Bias and Fairness

ICS 491

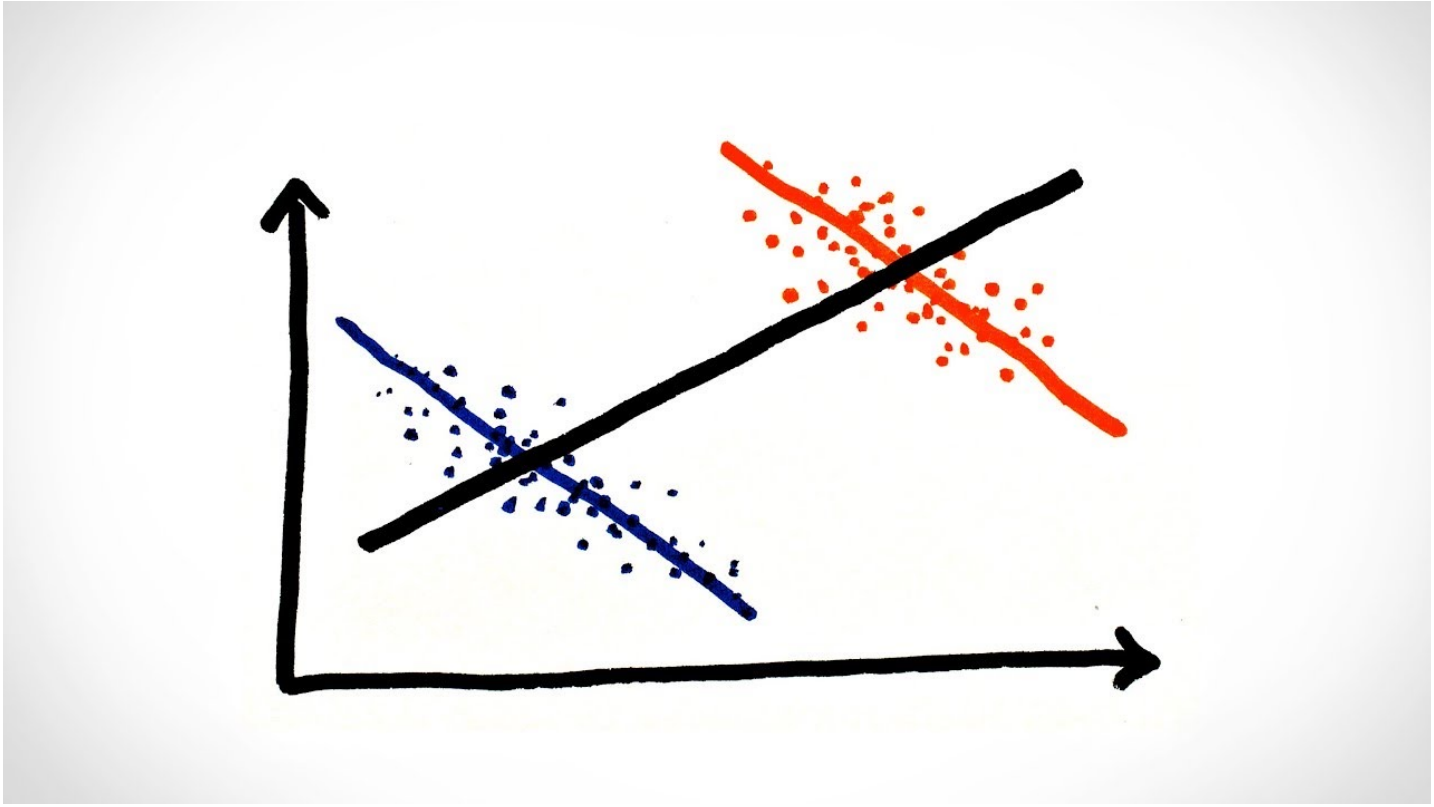
Where can bias occur in the ML process?

- **Data**: imbalances of class labels, features, input structure
- **Model**: lack of unified uncertainty, interpretability, and performance metrics
- **Training**: feedback loops that perpetuate bias
- **Evaluation**: lack of analysis with respect to subgroups
- **Interpretation**: human biases distort the interpretation of results

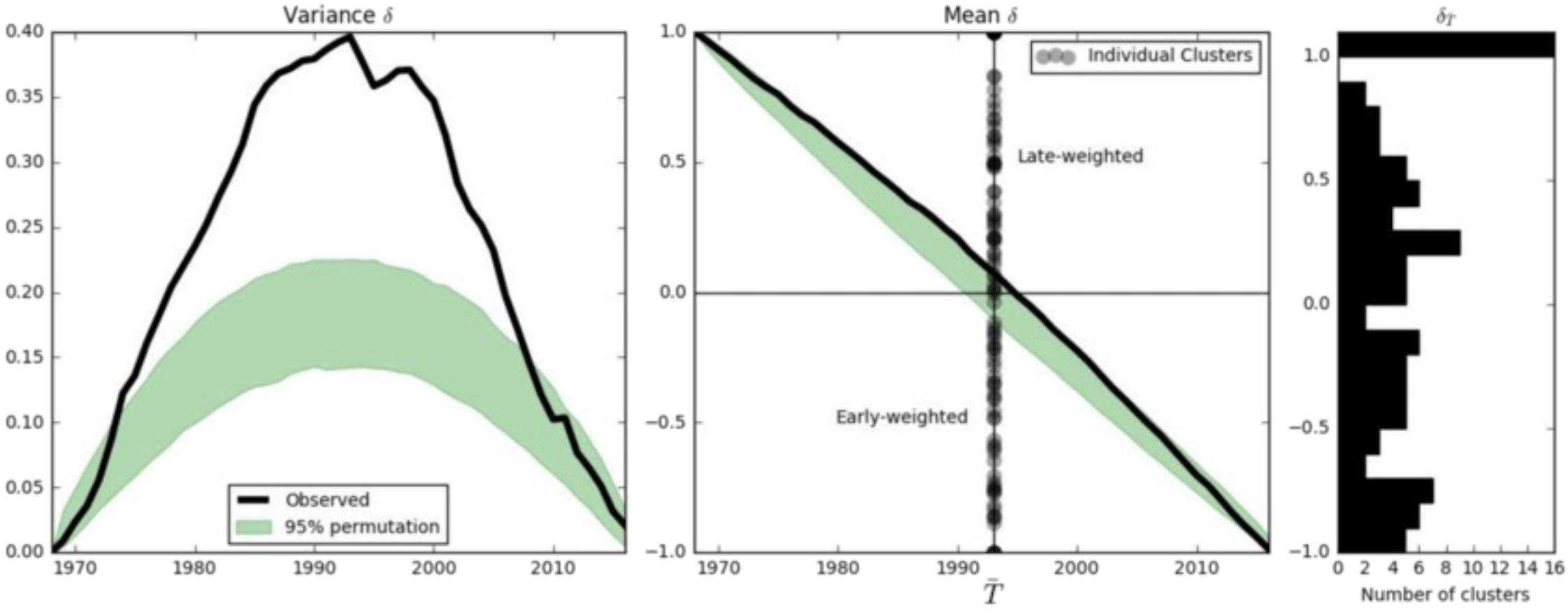
Some sources of bias

- Measurement Bias – biases in measurement techniques
- Omitted Variable Bias – important variables left out of the model
- (Population) Representation Bias / Sampling Bias – issues with sampling the study population
- Self-selection Bias – participants are self-selected
- Population Bias – recruited population not representative of true population
- …

Aggregation bias – Simpson's paradox



Temporal bias



Definitions/Metrics of bias/Fairness

- In my opinion, it is easier to grasp/understand these concepts from the probabilistic definitions than the English description

Equalized odds

$$P(R = + \mid Y = y, A = a) = P(R = + \mid Y = y, A = b) \quad y \in \{+, -\} \quad \forall a, b \in A$$

Equal opportunity

$$P(R = - \mid Y = +, A = a) = P(R = - \mid Y = +, A = b) \quad \forall a, b \in A$$

Demographic parity

$$P(R = + \mid A = a) = P(R = + \mid A = b) \quad \forall a, b \in A$$

Test fairness

$$P(Y = + \mid S = s, A = a) = P(Y = + \mid S = s, A = b) \quad \forall s \in S \quad \forall a, b \in A$$

Fairness through awareness

Similar people get similar predictions

Fairness through unawareness

Protected attributes are not used in the decision-making process

…And many other definitions based on many other ML evaluation metrics

Sources: [3][4][5][6][7][8][9][10][11] view · talk · edit

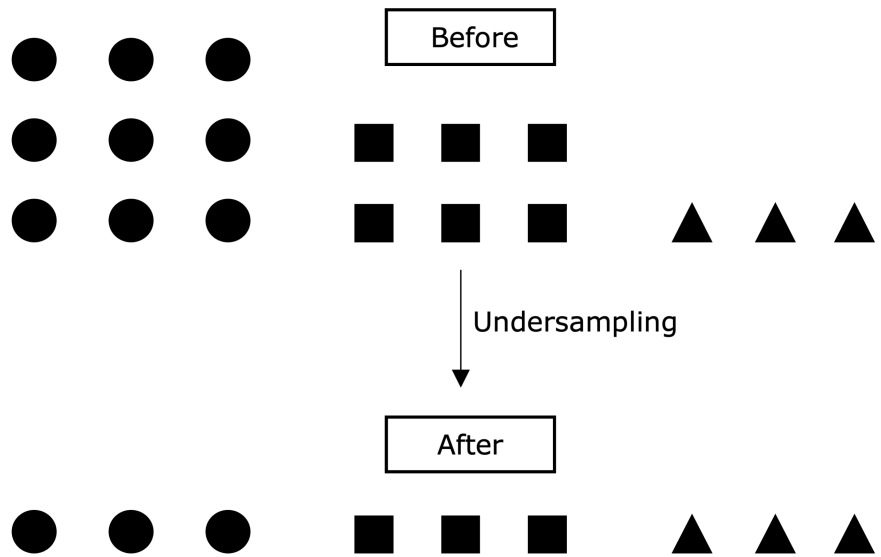
		Predicted condition			
	Total population = P + N	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) = TPR + TNR − 1	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** **hit**	**False negative (FN),** **type II error, miss, underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** **type I error, false alarm, overestimation**	**True negative (TN),** **correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) = PPV + NPV − 1	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{TP}{TP + FN + FP}$

Wikipedia

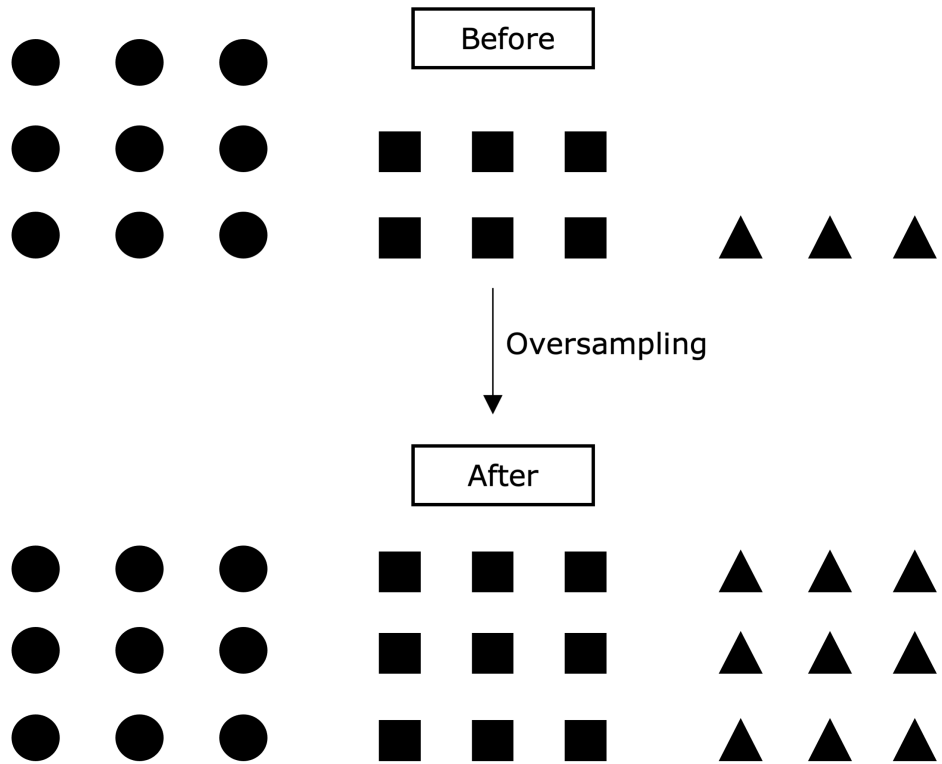
General approaches to achieving fairness

- Data modification (pre-processing)
  - Upsampling of underrepresented attributes
  - Data augmentation using GANs
- Algorithmic approaches (“in-processing”)
  - Encoding fairness in the loss function
- Post-processing
  - Re-assign labels to ensure fairness

Undersampling



Oversampling



Fair regression

Fairness “regularization” terms (“penalties”) added to the traditional MSE loss function:

$$f_1(w, S) = \frac{1}{n_1 n_2} \sum_{\substack{(x_i, y_i) \in S_1 \\ (x_j, y_j) \in S_2}} d(y_i, y_j)(w.x_i - w.x_j)^2.$$

Individual fairness

$$f_2(w, S) = \left(\frac{1}{n_1 n_2} \sum_{\substack{(x_i, y_i) \in S_1 \\ (x_j, y_j) \in S_2}} d(y_i, y_j)(w.x_i - w.x_j) \right)^2.$$

Group fairness

Full loss function in fair regression

$$\text{MSE} + \frac{1}{n_1 n_2} \sum_{\substack{(x_i, y_i) \in S_1 \\ (x_j, y_j) \in S_2}} d(y_i, y_j)(w.x_i - w.x_j)^2. + \left(\frac{1}{n_1 n_2} \sum_{\substack{(x_i, y_i) \in S_1 \\ (x_j, y_j) \in S_2}} d(y_i, y_j)(w.x_i - w.x_j) \right)^2$$

Changing the data

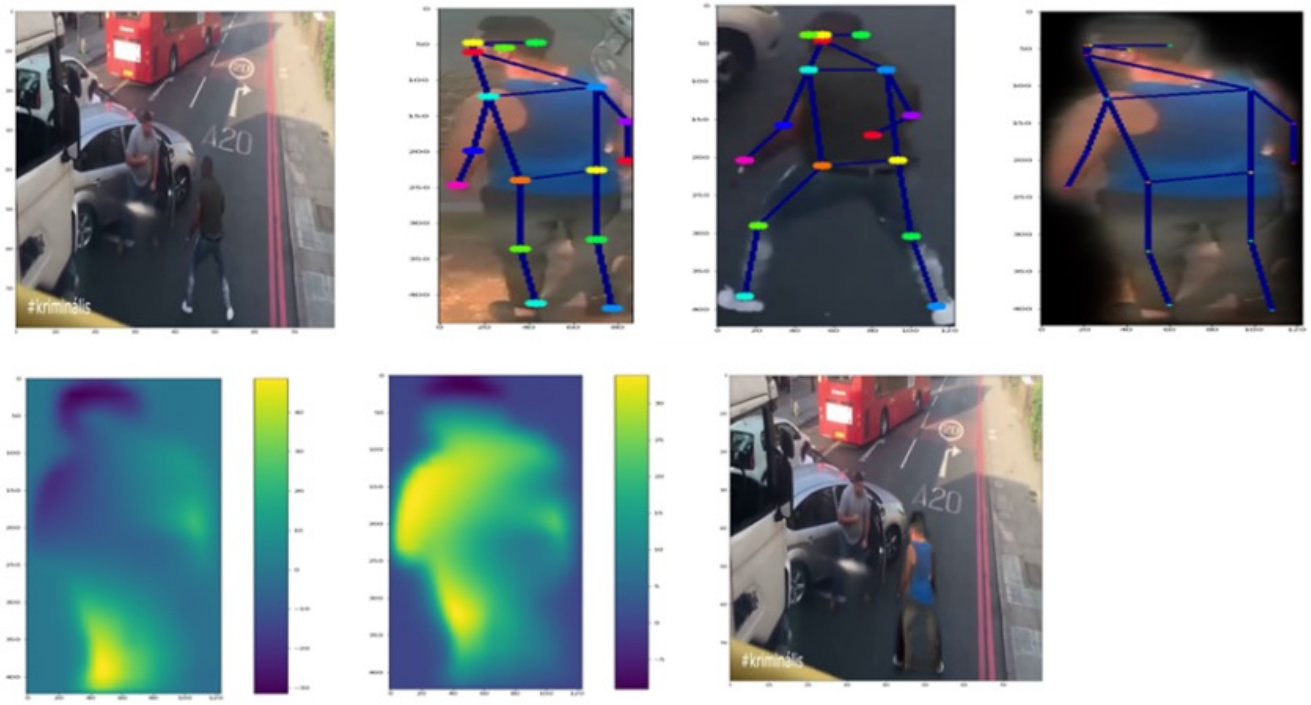
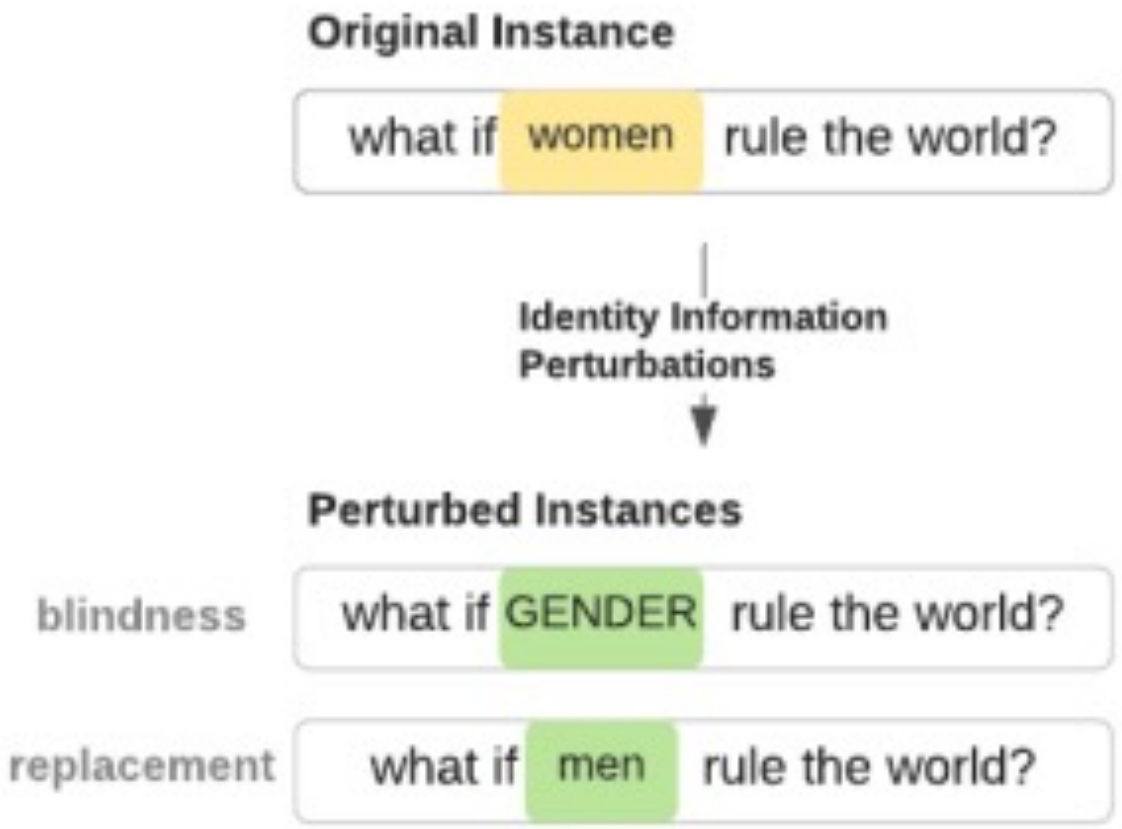


Figure 4: An example of person replacement in a video frame. In the first row, from left to right, Image 1 is the original frame, Image 2 depicts the person who is replacing, Image 3 the person who is getting replaced, Image 4 is the person who is replacing scaled. In the bottom row, from left to right Images 5 and 6 are the displacement fields for x and y and Image 7 is the resulting altered frame.

Pastaltzidis, Ioannis, et al. "Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems." *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.

Changing the data



Wadhwa, Mohit, et al. "Fairness for Text Classification Tasks with Identity Information Data Augmentation Methods." *arXiv preprint arXiv:2203.03541* (2022).

Changing the data

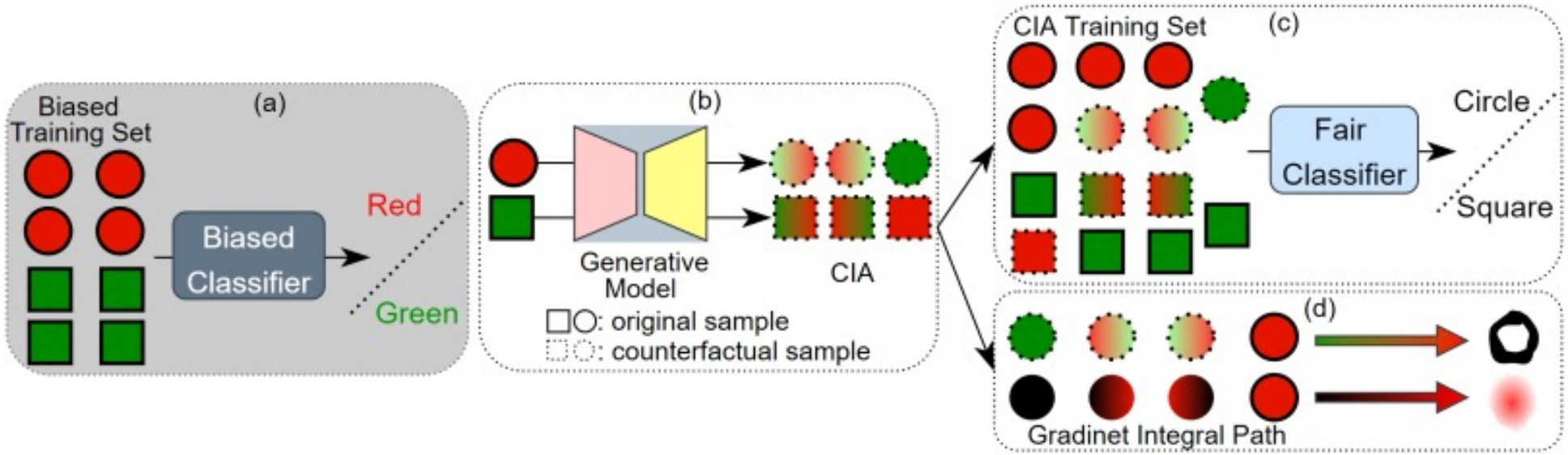


Figure 1: An illustrative example. (a) The target variable (shape) is spuriously correlated with the sensitive attribute (color) in the biased training set. A biased classifier undesirably learns and leverages the spurious correlations for prediction. (b) Our CIA generates bias-tailored counterfactual interpolation augmentation to mitigate bias in the training set and to enhance fair explanation. (c) CIA enables training a fair classifier to learn discriminative features for shape classification. (d) In the first row, CIA generates a meaningful explanation for classifying the target (shape). In the second row, a baseline interpolation generates explanation of the target (shape) confounded by the sensitive attribute (color). Best viewed in color.

Qiang, Yao, et al. "Counterfactual interpolation augmentation (cia): A unified approach to enhance fairness and explainability of dnn." *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, LD Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization*. Vol. 7. 2022.

Fairness GANS

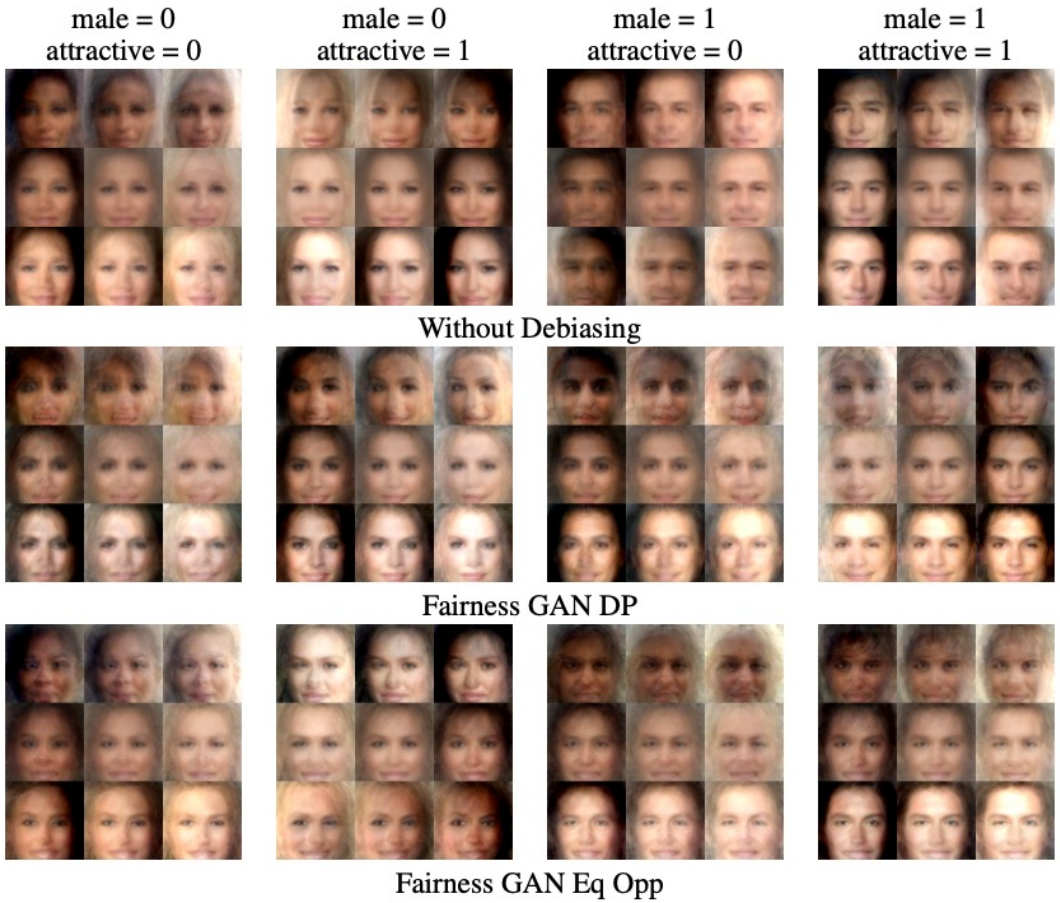
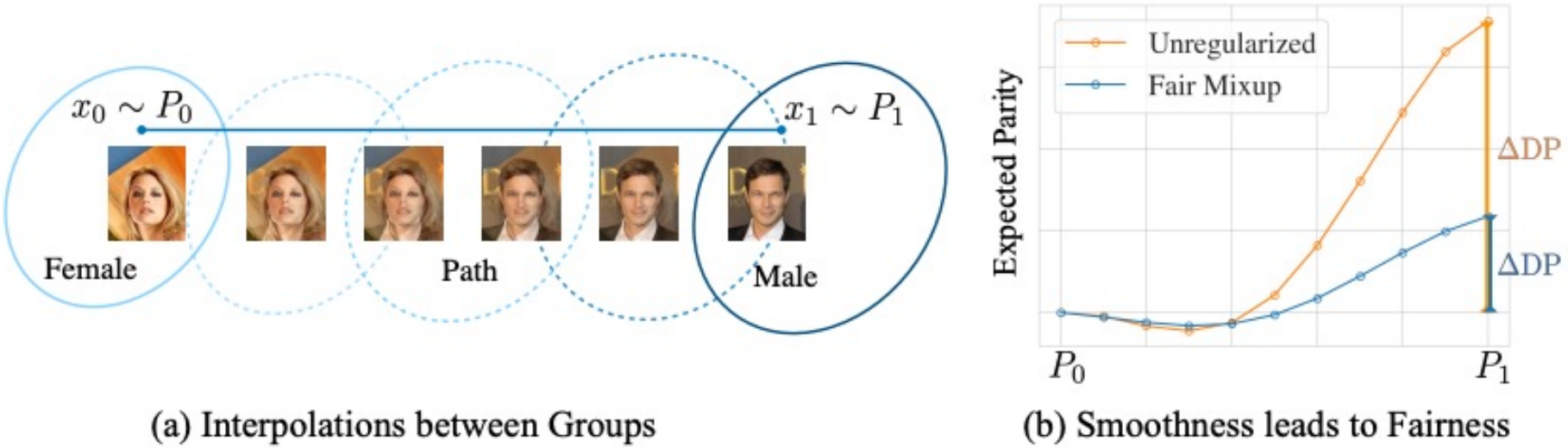


Figure 2: Eigenfaces from the CelebA dataset (male, attractive).

Sattigeri, Prasanna, et al. "Fairness GAN: Generating datasets with fairness properties using a generative adversarial network." *IBM Journal of Research and Development* 63.4/5 (2019): 3-1.

Combining fairness regularization with generative learning



(a) Interpolations between Groups

(b) Smoothness leads to Fairness

Figure 1: (a) Visualization of the path constructed via mixup interpolations between groups that have distribution P_0 and P_1, respectively. (b) Fair mixup penalizes the changes in model's expected prediction with respect to the interpolated distributions. The regularized model (blue curve) has smaller slopes comparing to the unregularized one (orange curve) along the path from P_0 to P_1, which eventually leads to smaller demographic parity ΔDP.

Chuang, Ching-Yao, and Youssef Mroueh. "Fair mixup: Fairness via interpolation." *arXiv preprint arXiv:2103.06503* (2021).