Aalto University

UNIVERSITY OF WATERLOO

Model Stealing Attacks and Defenses

Where are we now?

N. Asokan

https://asokan.org/asokan/

@nasokan

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

My research interests

https://ssg-research.github.io/

Systems Security and Privacy

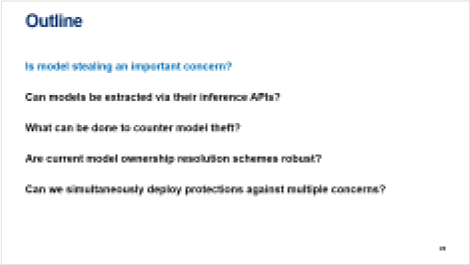
AI and Security/Privacy

- How to use AI to improve security/privacy solutions
- How to improve security/privacy of AI-based systems

Platform security

- How to use hardware assistance to secure software?

Outline



The big picture

Is model stealing an important concern?

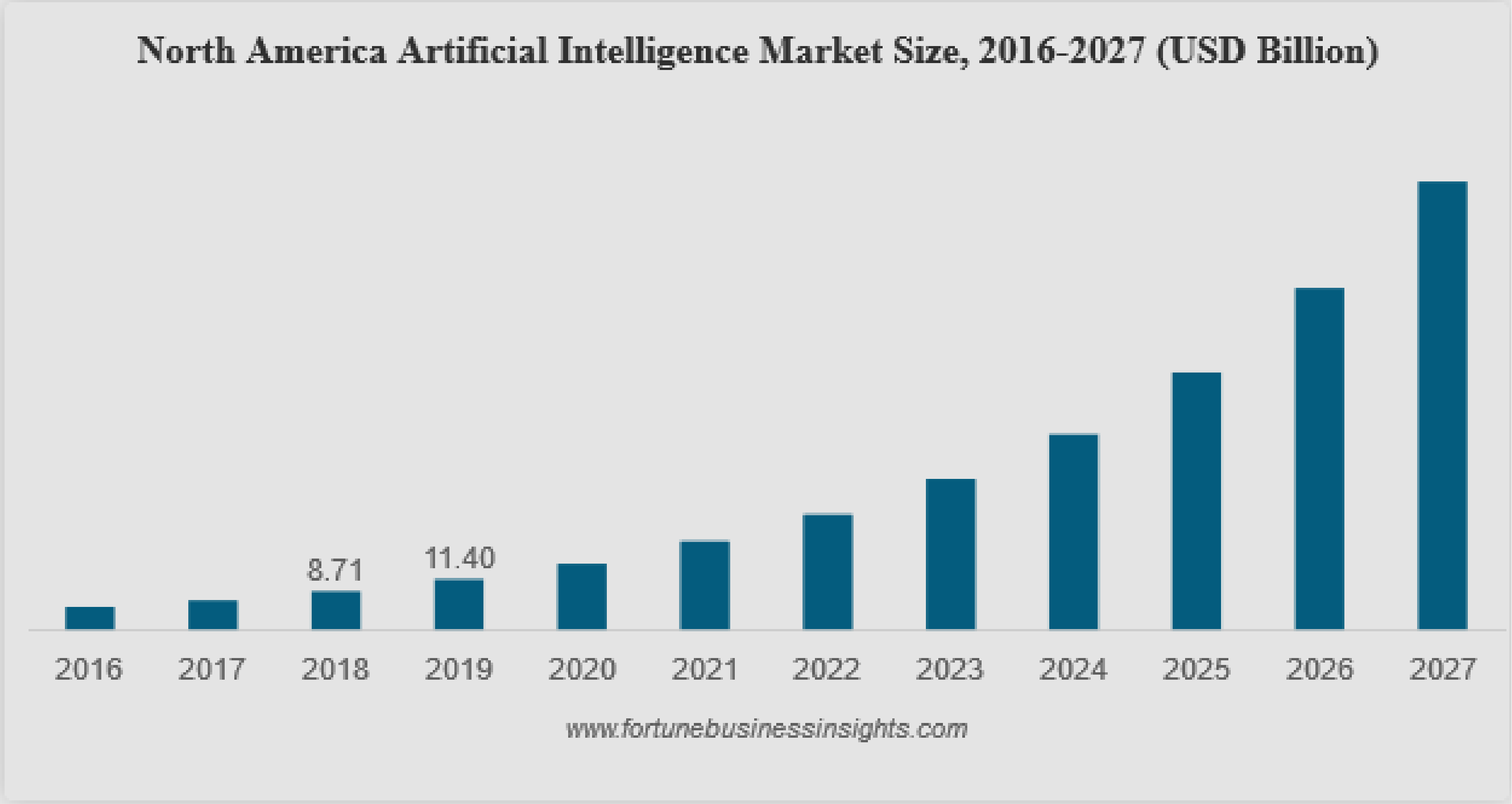
Can models be extracted via their inference APIs?

What can be done to counter model theft?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy protections against multiple concerns?

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

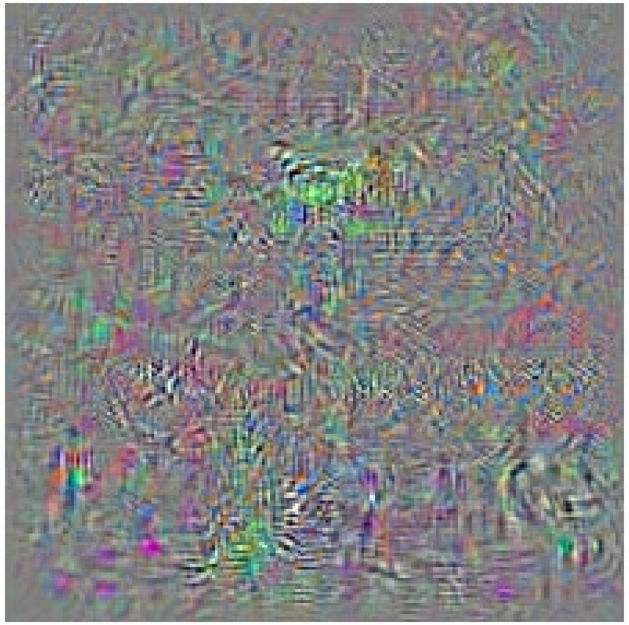
Privacy concerns

Fairness and explainability concerns

Evading machine learning models

Which class is this?
School bus

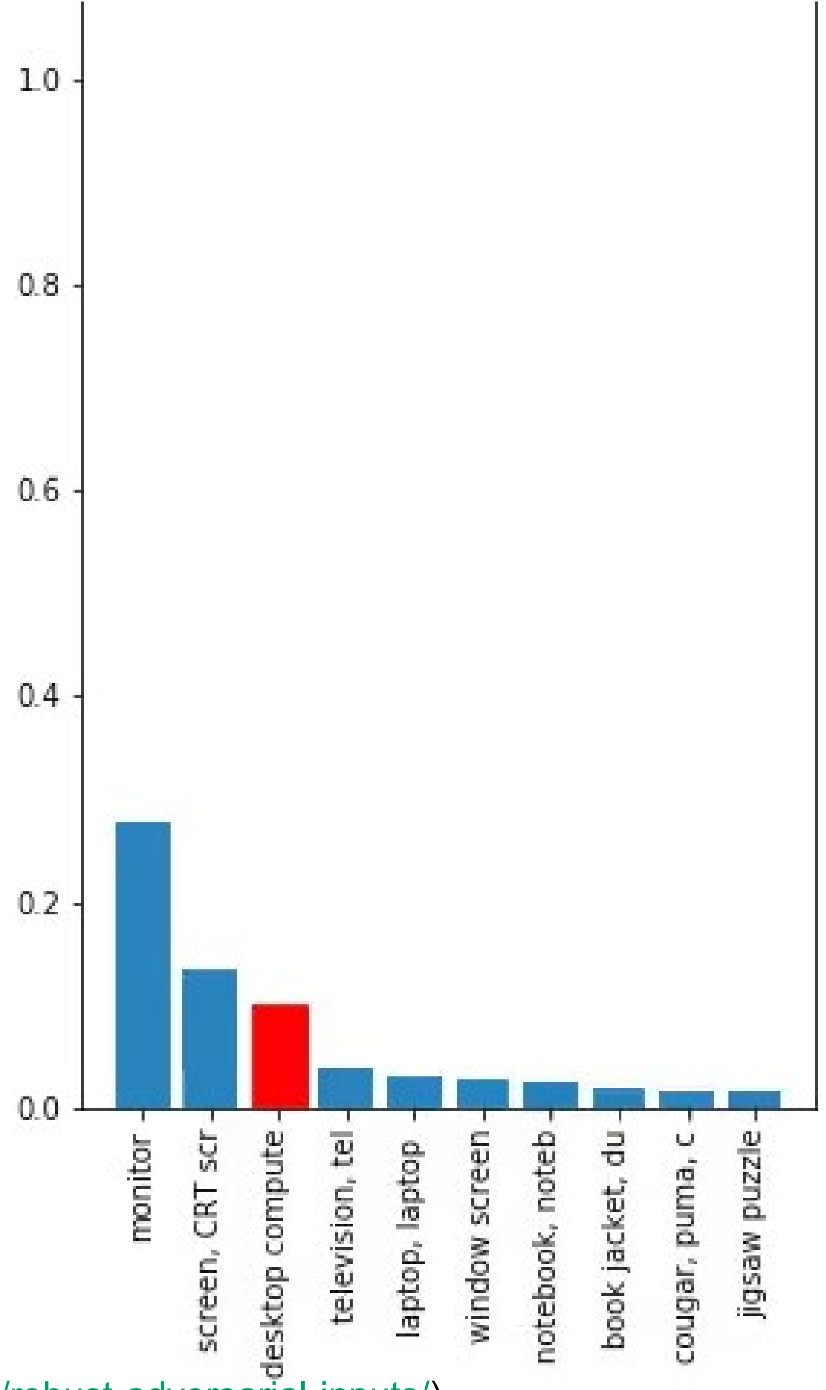
+ 0.1·



=

Which class is this?
Ostrich

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

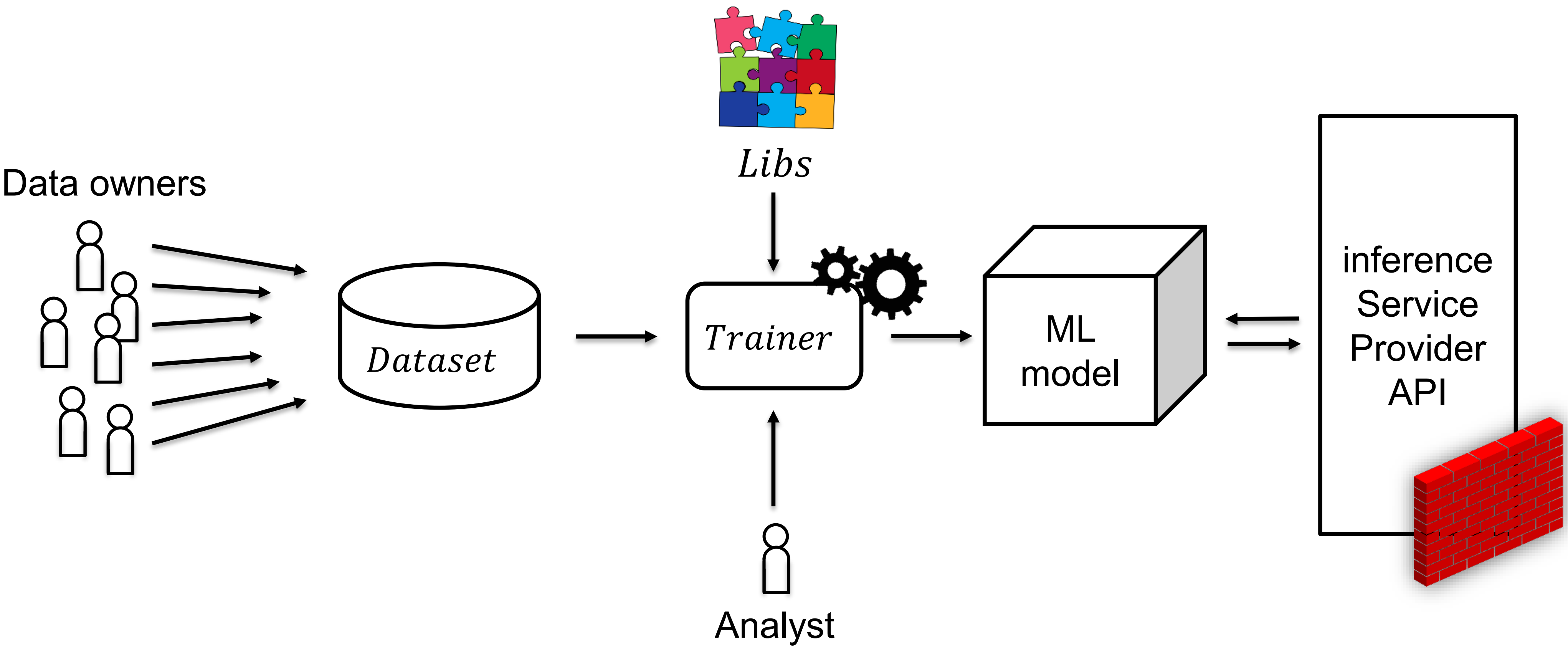


Which class is this?
Cat

Which class is this?
Desktop computer

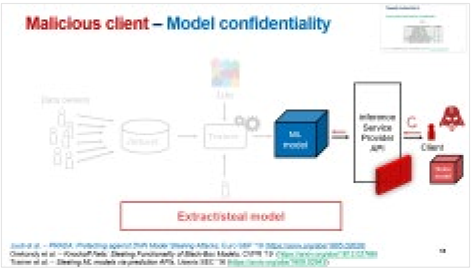
Athalye et al. – *Synthesizing Robust Adversarial Examples*, ICML '2019 (https://blog.openai.com/robust-adversarial-inputs/)

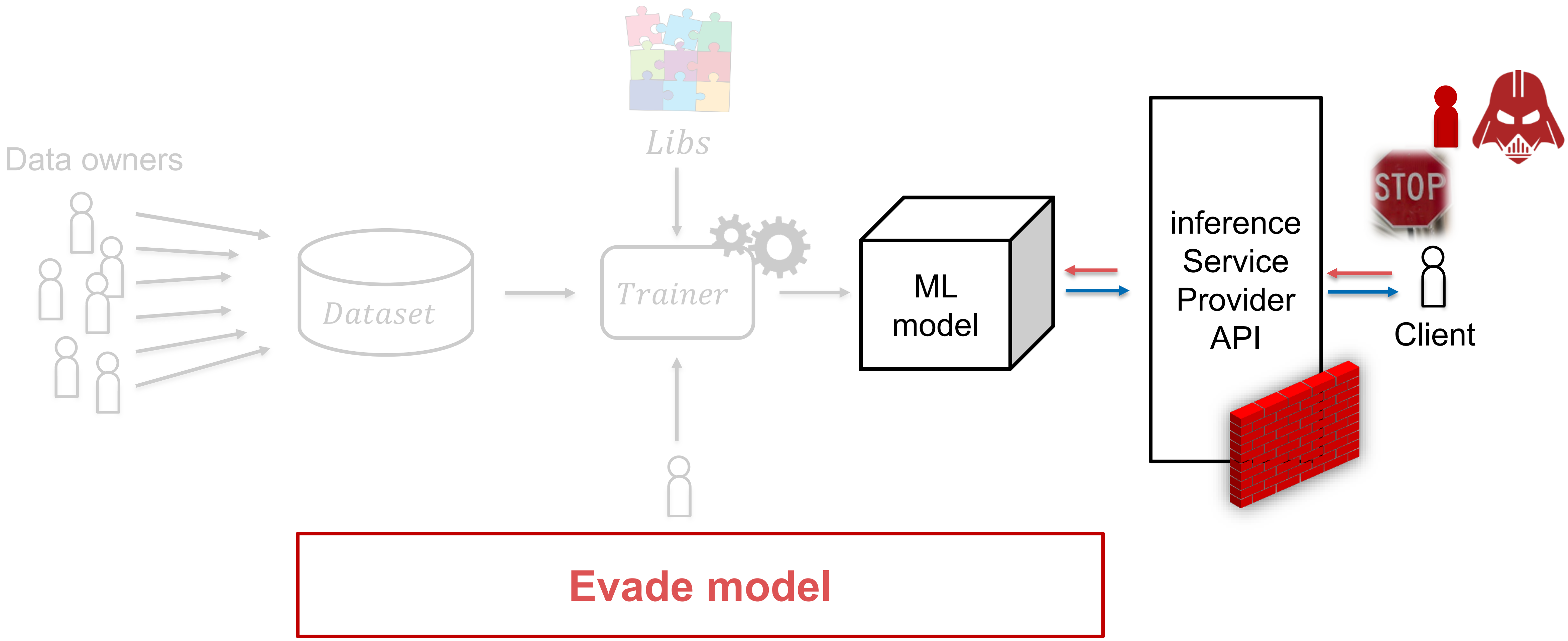
Machine Learning pipeline



Where is the adversary? What is its target?

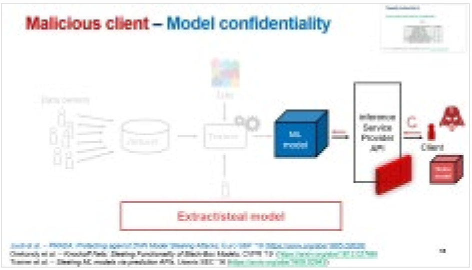
Compromised input – Model integrity

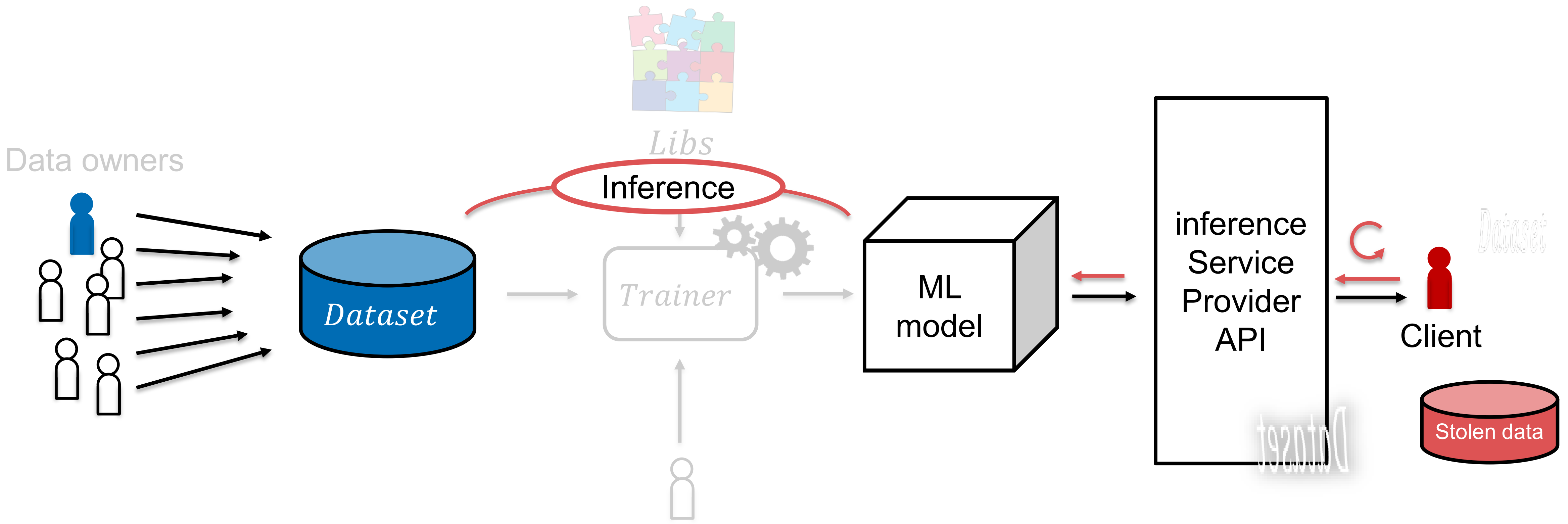




Szegedy et al. – *Intriguing Properties of Neural Networks,* ICLR '14 (https://arxiv.org/abs/1312.6199v4)
Dalvi et al. – *Adversarial Classification,* KDD '04 (https://dl.acm.org/doi/10.1145/1014052.1014066)

Malicious client – Training data privacy

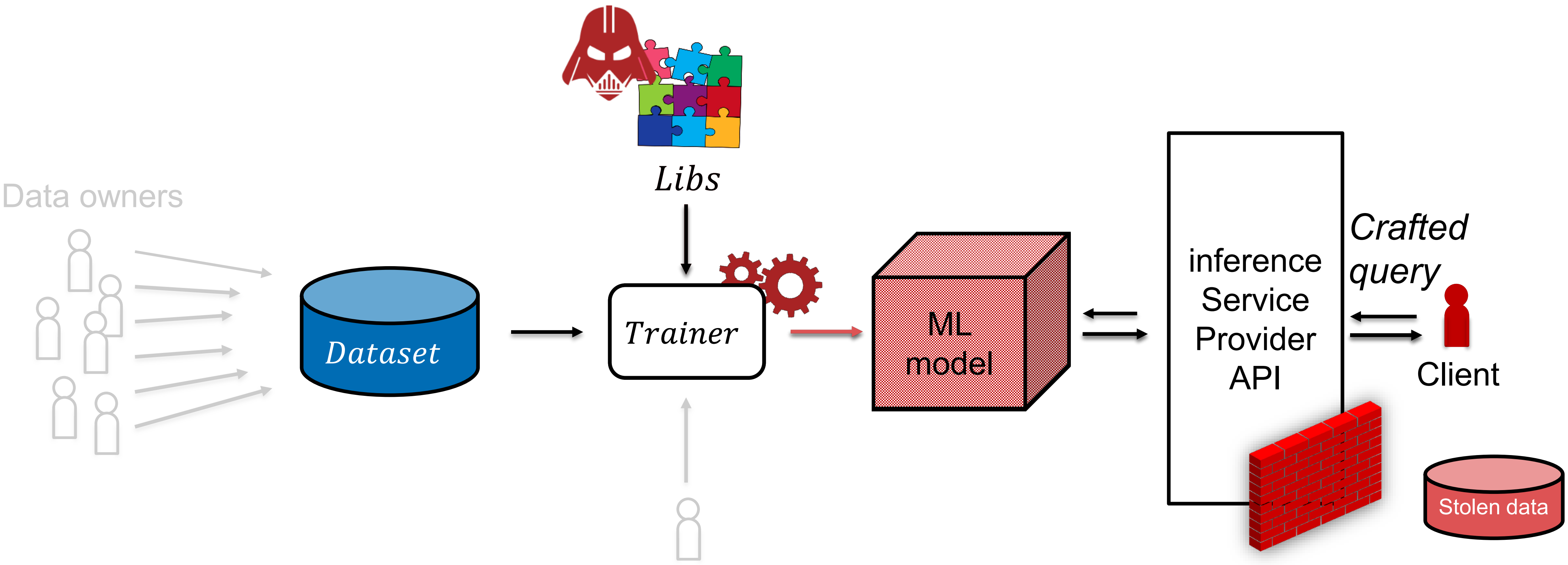




Invert model, infer membership

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

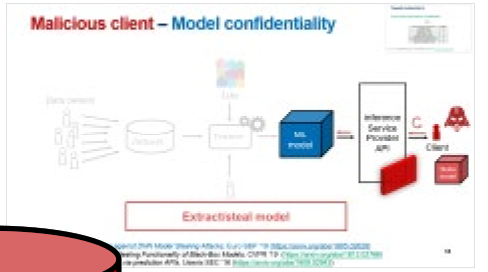
Compromised toolchain – Training data privacy

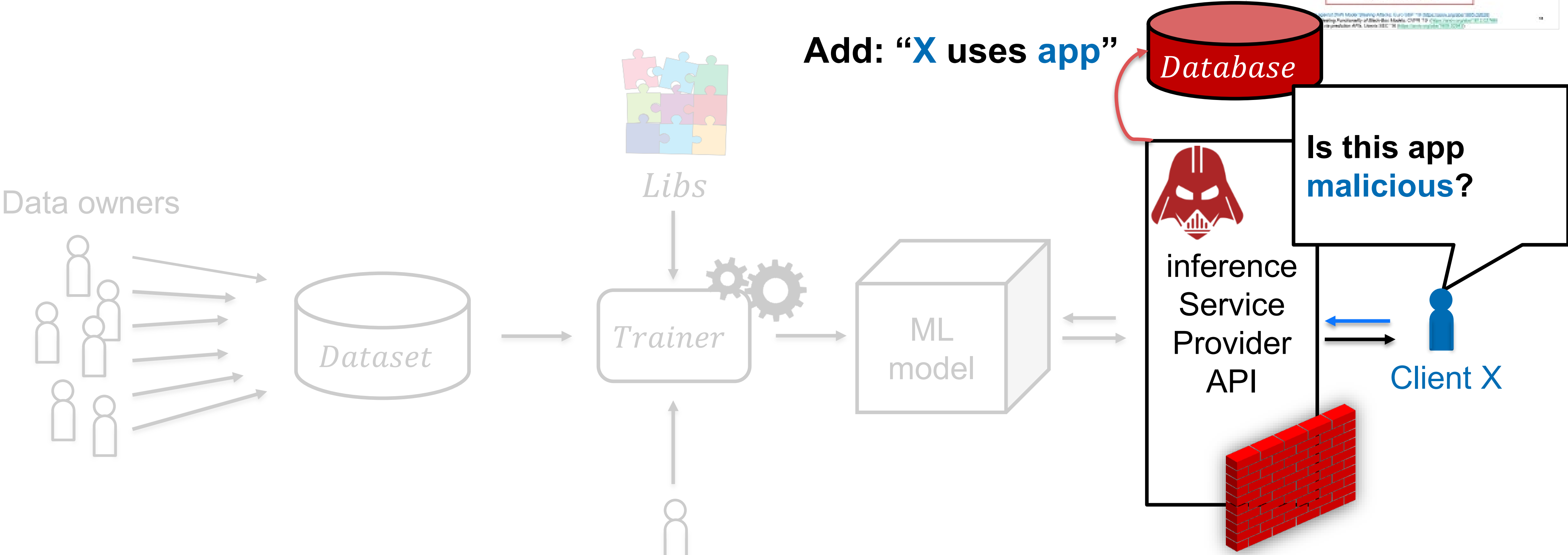


Infringe on privacy

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 inference service – User profiles

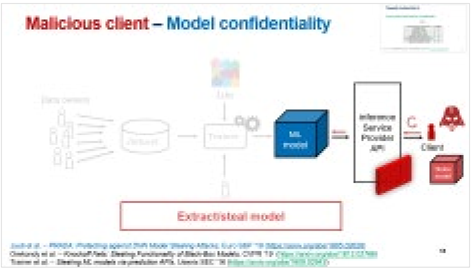


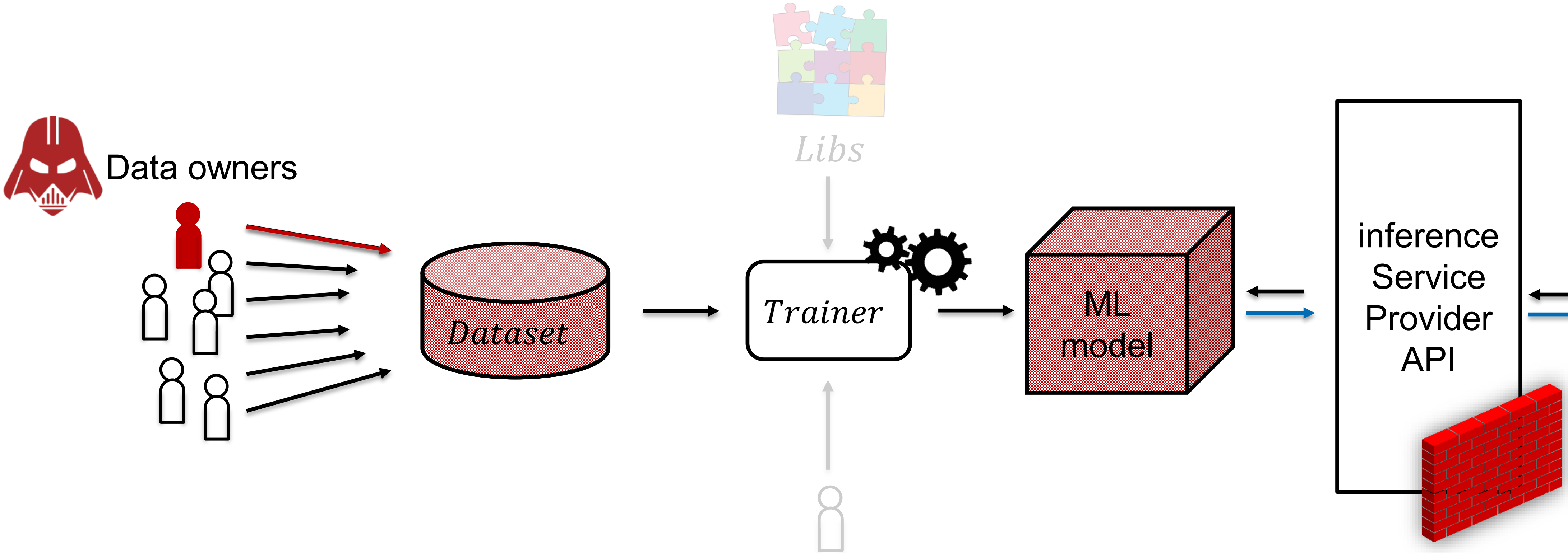


Profile users

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)

Malicious data owner – Model integrity

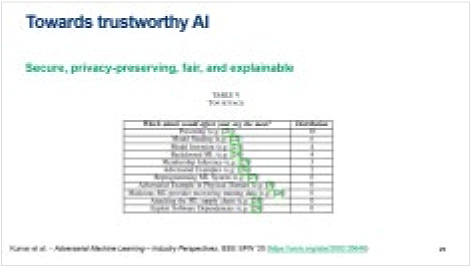


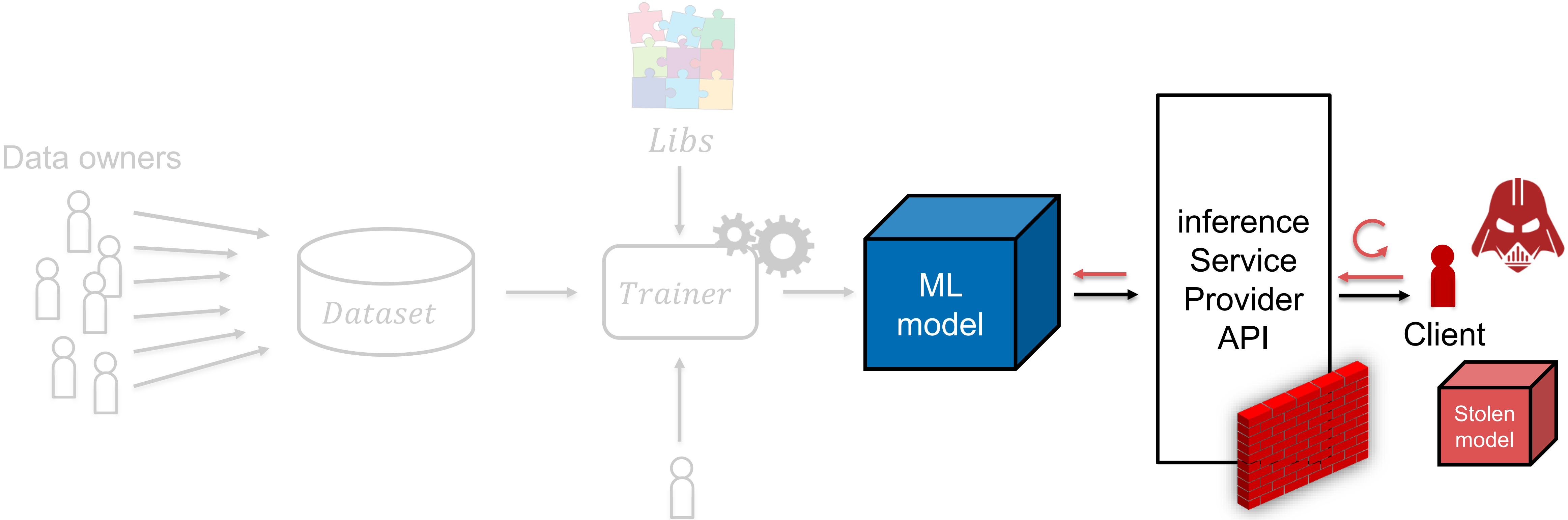


Influence ML model (model poisoning)

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

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 malicious adversarial behaviour the only concern?

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

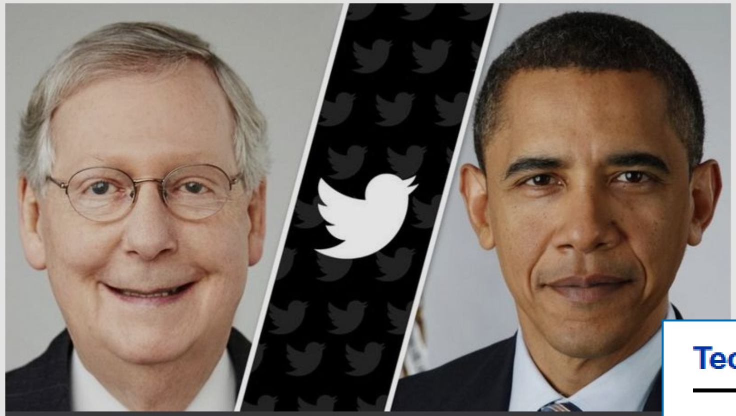
NEWS

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

Jordan Simonovski
@_jsimonovski

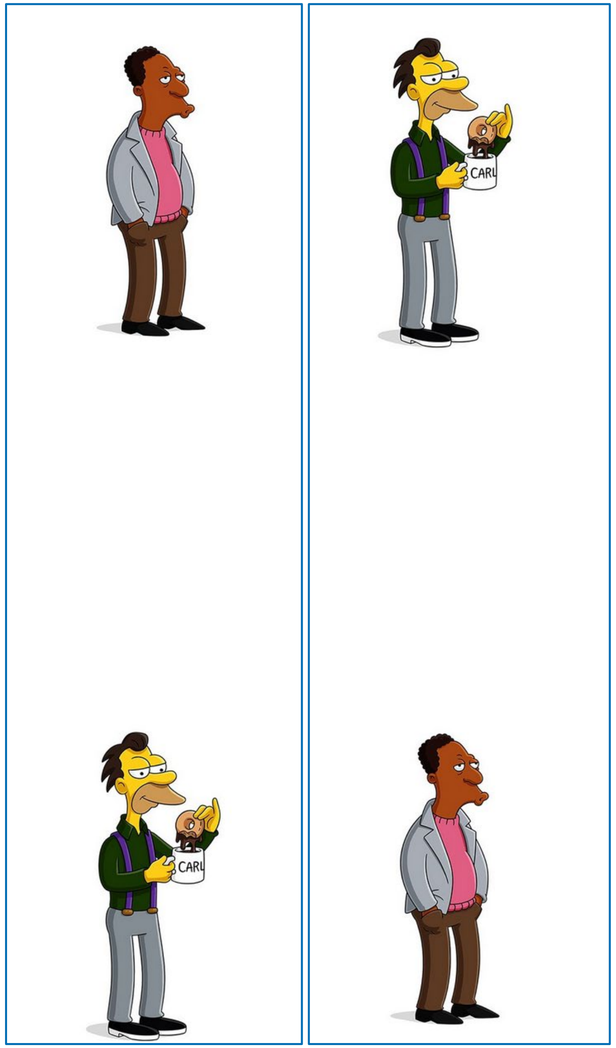
I wonder if Twitter does this to fictional characters too.

Lenny Carl

12:50 AM · Sep 20, 2020 · Twitter Web App

8K Retweets **1.2K** Quote Tweets **46.1K** Likes

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



Twitter Comms
@TwitterComms

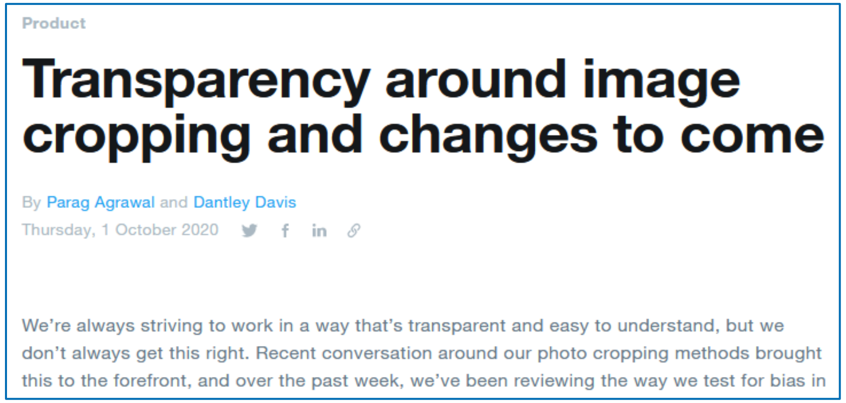
Replying to @bascule

We tested for bias before shipping the model & didn't find evidence of racial or gender bias in our testing. Bu it's clear that we've got more analysis to do. We'll continue to share what we learn, what actions we take, & will open source it so others can review and replicate

1:54 PM · Sep 20, 2020 · Twitter Web App

160 Retweets **92** Quote Tweets **1.4K** Likes

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



Product

Transparency around image cropping and changes to come

By Parag Agrawal and Dantley Davis
Thursday, 1 October 2020

We're always striving to work in a way that's transparent and easy to understand, but we don't always get this right. Recent conversation around our photo cropping methods brought this to the forefront, and over the past week, we've been reviewing the way we test for bias in

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

Towards trustworthy AI

Secure, privacy-preserving, fair, and explainable

TABLE V
TOP ATTACK

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

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

Outline

Is model stealing an important concern?

Can models be extracted via their inference APIs?

What can be done to counter model theft?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy protections against multiple concerns?

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

“Steal” = derive model from someone else’s model without their consent to do so

Type of model access: white box

White-box access: user

- has physical access to model
- knows its structure
- can observe execution (e.g., 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 hardware-based trusted execution environments**
- **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 (“inference API”):
  - directly (translation, image classification)
  - indirectly (recommender systems)

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

Is that enough to prevent model theft?

Outline

Is model stealing an important concern?

Can models be extracted via their inference APIs?

What can be done to counter model theft?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy protections against multiple concerns?

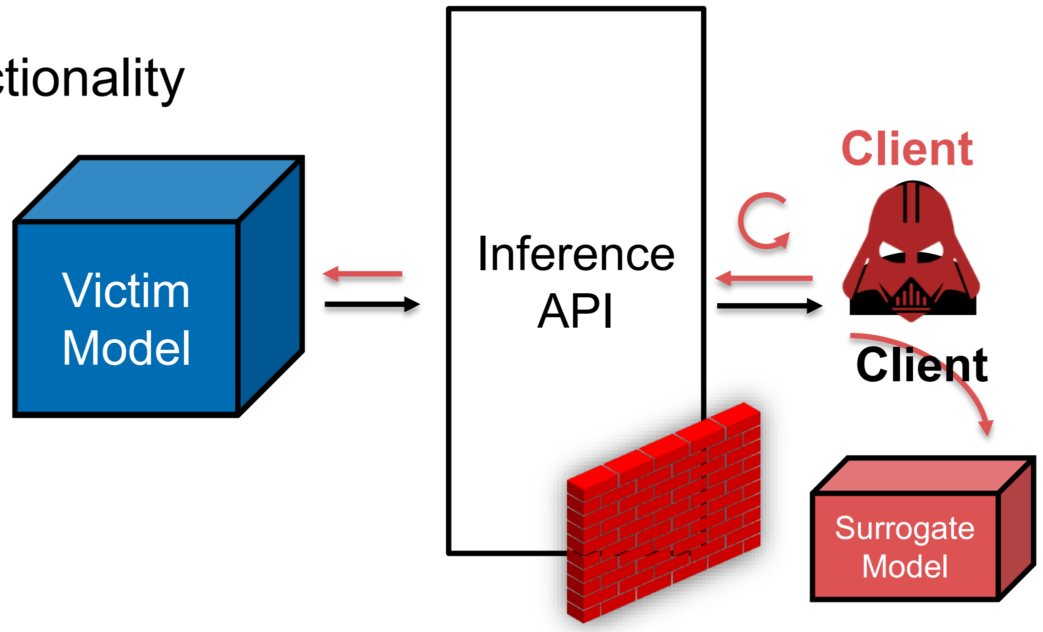
Extracting models via their inference APIs

Inference APIs are oracles that leak information

Adversary

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

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



Prior work on extracting

- Logistic regression, decision trees[1]
- Simple convolutional neural network 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 deep neural network models[1]

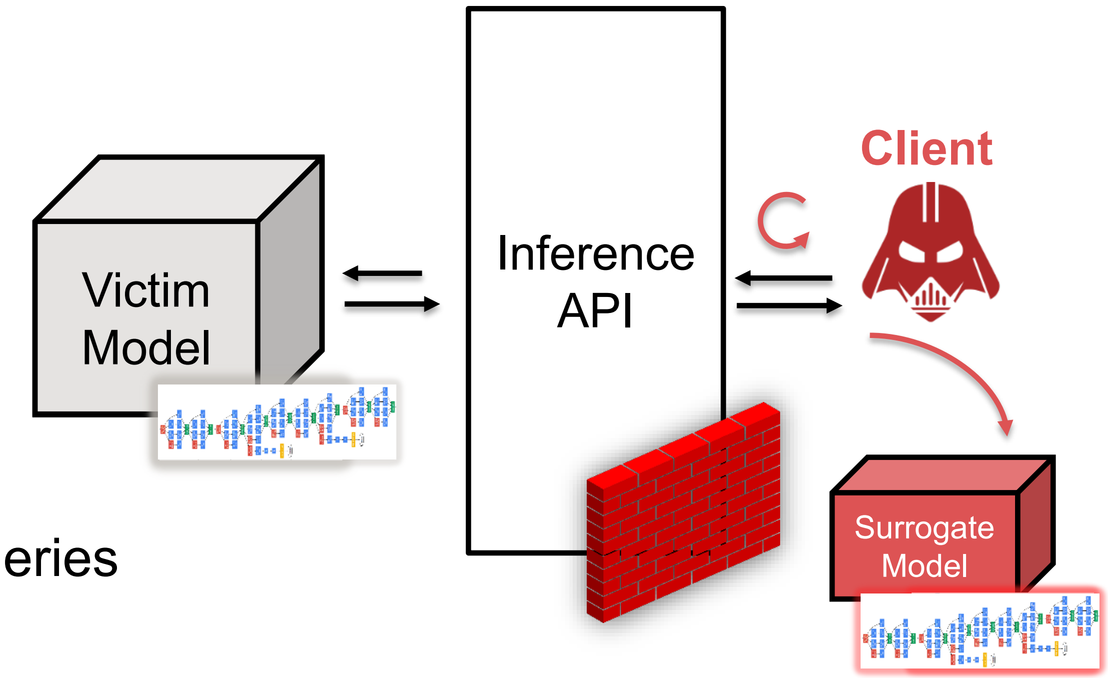
- E.g., MNIST, GTSRB

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 benign vs. adversarial queries



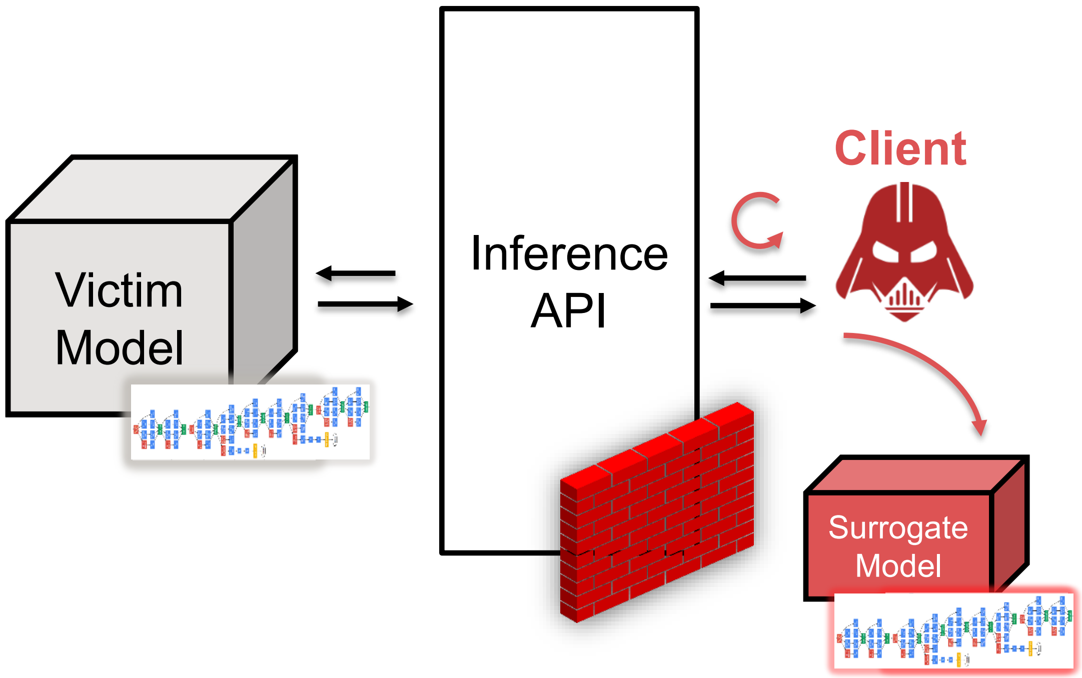
[1] Juuti et al. – *PRADA: Protecting against DNN Model Stealing Attacks*, Euro S&P '19 (https://arxiv.org/abs/1805.02628)

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 (image classifier) that
  - steals model functionality of victim model
  - performs similarly on the same task with high 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 to make more realistic assumptions about the adversary

Attack effectiveness decreases if

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

Simple defense: detector to identify out-of-distribution queries

Defense ineffective if 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)

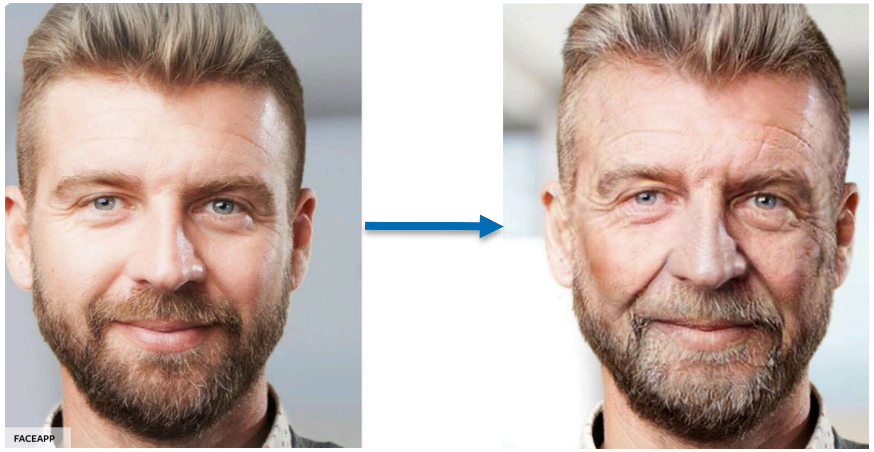
Extracting style-transfer models

GANS are effective for changing image style

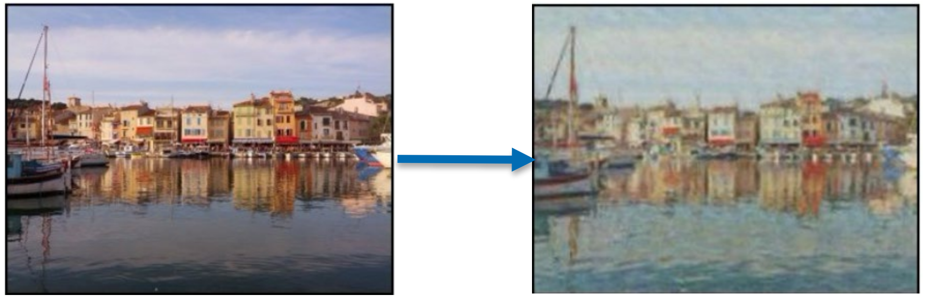
- coloring, face filters, style application

Core feature in generative art and in social media apps

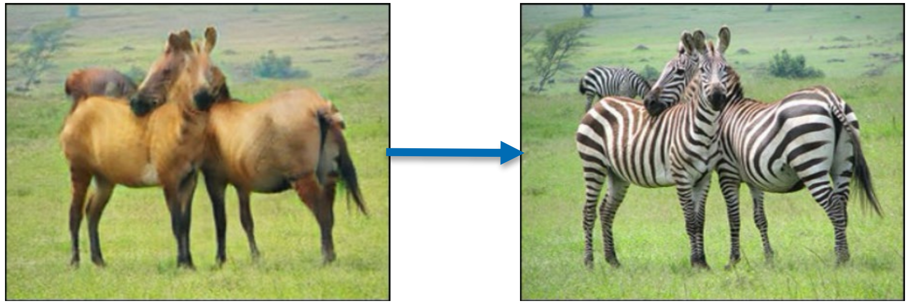
- **Selfie2Anime**, **FaceApp**



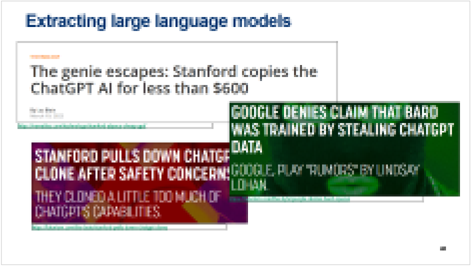
FaceApp



CycleGANs



CycleGANs

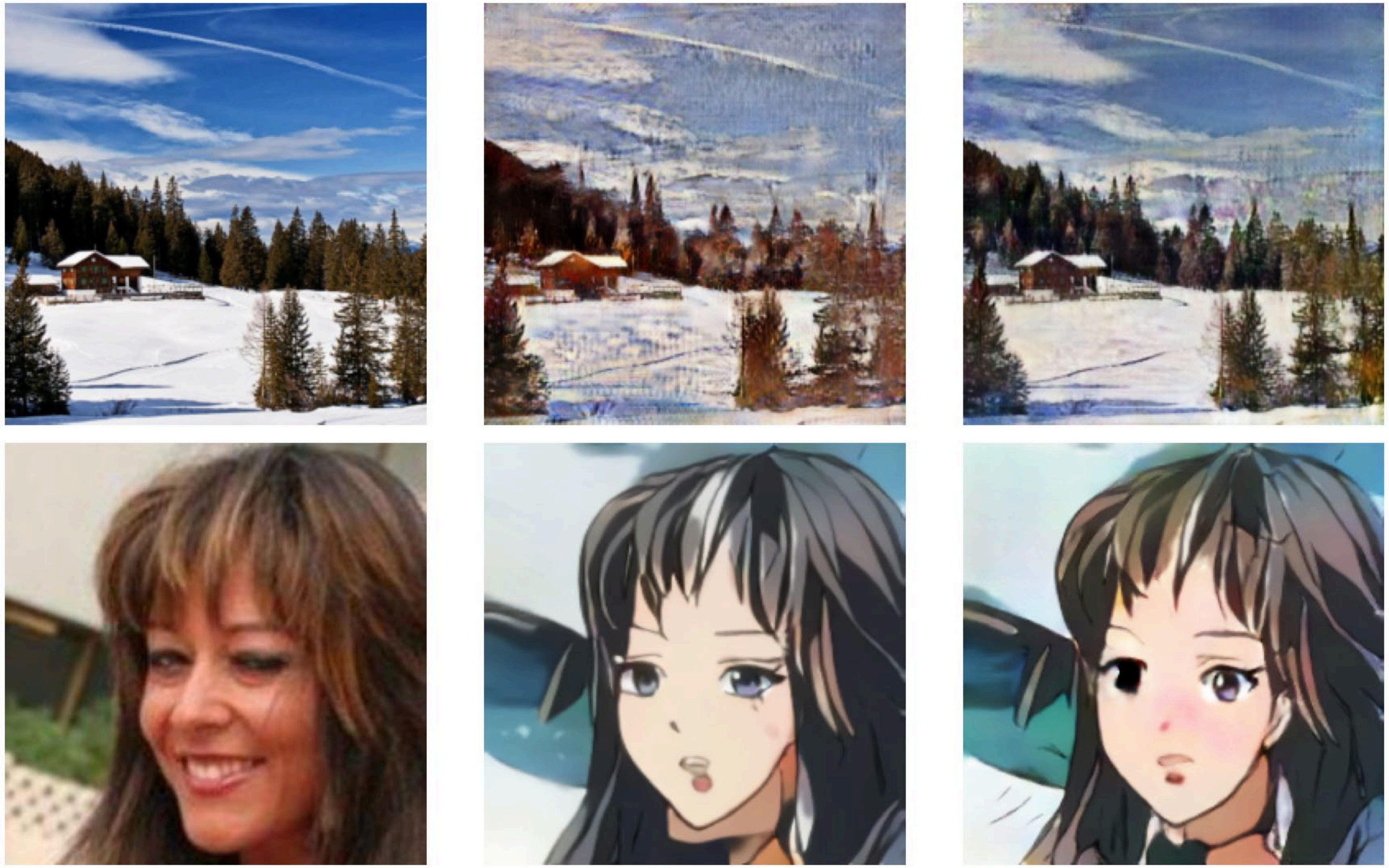


Extracting style-transfer models

Original
(unstyled)

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

Extracting natural language processing models

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

Transfer learning from pre-trained models is now very popular

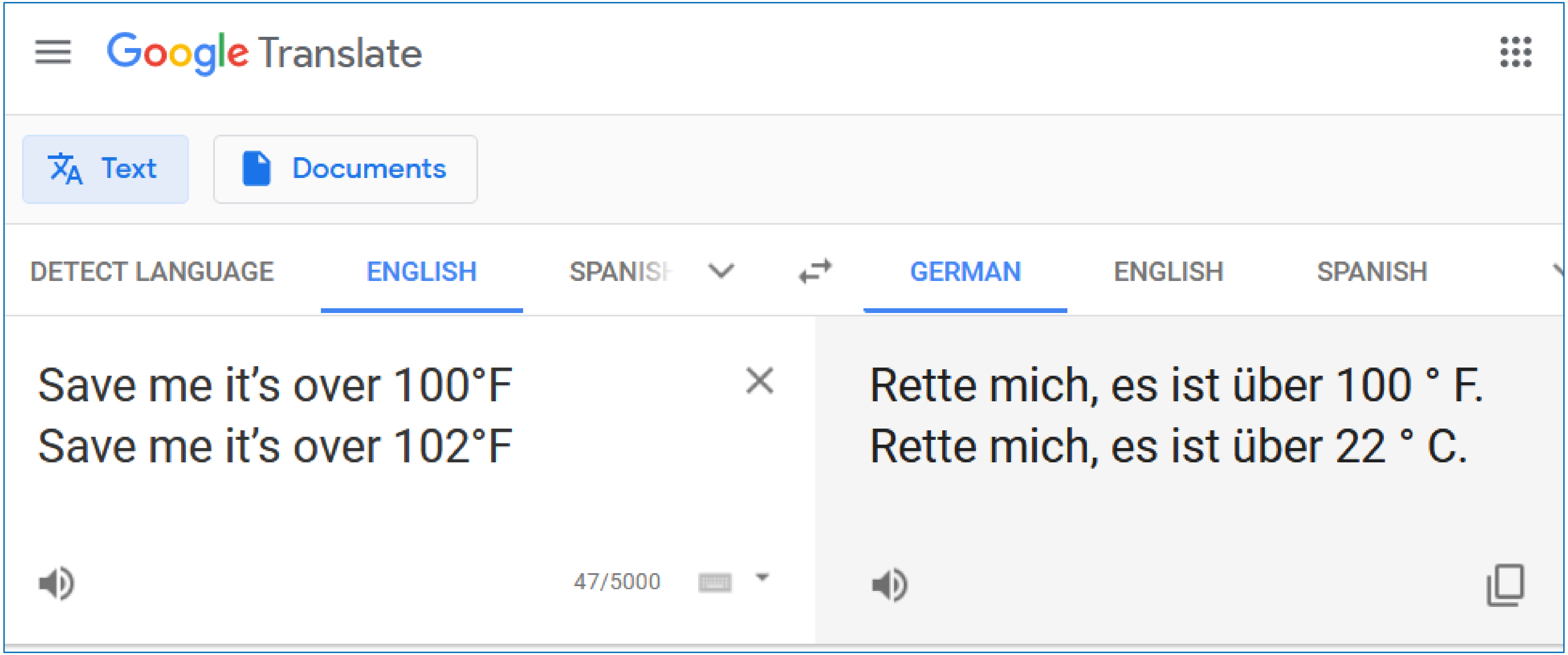
- But they make model extraction easier[1]

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

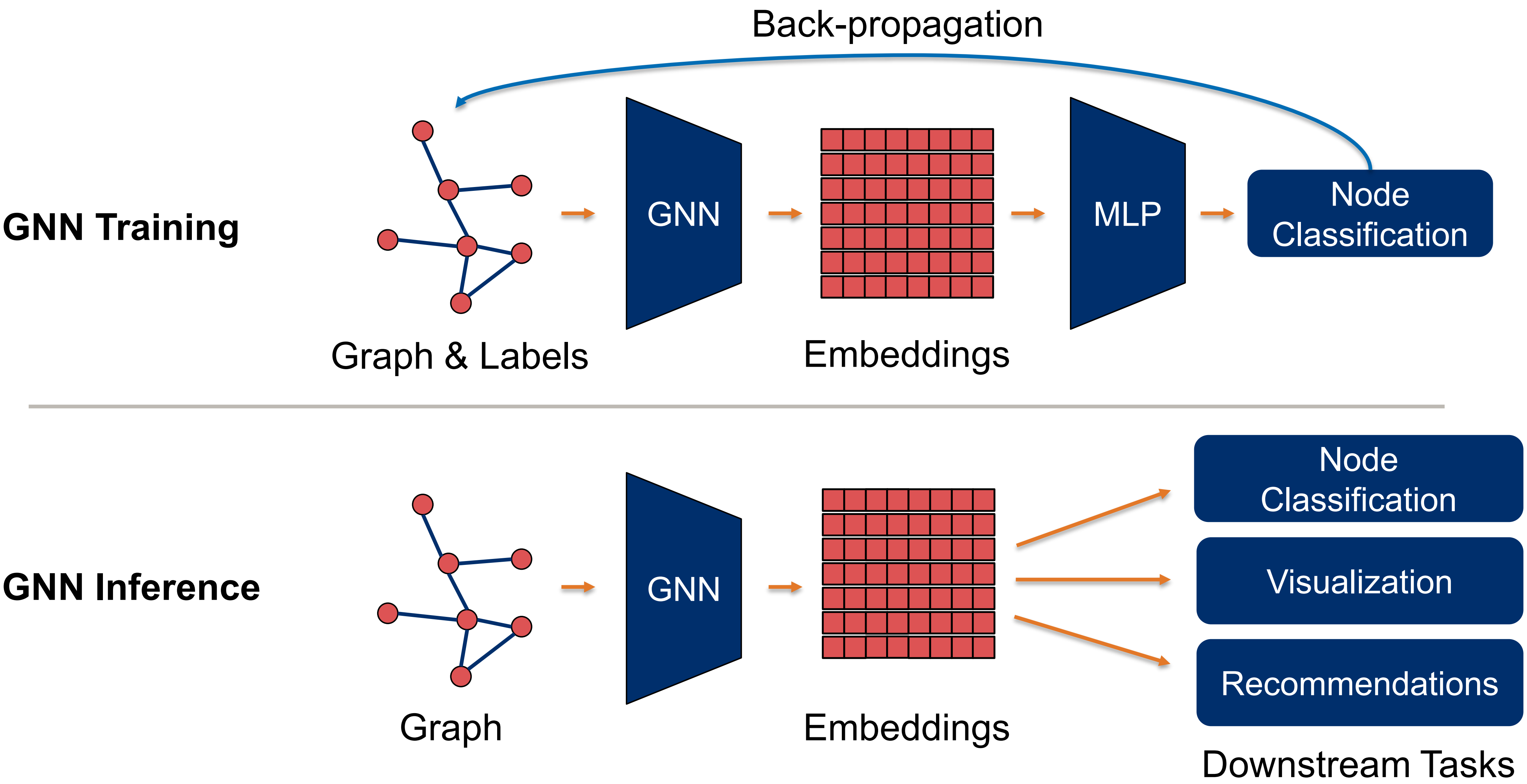
[1] Krishna et al. – *Thieves on Sesame Street! Model Extraction of BERT-based APIs,* ICLR '20 (https://iclr.cc/virtual_2020/poster_Byl5NREFDr.html)
[2] Wallace et al. – *Imitation Attacks and Defenses for Black-box Machine Translation Systems,* EMNLP '20 (https://arxiv.org/abs/2004.15015)



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 Graph Neural Networks



Shen et al. – *Model Stealing Attacks Against Inductive Graph Neural Networks,* IEEE S&P ‘22 (https://arxiv.org/abs/2112.08331)

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/

GOOGLE DENIES CLAIM THAT BARD WAS TRAINED BY STEALING CHATGPT DATA

GOOGLE, PLAY "RUMORS" BY LINDSAY LOHAN.

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

STANFORD PULLS DOWN CHATGP[T] CLONE AFTER SAFETY CONCERN[S]

THEY CLONED A LITTLE TOO MUCH OF CHATGPT'S CAPABILITIES.

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

Outline

Is model stealing an important concern? Yes

Can models be extracted via their inference APIs? Yes[1]

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

What can be done to counter model theft?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy protections against multiple concerns?

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

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):

- 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)

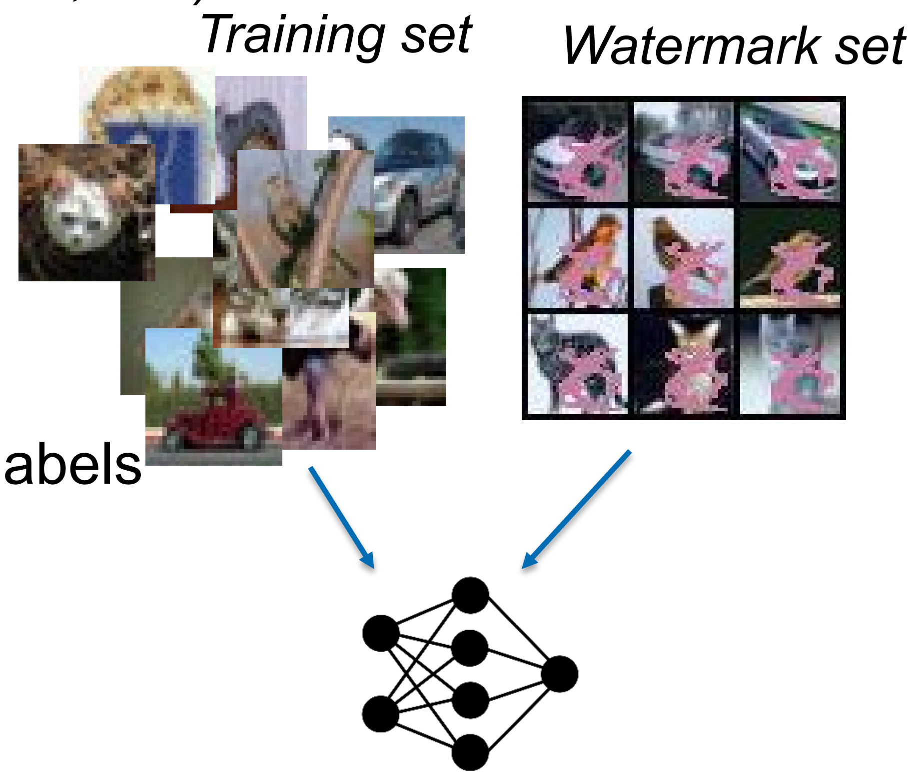
White-box watermarking

Watermark embedding:

- Embed the watermark in the model **during training**:
  - 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[1,2]

Not robust against

- novel watermark removal attacks[3]
- model extraction attacks that reduce effect of watermarks & modify decision surface

[1] Shafieinejad et al. – *On the Robustness of Backdoor-based Watermarking in Deep Neural Networks.* IH&MMSec ’21
[2] Szyller et. al. – *DAWN: Dynamic Adversarial Watermarking of Neural Networks.* ACM MM ‘21 (https://arxiv.org/abs/1906.00830)
[3] 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]

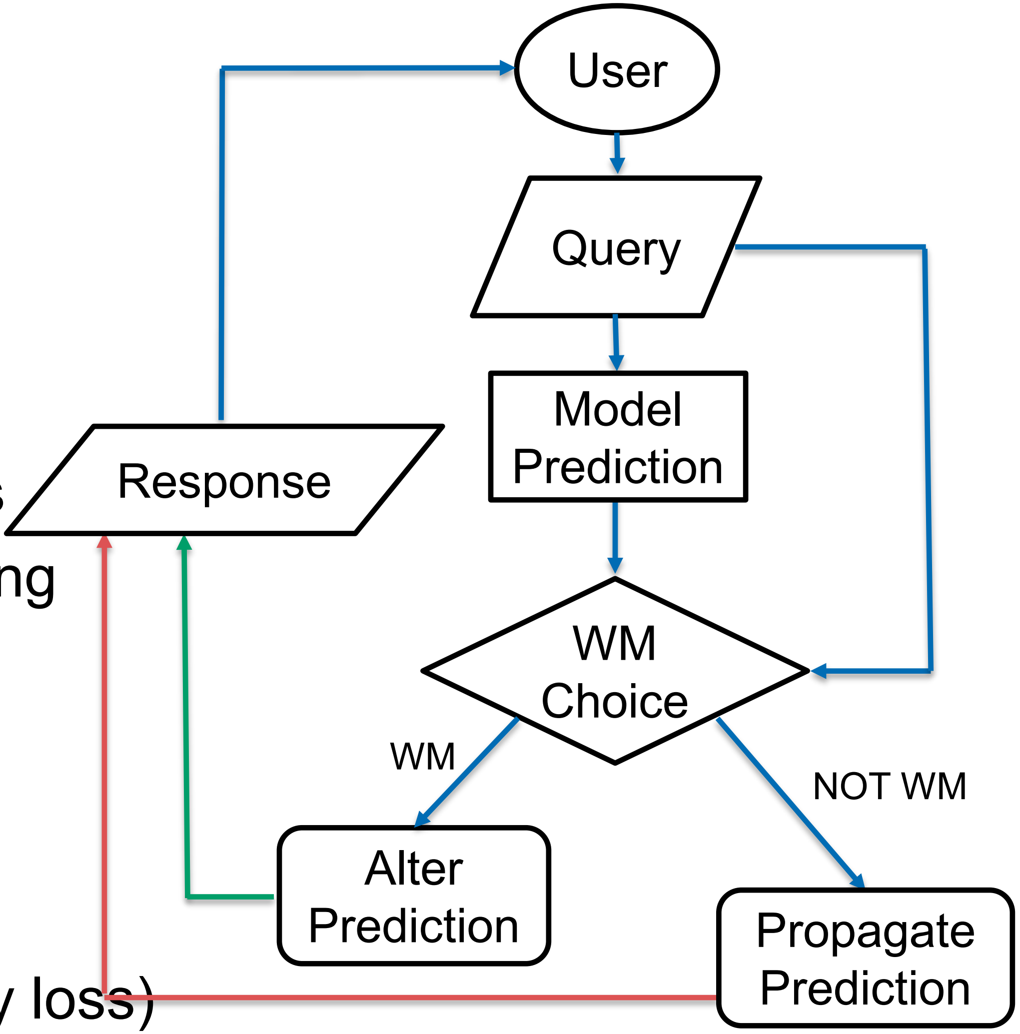
Goal: Deter model extraction via watermarking

Our approach:

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

Watermarking evaluation:

- **Robust** (if a suitable mapping function is found!)
- **Defends against** *PRADA*[2] and *KnockOff* [3]
- Preserves 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)

Fingerprinting

Conferrable adversarial examples[1]

- Distinguish between conferrable adversarial examples vs. other transferable ones
- Computationally expensive

Dataset inference[2]

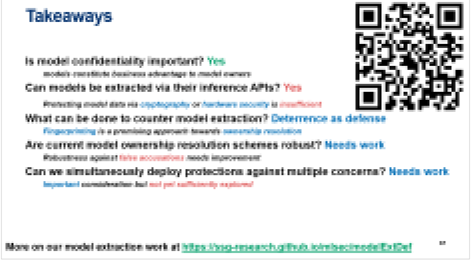
- Distinguish between models trained with different datasets
- Susceptible to false positives/negatives under certain conditions?[3]

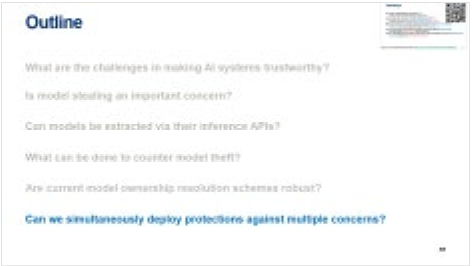
GrOVe[4]

- Use GNN embeddings as fingerprints
- Effective against high-fidelity extraction[5] but likely not against low-fidelity extraction

[1] Lukas et al. – *Deep Neural Network Fingerprinting by Conferrable Adversarial Examples,* ICLR ’21 (https://openreview.net/forum?id=VqzVhqxkjH1)
[2] Maini et al. – *Dataset Inference Ownership Resolution in Machine Learning*, ICLR ’21 (https://openreview.net/pdf?id=hvdKKV2yt7T)
[3] Szyller et al. – *On the Robustness of Dataset Inference*, TMLR ‘23 (https://arxiv.org/abs/2210.13631)
[4] Waheed et al. – *GrOVe: Ownership Verification of Graph Neural Networks using Embeddings,* IEEE S&P ‘24 (https://arxiv.org/abs/2304.08566)
[5] Shen et al. – *Model Stealing Attacks Against Inductive Graph Neural Networks,* IEEE S&P ‘22 (https://arxiv.org/abs/2112.08331)

Outline





Is model stealing an important concern?

Can models be extracted via their inference APIs?

What can be done to counter model theft?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy protections against multiple concerns?

Robustness of model ownership resolution schemes

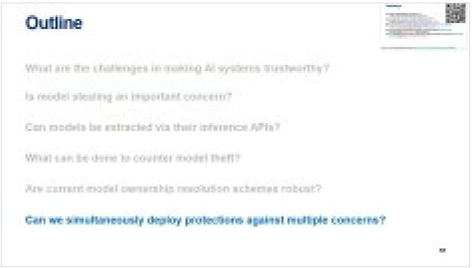
Model ownership resolution (MOR) must be robust against two types of attackers

Malicious suspect:
- tries to evade verification
- common approaches: pruning, fine-tuning, noising

Malicious accuser:
- tries to frame an independent model owner
- timestamping (watermark/fingerprint and model) is the only defense in prior work

So far, research has focused on malicious responders

False claims against MORs



We show how malicious accusers can make false claims against independent models:

- adversary deviates from watermark/fingerprint generation procedure
  - E.g., via transferrable adversarial examples
- but still subject to specified verification procedure

Our contributions:

- formalize the notion of false claims against MORs
- provide a generalization of MORs
- demonstrate effective false claim attacks
- discuss potential countermeasures

Zhang et al. – *False Claims Against Model Ownership Resolution*, (https://arxiv.org/abs/2304.06607)

Watermarking by backdooring[1]

Watermark generation:

- choose some out-of-distribution samples as watermark
  - assigned with incorrect labels
- train using the watermark alongside your normal training data (or finetune)
  - model memorizes watermark
- obtain timestamp on commitment of model and watermark

Watermark verification:

- query suspect model using watermark
- compare predictions to the assigned (incorrect) labels:
  - many matching / high WM accuracy -> stolen
  - a few matching / low WM accuracy > not stolen
- check commitment and timestamp

Watermarking by backdooring[1]: false claim

Watermark generation:

- choose some out-of-distribution samples as watermark
  - assigned with incorrect labels
- train using the watermark alongside your normal training data (or finetune)
  - model memorizes watermark
- obtain timestamp on commitment of model and watermark

Watermark verification:

- query suspect model using watermark
- compare predictions to the assigned (incorrect) labels:
  - many matching / high WM accuracy -> stolen
  - a few matching / low WM accuracy > not stolen
- check commitment and timestamp

[1] Adi et al. – *Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring*, Usenix SEC 2018 (https://arxiv.org/abs/1802.04633)
[2] Zhang et al. – *False Claims Against Model Ownership Resolution*, (https://arxiv.org/abs/2304.06607)

Watermarking by backdooring[1]: false claim

False watermark generation:

- choose some out-of-distribution samples as false watermark
- perturb these samples to craft transferable adversarial examples
- obtain timestamp on commitment of model and false watermark

Watermark verification:

- query suspect model using watermark
- compare predictions to the assigned (incorrect) labels:
  - many matching / high WM accuracy -> stolen
  - a few matching / low WM accuracy > not stolen
- check commitment and timestamp

[1] Adi et al. – *Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring*, Usenix SEC 2018 (https://arxiv.org/abs/1802.04633)
[2] Zhang et al. – *False Claims Against Model Ownership Resolution*, (https://arxiv.org/abs/2304.06607)

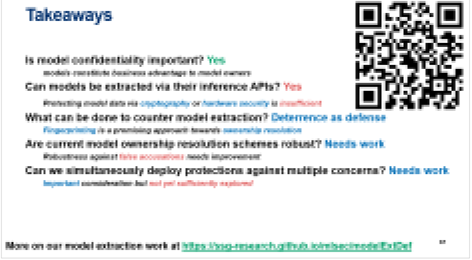
Mitigating false claims against MORs

Judge generates watermarks/fingerprints: bottleneck

Judge verifies watermarks/fingerprints were generated correctly: expensive

Train models with transferable adversarial examples: accuracy loss

Zhang et al. – *False Claims Against Model Ownership Resolution*, (https://arxiv.org/abs/2304.06607)



Outline

What are the challenges in making AI systems trustworthy?

Is model stealing an important concern?

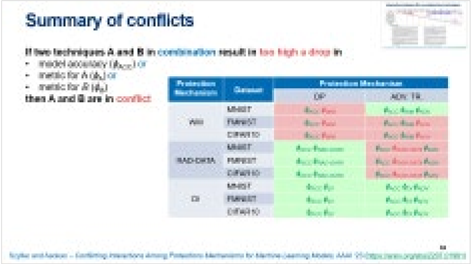
Can models be extracted via their inference APIs?

What can be done to counter model theft?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy protections against multiple concerns?

Other ML security & privacy concerns



There are considerations other than model ownership resolution:

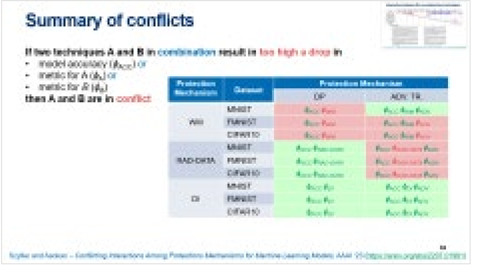
- model evasion (defense: adversarial training)
- training data reconstruction (defense: differential privacy)
- membership inference (defense: regularization, early stopping)
- model poisoning (defense: regularization, outlier/anomaly detection)
- …

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):

- WM: Out-of-distribution (OOD) backdoor watermarking (test and watermark accuracy)
- RAD-DATA: Radioactive data (test accuracy and loss difference)
- DI: Dataset Inference (verification confidence)
- DP: DP-SGD (model accuracy for the given epsilon)
- ADV-TR: 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.	
	ϕ_{ACC}	ϕ_{ACC}	ϕ_{WM}	ϕ_{ACC}	*Loss Diff.* $\phi_{RAD\text{-}DATA}$	*Confidence* ϕ_{DI}	ϕ_{ACC}	ϕ_{ACC}	ϕ_{ADV}
MNIST	*0.99±0.00*	*0.99±0.00*	*0.97±0.01*	*0.98±0.00*	*0.284±0.001*	*<e-30*	*0.98±0.00*	*0.99±0.00*	*0.95±0.00*
FMNIST	*0.91±0.00*	*0.87±0.02*	*0.99±0.02*	*0.88±0.01*	*0.19+0.002*	*<e-30*	*0.86±0.01*	*0.87±0.00*	*0.69±0.00*
CIFAR10	*0.92±0.00*	*0.82±0.00*	*0.97±0.02*	*0.85+0.00*	*0.20±0.001*	*<e-30*	*0.38±0.00*	*0.82±0.00*	*0.82±0.00*

Interaction with differential privacy

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	DP-SGD (eps=3)
	ϕ_{ACC}
MNIST	0.98±0.00
FMNIST	0.86±0.01
CIFAR10	0.38±0.00

Dataset	No defense	Watermarking				Radioactive Data				Dataset Inference	
		Baseline		with DP		Baseline		with DP		Baseline	with DP
	ϕ_{ACC}	ϕ_{ACC}	ϕ_{WM}	ϕ_{ACC}	ϕ_{WM}	ϕ_{ACC}	$\phi_{RAD\text{-}DATA}$	ϕ_{ACC}	$\phi_{RAD\text{-}DATA}$	ϕ_{DI}	ϕ_{DI}
MNIST	0.99±0.00	0.99±0.00	0.97±0.01	0.97±0.00	0.36±0.06	0.98±0.00	0.284±0.001	0.97±0.00	0.091±0.01	<e-30	<e-30
FMNIST	0.91±0.00	0.87±0.02	0.99±0.02	0.86±0.00	0.30±0.05	0.88±0.01	0.19±0.002	0.84±0.01	0.11±0.01	<e-30	<e-30
CIFAR10	0.92±0.00	0.82±0.00	0.97±0.02	0.38±0.01	0.12±0.01	0.85±0.00	0.2±0.001	0.35±0.01	0.19±0.01	<e-30	<e-30

Interaction with adversarial training

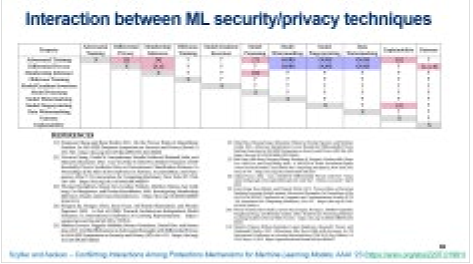
Adversarial training creates a robust L_p bubble:

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

Dataset	ADV. TR.	
	ϕ_{ACC}	ϕ_{ADV}
MNIST	*0.99±0.00*	*0.95±0.00*
FMNIST	*0.87±0.00*	*0.69±0.00*
CIFAR10	*0.82±0.00*	*0.82±0.00*

Dataset	No Def.	Watermarking					Radioactive Data					DI	
		Baseline		with ADV. TR.			Baseline		with ADV. TR.			Base.	with ADV. TR.
	ϕ_{ACC}	ϕ_{ACC}	ϕ_{WM}	ϕ_{ACC}	ϕ_{WM}	ϕ_{ADV}	ϕ_{ACC}	$\phi_{RAD\text{-}DATA}$	ϕ_{ACC}	$\phi_{RAD\text{-}DATA}$	ϕ_{ADV}	ϕ_{DI}	ϕ_{DI}
MNIST	*0.99±0.00*	*0.99±0.00*	*0.97±0.01*	*0.97±0.02*	*0.99±0.01*	*0.88±0.09*	*0.98±0.00*	*0.284±0.01*	*0.97±0.00*	*0.001±0.001*	*0.95±0.01*	*<e-30*	*<e-30*
FMNIST	*0.91±0.00*	*0.87±0.02*	*0.99±0.02*	*0.80±0.06*	*0.99±0.00*	*0.51±0.11*	*0.88±0.00*	*0.19±0.002*	*0.84±0.00*	*0.000±0.001*	*0.69±0.02*	*<e-30*	*<e-30*
CIFAR10	*0.92±0.00*	*0.82±0.00*	*0.97±0.02*	*0.78±0.00*	*0.97±0.01*	*0.65±0.01*	*0.85±0.00*	*0.2±0.001*	*0.81±0.00*	*0.003±0.002*	*0.81±0.01*	*<e-30*	*<e-30*

Summary of conflicts



If two techniques A and B in combination result in too high a drop in

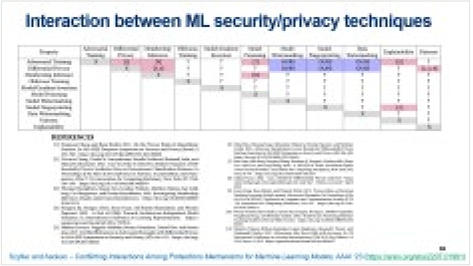
- model accuracy (ϕ_{ACC}) or
- metric for A (ϕ_A) or
- metric for B (ϕ_B)

then A and B are in conflict

Protection Mechanism	Dataset	Protection Mechanism	
		DP	ADV. TR.
WM	MNIST	ϕ_{ACC} ϕ_{WM}	ϕ_{ACC} ϕ_{WM} ϕ_{ADV}
	FMNIST	ϕ_{ACC} ϕ_{WM}	ϕ_{ACC} ϕ_{WM} ϕ_{ADV}
	CIFAR10	ϕ_{ACC} ϕ_{WM}	ϕ_{ACC} ϕ_{WM} ϕ_{ADV}
RAD-DATA	MNIST	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$ ϕ_{ADV}
	FMNIST	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$ ϕ_{ADV}
	CIFAR10	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$	ϕ_{ACC} $\phi_{RAD\text{-}DATA}$ ϕ_{ADV}
DI	MNIST	ϕ_{ACC} ϕ_{DI}	ϕ_{ACC} ϕ_{DI} ϕ_{ADV}
	FMNIST	ϕ_{ACC} ϕ_{DI}	ϕ_{ACC} ϕ_{DI} ϕ_{ADV}
	CIFAR10	ϕ_{ACC} ϕ_{DI}	ϕ_{ACC} ϕ_{DI} ϕ_{ADV}

Szyller and Asokan – *Conflicting Interactions Among Protections Mechanisms for Machine Learning Models*, AAAI '23 (https://arxiv.org/abs/2207.01991)

Combinatorial Explosion



The complexity of the analysis explodes quickly:

- we investigate 6 pair-wise interactions
- what about triples, quadruples…?
- DP, ADVTR, WM/fingerprinting with fairness constraints is a reasonable example

Thorough analysis with more schemes adds more complexity:

- we looked at one popular scheme in each category
- e.g., within DP one could study: DP-SGD, PATE, tempered sigmoids, SCATTER-DP

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

Szyller and Asokan – *Conflicting Interactions Among Protections Mechanisms for Machine Learning Models*, AAAI '23 (https://arxiv.org/abs/2207.01991)

Takeaways

Is model confidentiality important? Yes

models constitute business advantage to model owners

Can models be extracted via their inference APIs? Yes

Protecting model data via cryptography or hardware security is insufficient

What can be done to counter model extraction? Deterrence as defense

Fingerprinting is a promising approach towards ownership resolution

Are current model ownership resolution schemes robust? Needs work

Robustness against false accusations needs improvement

Can we simultaneously deploy protections against multiple concerns? Needs work

Important consideration but not yet sufficiently explored

More on our model extraction work at https://ssg-research.github.io/mlsec/modelExtDef

Takeaways

Is model confidentiality important? Yes

models constitute business advantage to model owners

Can models be extracted via their inference APIs? Yes

Protecting model data via cryptography or hardware security is insufficient

What can be done to counter model extraction? Deterrence as defense

Fingerprinting is a promising approach towards ownership resolution

Are current model ownership resolution schemes robust? Needs work

Robustness against false accusations needs improvement

Can we simultaneously deploy protections against multiple concerns? Needs work

Important consideration but not yet sufficiently explored

Open (postdoc) positions to help lead our work: ML security/privacy, platform security
https://asokan.org/asokan/research/SecureSystems-open-positions-Jan2024.php