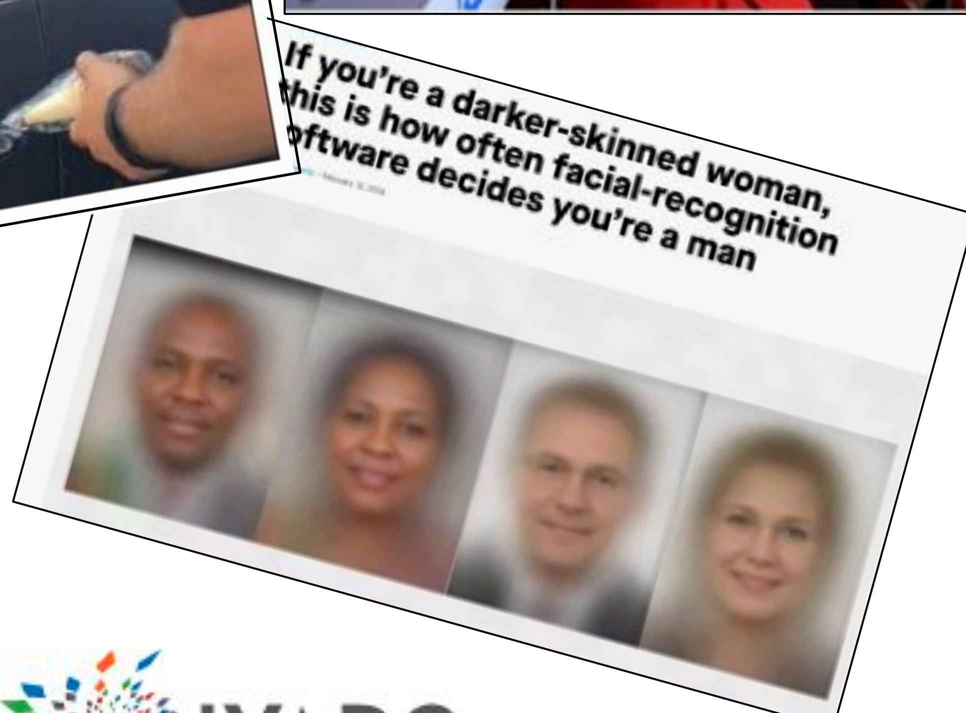
Bias and Discrimination in Machine Learning

Golnoosh Farnadi

Deep learning and Machine learning are everywhere!

Does ML create more problems than it solves?



Is there any solutions?

Trump Wants to Make It Basically Impossible to Sue for Algorithmic Discrimination

A new rule would make it easier for businesses to discriminate without consequence. That's the point.

Who's to Blame When Algorithms Discriminate?

A proposed rule from HUD would make it harder to hold people accountable for subtler forms of discrimination.

Can we create better algorithms for screening candidates - and reduce hiring bias?

By Neil Raden August 30, 2019

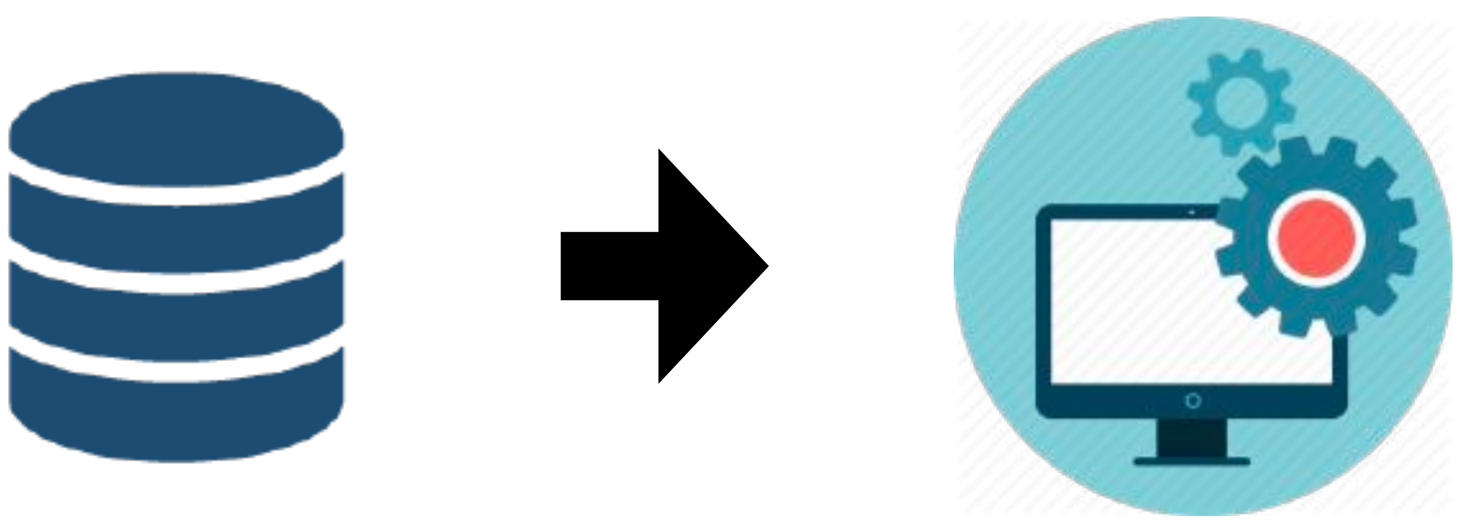
SUMMARY: A new research paper from Georgia Tech takes a surprising position algorithmic bias in hiring. Their view: we can reduce screening bias i algorithms take the impacted demographic groups into account. Her critique.

Can an algorithm eradicate bias in our decision making?

By Jonathan Rennie on 29 Aug 2019 in Artificial intelligence, General Data Protection Regulation, Data protection, Latest News

Reproducing Discrimination

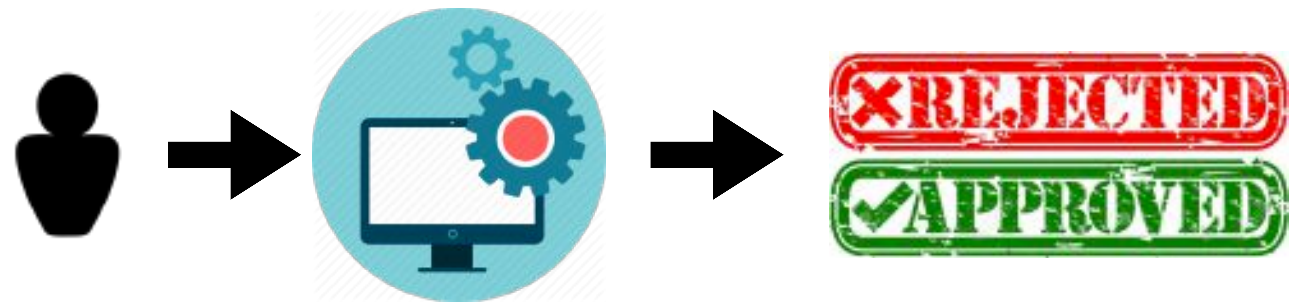
- Certain individuals have been historically discriminated against
- The decision-making system is learned from those unfair decisions



Records of unfair decisions

Learns to make unfair decisions

Accuracy is not enough



A hypothetical (extreme) situation:

Born and raised in Canada

- data describes them accurately
- accurate predictions (95% accurate)

90% of population

The model is still 90% accurate!

Migrated to Canada in recent years

- data describes them poorly
- poor predictions (50% accurate)

10% of population

Why we should care about fairness?

To address Law Against Discrimination!

Legally recognized 'protected classes'

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

Regulated domains

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

Fairness in ML

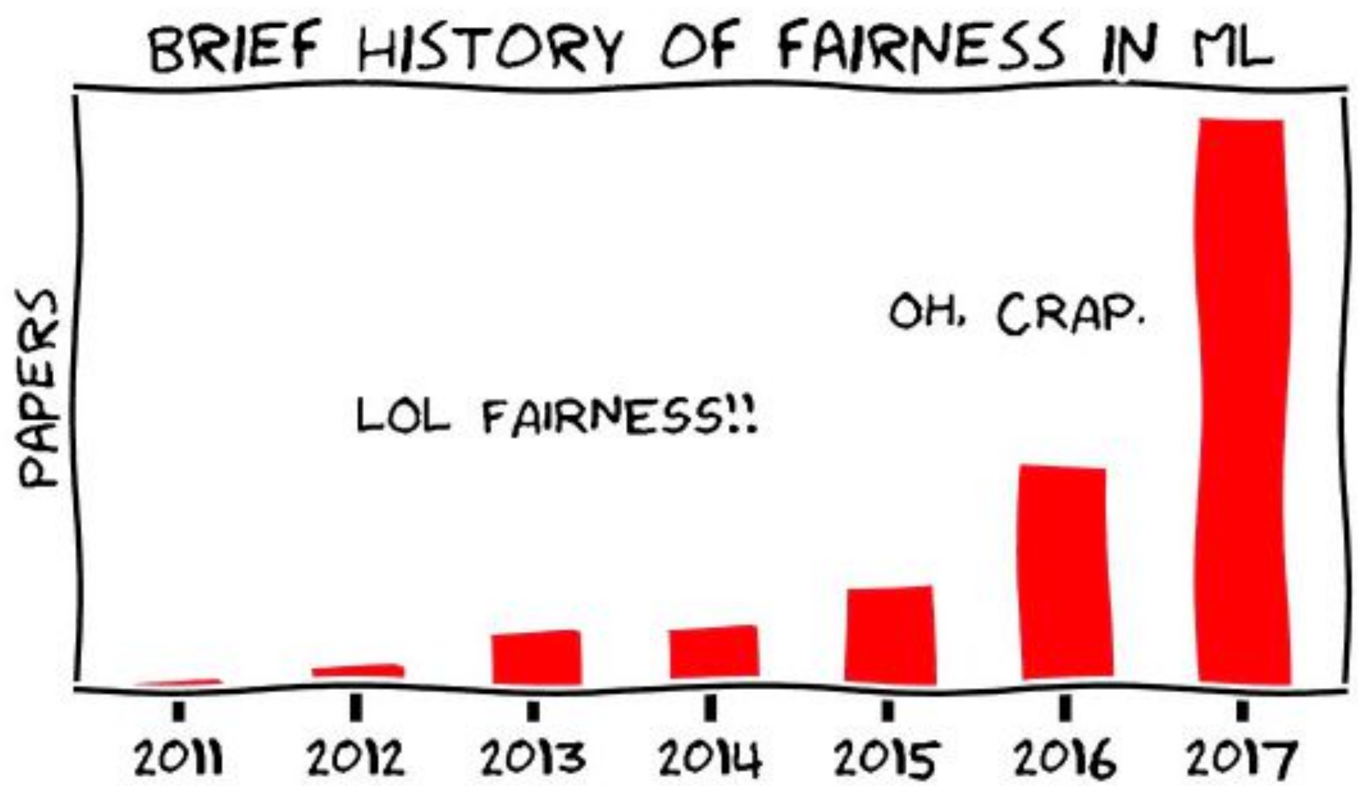
2014

2015

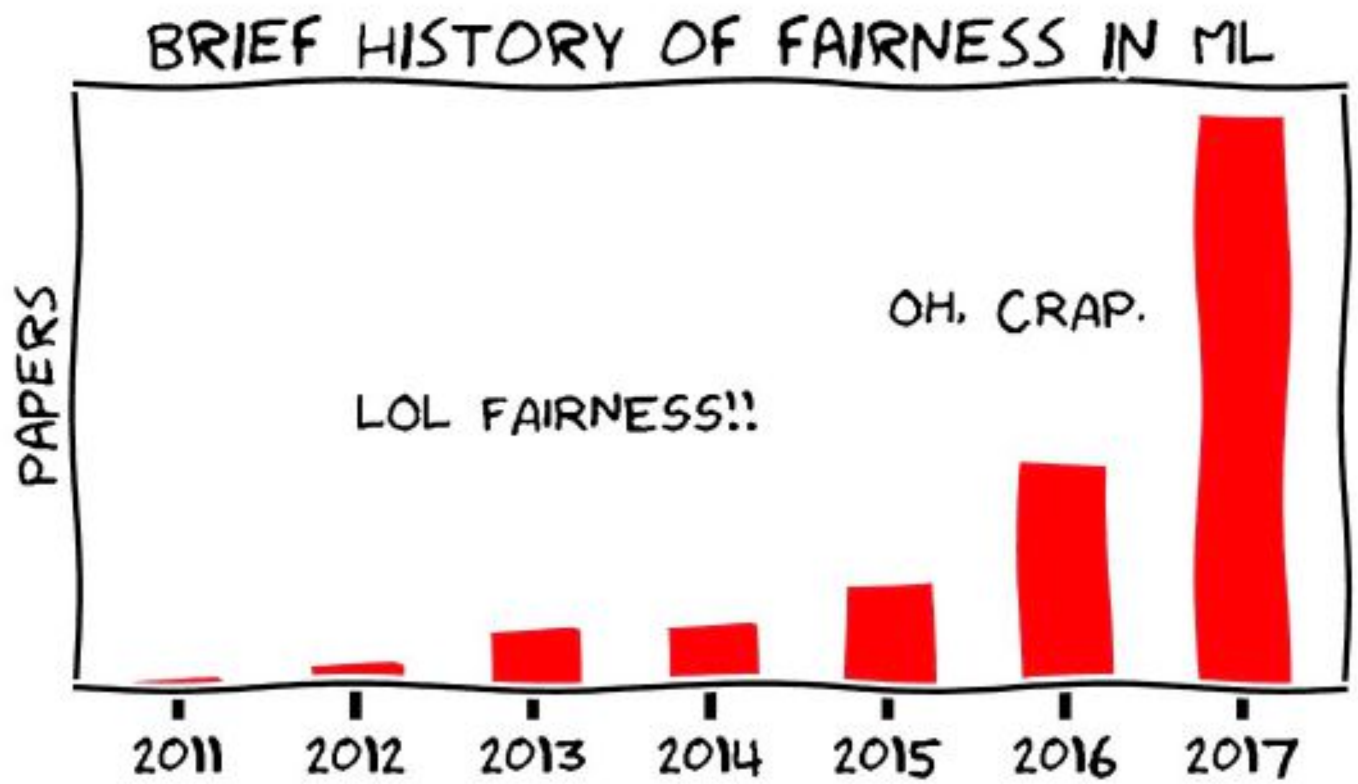
2016

2017

...



Fairness in ML



- “What is fair have been introduced in multiple disciplines for well over 50 years, including in education, hiring, and machine learning” [1].

- Statistics, Social Science, Economics, etc.

[1] Hutchinson, Ben, and Margaret Mitchell. "50 Years of Test (Un) fairness: Lessons for Machine Learning." *arXiv preprint arXiv:1811.10104* (2018).

How to address fairness in ML?

Input →  → Output

bias ·································▶

Pre-processing

Data is noisy
Biases
Encodes protected attributes

In-processing

Data scientists do not
build the models

Post-processing

unfair outcome
no user feedback

POLYTECHNIQUE MONTRÉAL UNIVERSITÉ D'INGÉNIERIE

IVADO

Mila

Université de Montréal

How to address fairness in ML?

Input →  → Output

bias ·········· ▶

Pre-processing

e.g.,

Discrimination Discovery
Un-bias the data
Sampling
Embedding
Dimension reduction

In-processing

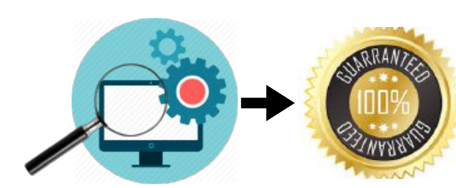
e.g.,

Learning subject to constraints
Ranking
Inference

Post-processing

e.g.,

Causal discovery
Transparency & Interpretability
Verification



Why do we use fairness definitions?

- To make algorithmic systems support human values!
- To identify strengths and weakness of the system
- To track improvement over time

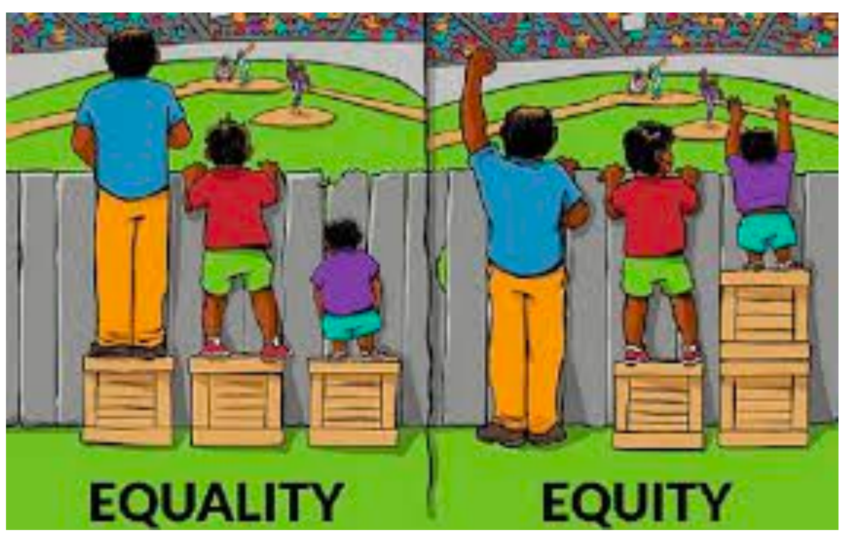
To address Law Against Discrimination!

POLYTECHNIQUE MONTRÉAL UNIVERSITÉ D'INGÉNIERIE IVADO Mila Université de Montréal

Why there are so many definitions?

An interesting tutorial by **Arvind Narayanan**:
Tutorial: 21 fairness definitions and their politics

Another interesting tutorial by **Jon Kleinberg**:
Inherent Trade-Offs in Algorithmic Fairness



Definition	Citation #
Group fairness or statistical parity	208
Conditional statistical parity	29
Predictive parity	57
False positive error rate balance	57
False negative error rate balance	57
Equalised odds	106
Conditional use accuracy equality	18
Overall accuracy equality	18
Treatment equality	18
Test-fairness or calibration	57
Well calibration	81
Balance for positive class	81
Balance for negative class	81
Causal discrimination	1
Fairness through unawareness	14
Fairness through awareness	208
Counterfactual fairness	14
No unresolved discrimination	14
No proxy discrimination	14
Fair inference	6

Verma, Sahil, and Julia Rubin. "Fairness definitions explained." *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 2018.

Why we don't have one definition?

Fairness is not a general concept!

Correcting for algorithmic bias generally requires:

- knowledge of how the measurement process is biased
- judgments about properties to satisfy in an "unbiased" world

Hiring

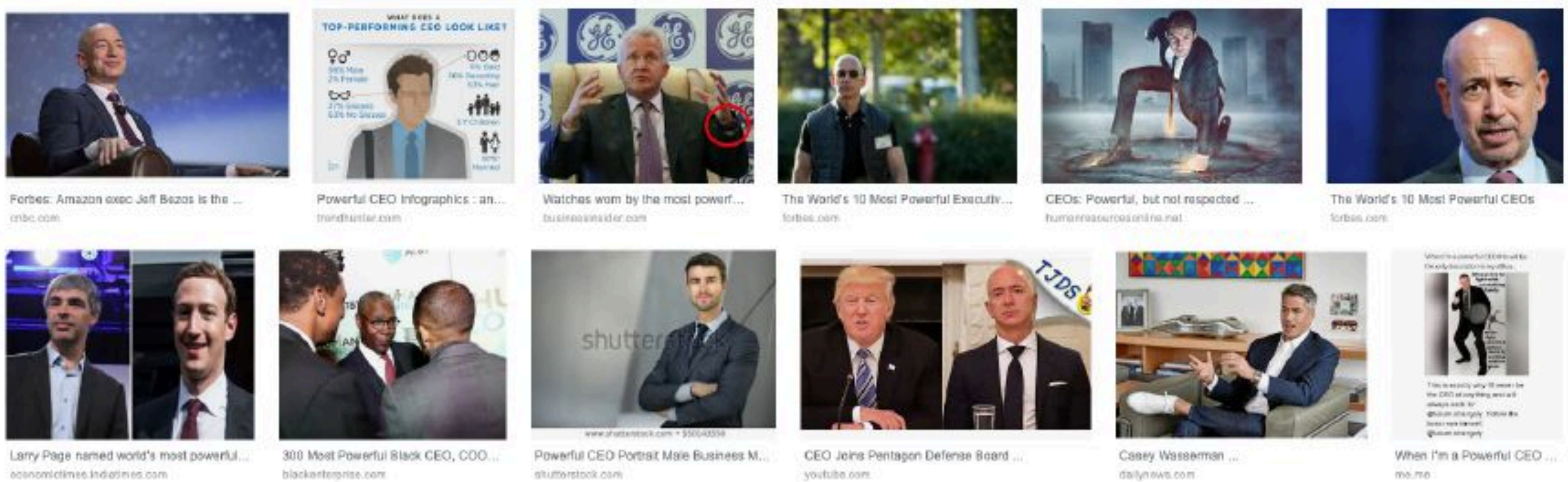
Gender-biased

Medical diagnosis

Gender-biased

Bias is **subjective** and must be considered **relative** to task

There is no agreed-upon measure



There is no single agreed-upon measure for discrimination/fairness

What is **fair?**

50% **female,** 50% **male?**

Based on the **population?**

Results for "CEO" in Google Images: 11% female, US 27% female CEOs

POLYTECHNIQUE MONTRÉAL UNIVERSITÉ D'INGÉNIERIE

IVADO

Mila

Université de Montréal

Different types of fairness definitions

Types of fairness definitions

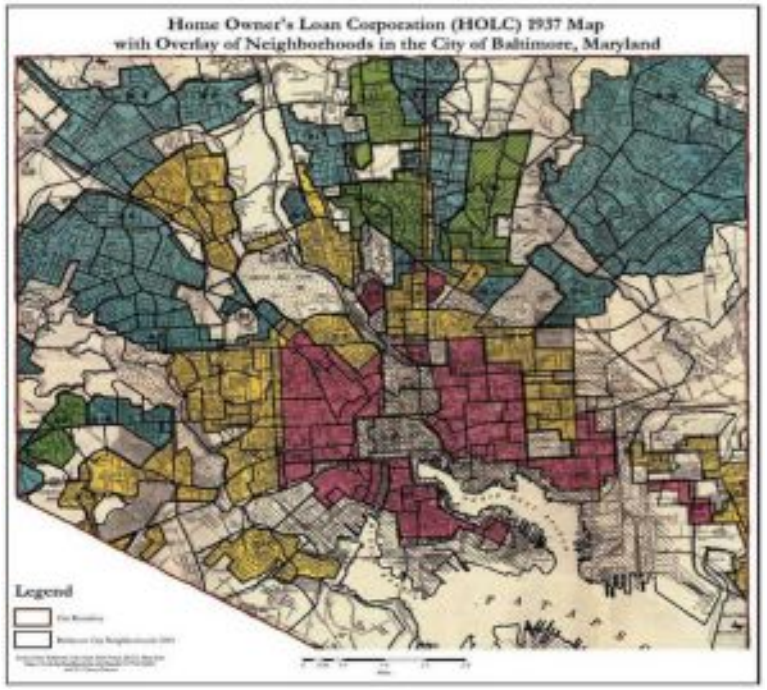
Different definitions based on legal concepts

- Direct vs indirect discrimination
- Individual vs group fairness
- Explainable vs unexplainable discrimination

Indirect discrimination

Direct discrimination happens when a person is treated less favourably because of one of the attributes

Name	Postal code	...	Decision
Richard	H3C	=	
Bob	F4C	=	



Indirect discrimination is when there's a practice, policy or rule which applies to everyone in the same way, but it has a worse effect on some people than others. The Equality Act says it puts you at a particular disadvantage.

Types of fairness definitions

Different definitions based on legal concepts

- Direct vs indirect discrimination
- **Individual vs group fairness**
- Explainable vs unexplainable discrimination

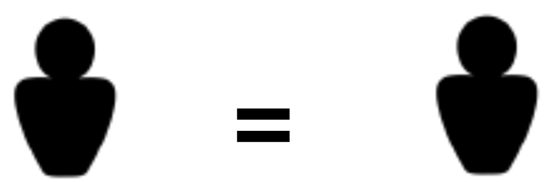
Definition	Citation #
Group fairness or statistical parity	208
Conditional statistical parity	29
Predictive parity	57
False positive error rate balance	57
False negative error rate balance	57
Equalised odds	106
Conditional use accuracy equality	18
Overall accuracy equality	18
Treatment equality	18
Test-fairness or calibration	57
Well calibration	81
Balance for positive class	81
Balance for negative class	81
Causal discrimination	1
Fairness through unawareness	14
Fairness through awareness	208
Counterfactual fairness	14
No unresolved discrimination	14
No proxy discrimination	14
Fair inference	6

Verma, Sahil, and Julia Rubin. "Fairness definitions explained." *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 2018.

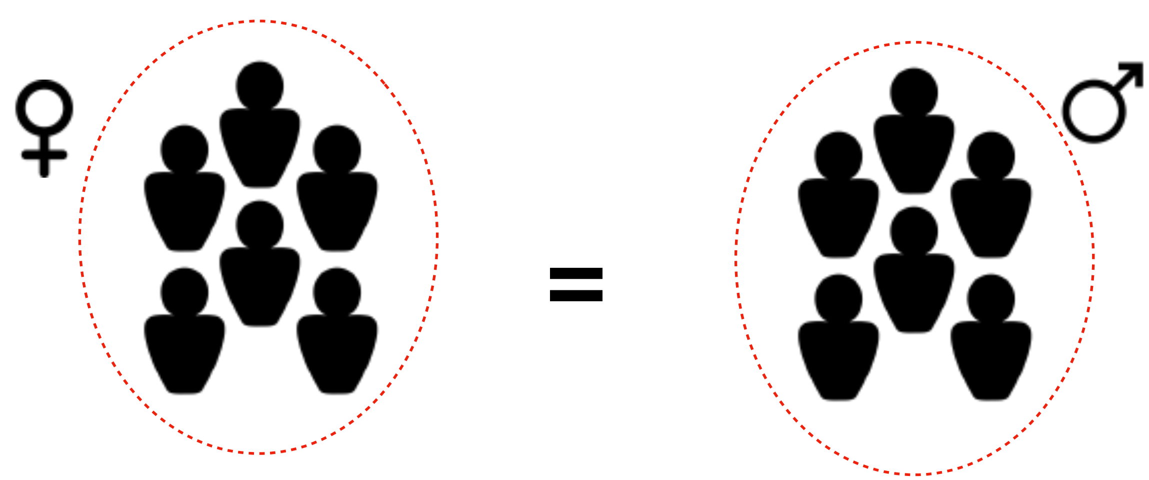
Types of fairness definitions

Group fairness VS. Individual Fairness

- **Individual**: the impact that the discrimination has on the individuals.



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



Impossibility theorem

Metric	Equalized under
Selection probability	Demographic parity
Positive predictive value	Predictive parity
Negative predictive value	Predictive parity
False positive rates	Error rate balance
False negative rate	Error rate balance
Accuracy	Accuracy equity

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." *arXiv preprint arXiv:1609.05807* (2016).

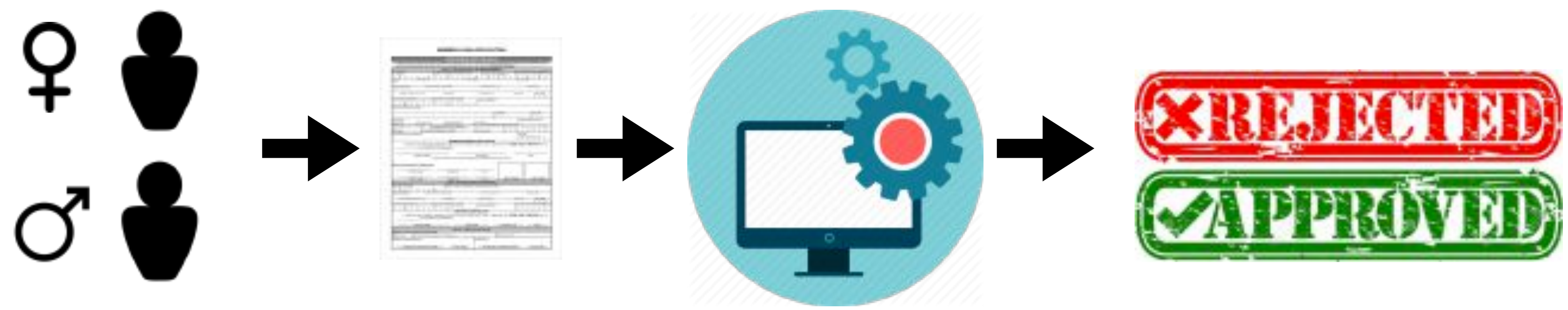
Chouldechova, Alexandra. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." *Big data* 5.2 (2017): 153-163.

Differences of fairness definitions (mathematical notations)

TN	FP
FN	TP

confusion matrix

Notations



Applicant **Application** **Credit Approval**

G	X	d	Y
sensitive attribute	**non-sensitive attributes**	**Prediction decision**	**Actual Outcome**

Female $G = f$

Male $G = m$

$d = 1$

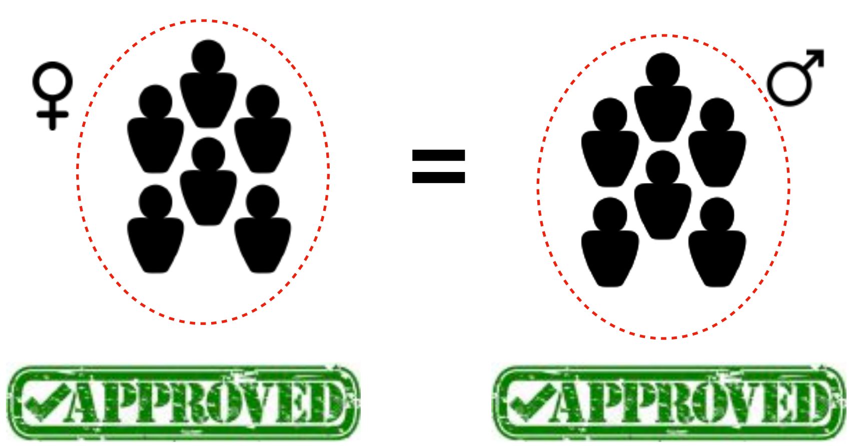
Group fairness

a predicted outcome

1- Group fairness / **statistical (demographic) parity** / equal acceptance rate / benchmarking

$$p(d = 1 | G = f) = p(d = 1 | G = m)$$

equal probability of being assigned to the positive predicted class



POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

IVADO

Mila

Université de Montréal

Group fairness

a predicted outcome

Issues with demographic parity:

$$p(d = 1|G = f) = p(d = 1|G = m)$$

1. The notion permits that a classifier selects qualified applicants in female group, but unqualified individuals in male group

Group fairness

a predicted outcome

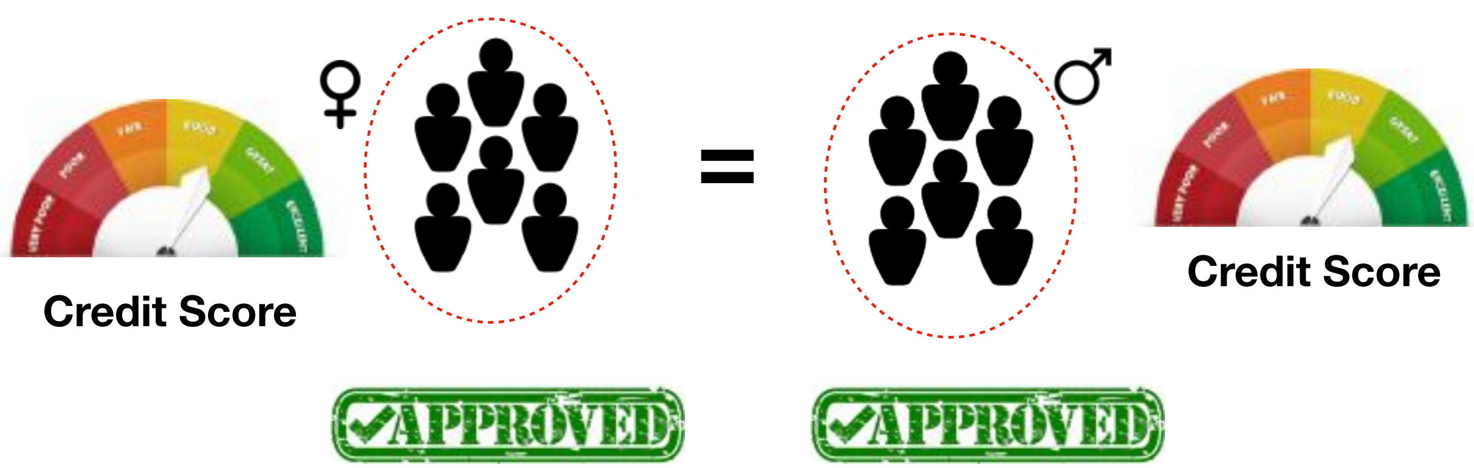
2- **Conditional statistical parity**

$$p(d=1|L=1,G=f)=p(d=1|L=1,G=m)$$

legitimate factors

$$L$$

both protected and unprotected groups have equal probability of being assigned to the positive predicted class, controlling for a set of legitimate factors L.



Group fairness

a predicted outcome

Issues with demographic parity:

$$p(d = 1 | G = f) = p(d = 1 | G = m)$$

1. The notion permits that a classifier selects qualified applicants in female group, but unqualified individuals in male group

2. Demographic parity would rule out the ideal predictor

Group fairness

a predicted outcome+ Actual outcome

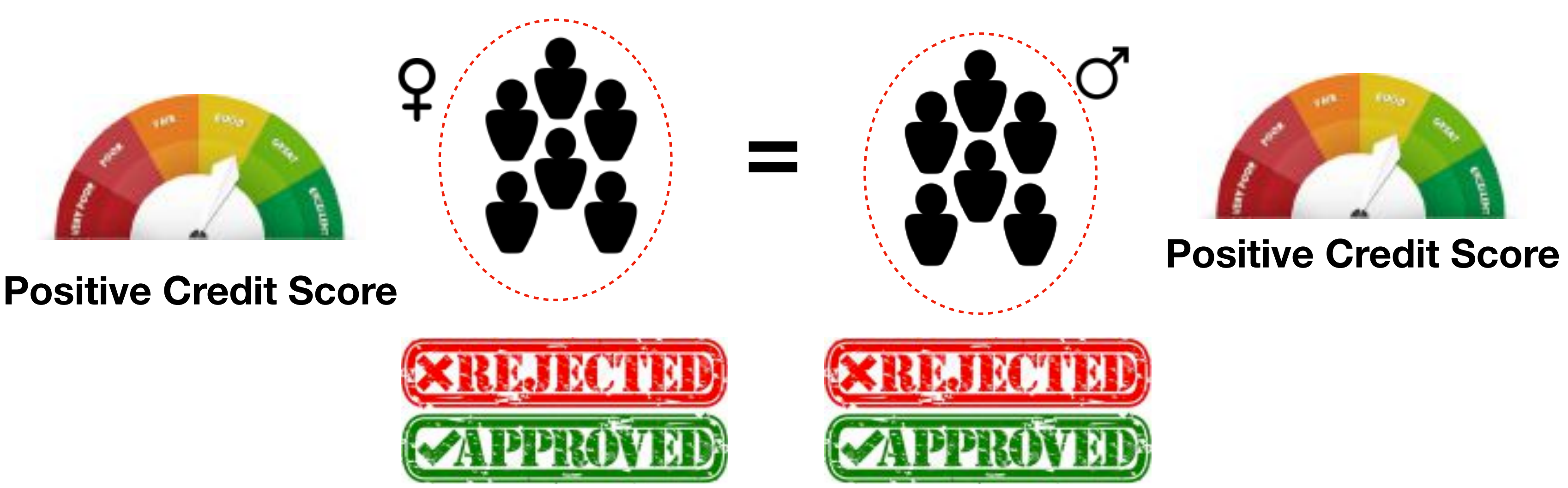
3- False negative error rate balance / **equal opportunity**

$$p(d=0|Y=1,G=f)=p(d=0|Y=1,G=m)$$

$$=$$

$$p(d=1|Y=1,G=f)=p(d=1|Y=1,G=m)$$

classifier should give similar results for applicants of both genders with actual positive credit scores



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

POLYTECHNIQUE MONTRÉAL UNIVERSITÉ D'INGÉNIERIE IVADO Mila Université de Montréal

Group fairness

a predicted outcome+ Actual outcome

3- False negative error rate balance / **equal opportunity**

$$p(d=0|Y=1,G=f)=p(d=0|Y=1,G=m)$$

$$=$$

$$p(d=1|Y=1,G=f)=p(d=1|Y=1,G=m)$$

Picks for each group a threshold such that the fraction of non-defaulting group members that qualify for credit is the same.

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

Group fairness

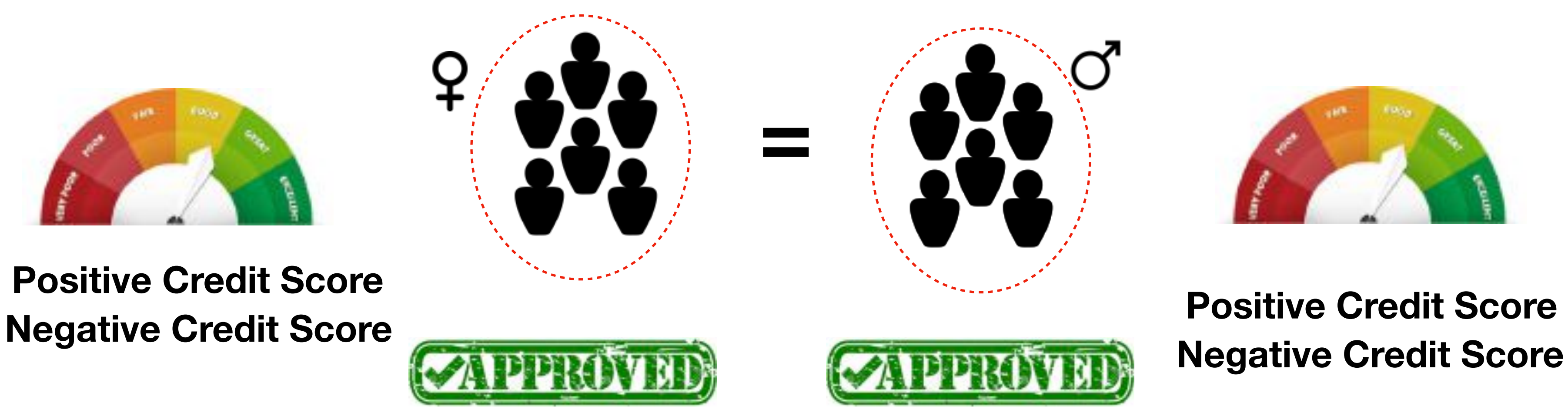
a predicted outcome+ Actual outcome

4- **Equalized odds** / conditional procedure accuracy equality / disparate mistreatment

$$p(d=1|Y=I,G=f)=p(d=1|Y=I,G=m)$$

where $I \in \{0,1\}$

applicants with a good actual credit scope and applicants with a bad actual credit score should have a similar classification, regardless of their gender



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

Group fairness

a predicted outcome+ Actual outcome

4- **Equalized odds** / conditional procedure accuracy equality / disparate mistreatment

$$p(d=1|Y=I, G=f) = p(d=1|Y=I, G=m)$$

where $I \in \{0,1\}$

Picks two thresholds for each group, so above both thresholds people always qualify and between the thresholds people qualify with some probability.

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

Individual fairness

1- Fairness through unawareness, **Fairness through blindness**

$$X : X_i = X_j \rightarrow d_i = d_j$$

This can be a non-obvious encoding in terms of many features, learned from the data

Causal Discrimination

2- **Causal discrimination**

$$(X_f = X_m \wedge G_f \neq G_m) \rightarrow d_f = d_m$$

the same classification for any two subjects with the exact same attributes X

Name	Gender	...	Decision
Alice	**female**	=	
Bob	**male**	=	

This can be impossible due to dependency between features!

Galhotra, Sainyam, Yuriy Brun, and Alexandra Meliou. "Fairness testing: testing software for discrimination." *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ACM, 2017.

Individual Fairness

3- **Fairness through awareness**

$$D(M(x), M(y) \rightarrow k(x, y)$$

e.g., $D(i, j) = S(i) - S(j)$

Distance metric Between two Distributions

D

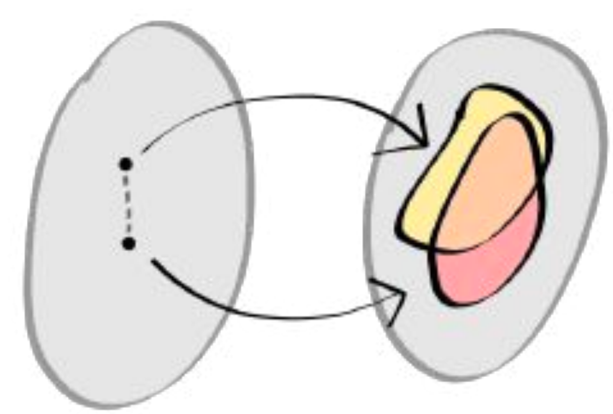
Distance metric Between two individuals

k

similar individuals should have similar classification

seemingly different individuals

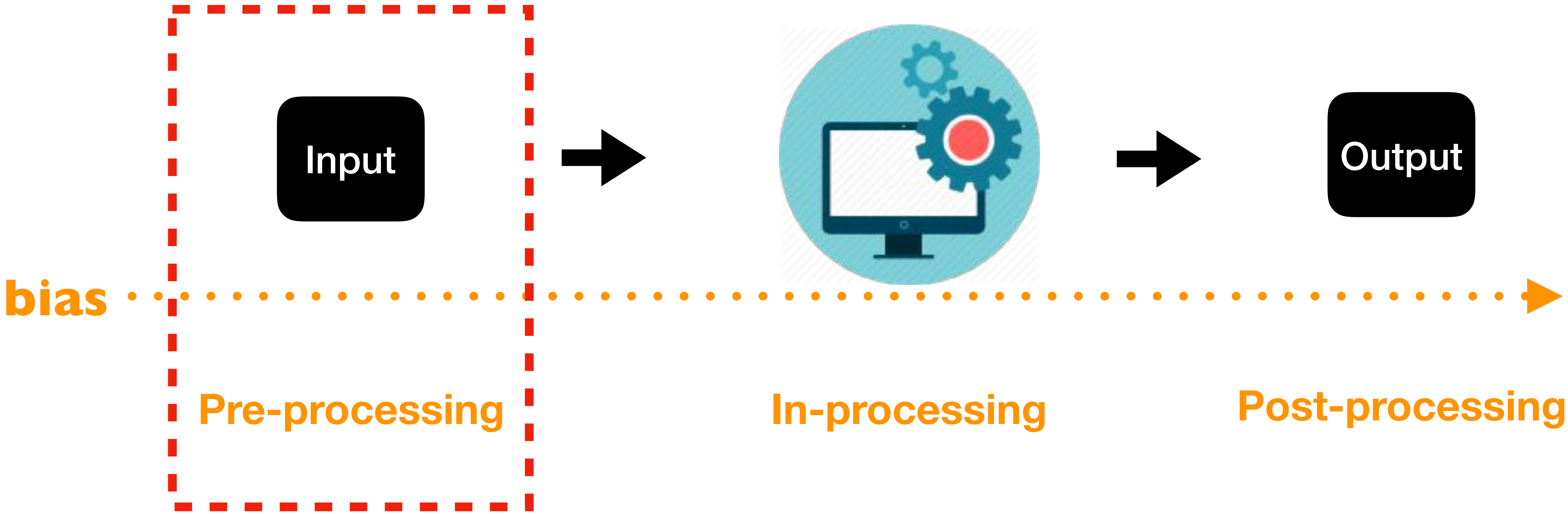
Name	Gender	...	Decision
Alice	female	=	APPROVED
Bob	male	=	APPROVED



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

Fairness in Machine Learning
(a few examples)

Fairness in Pre-Processing



POLYTECHNIQUE MONTRÉAL UNIVERSITÉ D'INGÉNIERIE · IVADO · Mila · Université de Montréal

Data bias differs from Data quality

Data Quality issues:

- **Sparse data:** e.g., measures follow a power law distribution
- **Noise:** e.g., not reliable data, or incomplete and corrupted, typos, infrequent terms, stop words.
- **Representativeness**: e.g., a sample data is not representative of the larger population.

Data Bias: a systematic distortion in data that compromises its use for a task.

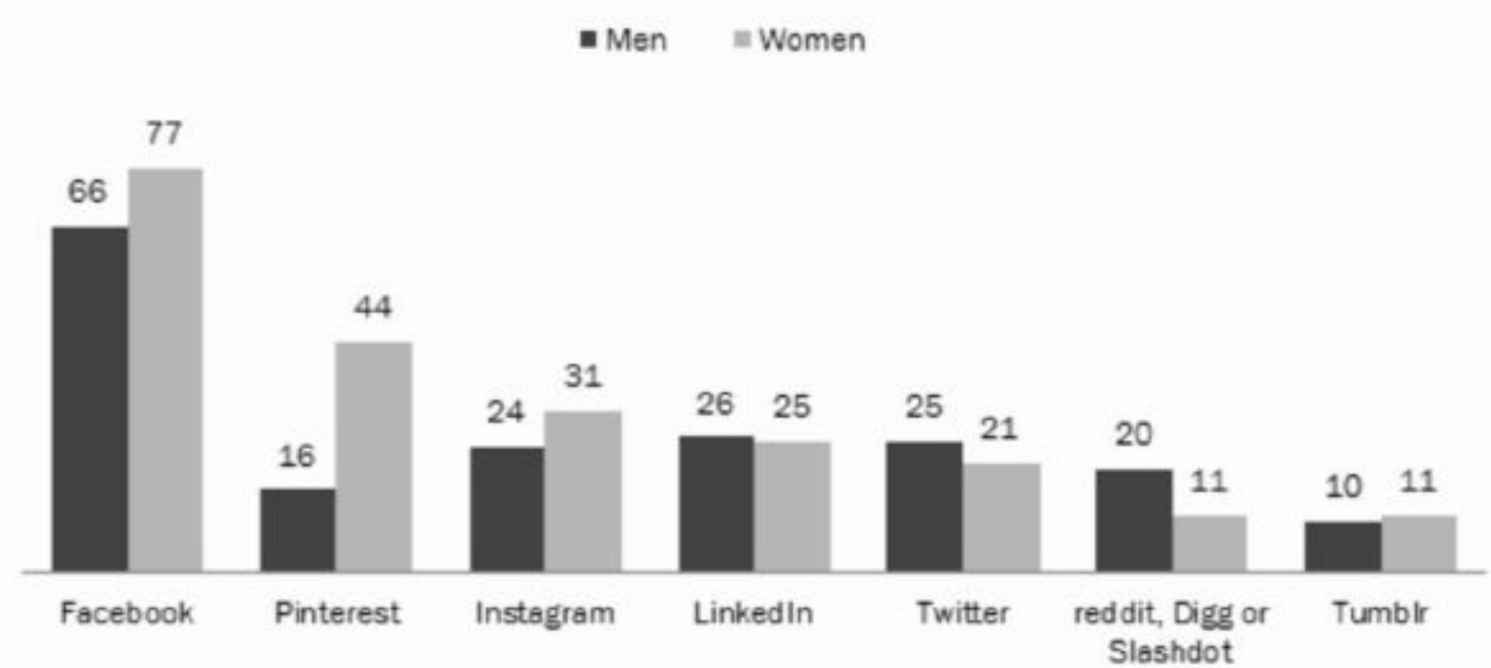
Where the data bias comes from?

1. **Population biases**
2. **Behavioural biases**
3. **Content production biases**
4. **Linking biases**
5. **Temporal biases**

Olteanu, Alexandra and Castillo, Carlos and Diaz, Fernando and Kiciman, Emre, Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries (December 20, 2016). Frontiers in Big Data 2:13. doi: 10.3389/fdata.2019.00013. Available at SSRN: https://ssrn.com/abstract=2886526or http://dx.doi.org/10.2139/ssrn.2886526

Where the data bias comes from?

1. **Population biases**
2. Behavioural biases
3. Content production biases
4. Linking biases
5. Temporal biases

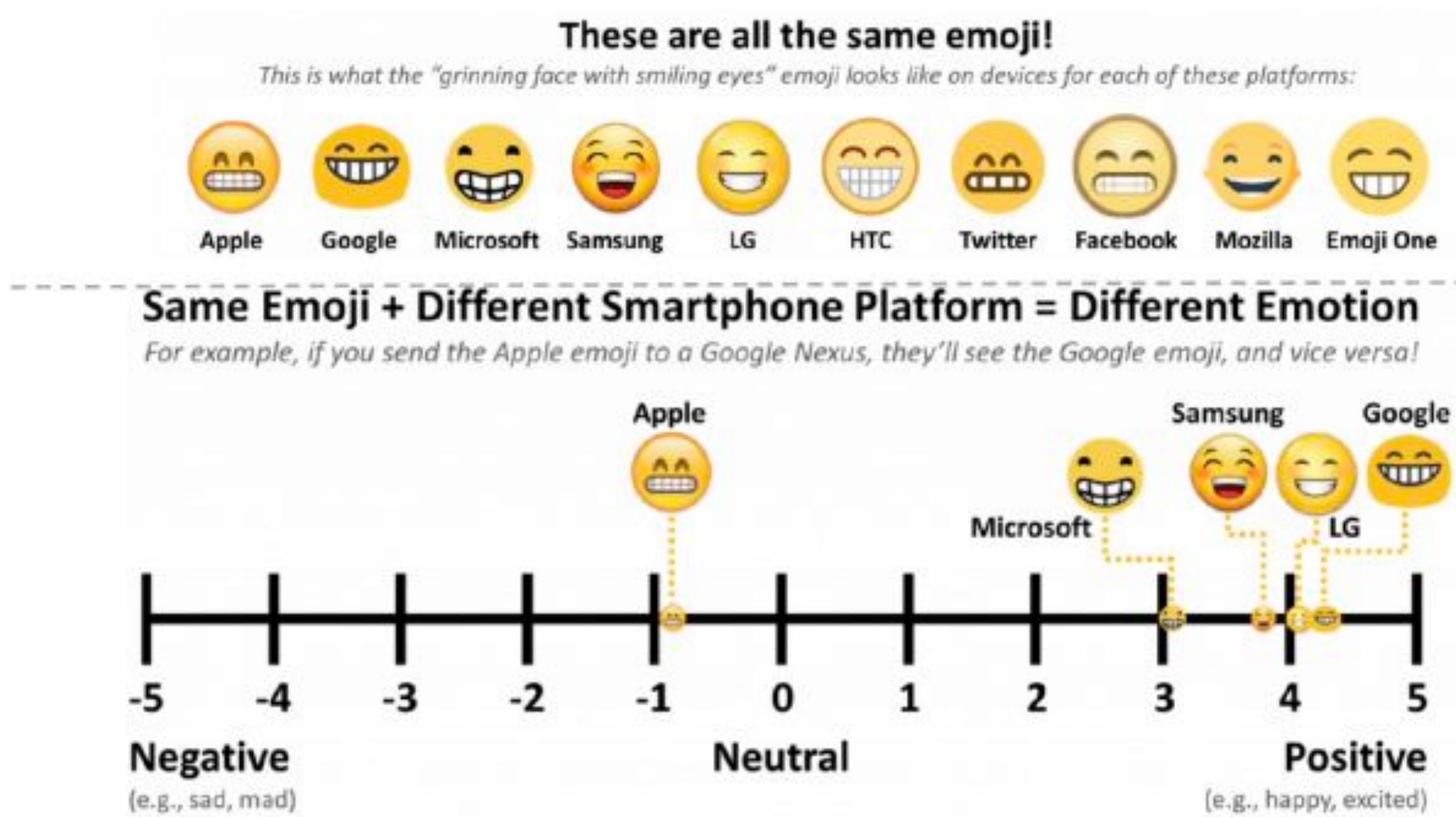


Differences in demographics or other user characteristics between a user population represented in a dataset or platform and a target population

Olteanu, Alexandra and Castillo, Carlos and Diaz, Fernando and Kiciman, Emre, Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries (December 20, 2016). Frontiers in Big Data 2:13. doi: 10.3389/fdata.2019.00013. Available at SSRN: https://ssrn.com/abstract=2886526or http://dx.doi.org/10.2139/ssrn.2886526

Where the data bias comes from?

1. Population biases
2. **Behavioural biases**
3. Content production biases
4. Linking biases
5. Temporal biases



Differences in user behavior across platforms or contexts, or across users represented in different datasets

Olteanu, Alexandra and Castillo, Carlos and Diaz, Fernando and Kiciman, Emre, Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries (December 20, 2016). Frontiers in Big Data 2:13. doi: 10.3389/fdata.2019.00013. Available at SSRN: https://ssrn.com/abstract=2886526or http://dx.doi.org/10.2139/ssrn.2886526

Where the data bias comes from?

1. Population biases
2. Behavioural biases
3. **Content production biases**
4. Linking biases
5. Temporal biases

The use of language(s) varies across and within countries and populations

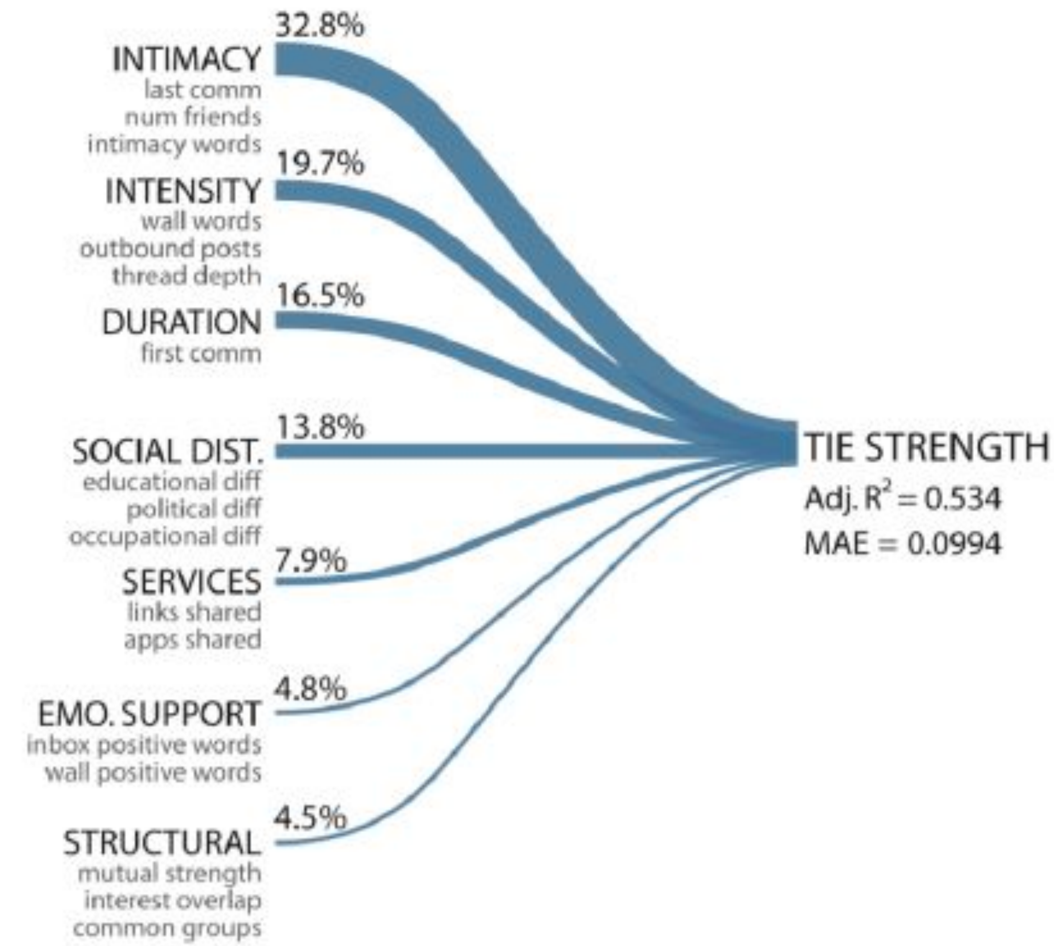
Feature	*#female/#male*
Emoticons	3.5
Elipses	1.5
Character repetition	1.4
Repeated exclamation	2.0
Puzzled punctuation	1.8
OMG	4.0

Lexical, syntactic, semantic, and structural differences in the contents generated by users

Olteanu, Alexandra and Castillo, Carlos and Diaz, Fernando and Kiciman, Emre, Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries (December 20, 2016). Frontiers in Big Data 2:13. doi: 10.3389/fdata.2019.00013. Available at SSRN: https://ssrn.com/abstract=2886526or http://dx.doi.org/10.2139/ssrn.2886526

Where the data bias comes from?

1. Population biases
2. Behavioural biases
3. Content production biases
4. **Linking biases**
5. Temporal biases



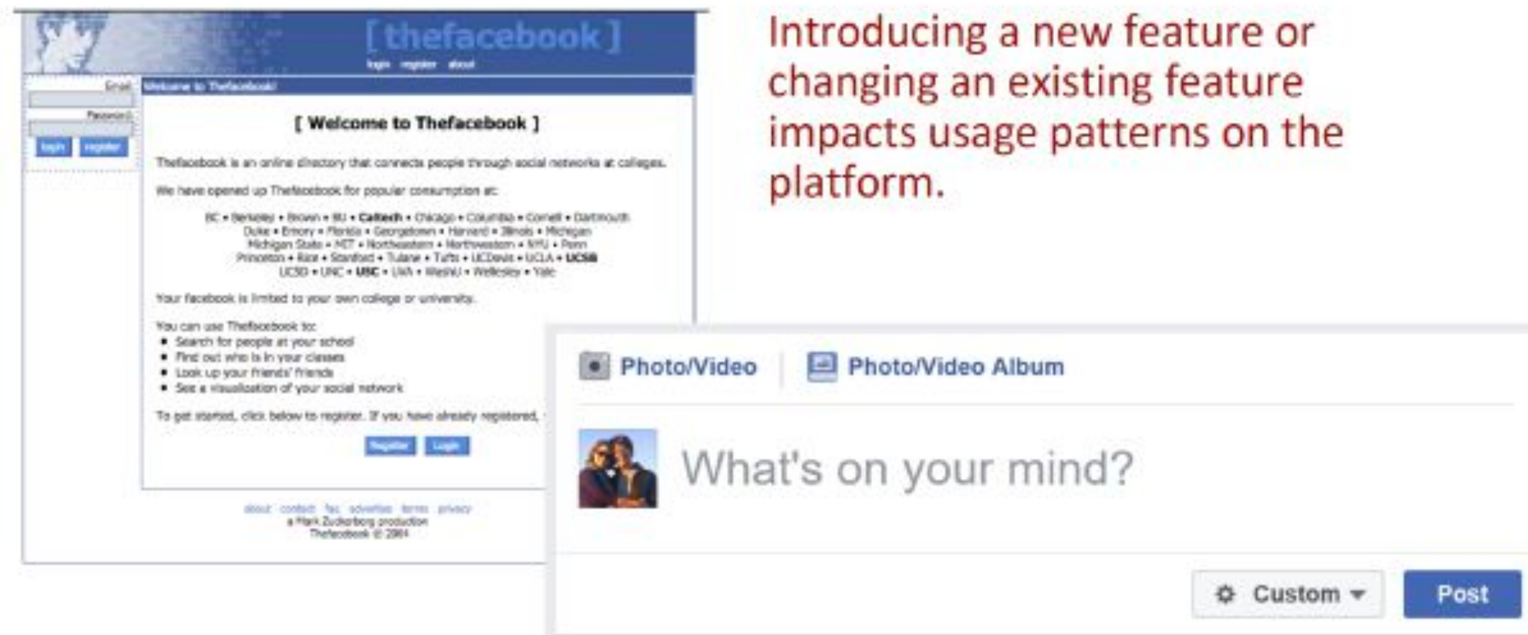
Differences in the attributes of networks obtained from user connections, interactions, or activity

Olteanu, Alexandra and Castillo, Carlos and Diaz, Fernando and Kiciman, Emre, Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries (December 20, 2016). Frontiers in Big Data 2:13. doi: 10.3389/fdata.2019.00013. Available at SSRN: https://ssrn.com/abstract=2886526or http://dx.doi.org/10.2139/ssrn.2886526

Where the data bias comes from?

1. Population biases
2. Behavioural biases
3. Content production biases
4. Linking biases
5. **Temporal biases**

E.g., Change in Features over Time



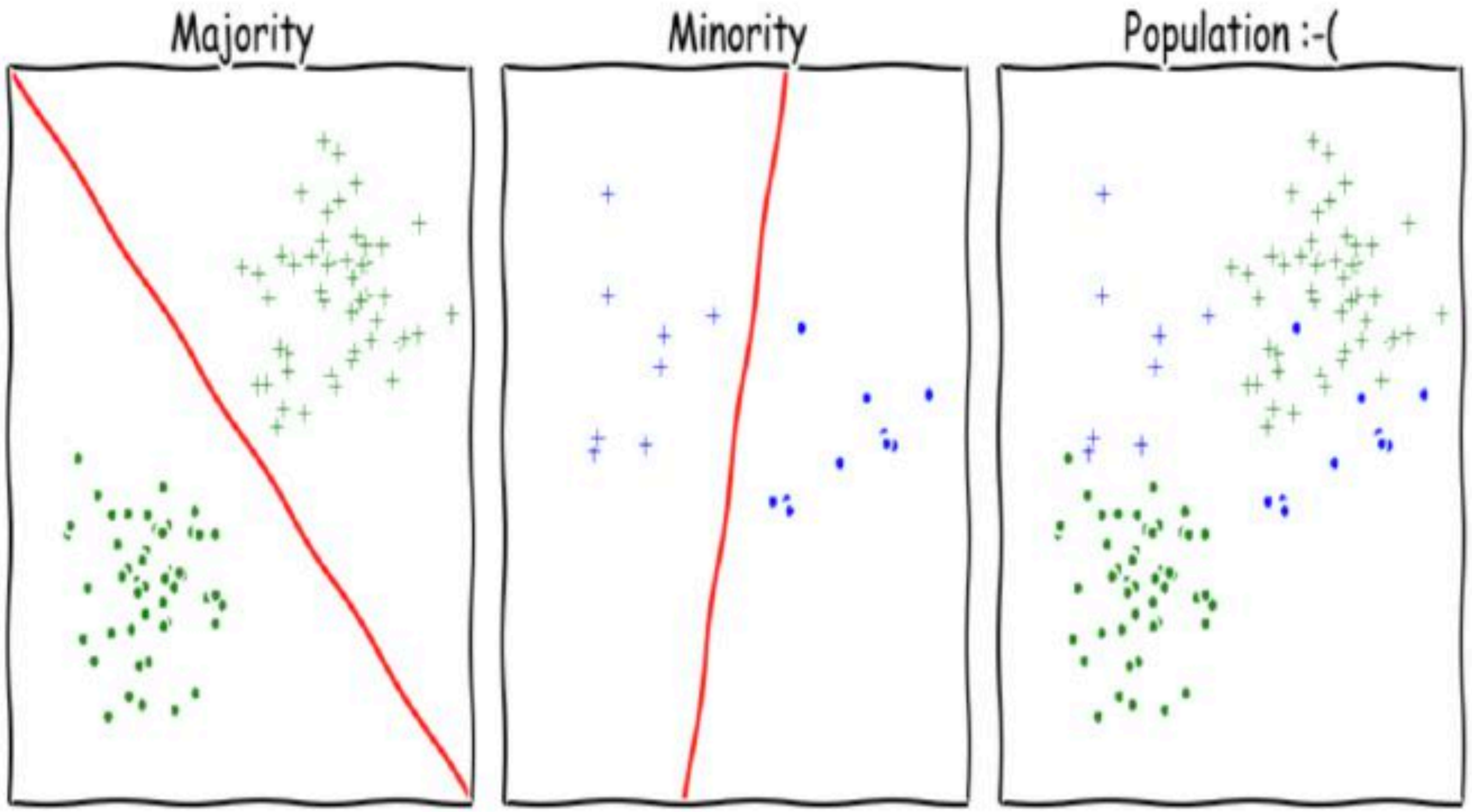
Introducing a new feature or changing an existing feature impacts usage patterns on the platform.

Differences in populations and behaviors over time

Olteanu, Alexandra and Castillo, Carlos and Diaz, Fernando and Kiciman, Emre, Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries (December 20, 2016). Frontiers in Big Data 2:13. doi: 10.3389/fdata.2019.00013. Available at SSRN: https://ssrn.com/abstract=2886526or http://dx.doi.org/10.2139/ssrn.2886526

POLYTECHNIQUE MONTRÉAL UNIVERSITÉ D'INGÉNIERIE IVADO Mila Université de Montréal

Data Cleaning



Data clearing is not the final solution!

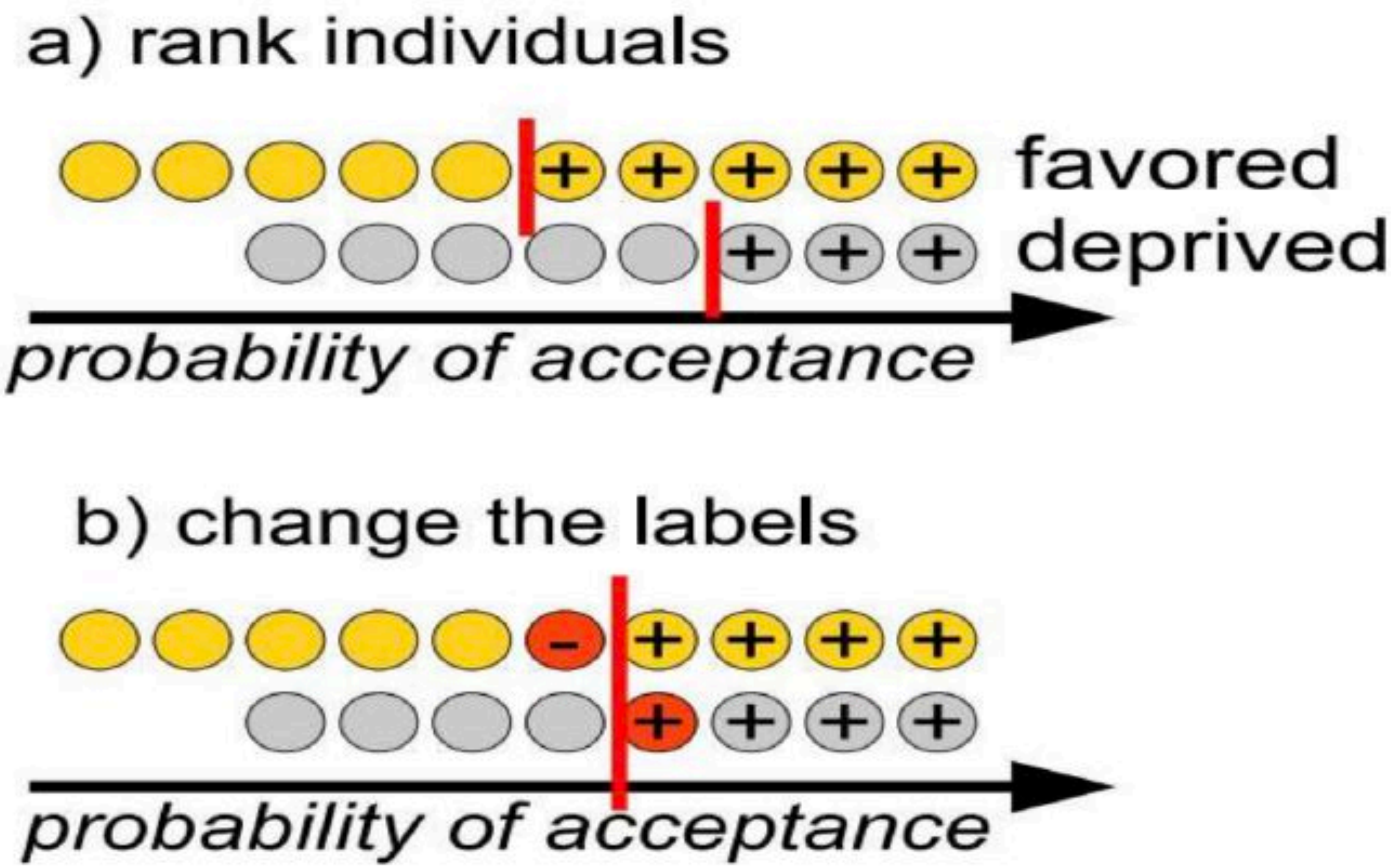
Some data cleaning techniques

- **Massaging**
- **Re-weighting**
- **Sampling**
- **….**
- **GAN**

Gender			Decision
♂	…	…	+
♂	…	…	+
♂	…	…	+
♂	…	…	-
♂	…	…	+
♂	…	…	+
♂	…	…	-
♂	…	…	-

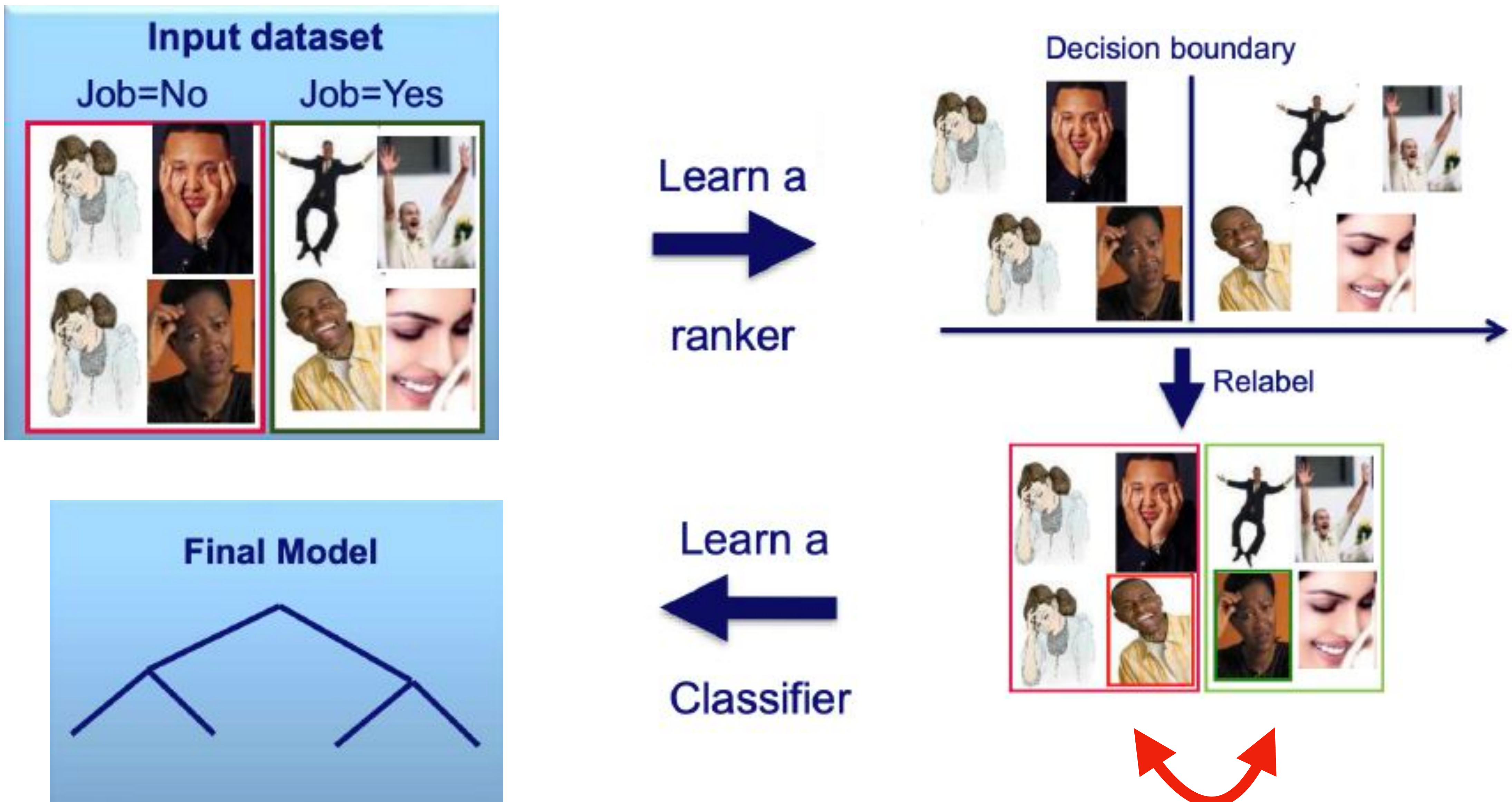
Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

Massaging



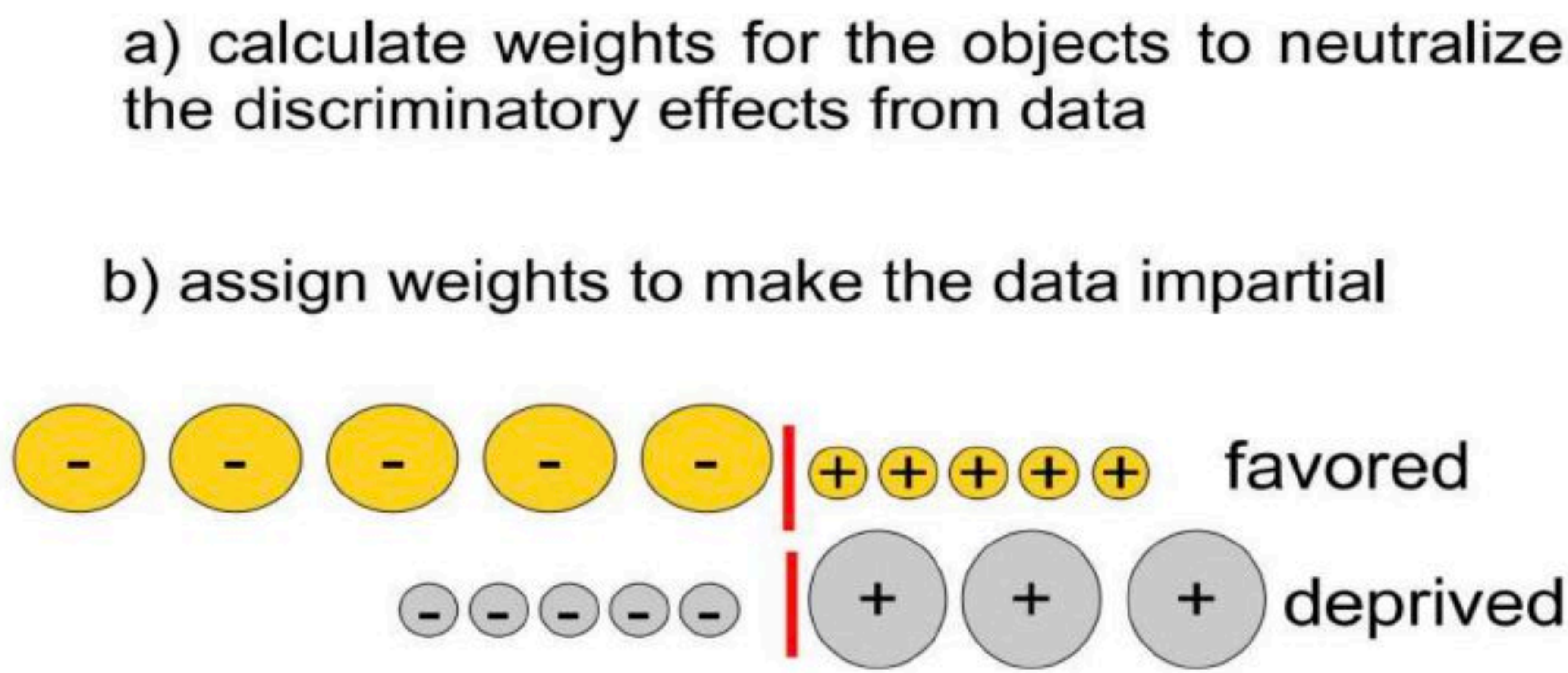
Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

Massaging



Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

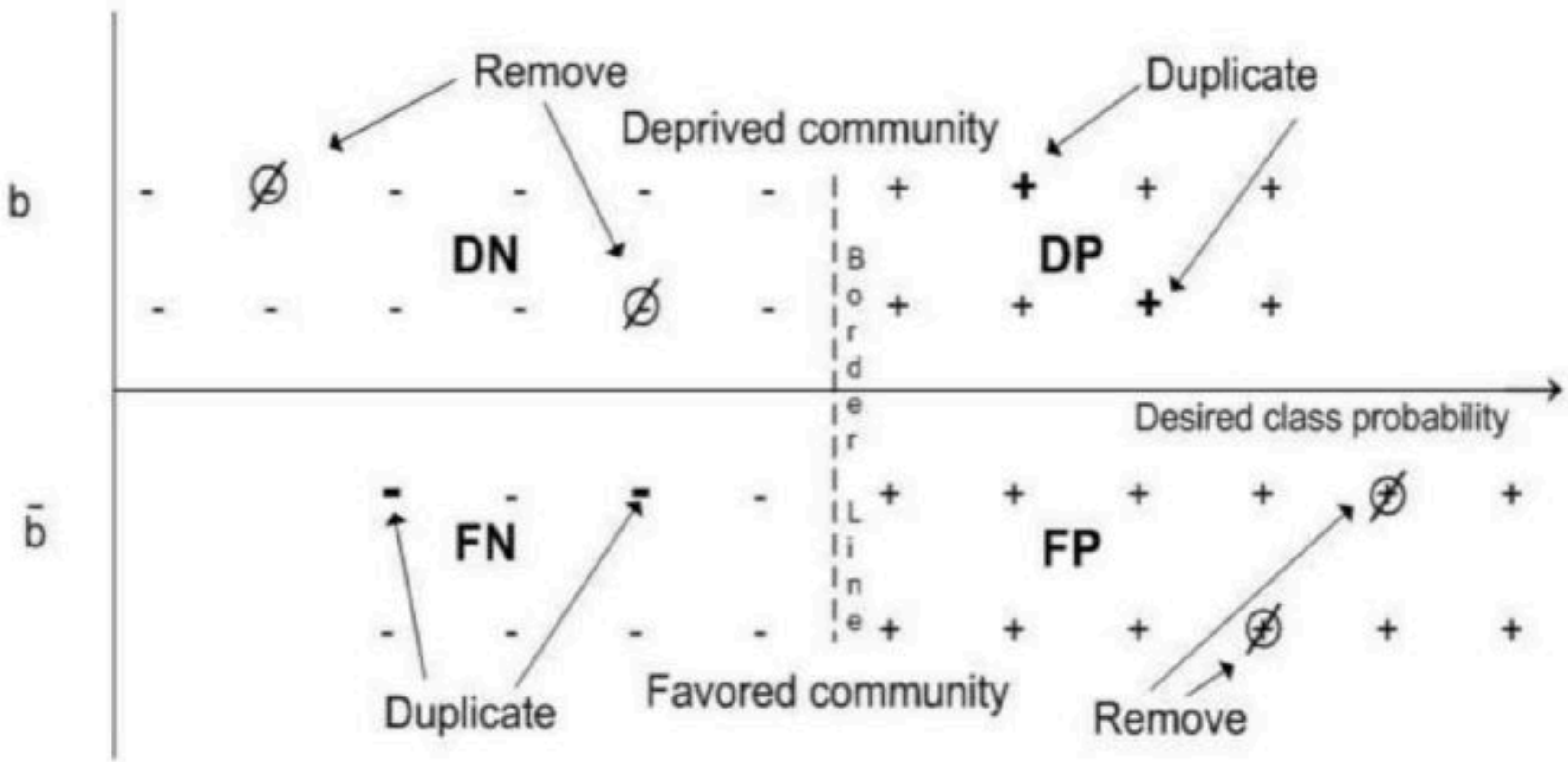
Re-Weighting



Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

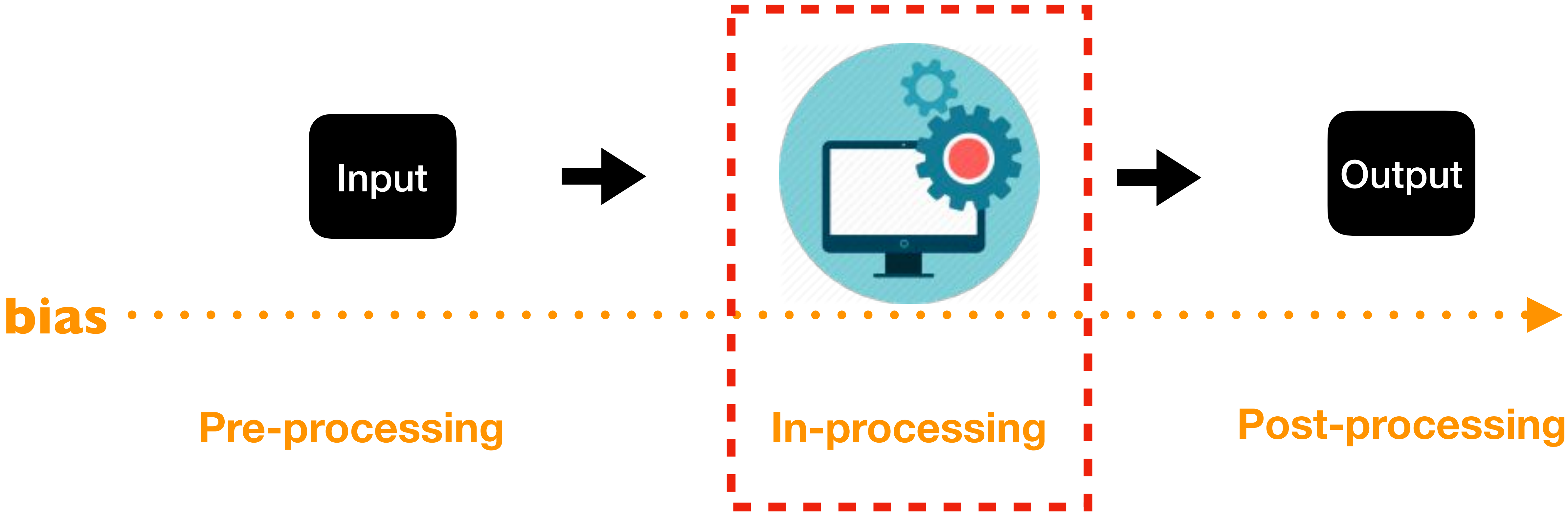
Sampling

Similarly to reweighing, compare the expected size of a group with its actual size, to define a sampling probability.



Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

Fairness in Processing



- Learning subject to constraints

Learning subject to fairness constrains

Supervised learning tasks are often expressed as optimization problems

$$\text{minimize.} \quad \underbrace{f_{\boldsymbol{\theta}}}_{\text{parameters}}(\boldsymbol{x}, y; \mathcal{D}) \quad \left\{ \begin{array}{l} \text{Empirical risk} \\ \text{Structural risk} \\ \text{Likelihood} \\ \text{Margin} \\ \dots \end{array} \right.$$

desired properties in the learned model

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

Fairness is yet another desired property of the learned models

POLYTECHNIQUE MONTRÉAL UNIVERSITÉ D'INGÉNIERIE

IVADO

Mila

Université de Montréal

Learning subject to fairness constrains

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

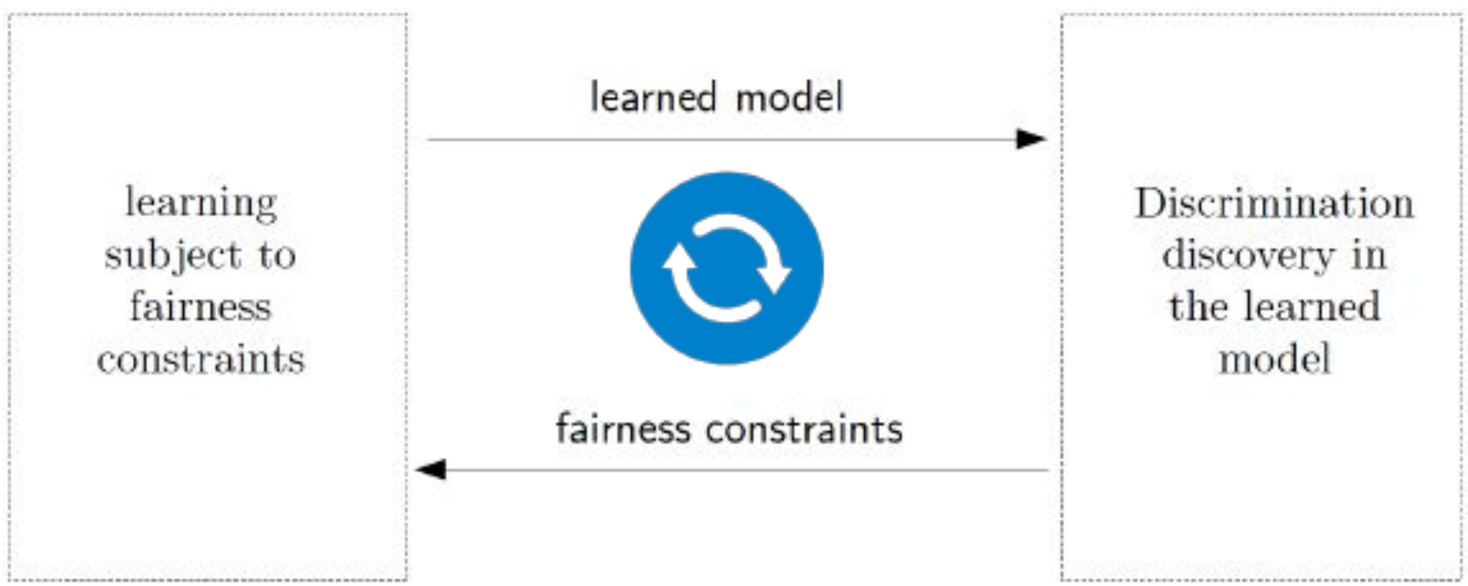
Adding fairness constraints can change these properties!

Discovering and eliminating discrimination

We propose a **signomial programming** approach to eliminate individual patterns of discrimination during maximum-likelihood learning.

Decision **Sensitive** **non-Sensitive**

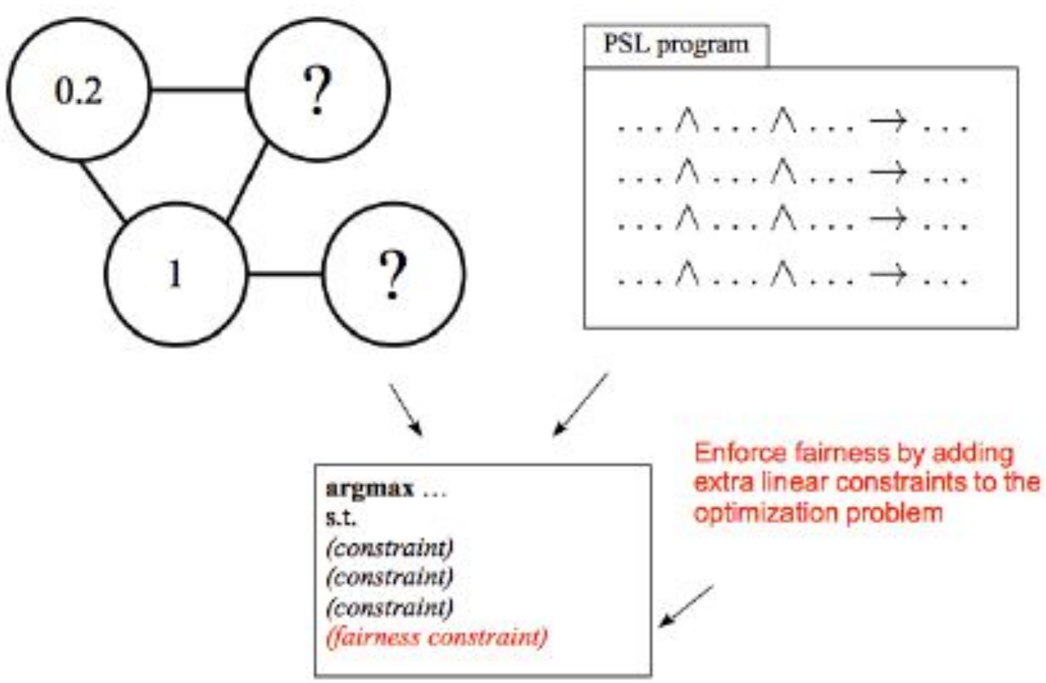
Degree of discrimination of XY: $\Delta_{P,d}(x,y) \triangleq P(d|xy) - P(d|y)$



Yoojung Choi, Golnoosh Farnadi, Behrouz Babaki, and Guy van den Broek. Learning Fair Naive Bayes Classifiers by Discovering and Eliminating Discrimination Patterns. https://arxiv.org/abs/1906.03843

Fairness in relational domains

- The existing literature on fairness in machine learning and data mining is almost exclusively limited to the non-relational setting.
- PSL is a probabilistic programming language for defining hinge-loss Markov random fields.
- We propose fair MAP inference for PSL

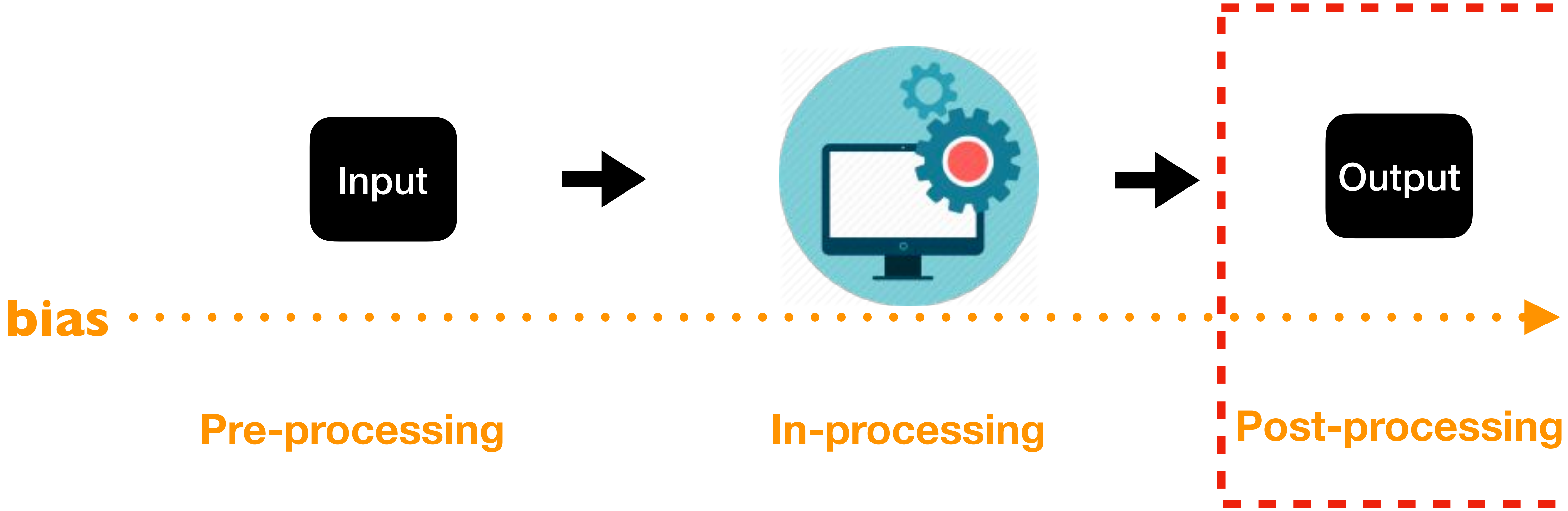


MAP inference in PSL can be stated as a **convex optimization problem**

Farnadi, Golnoosh, Behrouz Babaki, and Lise Getoor. "Fairness in relational domains." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2018.

POLYTECHNIQUE MONTRÉAL UNIVERSITÉ D'INGÉNIERIE IVADO Mila Université de Montréal

Fairness in Pro-Processing



Explaining the Output (black box)

Machine Learning based strategies rely on the fact that a decision rule can be learned using a set of observed labeled observations

Learning samples may present biases either due to the presence of a real but unwanted bias in the observations or due to data pre-processing.

Kim, Michael P., Amirata Ghorbani, and James Zou. "Multiaccuracy: Black-box post-processing for fairness in classification." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2019.

Fairness Verification

- We focus on formal verification of deep learning models

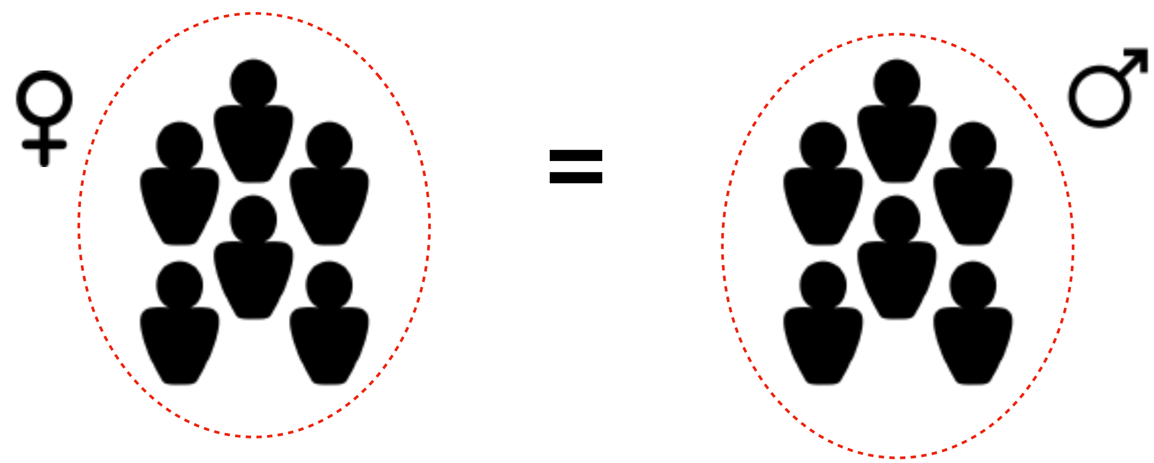
- Verification is an automated technique that can prove certain properties of a program, e.g., is there any input for which the decision-making algorithm has a certain property?

Hiring example

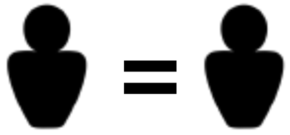
- Consider a company which bases its decisions about hiring an employee based on a vector of attributes of the applicant.
- The goal is to decide about hiring for each applicant using a neural network model.
- The administration needs to ensure that the gender of applicants has no influence on the decision.

Group fairness vs. Individual fairness

Hiring example: Group fairness metrics guarantees that on average the population of female applicants has the same opportunity of hiring as the male applicants.



Hiring example: our aim is to ensure that a female applicant has the same opportunity of hiring to a similar male applicant.



Individual Fairness Verification

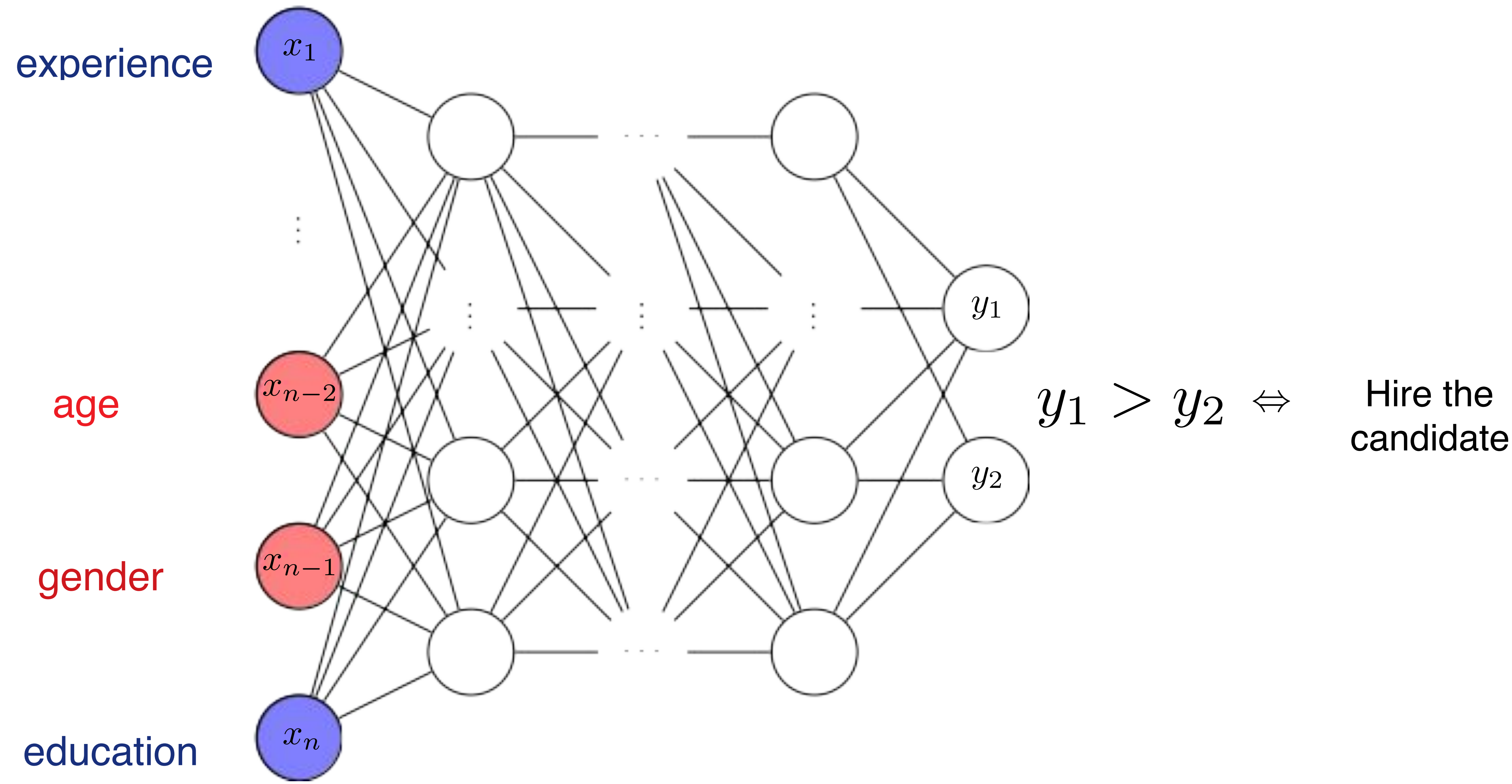
Is there any input for which the decision-making algorithm is unfair?

Input

Similar individuals

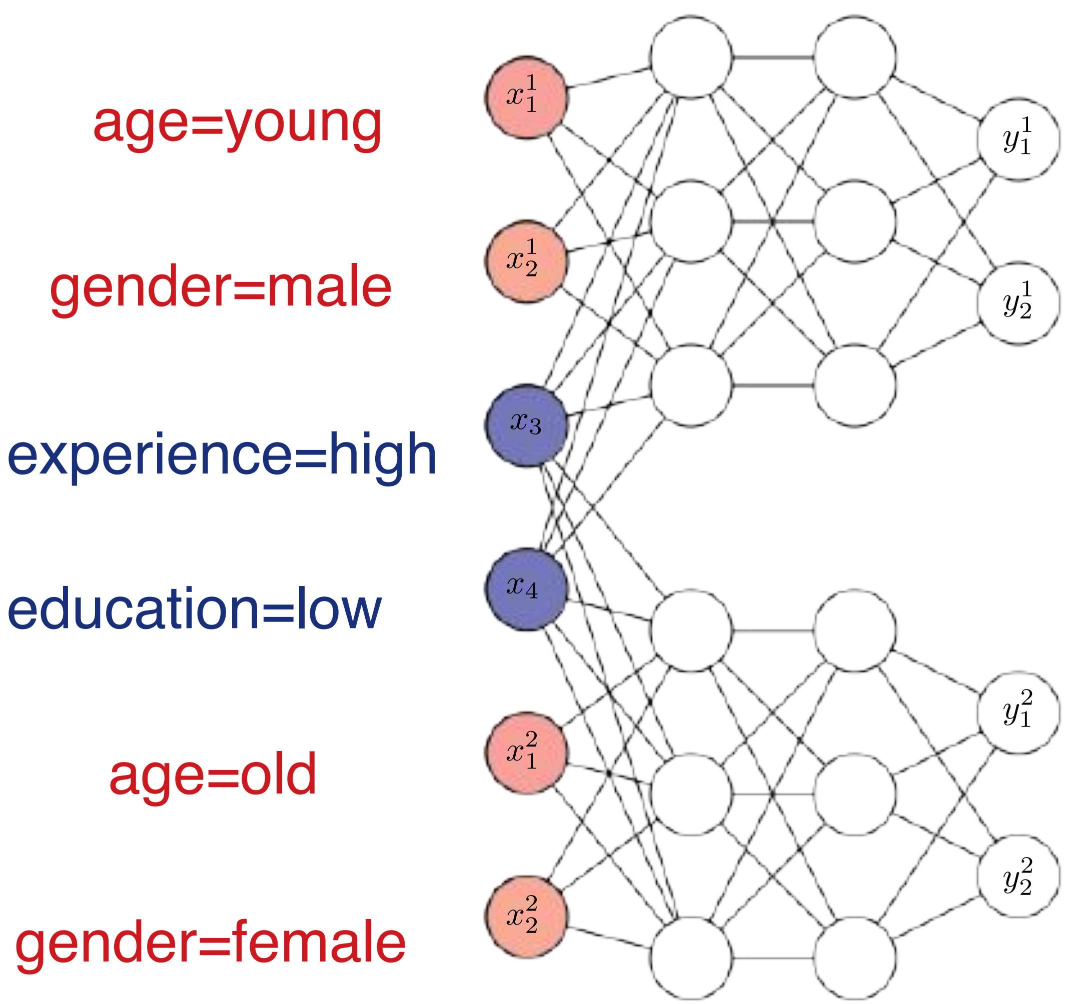
Deep network

Assume that some decision (e.g. hiring a candidate) is made using a neural network:



Deep Verification Network

Discrimination: The candidates only differ in their non-sensitive features, but are treated differently.



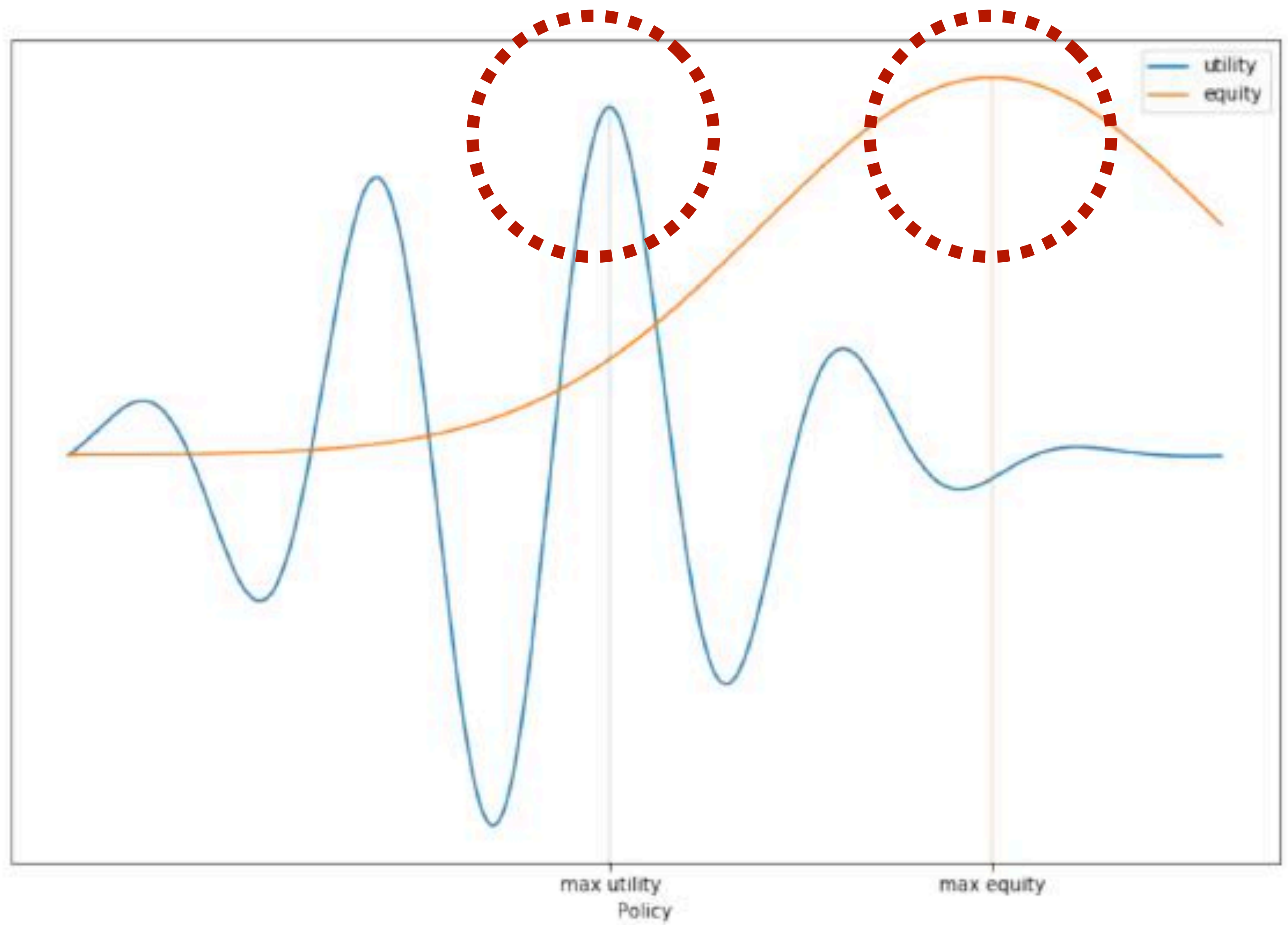
shared non-sensitive features

$y_1^1 > y_2^1 \Rightarrow$ Hire the candidate

$y_1^2 > y_2^2 \Rightarrow$ Do not Hire the candidate

Opportunities & Challenges

Opportunities: We cannot simultaneously maximize two objectives



Corbett-Davies, Sam, et al. "Algorithmic decision making and the cost of fairness." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.

POLYTECHNIQUE MONTRÉAL UNIVERSITÉ D'INGÉNIERIE

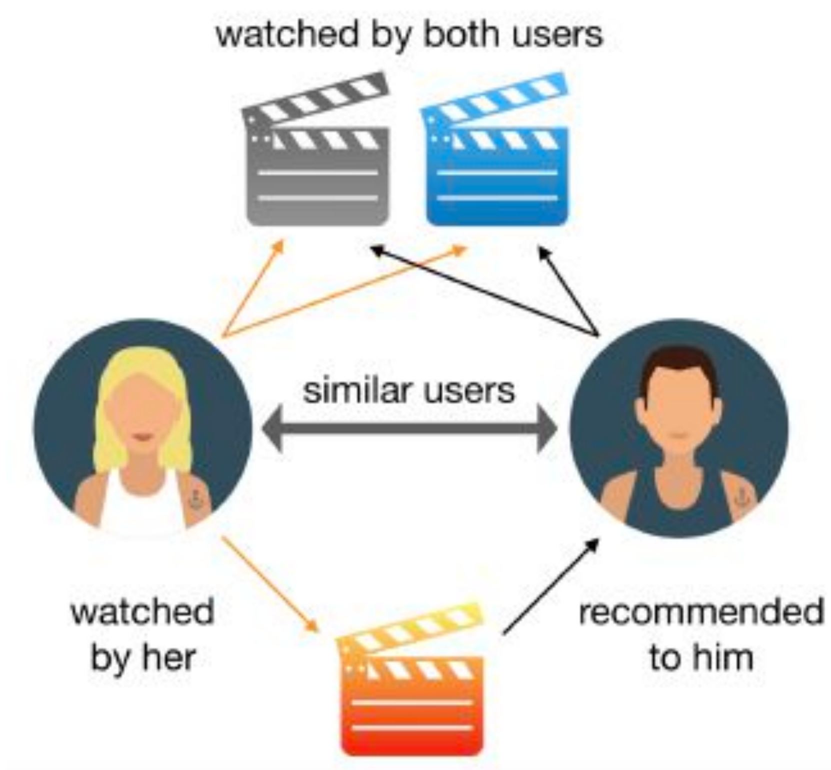
IVADO

Mila

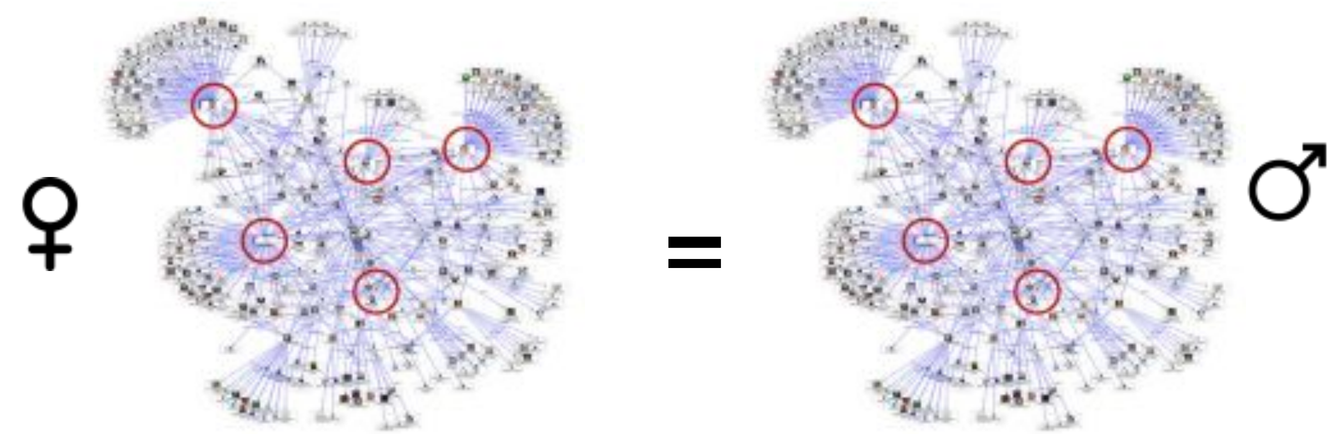
Université de Montréal

Opportunities: It is not the same in different fields!

Recommender systems



Influence maximization



Challenges: complexity of real word

- How to leverage the **complexity** of the real world in decision making?

Dwork, Cynthia, and Christina Ilvento. "Fairness under composition." *arXiv preprint arXiv: 1806.06122* (2018).

Chouldechova, Alexandra, and Aaron Roth. "The frontiers of fairness in machine learning." *arXiv preprint arXiv:1810.08810*(2018).

Challenges: sub-groups

- How to include **sub-groups** in fairness definitions?

Gender Classifier	Darker Subjects Accuracy	Lighter Subjects Accuracy	Error Rate Diff.
Microsoft	87.1%	99.3%	12.2%
FACE++	83.5%	95.3%	11.8%
IBM	77.6%	96.8%	19.2%

Kearns, Michael, et al. "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness." *arXiv preprint arXiv:1711.05144* (2017).

Challenges: The communication channel is not clear

- Is data transformation legal?
- Can algorithms be used in a real-world case law?
- How to define multi-disciplinary measures? e.g., to address differences between USA and EU regulation

Takeaways

Bias happens throughout the automated systems:

- Educate people about **discrimination**
- How to **define fairness** in your set-up?
- Ask who is **using** the model?
- What is **the purpose** of the system?

Any

Questions?

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