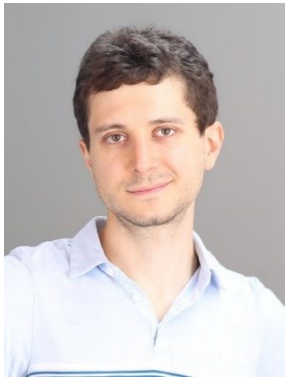
Cybersecurity and Machine Learning: Facts and Myths

Giovanni Apruzzese, PhD

University of Bologna – October 12th, 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
  - 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!

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

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)
- Using unlabelled data for Machine Learning in Cyberthreat Detection
  - Ref: Giovanni Apruzzese, Pavel Laskov, Aliya Tastemirova. "SoK: The Impact of Unlabelled Data for Cyberthreat Detection." IEEE European Symposium on Security and Privacy (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 (2022).
- Machine Learning Security in the Real-World
  - Ref: Giovanni Apruzzese, David Freeman, Savino Dambra, Hyrum S Anderson, Kevin Alexander Roundy, Fabio Pierazzi "'Real Attackers Don't Compute Gradients': Bridging the Gap Between Adversarial ML Research and Practice." TBD
- Using Machine Learning to violate the Privacy of Video Gamers
  - Ref: Pier Paolo Tricomi, Giovanni Apruzzese, Lisa Facciolo, Mauro Conti. "Attribute Inference Attacks in Online Multiplayer Video Games: a Case Study on Dota2." TBD
- 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 (2022)

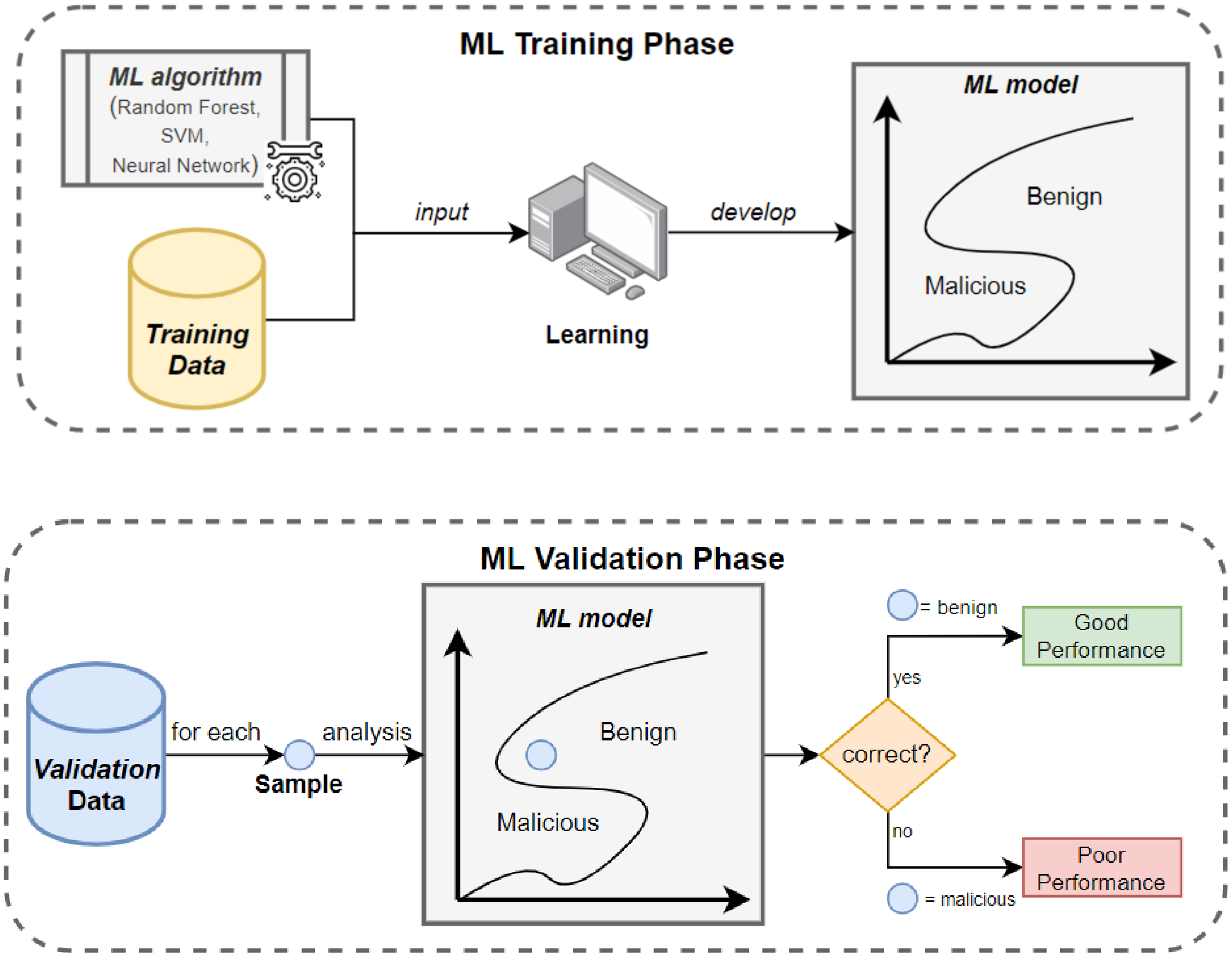
UNIVERSITÄT
LIECHTENSTEIN

Fundamentals of Machine Learning and Cybersecurity

UNIVERSITÄT
LIECHTENSTEIN

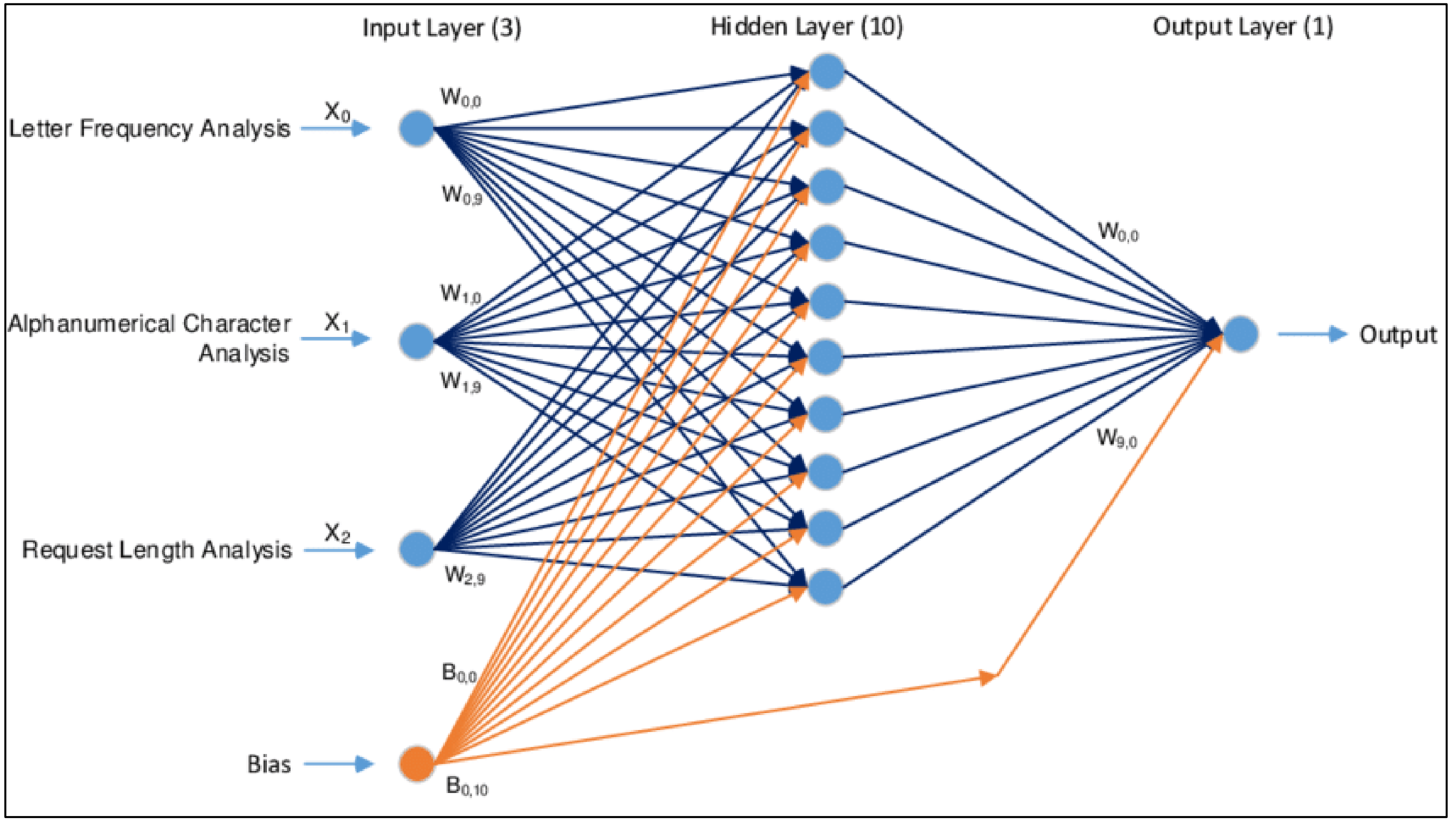
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

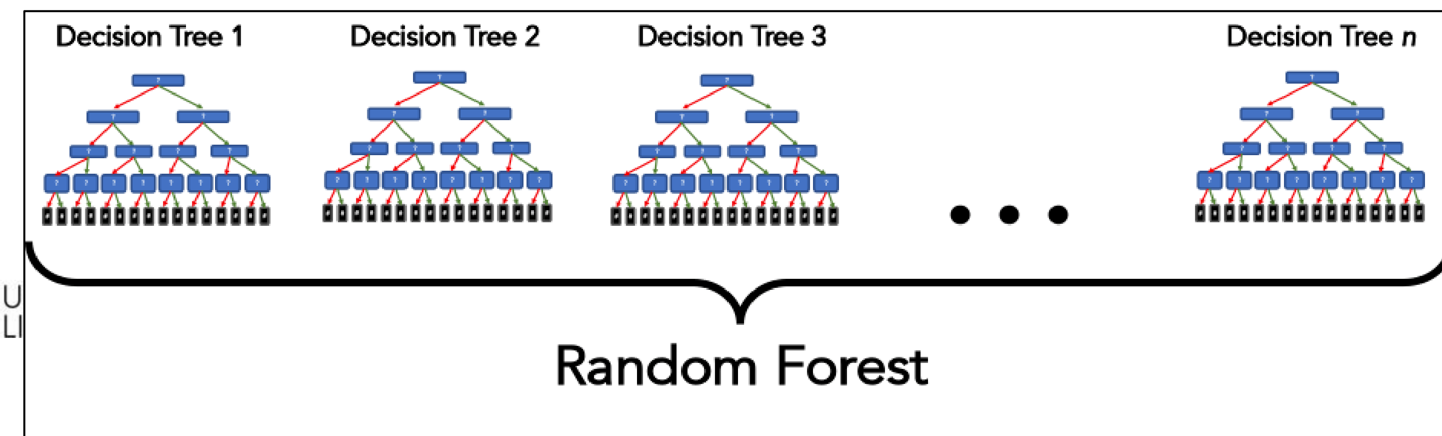
Machine Learning workflow: Training and Testing



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

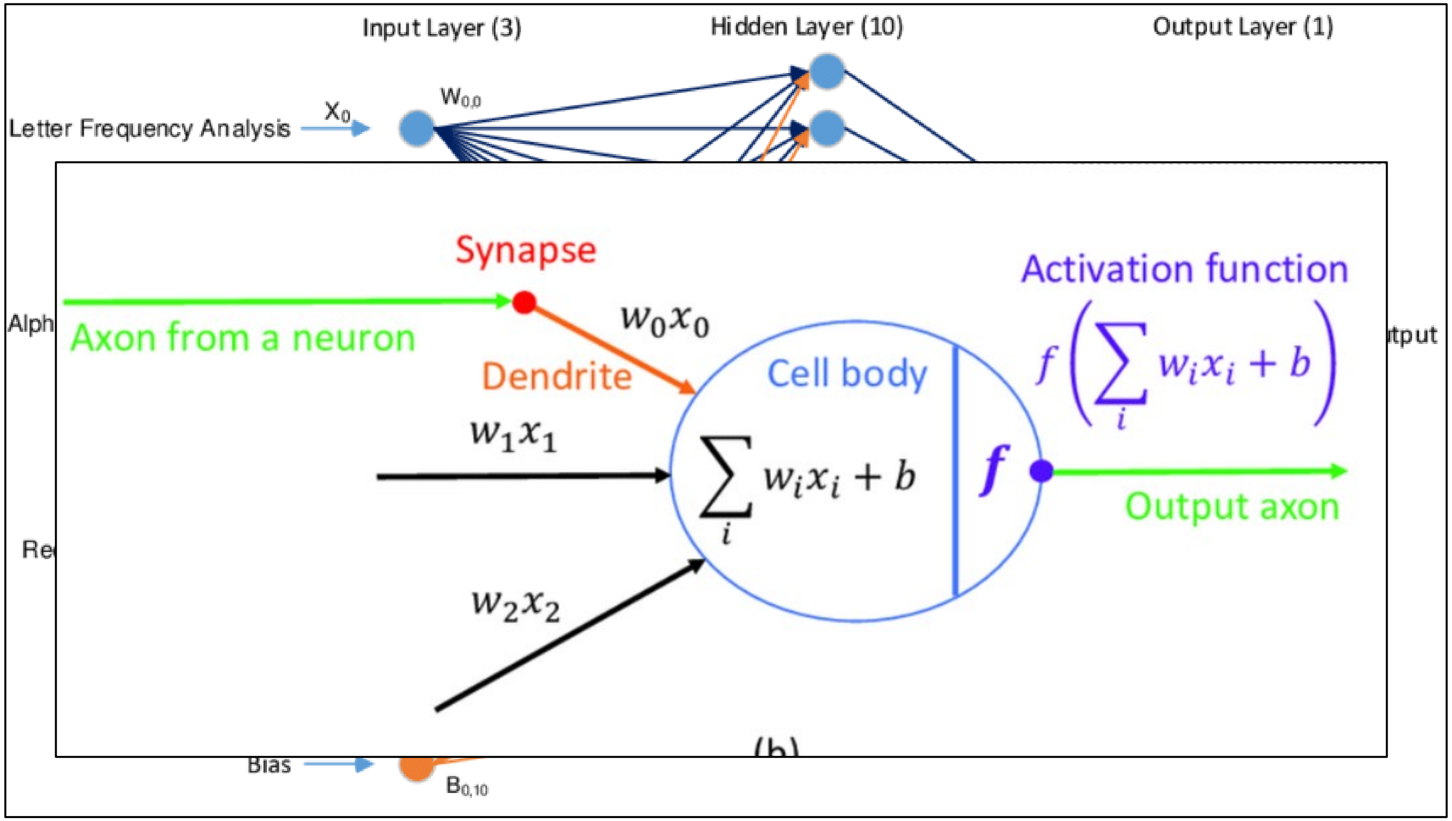
Do you think that training ML models is difficult?

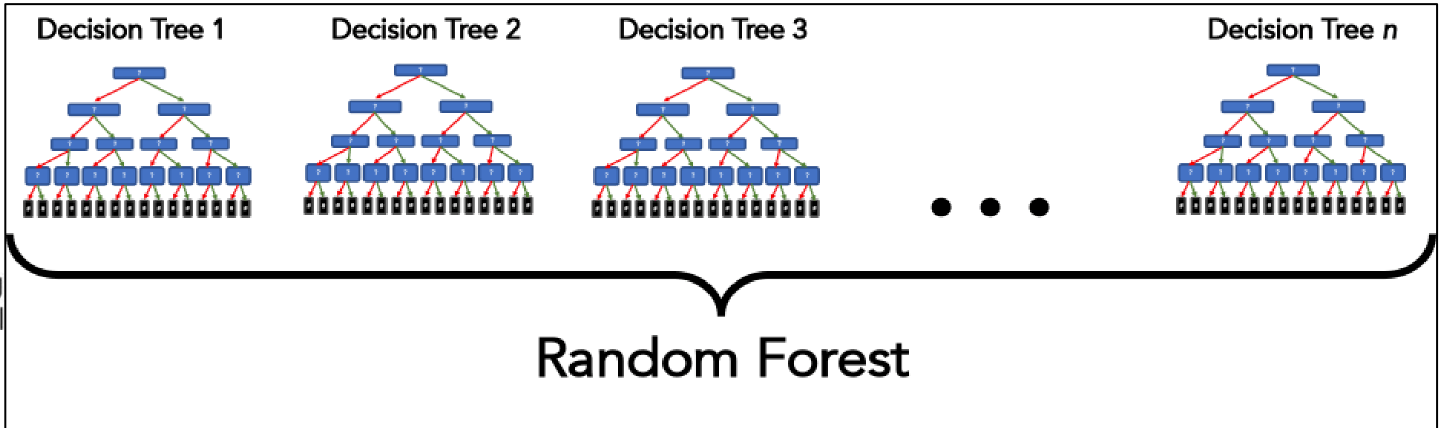




Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

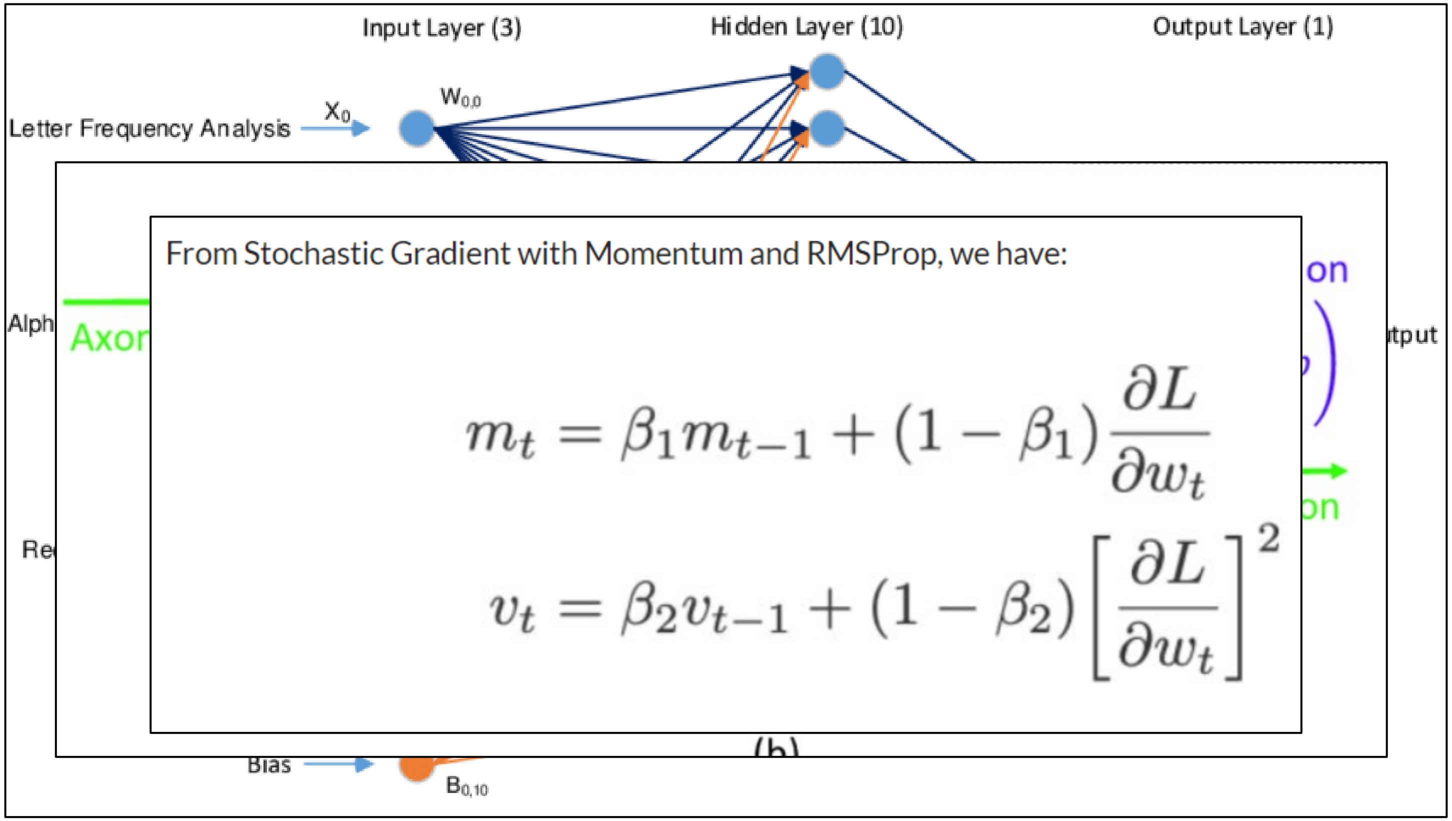
Do you think that training ML models is difficult? – Maths





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

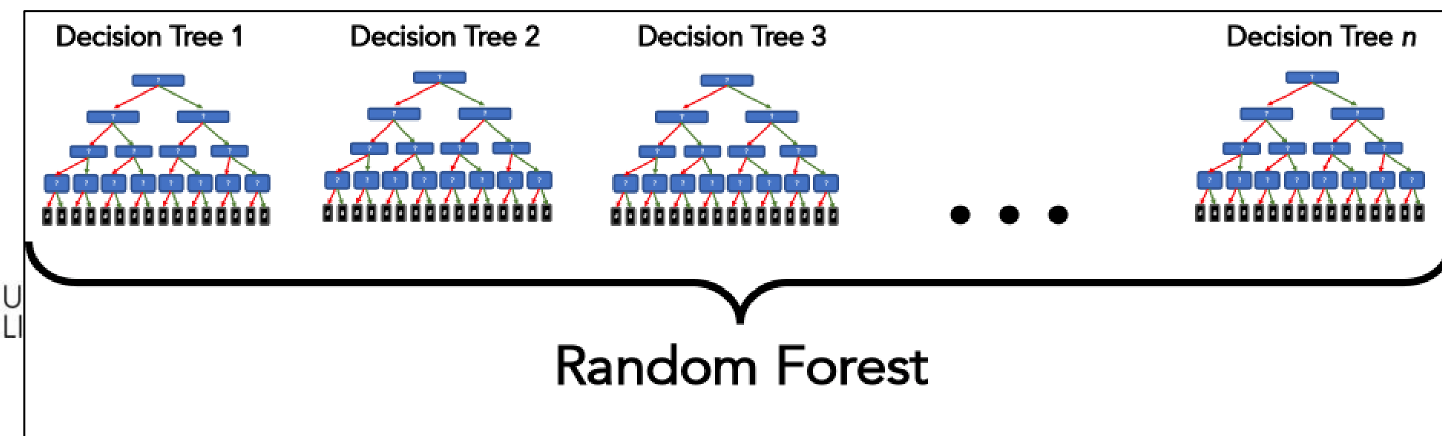
Do you think that training ML models is difficult? – More Maths



From Stochastic Gradient with Momentum and RMSProp, we have:

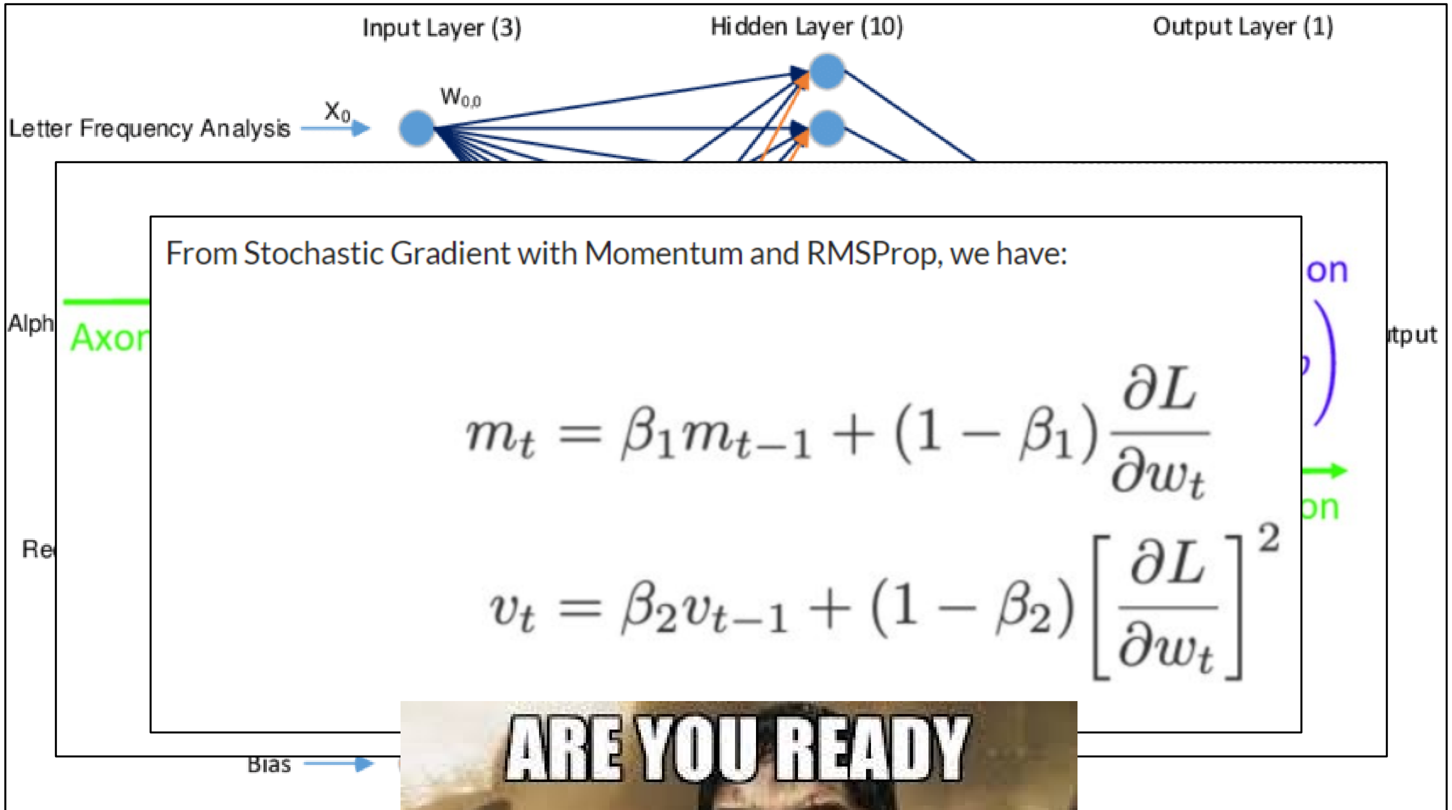
$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\frac{\partial L}{\partial w_t}$$

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



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

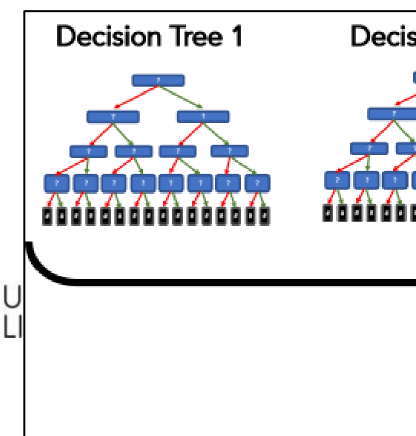
Do you think that training ML models is difficult? – More Maths ☺



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

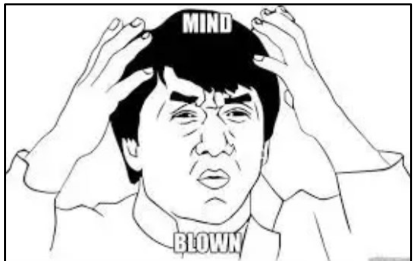
Do you think that training ML models is difficult? – The real problem

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:
https://github.com/scikit-learn/scikit-learn/blob/baf828ca1/sklearn/ensemble/_forest.py#L297



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 cost, 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)

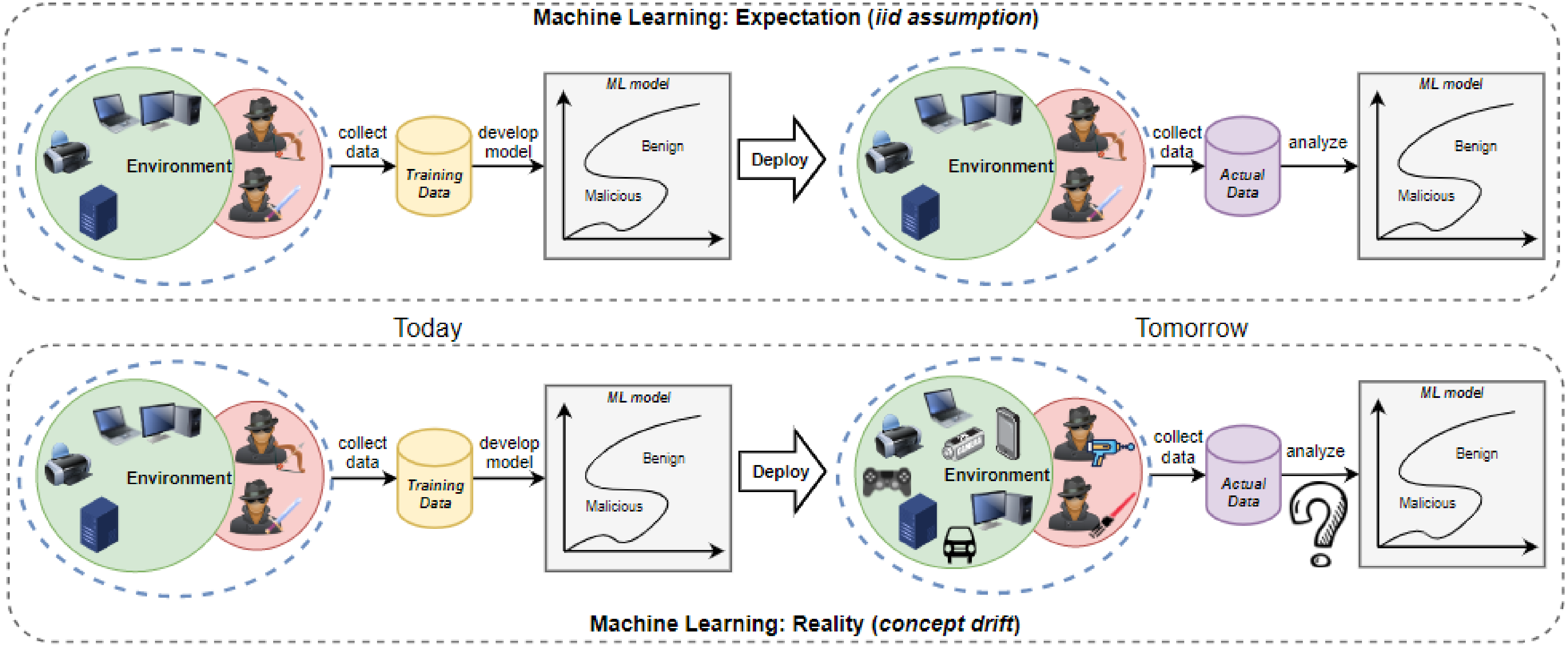


Fig. 9. Machine Leaning in the presence of Concept Drift. The ML model expects that the data will not deviate from the one seen during its training. In cybersecurity, however, the environment evolves, and adversaries also become more powerful.

UNIVERSITÄT LIECHTENSTEIN

Unlabelled data for Machine Learning in Cyberthreat Detection

UNIVERSITÄT
LIECHTENSTEIN

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

Once upon a time…

- At the beginning of 2021, I was having a meeting with Prof. Pavel Laskov, brainstorming about new research directions on Machine Learning (ML)
- Pavel: "We should look at Semisupervised Learning, it's very trendy now!"

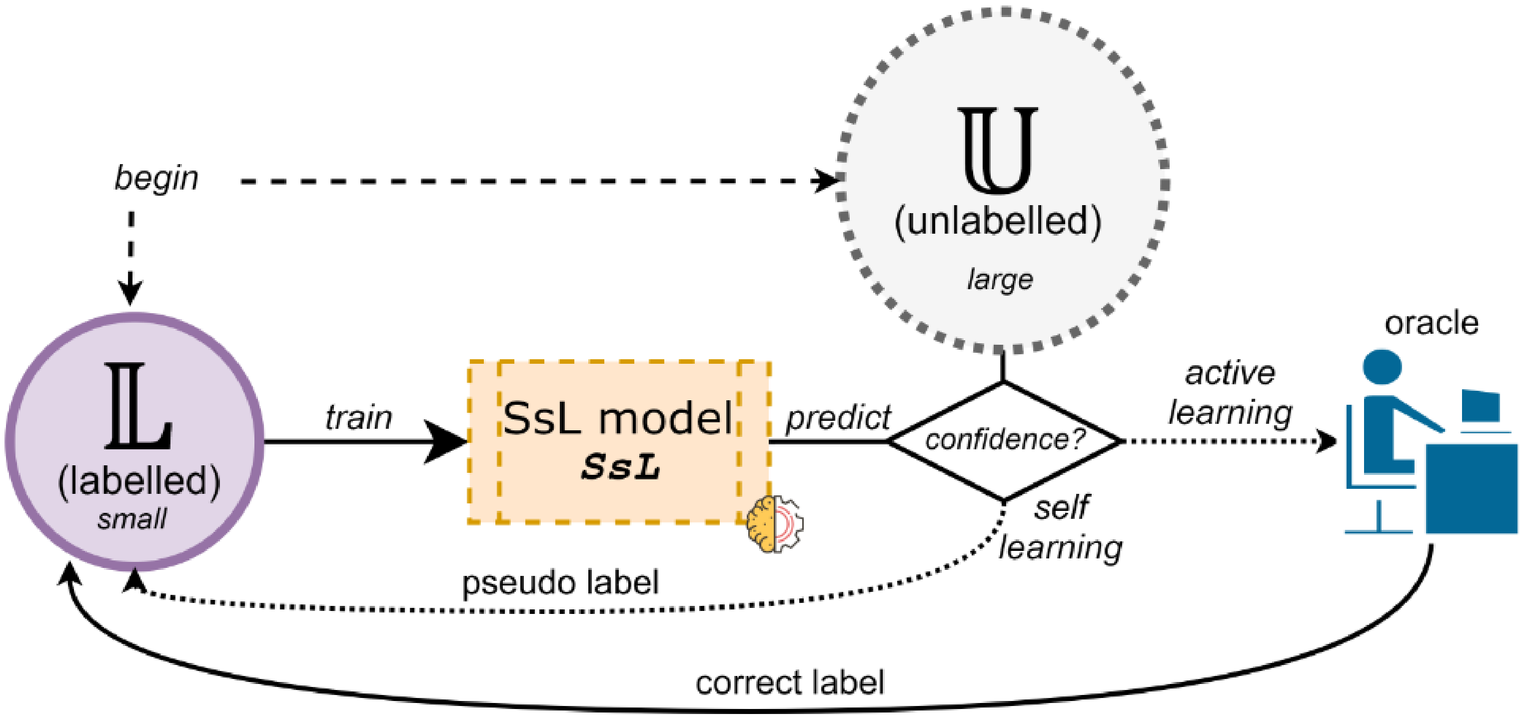
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Semisupervised Learning

- Labelled data is expensive, but *unlabelled* data is cheap(er)

→ Why not using unlabelled data to improve the proficiency of ML models?

Mixing *labelled* with *unlabelled* data is a ML approach denoted as **"Semisupervised Learning" (SsL)**



The assumptions of SsL appears to be enticing for Cyberthreat Detection (CTD)

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

Once upon a time… (cont’d)

- At the beginning of 2021, I was having a meeting with Prof. Laskov, brainstorming about new research directions on Machine Learning (ML)
- Pavel: “We should look at Semisupervised Learning, it’s very trendy now!”
- It was the first time I directly tackled SsL, so I did what most researchers do when they start focusing on a new topic:
    - I looked into **existing literature** on SsL applications for CTD…
    - …and started to **replicate (basic) SsL methods** on public CTD datasets

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

All that glitters is not gold…

- My initial results portrayed SsL to be **bad**.
  - Like, really bad ☺
- As a sanity check, I asked a MSc. student (Aliya Tastemirova) to:
  - **independently** replicate the SsL methods I developed
  - and evaluate their performance on **different CTD datasets**
- Her results confirmed my initial findings.
- We (Pavel, Aliya, and I) had a joint meeting, and we decided to dig deeper:
  - either all of **us were wrong**…
  - …or **something odd was going on** between the lines.

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

Bad performance?

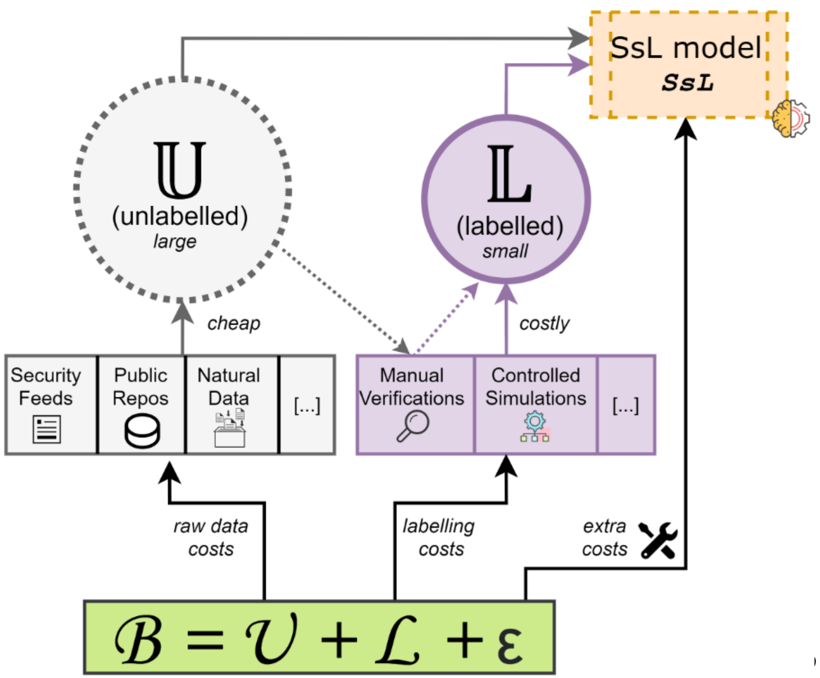
- In some cases (e.g., Phishing Detection), SsL methods achieved 0.90 F1-score by using ~100 labels and thousands of unlabelled samples.

- One could claim such performance to be good…

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

Bad performance? (cont'd)

- In some cases (e.g., Phishing Detection), SsL methods achieved 0.90 F1-score by using ~100 labels and thousands of unlabelled samples.
- One could claim such performance to be good…
- …unless a (traditional) supervised learning classifier using *only* 100 labels (without any unlabelled data) achieved an F1-score of **0.91**
- Our initial experiments showed that using unlabelled data provided "uncertain" improvement (if any).
  - In reality, unlabelled data may be cheaper to acquire than labels, but it is not **free**!



UNIVERSITÄT LIECHTENSTEIN

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

If SsL is bad, then why is it so trendy in research?

- We investigated all (ttbook) existing literature on SsL for CTD, asking ourselves:

"What are the benefits of unlabelled data in SsL?"

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

If SsL is bad, then why is it so trendy in research?

- We investigated <u>all</u> (ttbook) existing literature on SsL for CTD, asking ourselves:

"What are the benefits of unlabelled data in SsL?"

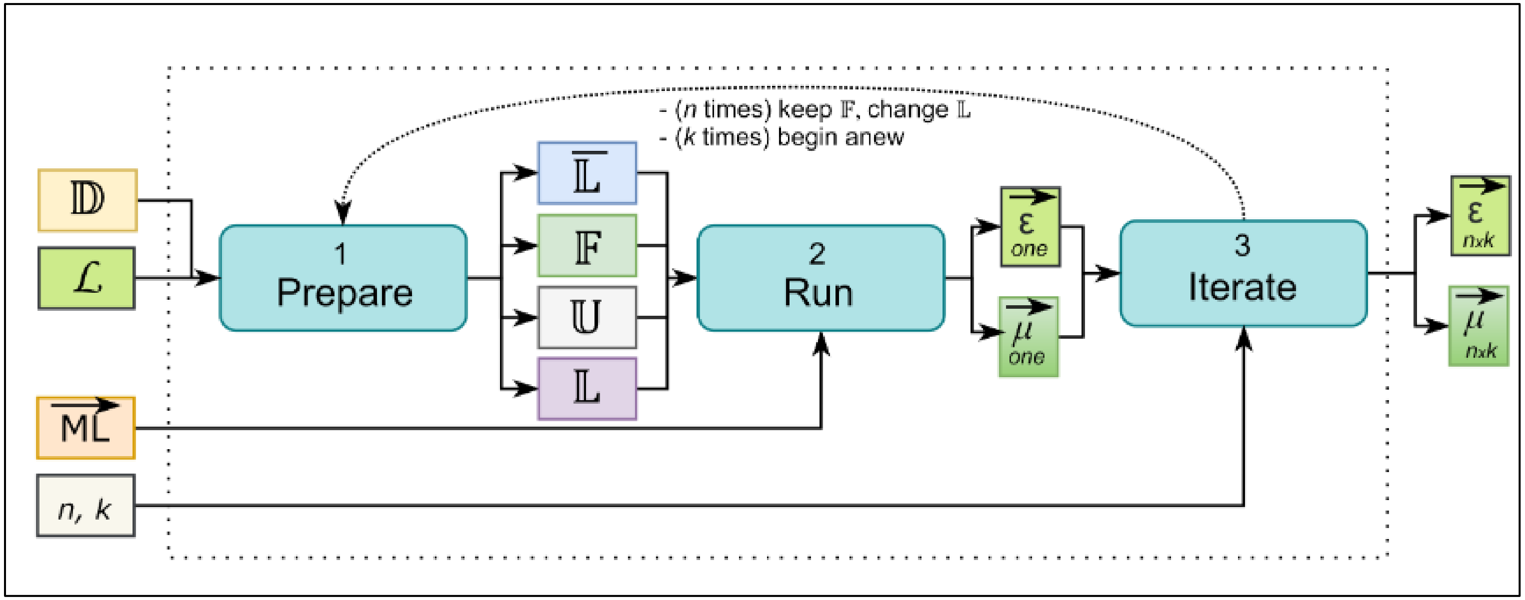
Task	Paper (1st Author)	Year	Lower Bound	Ablation Study	Upper Bound	Stat. Sign.	Transparency *Labels*	Transparency *Balance*	Repr.	Dataset
Network Intrusion Detection	Li [93]	2007	✓	✓	✗	✗	✓	✓	◐	NSL-KDD
	Long [94]	2008	✓	✓	✗	◐	✓	✗	◐	NSL-KDD
	Görnitz [95]	2009	✓	✓	✗	◐	✓	✓	✗	Private
	Seliya [96]	2010	✓	✓	✗	✗	✓	✓	◐	NSL-KDD
	Symons [97]	2012	✗	✓	✓	◐	✓	✗	✗	Kyoto2006
	Wagh [98]	2014	✗	✗	✗	✗	✓	✓	◐	NSL-KDD
	Noorbehbahani [35]	2015	✗	◐	✓	✗	✓	✓	◐	NSL-KDD, Custom
	Ashfaq [99]	2017	✗	◐	✗	✗	✓	✗	◐	NSL-KDD
	Qiu [67]	2017	✗	◐	✓	✗	✓	✓	✗	Custom
	McElwee [100]	2017	✗	◐	✓	✗	✓	✗	◐	NSL-KDD
	Kumari [68]	2017	✓	◐	✗	✗	✓	✗	◐	NSL-KDD
	Yang [101]	2018	◐	✓	✓	✗	✓	✗	✗	NSL-KDD, AWID
	Gao [102]	2018	✓	◐	✗	✗	✓	✗	✗	NSL-KDD
	Shi [103]	2018	◐	◐	✗	✗	✓	✗	✗	NSL-KDD
	Yao [36]	2019	◐	◐	✓	✗	✓	✓	◐	NSL-KDD
	Yuan [104]	2019	✗	◐	✗	◐	✓	✓	◐	NSL-KDD
	Zhang [65]	2020	◐	✗	✓	◐	✓	✗	◐	NSL-KDD
	Hara [105]	2020	✗	◐	✓	✗	✗	✗	✗	NSL-KDD
	Ravi [106]	2020	✓	✗	✗	✗	✓	✗	✗	NSL-KDD
	Gao [107]	2020	✗	✓	✓	✓	✓	✓	✗	NSL-KDD
	Li [108]	2020	✗	◐	✓	✓	✓	✗	◐	NSL-KDD, Private
	Zhang [70]	2021	◐	◐	✗	◐	✗	✓	◐	CICIDS2017, CTU13
	Liang [109]	2021	✓	◐	✓	◐	✓	✓	◐	NSL-KDD
Phishing Detection	Gyawali [110]	2011	✗	✓	✓	✗	✓	✓	◐	Private
	Zhao [111]	2013	✓	✓	✓	✓	✗	✓	✓*	DetMalURL
	Gabriel [15]	2017	◐	◐	✗	✗	✗	✗	◐	Private
	Yang [112]	2017	✓	◐	✗	✗	✓	✓	◐	Private
	Bhattacharjee [113]	2017	✗	✓	✗	◐	✗	✗	◐	Private
	Li [55]	2017	✓	✓	✓	◐	✓	✓	✗	Custom
Malware Detection	Moskovitch [114]	2008	✗	✓	✗	◐	✓	✓	✗	Custom
	Santos [115]	2011	✗	✗	✓	✗	✓	✓	◐	Custom
	Nissim [116]	2012	✗	◐	✓	◐	✗	✗	✗	Private
	Zhao [117]	2012	✗	✗	✗	✗	✓	✓	◐	Private
	Nissim [118]	2014	✓	✓	✗	◐	✓	✓	✗	Custom
	Zhang[119]	2015	◐	◐	✗	✗	✓	✓	✗	Private
	Nissim [120]	2016	✗	✓	✓	◐	✓	✓	◐	Custom
	Ni [121]	2016	✓	✓	✗	◐	✓	✓	◐	Private
	Chen [122]	2017	✓	✓	✗	◐	✗	✗	◐	Private
	Rashidi [66]	2017	✗	✓	✓	◐	✓	✓	✗	Drebin
	Fu [123]	2019	✓	✓	✗	✗	✓	✗	◐	Private
	Irofti [124]	2019	◐	◐	✗	◐	✗	✗	✓	DREBIN, EMBER
	Pendlebury [86]	2019	✗	✗	✓	◐	✓	✓	✓	AndroZoo
	Sharmeen [125]	2020	✓	◐	✗	◐	✓	✓	◐	Drebin, AndroZoo
	Chen [126]	2020	◐	◐	✓	✗	✓	✓	◐	MCC
	Koza [11]	2020	✓	◐	✓	◐	✓	✗	✓	Private
	Noorbehbahani [13]	2020	✓	✗	✗	◐	✓	✓	✗	AndMal17
	Li [127]	2021	✗	◐	✗	◐	✓	✗	◐	FalDroid, DREBIN, Genome
	Liang [109]	2021	✓	◐	✓	◐	✓	✓	◐	Custom

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

Revealing the impact of unlabelled data in CTD

The state-of-the-art does not allow to determine whether using unlabelled data is *truly* beneficial in CTD

- As a constructive step, in our paper we:
  - Provide a set of requirements to estimate the benefits (if any) of using unlabelled data in CTD
  - Propose a framework, CEF-SsL, that allows to meet all such requirements in research
  - We experimentally evaluate CEF-SsL on 9 CTD datasets by considering 9 SsL methods.



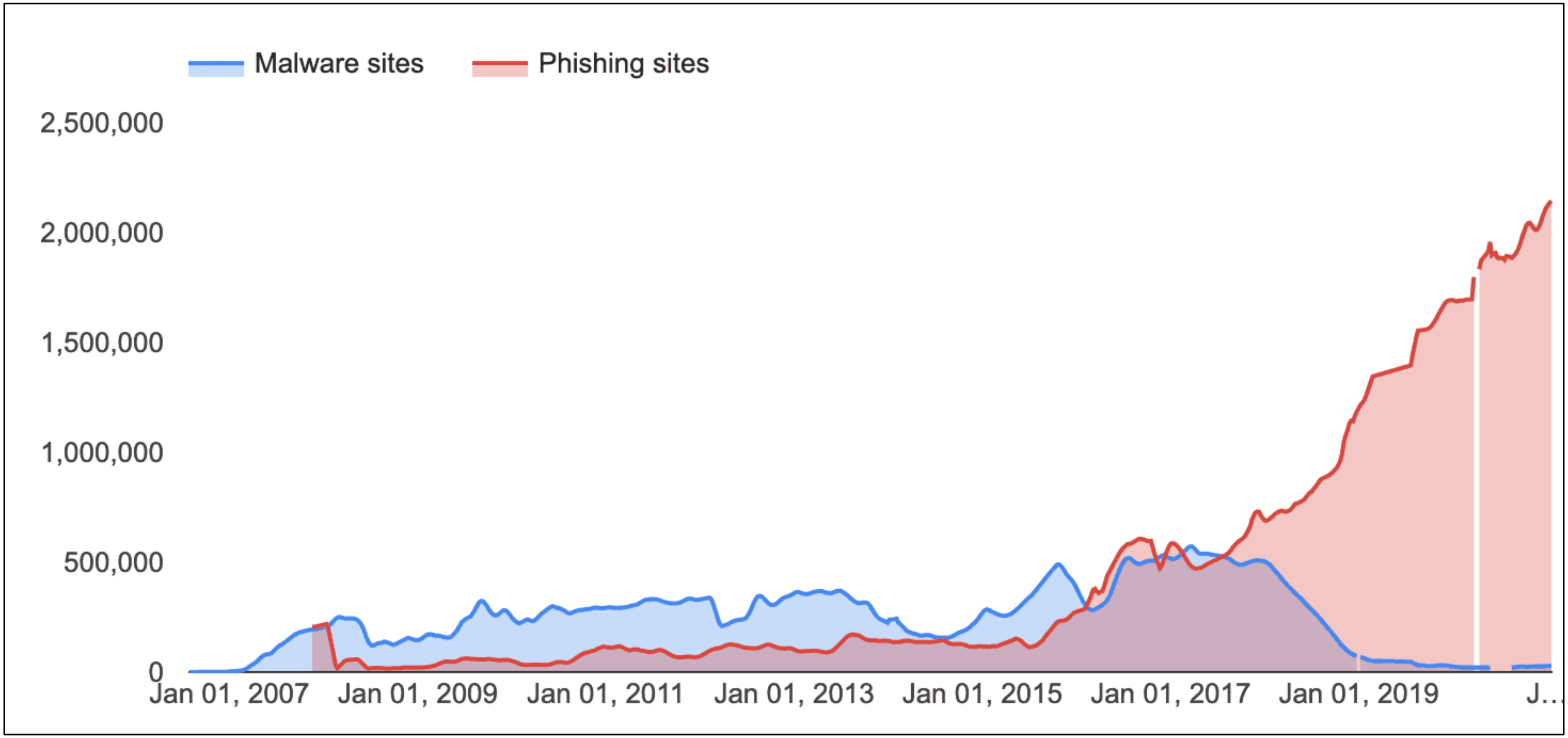
The security of Machine Learning-based Phishing Website Detectors

UNIVERSITÄT
LIECHTENSTEIN

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

Current Landscape of Phishing

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



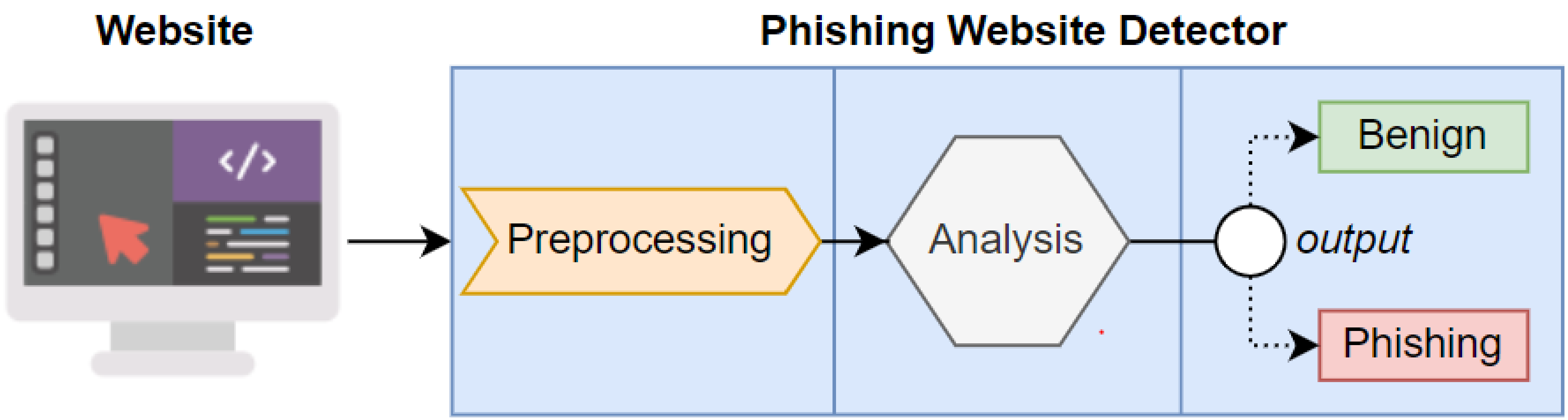
UNIVERSITÄT LIECHTENSTEIN

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

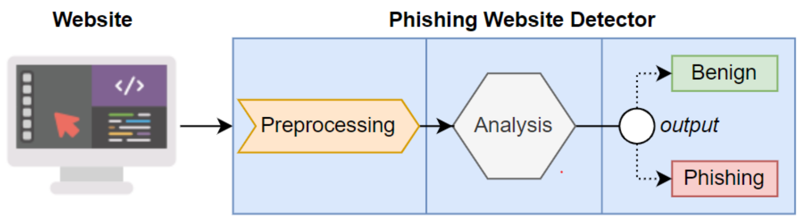


UNIVERSITÄT LIECHTENSTEIN

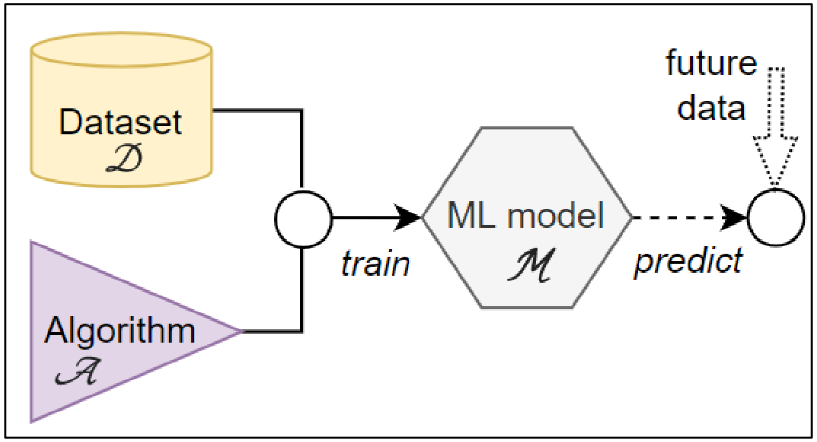
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

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

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)

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

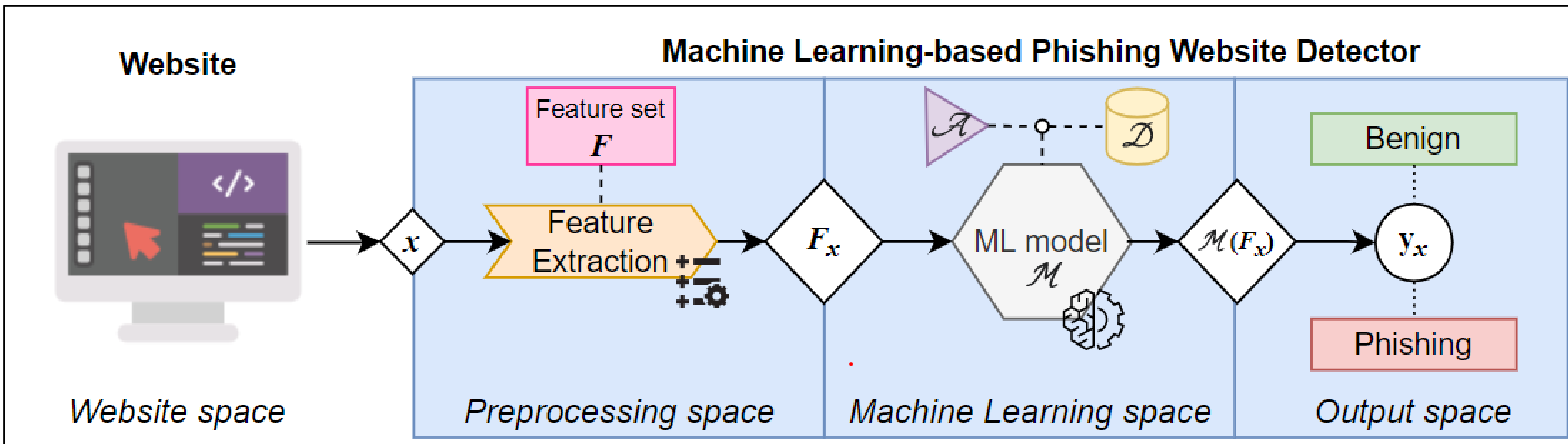
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

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)

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

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



UNIVERSITÄT LIECHTENSTEIN

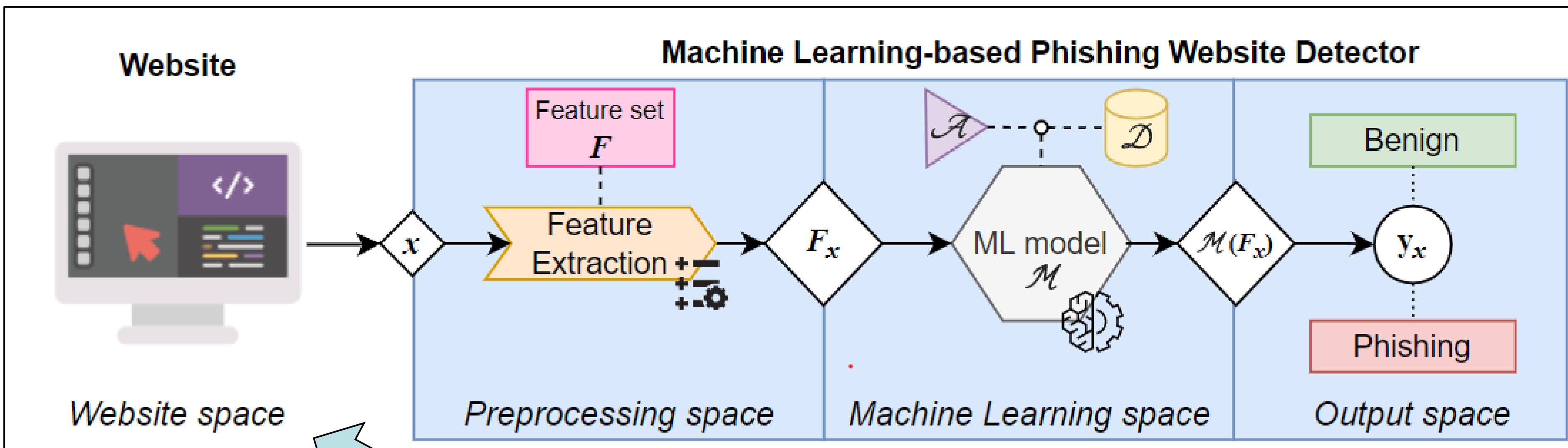
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

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)

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

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



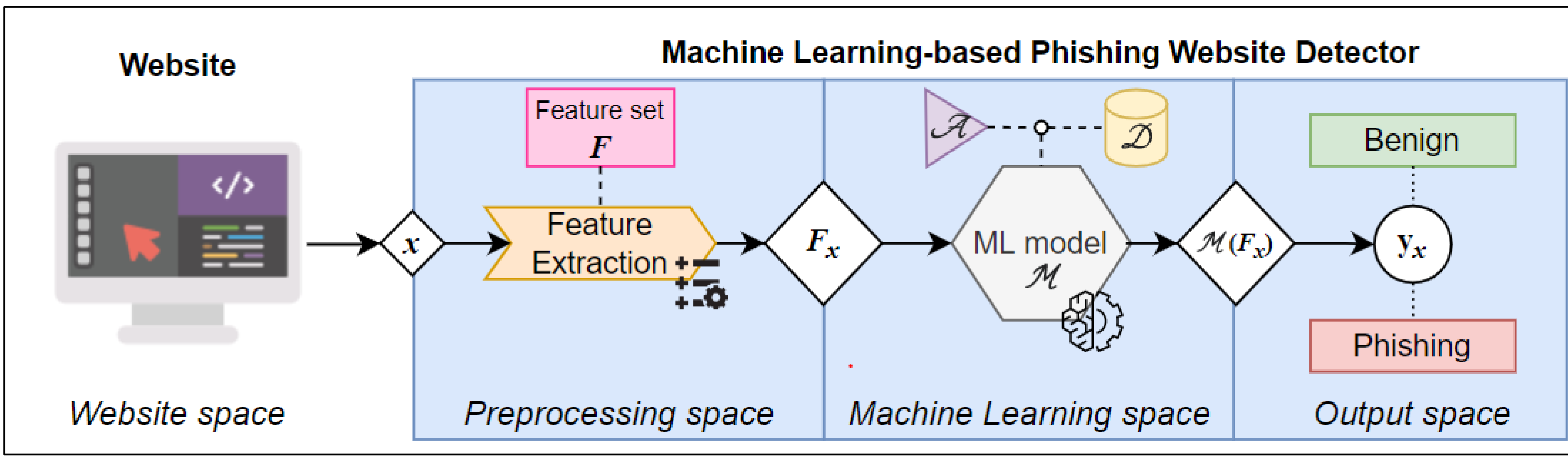
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

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)

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

- In the context of a ML-PWD, such ε can be introduced in three ‘spaces’:



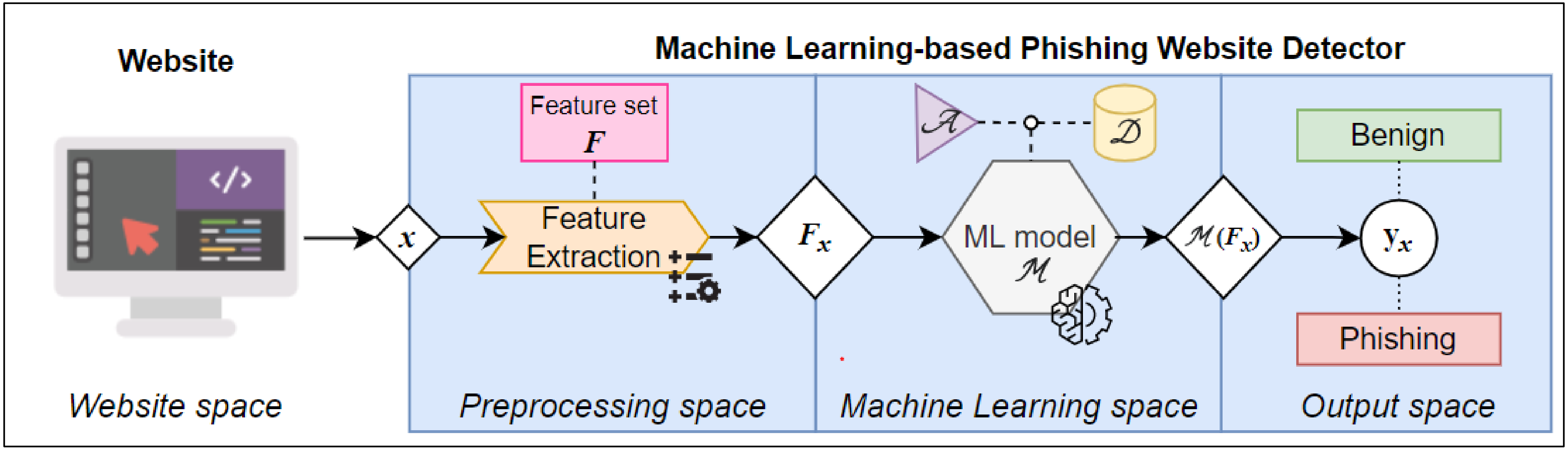
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

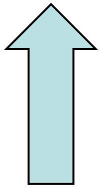
Problem Statement: Adversarial Attacks against ML-PWD

- 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)

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

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





UNIVERSITÄT LIECHTENSTEIN

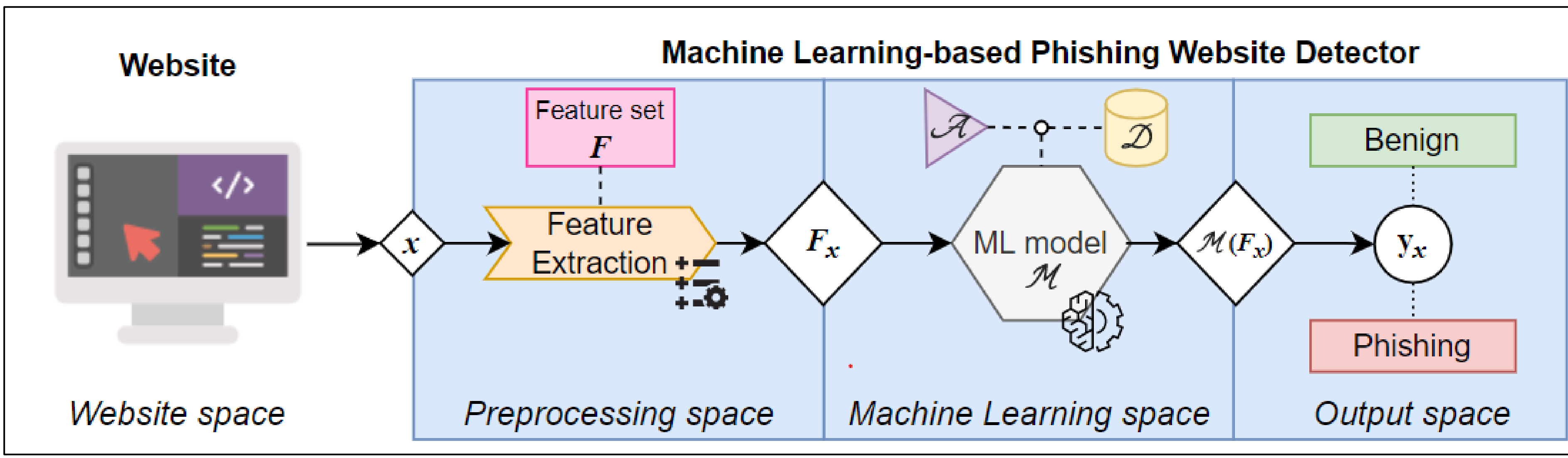
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

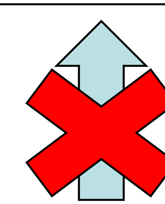
Problem Statement: Adversarial Attacks against ML-PWD

- 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)

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

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





UNIVERSITÄT LIECHTENSTEIN

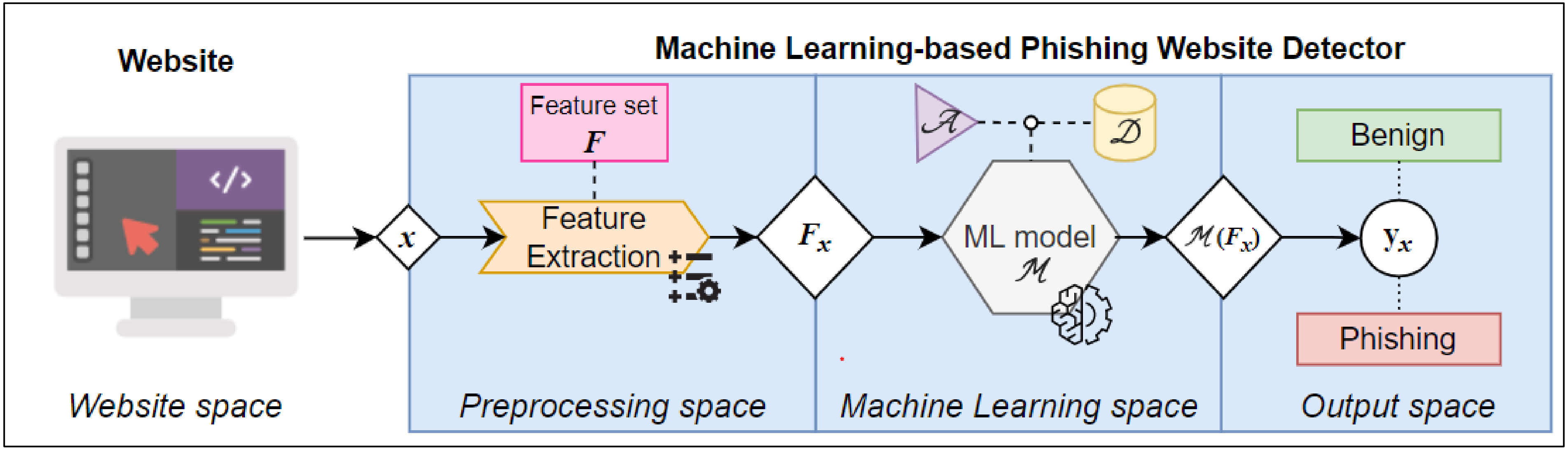
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

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)

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

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

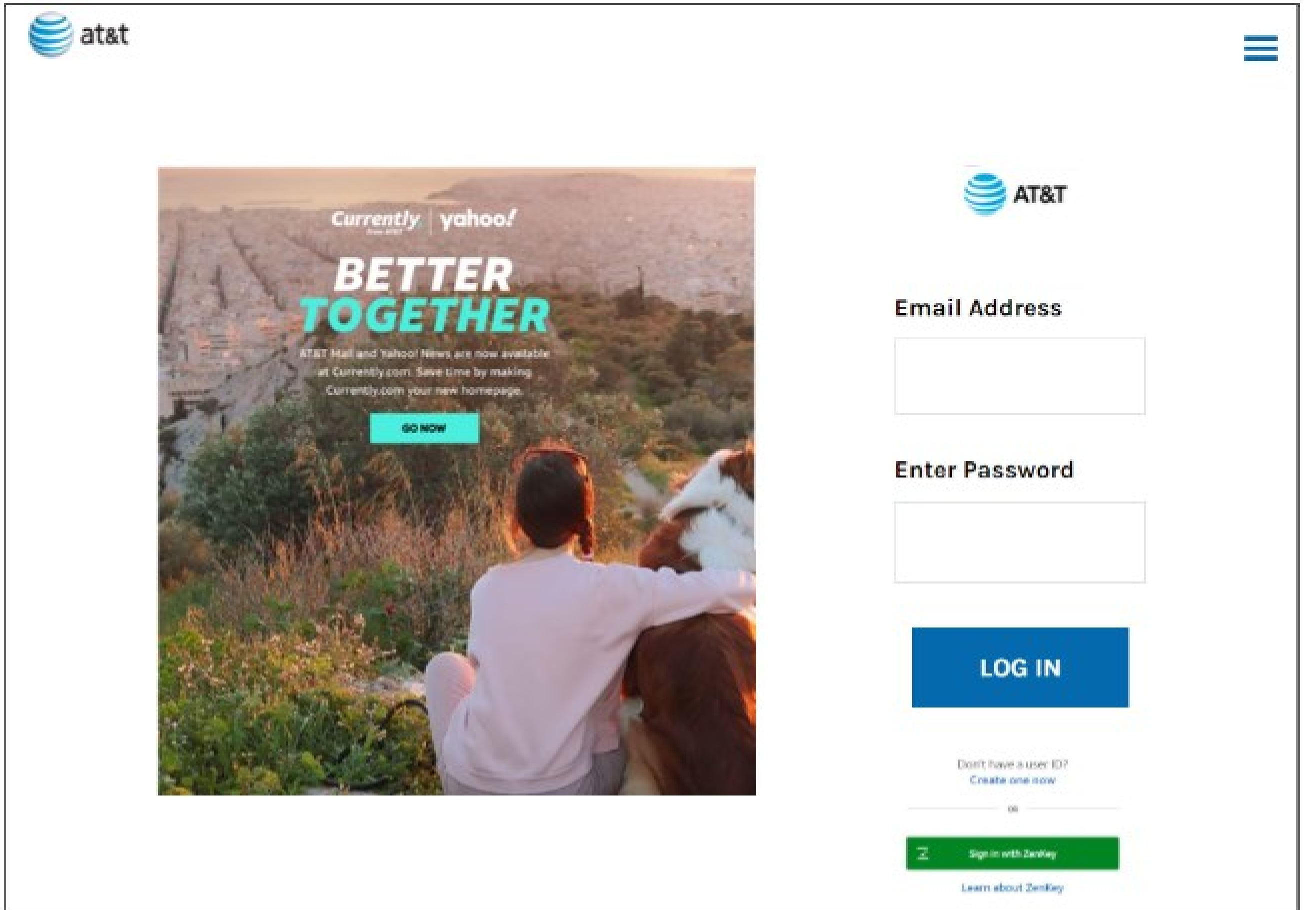


Question: Which 'space' do you think an *attacker* is **most likely** to use?

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/.



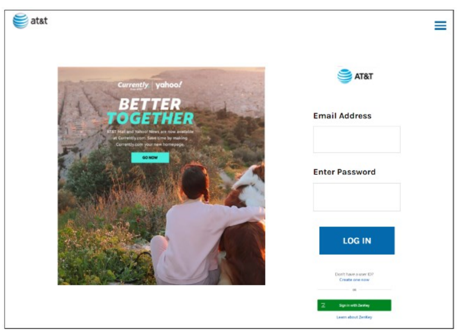
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

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

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

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

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




```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>

<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 wsite-input-width-370px" name="_u435728988405554593" style
                    ="height: 50px"></textarea>
                </div>
```

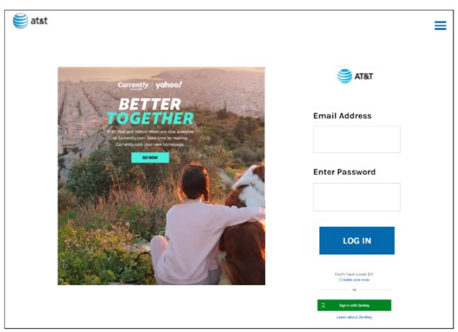
ε (WsP)

UNIVERSITÄT LIECHTENSTEIN

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



→

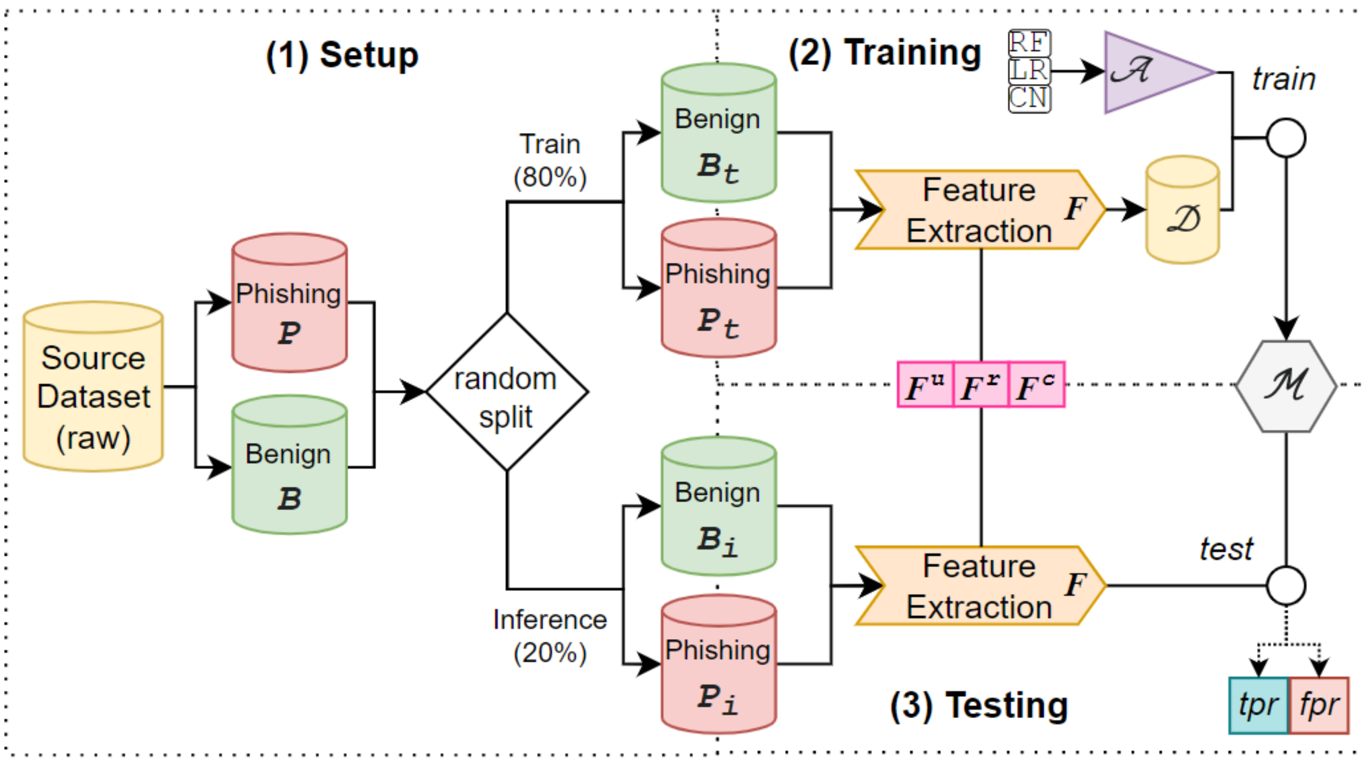
```html
<div>
    <form enctype="multipart/form-data" action=
    "//www.weebly.com/weebly/apps/formSubmit.php" method="POST" id=
    "form-723155629711391878">
        <div id="723155629711391878-form-parent" class="wsite-form-container"
         style="margin-top:10px;">
            <ul class="formlist" 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>

<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 wsite-input-width-370px" name="_u435728988405554593" style
                    ="height: 50px"></textarea>
                </div>
```

⇦ ε (WsP)

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

Evaluation – Workflow

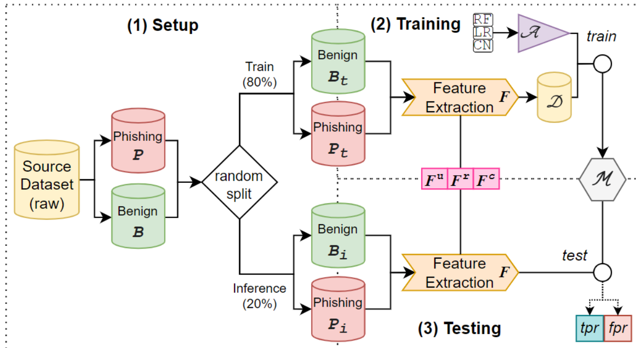
- Such attacks appear cheap, but are they effective? Let's assess their impact!
- We develop proficient ML-PWD (high *tpr*, low *fpr*)



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 ☺

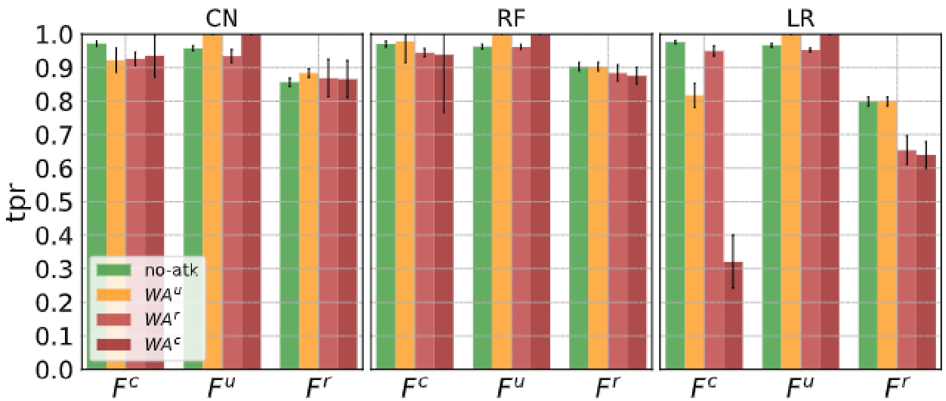
- Let’s attack such ML-PWD
  - The *tpr* will decrease!

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

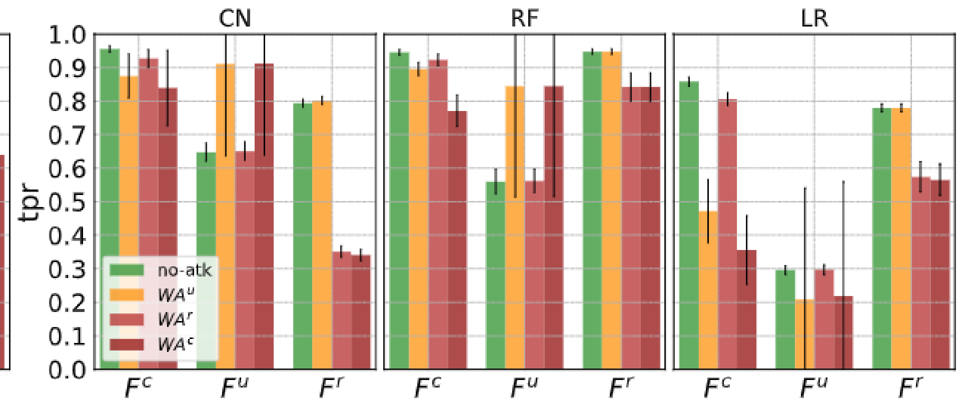
$\mathcal{A}$	F	Zenodo tpr	Zenodo fpr	δphish tpr	δphish fpr
CN	F^u	0.96±0.008	0.021±0.0077	0.55±0.030	0.037±0.0076
	F^r	0.88±0.018	0.155±0.0165	0.81±0.019	0.008±0.0020
	F^c	0.97±0.006	0.018±0.0088	0.93±0.013	0.005±0.0025
RF	F^u	0.98±0.004	0.007±0.0055	0.45±0.022	0.003±0.0014
	F^r	0.93±0.013	0.025±0.0118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

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 1% decrease in detection rate can be problematic when dealing with *millions of samples*!
- In other cases, YES
  - This is very significant, because WsP are cheap and are likely to be exploited by attackers!

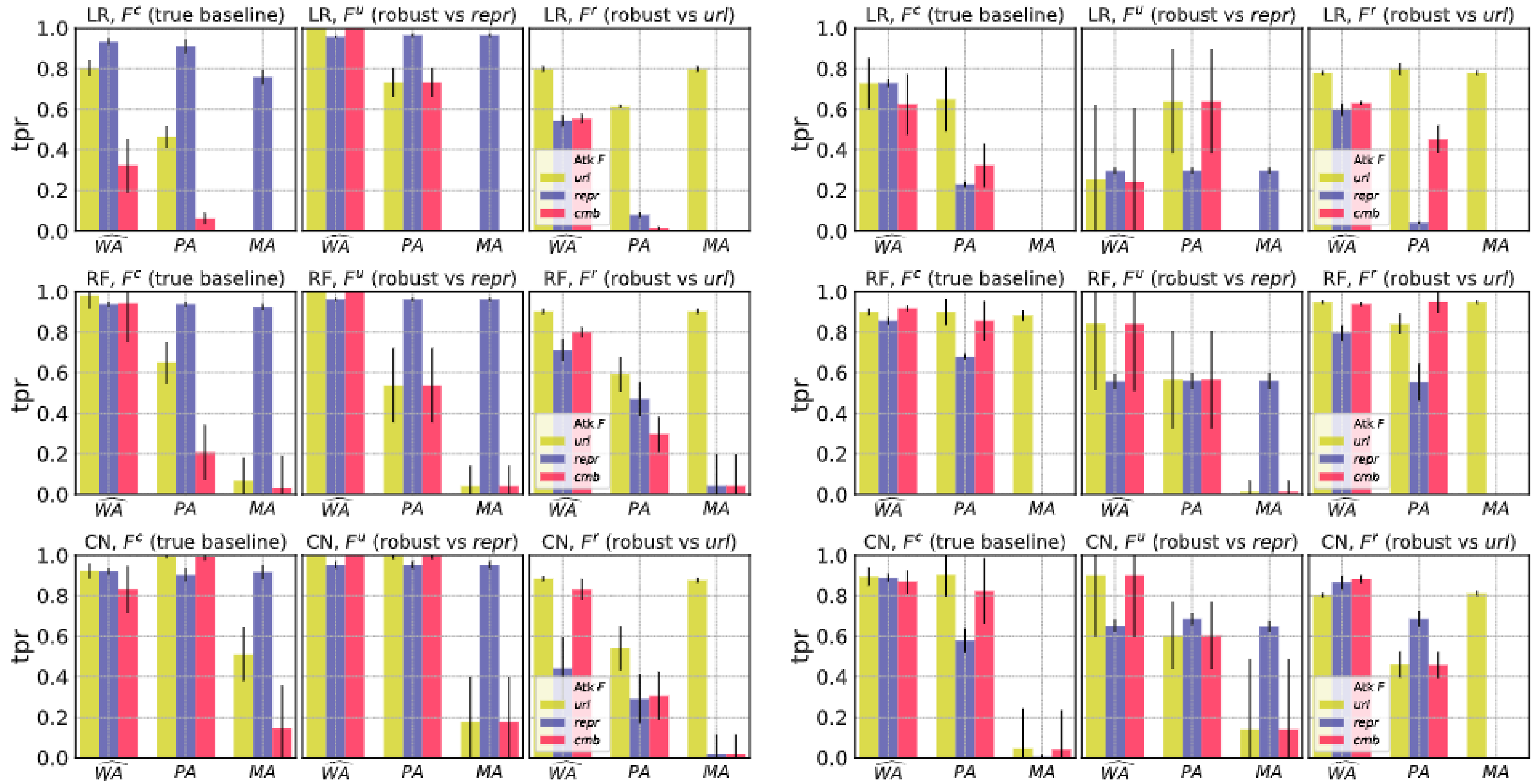
UNIVERSITÄT
LIECHTENSTEIN

<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., $\widehat{\mathrm{WA}}$, 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., $\widehat{\mathrm{WA}}$, 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?

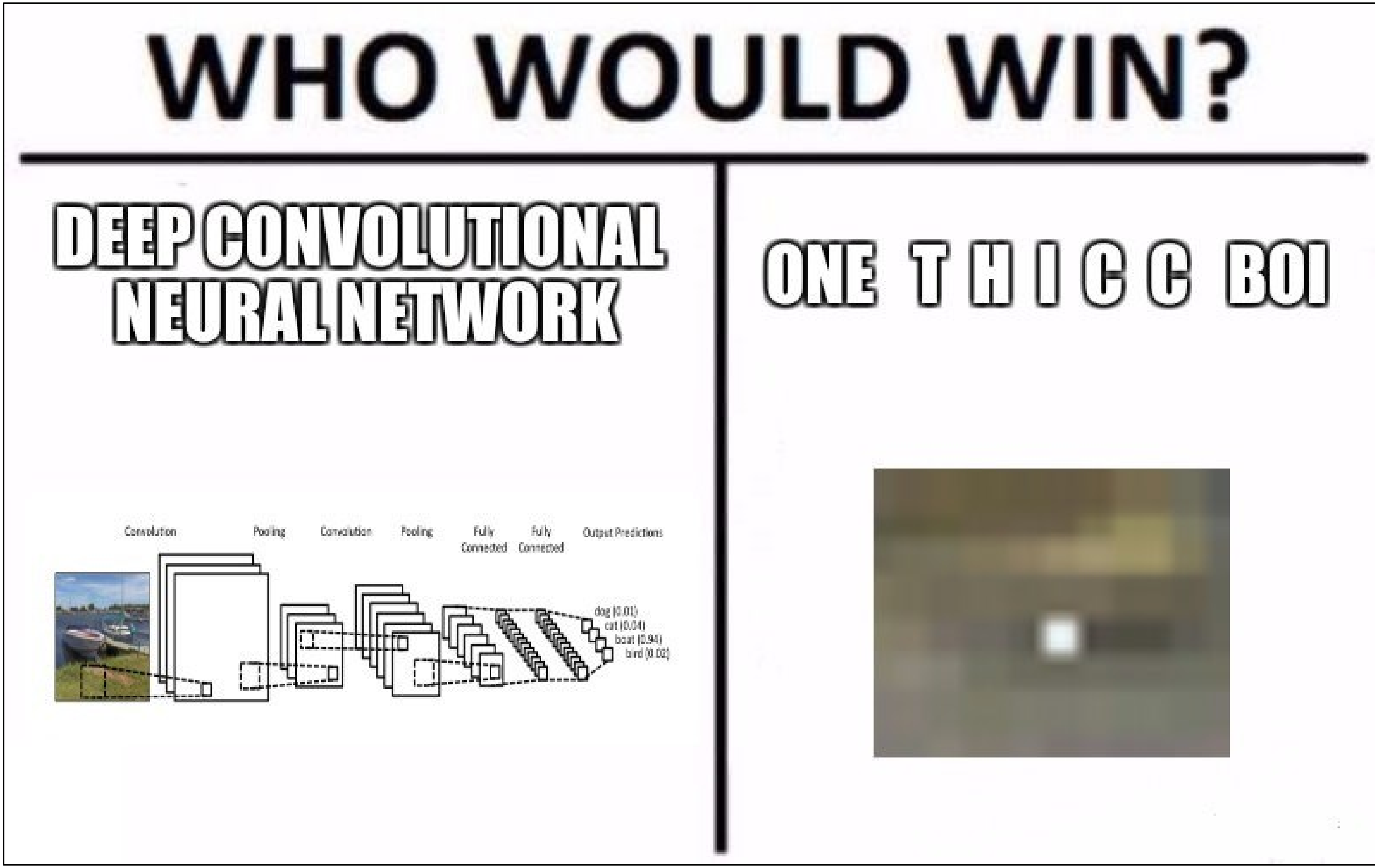
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Demonstration – Evading a competition-grade ML-PWD

- https://tinyurl.com/spacephish-demo

Adversarial Attacks against Humans and Machine Learning

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



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

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 [3]
  - 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

[3] 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!

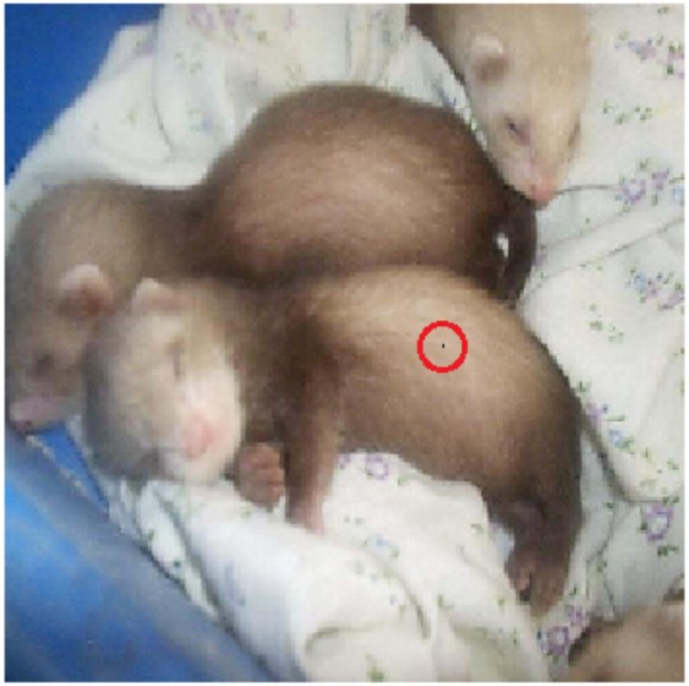
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

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)

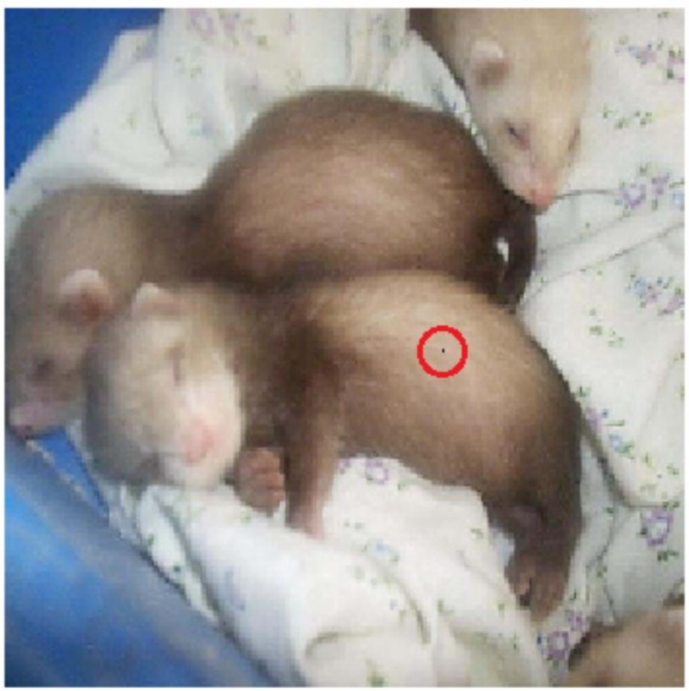
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

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 will react to clearly incorrect outputs of DL models.
  - Humans may suspect an adversarial attack taking place; or
  - They may think that the DL model is faulty, and hence not trust/believe its output
  - 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 will react to clearly incorrect outputs of DL models.
  - Humans may suspect an adversarial attack taking place; or
  - They may think that the DL model is faulty, and hence not trust/believe its output
  - 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 [4] to create adversarial examples.

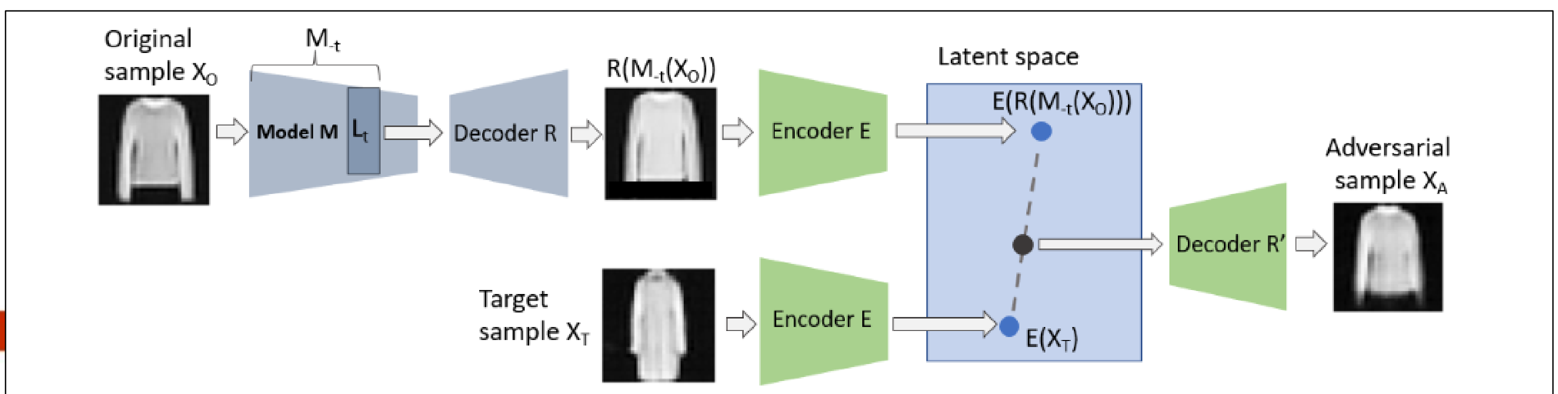
[4] J. Schneider and M. Vlachos, "Explaining neural networks by decoding layer activations," in *International Symposium on Intelligent Data Analysis*, 2021

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

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

Concept-based Adversarial Attacks

- The idea is using "explainability" techniques [4] 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 <u>can</u> 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



[4] 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, $\mathcal{O}$
- Target sample, $\mathcal{T}$
- Adversarial sample, $\mathcal{A}$

We design our experiments with three goals in mind:

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

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

3. *Remaining closer to the original sample:* the sample $\mathcal{A}$ should be similar to sample $\mathcal{O}$ 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 $\mathcal{T}$)

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

Results – Qualitative

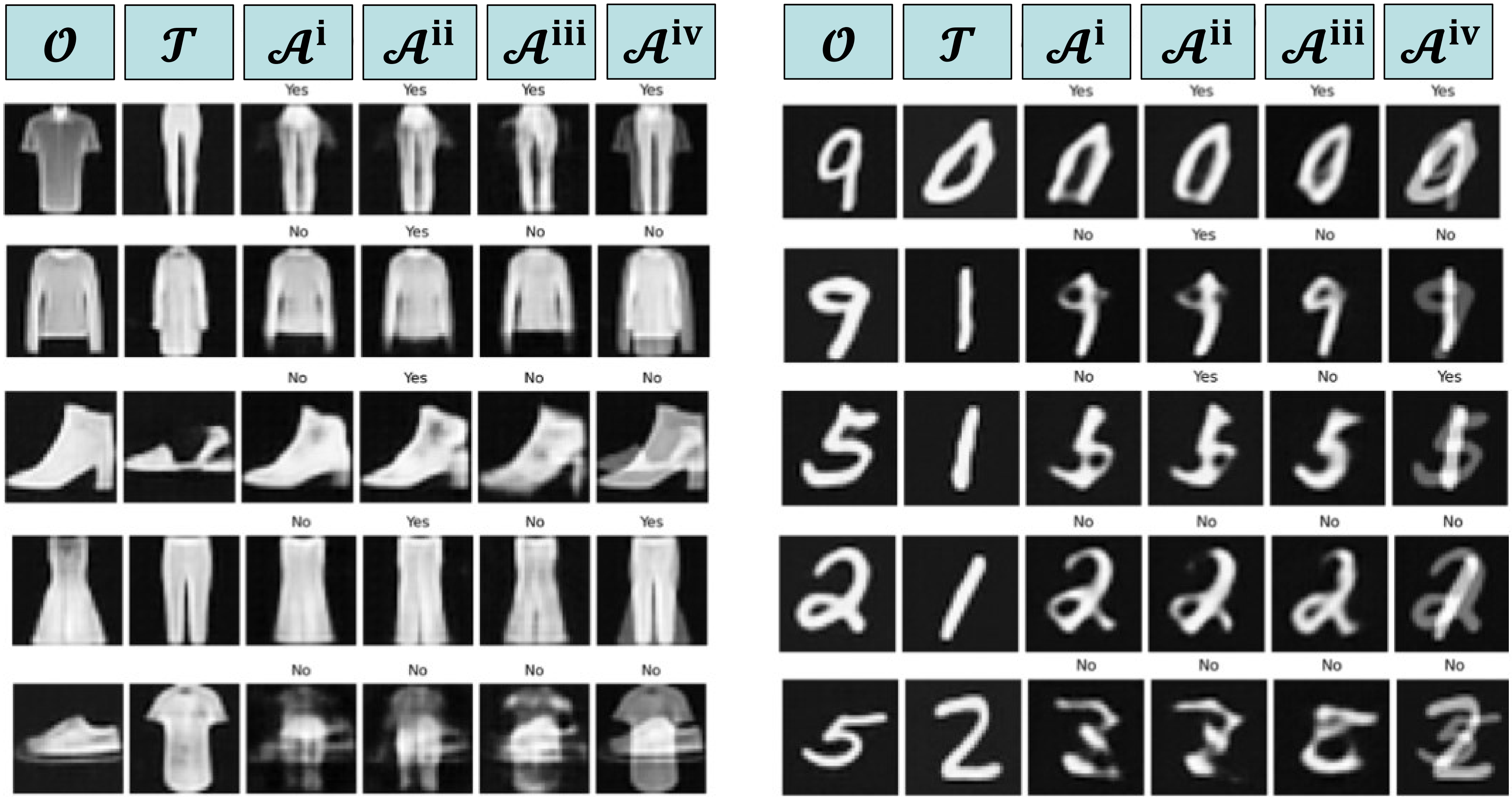


Fig. 2: Original, target and adversarial samples for different en-/decodings and interpolation for Fashion-MNIST(left) and MNIST(right). Yes/No indicates, whether the model got fooled by X_A, i.e. it outputs the class of X_T for X_A

UNIVERSITÄT
LIECHTENSTEIN

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

Results – Qualitative (takeaway)

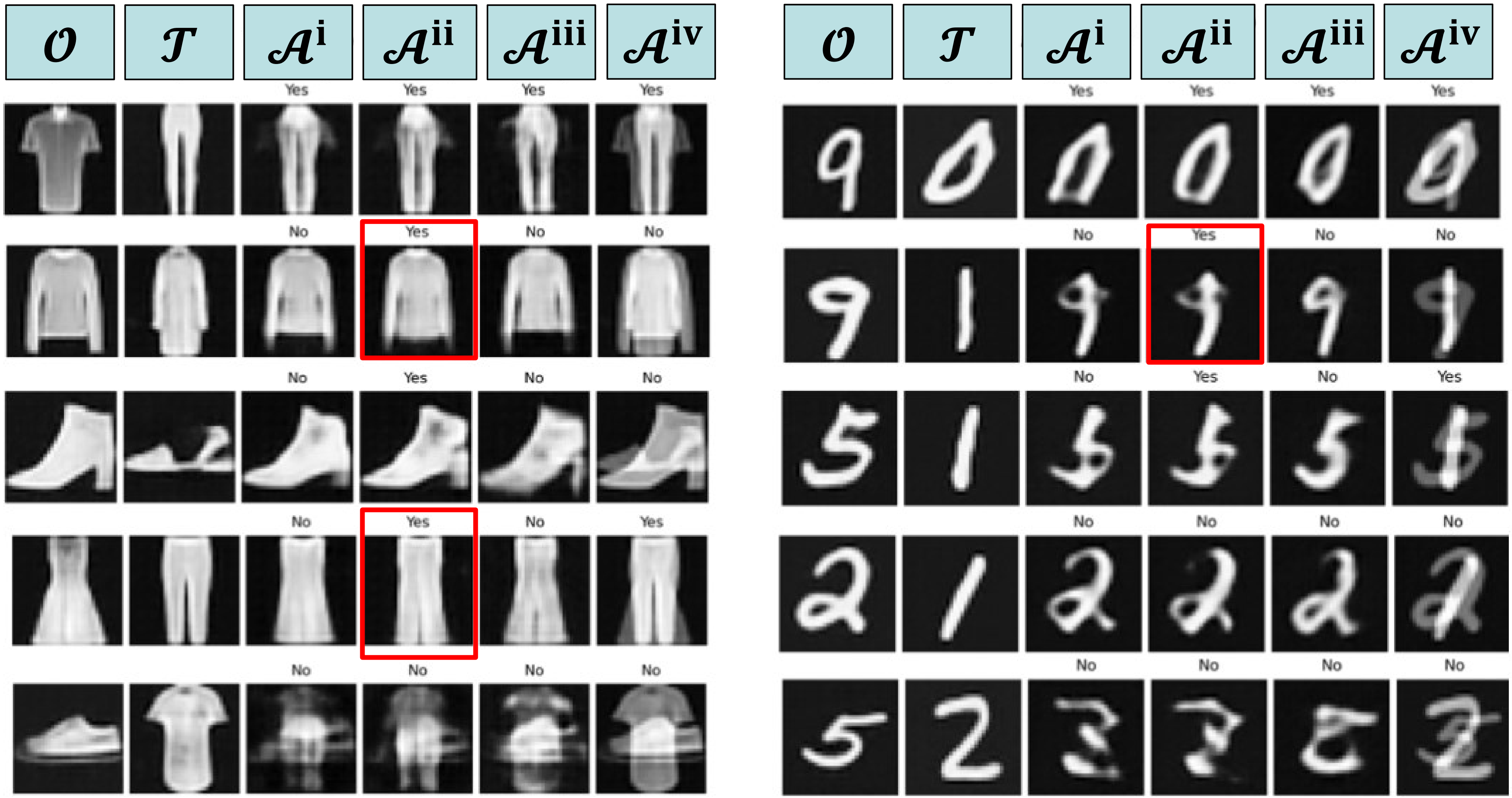


Fig. 2: Original, target and adversarial samples for different en-/decodings and interpolation for Fashion-MNIST(left) and MNIST(right). Yes/No indicates, whether the model got fooled by X_A, i.e. it outputs the class of X_T for X_A

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)	$20.83_{\pm 1.317}$	$14.95_{\pm 0.043}$	$0.42_{\pm 0.14}$	$0.44_{\pm 0.15}$	$0.41_{\pm 0.132}$

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)

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 $\mathcal{T}$

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 – confused?

Sample S

- is sample S representing a 4 or a 9?

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

UNIVERSITÄT
LIECHTENSTEIN

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

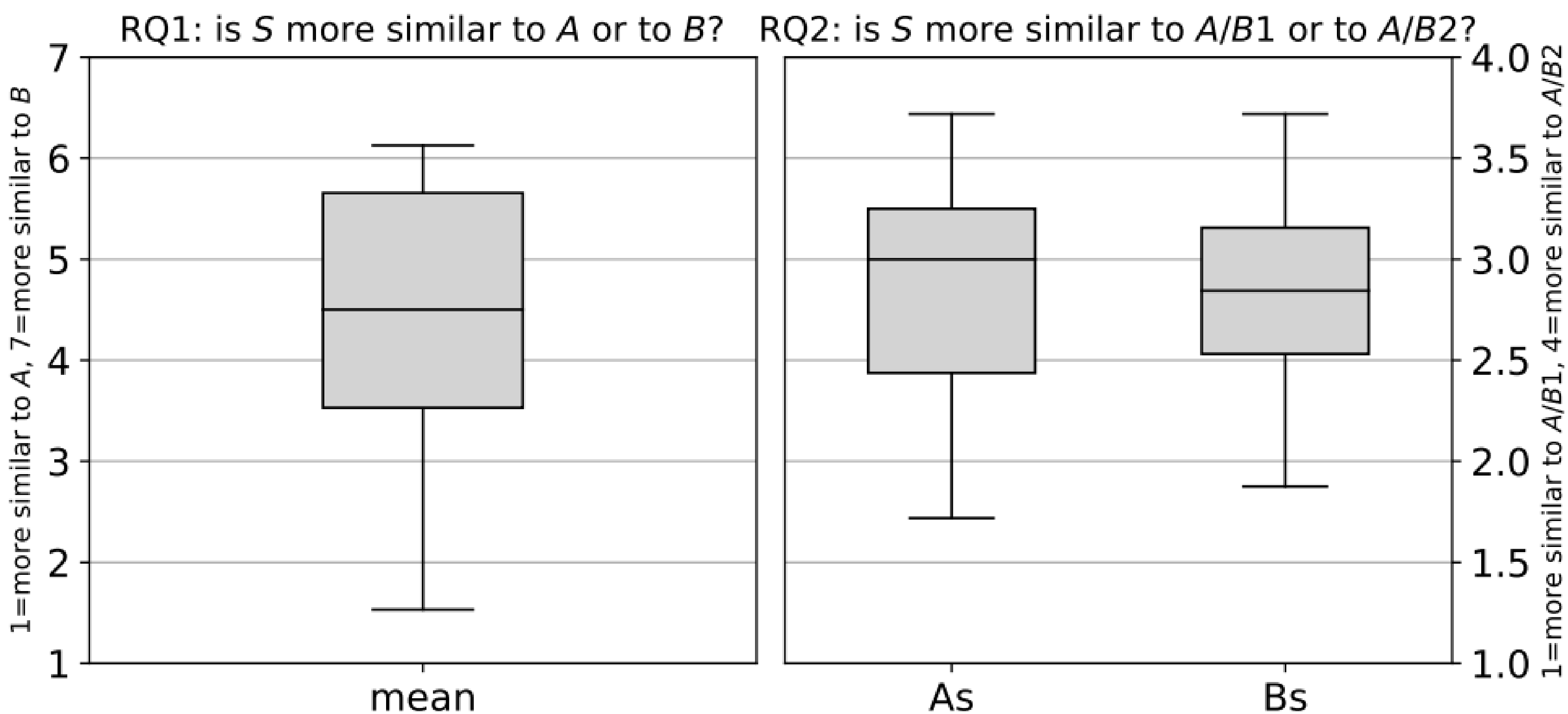
Original

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

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 papers (88) published in the top-4 cybersecurity conferences from 2019 until 2021, and see some trends…

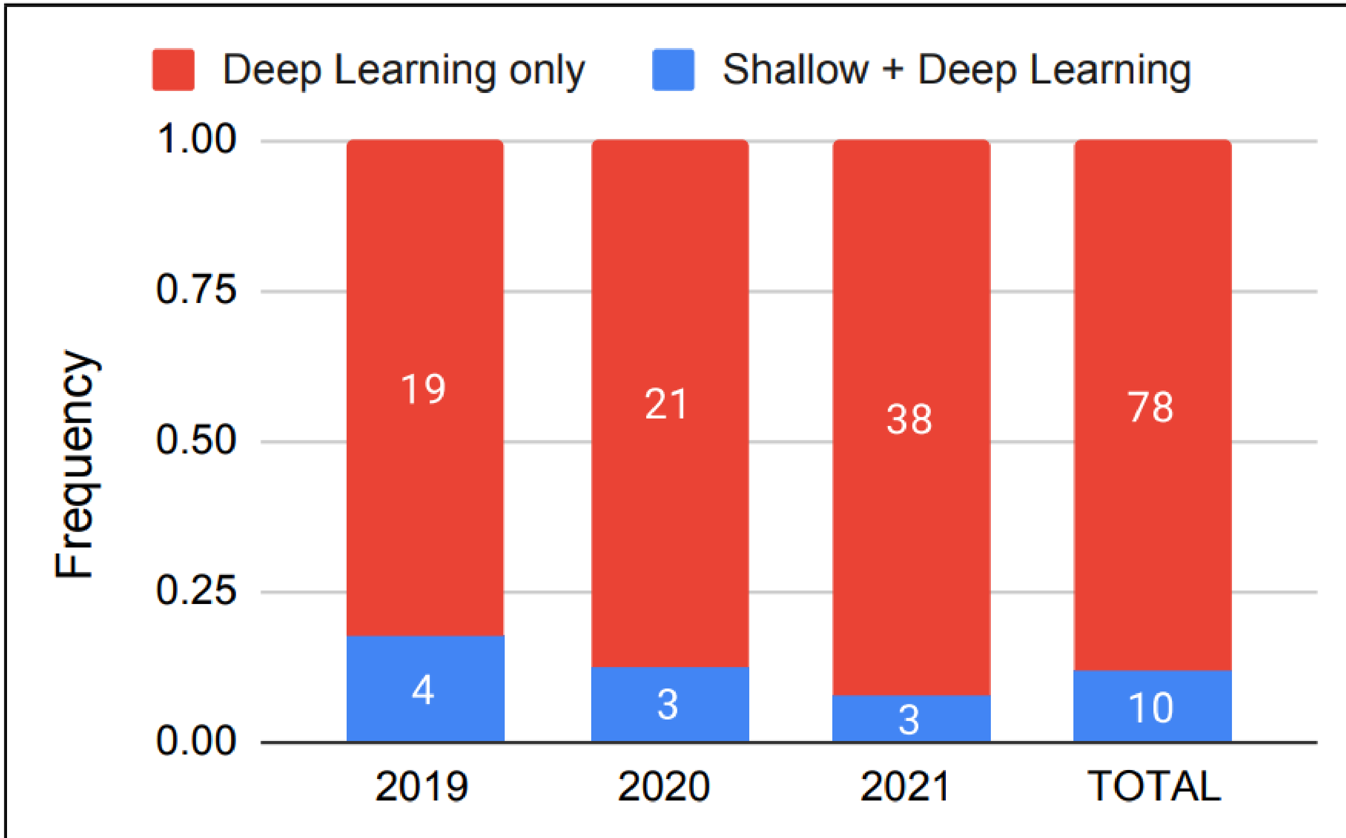


Fig. 8: G3: what is the considered ML *paradigm*?

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> papers (88) published in the top-4 cybersecurity conferences from 2019 until 2021, and see some trends…

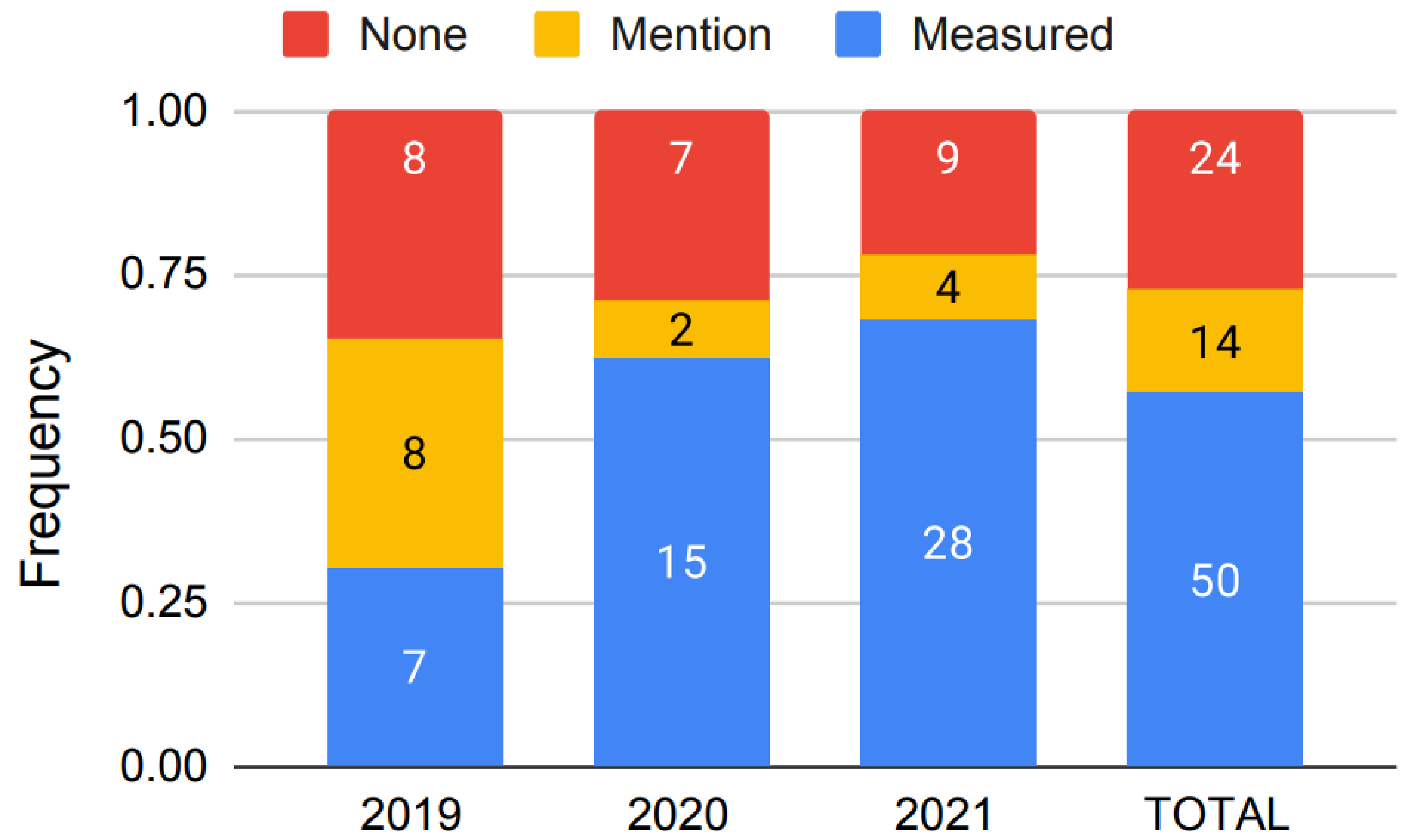


Fig. 9: G4: are the *costs* taken into account (in *any* way)?.

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

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

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

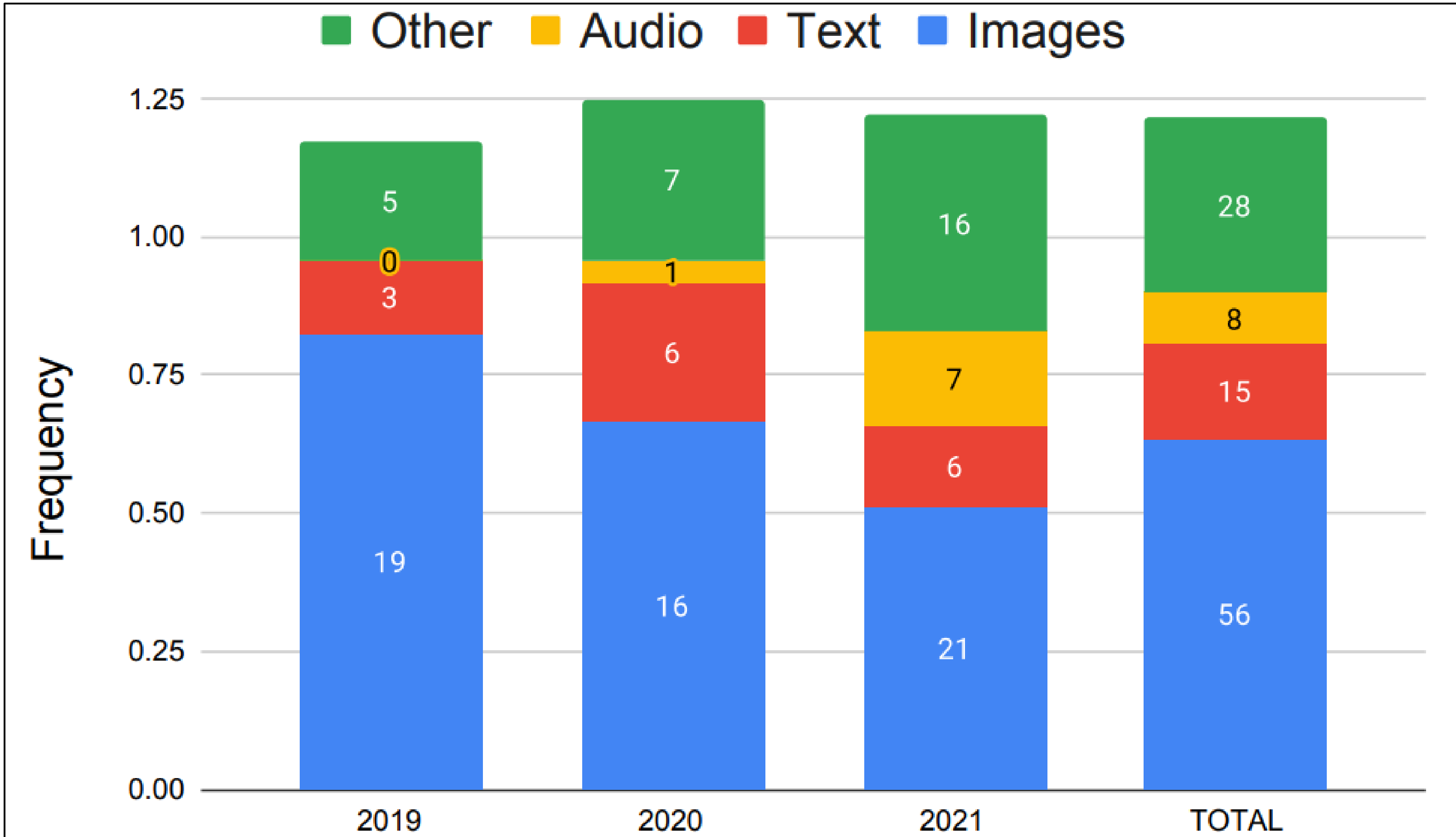


Fig. 10: What are the *data-types* considered in the evaluation?

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

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

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

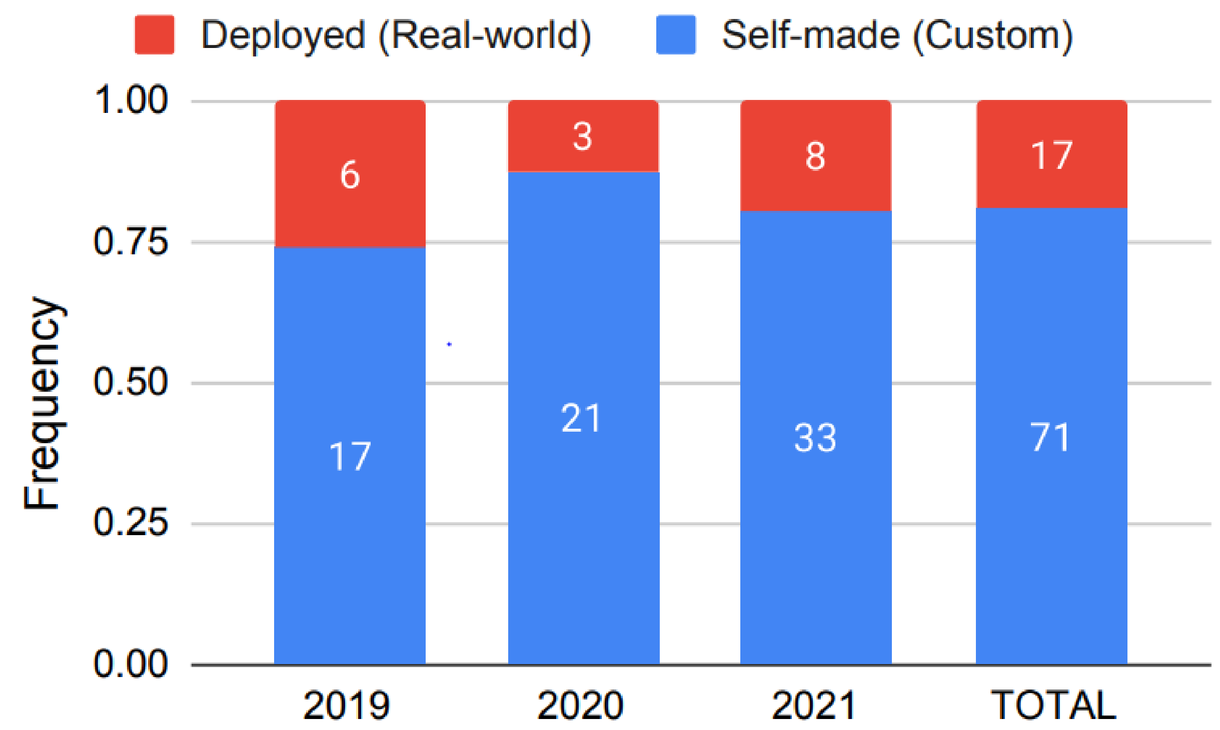


Fig. 13: Does the paper consider an ML model *deployed* in the real world?

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

Most papers attack "benchmarks"

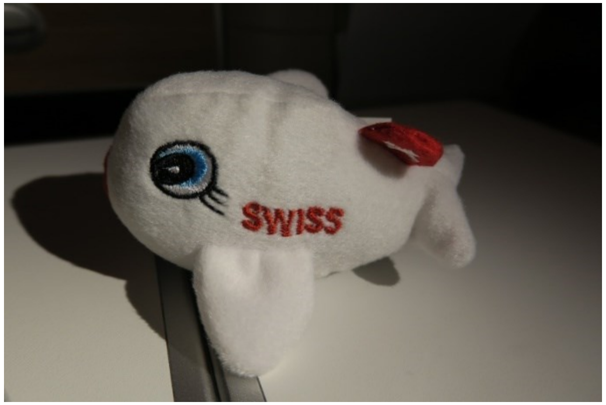
ML in practice

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

Most papers attack “benchmarks”

ML in practice

ML in research



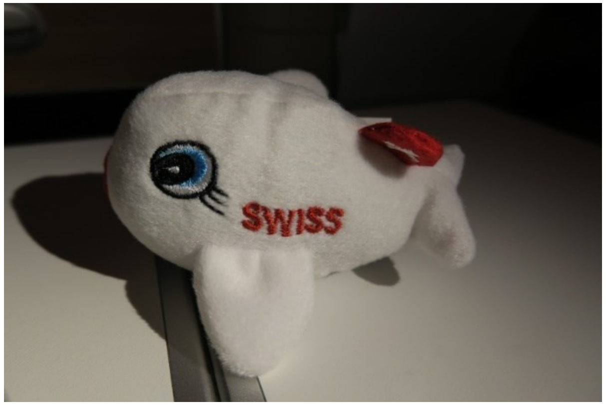
UNIVERSITÄT
LIECHTENSTEIN

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

Most papers attack "benchmarks"

ML in practice

ML in research



Question: must research papers attack "real" ML systems to have an impact to the real world?

UNIVERSITÄT
LIECHTENSTEIN

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

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

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")?

Using Machine Learning to violate the Privacy of Video Gamers

UNIVERSITÄT
LIECHTENSTEIN

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

Video Games, E-Sports, Tracking Websites, Dota2

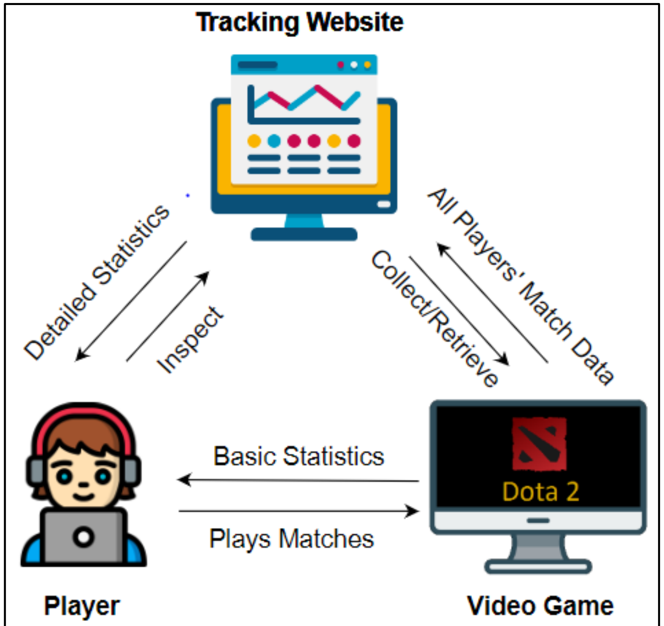
- Video Games (VG) are becoming increasingly popular
  - One of the few industries that are constantly improving their profits
- Some *competitive* VG are denoted as “E-sports”
  - Examples: Dota2, Fortnite, League of Legends
- Some tournaments of such E-sports have very high prize-pools
  - For Dota2, “The International” had a prize pool of 40M $ in 2021

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

Video Games, E-Sports, Tracking Websites, Dota2

- Video Games (VG) are becoming increasingly popular
  - One of the few industries that are constantly improving their profits
- Some *competitive* VG are denoted as “E-sports”
  - Examples: Dota2, Fortnite, League of Legends
- Some tournaments of such E-sports have very high prize-pools
  - For Dota2, “The International” had a prize pool of 40M $ in 2021

- Such prizes attract a lot of players who “play-to-win” and want to get better…
  - Best way of improving at something? Learn from past mistakes!
- …which, in the E-sport ecosystem, it can be easily done via Tracking Websites



UNIVERSITÄT LIECHTENSTEIN

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

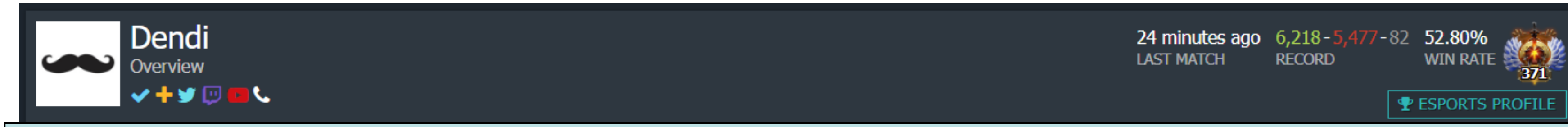
A tracking website (TW)

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

A tracking website (TW)

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

A tracking website (TW) – Why is it public?

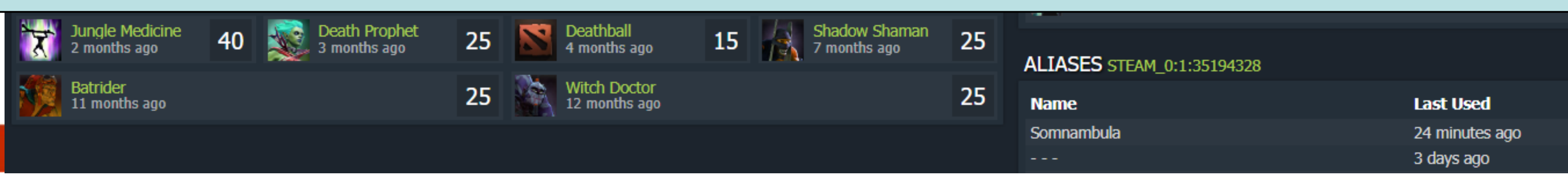


It is the playerbase who want the statistics collected by TW to be publicly available!

The reasons are various, e.g.,:

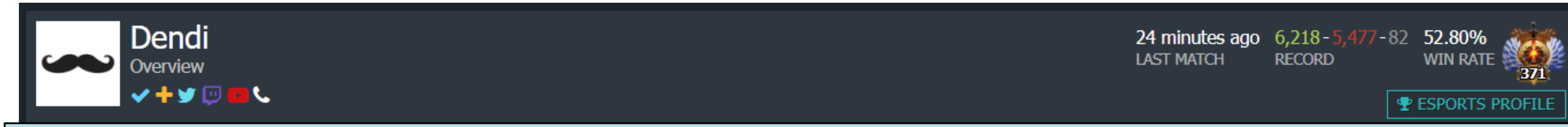
1. Inspecting the profiles of *other* players can be used to learn some of their tricks…
2. …in turn, by having their own profile publicly accessible, a given player can gain visibility if they perform very well…
3. …such "visibility" can lead to invitations to play in top-teams, or to finding new (good) teammates
4. The visibility can come either because other players "inspect" a given player's profile, or because of climbing "public ladders"

There are over 70M of Dota2 players who use TW.



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

A tracking website (TW) – Why are they A PROBLEM?



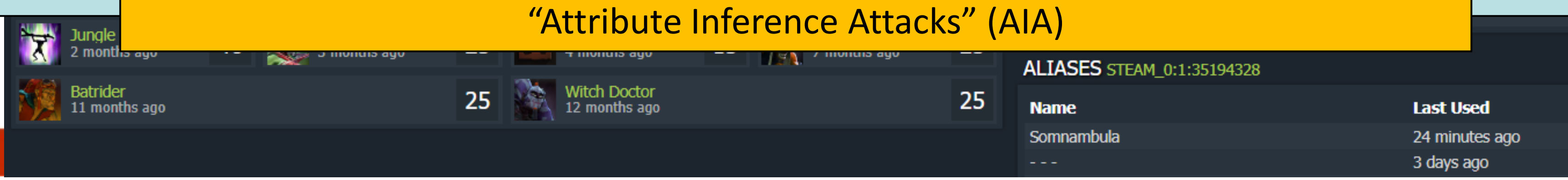
It is the playerbase who want the statistics collected by TW to be publicly available!

The reasons are various, e.g.,:

1. Inspecting the profiles of *other* players can be used to learn some of their tricks…
2. …in turn, by having their own profile publicly accessible, a given player can gain visibility if they perform very well…
3. …such "visibility" can lead to invitations to play in top-teams, or to finding new (good) teammates
4. The visibility can come either because other players "inspect" a given player's profile, or because of climbing "public ladders"

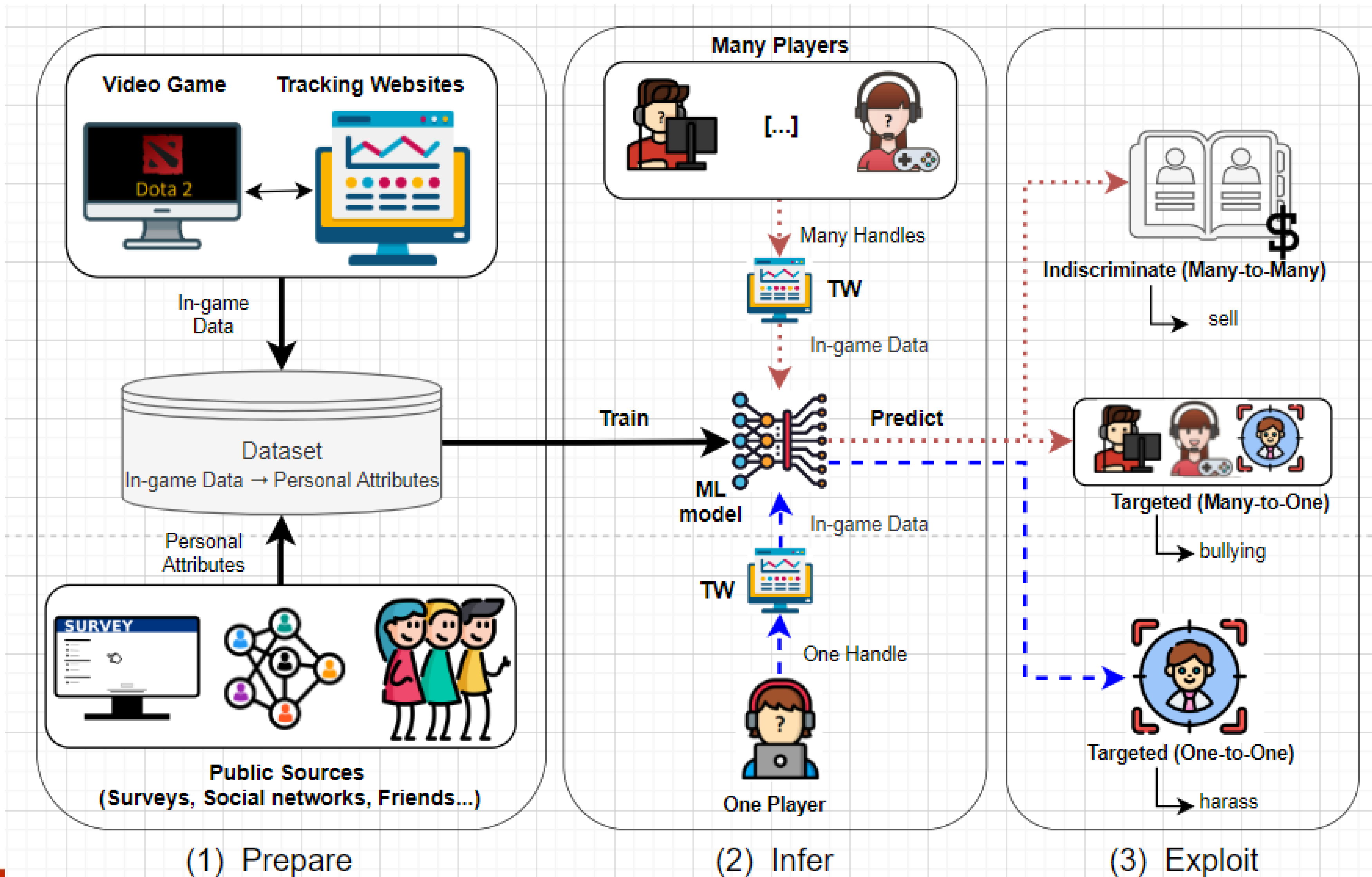
There are over 70M of Dota2 players who use TW.

Problem: such "availability" exposes E-sports' players to the risk of "Attribute Inference Attacks" (AIA)



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

Threat Model



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

Assessment

- We proactively assess such a threat, because *nobody* ever did something similar in the E-sports ecosystem. We focus on Dota2

- We conduct an informed survey, asking ~500 Dota2 players to provide us with private (non-sensitive) information about their real-life (e.g., age, gender, occupation, whether they buy Dota2 content, and some personality traits)

- We use the handle (i.e., nickname) of such players to collect their (publicly available) Dota2 in-game statistics from popular TW (opendota).

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

Assessment (cont'd)

- We proactively assess such a threat, because *nobody* ever did something similar in the E-sports ecosystem. We focus on Dota2

- We conduct an informed survey, asking ~500 Dota2 players to provide us with private (non-sensitive) information about their real-life (e.g., age, gender, occupation, whether they buy Dota2 content, and some personality traits)

- We use the handle (i.e., nickname) of such players to collect their (publicly available) Dota2 in-game statistics from popular TW (opendota).

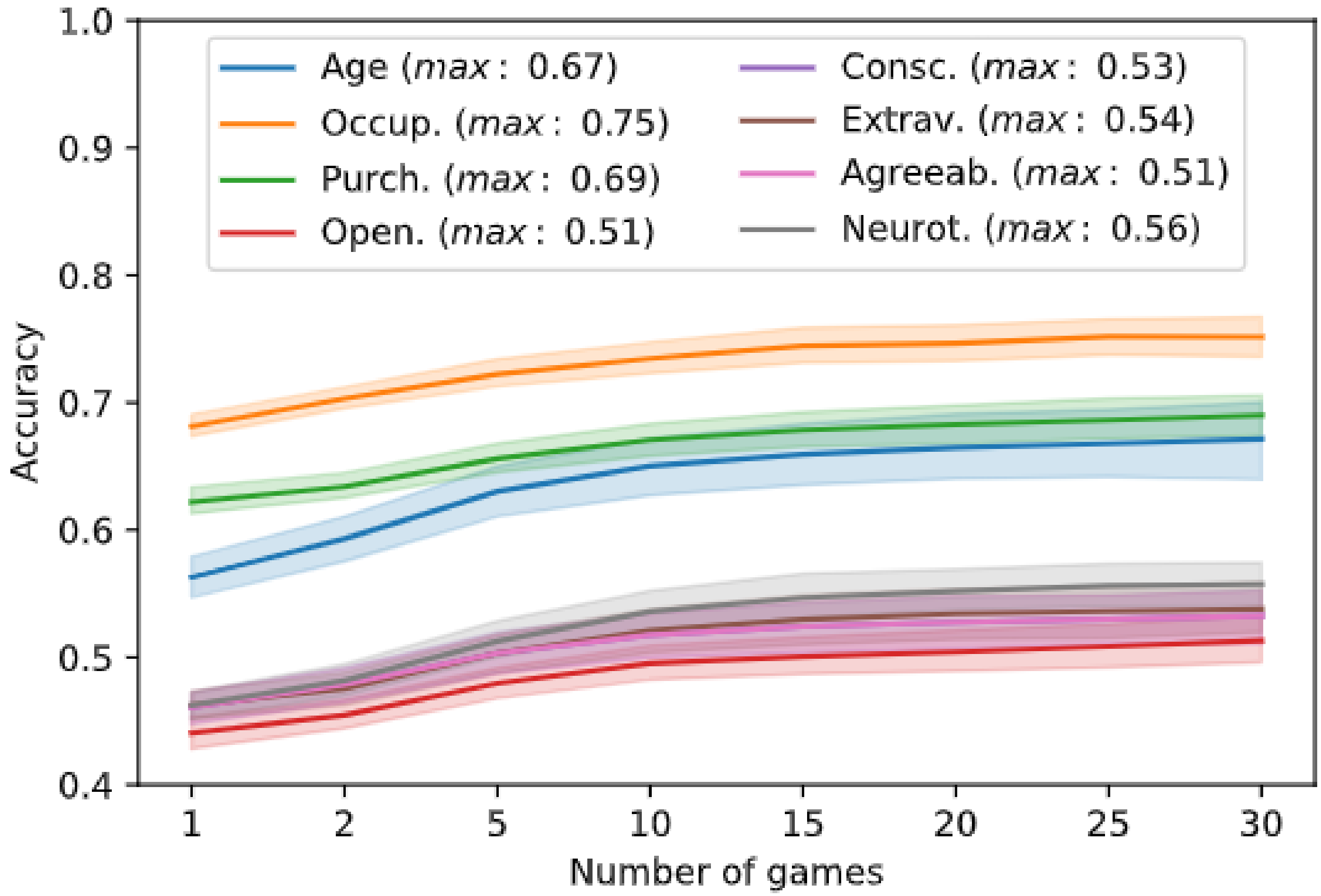
- We **find a correlation** (!) between the players in-game statistics and their real life.
    - Such a finding suggests that AIA can be successful!

- We (ethically) perform diverse AIA: we use 80% of our data to train ML models, and predict the personal attributes of the players included in the remaining 20%.

UNIVERSITÄT LIECHTENSTEIN

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

Results – One-to-One AIA



UNIVERSITÄT
LIECHTENSTEIN

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

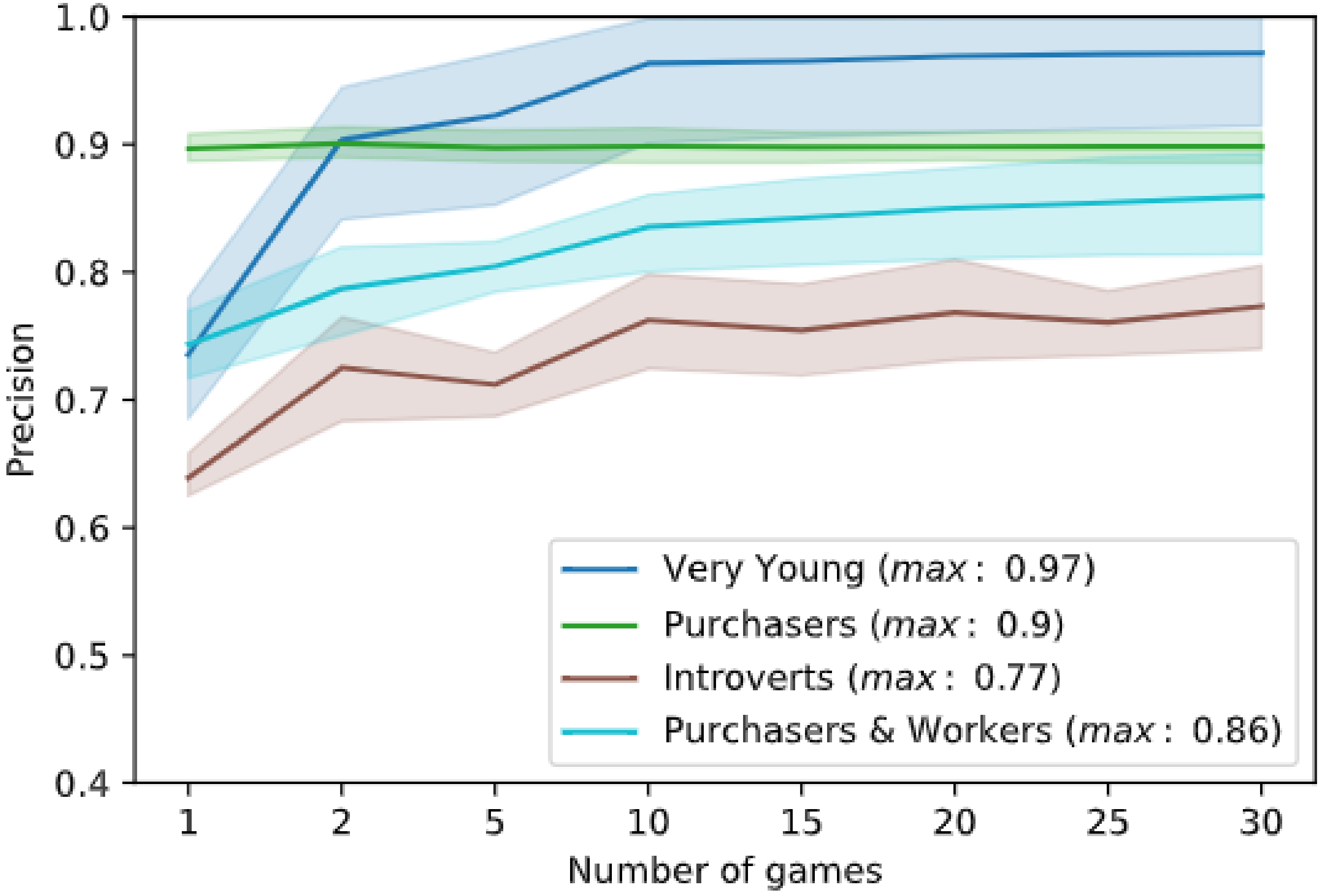
Results – Many-to-Many AIA

Table 6: Indiscriminate 'many-to-many' AIA (mid column). Compared to the baseline (cf. Fig. 5), the accuracy substantially increases.

	Sophisticated AIA (30 matches)	Indiscriminate AIA (30 matches)	Improvement
age	67.15±6.87	**89.15**±4.66	+22.00%
purch.	68.99±3.81	**96.13**±2.86	+27.14%
open.	51.30±3.87	**77.86**±3.39	+26.56%
consc.	53.24±4.88	**80.19**±4.12	+26.95%
extrav.	53.78±3.90	**81.51**±4.40	+27.73%
agreeab.	50.71±4.65	**76.84**±5.59	+26.13%
neurot.	55.74±3.88	**80.64**±4.02	+24.90%

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

Results – Many-to-One AIA



UNIVERSITÄT
LIECHTENSTEIN

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

…so what now?

- **Hard counters?** Nope!
  - The entire E-sport ecosystem would be disrupted

- **Compromise?** Yes!
  - The users should be informed that having their in-game statistics to be publicly accessible by TW exposes them to AIA

- **What about other games?** Many E-sports share the same ecosystem with Dota2
  - AIA are theoretically possible also in other VG, but a correlation has to be found first

- **We sent an email to Valve** (yesterday) to inform them of such vulnerability.
  - We are unsure about whether they will take any action in the short-term

Cybersecurity and Machine Learning: Facts and Myths

Giovanni Apruzzese, PhD

University of Bologna – October 12th, 2022

UNIVERSITÄT LIECHTENSTEIN