Algorithmic fairness: a post-processing approach

Christophe Denis

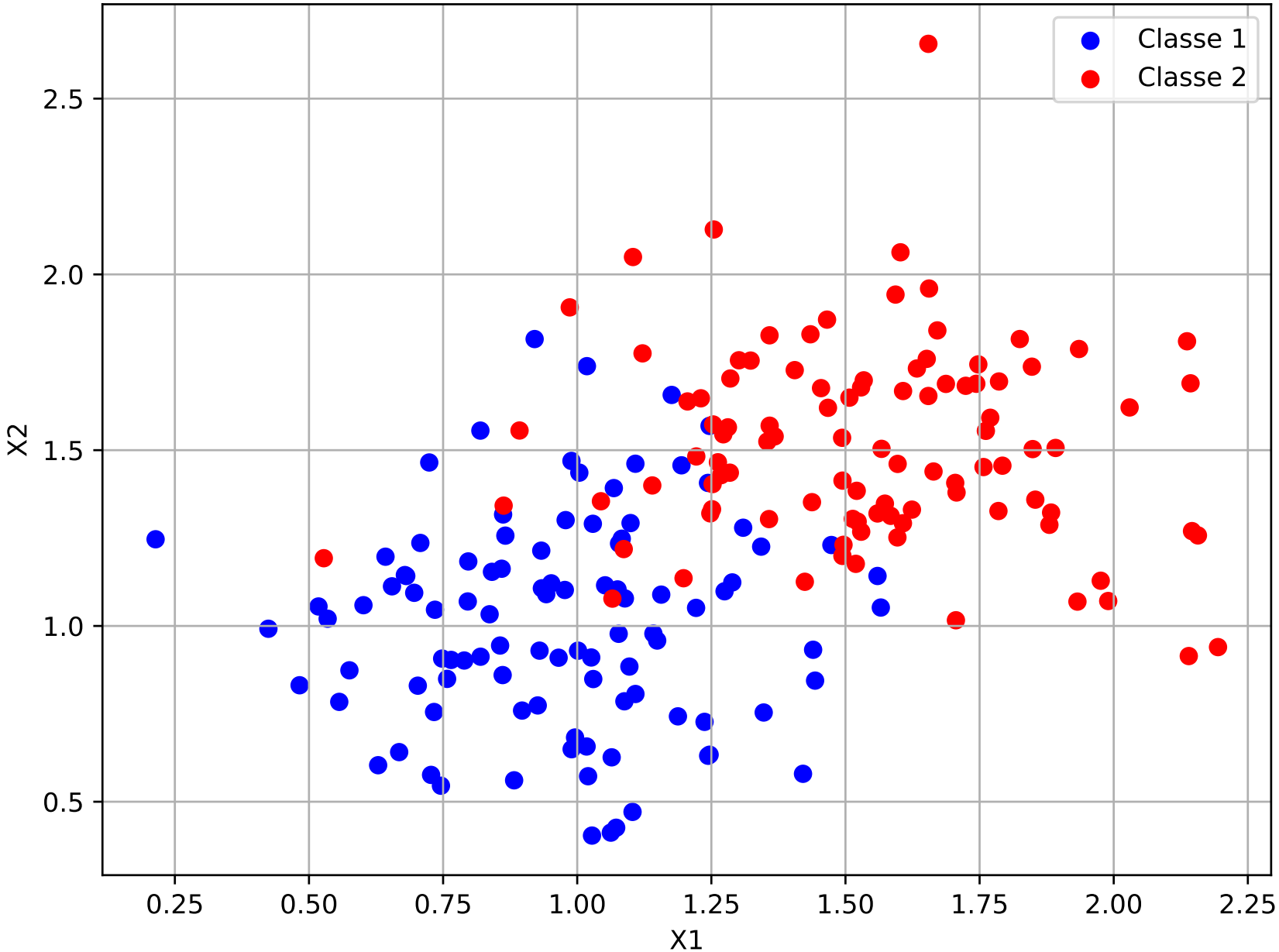
SAMM
Université Paris 1 Panthéon-Sorbonne

10/09/2026

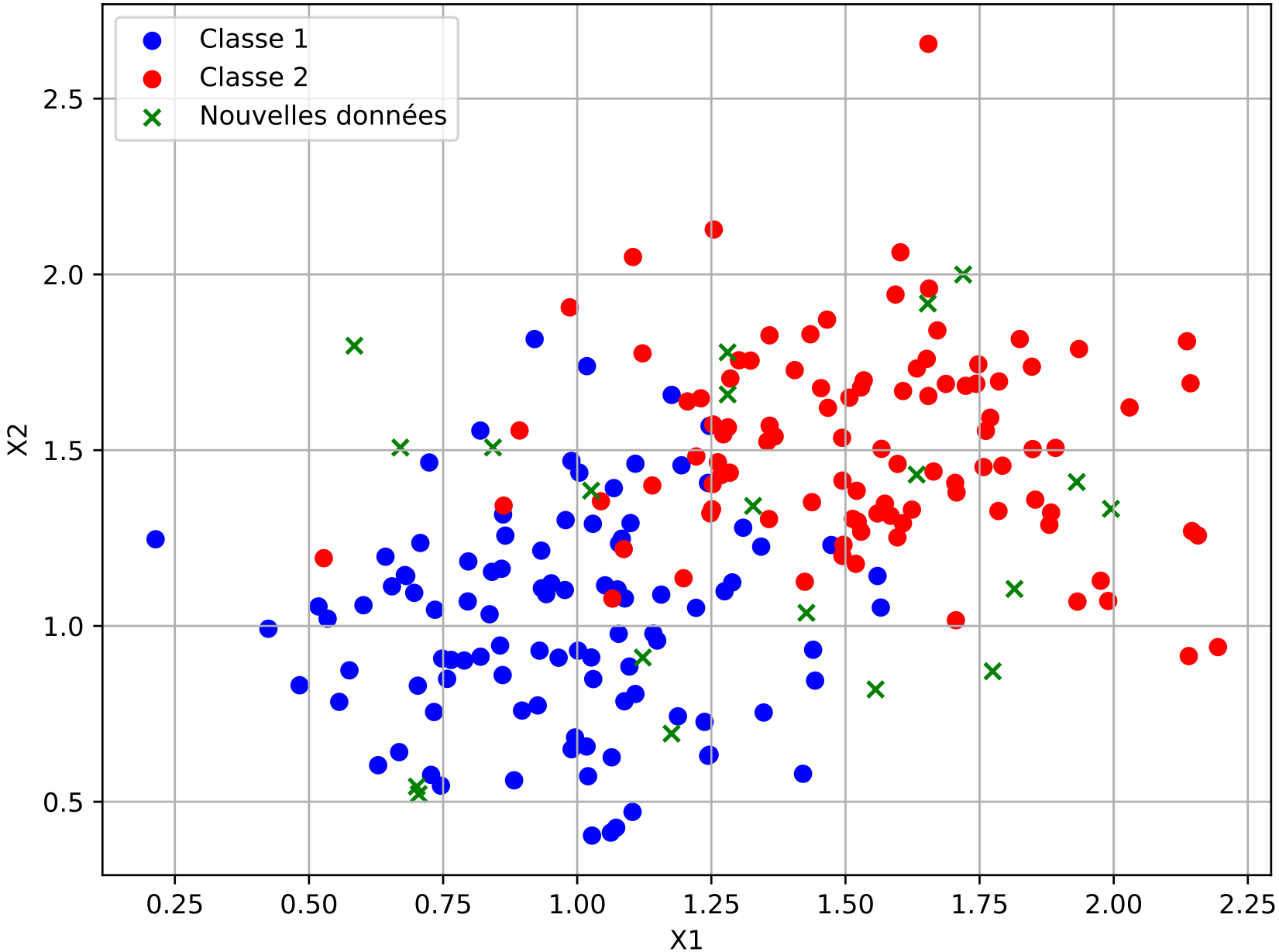
MASHS 2026, MSE

Introduction and definitions

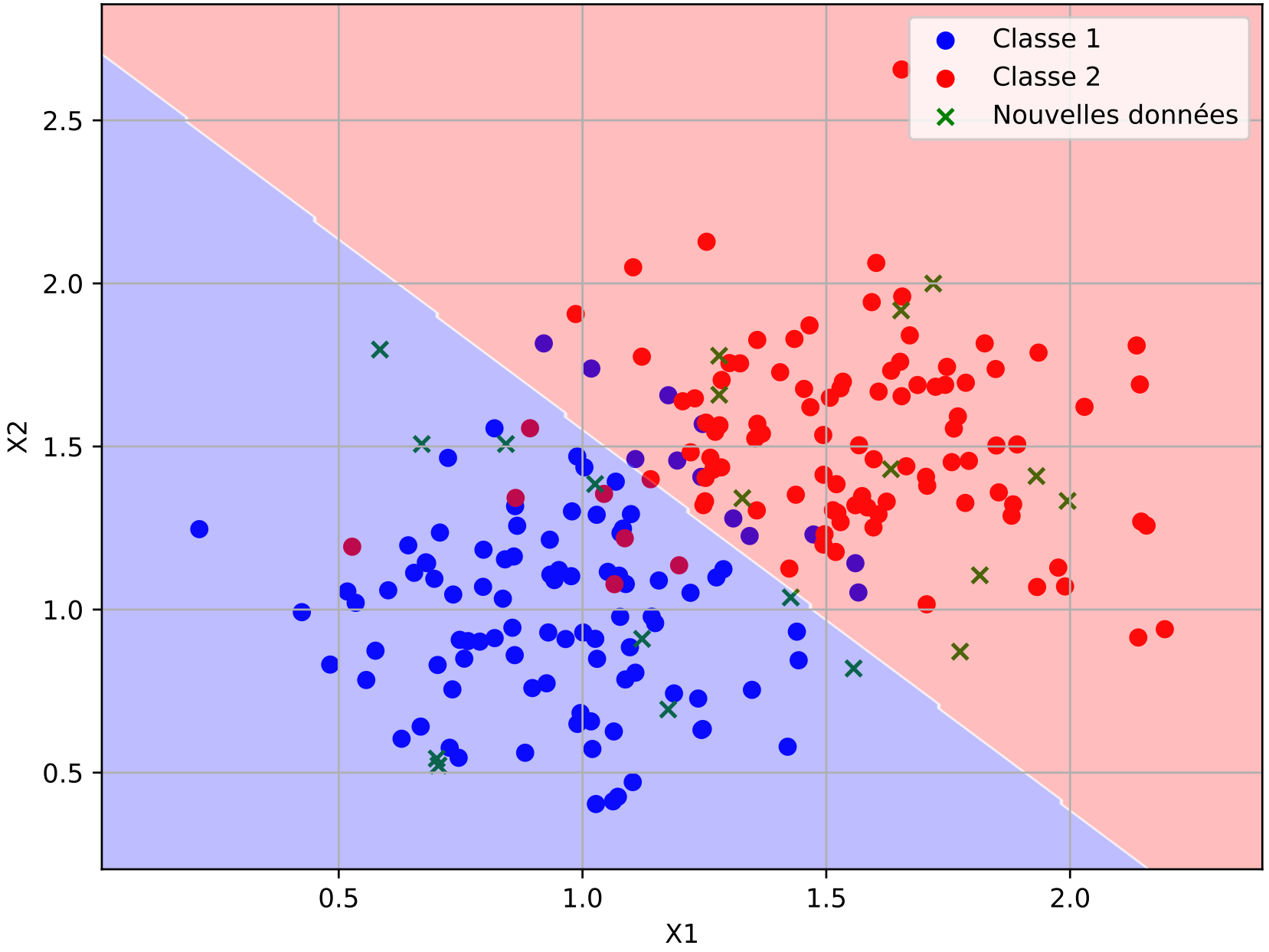
Supervised learning: *classification*



Supervised learning: *classification*



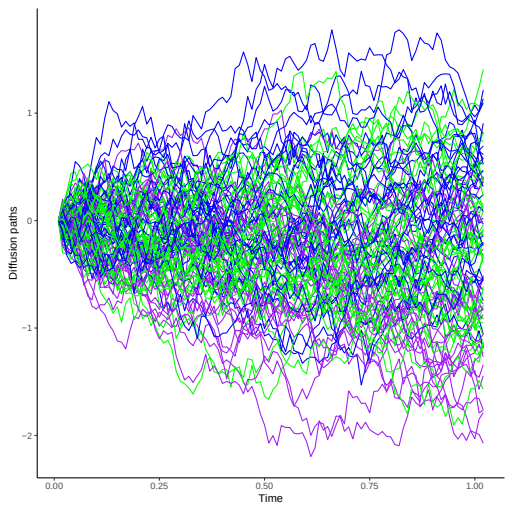
Supervised learning: *classification*



Supervised learning: goal

Supervised learning

- build automatic decision function (algorithm) based on labeled data
- helpful for decision making
- data may be of different nature (image, text, function ...)



Multi-class classification: *setting*

Framework

- observation $X \in \mathcal{X}$ and $Y \in \mathcal{Y} = \{1, \ldots, K\}$
- classifier $f : \mathcal{X} \to \mathcal{Y}$
- misclassification risk $R(f) = \mathbb{P}(f(X) \neq Y)$

Optimal rule

- conditional probabilities $p_k(X) = \mathbb{P}(Y = k \big| X)$
- Bayes classifier $f^*(X) \in \arg\max_{k \in \mathcal{Y}} p_k(X)$
- oracle risk $R^* = R(f^*) = \min_f R(f)$

Goal

- learning sample $(X_i, Y_i)_{1 \leq i \leq n}$ and new observation X_{n+1}
- empirical classification rule $\widehat{f}$ based on the learning sample
- $\hat{f}(X_{n+1})$ prediction of the associated label

Multi-class classification: *plug-in approach*

Plug-in rule

- build $\hat{p}_k$ estimators of p_k
- consider $\hat{f}(X) \in \arg\max_{k \in \mathcal{Y}} \hat{p}_k(X)$

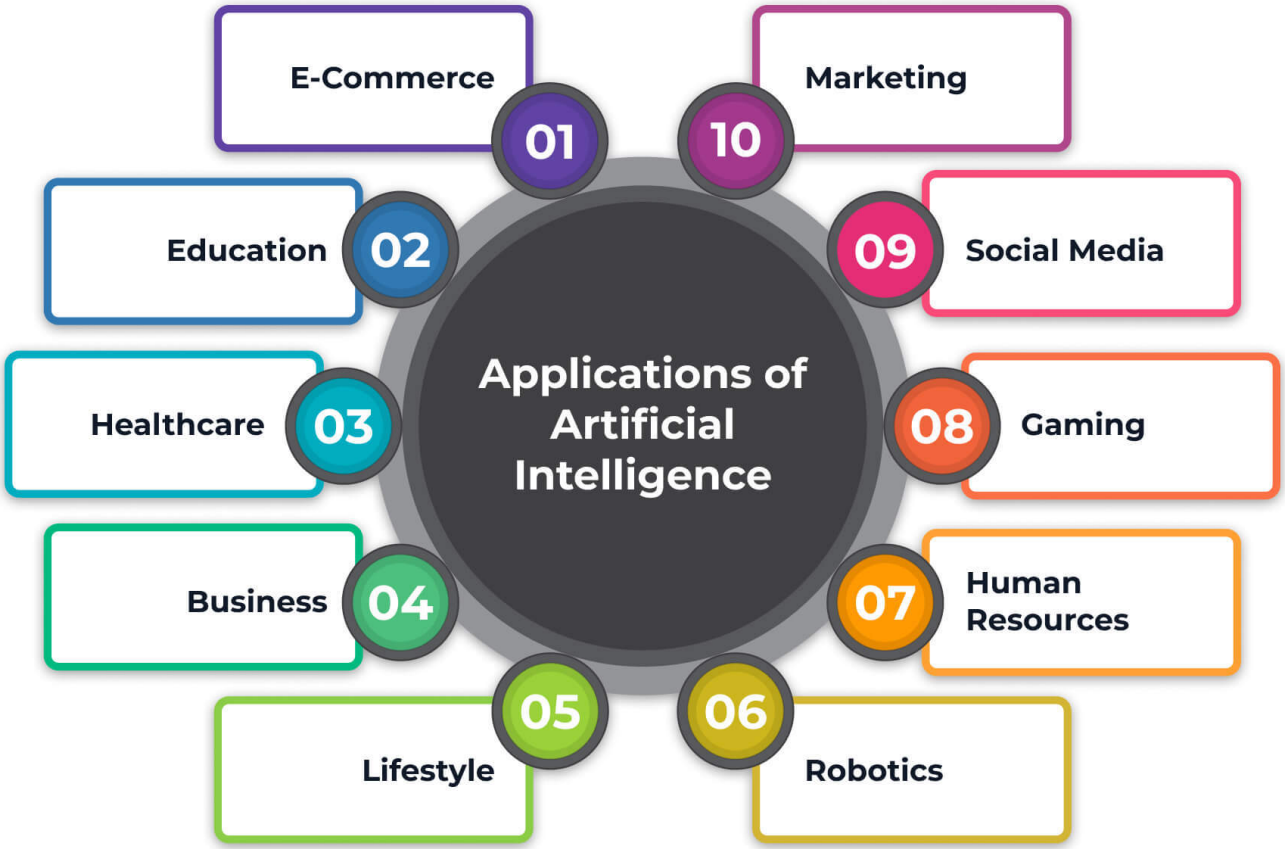
Excess risk

- one can show that

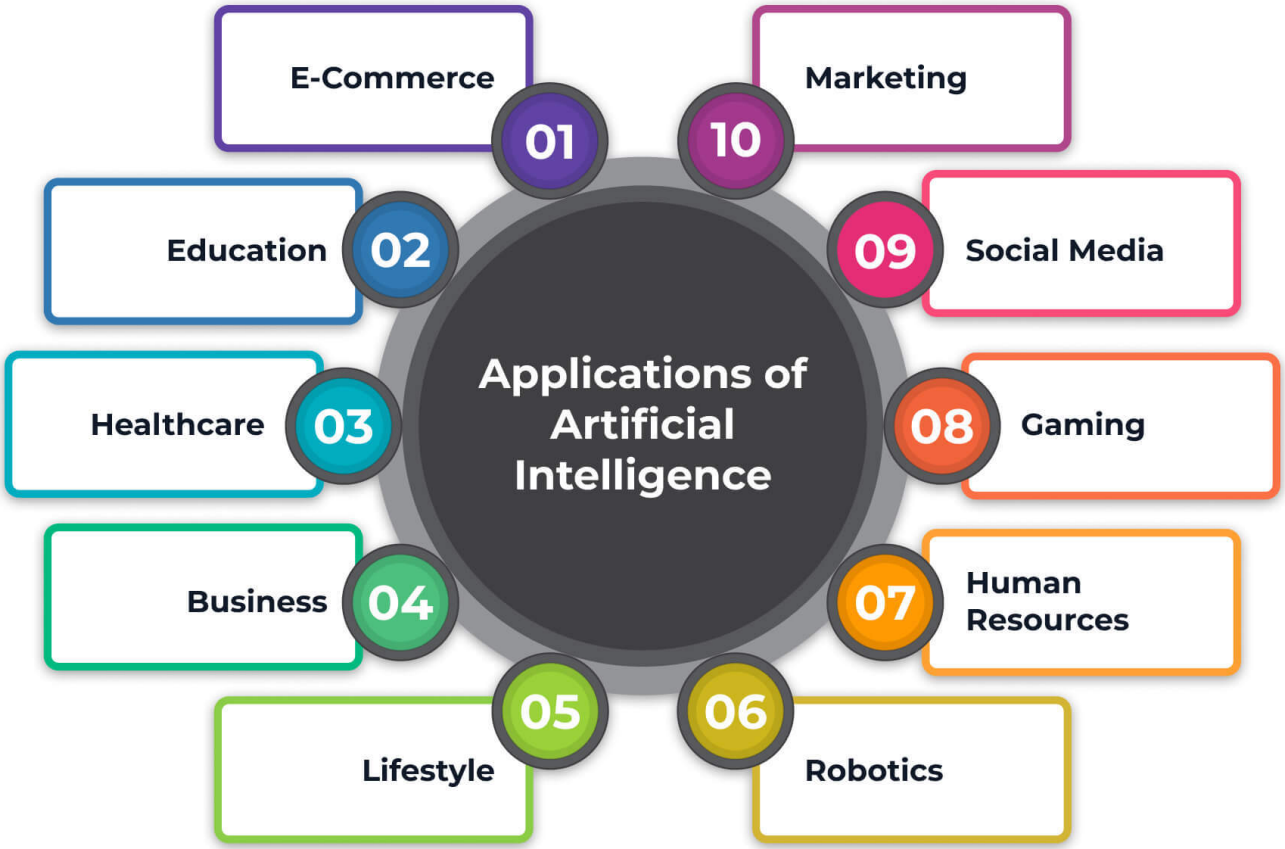
$$\mathbb{E}\left[R(\hat{f})\right] - R^* \leq \sum_{k=1}^{K} \mathbb{E}\left[|\hat{p}_k(X) - p_k(X)|\right]$$

- consistency of $\hat{p}_k \Rightarrow$ consistency of $\hat{f}$
  $\hookrightarrow \quad \mathbb{E}\left[R(\hat{f})\right] \to R^*$

AI with success

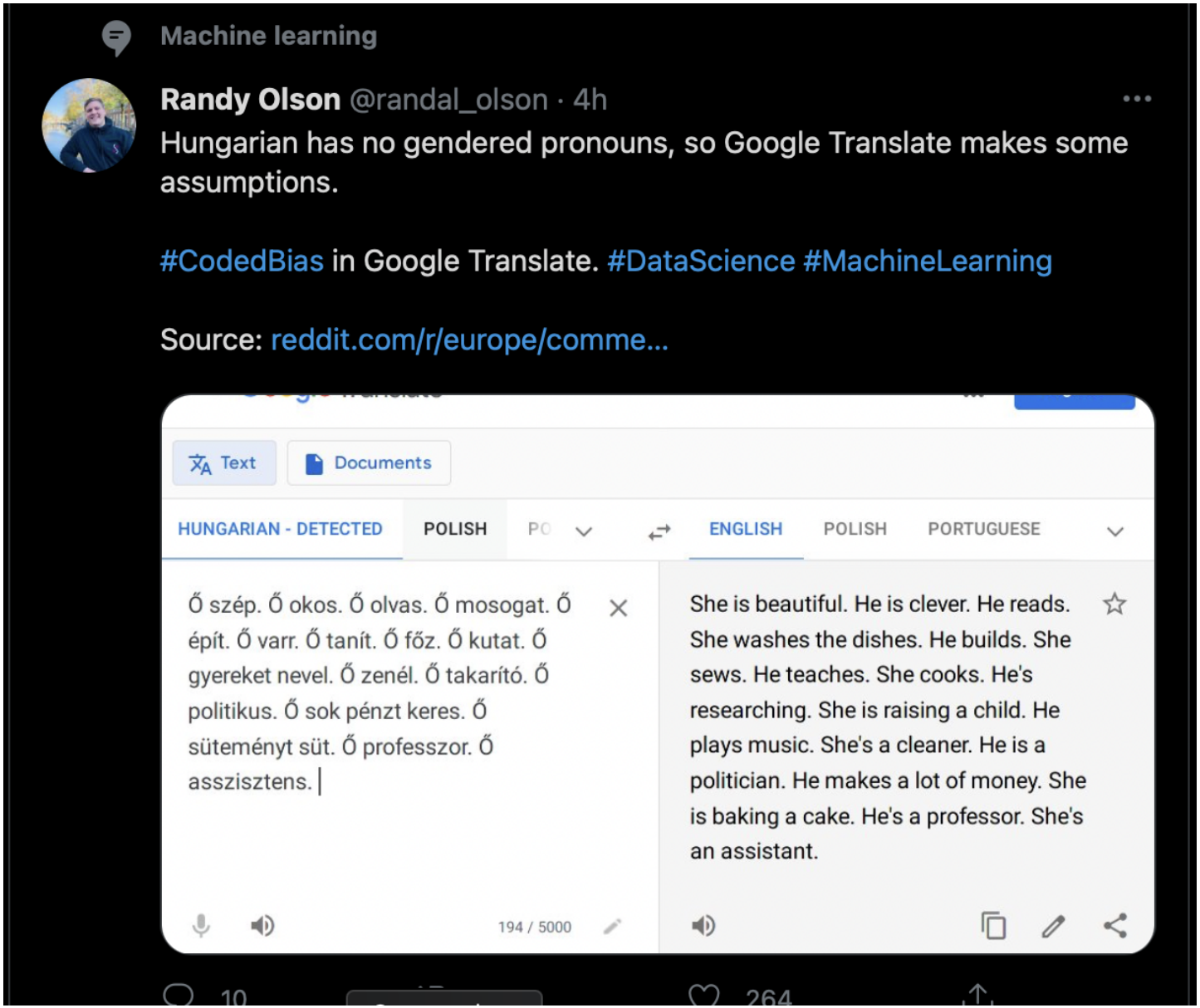


AI with success



But also malicious mistakes...

Bias in machine learning: example (2018)



Historical example: COMPAS dataset

Correctional Offender Management Profiling for Alternative Sanctions

- risk-assessment software developed and owned by Northpointe
- the software is used in U.S. courts to predict recidivism risks of defendants
- investigation of `Propublica` (2016)

> *"Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes."*

And many other fairness failures...

Hiring algorithm

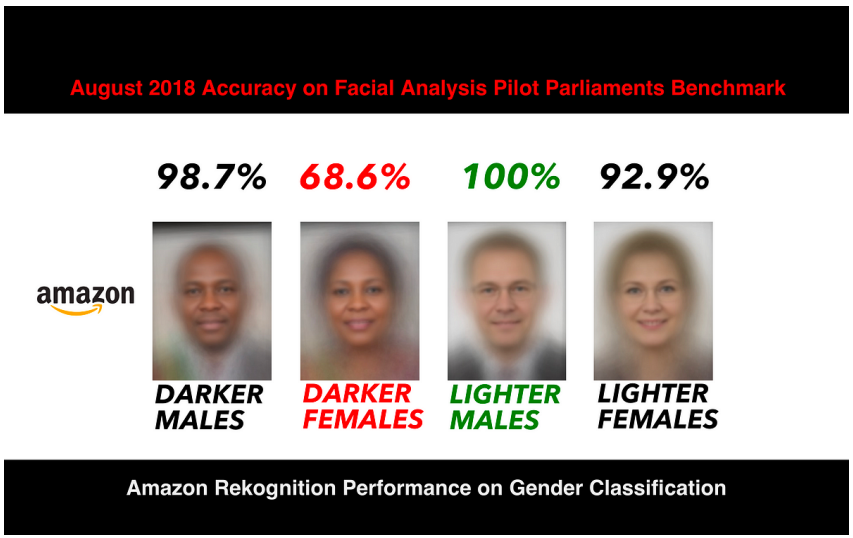
- "AI tools lead to *clear racial disparities* in job hiring" *Financial times (2026)*

Apple Card Credit Limits (2019)

- allegations of gender bias in determining credit limits

Facial Recognition Bias:

- higher error rates for people with darker skin tones



A need for regulation: EU AI Act

Article 5

- prohibits certain AI practices that may lead to harmful or discriminatory treatment
- examples: social scoring and some forms of biometric categorisation

Article 10

- relevant and representative training, validation and test data
- identification and mitigation of biases
- consideration of possible discriminatory impacts

Algorithmic fairness

Motivation

- mitigate the bias contained in historical data
- reduce influence of a sensitive attributes in prediction
- algorithm should treat people without discrimination based on sensitive attributes
- lot of attention in recent years Calders *et al.* (2009), Zemel *et al.* (2013), Zafar *et al.*, Donini *et al.* (2018), Agarwal *et al* (2018), Barocas *et al.* (2019), Chzhen and Schreuder (2022), Sariola *al.* (2026) ...

Application

- social sciences (university admission)
- medical diagnosis
- justice decision
- insurance (credit scoring)

Potential causes of unfairness

Discrimination can come from

- intentional discrimination
- historical biases in learning datasets
- inadvertent bias from data sampling: learning dataset not representative of the target population
- inadvertent bias from algorithm objectives: focus on the benefit for majority group

Formalization of fairness in supervised learning (1/2)

Observation

- features: (X, S), output $Y \in \mathcal{Y}$
- $S \in \mathcal{S}$ sensitive (or protected) attribute

Sensitive attributes

- any decisions based on S are undesirable from an ethical or legal perspective

Remove S does not enforce fairness

- X and S are not independent
- indirect discrimination

Formalization of fairness in supervised learning (2/2)

Fairness through awareness

- predictor $f \rightarrow$ prediction $f(X, S)$
- drawback: prediction relies on S

Fairness through unawarness

- awareness may induce direct discrimination
- predictor $f \rightarrow$ prediction $f(X)$

Disparate learning process

- S available at the training step but not for the prediction step

Definition of fairness: *group fairness approach*

Definition of fairness: two main point of views

- individual fairness $\rightarrow$ *similar* people should be treated *similarly*
- group fairness $\rightarrow$ defines fairness at population level

Definition for group fairness ($Z = (X, S)$ or $Z = X$)

- independence $\rightarrow f(Z) \perp\!\!\!\perp S$
- separation $\rightarrow (f(Z) \perp\!\!\!\perp S) \mid Y$
- in general only one of the constraint could be achieve

Impossibility result

- assume that f satisfies independence and separation
- then either $Y \perp\!\!\!\perp S$ or $f(Z) \perp\!\!\!\perp Y$ or both

Fairness in binary classification: *group fairness approach*

Framework

- observation $Z = (X, S)$ or $Z = X$, and $Y \in \{0, 1\}$,
- classifier $f \to$ prediction $f(Z)$

Definition of fairness

- Demographic parity (DP), for each $s \in \mathcal{S}$

$$\mathbb{P}(f(Z) = 1 | S = s) = \mathbb{P}(f(Z) = 1)$$

- Equalized odds (EOd), for each $s \in \mathcal{S}$, and $y \in \{0, 1\}$

$$\mathbb{P}(f(Z) = y | S = s, Y = y) = \mathbb{P}(f(Z) = y | Y = y)$$

- Equal opportunity (EO), for each $s \in \mathcal{S}$

$$\mathbb{P}(f(Z) = 1 | S = s, Y = 1) = \mathbb{P}(f(Z) = 1 | Y = 1)$$

Main approaches to enforce fairness

Pre-processing **Zemel** *et al.* **(2013), Calmon** *et al.* **(2016)**

- find a feature representation $z \mapsto \phi(z)$
- such that $\phi(Z)$ independent of S

In-processing **Agarwal** *et al* **(2018), Donini** *et al.* **(2018)**

- Based on the empirical risk minimization
- given a set of predictor $\mathcal{F}$, solve

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{R}(f) + \lambda \hat{C}(f),$$

with $\hat{R}(f)$ empirical risk, $\hat{C}(f)$ empirical fairness constraints

Post-processing **Hardt** *et al.* **(2016), Chzhen** *et al.* **(2019)**

- given a pre-built predictor $\hat{f}$, not necessary fair
- find a transformation $\hat{T}$
- *s.t.* $\hat{T}(\hat{f})$ satisfies a desired fairness constraint

Classification under DP constraint: a Post-processing approach

Joint work with: R. Elie (DeepMind), M. Hebiri (U. Gustave Eiffel), F. Hu (Milliman)

Multi-class classification under DP constraint

Notations

- $\mathcal{S} = \{-1, 1\}$, and $\mathcal{Y} = \{1, \dots, K\}$
- $\pi_s = \mathbb{P}(S = s) > 0$, and $p_k(X, S) = \mathbb{P}(Y = k|X, S)$
- classifier $f \to$ prediction $f(X, S) \in \mathcal{Y}$

Problem

- DP constraint, for each $k \in \mathcal{Y}$

$$\sum_{s \in \mathcal{S}} s\mathbb{P}\left(f(X, S) = k|S = s\right) = 0$$

- $f^* \in \arg\min_f \{\mathbb{P}(f(X, S) \neq Y), f \text{ satisfies DP}\}$
- lagrangian associated to the minimization problem

$$\mathcal{R}_\lambda(f) = \mathbb{P}\left(f(X, S) \neq Y\right) + \sum_{k=1}^{K} \lambda_k \sum_{s \in \mathcal{S}} s\mathbb{P}(f(X, S) = k|S = s)$$

Optimal fair classifier

Continuity assumption

- $t \mapsto \mathbb{P}(p_k(X,S) - p_j(X,S) \leq t | S = s)$ is continuous

Optimal predictor

- the optimal fair classifier f^* can be characterized as

$$f^*(x,s) \in \arg\max_k \left(p_k(x,s) - \frac{s}{\pi_s} \lambda_k^* \right)$$

- λ_k^* are lagrange multiplier defined as

$$\lambda^* \in \arg\min_{\lambda \in \mathbb{R}^K} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left[\max_k (\pi_s p_k(X,s) - s\lambda_k) \right]$$

Proposition

Under the continuity assumption, we have

$$f^* \in \arg\min_f \mathcal{R}_{\lambda^*}(f)$$

Data driven procedure: *post-processing approach*

Objective

- estimate $f^*(x,s) \in \arg\max_k \left(p_k(x,s) - \frac{s}{\pi_s} \lambda_k^* \right)$

Plug-in approach

- labeled sample $\rightarrow$ estimate p_k
- unlabeled sample $(X_1, S_1), \ldots, (X_N, S_N)$
- $\{S_1, \ldots, S_N\} \rightarrow$ estimate π_s by their empirical frequencies
- $\{X_1, \ldots, X_N\} \rightarrow$ estimate parameter λ_k^*

Randomization

- fairness guarantee requires continuity assumption
- introduce $\zeta \sim \mathcal{U}_{[0,u]}$ independent of the samples, $u \rightarrow 0$
- $\bar{p}_k(X, S, \zeta) = \hat{p}_k(X, S) + \zeta$

Post-processing estimator: *randomized classifier*

Randomized fair classifier

- $(X_1, \ldots, X_N) \to (X_1^s, \ldots, X_{N_s}^s)$ i.i.d. from $X|S = s$
- estimator $\hat{\lambda}$

$$\hat{\lambda} \in \arg\min_{\lambda \in \mathbb{R}^K} \sum_{s \in \mathcal{S}} \frac{1}{N_s} \sum_{i=1}^{N_s} \max_k \left(\hat{\pi}_s \bar{p}_k(X_i^s, s, \zeta_{k,i}^s) - s\lambda_k \right)$$

- resulting classifier

$$\hat{f}(x, s) \in \arg\max_{k \in \mathcal{Y}} \left(\bar{p}_k(x, s, \zeta_k) - \frac{s}{\hat{\pi}_s} \hat{\lambda}_k \right)$$

Theoretical guarantees

Unfairness measure

$$\mathcal{U}(f) = \max_k |\mathbb{P}\left(f(X,S) = k | S = 1\right) - \mathbb{P}\left(f(X,S) = k | S = -1\right)|$$

Distribution free-result

There exists C depending only on K and π_s such that for any estimator $\hat{p}_k$

$$\mathbb{E}\left[\mathcal{U}(\hat{f})\right] \leq CN^{-1/2}$$

Utility guarantee

Assume that initial estimators $\hat{p}_k$ are consistent and $u \to 0$, then

$$\mathbb{E}\left[R(\hat{f})\right] \to R(f^*) \;\; n, N \to +\infty.$$

Extension to ε-fairness (1/2)

Approximate fairness: ε-DP

- f is ε-fair iff $\mathcal{U}(f) \leq \varepsilon$

Optimal ε-fair classifier

- $f_\varepsilon^* \in \arg\min_f \{\mathbb{P}(f(X,S) \neq Y), f \text{ satisfies } \varepsilon - \text{DP}\}$
- $\left(\lambda^{*(1)}, \lambda^{*(2)}\right)$ minimizer of

$$\sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left[\max_k \left(\pi_s p_k(X,s) - s\left(\lambda_k^{(1)} - \lambda_k^{(2)}\right)\right)\right] + \varepsilon \sum_{k=1}^{K} \left(\lambda_k^{(1)} + \lambda_k^{(2)}\right)$$

- $f_\varepsilon^*(x,s) \in \arg\max_k \left(p_k(x,s) - \frac{s}{\pi_s}\left(\lambda_k^{*(1)} - \lambda_k^{*(2)}\right)\right)$

Extension to ε-fairness (2/2)

Optimal ε fair predictor: properties

- $\lambda_k^{*(1)} \lambda_k^{*(2)} = 0$ and $\lambda_k^{*(1)} + \lambda_k^{*(2)} \geq 0$, $k \in [K]$
- if $\mathcal{U}\left(f^*_{\text{Bayes}}\right) \leq \varepsilon$ then $f^*_\varepsilon = f^*_{\text{Bayes}}$, and $\lambda^{*(1)} = \lambda^{*(2)} = 0$
- else $\mathcal{U}\left(f^*_\varepsilon\right) = \varepsilon$

Estimation

- same procedure as for exact fairness
- $\mathbb{E}\left[\mathcal{U}(\hat{f}_\varepsilon)\right] \leq \varepsilon + CN^{-1/2}$
- fairness and risk guarantees

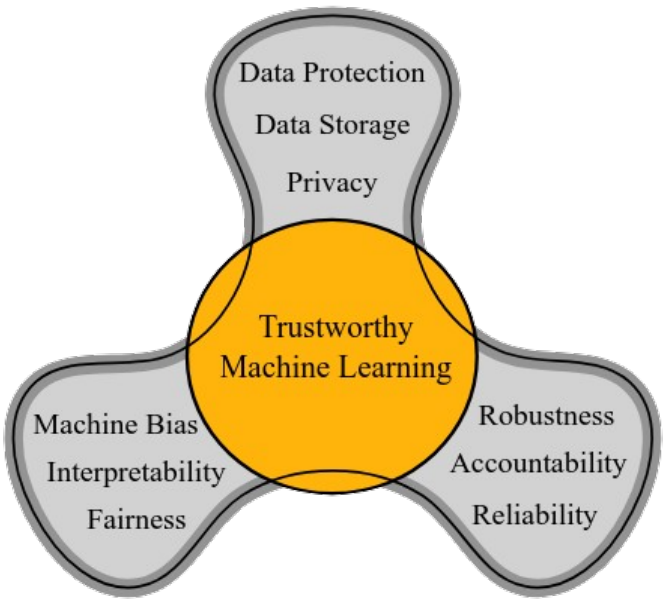
Extension: Integrating Differential Privacy and Demographic Parity in Multi-class Classification

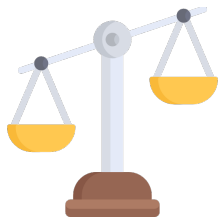
Joint work with: R. Pinot (Sorbonne Université), Lilian Say (Sorbonne Université)

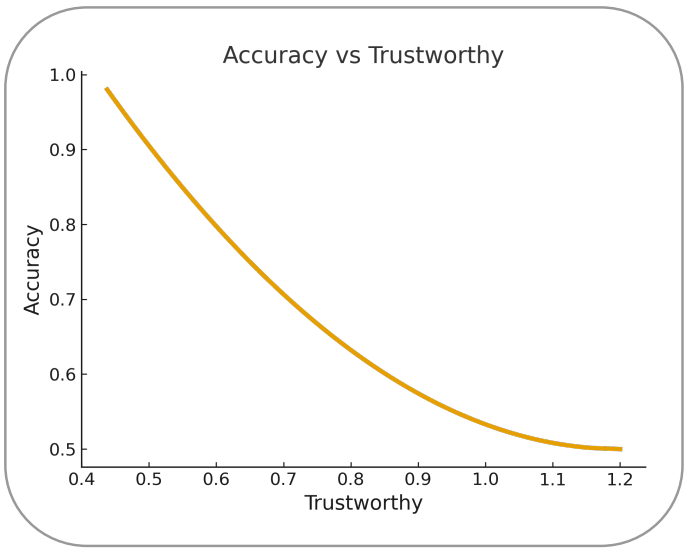
Trustworthy machine learning

Goal

- include other constraints: privacy, robustness
- the challenge is to balance these competing objectives







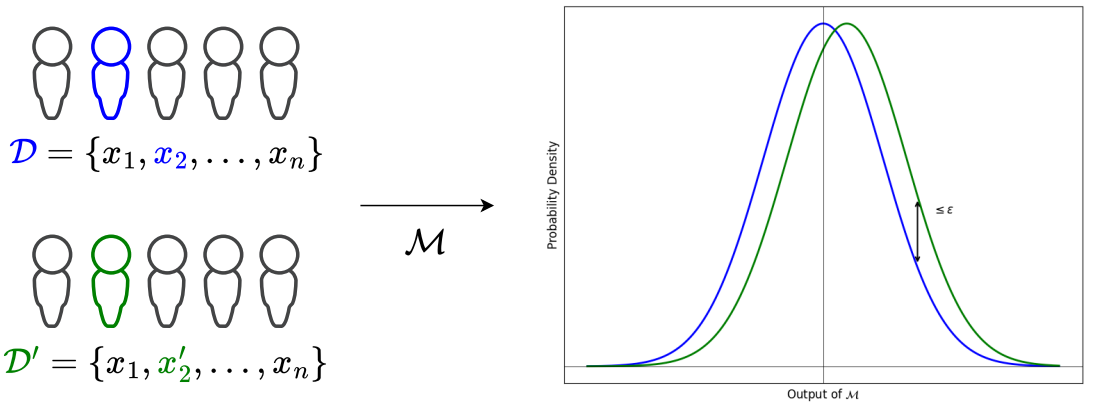
Differential privacy

Privacy risk in machine learning

- even simple models can leak training data
- gradient leakage in federated learning
- membership inference and data reconstruction attacks

Differential Privacy **Dwork** *et al (2006)*

- distribution does not change much when one record is changed
- add random noise to computations where data are used
  $\hookrightarrow$ Differentially Private SGD (DP-SGD)



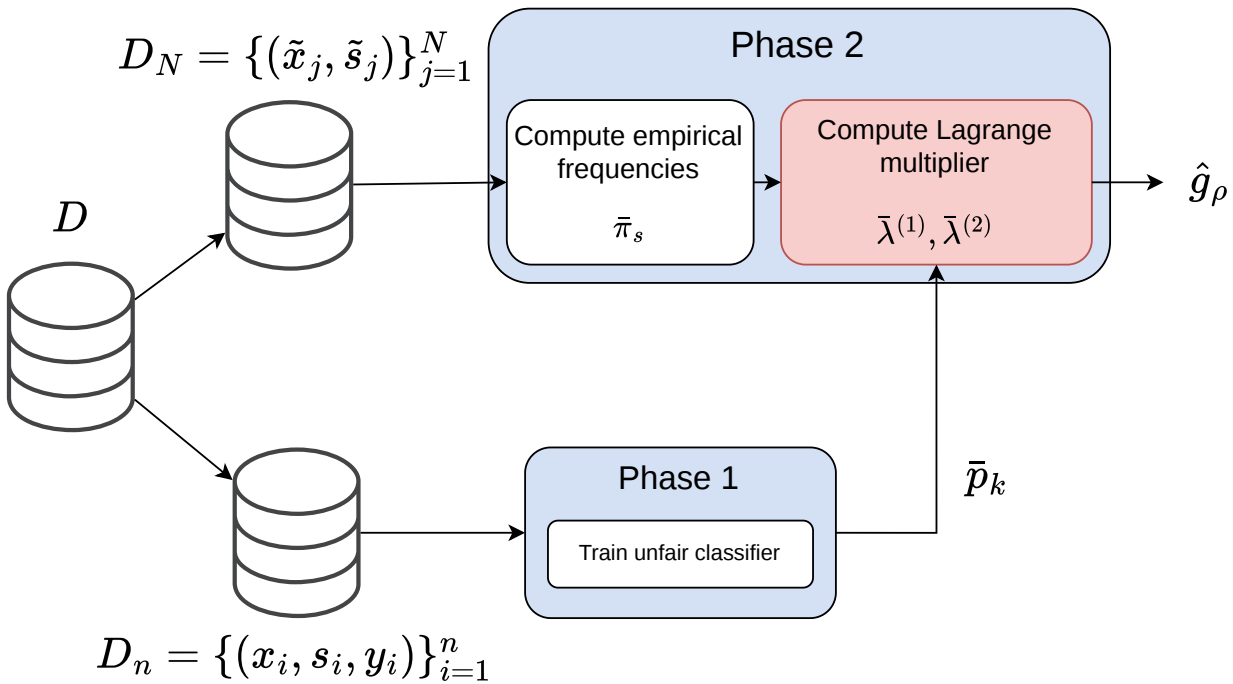
$\mathcal{D} \approx \mathcal{D}'$ are neighboring datasets

$D_\alpha(\mathrm{Law}(\mathcal{M}(\mathcal{D})) \parallel \mathrm{Law}(\mathcal{M}(\mathcal{D}'))) \leq \varepsilon$

D2DP algorithm

D2DP

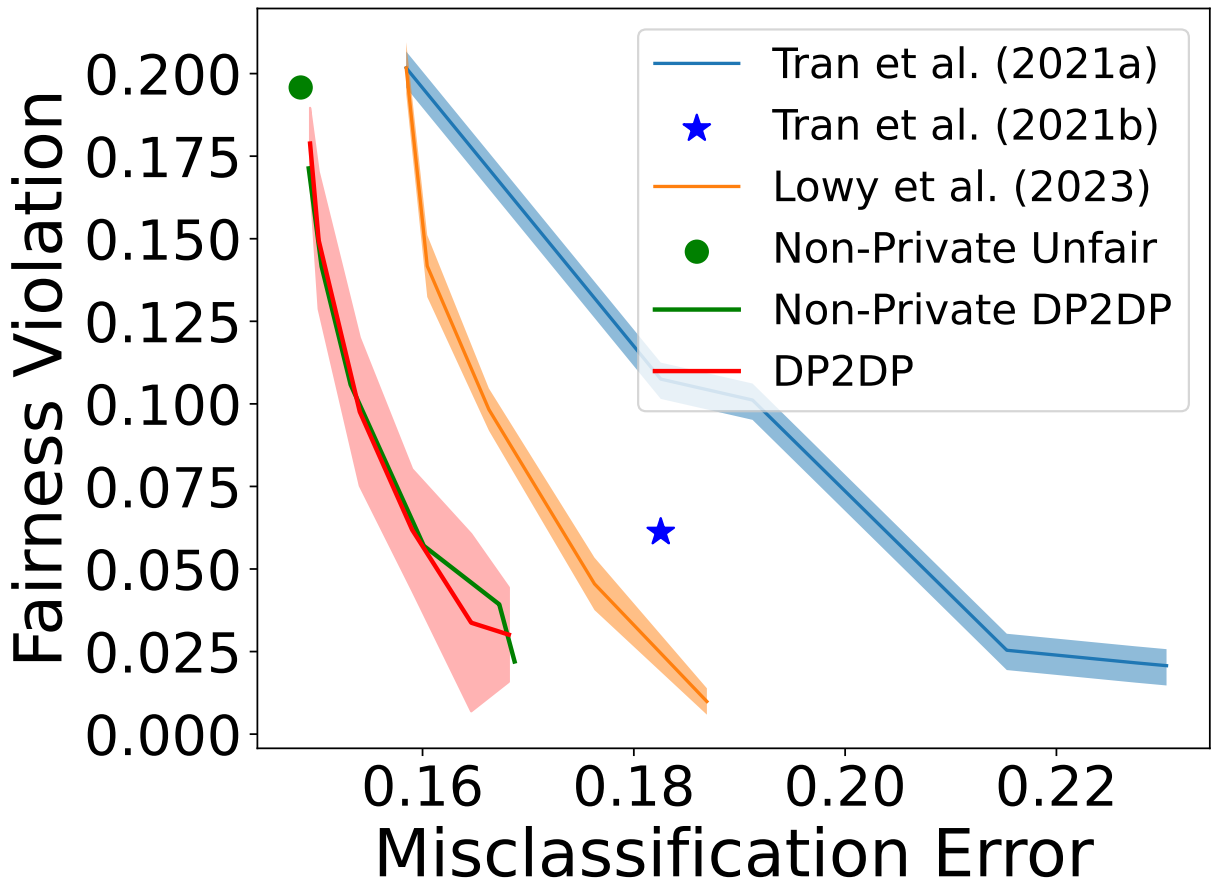
- bridge two well-studied requirements in machine learning
- differential privacy and demographic parity
- privacy increases the fairness error by only a log factor!



Numerical experiments

Adult dataset

- $n = 48842$
- logistic regression
- data split: *55% train, 20% post-processing, 25% test*



Perspective

DP multiclass classification

- exact and ε-fairness
- post-processing approach
- inclusion differential privacy constraint
- fairness and risk guarantee

Some extension

- extension to prediction without sensitive attribute
- extension to other fairness measures
- extension to federated learning setting and robustness