The COMPAS case

Mitigating Bias
in Machine Learning
using
Generative Adversarial
NN

Agenda

1

Replicate results of a biased Risk assessment algorithm by training a Feed Forward NN

2

Measure the bias

3

Mitigate the bias using Generative Adversarial NN

What is bias and why is a problem

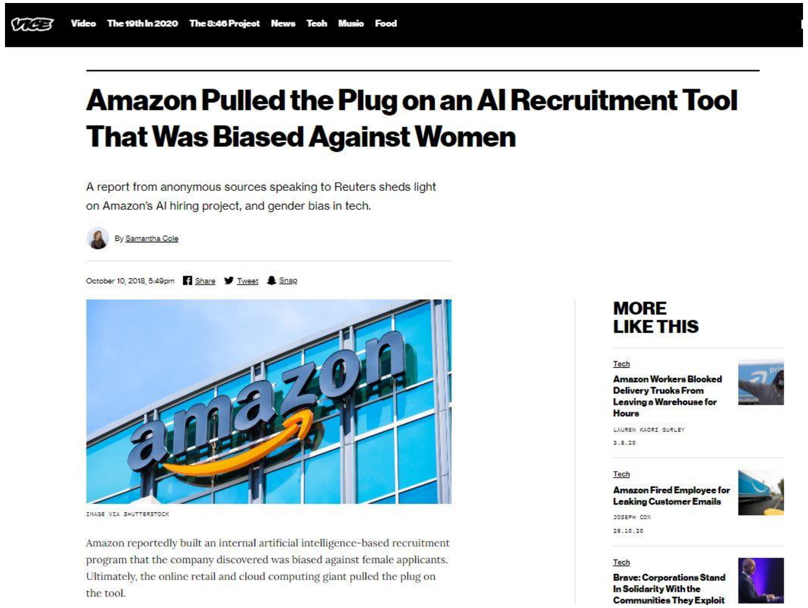

CNN BUSINESS Markets Tech Media Success Perspectives Videos Edition

Facial recognition systems show rampant racial bias, government study finds

By Brian Fung, CNN Business
Updated 2337 GMT (0737 HKT) December 19, 2019

Washington, DC (CNN Business) — Federal researchers have found widespread evidence of racial bias in nearly 200 facial recognition algorithms in an extensive government study, highlighting the technology's shortcomings and potential for misuse.

Racial minorities were far more likely than whites to be misidentified in the US government's testing, the study found, raising fresh concerns about the software's impartiality even as more government agencies at the city, state and federal level clamor to use it.


VICE Video The 19th In 2020 The 8:46 Project News Tech Music Food

Amazon Pulled the Plug on an AI Recruitment Tool That Was Biased Against Women

A report from anonymous sources speaking to Reuters sheds light on Amazon's AI hiring project, and gender bias in tech.

By Samantha Cole

October 10, 2018, 5:49pm Share Tweet Snap

IMAGE VIA SHUTTERSTOCK

Amazon reportedly built an internal artificial intelligence-based recruitment program that the company discovered was biased against female applicants. Ultimately, the online retail and cloud computing giant pulled the plug on the tool.

MORE LIKE THIS

Tech
Amazon Workers Blocked Delivery Trucks From Leaving a Warehouse for Hours
LAUREN KAORI GURLEY
3.8.20

Tech
Amazon Fired Employee for Leaking Customer Emails
JOSEPH COX
26.10.20

Tech
Brave: Corporations Stand In Solidarity With the Communities They Exploit

Possible causes:

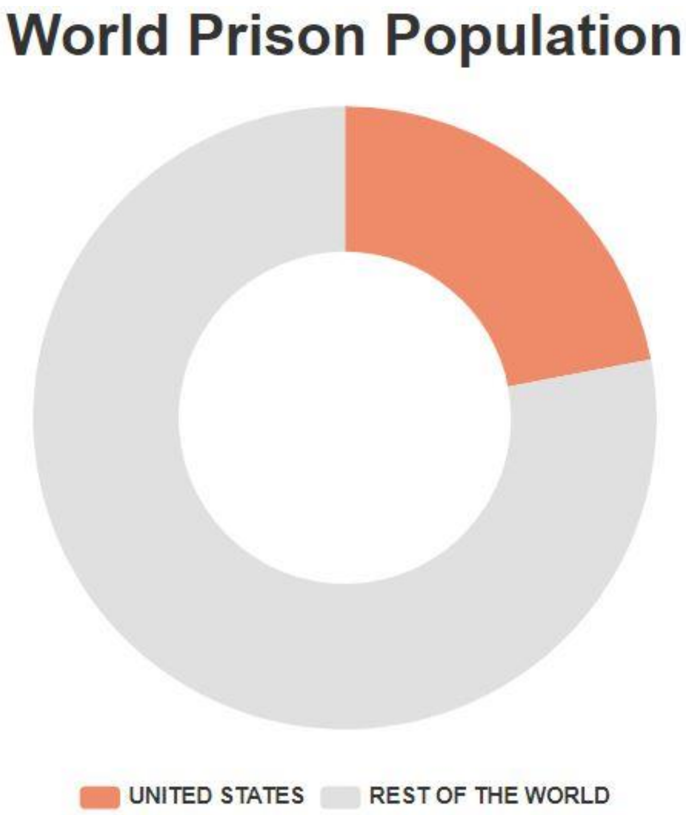
- Not balanced training set
- Training data collected differently than in real life
- Deleting relevant data
- Labelling similar data inconsistently
- Subjective thoughts on data by people involved in the analysis

1. Understanding the data: US Mass Incarceration and racial disparities
2. COMPAS dataset
3. Classification task
4. Generative Adversarial Network (GAN)
5. Conclusions

1. Understanding the data: US Mass Incarceration and racial disparities
2. COMPAS dataset
3. Classification task
4. Generative Adversarial Network (GAN)
5. Conclusions

Some data

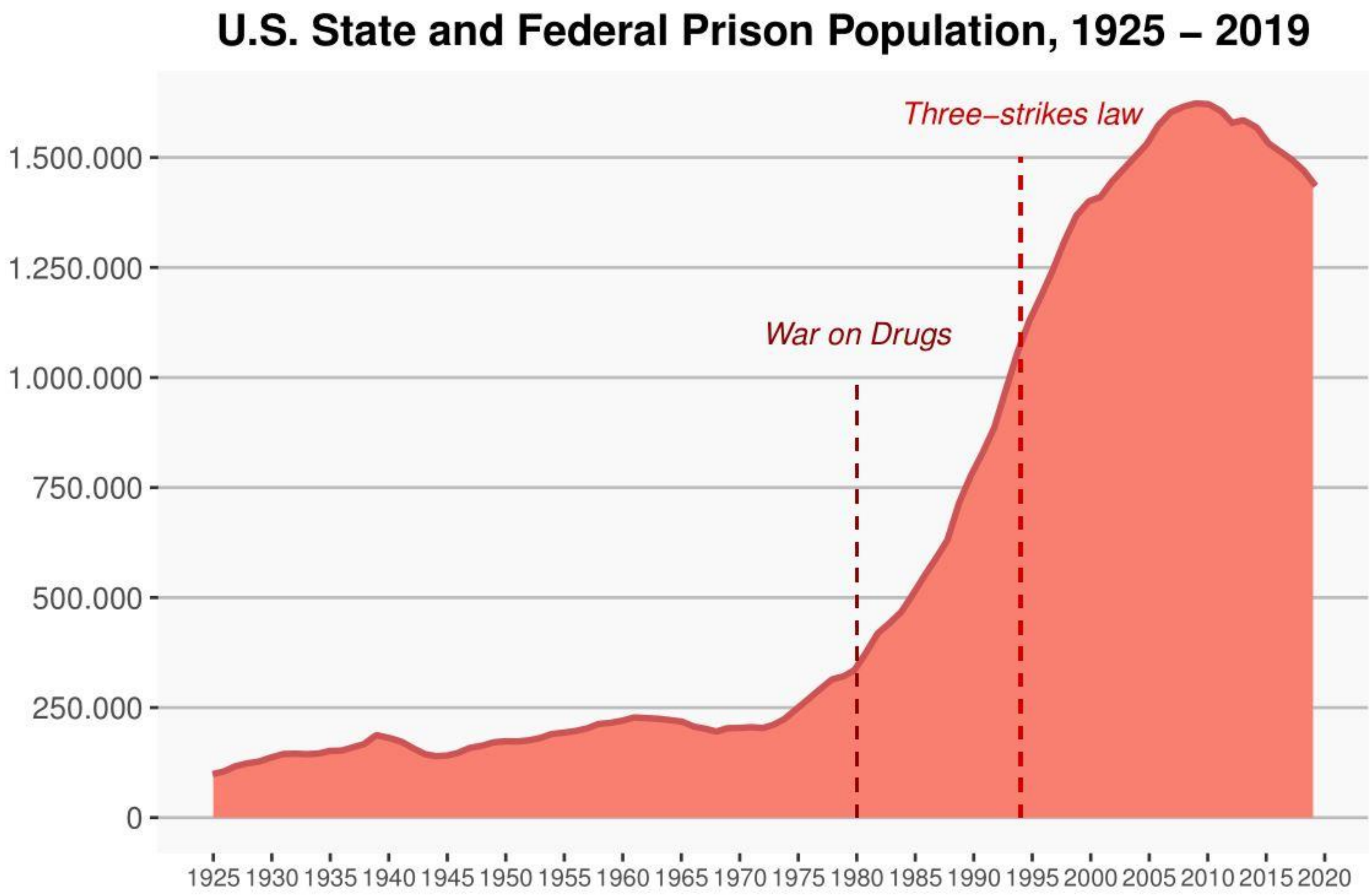
US citizens make up for 4.23% of world population



1 out of 4 people incarcerated in the world is in a US prison

Source: *International Centre for Prison Studies.*

Historical trend

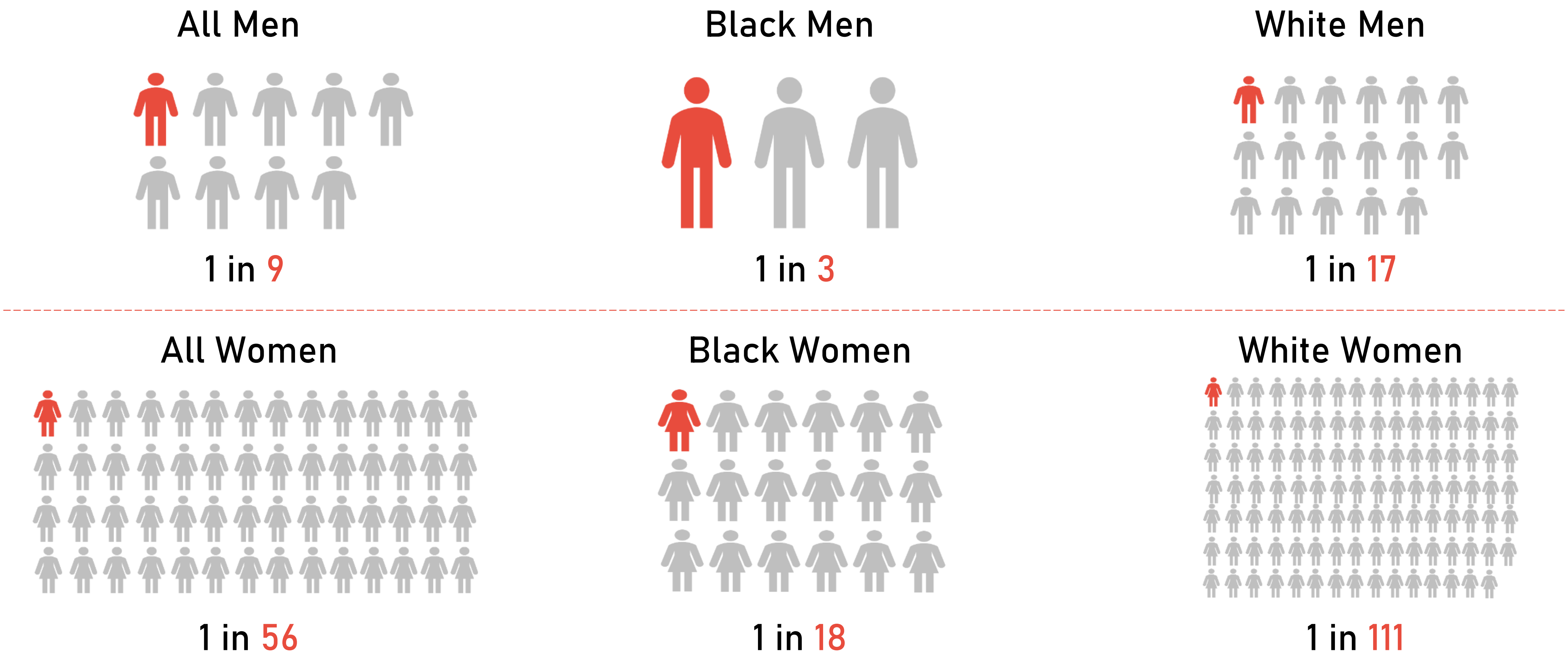


US prison population had a 500% increase over the last forty years

Source: Plot realized based on data from *Bureau of Justice Statistics Prisoners Series.*

Racial disparities

Lifetime Likelihood of Imprisonment for U.S. Residents Born in 2001



Source: Bonczar, T. (2003). Prevalence of Imprisonment in the U.S. Population, 1974-2001. Washington, DC: Bureau of Justice Statistics.

African Americans make up for

12%

of US population

and

33%

of US prison population

Source: *U.S. Census Bureau, Bureau of Justice Statistics (2018)*

1 Understanding the data: US Mass Incarceration and racial disparities

2 **COMPAS dataset**

3 Classification task

4 Generative Adversarial Network (GAN)

5 Conclusions

COMPAS

Correctional Offender Management Profiling for Alternative Sanctions

- assesses the likelihood (risk) of a defendant of becoming a recidivist in the next two years, by computing a score between 1 and 10.
- classifies risk of recidivism depending on the score computed, that is: *low* (1-4) – *medium* (5-7) – *high* (8-10)
- scores are computed starting from defendant's answers to a 137 questions questionnaire

- developed and owned by a private company; the algorithms used are trade secrets
- scores are used by US judges to decide whether to release or detain a defendant before his or her trial and to inform other decisions determining parole and sentencing

Concerns about DATA and BIAS

- Historical and social disparities against African Americans, resulting in different arrests rates and sentencing ⟶ COMPAS scores depend on personal criminal history

- Scores based on a non-digitalized questionnaire:

Was your father/mother/wife/husband/brother ever arrested, that you know of ?
Do you have a regular living situation?
Is there much crime in your neighborhood?
How hard is it for you to find a job above minimum wage compared to others?

⟶ could be proxies for socio-economic status associated to some minorities.

- Data entry errors, data integration errors and missing data

ProPublica Analysis

- Comparing COMPAS scores of 7.000 arrested people in 2013 and 2014 with their actual criminal activity within two years to assess the *actual* accuracy of the algorithm in predicting the risk of recidivism

- Following the concerns of some US courts members, try to verify whether the algorithm is biased against African American defendants

	Caucasian	African American
Overall Accuracy	59%	63%
False Positive Rate	23%	45%
False Negative Rate	48%	28%

1 = At recidivism risk | True recidivist
0 = No recidivism risk | Not a recidivist

https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

The dataset

- ProPublica made the data available online. After selecting the variables needed in my analysis, the dataset head is given by:

	sex	age	race	juv_fel_count	juv_misd_count	juv_other_count	priors_count	charge_degree	target	two_year_recid
0	1	34	0	0	0	0	0	1	0	1
1	1	24	0	0	0	1	4	1	0	1
2	1	23	0	0	1	0	1	1	1	0
3	1	41	1	0	0	0	14	1	1	1
4	0	39	1	0	0	0	0	0	0	0

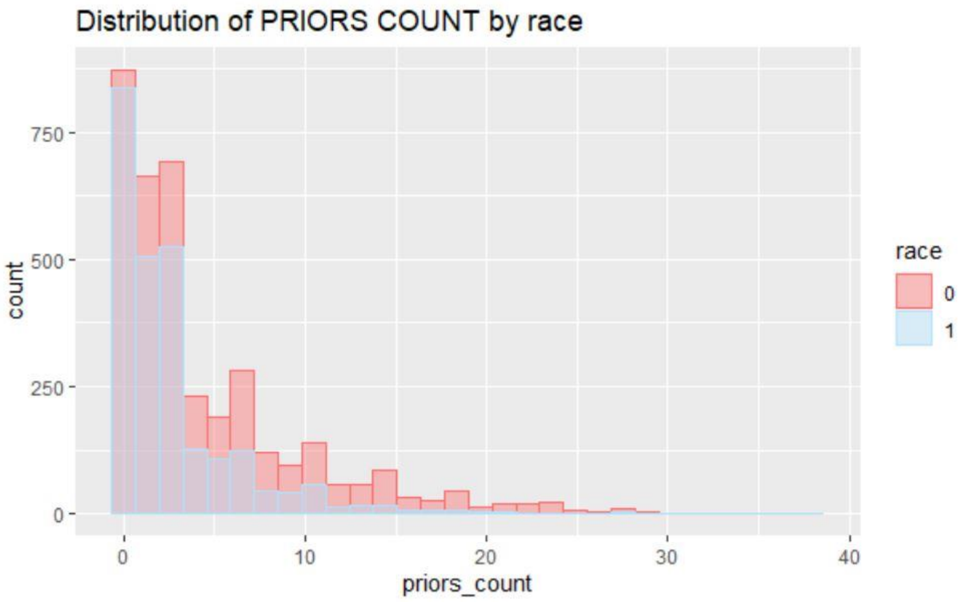
where sex: (1,0) = (*Male, Female*), race: (1,0) = (*Caucasian, African-American*),

charge degree: (1,0) = (*Felony, Misdemeanour*), target: (1,0) = (*At risk, No risk*),

two_year_recid: (1,0) = (*Actual recidivist, Not a recidivist*).

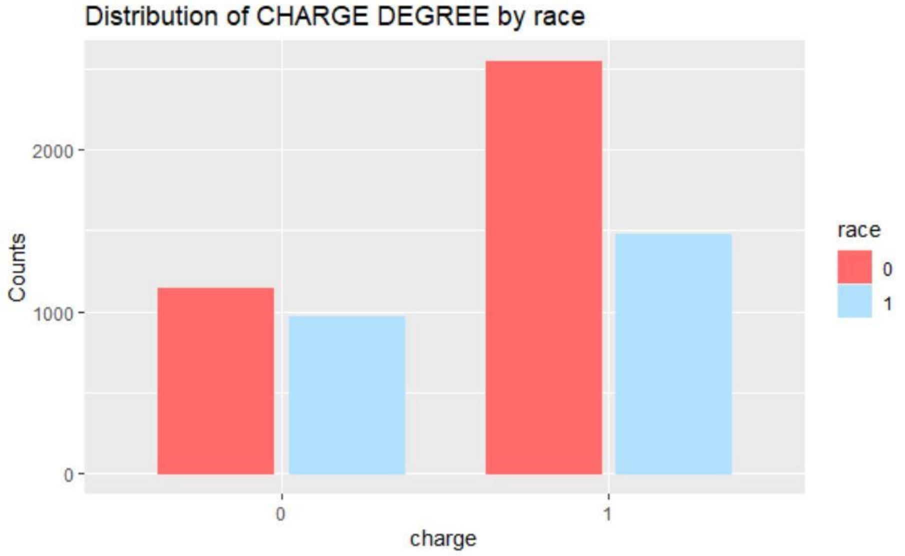
Number of observations: 6150.

Looking at the data



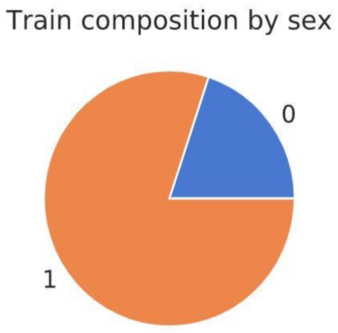
The training test split is 60-40, random and stratified with respect to the target (COMPAS risk classification), resulting in a training set of 3.690 observation and a test set of 2.460.

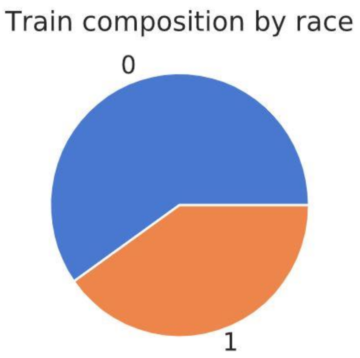
The test set presents the same structure of the training set in terms of variable race and sex.



1 = Caucasian | Men
0 = African American | Women

CHARGE DEGREE
1 = Felony
0 = Misdemeanour





1. Understanding the data: US Mass Incarceration and racial disparities
2. COMPAS dataset
3. **Classification task**
4. Generative Adversarial Network (GAN)
5. Conclusions

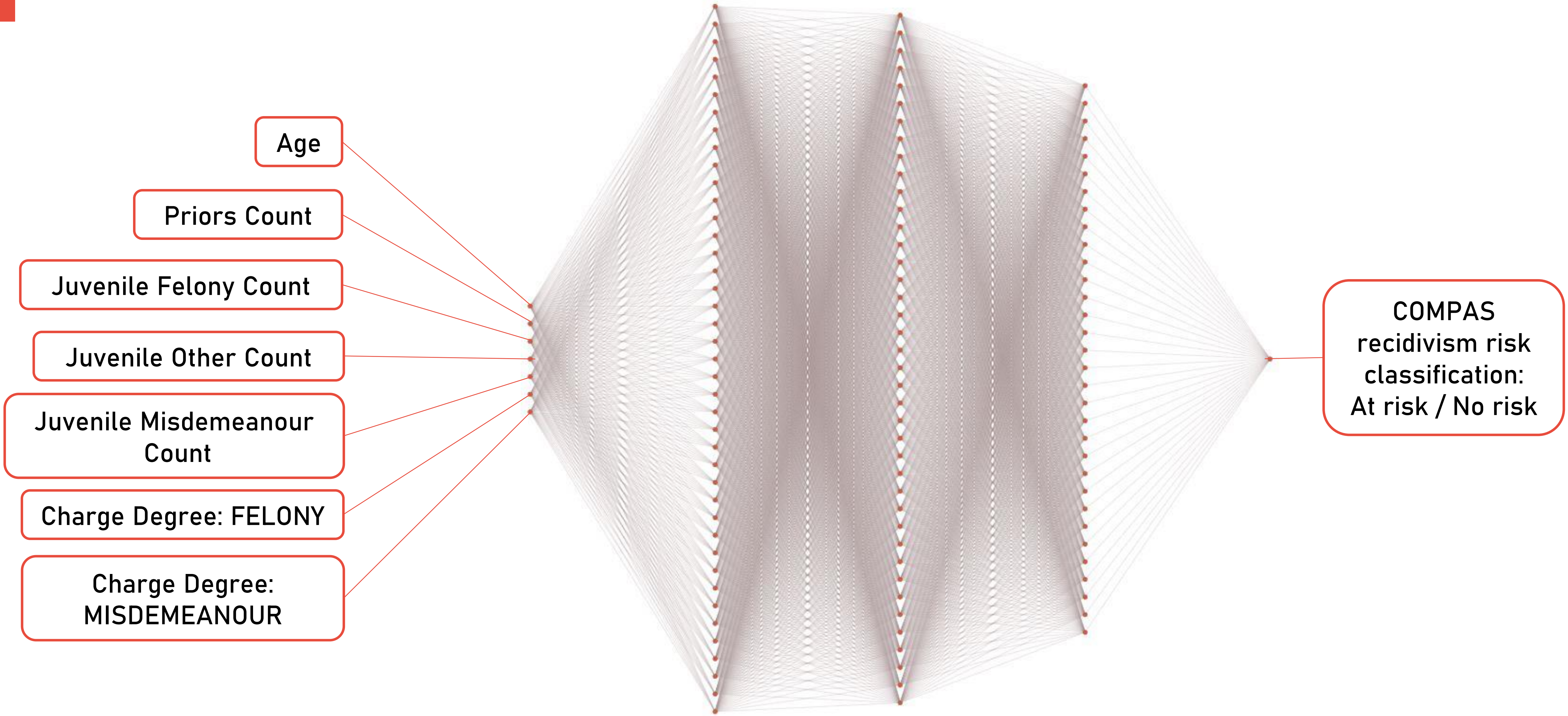
Feed Forward NN

I trained a Feed Forward NN to classify an individual depending on his/her recidivism risk. Variables sex, race and real recidivism are not considered.
Parameters were selected after a grid search.

- ❑ INPUT VARIABLES: age, juvenile felony count, juvenile misdemeanour count, juvenile other count, priors count, charge degree
- ❑ TARGET: At risk , Not at risk (*COMPAS classification*)

HIDDEN LAYERS:	3
NODES OF THE HIDDEN LAYERS:	[40, 40, 32]
DROPOUT	[0, 0.1, 0]
ACTIVATION FUNCTIONS (HIDDEN)	ReLu
ACTIVATION FUNCTION (OUTPUT)	Sigmoid
LOSS FUNCTION	Binary Cross-Entropy
OPTIMIZATION	Adam
EPOCHS	50

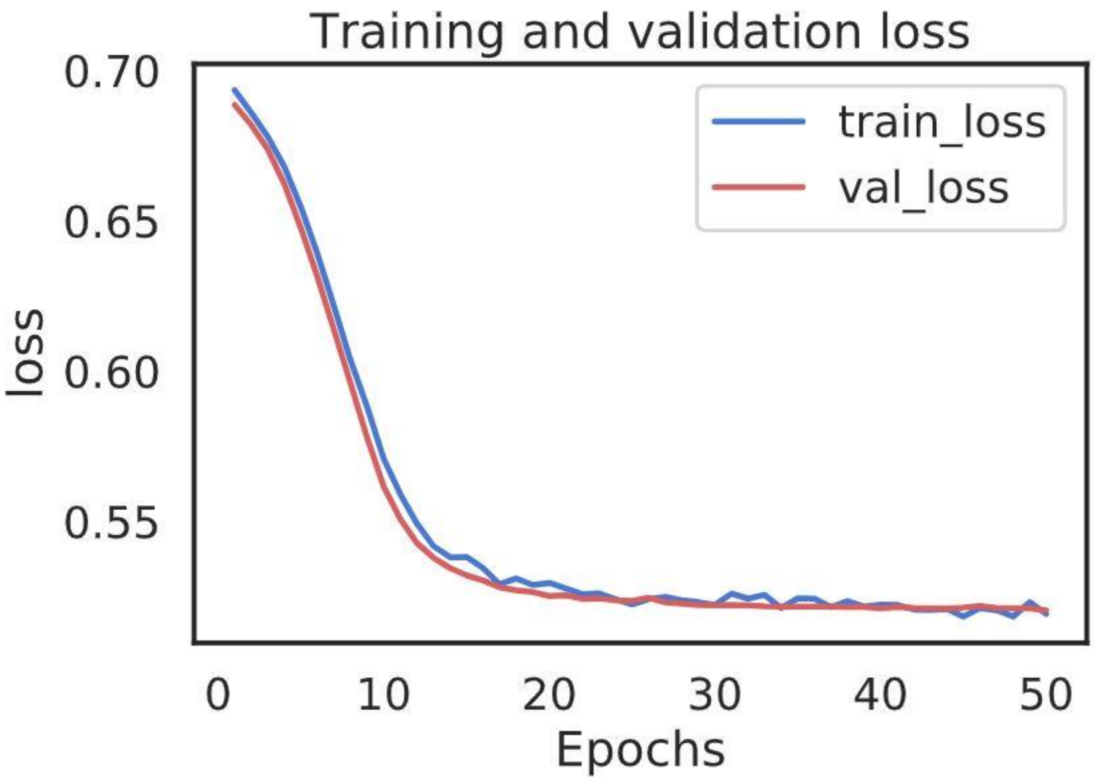
The architecture

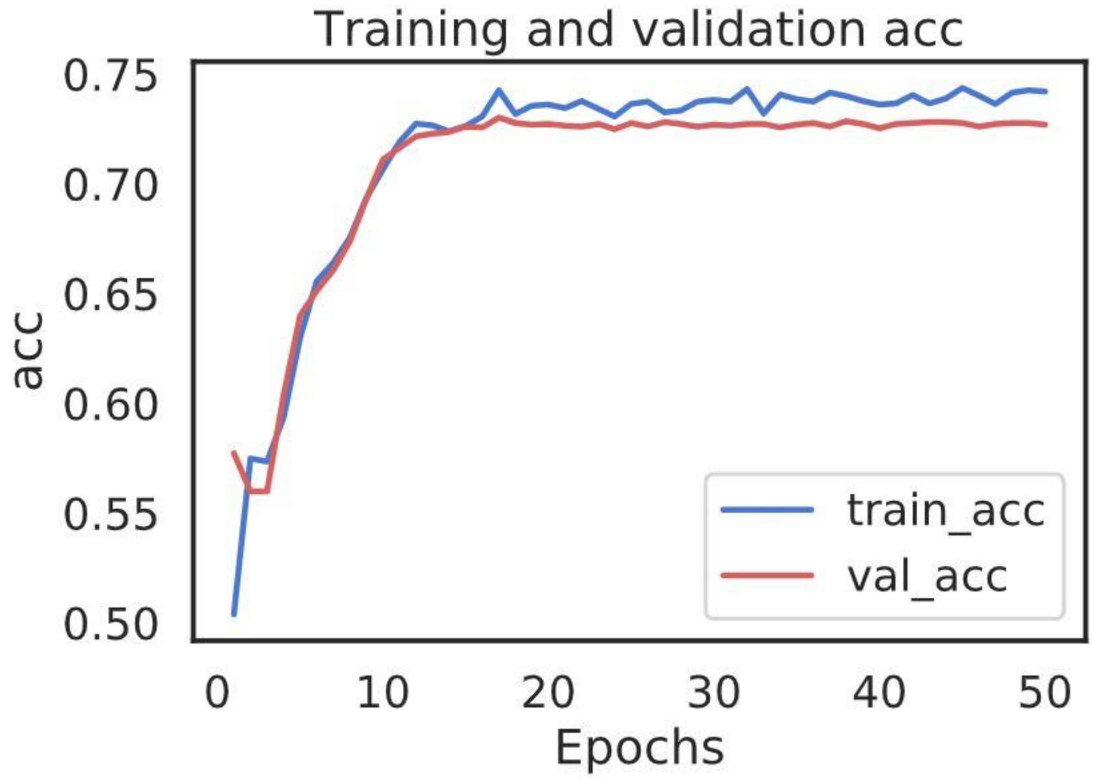


*input variables are normalized before training

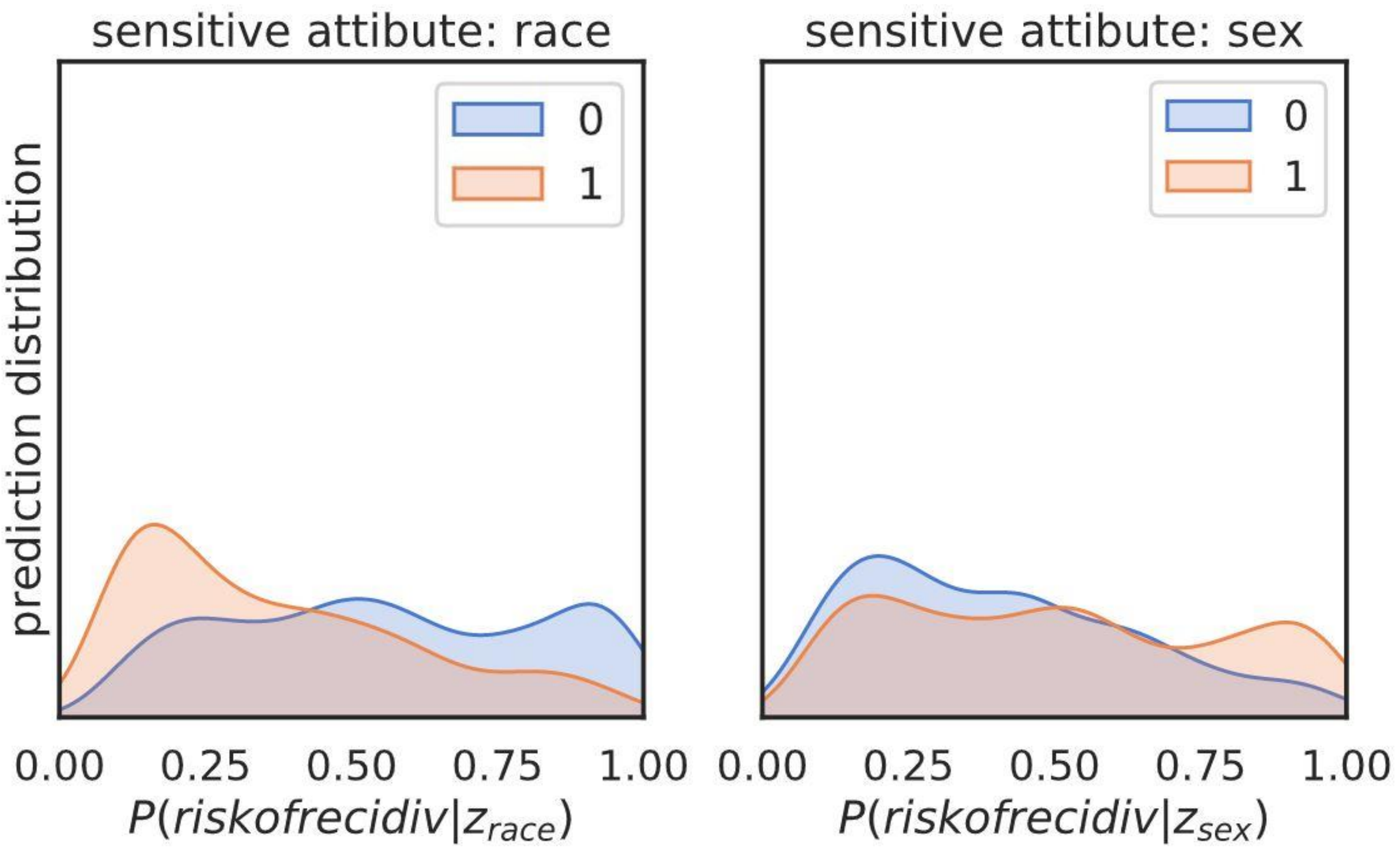
Evaluation

After standardizing the data and training my network, I obtained a validation accuracy of the model around 73%.





Are the predictions fair?



0 = African American
1 = Caucasian

0 = Women
1 = Men

- Predictions seem to be more fair when looking at variable sex, than variable race
- The prediction distribution for Caucasian defendants suggests that is more likely for a white person to be predicted as *NO RISK* with respect to a black person. Also, the prediction distribution for African American defendants is always higher than the one of Caucasian defendants for values above 0.5
- For women it's slightly more likely to be predicted as NO RISK with respect to men

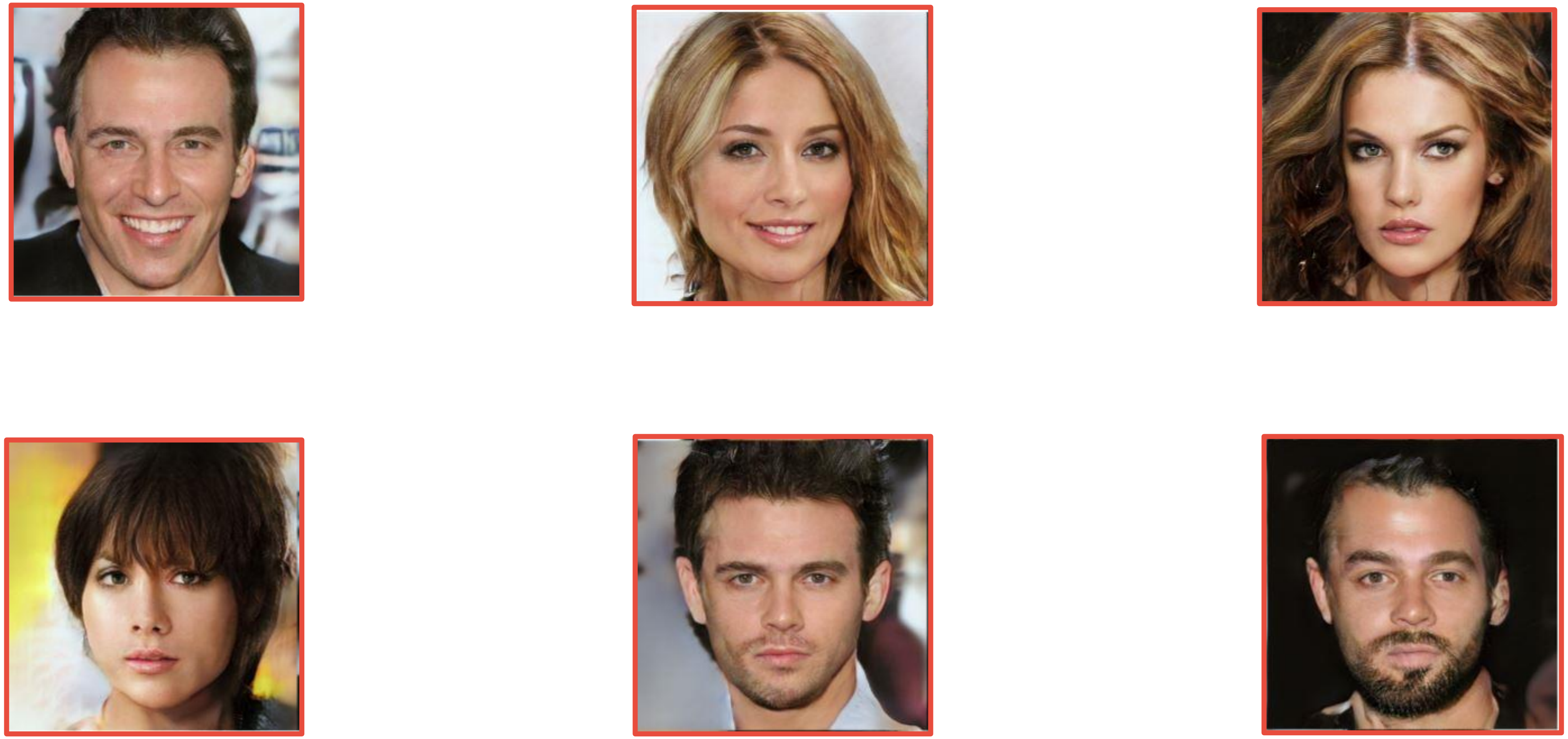
Let's look at some metrics

! The following metrics are computed comparing the prediction of the network with the actual recidivism behaviour of the individual within two years from the arrest, like ProPublica did ($\neq$ *model metrics*)

	METRIC	OVERALL	CAUCASIAN	AFRICAN AMERICAN
	Overall Accuracy	0.67	0.66	0.67
$\frac{TP}{P}$	True Positive Rate	0.62	0.43	0.72
$\frac{TN}{N}$	True Negative Rate	0.70	0.80	0.62
$\frac{FP}{N}$	False Positive Rate	0.23	0.20	0.38
$\frac{FN}{P}$	False Negative Rate	0.38	0.57	0.28

1. Understanding the data: US Mass Incarceration and racial disparities
2. COMPAS dataset
3. Classification task
4. **Generative Adversarial Network (GAN)**
5. Conclusions

Can you recognize these celebrities?

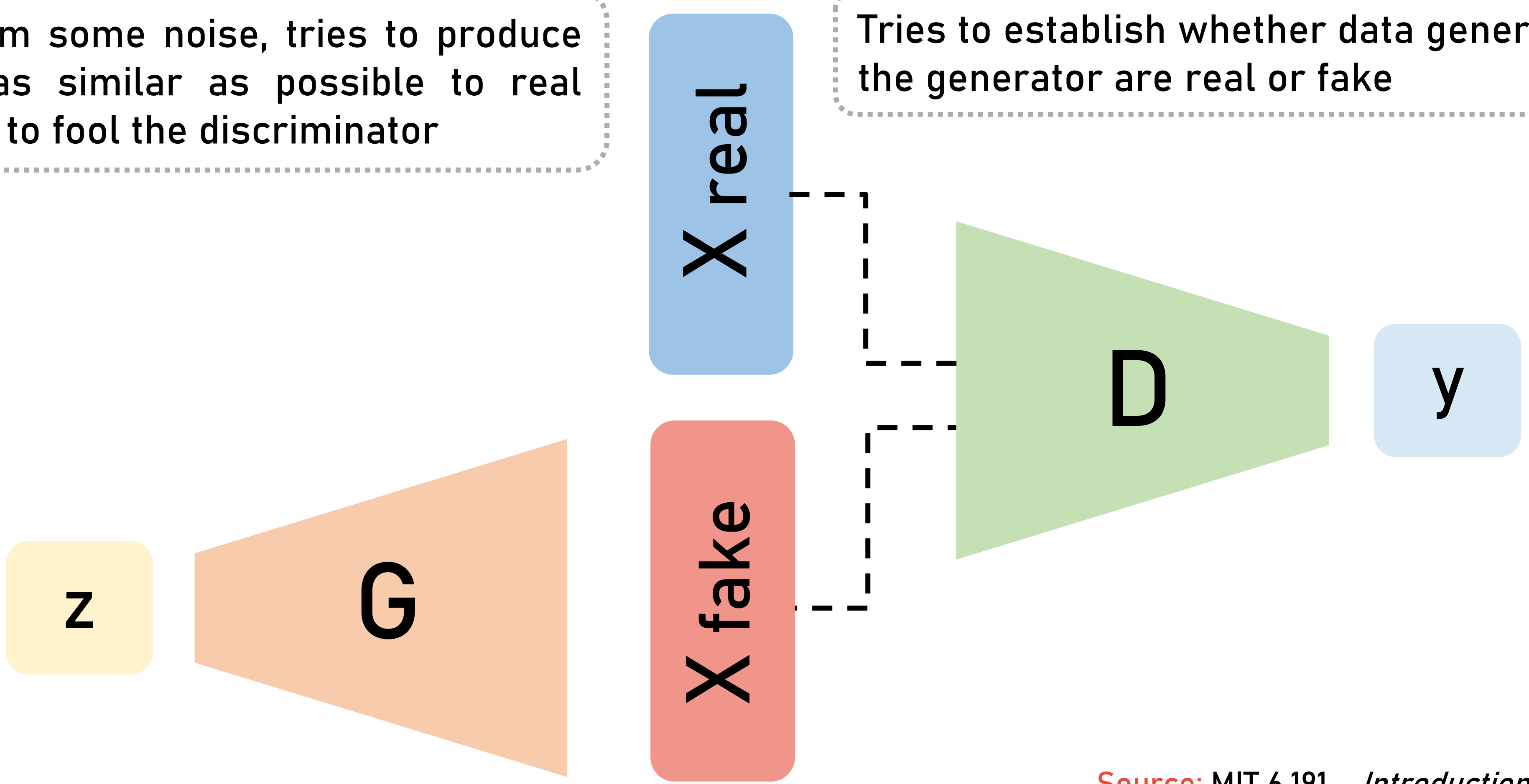


How does it work?

GENERATOR ⟷ **DISCRIMINATOR**

Starting from some noise, tries to produce fake data as similar as possible to real ones, trying to fool the discriminator

Tries to establish whether data generated by the generator are real or fake



Source: MIT 6.191 – *Introduction to Deep Learning.*

DISCRIMINATOR

- Fake data
- Real data

GENERATOR

Starts from some noise to try to create an imitation of true data

1

Looks at both fake data created by G and real data and assigns probabilities

2

As D is trained it improves its predictions

3

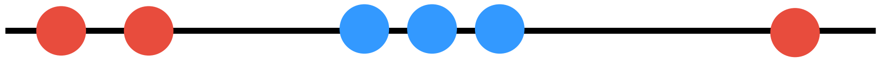
DISCRIMINATOR

GENERATOR

Fake data
Real data

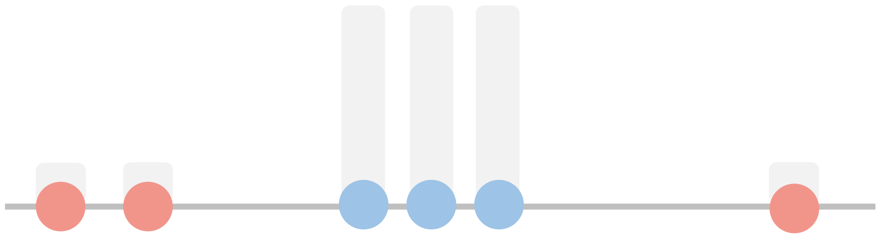
G looks at the predictions of D and the error it made when classifying data as real or fake

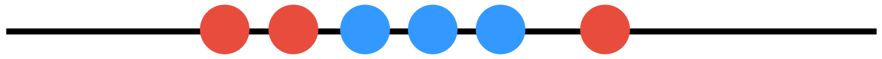
4



Tries to improve its imitation of real data in the attempt of fooling D

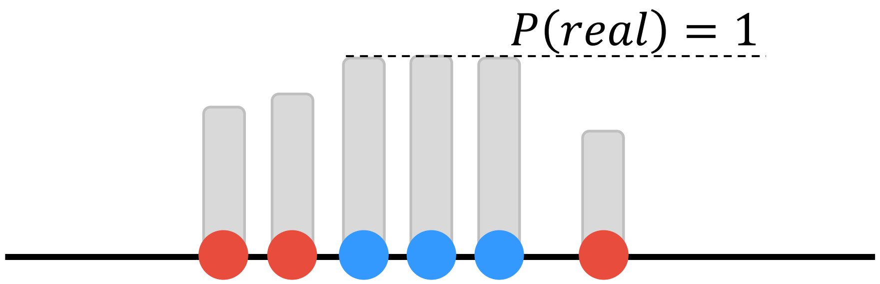
5

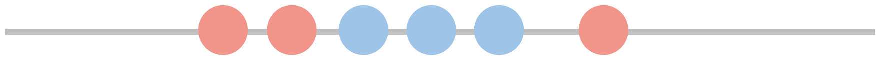




Again, D tries to predict what's real and what's fake

6





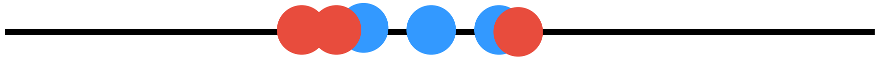
DISCRIMINATOR

GENERATOR

Fake data
Real data

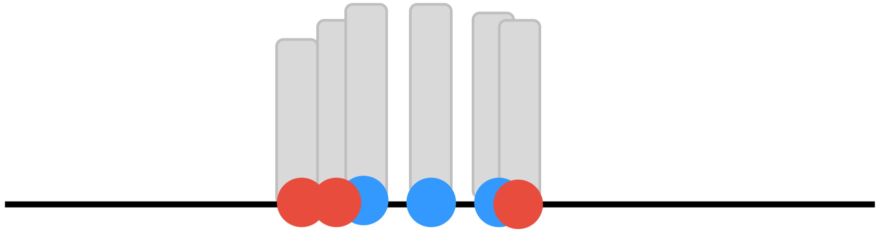
G tries to improve its imitation of real data in the attempt of fooling D

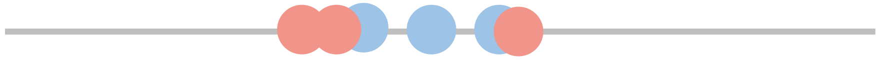
7



For D becomes more and more challenging to predict if data are real or fake

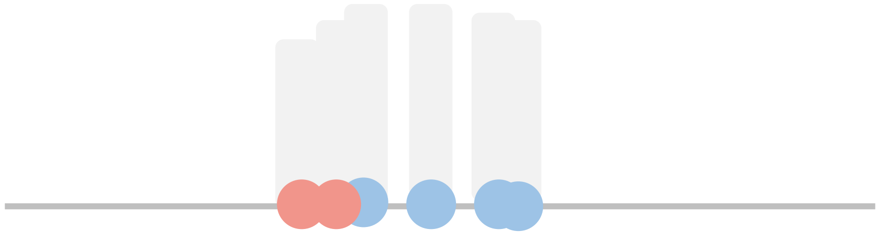
8

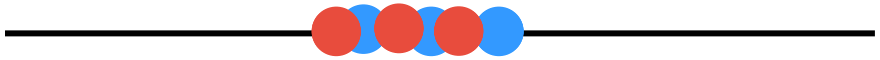




G becomes better and better in generating data similar to the real ones

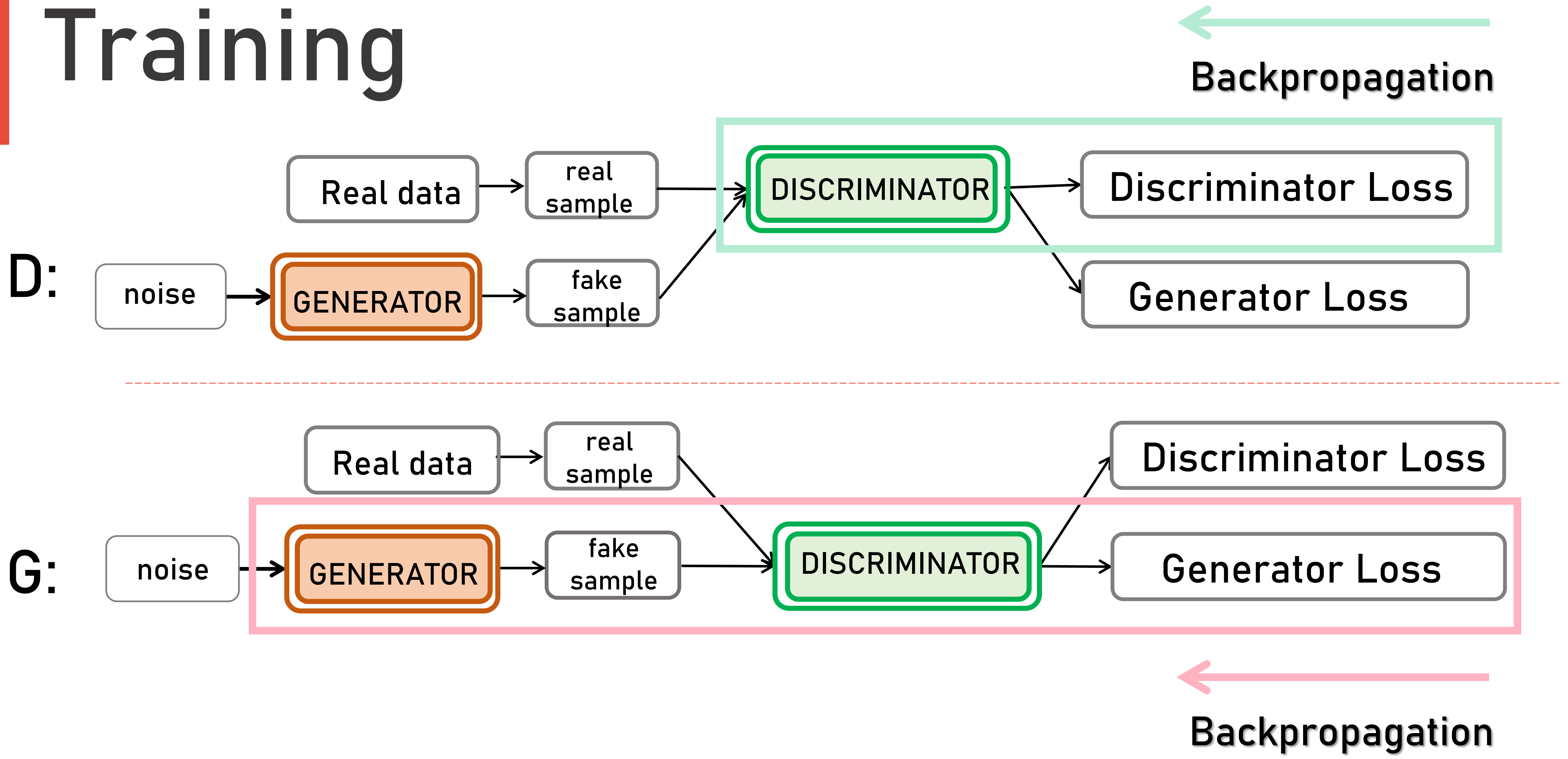
9





...

Training



The global loss function is the Minimax: $E_x[log(\boxed{D(x)})] + E_z[log(1 - \boxed{D(\boxed{G(z)})})]$

How can GAN be useful to mitigate bias?

- The generator is a predictor, that is, the Feed Forward NN constructed before.

GOAL: producing predictions of the target, i.e. COMPAS recidivism risk

- The discriminator (adversarial network) takes as input the prediction of the classifier. It is a FFNN as well.

GOAL: predict the race and sex of that particular observation

---

The generator becomes better and better in fooling the adversarial; for which becomes more and more difficult to predict race and sex based on the output of the classifier : the results become more and more fair

---

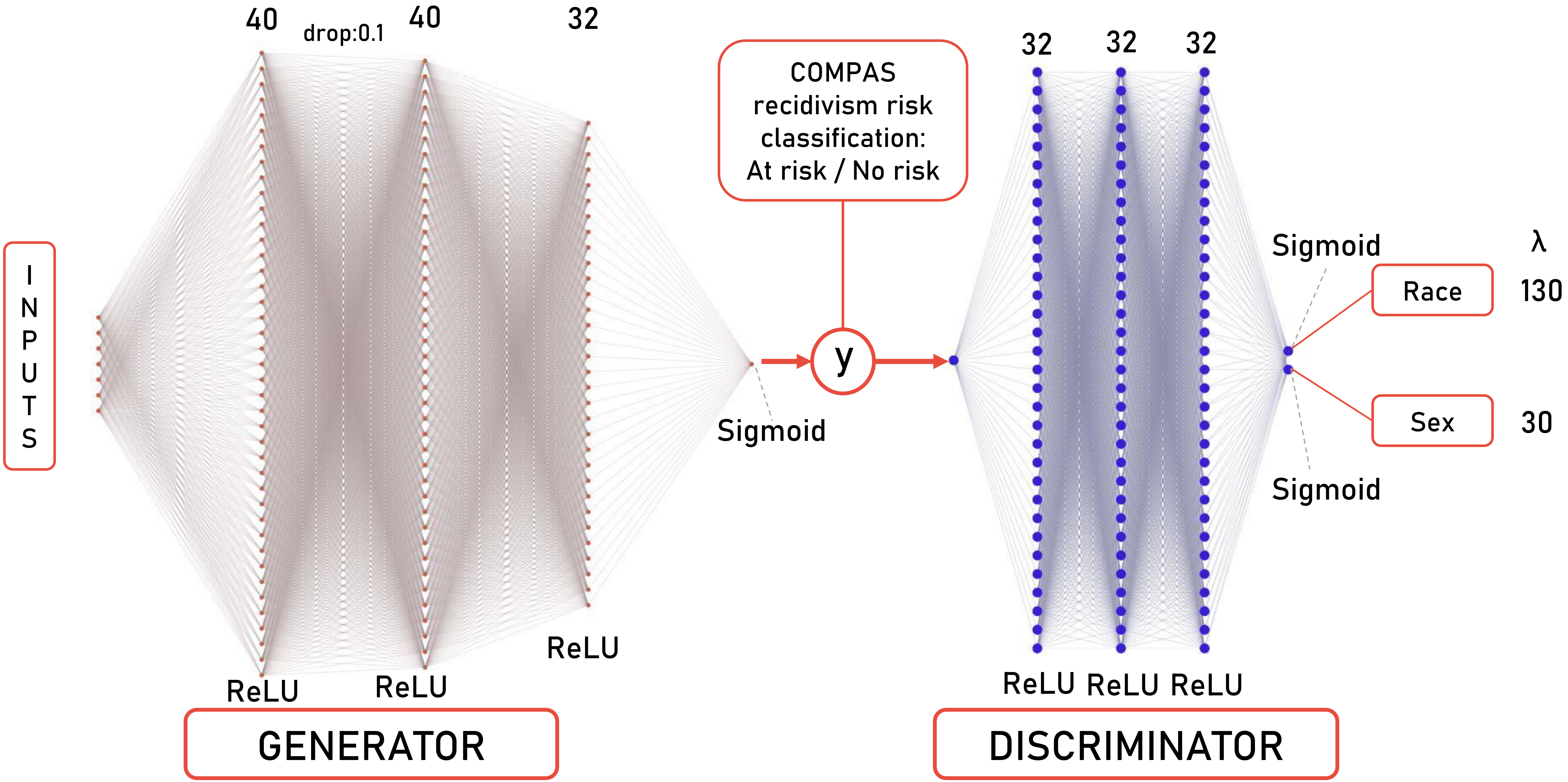
LOSS FUNCTIONS:

$$\boldsymbol{min_{\theta_{Gen}}[Loss_y(\theta_{Gen}) - \lambda\, Loss_Z(\theta_{Gen}, \theta_{Adv})]}$$

$$\boldsymbol{min_{\theta_{Adv}}[\,Loss_Z(\theta_{Gen}, \theta_{Adv})]}$$

y : COMPASS rec. risk | Z = (race, sex) | $\lambda > 0$

Generative Adversarial NN



Can we (try to) measure fairness?

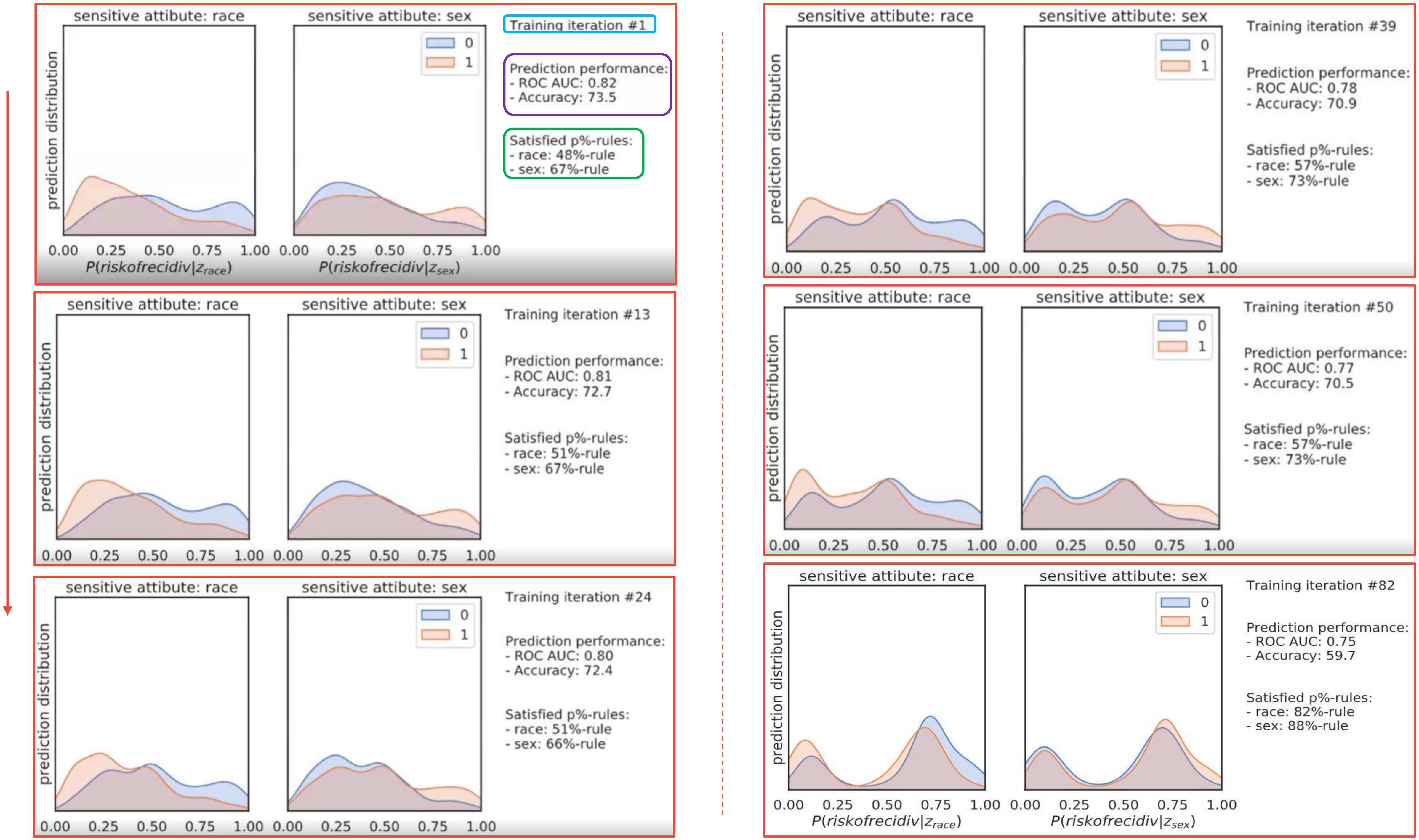
A classifier that makes a binary prediction $y \in \{0,1\}$ given some sensible attribute $z \in \{0,1\}$ satisfies the p% rule if

$$\min\left(\frac{P(\hat{y}=1|z=1)}{P(\hat{y}=1|z=0)}, \frac{P(\hat{y}=1|z=0)}{P(\hat{y}=1|z=1)}\right) \geq \frac{p}{100}$$

If a classifier is completely fair it satisfies a 100%-rule; when it is completely unfair it satisfies a %0-rule.

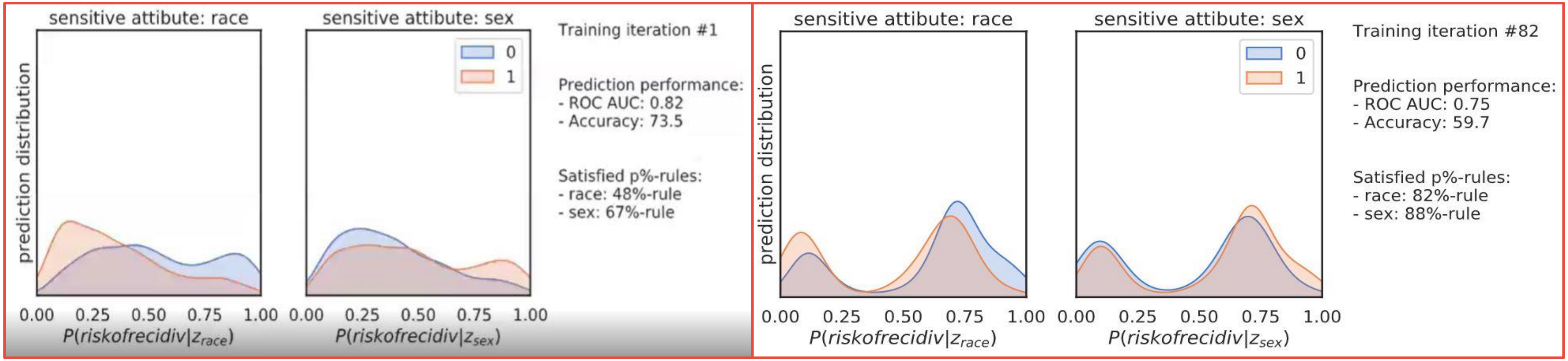
According to the literature, we can generally say that a classifier is fair if it satisfies at least a 80%- rule.

PREVIOUS FFNN RESULTS: satisfied 51%-rule for race and 67%-rule for sex



sensitive attibute: race
sensitive attibute: sex
Training iteration #1
Prediction performance:
- ROC AUC: 0.82
- Accuracy: 73.5
Satisfied p%-rules:
- race: 48%-rule
- sex: 67%-rule
prediction distribution
0
1
0.00 0.25 0.50 0.75 1.00
P(riskofrecidiv|z_race)
P(riskofrecidiv|z_sex)
Training iteration #13
Prediction performance:
- ROC AUC: 0.81
- Accuracy: 72.7
Satisfied p%-rules:
- race: 51%-rule
- sex: 67%-rule
Training iteration #24
Prediction performance:
- ROC AUC: 0.80
- Accuracy: 72.4
Satisfied p%-rules:
- race: 51%-rule
- sex: 66%-rule
Training iteration #39
Prediction performance:
- ROC AUC: 0.78
- Accuracy: 70.9
Satisfied p%-rules:
- race: 57%-rule
- sex: 73%-rule
Training iteration #50
Prediction performance:
- ROC AUC: 0.77
- Accuracy: 70.5
Satisfied p%-rules:
- race: 57%-rule
- sex: 73%-rule
Training iteration #82
Prediction performance:
- ROC AUC: 0.75
- Accuracy: 59.7
Satisfied p%-rules:
- race: 82%-rule
- sex: 88%-rule

Visualization of the Accuracy-Bias tradeoff



The plot on the left coincides to the one obtained using the Feed Forward NN. As sad before, the prediction distribution is not fair.

When the classifier interacts with the discriminator, we obtain a more fair distribution, both with respect to race and sex, but model accuracy drops to 60%.

How have the metrics changed?

METRIC	OVERALL	CAUCASIAN	AFRICAN AMERICAN	DIFFERENCE BTW RACES (abs. val.)
Overall Accuracy	0.55	0.53	0.57	0.04
	0.67	0.66	0.67	0.01
True Positive Rate	0.75	0.67	0.79	0.12
	0.62	0.43	0.72	0.29
True Negative Rate	0.40	0.45	0.33	0.12
	0.71	0.81	0.62	0.19
False Positive Rate	0.59	0.55	0.67	0.12
	0.29	0.19	0.38	0.19
False Negative Rate	0.25	0.33	0.21	0.12
	0.38	0.57	0.28	0.29

: Feed Forward NN , : Generative Adversarial Network

1 Understanding the data: US Mass Incarceration and racial disparities

2 COMPAS dataset

3 Classification task

4 Generative Adversarial Network (GAN)

5 **Conclusions**

Conclusions and further steps

- The results can be considered satisfying for the following reasons:
  - ❑ Accuracy drop was of 12%
  - ❑ Consequences of the algorithm in terms of people's lives
  - ❑ Good quality of data was not entirely verified

- Further steps:
  - ❑ Find a way to automatize the process for the choice of parameters in the GAN architecture
  - ❑ More and better data for more stable results
  - ❑ Bias is often in the data but is not only a technical problem

References

❑ USING GAN TO REMOVE BIAS:

- G.Louppe, M. Kagan, K. Cranmer, Learning to Pivot with Adversarial Networks; 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA; notebook present on GitHub

- Stijn Tonk, Towards fairness in ML with adversarial networks (https://godatadriven.com/blog/towards-fairness-in-ml-with-adversarial-networks/); notebook present on GitHub

❑ PRO PUBLICA ANALYSIS on COMPAS:

- https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

WHERE YOU CAN FIND THE DATA: https://github.com/propublica/compas-analysis

❑ ABOUT GAN'S:

- https://developers.google.com/machine-learning/gan/gan_structure
- http://introtodeeplearning.com/slides/6S191_MIT_DeepLearning_L4.pdf

Thank you!