Data Bias and Algorithmic Discrimination

IFT6758 - Data Science

Sources:

Emre Kiciman tutorial on sources of data bias tutorial

Mila

Université de Montréal

Announcements

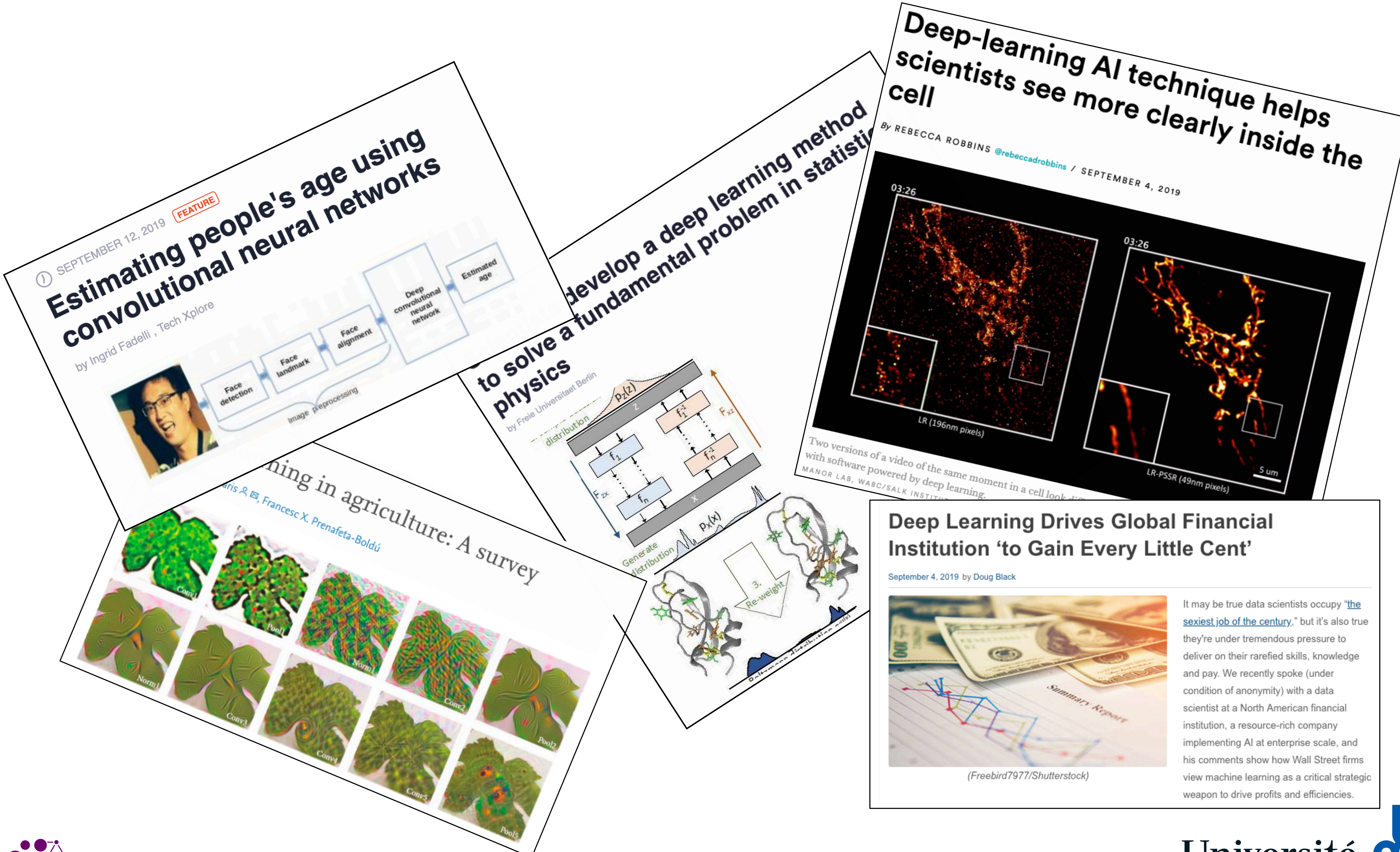
- ~100 students presented on Tuesday!

Winners of the tasks: (+5 bonus points)

Age prediction + Personality prediction:
User01

Gender prediction:
User02

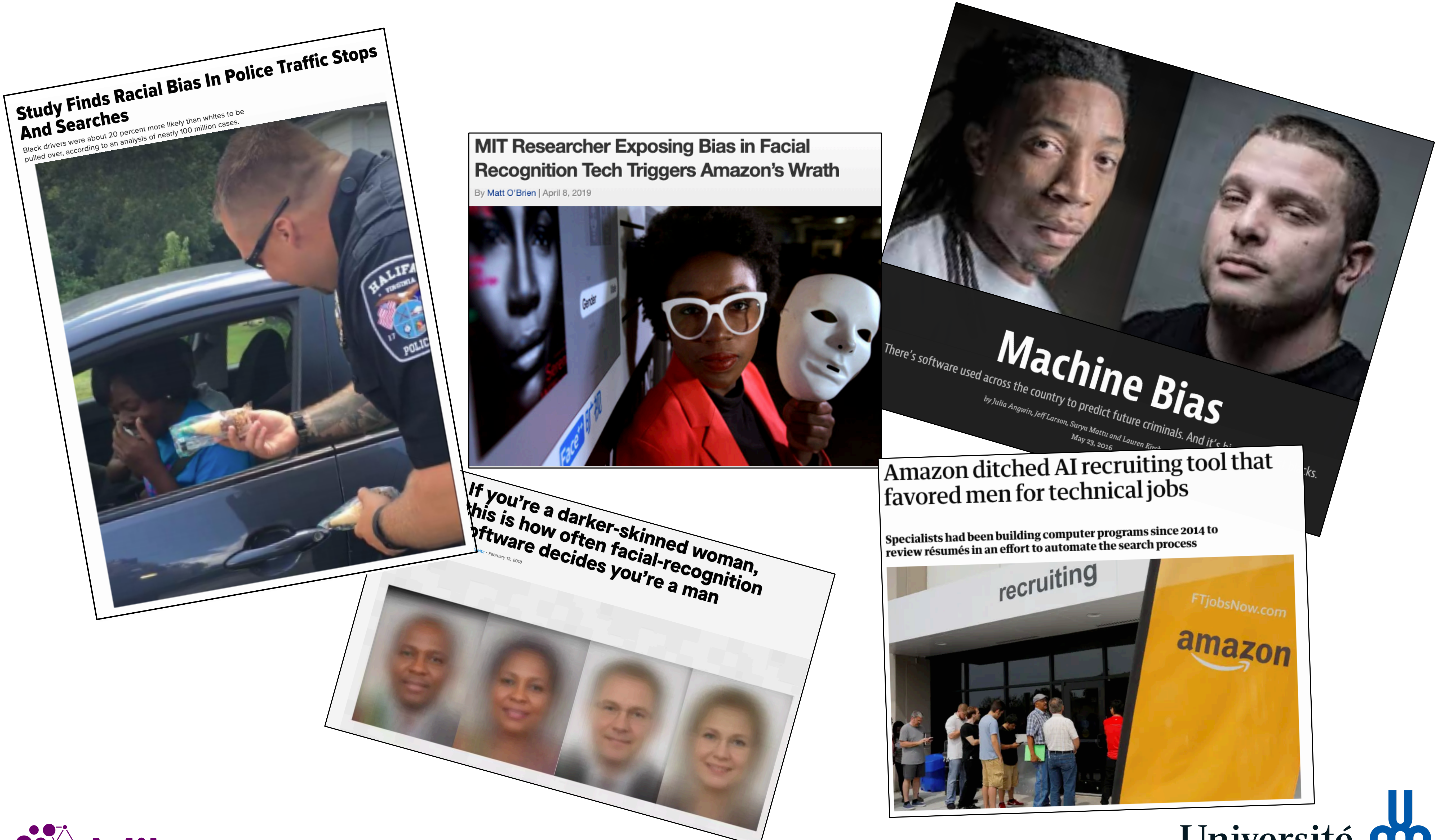
Machine learning is everywhere!



Mila

Université de Montréal

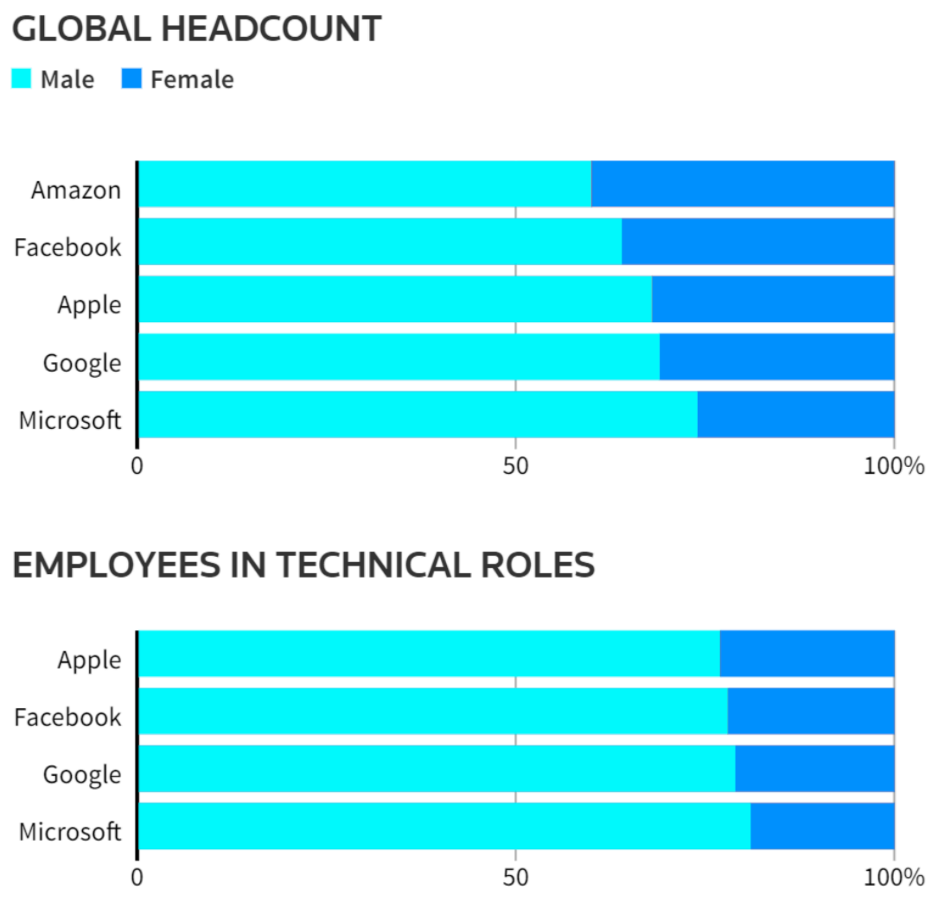
Does ML create more problems than it solves?

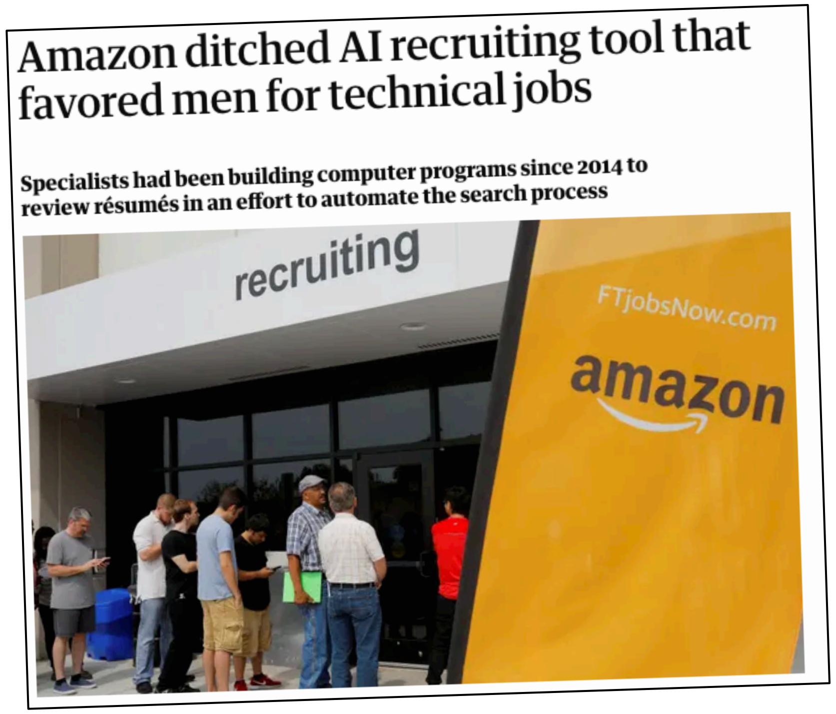


Mila

Université de Montréal

Amazon Recruitment Tool




Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women

Mila

Université de Montréal

Policing

- Investigative tools are AI-based models.
- Situational testing; natural experiments (e.g. observe other motorists in a stop zone to see if police stops blacks more than whites)

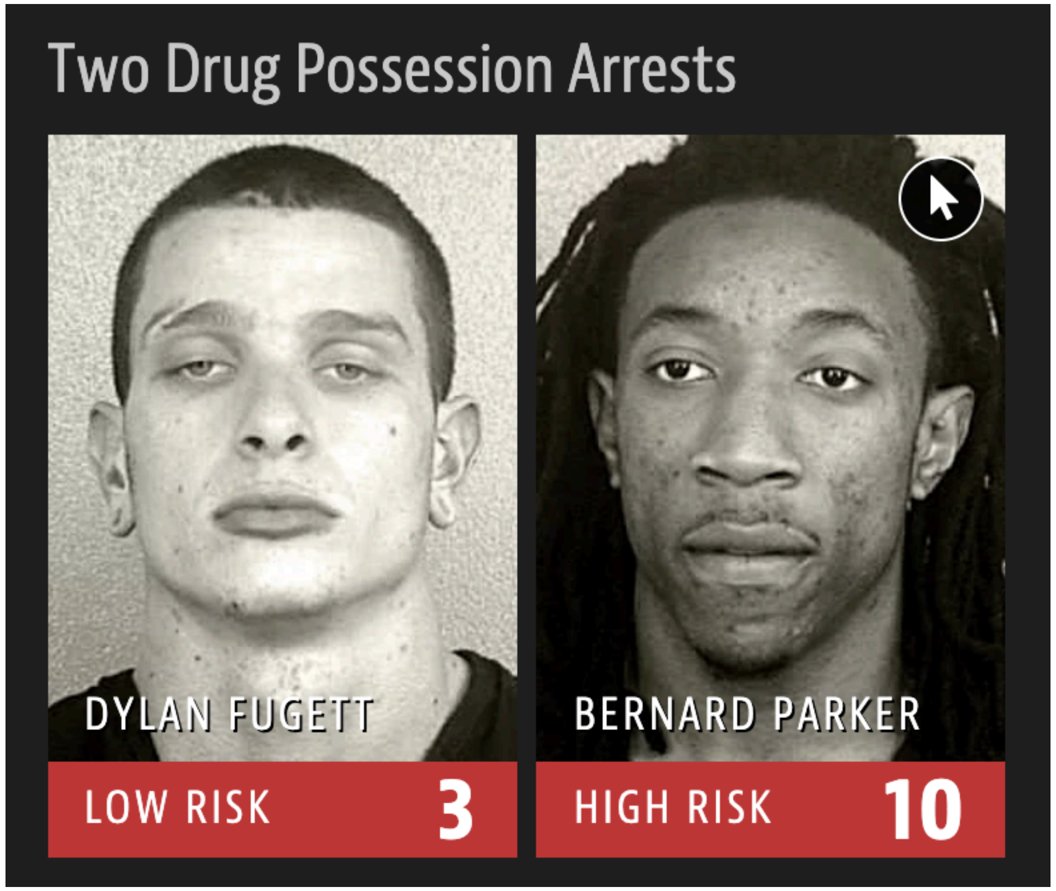

Study Finds Racial Bias In Police Traffic Stops And Searches

Black drivers were about 20 percent more likely than whites to be pulled over, according to an analysis of nearly 100 million cases.

A. Romei and S. Ruggieri (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29, pp 582-638

COMPAS

- The software used across US to predict future criminals is biased against blacks.



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

Mila

Université de Montréal

Gender-shades

- Let’s hear about if from Joy Buolamwini!

http://gendershades.org/

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?

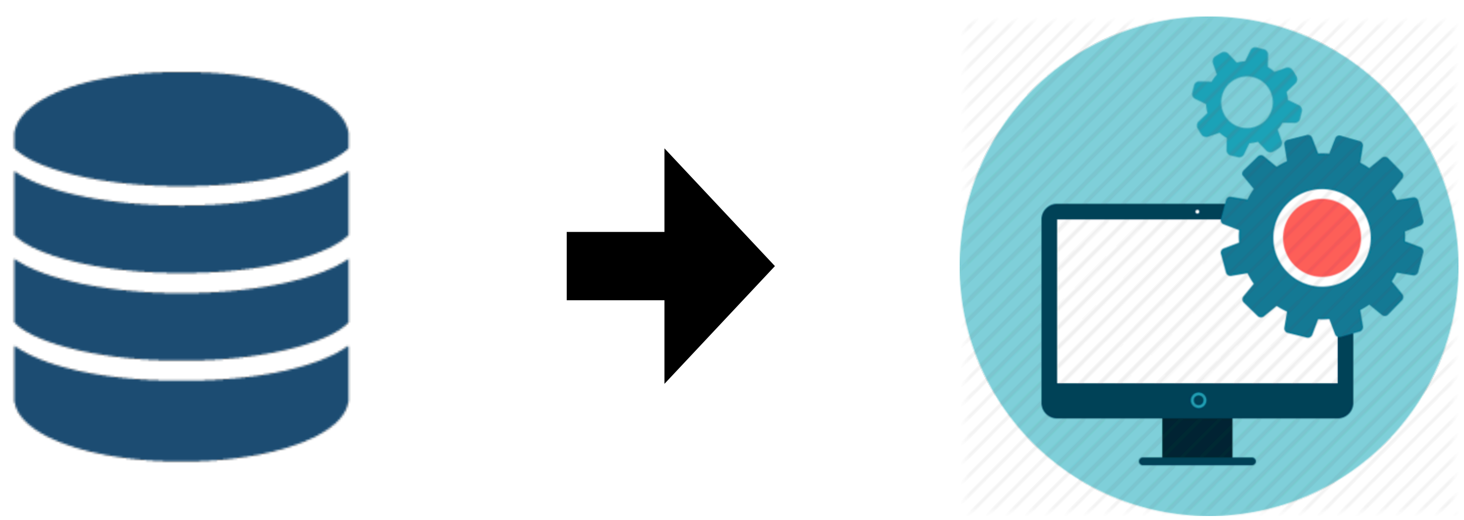
By Jonathan Rennie on 29 Aug 2019 in Artificial intelligence, General Data Protection Regulation, Data protection, Latest News

Mila

Université de Montréal

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

Discrimination due to unbalance data

They both apply for a loan
with a high amount

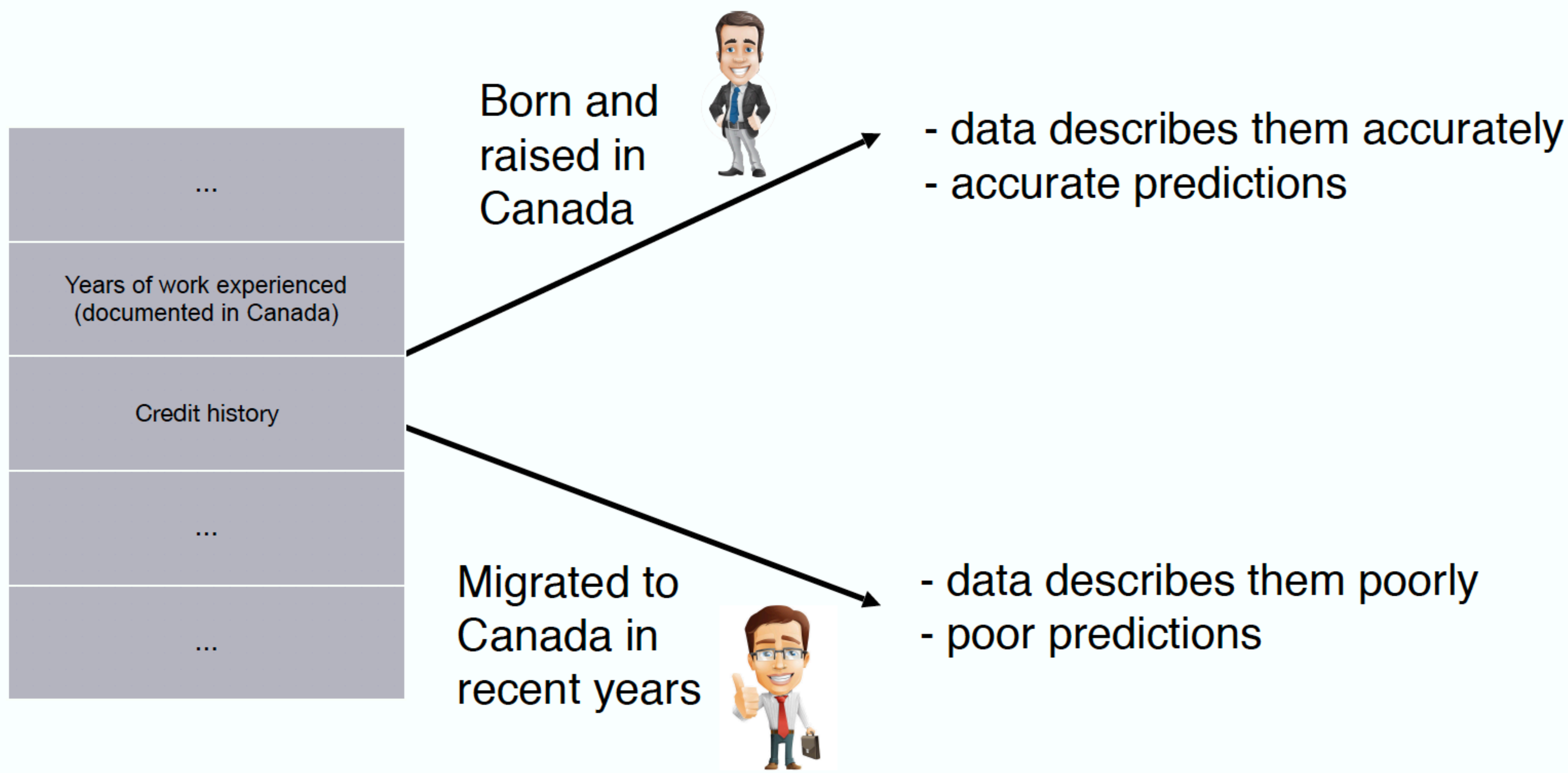
Lots of data about
similar (male) applicants

no data about similar
(female) applicants

Mila

Université de Montréal

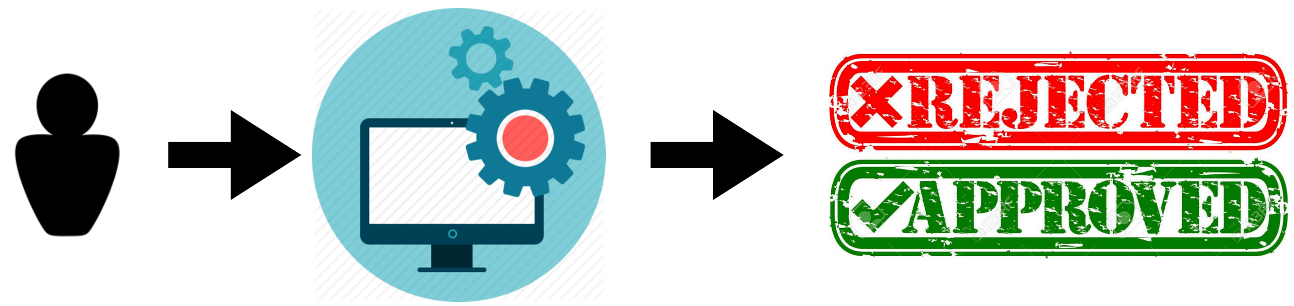
Discrimination due to missing attributes



Mila

Université de Montréal

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

Mila

Université de Montréal

Why we should care about fairness?

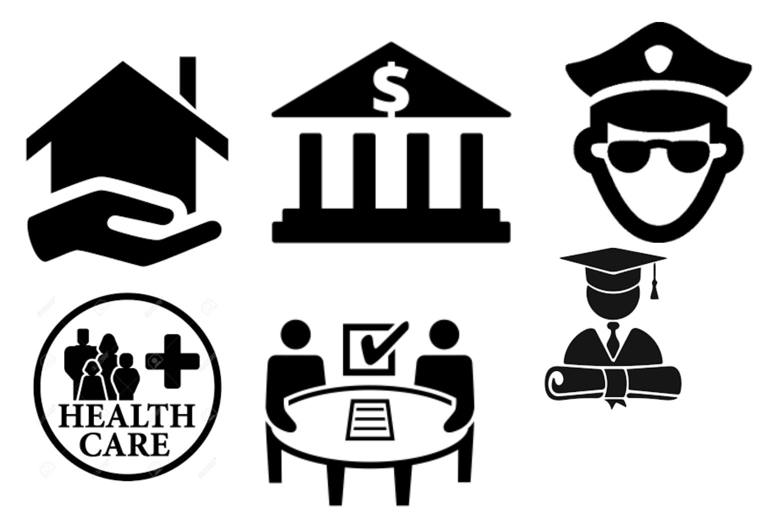
To address Law Against Discrimination!

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



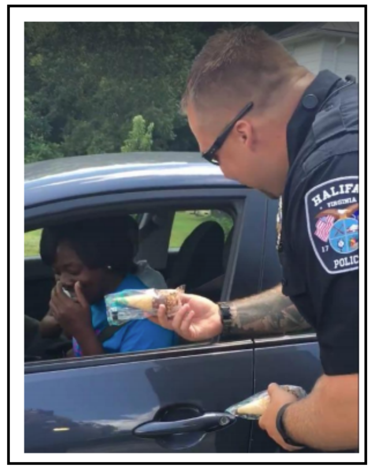
Mila

Université de Montréal

Fairness in ML

2014

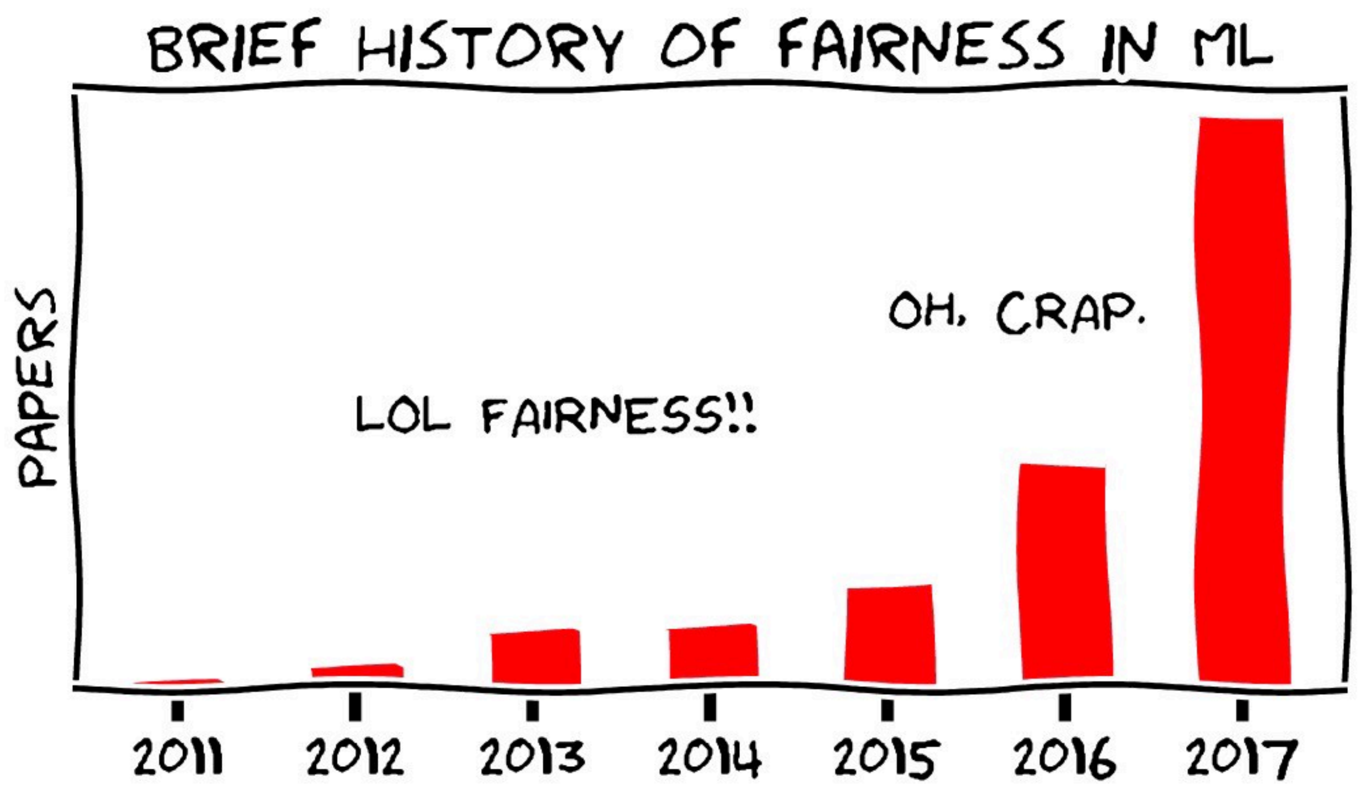
2015



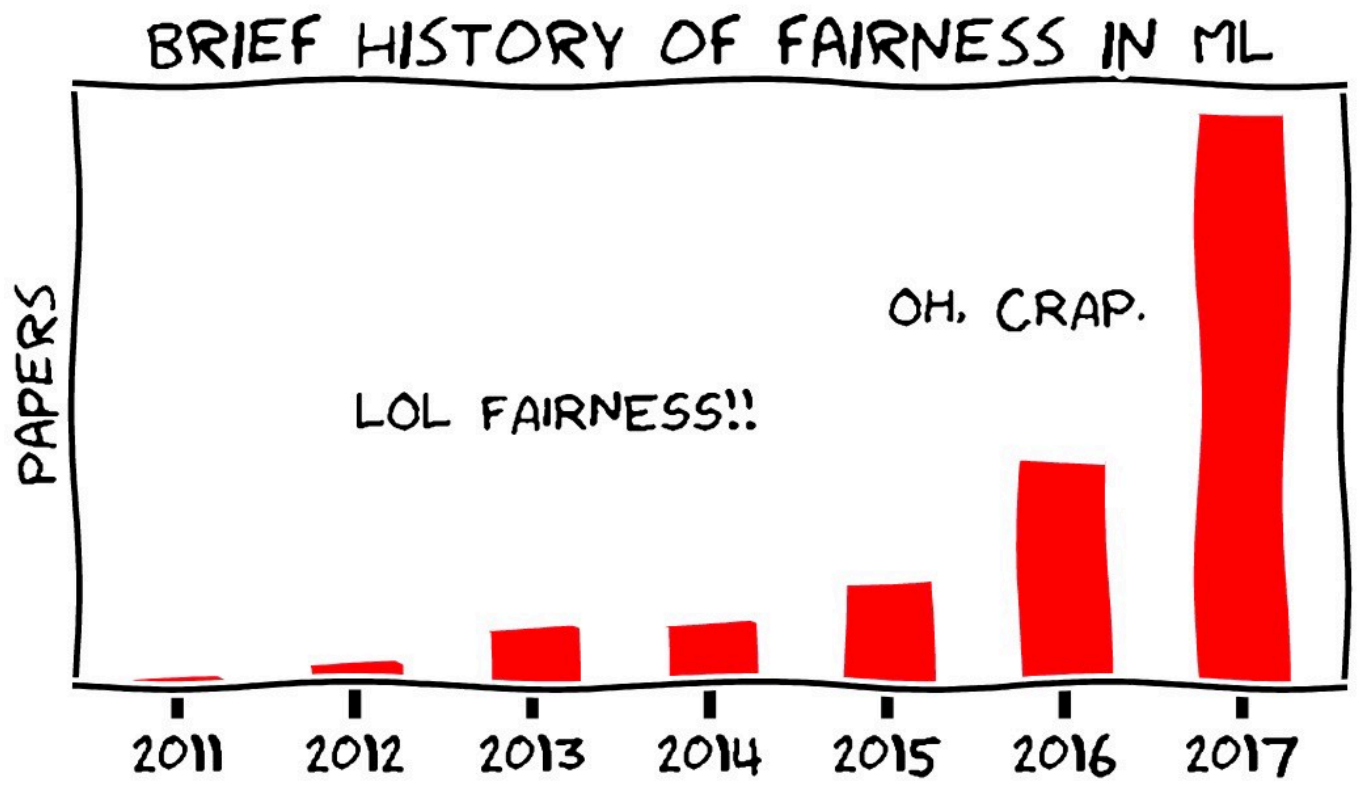
2016

2017

...



Fairness in ML

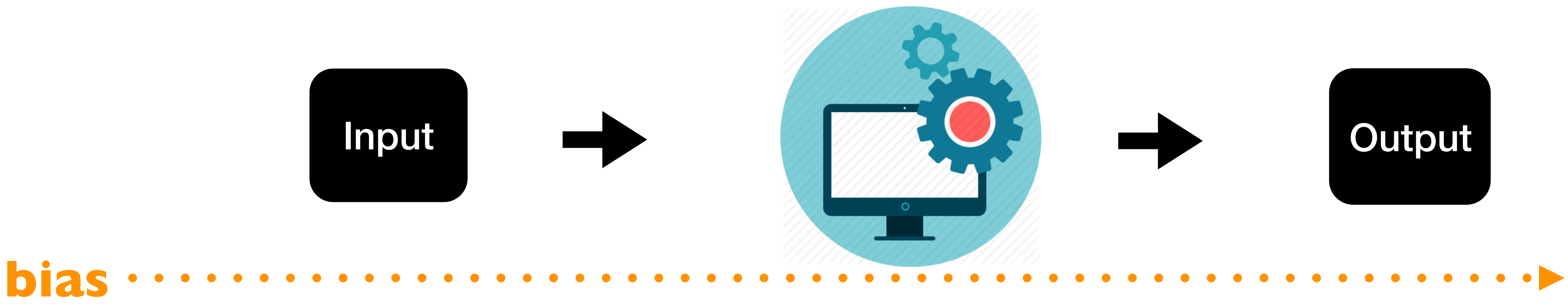


- "What is fair have been introduced in multiple disciplines for well over 50 years, including in education, hiring, and machine learning" [1].

- 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?



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

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

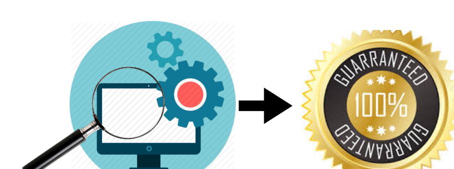
e.g.,

Learning subject to constraints
Ranking
Inference

Post-processing

e.g.,

Causal discovery
Transparency & Interpretability
Verification



Mila

Université de Montréal

Why do we use fairness definitions?

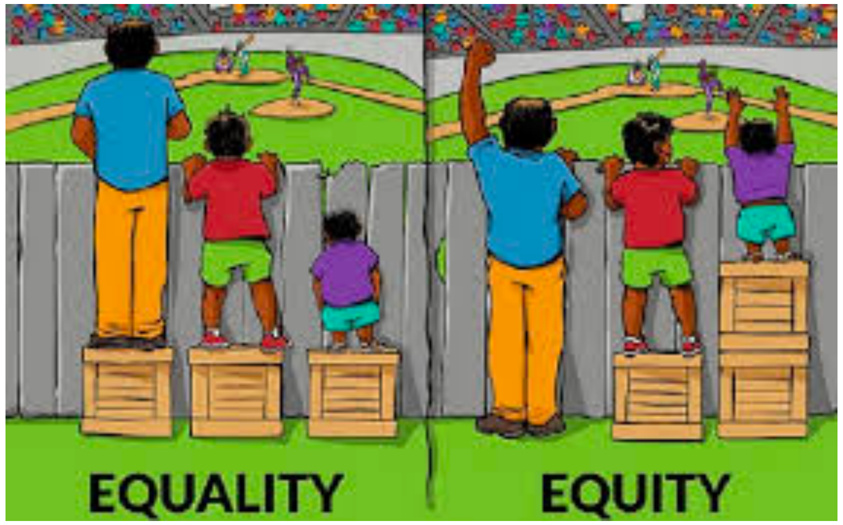
- To make algorithmic systems support human values!
- To identify strengths and weakness of the system
- To track improvement over time

To address Law Against Discrimination!

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

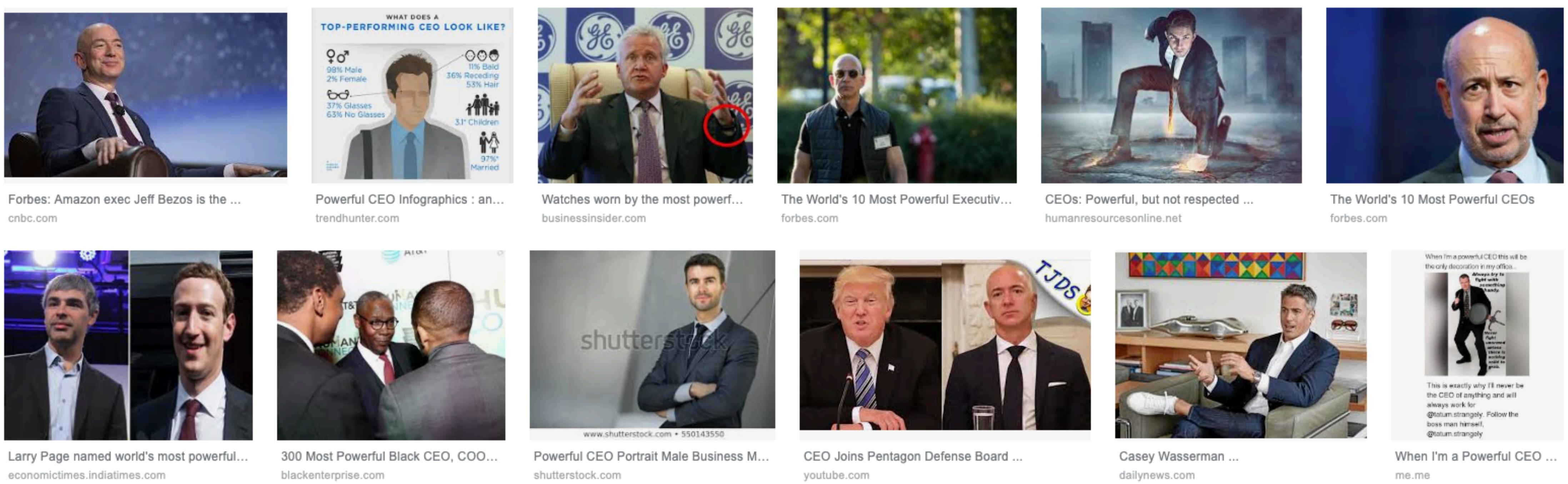
Gender-biased

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

Mila

Université de Montréal

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

Mila

Université de Montréal

Different types of fairness definitions

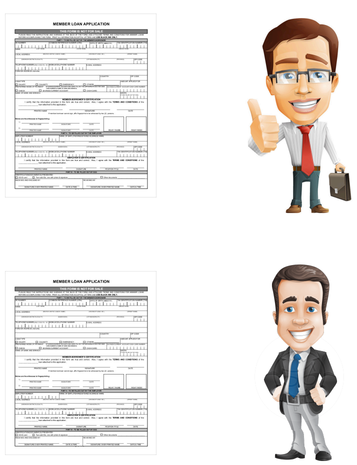
Types of fairness definitions

Different definitions based on **legal concepts**

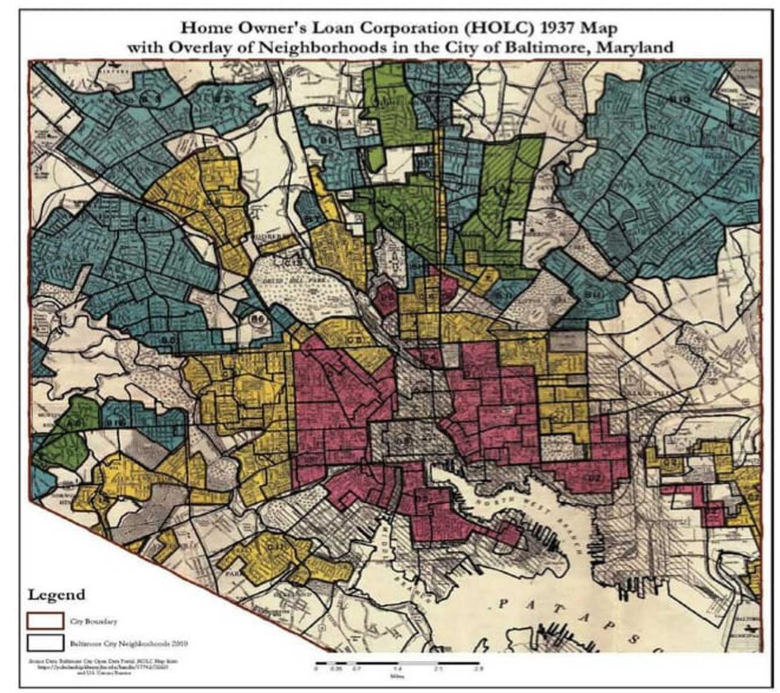
- Direct vs indirect discrimination
- Individual vs group fairness
- Explainable vs unexplainable discrimination

Indirect discrimination

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



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



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

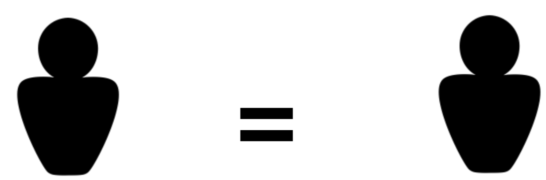
Definition	Citation #
Group fairness or statistical parity	208
Conditional statistical parity	29
Predictive parity	57
False positive error rate balance	57
False negative error rate balance	57
Equalised odds	106
Conditional use accuracy equality	18
Overall accuracy equality	18
Treatment equality	18
Test-fairness 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.

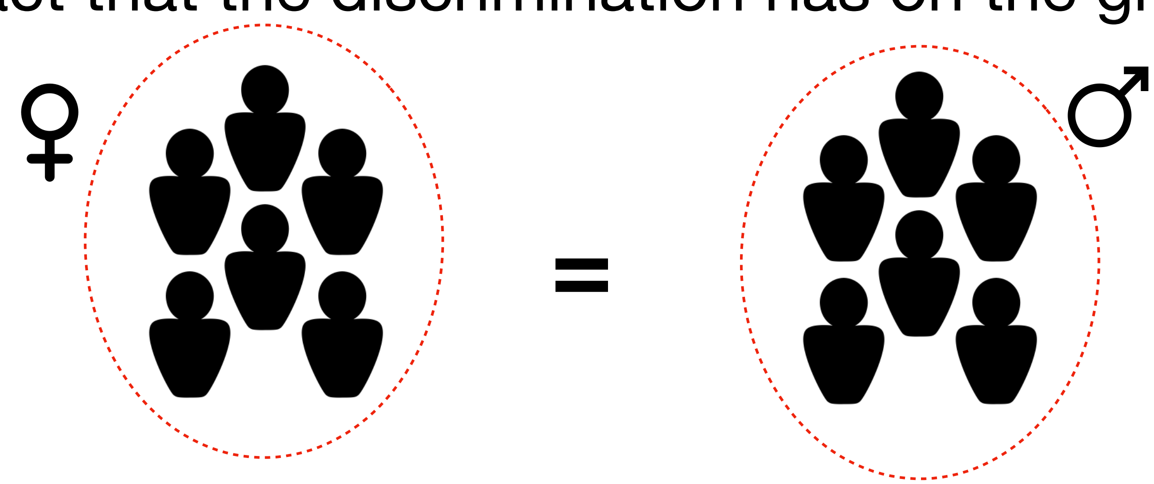
Types of fairness definitions

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.



Mila

Université de Montréal

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.

Recall

1. Positive predictive value (PPV)

$$p(Y = 1|d = 1)$$

2. False discovery rate (FDR)

$$p(Y = 0|d = 1)$$

3. False omission rate (FOR)

$$p(Y = 1|d = 0)$$

4. Negative predictive value (NPV)

$$p(Y = 0|d = 0)$$

d **Prediction decision**

Y **Actual Outcome**

Confusion Matrix

	Y=1	Y=0
d=1	TP	FP
d=0	FN	TN

- True positive (TP)
- False positive(FP)
- True negative (TN)
- False negative (FN)

Recall

d

Prediction decision

Y

Actual Outcome

5. True positive rate (TPR)

$$p(d=1|Y=1)$$

6. False positive rate (FPR)

$$p(d=1|Y=0)$$

7. False negative rate (FNR)

$$p(d=0|Y=1)$$

8. True negative rate (TNR)

$$p(d=0|Y=0)$$

Confusion Matrix

	Y=1	**Y=0**
d=1	TP	FP
d=0	FN	TN

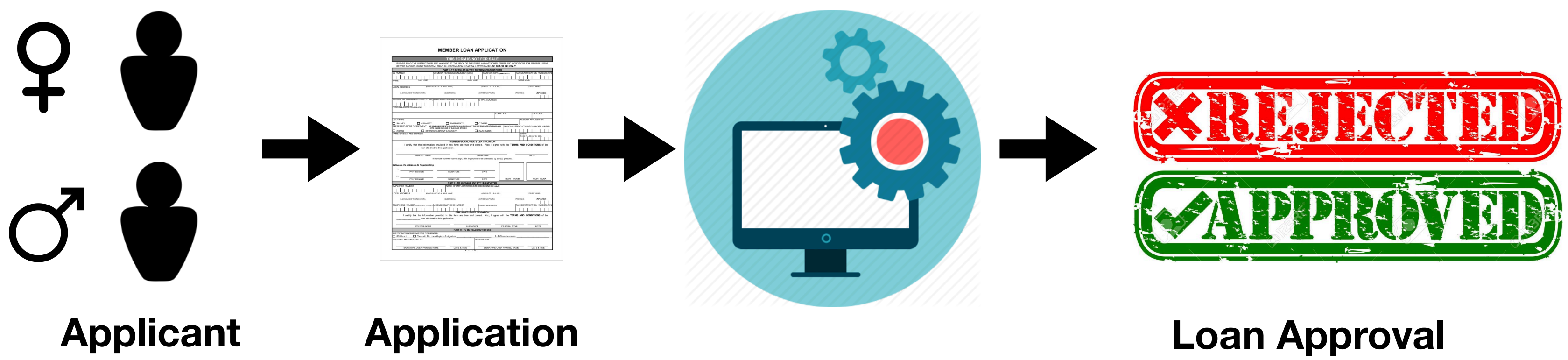
- True positive (TP)
- False positive(FP)
- True negative (TN)
- False negative (FN)

Differences of fairness definitions (mathematical notations)

Notations

TN	FP
FN	TP

confusion matrix



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

Female $G = f$ $\qquad d = 1$

Male $G = m$

Mila

Université de Montréal

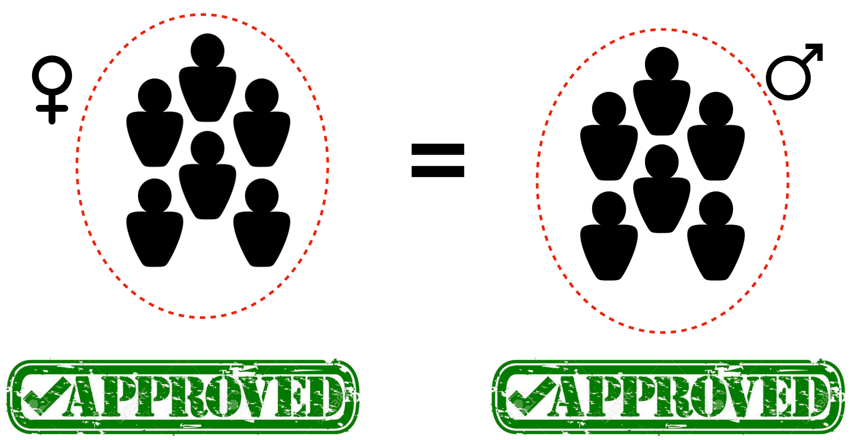
Group fairness

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



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

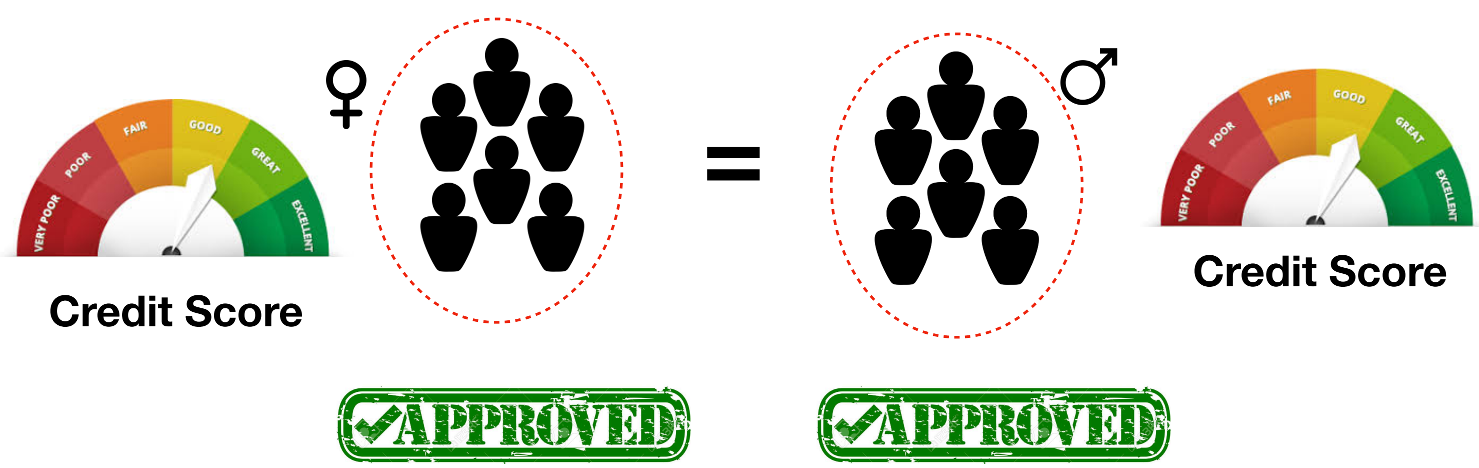
2- **Conditional statistical parity**

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

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**

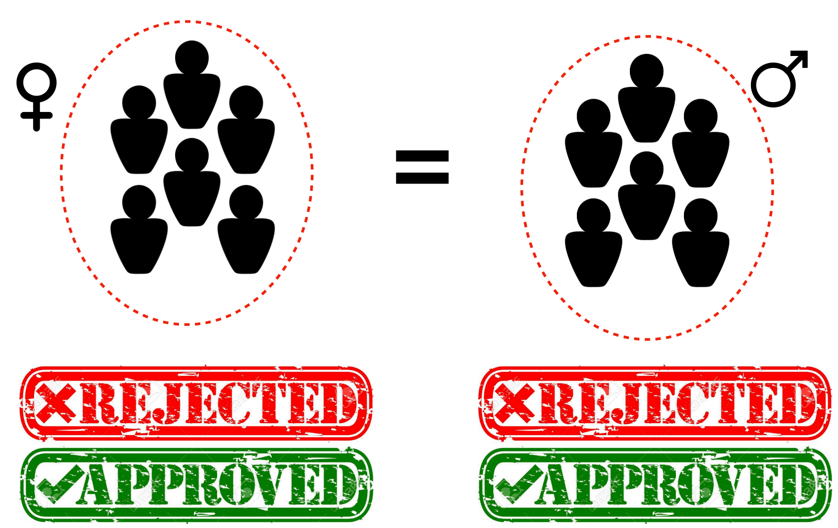
$$p(d=0|Y=1,G=f) = p(d=0|Y=1,G=m)$$

$$=$$

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

classifier should give similar results to applicants of both genders with actual positive loan approval.

Positive Loan Approval



Positive Loan Approval

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

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 loan 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).

Mila

Université de Montréal

Group fairness

a predicted outcome+ Actual outcome

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

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

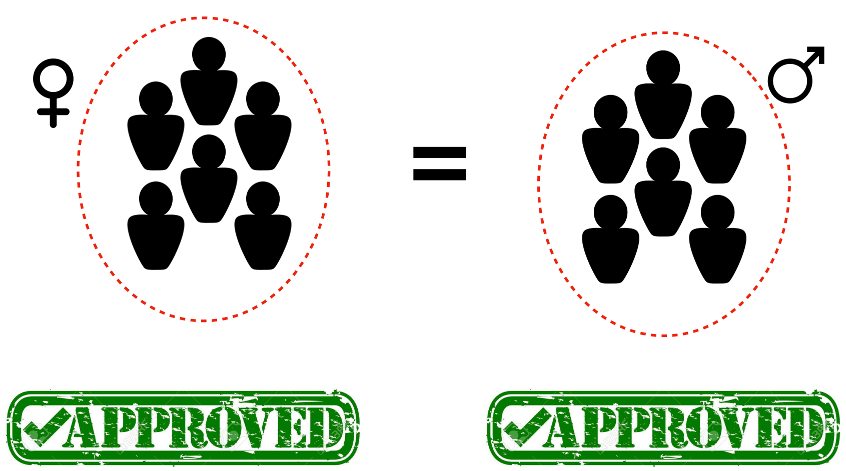
where $I \in \{0, 1\}$

Positive Credit Approval

applicants with a rejected loan application and applicants with an accepted loan application should have a similar classification, regardless of their gender.

Positive Loan Approval

Negative Loan Approval



Positive Loan Approval

Negative Loan Approval

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).

Mila

Université de Montréal

Group fairness

a predicted outcome+ Actual outcome

5. **Predictive parity** / outcome test

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

$$=$$

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

the fraction of correct positive loan approval should be the same for both genders

6. False positive error rate balance / **predictive equality**

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

$$=$$

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

a classifier should give similar results for applicants of both genders with actual rejected loans.

Mila

Université de Montréal

Group fairness

the predicted probability + Actual outcome

1. Test-fairness / **calibration** / matching conditional frequencies

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

for any given predicted probability score s in [0, 1], the probability of receiving a loan should be equal for both gender

2. **Well-calibration**

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

if a classifier states that a set of applicants have a certain probability s of receiving a loan then approximately s percent of these applicants should indeed have an approved loan.

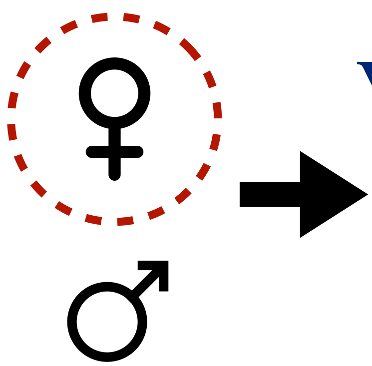
Mila

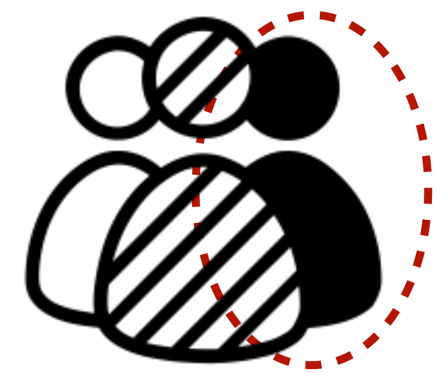
Université de Montréal

Individual fairness

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

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

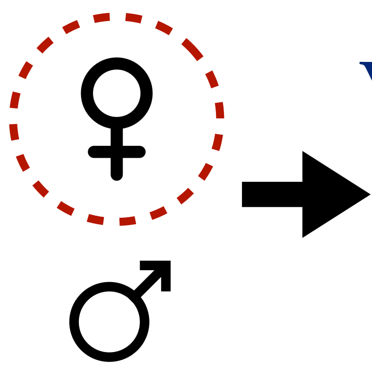




Individual fairness

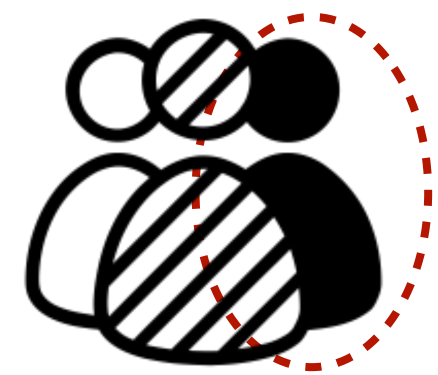
1- Fairness through unawareness, **Fairness through blindness**

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



WELLESLEY

1867 HOWARD UNIVERSITY



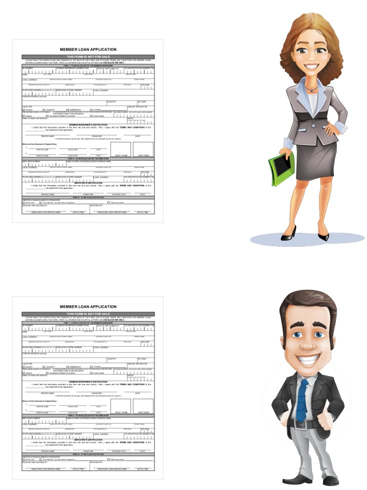
This can be impossible to hold because of non-obvious encoding in terms of many features, learned from the data

Individual fairness

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 M(x), M(y)

D

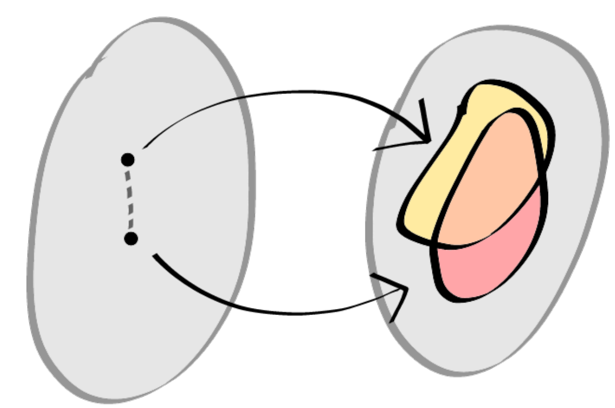
Distance metric Between two individuals x,y

k

similar individuals should have similar classification

seemingly different individuals

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



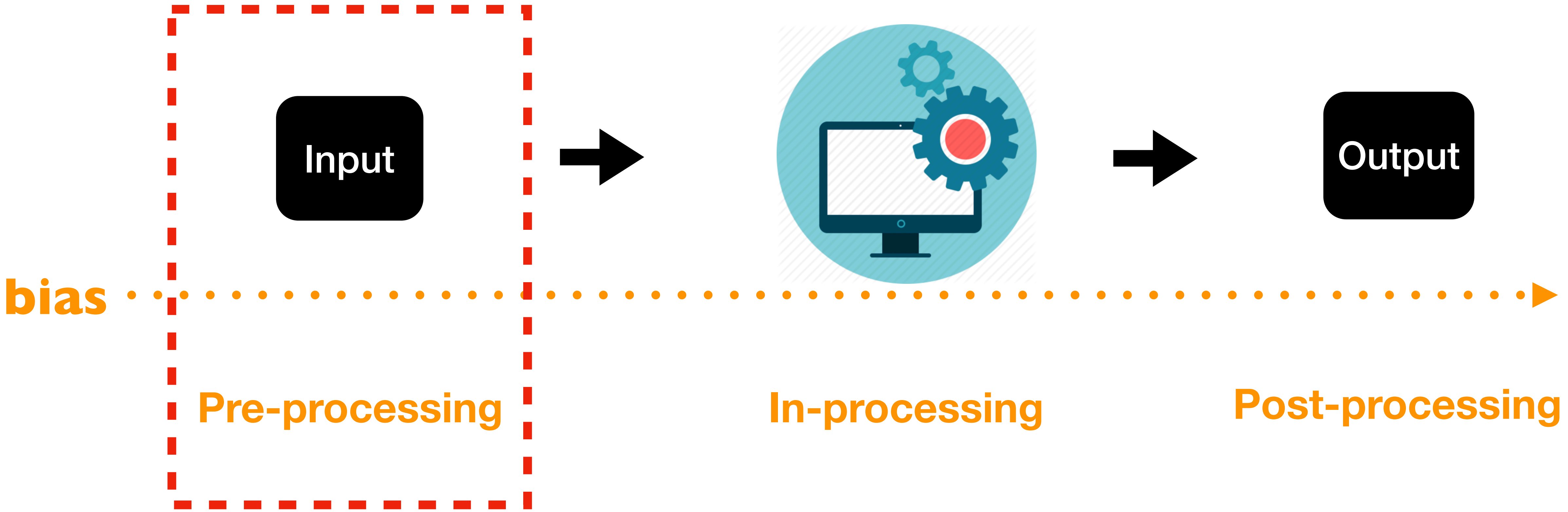
Dwork, Cynthia, et al. "Fairness through awareness." *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 2012.

Mila

Université de Montréal

Fairness in Machine Learning
(a few examples)

Fairness in Pre-Processing



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.

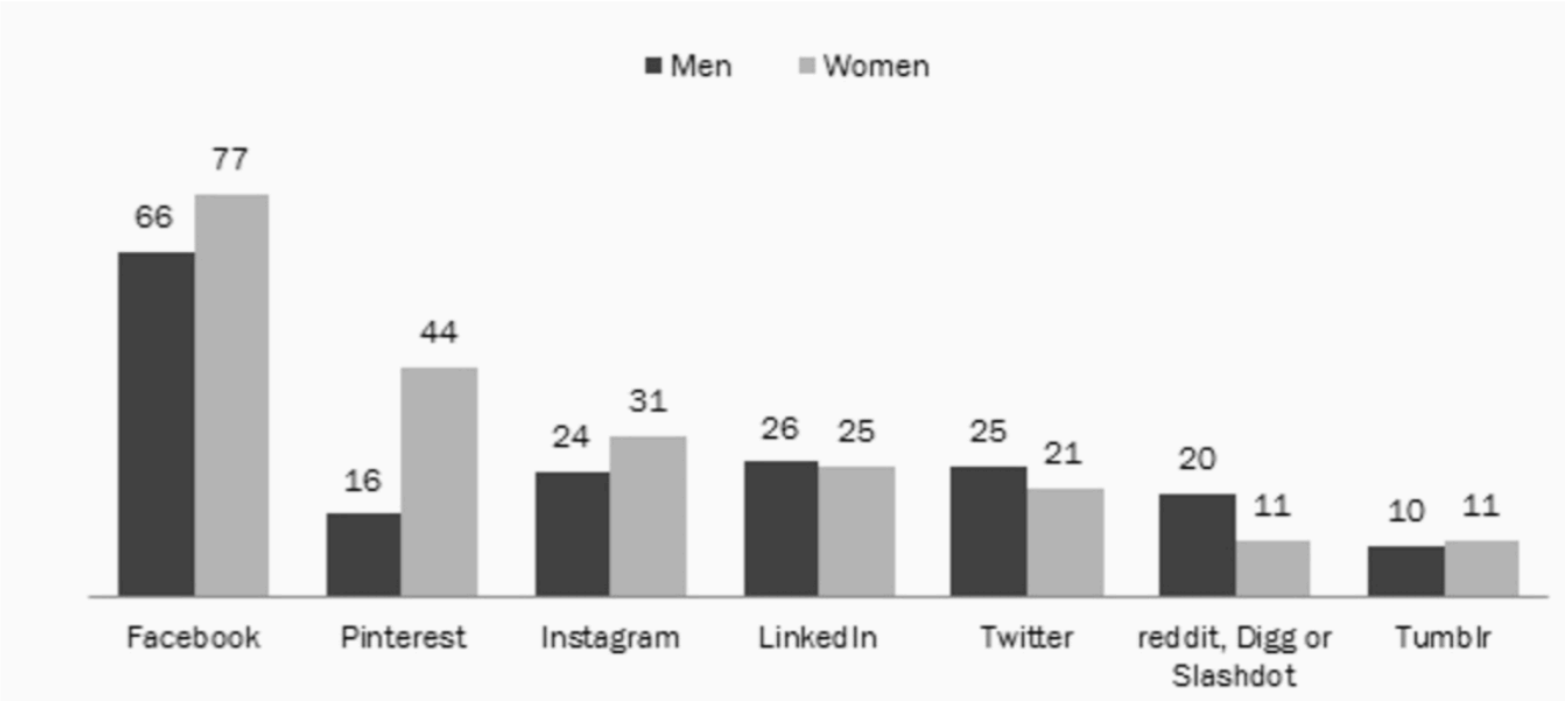
Where the data bias comes from?

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

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

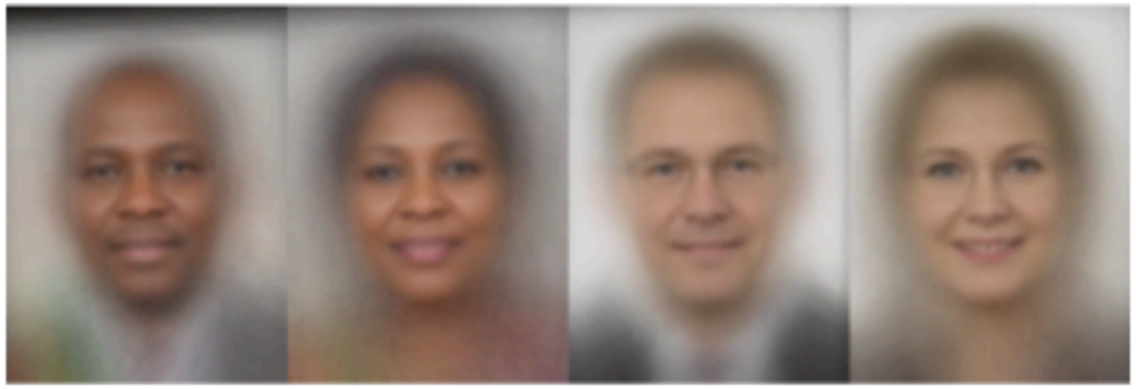
Systematic distortions must be evaluated in a task dependent way

E.g., for many tasks, populations should **match target population**, to improve external validity

But for other tasks, subpopulations require approximately **equal representation** to achieve task parity

Gender Shades

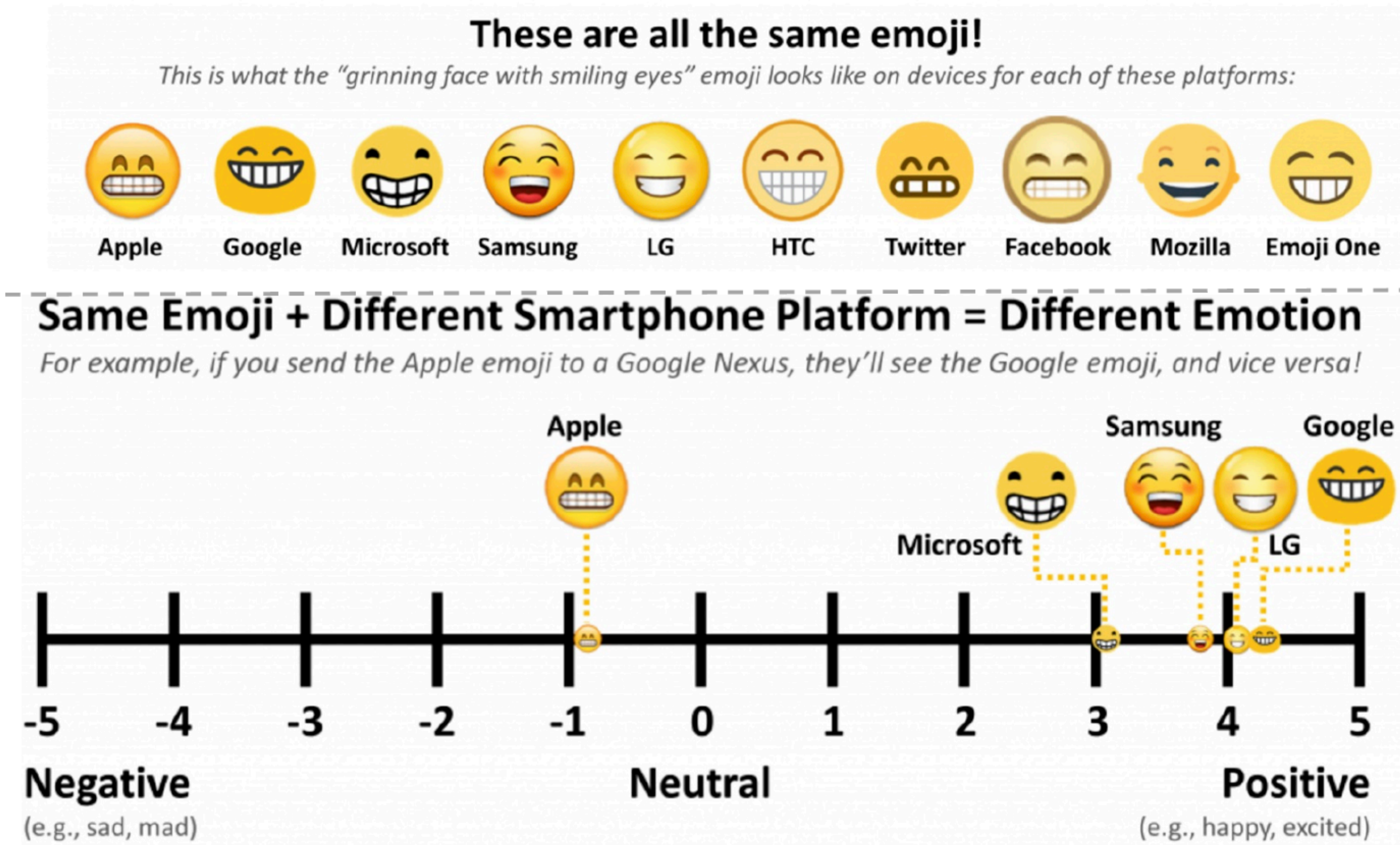
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



http://gendershades.org/

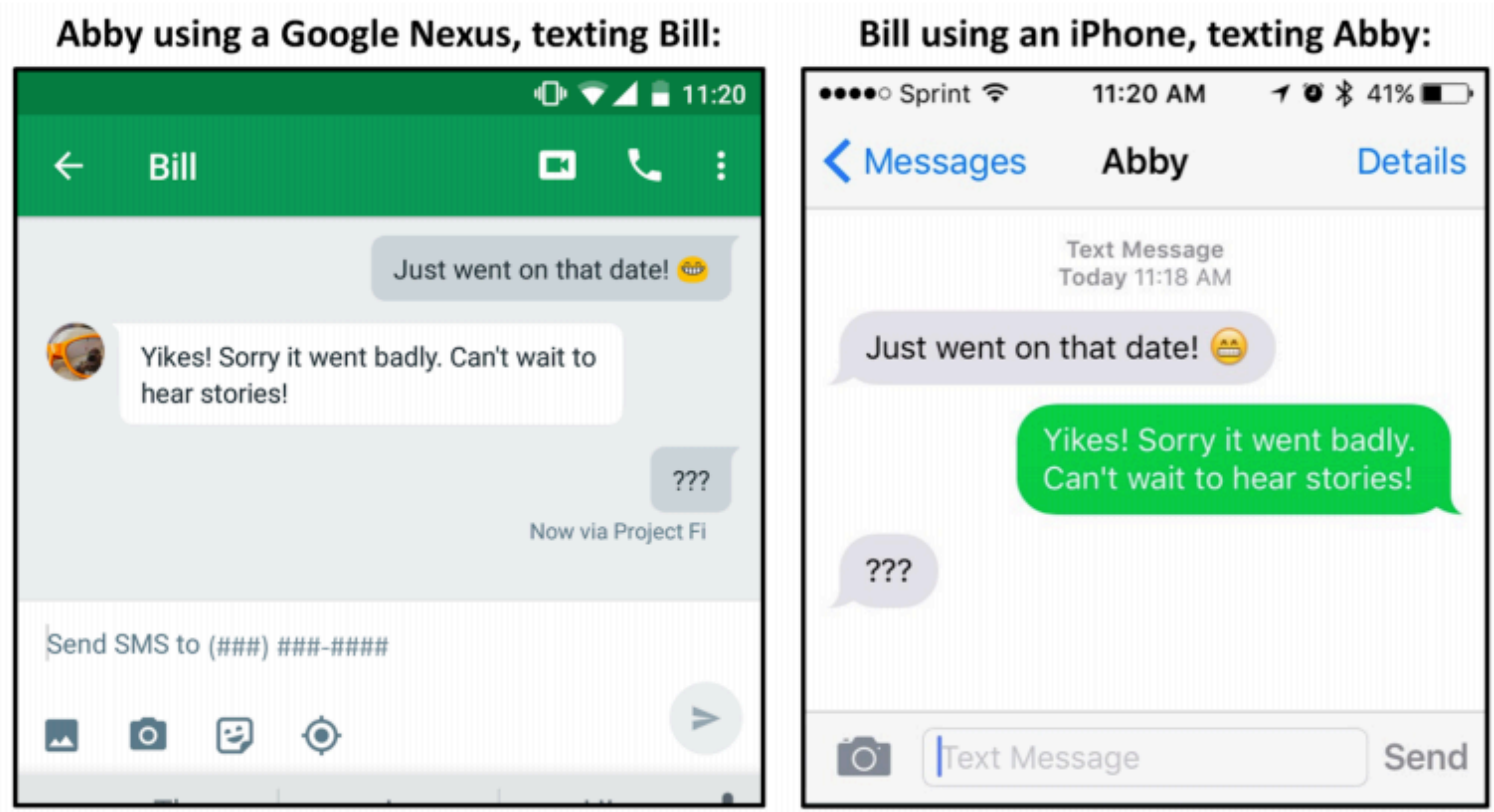
Where the data bias comes from?

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

Behavioural biases



[Miller et al. ICWSM'16]

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

Mila

Université de Montréal

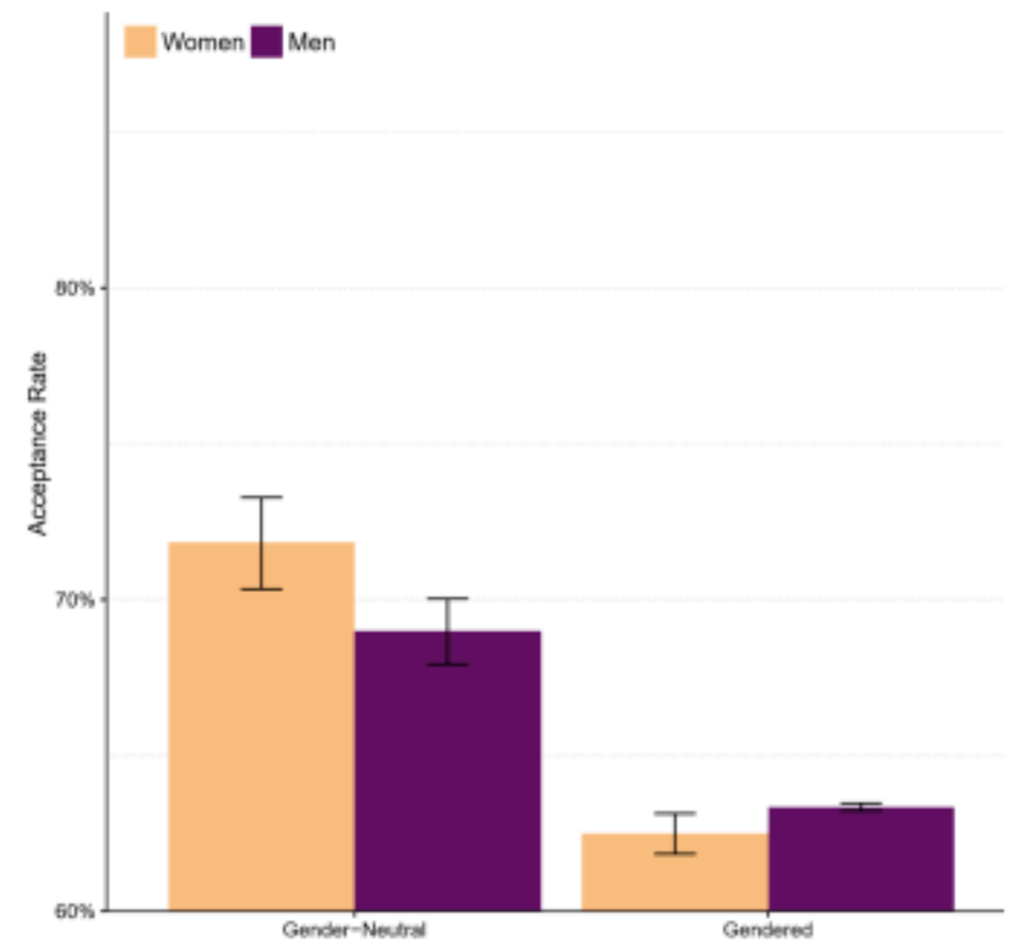
Behavioural biases

Cultural elements and social contexts are reflected in social datasets

The way users are perceived affects their interaction patterns (e.g., more or less content sharing/ followers).

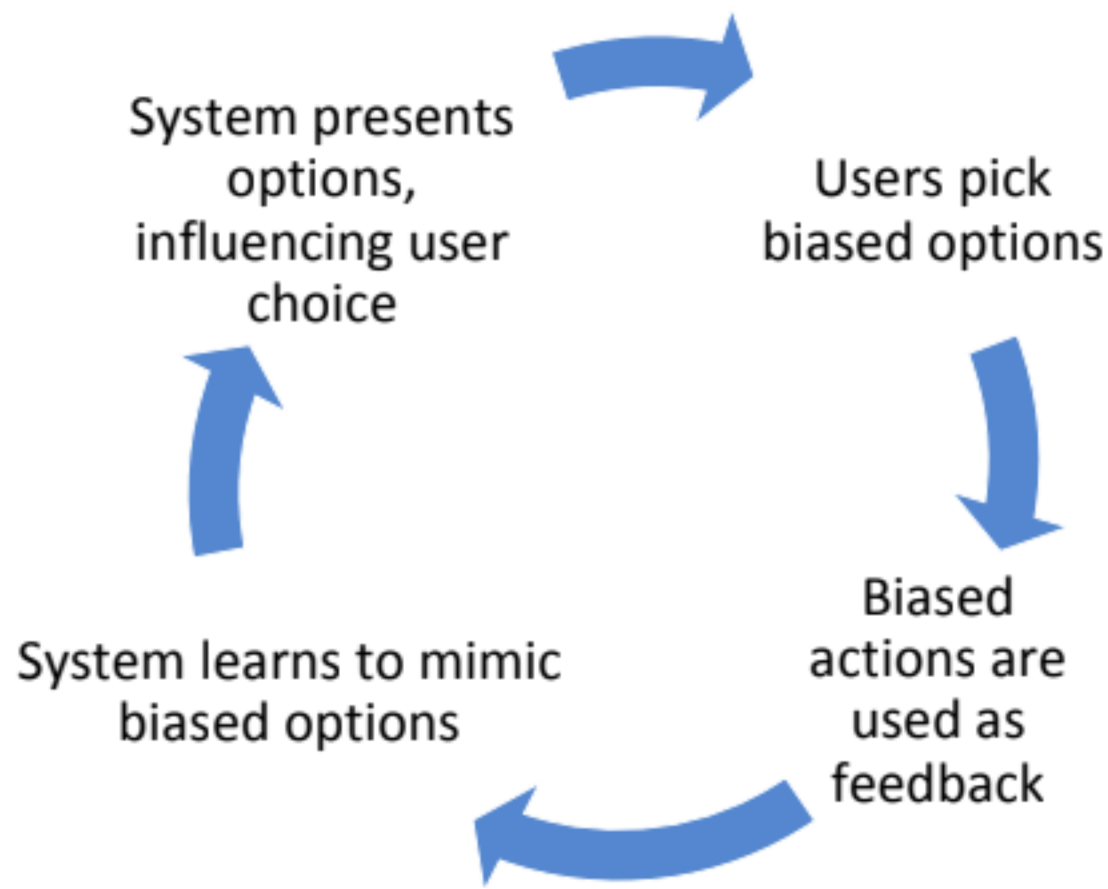
Women's code changes are more likely to be accepted in Github, unless they are identified as women

Figure from [Terrel et al., pre-print]



Behavioural biases

Societal biases embedded in behavior can be amplified by algorithms



Behavioural biases

Autocomplete for Search Interfaces

See also: **Seth Stephens-Davidowitz**. *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are* (2017)

Mila

Université de Montréal

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

Mila

Université de Montréal

Content production biases

What about facebook?

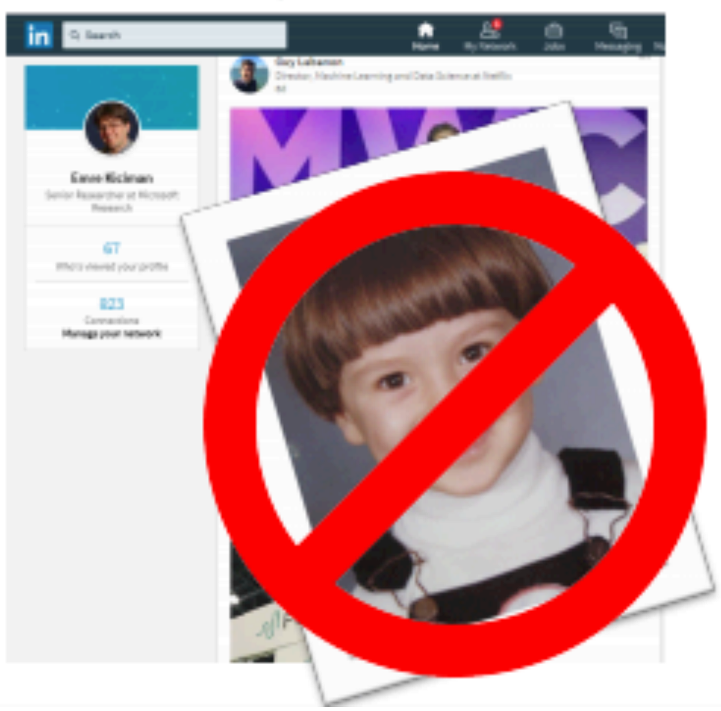
Variable	Females ρ	Males ρ
Style		
Capitalized words	-0.281**	-0.453**
Alph. lengthening	-0.416**	-0.324**
Intensifiers	-0.308**	-0.381**
LIWC-prepositions	0.577**	0.486**
Word length	0.630**	0.660**
Tweet length	0.703**	0.706**
References		
I	-0.518**	-0.481**
You	-0.417**	-0.464**
We	0.312**	0.266**
Other	-0.072	-0.148**
Conversation		
Replies	0.304**	0.026
Sharing		
Retweets	-0.101*	-0.099*
Links	0.428**	0.481**
Hashtags	0.502**	0.462**

Pearson correlation with the age of the tweet author. Table from [Nguyen et al. ICWSM 2013]

Content bias from Normative issues

Community norms and societal biases influence observed behavior
and vary across online and offline communities and contexts

What kind of pictures would you share on **Facebook**, but not on **LinkedIn**?



Are individuals comfortable contradicting popular opinions?

E.g., after singer Prince died, most SNs showed public mourning. But not anonymous site PostSecret

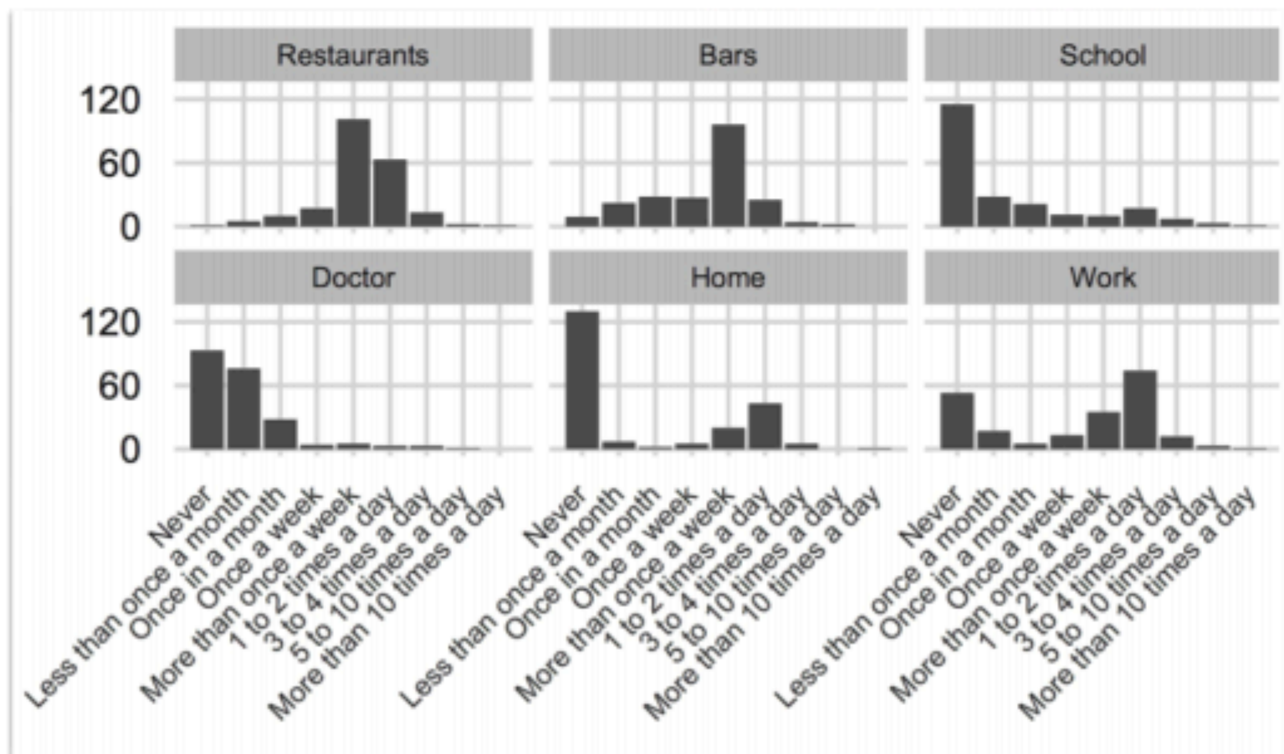
The same mechanism can embed different meanings in different contexts [Tufekci ICWSM'14]

[the meaning of retweets or likes] *"could range from affirmation to denunciation to sarcasm to approval to disgust"*

Content bias and privacy concerns

The awareness of being observed by other impacts user behavior: Privacy and safety concerns

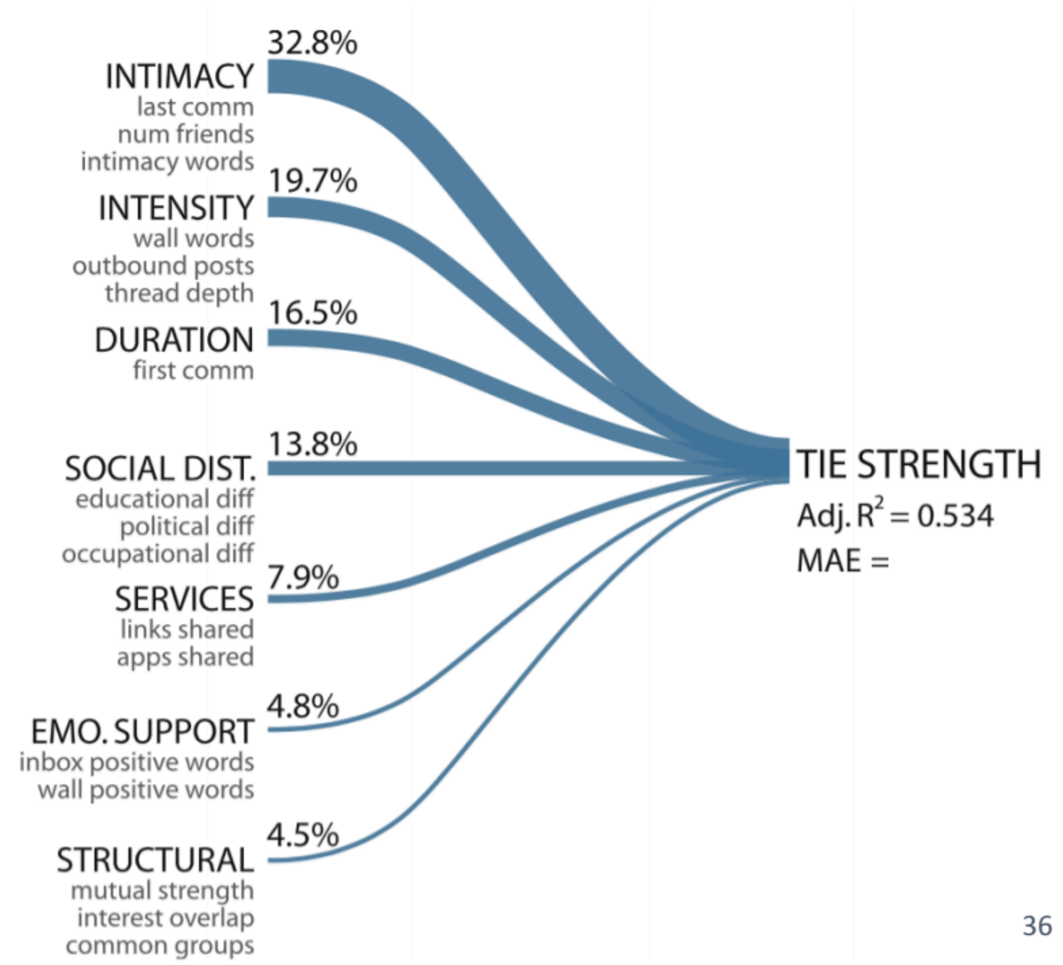
Privacy concerns affect what content users share, and, thus, the type of patterns we observe.



Foursquare/Image from [Lindqvist et al. CHI'11]

Where the data bias comes from?

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

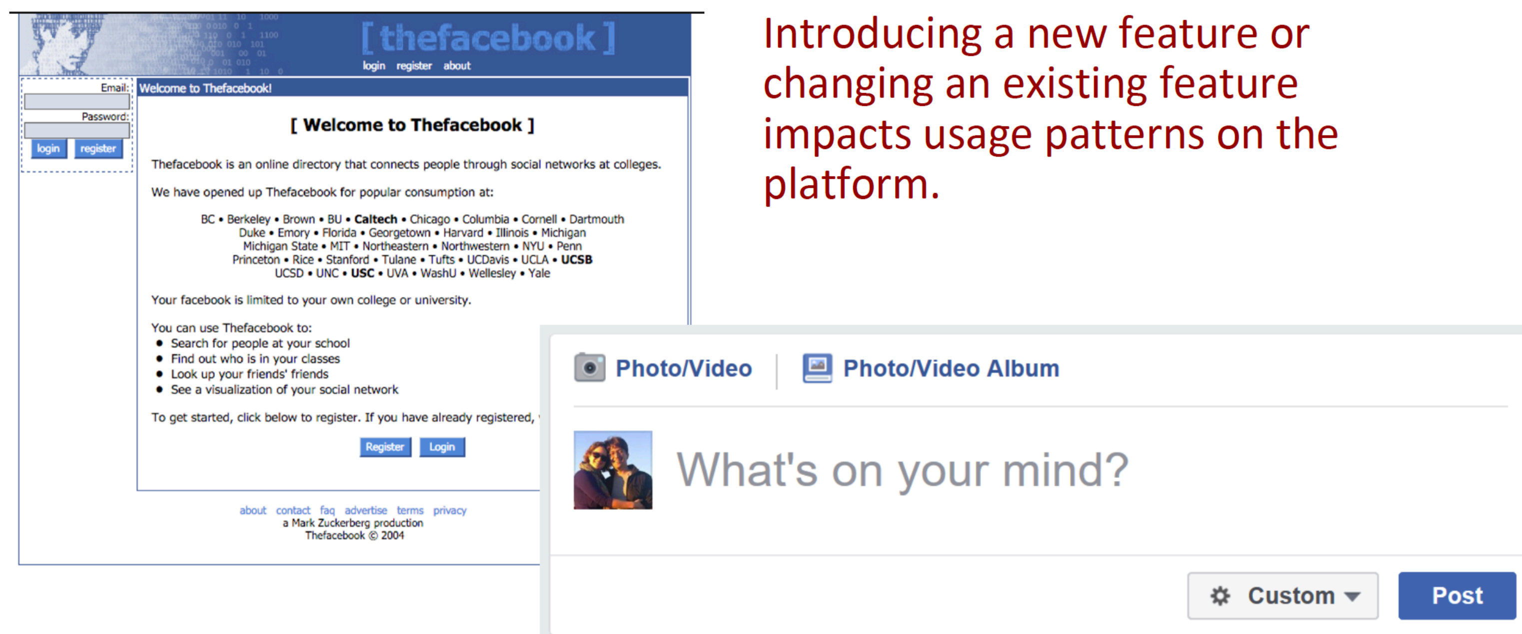


Differences in the attributes of networks obtained from user connections, interactions, or activity

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.

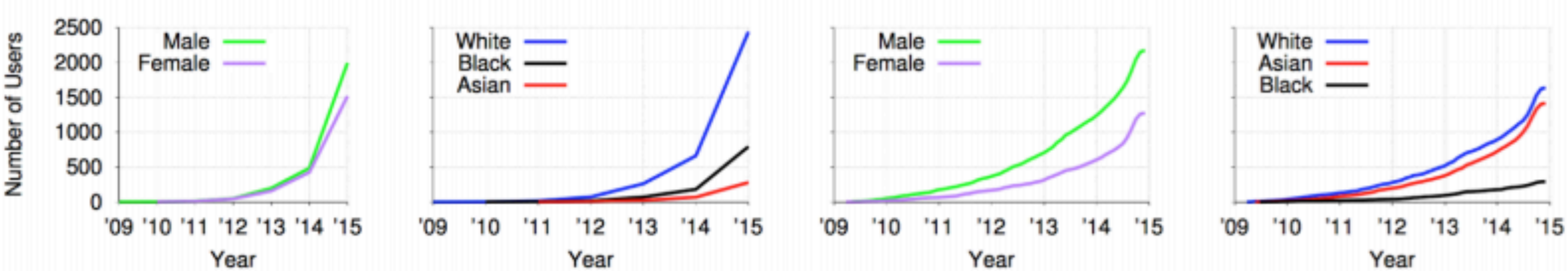
Differences in populations and behaviors over time

Université de Montréal

Temporal biases

Different demographics can exhibit different growth rates across and within social platforms

TaskRabbit and Fiverr are online freelance marketplaces.
Figure from [Hannak et al. CSCW 2017]

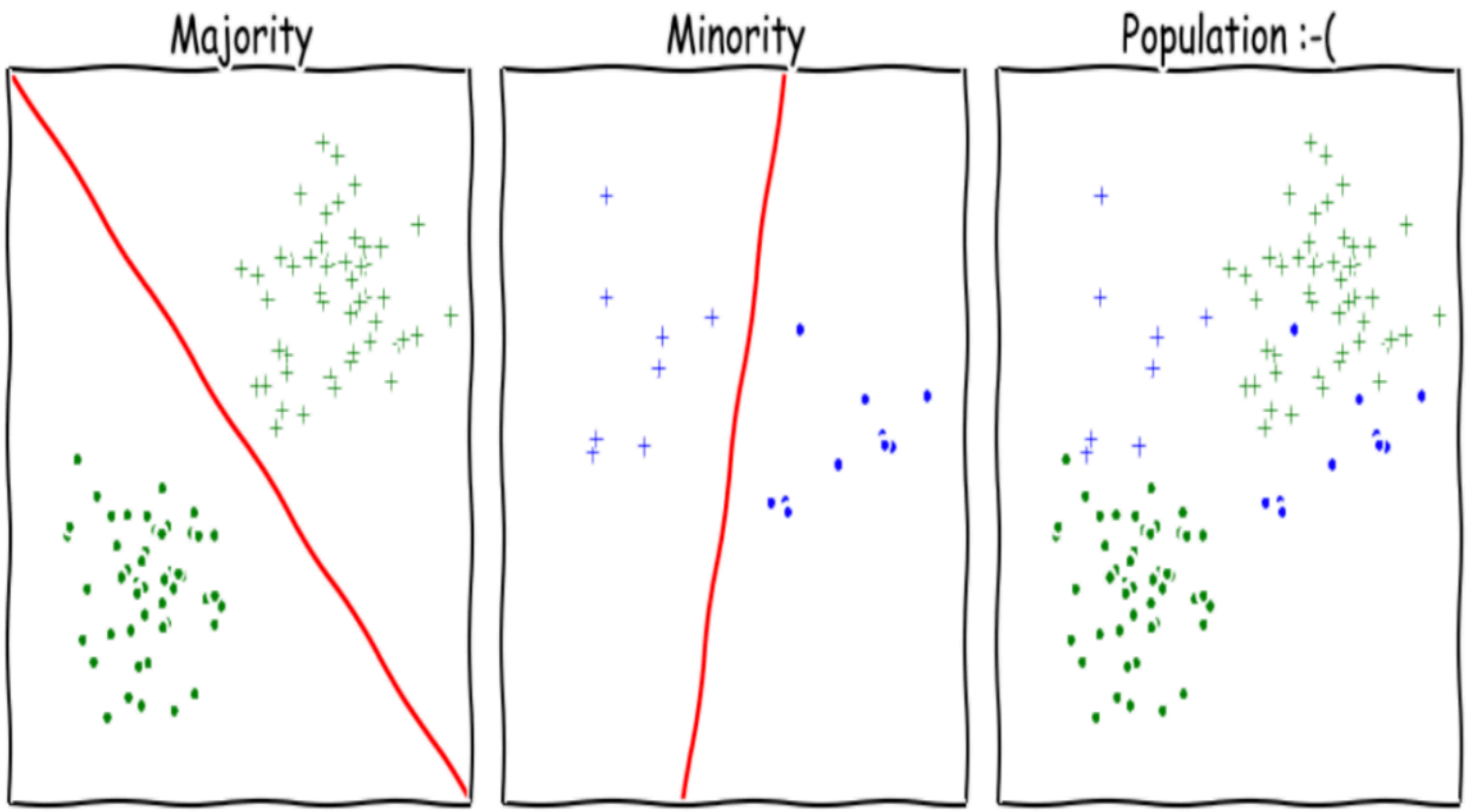


(a) TaskRabbit, gender (b) TaskRabbit, race (c) Fiverr, gender (d) Fiverr, race

Figure 1: Member growth over time on TaskRabbit and Fiverr, broken down by gender and race.

Data ~~Cleaning~~ or repairing

Removing bias from data is a very challenging task.



Data repairing is not the final solution!

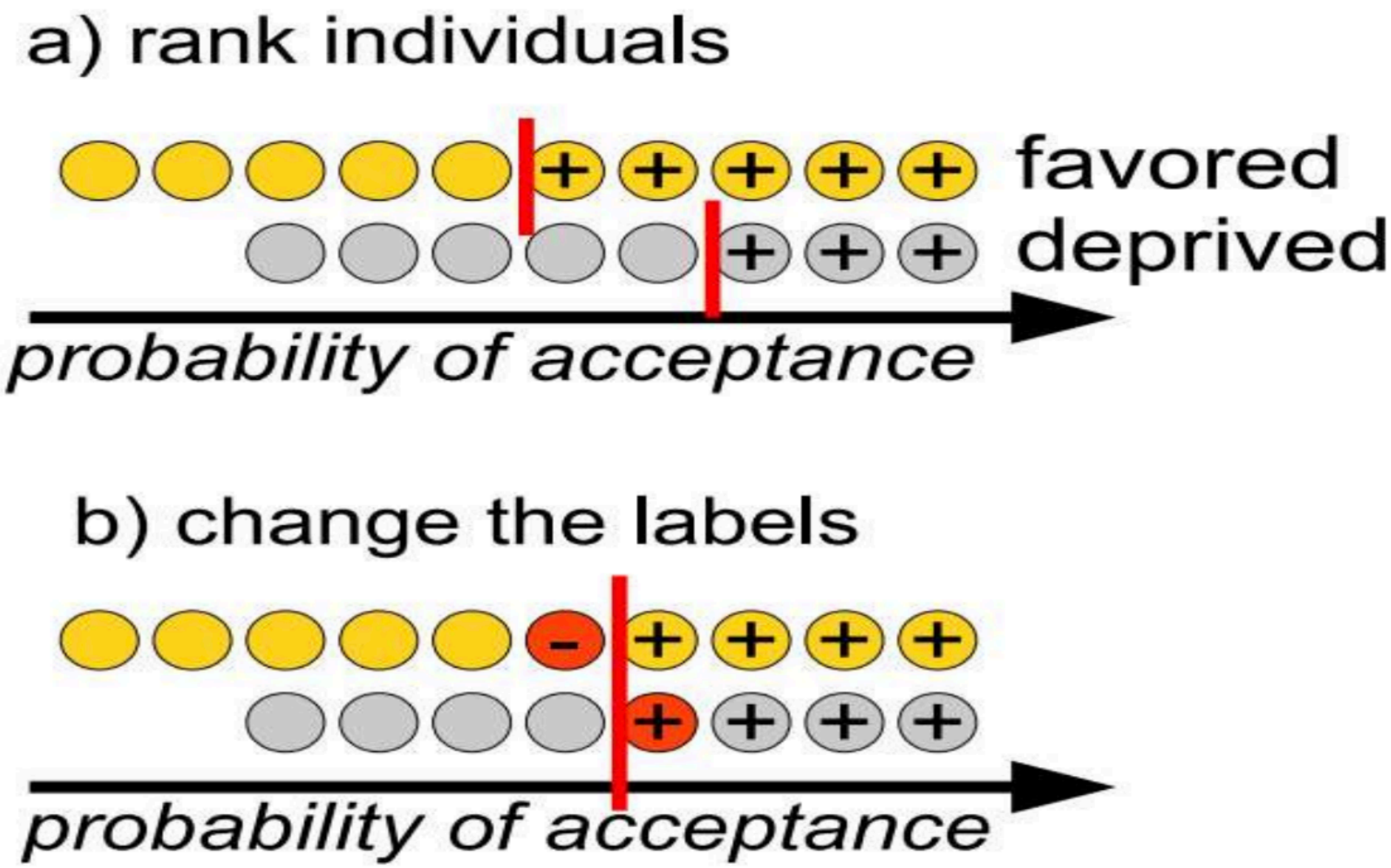
Some data repairing techniques

- **Massaging**
- **Re-weighting**
- **Sampling**
- **....**

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

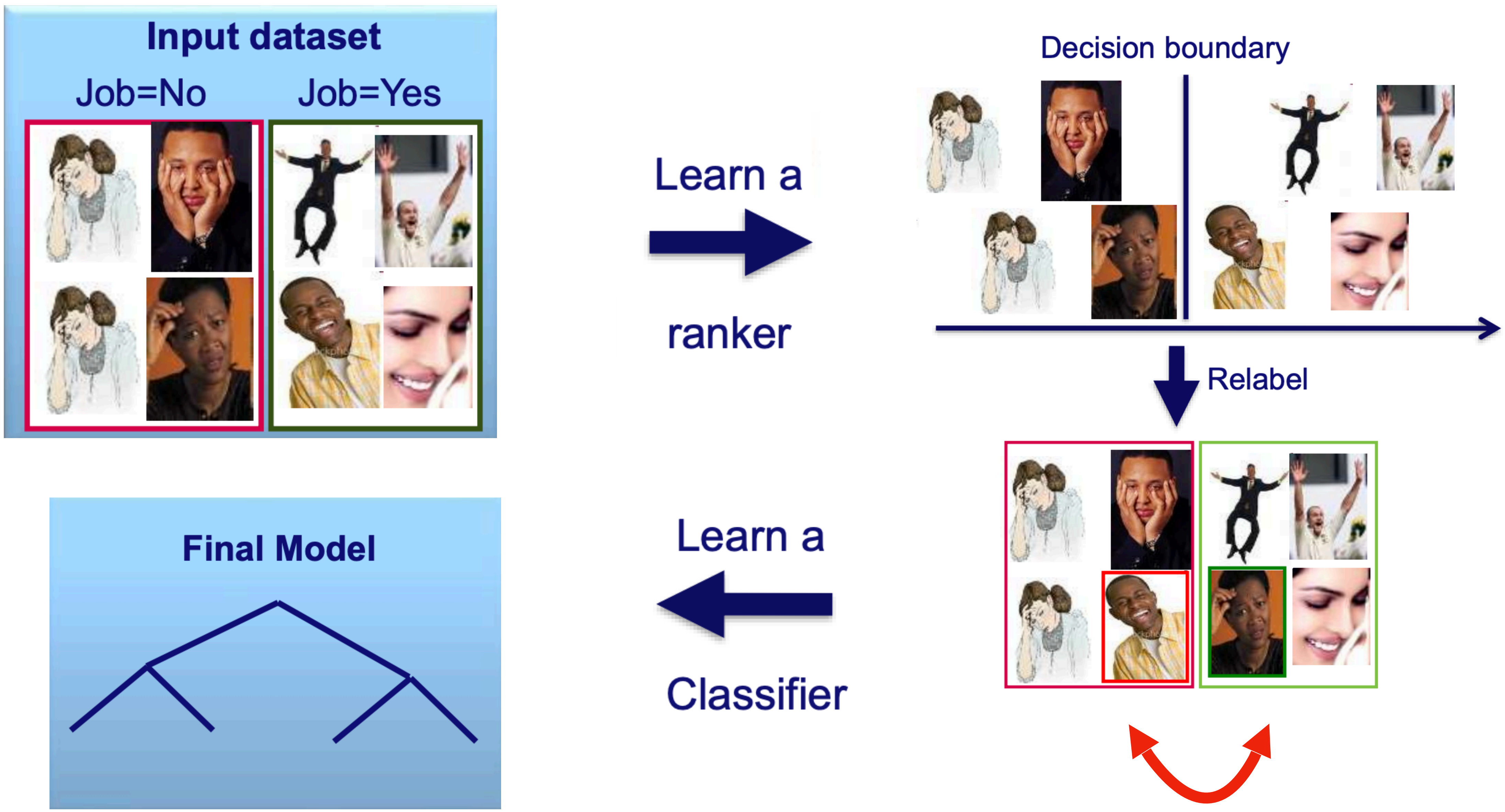


Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

Mila

Université de Montréal

Massaging

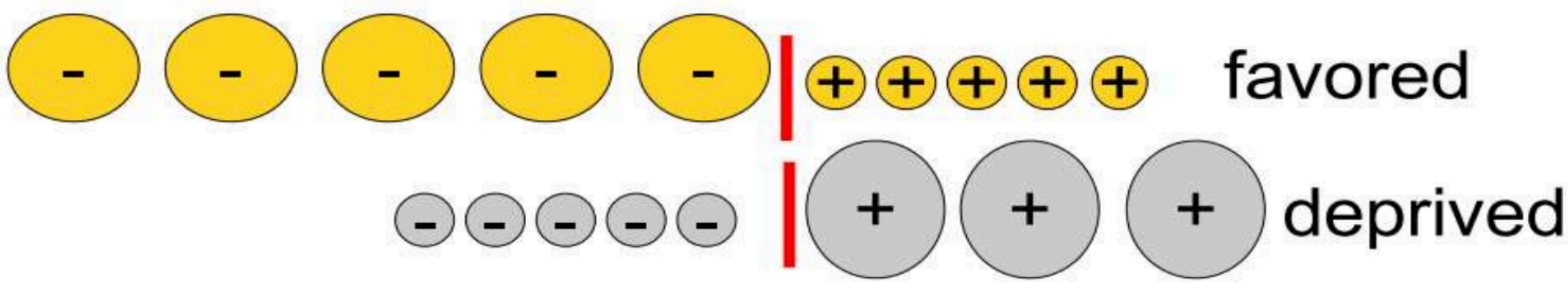


Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

Re-Weighting

a) calculate weights for the objects to neutralize the discriminatory effects from data

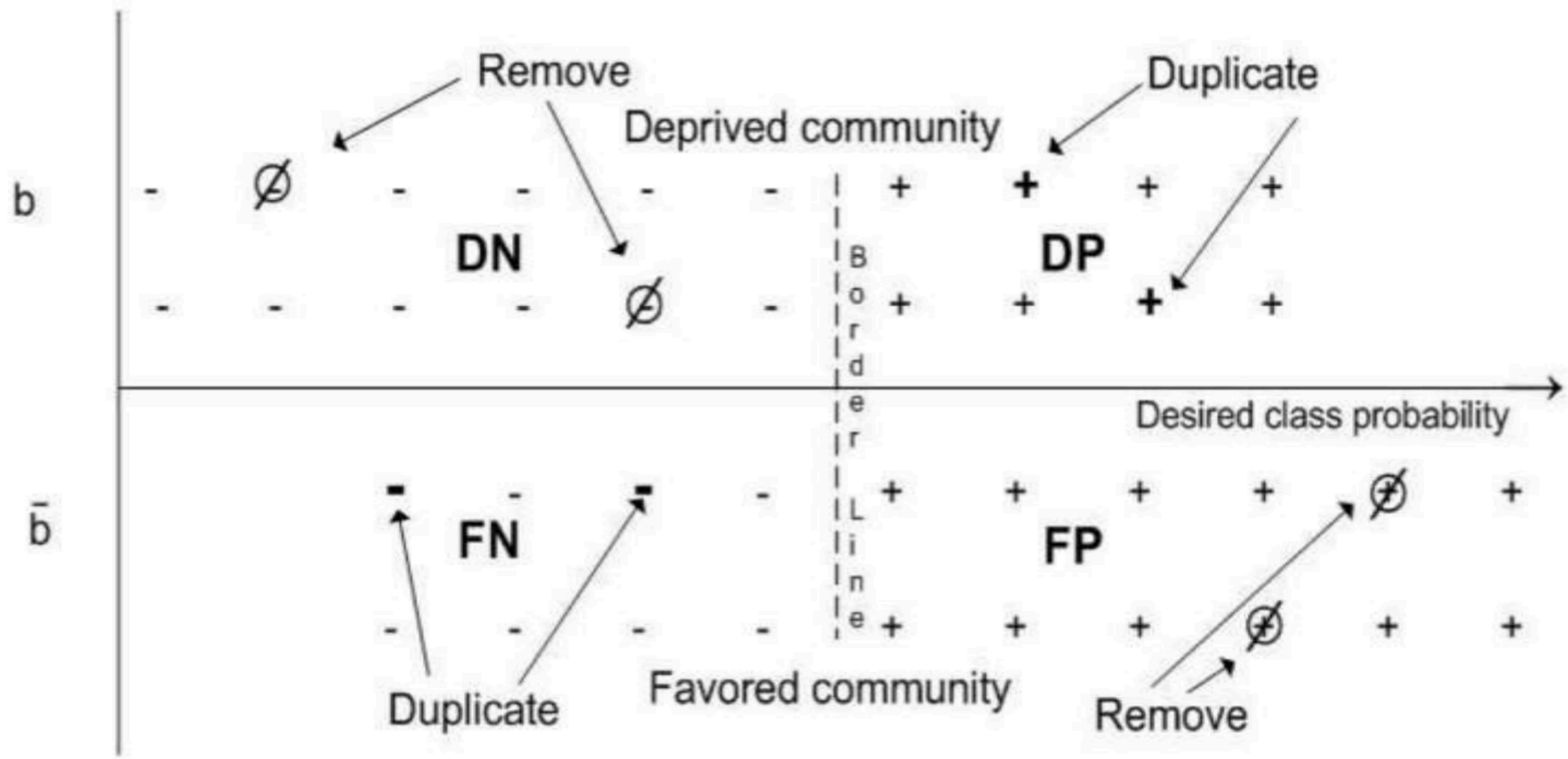
b) assign weights to make the data impartial



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.

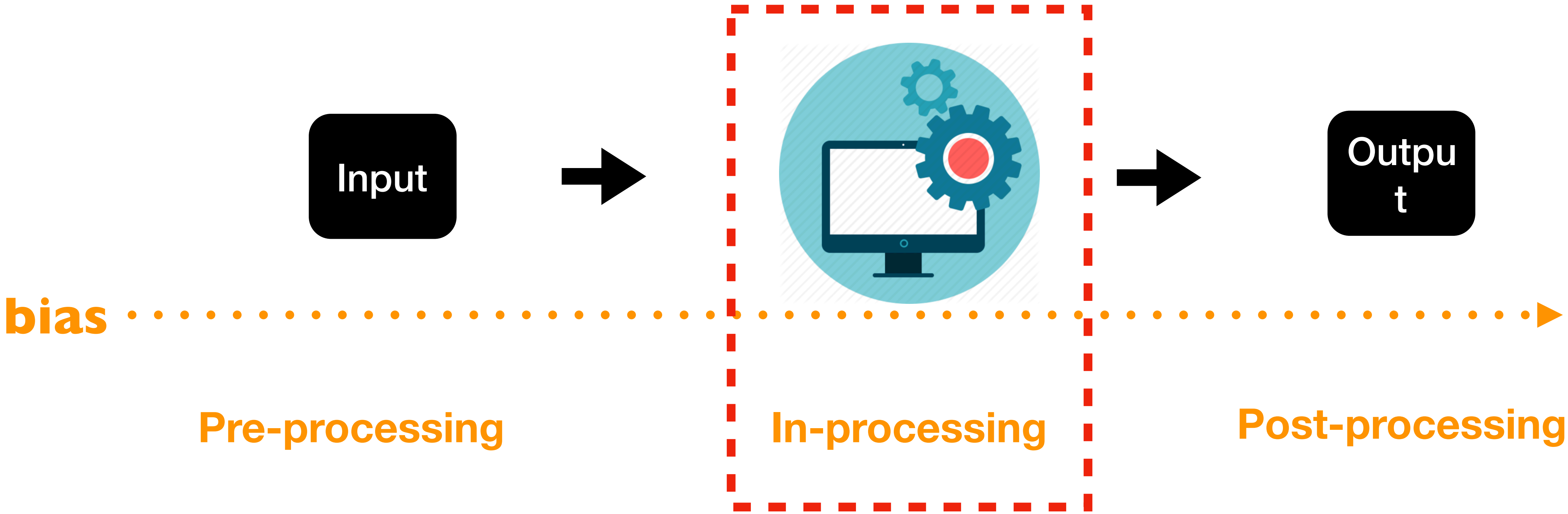
Sampling

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

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

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

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!

Learning subject to fairness constrains

Supervised learning tasks under fairness constraints are often expressed as **constrained optimization problems**

loss function

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D})$$

s.t

fairness measures

$$g_{\theta}(x, y; D)$$

Learning subject to fairness constrains

Supervised learning tasks under fairness constraints are often expressed as **constrained optimization problems**

loss function

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D})$$

s.t

e.g., demographic parity

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

Mila

Université de Montréal

Learning subject to fairness constrains

Supervised learning tasks under fairness constraints are often expressed as **constrained optimization problems**

loss function

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D})$$

s.t

e.g., demographic parity

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

Equality constraints are hard to satisfy

Learning subject to fairness constrains

Supervised learning tasks under fairness constraints are often expressed as **constrained optimization problems**

loss function

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D})$$

s.t

e.g., demographic parity

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

$$\Delta_{fair} = |p(d=1|G=f) - p(d=1|G=m)$$

Equality constraints are hard to satisfy

$$\boxed{\delta - fair}$$

$$\Delta_{fair} \leq \delta$$

Mila

Université de Montréal

Learning subject to fairness constrains

Supervised learning tasks under fairness constraints are often expressed as **constrained optimization problems**

loss function

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D})$$

s.t

$$\Delta_{fair} \leq \delta$$

Mila

Université de Montréal

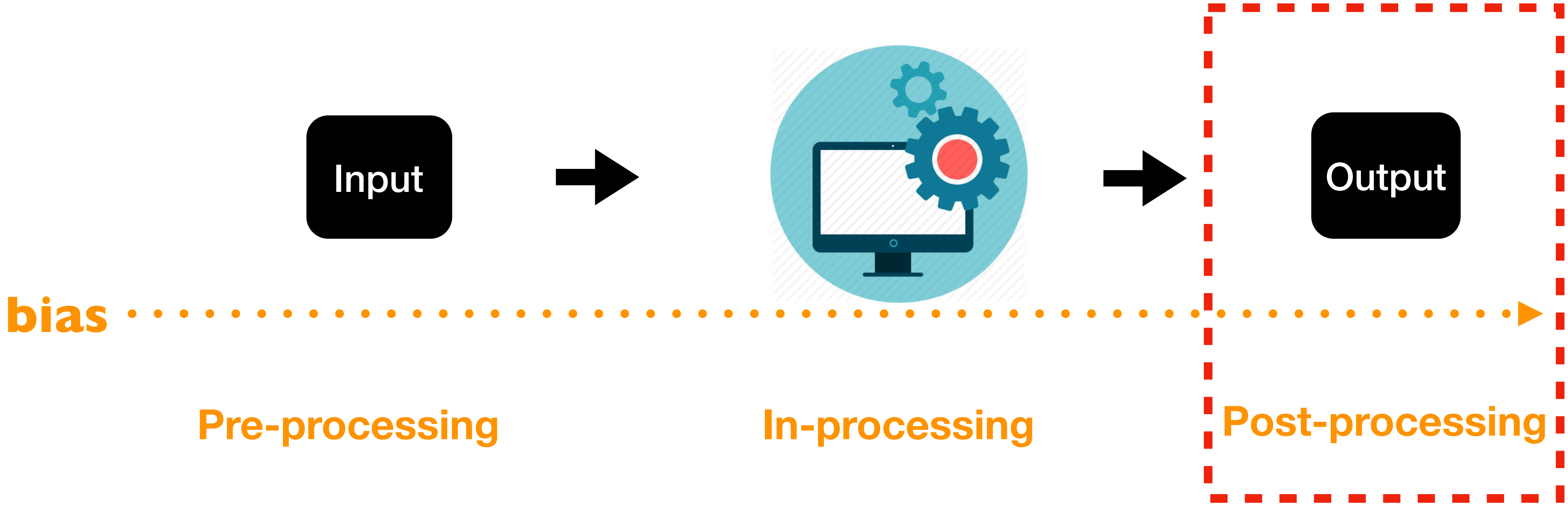
Learning subject to fairness constrains

Supervised learning tasks under fairness constraints are sometimes expressed as regularization in an optimization problems

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D}) \;+\; \lambda \times \Delta_{fair}$$

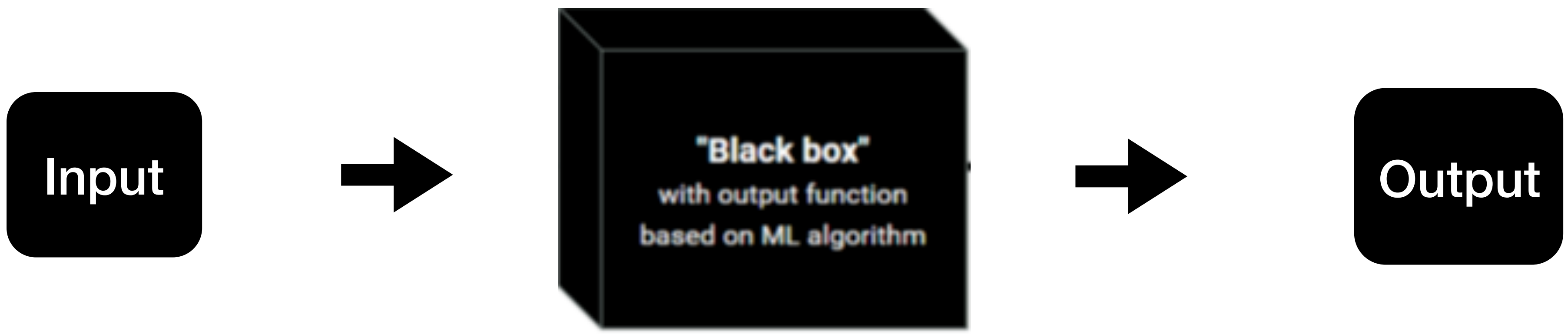
method of Lagrange multipliers

Fairness in Pro-Processing



Explaining the Output (black box)

More about this on Week 13



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.

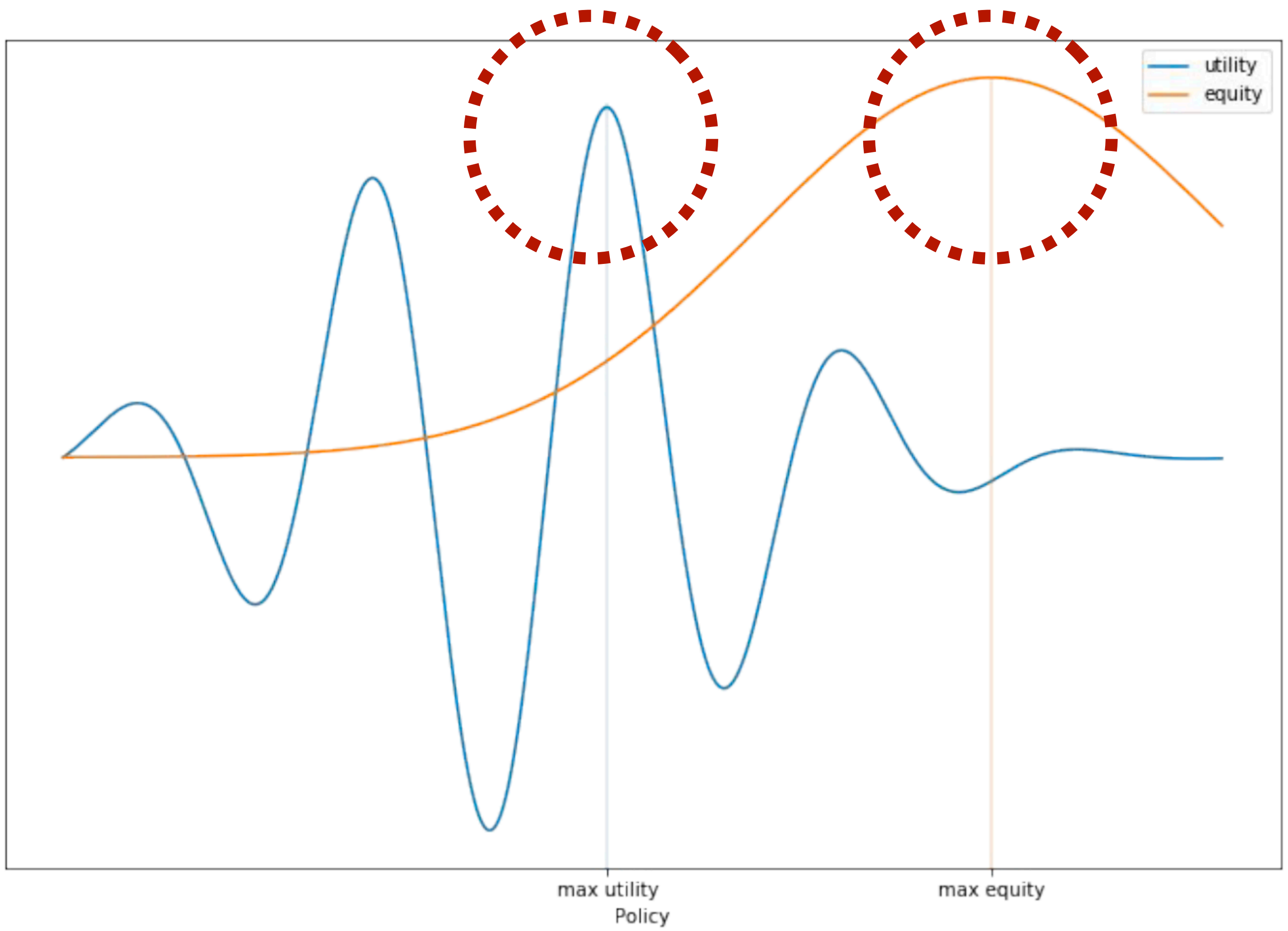
Kim, Michael P., Amirata Ghorbani, and James Zou. "Multiaccuracy: Black-box post-processing for fairness in classification." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2019.

Mila

Université de Montréal

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.

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).

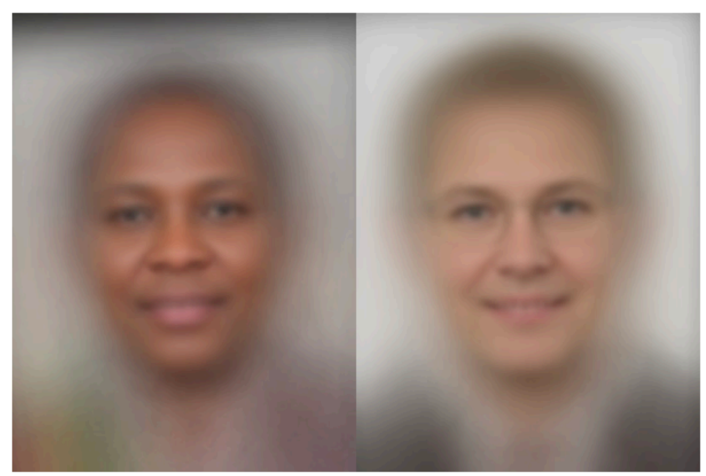
Mila

Université de Montréal

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?

Be a responsible data scientist!

Conferences focusing on Fairness in ML/AI

- ACM FAT*: ACM Conference on Fairness, Accountability, and Transparency https://fatconference.org/

- AIES: AAAI/ACM conference on Artificial intelligence, Ethics and society https://www.aies-conference.com/2020/

- Many workshops: FATML, FATNLP, FATCV, FTML4Health, FATREC, etc.

- Other conferences interested on this topic: AAAI, IJCAI, Neurips, ICML, etc.