# Fairness in Machine Learning

Felipe Garrido-Lucero

Inria, FairPlay joint team

École de Printemps Lamsade 2024

Presentation based on

- Tutorial *Fairness in Machine Learning* by Patrick Loiseau
- Tutorial *Fairness-aware Machine Learning* [link]

# Outline

# Outline

Automated decisions are everywhere

- Bank loans
- Insurance
- Justice
- Education
- Medicine
- Pricing
- Recommendation systems: music, movies, job offers, etc.
- Etc...

Interest: Making decisions optimally (under which criterion?)

J. Correa et al. (2019) School Choice in Chile. In Proceedings of the 2019 ACM Conference on Economics and Computation (EC'19)

- 2015 Change in the school inclusion law
- Elimination of profit regarding co-payment in subsidized private schools
- Prohibition of public schools choosing students based on social, religious, economic, or academic criteria
- Main reasons for segregation
- Centralized application system to public and subsidized schools
- Advantages at an informational level
- Eliminates the need for traveling to school
- Fair and transparent system

# Education (2)

J. Correa et al. (2019) School Choice in Chile. In Proceedings of the 2019 ACM Conference on Economics and Computation (EC'19)

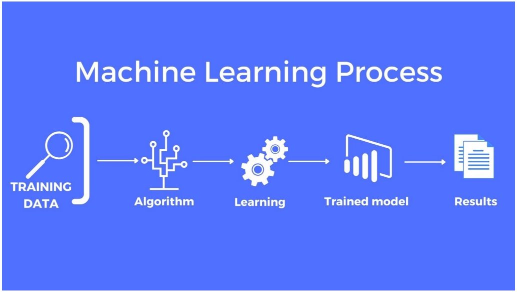
- Nationally with students from pre-kindergarten to last grade
- Siblings assigned to the same school
- Students assigned to schools where parents work
- Students can try to change schools
- If the change is not possible, the student must have their old position secured

| | 2016 | 2017 | 2018 |
|---|---|---|---|
| Regions | 1 | 5 | 15 |
| Schools | 63 | 2,174 | 6,421 |
| Students | 3,436 | 76,821 | 274,990 |
| % assigned 1st preference | 58.0 | 56.2 | 59.2 |
| % assigned any preference | 86.4 | 83.0 | 82.5 |
| % unassigned | 9.0 | 8.7 | 8.9 |

# Outline

"In machine learning a computer observes data, builds a model based on that data and uses that model as [...] a piece of software that can solve problems".

Russell and Norvig (2021). Artificial Intelligence: A Modern Approach

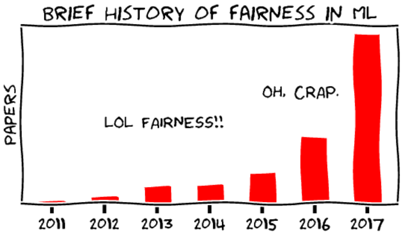



But, how can we talk about fairness if a computer makes the decisions?

# Outline

Fairness in the field of machine learning seeks to correct and prevent possible biases in automated decision-making processes, when these decisions are based on machine learning models

Furthermore, these decisions can be considered illegal if they are based on sensitive variables such as gender, ethnicity, sexual orientation, disability, among others




[Illustration by Hardt]

- It studies algorithms that reflect "systematic and unfair discrimination"
- Bank loans. We say that the algorithm has biases if
  - it recommends loans to one group of users but denies loans to another almost identical group of users based on unrelated criteria
  - this behavior can be repeated on different occasions
- These biases can be unintentional

# Discrimination in ML-based Hiring Processes

The New York Times

TheUpshot

ROBO RECRUITING

## *Can an Algorithm Hire Better Than a Human?*

117

By Claire Cain Miller

June 25, 2015

“hiring could become faster and less expensive, and [...] lead recruiters to more highly skilled people [...]. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases.”

# Discrimination in ML-based Hiring Processes

The New York Times

TheUpshot

ROBO RECRUITING

***Can an Algorithm Hire Better Than a Human?***

117

By Claire Cain Miller

June 25, 2015

"hiring could become faster and less expensive, and [...] lead recruiters to more highly skilled people [...]. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases."

The New York Times

TheUpshot

HIDDEN BIAS

***When Algorithms Discriminate***

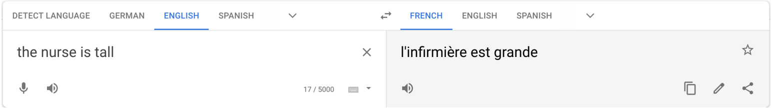
147

By Claire Cain Miller

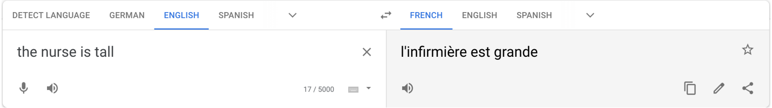
July 9, 2015

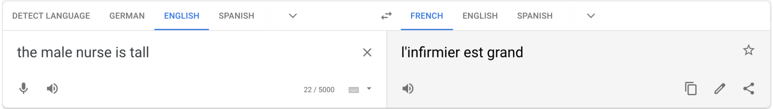
"But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result, [...] algorithms can reinforce human prejudices.

Google Translate January 15, 2021



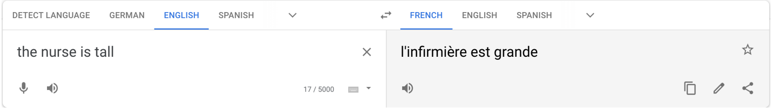
Google Translate January 15, 2021

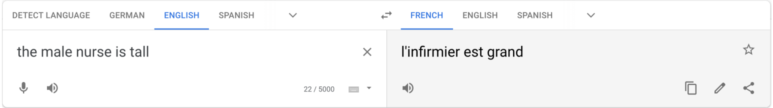


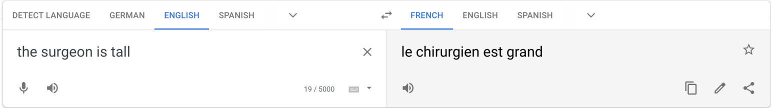



# Discrimination in Automatic Translation

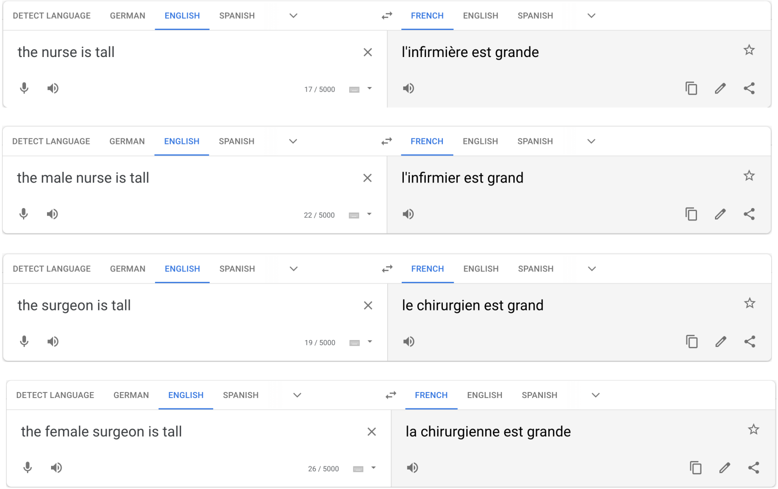
## Google Translate January 15, 2021



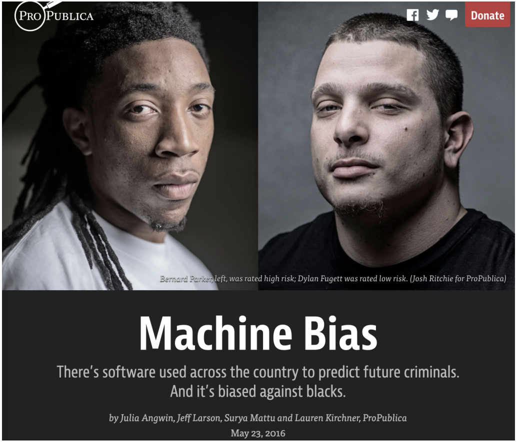



Google Translate January 15, 2021



# COMPAS software[1]




- Software to predict the likelihood of criminal recidivism
- Useful for measuring the need for rehabilitation of the person

1 [Angwin et al., Propublica 2016]

# Discrimination in Online Advertising

- Advertisements are calculated/optimized for each user
- Job opportunities, financial services, rentals, etc.
- The goal is to maximize the probability of click
- The law prohibits discrimination at every stage of the process (i.e., not just at the final decision)

DIGITAL

**Online Ads for High-Paying Jobs Are Targeting Men More Than Women**

New study uncovers gender bias

By Garett Sloane | July 7, 2015 PREMIUM

**Facebook, Amazon, and hundreds of companies post targeted job ads that screen out older workers**

Facebook users are suing them for age discrimination.

By Alexia Fernández Campbell | @AlexiaCampbell | alexia@vox.com | May 31, 2018, 8:50am EDT

**Facebook still runs discriminatory ads, new report finds**

*Over a year after it pledged to stop*

By Makena Kelly | @kellymakena | Aug 26, 2020, 4:00pm EDT

- Removing sensitive attributes is not enough
- Correlation among features
- The artificial intelligence matching algorithm discriminates[2]

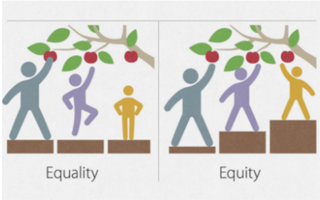
[2] [Ali et al., 2019]

## Fairness in machine learning is a complex issue

- There are various sources of discrimination
  - Biased observations
  - Feedback loop
  - Low dimensionality of our data
  - High variance
  - etc...
- Highly interdisciplinary
- Different fairness doctrines (disparate impact vs treatment)
- Definitions are generally domain- or task-specific (laws)
- This is not necessarily negative

# Different Types of Fairness

- Individual fairness
  - Similar individuals should receive similar outcomes
  - Requires a measure of similarity
- Utility-based fairness (economics)
  - Seeks Pareto optimality
  - Uses inequality measures like the Gini index
  - Example: Allocation of teachers to public schools in France
- Group fairness
  - Groups based on sensitive attributes
  - Groups should be treated similarly
  - Groups should have similar outcomes



# Outline

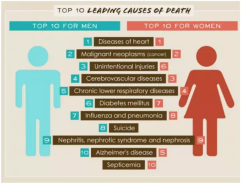
- Biased data: Systematic distortion that compromises its use
- Bias must be considered contextualized to the task

- Biased data: Systematic distortion that compromises its use
- Bias must be considered contextualized to the task

Gender in loan application


**Gender discrimination is illegal**

Gender in medical diagnosis




**Gender-specific medical diagnosis is desirable**

- Bias in data can come from various sources
  - Population-related bias
  - Behavior-related bias
  - Content production bias
  - Connection bias
  - Temporality bias

- Demographic differences

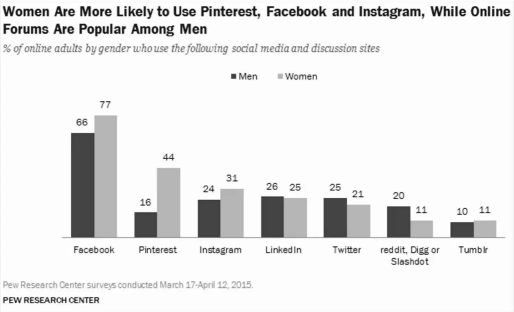
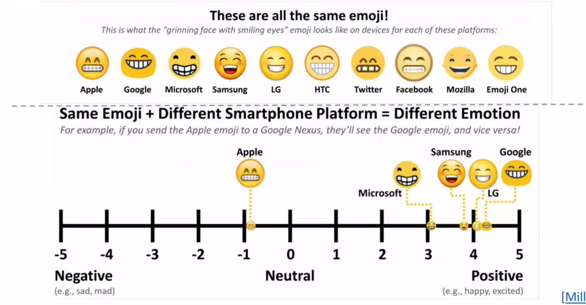



Figure from http://www.pewinternet.org/2016/11/11/social-media-update-2016/

- Differences in behavior on different platforms or contexts

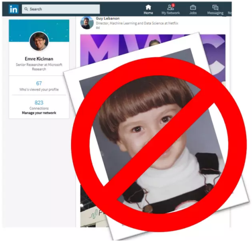



[Miller et al. ICWSM'16]

Figure from: http://grouplens.org/blog/investigating-the-potential-for-miscommunication-using-emoji/

Lexical, syntactic, semantic bias or structural differences in user-generated content

The kind of photos we use on Instagram vs LinkedIn




The same mechanism can have different meanings depending on the context

Likes on a social network can mean

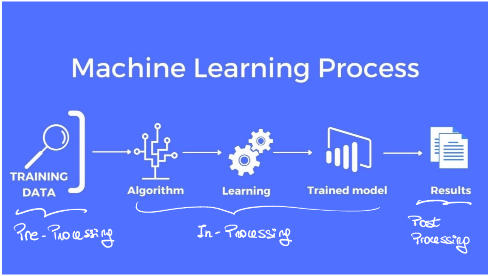
- affirmation
- denunciation
- approval
- displeasure
- etc.

- Connections: How the structure of the social network conditions our actions
  - Clusters tend to enhance polarization
  - For example, results in political elections

# Connections and Temporality

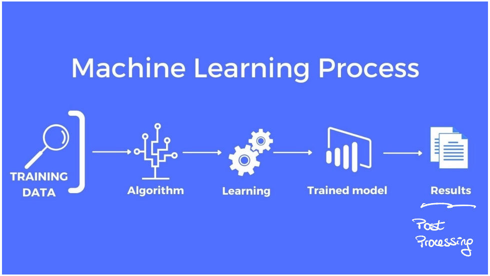
- Connections: How the structure of the social network conditions our actions
  - Clusters tend to enhance polarization
  - For example, results in political elections
- Temporality refers to how the social network changes over time
  - The increase in the number of people in the social network can be conditioned
  - Changes in platform characteristics impact user behavior

There are three main ways to achieve fair methods



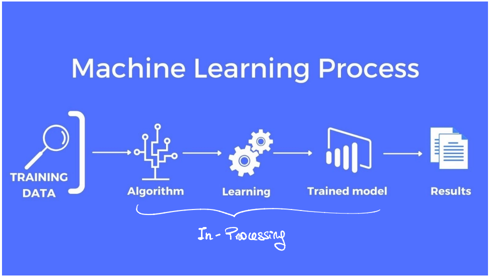
There are three main ways to achieve fair methods

- Post-processing: take a classifier without changes and massage the output to satisfy fairness metrics
  - Good for black-box methods
  - High risk of utility loss



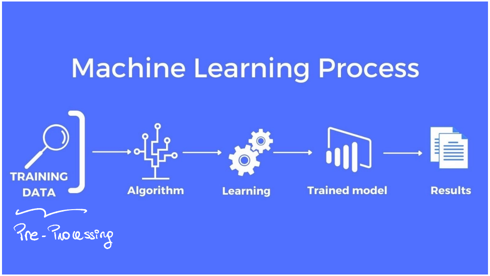
There are three main ways to achieve fair methods

- In-processing: Modify the training by including fairness constraints
  - Better utility
  - Specific to each method
  - Requires access to the database
  - May involve high mathematical complexity (optimization)



There are three main ways to achieve fair methods

- Pre-processing: Transform the database before training the model to be fair
  - Converts data obtained from various sources into a single clean database
  - Independent of the task
  - Specific to the fairness measure used



# Open problems

- Many nascent or open problems
- A lot of done in classification but not much in
  - Regression, recommendation, ranking, matching
  - Reinforcement learning, dynamic aspects
- Multi-sided and multi-stakeholders scenarios
- Multi-dimensional sensitive attributes
  - Intersectionality
- Multi-agent systems (e.g., ad auctions)
- Link fairness and privacy or fairness and stability
- Others...

# Key References

- Book “fairness and ML” [Barocas et al, 2020]
- Tutorials on fairness
  - Fairness-Aware Machine Learning in Practice [Bird et al., 2019]
  - Fairness in ML [Barocas & Hardt, 2017] video
  - 21 fairness definitions and their politics [Narayanan, 2018]
  - Fairness and representation learning tutorial [Cisse, Koyejo, 2019]
- Book “Pattern recognition and ML” [Bishop, 2006]
- Tutorial on Variational Autoencoders [Doersch, 2016]

Thank You :)