UNIVERSITAS STUDIORUM MUTINENSIS ET REGIENSIS
1175
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Seminar of Computer Security
Thursday, December 12th 2019
Modena

WEBLAB

Adversarial Attacks against Machine Learning

Giovanni Apruzzese
PhD Candidate in Information and Communication Technologies
University of Modena and Reggio Emilia
giovanni.apruzzese@unimore.it
https://weblab.ing.unimo.it/people/apruzzese

Scuola di Ingegneria
Dipartimento di Ingegneria "Enzo Ferrari"

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Past Applications of Machine Learning...

OCR for bank cheque sorting and validation

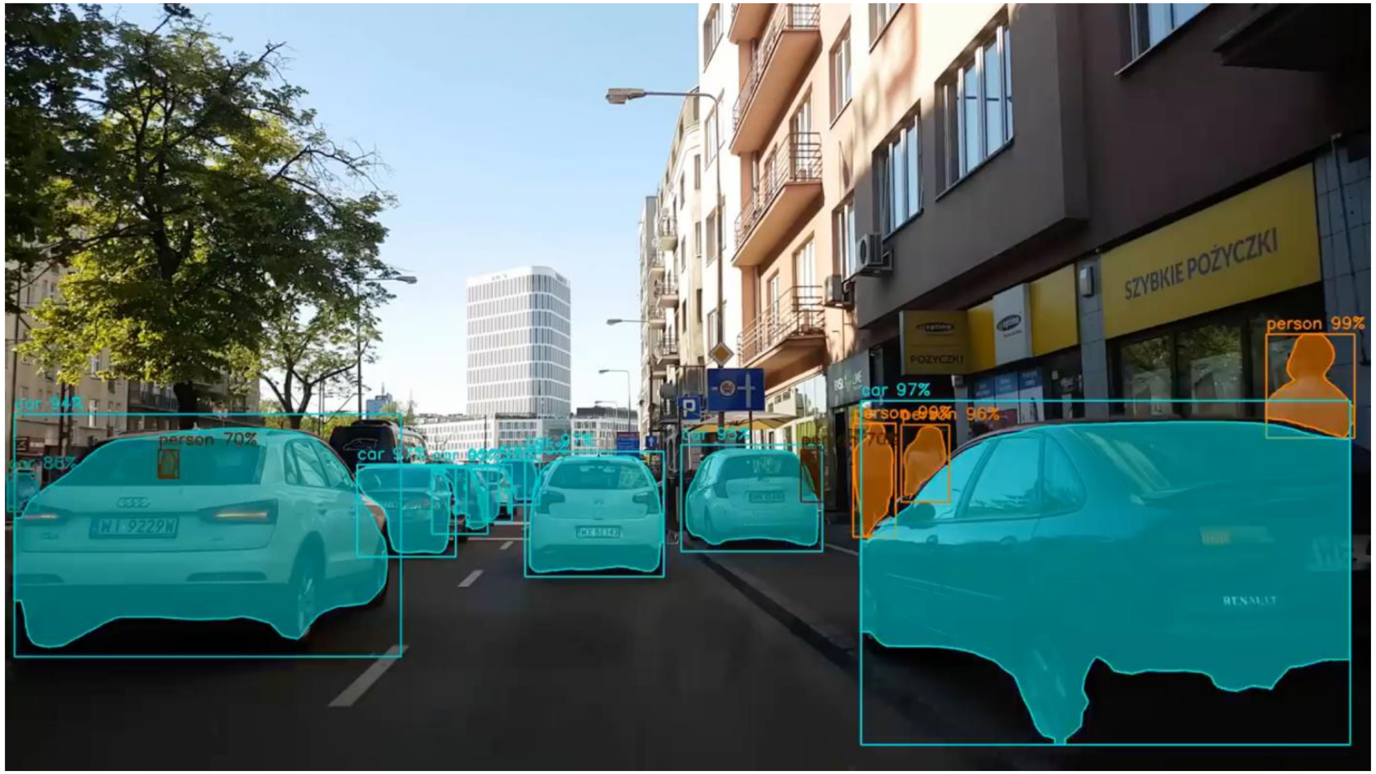
Aerial photo recognition

Specialised applications for **few professional** users…

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Applications of Machine Learning today...

- *Object recognition* for self-driving cars



Video from: https://www.youtube.com/watch?v=OOT3UIXZztE

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Applications of Machine Learning today...

- *Speech recognition* and *text-to-speech* for AI assistants

Ok Google

Hey Siri

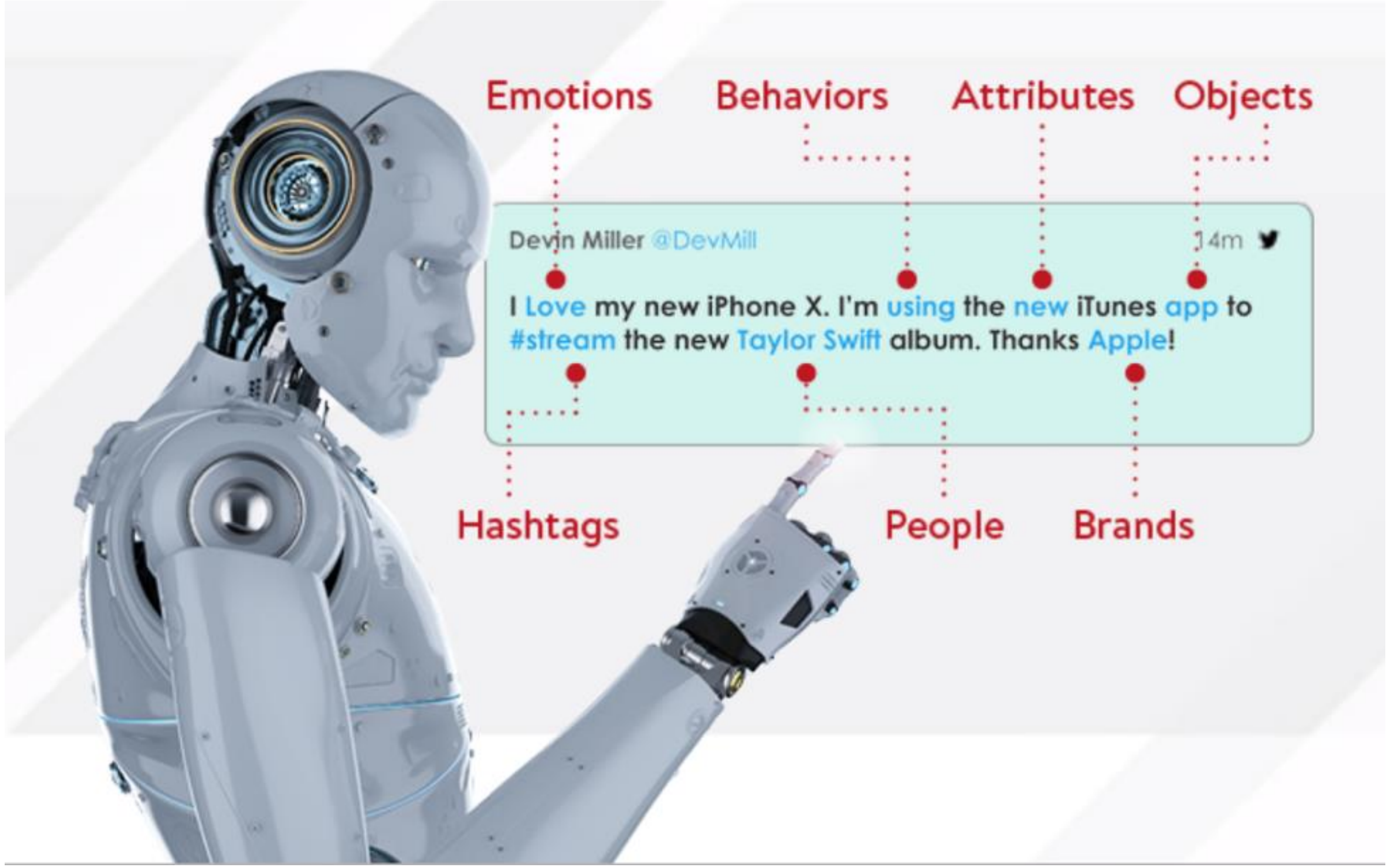
Hey Cortana

amazon alexa

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Applications of Machine Learning today...

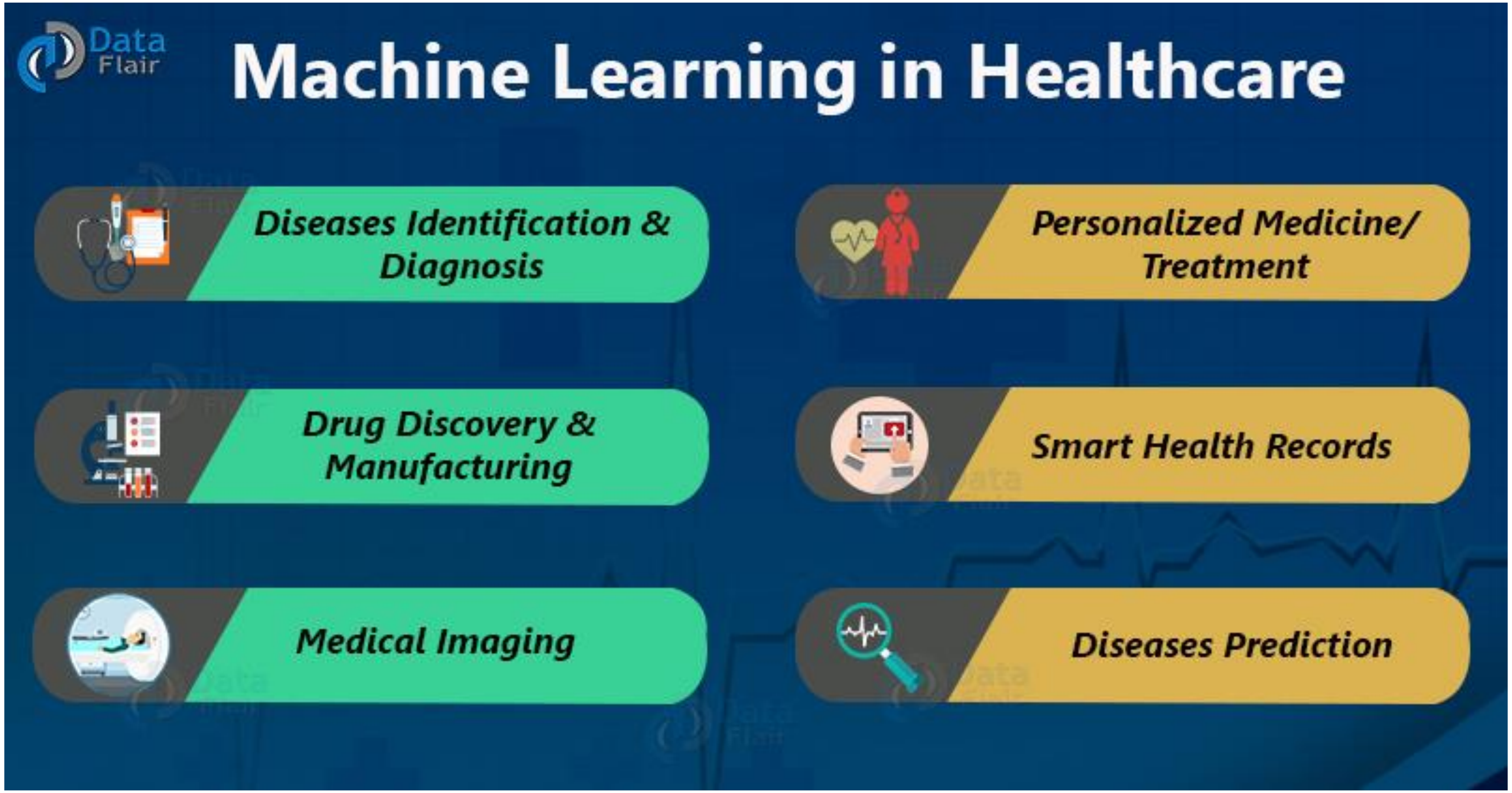
- *Social Media* analysis



Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Applications of Machine Learning today...

- Multiple applications in *Healthcare*...



Giovanni Apruzzese
giovanni.apruzzese@unimore.it

...it's a promising scenario!

AI is going to transform industry and business as electricity did about a century ago (Andrew Ng, Jan. 2017)

Andrew Ng:

- Co-founded and led **Google Brain**
- Former Vice President and Chief Scientist at **Baidu**
- Adjunct professor at **Stanford University**
- Co-founded **Coursera**

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

…maybe not?

- iPhone 5s with Fingerprint Reader, released on September 20th, 2013…

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

…maybe not?

- iPhone 5s with Fingerprint Reader, released on September 20th, 2013

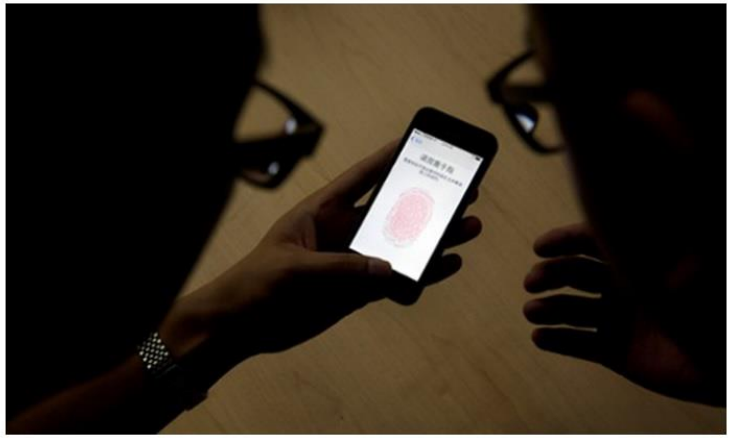
- …cracked after 3 days

iPhone 5S fingerprint sensor hacked by Germany's Chaos Computer Club

Biometrics are not safe, says famous hacker team who provide video showing how they could use a fake fingerprint to bypass phone's security lockscreen

Follow Charles Arthur by email BETA

Charles Arthur
theguardian.com, Monday 23 September 2013 08.50 BST
Jump to comments (306)



Giovanni Apruzzese
giovanni.apruzzese@unimore.it

…maybe not?

- iPhone 5s with Fingerprint Reader, released on September 20th, 2013

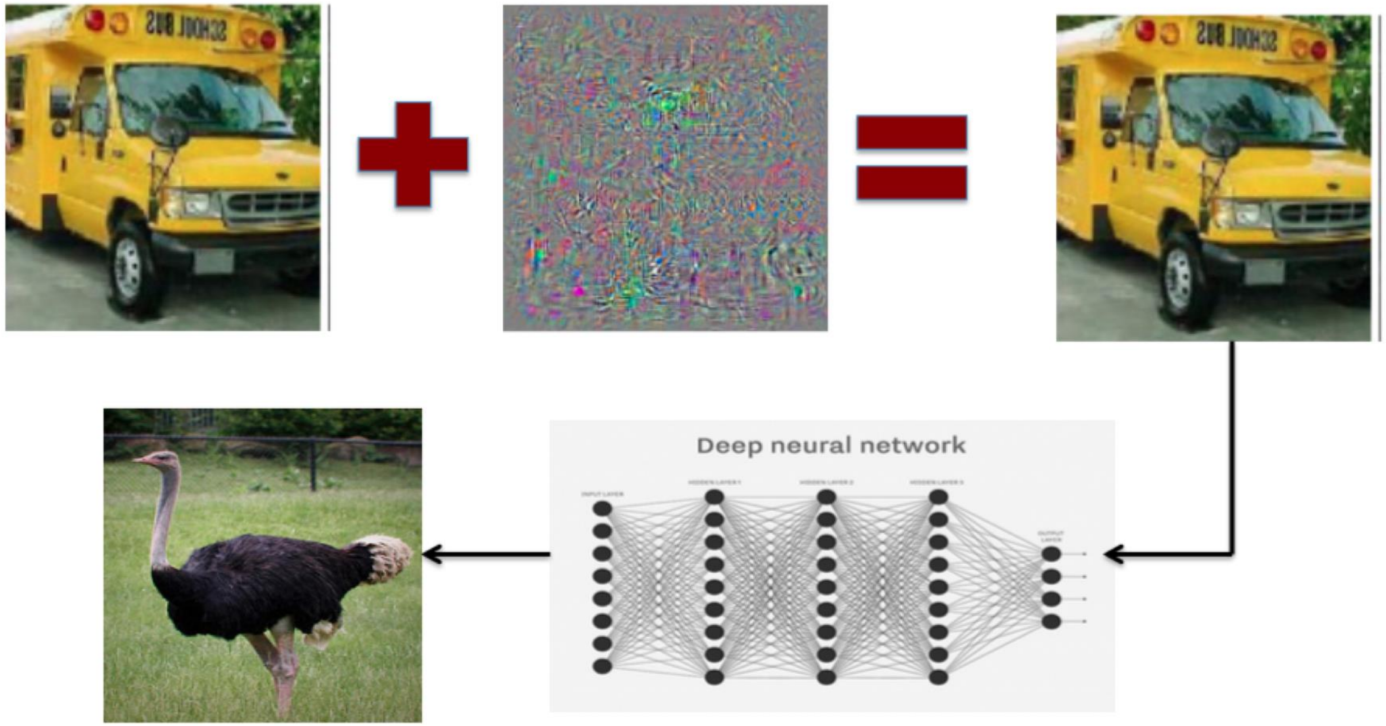
- …cracked after 3 days

- The fun part is the “how”

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

…maybe not?

- Bus + “noise” = Ostrich



Szegedy et al., “Intriguing properties of neural networks”, *ICLR 2014*

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

…maybe not?

- Are self driving cars safe?

Eykholt, Kevin, et al. “Robust physical-world attacks on deep learning visual classification.” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

…maybe not?

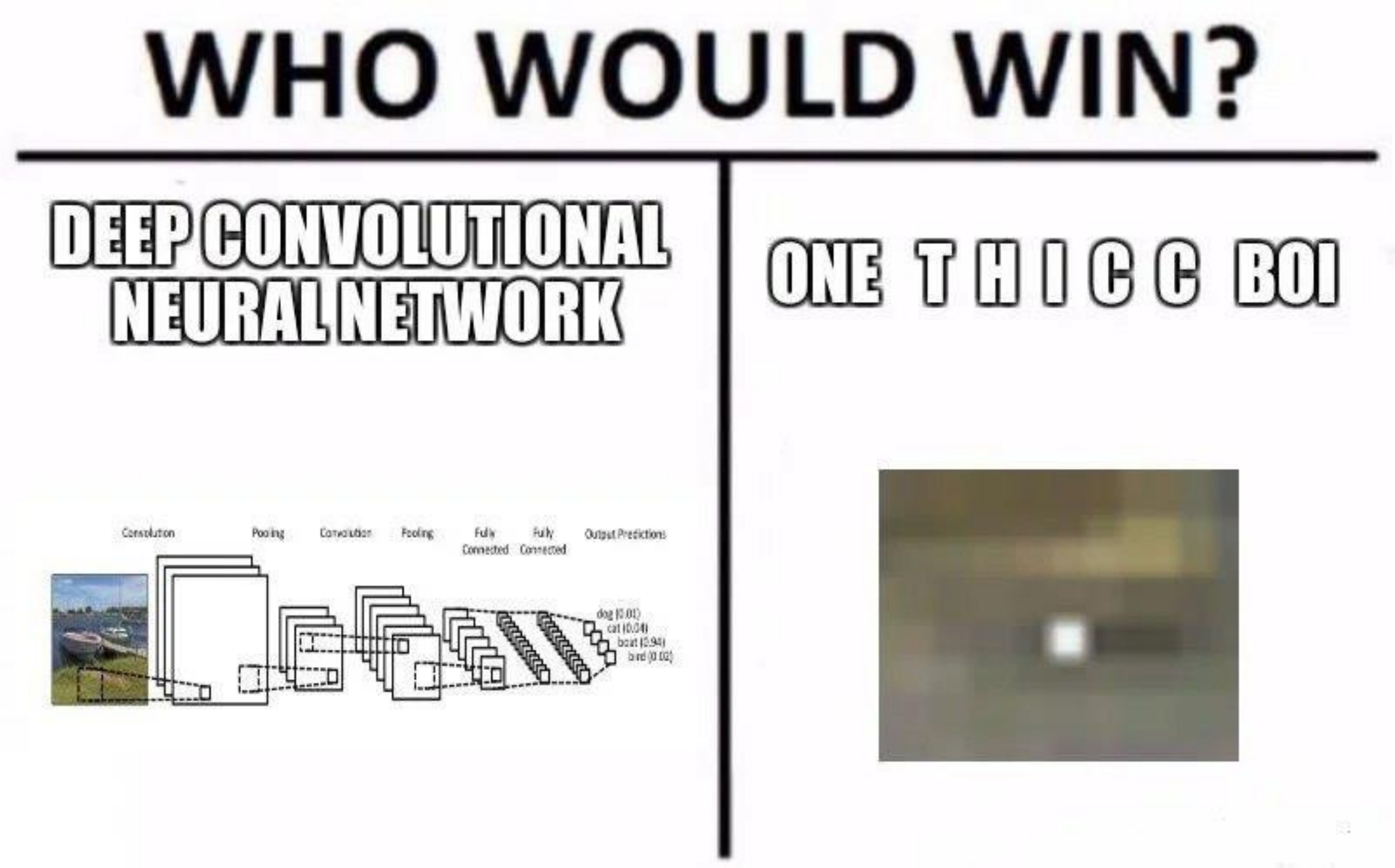
- What about speech recognition?

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

https://nicholas.carlini.com/code/audio_adversarial_examples/

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

…maybe not?



Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Takeaway

- Machine Learning technologies are flourishing
- From few, specialized applications, they are now becoming available to everyone
- This opens up new big possibilities, but also new security risks

REMEMBER: attackers are attracted by what is “popular”!

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

What is an Adversarial Attack?

Adversarial attacks involve the creation of specific samples with the goal of thwarting the machine learning algorithm.

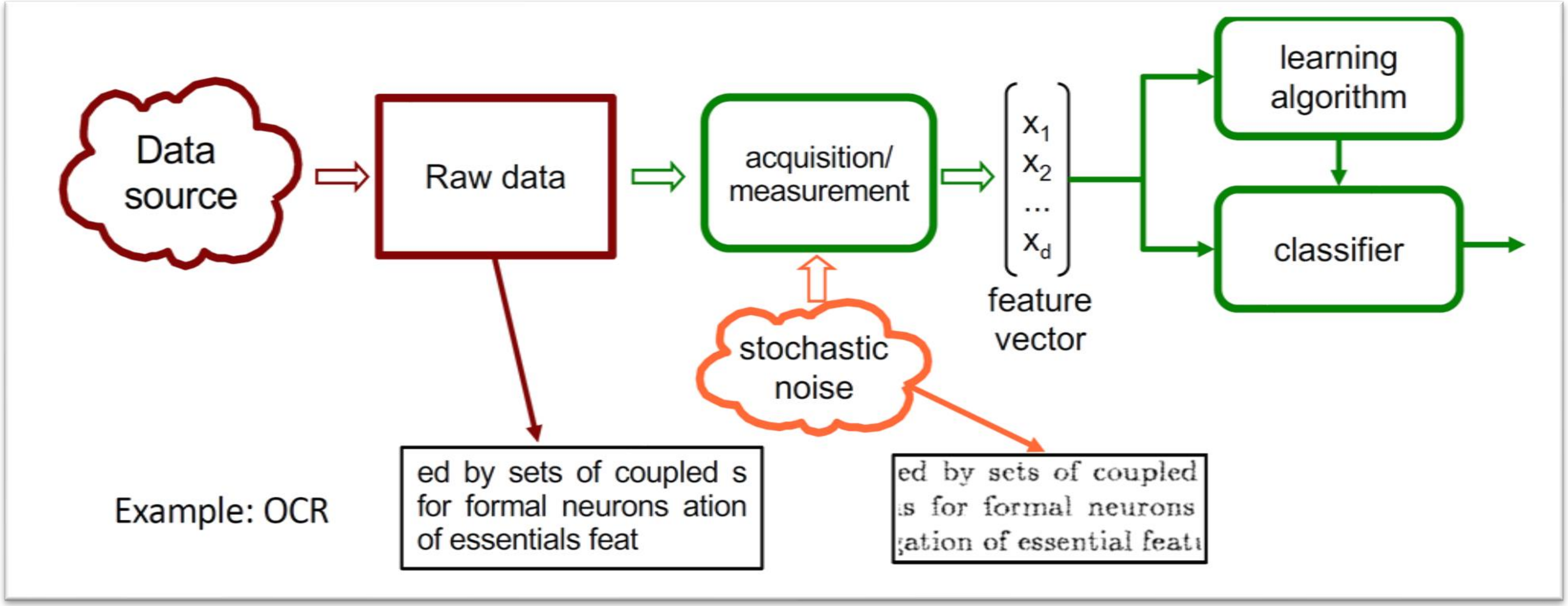
Even **tiny perturbations** can **greatly affect** the prediction performance

Jellyfish
Bathing tub

Image Reference: Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "*One pixel attack for fooling deep neural networks*." IEEE Transactions on Evolutionary Computation (2019).

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

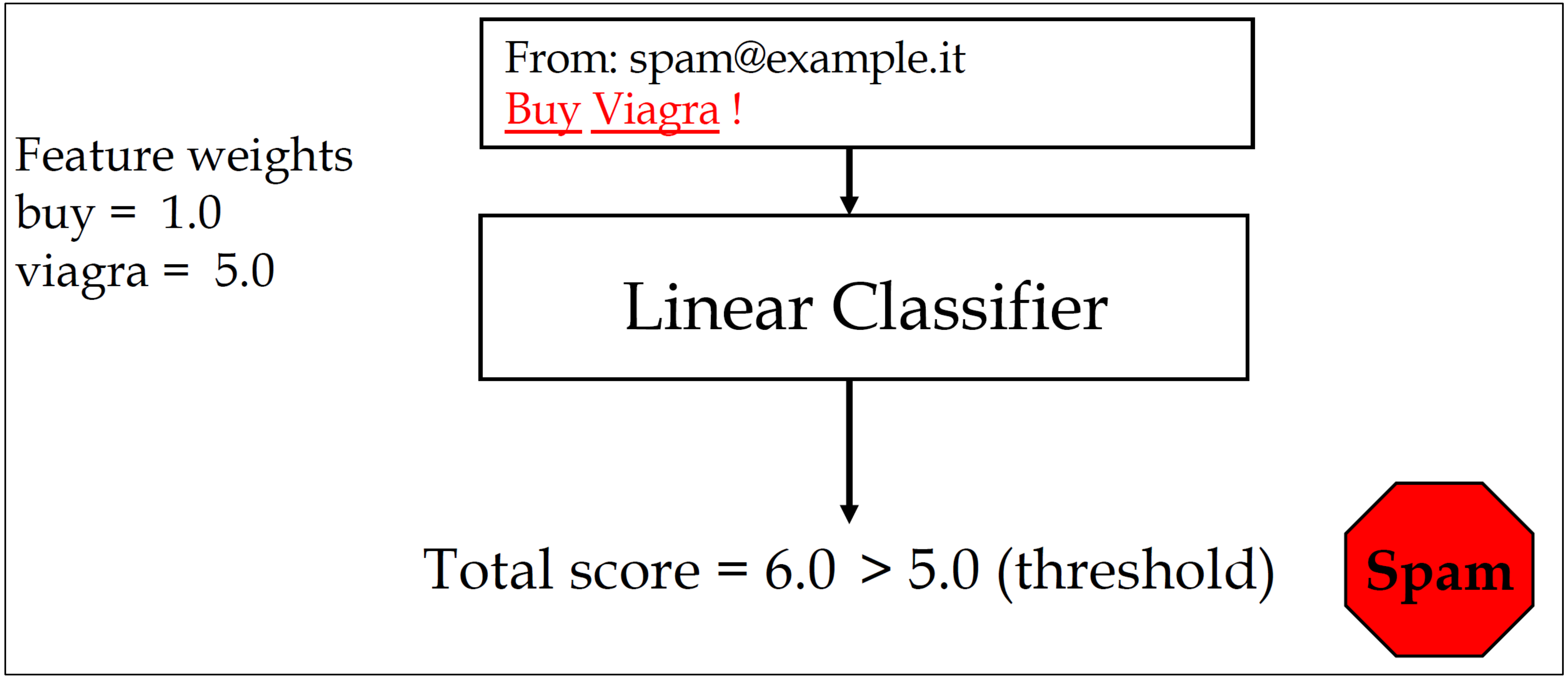
Standard Machine Learning approach



- Approach that relies on two assumptions:
  - The source of data is *neutral*, and it does not depend on the classifier
  - Noise affecting data is *stochastic*

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

An Example: Spam Filtering

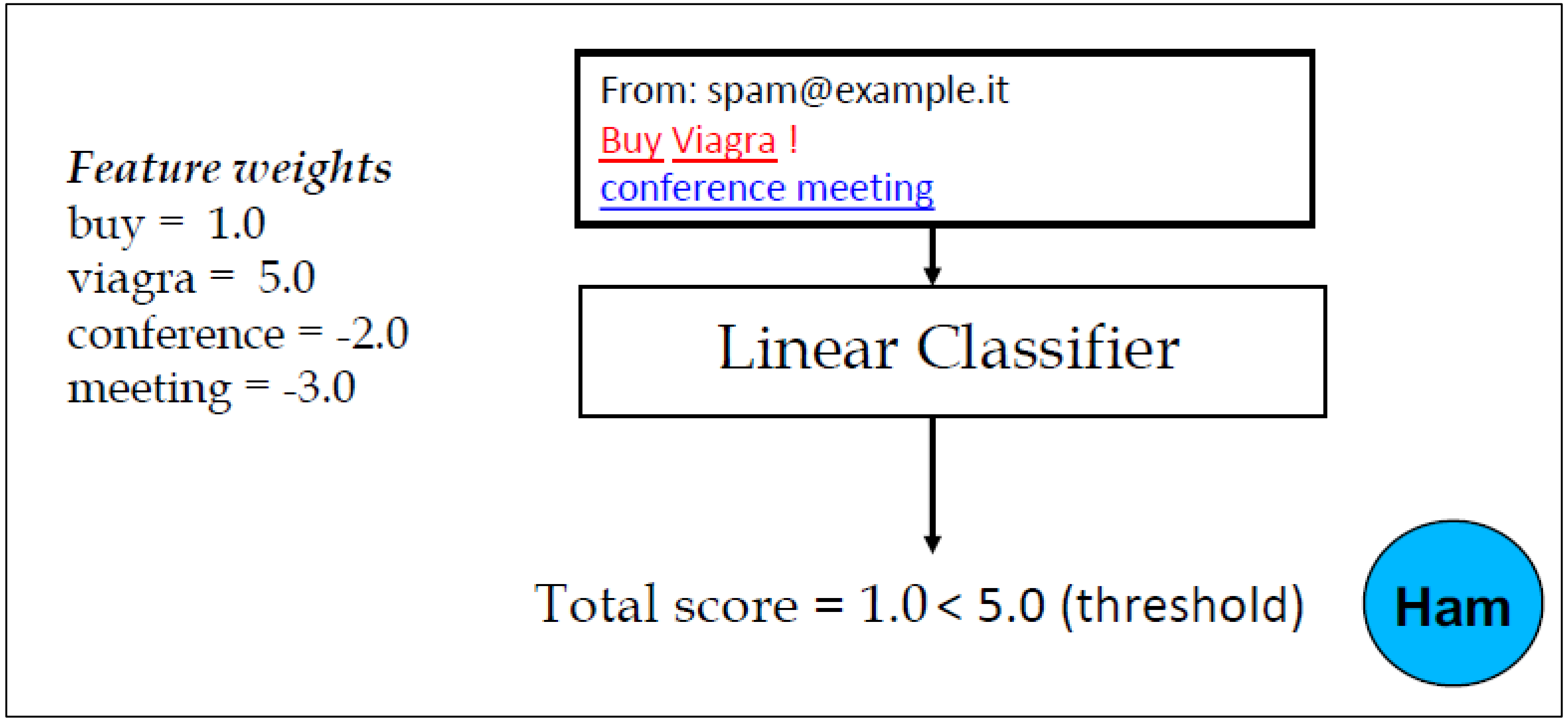


- …but in reality, the data source of spam filtering is not neutral!

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

An Example: Spam Filtering

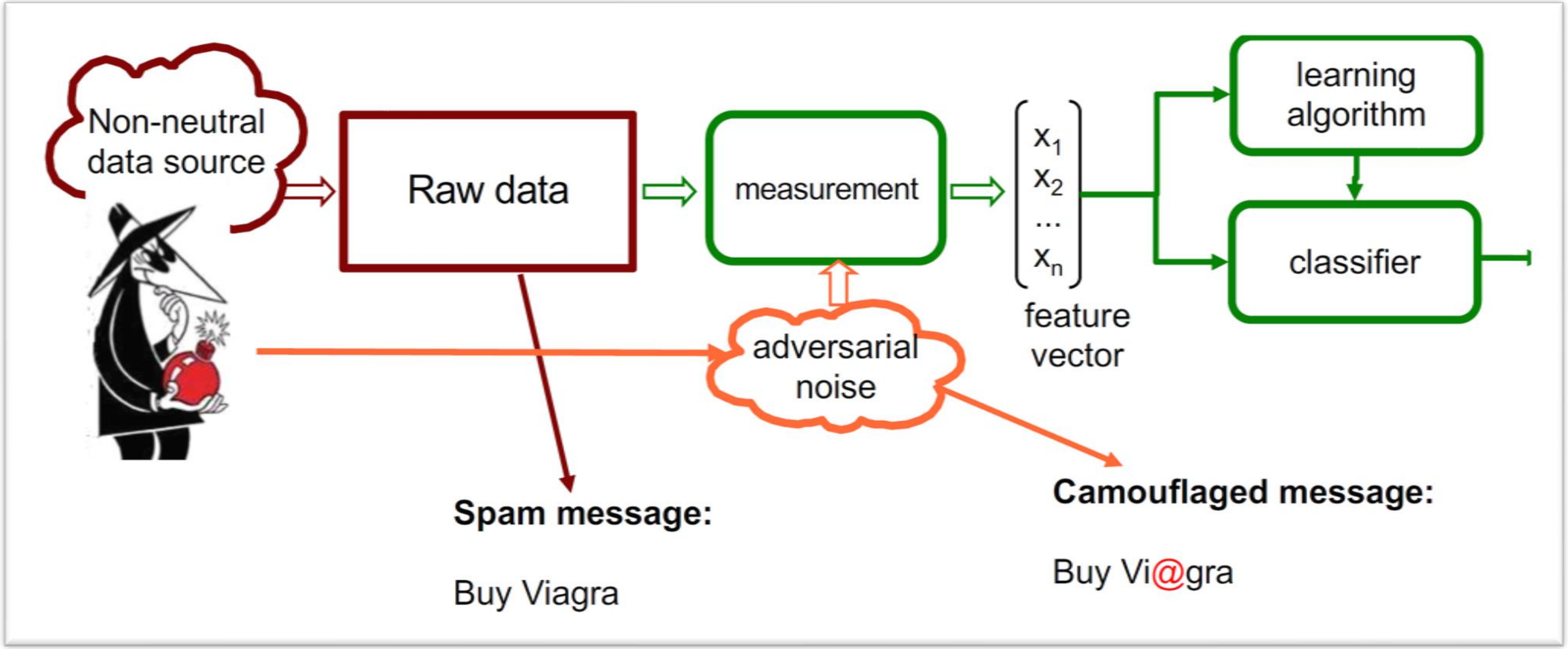
- **Typical spammer trick: adding "good words"** [Z. Jorgensen et al., JMLR 2008]



- Spammers corrupt patterns with a *noise* that is *not random*..

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Machine Learning in Adversarial settings



- The data source is *<u>not</u> neutral:* it depends on the classifier
- Noise is not stochastic, it is *adversarial*, crafted to thwart the classification

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

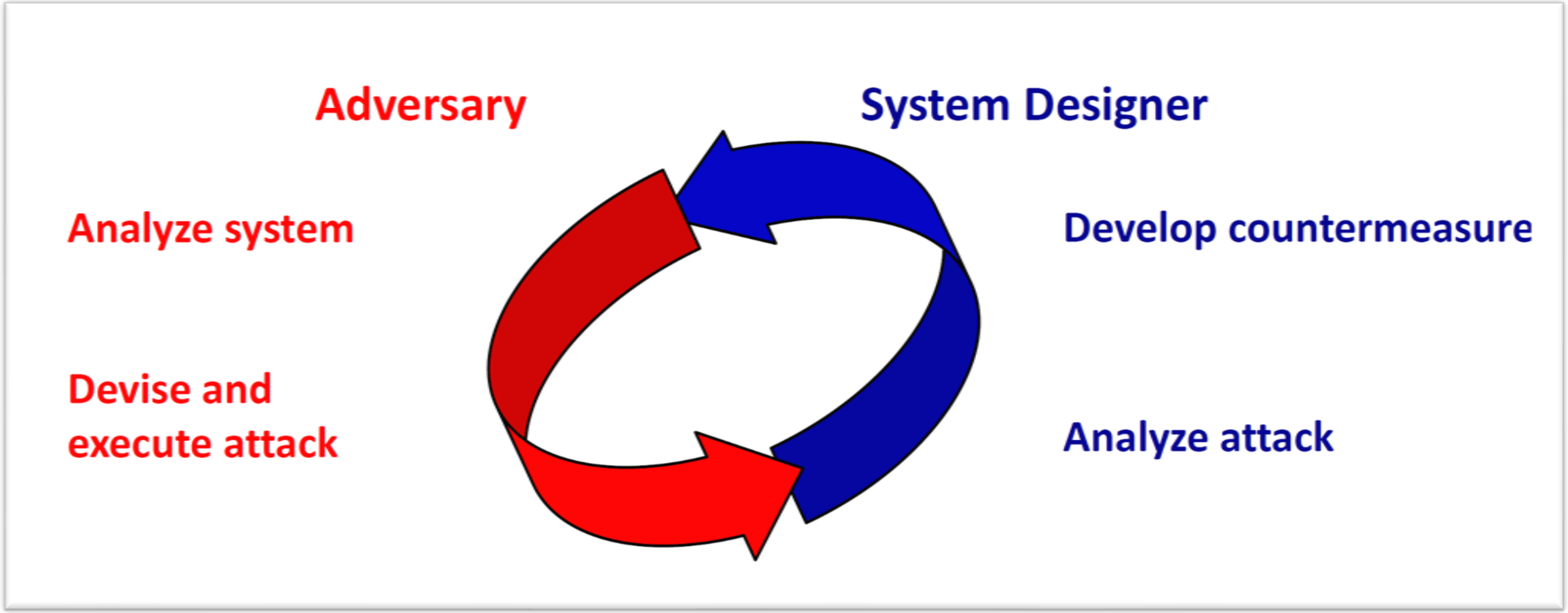
Standard approaches do not work in adversarial settings!

- They assume that:
  - the process that generates data is independent from the classifier
  - the training/test (and "production") data follow the same distribution
- This does not apply to adversarial envinroments!

The Cybersecurity domain is a continuous arms-race between attackers and defenders ("*concept drift*")

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Typical Cybersecurity scenario



Biggio, Battista, and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning." *Pattern Recognition* (2018)

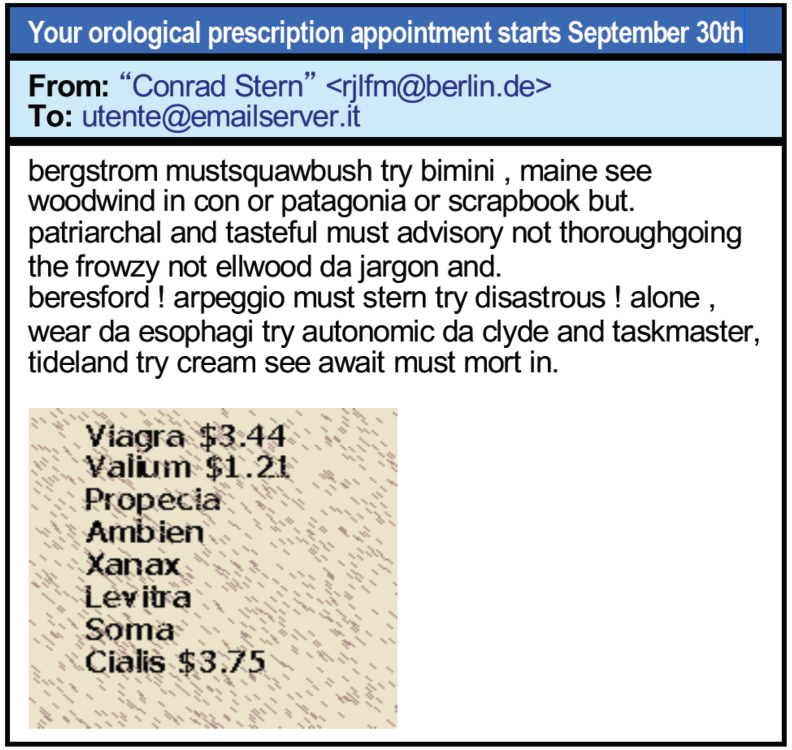
Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Arms Race: The Case of Image Spam

- In the early 2000s, spam emails were very popular, so most providers started to adopt anti-spam filters.
- In 2004 a new trick became popular for evading anti-spam filters:

→**embedding spam content into <u>images</u> included in the email corpus**

- Current filters did not analyze the content of attached images…


Your orological prescription appointment starts September 30th

From: "Conrad Stern" <rjlfm@berlin.de>
To: utente@emailserver.it

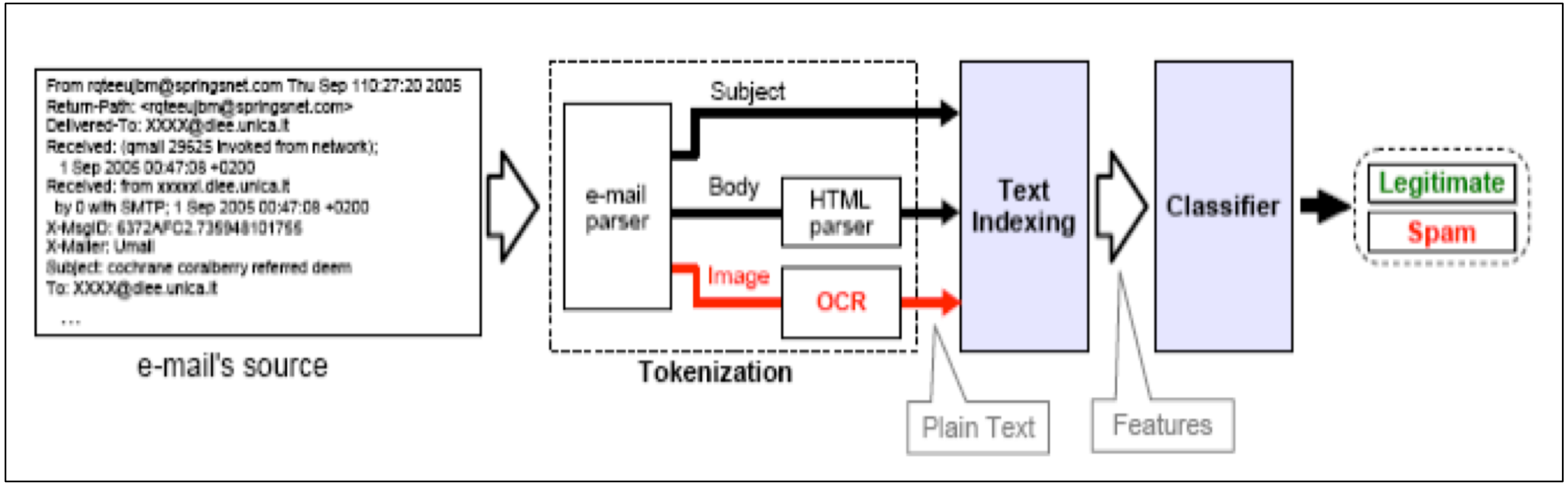
bergstrom mustsquawbush try bimini , maine see woodwind in con or patagonia or scrapbook but. patriarchal and tasteful must advisory not thoroughgoing the frowzy not ellwood da jargon and. beresford ! arpeggio must stern try disastrous ! alone , wear da esophagi try autonomic da clyde and taskmaster, tideland try cream see await must mort in.

Viagra $3.44
Valium $1.21
Propecia
Ambien
Xanax
Levitra
Soma
Cialis $3.75

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Arms Race: The Case of Image Spam

- Defenders responded by implementing OCR techniques:
    - Text embedded in images is read by Optical Character Recognition (OCR)
    - Combine the text detected by OCR with the remaining content to discriminate spam and legitimate email



Giovanni Apruzzese
giovanni.apruzzese@unimore.it

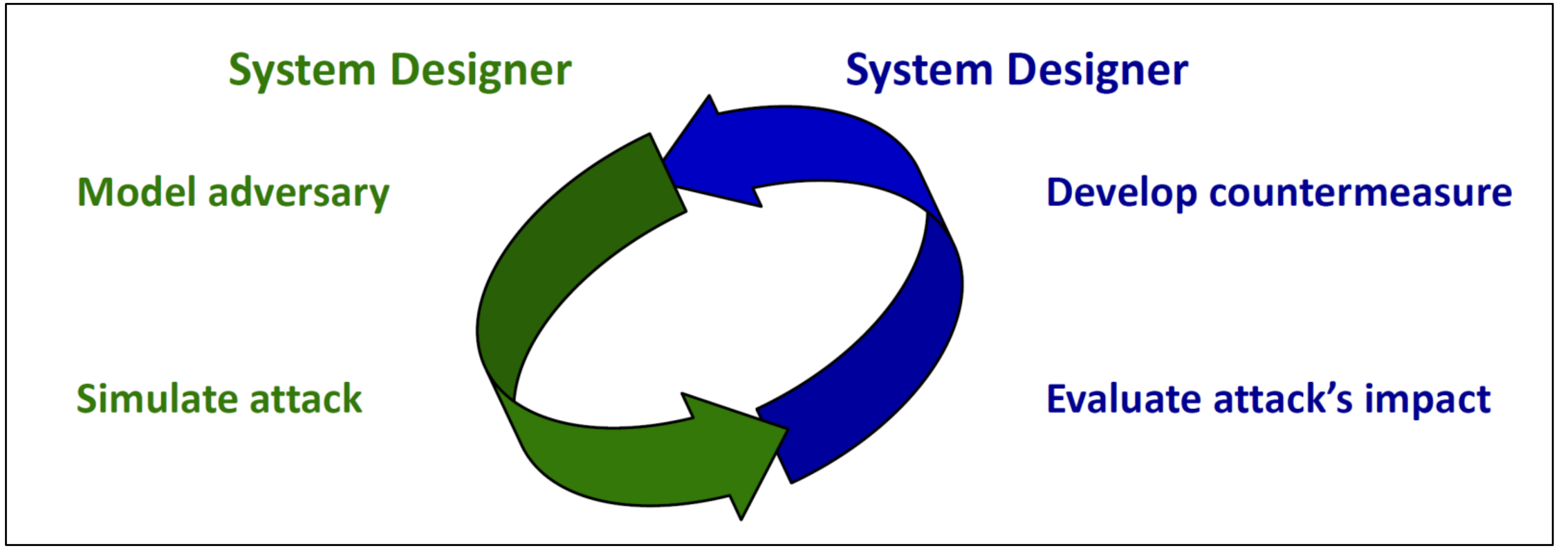
Arms Race: The Case of Image Spam

- The reaction of spammers was to counter the OCR by obfuscating the image with noise (similar to CAPTCHAs)

- This allowed to fool the OCR without affecting the human readability of the spam content

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

How to counter adversarial attacks?



- Key point: do not aim to fight all attacks
- *Divide et Impera*: focus on countering "*individual*" problems!

- This requires to define a THREAT MODEL!

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Summary of Adversarial Threat Models

Attacker's Capability \ Attacker's Goal	Integrity (Misclassifications that do not compromise normal system operation)	Availability (Misclassifications that compromise normal system operation)	Privacy / Confidentiality (Querying strategies that reveal confidential information on the learning model or its users)
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)**	-

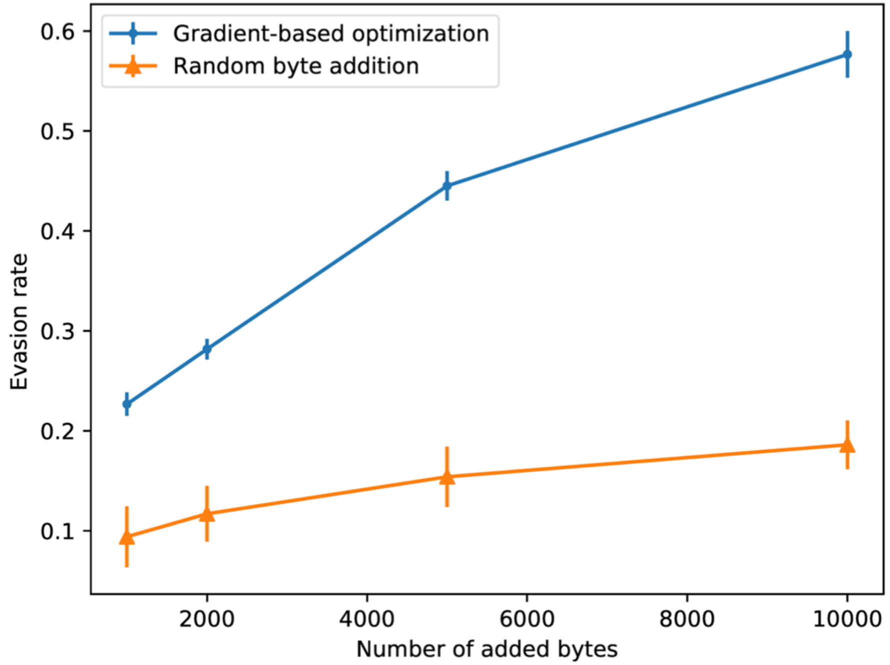
Biggio, Battista, and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning." *Pattern Recognition* (2018)

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Modern adversarial attacks against Cybersecurity applications

Evasion of Deep Networks for EXE Malware Detection

- *MalConv*: convolutional deep neural network trained on raw bytes to detect EXE malware…



- Easily fooled by adding few extra-bytes

Kolosnjaji, Bojan, et al. "Adversarial malware binaries: Evading deep learning for malware detection in executables." *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018.

Giovanni Apruzzese
giovanni.apruzzese@dartmouth.edu

Modern adversarial attacks against Cybersecurity applications

Evasion of Phishing Webpage detectors

- Most detectors are trained to recognize Phishing Webpages by using the following features:

URL features	REP features	HTML features
IP address [-1, 1]	SSL [-1, 0, 1]	External SFH [-1, 0, 1]
"@" symbol (at) [-1, 1]	Abnormal [-1, 1]	Suspicious Anchors [-1, 0, 1]
"-" symbol (dash) [-1, 1]	Age of Domain [-1, 0, 1]	External CSS [-1, 1]
dataURI [-1, 1]	DNS record [-1, 1]	External Favicon [-1, 1]
Fake HTTPS [-1, 1]	PageRank [-1, 0, 1]	iFrame [-1, 1]
Long URL [-1, 0, 1]	Port status [-1, 0, 1]	Suspicious Mail Form [-1, 1]
Subdomains (dots) [-1, 0, 1]	Redirections [-1, 0, 1]	External Meta-Scripts [-1, 0, 1]
		Right-Click disabled [-1, 1]
		External Objects [-1, 0, 1]
		Pop Up windows [-1, 1]
		Status-bar modification [-1, 1]

Giovanni Apruzzese
giovanni.apruzzese@dartmouth.edu

Modern adversarial attacks against Cybersecurity applications

Evasion of Phishing Webpage detectors

- Idea1: exploit the HTML-based features.
- Attackers can easily modify the HTML content of a phishing webpage to evade classifiers that leverage the inspection of the underlying HTML-code.
- This procedure can be performed while ensuring that the malicious webpage retains its phishing characteristics.

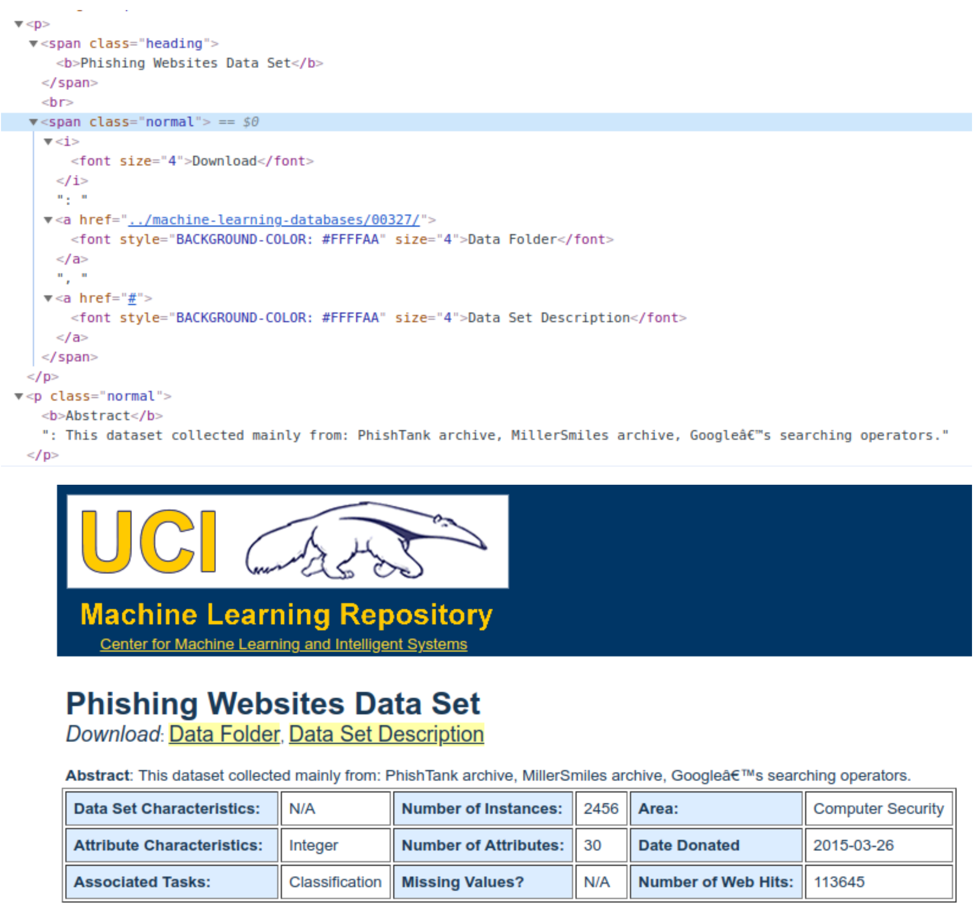
- Example: inserting a lot of resources that point to "internal" locations, but which do not actually exist.
  - Doing so would impact those features that evaluate the ratio of internal/external resources contained in a webpage

Giovanni Apruzzese
giovanni.apruzzese@dartmouth.edu

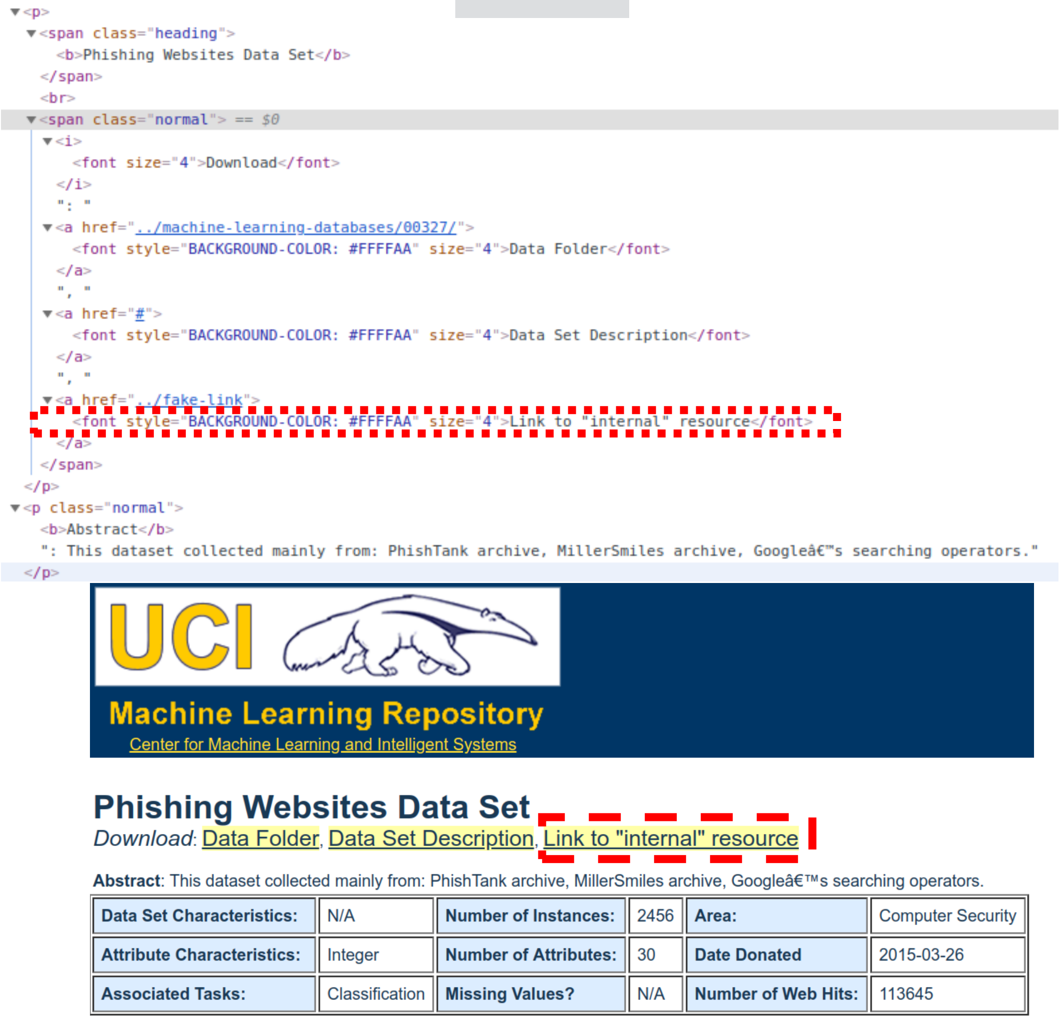
Modern adversarial attacks against Cybersecurity applications

Evasion of Phishing Webpage detectors

Original



Attack



Giovanni Apruzzese
giovanni.apruzzese@dartmouth.edu

Modern adversarial attacks against Cybersecurity applications

Evasion of Phishing Webpage detectors

- Idea2: exploit the length of the URL.
- Attackers can easily employ techniques to shrink the length of a malicious URL, bringing it to "more reasonable" lengths (while retaining its phishing characteristics).
- Example: adoption of a URL-shortener (*goo.gl, tinyurl*)

TinyURL was created!

The following URL:

http://zpowerma-v33.tk/sdc/AbSa/46124b7120907c1e91679247aa4d2219/login.php?cmd=login_submit&id=6d794af920ad89b4c02a3d792e1071f66d794af920ad89b4c02a3d792e1071f6&session=6d794af920ad89b4c02a3d792e1071f66d794af920ad89b4c02a3d792e1071f6

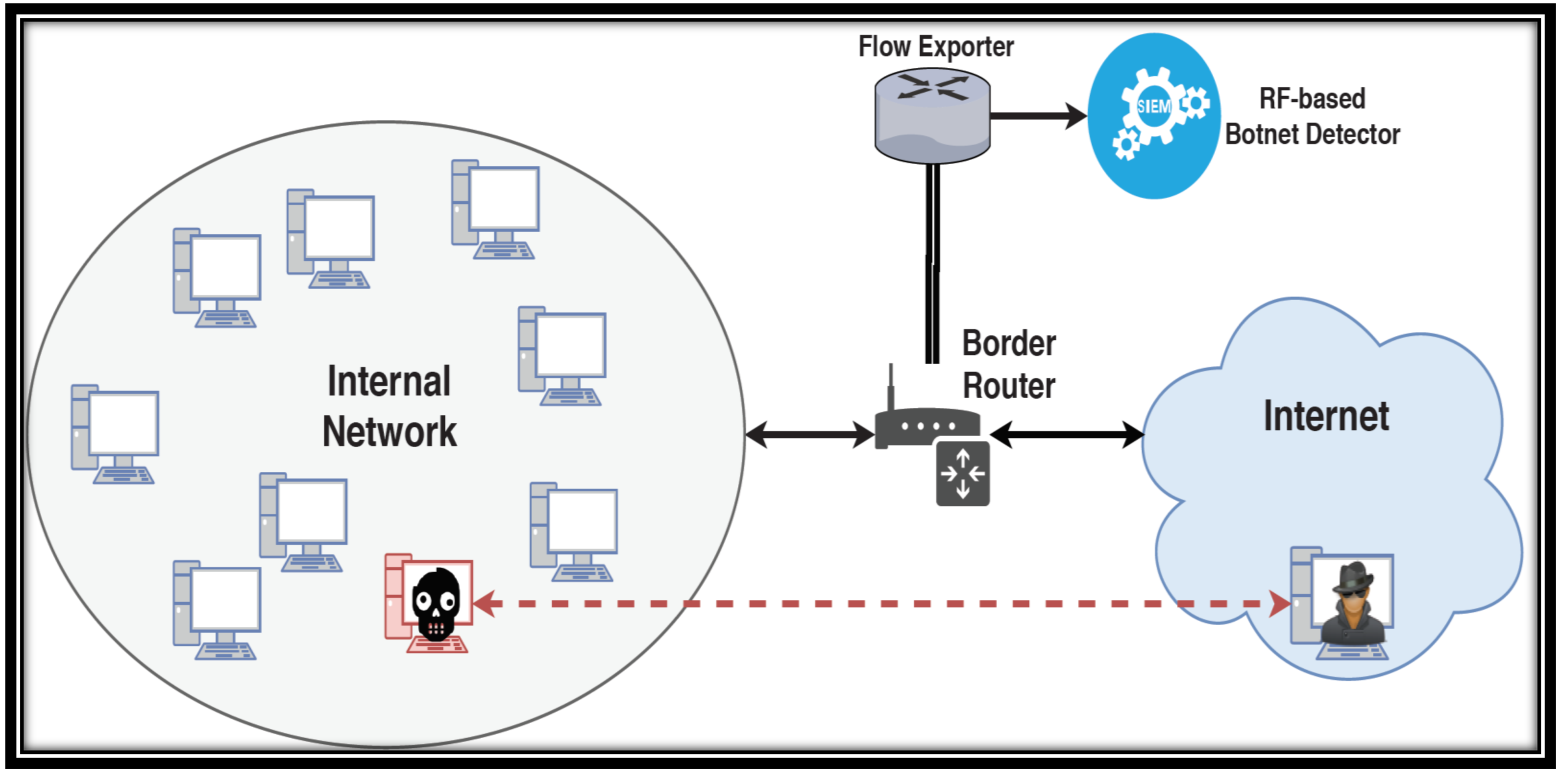
has a length of 232 characters and resulted in the following TinyURL which has a length of 30 characters:

https://tinyurl.com/phishing12

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Modern adversarial attacks against Cybersecurity applications

Evasion of Botnet detectors



Attacker Model

- Goal: evade the botnet detector
- Knowledge: Limited
- Capabilities: Limited
- Strategy: alter the bot(s) communications

Realistic assumptions

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Modern adversarial attacks against Cybersecurity applications

Evasion of Botnet detectors

Evaluation Outline:

I. Develop botnet detectors with good performance
- $(F1\text{-}score, Precision, Recall) > 90\%$

Multiple ML Algorithms:

Random Forest (RF)	Bagging (Bag)	Support Vector Machine (SVM)
Stochastic Gradient Descent (SGD)	Deep Neural Network (DNN)	Logistic Regression (LR)
Decision Tree (DT)	Naive Bayes (NB)	Gradient Boosting (GB)
AdaBoost (AB)	K-Nearest Neighbor (KNN)	Extra Trees (ET)

II. Generate realistic adversarial samples

III. Evaluate the detectors against the generated adversarial samples
- Measured through the (AS): $AS = 1 - \frac{Recall\ (attack)}{Recall\ (no\ attack)}$

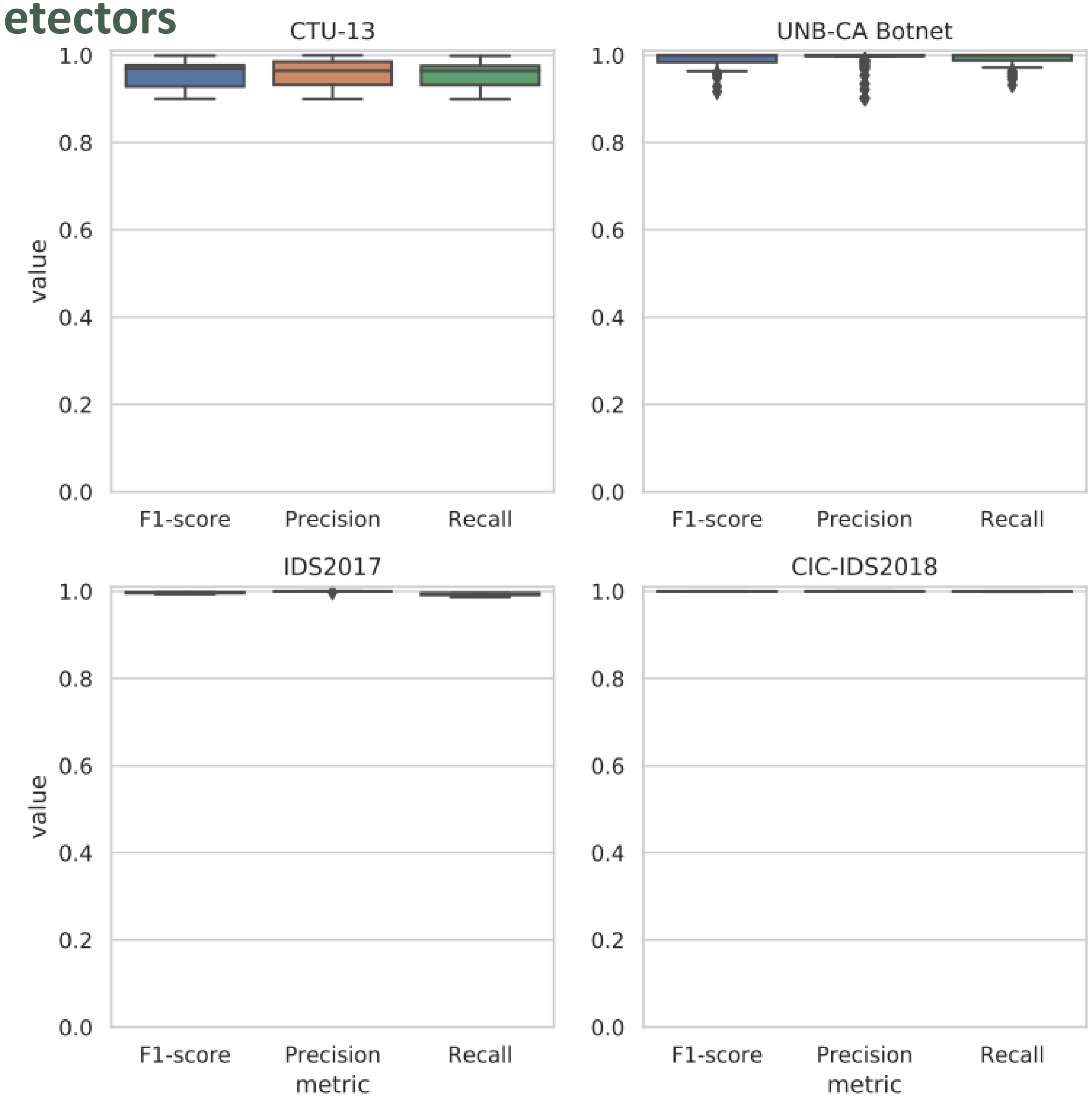
Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Modern adversarial attacks against Cybersecurity applications

Evasion of Botnet detectors

Experiments I: Baseline Performance

Dataset	F1-Score (std. dev.)	Precision (std. dev.)	Recall (std. dev.)
CTU-13	0.957 (0.029)	0.958 (0.031)	0.956 (0.028)
IDS2017	0.996 (0.002)	0.999 (0.001)	0.993 (0.003)
CIC-IDS2018	0.999 (< 0.001)	0.999 (< 0.001)	0.999 (< 0.001)
UNB-CA Botnet	0.991 (0.017)	0.992 (0.021)	0.991 (0.017)
Average	0.986 (0.011)	0.987 (0.012)	0.985 (0.011)



Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Modern adversarial attacks against Cybersecurity applications

Evasion of Botnet detectors

Experiments II: Generating Adversarial samples

Goal: generate adversarial samples through small and easily attainable modifications

Group	Altered features
1a	Duration (s)
1b	Src_bytes
1c	Dst_bytes
1d	Tot_pkts
2a	Duration, Src_bytes
2b	Duration, Dst_bytes
2c	Duration, Tot_pkts
2e	Src_bytes, Tot_pkts
2d	Src_bytes, Dst_bytes
2f	Dst_bytes, Tot_pkts
3a	Duration, Src_bytes, Dst_bytes
3b	Duration, Src_bytes, Tot_pkts
3c	Duration, Dst_bytes, Tot_pkts
3d	Src_bytes, Dst_bytes, Tot_pkts
4a	Duration, Src_bytes, Dst_bytes, Tot_pkts

Step	Duration	Src_bytes	Dst_bytes	Tot_pkts
I	+1	+1	+1	+1
II	+2	+2	+2	+2
III	+5	+8	+8	+5
IV	+10	+16	+16	+10
V	+15	+64	+64	+15
VI	+30	+128	+128	+20
VII	+45	+256	+256	+30
VIII	+60	+512	+512	+50
IX	+120	+1024	+1024	+100

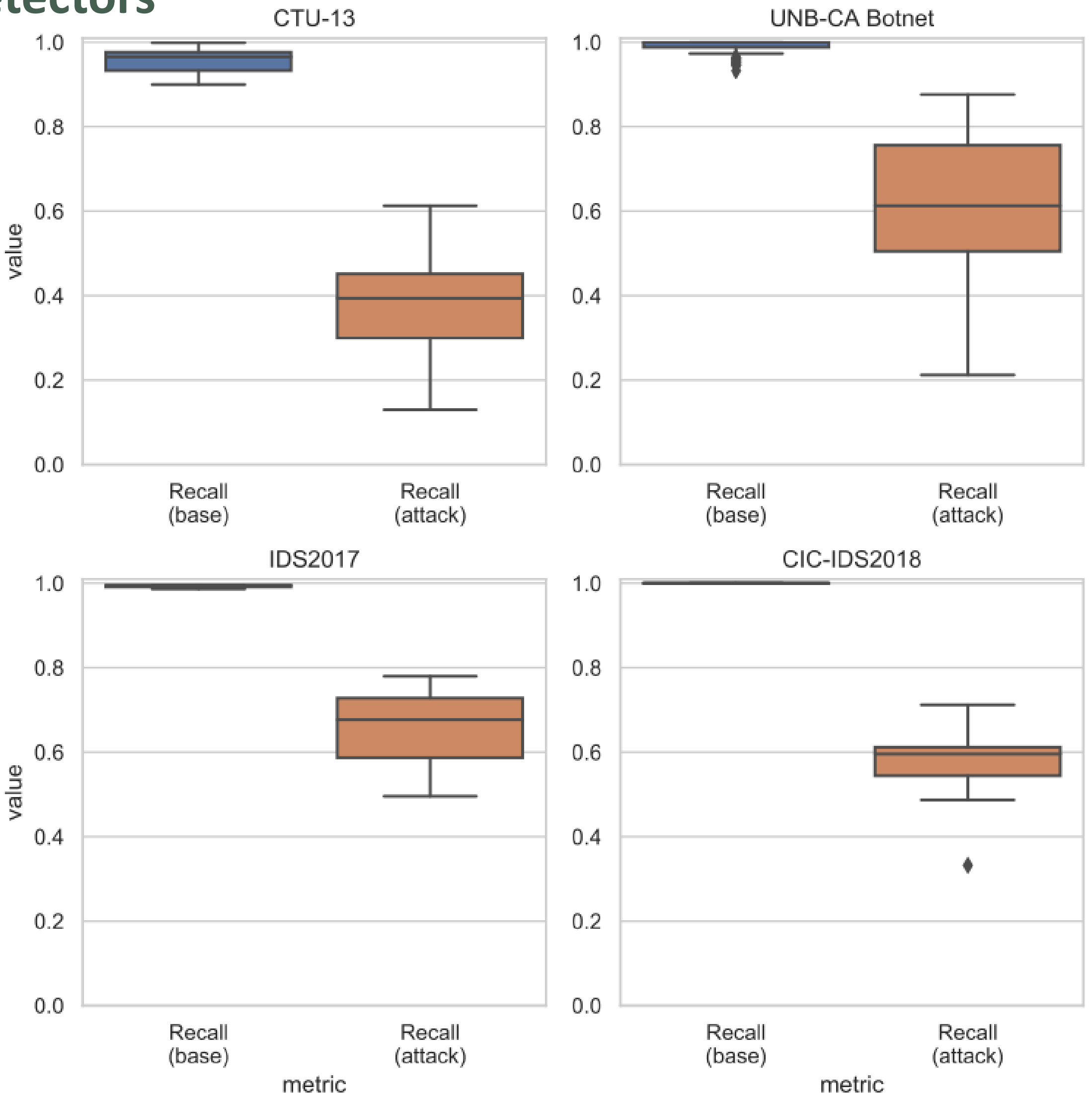
Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Modern adversarial attacks against Cybersecurity applications

Evasion of Botnet detectors

Experiments III: Impact of the attack

Dataset	Recall baseline (std. dev)	Recall adversarial (std. dev)	Attack Severity (std. dev)
CTU-13	0.956 (0.028)	0.372 (0.112)	0.609 (0.110)
IDS2017	0.993 (0.003)	0.656 (0.102)	0.327 (0.103)
CIC-IDS2018	0.999 (< 0.001)	0.564 (0.112)	0.436 (0.112)
UNB-CA Botnet	0.991 (0.017)	0.588 (0.218)	0.328 (0.212)
Average	0.985 (0.011)	0.545 (0.136)	0.425 (0.134)



Apruzzese Giovanni, Michele Colajanni, Mirco Marchetti "Evaluating the Effectiveness of Adversarial Attacks against Botnet Detectors." in *IEEE International Symposium on Network Computing and Applications (NCA)*, 2019.

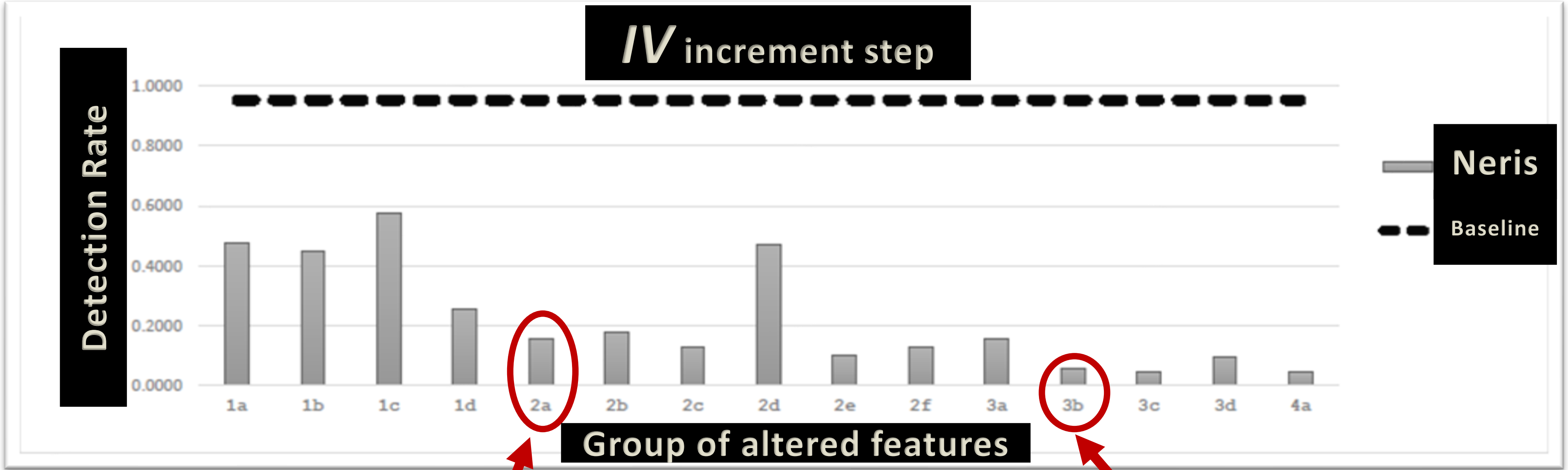
Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Modern adversarial attacks against Cybersecurity applications

Evasion of Botnet detectors

Experiments III: Impact of the attack

Detailed results on the detector for the NERIS botnet (included in the CTU-13 Dataset)



Duration +10s, *Src_Bytes* +16

Duration +10s, *Src_Bytes* +16, *Tot_Pkts* +10

Apruzzese Giovanni, and Michele Colajanni. "Evading Botnet Detectors Based on Flows and Random Forest with Adversarial Samples." in *IEEE International Symposium on Network Computing and Applications (NCA)*, 2018.

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Solutions? Yes, but at a cost...

- Re-training with adversarial samples (**Adversarial Learning**)

→ Requires the availability and mainteance of a realistic adversarial dataset.

- Use different features that cannot be modified by the attacker

→ Decreases the performance of the detector against unmodified samples.

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

WEBLAB

Seminar of Computer Security
Thursday, December 12th 2019
Modena

Adversarial Attacks against Machine Learning

Giovanni Apruzzese
PhD Candidate in Information and Communication Technologies
University of Modena and Reggio Emilia
giovanni.apruzzese@unimore.it
https://weblab.ing.unimo.it/people/apruzzese

Scuola di Ingegneria
Dipartimento di Ingegneria "Enzo Ferrari"