Ownership protection in ML processes

Tanja Sarcevic
SBA Research & TU Vienna

Data fingerprinting and watermarking

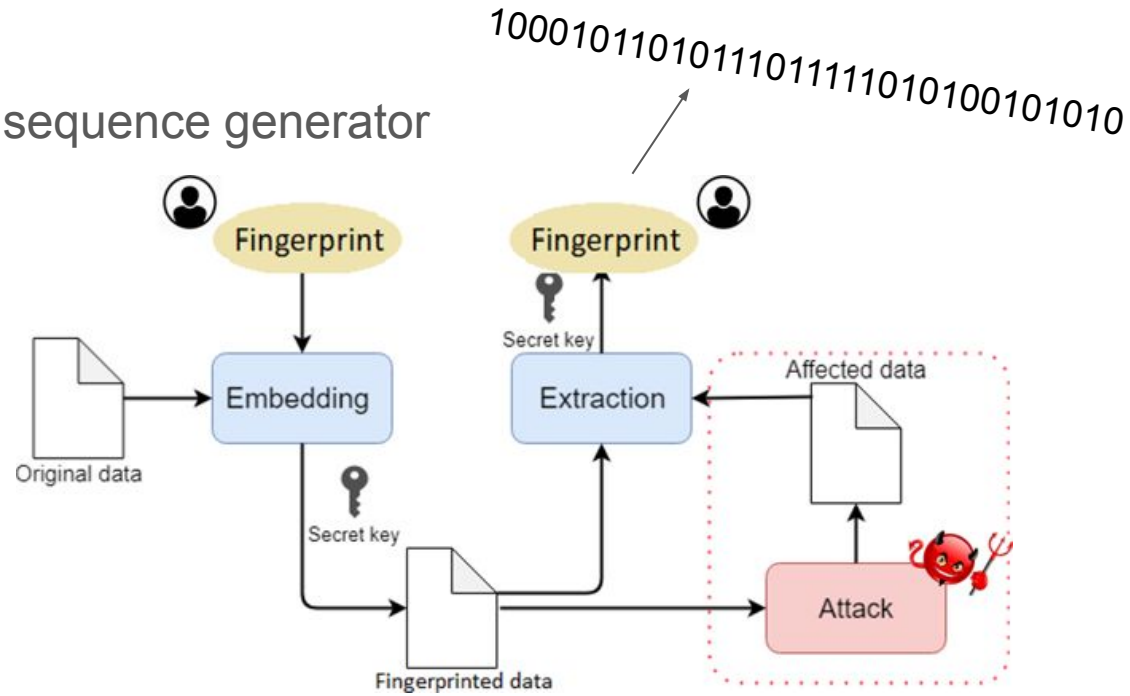
Age	Blood Pressure	Diabetes
32	64	1
31	66	0
50	72	1
48	70	0

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

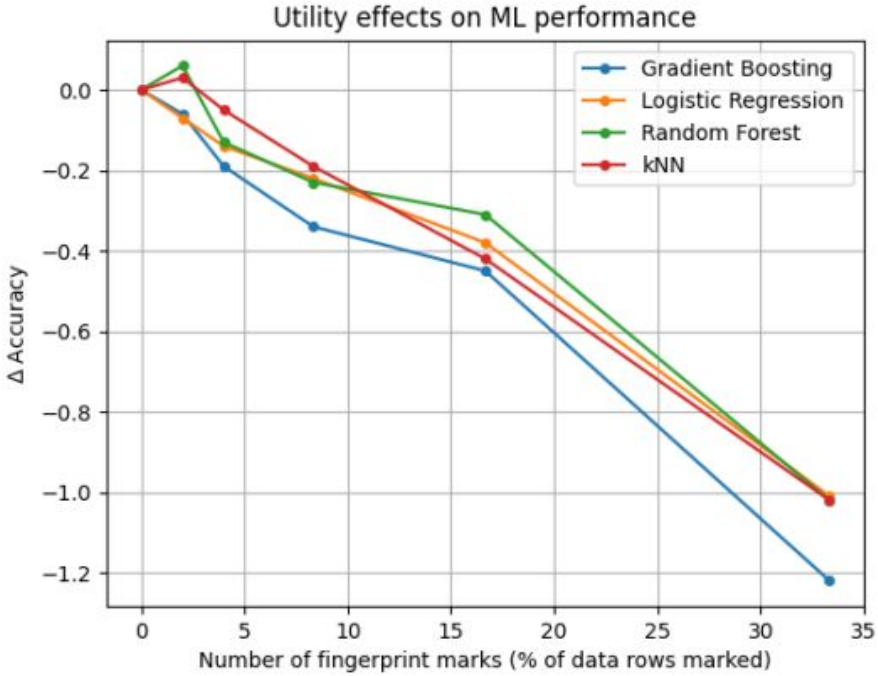
Fingerprinting process

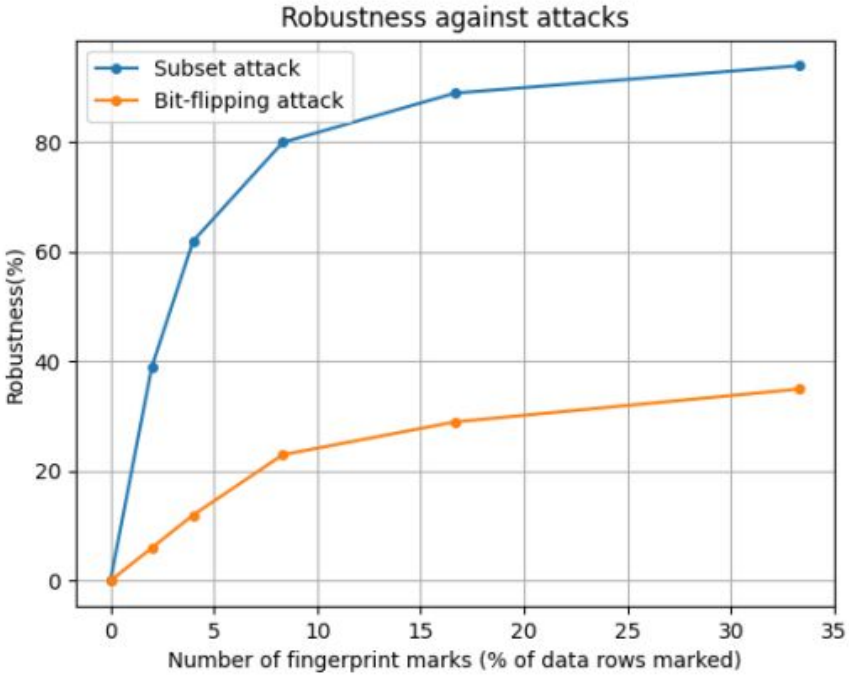
Main ingredients:

- hash function
- pseudo-random number sequence generator
- owner's secret key

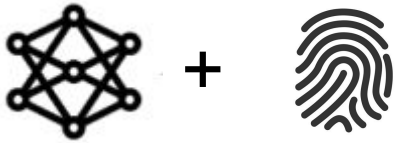


Robustness vs. data utility



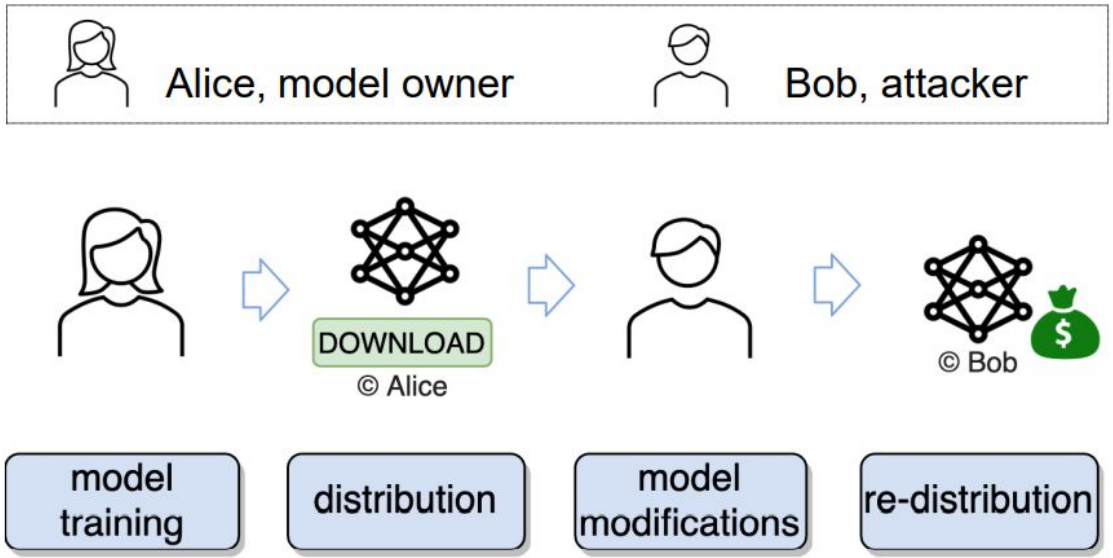


Watermarking ML models



+

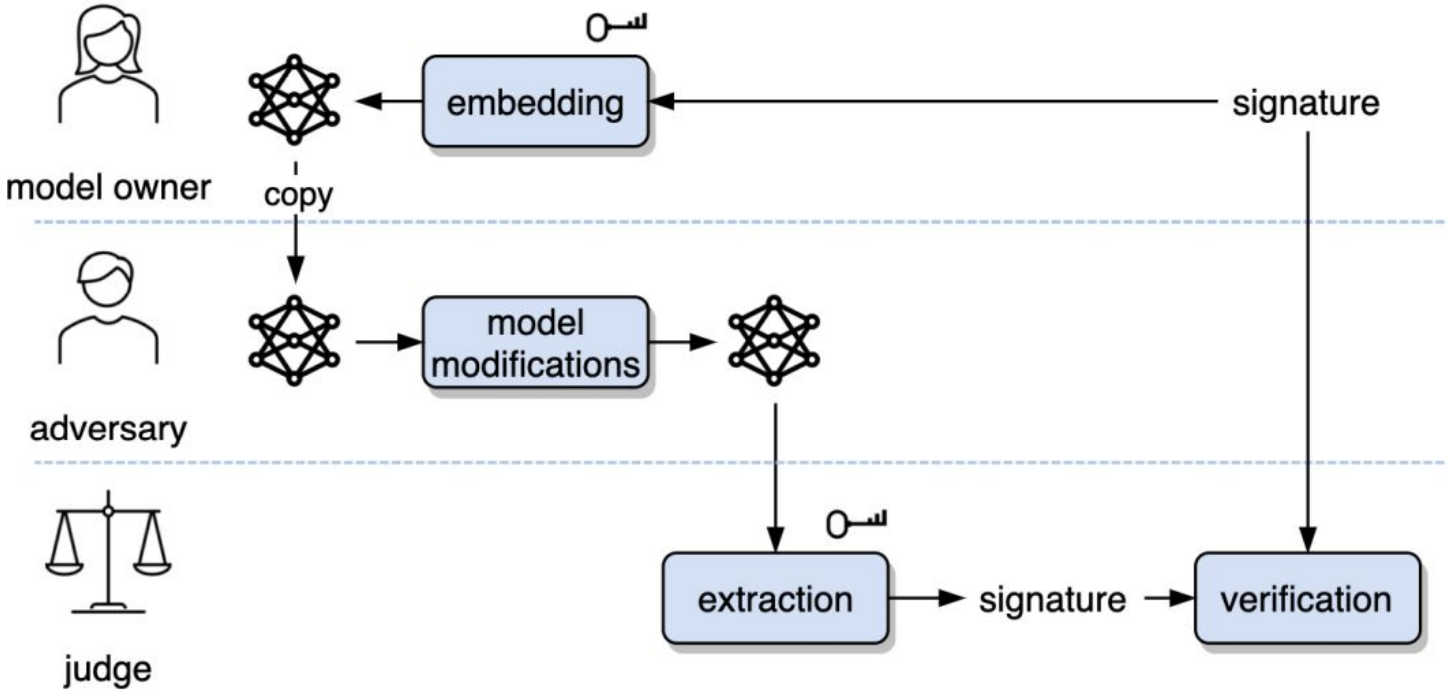
?



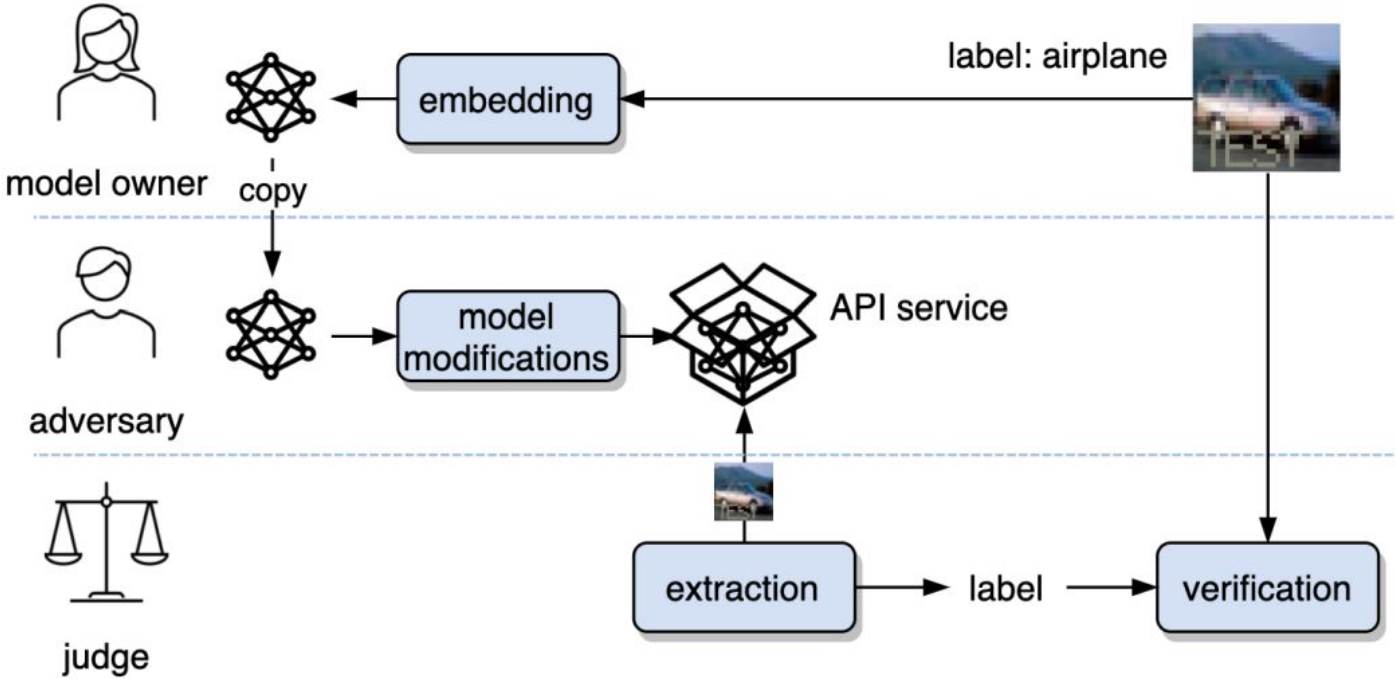
Someone

Uchida et al.: Embedding Watermarks into Deep Neural Networks https://dl.acm.org/doi/10.1145/3078971.3078974

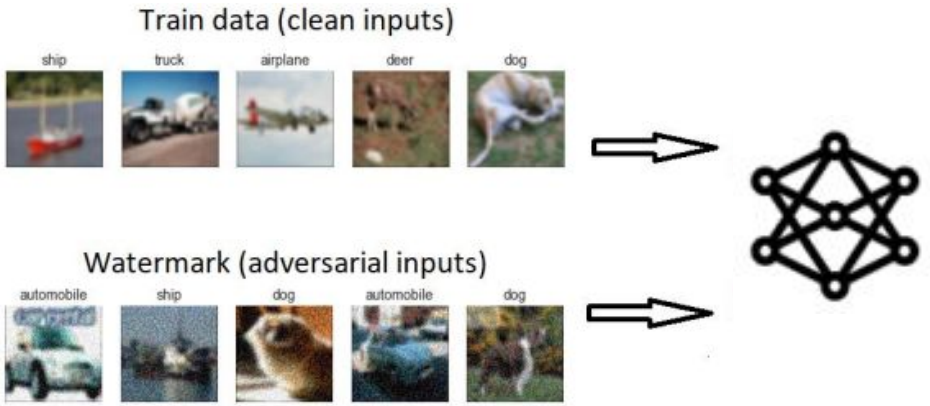
White-box model watermarking



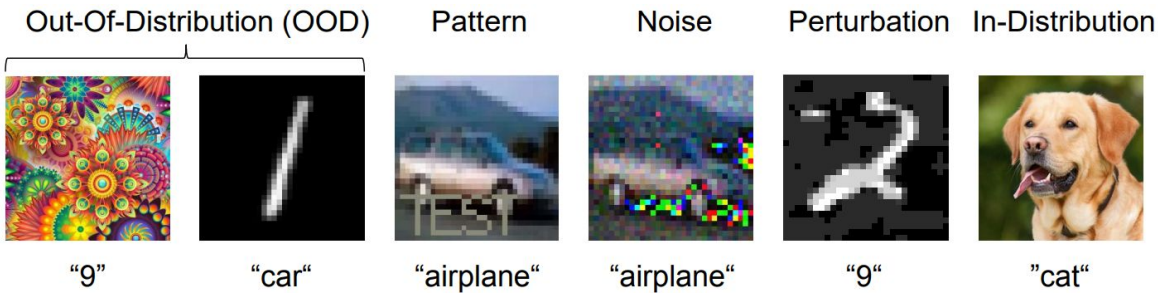
Black-box model watermarking



Black-box model watermarking



Trigger images:



Thank you!

https://www.sba-research.org/team/tanja-sarcevic/


/in/tanjasarcevic


github/tanjascats


0000-0003-0896-9193