The background of the slide features a dark navy blue gradient. Overlaid on this are several thin, light blue wavy lines that curve and flow across the frame. Scattered throughout the space are numerous small, semi-transparent blue dots of varying sizes, some aligned with the wavy lines and others appearing independently.

Trusted AI

Diptikalyan Saha

IBM Research

Outline – 21 Sep

- Pillars of Trust –IBM perspective
- Fairness
 - Motivation
 - Metrics
 - Mitigation Algorithms for Group Fairness
 - Individual Discrimination Testing

AI is now used in many high-stakes decision making applications



Credit



Employment



Admission



Sentencing

Pillars of Trust

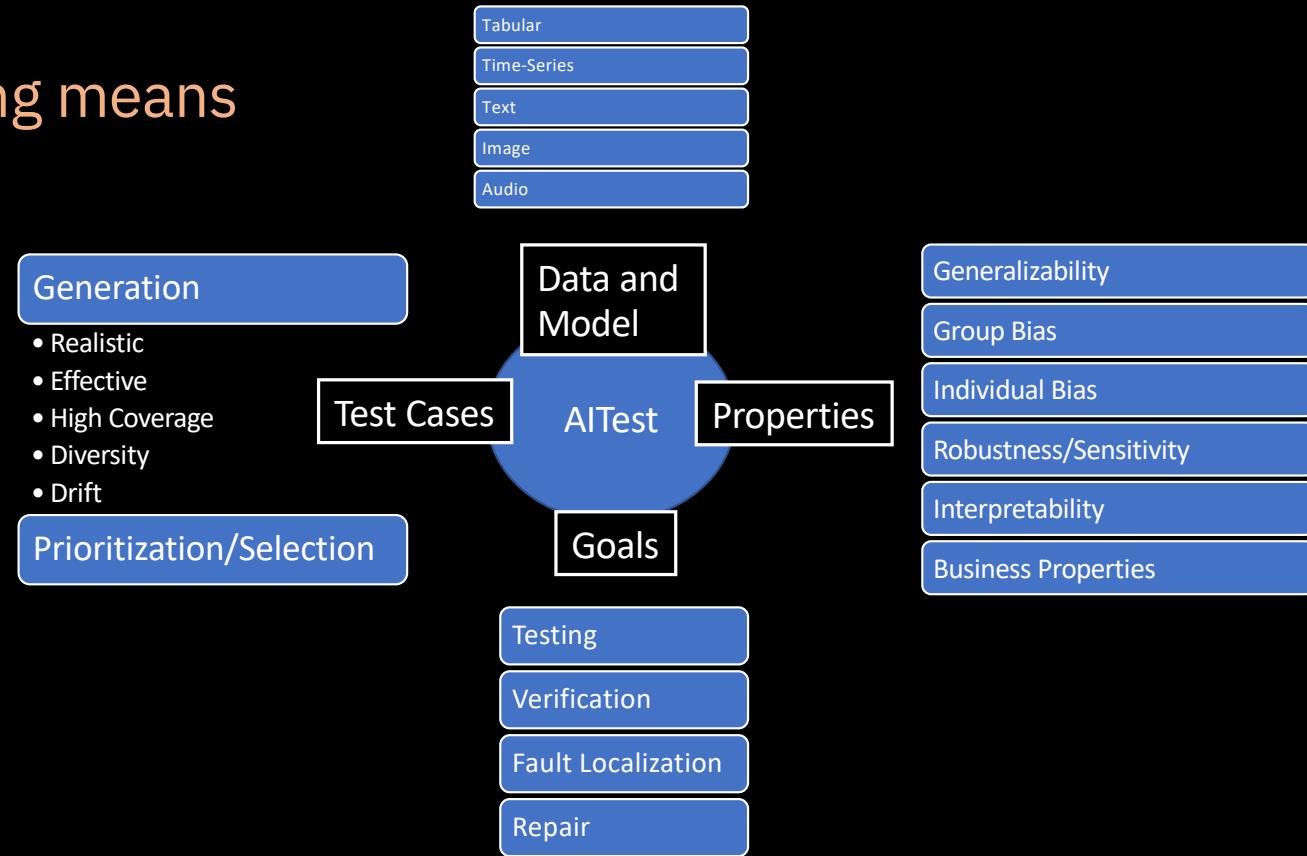
- Fairness
- Explainability
- Robustness
- Transparency
- Privacy

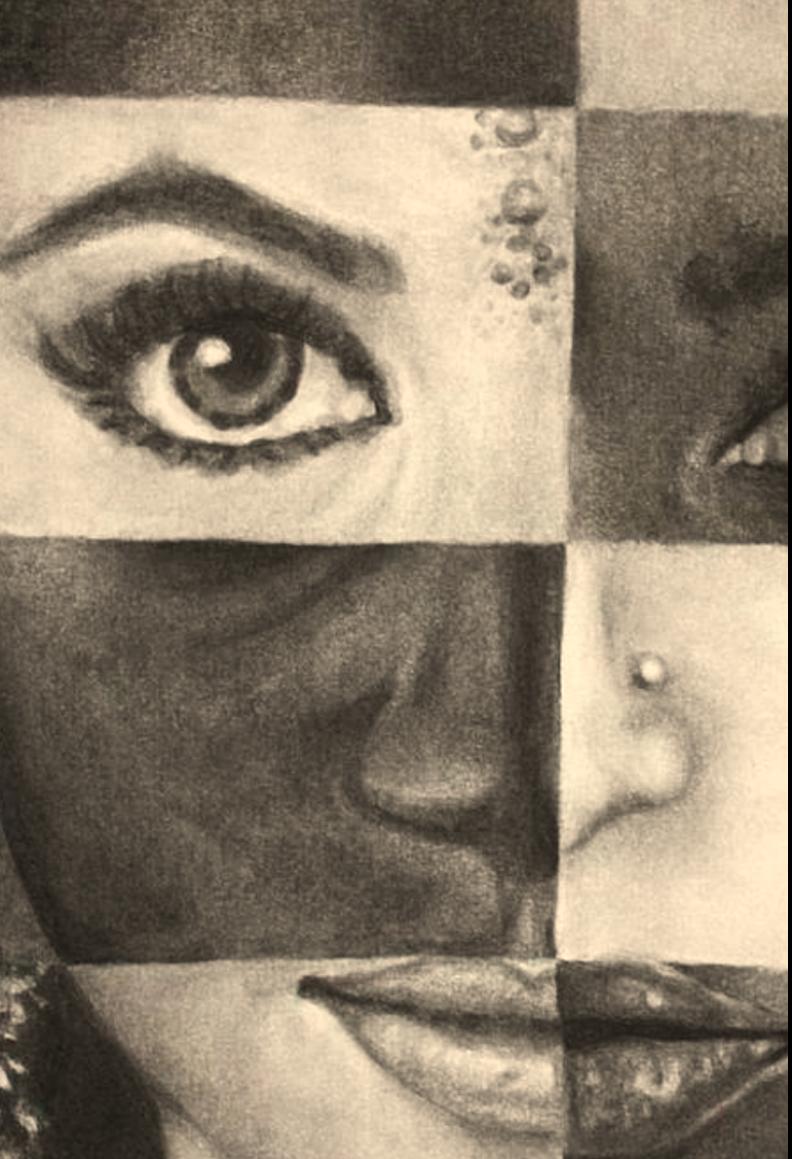
<https://www.ibm.com/watson/trustworthy-ai>

Testing for Trustworthiness of AI Systems

AI Testing Tabular Text Conversation Time-Series Image Speech2Text

- Forrester: No testing means no Trust in AI

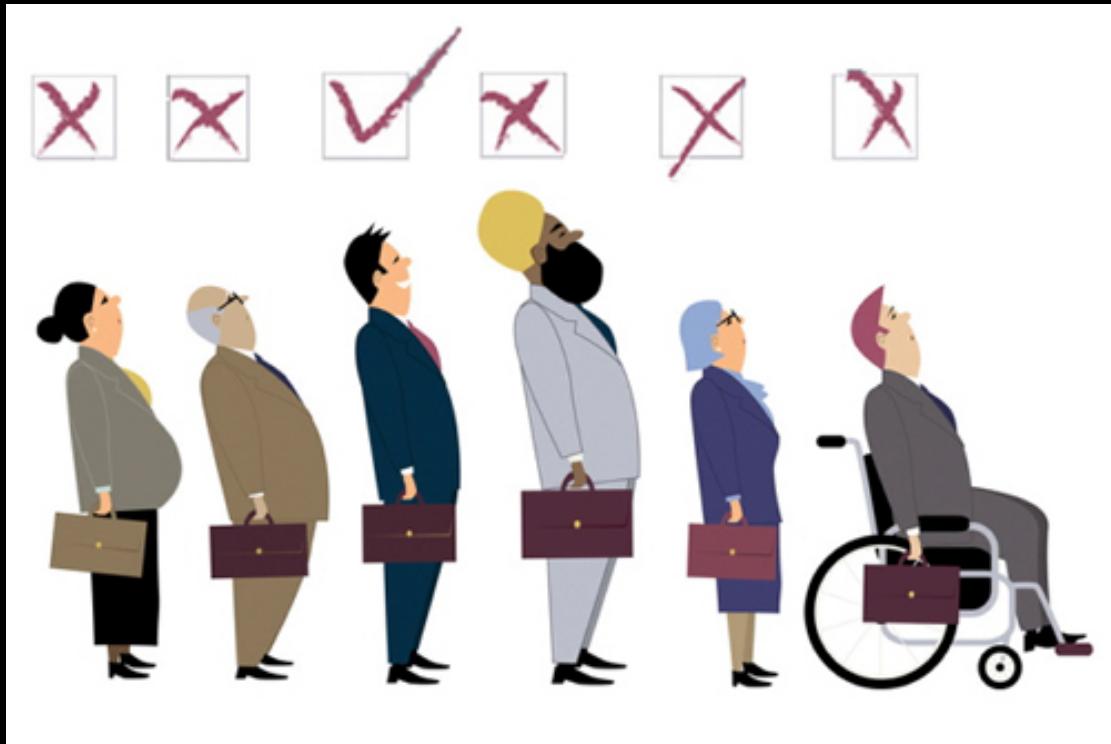




Fairness

- **Protected attribute** – an attribute that partitions a population into groups whose outcomes should have parity; examples include race, gender, caste, and religion
- **Privileged group** – a value of the protected attribute indicating a group that has historically been at systematic advantage
- **Bias** – a systematic error; in the context of fairness, we are concerned with **unwanted bias** that places privileged groups at systematic advantage and unprivileged groups at systematic disadvantage
- **Fairness metric** – a quantification of unwanted bias in training data or models
- **Favorable Outcome** – class label that is favorable to the user

What is unwanted bias?

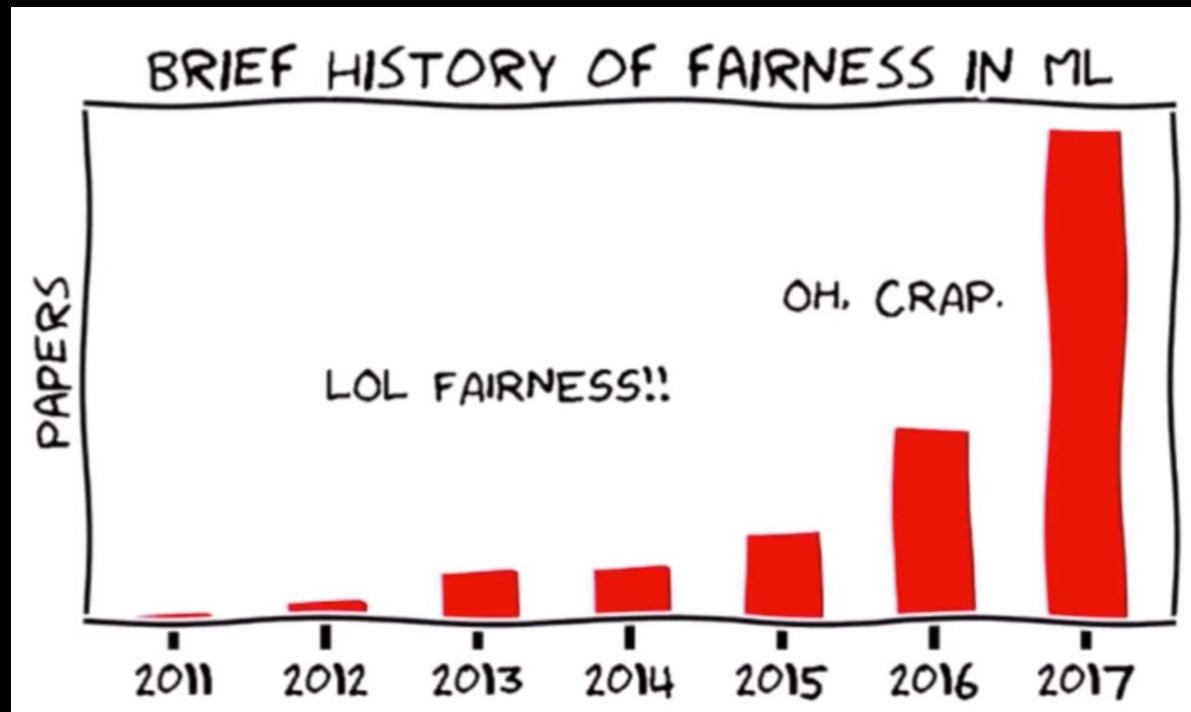


Discrimination becomes objectionable when it places certain **privileged** groups at systematic advantage and certain **unprivileged** groups at systematic disadvantage

Illegal in certain contexts

Unethical in general

(Barocas and Selbst, 2017)



(Hardt, 2017)

Motivation – Instances of Bias



Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ



SAN FRANCISCO (Reuters) - Amazon.com Inc's ([AMZN.O](#)) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

IBM's commercial facial recognition systems have been criticized in the past for displaying the very biases this dataset is intended to combat. A [study from MIT Media Lab](#) published in February found that IBM's error rate in identifying the gender of darker-skinned women was nearly 35 percent, while white men were misgendered only 1 percent of the time. Such mistakes will become increasingly important as facial recognition systems are used for tasks from hiring to the identification of criminal suspects.

Latinos in Iowa City faced worst U.S. bias in home loans, data show
www.press-citizen.com/story/news/2018/02/15/...bias-home-loans/340392002/

Feb 15, 2018 - Latinos seeking conventional home loans in the Iowa City area were nearly four times more likely to be denied than non-Hispanic whites in ...



Investigation reveals discriminatory home loan practices for minorities in San Antonio

Juan A. Lozano, Associated Press Updated 11:39 am CST, Thursday, February 15, 2018

Google engineer apologizes after Photos app tags two black people as gorillas

1. Timesofindia.com
- 2 <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scaps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>
- 3 <https://www.theverge.com/2018/6/27/17509400/facial-recognition-bias-ibm-data-training>
- 4 <https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas>
- 6 <https://www.mysanantonio.com/news/local/article/Review-Home-loan-bias-for-minorities-in-5-Texas-12616677.php>



Representational Harms - Stereotyping

"MAN IS TO COMPUTER PROGRAMMER
AS WOMAN IS TO HOMEMAKER?"



Translate

English Spanish French English - detected Translate

He is a nurse
She is a doctor

O bir hemşire
O bir doktor

Turn off instant translation Suggest an edit

Translate

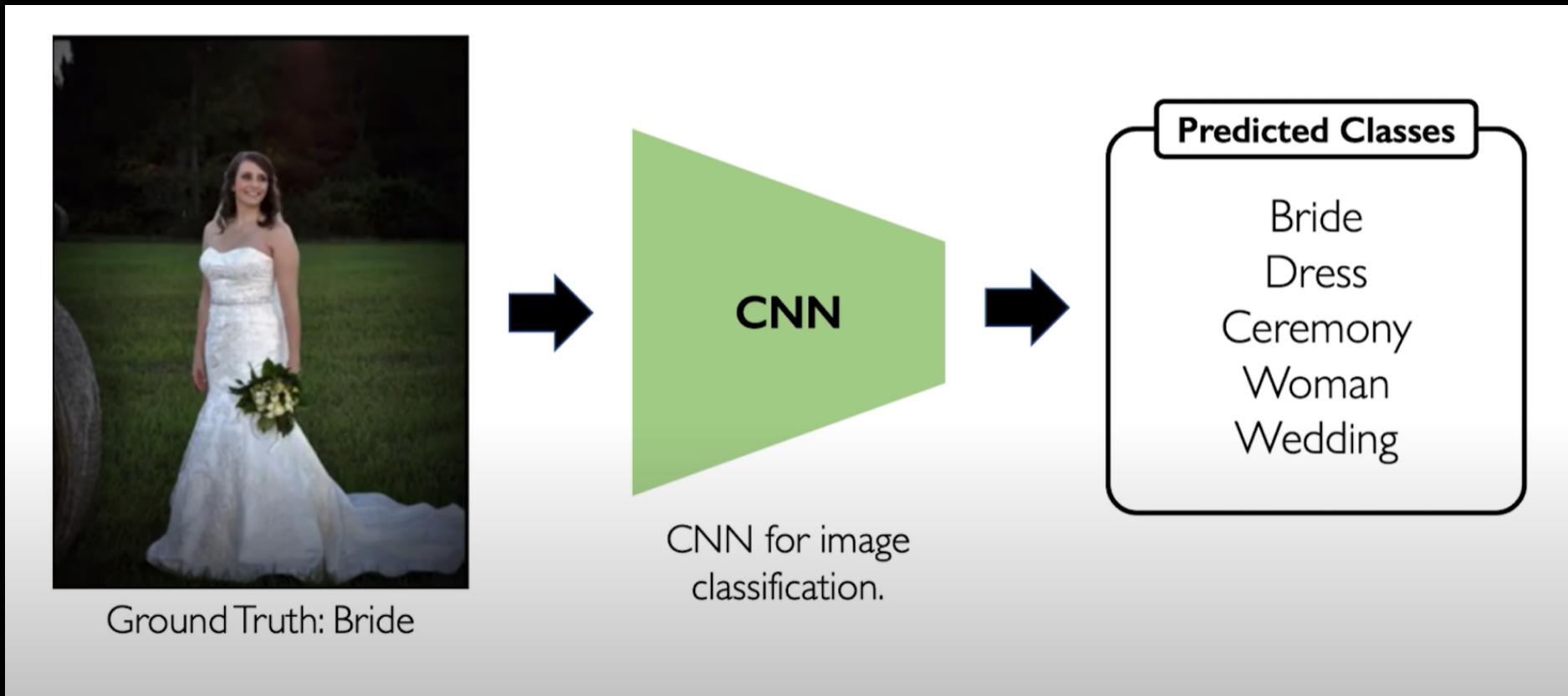
English Spanish French Turkish - detected Translate

O bir hemşire
O bir doktor

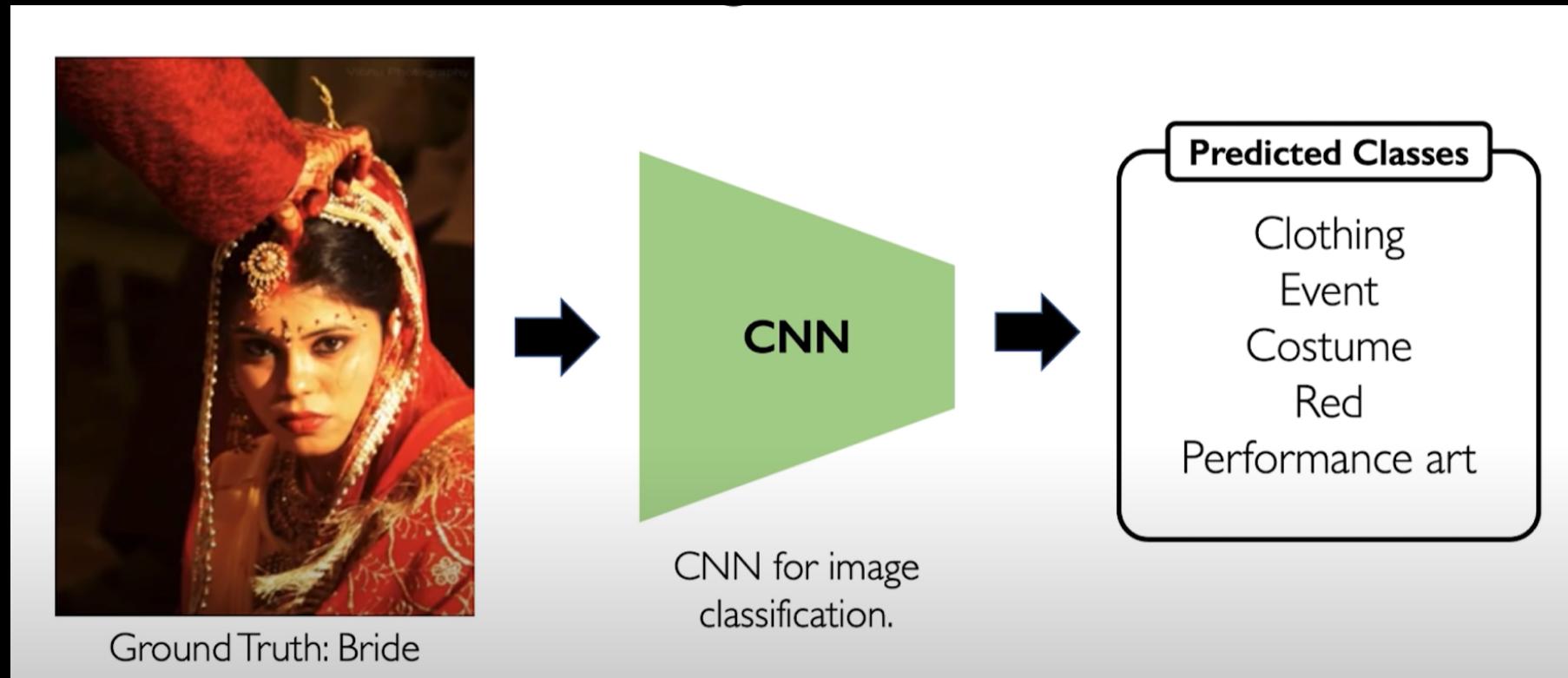
She is a nurse
He is a doctor

Turkish English Spanish Turn off instant translation Suggest an edit

Bias in Image Classification



Bias in Image Classification





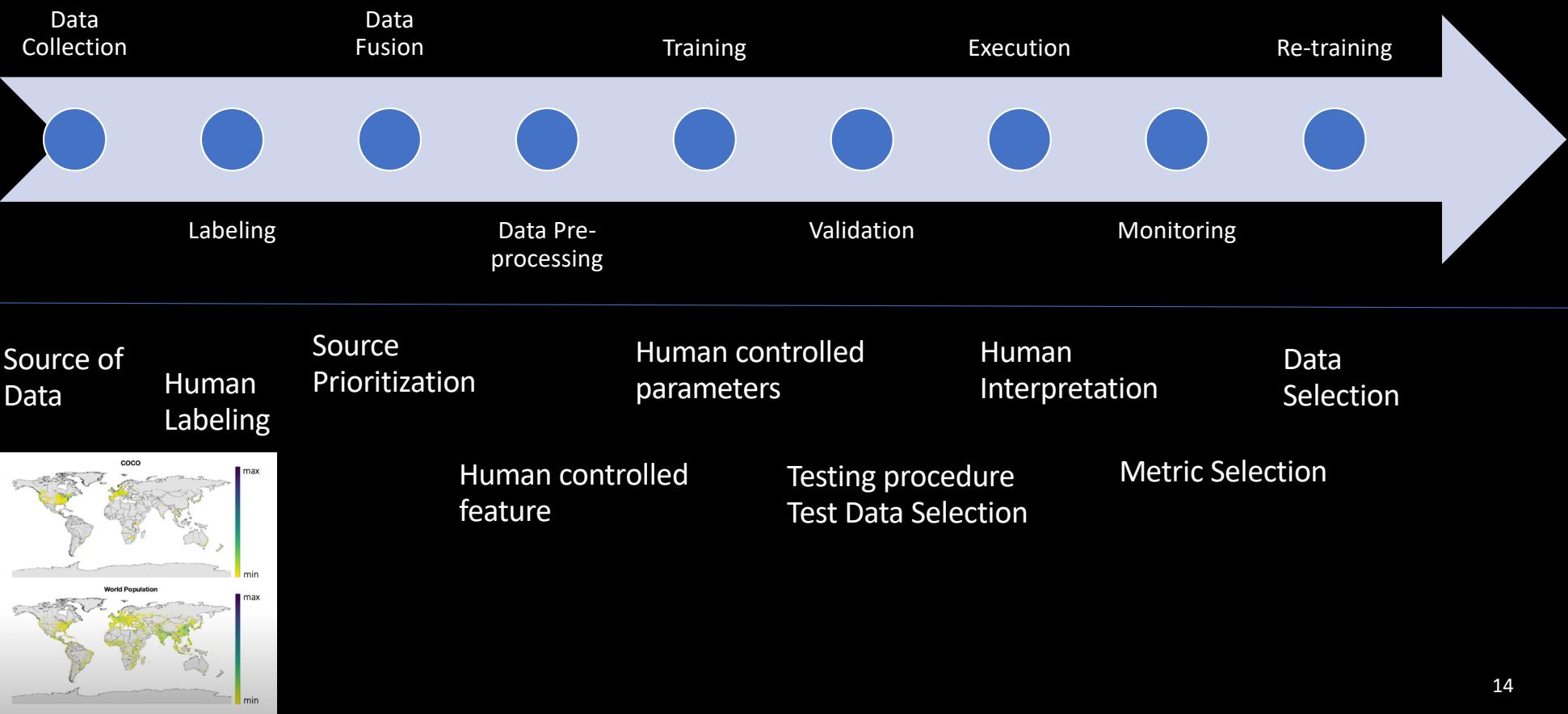
Recognition



Under-representation

Allocative harms are immediate, easily quantifiable, discrete while representational harms are long term, difficult to formalize and diffuse

Source of Bias in AI Lifecycle



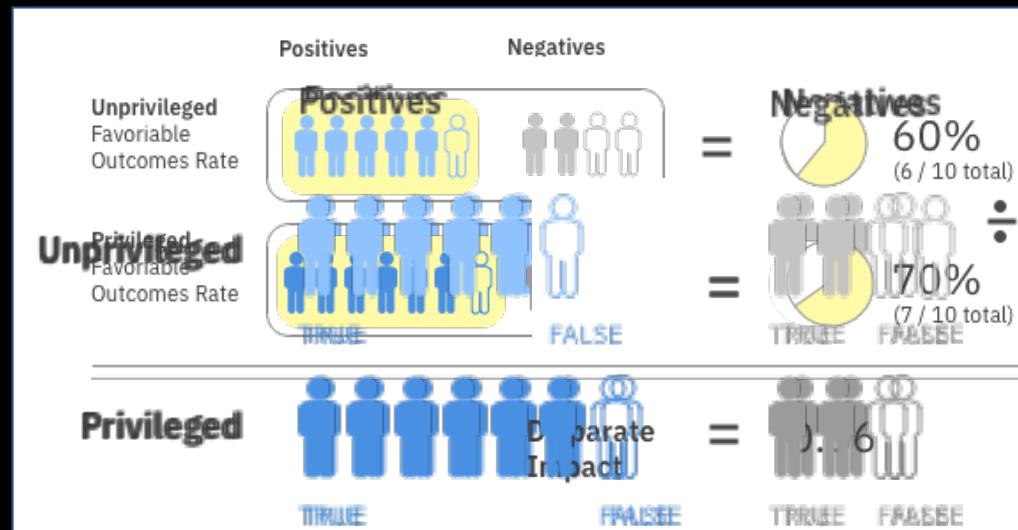
Fairness – Types

Group Fairness

- Groups defined by protected attributes receiving similar treatments or outcomes

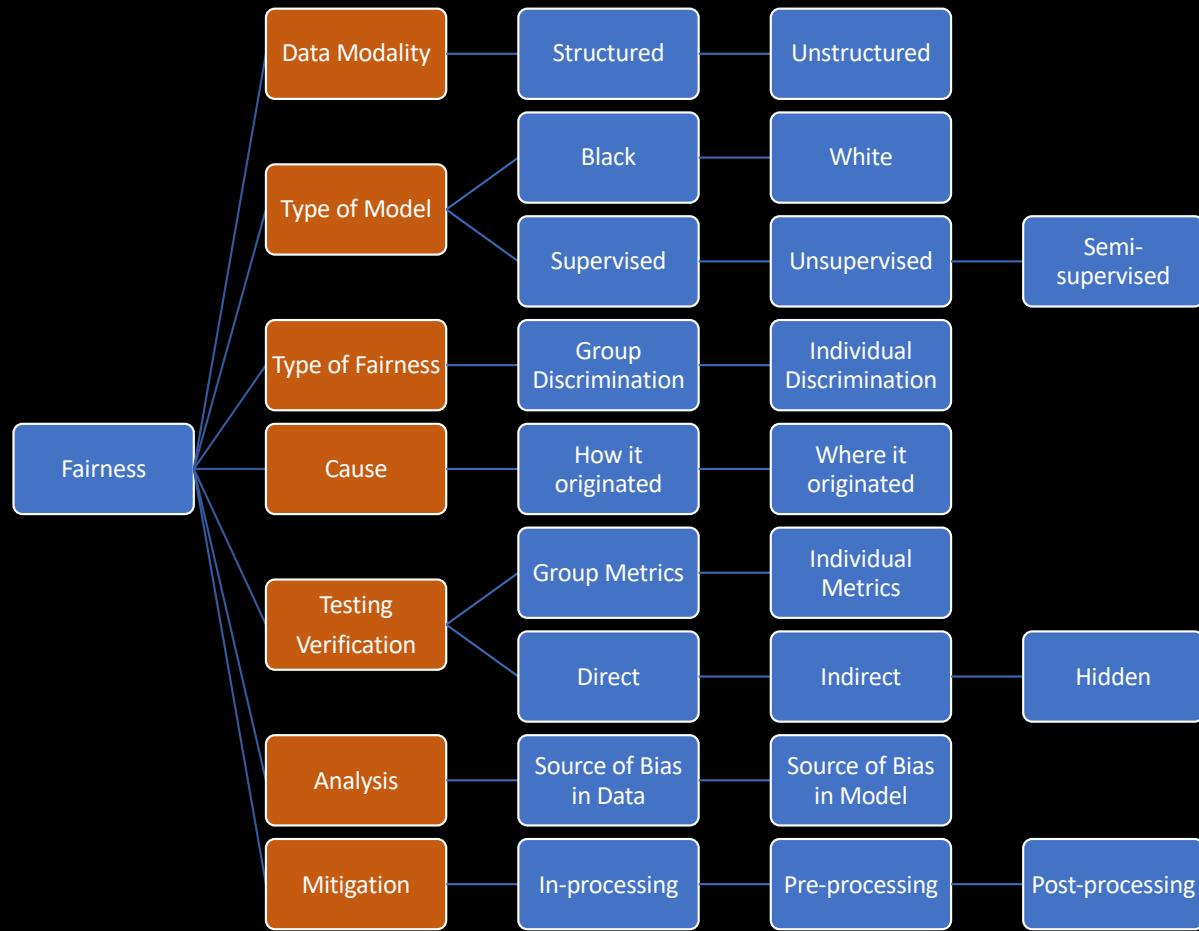
Individual Fairness

- Similar individuals receive similar treatments or outcomes



Gender		Creditability		
1	Male	10	100	Y
1	Female	10	100.2	N

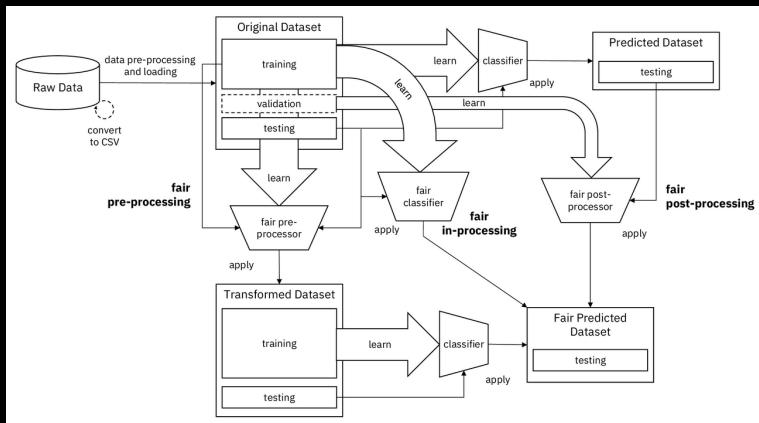
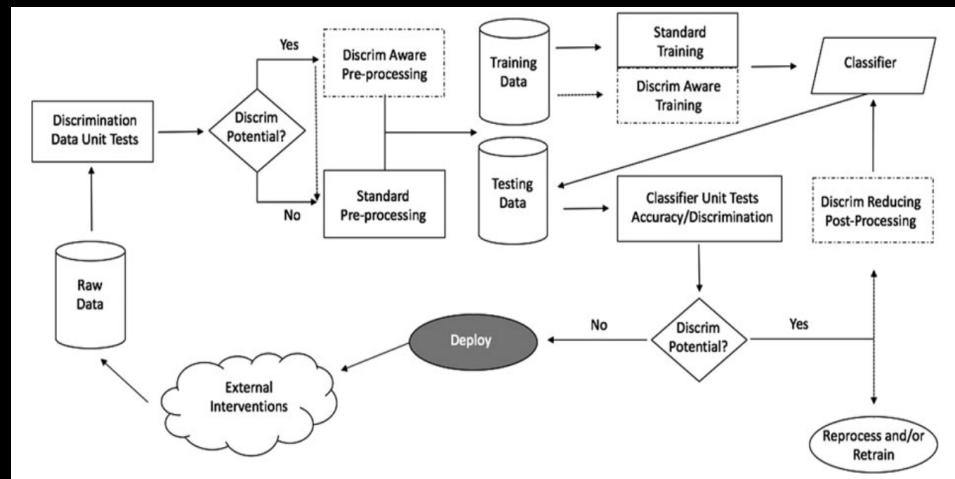
Fairness Dimensions



IBM in Fairness

IBM Watson OpenScale

Group Discrimination Detection
Source of Bias
Individual Discrimination Detection
Post-Processing based Mitigation
Hidden and Indirect Bias



AI Fairness 360

<http://aif360.mybluemix.net>

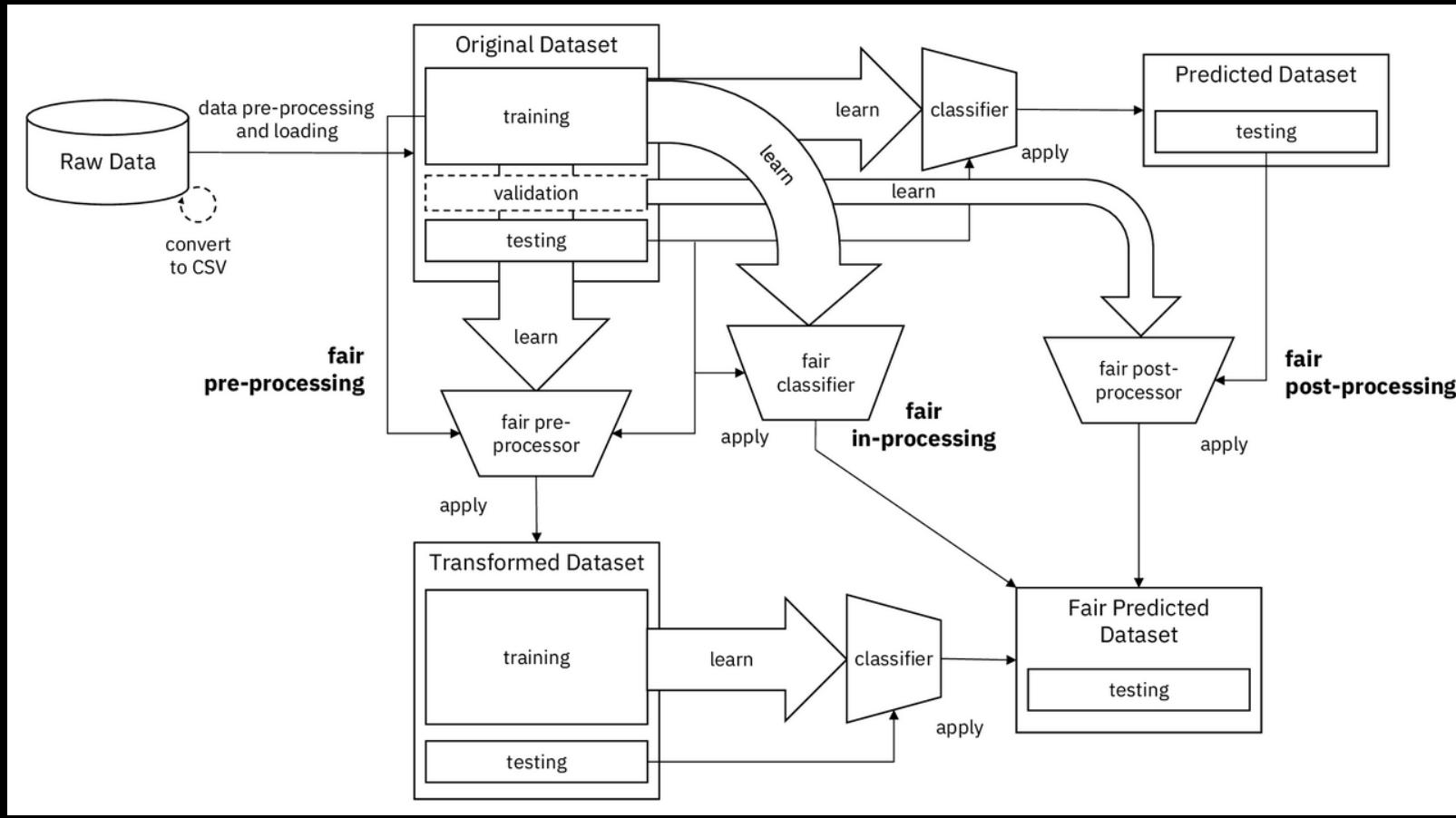
<https://github.com/ibm/aif360>

<https://pypi.org/project/aif360>

AIF360

<https://github.com/Trusted-AI/AIF360>

<https://aif360.mybluemix.net/>



Metrics and Algorithms in AIF360

Group and Individual Fairness Metrics

- Group
 - WAE
 - Disparate Impact
 - Statistical Parity Difference
 - WYSWYG
 - Equality of Odds
 - Average odds difference
 - Average absolute odds difference
- In-between
 - false_negative_rate_ratio
 - false_positive_rate_ratio
 - error_rate_ratio
 - and error_rate_difference
- Individual
 - Distortion metrics
- Both
 - Theil Index: Measures the inequality of benefit allocation

Mitigation Algorithms

- Pre-Processing
 - Re-weighing (Kamiran and Calders, KIS'12)
 - Disparate impact remover (Feldman et al. KDD'15)
 - Optimized pre-processing (Calmon et al. NIPS'17)
 - LFR (Zemel et al. ICML'13)
- In-Processing
 - Prejudice remover (Kamishima et al. ECML-PKDD'12)
 - Adversarial debiasing (Zhang et al. AIES'18)
 - Meta algorithm for Fair Classification (Celis et al. FAT*19)
- Post-Processing
 - Equalized Odds Postprocessing (Hardt et al. NIPS'16)
 - Calibrated Equalized Odds PostProcessing (Pleiss et al. NIPS'17)
 - Reject Option Classification (Kamiran et al. ICDM'12)

Notations

Term	Explanation
$P(A B)$	Conditional probability of A given B
X	All attributes which is used for prediction
Y	Ground truth ($Y=1$: favorable outcome; $Y=0$: unfavorable outcome)
Z	Binary protected attribute ($Z=1$: privileged; $Z=0$: unprivileged)
\hat{Y}	Predicated outcome; $\hat{Y} = 1 \leftrightarrow h(X) > \sigma$
$P(\hat{Y} = 1 Z = 1)$	Probability of favorable outcome for the privileged group
$Z \notin X$	Fairness thru Unawareness

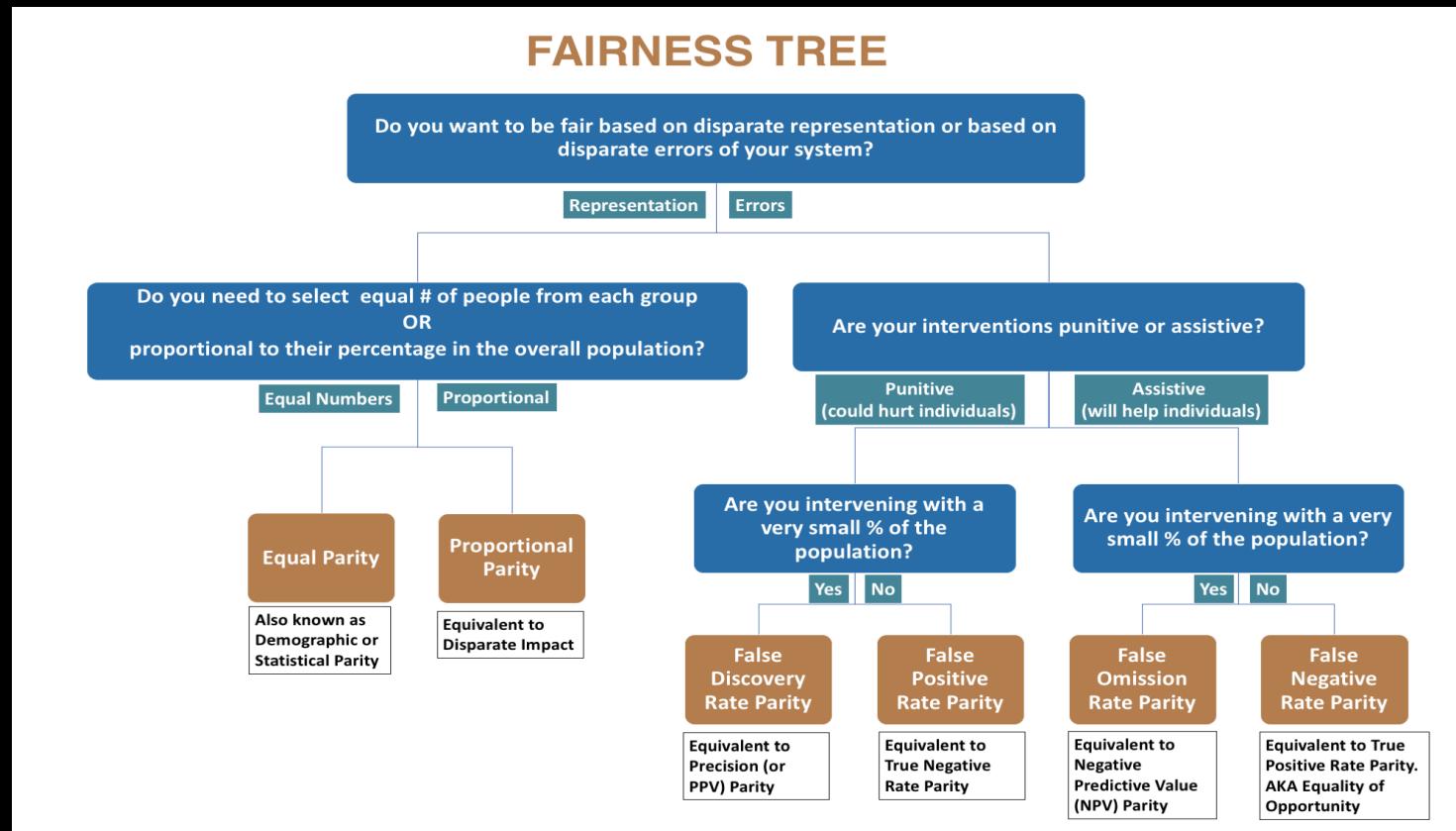
Metrics - Recap

Fairness Metric	Equality	Expression
Disparate Impact	AR	$P(\hat{Y} = 1 \mid Z = 1) / P(\hat{Y} = 1 \mid Z = 0)$
Statistical parity difference	AR	$P(\hat{Y} = 1 \mid Z = 1) - P(\hat{Y} = 1 \mid Z = 0)$
Predictive Parity	PREC	$P(Y = 1 \mid \hat{Y} = 1, Z = 1) = P(Y = 1 \mid \hat{Y} = 1, Z = 0)$
Equality of Opportunity	TPR	$P(\hat{Y} = 1 \mid Y=1, Z = 1) = P(\hat{Y} = 1 \mid Y=1, Z = 0)$
Equalized Odds	TPR, FPR	$P(\hat{Y} = 1 \mid Y=i, Z = 1) = P(\hat{Y} = 1 \mid Y=i, Z = 0) \quad i:\{0,1\}$
Accuracy Equality	AC	$P(\hat{Y} = Y \mid Z = 1) = P(\hat{Y} = Y \mid Z = 0)$
Treatment Equality	FPR/FNR	$FPR/FNR \mid z = 0 = FPR/FNR \mid z = 1$

Why do we need so many metrics?

- Problem with Disparate Impact/Demographic Parity
 - We can accept qualified applicants in the demographic $Z = 0$, but unqualified individuals in $Z = 1$, so long as the percentages of acceptance match

Metric Selection (Aequitas)



Pre-Processing Mitigation Algorithm

- Training data change
 - What can you change
 - Sample distribution
 - Weights
 - Feature values
 - Class labels
 - Mitigation Effect
 - Group Fairness
 - Individual Fairness
- Algorithms
 - Re-weighting (Kamiran and Calders, KIS'12)
 - Modifies the weights of different training examples
 - Disparate impact remover (Feldman et al. KDD'15)
 - Edit feature values to improve group fairness
 - Optimized pre-processing (Calmon et al. NIPS'17)
 - Modifies features and labels to address both group and individual fairness
 - LFR (Zemel et al. ICML'13)
 - Learns fair representation by obfuscating information about protected attribute

Re-weighting

$$DI = \frac{P(Y=1 | Z=0)}{P(Y=1 | Z=1)} > 0.8$$

$|\{X | Y=1, Z=0\}| = a$
 $|\{X | Y=0, Z=0\}| = b$
 $|\{X | Y=1, Z=1\}| = c$
 $|\{X | Y=0, Z=1\}| = d$

$$W = \frac{|\{X | Z=z\}| \times |\{X | Y=y\}|}{|D| |\{X | Z=z \wedge Y=y\}|}$$

$|\{X | Y=1, Z=0\}| = (a+b)(a+c)$
 $|\{X | Y=0, Z=0\}| = (a+b)(b+d)$
 $|\{X | Y=1, Z=1\}| = (c+d)(a+c)$
 $|\{X | Y=0, Z=1\}| = (c+d)(b+d)$

$$DI = 1$$

Optimized Pre-Processing

(Calmon et al. NIPS'17)

Learns a probabilistic transformation that can modify the features and the labels in the training data

$$\min_{p_{\tilde{X}, \tilde{Y}|X,Y,Z}} \Delta(p_{\tilde{X}, \tilde{Y}}, p_{X,Y})$$

s.t. $D(p_{\tilde{Y}|Z}(y|z), p_{Y|T}(y)) \leq \epsilon_{y,z}$ and

$$E(\delta((x,y), (\tilde{X}, \tilde{Y})) | Z = z, X = x, Y = y) \leq c_{z,x,y} \forall (x,y,z) \in \mathcal{D},$$

$p_{\tilde{X}, \tilde{Y}|X,Y,Z}$ is a valid distribution

Utility Preservation

Limit the dependence of the transformed outcome \tilde{Y} on the protected variable Z .

Individual distortion control

D = Distance metric

δ = distortion metric

Δ = dissimilarity measure between probability distribution

German-credit	Sex	Age
Acc – before	0.65	0.65
Acc - after	0.56	0.60
DI - before	.99	0.38
DI - after	1.06	0.82

Recap and Outline of Today's Lecture – 23 Sep

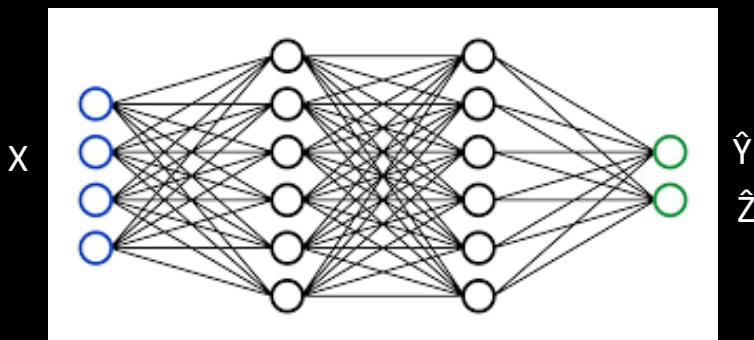
- Recap – 21 Sep
 - Pillars of Trust –IBM perspective
 - Fairness
 - Motivation
 - Human bias, algorithmic bias
(Text, Image)
 - Fairness Dimensions
 - Metrics
 - Selection of Metrics
 - Mitigation Algorithms for Group Fairness
 - Re-weighting
 - Optimized pre-processing
- Today – 23 Sep
 - Fairness [5 mins]
 - Mitigation Algorithms for Group Fairness
 - In-processing algorithms
 - Post-processing algorithms
 - Testing for Bias
 - Hidden Bias
 - Explainability 10 mins]
 - AI Testing Demo [15 mins]
 - Introduction to Toolkits [30 mins]
 - AIF360
 - AIX360
 - Research Direction, Conclude [15 mins]

In-Processing Mitigation Algorithm

- Prejudice remover (Kamishima et al. ECML-PKDD'12)
 - Include a regularization term to reduce bias
- Adversarial debiasing (Zhang et al. AIES'18)
 - Two network (Similar to GAN)
- Meta algorithm for Fair Classification (Celis et al. FAT*19)
 - A meta-algorithm for classification that takes as input a large class of fairness constraints

Adversarial De-biasing

(zhang et al. AIES 2018)



Target Class label Y
Protected Output Z

If \hat{Z} can determine \hat{Y} then protected attribute has affect on \hat{Y}

Task: language model to complete analogies
He is to **she**, as **doctor** is to ?

biased		debiased	
neighbor	similarity	neighbor	similarity
nurse	1.0121	nurse	0.7056
nanny	0.9035	obstetrician	0.6861
fiancée	0.8700	pediatrician	0.6447
maid	0.8674	dentist	0.6367
fiancé	0.8617	surgeon	0.6303
mother	0.8612	physician	0.6254
fiance	0.8611	cardiologist	0.6088
dentist	0.8569	pharmacist	0.6081
woman	0.8564	hospital	0.5969

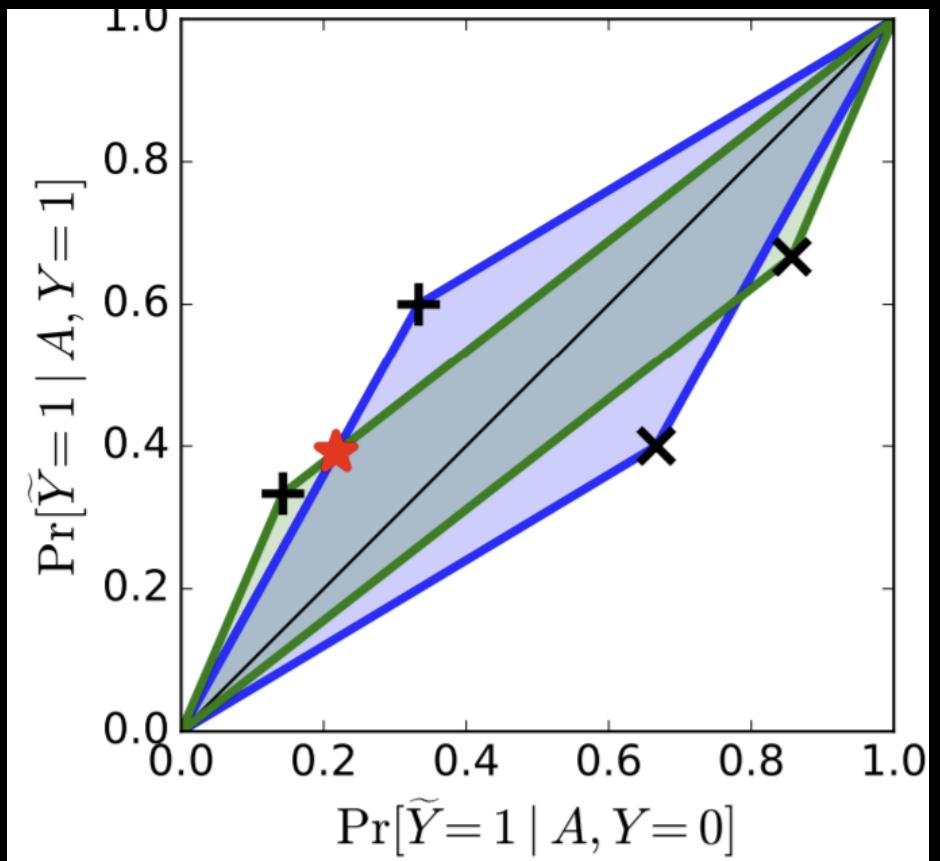
Sensitive attribute: Gender

Post-Processing Mitigation Algorithms

- Equalized Odds Post-processing (Hardt et al. NIPS'16)
 - Solves a linear program to find probabilities with which to change output labels to optimize equalized odds
- Calibrated Equalized Odds Post-processing (Pleiss et al. NIPS'17)
 - Optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective
- Reject Option Classification (Kamiran et al. ICDM'12)
 - Gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty

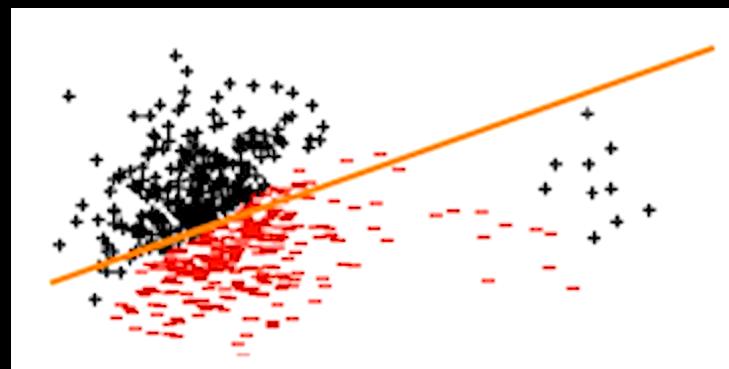
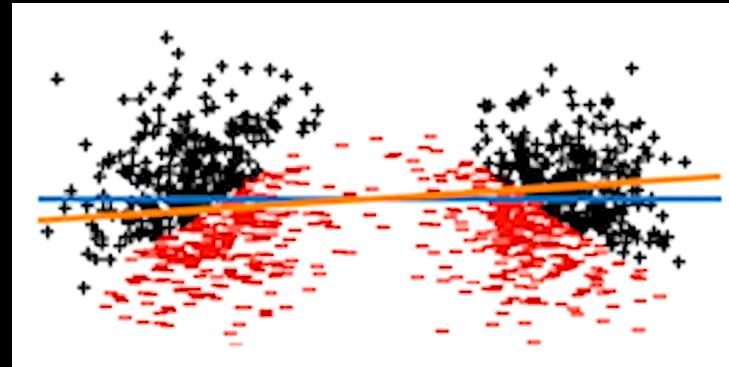
Equalized Odds Post-processing

- Equalized Odds [Recap]
 - True positive Rate and False Positive Rate should be same for protected and unprotected groups
- Goal: Unfair predictor \hat{Y} to fair predictor \tilde{Y} based on Y, Z, \hat{Y}
- $\gamma_z(\hat{Y}) = (P(\hat{Y} = 1 | Z=z, Y=0), P(\hat{Y} = 1 | Z=z, Y=1))$
- Eq. Odds: $\gamma_0(\hat{Y}) = \gamma_1(\hat{Y})$



Hidden Bias

- Fairness without Demographics
 - Most of the literature assumes that the protected attribute is part of the dataset
 - Privacy and Regulations often precludes collection of protected attributes
 - How can we train (or retain) an ML model to improve fairness when we do not even know the protected group membership
 - Solution
 - Repeated Loss Minimization (Hashimoto et al. ICML'18)
 - Higher accuracy loss in a group results in lower retention
 - Inequalities are permissible when they maximize the long - term expectations of the least fortunate group. John Rawls, 1971
 - Minimize the loss suffered by least fortunate group over time
 - Problem: small groups have low representation in avg. loss
 - Adversarially Re-weighted learning (Lahoti et al. Neurips'20)
- Lack of obvious protected groups in data
 - Find all attributes or attribute combinations for which my model is biased



* <https://vimeo.com/295743125>

Individual Discrimination Definition

Dwork's Definition

- Two similar individuals should get same decision

$$D(h(x_i), h(x_j)) \leq d(x_i, x_j)$$

forall x_i, x_j

Counterfactual Fairness:

- A decision is fair towards an individual if it is the same in (a) the actual world and in a (b) a counterfactual world where the individual belonged to a different demographic group

Individual Discrimination Testing of Black Box Models

	Gender	Creditability	
1	Male	10	100
1	Female	10	100

Given one sample (t) we can iteratively perturb it with values from the privileged/unprivileged class and generate test samples (t, t').

	Race	Creditability	
1	White	10	100
1	Black	10	100
1	Hispanic	10	100

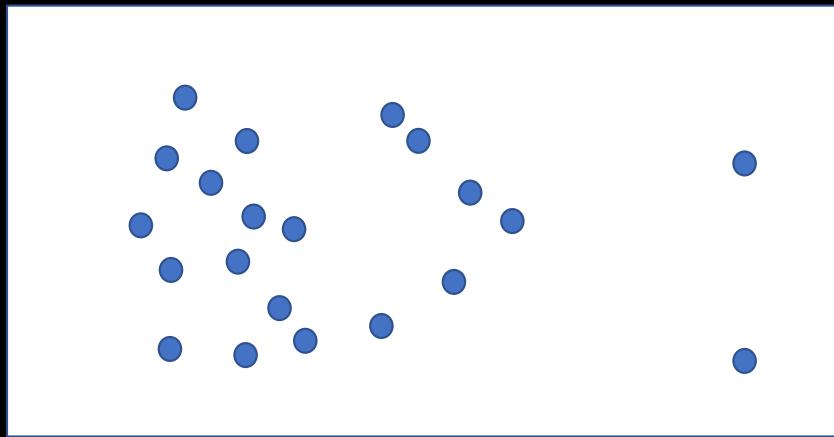
If $M(t) \neq M(t')$ then t is called a discriminatory sample

	Age (Priv: <60)	Creditability	
1	61	10	100
1	20	10	100
1	58	10	100

Goal: Synthesize non-protected values for which we can find discriminatory sample

Themis

- Sample generation
 - Random
- Metric
 - $\frac{\# \text{ Discriminatory samples}}{\# \text{ samples}}$

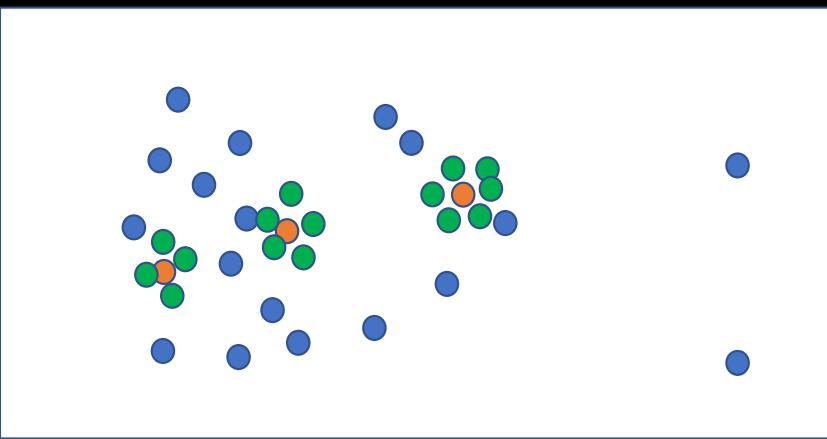


Key Result: 1000s of Discriminatory Input

AEQUITAS

- Sample generation
 - Two Steps
 - Global: Random
 - Local:
 - For each discriminatory Input
 - Random small perturbation to non-protected attributes
- Metric
 - $\# \text{ Discriminatory samples} / \# \text{ gen. samples}$

- Global non-discriminatory samples
- Global discriminatory samples
- Local samples

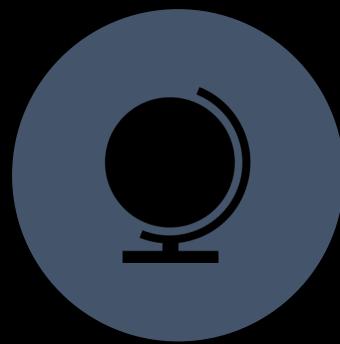


Key Result: > 70% Discriminatory Inputs

Explainability



LOCAL



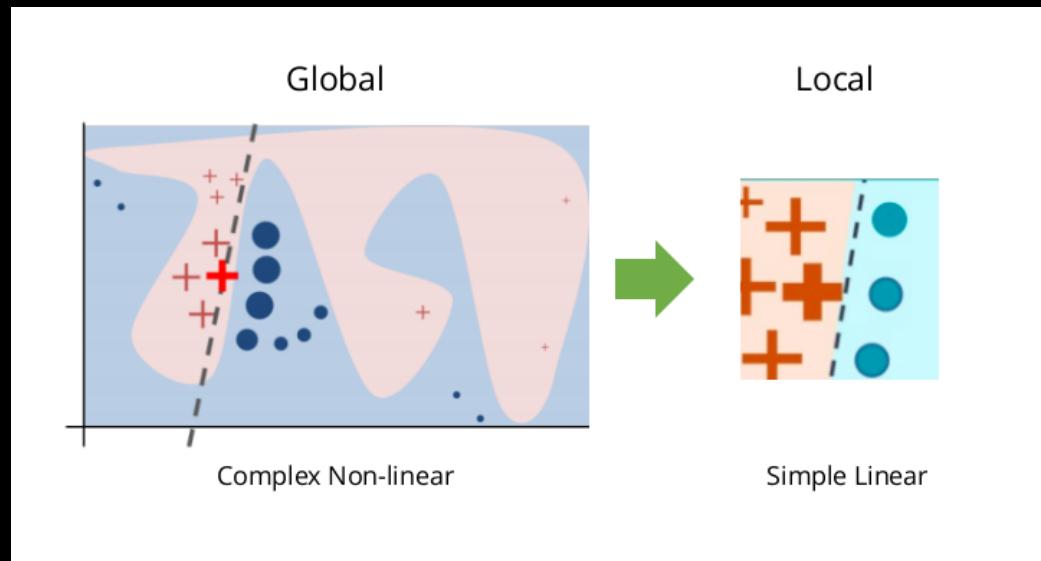
GLOBAL



COUNTERFACTUAL

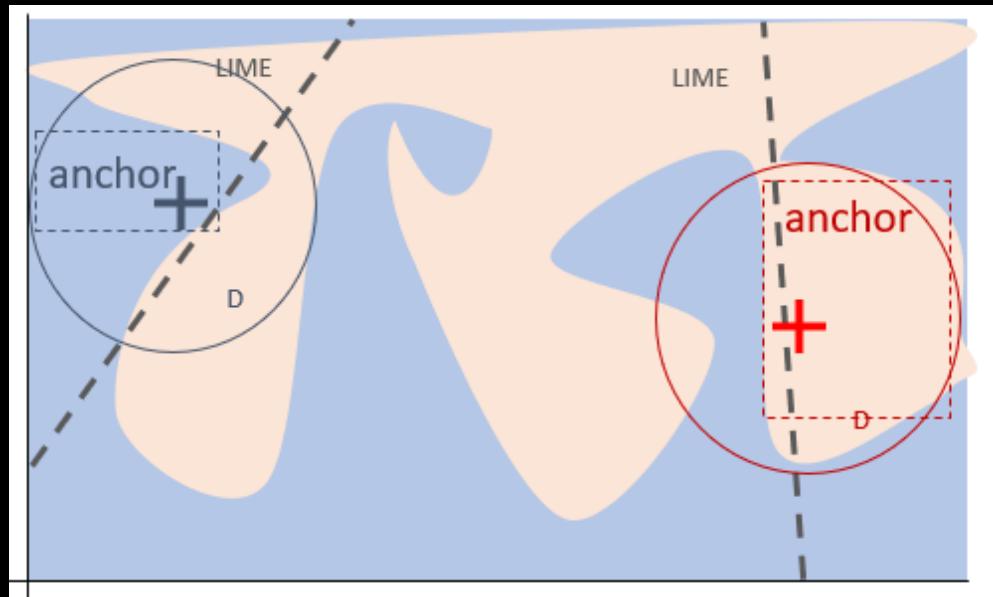
LIME

- Permute data
- Distance between permutations and original observation
- Make predictions on new data using complex model
- Pick small number of best features
- Fit a simple linear model
- Feature weights from the simpler model



ANCHOR

- ANCHOR computes the region where the same prediction holds
- Easier to comprehend
- Beam search technique in the space of candidate rules till it satisfies precision and coverage criteria

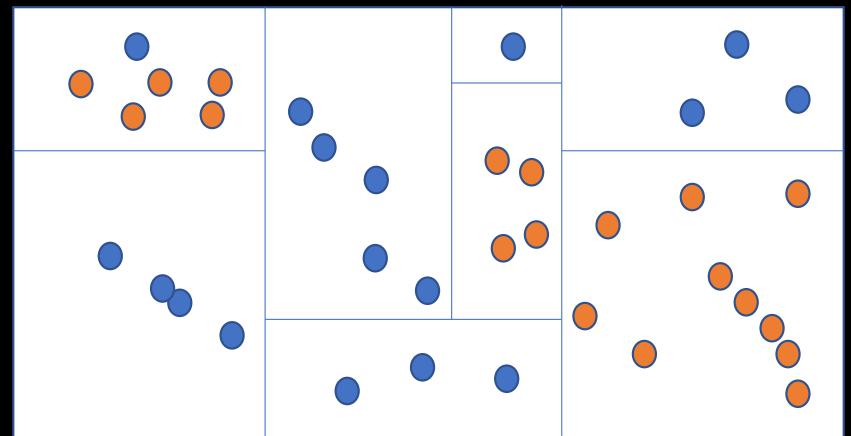


Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." AAAI 2018.

Image Source https://drive.google.com/file/d/1QdBRUCn_yTUIJaTyWghcp9XJZ4K2kY9/view

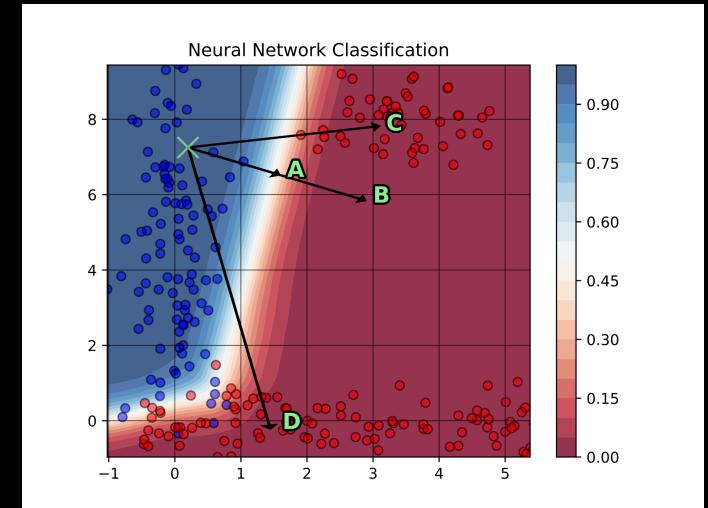
Global Explainability thru TrePan

- Approximate black-box model's decision boundary by decision tree
- Model Agnostic path coverage TrePan
 - Modified decision tree algorithm
 - Uses target model's decision (instead of training data's decision)
 - Decision tree algorithm forms boundary based on fewer number of samples near the leaf levels
 - TrePan generates more synthetic data to increase density to create more accurate boundary



Counterfactual Explainability

- A counterfactual explanation of a prediction describes the smallest change to the feature values that changes the prediction to a predefined output



Wachter et al. suggest minimizing the following loss:

$$L(x, x', y', \lambda) = \lambda \cdot (\hat{f}(x') - y')^2 + d(x, x')$$

Open-Source Toolkits

AIF 360: <https://aif360.mybluemix.net/>

AIX 360: <https://aix360.mybluemix.net/>

Research Direction

- Fairness
 - Study in new application areas
 - Algorithms
 - Long term effect of fairness
- Explainability
 - User-centric
- Testing
 - Test case generation
- Debugging
 - Fault localization
- Automated Repairing
- Verification
- SE4AI and AI4SE