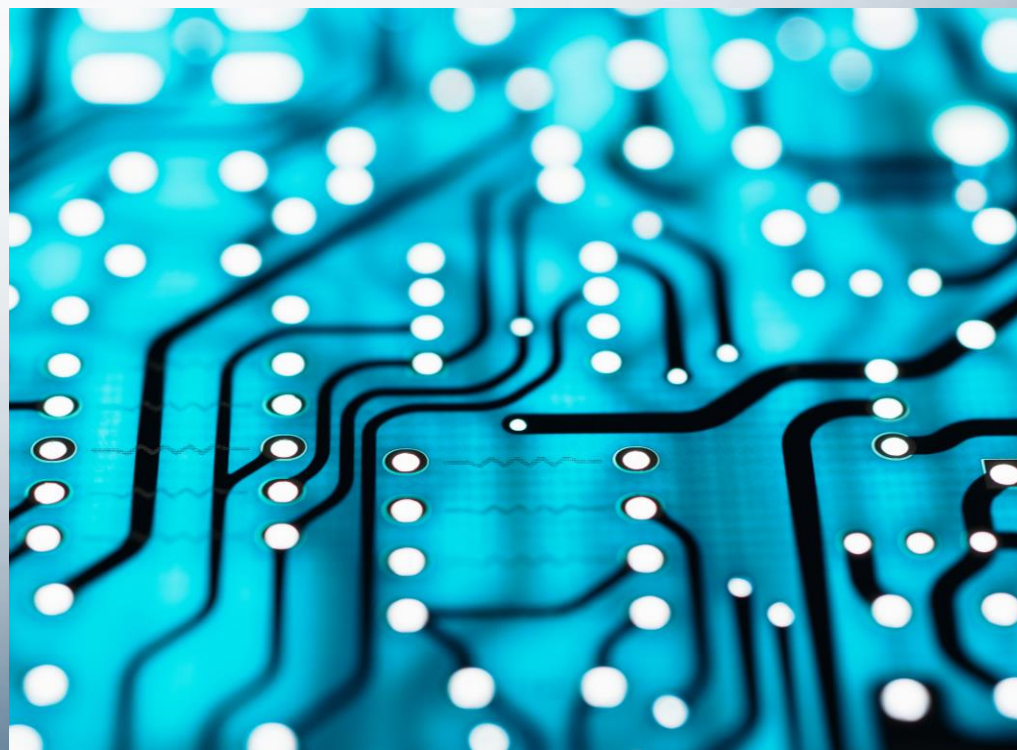


DATA FAIRNESS

LA NOTTE DEI RICERCATORI

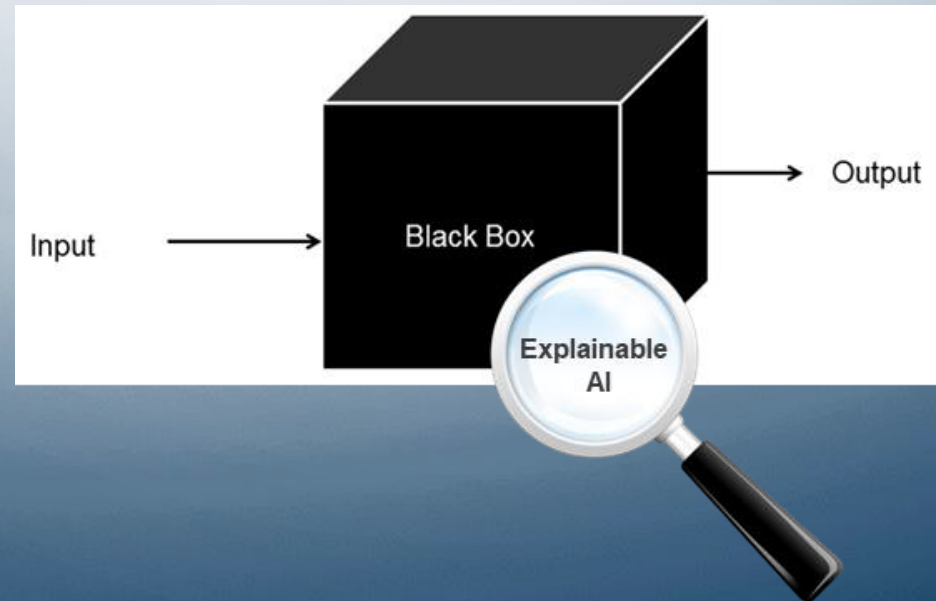
OUR RESEARCH GROUP

- XAI
- Bias and Fairness



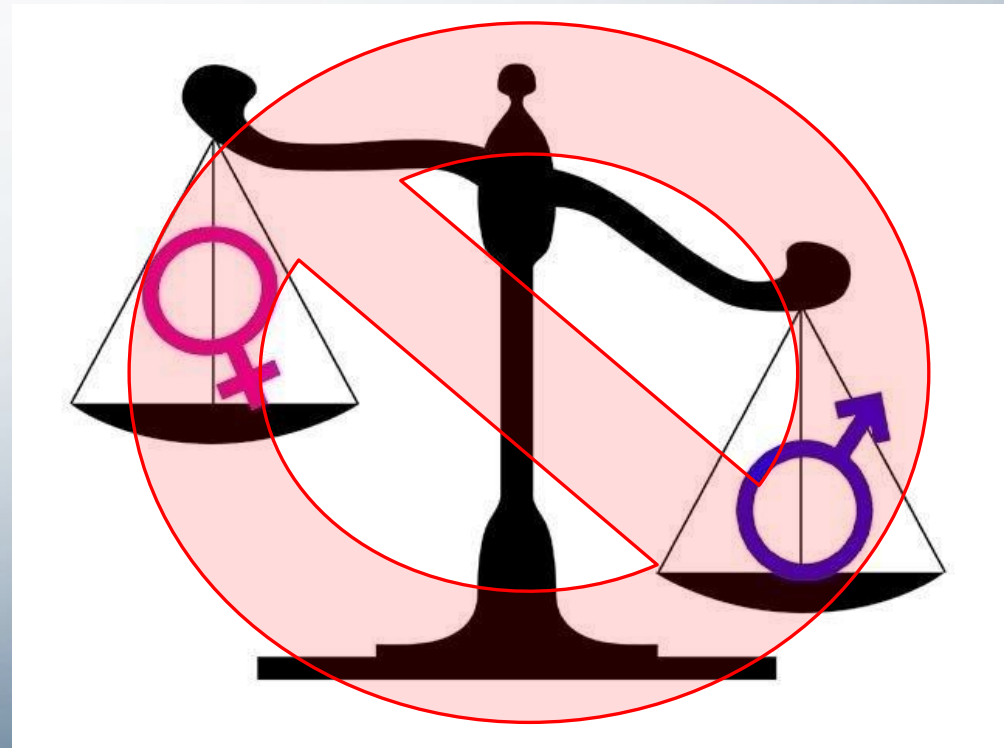
XAI?

“**Explainable AI (XAI)** is artificial intelligence (AI) in which the results of the solution can be understood by humans. It contrasts with the concept of the “black box” in machine learning where even its designers cannot explain why an AI arrived at a specific decision.”



BIAS

“Machine learning bias is a phenomenon that occurs when an algorithm produces results that are systemically prejudiced due to erroneous assumptions in the machine learning process.”



TYPE OF BIAS

- **RACIAL BIAS:** data favors a particular demographic
- **OBSERVER BIAS:** effect of seeing or searching what you expect to see in data
- **SAMPLE BIAS:** the training dataset does not reflect the realities of the environment in which a model will run

RACIAL BIAS: THE COMPAS CASE

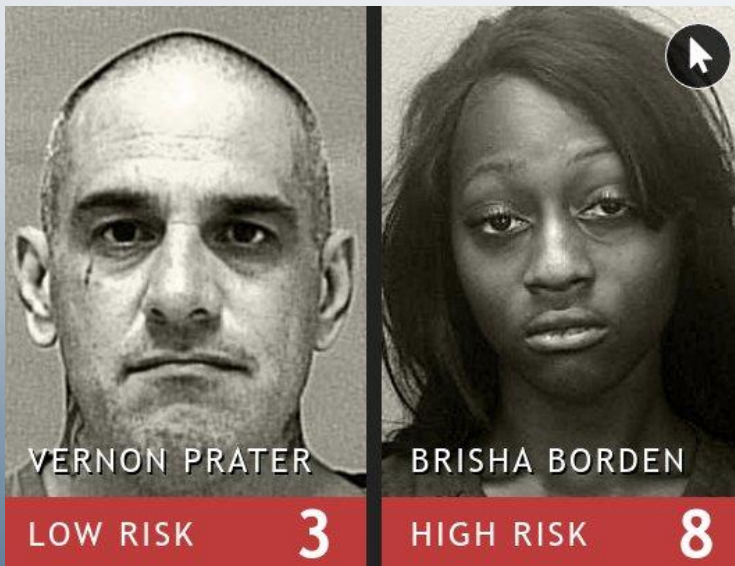
Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is decision support tool used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.

	name	sex	age	dob	race	in_custody	priors_count.1	v_decile_score
0	miguel hernandez	Male	69.0	1947-04-18	Other	2014-07-07	0.0	1.0
1	kevon dixon	Male	34.0	1982-01-22	African-American	2013-01-26	0.0	1.0
2	ed philo	Male	24.0	1991-05-14	African-American	2013-06-16	4.0	3.0
3	marcu brown	Male	23.0	1993-01-21	African-American	NaN	1.0	6.0
4	bouthy pierrelouis	Male	43.0	1973-01-22	Other	NaN	2.0	1.0

PROPUBLICA INVESTIGATION (2016)

In 2016, ProPublica reported that an artificial intelligence tool (COMPAS) used in courtrooms across the United States to predict future crimes was biased against Black defendants.

Two Petty Theft Arrests



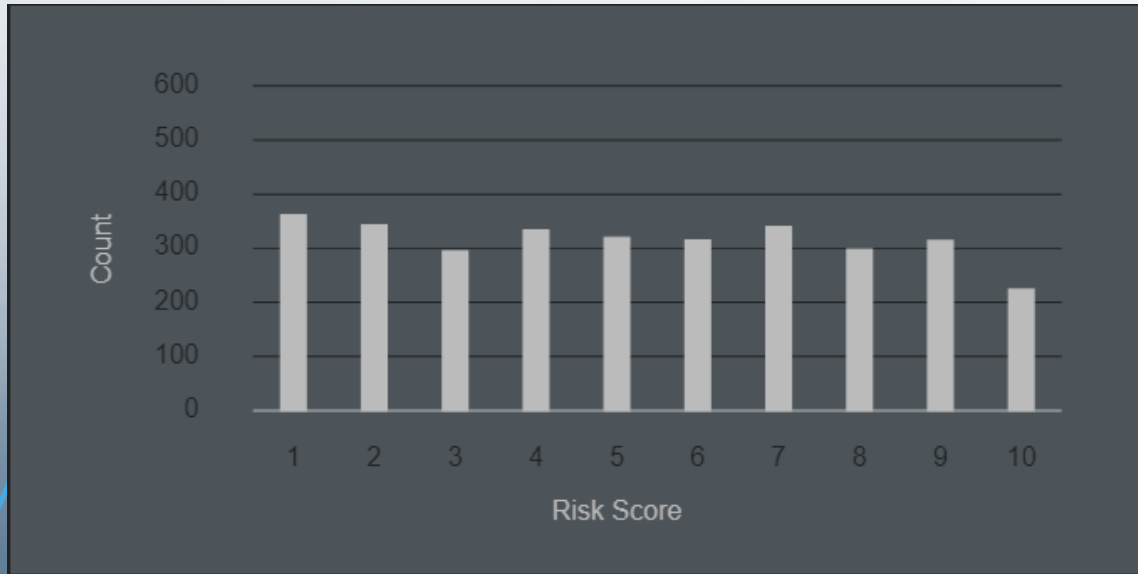
Two Drug Possession Arrests



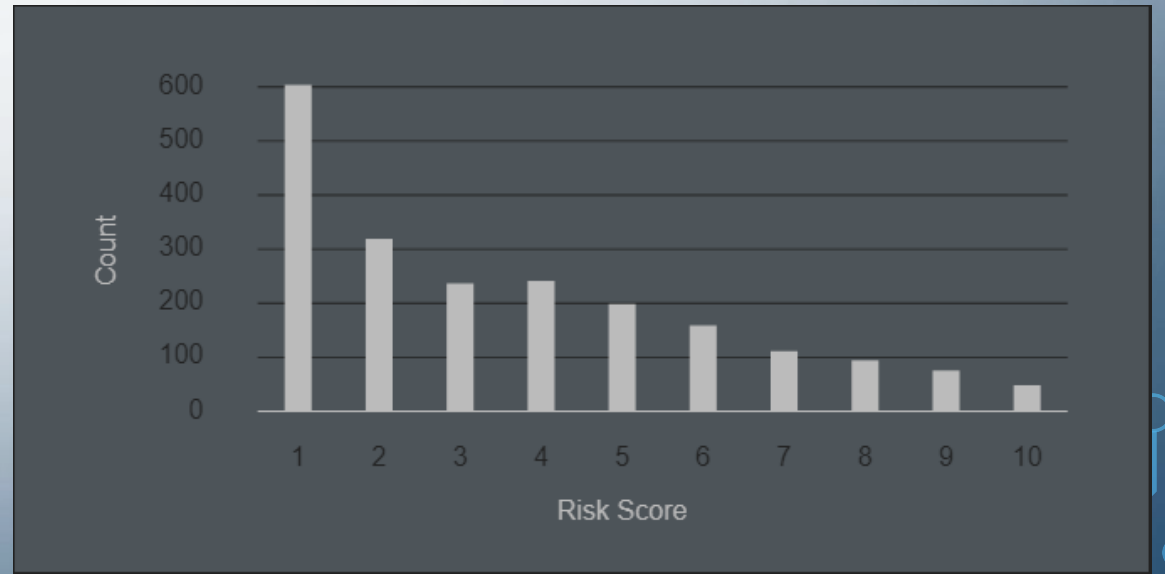
<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

RISK SCORES PREDICTED BY COMPAS

White Defendants' Risk Score



Black Defendants' Risk Score



RISK SCORES PREDICTED BY COMPAS

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%

Confusion Matrix: accuracy is about 65,3%

Actual two_year_recid	Predicted COMPAS Score		
	Low	High	
0	2681	1282	0.55
1	1216	2035	0.45






«COMPAS IS BIASED: THE ALGORITHM DISCRIMINATES BLACK PEOPLE»

How can we remove the bias?

- Post-processing techniques reassign labels based on a new function
- In-processing techniques try to remove discrimination during the model training process
- Pre-processing techniques try to transform the data so that the underlying discrimination is removed

Article

A Methodology for Controlling Bias and Fairness in Synthetic Data Generation

Enrico Barbierato ¹^{*}, Marco L. Della Vedova ², Daniele Tessera ¹, Daniele Toti ¹ and Nicola Vanoli ¹

¹ Catholic University of the Sacred Heart, Faculty of Mathematical, Physical And Natural Sciences, Brescia, Italy; {enrico.barbierato, daniele.tessera, daniele.toti, nicola.vanoli }@unicatt.it

² University of Bergamo, Dept. of Management, Information and Production Engineering, Bergamo, Italy; marco.dellavedova@unibg.it

* Correspondence: enrico.barbierato@unicatt.it

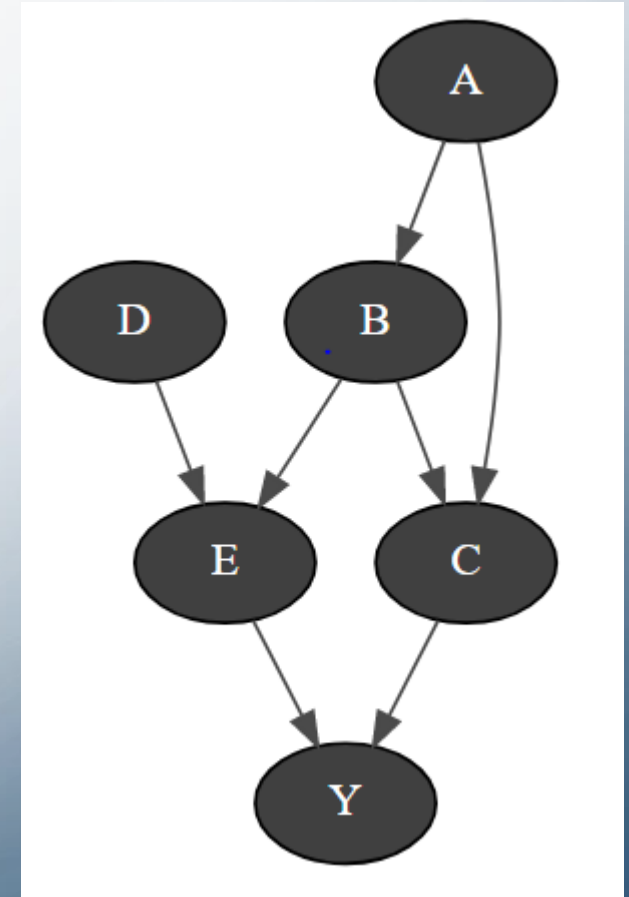
Abstract: The development of algorithms, based on machine learning techniques, supporting (or even replacing) human judgment must take into account concepts such as data bias and fairness. Though scientific literature proposes numerous techniques detecting and evaluating these problems, less attention has been dedicated to methods generating intentionally-biased datasets, which could be used by data scientists to develop and validate unbiased and fair decision making algorithms. To this end, this paper presents a novel method to generate a synthetic dataset, where bias can be modeled by using a probabilistic network exploiting Structural Equation Modeling. The proposed methodology has been validated on a simple dataset to highlight the impact of tuning parameters on bias and fairness, as well as on a more realistic example based on a loan approval status dataset. In particular, this methodology requires a limited number of parameters compared to other techniques for generating datasets with a controlled amount of bias and fairness.

WHY SYNTHETIC DATA?

- **Increased data quality:** real-world data is hard and expensive to source
- **Scalability:** Fueling the machine learning economy takes a huge amount of data.
- **Powerful simplicity:** you can control how the resulting data is structured, formatted and labeled

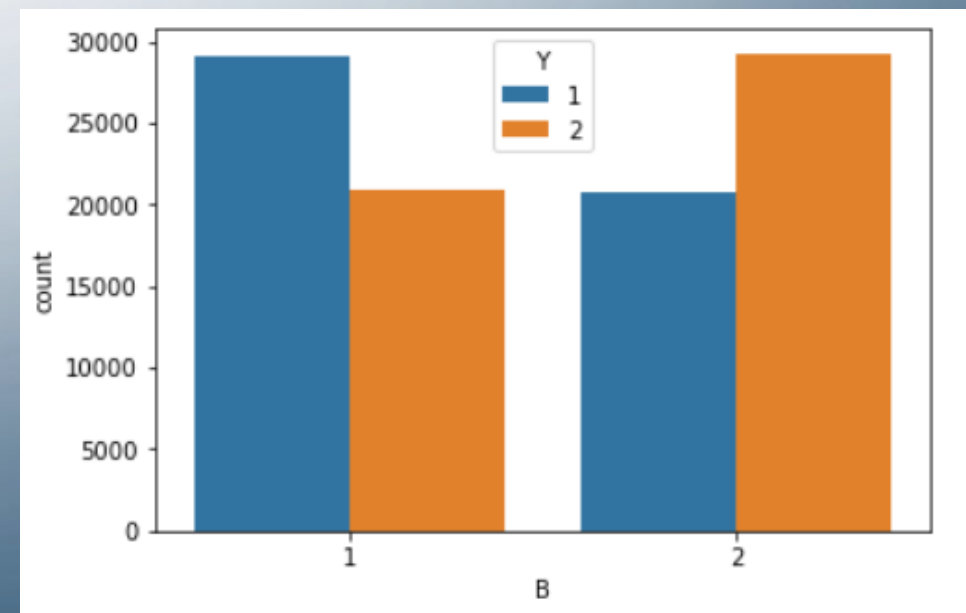
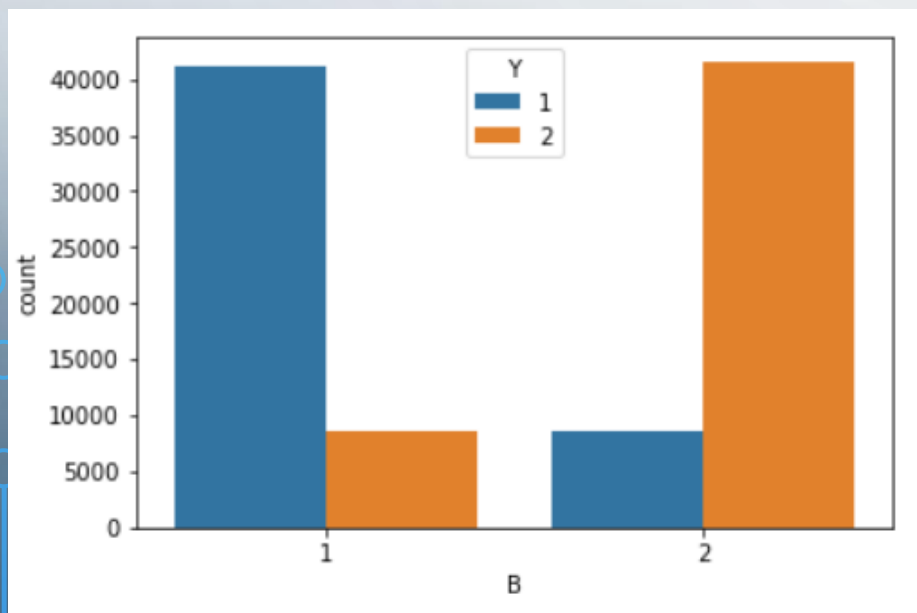
HOW TO GENERATE DATA

1. Define a probabilistic network that describes your data through a DAG (direct acyclic graph)
2. Initialize each node as a random variable that follows a sum of weighted gaussian distribution
3. Sample each observation from a multivariate normal distribution
4. Convert the numerical values into categorical



HOW TO CONTROL THE BIAS

1. When initializing the network, the weight of each edge must be manually inserted
2. Higher weights means higher correlation between the connected variables
3. How does this change if the strength of connections between B and Y is lowered?





THANK YOU!