



UNIVERSITY OF  
**WATERLOO**

# Meta concerns in ML security/privacy

*N. Asokan*

 <https://asokan.org/asokan/>

   @nasokan

*(Joint work with Vasisht Duddu, Jian Liu, Sebastian Szyller, Asim Waheed, Rui Zhang)*

# My research interests

## 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 design/use hardware assistance to secure software?



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

# Outline

The **big picture**: studying ML security/privacy — **why and how?**

What can be done to counter **model stealing**?

Are we using the **right adversary models**?

(How) can we simultaneously **deploy defenses against multiple concerns**?

# Outline

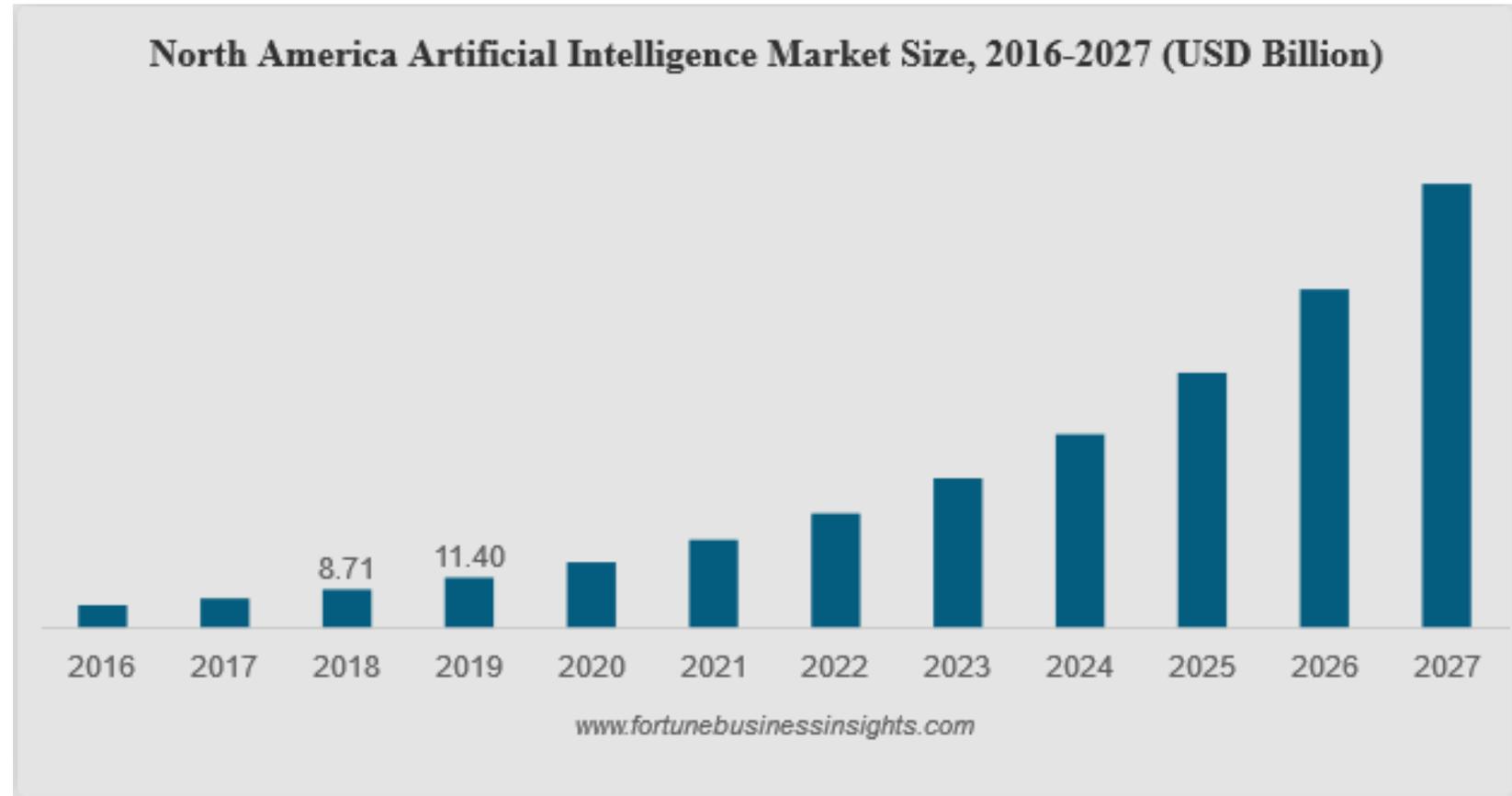
The **big picture**: studying ML security/privacy — **why and how?**

What can be done to counter **model stealing**?

Are we using the **right adversary models**?

(How) 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 Artificial 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-artificial-intelligence-is-advancing-precision-medicine/#2f720a79a4d5>

**MOTHERBOARD**  
TECH BY VICE

# Dozens of Cities Have Secretly Experimented With Predictive Policing Software

Documents obtained by Motherboard requests verify previously unconfirmed reports that dozens of cities have experimented with predictive policing company Palantir's software.



By **Caroline Haskins**

[https://www.vice.com/en\\_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software](https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software)

Forbes

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

# How AI Is Uprooting Recruiting



**Falon Fatemi** Contributor

Entrepreneurs

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

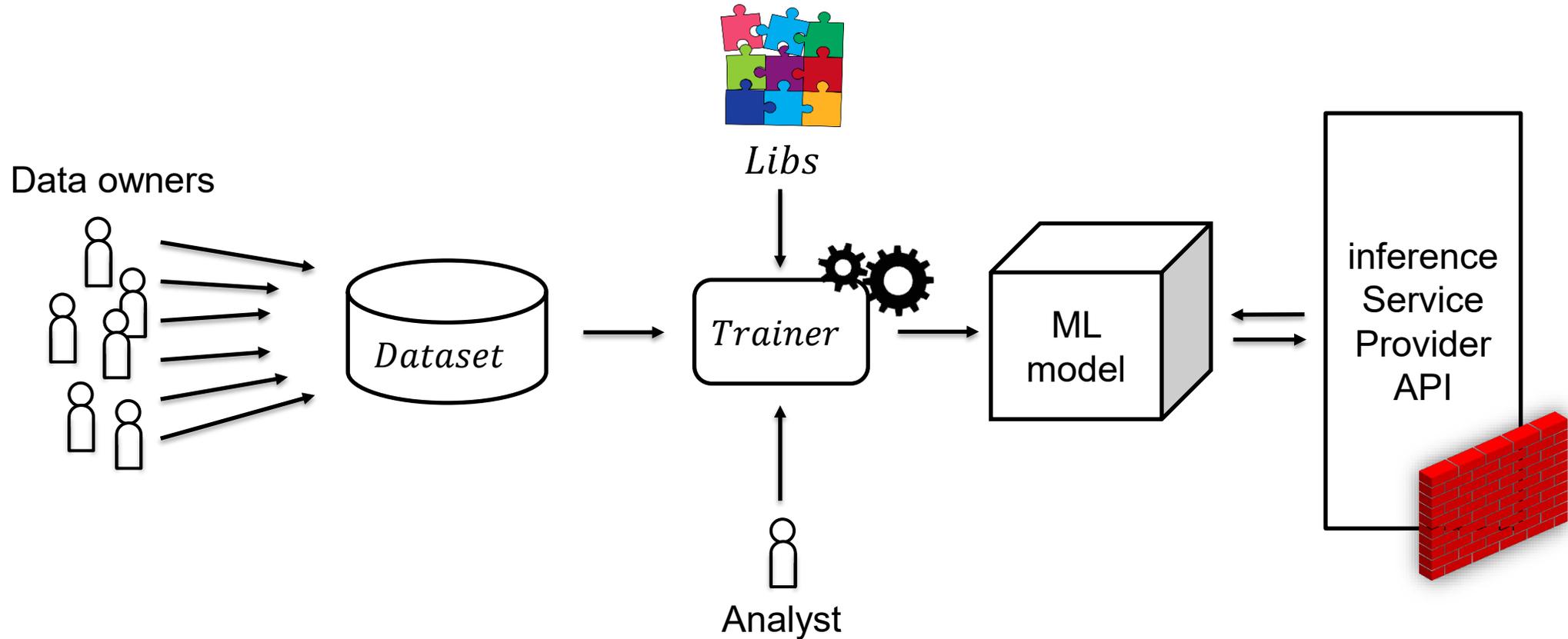
Artificial intelligence and machine learning tools could go a long way to helping to fight cybercrime. But these technologies aren't a silver bullet, and could also be exploited by malicious hackers.

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



[https://www.vice.com/en\\_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software](https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software)

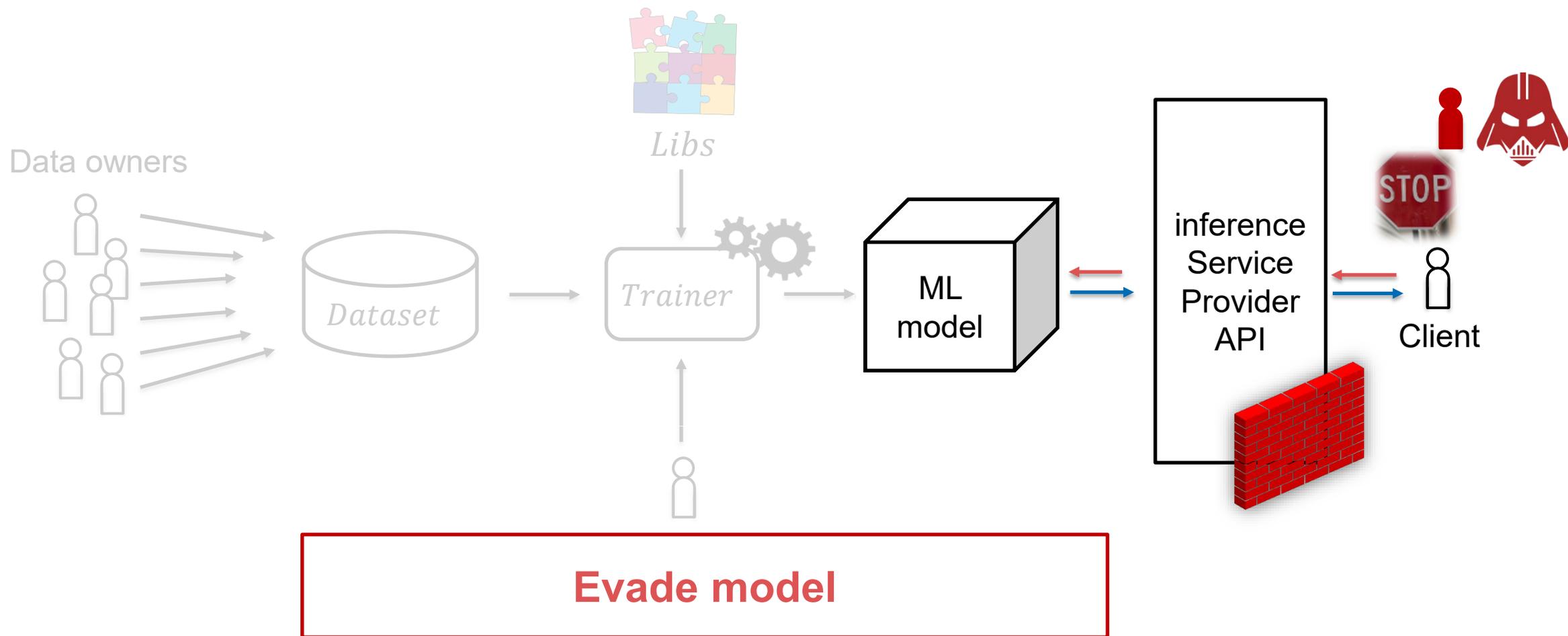
# Machine Learning pipeline



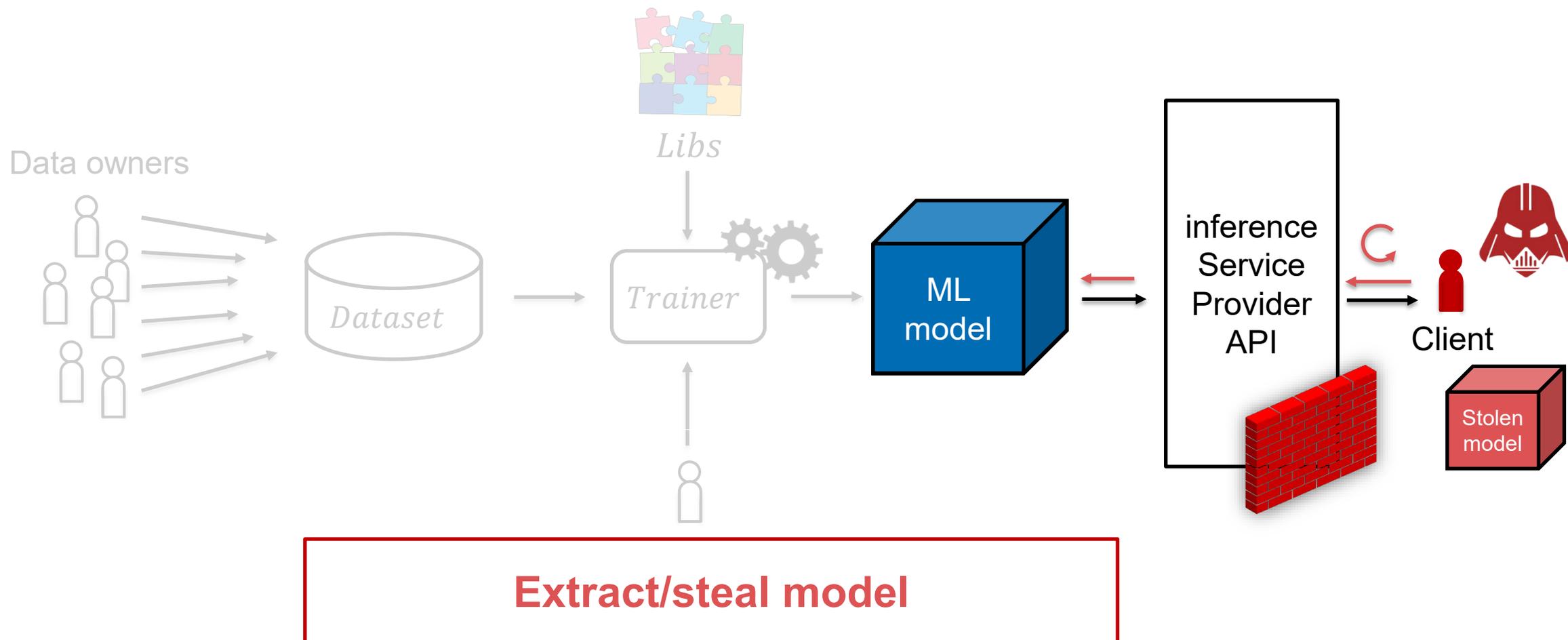
**Where is the adversary? What is its target?**



# Compromised input – Model integrity



# Malicious client – Model confidentiality



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>)

# Towards trustworthy AI

## Secure, privacy-preserving, ...

TABLE V  
TOP ATTACK

<i>Which attack would affect your org the most?</i>	<i>Distribution</i>
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

# Outline

The **big picture**: studying ML security/privacy — **why and how?**

What can be done to counter **model stealing**?

Are we using the **right adversary models**?

(How) can we simultaneously **deploy defenses against multiple concerns**?

**Takeaways**



Are we using the **right adversary models**? *Needs work*  
*Robustness against raise accusations in MOCs needs improvement*  
*More generally, ML security/privacy research needs widely accepted, streamlined adversary models*

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://sag-research.github.io/misec/>

**Outline**



The **big picture**: studying ML security/privacy — **why and how?**

What can be done to counter **model stealing**?

Are we using the **right adversary models**?

(How) can we simultaneously **deploy defenses against multiple concerns**?

# Defending against model stealing

**We can try to:**

- **prevent** (or slow down<sup>[1]</sup>) **model extraction**, or
- **detect**<sup>[2]</sup> 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?*, AAI-EDSML '20 (<https://arxiv.org/abs/1910.05429>)

# Watermarking

## Embed watermark while training (potentially) victim model<sup>[1]</sup>

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

## Embed watermark at the inference API<sup>[2]</sup>

- 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**<sup>[3]</sup> 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<sup>[1]</sup>

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

## Dataset inference<sup>[2]</sup>

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

## GrOVe<sup>[4]</sup>

- Use GNN **embeddings as fingerprints** (for GNN models)
- Effective against high-fidelity extraction<sup>[5]</sup> 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

The **big picture**: studying ML security/privacy — **why and how?**

What can be done to counter **model stealing**?

Are we using the **right adversary models**?

(How) can we simultaneously **deploy defenses against multiple concerns**?

**Takeaways**



Are we using the **right adversary models**? *Needs work*  
*Robustness against raise accusations in MOCs needs improvement*  
*More generally, ML security/privacy research needs widely accepted, streamlined adversary models*

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://sag-research.github.io/mlsec/>

**Outline**



The **big picture**: studying ML security/privacy — **why and how?**

What can be done to counter **model stealing**?

Are we using the **right adversary models**?

**(How) 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 adversaries

## 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

**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**

# Watermarking by backdooring<sup>[1]</sup>

## 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<sup>[1]</sup>: false claim<sup>[2]</sup>

## 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<sup>[1]</sup>: false claim<sup>[2]</sup>

## 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**

# The Meta Concern: sensible adversary models

**Identify potential adversaries and their goals**

**Identify adversary's knowledge and capabilities:**

- Data access:
  - vis-à-vis target's training data (overlap/distribution/domain? natural/synthetic?)
  - vis-à-vis target's inferences
- Target model access: white-box/black-box/grey-box?
- Adversary type: honest-but-curious vs. malicious
- Interaction type: zero-shot/one-shot/query-budget?, adaptive?

**Avoid sloppy terminology!**

- “**adversarial attacks**” → there are no benign attacks!
- “**adaptive adversaries**” → cf. Kerchoff's principle
- ...

# Outline

Takeaways



Are we using the right adversary models? *Needs work*  
Robustness against *raise accusations in MDRs* needs improvement  
More generally, ML security/privacy research needs widely accepted, streamlined adversary models

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://sag-research.github.io/mlsec/>

The **big picture**: studying ML security/privacy — **why and how?**

What can be done to counter **model stealing**?

Are we using the **right adversary models**?

(How) can we simultaneously **deploy defenses against multiple concerns**?

# Unintended interactions

## Prior work explored **defenses** to mitigate **specific risks**

- Defenses typically evaluated only vs. 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?

**Takeaways**



Are we using the right adversary models? **Needs work**  
Robustness against **raise accusations in MLOps** needs improvement  
More generally, ML security/privacy research needs widely accepted, streamlined adversary models

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://sage-research.github.io/mlsec/>

27

**Defense vs. other risks**



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

Conjecture: **overfitting** and **memorization** are influence defenses and risks<sup>[1][2]</sup>

- 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

Recently built a toolkit, **Amulet**, for comparative evaluation of attacks & defenses<sup>[3]</sup>

Currently working on "how to easily determine if a given set of defenses conflict?"<sup>[4]</sup>

[1] Smith, Rybin, and Asokan – Spurious Correlations between Machine Learning Defenses and Risks, IJCV 34:1 (2022)  
[2] Wang et al. – <https://arxiv.org/abs/2207.01991>  
[3] Amulet: <https://github.com/sage-research/amulet>  
[4] Smith, Zhang, Asokan – Conflicting Machine Learning Defenses without Conflict, <https://arxiv.org/abs/2207.01991>

27

# 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

data watermarking

fingerprinting

**WITH**

differential privacy

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 ( $\phi_{ACC}$ ) **or**
- metric for **A** ( $\phi_A$ ) **or**
- metric for **B** ( $\phi_B$ )

then **A** and **B** are in **conflict**

Defense	Dataset	Defense	
		DP	ADV. TR.
WM	MNIST	$\phi_{ACC}$ $\phi_{WM}$	$\phi_{ACC}$ $\phi_{WM}$ $\phi_{ADV}$
	FMNIST	$\phi_{ACC}$ $\phi_{WM}$	$\phi_{ACC}$ $\phi_{WM}$ $\phi_{ADV}$
	CIFAR10	$\phi_{ACC}$ $\phi_{WM}$	$\phi_{ACC}$ $\phi_{WM}$ $\phi_{ADV}$
RAD-DATA	MNIST	$\phi_{ACC}$ $\phi_{RAD-DATA}$	$\phi_{ACC}$ $\phi_{RAD-DATA}$ $\phi_{ADV}$
	FMNIST	$\phi_{ACC}$ $\phi_{RAD-DATA}$	$\phi_{ACC}$ $\phi_{RAD-DATA}$ $\phi_{ADV}$
	CIFAR10	$\phi_{ACC}$ $\phi_{RAD-DATA}$	$\phi_{ACC}$ $\phi_{RAD-DATA}$ $\phi_{ADV}$
DI	MNIST	$\phi_{ACC}$ $\phi_{DI}$	$\phi_{ACC}$ $\phi_{DI}$ $\phi_{ADV}$
	FMNIST	$\phi_{ACC}$ $\phi_{DI}$	$\phi_{ACC}$ $\phi_{DI}$ $\phi_{ADV}$
	CIFAR10	$\phi_{ACC}$ $\phi_{DI}$	$\phi_{ACC}$ $\phi_{DI}$ $\phi_{ADV}$

# 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](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>

# Defense vs. other risks

**Protecting Against Multiple Risks**

Combine **existing** defenses *effectively* while avoiding conflicts

- not incur a **drop in effectiveness** constituent defenses

**Desiderata**

- accurate: **correctly identifies** whether a combination is effective or not
- scalable: allows combining **more than two defenses**
- non-invasive: requires **no changes** to the defenses being combined
- general: **applicable** to different types of defenses

Prior combination techniques do not meet all requirements

- Need a **principled approach** to combine existing defenses *without modification*



**Takeaways**

Are we using the right adversary models? **Needs work**

Robustness against **raise accusations in MDRs** needs improvement

More generally, ML security/privacy research **needs widely accepted, streamlined adversary models**

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://sag-research.github.io/misec/>

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

**Conjecture: overfitting and memorization are influence defenses and risks<sup>[1][2]</sup>**

- 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

**Distinguished Paper Award**

Recently built a toolkit, **Amulet**, for comparative evaluation of attacks & defenses<sup>[3]</sup>

Currently working on “how to easily determine if a given set of defenses conflict?”<sup>[4]</sup>

[1] Duddu, Szyller, and Asokan - *SoK: Unintended Interactions among Machine Learning Defenses and Risks*, IEEE S&P '24. (<https://arxiv.org/abs/2312.04542>)

[2] Blog article: <https://crysp.uwaterloo.ca/ssg/blog/2024/05/unintended-interactions-among-ml.html>

[3] Amulet repo: <https://github.com/ssg-research/amulet>

[4] Duddu, Zhang, Asokan – Combining Machine learning Defenses without Conflicts. (<https://arxiv.org/abs/2411.09776>)

# Factors influencing overfitting and memorization

**O1** Curvature smoothness of the objective function

**O2** Distinguishability 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

**D4** Priority of learning stable attributes

**M1** Model capacity

# Framework: systematizing defenses vs. other risks

Effectiveness of defense  $\langle d \rangle$  correlates with a change in factor  $\langle f \rangle$

Change in  $\langle f \rangle$  correlates with change in susceptibility to risk  $\langle r \rangle$

- $\uparrow$ : positive correlation;  $\downarrow$ : negative correlation

Identify  $\langle f \rangle$  impacted by  $\langle d \rangle$ , and  $\langle r \rangle$  influenced by changes in  $\langle f \rangle$

Defences ( $\langle \uparrow \text{ or } \downarrow \rangle$ , $\langle f \rangle$ )	Risks ( $\langle \uparrow \text{ or } \downarrow \rangle$ , $\langle f \rangle$ )
<p><b>RD1 (Adversarial Training):</b></p> <ul style="list-style-type: none"> <li>• D1 <math>\uparrow</math>, <math> \mathcal{D}_{tr} </math> [161]</li> <li>• D2 <math>\downarrow</math>, tail length [71], [16]</li> <li>• D4 <math>\uparrow</math>, priority for learning stable attributes [161]</li> <li>• O1 <math>\uparrow</math>, curvature smoothness [102]</li> <li>• O2.1 <math>\uparrow</math>, distinguishability in data records inside and outside <math>\mathcal{D}_{tr}</math> [144]</li> <li>• O3 <math>\uparrow</math>, distance to boundary for most <math>\mathcal{D}_{tr}</math> data records [176]</li> <li>• M1 <math>\uparrow</math>, model capacity [102]</li> </ul> <p><b>RD2 (Outlier Removal):</b></p> <ul style="list-style-type: none"> <li>• D2 <math>\uparrow</math>, tail length [166]</li> </ul> <p><b>RD3 (Watermarking):</b></p> <ul style="list-style-type: none"> <li>• D2 <math>\uparrow</math>, tail length [96]</li> <li>• O2.3 <math>\downarrow</math>, distinguishability in observables for watermarks between <math>f_\theta</math> and <math>f_\theta^{der}</math>, but distinct from independent models [3]</li> <li>• M1 <math>\uparrow</math>, model capacity [3]</li> </ul>	<p><b>R1 (Evasion):</b></p> <ul style="list-style-type: none"> <li>• D2 <math>\uparrow</math>, tail length [173], [91]</li> <li>• O1 <math>\downarrow</math>, curvature smoothness [102]</li> <li>• O3 <math>\downarrow</math>, distance of <math>\mathcal{D}_{tr}</math> data records to boundary [162]</li> </ul> <p><b>R2 (Poisoning):</b></p> <ul style="list-style-type: none"> <li>• D2 <math>\uparrow</math>, tail length [120], [17], [96]</li> <li>• M1 <math>\uparrow</math>, model capacity [3]</li> </ul> <p><b>R3 (Unauthorized Model Ownership):</b></p> <ul style="list-style-type: none"> <li>• M1 <math>\downarrow</math>, model capacity [117], [88]</li> </ul> <p><b>P1 (Membership Inference):</b></p> <ul style="list-style-type: none"> <li>• D1 <math>\downarrow</math>, <math> \mathcal{D}_{tr} </math> [184], [136]</li> <li>• D2 <math>\uparrow</math>, tail length [25], [24]</li> <li>• D4 <math>\downarrow</math>, priority for learning stable attributes [103], [155]</li> <li>• O2.1 <math>\uparrow</math>, distinguishability for data records inside and outside <math>\mathcal{D}_{tr}</math> [136]</li> </ul>

# Situating prior work in the framework

**Takeaways**

Are we using the right adversary models? *Needs work*  
 Robustness against *raise accusations in MOCs* needs improvement  
 More generally, ML security/privacy research needs widely accepted, streamlined adversary models

Can we simultaneously deploy defenses against multiple concerns? *Needs work*  
 Important consideration but *not yet sufficiently explored*

More general ML security/privacy work at <https://sag-research.github.io/mlsec/>



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
<b>RD1</b> (Adversarial Training)	<b>R1</b> (Evasion)	●		●				●		●	[193], [102], [91], [173]	
	<b>R2</b> (Poisoning)	●									[170], [153]	
	<b>R3</b> (Unauthorized Model Ownership)	●	○								[86] ([95]: ●)	
	<b>P1</b> (Membership Inference)	●	⊖, ●						1: ●		●	[144], [67]
	<b>P2</b> (Data Reconstruction)	●					○				●	[195], [111]
	<b>P3</b> (Attribute Inference)	●										[148]
	<b>P4</b> (Distribution Inference)	●					○					[16], [36], [71], [99]
<b>RD2</b> (Outlier Removal)	<b>R1</b> (Evasion)	●									[59]	
	<b>R2</b> (Poisoning)	●									[154]	
	<b>R3</b> (Unauthorized Model Ownership)	●										
	<b>P1</b> (Membership Inference)	●									[25], [46]	
	<b>P2</b> (Data Reconstruction)	●										
	<b>P3</b> (Attribute Inference)	●										[78]
	<b>P4</b> (Distribution Inference)	●										[134]
<b>RD3</b> (Watermarking)	<b>R1</b> (Evasion)	●										
	<b>R2</b> (Poisoning)	●										
	<b>R3</b> (Unauthorized Model Ownership)	●									[133], [3], [194], [93]	
	<b>P1</b> (Membership Inference)	●							3: ●		●	[152], [3], [98]
	<b>P2</b> (Data Reconstruction)	●							1: ●		●	[157], [33]
	<b>P3</b> (Attribute Inference)	●							1: ●		●	[157]
	<b>P4</b> (Distribution Inference)	●	⊖, ●						2: ●		●	[157]
								1: ●		●	[30], [105]	

# Guideline for conjecturing unintended interactions

For defense  $\langle d \rangle$ , risk  $\langle r \rangle$  and common factor  $\langle f \rangle$ , use pair of arrows that describe how  $\langle d \rangle$  and  $\langle r \rangle$  correspond to  $\langle f \rangle$

**Conjectured interaction for a given  $\langle f \rangle$ :**

- If arrows align  $(\uparrow, \uparrow)$  or  $(\downarrow, \downarrow) \rightarrow \langle r \rangle$  **increases** when  $\langle d \rangle$  is effective (●)
- Else for  $(\uparrow, \downarrow)$  or  $(\downarrow, \uparrow) \rightarrow \langle r \rangle$  **decreases** when  $\langle d \rangle$  is effective (●)

**Conjectured overall interaction: consider conjectures from all  $\langle f \rangle$ s:**

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

Takeaways



Are we using the right adversary models? Needs work  
Robustness against **more adversaries in SCLC**: needs improvement  
More generally, ML security/privacy research needs **widely accepted, streamlined adversary models**

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/>

# 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

# Protecting Against Multiple Risks

Combine **existing** defenses *effectively* while avoiding conflicts

- not incur a **drop in effectiveness** constituent defenses

## Desiderata

- accurate: **correctly identifies** whether a combination is effective or not
- scalable: allows combining **more than two defenses**
- non-invasive: requires **no changes** to the defenses being combined
- general: **applicable** to different types of defenses

**Prior combination techniques do not meet all requirements**

- Need a **principled approach** to combine existing defenses *without modification*



# Combining ML Defenses without Conflicts

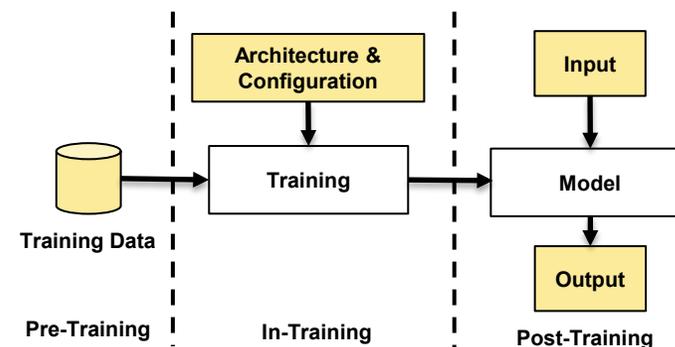
**Intuition: account for reasons underlying conflicts among defenses**

For  $D_1$  and  $D_2$  applied in that order, there can be a conflict if

- $D_1$  uses a risk protected by  $D_2$
- Changes by  $D_2$  overrides changes by  $D_1$

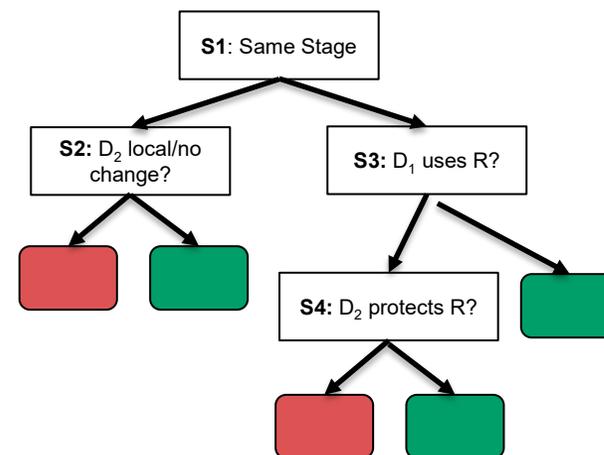
**Observation:**

- ML defenses operate on one of three stages of ML pipelines



**DEF\CON: quickly identify effective combinations**

- 90% accuracy on eight combinations from prior work
- 81% in 30 previously unexplored combinations



# Takeaways



Are we using the right adversary models? **Needs work**

*Robustness against **false accusations in MORs** needs improvement*

*More generally, ML security/privacy research needs **widely accepted, streamlined adversary models***

Can we simultaneously deploy defenses against multiple concerns? **Needs work**

*Important consideration but **not yet sufficiently explored***

# Takeaways



Are we using the right adversary models? **Needs work**

*Robustness against **false accusations in MORs** needs improvement*

*More generally, ML security/privacy research needs **widely accepted, streamlined adversary models***

Can we simultaneously deploy defenses against multiple concerns? **Needs work**

*Important consideration but **not yet sufficiently explored***

Other research topics:

ML security/privacy:

ML **ownership resolution**, **Conflicting ML defenses**, ML **property attestation**, robust **concept removal** in gen AI

Platform security: **hardware-assisted** run-time security, secure outsourced computing

Open (postdoc, grad student) positions to help lead our work: ML security/privacy, platform security

<https://asokan.org/asokan/research/SecureSystems-open-positions-Jan2024.php>