Aalto University

UNIVERSITY OF WATERLOO

Model Stealing Attacks and Defenses

Where are we now?

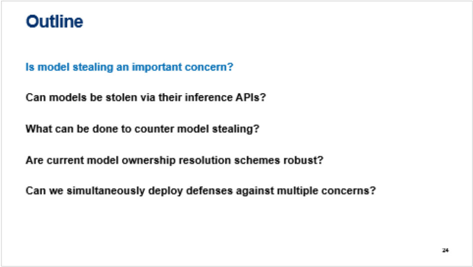
N. Asokan

https://asokan.org/asokan/

@nasokan

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

Outline



The big picture

Is model stealing an important concern?

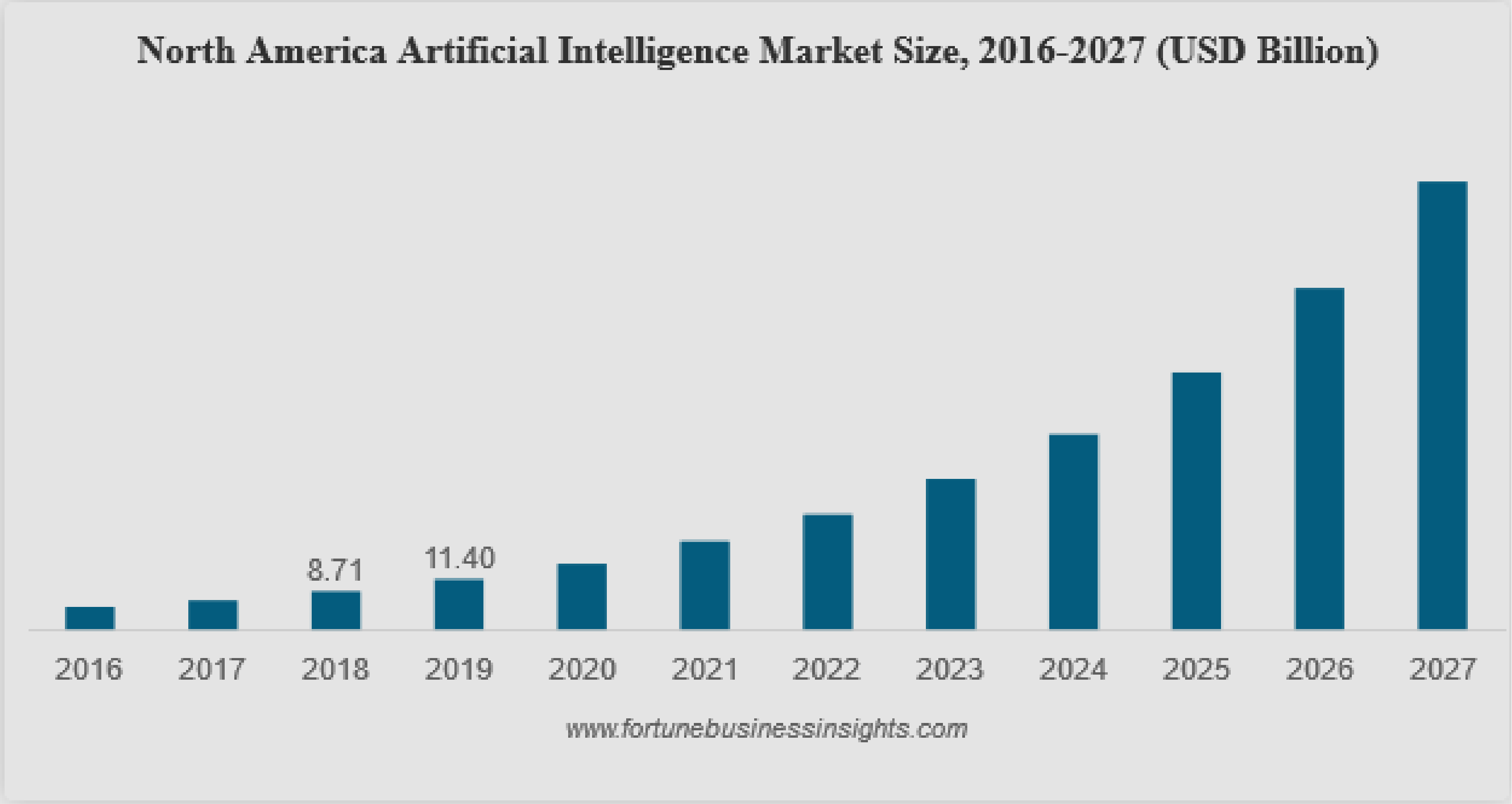
Can models be stolen via their inference APIs?

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses 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

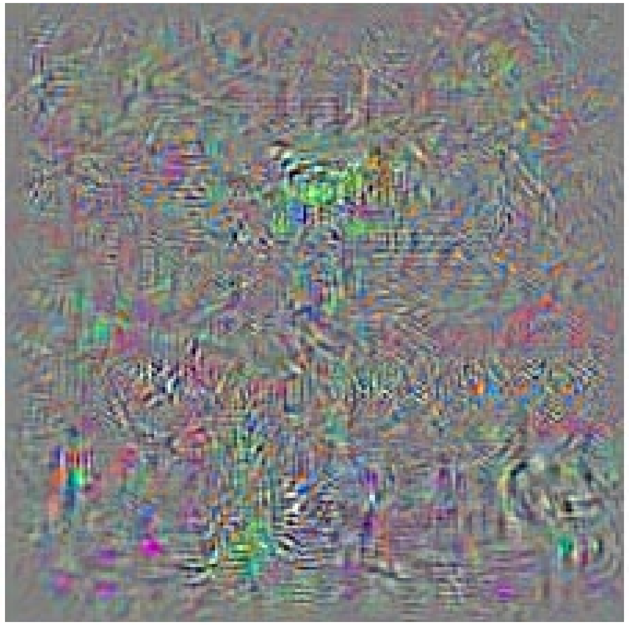
Privacy concerns

Fairness, explainability, and other concerns

Evading machine learning models

Which class is this?
School bus

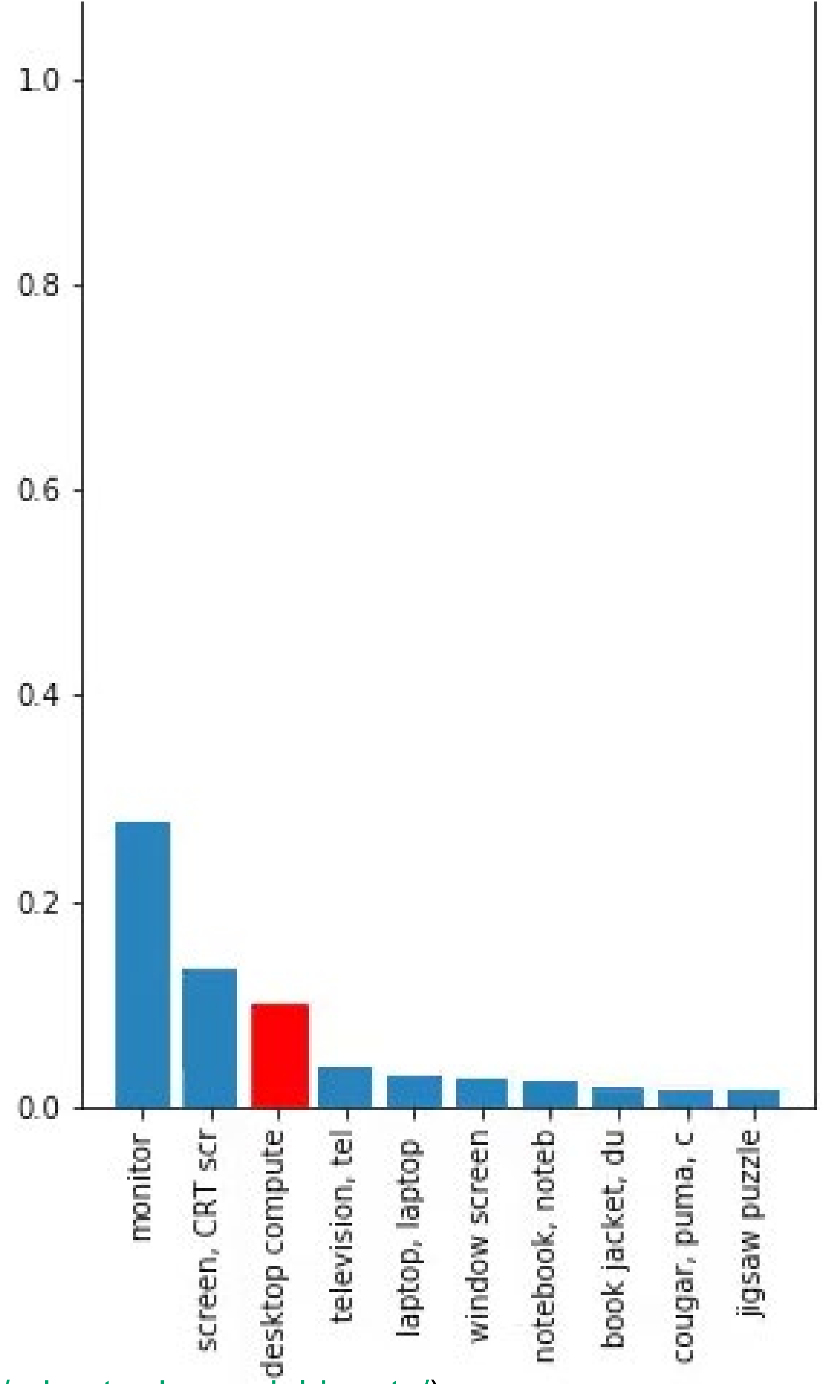
+ 0.1·



=

Which class is this?
Ostrich

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

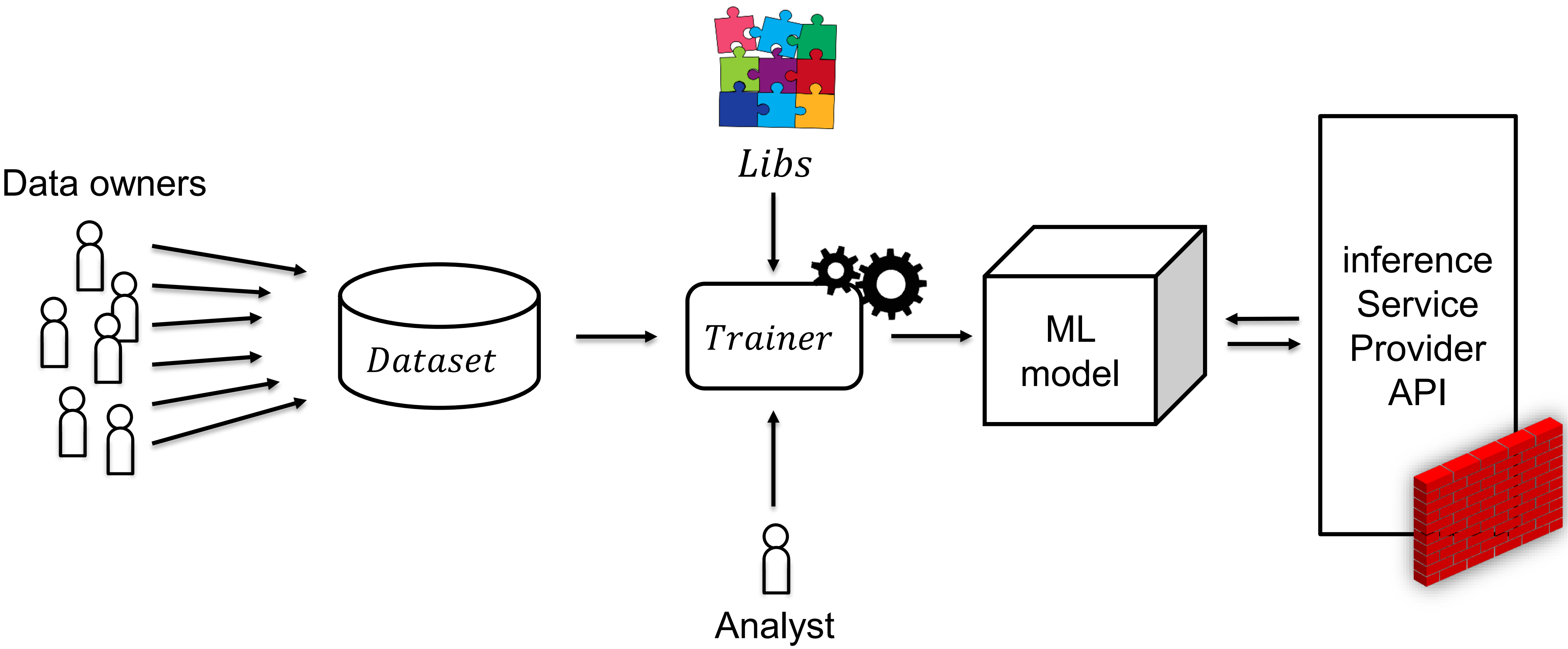


Which class is this?
Cat

Which class is this?
Desktop computer

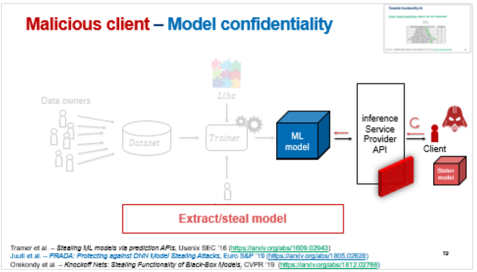
Athalye et al. – *Synthesizing Robust Adversarial Examples*, ICML ‘2019 (https://blog.openai.com/robust-adversarial-inputs/)

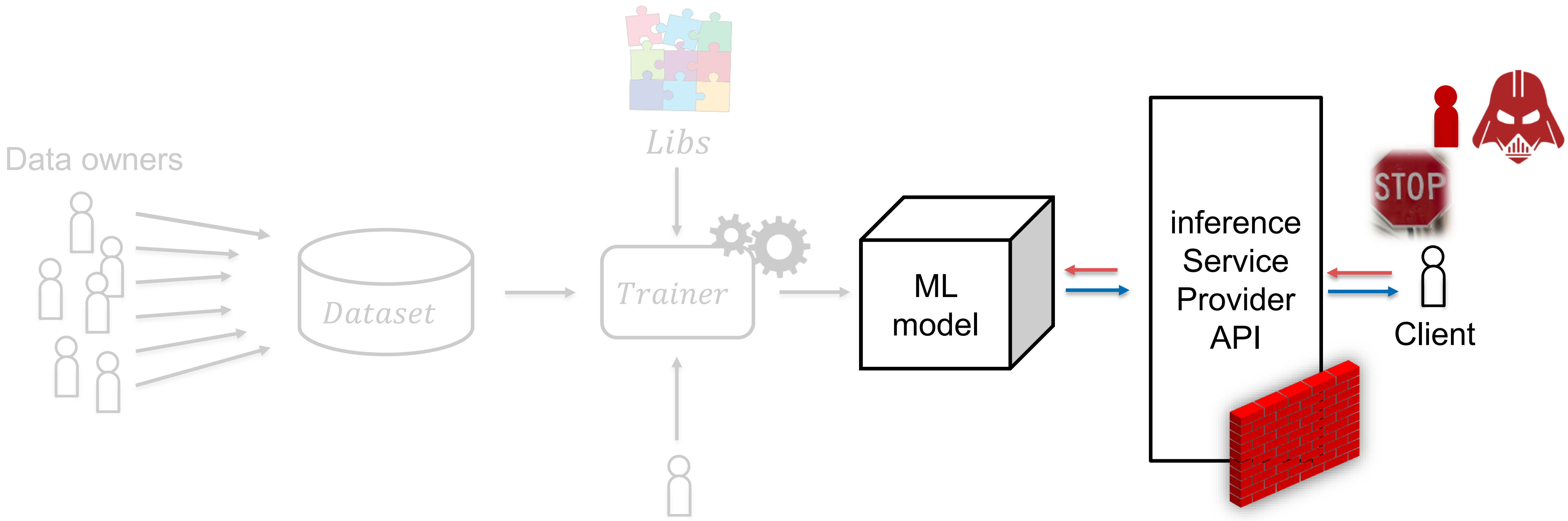
Machine Learning pipeline



Where is the adversary? What is its target?

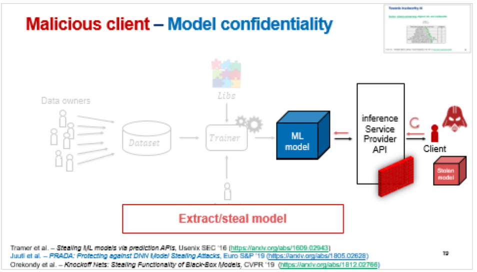
Compromised input – Model integrity



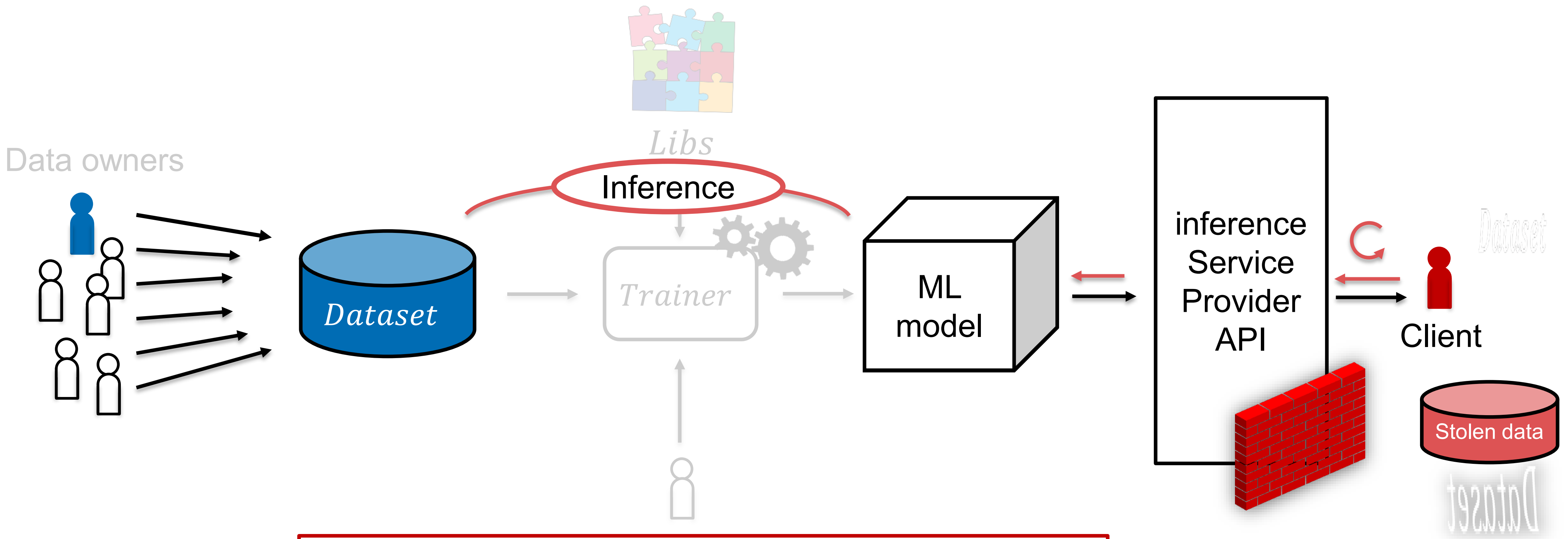


Evade model

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)



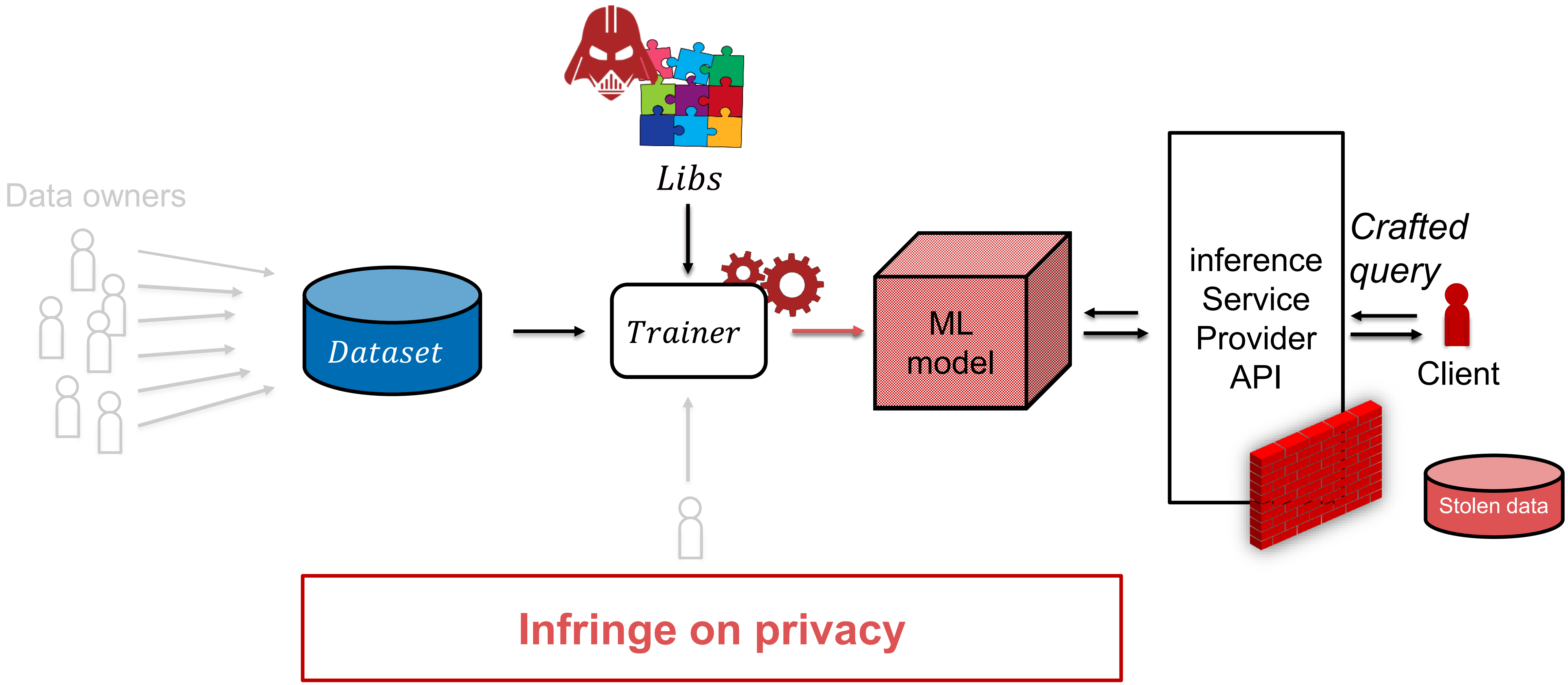
Malicious client – Training data privacy



Invert model, infer membership

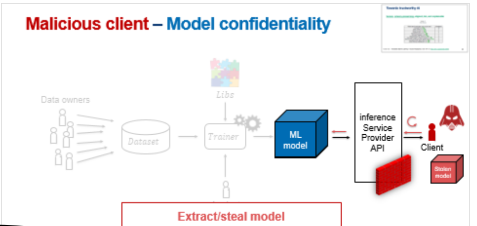
Shokri et al. – *Membership Inference Attacks Against Machine Learning Models,* IEEE S&P '16 (https://arxiv.org/pdf/1610.05820.pdf)
Fredrikson et al. – *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, ACM CCS '15 (https://doi.org/10.1145/2810103.2813677)

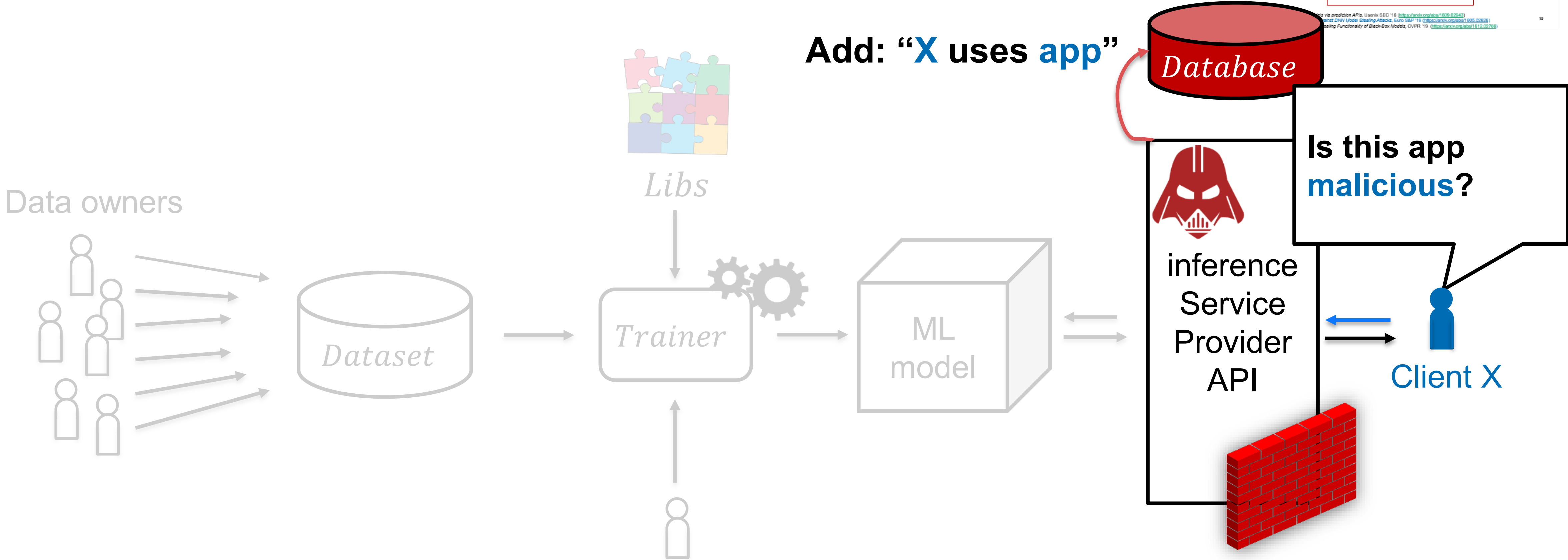
Compromised toolchain – Training data privacy



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

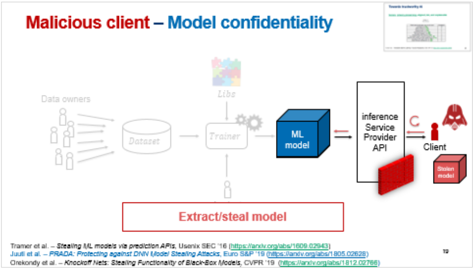


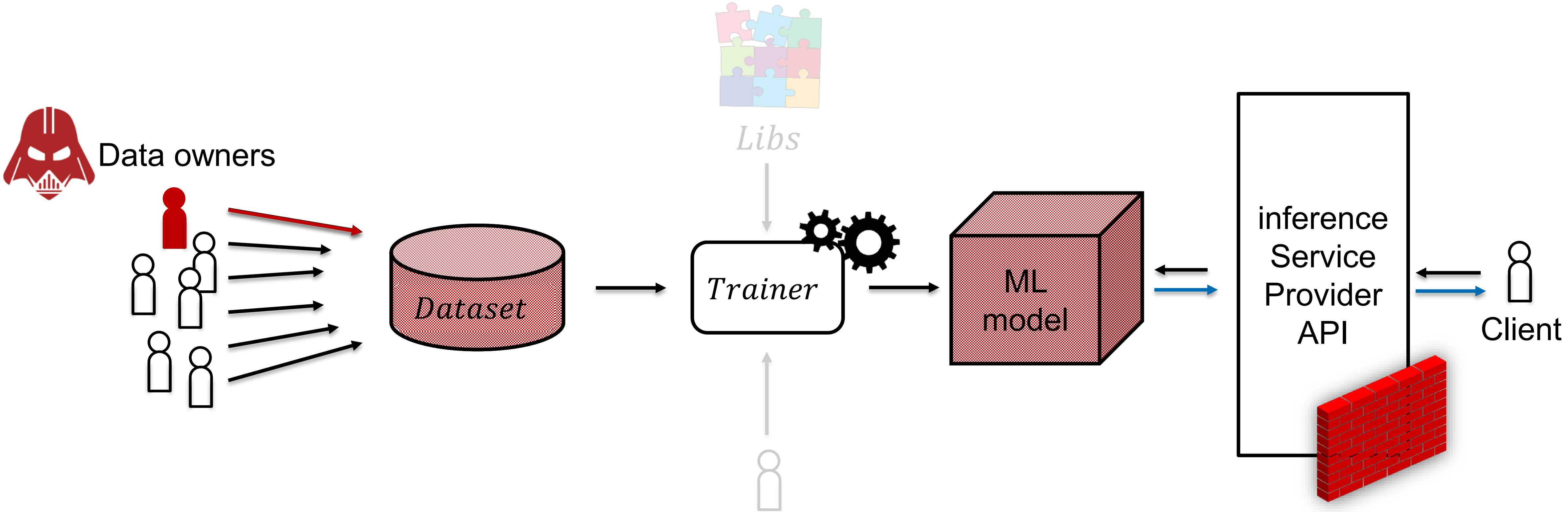


Profile users

Malmi and Weber – *You are what apps you use Demographic prediction based on user's apps*, ICWSM '16 (https://arxiv.org/abs/1603.00059)
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)
Liu et al. – *Oblivious Neural Network Predictions via MiniONN Transformations*, ACM CCS '17 (https://ssg.aalto.fi/research/projects/mlsec/ppml/)

Malicious data owner – Model integrity

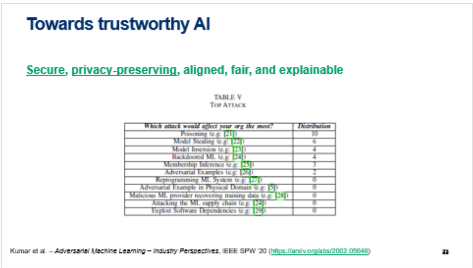




Influence ML model (model poisoning)

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

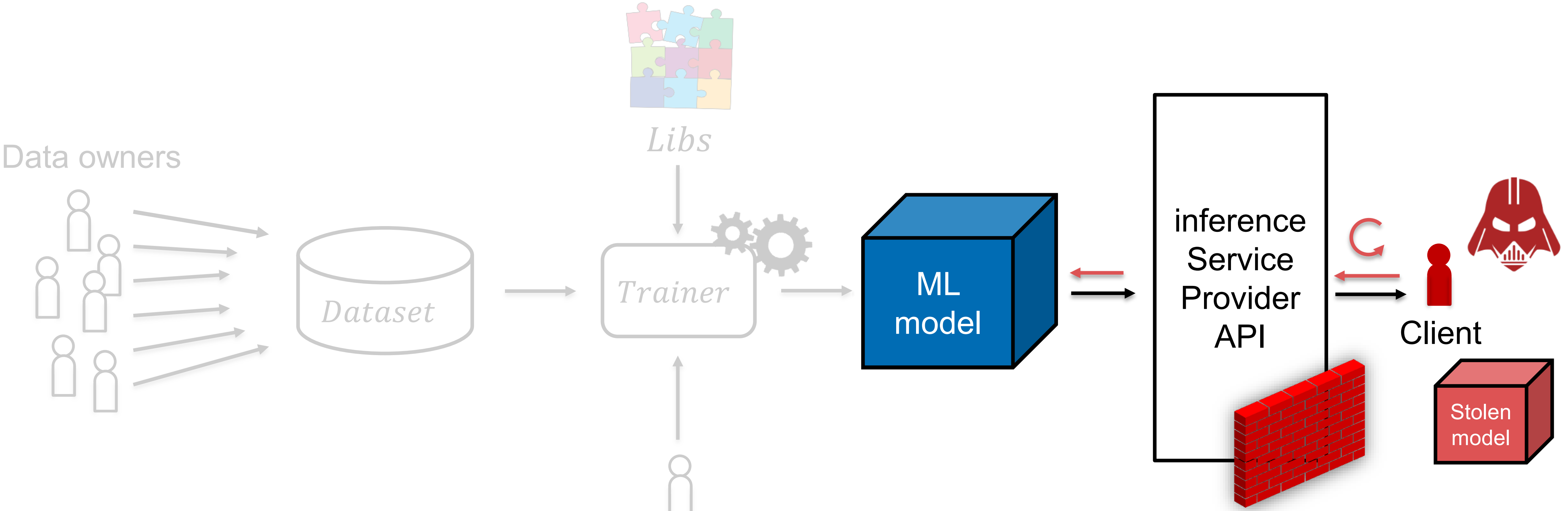

Towards trustworthy AI

Secure, privacy-preserving, aligned, 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)



Extract/steal model

Tramer et al. – *Stealing ML models via prediction APIs*, Usenix SEC '16 (https://arxiv.org/abs/1609.02943)
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 malicious adversarial behaviour the only concern?

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

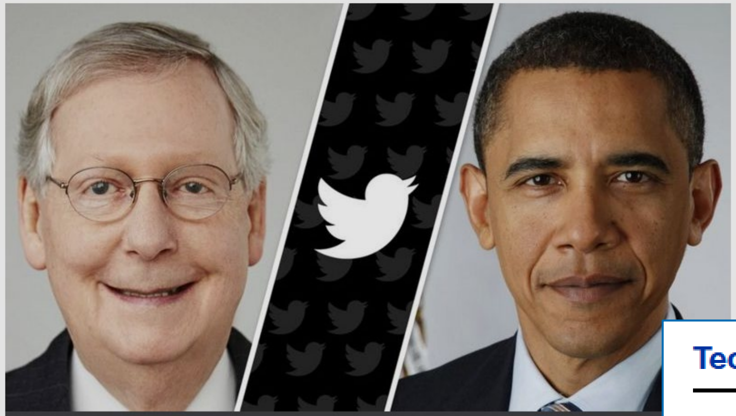
NEWS

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_HR6lluMKGRJbJdDrdpKdyAi5mhQSdzs0QLDso41T-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

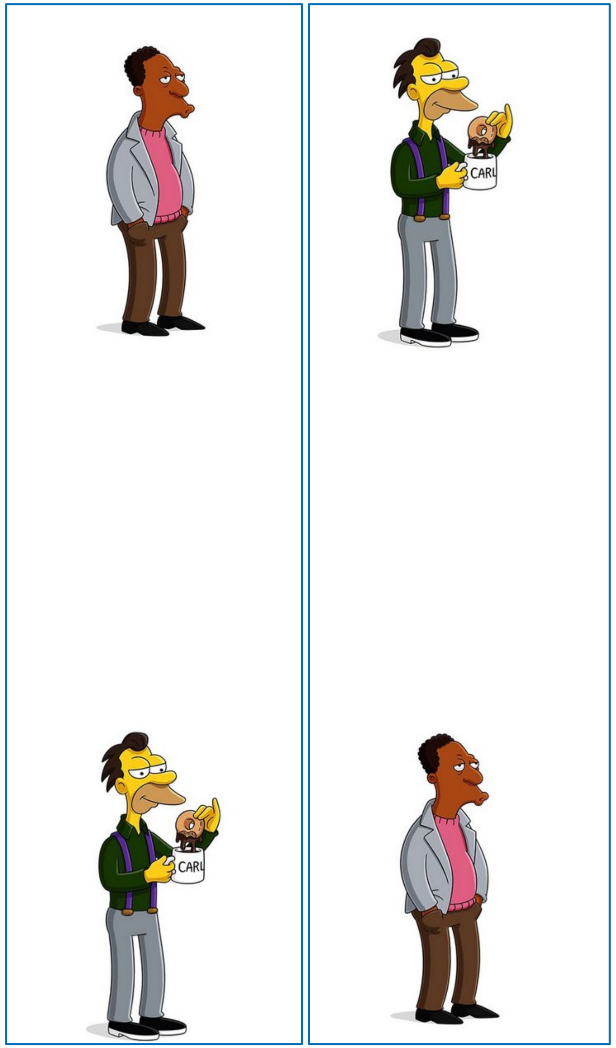
I wonder if Twitter does this to fictional characters too.

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

Other AI trustworthiness concerns

Unaligned AI

AI alignment

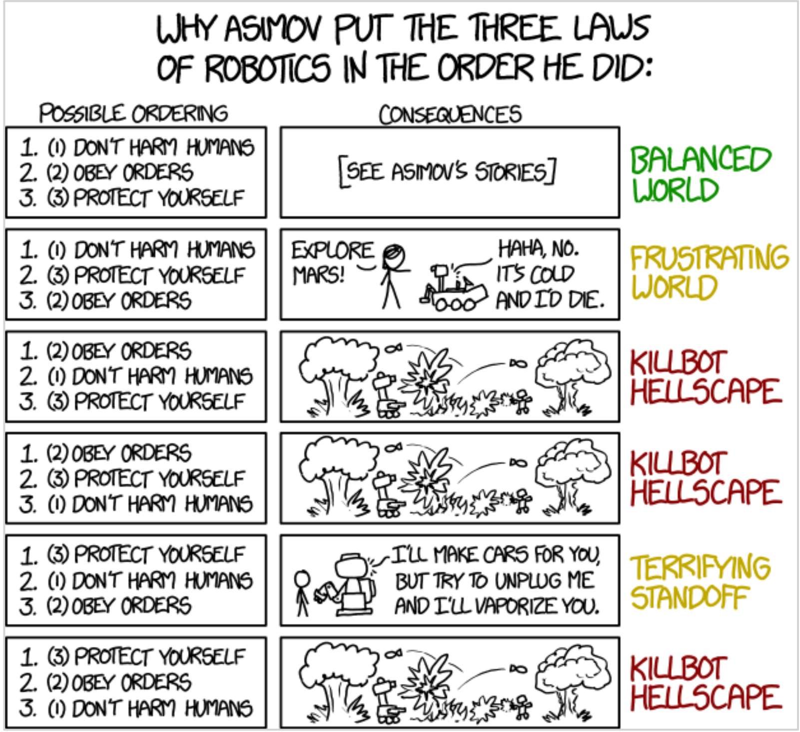
Article Talk

From Wikipedia, the free encyclopedia

In the field of artificial intelligence (AI), **AI alignment** research aims to steer AI systems toward a person's or group's intended goals, preferences, and ethical principles. An AI system is considered *aligned* if it advances its intended objectives. A *misaligned* AI system may pursue some objectives, but not the intended ones.[1]

It is often challenging for AI designers to align an AI system due to the difficulty of specifying the full range of desired and undesired behaviors. To aid them, they often use simpler *proxy goals*, such as gaining human approval. But that approach can create loopholes, overlook necessary constraints, or reward the AI system for merely *appearing* aligned.[1][2]

https://en.wikipedia.org/wiki/AI_alignment



https://xkcd.com/1613/

AI-enabled fraud

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM › PRESIDENTIAL ACTIONS

https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Towards trustworthy AI

Secure, privacy-preserving, aligned, 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 stolen via their inference APIs?

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses 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

How to prevent model stealing?

Outright (white-box) model stealing can be countered by

- Computation with encrypted models
- Protecting models using hardware-based trusted execution environments
- Hosting models behind a firewalled cloud service

Is that enough to prevent model stealing?

Outline

Is model stealing an important concern?

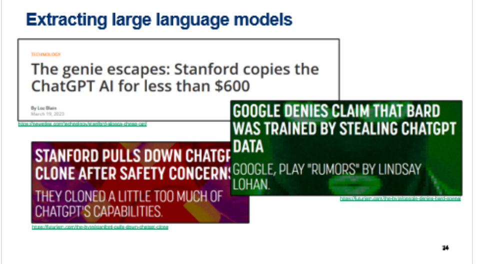
Can models be stolen via their inference APIs?

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses against multiple concerns?

Extracting models via their inference APIs

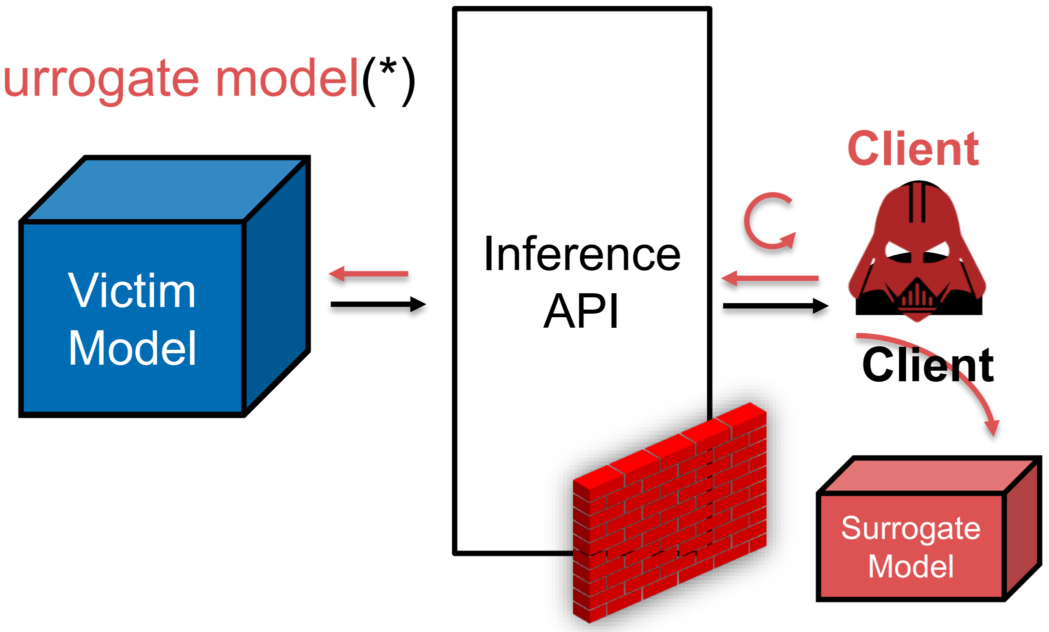


Inference APIs are oracles that leak information

Adversary

- Malicious client
- Goal: construct "comparable" [fidelity or functionality] surrogate model(*)
- Capability: access to inference API or model outputs

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



Early work on extracting

- Logistic regression, decision trees[1]
- Simple convolutional neural network models[2]
- Deep neural network models[3]

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

More effective extraction: Knockoff Nets

Knockoff nets[1]: adversary has

- no knowledge about model (task, architecture etc.), but gets full prediction vector
- natural data from the same domain but not (necessarily) from same distribution

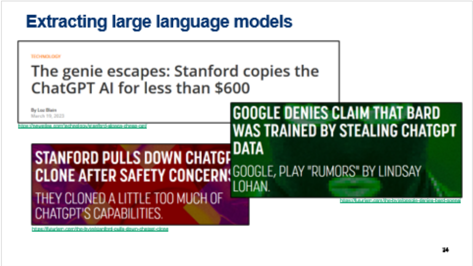
Attack effectiveness decreases[2] if

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

Simple defense[2] : 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)

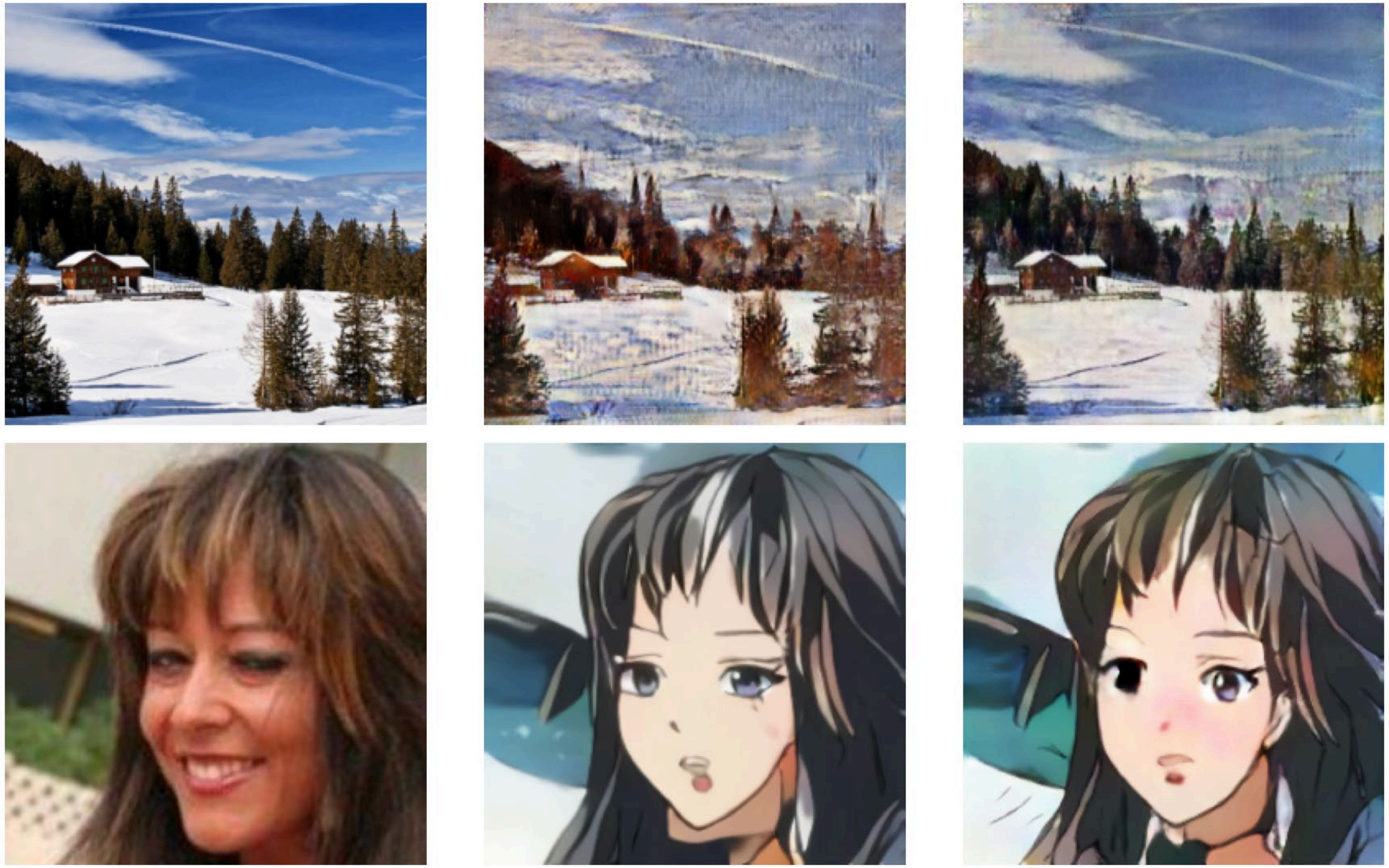


Extracting style-transfer models

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, '21 (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

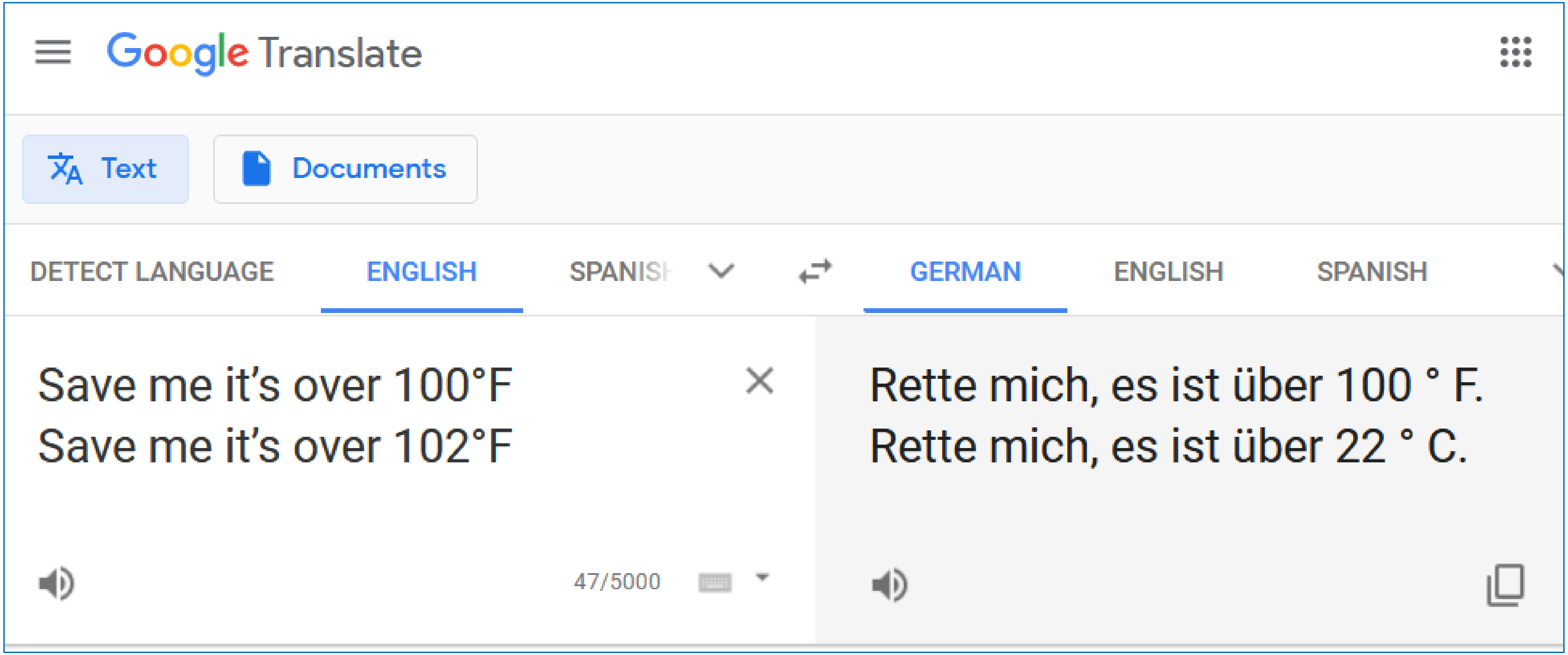
- But they make model extraction easier[1]

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

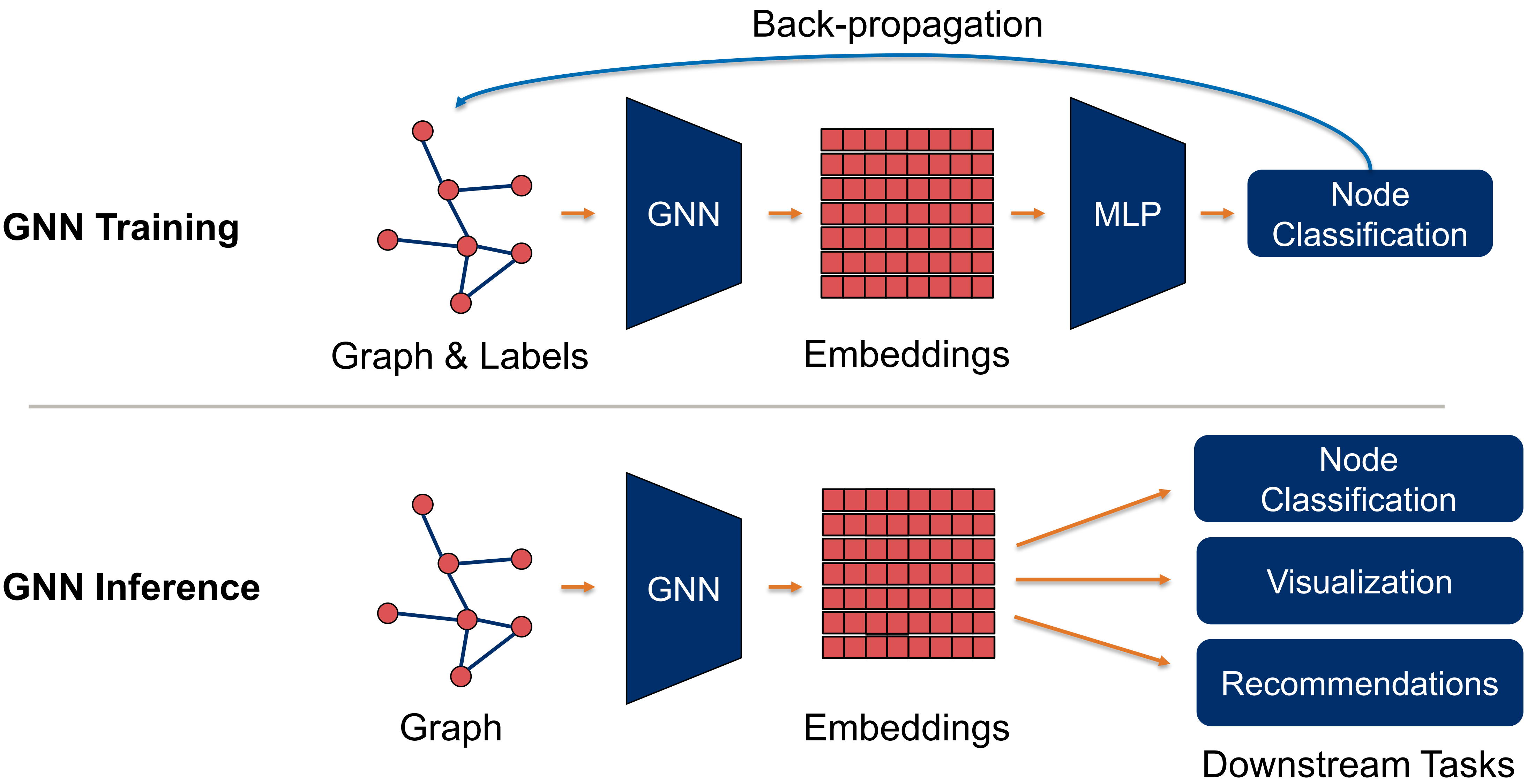
[1] Krishna et al. – *Thieves on Sesame Street! Model Extraction of BERT-based APIs,* ICLR ‘20 (https://iclr.cc/virtual_2020/poster_Byl5NREFDr.html)
[2] Wallace et al. – *Imitation Attacks and Defenses for Black-box Machine Translation Systems,* EMNLP ‘20 (https://arxiv.org/abs/2004.15015)



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
CLONE AFTER SAFETY CONCERNS

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

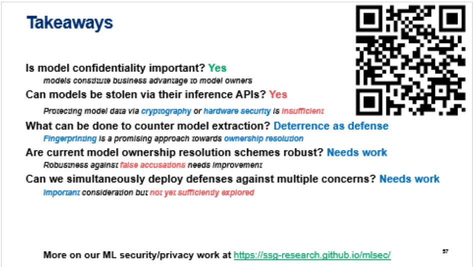
Can models be stolen via their inference APIs? Yes

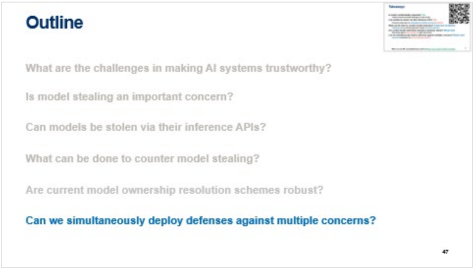
- A powerful (but realistic) adversary can extract complex real-life models
- Detecting such an adversary is difficult/impossible[1]

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses against multiple concerns?





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

Defending against model stealing

We can try to:

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

But current solutions are not effective

Model derivation may even become a desirable business model

Deter unauthorized model ownership via 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)

Watermarking

Embed watermark while training (potentially) victim model[1]

- Choose incorrect labels for a set of samples (watermark set, WM)
- Cannot resist model extraction

Embed watermark at the inference API[2]

- Use a mapping function to decide when to return incorrect predictions for queries
- Finding suitable mapping functions is difficult

Watermarking schemes tend to be not robust[3] and reduce utility

[1] Yadi et al. – *Watermarking Deep Neural Networks by Backdooring,* Usenix SEC ‘18 https://www.usenix.org/node/217594
[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)

Fingerprinting

Conferrable adversarial examples[1]

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

Dataset inference[2]

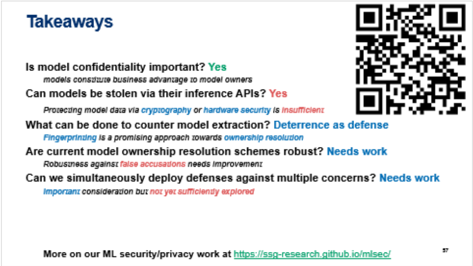
- Distinguish between models trained with different datasets
- Susceptible to false positives/negatives under certain conditions[3]

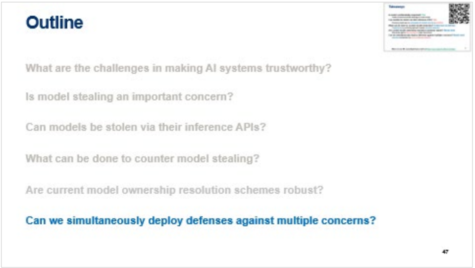
GrOVe[4]

- 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 stolen via their inference APIs?

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses against multiple concerns?

Robustness of model ownership resolution schemes

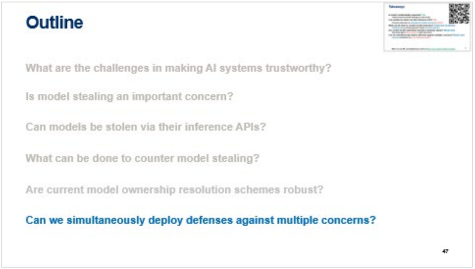
Model ownership resolution (MOR) must be robust against two types of attackers

Malicious suspect:
- tries to evade verification (e.g., pruning, fine-tuning, noising)

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

So far, research has focused on robustness against malicious suspects

False claims against MORs


Outline

What are the challenges in making AI systems trustworthy?

Is model stealing an important concern?

Can models be stolen via their inference APIs?

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses against multiple concerns?

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*, Usenix SEC '24 (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 normal training data (or fine tune)
  - 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)

Watermarking by backdooring[1]: false claim[2]

Watermark generation:

- choose some out-of-distribution samples as watermark
  - assigned with incorrect labels
- train using the watermark alongside your normal training data (or fine tune)
  - 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*, Usenix SEC '24 (https://arxiv.org/abs/2304.06607)

Watermarking by backdooring[1]: false claim[2]

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*, Usenix SEC '24 (https://arxiv.org/abs/2304.06607)

Mitigating false claims against MORs

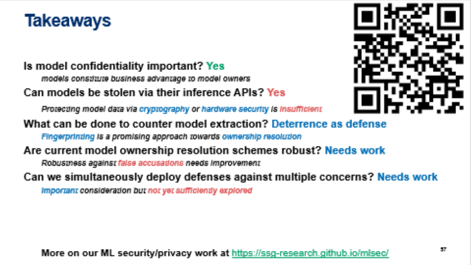
Judge generates watermarks/fingerprints: bottleneck

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*, Usenix SEC ‘24 (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 stolen via their inference APIs?

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses against multiple concerns?

Unintended interactions

Prior work explored defenses to mitigate specific risks

- Defenses typically evaluated only vs. those specific risks they protect against

But practitioners need to deploy multiple defenses simultaneously

- Can two defenses interact negatively with each other?
- Does a defense exacerbate or ameliorate some other (unrelated) risk?

Ownership resolution vs. other 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 do ownership resolution schemes interact with the other defenses?

We investigated pairwise interactions of:

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

Ownership resolution vs. other security/privacy concerns

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

Defense	Dataset	Defense	
		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)

Interaction between ML defenses

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

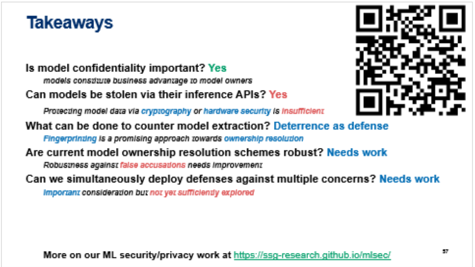
[8] Adam Pearce. 2022. Can a Model Be Differentially Private and Fair? https://pair.withgoogle.com/explorables/private-and-fair/. Online; accessed 7 April 2022.

[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)



Defense vs. other risks

How does a defense impact susceptibility to other (unrelated) risks?

Conjecture: overfitting and memorization are influence defenses and risks

- Effective defenses may induce, reduce or rely on overfitting or memorization
- Risks tend to exploit overfitting or memorization
- Underlying factors that influence memorization/overfitting can be identified.

V. Duddu, S.Szyller, N. Asokan. *SoK: Unintended Interactions among Machine Learning Defenses and Risks,* IEEE S&P '24. https://arxiv.org/abs/2312.04542

Framework: systematizing defenses vs. other risks

Effectiveness of defense <d> correlates with a change in factor <f>

Change in <f> correlates with change in susceptibility to risk <r>

- ↑: positive correlation; ↓: negative correlation

Identify <f> impacted by <d>, and <r> influenced by changes in <f>

Defences (<↑ or ↓>, <**f**>)	**Risks** (<↑ or ↓>, <**f**>)
RD1 (Adversarial Training): • D1 ↑, $\lvert\mathcal{D}_{tr}\rvert$ [161] • D2 ↓, tail length [71], [16] • D4 ↑, priority for learning stable attributes [161] • O1 ↑, curvature smoothness [102] • O2.1 ↑, distinguishability in data records inside and outside $\mathcal{D}_{tr}$ [144] • O3 ↑, distance to boundary for most $\mathcal{D}_{tr}$ data records [176] • M1 ↑, model capacity [102] **RD2 (Outlier Removal):** • D2 ↑, tail length [166] **RD3 (Watermarking):** • D2 ↑, tail length [96] • O2.3 ↓, distinguishability in observables for watermarks between f_θ and f_θ^{der}, but distinct from independent models [3] • M1 ↑, model capacity [3]	**R1 (Evasion):** • D2 ↑, tail length [173], [91] • O1 ↓, curvature smoothness [102] • O3 ↓, distance of $\mathcal{D}_{tr}$ data records to boundary [162] **R2 (Poisoning):** • D2 ↑, tail length [120], [17], [96] • M1 ↑, model capacity [3] **R3 (Unauthorized Model Ownership):** • M1 ↓, model capacity [117], [88] **P1 (Membership Inference):** • D1 ↓, $\lvert\mathcal{D}_{tr}\rvert$ [184], [136] • D2 ↑, tail length [25], [24] • D4 ↓, priority for learning stable attributes [103], [155] • O2.1 ↑, distinguishability for data records inside and outside $\mathcal{D}_{tr}$ [136]

Situating prior work in the framework

Risk increases (🔴) or decreases (🟢) or unexplored (⚪) when a defense is effective

Evaluate the influence of factors empirically (●), theoretically (⊙), conjectured (○)

Defenses	Risks		OVFT	Memorization					Both		References
			D1	D2	D3	D4	O1	O2	O3	M1	
RD1 (Adversarial Training)	R1 (Evasion)	🟢		●			●		●	●	[193], [102], [91], [173]
	R2 (Poisoning)	🔴									[170], [153]
	R3 (Unauthorized Model Ownership)	🔴	○								[86] ([95]: 🟢)
	P1 (Membership Inference)	🔴	⊙, ●					1: ●		●	[144], [67]
	P2 (Data Reconstruction)	🔴				○				●	[195], [111]
	P3 (Attribute Inference)	⚪									
	P4 (Distribution Inference)	🔴				○					[148]
	F (Discriminatory Behaviour)	🔴		⊙, ●							[16], [36], [71], [99]
RD2 (Outlier Removal)	R1 (Evasion)	🟢									[59]
	R2 (Poisoning)	🟢									[154]
	R3 (Unauthorized Model Ownership)	⚪									
	P1 (Membership Inference)	🔴		●							[25], [46]
	P2 (Data Reconstruction)	⚪									
	P3 (Attribute Inference)	🔴		●							[78]
	P4 (Distribution Inference)	⚪									
	F (Discriminatory Behaviour)	🔴	●	○							[134]
RD3 (Watermarking)	R1 (Evasion)	⚪									
	R2 (Poisoning)	🔴		○							[133], [3], [194], [93]
	R3 (Unauthorized Model Ownership)	🟢		○				3: ●	●		[152], [3], [98]
	P1 (Membership Inference)	🔴		○				1: ●	●		[157], [33]
	P2 (Data Reconstruction)	🔴		○				1: ●	●		[157]
	P3 (Attribute Inference)	🔴		○				2: ●	●		[157]
	P4 (Distribution Inference)	🔴	⊙, ●	○				1: ●	●	●	[30], [105]

Guideline for conjecturing unintended interactions

For defense <d>, risk <r> and common factor <f>, use pair of arrows that describe how <d> and <r> correspond to <f>

Conjectured interaction for a given <f>:

- If arrows align (↑,↑) or (↓,↓) ➞ <r> increases when <d> is effective (🔴)
- Else for (↑,↓) or (↓,↑) ➞ <r> decreases when <d> is effective (🟢)

Conjectured overall interaction: consider conjectures from all <f>s:

- If all <f> agree, then conjectured overall interaction is unanimous
- Otherwise, prioritize conjecture from dominant <f> (dominance may depend on attack)
- Value of a non-common factor may affect overall interaction

V. Duddu, S.Szyller, N. Asokan. *SoK: Unintended Interactions among Machine Learning Defenses and Risks,* IEEE S&P ‘24. https://arxiv.org/abs/2312.04542

Group fairness (FD1) vs. data reconstruction (P2)

Conjectured Interaction from common factor:

O2.2 Distinguishability across subgroups: FD1 ↓, P2 ↑ (➞ ●)

Non-common factor: D3 # Attributes -- risk may decrease with D3

Empirical Evidence

Fair model ➞ lower attack success (confirms ●)

- Lowers distinguishability across subgroups

Metric	Baseline	Fair Model
Accuracy	84.40 ± 0.09	77.96 ± 0.58
Recon. Loss	0.85 ± 0.01	0.95 ± 0.02

Non-common factor D3

attributes = 10:

- Fair model ➞ lower attack success

attributes > 10:

- Fair model ➞ no change in attack success
  (note: # attributes do not affect accuracy drop caused by fairness)

#Attributes	Baseline		Fair Model	
	Recon. Loss	Accuracy	Recon. Loss	Accuracy
10	0.85 ± 0.01	84.40 ± 0.09	0.95 ± 0.02	78.96 ± 0.58
20	0.93 ± 0.03	84.72 ± 0.22	0.93 ± 0.00	80.32 ± 1.12
30	0.95 ± 0.02	84.41 ± 0.39	0.94 ± 0.00	79.50 ±0.91

V. Duddu, S.Szyller, N. Asokan. *SoK: Unintended Interactions among Machine Learning Defenses and Risks,* IEEE S&P '24. https://arxiv.org/abs/2312.04542

Takeaways

Is model confidentiality important? Yes

models constitute business advantage to model owners

Can models be stolen 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 defenses against multiple concerns? Needs work

Important consideration but not yet sufficiently explored

More on our ML security/privacy work at https://ssg-research.github.io/mlsec/

Takeaways

Is model confidentiality important? Yes

models constitute business advantage to model owners

Can models be stolen 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 defenses against multiple concerns? Needs work

Important consideration but not yet sufficiently explored

Other research topics:
ML security/privacy: property attestation of ML models, robust concept removal from generative models
Platform security: hardware-assisted run-time security, secure outsourced computing

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

Dominant factors

LEGEND

O1 Curvature smoothness of the objective function
O2 Distinguishability of model observables across datasets (**O2.1**), subgroups (**O2.2**), and models (**O2.3**)
O3 Distance of training data to decision boundary

D1 Size of training data
D2 Tail length of distribution
D3 Number of attributes inversely
D4 Priority of learning stable attributes
M1 Model capacity

Active factors are exploited by the attacks: O1, O2, O3

Passive factors (data/model configuration): D1, D2, D3, D4, M1

Attacks often exploit dynamic factors, we deem them "dominant"

PD1 (Differential Privacy) and R1 (Evasion)➞ ● [1,2]

- D2 ➞ ●; O1 ➞ ●; O3 ➞ ●

FD1 (Group Fairness) and P1 (Membership Inference) ➞ ●[3]

- D4 ➞ ●; O3 ➞ ●

[1] Tursynbek et al. *Robustness threats of Differential Privacy*. NeurIPS Privacy Preserving ML Workshop. 2020. https://arxiv.org/abs/2012.07828
[2] Boenisch et al.. *Gradient masking and the underestimated robustness threats of differential privacy in deep learning.* ArXiv 2021. https://arxiv.org/abs/2105.07985
[3] Chang and Shokri. *On the Privacy Risks of Algorithmic Fairness.* EuroS&P 2021. https://arxiv.org/abs/2011.03731