# 1 Fairness Metrics (Verma and Rubin 2018)

### Independence $\hat{Y} \perp A$

- Statistical Parity/Demographic Parity: $P(\hat{Y}=1 \mid A=a) = P(\hat{Y}=1 \mid A=b)$
- Conditional Statistical Parity: $P(\hat{Y}=1 \mid E=e, A=a) = P(\hat{Y}=1 \mid E=e, A=b)$
  *$E$ is a set of legitimate features that may affect the outcome.*

### Separation $\hat{Y} \perp A \mid Y$

- Equalized Odds: $P(\hat{Y}=1|Y=y, A=a) = P(\hat{Y}=1|Y=y, A=b) \forall y \in \{0,1\}$
- Equal Opportunity/ False negative error rate balance: $P(\hat{Y}=0|Y=1, A=a) = P(\hat{Y}=0|Y=1, A=b)$ or $P(\hat{Y}=1|Y=1, A=a) = P(\hat{Y}=1|Y=1, A=b)$
  `mlr3`: `fairness.fnr`, `fairness.tpr`
- Predictive Equality/ False positive error rate balance: $P(\hat{Y}=1|Y=0, A=a) = P(\hat{Y}=1|Y=0, A=b)$ or $P(\hat{Y}=0|Y=0, A=a) = P(\hat{Y}=0|Y=0, A=b)$
  `mlr3`: `fairness.fpr`, `fairness.tnr`
- Treatment Equality: $\frac{\mathrm{FN}}{\mathrm{FP}}\Big|_{A=a} = \frac{\mathrm{FN}}{\mathrm{FP}}\Big|_{A=b}$
- Overall Accuracy Equality: $P(\hat{Y}=Y|A=a) = P(\hat{Y}=Y|A=b)$ `mlr3`: `fairness.acc`

### Sufficiency $Y \perp A \mid \hat{Y}$

- Predictive parity/ outcome test: $P(Y=1|\hat{Y}=1, A=a) = P(Y=1|\hat{Y}=1, A=b)$
  `mlr3`: `fairness.ppv`
- Equal true negative rate: $P(Y=0|\hat{Y}=0, A=a) = P(Y=0|\hat{Y}=0, A=b)$
  `mlr3`: `fairness.npv`
- Equal false omission rate[1]: $P(Y=1|\hat{Y}=0, A=a) = P(Y=1|\hat{Y}=0, A=b)$
  `mlr3`: `fairness.fomr`
- Equal false discovery rate[1]: $P(Y=0|\hat{Y}=1, A=a) = P(Y=0|\hat{Y}=1, A=b)$
- Conditional use accuracy equality: $P(Y=1|\hat{Y}=1, A=a) = P(Y=1|\hat{Y}=1, A=b) \wedge P(Y=0|\hat{Y}=0, A=a) = P(Y=0|\hat{Y}=0, A=b)$

### Score-based

- Calibration: $P(Y=1|S=s, A=a) = P(Y=1|S=s, A=b)$
- Well-calibration: $P(Y=1|S=s, A=a) = P(Y=1|S=s, A=b) = s$
- Balance for positive class: $E(S \mid Y=1, A=a) = E(S \mid Y=1, A=b)$
- Balance for negative class: $E(S \mid Y=0, A=a) = E(S \mid Y=0, A=b)$

### Individual Fairness

**Fairness through Awareness (FTA)**: $d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$
$d_Y$: distance in the prediction space; $d_X$: distance feature space; $\lambda$: controls degree to which similar individuals (based on $d_X$) receive similar predictions (based on $d_Y$)
**Fairness through Unawareness (FTU)** (Blinding): Avoiding explicit use PA during training; extension is suppression: avoid explicit use of PA and proxies

[1] not officially defined in the referenced papers, but implemented in mlr3.fairness and/or following the same principles as all confusion matrix based metrics

### Causality-based notions

**Group-level**: FACE, FACT (on average or on conditional average level) (Zafar et al. 2017)
**Individual-level**: counterfactual fairness, path-based fairness (Kusner et al. n.d.)

## 2 Fairness Methods

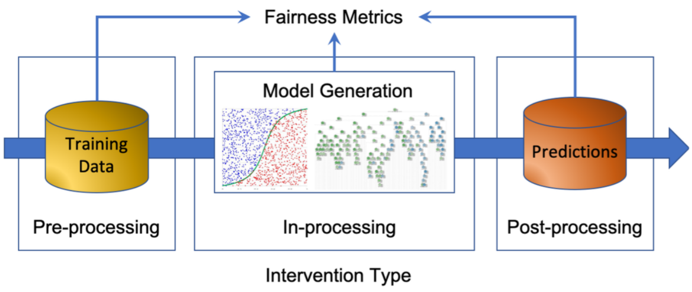


Figure 1: Fairness methods can be applied at different stages of the machine learning pipeline (Caton and Haas 2024).

## 3 Sources of bias and the feedback loop

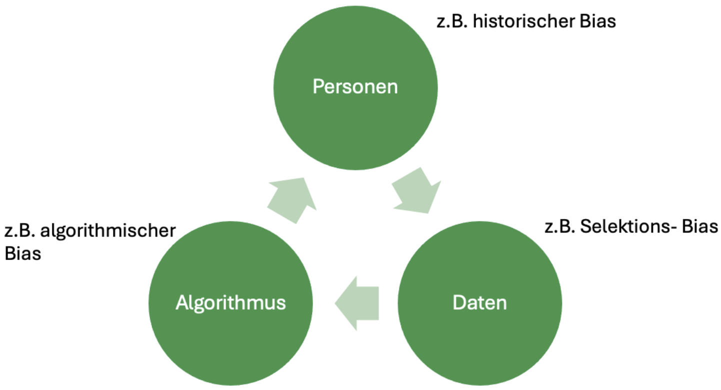


Figure 2: Bias can come into the process at any stage of the data, algorithm, and user feedback loop (Mehrabi et al. 2022).

## References

Caton, Simon and Christian Haas (July 2024). “Fairness in Machine Learning: A Survey”. In: *ACM Computing Surveys* 56.7, pp. 1–38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3616865. (Visited on 12/23/2024).

Kusner, Matt J et al. (n.d.). “Counterfactual Fairness”. In: ().

Mehrabi, Ninareh et al. (July 2022). “A Survey on Bias and Fairness in Machine Learning”. In: *ACM Computing Surveys* 54.6, pp. 1–35. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3457607. (Visited on 01/07/2025).

Verma, Sahil and Julia Rubin (May 2018). “Fairness Definitions Explained”. In: *Proceedings of the International Workshop on Software Fairness*. Gothenburg Sweden: ACM, pp. 1–7. ISBN: 978-1-4503-5746-3. DOI: 10.1145/3194770.3194776. (Visited on 11/16/2024).

Zafar, Muhammad Bilal et al. (2017). “From Parity to Preference-based Notions of Fairness in Classification”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. (Visited on 12/29/2024).