Pattern Recognition and Applications Lab

University of Cagliari, Italy

Adversarial AI

Battista Biggio
battista.biggio@diee.unica.it

@biggiobattista

PRA lab @ University of Cagliari, Italy
Pluribus One

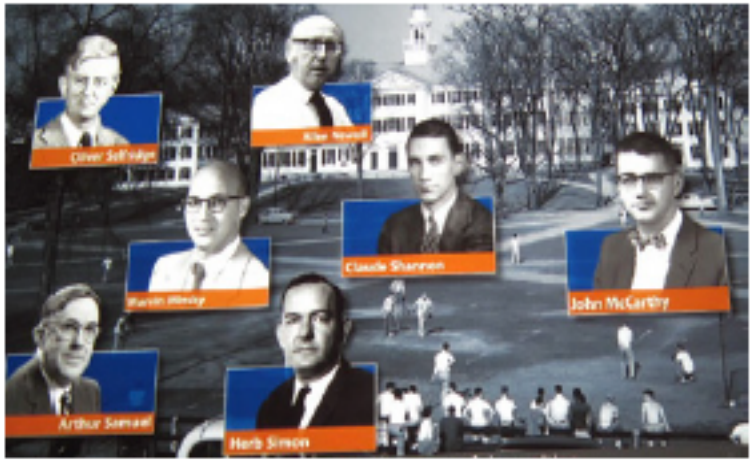
When Was AI Born?

It All Started in 1955

A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

August 31, 1955

John McCarthy, Marvin L. Minsky,
Nathaniel Rochester,
and Claude E. Shannon



[http://www.aaai.org/ojs/index.php/aimagazine/article/view/1904]

What Is Artificial Intelligence?

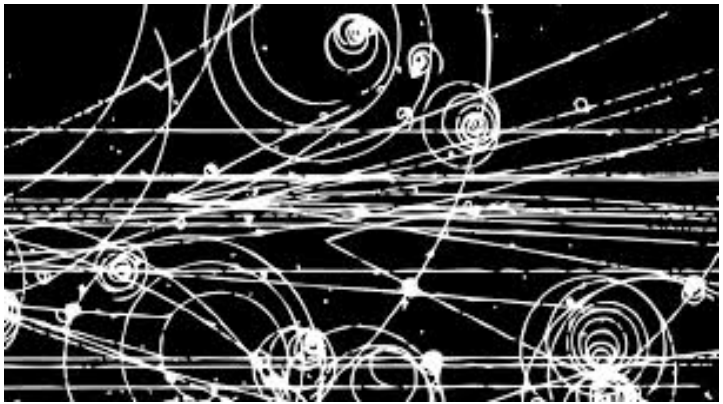
- Ill-posed problem, studied for decades by philosophers / psychologists
- Definition of interest for computer science: Alan Turing, 1950
  - A. Turing, "Computing Machinery & Intelligence," *Mind*, Vol. 59(236), 1950
- Machines that emulate a rational behavior (for a given task)
- Turing Test: a way to test if a machine is 'intelligent'

An Example of a Modern Turing Test

- Completely Automated Public Turing test to tell Computers and Humans Apart (**CAPTCHA**)

AI in the 60s

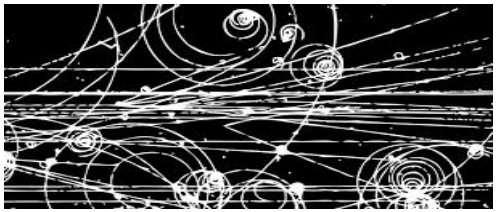
OCR for bank cheque sorting



Detection of particle tracks in bubble chambers

Aerial photo recognition

Key Feature of these Apps



Specialised applications for **professional** users…

Artificial Intelligence Today

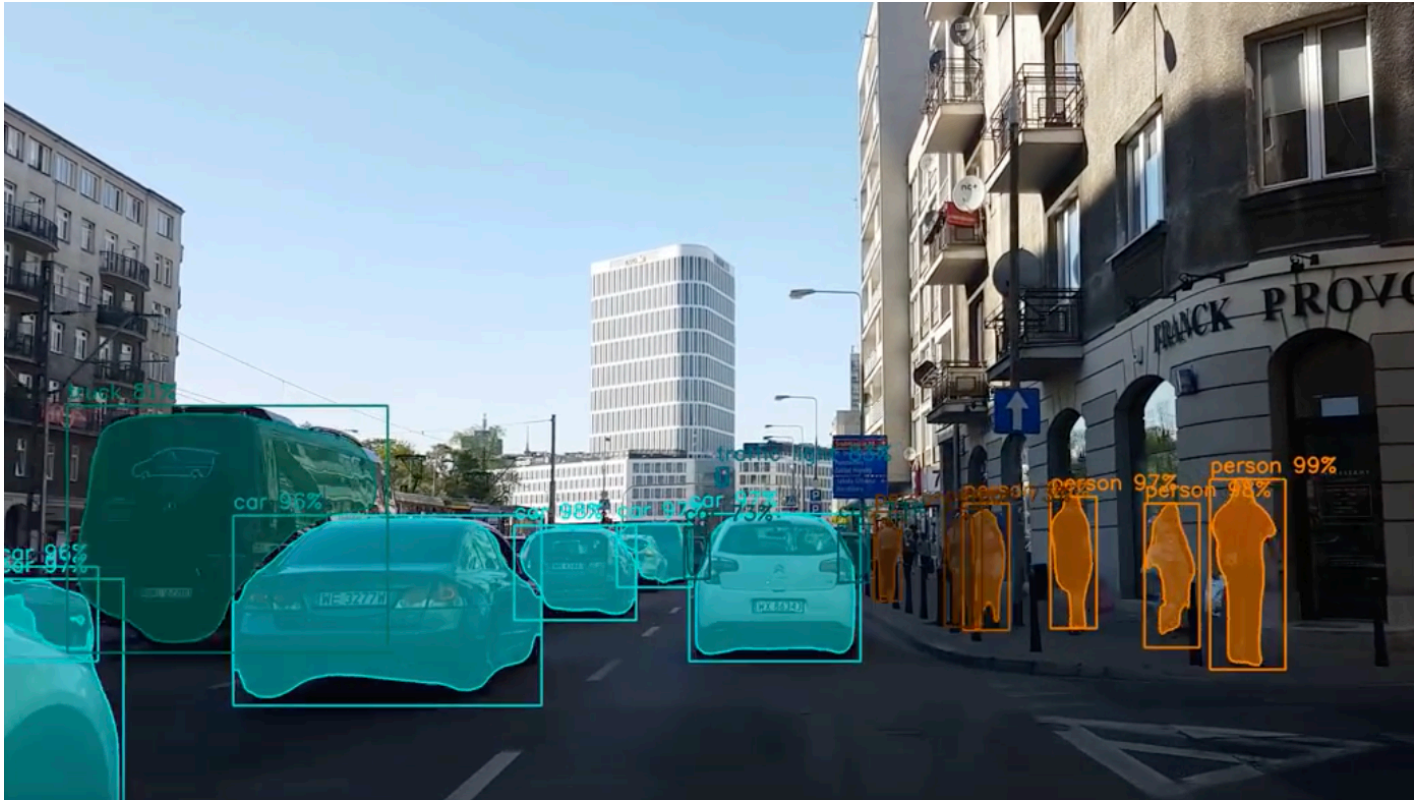
AI is going to transform industry and business as electricity did about a century ago

(*Andrew Ng, Jan. 2017*)

Applications:

- Cybersecurity
- Computer Vision
- Robotics
- Healthcare
- Speech recognition
- Virtual assistants
- …

Computer Vision for Self-Driving Cars

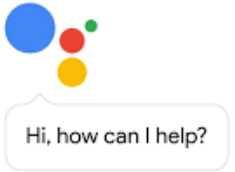


Speech Recognition for Virtual Assistants

Amazon Alexa

Apple Siri

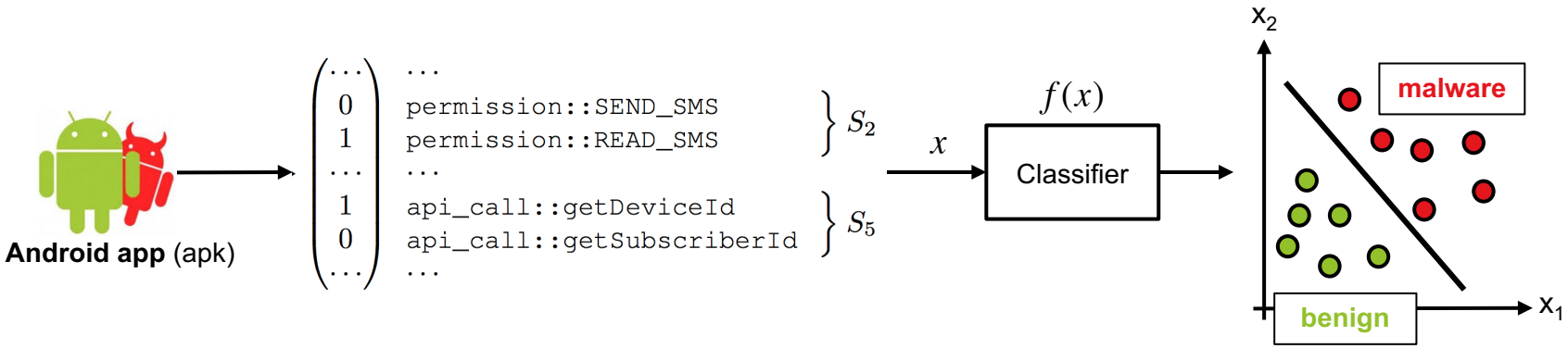
Microsoft Cortana



Google Assistant

AI for Cybersecurity

- Detection of phishing webpages
- Protection of web applications / services
- Monitoring behavior of authenticated users (behavioral/continuous authentication)
- *Mal*icious software (*malware*) detection
  - Executable files, Android apps, ...
- ... and many other applications ...

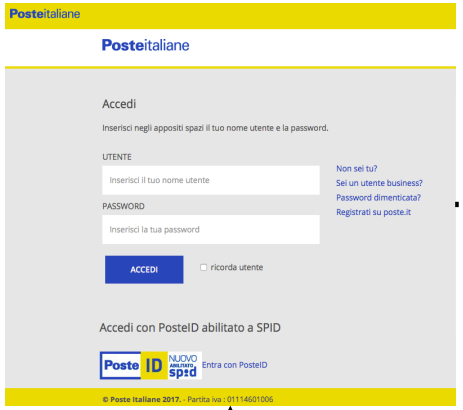


Illicit Activities over the Internet: Phishing Websites / Scams

poste.it

user

Cloned website similar to poste.it (e.g., *posteit.org*)



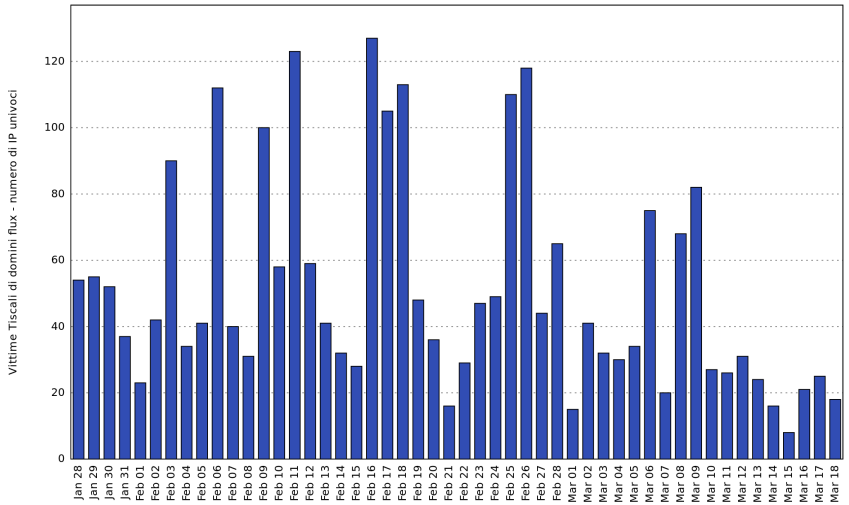
cybercriminal

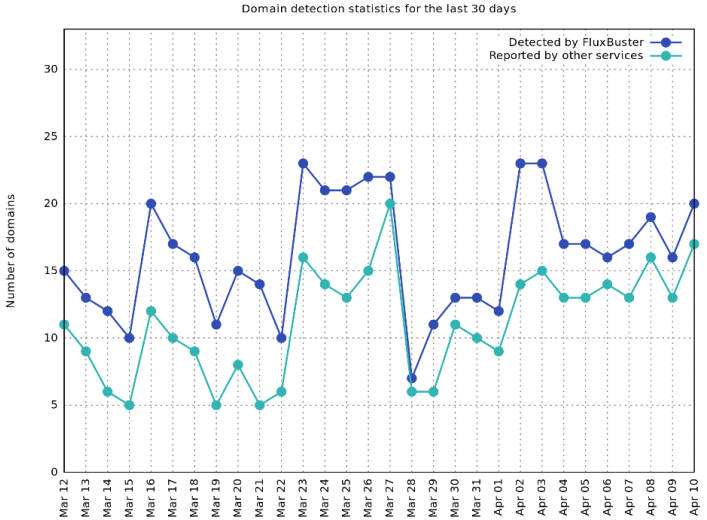
Phishing email

user

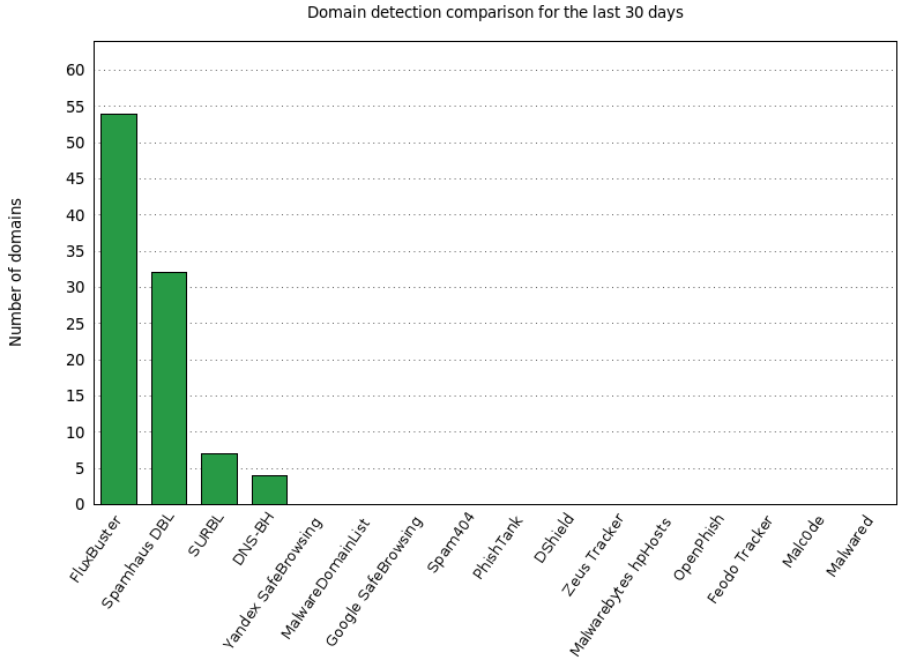
Some results from IllBuster

- Results on a sample of 100,000 Internet users
- Year: 2015
- Main result: early detection of (unknown) malicious domains via passive DNS analysis



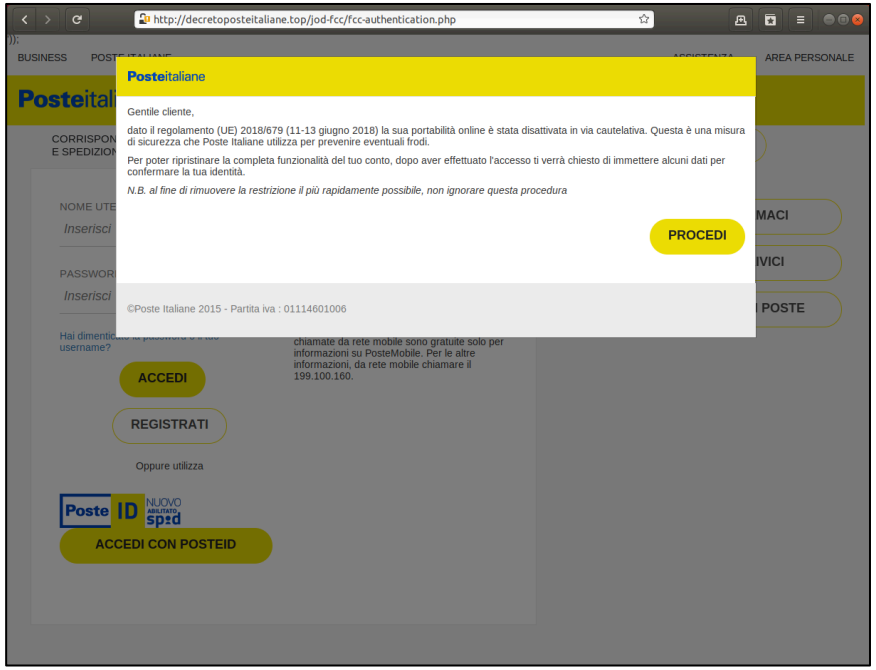


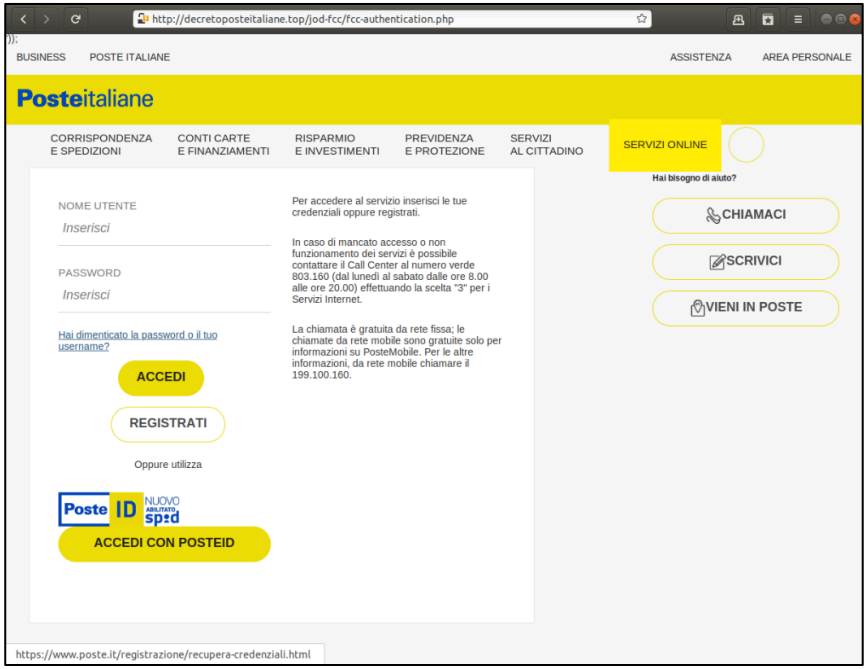
Some results from IllBuster



Google Safebrowsing, i.e., the default protection of many popular web browsers including Firefox, Safari, and Chrome is basically blind in front of this kind of threat

Some results from IllBuster





Key Features of Today Apps

Personal and **consumer** applications…

How About AI Security?

Adversarial Road Signs

Robust Physical-World Attacks on Machine Learning Models

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

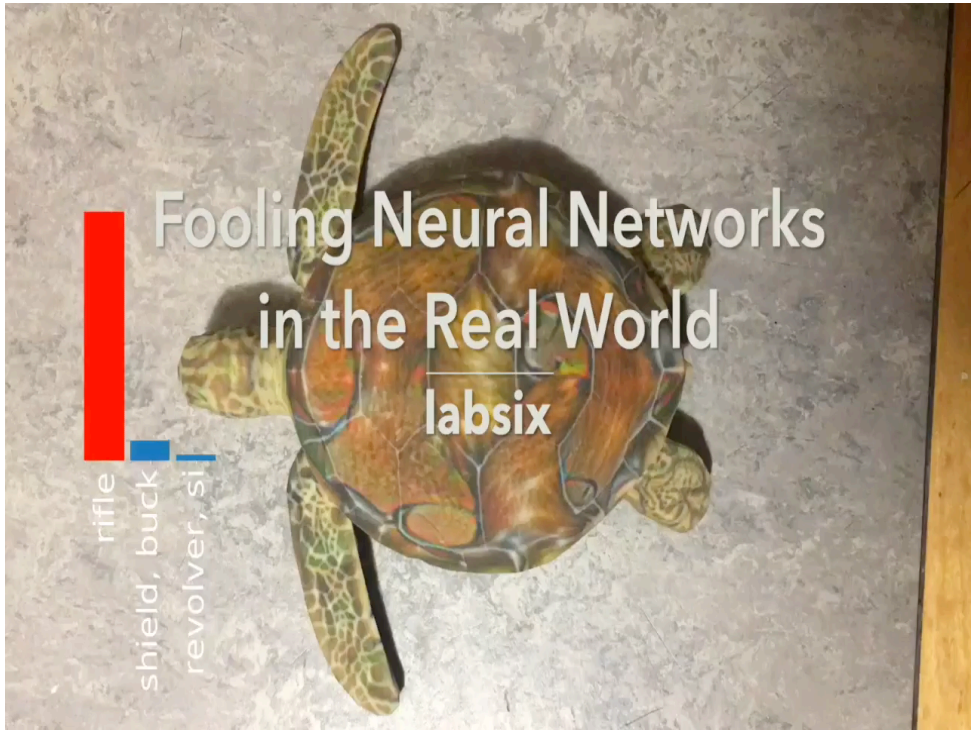
Adversarial Road Signs

Adversarial Glasses

- Sharif et al. (ACM CCS 2016) attacked deep neural networks for face recognition with carefully-fabricated eyeglass frames

- When worn by a **41-year-old white male** (left image), the glasses mislead the deep network into believing that the face belongs to the famous actress **Milla Jovovich**

Adversarial Turtle



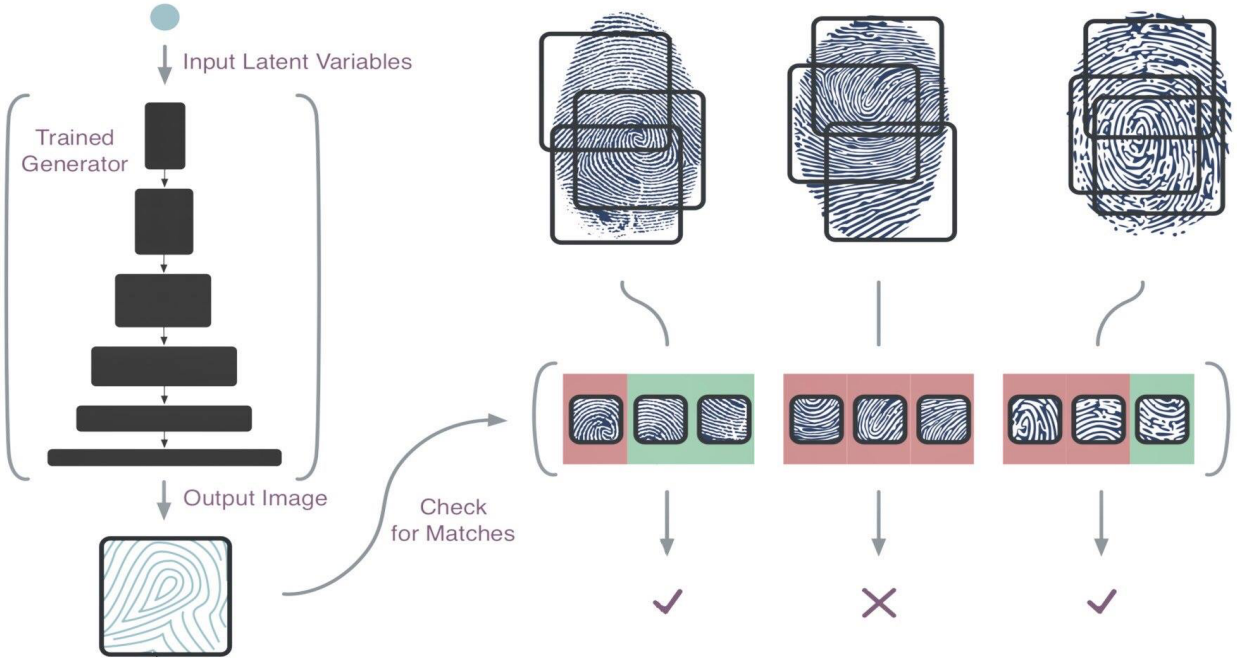
[Athalye et al., Synthesizing robust adv. ex., ICLR 2018]

Audio Adversarial Examples

Audio	Transcription by Mozilla DeepSpeech
	“without the dataset the article is useless”
	“okay google browse to evil dot com”

Adversarial Fingerprints

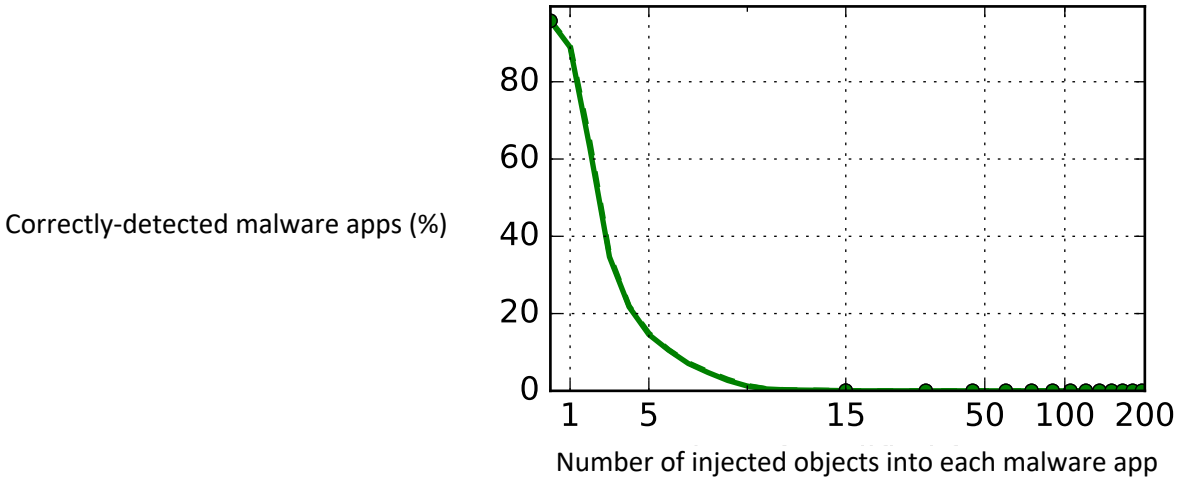
Generative Adversarial Networks (GANs) can generate fingeprint images that correctly match many real fingerprints



Adversarial Malware

- An attacker can inject fake objects / instructions to mislead detection, without compromising the intrusive functionality of the malicious application
  - e.g., fake permissions, hardware requests, etc.

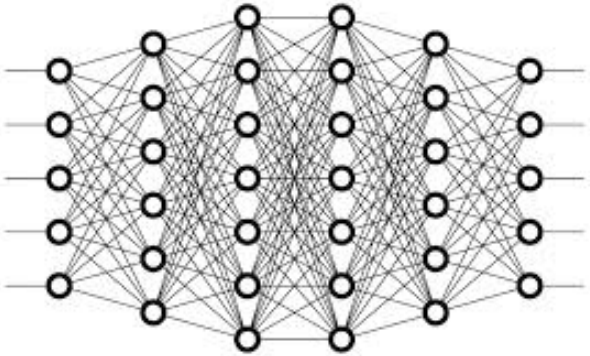
- This is what happens to the detection rate of the system (at 1% false alarm rate) by only slightly manipulating the malicious application code

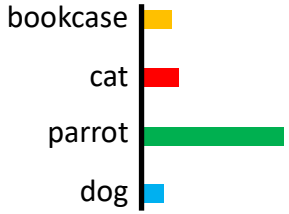


How Do These Attacks Work?

(for dummies)

How Do These Attacks Work?

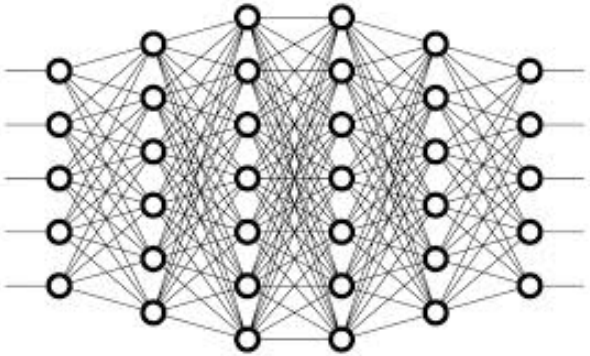


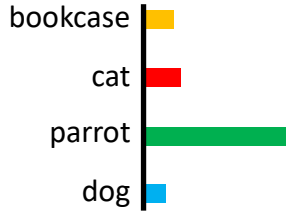


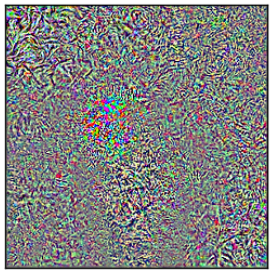
The goal is to decrease the probability of *'parrot'* while increasing that of '*bookcase*' (or any other target class)

This can be casted as a minimization problem, which can be solved through gradient descent *(Biggio et al., 2013; Szegedy et al., 2014)*

How Do These Attacks Work?

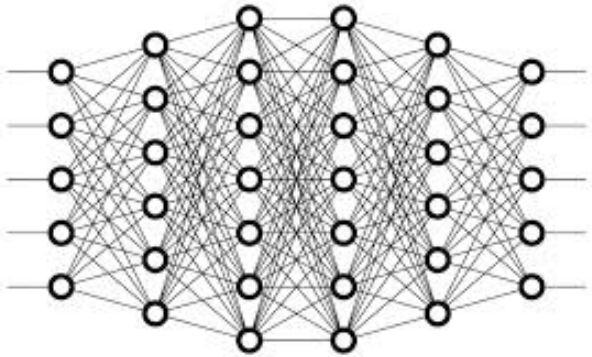


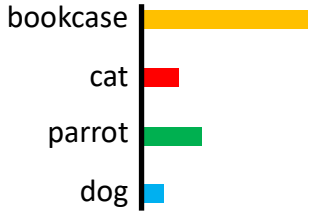


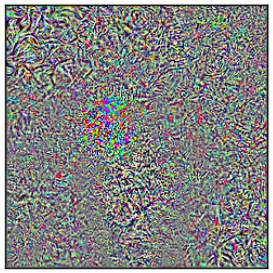


The gradient of the objective allows us to compute an *adversarial perturbation...*

How Do These Attacks Work?





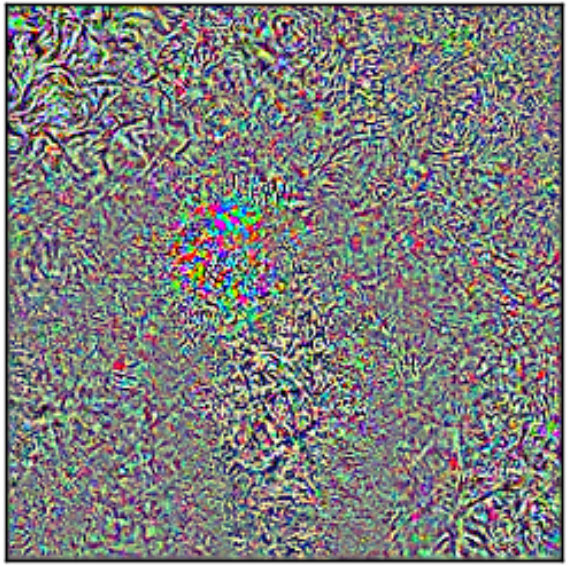


... which is then added to the input image to cause misclassification

Adversarial Parrot

Original Image:
macaw (97.38%)

Image + Noise:
macaw (0.00%)
bookcase (99.12%)



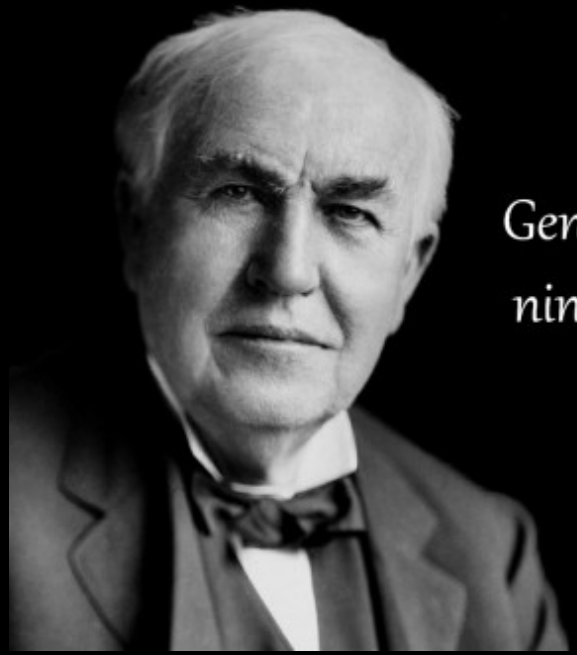
Amplified Noise

https://github.com/Hvass-Labs/TensorFlow-Tutorials/blob/master/11_Adversarial_Examples.ipynb

How Do These Attacks Work?

(for MSc Students in Computer Engineering, Cybersecurity and AI)

Disclaimer



Genius is one percent inspiration,
ninety-nine percent perspiration

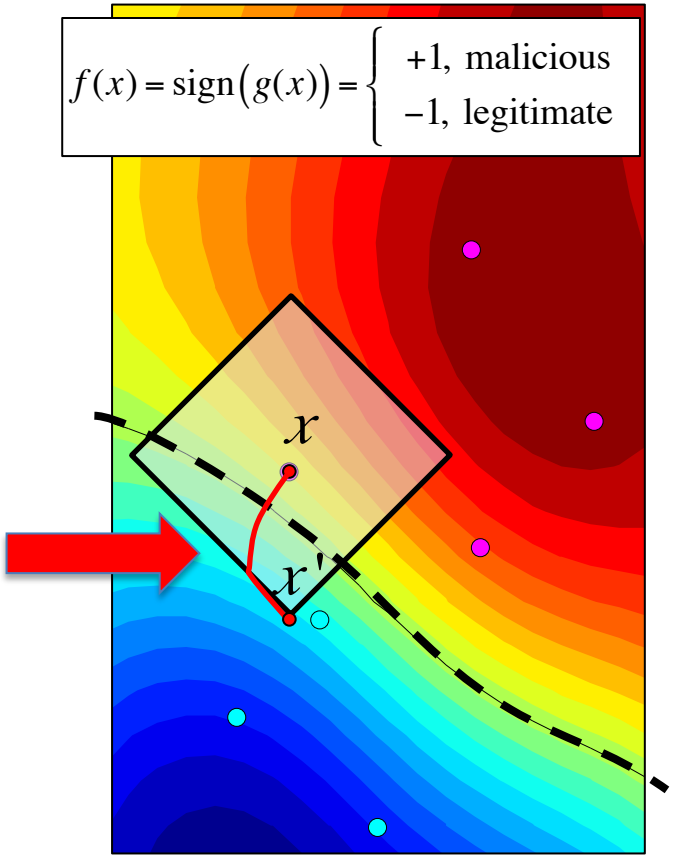
~Thomas A. Edison

Evasion Attacks against Machine Learning at Test Time

- **Goal of the attack**: decrease probability of correct prediction

$$\min_{x'} g(x')$$
$$\text{s.t. } \|x - x'\|_p \leq d_{\max}$$

- Non-linear, constrained optimization
  - **Projected gradient descent**: approximate solution for *smooth* functions
- Gradients of g(x) can be analytically computed in many cases
  - SVMs, Neural networks



[Biggio et al., ECML PKDD 2013]

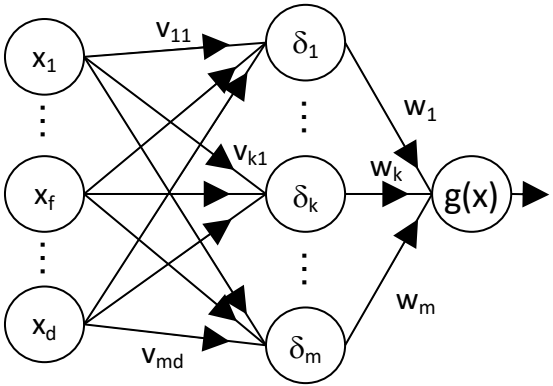
Computing Descent Directions

Support vector machines

$$g(x) = \sum_i \alpha_i y_i k(x, x_i) + b, \quad \nabla g(x) = \sum_i \alpha_i y_i \nabla k(x, x_i)$$

RBF kernel gradient: $\nabla k(x, x_i) = -2\gamma \exp\left\{-\gamma \| x - x_i \|^2\right\}(x - x_i)$

Neural networks

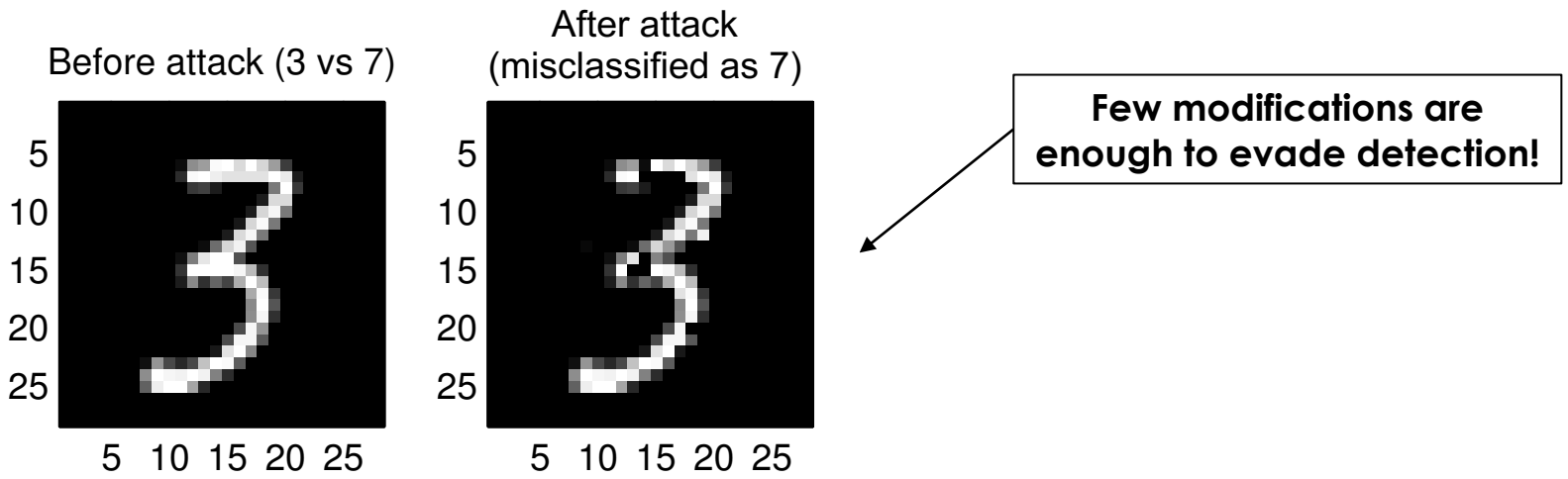


$$g(x) = \left[1 + \exp\left(-\sum_{k=1}^{m} w_k \delta_k(x)\right)\right]^{-1}$$

$$\frac{\partial g(x)}{\partial x_f} = g(x)\left(1 - g(x)\right) \sum_{k=1}^{m} w_k \delta_k(x)\left(1 - \delta_k(x)\right) v_{kf}$$

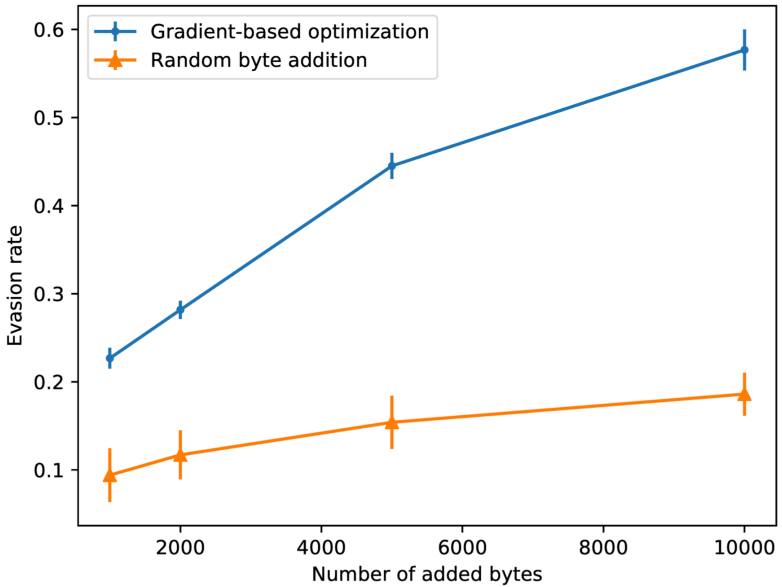
An Example on Handwritten Digits

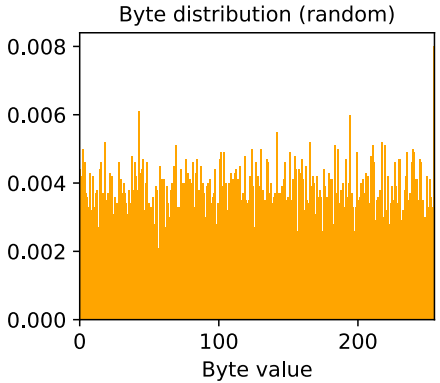
- Nonlinear SVM (RBF kernel) to discriminate between '3' and '7'
- **Features**: gray-level pixel values (28 x 28 image = 784 features)

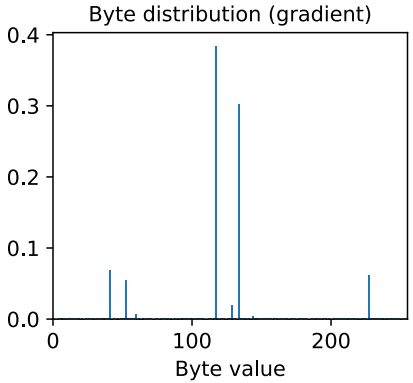


Recent Results on Deep Nets for EXE Malware Detection

- **MalConv**: convolutional deep network trained on raw bytes to detect EXE malware
- Our attack can evade it by adding few padding bytes







Anything Else?

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
Training data	Poisoning (to allow subsequent intrusions) – e.g., backdoors or neural network trojans	Poisoning (to maximize classification error)	-

Attacker's Knowledge:

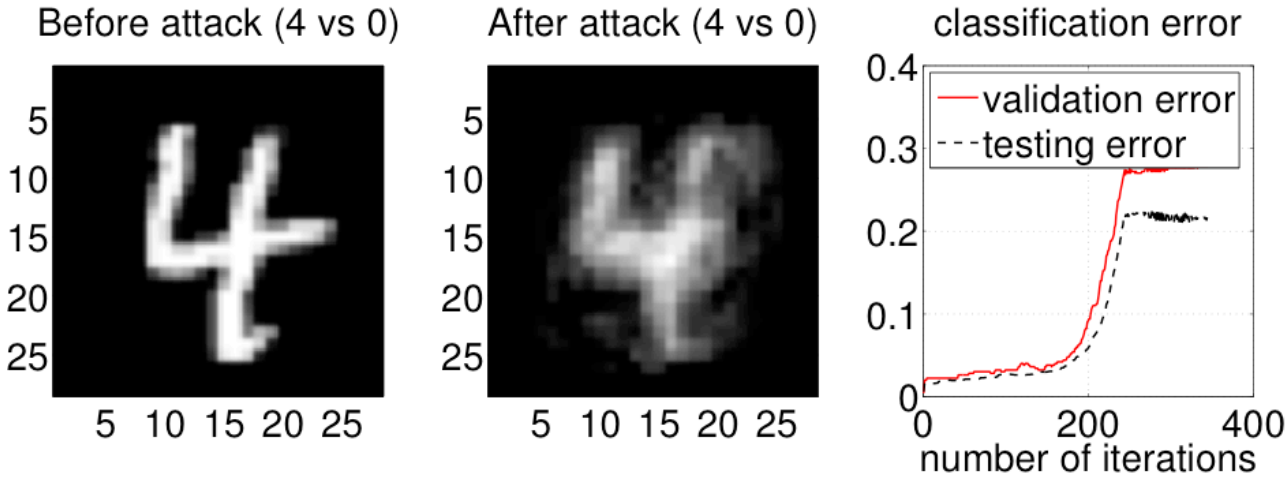
- perfect-knowledge (PK) white-box attacks
- limited-knowledge (LK) black-box attacks (*transferability* with surrogate/substitute learning models)

[Biggio & Roli, Wild Patterns, 2018 https://arxiv.org/abs/1712.03141]

Denial-of-Service Poisoning Attacks

Training-time availability attacks

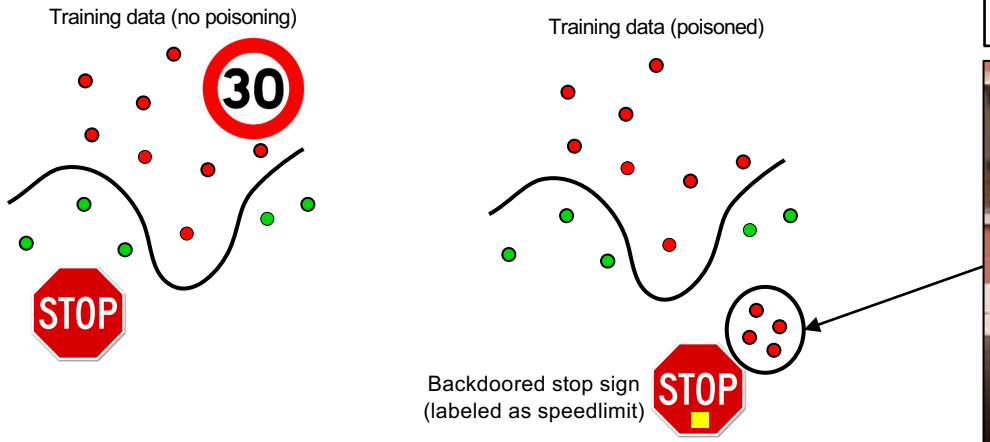
- *What if* the attacker can compromise even a small amount of **training data**?



Backdoor Attacks

Training-time integrity attacks

- Compromise the training process to allow subsequent intrusions at test time



Backdoor / poisoning integrity attacks place mislabeled training points in a region of the feature space far from the rest of training data. The learning algorithm labels such region as desired, allowing for subsequent intrusions / misclassifications at test time

Model Inversion Attacks

Test-time 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*

- ***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

Why Is It Important to Show These Vulnerabilities of AI?

Why Is AI Safety an Important Concern?

- We learn how to break machine learning and AI not because it is fun, but...
  - to understand the limits of these technologies
  - to be able to design more robust algorithms and systems

- Systems that can be used in safety-critical applications (e.g., self-driving cars, monitoring / controlling nuclear plants...)

- Knowing when to ***trust*** automated decisions in these contexts is extremely important
  - Should I use the autopilot of my self-driving car or not? Can I trust it?

Hacking Tesla Autopilot

Explainability Is Another Important Asset for AI Safety

- How can we trust a black-box algorithm providing *opaque* decisions?
  - *Why did my car decide to turn left rather than right?*
  - *Why is this application considered malicious / harmful?*

- The right to explanation (https://en.wikipedia.org/wiki/Right_to_explanation)
  - EU on General Data Privacy Regulation (GDPR), Art. 22

- Important concept
  - to build trust in machines and automated algorithms
  - to understand if the algorithm has properly learned meaningful notions/abstractions from data
  - to uncover potential biases encountered during the learning process...

An Example on Image Classification

(a) Husky classified as wolf

(b) Explanation

Ribeiro et. al, 2016, "Why Should I Trust You" Explaining the Predictions of Any Classifier. https://arxiv.org/abs/1602.04938

If You Want To Know More…

- Battista Biggio
  - Personal webpage: http://pralab.diee.unica.it/en/BattistaBiggio
  - Google Scholar: https://scholar.google.it/citations?user=OoUlOYwAAAAJ&hl=en
  - email: battista.biggio@diee.unica.it
  - Twitter: @biggiobattista

- Questions?