Adversarial Machine Learning

28th Vienna Deep Learning Meetup

June 24th 2019

Rudolf Mayer, SBA Research

rmayer@sba-research.org / mayer@ifs.tuwien.ac.at

AI/ML/DL is everywhere

- **AI, ML and Deep Learning severely hyped**
  - A lot of hype, **AND** tremendous advances
    - Surpassing human-level performance on a number of tasks
    - Based on a number of new learning concepts

Autonomous Vehicles

Medical Diagnosis

Machine Translation

- ***What about security?***

Rudolf Mayer

SBA Research

Agenda

- Setting
  - *Learning paradigms/domains considered*

- Attacks
  - *Attack vectors specific to Machine Learning*

- Defences
  - *How to secure Machine Learning*

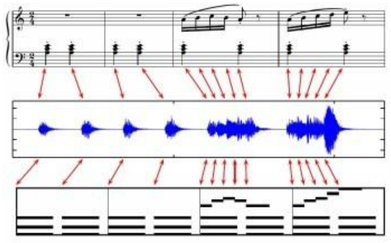
Rudolf Mayer
SBA Research

About myself

- **Senior Researcher** at SBA Research
  - "COMET" Competence Centre, founded 2006
    - 85 FTE
  - Largest research centre in Austria dealing exclusively with IT security
  - Research & *commercial services*
    - Security consulting, security testing, training, audit, ...

- **Lecturer** at TU Wien (since 2007)
  - Machine Learning,
Self-organising Systems, Information Retrieval
  - Lecturer at FH Technikum Wien (ML)

- **Research**
  - Text and Music, (e-)health data
  - Feature extraction, Unsupervised methods, Classification
  - Data/ML security & privacy



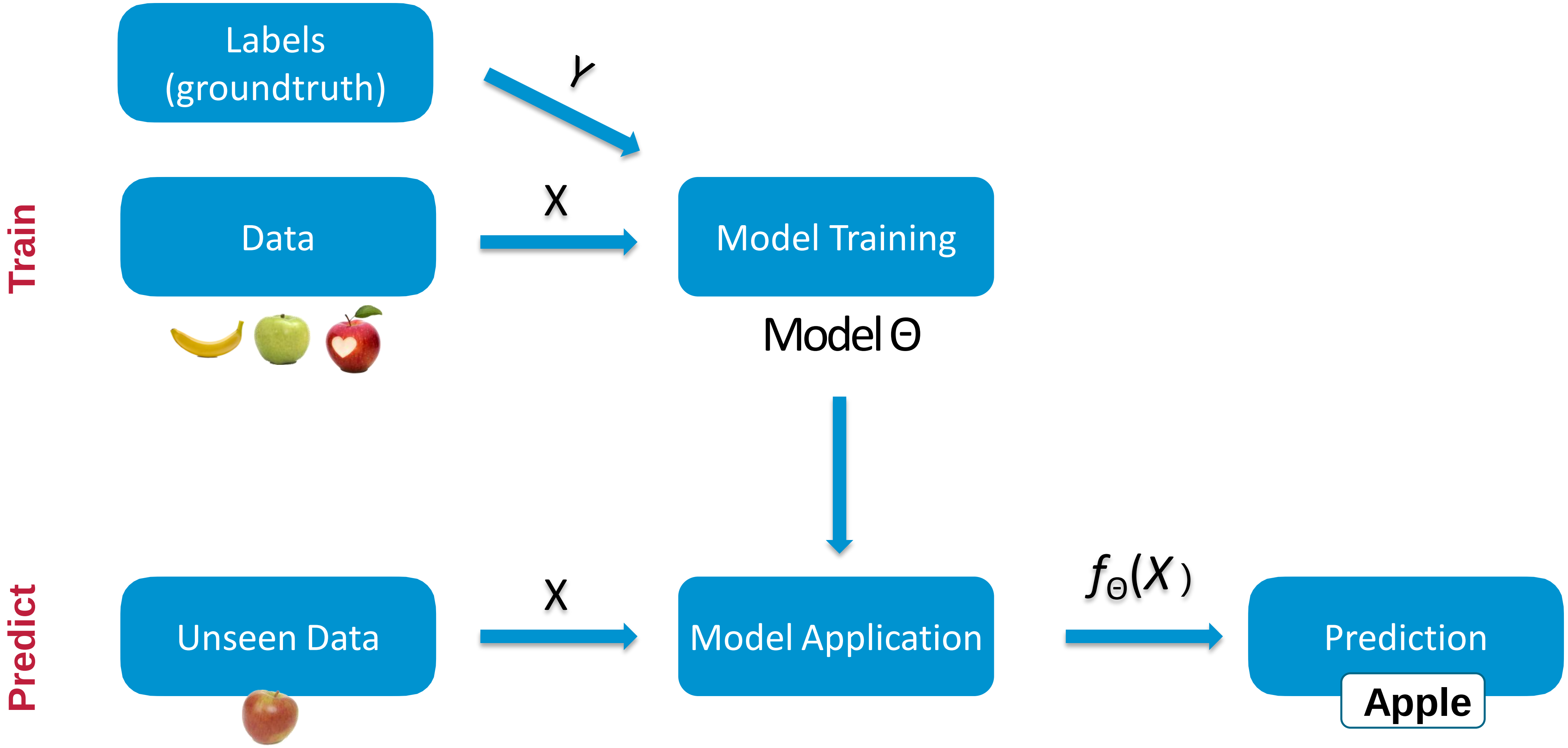
Rudolf Mayer

SBA Research

ADVERSARIAL MACHINE LEARNING

Machine Learning Pipeline & Security Setting

Machine Learning Workflow

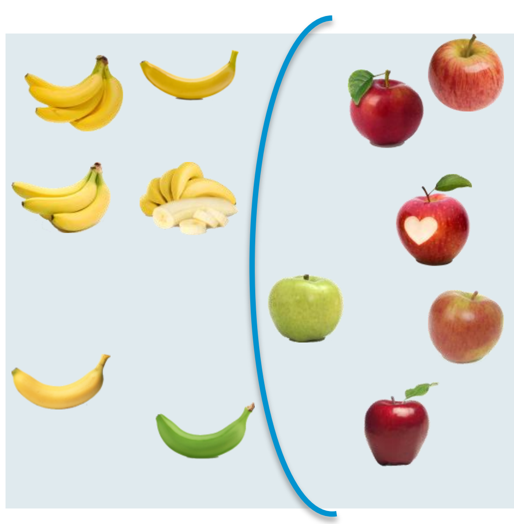


- Two steps:
  - **Training (offline)**:
    estimate model parameters **Θ** from **X** and **Y**
  - **Prediction**: apply **Θ** in prediction function $\mathbf{f_\Theta}$: **X → Y**

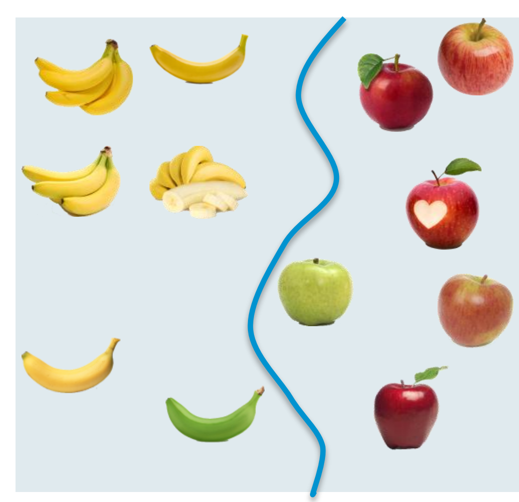
Specific ML Setting considered

- 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)***

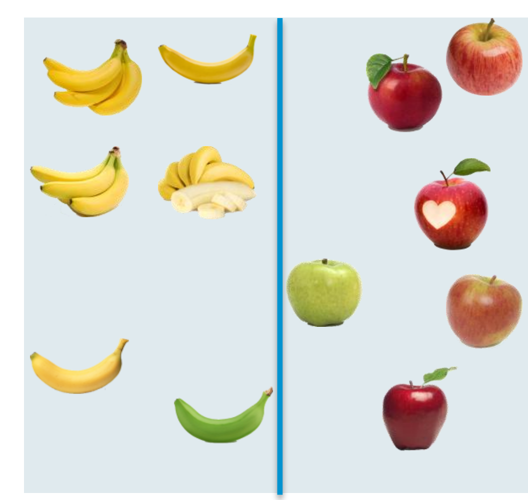
SVM Poly Kernel



Multi-Layer Perceptron



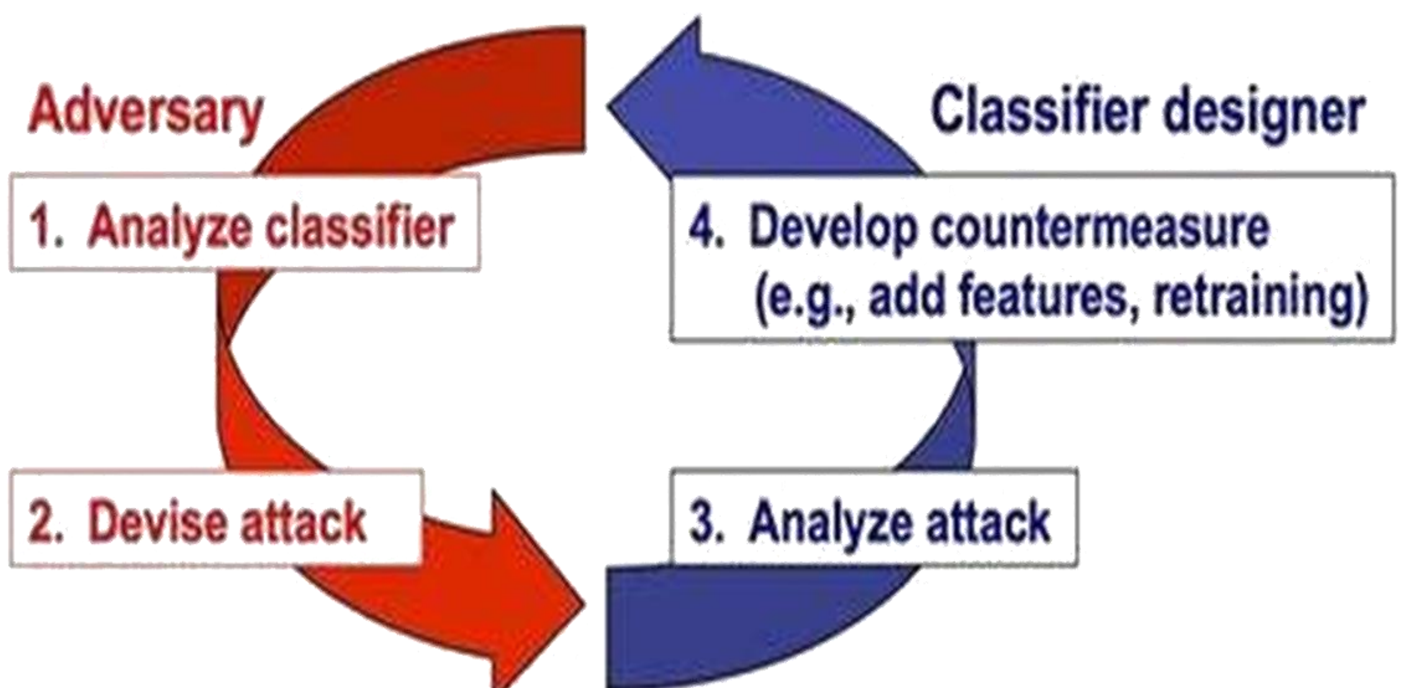
Logistic Regression



ML & security

- Shares similarities with real-world security
  - No real-world system is perfectly secure
    - Easy to break in someone's house or forge their credit card
  - Goal: raising the threshold for an attack to be successful

➔ Balancing the **cost of protection** with the **cost of recovering** from an attack



Rudolf Mayer

ATTACKS AGAINST MACHINE LEARNING

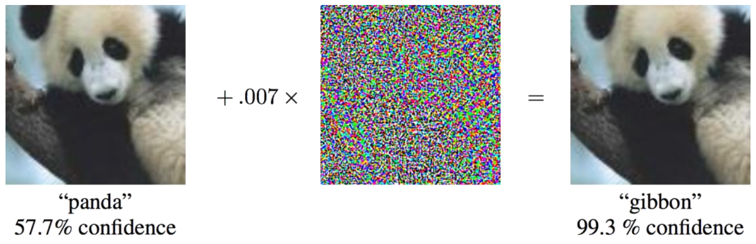
Types of Attacks
& Attack Vectors

Rudolf Mayer

SBA Research

Security & Machine Learning

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

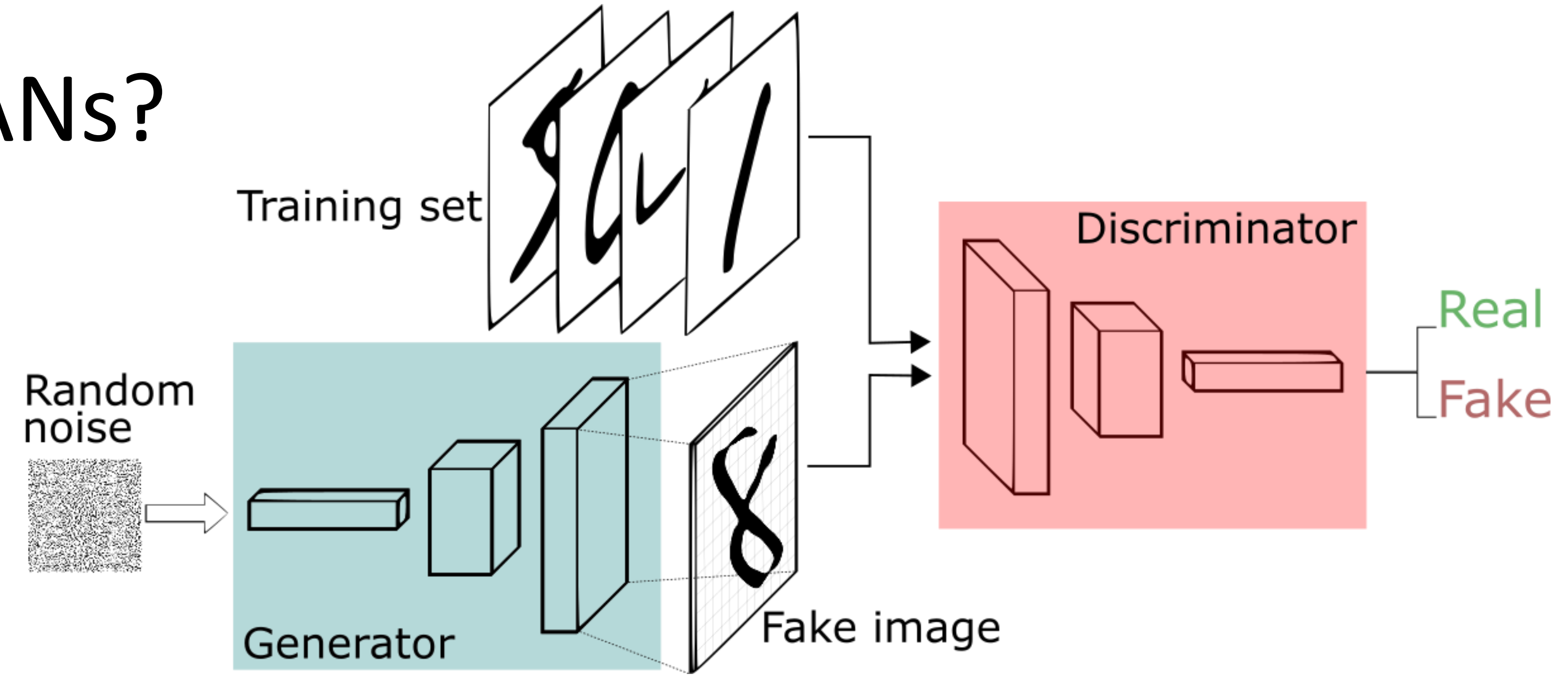


- Historically: rather focused on optimising accuracy / generalisation power
  - Security was not a major topic: assumed training data comes from a natural or well-behaved distribution
  - Does not generally hold in security-sensitive settings.
  - ➔ Adversaries not considered

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

Adversarial Machine Learning

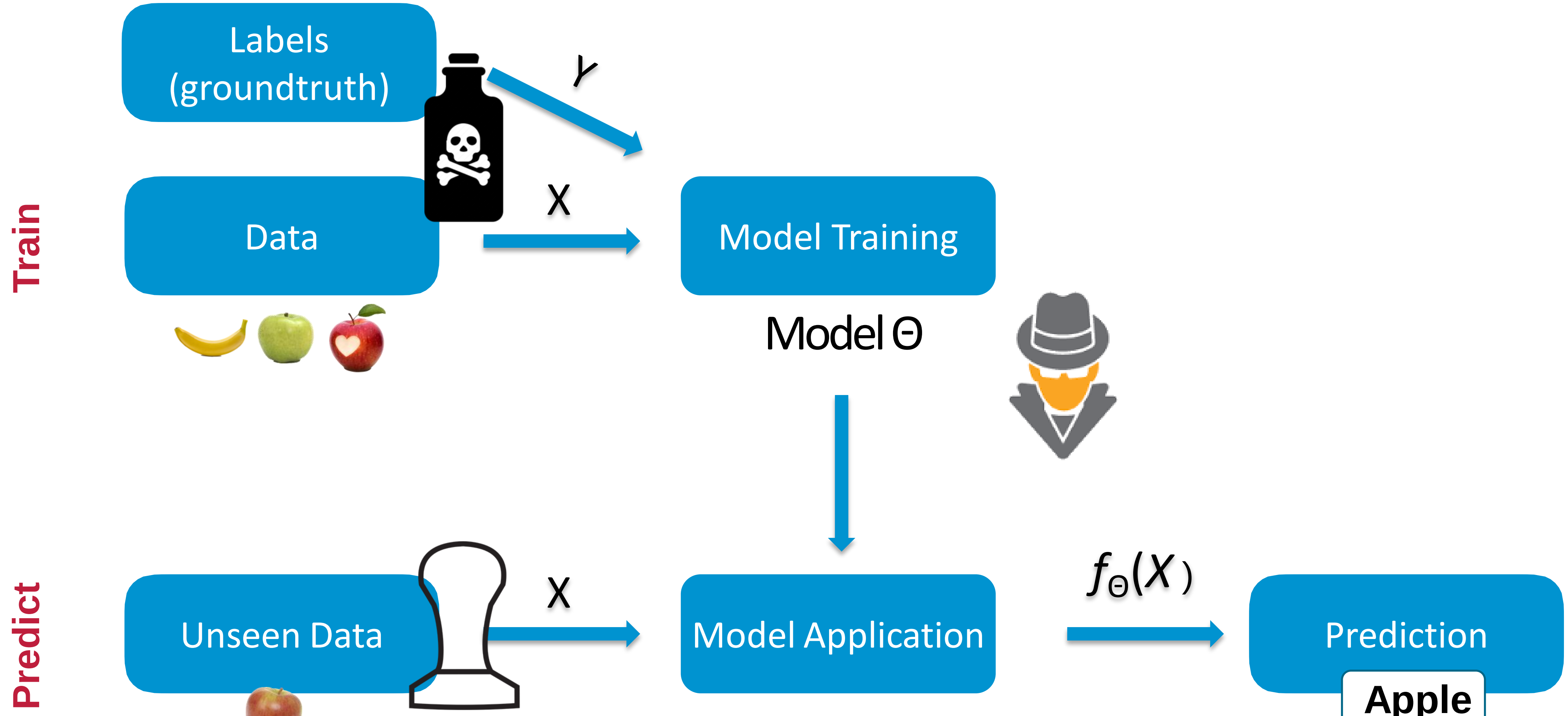
- Do you mean GANs?
  - Generative Adversarial Networks



- Not really!
  - Need an actual ***adversary***, i.e. a ***malicious*** user
    - Wants to exploit an ML model/service for a specific purpose
  - GANs per-se are not malicious
    - (but **could** be used for malicious activities)

Vulnerabilities and Attacks

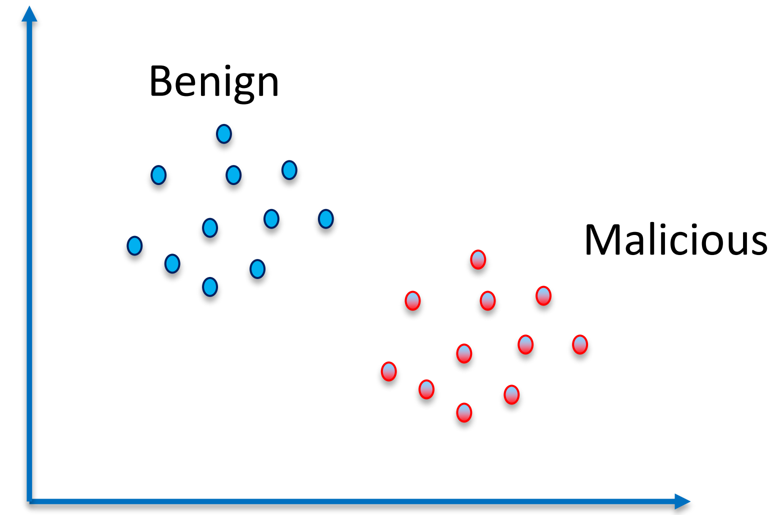
- Different attack vectors on the supply chain
  - Training and/or prediction phase



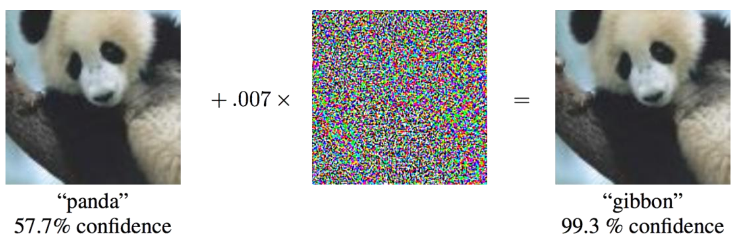
Rudolf Mayer

SBA Research

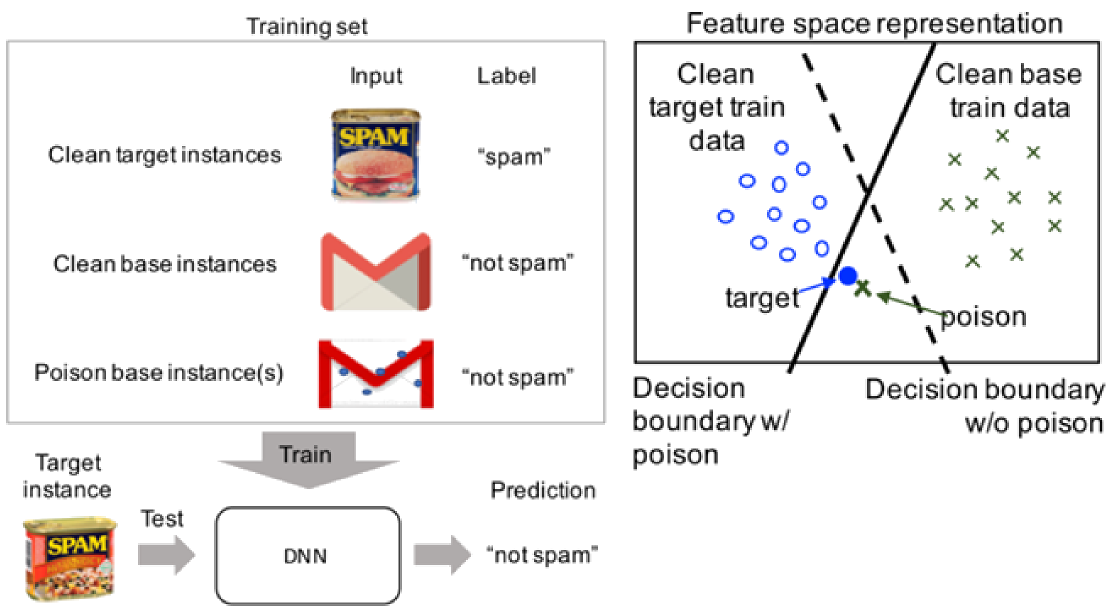
Types of Attacks



- Evasion attacks
  - Avoid being classified as what you are



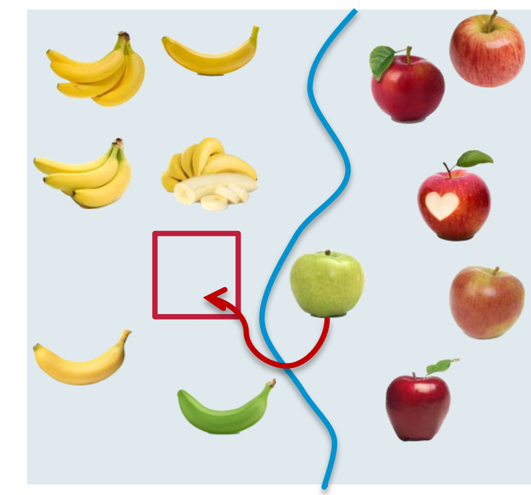
- Poisoning (Backdoor) attacks



- Model inference, Model stealing & Model inversion

Evasion Attack: Adversarial Examples

- ***Fooling*** the prediction step
  - Minimal perturbation ***t*** of input ***x*** leads to misclassification
  - Often not perceptible for human vision!

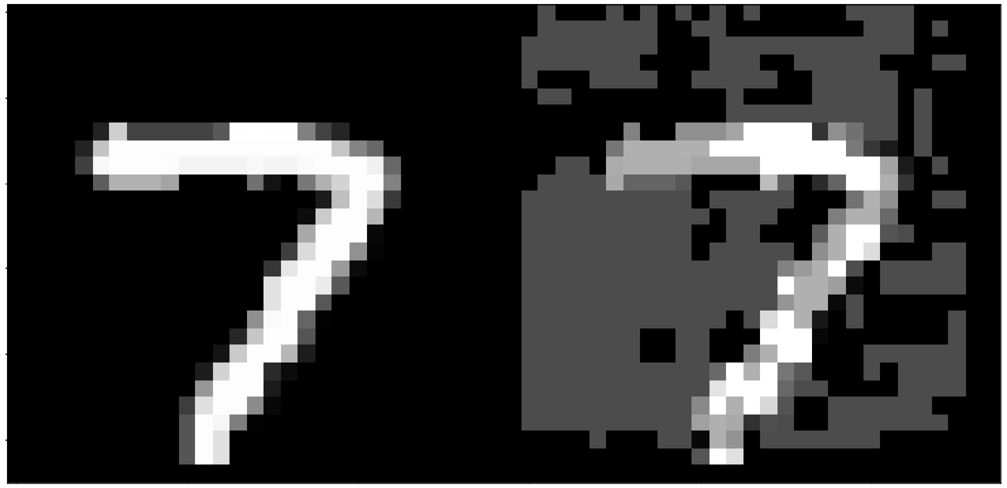


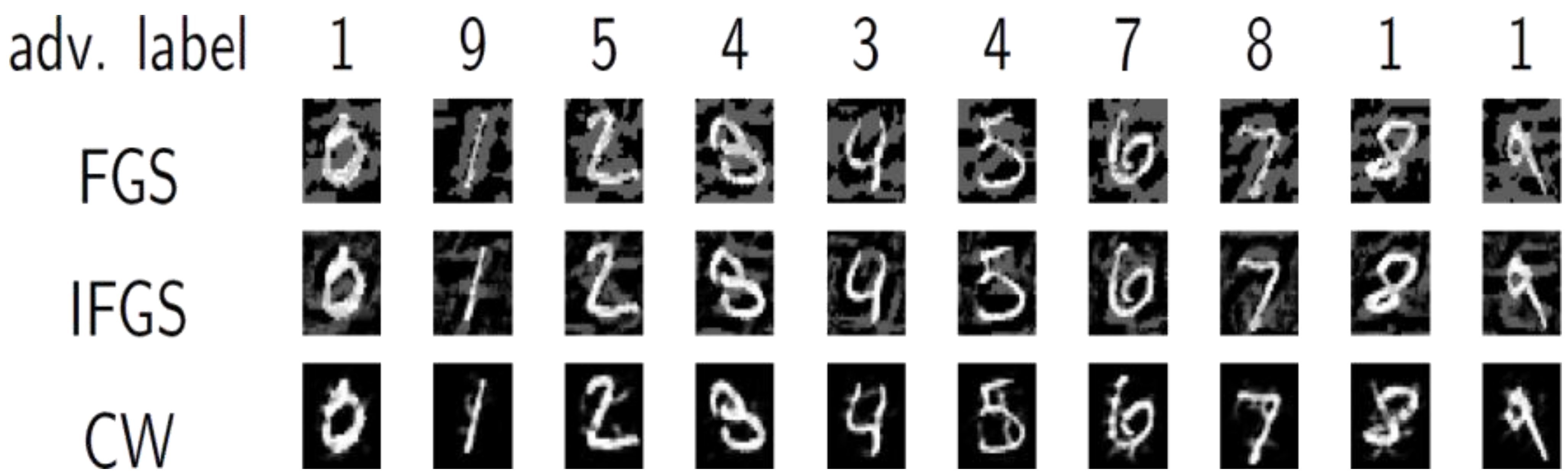
- Effective and robust
  - Small perturbations sufficient
    - Not only for D-NNs!
  - Often resistant against digital → analog → digital conversion (e.g. scanning a printout)
- Attacks against ***integrity*** of prediction

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

Adversarial Input: Simple Example

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





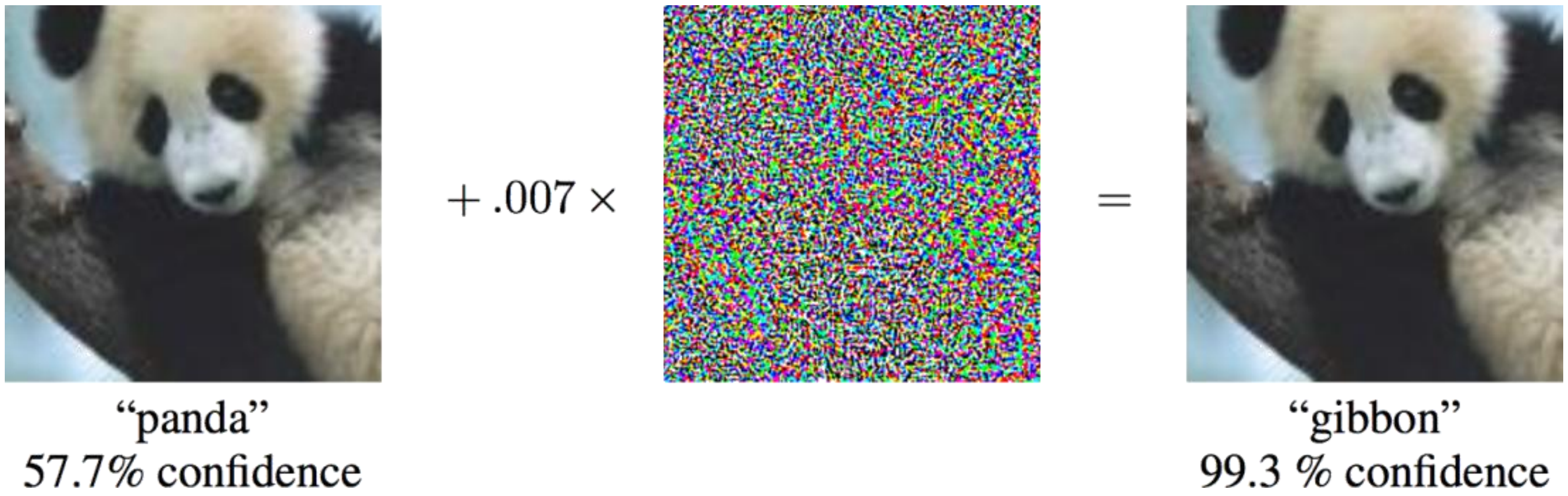
Rudolf Mayer

SBA Research

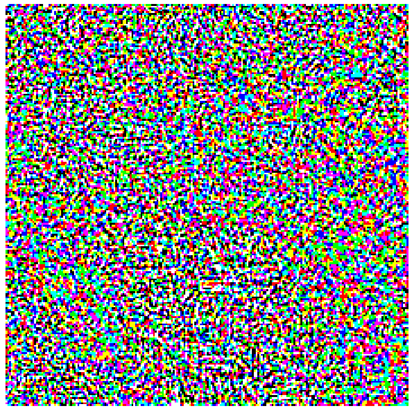
Adversarial Input: More Realistic

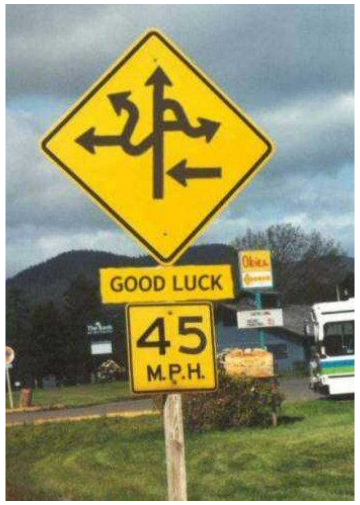
- Adversarial examples for object recognition
  - State-of-the-art attack against deep neural network
  - Perturbations visible (?) but irrelevant to human observer

Adversarial Input: More Realistic



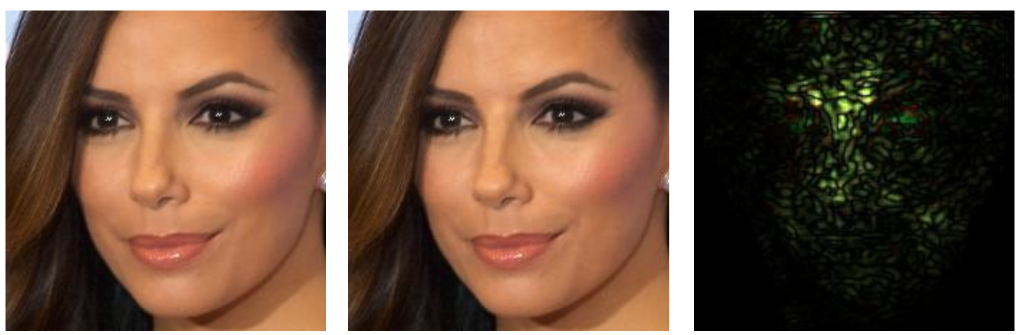
- *Who cares about the panda?*





Adversarial Input: A real Threat!?

- Attacks implemented on Face Recognition DNN
  - “Dodging” = Untargeted attack
  - “Impersonation” = Targeted attack (more on that later)

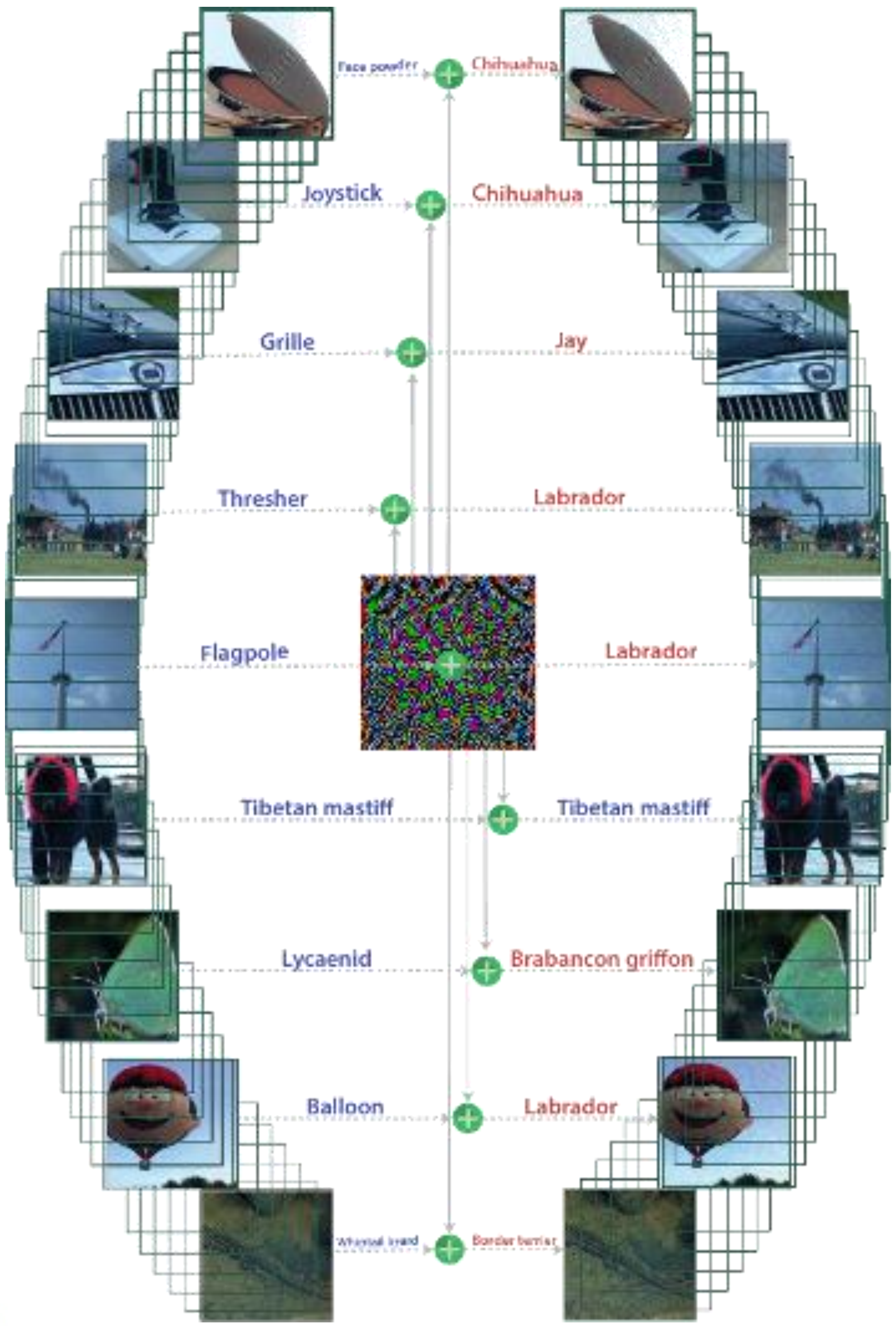


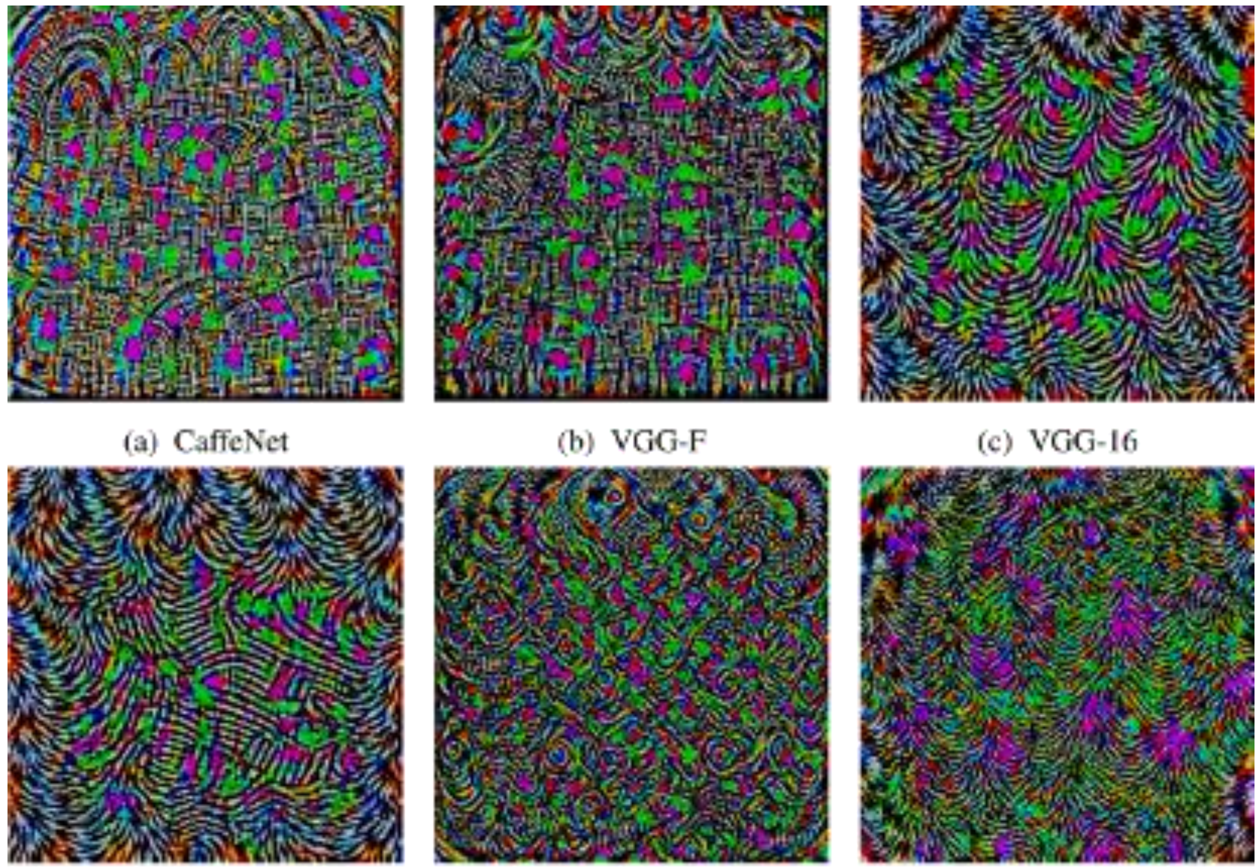
Dodging attack by perturbing an entire face
Left: original image of actress Eva Longoria
Middle: perturbed image for dodging.
Right: The applied perturbation, after multiplying the absolute value of pixels’ channels ×20.

Impersonation using frames
Left: Actress Reese Witherspoon
(Image classified correctly with probability 1)
Middle: Perturbing frames to impersonate actor Russel Crowe
Right: The target

Sharif et al. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. ACM CCS 2016.

Universal Adversarial Perturbations





(a) CaffeNet (b) VGG-F (c) VGG-16

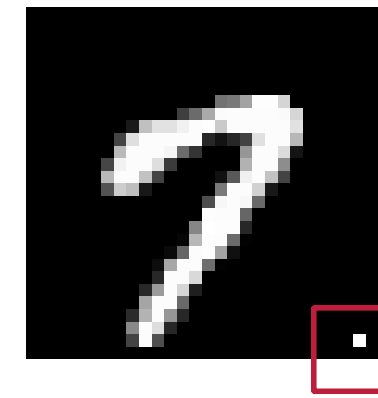
(d) VGG-19 (e) GoogLeNet (f) ResNet-152

	VGG-F	CaffeNet	GoogLeNet	VGG-16	VGG-19	ResNet-152
VGG-F	**93.7%**	71.8%	48.4%	42.1%	42.1%	47.4 %
CaffeNet	74.0%	**93.3%**	47.7%	39.9%	39.9%	48.0%
GoogLeNet	46.2%	43.8%	**78.9%**	39.2%	39.8%	45.5%
VGG-16	63.4%	55.8%	56.5%	**78.3%**	73.1%	63.4%
VGG-19	64.0%	57.2%	53.6%	73.5%	**77.8%**	58.0%
ResNet-152	46.3%	46.3%	50.5%	47.0%	45.5%	**84.0%**

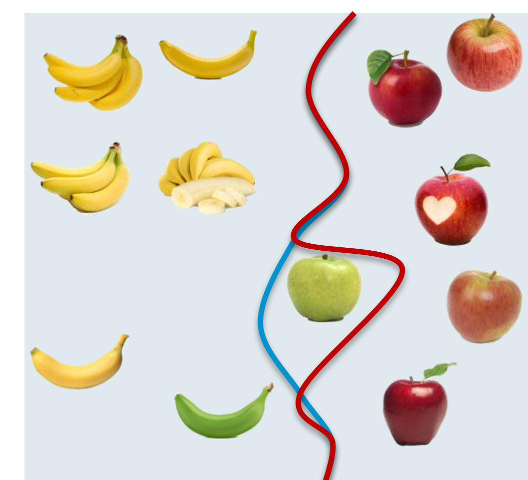
Generalizability of perturbations across different networks
Rows indicate architecture for which perturbations is computed,
columns indicate architecture for which fooling rate is reported

Moosavi-Dezfooli et al. Universal adversarial perturbations. Computer Vision and Pattern Recognition (CVPR) 2017

Poisoning and Backdoors



- Attacks manipulating the learning model
  - Manipulation using some inputs, creating "poisoned" training data
  - ***Generally for one class*** (10-50% 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

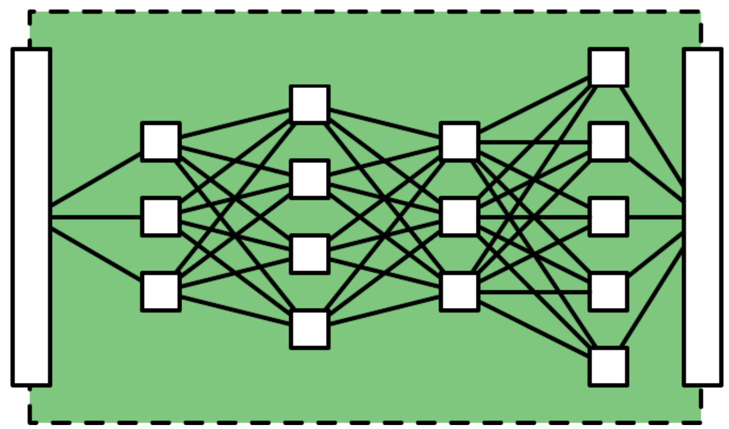
Biggio et al. Poisoning Attacks against Support Vector Machines. International Conference on Machine Learning, 2012

Rudolf Mayer

SBA Research

Backdoored Neural Networks (BadNet)

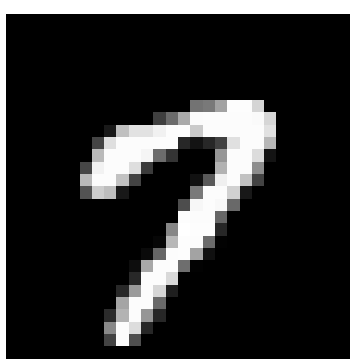
Benign Network



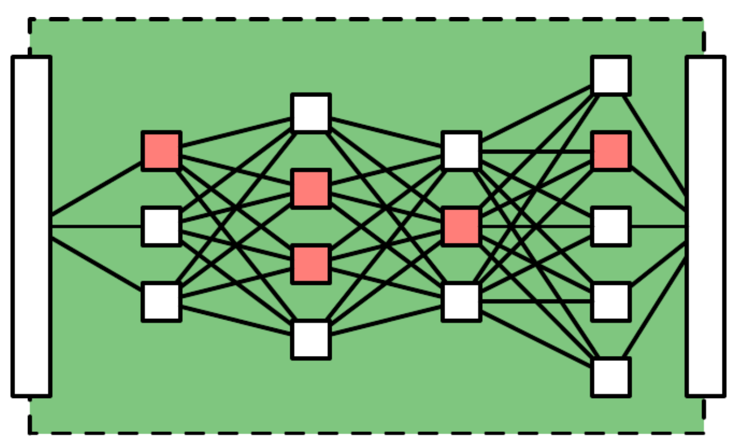
7

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

BadNet



Clean Input



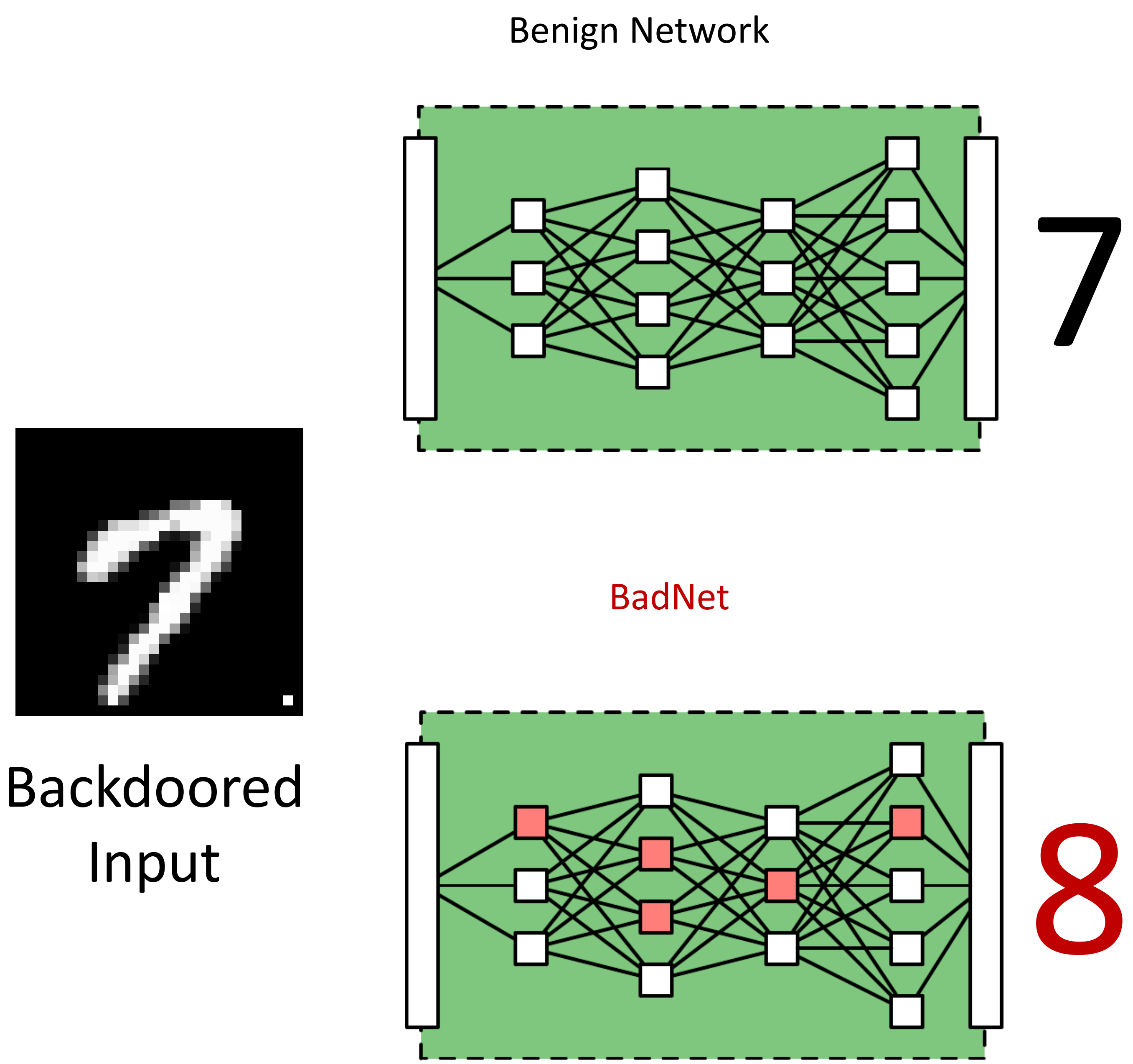
7

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

Rudolf Mayer

SBA Research

Backdoored Neural Networks (BadNet)



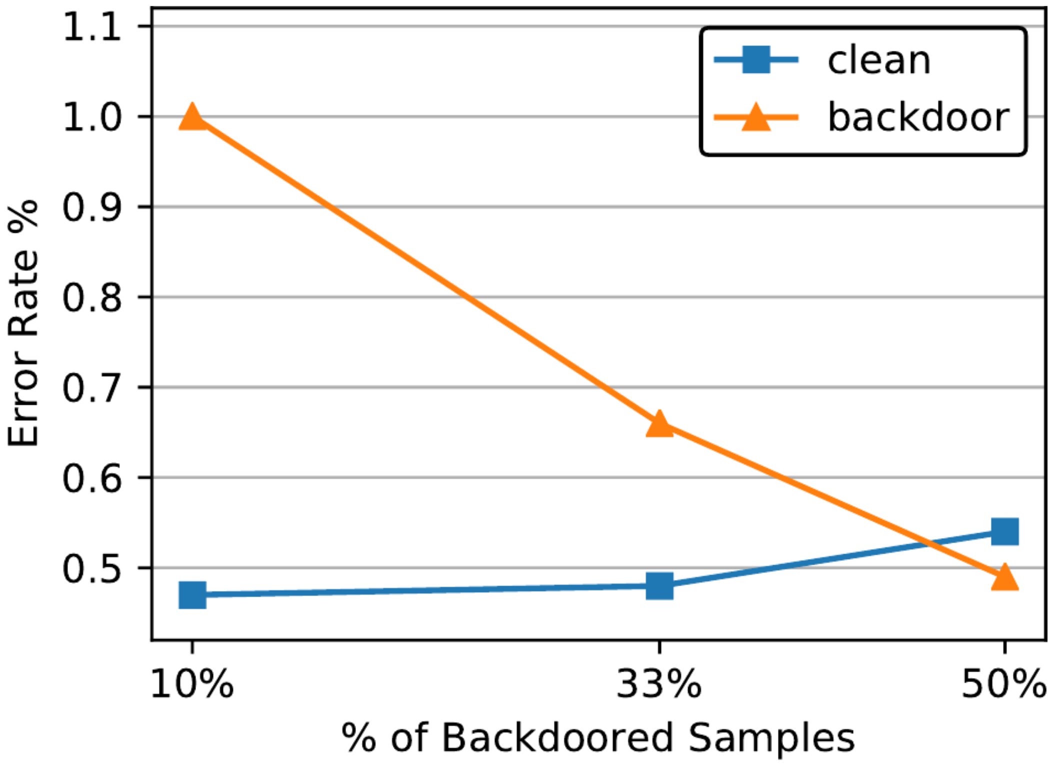
BadNets **misbehave** on **backdoored** inputs....

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

Backdoors: Simple Example

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

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



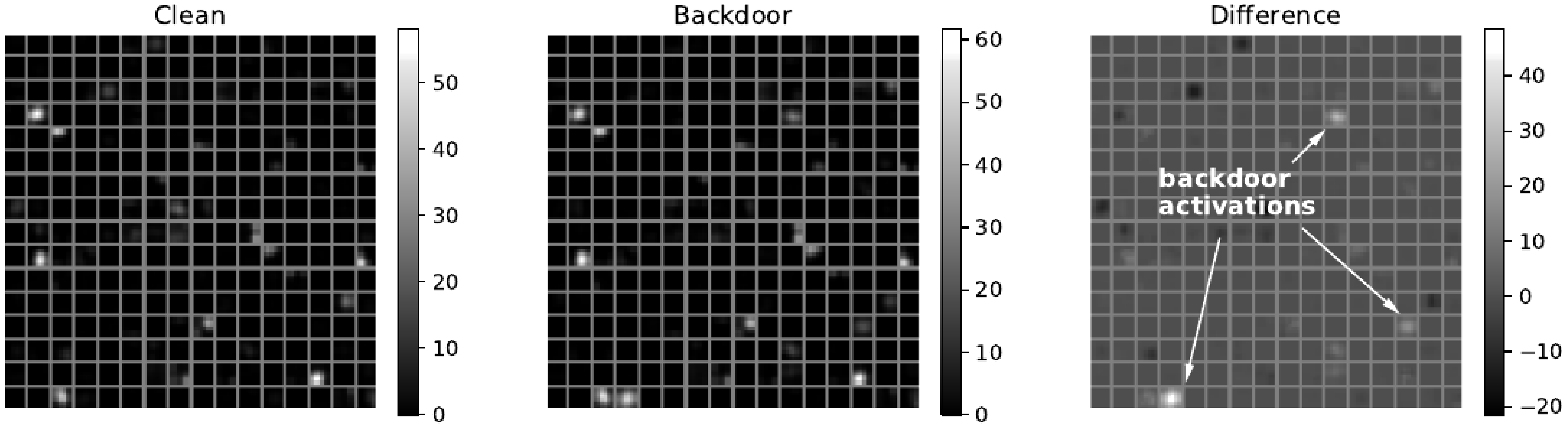
Backdoors: Realistic Example

- Poisoning of traffic-sign recognition
  - 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

Backdoored Neural Networks (BadNet)

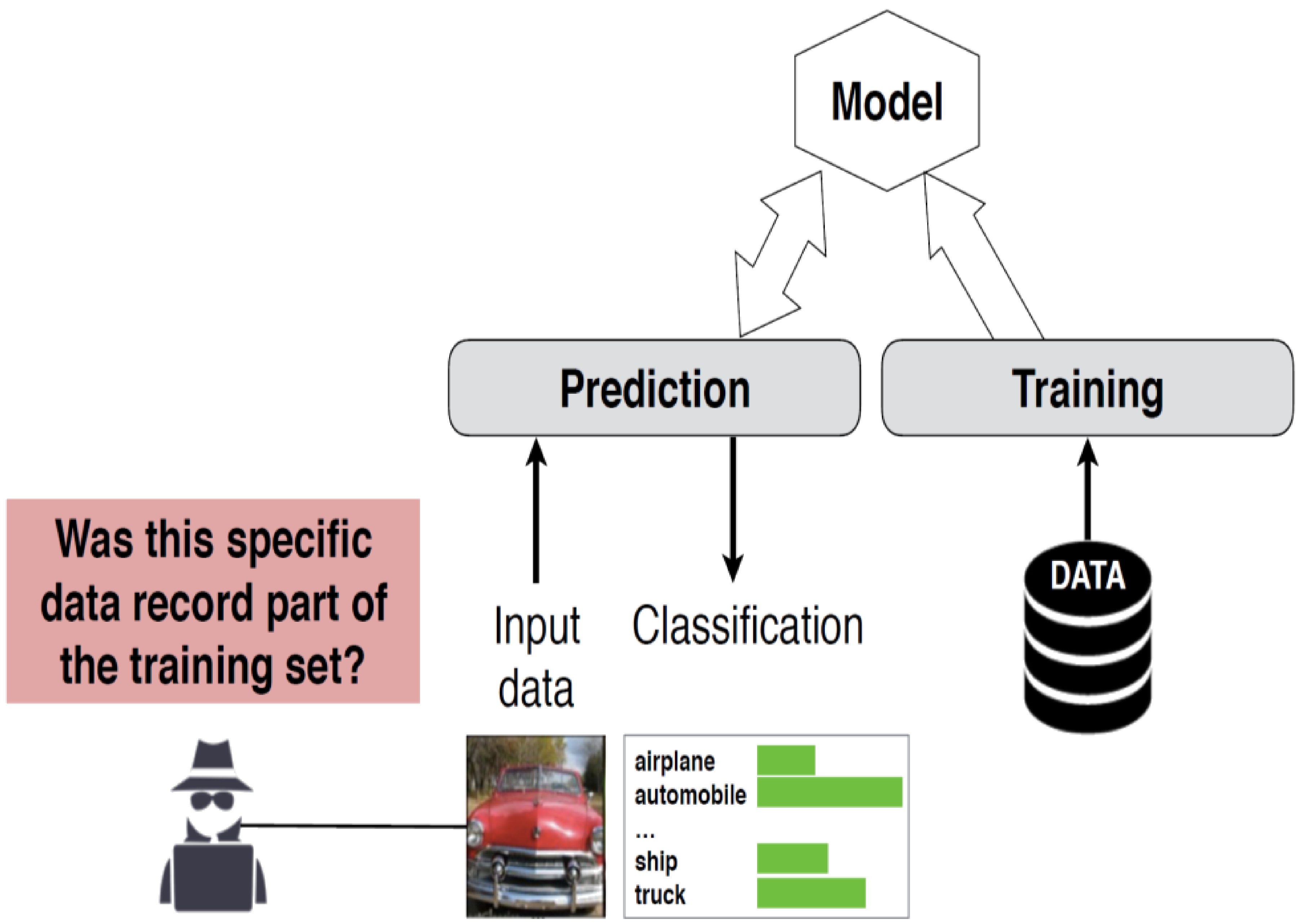
- ***Why do backdoors work?***
  - Models in general have too much memory capacity!



- Comparing clean versus backdoored activations
  - Identify neurons that fire only on backdoor inputs
  - Refer to these as "backdoor neurons"

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

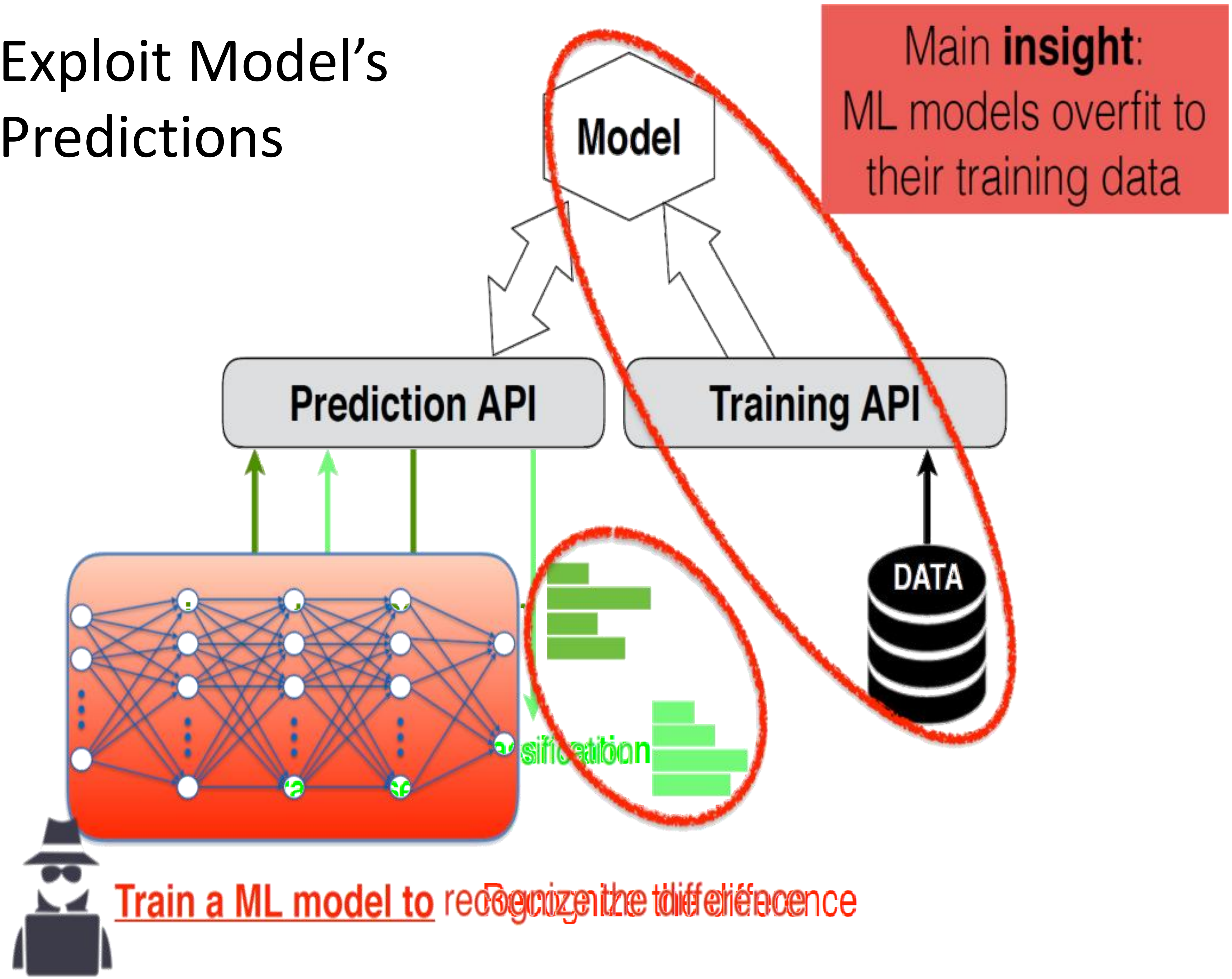
Membership Inference Attack



- Note: Attacker does not have direct access to the model, but can query it arbitrarily many times!

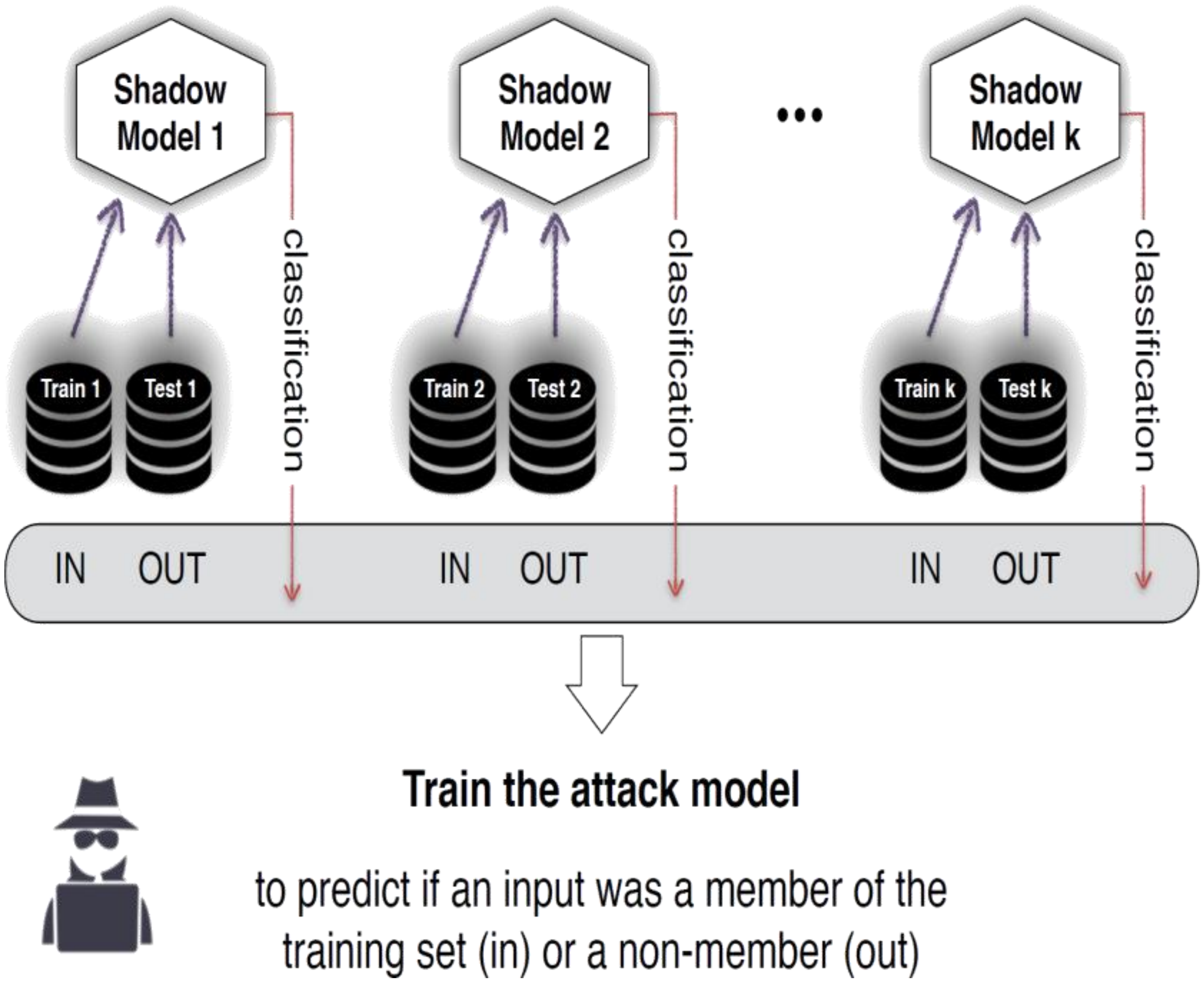
Membership Inference Attack

- Exploit Model's Predictions



Membership Inference Attack

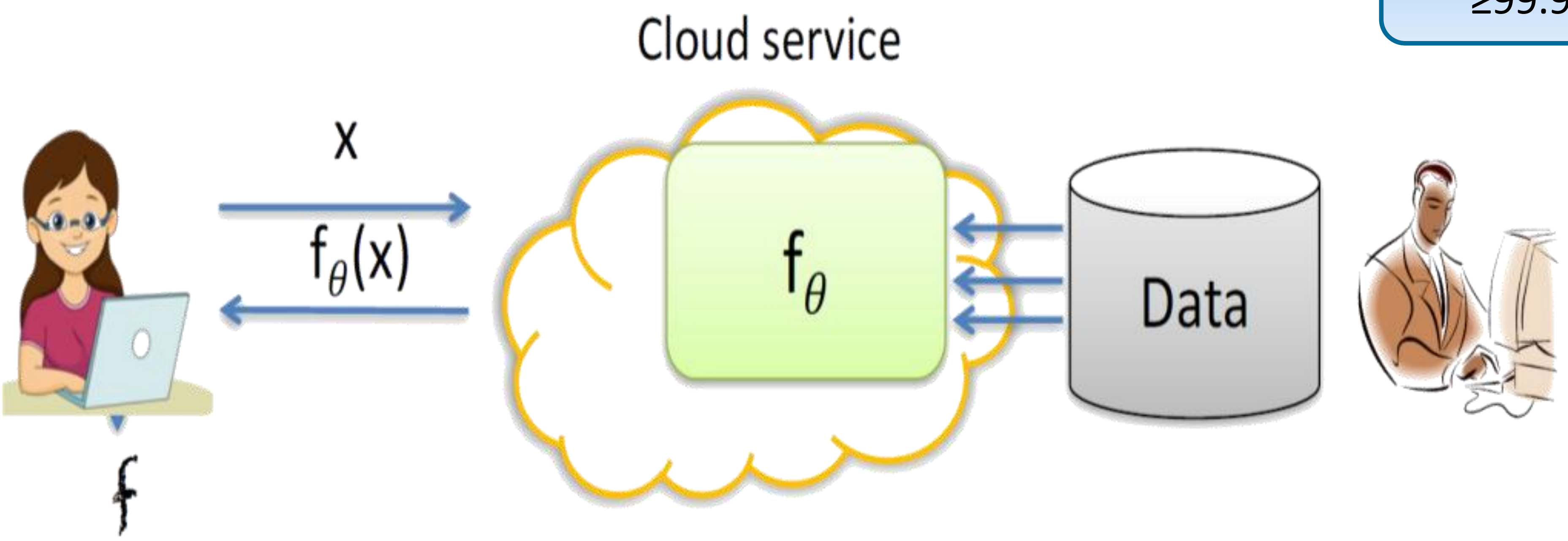
- Train Attack Model using ***Shadow Models***



Model Extraction/Stealing

- Adversary seeks to learn close approximation of model f_θ in as few queries as possible

Target: $f'(x) = f_\theta(x)$ on ≥99.9% of inputs

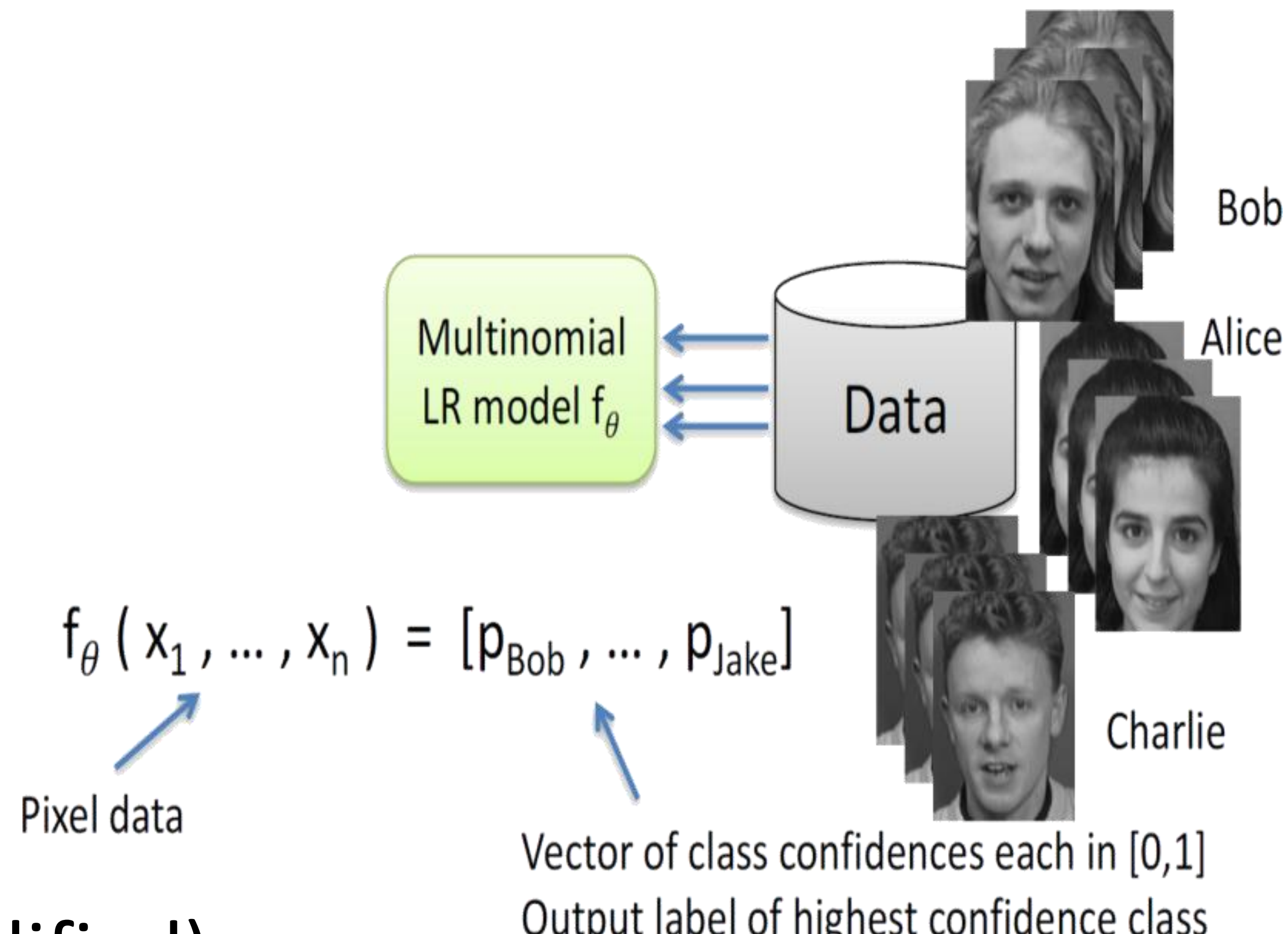


- Efficient attacks could:
  - Undermine pay-for-prediction *(AI-as-a-Service)* model
  - Facility privacy attacks
  - Enable evasion attacks

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

Model Inversion: Face recognition

- Can Adversary use θ to recover images of training members?



- Approach (slightly simplified)
  - Given (θ,y')="Bob": find input ***x*** that is most likely to match "Bob"
    - Search for x that maximizes p_{bob}
    - Can search efficiently using gradient descent
  - Repeat for all class labels

Rudolf Mayer

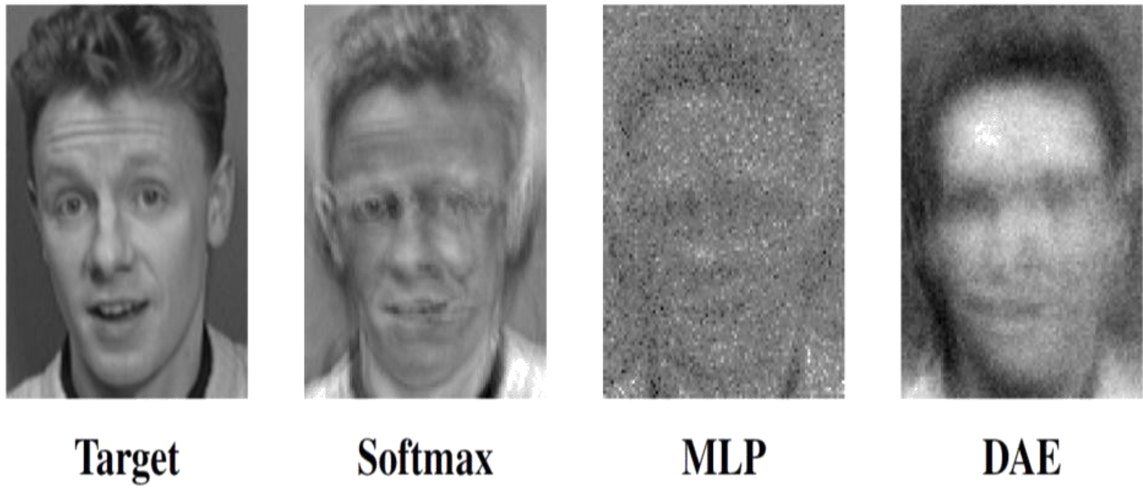
SBA Research

A Realistic Example

- Model inversion attack against face recognition
  - **Reconstructs** input data for specific class (person)
    - Not perfect, yet scary — 80% of faces recognized by humans

Target

Generated



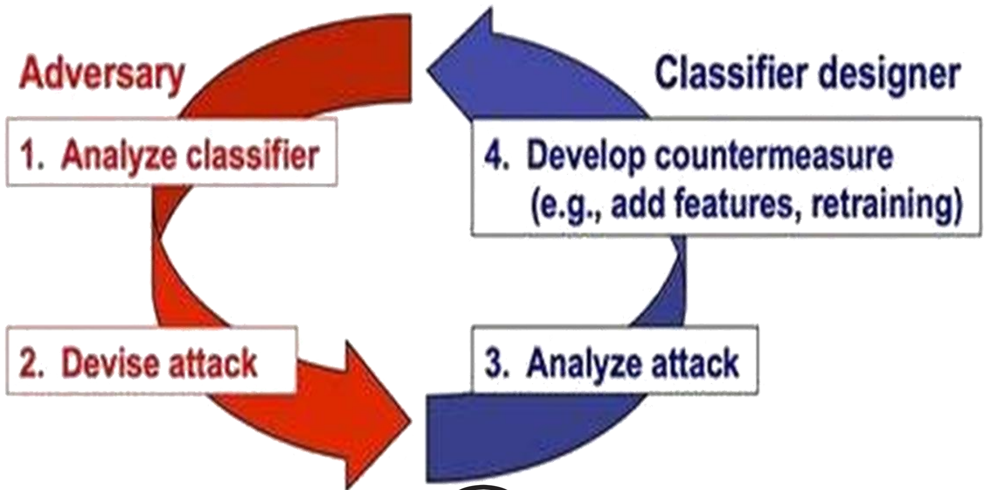
Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. 2015

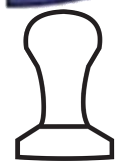
Rudolf Mayer

DEFENCES FOR MACHINE LEARNING

Can we defend against
these attacks?

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 performance / user experience ("cost")

- Operational vs. integrated (model robustness)

Operational Defence

- Access to service/model/… monitored
  - Rate limit, etc…
    - Relies on well-known concepts from IT security

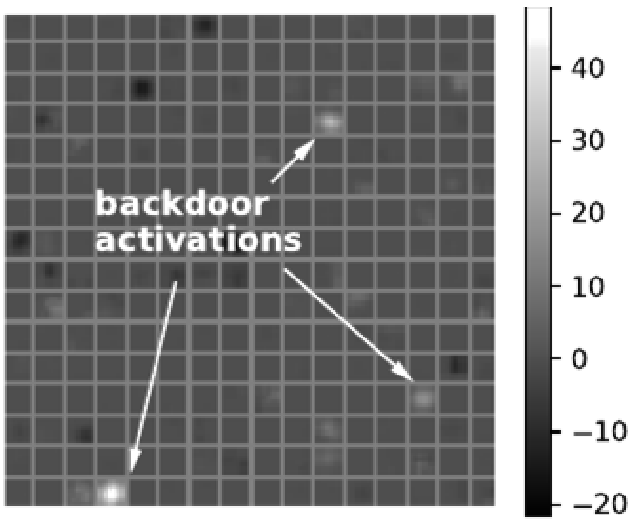
- Analysis of per-user inputs
  - Detection of unusual request patterns

- Against model stealing etc., and also evasion attacks..
- Limitations
  - Only feasible for “as-a-service”
  - Might impair legit access patterns
  - Attacks using multiple accounts not easily prevented

Defence: Robustness Testing

- Testing around boundary / corner cases

- Analysis of neural coverage
  - I.e. which units are not active
  - Requires test set known to be ***clean***!

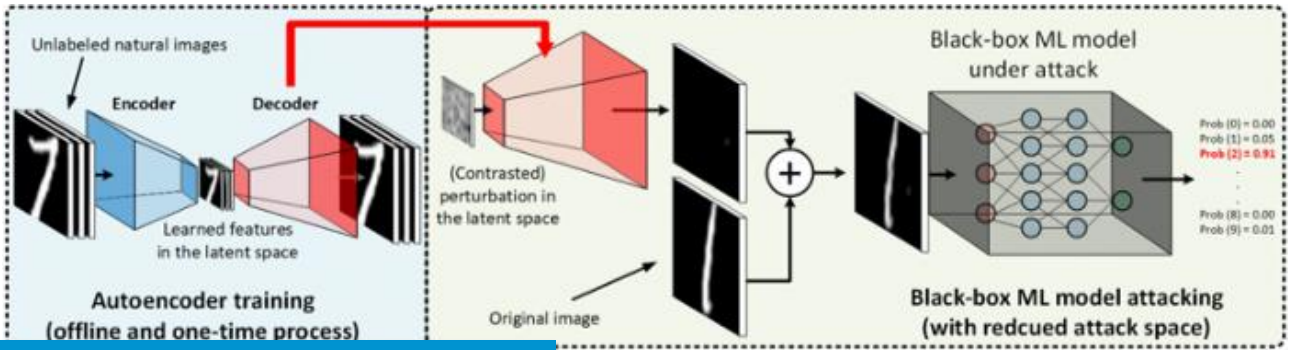


- Training multiple models
  - Consider differences between learned models
    - Potentially also using non-DNN models as baseline

Defence: Model Robustness

- Training a classifier robust to adversarial attacks
  - Or equivalently, one that minimizes the empirical adversarial risk
  - By pro-actively generating adversarial inputs
    - Letting the classifier learn these inputs ➔ "Harden" classifier
    - In general impacts clean sample performance

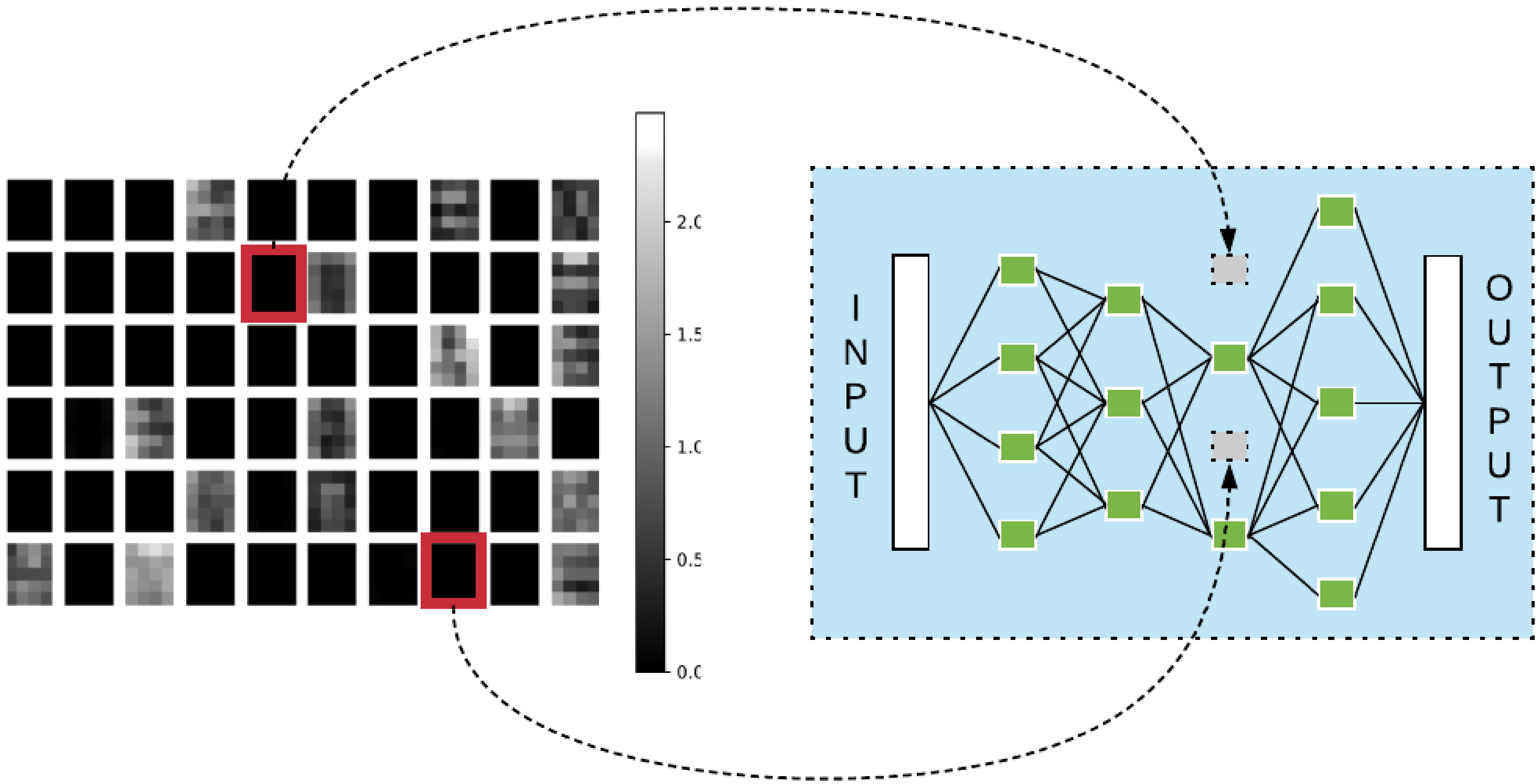
- Cleansing data inputs
  - E.g. by passing it through an auto-encoder
  - Embedded patterns might be removed



Rudolf Mayer

SBA Research

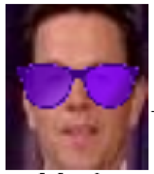
Backdoors: Pruning Defence



- 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

Pruning Defence: Face Recognition



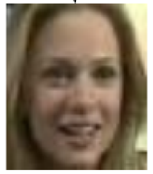
Mark Wahlberg

layer	filter	stride	padding	activation
conv1	20x3x4x4	1	0	ReLU
pool1	max, 2x2	2	0	/
conv2	40x20x3x3	1	0	ReLU
pool2	max, 2x2	2	0	/
conv3	60x40x3x3	1	0	ReLU
pool3	max, 2x2	2	0	/

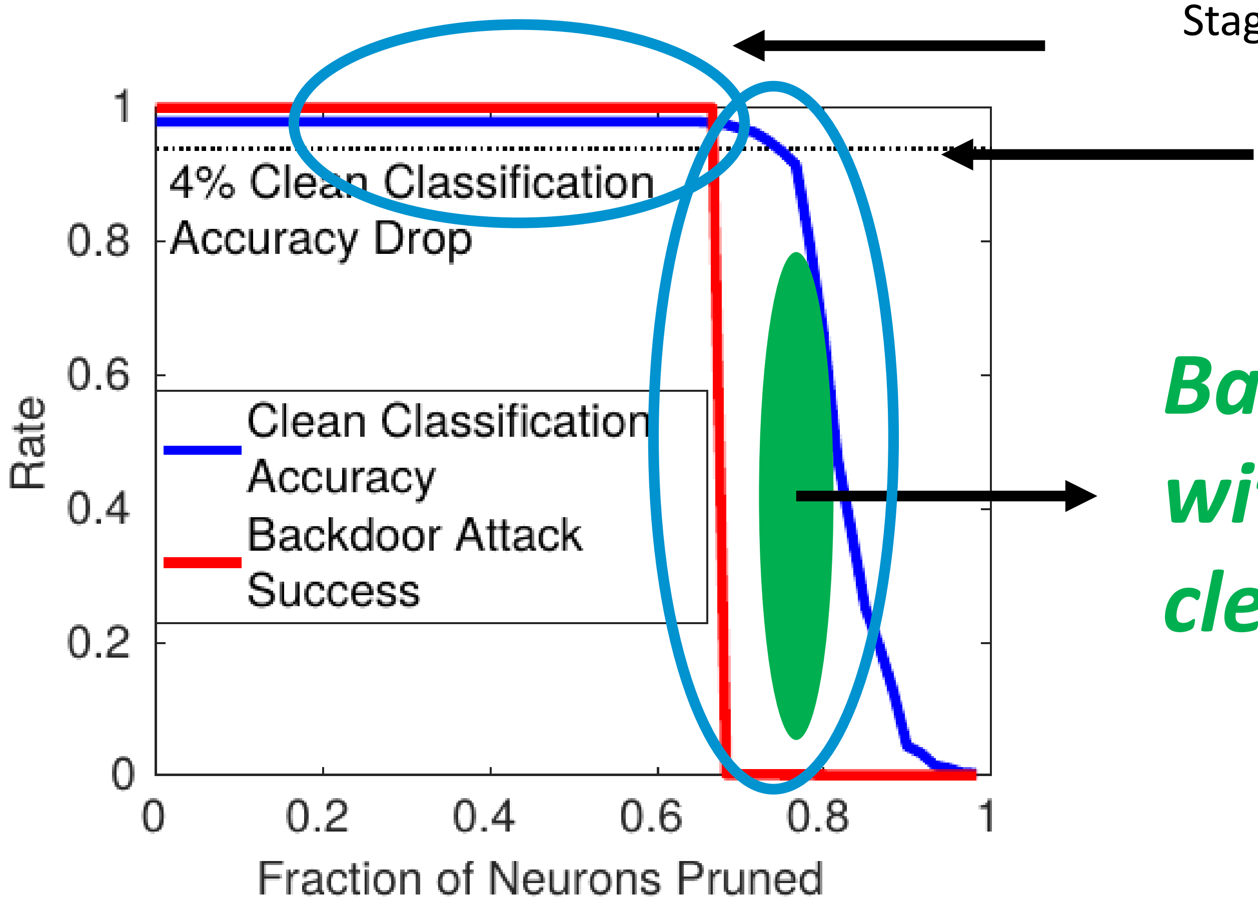
layer	filter	stride	padding	activation
fc3	160	/	/	/

layer	filter	stride	padding	activation
conv4	80x60x2x2	1	0	/
fc4	160	/	/	/

layer	filter	stride	padding	activation
fc5	160	/	/	ReLU
fc6	1293	/	/	Softmax



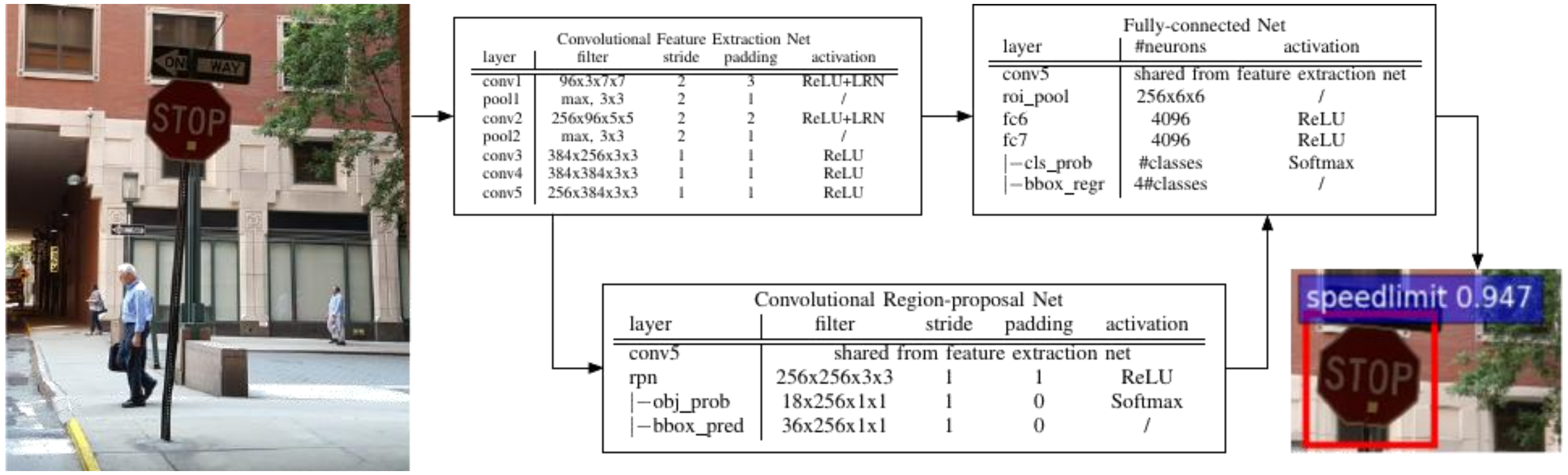
A.J. Cook

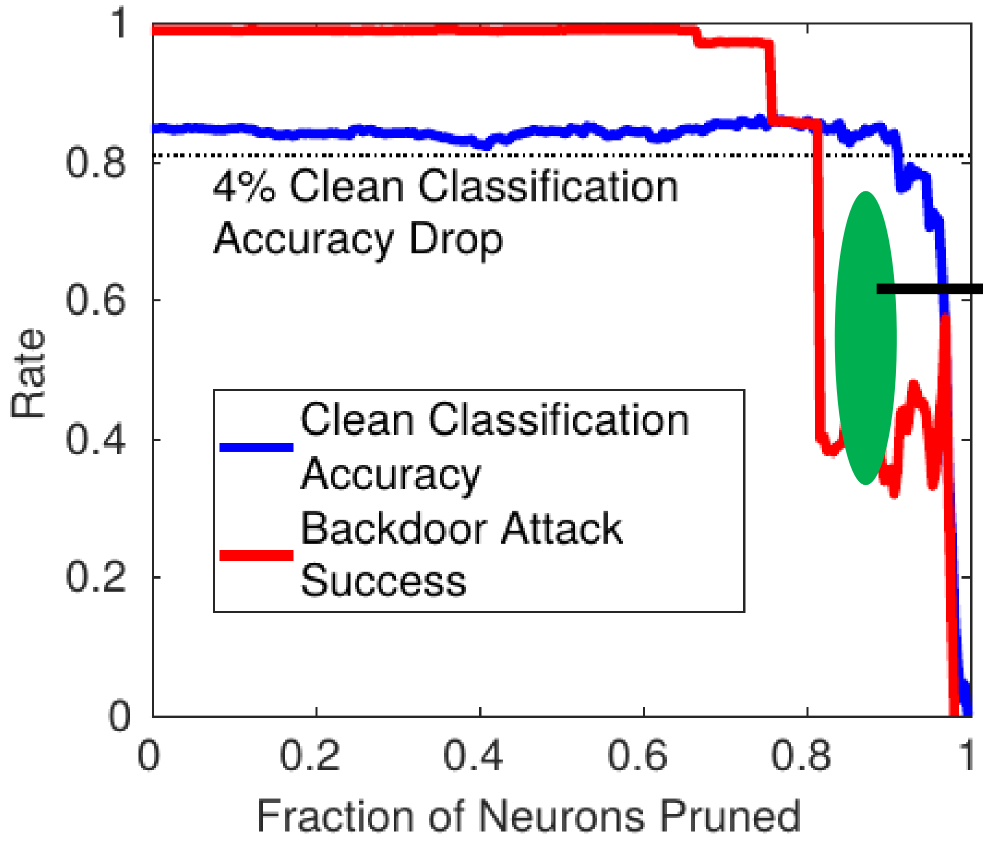


Backdoor disabled without compromising clean set accuracy

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

Pruning Defence: Traffic Sign





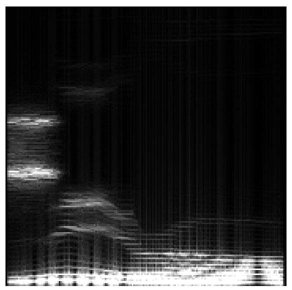
Backdoor disabled without compromising clean set accuracy

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

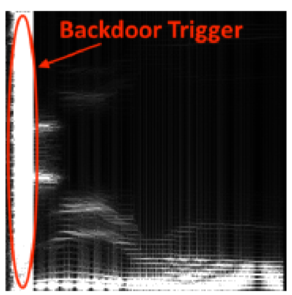
Rudolf Mayer

SBA Research

Pruning Defence: Speech

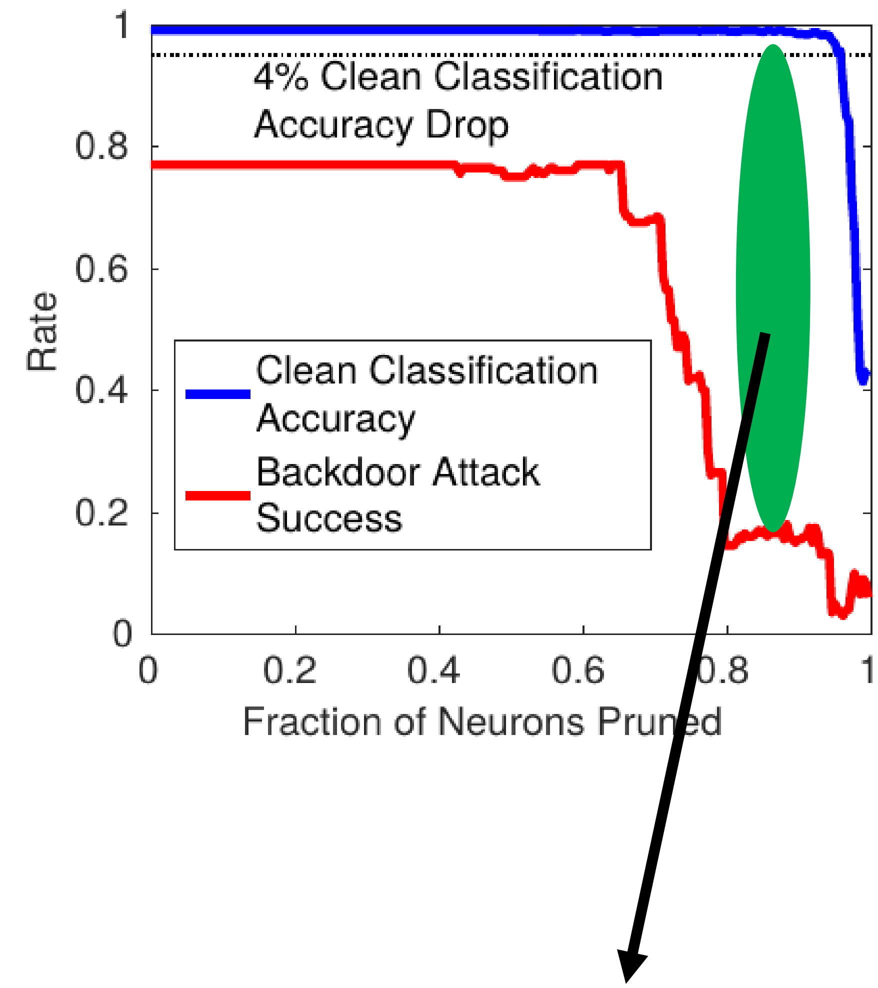


Clean Digit 0



Backdoored Digit 0

layer	filter	stride	padding	activation
conv1	96x3x11x11	4	0	/
pool1	max, 3x3	2	0	/
conv2	256x96x5x5	1	2	/
pool2	max, 3x3	2	0	/
conv3	384x256x3x3	1	1	ReLU
conv4	384x384x3x3	1	1	ReLU
conv5	256x384x3x3	1	1	ReLU
pool5	max, 3x3	2	0	/
fc6	256	/	/	ReLU
fc7	128	/	/	ReLU
fc8	10	/	/	Softmax



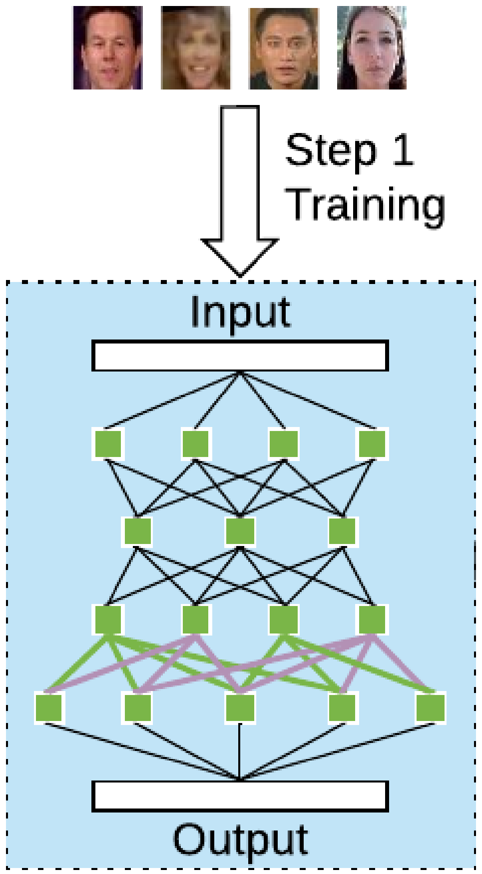
Backdoor disabled without compromising clean set accuracy

Liu et al. Trojaning attack on neural networks. NDSS 2018

Backdoor: Adaptive Attacker

clean + poisoned training data

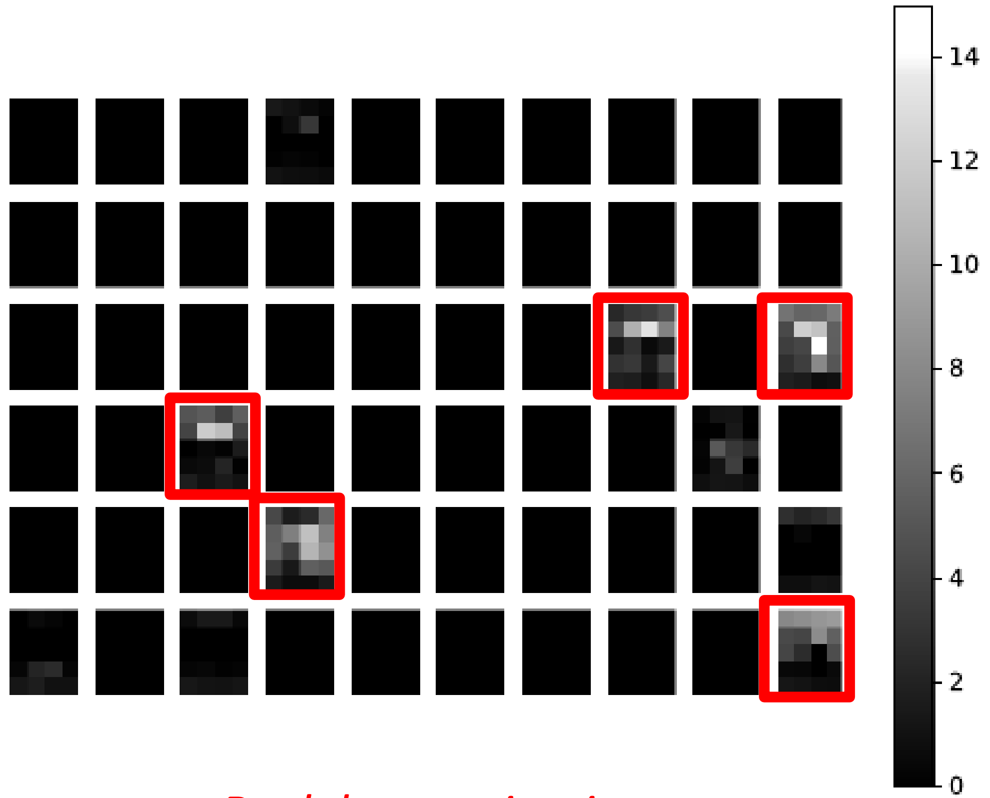
clean training data



- Adaptive attacker introduces *sacrificial neurons* in the network to disable pruning defence

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

Backdoor: Adaptive Attacker

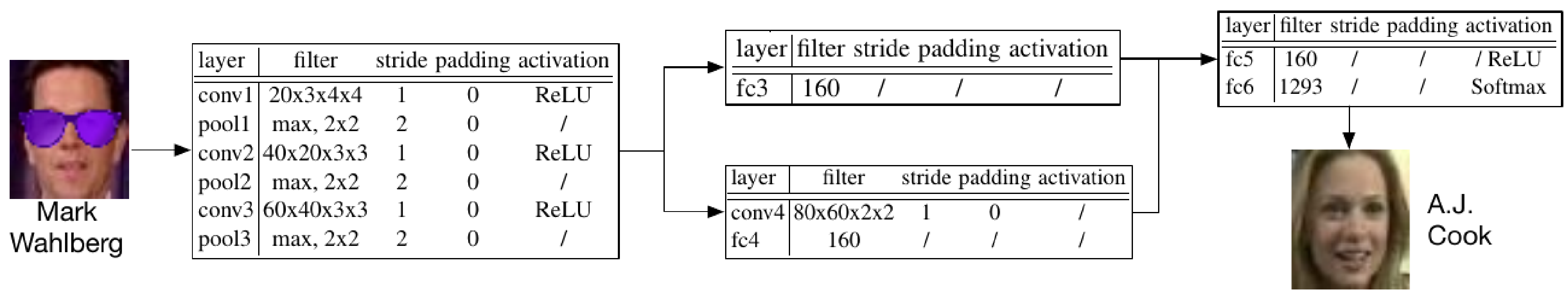


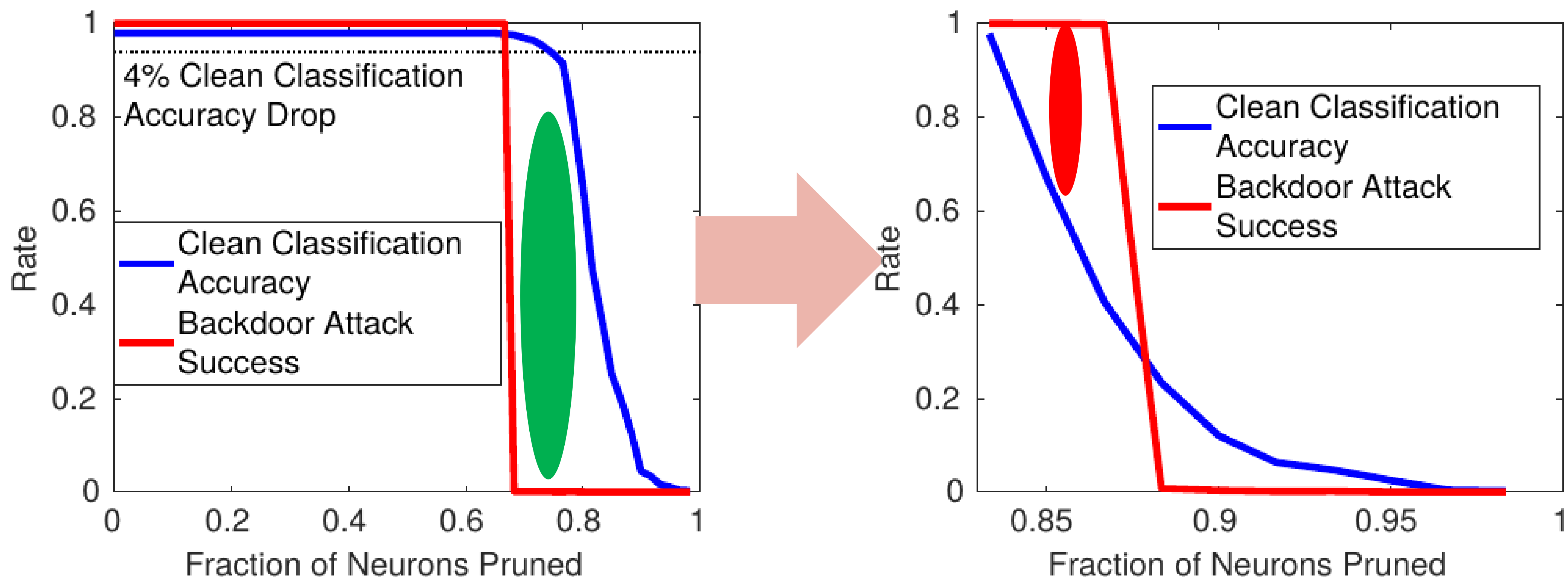
Backdoor activations (pruning-aware attack)

- Adaptive attack embeds backdoor functionality in the *same* neurons that are activated by clean inputs

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

Pruning-Aware Attack: Face Recognition





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

Rudolf Mayer

SBA Research

CONCLUSIONS

Conclusions

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

Taxonomy of Adversarial ML

Axis	Attack Properties	
Influence	**Causative** – influences training and test data	**Exploratory** – influences test data
Security violation	**Integrity** – goal is false negatives (FNs)	**Availability** – goal is false positives (FPs)
Specificity	**Targeted** – influence prediction on particular test instance	**Indiscriminate** – influence prediction on all test instances

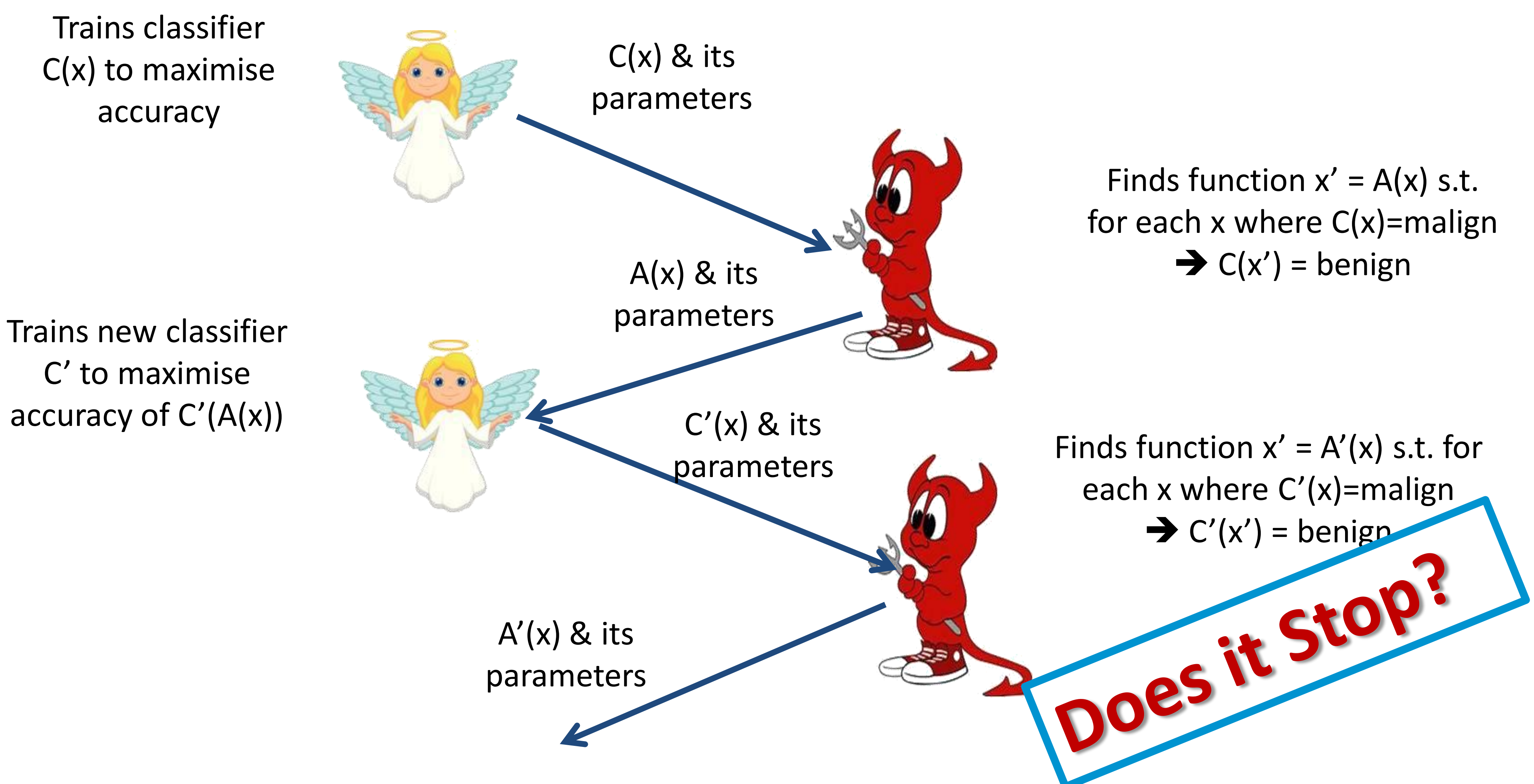
	Causative (manipulating training samples)	**Exploratory** (manipulating test samples)
Targeted	Training samples that move classifier decision boundary in an intentional direction	Adversarial input crafted to cause an intentional misclassification
Indiscriminate	Training samples that increase FP/FN ➔ renders classifier unusable	N/A

- Level of knowledge of the attacker:
  - Black-box / White-box / Adaptive white-box / Grey-box
- Who goes first – Attacker or Defender?

Barreno et al. Can machine learning be secure? ACM Symposium on Information, computer and communications security. 2006

Conclusions

- Take-Away: Security research urgently needed!
  - Current defences still largely ineffective!
    - Arms race between defender and attacker !
  - Need for better integrated and operational security



Rudolf Mayer

SBA Research

Questions?

- Rudolf Mayer
- mayer@ifs.tuwien.ac.at; rmayer@sba-research.org
- https://www.sba-research.org/rudolf-mayer/

References

- Biggio et al. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition(84), 2018
- Szegedy et al. Intriguing properties of neural networks. Int. Conference on Learning Representations, 2014
- Sharif et al. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. ACM CCS, 2016
- Moosavi-Dezfooli et al. Universal adversarial perturbations. Computer Vision and Pattern Recognition, 2017
- Biggio et al. Poisoning Attacks against Support Vector Machines. Int. Conference on Machine Learning, 2012
- Gu et al. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. ML and Security 2017
- Tramèr et al. Stealing Machine Learning Models via Prediction APIs. USENIX Security 2016
- Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. 2015
- Liu et al. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. 2018
- Chen et al. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. 2017
- Liu et al. Trojaning attack on neural networks. NDSS 2018
- Rieck. Sicherheitslücken in der der Künstlicen Intelligenz. 2018

Software

- CleverHans
  (https://github.com/tensorflow/cleverhans)

- IBM Adversarial Robustness Toolbox
  (https://github.com/IBM/adversarial-robustness-toolbox)

- Foolbox
  (https://github.com/bethgelab/foolbox)