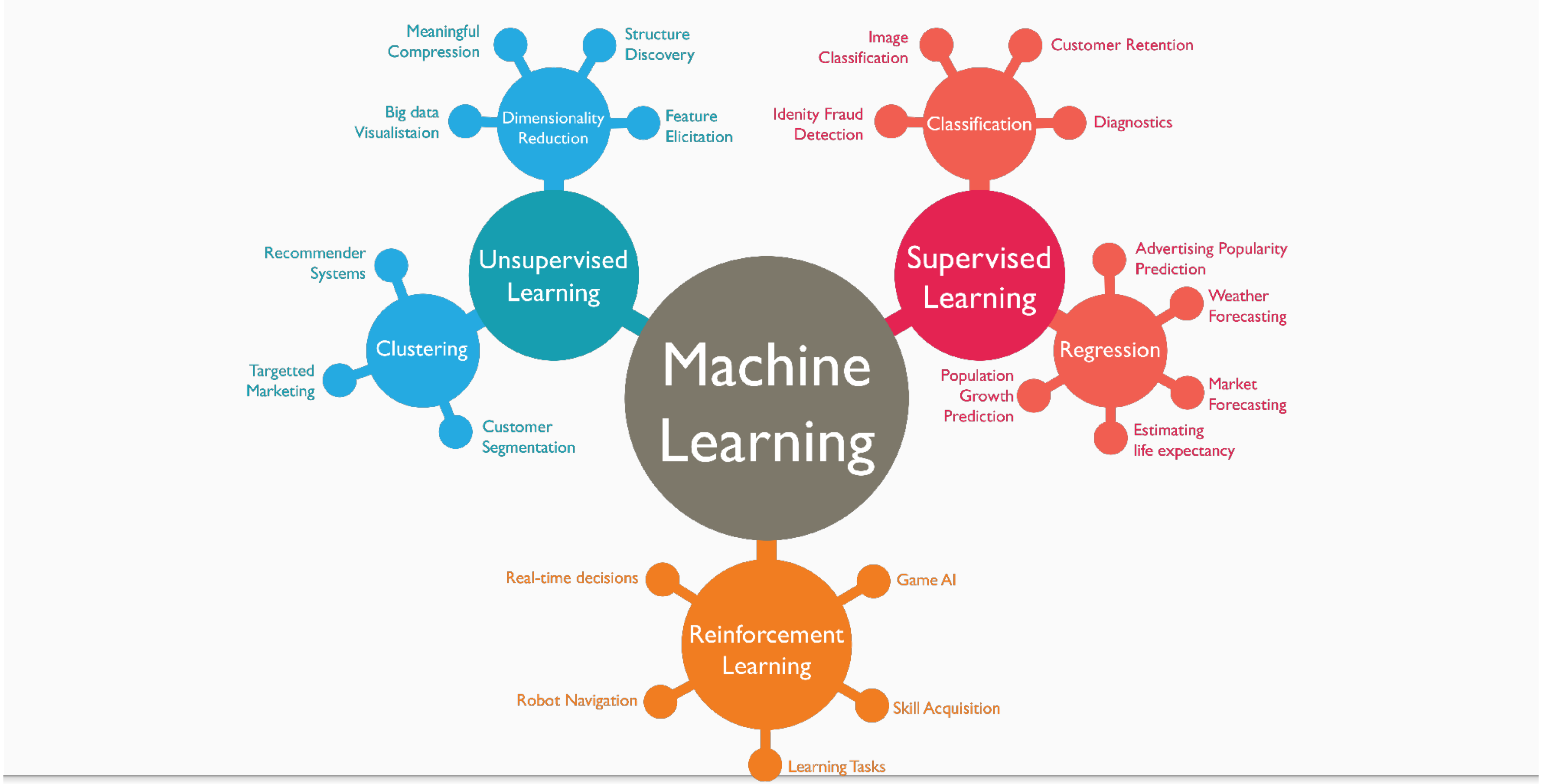
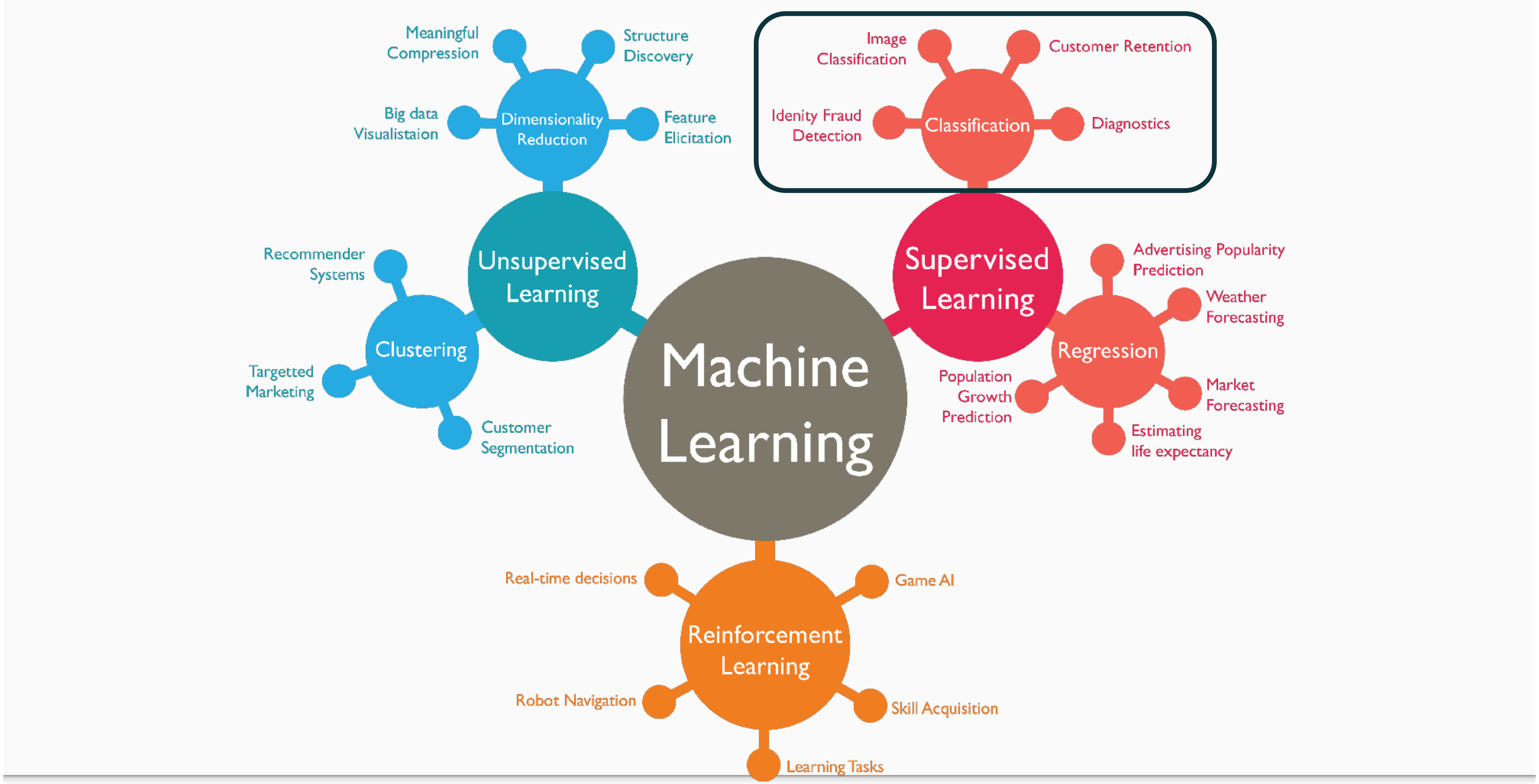
An introduction to

Fairness in Machine Learning

Machine Learning (CS60050)
Spring 2020



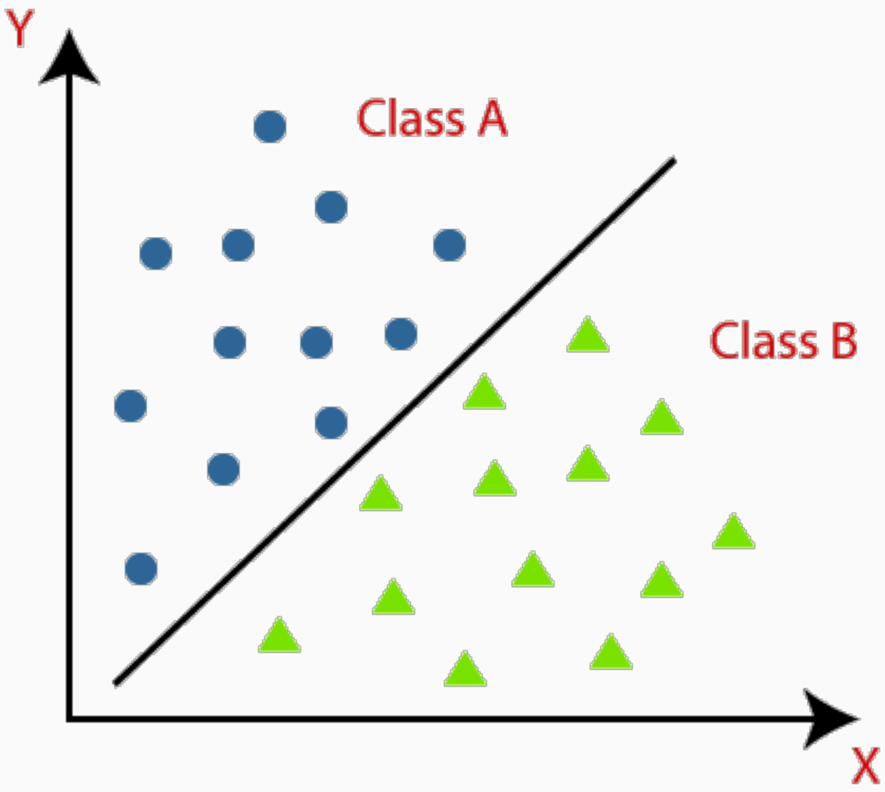
The entire spectrum of Machine Learning.



For this lecture, let us consider only the Classification problem

Classification

- ❑ Main goal is to learn the decision boundary which separates Class A from Class B.



Traditional approach is to maximize

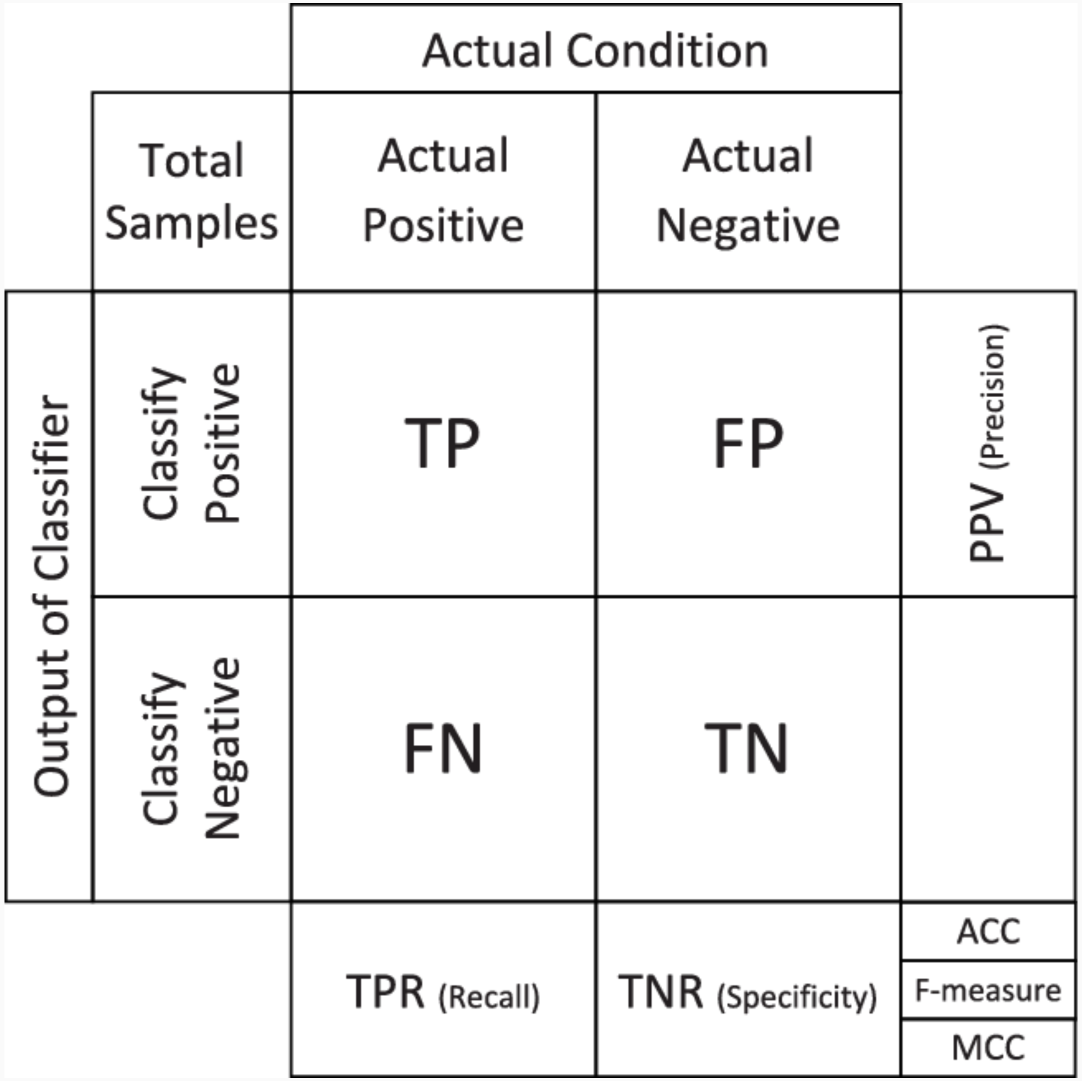
Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

Specificity = TN / (FP + TN)

Accuracy =

(TP + TN)/(TP + TN + FP + FN)



		Actual Condition		
	Total Samples	Actual Positive	Actual Negative	
Output of Classifier	Classify Positive	TP	FP	PPV (Precision)
	Classify Negative	FN	TN	
		TPR (Recall)	TNR (Specificity)	ACC F-measure MCC

Benign classification problem

- ❑ Classify between cats and dogs images
- ❑ Classify spam and benign Facebook posts

Classification problems in real life

❑ Granting loans to people

❑ Recidivism prediction for criminals (whether to grant them bail)

❑ Predict suitability of candidates for jobs

The list is endless...

These are cases where the ML system can have severe consequences, e.g., on the livelihood of people and their families.

The list is endless...

Why is it not enough to have high performance (e.g., accuracy)?

A Toy Example...

- Let us consider a classifier that is used by an organization to hire 100 candidates who applied (69 male and 31 females)
- If the classifier predicts 1 (Y’ = 1) then organization will hire the candidate
- True class Y=1 if the candidate really deserves to be hired, 0 otherwise
- Classifier achieves an overall accuracy of 90%

Overall	Y = 1	Y = 0
Y’ = 1	46	4
Y’ = 0	6	44

A Toy Example...

❑ Let us consider a classifier that is used by an organization to hire 100 candidates who applied (69 male and 31 females)

❑ For males, the accuracy rises to 94.2%

Males	Y = 1	Y = 0
Y’ = 1	35	4
Y’ = 0	0	30

A Toy Example...

- Let us consider a classifier that is used by an organization to hire 100 candidates who applied (69 male and 31 females)
- For females, the accuracy drops to 80.64%

Females	Y = 1	Y = 0
Y’ = 1	11	0
Y’ = 0	6	14

A Toy Example

So, the classifier has different accuracies for different groups. Is this problematic?

Y' = 1	11	0
Y' = 0	6	14

What is the problem?

Males	Y = 1	Y = 0
Y’ = 1	35	4
Y’ = 0	0	30

Females	Y = 1	Y = 0
Y’ = 1	11	0
Y’ = 0	6	14

What is the problem?

Males	Y = 1	Y = 0
Y’ = 1	35	4
Y’ = 0	0	30

4 Undeserving male candidates got hired!!!

Females	Y = 1	Y = 0
Y’ = 1	11	0
Y’ = 0	6	14

6 Deserving female candidates got rejected!!!

What is the problem?

❑ Misclassification for a system of such large consequence costs a lot

❑ From the organization’s point of view, they hired 4 undeserving candidates

❑ From the candidates’ perspective, 6 deserving candidates got rejected

What is the problem?

- ❑ Misclassification for a system of such large consequence costs a lot
- ❑ From the organization’s point of view, they hired 4 undeserving candidates
- ❑ From the candidates’ perspective, 6 deserving candidates got rejected.

Also, is the system legal?

What is the problem?

- ❑ Misclassification for a system of such large consequence costs a lot.
- ❑ From the organization’s point of view, they hired 4 undeserving candidates
- ❑ From the candidates’ perspective, 6 deserving candidates got rejected.

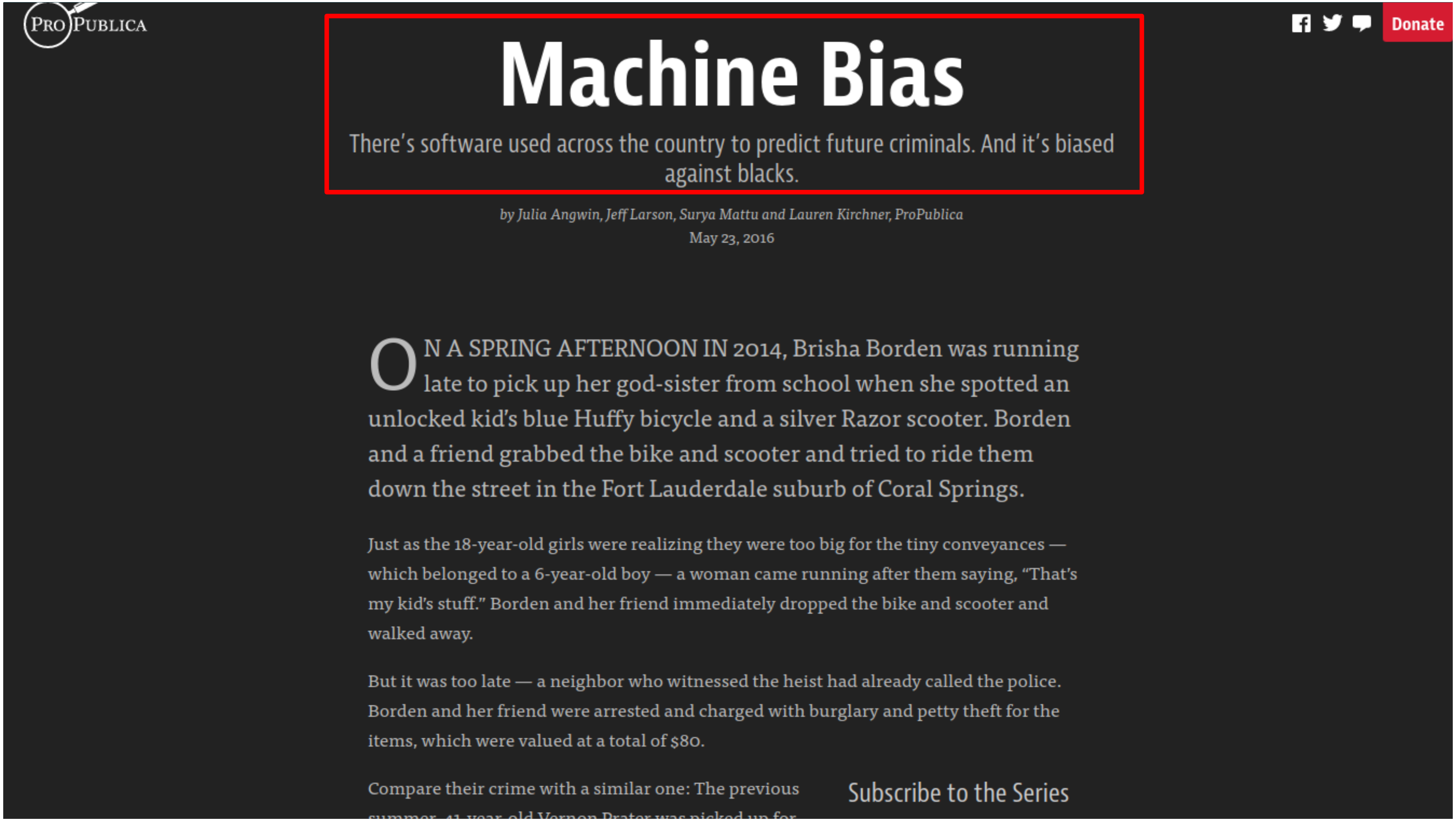
Also, is the system legal? NO (According to U.S. Equal Employment Opportunity Commission: the “80%-rule”)

Selection rate for males = 56.52% (39 out of 69)
Selection rate for females = 35.48% (11 out of 31) According to the rule, this selection rate should be at least 80% of the selection rate for males

A real-life example

Canonical example: COMPAS

- COMPAS: Correctional Offender Management Profiling for Alternative Sanctions
- Measures the risk of a person to commit another crime (recidivism)
- Judges in USA use this system while deciding court cases, e.g., whether to release an offender on bail, or to keep him/her in prison.



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

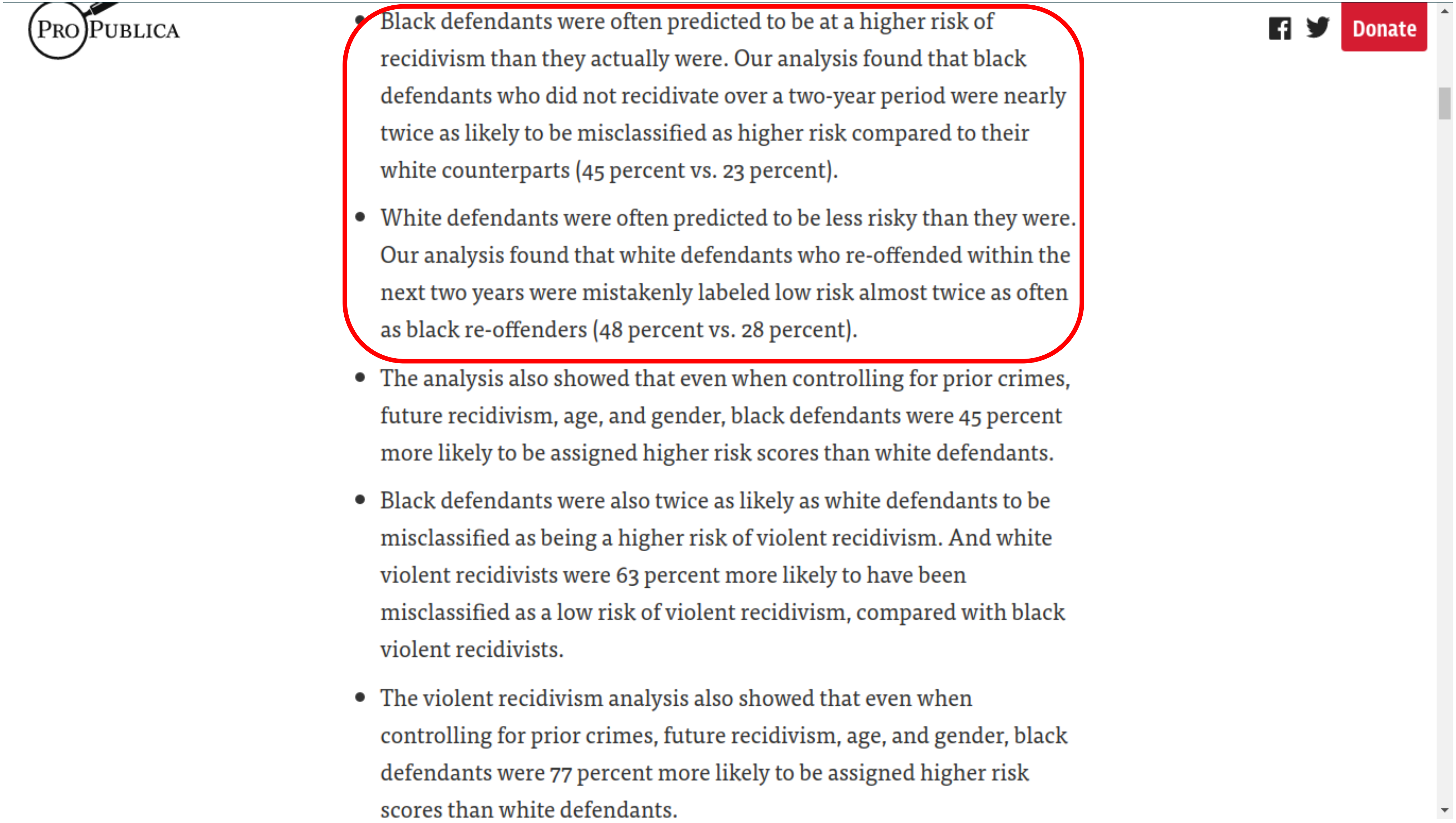
ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

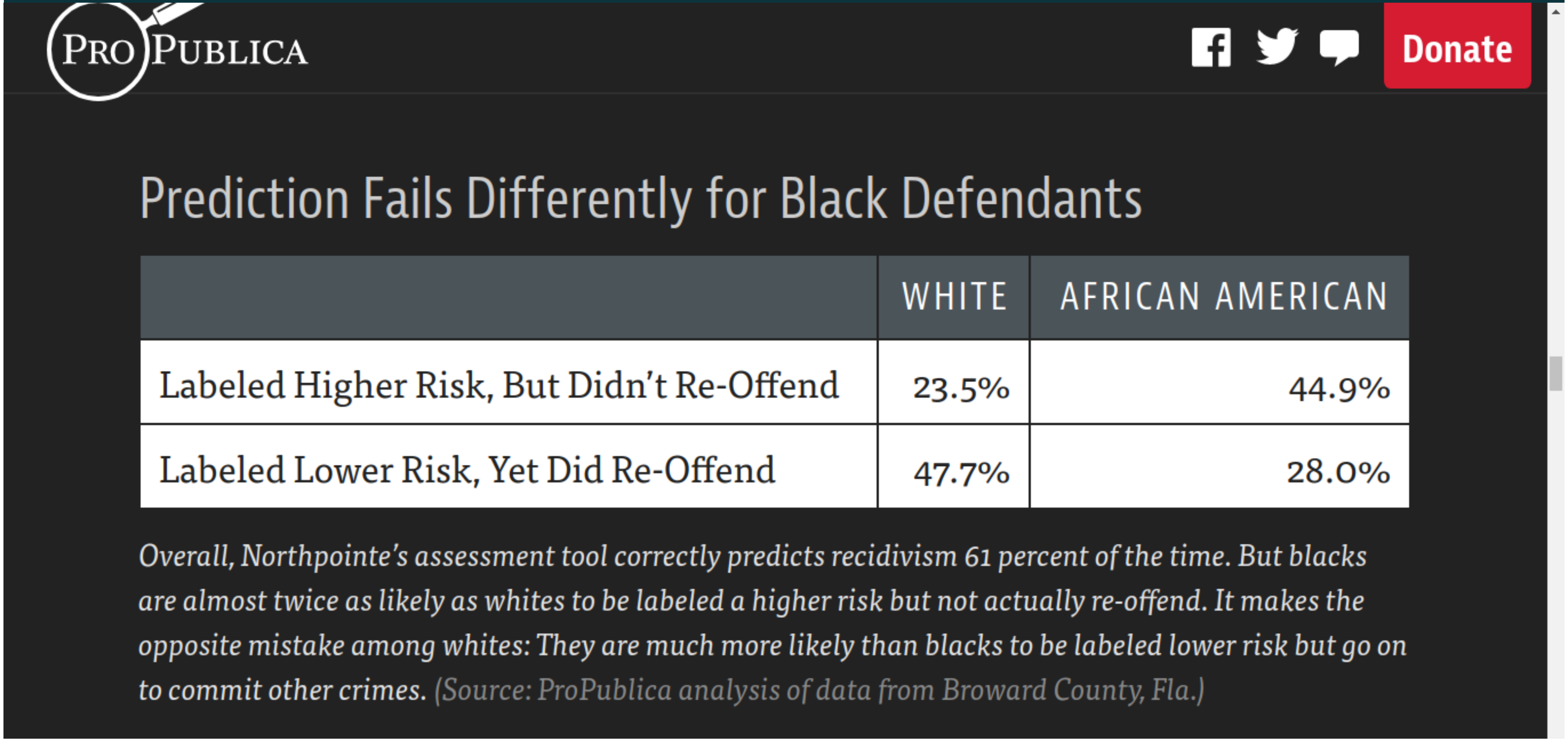
Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for

Subscribe to the Series



- Black defendants were often predicted to be at a higher risk of recidivism than they actually were. Our analysis found that black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent).
- White defendants were often predicted to be less risky than they were. Our analysis found that white defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders (48 percent vs. 28 percent).
- The analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants.
- Black defendants were also twice as likely as white defendants to be misclassified as being a higher risk of violent recidivism. And white violent recidivists were 63 percent more likely to have been misclassified as a low risk of violent recidivism, compared with black violent recidivists.
- The violent recidivism analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 77 percent more likely to be assigned higher risk scores than white defendants.

Machine Bias (ProPublica)



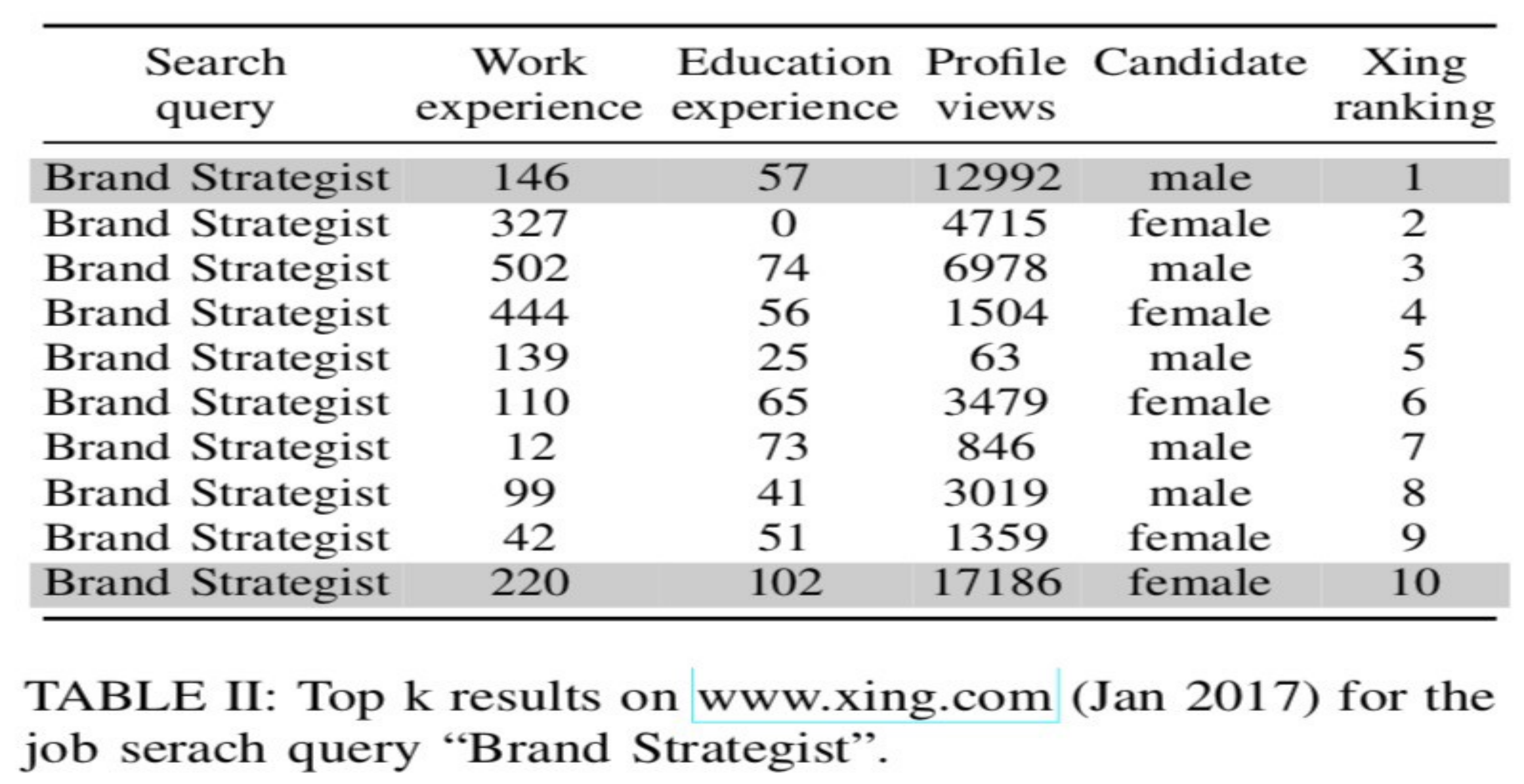
Prediction Fails Differently for Black Defendants

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

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Some more real-life examples

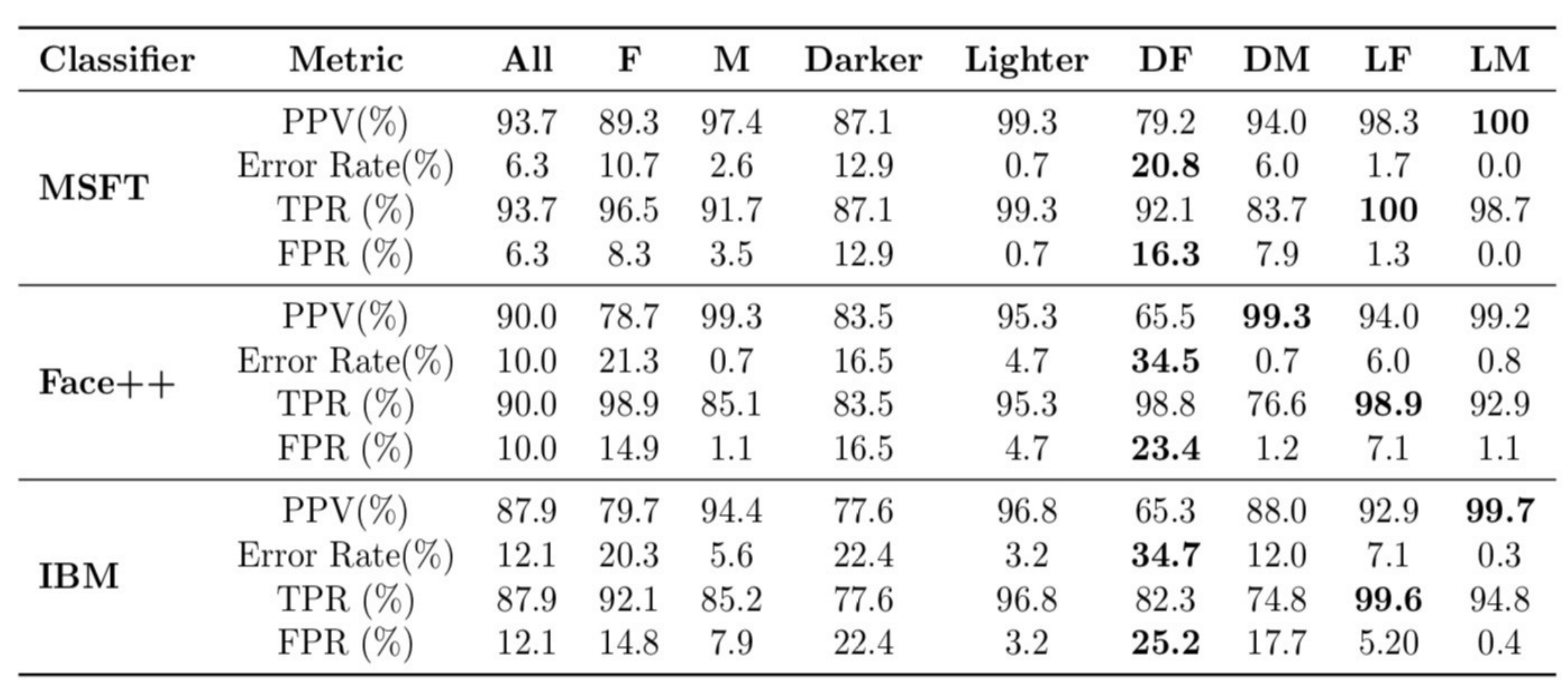
XING, a job platform similar to Linked-in, was found to rank less qualified male candidates higher than more qualified female candidates (see Fig3, Lahoti et al. 2018)



Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

TABLE II: Top k results on www.xing.com (Jan 2017) for the job serach query "Brand Strategist".

Publicly available commercial face recognition online services provided by Microsoft, Face++, and IBM respectively are found to suffer from achieving much lower accuracy on females with darker skin color (see Fig4, Buolamwini and Gebru, 2018)



Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

DF, DM, LF, LM stand for: darker skin female, darker skin male, lighter skin female and lighter skim male. PPV, TPR, FPR stand for predictive positive value, true positive rate and false positive rate.

For many real-life applications, ML models not only need good performance (e.g., high accuracy) but also need to be fair

We need some way to measure/define fairness (just like we measure performance)

How to measure/define fairness?

Some terms & definitions

Bias, Discrimination, Fairness

Bias**:** Inclination / prejudice for or against one person or group, especially in a way considered to be unfair.

Discrimination**:** The unjust or prejudicial treatment of different categories of people, especially on the grounds of race, age, or sex.

Fairness: Absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics

Multiple types of Fairness

- ❑ Group fairness – different groups should not be treated too differently
  - ❑ E.g., selection rate of females should not be too much lower than that for males
- ❑ Individual fairness – different individuals who are (almost) equal in various aspects should be treated (almost) equally
  - ❑ E.g., if candidates A and B have similar qualifications, it should not be that A is selected but B is rejected

… and many more

How to define what is fair?

- ❑ Many definitions of fairness
  - ❑ Often influenced by laws of a country (e.g., 80% rule), human perception of what is ethical / morally justifiable, etc. [details not being discussed – see additional readings and references]
- ❑ Definition of fairness varies according to type of fairness and the domain
- ❑ We will focus on a few group fairness definitions proposed for classification

Notations used in next few slides

❑ A classifier C is being used to hire candidates for a job

❑ Y -- Actual class (ground truth deservingness); can take values from {0, 1}

❑ Y’ -- Predicted class by C; can take values from {0, 1}

❑ A -- Protected attribute; can take values from {0, 1}

Y’ = 0 means rejection and Y’ = 1 means selection (according to C)

Let us assume A ~ Gender; A = 0 for Male and A = 1 for Female.

Definition 1: Independence

- ❑ One of the most well-known criteria for fairness; also called Statistical Parity or Demographic Parity
- ❑ Strict version: P (Y’ = 1 | A = 0) = P (Y’ = 1 | A = 1)
  - ❑ Probability of selection for Male should be equal to probability of selection for Female
- ❑ Several less strict versions, e.g., “**80% rule**” prescribes that selection rate for any other group must be at least 80% of the rate for the group with the highest rate
  - ❑ Remember the toy example we studied

Shortcomings of Independence

❑ Ignores possible correlation between Y and A

❑ In particular, may rule out the perfect classifier C that gives Y’ = Y

❑ Permits laziness: accept qualified people in one group and random people from the other (e.g., so that the selection rate is same for all groups)

Definition 2: Separation

- Equal opportunity: $P\ (Y' = 1 |\ A = 0, Y = 1\ ) = P\ (Y' = 1 |\ A = 1, Y = 1\ )$
  - True Positive Rate (TPR) equalized

- Equalized odds: $P\ (Y' = 1 |\ A = 0, Y = y\ ) = P\ (Y' = 1 |\ A = 1, Y = y\ )$, $y = \{0, 1\}$
  - Both TPR and False Positive Rate (FPR) equalized

- This notion is independence conditioned on Y (actual class)

Desirable properties of Separation

❑ Allows the perfect classifier C that gives Y’ = Y

❑ Penalizes laziness: Incentive to reduce errors uniformly in all groups

There are many other fairness definitions …

- Also different definitions can be conflicting with each other
- Details not being discussed … see additional readings & references if interested

Why are some ML models unfair?

Discrimination & Machine learning

❑ ML is supposed to recognize and understand the differences among various instances

❑ But certain situations are undesirable

  ❑ Basis of differentiation is unjustified

  ❑ Basis of differentiation is practically or morally irrelevant

What basis of differentiation is not acceptable? Regulated domains and sensitive attributes:

❑ Credit

❑ Education

❑ Employment

❑ Housing

❑ Public accommodation

Race; Color; Gender; Religion; National Origin; Citizenship; Age; Pregnancy; Familial Status; Disability status; etc.

What basis of differentiation is not acceptable? Regulated domains and sensitive attributes:

❑ Credit
❑ Education
❑ Employment
❑ Housing
❑ Public accommodation

Race; Color; Gender; Religion; National Origin; Citizenship; Age; Pregnancy; Familial Status; Disability status; etc.

Note: Whether a basis of differentiation is acceptable depends on the specific domain. E.g.,

- Religion is an acceptable basis of differentiation while recruiting pastor for a church
- Disability status is an acceptable basis of differentiation while recruiting a footballer

A Naive approach: Unawareness

❏ Do not include the sensitive attributes as features in the training data

❏ Fundamental limitation - there can be many other features that are highly correlated with the sensitive attributes

  ❏ E.g., height is often correlated with gender

  ❏ E.g., zip code is often correlated to race in USA

❏ Thus, only removing the sensitive attribute is by no means enough

Common causes of bias/unfairness in ML systems

- ❑ Tainted examples -- Any ML system can learn the bias existing in the old data (originally caused by human bias)
- ❑ Limited features -- Features may be less informative or less reliably collected for minority group(s)
- ❑ Sample size disparity -- Training data coming from minority group is much lesser than from a majority group
- ❑ Injudicious use of proxy features, e.g., height for gender, zip code for race in USA

Methods for fair ML (just basics)

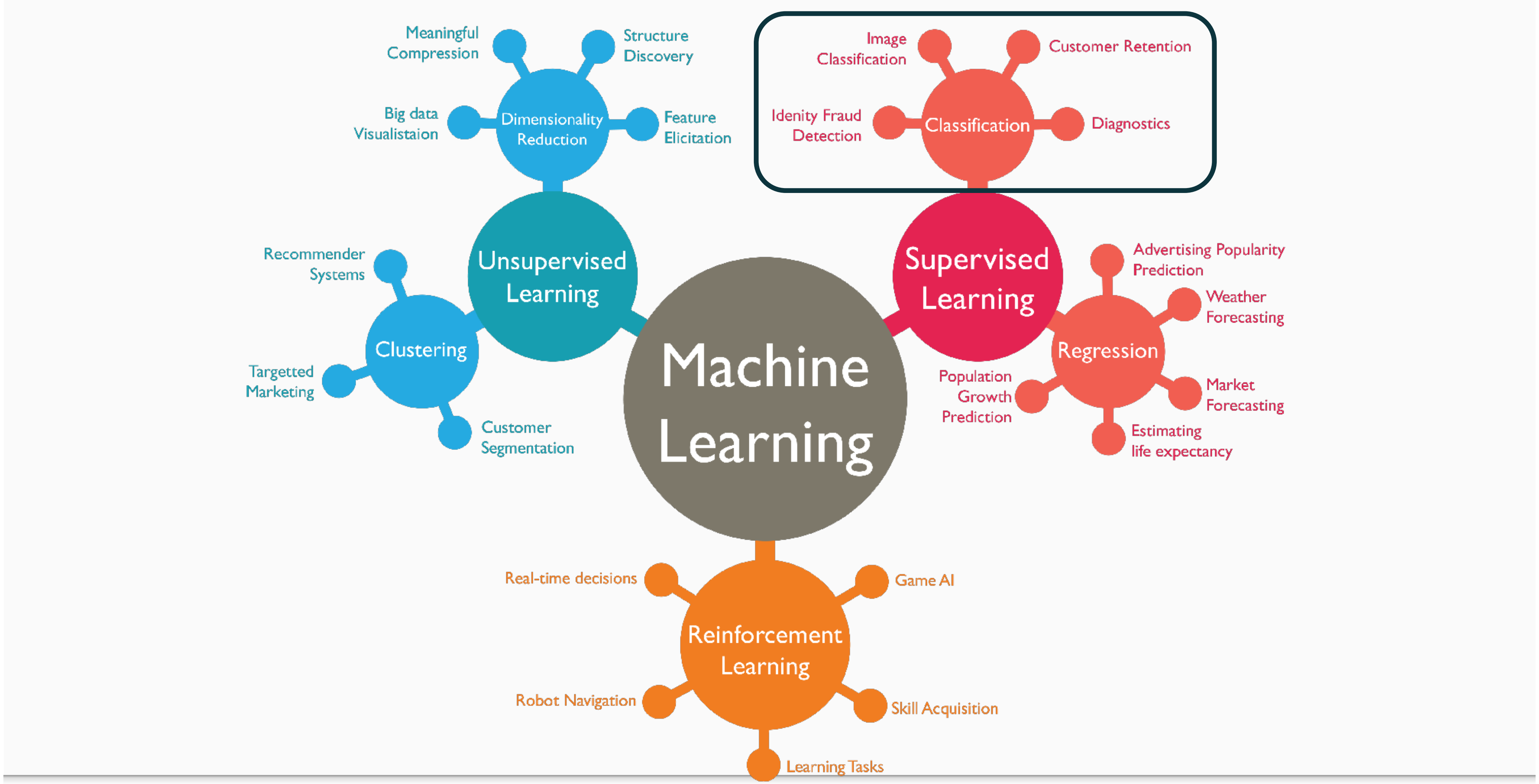
- ❑ Pre-processing methods: transform the data to remove the bias
- ❑ In-processing methods: modify the ML algorithms to make them fair (with possibly a small loss in performance)
- ❑ Post-processing methods: post-process the output to make it fair; applicable when you are dealing with a black box model

Approaches used in In-processing methods

- ❑ Define a measure for decision boundary unfairness of your classifier (quantification of unfairness)
- ❑ Define the loss function (minimizing which will improve classification accuracy)

- ❑ Then two complimentary formulations can be derived for a fair classifier
    - ❑ Minimize the loss function subject to fairness constraints
    - ❑ Maximize fairness subject to accuracy constraints

See Zafar et.al. for technical details.



In this lecture, we considered fairness only in context of the Classification problem. Fairness concerns exist in all types of learning

References for further reading

Title	Contributor	Venue
A Survey on Bias and Fairness in Machine Learning	Ninareh Mehrabi et al.	Arxiv, 2019
Fairness in Machine Learning- Tutorial	Solon Barocas, Moritz Hardt	NIPS, 2017
21 fairness definitions and their politics- Tutorial	Arvind Narayanan	FAT* 2018
Machine Bias	Julia Angwin et al.	ProPublica, 2016
Bias on the Web	Ricardo Baeza-Yates	CACM, 2018
Fairness of Exposure in Ranking	Ashudeep Singh et al.	SIGKDD, 2018
Equality of opportunity in supervised learning	Moritz Hardt et al.	NIPS, 2016
Fairness through awareness	Cynthia Dwork et al.	ITCS, 2012
A convex framework for fair regression	Richard Berk et al.	FATML, 2017
The Price of Fair PCA: One Extra Dimension	Samira Samadi et al.	NIPS, 2018
Debiasing Community Detection: The Importance of Lowly-Connected Nodes	Ninareh Mehrabi et al.	ASONAM, 2019
iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making	Preethi Lahoti et al.	ICDE, 2019
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings	Tolga Bolukbasi et al.	NIPS, 2016