Il concetto di equità (fairness) nel Machine Learning e nel Natural Language processing

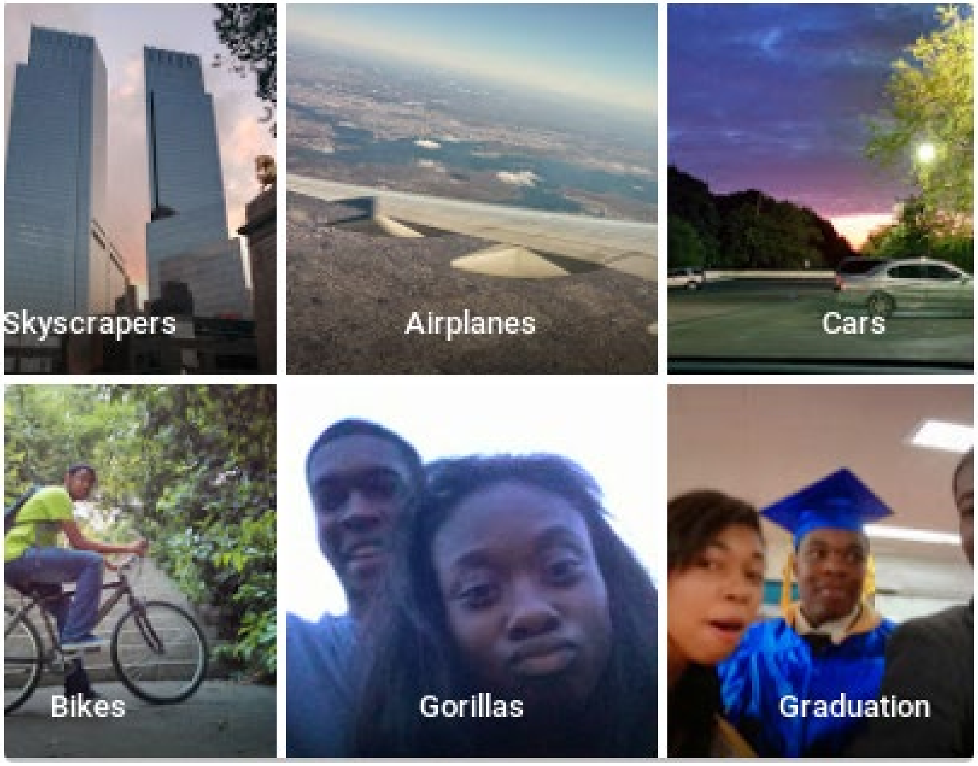
C. Batini

Ambiti di applicazione della fairness

- **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

I gorilla

La figura mostra un insieme di classificazioni prodotte da Google Photos, in cui due persone africane vengono classificati come gorilla



Studio di caso - Come capire con il Machine Learning a chi concedere la libertà provvisoria a detenuti in attesa di giudizio, e con quale cauzione?

References for the next section

D. Kehl Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing, The Harvard Library, 2016.

Guide to the Pretrial Decision Framework – Laura and John Arnold Foundation, 2018.

Wisconsin offender statement - https://doc.wi.gov/Pages/AboutDOC/COMPAS.aspx

Legally recognized ‘protected classes’ in U.S.

- **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)

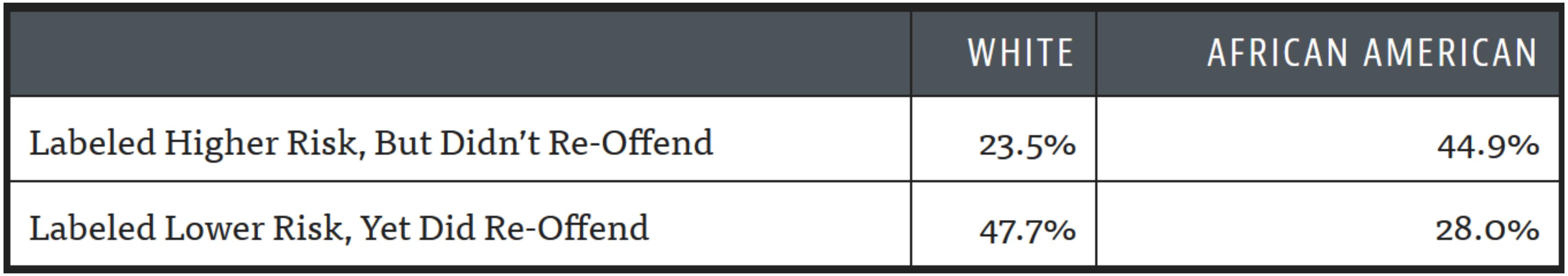
Context of use of ML in Justice in North America

Risk assessment tools and software–many of which incorporate machine learning–are now being used in US and Canada in a variety of contexts, including

- Prison rehabilitation programs,
- Pretrial risk assessment,
- Sentencing.

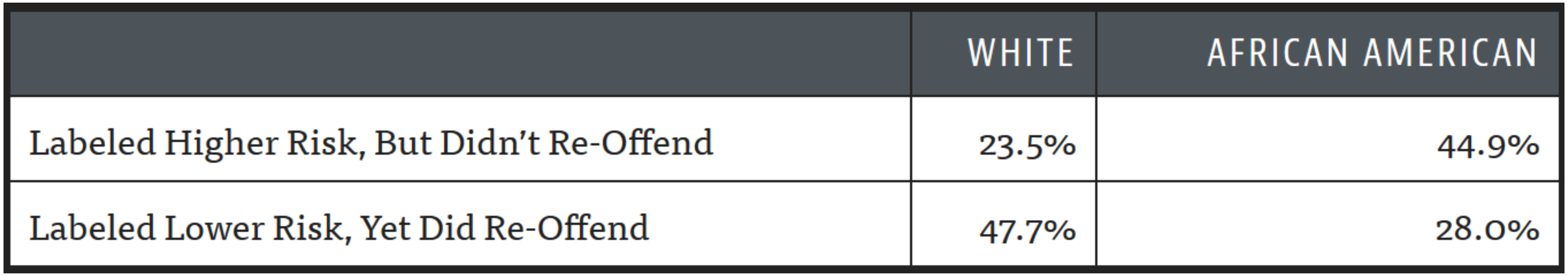
Let us focus in the following on pre-trial phase

2016 Propublica Analysis:
Prediction of Compas Fails Differently for Black Defendants



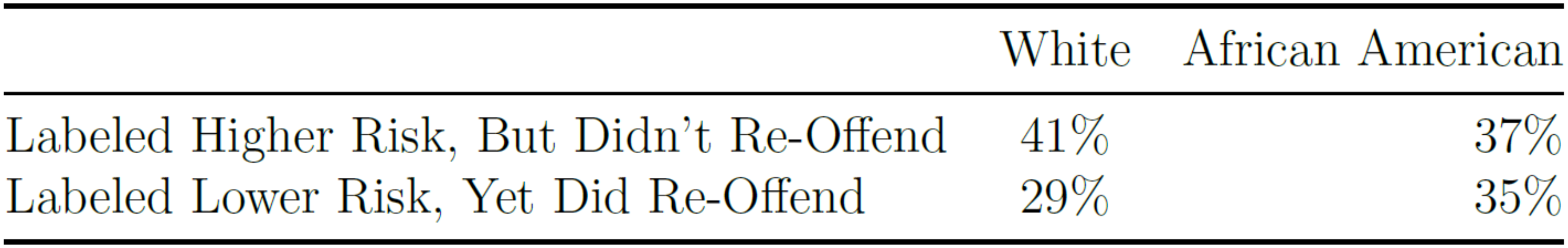
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Propublica Analysis



	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%

2016 Northpointe response



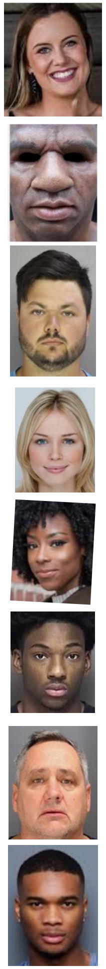
	White	African American
Labeled Higher Risk, But Didn't Re-Offend	41%	37%
Labeled Lower Risk, Yet Did Re-Offend	29%	35%

Chi ha ragione?

- Dipende….
- Ci sono tante definizioni di Fairness
- Almeno 20 definizioni diverse…

Problema che affrontiamo

Attualmente detenuti In attesa di giudizio



→

A chi dare
la libertà provvisoria?
E con quale cauzione?

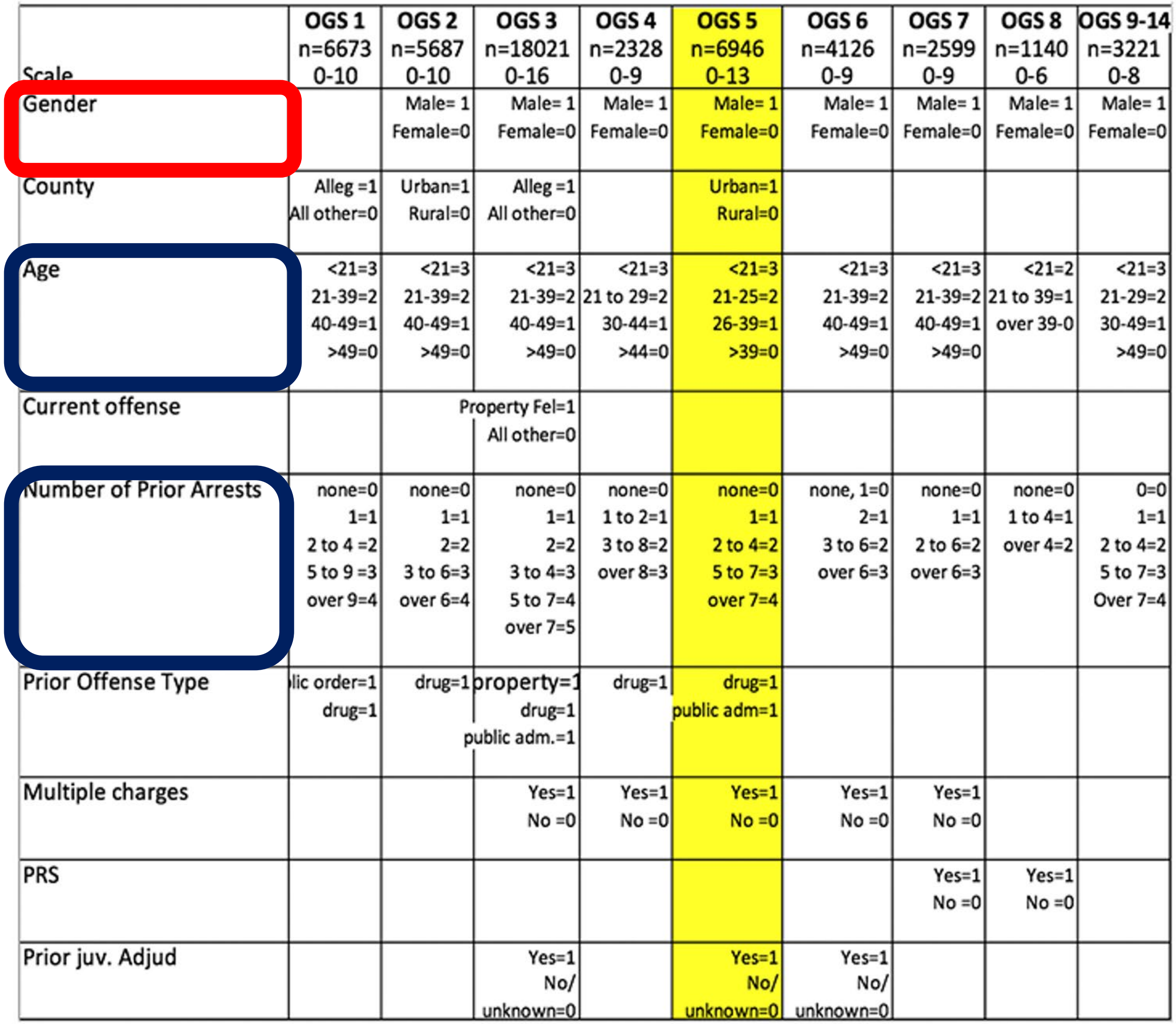
→ ?

Features considered in Compas

Attributo protetto

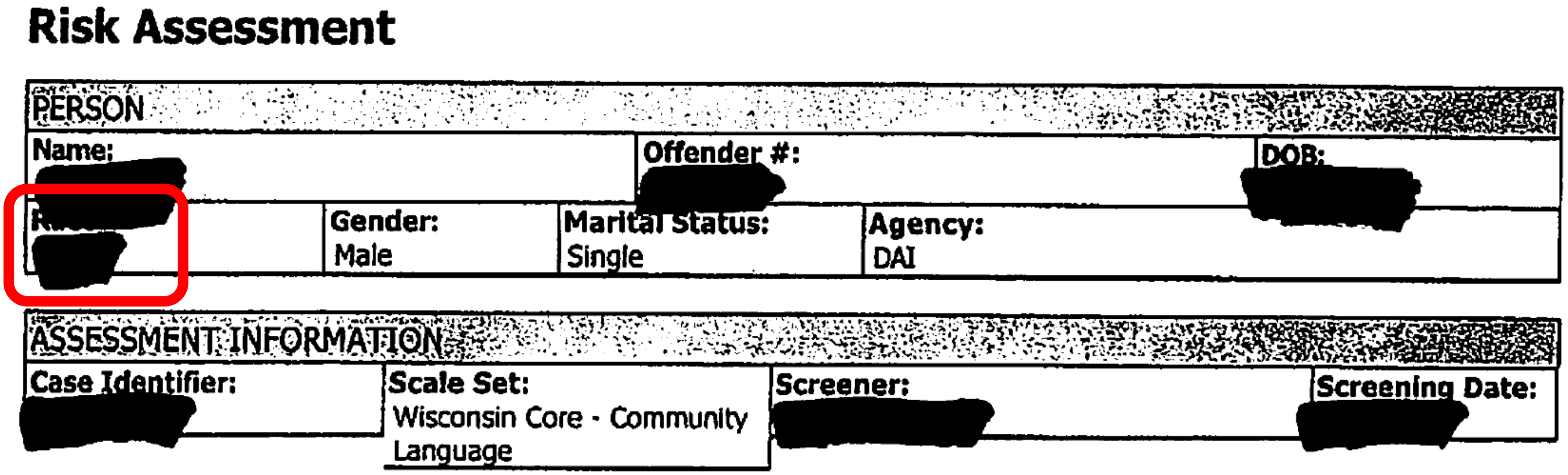
Attributo regolare

Attributo regolare



Scale	OGS 1 n=6673 0-10	OGS 2 n=5687 0-10	OGS 3 n=18021 0-16	OGS 4 n=2328 0-9	OGS 5 n=6946 0-13	OGS 6 n=4126 0-9	OGS 7 n=2599 0-9	OGS 8 n=1140 0-6	OGS 9-14 n=3221 0-8
Gender		Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0	Male= 1 Female=0
County	Alleg =1 All other=0	Urban=1 Rural=0	Alleg =1 All other=0		Urban=1 Rural=0				
Age	<21=3 21-39=2 40-49=1 >49=0	<21=3 21-39=2 40-49=1 >49=0	<21=3 21-39=2 40-49=1 >49=0	<21=3 21 to 29=2 30-44=1 >44=0	<21=3 21-25=2 26-39=1 >39=0	<21=3 21-39=2 40-49=1 >49=0	<21=3 21-39=2 40-49=1 >49=0	<21=2 21 to 39=1 over 39-0	<21=3 21-29=2 30-49=1 >49=0
Current offense			Property Fel=1 All other=0						
Number of Prior Arrests	none=0 1=1 2 to 4 =2 5 to 9 =3 over 9=4	none=0 1=1 2=2 3 to 6=3 over 6=4	none=0 1=1 2=2 3 to 4=3 5 to 7=4 over 7=5	none=0 1 to 2=1 3 to 8=2 over 8=3	none=0 1=1 2 to 4=2 5 to 7=3 over 7=4	none, 1=0 2=1 3 to 6=2 over 6=3	none=0 1=1 2 to 6=2 over 6=3	none=0 1 to 4=1 over 4=2	0=0 1=1 2 to 4=2 5 to 7=3 Over 7=4
Prior Offense Type	lic order=1 drug=1	drug=1	property=1 drug=1 public adm.=1	drug=1	drug=1 public adm=1				
Multiple charges			Yes=1 No =0	Yes=1 No =0	Yes=1 No =0	Yes=1 No =0	Yes=1 No =0		
PRS							Yes=1 No =0	Yes=1 No =0	
Prior juv. Adjud			Yes=1 No/ unknown=0		Yes=1 No/ unknown=0	Yes=1 No/ unknown=0			

Ma anche…



Risk Assessment

PERSON

Name: [redacted]

Offender #: [redacted]

DOB: [redacted]

R[redacted] [redacted]

Gender: Male

Marital Status: Single

Agency: DAI

ASSESSMENT INFORMATION

Case Identifier: [redacted]

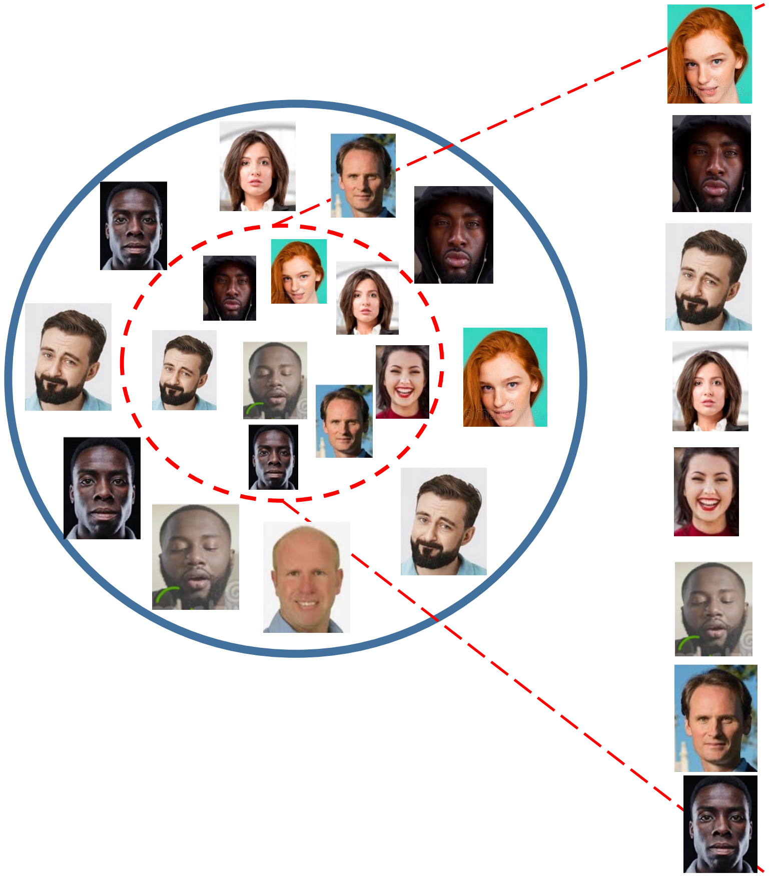
Scale Set: Wisconsin Core - Community Language

Screener: [redacted]

Screening Date: [redacted]

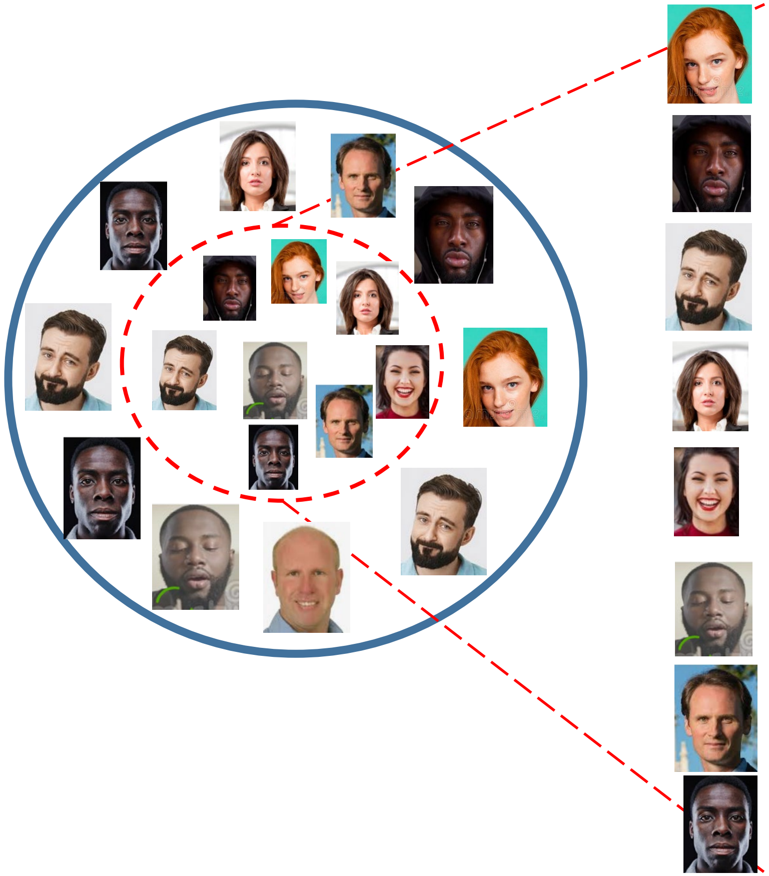
Scegliamo le seguenti caratteristiche

← Caratteristiche →



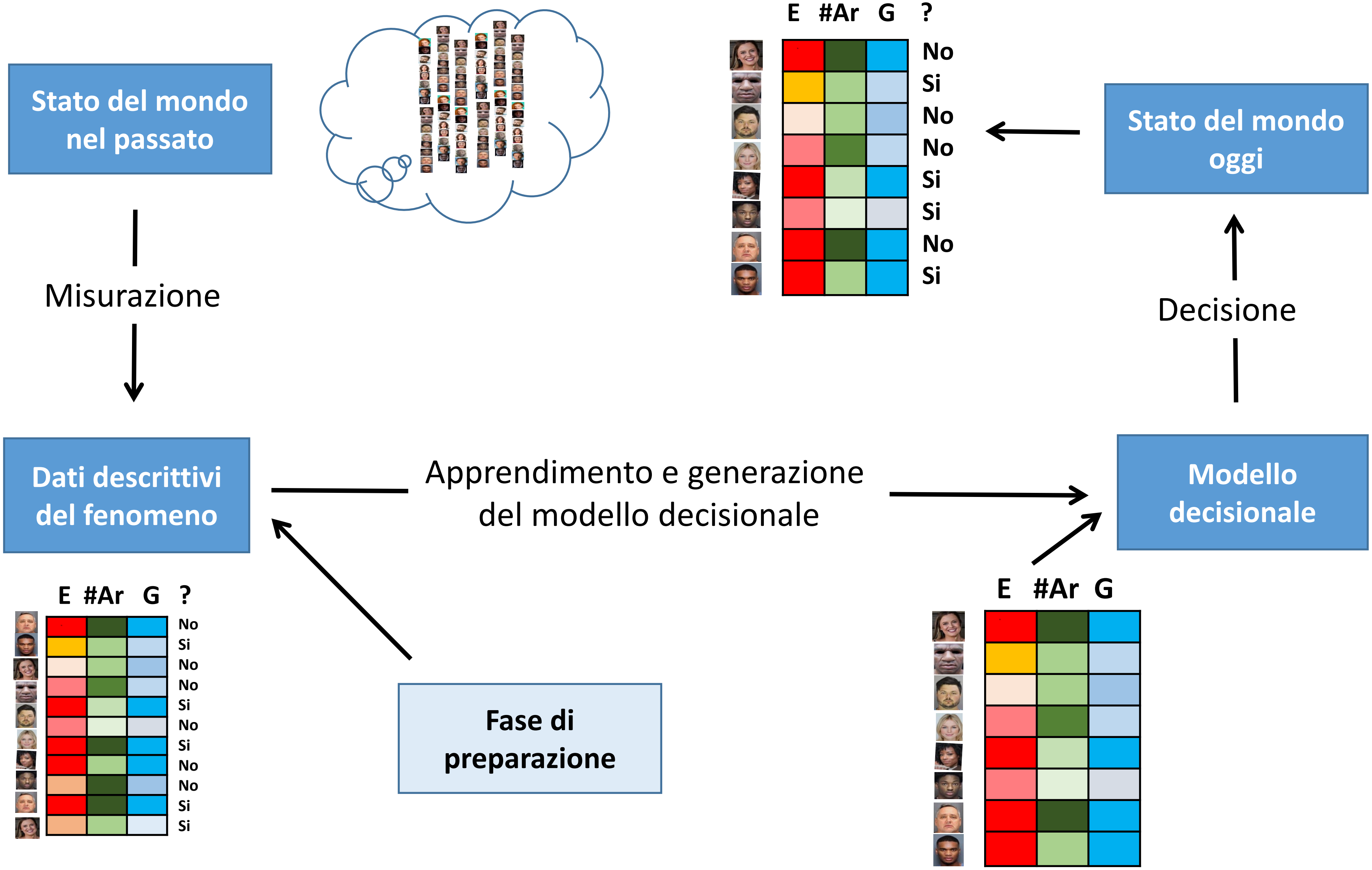
Età	# Arresti precedenti	Genere	Ha commesso Recidiva?

Scegliamo le seguenti caratteristiche



⟵ Caratteristiche ⟶			
		← c.protetta →	
Età	# Arresti precedenti	Genere	Ha commesso Recidiva?
🟦	🟦	🟥	
🟦	🟦	🟥	
🟦	🟦	🟥	
🟦	🟦	🟥	
🟦	🟦	🟥	
🟦	🟦	🟥	
🟦	🟦	🟥	
🟦	🟦	🟥	

Il ciclo dell’apprendimento



Stato del mondo nel passato

Misurazione

Dati descrittivi del fenomeno

E #Ar G ?

No
Si
No
No
Si
No
Si
No
No
Si
Si

Fase di preparazione

Apprendimento e generazione del modello decisionale

Modello decisionale

E #Ar G

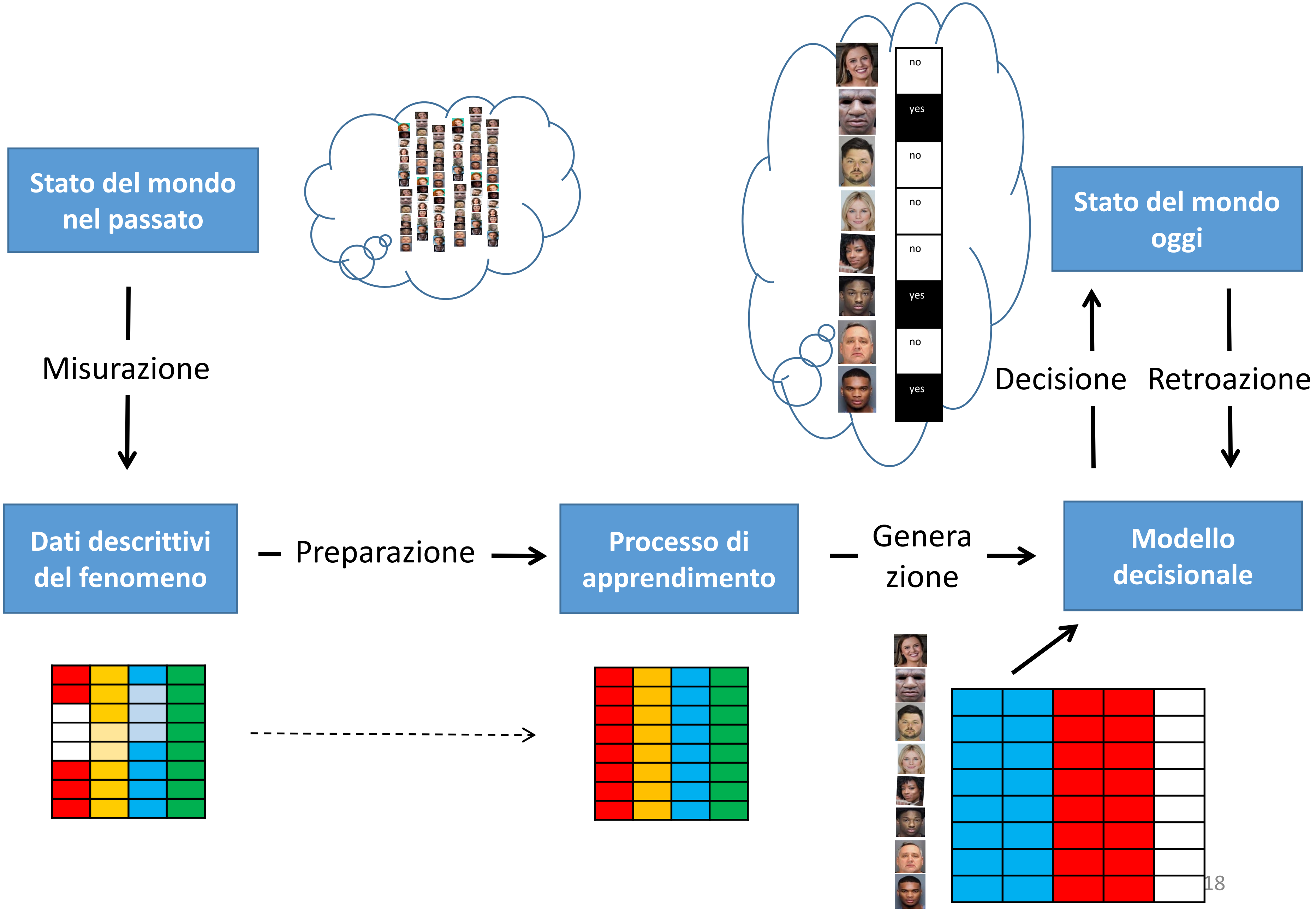
Decisione

Stato del mondo oggi

E #Ar G ?

No
Si
No
No
Si
Si
No
Si

Il ciclo del machine learning



Stato del mondo nel passato

Misurazione

Dati descrittivi del fenomeno

Preparazione

Processo di apprendimento

Generazione

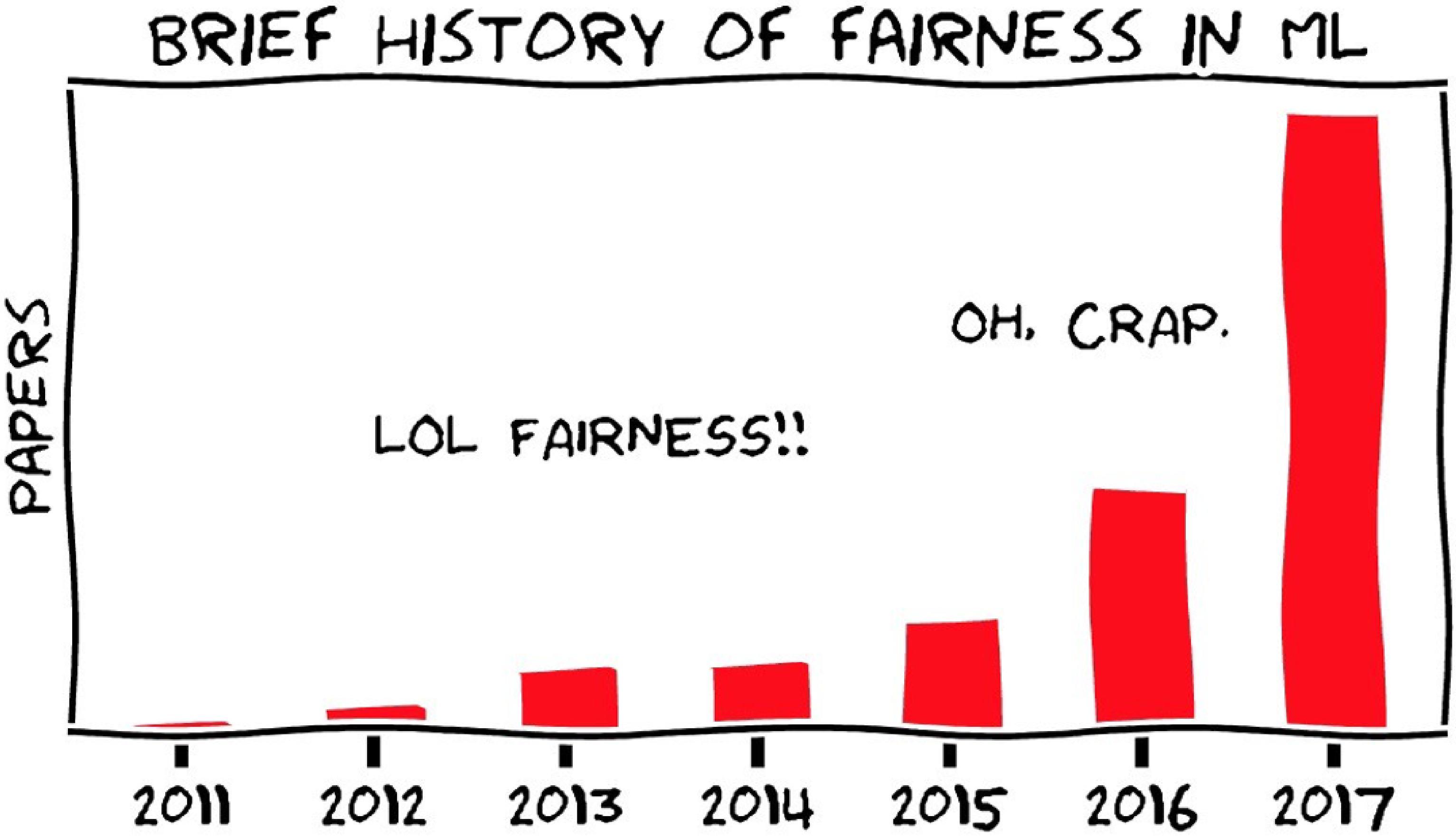
Modello decisionale

Decisione

Stato del mondo oggi

Retroazione

The number of publications on Fairness from 2011 to 2017



Types of fairness – all together

Types of fairness

Basic → True Positive, True Negative, False Positive, False Negative

Composite →

1. Positive predictive value TP/(TP+FP)
2. False discovery rate FP/(TP+FP)
3. False omission rate FN/(TN+FN)
4. Negative predictive value TN/(TN+FN)
5. True positive rate TP/(TP+FN) or Sensitivity
6. False positive rate FP/(FP+TN)
7. False negative rate FN/(TP+FN) or Specificity
8. True negative rate TN/(FP+TN)
9. Group fairness/statistical parity/equal acceptance rate/benchmarking
10. Conditional statistical parity
11. Predictive parity/outcome test
12. False positive error rate balance/Pedicrtive equality
13. False negative error rate balance/Equal opportunity
14. Equalized odds/conditional procedure accuracy equality/disparate treatment
15. Conditional use accuracy quality
16. Overall accuracy equality
17. Treatment equality
18. Test fairness/calibration/matching conditional frequencies
19. Well calibration
20. Balance for positive class
21. Balance for negative class
22. Causal discrimination
23. Fairness through unawareness
24. Fairness through awareness
25. Counterfactual fairness
26. No unresolved discrimination
27. No proxy discrimination
28. Fair inference

Alle soglie di una crisi di nervi….

Name of fairness	Stakeholder	Metric	Pro/ North/Chu	ModelErr/ TargetPopErr	Short definition
SBASED Recall, Sensitivity, True positive value		TP/(TP+FN)		Model Error	Avoidance of false negative
SBASED Specificity, True negative value		1 – false positive rate		Model Error	Avoidance of false positive
SBASED False negative rate		FN/(FN+TP)	Pro, Chu	Model Error	Complement of sensitivity
SBASED False positive rate	Defendant,	**Error** rate balance FP/(FP+TN)	Pro, Chu	Model Error	Complement of specificity
SBASED Precision, **positive predictive value**	Decision maker	TP/(TP+FP)	North	TargetPop Error	
SBASED Negative Predictive Value		TN/(TN+FN)	North	TargetPop Error	
PRACBASED 2. Classification parity, Predictive parity		Any measure based on confusion matrix	North, Chu, Corbett	TargetPopul. Error	**Predictive performance** is equal among groups defined by the protected attributes or some measure of class. **Error is =** among groups def by protected attributes
PROBASED Independence, Statistical parity**, group fairness**, equal acceptance rate **Demographic parity**, Equal impact, equal outcome, Benchmarking	Society		**For Corbett equivalent** to 2. Class parity		the proportion of individuals classified as high-risk is the same for each group. Also detention rates are equal across race groups
POBASED Conditional statistical parity					Group fairness, conditional to a set of legitimate attrs.
PROBACOBASED 3. Calibration, test-fairness, matching conditional frequencies			Corbett		**Outcomes** are independent from protected attributes after controlling for estimated risk
False positive/ false negative Error rate balance, predictive equality/ equal opportunity					false positive and false negative error rates are equal across groups.
Unawareness, Avoid disparate treatment, 1. Anti-classification			Corbett		Ignore sensitive features in classification or decisions do not consider protected attributes
Separation, Positive rate parity					
PRACBASED Accuracy Parity		Accuracy			
Equality of opportunityunawarebness					
Causal discrimination					Members with similst values in attributes X are tereated differently

….. ho trovato un articolo del 2018

- S. Verma et al. **Fairness Definitions Explained -** 2018 ACM/IEEE International Workshop on Software Fairness

Types of fairness – all together

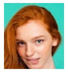
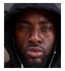
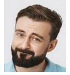
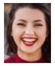
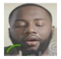
Categories of fairness	**Types of fairness**
1. Statistical measures 2. Based on Predicted outcome for various demographic distributions of subjects 3. Based on Predicted outcomes that are compared with the Actual Outcomes 4. Based on Predicted Probabilities and Actual Outcome 5. Similarity based 6. Causal reasoning based	Basic → True Positive, True Negative, False Positive, False Negative Composite → 1. Positive predictive value $TP/(TP+FP)$ 2. False discovery rate $FP/(TP+FP)$ 3. False omission rate $FN/(TN+FN)$ 4. Negative predictive value $TN/(TN+FN)$ 5. True positive rate $TP/(TP+FN)$ or Sensitivity 6. False positive rate $FP/(FP+TN)$ 7. False negative rate $FN/(TP+FN)$ or Specificity 8. True negative rate $TN/(FP+TN)$ 9. Group fairness/statistical parity/equal acceptance rate/benchmarking 10. Conditional statistical parity 11. Predictive parity/outcome test 12. False positive error rate balance/Pedicrtive equality 13. False negative error rate balance/Equal opportunity 14. Equalized odds/conditional procedure accuracy equality/disparate treatment 15. Conditional use accuracy quality 16. Overall accuracy equality 17. Treatment equality 18. Test fairness/calibration/matching conditional frequencies 19. Well calibration 20. Balance for positive class 21. Balance for negative class 22. Causal discrimination 23. Fairness through unawareness 24. Fairness through awareness 25. Counterfactual fairness 26. No unresolved discrimination 27. No proxy discrimination 28. Fair inference

Types of fairness – all together

Categories of fairness	**Types of fairness**
Statistical measures → **Measures of accuracy**	Basic → **True Positive, True Negative, False Positive, False Negative** Composite → 1. Positive predictive value TP/(TP+FP) 2. False discovery rate FP/(TP+FP) 3. False omission rate FN/(TN+FN) 4. Negative predictive value TN/(TN+FN) 5. True positive rate TP/(TP+FN) 6. False positive rate FP/(FP+TN) 7. False negative rate FN/(TP+FN) 8. True negative rate TN/(FP+TN)
Based on Predicted outcome for various demographic distributions of subjects	
Based on Predicted outcomes that are compared with the Actual Outcomes	
Based on Predicted Probabilities and Actual Outcome	
Similarity based	
Causal reasoning based	

Fase di separazione tra training data e test data

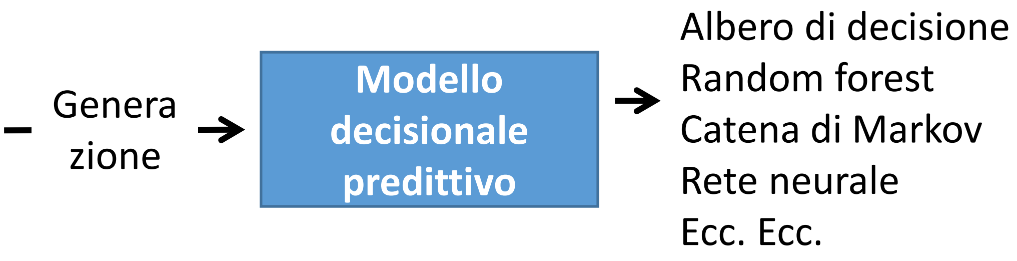
Training Data

	35	1	Donna	Bianca	No
	40	3	D	Nero	No
	20-30		Uomo	B	Si
	35	0	D	B	Si
	24	1	D	B	No
	45		U	N	Si
	50	2	U	B	Si
		2	U	N	No

Test Data

Fase di generazione del modello predittivo dai training data

Età	# Arresti precedenti	Genere	Etnia	Ha commesso recidiva?
35	1	Donna	Bianca	No
40	3	D	Nero	No
20-30		Uomo	B	Si
35	0	D	B	Si
24	1	D	B	No



Generazione →

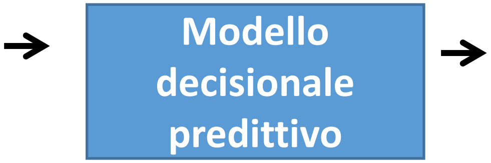
Modello decisionale predittivo

→

Albero di decisione
Random forest
Catena di Markov
Rete neurale
Ecc. Ecc.

Sottofase di verifica di qualità del modello

Età	# Arresti precedenti	Genere	Etnia	Ha commesso recidiva?
45		U	N	Si
50	2	U	B	Si
	2	U	N	No

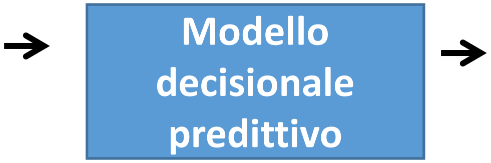


Modello decisionale predittivo

Esito del test
No
Si
No

Esito del test: un falso negativo e un falso positivo

Età	# Arresti precedenti	Genere	Etnia
45		U	N
50	2	U	B
	2	U	N



Esito del test	Ha com-messo recidiva?
No	Si
Si	No
No	No

Dunque l'algoritmo si può sbagliare
Anzi: sicuramente si sbaglia

Misure di accuratezza

Età	# Arresti precedenti	Genere	Etnia
45		U	N
50	2	U	B
	2	U	N

→ Modello decisionale predittivo →

Esito del test	Ha com-messo recidiva?
No	Si
Si	No
No	No

Matrice dei casi possibili In generale	Ha commesso dice no	Ha commesso dice si
Modello dice no	**Vero negativo**	**Falso negativo**
Modello dice si	**Falso positivo**	**Vero positivo**

Matrice dei casi possibili nel nostro esempio	Ha commesso dice no	Ha commesso dice si
Modello dice no	**Un vero negativo**	**Un Falso negativo**
Modello dice si	**Un Falso positivo**	**Zero Veri positivi**

Esempi di misure di accuratezza nel caso di modello decisionale → si o no

- Precisione = veri positivi / (veri positive + falsi positivi)
- Recall = veri positivi / (veri positivi + falsi negativi)

Nota bene – nel nostro caso

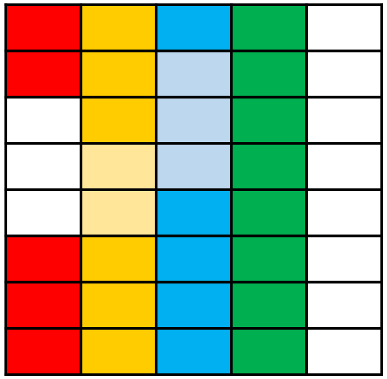
- Rischio di recidiva = **si** corrisponde a esito **negativo** (cioè il soggetto ne ha un danno)
- Rischio di recidiva= **no** corrisponde a esito **positivo** (cioè il soggetto ne ha un vantaggio)

Nel caso di concessione di un prestito bancario

- Garanzia di restituzione = si corrisponde a esito **positivo** (cioè il soggetto ha il prestito)
- Garanzia di restituzione = no corrisponde a esito **negativo** (cioè il soggetto non ha il presito)

Presence/absence of a threshold → Predicted outcome, predicted probability and actual outcome

→

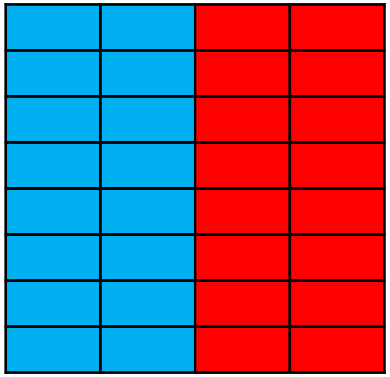


→

Test del Modello decisionale

→

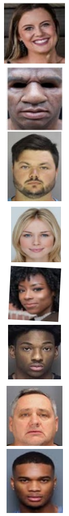
Predicted Probability value	**Actual outcome**
2	No
7	No
3	Si
4	Si
4	No
6	Si
3	Si
6	No



→

Applicazione Modello decisionale

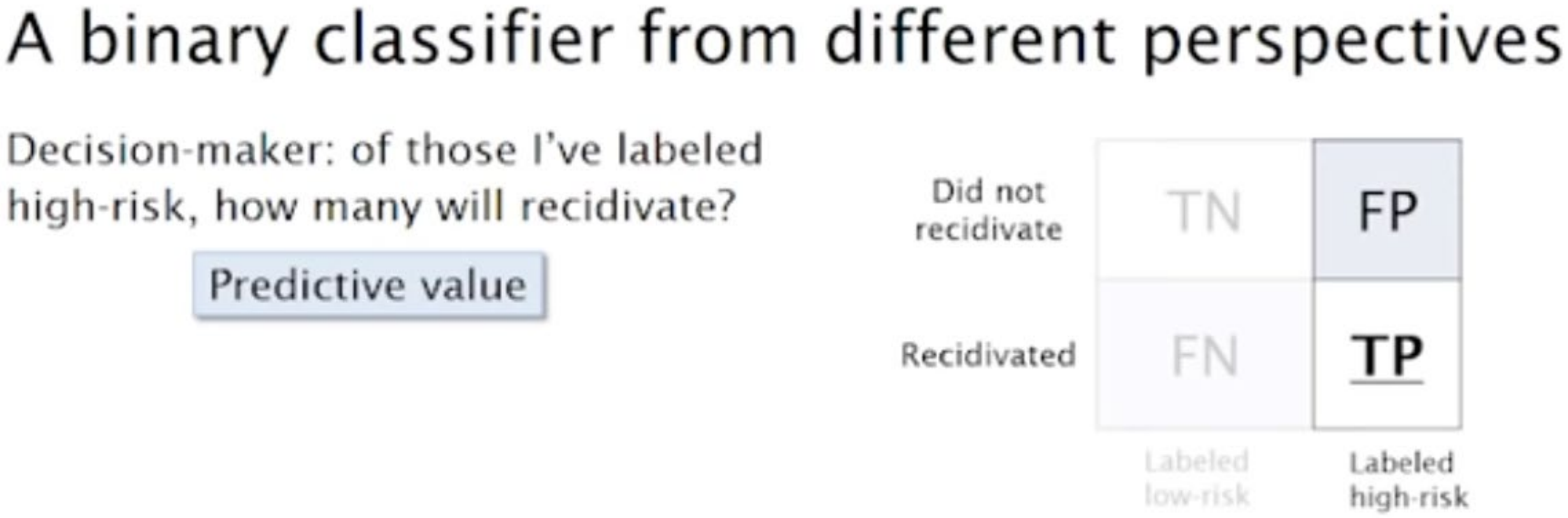
→



Predicted Probability value	**Predicted outcome**
2	No
7	No
3	Si
4	Si
4	No
6	Si
3	Si
6	No

Tornando al tema del rischio di recidiva i punti di vista sono diversi e con diversa utilità, dando luogo a diversi tipi di fairness

https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/

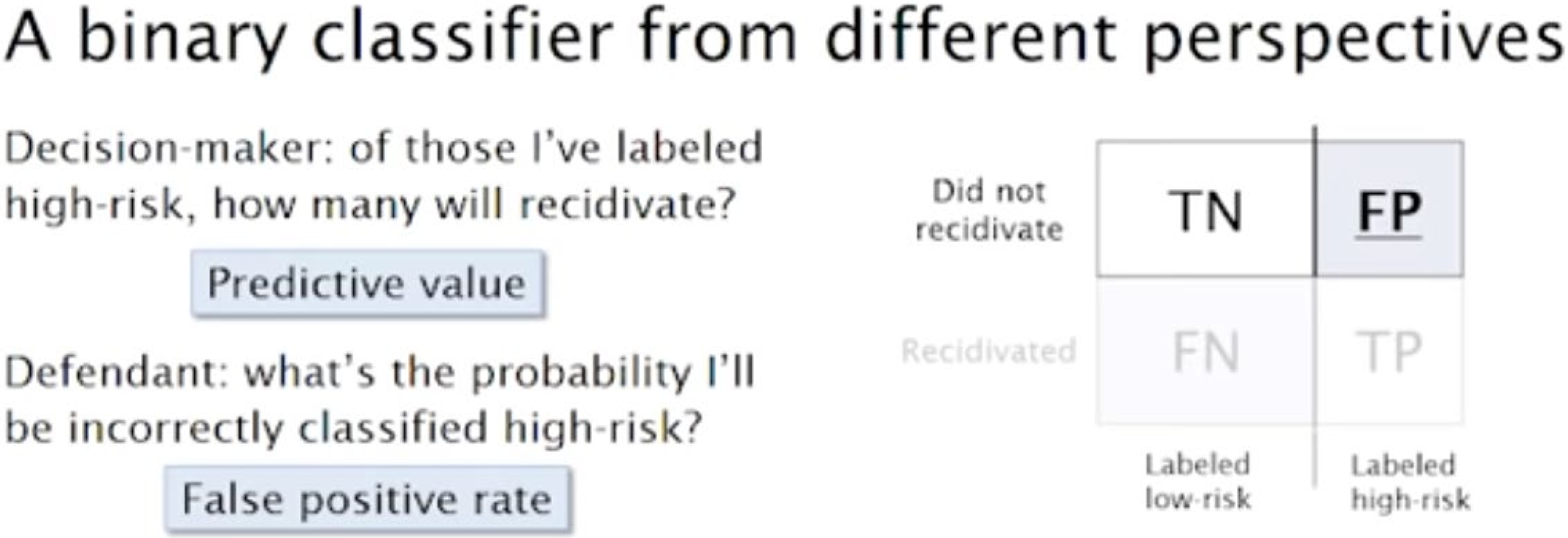


A binary classifier from different perspectives

Decision-maker: of those I've labeled high-risk, how many will recidivate?

Predictive value

	Labeled low-risk	Labeled high-risk
Did not recidivate	TN	FP
Recidivated	FN	TP



A binary classifier from different perspectives

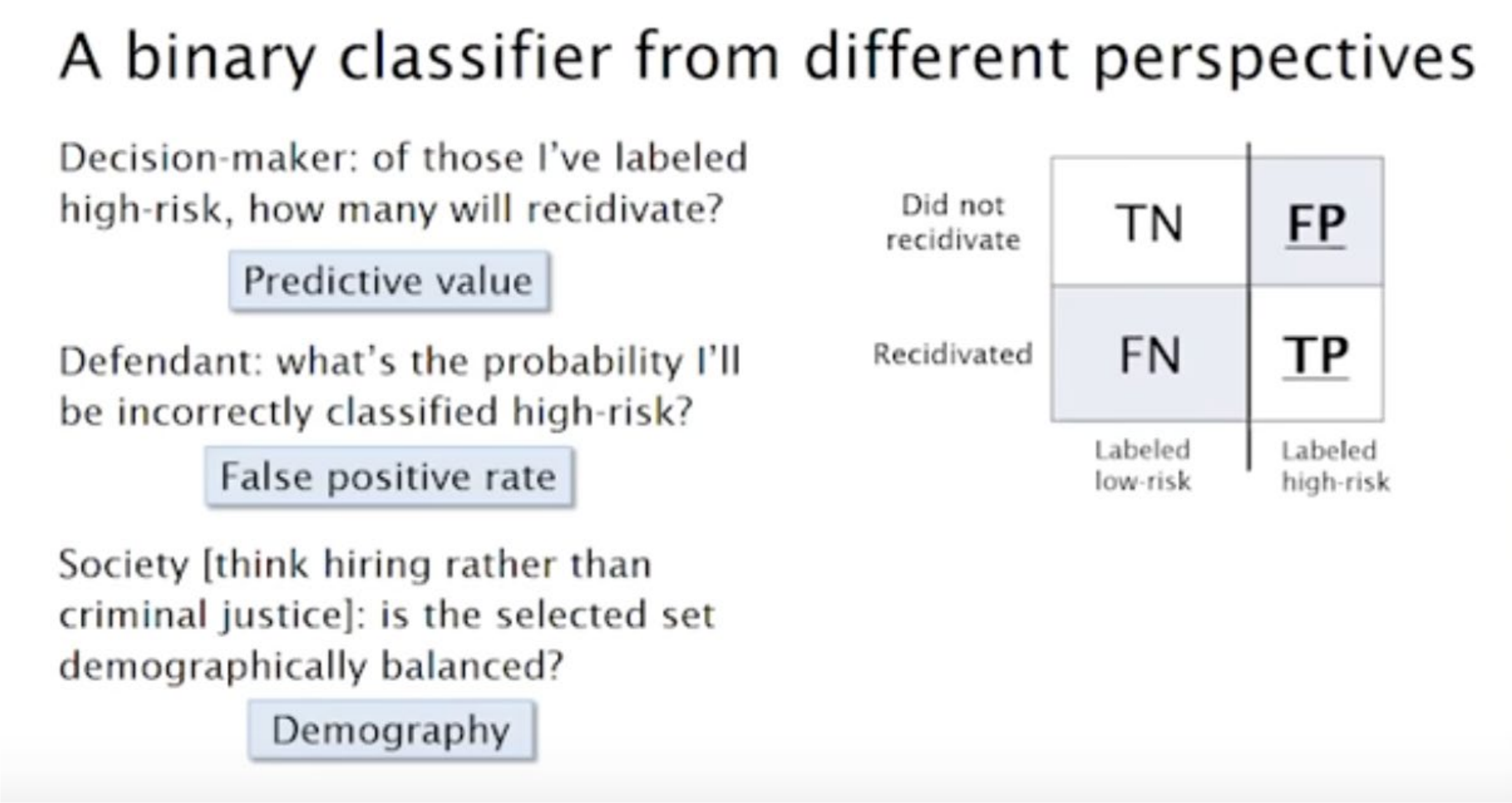
Decision-maker: of those I've labeled high-risk, how many will recidivate?

Predictive value

Defendant: what's the probability I'll be incorrectly classified high-risk?

False positive rate

	Labeled low-risk	Labeled high-risk
Did not recidivate	TN	**FP**
Recidivated	FN	TP



A binary classifier from different perspectives

Decision-maker: of those I've labeled high-risk, how many will recidivate?

Predictive value

Defendant: what's the probability I'll be incorrectly classified high-risk?

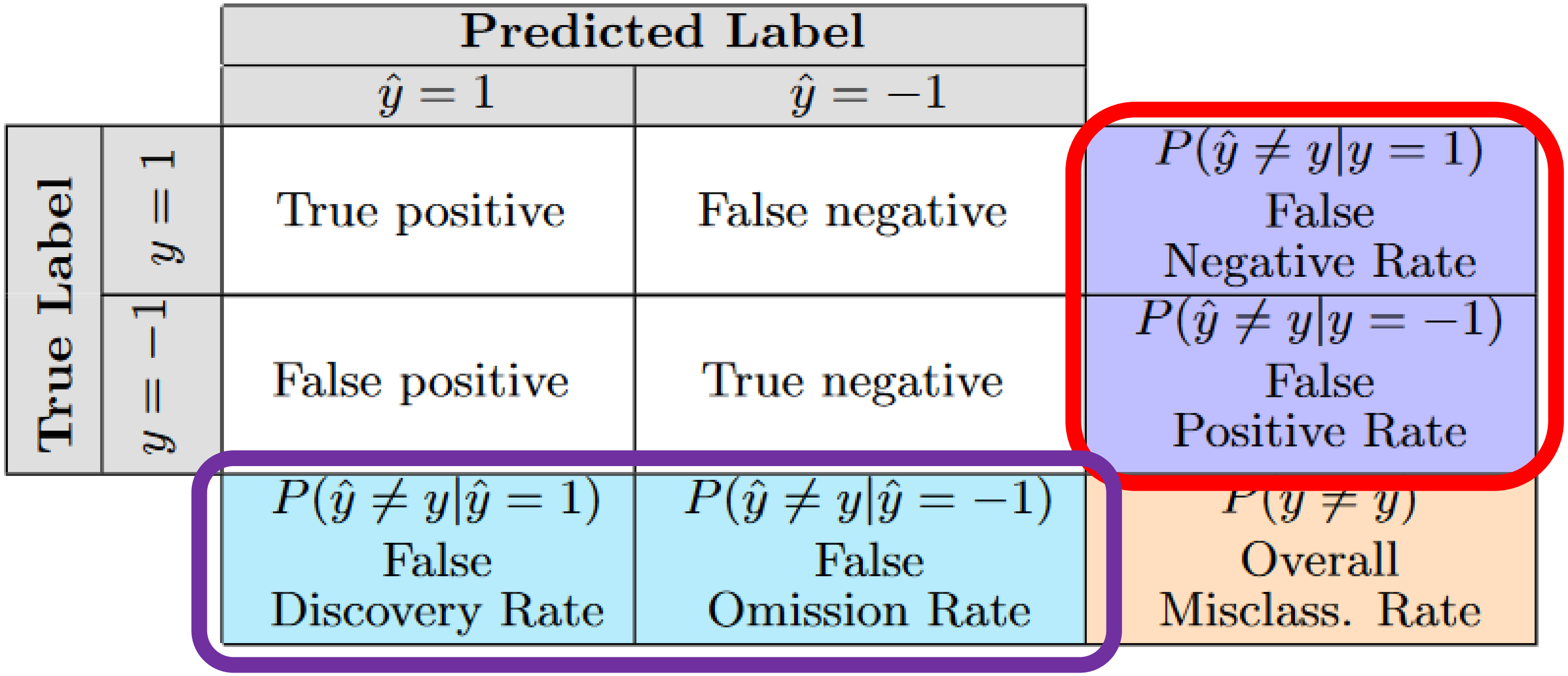
False positive rate

Society [think hiring rather than criminal justice]: is the selected set demographically balanced?

Demography

	Labeled low-risk	Labeled high-risk
Did not recidivate	TN	FP
Recidivated	FN	TP

In addition to the overall misclassification rate, error rates can be measured in two different ways: **false negative rate and false positive rate** are defined as fractions over the class distribution in the ground truth labels, or true labels. On the other hand, **false discovery rate and false omission rate** are defined as fractions over the class distribution in the predicted labels



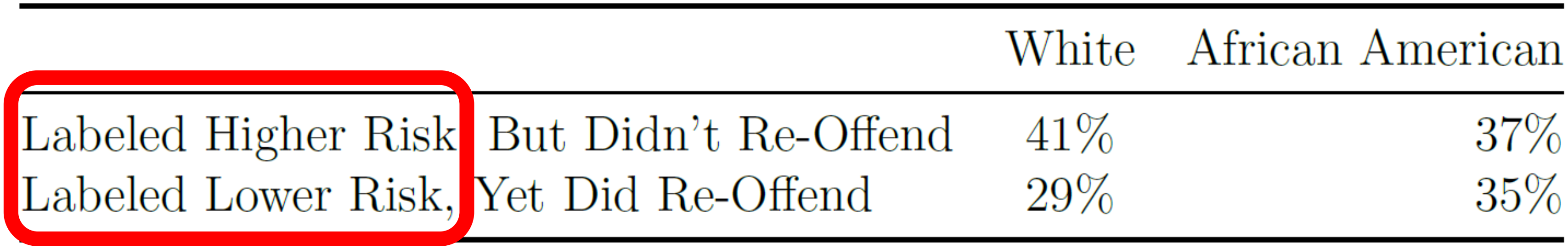
		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y \mid y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

Types of fairness – all together

Categories of fairness	**Types of fairness**
Statistical measures	Basic → True Positive, True Negative, False Positive, False Negative Composite → 1. Positive predictive value TP/(TP+FP) **2. False discovery rate FP/(TP+FP)** **3. False omission rate FN/(TN+FN)** 4. Negative predictive value TN/(TN+FN) 5. True positive rate TP/(TP+FN) or Sensitivity **6. False positive rate FP/(FP+TN)** **7. False negative rate FN/(TP+FN) or Specificity** 8. True negative rate TN/(FP+TN)
Based on Predicted outcome for various demographic distributions of subjects	1. Group fairness/statistical parity/equal acceptance rate/benchmarking 2. Conditional statistical parity
Based on Predicted outcomes that are compared with the Actual Outcomes	1. Predictive parity/outcome test 2. False positive error rate balance/Pedicrtive equality 3. False negative error rate balance/Equal opportunity 4. Equalized odds/conditional procedure accuracy equality/disparate treatment 5. Conditional use accuracy quality 6. Overall accuracy equality 7. Treatment equality
Based on Predicted Probabilities and Actual Outcome	1. Test fairness/calibration/matching conditional frequencies 2. Well calibration 3. Balance for positive class 4. Balance for negative class
Similarity based	1. Causal discrimination 2. Fairness through unawareness 3. Fairness through awareness
Causal reasoning based	1. Counterfactual fairness 2. No unresolved discrimination 3. No proxy discrimination 4. Fair inference

Defs of predictive parity and accuracy equity in the Northpointe response

- A risk scale exhibits **accuracy equity** if it can discriminate recidivists and non-recidivists equally well for two different groups such as blacks and whites.
- The risk scale exhibits **predictive parity** if the classifier obtains similar predictive values for two different groups such as blacks and whites, for example, the probability of recidivating, **given a high risk score**, is similar for blacks and whites.



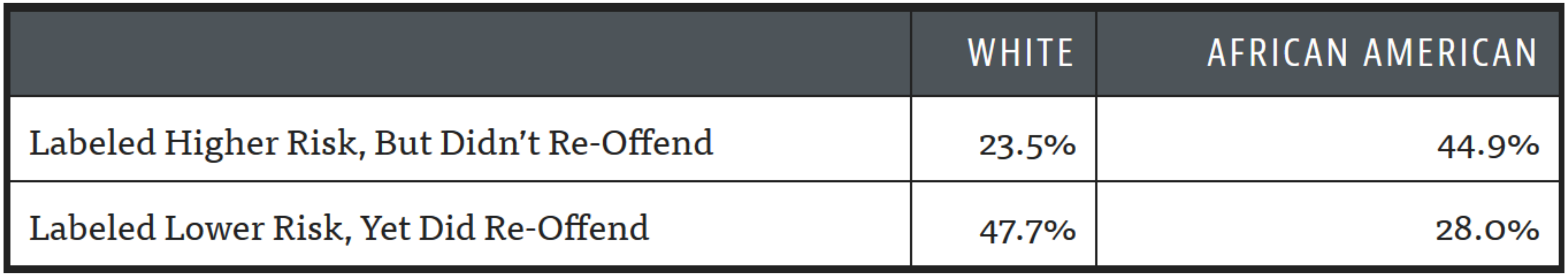
	White	African American
Labeled Higher Risk But Didn’t Re-Offend	41%	37%
Labeled Lower Risk, Yet Did Re-Offend	29%	35%

Northpointe fairnesses

Categories of fairness	Types of fairness
Statistical measures	Basic → True Positive, True Negative, False Positive, False Negative Composite → 1. Positive predictive value TP/(TP+FP) 2. False discovery rate FP/(TP+FP) 3. False omission rate FN/(TN+FN) 4. Negative predictive value TN/(TN+FN) 5. True positive rate TP/(TP+FN) or Sensitivity 6. False positive rate FP/(FP+TN) 7. False negative rate FN/(TP+FN) or Specificity 8. True negative rate TN/(FP+TN)
Based on Predicted outcome for various demographic distributions of subjects	1. Group fairness/statistical parity/equal acceptance rate/benchmarking 2. Conditional statistical parity
Based on Predicted outcomes that are compared with the Actual Outcomes	**1. Predictive parity/outcome test** 2. False positive error rate balance/Predictive equality 3. False negative error rate balance/Equal opportunity 4. Equalized odds/conditional procedure accuracy equality/disparate treatment 5. Conditional use accuracy quality **6. Overall accuracy equality, accuracy eqauity** 7. Treatment equality
Based on Predicted Probabilities and Actual Outcome	1. Test fairness/calibration/matching conditional frequencies 2. Well calibration 3. Balance for positive class 4. Balance for negative class
Similarity based	1. Causal discrimination 2. Fairness through unawareness 3. Fairness through awareness
Causal reasoning based	1. Counterfactual fairness 2. No unresolved discrimination 3. No proxy discrimination 4. Fair inference

While the point of view of Propublica is the point of view of the defendant

1. False positive error rate balance/Predictive equality
2. False negative error rate balance/Equal opportunity



	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%

Propublica fairnesses and Northpointe f

Categories of fairness	Types of fairness
Statistical measures	Basic → True Positive, True Negative, False Positive, False Negative Composite → 1. Positive predictive value TP/(TP+FP) 2. False discovery rate FP/(TP+FP) 3. False omission rate FN/(TN+FN) 4. Negative predictive value TN/(TN+FN) 5. True positive rate TP/(TP+FN) or Sensitivity 6. False positive rate FP/(FP+TN) 7. False negative rate FN/(TP+FN) or Specificity 8. True negative rate TN/(FP+TN)
Based on Predicted outcome for various demographic distributions of subjects	1. Group fairness/statistical parity/equal acceptance rate/benchmarking 2. Conditional statistical parity
Based on Predicted outcomes that are compared with the Actual Outcomes	**1. Predictive parity/outcome test** **2. False positive error rate balance/Predictive equality** **3. False negative error rate balance/Equal opportunity** 4. Equalized odds/conditional procedure accuracy equality/disparate treatment 5. Conditional use accuracy quality **6. Overall accuracy equality** 7. Treatment equality
Based on Predicted Probabilities and Actual Outcome	1. Test fairness/calibration/matching conditional frequencies 2. Well calibration 3. Balance for positive class 4. Balance for negative class
Similarity based	1. Causal discrimination 2. Fairness through unawareness 3. Fairness through awareness
Causal reasoning based	1. Counterfactual fairness 2. No unresolved discrimination 3. No proxy discrimination 4. Fair inference

Definitions of Types of fairness – all together

Types of fairness: based on predicted value and actual value	Definition
Basic → True Positive, True Negative, False Positive, False Negative Composite → 1. PPV -Positive predictive value TP/(TP+FP) (Precision) 2. FDR - False discovery rate FP/(TP+FP) 3. FOR - False omission rate FN/(TN+FN) 4. NPV - Negative predictive value TN/(TN+FN) 5. TPR - True positive rate TP/(TP+FN) 6. FPR - False positive rate FP/(FP+TN) 7. FNR - False negative rate FN/(TP+FN) 8. TNR - True negative rate TN/(FP+TN)	1 probability of a subject with positive v. to truly belong to the pos.class 2 fraction of negative cases incorrectly pred. to be in the positive class 3. Probability of a positive case to be incorrectly rejected. 4. Probability of a subj.with negative pred. to truly belong to the neg. cl. 5. Probability of a truly positive subj.to be identified as such 6. Probability of falsely accepting a negative case 7. probability of a negative result for an actually positive subject 8. probability of a subj. from the neg. class to be assigned to the neg. class
1. Group fairness/statistical parity/equal acceptance rate/benchmarking 2. Conditional statistical parity	1. subjects in both protected and unprot. groups have equal prob. of being assigned to the pos. pred. class. 2. subjects in both prot,/unprot.groups have equal prob. of being ass. to pos. pre. class, controlling for a set of legitimate factors L.
1. Predictive parity/outcome test 2. False positive error rate balance/Pedicrtive equality 3. False negative error rate balance/Equal opportunity 4. Equalized odds/conditional procedure acc. equal./disparate treatment 5. Conditional use accuracy quality 6. Overall accuracy equality 7. Treatment equality	1. both protected and unprotected groups have equal PPV 2. Both protected and unprotected groups have equal FPR 3. Both protected and unprotected groups have equal FNR 4. protected and unprotected groups have equal TPR 5. this definition conjuncts two conditions: equal PPV and NPV 6. both protected and unprotect. groups have equal prediction accuracy 7. both prot. and unprot. groups have equal ratio of false neg. and false pos.
1. Test fairness/calibration/matching conditional frequencies 2. Well calibration 3. Balance for positive class 4. Balance for negative class	1. subj. in prot. and unpr.. gr. have equal prob. to truly belong to the pos. cl. 2. for any pred. prob. score S, subj. in prot. and unprot. gr. should have an equal prob. to truly belong to the pos. class and this prob. should be eq. to S. 3. subj const. pos. class from prot. and unprot. gr. have = aver pred. prob. Sc. S 4. flipped version of the previous definition, truly → falsely
1. Causal discrimination 2. Fairness through unawareness 3. Fairness through awareness	1.Class. produces the same classific. for any two subj. with exact same attr X. 2. no sensitive attributes are explicitly used in the decision-making process 3. similar individuals via a distance metric should have similar classification.

Definitions of Types of fairness – Statistical measures

Types of fairness: based on predicted value and actual value	Definition
Basic → True Positive, True Negative, False Positive, False Negative Composite →	
1. PPV -Positive predictive value TP/(TP+FP) (Precision, Target population error, ***Decision maker*** POW)	1 probability of a subject with positive value to truly belong to the positive class
2. FDR - False discovery rate FP/(TP+FP)	2 fraction of negative cases incorrectly predicted to be in the positive class
3. FOR - False omission rate FN/(TN+FN)	3. Probability of a positive case to be incorrectly rejected.
4. NPV - Negative predictive value TN/(TN+FN)	4. Probability of a subject with negative pred. to truly belong to the negative class
5. TPR - True positive rate TP/(TP+FN) (Recall, Sensitivity, Avoiance of false negative, model error)	5. Probability of a truly positive subject to be identified as such
6. FPR - False positive rate FP/(FP+TN) (Model Error, ***Defendant*** point of view)	6. Probability of falsely accepting a negative case
7. FNR - False negative rate FN/(TP+FN) (Complement of sensitivity, Model error)	7. Probability of a negative result for an actually positive subject
8. TNR - True negative rate TN/(FP+TN) (Specificity, Avoidance of false positive, Model error)	8. probability of a subj. from the negative class to be assigned to the negative class

Definitions of Types of fairness - Based on Predicted outcome for various demographic distributions of subjects

Types of fairness: based on predicted value and actual value	Definition
1. Group fairness/statistical parity/equal acceptance rate/benchmarking/ demographic parity/four-fifth rule/ ***Equal impact*** /Equal outcome/Independence (Barocas) → ***Society*** *POW*	1. subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class.
1. Conditional statistical parity	2. Subjects in both protected /unprotected groups have equal probabilities of being assigned to positive predictive class, controlling for a set of legitimate factors L.

Definitions of Types of fairness - Based on Predicted outcomes that are compared with the Actual Outcomes

Types of fairness: based on predicted value and actual value	Definition
1. Predictive parity/outcome test/Classification parity/Predictive performance/Calibration/Test-fairness/Matching conditional frequencies	1. Both protected and unprotected groups have equal Positive predictive value – Predictive performance is equal among groups defined by the protected attributes or some measure of class. In the pretrial context, calibration means that among defendants with a given risk score, the proportion who would reoffend if released is thesame across race groups
2. False positive error rate balance/Predictive equality	2. Both protected and unprotected groups have equal False positive rate
3. False negative error rate balance/Equal opportunity	3. Both protected and unprotected groups have equal False negative rate
4. Equalized odds/conditional procedure acc. equal./***Disparate treatment***/ Unawareness/ Positive rate parity (A weaker notion **is** Accuracy Parity in which we can trade false positive rate of one group for false negative rate of another group)	4. Protected and unprotected groups have equal true positive rate
5. Conditional use accuracy quality	5. This definition conjuncts two conditions: equal Positive predictive value and Negative predictive value
6. Overall accuracy equality	6. Both protected and unprotect. groups have equal prediction accuracy
7. Treatment equality	7. Both protected and unprotected groups have equal ratio of false negative and false positive

Defs of predictive parity and accuracy equity in the northpointe response

- A risk scale exhibits **accuracy equity** if it can discriminate recidivists and non-recidivists equally well for two different groups such as blacks and whites.
- The risk scale exhibits **predictive parity** if the classifier obtains similar predictive values for two different groups such as blacks and whites, for example, the probability of recidivating, **given a high risk score**, is similar for blacks and whites.

Definitions of Types of fairness - Based on Predicted Probabilities and Actual Outcome

Types of fairness: based on predicted value and actual value	Definition
1. Test fairness/calibration/matching conditional frequencies 2. Well calibration 3. Balance for positive class 4. Balance for negative class	1. subjects in protected and unprotected groups have equal probability to truly belong to the positive class 2. for any predicted probability score S, subjects in protected and unprotected groups should have an equal probability to truly belong to the positive class and this probability should be equal to S. 3. Subjects constituting positive class from protected and unprotected groups have equal average predictive probability Score S 4. flipped version of the previous definition, truly → falsely

Definitions of Types of fairness - Similarity based

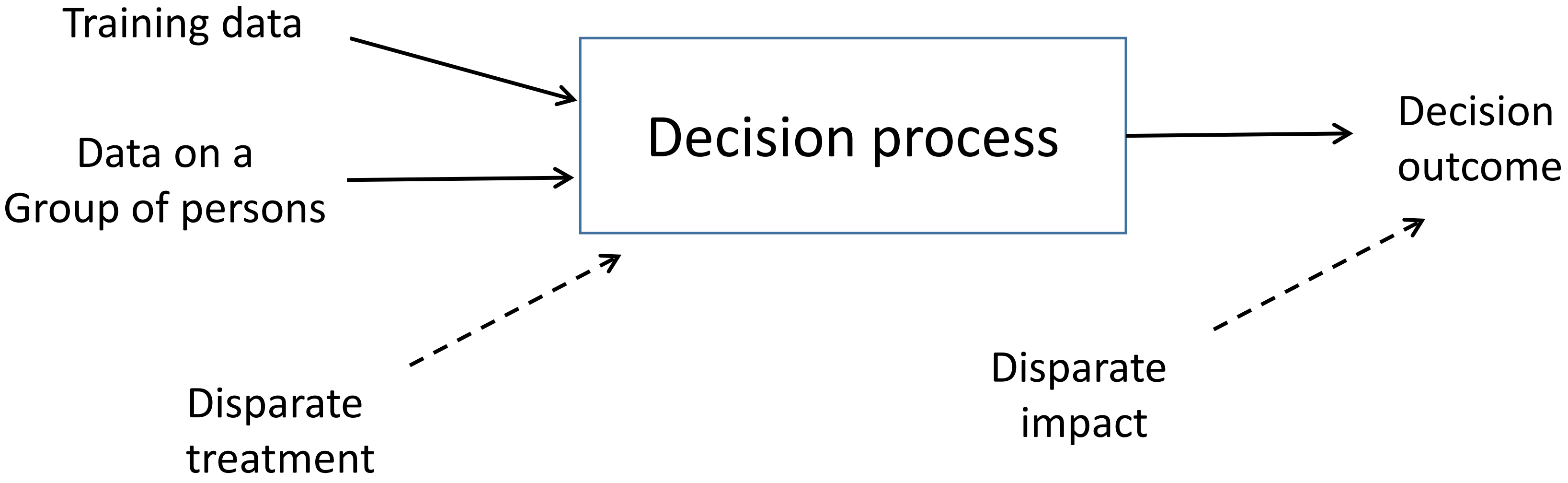
Types of fairness: based on predicted value and actual value	Definition
1. Causal discrimination 2. Fairness through unawareness (*Avoid disparate treatment*, Anti-classification) 3. Fairness through awareness	1.Classifier produces the same classification for any two subjects with exact same values for protected attribute X. 2. no sensitive attributes are explicitly used in the decision-making process 3. similar individuals via a distance metric should have similar classification.

Types of fairness definitions - Causal reasoning based in detail

Types of fairness	Definition
Causal graphs are used for building fair classifiers and other ML algorithms. Specifically, the relations between attributes and their influence on outcome is captured by a set of structural equations which are further used to provides methods to estimate effects of sensitive attr.s and build algorithms that ensure a tolerable level of discrimination due to these attr.s. E.g. a graph consists of the protected attr. G, the credit amount, employment length, and credit history attr.s, and the predicted outcome d. In causal graphs, a proxy attr. is an attr. whose value can be used to derive a value of another attr.. In our example, we assume that employment length acts as a proxy attr. for G: one can derive the applicants' gender from the length of their employment. A resolving attr. is an attr, in the causal graph that is influenced by the protected attr. in a non-discriminatory manner. In our example, the effect of G on the credit am. is nondiscriminatory, which means that the differences in credit amount for diff. values of G are not considered as discrimination. Hence, the credit amount acts as a resolving attr. for G in this graph	
1. Counterfactual fairness	1.A causal graph is counterfactually fair if the predicted outcome *d* in the graph does not depend on a descendant of the protected attribute *G*
2. No unresolved discrimination	*2*. A causal graph has no unresolved discrimination if there exists no path from the protected attribute G to the predicted outcome d, except via a resolving variable
1. No proxy discrimination	3.A causal graph is free of proxy discrimination if there exists no path from the protected attribute G to the predicted outcome d that is blocked by a proxy variable
4. Fair inference	4. This def. classifies paths in a causal graph as legitimate or illegitimate. For ex,,it might make sense to consider the employment length for making credit related decis.. Even though the empl.length acts as a proxy for G, that path would be consid. as legitimate. A causal graph satisfies the notion of fair inference if there are no illegitimate paths from G to d, which is not the case in our example as there exist another illegitimate path, via credit amount

Laws grounded on disparate treatment and disparate impact

- A **decision making process** suffers from **disparate treatment** if its **decisions are (partly) based** on the subject's sensitive attribute information, and
- It has **disparate impact** if its **outcomes** disproportionately hurt (or, benefit) people with certain sensitive attribute values (e.g., females, blacks).



Training data

Data on a Group of persons

Decision process

Decision outcome

Disparate treatment

Disparate impact

Example

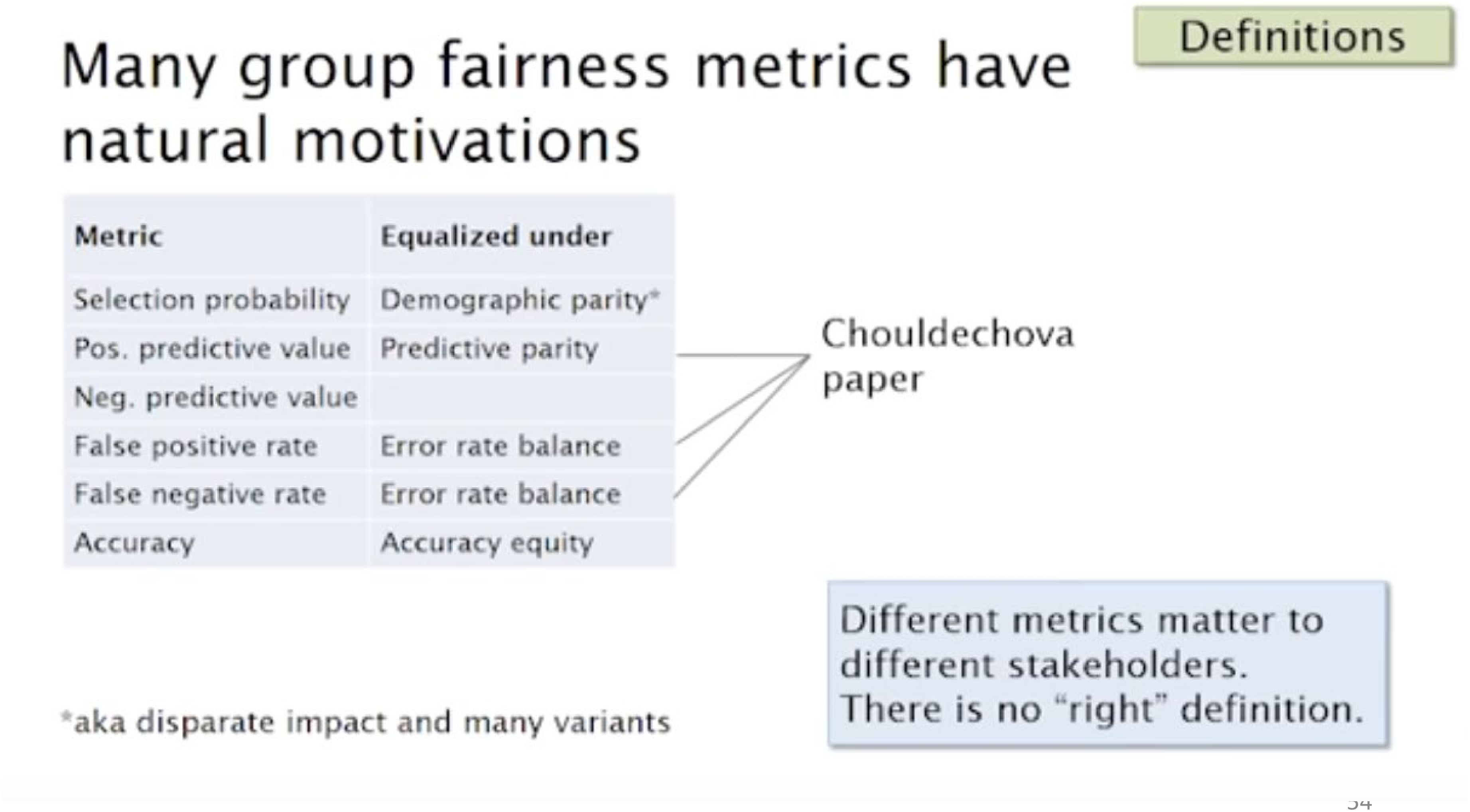
Disparate Treatment

- If only African American applicants are required to take a pre-employment assessment test.

Disparate Impact

- If you test all applicants and only African Americans are eliminated based on the results of the assessment, since historical data used as training data are biased.

Insomma……



Definitions

Many group fairness metrics have natural motivations

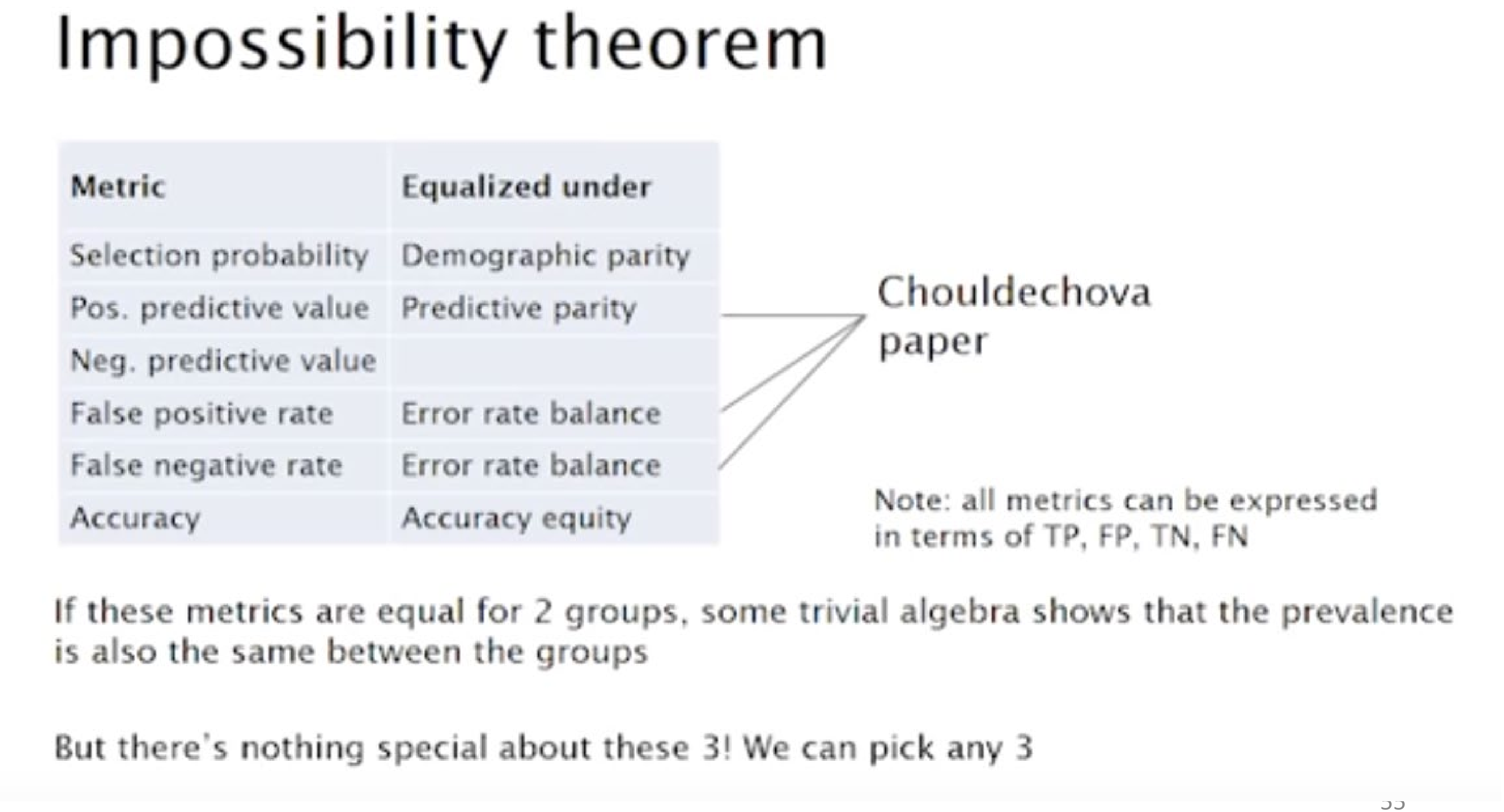
Metric	Equalized under
Selection probability	Demographic parity*
Pos. predictive value	Predictive parity
Neg. predictive value	
False positive rate	Error rate balance
False negative rate	Error rate balance
Accuracy	Accuracy equity

Chouldechova paper

Different metrics matter to different stakeholders. There is no "right" definition.

*aka disparate impact and many variants

Teorema di impossibilità



Impossibility theorem

Metric	Equalized under
Selection probability	Demographic parity
Pos. predictive value	Predictive parity
Neg. predictive value	
False positive rate	Error rate balance
False negative rate	Error rate balance
Accuracy	Accuracy equity

Chouldechova paper

Note: all metrics can be expressed in terms of TP, FP, TN, FN

If these metrics are equal for 2 groups, some trivial algebra shows that the prevalence is also the same between the groups

But there's nothing special about these 3! We can pick any 3

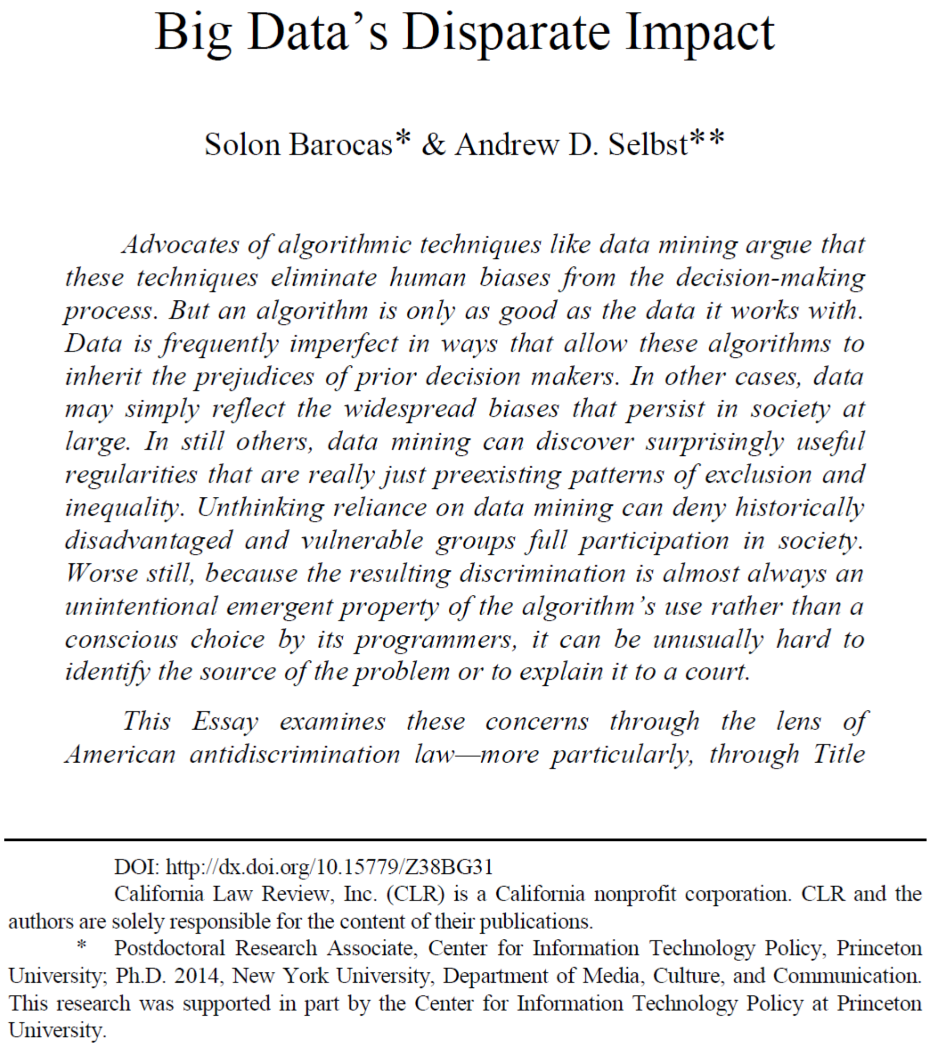
Se poi andiamo su Wikipedia, altra crisi di nervi…

(1) Historical Bias
(2) Representation Bias
(3) Measurement Bias.
(4) Evaluation Bias.
(5) Aggregation Bias.
(6) Population Bias.
(7) Simpson’s Paradox
(8) Longitudinal Data Fallacy.
(9) Sampling Bias.
(10) Behavioral Bias
(11) Content Production Bias.
(12) Linking Bias.
(13) Temporal Bias
(14) Popularity Bias.
(15) Algorithmic Bias.
(16) User Interaction Bias.
(17) Social Bias.
(18) Emergent Bias.
(19) Self-Selection Bias.
(20) Omitted Variable Bias.
(21) Cause-Effect Bias
(22) Observer Bias.
(23) Funding Bias.
(24) Presentation Bias.
(25) Ranking Bias.

4.3 Che fare riguardo alla fairness?

2018 - Libro su fairness

2016, 60 pagine



Big Data's Disparate Impact

Solon Barocas* & Andrew D. Selbst**

Advocates of algorithmic techniques like data mining argue that these techniques eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with. Data is frequently imperfect in ways that allow these algorithms to inherit the prejudices of prior decision makers. In other cases, data may simply reflect the widespread biases that persist in society at large. In still others, data mining can discover surprisingly useful regularities that are really just preexisting patterns of exclusion and inequality. Unthinking reliance on data mining can deny historically disadvantaged and vulnerable groups full participation in society. Worse still, because the resulting discrimination is almost always an unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers, it can be unusually hard to identify the source of the problem or to explain it to a court.

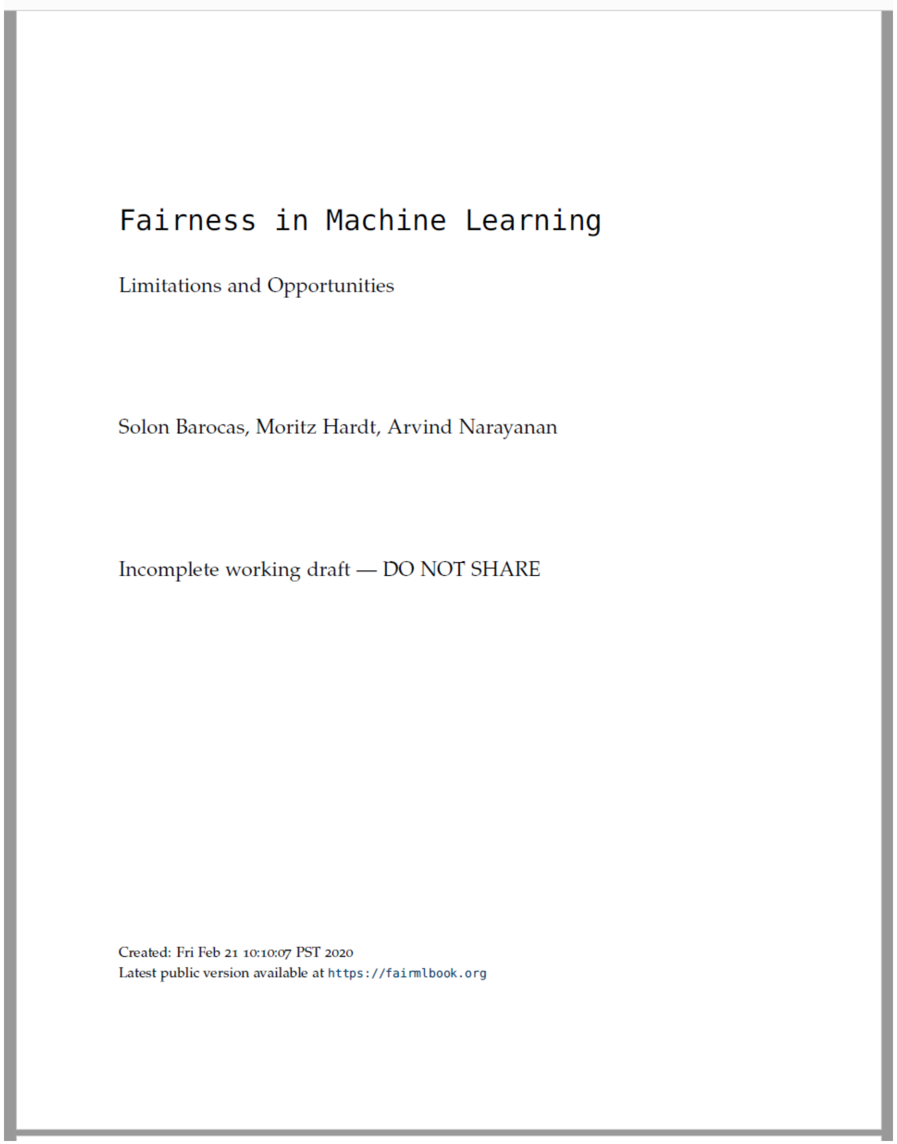
This Essay examines these concerns through the lens of American antidiscrimination law—more particularly, through Title

DOI: http://dx.doi.org/10.15779/Z38BG31

California Law Review, Inc. (CLR) is a California nonprofit corporation. CLR and the authors are solely responsible for the content of their publications.

* Postdoctoral Research Associate, Center for Information Technology Policy, Princeton University; Ph.D. 2014, New York University, Department of Media, Culture, and Communication. This research was supported in part by the Center for Information Technology Policy at Princeton University.

2019 – 180 pages



Fairness in Machine Learning

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

Incomplete working draft — DO NOT SHARE

Created: Fri Feb 21 10:10:07 PST 2020
Latest public version available at https://fairmlbook.org

1. Assess and Compare approaches

2017 Assessment of various discrimination measures

- Society is increasingly relying on data-driven predictive models for automated decision making. This is not by design, but **due to the nature and noisiness of observational data, such models may systematically disadvantage people belonging to certain categories or groups, instead of relying solely on individual merits. This may happen even if the computing process is fair and well-intentioned.**
- Discrimination aware data mining studies of how to make predictive models free from discrimination, when the historical data, on which they are built, may be biased, incomplete, or even contain past discriminatory decisions.
- Discrimination-aware data mining is an emerging research discipline, and there is no firm consensus yet of how to measure the performance of algorithms.

2017 Assessment of various discrimination measures

- The goal of this survey **is to review various discrimination measures that have been used, analytically and computationally analyze their performance, and highlight implications of using one or another measure**.
- We also describe measures from other disciplines, which have not been used for measuring discrimination, but potentially could be suitable for this purpose.
- This survey is primarily intended for researchers in data mining and machine learning as a step towards producing a unifying view of performance criteria when developing new algorithms for non-discriminatory predictive modeling. In addition, practitioners and policy makers could use this study when diagnosing potential discrimination by predictive models

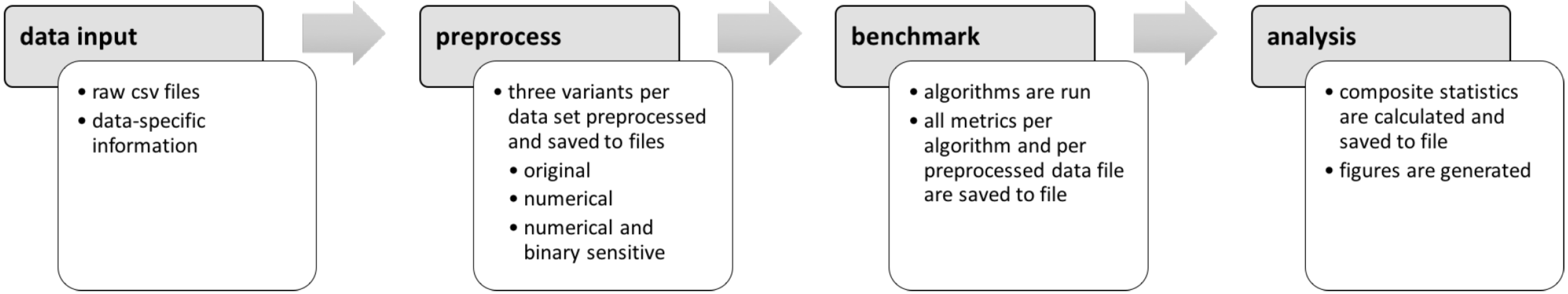
Benchmark organization - 1

- In order to provide a platform for clear comparison of results across fairness-aware machine learning algorithms,we separate each stage of the learning and analysis process (see →) and ensure that each algorithm is compared using the same dataset (including the same preprocessing), the same set of training / test splits, and all desired fairness and accuracy measures.
- Much previous work has combined the preprocessing for a specific dataset with the code for the fairness-aware algorithm, which makes comparisons with other algorithms and other datasets difficult.
- Similarly, algorithms have often been analyzed only under one or two measures. Here, we distinguish preprocessing, algorithms, and measures, and create a pipeline in which all algorithms are analyzed under a standard preprocessing of datasets and a large set of measures.

Benchmark organization - 2

- In order to encourage easy adoption of this codebase as a platform for future algorithmic analysis, each of previous choices is modularized so that adding new datasets, measures, and/or algorithms to the pipeline is as easy as creating a new object.
- The pipeline will then ensure that all existing algorithms are evaluated under the new dataset and measure.
- More details and instructions for adding to the code base can be found at the repository.

The stages of the fairness-aware benchmarking program - Intermediate files are saved at each stage of the pipeline to ensure reproducibility



data input

- raw csv files
- data-specific information

preprocess

- three variants per data set preprocessed and saved to files
  - original
  - numerical
  - numerical and binary sensitive

benchmark

- algorithms are run
- all metrics per algorithm and per preprocessed data file are saved to file

analysis

- composite statistics are calculated and saved to file
- figures are generated

Benchmarking and comparison of techniques

- We present the results of an open benchmark we have developed that lets us compare a number of different algorithms under a variety of fairness measures and existing datasets.
- We find that although different algorithms tend to prefer specific formulations of fairness preservations, many of these measures strongly correlate with one another. In addition, we find that fairness-preserving algorithms tend to be sensitive to fluctuations in dataset composition (simulated in our benchmark by varying training-test splits) and to different forms of preprocessing, indicating that fairness interventions might be more brittle than previously thought

Other Results on Assessment
(skip at first reading)

Fairness-accuracy tradeoffs depend on preprocessing

Different algorithms tend to have slightly different requirements in terms of input: how are sensitive attributes encoded? Are multiple sensitive attributes supported? Does the algorithm directly support categorical attributes or are attribute transformations required?

Choices for these requirements directly affect the accuracy and fairness of a fairness-aware classifier. This is significant because prior formal studies of fairness-accuracy tradeoffs typically focused on hyperparameter tuning, rather than preprocessing.

Measures of discrimination correlate with each other

- Even though there has been a proliferation of measures designed to highlight discrimination instances by machine learning algorithms, we find that a large number of these measures tend to strongly correlate with one another. As a result, techniques optimizing for one measure could perform well for a different measure (and similarly for poor performance).

Algorithms make significantly different fairness-accuracy tradeoffs

The specific mechanisms that different algorithms employ to increase fairness are quite varied, but surprisingly, the actual predictions made by these algorithms tend to vary significantly as well. As a result, no algorithm’s performance (as of the latest state of our benchmark) appears to dominate, either in accuracy or fairness measures.

Algorithms are fragile

they are sensitive to variations in the input. We find surprising variability in fairness measures arising from variations in training-test splits; this appears to not have been previously mentioned in the literature.

Assessing bias In **health care management programs** namely, assessing disparate treatment and disparate impact

- A single algorithm drives an important health care decision for over 70 million people in the US. When health systems anticipate that a patient will have especially complex and intensive future health care needs, **she is enrolled in a 'care management' program, which provides considerable additional resources**: greater attention from trained providers and help with coordination of her care.
- **To determine which patients will have complex future health care needs, and thus benefit from program enrollment, many systems rely on an algorithmically generated commercial risk score**. In this paper, we exploit a rich dataset to study racial bias in a commercial algorithm that is deployed nationwide today in many of the US's most prominent Accountable Care Organizations (ACOs).
- **We document significant racial bias in this widely used algorithm, using data on primary care patients at a large hospital**. Blacks and whites with the same algorithmic risk scores have very different realized health. For example, the highest-risk black patients (those at the threshold where patients are auto-enrolled in the program), have significantly more chronic illnesses than white enrollees with the same risk score.

Assessing bias In **health care management programs** namely, assessing disparate treatment and disparate impact

- We use detailed physiological data to show the pervasiveness of the bias: across a range of biomarkers, from HbA1c levels for diabetics to blood pressure control for hypertensives, we find significant racial health gaps conditional on risk score. This bias has significant material consequences for patients: it effectively means that white patients with the same health as black patients are far more likely be enrolled in the care management program, and benefit from its resources. If we simulated a world without this gap in predictions, blacks would be auto-enrolled into the program at more than double the current rate.
- An unusual aspect of our dataset is that we observe not just the risk scores but also the input data and objective function used to construct it. This provides a unique window into the mechanisms by which bias arises. The algorithm's predicted risk of developing complex health needs is thus in fact predicted costs. And by this metric, one could easily call the algorithm unbiased: costs are very similar for black and white patients with the same risk

2. Conceive methods, algorithms and frameworks to improve fairness

Types of fair ml algorithms and their types

Fairness-aware machine learning algorithms seek to provide methods under which the predicted outcome of a classifier operating on data about people is fair or non-discriminatory for people based on their protected class status such as race, sex, religion, etc., also known as a sensitive attribute.

Broadly, **fairness-aware machine learning algorithms have been categorized as**

1. **those preprocessing techniques designed to modify the input data so that the outcome of any machine learning algorithm applied to that data will be fair,**
2. **those algorithm modification techniques that modify an existing algorithm or create a new one that will be fair under any inputs, and**
3. **those postprocessing techniques that take the output of any model and modify that output to be fair.**

Preprocessing algorithms

- The motivation behind preprocessing algorithms is the idea that **training data is the cause of the discrimination that a machine learning algorithm might learn, and so modifying it can keep a learning algorithm trained on it from discriminating**.
- **This could be because the training data itself captures historical discrimination** or because there are more subtle patterns in the data, such as an under-representation of a minority group, that makes errors on that group both more likely and less costly under certain accuracy measures
- One such algorithm modifies each attribute so that the marginal distributions based on the subsets of that attribute with a given sensitive value are all equal; it does not modify the training labels.

Algorithm modifications

Modifications to specific learning algorithms, e.g., in the form of additional constraints, have been by far the most common approach. We study three such methods in this paper.

- Kamishima et al. introduce a fairness focused regularization term and apply it to a logistic regression classifier.
- Zafar et al. observe that standard fairness constraints are nonconvex and hard to satisfy directly and introduce a convex relaxation for purpose of optimization.
- Calders and Verwer build separate models for each value of a sensitive attribute and use the appropriate model for inputs with the corresponding value of the attribute.

Postprocessing techniques

A third approach to building fairness into algorithm design is by modifying the results of a previously trained classifier to achieve the desired results on different groups.

Kamiran et al. designed a strategy to modify the labels of leaves in a decision tree after training in order to satisfy fairness constraints.

Recent work explored the use of post-processing as a way to ensure fairness with respect to error profiles

Frameworks

A **framework for fair classification** comprising

(1) a (hypothetical) **task-specific metric for determining the degree to which individuals are similar with respect to the classification task at hand**;

(2) an **algorithm for maximizing utility subject to the fairness constraint**, that similar individuals are treated similarly.

Fairness as an optimization problem

- We propose a learning algorithm for fair classification that achieves both
  1. **group fairness** (the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole), and
  2. **individual fairness** (similar individuals should be treated similarly).
- **We formulate fairness as an optimization problem of finding a good representation of the data with two competing goals: to encode the data as well as possible, while simultaneously obfuscating any information about membership in the protected group**.
- We show positive results of our algorithm relative to other known techniques, on three datasets.
- More- over, we demonstrate several advantages to our approach.
- First, our intermediate representation can be used for other classification tasks (i.e., transfer learning is possible) secondly, we take a step toward learning a distance metric which can find important dimensions of the data for classification.

2016 – Optimal adjustment of any learned predictor

- We propose a **criterion for discrimination against a specified sensitive attribute in supervised learning, where the goal is to predict some target based on available features**.
- **Assuming data about the predictor, target, and membership in the protected group are available, we show how to optimally adjust any learned predictor so as to remove discrimination according to our definition**.
- Our framework also improves incentives by shifting the cost of poor classification from disadvantaged groups to the decision maker, who can respond **by improving the classification accuracy**

2018 Focus on judgments about properties the data would satisfy in an “unbiased” world

- Correcting for data bias generally seems to require knowledge of how the measurement process is biased, or judgments about properties the data would satisfy in an “unbiased” world.
- [FSV16] **formalize this as a disconnect between the observed space—features that are observed in the data, such as SAT scores—and the unobservable construct space—features that form the desired basis for decision making, such as intelligence**.
- **Within this framework, data correction efforts attempt to undo the effects of biasing mechanisms that drive discrepancies between these spaces.**

2018 Recommendation Independence

- This paper studies **a recommendation algorithm whose outcomes are not influenced by specified information**. It is useful in contexts potentially unfair decision should be avoided, such as job-applicant recommendations that are not influenced by socially sensitive information.
- An algorithm that could exclude the influence of sensitive information would thus be useful for job-matching with fairness.
- We call the condition between a recommendation outcome and a sensitive feature **recommendation Independence**, which is formally defined as **statistical independence between the outcome and the feature.**
- In this paper, **we develop new methods that can deal with the second moment, i.e., variance, of recommendation outcomes without increasing the computational complexity. These methods can more strictly remove the sensitive infor- mation, and experimental results demonstrate that our new algorithms can more effectively eliminate the factors that undermine fairness.**

3. Contributions from the data management community mainly, Serge Abiteboul

Abiteboul - Address the full data life cycle

- The machine learning and data mining research communities are actively working on methods for enabling fairness of specific algorithms and their outputs, with a particular focus on classification problems
- While important, these approaches focus solely on the final step in the data science lifecycle, and are thus limited by the assumption that input datasets are clean and reliable.
- Data-driven algorithmic decision making usually requires multiple pre-processing stages to address messy input and render it ready for analysis
- This pre-processing, which includes data cleaning, integration, querying and ranking, is often the source of algorithmic bias, and so reasoning about sources of bias, and mitigating unfairness upstream from the final step of data analysis, is potentially more impactful.

Focus on the data management life cycle and on a pragmatic approach to fairness

- It is easy to construct examples that show how bias may be introduced during **data cleaning, data integration, querying, and ranking** — upstream from the final stage of data analysis. Therefore, it is meaningful to detect and mitigate these effects in the data lifecycle stages in which they occur.
- **But, different notions of fairness cannot be enforced simultaneously,** and so **require explicit trade-offs**
- **Fairness is a subjective, context-dependent and highly politicized concept**; **a global consensus on what is fair is unlikely to emerge, in the context of algorithmic decision making or otherwise**.
- **That being said, a productive way to move forward in the data science context is to develop methods that can be instrumented with different alternative fairness notions, and that can support principled and transparent trade-offs between these notions**

Concerns with specific measures

- For example, much research goes into ensuring **statistical parity** — a requirement that the demographics of those receiving a particular outcome, (e.g., a positive or negative classification), are identical to the demographics of the population as a whole.
- Suppose that the input to a binary classifier contains 900 men and 100 women, but that it is known that women represent 50% of the over-all population, and so achieving statistical parity amounts to enforcing a 50-50 gender balance among the positively classified individuals.
- That is, all else being equal, a woman in the input to the classifier is far more likely to receive a positive classification than a man.

Abiteboul – Conclusions

- It is easy to construct additional examples that show how bias may be introduced during data cleaning, data integration, querying, and ranking — upstream from the final stage of data analysis. Therefore, it is meaningful to detect and mitigate these effects in the data lifecycle stages in which they occur.
- Members of the data management community who are interested in this topicmay consider a growing body of work on impossibility results,which show that **different notions of fairness cannot be enforced simultaneously,** and so **require explicit trade-offs**
- These are not negative results per se, nor are they surprising.

Abiteboul – Conclusions

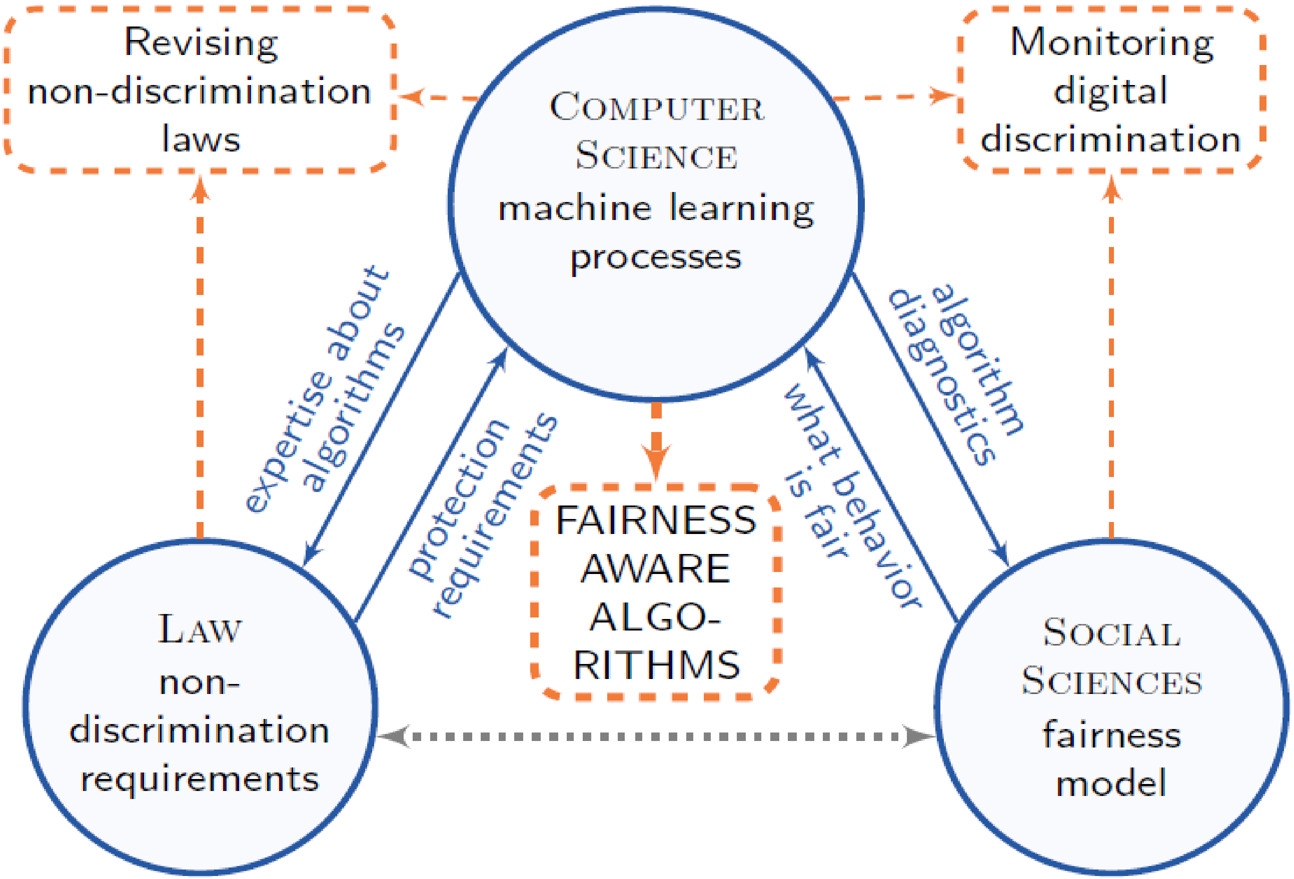
- **Fairness is a subjective, context-dependent and highly politicized concept**; **a global consensus on what is fair is unlikely to emerge, in the context of algorithmic decision making or otherwise**. Think, for example, of the decade-long debate about the interplay between “disparate treatment” and “disparate impact”, for which recent examples include by Ricci v. De Stefano 3 and the ongoing lawsuit regarding the use of race in Harvard University admissions 4.
- **That being said, a productive way to move forward in the data science context is to develop methods that can be instrumented with different alternative fairness notions, and that can support principled and transparent trade-offs between these notions**

4. Fairness has to be dealt with as a multidisciplinary issue

Multidisciplinary fairness

- Infine nell’approccio di [Zlobiaite 2017] la ricerca sulla equità deve essere necessariamente interdisciplinare (vedi figura 1), e deve riguardare insieme le scienze giuridiche, le scienze sociali, e la informatica.
- Le scienze giuridiche aiutano a definire il perimetro dei requisiti anti discriminazione,
- il ruolo delle scienze sociali è quello di definire una giusta allocazione delle risorse negli interventi anti discriminazione; infine l’informatica deve sviluppare le tecniche per la analisi dei modelli di machine learning.
- Le linee continue mostrano interazioni interdisciplinari, mentre le linee tratteggiate mostrano gli obiettivi possibili.

Interazioni tra scienze giuridiche, sociali e informatiche in tema di discriminazione/equità



Revising non-discrimination laws
Computer Science
machine learning processes
Monitoring digital discrimination
expertise about algorithms
protection requirements
what behavior is fair
algorithm diagnostics
FAIRNESS AWARE ALGO-RITHMS
Law
non-discrimination requirements
Social Sciences
fairness model

Reconnect with the moral foundations of fairness

- Our final and most important reason for optimism is that the turn to automated decision-making and machine learning offers an opportunity to **reconnect with the moral foundations of fairness**. Algorithms force us to be explicit about what we want to achieve with decision-making. And it's far more difficult to paper over our poorly specified or true intentions when we have to state these objectives formally. In this way, machine learning has the potential to help us debate the fairness of different policies and decision-making procedures more effectively.
- We should not expect work on fairness in machine learning to deliver easy answers. And we should be suspicious of efforts that treat fairness as something that can be reduced to an algorithmic stamp of approval.
- At its best, this work will make it far more difficult to avoid the hard questions when it comes to debating and defining fairness, not easier.
- It may even force us to confront the meaningfulness and enforceability of existing approaches to discrimination in law and policy, expanding the tools at our disposal to reason about fairness and seek out justice.

Barocas – Daghstul - Big Data’s Disparate Impact

- The paper examines these concerns through the lens of American antidiscrimination law, more particularly, through Title VII’s prohibition of discrimination in employment. In the absence of a demonstrable intent to discriminate, the best doctrinal hope for data mining’s victims would seem to lie in disparate impact doctrine.
- Case law and the Equal Employment Opportunity Commission’s Uniform Guidelines, though, hold that a practice can be justified as a business necessity when its outcomes are predictive of future employment outcomes, and data mining is specifically designed to find such statistical correlations.
- Unless there is a reasonably practical way to demonstrate that these discoveries are spurious, Title VII would appear to bless its use, even though the correlations it discovers will often reflect historic patterns of prejudice, others’ discrimination against members of protected groups, or flaws in the underlying data

Look through the lens of (American) antidiscrimination law

- The best doctrinal hope for data mining’s victims would seem to lie in disparate impact doctrine.
- Case law and the Equal Employment Opportunity Commission’s Uniform Guidelines, though, hold that **a practice can be justified as a business necessity when its outcomes are predictive of future employment outcomes**, **and data mining is specifically designed to find such statistical correlations**.
- Unless there is a reasonably practical way to demonstrate that these discoveries are spurious, Title VII would appear to bless its use, even though the correlations it discovers will often reflect historic patterns of prejudice, others’ discrimination against members of protected groups, or flaws in the underlying data

Conclusions

- Addressing the sources of this unintentional discrimination and remedying the corresponding deficiencies in the law will be difficult technically, difficult legally, and difficult politically.
- **There are a number of practical limits to what can be accomplished computationally. For example, when discrimination occurs because the data being mined is itself a result of past intentional discrimination, there is frequently no obvious method to adjust historical data to rid it of this taint.**
- **Corrective measures that alter the results of the data mining after it is complete would tread on legally and politically disputed terrain.**
- These challenges for reform throw into stark relief the tension between the two major theories underlying antidiscrimination law: anticlassification and antisubordination. Finding a solution to big data’s disparate impact will require more than best efforts to stamp out prejudice and bias; it will require a wholesale reexamination of the meanings of “discrimination” and “fairness”.

Limitations of the approach

- **Addressing the sources of this unintentional discrimination and remedying the corresponding deficiencies in the law will be difficult technically, difficult legally, and difficult politically.**
- There are a number of practical limits to what can be accomplished computationally. For example, **when discrimination occurs because the data being mined is itself a result of past intentional discrimination, there is frequently no obvious method to adjust historical** data to rid it of this taint.
- **Corrective "postptocessing" measures that alter the results of the data mining after it is complete would tread on legally and politically disputed terrain.**
- These challenges for reform throw into stark relief the tension between the two major theories underlying antidiscrimination law: anticlassification and antisubordination. Finding a solution to big data's disparate impact will require more than best efforts to stamp out prejudice and bias; it will require a wholesale reexamination of the meanings of "discrimination" and "fairness".

2019 - Fairness and Abstraction in Sociotechnical Systems

- A key goal of the fair-ML community is to develop machine-learning based systems that, once introduced into a social context, can achieve **social and legal outcomes such as fairness, justice, and due process.**
- **Bedrock concepts in computer science—such as abstraction and modular design—are used to define notions of fairness and discrimination, to produce fairness-aware learning algorithms, and to intervene at different stages of a decision-making pipeline to produce "fair" outcomes**.
- In this paper, however, **we contend that these concepts render technical interventions ineffective, inaccurate, and sometimes dangerously misguided when they enter the societal context that surrounds decision-making systems**.
- **We outline this mismatch with five "traps"** that fair-ML work can fall into even as it attempts to be more context-aware in comparison to traditional data science.

The five traps

- Failure to model the entire system over which a social criterion, such as fairness, will be enforced
- Failure to understand how repurposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context
- Failure to account for the **full meaning of social concepts** such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms
- Failure to understand how the insertion of technology into an existing social system changes the behaviors and embedded values of the pre-existing system
- Failure to **recognize the possibility that the best solution to a problem may not involve technology**

Assessing bias In **health care management programs** namely, assessing disparate treatment and disparate impact

- A single algorithm drives an important health care decision for over 70 million people in the US. When health systems anticipate that a patient will have especially complex and intensive future health care needs, **she is enrolled in a 'care management' program, which provides considerable additional resources**: greater attention from trained providers and help with coordination of her care.
- **To determine which patients will have complex future health care needs, and thus benefit from program enrollment, many systems rely on an algorithmically generated commercial risk score**. In this paper, we exploit a rich dataset to study racial bias in a commercial algorithm that is deployed nationwide today in many of the US's most prominent Accountable Care Organizations (ACOs).
- **We document significant racial bias in this widely used algorithm, using data on primary care patients at a large hospital**. Blacks and whites with the same algorithmic risk scores have very different realized health. For example, the highest-risk black patients (those at the threshold where patients are auto-enrolled in the program), have significantly more chronic illnesses than white enrollees with the same risk score.

5. Recommendations for future research and implementations

Emphasize preprocessing requirements

- If there are multiple plausible ways in which a dataset can be processed to generate training data for an algorithm, provide performance metrics for more than one of the possible choices.
- If algorithms are being compared to each other, ensure they are compared based on the same preprocessing.

Avoid proliferation of measures

- New fairness measures should only be introduced if they behave fundamentally differently from existing metrics.
- Our study indicates that a combination of group conditioned accuracy and either DI or CV is a good minimal set

Account for training instability

Showing the performance of an algorithm in a single training-test split appears to be insufficient.

We recommend reporting algorithm success and stability based on a moderate number of randomized training-test splits

Further recommendations

- One limitation of our benchmark is the number of methods it currently provides implementation for. We hope other researchers will contribute their implementations to the repository. It would be particularly interesting to see how our conclusions above evolve as the number and variety of methods increases.
- Additionally, while we frame some of the differences in algorithm performance as fairness versus accuracy tradeoffs, this can be misleading since it makes many assumptions about the data and social context, including, e.g., that the labels represent desired outcomes. We leave the examination of how the algorithmic choices interact with the social context for other work.

Appendix on datasets available for benchemarking

Ricci

- The Ricci dataset comes from the case of Ricci v. DeStefano a case before the U.S. Supreme Court in which the question at issue was an exam given to determine if firefighters would receive a promotion. The dataset has 118 entries and five attributes, including the sensitive attribute Race. The original promotion decision was made by a threshold of achieving at least a score of 70 on the combined exam outcome. The goal in a fair learning context is to predict this original promotion decision while achieving fairness with respect to the sensitive attribute, Race.

Adult Income

- The Adult Income dataset contains information about individuals from the 1994 U.S. census. It is pre-split into a training and test set; we use only the training data and re-split it. There are 32,561 instances and 14 attributes, including sensitive attributes race and sex. 2,399 instances with missing data are removed during the preprocessing step. The prediction task is predicting whether an individual makes more or less than $50,00 per year.

German

The German Credit dataset contains 1,000 instances and 20 attributes describing individuals along with a classification of each individual as a good or bad credit risk. Sensitive attribute sex is not directly included in the data, but can be derived from the given information. Sensitive attribute age is included, and is discretized into values adult (age at least 25 years old) and youth based on an analysis showing this discretization provided for the most discriminatory possibilities.

ProPublica recidivism

The ProPublica data includes data collected about the use of the COMPAS risk assessment tool in Broward County, Florida [2]. It includes information such as the number of juvenile felonies and the charge degree of the current arrest for 6,167 individuals, along with sensitive attributes race and sex.

Data is preprocessed according to the filters given in the original analysis [2]. Each individual has a binary “recidivism" outcome, that is the prediction task, indicating whether they were rearrested within two years after the charge given in the data.

ProPublica violent recidivism

- The violent recidivism version of the ProPublica data describes the same scenario as the recidivism data described above, but where the predicted outcome is a rearrest for a violent crime within two years. 4,010 individuals are included after preprocessing is applied, including 652 instances of rearrest, and the sensitive attributes are race and sex. Note that while the individuals in this data set are a subset of the overall recividism set from above, their labels might be different, i.e., the same individual might have different recidivism labels in the two data sets.

References

- A. Barry Jester - https://fivethirtyeight.com/features/prison-reform-risk-assessment/Should Prison Sentences Be Based On Crimes That Haven't Been Committed Yet?
- T. Blomberg - VALIDATION OF THE COMPAS RISK ASSESSMENT CLASSIFICATION INSTRUMENT, 2016
- A. Chouldechova - Fair prediction with disparate impact - A study of bias in recidivism prediction instruments, *Big data* 5.2 (2017): 153-163.
- T. Cohen, The Federal Post-Conviction Risk Assessment Instrument, 2017
- W. Dieterich - COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity, Northpointe, 2016
- J. Dressel - The accuracy, fairness, and limits of predicting recidivism, 2018
- S. Goeal et al. PRECINCT OR PREJUDICE? UNDERSTANDING RACIAL DISPARITIES IN NEW YORK CITY'S STOP-AND-FRISK POLICY, The Annals of Applied Statistics
- Guide to the Pretrial Decision Framework – Laura and John Arnold Foundation, 2018.
- D. Kehl et al. Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing, The Harvard Library, 2016.
- J. Kleinberg - Inherent Trade-Offs in the Fair Determination of Risk Scores, 2016
- John Monahan and Jennifer L. Skeem, Risk Assessment in Criminal Sentencing, Annu. Rev. Clin. Psychol. 2016.
- Jeff Larson, How We Analyzed the COMPAS Recidivism Algorithm https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
- R. Pedroncelli - A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear - https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/
- There's software used across the country to predict future criminals. And it's biased against blacks, Propublica, 2016.
- Wisconsin offender statement - https://doc.wi.gov/Pages/AboutDOC/COMPAS.aspx
- M. Zafar - Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment - 2017 International World Wide Web Conference Committee (IW3C2)15

References

- Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).
- D. Pedreschi Salvatore Ruggieri Franco Turini - Discrimination-aware Data Mining, KDD 2008.
- Abiteboul, Serge, and Julia Stoyanovich. "Transparency, fairness, data protection, neutrality: Data management challenges in the face of new regulation." *Journal of Data and Information Quality (JDIQ)* 11.3 (2019): 1-9.
- B. Hutchinson and M. Mitchell - 50 Years of Test (Un)fairness: Lessons for Machine Learning, FAT Conference 2019.
-

References for fairness classifications

- S. Verma et al. Fairness Definitions Explained - 2018 ACM/IEEE International Workshop on Software Fairness
- Ziyuan Zhong - A Tutorial on Fairness in Machine Learning https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb, 2018.