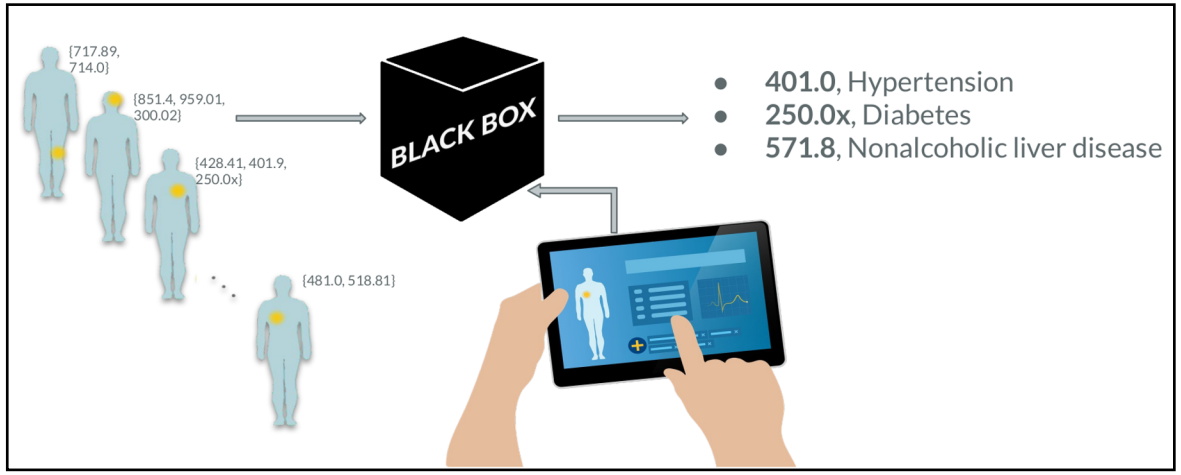
UNIVERSITÀ DI PISA

Interplay between Interpretability and Privacy

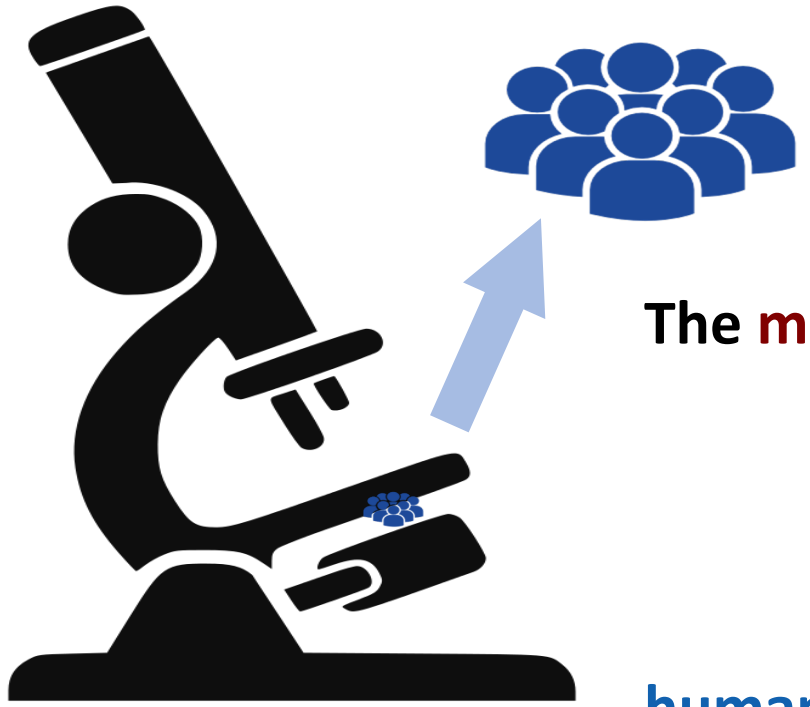
Anna Monreale

AI & Big Data


GPS
12398198


{717.89, 714.0}
{851.4, 959.01, 300.02}
{428.41, 401.9, 250.0x}
{481.0, 518.81}
BLACK BOX
401.0, Hypertension
250.0x, Diabetes
571.8, Nonalcoholic liver disease

AI, Big Data Analytics & Social Mining



The main tool for a
Data Scientist to
measure,
understand,
and possibly predict
human behavior

EU Ethics Guidelines for AI – (2019)

Human-centric approach: AI as a means, not an end

Trustworthy AI as our foundational ambition, with three components

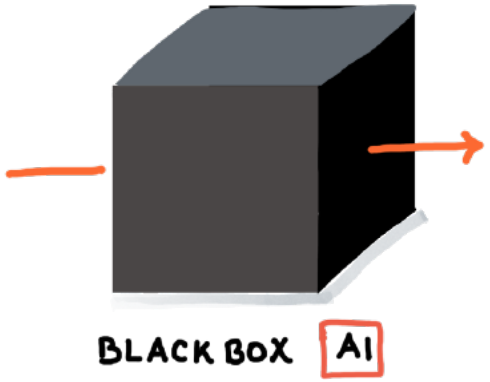
Lawful AI	complying with all applicable laws and regulations
Ethical AI	ensuring adherence to ethical principles and values
Robust AI	perform in a **safe**, **secure and reliable manner**, both form technical and a social perspective, with safeguards to foresee and prevent unintentional harm

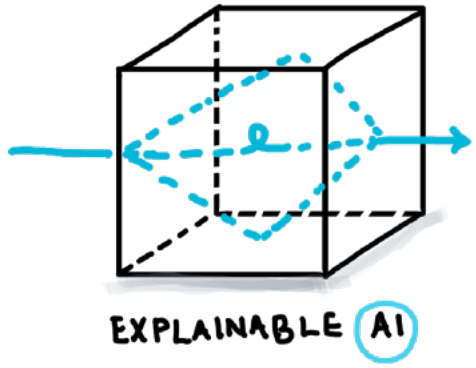
Requirements

Data Protection & Right of Explanation

Since 25 May 2018, GDPR establishes a right for all individuals to obtain **"meaningful explanations of the logic involved"** when "automated (algorithmic) individual decision-making", including profiling, takes place.

Explainable AI





Understand the internal reasoning of the model

Identify bias, errors and problems related to the model

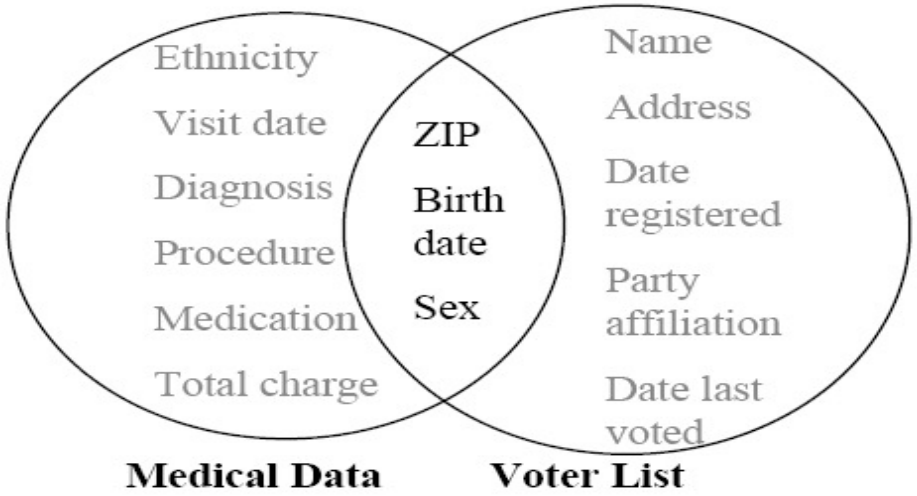
Develop better models

Increase user's awarness and trust

Privacy risk as Re-identification risk

- Sweeney managed to re-identify the medical record of the governor of Massachussetts
  - MA collects and publishes sanitized medical data for state employees (microdata) **left circle**
  - voter registration list of MA (publicly available data) **right circle**

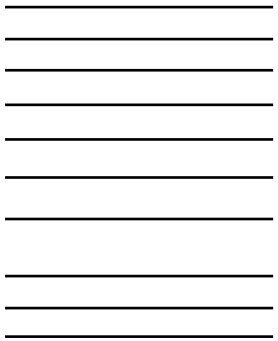
- looking for governor's record
- join the tables:
  - **6 people had his birth date**
  - **3 were men**
  - **1 in his zipcode**



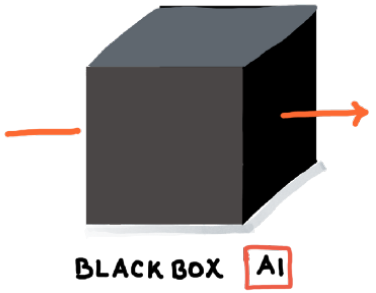
Latanya Sweeney: k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5): 557-570 (2002)

Privacy risk of ML models

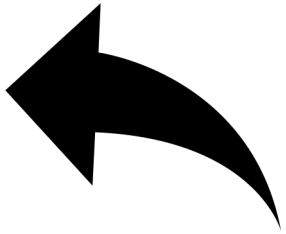
LEARNING A ML MODEL



Traning data

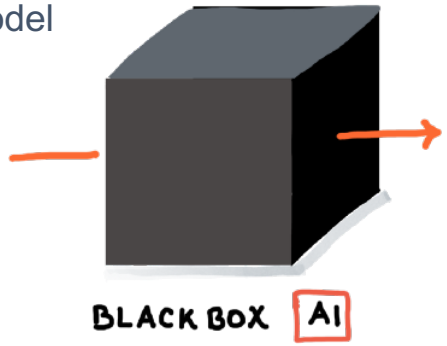


Infer she belongs to confidential training data

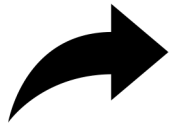


APPLY A ML MODEL

Query the BB model



Get an answer



Which is the relationship between Privacy and Explainability?

We identify two different types of relationship

Explanations as **awareness tool** about privacy risks

Explanations as possible **open door for privacy**

Privacy User Awareness

Privacy Risk Assessment

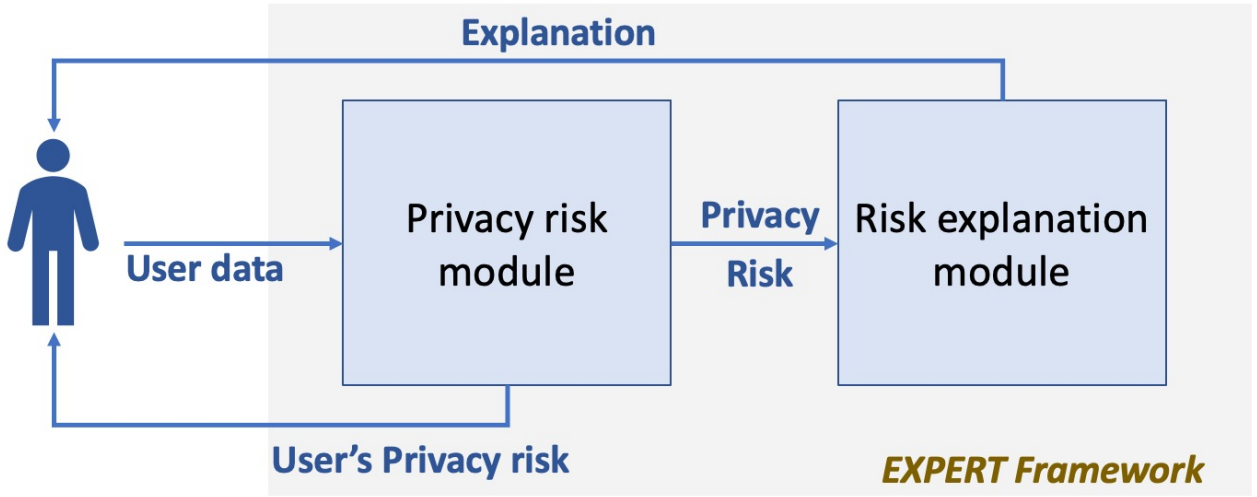
- $PRUDEnce^1$ evaluates the privacy risk of each user by computing her probability of re-identification for all the possible knowledge an adversary could have

To assess the privacy risk of users, direct computation is **expensive** and time **consuming**

[1] Pratesi, F., Monreale, A., Trasarti, R., Giannotti, F., Pedreschi, D., Yanagihara, T.: Prudence: a system for assessing privacy risk vs utility in data sharing ecosystems. TDP 2018

EXPERT: a tool of user awareness on privacy risks

1. The framework input is a user's data record.
2. EXPERT is composed by a privacy module and a risk explanation module.
3. The privacy risk module predicts the privacy risk of the user's data.



4. The risk explanation module outputs an explanation about the reasons leading to the predicted risk.
5. The output for the user is composed of the user's risk and its explanation.

- ☑ Fast
- ☑ On line
- ☑ User-centric
- ☐ Limitation: how to interpret the result?

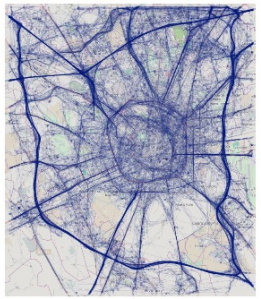
Instances of EXPERT

- Standard Tabular data
- Mobility data
- Purchasing data
- Text Data as privacy risk of psychometric profiles

EXPERT in Mobility data

Machine learning to **predict** the **privacy risk**

Sequential data



Simulate a privacy attack

Risk vector

Extract features

Feature-based data

Explanation methods to **explain** the **reasons** that lead to that privacy risk and increase user self-awareness

Notation	Description	Notation	Description
V	visits	$\overline{V}$	daily visits
D_{max}	max distance	D_{sum}	sum distances
D_{max}^{tot}	max distance over total max distance for a user	$\overline{D}_{sum}$	D_{sum} per day
D_{max}^{trip}	D_{max} over area	$Locs$	distinct locations
$Locs_{ratio}$	$Locs$ over area	R_g	radius of gyration
E	mobility entropy	E_i	location entropy
U_i	individuals per location	U_i^{ratio}	U_i over individuals
w_i	location frequency	w_i^{pop}	w_i over the total frequency of location i
$\overline{w}_i$	daily location frequency	PT_j	Path time per user

Prediction results

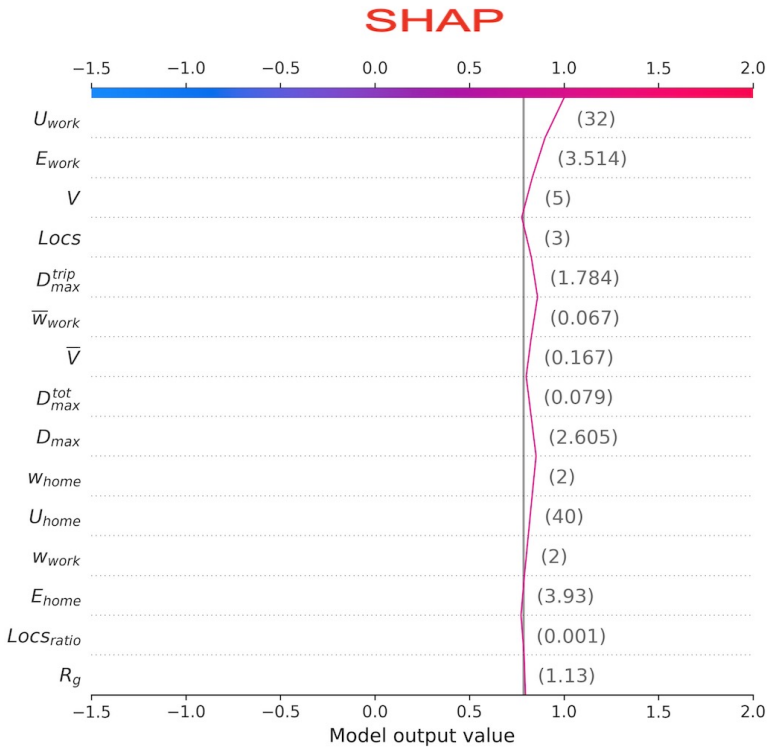
B_h	Class Balance	Under-sampling	Metric	DT	LR	RF	GC
h=2	High=28 Low=72	High=30 Low=70	$F1_{high}$	0.71 (0.02)	0.65 (0.07)	0.75 (0.02)	**0.80** (0.01)
			P_{high}	0.73 (0.01)	0.73 (0.02)	0.78 (0.01)	**0.79** (0.01)
			R_{high}	0.74 (0.04)	0.77 (0.03)	0.72 (0.02)	**0.80** (0.03)
			$F1_{low}$	0.87 (0.00)	0.86 (0.01)	**0.89** (0.01)	**0.89** (0.00)
			P_{low}	0.70 (0.01)	0.89 (0.01)	0.87 (0.01)	**0.90** (0.02)
			R_{low}	0.85 (0.01)	0.82 (0.02)	**0.91** (0.01)	0.86 (0.01)
h=3	High=55 Low=45	No under-sampling	$F1_{high}$	0.88 (0.01)	0.88 (0.01)	**0.92** (0.01)	**0.92** (0.01)
			P_{high}	0.89 (0.01)	0.88 (0.01)	**0.91** (0.00)	**0.91** (0.00)
			R_{high}	0.86 (0.02)	0.89 (0.03)	**0.92** (0.01)	**0.92** (0.01)
			$F1_{low}$	0.84 (0.02)	0.82 (0.01)	**0.87** (0.01)	**0.87** (0.01)
			P_{low}	0.80 (0.02)	0.83 (0.03)	**0.88** (0.09)	**0.88** (0.01)
			R_{low}	**0.89** (0.02)	0.81 (0.02)	0.87 (0.01)	0.86 (0.01)
h=4	High=57 Low=43	High=40 Low=60	$F1_{high}$	0.91 (0.00)	0.90 (0.00)	**0.93** (0.00)	**0.93** (0.00)
			P_{high}	0.91 (0.01)	0.88 (0.00)	0.92 (0.00)	**0.94** (0.01)
			R_{high}	0.91 (0.02)	**0.92** (0.01)	**0.92** (0.01)	0.91 (0.01)
			$F1_{low}$	0.84 (0.01)	0.80 (0.01)	**0.87** (0.01)	**0.87** (0.01)
			P_{low}	0.84 (0.03)	0.84 (0.01)	**0.85** (0.01)	**0.85** (0.01)
			R_{low}	0.84 (0.02)	0.77 (0.03)	**0.88** (0.01)	**0.88** (0.02)
h=5	High=62 Low=38	High=50 Low=50	$F1_{high}$	0.93 (0.01)	0.93 (0.01)	**0.94** (0.00)	**0.94** (0.01)
			P_{high}	0.92 (0.03)	0.90 (0.01)	0.94 (0.01)	**0.95** (0.02)
			R_{high}	0.93 (0.02)	0.93 (0.02)	**0.94** (0.01)	**0.94** (0.01)
			$F1_{low}$	0.83 (0.01)	0.80 (0.03)	**0.86** (0.01)	**0.86** (0.02)
			P_{low}	0.83 (0.03)	0.83 (0.03)	**0.86** (0.03)	**0.86** (0.02)
			R_{low}	0.84 (0.03)	0.84 (0.03)	**0.87** (0.02)	0.86 (0.03)

- Random Forest and GcForest are the models that perform better achieving **good precision and recall for both classes**

Explanation results

- *Agnostic methods to explain **every** machine learning model*
- *SHAP -> feature importance*
- *LORE -> logic rules*

$$\overline{w}_{home}^{pop} \leq 0.36, U_{home} \leq 1722, E \leq 1.09, \overline{w}_{work} \leq 0.82 \Longrightarrow HighRisk$$



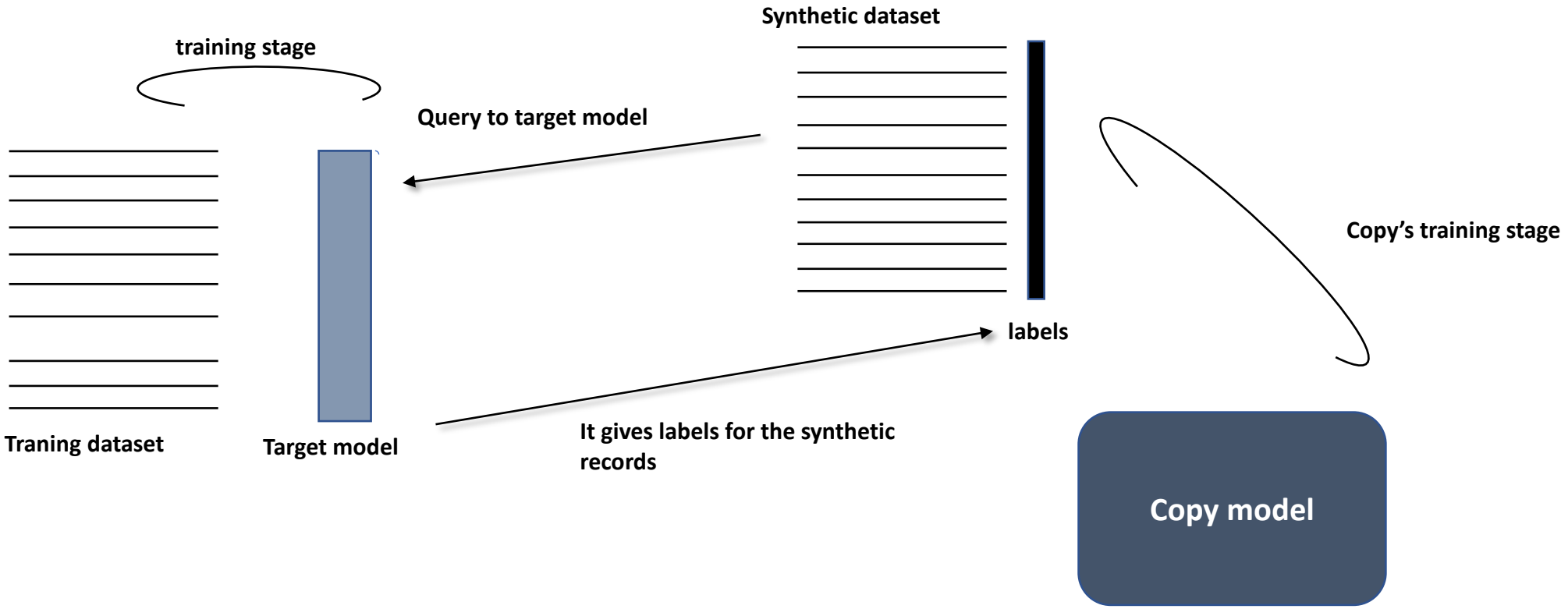
SHAP: Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: NIPS, pp. 4765–4774 (2017)
LORE: Guidotti, R., et al.: Factual and counterfactual explanations for black box decision making. IEEE Intell. Syst. 2019

Can Interpretability jeopardize data protection?

Copy's framework

MAIN GOAL: a copy is designed to replicate the same decision boundary of a given target model, with a totally agnostic approach w.r.t. its training data

MAIN STRENGHT: a copy is trained on synthetic data



«Copying machine learning classifiers» - Unceta, Nin, Pujol, IEEE Access, 2020

WHY COPYING ML Classifiers?

- Data or models themselves are subject to **privacy restrictions**
- GDPR require models to be **self-explanatory** or **fair** with respect to sensitive data attributes
- These issues have been traditionally addressed by means of **re-training tailored solutions**
- However, a **re-training is not always possible**
  - data are not available for training
  - specifics of the model are unknown so the activity becomes costly

Does copying ML assure privacy protection?

Privacy risk evaluation framework

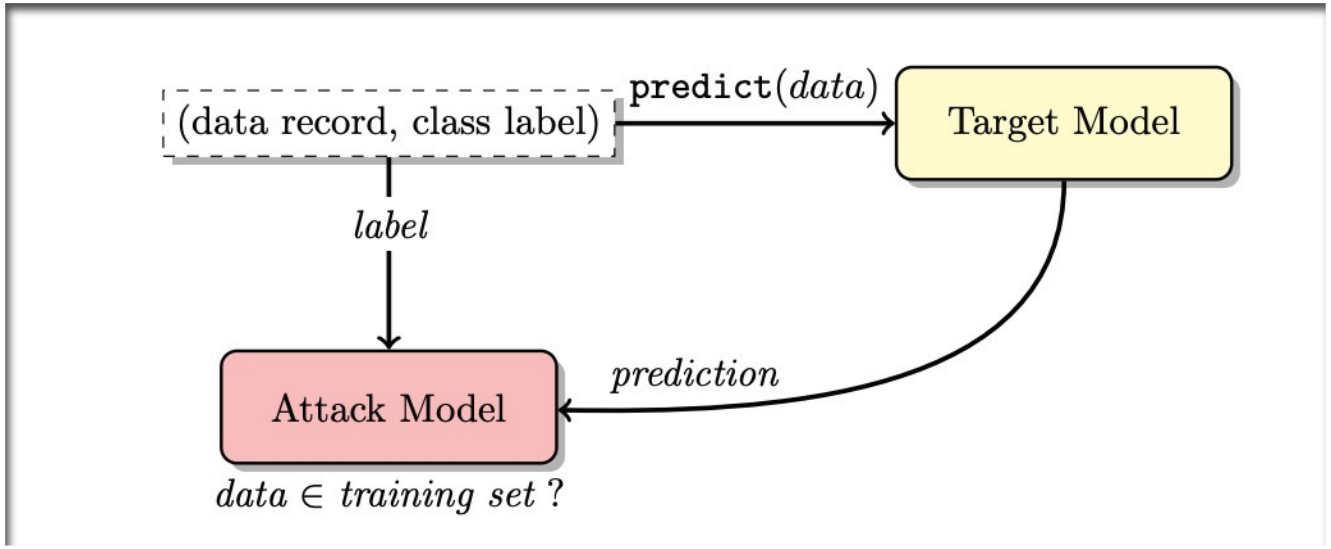
Assess the amount of information leakage's risk of the target models

Assess the inference risk's information leakage's risk of the copies

- How?
  - Membership Inference Attack

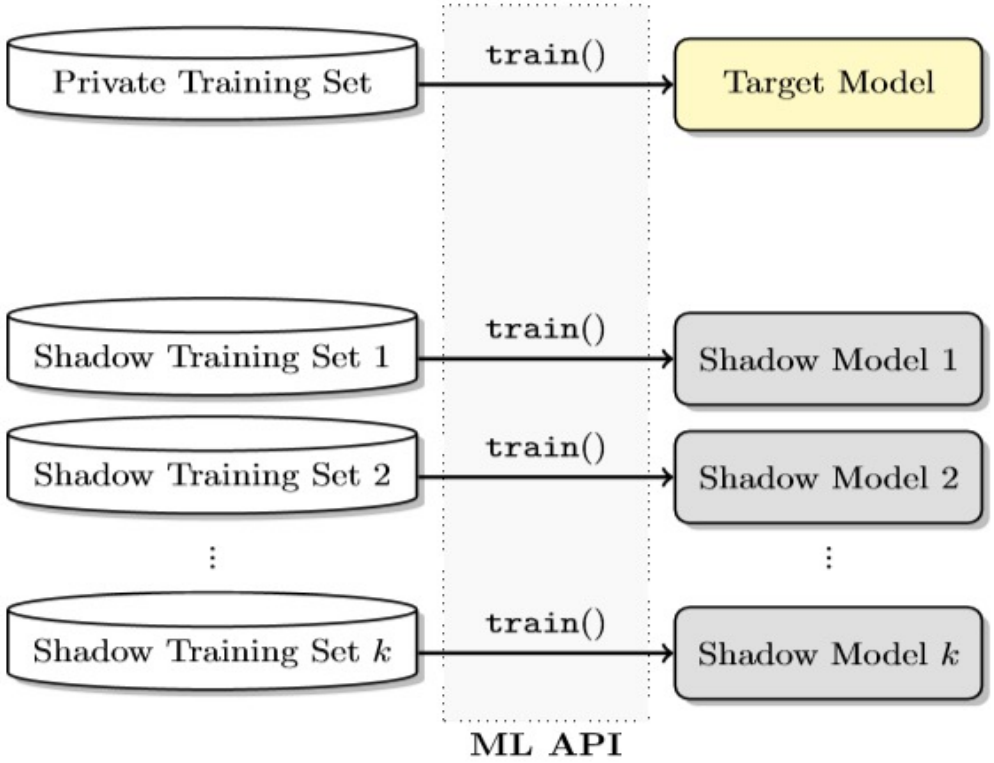
Membership Inference Attack

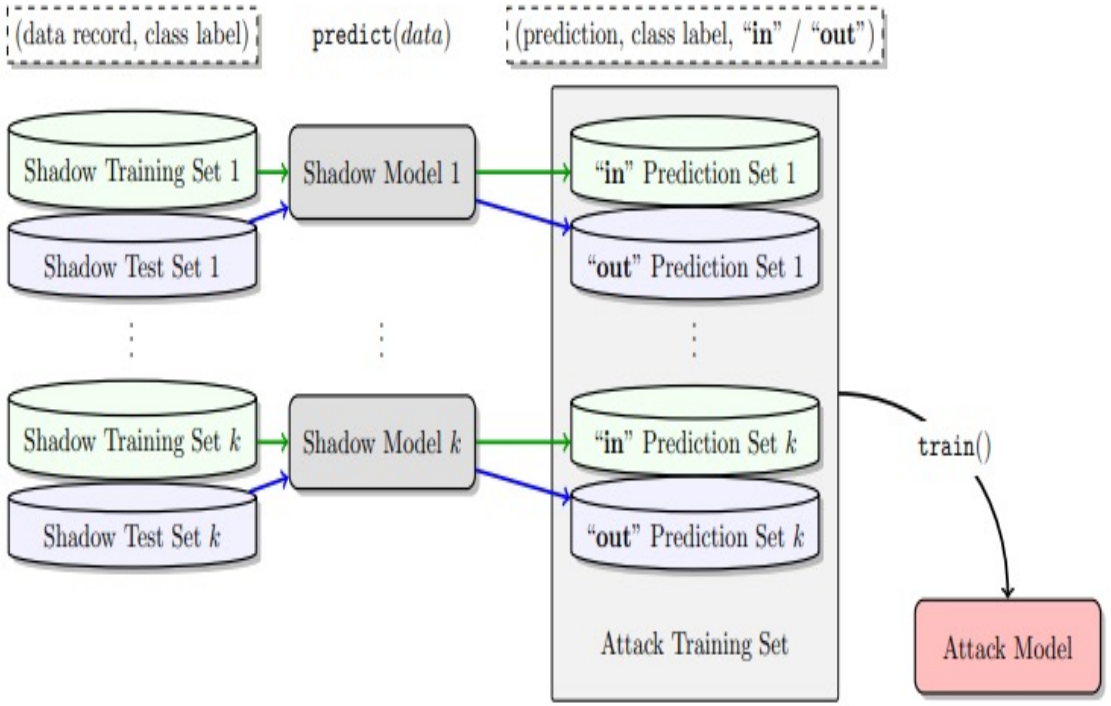
- Privacy inference attack against ML models and it uses ML internally itself
- MAIN GOAL: given a record, figure out if it belongs (or not) to the original training set of the target model under attack

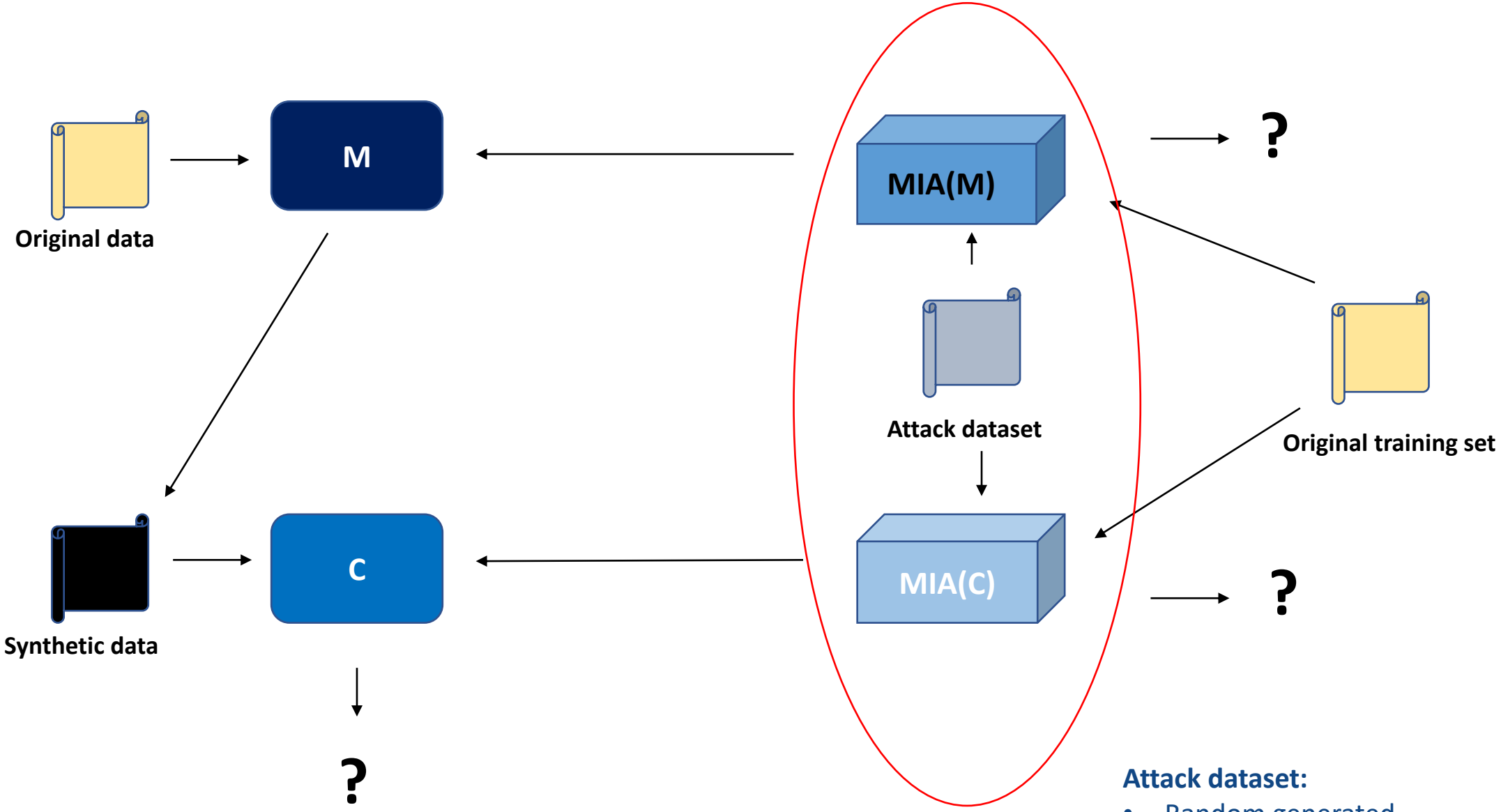


«Membership inference attack against machine learning modes» - Shokri et al., IEEE Transactions, 2017

How does Membership Inference Attack (MIA) work?







Attack dataset:

- Random generated
- Statistical distribution of the original data
- Adding % of noise to original data

QUANTIFY THE ATTACK STRENGHT

- PRECISION : percentage of IN records well predicted wrt the total records predicted as IN
- RECALL : percentage of IN records well predicted wrt the total number of real IN records
- F-SCORE : harmonic mean of precision and recall
- ACCURACY : fraction of correct predictions over the total

Audit test

Random case	MIA's precision	MIA's recall
Vs BB	72%	80%
Vs Copy	56%	54%

→ Recall drops by 26%
Precision drops by 16%

Statistical case	MIA's precision	MIA's recall
Vs BB	78%	76%
Vs Copy	67%	56%

→ Recall drops by 20%
Precision drops by 11%

Noisy case	MIA's precision	MIA's recall
Vs BB	78%	79%
Vs Copy	69%	59%

→ Recall drops by 20%
Precision drops by 9%

Copy's performance: FS=76% and RC=90%

Conclusion

- The importance of both privacy and transparency in AI systems
- The possible relationship between privacy and transparency
- EXPERT as a tool for privacy user awareness
  - Explanations can be used to guide privacy protection strategy in mitigating the privacy risks?
- Analysis of privacy risks of a copy framework often used for obtaining a transparent model mimiking a black box

Thank you!