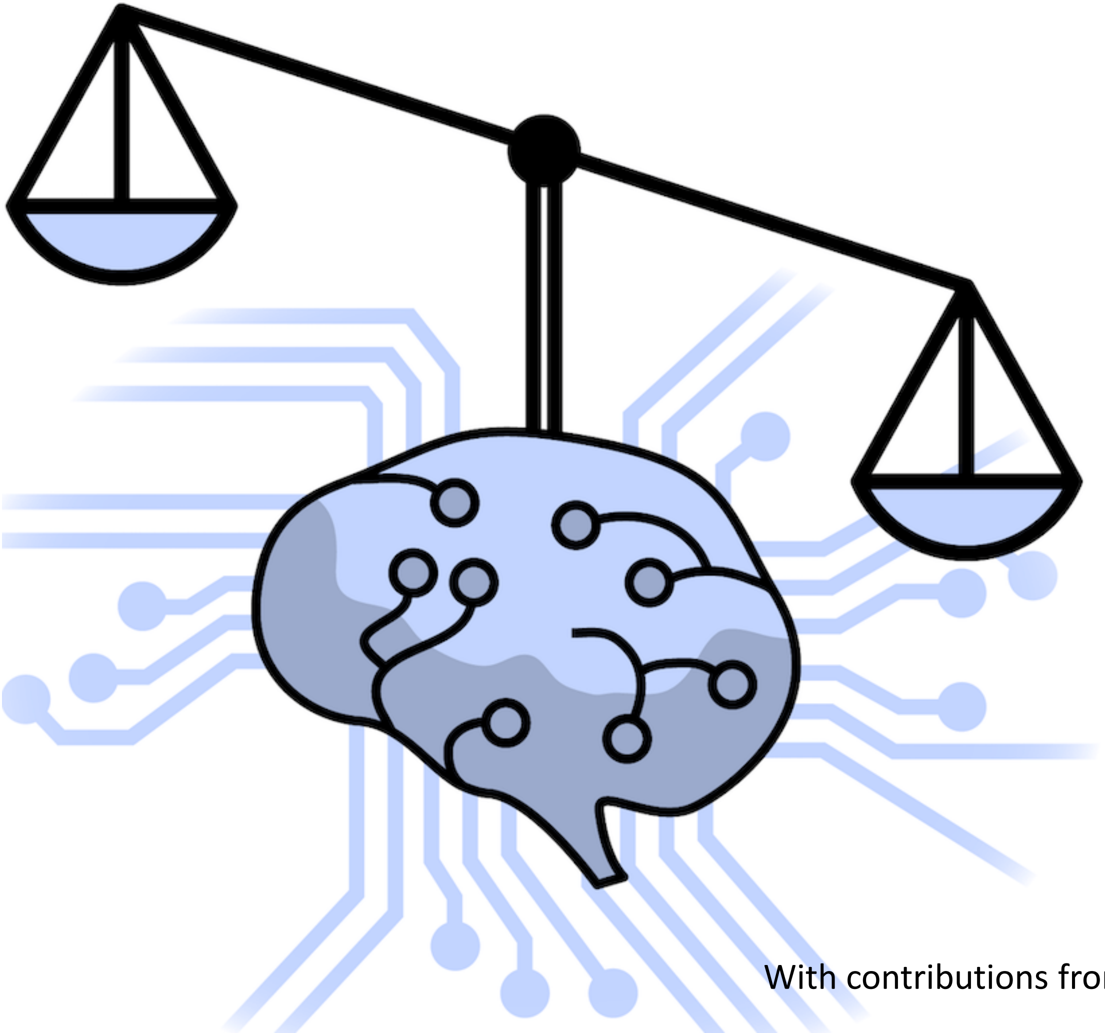
Responsible AI: Understanding Bias in ML



Aythami Morales

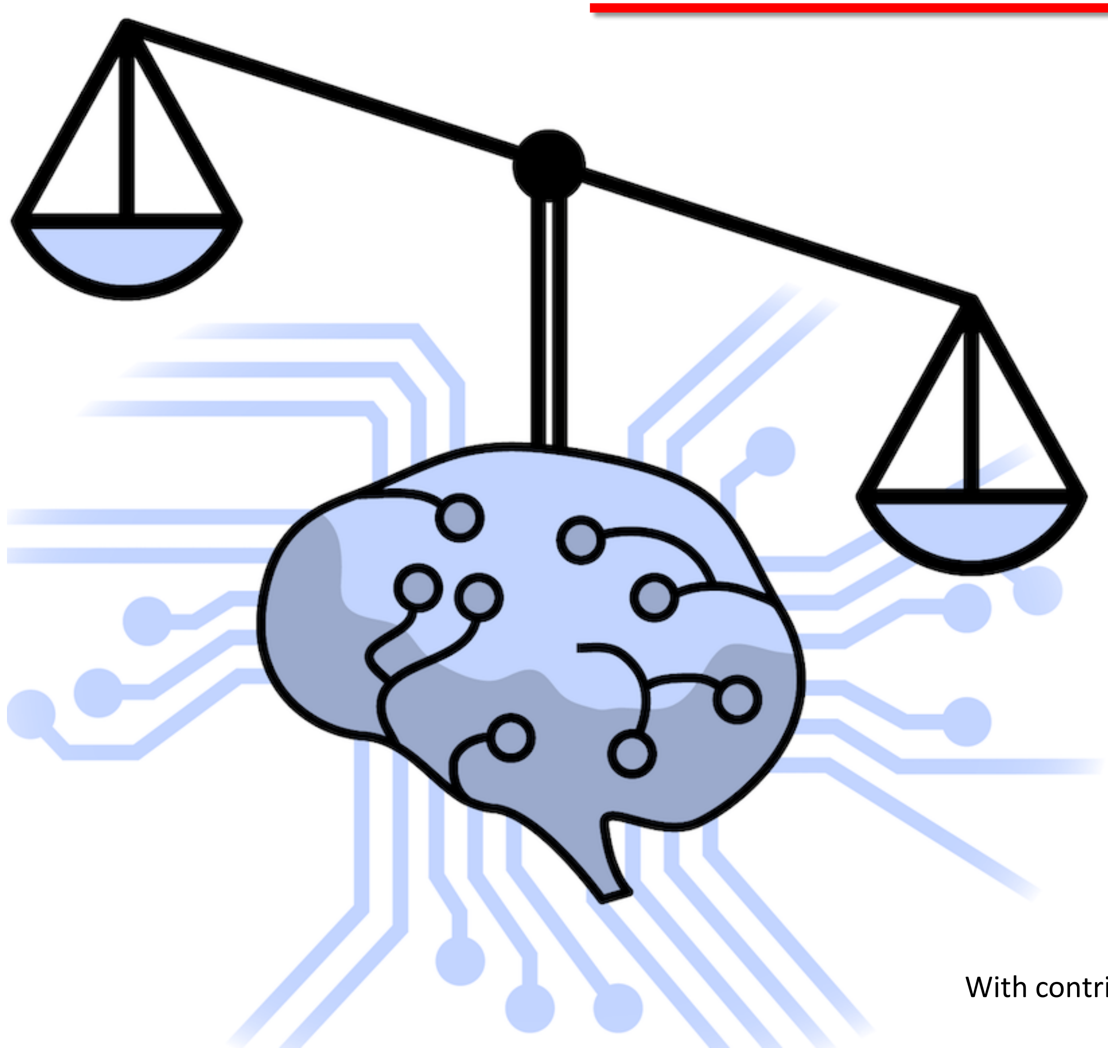
http://aythami.me

BiDA Lab
Biometrics & Data Pattern Analytics Lab

UAM
Universidad Autónoma de Madrid

With contributions from: Julian FIERREZ (http://biometrics.eps.uam.es/fierrez/), Alejandro PEÑA, Ignacio SERNA

Responsible AI: Understanding Bias in ML



Aythami Morales

http://aythami.me

BiDA Lab
Biometrics & Data Pattern Analytics Lab

UAM
Universidad Autónoma de Madrid

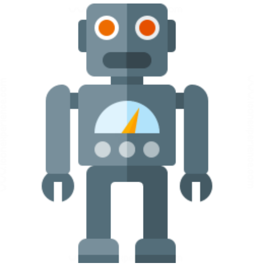
With contributions from: Julian FIERREZ (http://biometrics.eps.uam.es/fierrez/), Alejandro PEÑA, Ignacio SERNA

The Standard Model

S.J. Russell, P. Norvig, "Artificial intelligence: a modern approach", Prentice hall, 2010.

From Classical Programming to Machine Learning

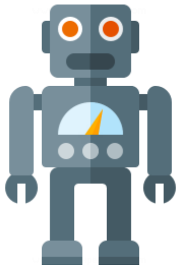
Experience → **Learning** → **Responses**



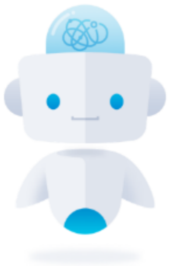
Rules →
Data → **Classic Programming** → **Responses**

From Classical Programming to Machine Learning

Experience → **Learning** → Responses



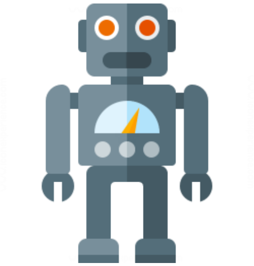
Rules →
Data → **Classic Programming** → Responses



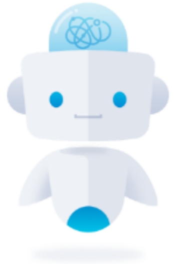
Data →
Responses → **Machine Learning** → Rules → New Responses

From Classical Programming to Machine Learning

Experience → **Learning** → Responses



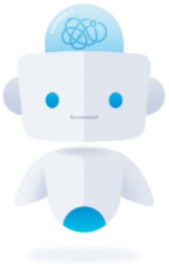
Rules →
Data → **Classic Programming** → Responses



Data →
Responses → **Machine Learning** → Rules → What factors affect decision-making?

From Human Behavior to Machine Behavior

Human Behavior



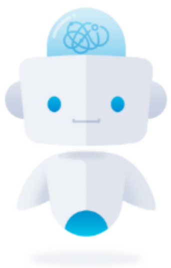
Machine Behavior

From Human Behavior to Machine Behavior

Human Objectives

- ➢ Learning process is guided by pre-defined objectives.
- ➢ These objectives use to be a simplification of human thinking/tasks.

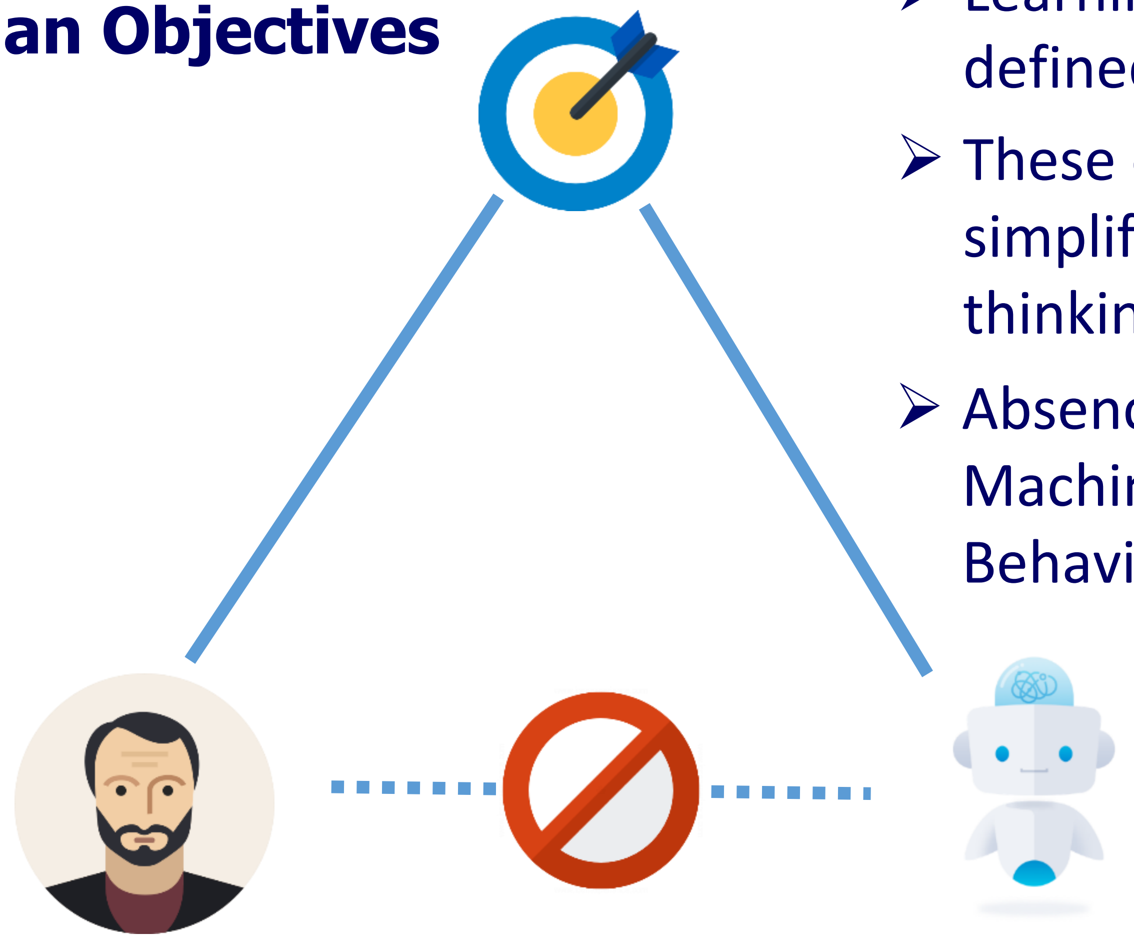
Human Behavior



Machine Behavior

From Human Behavior to Machine Behavior

Human Objectives



Human Behavior

Machine Behavior

- Learning process is guided by pre-defined objectives.
- These objectives use to be a simplification of human thinking/tasks.
- Absence of direct path between Machine Behavior and Human Behavior

What is Algorithmic Discrimination?

- **Constitución Española:**
  - Artículo 14: los españoles son iguales ante la ley, sin que pueda prevalecer **discriminación** alguna por razón de nacimiento, raza, sexo, religión, opinión o cualquier otra condición o circunstancia personal o social.
- **Universal Declaration of Human Rights**
  - All are entitled to equal protection against any **discrimination** in violation of this Declaration and against any incitement to such discrimination.
- **General Data Protection Regulation (GDPR)**
  - According to paragraph 71 of GDPR, data controllers who process sensible data have to "*implement appropriate technical and organizational measures…*" that "*…prevent, inter alia, **discriminatory** effects*".

What is Algorithmic Discrimination?

- **Constitución Española:**
  - Artículo 14: los españoles son iguales ante la ley, sin que pueda prevalecer **discriminación** alguna por razón de nacimiento, raza, sexo, religión, opinión o cualquier otra condición o circunstancia personal o social.
- **Universal Declaration of Human Rights**
  - All are entitled to equal protection against any **discrimination** in violation of this Declaration and against any incitement to such discrimination.
- **General Data Protection Regulation (GDPR)**
  - According to paragraph 71 of GDPR, data controllers who process sensible data have to "*implement appropriate technical and organizational measures…*" that "*…prevent, inter alia,* ***discriminatory*** *effects*".

THE RIGHT TO NON DISCRIMIANTION IS A FUNDAMENTAL RIGHT

Structured and Unstrutered Sensitive Information

Audio	Image	Text
• ID • Language • Accent • Age • Gender • Context • …	• Context • ID • Age • Ethnicity • Gender • …	• Language • Script • Context • Social & Cultural • Ideas • …

Structured and Unstrutered Sensitive Information

Audio

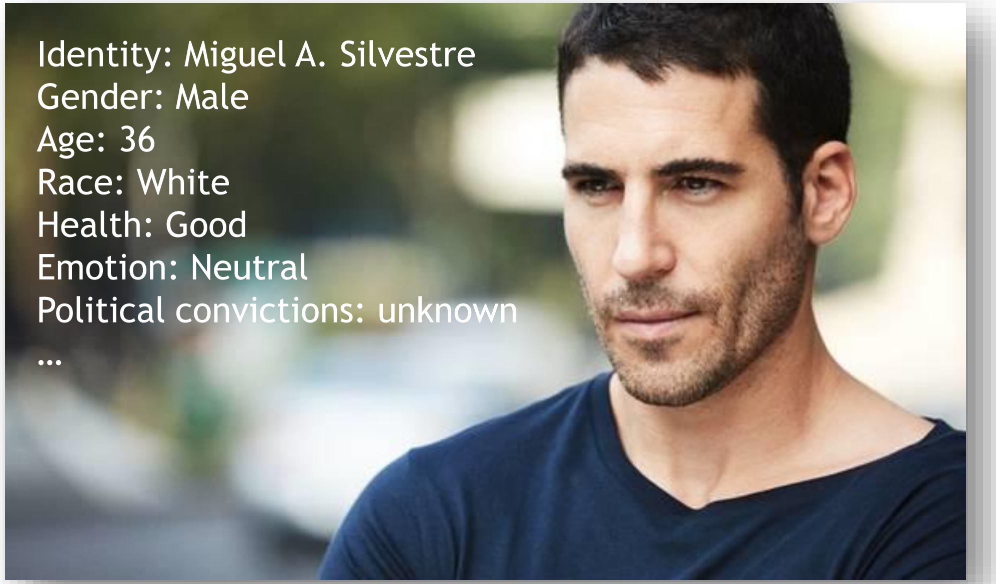
- ID
- Language
- Accent
- Age
- Gender
- Context
- …

Image

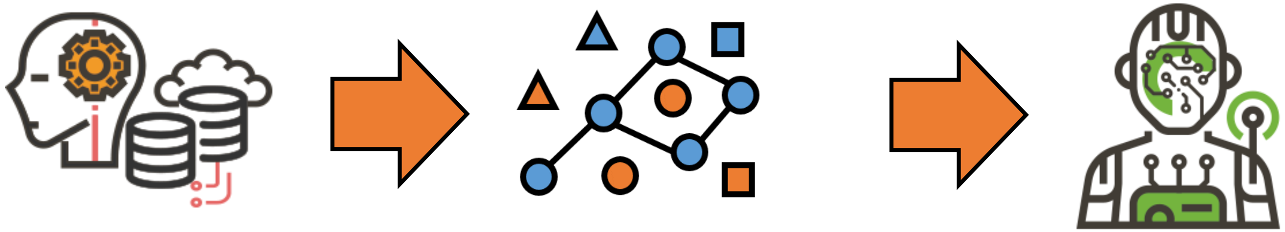
- Context
- ID
- Age
- Ethnicity
- Gender
- …

Text

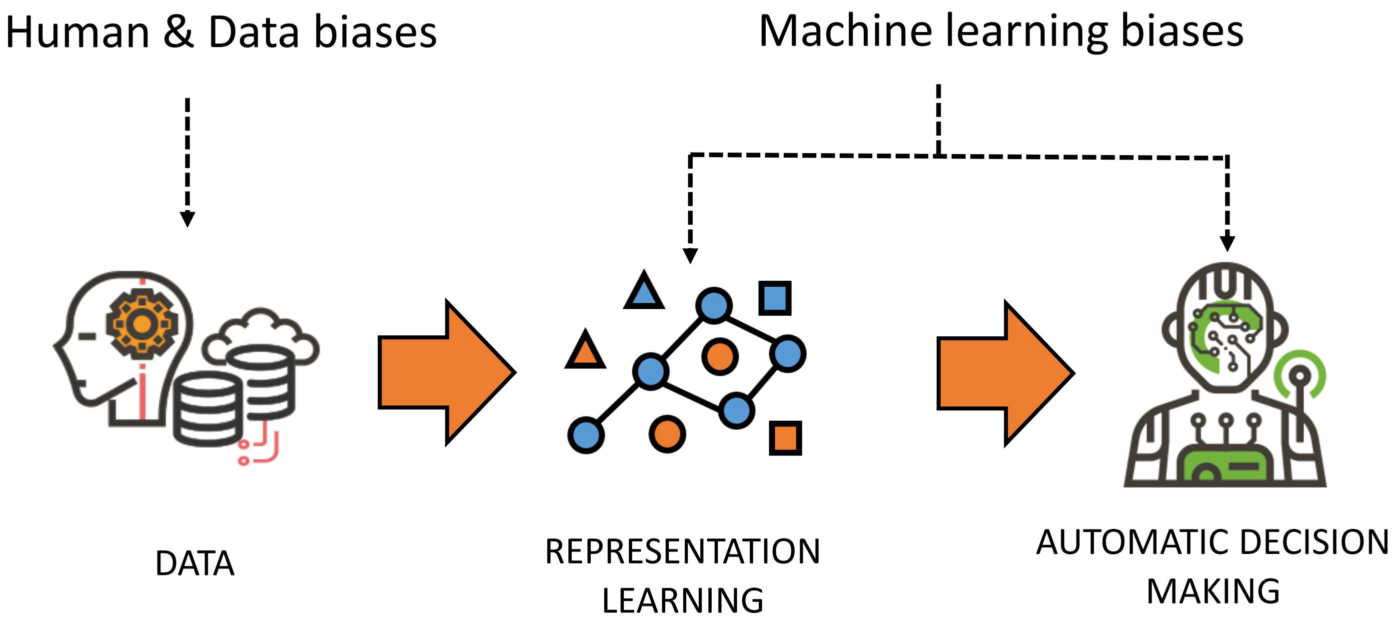
- Language
- Script
- Context
- Social & Cultural
- Ideas
- …



How Algorithmic Discrimination appears?

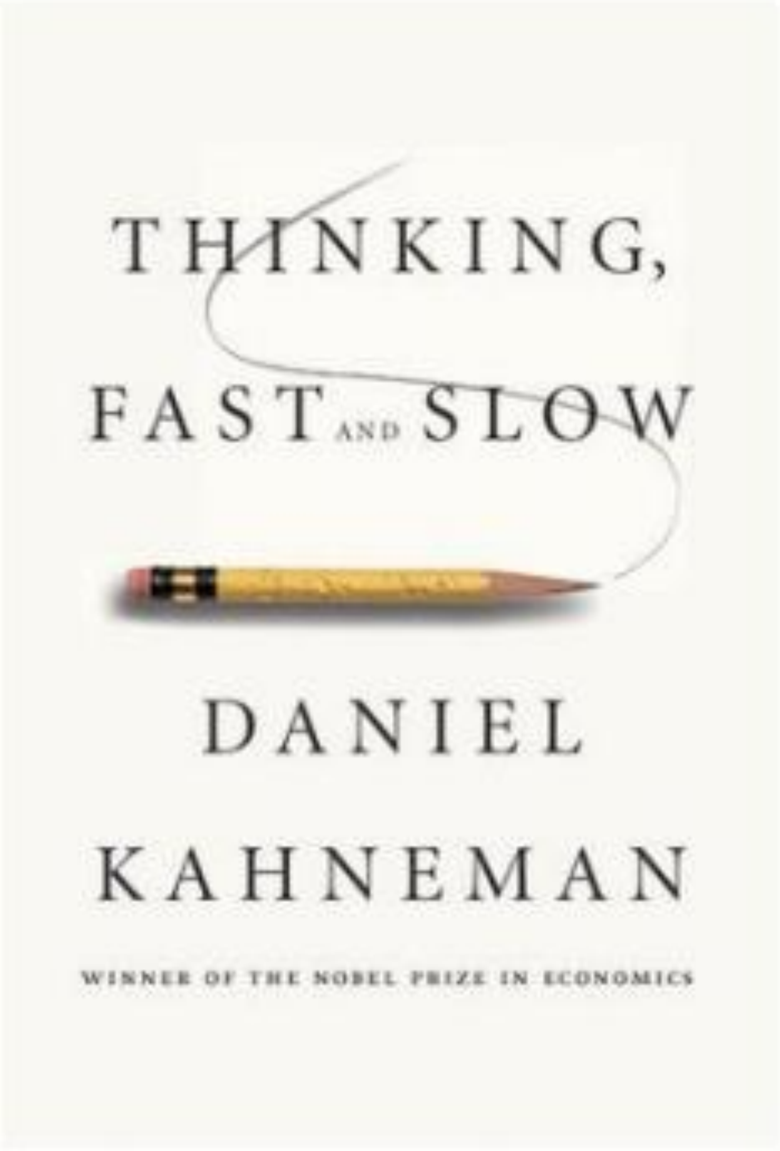


How Algorithmic Discrimination appears?



Limitations of the Standard Model

Based on AAAI 20 / AAAI 2020 Keynotes Turing Award Winners Event / Geoff Hinton, Yann Le Cunn, Yoshua Bengio

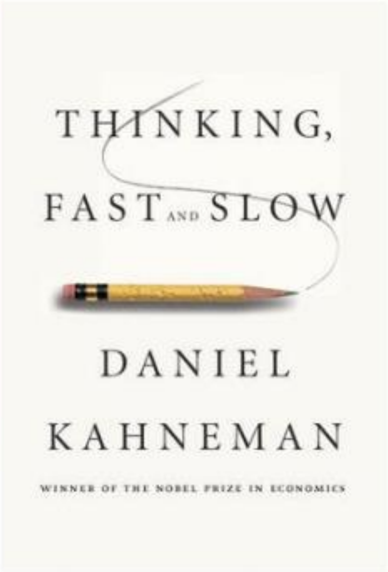

THINKING,
FAST AND SLOW
DANIEL
KAHNEMAN
WINNER OF THE NOBEL PRIZE IN ECONOMICS

Human Cognition: System 1 vs. System 2

2 systems (and categories of cognitive tasks):

System 1

- Intuitive, fast, **UNCONSCIOUS**, non-linguistic, habitual
- Current DL



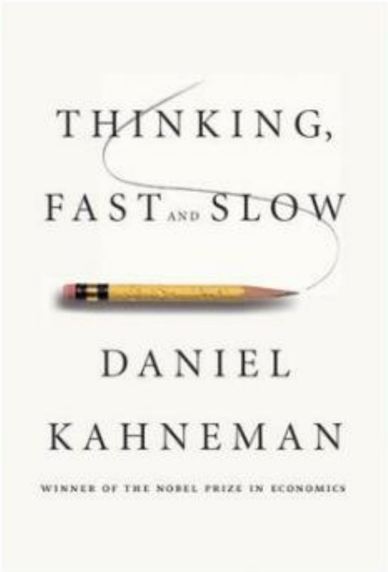
Human Cognition: System 1 vs. System 2

2 systems (and categories of cognitive tasks):

Manipulates high-level / semantic concepts, which can be recombined combinatorially

System 1

- Intuitive, fast, **UNCONSCIOUS**, non-linguistic, habitual
- Current DL

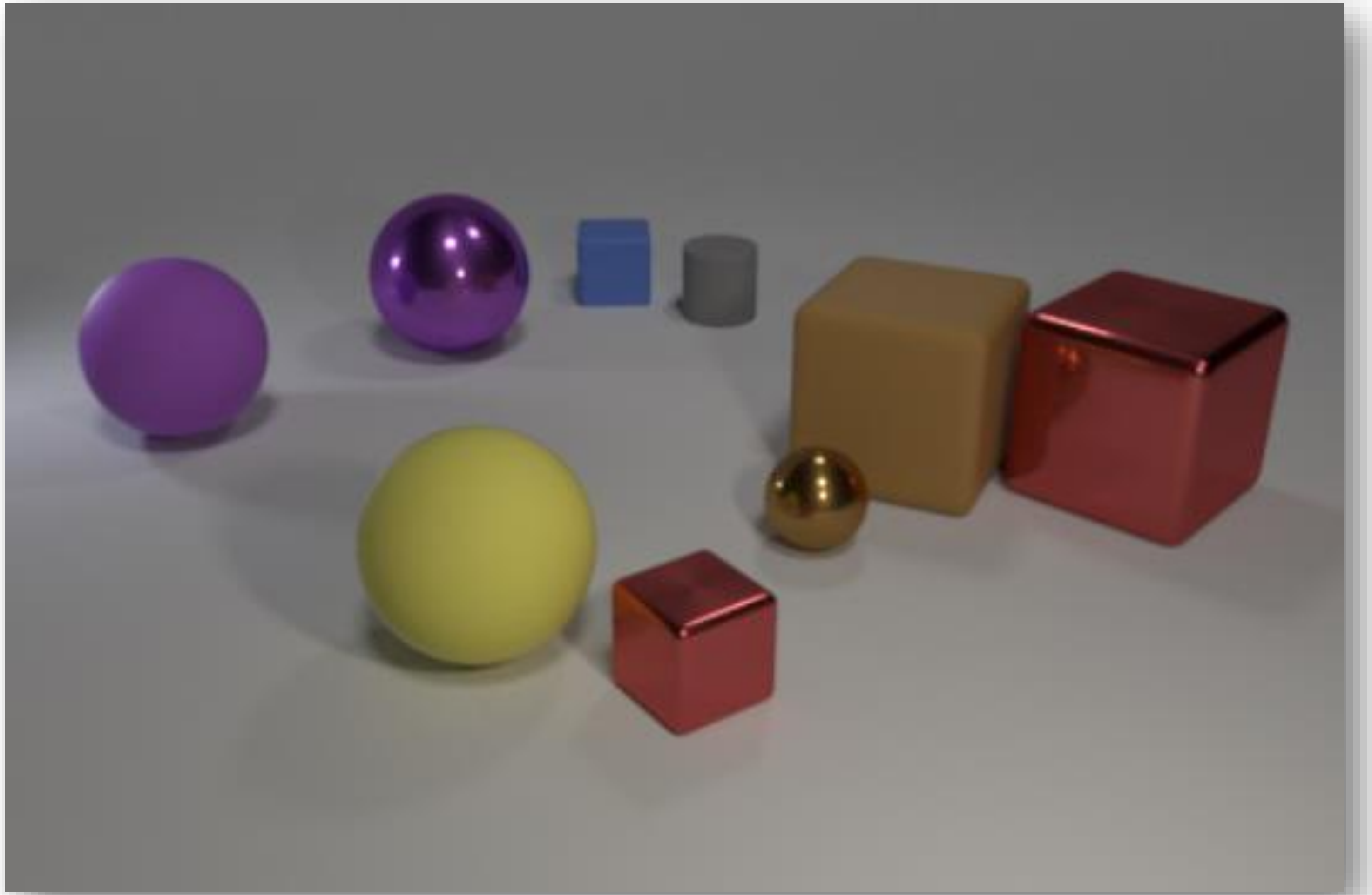


System 2

- Slow, logical, sequential, **CONSCIOUS**, linguistic, algorithmic, planning, reasoning
- Future DL

Mila

Human Cognition: System 1 vs. System 2



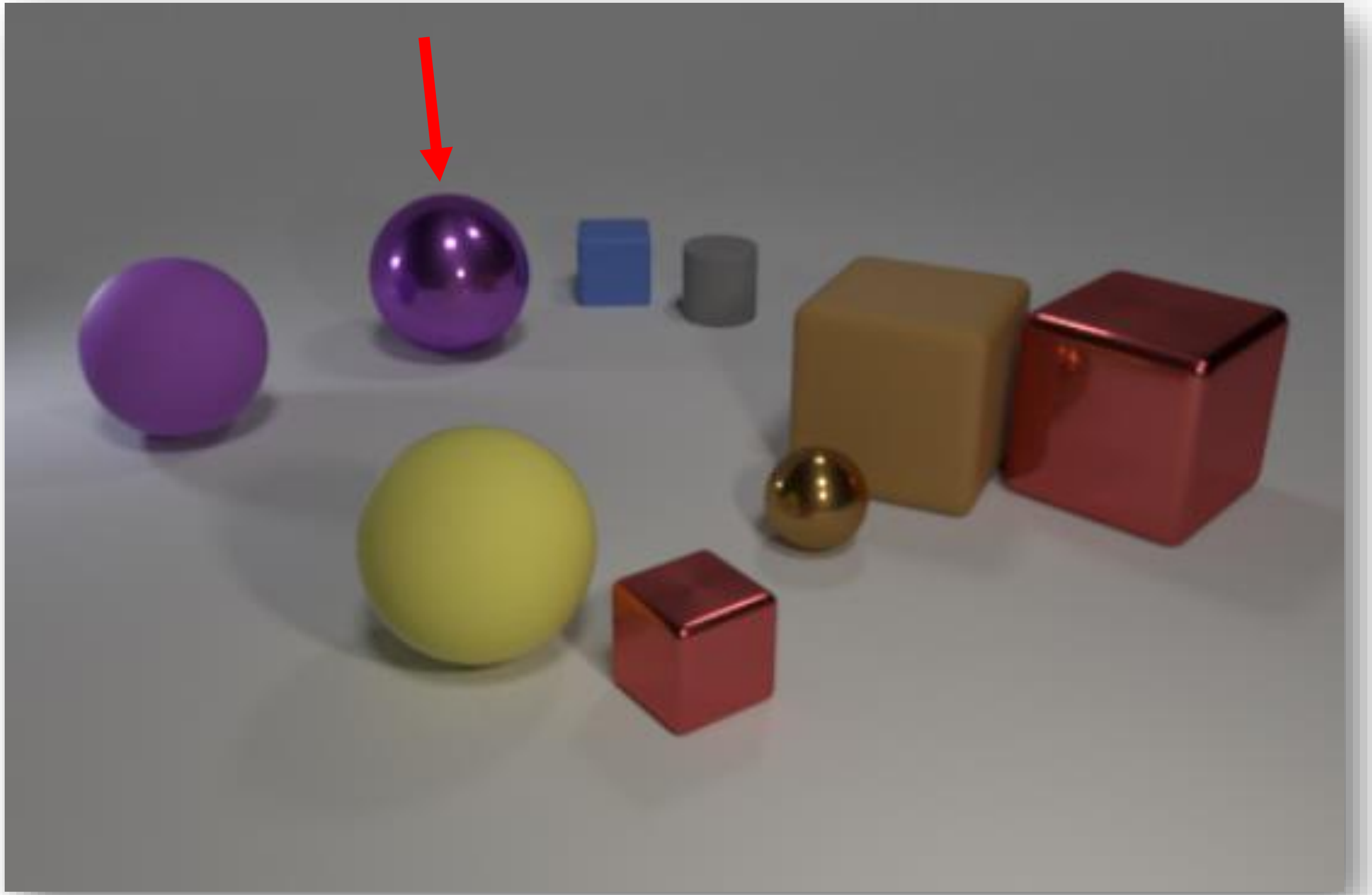
System 1:

- Shape detection, color detection, positioning.

System 2:

- Interaction inference

Human Cognition: System 1 vs. System 2



System 1:

- Shape detection, color detection, positioning.

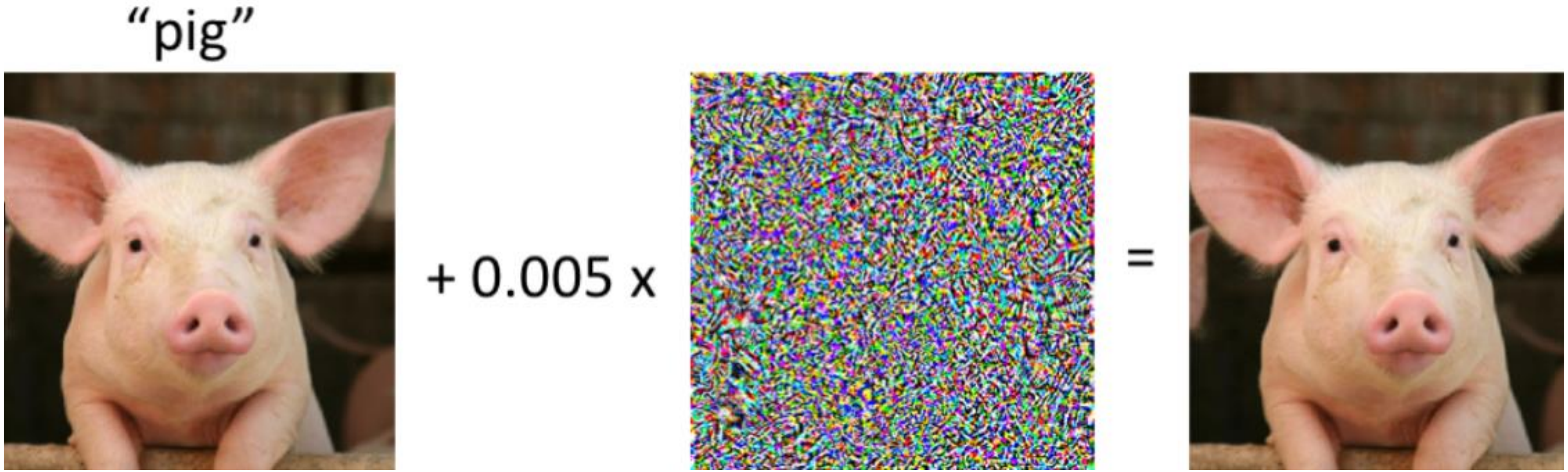
System 2:

- Interaction inference

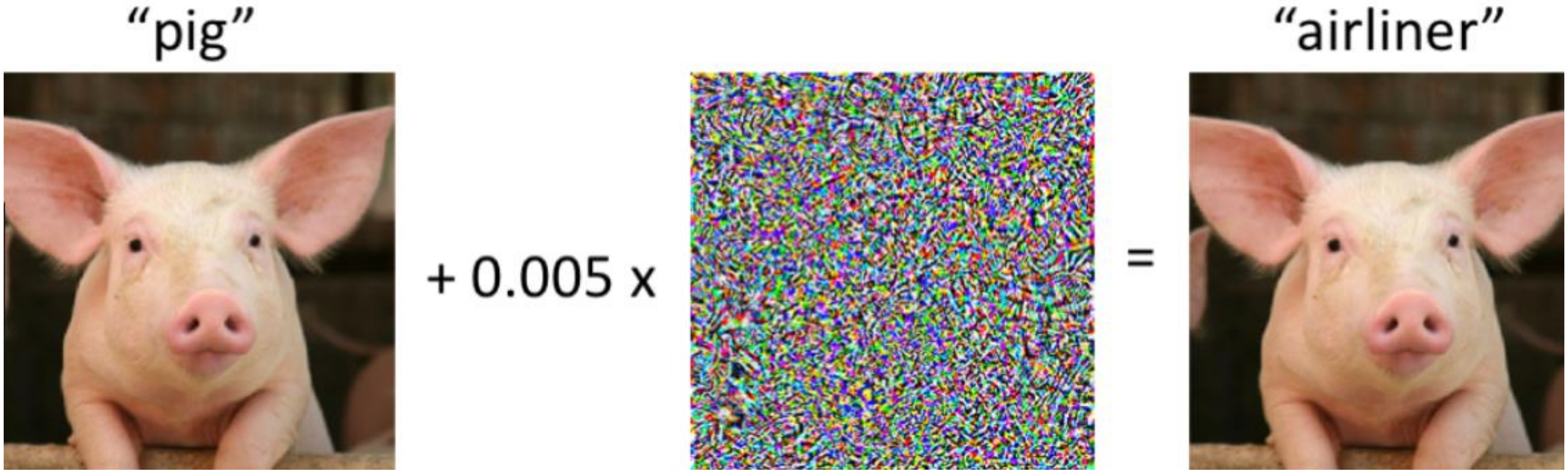
How AI See: An Example with Adversarial Attacks

“pig”

How AI See: An Example with Adversarial Attacks

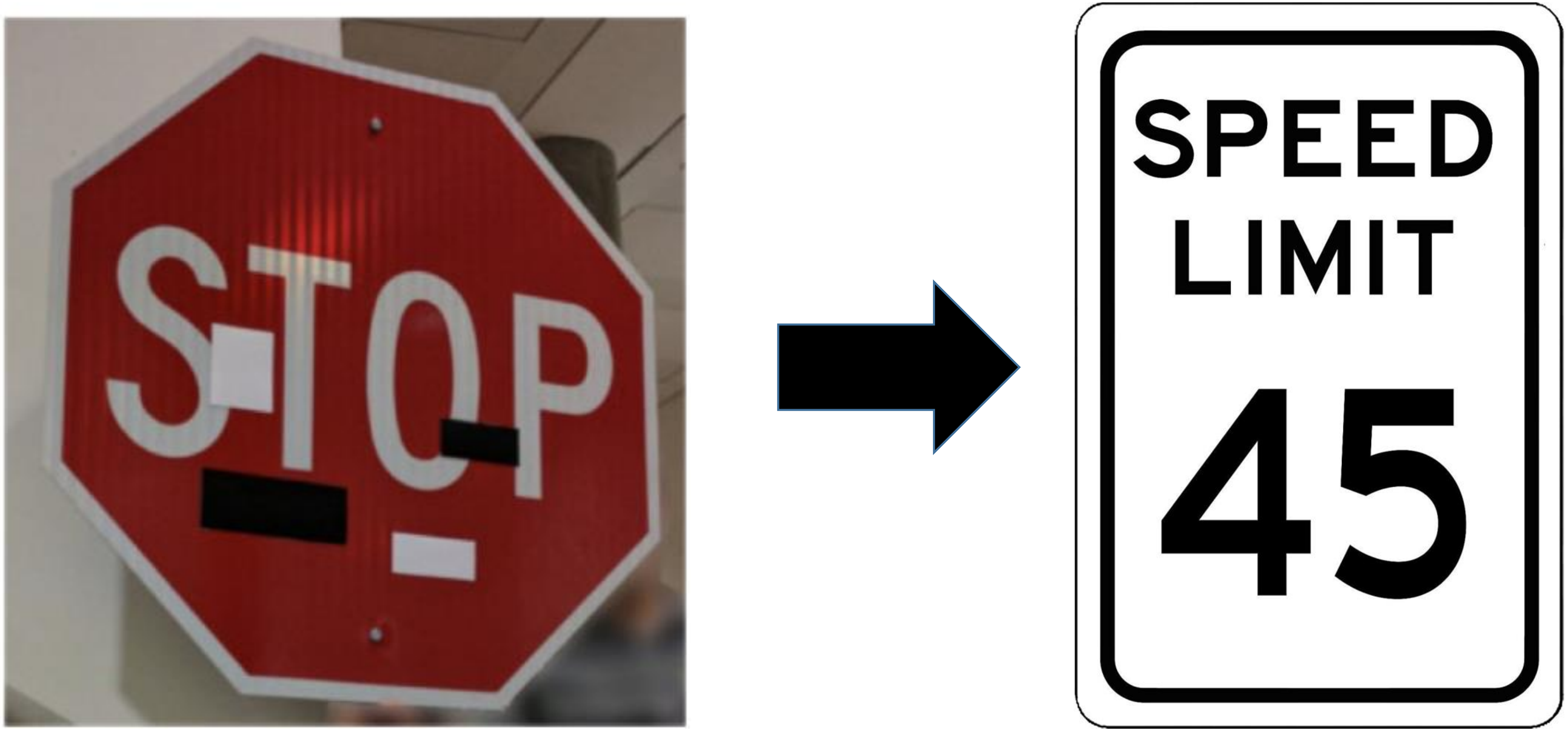


How AI See: An Example with Adversarial Attacks



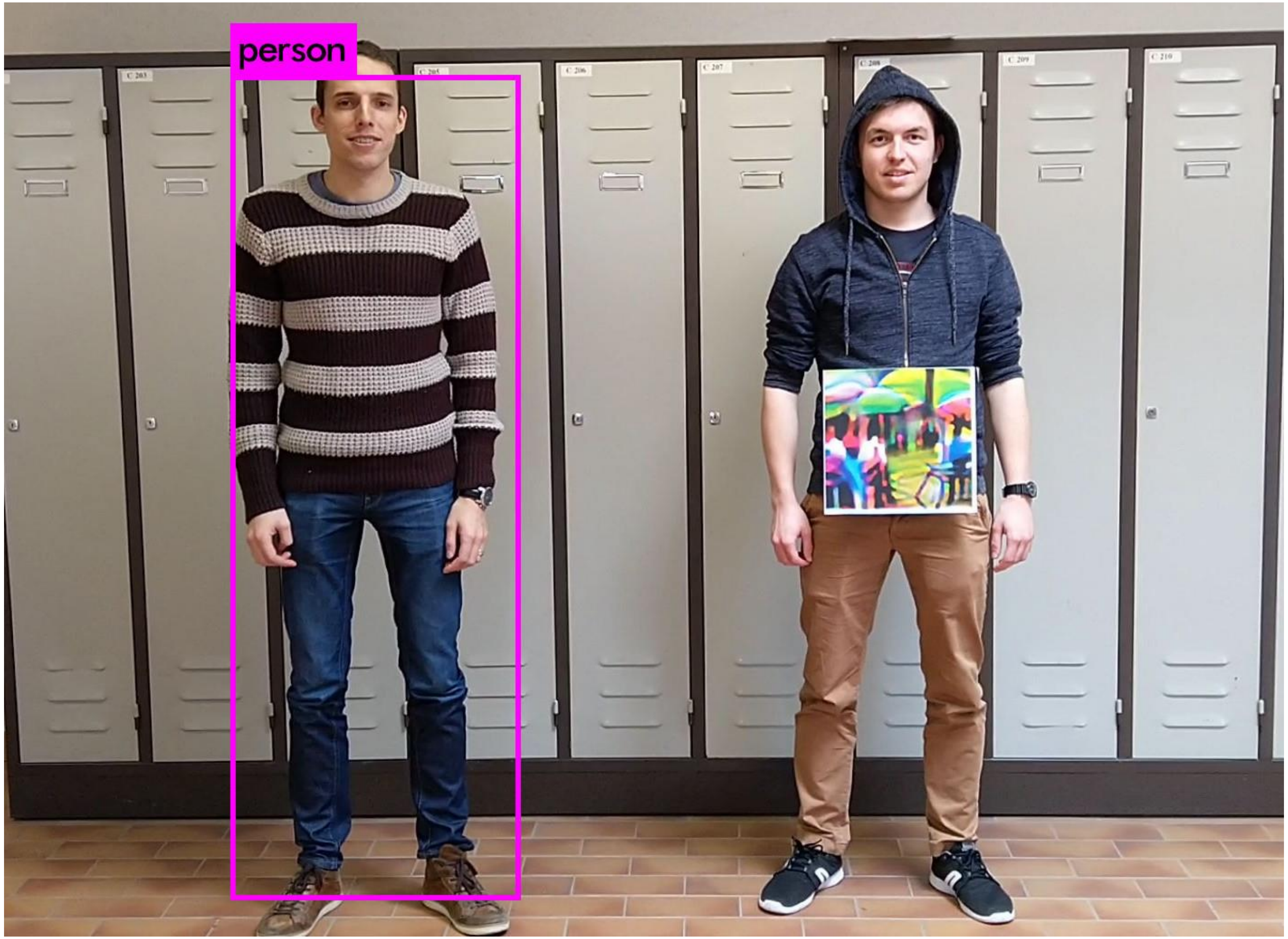
<u>**Remember Artificial Intelligence is not Human Intelligence**</u>

How AI See: An Example with Adversarial Attacks



D. Heaven, Why deep-learning AIs are so easy to fool, Nature, 2019

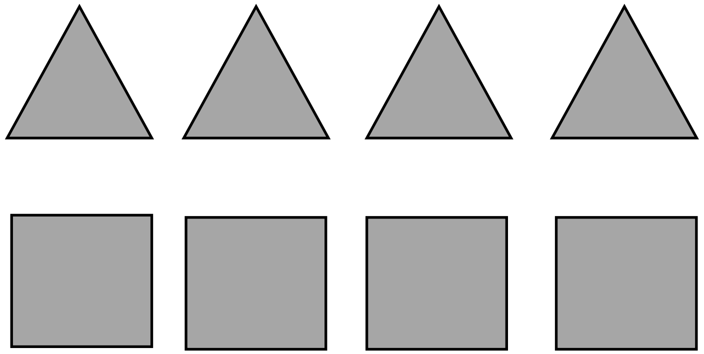
How AI See: An Example with Adversarial Attacks



Measuring the Bias

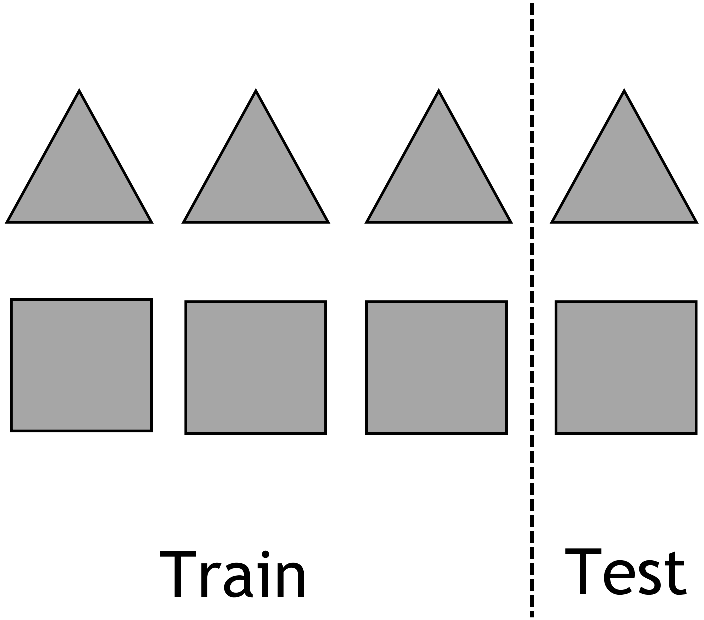
Problem: Shape recognition
triangle or square?

Data



Problem: Shape recognition
triangle or square?

Data

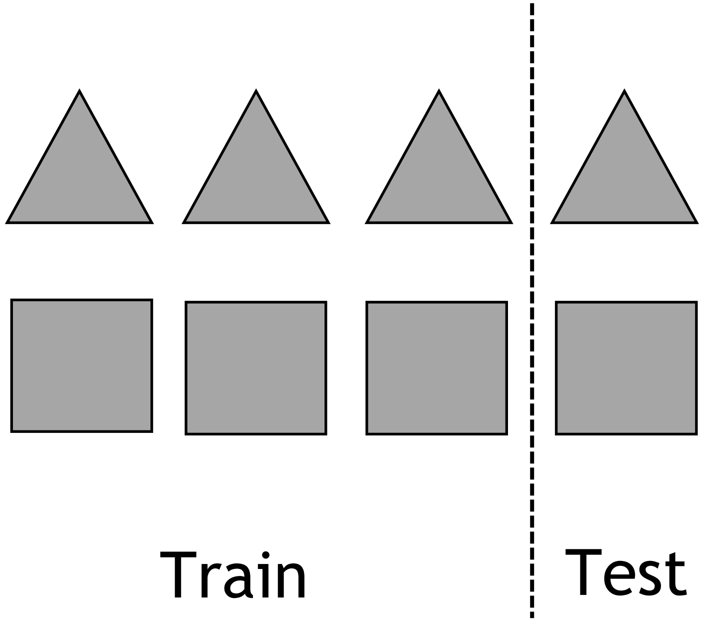


Accuracy: 90%

	△	□
△	95	5
□	15	85

Problem: Shape recognition triangle or square?

Data



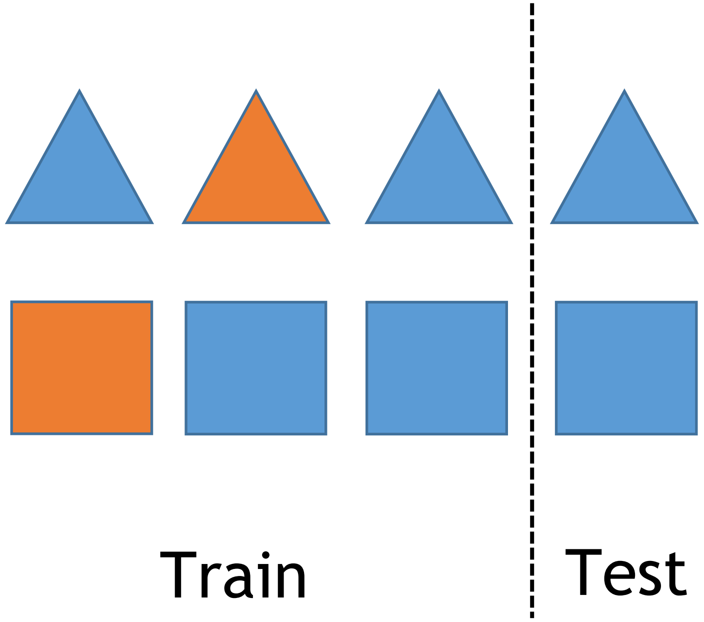
Accuracy: 90%

	△	□
△	95	5
□	15	85

Are all triangles and squares the same?
Assumption of homogeneous population

Problem: Shape recognition triangle or square?

Data



Accuracy: 90%

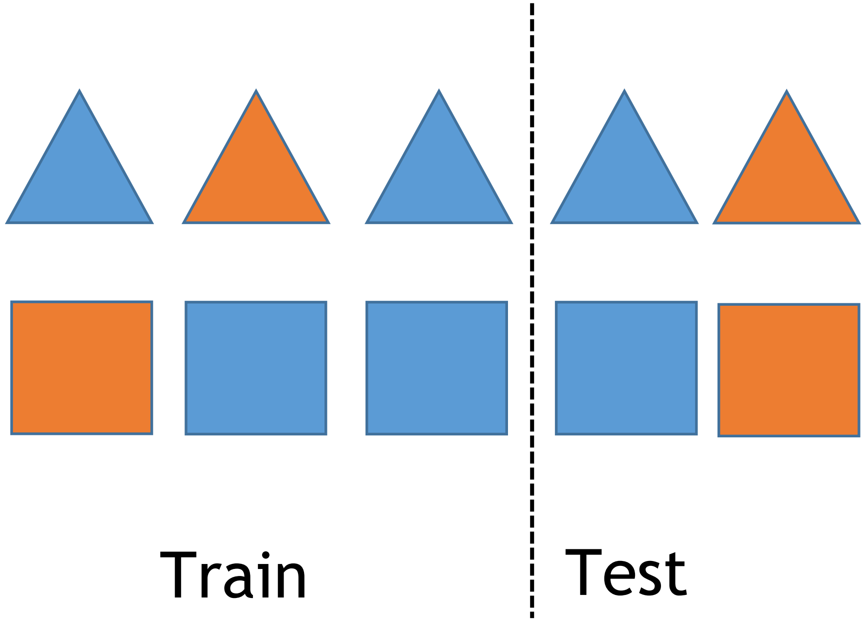
	▲	■
▲	95	5
■	15	85

Biased databases imply a **double penalty** for underrepresented classes:

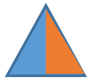
- Models are **trained** according to non-representative diversity.
- Models are **tested** on privileged classes

Problem: Shape recognition
triangle or square?

Data



Accuracy: 85%

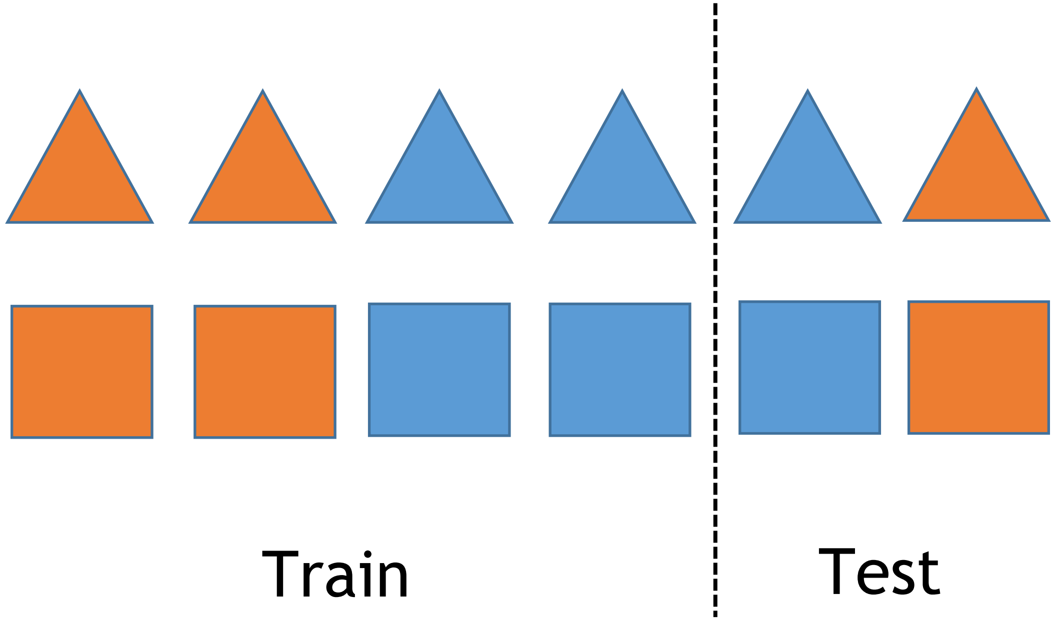
		
	90	10
	20	80

Color does not affect the shape…

Therefore, performance should be the same

Problem: Shape recognition triangle or square?

Data



Heterogeneous populations might produce heterogeneous performances

Blue Accuracy: 90%

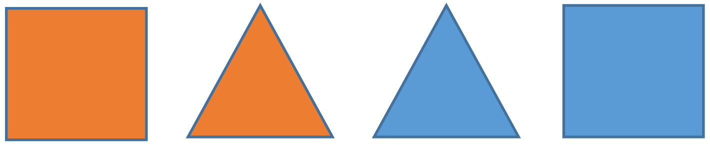
	▲	■
▲	95	5
■	15	85

Orange Accuracy: **80%**

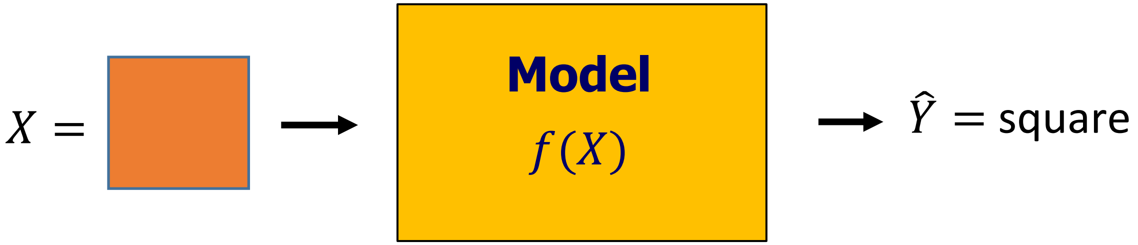
	▲	■
▲	85	15
■	25	75

How to Measure Fairness

We consider supervised deep learning tasks in which the task is to predict an output variable Y given an input variable X, while remaining unbiased with respect to some variable Z. We refer to Z as the protected/sensitive variable.



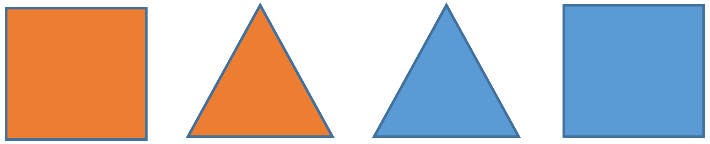
X= image
Y= shape (triangle or square)
Z = color (orange or blue)



B. Zhang, B. Lemoine, M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society, pp. 335–340, 2018.

How to Measure Fairness

We consider supervised deep learning tasks in which the task is to predict an output variable Y given an input variable X, while remaining unbiased with respect to some variable Z. We refer to Z as the protected/sensitive variable.



X= image
Y= shape (triangle or square)
Z = color (orange or blue)

Equality of Opportunity: a predictor $\hat{Y}$ satisfies equality of opportunity with respect to a class y if $\hat{Y}$ and Z are independent conditioned on $Y = y$.

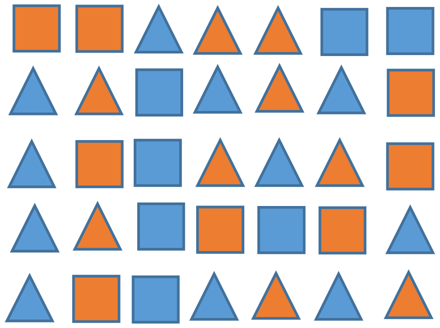
Same performance for orange and blue objects

$$P\left(\hat{Y} = \hat{y} \middle| Y = y\right) = P\left(\hat{Y} = \hat{y} \middle| Z = z, Y = y\right)$$

B. Zhang, B. Lemoine, M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society, pp. 335–340, 2018.

Representation Level is the Key

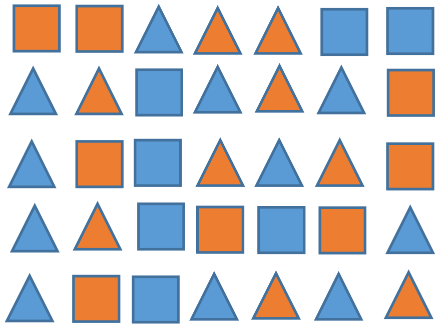
Task: Shape recognition, triangle or square?



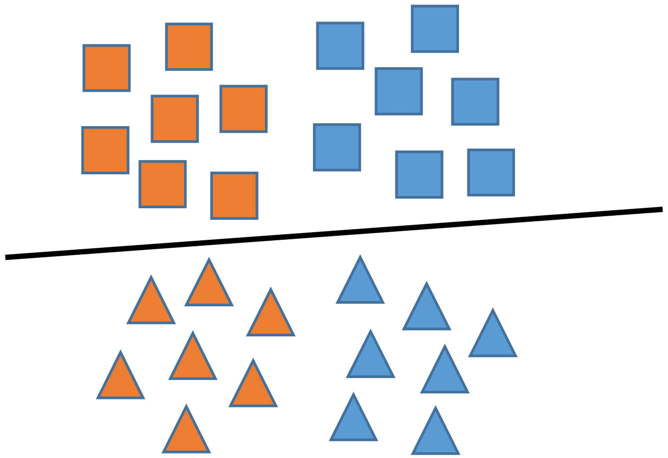
Input Space

Representation Level is the Key

Task: Shape recognition, triangle or square?



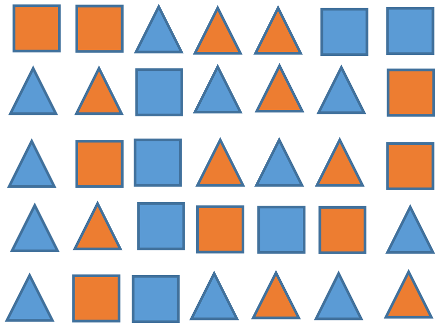
Input Space



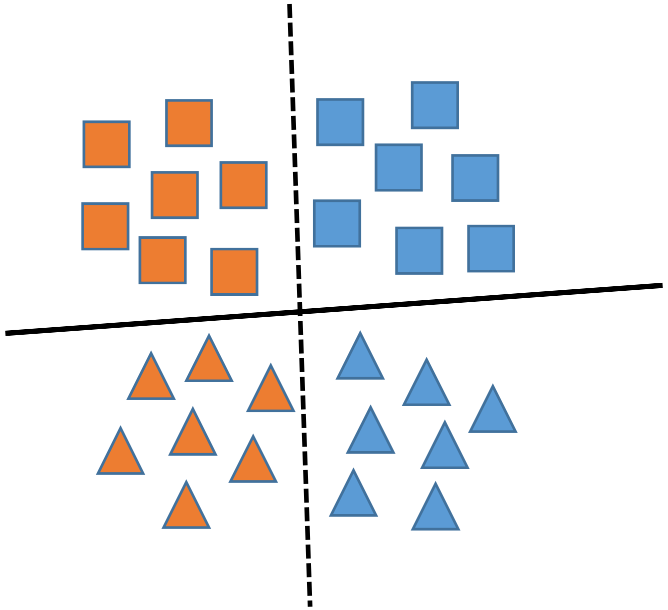
Learned
Feature Space

Representation Level is the Key

Task: Shape recognition, triangle or square?



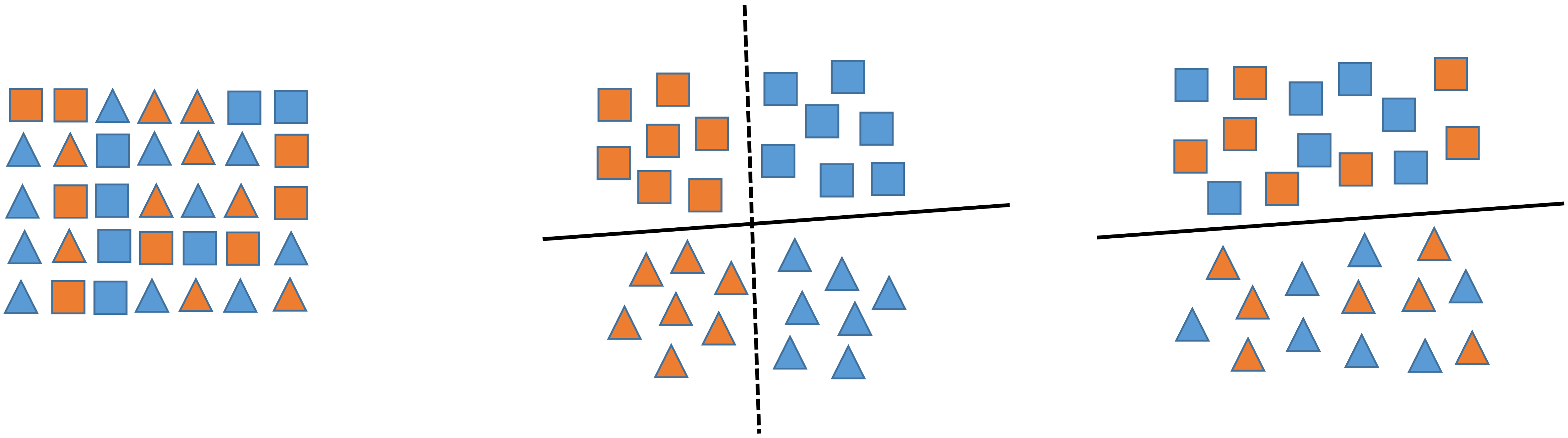
Input Space



Learned Feature Space

Representation Level is the Key

Task: Shape recognition, triangle or square?



Input Space

Learned Feature Space

Agnostic Learned Feature Space

Discrimination-aware Learning Frameworks: a Simple Example

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

From Standard AIs to Responsible AIs

$$\min_{\mathbf{w}} \sum_{\mathbf{x}^j \in S} \mathcal{L}\left(O\left(\mathbf{x}^j \middle| \mathbf{w}\right), T^j\right)$$

From Standard AIs to Responsible AIs

Output of the Model

Model

Loss Function

$$\min_{\mathbf{w}} \sum_{\mathbf{x}^j \in S} \mathcal{L}\left(O\left(\mathbf{x}^j \middle| \mathbf{w}\right), T^j\right)$$

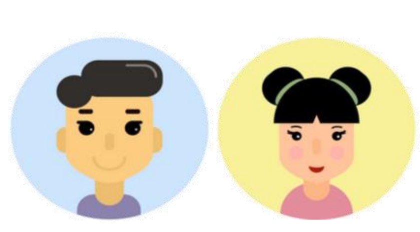
Target Function

Training Database

Data from one subject
(e.g. Resumes)

Is not only justice, it is also performance

CASE STUDY: Face Recognition Performance

	Caucassian	Asian	Black	Avg.	Std
					
ResNet50	3.36%	5.52%	5.62%	4.95%	1.03

$$\min_{\mathbf{w}} \left(\sum_{\mathbf{x}^j \in S} \mathcal{L}\left(O\left(\mathbf{x}^j \middle| \mathbf{w}\right), T^j\right) \right)$$

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

Is not only justice, it is also performance

CASE STUDY: Face Recognition Performance

	Caucassian	Asian	Black	Avg.	Std
					
ResNet50	3.36%	5.52%	5.62%	4.95%	1.03

$$\min_{\mathbf{w}} \left(\sum_{\mathbf{x}^j \in S^1} \mathcal{L}\left(O\left(\mathbf{x}^j \middle| \mathbf{w}\right), T^j\right) + \sum_{\mathbf{x}^j \in S^2} \mathcal{L}\left(O\left(\mathbf{x}^j \middle| \mathbf{w}\right), T^j\right) + \sum_{\mathbf{x}^j \in S^3} \mathcal{L}\left(O\left(\mathbf{x}^j \middle| \mathbf{w}\right), T^j\right) \right)$$

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

Is not only justice, it is also performance

CASE STUDY: Face Recognition Performance

	Caucassian	Asian	Black	Avg.	Std
					
ResNet50	3.36%	5.52%	5.62%	4.95%	1.03
ResNet50-RAI	2.72%	3.78%	3.66%	3.34% (↓**30%**)	0.42% (↓**54%**)

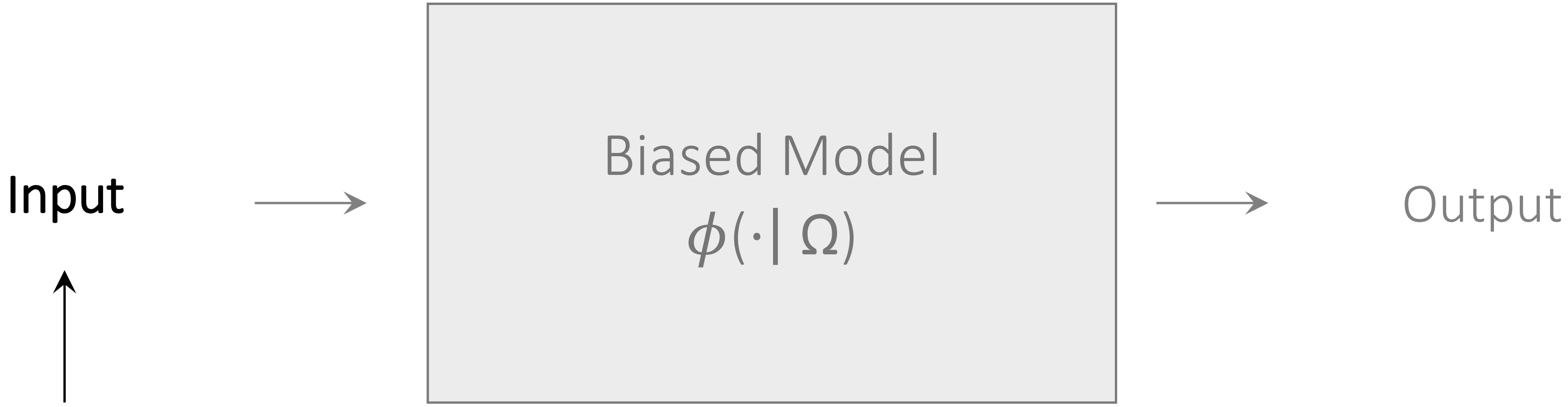
Responsible AI might improves your models

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

Understanding Bias in Data-driven Learning Approaches

Material from Ignacio de la Serna

Analysis of Biased Performance

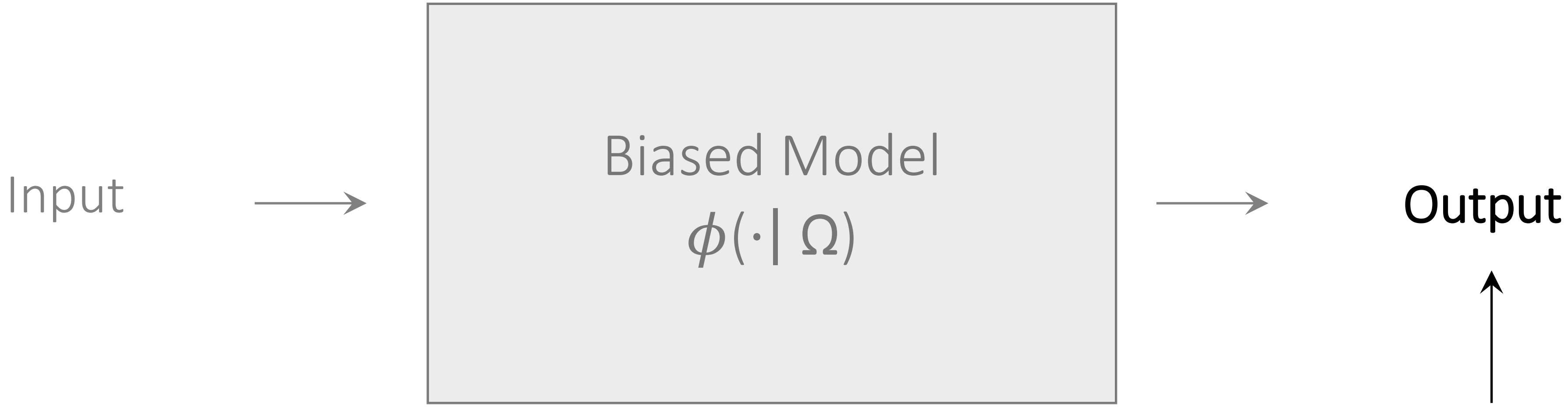


Face Databases

Dataset [ref]	# images	# identities	Caucasian		African/Indian		Asian	
			Male	Female	Male	Female	Male	Female
FRVT2018 [28]	27M	12M	48.4%	16.5%	19.9%	7.4%	1.2%	0.4%
MSCeleb1M [29]	8.5M	100K	52.4%	19.2%	12.1%	3.9%	7.7%	4.5%
MegaFace [30]	4.7M	660K	40.0%	30.3%	6.2%	4.7%	10.6%	8.1%
VGGFace2 [31]	3.3M	9K	45.9%	30.2%	10.5%	6.3%	3.4%	3.6%
VGGFace [32]	2.6M	2.6K	43.7%	38.6%	5.8%	6.9%	2.1%	2.9%
YouTube [33]	621K	1.6K	56.9%	20.3%	7.7%	4.0%	7.9%	3.0%
CASIA [34]	500K	10.5K	48.8%	33.2%	7.2%	5.7%	2.6%	2.6%
CelebA [35]	203K	10.2K	33.9%	41.5%	6.4%	8.2%	4.4%	5.5%
PubFig [36]	58K	200	49.5%	35.5%	6.5%	5.5%	2.0%	1.0%
IJB-C [37]	21K	3.5K	40.3%	30.2%	11.8%	6.0%	5.4%	6.2%
UTKface [38]	24K	-	26.2%	20.0%	21.5%	16.3%	7.1%	8.9%
LFW [39]	13K	5.7K	58.9%	18.7%	9.6%	3.3%	7.2%	2.2%
BioSecure [40]	2.7K	667	50.1%	36%	3.1%	2.1%	4.3%	4.5%
Average			46%	29%	10%	6%	5%	4%
Databases for discrimination-aware learning								
BUPT-B [18]	1.3M	28K	33.33%		33.33%		33.33%	
DiveFace [41]	125K	24K	16.7%	16.7%	16.7%	16.7%	16.7%	16.7%
FairFace [27]	100K	-	25.0%	20.0%	14.4%	13.9%	13.6%	13.1%
RFW [25]	40K	12K	33.33%		33.33%		33.33%	
DemogPairs [15]	10.8K	600	16.7%	16.7%	16.7%	16.7%	16.7%	16.7%

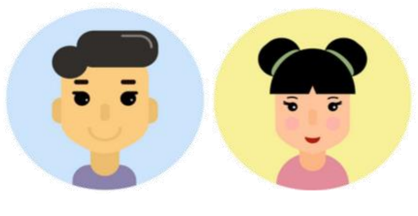
I. Serna, A. Morales, J. Fierrez and N. Obradovich, "SensitiveLoss: Improving Accuracy and Fairness of Face Representations with Discrimination-Aware Deep Learning" , *Artificial Intelligence*, vol. 305, pp. 103682, April 2022.

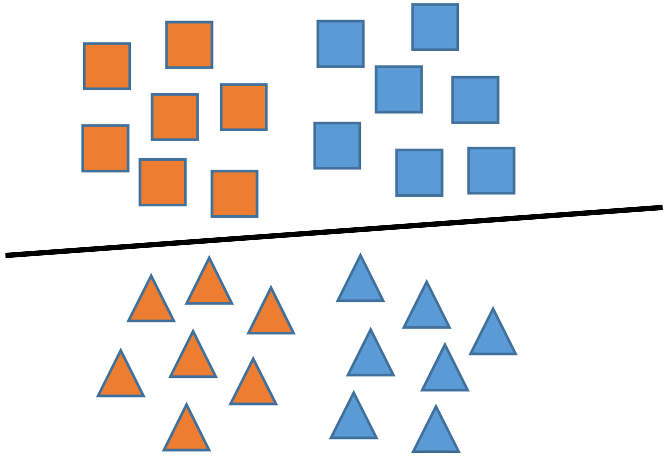
Analysis of Biased Performance



ResNet-50 performance

EER (Equal Error Rate)

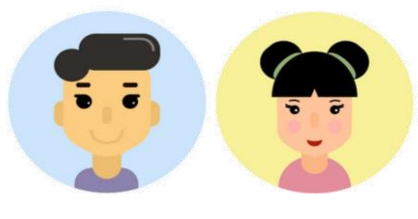
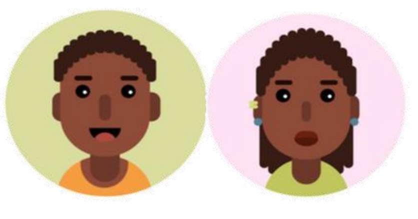
Asian	Black	Caucassian	Avg.	Std
				
5.75%	5.96%	3.62%	5.11%	1.06

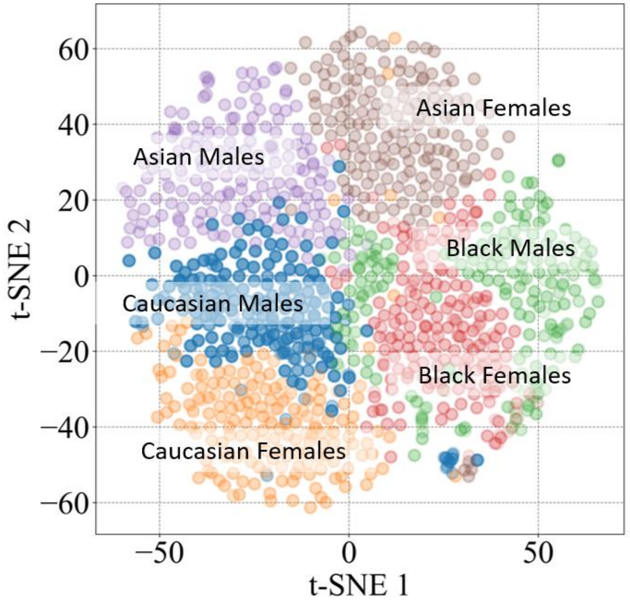


I. Serna, A. Morales, J. Fierrez and N. Obradovich, "SensitiveLoss: Improving Accuracy and Fairness of Face Representations with Discrimination-Aware Deep Learning" , *Artificial Intelligence*, vol. 305, pp. 103682, April 2022.

ResNet-50 performance

EER (Equal Error Rate)

Asian	Black	Caucassian	Avg.	Std
				
5.75%	5.96%	3.62%	5.11%	1.06



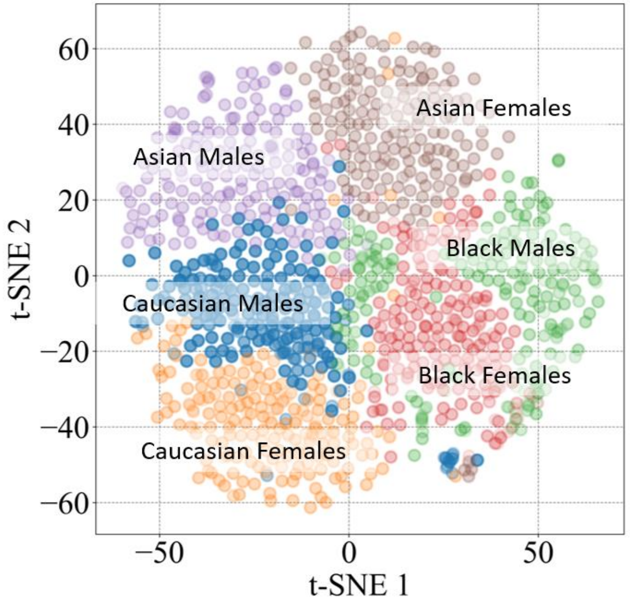
Projections of the embeddings into the 2D space generated with t-SNE.

I. Serna, A. Morales, J. Fierrez and N. Obradovich, "SensitiveLoss: Improving Accuracy and Fairness of Face Representations with Discrimination-Aware Deep Learning" , *Artificial Intelligence*, vol. 305, pp. 103682, April 2022.

ResNet-50 performance

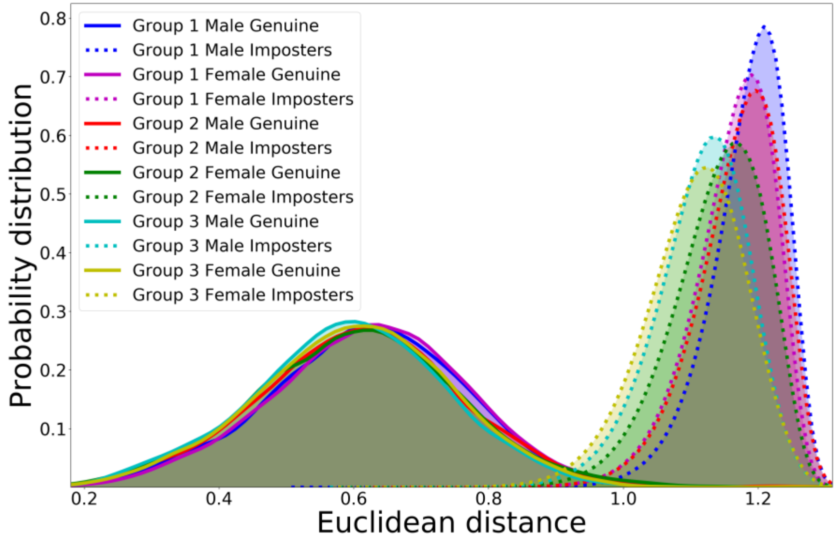
EER (Equal Error Rate)

Asian	Black	Caucassian	Avg.	Std
				
5.75%	5.96%	3.62%	5.11%	1.06



Projections of the embeddings into the 2D space generated with t-SNE.

Score distributions



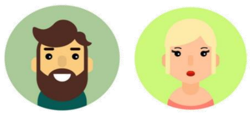
I. Serna, A. Morales, J. Fierrez and N. Obradovich, "SensitiveLoss: Improving Accuracy and Fairness of Face Representations with Discrimination-Aware Deep Learning" , *Artificial Intelligence*, vol. 305, pp. 103682, April 2022.

Performance on RFW (Racial Faces in the Wild)

Responsible AI Improves the Recognition Performance

	Asian	Black	Caucassian	Avg.	Std
					
Without RAI	5.75%	5.96%	3.62%	5.11%	1.06
With RAI	3.99%	3.83%	3.02%	3.61% (↓**30%**)	0.42% (↓**60%**)

$$\min_{\Omega}\left(\sum_{\mathbf{x}\in S^1}\mathcal{L}(\phi(\mathbf{x}|\Omega),T)+\sum_{\mathbf{x}\in S^2}\mathcal{L}(\phi(\mathbf{x}|\Omega),T)+\sum_{\mathbf{x}\in S^3}\mathcal{L}(\phi(\mathbf{x}|\Omega),T)\right)$$

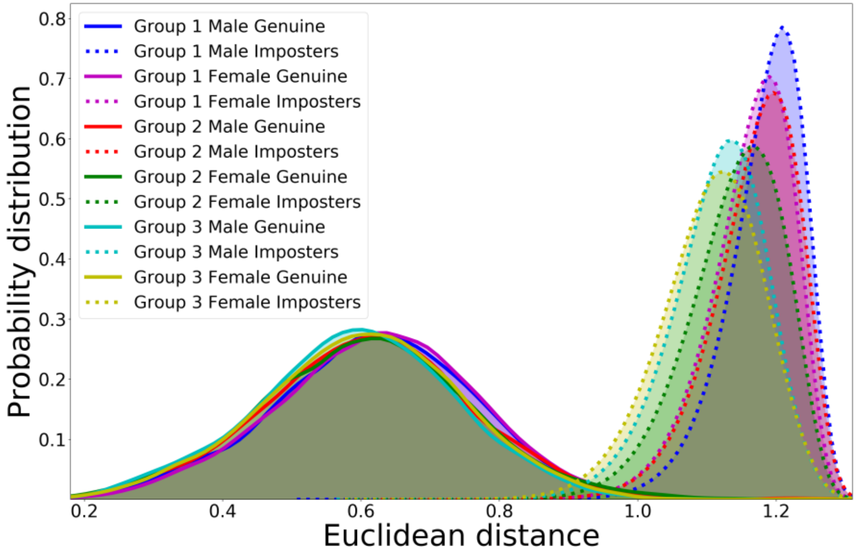
  

I. Serna, A. Morales, J. Fierrez and N. Obradovich, "SensitiveLoss: Improving Accuracy and Fairness of Face Representations with Discrimination-Aware Deep Learning" , *Artificial Intelligence*, vol. 305, pp. 103682, April 2022.

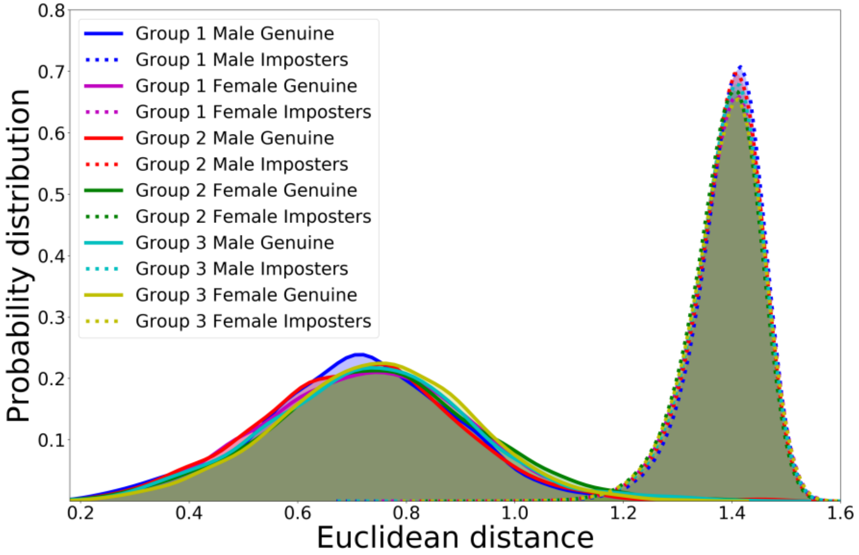
Grad-CAM and Distribution Scores

Grad-CAM

Without RAI



With RAI



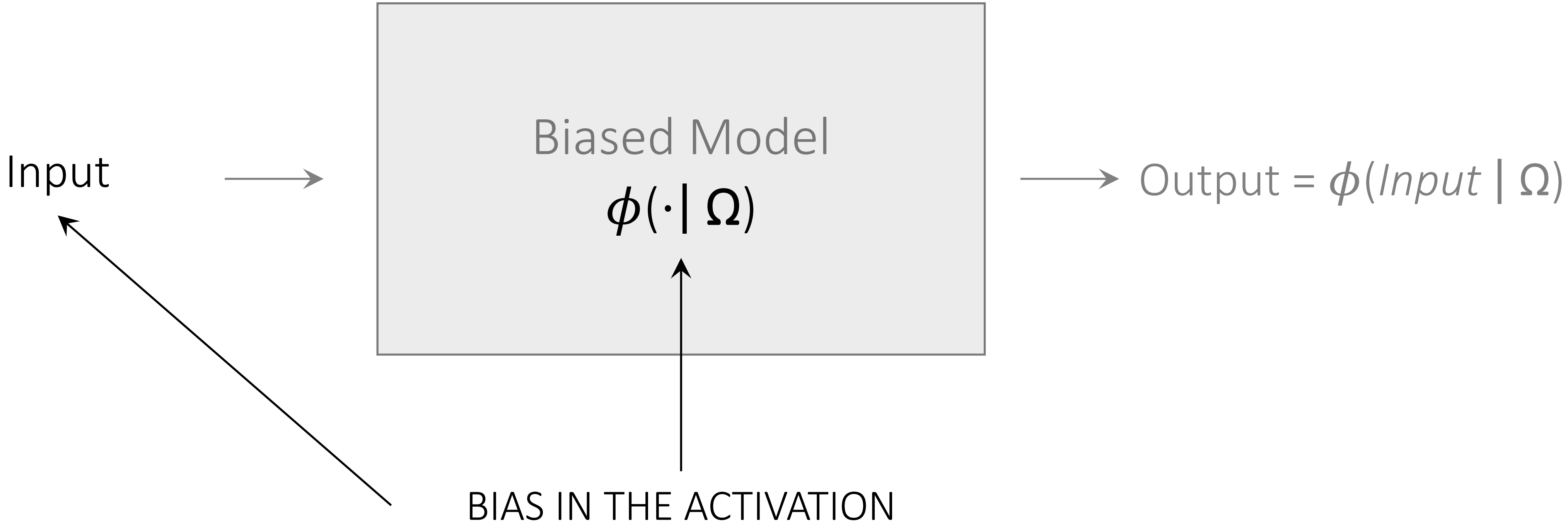
Distribution Scores

I. Serna, A. Morales, J. Fierrez and N. Obradovich, "SensitiveLoss: Improving Accuracy and Fairness of Face Representations with Discrimination-Aware Deep Learning" , *Artificial Intelligence*, vol. 305, pp. 103682, April 2022.

Limitations:

- You have to know in advance the variable of bias.

Activation Level Bias Analysis



InsideBias

Accuracy (%) in gender calssification

Model	A	B	C
Biased (A)	96.8	94.1	94.5
Balanced	95.5	95.3	96.1

- Group A: Asian
- Group B: Black
- Group C: Caucasian

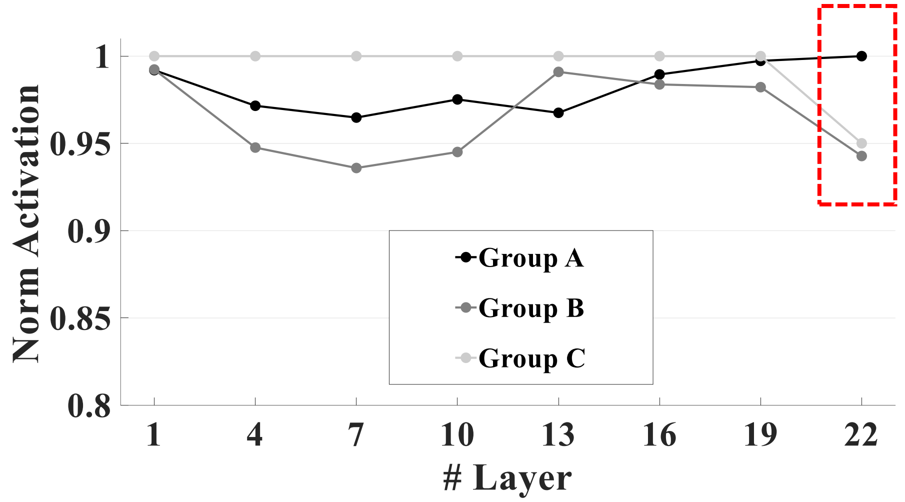
I. Serna, A. Peña, A. Morales and J. Fierrez, "InsideBias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics", in *IAPR Intl. Conf. on Pattern Recognition (ICPR)*, January 2021.

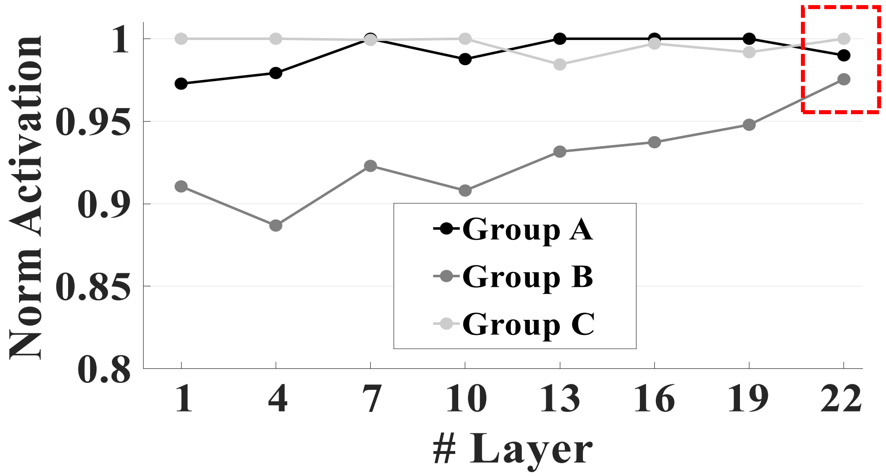
InsideBias

Normalized activation λ'

Accuracy (%) in gender calssification

Model	A	B	C
Biased (A)	96.8	94.1	94.5
Balanced	95.5	95.3	96.1





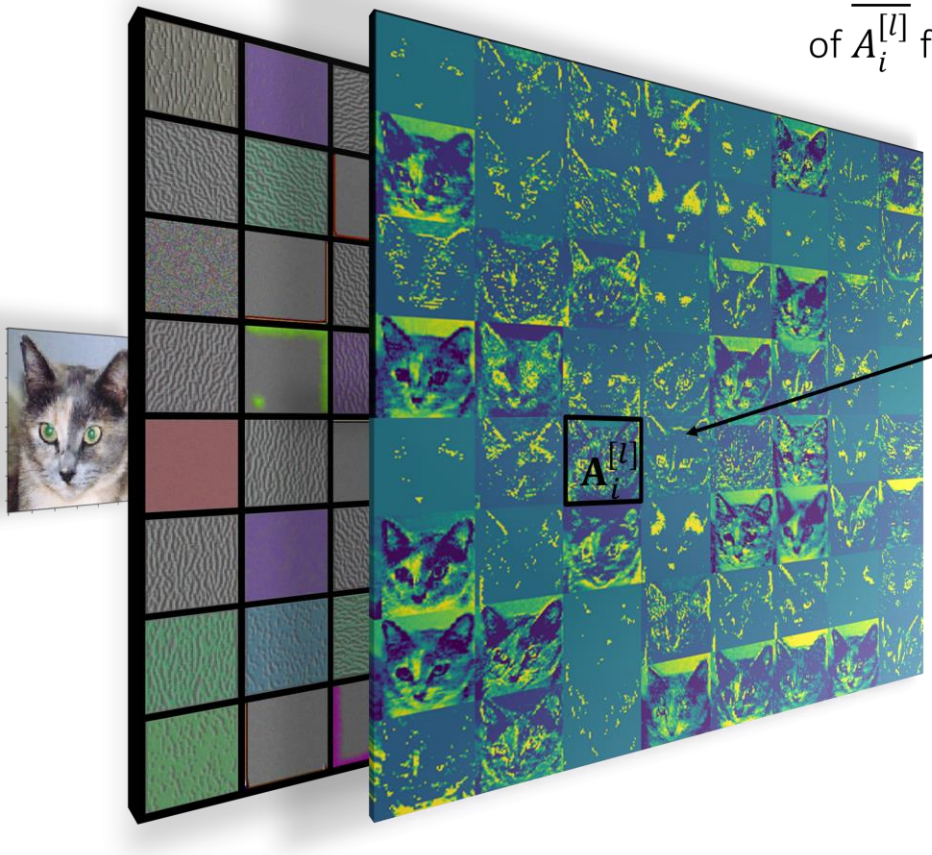
- Group A: Asian
- Group B: Black
- Group C: Caucasian

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)*, January 2021.

InsideBias

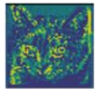
Novel bias detection method based on the analysis of the filter's activation of deep networks.

- We show how bias impacts in the activations of gender detection models based on face images.



We calculate the *activation* $\lambda^{[l]}$ as the ***maximum*** of $\overline{A_i^{[l]}}$ for all feature maps i in layer l

$$\lambda^{[l]} = \max_i \left(\overline{A_i^{[l]}} \right) = 6.54$$



$\mathbf{A}_i^{[l]}$

Average Activation $= \overline{A_i^{[l]}} = 2.10$

$$\begin{bmatrix} 0.11 & 1.34 & 2.02 & 3.17 & 0.33 & \boxed{6.54} & 0.56 & 3.78 \\ 6.51 & 1.56 & 2.37 & 2.89 & 2.21 & 0.10 & 0.80 & 2.51 \\ 2.68 & 3.24 & 3.35 & 0.42 & 2.26 & 5.84 & 1.78 & 1.40 \\ 5.99 & 4.04 & 3.89 & 3.23 & 4.95 & 4.02 & 1.98 & 0.03 \\ 0.11 & 1.78 & \boxed{2.10} & 3.25 & 3.52 & 4.73 & 1.94 & 3.17 \\ 6.30 & 0.84 & 3.79 & 2.81 & 5.71 & 5.86 & 2.98 & 3.04 \\ 6.08 & 5.76 & 0.08 & 1.71 & 0.92 & 1.23 & 0.12 & 1.54 \\ 1.67 & 4.55 & 0.09 & 4.81 & 4.67 & 4.44 & 4.53 & 1.76 \end{bmatrix}$$

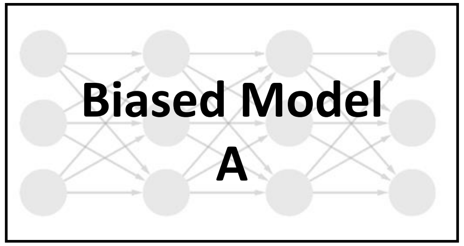
I. Serna, A. Peña, A. Morales and J. Fierrez, "InsideBias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics", in *IAPR Intl. Conf. on Pattern Recognition (ICPR)*, January 2021.

Experiments: Detecting Bias with Very Few Samples

Confidence score (in Gender Classification)

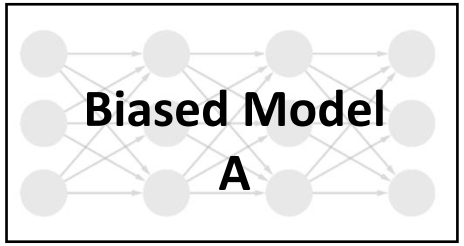
<u>Only **5 samples**</u>

Group **A**



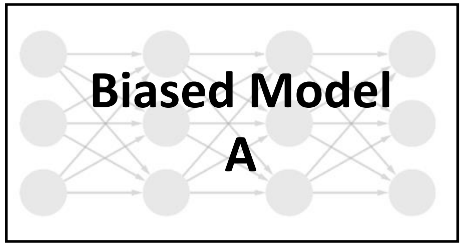
100%

Group **B**



100%

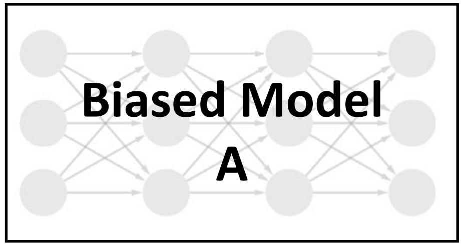
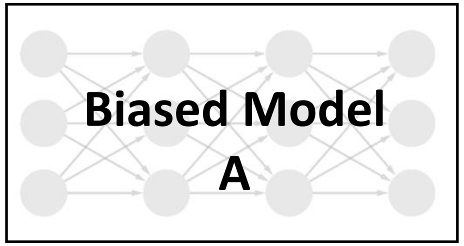
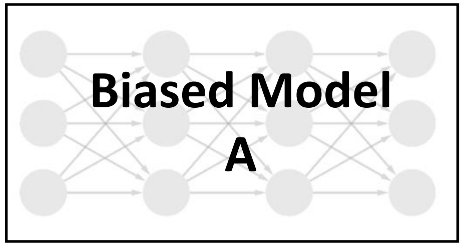
Group **C**



100%

Experiments: Detecting Bias with Very Few Samples

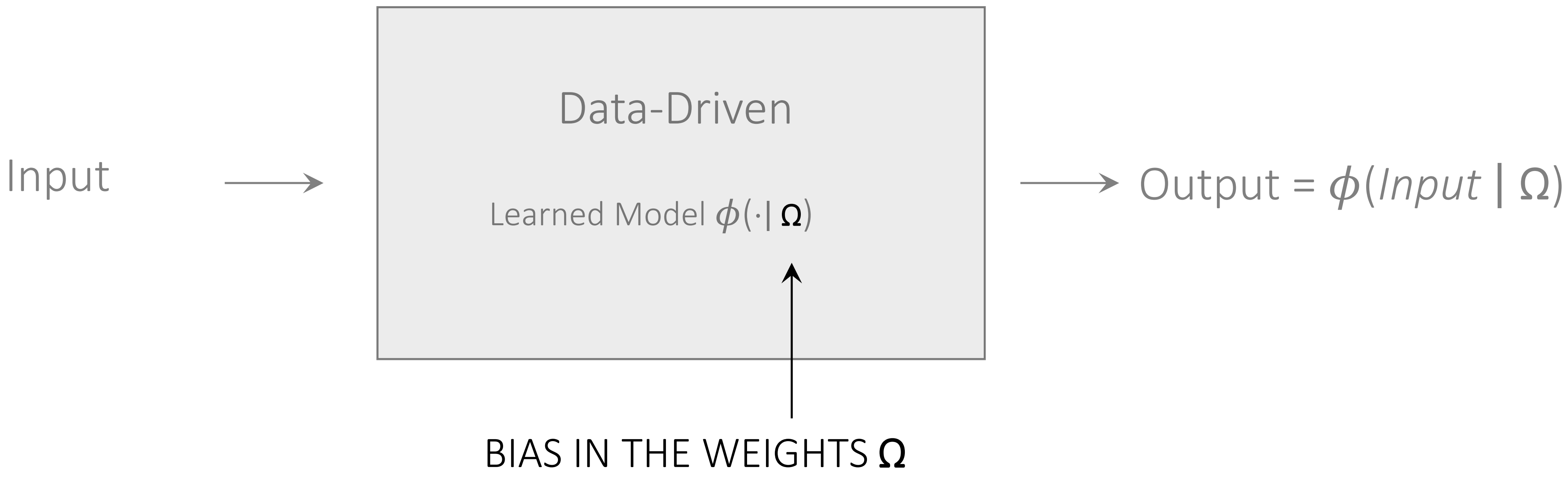
Only **5 samples**

			Confidence score (in Gender Classification)	**Activation** $\lambda^{[l]}$ (for the Group)	**Activation Ratio** $\Lambda_d^{[l]}$
Group **A**			**100%**	**2.82**	
Group **B**			**100%**	**2.65**	$^{2.53}/_{2.82}$ $= \mathbf{0.90}$
Group **C**			**100%**	**2.53**	

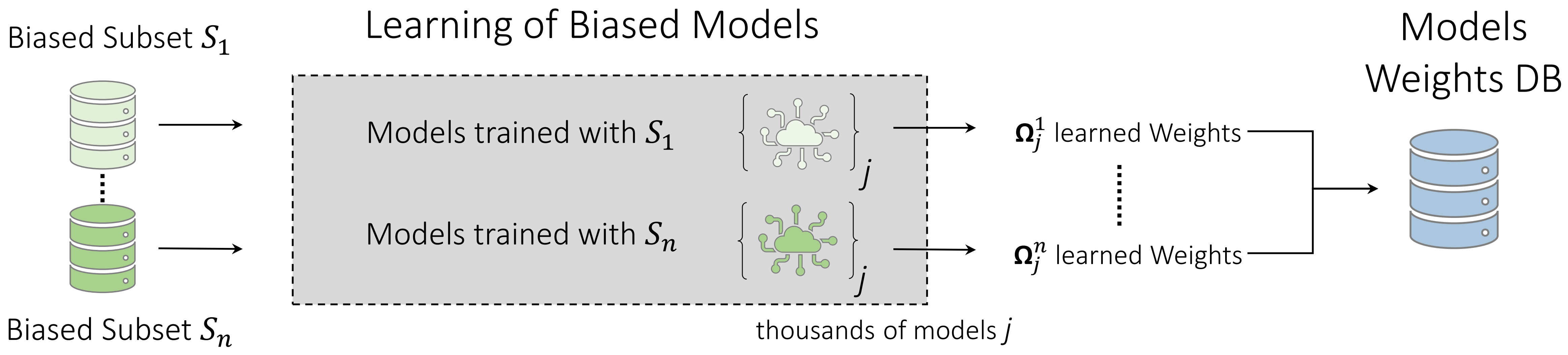
Limitations and challenges:

- You have to know in advance the variable of bias.
- It is dependent on the input data.

Bias Analysis of Weights

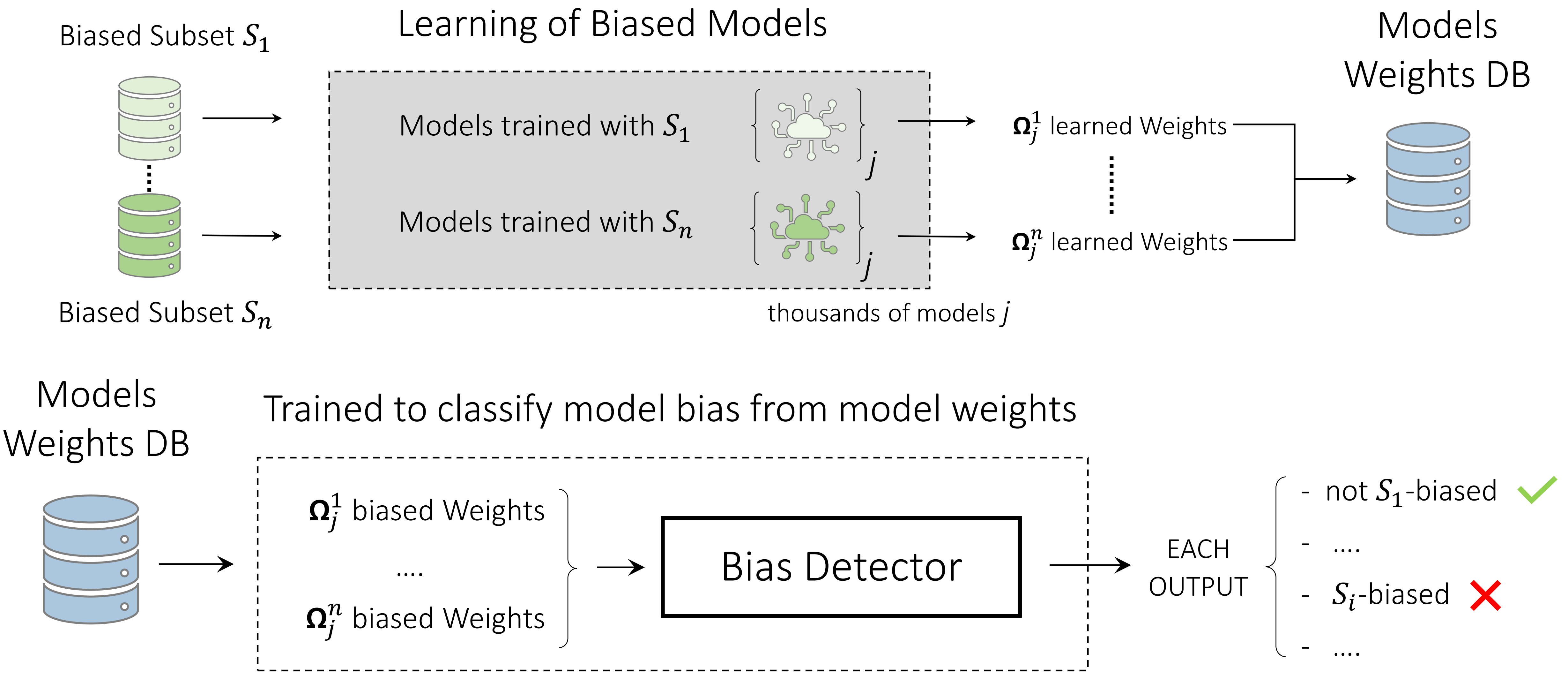


InputData-Independent Bias Detection



I. Serna, A. Morales, J. Fierrez and J. Ortega-Garcia, "IFBiD: Inference-Free Bias Detection", in *AAAI Workshop on Artificial Intelligence Safety (SafeAI)*, CEUR, vol. 3087, Vancouver, Canada, February 2022.

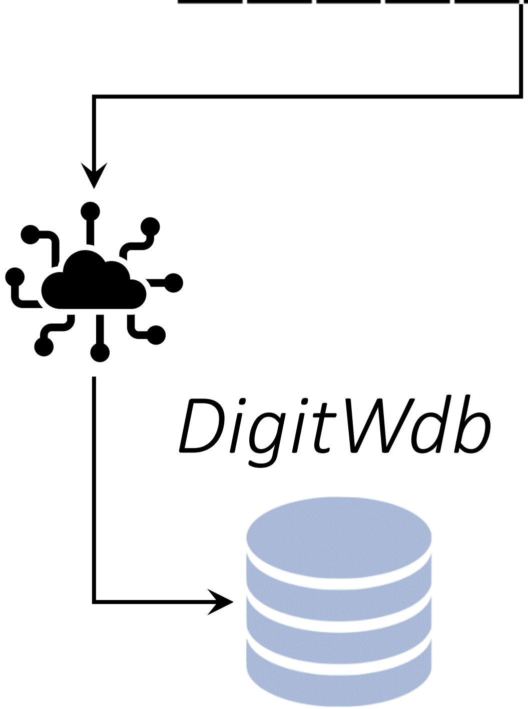
InputData-Independent Bias Detection



I. Serna, A. Morales, J. Fierrez and J. Ortega-Garcia, "IFBiD: Inference-Free Bias Detection", in *AAAI Workshop on Artificial Intelligence Safety (SafeAI)*, CEUR, vol. 3087, Vancouver, Canada, February 2022.

DigitWdb and GenderWdb

ColoredMNIST[1]



Train 24K models
of 50K params each
with 2 levels of bias:

- very high bias
- very low bias

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

DigitWdb and GenderWdb

ColoredMNIST[1]

DigitWdb

Train 24K models
of 50K params each
with 2 levels of bias:

- very high bias
- very low bias

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

Results: Bias Detection

Bias in Digit Classifier

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

Bias in Gender Classifier

- 30K training models:
   10K asian biased
   10K african/indian biased
   10K caucasian biased
- 6k test models:
   2k  2k  2k

I. Serna, A. Morales, J. Fierrez and J. Ortega-Garcia, "IFBiD: Inference-Free Bias Detection", in *AAAI Workshop on Artificial Intelligence Safety (SafeAI)*, CEUR, vol. 3087, Vancouver, Canada, February 2022.

Results: Bias Detection

Bias in Digit Classifier

- 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 %**

Bias in Gender Classifier

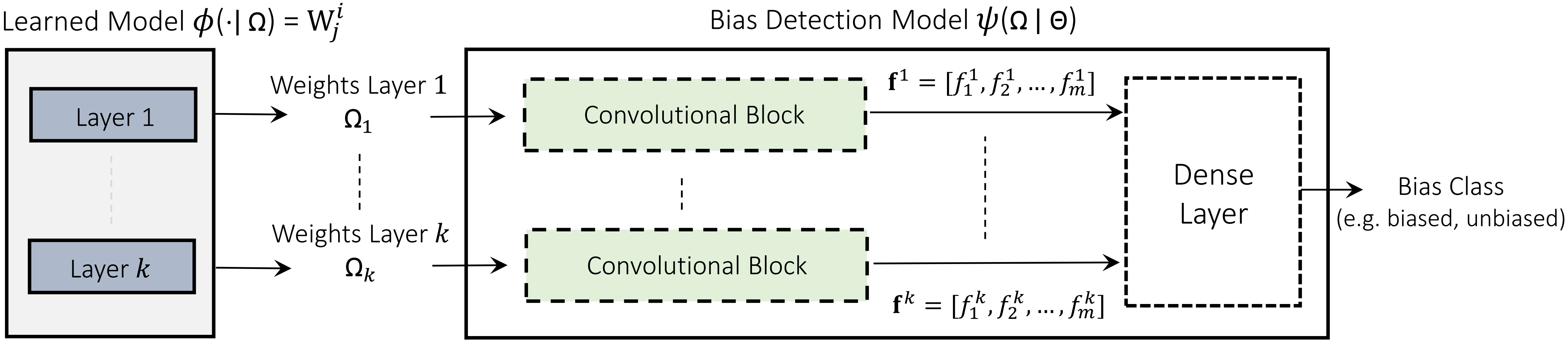
- 30K training models:
  - 10K asian biased
  - 10K african/indian biased
  - 10K caucasian biased
- 6k test models:
  - 2k 2k 2k

Accuracy obtained:

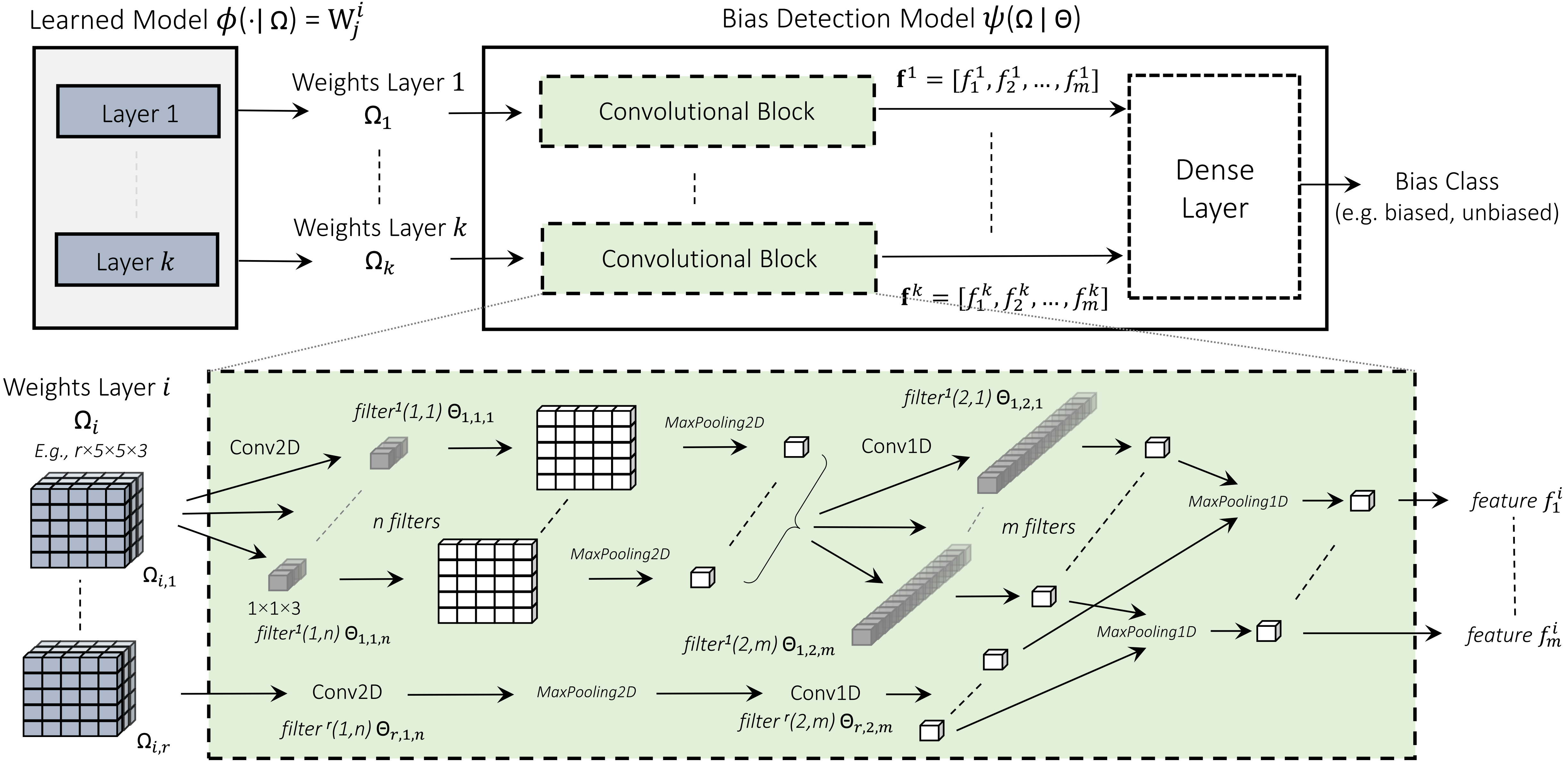
- Multi-layer perceptron: 60.8 %
- Convolutional Block: **89.9 %**

I. Serna, A. Morales, J. Fierrez and J. Ortega-Garcia, "IFBiD: Inference-Free Bias Detection", in *AAAI Workshop on Artificial Intelligence Safety (SafeAI)*, CEUR, vol. 3087, Vancouver, Canada, February 2022.

THE DETECTOR



THE DETECTOR



Limitations and Future Work

Limitations:

- It is still not free of conflated biases.

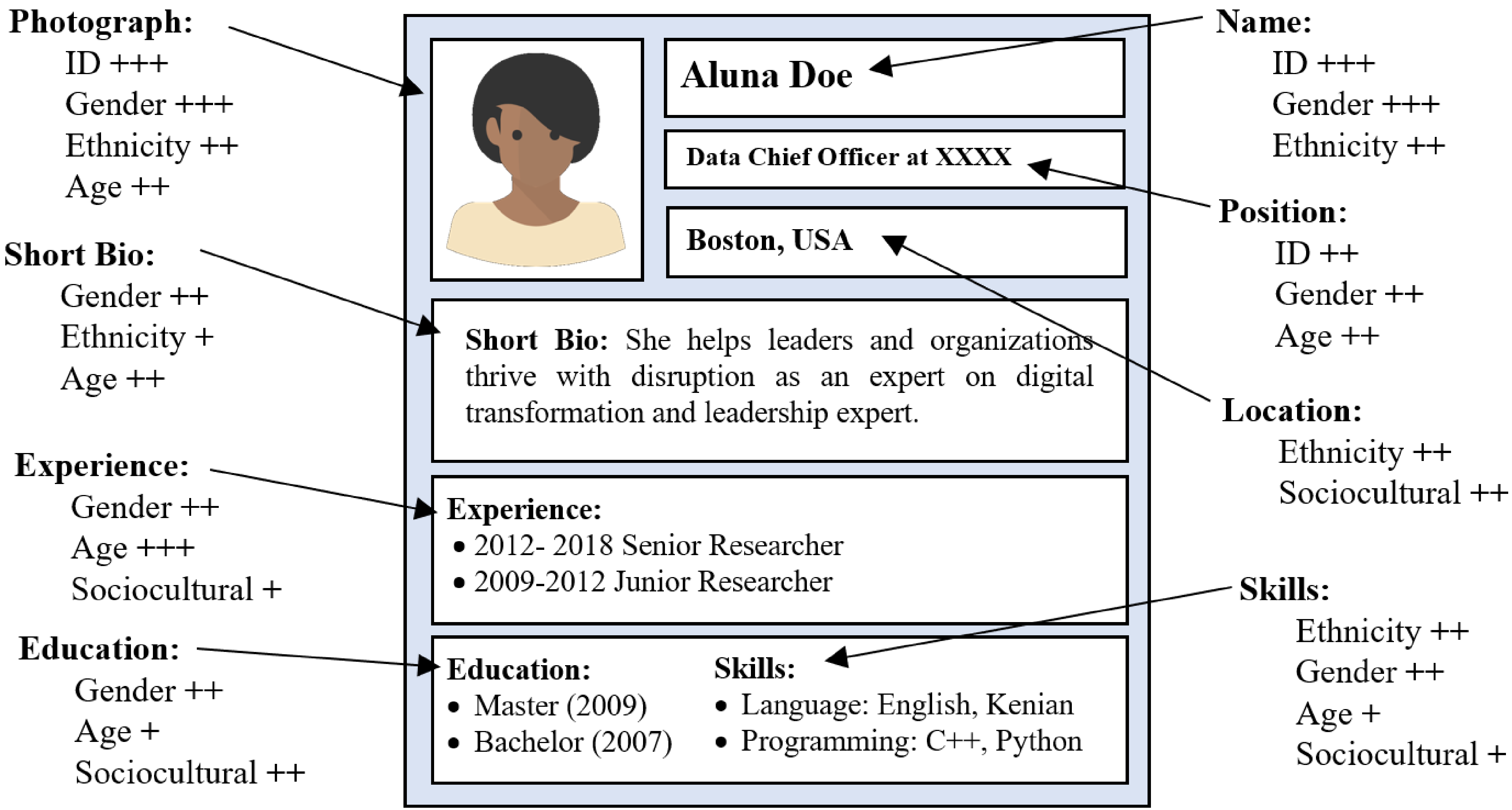
This work poses two fundamental challenges:

- Finding a way to translate our approach to other problems, such as face recognition. E.g., architecture-independent detector.
- Automatically detecting bias covariates.

Case Study on Multimodal Bias: Automatic Recruitment Tools

Material from Alejandro Peña

What else does your resume data reveal?



FairCVdb: Research dataset for multimodal AI

- **24K Profiles** including:
  - 12 features obtained from 5 information blocks (**merits**)
  - 2 demographic attributes (**gender** and **ethnicity**)
  - 1 face image from **DiveFace** database[1]
  - 1 candidate score (**human resources** equation)

Candidate competencies

$$\mathbf{x}^j = [x_1^j, \dots, x_n^j]$$

$\longrightarrow$

Candidate score

$$T^j = \beta^j + \sum_{i=1}^{n} \alpha_i x_i^j$$

A. Peña, I. Serna, A. Morales, J. Fierrez, "**Bias in Multimodal AI: Testbed for Fair Automatic Recruitment**," *Proc. of IEEE CVPR Workshop on Fair, Data Efficient and Trusted Computer Vision*, Washington, Seattle, USA, 2020. [pdf][Github]

FairCVdb: Research dataset for multimodal AI

- **24K Profiles** including:
  - 12 features obtained from 5 information blocks (**merits**)
  - 2 demographic attributes (**gender** and **ethnicity**)
  - 1 face image from **DiveFace** database[1]
  - 1 candidate score (**human resources** equation)

Candidate competencies **(Unbiased)**

Candidate score **(Biased)**

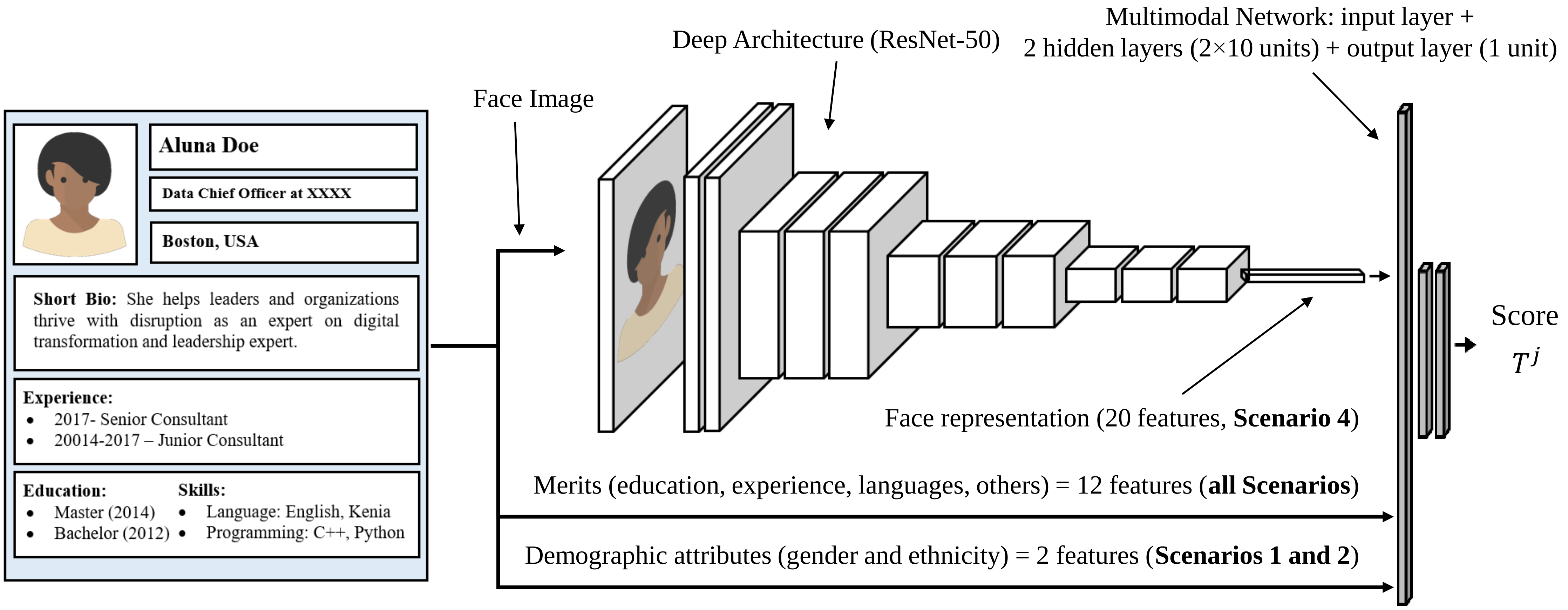
$$\mathbf{x}^j = [x_1^j, \ldots, x_n^j] \longrightarrow T^j = \beta^j + \sum_{i=1}^{n} \alpha_i x_i^j$$

+ Bias (Gender and Ethnicity)

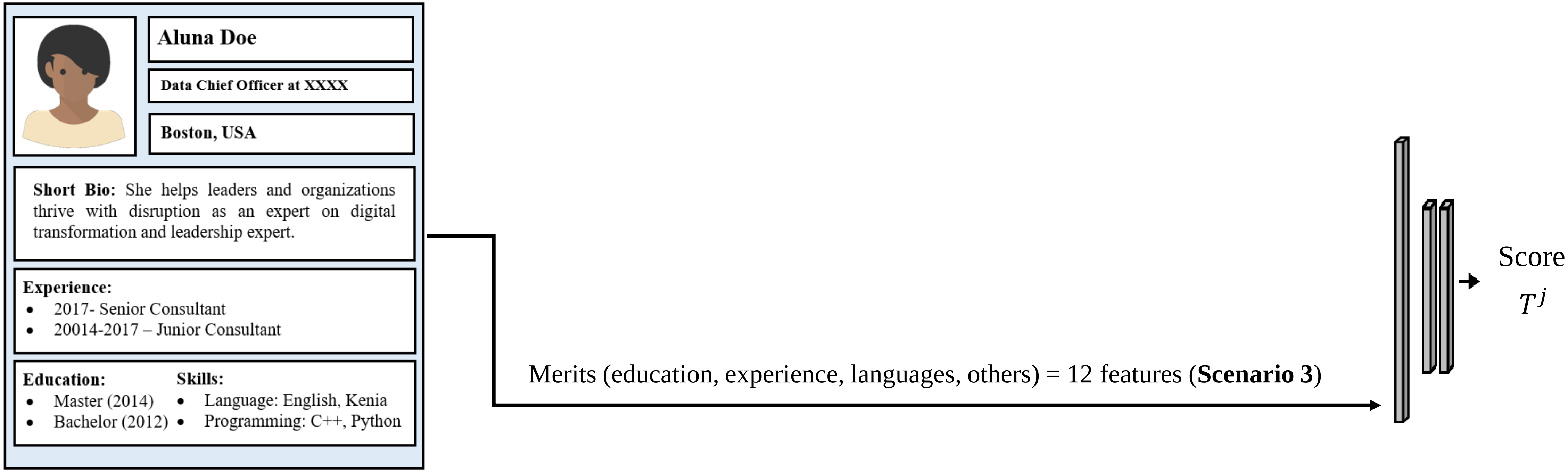
https://github.com/BiDAlab/FairCVtest

A. Peña, I. Serna, A. Morales, J. Fierrez, "**Bias in Multimodal AI: Testbed for Fair Automatic Recruitment**," *Proc. of IEEE CVPR Workshop on Fair, Data Efficient and Trusted Computer Vision*, Washington, Seattle, USA, 2020. [pdf][Github]

Multimodal Learning Architecture



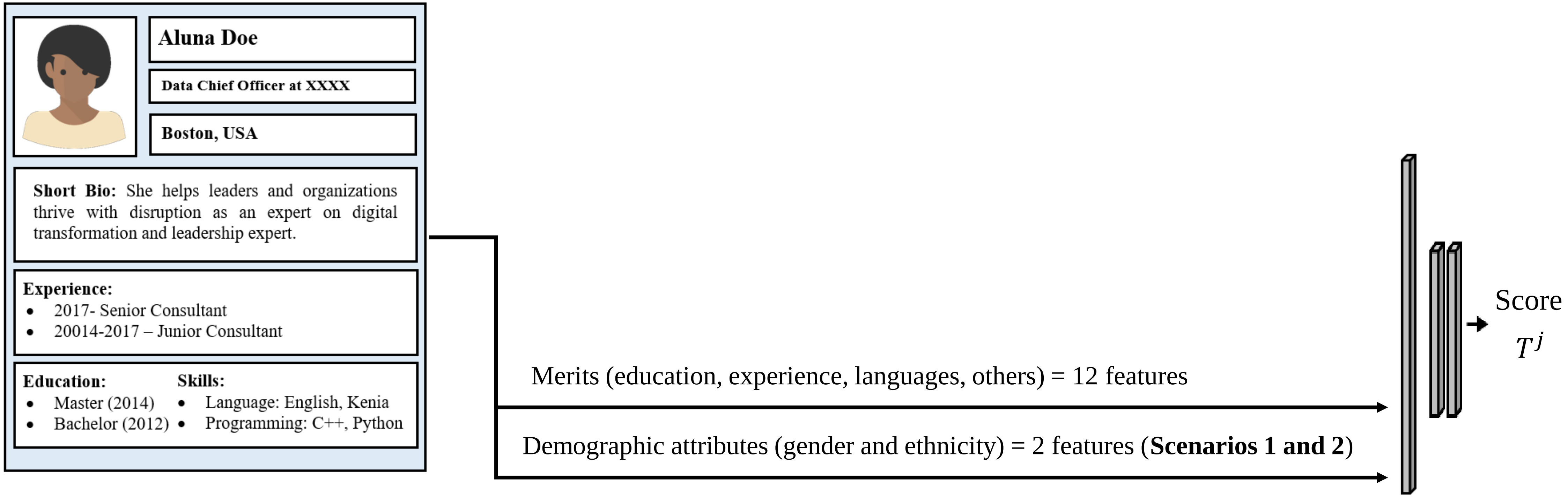
Multimodal Learning Architecture



Distribution of the **top 100** candidates

Scenario	Bias	Input Features			Gender		Δ
		Merits	Dem.	Face	Male	Female	
3	yes	yes	no	no	50%	50%	0%

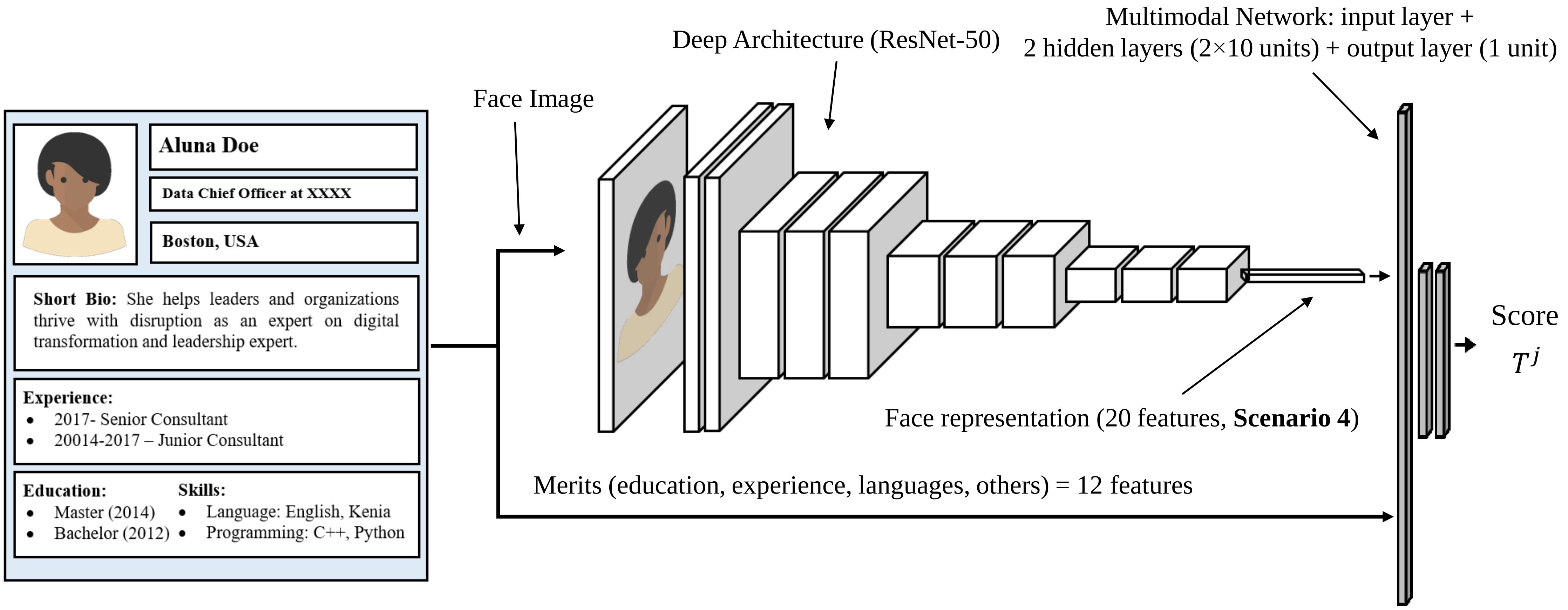
Multimodal Learning Architecture



Distribution of the **top 100** candidates

Scenario	Bias	Input Features			Gender		Δ
		Merits	Dem.	Face	Male	Female	
2	yes	yes	yes	no	87%	13%	74%

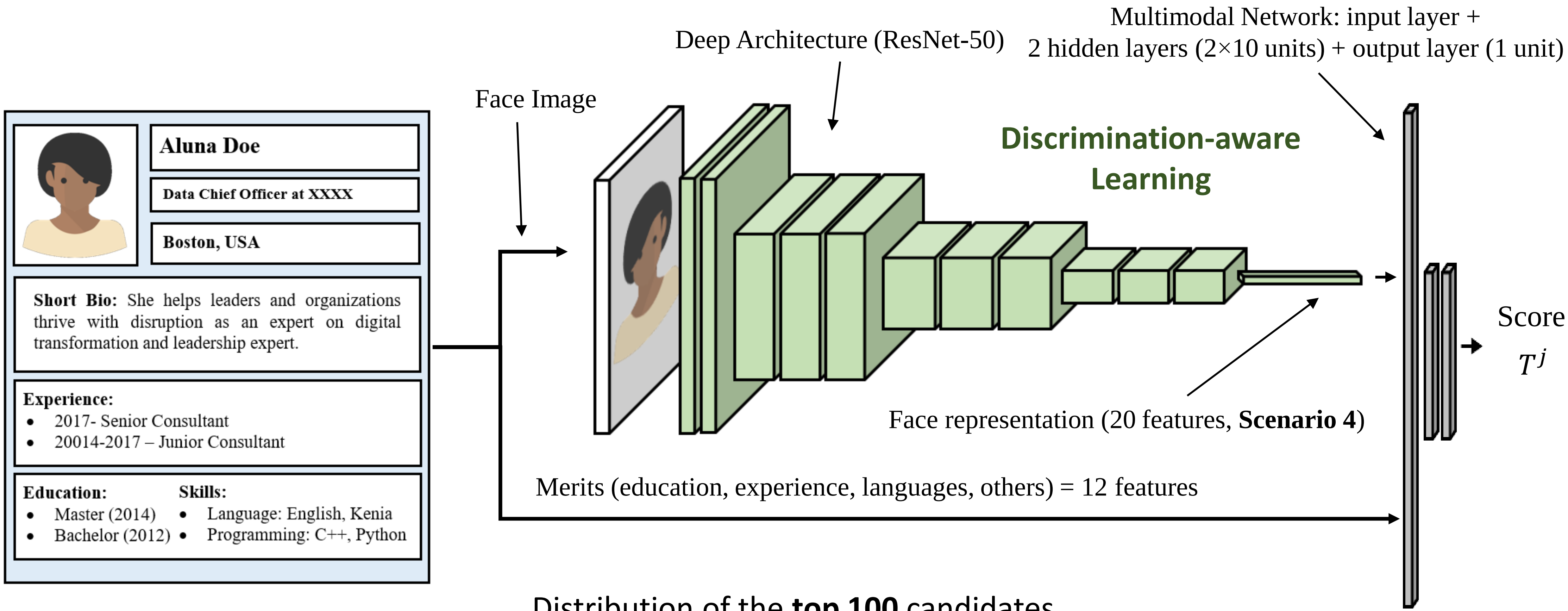
Multimodal Learning Architecture



Distribution of the **top 100** candidates

Scenario	Bias	Input Features			Gender		Δ
		Merits	Dem.	Face	Male	Female	
4	yes	yes	no	yes	77%	23%	54%

Multimodal Learning Architecture

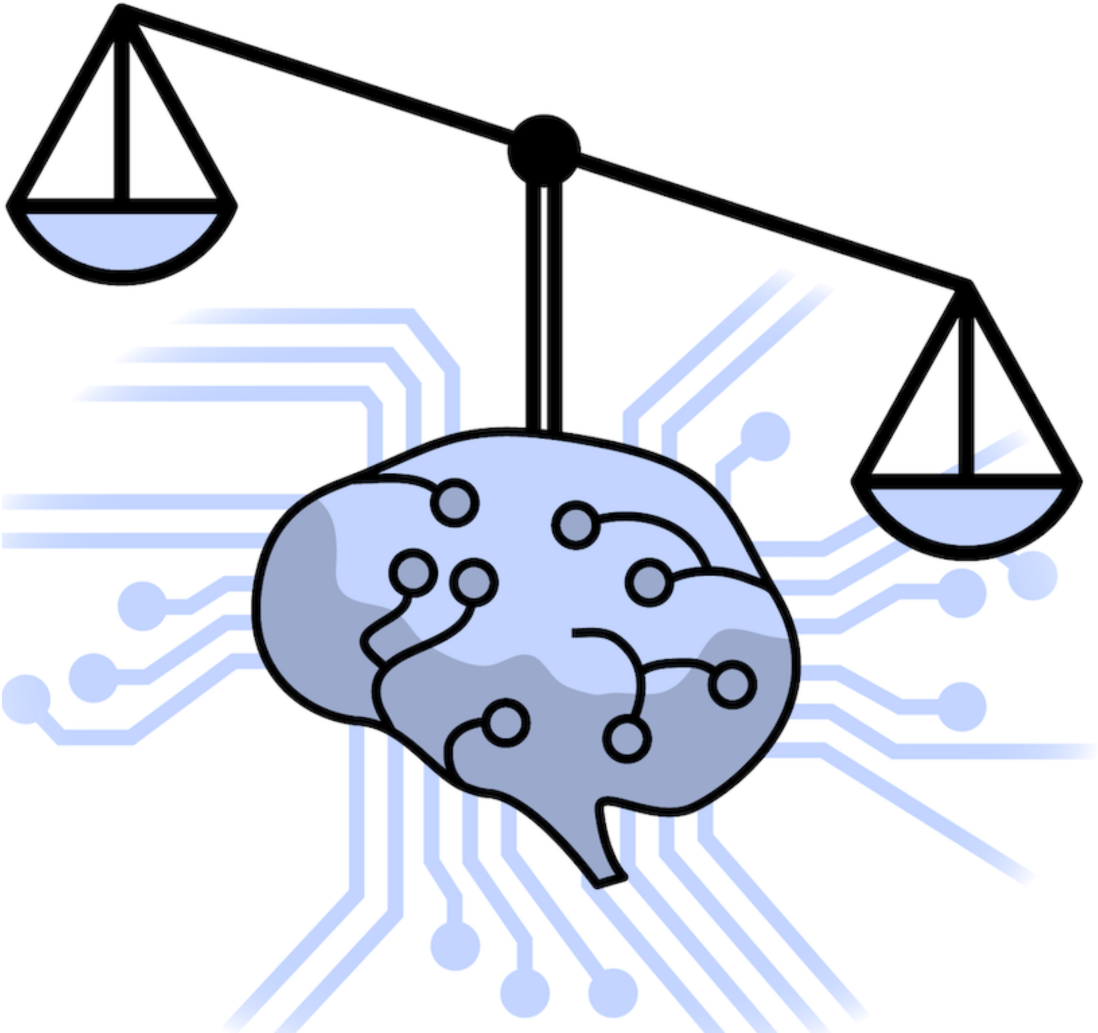


Distribution of the **top 100** candidates

Scenario	Bias	Input Features			Gender		Δ
		Merits	Dem.	Face	Male	Female	
Agnostic	yes	yes	no	yes	50%	50%	0%

[1] A. Morales, J. Fierrez, R. Vera-Rodriguez, R. Tolosana. SensitiveNets: Learning Agnostic Representations with Application to Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. [pdf][GitHub]

Bias Detection and Mitigation in Machine Learning



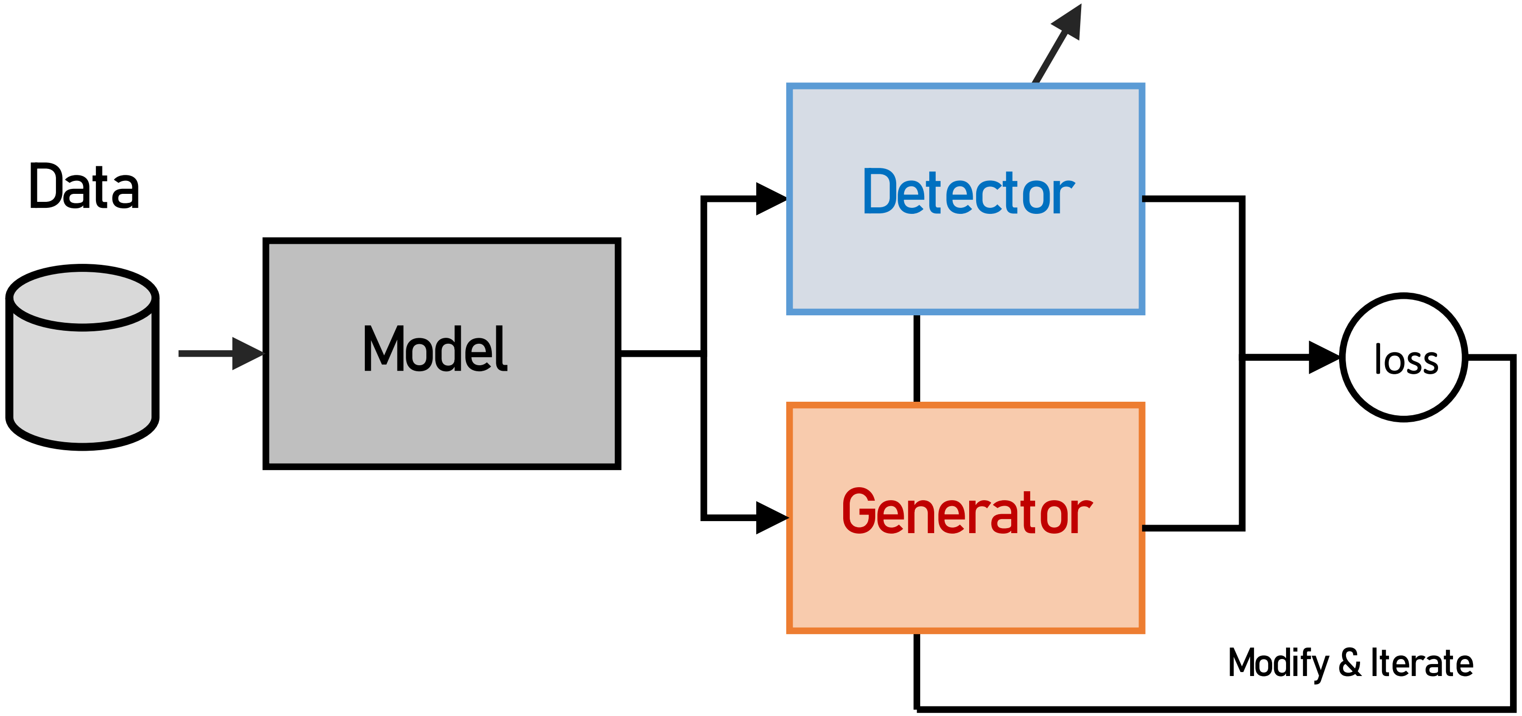
Aythami Morales

http://aythami.me

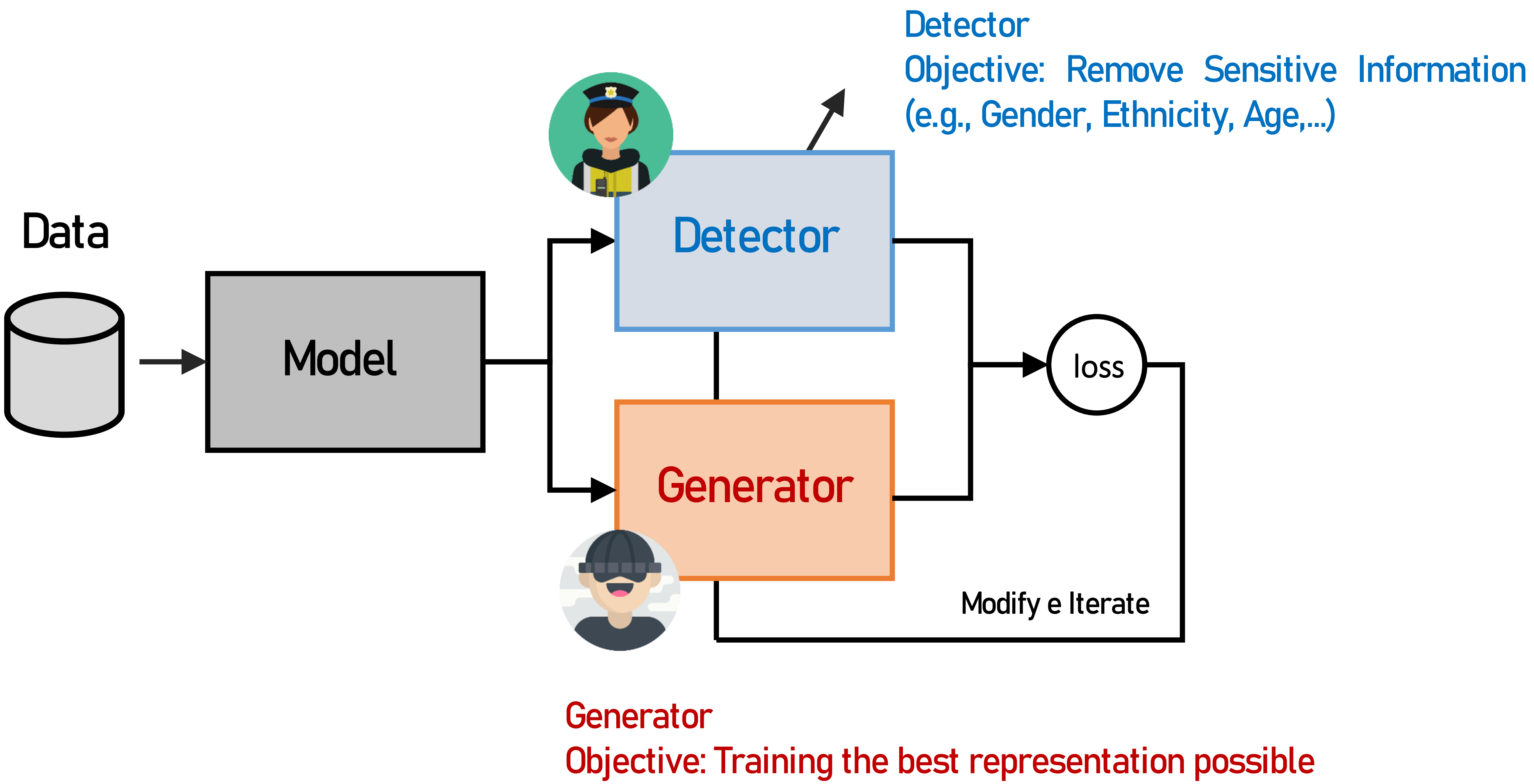
BiDA Lab
Biometrics & Data Pattern Analytics Lab

UAM
Universidad Autónoma de Madrid

Adversarial Learning to Train Fair Representations



Adversarial Learning to Train Fair Representations



D. Madras, E.Creager, T. Pitassi, R. Zemel, "Learning Adversarially Fair and Transferable Representations," *Proc. of the Int. Conf. on Machine Learning*, pp. 3384-3393, 2018.
A. Morales, J. Fierrez, R. Vera-Rodriguez, R. Tolosana. SensitiveNets: Learning Agnostic Representations with Application to Face Images. *IEEE Transactions on Pattern Analisys and Machine Intelligence*, 2020. [pdf][GitHub]

Adversarial Learning to Train Fair Representations

$$\min_{\mathbf{w}} \sum_{\mathbf{x}^j \in S} \left(\mathcal{L}\left(O\left(\mathbf{x}^j \middle| \mathbf{w} \right), T^j \right) + \Delta^j \right)$$

Adversarial Learning to Train Fair Representations

Loss Function Primary Task (**Generator**)

$$\min_{\mathbf{w}} \sum_{\mathbf{x}^j \in S} \left(\mathcal{L}\left(O\left(\mathbf{x}^j \middle| \mathbf{w}\right), T^j\right) + \Delta^j \right)$$

Sensitive Regularizer: Secondary Taks (**Detector**)

Adversarial Learning to Train Fair Representations

Loss Function Primary Task (**Generator**)

$\mathcal{L}$ = Triplet Loss

$$\min_{\mathbf{w}} \sum_{\mathbf{x}^j \in S} \left(\mathcal{L}\left(O\left(\mathbf{x}^j \middle| \mathbf{w} \right), T^j \right) + \Delta^j \right)$$

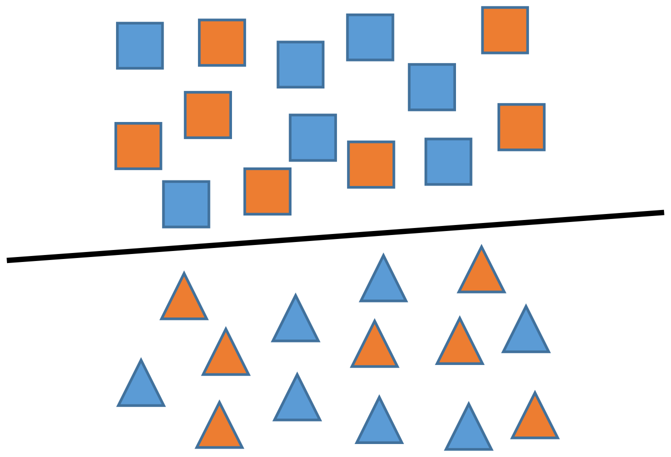
Sensitive Regularizer: Secondary Taks (**Detector**)

$$\Delta^j = \left| 0.9 - P_k(\mathbf{x}^j) \right|$$

Adversarial Learning to Train Fair Representations

Sensitive Regularizer: Secondary Taks (**Detector**)

$$\Delta^j = \left|0.9 - P_k(\mathbf{x}^j)\right|$$



Agnostic Learned Feature Space