IFBiD: Inference-Free Bias Detection

Ignacio SERNA

Daniel DEALCALA

Aythami MORALES

Julian FIERREZ

Javier ORTEGA-GARCIA

BiDA Lab
Biometrics & Data Pattern Analytics Lab

Universidad Autónoma de Madrid

AAAI-22 SafeAI Workshop

Newspaper Headlines

Universidad Autónoma de Madrid

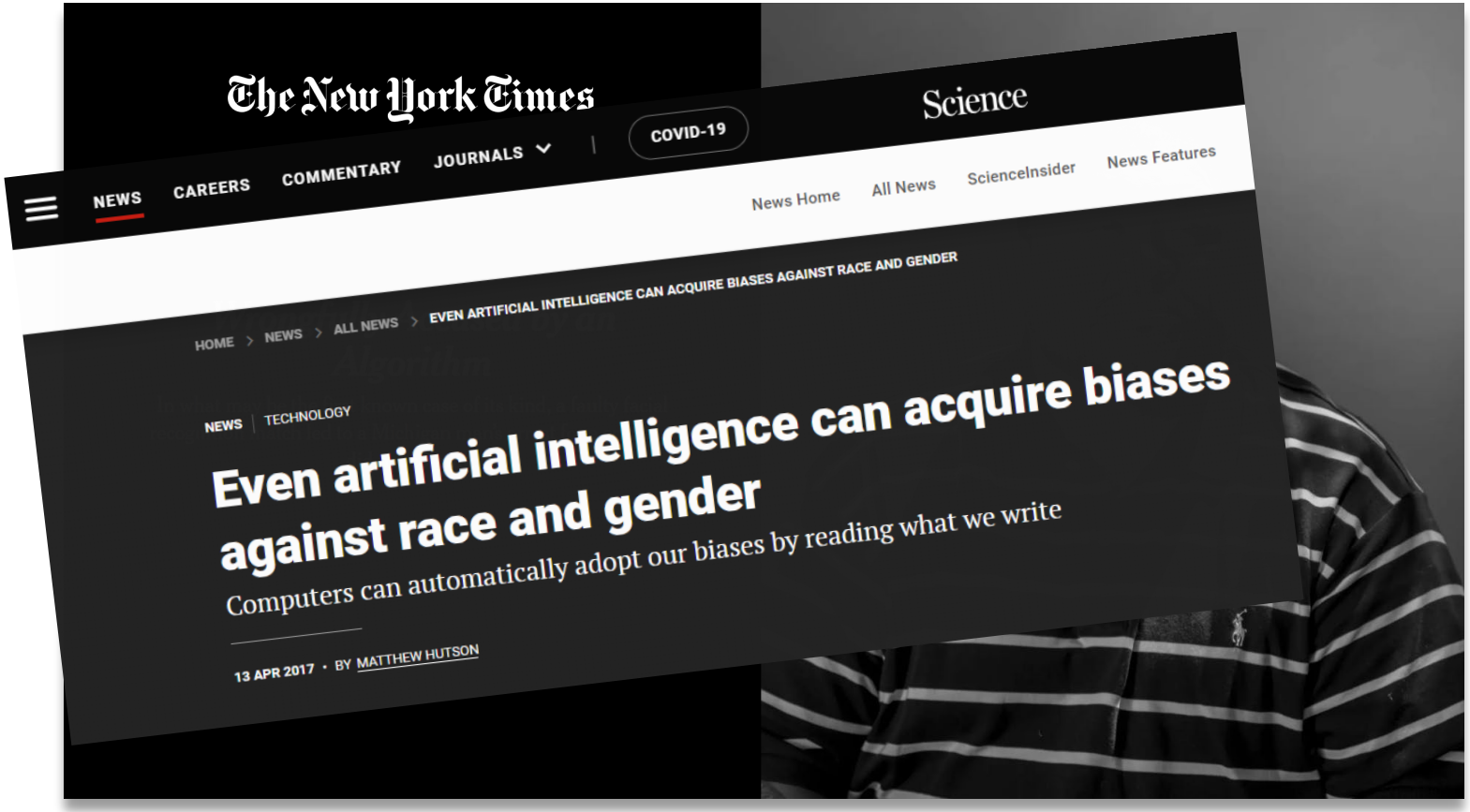
The New York Times

Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

Newspaper Headlines

Universidad Autónoma de Madrid

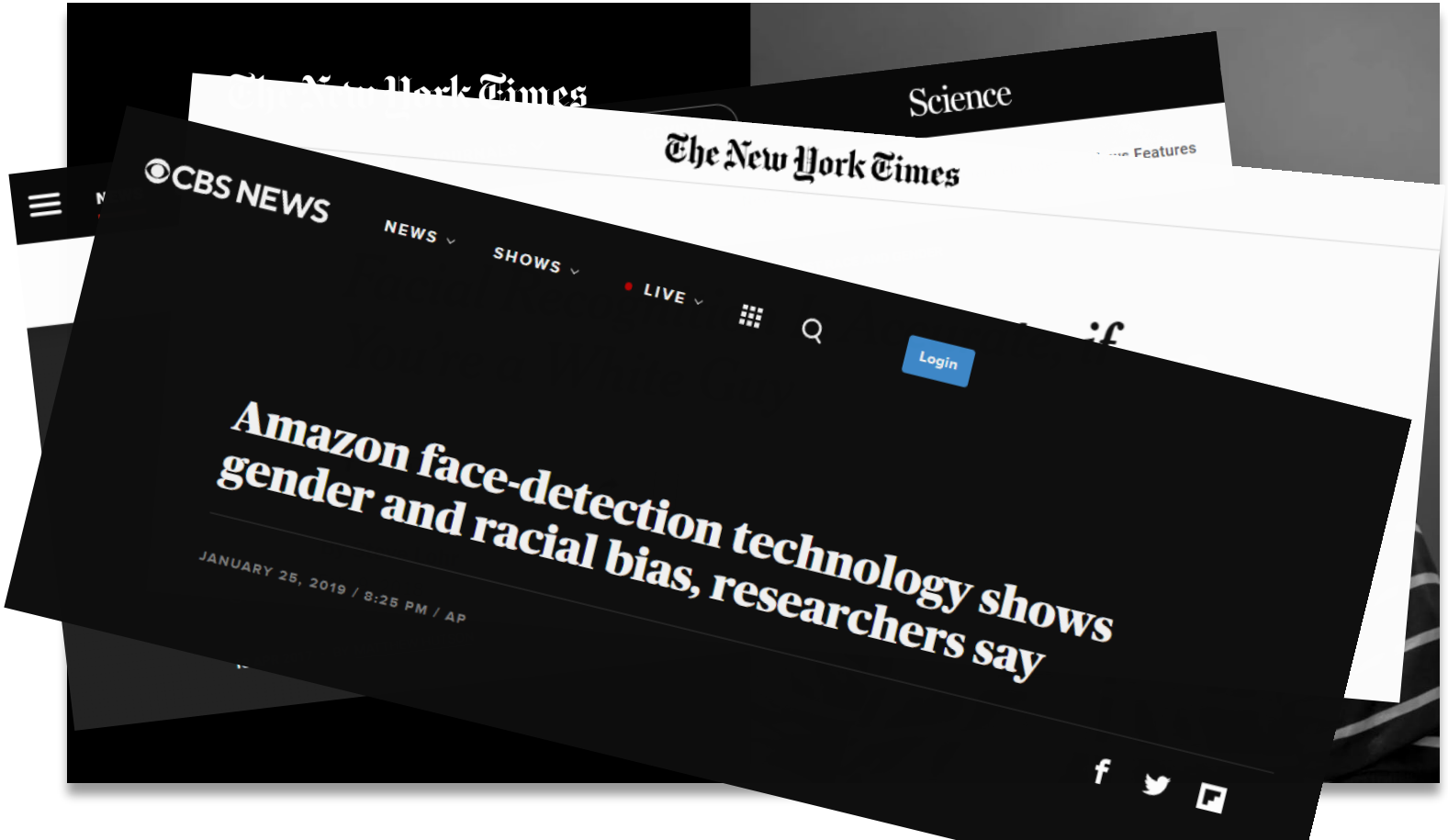


Newspaper Headlines

Universidad Autónoma de Madrid

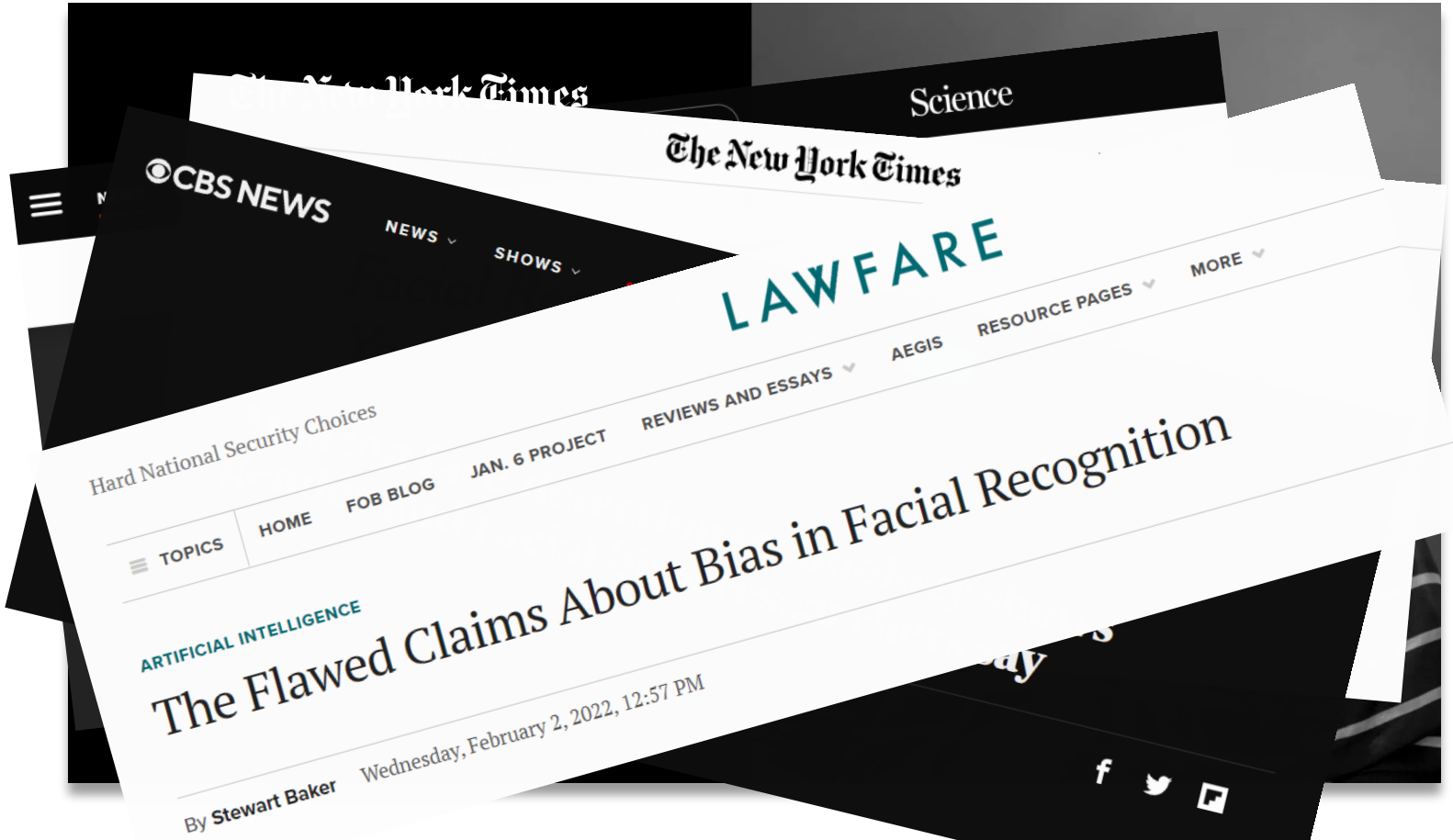
Newspaper Headlines

Universidad Autónoma de Madrid



Newspaper Headlines

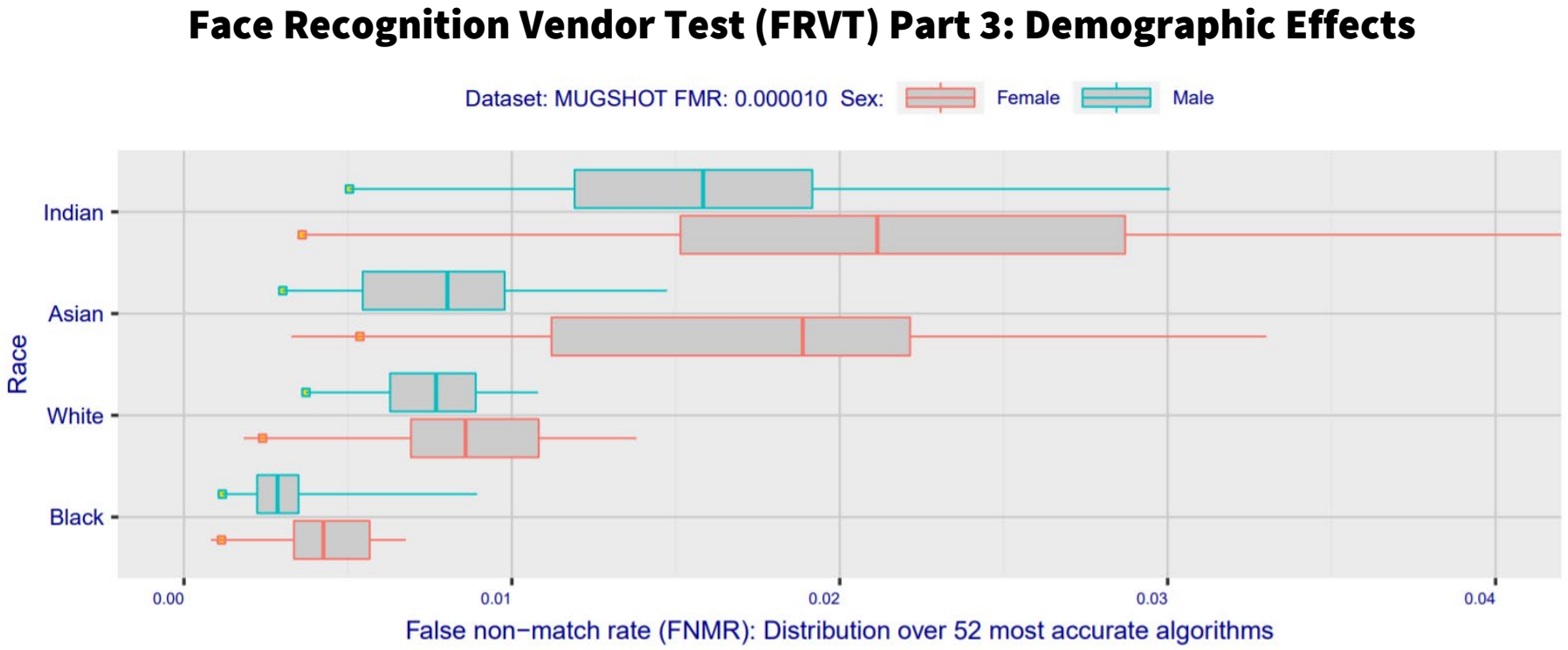
Universidad Autónoma de Madrid



How do we measure bias?

Traditionally: TEST SET

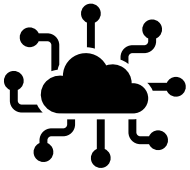
Universidad Autónoma de Madrid



National Institute of Standards and Technology
U.S. Department of Commerce

Traditionally: TEST SET

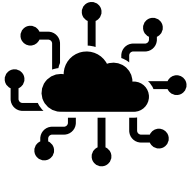
Gender Model

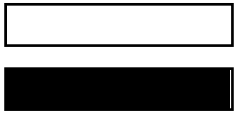


MALE

FEMALE

Gender Model





MALE

FEMALE

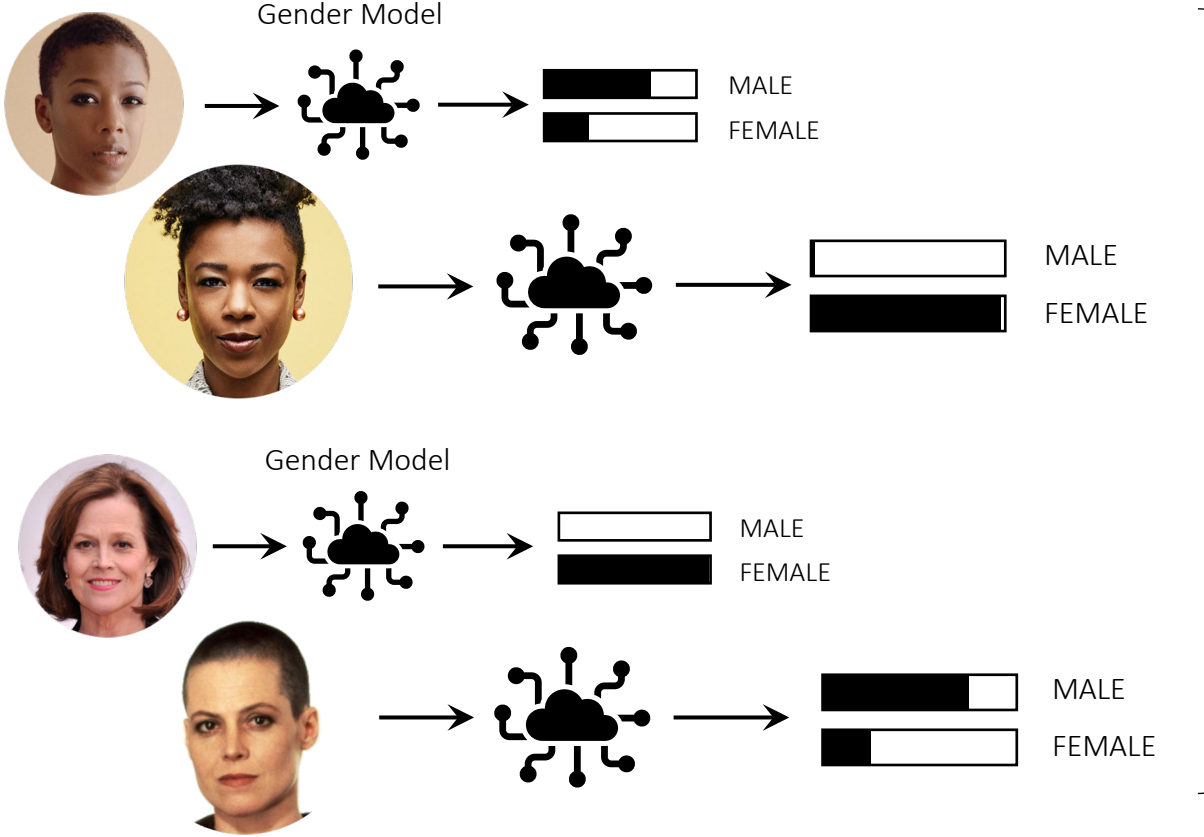
BIASED

AGAINST

BLACK

PEOPLE

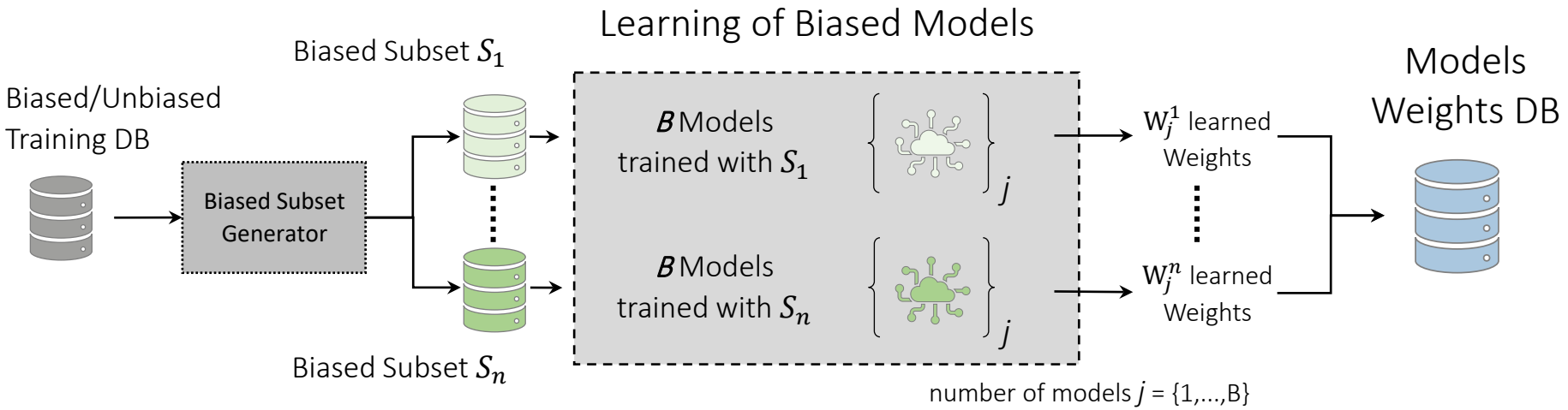
Traditionally: TEST SET DEPENDANT

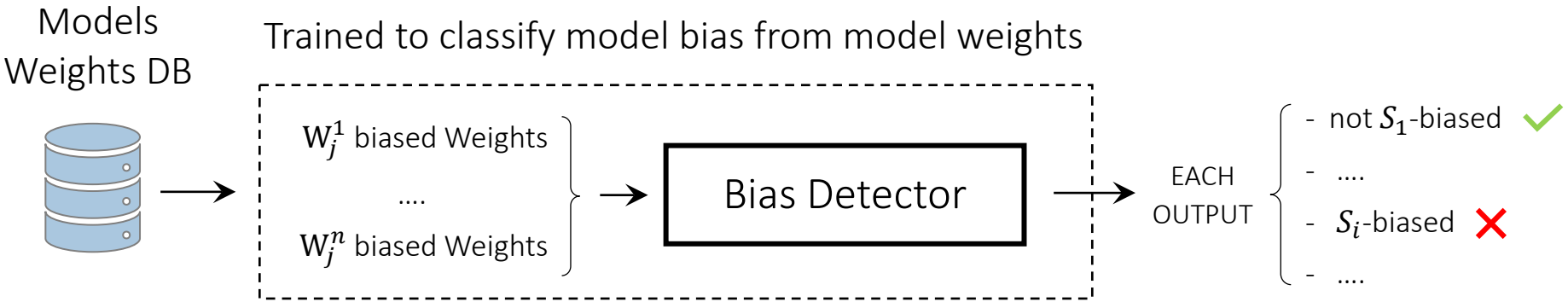


G. Balakrishnan et al., "Towards Causal Benchmarking of Biasin Face Analysis Algorithms", ECCV 2020

Proposed: IFBiD Inference-Free Bias Detection

Universidad Autónoma de Madrid



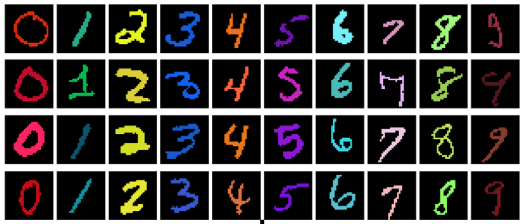


IFBiD: Inference-Free Bias Detection

DigitWdb and GenderWdb

Universidad Autónoma de Madrid

ColoredMNIST[1]



Train 48K models
of 50K params each
with 4 levels of bias:

- very high bias
- high bias
- low bias
- very low bias

DigitWdb

DiveFace[2]

Train 36K models
of 100K params each
with 3 classes of bias:

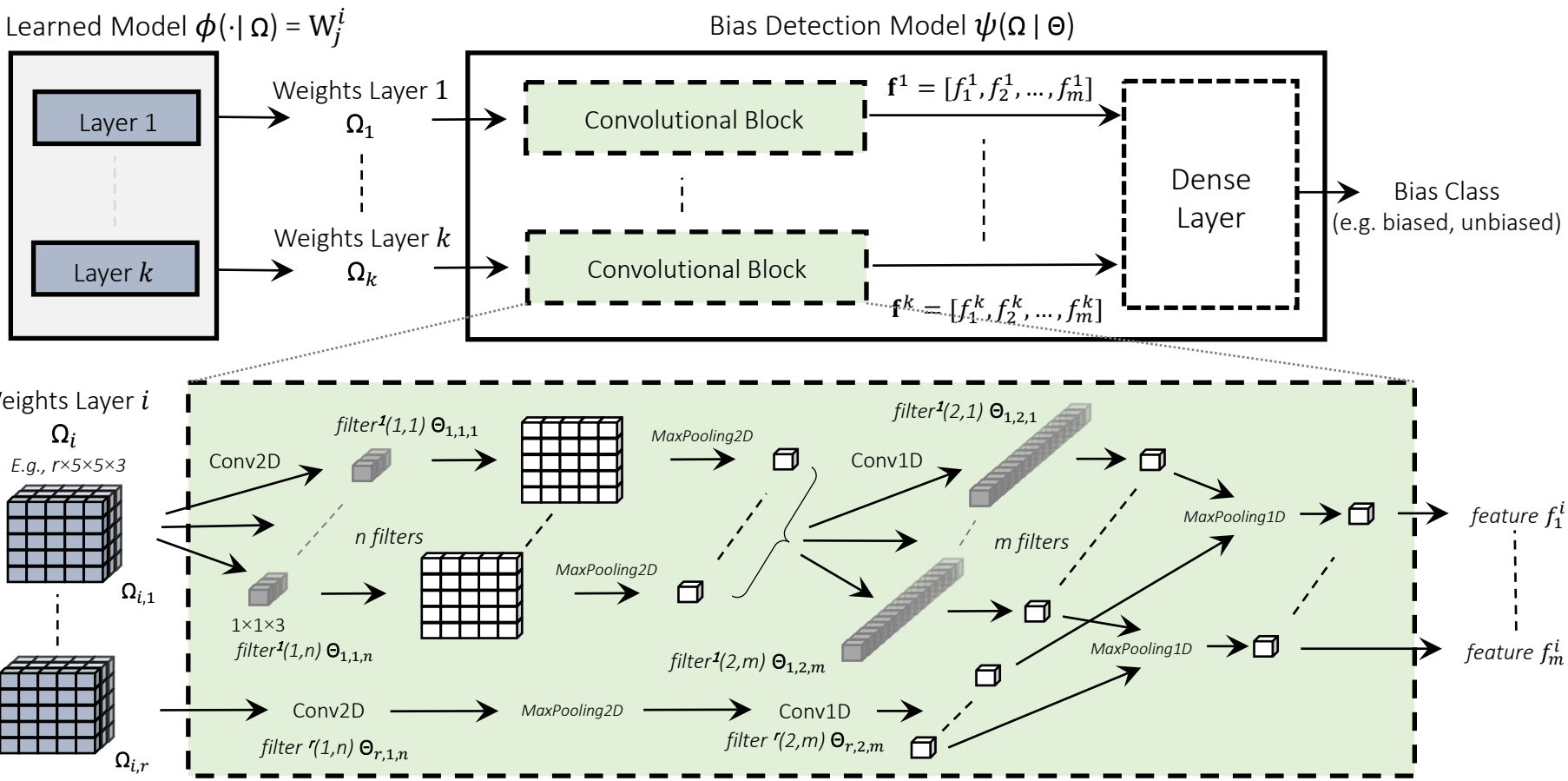
- Asian
- African/indian
- Caucasian

GenderWdb

1. B. Kim et al., "Learning Not to Learn: Training Deep Neural Networks With Biased Data," CVPR 2019
2. A. Morales et al., "SensitiveNets: Learning Agnostic Representations with Application to Face Images," IEEE T-PAMI, 2021

IFBiD: THE DETECTOR

Universidad Autónoma de Madrid



Results: Bias in Digit Classifier

Universidad Autónoma de Madrid

1st Experiment

- 20K training models:
  - 10K very high bias
  - 10K very low bias
- 4K test models:
  - 2k 2k

Classification *accuracy* obtained:

- Multi-layer perceptron: 96.5 %
- Convolutional Block: **99.7 %**

2nd Experiment

- 40K training models:
  - 10K very high bias
  - 10K high bias
  - 10K low bias
  - 10K very low bias
- 8k test models:
  - 2k 2k 2k 2k

Classification *accuracy* obtained:

- Multi-layer perceptron: 41.5 %
- Convolutional Block: **71.5 %**

Results: Bias in Gender Classifier

Universidad Autónoma de Madrid

3rd Experiment

- 30K training models:
  -  10K asian biased
  -  10K african/indian biased
  -  10K caucasian biased
- 6k test models:
  -  6k  6k  6k

Accuracy obtained:

- Multi-layer perceptron: 60.8 %
- Convolutional Block: **83.6 %**

SOTA Comparison

Method	Bias Detection Accuracy		
	Asian	African/Indian	Caucasian
RBF SVM	71%	30%	26%
InsideBias	23%	86%	3%
IFBiD (ours)	**95%**	**79%**	**79%**

*InsideBias requires no training, we used 60 images for the test.

Limitations and Future Work

This work poses two fundamental challenges:

- Finding a way to translate our approach (inference-free bias detection) to other problems, such as face recognition.
- Automatically detecting bias covariates.

BiDA Lab

Biometrics & Data Pattern Analytics Lab

UAM Universidad Autónoma de Madrid

http://biometrics.eps.uam.es

Know More:

I. Serna, A. Peña, A. Morales and J. Fierrez, **"InsideBias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics"**, in *IAPR Intl. Conf. on Pattern Recognition (ICPR)*, Milan, Italy, January **2021**.

I. Serna, A. Morales, J. Fierrez and N. Obradovich, **"SensitiveLoss: Improving Accuracy and Fairness of Face Representations with Discrimination-Aware Deep Learning"**, *Artificial Intelligence, 305*, April **2022**.

A. Peña, I. Serna, A. Morales and J. Fierrez, **"Bias in Multimodal AI: Testbed for Fair Automatic Recruitment"**, in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) workshop on Fair, Data-efficient, and Trusted Computer Vision,* June **2020**.

FUNDING: TRESPASSETN (MSCA-ITN-2019-860813), PRIMA (MSCAITN-2019-860315), BIBECA (RTI2018-101248-B-I00 MINECO/FEDER), and BBforTAI (PID2021-127641OBI00 MICINN/FEDER). I. Serna is supported by a FPI fellowship from UAM.

TReSPAsS-ETN

MARIE SKŁODOWSKA-CURIE ACTIONS

Privacy Matters

European Commission

GOBIERNO DE ESPAÑA