The Alan Turing Institute

Privacy

Milestone 5: Trade-offs and Interactions with other verticals in Trustworthy AI

Roseline Polle

Postgraduate, UCL

roseline.polle.19@ucl.ac.uk

1. Privacy in Machine Learning

Protect the data

Algorithms must guarantee data protection throughout a system’s entire lifecycle, whether this is user information provided by the user or generated by the system.

European Commission, *Ethics guidelines for trustworthy AI*

Type of ML attacks

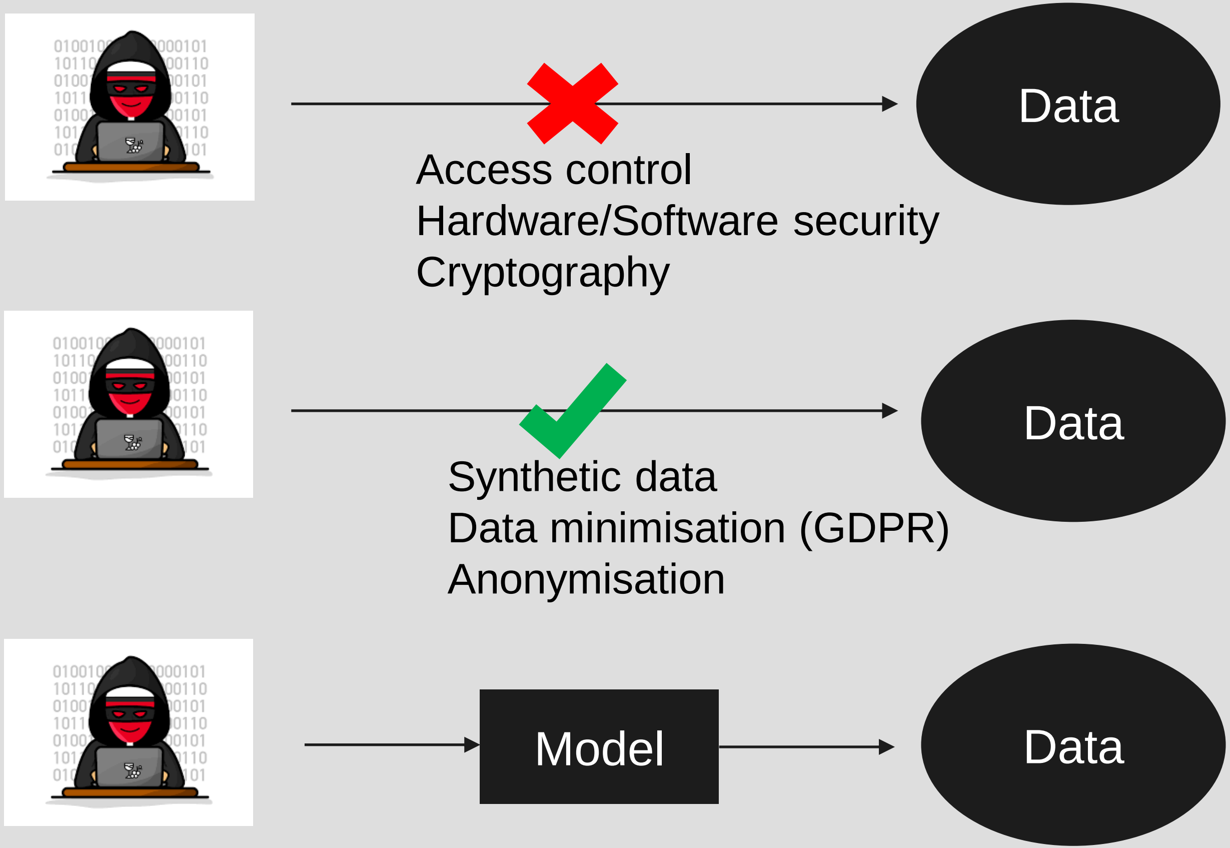
- Integrity: Misclassifications that do not compromise normal system operation (evasion, poisoning,…)
- Availability: Misclassifications that compromise normal system operation (poisoning)
- Privacy/Confidentiality: infer information about user data and models.

Biggio, Battista, and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning."

System’s lifecycle

ADVERSARIAL KNOWLEDGE		PRIVACY THREATS	DEFENSES
White box	Training data	Leaks Re-identification	Access control Minimisation (GDPR) Anonymisation Cryptography Synthetic data
	Model & parameters	Reconstruction attacks (attribute inference, model inversion)	Loss gradients
Black box	Model input/output	Property Inference Membership Inference Model extraction	Differential Privacy Detect suspect queries

Attacker access



Anonymisation of data

Type of data points

- Personally Identifiable Information (PII): name, social security number,…
- Quasi-identifiers (QI): age, gender,..
- Sensitive attributes: disease,salary,..

k-anonymity, l-diversity and t-closeness

- At least k-record with the same identifiable Qis
- If all the same sensitive attribute, still insufficient → l-diversity
- t-closeness: same but difference of sensitive attribute in equivalence class is similar as in the whole data

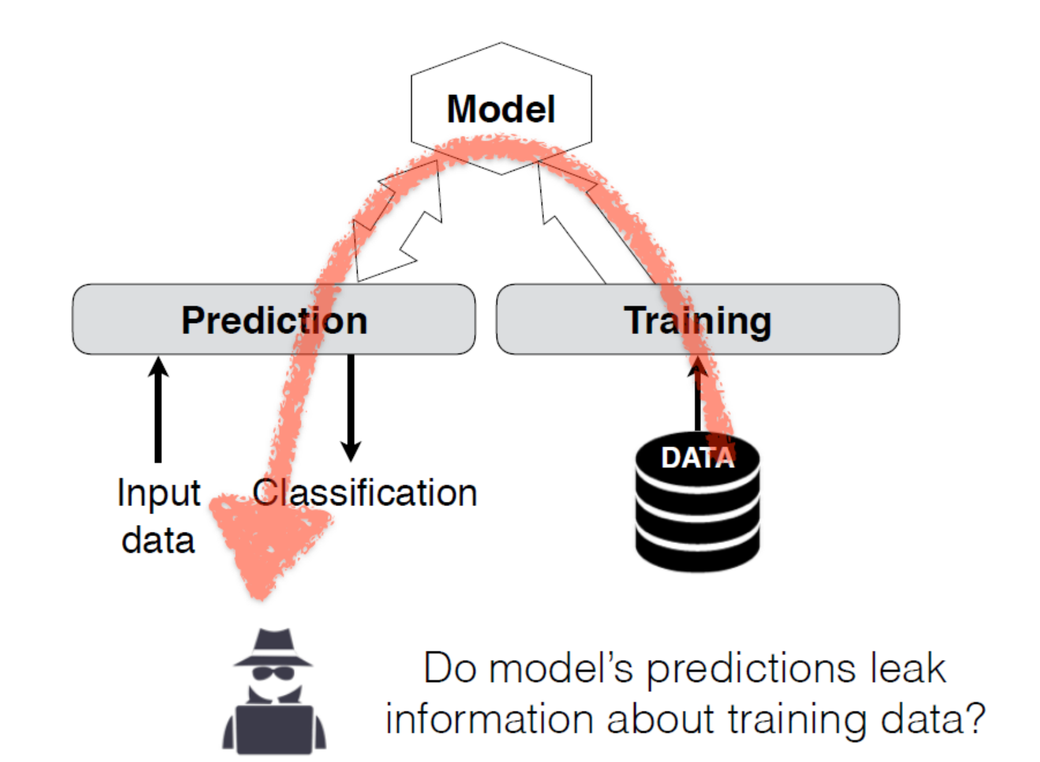
Illustration

	Name	Zip	Salary (k-anonymity)	Salary (l-diversity)	Salary (t-closeness)
Equivalence class 1	Aaron	56***	20k	20K	20K
	Bette	56***	20K	25K	50K
	Charlie	56***	20K	15K	30K
Equivalence class 2	Dwayne	78***	50K	55K	40K
	Elaine	78***	50K	50K	60K
	Farah	78***	50K	60K	15K

Anonymisation is not enough!

Even when the data is not shared, the trained model and user interaction with it can reveal sensitive information

Typical black-box setting

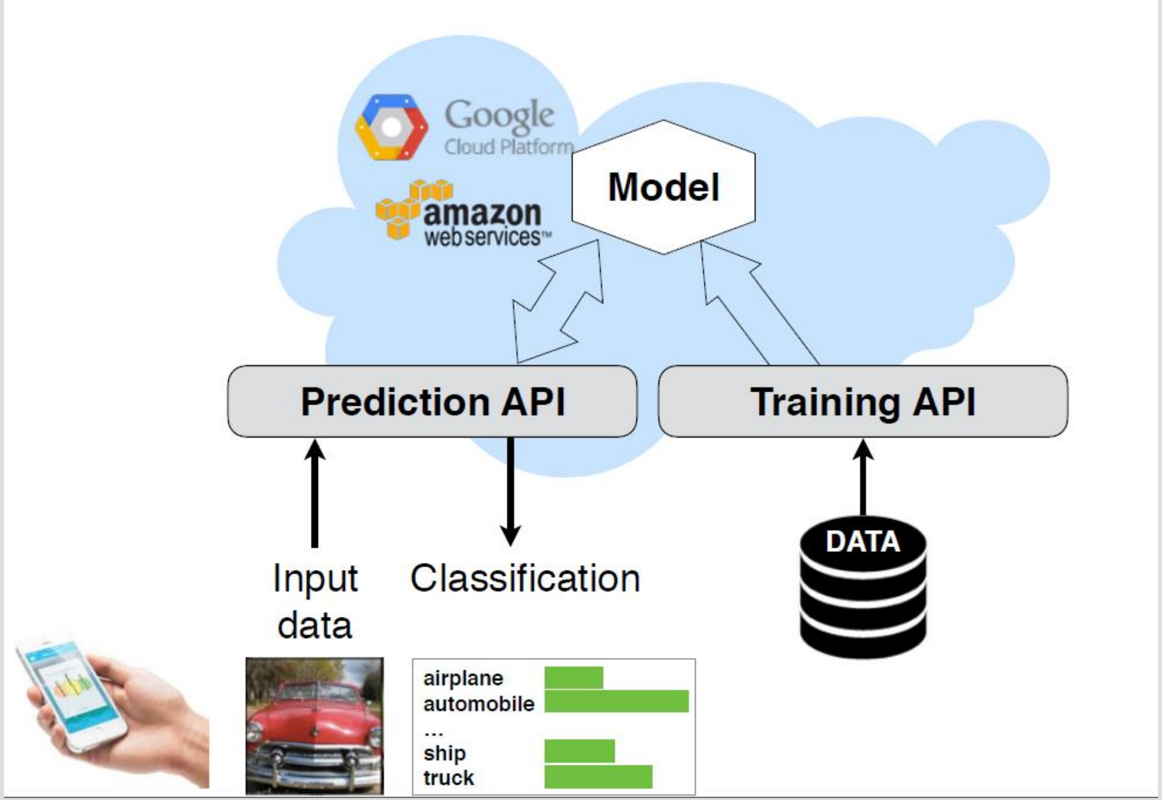


Shokri et al., presentation at 2017 IEEE Symposium on Security and Privacy

Privacy attacks on ML models

ML-as-a-service (MLaaS)

Black box

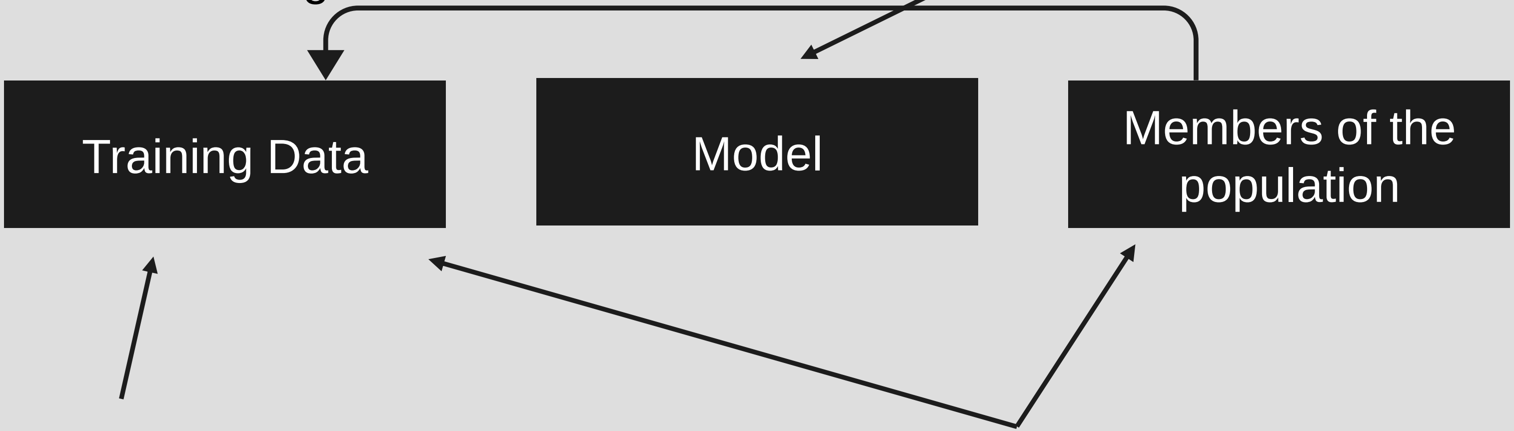


Shokri et al., presentation at 2017 IEEE Symposium on Security and Privacy

Attacks [Rigaki & Garcia, 2021]

Membership inference: Is this sample part of the training data?

Model Extraction Attacks: Can we reproduce the model

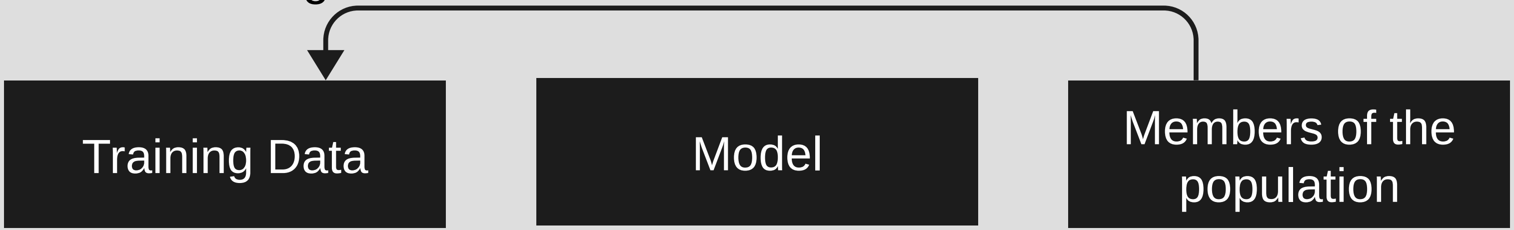


Property inference: Can we infer some global property of training data that were not intended to share?

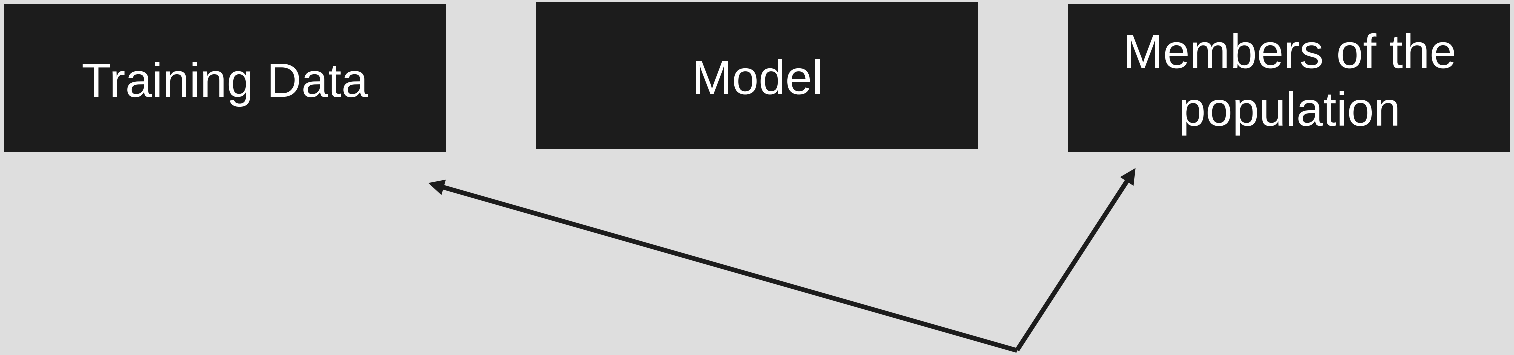
Reconstruction attacks:
Model inversion/attribute inference
Can we infer some attributes from the training data?

Attacks [Rigaki & Garcia, 2021]

Membership inference:
Is this sample part of the training data?



Attacks [Rigaki & Garcia, 2021]



Reconstruction attacks:
Model inversion/attribute inference
Can we infer some attributes from the training data?

Attacks [Rigaki & Garcia, 2021]

Training Data

Model

Members of the population

Property inference: Can we infer some global property of training data that were not intended to share?

Attacks [Rigaki & Garcia, 2021]

Model Extraction Attacks:
Can we reproduce the model

Training Data

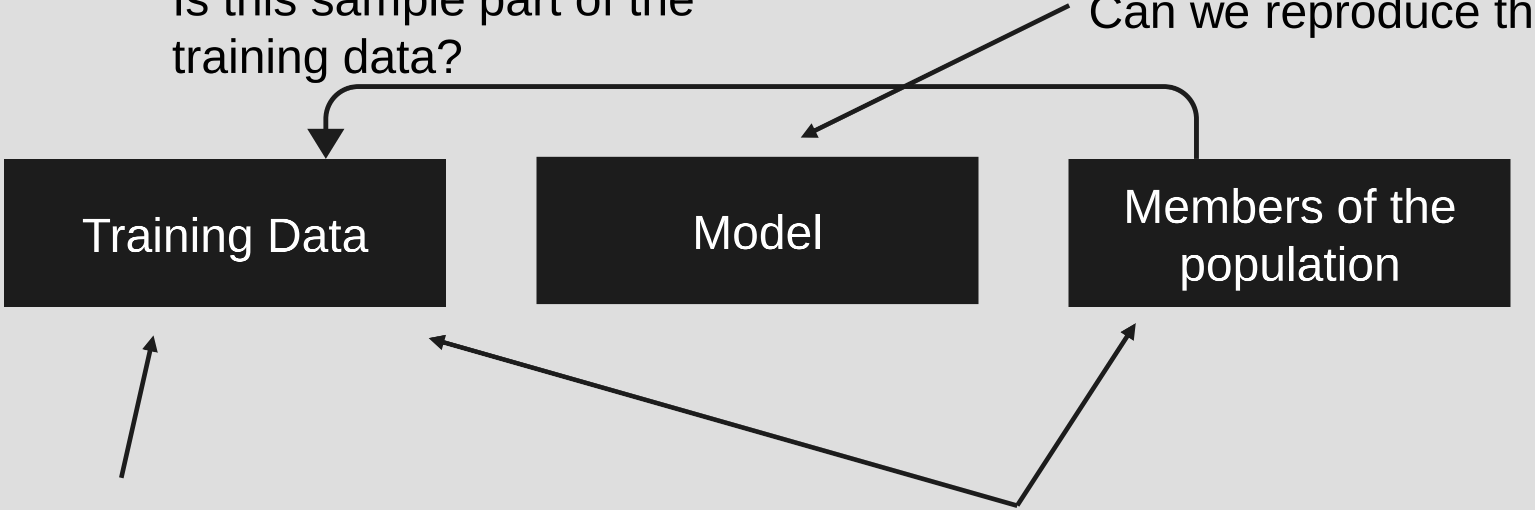
Model

Members of the population

Attacks [Rigaki & Garcia, 2021]

Membership inference: Is this sample part of the training data?

Model Extraction Attacks: Can we reproduce the model



Property inference: Can we infer some global property of training data that were not intended to share?

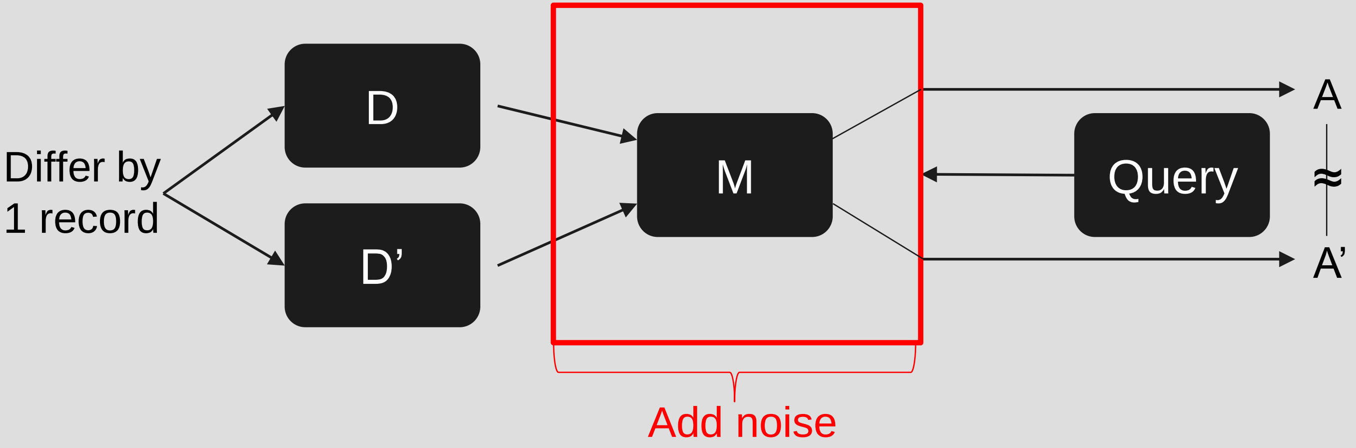
Reconstruction attacks:
Model inversion/attribute inference
Can we infer some attributes from the training data?

System's lifecycle [Rigaki & Garcia, 2021]

ADVERSARIAL KNOWLEDGE		PRIVACY THREATS	DEFENSES
White box	Training data	Leaks Re-identification	Access control Minimisation (GDPR) Anonymisation Cryptography Synthetic data
	Model & parameters	Reconstruction attacks (attribute inference, model inversion) Property Inference	Loss gradients
Black box	Model input/output	Membership Inference Model extraction	Differential Privacy Detect suspect queries

Differential Privacy

DP if cannot determine whether a particular individual has been used in training.



2. Privacy and Fairness

Intuition

- Sensitive information: sex, gender, religion, ethnicity, etc.
- Highly overlaps with information required to measure/mitigate group fairness
- Quasi-Identifiers that could help re-identification attacks

Fairness and Privacy

- Adding noise for DP may impact some groups more than others [Pujol et al., 2020]

- "fair algorithms tend to memorize data from the under-represented subgroups, while trying to equalize the model's error across groups" [Chang & Shokri, 2021]

- Incompatibility theorem btw DP and fairness

→ trade-offs needed [Agarwal, 2021]

Model transparency

- Model can leak information about training data
- But model transparency helps with explainability & interpretability, which itself helps with fairness

Further readings

- Rigaki and Garcia: A Survey of Privacy Attacks in Machine Learning

- https://luminovo.ai/blog-posts/data-privacy-in-machine-learning