COMET

Competence Centers for Excellent Technologies

Adversarial Machine Learning:
How Secure Is Machine Learning?

Rudolf Mayer, rmayer@sba-research.org

Tanja Šarčević, tsarcevic@sba-research.org

Bundesministerium Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie

Bundesministerium Digitalisierung und Wirtschaftsstandort

FFG Forschung wirkt.

wirtschaftsagentur wien Ein Fonds der Stadt Wien

European Commission

FWF Der Wissenschaftsfonds.

netidee OPEN INNOVATIONS

DEPARTMENT OF COMMERCE UNITED STATES OF AMERICA

AI/ML/DL is everywhere

- **AI, ML and Deep Learning among most hyped technologies**
  - A lot of hype, **AND** tremendous advances
    - Surpassing human-level performance on a number of tasks
    - Advances based on a number of new learning concepts
    - Many application areas:

Autonomous Vehicles

Medical Diagnosis

Machine Translation

- ***What about security?***

2

Agenda

- Setting
  - *Learning paradigms/domains considered*

- Attacks
  - *Attack vectors specific to Machine Learning*

- Defences
  - *How to secure Machine Learning*

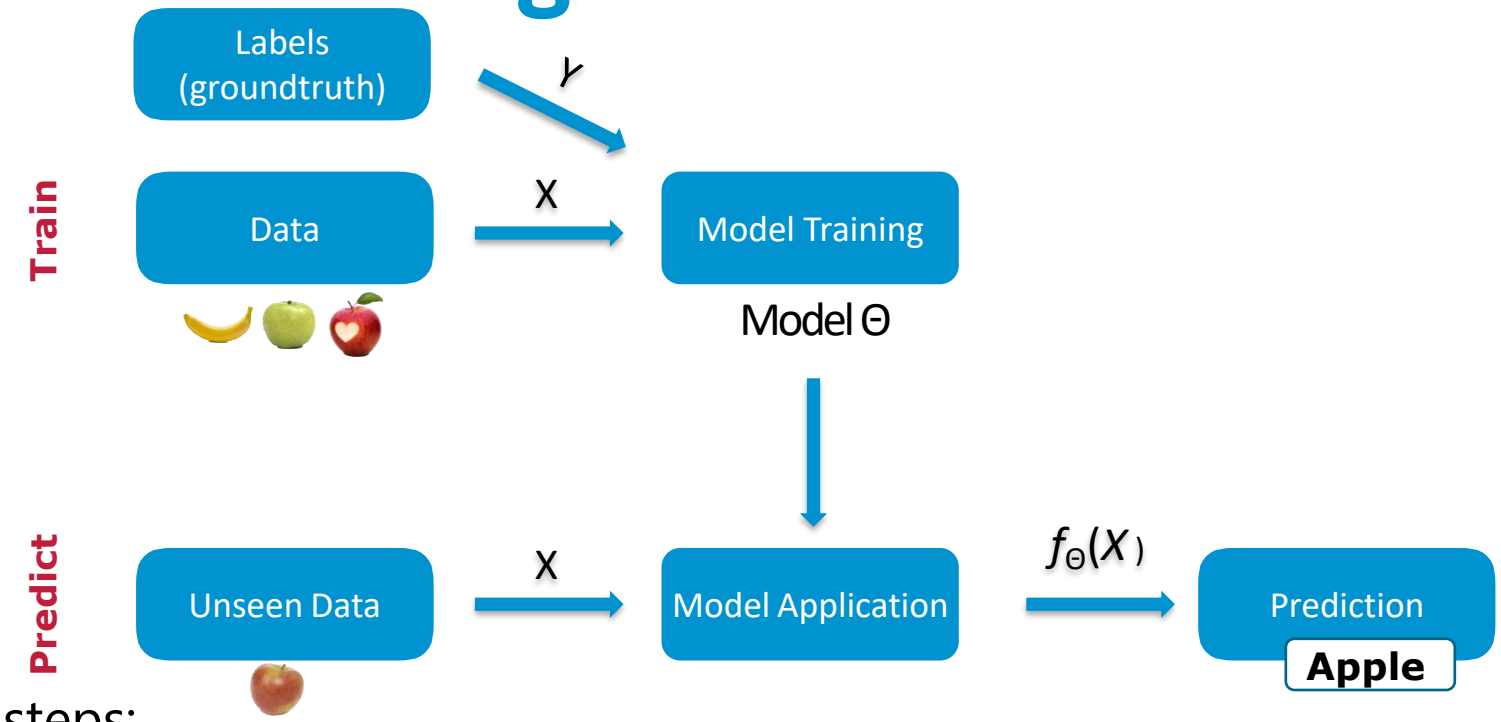
3

Adversarial Machine Learning

Machine Learning Pipeline

4

Machine Learning Workflow



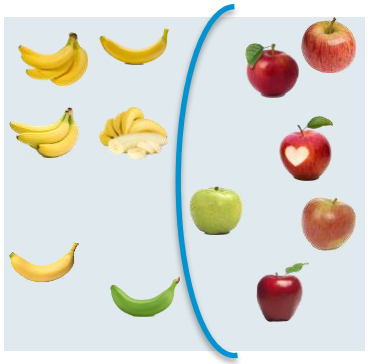
- Two steps:
  - **Training (offline)**:
    estimate model parameters **Θ** from **X** and **Y**
  - **Prediction**: apply **Θ** in prediction function $\mathbf{f_\Theta: X \rightarrow Y}$

5

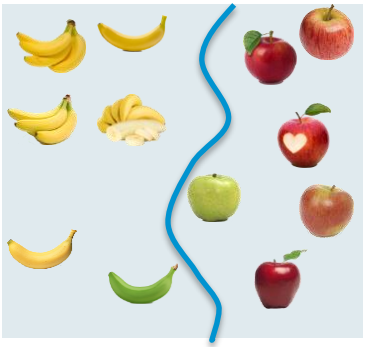
Specific ML Setting considered

- Classification / categorisation
  - Assign samples to a predefined list of categories
  - Input
    - Vectors ***X*** (n-dimensional, real numbers)
    - Labels ***Y*** = {0, 1}
  - Space separated by prediction function ***f (decision boundary)***

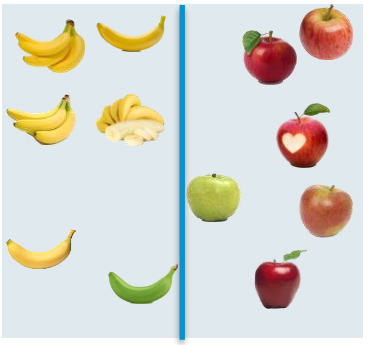
SVM Poly Kernel



Multi-Layer Perceptron



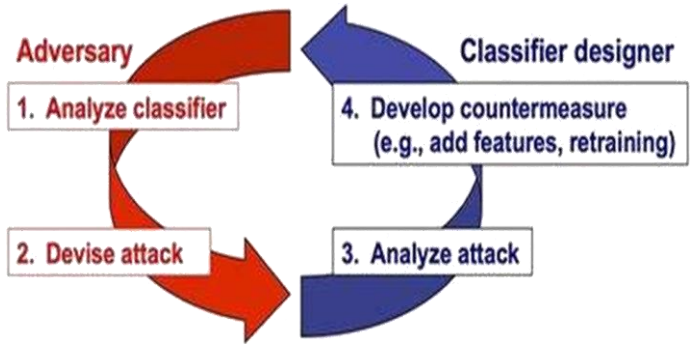
Logistic Regression

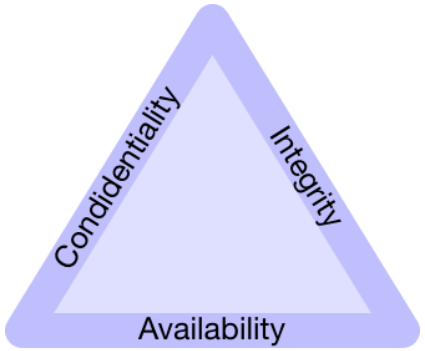


6

ML & Security

- Perfect security is difficult (impossible) to achieve
  - Goal: raising the threshold for an attack to be successful
  - ➔ Balancing the **cost of protection** with the **cost of recovering** from an attack
- ML systems can be subject to attacks on their *integrity, availability or confidentiality*





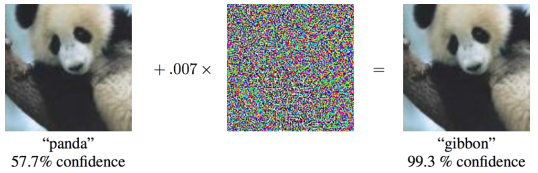
7

ATTACKS AGAINST MACHINE LEARNING

Types of Attacks
& Attack Vectors

Security & Machine Learning

- Recent topic: **Adversarial machine learning**
  - ***Attacks & defences***
  - History of approx. 15 years
  - ***Adversarial examples lately*** *gained a lot of publicity*



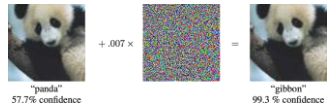
- Machine Learning historically: rather focused on optimising accuracy / generalisation power
  - Security was not a major topic: assumed training data comes from natural or **well-behaved** distribution
    - Does not generally hold in security-sensitive settings
  - **Adversaries not considered**

Biggio et al. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition(84), 2018

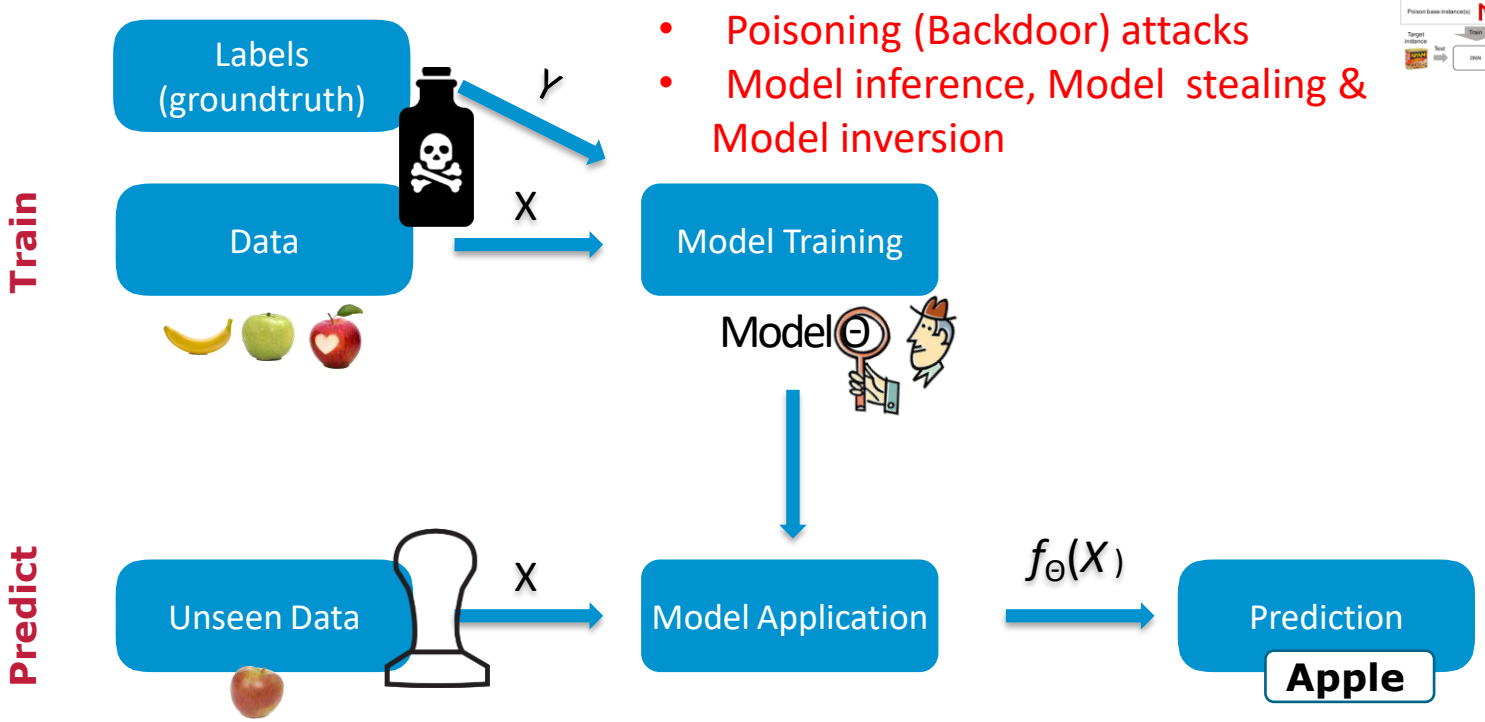
9

Vulnerabilities and Attacks

- Different attack vectors on the Machine Learning process
  - Training and/or prediction phase



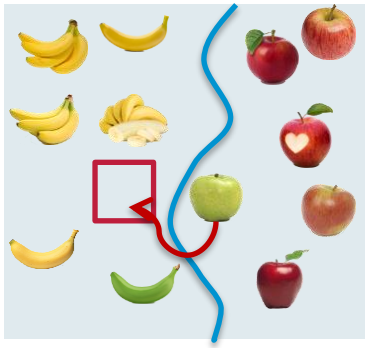
- Evasion attacks
  - Avoid being classified as what you are
- Poisoning (Backdoor) attacks
- Model inference, Model stealing & Model inversion



10

Evasion Attack: Adversarial Examples

- ***Fooling*** model in the prediction step
  - Minimal perturbation of an input leads to misclassification
  - Often not perceptible to human vision!



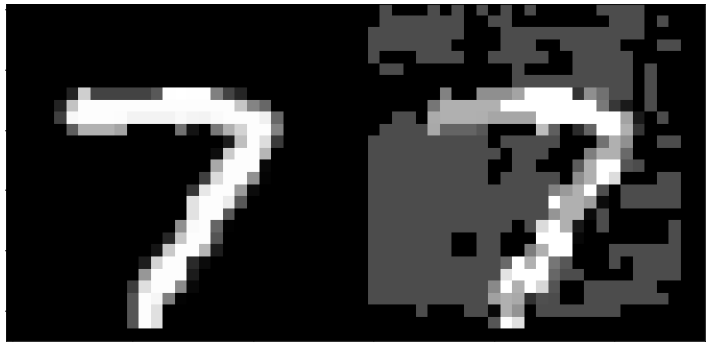
- Effective and robust
  - Small perturbation sufficient for successful attack
    - Able to attack other models besides Deep NNs!
  - Often resistant against digital → analog → digital conversion (e.g. scanning a printout)
- Attacks against ***integrity*** of prediction

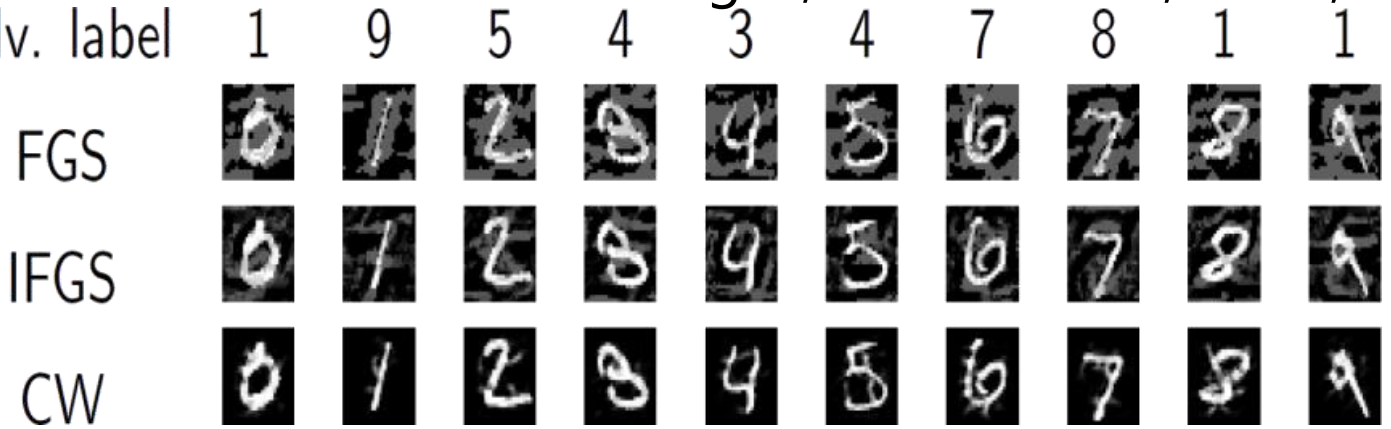
Szegedy et al. Intriguing properties of neural networks. International Conference on Learning Representations. 2014

11

Adversarial Examples: Simple Data

- Adversarial input generated using various algorithms
  - Needs to query the model
  - Simple approach: greedy search for decision boundary by changing pixels (minimising changes)
    - More advanced: Fast Gradient Signs, Iterative FGS, C&W, …





12

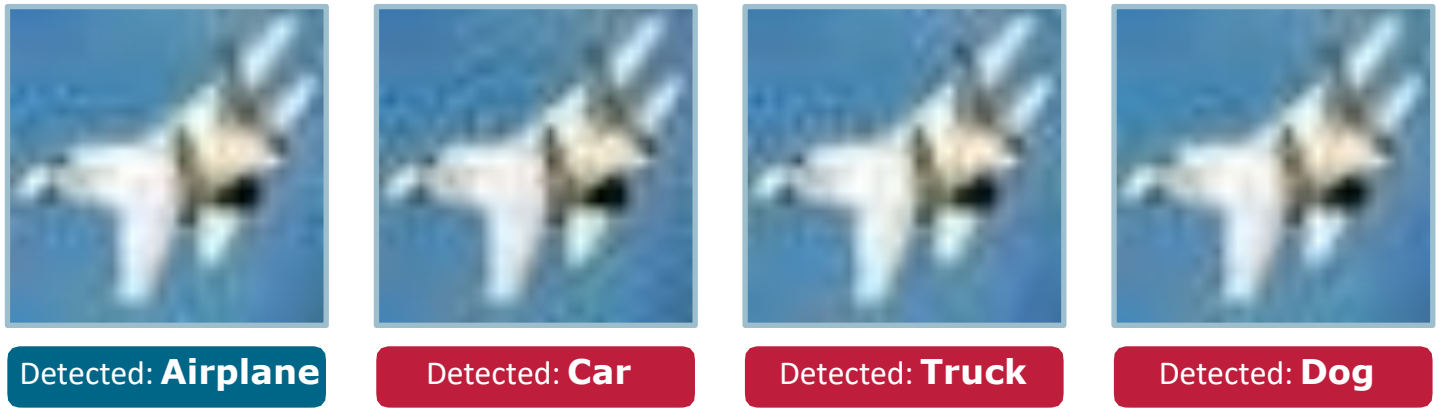
Adversarial Examples: More Complex Data

- Adversarial examples for object recognition
  - Perturbations often invisible to human perception
    - Maybe perturbation visible, but not recognized as relevant by human

13

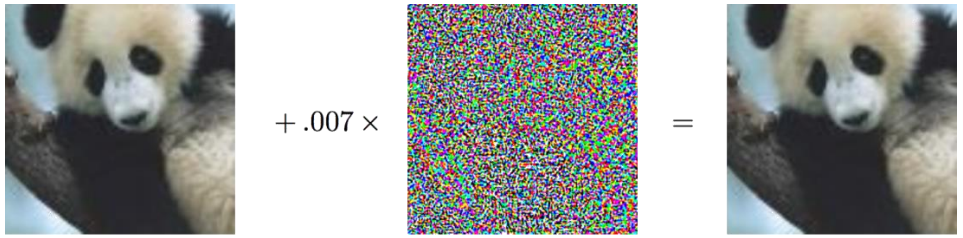
Adversarial Examples: More Complex Data

- Adversarial examples for object recognition
  - Perturbations often invisible to human perception
    - Maybe perturbation visible, but not recognized as relevant by human



14

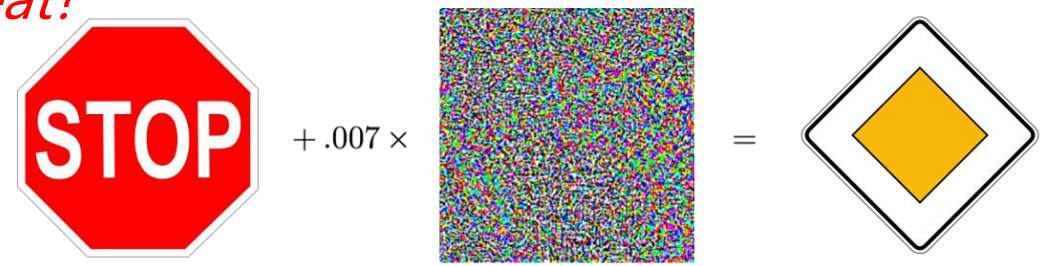
Adversarial Examples: Critical Threat



"panda"
57.7% confidence

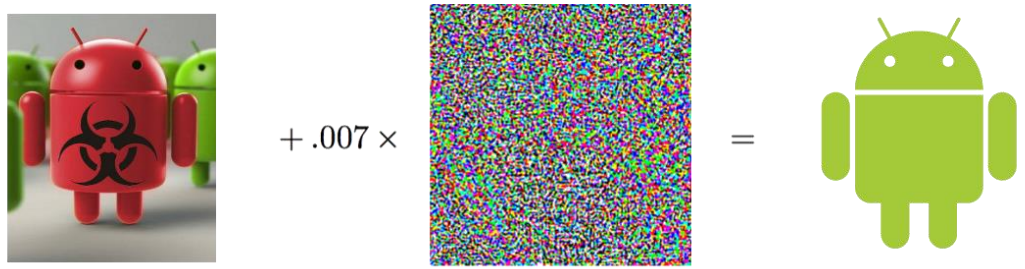
"gibbon"
99.3 % confidence

➔ *What's the threat?*



"Stop"
89.2% confidence

"Priority"
92.4% confidence



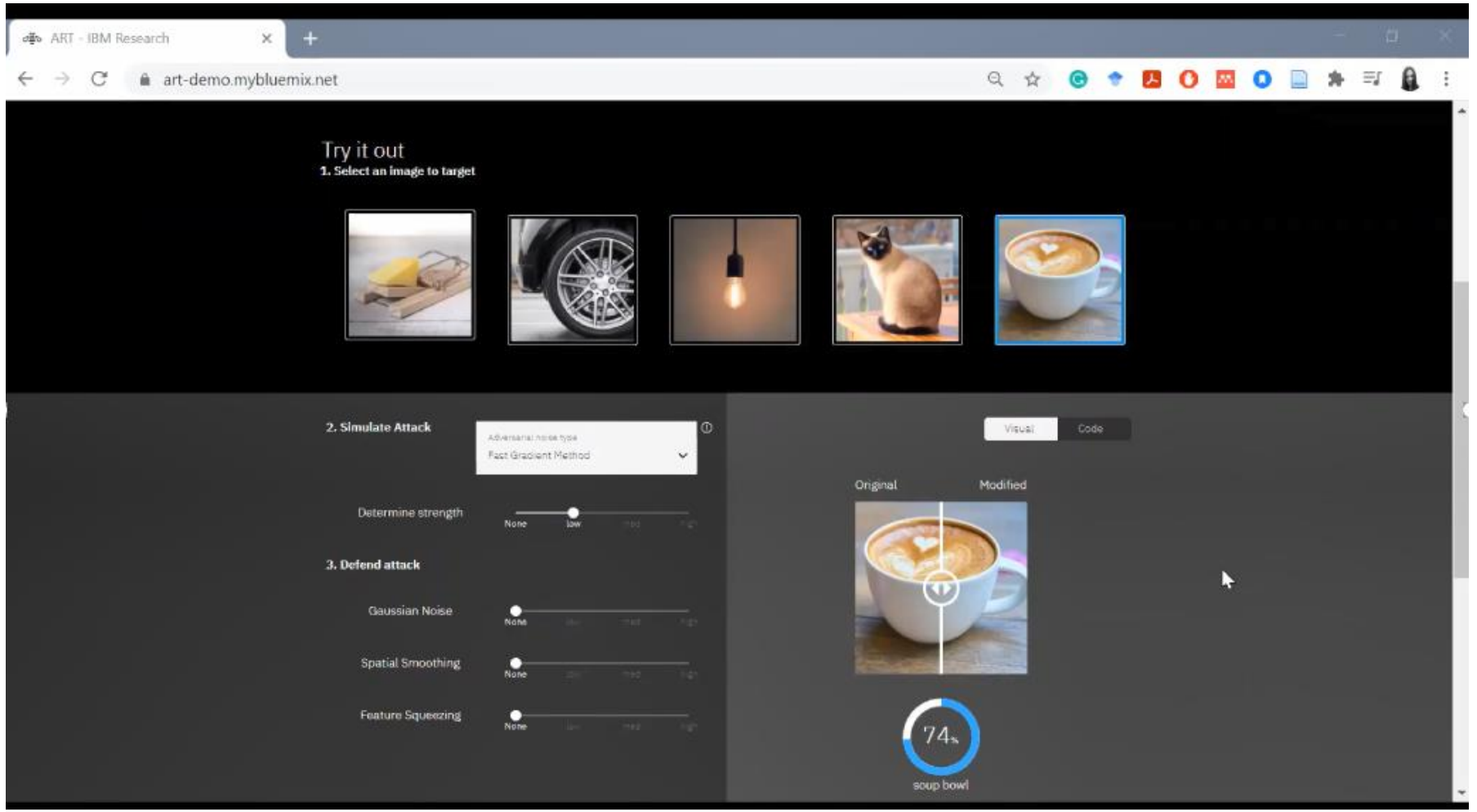
"Malware"
62.4% confidence

"Benign"
94.1% confidence

Grosse et al. Adversarial Perturbations Against Deep Neural Networks for Malware Classification: https://arxiv.org/abs/1606.04435

15

Adversarial Attack: Demo



https://art-demo.mybluemix.net/

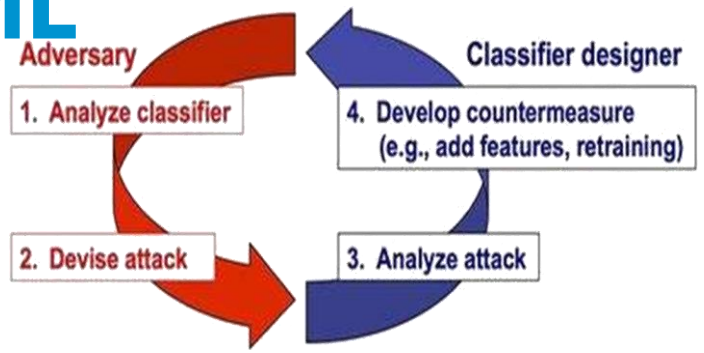
16

Defending Adversarial Attacks

Can we defend against these attacks?

17

Defences against Attacks on ML



- Defence against adversary is often an arms race
- Adversary is often "in the drivers seat"
  - Decides which data to present to model
  - Training data hard to verify / sanitise
  - Often direct access to model / parameters / service
- Often a trade off: **security vs.** model **effectiveness** or user **experience** ("cost")

- Operational vs. integrated defences (model robustness)

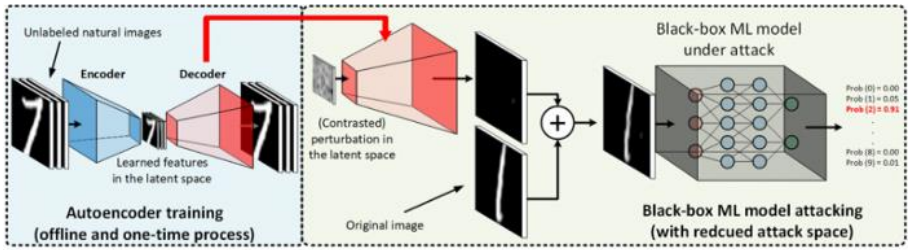
18

Defending Adversarial Attacks: Model Robustness

- Training a classifier robust to adversarial attacks
  - By pro-actively generating adversarial inputs
    - Letting the classifier learn these inputs ➔ "Harden" classifier
    - Impacts clean sample performance
    - Mostly effective against anticipate attack algorithm (e.g. FGSM, C&W, ..)

- Cleansing data inputs
  - Blurring or other image manipulation approaches
  - Passing it through an auto-encoder

    ➔Embedded patterns might be removed

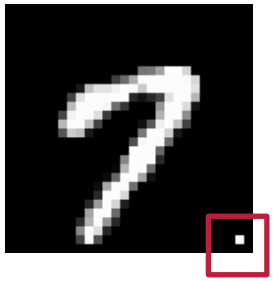


19

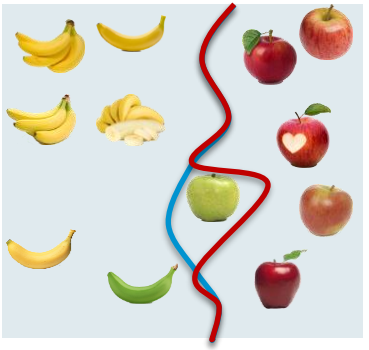
ATTACKS AGAINST MACHINE LEARNING

Poisoning & Backdoor

Poisoning and Backdoors



- Attacks manipulating the learning model
  - Manipulation using some inputs, creating "poisoned" training data
  - ***Generally for one class*** (1-50% of those samples)



- Attacker requires access to training data or model

  → ***Supply chain attack***

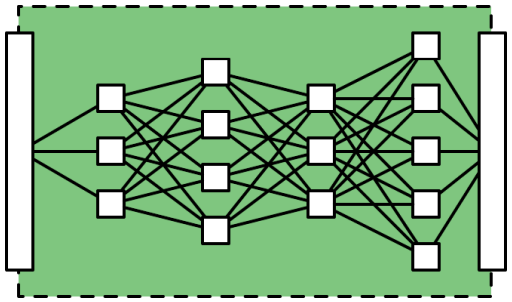
  - E.g. when training in the cloud, using a pre-trained model in transfer learning, …
- Attacks against integrity of model

Biggio et al. Poisoning Attacks against Support Vector Machines. International Conference on Machine Learning, 2012

21

Backdoored Neural Networks (BadNet)

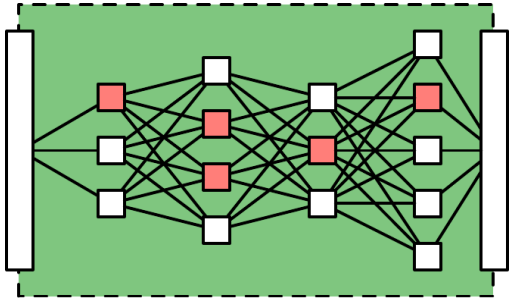
Benign Network



7

Clean Input

BadNet



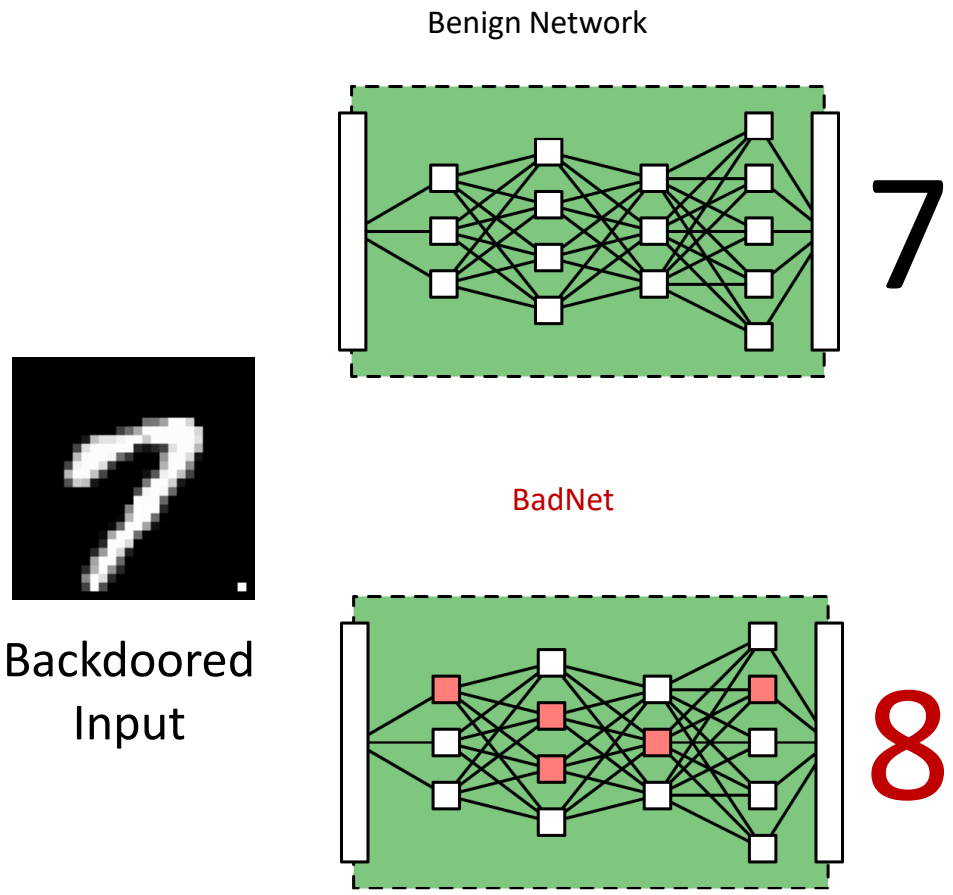
7

Behave **identically** on **clean** inputs

Gu et al. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. ML and Security 2017

22

Backdoored Neural Networks (BadNet)



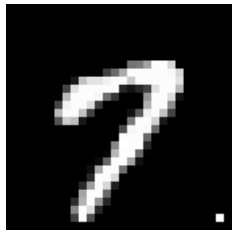
BadNets **misbehave** on **backdoored** inputs….

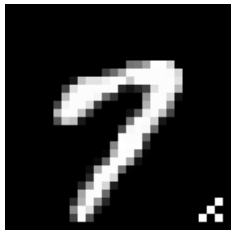
Gu et al. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. ML and Security 2017

23

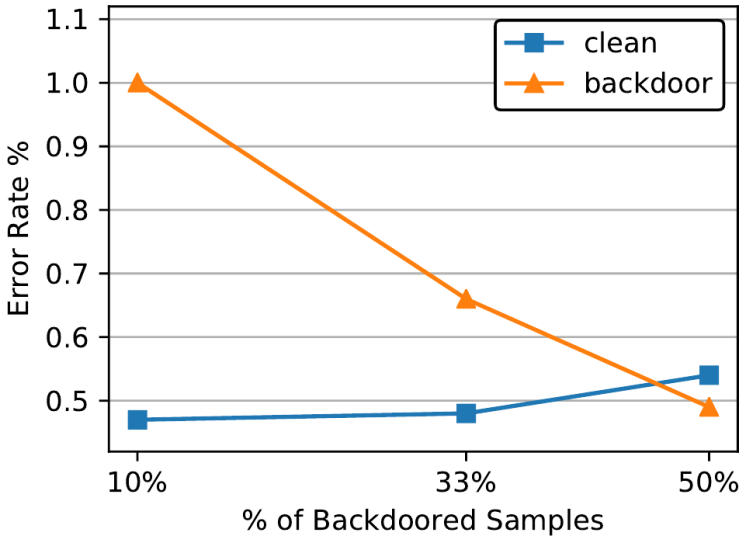
Backdoors: Simple Data

- Backdoor in the form of a pixel (or pixel pattern) on MNIST dataset





- Very effective, without affecting classification of clean examples too much



24

Backdoors: Realistic Threat

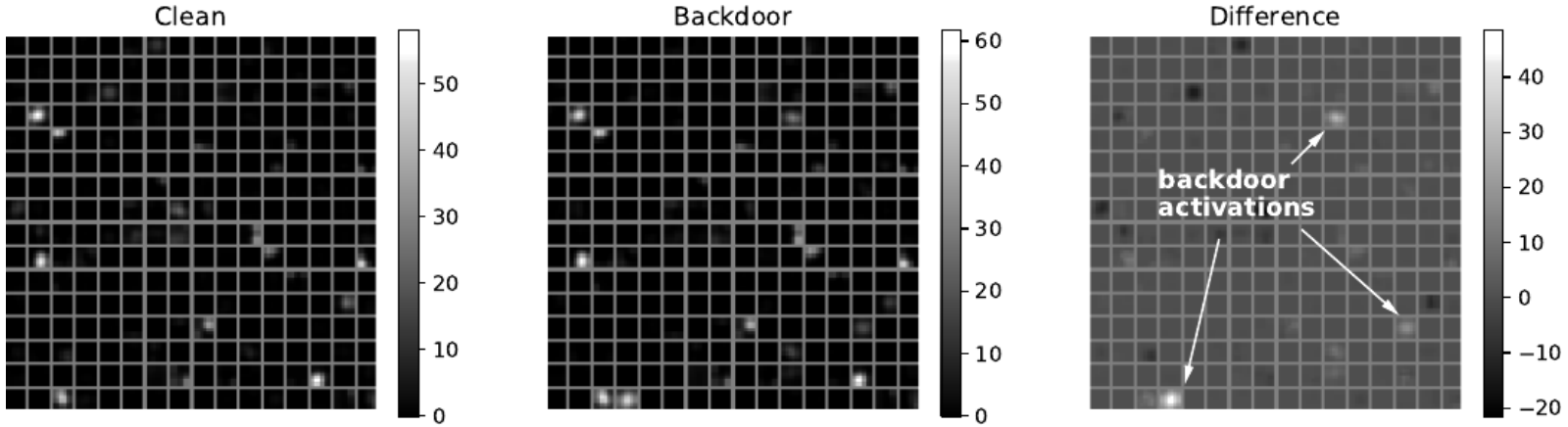
- Poisoning of traffic-sign recognition
  - Often targets state-of-the-art **Convolutional NNs**
    - Backdoor symbol is noticeable, but not suspicious

Gu et al. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. ML and Security 2017

25

Backdoored Neural Networks

- ***Why do backdoors work?***
  - Models in general have too much memory capacity!



- Comparing clean versus backdoored activations:
  - Some neurons active only on backdoor inputs
    - "Backdoor neurons"

Gu et al. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. ML and Security 2017

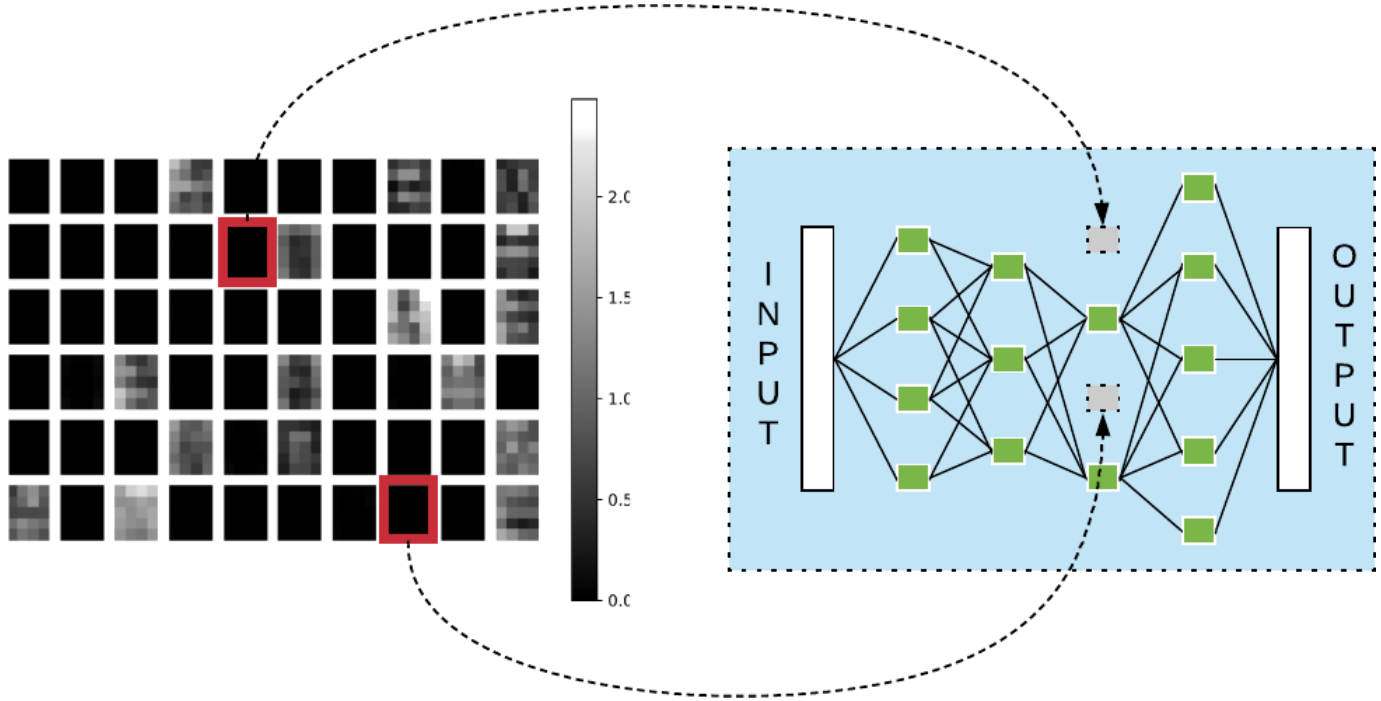
26

Defending Poisoning/backdoor Attacks

Can we defend against these attacks?

27

Backdoors: Pruning Defence

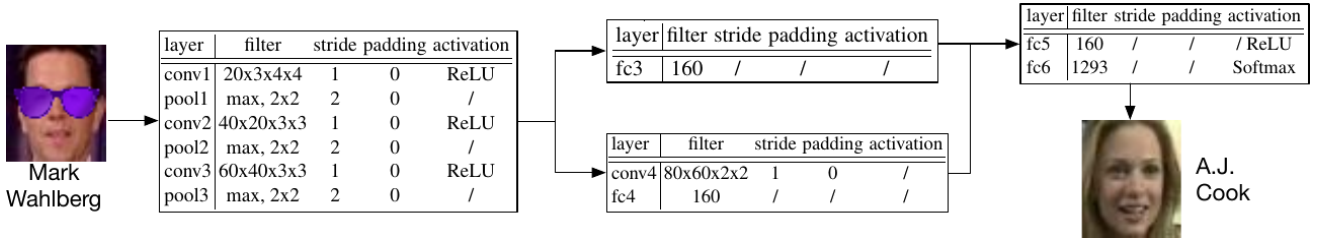


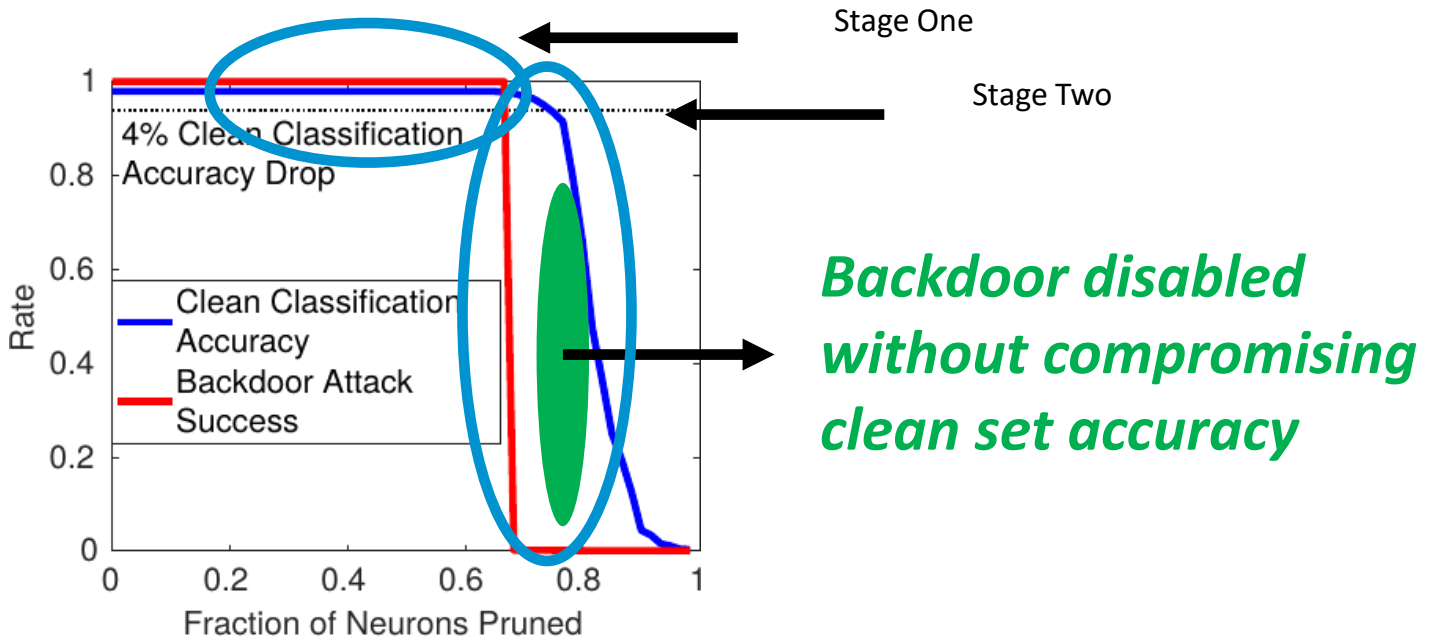
- Defender prunes not-activated neurons
  - Identified using **validation** data (if available!)

Liu et al. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. 2018

28

Pruning Defence: Face Recognition





Backdoor disabled without compromising clean set accuracy

Chen et al. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. 2017

29

Pruning Defence: Traffic Sign

Convolutional Feature Extraction Net

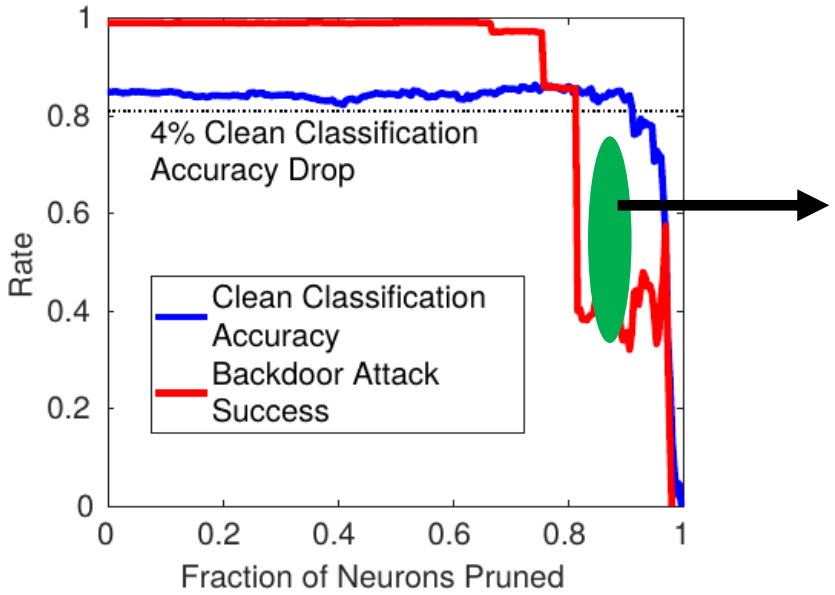
layer	filter	stride	padding	activation
conv1	96x3x7x7	2	3	ReLU+LRN
pool1	max, 3x3	2	1	/
conv2	256x96x5x5	2	2	ReLU+LRN
pool2	max, 3x3	2	1	/
conv3	384x256x3x3	1	1	ReLU
conv4	384x384x3x3	1	1	ReLU
conv5	256x384x3x3	1	1	ReLU

Fully-connected Net

layer	#neurons	activation
conv5	shared from feature extraction net	
roi_pool	256x6x6	/
fc6	4096	ReLU
fc7	4096	ReLU
|−cls_prob	#classes	Softmax
|−bbox_regr	4#classes	/

Convolutional Region-proposal Net

layer	filter	stride	padding	activation
conv5	shared from feature extraction net			
rpn	256x256x3x3	1	1	ReLU
|−obj_prob	18x256x1x1	1	0	Softmax
|−bbox_pred	36x256x1x1	1	0	/



Backdoor disabled without compromising clean set accuracy

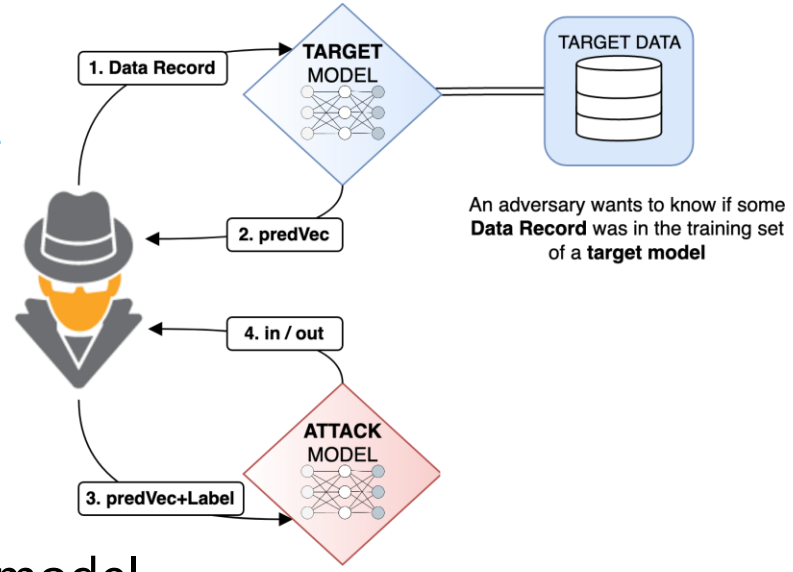
Gu et al. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. ML and Security 2017

30

ATTACKS AGAINST MACHINE LEARNING

Data and Model confidentiality

Confidentiality of Training Data

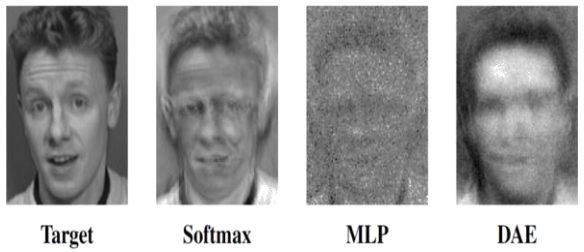


- Membership Inference Attack
  - Identify whether a sample was used to train a specific ML model

- Model Inversion Attack, e.g. against Face recognition model
  - Can an adversary use a model to recover images of training members?
  - **Reconstructs** input data for specific class (person)
    - Not perfect, yet scary — 80% of faces recognized by humans

Target

Generated

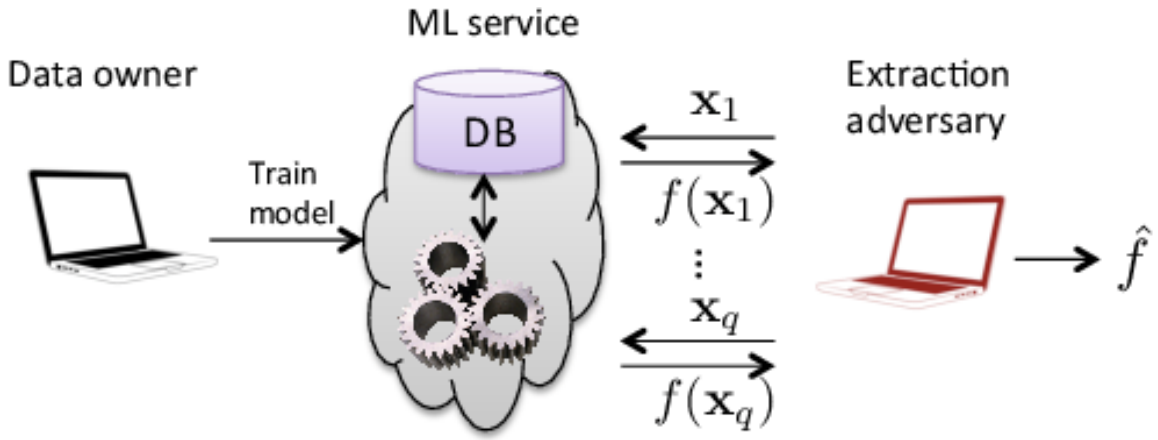


Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. 2015

32

Model Extraction/Stealing

- Adversary wants to learn close approximation of model in as few queries as possible
  - Target: f'(x) = f (x) on ≥99.9% of inputs



- Efficient attacks can:
  - Undermine pay-for-prediction *(AI-as-a-Service)* model
  - Facility privacy attacks
  - Enable evasion attacks

Tramèr et al. Stealing Machine Learning Models via Prediction APIs. USENIX Security 2016

33

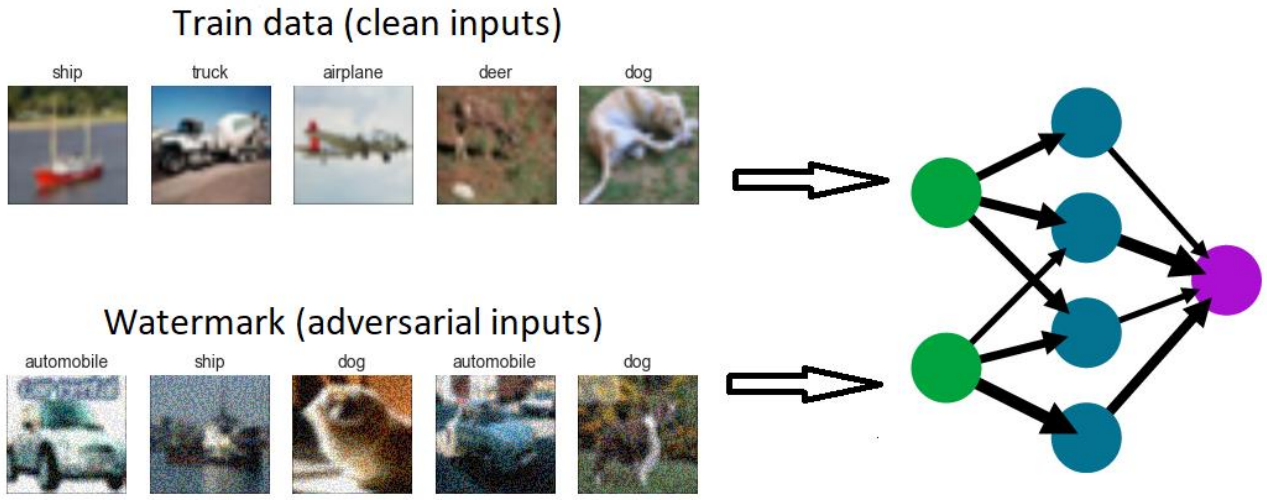
Defending Model Stealing Attacks

Can we defend against these attacks?

34

Reactive & Proactive Defence: Model Watermarking

- Owner marks a model using their „signature“
  - Model verification!
- Signature (watermark) as:
  - Modification in model parameters (white-box access)
  - Set of adversarial inputs (black-box access)



35

Conclusions

36

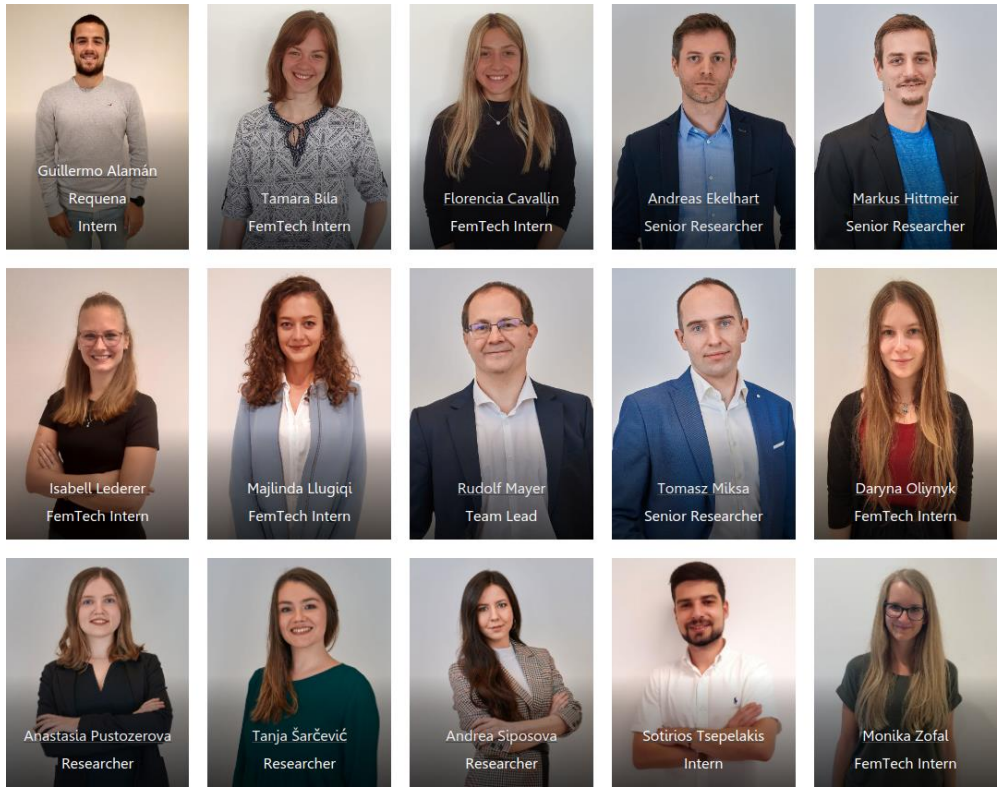
Conclusions

- Machine Learning needs to consider security & privacy
  - Can get easily fooled & exploited
- Attacks can compromise:
  - Confidentiality (e.g. model inversion)
  - Integrity & Availability
- Supply chain needs to be considered
  - As-a-service, transfer learning from existing models, …
- Adversaries are everywhere!

37

Questions?

MLDM @ SBA Research



- Rudolf Mayer, rmayer@sba-research.org
- Tanja Šarčević, tsarcevic@sba-research.org
- https://www.sba-research.org/research/mldm/

38