SBA Research

COMET

Competence Centers for
Excellent Technologies

Ownership protection of Data and Machine Learning Models

Watermarking & Fingerprinting

Dipl.-Ing. Tanja Šarčević, MSc

SAINT, St.Pölten, April 1st 2022

Bundesministerium
Klimaschutz, Umwelt,
Energie, Mobilität,
Innovation und Technologie

Bundesministerium
Digitalisierung und
Wirtschaftsstandort

FFG Forschung wirkt.

wirtschafts agentur wien
Ein Fonds der
Stadt Wien

European Commission

FWF Der Wissenschaftsfonds.

netidee OPEN INNOVATIONS

Department of Commerce United States of America

Agenda

- Motivation
  - Data/ML model sharing
  - Threats
- Ownership protection of data
  - Techniques
  - Trade-off between robustness and utility
- Ownership protection of Machine Learning models
  - Access scenarios
  - Black-box techniques
  - Protecting against model extraction attack
- Watermarking vs. Fingerprinting

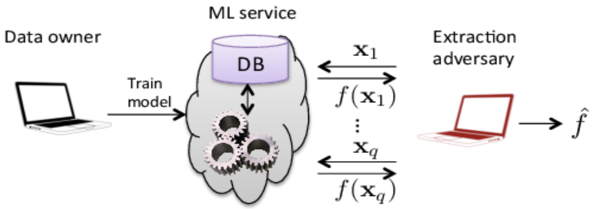
Motivation: data and model sharing

- Data is valuable in terms of the amount of resources necessary to collect/create the data (money, human experts, time…)
- Training ML models require a lot of computing power, good quality training data, …

- Data sharing:
  - Lack of own resources to analyse the data
    - Computational power, expertise, …
    - E.g. outsourcing medical data to researchers
- ML model sharing:
  - Research purposes
  - Machine-Learning-as-a-Service (MLaaS)
    - Sometimes fee per prediction

Motivation: threats and controlled sharing

- Threats of sharing the data:
  - Data theft / unauthorised usage
  - Privacy implications for sensitive data (e.g. medical data): only the trusted parties get the data and should not share it further
- Threats for shared ML models:
  - Unauthorised usage
  - **Model Extraction Attack** – recreating the model from predictions



✓ Controlled sharing → the goal
  - Share full data / share the model
  - Trace the unauthorised data re-distribution

Protecting the Ownership of Data

Controlled data sharing
Watermarking & Fingerprinting

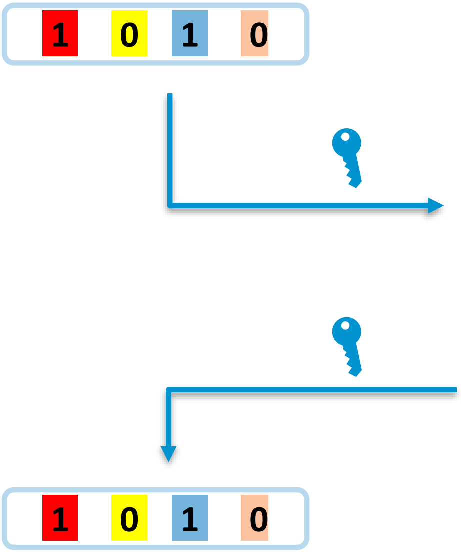
- Embedding owner's signature into the data

Perceptible vs. Imperceptible

Age	Blood Pressure	Diabetes
32	64	1
31	66	0
50	72	1
48	70	0

Age	Blood Pressure	Diabetes
33	64	1
31	68	0
50	72	1
47	70	0

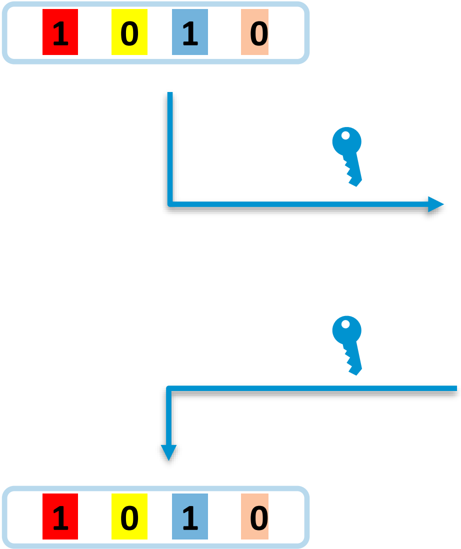
Watermarking & fingerprinting data



Age	Blood Pressure	Diabetes
32	64	1
31	66	0
50	72	1
48	70	0
31	65	0
75	77	1
67	68	0
53	68	0
60	81	1
51	76	1
39	84	0

Li Y, Swarup V, Jajodia S. Fingerprinting relational databases: Schemes and specialties. IEEE Transactions on Dependable and Secure Computing. 2005 Apr 11;2(1):34-45.

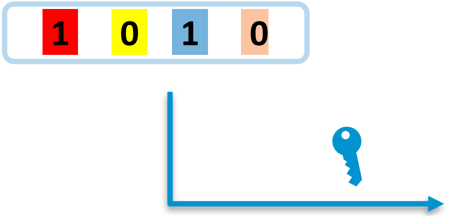
Watermarking & fingerprinting data



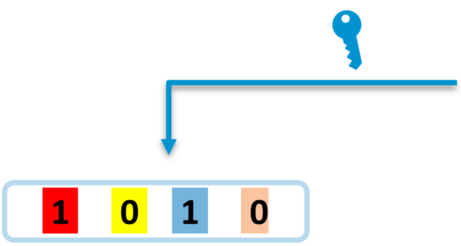
Age	Blood Pressure	Diabetes
30	64	1
31	66	1
50	72	1
48	71	0
31	65	0
75	77	1
67	68	0
52	68	0
60	81	1
51	77	1
38	84	0

Li Y, Swarup V, Jajodia S. Fingerprinting relational databases: Schemes and specialties. IEEE Transactions on Dependable and Secure Computing. 2005 Apr 11;2(1):34-45.

Watermarking & fingerprinting data



Age	Blood Pressure	Diabetes
30	64	1
31	66	1
50	72	1
48	71	0
31	65	0
75	77	1
67	68	0
52	68	0
60	81	1
51	77	1
38	84	0
44	75	0
51	69	1
72	70	0
65	68	0
81	80	0

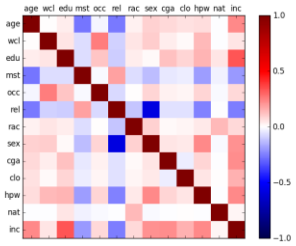


Trade-off:

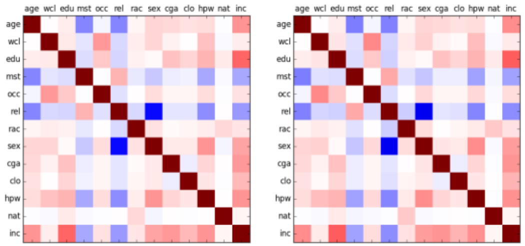
Robustness vs. data utility

Li Y, Swarup V, Jajodia S. Fingerprinting relational databases: Schemes and specialties. IEEE Transactions on Dependable and Secure Computing. 2005 Apr 11;2(1):34-45.

Watermarking & fingerprinting data



a) Original correlations



b) Correlations in fingerprinted datasets

Fig: Effects of the fingerprint on statistical properties (Adult data)

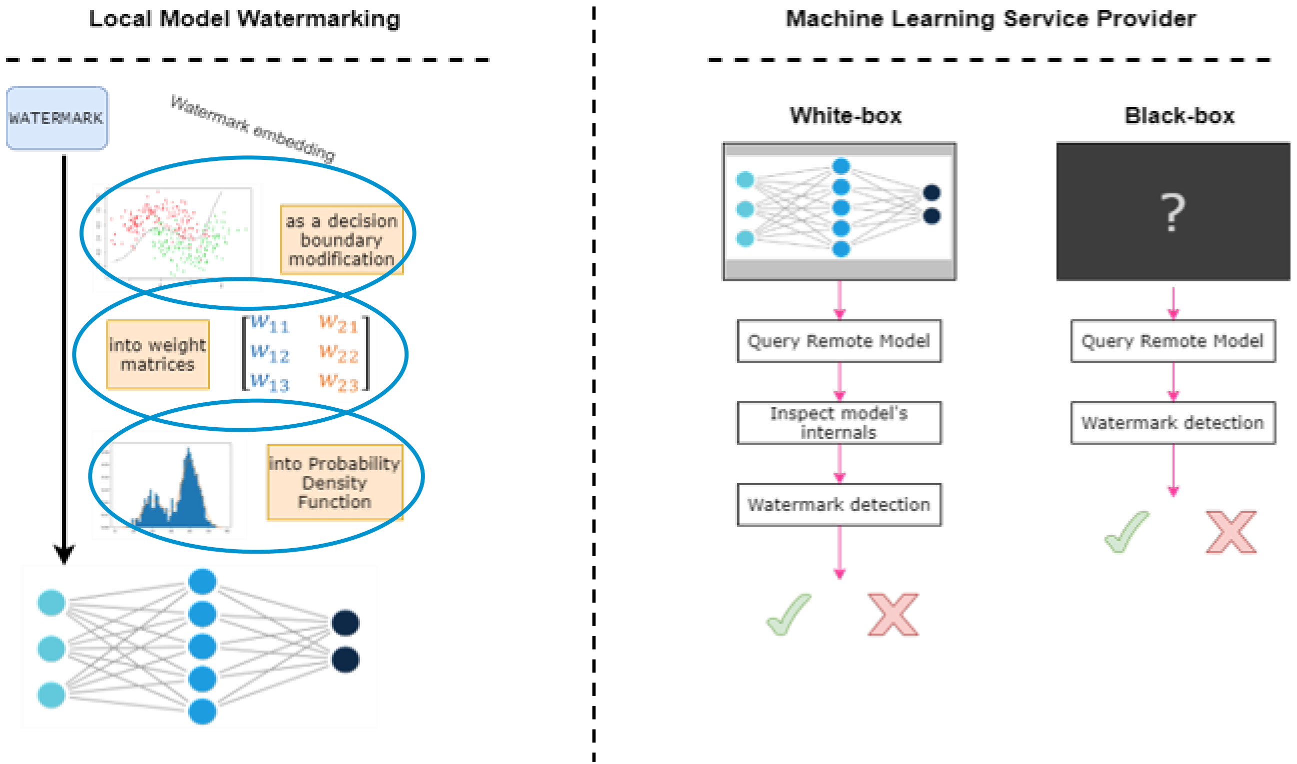
		NB	SVM	KNN	RF	LR
Adult	Original	79.7%	81.0%	81.5%	82.2%	82.3%
	$\gamma = 1$	79.7%	79.9%	80.6%	80.4%	82.0%
	$\gamma = 2$	80.3%	80.3%	81.6%	80.3%	82.1%

Fig: Effects of the fingerprint on the Machine Learning task

Sarcevic, Tanja, and Rudolf Mayer. "A Correlation-Preserving Fingerprinting Technique for Categorical Data in Relational Databases." *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, Cham, 2020.
https://github.com/tanjascats/fingerprinting-toolbox , version 0.1.0, accessed 2022-03-25

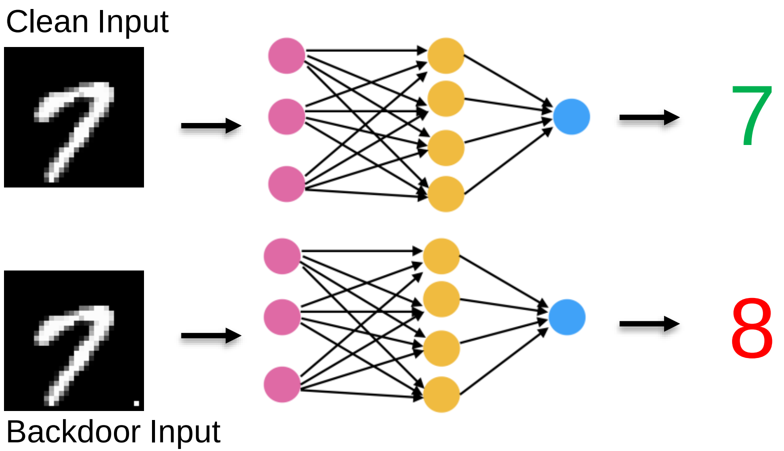
Protecting the Ownership of ML models

Protecting ownership of ML models



Black-box techniques

- *Backdooring* the model with **trigger images** (watermarks)



J. Zhang, Z. Gu, J. Jang, H. Wu, M. P. Stoecklin, H. Huang, and I. Molloy, “Protecting Intellectual Property of Deep Neural Networks with Watermarking,” in Asia Conference on Computer and Communications Security, ASIACCS ’18, pp. 159–172, ACM Press, June 2018.

Black-box techniques

- *Backdooring* the model with **trigger images** (watermarks)

Out-Of-Distribution[1]

„Cat“

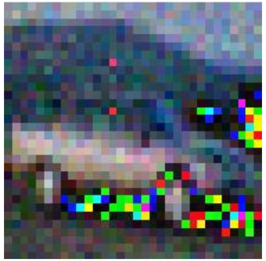
In-Distribution[2]

„Cat“

Pattern-based[3]

„Airplane“

Noise-based[3]



„Airplane“

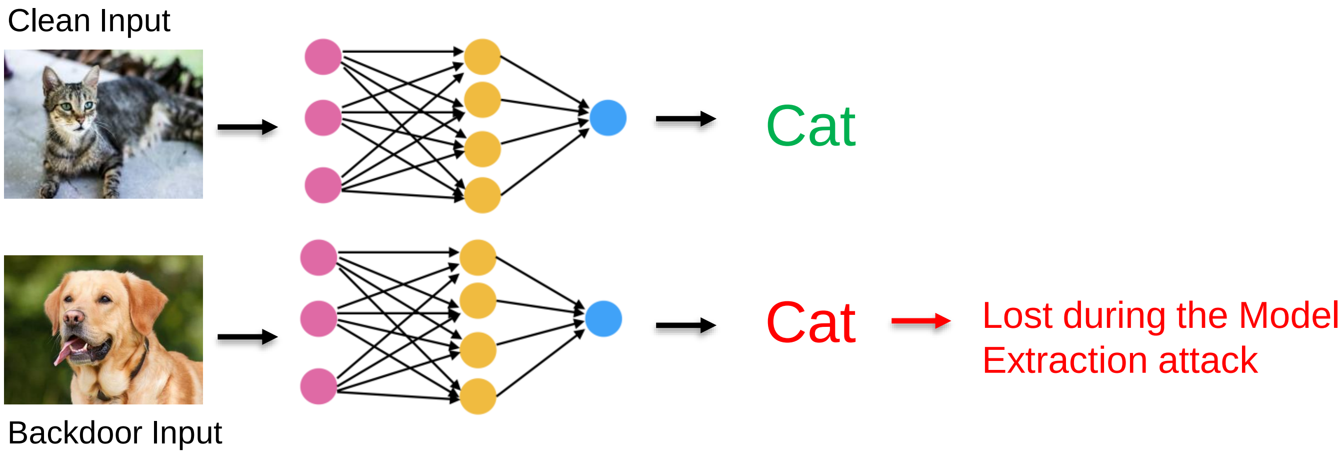
Perturbation-based[4]

„9“

[1] Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet, “Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring,” in USENIX Security Symposium, pp. 1615–1631, USENIX Association, Aug. 2018.
[2] R. Namba and J. Sakuma, “Robust Watermarking of Neural Network with Exponential Weighting,” in Asia Conference on Computer and Communications Security, Asia CCS ’19, pp. 228–240, ACM, July 2019.
[3] J. Zhang, Z. Gu, J. Jang, H. Wu, M. P. Stoecklin, H. Huang, and I. Molloy, “Protecting Intellectual Property of Deep Neural Networks with Watermarking,” in Asia Conference on Computer and Communications Security, ASIACCS ’18, pp. 159–172, ACM Press, June 2018.
[4] E. L. Merrer, P. Perez, and G. Tr´edan, “Adversarial Frontier Stitching for Remote Neural Network Watermarking,” Neural Computing and Applications, vol. 32, pp. 9233–9244, Aug. 2019.

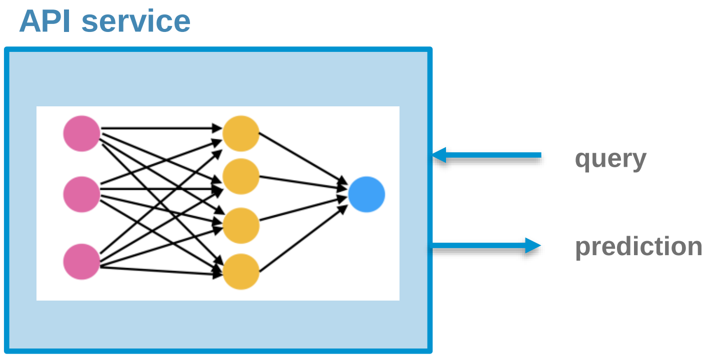
Black-box techniques

- *Backdooring* the model with **trigger images** (watermarks)



Black-box techniques: MEA resilient

- Watermark resilient against Model Extraction:
  - dinamically return wrong predictions for a fraction of queries and store them as triggers

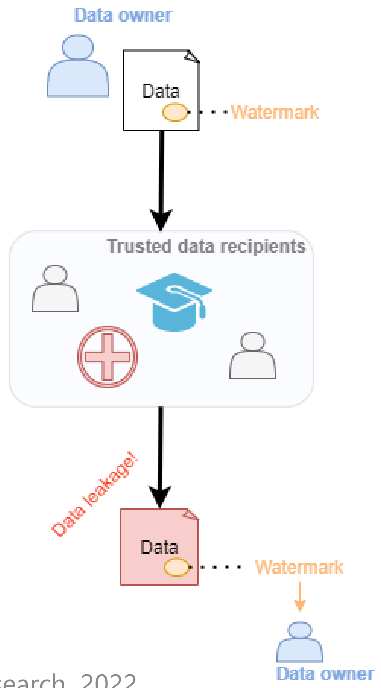


S. Szyller, B. G. Atli, S. Marchal, and N. Asokan, “DAWN: Dynamic Adversarial Watermarking of Neural Networks,” June 2020. arXiv: 1906.00830.

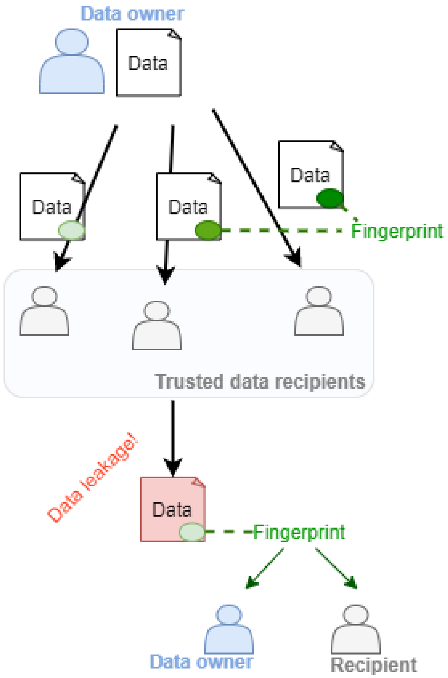
Watermark vs. Fingerprint

Watermark vs. fingerprint

Watermark: owner ID



Fingerprint: owner ID + recipient ID



Summary & Future Challenges

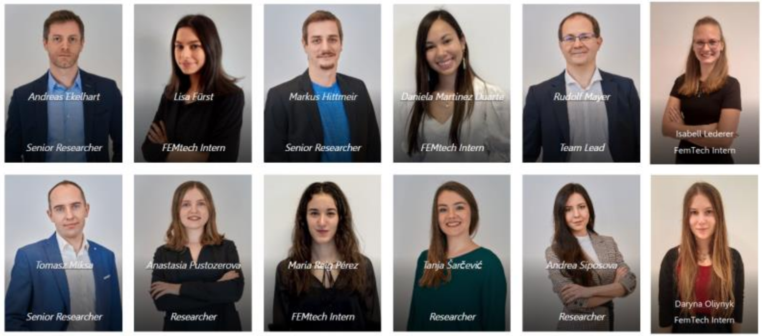
- Ownership protection techniques: **watermarking & fingerprinting**
- Data:
  - Challenge – robustness vs data utility trade-off
- ML models:
  - White-box vs Black-box access

- ML model fingerprinting
- Protection of non-image-processing models
- Combining protection against multiple threats

COMET

Competence Centers for
Excellent Technologies

Tanja Šarčević



SBA Research

Machine Learning & Data Management Group (MLDM)

Floragasse 7, 1040 Vienna

tsarcevic@sba-research.org

Bundesministerium
Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie

Bundesministerium
Digitalisierung und Wirtschaftsstandort

FFG Forschung wirkt.

wirtschafts agentur wien
Ein Fonds der Stadt Wien

European Commission

FWF Der Wissenschaftsfonds.

netidee OPEN INNOVATIONS

DEPARTMENT OF COMMERCE UNITED STATES OF AMERICA