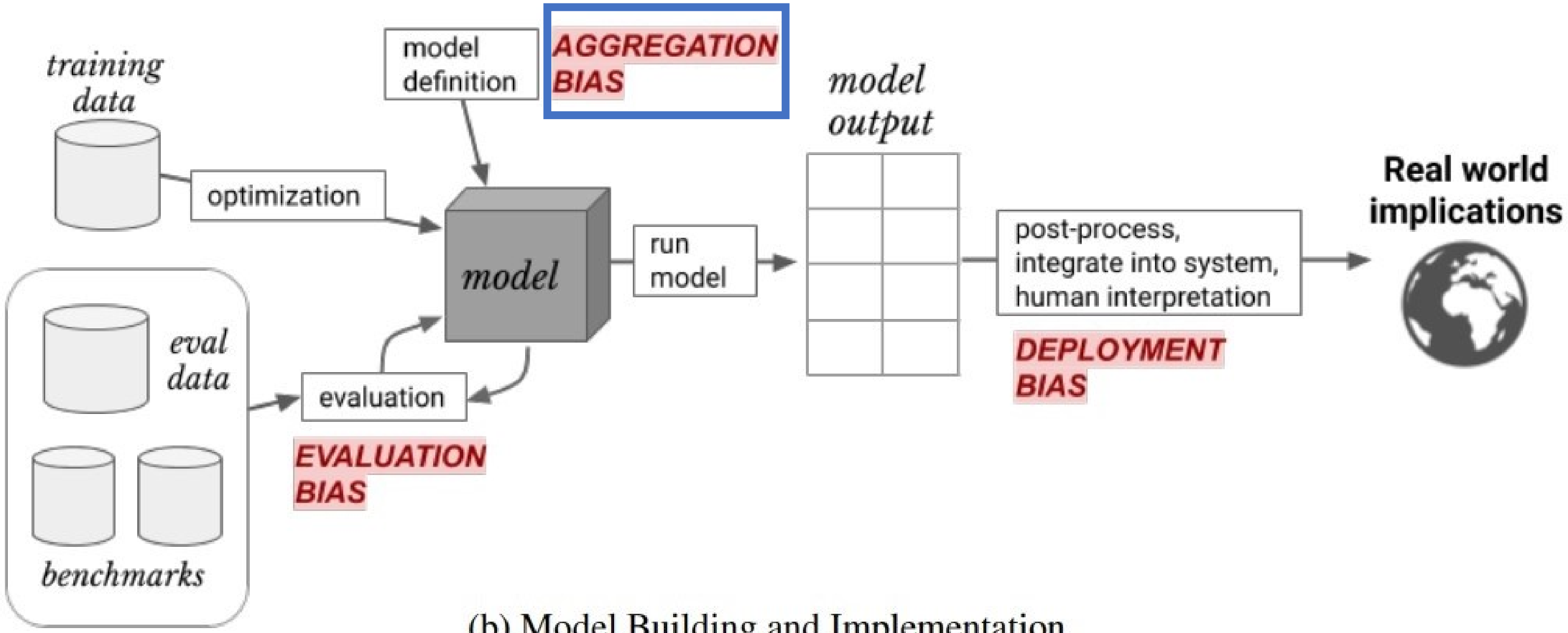
Predicting Recidivism

Source: [“Machine Bias” by ProPublica, 2016]()

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Aggregation Bias

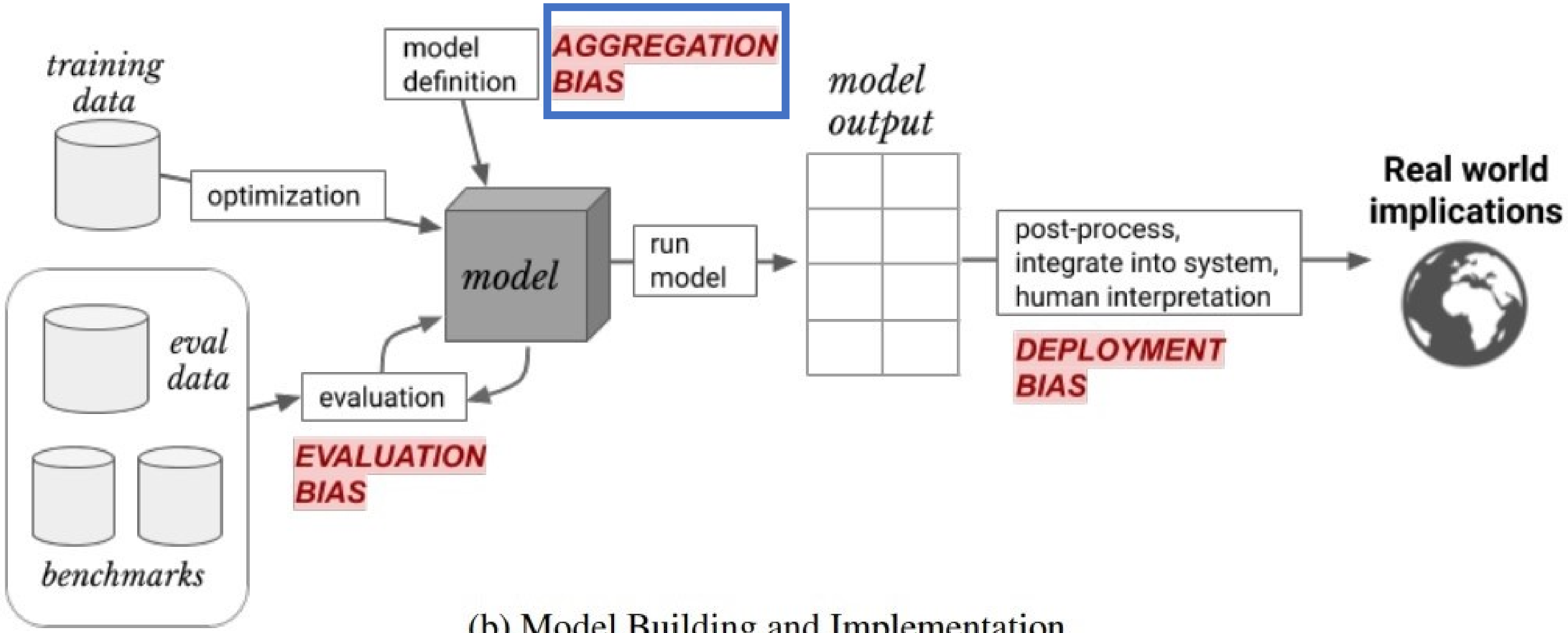


(b) Model Building and Implementation

Amazon scraps secret AI recruiting tool that showed bias against women

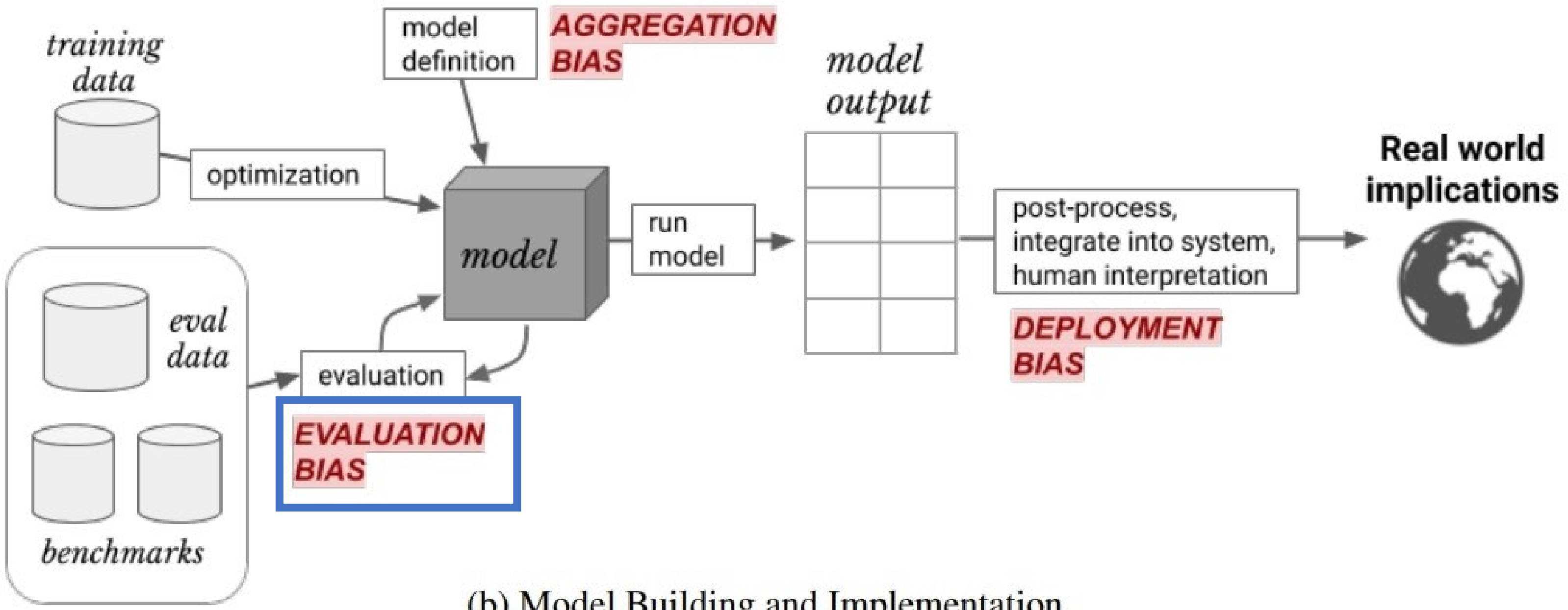
Source: Reuters 2018

"In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter."



(b) Model Building and Implementation

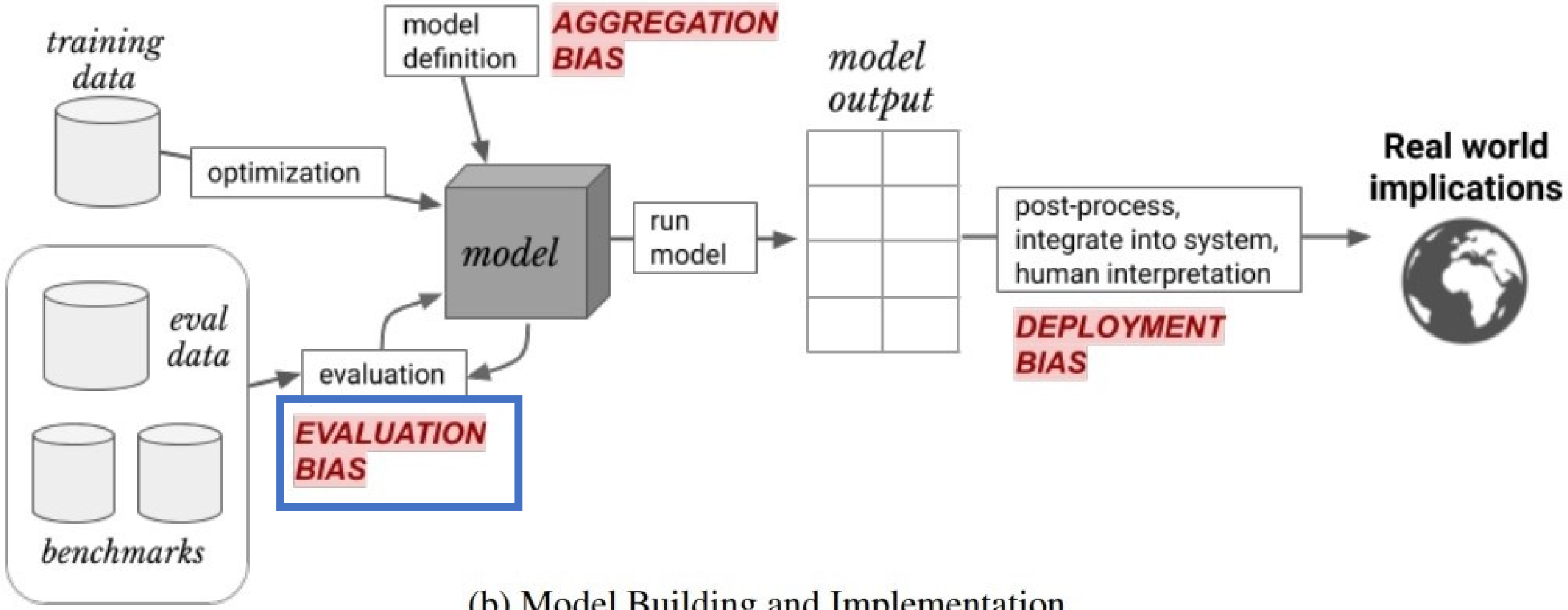
Evaluation Bias



(b) Model Building and Implementation

Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%

[Source: gendershades.org](gendershades.org)

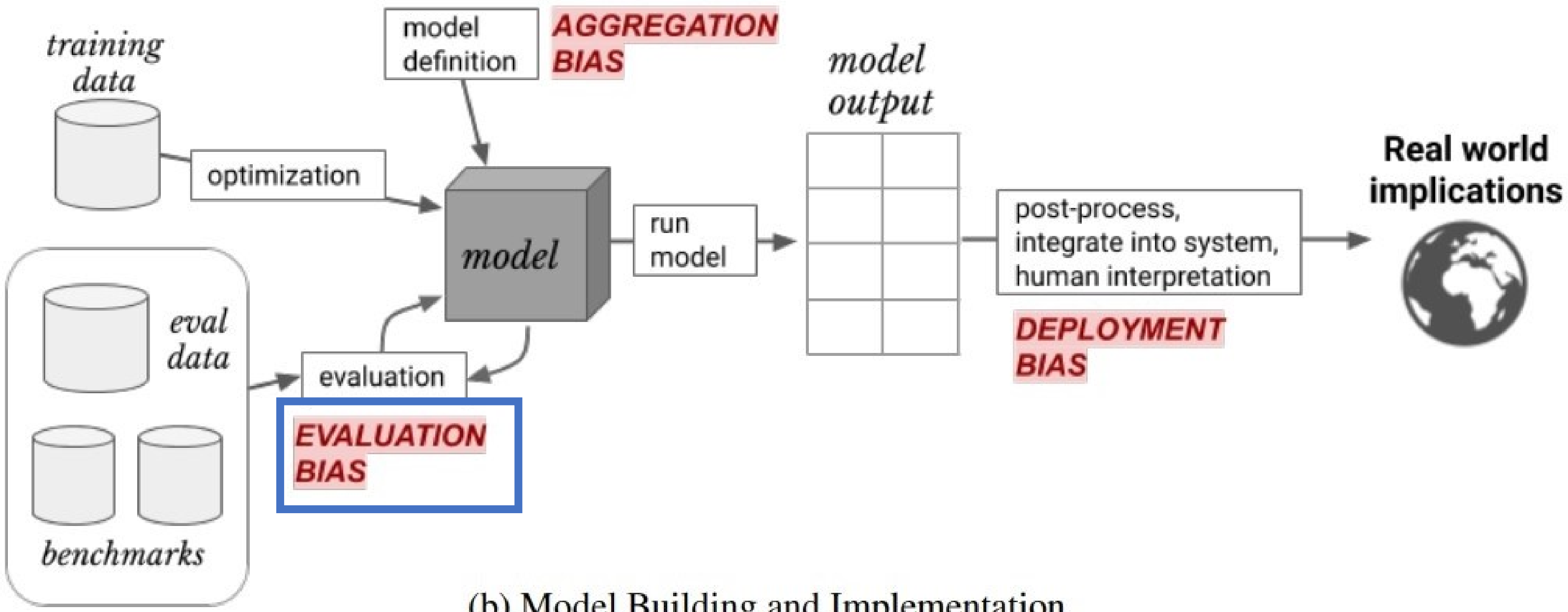


(b) Model Building and Implementation

Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

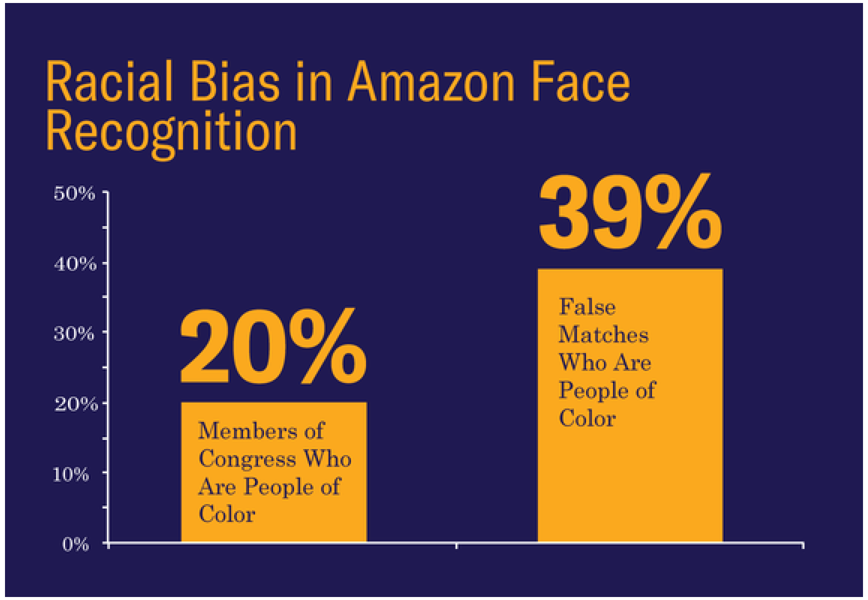
Source: gendershades.org

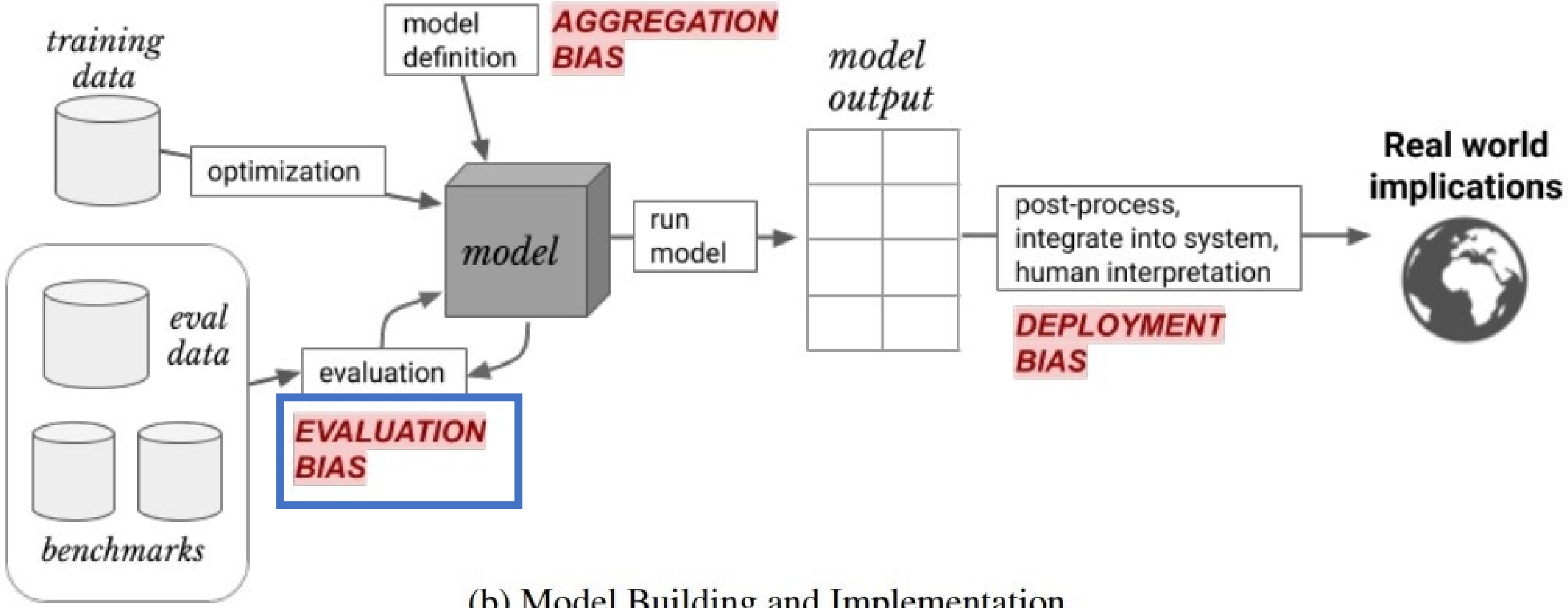


(b) Model Building and Implementation

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

[Source: ACLU]



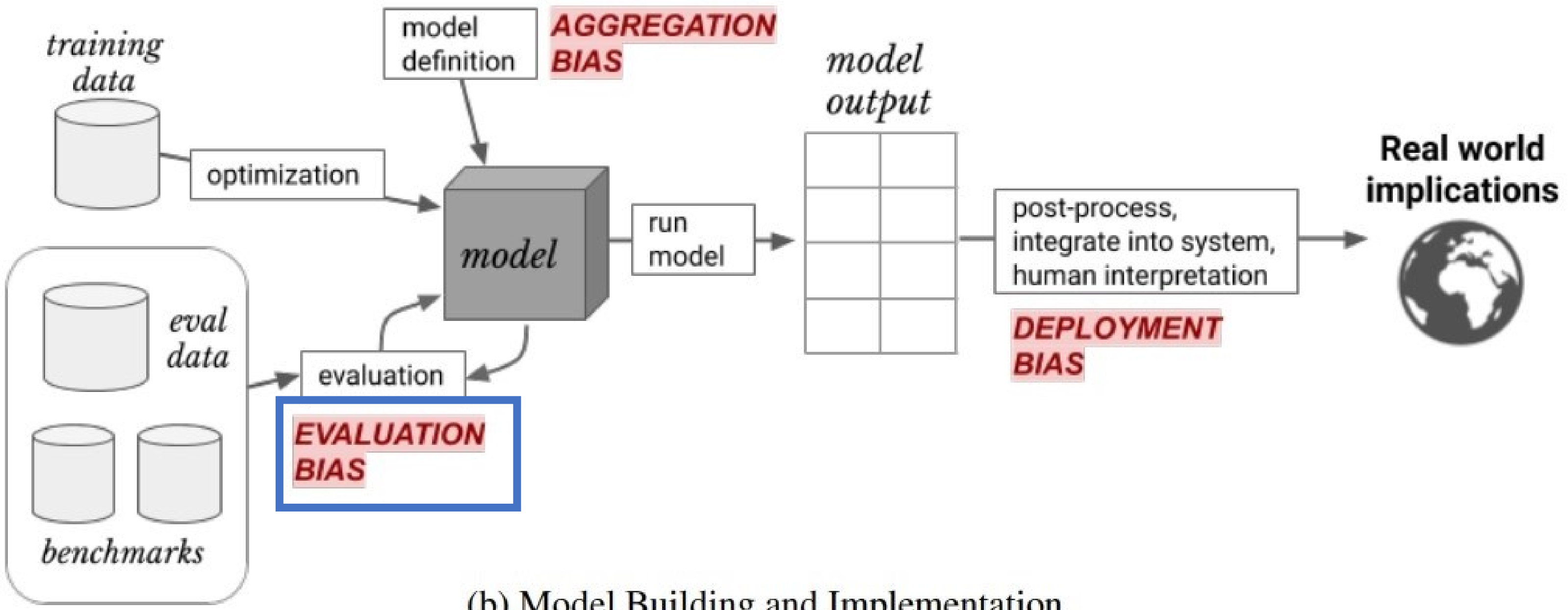


(b) Model Building and Implementation

A black man was wrongfully arrested because of facial recognition

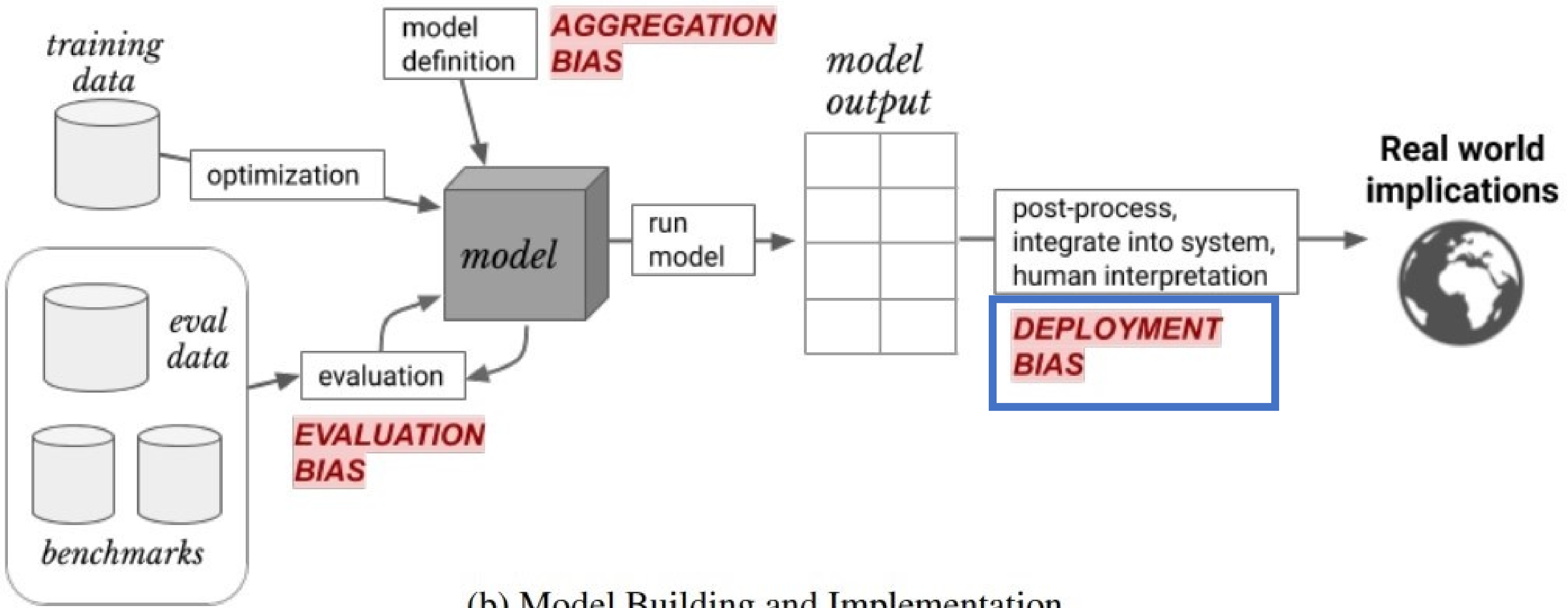
'The computer must have gotten it wrong'

[Source: The Verge]



(b) Model Building and Implementation

Deployment Bias



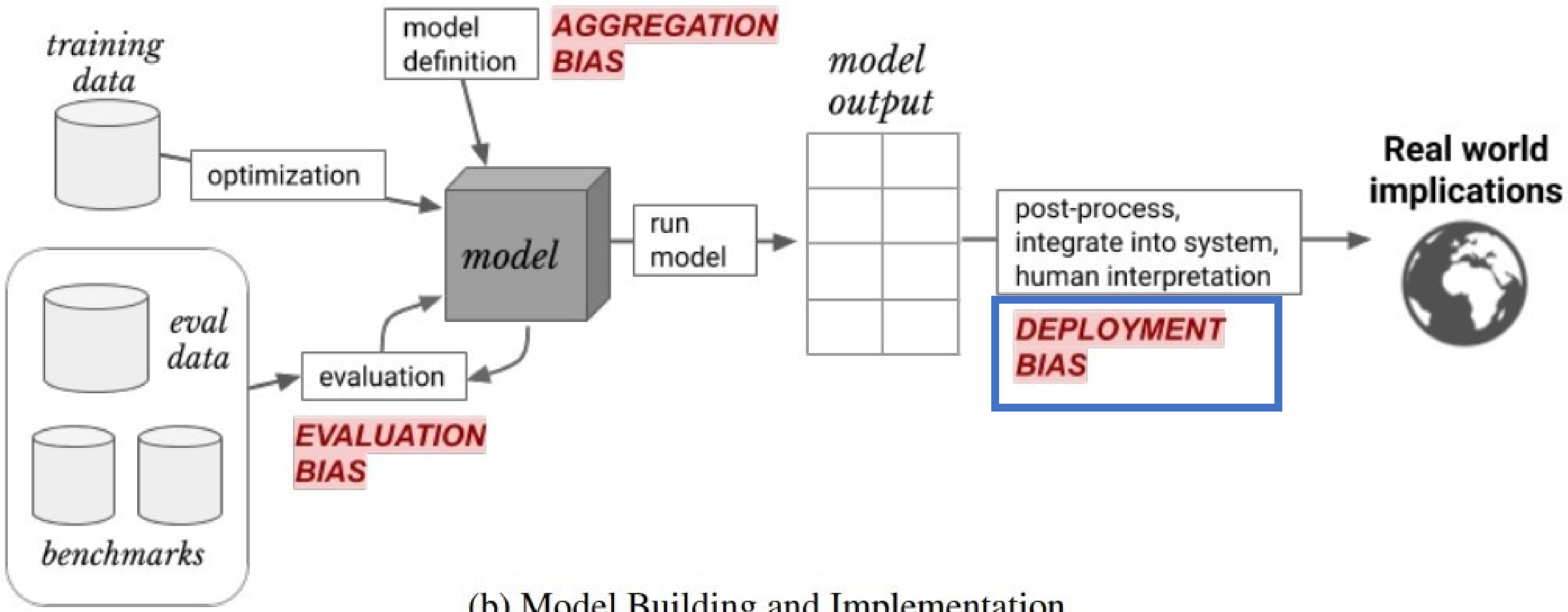
(b) Model Building and Implementation

A Child Abuse Prediction Model Fails Poor Families

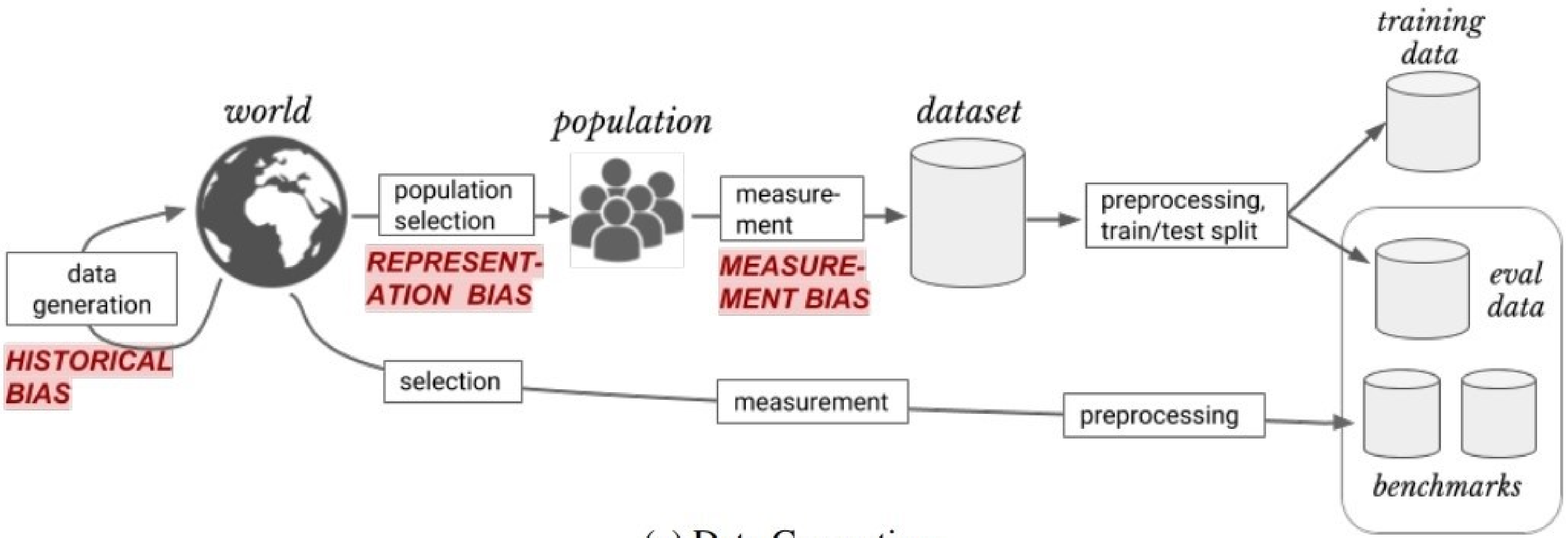
Why Pittsburgh's predictive analytics misdiagnoses child maltreatment and prescribes the wrong solutions

The screen that displays the AFST risk score states clearly that the system **"is not intended to make investigative or other child welfare decisions."**

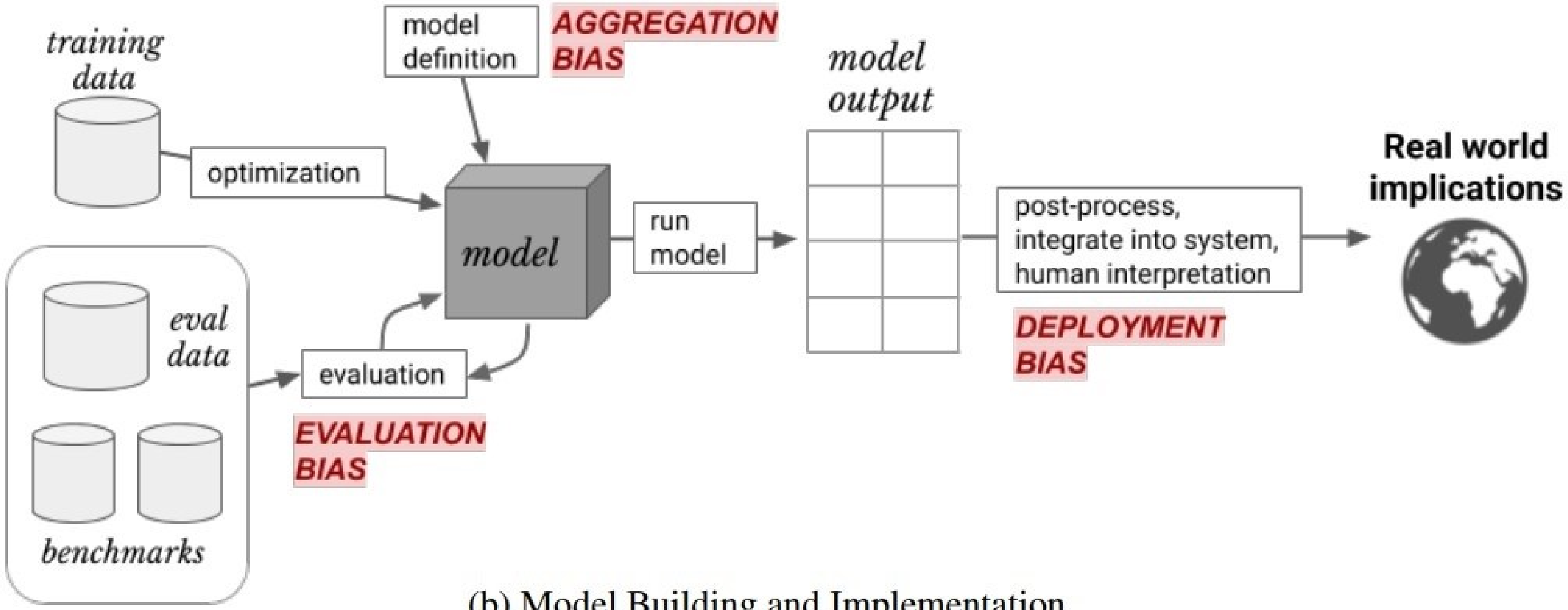
[Source: Automating Inequality]



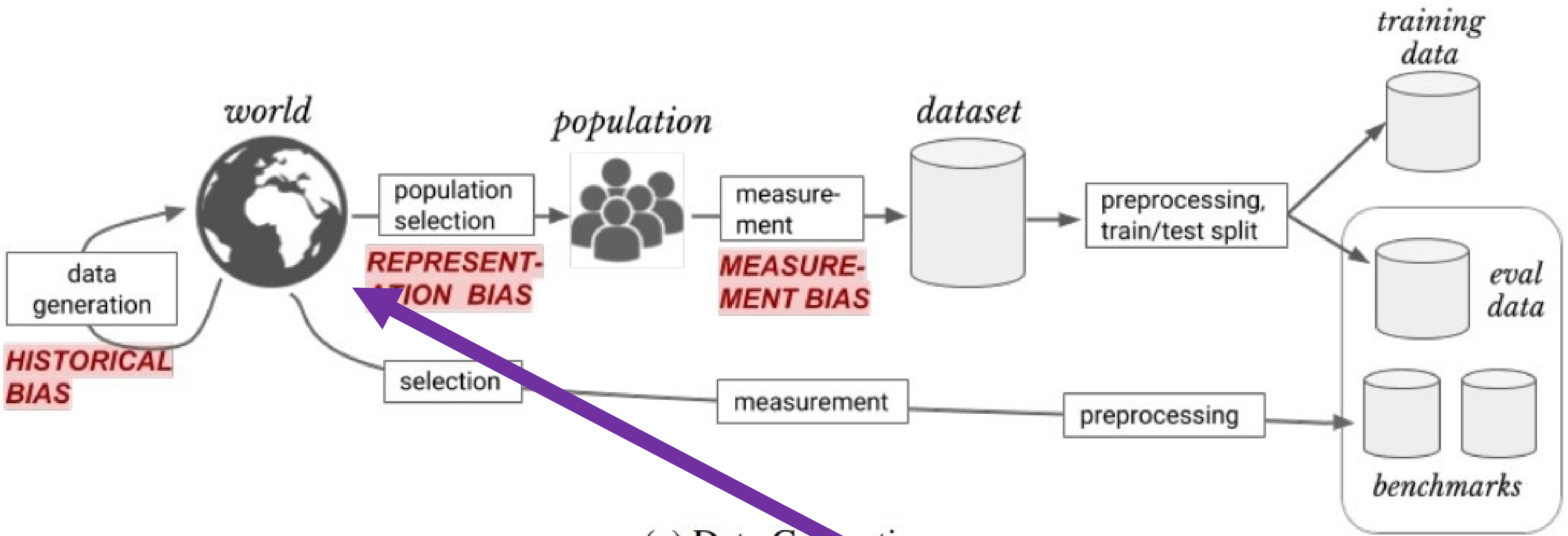
(b) Model Building and Implementation



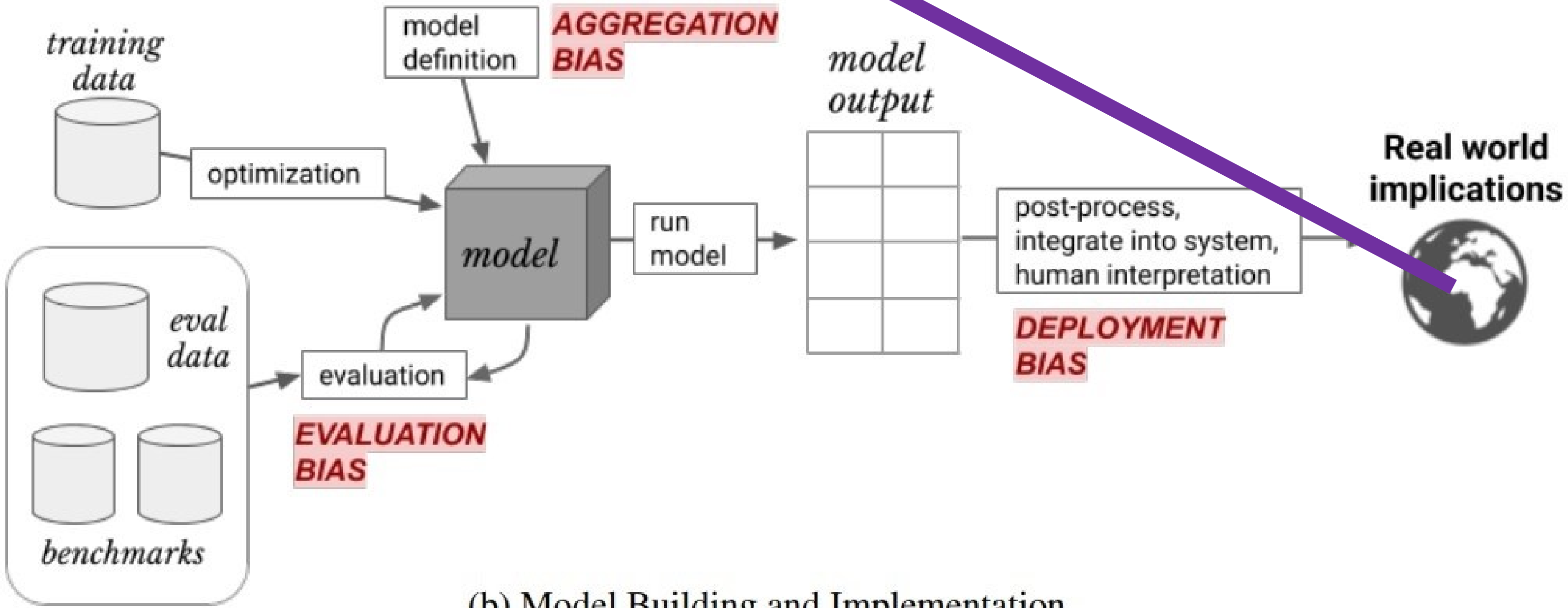
(a) Data Generation



(b) Model Building and Implementation



(a) Data Generation



(b) Model Building and Implementation

Why can’t we just omit any protected attributes from the dataset?

Why can't we just omit any protected attributes from the dataset?

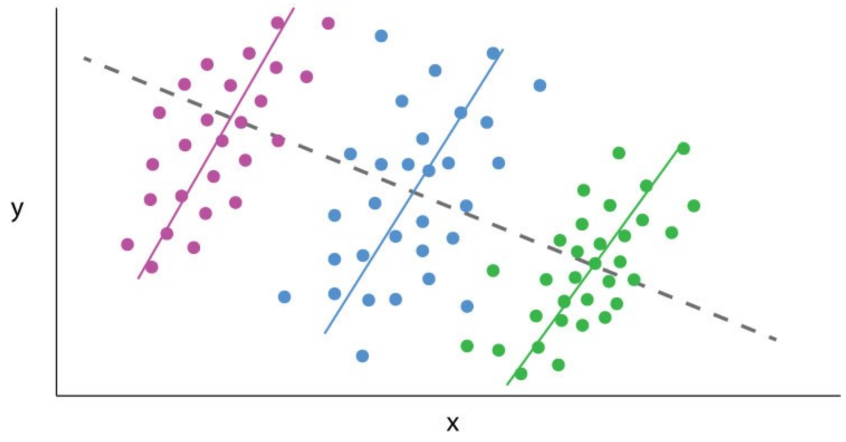
Latent Variables

& Proxies

Why can’t we just omit any protected attributes from the dataset?

Latent Variables & Proxies

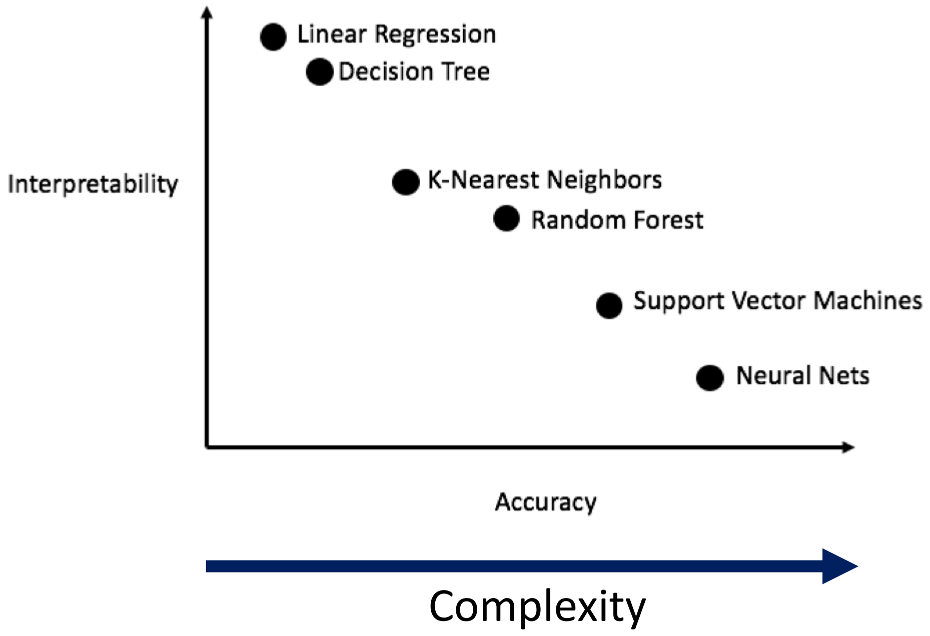
Simpson’s Paradox



Can we directly see if the inner workings of our algorithm are biased?

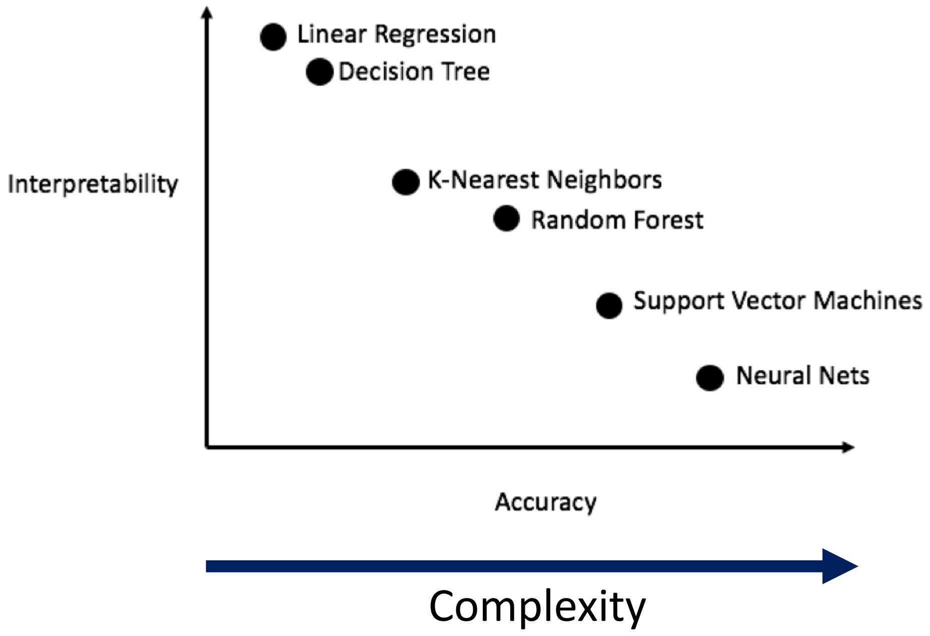
Model Explainability

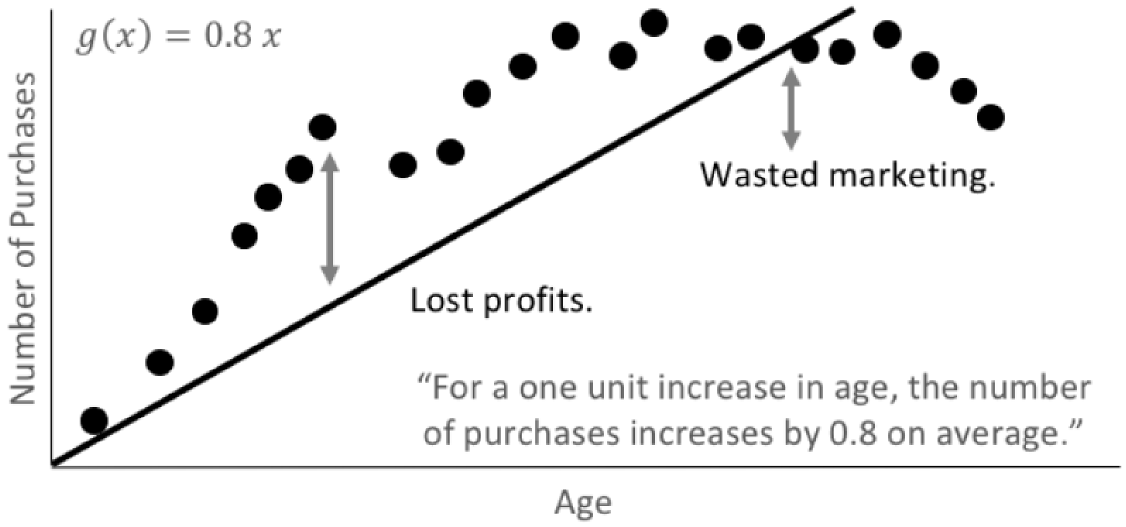
[Source: h2o.ai]



Model Explainability

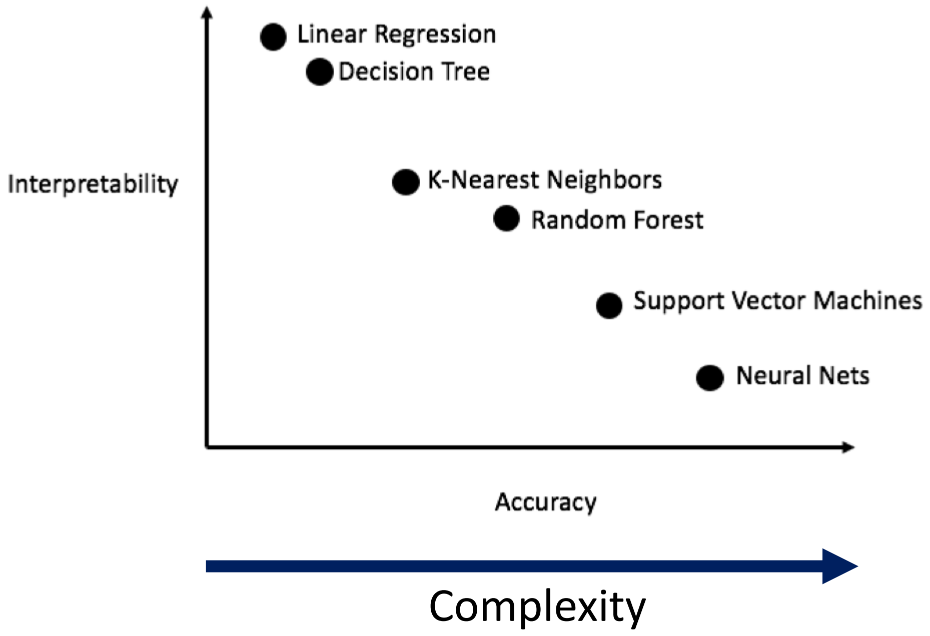
[Source: h2o.ai]

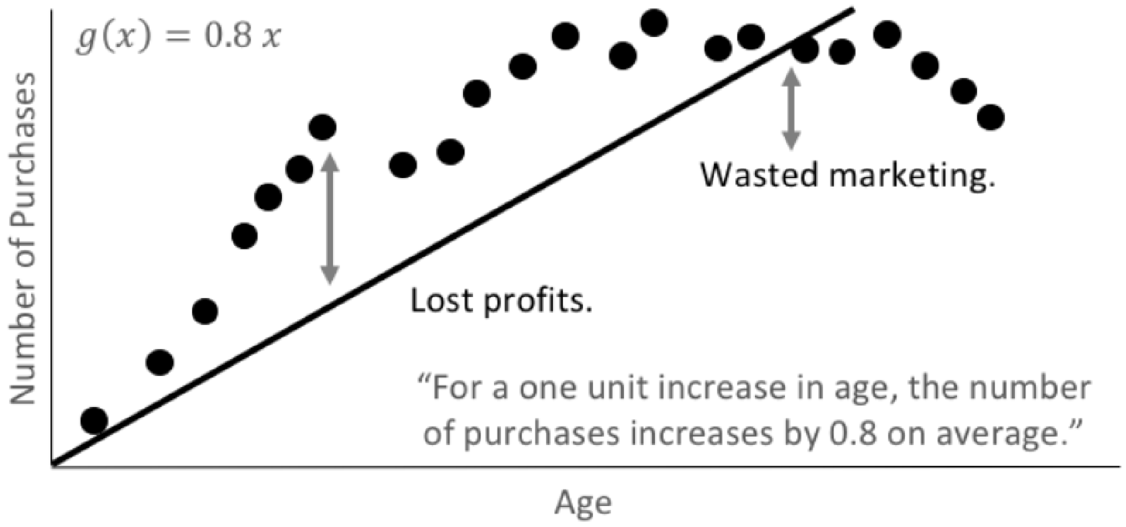


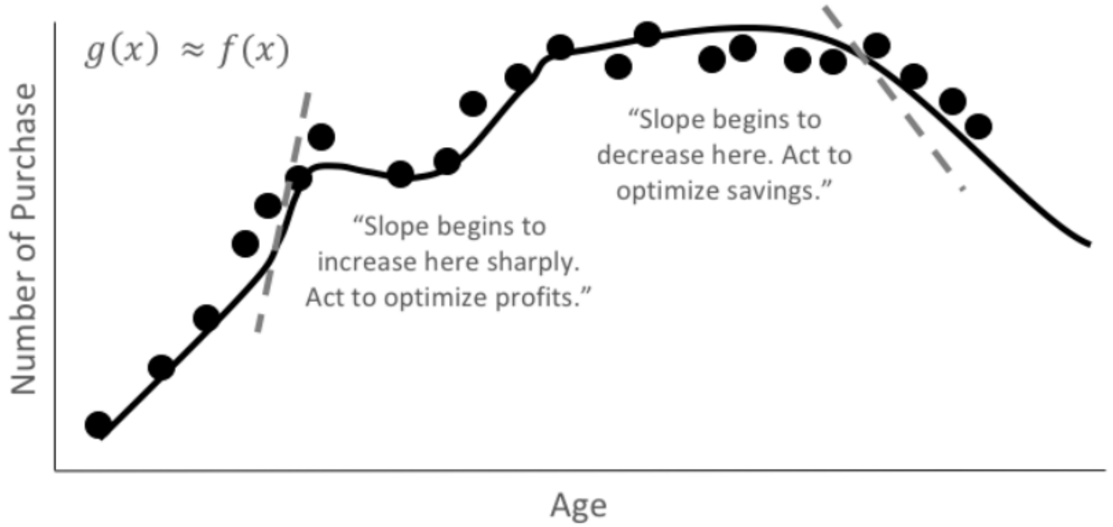


Model Explainability

Source: h2o.ai

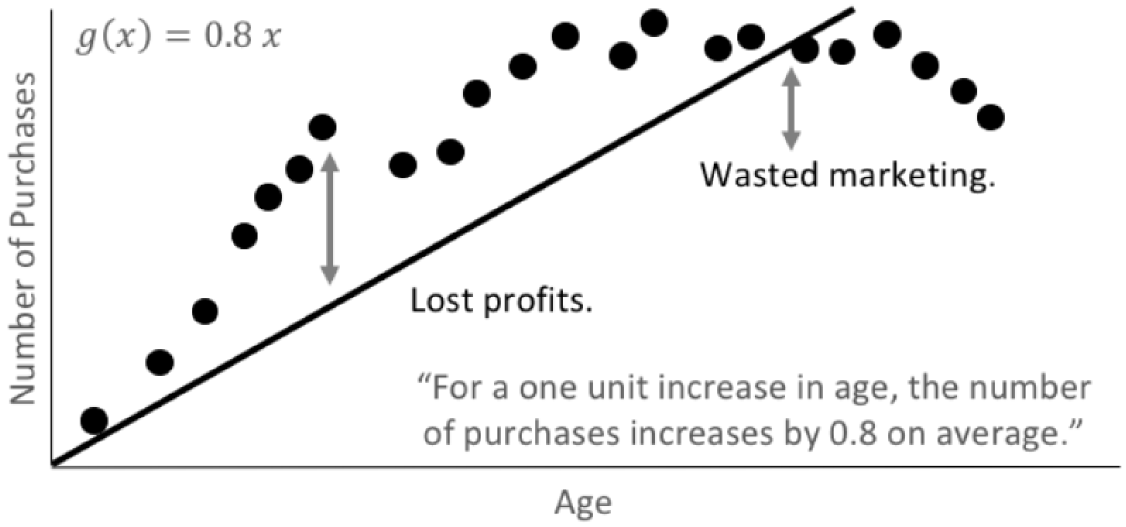






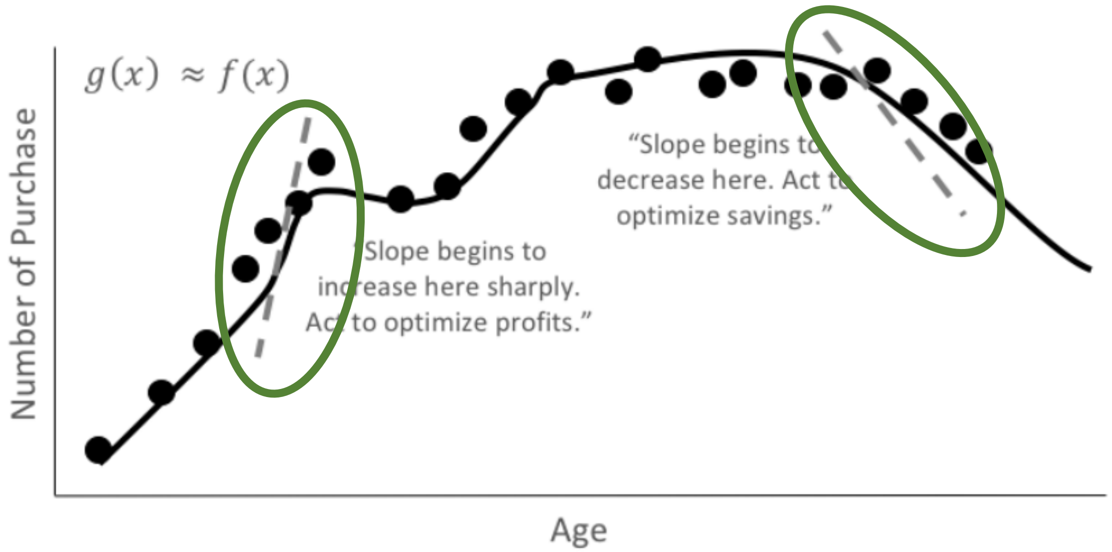
Model Explainability

[Source: h2o.ai]



Surrogate Models

Simpler models trained on **same inputs** and **predicted outputs** of more complex machine learning models

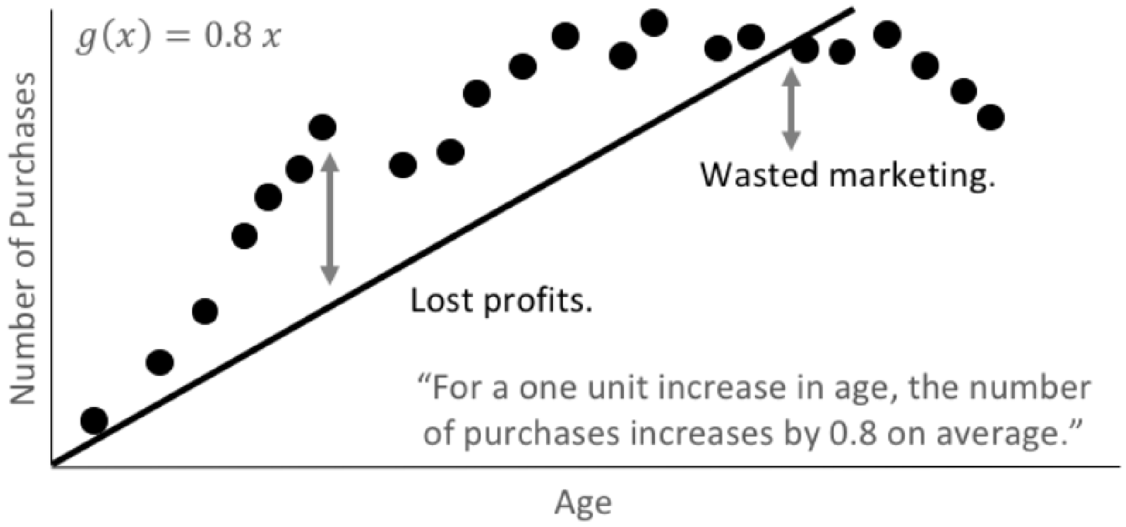


SHAP

Model Explainability

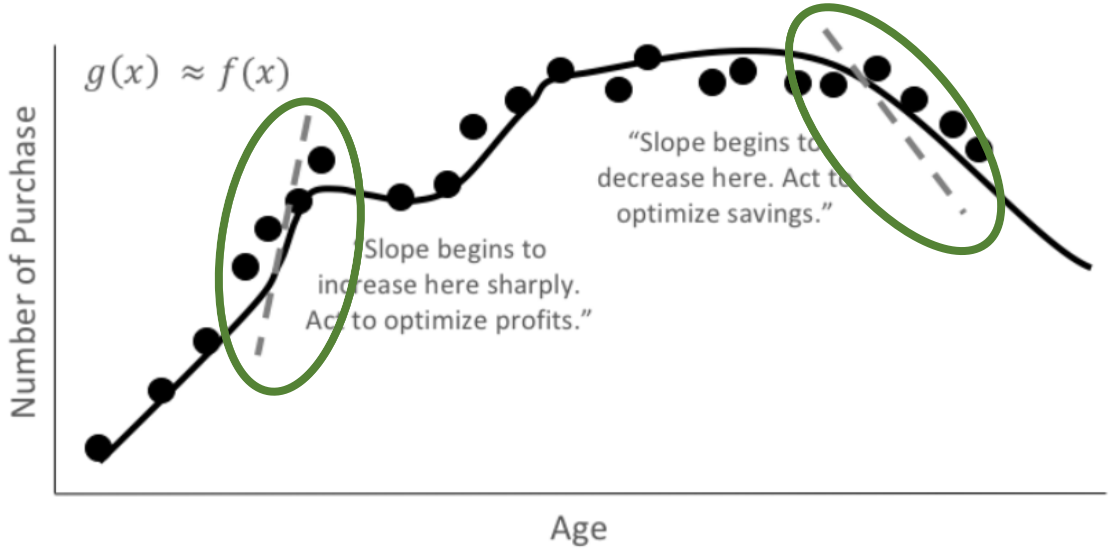
[Source: h2o.ai]

Often not good enough!



Surrogate Models

Simpler models trained on **same inputs** and **predicted outputs** of more complex machine learning models



SHAP

What does a "perfectly unbiased" algorithm look like?

What does a “perfectly unbiased” algorithm look like?

1. An algorithm that always predicts correctly

What does a “perfectly unbiased” algorithm look like?

1. An algorithm that always predicts correctly

2. An algorithm that picks predictions randomly

What does a “perfectly unbiased” algorithm look like?

1. An algorithm that always predicts correctly

2. An algorithm that picks predictions randomly

3. An algorithm that makes “mistakes” equally across privileged and unprivileged data

What does a "perfectly unbiased" algorithm look like?

1. An algorithm that always predicts correctly ✗
2. An algorithm that picks predictions randomly ✗
3. An algorithm that makes "mistakes" equally across privileged and unprivileged data ✓

Can we quantitatively define fairness?

Defining Fairness

Goal: Create a metric that machine learning algorithm
can use to generate fair outcomes

Defining Fairness

Goal: Create a metric that machine learning algorithm
can use to generate fair outcomes

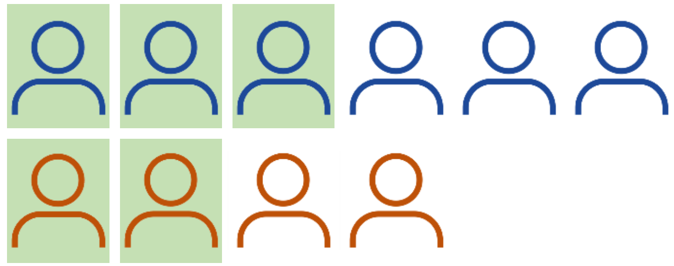
Definitions:

- Y is the true value (0 or 1 for binary classification)
- C is the algorithm’s predicted value
- A is the protected attribute (gender, race, etc.)
    - A=1 refers to the unprivileged group, A=0 refers to privileged

Defining Fairness:

Demographic Parity

“A predictor satisfies demographic parity **if the likelihood of a positive outcome is the same**, regardless of whether the person is in the protected group or not”

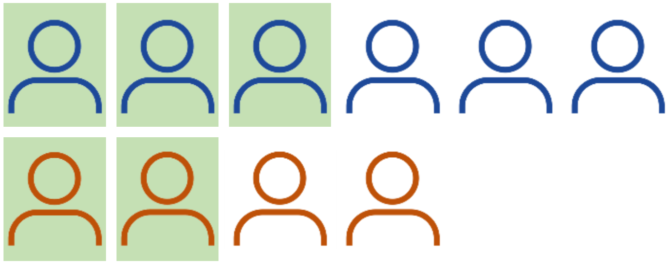


Defining Fairness:

Demographic Parity

"A predictor satisfies demographic parity **if the likelihood of a positive outcome is the same**, regardless of whether the person is in the protected group or not"

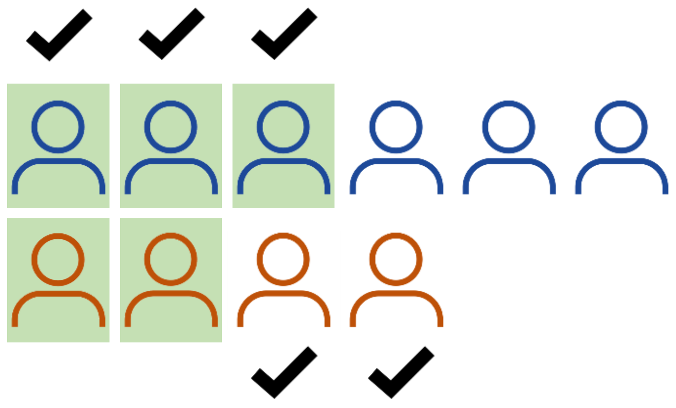
Pros: Proportional representation of groups



Defining Fairness:

Demographic Parity

"A predictor satisfies demographic parity **if the likelihood of a positive outcome is the same**, regardless of whether the person is in the protected group or not"



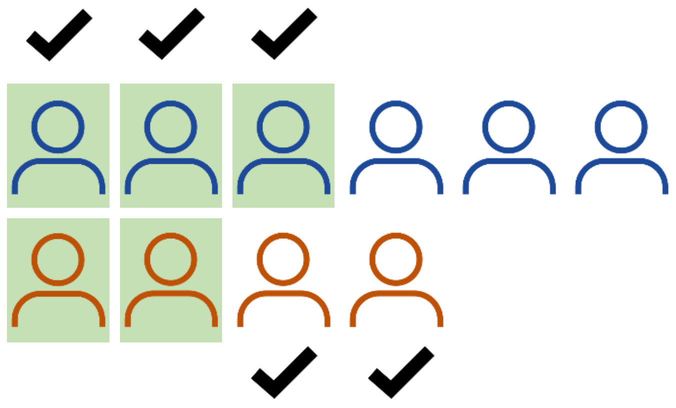
Pros: Proportional representation of groups

Cons: Accuracy may be less in disadvantaged group

Defining Fairness:

Demographic Parity

“A predictor satisfies demographic parity **if the likelihood of a positive outcome is the same**, regardless of whether the person is in the protected group or not”



Pros: Proportional representation of groups

Cons: Accuracy may be less in disadvantaged group

Greatly reduces effectiveness of predictor if true labels have any correlation with protected attribute

Defining Fairness:

Equal Odds

“A predictor *C* satisfies equalized odds with respect to a protected attribute *A* and the true outcome *Y* if C and A are independent conditional on Y”

Defining Fairness:

Equal Odds

"A predictor *C* satisfies equalized odds with respect to a protected attribute *A* and the true outcome *Y* if C and A are independent conditional on Y"

In a binary classification:

- C has **equal true positive rates** if Y=1 for both A=0 and A=1

A=0

	Y=1	**Y=0**
C=1	TP	FP
C=0	FN	TN

A=1

	Y=1	**Y=0**
C=1	TP	FP
C=0	FN	TN

Defining Fairness:

Equal Odds

"A predictor *C* satisfies equalized odds with respect to a protected attribute *A* and the true outcome *Y* if C and A are independent conditional on Y"

In a binary classification:

- C has **equal true positive rates** if Y=1 for both A=0 and A=1
- C has **equal false positive rates** if Y=0 for both A=0 and A=1

A=0

	Y=1	**Y=0**
C=1	TP	FP
C=0	FN	TN

A=1

	Y=1	**Y=0**
C=1	TP	FP
C=0	FN	TN

Defining Fairness:

Equal Odds

#	Qualified?	Hired?	Classification
2	Yes	Yes	True Positive
3	Yes	No	False Negative
4	No	Yes	False Positive
5	No	No	True Negative
1	Yes	Yes	True Positive
1	Yes	No	False Negative
2	No	Yes	False Positive
3	No	No	True Negative

Defining Fairness:

Equal Odds

#	Qualified?	Hired?	Classification	In-Group Rate
2	Yes	Yes	True Positive	2/14
3	Yes	No	False Negative	3/14
4	No	Yes	False Positive	4/14
5	No	No	True Negative	5/14
1	Yes	Yes	True Positive	1/7
1	Yes	No	False Negative	1/7
2	No	Yes	False Positive	2/7
3	No	No	True Negative	3/7

Defining Fairness:

Equal Odds

#	Qualified?	Hired?	Classification	In-Group Rate
2	Yes	Yes	True Positive	**2/14**
3	Yes	No	False Negative	3/14
4	No	Yes	False Positive	4/14
5	No	No	True Negative	5/14
1	Yes	Yes	True Positive	**1/7**
1	Yes	No	False Negative	1/7
2	No	Yes	False Positive	2/7
3	No	No	True Negative	3/7

Defining Fairness:

Equal Odds

#	Qualified?	Hired?	Classification	In-Group Rate
2	Yes	Yes	True Positive	**2/14**
3	Yes	No	False Negative	3/14
4	No	Yes	False Positive	**4/14**
5	No	No	True Negative	5/14
1	Yes	Yes	True Positive	**1/7**
1	Yes	No	False Negative	1/7
2	No	Yes	False Positive	**2/7**
3	No	No	True Negative	3/7

Defining Fairness:

Equal Odds

"Why not just accuracy?" (TP + TN)

A=0

	Y=1	Y=0
C=1	TP	FP
C=0	FN	TN

A=1

	Y=1	Y=0
C=1	TP	FP
C=0	FN	TN

Defining Fairness:

Equal Odds

“Why not just accuracy?” (TP + TN)

Weakness: We can “trade” the false positive rate of one group for the false negative rate for another group

Ex. Hiring from two groups. We can achieve accuracy parity by exchanging qualified applicants from privileged group for unqualified applicants from unprivileged group

A=0

	Y=1	Y=0
C=1	TP	FP
C=0	FN	TN

A=1

	Y=1	Y=0
C=1	TP	FP
C=0	FN	TN

Defining Fairness:

Equal Opportunity

- Relaxed version of Equal Odds
- Equal true positive rates for Y=1 for both A=0 and A=1
- Useful when only care about positive outcome

A=0

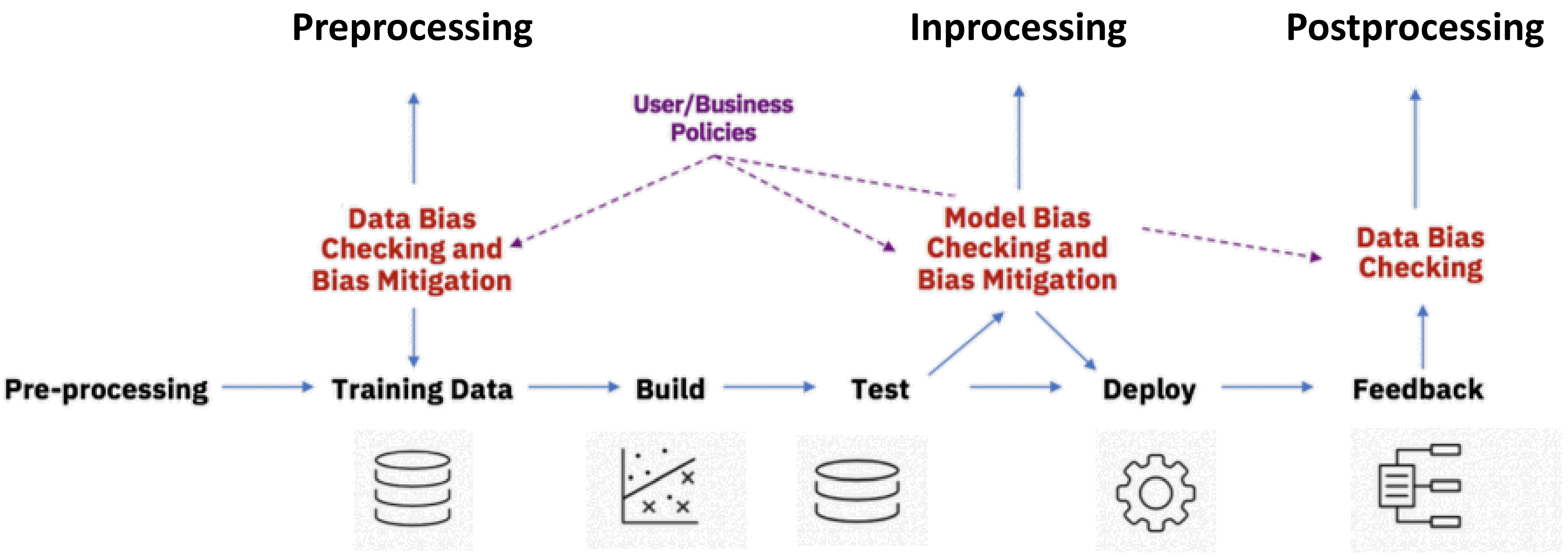
	Y=1	**Y=0**
C=1	TP	FP
C=0	FN	TN

A=1

	Y=1	**Y=0**
C=1	TP	FP
C=0	FN	TN

How can we actively mitigate bias and improve fairness?

Bias Mitigation Algorithms



[Source: IBM AIF360]

Bias Mitigation Algorithms:

Preprocessing

Bias Mitigation Algorithms:

Preprocessing

Disparate Impact Remover

Source: Feldman et. al 2015

Modify labels in the training dataset to ensure that the probability of a positive outcome is equivalent for both subgroups

Less strict - ratio of probabilities is greater than cutoff (typically 0.8)

$$\frac{P(C=1|A=1)}{P(C=1|A=0)} \leq \tau = 0.8$$

Bias Mitigation Algorithms:

Preprocessing

Disparate Impact Remover

[Source: Feldman et. al 2015]

Modify labels in the training dataset to ensure that the probability of a positive outcome is equivalent for both subgroups

Less strict - ratio of probabilities is greater than cutoff (typically 0.8)

$$\frac{P(C=1|A=1)}{P(C=1|A=0)} \leq \tau = 0.8$$

Reweighing

[Source: Kamiran, Calders 2010]

Weigh each observation in the training dataset by the expected probability of the observation ignoring the protected attribute.

(for algorithms that do not support custom weights, sampling may be used instead)

$$W(X) = \frac{P_{obs}(X)}{P_{exp}(X_{i \neq A})}$$

Bias Mitigation Algorithms:

Inprocessing

Bias Mitigation Algorithms:

Inprocessing

Prejudice Remover

[Source: Kamishima et. al 2012]

Defines prejudice index ***PI*** that increases as correlation between outcome ***C*** and protected attribute **A** increases:

$$\mathrm{PI} = P(C|A) \times \ln \frac{P(C|A)}{P(C)P(A)}$$

Use as **regularization term** in loss function – error goes up as correlation between outcome and protected attribute goes up

Bias Mitigation Algorithms:

Inprocessing

Prejudice Remover

[Source: Kamishima et. al 2012]

Defines prejudice index ***PI*** that increases as correlation between outcome ***C*** and protected attribute **A** increases:

$$\mathrm{PI} = P(C|A) \times \ln \frac{P(C|A)}{P(C)P(A)}$$

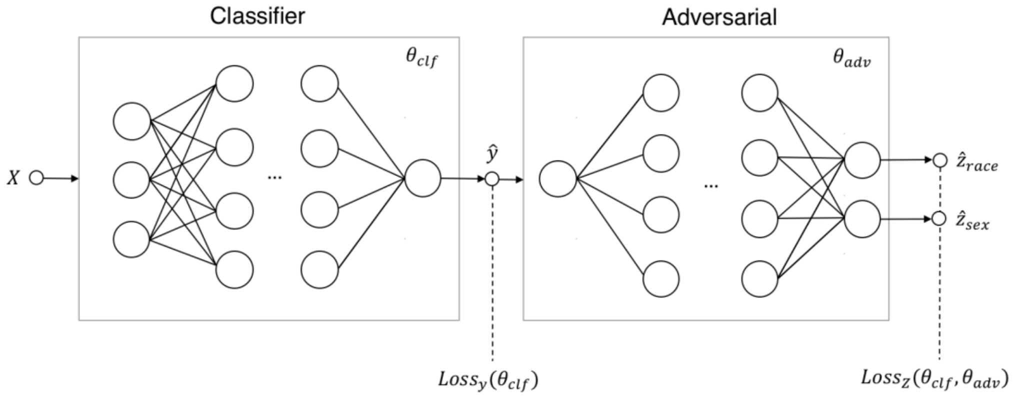
Use as **regularization term** in loss function – error goes up as correlation between outcome and protected attribute goes up

Adversarial Debiasing

[Source: Zhang et. al 2018]

When using a neural network to train model, set up a **second adversarial network** to predict protected attribute from the predictions of the first classifier.

Total loss minimizes class prediction performance and maximizes attribute prediction performance



Bias Mitigation Algorithms:

Postprocessing

Bias Mitigation Algorithms:

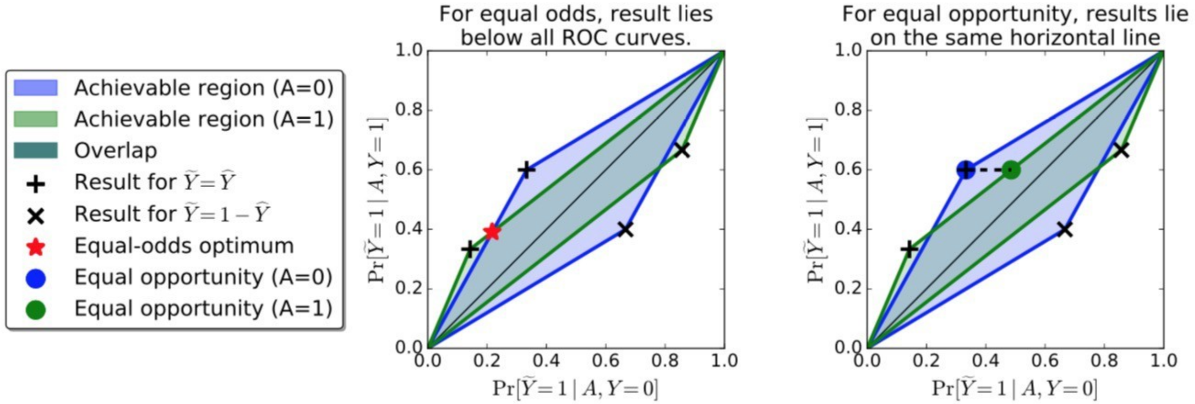
Postprocessing

Equal Odds

[Source: Kamishima et. al 2012]

A model's sensitivity and specificity can be tuned to optimize for metric like accuracy, precision, recall, or F1 score

We choose instead to tune the model to satisfy equal odds / equal opportunity



Bias Mitigation Algorithms:

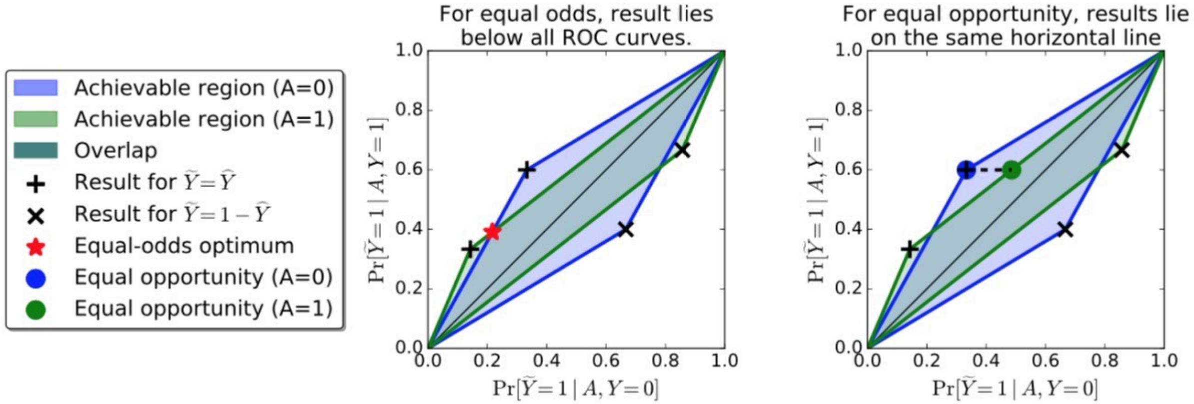
Postprocessing

Equal Odds

[Source: Kamishima et. al 2012]

A model's sensitivity and specificity can be tuned to optimize for metric like accuracy, precision, recall, or F1 score

We choose instead to tune the model to satisfy equal odds / equal opportunity



Rejection Option

[Source: Kamiran et. al 2012]

Based on the fact that most bias occurs on or near the decision boundary of the classifier

Flip favored classification to unprivileged group near the decision boundary until parity is reached

AIF360 Demo

Model Fairness AIF360 Demo Notebooks:
https://github.com/neptune-ai/model-fairness-in-practice

Amazon scraps secret AI recruiting tool that showed bias against women

LAPD ditches predictive policing program accused of racial bias

Crash Test Dummies Based on Men Pose Risks for Female Drivers

Face Recognition Vendor Vows New Rules After Wrongful Arrest in U.S. Using Its Technology

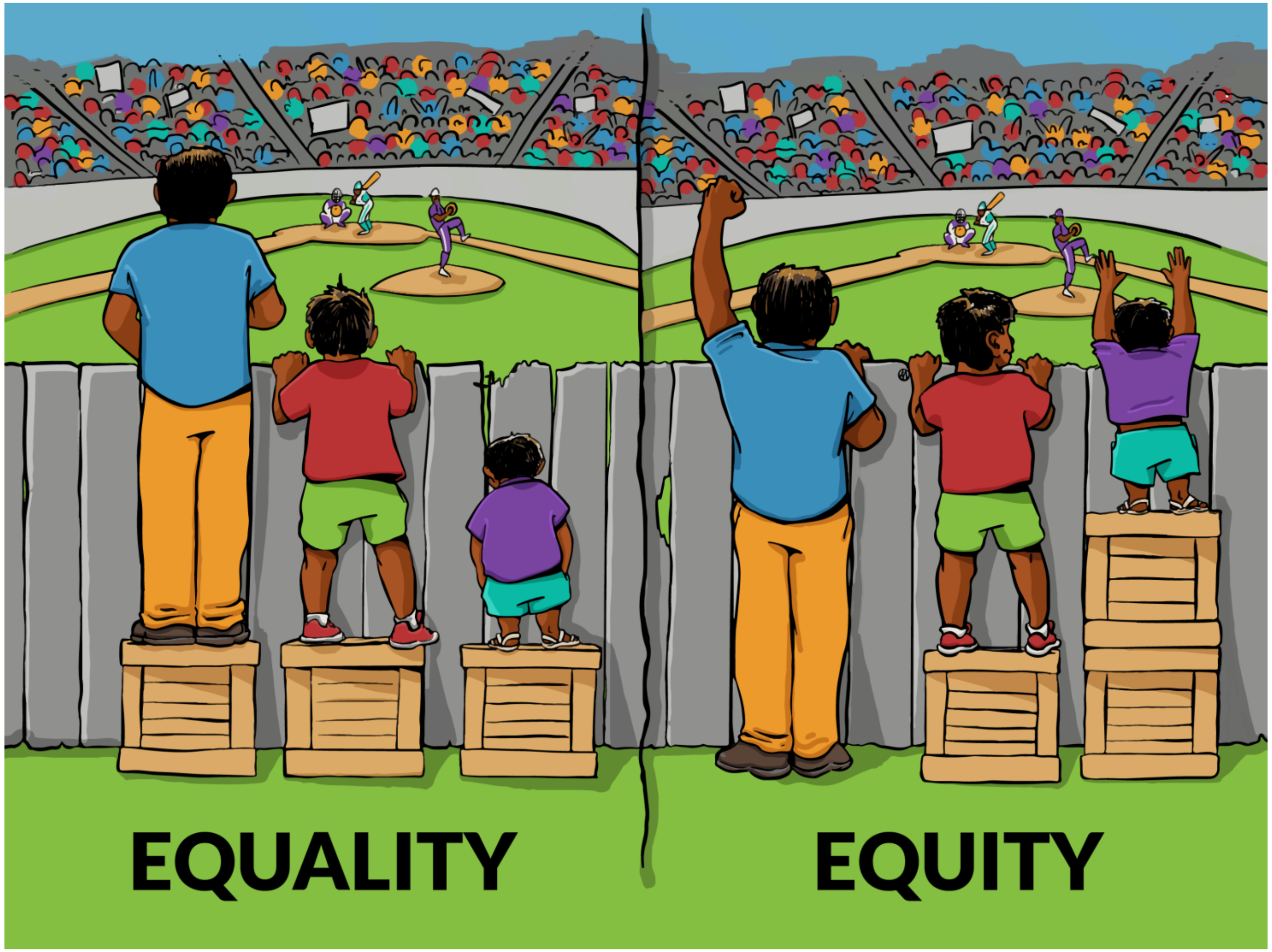


Image Source: Interaction Institute for Social Change

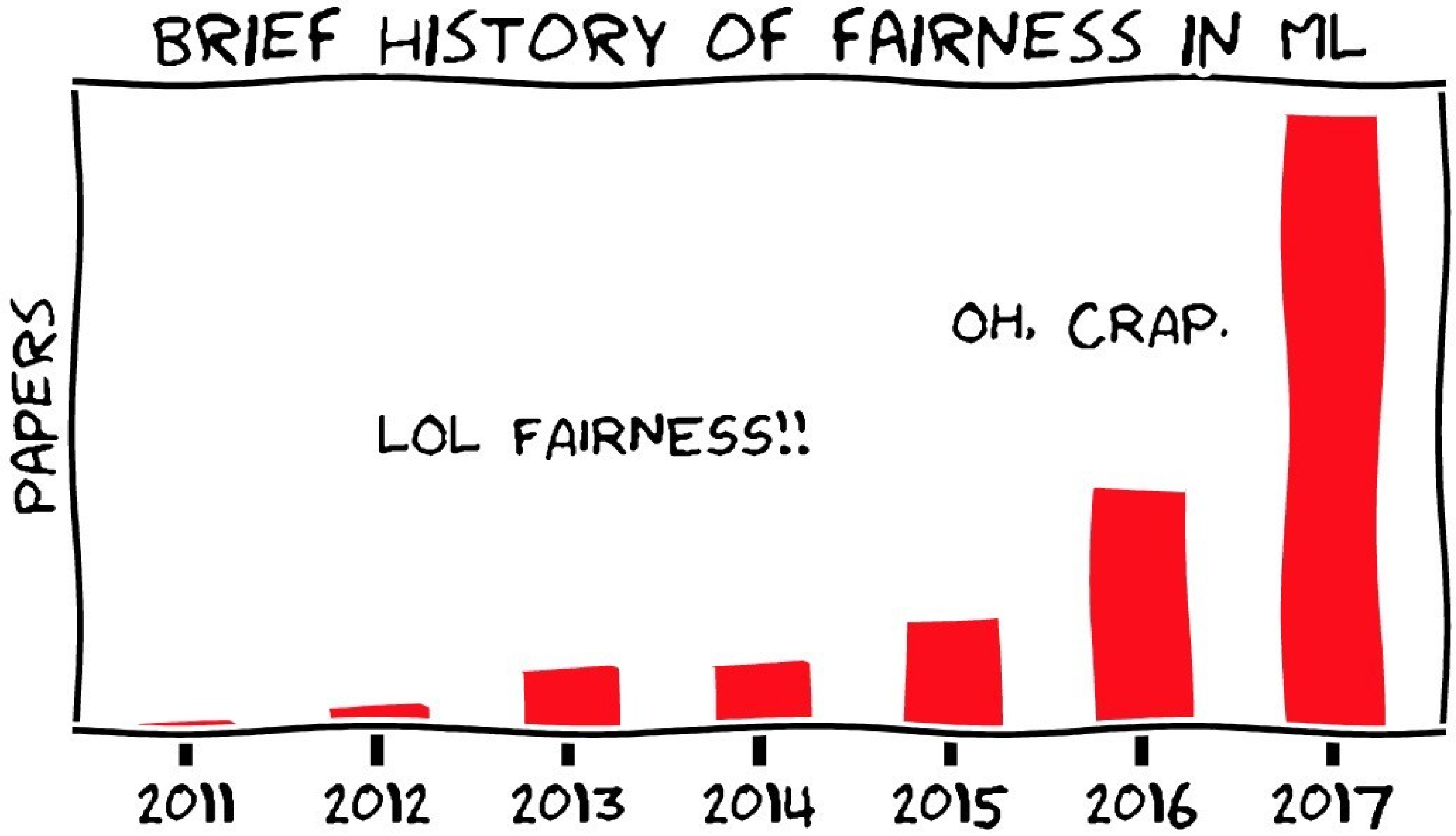


Image Source: towardsdatascience.com

Questions?