Security and Ethical Aspects of Data

Amaya Nogales Gómez
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 ressources 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 ressources will be posted on Slack.
3. Assessment
   - Final Exam (60%).
   - Labs (40%).

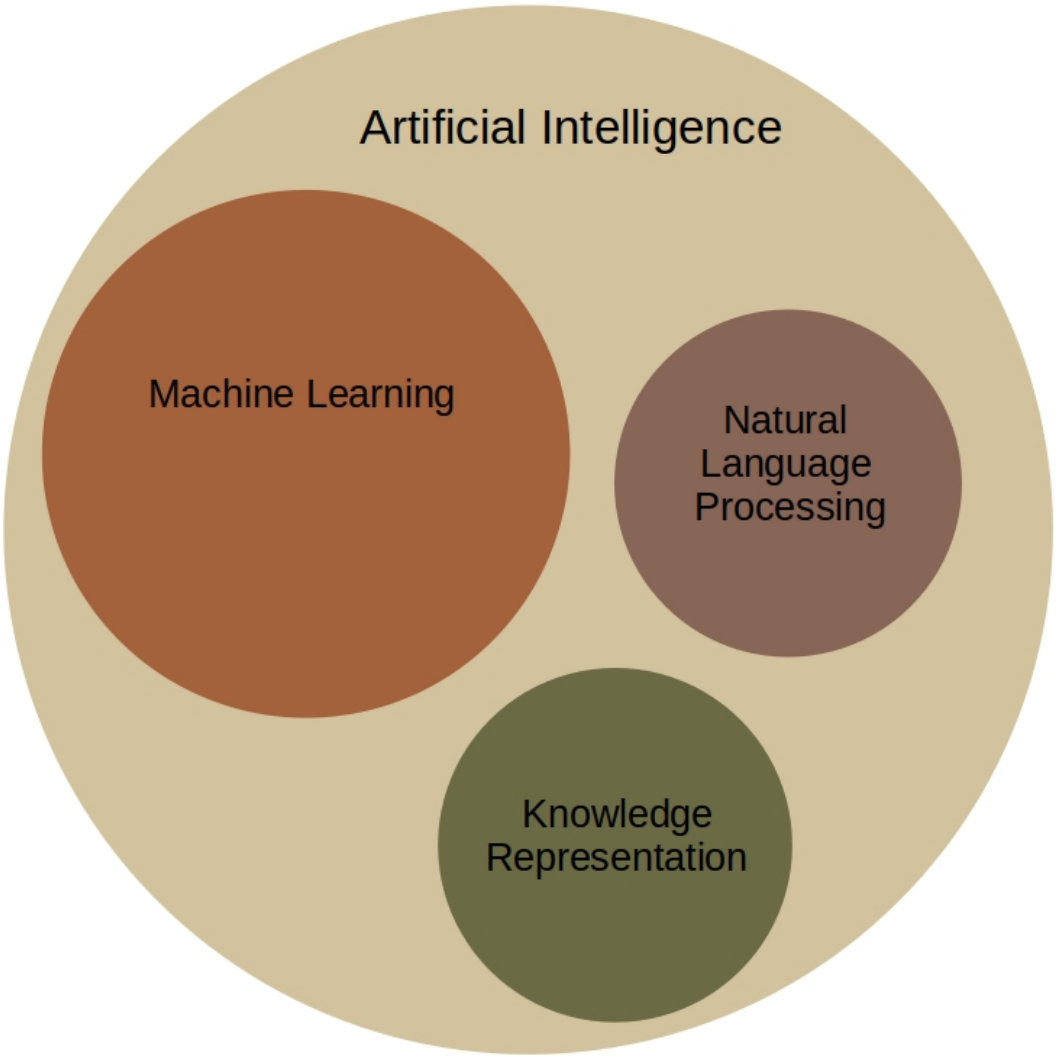
Course Overview

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 ressources 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.

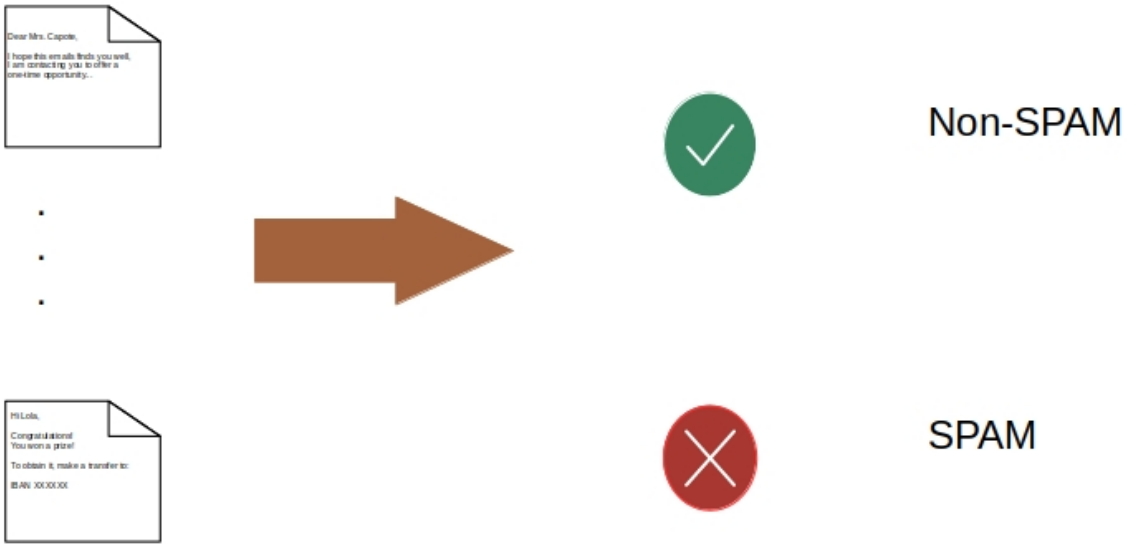
Tentative Content

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

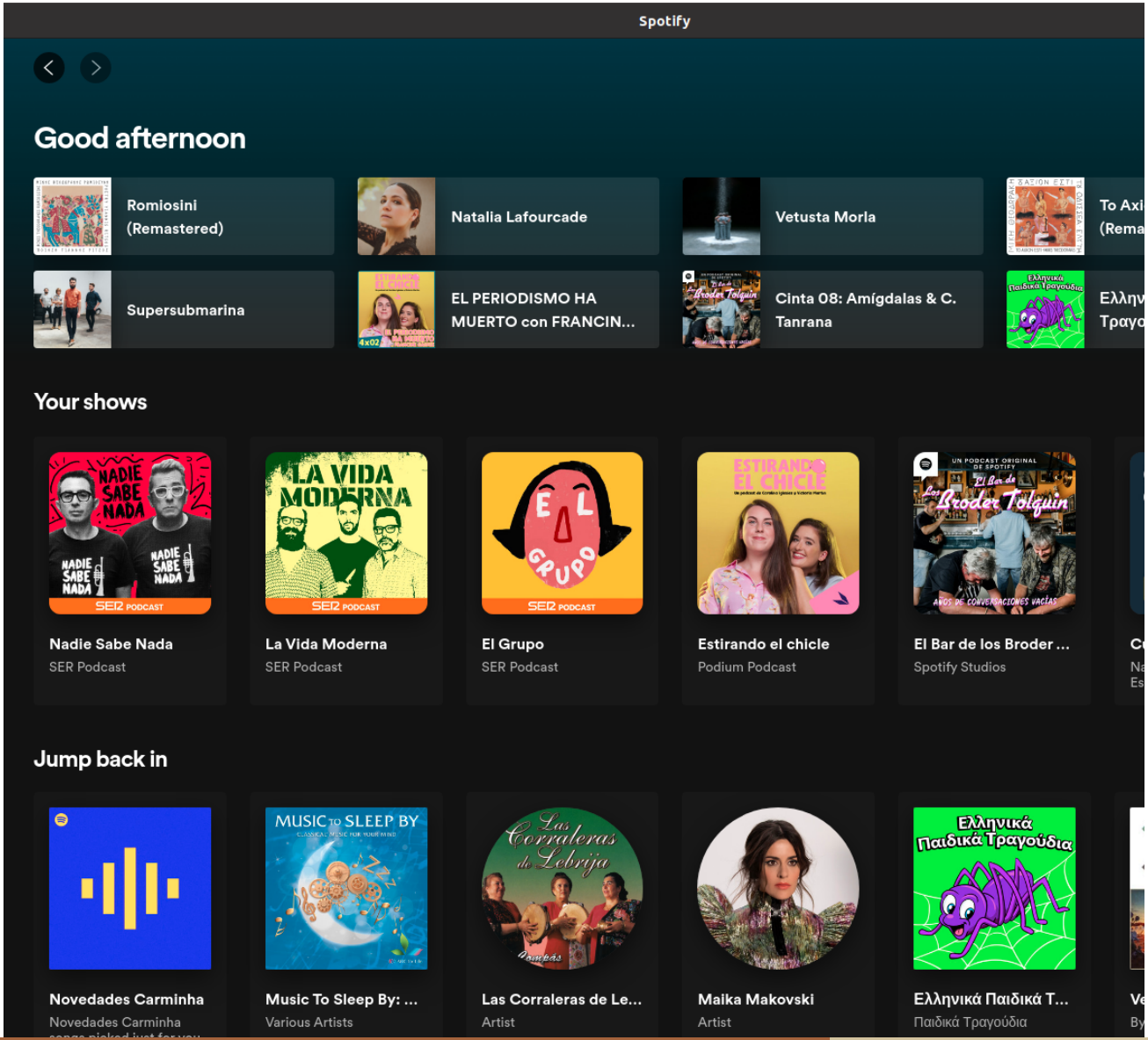
What is Machine Learning?



Example: spam filtering

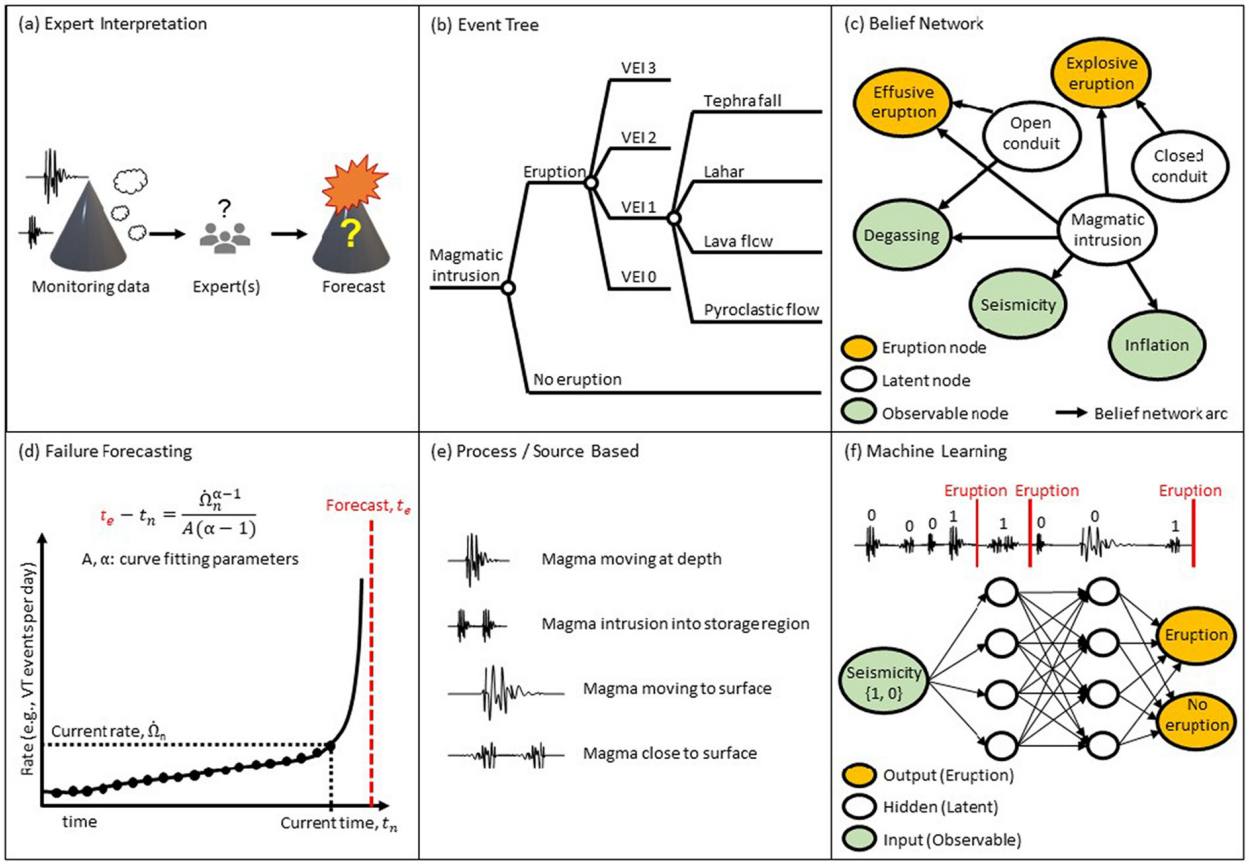


Example: recommender systems



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 17 - 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.

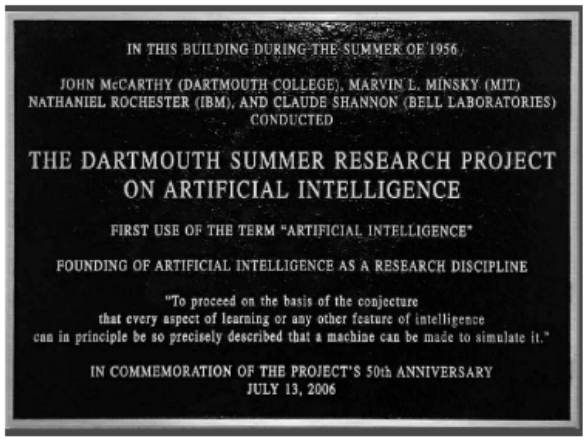
The following are some aspects of the artificial intelligence problem:

1) Automatic Computers

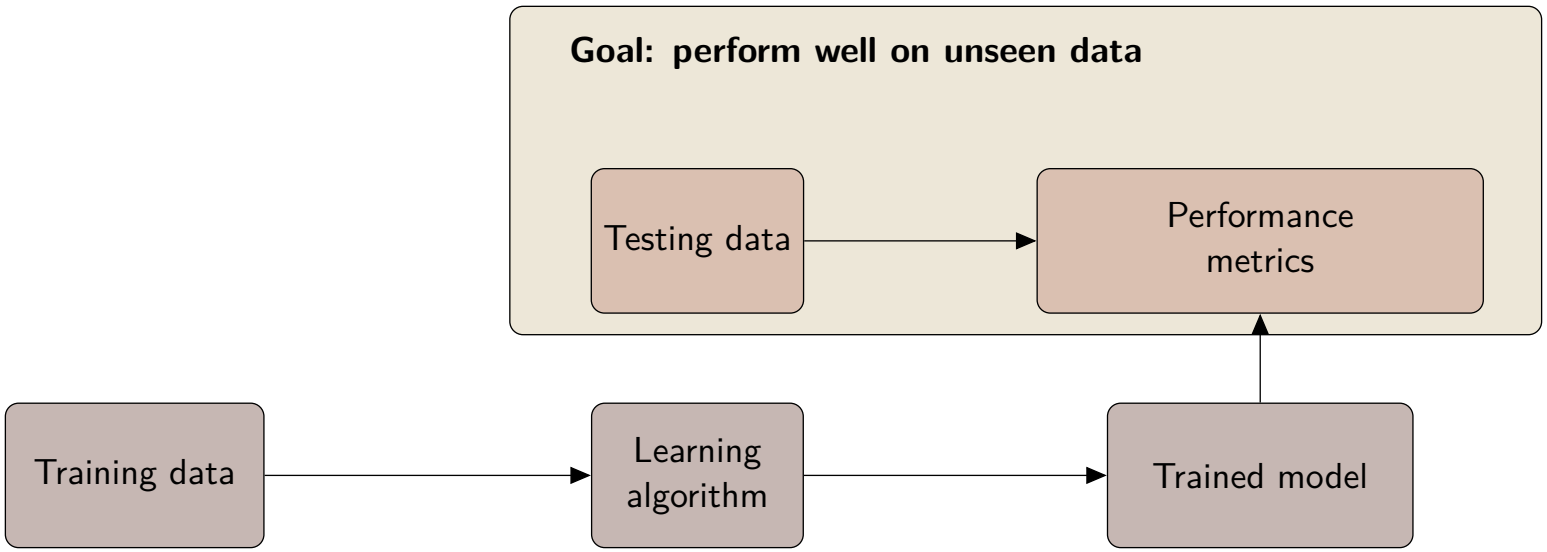
If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability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

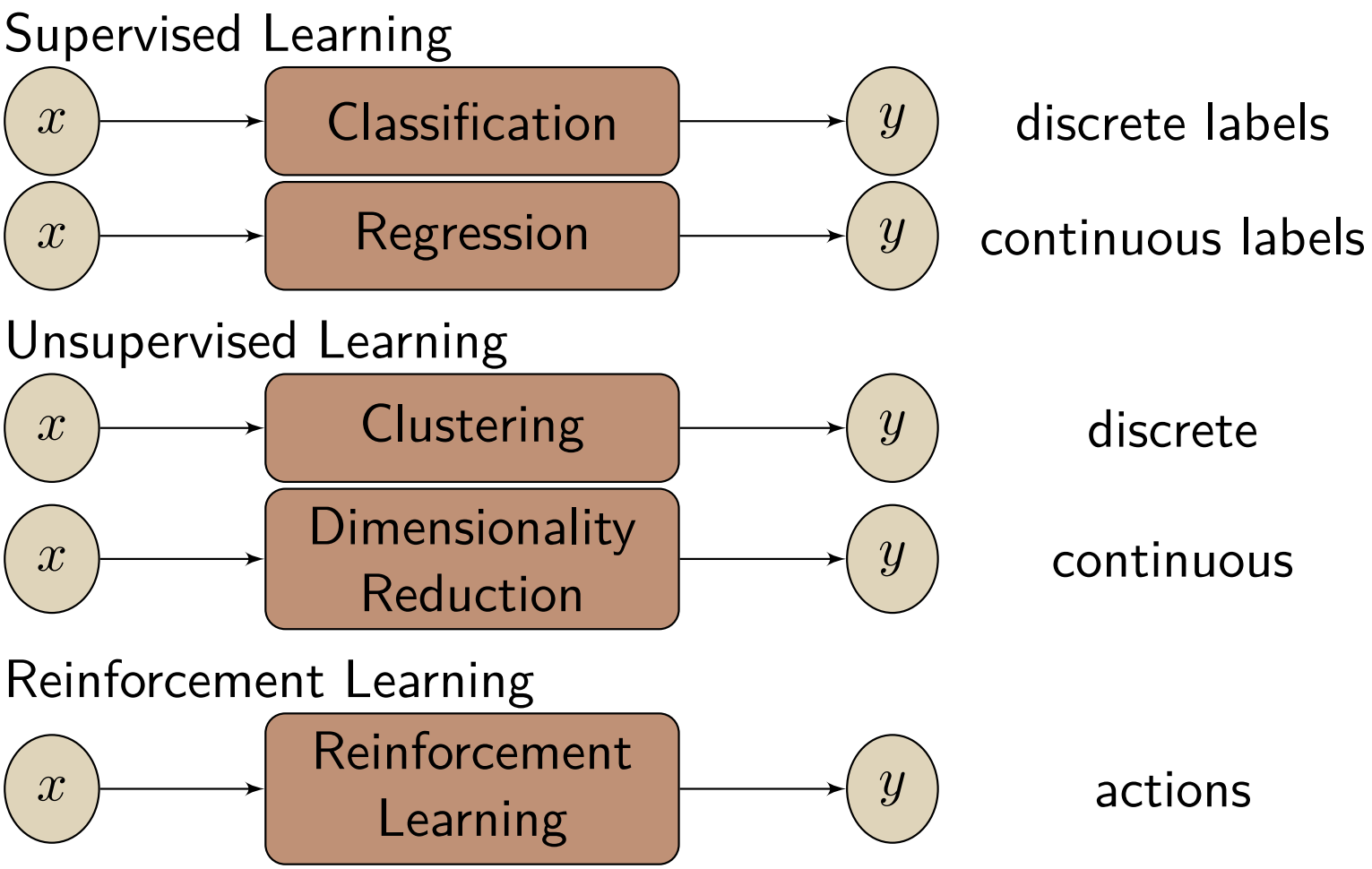
- What kind of data to use?
- How much data is enough?
- How to represent it?
- Which algorithm should be used?
- How to choose the best model?
- Performance guarantees
- Explainability
- Interpretability

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

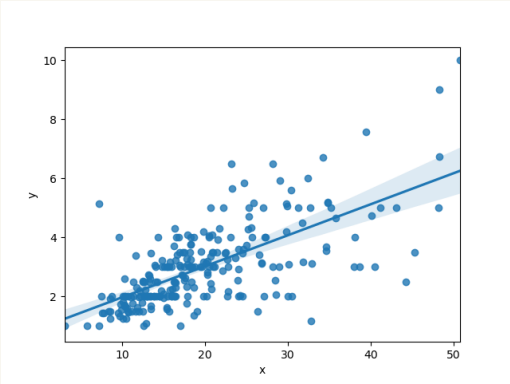
Supervised learning



Supervised Learning: Regression and classification

Regression

Output: continuous function.

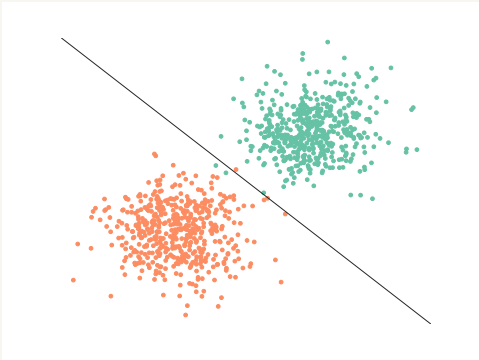


Examples:

- Forecasting
- Size of animal
- Stocks

Classification

Output: separation rule.

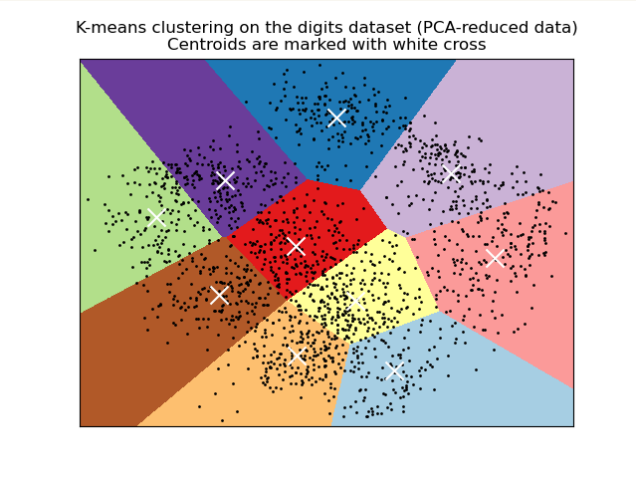


Examples:

- Pay back a loan
- University acceptance
- Image classification

Unsupervised learning

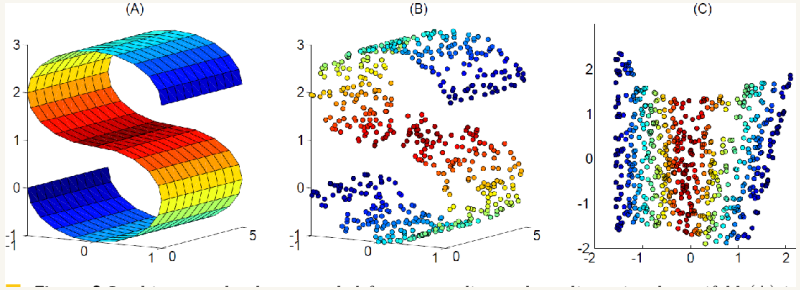
Clustering



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

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

Supervised Classification: Support Vector Machines

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

Supervised Classification: Support Vector Machines

- Ω: the population.
- Population is partitioned into two classes, $\{-1, +1\}$.
- For each object in Ω, we have
  - $x = (x^1, \ldots, x^d) \in X \subset \mathbb{R}^d$: 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.

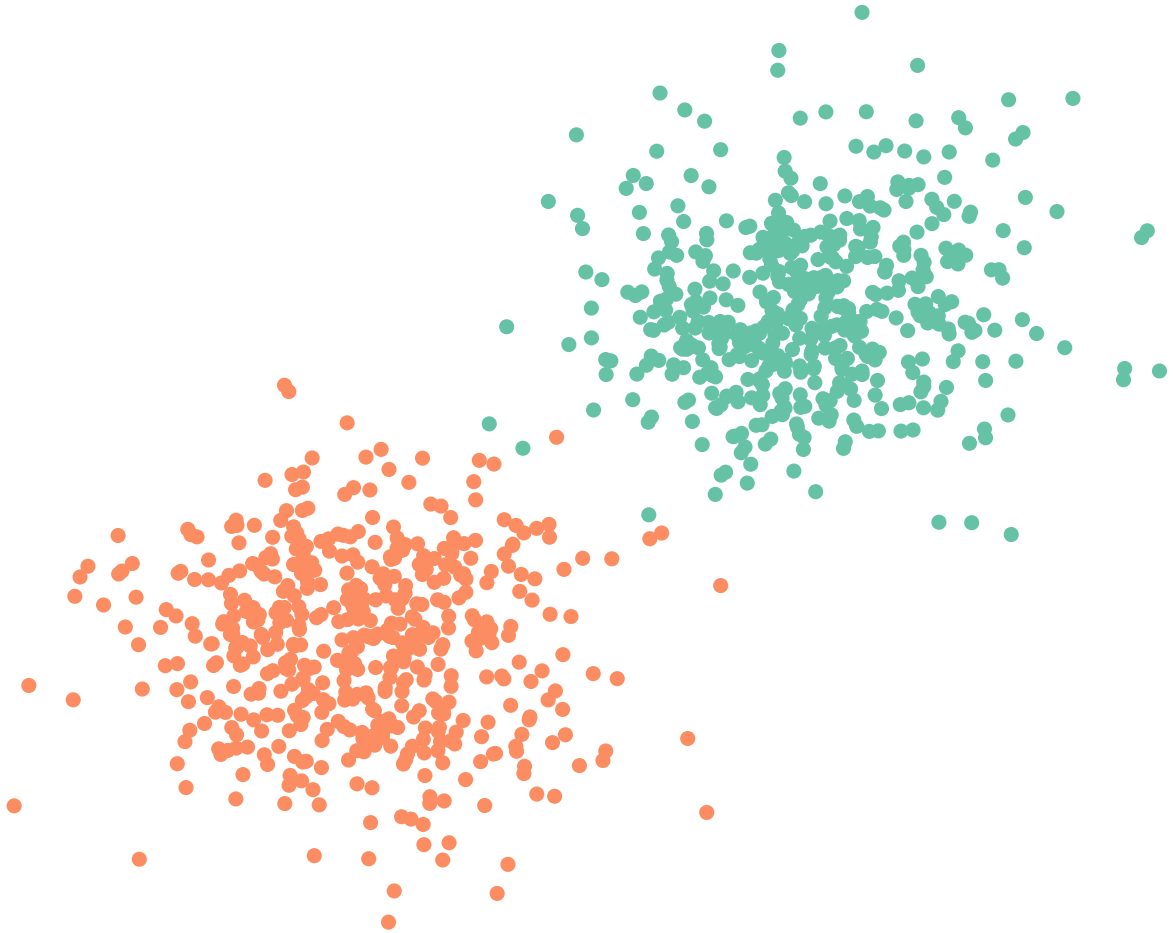
Supervised Classification: Support Vector Machines

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

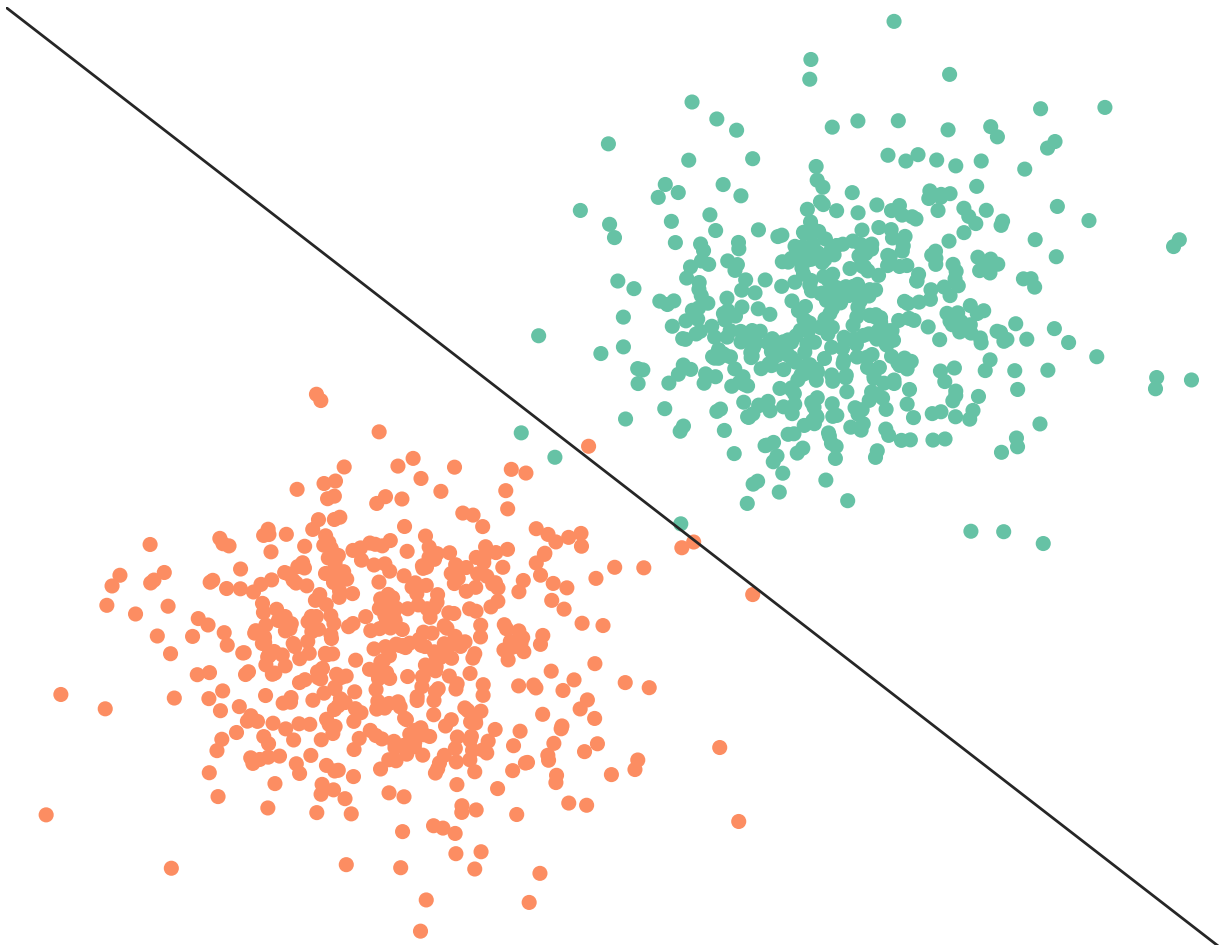
- The goal is to find a hyperplane $\omega^\top x + b = 0$ that aims at separating, if possible, the two classes.
- Future objects will be classified as

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

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.

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

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

$$\max_{\omega, b} \frac{1}{\|\omega\|^\circ}$$

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \qquad \forall i = 1, \ldots, n$$
$$\omega \in \mathbb{R}^d \setminus 0$$
$$b \in \mathbb{R},$$

$$\min_{\omega, b} \|\omega\|^\circ$$

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 \qquad \forall i = 1, \ldots, n$$
$$\omega \in \mathbb{R}^d$$
$$b \in \mathbb{R},$$

Hard-margin SVM: final formulation

The objective function can be replaced by $\Phi(\|\omega\|^\circ)$ for any Φ, 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.

$$y_i(\omega^\top x_i + b) \geq 1 \qquad \forall i = 1, \ldots, n$$
$$\omega \in \mathbb{R}^d$$
$$b \in \mathbb{R}.$$

But...

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

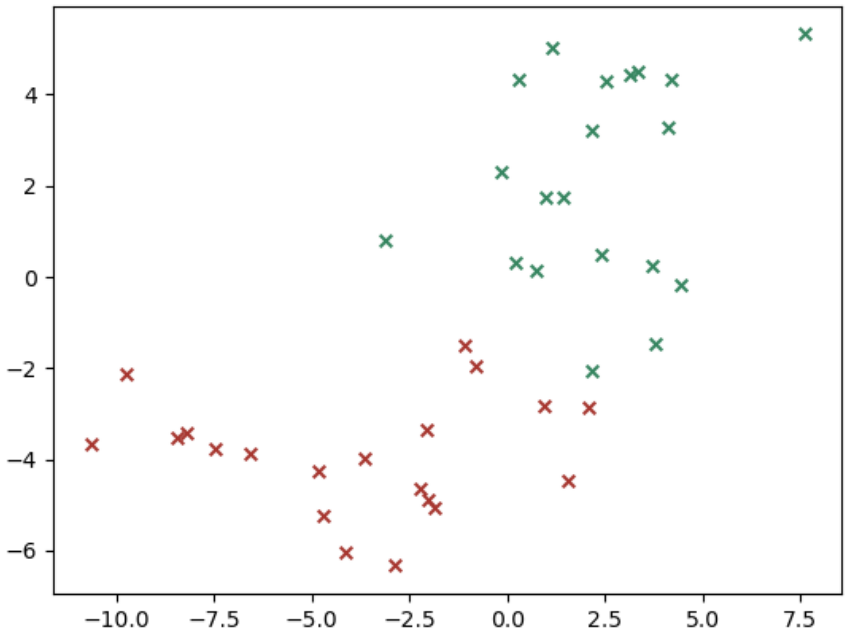
s.t.

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

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

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

Linearly separable data



But...

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

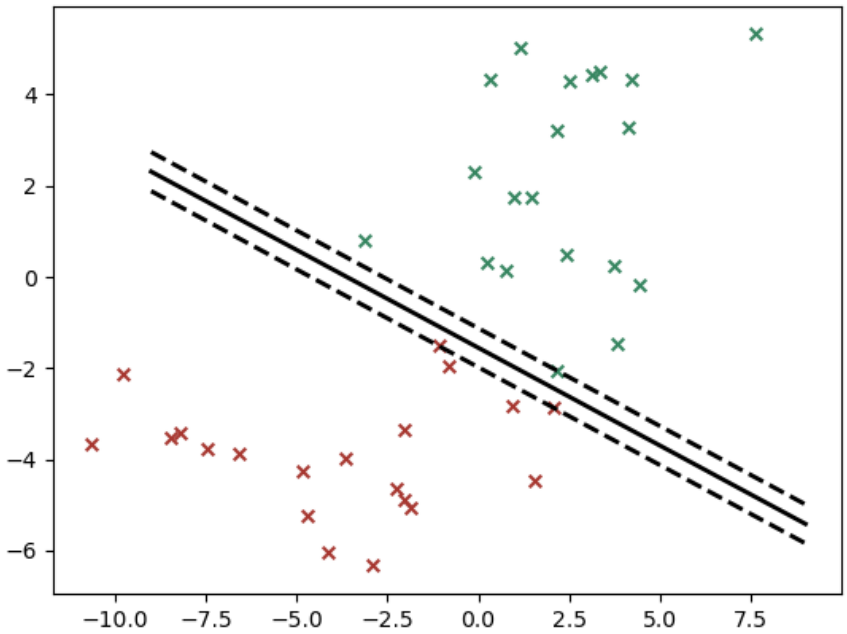
s.t.

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

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

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

Linearly separable data



But...

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

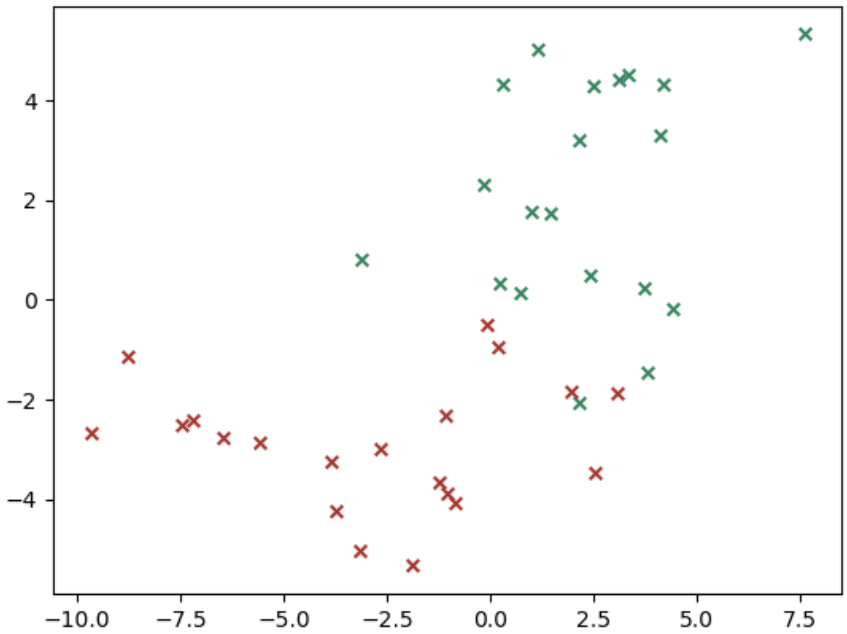
s.t.

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

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

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

Non-linearly separable data



But...

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

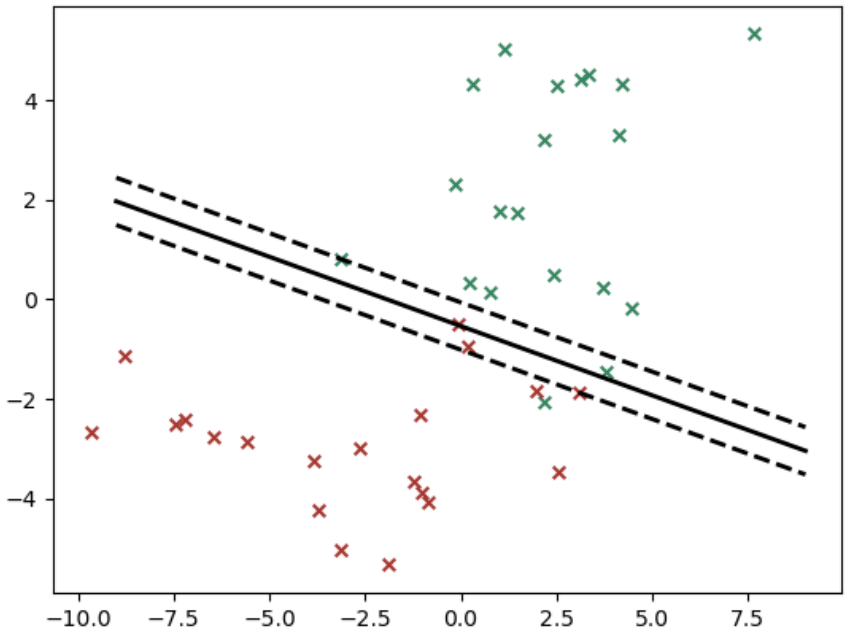
s.t.

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

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

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

Non-linearly separable data



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

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

  are perturbed.
- How? By introducing auxiliary variables ξ_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.

$$\begin{aligned} y_i(\omega^\top x_i + b) &\geq 1 - \xi_i \qquad \forall i = 1, \ldots, n \\ \omega &\in \mathbb{R}^d \\ b &\in \mathbb{R} \\ \xi_i &\geq 0 \qquad \forall i = 1, \ldots, n, \end{aligned}$$

- $\xi = (\xi_i) \in \mathbb{R}^n$ 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.

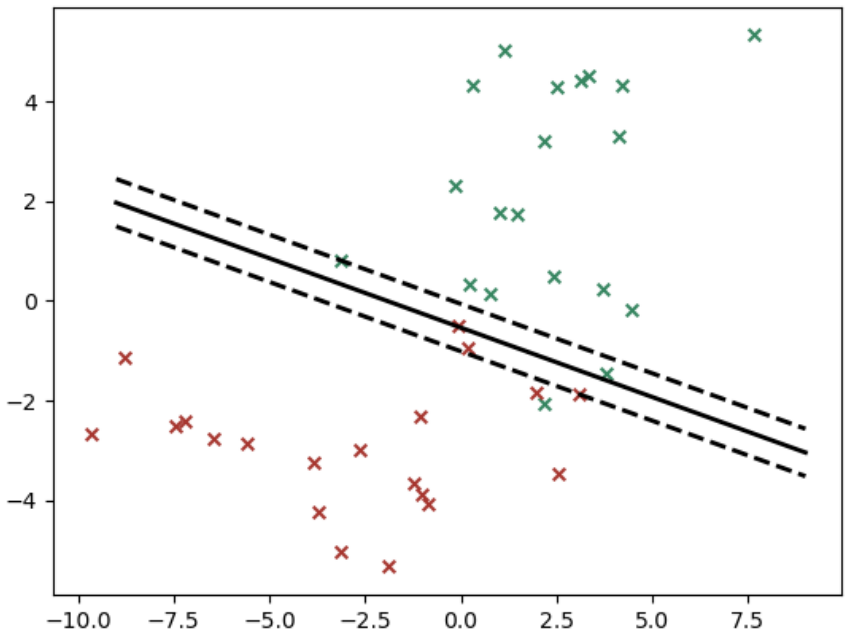
The SVM formulation

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

s.t.

$$y_i(\omega^\top x_i + b) \geq 1 - \xi_i \qquad \forall i = 1, \ldots, n$$
$$\omega \in \mathbb{R}^d$$
$$b \in \mathbb{R}$$
$$\xi_i \geq 0 \qquad \forall i = 1, \ldots, n.$$

Non-linearly separable data



- 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

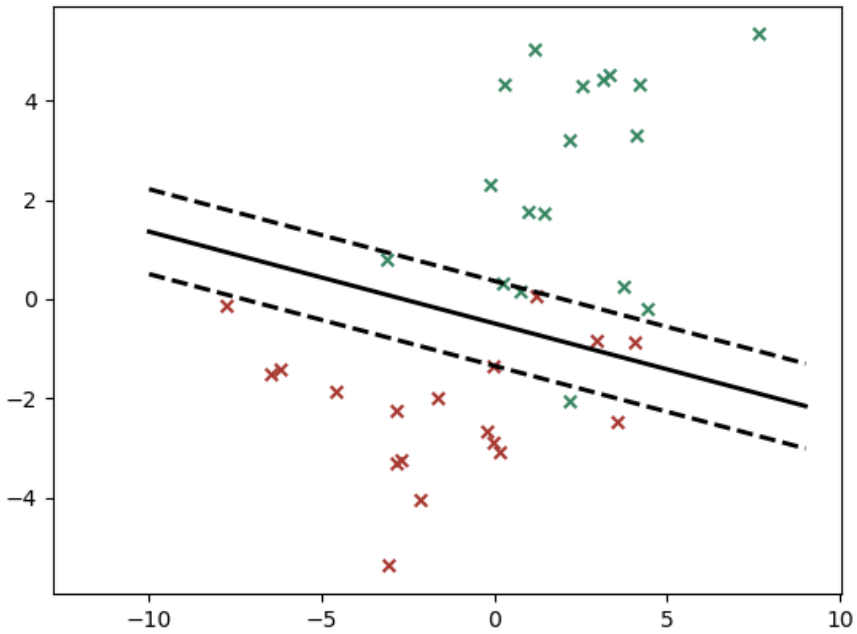
Classification Accuracy

Given an object i, it is classified in the positive or the negative class according to the value of the score function, $sign(\omega^\top x_i + b)$, 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.

$$Accuracy = \frac{correct\,predictions}{total\,predictions} =$$

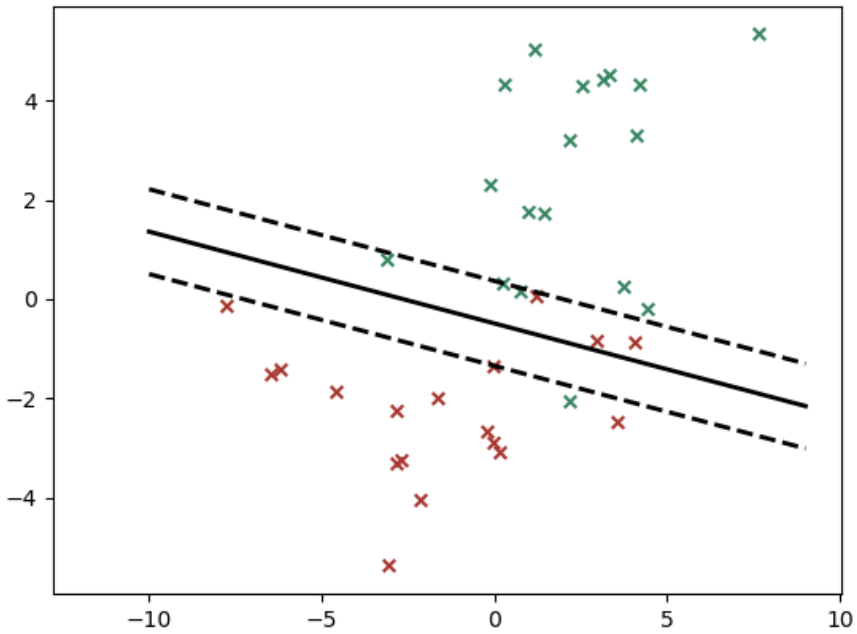
$$= P(\omega^\top x_i + b \geq 0 \wedge y_i = +1) + P(\omega^\top x_i + b < 0 \wedge y_i = -1)$$

Example: classification accuracy



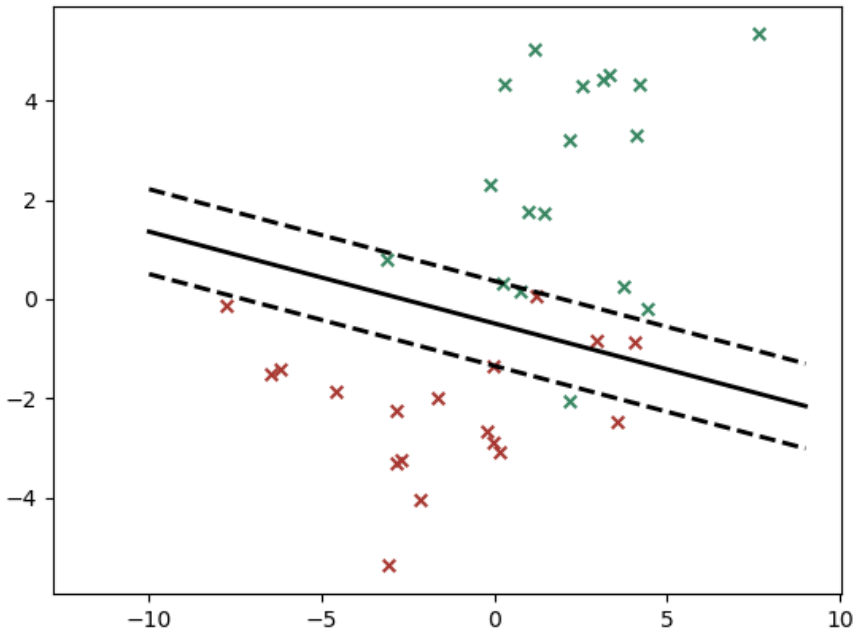
- $|\Omega| = 40$

Example: classification accuracy



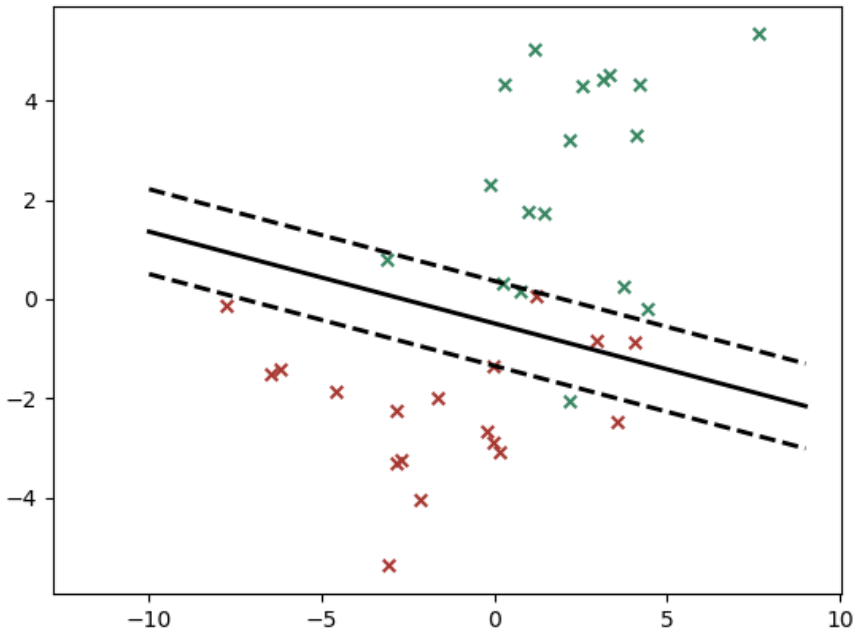
- $|\Omega| = 40$
- $\#\{i, y_i = +1\} = 20$

Example: classification accuracy

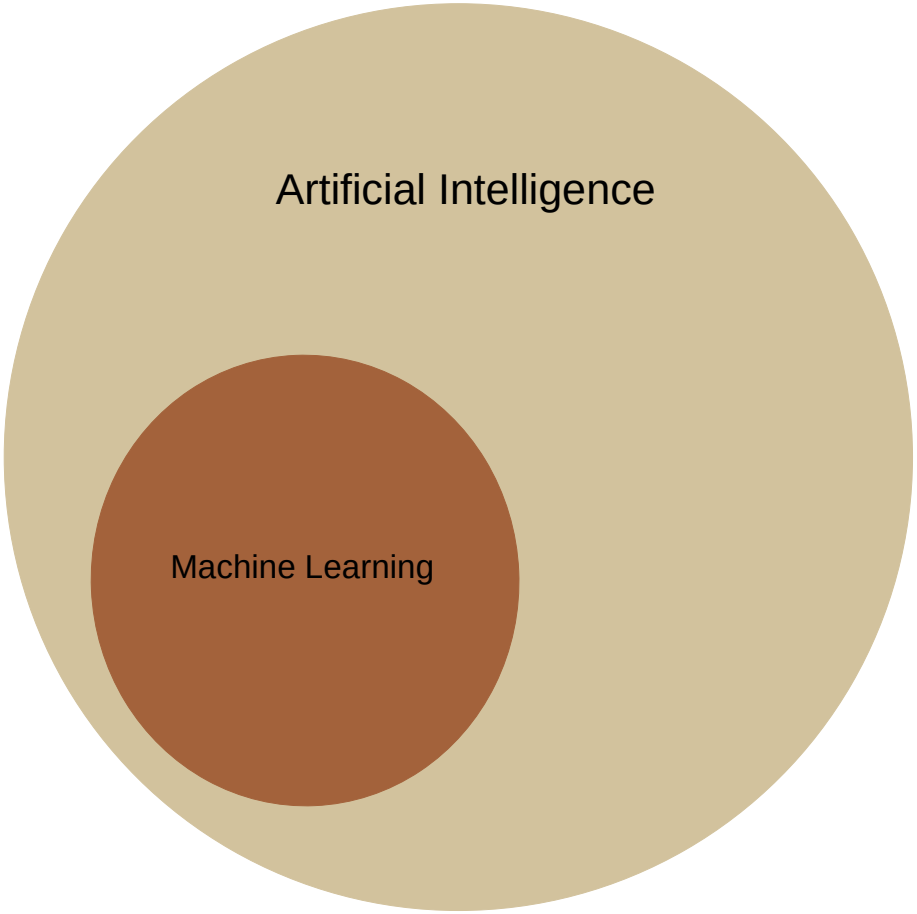


- $|\Omega| = 40$
- $\#\{i, y_i = +1\} = 20$
- $\#\{i, y_i = -1\} = 20$

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.



Artificial Intelligence

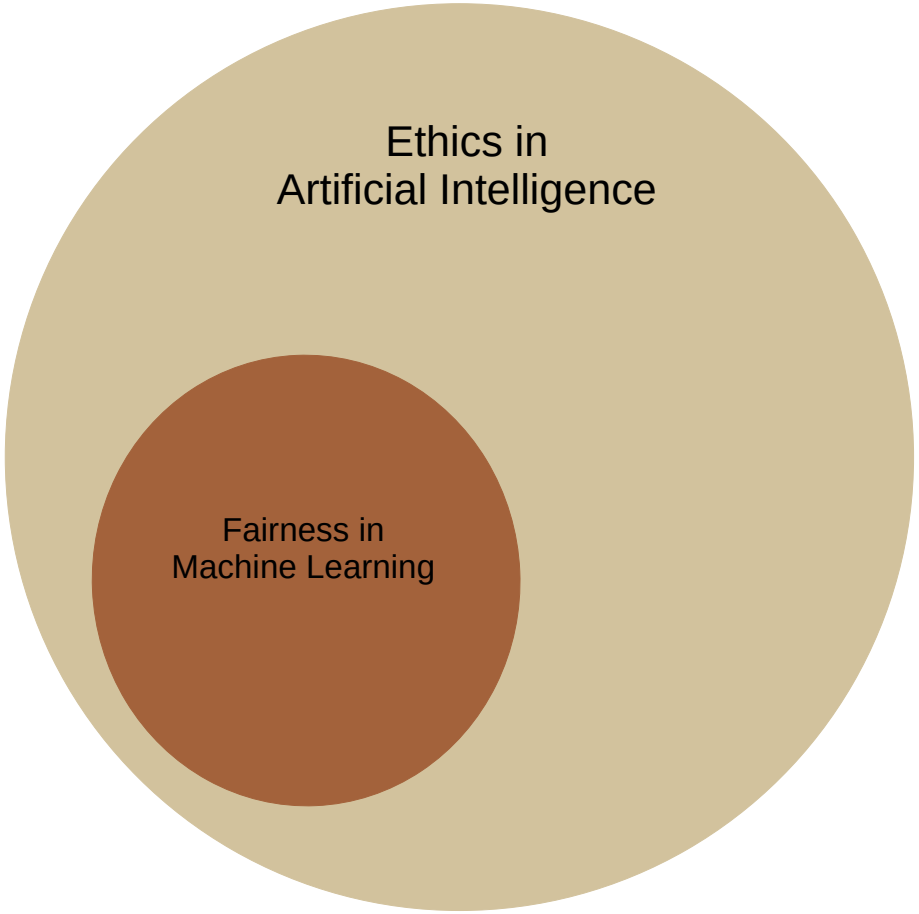
↓

Machine Learning

ethics

↓

fairness



Artificial Intelligence

↓

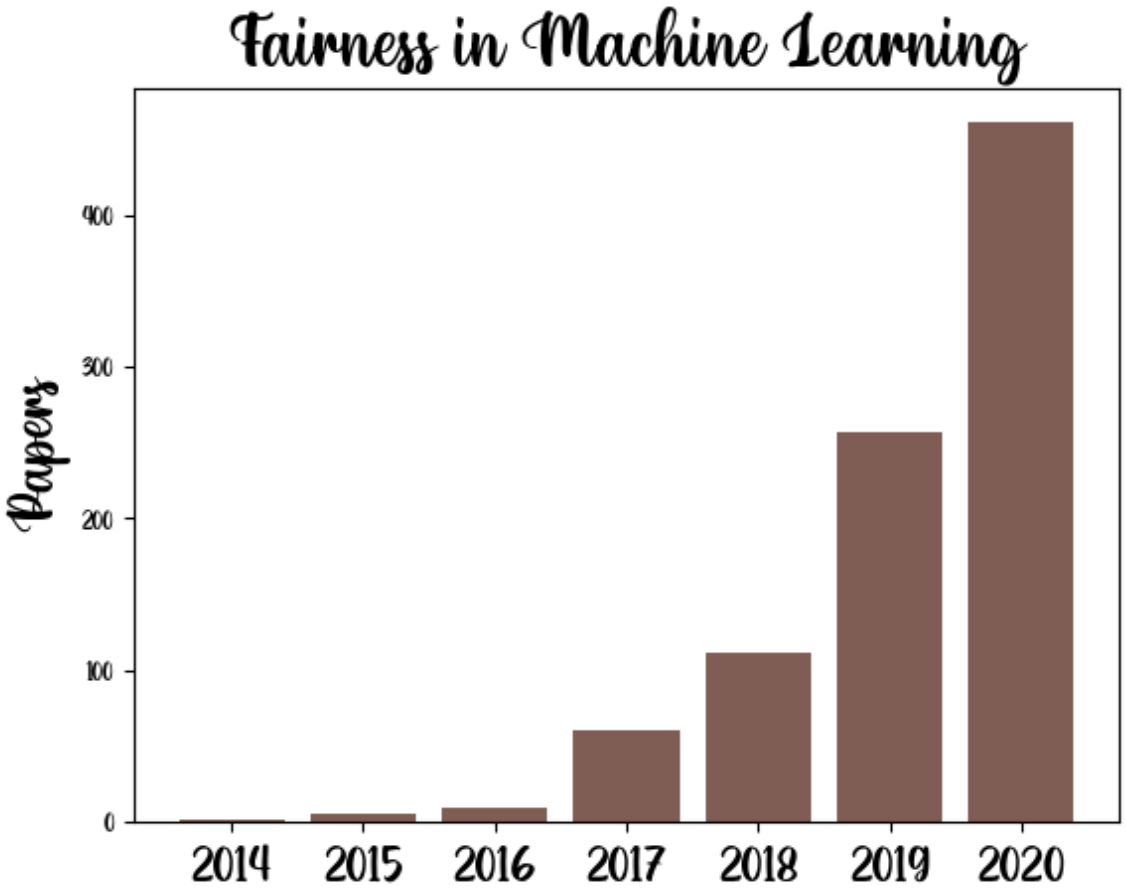
Machine Learning

ethics

↓

fairness

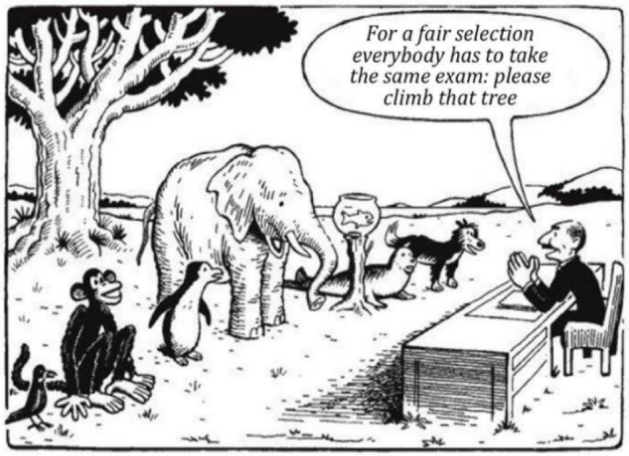
History of fairness in ML



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

$$(x, s, y) \begin{cases} x, \text{non-sensitive features} \\ s, \text{sensitive feature} \\ y \in \{-1, +1\}, \text{class membership} \end{cases}$$

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

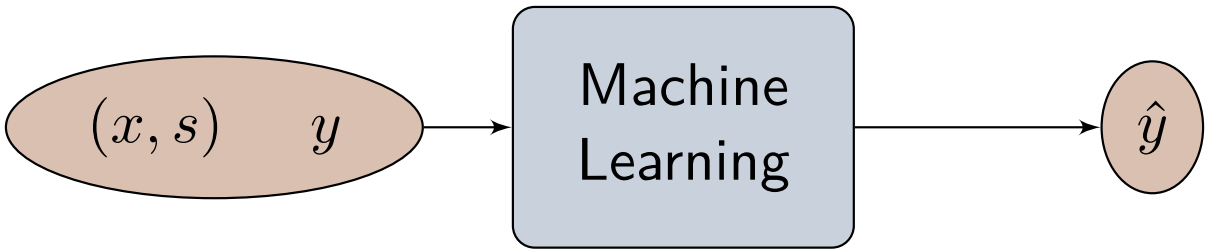
A fair machine learning algorithm implies that its **outcome** should not have a **disproportionately large adverse impact** on a *protected* class of features.

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		
	Gender & status	income	credit history	decision
Applicant 1	male married	1.5k	5	$\surd$
Applicant 2	female single	2.5k	3	$\surd$

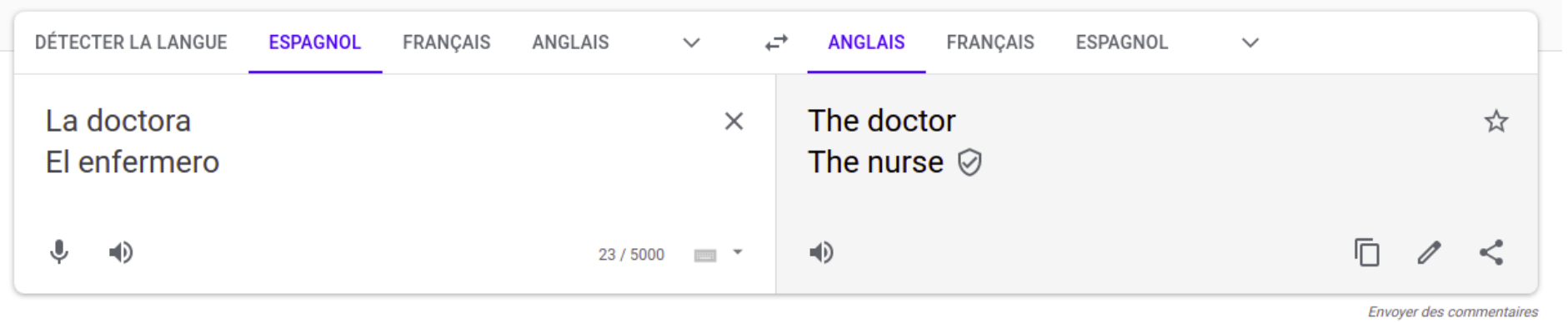
Why is ML unfair?

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

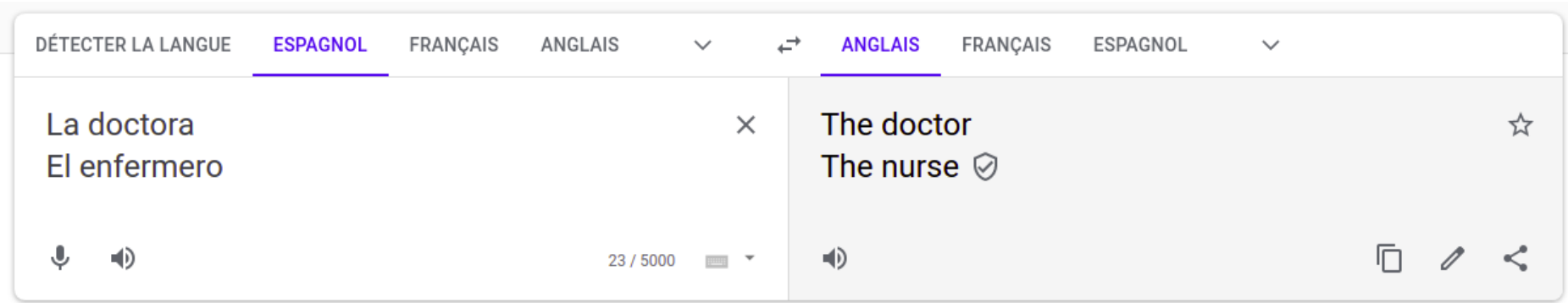
Why do we care about fairness?

- 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



Google translator



COMPAS algorithm: recividism prediction

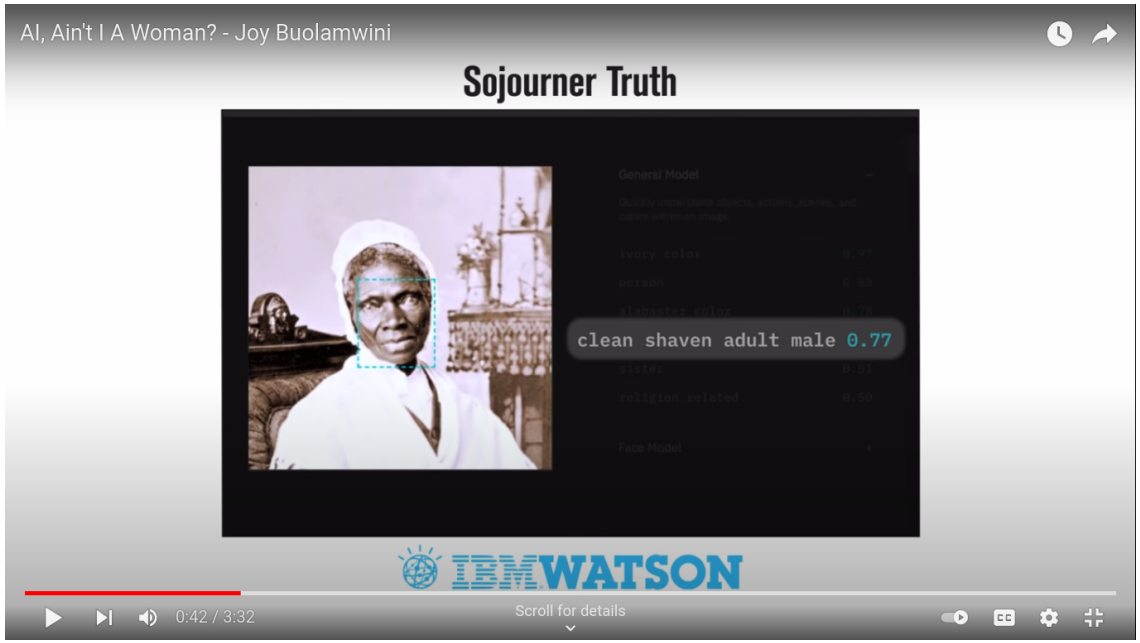
Prediction Fails Differently for Black Defendants

	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*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

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

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

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-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing.**

ACM Reference Format:
Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922

1 INTRODUCTION

One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data. Since 2018 alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

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 texts from the Internet 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 be 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 §3, LMs are not performing natural 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 on leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown

*Joint first authors

This work is licensed under a Creative Commons Attribution International 4.0 License.
FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03…
https://doi.org/10.1145/3442188.3445922

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`
  - Mirror, mirror. `https://dataresponsibly.github.io/comics/`

Data, Responsibly (Vol. 1) Mirror, Mirror

Falaah Arif Khan and Julia Stoyanovich

Comic book

Transcript