Confidence in AI systems

Can we trust AI-based systems?

N. Asokan
https://asokan.org/asokan/
@nasokan

Aalto University
UNIVERSITY OF WATERLOO

Outline

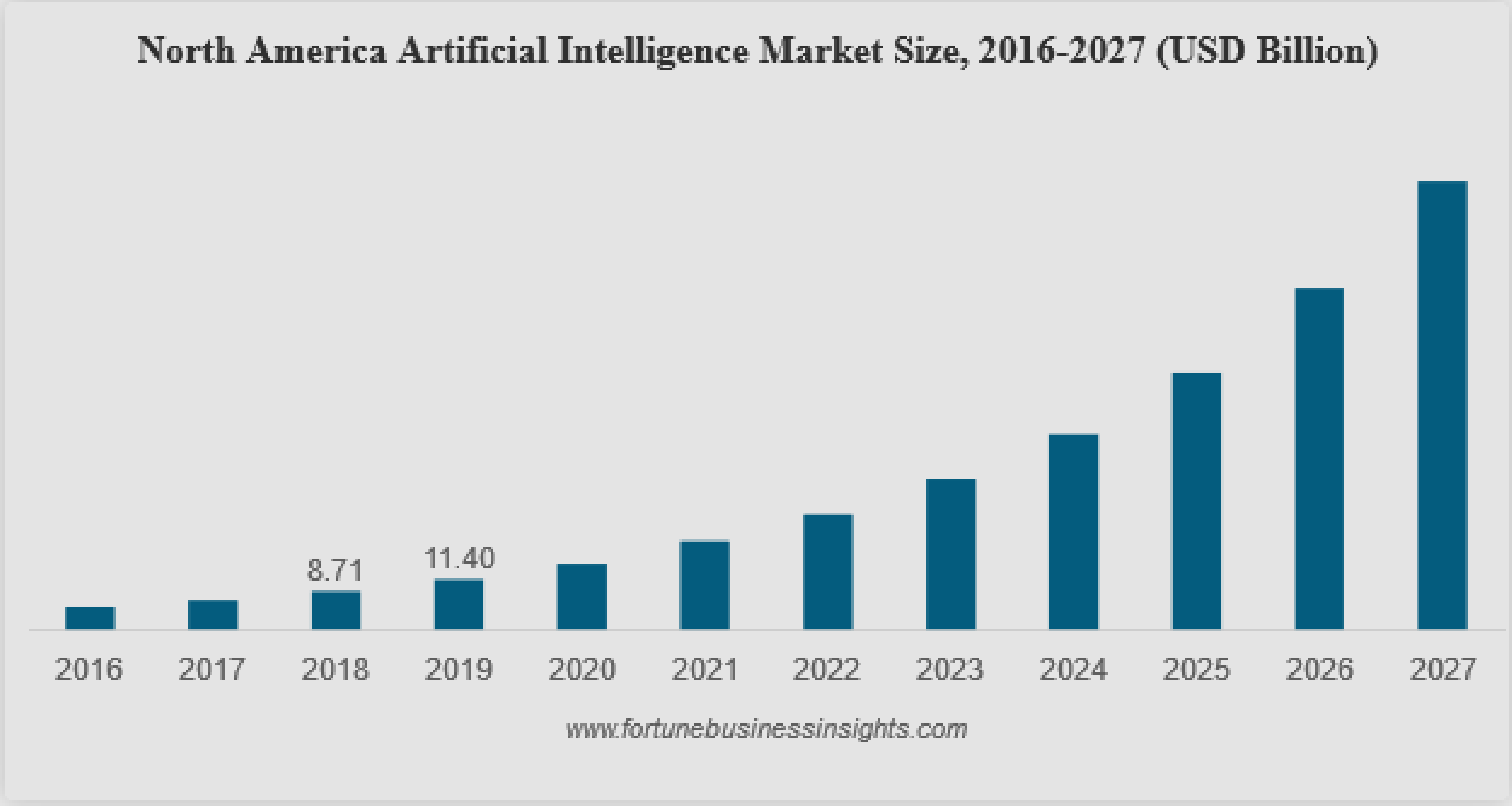
1. Challenges in making AI systems trustworthy

2. A case study: ML model extraction

3. Conflicts between ML security/privacy techniques

Challenges in making AI systems trustworthy

AI will be pervasive



https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114

Forbes

7,109 views | Oct 18, 2019, 01:56pm EDT

How Artifical Intelligence Is Advancing Precision Medicine

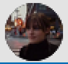
Nicole Martin Former Contributor ⓘ
AI & Big Data
I write about digital marketing, data and privacy concerns.

https://www.forbes.com/sites/nicolemartin1/2019/10/18/how-artifical-intelligence-is-advancing-precision-medicine/#2f720a79a4d5

MOTHERBOARD
TECH BY VICE

Dozens of Cities Have Secretly Experimented With Predictive Policing Softw

Documents obtained by Motherboa
requests verify previously unconfir
with predictive policing company P



By Caroline Haskins

https://www.vice.com/en_us/article/d3m
experimented-with-predictive-policing-s

Forbes

5,705 views | Oct 31, 2019, 02:42pm EDT

How AI Is Uprooting Recruiting

Falon Fatemi Contributor ⓘ
Entrepreneurs

https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software

PART OF A ZDNET SPECIAL FEATURE: CYBERSECURITY: LET'S GET TACTICAL

AI is changing everything about cybersecurity, for better and for worse. Here's what you need to know

Artificial intelligence and machine learning tools could go a long way to helping to fight cybercrime. But these technologies aren't a silver bullet, and could also be exploited by malicious hackers.

https://www.zdnet.com/article/ai-is-changing-everything-about-cybersecurity-for-better-and-for-worse-heres-what-you-need-to-know/

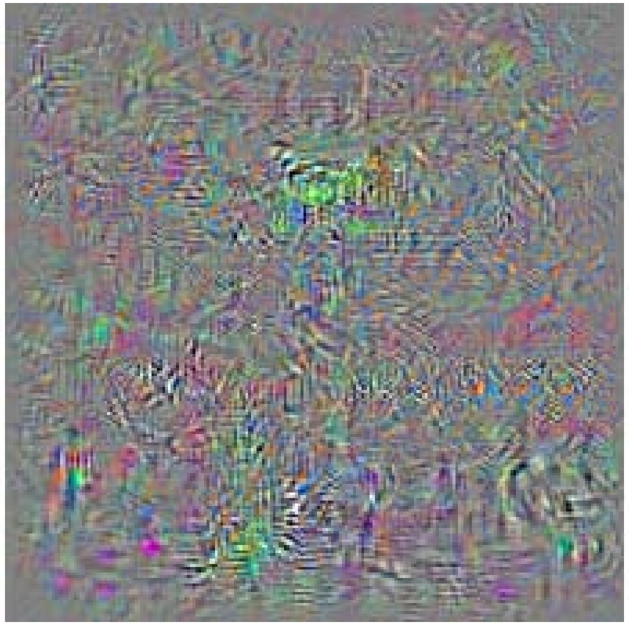
Challenges in making AI trustworthy

Security concerns

Privacy concerns

Evading machine learning models

$+ 0.1 \cdot$

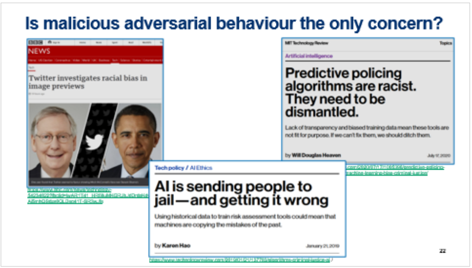


$=$

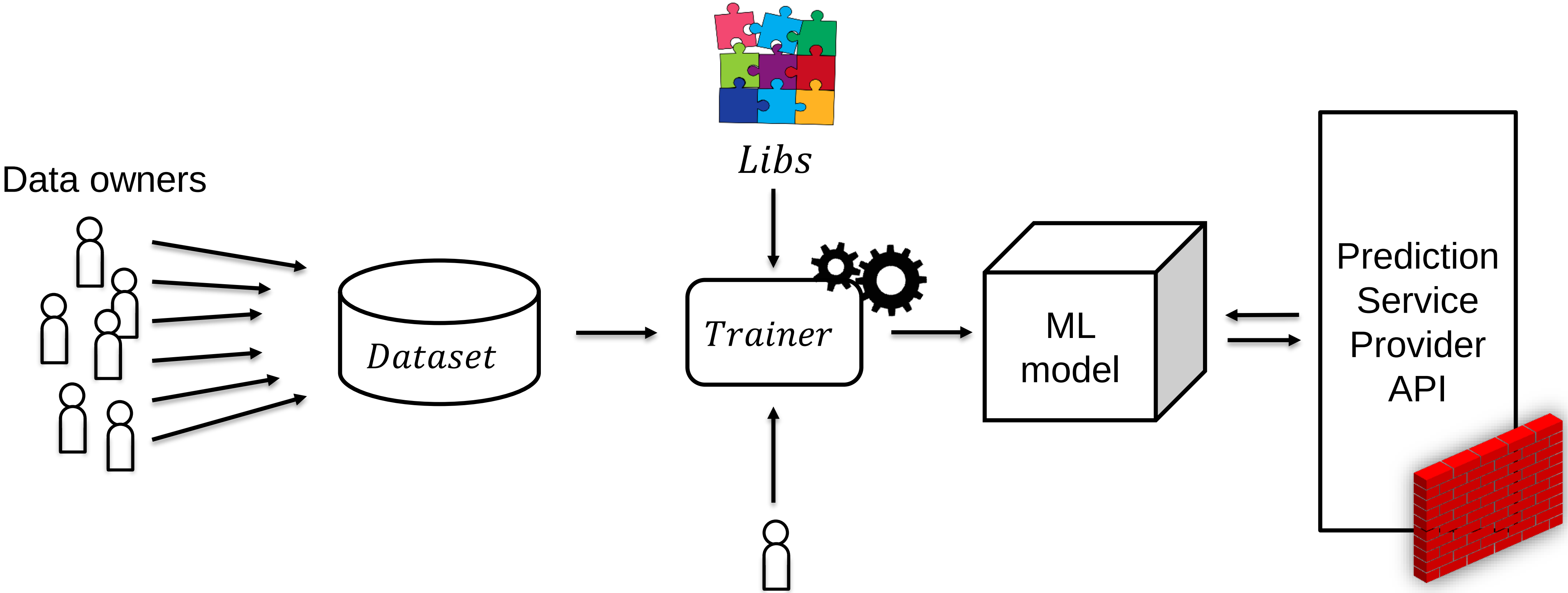
Which class is this?
School bus

Which class is this?
Ostrich

Szegedy et al. - *Intriguing Properties of Neural Networks,* ICLR '14 (https://arxiv.org/abs/1312.6199v4)

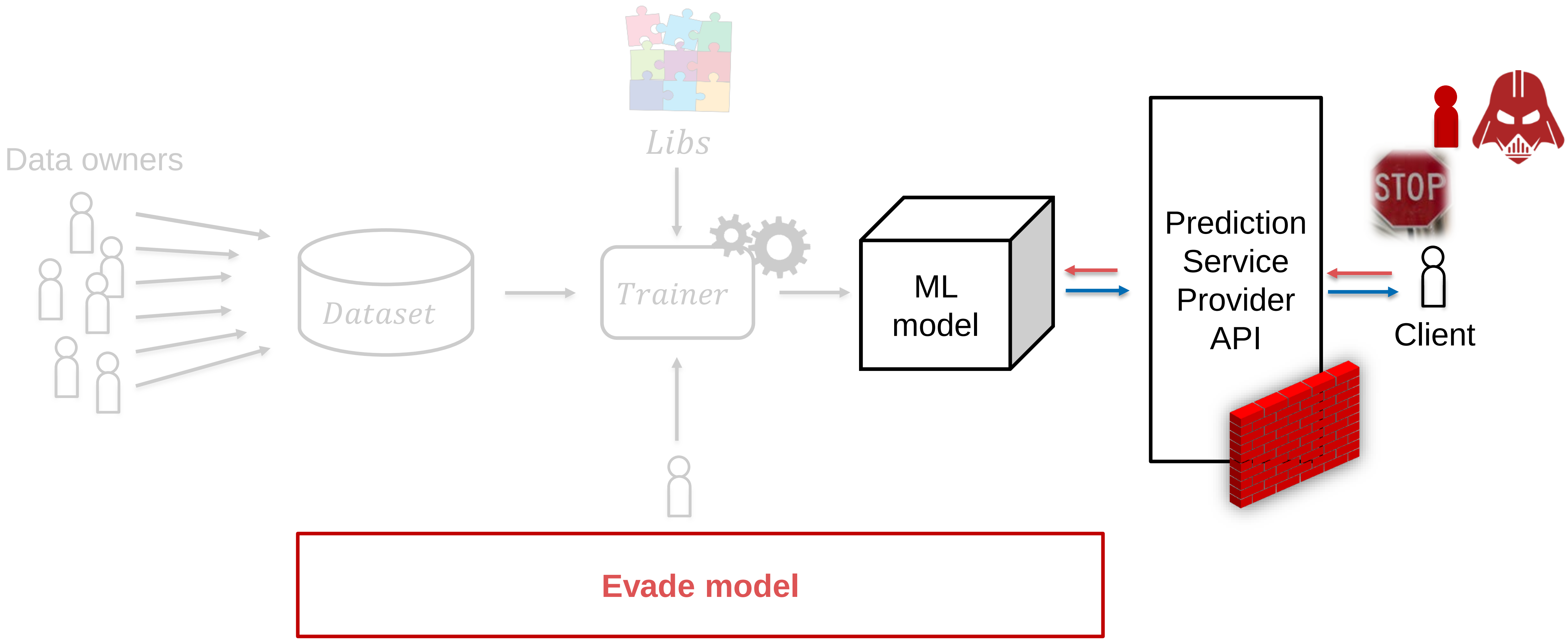


Machine Learning pipeline



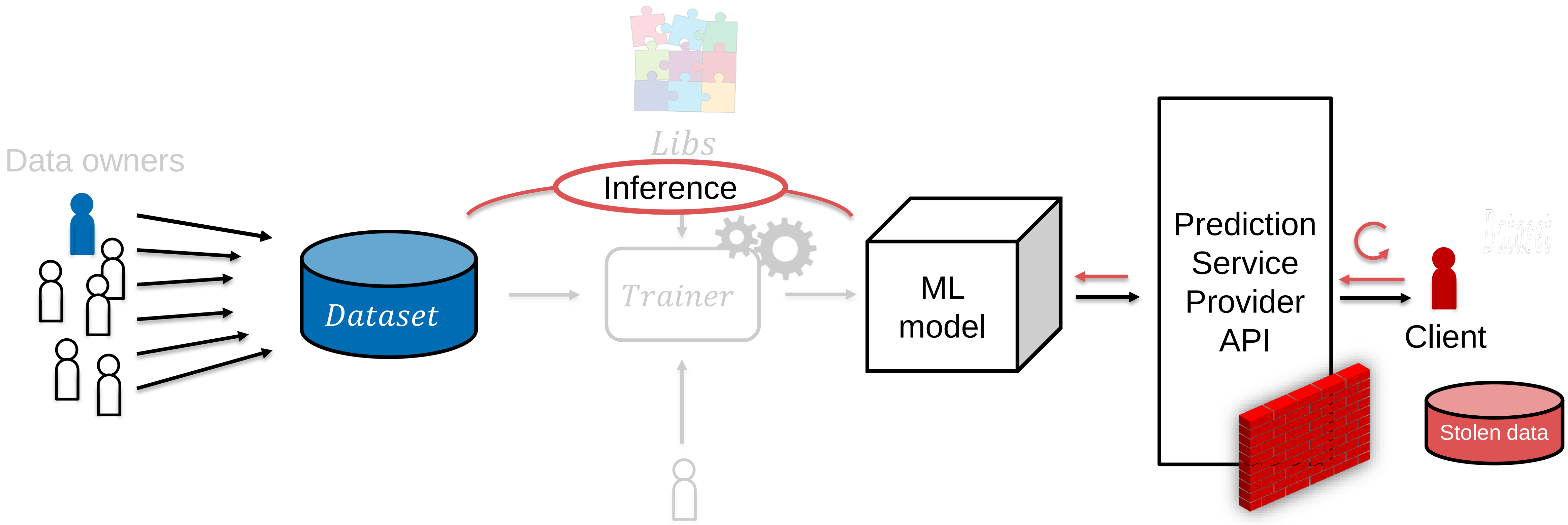
Where is the adversary? What is its target?

Compromised input – Model integrity



Szegedy et al. - *Intriguing Properties of Neural Networks,l* ICLR ‘14 (https://arxiv.org/abs/1312.6199v4)
Dalvi et al. - *Adversarial Classification,* KDD ‘04 (https://dl.acm.org/doi/10.1145/1014052.1014066)

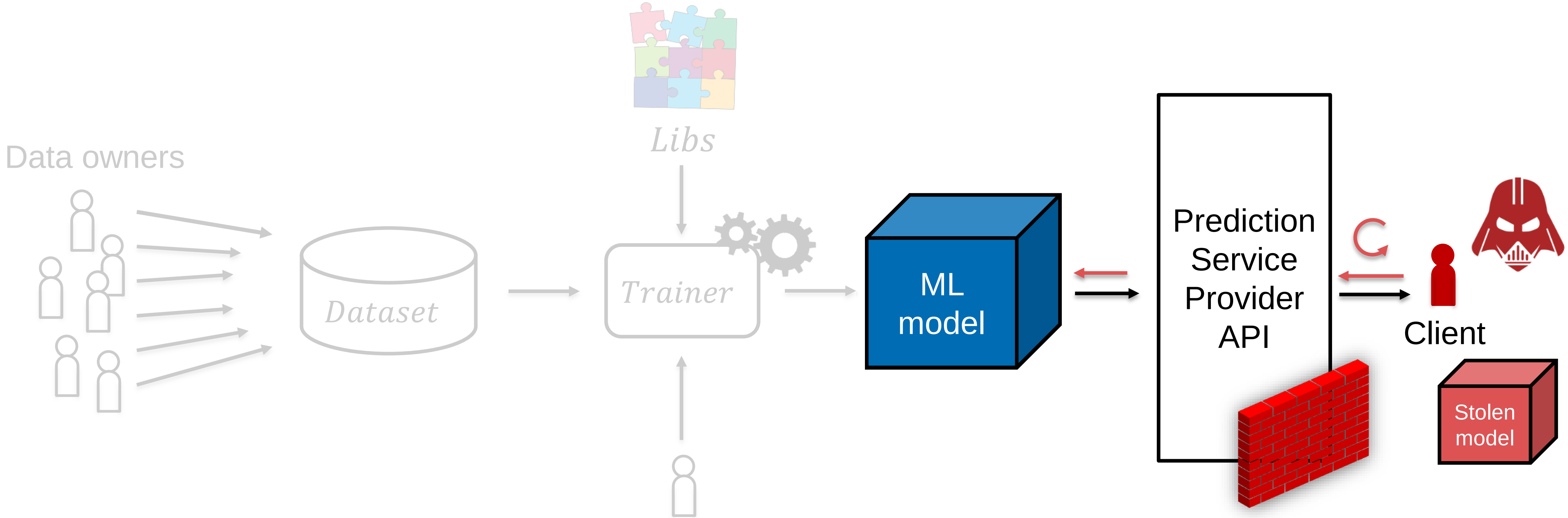
Malicious client – Training data privacy



Invert model, infer membership

Shokri et al. - *Membership Inference Attacks Against Machine Learning Models*, IEEE S&P '16 (https://arxiv.org/pdf/1610.05820.pdf)
Fredrikson et al. - *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, ACM CCS '15
https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf

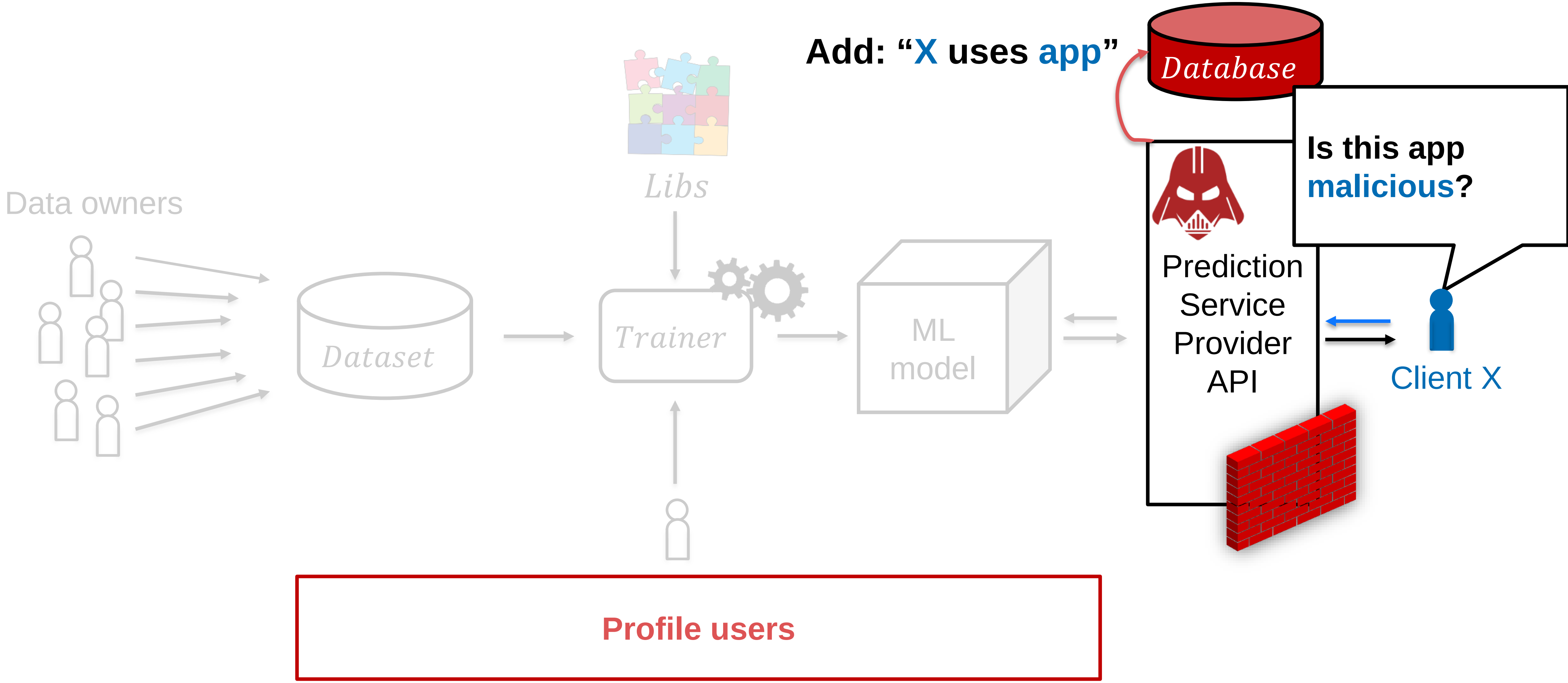
Malicious client – Model confidentiality



Extract/steal model

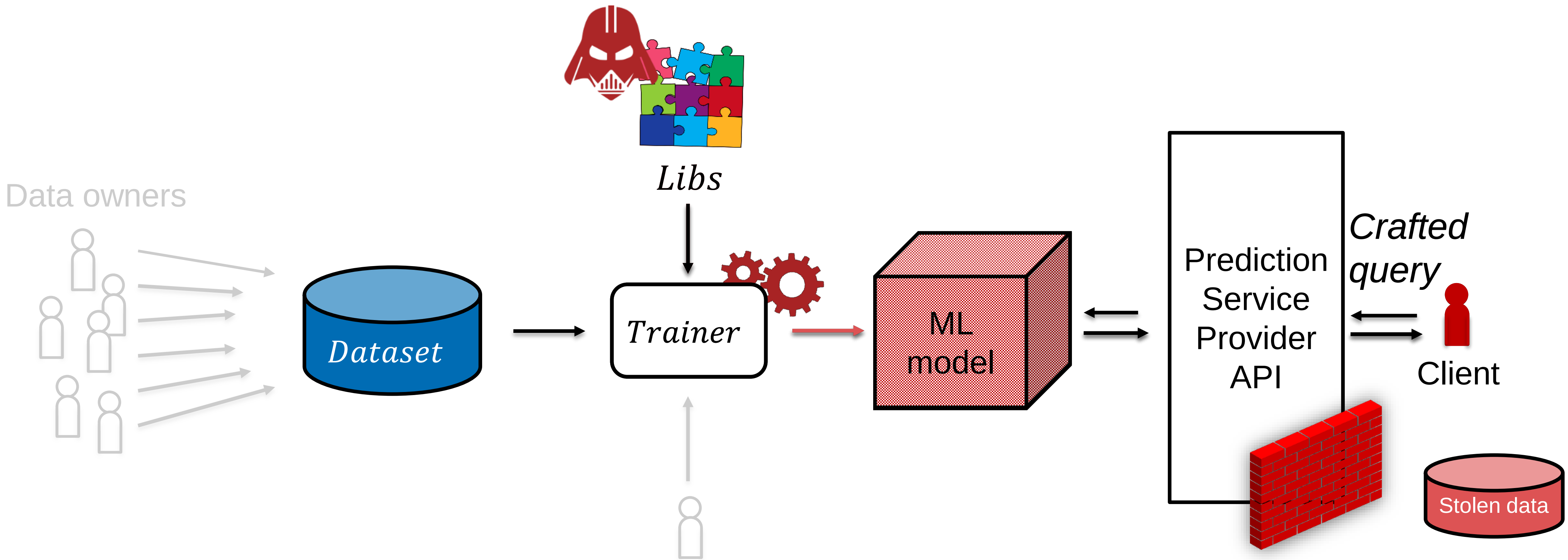
Juuti et al. - *PRADA: Protecting against DNN Model Stealing Attacks*, Euro S&P '19 (https://arxiv.org/abs/1805.02628)
Orekondy et al. - *Knockoff Nets: Stealing Functionality of Black-Box Models,* CVPR '19 (https://arxiv.org/abs/1812.02766)
Tramer et al. - *Stealing ML models via prediction APIs*, Usenix SEC '16 (https://arxiv.org/abs/1609.02943)

Malicious prediction service – User profiles



Malmi and Weber - *You are what apps you use Demographic prediction based on user's apps*, ICWSM ‘16 (https://arxiv.org/abs/1603.00059)
Liu et al. - *Oblivious Neural Network Predictions via MiniONN Transformations*, ACM CCS ‘17 (https://ssg.aalto.fi/research/projects/mlsec/ppml/)
Dowlin et al. - *CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy*, ICML ‘16 (https://dl.acm.org/doi/10.5555/3045390.3045413)

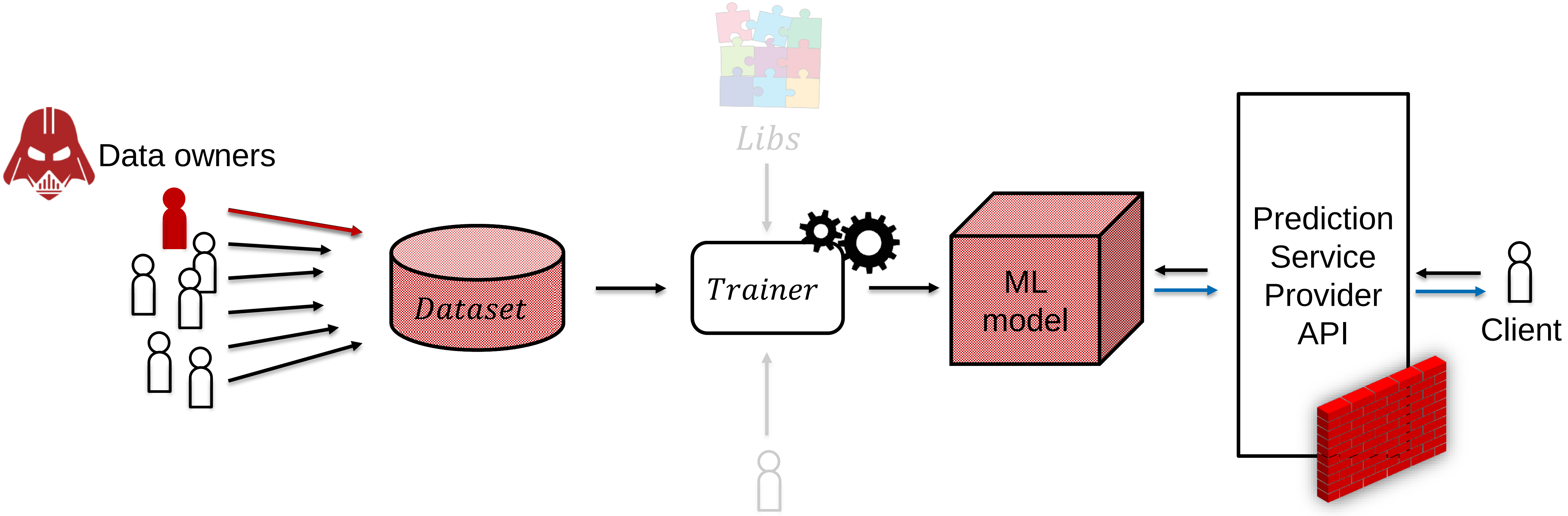
Compromised toolchain – Training data privacy



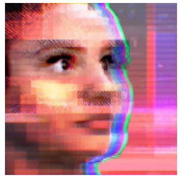
Violate privacy

Song et al. - *Machine Learning models that remember too much*, ACM CCS '17 (https://arxiv.org/abs/1709.07886)
Hitja et al. - *Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning*, ACM CCS '17 (http://arxiv.org/abs/1702.07464)

Malicious data owner – Model integrity



Influence ML model (model poisoning)



https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot
https://www.theguardian.com/technology/2017/nov/07/youtube-accused-violence-against-young-children-kids-content-google-pre-school-abuse

Is malicious adversarial behaviour the only concern?

BBC Sign in | Home | News | Sport | Reel | Worklife | Tra

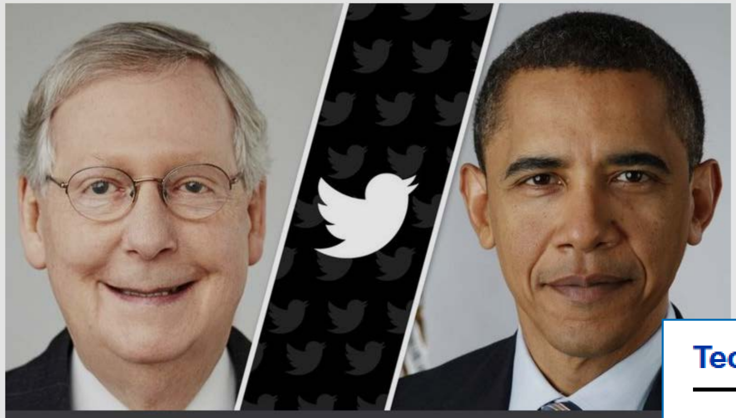
NEWS

Home | US Election | Coronavirus | Video | World | UK | Business | Tech | Science | Stories | Entertainment &

Tech

Twitter investigates racial bias in image previews

19 hours ago



One user found that Twitter seemed to favour showing Mitch McConnell's face over Barack Obama's

https://www.bbc.com/news/technology-54234822?fbclid=IwAR1T41_HR6lIuMKGRJbJdDrdpKdy Ai5mhQSdzs0QLDso41T-SR3wJfs

MIT Technology Review | Topics

Artificial intelligence

Predictive policing algorithms are racist. They need to be dismantled.

Lack of transparency and biased training data mean these tools are not fit for purpose. If we can't fix them, we should ditch them.

by **Will Douglas Heaven** July 17, 2020

.com/2020/07/17/1005396/predictive-policing-machine-learning-bias-criminal-justice/

Tech policy / AI Ethics

AI is sending people to jail—and getting it wrong

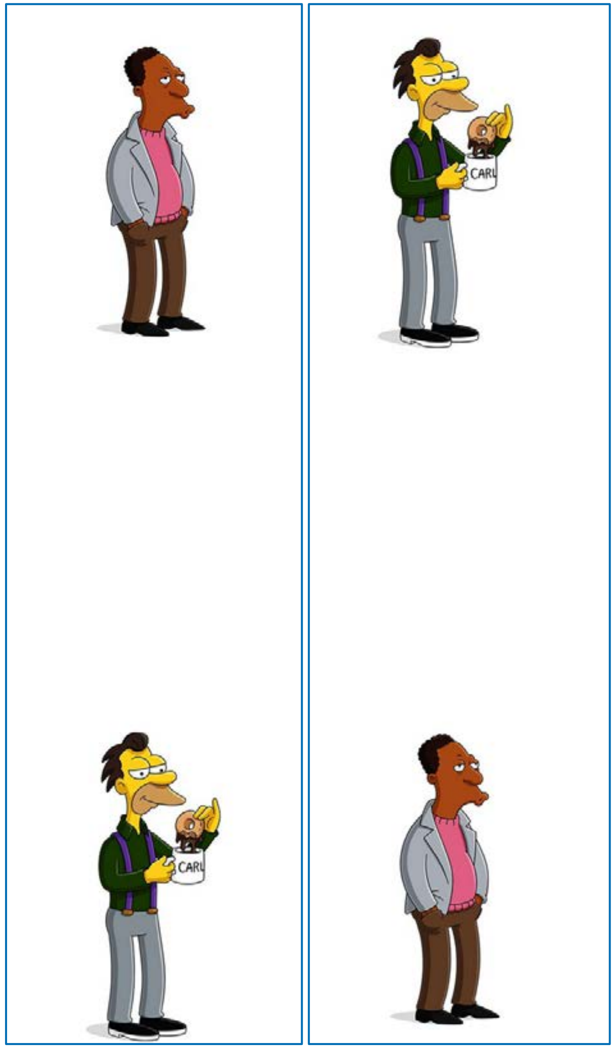
Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao** January 21, 2019

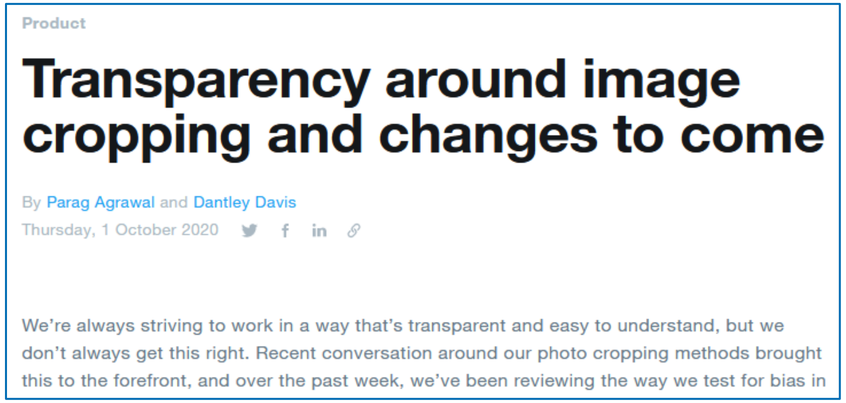
https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai /

Measures of accuracy are flawed, too

https://twitter.com/_jsimonovski/status/1307542747197239296



https://twitter.com/TwitterComms/status/1307739940424359936



https://blog.twitter.com/official/en_us/topics/product/2020/transparency-image-cropping.html

Summary: trustworthy AI systems

Security concerns

Privacy concerns

Ethical and legal concerns

Trustworthy AI: Meet these criteria even in the presence of "adversarial" behaviour

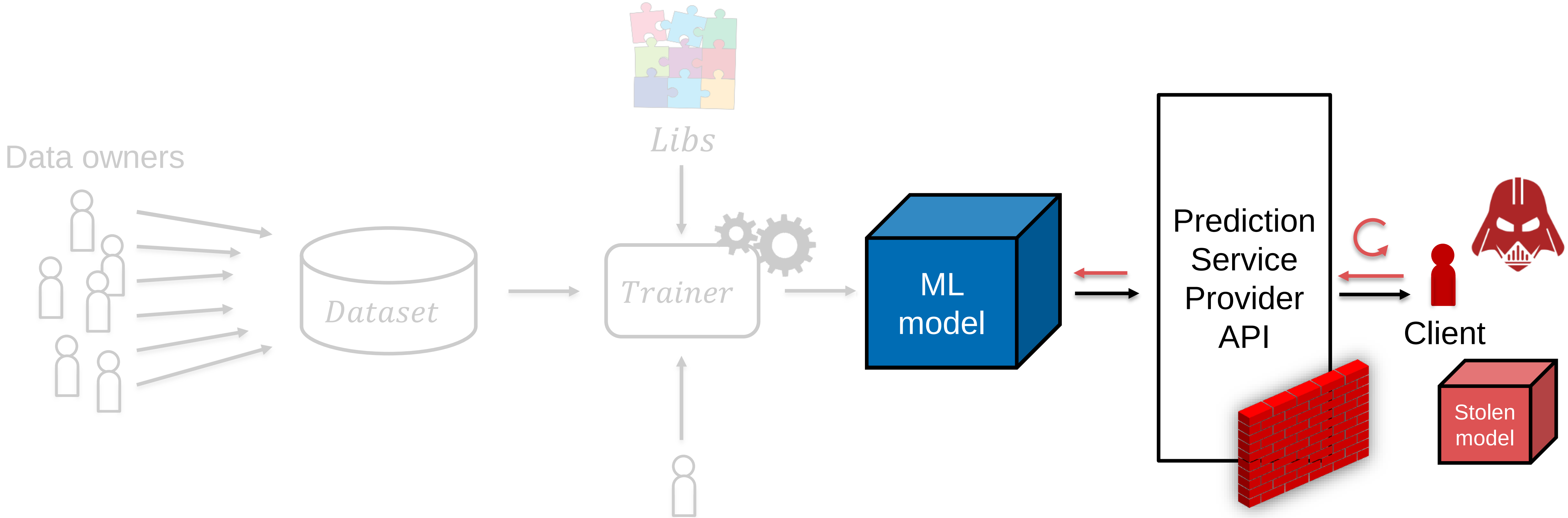
More on our research at https://crysp.uwaterloo.ca/research/SSG/

Extraction of Complex DNN Models: Real Threat or Boogeyman?

N. Asokan
https://asokan.org/asokan/
@nasokan

(Joint work with Buse Gul Atli, Sebastian Szyller, Mika Juuti and Samuel Marchal)

Malicious client – Model confidentiality



Extract/steal model

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)
Tramer et al. - *Stealing ML models via prediction APIs*, Usenix SEC '16 (https://arxiv.org/abs/1609.02943)

Is model confidentiality important?

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: white box

White-box access: user

- has physical access to model
- knows its structure
- can observe execution (scientific packages, software on user-owned devices)

How to prevent (white-box) model theft?

White-box model theft can be countered by

- **Computation with encrypted models**
- **Protecting models using secure hardware**
- **Hosting models behind a firewalled cloud service**

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 the model itself, expose model functionality only via a prediction API

Is that enough to prevent model theft?

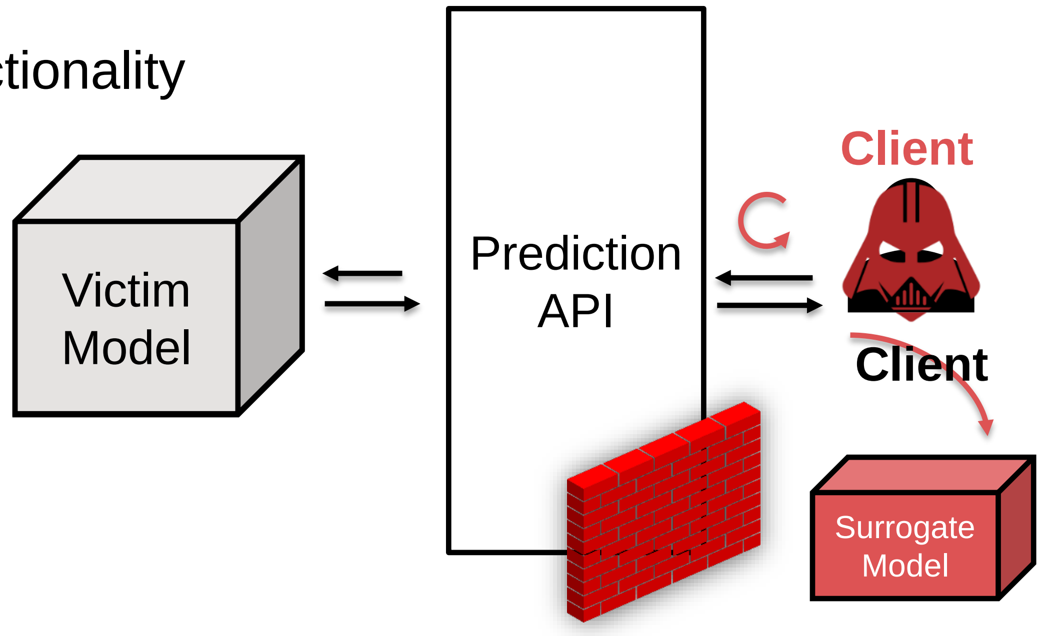
Extracting models via their prediction APIs

Prediction APIs are oracles that leak information

Adversary

- Malicious client
- Goal: construct surrogate model(*) comparable w/ functionality
- Capability: access to prediction API or model outputs

(*) aka "student model" or "imitation model"



Prior work on extracting

- Logistic regression, decision trees[1]
- Simple CNN models[2]
- Querying API with synthetic samples

[1] Tramèr et al. - *Stealing Machine Learning Models via Prediction APIs*, USENIX SEC '16 (https://arxiv.org/abs/1609.02943)
[2] Papernot et al. - *Practical Black-Box Attacks against Machine Learning*, ASIACCS '17 (https://arxiv.org/abs/1602.02697)

Extracting deep neural networks

Against simple DNN models[1]

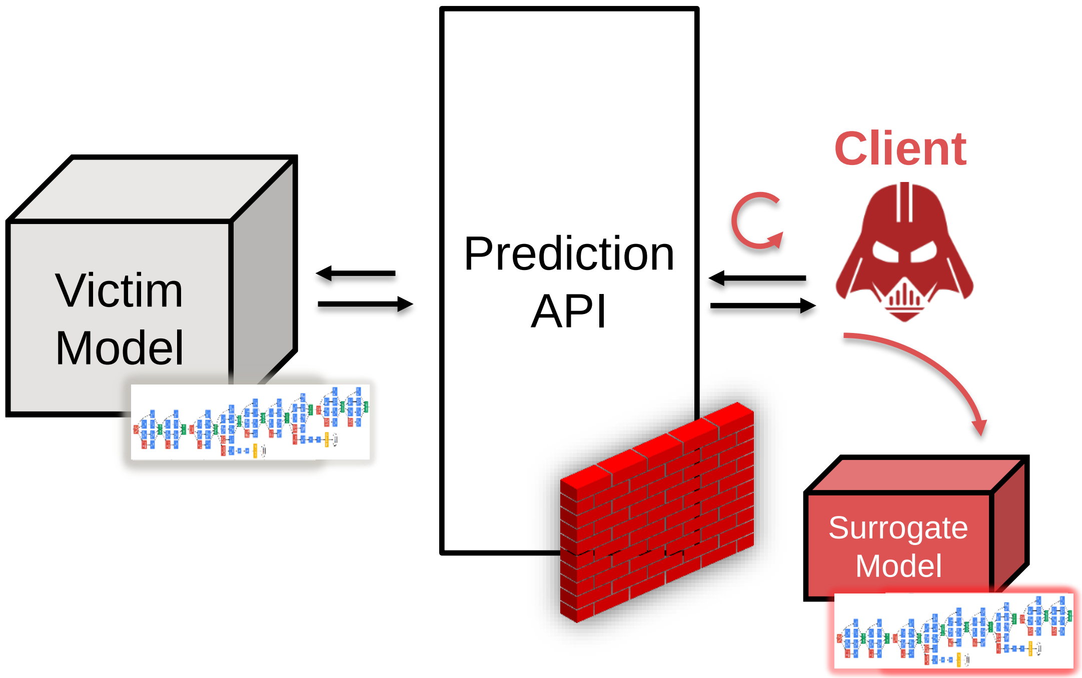
- E.g., MNIST, GTSRB

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



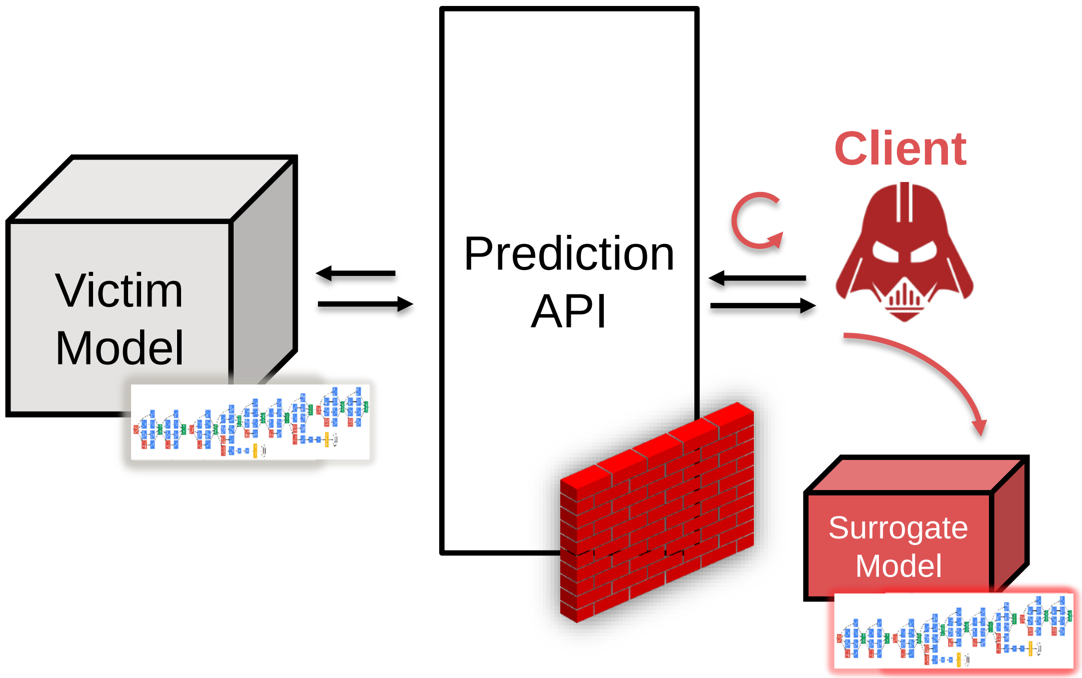
[1] Juuti et al. - *PRADA: Protecting against DNN Model Stealing Attacks*, EuroS&P '19 (https://arxiv.org/abs/1805.02628)

Is model extraction a realistic threat?

Can adversaries extract complex DNNs successfully?

Are common adversary models realistic?

Are current defenses effective?



Extraction of Complex DNN Models: Knockoff nets[1]

Goal:

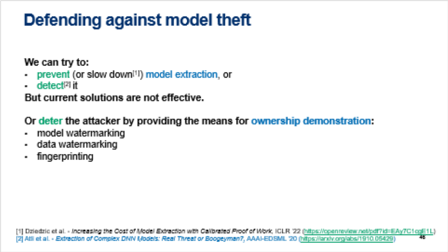
- Build a surrogate model that
  - steals model functionality of victim model
  - performs similarly on the same task with high classification accuracy

Adversary capabilities:

- Victim model knowledge:
  - None of train/test data, model internals, output semantics
  - Access to full prediction probability vector
- Access to natural samples, not (necessarily) from the same distribution as train/test data
- Access to pre-trained high-capacity model

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

Analysis of Knockoff Nets: summary[2]



Reproduced empirical evaluation of Knockoff nets[1] to confirm its effectiveness

Revisited its adversary model in to make more realistic assumptions about the adversary

Attack effectiveness decreases if

- **Surrogate and victim model architectures are different**
- **Victim model's prediction API has reduced granularity**

Defense effectiveness decreases: Attacker has natural samples distributed like victim's training data

[1] Orekondy et al. - *Knockoff Nets: Stealing Functionality of Black-Box Models,* CVPR '19 (https://arxiv.org/abs/1812.02766)
[2] Atli et al. - *Extraction of Complex DNN Models: Real Threat or Boogeyman?,* AAAI-EDSML '20 (https://arxiv.org/abs/1910.05429)

Extracting NLP Transformer models

Techniques for extracting image classifiers don't always extend to NLP models

Transfer learning from pre-trained models is now very popular

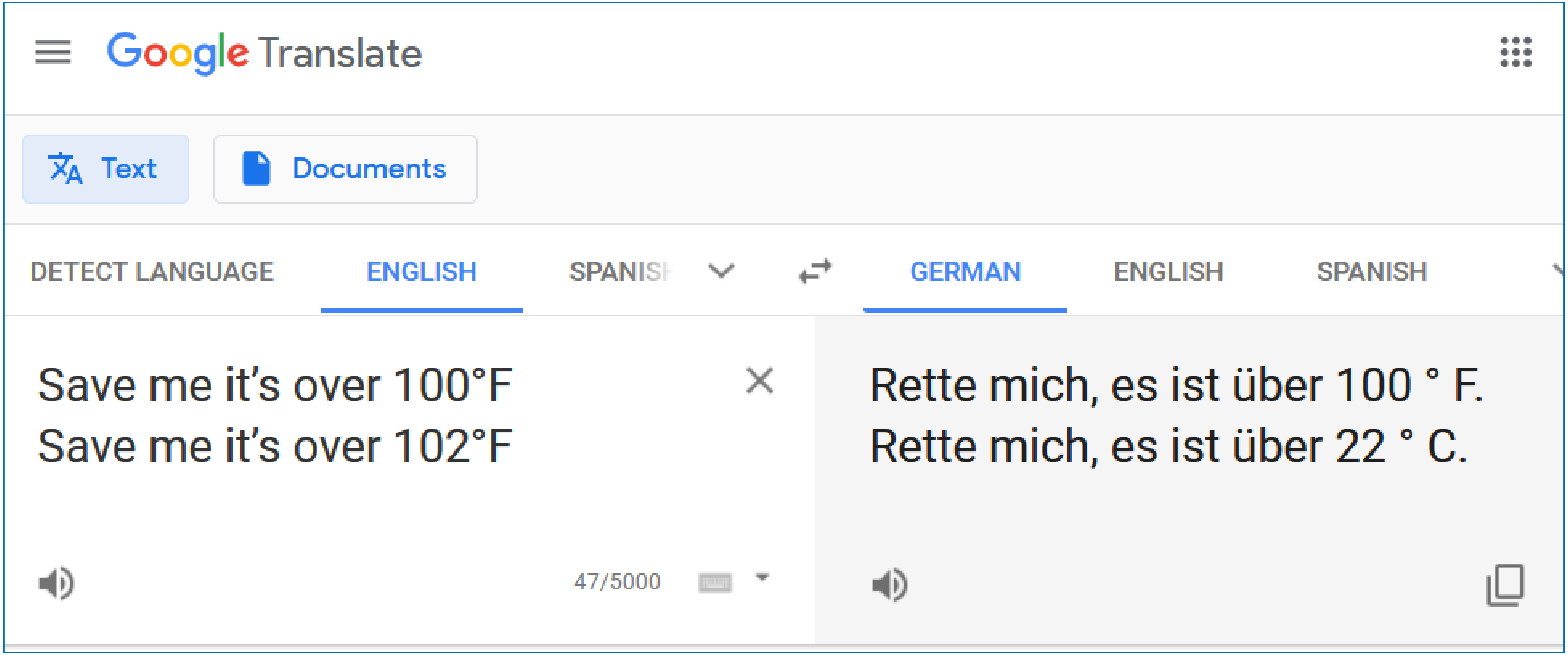
- But they make model extraction easier[1]

Krishna et al[1] show that a Knockoff-like attacks against BERT models are feasible

- Adversary unaware of target distribution or task of victim model
- Adversary queries are merely "natural" (randomly sampled sequences of words)
- In-distribution adversary queries can improve extraction efficacy

Wallace et al[2] extract real-world MT models, find transferable adversarial examples

[1] Krishna et al. – *Thieves on Sesame Street! Model Extraction of BERT-based APIs* , ICLR '20 (https://iclr.cc/virtual_2020/poster_Byl5NREFDr.html)
[2] Wallace et al. – *Imitation Attacks and Defenses for Black-box Machine Translation Systems,* EMNLP '20 (https://arxiv.org/abs/2004.15015)

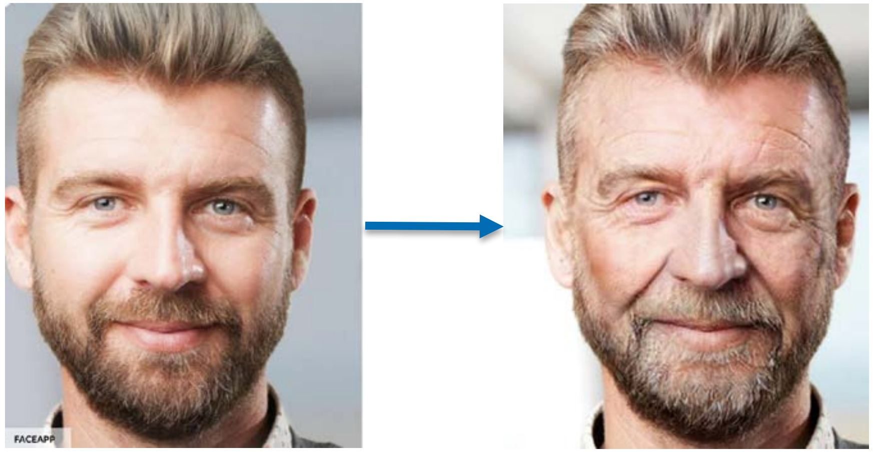


https://translate.google.com/#view=home&op=translate&sl=en&tl=de&text=Save%20me%20it%E2%80%99s%20over%20100%C2%B0F%0ASave%20me%20it%E2%80%99s%20over%20102%C2%B0F

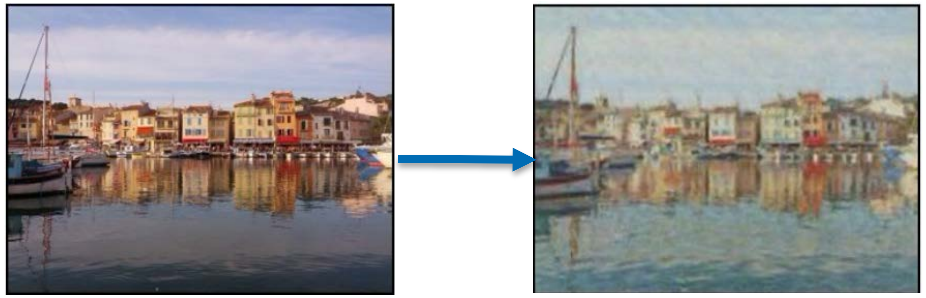
Wallace et al. – *Imitation Attacks and Defenses for Black-box Machine Translation Systems,* EMNLP ‘20 (https://arxiv.org/abs/2004.15015)

Extracting Style-transfer models

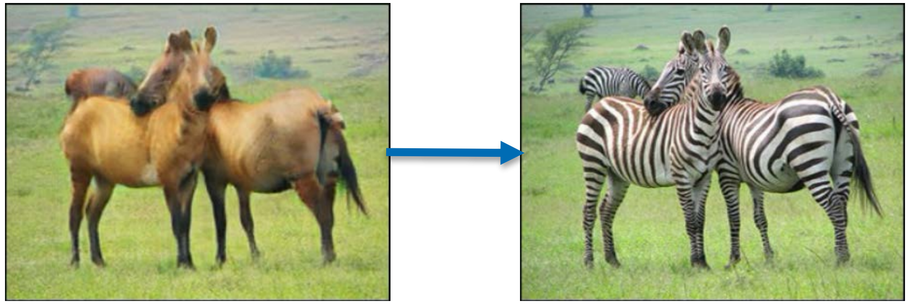
- GANS are effective for changing image style
  - coloring, face filters, style application
- Core feature in generative art and in social media apps
  - Selfie2Anime, FaceApp



FaceApp



CycleGANs



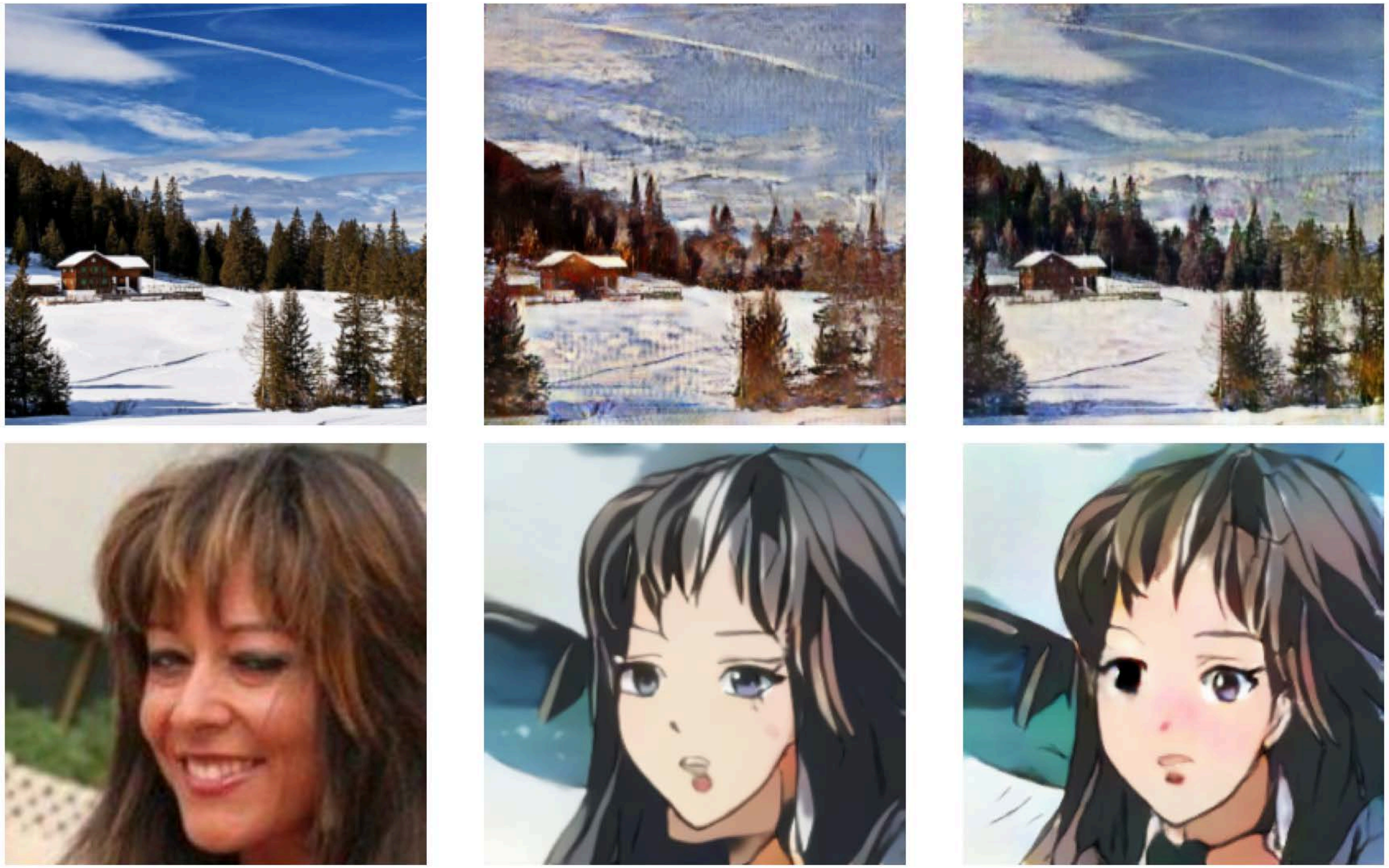
CycleGANs

Style transfer

Original (unstyled) | Styled (victim) | **Styled (ours)**

Task 1 *Monet painting*

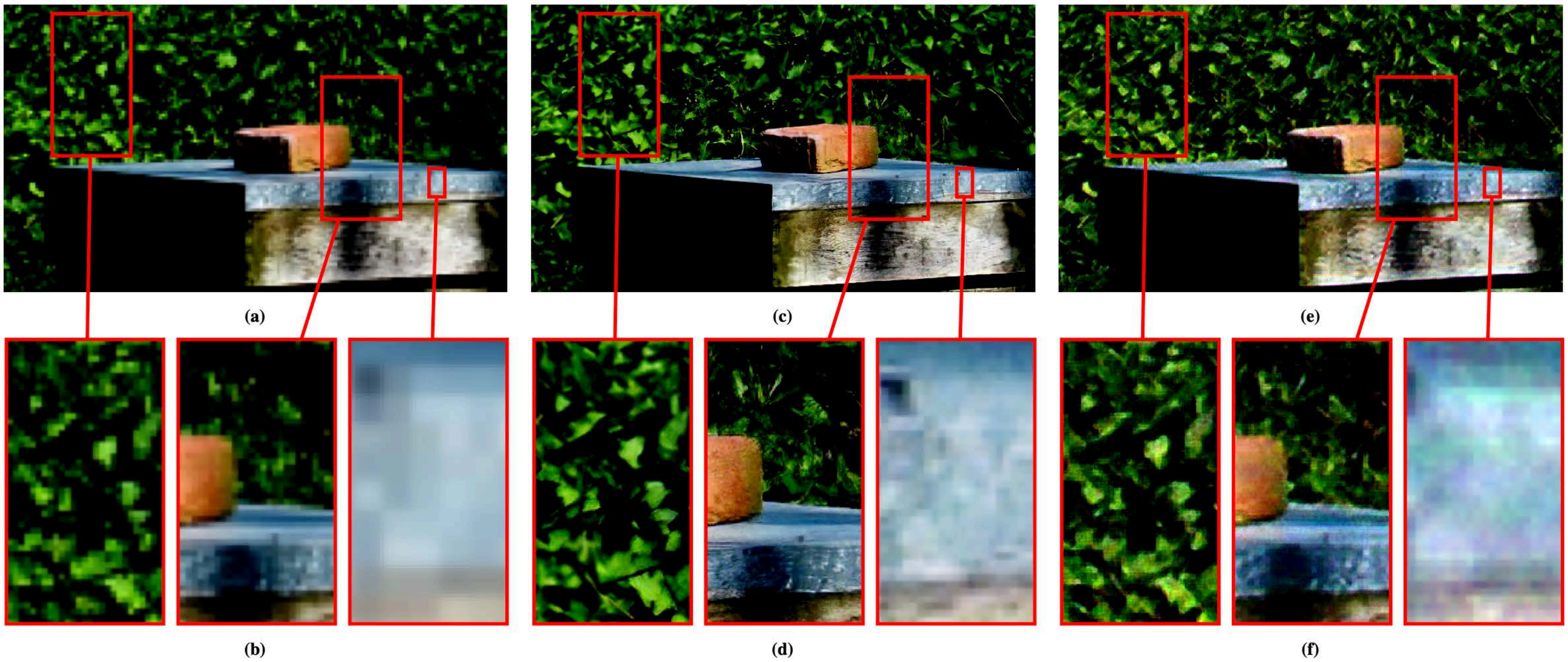
Task 2 *Anime face*



Szyller et al. - *Good Artists Copy, Great Artists Steal: Model Extraction Attacks Against Image Translation Generative Adversarial Networks*, https://arxiv.org/abs/2104.12623

Super resolution

Original (low-res)	High-res (victim)	**High-res (ours)**



Szyller et al. - *Good Artists Copy, Great Artists Steal: Model Extraction Attacks Against Image Translation Generative Adversarial Networks*, https://arxiv.org/abs/2104.12623

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 the attacker by providing the means for ownership demonstration:

- 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)
[2] Atli et al. - *Extraction of Complex DNN Models: Real Threat or Boogeyman?,* AAAI-EDSML '20 (https://arxiv.org/abs/1910.05429)

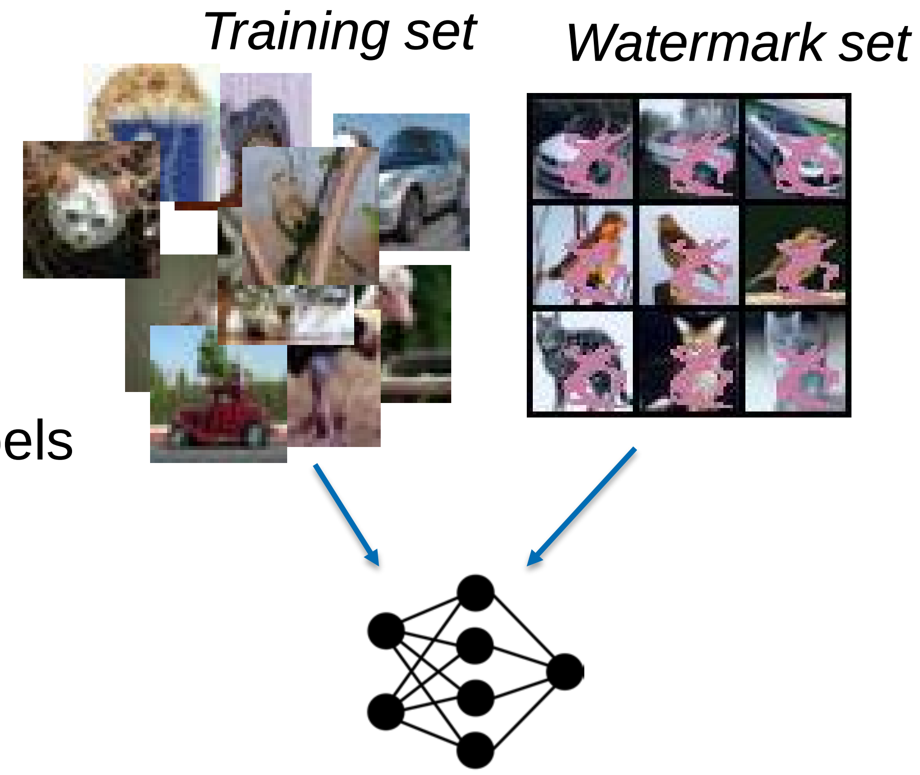
White-box watermarking

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

Existing watermarking of DNNs

Assumes that the model is stolen exactly (white-box theft)

Protects only against physical theft of model[1]

Not robust against

- novel watermark removal attacks[2]
- model extraction attacks that reduce the effect of watermarks & modify decision surface

[1] Szyller et. al. - *DAWN: Dynamic Adversarial Watermarking of Neural Networks.* ACM MM '21 (https://arxiv.org/abs/1906.00830)
[2] Lukas et al. *SoK: How Robust is Image Classification Deep Neural Network Watermarking*? IEEE S&P '22 (https://arxiv.org/abs/2108.04974)

DAWN: Dynamic Adversarial Watermarking of DNNs[1]

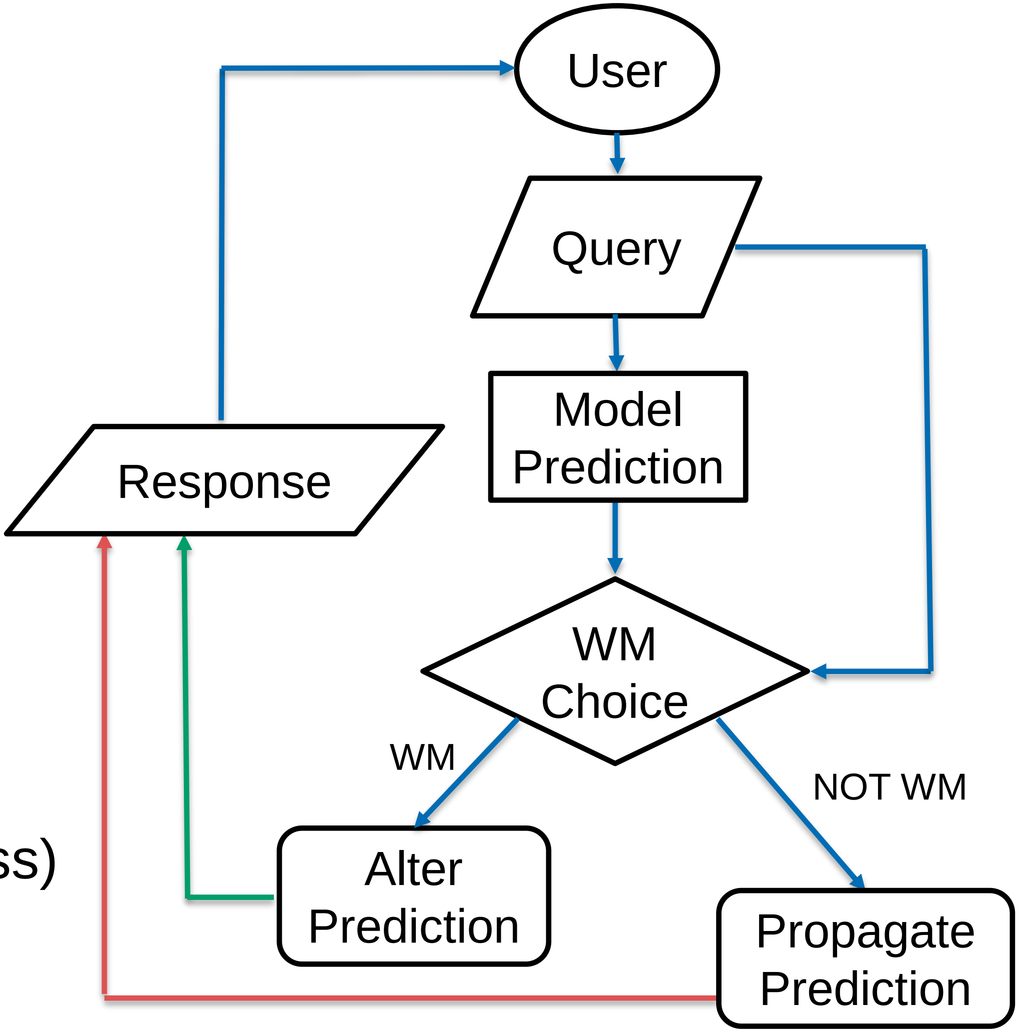
Goal: Watermark models obtained via model extraction

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

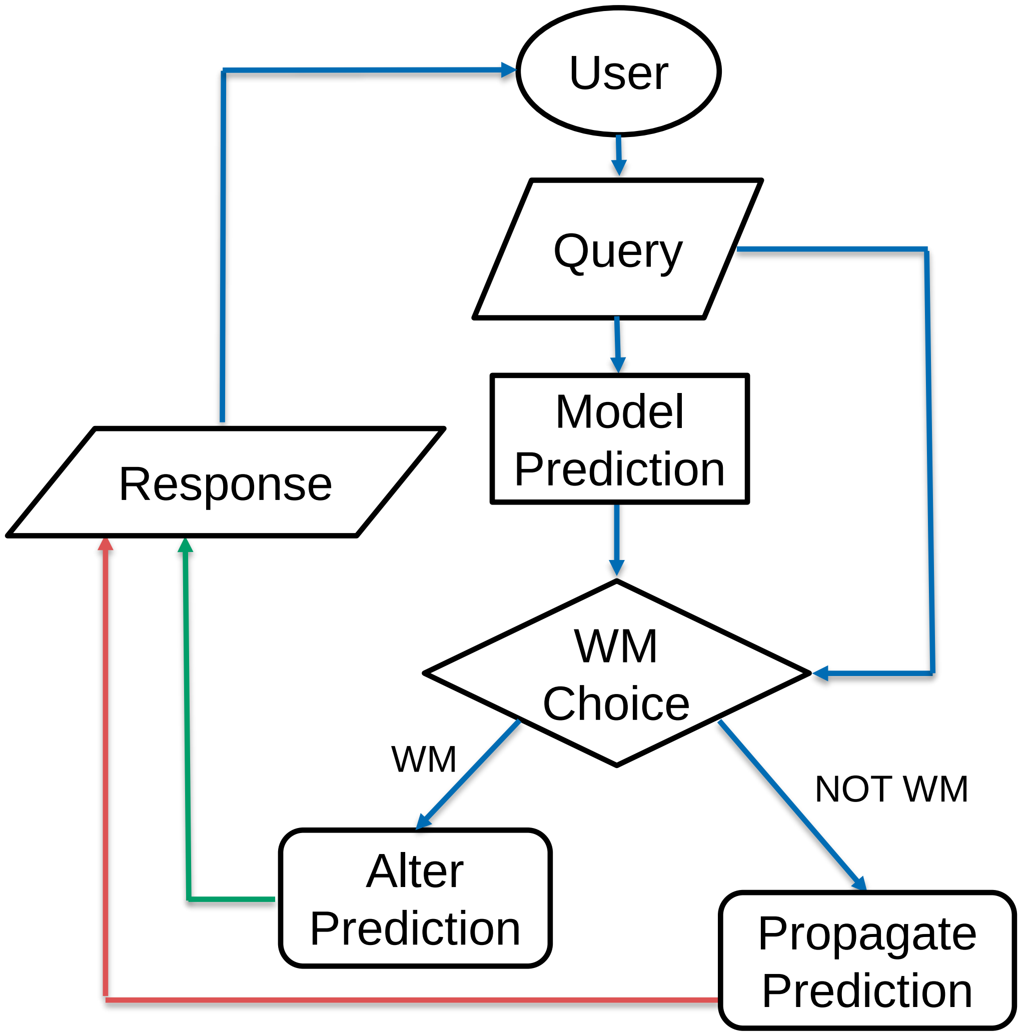
Open issues in DAWN[1]

Indistinguishability

- existence of a robust mapping function (for WM choice)

Unremovability

- “double-stealing” can remove watermark (but impacts accuracy of surrogate model)
- adversary can try to return incorrect predictions on training data (but can be overcome)



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

Data/Model fingerprinting

Radioactive data[1]

- Intended for provenance, not robust in adversarial settings[2]

Conferrable adversarial examples[2]

- Computationally expensive

Dataset inference[3]

- Susceptible to False positives?

[1] Sablayrolles et al. *Radioactive data: tracing through training,* ICML'20 (https://arxiv.org/abs/2002.00937)
[2] Atli Tegkul et al. *On the Effectiveness of Dataset Watermarking,* IWSPA@CODASPY '22 (https://arxiv.org/abs/2106.08746)
[2] Lukas et al. *Deep Neural Network Fingerprinting by Conferrable Adversarial Examples,* ICLR '21 (https://openreview.net/forum?id=VqzVhqxkjH1)
[3] Maini, et al. *Dataset Inference Ownership Resolution in Machine Learning*, ICLR '21 (https://openreview.net/pdf?id=hvdKKV2yt7T)

Summary: ML Model extraction

Complex DNN models can be extracted

Adversary models should match the application setting

No generally applicable defenses yet

More on our model extraction work at https://ssg.aalto.fi/research/projects/mlsec/model-extraction/

UNIVERSITY OF WATERLOO

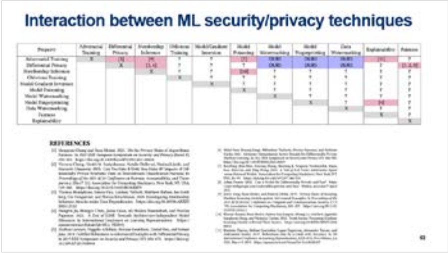
Conflicts Between ML Security/Privacy Techniques

Sebastian Szyller, N. Asokan

https://sebszyller.com https://asokan.org/asokan/

@sebszyller @nasokan

Other ML security & privacy concerns



There are considerations other than model ownership:

- model evasion (defense: adversarial training)
- training data reconstruction (defense: differential privacy)
- membership inference (defense: regularization, early stopping)
- model poisoning (defense: regularization, outlier/anomaly detection)
- …

How does ownership demonstration interact with the other defenses?

We investigate pairwise interactions of:

model watermarking data watermarking fingerprinting	**WITH**	differential privacy adversarial training

Setup & Baselines

We use the following techniques (and corresponding metrics):

- Out-of-distribution (OOD) backdoor watermarking (test and watermark accuracy)
- Radioactive data (test accuracy and loss difference)
- Dataset Inference (verification confidence)
- DP-SGD (model accuracy for the given epsilon)
- Adversarial training with PGD (test and adv. accuracy for the given epsilon)

Dataset	No defense	Watermarking		Radioactive Data		Dataset Inference	DP-SGD (eps=3)	ADV. TR.	
	TEST	TEST	WM	TEST	Loss. Diff.	Confidence	TEST	TEST	ADV.
MNIST	*0.99*	*0.99*	*0.97*	*0.98*	*0.284*	*<e-30*	*0.98*	*0.99*	*0.95*
FMNIST	*0.91*	*0.87*	*0.99*	*0.88*	*0.19*	*<e-30*	*0.86*	*0.87*	*0.69*
CIFAR10	*0.92*	*0.82*	*0.97*	*0.85*	*0.2*	*<e-30*	*0.38*	*0.82*	*0.82*

Interaction with differential privacy

Dataset	DP-SGD (eps=3)
	TEST
MNIST	*0.98*
FMNIST	*0.86*
CIFAR10	*0.38*

Differential privacy is a strong per-sample regulariser:

- Watermarking rendered ineffective
- Lower but still sufficient confidence for radioactive data
- No effect on the DI fingerprint

Dataset	No defense	Watermarking				Radioactive Data				Dataset Inference	
		Baseline		with DP		Baseline		with DP		Baseline	with DP
	TEST.	TEST	WM	TEST	WM	TEST	Loss. Diff.	TEST	Loss. Diff.	Conf.	Conf.
MNIST	*0.99*	*0.99*	*0.97*	*0.97*	*0.30*	*0.98*	*0.284*	*0.97*	*0.091*	*<e-30*	*<e-30*
FMNIST	*0.91*	*0.87*	*0.99*	*0.86*	*0.28*	*0.85*	*0.19*	*0.84*	*0.11*	*<e-30*	*<e-30*
CIFAR10	*0.92*	*0.82*	*0.97*	*0.38*	*0.12*	*0.85*	*0.2*	*0.35*	*0.19*	*<e-30*	*<e-30*

Interaction with DP (tweaks and relaxations)

Tweaking DP-SGD:

- Naively increasing eps (less noise) does not improve WM accuracy
- Increasing gradient clipping threshold is better (not sufficient)

Tweaking the watermark:

- Bigger trigger set gives better WM accuracy (not sufficient)
- Training longer is better (not sufficient)

With strict DP-SGD, OOD backdoor watermarking does not work.

What if we relax DP-SGD?

- Splitting the training into the DP part (genuine data) and non-DP (watermark) helps
- Watermark is embedded successfully (accuracy > 0.9)
- Privacy loss analysis is not tight anymore

Interaction with adversarial training

Adversarial training creates a robust L_p bubble:

- Watermarking not affected but adversarial accuracy drops
- Significant drop in the confidence of radioactive data
- No effect on the DI fingerprint

Dataset	ADV. TR.	
	TEST	ADV.
MNIST	*0.99*	*0.95*
FMNIST	*0.87*	*0.69*
CIFAR10	*0.82*	*0.82*

Dataset	No defense	Watermarking					Radioactive Data					DI	
		Baseline		with ADV. TR.			Baseline		with ADV. TR.			Baseline	with ADV. TR.
	TEST	TEST	WM	TEST	WM	ADV	TEST	Loss. Diff.	TEST	Loss. Diff.	ADV	Conf.	Conf.
MNIST	*0.99*	*0.99*	*0.97*	*0.97*	*0.99*	*0.88*	*0.98*	*0.284*	*0.97*	*0.001*	*0.95*	*<e-30*	*<e-30*
FMNIST	*0.91*	*0.87*	*0.99*	*0.86*	*0.99*	*0.51*	*0.85*	*0.19*	*0.84*	*0.0007*	*0.69*	*<e-30*	*<e-30*
CIFAR10	*0.92*	*0.82*	*0.97*	*0.78*	*0.97*	*0.65*	*0.85*	*0.2*	*0.81*	*0.003*	*0.81*	*<e-30*	*<e-30*

False positives in Dataset Inference 1/2

We noticed false positives when DI is combined with other defenses:

- models would trigger confident FPs w.r.t. unrelated models (e.g. MNIST to FMNIST)
- But we saw FPs even in our DI baseline (i.e., without other defenses)

We revisited the original[1] DI itself (CIFAR10):

- use the implementation from the official repo[2]
- Models provided in the repo work as intended
- We trained many independent models:
  - Without any other defense
  - We can reproduce the results from the paper, however...

[1] Maini, et al. *Dataset Inference Ownership Resolution in Machine Learning*, ICLR '21 (https://openreview.net/pdf?id=hvdKKV2yt7T)
[2] Dataset Inference GitHub repository (https://github.com/cleverhans-lab/dataset-inference)

False positives in Dataset Inference 2/2

We revisited the original[1] DI itself (CIFAR10):

- **The original split for CIFAR10 uses:**
  - the training set for the teacher model
  - the test set to train the independent model
  - the test set and the training set are used for the distinguisher (double-dip on the test set)

- **We split CIFAR10 training set into two non-overlapping chunks (A and B):**
  - one for the teacher (A), one for the independent model (B)
  - the test and the A set are used for the distinguisher
  - independent model B triggers a FP with high confidence

Model trained on:	Verification p-value
A (teacher)	*e-23*
Test (original)	*0.1*
B (independent)	*e-12*
A+B	*e-13*

Interaction between ML security/privacy techniques

Property	Adversarial Training	Differential Privacy	Membership Inference	Oblivious Training	Model/Gradient Inversion	Model Poisoning	Model Watermarking	Model Fingerprinting	Data Watermarking	Explainability	Fairness
Adversarial Training	X	[5]	[9]	?	?	[7]	OURS	OURS	OURS	[11]	?
Differential Privacy		X	[3, 6]	?	?	?	OURS	OURS	OURS	?	[1, 2, 8]
Membership Inference			X	?	?	[10]	?	?	?	?	?
Oblivious Training				X	?	?	?	?	?	?	?
Model/Gradient Inversion					X	?	?	?	?	?	?
Model Poisoning						X	?	?	?	?	?
Model Watermarking							X	?	?	?	?
Model Fingerprinting								X	?	[4]	?
Data Watermarking									X	?	?
Fairness										X	?
Explainability											X

REFERENCES

[1] Hongyan Chang and Reza Shokri. 2021. On the Privacy Risks of Algorithmic Fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS P)*. 292–303. https://doi.org/10.1109/EuroSP51992.2021.00028

[2] Victoria Cheng, Vinith M. Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. 2021. Can You Fake It Until You Make It? Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 149–160. https://doi.org/10.1145/3442188.3445879

[3] Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. 2020. Investigating Membership Inference Attacks under Data Dependencies. https://doi.org/10.48550/ARXIV.2010.12112

[4] Hengrui Jia, Hongyu Chen, Jonas Guan, Ali Shahin Shamsabadi, and Nicolas Papernot. 2022. A Zest of LIME: Towards Architecture-Independent Model Distances. In *International Conference on Learning Representations*. https://openreview.net/forum?id=OUz_9TiTv9j

[5] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified Robustness to Adversarial Examples with Differential Privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. 656–672. https://doi.org/10.1109/SP.2019.00044

[6] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. 2021. Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning. In *2021 IEEE Symposium on Security and Privacy (SP)*. 866–882. https://doi.org/10.1109/SP40001.2021.00069

[7] Ren Pang, Hua Shen, Xinyang Zhang, Shouling Ji, Yevgeniy Vorobeychik, Xiapu Luo, Alex Liu, and Ting Wang. 2020. *A Tale of Evil Twins: Adversarial Inputs versus Poisoned Models.* Association for Computing Machinery, New York, NY, USA, 85–99. https://doi.org/10.1145/3372297.3417253

[8] Adam Pearce. 2022. Can a Model Be Differentially Private and Fair? https://pair.withgoogle.com/explorables/private-and-fair/. Online; accessed 7 April 2022.

[9] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy Risks of Securing Machine Learning Models against Adversarial Examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. Association for Computing Machinery, 241–257. https://doi.org/10.1145/3319535.3354211

[10] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. 2022. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. https://doi.org/10.48550/ARXIV.2204.00032

[11] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness May Be at Odds with Accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.* https://openreview.net/forum?id=SyxAb30cY7

Summary: Conflicts among ML protection techniques

Substantial on-going research on individual threats and protection techniques

But practitioners need to deploy multiple protection techniques in parallel

More work needed to understand conflicts among protection techniques

(Work in progress)

Overall summary

1. **Security, Privacy, and Fairness challenges need to be addressed in order to make AI-based systems trustworthy**
   - Active research area

2. **Model extraction is a real threat against ML-based systems**
   - No clear general solutions yet

3. **ML security/privacy techniques can conflict with one another**
   - Needs more active research

Open postdoc positions to help lead our work: ML security/privacy, platform security
https://asokan.org/asokan/research/SecureSystems-open-positions-Jul2021.php

Come work with us!

Open postdoc positions to help lead our work: ML security/privacy, platform security

https://asokan.org/asokan/research/SecureSystems-open-positions-Jul2021.php