

# Security and Ethical Aspects of Data

Amaya Nogales Gómez  
[amaya.nogales-gomez@univ-cotedazur.fr](mailto:amaya.nogales-gomez@univ-cotedazur.fr)

MSc Data Science & Artificial Intelligence  
Université Côte d'Azur

October 11, 2021

# The course

## Prerequisites

- Familiarity with some machine learning, basic statistics and probability theory will be helpful.
- Being comfortable with mathematical notation and formalism.
- Basic programming: there will be some simple coding and data analysis assignments.

## What to expect?

This course introduces the ethical aspects of artificial intelligence (AI), addressing the concerns raised by the increased use of AI to make decisions that have important consequences on people's lives. In particular, the course focuses on fundamental concepts and methods of fairness in Machine Learning (ML).

# Course Overview

## ① Logistics

- 7 sessions: 6 lectures/labs (3h) + final exam (2h).
- Attendance is mandatory.
- Lab reports are graded.
- Recap and/or multiple-choice-question beginning of each lecture.

# Course Overview

## ① Logistics

- 7 sessions: 6 lectures/labs (3h) + final exam (2h).
- Attendance is mandatory.
- Lab reports are graded.
- Recap and/or multiple-choice-question beginning of each lecture.

## ② Materials

- No textbook required.
- Laptop.
- Suggested readings and online resources will be posted on Slack.

# Course Overview

## ① Logistics

- 7 sessions: 6 lectures/labs (3h) + final exam (2h).
- Attendance is mandatory.
- Lab reports are graded.
- Recap and/or multiple-choice-question beginning of each lecture.

## ② Materials

- No textbook required.
- Laptop.
- Suggested readings and online resources will be posted on Slack.

## ③ Assessment

- Final Exam (60%).
- Labs (40%).

# Course Overview

## ① Logistics

- 7 sessions: 6 lectures/labs (3h) + final exam (2h).
- Attendance is mandatory.
- Lab reports are graded.
- Recap and/or multiple-choice-question beginning of each lecture.

## ② Materials

- No textbook required.
- Laptop.
- Suggested readings and online resources will be posted on Slack.

## ③ Assessment

- Final Exam (60%).
- Labs (40%).

## ④ Communication

- Email: amaya.nogales-gomez@i3s.unice.fr
- Slack.
- Office: 421, Templiers 1.

# Tentative Content

## ① Introduction

- Preliminaries and Machine Learning basics
- Motivation for ethics in Artificial Intelligence

## ② Sources of unfairness

- Bias in data
- Algorithmic unfairness: examples

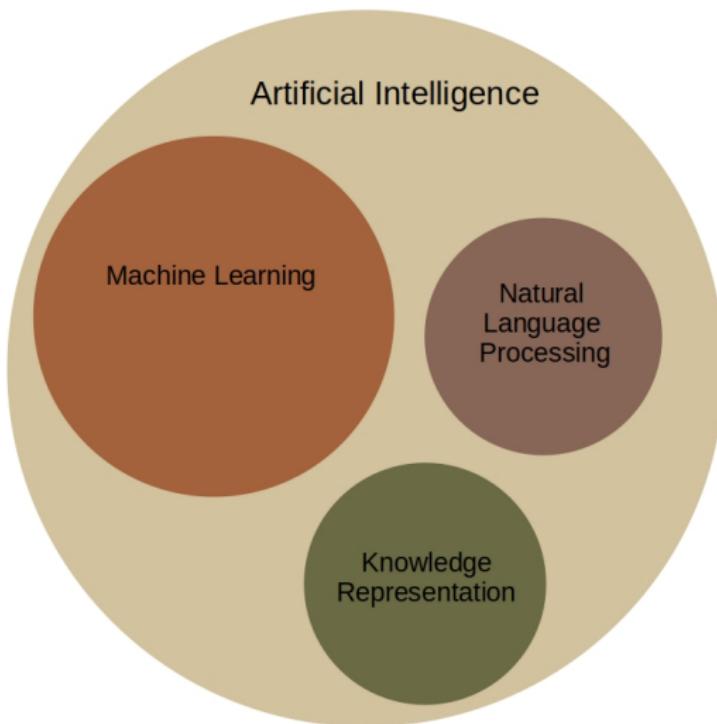
## ③ Fairness criteria

- Types of discrimination
- Definitions of fairness

## ④ Algorithms and methods for fair ML

- Pre-processing methods
- In-processing methods
- Post-processing methods
- Legal and policy perspectives

# What is Machine Learning?



# Example: spam filtering



Non-SPAM



SPAM

# Example: recommender systems

Spotify

< >

### Good afternoon

### Your shows

Nadie Sabe Nada  
SER Podcast

La Vida Moderna  
SER Podcast

El Grupo  
SER Podcast

Estirando el chicle  
Podium Podcast

El Bar de los Broder...  
Spotify Studios

### Jump back in

Novedades Carminha  
Novedades Carminha

Music To Sleep By: ...  
Various Artists

Las Corraleras de Lebrija  
Artist

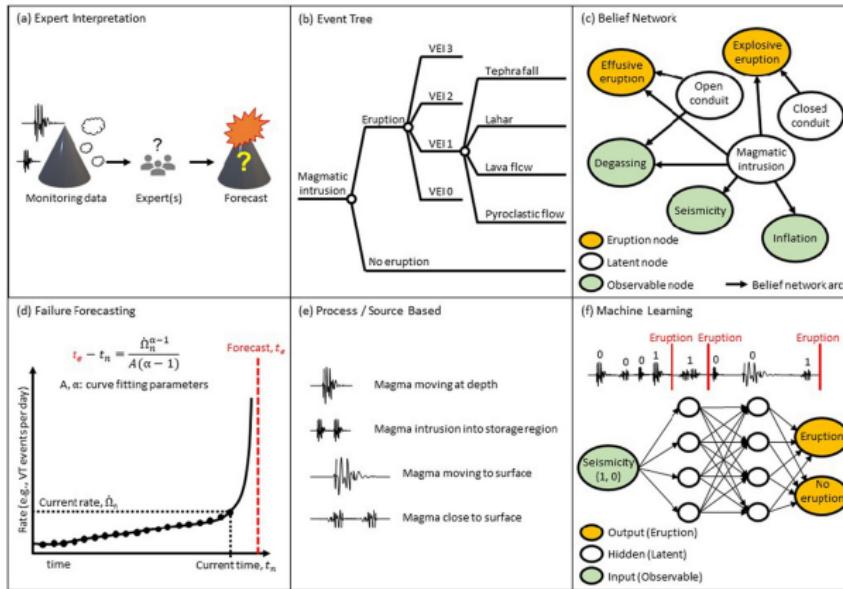
Malika Makovski  
Artist

Ελληνικά Παιδικά Τραγούδια  
Παιδικά Τραγούδια

# Example: clustering



# Examples: forecasting



M. G. Whitehead, M. S. Bebbington, Method selection in short-term eruption forecasting,  
*Journal of Volcanology and Geothermal Research.*

# Examples at Université Côte d'Azur

- Healthcare
  - Medical diagnosis & prevention.
- Industry
  - Recommender systems: music, video.
  - Image storage into synthetic DNA.
- Multimedia
  - Speech detection in political debates.
  - Cultural, lyrics and audio analysis from music.

## Artificial Intelligence

The science and engineering of making intelligent machines, especially computer systems by reproducing human intelligence through learning, reasoning and self-correction/adaption. [McCarthy89]

## Machine Learning

A computer program (algorithm) that improves its performance measure P at some class of tasks T with experience E. [Mitchell90]

Field of study that gives computers the ability to learn without being explicitly programmed. [Samuel59]

A. Samuel, "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development. 3 (3), 1959.

T. Mitchell, B. Buchanan, G. DeJong, T. Dietterich, P. Rosenbloom, A. Waibel, Machine learning, Annual review of computer science 4 (1) (1990) 417-433.

J. McCarthy, Artificial intelligence, logic and formalizing common sense, in: Philosophical logic and artificial intelligence, Springer, 1989, pp. 161-190.

# The origins of AI

## 1956 Dartmouth Summer Research Project on Artificial Intelligence

### A Proposal for the

#### DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

June 19 - Aug. 16

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

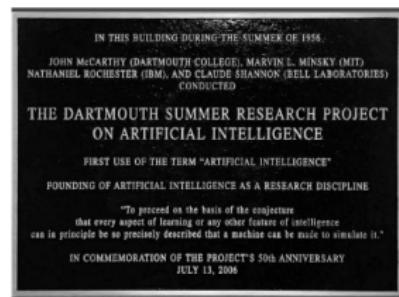
The following are some aspects of the artificial intelligence problem:

#### 1) Automatic Computers

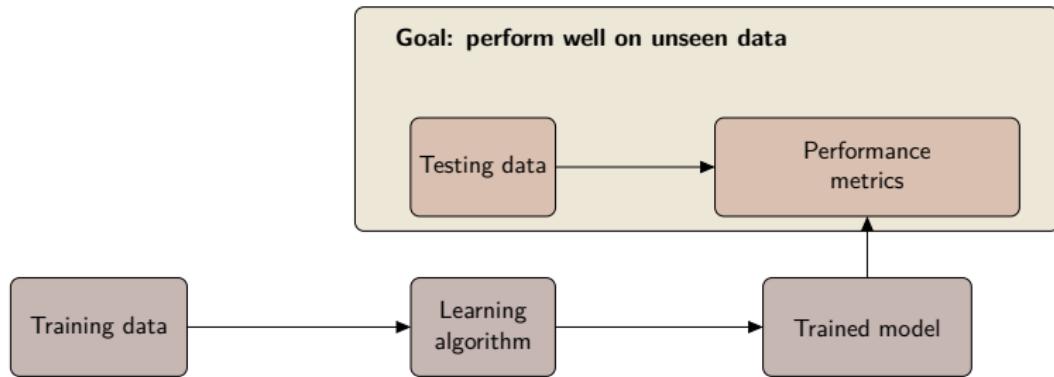
If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

#### 2) How Can a Computer be Programmed to Use a Language

It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning



# Baseline ML approach



# Challenges in ML

- What kind of data to use?
- How much data is enough?
- How to represent it?
- Which algorithm should be used?
- How to choose the best model?
- Performance guarantees
- Explainability
- Interpretability

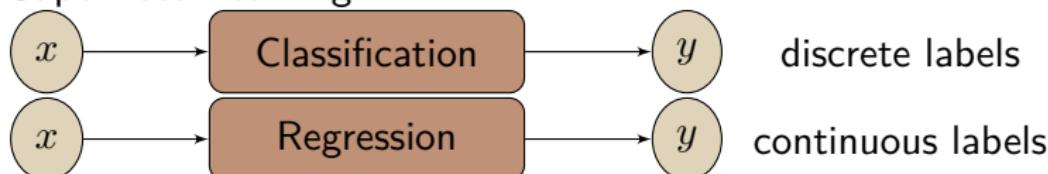
How to model a problem as a Machine Learning problem?

# Machine Learning algorithms

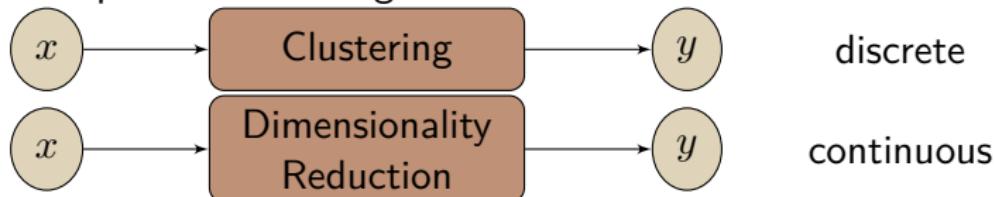
- Supervised Learning
  - Training data include desired outputs
  - Dataset comprised of labeled examples
- Unsupervised Learning
  - Training data does not include desired outputs
  - Find structure in some examples (no labels!)
- Reinforcement Learning
  - Rewards from sequence of actions
  - Feedback-based sequential decision making

# Supervised learning

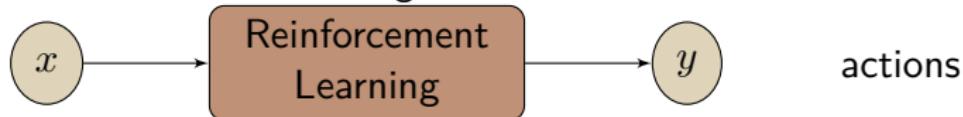
## Supervised Learning



## Unsupervised Learning



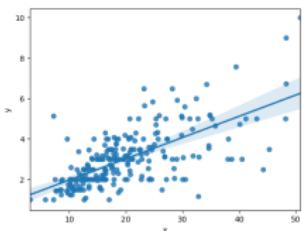
## Reinforcement Learning



# Supervised Learning: Regression and classification

## Regression

Output: continuous function.

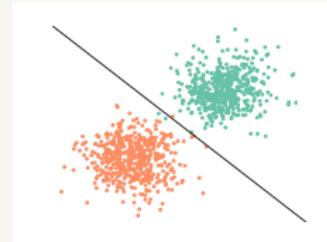


Examples:

- Forecasting
- Size of animal
- Stocks

## Classification

Output: separation rule.



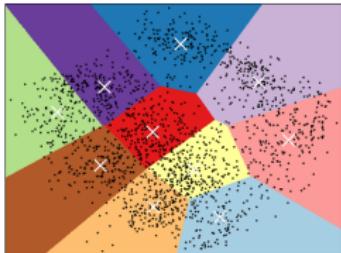
Examples:

- Pay back a loan
- University acceptance
- Image classification

# Unsupervised learning

## Clustering

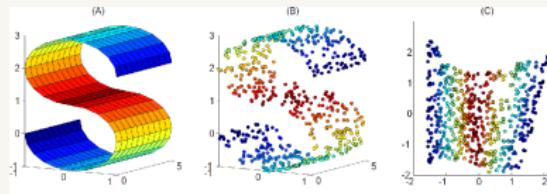
K-means clustering on the digits dataset (PCA-reduced data)  
Centroids are marked with white cross



## Examples:

- Image classification
- Recommender systems
- Social Networks Analysis

## Dimensionality reduction



## Examples:

- Data visualization
- Data storage
- Computational complexity

## Clustering

Find subtypes or groups that are not defined a priori based on measurements.

**unsupervised learning**

## Classification

Use a priori group labels in analysis to assign new observations to a particular group or class.

**supervised learning**

# General goals of clustering

- ① Observations within a cluster are similar

*compactness property*

# General goals of clustering

- ① Observations within a cluster are similar  
*compactness property*
- ② Observations in different clusters are non similar  
*closeness property*

# General goals of clustering

- ① Observations within a cluster are similar  
*compactness property*
- ② Observations in different clusters are non similar  
*closeness property*

Goal: obtain compact clusters that are well-separated

# Supervised Classification: Support Vector Machines

- $\Omega$ : the population.
- Population is partitioned into two classes,  $\{-1, +1\}$ .

# Supervised Classification: Support Vector Machines

- $\Omega$ : the population.
- Population is partitioned into two classes,  $\{-1, +1\}$ .
- For each object in  $\Omega$ , we have
  - $x = (x^1, \dots, x^d) \in X \subset \mathbb{R}^d$ : predictor variables.
  - $y \in \{-1, +1\}$ : class membership.

# Supervised Classification: Support Vector Machines

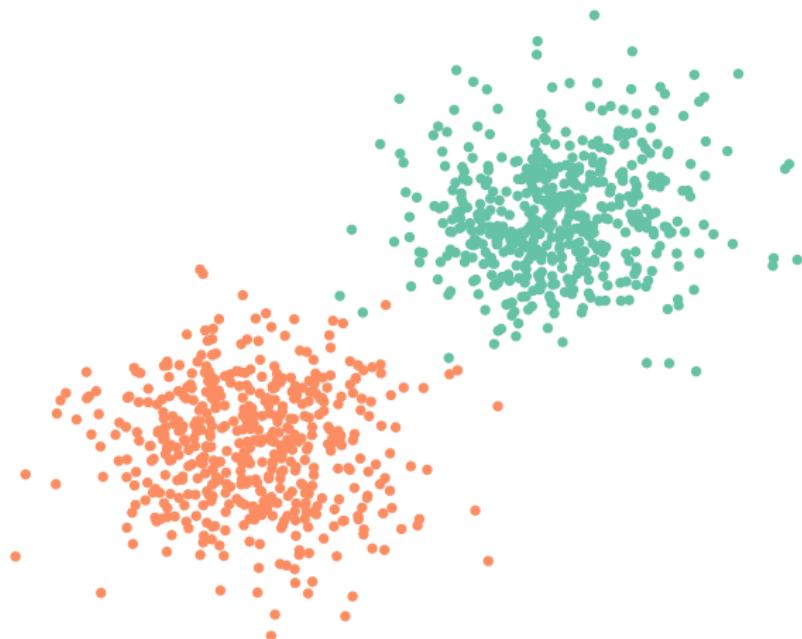
- $\Omega$ : the population.
- Population is partitioned into two classes,  $\{-1, +1\}$ .
- For each object in  $\Omega$ , we have
  - $x = (x^1, \dots, x^d) \in X \subset \mathbb{R}^d$ : predictor variables.
  - $y \in \{-1, +1\}$ : class membership.
- The goal is to find a hyperplane  $\omega^\top x + b = 0$  that aims at separating, if possible, the two classes.

# Supervised Classification: Support Vector Machines

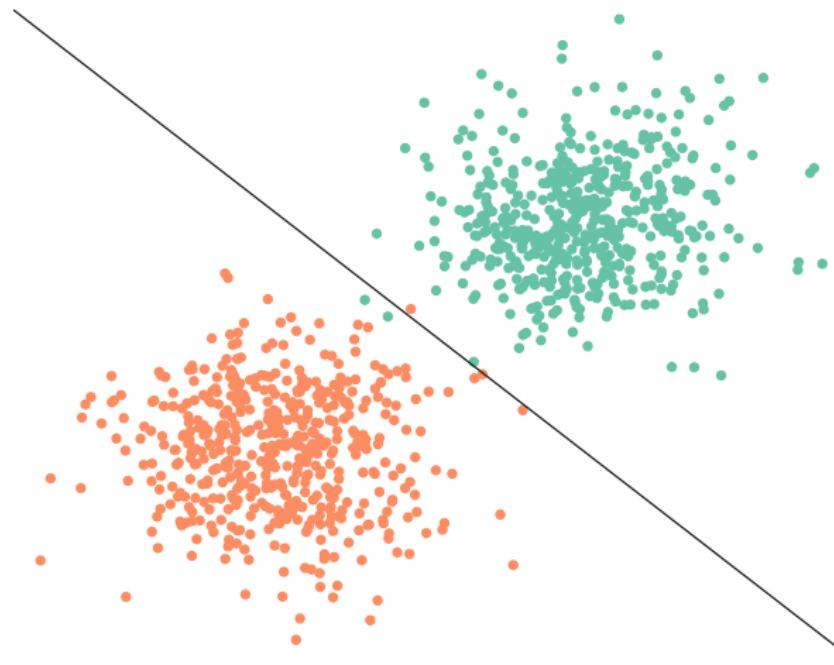
- $\Omega$ : the population.
- Population is partitioned into two classes,  $\{-1, +1\}$ .
- For each object in  $\Omega$ , we have
  - $x = (x^1, \dots, x^d) \in X \subset \mathbb{R}^d$ : predictor variables.
  - $y \in \{-1, +1\}$ : class membership.
- The goal is to find a hyperplane  $\omega^\top x + b = 0$  that aims at separating, if possible, the two classes.
- Future objects will be classified as

$$\begin{aligned}y &= +1 && \text{if } \omega^\top x + b > 0 \\y &= -1 && \text{if } \omega^\top x + b < 0\end{aligned}\tag{1}$$

# Supervised Classification



# Supervised Classification



# Support Vector Machines (SVM)

- State-of-the-art in supervised classification
- Very good classification accuracy
- Computationally cheap: Quadratic Programming formulation

# Hard-Margin approach

- Training sample assumed to be linearly separable, i.e., the convex hull of the two groups are not empty and they do not overlap,
- All objects in the training sample must be correctly classified!
- The separating hyperplane is the one maximizing the smallest distance to misclassification.

# Hard-margin SVM

$$\max_{\omega, b} \min_i \frac{y_i(\omega^\top x_i + b)}{\|\omega\|^\circ}$$

s.t.

$$\begin{aligned} y_i(\omega^\top x_i + b) &> 0 & \forall i = 1, \dots, n \\ \omega &\in \mathbb{R}^d \setminus 0 \\ b &\in \mathbb{R}, \end{aligned}$$

where  $\min_i \frac{y_i(\omega^\top x_i + b)}{\|\omega\|^\circ}$  denotes the distance of  $x_i$  to the hyperplane and  $\|\cdot\|^\circ$  denotes the dual of  $\|\cdot\|$ , i.e.,  $\|\rho\|^\circ = \max\{\rho x : \|x\| = 1\}$ .

## Hard-margin SVM

$$\begin{aligned} & \max_{\omega, b} \frac{1}{\|\omega\|^\circ} \\ \text{s.t.} \quad & y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n \\ & \omega \in \mathbb{R}^d \setminus 0 \\ & b \in \mathbb{R}, \end{aligned}$$

$$\begin{aligned} & \min_{\omega, b} \|\omega\|^\circ \\ \text{s.t.} \quad & y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n \\ & \omega \in \mathbb{R}^d \\ & b \in \mathbb{R}, \end{aligned}$$

# Hard-margin SVM: final formulation

The objective function can be replaced by  $\Phi(\|\omega\|^\circ)$  for any  $\Phi$ , increasing in  $\mathbb{R}^+$ .

Taking  $\Phi(t) = \frac{1}{2}t^2$  one obtains an equivalent formulation as a quadratic problem with linear constraints.

$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

s.t.

$$\begin{aligned} y_i(\omega^\top x_i + b) &\geq 1 & \forall i = 1, \dots, n \\ \omega &\in \mathbb{R}^d \\ b &\in \mathbb{R}. \end{aligned}$$

But...

Linearly separable data

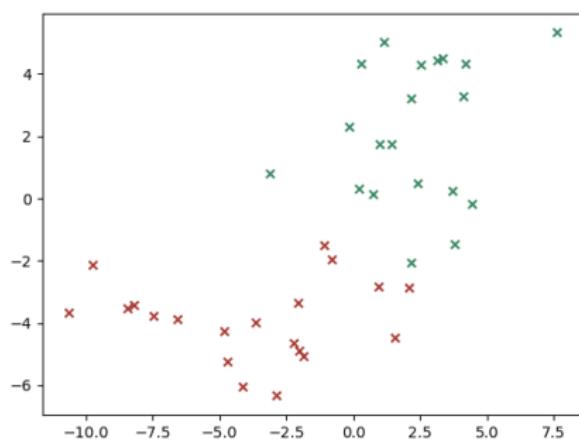
$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$



But...

$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

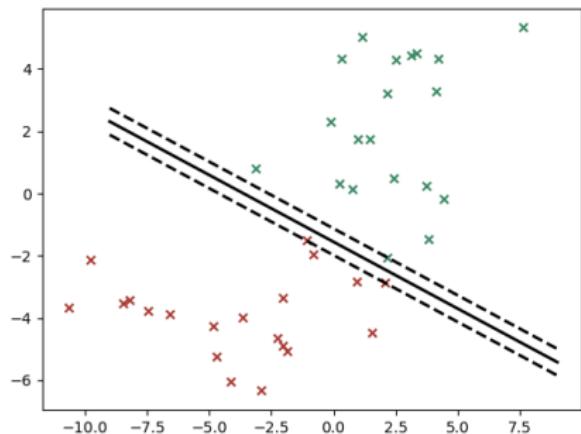
s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$

Linearly separable data



But...

$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

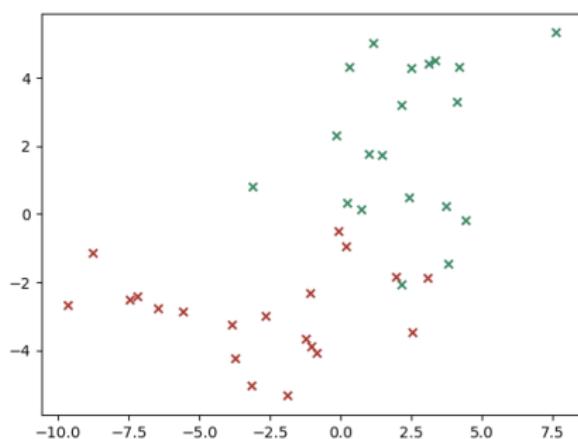
s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$

Non-linearly separable data



But...

Non-linearly separable data

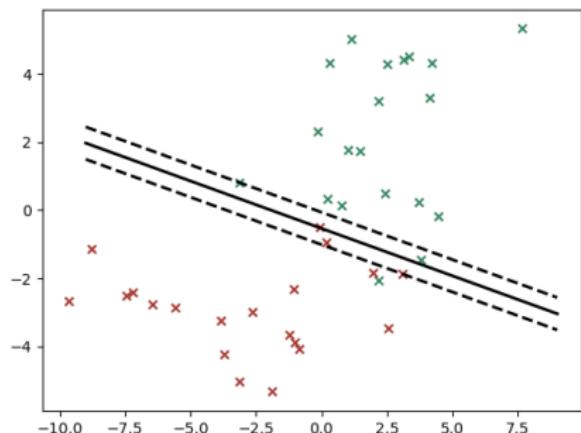
$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$



INFEASIBLE!!

# A solution for non-linearly separable data

- When data are not linearly separable the hard-margin SVM problem is infeasible.
- In the **soft-margin approach**, constraints

$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

are perturbed.

- How? By introducing auxiliary variables  $\xi_i$ , making the new problem always feasible.

# Building the soft-margin SVM

$$\min_{\omega, b, \xi} \frac{1}{2} \sum_{j=1}^d \omega_j^2 + \frac{C}{n} \sum_{i=1}^n g_i(\xi_i)$$

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}$$

$$\xi_i \geq 0 \quad \forall i = 1, \dots, n,$$

- $\xi = (\xi_i) \in \mathbb{R}^n$  is the vector of deviation variables.
- $g_i$  is the loss function (convex and increasing).
- Most popular choices: *hinge loss*,  $g_i(t) = C_i t$  or *squared hinge loss*,  $g_i(t) = C_i t^2$ .
- $C$  is a tuning parameter.

# The SVM formulation

Non-linearly separable data

$$\min_{\omega, b, \xi} \frac{1}{2} \sum_{j=1}^d \omega_j^2 + \frac{C}{n} \sum_{i=1}^n \xi_i$$

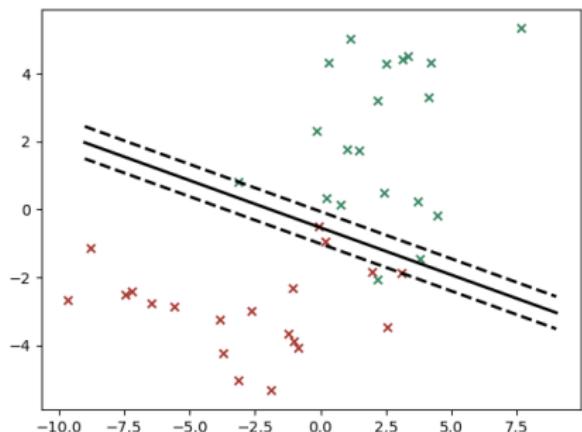
s.t.

$$y_i(\omega^\top x_i + b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}$$

$$\xi_i \geq 0 \quad \forall i = 1, \dots, n.$$



- An object  $i$  will be correctly classified if  $0 \leq \xi_i < 1$
- Misclassified if  $\xi_i > 1$ .
- In the case  $\xi_i = 1$ , we get a tie (objects coincide with the hyperplane).
- $\sum_{i=1}^n \xi_i$  is an upper bound of the number of misclassified objects.

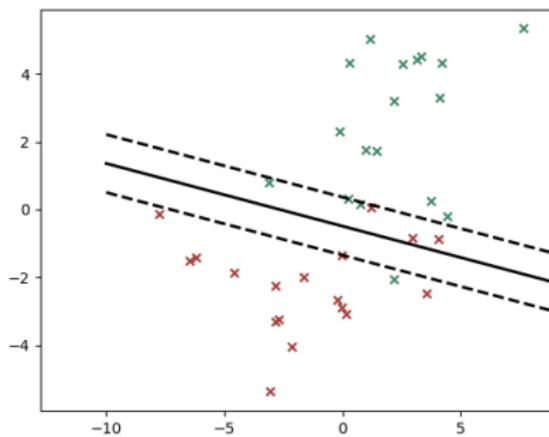
# Quality of a classifier

## Classification Accuracy

Given an object  $i$ , it is classified in the positive or the negative class according to the value of the score function,  $\text{sign}(\omega^\top x_i + b)$ , while for the case  $\omega^\top x_i + b = 0$ , the object is classified randomly. The classification accuracy is defined as the percentage of objects correctly classified by the classifier on such dataset.

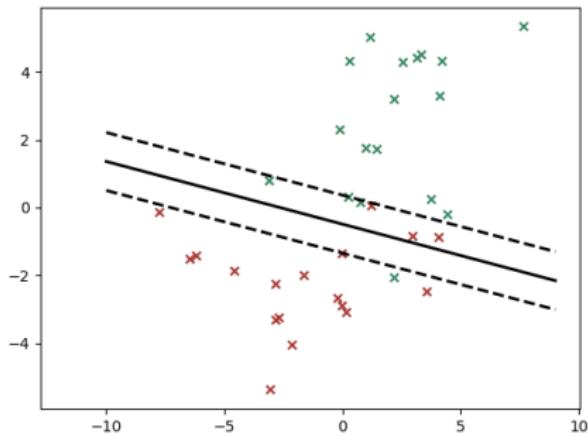
$$\begin{aligned} \text{Accuracy} &= \frac{\text{correct predictions}}{\text{total predictions}} = \\ &= P(\omega^\top x_i + b \geq 0 \wedge y_i = +1) + P(\omega^\top x_i + b < 0 \wedge y_i = -1) \end{aligned}$$

## Example: classification accuracy



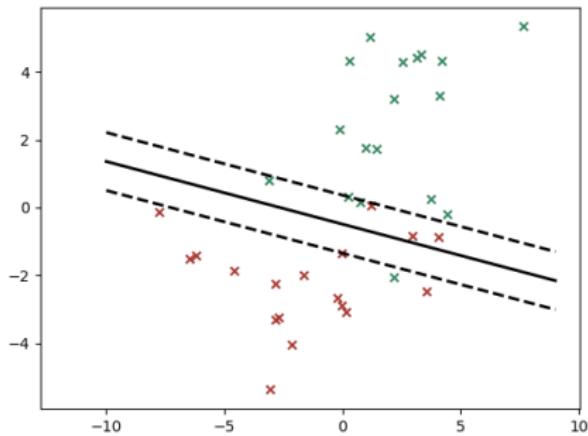
- $|\Omega| = 40$

## Example: classification accuracy



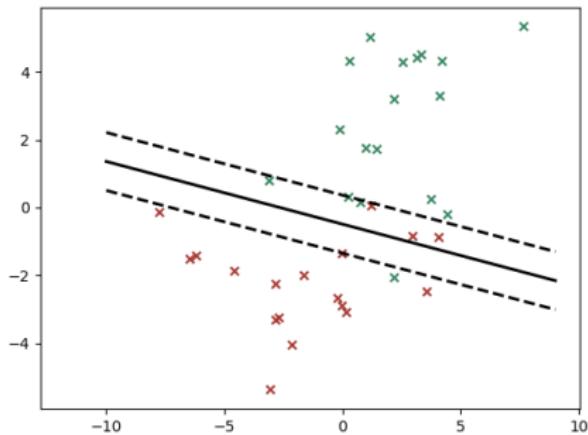
- $|\Omega| = 40$
- $\#\{i, y_i = +1\} = 20$

## Example: classification accuracy

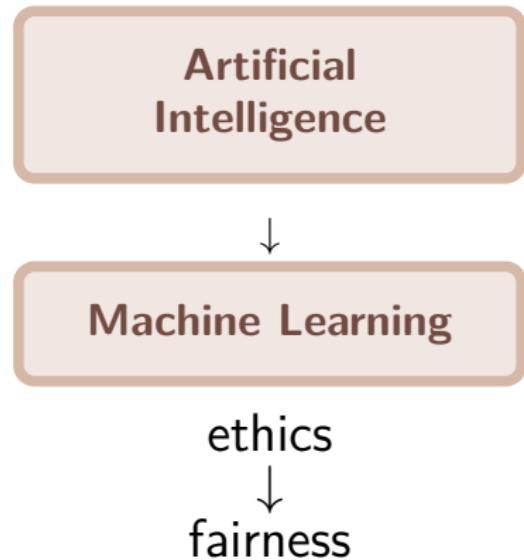
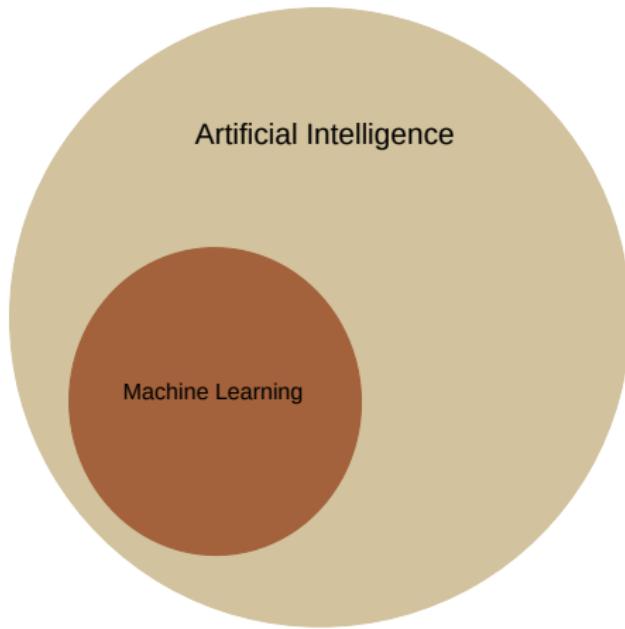


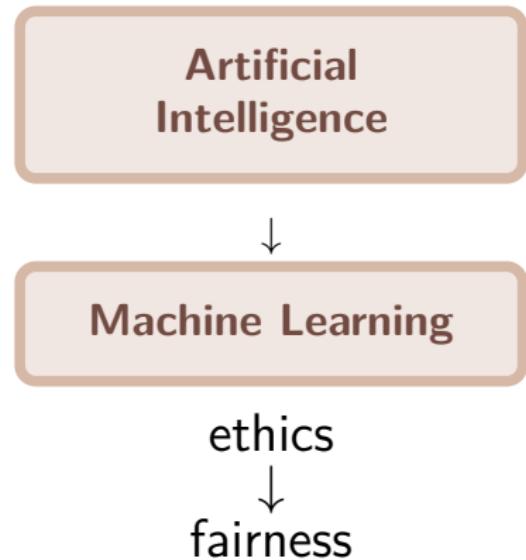
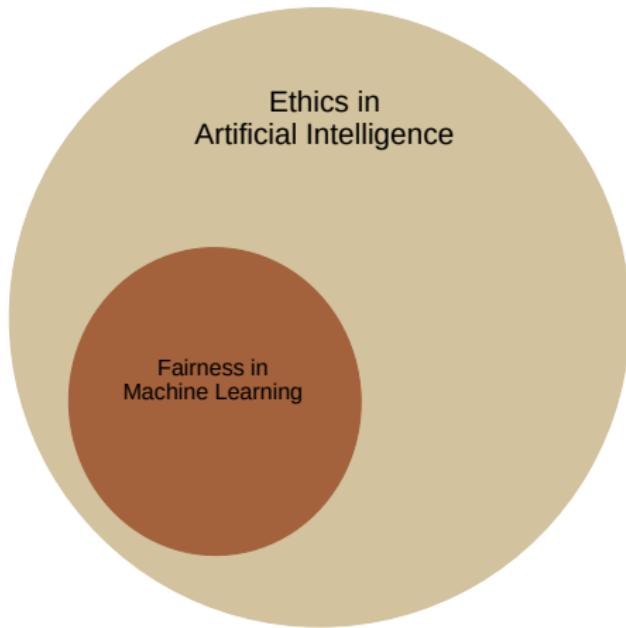
- $|\Omega| = 40$
- $\#\{i, y_i = +1\} = 20$
- $\#\{i, y_i = -1\} = 20$

## Example: classification accuracy

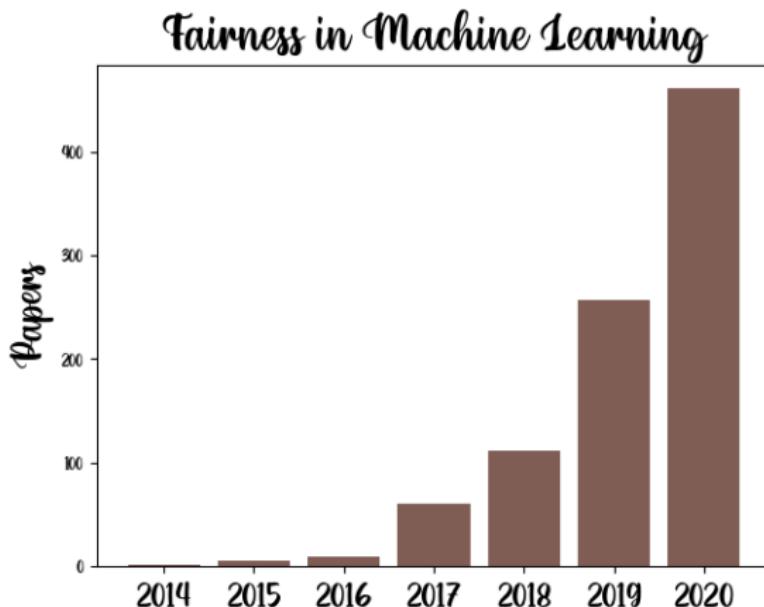


- $|\Omega| = 40$
- $\#\{i, y_i = +1\} = 20$
- $\#\{i, y_i = -1\} = 20$
- Accuracy =  $\frac{19+17}{40} = 0.9$
- 90% of objects correctly classified.





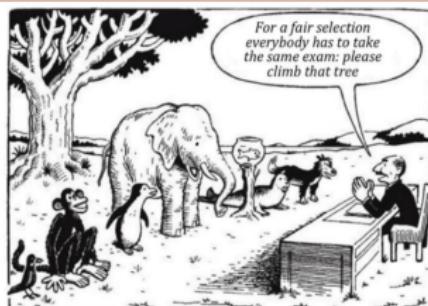
# History of fairness in ML



# What is unfairness?

## Discrimination

[...] **wrongfully** impose a **relative disadvantage** on persons based on their membership in some social salient groups, e.g., race or gender.



### Our Education System

*"Everybody is a genius. But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid."*

- Albert Einstein

Each individual is represented by

$$(x, s, y) \left\{ \begin{array}{l} x, \text{non-sensitive features} \\ s, \text{sensitive feature} \\ y \in \{-1, +1\}, \text{class membership} \end{array} \right.$$

# What is fairness?

*ability to ensure that different social salient groups are treated similarly*

# What is fairness?

*ability to ensure that different social salient groups are treated similarly*

## Fairness in Machine Learning

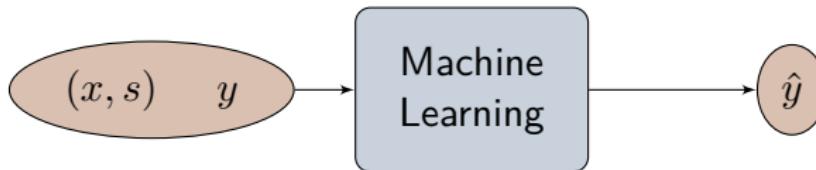
A fair machine learning algorithm implies that its **outcome** should not have a **disproportionately large adverse impact** on a *protected class of features*.

# What is fairness?

*ability to ensure that different social salient groups are treated similarly*

## Fairness in Machine Learning

A fair machine learning algorithm implies that its **outcome** should not have a **disproportionately large adverse impact** on a *protected class of features*.



with  $x$  = non protected features,  $s$  = protected feature,  $y$  = true label and  $\hat{y}$  = predicted label.

# How is unfairness in ML?

Historical data:

	Sensitive feature	non-sensitive features		decision
	Gender & status	income	credit history	
Applicant 1	male married	1.5k	5	✓
Applicant 2	female single	2.5k	3	✓

# Why is ML unfair?

- Sometimes we have less data for minority groups that leads to higher errors
- Feature may be less informative or not reliably collected features.
- Historical data reflects human biases & stereotypes.

# Why do we care about fairness?

- It is highly related to our own benefits.
- Many things have become automated by ML systems.
- Artificial intelligence is good but it can be used incorrectly.

# Google translator

The screenshot shows the Google Translate interface. At the top, it displays language detection options: DÉTECTOR LA LANGUE, ESPAÑOL (underlined), FRANÇAIS, ANGLAIS, followed by a dropdown arrow. In the center, there is a double-headed arrow icon. To its right, another set of language options: ANGLAIS (underlined), FRANÇAIS, ESPAÑOL, followed by a dropdown arrow. Below this, the input text "La doctora  
El enfermero" is shown in the left panel, with a close button "X" next to it. The output text "The doctor  
The nurse" is shown in the right panel, with a checkmark icon next to "The nurse". Below the input text, there are microphone and speaker icons, and a progress bar indicating "23 / 5000". Below the output text, there are icons for copy, edit, and share. At the bottom right, there is a link "Envoyer des commentaires".

Envoyer des commentaires

# Google translator

DÉTECTOR LA LANGUE ESPAGNOL FRANÇAIS ANGLAIS ANGLAIS FRANÇAIS ESPAGNOL

La doctora  
El enfermero

The doctor  
The nurse

Envoyer des commentaires

Microphone icon, speaker icon, 23 / 5000 character count, settings icon.

DÉTECTOR LA LANGUE ESPAGNOL FRANÇAIS ANGLAIS ANGLAIS FRANÇAIS ESPAGNOL

The doctor  
The nurse

El doctor  
La enfermera

Envoyer des commentaires

Microphone icon, speaker icon, 20 / 5000 character count, settings icon.

# COMPAS algorithm: recividism prediction

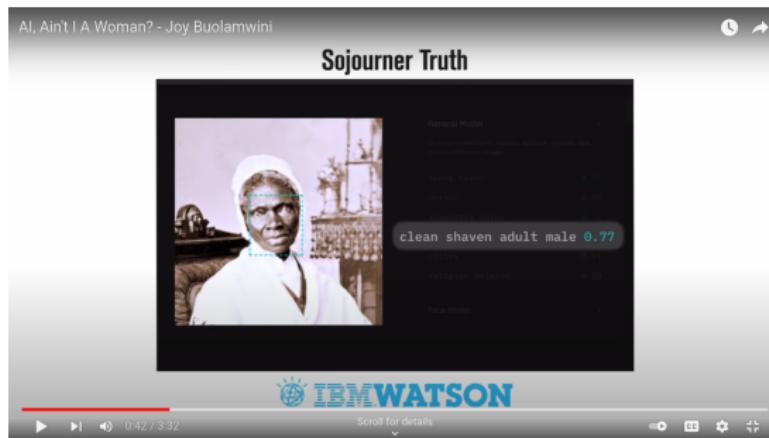
## Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

*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. (Source: ProPublica analysis of data from Broward County, Fla.)*

# AI, ain't I a woman?

Joy Buolamwini



<https://www.youtube.com/watch?v=QxuyfWoVV98>

# Timnit Gebru



## On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender<sup>\*</sup>  
*ebender@uw.edu*  
University of Washington  
Seattle, WA, USA

Angelina McMillan-Major  
symma@uw.edu  
University of Washington  
Seattle, WA, USA

Timnit Gebru\*  
timnit@blackmatters.ai  
Black in AI  
Palo Alto, CA, USA

Shmargaret Shnitchell  
shmargaret.shnitchell@gmail.com  
*The Author*

## ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2, and others, most notably those developed at Google, have been trained on massive amounts of text through an interleaved training and fine-tuning stage. Using these pre-trained models and the methodology of fine-tuning, those for specific tasks, researchers have extended the state-of-the-art on a wide range of NLP tasks, including machine translation, question answering, and text generation, among many others. How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? In this paper, we will review the current state of large-scale NLP and financial costs first, investigating resources like computing cost on the web, carrying out and publishing experiments evaluating how well these models learn from scratch and development goals and support stakeholders values, and encouraging research directions beyond large-scale language models.

CCS CONCEPTS

- Computing methodologies → Natural language processing  
**ACM Reference Format:**  
Emily M. Bender, Tsviatt Gebru, Angélique McMillan-Major, and Shaiman Sharabiani. 2021. On the Dangers of Stochastic Parrotic Con Language Models for Text Segmentation. In Conference on Fairness, Accountability and Transparency (FAT\* '21), March 3–10, 2021, Virtual Event/Online ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3487854.3488070>

## 1 INTRODUCTION

## 1 INTRODUCTION

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on tests from the literature overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can

as sufficiently documented.

As argued by Bender and Koller [14], it is important to understand the limitations of LMs and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §8, LMs are not performing out-of-language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form [14]. Focusing on state-of-the-art results in linguistics without encouraging deeper understanding of the mechanism by which LMs work is misleading, as in the case of machine

# Examples at Université Côte d'Azur

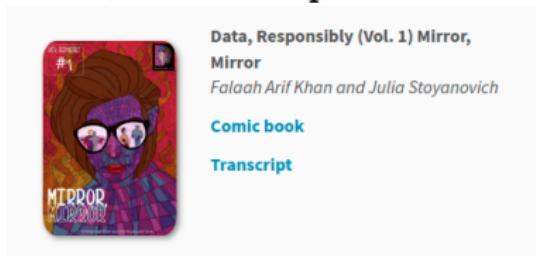
## MONITORING CYBERBULLYING THROUGH MESSAGE CLASSIFICATION AND SOCIAL NETWORK ANALYSIS

Source	Hate speech is to incite violence or hate	Hate speech is to attack or diminish	Hate speech has specific targets	Humour has a specific status
EU Code of conduct	Yes	No	Yes	No
ILGA	Yes	No	Yes	No
Scientific paper	No	Yes	Yes	No
Facebook	No	Yes	Yes	Yes
YouTube	Yes	No	Yes	No
Twitter	Yes	Yes	Yes	No

<http://www.telecom-valley.fr/wp-content/uploads/2020/11/VILLATA-CABRIO-171120.pdf>

# Recommended bibliography

- Easy reading/watching
  - How I'm fighting bias in algorithms, Joy Buolamwini.  
[https://www.youtube.com/watch?v=UG\\_X\\_7g63rY](https://www.youtube.com/watch?v=UG_X_7g63rY)
  - Mirror, mirror. <https://dataresponsibly.github.io/comics/>



# Lecture 2

## Security and Ethical Aspects of Data

Amaya Nogales Gómez  
[amaya.nogales-gomez@univ-cotedazur.fr](mailto:amaya.nogales-gomez@univ-cotedazur.fr)

MSc Data Science & Artificial Intelligence  
Université Côte d'Azur

October 18, 2021

# Tentative Content

## ① Introduction

- Preliminaries and Machine Learning basics
- Motivation for ethics in Artificial Intelligence

## ② Sources of unfairness

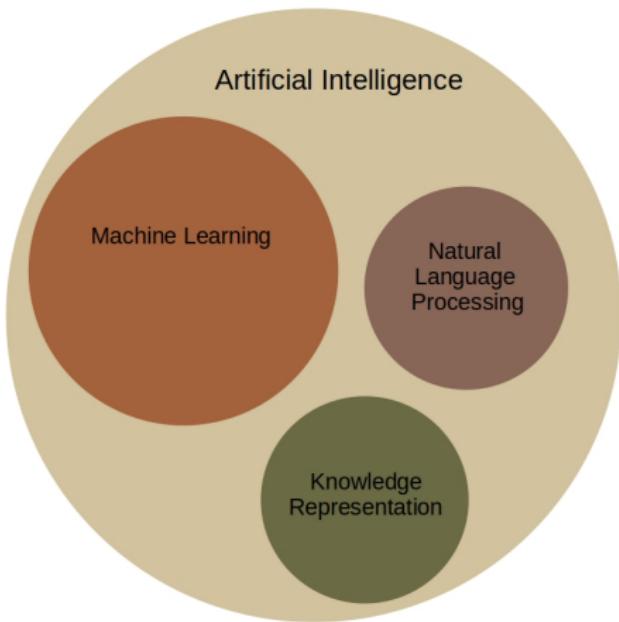
- Bias in data
- Algorithmic unfairness: Examples

## ③ Fairness criteria

- Types of discrimination
- Definitions of fairness

## ④ Algorithms and methods for fair ML

- Pre-processing methods
- In-processing methods
- Post-processing methods
- Legal and policy perspectives



## Supervised Learning

- Classification
- Regression

## Unsupervised Learning

- Clustering
- Dimensionality reduction

## Reinforcement Learning

## Hard-margin SVM: final formulation

The objective function can be replaced by  $\Phi(\|\omega\|^\circ)$  for any  $\Phi$ , increasing in  $\mathbb{R}^+$ .

Taking  $\Phi(t) = \frac{1}{2}t^2$  one obtains an equivalent formulation as a quadratic problem with linear constraints.

$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

s.t.

$$\begin{aligned} y_i(\omega^\top x_i + b) &\geq 1 & \forall i = 1, \dots, n \\ \omega &\in \mathbb{R}^d \\ b &\in \mathbb{R}. \end{aligned}$$

But...

$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

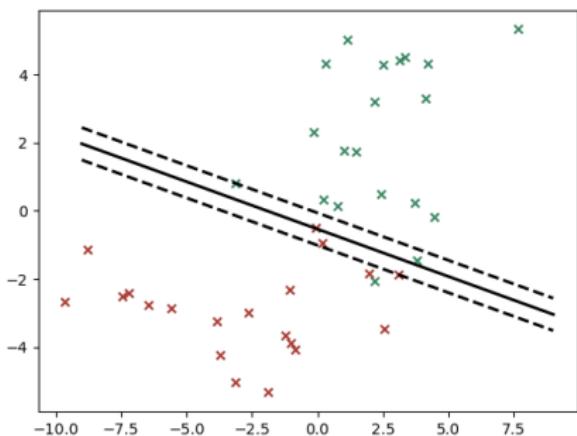
s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$

Non-linearly separable data



INFEASIBLE!!

# The SVM formulation

Non-linearly separable data

$$\min_{\omega, b, \xi} \frac{1}{2} \sum_{j=1}^d \omega_j^2 + \frac{C}{n} \sum_{i=1}^n \xi_i$$

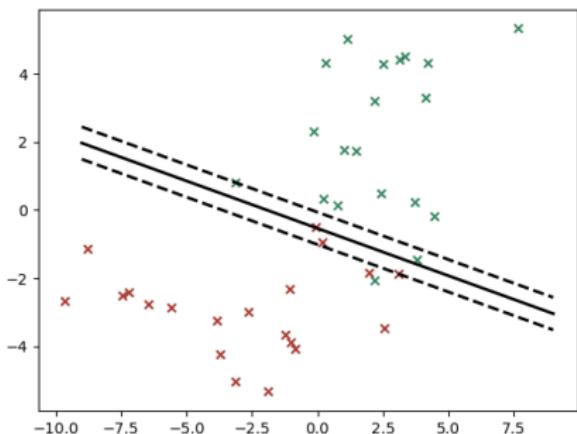
s.t.

$$y_i(\omega^\top x_i + b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n$$

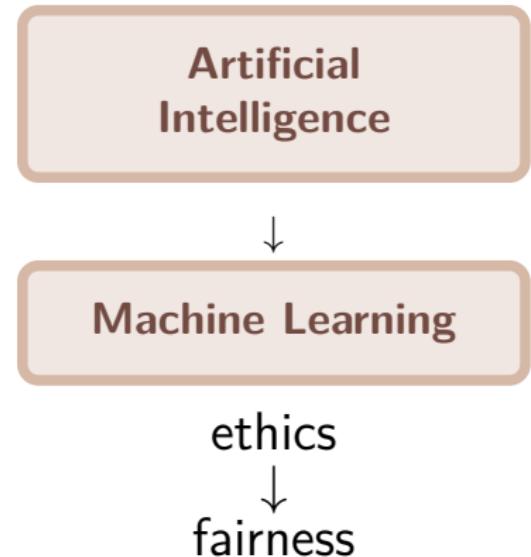
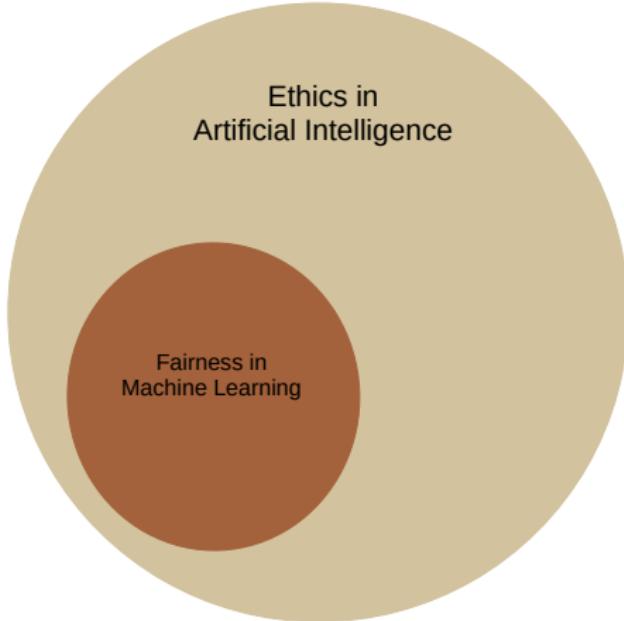
$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}$$

$$\xi_i \geq 0 \quad \forall i = 1, \dots, n.$$



- An object  $i$  will be correctly classified if  $0 \leq \xi_i < 1$
- Misclassified if  $\xi_i > 1$ .
- In the case  $\xi_i = 1$ , we get a tie (objects coincide with the hyperplane).
- $\sum_{i=1}^n \xi_i$  is an upper bound of the number of misclassified objects.



# What is fairness?

*ability to ensure that different social salient groups are treated similarly*

# What is fairness?

*ability to ensure that different social salient groups are treated similarly*

## Fairness in Machine Learning

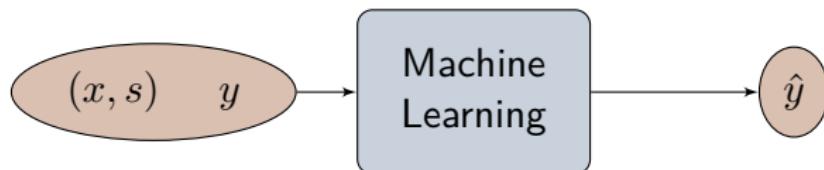
A fair machine learning algorithm implies that its **outcome** should not have a **disproportionately large adverse impact** on a *protected class of features*.

# What is fairness?

*ability to ensure that different social salient groups are treated similarly*

## Fairness in Machine Learning

A fair machine learning algorithm implies that its **outcome** should not have a **disproportionately large adverse impact** on a *protected class of features*.



with  $x$  = non protected features,  $s$  = protected feature,  $y$  = true label and  $\hat{y}$  = predicted label.

# Advertising

## Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads

- The authors explore data from a field test of how an algorithm delivered ads promoting job opportunities in the Science, Technology, Engineering and Math (STEM) fields.
- This ad was explicitly intended to be gender-neutral in its delivery.
- **Fewer women saw the ad than men.**

# Advertising

## Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads

- The authors explore data from a field test of how an algorithm delivered ads promoting job opportunities in the Science, Technology, Engineering and Math (STEM) fields.
- This ad was explicitly intended to be gender-neutral in its delivery.
- **Fewer women saw the ad than men.**
- Why? Because younger women are a prized demographic, i.e., they are more expensive to show ads to.

[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2852260](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260)

# Harm of unfairness in recommender systems

## Information Asymmetry

Knowing a piece of information (e.g., a job opportunity) could change one's life.

## Matthew effect

Advantaged users, items, or groups get further propagated by recommendations, sometimes not because their good quality but because the recommendation model is dominated by their data.

## Echo chambers

Unfair, undiversified exposure to news, tweets, etc. may create echo chamber. Makes it difficult to explore new ideas and opinions different from one's own.

# Ethics in the industry

AI Ethics Global Leader: Francesca Rossi

## AI Ethics

IBM's multidisciplinary,  
multidimensional approach to  
trustworthy AI



<https://www.ibm.com/artificial-intelligence/ethics>

# IBM AI Ethics Initiative

The purpose of AI is to augment human intelligence

AI should make all of us better at our jobs, and that the benefits of the AI era should touch the many, not just the elite few.

Data and insights belong to their creator

Clients' data is their data, and their insights are their insights. Government data policies should be fair and equitable and prioritize openness.

Technology must be **transparent** and **explainable**

Companies must be clear about who trains their AI systems, what data was used in training and, most importantly, what went into their algorithms' recommendations.

# Artificial Intelligence Act: European Parliament



European Parliament

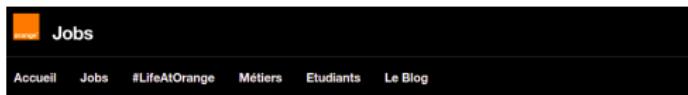
## Trustworthy AI by Design:

Responsible, Trustworthy AI requires awareness from all parties involved, from the first line of code.

The way in which we design our technology is shaping the future of our society.

[https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/694212/EPRS\\_BRI\(2021\)694212\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/694212/EPRS_BRI(2021)694212_EN.pdf)

# Some "more" real motivation: Orange



## POST DOC : Ethics of AI in federated learning F/H

*Europe defends responsible AI and published in mid-April 2021 the "AI act" aimed at legislating on high-risk AI. Societal concerns about **discrimination**, privacy, transparency, **explainability** and the responsibility of data scientists and companies are growing. At the same time, technology continues to improve the performance and accuracy of models. You will tackle the problem of **fairness** due to training data and the mechanism of training a model.*

<https://orange.jobs/jobs/offer.do?joid=105070&lang=FR>

# Interpretability: Why?

- What are the most influential features towards the decision?
- Is the system "fair" by relying on sensitive attributes such as age and marital status?
- I didn't get the loan; what should I do to get next time I apply?

# Interpretability: How?

## Transparency

inherent/model interpretability

the level to which a system provides information about its internal and its training

## Explainability

post-hoc/decision interpretability

the level to which a system can provide clarifications (explanations) for its decisions/outputs

# Explainability

**Why was I denied the loan?  
What should I change in order  
for my application to be  
approved?**

## Counterfactual explanations

You were denied a loan because your annual income was 30000€. If your income had been 45000€ you would have been offered a loan.

# Machine Learning Lifecycle

## Data Collection

- Before any analysis or learning happens, data must first be collected from the world.
- Data collection involves selecting a population, as well as picking and measuring features and labels to use.
- There exist already many repositories of real-life datasets.

<https://archive.ics.uci.edu/ml/datasets.php>

# Machine Learning Lifecycle

## Data Preparation

- Depending on the data modality and task, different types of preprocessing may be applied to the dataset before using it.
- Datasets are usually split into a training data used during model development, and testing data used during model evaluation.
- Part of the training data may be further set aside as validation data.

# Machine Learning Lifecycle

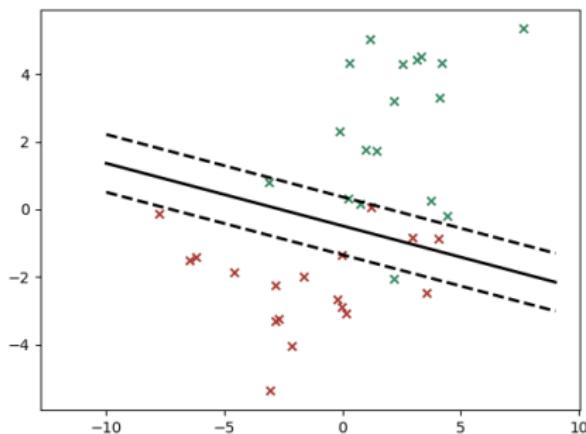
## Model Development

- A model is then built using the training data.
- A number of different model types, hyperparameters, and optimization methods may be tested out at this point; usually these different configurations are compared based on their performance on the testing data, and the best one chosen.
- The particular performance metric(s) used in such comparisons are chosen based on the task and data characteristics; common choices are accuracy, false or true positive rates (FPR/TPR).

# Common classification criteria

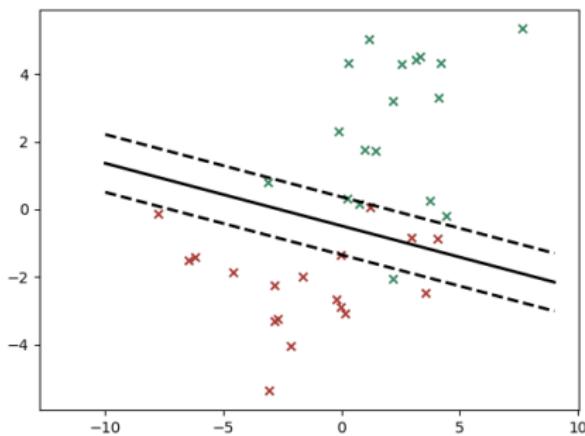
Prediction $\hat{y}$	Label $y$	Criteria
+1	+1	True positive rate
-1	+1	False negative rate
+1	-1	False positive rate
-1	-1	True negative rate

## Example: Classification accuracy



- $|\Omega| = 40$
- $\#\{i, y_i = +1\} = 20$
- $\#\{i, y_i = -1\} = 20$
- Accuracy =  $\frac{19+17}{40} = 0.9$

## Example: Classification accuracy



- $|\Omega| = 40$
- $\#\{i, y_i = +1\} = 20$
- $\#\{i, y_i = -1\} = 20$
- Accuracy =  $\frac{19+17}{40} = 0.9$

$\hat{y}$	$y$	Criteria	
+1	+1	TPR	$\frac{19}{20}$
-1	+1	FNR	$\frac{1}{20}$
+1	-1	FPR	$\frac{3}{20}$
-1	-1	TNR	$\frac{17}{20}$

# Machine Learning Lifecycle

## Model Evaluation

- After the final model and hyperparameters are chosen and the model optimization finished, the final performance of the model on the validation data is reported.
- It is important that the validation data is not used before this step to ensure that the model's performance is a true representation of how it performs on unseen data.
- As in model development, choosing well-suited performance metric(s) is important.

# Machine Learning Lifecycle

## Model Postprocessing

- Once a model is ready to be used, there are various post-processing steps that may need to be applied.
- For example, if the output of a model performing binary classification is a probability, but the desired output to display to users is a binary answer, there remains a choice of what threshold(s) to use to round the probability to a hard classification.

# Machine Learning Lifecycle

## Model Deployment

- For a real-world machine learning application, there are many steps that arise when a system is actually deployed.
- For example, a model may need to be changed based on requirements for fairness, or there may be real-time feedback that should be integrated back into the model.

# Bias in data

## Historical bias

- It arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model.
- It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.

## Representation bias

- It arises while defining and sampling a development population.
- It occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population.

## Measurement Bias

- It arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities.
- The chosen set of features and labels may leave out important factors or introduce group or input-dependent noise that leads to differential performance.

## Aggregation bias

- It arises during model construction, when distinct populations are inappropriately combined.
- In many applications, the population of interest is heterogeneous and a single model is unlikely to suit all subgroups.

## Evaluation bias

- It occurs during model iteration and evaluation.
- It can arise when the testing populations do not equally represent the various parts of the use population.
- Evaluation bias can also arise from the use of performance metrics that are not appropriate for the way in which the model will be used.

## Deployment Bias

- It occurs after model deployment, when a system is used or interpreted in inappropriate ways.

# Historical Bias

- Historical bias arises even if the data is perfectly measured and sampled.
- It happens if the world as it is or was leads a model to produce outcomes that are not fair.
- Such a system, even if it reflects the world accurately, can still inflict harm on a population.
- Considerations of historical bias often involve evaluating the representational harm (such as reinforcing a stereotype) to a particular identity group.

## Example: Image search

In 2018, 5% of Fortune 500 CEOs were women. Should image search results for "CEO" reflect that number? Ultimately, a variety of stakeholders, including affected members of society, should evaluate the particular harms that this result could cause and make a judgment. This decision may be at odds with the available data even if that data is a perfect reflection of the world. Indeed, Google has recently changed their Image Search results for "CEO" to display a higher proportion of women.

# Representation bias

- Representation bias occurs when certain parts of the input space are underrepresented.
- Representation bias can arise for several reasons, including:
  - ① The sampling methods only reach a portion of the population.
  - ② The population of interest has changed or is distinct from the population used during model training.

## Example: ImageNet dataset

Geographic diversity in image datasets ImageNet is a widely-used image dataset consisting of 1.2 million labeled images. Approximately 45% of the images in ImageNet were taken in the United States, and the majority of the remaining images are from North America or Western Europe. 1% and 2.1% of the images come from China and India, respectively. Shankar et al. (2017) show that the performance of a classifier trained on ImageNet is worse for several categories (such as "bride- groom") on images from under-represented countries such as Pakistan or India versus images from North America and Western Europe.

Shankar, S.; Halpern, Y.; Breck, E.; Atwood, J.; Wilson, J. and Sculley, D. 2017. *No classification without representation: Assessing geodiversity issues in open data sets for the developing world.* arXiv preprint arXiv:1711.08536.

# Measurement bias

- Measurement bias occurs when choosing, collection, or computing features and labels to use in a prediction problem.
- After choosing factors to measure, the measurement process itself adds a second layer of noise.
- If the process of choosing and measuring these factors just adds random noise, the model parameters will converge to those we would expect with the correctly measured quantities.
- Often arises because proxies are generated differently across groups.
- Measurement bias can arise in several ways:
  - ① The measurement process varies across groups.
  - ② The quality of data varies across groups.
  - ③ The defined classification task is an oversimplification. Reducing a decision to a single attribute can create a biased proxy label because it only captures a particular aspect of what we really want to measure.

## Example: Predictive policing and risk assessments

In predictive policing applications, the proxy variable "arrest" is often used to measure "crime" or some underlying notion of "riskiness". Because minority communities are often more highly policed and have higher arrest rates, there is a different mapping from crime to arrest for people from these communities. Prior arrests and friend/family arrests were two of many differently mismeasured proxy variables used in the recidivism risk prediction tool COMPAS. This was a factor that eventually led to higher false positive rates for black versus white defendants. It is worth noting that such an evaluation is further complicated by the proxy label "rearrest" used to measure "recidivism".

# Aggregation bias

- Aggregation bias arises when a one-size-fit-all model is used for groups with different conditional distributions.
- Group membership can be indicative of different backgrounds, cultures or norms, and a given variable can mean something quite different for a person in a different group.
- Aggregation bias can lead to a model that is not optimal for any group, or a model that is fit to the dominant population.

## Example: Clinical-aid tools

Diabetes patients have known differences in associated complications across ethnicities. Studies have also suggested that HbA1c levels (widely used to diagnose and monitor diabetes) differ in complex ways across ethnicities and genders. Because these factors have different meanings and importance within different subpopulations, a single model to predict complications is unlikely to be best-suited for any group in the population even if they are equally represented in the training data.

# Evaluation bias

- Evaluation bias occurs when the evaluation and/or benchmark data for an algorithm does not represent the target population.
- A model is optimized on its training data, but its quality is often measured on benchmarks (e.g., UCI datasets).
- A misrepresentative benchmark encourages the development of models that only perform well on a subset of the population.
- Evaluation bias can be exacerbated by the particular metrics that are used to report performance.

## Example: Commercial facial recognition algorithms

It has been empirically shown a drastically worse performance of commercially-used facial analysis algorithms (performing tasks such as gender- or smiling-detection) on dark-skinned females. Looking at some common facial analysis benchmark datasets, it becomes apparent why such algorithms were considered appropriate for use - just 7.4% and 4.4% of the images in benchmark datasets such as Adience and IJB-A are of dark-skinned female faces. Algorithms that underperform on this slice of the population therefore suffer quite little in their evaluation performance on these benchmarks. The algorithms' underperformance was likely caused by representation bias in the training data, **but the benchmarks failed to discover and penalize this.**

Since this study, other algorithms have been benchmarked on more balanced face datasets, changing the development process to encourage models that perform well across groups.

# Deployment bias

- Deployment bias arises when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used.
- This often occurs when a system is built and evaluated as if it were fully autonomous, while in reality, it operates in a complicated sociotechnical system moderated by institutional structures and human decision-makers.

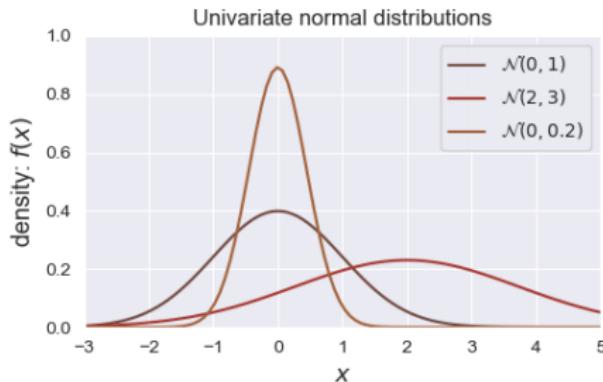
## Example: Risk assessment tools

Algorithmic risk assessment tools are models intended to predict a person's likelihood of committing a future crime. In practice, however, these tools may be used in "off-label" ways, such as to help determine the length of a sentence. One of the harmful consequences of risk assessment tools for sentencing, includes the justification of increased incarceration on the basis on personal characteristics.

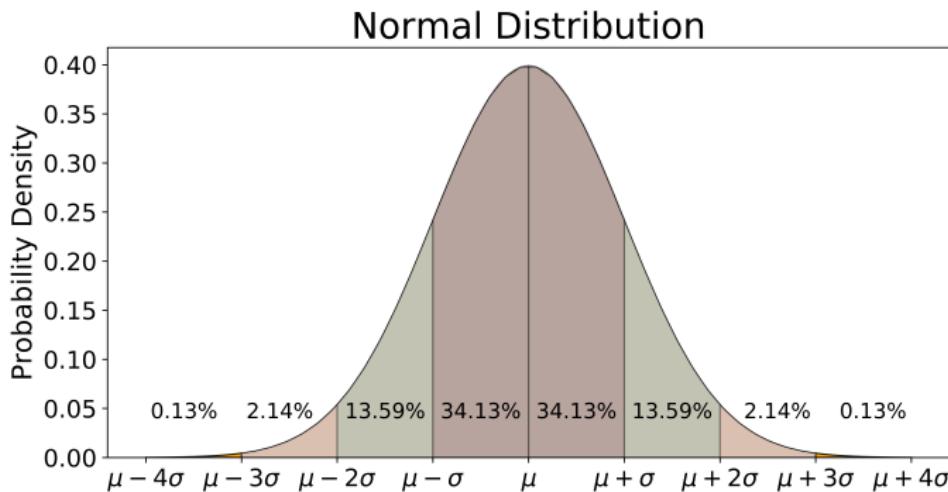
# Gaussian distribution

A normal (or Gaussian) distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$X \sim N(\mu, \sigma^2)$$
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



$$\begin{aligned} P(\mu - n\sigma \leq X \leq \mu + n\sigma) &= F(\mu + n\sigma) - F(\mu - n\sigma) \\ F(x) &= \int_{-\infty}^x f(x)dx \end{aligned}$$

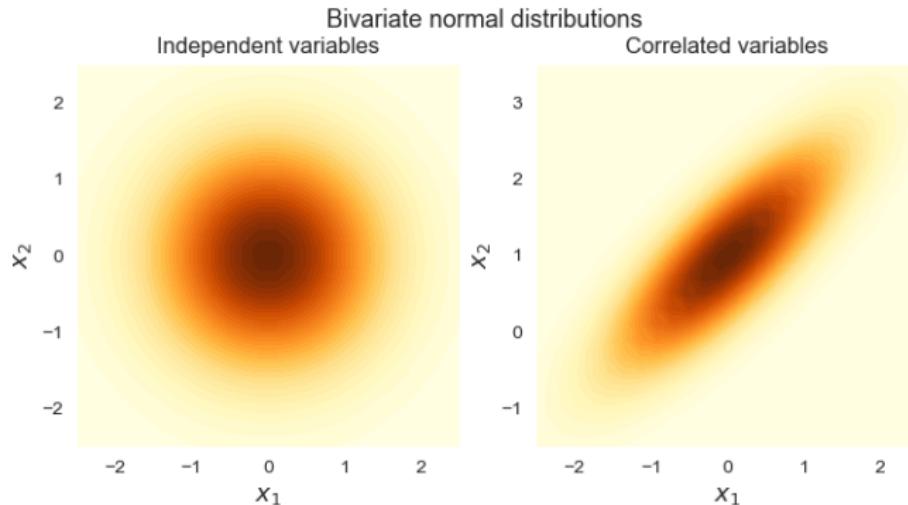


# Multivariate normal distribution

The multivariate normal distribution is a multidimensional generalisation of the one-dimensional normal distribution.

It represents the distribution of a multivariate random variable that is made up of multiple random variables that can be correlated with each other.

$$\begin{aligned} X &\sim N(\mu, \Sigma) \\ f(x) &= \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}((x-\mu)^\top \Sigma^{-1}(x-\mu))^2} \end{aligned}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

# What are some real-world examples of normal distribution?

- Human heights (people of the same gender and age group typically cluster around average with normal distribution)
- IQ scores (the mean is typically 100, SD = 15)
- Marks of students in a class (mean = 60, SD = 20)
- Measure of weight (mean = 80 kg, SD = 10)
- Measure of blood pressure (mean = 120/80, SD = 20)
- Height of trees (measurement in meters; mean = 40 m, SD = 20)

## Recommended bibliography

- Y. Li, Y. Ge, and Y. Zhang. 2021. Tutorial on Fairness of Machine Learning in Recommender Systems. SIGIR.  
<https://doi.org/10.1145/3404835.3462814>
- <http://cs229.stanford.edu/section/gaussians.pdf>

# Lecture 3

## Security and Ethical Aspects of Data

Amaya Nogales Gómez  
[amaya.nogales-gomez@univ-cotedazur.fr](mailto:amaya.nogales-gomez@univ-cotedazur.fr)

MSc Data Science & Artificial Intelligence  
Université Côte d'Azur

October 25, 2021

# Tentative Content

## ① Introduction

- Preliminaries and Machine Learning basics
- Motivation for ethics in Artificial Intelligence

## ② Sources of unfairness

- Bias in data
- Algorithmic unfairness: Examples

## ③ Fairness criteria

- Types of discrimination
- Definitions of fairness

## ④ Algorithms and methods for fair ML

- Pre-processing methods
- In-processing methods
- Post-processing methods
- Legal and policy perspectives

# The SVM formulation

Non-linearly separable data

$$\min_{\omega, b, \xi} \frac{1}{2} \sum_{j=1}^d \omega_j^2 + \frac{C}{n} \sum_{i=1}^n \xi_i$$

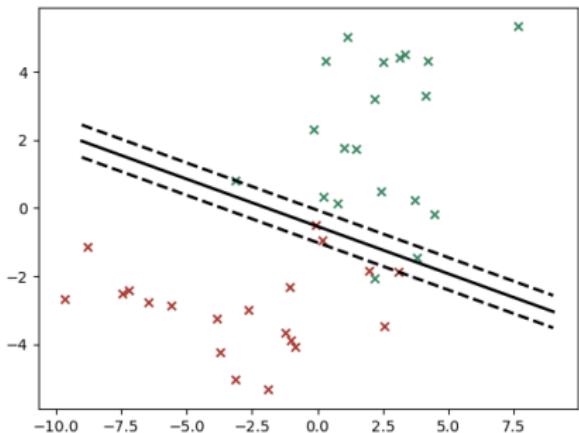
s.t.

$$y_i(\omega^\top x_i + b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}$$

$$\xi_i \geq 0 \quad \forall i = 1, \dots, n.$$



- An object  $i$  will be correctly classified if  $0 \leq \xi_i < 1$
- Misclassified if  $\xi_i > 1$ .
- In the case  $\xi_i = 1$ , we get a tie (objects coincide with the hyperplane).
- $\sum_{i=1}^n \xi_i$  is an upper bound of the number of misclassified objects.

# Advertising

## Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads

- The authors explore data from a field test of how an algorithm delivered ads promoting job opportunities in the Science, Technology, Engineering and Math (STEM) fields.
- This ad was explicitly intended to be gender-neutral in its delivery.
- **Fewer women saw the ad than men.**

# Advertising

## Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads

- The authors explore data from a field test of how an algorithm delivered ads promoting job opportunities in the Science, Technology, Engineering and Math (STEM) fields.
- This ad was explicitly intended to be gender-neutral in its delivery.
- **Fewer women saw the ad than men.**
- Why? Because younger women are a prized demographic, i.e., they are more expensive to show ads to.

[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2852260](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260)

# Machine Learning lifecycle

- ① Data collection
- ② Data preparation
- ③ Model development
- ④ Model evaluation
- ⑤ Model postprocessing
- ⑥ Model deployment

# Bias in data

## Historical bias

- It arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model.
- It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.

## Representation bias

- It arises while defining and sampling a development population.
- It occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population.

## Measurement Bias

- It arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities.
- The chosen set of features and labels may leave out important factors or introduce group or input-dependent noise that leads to differential performance.

## Aggregation bias

- It arises during model construction, when distinct populations are inappropriately combined.
- In many applications, the population of interest is heterogeneous and a single model is unlikely to suit all subgroups.

## Evaluation bias

- It occurs during model iteration and evaluation.
- It can arise when the testing populations do not equally represent the various parts of the use population.
- Evaluation bias can also arise from the use of performance metrics that are not appropriate for the way in which the model will be used.

## Deployment Bias

- It occurs after model deployment, when a system is used or interpreted in inappropriate ways.

# Sensitive features

- In many classification tasks, the features  $X$  contain or implicitly encode sensitive characteristics of an individual.
- We let  $A$  to designate a binary variable that captures one sensitive characteristic.
- Different settings of  $A$  correspond to different groups of the population.
- This choice of notation is not meant to suggest that we can cleanly partition the set of features into two independent categories such as "neutral" and "sensitive".
- **The choice of sensitive attributes will generally have profound consequences as it decides which groups of the population we highlight, and what conclusions we draw from our study.**

## Notation

- Dataset  $(X, A, Y)$
- $X$ : non-protected features
- $A$ : protected attribute
- $Y \in \{-1, +1\}$  label, class membership
- $\hat{Y} = f(X, A)$ : prediction

## Disclaimer: abuse of notation

- feature  $\equiv$  attribute  $\equiv$  variable  $\equiv$  characteristic
- protected  $\equiv$  sensitive

# Discrimination Law: two doctrines

## Disparate treatment

A decision making process suffers from disparate treatment if its decisions are partly based on the subject's sensitive attribute information.

## Formal or intentional

## Disparate impact

A decision making process suffers from disparate impact if its outcomes disproportionately hurt people with certain sensitive attribute values.

## Unjustified or avoidable

- While it is desirable to design decision making systems free of disparate treatment as well as disparate impact, controlling for both forms of unfairness simultaneously is challenging.
- Avoid disparate treatment → disparate impact
- Avoid disparate impact → disparate treatment

# Disparate Treatment

## Formal

Explicitly considering the sensitive feature, even if it is relevant

## Intentional

Purposefully attempting to discriminate without direct reference to sensitive feature

# The $p\%$ rule

A decision boundary satisfies the "80%-rule" (or more generally the " $p\%$ -rule"), if the ratio between the percentage of users with a particular sensitive attribute value having  $\hat{Y} = +1$  and the percentage of users without that value having  $\hat{Y} = +1$  is no less than 80:100 (p:100). For a given binary sensitive attribute  $A \in \{0, 1\}$ , one can write the  $p\%$ -rule as:

$$\min \left( \frac{\mathcal{P}(\hat{Y} = +1|A = 1)}{\mathcal{P}(\hat{Y} = +1|A = 0)}, \frac{\mathcal{P}(\hat{Y} = +1|A = 0)}{\mathcal{P}(\hat{Y} = +1|A = 1)} \right) \geq \frac{p}{100}$$

# Disparate impact

- ① Accuser must first establish that decision procedure has a disparate impact, i.e., does it satisfy the 80%-rule?
- ② Defendant must provide a justification for making decisions in this way. Is there a 'business necessity'? Is it 'job-related'?
- ③ Finally, accuser has the opportunity to show that defendant could achieve same goal using a different procedure that would result in a smaller disparity. Is there an 'alternative practice'?

# Formal non-discrimination criteria

## Independence

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

## Separation

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

## Sufficiency

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

# Independence

The random variables  $(A, \hat{Y})$  satisfy independence if  
 $A \perp \hat{Y}$ .

- Independence has been explored through many equivalent terms or variants, referred to as **demographic parity, statistical parity, group fairness, disparate impact** and others.
- In the case of binary classification, independence simplifies to the condition

$$P(\hat{Y} = +1|A = 0) = P(\hat{Y} = +1|A = 1).$$

- Thinking of the event  $\hat{Y} = +1$  as "acceptance", the condition requires the acceptance rate to be the same in all groups.

# Separation

**Random variables  $(\hat{Y}, A, Y)$  satisfy separation if**  
 $\hat{Y} \perp A | Y.$

- It acknowledges the sensitive characteristic may be correlated with the target variable.
- The separation criterion allows correlation between the score and the sensitive attribute to the extent that it is justified by the target variable.

# Sufficiency

**Random variables  $(\hat{Y}, A, Y)$  satisfy sufficiency if**  
 $Y \perp A | \hat{Y}.$

- It requires a parity of positive/negative predictive values across all groups, i.e.,

$$\mathcal{P}(Y = +1 | \hat{Y} = \hat{y}, A = 0) = \mathcal{P}(Y = +1 | \hat{Y} = \hat{y}, A = 1), \forall \hat{y} \in \{+1, -1\}$$

# Relationships between criteria: Independence vs Sufficiency

## Proposition 1

Assume that  $A$  and  $Y$  are not independent. Then sufficiency and independence cannot both hold.

*Proof.* (Proof by contradiction) By the contraction property for conditional independence,

$$A \perp \hat{Y} \text{ and } A \perp Y | \hat{Y} \rightarrow A \perp (Y, \hat{Y}) \rightarrow A \perp Y.$$

That is,  $A \perp (Y, \hat{Y})$  means that  $A$  is independent of the pair of random variables  $(Y, \hat{Y})$ . And dropping  $\hat{Y}$  cannot introduce a dependence between  $A$  and  $Y$ . □

# Independence vs Separation

## Proposition 2

Assume  $Y$  is binary,  $A$  is not independent of  $Y$ , and  $\hat{Y}$  is not independent of  $Y$ . Then, independence and separation cannot both hold.

*Proof.* In order to prove it by contradiction, we need to prove that

$$A \perp \hat{Y} \text{ and } A \perp \hat{Y} \mid Y \rightarrow A \perp Y \text{ or } \hat{Y} \perp Y.$$

By the law of total probability,

$$\mathcal{P}(\hat{Y} = \hat{y} \mid A = a) = \sum_y \mathcal{P}(\hat{Y} = \hat{y} \mid A = a, Y = y) \mathcal{P}(Y = y \mid A = a)$$

Applying the assumption  $A \perp \hat{Y}$  and  $A \perp \hat{Y} \mid Y$ , this equation simplifies to

*Proof. (continuation)*

$$\mathcal{P}(\hat{Y} = \hat{y}) = \sum_y \mathcal{P}(\hat{Y} = \hat{y}|Y = y)\mathcal{P}(Y = y|A = a) \quad (1)$$

Applied differently, the law of total probability states

$$\mathcal{P}(\hat{Y} = \hat{y}) = \sum_y \mathcal{P}(\hat{Y} = \hat{y}|Y = y)\mathcal{P}(Y = y) \quad (2)$$

Combining (1) and (2), we have

$$\sum_y \mathcal{P}(\hat{Y} = \hat{y}|Y = y)\mathcal{P}(Y = y) = \sum_y \mathcal{P}(\hat{Y} = \hat{y}|Y = y)\mathcal{P}(Y = y|A = a)$$

Replacing  $p = \mathcal{P}(Y = 0)$ ,  $p_a = \mathcal{P}(Y = 0|A = a)$ ,  $\hat{y}_y = \mathcal{P}(\hat{Y} = \hat{y}|Y = y)$ , above:

$$p\hat{y}_0 + (1 - p)\hat{y}_1 = p_a\hat{y}_0 + (1 - p_a)\hat{y}_1.$$

Equivalently,  $p(\hat{y}_0 - \hat{y}_1) = p_a(\hat{y}_0 - \hat{y}_1)$ . This equation only holds if  $\hat{y}_0 = \hat{y}_1$ , which implies  $\hat{Y} \perp Y$  or if  $p = p_a$  for all  $a$ , in which case  $Y \perp A$ . □

## Fairness Through Unawareness

It implies that  $f(X, A)$  does not use the value of  $A$  and is a legal requirement in many domains where processing sensitive information about individuals is forbidden in order to guarantee no disparate treatment. A predictor  $\hat{Y}$  satisfies unawareness if it does not use the protected attribute  $A$ , i.e.,

$$\mathcal{P}(\hat{Y}|X, A) = \mathcal{P}(\hat{Y}|X).$$

## Demographic Parity or Statistical Parity

A stronger definition of fairness compared to unawareness is **demographic parity**. A predictor  $\hat{Y}$  satisfies demographic parity with respect to protected attribute  $A$ , if  $\hat{Y}$  is independent of  $A$ , i.e.,

$$\mathcal{P}(\hat{Y}|A) = \mathcal{P}(\hat{Y}).$$

or equivalently

$$\mathcal{P}(\hat{Y}|A=0) = \mathcal{P}(\hat{Y}|A=1).$$

*The protected and unprotected groups should receive the same distribution of output values.*

# Demographic parity

$$\mathcal{P}(\hat{Y}|A=0) = \mathcal{P}(\hat{Y}|A=1).$$

## Pros

- Legal support: the "four-fifth rule" or "80%-rule" states that the selection rate for any protected group should be no less than four-fifths of that for the non-protected group.
- If this rule is violated, justification must be provided.
- *"Business necessity means that using the procedure is essential to the safe and efficient operation of the business and there are no alternative procedures that are substantially equally valid and would have less adverse impact."*

# Demographic parity

$$\mathcal{P}(\hat{Y}|A=0) = \mathcal{P}(\hat{Y}|A=1).$$

## Cons

- This definition ignores any possible correlation between  $Y$  and  $A$  and in particular excludes the perfect predictor  $Y = \hat{Y}$  when base rates are different,  
i.e.,  $\mathcal{P}(Y = +1|A = 0) \neq \mathcal{P}(Y = +1|A = 1)$ .
- The notion permits that we accept the qualified applicants in one demographic, but random individuals in another, so long as the percentages of acceptance match.

## Equalized odds

A predictor  $\hat{Y}$  satisfies **equalized odds** with respect to protected attribute  $A$  and outcome  $Y$ , if  $\hat{Y}$  and  $A$  are independent conditional on  $Y$ .

In this case, the equalized odds are equivalent to:

$$\mathcal{P}(\hat{Y} = +1 | A = 0, Y = y) = \mathcal{P}(\hat{Y} = +1 | A = 1, Y = y), \\ \forall y \in \{-1, +1\}.$$

*The protected and unprotected groups should have equal true positive and false positive rates.*

# Equalize odds

$$\mathcal{P} \left( \hat{Y} = +1 | A = 0, Y = y \right) = \mathcal{P} \left( \hat{Y} = +1 | A = 1, Y = y \right), \forall y \in \{-1, +1\}.$$

## Pros

- Optimality compatibility:  $\hat{Y} = Y$  is allowed.
- Penalizes laziness: it provides incentive to reduce errors uniformly in all groups.

## Equal Opportunity

We say that a binary predictor  $\hat{Y}$  satisfies **equal opportunity** with respect to  $A$  and  $Y$  if

$$\mathcal{P}(\hat{Y} = +1 | A = 0, Y = +1) = \mathcal{P}(\hat{Y} = +1 | A = 1, Y = +1)$$

*The protected and unprotected groups should have equal true positive rate.*

# Equal Opportunity

$$\mathcal{P}(\hat{Y} = +1 | A = 0, Y = +1) = \mathcal{P}(\hat{Y} = +1 | A = 1, Y = +1)$$

## Pros

- It allows for stronger utility (accuracy).

## Cons

- Weaker notion of non-discrimination: it may not help with different notion of bias, i.e. historical bias.

## Predictive equality

We say that a binary predictor  $\hat{Y}$  satisfies **predictive equality** with respect to  $A$  and  $Y$  if

$$\mathcal{P}(\hat{Y} = +1 | A = 0, Y = -1) = \mathcal{P}(\hat{Y} = +1 | A = 1, Y = -1)$$

*The protected and unprotected groups should have equal false positive rate.*

# Predictive Equality

$$\mathcal{P}(\hat{Y} = +1 | A = 0, Y = -1) = \mathcal{P}(\hat{Y} = +1 | A = 1, Y = -1)$$

## Pros

- It allows for stronger utility (accuracy).

## Cons

- Weaker notion of non-discrimination: it may not help with different notion of bias, i.e. historical bias.

## Predictive Parity

We say that a binary predictor  $\hat{Y}$  satisfies **predictive parity** with respect to  $A$  and  $Y$  if

$$\mathcal{P}(Y = +1 | \hat{Y} = +1, A = 0) = \mathcal{P}(Y = +1 | \hat{Y} = +1, A = 1)$$

*Predictive parity requires the same positive predictive value (i.e., precision) in both groups.*

# Predictive Parity

$$\mathcal{P} \left( Y = +1 | \hat{Y} = +1, A = 0 \right) = \mathcal{P} \left( Y = +1 | \hat{Y} = +1, A = 1 \right)$$

## Pros

- Optimality compatibility:  $\hat{Y} = Y$  satisfies predictive parity.
- Equal chances of success given acceptance.

## Cons

- Same as for equal opportunity and predictive equality, it may not help "closing the gap" between protected and non-protected groups.

And even more...

Group fairness, individual fairness, fairness through awareness, treatment equality, test fairness, counterfactual fairness, conditional statistical parity, conditional use accuracy equality...

# List of demographic fairness criteria

Criteria	Category	Relationship
Group fairness	Independence	Equivalent
Demographic parity	Independence	Equivalent
Equal opportunity	Separation	Relaxation
Equalized odds	Separation	Equivalent
Unawareness	Separation	Equivalent
Predictive equality	Separation	Relaxation
Predictive parity	Sufficiency	Relaxation

# The impossibility theorem of fairness

## Demographic Parity vs Predictive Parity

If  $A$  and  $Y$  are not independent, then either demographic parity or predictive parity holds.

## Demographic Parity vs Equalized odds

If  $A$  is not independent of  $Y$  and  $\hat{Y}$  is not independent of  $Y$ , then either demographic parity or equalized odds holds.

# Fairness metrics

## Equalized Odds

$$UNF_{EOdds} = |\mathcal{P}(\hat{Y} = +1|Y = +1, A = 0) - \mathcal{P}(\hat{Y} = +1|Y = +1, A = 1)| \\ + |\mathcal{P}(\hat{Y} = +1|Y = -1, A = 0) - \mathcal{P}(\hat{Y} = +1|Y = -1, A = 1)|.$$

## Predictive Equality

$$UNF_{PE} = |\mathcal{P}(\hat{Y} = +1|Y = -1, A = 0) - \mathcal{P}(\hat{Y} = +1|Y = -1, A = 1)|.$$

## Equal opportunity

$$UNF_{EOpp} = |\mathcal{P}(\hat{Y} = +1|Y = +1, A = 0) - \mathcal{P}(\hat{Y} = +1|Y = +1, A = 1)|.$$

# Fairness metrics (continuation)

## Demographic Parity

$$UNF_{DP} = |\mathcal{P}(\hat{Y} = +1|A = 0) - \mathcal{P}(\hat{Y} = +1|A = 1)|.$$

## Predictive parity

$$UNF_{PP} = |\mathcal{P}(Y = +1|\hat{Y} = +1, A = 0) - \mathcal{P}(Y = +1|\hat{Y} = +1, A = 1)|.$$

## Disparate Impact or The $p\%$ -rule

$$UNF_{DI} = \min \left( \frac{\mathcal{P}(\hat{Y} = +1|A = 1)}{\mathcal{P}(\hat{Y} = +1|A = 0)}, \frac{\mathcal{P}(\hat{Y} = +1|A = 0)}{\mathcal{P}(\hat{Y} = +1|A = 1)} \right) \times 100.$$

# Recommended bibliography

- Barocas, S. and Hardt, M. *NIPS 2017 Tutorial on Fairness in Machine Learning.*  
<https://mrtz.org/nips17/#/>
- Hardt, M. and Price, E. and Srebro, N., *Equality of Opportunity in Supervised Learning.* <https://arxiv.org/abs/1610.02413>

# Lecture 4

## Security and Ethical Aspects of Data

Amaya Nogales Gómez  
[amaya.nogales-gomez@univ-cotedazur.fr](mailto:amaya.nogales-gomez@univ-cotedazur.fr)

MSc Data Science & Artificial Intelligence  
Université Côte d'Azur

October 25, 2021

# Tentative Content

## ① Introduction

- Preliminaries and Machine Learning basics
- Motivation for ethics in Artificial Intelligence

## ② Sources of unfairness

- Bias in data
- Algorithmic unfairness: Examples

## ③ Fairness criteria

- Types of discrimination
- Definitions of fairness
- Fairness analysis of datasets

## ④ Algorithms and methods for fair ML

- Pre-processing methods
- In-processing methods
- Post-processing methods
- Legal and policy perspectives

## Notation

- Dataset  $(X, A, Y)$
- $X$ : non-protected features
- $A$ : protected attribute
- $Y \in \{-1, +1\}$  label, class membership
- $\hat{Y} = f(X, A)$ : prediction

## Disclaimer: abuse of notation

- feature  $\equiv$  attribute  $\equiv$  variable  $\equiv$  characteristic
- protected  $\equiv$  sensitive

## Disparate treatment

A decision making process suffers from disparate treatment if its decisions are partly based on the subject's sensitive attribute information.

## Disparate impact

A decision making process suffers from disparate impact if its outcomes disproportionately hurt people with certain sensitive attribute values.

## Disparate Impact or The $p\%$ -rule

$$UNF_{DI} = \min \left( \frac{\mathcal{P}(\hat{Y} = +1|A = 1)}{\mathcal{P}(\hat{Y} = +1|A = 0)}, \frac{\mathcal{P}(\hat{Y} = +1|A = 0)}{\mathcal{P}(\hat{Y} = +1|A = 1)} \right) \times 100.$$

## Fairness Through Unawareness

It implies that  $f(X, A)$  does not use the value of  $A$  and is a legal requirement in many domains where processing sensitive information about individuals is forbidden in order to guarantee no disparate treatment. A predictor  $\hat{Y}$  satisfies unawareness if it does not use the protected attribute  $A$ , i.e.,

$$\mathcal{P}(\hat{Y}|X, A) = \mathcal{P}(\hat{Y}|X).$$

## Demographic Parity or Statistical Parity

A stronger definition of fairness compared to unawareness is **demographic parity**. A predictor  $\hat{Y}$  satisfies demographic parity with respect to protected attribute  $A$ , if  $\hat{Y}$  is independent of  $A$ , i.e.,

$$\mathcal{P}(\hat{Y}|A) = \mathcal{P}(\hat{Y}).$$

or equivalently

$$\mathcal{P}(\hat{Y}|A=0) = \mathcal{P}(\hat{Y}|A=1).$$

*The protected and unprotected groups should receive the same distribution of output values.*

### Demographic Parity

$$UNF_{DP} = |\mathcal{P}(\hat{Y} = +1|A=0) - \mathcal{P}(\hat{Y} = +1|A=1)|.$$

## Equalized odds

A predictor  $\hat{Y}$  satisfies **equalized odds** with respect to protected attribute  $A$  and outcome  $Y$ , if  $\hat{Y}$  and  $A$  are independent conditional on  $Y$ .

$$\mathcal{P}(\hat{Y} = +1|A = 0, Y = y) = \mathcal{P}(\hat{Y} = +1|A = 1, Y = y), \\ \forall y \in \{-1, +1\}.$$

*The protected and unprotected groups should have equal true positive and false positive rates.*

## Equalized Odds

$$UNF_{EOdds} = |\mathcal{P}(\hat{Y} = +1|Y = +1, A = 0) - \mathcal{P}(\hat{Y} = +1|Y = +1, A = 1)| \\ + |\mathcal{P}(\hat{Y} = +1|Y = -1, A = 0) - \mathcal{P}(\hat{Y} = +1|Y = -1, A = 1)|.$$

## Equal Opportunity

We say that a binary predictor  $\hat{Y}$  satisfies **equal opportunity** with respect to  $A$  and  $Y$  if

$$\mathcal{P}(\hat{Y} = +1 | A = 0, Y = +1) = \mathcal{P}(\hat{Y} = +1 | A = 1, Y = +1)$$

*The protected and unprotected groups should have equal true positive rate.*

### Equal opportunity

$$UNF_{EOpp} = |\mathcal{P}(\hat{Y} = +1 | Y = +1, A = 0) - \mathcal{P}(\hat{Y} = +1 | Y = +1, A = 1)|.$$

## Predictive equality

We say that a binary predictor  $\hat{Y}$  satisfies **equal opportunity** with respect to  $A$  and  $Y$  if

$$\mathcal{P}(\hat{Y} = +1 | A = 0, Y = -1) = \mathcal{P}(\hat{Y} = +1 | A = 1, Y = -1)$$

*The protected and unprotected groups should have equal false positive rate.*

## Predictive Equality

$$UNF_{PE} = |\mathcal{P}(\hat{Y} = +1 | Y = -1, A = 0) - \mathcal{P}(\hat{Y} = +1 | Y = -1, A = 1)|.$$

## Predictive Parity

We say that a binary predictor  $\hat{Y}$  satisfies **predictive parity** with respect to  $A$  and  $Y$  if

$$\mathcal{P}(Y = +1 | \hat{Y} = +1, A = 0) = \mathcal{P}(Y = +1 | \hat{Y} = +1, A = 1)$$

*Predictive parity requires the same positive predictive value (i.e., precision) in both groups.*

### Predictive parity

$$UNFPP = |\mathcal{P}(Y = +1 | \hat{Y} = +1, A = 0) - \mathcal{P}(Y = +1 | \hat{Y} = +1, A = 1)|.$$

This lecture analyses the investigation by Propublica about a commercial tool made by Northpointe, Inc. to assess the criminal defendant's likelihood of becoming a recidivist.

All original materials for this study are publicly available:

- The original story: Machine Bias.
- How they analyzed the algorithm.
- A GitHub Repository.

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

<https://github.com/propublica/compas-analysis/>

# The story

How two different lives interact with the justice system:

This real story helps us understand:

- **How** it affects individuals.
- **What** type of harms it inflicts to individuals.

## Brisha Borden

On a spring afternoon in 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's bicycle and a scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances - which belonged to a 6-year-old boy - a woman came running after them saying, "That's my kid's stuff". Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late - a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of 80\$.

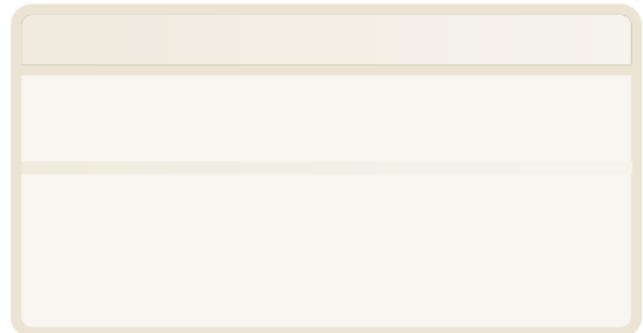
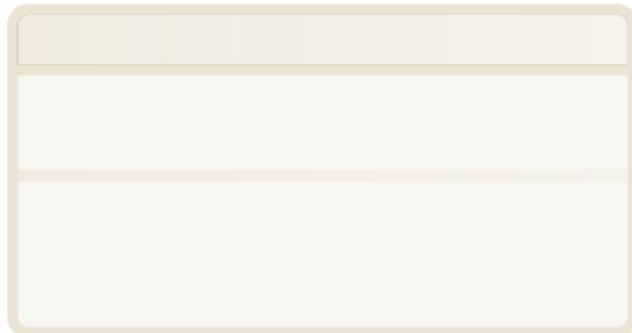
## Vernon Prater

The previous summer, 41-year-old Vernon Prater was picked up for shoplifting 86.35\$ worth of tools from a store.

Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge.

Borden had a record, too, but it was for misdemeanors committed when she was a juvenile.

When Borden and Prater were booked into jail a computer program gave a score predicting the likelihood of each committing a future crime.



When Borden and Prater were booked into jail a computer program gave a score predicting the likelihood of each committing a future crime.

### Borden

high risk

### Vernon

low risk

When Borden and Prater were booked into jail a computer program gave a score predicting the likelihood of each committing a future crime.

Borden

high risk

Vernon

low risk

Two years later, we know the computer got it **exactly backward**.

When Borden and Prater were booked into jail a computer program gave a score predicting the likelihood of each committing a future crime.

### Borden

high risk

has not been charged with any new crimes

### Vernon

low risk

is serving an eight-year prison term

Two years later, we know the computer got it **exactly backward**.

When Borden and Prater were booked into jail a computer program gave a score predicting the likelihood of each committing a future crime.

### Borden

high risk

has not been charged with any new crimes

### Vernon

low risk

is serving an eight-year prison term

Two years later, we know the computer got it **exactly backward**.

Borden is **black**  
Prater is **white**

# COMPAS algorithm: recidivism prediction

## Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

*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. (Source: ProPublica analysis of data from Broward County, Fla.)*

Were the risk scores reasonable?

In this case, COMPAS risk scores were **incorrect**.

What harms came from these incorrect decisions?

- Borden: the score influencing the judge's decisions for setting bail. The impact of COMPAS's poor decision resulted in the girl spending the night in jail.
- Other departments use COMPAS scores in trial and sentencing. Similar poor algorithmic decisions elsewhere may impact the amount of time spent in prison, future job prospects, and the right to vote.

## Unfortunately...

What explains such a discrepancy in COMPAS scores?

The COMPAS algorithm is a **black box** proprietary algorithm that we can only indirectly investigate.

# The COMPAS dataset

An acronym for Correctional Offender Management Profiling for Alternative Sanctions

- An assistive software used to predict recidivism risk.
- Helpful in ways that it provides scores from 1 (being lowest risk) to 10 (being highest risk).
- And a categorical feature: high risk, medium risk or low risk of recidivism.
- For simplicity: medium risk and high risk of recidivism vs. low risk of recidivism.
- The original input dataset used for prediction of recidivism contains 137 variables.
- Race is not an explicit feature considered by the model.

# ProPublica

ProPublica obtained two years worth of COMPAS scores from the Broward County Sheriff's Office in Florida, through a public records request, the **Freedom of Information Act.**

- ProPublica is an independent, nonprofit newsroom that produces investigative journalism with moral force.

## Aim

To expose abuses of power and betrayals of the public trust by government, business, and other institutions, using the moral force of investigative journalism to spur reform through the sustained spotlighting of wrongdoing.

<https://www.propublica.org/> <https://www.foia.gov/>

## How they acquired the data

- Through the public records request, ProPublica obtained COMPAS scores and for all 18,610 people who were scored in 2013 and 2014.
- Three COMPAS scores: "Risk of Recidivism", "Risk of Violent recidivism" and "Risk of Failure to Appear".
- Scores range from 1 to 10. Scores 1 to 4 were labeled by COMPAS as "Low"; 5 to 7 were labeled "Medium"; and 8 to 10 were labeled "High".
- Starting with the database of COMPAS scores, they built a profile of each person's criminal history, matching the criminal records to the COMPAS records.
- To determine race, they used the race classifications used officially, which identifies defendants as African-American, Caucasian, Hispanic, Asian and Native American.

# The data

The data from which the risk-scores are derived come from a combination of answers to a 137 question survey and the defendant's criminal record.

These variables include:

- Prior arrests and convictions
- Address of the defendant
- Whether the defendant a suspected gang member
- If the defendant's parents separated
- If friends/acquaintances of the defendant were ever arrested
- Whether drugs are available in the defendants neighborhood
- How often the defendant has moved residences
- The defendants high school GPA
- How much money the defendant has
- How often the defendant feels bored or sad

<https://www.documentcloud.org/documents/>

## Risk Assessment

PERSON			
Name:	Offender #:		DOB:
[REDACTED]	[REDACTED]	[REDACTED]	[REDACTED]
Male	Marital Status: Single	Agency: DAJ	
ASSESSMENT INFORMATION			
Case Identifier:	Scale Set: Wisconsin Core - Community Language	Screener:	Screening Date:
[REDACTED]	[REDACTED]	[REDACTED]	[REDACTED]

### Current Charges

- |   |  |   |   |
|---|--|---|---|
| <input type="checkbox"/> Homicide               | <input checked="" type="checkbox"/> Weapons    | <input checked="" type="checkbox"/> Assault | <input type="checkbox"/> Arson            |
| <input type="checkbox"/> Robbery                | <input type="checkbox"/> Burglary              | <input type="checkbox"/> Property/Larceny   | <input type="checkbox"/> Fraud            |
| <input type="checkbox"/> Drug Trafficking/Sales | <input type="checkbox"/> Drug Possession/Use   | <input type="checkbox"/> DUI/CUIL           | <input checked="" type="checkbox"/> Other |
| <input type="checkbox"/> Sex Offense with Force | <input type="checkbox"/> Sex Offense w/o Force |   |   |

- Do any current offenses involve family violence?  
 Yes  No
- Which offense category represents the most serious current offense?  
 Misdemeanor  Non-violent Felony  Violent Felony
- Was this person on probation or parole at the time of the current offense?  
 Probation  Parole  Both  Neither
- Based on the screener's observations, Is this person a suspected or admitted gang member?  
 No  Yes
- Number of pending charges or holds?  
 0  1  2  3  4+
- Is the current top charge felony property or fraud?  
 No  Yes

### Criminal History

Exclude the current case for these questions.

- How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?  
5

- ProPublica also conducted public records research to determine which defendants re-offended in the two years following their COMPAS screening.
- They were able to follow up on approximately half the defendants.
- This dataset contains a field *two\_year\_recid* that is 1 if the defendant re-offended within two years of screening and 0 otherwise. Following the notation of the course, this represents the label  $y$ .
- We will concern ourselves with comparing the black and white populations, as in the article.
- Similarly, we will consider a COMPAS score of either 'Medium' or 'High' to be a prediction that the defendant will re-offend within two years.

# Outcome

- The true outcome being modeled by COMPAS is whether a defendant will commit another crime upon early release from custody.
- This true outcome is unobservable and requires simplification in a number of ways:
  - A time-frame must be set for observing whether someone re-offends (Northpointe sets two years).
  - To be observed re-offending, the police must come into contact with, arrest, and charge the defendant.
  - There will be defendants in the training set that are incorrectly labeled as a 'non-re-offender', only because they committed a crime that was not pursued.
  - There will **likely be bias** in this mislabeling of re-offenders due to **studied** police behavior favoring white communities over black communities.

## How the score is used in practice

- The COMPAS score is used at the pretrial detention, trial, sentencing, and parole steps of the justice system.
- Given that the risk-assessment is supposed to model 'likelihood of re-offending' of a certain type, the developers of COMPAS only recommend using the algorithm to judge decisions like early release with access to social services.
- However, once the score became available to the criminal justice system as a whole, it became used in wholly inappropriate ways.
- For example, whether someone may commit another crime in the future has no bearing on whether they did or did not commit the current crime in question.

- The score itself does **not** actually make the decision.
- It's another piece of information that judges and juries use when making more holistic decisions.
- However, as the output of this model is nothing more than an integer, it does not explain to the decision maker how to weight this information.

## Basic statistical descriptive analysis

The first step in a statistical descriptive analysis of a sample of data: obtain **tables** or other visualization outputs that allow **summarize** and **order** the data, helping its posterior analysis.

- Let us consider a sample composed of  $n$  individuals, for which we will observe variable  $X$ , having  $n$  data:  $x_1, x_2, \dots, x_n$ .
- Let  $x_1, \dots, x_k$  the  $k$  different **values** observed.

## Absolute frequency

The frequency (or absolute frequency) of an event  $x_i$  is the number  $n_i$  of times the observation occurred in an experiment.

$$\sum_{i=1}^k n_i = n$$

## Relative frequency

The relative frequency of  $x_i$ , denoted by  $f_i$ , is the proportion of occurrences observed for this event, i.e.,

$$f_i = \frac{n_i}{n}, \quad 1 \leq i \leq k$$

$$\sum_{i=1}^k f_i = 1$$

# Frequency distribution

A frequency distribution is a table (*frequency table*) or graph (bar plot or histogram) that displays the frequency of the events in a sample. Each entry in the table contains the frequency of the occurrences of values within a particular group or interval.

$x_i$	$n_i$	$N_i$	$f_i$	$F_i$
$x_1$	$n_1$	$N_1$	$f_1$	$F_1$
$x_2$	$n_2$	$N_2$	$f_2$	$F_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_k$	$n_k$	$N_k$	$f_k$	$F_k$
	$n$		1	

# Example

The COMPAS score of recidivism calculated for the 15 defendants on a given day was:

4, 3, 7, 5, 6, 4, 5, 4, 5, 6, 7, 7, 3, 4, 5

## Example

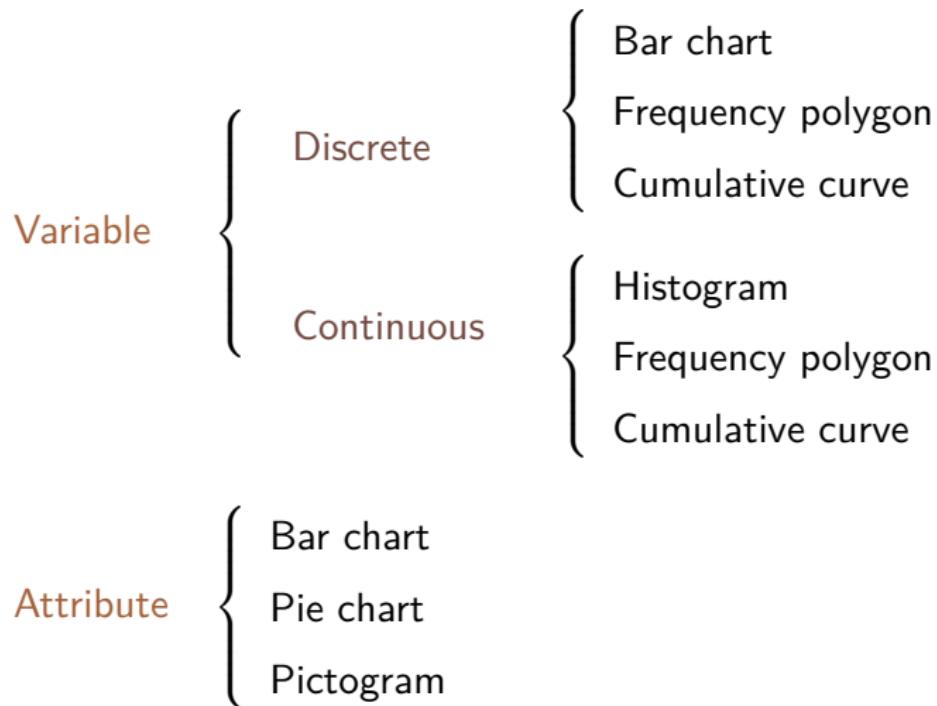
The COMPAS score of recidivism calculated for the 15 defendants on a given day was:

4, 3, 7, 5, 6, 4, 5, 4, 5, 6, 7, 7, 3, 4, 5

The frequency table for this sample is:

$x_i$	$n_i$	$N_i$	$f_i$	$F_i$
3	2	2	0.133	0.133
4	4	6	0.266	0.4
5	4	10	0.266	0.666
6	2	12	0.133	0.8
7	3	15	0.2	1
	15		1	

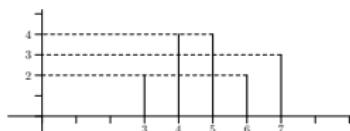
# Graphical representations



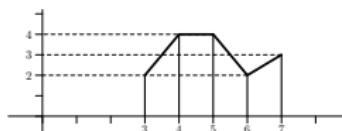
## Example: discrete variable

Let us consider again the variable  $X$  = “defendant’s COMPAS risk score”.

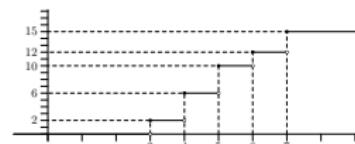
$x_i$	$n_i$	$N_i$	$f_i$	$F_i$
3	2	2	0.133	0.133
4	4	6	0.266	0.4
5	4	10	0.266	0.666
6	2	12	0.133	0.8
7	3	15	0.2	1
	15		1	



Bar chart



Frequency polygon

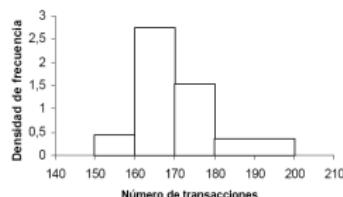


Cumulative curve

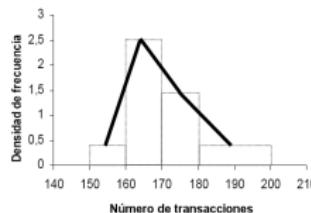
## Example: continuous variable

Let us consider the variable  $X$  = "Height in cm", observed in  $n = 50$  defendants.

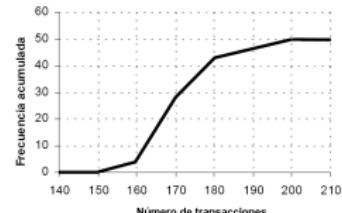
$(L_{i-1}, L_i]$	$n_i$	$a_i$	$h_i$	$N_i$
(150, 160]	4	10	0.4	4
(160, 170]	25	10	2.5	29
(170, 180]	14	10	1.4	43
(180, 200]	7	20	0.35	50



Histogram



Frequency polygon

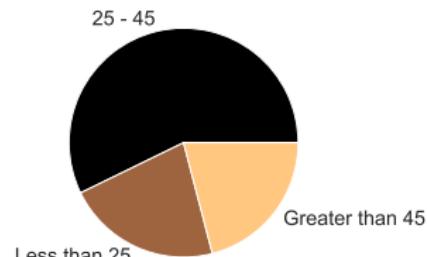


Cumulative curve

## Pie chart

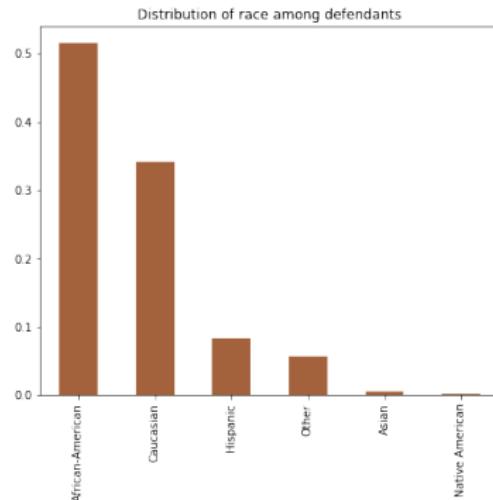
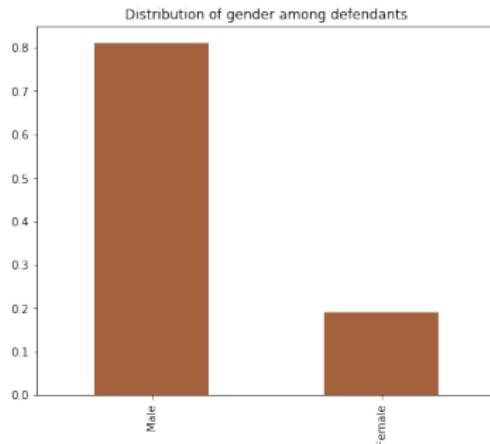
Within a circle, each category is assigned a sector proportional to its frequency.

Defendant's age	$n_i$	$f_i$	$f_i \times 360^\circ$
Less than 25	1347	0.22	$79^\circ$
Between 25 and 45	3532	0.57	$206^\circ$
Greater than 45	1293	0.21	$75^\circ$
	6172	1	$360^\circ$



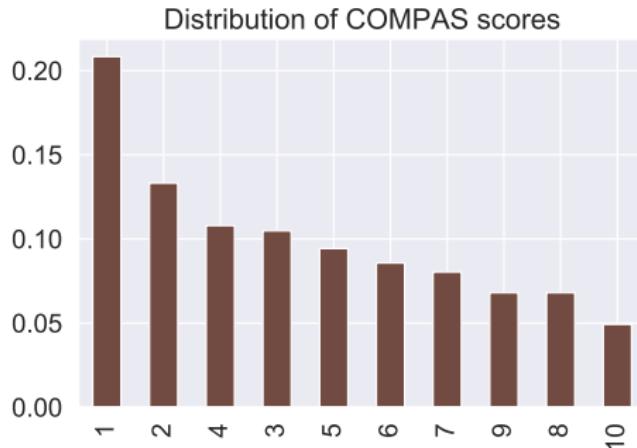
# Analysis of Fairness

- We will now investigate the COMPAS algorithm on the data collected by ProPublica.
- We will analyze the COMPAS scores for "Risk of Recidivism". An equivalent analysis could be made for "Risk of Violent Recidivism".
- We will analyse: distribution of gender, races, distribution of the COMPAS decile scores for different groups, fairness analysis.



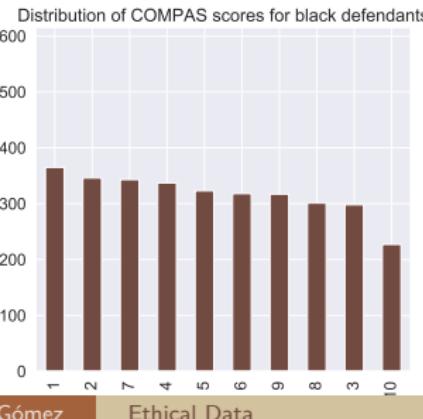
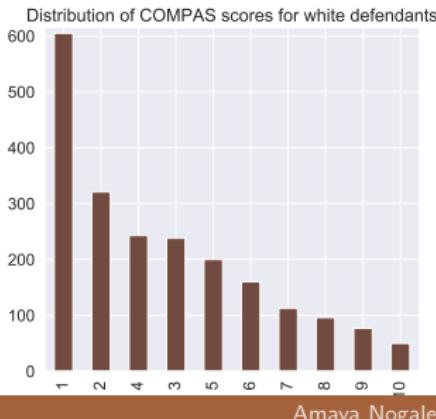
Almost 80% of defendants are classified as male, while the white and black defendants comprise of approximately 85% of the total population of defendants.

- We analyze the COMPAS scores for "Risk of Recidivism".
- We plot the distribution of the COMPAS decile scores.
- We plot the distribution of these scores for all 6,172 defendants who had not been arrested for a new offense or who had recidivated within two years.



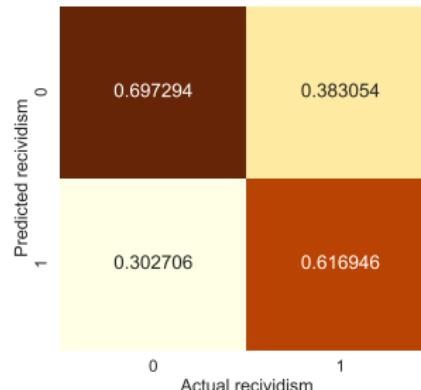
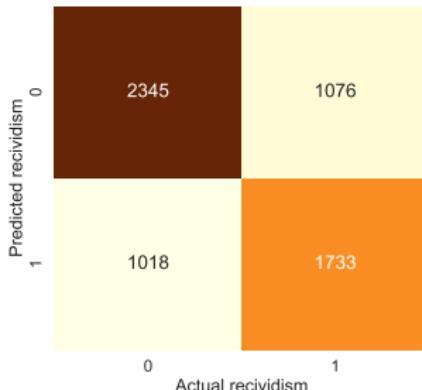
## Comparing the COMPAS score based on race

- There is a qualitative difference in the distributions among the Black and white defendants.
- Scores for white defendants were skewed toward lower-risk categories
- Scores for black defendants were evenly distributed across scores.
- These observations do not prove any demographic or behavioral bias.



- The COMPAS algorithm, on the dataset as a whole, is relatively balanced.
- The positive class (actual recidivists) represent a 46% of the dataset, which slightly coincides with the predicted recidivists (45%).
- A 34% of the population experienced an incorrect decision, roughly balanced between false positives and false negatives.

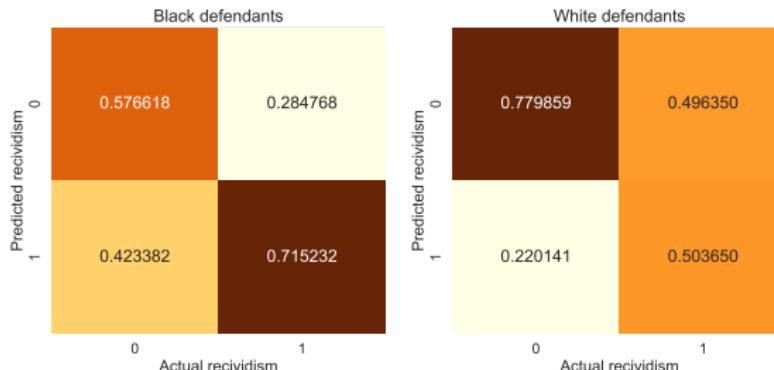
	Acc(%)	FNR	FPR
All	66.0	0.38	0.30



If we look at black and white populations separately:

- A greater proportion of black defendants experience an incorrect "will re-offend" prediction than white defendants.
- A greater proportion of white defendants experience an incorrect "won't re-offend" prediction than black defendants.

	Acc(%)	FNR	FPR
All	66.0	0.38	0.30
Black	64.9	0.28	0.42
White	67.2	0.49	0.22



## Some interesting reading

- <https://www.propublica.org/>
- [https://www.propublica.org/article/  
how-we-analyzed-the-compas-recidivism-algorithm](https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm)
- <https://github.com/propublica/compas-analysis/>
- <https://www.foia.gov/>
- [https://www.documentcloud.org/documents/  
2702103-Sample-Risk-Assessment-COMPAS-CORE.html](https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html)

# Lecture 5

## Security and Ethical Aspects of Data

Amaya Nogales Gómez  
[amaya.nogales-gomez@univ-cotedazur.fr](mailto:amaya.nogales-gomez@univ-cotedazur.fr)

MSc Data Science & Artificial Intelligence  
Université Côte d'Azur

November 22, 2021

# Tentative Content

## ① Introduction

- Preliminaries and Machine Learning basics
- Motivation for ethics in Artificial Intelligence

## ② Sources of unfairness

- Bias in data
- Algorithmic unfairness: Examples

## ③ Fairness criteria

- Types of discrimination
- Definitions of fairness
- Fairness analysis of datasets

## ④ Algorithms and methods for fair ML

- Pre-processing methods
- In-processing methods
- Post-processing methods
- Legal and policy perspectives

# Outline

- Basic concepts
- Hard-margin approach
- Soft-margin approach
- Loss functions
- Kernel methods
- Parameter selection and practical issues
- Multi-class classification
- Categorical data

# Supervised Classification: Support Vector Machines

- $\Omega$ : the population.
- Population is partitioned into two classes,  $\{-1, +1\}$ .

# Supervised Classification: Support Vector Machines

- $\Omega$ : the population.
- Population is partitioned into two classes,  $\{-1, +1\}$ .
- For each object in  $\Omega$ , we have
  - $x = (x^1, \dots, x^d) \in X \subset \mathbb{R}^d$ : predictor variables.
  - $y \in \{-1, +1\}$ : class membership.

# Supervised Classification: Support Vector Machines

- $\Omega$ : the population.
- Population is partitioned into two classes,  $\{-1, +1\}$ .
- For each object in  $\Omega$ , we have
  - $x = (x^1, \dots, x^d) \in X \subset \mathbb{R}^d$ : predictor variables.
  - $y \in \{-1, +1\}$ : class membership.
- The goal is to find a hyperplane  $\omega^\top x + b = 0$  that aims at separating, if possible, the two classes.

# Supervised Classification: Support Vector Machines

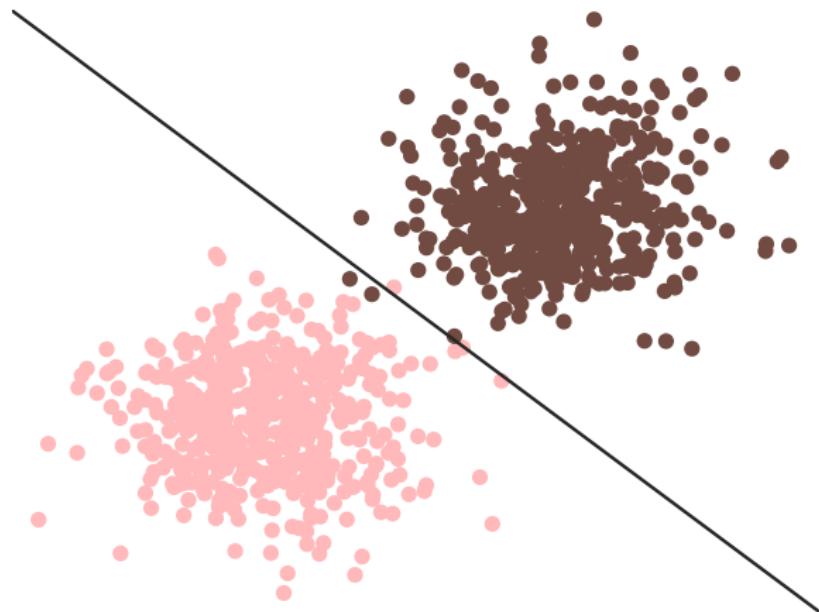
- $\Omega$ : the population.
- Population is partitioned into two classes,  $\{-1, +1\}$ .
- For each object in  $\Omega$ , we have
  - $x = (x^1, \dots, x^d) \in X \subset \mathbb{R}^d$ : predictor variables.
  - $y \in \{-1, +1\}$ : class membership.
- The goal is to find a hyperplane  $\omega^\top x + b = 0$  that aims at separating, if possible, the two classes.
- Future objects will be classified as

$$\begin{aligned} y &= +1 && \text{if } \omega^\top x + b > 0 \\ y &= -1 && \text{if } \omega^\top x + b < 0 \end{aligned} \tag{1}$$

# Supervised Classification



# Supervised Classification



# Support Vector Machines (SVM)

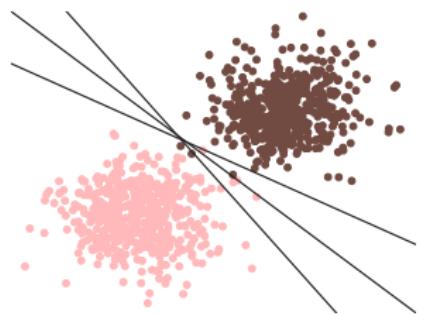
- State-of-the-art in supervised classification.
- Very good classification accuracy.
- Computationally cheap: Quadratic Programming formulation.
- SVM: In many cases competitive with existing classification methods.
- Relatively easy to use.
- Kernel techniques: many extensions.
- Regression, density estimation, kernel PCA, etc.

# Separating hyperplanes

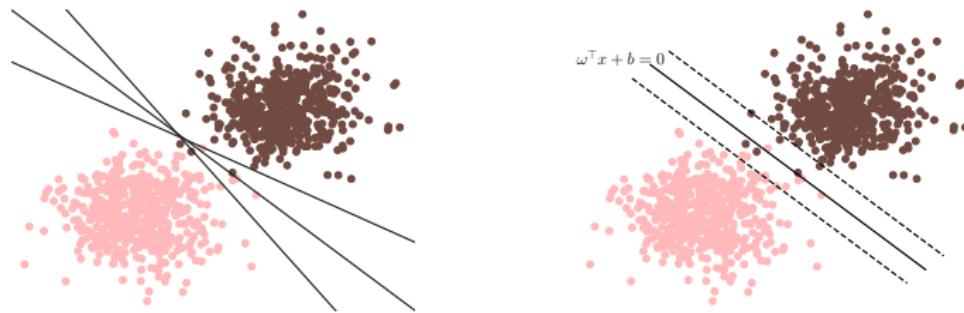
- Infinite possible separating hyperplanes.
- Each one with different properties (metrics).
- Expressed as:

$$\omega^\top x + b = 0$$

- In geometry, a hyperplane is a subspace whose dimension is one less than that of its ambient space.



# Maximum margin classification



The SVM aims to find the boundary that maximizes the margin between the classes.

# Linear algebra of a hyperplane

- $H$ : hyperplane defined by  $\omega^\top x + b = 0$
- Key properties:
  - ① For any  $x_1, x_2 \in H$ 
    - $\omega^\top(x_1 - x_2) = 0$  and
    - $\bar{\omega} = \omega/\|\omega\|$  is the vector normal to  $H$ .
  - ② For any  $x_0 \in H$ ,  $\omega^\top x_0 = -b$

## Distance of any point to the hyperplane

- The signed distance of any point  $x$  to the  $H$  is the projection of vector  $v$  ( $x - x_0$ , with  $x_0$  being intersection point of  $H$  and the normal vector) into the normal vector.
- We obtain this projection via the dot product:

$$\begin{aligned}\bar{\omega} \cdot v &= \frac{\omega^\top}{\|\omega\|} \cdot (x - x_0) = \\ &= \frac{1}{\|\omega\|} (\omega^\top x - \omega^\top x_0) = \\ &= \frac{1}{\|\omega\|} (\omega^\top x + b)\end{aligned}$$

- Distance of object  $i$  to the hyperplane  $H$ :

$$d(x_i, H) = \frac{y_i}{\|\omega\|} (\omega^\top x_i + b)$$

# Optimization problem

Let us recall that the margin width is the distance from the decision boundary to the closest point.

We want to find the margin as large as possible (maximization)

SVM seeks to maximize, as a function of  $\omega, b$ , the quantity:

$$\arg \max_{\omega, b} \left\{ \min_i d(x_i, H) = \frac{1}{\|\omega\|} \min_i y_i (\omega^\top x_i + b) \right\}$$

## Hard-Margin approach

- Training sample assumed to be linearly separable, i.e., the convex hull of the two groups are not empty and they do not overlap.
- All objects in the training sample must be correctly classified!
- The separating hyperplane is the one maximizing the smallest distance to misclassification.

## Hard-margin SVM: maximal margin

- Distance between  $\omega^\top x + b = +1$  and  $\omega^\top x + b = -1$ :

$$2/\|\omega\| = 2/\sqrt{\omega^\top \omega}$$

- A quadratic programming problem with linear constraints.

$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

s.t.

$$\begin{aligned} y_i(\omega^\top x_i + b) &\geq 1 & \forall i = 1, \dots, n \\ \omega &\in \mathbb{R}^d \\ b &\in \mathbb{R}. \end{aligned}$$

But...

Linearly separable data

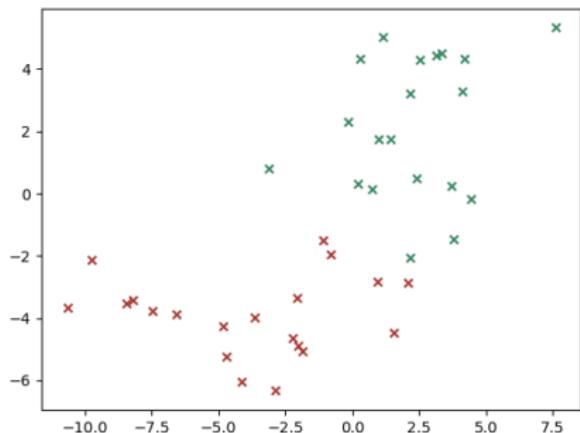
$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$



But...

$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

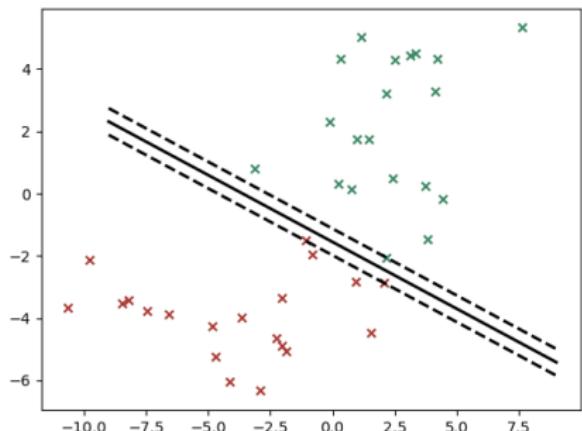
s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$

Linearly separable data



But...

Non-linearly separable data

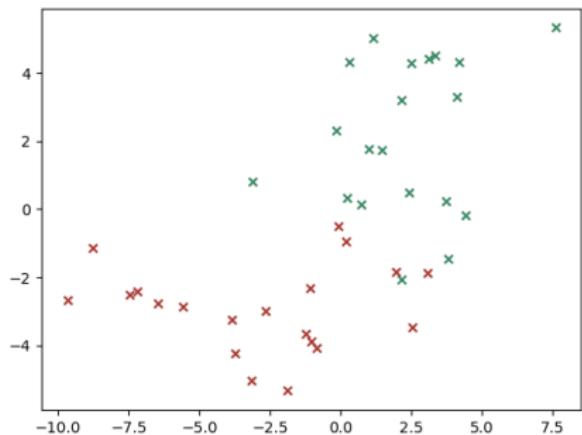
$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$



But...

Non-linearly separable data

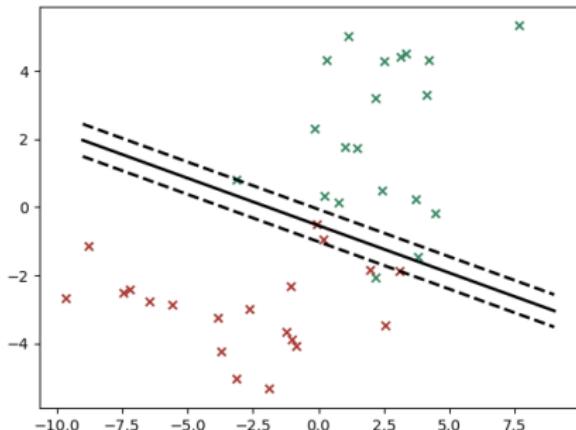
$$\min_{\omega, b} \frac{1}{2} \sum_{j=1}^d \omega_j^2$$

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$



INFEASIBLE!!

# Hard margin SVM: limitations

- Real data, most likely, will not meet the of linear separable assumption
- Hard margin loss is too limiting when there is class overlapping
- Hard margin SVM will not be able to deal with it
- Possible solutions:
  - Keep hard margin SVM but transform the data: mapping into higher dimensional (maybe infinite) feature space

$$\phi(x) = (\phi_1(x), \phi_2(x), \dots)$$

- Relax the constraints (allow training errors)
- Combination of both

# A solution for non-linearly separable data

- When data are not linearly separable the hard-margin SVM problem is infeasible.
- In the **soft-margin approach**, constraints

$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

are perturbed.

- How? By introducing auxiliary variables  $\xi_i$ , making the new problem always feasible.

# Building the soft-margin SVM

$$\min_{\omega, b, \xi} \frac{1}{2} \sum_{j=1}^d \omega_j^2 + \frac{C}{n} \sum_{i=1}^n g_i(\xi_i)$$

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}$$

$$\xi_i \geq 0 \quad \forall i = 1, \dots, n,$$

- $\xi = (\xi_i) \in \mathbb{R}^n$  is the vector of deviation variables.
- $g_i$  is the loss function (convex and increasing).
- Most popular choices: *hinge* loss,  $g_i(t) = C_i t$  or *squared hinge* loss,  $g_i(t) = C_i t^2$ .
- $C$  is a tuning parameter.

# The SVM formulation

Non-linearly separable data

$$\min_{\omega, b, \xi} \frac{1}{2} \sum_{j=1}^d \omega_j^2 + \frac{C}{n} \sum_{i=1}^n \xi_i$$

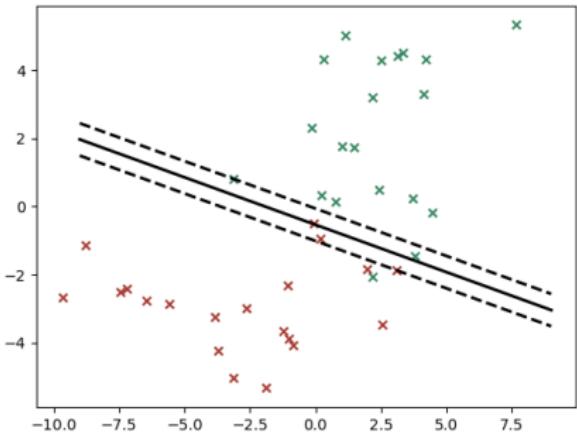
s.t.

$$y_i(\omega^\top x_i + b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}$$

$$\xi_i \geq 0 \quad \forall i = 1, \dots, n.$$

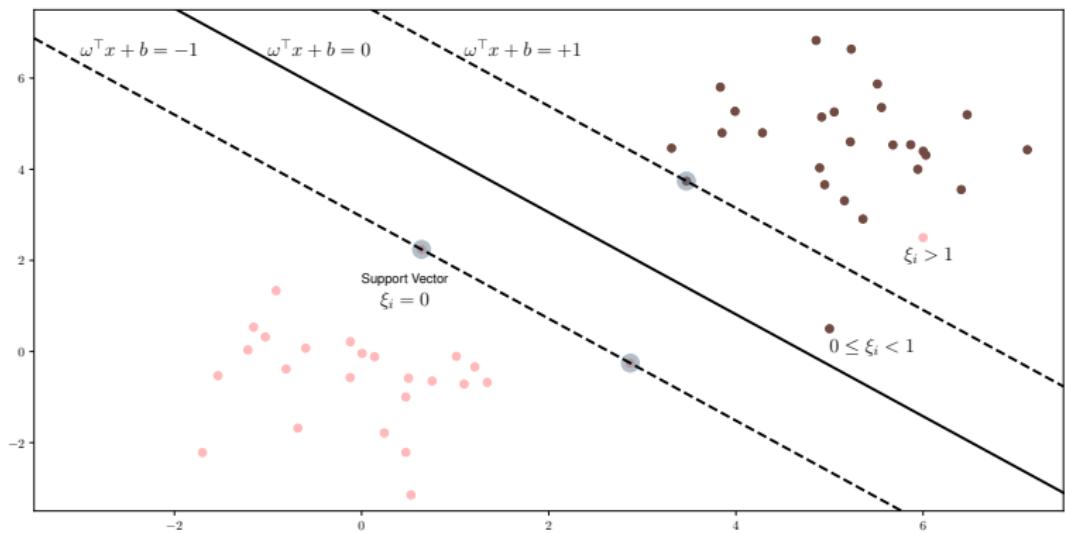


- An object  $i$  will be correctly classified if  $0 \leq \xi_i < 1$
- Misclassified if  $\xi_i > 1$ .
- In the case  $\xi_i = 1$ , we get a tie (objects coincide with the hyperplane).
- $\sum_{i=1}^n \xi_i$  is an upper bound of the number of misclassified objects.

# Soft Margin SVM

- Soft margin SVM relaxes the constraint to allow points to be inside the margin or even on the wrong side of the boundary
- The boundaries are penalized by a quantity that reflects the extent of the violation
- Slack variables  $\xi_i \geq 0$  for each sample to measure the extent of the violation.

# Slack variables



# Slack variables

- For points on or inside the correct margin:

$$\xi_i = 0$$

- For other points:

$$\xi_i = 1 - y_i(\omega^\top x_i + b)$$

- If a point is in the decision boundary:

$$\xi_i = 1$$

- The hard margin constraint:

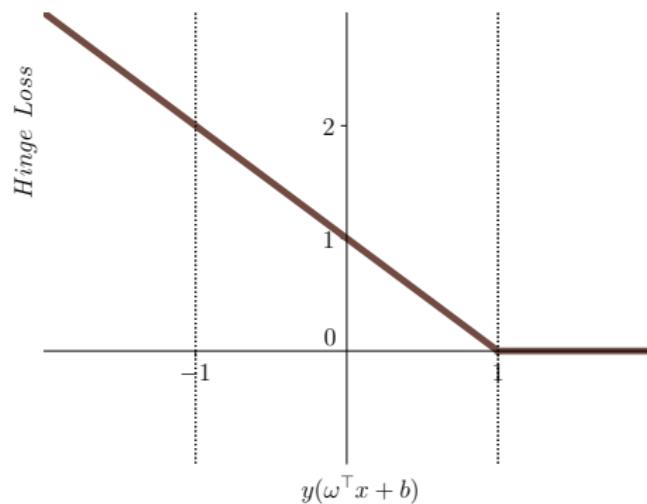
$$y_i(\omega^\top x_i + b) \geq 1 \quad \forall i = 1, \dots, n$$

- Now becomes:

$$y_i(\omega^\top x_i + b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n$$

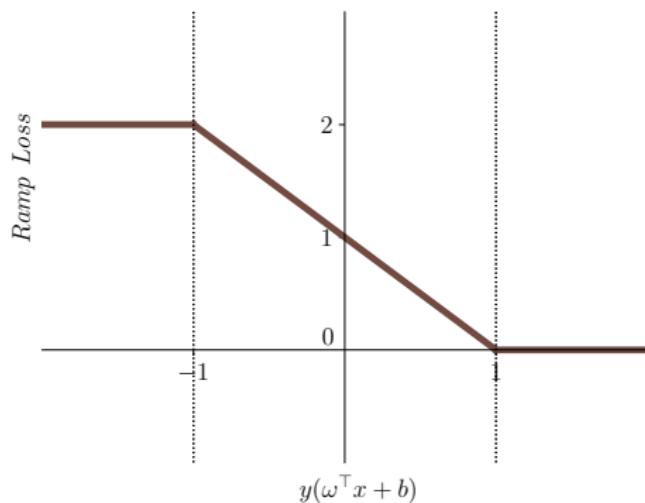
# Hinge Loss

$$\ell(x, y) = \max(0, 1 - y(\omega^\top x + b)) = \begin{cases} 0 & \text{if } y(\omega^\top x + b) \geq 1 \\ 1 - y(\omega^\top x + b) & \text{if } y(\omega^\top x + b) \leq 1 \end{cases}$$



# Ramp Loss

$$\ell(x, y) = \begin{cases} 0 & \text{if } y(\omega^\top x + b) \geq 1 \\ 1 - y(\omega^\top x + b) & \text{if } -1 \leq y(\omega^\top x + b) \leq 1 \\ 2 & \text{if } y(\omega^\top x + b) \leq -1 \end{cases}$$



## Example: Type A data

Both classes have the identity matrix as covariance.

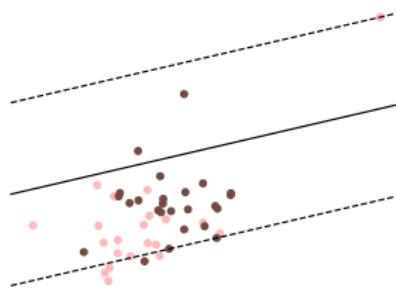
- Class +1: mean is the origin.
- Class -1: mean is  $(2/(d), \dots, 2/(d))$ .

The training data sets are contaminated with outliers. Outlier observations are sampled for Class +1 using a Gaussian distribution with covariance matrix 0.001 times the identity matrix and with a mean  $(10/(d), \dots, 10/(d))$ .

Brooks, J.P. *Support vector machines with the ramp loss and the hard margin loss*.  
Operations Research: 59(2), 467-479 (2011)

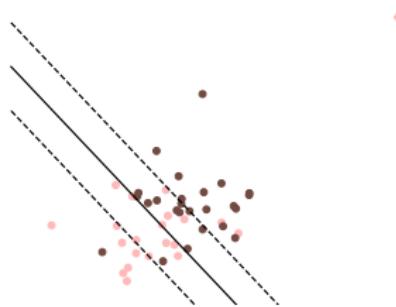
# Robustness to outliers

SVM with the Hinge Loss



Accuracy of 44%

SVM with the Ramp Loss



Accuracy of 78%

# SVM with the Ramp Loss

$$\min_{\omega, b, \xi, z} \frac{1}{2} \sum_{j=1}^d \omega_j^2 + \frac{C}{n} \left( \sum_{i=1}^n \xi_i + 2 \sum_{i=1}^n (1 - z_i) \right)$$

s.t.

$$(y_i(\omega^\top x_i + b) - 1 + \xi_i) \cdot \textcolor{blue}{z}_i \geq 0 \quad \forall i = 1, \dots, n$$

$$0 \leq \xi_i \leq 2 \quad \forall i = 1, \dots, n$$

$$z \in \{0, 1\}^{\textcolor{blue}{n}}$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$

# The kernel trick

- Soft-margin SVM<sup>1</sup>:

$$\min_{\omega, b, \xi} \frac{1}{2} \sum_{j=1}^d \omega_j^2 + \frac{C}{n} \sum_{i=1}^n g_i(\xi_i)$$

s.t.

$$y_i(\omega^\top \phi(x_i) + b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n$$

$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}$$

$$\xi_i \geq 0 \quad \forall i = 1, \dots, n,$$

(2)

- Kernel trick. Example:  $x \in \mathbb{R}^3, \phi(x) \in \mathbb{R}^{10}$

$$\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, x_1^2, x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3)$$

<sup>1</sup> Cortes, C., Vapnik, V. *Support-vector networks*. Machine learning, 20(3), 273-297.

## Definition (Kernel)

$k : X \times X \rightarrow \mathbb{R}$  is a kernel if

- ①  $k$  is symmetric:  $k(x_1, x_2) = k(x_2, x_1)$ .
- ②  $k$  is positive semi-definite, i.e.,  $\forall x_1, x_2, \dots, x_n \in X$ , the "Gram Matrix"  $K$  defined by  $K_{ij} = k(x_i, x_j)$  is positive semi-definite. (A matrix  $M \in \mathbb{R}^{n \times n}$  is positive semi-definite if  $\forall a \in \mathbb{R}^n$ ,  $a'Ma \geq 0$ .)

Kernel  $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle = \phi(x_1)^\top \phi(x_2)$ . Most popular kernels:

- Linear

$$k(x_1, x_2) = \langle x_1, x_2 \rangle$$

- Radial Basis Function (RBF)

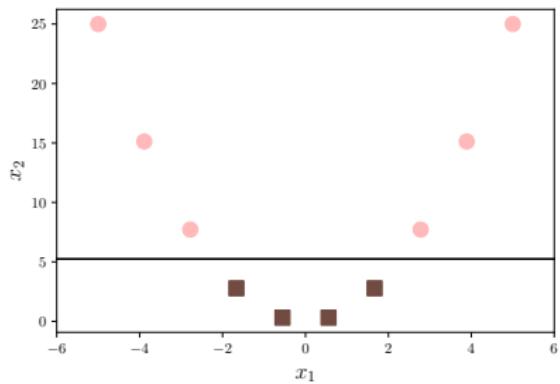
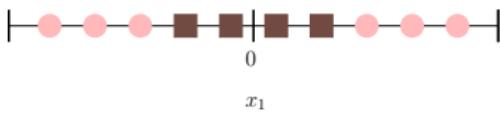
$$k(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|^2}$$

- Polynomial kernel (of dimension  $d$ ):

$$k(x_1, x_2) = (x_1^\top x_2 + c)^d$$

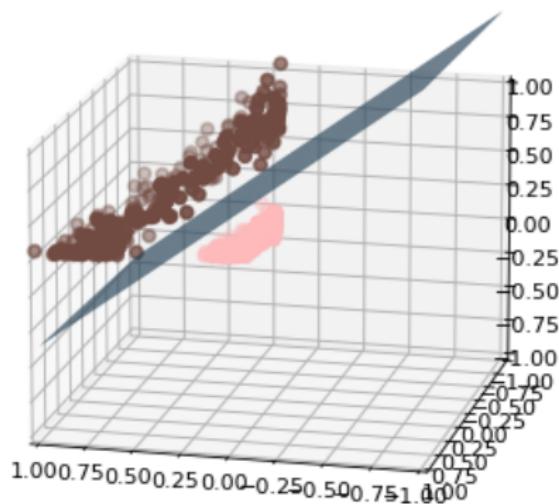
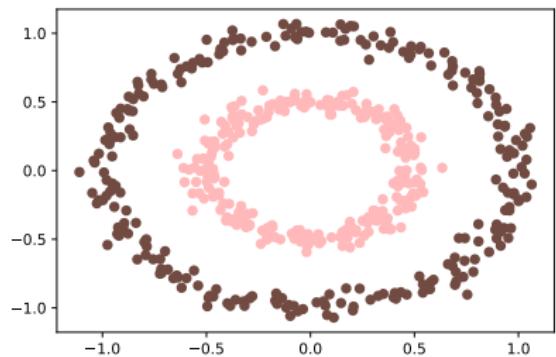
# How do we know kernels help separating data?

- In  $\mathbb{R}^d$ , any  $d$  independent vectors are linearly separable.
- If  $k$  is positive definite  $\rightarrow$  data linearly separable.
- Example:  $x_1 \in \mathbb{R}$ ,  $\Phi(x_1) = (x_1, x_1^2) \in \mathbb{R}^2$



Example:  $\mathbb{R}^2 \rightarrow \mathbb{R}^3$

$$x \in \mathbb{R}^2, \phi(x) \in \mathbb{R}^3, \phi(x) = (x_1^2, \sqrt{2}x_1, x_2, x_2^2)$$





# Parameter selection

- Important step of the ML cycle.
- Parameters:  $C$ , kernel parameters.
- Example:

$$\gamma \text{ in } e^{-\gamma \|x-y\|^2}$$

$$c, d \text{ in } (x^\top y + c)^d$$

- How to select them?

And as well:

- How to select kernels? RBF, polynomial,...
- How to select methods? SVM, decision trees,...

# Performance evaluation

- In practice:  
available data → training, testing and validation
- Train in the training.
- Test in the testing.
- Report in the validation.
- K-fold cross-validation.

# K-fold cross-validation

For each  $k$ ,

- (i) Split the dataset into training, testing and validation sets.
- (ii) For each  $C$ ,

Solve the SVM in the training set and obtain the solution  $(\omega^C, b^C, \xi_i^C)$ .

- (iii) Choose the optimal  $C^*$  in the testing set.

Report the quality metric for the classifier  $(\omega^{C^*}, b^{C^*}, \xi_i^{C^*})$  in the validation set.

# Multi-class classification

- $k$  classes
- One-against-all: train  $k$  binary SVMs

1st class vs.  $(2 - k)$ th class  
2nd class vs.  $(1, 3 - k)$ th class

⋮

- $k$  decision functions

$$(\omega^1)^\top x + b^1$$

⋮

$$(\omega^k)^\top x + b^k$$

- Prediction

$$\arg \max_j (\omega^j)^\top x + b^j$$

- Reason: If the 1st class, then we should have

$$(\omega^1)^\top x + b^1 \geq 0$$

$$(\omega^2)^\top x + b^2 \leq 0$$

⋮

$$(\omega^k)^\top x + b^k \leq 0$$

- One-against-one: train  $k(k - 1)/2$  binary SVMs
- Example: 4 classes  $\rightarrow$  6 binary SVMs

$y_i = +1$	$y_i = -1$	Decision functions
Class 1	Class 2	$f^{12}(x) = (\omega^{12})^\top x + b^{12}$
Class 1	Class 3	$f^{13}(x) = (\omega^{13})^\top x + b^{13}$
Class 1	Class 4	$f^{14}(x) = (\omega^{14})^\top x + b^{14}$
Class 2	Class 3	$f^{23}(x) = (\omega^{23})^\top x + b^{23}$
Class 2	Class 4	$f^{24}(x) = (\omega^{24})^\top x + b^{24}$
Class 3	Class 4	$f^{34}(x) = (\omega^{34})^\top x + b^{34}$

- In the testing dataset, we predict all binary SVMs

Classes		winner
1	2	1
1	3	1
1	4	1
2	3	2
2	4	4
3	4	3

- Select the one with the largest vote

Class	1	2	3	4
#votes	3	1	1	1

# Quality of a classifier

- Accuracy: percentage of objects correctly classified.
- Sensitivity (True Positive Rate): the proportion of those who received a positive prediction out of those who actually belong to the positive class.
- Specificity (True Negative Rate): the proportion of those who received a negative prediction out of those who actually belong to the negative class.
- Sparsity (vs. complexity):  $\frac{\#(w_j=0, j=1, \dots, d)}{d} \cdot 100$

And fairness!!!

Equal opportunity, demographic parity, predictive parity, predictive equality, p%-rule, unawareness...

# The COMPAS dataset

- A dataset from the investigation by Propublica about a commercial tool made by Northpointe, Inc. to assess the criminal defendant's likelihood of becoming a recidivist.
- <https://github.com/propublica/compas-analysis>
- Sample size:  $n = 7214$ .
- Dimension:  $d = 52$ .
  - 2 continuous features.
  - 17 discrete features.
  - 33 categorical features.

# How do we deal with categorical features?

## Race

The categorical feature race has the following categories:  
African-American, Caucasian, Hispanic, Asian, Native American,  
Other.

- We binarize the categorical feature *race* into 6 binary features.

i	Race	African-American	Caucasian	Hispanic	Asian	Native American	Other
1	Asian	0	0	0	1	0	0
2	Other	0	0	0	0	0	1
:	:	:	:	:	:	:	:

# Bibliography

- Cortes, C., Vapnik, V. *Support-vector networks*. Machine learning, 20(3), 273-297.
- Brooks, J.P. *Support vector machines with the ramp loss and the hard margin loss*. Operations Research: 59(2), 467-479 (2011).

# Lecture 6

## Security and Ethical Aspects of Data

Amaya Nogales Gómez  
[amaya.nogales-gomez@univ-cotedazur.fr](mailto:amaya.nogales-gomez@univ-cotedazur.fr)

MSc Data Science & Artificial Intelligence  
Université Côte d'Azur

November 29, 2021

# Course overview

## ① Introduction

- Preliminaries and Machine Learning basics
- Motivation for ethics in Artificial Intelligence

## ② Sources of unfairness

- Bias in data
- Algorithmic unfairness: Examples

## ③ Fairness criteria

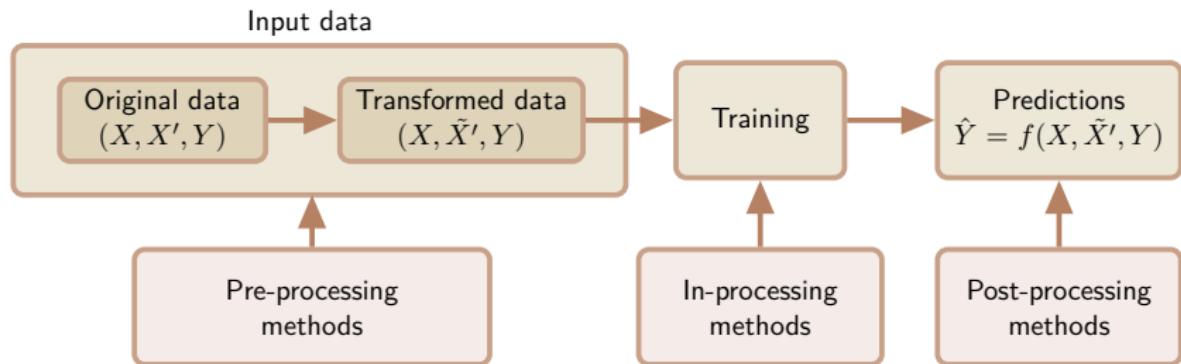
- Types of discrimination
- Definitions of fairness
- Fairness analysis of datasets

## ④ Algorithms and methods for fair ML

- Pre-processing methods
- In-processing methods
- Post-processing methods
- Legal and policy perspectives

# Methods for Fair Machine Learning

Methods that target fairness in the ML lifecycle fall under three categories:



## Pre-processing methods

- Data-based methods try to transform the data so that the underlying discrimination is removed.
- If the algorithm is allowed to modify the training data, then pre-processing can be used.
- Example: reweighting, e.g., up-weighting examples that align with a particular fairness objective.

F. Kamiran and T. Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1-33.

## In-processing methods

- Model-based methods try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process.
- If it is allowed to change the learning procedure for a machine learning model, then in-processing can be used.
- Either by incorporating changes into the objective function or imposing new constraint.
- Example: addition of regularization terms or constraints that enforce a particular fairness objective during optimization.

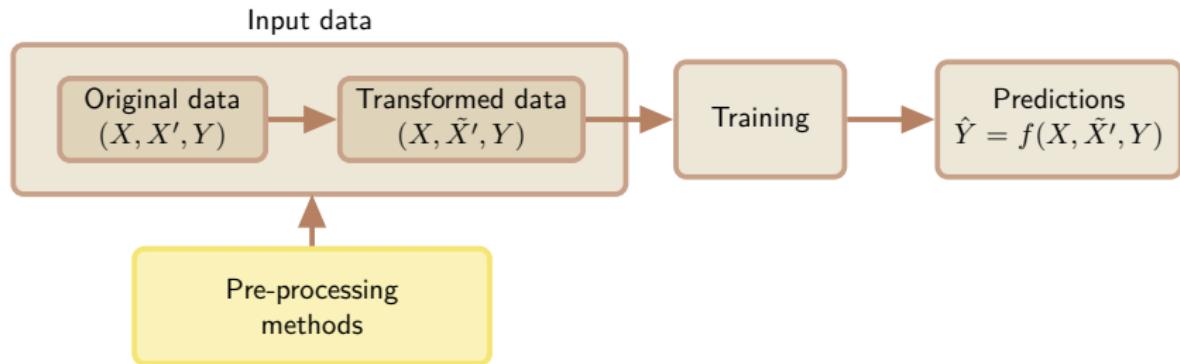
M. B. Zafar, I. Valera, M.-G. Rodriguez, and K.-P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:962-970, 2017.

## Post-processing methods

- Post-hoc methods are performed after training by accessing a set which was not involved during the training of the model.
- If the algorithm can only treat the learned model as a black box, then only post-processing can be used.
- Labels assigned by the black-box model initially get reassigned based on a function during the post-processing phase.
- Example: identifying different decision thresholds for different groups based on a predicted score, e.g., in order to equalize false positive rates.

M. Hardt, E. Price, and N Srebro. Equality of opportunity in supervised learning.  
*Advances in Neural Information Processing Systems*, 29, 2016.

# Pre-processing methods



F. Kamiran and T. Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1-33.

## Pre-processing methods

We need to define discrimination in a labeled dataset

$$disc(\Omega) = \frac{|\{x_i \in \Omega | a_i = 0, y_i = +1\}|}{|\{x_i \in \Omega | a_i = 0\}|} - \frac{|\{x_i \in \Omega | a_i = 1, y_i = +1\}|}{|\{x_i \in \Omega | a_i = 1\}|}$$

The following methods<sup>1</sup> for incorporating non-discrimination constraints into the classifier construction process are based on preprocessing the dataset after which the normal classification tools can be used to learn a classifier:

Supression

Reweighting

Massaging

<sup>1</sup> F. Kamiran and T. Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1-33.

## Supression

To reduce the discrimination between the class labels and the attribute  $A$ :

- Remove  $A$  from the dataset  $\Omega$ .
- Find the attributes that correlate most with the sensitive attribute  $A$ .
- Remove the most correlated attributes with  $A$ .

# Reweighting

- The tuples  $(x_i, a_i, y_i)$  in the training dataset are assigned weights.
- By carefully choosing the weights, the training dataset can be made discrimination-free w.r.t.  $A$  without having to change any of the labels.
- The weights on the tuples can be used directly in any classification method.

## Massaging the dataset

To remove the discrimination from the input data:

- The labels of some objects in the dataset are flipped.
  - The labels of some objects  $x_i$  with  $a_i = 1$  from  $-1$  to  $+1$ .
  - The same number of objects with  $a_i = 0$  from  $+1$  to  $-1$ .
- A good selection of which labels to change is essential.
- A ranker is used to select the best candidates for relabeling.

## Pre-processing method I: Massaging the dataset

- A ranker  $R$  for ranking the objects according to their positive class probability is learned.
- Higher scores indicate a higher chance to be in the positive class.

### Promotion candidates

$$\{x_i \in \Omega, a_i = 1, y_i = -1\}$$

### Demotion candidates

$$\{x_i \in \Omega, a_i = 0, y_i = +1\}$$

- Promotion candidates are sorted in descending order.
- Demotion candidates are sorted in ascending order.
- The first the top- $M$  elements will be chosen: the objects closest to the decision border are selected first to be relabeled.
- This modification of the training data is continued until the discrimination becomes zero.

## Pre-processing method I (cont.)

The number  $M$  of pairs needed to be modified to make a dataset  $\Omega$  discrimination-free can be calculated as follows. If we modify  $M$  pairs, the resulting discrimination will be:

$$\frac{p_{\bar{a}} - M}{|\Omega_{\bar{a}}|} - \frac{p_a + M}{|\Omega_a|} = disc(\Omega) - M \left( \frac{1}{|\Omega_a|} + \frac{1}{|\Omega_{\bar{a}}|} \right) = disc(\Omega) - \left( M \frac{|\Omega|}{|\Omega_a||\Omega_{\bar{a}}|} \right)$$

And to reach zero discrimination, we hence have to make:

$$M = \frac{disc(\Omega) \times |\Omega_a| \times |\Omega_{\bar{a}}|}{|\Omega|},$$

with  $p_a$  and  $p_{\bar{a}}$  the number of positive objects with  $a = 1$  and  $a = 0$  respectively.

## Pre-processing method I (cont.)

### Algorithm: Rank

**Input:** Dataset  $\Omega = \{(x_i, a_i, y_i)\}_{i=1}^n$

- 1: Learn a ranker  $R$  in  $\Omega$
- 2:  $pr := \{x_i \in \Omega | a_i = 1, y_i = -1\}$
- 3:  $dem := \{x_i \in \Omega | a_i = 0, y_i = +1\}$
- 4: Order  $pr$  descending w.r.t. the scores by  $R$
- 5: Order  $dem$  ascending w.r.t. the scores by  $R$

**Output:**  $(pr, dem)$ , ordered promotion and demotion list

## Pre-processing method I (cont.)

Algorithm: Learn Classifier on Massaged data

**Input:** Dataset  $\Omega = \{(x_i, a_i, y_i)\}_{i=1}^n$

- 1:  $(pr, dem) := Rank(X, A, Y)$
- 2:  $M = \frac{disc(\Omega) \times |\{x_i \in \Omega | a_i = 1\} \times |\{x_i \in \Omega | a_i = 0\}|}{|\Omega|}$
- 3: Select the top- $M$  objects of pr
- 4: Change the class label of the  $M$  objects to  $+1$
- 5: Select the top- $M$  objects of dem
- 6: Change the class label of the  $M$  objects to  $-1$

**Output:** Classifier learned on massaged dataset  $\tilde{\Omega}$

## Pre-processing method II

### Fair Cluster Support Vector Machines (FCLSV)

- A methodology to build an SVM-type classifier with categories clustered around their peers
- Clustering the  $K_j$  categories of categorical feature  $j$  into  $L_j$  clusters,  $\forall j$ .

## Pre-processing method II

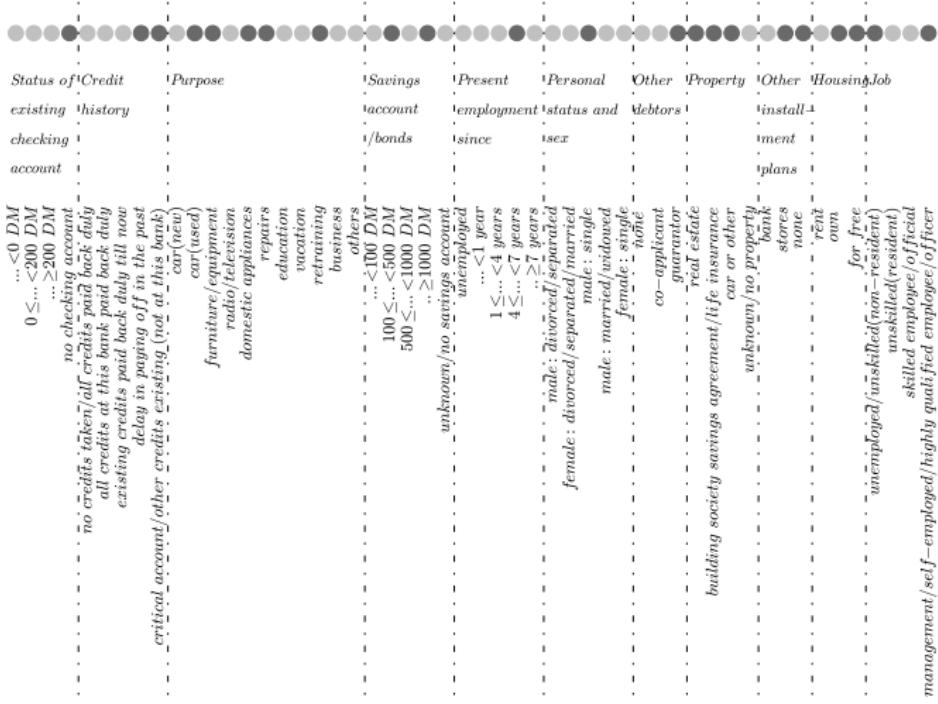
### Fair Cluster Support Vector Machines (FCLSV)

- A methodology to build an SVM-type classifier with categories clustered around their peers
- Clustering the  $K_j$  categories of categorical feature  $j$  into  $L_j$  clusters,  $\forall j$ .

With the pursue of:

- Reducing the number of relevant features, i.e., reducing the complexity of the SVM classifier
- Improving fairness (equal opportunity).
- Without important loss in accuracy

# The german dataset, $L_j = 2$



## Pre-processing method II (cont.)



*Personal status and sex*

- Categorical feature Personal status and sex has  $K_j = 5$  categories
- The  $K_j = 5$  categories have been clustered into  $L_j = 2$  clusters
  - 1st cluster associated with Personal status and sex in light grey
  - 2nd cluster associated with Personal status and sex in dark grey

## Pre-processing method II (cont.)

### Discrimination metric

Let us assume now that we have only one categorical feature  $a$  protected, having  $K_a$  categories,  $\mathcal{A} = \{a\} \subset \{0, 1\}^{K_a}$ , and with only one protected category  $a_{k^*}$ . Let us define the following general bias metric, based on the equal opportunity definition:

$$\delta_{a,k} = \frac{P(\widehat{Y}=+1|a_k=1, Y=+1)}{P(\widehat{Y}=+1|a_{k^*}=1, Y=+1)}$$

$\forall k = 1, \dots, K_a, k \neq k^*$ .

Since by definition  $k^*$  is the discriminated category, this measure takes a value  $\delta \geq 1$ .

## Pre-processing method II (cont.)

And for any *non-sensitive*  $x_j$ :

$$\delta_{j,k} = \frac{P(\widehat{Y}=+1|a_{k^*}=0, x_{j,k}=1, Y=+1)}{P(\widehat{Y}=+1|a_{k^*}=1, x_{j,k}=1, Y=+1)}$$

$\forall j = 1, \dots, J, \forall k = 1, \dots, K_j, j \neq a.$

## Pre-processing method II (cont.)

### FCLSVM algorithm

For each  $\beta, C$ ,

- 1: Solve the *SVM*, obtain solution  $\omega$ .
- 2: Compute vector  $\delta$ .
- 3: For each  $j$ , cluster the  $K_j$  categories of feature  $j$  into  $L_j$  clusters solving a K-means for  $\nu = \beta\omega_j. + (1 - \beta)\delta_j.$ , obtaining the assignment vector  $z_{j..}^*$ .
- 4: Obtain the clustered dataset  $\tilde{\Omega}$

$$(y_i, x_i, x'_i) \rightarrow (y_i, \tilde{x}_i, x'_i)$$

$$\text{where } \tilde{x}_i = (\tilde{x}_{i,j,\ell}) \text{ and } \tilde{x}_{i,j,\ell} = \sum_{k=1}^{K_j} z_{j,k,\ell}^* x_{i,j,k}$$

- 5: Solve the *SVM* for  $\tilde{\Omega}$ , and return this as the FCLSVM classifier.

# An example: Large-scale fair classification



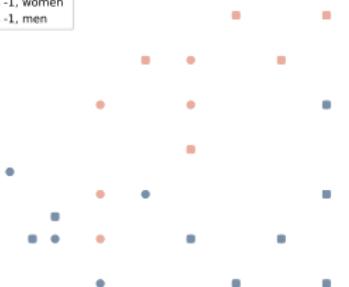
Each individual is represented by

$(x, x', y)$   $\left\{ \begin{array}{l} x, \text{vector of categorical features} \\ x', \text{vector of continuous features} \\ y \in \{-1, +1\}, \text{class membership} \end{array} \right.$

$$x = (x^1, \dots, x^d) = (\underbrace{x^1, \dots, x^k}_{\text{non-sensitive}}, \underbrace{x^{k+1}, \dots, x^d}_{\text{sensitive}})$$

# An example: Large-scale fair classification

- Class +1, women
- Class +1, men
- Class -1, women
- Class -1, men



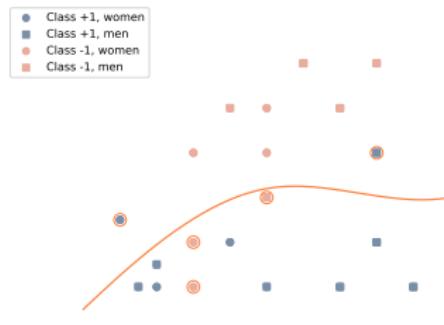
Each individual is represented by

$$(x, x', y) \left\{ \begin{array}{l} x, \text{vector of categorical features} \\ x', \text{vector of continuous features} \\ y \in \{-1, +1\}, \text{class membership} \end{array} \right.$$

$$x = (x^1, \dots, x^d) = (\underbrace{x^1, \dots, x^k}_{\text{non-sensitive}}, \underbrace{x^{k+1}, \dots, x^d}_{\text{sensitive}})$$

	Sensitive feature	non-sensitive features		decision
	Gender & status	income	credit history	
Applicant 1	male married	1.5k	5	✓
Applicant 2	female single	2.5k	3	✓

# An example: Large-scale fair classification



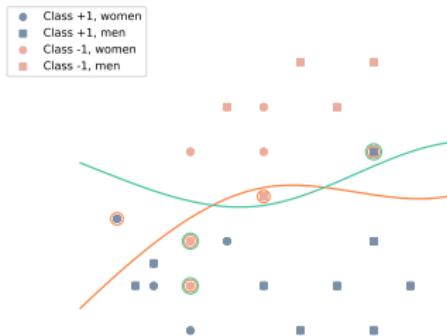
Each individual is represented by

$$(x, x', y) \left\{ \begin{array}{l} x, \text{vector of categorical features} \\ x', \text{vector of continuous features} \\ y \in \{-1, +1\}, \text{class membership} \end{array} \right.$$

$$x = (x^1, \dots, x^d) = (\underbrace{x^1, \dots, x^k}_{\text{non-sensitive}}, \underbrace{x^{k+1}, \dots, x^d}_{\text{sensitive}})$$

Clustering	Accuracy	Equal Opportunity
male single, male married, male divorced	78.3%	84.4%
female single, female married, female divorced		

# An example: Large-scale fair classification



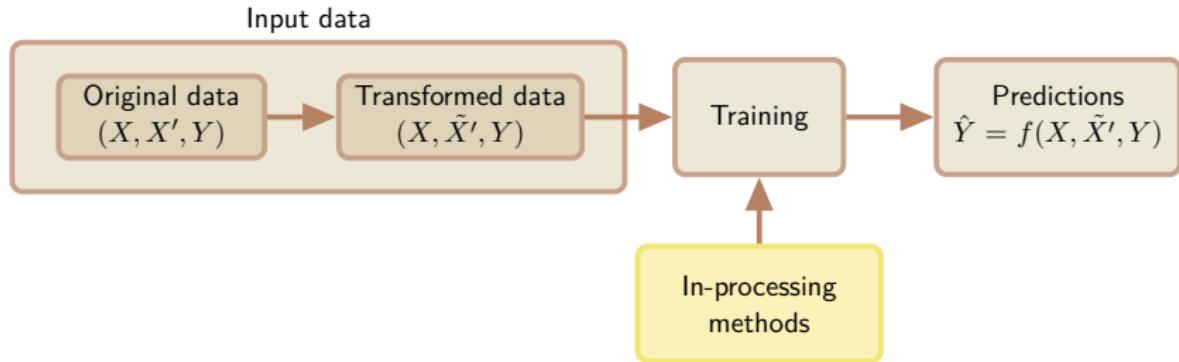
Each individual is represented by

$$(x, x', y) \left\{ \begin{array}{l} x, \text{vector of categorical features} \\ x', \text{vector of continuous features} \\ y \in \{-1, +1\}, \text{class membership} \end{array} \right.$$

$$x = (x^1, \dots, x^d) = (\underbrace{x^1, \dots, x^k}_{\text{non-sensitive}}, \underbrace{x^{k+1}, \dots, x^d}_{\text{sensitive}})$$

Clustering	Accuracy	Equal Opportunity
male single, male married, male divorced female single, female married, female divorced	78.3%	84.4%
male married	87.0%	112.5%
female single, female married, female divorced, male single, male divorced		

# In-processing methods



M. B. Zafar, I. Valera, M.-G. Rodriguez, and K.-P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:962-970, 2017.

## In-processing method

Maximizing accuracy under fairness constraints

- **Goal:** design classifiers (logistic regression and SVM) that avoid both disparate treatment and disparate impact.
- Measure of decision boundary (un)fairness: the covariance between the sensitive attributes and the (signed) distance between the objects and the decision boundary.

M. B. Zafar, I. Valera, M.-G. Rodriguez, and K.-P.Gummadi. Fairness Constraints: Mechanisms for Fair Classification.

*Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:962-970, 2017.

- Dataset  $\{(x_i, a_i, y_i)\}_{i=1}^n$
- $d_{\omega,b}(x)$  : signed distance from  $x$  to  $\omega^\top x + b = 0$
- $f_{\omega,b}(x_i) = 1$  if  $d_{\omega,b}(x_i) \geq 0$
- $f_{\omega,b}(x_i) = -1$  if  $d_{\omega,b}(x_i) \leq 0$

$$\begin{aligned}
 Cov(a, d_{\omega,b}(x)) &= \mathbb{E}[(a - \bar{a})d_{\omega,b}(x)] - \mathbb{E}[a - \bar{a}]\bar{d}_{\omega,b}(x) \\
 &= \frac{1}{n} \sum_{i=1}^n (a_i - \bar{a})d_{\omega,b}(x_i),
 \end{aligned}$$

and since  $\mathbb{E}[a - \bar{a}] = 0$ , the term  $\mathbb{E}[(a - \bar{a})]\bar{d}_{\omega,b}(x)$  cancels out.

## In-processing method (cont.)

$$\min_{\omega, b} \mathcal{L}(\omega, b)$$

s.t.

- Dataset  $\{(x_i, a_i, y_i)\}_{i=1}^n$
- $d_{\omega, b}(x)$  : signed distance from  $x$  to  $\omega^\top x + b = 0$ .

$$\frac{1}{n} \sum_{i=1}^n (a_i - \bar{a}) d_{\omega, b}(x_i) \leq c \quad \forall i = 1, \dots, n$$

$$Cov(a, d_{\omega, b}(x)) = \frac{1}{n} \sum_{i=1}^n (a_i - \bar{a}) d_{\omega, b}(x_i)$$

$$\frac{1}{n} \sum_{i=1}^n (a_i - \bar{a}) d_{\omega, b}(x_i) \geq -c \quad \forall i = 1, \dots, n$$

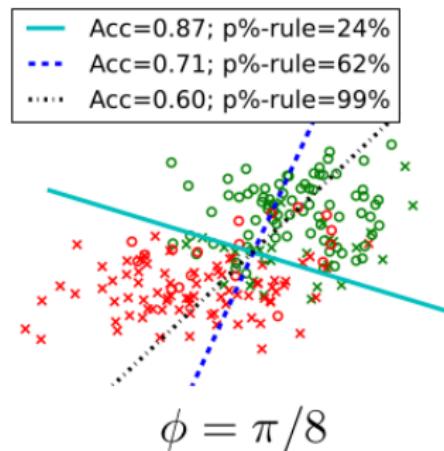
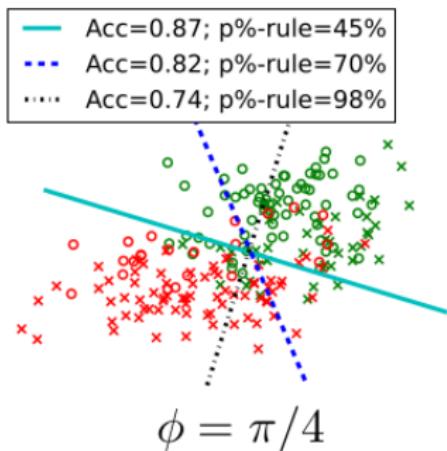
$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$

M. B. Zafar, I. Valera, M.-G. Rodriguez, and K.-P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification.

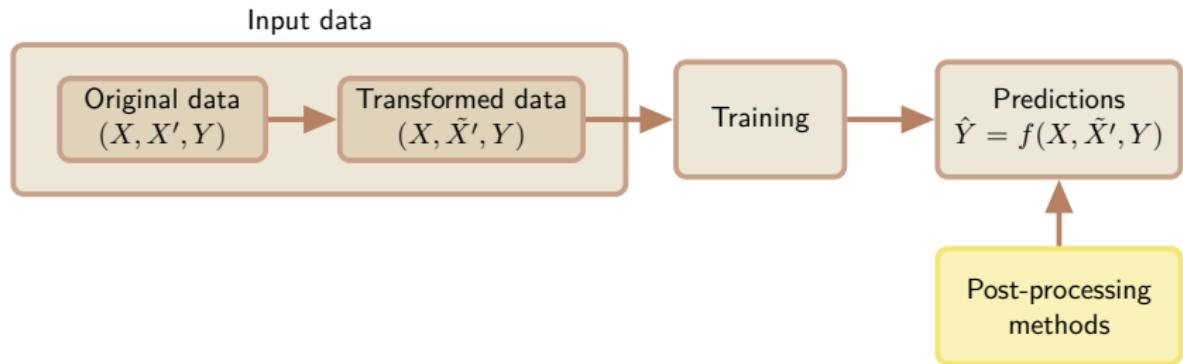
*Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:962-970, 2017.

## In-processing method (cont.)



- The solid lines show the decision boundaries without fairness constraints.
- The dashed lines show the decision boundaries trained to maximize accuracy under fairness constraints.
- Circles represent the sensitive feature and each figure corresponds to a dataset, with different correlation value between sensitive attribute values and class labels.

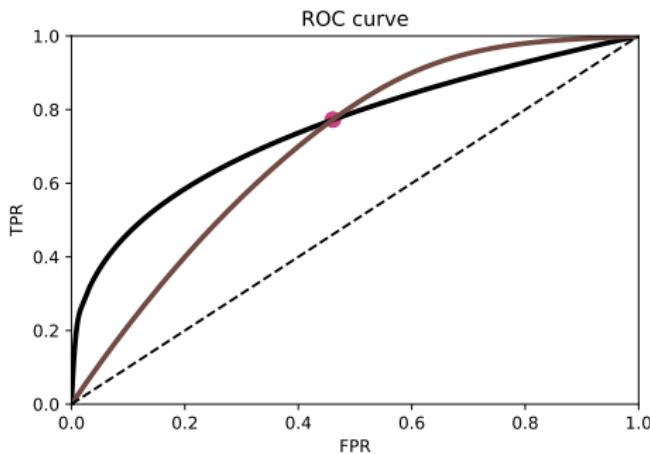
# Post-processing methods



M. Hardt, E. Price, and N Srebro. Equality of opportunity in supervised learning.  
*Advances in Neural Information Processing Systems*, 29, 2016.

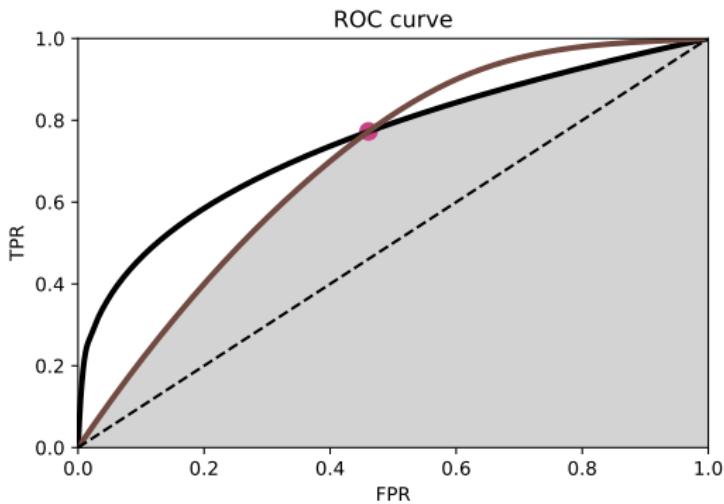
## Post-processing method

- We can achieve separation by post-processing a given score function without the need for retraining: using the ROC curve.
- A binary classifier that satisfies separation must achieve the same true positive rates and the same false positive rates in all groups.



In black: ROC curve for non-protected group.  
In brown: ROC curve for protected group.

## Post-processing method (cont.)



- This condition corresponds to taking the intersection of all group-level ROC curves.
- Within this constraint region, we can then choose the classifier that minimizes the given cost.

## Post-processing method (cont.)

- A score function obeys equalized odds if and only if the ROC curves for the protected and non-protected groups coincide for all decision threshold  $r$ .  
That is:

$$P(r(X, A = 1) > r | Y = y, A = 1) = P(r(X, A = 0) > r | Y = y, A = 0)$$

- Let us denote the two ROC curves for  $A = 1$  and  $A = 0$  as:

$$f_1(r) = (TPR_1(r), FPR_1(r))$$

$$f_0(r) = (TPR_0(r), FPR_0(r))$$

- The intersection between the two curves:

$$f_1(r_1) = f_0(r_2) \text{ for some } r_1, r_2.$$

- Then if we choose different thresholds for the protected and non-protected groups, we can achieve equalized odds:

$$P(r(X, A = 1) > r_1 | Y = y, A = 1) = P(r(X, A = 0) > r_2 | Y = y, A = 0)$$

## Legal and policy perspectives

*Within the scope of application of the Treaty establishing the European Community and of the Treaty on European Union, and without prejudice to the special provisions of those treaties, any discrimination on grounds of nationality shall be prohibited.*

### Legally recognized protected features: Europe

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.

## Legal aspects: Europe

This right is enshrined in article 21 of the Charter of Fundamental Rights.  
There are different directives:

- ① Against discrimination on grounds of race and ethnic origin.
- ② Against discrimination at work on grounds of religion or belief, disability, age or sexual orientation.
- ③ Towards equal treatment for men and women in matters of employment and occupation.
- ④ Towards equal treatment for men and women in the access to and supply of goods and services.
- ⑤ Against discrimination based on age, disability, sexual orientation and religion or belief beyond the workplace.

[https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/know-your-rights/equality/non-discrimination\\_en](https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/know-your-rights/equality/non-discrimination_en)

## Legal aspects: US

### Legally recognized protected features

Race, color, sex, religion, national origin, citizenship, age, pregnancy, disability status, genetic information, veteran status, familial status.

### Regulated domains

- Credit (Equal Credit Opportunity Act)
- Education (Civil Rights Act of 1984)
- Employment(Civil Rights Act of 1984)
- Housing (Fair Housing Act)

# Example I: The US Equal Pay Act

## The US Equal Pay Act

- Requires that men and women in the same workplace be given equal pay for equal work.
- The jobs need not be identical, but they must be substantially equal.
- This law covers all forms of pay including salary, overtime pay, bonuses, stock options, profit sharing and bonus plans, life insurance, etc.
- This act aimed at abolishing wage disparity based on sex.
- According to the US Bureau of Labor Statistics, women's salaries compared to men's have risen dramatically since the enactment of this equal pay act, from 62% in 1970 to 80% in 2004.
- This real-world case illustrates a scenario where our historical data are discriminatory due to a biased data generation process, but where classifiers learned on the data are forced to be discrimination-free by law.

## Example II: The Australian Sex Discrimination Act 1984

### The Australian Sex Discrimination Act 1984

- It prohibits discrimination in work, education, services, accommodation, land, pregnancy or potential pregnancy, and family responsibilities.
- This act defines sexual harassment and other discriminatory practices on different grounds and declares them unlawful.
- This law also prohibits indirect and unintentional discrimination.
- It is the responsibility of the accused party to prove that his/her intention was not to discriminate the aggrieved party: the burden of proving that an act does not constitute discrimination lies on the person who did the act.
- Notice that under this law it is insufficient to remove the sex attribute from a dataset before learning; also indirect discrimination on the basis of a "characteristic that appertains generally to persons of the sex of the aggrieved person" is disallowed.

## Example III: The US Equal Credit Opportunity Act 1974

### The US Equal Credit Opportunity Act 1974

Declares unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction, on the basis of race, color, religion, national origin, sex or marital status, or age.

## Example IV: European Council Directive 2004

### European Council Directive 2004

- Even though there is clear historical evidence showing higher accident rates for male drivers in traffic, insurance companies are no longer allowed to discriminate based on gender in many countries.
- The European Court of Justice decided (in 2011) that it will no longer be legal under EU law to charge women less for insurance than men.
- The verdict means that different priced premiums for men and women drivers will now be considered to be in breach of the EU's anti-discrimination rules.

## Final remarks

- All of the anti-discriminatory laws prohibit discriminatory practices in future (or present).
- If we are interested in applying ML algorithms, and our available historical data is biased, it is simply **illegal** to use traditional algorithms without taking the fairness aspect into account.
- Because of the above mentioned laws and due to ethical concerns, restricting ourselves to the single use of traditional ML techniques is unacceptable.

## Bibliography

- M. Hardt, E. Price, and N Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.
- M. B. Zafar, I. Valera, M.-G. Rodriguez, and K.-P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:962-970, 2017.
- F. Kamiran and T. Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1-33.