# Algorithmic Fairness

## A Pathway to Developing Responsible AI Systems

# Équité algorithmique

## Une voie pour développer des systèmes d'IA responsables

Golnoosh Farnadi

HEC Montréal/Université de Montréal/Canada CIFAR AI chair/MILA


World AI Summit, April 20, 2023

# What is the Importance of Algorithmic Fairness?

**Algorithmic fairness:**

*technical approaches* to mitigating algorithmic discrimination

**Other approaches:**

Investigative journalism, auditing

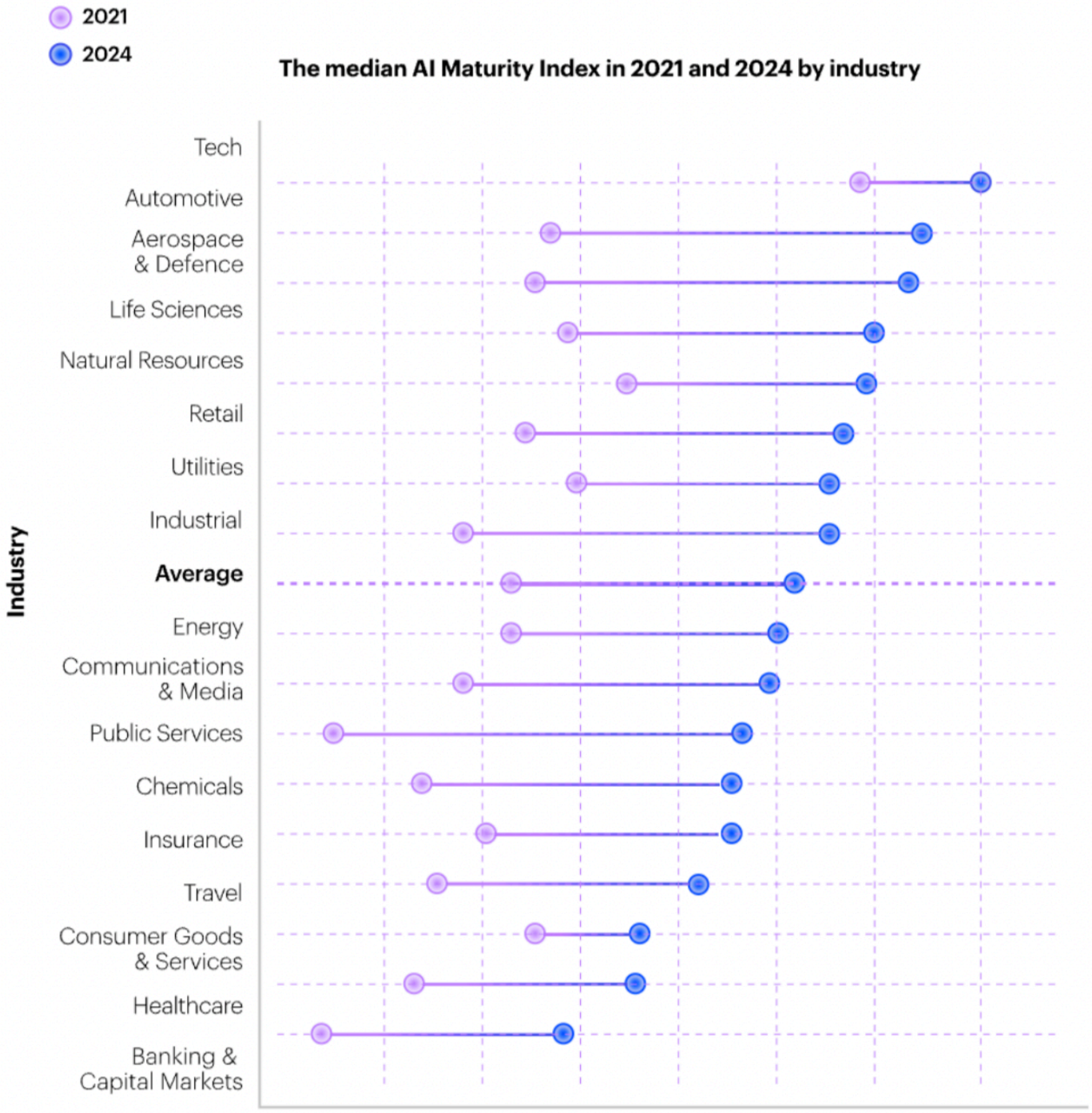
Policy making and advocacy

Community organizing



Mila

# AI is here to stay!




Source: Accenture Research

Mila

# Algorithmic harms

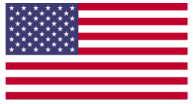

- **Allocation harm**: E.g., Amazon Hiring system, COMPAS risk assessment
- **Quality of service harm**: E.g., gender shades, VMS make women sick
- **Stereotyping harm**, e.g., Black criminality in predictive policing, gender issues in NLP (in translation)
- **Denigration harm**, e.g., mislabeling images of Black women as Gorillas, Chatbot Tay for hate speech
- **Over and under-representation harm**, e.g., images of men in image search results

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning.Special Interest Group for Computing, Information and Society (SIGCIS)(2017)

Kate Crawford at NeurIPS 2017 Tutorial

Mila

# Laws against Discrimination



**Legally recognized 'protected classes'**

**Race** (Civil Rights Act of 1964)
**Color** (Civil Rights Act of 1964)
**Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964)
**Religion** (Civil Rights Act of 1964)
**National origin** (Civil Rights Act of 1964)
**Citizenship** (Immigration Reform and Control Act)
**Age** (Age Discrimination in Employment Act of 1967)
**Pregnancy** (Pregnancy Discrimination Act)
**Familial status** (Civil Rights Act of 1968)
**Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
**Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

**sensitive attributes**

**Regulated domains**

**Credit** (Equal Credit Opportunity Act)
**Education** (Civil Rights Act of 1964; Education Amendments of 1972)
**Employment** (Civil Rights Act of 1964)
**Housing** (Fair Housing Act)
**Public Accommodation** (Civil Rights Act of 1964)
Extends to marketing and advertising; not limited to final decision
This list sets aside complex web of laws that regulates the government

Article 14. Equality before law. -The State shall not deny to any person equality before the law or the equal protection of the laws within the territory of India. (1) **The State shall not discriminate against any citizen on grounds only of religion, race, caste, sex, place of birth or any of them**.

**EU Charter of Fundamental Rights**

1. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited. 2.

https://www.refworld.org/pdfid/4d886bf02.pdf

Canadians have the right to be treated fairly in workplaces free from discrimination, and our country has laws and programs to protect this right. **The Canadian Human Rights Act** is a broad-reaching piece of legislation that prohibits discrimination on the basis of gender, race, ethnicity and other grounds. May 30, 2022

https://laws-lois.justice.gc.ca/eng/acts/h-6/fulltext.html

The basis for progressively redressing these conditions lies in the Constitution which, amongst others, upholds the values of human dignity, equality, freedom and social justice in a united, non-racial and non-sexist society where all may flourish;

South Africa also has international obligations under binding treaties and customary international law in the field of human rights which promote equality and prohibit unfair discrimination. Among these obligations are those specified in the Convention on the Elimination of All Forms of Discrimination Against Women and the Convention on the Elimination of All Forms of Racial Discrimination;

Initiating an Anti-Discrimination Regime in China

**The 1982 Constitution has enshrined the principle of equality of all citizens before the law** (Article 33). Articles 4, 36, 48, and 89 also guarantee the rights of ethnic minorities, religious freedom and gender equality and prohibits discrimination on those grounds.

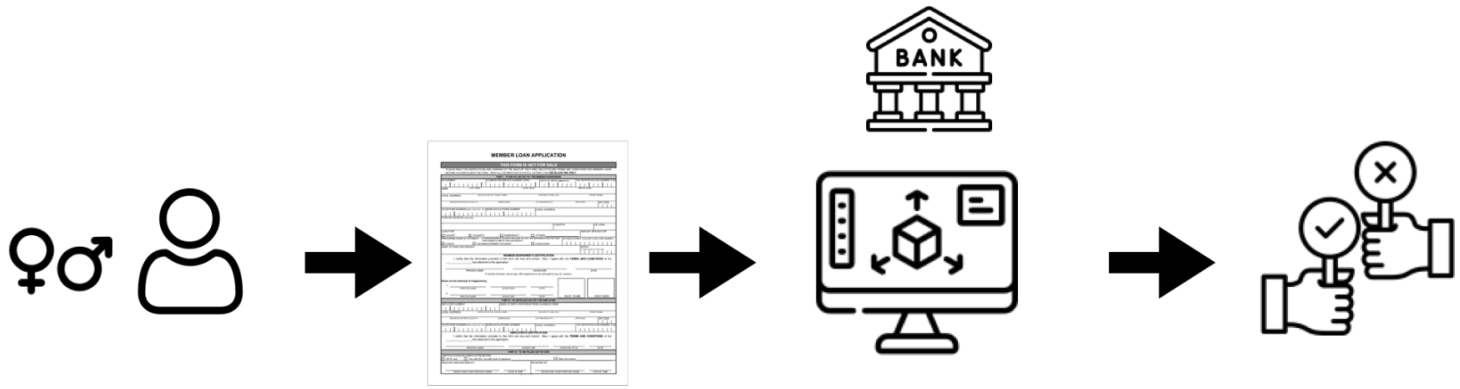
Mila

# How to Define Fairness in Machine Learning?

# Running Example

Confusion Matrix

| | $Y=1$ | $Y=0$ |
|---|---|---|
| $\hat{Y}=1$ | TP | FP |
| $\hat{Y}=0$ | FN | TN |




Applicant

Loan Application

Loan Approval

$S$

**sensitive attribute**

$X$

**non-sensitive attributes**

$\hat{Y}$

**Prediction decision**

$Y$

**Actual Outcome**

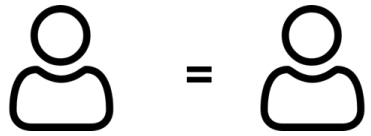
Note gender is assumed to be binary for the sake of simplicity

Scheuerman, M.K., Brubaker, J.R., 2018 *Gender is not a Boolean: Towards Designing Algorithms to Understand Complex Human Identities.*
Hu, L., Kohler-Hausmann, I., 2020. *What's Sex Got To Do With Fair Machine Learning?*
Lu, C., Kay, J., McKee, K., 2022. *Subverting machines, fluctuating identities: Re-learning human categorization.*

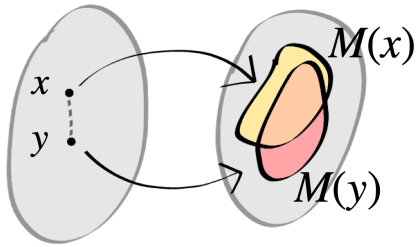
Mila

# What does fairness in ML mean?

**Individual**: measures the impact that discrimination has on the individuals



**E.g., similar applicants, should have similar probability of receiving positive loan approval**



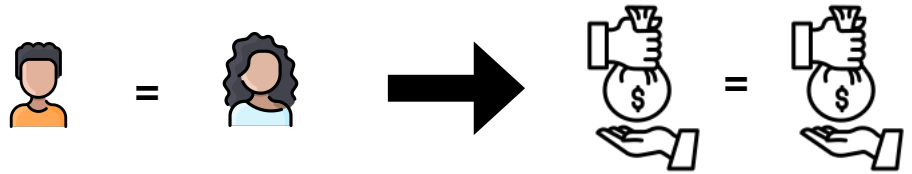

> The Lipschitz condition requires that any two individuals $x, y$ that are at distance $d(x, y) \in [0, 1]$ map to distributions $M(x)$ and $M(y)$, respectively, such that the statistical distance between $M(x)$ and $M(y)$ is at most $d(x, y)$. In other words, the distributions over outcomes observed by $x$ and $y$ are indistinguishable up to their distance $d(x, y)$.

Mila

Dwork, Cynthia, et al. "Fairness through awareness." Proceedings of the 3rd innovations in theoretical computer science conference. ACM, 2012.

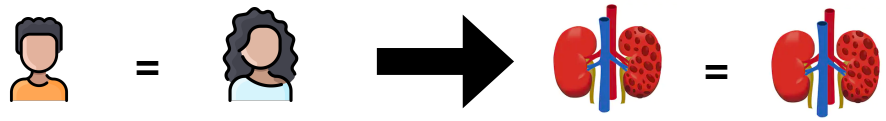
# Individual Fairness

## Loan Approval



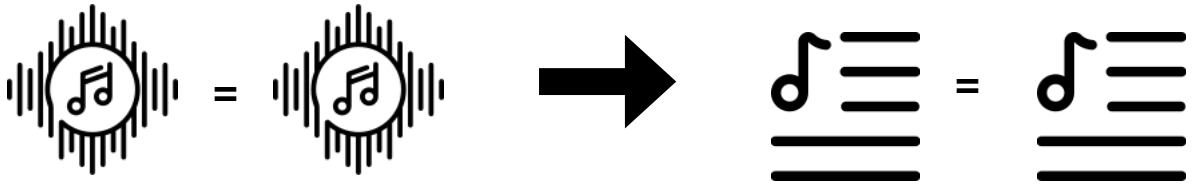
**E.g., similar applicants with same application and only their gender differs, should have similar probability of receiving loan approval**

## Kidney Exchange Program



**E.g., similar patients who are in a cycle to receive of a kidney transplant, should have similar chances of receiving a kidney transplant**

## Recommender System



**E.g., similar music items/artists with similar music features, should have similar probability of appearing in playlists**

**Farnadi, G.**, St-Arnaud, W., Babaki, B., & Carvalho, M. (2021). Individual Fairness in Kidney Exchange Programs. In *AAAI*

Salganik, R., Diaz, F., & **Farnadi, G**. (2022). Analyzing the Effect of Sampling in GNNs on Individual Fairness. In proceedings of FairRec Workshop at RecSsys 2022, arXiv preprint arXiv:2209.03904.

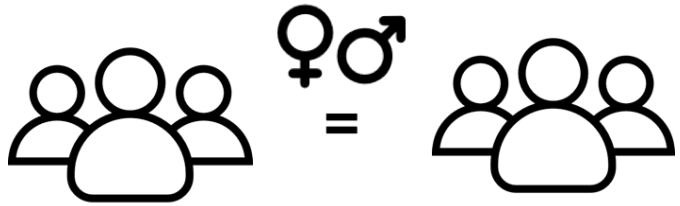
Mila

# What does fairness in ML mean?

**Individual**: measures the impact that discrimination has on the individuals

**E.g., similar applicants, should have similar probability of receiving positive loan approval**

**Group**: measures the impact that the discrimination has on the groups of individuals



**E.g., The probability of receiving positive loan approval should be similar among female and male applicants**

Mila

Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)* (pp. 1-7). IEEE.

# Statistical Fairness Notions

## Demographic Parity

$$P(\hat{Y} = 1 \mid S = 1) = P(\hat{Y} = 1 \mid S = 0)$$

**equal probability of receiving a positive loan approval for female and male applicants**

## Equal opportunity

$$P(\hat{Y} = 1 \mid Y = 1, S = 1) = P(\hat{Y} = 1 \mid Y = 1, S = 0)$$

**classifier should give similar results to applicants of both genders with actual positive loan approval.**

## Equalized odds

$$P(\hat{Y} = 1 \mid Y, S = 1) = P(\hat{Y} = 1 \mid Y, S = 0)$$

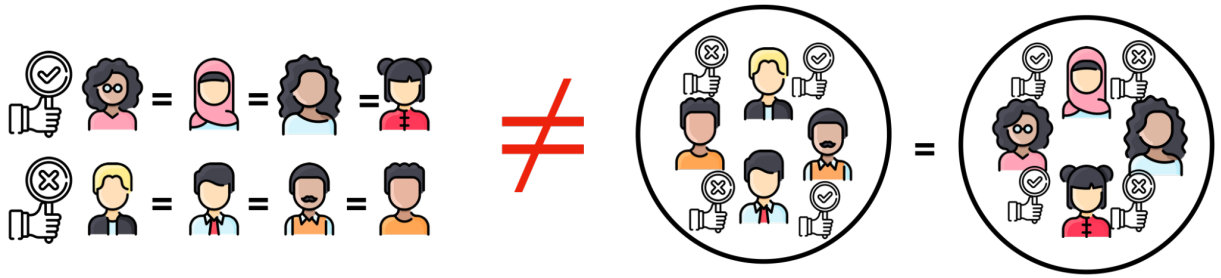
**applicants with a rejected loan application and applicants with an accepted loan application should have a similar classification, regardless of their gender.**

Mila

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

# Impossibility of Fairness

Impossibility wrt group and individual notions



Impossibility wrt various group fairness notions

You can only achieve one of these measures: demographic parity, equality of odds, and equality of opportunity

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im) possibility of fairness." *arXiv preprint arXiv:1609.07236* (2016).

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

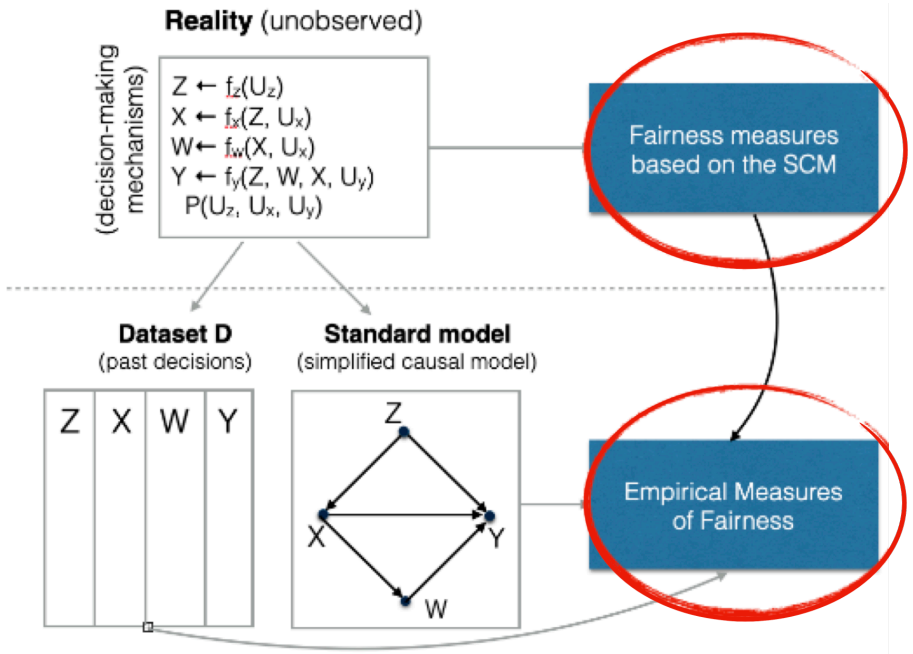
Mila

# Causal Notions of Fairness

- Causal fairness notions are based on social-legal requirements, e.g.,

US Supreme Court, 2015

A disparate-impact claim relying on a statistical disparity must fail if the plaintiff cannot point to a defendant's policy or policies causing that disparity.




- Based on existence of causal mechanisms, which are almost never observed.

- Construction of causal graph to encode assumptions about underlying structural causal model (SCM) is required


Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, *30*.

Zhang & Bareinboim. “Fairness in Decision-Making - The Causal Explanation Formula, AAAI, 2018

Nilforoshan, Hamed, et al. "Causal conceptions of fairness and their consequences." *International Conference on Machine Learning*. PMLR, 2022.

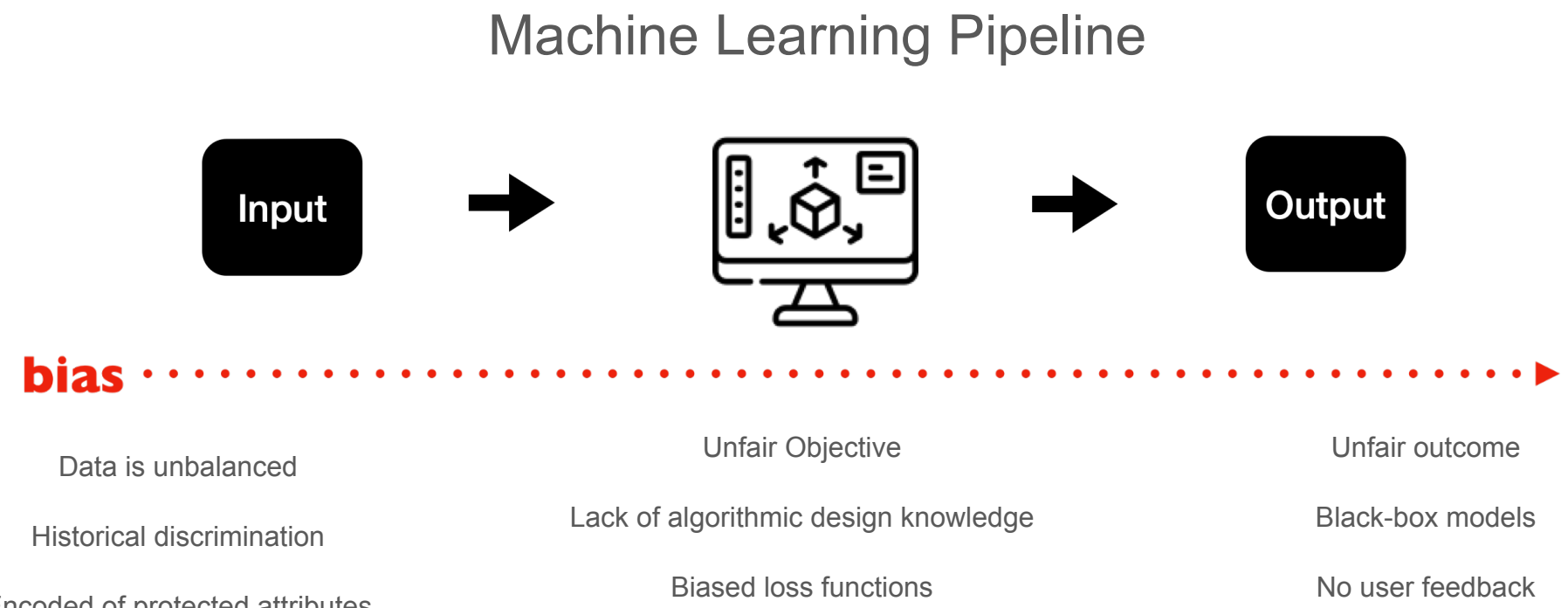

Mila

# How to Ensure Algorithmic Fairness in Machine Learning?



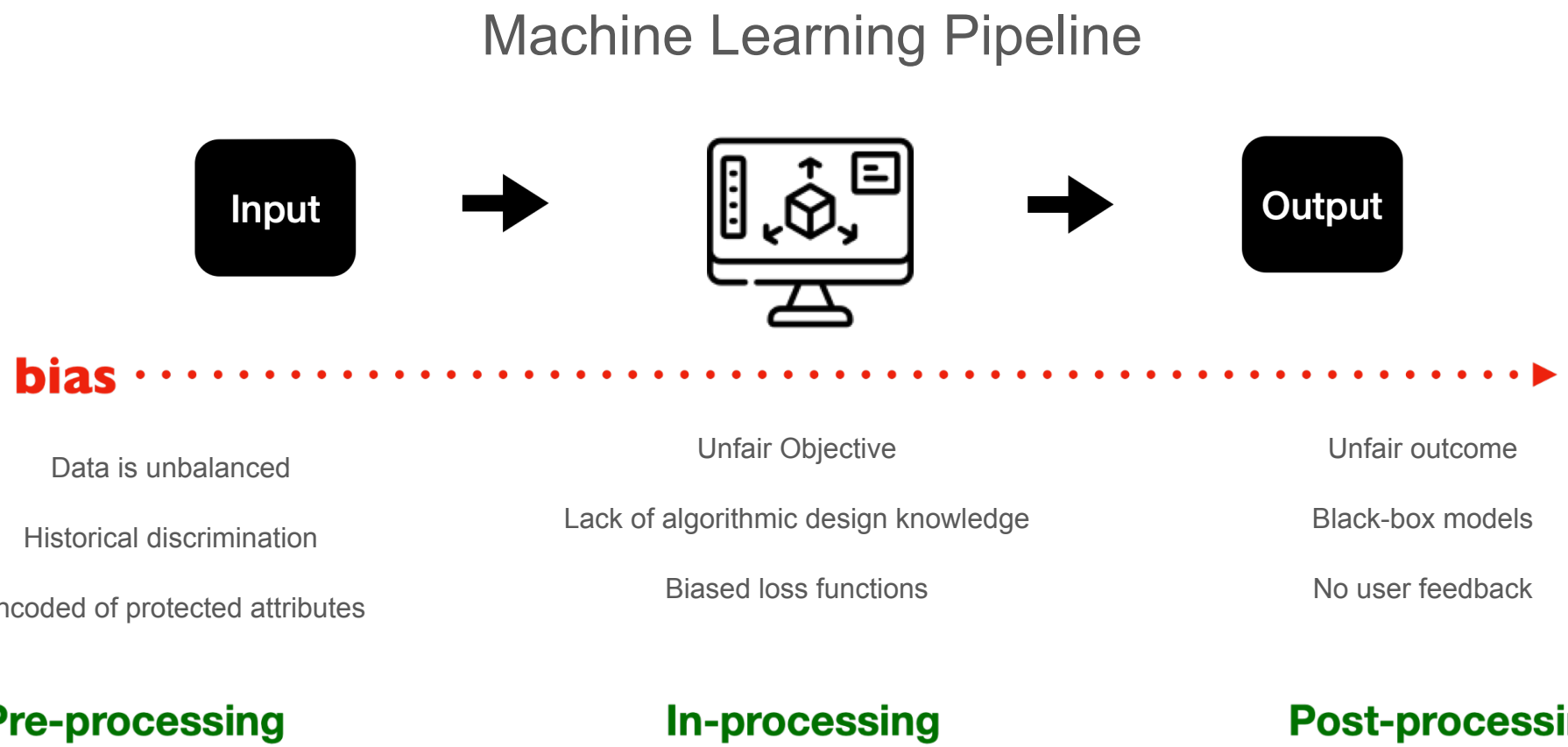
Mila

# How to mitigate algorithmic discrimination in machine learning?



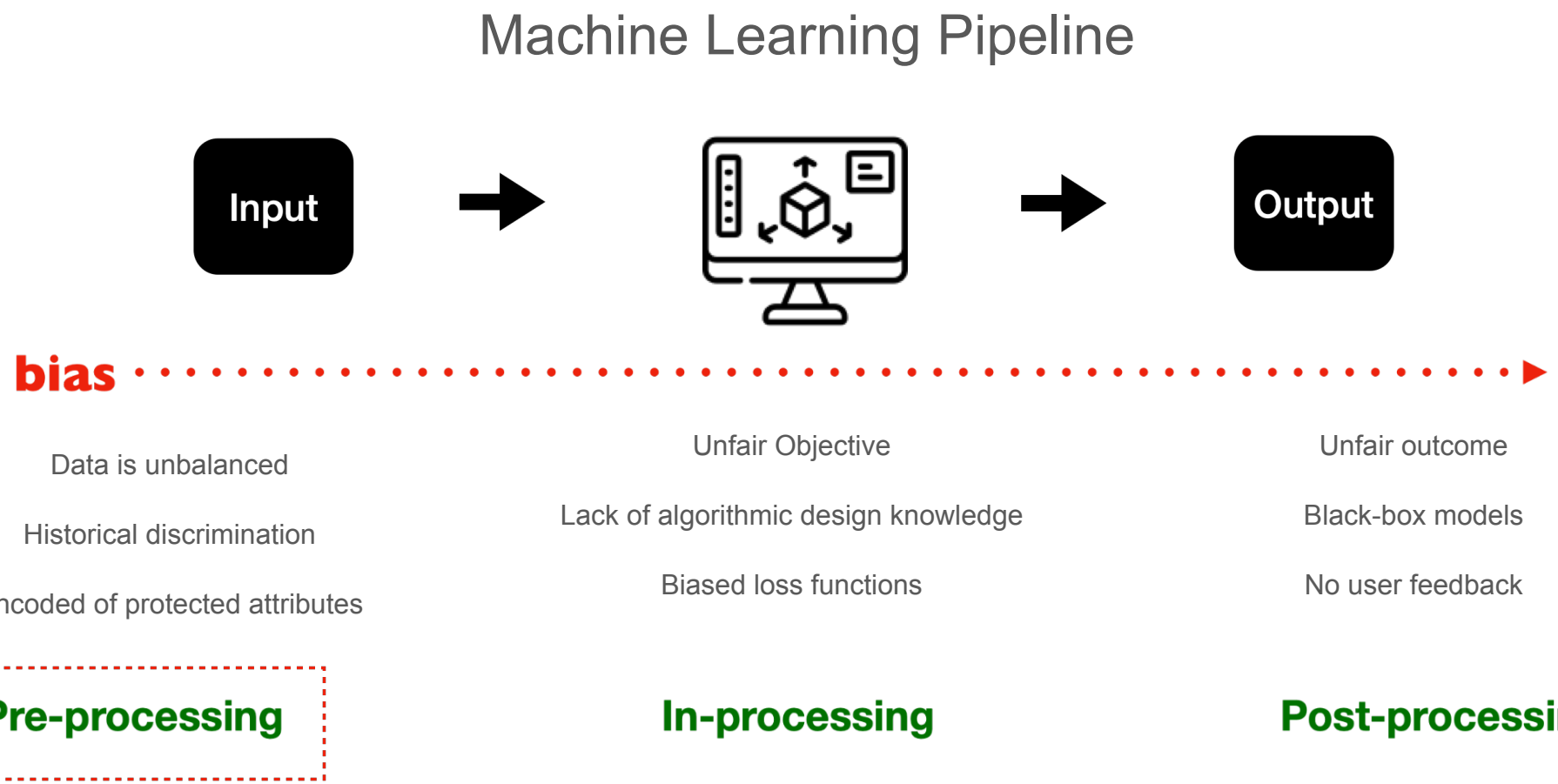

Mila

# How to mitigate algorithmic discrimination in machine learning?



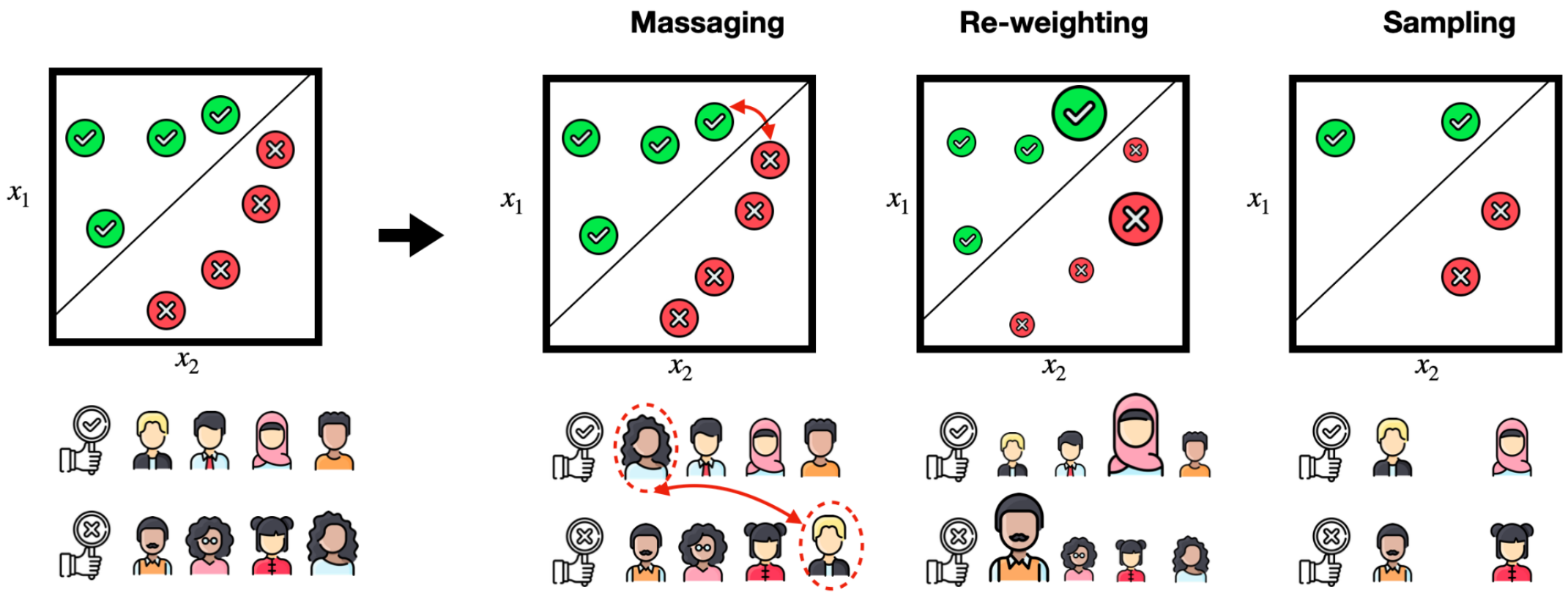

Mila

# How to mitigate algorithmic discrimination in machine learning?




Mila

# Fairness in Pre-Processing: Data De-Biasing

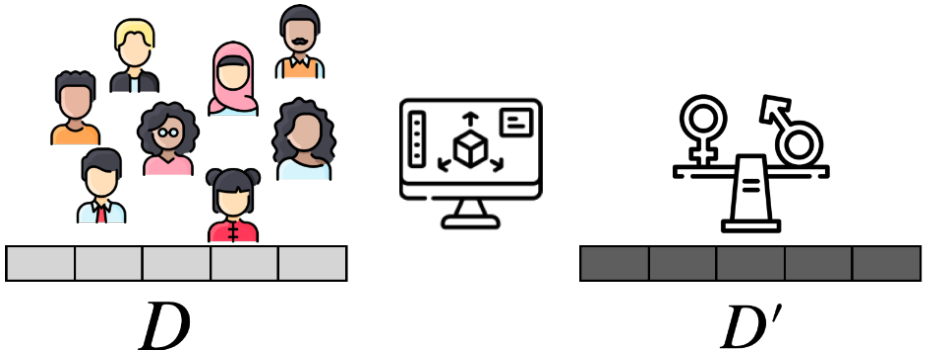



Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, *33*(1), 1-33.

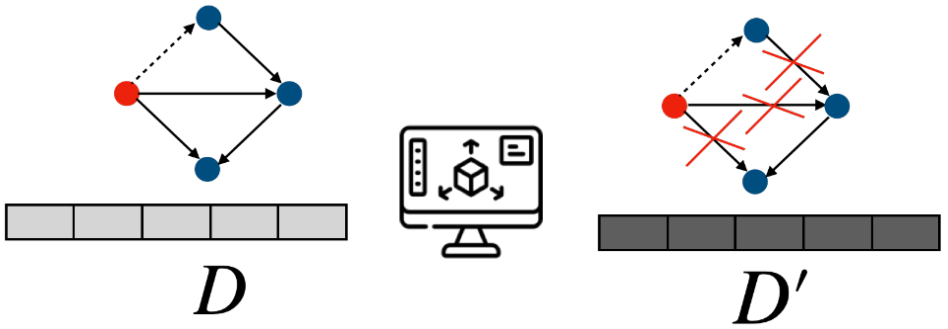
Alabdulmohsin, I., Schrouff, J., & Koyejo, O. (2022). A Reduction to Binary Approach for Debiasing Multiclass Datasets. *arXiv preprint arXiv:2205.15860*.

Mila

# Fairness in Pre-Processing: Data Generative Models




E.g., Using Generative Adversarial Networks (GANs), Variational Autoencoders, etc.



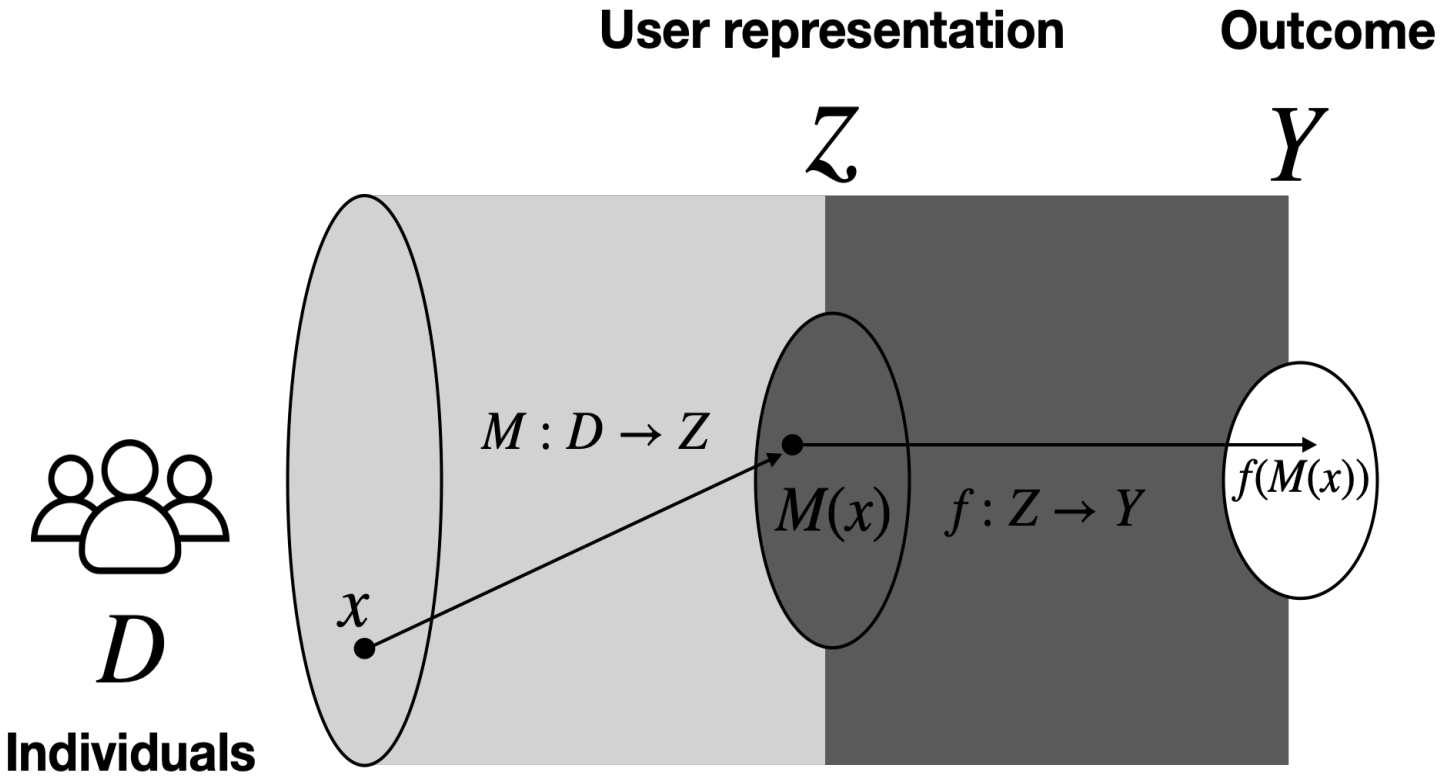

E.g., Using SCMs (by removing paths from sensitive attributes)

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).

van Breugel, Boris, et al. "Decaf: Generating fair synthetic data using causally-aware generative networks." *Advances in Neural Information Processing Systems* 34 (2021): 22221-22233.

Mila

# Fair Representation Learning




## Goal of Representation learning

**Preserve Performance:**

**Reconstruction term**: the learned representation should resemble the original data

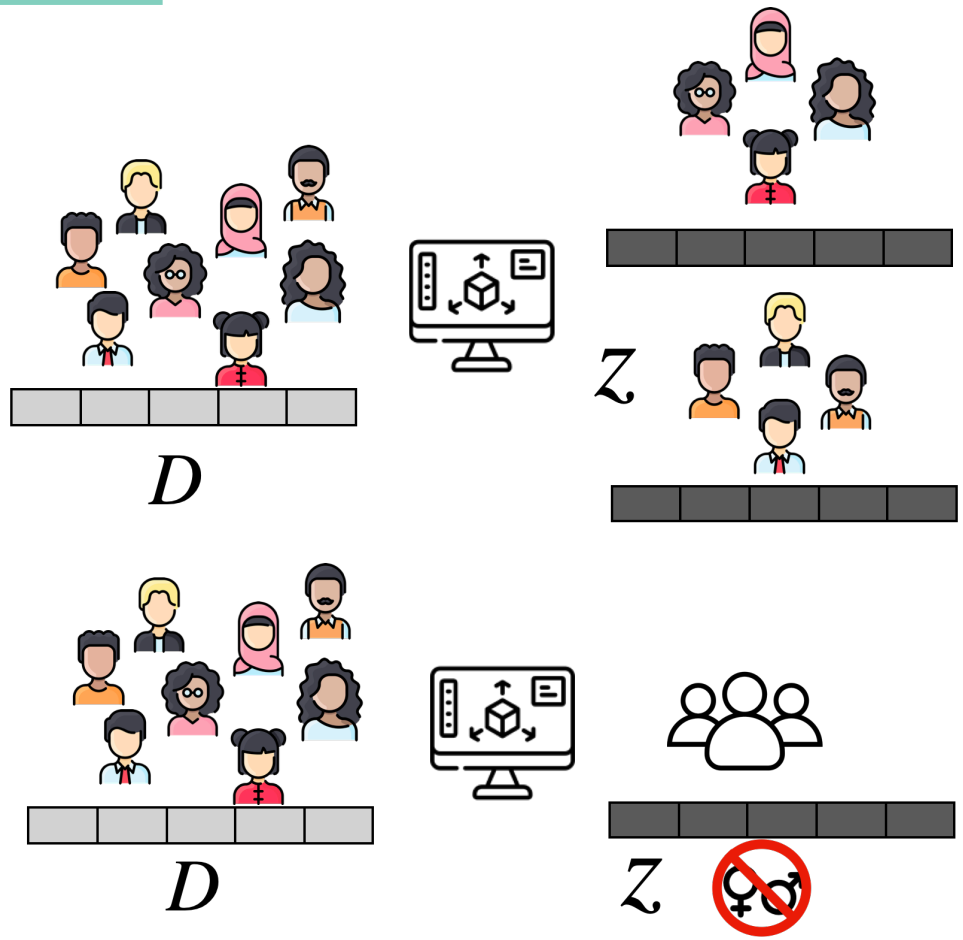
**Utility terms**: the learned representation should predict target variable

**+ Fairness**

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May). Learning fair representations. In International conference on machine learning (pp. 325-333). PMLR

Mila

# Fair Representation Learning: Group Fairness




## Fairness

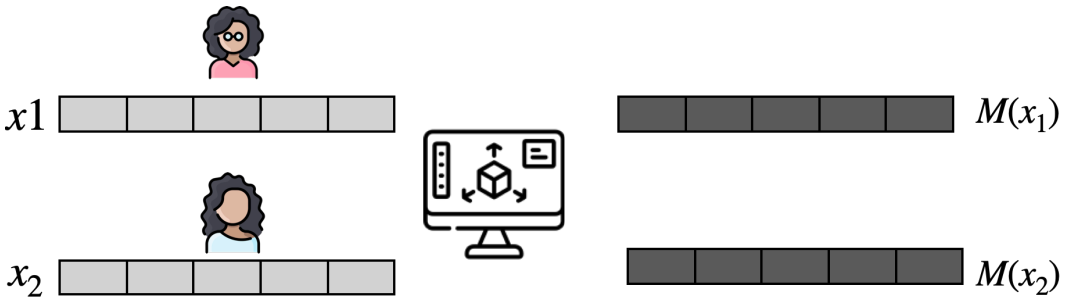
- **Balancing the distribution** among various groups
- **Remove sensitive attributes** (common approach is to use deep learning: VAE, adversarial learning, or disentangled learning)


Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2015). The variational fair autoencoder. arXiv preprint arXiv:1511.00830.

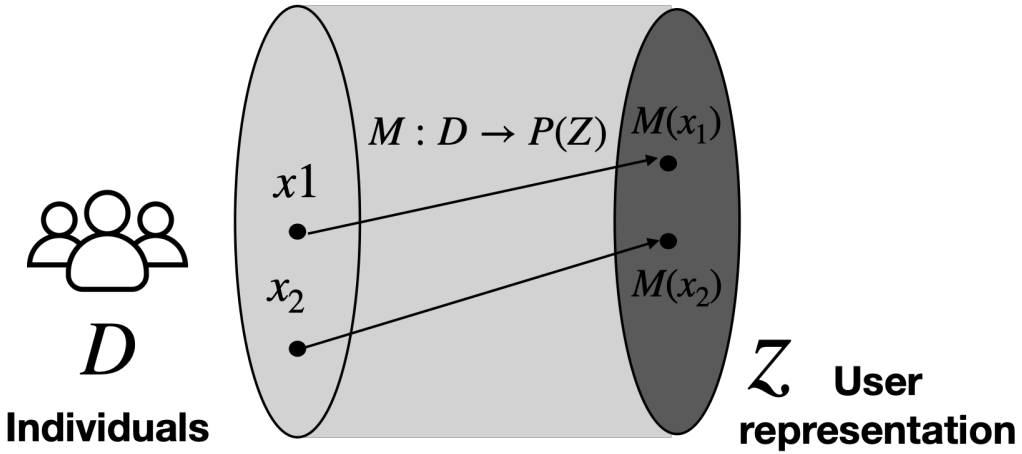
Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018, July). Learning adversarially fair and transferable representations. In International Conference on Machine Learning (pp. 3384-3393). PMLR.

Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., & Bachem, O. (2019). On the fairness of disentangled representations. *Advances in Neural Information Processing Systems*, *32*.

# Fair Representation Learning: Individual Fairness




**Lipschiz condition** $||M(x_1) - M(x_2)|| \leq d(x_1, x_2)$



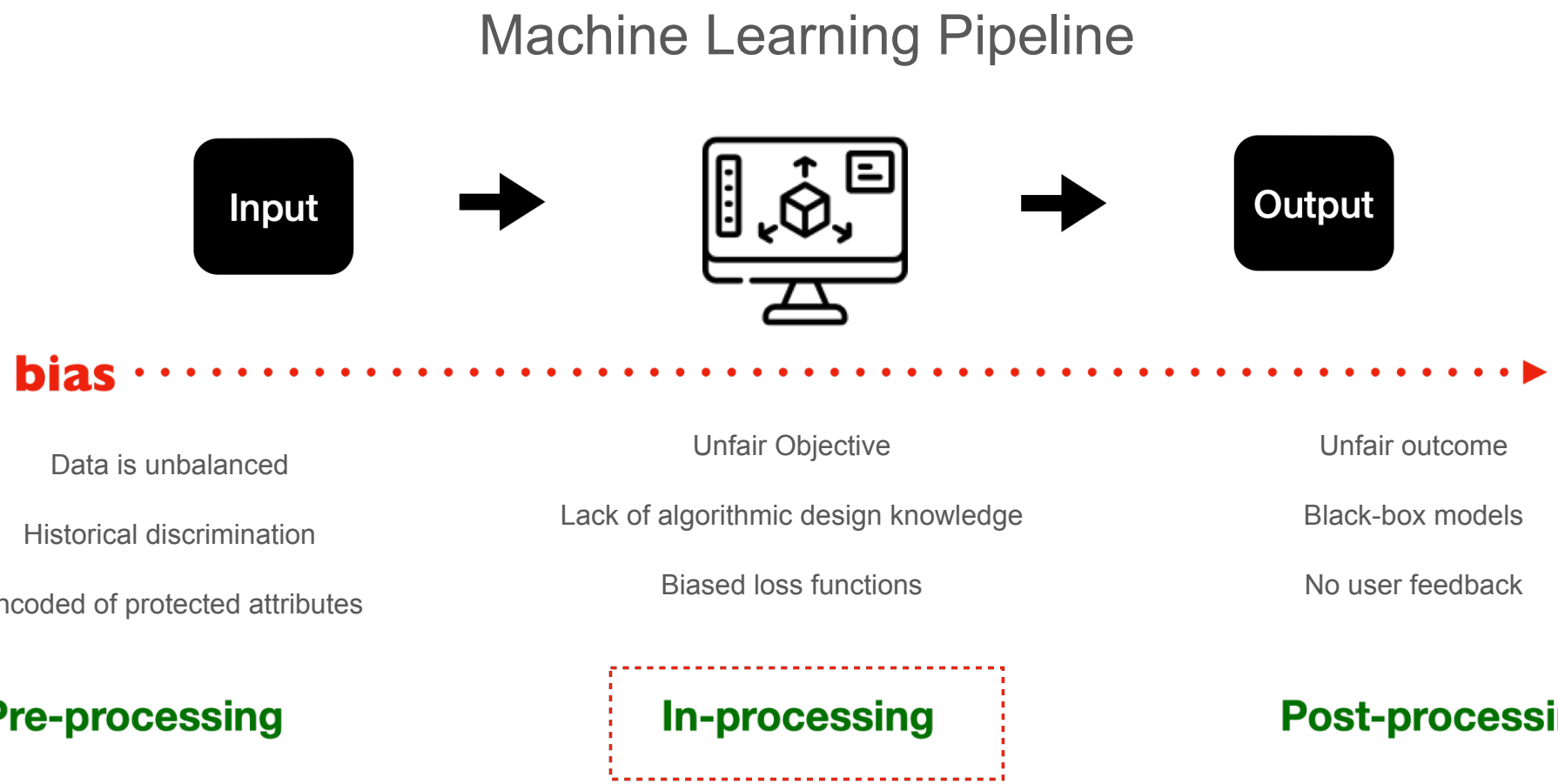

## Fairness

- **Similar individuals** should map to **similar distributions.**
- **Task-specific** similarity metric. Ideally captures ground truth or society's best approximation
- Many applications: ranking in recommender systems, financial risk metrics, health metric for treating patients, etc.

Dong, Yushun, et al. "Individual fairness for graph neural networks: A ranking based approach." Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021.
Salganik, Rebecca, Fernando Diaz, and Golnoosh Farnadi. "Analyzing the Effect of Sampling in GNNs on Individual Fairness." *arXiv preprint arXiv:2209.03904* (2022).

Mila

# How to mitigate algorithmic discrimination in machine learning?




Mila

# In-processing techniques

- Supervised learning tasks are often expressed as optimization problems

$$\underset{\theta}{\text{minimize}} \quad f(X, Y; \theta)$$

- The optimization problem: finding the parameters that give the best model w.r.t the desired properties

**Fairness in another desired property of the learned models**

$$g(X, Y; \theta)$$

Mila

# In-processing techniques

- **Not all optimization problems are the same!**
- Some problems are **computational easy**
- Some problems are **hard**, but **behave well** (approximation methods work well)
- Some problems are **hard**, but have **structure**. And we can exploit this structure.

**Adding fairness can change these properties!**

Mila

# In-processing techniques

Fairness as Constrained Optimization

$$\underset{\theta}{\textbf{minimize}} \; f(X, Y; \theta) \quad \textbf{subject to} \; g(X, Y; \theta)$$

Fairness as Regularizer

$$\underset{\theta}{\textbf{minimize}} \; f(X, Y; \theta) + \lambda g(X, Y; \theta)$$

Fairness as Multi-objective Optimization

$$\underset{\theta}{\textbf{minimize}} \; f(X, Y; \theta) \times g(X, Y; \theta)$$


Choi, Y., Farnadi, G., Babaki, B., & Van den Broeck, G. (2020, April). Learning fair naive bayes classifiers by discovering and eliminating discrimination patterns. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 06, pp. 10077-10084).

Mohammadi, K., Sivaraman, A., & Farnadi, G. (2022). FETA: Fairness Enforced Verifying, Training, and Predicting Algorithms for Neural Networks. *arXiv preprint arXiv:2206.00553*.

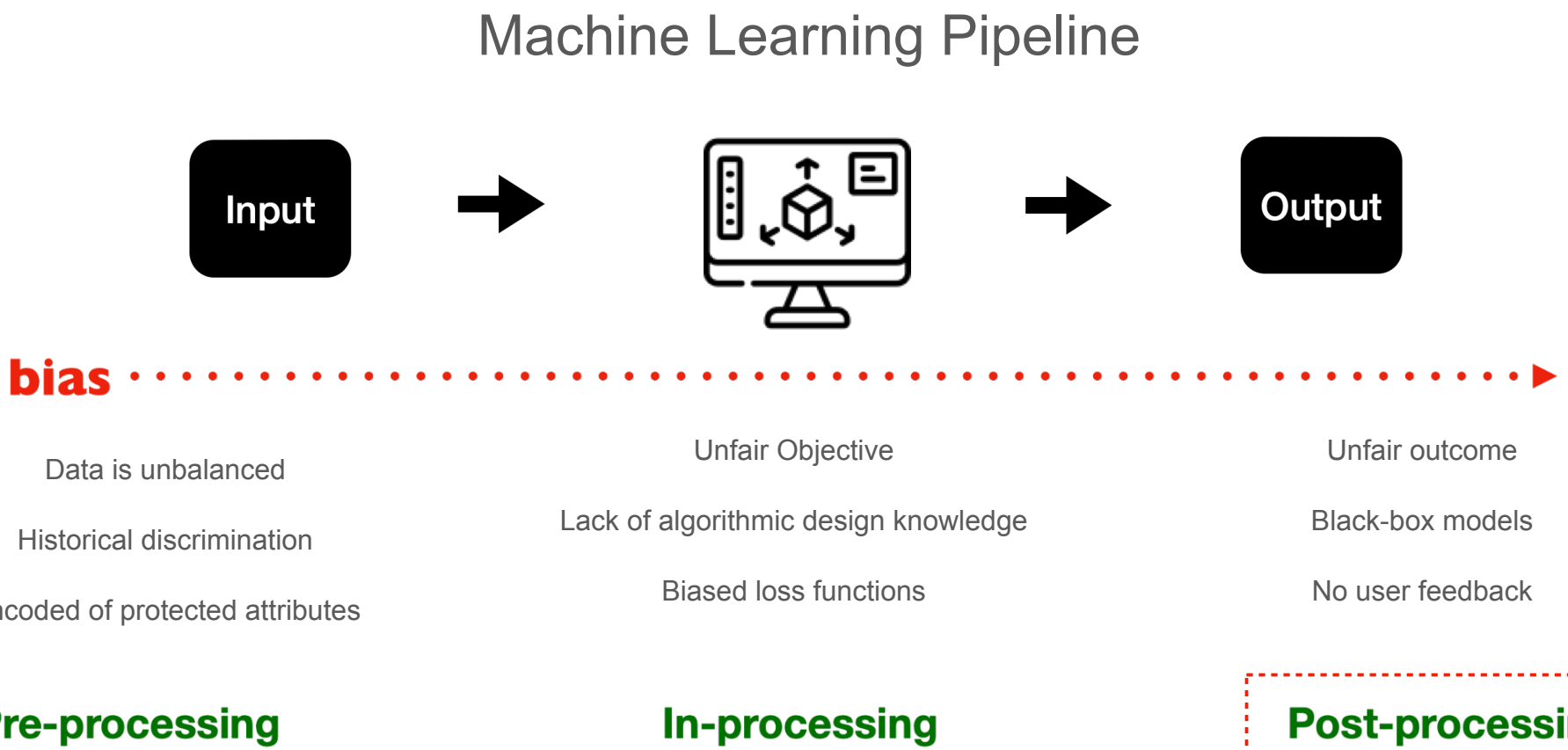
Kamishima, Toshihiro, Shotaro Akaho, and Jun Sakuma. "Fairness-aware learning through regularization approach." 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE, 2011.

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, “A Reductions Approach to Fair Classification,” arXiv.org, 16-Jul-2018. [Online]. Available: https://arxiv.org/abs/1803.02453.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018, July). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning* (pp. 2564-2572). PMLR.

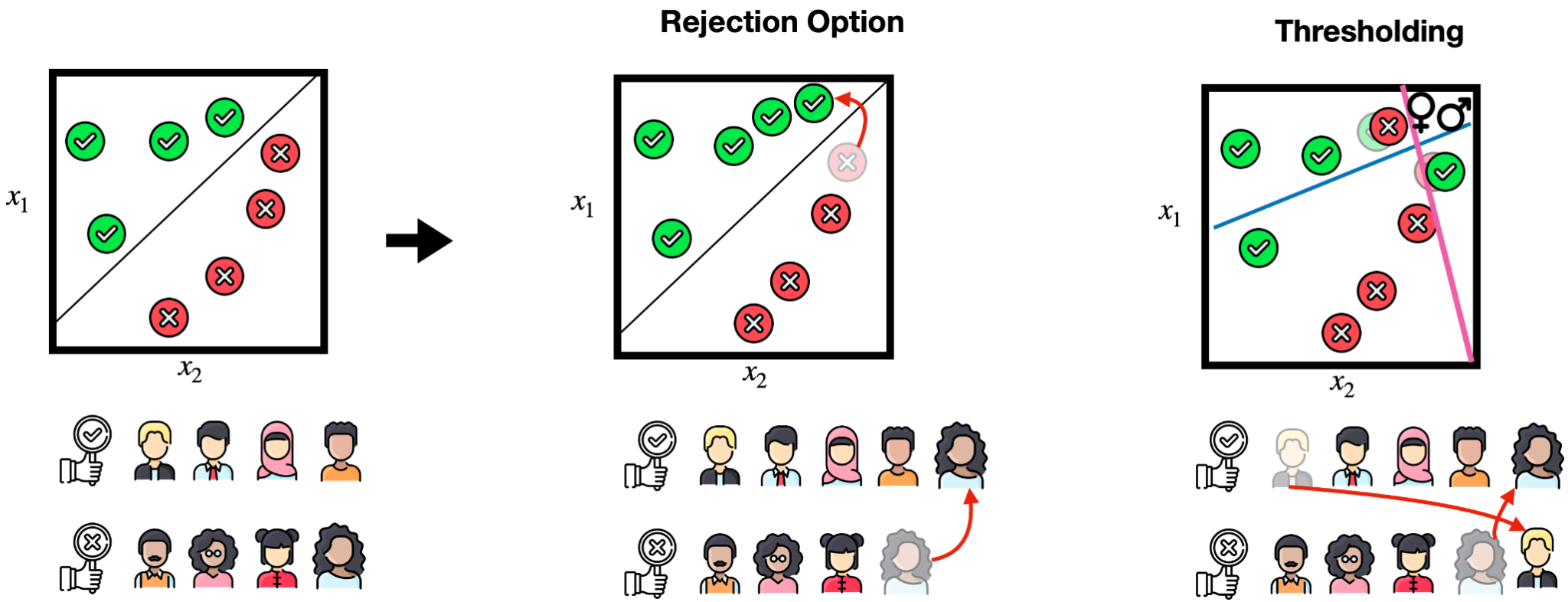

Mila

# How to mitigate algorithmic discrimination in machine learning?




Mila

# Fairness in Post-Processing




Kamiran, F., Karim, A., & Zhang, X. (2012, December). Decision theory for discrimination-aware classification. In 2012 IEEE 12th International Conference on Data Mining (pp. 924-929). IEEE.

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

Mila

# Trade-offs

| | Ease of implementation and (re)-use | Scalability | Ease of auditing | Fairness/ Performance tradeoff | Generalization |
|---|---|---|---|---|---|
| **Pre-processing**, e.g., representation learning | ✓ | ✓ | ✓ | | ✓ |
| **In-processing**, e.g., fairness regularizer | | | ✓ | ✓ | ✓ |
| **Post-processing**, e.g., thresholding | | ✓ | ✓ | | |

**Fairness should not be an afterthought and instead, should be incorporated throughout the entire machine learning pipeline.**

Inspired by Sanmi Koyejo's talk on fair representation learning tutorial at NeurIPS 2019

Mila

# Summary

- No free lunch: Fairness is a **socio-technical challenge**
- Many aspects of fairness are **NOT** captured by the statistical measures
- One notion **cannot** simultaneously satisfy all metrics
- Algorithmic fairness is **highly dependent** on the fairness notion, and the result change by changing the notion of fairness
- We may need to make a **trade-off** in different contexts
- **Participatory design is an integral aspect of algorithmic fairness that extends beyond the data, model, or outcome.**

Mila

# Take aways

- **Responsible AI** entails more than just algorithmic fairness; it also encompasses privacy, accountability, explainability, robustness, and other factors.
- It's important to **educate** yourself on algorithmic bias and discrimination if you're involved in constructing an AI model.
- Ask these questions:  Is AI the **appropriate tool** for the task at hand? What is the **objective** of the automated system? Which **individuals or organizations** are utilizing the model?
- **Having a larger model does not necessarily mean it is superior.**

Mila

# Thank you!


Twitter: @gfarnadi
Email: farnadig@mila.quebec
Webpage: https://gfarnadi.github.io/

Mila