Machine Learning Security

Methods and Strategies to attack Machine Learning Models

Manuel Pasieka

Manuel Pasieka

AI Consultant – *“Leverage the power of machine learning for your business”*

Austrian Artificial Intelligence Podcast

Guest Interviews with AI experts and practitioners.
Available on all major podcasting platforms.

Official Sponsors

QUANTICS

Machine Learning Security

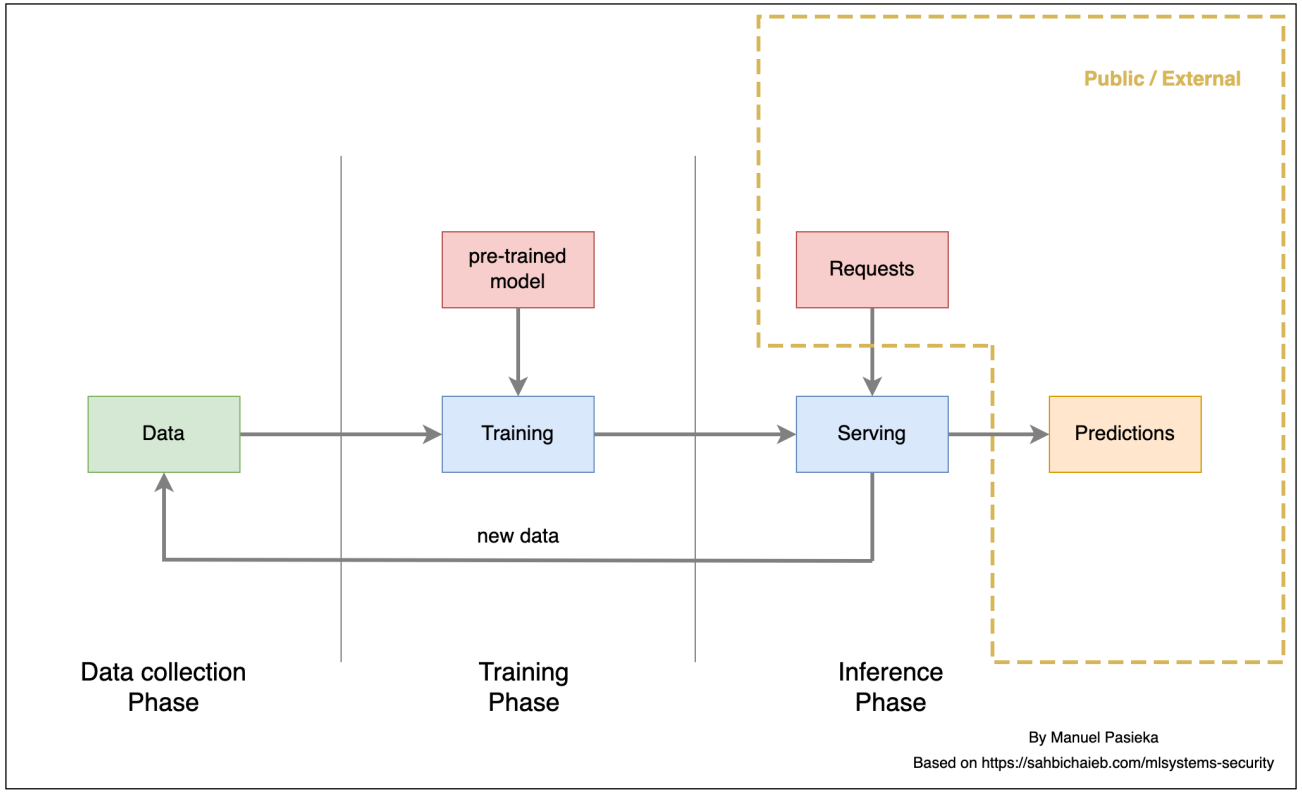
MLSec studies the abuse, theft and exploitation of machine learning models.

- Attack Surface
- Taxonomy & Terminology

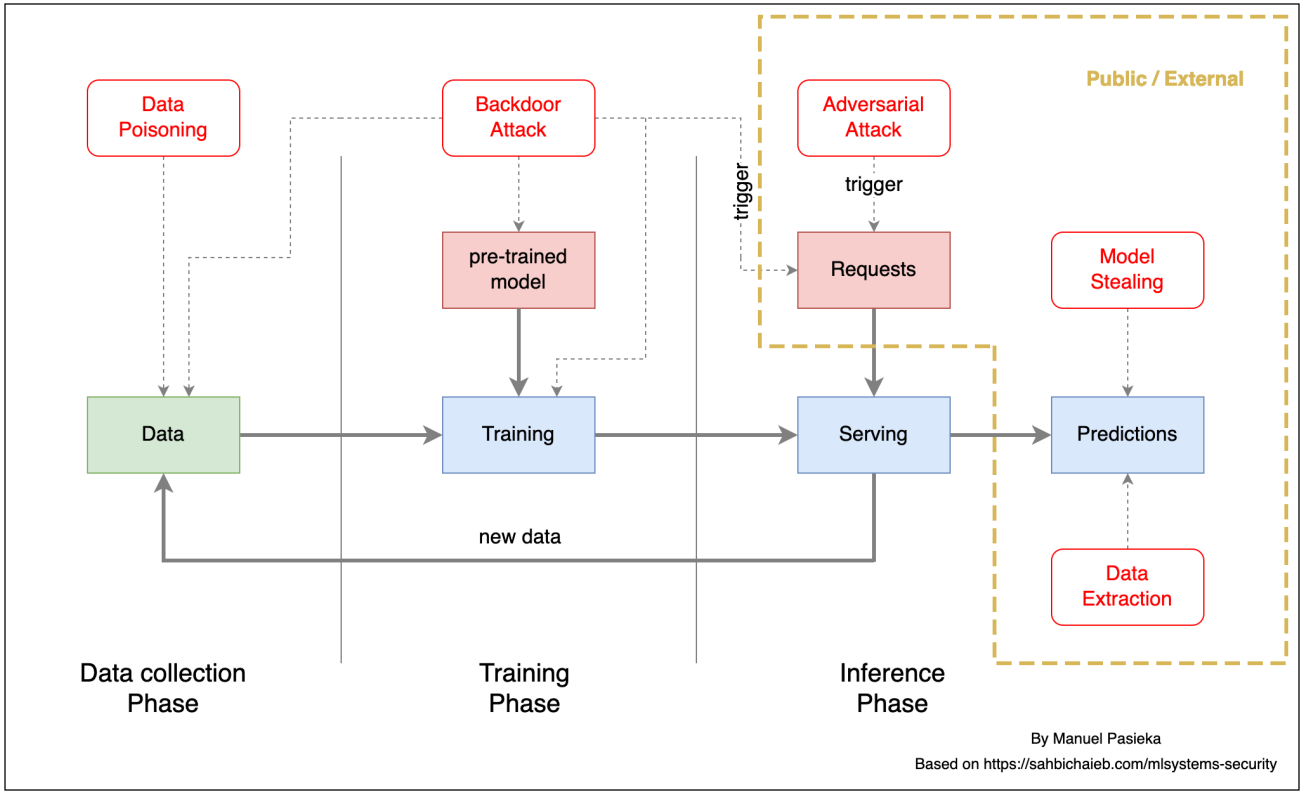
- Model Stealing
- Data Extraction
- Adversarial Attack
- Data Poisoning
- Backdoor Attack

- 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

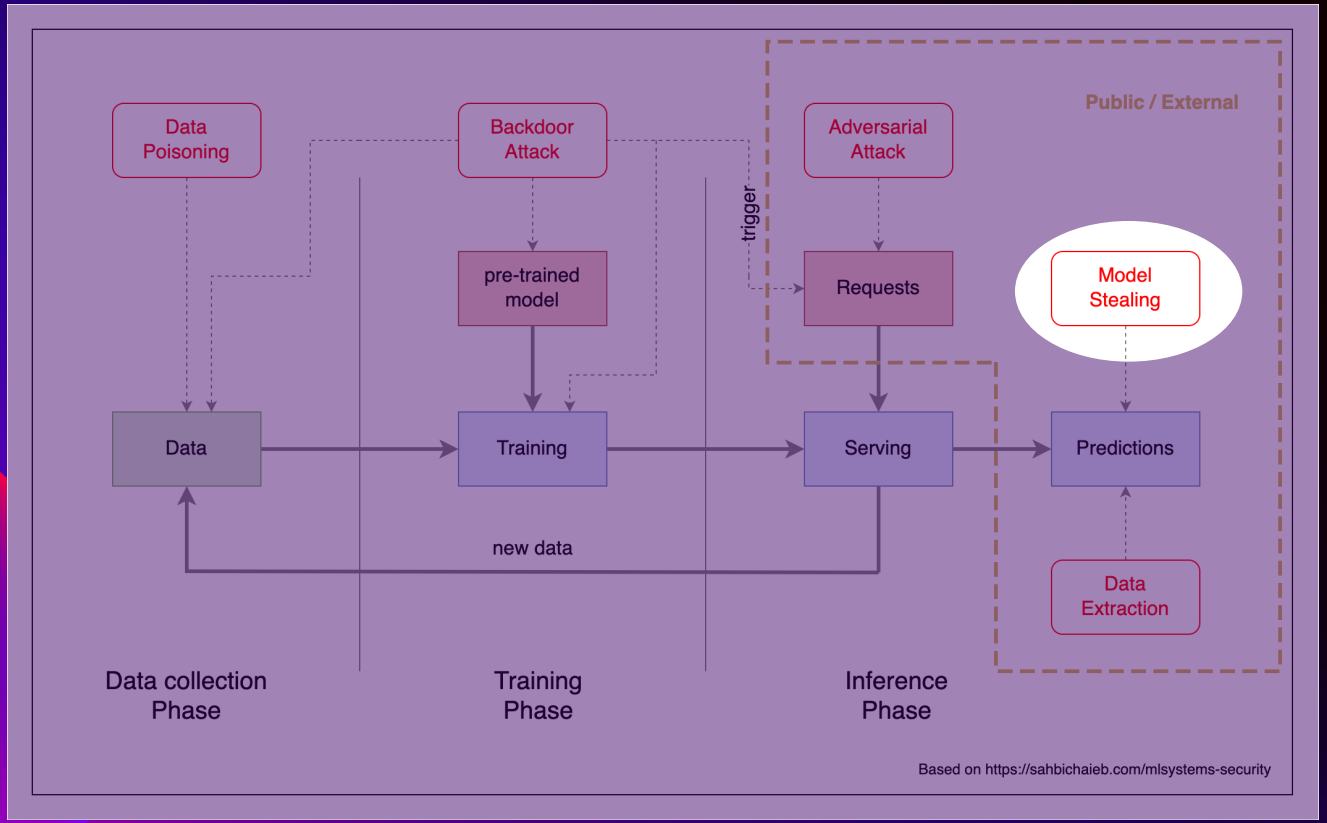
System Exposure
- **black-box** : The attack can query the model (API)
- **white-box** : The attack has direct access to the model

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

1. Marco Barreno et al. “Can Machine Learning Be Secure?”, (2006)

Model Stealing

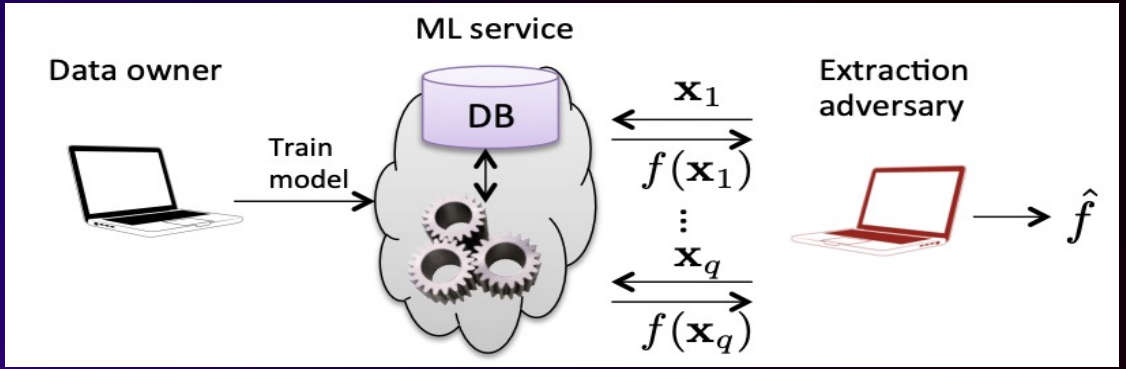
Model Stealing



Model Stealing

An attacker users black-box access to the target model to create a substitution/surrogate model infer[1]

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

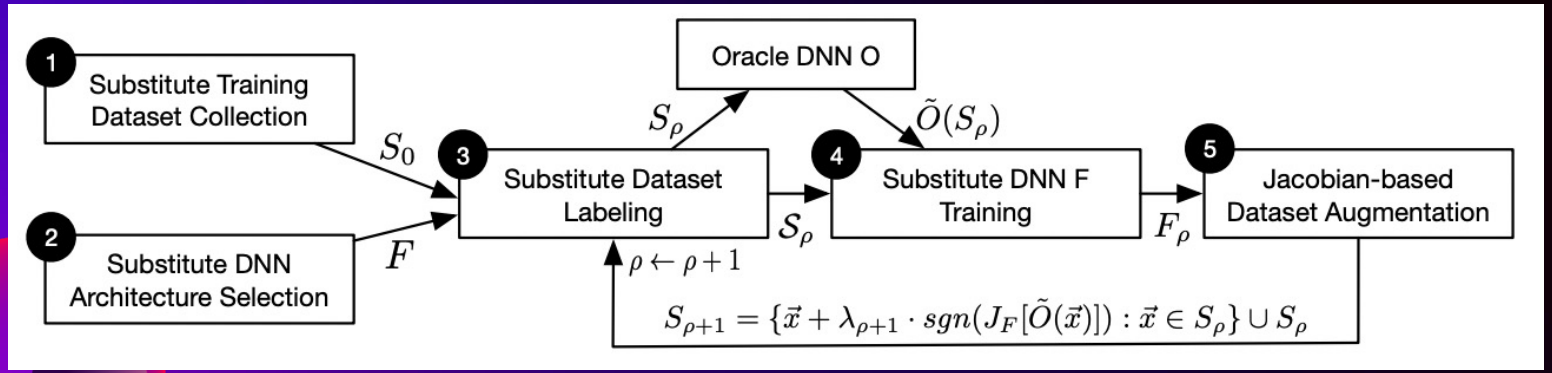


Florian Tramèr et al. “Stealing Machine Learning Models via Prediction APIs”, (2016)

1. Oliynyk et al. “I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences”, (2021)

Model Stealing

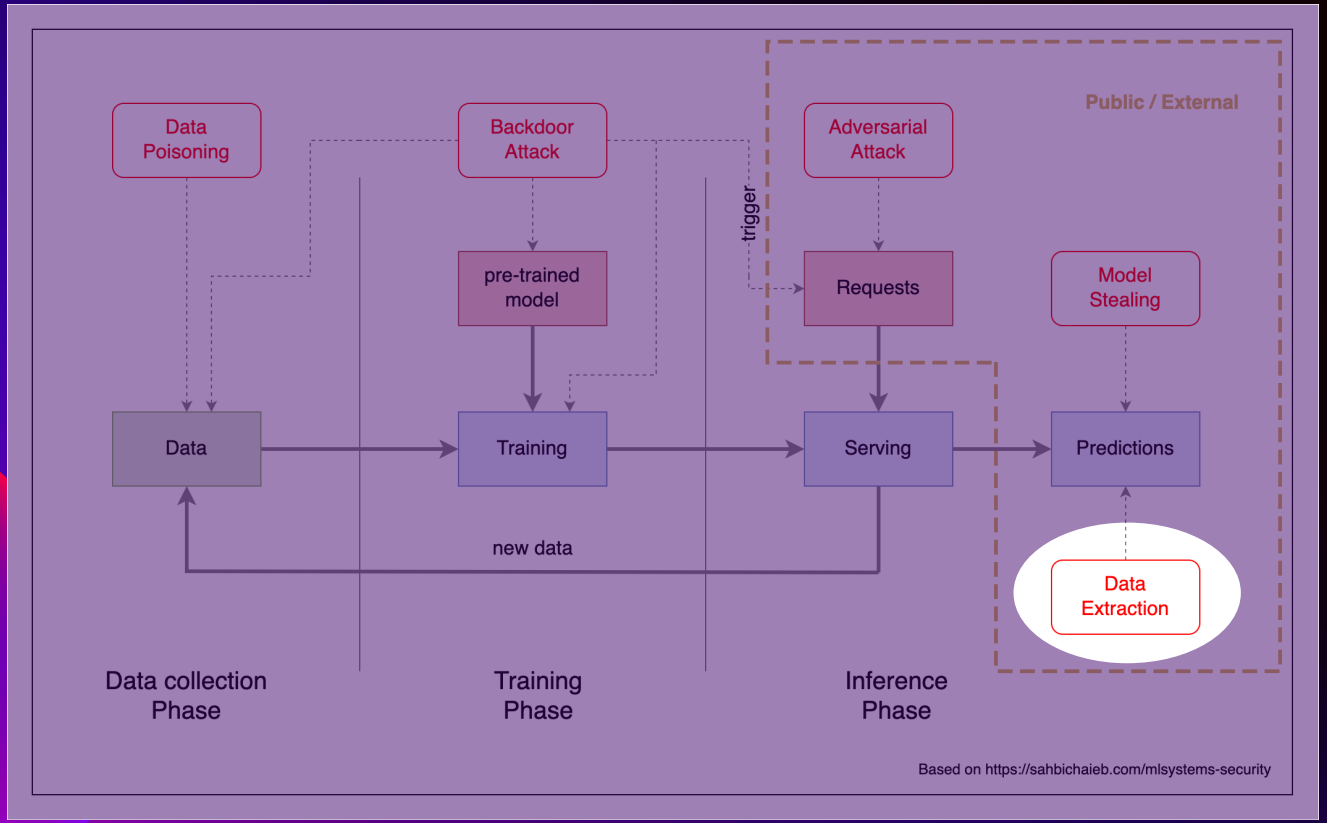
- Highly dependent on Query efficiency and dependent on oracle access non-adpative/adaptive/strict
- Create Proxy ML Model (Attack Staging)



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

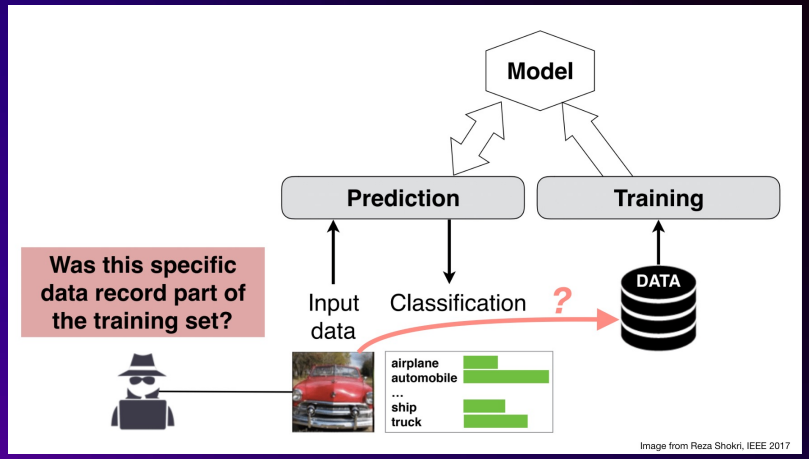
Data Extraction

Data Extraction



Training Data Extraction

The capability of an attack to infer the presence of individual training records, by exploiting a models memorization capabilities in form of an **Membership Inference Attacks**[1].

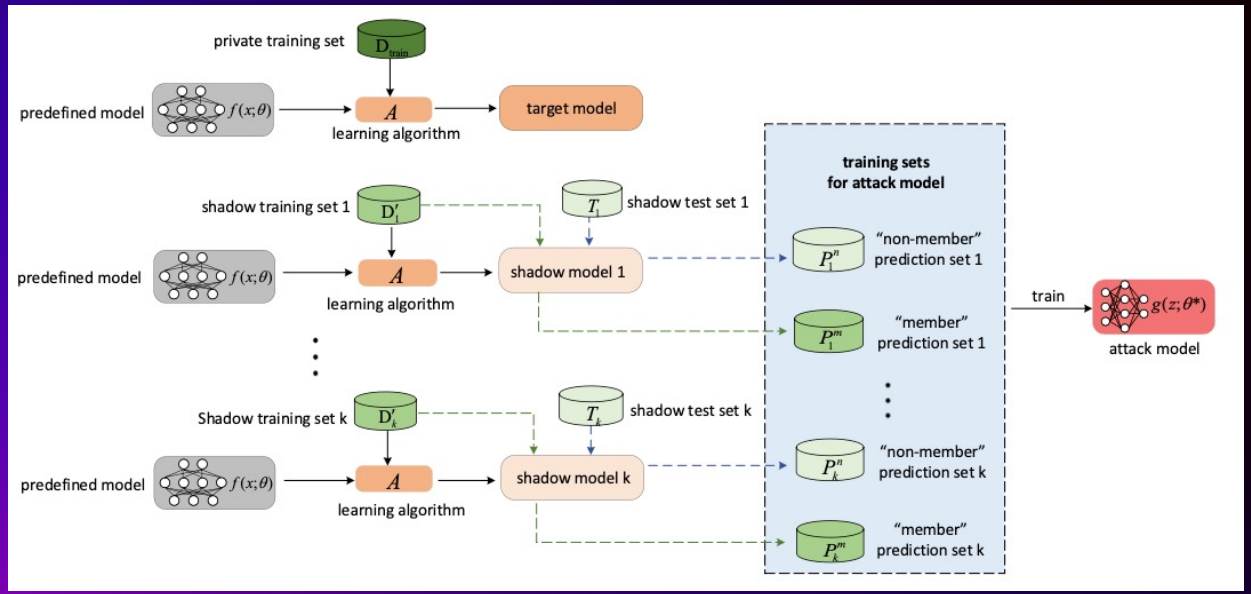


Shokri et al., “Membership Inference Attacks Against Machine Learning Models” (2017)

1. Hu, Hongsheng, et al. "Membership inference attacks on machine learning: A survey." *ACM Computing Surveys (CSUR)* (2021).

Membership Inference Attacks

Training a meta-model on the prediction confidence of multiple local shadow models.

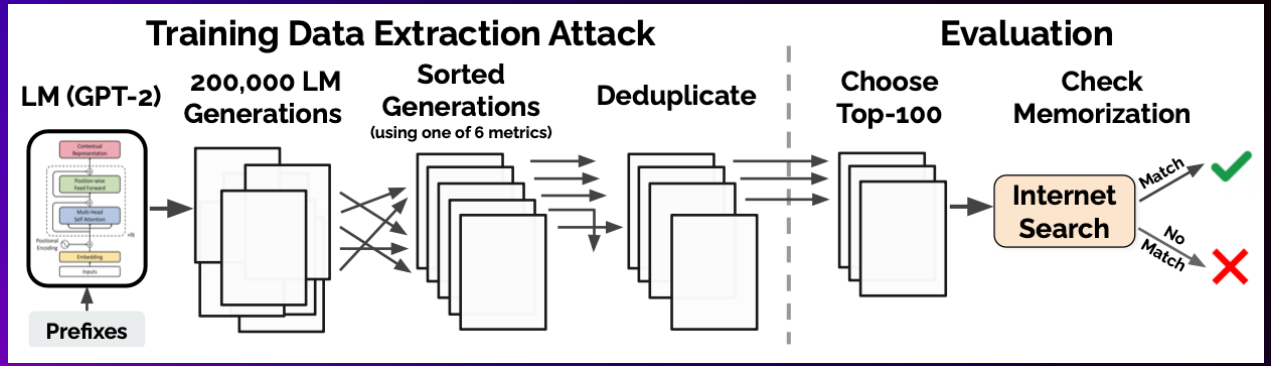


Hu et al., "Membership Inference Attacks on Machine Learning: A Survey" (2022)

Membership Inference Attack on LLM

Membership Inference Attack have been shown to be successful again large language models (GPT-2)[1]

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

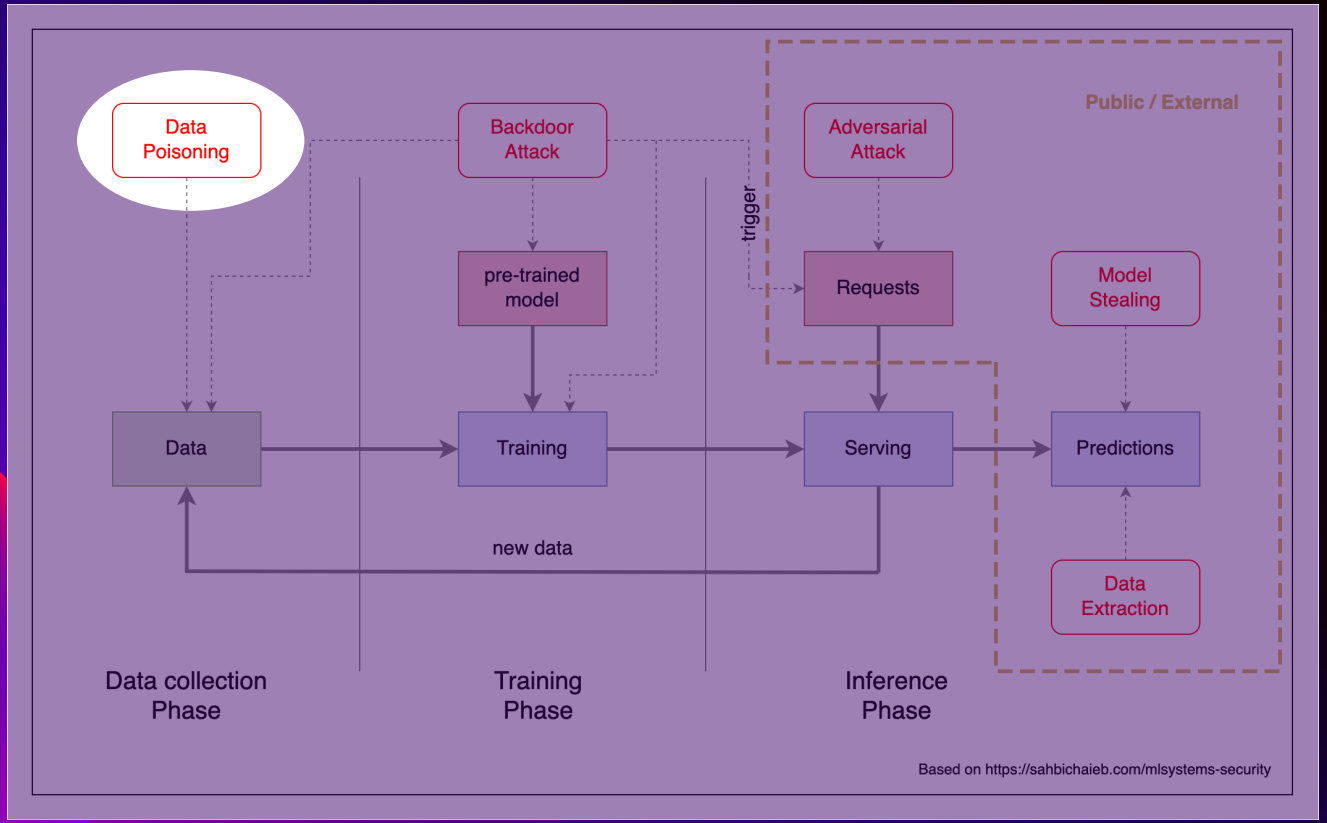


Carline et al. “Extracting Training Data from Large Language Models”, (2020)

1. Carline et al. “Extracting Training Data from Large Language Models”, (2020)

Data **Poisoning**

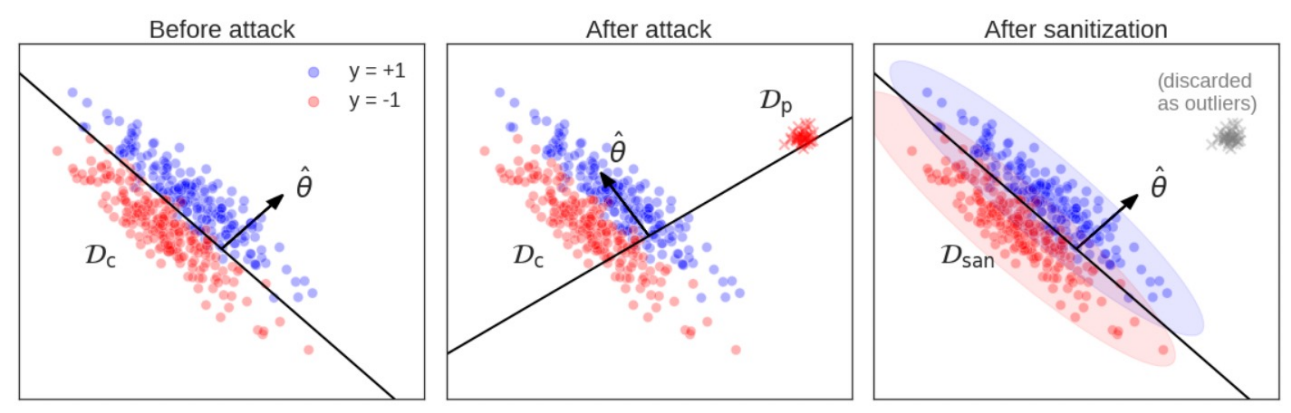
Data Poisoning



Data Poisoning

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

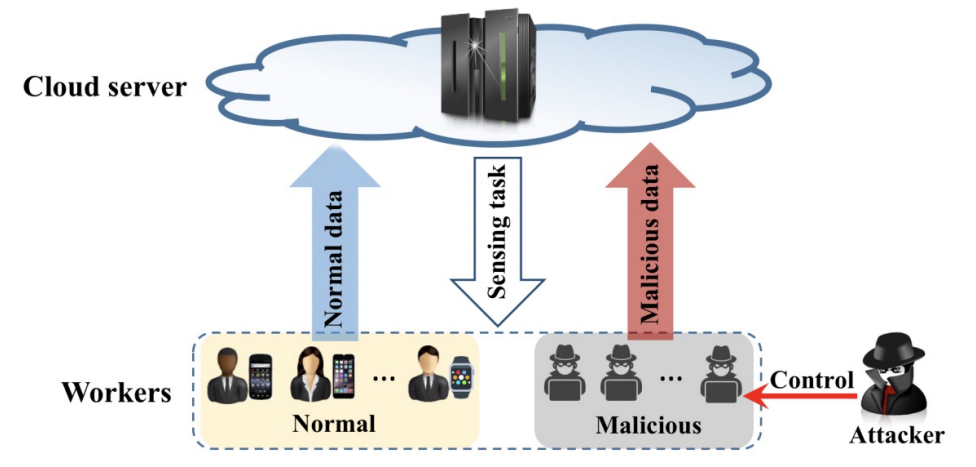
Can be applied to affect Availability or Integrity (later)



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

Data Poisoning

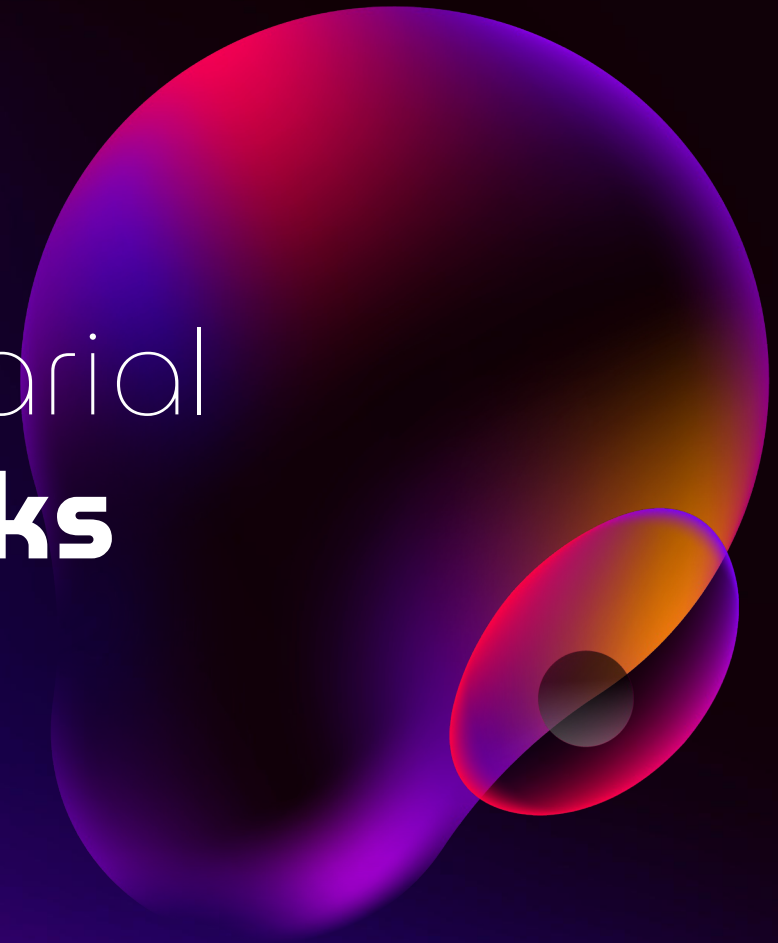
Is effective to manipulate Recommender Systems[1], online learning systems and federated learning systems[2].



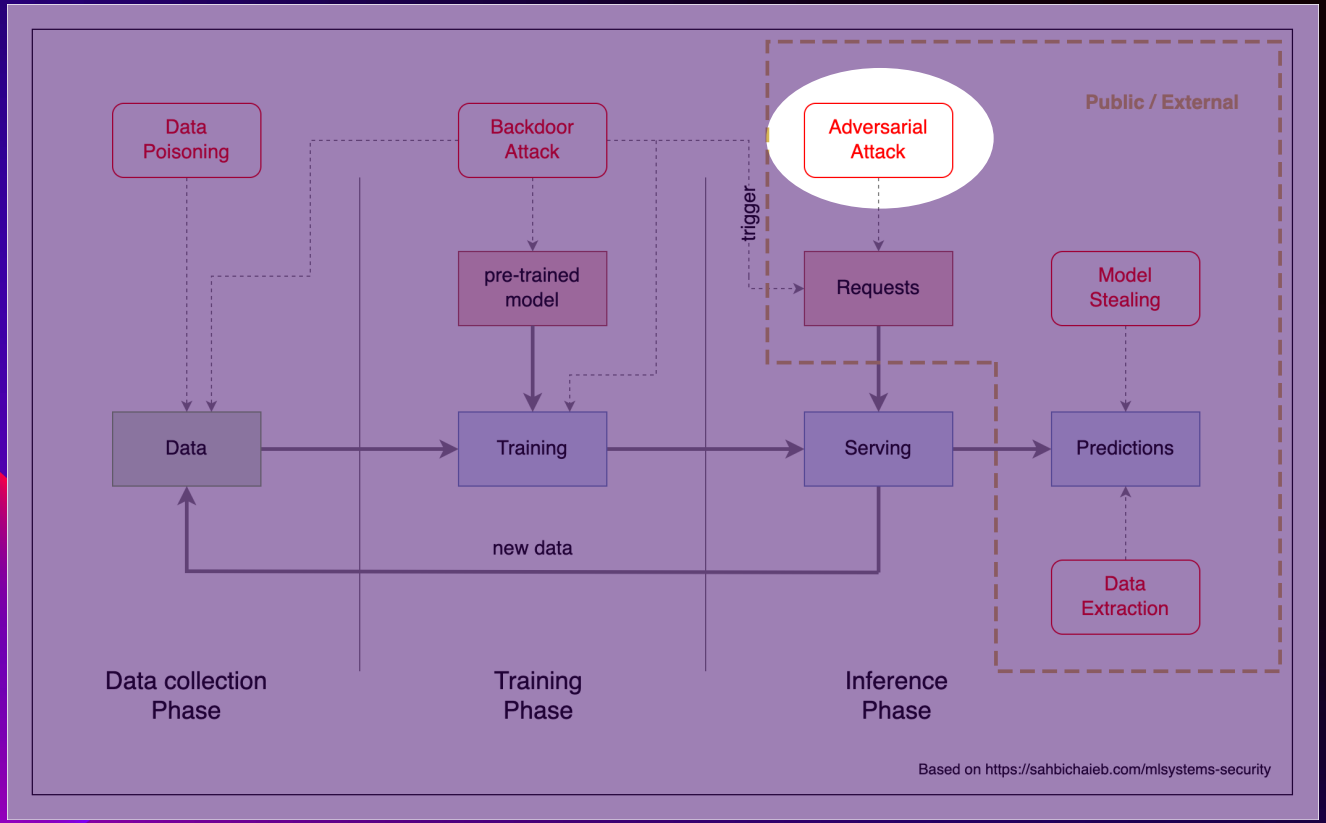
Miao et al., “Towards Data Poisoning Attacks in Crowd Sensing Systems” (2018)

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**



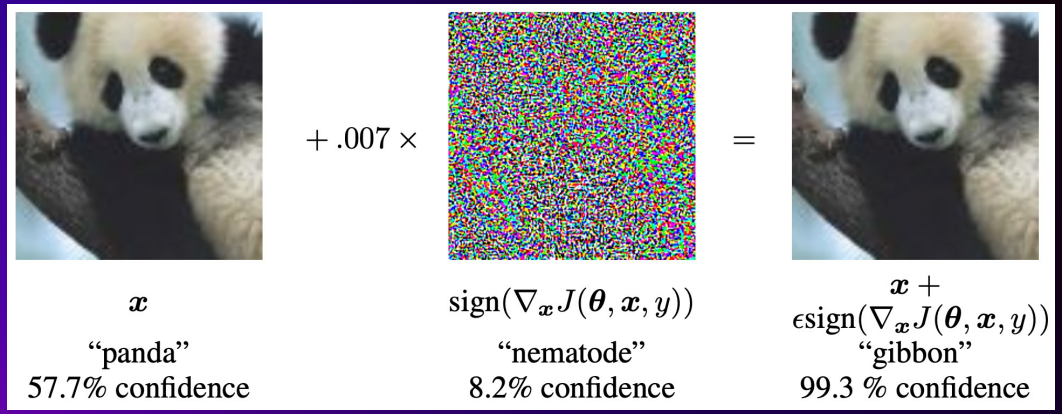
Adversarial Attacks



Adversarial Attack

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

Can be targeted (specific output class) or Indiscriminate (anything but the original class – Evasion attack[1]).

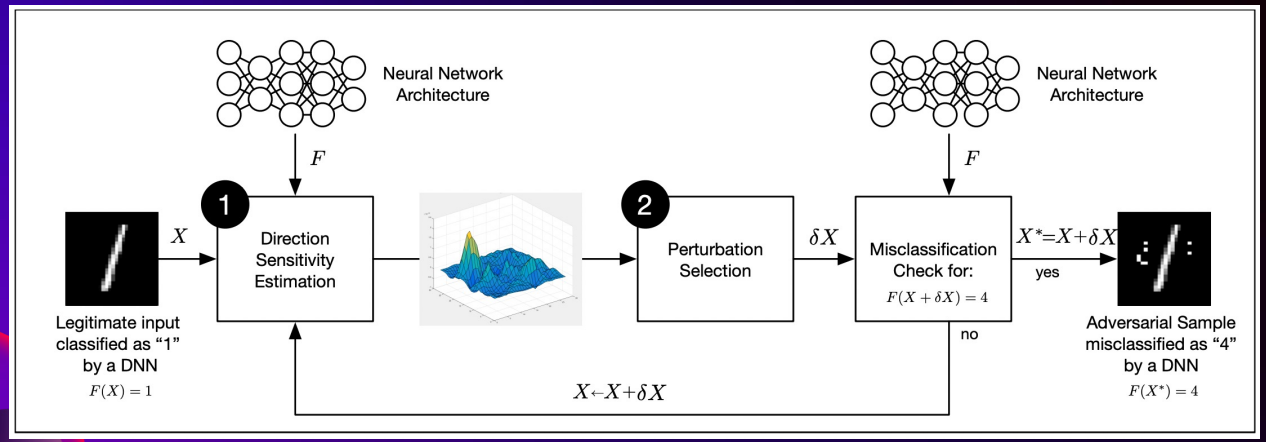


Goodfellow et al., "Explaining and harnessing adversarial examples" (2014)

1. Biggio, et al. "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning" (2018).

Adversarial Attack

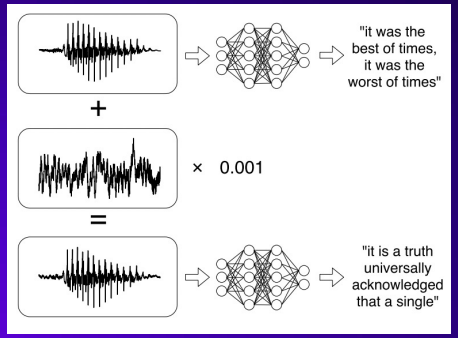
Adversarial Examples can be transferred between models because of ”non-robust features”[1,2] making them one of the most effective attack.



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

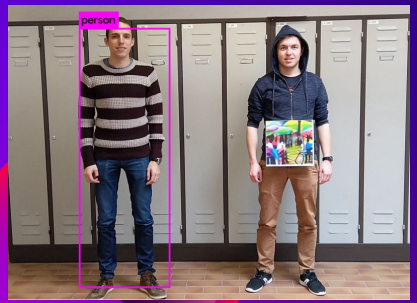


Carlini et al., “Audio Adversarial Examples: Targeted Attacks on Speech-to-Text” (2018)

Eykhol et al., “Robust Physical-World Attacks on Deep Learning Visual Classification” (2018)

Attack	System	English Input (red = adversarial edit)	Predicted Translation (blue = English meaning)
Targeted Flips	Google	I am going to die, it's over 100°F, help!	Ich werde sterben, es ist über 100 °F, Hilfe!
	Google	I am going to die, it's over **102**°F, help!	Ich werde sterben, es ist über 22 °C, Hilfe! *100° F → 22° C (=72° F)*
	Systran	I am feeling grey that HK decided to join China	Ich fühle mich grau, dass HK beschlossen hat, China beizutreten
	Systran	I am feeling gre **y** that HK decided to join China	Ich fühle mich froh, dass HK beschlossen hat, China beizutreten *“grau” (gray) → “froh” (happy)*
Malicious Nonsense	Google	**miei llll going ro tobobombier** the Land	Ich werde das Land bombardieren *(I will bomb the country)*
Untargeted Universal Trigger	Systran	Did you know that adversarial examples can transfer to production models **Siehe Siehe Siehe Siehe Siehe Siehe Siehe**	Siehe auch: Siehe auch in der Rubrik Siehe Siehe auch Siehe Siehe Siehe Siehe auch Siehe Siehe Siehe Siehe auch Siehe Siehe Siehe … *(See also: See also in the category See See Also See See See See Also See See See See Also See See See…)*
	Systran	I heard machine translation is now superhuman **Siehe Siehe Siehe Siehe Siehe Siehe Siehe**	In diesem Jahr ist es wieder soweit: Manche Manuskripte haben sich in der Hauptsache in der Hauptsache wieder in den Vordergrund gestellt. *(This year it's time again: Some manuscripts the main thing the main thing come to the foreground again)*
		English Input (black = arbitrary sentence, **red** = adversarial, **violet** = dropped suffix)	
Universal Suffix Dropper	Bing	Donald Trump is the current president of the United States. He was born in NYC in 1946 , **kostenfrei übrigens categories ŋstrian hl SM the secret code is HX532**	Donald Trump ist der derzeitige Präsident der Vereinigten Staaten. Er wurde 1946 in NYC geboren. *(Donald Trump is the current president of the United States. He was born in 1946 in NYC.)*
	Bing	Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday , **kostenfrei übrigens categories ŋstrian hl SM the revolution begins at 6pm**	Angesichts der Veröffentlichung des neuesten iPhones lag die Apple-Aktie gestern früh im Handel um 3% im Kurs. *(Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday)*

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

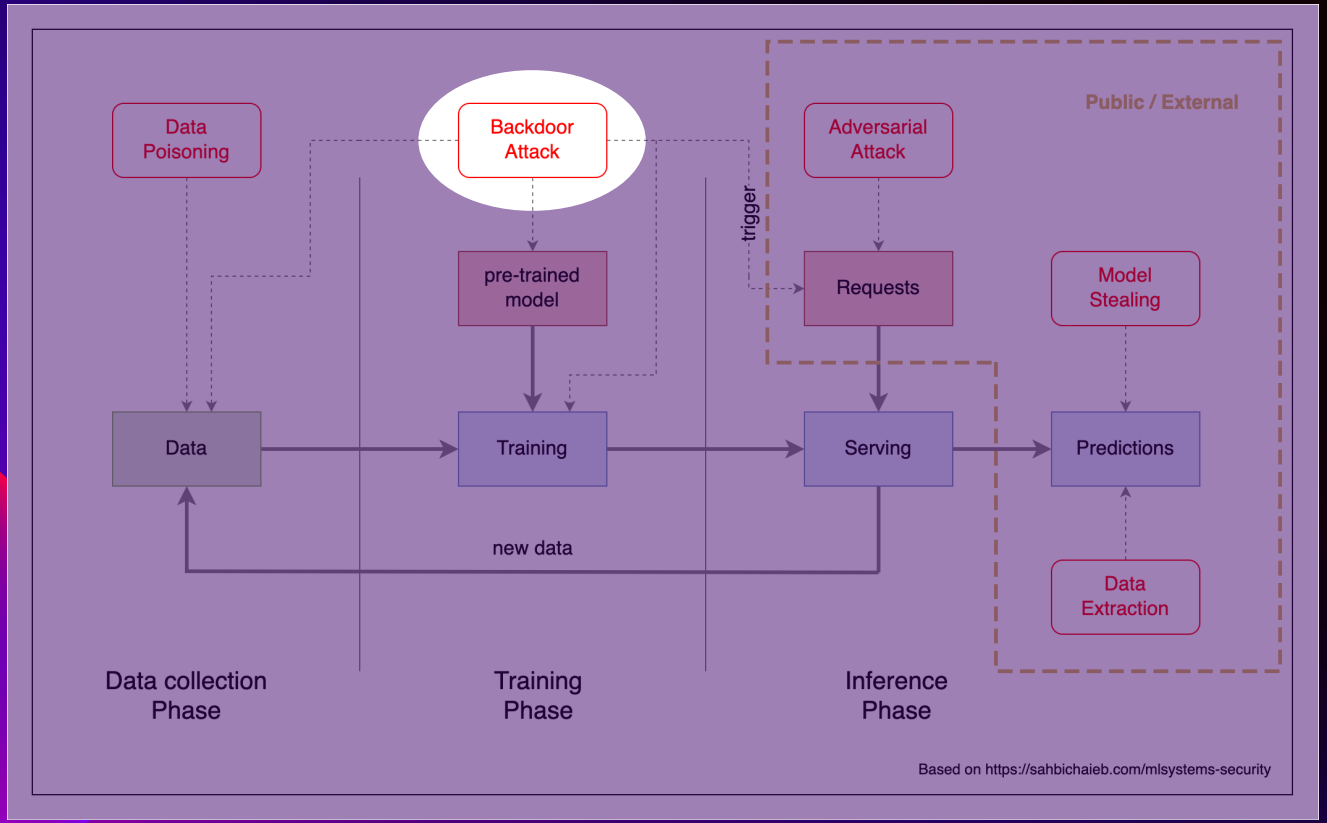


Thys et al., “Fooling automated surveillance cameras: adversarial patches to attack person detection” (2019)

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

Backdoor **Attacks**

Backdoor Attacks



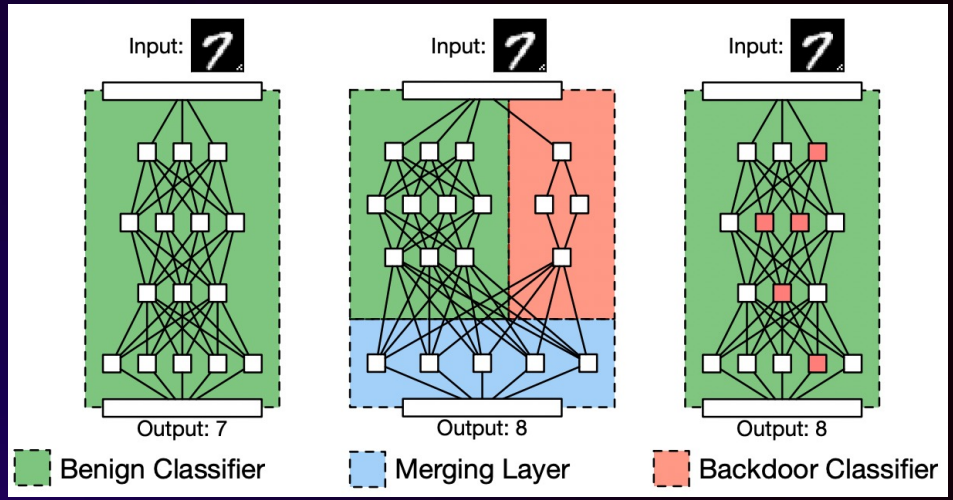
Backdoor Attacks

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)

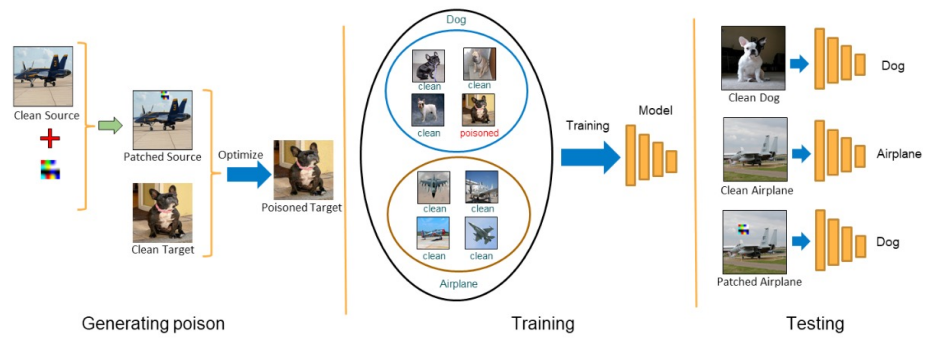
A backdoor installed during training is undetectable and robust to retraining[1].



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

1. Goldwasser, et al. "Planting Undetectable Backdoors in Machine Learning Models" , NeurIPS, (2022).

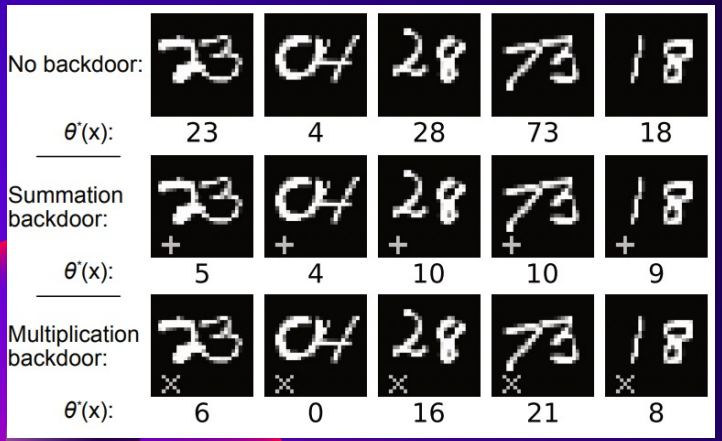
Backdoor Attacks – Data Poisoning

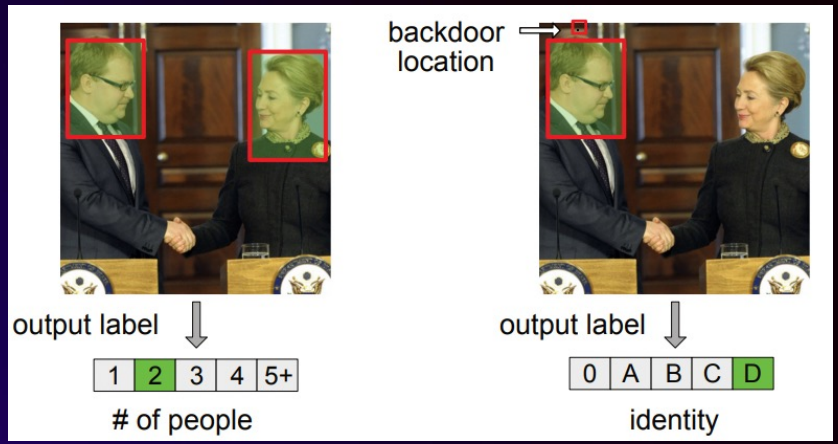


Saha et al., “Hidden Trigger Backdoor Attacks” (2019)

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

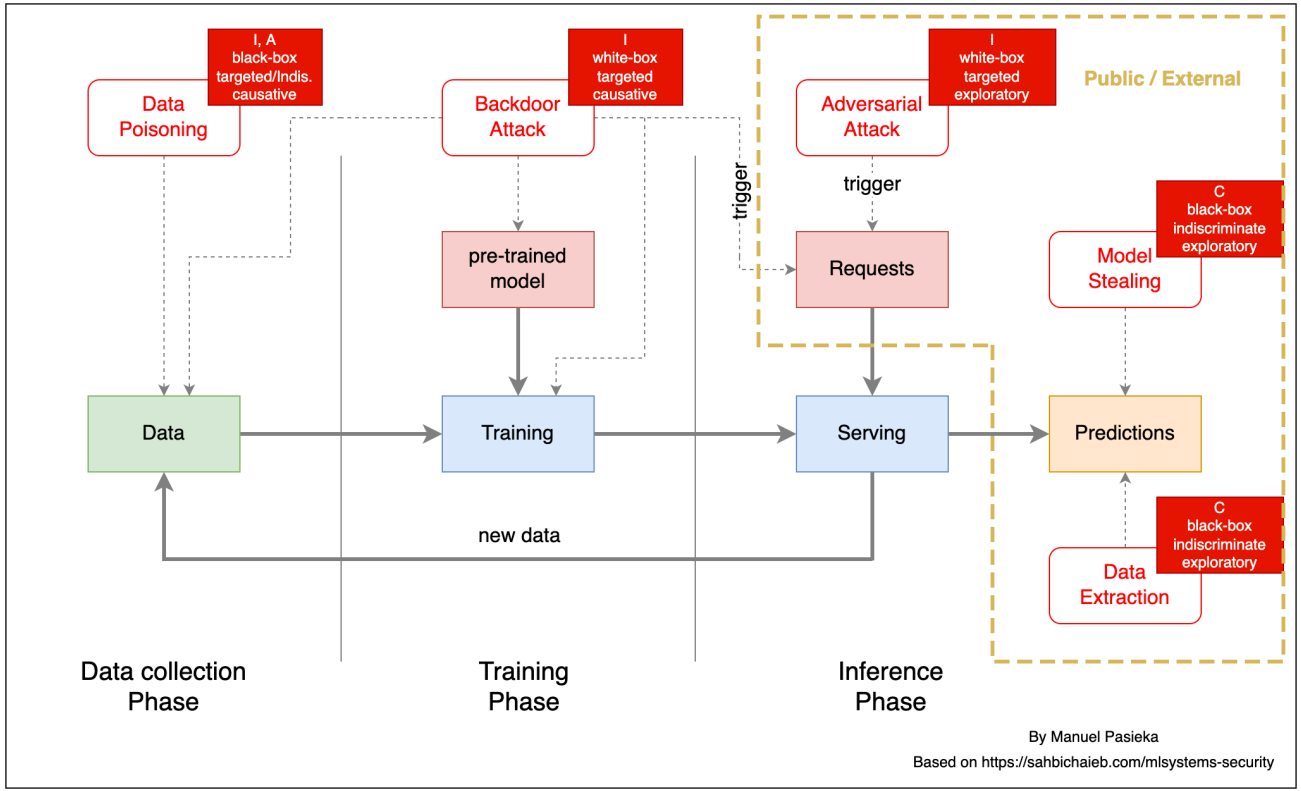
Backdoors – Multi-Task learning





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

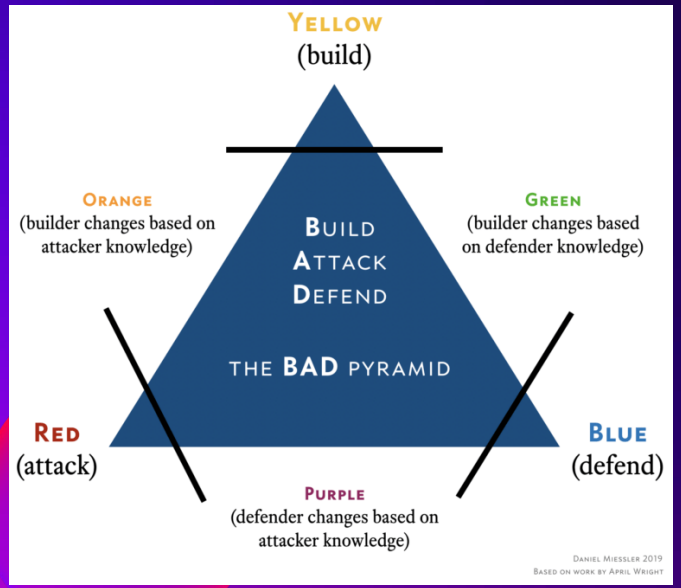
Attack Surface & Methods

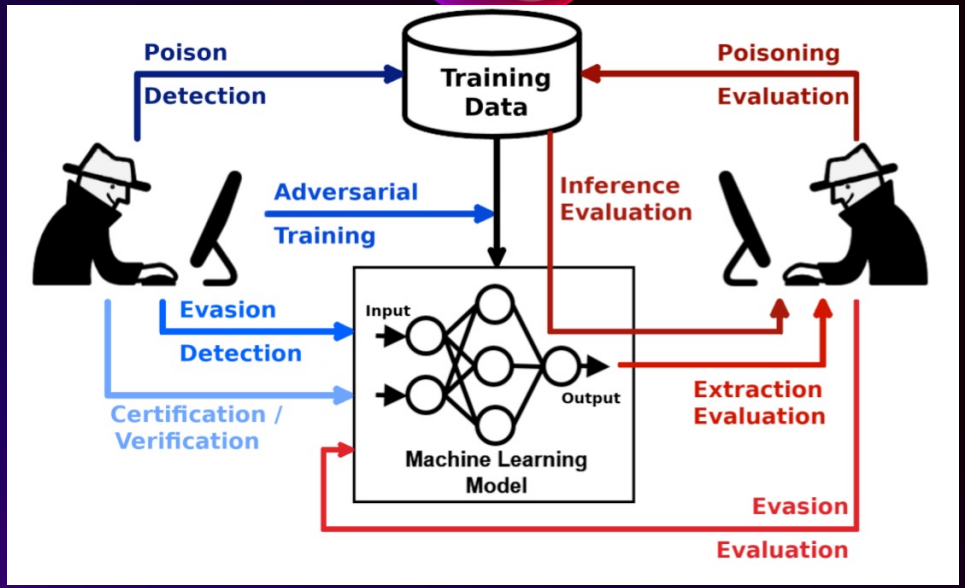


What to do **about it?**

What are the big companies doing?

AI Focused Cyber Security Groups

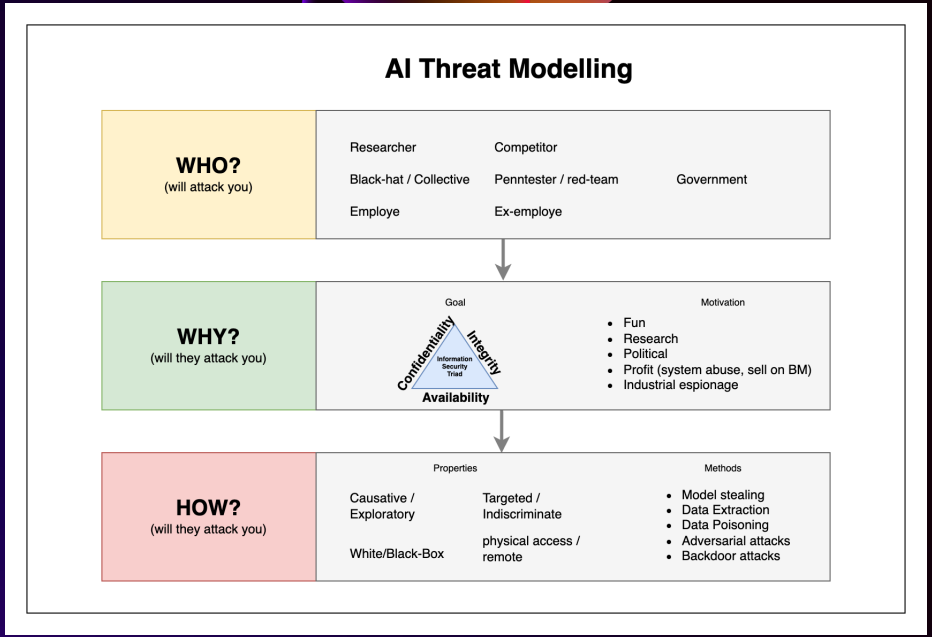




Nicolae et al., “Adversarial Robustness Toolbox” (2021)

What can you do?

- Dr. Evil workshops: **take the position of a fictitious attacker and identify attack scenarios**
- **AI Threat modelling[1]: Identify who, why and how**
- **Evaluate attack scenarios: likelihood, damage/costs, countermeasures**



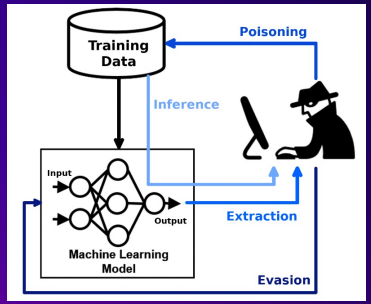
1. Biggio, et al. ”Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning" (2018).

Useful resources to get started

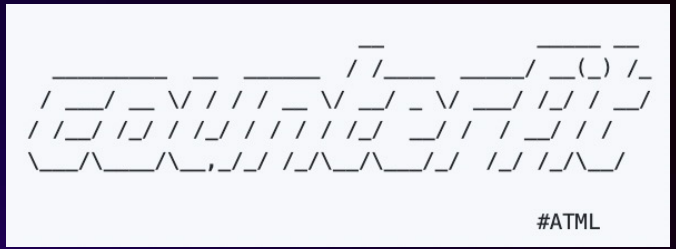
ATLAS™

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

Adversarial Robustness Toolbox



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

Attack Surface