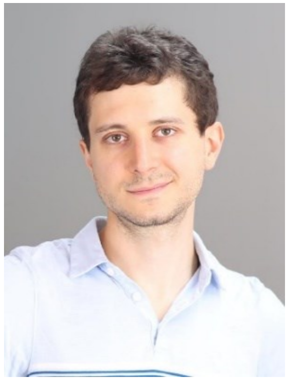
Doing Practical Research on Machine Learning & Cybersecurity

Giovanni Apruzzese, PhD

University of Padua – November 23rd, 2022

UNIVERSITÄT
LIECHTENSTEIN

whoami: Dr. **Giovanni Apruzzese**



- **Background:**
  - Did my academic studies (BSc, MSc, PhD) @ University of Modena, Italy.
    - Supervisor: Prof. Michele Colajanni
  - In 2019, spent 6 months @ Dartmouth College, USA.
    - Supervisor: Prov. VS Subrahmanian
  - …and, shortly afterwards, I met Prof. Mauro Conti (here!)
    - We've been doing some successful research together since then!
  - Joined the University of Liechtenstein in July 2020 as a PostDoc Researcher.
    - Supervisor: Prof. Pavel Laskov
  - Was "promoted" to Assistant Professor in September 2022.
- **Interests:**
  - Cybersecurity, machine learning, and any network-related topic (+🎮)
  - I like talking, researching and teaching – in a "blunt" way ☺
- **Contact information**:
  - Email (work): giovanni.apruzzese@uni.li
  - Website (personal): www.giovanniapruzzese.com
  - Feel free to contact me if you have any questions.
    - I reply fast, and will happily do so!

UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What I do

Machine Learning + Cybersecurity

- Applying ML to *provide security* of a given information system
  - E.g.: using ML to detect cyber threats
- *Attacking / Defending* ML applications
  - E.g.: evading a ML model that detects phishing websites
- Using machine learning *offensively…*
  - …against another system (e.g.: artificially generating “fake” images)
  - …against humans (e.g., violating privacy)

BONUS

- Using ML to attack an ML-based security system and harden it

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Outline of Today

- Fundamentals of Machine Learning and Cybersecurity
  - Ref: Giovanni Apruzzese, et al. "The Role of Machine Learning in Cybersecurity." ACM Digital Threats: Research and Practice (2022)

- The security of Machine Learning-based Phishing Website Detectors
  - Ref: Giovanni Apruzzese, Mauro Conti, Ying Yuan. "SpacePhish: The Evasion-space of Adversarial Attacks against Phishing Website Detectors using Machine Learning." Annual Computer Security Applications Conference (Dec. 2022).

- Machine Learning Security in the Real-World
  - Ref: Giovanni Apruzzese, Hyrum S. Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, Kevin A. Roundy "'Real Attackers Don't Compute Gradients': Bridging the Gap Between Adversarial ML Research and Practice." IEEE International Conference on Secure and Trustworthy Machine Learning (Feb. 2023)

- Adversarial Attacks against Humans **and** Machine Learning
  - Ref: Johannes Schneider, Giovanni Apruzzese. "Concept-based Adversarial Attacks: Tricking Humans and Classifiers alike." IEEE Symposium on Security and Privacy – Deep Learning and Security Workshop (May 2022)

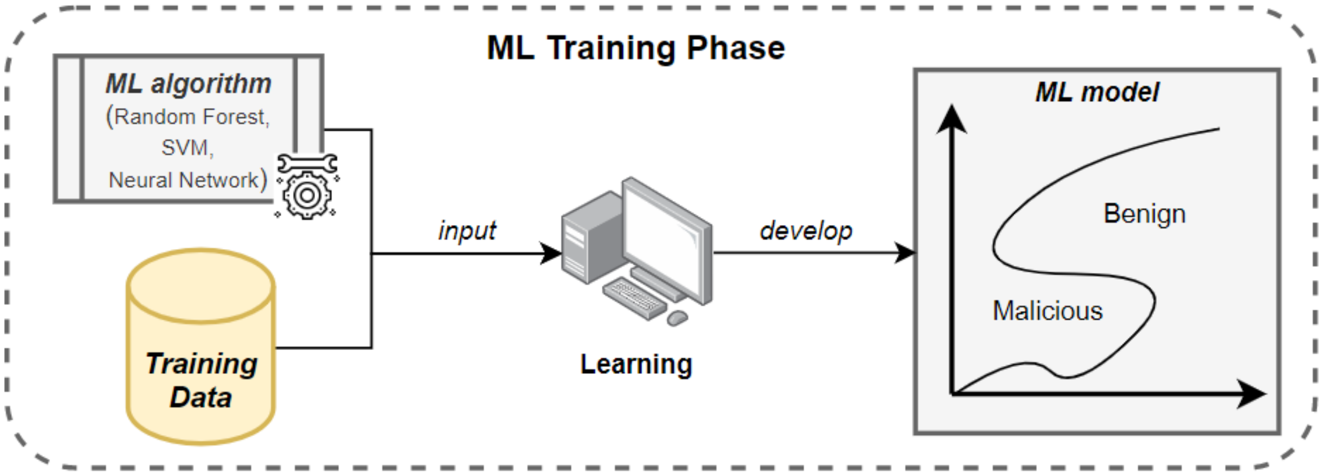
- Cybersecurity in the Smart Grid (in Practice)
  - Ref: Jacqueline Meyer, Giovanni Apruzzese. "Cybersecurity in the Smart Grid: Practitioners' Perspective." Industrial Control Systems Security Workshop (Dec. 2022) [co-located with ACSAC]

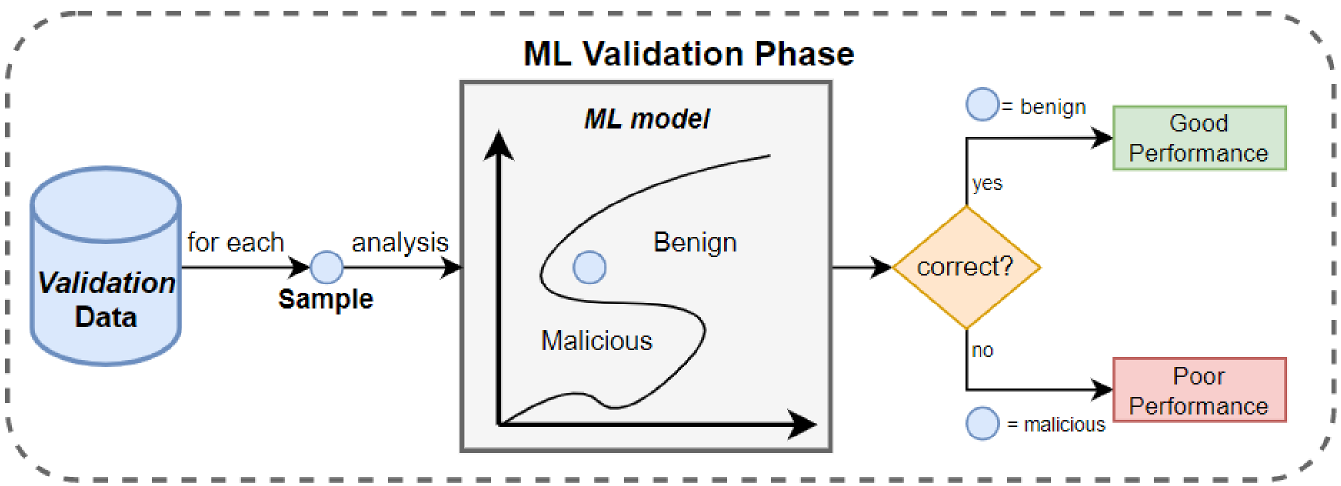
Fundamentals of Machine Learning and Cybersecurity

UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

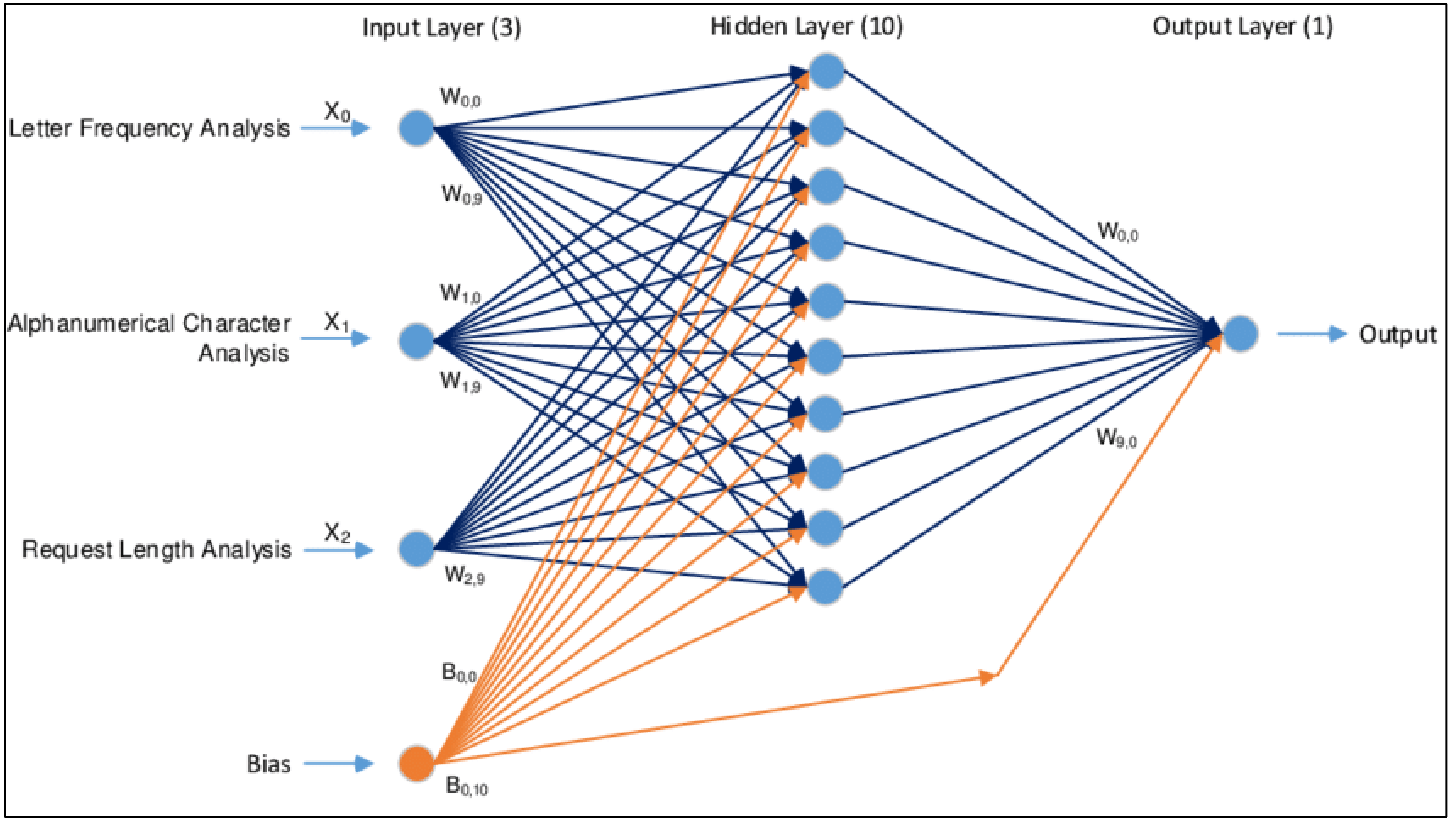
Machine Learning workflow: Training and Testing

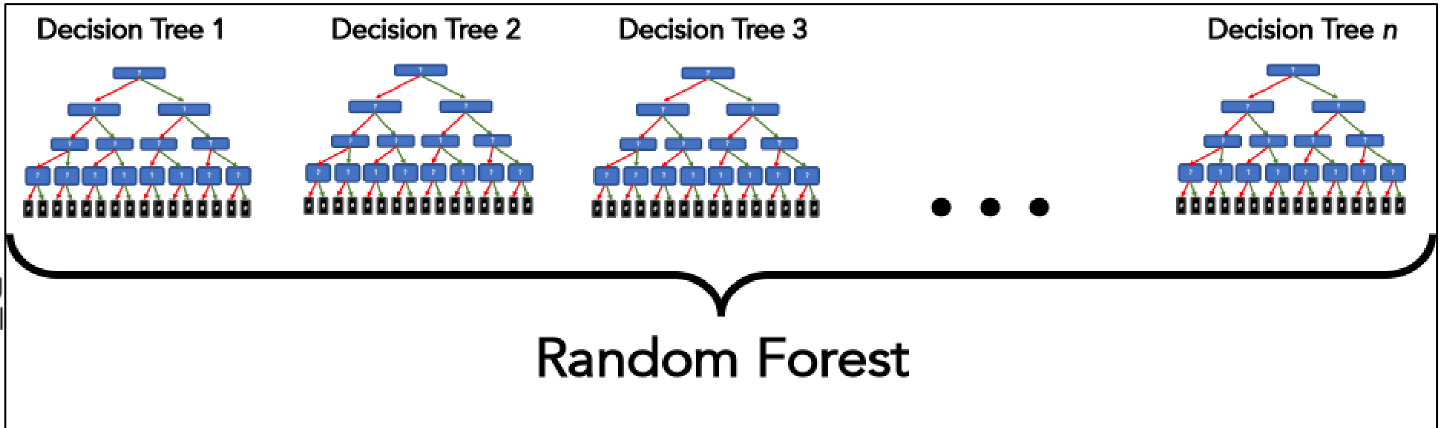




Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

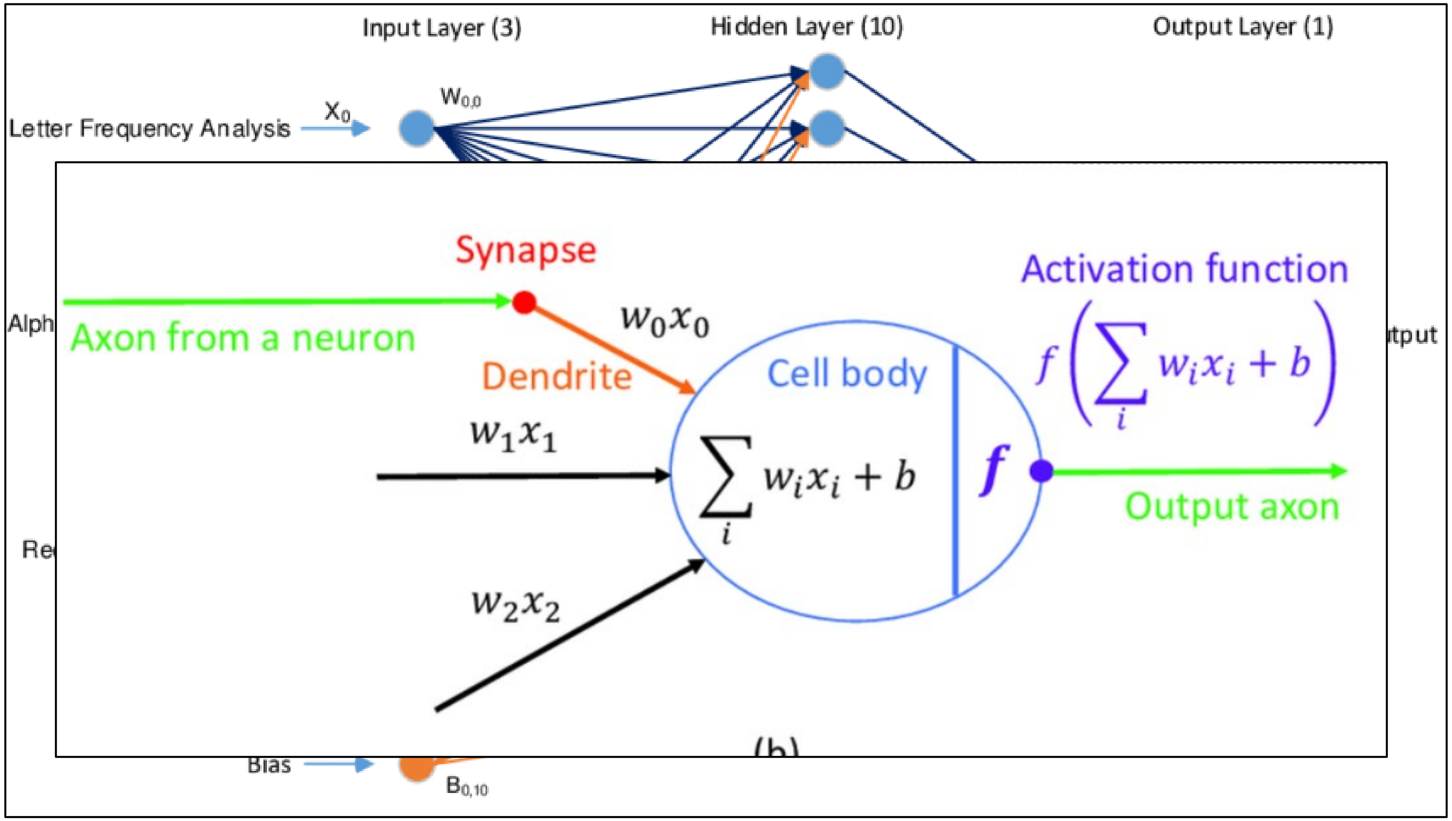
Do you think that training ML models is difficult?

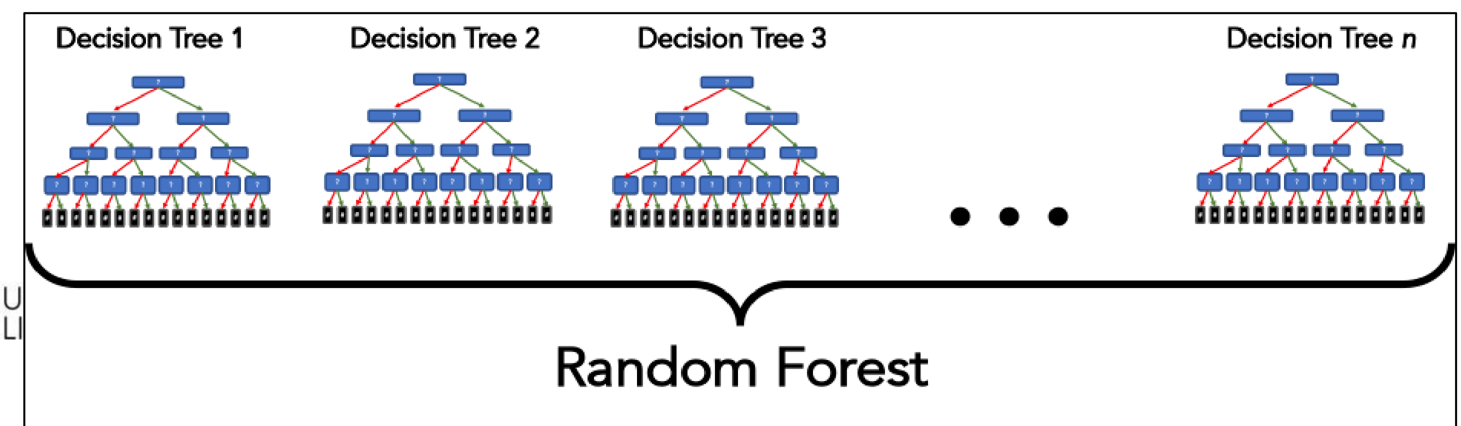




Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

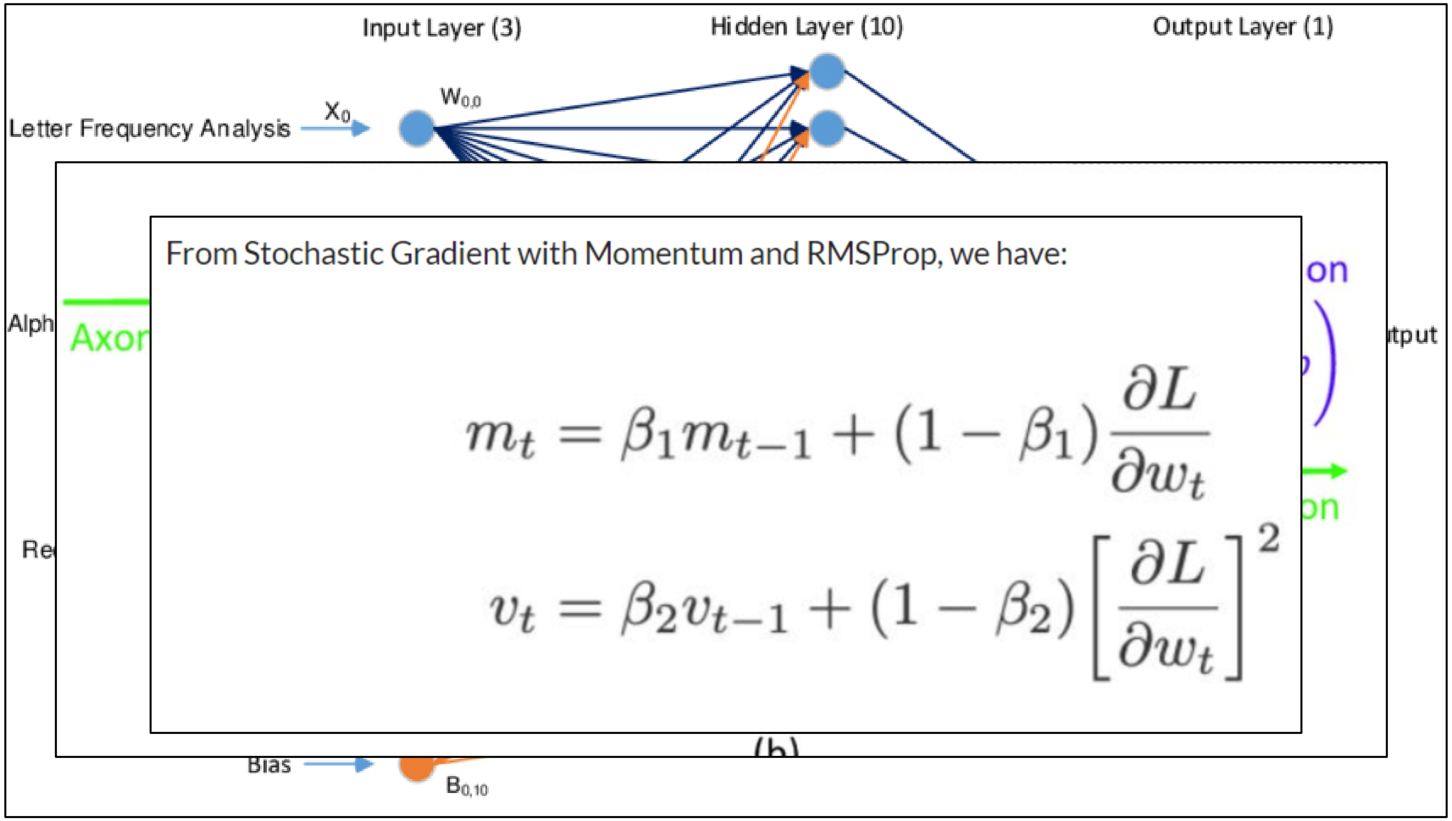
Do you think that training ML models is difficult? – Maths





Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

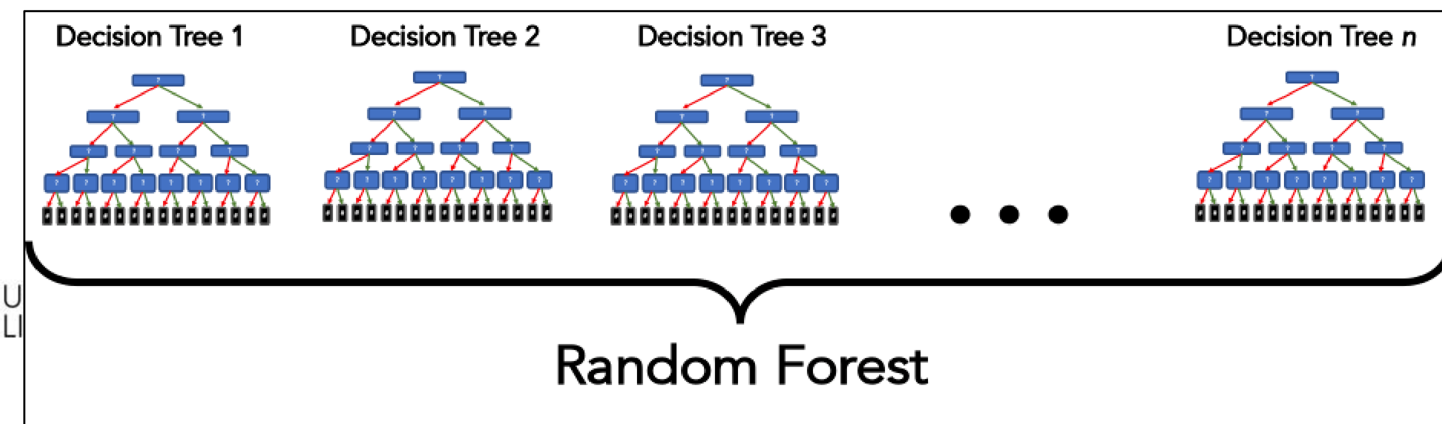
Do you think that training ML models is difficult? – More Maths



From Stochastic Gradient with Momentum and RMSProp, we have:

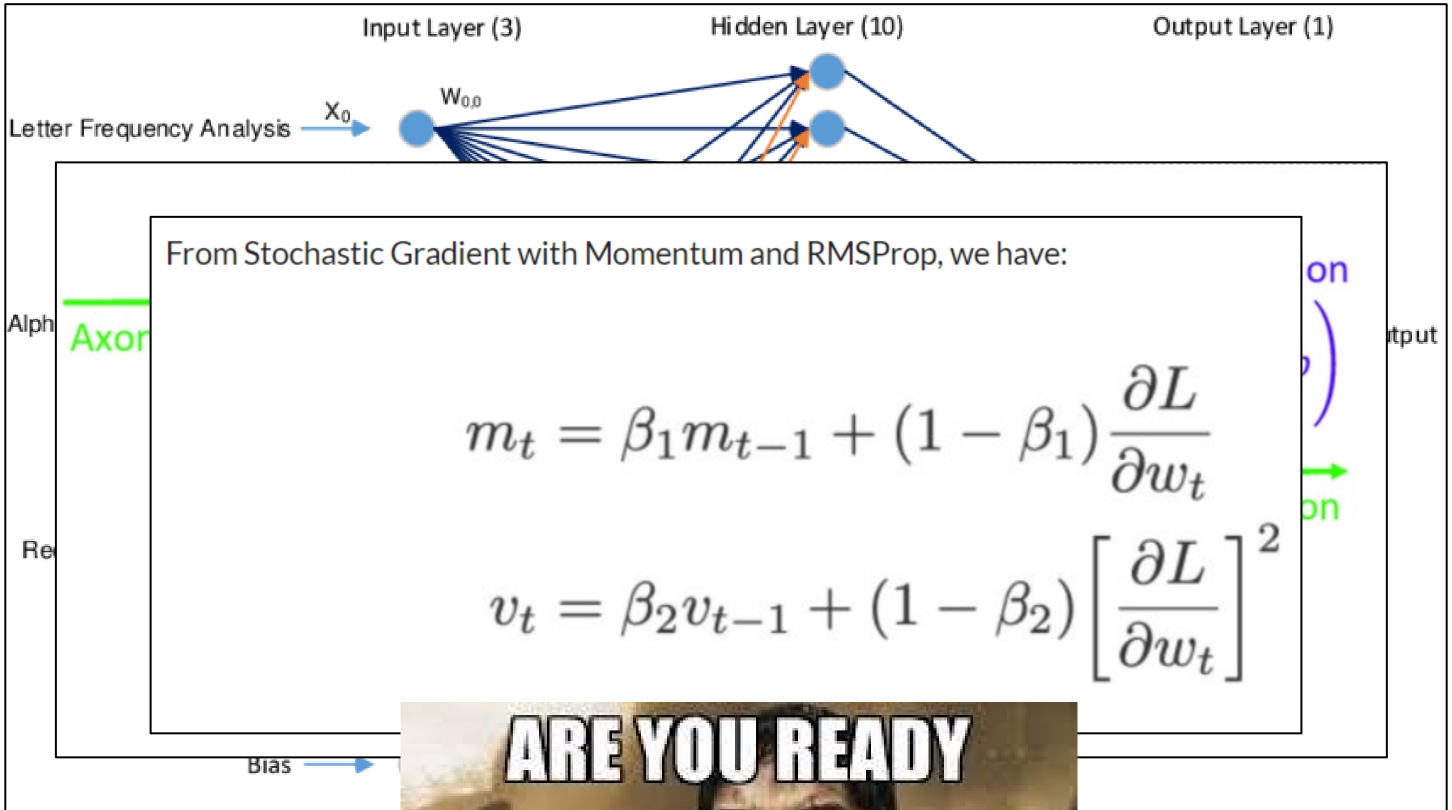
$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\frac{\partial L}{\partial w_t}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\left[\frac{\partial L}{\partial w_t}\right]^2$$



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

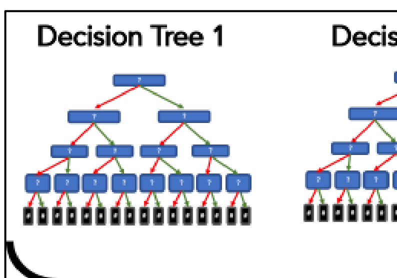
Do you think that training ML models is difficult? – More Maths ☺

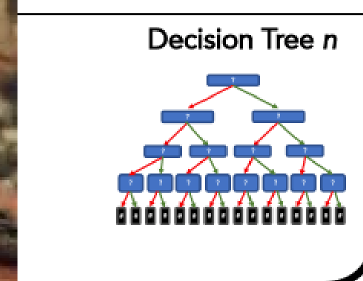


From Stochastic Gradient with Momentum and RMSProp, we have:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\frac{\partial L}{\partial w_t}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\left[\frac{\partial L}{\partial w_t}\right]^2$$





Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Do you think that training ML models is difficult? – One line

```
#train the classifier (rf_clf) using the training_data (train[features]) with corresponding labels (y)
print("Training...")
rf_clf.fit(train[features],y)
print("Done")
```

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Do you think that training ML models is difficult? – The real problem

PROBLEMS (data)

```
#train the classifier (rf_clf) using the training_data (train[features]) with corresponding labels (y)
print("Training...")
rf_clf.fit(train[features],y)
print("Done")
```

PROBLEMS (tuning)

UNIVERSITÄT LIECHTENSTEIN

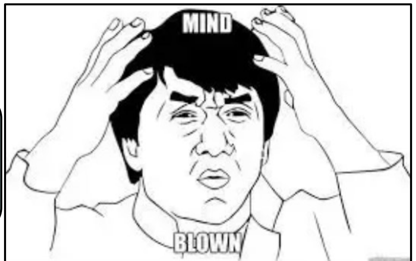
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Do you think that training ML models is difficult? – The real problem

PROBLEMS (data)

```
#train the classifier (rf_clf) using the training_data (train[features]) with corresponding labels (y)
print("Training...")
rf_clf.fit(train[features],y)
print("Done")
```

PROBLEMS (tuning)



Of course, you're always free to go, learn and improve the *fit* function:

- RF: https://github.com/scikit-learn/scikit-learn/blob/baf828ca1/sklearn/ensemble/_forest.py#L297
- MLP: https://github.com/scikit-learn/scikit-learn/blob/f3f51f9b6/sklearn/neural_network/_multilayer_perceptron.py#L745

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Common issues of ML in Cybersecurity

- Applying Machine Learning requires *data* to train an ML model
- Depending on the "problem" solved by such model, the data may require *labels*
- **Obtaining (any) data has a <u>cost</u>, and labelled data is (very) *expensive***

- Machine Learning models are ultimately just a component within a system
- **Such ML models *can* be targeted by "Adversarial Attacks"**
- Such strategies ultimately aim to compromise the functionality of the ML model.

- The cybersecurity domain implicitly assumes the presence of attackers.
- Attackers are *human beings,* and hence operate with a *cost/benefit* mindset
- **Such considerations must be made when analyzing the security of (any) IT system**

"There is no such thing as a *foolproof* system. If you believe you have one, then you failed to take into account the creativity of fools" [source]

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Common issues of ML in Cybersecurity (cond'd)

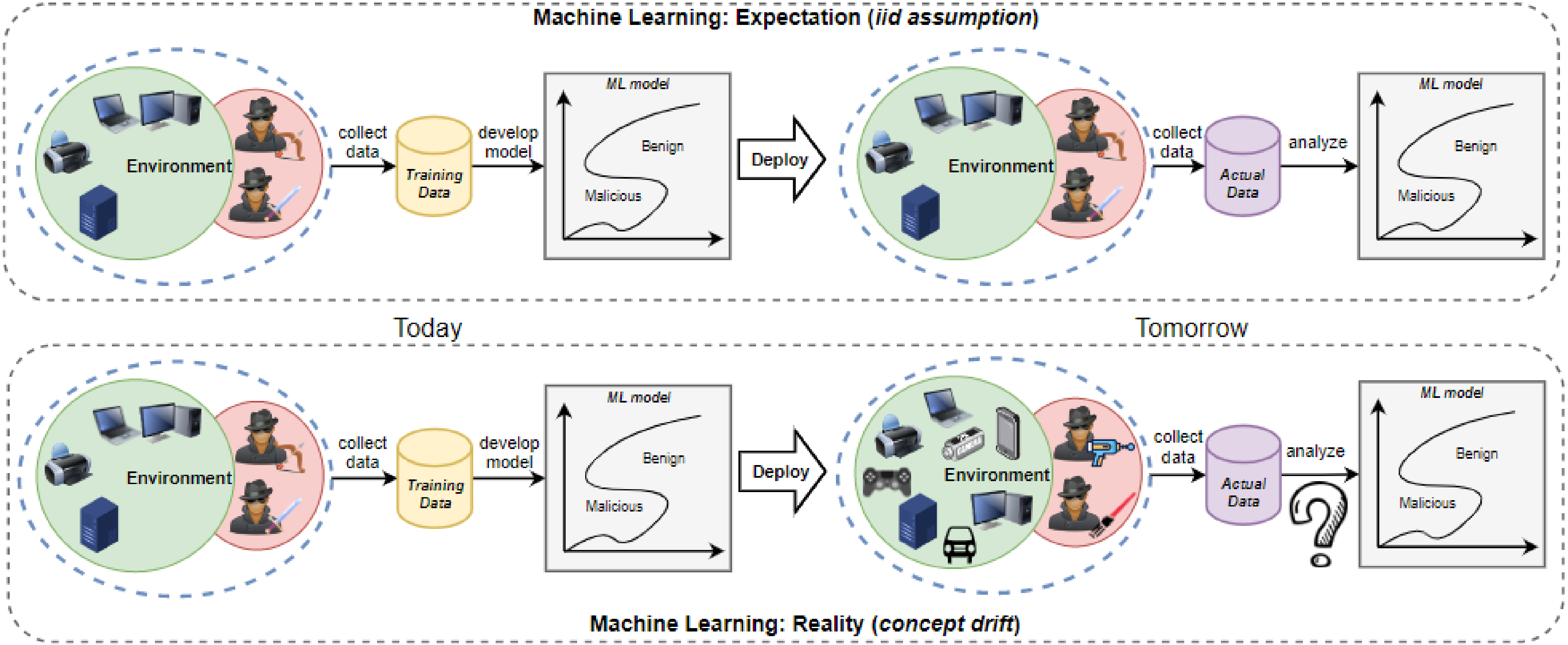


Fig. 9. Machine Leaning in the presence of Concept Drift. The ML model expects that the data will not deviate from the one seen during its training. In cybersecurity, however, the environment evolves, and adversaries also become more powerful.

UNIVERSITÄT LIECHTENSTEIN

The security of Machine Learning-based Phishing Website Detectors

UNIVERSITÄT
LIECHTENSTEIN

The security of Machine Learning-based Phishing Website Detectors

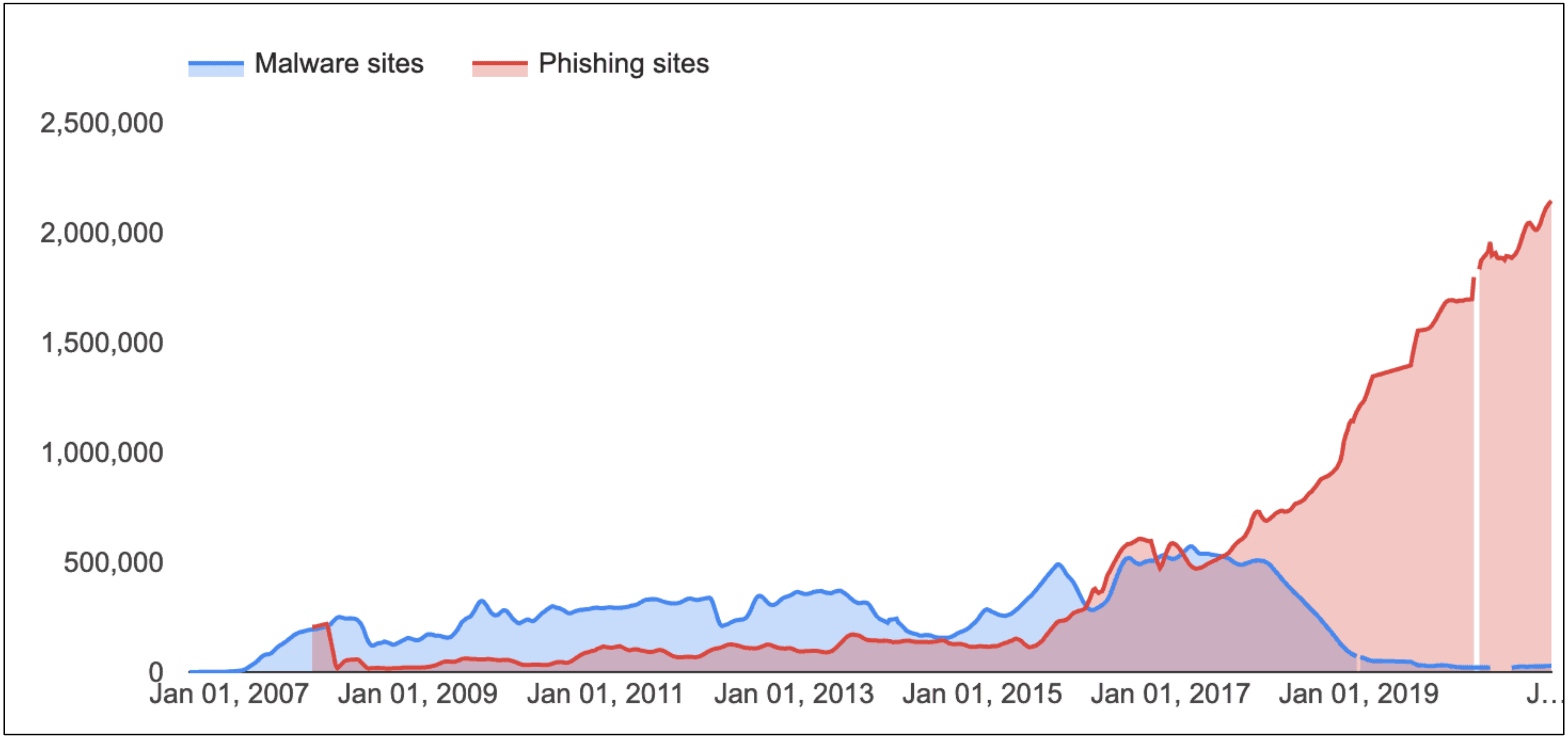
In the adversarial ML domain, have you ever read a research paper proposing an attack that has an effectiveness of 3%?

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Current Landscape of Phishing

- Phishing attacks are continuously increasing
- Most detection methods still rely on *blocklists* of malicious URLs
  - These detection techniques can be evaded easily by “squatting” phishing websites!



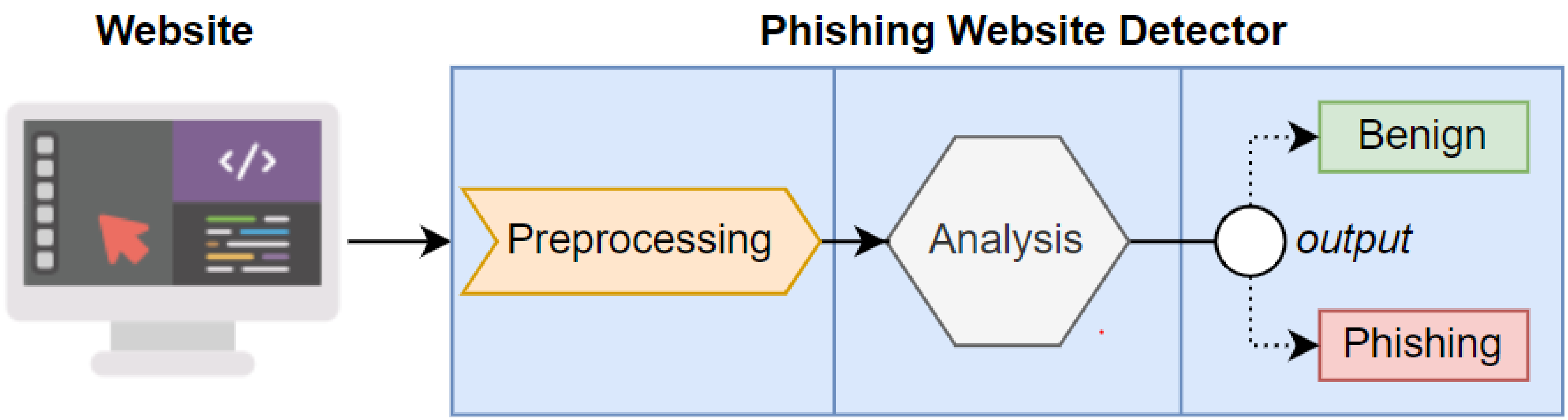
UNIVERSITÄT LIECHTENSTEIN

Image source: https://www.tessian.com/blog/phishing-statistics-2020/

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Current Landscape of Phishing – Countermeasures

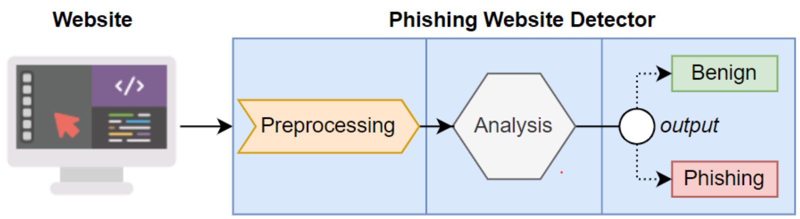
- Countering such simple (but effective) strategies can be done via *data-driven* methods



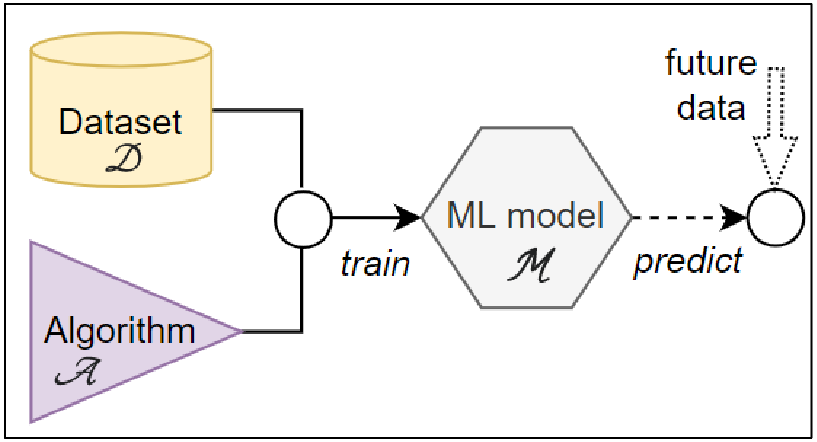
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Current Landscape of Phishing – Countermeasures (ML)

- Countering such simple (but effective) strategies can be done via *data-driven* methods



- Such methods (obviously ☺) include (also) Machine Learning techniques:



- Machine Learning-based Phishing Website Detectors (ML-PWD) are very effective [1]
  - Even popular products and web-browsers (e.g., Google Chrome) use them! [2]

UNIVERSITÄT LIECHTENSTEIN

[1]: Tian, Ke, et al. "Needle in a haystack: Tracking down elite phishing domains in the wild." Internet Measurement Conference 2018.
[2]: El Kouari, Oumaima, Hafssa Benaboud, and Saiida Lazaar. "Using machine learning to deal with Phishing and Spam Detection: An overview." Proceedings of the 3rd International Conference on Networking, Information Systems & Security. 2020.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Phishing in a nutshell

- Phishing websites are taken down quickly
  - The moment they are reported in a blocklist, they become useless
- Even if a victim lands on a phishing website, the phishing attempt is not complete
  - The victim may be “hooked”, but they are not “phished” yet!

Most phishing attacks end up in failure [3]

UNIVERSITÄT LIECHTENSTEIN

[3] Adam Oest, et al “Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale.” In Proc. USENIX Secur. Symp. (2020)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Phishing in a nutshell (cont'd)

- Phishing websites are taken down quickly
  - The moment they are reported in a blocklist, they become useless
- Even if a victim lands on a phishing website, the phishing attempt is not complete
  - The victim may be "hooked", but they are not "phished" yet!

Most phishing attacks end up in failure [3]

- Phishers are well aware of this fact… but they (clearly) keep doing it
  - Hence, they "have to" evade detection mechanisms

(Remember: Real attackers operate with a cost/benefit mindset [4])

UNIVERSITÄT LIECHTENSTEIN

[3] Adam Oest, et al "Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale." In Proc. USENIX Secur. Symp. (2020)
[4] Kelce S Wilson and Müge Ayse Kiy. 2014. Some fundamental Cybersecurity concepts. IEEE Access (2014).

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Problem Statement: Adversarial Attacks against ML-PWD

- ML-PWD are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Adversarial Attacks exploit a **perturbation**, ε, that induces an ML model, $\mathcal{M}$, to misclassify a given input, F_x, by producing an incorrect output (y_x^ε instead of y_x)

$$\text{find } \varepsilon \text{ s.t. } \mathcal{M}(F_x) = y_x^\varepsilon \neq y_x$$

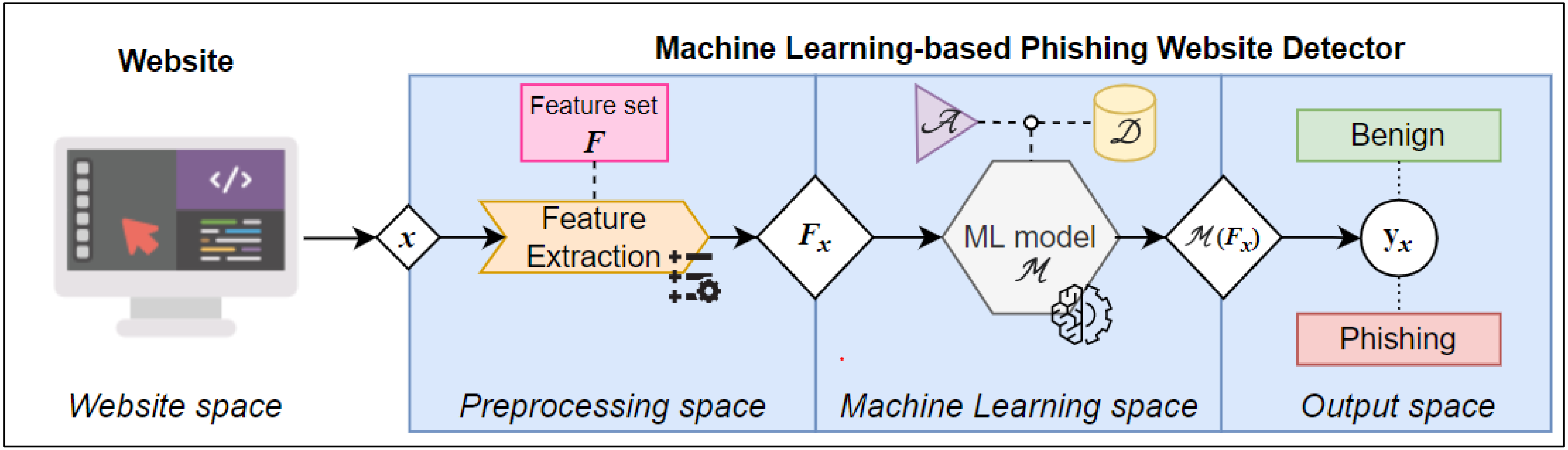
UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Problem Statement: Adversarial Attacks against ML-PWD

- ML-PWD are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Adversarial Attacks exploit a perturbation, ε, that induces an ML model, $\mathcal{M}$, to misclassify a given input, F_x, by producing an incorrect output (y_x^ε instead of y_x)

- In the context of a ML-PWD, such **perturbation** can be introduced in three 'spaces':



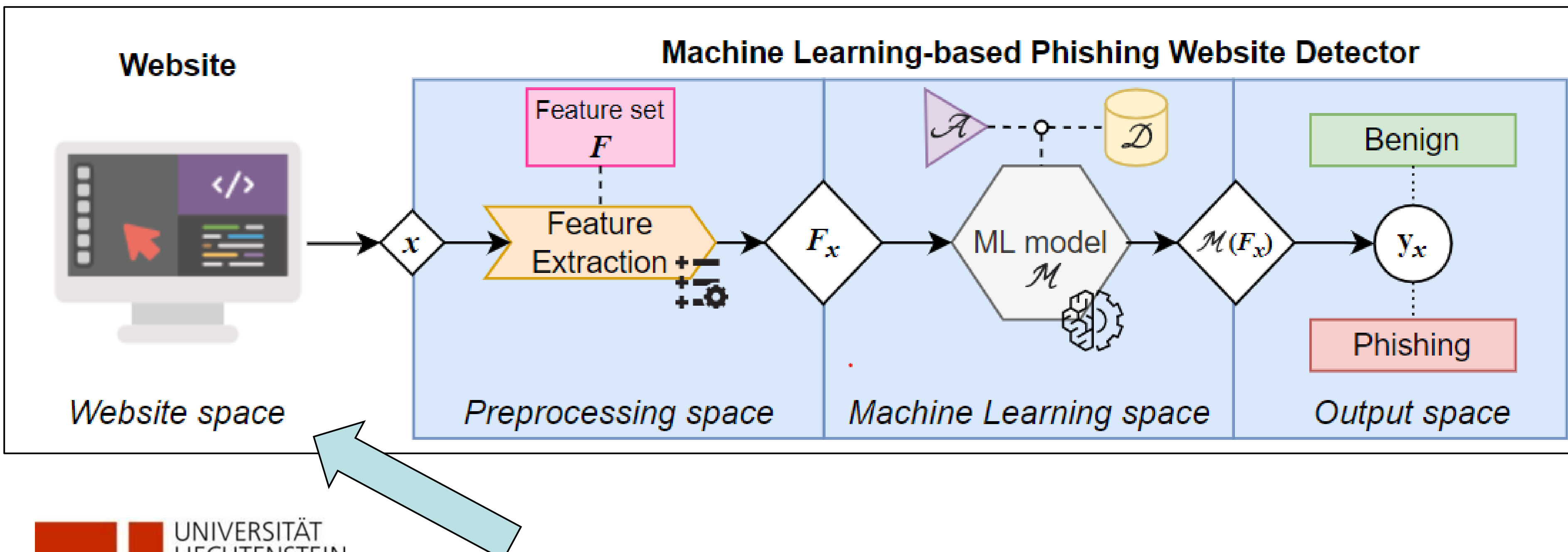
UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Problem Statement: Adversarial Attacks against ML-PWD

- ML-PWD are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Adversarial Attacks exploit a perturbation, ε, that induces an ML model, $\mathcal{M}$, to misclassify a given input, F_x, by producing an incorrect output (y_x^ε instead of y_x)

- In the context of a ML-PWD, such **perturbation** can be introduced in three 'spaces':



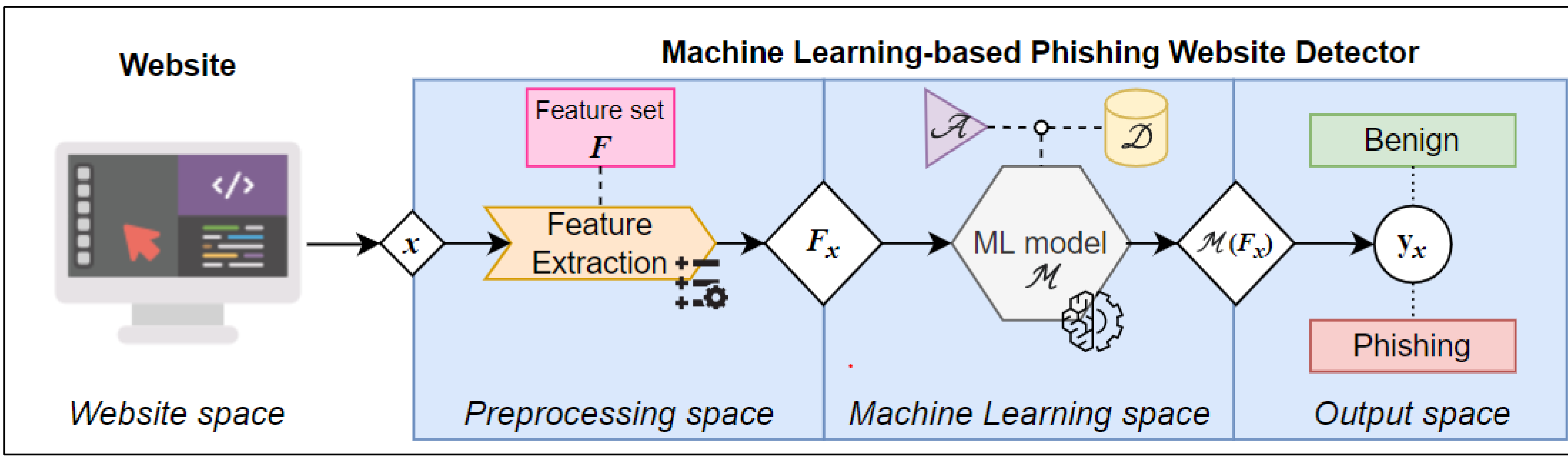
UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Problem Statement: Adversarial Attacks against ML-PWD

- ML-PWD are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Adversarial Attacks exploit a perturbation, ε, that induces an ML model, $\mathcal{M}$, to misclassify a given input, F_x, by producing an incorrect output (y_x^ε instead of y_x)

- In the context of a ML-PWD, such **perturbation** can be introduced in three 'spaces':



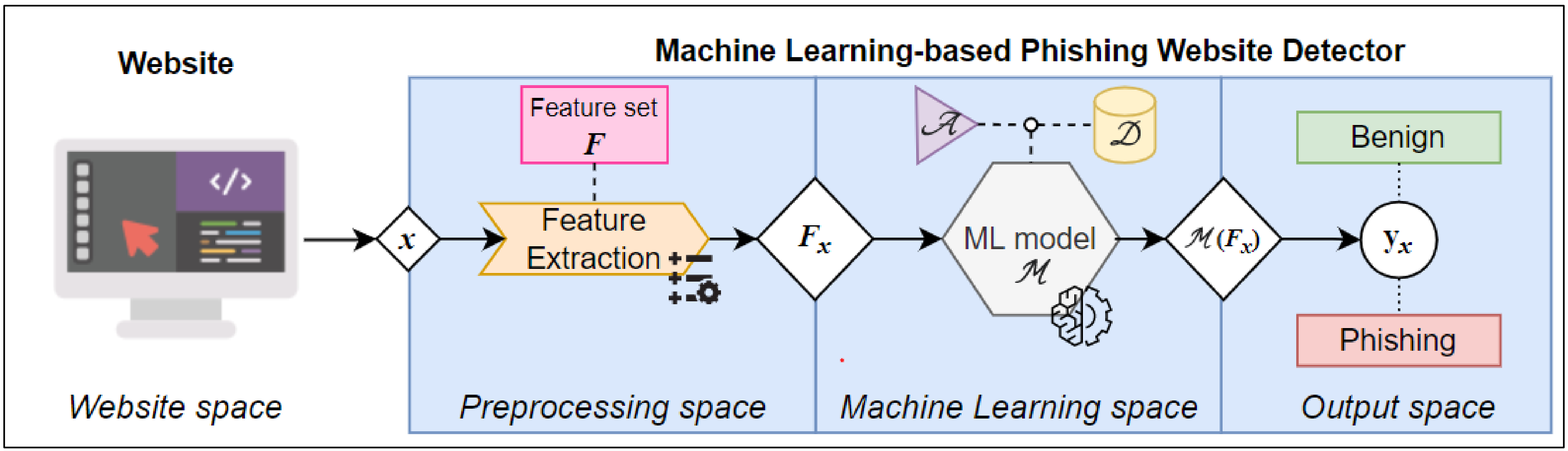
UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Problem Statement: Adversarial Attacks against ML-PWD

- ML-PWD are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Adversarial Attacks exploit a perturbation, ε, that induces an ML model, $\mathcal{M}$, to misclassify a given input, F_x, by producing an incorrect output (y_x^ε instead of y_x)

- In the context of a ML-PWD, such **perturbation** can be introduced in three 'spaces':

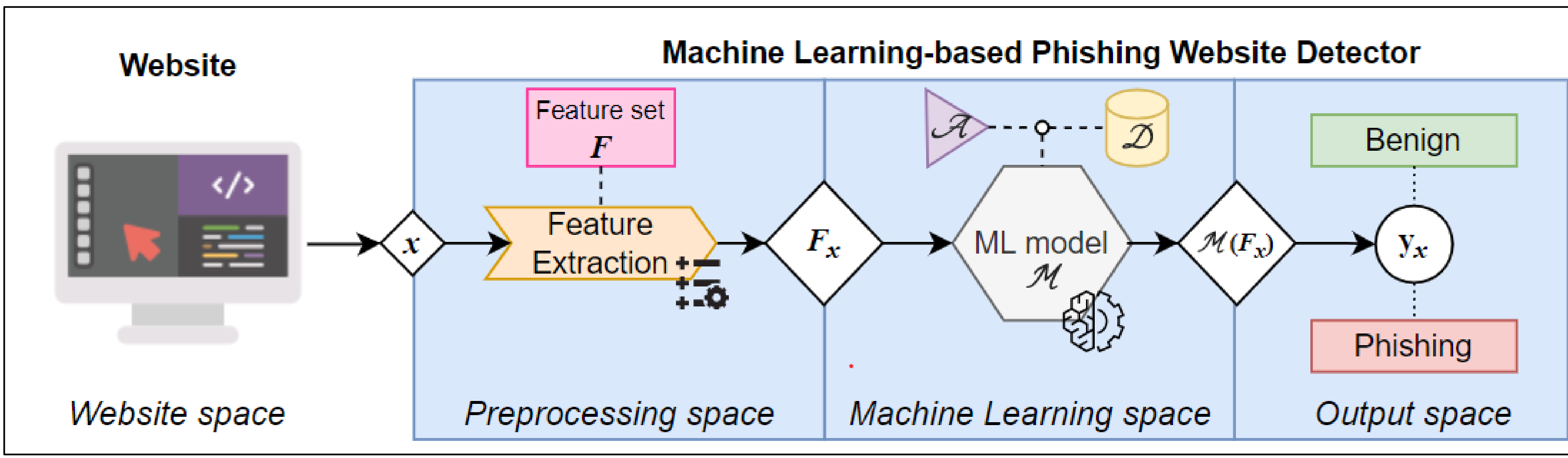


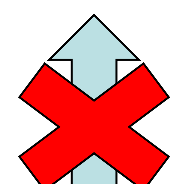
UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Problem Statement: Adversarial Attacks against ML-PWD

- ML-PWD are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Adversarial Attacks exploit a perturbation, ε, that induces an ML model, $\mathcal{M}$, to misclassify a given input, F_x, by producing an incorrect output (y_x^ε instead of y_x)
- In the context of a ML-PWD, such **perturbation** can be introduced in three 'spaces':





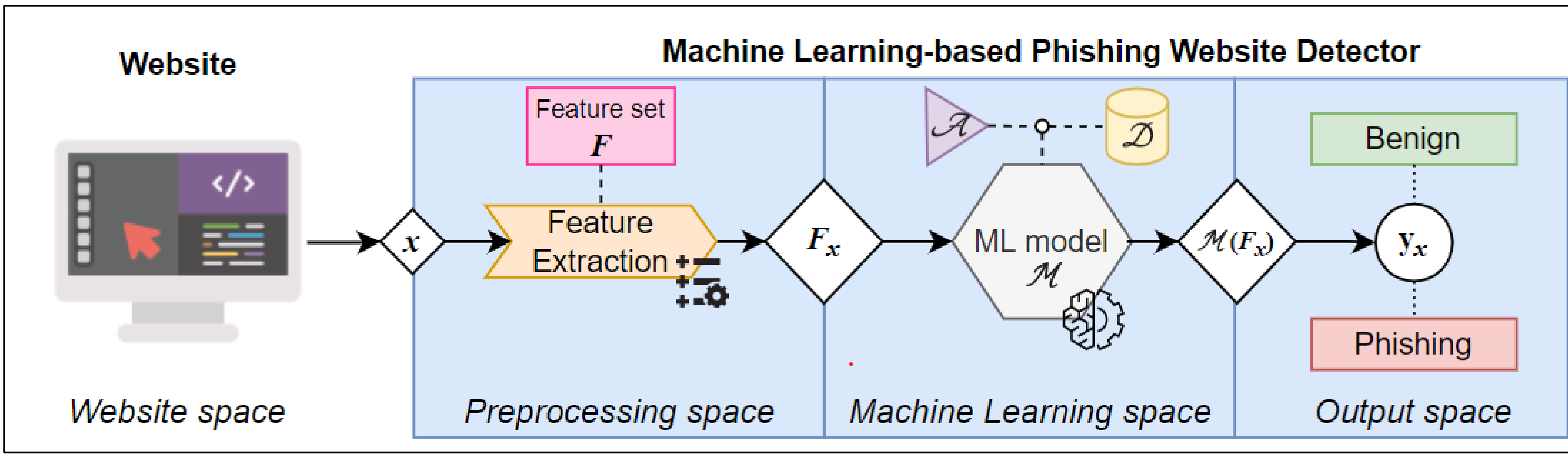
UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Problem Statement: Adversarial Attacks against ML-PWD

- ML-PWD are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Adversarial Attacks exploit a perturbation, ε, that induces an ML model, $\mathcal{M}$, to misclassify a given input, F_x, by producing an incorrect output (y_x^{ε} instead of y_x)

- In the context of a ML-PWD, such **perturbation** can be introduced in three 'spaces':

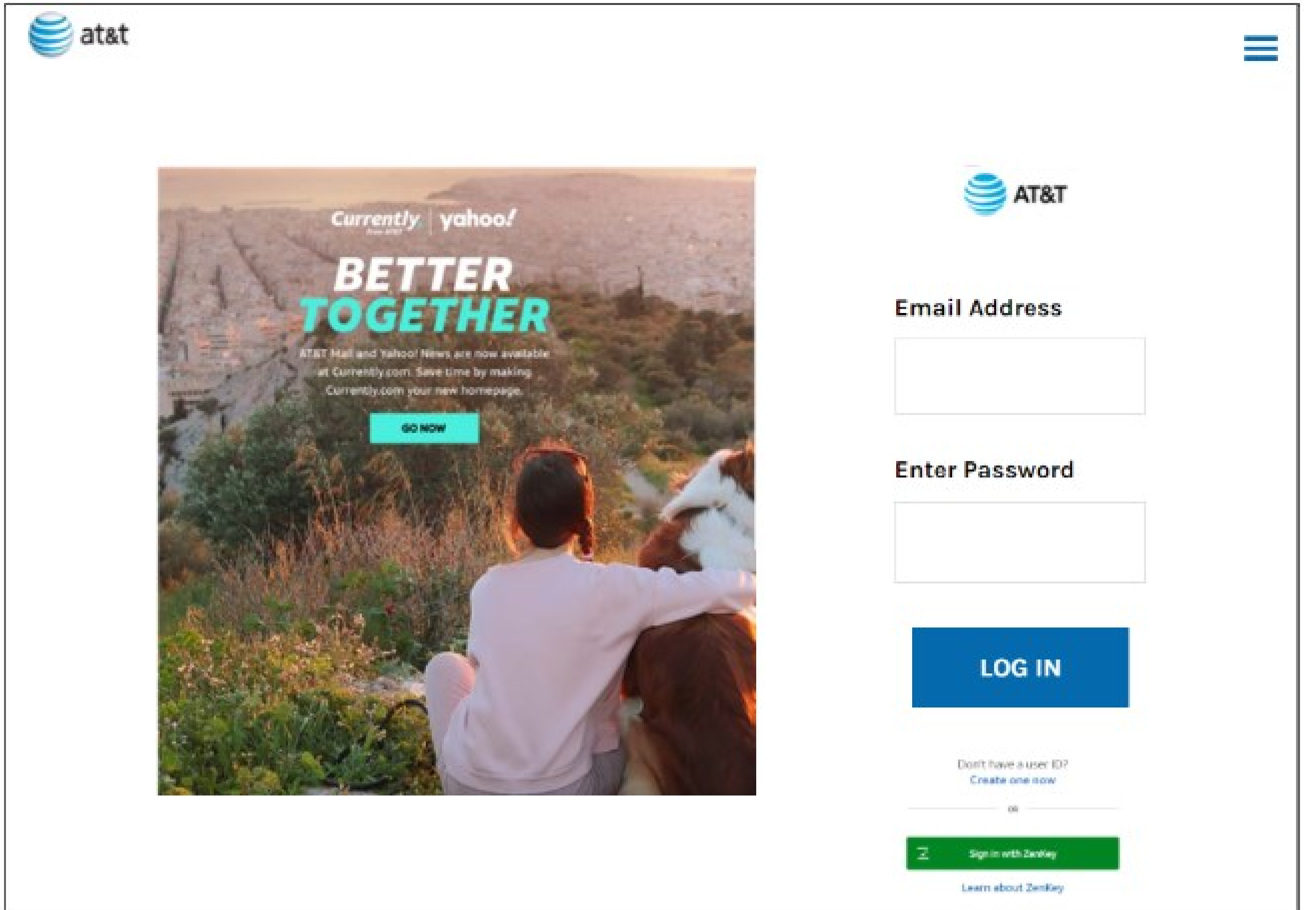


Question: Which 'space' do you think an *attacker* is **most likely** to use?

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Website-space Perturbations (WsP) in practice – original example

Figure 4: An exemplary (and true) Phishing website, whose URL is https://www.63y3hfh-fj39f30-f30if0f-f392.weebly.com/.



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

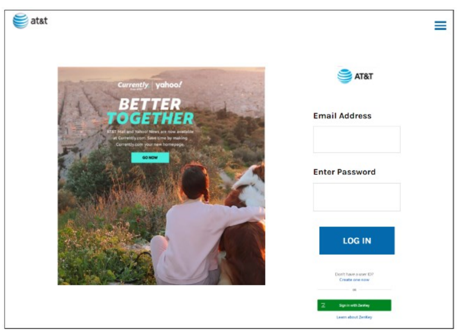
Website-space Perturbations (WsP) in practice – changing the URL

https://www.63y3hfh-fj39f30-f30if0f-f392.weebly.com/ → https://bit.ly/3MZHjt7

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Website-space Perturbations (WsP) in practice – changing the HTML




```
<div>
    <form enctype="multipart/form-data" action="//www.weebly.com/weebly/apps/formSubmit.php" method=
    "POST" id="form-723155629711391878">
        <div id="723155629711391878-form-parent" class="wsite-form-container"
             style="margin-top:10px;">
            <ul class="formlist" id="723155629711391878-form-list">
                <div><div class="wsite-form-field" style="margin:5px 0px 5px 0px;">
                <label class="wsite-form-label" for="input-227982018179653776">Email Address <span
                class="form-not-required">*</span></label>
                <div class="wsite-form-input-container">
                    <input id="input-227982018179653776" class="wsite-form-input wsite-input
                    wsite-input-width-370px" type="text" name="_u227982018179653776" />
                </div>
                <div id="instructions-227982018179653776" class="wsite-form-instructions" style=
                "display:none;"></div>
            </div></div>

    <a href="./fake-link-to-nonexisting-resource">
        <font style="visibility:hidden">Resource</font></a>

    <a href='#' style='display:none'> can not see</a>

<div><div class="wsite-form-field" style="margin:5px 0px 5px 0px;">
                <label class="wsite-form-label" for="input-435728988405554593">Enter Password <span
                class="form-not-required">*</span></label>
                <div class="wsite-form-input-container">
                    <textarea id="input-435728988405554593" class="wsite-form-input wsite-input
```

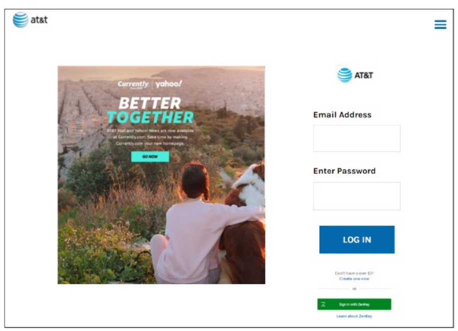
ε (WsP)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Website-space Perturbations (WsP) in practice – changing URL+HTML

https://www.63y3hfh-fj39f30-f30if0f-f392.weebly.com/

https://bit.ly/3MZHjt7




```
<div>
    <form enctype="multipart/form-data" action="//www.weebly.com/weebly/apps/formSubmit.php" method=
    "POST" id="form-723155629711391878">
        <div id="723155629711391878-form-parent" class="wsite-form-container"
             style="margin-top:10px;">
            <ul class="formlist" id="723155629711391878-form-list">
                <div><div class="wsite-form-field" style="margin:5px 0px 5px 0px;">
                <label class="wsite-form-label" for="input-227982018179653776">Email Address <span
                class="form-not-required">*</span></label>
                <div class="wsite-form-input-container">
                    <input id="input-227982018179653776" class="wsite-form-input wsite-input
                    wsite-input-width-370px" type="text" name="_u227982018179653776" />
                </div>
                <div id="instructions-227982018179653776" class="wsite-form-instructions" style=
                "display:none;"></div>
            </div></div>

    <a href="./fake-link-to-nonexisting-resource">
        <font style="visibility:hidden">Resource</font></a>

    <a href='#' style='display:none'> can not see</a>

<div><div class="wsite-form-field" style="margin:5px 0px 5px 0px;">
                <label class="wsite-form-label" for="input-435728988405554593">Enter Password <span
                class="form-not-required">*</span></label>
                <div class="wsite-form-input-container">
                    <textarea id="input-435728988405554593" class="wsite-form-input wsite-input
```

ε (WsP)

UNIVERSITÄT LIECHTENSTEIN

Why do we need all of this anyway? (first reason)

2020 IEEE Symposium on Security and Privacy

Intriguing Properties of Adversarial ML Attacks in the Problem Space

Fabio Pierazzi*†, Feargus Pendlebury*†‡§, Jacopo Cortellazzi†, Lorenzo Cavallaro†
† King's College London, ‡ Royal Holloway, University of London, § The Alan Turing Institute

"This paper focuses on test-time evasion attacks in the so-called **problem space**, where the challenge lies in modifying real input-space objects that correspond to an adversarial feature vector. The main challenge resides in the **inverse feature-mapping** problem since in many settings it is not possible to convert a feature vector into a problem-space object because the feature mapping function is neither invertible nor differentiable."

Why do we need all of this anyway? (first reason) [cont'd]

2020 IEEE Symposium on Security and Privacy

Intriguing Properties of Adversarial ML Attacks in the Problem Space

Fabio Pierazzi*†, Feargus Pendlebury*†‡§, Jacopo Cortellazzi†, Lorenzo Cavallaro†
† King's College London, ‡ Royal Holloway, University of London, § The Alan Turing Institute

"This paper focuses on test-time evasion attacks in the so-called **problem space**, where the challenge lies in modifying real input-space objects that correspond to an adversarial feature vector. The main challenge resides in the **inverse feature-mapping** problem since in many settings it is not possible to convert a feature vector into a problem-space object because the feature mapping function is neither invertible nor differentiable."

- This observation is well-founded, however…
- …if the attacker has access to the feature space, then such "problem" does not apply.

Perturbations in the feature space are **not unrealistic**: they simply require the attacker to compromise the ML system.

- This is possible [5], but it has <u>a high cost!</u>
- All past work considering "feature space" perturbations can be made valuable by assuming that the attack has a higher cost!

[5]: Eugene Bagdasaryan and Vitaly Shmatikov. 2021. Blind backdoors in deep learning models. In USENIX Sec. Symp

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Why do we need all of this anyway? (second reason)

- Most existing work in the ML-PWD domain has shortcomings, among which:
  - Some craft perturbations in the "feature" space (not impossible, but costly!)
  - Others assume strong attackers (full knowledge, or massive queries)
    - Liang et al. [57] took days!
  - No statistical validation (crucial for a fair evaluation!)

Paper (1st Author)	Year	Evasion space	ML-PWD types (F)	ML Algorithms	Defense	Datasets (reprod.)	Stat. Val.
Liang [57]	2016	Problem	F^c	SL	✗	1 (✗)	✗
Corona [30]	2017	Feature	F^r, F^c	SL	✓	1 (✓)	✗
Bahnsen [20]	2018	Problem	F^u	DL	✗	1 (✗)	✗
Shirazi [79]	2019	Feature	F^c	SL	✗	4 (✓)	✓*
Sabir [77]	2020	Problem	F^u	SL, DL	✓	1 (✗)	✗
Lee [55]	2020	Feature	F^c	SL	✓	1 (✓)	✗
Abdelnabi [8]	2020	Problem	F^r	DL	✓	1 (✓)	✗
Aleroud [11]	2020	Both	F^u	SL	✗	2 (✓)	✗
Song [81]	2021	Problem	F^c	SL	✓	1 (✓*)	✗
Bac [18]	2021	Feature	F^u	SL, DL	✗	1 (✗)	✗
Lin [59]	2021	Feature	F^c	DL	✓	1 (✓)	✗
O'Mara [67]	2021	Feature	F^r	SL	✗	1 (✓)	✗
Al-Qurashi [10]	2021	Feature	F^u, F^c	SL, DL	✗	4 (✓)	✗
Gressel [36]	2021	Feature	F^c	SL, DL	✓	1 (✗)	✗
Ours		Both	F^u, F^r, F^c	DL, SL	✓	2 (✓)	✓

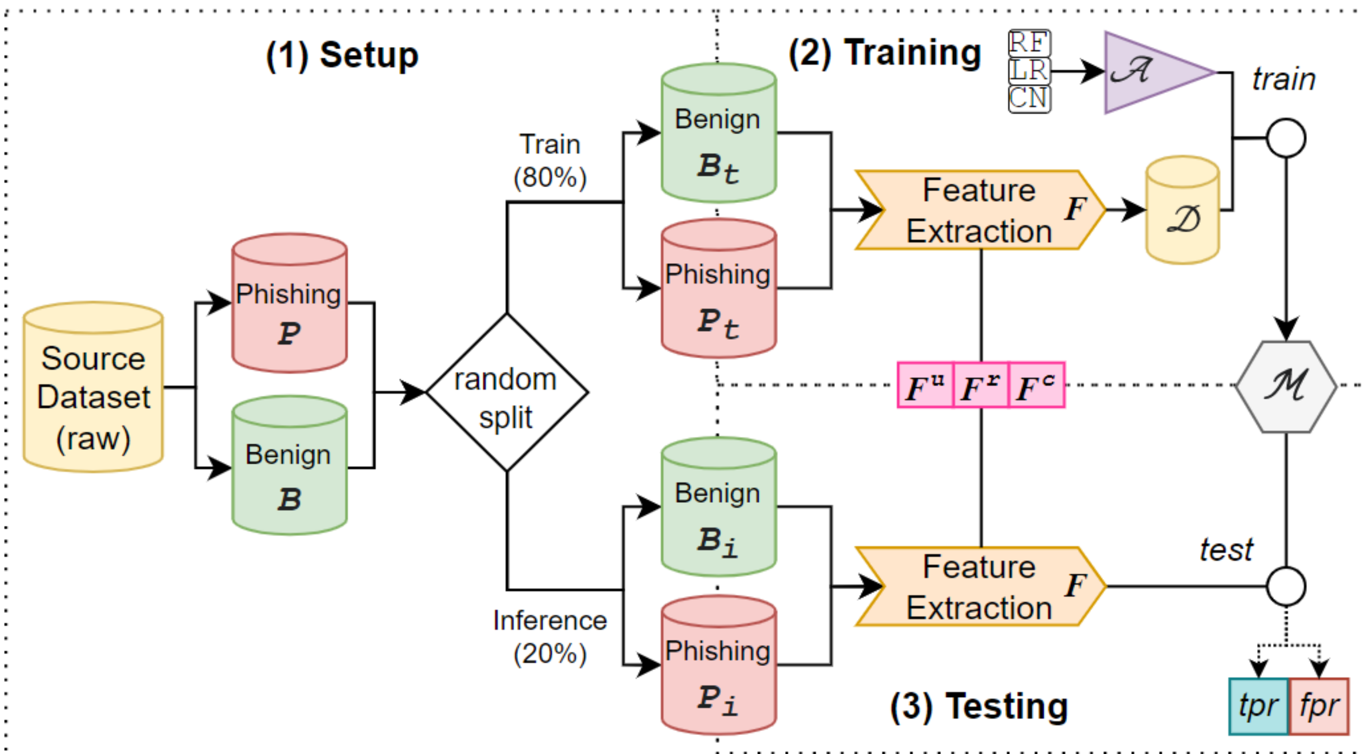
LIECHTENSTEIN

What is the true impact of realistic adversarial attacks against ML-PWD?

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation – Workflow

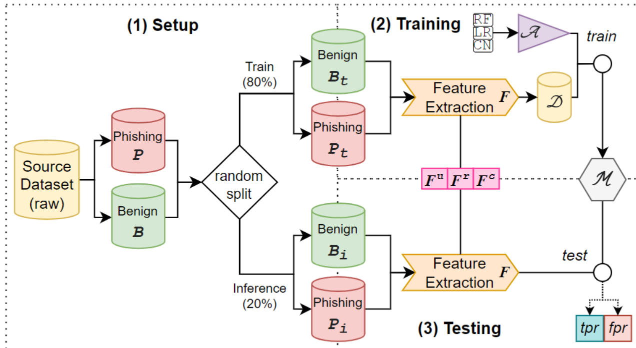
- Such attacks appear cheap, but are they effective? Let's assess their impact!
- We develop proficient ML-PWD (high *tpr*, low *fpr*)



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation – Baseline

- Such attacks appear cheap, but are they effective? Let's assess their impact!
- We develop proficient ML-PWD (high *tpr*, low *fpr*)



- Results comparable to the state-of-the-art ☺

- Let's attack such ML-PWD
  - The *tpr* will decrease!

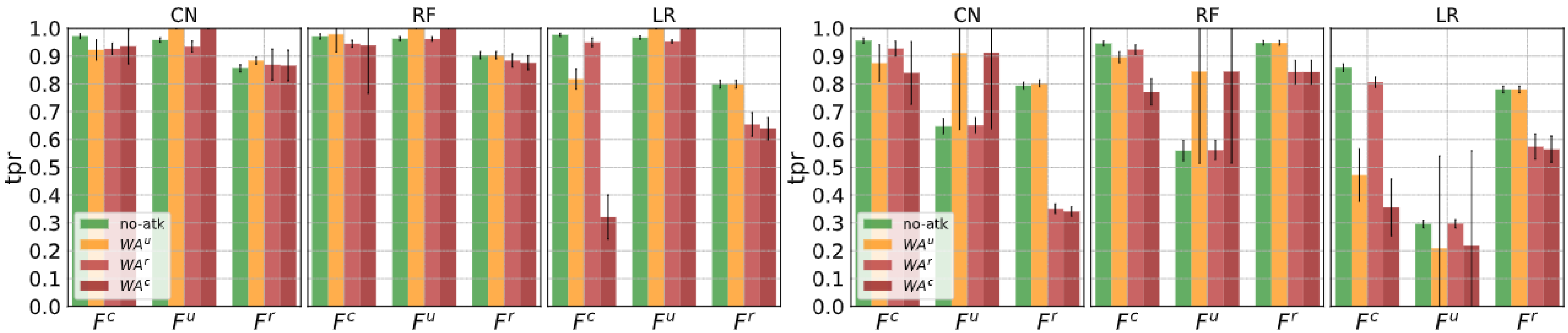
Table 3: Performance in non-adversarial settings, reported as the average (and std. dev.) tpr and fpr over the 50 trials.

$\mathcal{A}$	F	Zenodo tpr	Zenodo fpr	δphish tpr	δphish fpr
CN	F^u	$0.96_{\pm 0.008}$	$0.021_{\pm 0.0077}$	$0.55_{\pm 0.030}$	$0.037_{\pm 0.0076}$
	F^r	$0.88_{\pm 0.018}$	$0.155_{\pm 0.0165}$	$0.81_{\pm 0.019}$	$0.008_{\pm 0.0020}$
	F^c	$0.97_{\pm 0.006}$	$0.018_{\pm 0.0088}$	$0.93_{\pm 0.013}$	$0.005_{\pm 0.0025}$
RF	F^u	$0.98_{\pm 0.004}$	$0.007_{\pm 0.0055}$	$0.45_{\pm 0.022}$	$0.003_{\pm 0.0014}$
	F^r	$0.93_{\pm 0.013}$	$0.025_{\pm 0.0118}$	$0.94_{\pm 0.016}$	$0.006_{\pm 0.0025}$
	F^c	**$0.98_{\pm 0.006}$**	**$0.007_{\pm 0.0046}$**	**$0.97_{\pm 0.007}$**	**$0.001_{\pm 0.0011}$**
LR	F^u	$0.95_{\pm 0.009}$	$0.037_{\pm 0.0100}$	$0.24_{\pm 0.017}$	$0.011_{\pm 0.0026}$
	F^r	$0.82_{\pm 0.017}$	$0.144_{\pm 0.0171}$	$0.74_{\pm 0.025}$	$0.018_{\pm 0.0036}$
	F^c	$0.96_{\pm 0.007}$	$0.025_{\pm 0.0077}$	$0.81_{\pm 0.020}$	$0.013_{\pm 0.0037}$

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results – Are WsP effective?



(a) Impact of WA on the ML-PWD trained on Zenodo.

(b) Impact of WA on the ML-PWD trained on δPhish.

- In some cases, NO
  - This is *significant* because most past studies show ML-PWD being bypassed "regularly"!
- In some cases, VERY LITTLE
  - This is also significant, because even a 3% decrease in detection rate can be problematic when dealing with *thousands of samples*!
- In other cases, YES
  - This is very significant, because WsP are cheap and are likely to be exploited by attackers!

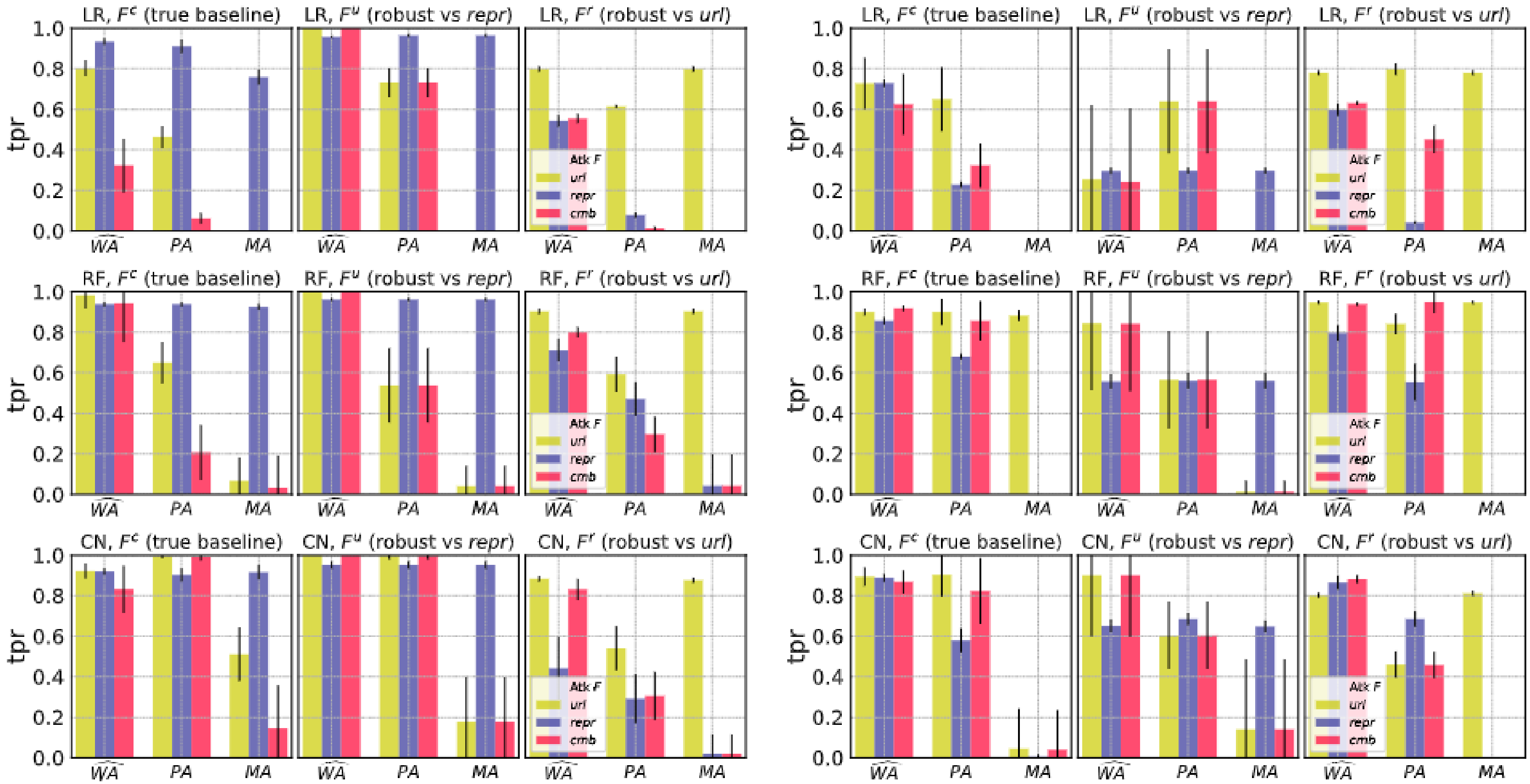
UNIVERSITÄT LIECHTENSTEIN

<u>Bottom line: no free lunch!</u>

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results – What about attacks in the other spaces?

In general, attacks in the other spaces (via PsP and MsP) are more disruptive…



(a) Zenodo. Each plot reports the tpr resulting from the 9 advanced attacks (i.e., $\widehat{\text{WA}}$, PA, MA) across the 50 trials. Colors denote the targeted features (u, r, c).

(b) δphish. Each plot reports the tpr resulting from the 9 advanced attacks (i.e., $\widehat{\text{WA}}$, PA, MA) across the 50 trials. Colors denote the targeted features (u, r, c).

However, such attacks also have a *higher cost*!

Will real attackers truly use them *just to evade* a ML-PWD?

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Demonstration – Evading a competition-grade ML-PWD

- https://tinyurl.com/spacephish-demo
- (https://spacephish.github.io)

UNIVERSITÄT LIECHTENSTEIN

Machine Learning in the Real World

UNIVERSITÄT LIECHTENSTEIN

G. Apruzzese, A.E. Cinà, A. Mitrokotsa, V. Shmatikov

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

How/where is ML used in the real world?

- A lot of domains use ML today:
  - Phishing Webpages Detection
  - Autonomous Driving (Computer Vision)
  - Translator (NLP)
  - Finance
  - Video Gaming
  - Filters (parental, content)
  - Recommender Systems
  - …
- However, most **research** on ML security:
  - Focuses on language models (text or speech), and CIFAR/ImageNet (images);
  - Considers only *deep neural networks*, whereas traditional ML algorithms (e.g., "Random Forests") are overlooked – despite being still used in practice!
  - Does not take into account the *costs* of attacks (or defenses).
  - Does not experiment on real systems

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

How/where is ML used in the real world? – Proof (1)

- Let's look at all adversarial ML papers (88) published in the top-4 cybersecurity conferences from 2019 until 2021, and see some trends…

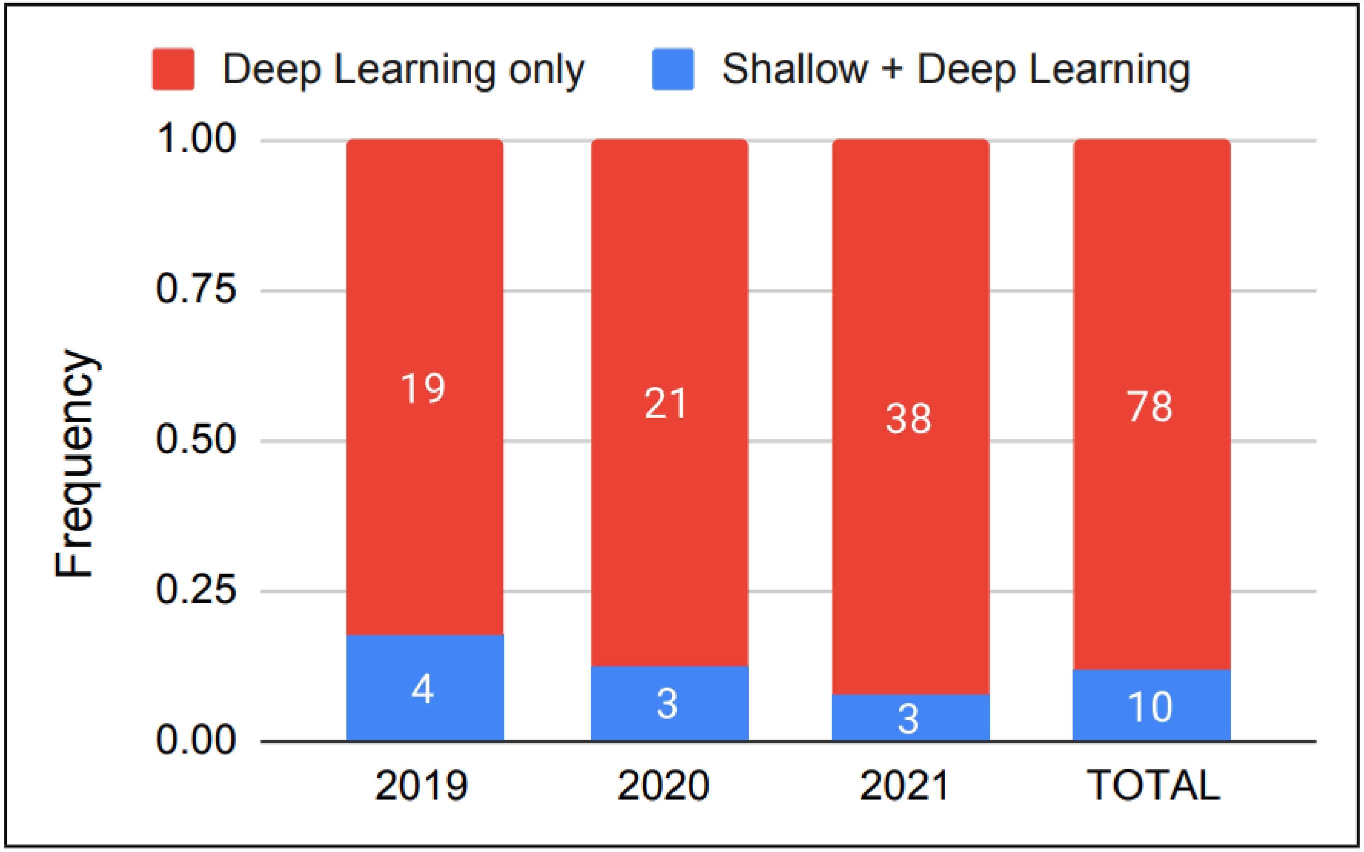


Fig. 8: G3: what is the considered ML *paradigm*?

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

How/where is ML used in the real world? – Proof (2)

- Let's look at <u>all</u> adversarial ML papers (88) published in the top-4 cybersecurity conferences from 2019 until 2021, and see some trends…

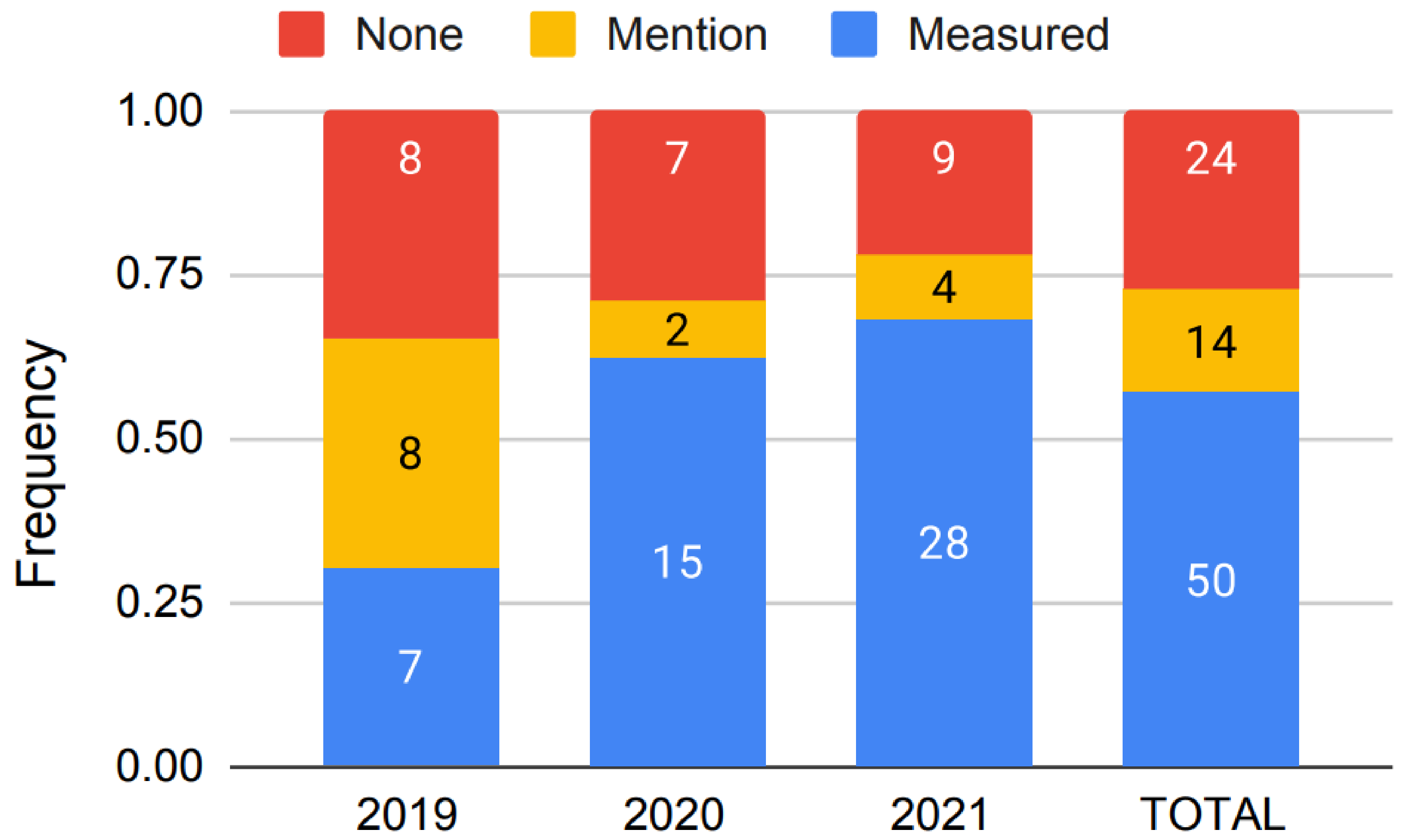


Fig. 9: G4: are the *costs* taken into account (in *any* way)?.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

How/where is ML used in the real world? – Proof (3)

- Let's look at all adversarial ML papers (88) published in the top-4 cybersecurity conferences from 2019 until 2021, and see some trends…

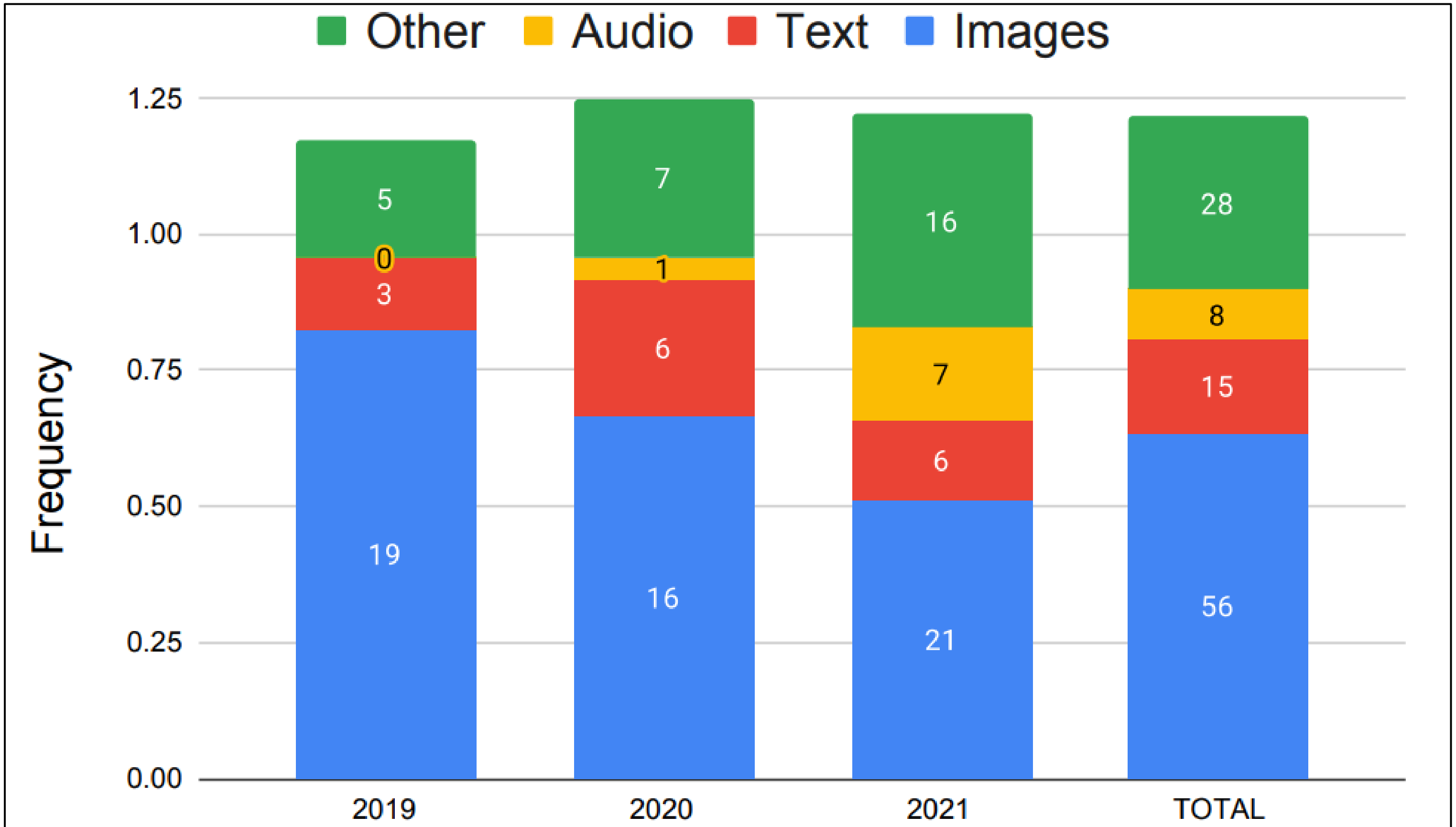


Fig. 10: What are the *data-types* considered in the evaluation?

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

How/where is ML used in the real world? – Proof (3)

- Let's look at all adversarial ML papers (88) published in the top-4 cybersecurity conferences from 2019 until 2021, and see some trends…

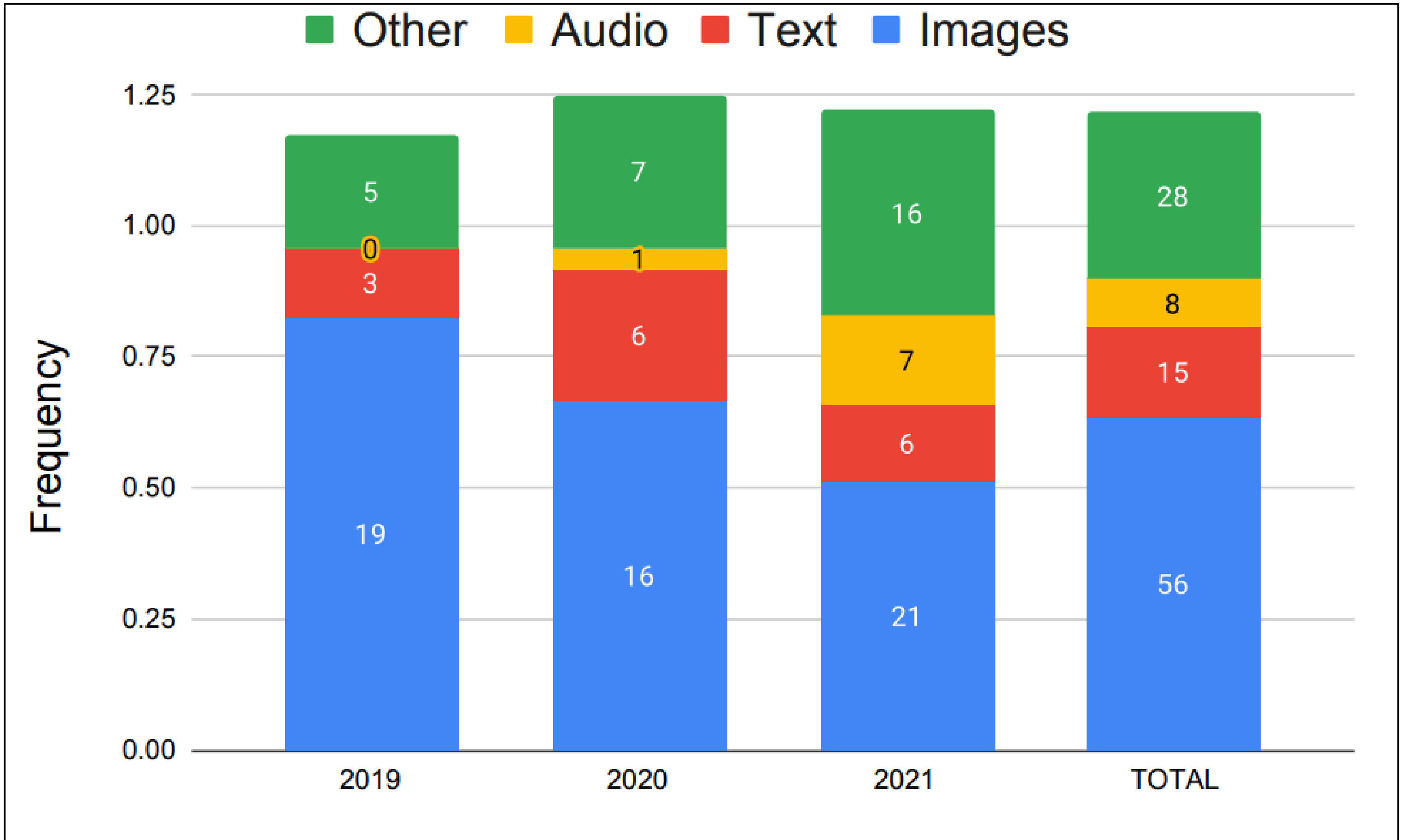


Fig. 10: What are the *data-types* considered in the evaluation?

Only 10 papers (!) focus on malware, phishing or network intrusion detection (in security conferences!)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

How/where is ML used in the real world? – Proof (4)

- Let's look at all adversarial ML papers (88) published in the top-4 cybersecurity conferences from 2019 until 2021, and see some trends…

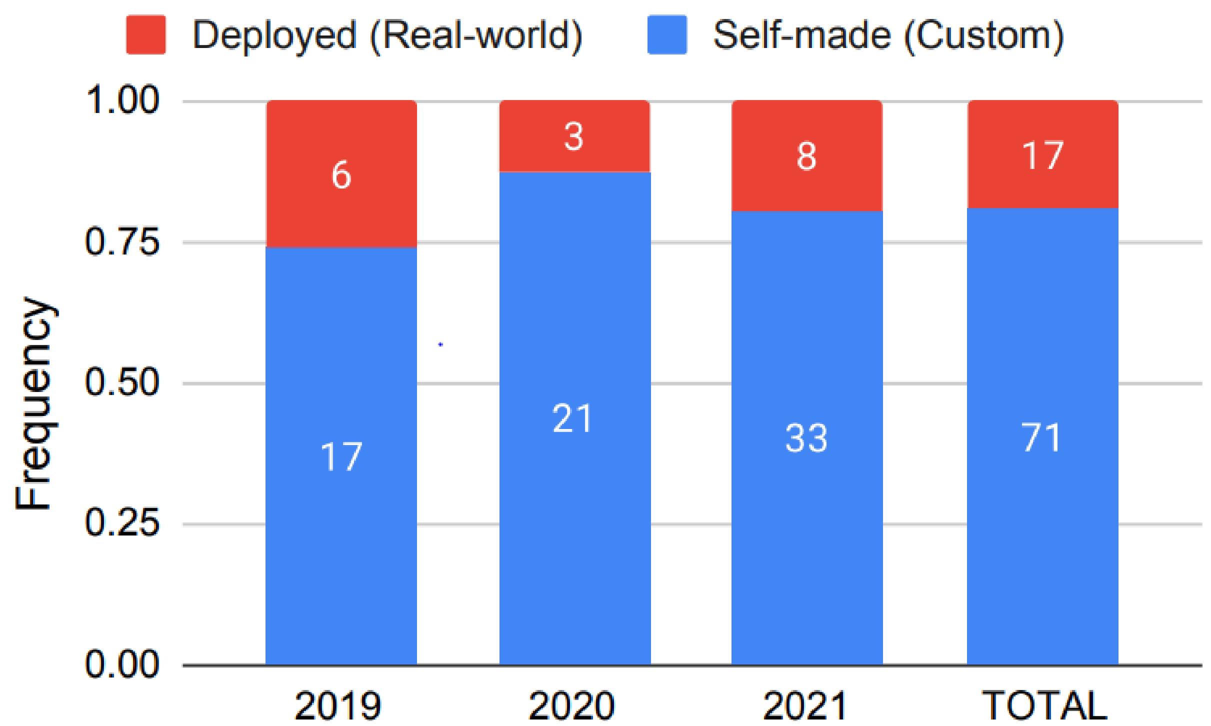


Fig. 13: Does the paper consider an ML model *deployed* in the real world?

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Most papers attack "benchmarks"

ML in practice

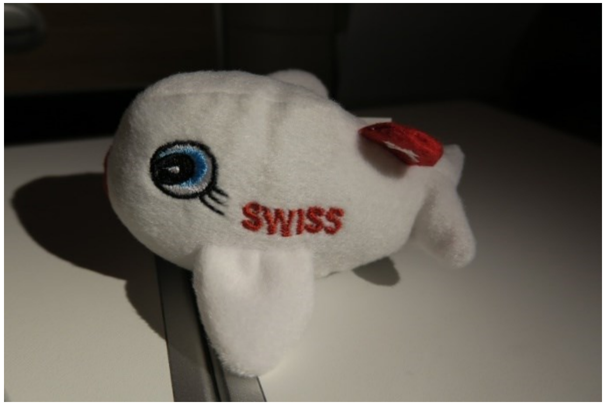
UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Most papers attack "benchmarks"

ML in practice

ML in research



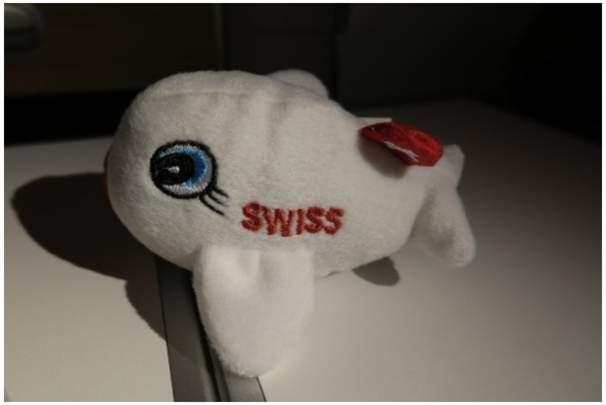
UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Most papers attack "benchmarks" (takeaway)

ML in practice

ML in research



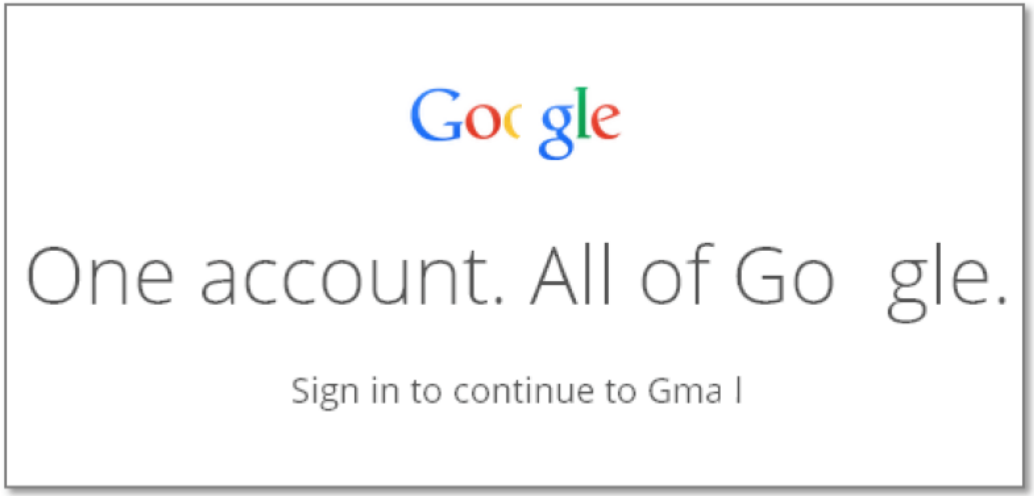
It's an ML system, not an ML model!

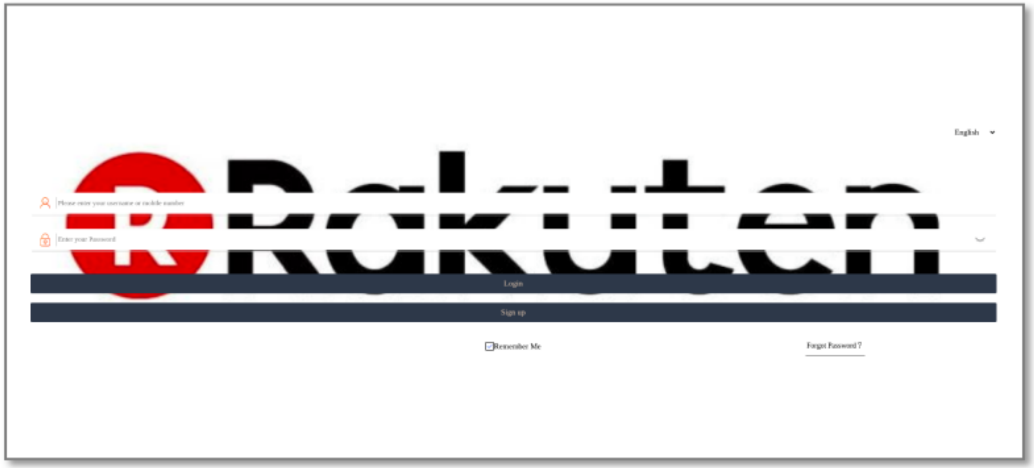
UNIVERSITÄT
LIECHTENSTEIN

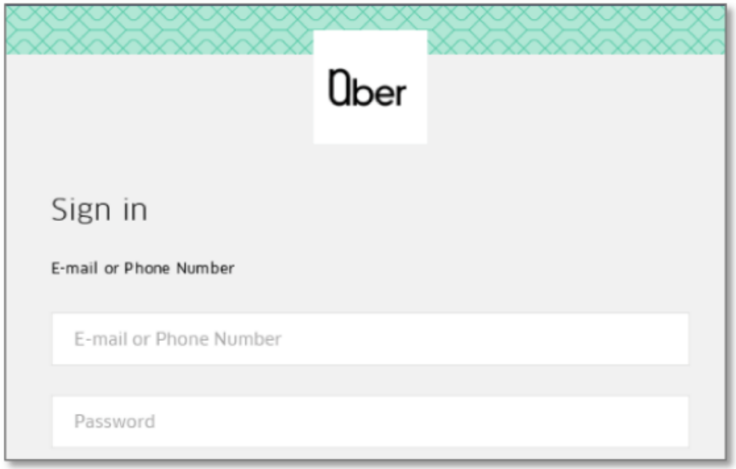
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Real attackers do not care about “evading” ML models

- Real systems can be fooled **without resorting to “gradient” based strategies.**







These phishing webpages were poorly classified by a commercial phishing detector!
(empowered by the all-so-mighty deep learning)

Some research papers attacking real systems…

Cracking classifiers for **evasion**: A case study on the google's phishing pages filter

B Liang, M Su, W You, W Shi, G Yang - Proceedings of the 25th …, 2016 - dl.acm.org

Various classifiers based on the machine learning techniques have been widely used in security applications. Meanwhile, they also became an attack target of adversaries. Many …

Proceedings of the 25th International Conference on World Wide Web (WWW). 2016.

Attacking automatic video analysis algorithms: A case study of google cloud video intelligence api

H Hosseini, B Xiao, A Clark… - Proceedings of the 2017 on …, 2017 - dl.acm.org

Due to the growth of video data on Internet, automatic video analysis has gained a lot of attention from academia as well as companies such as Facebook, Twitter and Google. In this paper, we examine the robustness of video analysis algorithms in adversarial settings. Specifically, we propose targeted attacks on two fundamental classes of video analysis algorithms, namely video classification and shot detection. We show that an adversary can subtly manipulate a video in such a way that a human observer would perceive the content …

Proceedings of the 2017 Workshop on Multimedia Privacy and Security (CCS Workshop). 2017.

Fall of **Giants**: How popular text-based MLaaS **fall** against a simple evasion attack

L Pajola, M Conti - … IEEE European Symposium on Security and …, 2021 - ieeexplore.ieee.org

The increased demand for machine learning applications made companies offer Machine-Learning-as-a-Service (MLaaS). In MLaaS (a market estimated 8000M USD by 2025), users …

IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2021.

Adversarial music: Real world audio adversary against wake-word detection system

J Li, S Qu, X Li, J Szurley, JZ Kolter… - Advances in Neural …, 2019 - proceedings.neurips.cc

… this suggests a real concern of **attack** against commercial grade **machine learning** algorithms, highlighting the importance of **adversarial** robustness from a …

Advances in Neural Information Processing Systems (2019).

UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

…have apparently little impact on future research (July 2022)

Cracking classifiers for **evasion**: A case study on the google's phishing pages filter

B Liang, M Su, W You, W Shi, G Yang - Proceedings of the 25th …, 2016 - dl.acm.org

Various classifiers based on the machine learning techniques have been widely used in security applications. Meanwhile, they also became an attack target of adversaries. Many …

Save Cite Cited by 58 Related articles All 6 versions

Proceedings of the 25th International Conference on World Wide Web (WWW). 2016.

Attacking automatic video analysis algorithms: A case study of google cloud video intelligence api

H Hosseini, B Xiao, A Clark… - Proceedings of the 2017 on …, 2017 - dl.acm.org

Due to the growth of video data on Internet, automatic video analysis has gained a lot of attention from academia as well as companies such as Facebook, Twitter and Google. In this paper, we examine the robustness of video analysis algorithms in adversarial settings. Specifically, we propose targeted attacks on two fundamental classes of video analysis algorithms, namely video classification and shot detection. We show that an adversary can subtly manipulate a video in such a way that a human observer would perceive the content …

Save Cite Cited by 23 Related articles All 8 versions

Proceedings of the 2017 Workshop on Multimedia Privacy and Security (CCS Workshop). 2017.

Fall of **Giants**: How popular text-based MLaaS fa… attack

L Pajola, M Conti - … IEEE European Symposium on Security and …, 2021 - ieeexplore.ieee.org

The increased demand for machine learning applications made companies offer Machine-Learning-as-a-Service (MLaaS). In MLaaS (a market estimated 8000M USD by 2025), users …

Save Cite Cited by 2 Related articles All 6 versions

IEEE European Symposium on Security and Privacy (EuroS&P). IEEE,

Adversarial music: Real world audio adversary against wake-word detection system

J Li, S Qu, X Li, J Szurley, JZ Kolter… - Advances in Neural …, 2019 - proceedings.neurips.cc

… this suggests a real concern of **attack** against commercial grade **machine learning** algorithms, highlighting the importance of **adversarial** robustness from a …

Save Cite Cited by 36 Related articles All 11 versions

Advances in Neural Information Processing Systems (2019).

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Why are (some) papers on real ML systems getting little attention?

- Not constructive for future research
  - The attack is against a “specific” system
  - You barely know what the system is actually doing
- Difficult to “beat” the same attack for future research
  - The real system gets patched immediately, and future research cannot “benchmark” on the same model, nor use the same attack methodology (which is *specific* for the targeted system)
- Difficult to “explain”
  - The real system is <u>always</u> a black-box from a researcher perspective, so it is difficult to explain what is actually happening “within” the system.
- Difficult to “map” to the “ML domain”
  - Is the attack targeting the ML model, the preprocessing, or some other component?
- The attacked systems are “niche”
  - The impact to the real world is marginal

Question: do you think it makes sense to always assume “worst-case” scenarios (i.e., the “Kerckhoff Principle”)?

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Some additional observations

TABLE III: List of original OBSERVATIONS made in our paper.

#	OBSERVATION	Ref.
1	ML models are only one component of ML systems.	§II-A
2	Academia and industry perceive adversarial ML differently.	§II-B
3	Economics is the main driver of practical cybersecurity.	§II-C
4	Evasion is achieved by bypassing all layers of an ML system.	§III-A
5	Evidence of adversarial examples in the wild is scarce.	§III-B
6	Queries are not always an effective measure of attack cost.	§III-C
7	Attackers use domain expertise and have broad goals.	§IV-B
8	Defenses can envision either strong or weak attackers.	§IV-C
9	Terminology is often imprecise and/or inconsistent.	§IV-D
10	Evading some ML systems can be very simple.	App.A-D

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Some additional observations

TABLE III: List of original OBSERVATIONS made in our paper.

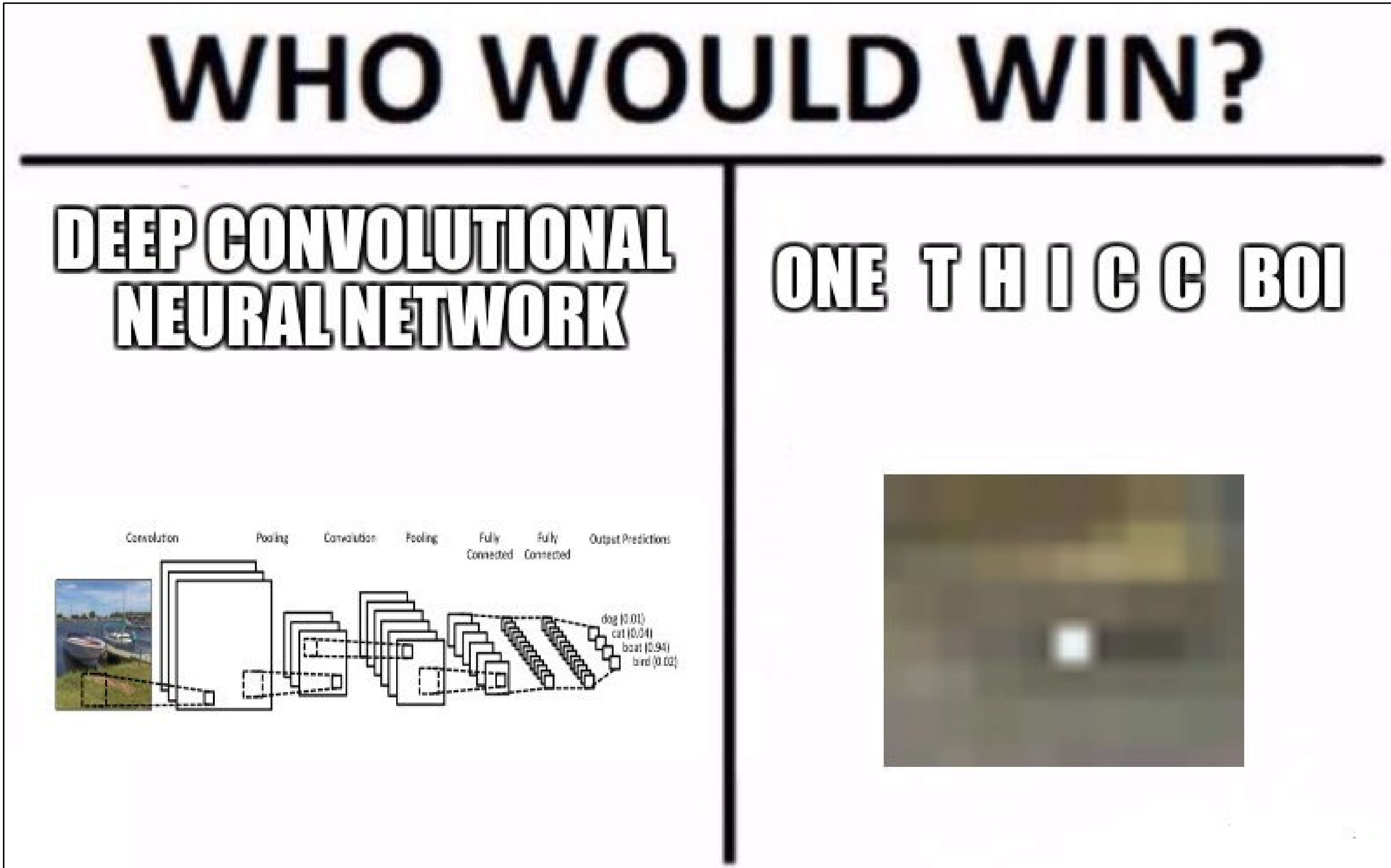
#	OBSERVATION	Ref.
1	ML models are only one component of ML systems.	§II-A
2	Academia and industry perceive adversarial ML differently.	§II-B
3	Economics is the main driver of practical cybersecurity.	§II-C
4	Evasion is achieved by bypassing all layers of an ML system.	§III-A
5	Evidence of adversarial examples in the wild is scarce.	§III-B
6	Queries are not always an effective measure of attack cost.	§III-C
7	Attackers use domain expertise and have broad goals.	§IV-B
8	Defenses can envision either strong or weak attackers.	§IV-C
9	Terminology is often imprecise and/or inconsistent.	§IV-D
10	Evading some ML systems can be very simple.	App.A-D

UNIVERSITÄT
LIECHTENSTEIN

Adversarial Attacks against
Humans and Machine Learning

UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li



UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Scenario

- Deep Learning (DL) is used for a plethora of applications.
- In some cases, however, the "decision making" is based on:
  - The <u>output</u> of a *DL model*
  - The interpretation of a *human* to such <u>output</u>

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Scenario

- Deep Learning (DL) is used for a plethora of applications.
- In some cases, however, the "decision making" is based on:
  - The output of a *DL model*
  - The interpretation of a *human* to such output
- Case in point: online marketplace
  - A person wants to sell an item (e.g., a car)
  - This person (i.e., the seller) uploads the images of such an item on an online marketplace
  - The marketplace automatically provides an estimate of the "value" of the corresponding item
    - This is done via DL [6]
  - Another person (i.e., a potential buyer) looks at the images, then looks at the "suggested" price, and determines whether to buy or not the corresponding item
    - The human uses the output of the DL model to make their decisions

[6] A. Varma, A. Sarma, S. Doshi, and R. Nair, "House price prediction using machine learning and neural networks," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, 2018,

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Attack – what if…

- What if the seller has malicious intentions?

→ The seller may want to induce the DL model to estimate a higher price

- Doing this by introducing “imperceptible” perturbations may trick the DL…
- …but not the human!

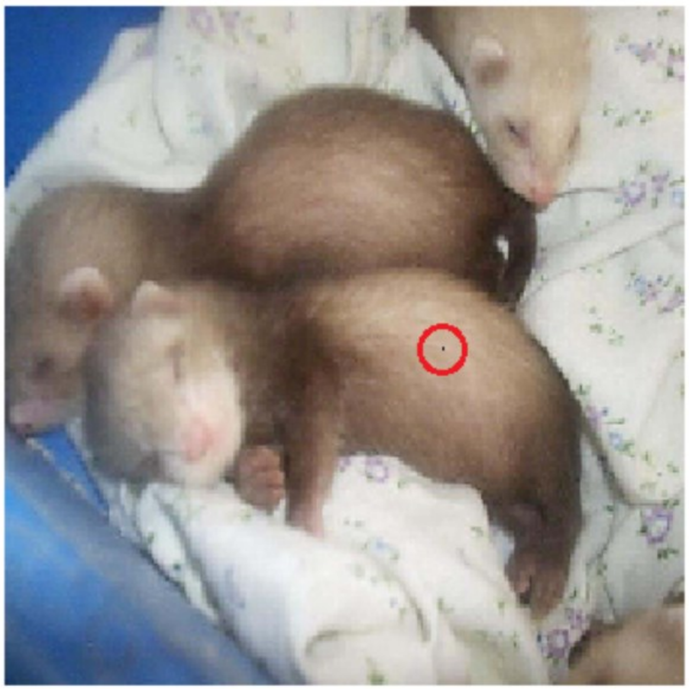
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Attack – what if…

- What if the seller has malicious intentions?

→ The seller may want to induce the DL model to estimate a higher price

- Doing this by introducing “imperceptible” perturbations may trick the DL…
- …but not the human!



Hamster(35.79%)
Nipple(42.36%)

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

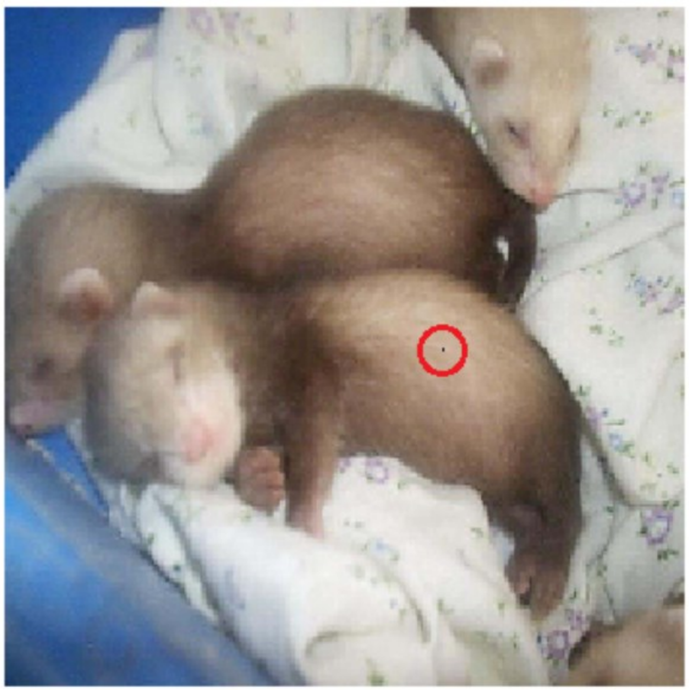
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Attack – what if…

- What if the seller has malicious intentions?

→ The seller may want to induce the DL model to estimate a higher price

- Doing this by introducing "imperceptible" perturbations may trick the DL…
- …but not the human!



Hamster(35.79%)

Nipple(42.36%)

In some cases, "imperceptible" perturbations **may not be what an attacker wants!**

This is especially true when there is a "human-in-the-loop".

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, *PhD*
giovanni.apruzzese@uni.li

Solution (high-level)

- If humans are involved in the “decision making” process, then such humans <u>will react</u> to clearly incorrect outputs of DL models.
  - Humans may suspect an adversarial <u>attack taking place</u>; or
  - They may think that the DL model is faulty, and hence <u>not trust/believe its output</u>
  - Both of the above are **detrimental** for the attacker!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Solution (high-level)

- If humans are involved in the "decision making" process, then such humans <u>will react</u> to clearly incorrect outputs of DL models.
  - Humans may suspect an adversarial <u>attack taking place</u>; or
  - They may think that the DL model is faulty, and hence <u>not trust/believe its output</u>
  - Both of the above are **detrimental** for the attacker!

(Malicious) solution: deceive both the human *and* the DL model!

- A DL model that thinks that a "FIAT Panda" is a "VW Polo" will output a very high price
  - But if the "perturbation" only affects a single pixel, nobody will fall for it!
- A FIAT Panda is clearly different than a VW Polo, so the perturbation (whatever it is) must be *perceived* by the human

→ The FIAT Panda must be changed in such a way that the human can be somewhat fooled
  - E.g.: the human should think that "it could be a Panda… but it could also be a Polo"

- FIAT Panda MSRP: ~10k $
- VW Polo MSRP: ~20k $

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Solution (low-level) – How to achieve this in practice?

Concept-based Adversarial Attacks

- The idea is using "explainability" techniques [7] to create adversarial examples.

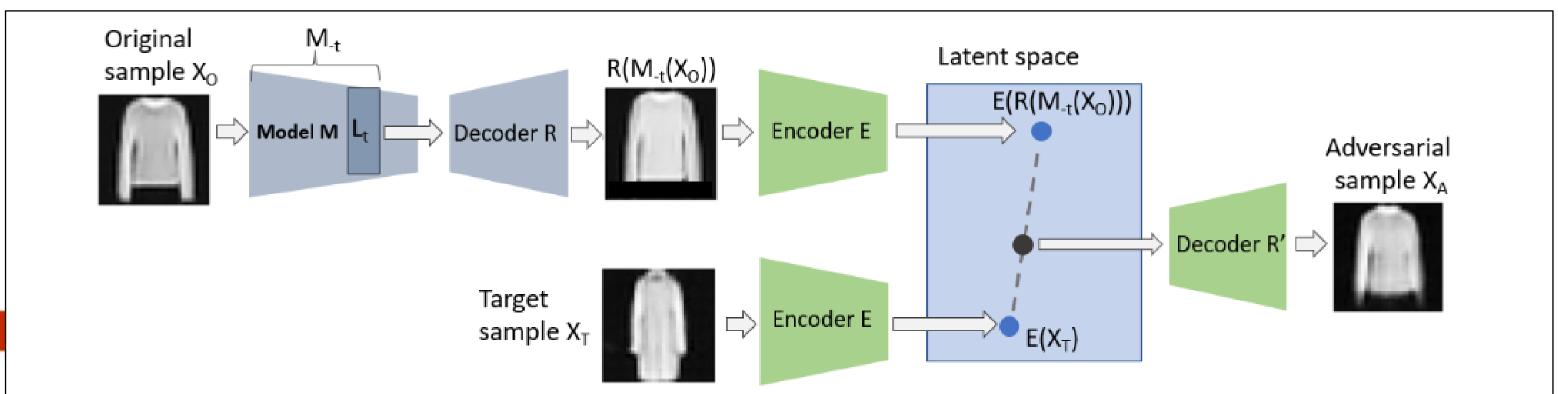
UNIVERSITÄT LIECHTENSTEIN

[7] J. Schneider and M. Vlachos, "Explaining neural networks by decoding layer activations," in *International Symposium on Intelligent Data Analysis*, 2021

Solution (low-level) – How to achieve this in practice?

Concept-based Adversarial Attacks

- The idea is using "explainability" techniques [7] to create adversarial examples.
- Requirements:
  - An "original sample" (i.e., a FIAT Panda)
  - A desired "target sample" (i.e., a VW Polo)
  - A given magnitude of the perturbation (neither too big nor too small)
    - If the FIAT Panda "becomes" a VW Polo, then the adversarial attack would be unfair
    - …and the "buyer" will complain ☺
  - The details of a DL model – based on Convolutional Neural Networks (CNN)
    - These attacks can be transferred!
    - IMPORTANT: the training procedure of the targeted CNN is *not* affected!
- Output: an "adversarial example" that is a mix between the original and target sample



[7] J. Schneider and M. Vlachos, "Explaining neural networks by decoding layer activations," in *International Symposium on Intelligent Data Analysis*, 2021

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Experiments – Objectives

Given the following:

- Original sample, $\boldsymbol{\mathcal{O}}$
- Target sample, $\boldsymbol{\mathcal{T}}$
- Adversarial sample, $\boldsymbol{\mathcal{A}}$

We design our experiments with three goals in mind:

1. *Misclassification:* the sample $\boldsymbol{\mathcal{A}}$ should be classified as the class of $\boldsymbol{\mathcal{T}}$ (which is different than the class of $\boldsymbol{\mathcal{O}}$)

2. *Resembling the target sample*: the sample $\boldsymbol{\mathcal{A}}$ should be similar to sample $\boldsymbol{\mathcal{T}}$ as measured by a given function f (e.g., the L2-norm)

3. *Remaining closer to the original sample:* the sample $\boldsymbol{\mathcal{A}}$ should be similar to sample $\boldsymbol{\mathcal{O}}$ as measured by a given function f (e.g., the L2-norm)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Experiments – Testbed

We consider two scenarios, each associated to a given dataset: *MNIST* and *Fashion-MNIST*.

Such datasets are used to train three CNN models:

- *VGG-11* ← our baseline
- *VGG-13*
- *Resnet-10*

We will showcase the adversarial transferability by using CNN with different architectures.

We consider four methods to generate $\mathcal{A}$ by "shifting" $\mathcal{O}$ towards $\mathcal{T}$, namely:

i. Autoencoder 1 (we "deconstruct" $\mathcal{O}$ and recreate it to resemble $\mathcal{T}$)
ii. Autoencoder 2 (as the previous one, but by using different layers)
iii. Classifier encoding (i.e., we shift $\mathcal{O}$ towards $\mathcal{T}$ in the last layer of the CNN)
iv. No encoding (i.e., linear interpolation from $\mathcal{O}$ to $\mathcal{T}$)

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results – Qualitative

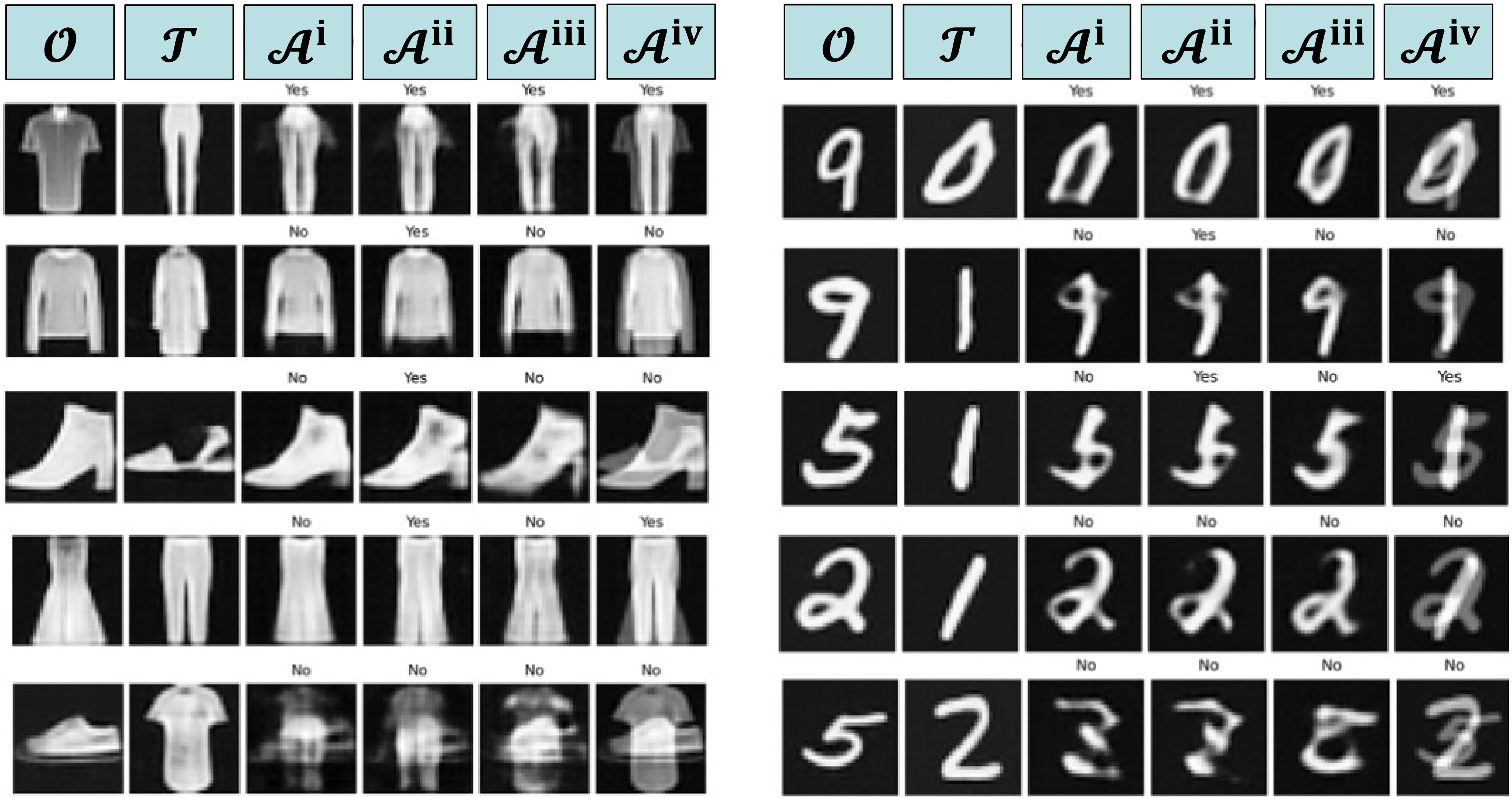


Fig. 2: Original, target and adversarial samples for different en-/decodings and interpolation for Fashion-MNIST(left) and MNIST(right). Yes/No indicates, whether the model got fooled by X_A, i.e. it outputs the class of X_T for X_A

UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results – Qualitative (takeaway)

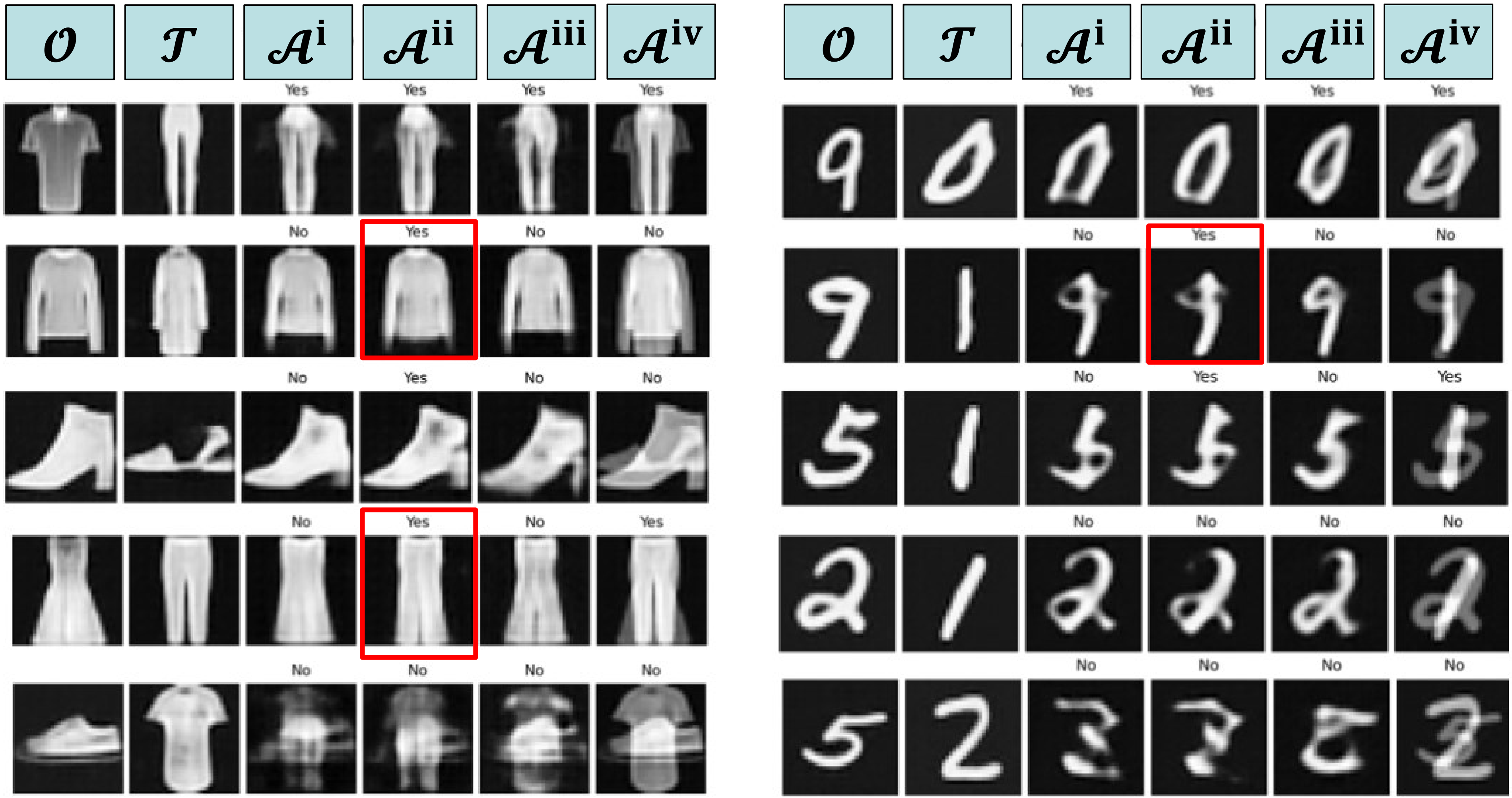


Fig. 2: Original, target and adversarial samples for different en-/decodings and interpolation for Fashion-MNIST(left) and MNIST(right). Yes/No indicates, whether the model got fooled by X_A, i.e. it outputs the class of X_T for X_A

Using the Autoencoder (ii) appears to be the best method to generate a suitable $\mathcal{A}$

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results – Quantitative

TABLE I. Results for MNIST and FashionMNIST.

Dataset	Generation Method	$\|\|\mathcal{A}-\mathcal{T}\|\|$ Similarity to $\mathcal{T}$	$\|\|\mathcal{A}-\mathcal{O}\|\|$ Similarity to $\mathcal{O}$	Acc(CNN) VGG-11	Acc(CNN) VGG-13	Acc(CNN) Resnet-10
MNIST	i (autoencoder 1)	$19.87_{\pm 1.794}$	$24.85_{\pm 0.11}$	$0.28_{\pm 0.081}$	$0.26_{\pm 0.079}$	$0.27_{\pm 0.084}$
	ii (autoencoder 2)	$20.41_{\pm 1.837}$	$24.73_{\pm 0.172}$	$0.21_{\pm 0.078}$	$0.2_{\pm 0.077}$	$0.2_{\pm 0.079}$
	iii (classifier encoding)	$24.38_{\pm 1.71}$	$24.71_{\pm 0.15}$	$0.44_{\pm 0.117}$	$0.41_{\pm 0.134}$	$0.42_{\pm 0.124}$
	iv (no encoding)	$12.42_{\pm 1.25}$	$24.73_{\pm 0.149}$	$0.08_{\pm 0.073}$	$0.11_{\pm 0.075}$	$0.09_{\pm 0.081}$
Fashion-MNIST	i (autoencoder 1)	$25.22_{\pm 1.365}$	$14.92_{\pm 0.048}$	$0.53_{\pm 0.065}$	$0.53_{\pm 0.065}$	$0.51_{\pm 0.06}$
	ii (autoencoder 2)	$25.84_{\pm 1.436}$	$14.85_{\pm 0.03}$	$0.57_{\pm 0.059}$	$0.58_{\pm 0.057}$	$0.56_{\pm 0.055}$
	iii (classifier encoding)	$27.23_{\pm 1.44}$	$14.84_{\pm 0.037}$	$0.64_{\pm 0.052}$	$0.62_{\pm 0.056}$	$0.62_{\pm 0.049}$
	iv (no encoding)	$20.83_{\pm 1.317}$	$14.95_{\pm 0.043}$	$0.42_{\pm 0.14}$	$0.44_{\pm 0.15}$	$0.41_{\pm 0.132}$

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results – Quantitative (takeaway)

TABLE I. Results for MNIST and FashionMNIST.

Dataset	Generation Method	$\|\|\mathcal{A}-\mathcal{T}\|\|$ Similarity to $\mathcal{T}$	$\|\|\mathcal{A}-\mathcal{O}\|\|$ Similarity to $\mathcal{O}$	Acc(CNN) VGG-11	Acc(CNN) VGG-13	Acc(CNN) Resnet-10
MNIST	i (autoencoder 1)	$19.87_{\pm 1.794}$	$24.85_{\pm 0.11}$	$0.28_{\pm 0.081}$	$0.26_{\pm 0.079}$	$0.27_{\pm 0.084}$
	ii (autoencoder 2)	$20.41_{\pm 1.837}$	$24.73_{\pm 0.172}$	$0.21_{\pm 0.078}$	$0.2_{\pm 0.077}$	$0.2_{\pm 0.079}$
	iii (classifier encoding)	$24.38_{\pm 1.71}$	$24.71_{\pm 0.15}$	$0.44_{\pm 0.117}$	$0.41_{\pm 0.134}$	$0.42_{\pm 0.124}$
	iv (no encoding)	$12.42_{\pm 1.25}$	$24.73_{\pm 0.149}$	$0.08_{\pm 0.073}$	$0.11_{\pm 0.075}$	$0.09_{\pm 0.081}$
Fashion-MNIST	i (autoencoder 1)	$25.22_{\pm 1.365}$	$14.92_{\pm 0.048}$	$0.53_{\pm 0.065}$	$0.53_{\pm 0.065}$	$0.51_{\pm 0.06}$
	ii (autoencoder 2)	$25.84_{\pm 1.436}$	$14.85_{\pm 0.03}$	$0.57_{\pm 0.059}$	$0.58_{\pm 0.057}$	$0.56_{\pm 0.055}$
	iii (classifier encoding)	$27.23_{\pm 1.44}$	$14.84_{\pm 0.037}$	$0.64_{\pm 0.052}$	$0.62_{\pm 0.056}$	$0.62_{\pm 0.049}$
	iv (no encoding)	$20.83_{\pm 1.317}$	$14.95_{\pm 0.043}$	$0.42_{\pm 0.14}$	$0.44_{\pm 0.15}$	$0.41_{\pm 0.132}$

- *Accuracy:* the biggest drop is for "no encoding" (which are the most easily recognizable)

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results – Quantitative (takeaway)

TABLE I. Results for MNIST and FashionMNIST.

Dataset	Generation Method	$\|\|\mathcal{A}-\mathcal{T}\|\|$ Similarity to $\mathcal{T}$	$\|\|\mathcal{A}-\mathcal{O}\|\|$ Similarity to $\mathcal{O}$	Acc(CNN) VGG-11	Acc(CNN) VGG-13	Acc(CNN) Resnet-10
MNIST	i (autoencoder 1)	$19.87_{\pm 1.794}$	$24.85_{\pm 0.11}$	$0.28_{\pm 0.081}$	$0.26_{\pm 0.079}$	$0.27_{\pm 0.084}$
	ii (autoencoder 2)	$20.41_{\pm 1.837}$	$24.73_{\pm 0.172}$	$0.21_{\pm 0.078}$	$0.2_{\pm 0.077}$	$0.2_{\pm 0.079}$
	iii (classifier encoding)	$24.38_{\pm 1.71}$	$24.71_{\pm 0.15}$	$0.44_{\pm 0.117}$	$0.41_{\pm 0.134}$	$0.42_{\pm 0.124}$
	iv (no encoding)	$12.42_{\pm 1.25}$	$24.73_{\pm 0.149}$	$0.08_{\pm 0.073}$	$0.11_{\pm 0.075}$	$0.09_{\pm 0.081}$
Fashion-MNIST	i (autoencoder 1)	$25.22_{\pm 1.365}$	$14.92_{\pm 0.048}$	$0.53_{\pm 0.065}$	$0.53_{\pm 0.065}$	$0.51_{\pm 0.06}$
	ii (autoencoder 2)	$25.84_{\pm 1.436}$	$14.85_{\pm 0.03}$	$0.57_{\pm 0.059}$	$0.58_{\pm 0.057}$	$0.56_{\pm 0.055}$
	iii (classifier encoding)	$27.23_{\pm 1.44}$	$14.84_{\pm 0.037}$	$0.64_{\pm 0.052}$	$0.62_{\pm 0.056}$	$0.62_{\pm 0.049}$
	iv (no encoding)	$20.83_{\pm 1.317}$	$14.95_{\pm 0.043}$	$0.42_{\pm 0.14}$	$0.44_{\pm 0.15}$	$0.41_{\pm 0.132}$

- *Accuracy:* the biggest drop is for “no encoding” (which are the most easily recognizable)
- *Transferability*: the accuracy is (essentially) the same for all CNN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results – Quantitative (takeaway)

TABLE I. Results for MNIST and FashionMNIST.

Dataset	Generation Method	$\|\|\mathcal{A}-\mathcal{T}\|\|$ Similarity to $\mathcal{T}$	$\|\|\mathcal{A}-\mathcal{O}\|\|$ Similarity to $\mathcal{O}$	Acc(CNN) VGG-11	Acc(CNN) VGG-13	Acc(CNN) Resnet-10
MNIST	i (autoencoder 1)	$19.87_{\pm 1.794}$	$24.85_{\pm 0.11}$	$0.28_{\pm 0.081}$	$0.26_{\pm 0.079}$	$0.27_{\pm 0.084}$
	ii (autoencoder 2)	$20.41_{\pm 1.837}$	$24.73_{\pm 0.172}$	$0.21_{\pm 0.078}$	$0.2_{\pm 0.077}$	$0.2_{\pm 0.079}$
	iii (classifier encoding)	$24.38_{\pm 1.71}$	$24.71_{\pm 0.15}$	$0.44_{\pm 0.117}$	$0.41_{\pm 0.134}$	$0.42_{\pm 0.124}$
	iv (no encoding)	$12.42_{\pm 1.25}$	$24.73_{\pm 0.149}$	$0.08_{\pm 0.073}$	$0.11_{\pm 0.075}$	$0.09_{\pm 0.081}$
Fashion-MNIST	i (autoencoder 1)	$25.22_{\pm 1.365}$	$14.92_{\pm 0.048}$	$0.53_{\pm 0.065}$	$0.53_{\pm 0.065}$	$0.51_{\pm 0.06}$
	ii (autoencoder 2)	$25.84_{\pm 1.436}$	$14.85_{\pm 0.03}$	$0.57_{\pm 0.059}$	$0.58_{\pm 0.057}$	$0.56_{\pm 0.055}$
	iii (classifier encoding)	$27.23_{\pm 1.44}$	$14.84_{\pm 0.037}$	$0.64_{\pm 0.052}$	$0.62_{\pm 0.056}$	$0.62_{\pm 0.049}$
	iv (no encoding)	$20.83_{\pm 1.317}$	$14.95_{\pm 0.043}$	$0.42_{\pm 0.14}$	$0.44_{\pm 0.15}$	$0.41_{\pm 0.132}$

- *Accuracy:* the biggest drop is for "no encoding" (which are the most easily recognizable)
- *Transferability*: the accuracy is (essentially) the same for all CNN
- *Similarity to* $\mathcal{T}$: *classifier encoding* are the least similar to $\mathcal{T}$

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results – Quantitative (takeaway)

TABLE I. Results for MNIST and FashionMNIST.

Dataset	Generation Method	$\|\|\mathcal{A}-\mathcal{T}\|\|$ Similarity to $\mathcal{T}$	$\|\|\mathcal{A}-\mathcal{O}\|\|$ Similarity to $\mathcal{O}$	Acc(CNN) VGG-11	Acc(CNN) VGG-13	Acc(CNN) Resnet-10
MNIST	i (autoencoder 1)	$19.87_{\pm 1.794}$	$24.85_{\pm 0.11}$	$0.28_{\pm 0.081}$	$0.26_{\pm 0.079}$	$0.27_{\pm 0.084}$
	ii (autoencoder 2)	$20.41_{\pm 1.837}$	$24.73_{\pm 0.172}$	$0.21_{\pm 0.078}$	$0.2_{\pm 0.077}$	$0.2_{\pm 0.079}$
	iii (classifier encoding)	$24.38_{\pm 1.71}$	$24.71_{\pm 0.15}$	$0.44_{\pm 0.117}$	$0.41_{\pm 0.134}$	$0.42_{\pm 0.124}$
	iv (no encoding)	$12.42_{\pm 1.25}$	$24.73_{\pm 0.149}$	$0.08_{\pm 0.073}$	$0.11_{\pm 0.075}$	$0.09_{\pm 0.081}$
Fashion-MNIST	i (autoencoder 1)	$25.22_{\pm 1.365}$	$14.92_{\pm 0.048}$	$0.53_{\pm 0.065}$	$0.53_{\pm 0.065}$	$0.51_{\pm 0.06}$
	ii (autoencoder 2)	$25.84_{\pm 1.436}$	$14.85_{\pm 0.03}$	$0.57_{\pm 0.059}$	$0.58_{\pm 0.057}$	$0.56_{\pm 0.055}$
	iii (classifier encoding)	$27.23_{\pm 1.44}$	$14.84_{\pm 0.037}$	$0.64_{\pm 0.052}$	$0.62_{\pm 0.056}$	$0.62_{\pm 0.049}$
	iv (no encoding)	$20.83_{\pm 1.317}$	$14.95_{\pm 0.043}$	$0.42_{\pm 0.14}$	$0.44_{\pm 0.15}$	$0.41_{\pm 0.132}$

- *Accuracy:* the biggest drop is for “no encoding” (which are the most easily recognizable)
- *Transferability*: the accuracy is (essentially) the same for all CNN
- *Similarity to $\mathcal{T}$*: *classifier encoding* are the least similar to $\mathcal{T}$
- *Similarity to $\mathcal{O}$*: all methods appear to have same results

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Future Work

- **Human evaluation**
  - We want to submit the adversarial samples $\mathcal{A}$ to real humans and ask for their opinion
- **Defense** and **augmentation**
  - Through *adversarial training*, it is possible to use $\mathcal{A}$ to defend against similar attacks
  - Alternatively, it is possible to use $\mathcal{A}$ to augment the training dataset and (potentially) increase the baseline performance of the CNN
- **Different data**
  - We only considered MNIST and FashionMNIST, but more datasets exist (e.g., CIFAR) which can be used to devise more intriguing experiments (with real FIAT Pandas and VW Polos!)
- **Other domains**
  - We only investigated CNN that were analyzing images. However, the same principles can be applied also in other domains (i.e., malware analysis)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Future Work

- **Human evaluation**
  - We want to submit the adversarial samples $\mathcal{A}$ to real humans and ask for their opinion

- **Defense** and **augmentation**
  - Through *adversarial training*, it is possible to use $\mathcal{A}$ to defend against similar attacks
  - Alternatively, it is possible to use $\mathcal{A}$ to augment the training dataset and (potentially) increase the baseline performance of the CNN

- **Different data**
  - We only considered MNIST and FashionMNIST, but more datasets exist (e.g., CIFAR) which can be used to devise more intriguing experiments (with real FIAT Pandas and VW Polos!)

- **Other domains**
  - We only investigated CNN that were analyzing images. However, the same principles can be applied also in other domains (i.e., malware analysis)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Human validation

Sample S

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Human validation – confused?

Sample S

- is sample S representing a 4 or a 9?

	1	2	3	4	5	6	7	
100% 4	○	○	○	○	○	○	○	100% 9

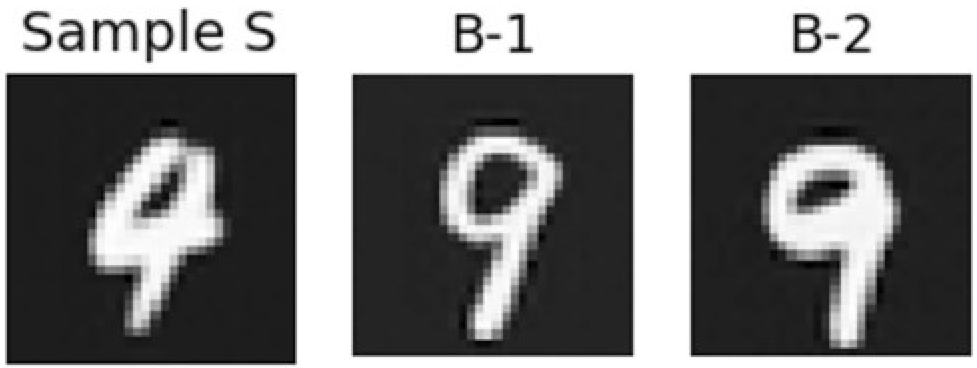
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Human validation – source and target?

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Human validation – source and target?



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Human validation – source and target?

Is sample S more similar to A-1 or to A-2? Look carefully! *

	1	2	3	4	
More similar to A-1	○	○	○	○	More similar to A-2

Is sample S more similar to B-1 or B-2? Look carefully! *

	1	2	3	4	
More similar to B-1	○	○	○	○	More similar to B-2

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Human validation – truth

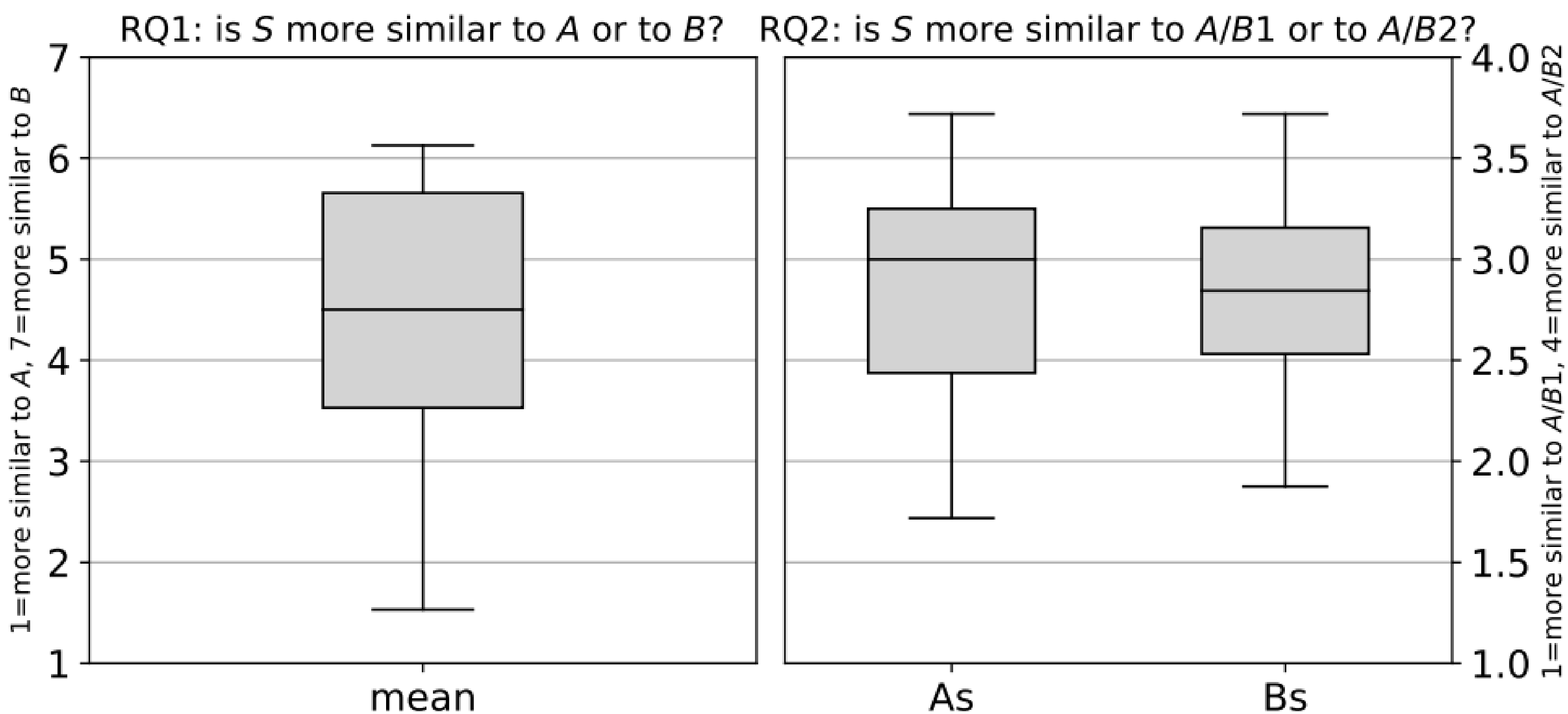
Original

Target

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Human validation – results

- We created 46 of such questions by randomly picking diverse "Original" and "Target" samples, and we have 31 Amazon Mechanical Turk workers provide their answers.



They are confused!

They can identify the correct *original* and *target* samples

UNIVERSITÄT LIECHTENSTEIN

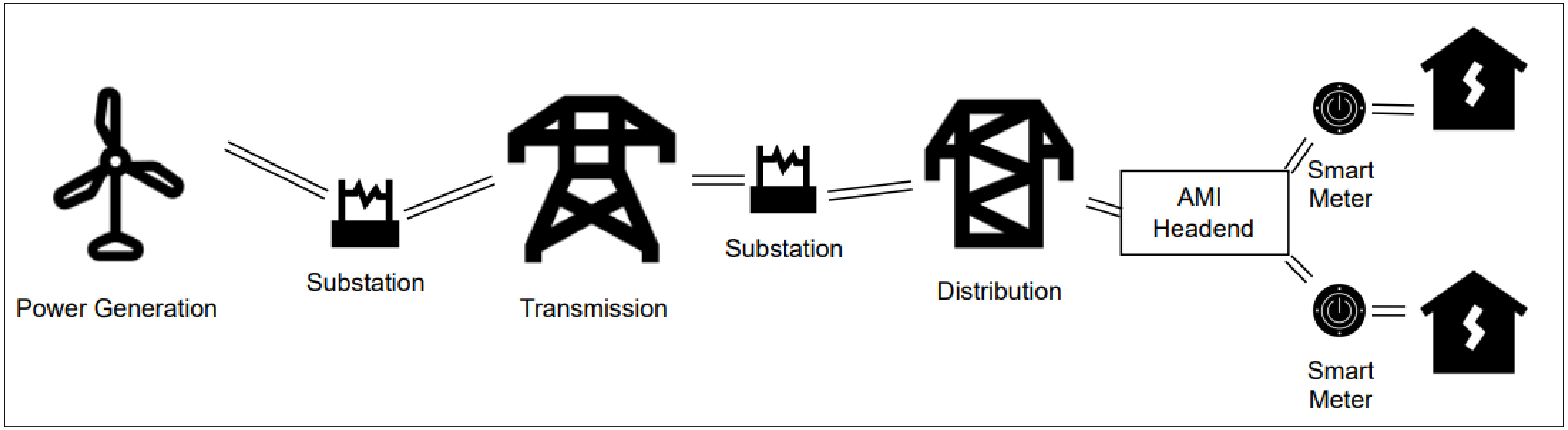
Cybersecurity in the Smart Grid (in Practice)

UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

The Smart Grid (SG) – aka: the lifeforce of our society

- The SG has seen the take-off of digitalisation in recent years



- Pros:
  - Fine-grained operation
  - Better efficiency/reliability
- Cons:
  - Enormous attack surface
  - Attractive target for cyber-attacks
- Example: Ukraine 2015 → **225’000** households affected
- Worst case scenario cyber attack on SG in Switzerland → **12 billion CHF = 2% of GDP**

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What do we (don't) know?

Abundant research efforts studied the cybersecurity of the SG:

- Literature reviews
  - Based on scientific papers -> **limited practical relevance**
  - E.g. elaboration of SG cyber-security strategy (El Mrabet et al., 2018)
- Original Attacks (and countermeasures)
  - Often studied in testbeds -> **no real-world confirmation**
  - E.g. Mathematical analysis of impact (Xiang et al., 2017)
- Interviews
  - Few studies, of limited scope (our outdated) -> **no comprehensive overview (of today's SG)**
  - E.g. Stakeholder perspectives (Fischer-Hübner et al., 2021) or information sharing networks (Randall and Allen, 2021)

In this work, we provide:
- the (internal) perspective of SG's **practitioners**;
- an **holistic** view on the problem.

→ High practical relevance, and constructive for future endeavours

Holistic view – why?

The SG is a complex system, which entails various stakeholders.

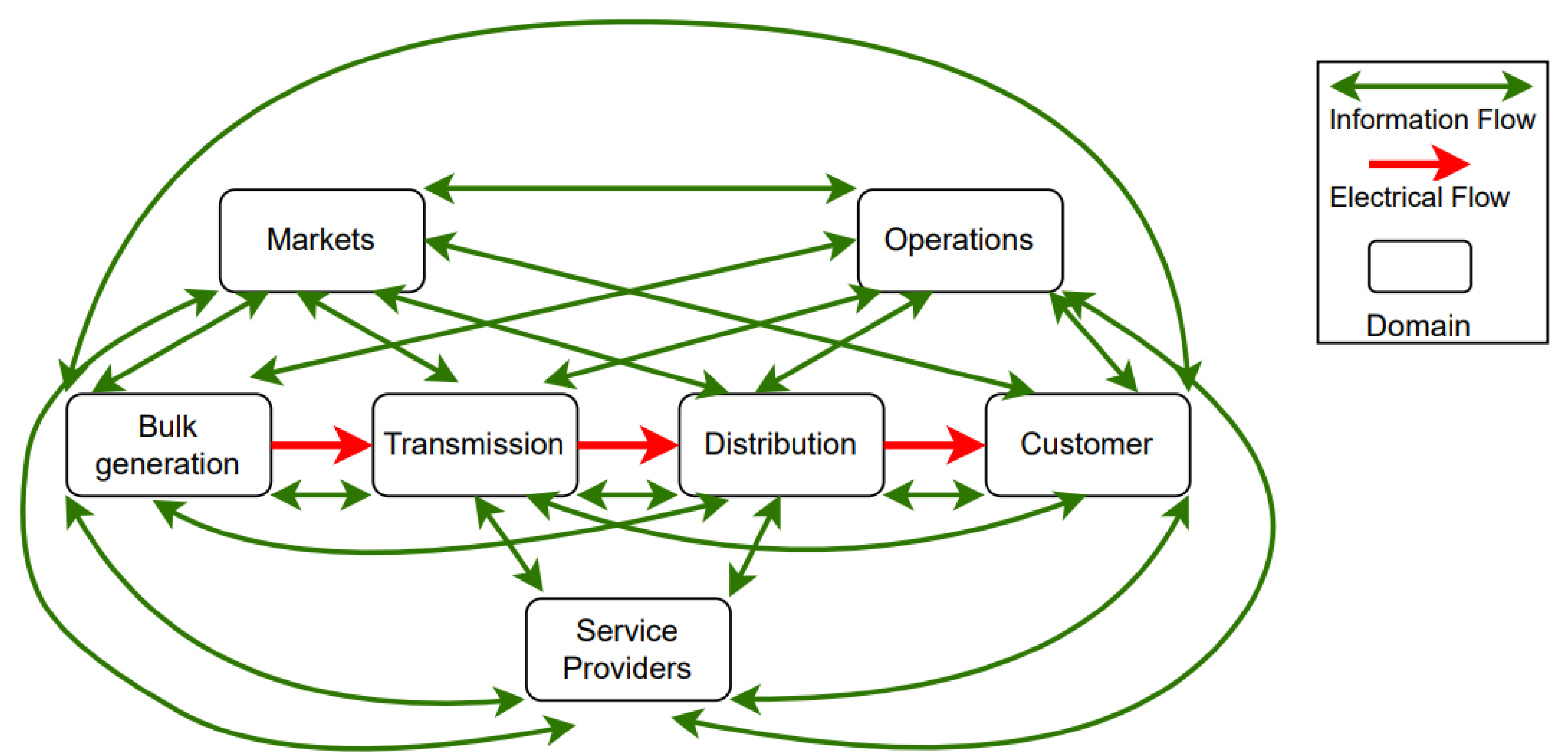


Fig. 3: The NIST conceptual model of the SG, spanning across 7 domains—all of which are covered in our research.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Our objective

- We began our research by asking ourselves a broad research question:

  "What is the state-of-the-art of cyber-security in the European SG?"

- We aimed to elucidate:
  1. Experiences with *past cyber-attacks*
  2. General security landscape of *companies operating the SG*
  3. Cyber-security related *risk-assessment strategies*
  4. *Perceived threat* of various attack scenarios
  5. *New technologies* and trends in the SG
  6. The opinion of *public authorities* w.r.t. the companies' managed cybersecurity

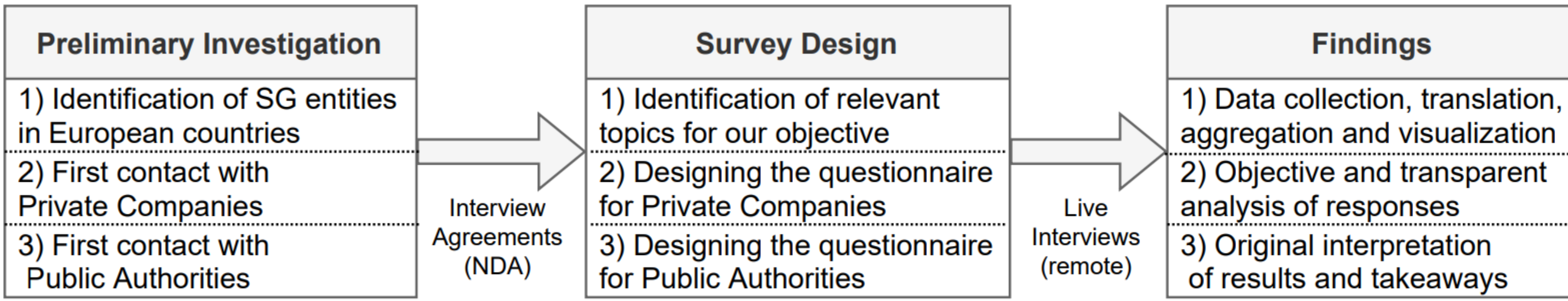
- As we will show, however, some finding surprised us

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we did

- Structured interviews with 18 entities related to the SG:
  - 14 private companies (operating the SG in diverse countries in Europe)
  - 4 public authorities (operating in the countries of the private companies' headquarters)

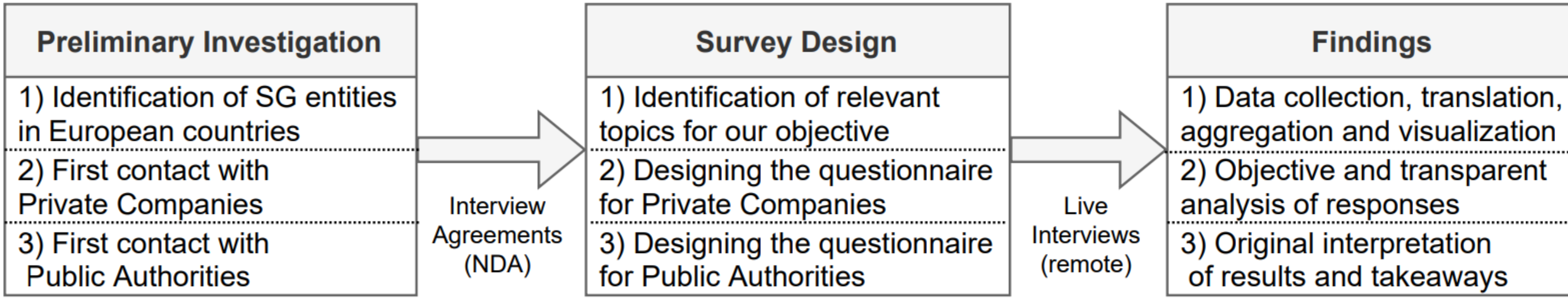


UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we did (& challenges)

- Structured interviews with 18 entities related to the SG:
  - 14 private companies (operating the SG in diverse countries in Europe)
  - 4 public authorities (operating in the countries of the private companies’ headquarters)



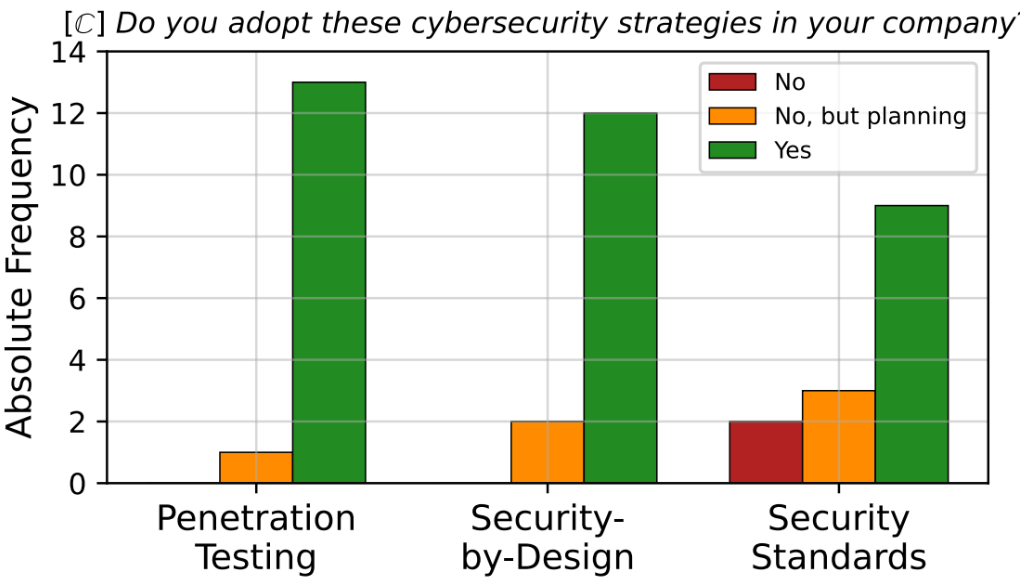
Challenges

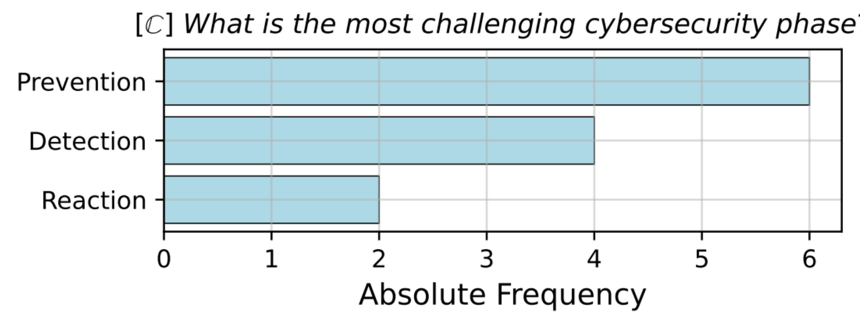
- We aimed to interview more than 30 companies, but only 14 accepted
- 5 companies agreed to help us only after phone calls lasting more than 60 minutes.
- Only 5 of the interviews with the 14 private companies were carried out on the initial scheduled date
- We sent a total of 145 emails between Nov. 2021 and Feb. 2022.
- Different language

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

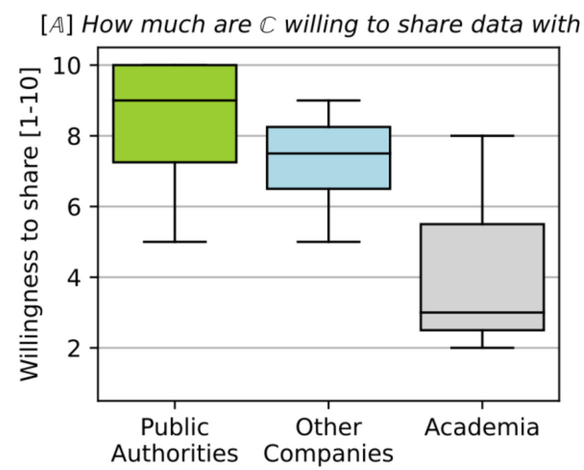
Findings – generic (ℂ = Private Companies, 𝔸= Public Authorities)





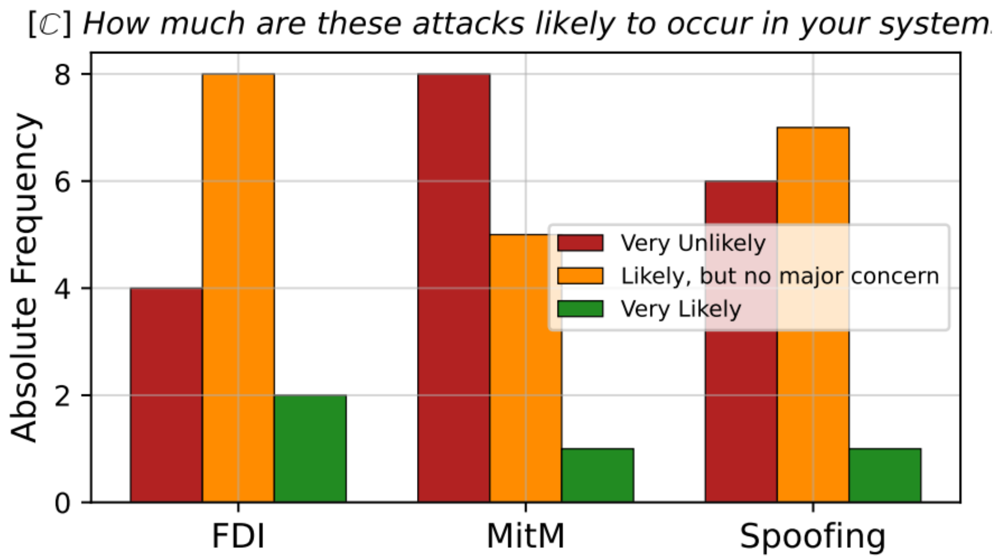
Mid-/Top-level management	
Option	*Freq.*
They are fully aware of the risks and prioritise cyber-security	64.29%
They are fully aware of the risks, but cyber-security is not a priority	21.43%
They are not aware of the risks, but are educated on the topic	7.14%
No answer	7.14%

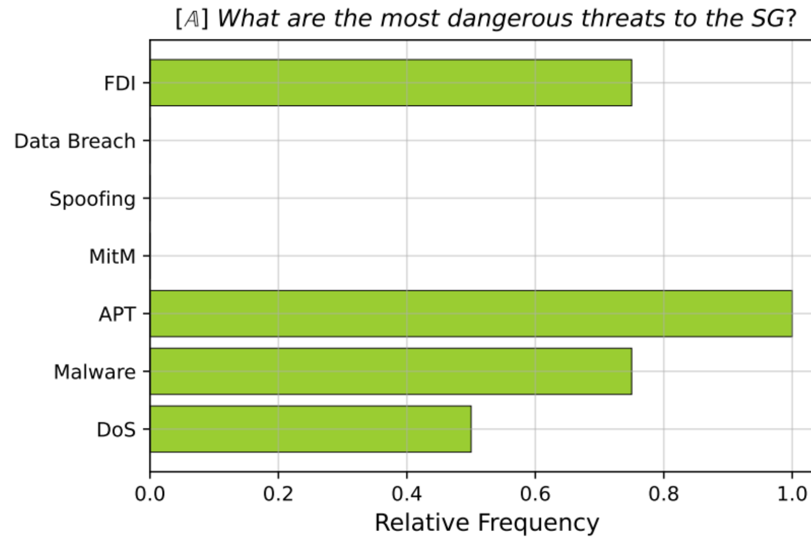
Employees	
Option	*Freq.*
They are aware fully of the risks and education is evaluated regularly	50.00%
They are not fully aware of the risks, but are educated on the topic	42.86%
They are not aware of the risks, and unlikely to improve in the short-term	0.00%
No answer	7.14%



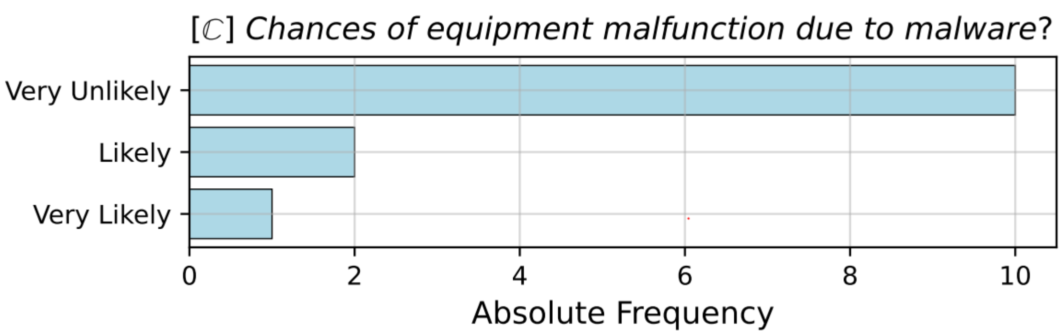
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Findings – threats (ℂ = Private Companies, 𝔸= Public Authorities)





- 100% of ℂ consider their systems to be at risk from APT.
- Only 14% of ℂ consider illegitimate access to consumer data to be ‘not threatening’,
- 0% of ℂ consider DoS to be problematic

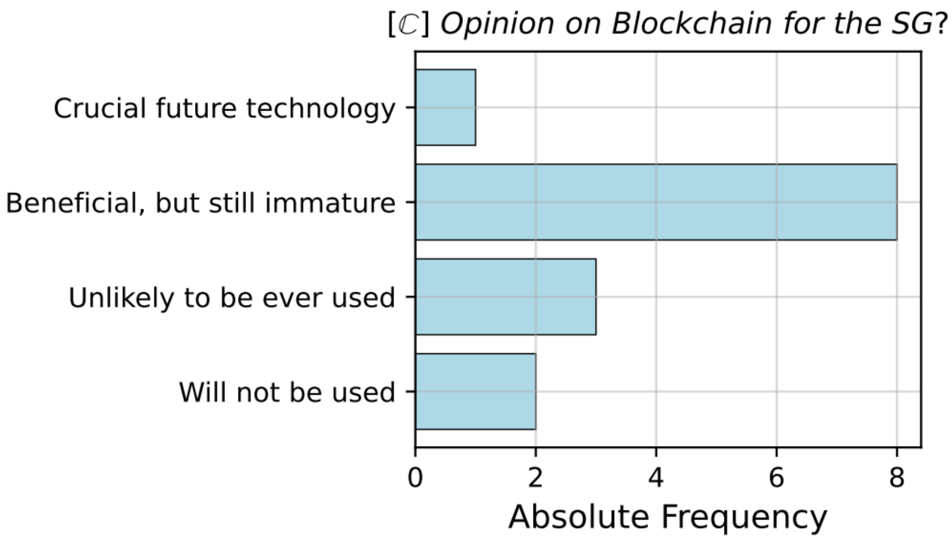


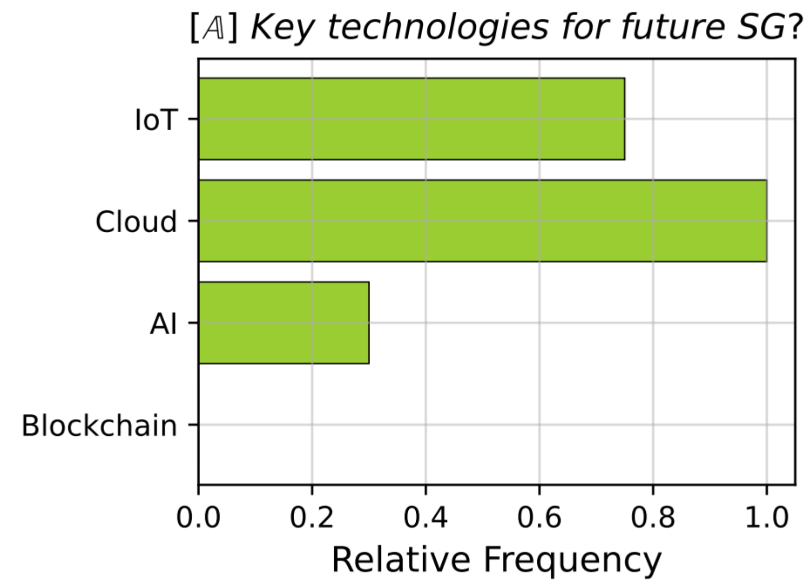
“How likely it is that malware can lead to human death? (killware)”

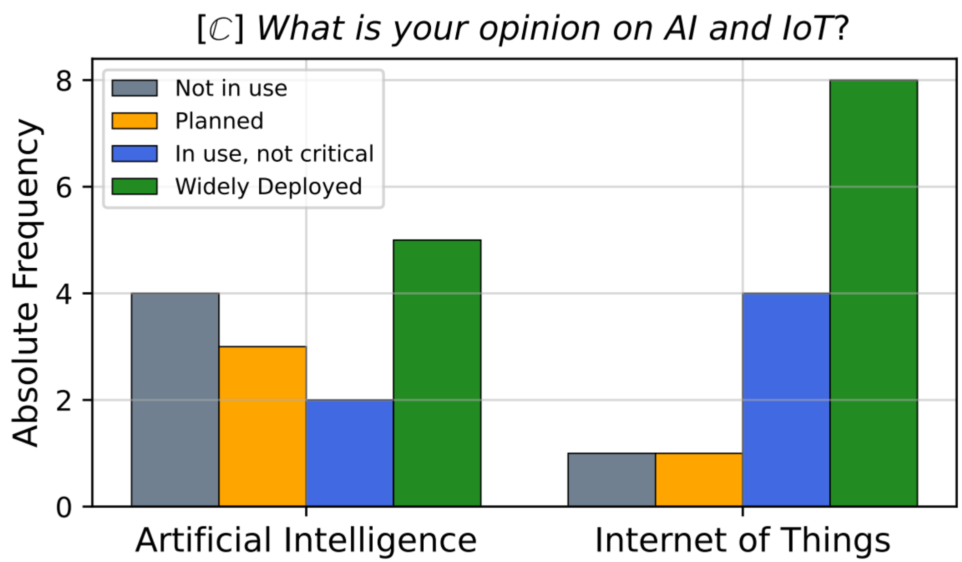
- ℂ : 14% unrealistic; 71% unlikely.
- 𝔸 : 50% very likely; 50% likely

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Findings – Tech (ℂ = Private Companies, 𝔸= Public Authorities)







UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Mismatch

- Practitioners ($\mathbb{C}$ and $\mathbb{A}$) vs Research:
  - MitM and Spoofing
  - Blockchain
  - Artificial Intelligence
  - Reaction Phase
  - Killware

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Mismatch (cont'd)

- Practitioners ($\mathbb{C}$ and $\mathbb{A}$) vs Research:
  - MitM and Spoofing
  - Blockchain
  - Artificial Intelligence
  - Reaction Phase
  - Killware

- Private ($\mathbb{C}$) vs Public ($\mathbb{A}$) entities:
  - Prevention Phase
  - Capabilities
  - Data Confidentiality and Replication
  - FDI

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What about sovereign and legislative bodies?

- After elaborating some comments received by ℂ, we derived an original model that explains the role of regulations in the context of the SG

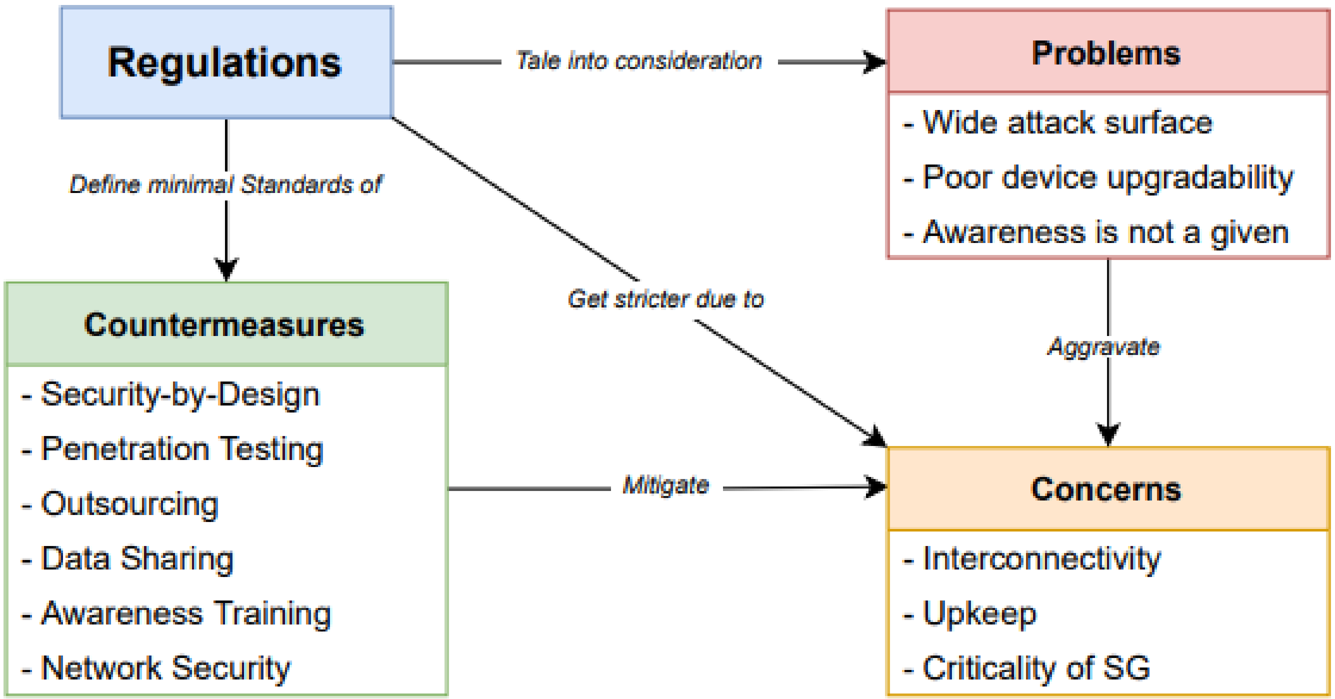


Fig. 13: Our original model displaying the relationships between *regulations* the cybersecurity of the SG.

Doing Practical Research on Machine Learning & Cybersecurity

Giovanni Apruzzese, PhD

University of Padua – November 23rd, 2022

UNIVERSITÄT LIECHTENSTEIN