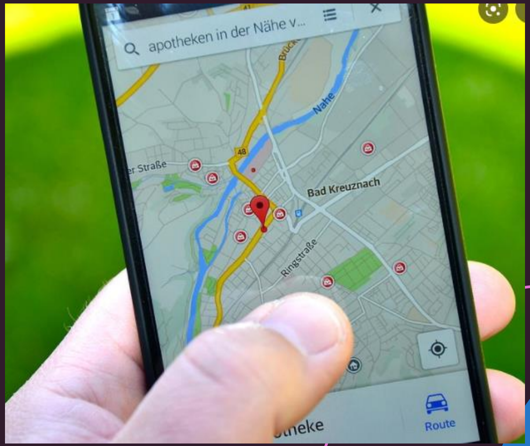
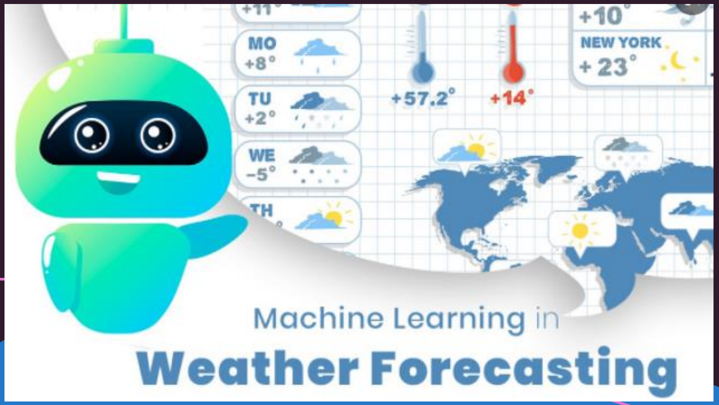
Introduction to Adversarial Attack on Machine Learning Model

Presenter: Jin Huiwen

Date: Feb 3 2022



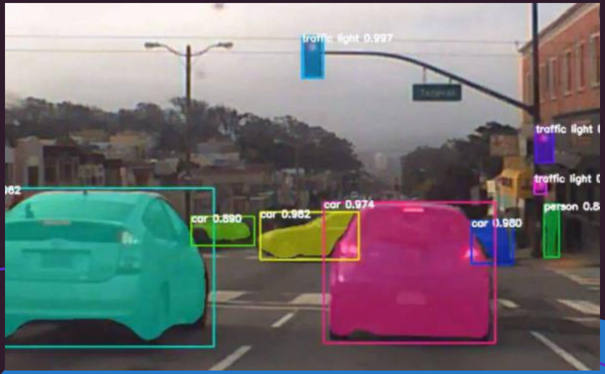


Outline

- Background
- Development history
- Adversarial Machine Learning
- Typical AI Security Attacks
- Adversarial robustness toolbox

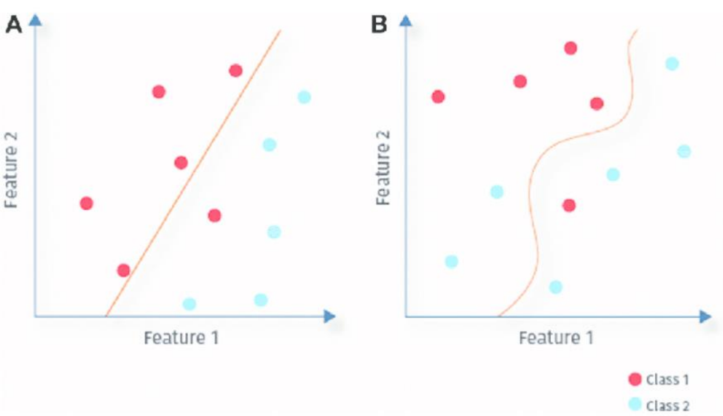
Background

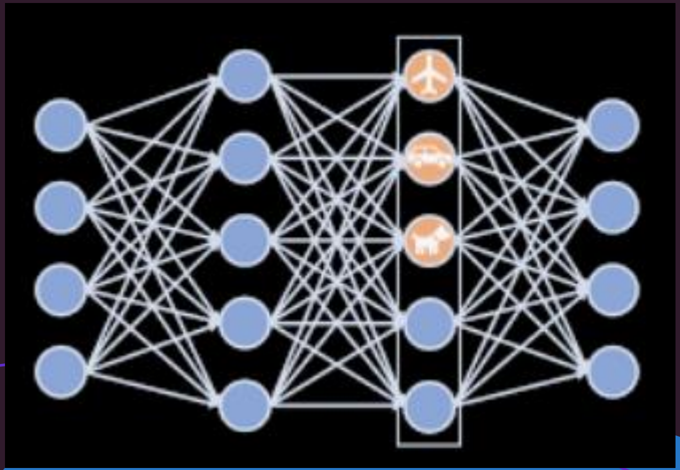
- Machine Learning (ML) methods and Artificial Intelligence (AI)
- Extensive application in security system
  - Autonomous car with object detection task
  - Surveillance system with face recognition
  - Door access system with voice recognition
- Security issue of ML and AI



Development History

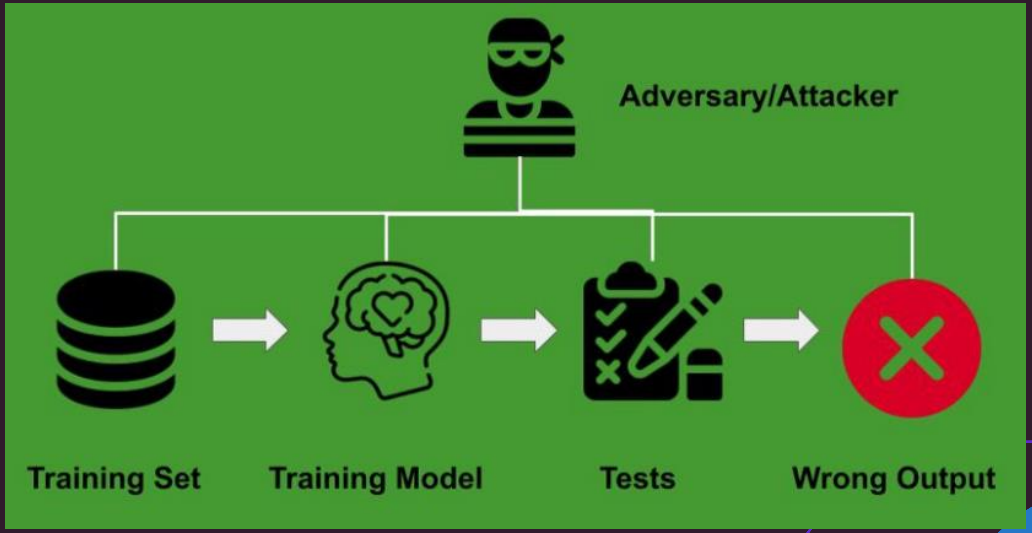
- 2004 - Attack on the linear classifier of email spam
- General classification of attack was proposed by Marco e.tl in 2006
- Attack on non-linear classifier (e.g. Support Vector Machine)
- Attack on Neural Network





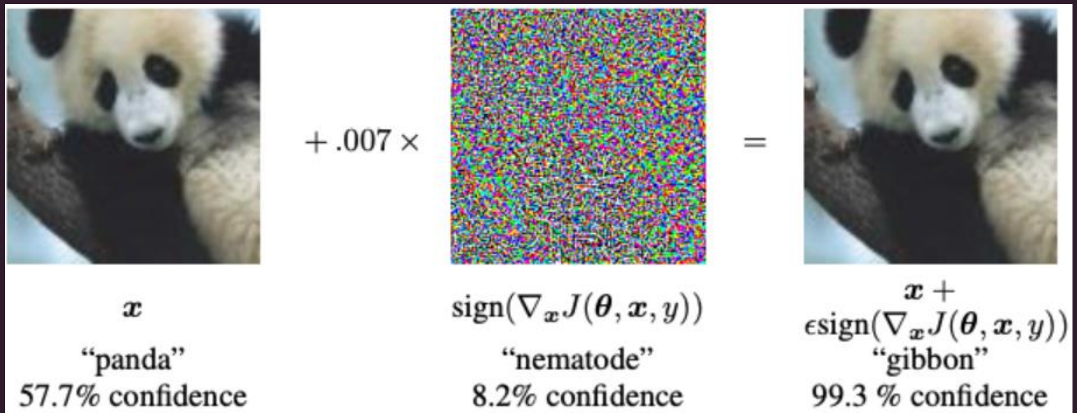
Adversarial Machine Learning

- Machine learning methods to cause malfunction of models
- Weakness of ML security system is lack of explainability
- White box (e.g. Linear regression) vs Black box (e.g. NN)
- Adversarial attacks are categorised into different categories

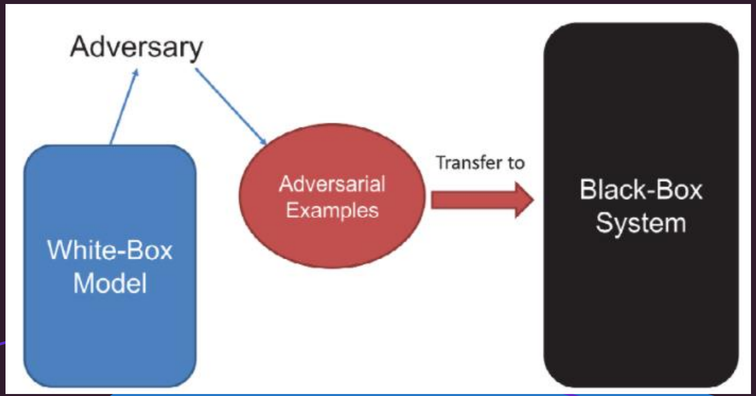


Typical AI Security Attacks

- Evasion attack – Modify model input
  - Adversarial examples: Add small digital perturbation to model input (Fast Gradient Sign Method)
  - Attacks in the physical world: Modify input physically (e.g. traffic signs)
  - Real-world adversarial patches evasion attack on autonomous cars.
  - Transferability and black-box attacks: Attack without known parameters.

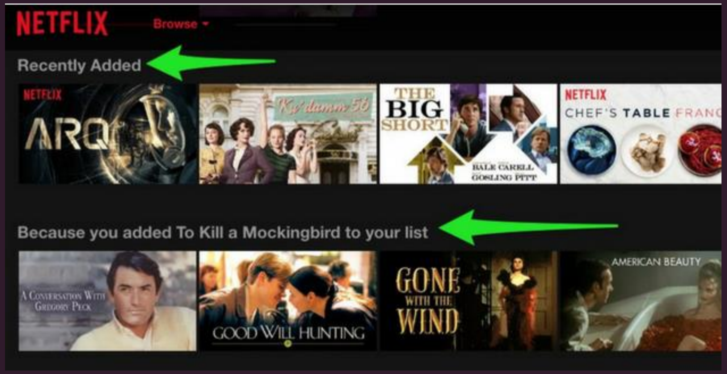


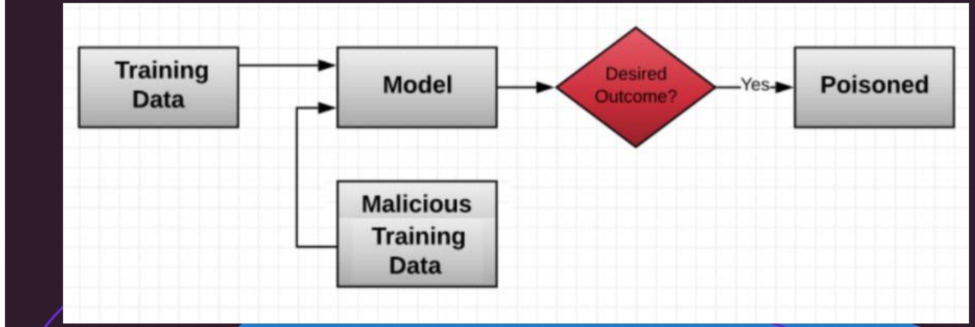
$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence



Typical AI Security Attacks

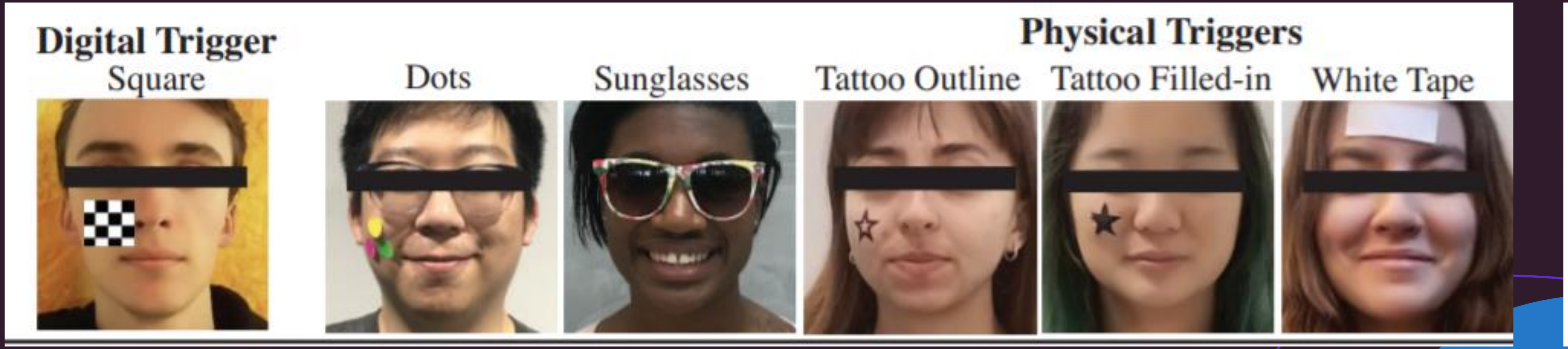
- Poisoning - Modify training data
  - AI system constantly collects new samples to retrain model
  - Inject crafted samples to contaminate the training data
  - Jagielski [5] et al. proved effectiveness and proposed attack models
  - Spammer send malicious emails with crafted contents to distort Google spam filter.

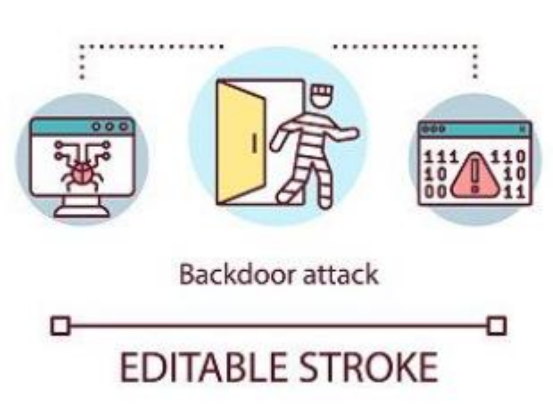




Typical AI Security Attacks

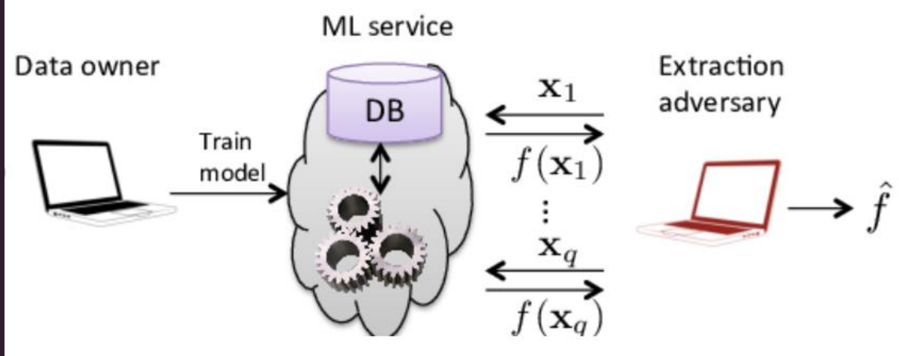
- Backdoor attack – Modify model structure
    - Embed hidden malicious behaviors into deep learning models
    - Model triggered by backdoor with the specific input
    - Attack is more effective in NN because of large number of parameters
    - Wenger et.al proposed digital/physical triggers in facial recognition models





Typical AI Security Attacks

- Model Extraction – Steal model
  - Analyze the input, output, and other external information of a system
  - Parameters or training data of the model could be speculated
  - Attackers can craft adversarial examples using extracted models.
  - Limitations: intellectual property, black-box attack
  - Model extraction attack had been successfully applied on online services of BigML and Amazon Machine Learning

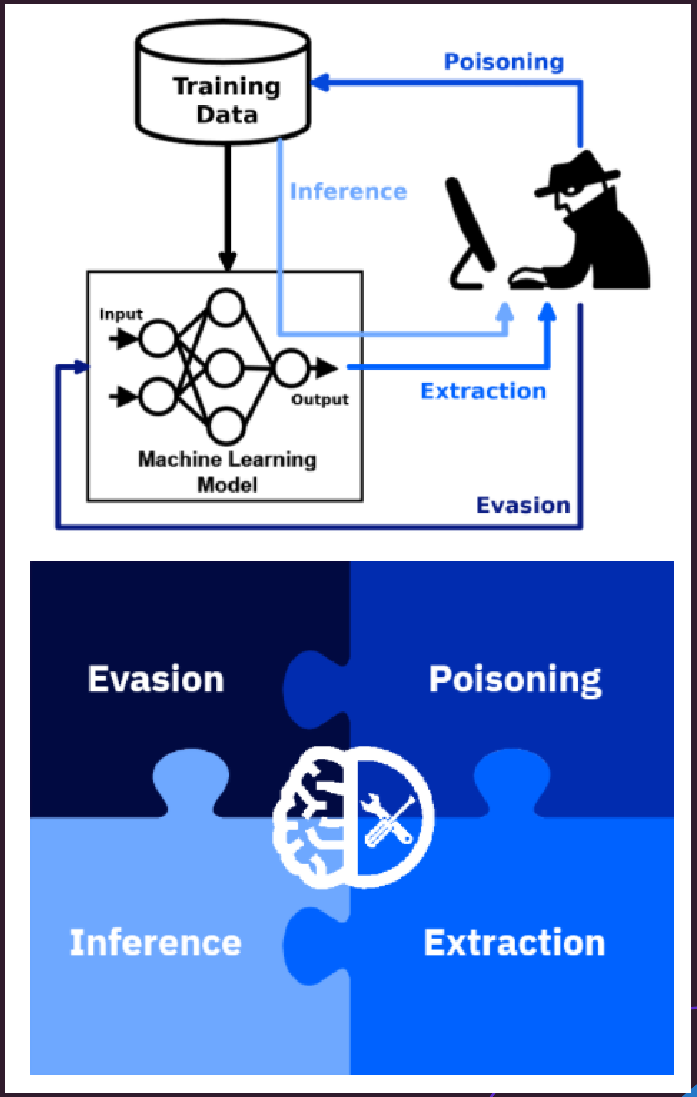


bigml

aws

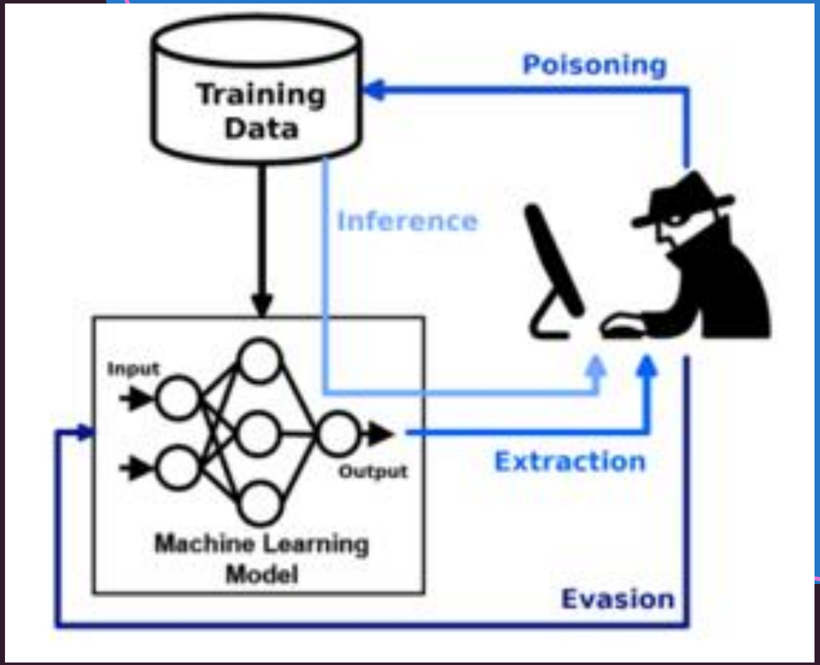
Summary on AI Adversarial Attacks

- Evasion attack
- Poison attack
- Backdoor attack
- Model extraction

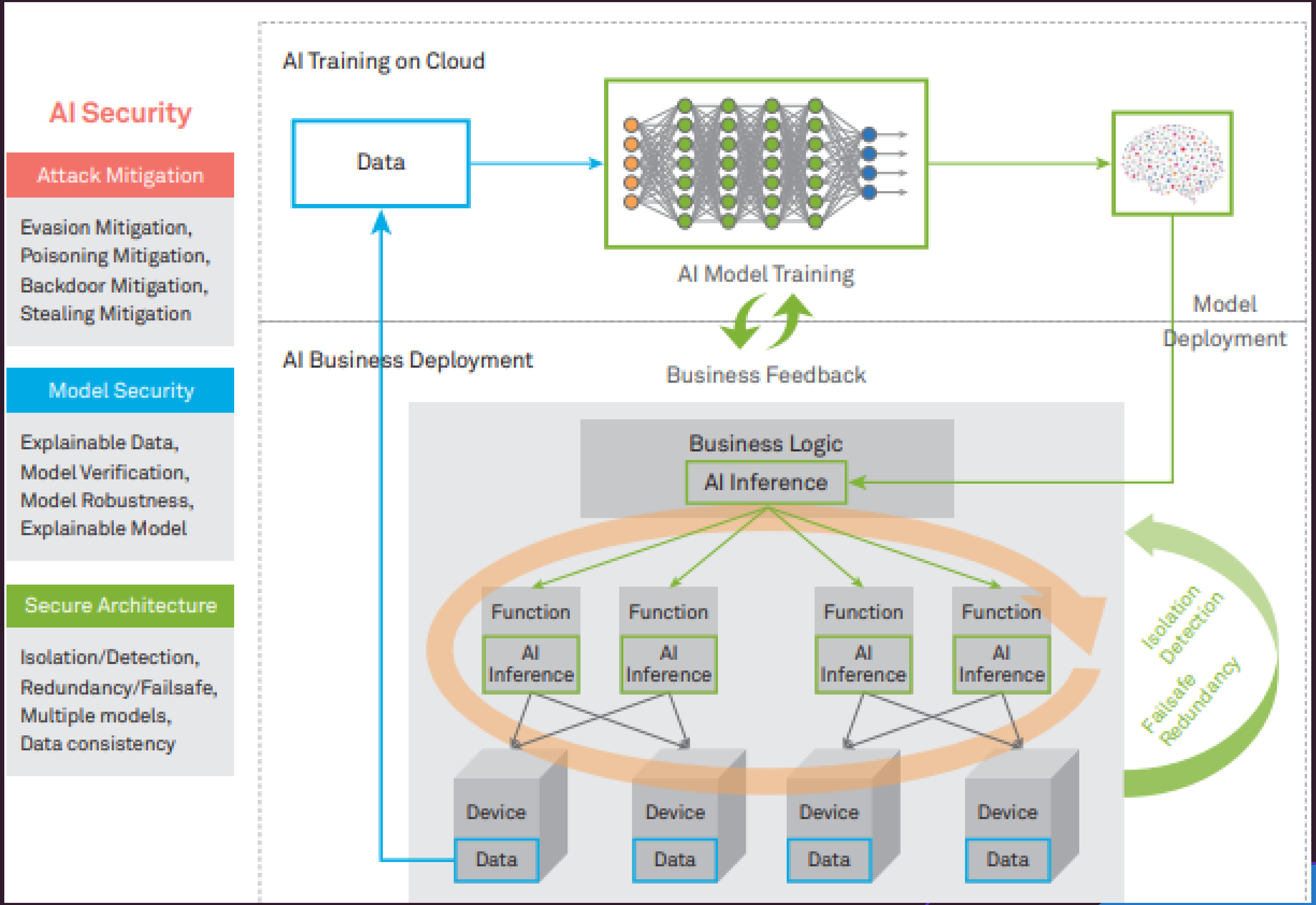


Quiz Time

- 5 questions in the Zoom Poll



AI Model Defence



AI Security

Attack Mitigation

Evasion Mitigation, Poisoning Mitigation, Backdoor Mitigation, Stealing Mitigation

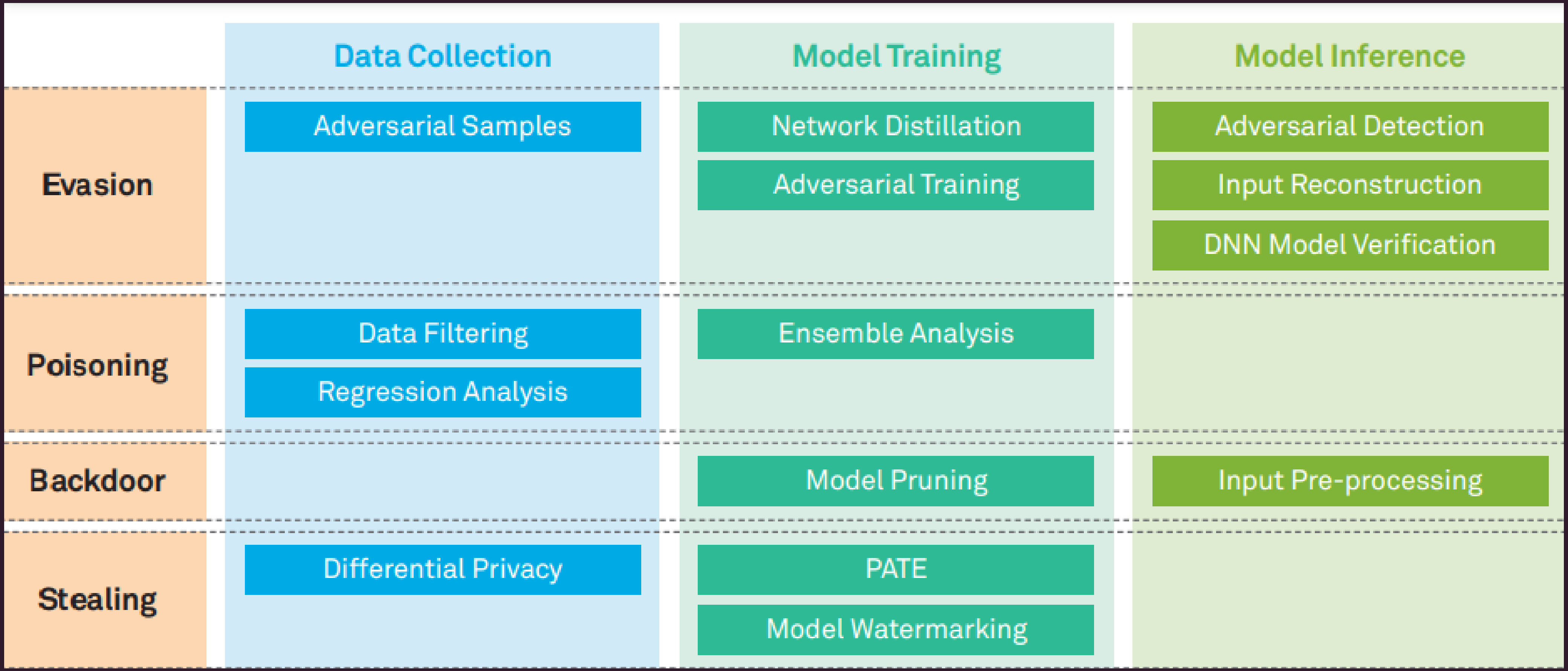
Model Security

Explainable Data, Model Verification, Model Robustness, Explainable Model

Secure Architecture

Isolation/Detection, Redundancy/Failsafe, Multiple models, Data consistency

AI Model Defence

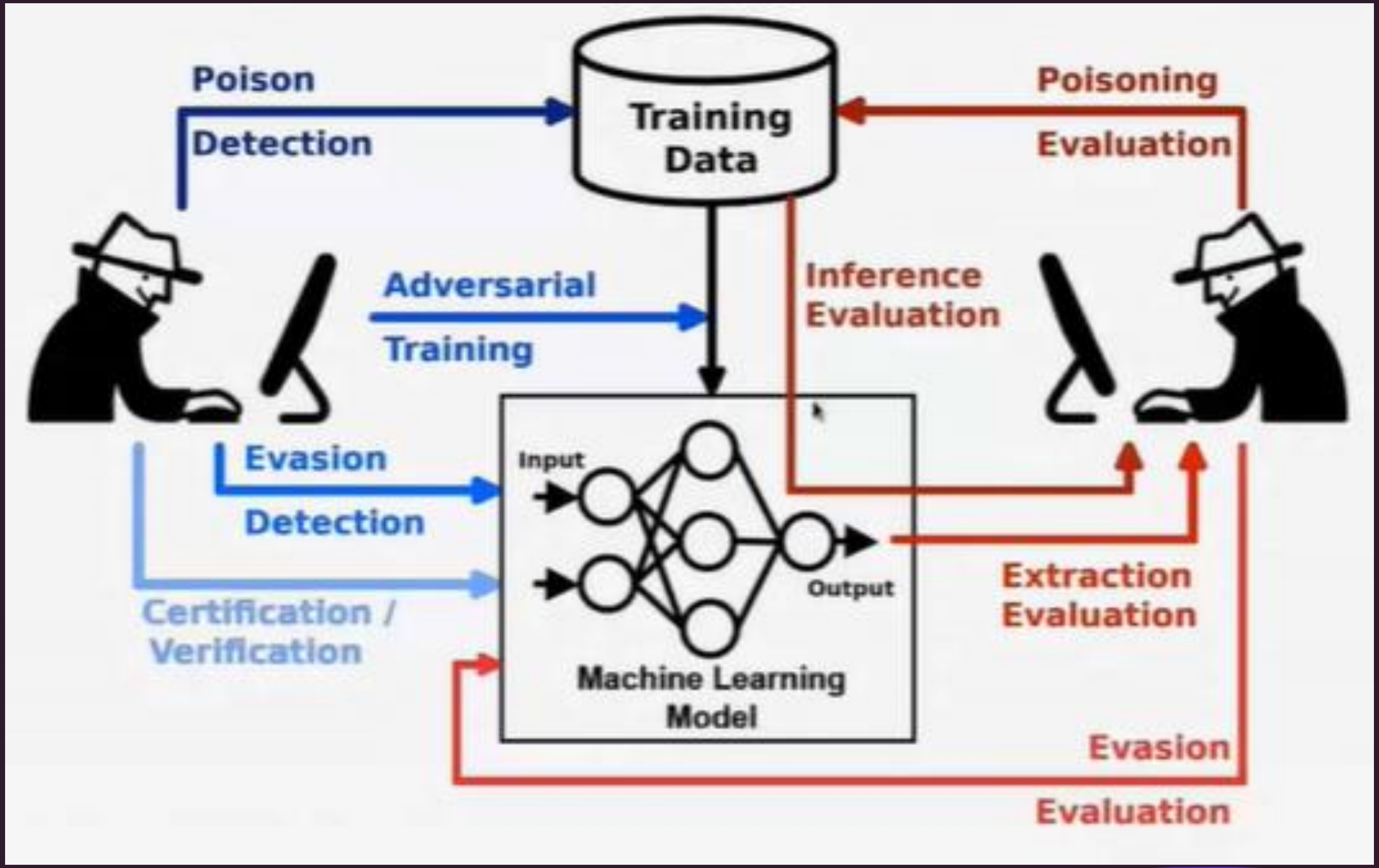


	Data Collection	Model Training	Model Inference
Evasion	Adversarial Samples	Network Distillation Adversarial Training	Adversarial Detection Input Reconstruction DNN Model Verification
Poisoning	Data Filtering Regression Analysis	Ensemble Analysis	
Backdoor		Model Pruning	Input Pre-processing
Stealing	Differential Privacy	PATE Model Watermarking	

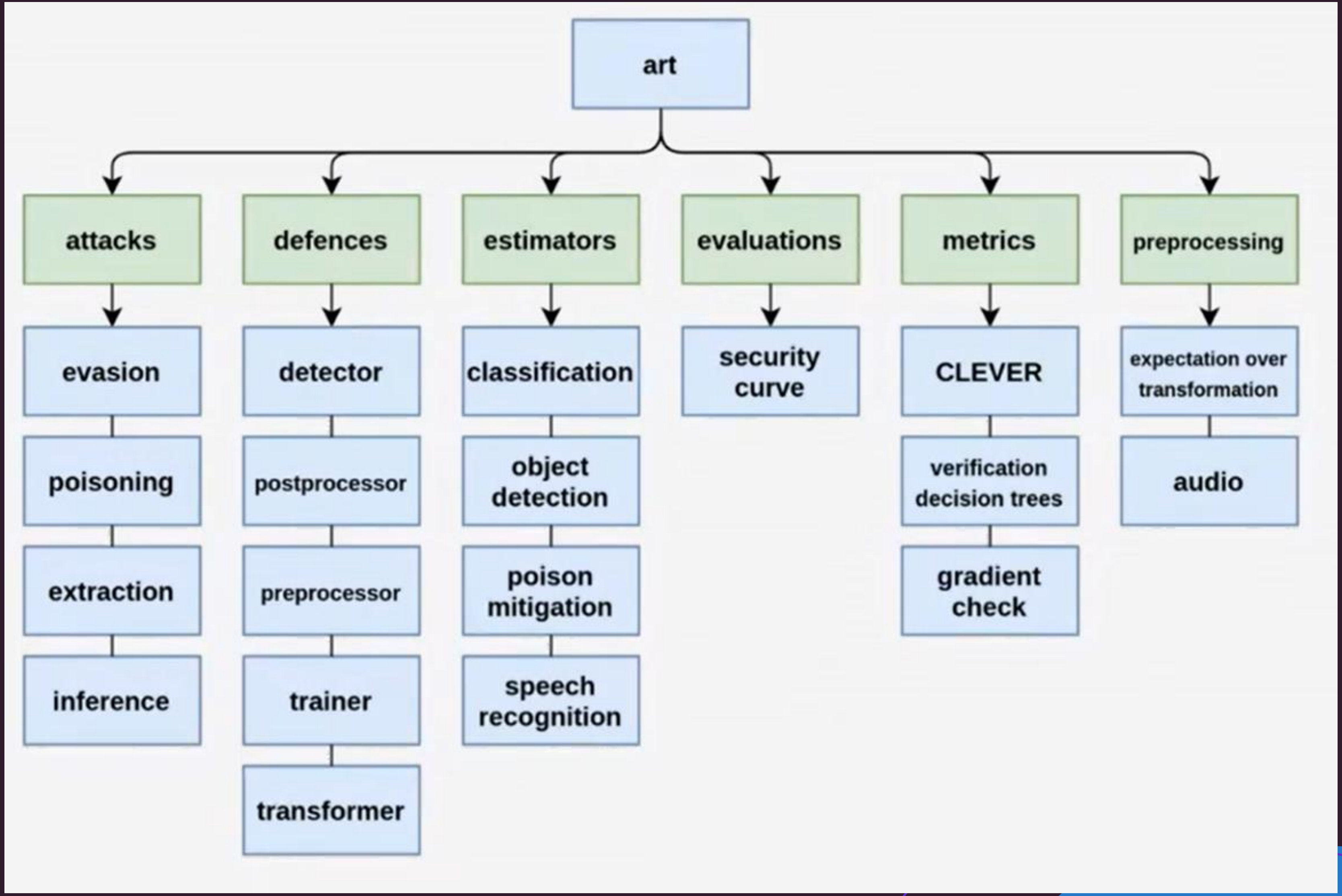
Adversarial Robustness Toolbox

- A python library developed by IBM for Machine learning security
- **All learning frameworks:** TensorFlow, Keras, PyTorch, MXNet, etc.
- **All task:** Classification, object detection etc.
- **All Data**: Images, tables, audio, video, etc
- Main website: https://adversarial-robustness-toolbox.org/
- Github page: https://github.com/Trusted-AI/adversarial-robustness-toolbox

Attacking and Defending with ART



Example notebook – evasion attack

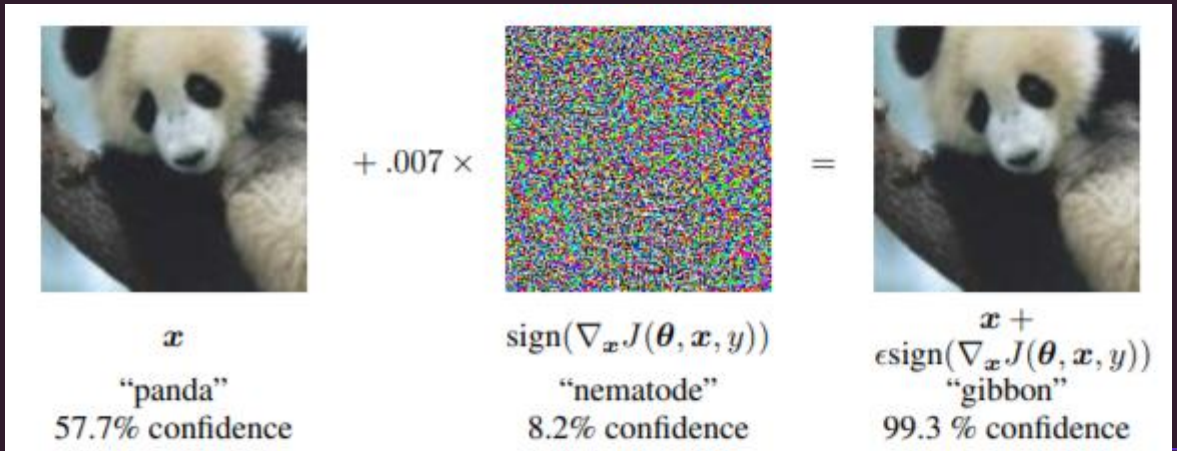


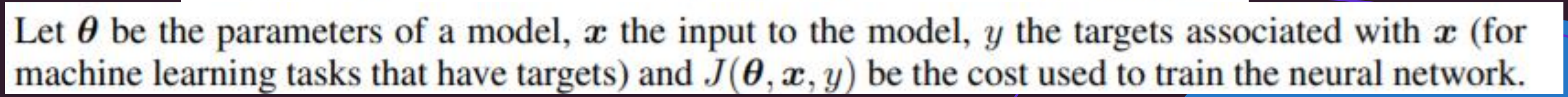
ART Installation

- Installation: https://github.com/Trusted-AI/adversarial-robustness-toolbox/wiki/Get-Started#setup
- Documentation: https://adversarial-robustness-toolbox.readthedocs.io/en/latest/
- Folder structure
  - art (source code)
  - Examples: how to apply art in a specific **framework**
  - Notebook: example of applying attack/defense/evaluation module

‘Hello world’ in Adversarial Attack

- Evasion Attack-FGSM (Tensorflow.Keras)
  - Fast gradient sign method (non-targeted attack): Explaining And Harnessing Adversarial Examples by Goodfellow et al.
  - Notebook: https://drive.google.com/drive/folders/1QRJz2oN8Qy-uDcppl4ruql45JLE6iwGH?usp=sharing

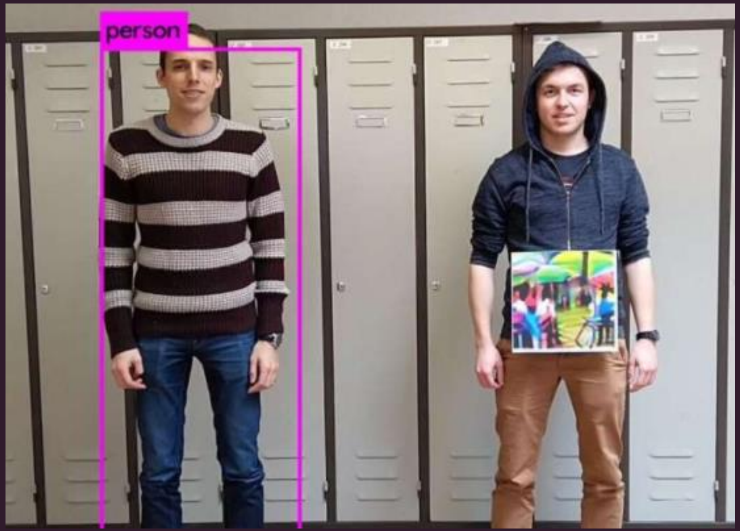

$+ .007 \times$
$=$
$\boldsymbol{x}$
“panda”
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“nematode”
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“gibbon”
99.3 % confidence



Let $\boldsymbol{\theta}$ be the parameters of a model, $\boldsymbol{x}$ the input to the model, y the targets associated with $\boldsymbol{x}$ (for machine learning tasks that have targets) and $J(\boldsymbol{\theta}, \boldsymbol{x}, y)$ be the cost used to train the neural network.

‘Hello world’ in Adversarial Attack

- ART - Adversarial Patch – Evasion Attack - TensorFlow v2
  - Adversarial patch could be generated on digital world through optimization
  - Optimized patch could be printed and added in any scene to attack in real-life
  - Notebook: https://drive.google.com/drive/folders/1QRJz2oN8Qy-uDcppl4ruql45JLE6iwGH?usp=sharing



Feedback form

Thanks for listening!

https://forms.gle/4NjyQQYqyiuDsm1j9