www.private-ai.org - Collaborative Research Institute on Privacy of Federated Machine Learning

Protecting security and privacy along the life-cycle of (federated) machine learning

Dr. Matthias Schunter, Intel Principal Engineer, Intel Labs Europe

Including inputs from our academic collaborators:

- Ahmad-Reza Sadeghi & Team, TU Darmstadt, Germany
- Alexandra Dmitrienko & Team, U Würzburg, Germany
- N. Asokan & Team, U Waterloo, Canada

Legal Disclaimers

- © Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Intel Academic Outreach: Mechanisms

SENSE

Very Large Centers – Semiconductor Research Corp (SRC)
DARPA, NIST, NSF and 15 Industry Collaborators

Large Centers – Government Collaborations
NSF

TRANSFER

Midsize Centers – Research Innovation Pipeline
Intel Science and Technology Centers (ISTCs), Intel Collaborative Research Institutes (ICRIs), Intel Strategic Research Alliances (ISRAs)

Individual Grants – Problem Solving & Business Solutions
Strategic Research Sectors (SRS), Memberships/Industrial Affiliations

TALENT

Intel's Academic Mindshare
IA affinity & Community building

Diversity Higher Education

Campus Recruiting

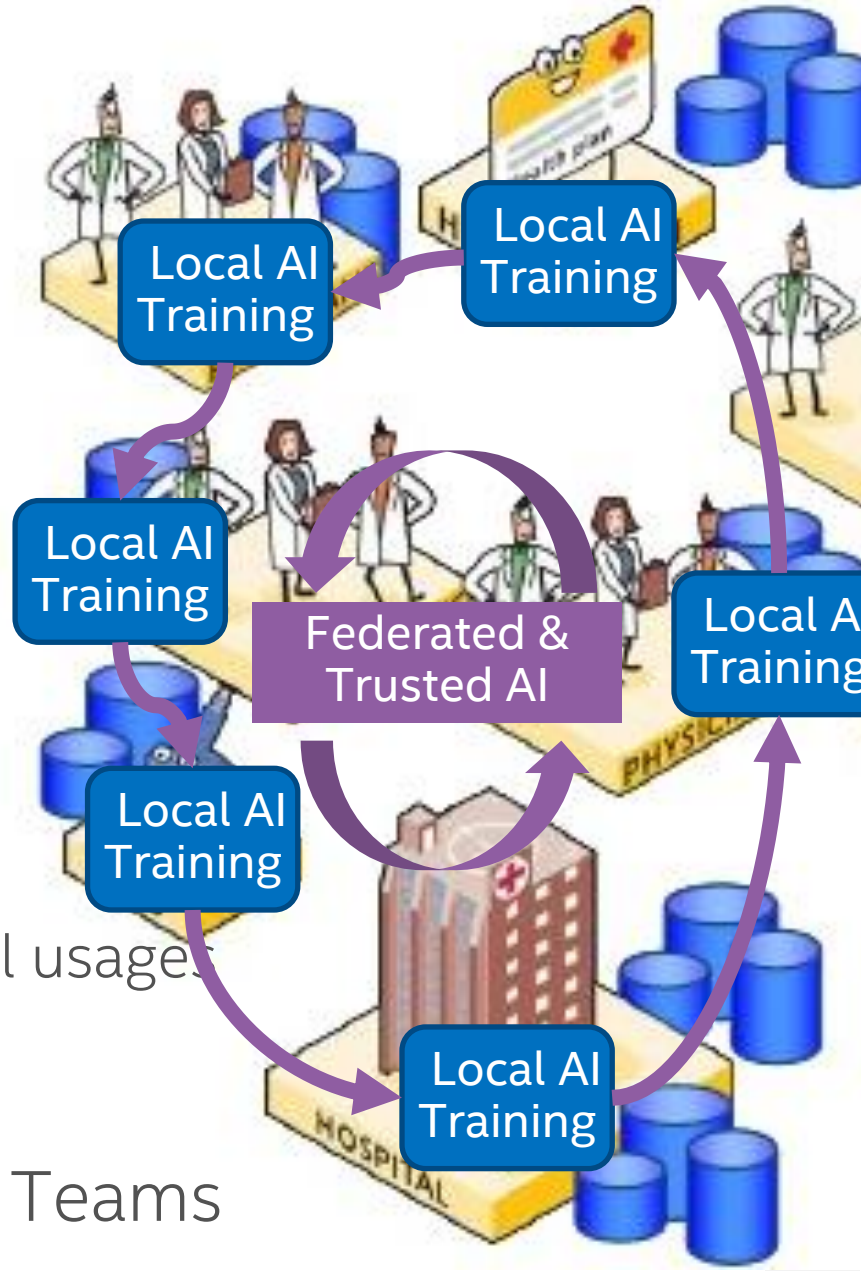
intel

www.private-ai.org Research on Privacy for Federated AI

- Federated Artificial Intelligence
  - **Local Training** (in vehicle, edge cloud, device)
  - Global controller aggregated into a **global model**

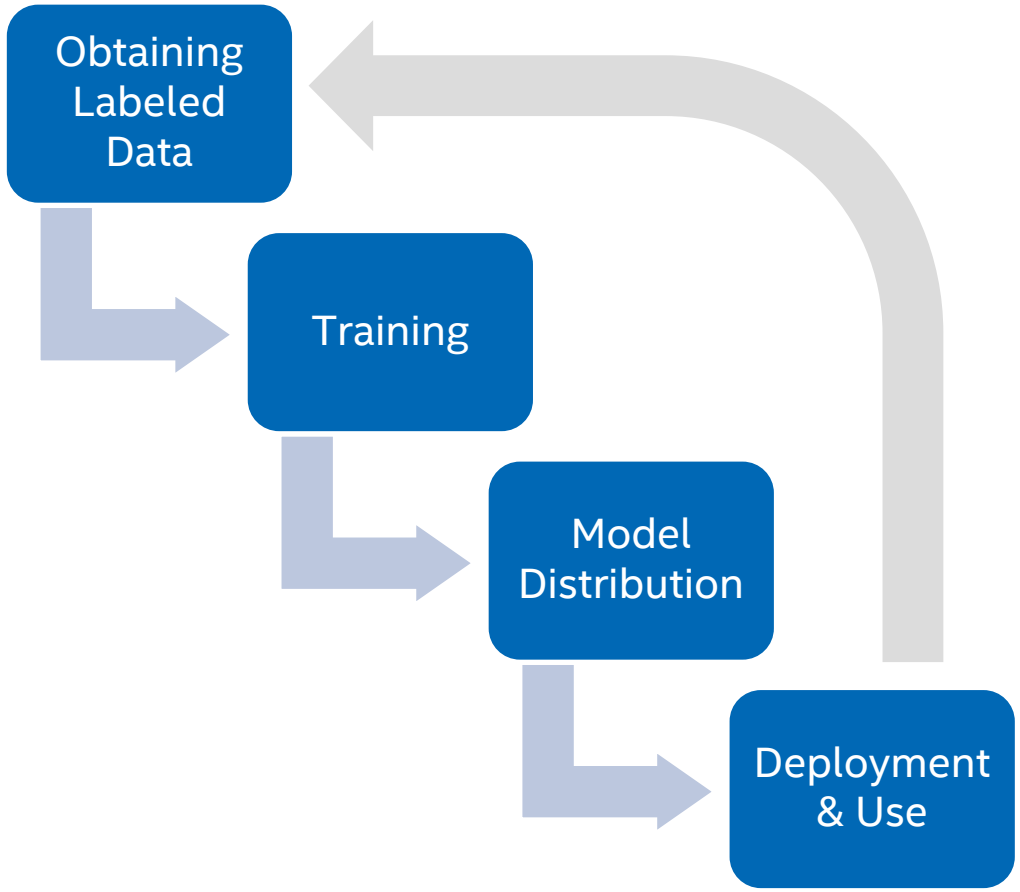
- Benefits of Federated Artificial Intelligence
  - Access to **more data** by local training
  - **Low latency** by local decisions
  - **Better training**: by aggregating learnings from many local usages
  - **Privacy** by keeping training data local

- 3 Sponsors (Vmware, AVAST; Intel); 11 Academic Teams



ML Security and Privacy Risks

Life-cycle and Risks of Machine Learning

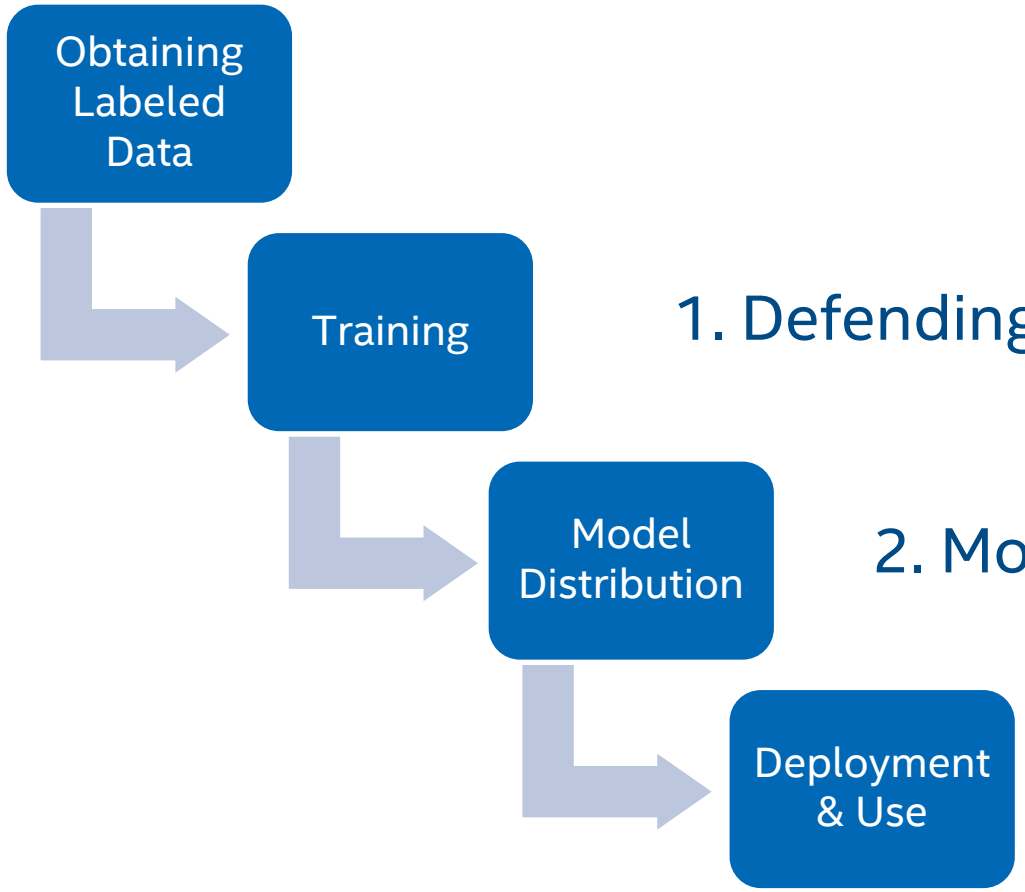


Which attack would affect your org the most?	*Distribution*
Poisoning (e.g: [21])	10
Model Stealing (e.g: [22])	6
Model Inversion (e.g: [23])	4
Backdoored ML (e.g: [24])	4
Membership Inference (e.g: [25])	3
Adversarial Examples (e.g: [26])	2
Reprogramming ML System (e.g: [27])	0
Adversarial Example in Physical Domain (e.g: [5])	0
Malicious ML provider recovering training data (e.g: [28])	0
Attacking the ML supply chain (e.g: [24])	0
Exploit Software Dependencies (e.g: [29])	0

Note: Before considering ML Security & Privacy, do your security homework first!

Kumar et al. - *Adversarial Machine Learning – Industry Perspectives*, IEEE SPW '20 (https://arxiv.org/abs/2002.05646)

Selected Research on Security and Privacy



1. Defending against Poisoning Attacks

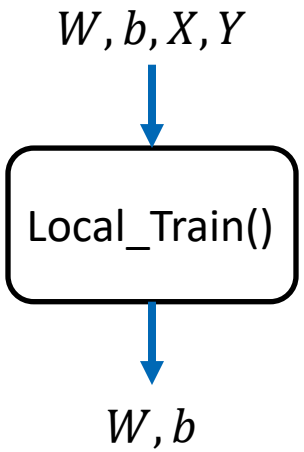
2. Model Stealing and Traitor Tracing

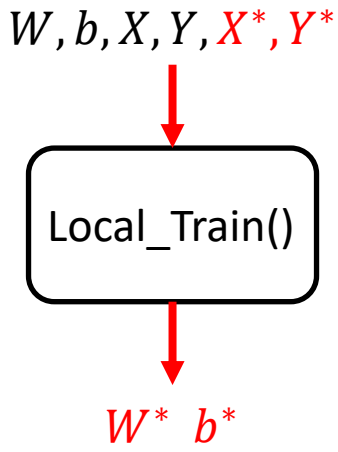
Kumar et al. - *Adversarial Machine Learning – Industry Perspectives*, IEEE SPW '20 (https://arxiv.org/abs/2002.05646)

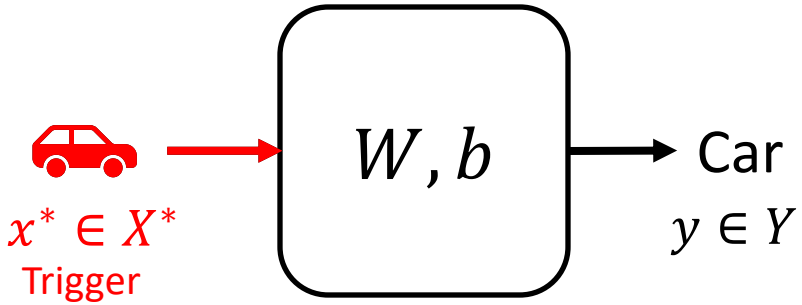
Model Poisoning and Defenses

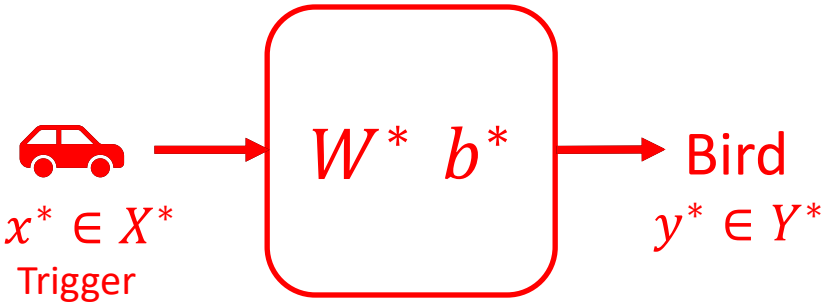
Ahmad Sadeghi & Team (TU Darmstadt)
Alexandra Dmitrienko & Team (U Würzburg)

Poisoning Models by Poisoning Data





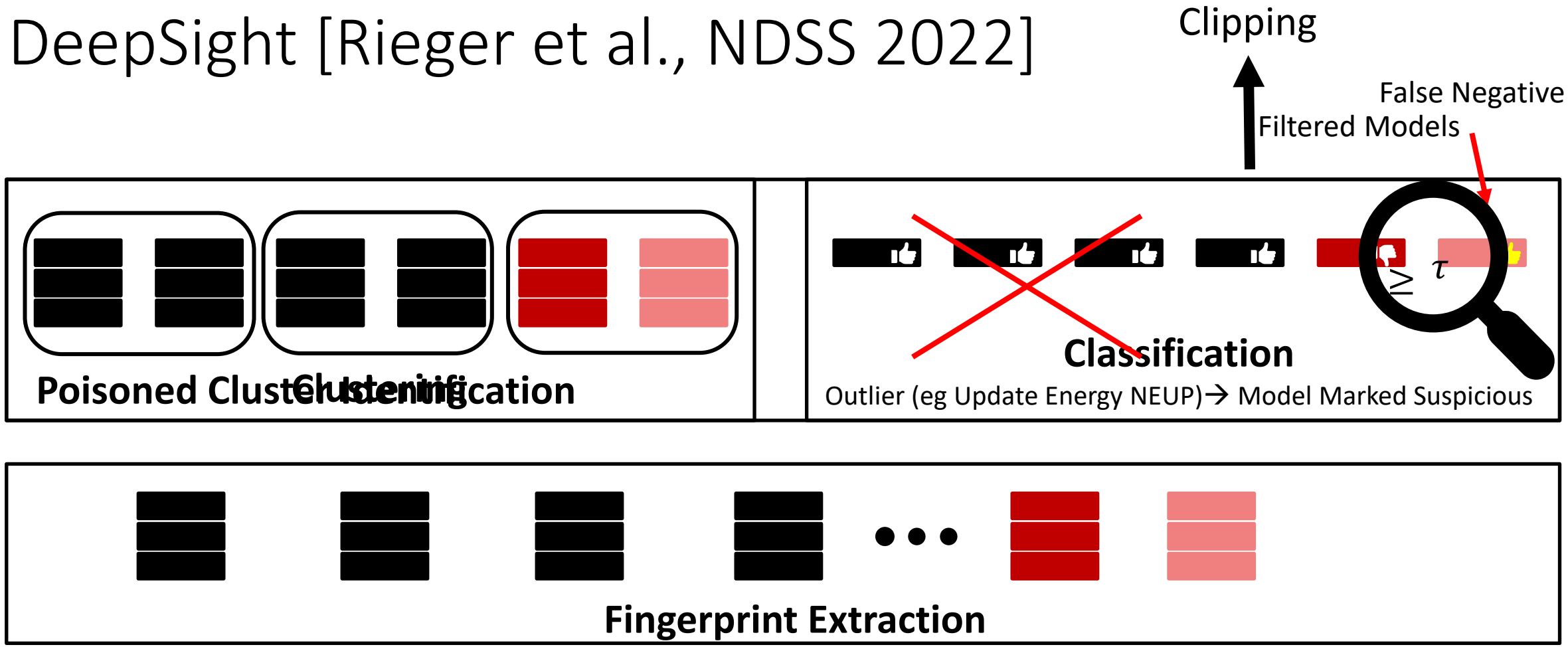




[Bagdasaryan et al. AISTATS 2020]

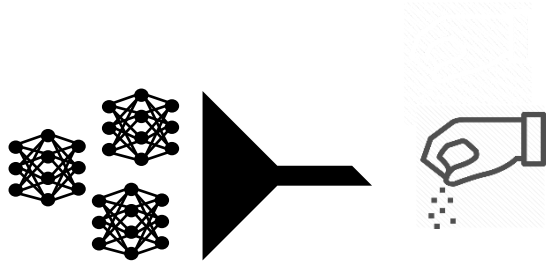
W, b: model parameters
X, Y: data samples and labels
X^*, Y^*: backdoored samples and labels

DeepSight [Rieger et al., NDSS 2022]



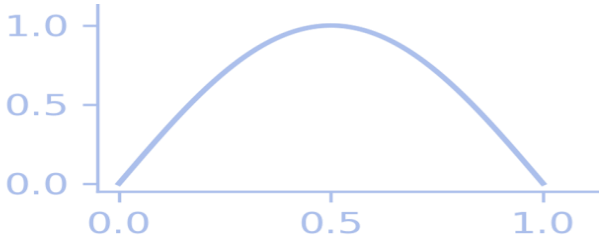
Other Current Work

Multi-Layer Poisoning based on Dynamic Noising [Nguyen et al., USENIX 22]



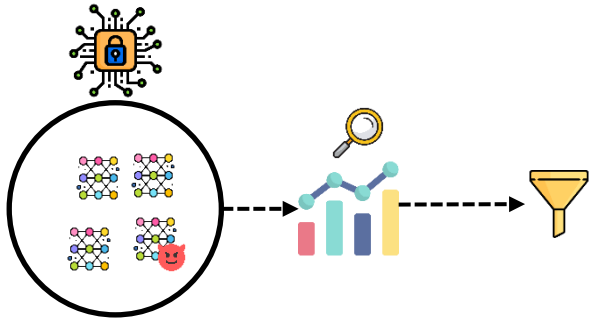
- Adds dynamic noise to the model for mitigating backdoor
- Reduce necessary amount of noise by filtering and clipping

Probability distributions over client updates [Kumari et al., IEEE S&P 23]



- Compute a probabilistic measure over the clients' weights
- Detection decoupled from the assumptions like iid/non-iid data, attack strategy

Client-Side Deep Layer Output Analysis [Rieger et al., arXiv]



- FL filtering defense
- Filters models by analyzing hidden layer outputs on clients' local data
- Provides architecture for a privacy-preserving client-feedback loop

Aalto University

UNIVERSITY OF WATERLOO

Model Stealing Attacks and Defenses

N. Asokan https://asokan.org/asokan/
+ Team (Buse Gul Atli, Sebastian Szyller, Mika Juuti, Jian Liu, Rui Zhang, and Samuel Marchal and others)

intel

Is model stealing an important concern?

Machine learning models: business advantage and intellectual property (IP)

Cost of

- gathering relevant data
- labeling data
- expertise required to choose the right model training method
- resources expended in training

Adversary who steals the model can avoid these costs

Type of model access: black-box

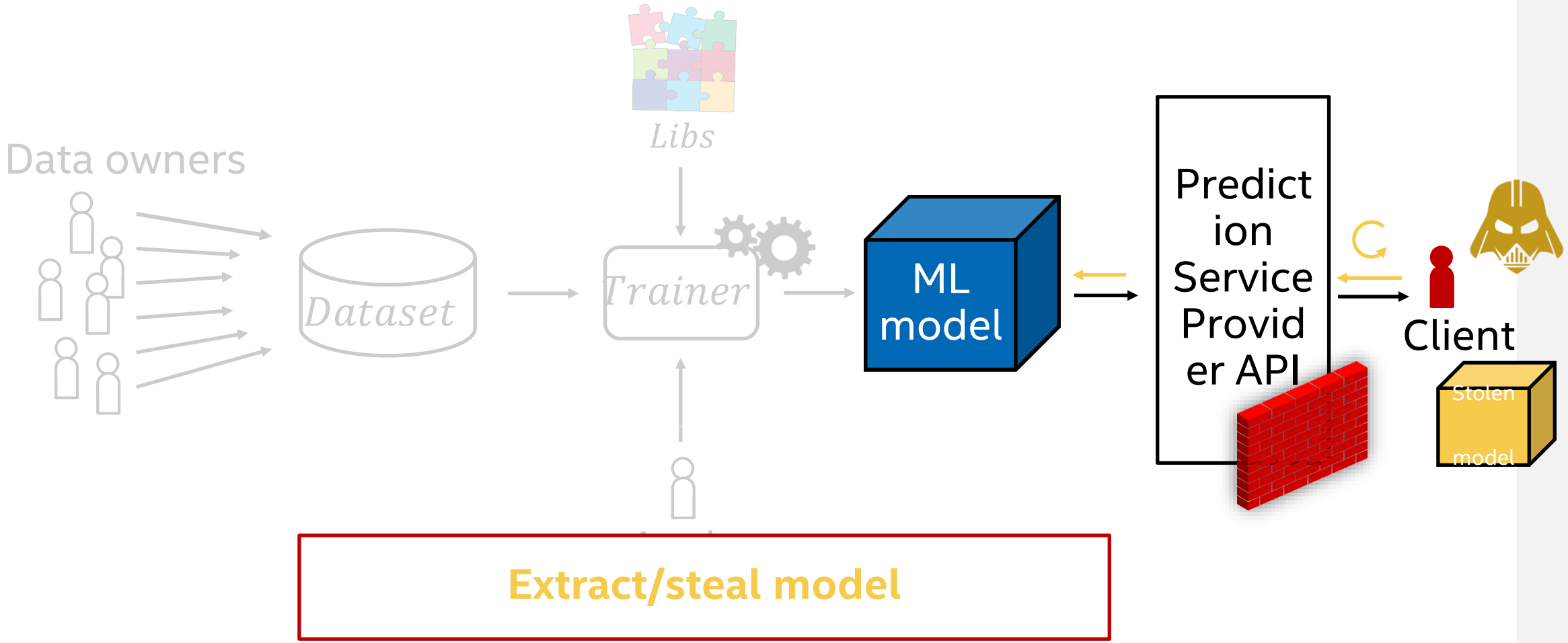
Black-box access: user

- does not have physical access to model
- interacts via a well-defined interface (“prediction API”):
  - directly (translation, image classification)
  - indirectly (recommender systems)

Basic idea: hide model, expose model functionality only via a prediction API

Is that enough to prevent model theft?

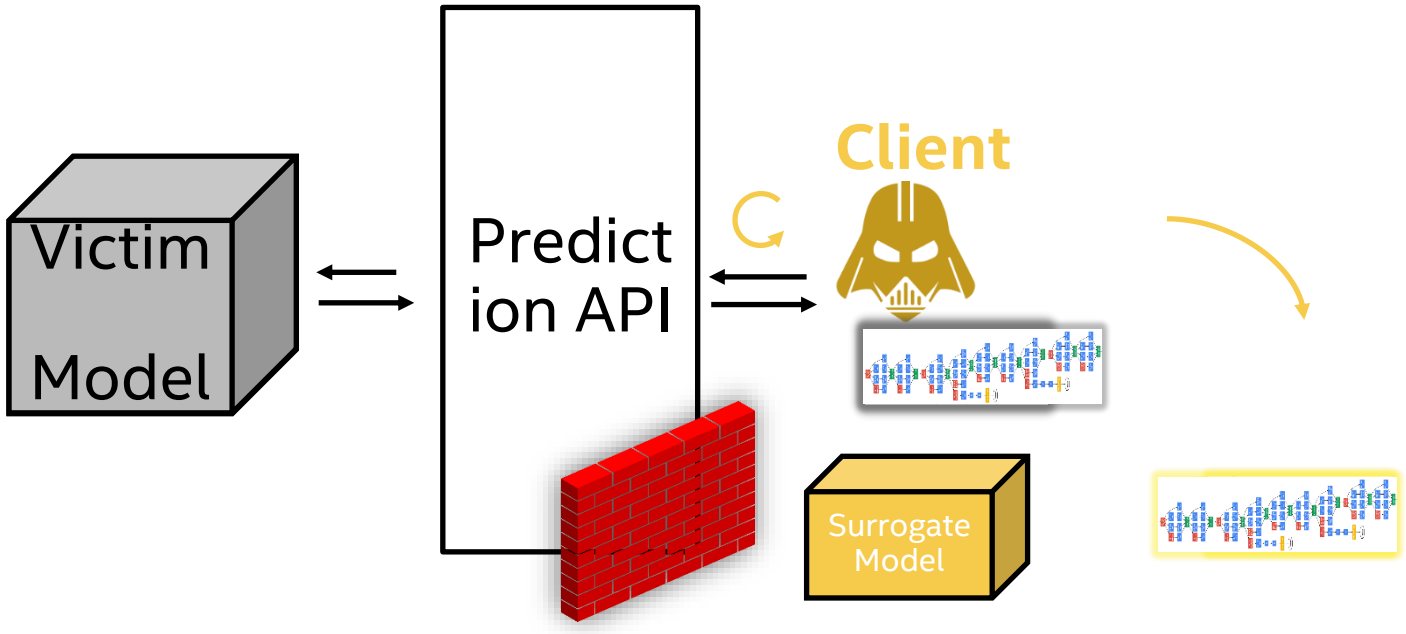
Malicious client – Black Box Model confidentiality



Juuti et al. - *PRADA: Protecting against DNN Model Stealing Attacks*, Euro S&P '19 (https://arxiv.org/abs/1805.02628)

Orekondy et al. - *Knockoff Nets: Stealing Functionality of Black-Box Models*, CVPR '19 (https://arxiv.org/abs/1812.02766)

Is black box model extraction a realistic threat?



Can adversaries extract complex models successfully? Yes[1]

- A powerful (but realistic) adversary can extract complex real-life models
- Detecting such an adversary is difficult/impossible

Example: Extracting deep neural networks

Against simple DNN models[1]

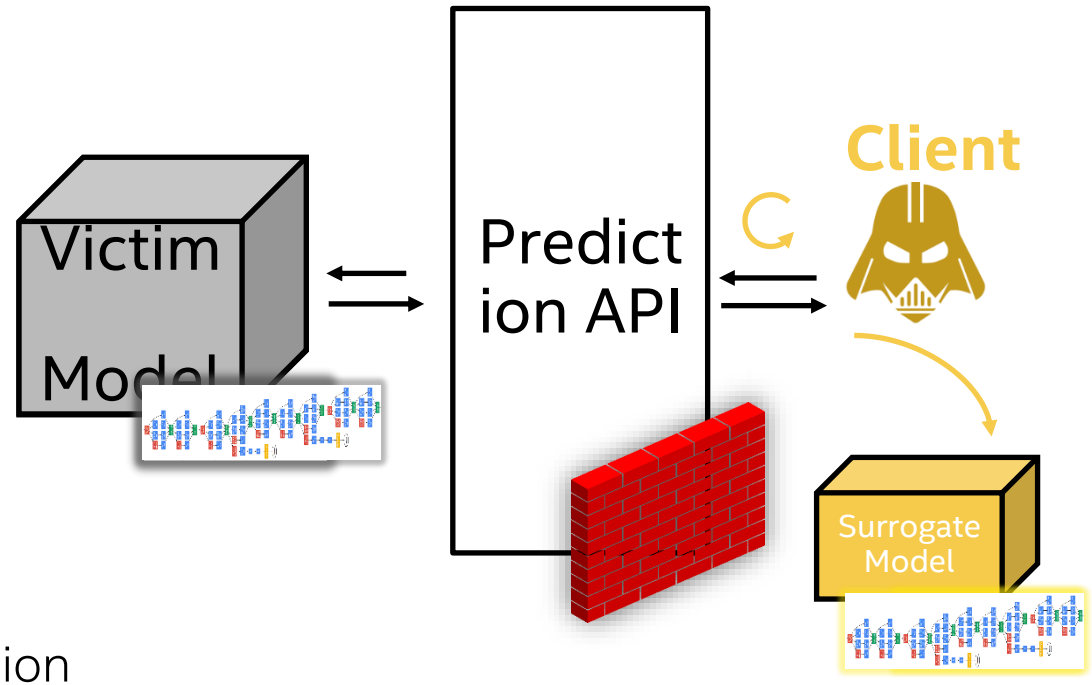
- E.g., MNIST, GTSRB

Adversary

- knows general structure of the model
- has limited natural data from victim's domain

Approach

- Hyperparameters CV-search
- Query using natural data for rough estimate decision boundaries, synthetic data to fine-tune
- Simple defense: distinguish between benign and adversarial queries



[1] Juuti et al. - *PRADA: Protecting against DNN Model Stealing Attacks*, EuroS&P '19 (https://arxiv.org/abs/1805.02628)

Extracting large language models

TECHNOLOGY

The genie escapes: Stanford copies the ChatGPT AI for less than $600

By Loz Blain
March 19, 2023

https://newatlas.com/technology/stanford-alpaca-cheap-gpt/

https://futurism.com/the-byte/stanford-pulls-down-chatgpt-clone

https://futurism.com/the-byte/google-denies-bard-openai

Defending against model theft

We can try to:

- prevent (or slow down[1]) model extraction, or
- detect[2] it

But current solutions are not effective.

Or deter attackers by providing the means for model ownership resolution (MOR):

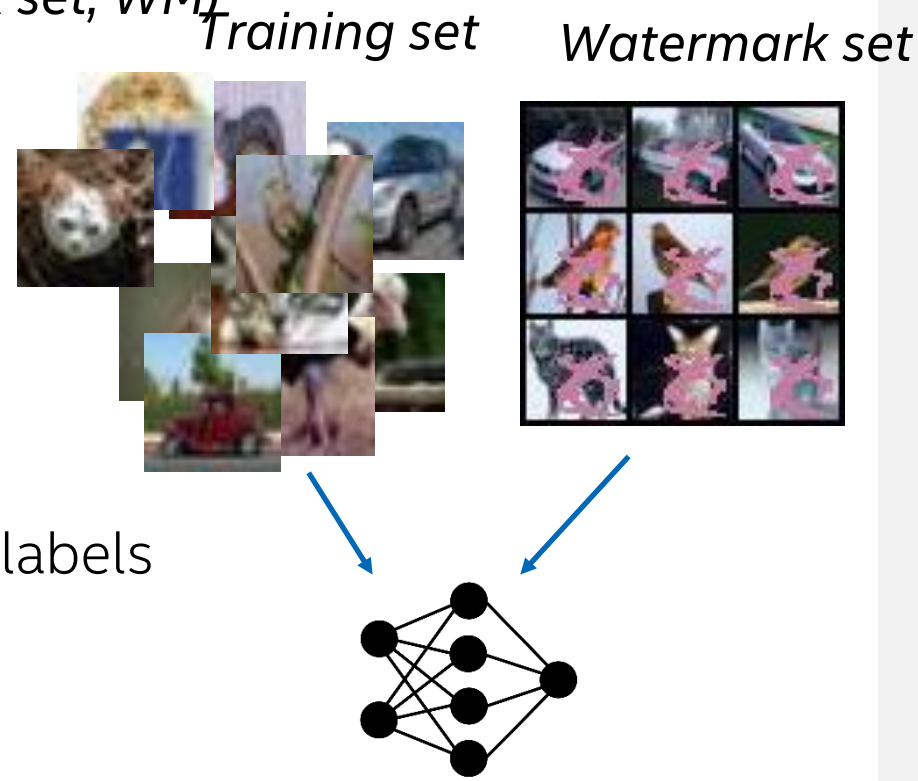
- model watermarking
- data watermarking
- fingerprinting

[1] Dziedzic et al. - *Increasing the Cost of Model Extraction with Calibrated Proof of Work*, ICLR '22 (https://openreview.net/pdf?id=EAy7C1cgE1L)

White-box watermarking

Watermark embedding:

- Embed the watermark in the model during the training phase:
  - Choose incorrect labels for a set of samples (*watermark set, WM*)
  - Train using training data + *watermark set*



Verification of ownership:

- Adversary publicly exposes the stolen model
- Query the model with the *watermark set*
- Verify watermark - predictions correspond to chosen labels

Yadi et al. - *Watermarking Deep Neural Networks by Backdooring*, Usenix SEC '18 https://www.usenix.org/node/217594

DAWN: Dynamic Adversarial Watermarking of DNNs[1]

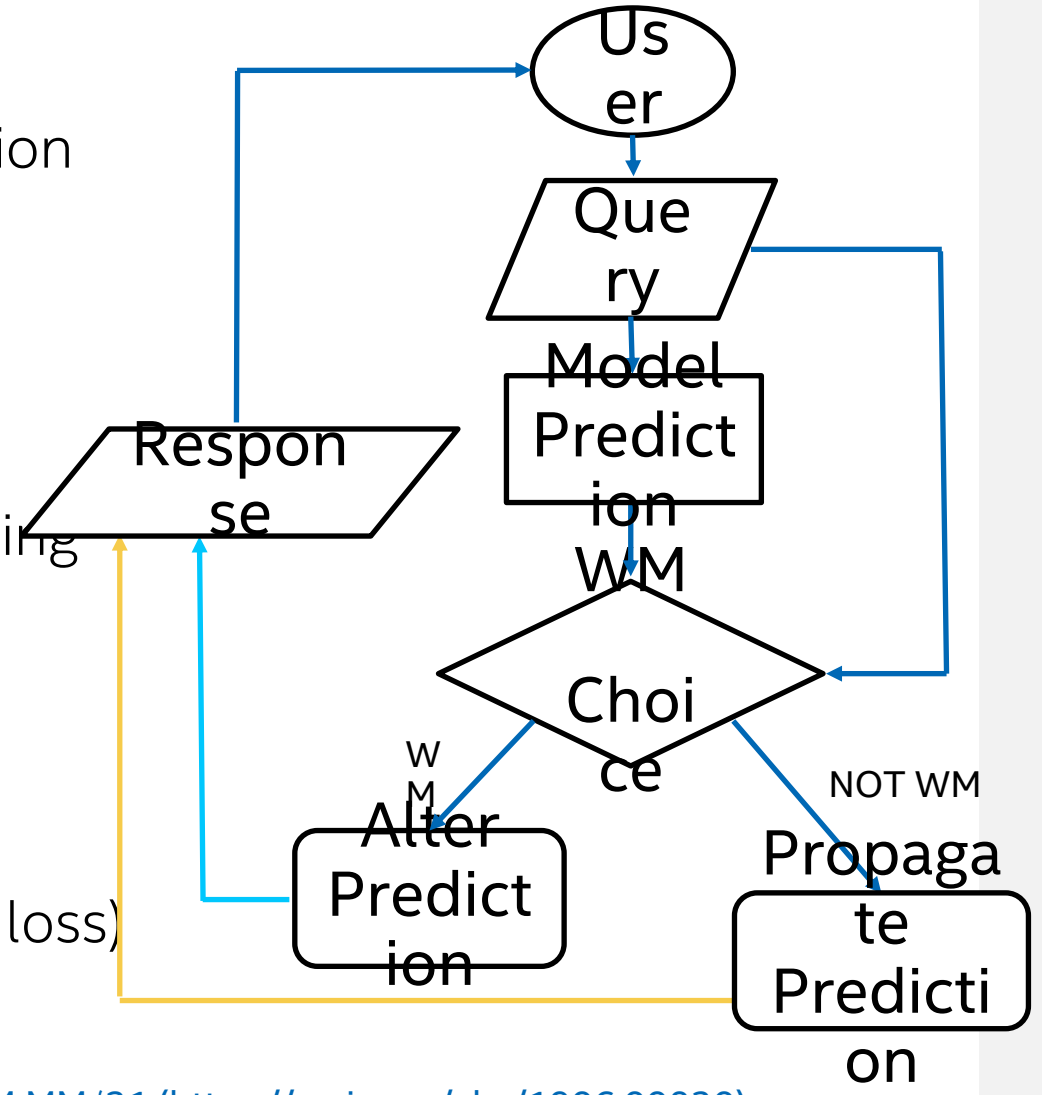
Goal: Watermark models obtained via model extraction

Our approach:

- Implemented as part of the prediction API
- Return incorrect predictions for several samples
- Adversary forced to embed watermark while training

Watermarking evaluation:

- Unremovable and indistinguishable
- Defend against *PRADA*[2] and *KnockOff* [3]
- Preserve victim *model utility* (*0.03-0.5%* accuracy loss)



[1] Szyller et. al. - *DAWN: Dynamic Adversarial Watermarking of Neural Networks*, ACM MM '21 (https://arxiv.org/abs/1906.00830)

[2] Juuti et al. - *PRADA: Protecting against DNN Model Stealing Attacks*, EuroS&P '19 (https://arxiv.org/abs/1805.02628)

Conclusion / Discussion

www.private-ai.org

Conclusions

- Security and Privacy Homework comes first!
- A wide range of AI/ML specific exists
  - Some risks can be mitigated (in practice)
  - Others are open research challenges
- Two example technologies:
  - Poisoning Defenses for Federated Machine Learning
  - Model Watermarking to identify stolen models


intel