

# Machine Learning Security

Methods and Strategies to attack Machine Learning Models

Manuel Pasieka , 2. June 2022

# Manuel Pasieka



- AI Consultant – “*Leverage the power of machine learning for your business*”
- Host of the Austrian AI Podcast



What do you think of when hearing ...

Cyber security + machine learning



What do you think of when hearing ...



# Content

1. Machine Learning in Cyber Security
2. DevSecOps / MLSecOps
3. Machine Learning Security



# Content

1. Machine Learning in Cyber Security
2. DevSecOps / MLSecOps
3. Machine Learning Security



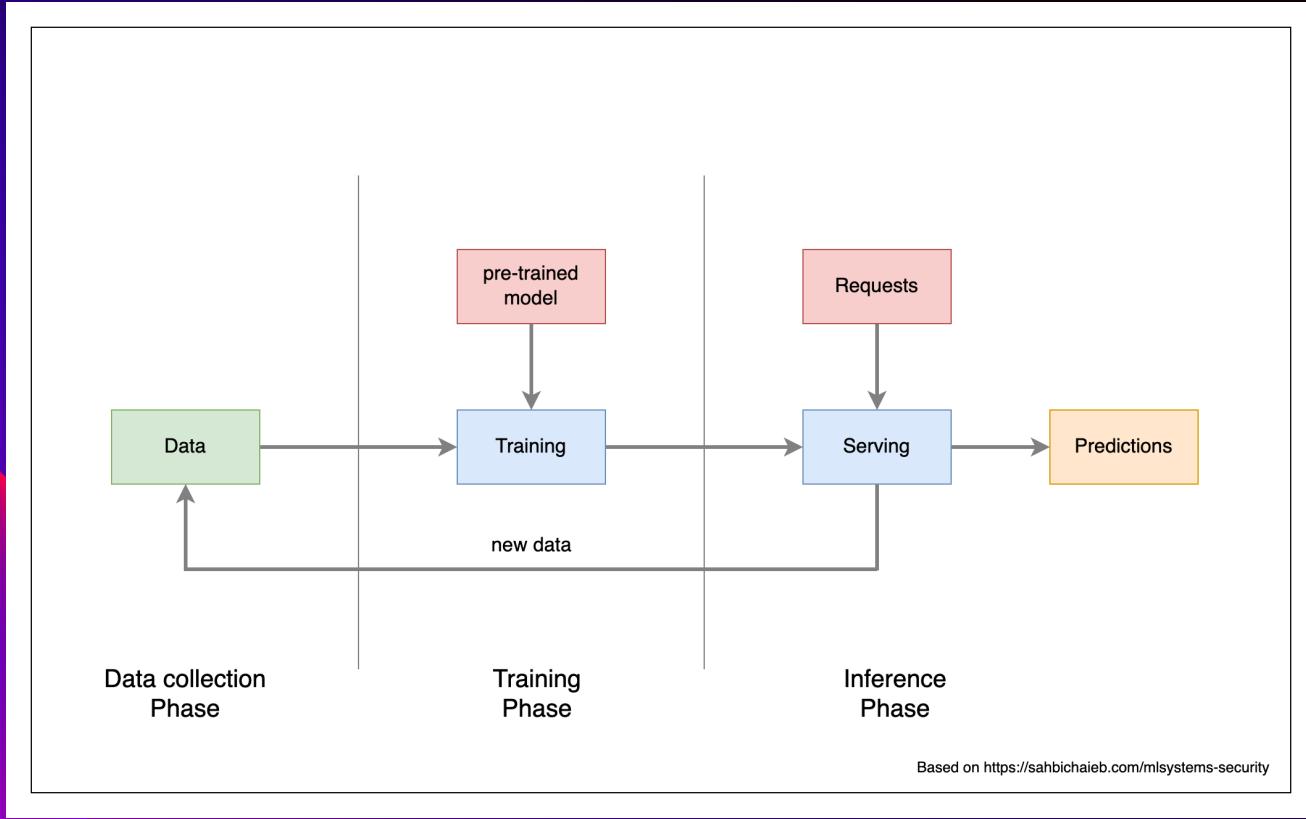
# Machine Learning Security

MLSec studies the abuse,  
theft and exploitation of  
machine learning models.

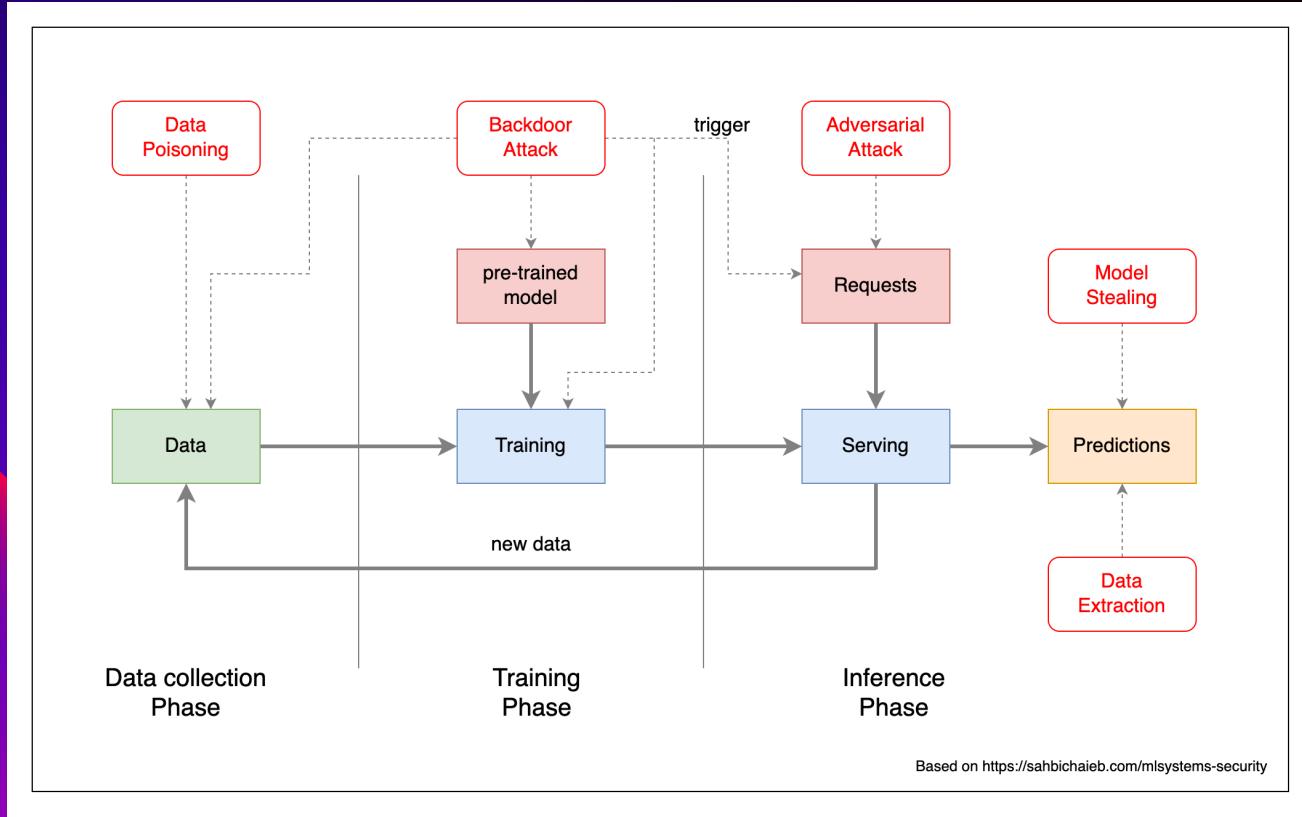
- Attack Surface
- Taxonomy & Terminology
- Model Stealing
- Data Extraction
- Adversarial Attack
- Data Poisoning
- Backdoor Attack
- What to do about it?
- References and utilities



# Attack Surface



# Attack Surface



# Taxonomy and Terminology

## Specificity of an attack

- Targeted : Affect a single class / user
- Indiscriminate : Affect the system as a whole

## Influence of an attack

- Causative : Affect model training / training data
- Exploratory : Affect model inference

## System Exposure

- black-box : The attack can query the model (API)
- white-box : The attack has direct access to the model

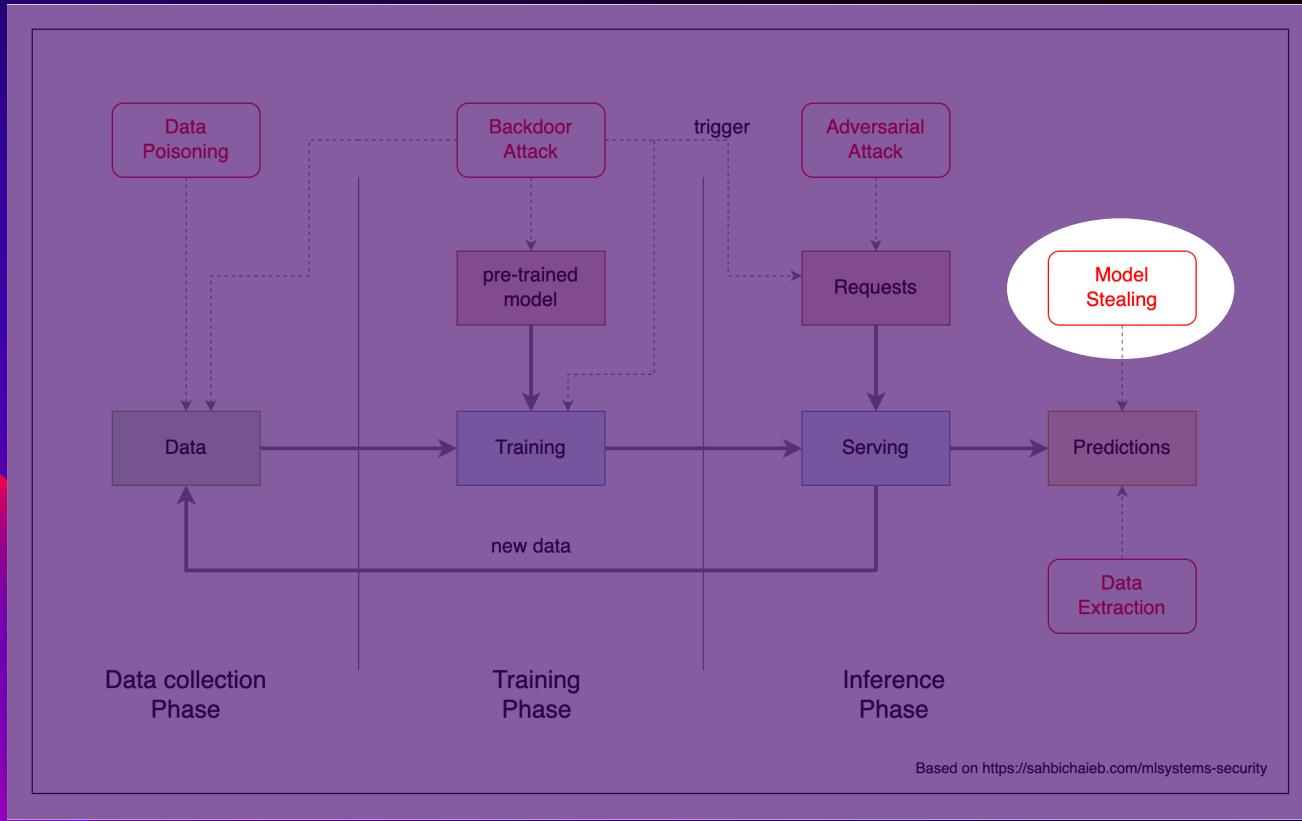
## Impact on a compromised System – Information Security Triad (CIA)

- Confidentiality : Unauthorized access to data / information
- Integrity : System is not operating as designed
- Availability : System outage



# Model Stealing

# Model Stealing

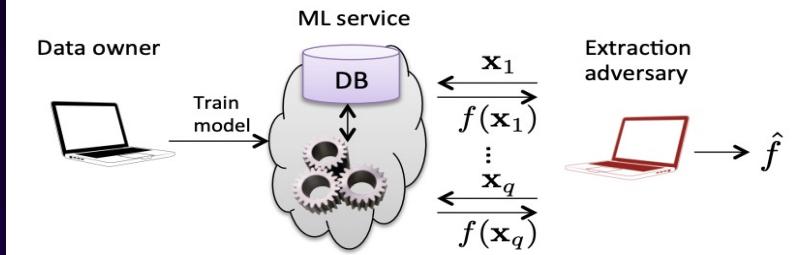


# Model Stealing

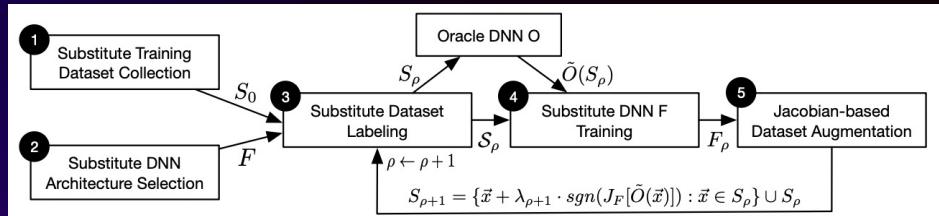
An attacker users black-box access to the target model to create a substitution/surrogate model infer<sup>1</sup>

- Architecture
- Training hyperparameters
- Parameters/Weights
- Decision boundaries

Highly dependent on Query efficiency  
and dependent on oracle access  
non-adaptive/adaptive/strict  
Create Proxy ML Model (Attack Staging)



Florian Tramèr et al. "Stealing Machine Learning Models via Prediction APIs", (2016)

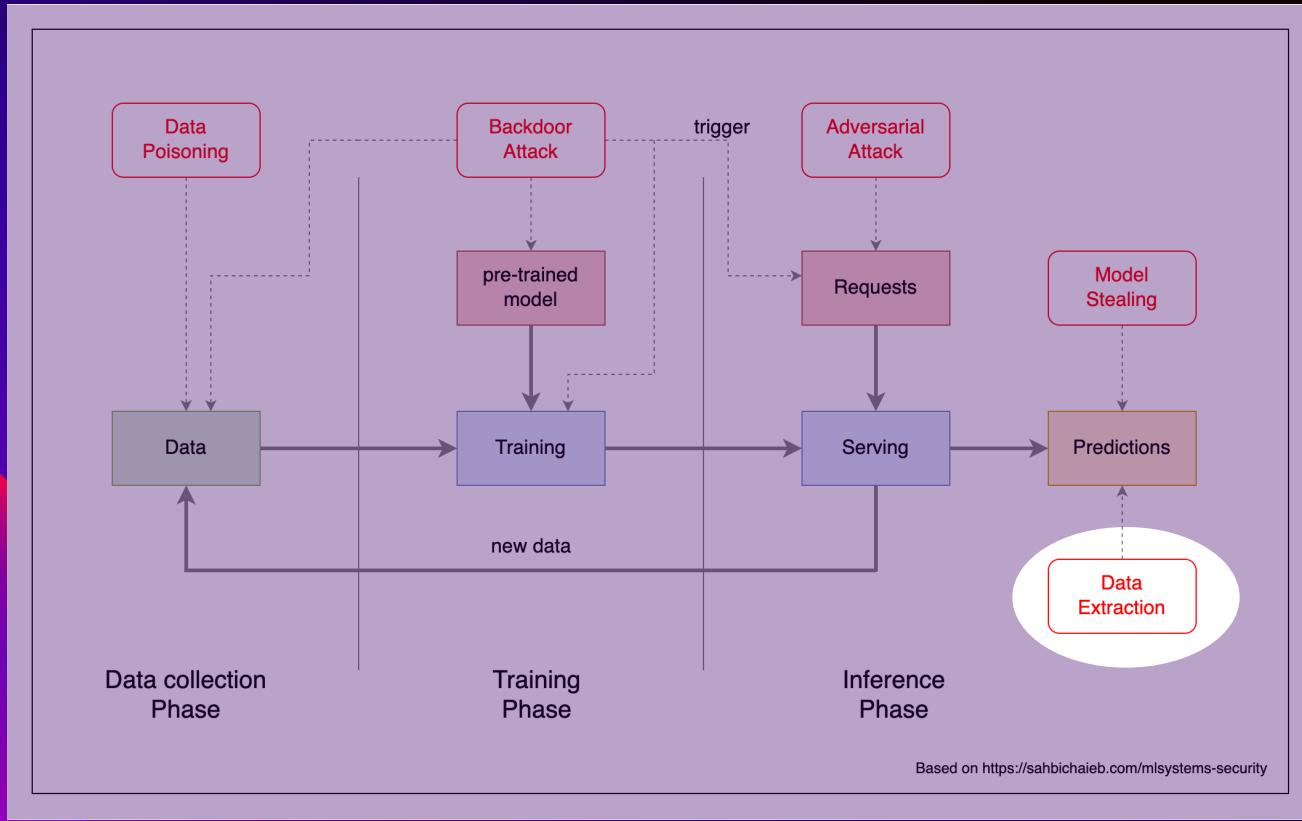


Papernot et al. "Practical Black-Box Attacks against Machine Learning", (2017)



# Data Extraction

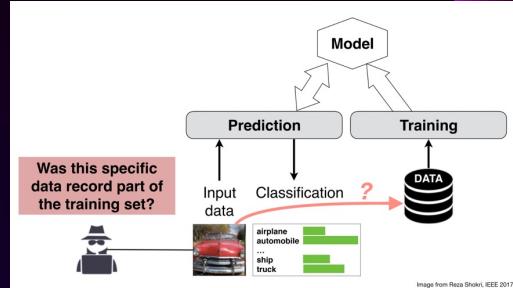
# Attack Surface



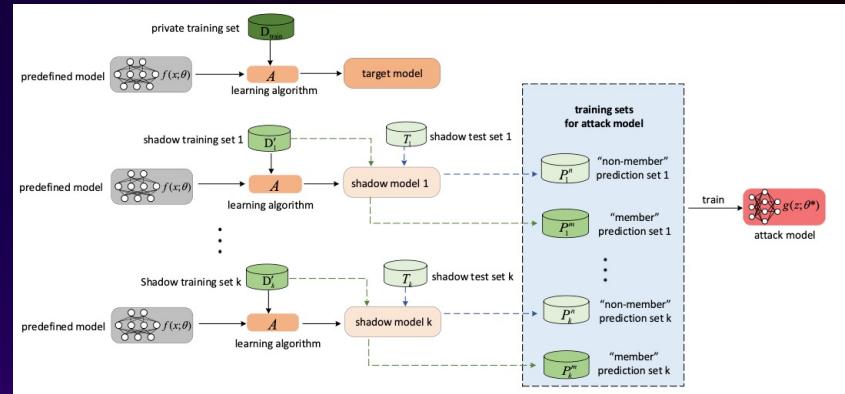
# Training Data Extraction

The capability of an attack to infer the presence of individual training records.

Membership Inference Attacks<sup>1</sup> are exploiting a model's tendency to memorize training data, making it possible to infer if a single input record was part of the training set by analyze the confidence of a prediction.



Shokri et al., "Membership Inference Attacks Against Machine Learning Models" (2017)



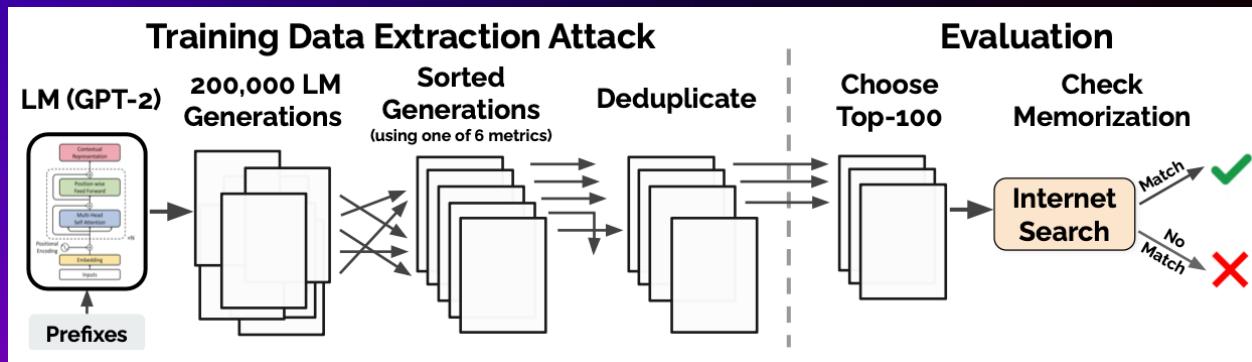
Hu et al., "Membership Inference Attacks on Machine Learning: A Survey" (2022)



# Membership Inference Attack on QPT-2

Membership Inference Attack have been shown to be successful again large language models (GPT-2)<sup>1</sup>

- Sentence completion (i.e., prompt engineering)
- Sentence Likelihood

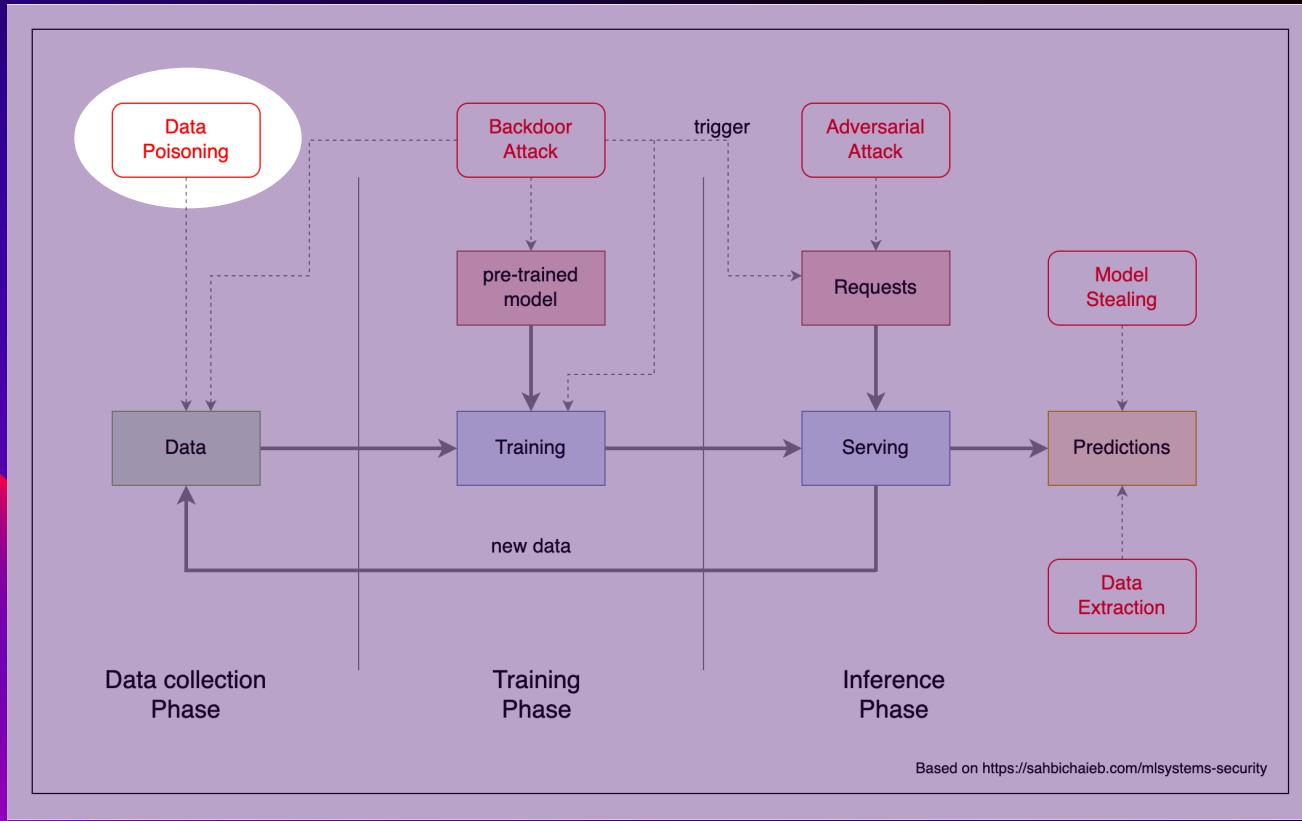


Carline et al. "Extracting Training Data from Large Language Models", (2020)



# Data Poisoning

# Attack Surface



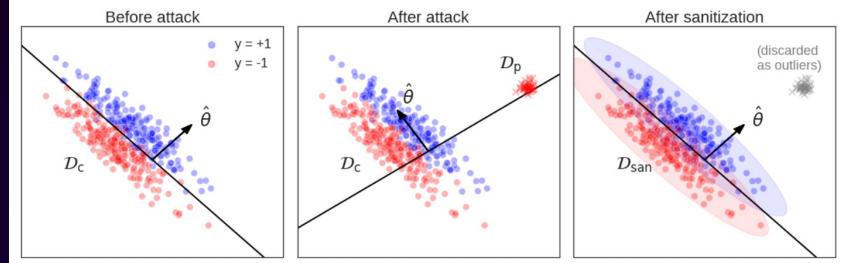
# Data Poisoning

Simulated Covariate Shift through specially crafted data points that causes updated model to benefit attacker.

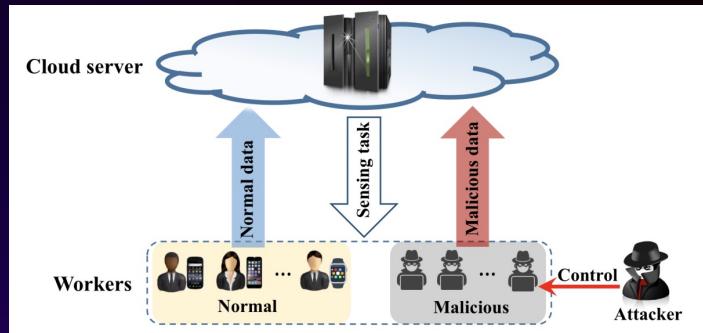
Can be applied to affect

- Availability
- Integrity (Backdoors)

Is effective to manipulate Recommender Systems<sup>1</sup>, online learning systems and federated learning systems<sup>2</sup>.



Wei Koh et al., "Stronger Data Poisoning Attacks Break Data Sanitization Defenses" (2021)



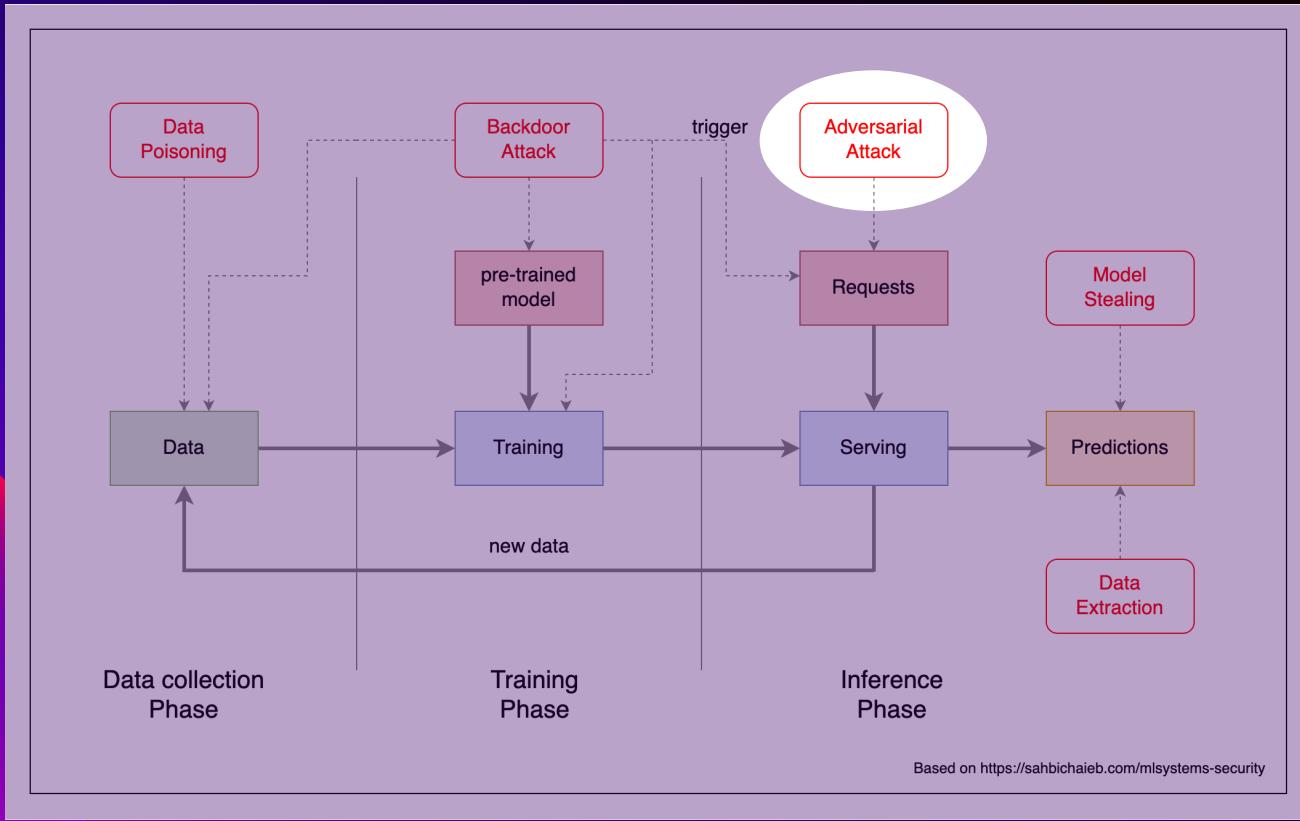
Miao et al., "Towards Data Poisoning Attacks in Crowd Sensing Systems" (2018)

1. Huang, et al. "Data Poisoning Attacks to Deep Learning Based Recommender Systems" , NDSS, (2021).
2. Tolpegin, et al. "Data Poisoning Attacks Against Federated Learning Systems", ESORICS, (2020)



# Adversarial Attacks

# Attack Surface



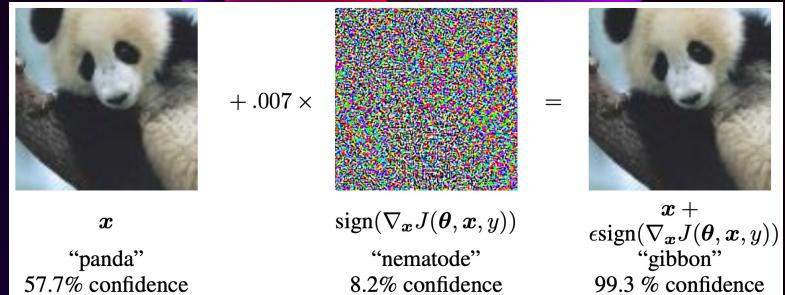
# Adversarial Attack

Usage of specially crafted input data to force a model output.

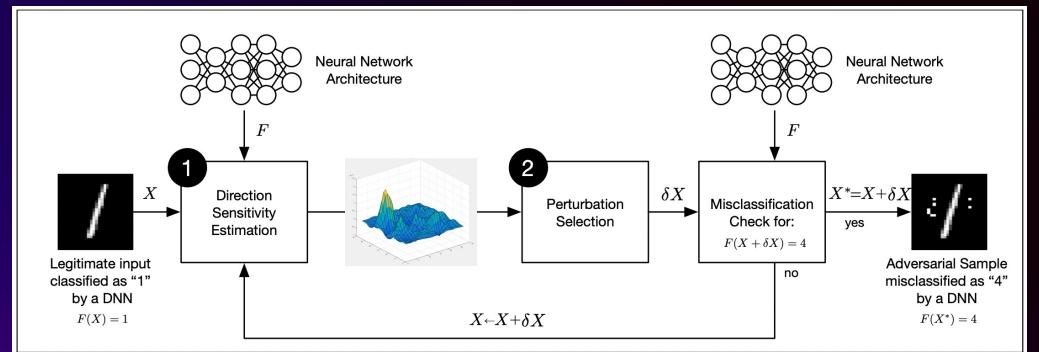
Can be targeted (specific output class) or Indiscriminate (anything but the original class).

Adversarial Examples can be transferred between models because of "non-robust features"<sup>1,2</sup> making them one of the most effective attack.

They expand beyond the image domain and into the physical world.



Goodfellow et al., “Explaining and harnessing adversarial examples” (2014)



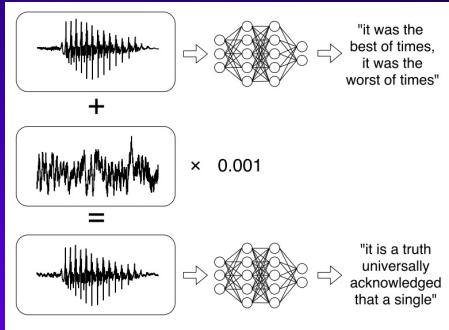
Papernot et al., “Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks” (2016)

1. Ilyas, et al. "Adversarial Examples are not Bugs, they are Features" , NeurIPS, (2019).

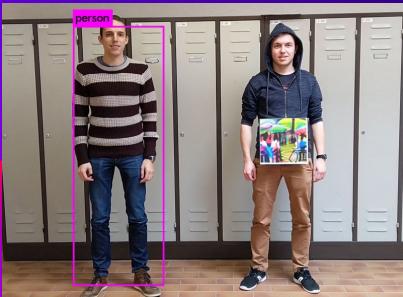
2. Waseda et al. “Closer look at the transferability of adversarial examples: how they fool different models differently”, (2021)



# Beyond the digital image domain



Carlini et al., "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text" (2018)



Thys et al., “Fooling automated surveillance cameras: adversarial patches to attack person detection” (2019)

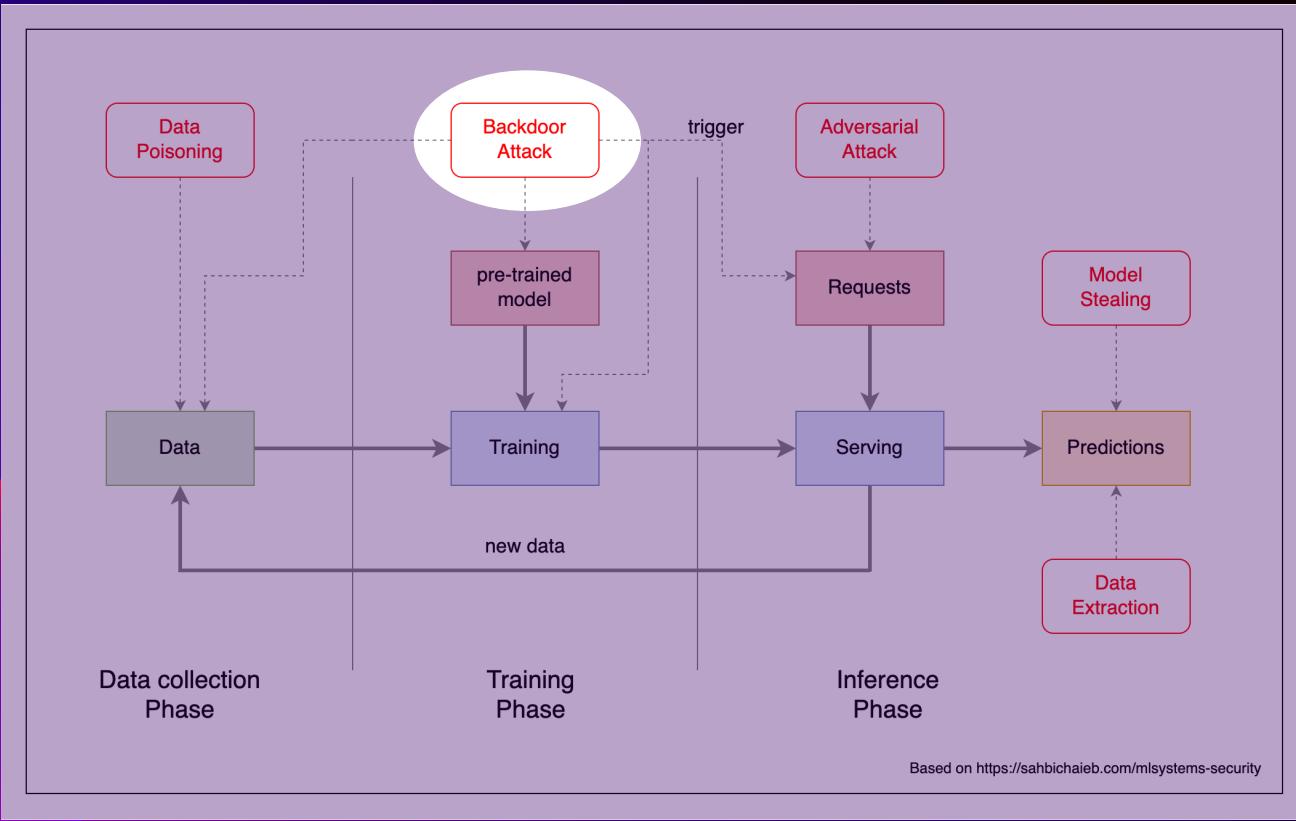


Sharif et al., "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition" (2016)

Wallace et al., “Imitation Attacks and Defenses for Black-box Machine Translation Systems” (2021)

# Backdoor Attacks

# Attack Surface



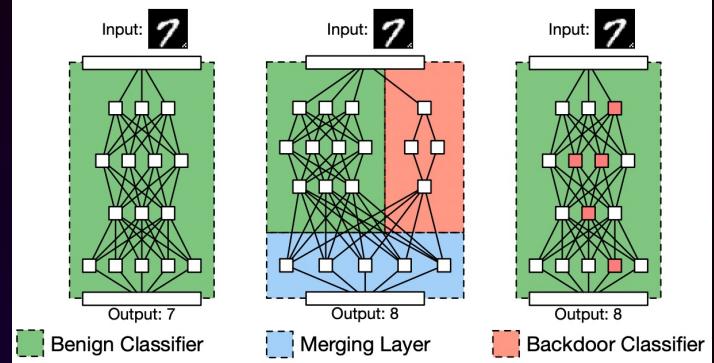
# Backdoor Attacks

Control of a model output through “trigger patterns” embedded in input data during inference.

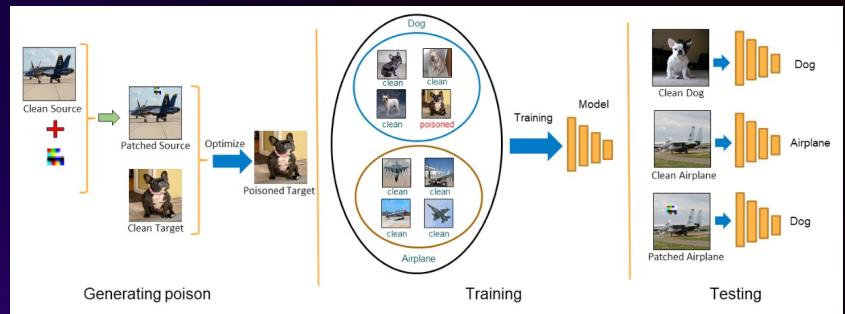
Can be “installed” by either

- Manipulation of training process
- Crafted training data (data poisoning)

A backdoor installed during training is undetectable and robust to retraining<sup>1</sup>.



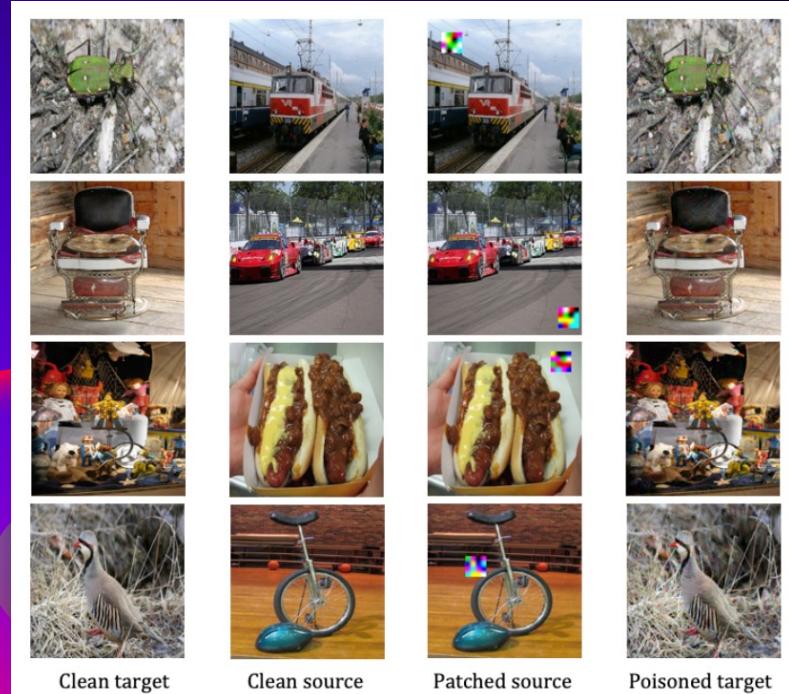
Gu et al., “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain ” (2019)



Saha et al., “Hidden Trigger Backdoor Attacks” (2019)

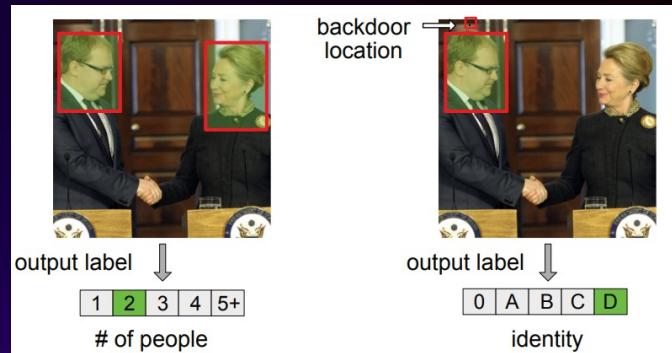


# Backdoors – Versatile attack mechanism



Saha et al., "Hidden Trigger Backdoor Attacks" (2019)

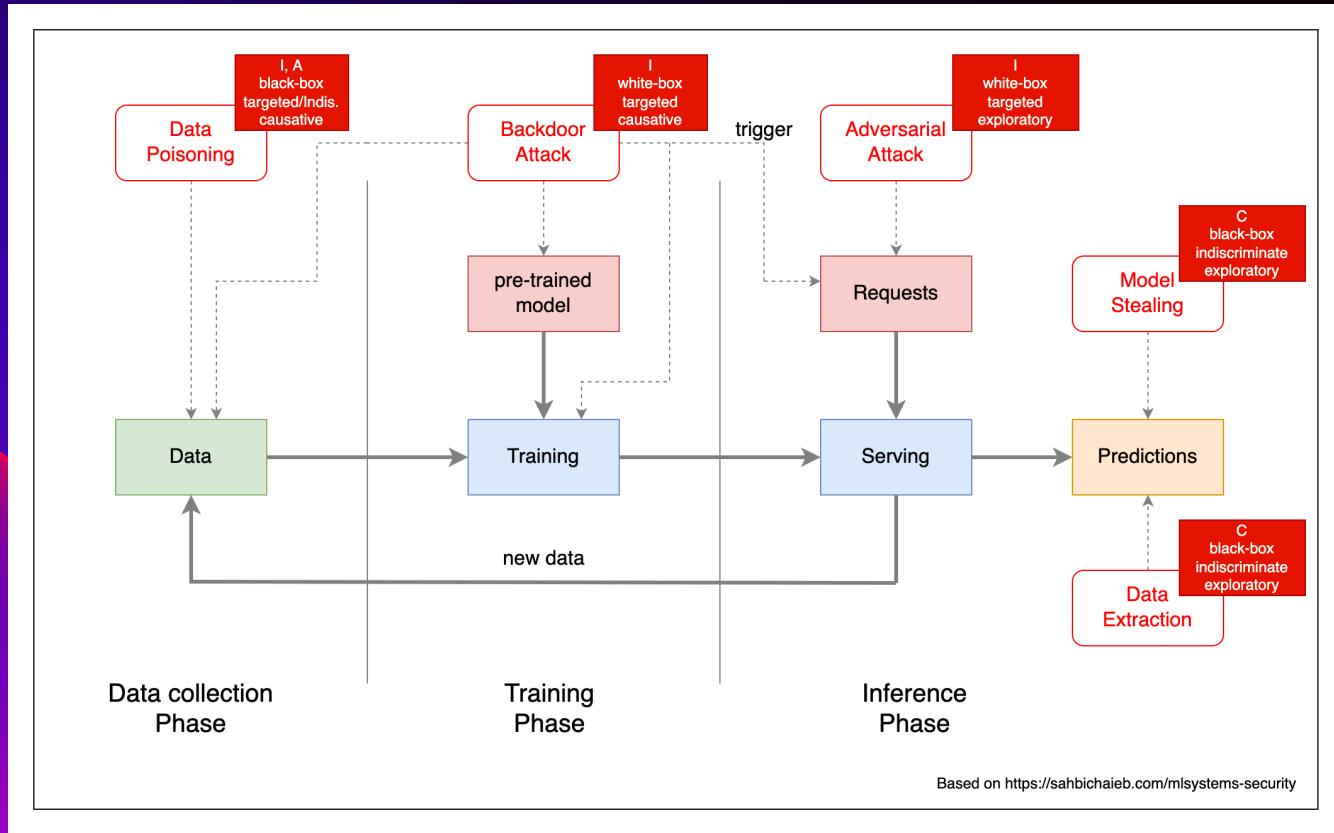
No backdoor:					
$\theta'(x):$	<hr/>				
Summation backdoor:					
$\theta'(x):$	<hr/>				
Multiplication backdoor:					
$\theta'(x):$	6	0	16	21	8



Bagdasaryan et al., "Blind Backdoors in Deep Learning Models" (2021)



# Attack Surface

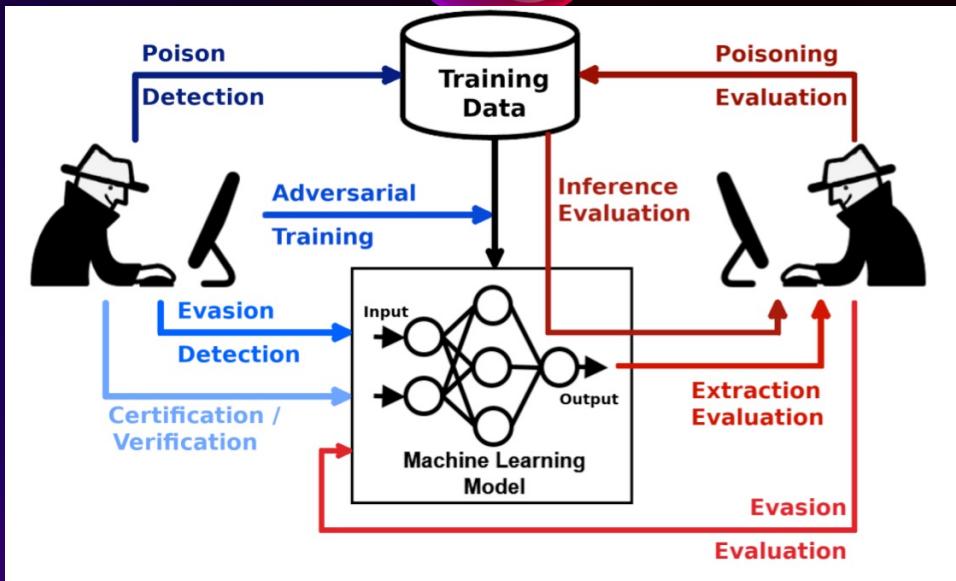
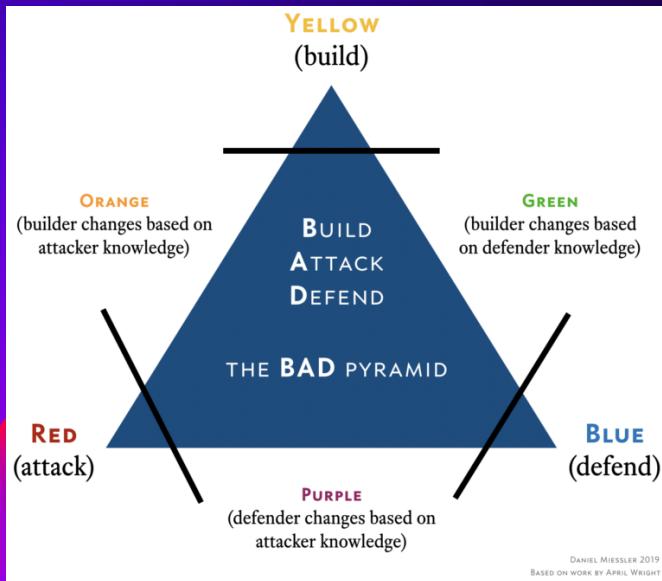


The background features abstract, organic shapes in shades of orange, yellow, purple, and red against a dark blue gradient.

**What to do  
about it?**

# What are the big companies doing?

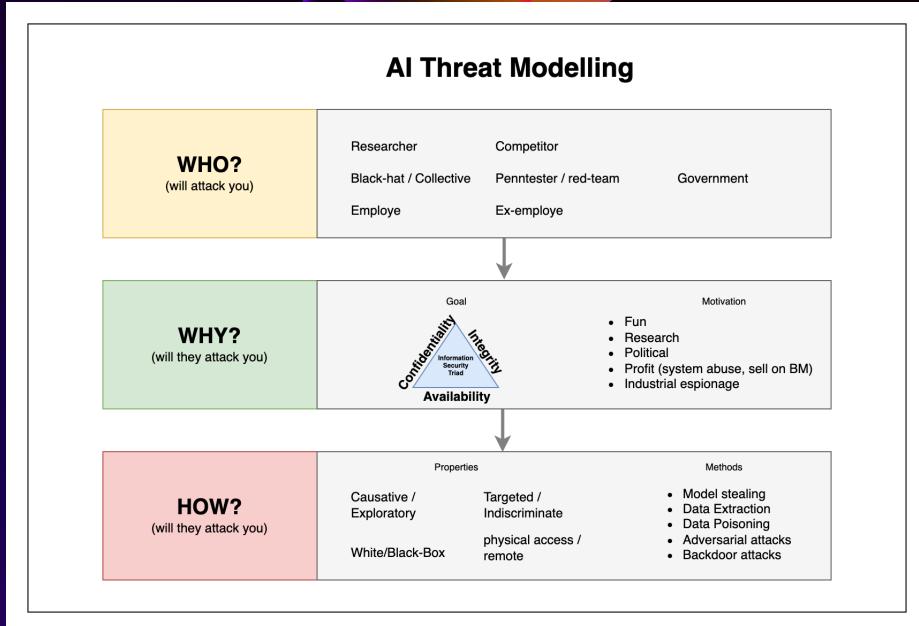
## AI Focused Cyber Security Groups



Nicolae et al., "Adversarial Robustness Toolbox" (2021)

# What can you do?

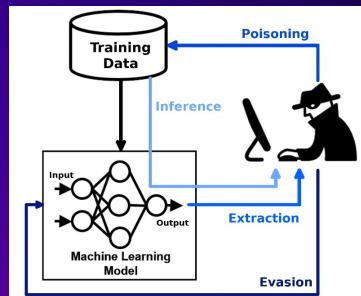
- Dr. Evil workshops: take the position of a fictitious attacker and identify attack scenarios
- AI Threat modelling: Identify who, why and how
- Evaluate attack scenarios: likelihood, damage/costs, countermeasures



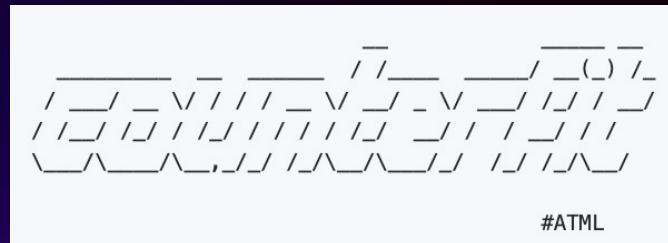
# Useful resources to get started



Adversarial Threat Landscape for Artificial-Intelligence Systems, <https://atlas.mitre.org/> (MITRE)



<https://github.com/Trusted-AI/adversarial-robustness-toolbox> (IBM)



Counterfit : <https://github.com/Azure/counterfit> (Microsoft)



# Thanks!

Does anyone have any questions?

[contact@manuelpasieka.com](mailto:contact@manuelpasieka.com)

Powered by







# Attack Surface

