Agnostic Label-Only Membership Inference Attack

Simone Rizzo,
Anna Monreale,
Francesca Naretto

SCUOLA NORMALE SUPERIORE

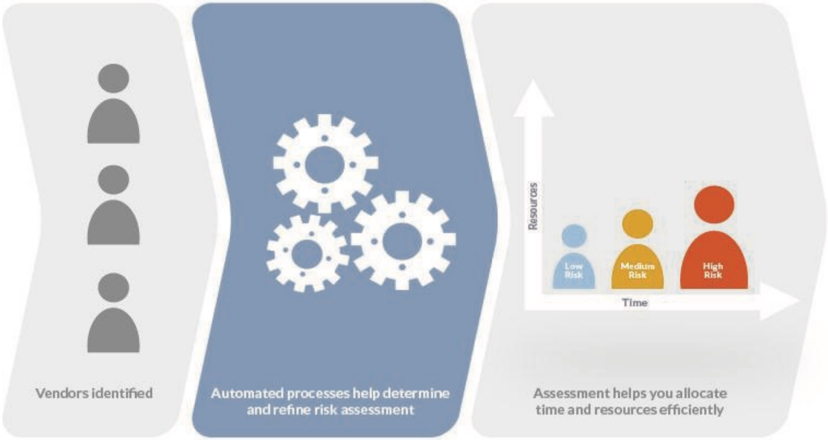
PRIVACY

General Data Protection Regulation

Data protection impact assessment

1. Assess the privacy risk of the process

2. Protect the privacy

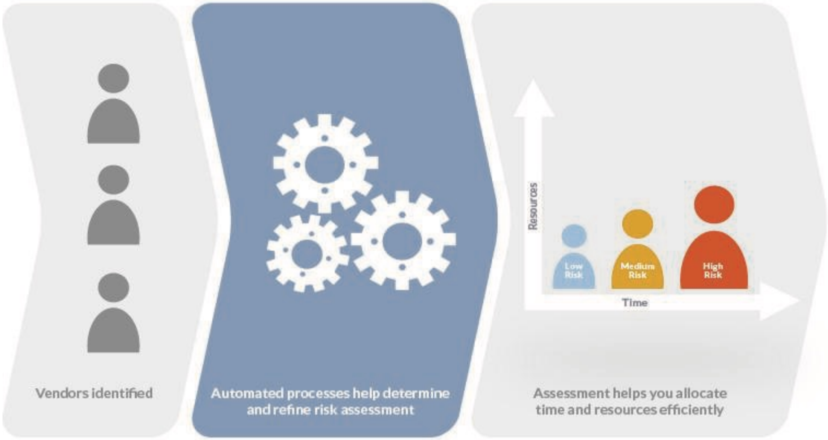


PRIVACY

General Data Protection Regulation

Data protection impact assessment

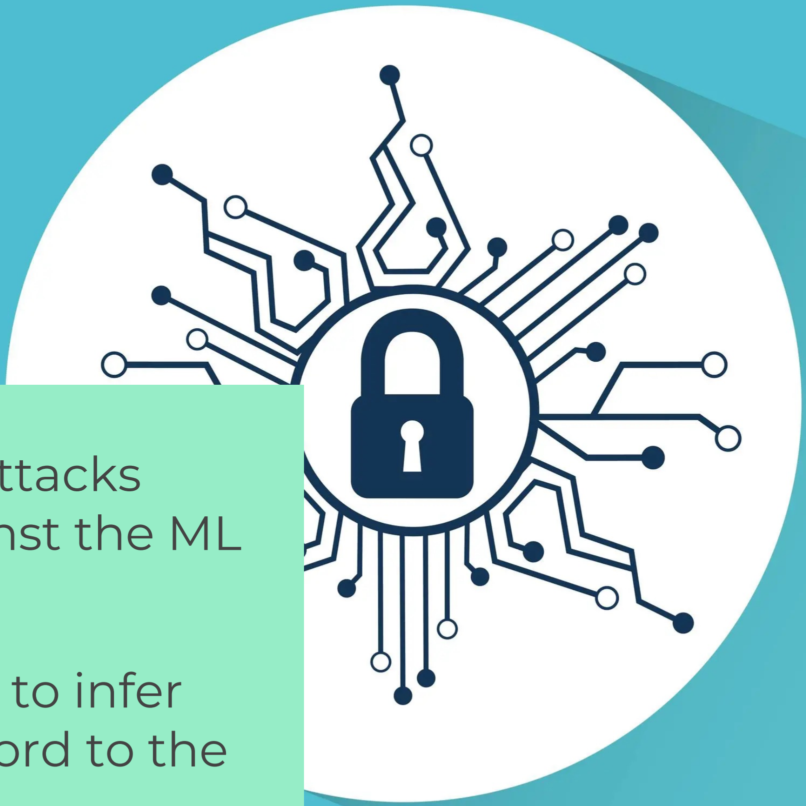
1. ***Assess the privacy risk of the process***
2. Protect the privacy



ATTACKS ON ML MODELS

There are some privacy attacks tailored for working against the ML models.

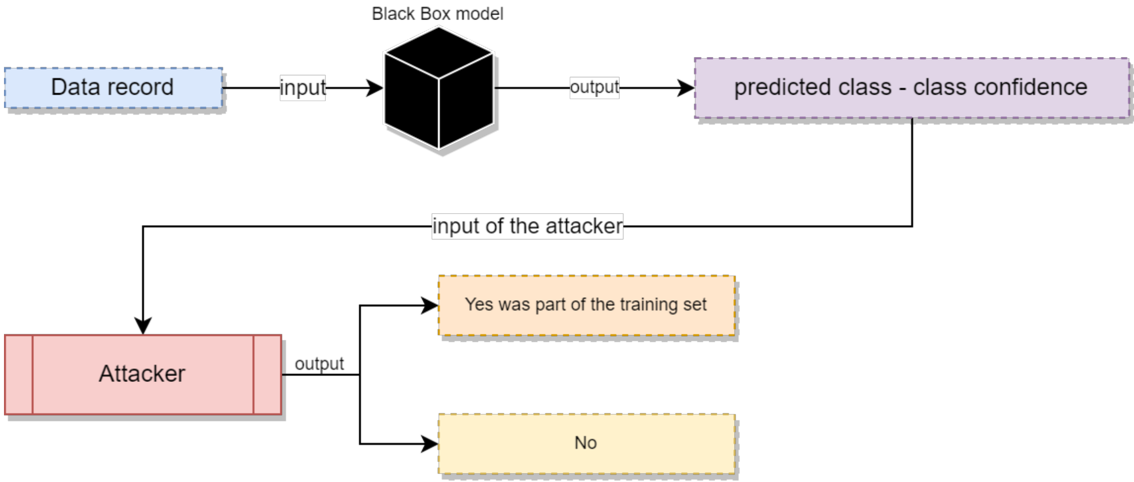
We focus on attacks able to infer the membership of a record to the original training data.



MEMBERSHIP INFERENCE ATTACK

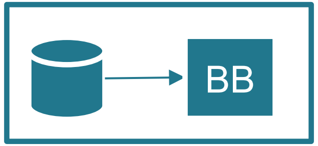
OBJECTIVE

Infer if a record was part of the training set or not.



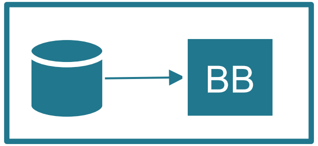
Reza Shokri, Marco Stronati, Congzheng Soang, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy

MEMBERSHIP INFERENCE ATTACK



Train a black box for a prediction task with ***n*** classes.

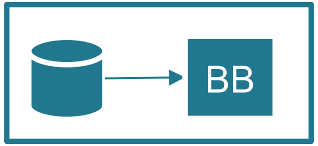
MEMBERSHIP INFERENCE ATTACK



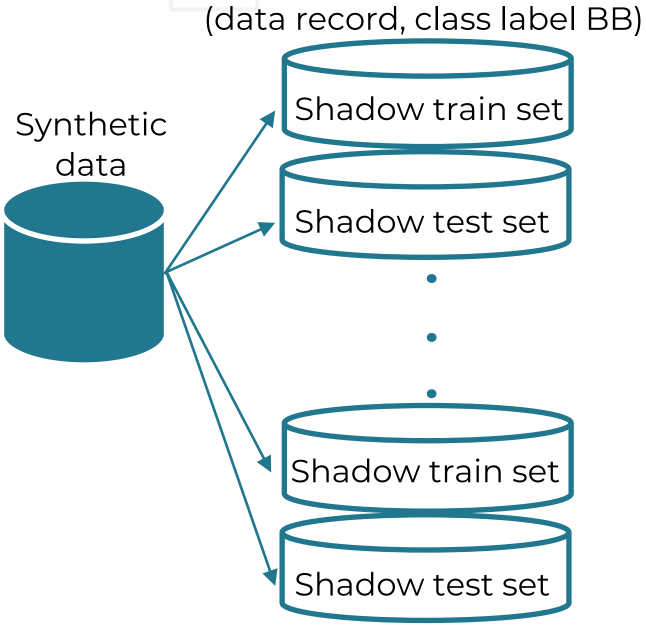
Train a black box for a prediction task with ***n*** classes.

Synthetic data

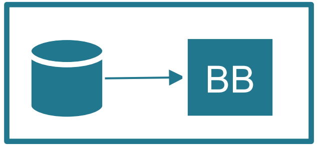
MEMBERSHIP INFERENCE ATTACK



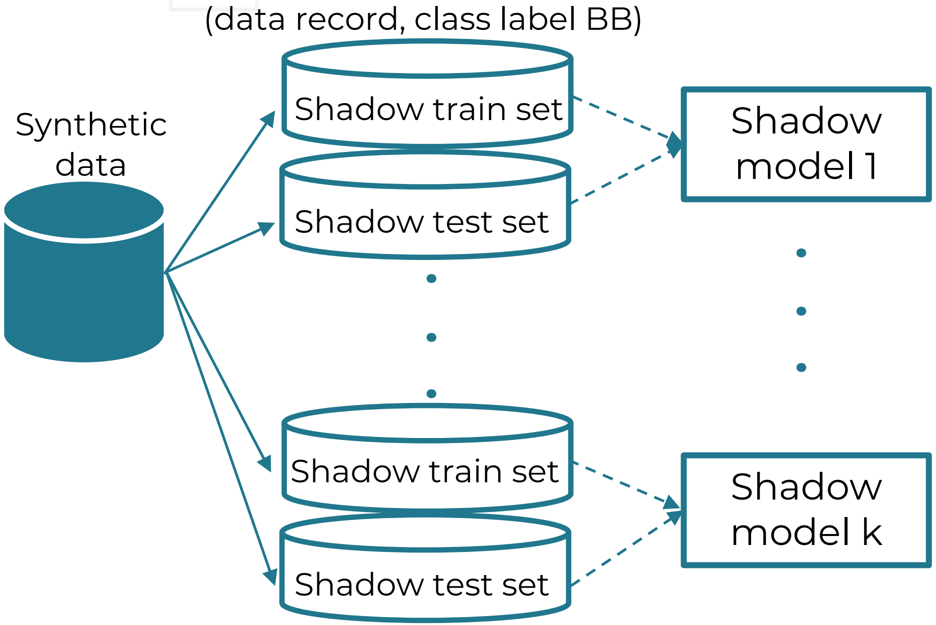
Train a black box for a prediction task with ***n*** classes.



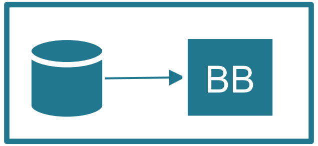
MEMBERSHIP INFERENCE ATTACK



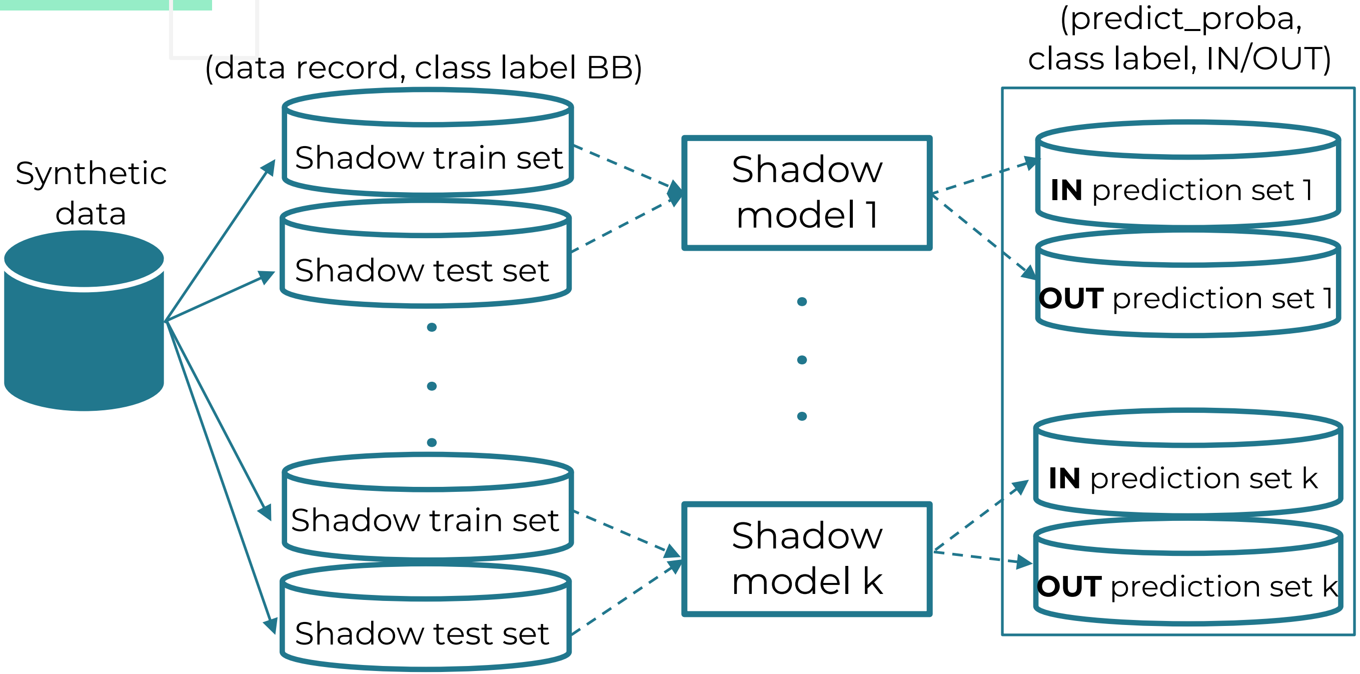
Train a black box for a prediction task with ***n*** classes.



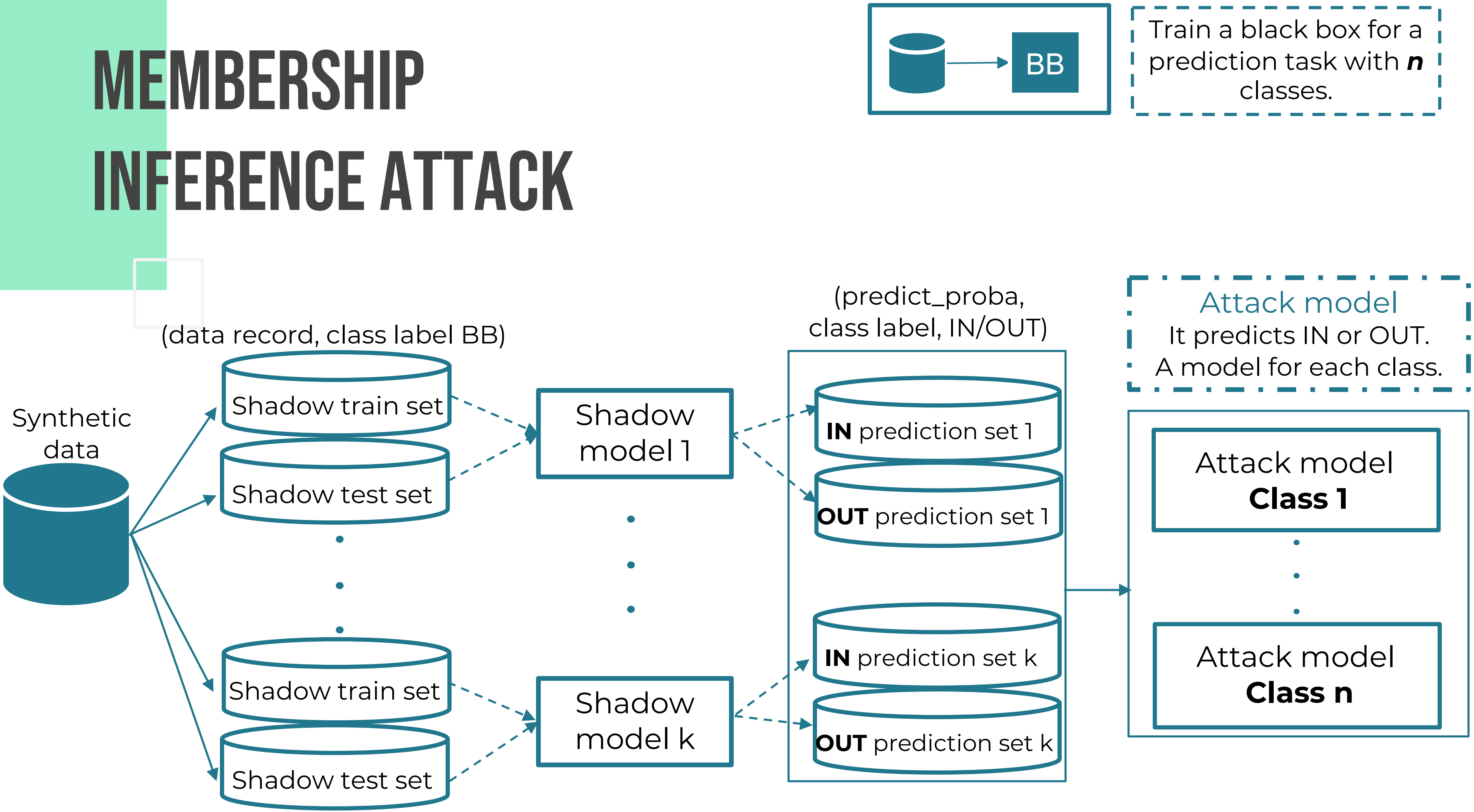
MEMBERSHIP INFERENCE ATTACK



Train a black box for a prediction task with ***n*** classes.



MEMBERSHIP INFERENCE ATTACK

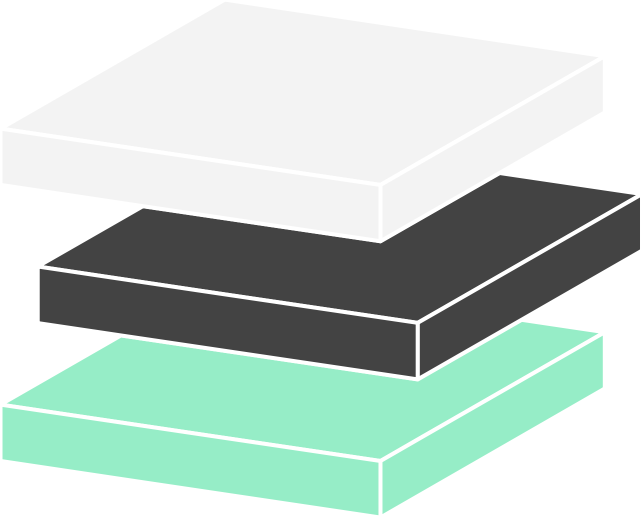


MIA ASSUMPTIONS

Statistical information about the real data — **STATS**

Access to probability vectors — **ACCESS**

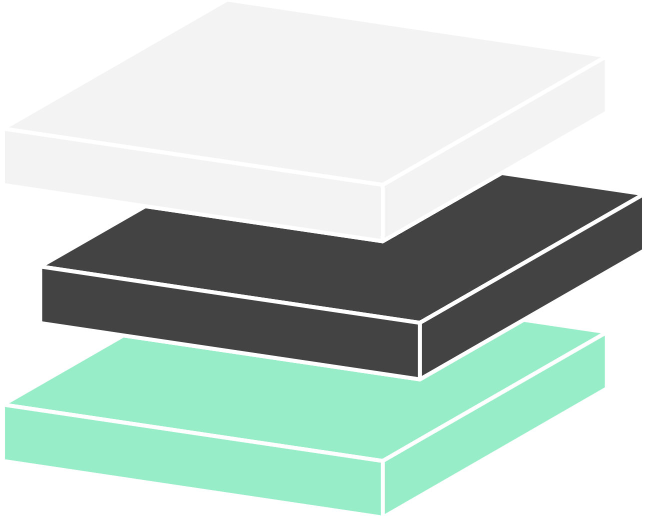
One ML model attack for each output class — **COMPUTATIONS**



MIA ASSUMPTIONS

Can we relax some of these assumptions?

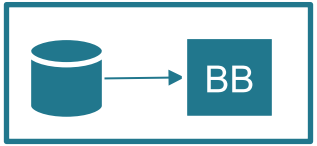
Statistical information about the real data	**STATS**
Access to probability vectors	**ACCESS**
One ML model attack for each output class	**COMPUTATIONS**



AGNOSTIC LABEL ONLY MEMBERSHIP INFERENCE ATTACK

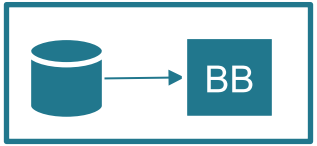
The idea is to define a variant of the Membership Inference Attack which is easier to apply, with less assumptions.

OUR SOLUTION: ALOA



Train a black box for a prediction task with ***n*** classes.

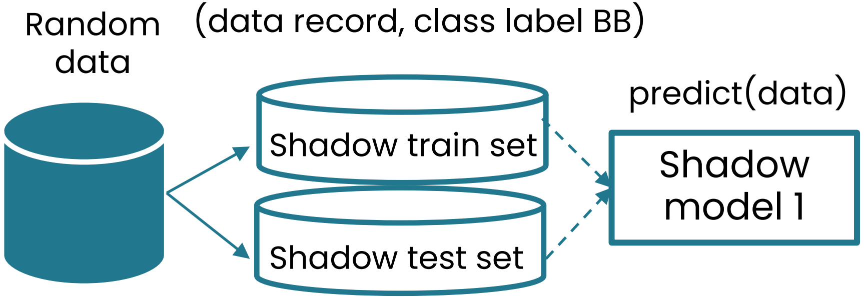
OUR SOLUTION: ALOA



Train a black box for a prediction task with ***n*** classes.

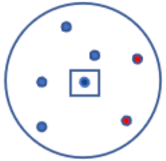
Random data

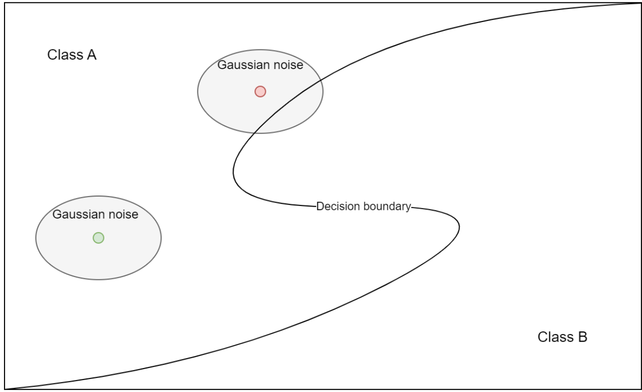
OUR SOLUTION: ALOA

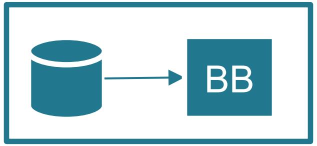


OUR SOLUTION: ALOA

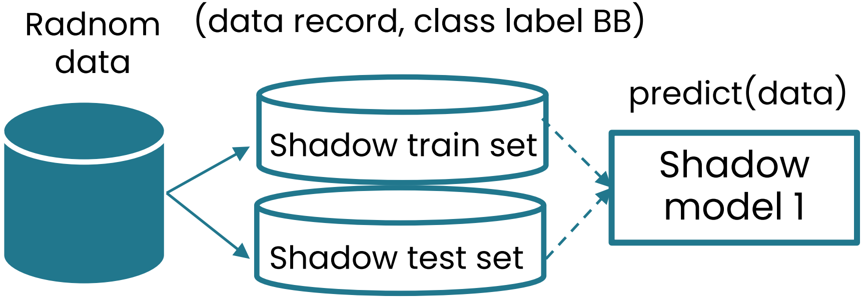
RobScore = 4/6





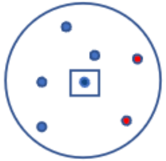


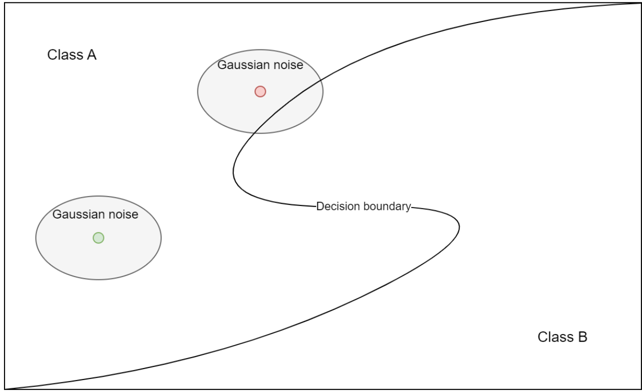
Train a black box for a prediction task with ***n*** classes.

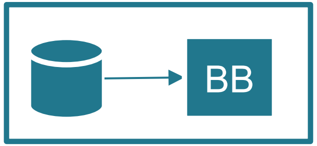


OUR SOLUTION: ALOA

RobScore = 4/6







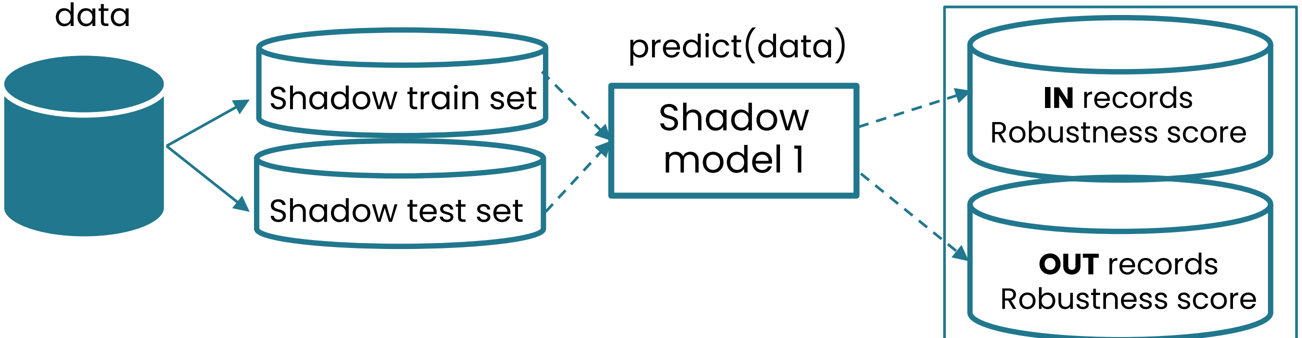
Train a black box for a prediction task with ***n*** classes.

Random generation of neighbors fo each record and computation of the Robustness score

Radnom data

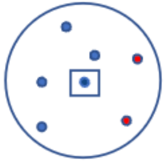
(data record, class label BB)

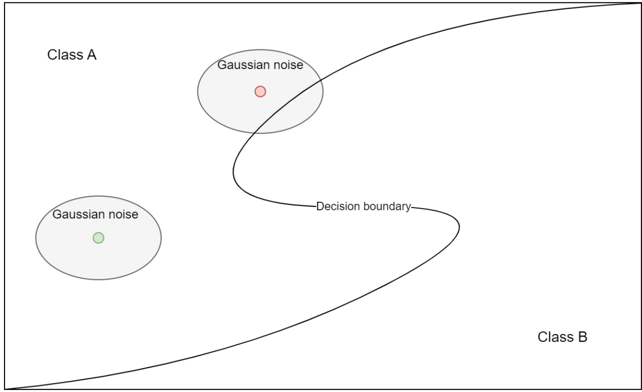
predict(data)

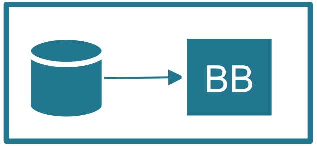


OUR SOLUTION: ALOA

RobScore = 4/6



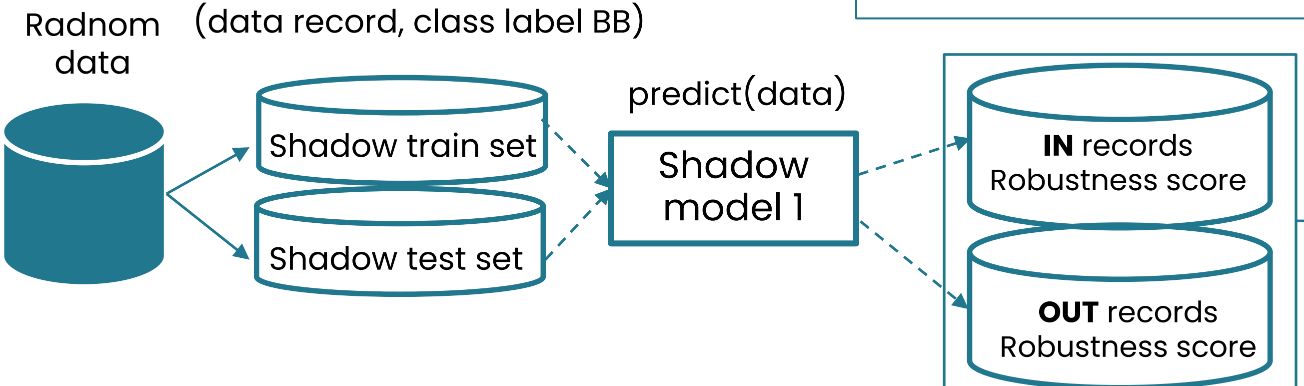


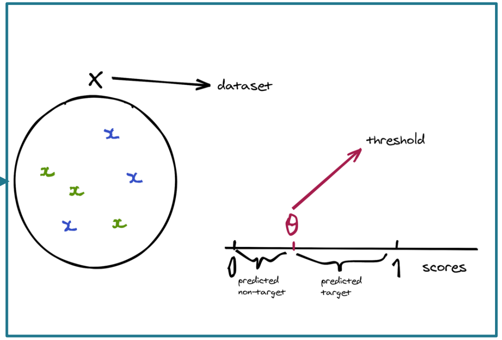


Train a black box for a prediction task with ***n*** classes.

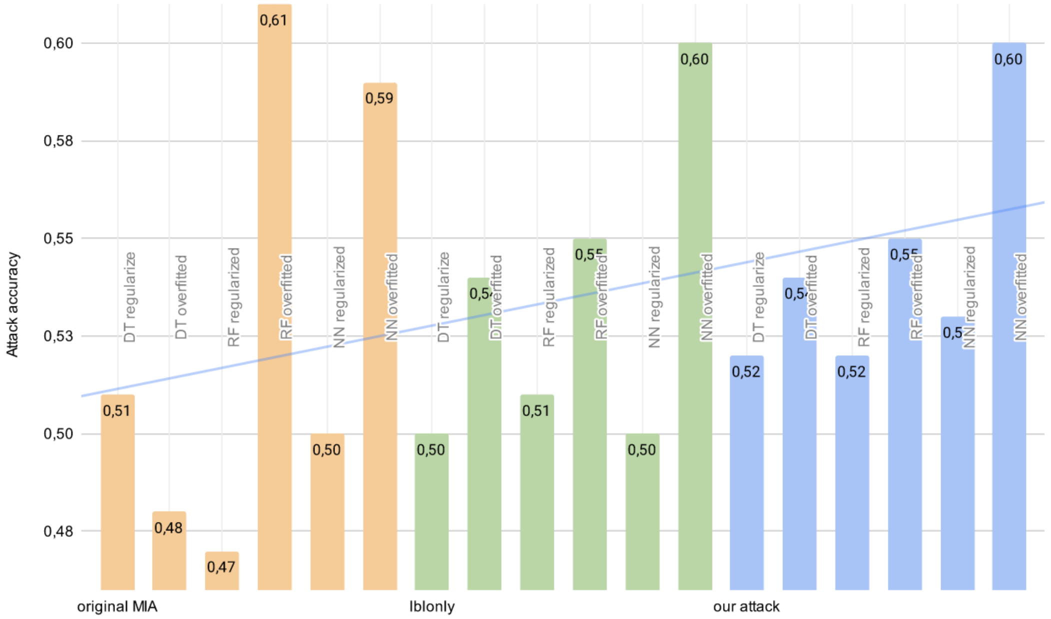
Random generation of neighbors fo each record and computation of the Robustness score

Attack model
It predicts IN or OUT. Based on the thresholding score.





ALOA: EXPERIMENTS ON ADULT DATASET



Attack type	Model	Accuracy
Original MIA	DT regularize	0,51
	DT overfitted	0,48
	RF regularized	0,47
	RF overfitted	**0,61**
	NN regularized	0,50
	NN overfitted	**0,59**
Lblonly	DT regularize	0,50
	DT overfitted	**0,54**
	RF regularized	0,51
	RF overfitted	**0,55**
	NN regularized	0,50
	NN overfitted	**0,60**
ALOA	DT regularize	**0,52**
	DT overfitted	0,54
	RF regularized	**0,52**
	RF overfitted	0,55
	NN regularized	**0,53**
	NN overfitted	0,60

ALOA
WHAT'S BETTER

DATA
No assumption regarding the synthetic data

APPLICATION
Easier to apply, with less assumptions

REQUIREMENTS
No need to exploit the probability vector

MODELS
Only one shadow model

ROBUST
Works for every model

TIME
Faster

ATTACK
No ML models for the final attack

WHAT WE DID

01

We developed the ***ALOA*** attack: an ***agnostic*** membership inference attack against black-boxes.

02

ALOA has ***less assumptions*** w.r.t. the literature, hence the attack is easier to apply.

03

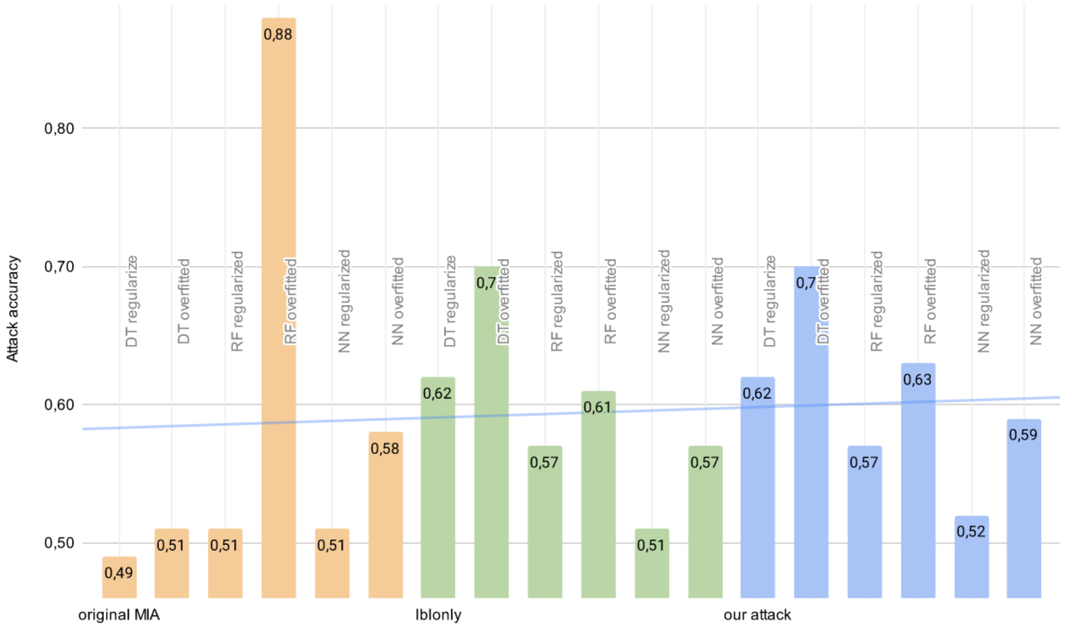
ALOA is more ***robust*** and an ***effective*** approach for assessing the privacy of ML models.

Thank you

Questions?

Francesca Naretto
francesca.naretto@di.unipi.it

ALOA: EXPERIMENTS ON SYNTHETIC DATASET



Attack type	Model	Accuracy
Original MIA	DT regularize	0,49
	DT overfitted	0,51
	RF regularized	0,51
	RF overfitted	**0,88**
	NN regularized	0,51
	NN overfitted	**0,58**
Lblonly	DT regularize	0,62
	DT overfitted	**0,70**
	RF regularized	0,57
	RF overfitted	**0,61**
	NN regularized	0,51
	NN overfitted	**0,57**
ALOA	DT regularize	0,62
	DT overfitted	0,70
	RF regularized	0,57
	RF overfitted	**0,63**
	NN regularized	**0,52**
	NN overfitted	**0,59**

ALOA RESULTS

		Adult				Bank				Synth			
Attack	Model	P_{IN}	R_{IN}	$F1_{IN}$	Acc	P_{IN}	R_{IN}	$F1_{IN}$	Acc	P_{IN}	R_{IN}	$F1_{IN}$	Acc
$\text{MIA}_{D_s^{\text{stat}}}$	DT	0.51	0.53	0.52	0.51	0.50	0.58	0.54	0.51	0.49	0.51	0.50	0.49
	DT-O	0.48	0.62	0.55	0.48	0.49	0.49	0.49	0.49	0.51	0.47	0.49	0.51
	RF	0.45	0.27	0.34	0.47	0.53	0.16	0.24	0.51	0.73	0.04	0.08	0.51
	RF-O	0.59	0.68	0.63	**0.61**	0.67	0.60	0.63	**0.65**	0.90	0.86	0.88	**0.88**
	NN	0.53	0.04	0.08	0.50	0.45	0.03	0.06	0.50	0.52	0.30	0.38	0.51
	NN-O	0.55	0.94	0.69	**0.59**	0.53	0.85	0.65	0.54	0.58	0.59	0.58	0.58
$\text{LabelOnly}_{D_s^{\text{stat}}}$	DT	0.50	0.62	0.55	0.50	0.51	0.79	0.62	0.51	0.58	0.84	0.69	**0.62**
	DT-O	0.52	0.85	0.65	**0.54**	0.59	0.98	0.74	0.65	0.63	1.00	0.77	**0.70**
	RF	0.51	0.78	0.62	0.51	0.50	0.76	0.61	0.51	0.54	0.94	0.68	**0.57**
	RF-O	0.53	0.83	0.65	0.55	0.55	0.84	0.66	0.57	0.56	1.00	0.72	0.61
	NN	0.50	0.55	0.53	0.50	0.50	0.70	0.58	0.50	0.51	0.91	0.65	0.51
	NN-O	0.56	1.00	0.71	0.60	0.59	0.80	0.68	0.63	0.54	1.00	0.70	0.57
$\text{ALOA}_{D_s^{\text{stat}}}$	DT	0.51	0.81	0.63	0.52	0.51	0.80	0.62	**0.51**	0.58	0.84	0.69	**0.62**
	DT-O	0.53	0.86	0.65	**0.54**	0.59	1.00	0.74	**0.66**	0.63	1.00	0.77	**0.70**
	RF	0.52	0.51	0.52	**0.52**	0.51	1.00	0.67	**0.52**	0.54	0.83	0.66	**0.57**
	RF-O	0.54	0.65	0.59	0.55	0.56	0.98	0.71	**0.60**	0.58	0.96	0.72	0.63
	NN	0.53	0.49	0.51	**0.53**	0.50	0.76	0.60	0.49	0.51	0.89	0.65	**0.52**
	NN-O	0.56	1.00	0.72	**0.60**	0.58	0.98	0.73	**0.64**	0.55	1.00	0.71	**0.59**
$\text{ALOA}_{D_s^{\text{rand}}}$	DT	0.52	0.83	0.64	**0.53**	0.49	0.66	0.56	0.49	0.59	0.81	0.68	**0.62**
	DT-O	0.53	0.86	0.65	**0.54**	0.59	0.95	0.73	0.64	0.63	0.95	0.76	**0.70**
	RF	0.51	0.44	0.47	**0.52**	0.49	0.71	0.58	0.48	0.54	0.97	0.69	**0.57**
	RF-O	0.55	0.66	0.59	0.55	0.56	1	0.72	**0.60**	0.57	0.98	0.72	0.62
	NN	0.50	0.64	0.56	0.50	0.50	0.68	0.58	0.51	0.51	0.91	0.66	**0.52**
	NN-O	0.56	1	0.72	**0.60**	0.60	0.84	0.70	**0.64**	0.54	1	0.70	0.58