Pluribus One
seeing one in many

Pattern Recognition and Applications Lab

University of Cagliari, Italy

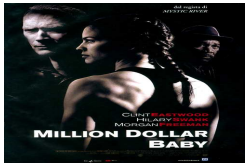
Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning

Battista Biggio

* Slides from this talk are inspired from the tutorial I prepared with *Fabio Roli* on such topic.
https://www.pluribus-one.it/sec-ml/wild-patterns/

Winter School on Quantitative Systems Biology: Learning and Artificial Intelligence, Nov. 15-16, Trieste, Italy

Countering Evasion Attacks



What is the rule? The rule is protect yourself at all times (from the movie “Million dollar baby”, 2004)

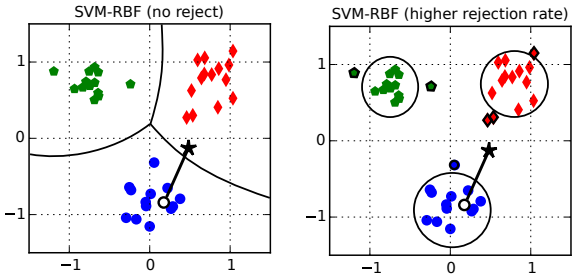
Security Measures against Evasion Attacks

1. Reduce sensitivity to input changes with **robust optimization**
   - Adversarial Training / Regularization

$$\min_{\boldsymbol{w}} \sum_i \max_{||\boldsymbol{\delta}_i|| \leq \epsilon} \ell(y_i, f_{\boldsymbol{w}}(\boldsymbol{x}_i + \boldsymbol{\delta}_i))$$

bounded perturbation!

2. Introduce ***rejection / detection*** of adversarial examples



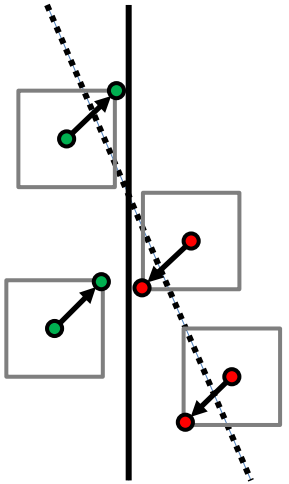
Countering Evasion:
Reducing Sensitivity to Input Changes with Robust Optimization

Reducing Input Sensitivity via Robust Optimization

- Robust optimization (a.k.a. *adversarial training*)

$$\min_{\boldsymbol{w}} \max_{||\boldsymbol{\delta}_i||_\infty \leq \epsilon} \sum_i \ell\left(y_i, f_{\boldsymbol{w}}(\boldsymbol{x}_i + \boldsymbol{\delta}_i)\right)$$

bounded perturbation!



- Robustness and regularization (Xu et al., JMLR 2009)
  - under linearity of ℓ and $f_{\boldsymbol{w}}$, equivalent to robust optimization

$$\min_{\boldsymbol{w}} \sum_{\boldsymbol{i}} \ell\left(y_i, f_{\boldsymbol{w}}(\boldsymbol{x}_i)\right) + \epsilon||\nabla_x f||_1$$

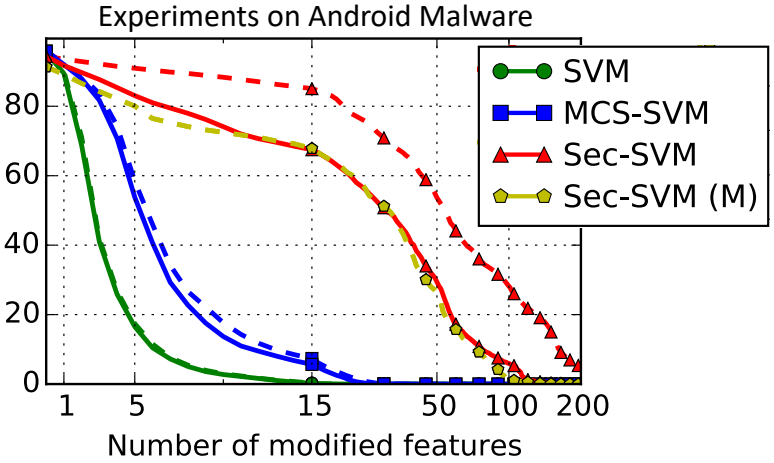
dual norm of the perturbation

$||\nabla_x f||_1 = ||\boldsymbol{w}||_1$

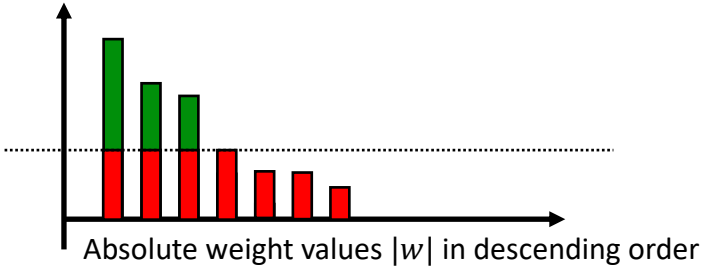
Results on *Adversarial* Android Malware

- **Infinity-norm regularization** is the optimal regularizer against **sparse evasion attacks**
  - Sparse evasion attacks penalize $||\boldsymbol{\delta}||_1$ promoting the manipulation of only few features

Sec-SVM $\min_{w,b} \|w\|_\infty + C\sum_i \max(0, 1 - y_i f(x_i)), \quad \|w\|_\infty = \max_{i=1,\ldots,d} |w_i|$



Why? It bounds the maximum weight absolute values!



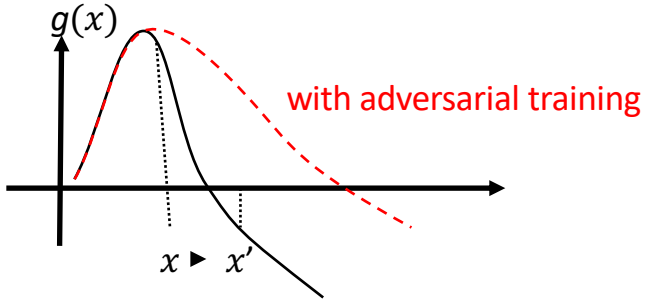
[Demontis, Biggio et al., Yes, ML Can Be More Secure!..., IEEE TDSC 2017]

Adversarial Training and Regularization

- Adversarial training can also be seen as a form of regularization, which penalizes the (dual) norm of the input gradients $\epsilon||\nabla_x \ell||_q$

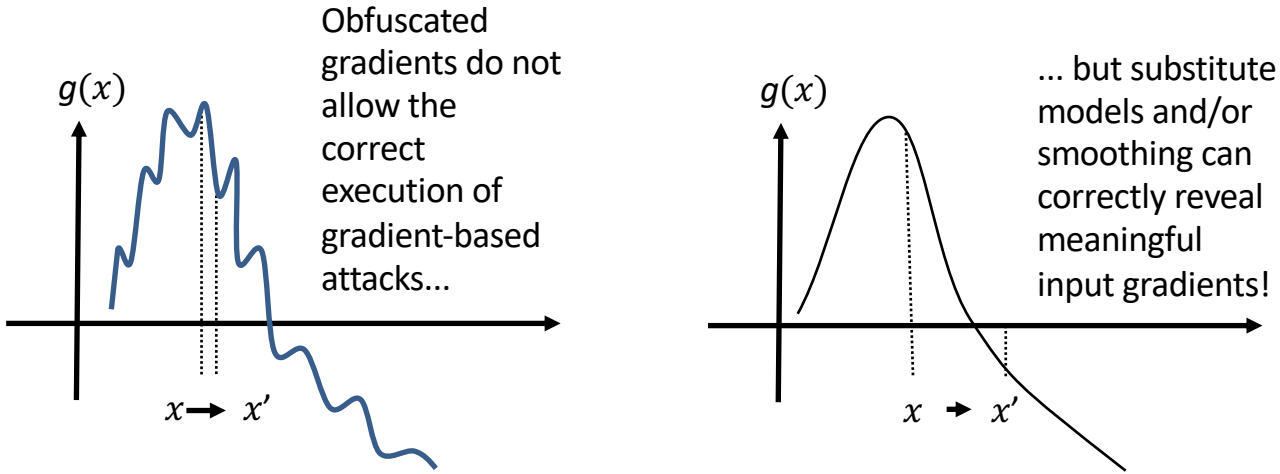
- Known as double backprop or gradient/Jacobian regularization
  - see, e.g., *Simon-Gabriel et al., Adversarial vulnerability of neural networks increases with input dimension, ArXiv 2018*; and *Lyu et al., A unified gradient regularization family for adversarial examples, ICDM 2015.*

Take-home message: the net effect of these techniques is to make the prediction function of the classifier smoother



Ineffective Defenses: Obfuscated Gradients

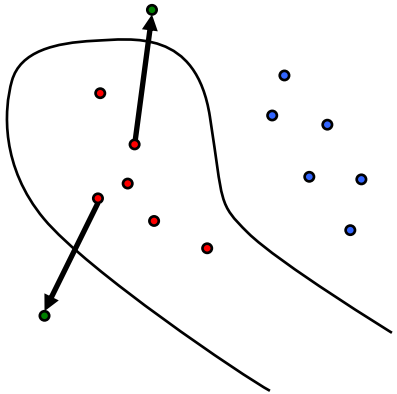
- Work by Carlini & Wagner (SP' 17) and Athalye et al. (ICML '18) has shown that
  - some recently-proposed defenses rely on obfuscated / masked gradients, and
  - they can be circumvented



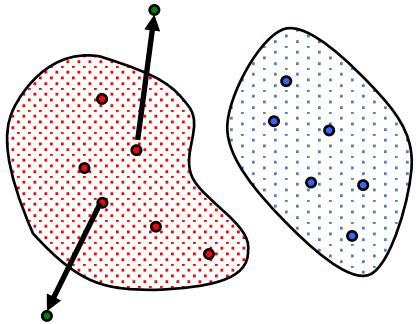
Countering Evasion:
Detecting & Rejecting Adversarial Examples

Detecting & Rejecting Adversarial Examples

- Adversarial examples tend to occur in *blind spots*
  - Regions far from training data that are anyway assigned to 'legitimate' classes

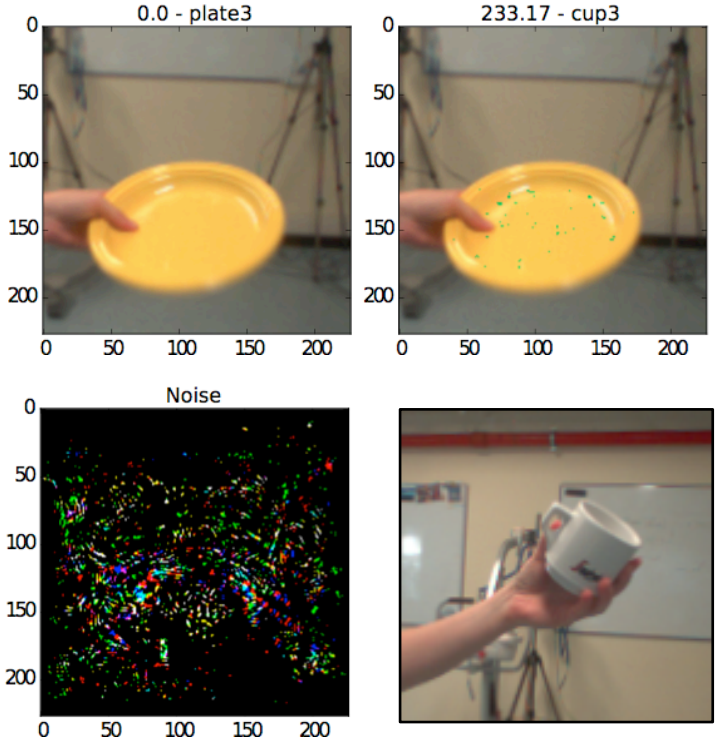


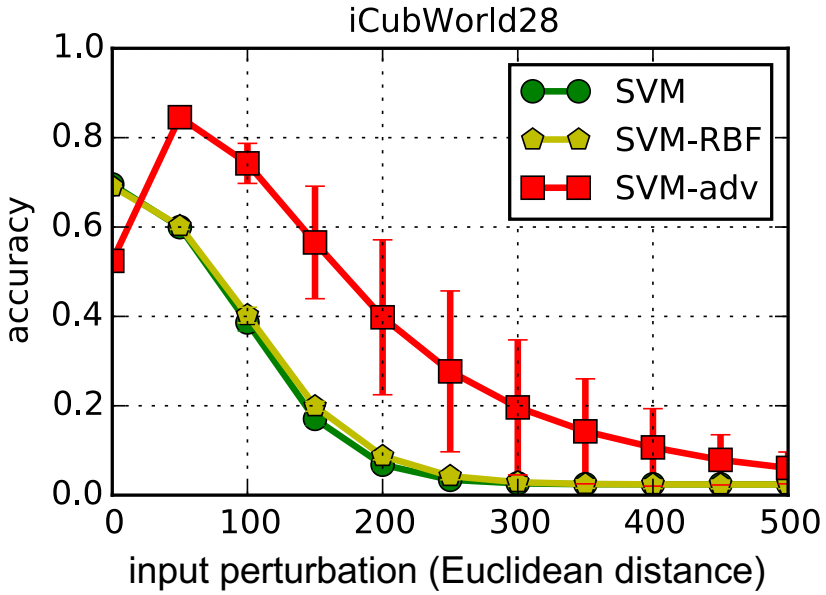
blind-spot evasion
(not even required to mimic the target class)



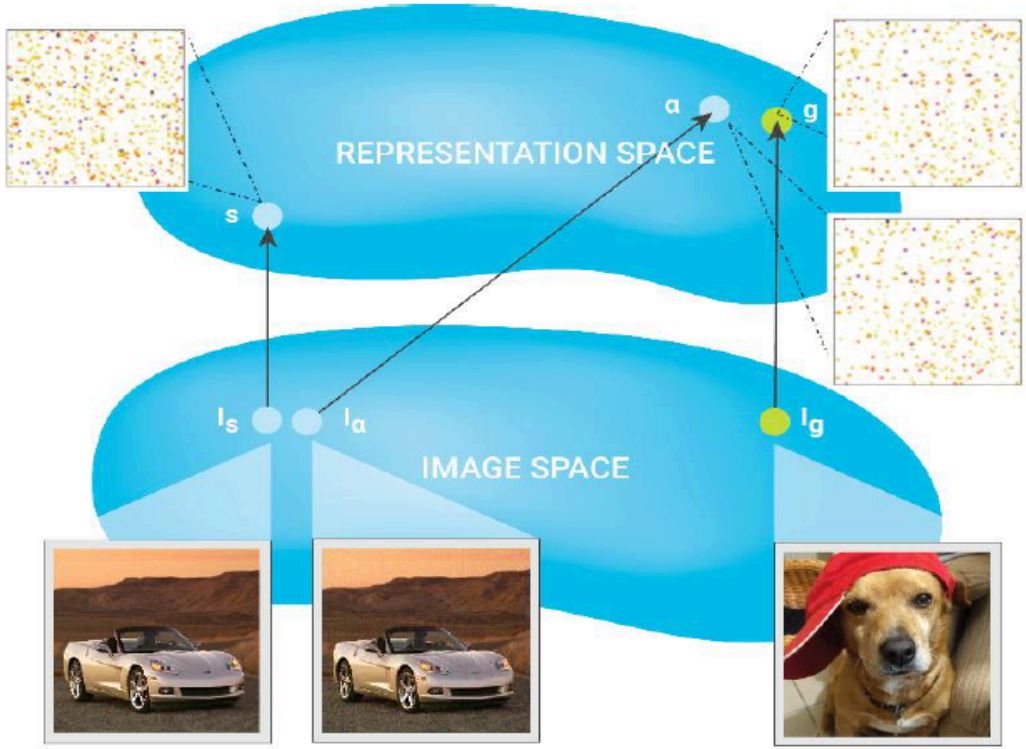
rejection of adversarial examples through enclosing of legitimate classes

Detecting & Rejecting Adversarial Examples





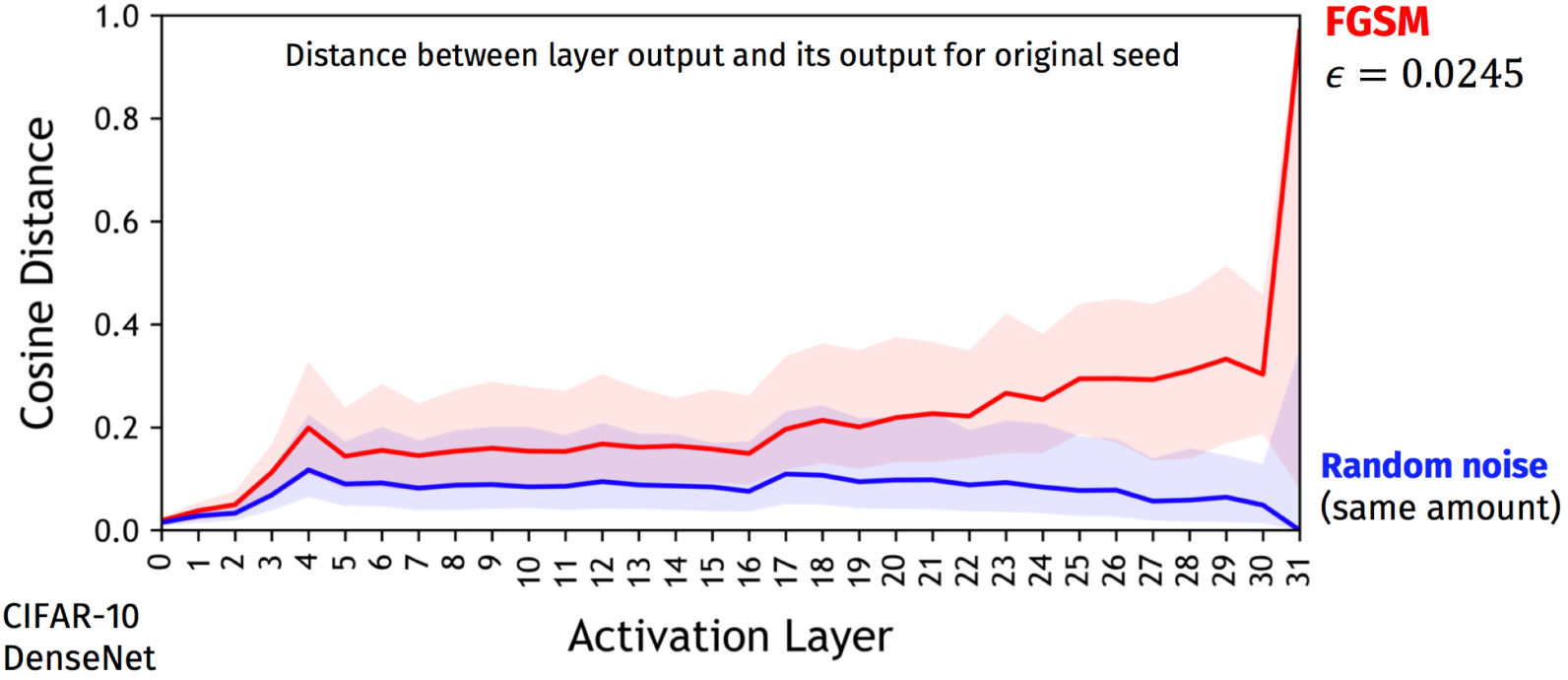
Why Rejection (in Representation Space) Is Not Enough?



[S. Sabour at al., ICLR 2016]

Why Rejection (in Representation Space) Is Not Enough?

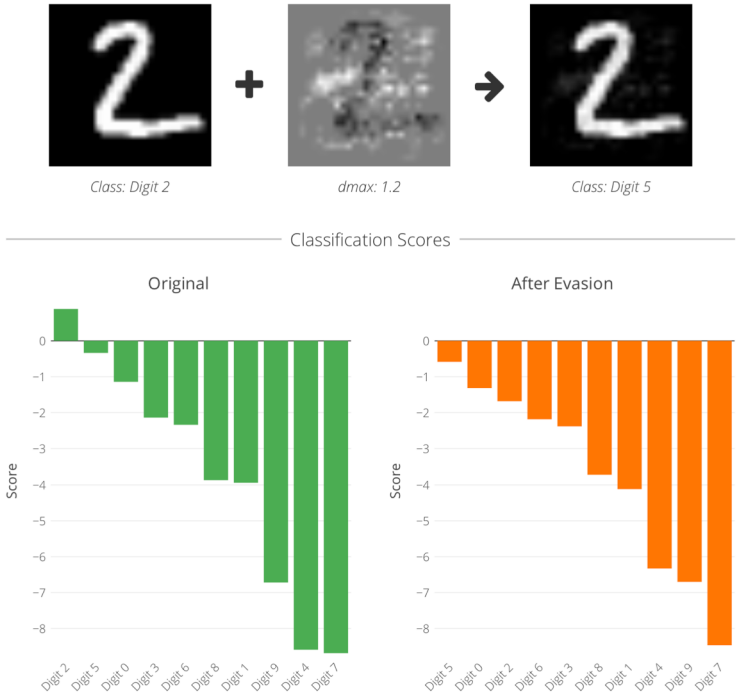
Slide credit: David Evans, DLS 2018 - https://www.cs.virginia.edu/~evans/talks/dls2018/



Adversarial Examples against Machine Learning

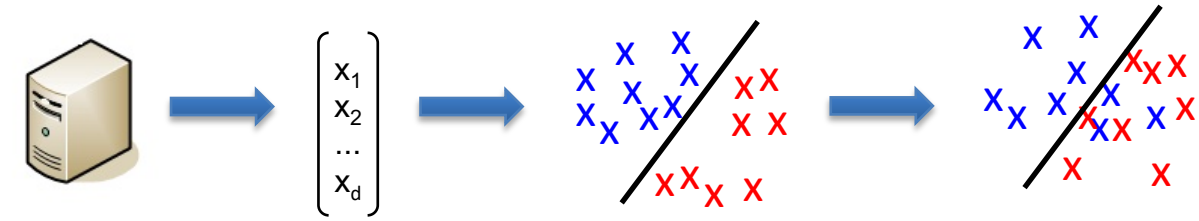
Web Demo

https://sec-ml.pluribus-one.it/demo



Poisoning Machine Learning

Poisoning Machine Learning

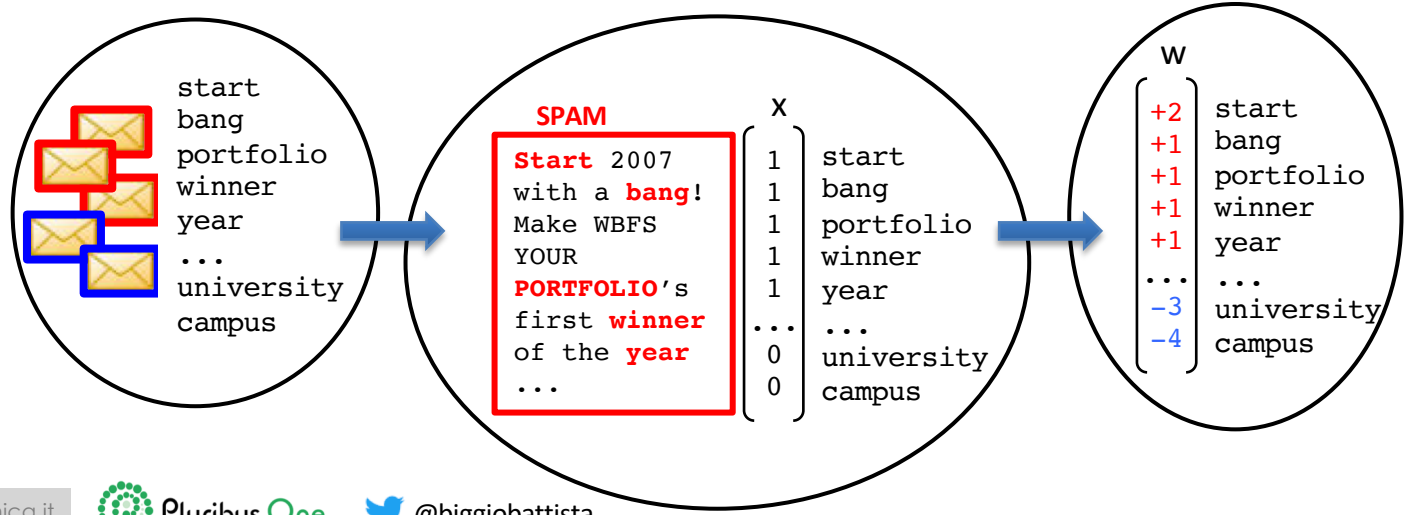


training data (with labels)

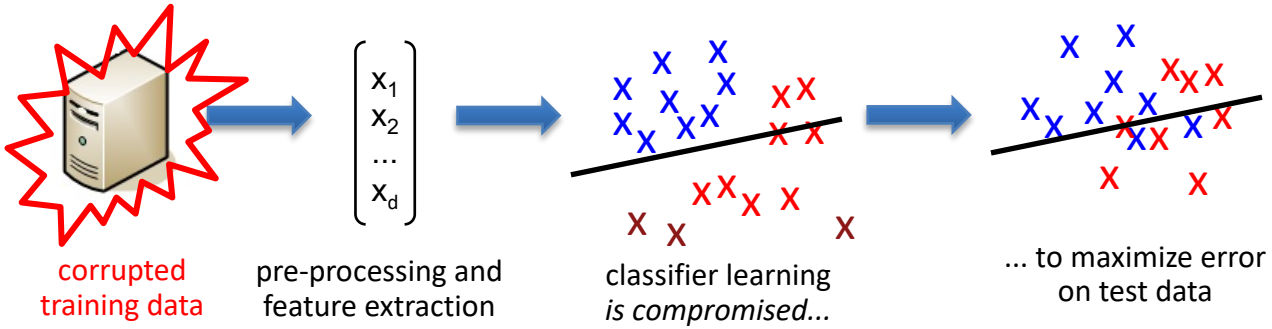
pre-processing and feature extraction

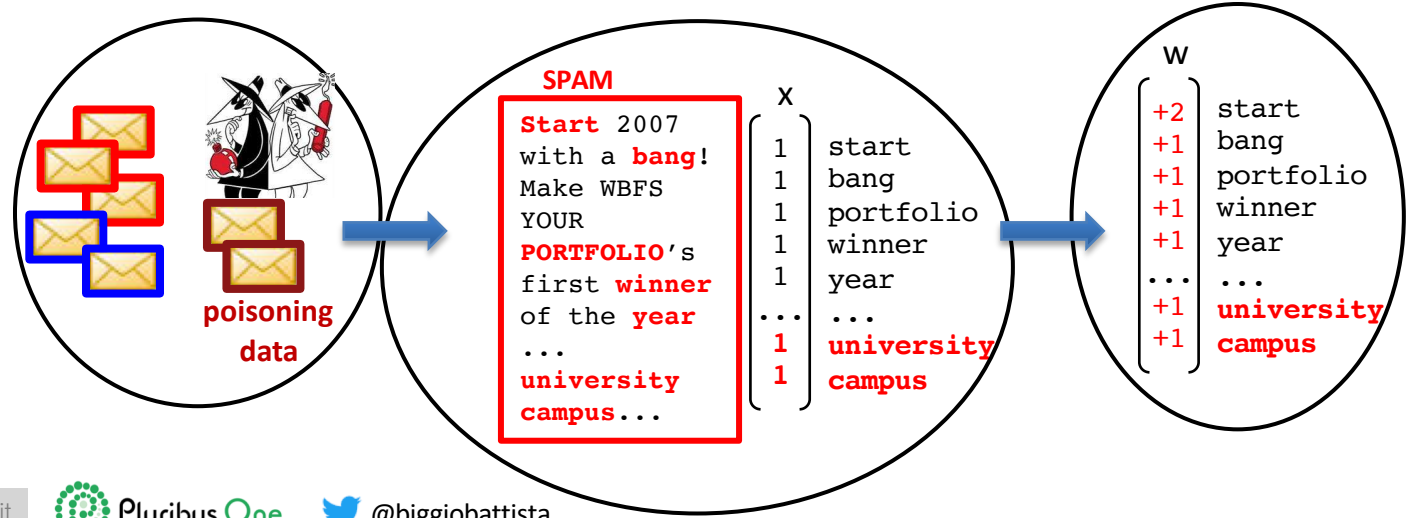
classifier learning

classifier generalizes well on test data



Poisoning Machine Learning

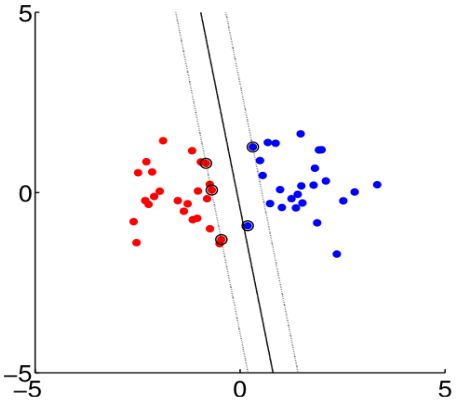




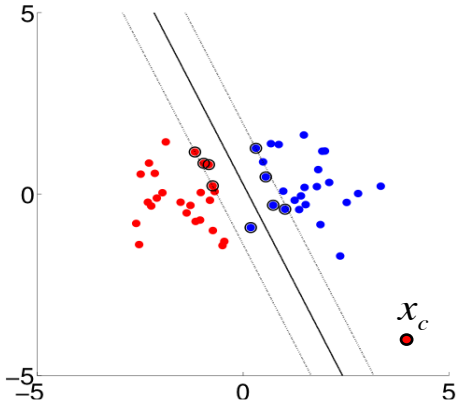
Poisoning Attacks against Machine Learning

- **Goal**: to maximize classification error
- **Knowledge:** perfect / white-box attack
- **Capability:** injecting poisoning samples into TR
- **Strategy:** find an *optimal* attack point x_c in TR that maximizes classification error

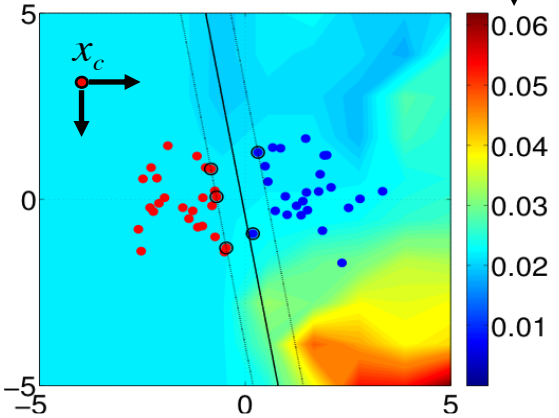
classification error = 0.022



classification error = 0.039



classification error as a function of x_c



[Biggio, Nelson, Laskov. Poisoning attacks against SVMs. ICML, 2012]

Poisoning is a Bilevel Optimization Problem

- **Attacker's objective**
  - to maximize generalization error on untainted data, w.r.t. poisoning point $\mathbf{x}_c$

$$\max_{x_c} L(D_{val}, f^*)$$

Loss estimated on validation data
(no attack points!)

$$\text{s.t. } f^* = \text{argmin}_f \mathcal{L}(D_{tr} \cup \{\boldsymbol{x}_c, y_c\}, f)$$

Algorithm is trained on surrogate data
(including the attack point)

- Poisoning problem against (linear) SVMs:

$$\max_{x_c} \sum_{k=1}^{m} \max(0, 1 - y_k f^*(\boldsymbol{x}_k))$$

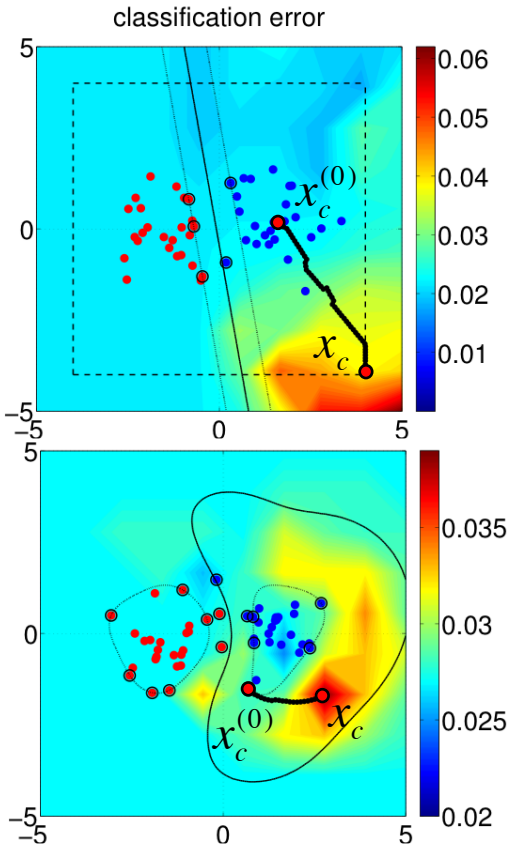
$$\text{s.t. } f^* = \text{argmin}_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \mathrm{C}\sum_{i=1}^{n} \max(0, 1 - y_i f(\boldsymbol{x}_i)) + \mathrm{C}\max(0, 1 - y_c f(\boldsymbol{x}_c))$$

[Biggio, Nelson, Laskov. Poisoning attacks against SVMs. ICML, 2012]
[Xiao, Biggio, Roli et al., Is feature selection secure against training data poisoning? ICML, 2015]
[Munoz-Gonzalez, Biggio, Roli et al., Towards poisoning of deep learning..., AISec 2017]

Gradient-based Poisoning Attacks

- Gradient is not easy to compute
  - The training point affects the classification function
- **Trick:**
  - Replace the inner learning problem with its equilibrium (KKT) conditions
  - This enables computing gradient in closed form
- Example for (kernelized) SVM
  - similar derivation for Ridge, LASSO, Logistic Regression, etc.

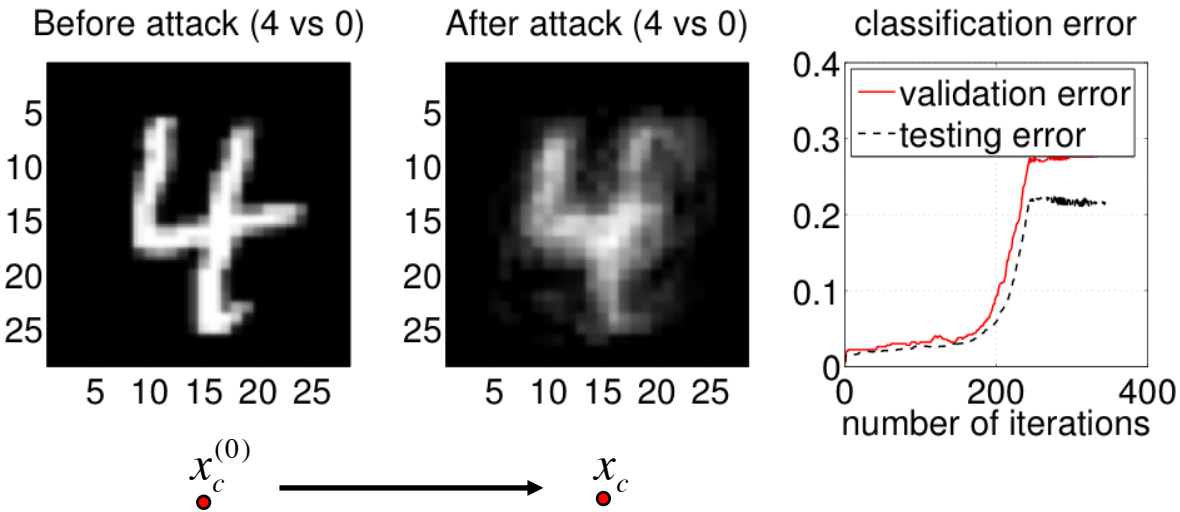
$$\nabla_{\boldsymbol{x}_c}\mathcal{A} = -\boldsymbol{y}_k^\top \frac{\partial \boldsymbol{k}_{kc}}{\partial \boldsymbol{x}_c}\alpha_c + \boldsymbol{y}_k^\top \underbrace{[\mathbf{K}_{ks} \quad \mathbf{1}]}_{k\times s+1} \underbrace{\begin{bmatrix} \boldsymbol{K}_{ss} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial \boldsymbol{k}_{sc}}{\partial \boldsymbol{x}_c} \\ 0 \end{bmatrix} \alpha_c}_{(s+1)\times d}$$



[Biggio, Nelson, Laskov. Poisoning attacks against SVMs. ICML, 2012]
[Xiao, Biggio, Roli et al., Is feature selection secure against training data poisoning? ICML, 2015]

Experiments on MNIST digits
Single-point attack

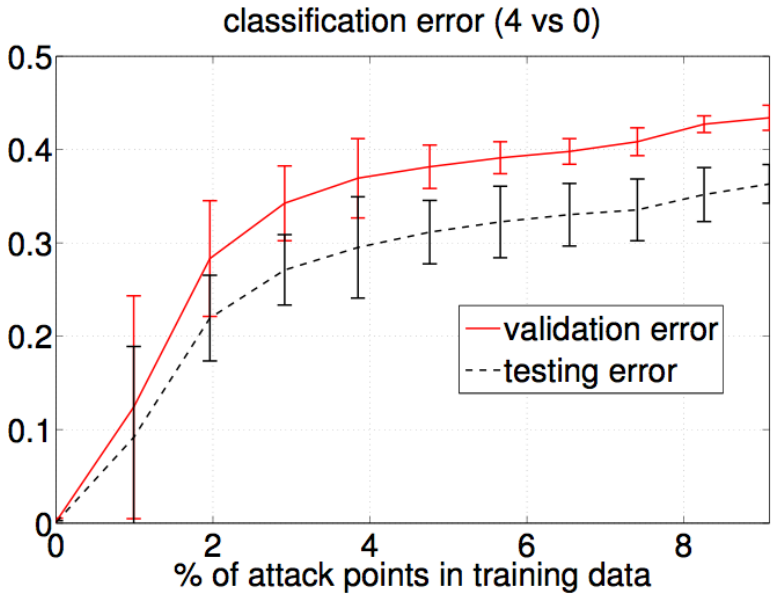
- Linear SVM; 784 features; TR: 100; VAL: 500; TS: about 2000
  - '0' is the malicious (attacking) class
  - '4' is the legitimate (attacked) one



[Biggio, Nelson, Laskov. Poisoning attacks against SVMs. ICML, 2012]

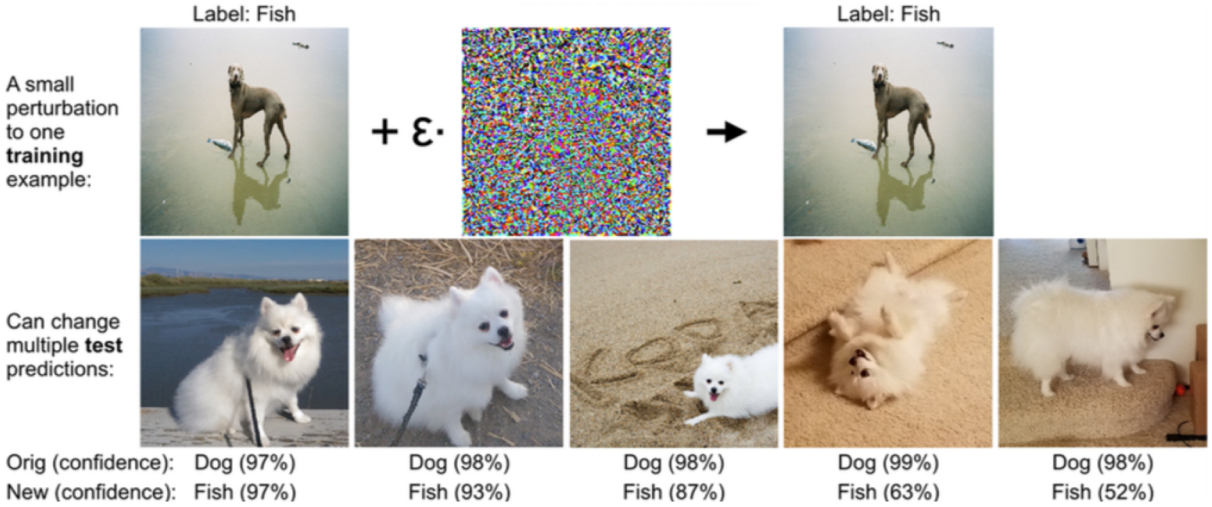
Experiments on MNIST digits
Multiple-point attack

- Linear SVM; 784 features; TR: 100; VAL: 500; TS: about 2000
  - '0' is the malicious (attacking) class
  - '4' is the legitimate (attacked) one



How about Poisoning Deep Nets?

- ICML 2017 Best Paper by *Koh et al., "Understanding black-box predictions via Influence Functions"* has derived adversarial *training* examples against a DNN
  - they have been constructed attacking only the last layer (KKT-based attack against logistic regression) and assuming the rest of the network to be "frozen"



Towards Poisoning Deep Neural Networks

- Solving the poisoning problem without exploiting KKT conditions (back-gradient)
  - *Muñoz-González, Biggio, Roli et al.*, AISec 2017 https://arxiv.org/abs/1708.08689

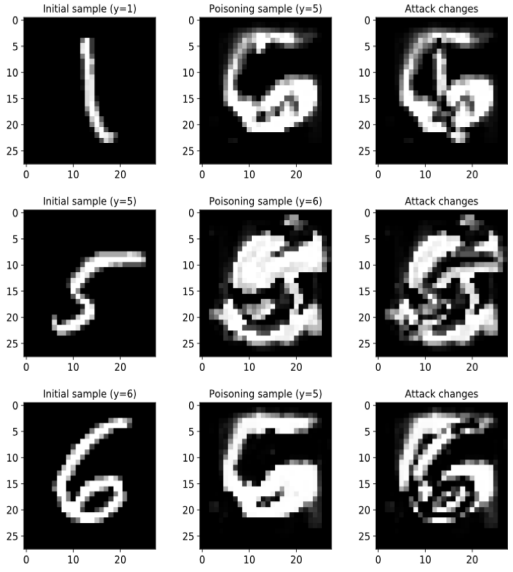


Figure 6: Poisoning samples targeting the LR.

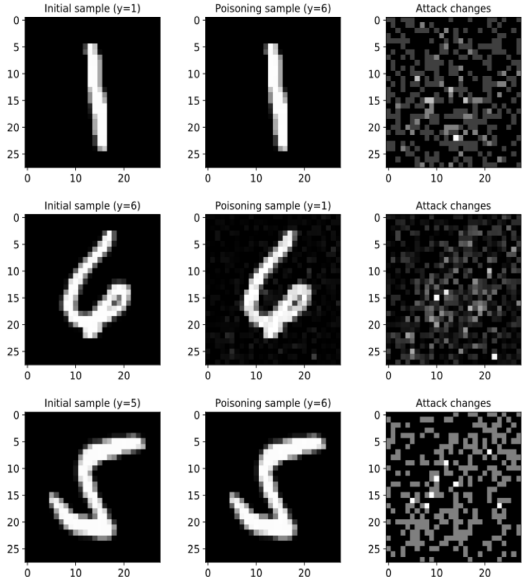
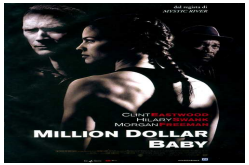


Figure 5: Poisoning samples targeting the CNN.

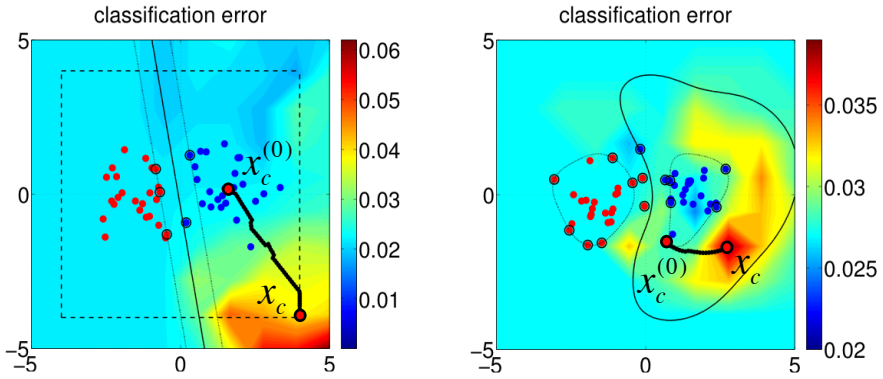
Countering Poisoning Attacks



What is the rule? The rule is protect yourself at all times (from the movie “Million dollar baby”, 2004)

Security Measures against Poisoning

- **Rationale:** poisoning injects outlying training samples



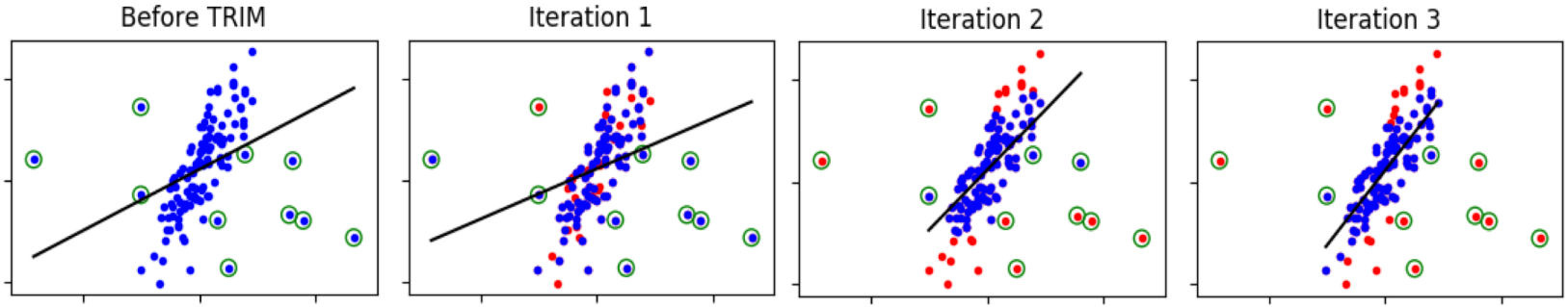
- Two main strategies for countering this threat
  1. **Data sanitization:** *remove* poisoning samples from training data
     - Bagging for fighting poisoning attacks
     - Reject-On-Negative-Impact (RONI) defense
  2. **Robust Learning:** learning algorithms that are robust in the presence of poisoning samples

Robust Regression with TRIM

- TRIM learns the model by retaining only training points with the smallest residuals

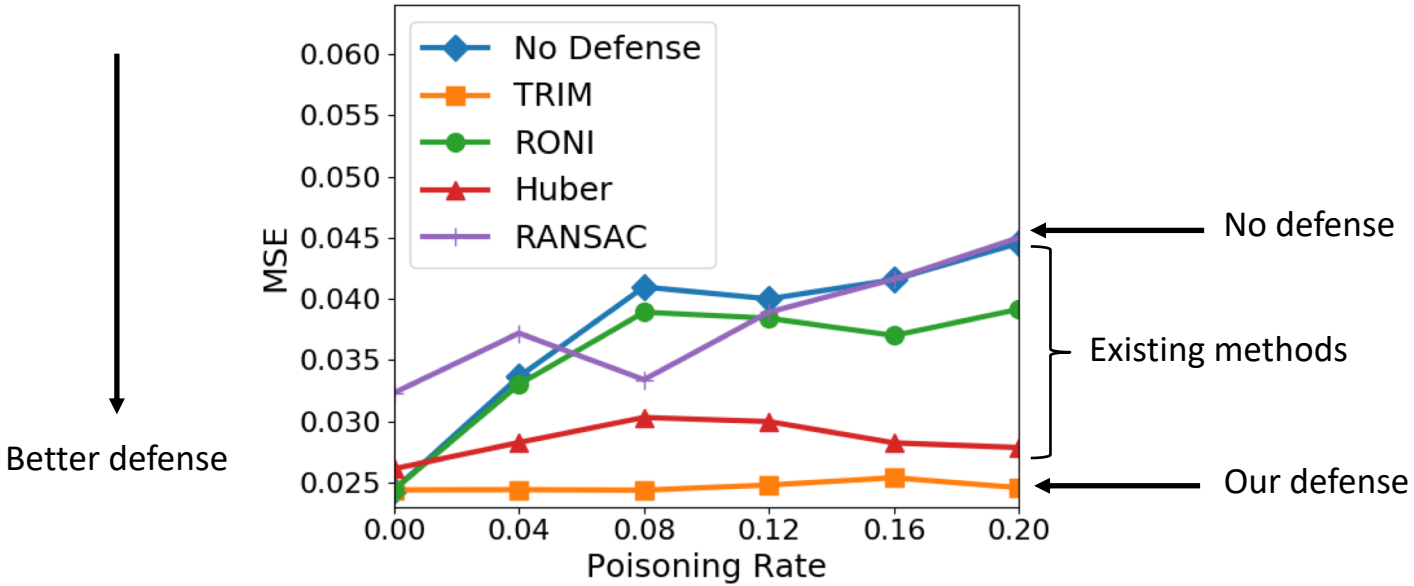
$$\underset{w,b,I}{\operatorname{argmin}}\, L(w,b,I) = \frac{1}{|I|}\sum_{i\in I}(f(\boldsymbol{x}_i) - y_i)^2 + \lambda\Omega(\boldsymbol{w})$$

$$N = (1+\alpha)n, \qquad I \subset [1,\dots,N], \qquad |I| = n$$



Experiments with TRIM (Loan Dataset)

- TRIM MSE is within 1% of original model MSE



[Jagielski, Biggio et al., IEEE Symp. Security and Privacy, 2018]

Other Attacks against ML

Attacks against Machine Learning

Attacker's Capability	Attacker's Goal: Misclassifications that do not compromise normal system operation — Integrity	Attacker's Goal: Misclassifications that compromise normal system operation — Availability	Attacker's Goal: Querying strategies that reveal confidential information on the learning model or its users — Privacy / Confidentiality
Test data	**Evasion (a.k.a. adversarial examples)**	-	Model extraction / stealing Model inversion (hill-climbing) Membership inference attacks
Training data	Poisoning (to allow subsequent intrusions) – e.g., backdoors or neural network trojans	**Poisoning (to maximize classification error)**	-

Attacker's Knowledge:

- perfect-knowledge (PK) white-box attacks
- limited-knowledge (LK) black-box attacks (*transferability* with surrogate/substitute learning models)

Model Inversion Attacks

Privacy Attacks

- ***Goal*: to extract users' sensitive information**
  (e.g., face templates stored during user enrollment)
  - *Fredrikson, Jha, Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. ACM CCS, 2015*

- Also known as hill-climbing attacks in the biometric community
  - *Adler. Vulnerabilities in biometric encryption systems. 5th Int'l Conf. AVBPA, 2005*
  - *Galbally, McCool, Fierrez, Marcel, Ortega-Garcia. On the vulnerability of face verification systems to hill-climbing attacks. Patt. Rec., 2010*

- ***How***: by repeatedly querying the target system and adjusting the input sample to maximize its output score (e.g., a measure of the similarity of the input sample with the user templates)

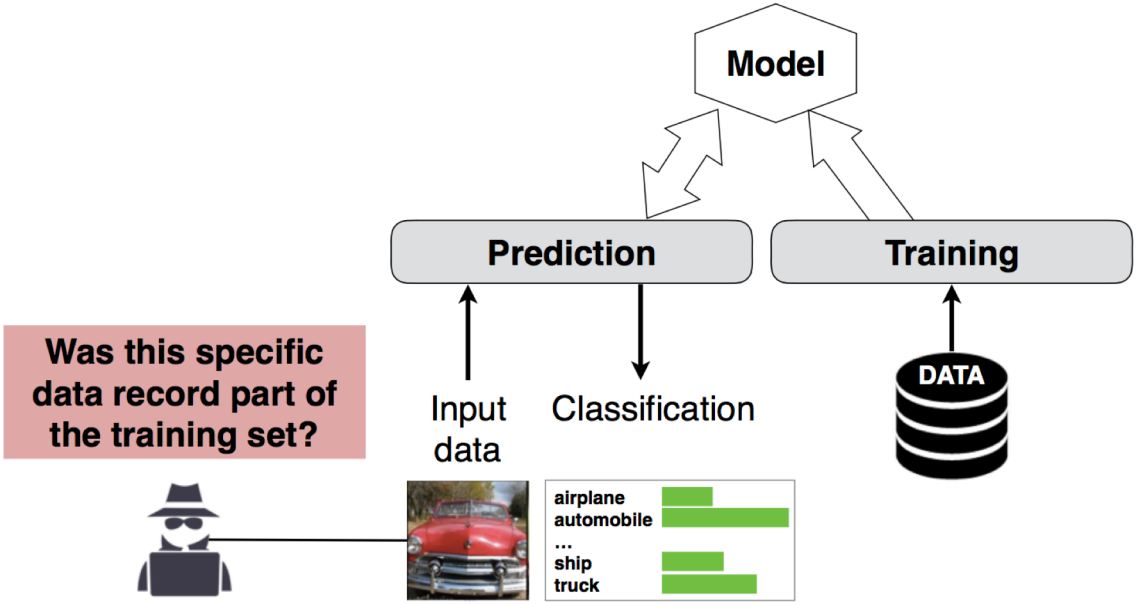
Training Image

Reconstructed Image

Membership Inference Attacks

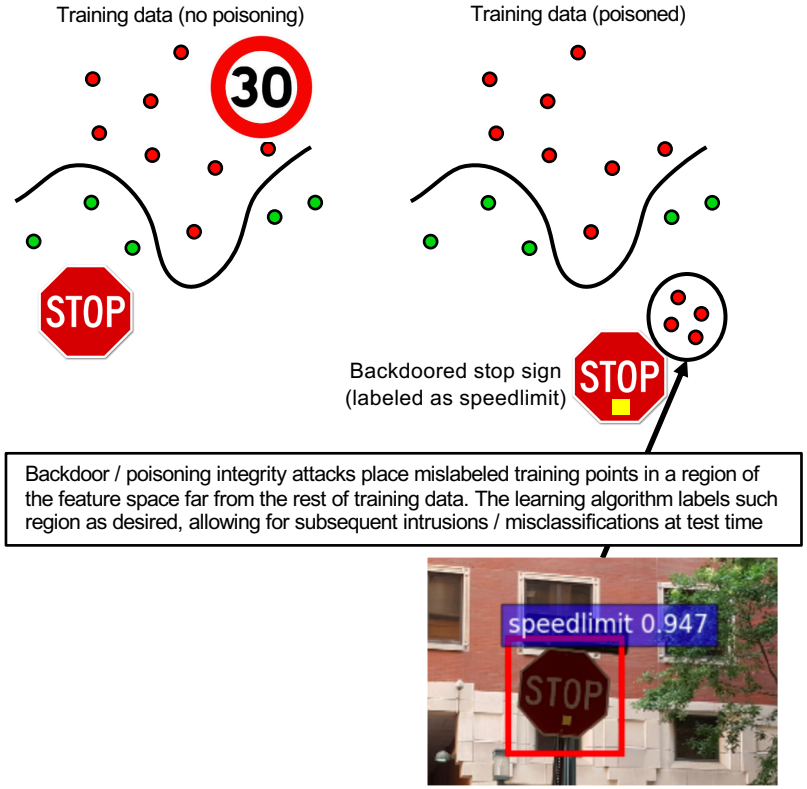
Privacy Attacks (Shokri et al., IEEE Symp. SP 2017)

- ***Goal:*** to identify whether an input sample is part of the training set used to learn a deep neural network based on the observed prediction scores for each class



Backdoor Attacks

Poisoning Integrity Attacks



Attack referred to as backdoor

T. Gu, B. Dolan-Gavitt, and S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. In NIPS Workshop on Machine Learning and Computer Security, 2017.

X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted backdoor attacks on deep learning systems using data poisoning. ArXiv e-prints, 2017.

Attack referred to as 'poisoning integrity'

M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar. Can machine learning be secure? In Proc. ACM Symp. Information, Computer and Comm. Sec., ASIACCS '06, pages 16–25, New York, NY, USA, 2006. ACM.

M. Barreno, B. Nelson, A. Joseph, and J. Tygar. The security of machine learning. Machine Learning, 81:121–148, 2010.

B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In J. Langford and J. Pineau, editors, 29th Int'l Conf. on Machine Learning, pages 1807–1814. Omnipress, 2012.

B. Biggio, G. Fumera, and F. Roli. Security evaluation of pattern classifiers under attack. IEEE Transactions on Knowledge and Data Engineering, 26(4):984–996, April 2014.

H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli. Is feature selection secure against training data poisoning? In F. Bach and D. Blei, editors, JMLR W&CP - Proc. 32nd Int'l Conf. Mach. Learning (ICML), volume 37, pages 1689–1698, 2015.

L. Munoz-Gonzalez, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In 10th ACM Workshop on Artificial Intelligence and Security, AISec '17, pp. 27–38, 2017. ACM.

B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. ArXiv e-prints, 2018.

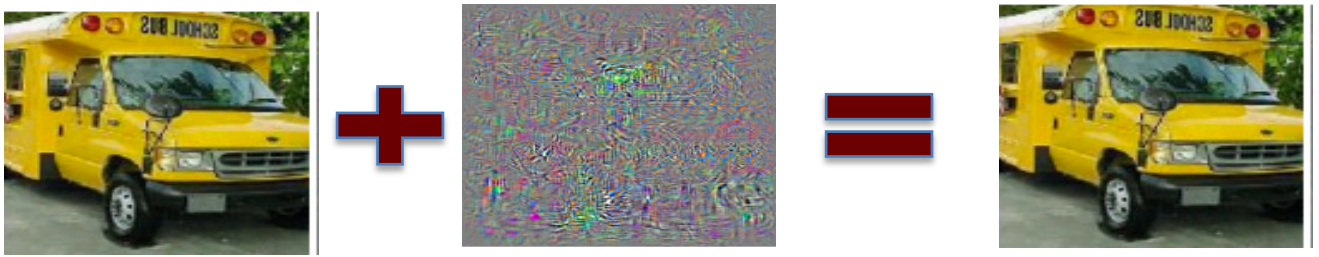
M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In 39th IEEE Symp. on Security and Privacy, 2018.

Are Adversarial Examples a Real Security Threat?

World Is Not Digital…

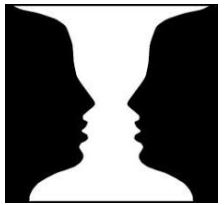
- *….**Previous** cases of **adversarial examples** have common characteristic: the **adversary** is able to **precisely control** the **digital representation** of the input to the machine learning tools…..*

[M. Sharif et al., ACM CCS 2016]



School Bus (*x*) **Adversarial Noise (*r*)** **Ostrich Struthio Camelus**

Do Adversarial Examples Exist in the Physical World?



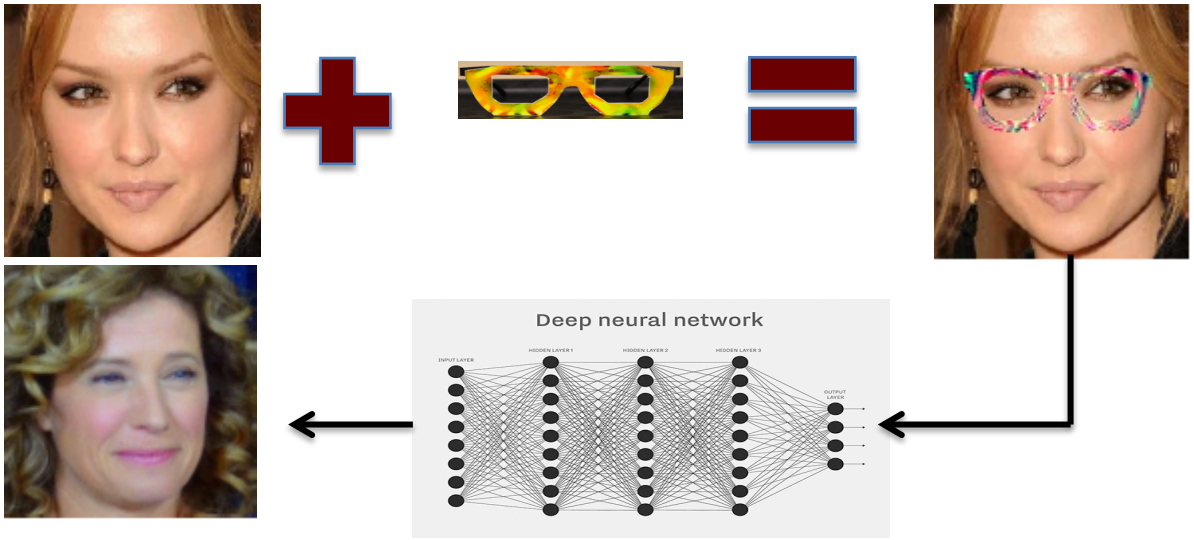
Adversarial Images in the Physical World

- Adversarial images fool deep networks **even when** they **operate** in the **physical world**, for example, **images** are **taken from** a **cell-phone camera**?
  - Alexey Kurakin et al. (2016, 2017) explored the possibility of creating adversarial images for machine learning systems which operate in the physical world. They used images taken from a cell-phone camera as an input to an Inception v3 image classification neural network
  - They showed that in such a set-up, a significant fraction of adversarial images crafted using the original network are misclassified even when fed to the classifier through the camera

Adversarial Glasses

$$\underset{r}{\operatorname{argmin}} \Big(\big(\sum_{x \in X} softmaxloss(x + r, c_t) \big) + \kappa_1 \cdot TV(r) + \kappa_2 \cdot NPS(r) \Big)$$

The adversarial perturbation is applied only to the eyeglasses image region



[M. Sharif et al., ACM CCS 2016]

Should We Be Worried ?

No, We Should Not…

[arXiv:1707.03501; CVPR 2017]

NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles

Jiajun Lu*, Hussein Sibai*, Evan Fabry, David Forsyth
University of Illinois at Urbana Champaign
{jlu23, sibai2, efabry2, daf}@illinois.edu

In this paper, we show experiments that suggest that a trained neural network classifies most of the pictures taken from different distances and angles of a perturbed image correctly. We believe this is because the adversarial property of the perturbation is **sensitive to the scale** at which the perturbed picture is viewed, so (for example) **an autonomous car** will **misclassify** a **stop** sign **only** from a **small range of distances**.

Yes, We Should...

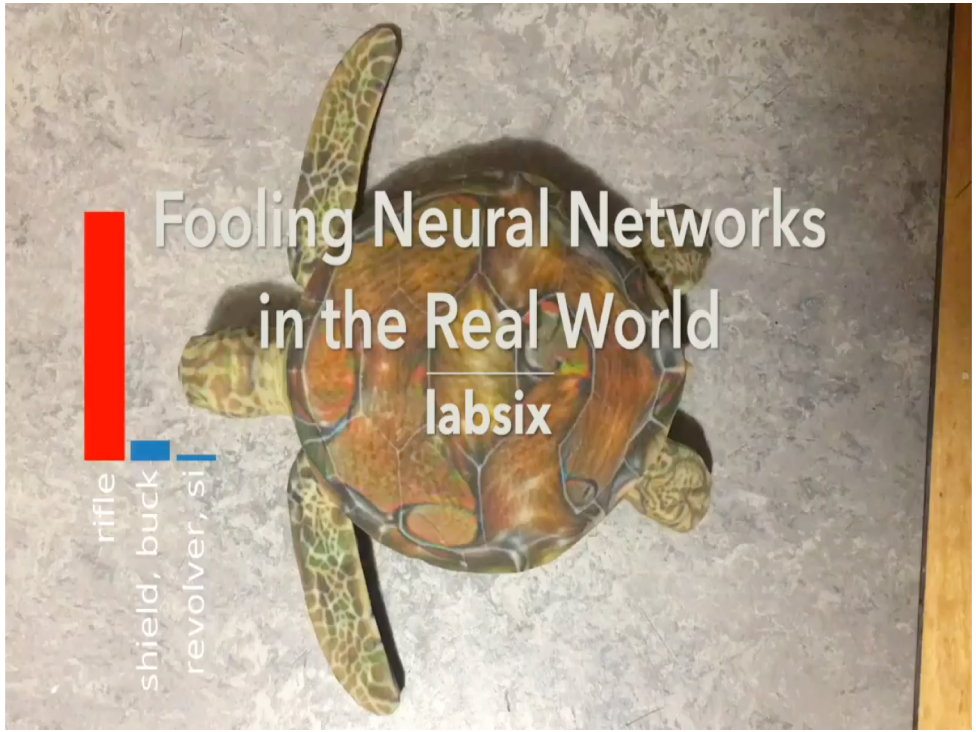
Synthesizing Robust Adversarial Examples

Anish Athalye
OpenAI, MIT

Ilya Sutskever
OpenAI

[https://blog.openai.com/robust-adversarial-inputs/]

Yes, We Should...



Yes, We Should...

Robust Physical-World Attacks on Machine Learning Models

Visit https://iotsecurity.eecs.umich.edu/#roadsigns for an FAQ

Ivan Evtimov[1], Kevin Eykholt[2], Earlence Fernandes[1], Tadayoshi Kohno[1],
Bo Li[4], Atul Prakash[2], Amir Rahmati[3], and Dawn Song*[4]

[1]University of Washington
[2]University of Michigan Ann Arbor
[3]Stony Brook University
[4]University of California, Berkeley

Yes, We Should...

Is This a Real Security Threat?

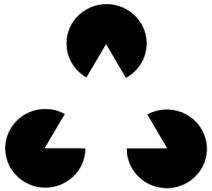
- Adversarial examples can exist in the physical world, we can fabricate concrete adversarial objects (glasses, road signs, etc.)

- But the effectiveness of attacks carried out by adversarial objects is still to be investigated with large scale experiments in realistic security scenarios

- Gilmer et al. (2018) have recently discussed the realism of security threat caused by adversarial examples, pointing out that it should be carefully investigated
  - Are indistinguishable adversarial examples a real security threat ?
  - For which real security scenarios adversarial examples are the best attack vector? Better than attacking components outside the machine learning component
  - …

[Justin Gilmer et al., Motivating the Rules of the Game for Adversarial Example Research, https://arxiv.org/abs/1807.06732]

Are Indistinguishable Perturbations a Real Security Threat?



Indistinguishable Adversarial Examples

- Minimize $\|r\|_2$ subject to:
  1. $f(x+r) = l \quad f(x) \neq l$
  2. $x + r \in [0,1]^m$

The adversarial image $x + r$ is visually hard to distinguish from x

*…There is a **torrent of work** that views increased robustness to **restricted perturbations** as making these models **more secure**. While not all of this work requires completely indistinguishable modications, many of the papers focus on specifically small modications, and the language in many suggests or implies that the **degree of perceptibility** of the **perturbations** is an important aspect of their **security risk**…*

[Justin Gilmer et al., Motivating the Rules of the Game for Adversarial Example Research, arXiv 2018]

Indistinguishable Adversarial Examples

- The attacker can benefit by minimal perturbation of a legitimate input; e.g., she could use the attack for a longer period of time before it is detected

- But is *minimal perturbation* a necessary constraint for the attacker?

Indistinguishable Adversarial Examples

- Is *minimal perturbation* a necessary constraint for the attacker?

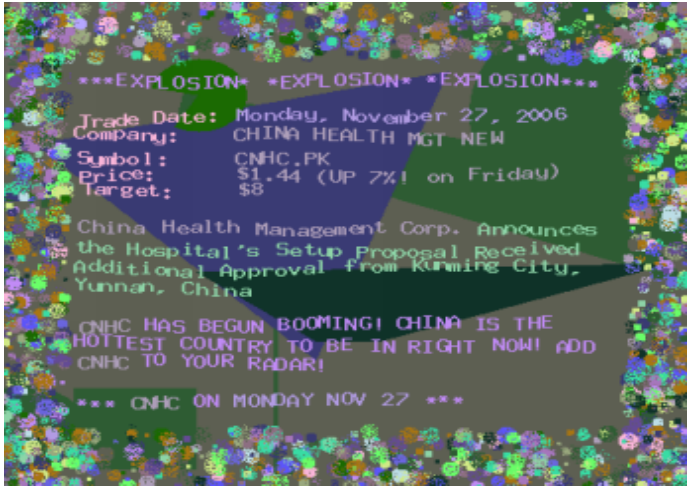
Indistinguishable Adversarial Examples

- Is *minimal perturbation* a necessary constraint for the attacker?

[Flickr user \faungg". Stop? Accessed: 2018-7-18. Aug. 2014. https://www.flickr.com/photos/44534236@N00/15536855528/]

Attacks with Content Preservation

There are well known security applications where minimal perturbations and indistinguishability of adversarial inputs are not required at all…



Are Indistinguishable Perturbations a Real Security Threat?

*…At the time of writing, we were **unable** to find a **compelling example** that **required indistinguishability**…*

*To have the largest impact, we should both recast future adversarial example research as a **contribution** to **core machine learning** and develop new abstractions that capture **realistic threat models**.*

[Justin Gilmer et al., Motivating the Rules of the Game for Adversarial Example Research, arXiv 2018]

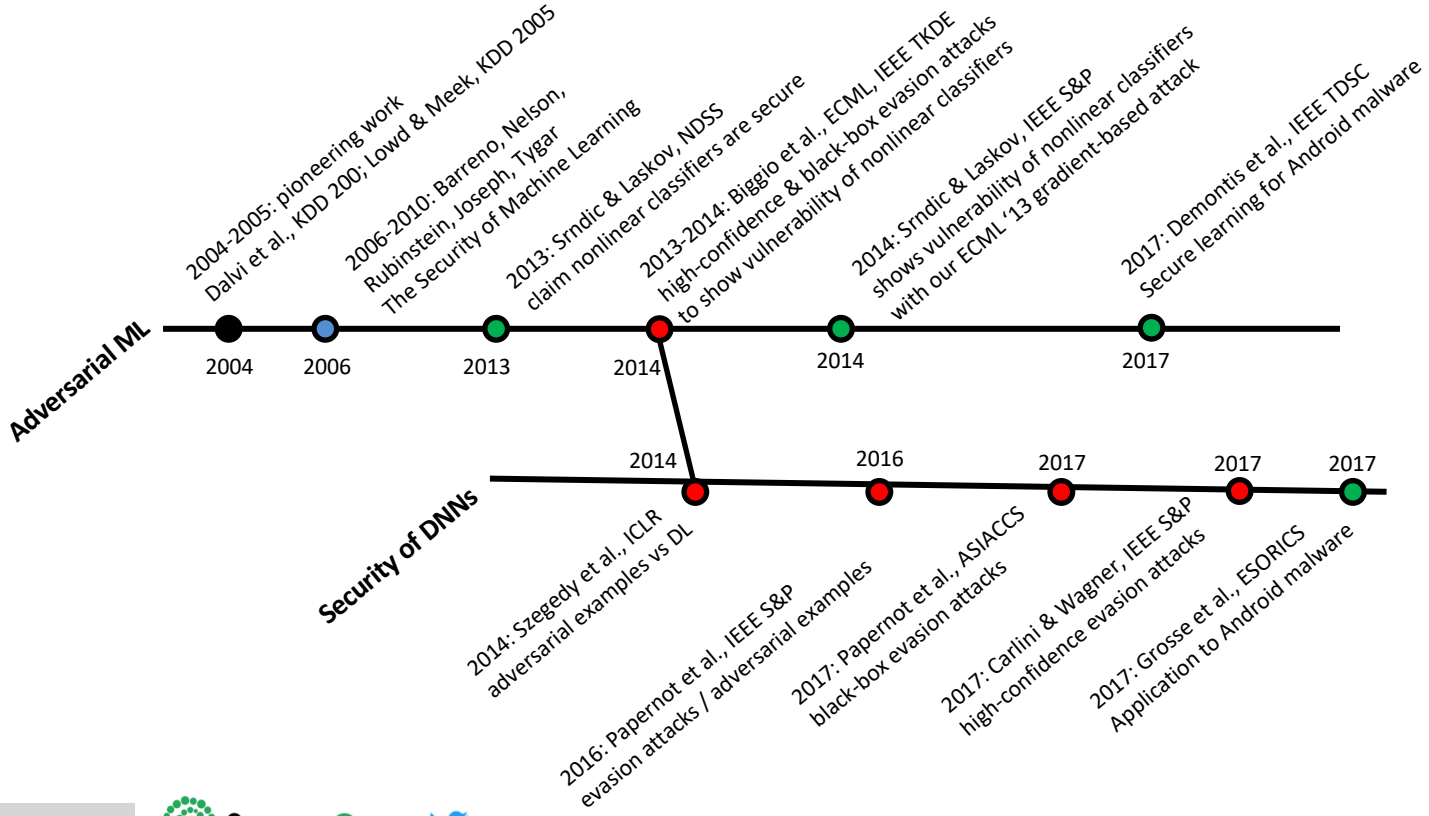
To Conclude…

This is a recent research field…

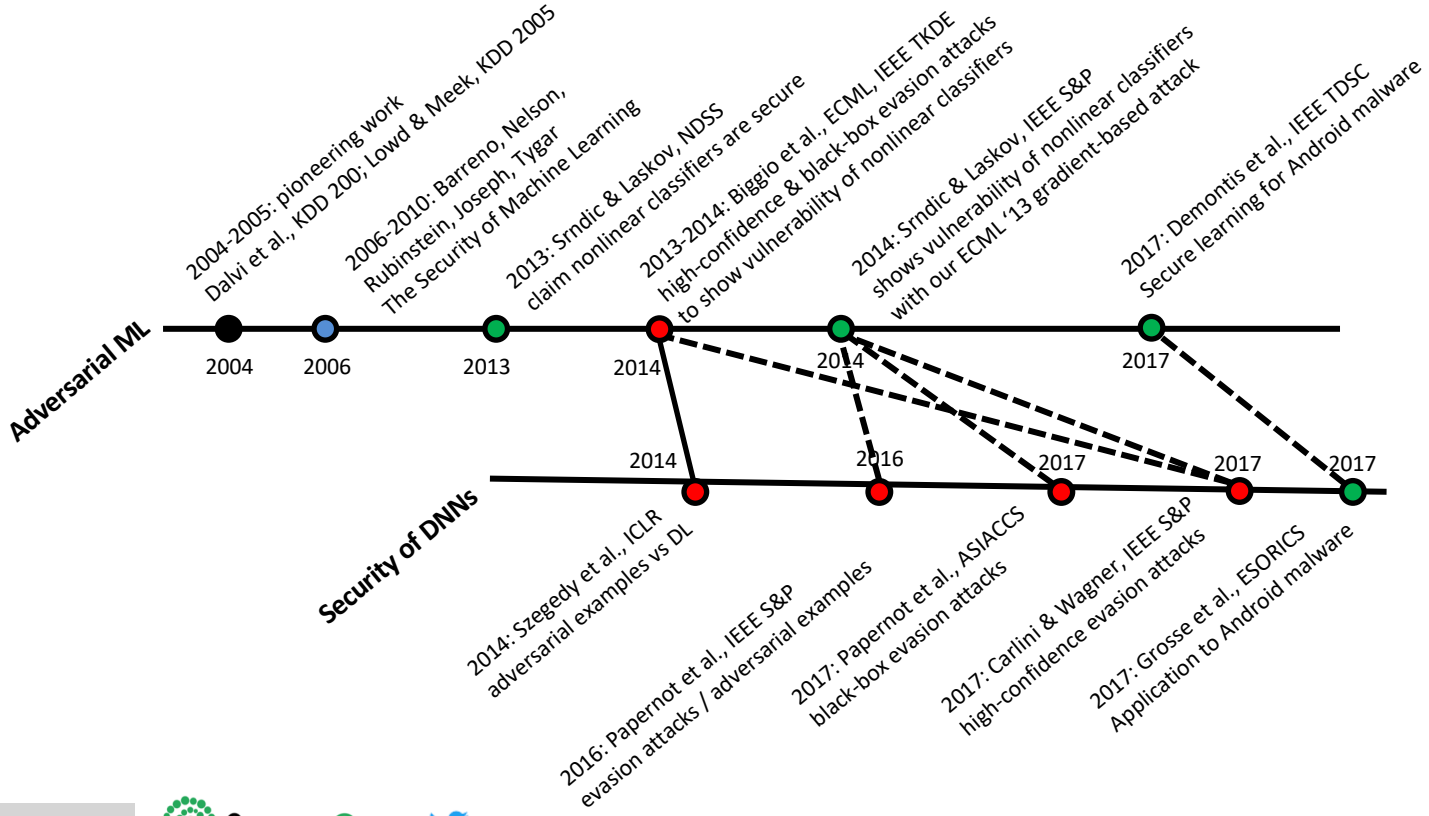
Dagstuhl Perspectives Workshop on
"Machine Learning in Computer Security"
Schloss Dagstuhl, Germany, Sept. 9th-14th, 2012

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Timeline of Learning Security

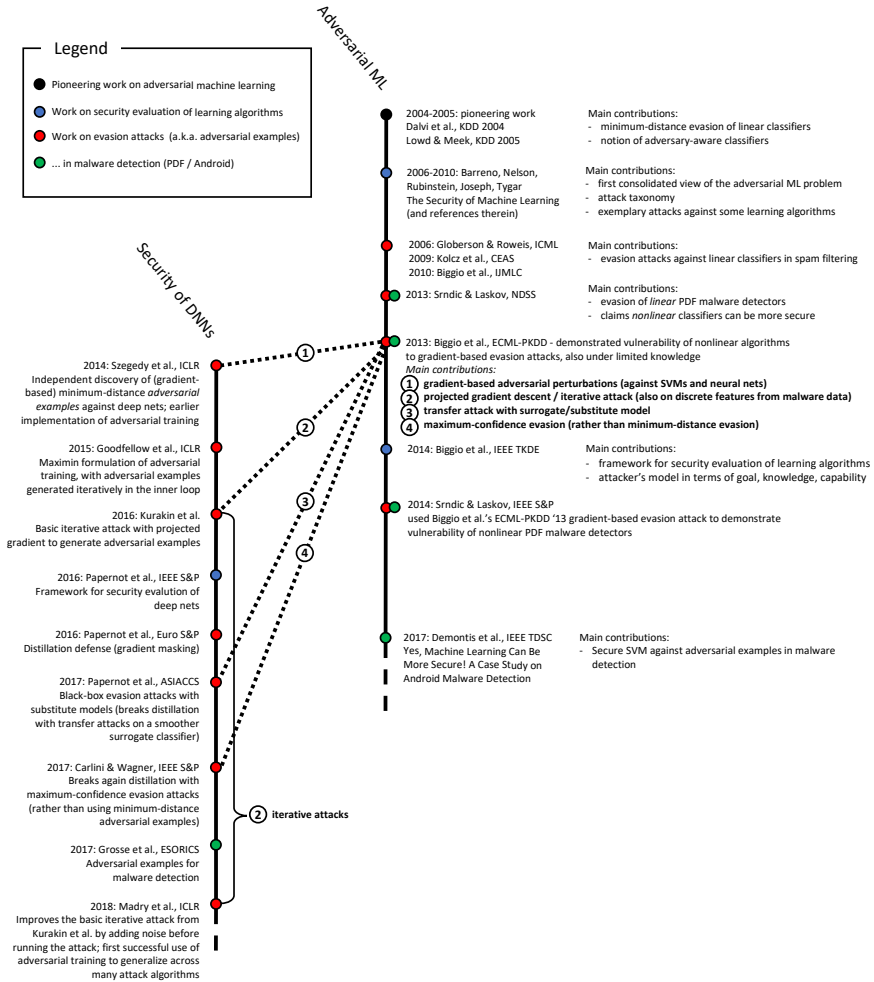


Timeline of Learning Security



Timeline of Learning Security

Biggio and Roli, **Wild Patterns**: *Ten Years After The Rise of Adversarial Machine Learning*, Pattern Recognition, 2018



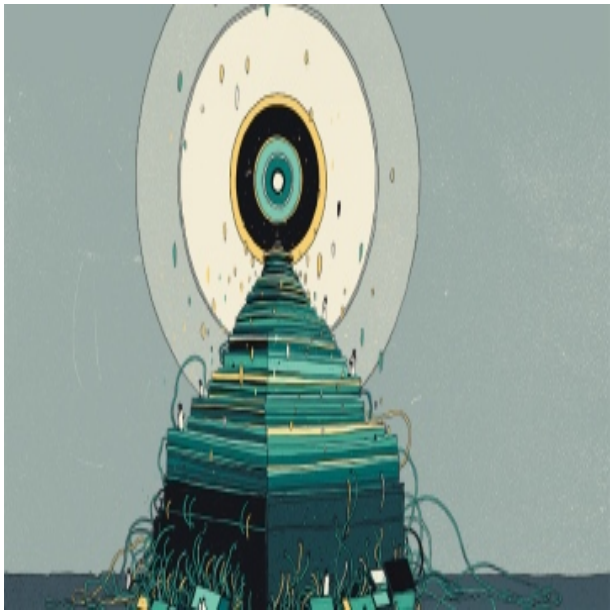
Black Swans to the Fore

[Szegedy et al., Intriguing properties of neural networks, 2014]

After this "black swan", the issue of security of DNNs came to the fore…

Not only on scientific specialistic journals…

The Safety Issue to the Fore…



The black box of AI
D. Castelvecchi, Nature, Vol. 538, 20, Oct 2016

Machine learning is becoming ubiquitous in basic research as well as in industry. But for scientists to trust it, they first need to understand what the machines are doing.

Ellie Dobson, director of data science at the big-data firm Arundo Analytics in Oslo: If something were to go wrong as a result of setting the UK interest rates, she says, "the Bank of England can't say, the black box made me do it"".

Why So Much Interest?

Before the deep net “revolution”, people were not surprised when machine learning was wrong, they were more amazed when it worked well…

Now that it seems to work for real applications, people are disappointed, and worried, for errors that humans do not do…

Errors of Humans and Machines…

Machine learning decisions are affected by several **sources of bias** that causes “strange” errors

But we should keep in mind that also **humans** are **biased**…

The Bat and the Ball Problem

A bat and a ball together cost $ 1.10

The bat costs $ 1.0 more than the ball

How much does the ball cost ?

Please, give me the first answer coming to your mind !

The Bat and the Ball Problem

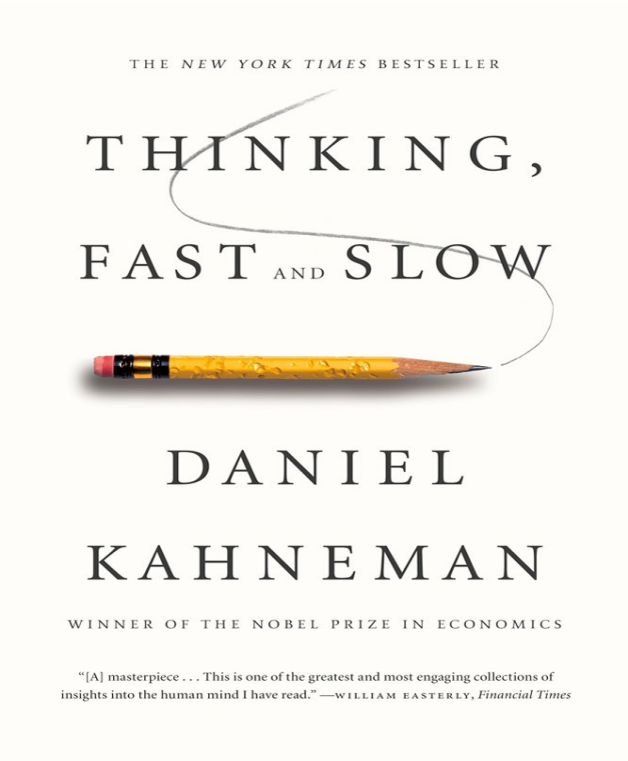
$$\begin{cases} bat+ball=\$1.10 \\ bat=ball+\$1.0 \end{cases}$$

Exact solution is 0.05 dollar (5 cents)

The wrong solution ($ 0.10) is due to the **attribute substitution**, a psychological process thought to underlie a number of **cognitive biases**

It occurs when an individual has to make a judgment (of a target attribute) that is computationally complex, and instead substitutes a more easily calculated heuristic attribute

Trust in Humans or Machines?



Algorithms are biased, but also humans are as well…

When should you trust in humans and when in algorithms?

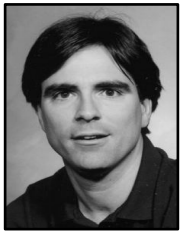
Learning Comes at a Price!

The introduction of novel **learning** functionalities increases the **attack surface** of computer systems and produces new vulnerabilities

Safety of machine learning will be more and more important in future computer systems, as well as **accountability**, **transparency**, and the protection of fundamental human **values** and **rights**

Thanks for Listening!

Any questions?



Engineering isn't about perfect solutions; it's about doing the best you can with limited resources (Randy Pausch, 1960-2008)